Genetic linkage mapping in complex pedigrees

Stuart Macgregor BSc MSc

Thesis submitted for the degree
of
Doctor of Philosophy

The University of Edinburgh
2003

Declaration

I declare that this thesis is the product of my own efforts and has not been submitted in any previous application for a degree. The research on which it is based is my own, except where stated in the text.

Stuart Macgregor

Abstract

Genetic linkage analysis is the primary method for the identification of loci contributing to complex disease susceptibility. Linkage analysis techniques can be applied to both disease status (discrete traits) and to quantitative trait measures (quantitative trait loci or QTL mapping). Such techniques will be most effective if they can be applied to all of the available data; in human, ecological and livestock genetics this often means families with complex pedigree structures. The analysis of complex pedigrees is more difficult, both in terms of model formulation and computational ease, than similar studies of small family structures such as affected sibling pairs (ASP). Univariate variance component (VC) techniques suitable for QTL analysis of both quantitative and qualitative (via a threshold model) traits are described. Extensions to the univariate VC methods are proposed, allowing QTL analyses of longitudinal data in complex pedigrees, with polynomial based covariance functions offering a parsimonious description of the covariance structure across measures. Computer simulations are used to show that, under a range of realistic scenarios, the longitudinal QTL method offers more power to detect QTL than univariate or repeated measures methods. The longitudinal method is subsequently applied to 330 extended families from the Framingham Heart Study, allowing the identification of QTL for a number of cardiovascular disease risk factors. The maximum LOD score (3.12) is obtained on chromosome 16 for Body Mass Index (BMI) and subsequent multivariate analyses showed that this QTL is most relevant to BMI at early ages. Threshold model based VC and parametric linkage analyses are applied to a set of Scottish families affected by psychiatric disease. The results from this analysis are in agreement with previous results implicating chromosome 1q42 in psychiatric disease susceptibility. The broad application of the VC techniques is further demonstrated by applying the techniques to a QTL mapping problem in a very large Red Deer (*Cervus Elaphus*) pedigree.

Linkage analysis is commonly used to identify candidate regions for further study. These candidate regions will be the chromosomal segments shared among related individuals with common diagnoses, with recombination events delineating the regions of interest. However in genetically complex traits, the relationship between phenotype and genotype is not one to one. The effect of changing the parameters defining the relationship between phenotype and genotype is investigated, both analytically and by computer simulation. Increasing the rate at which affected individuals without mutations in the disease region of interest occur in the sample (the phenocopy rate) is found to have a large

effect on the validity of the inferred region. This has implications in genetic studies of common disease (e.g. schizophrenia), where the phenocopy rate will often be non-zero.

The use of extended families for linkage mapping has become a controversial issue, with the field of psychiatric genetics somewhat polarised; a number of groups have collected mainly extended family data whilst others have focused on small ASP family structures. Whilst the advantages of a given study design will vary depending upon the unknown 'true' disease model, it is argued that extended families will often be more useful for locus identification than sib pair based studies. It is shown that the heterogeneity introduced by collecting large numbers of sib pairs from a number of different populations will impact significantly upon the power to detect the effects of any single gene.

Acknowledgements

I would like to thank my supervisors Peter Visscher, Sara Knott and Douglas Blackwood for their continued support throughout the last three years. Peter and Sara were particularly helpful in the final stages of writing up and Douglas' enthusiasm for genetics and psychiatry greatly aided my move into psychiatric genetics.

A debt of gratitude is owed to my sponsor company Organon for providing funding. Thank you also to the BBSRC for providing funding.

Thanks go to Ian White for reading some of the chapters and for general guidance on statistical and computational matters. Thanks also to Robin Thompson for ASREML advice. Some of this work was done in collaboration with the group at the Western General Hospital and the Royal Edinburgh Hospital; particular thanks go to David Porteous, Walter Muir, Kathy Evans, Kirsty Millar, Pippa Thompson and Maura Walker. Thanks also to Josephine Pemberton and Jon Slate for the opportunity to participate in some collaborative work on natural population data.

My PhD studies were made substantially easier as a result of the support of my friends and family. Particular thanks go to my mother for her continued support over my protracted student career. Thanks to Tim, Dan and Dan for great coffee times and random web nonsense and to my non-science friends for many good times.

I dedicate this thesis to my wife Jeanette. Her constant encouragement and love made this thesis possible.

Publications

The following publications have resulted as a direct outcome of the research described in this thesis:

1. S. Macgregor, P. M. Visscher, S. Knott, D. Porteous, W. Muir, K. Millar, and D. Blackwood. Is schizophrenia linked to chromosome 1q? Science, 298:2277a, 2002.

2. S. Macgregor, S.A. Knott, I. White, and P.M. Visscher. Longitudinal analysis of the Framingham data. BMC Genet, in press, 2003.

3. S. Macgregor, P. Visscher, S. Knott, P. Thompson, K. Millar, D. Porteous, W. Muir, and D. Blackwood. Schizophrenia, bipolar disorder and chromosome 1 linkage. Am. J. Med. Genet., 114:O28, 2002.

4. S. Macgregor, S.A. Knott, I. White, and P.M. Visscher. XIX International congress of genetics. In Genomes > The linkage to life, page 218, 2003.

5. J. Slate, P. M. Visscher, S. MacGregor, D. Stevens, M. L. Tate, and J. M. Pemberton. A genome scan for quantitative trait loci in a wild population of red deer (*Cervus elaphus*). Genetics, 162:1863-1873, 2002.

6. W.J. Gauderman, S. Macgregor, L. Briollais, K. Scurrah, M. Tobin, T. Park, D. Wang, S. Rao, S. John, and S. Bull. Longitudinal data analysis in pedigree studies. BMC Genet, in press, 2003.

7. R. Segurado, S. D. Detera-Wadleigh, D. F. Levinson, C. M. Lewis, M. Gill, J. I. Nurnberg, N. Craddock, J. R. DePaulo, M. Baron, E. S. Gershon, J. Ekholm, S. Cichon, G. Turecki, S. Claes, J. R. Kelsoe, P. R. Schofield, R. F. Badenhop, J. Morissette, H. Coon, D. Blackwood, L. A. McInnes, T. Foroud, H. J. Edenberg, T. Reich, J. P. Rice, A. Goate, M. G. McInnis, F. J. McMahon, J. A. Badner, L. R. Goldin, P. Bennett, V. L. Willour, P. P. Zandi, J. J. Liu, C. Gilliam, S. H. Juo, W. H. Berrettini, T. Yoshikawa, L. Peltonen, J. Lonnqvist, M. M. Nothen, J. Schumacher, C. Windemuth, M. Rietschel, P. Propping, W. Maier, M. Alda, P. Grof, G. A. Rouleau, J. Del-Favero, C. Van Broeckhoven, J. Mendlewicz, R. Adolfsson, M. A. Spence, H. Luebbert, L. J. Adams, J. A. Donald, P. B. Mitchell, N. Barden, E. Shink, W. Byerley, W. Muir, P. M. Visscher, S. Macgregor, H. Gurling, G. Kalsi, A. McQuillin, M. A. Escamilla, V. I. Reus, P. Leon, N. B. Freimer, H. Ewald, T. A. Kruse, O. Mors, U. Radhakrishna, J. L. Blouin, S. E. Antonarakis, and N. Akarsu. Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. Am. J. Hum. Genet., 73:49-62, 2003.

Contents

Chapter 1

Introduction

1.1 Identifying genetic variation underlying human disease

Many of the diseases that affect human populations are known to be subject to a degree of genetic control [38]. The World Health Organisation (WHO) places human diseases into two broad categories: one covering communicable disease (such as infectious and parasitic diseases) and one covering non-communicable disease (such as cancers and cardiovascular disease (CVD)) (`http://www3.who.int/whosis/menu.cfm?path=evidence,burden,burden_estimates`). In the developed world, the vast majority (~80% of all deaths) of the disease burden is a result of non-communicable disease; this is due in large part to superior health care and nutrition. Since the effects of communicable disease are small, in Europe and North America, cancers and CVD account for over two-thirds of all deaths and are the subject of intense research. Although many environmental risk factors have been identified, there is a (substantial) genetic component to most cancers [191] and to heart disease (e.g. [145], see also section 1.2.2 and chapter 4). Future research into reducing the impact of these diseases has therefore focused on the identification of the underlying disease genes. In particular, the pharmaceutical industry has invested heavily in genetics/genomics, with the hope being that an understanding of the genetic components of disease will lead to the identification of novel drug targets. In the developing world, the disease burden is mainly a result of communicable disease (~75% of all deaths): many of these deaths could be avoided with the provision of suitable health care and nutrition. If the effects of communicable disease can be reduced, genetic research may also have a substantial impact on public health in the developing world.

1.1.1 Mendelian Disease

The simplest diseases genetically are those that arise as a direct result of the genotype an individual has at a single (disease) locus. These are known as *Mendelian* diseases. For

such diseases a particular inheritance pattern can be clearly seen in families. If the (discrete) character or disease is expressed in individuals who have one or two copies of the disease allele, the character is *dominant*. Assuming the presence of the character does not affect the parents decision to have further children (ascertainment bias), the inheritance pattern observed in families will be distinctive, with half of the offspring of an affected parent exhibiting the character (assuming that the affected parent is heterozygote and the other parent is unaffected). When the character is only expressed in homozygotes, the character is *recessive* and again assuming no ascertainment bias, one quarter of the offspring of two heterozygous parents will exhibit the character. There are over 14000 Mendelian characters known in humans, with over 8000 mapped to a particular chromosome (`http://www.ncbi.nlm.nih.gov/omim/`). Techniques for mapping genes to particular chromosomal regions are described later in the chapter.

With the advent of molecular marker technology there has been an explosion of interest in the identification of genes underlying human Mendelian disease. In the initial stages of this era, marker information was used to test for phenotypic-genotypic correlations in genomic regions thought to play a biological role in a particular disease, the *candidate gene* (or functional cloning) paradigm. Subsequently, marker technology became sufficiently inexpensive to allow coverage of the whole genome in the search for genetic determinants, the *positional cloning* paradigm [42, 43]. The ready availability of genetic markers for this genome scan approach has allowed researchers to find more than 1000 genes associated with human disease [168]. In the vast majority of cases genomic regions have been found by first examining individuals in pedigrees and applying linkage analysis techniques. Subsequent to a successful linkage analysis, linked regions have often been narrowed (or *fine mapped*) by techniques such as Linkage Disequilibrium (LD) mapping (e.g. Cystic Fibrosis, Nijmegen breakage syndrome [220]). Linkage analysis (discussion in detail in section 1.3.1) relies upon the co-segregation of disease loci with nearby linked marker loci. LD (discussed in section 1.4) is a population level phenomenon, describing the degree to which the frequency of two alleles (at two different loci) differs from the expected frequency assuming they occurred independently. LD may exist between nearby loci if the alleles in two 'unrelated' individuals occur together (i.e. in a haplotype, see section 1.3.1) as a result of transmission from a common ancestor; these haplotypes, whittled down by many generations of recombination events, may be useful for the fine mapping of QTL.

A classic example of the efficacy of positional cloning in Mendelian disease mapping was in the recessively inherited disease Cystic Fibrosis (CF, OMIM 219700). Linkage analysis was used to map CF to chromosome 7q [234]. Although this result strongly implicated this region, due to the small number of recombination events available within the affected families, it was not possible to directly identify the gene responsible. LD based mapping techniques were then applied, allowing the identification of the gene responsible for CF [20]. This led to a significant increase in the understanding of CF pathogenesis. In the near future, techniques based on knowledge of the underlying biology, such as gene therapy, may allow more effective treatment of CF.

Diseases with Mendelian inheritance tend to be rare because the strong phenotypic effect of the disease mutation on fitness will ensure that the mutations will be rapidly purged from the population (assuming that the disease decreases fitness). Exceptions to this include cases such as CF where heterozygous advantage (i.e. where individuals heterozygous for the disease allele have greater fitness than the wild type homozygote) maintains it the disease at a frequency of about 1 in 2000 in the UK population. Similarly, the incidence of Duchenne Muscular dystrophy is maintained by high levels of recurrent mutation and late onset diseases such as Huntington disease only present after reproductive age [220].

1.1.2 Complex Disease

Given the successes with Mendelian disease, the linkage analysis -> LD analysis procedure was extended to deal with traits in which there is not a one to one correspondence between genotype and phenotype. Such traits are referred to as *complex traits*. Complex traits will be caused by a multitude of genetic factors; this may involve a number of genes of moderate effect, *oligogenes*, or a large number of genes of small effect, *polygenes*. The genes involved may only cause disease in the presence of a particular set of background genes or environmental circumstances. There may be genetic heterogeneity, with different populations having substantially different distributions of disease genes. There are two main types of genetic heterogeneity; *locus heterogeneity* and *allelic heterogeneity*. Locus heterogeneity refers to the situation where there are multiple disease susceptibility loci segregating; any single locus may contribute to disease susceptibility in a particular sub-population but be unnecessary for disease in another sub-population. Non-syndromic deafness is the classic example of locus heterogeneity with over 60 distinct loci reported to date [171]; in this case each individual mutation at one of the distinct loci is sufficient to cause the disease. For complex diseases, the locus heterogeneity model may be less clear cut, with individual loci only increasing the risk of disease, with affection determined as a result of a number of genetic and environmental factors. Allelic heterogeneity refers to the situation where there are multiple alleles at a single locus, each of which may be sufficient to cause disease. CF is the classic example of a disease exhibiting allelic heterogeneity, with some variants only arising in isolated populations [120]. Due to the effects of these multiple genetic and environment factors, progress in disease gene mapping of complex disease has been substantially slower than that observed in genetic studies of Mendelian disorders.

Complex traits commonly targeted for genetic analysis include schizophrenia, bipolar disorder, hypertension, cancer and diabetes. In terms of public health these common (defining common as being >1 case per 1000 people [231]) complex diseases are substantially more important than the numerous, but rare, Mendelian diseases. Even in cases in which there is a mechanism for the maintenance of a disease mutation in the population (for example as in the case of CF), Mendelian diseases rarely become common.

The potential of the positional cloning approach for a common disease gene mapping

has been been demonstrated in the case of Crohn's disease [106]. Crohn's disease is common (incidence ~1 in 1000), causing chronic inflammation of the gastro-intestinal tract. It is thought to result when particular environmental factors arise in genetically predisposed individuals [106]. In this case researchers were able to follow up a positive signal from linkage analysis with LD based analysis and, after some luck with sequencing a region only weakly associated with the disease, they were able to identify the variant conferring susceptibility to the disease.

To improve the efficacy of the positional cloning approach in complex disease various refinements have been utilised. Some of the successes have relied upon the identification of genes for particularly extreme forms of a disease. This approach is based upon the expectation that such subsets may be more genetically homogeneous (less *phenotypic heterogeneity*). A case in point is Alzheimer's disease; analysis of early onset cases allowed the identification of mutations causing the deposition of amyloid β peptides in plaques in the brain, leading to greater understanding of disease pathogenesis [94]. Another possible way forward involves identifying subsets of the disease which have near-Mendelian inheritance patterns; this approach has been successful in the identification of loci for breast cancer (genes BRCA1; OMIM 113705 [157], BRCA2; OMIM 600185 [257]).

Another promising approach for complex trait dissection involves finding pedigrees in which chromosomal abnormalities segregate with the phenotype of interest. One application of this approach considered a balanced chromosome 1;11 translocation in a Scottish family affected by major psychiatric disease [216]. Subsequent linkage analysis using the translocation as a marker generated a substantial test statistic, illustrating that it was very likely that the translocation interrupted a gene conferring susceptibility to psychiatric disease [26] (see also section 1.2.1).

1.2 Primary Research Areas

1.2.1 Psychiatric genetics

One of the foci of this thesis is psychiatric genetics. Mental health problems such as depression, anxiety and schizophrenia account for 12 percent of the United Kingdom National Health Service budget (`http://www.doh.gov.uk/dohreport/report2000/dr2000-11.html`). Twin studies comparing trait incidence in monozygotic and dyzygotic twins have shown that diseases such as schizophrenia have a strong genetic component [242]. The proportion of variance attributable to genetic factors is generally estimated at around 80% in schizophrenia and bipolar disorder [148]; this proportion is lower, but still substantial, for disorders such as anxiety (~35% [101]) and (unipolar or recurrent major) depression (~50% [147]). These proportions are particularly high when one considers the difficulty in unambiguously diagnosing these disorders. Interestingly, it has been suggested that disorders such as schizophrenia and bipolar disorder, which are generally regarded as clinically distinct, may share susceptibility genes [22, 24, 26]. The evidence

for a strong genetic component, together with suitable marker data, has led researchers to devote substantial resources to identifying the genes responsible for susceptibility, with the hope that knowledge of these genes will improve understanding of the pathophysiology of diseases such as schizophrenia and lead to more effective treatment. Even if a single gene, perhaps of minor effect on schizophrenia, could be unambiguously identified, the biochemical pathways and molecular mechanisms suggested by this gene might prove to be of relevance to the disorder in general [96].

When faced with the difficult task of identifying genes for complex disease appropriate study design is crucial. In psychiatric genetics there are two popular designs for linkage analysis. The first is based upon collections of affected sibling pairs (ASP). The second concentrates on large extended families. The rationale behind the ASP approach is that sibling pairs offer more power than other pairs of relatives (such as cousins) for relatively small effect sizes [189]; small effect in this case is defined as having a λ_s value less than say 2. λ_s is defined as the conditional probability an individual is affected by a disease given its sibling is affected, divided by the population prevalence of that disease. λ_s for schizophrenia is around 10 [188] but there are likely to be multiple susceptibility loci, giving a lower locus specific λ_s. Sib pair samples can also be less expensive to recruit than extended family samples, allowing larger sample sizes. The advantage of the extended family design lies with the potential for reduced *locus heterogeneity* within the sample; large families are unlikely to harbour more than one risk allele. Since schizophrenia is known to have multiple susceptibility loci, every care should be taken to minimise locus heterogeneity within a sample. As ASP samples are based on large numbers of unrelated families they will sample much more widely from the population(s) of interest. In particular, a few recent studies have considered large ASP meta-analyses. These meta-analyses include families from a diverse range of populations [137, 199, 134], making it unlikely that they all result from a single cause. The efficacy of each of the methods will depend upon the true (unknown) disease model. Genes detected by the ASP based design may be of greater relevance to the population in general than rarer genes detected in extended pedigrees. However, definite identification of even a single schizophrenia susceptibility gene is likely to be of substantial significance [96]. In the past some researchers have favoured the ASP design because the statistical analyses are more tractable; however, the methods described in this chapter, chapter 2 and chapter 5 illustrate that this should no longer be a concern. The effect that locus heterogeneity has on the power to detect any single susceptibility locus is discussed in chapter 6.

1.2.2 Cardiovascular disease and quantitative genetics

Another focus of this thesis is the analysis of quantitative risk factors affecting cardiovascular disease (CVD) . Globally, CVD accounts for a third of all deaths (`http://www.who.int/cardiovascular_diseases/priorities/en/`). Large scale epidemiological studies such as the Framingham heart study ([175], chapter 4) have shown that diseases such as heart disease are affected by a large range of factors such as smok-

ing, blood pressure, cholesterol, physical activity and poor diet. Some of these factors have been shown to have a genetic basis and hence have been targeted for genetic (linkage) analysis. Unlike the qualitative disease outcomes commonly studied in psychiatric genetics, factors such as blood pressure, cholesterol and obesity (often expressed as Body Mass Index or BMI; this is weight in kilo-grams divided by height in metres squared) are quantitative. Genomic locations containing one or more genes influencing a quantitative trait are commonly assessed in terms of the amount of variation they explain in the observed phenotype. Such genomic regions are referred to as *Quantitative Trait Loci* or QTL. A QTL with a large effect upon the trait may be composed of either two or more proximal genes modestly affecting the trait value (in the same direction) or, alternatively, a single gene with a large effect upon the trait. Separating the effects of nearby QTL requires large amounts of data and it will not usually be possible to distinguish between the effects of a number of small (linked) effects and a single larger effect. Genetic analysis of quantitative traits requires methods different to those commonly applied to qualitative traits. There have been many genetic linkage studies of quantitative traits truncated to be qualitative (e.g. truncation of blood pressure to a yes/no definition of hypertension [12]). However, there will be more power to detect loci affecting the trait when the underlying quantitative traits are analysed [70]. As with the ASP analysis methods for qualitative traits, historically some researchers favoured binary traits over quantitative traits because they were simpler to analyse. The methods described in chapter 2 show that quantitative traits can be effectively analysed without the need for truncation to a qualitative trait. Since many of the quantitative traits affecting CVD change throughout life, appropriate data sets should allow characterisation of the composition of the traits over time. Analysis methods suitable for longitudinal quantitative trait analysis are described in chapter 2 and applied to real and simulated data in chapters 4 and 3, respectively.

1.3 Analysis methods

A large number number of techniques exist for extracting linkage information from sets of genotyped relatives. In the thesis linkage methods, based on both qualitative and quantitative traits, are applied. The basic methods are introduced below with pointers given to related chapters.

1.3.1 Parametric analysis

Genetic linkage analysis is a technique for assessing the recombination frequency between loci in pedigrees. A few definitions are required to explain this further. A haplotype is a series of alleles found at adjoining loci on a chromosome. These haplotypes are broken down by the recombination events which occur when the gametes are formed in reproduction. By examining two distinct molecular markers it is often possible to count the proportion of gametes in a sample in which the parental haplotypes are conserved (or *non-recombinant*)

at these two loci. This haplotype conservation will occur most frequently if the two loci in question are close together on the same chromosome. Alternatively, if the loci are very far apart, or on different chromosomes, the two loci will segregate independently, with the alleles being both derived from the same parent 50% of the time. The probability that the haplotype is not conserved (or *recombinant*) is the *recombination fraction*, θ. This can be mapped from a fraction in [0,0.5] to a measure of genetic distance, m, in the range $[0,\infty)$ with a map function. Popular map functions are the Haldane map function

$$m = -\frac{1}{2}\ln(1-2\theta),$$

which ignores interference (a phenomenon which inhibits recombination events from occurring in close proximity), and the Kosambi map function

$$m = \frac{1}{4}\ln\left(\frac{1+2\theta}{1-2\theta}\right),$$

which takes interference into account.

Genetic linkage analysis can be employed for disease gene mapping by assuming a person's genotype can be inferred from their phenotype, usually by assuming the disease gene is dominant or recessive in its effect upon the phenotype. Assuming the haplotype formed by this inferred genotype and a molecular marker can be assessed in a sample of individuals, it is possible to gauge whether the disease gene is likely to be on the same chromosome as the molecular marker (i.e. $\theta < 0.5$), or some other chromosome (i.e. $\theta = 0.5$). If the recombination fraction between the marker and the putative disease locus is less than 0.5 the loci are said to be *linked*.

In cases where haplotype status (known as the phase of an individual) and hence the recombination events cannot be unambiguously determined, it is possible to write down a likelihood for the pedigree, incorporating the probabilities of the unknown elements (such as the initial haplotype status of the founder individuals), which can then be maximised to assess the distance between the putative disease locus and the marker. The importance of the disease locus can be assessed by applying tests based on likelihood theory.

Since some individuals may not have genotype information available, the population frequency of the genotypes can be factored into the likelihood. This population frequency is usually estimated either from genotyped founder individuals or from suitably matched unrelated individuals. By far the most common test statistic used in human genetics is a form of the likelihood ratio (LR) test known as the LOD score. The LOD score is defined as the base 10 logarithm of the ratio of the likelihood that the recombination fraction is some value θ to the likelihood under the hypothesis of no linkage (i.e. $\theta = 0.5$). LOD scores can be converted to traditional $2\ln(LR)$ statistics by multiplying them by $2\ln(10) \simeq 4.6$.

For example, consider the family in figure 1.1. The family is typed for a single molecular marker (with alleles A and a) in three generations, with all individuals assumed to be affected (shaded) or unaffected (unshaded) as a result of the genotype at a putative

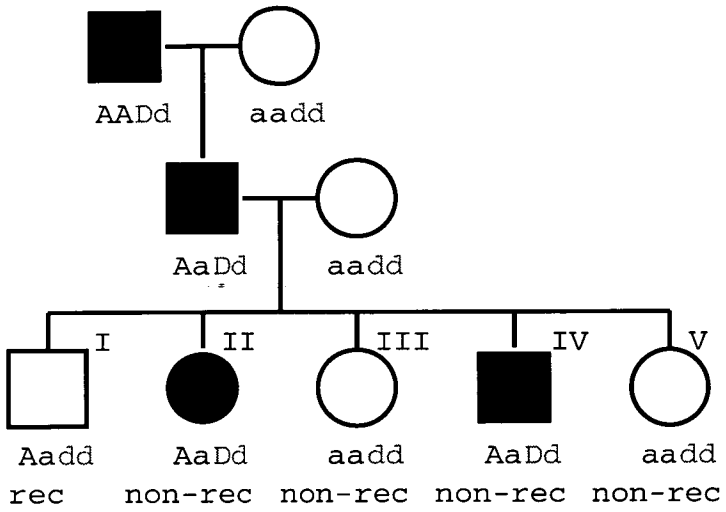


Figure 1.1: Likelihood calculation example

disease locus (assumed to have disease allele D and wild type allele d). Assuming dominant inheritance, the inferred disease genotype is shown in grey. In the grand-parental generation the haplotype status of the individuals are known (AD/Ad and ad/ad) but the recombination status cannot be determined. The father in the second generation must have inherited the haplotype 'ad' from his mother and the haplotype 'AD' from his father; the recombination status is not known however as both grandparents are homozygous at the typed marker. In the third generation the recombination status of the children can be determined. Since the father in the second generation inherited the 'AD' and 'ad' haplotypes, any children inheriting these haplotype must be non-recombinant (non-rec in figure 1.1) with respect to these two loci. Since the mother in the second generation always transmits an 'ad' haplotype (irrespective of recombination), children II to V must have inherited the 'AD' or the 'ad' haplotype from the father. Conversely, child I must have inherited an 'Ad' haplotype from his father (alongside the 'ad' haplotype from the mother). This means child I is a recombinant (rec in figure 1.1). The likelihood function is constructed based on this information. Since each gamete has a probability θ of being recombinant and (1-θ) of being non-recombinant the likelihood is

$$L(\theta) = (1-\theta)^4\theta^1. \tag{1.1}$$

Maximising the likelihood gives the maximum likelihood estimate (MLE) for θ, usually denoted $\widehat{\theta}$. The MLE is obtained by differentiating equation 1.1 and equating to zero. Computing $\log_{10}$ of the maximised likelihood divided by the likelihood evaluated assuming no linkage (θ=0.5) yields the LOD score. In this example the MLE is $\widehat{\theta} = 0.2$ and the LOD is $log_{10}\left(\frac{0.8^4 0.2}{0.5^5}\right) = log_{10}(2.62) = 0.42$.

If the same family was used but there was no information on the either the disease

status or the marker genotypes at the disease locus in the grandparents it would not be possible to determine the haplotype of the father. The likelihood must therefore be constructed allowing for the possible haplotypes carried by the father. Assuming the two possible haplotypes (i.e. AD, ad or Ad, aD) are equally likely the likelihood is

$$L(\theta) = \frac{1}{2}(1-\theta)^1\theta^4 + \frac{1}{2}(1-\theta)^4\theta^1.$$

In this case $\widehat{\theta} = \frac{1}{2} - \frac{\sqrt{3}}{6} \simeq 0.21$ and the LOD is 0.12. The estimate of θ is larger here (and the LOD is lower) than in the first case because there is greater uncertainty about the number of recombinations inferred to have occurred.

Even when the disease appears to be Mendelian it is not uncommon for environmental factors and/or genetic factors unlinked to the region to influence the phenotype of an individual [168]. Therefore, the specification of a set of penetrance parameters is integral to parametric linkage analysis based disease mapping. Penetrance is defined as the probability that a person with the risk genotype develops the disease; that is, P(disease | has risk genotype). Similarly, the probability that affected persons do not have the risk genotype of interest is the phenocopy rate; that is, P(disease | does not have risk genotype). Assuming the parental origin of the allele is of no consequence (no *imprinting*), there will be three possible risk genotypes (DD, Dd and dd) to specify (i.e. three penetrance parameters). If the effect of the locus on disease is thought to be dominant, then the penetrances for individuals with one or two copies of the disease allele should be set to be equal (i.e P(disease | DD) = P(disease | DD)). Under recessive inheritance, the penetrances for individuals with zero or one disease allele should be set to be equal (i.e P(disease | Dd) = P(disease | dd)). The penetrance values are factored into the likelihood function described in the last paragraph.

Choosing appropriate penetrance parameters is clearly non-trivial for diseases in which the Mendelian inheritance model is, at best, approximate. In fact, it can be shown that, in the presence of multiple trait loci, it is sometimes impossible to correctly assign penetrances for all loci simultaneously ([201], p121). Nonetheless, a number of studies have shown that only a small number of penetrance sets (e.g. recessive, dominant) are necessary to ensure the power to detect linkage is near optimal (e.g. [88]) when single marker analysis is performed. A separate analysis is performed for each penetrance set of interest. Furthermore, fitting a model with incorrect penetrances and disease allele frequencies has been shown not to increase the type I error rate for detection of linkage [255] (although erroneously specifying marker allele frequencies in ungenotyped founder individuals to be lower than they should be may lead to false positives [86]). What happens in practice is that when some of the parameters are mis-specified or there are genotyping errors, the lack of fit of the model is absorbed in the distance (the recombination fraction) between the putative disease locus and the marker of interest. Any model which is similar to the true model will still give evidence for linkage but, since the recombination fraction has absorbed some of the noise in the model, the recombination fraction will be overestimated

[201, 85]. This means that the parametric linkage method, despite its original conception as a method for the analysis of Mendelian disorders, can be effectively applied to linkage *detection* (but not, for example, location estimates) for complex diseases in which the inheritance pattern is not Mendelian. The serendipity of this result has led to some misconceptions however. Some researchers, accustomed to applying Mendelian models to non-Mendelian traits for linkage detection, have reported the regions derived from their family based complex disease samples as if they were known without error. The effect of inferring disease regions in this way is investigated in chapter 7.

For truly Mendelian disorders, where the model can be specified accurately, linkage analyses utilising multiple markers simultaneously (multipoint linkage) usually gives greatest power. This follows because the use of multiple markers reduces the chance of there being no marker information, useful for linkage, in the region of interest. However, since evidence for linkage and the distance of the putative disease region from the tested marker are assessed together (confounded), true linkage to a single marker under mis-specified parameters often results in the maximum LOD score occurring too far from the actual locus (i.e. over-estimation of the recombination fraction). If multiple markers are used then the maximum LOD is often shifted outside of the range of all of the markers. This happens because, in cases in which the over-estimation of the recombination fraction leads to recombination fraction estimates larger than the known distance between the some of the loci, the only location compatible with the markers and their pre-specified locations is before the first marker or beyond the last. That is, the maximum LOD score cannot be further from every marker simultaneously. As a result of this, the multipoint LOD score can be rather low when there is parameter mis-specification. For this reason, multipoint analyses are often not utilised in parametric analysis of complex disease data. Providing one does not require accurate estimates of the recombination fraction, two point (a single marker together with the putative disease locus) linkage analysis is often a robust method for linkage detection. Whilst an estimate of the disease gene location may be desirable, estimation of the position of the true locus is rather difficult in linkage analysis (see Chapter 7), not least because of the relatively small number of recombination events delineating the region of interest in realistically sized human genetic data sets. A possible alternative, avoiding the problems of parameter mis-specification and/or genotyping error effects, involves explicitly modelling the mis-specification with one extra parameter [85]. Another possibility is maximisation of the likelihood over all of the model parameters (penetrances, allele frequencies, recombination fraction) [47]; however interpretation of statistical significance is more difficult in such cases and requires alternative methods such as computer intensive simulations. Note that even in linkage analyses which only use the markers one at a time, multiple markers may still be very useful in the initial stages because they can be used to identify unlikely double recombination events between tightly linked sets of markers. Programs such as Merlin [1] flag such events as they are likely to be genotyping errors.

An important consideration in complex disease may be locus heterogeneity. Whilst

linkage studies are immune to the effects of allelic heterogeneity (since all families will show linkage to the same chromosomal region, irrespective of which mutation is present in that family), locus heterogeneity will dramatically reduce power to detect linkage (Chapter 6). The effects of recombination fraction heterogeneity (i.e. θ between the putative disease locus and the marker is <0.5 in some families and 0.5 in the other families) can be modelled in the likelihood formulation of parametric linkage analyses. Smith [210] proposed fitting an additional parameter, α, in the likelihood. A proportion, α, of the families are assumed to be 'linked' (i.e. the recombination fraction between the putative disease locus and the marker is less than 0.5) to the disease locus of primary interest. In the remaining $1-\alpha$ families, the disease status is segregating independently of the disease locus of primary interest (i.e. that linked to the marker being analysed); that is, the recombination fraction between the locus causing the disease in these families and the disease locus of primary interest is assumed to be 0.5 (or linked marker). The disease allele frequency and penetrances specified for this model are particular to the locus (marker) of interest in the $\alpha \times 100\%$ of the families that are 'linked'. Since the disease state of the other $(1-\alpha) \times 100\%$ of the families is not linked to this locus (marker), the parameters relating the disease phenotype to underlying genotype are of no relevance to these 'unlinked' families. In complex disease there may of course be many disease susceptibility loci but generally only one is of interest at any one time. The heterogeneity model only deals with one locus at any given time, with all other unlinked disease loci ignored in this formulation. It is important to note that whilst this heterogeneity model may be an improvement over the standard model (i.e. with $\alpha = 1$) when there are families affected as a result of mutations segregating at loci unlinked to the one of primary interest, allowing α to be less than 1 is not a panacea for poor study design (see also chapter 6).

The likelihood under the heterogeneity model is maximised over both α and θ. Call the likelihood with both parameters unrestricted L1 and the likelihood with either $\alpha = 0$ or $\theta = 0.5$ L0 (either condition is sufficient for the other to hold). The likelihood ratio test ($\log_{10}$ version) of L1 versus L0 is often referred to as the HLOD statistic.

Likelihood calculation can be achieved in arbitrary pedigree structures by use of the Elston-Stewart algorithm [66]. This algorithm uses a technique known as *peeling*. Peeling works by splitting the extended pedigree into nuclear families and calculating the likelihood for each nuclear family separately. The overall likelihood is calculated by summing over all the nuclear families, taking into account the possible genotypes of the individuals linking the nuclear families. The likelihood of any single nuclear family only need consider three individuals simultaneously (the mother, father and, sequentially, each child). If the nuclear family computation is inexpensive and there are a limited number of admissible genotypes for the linking individuals then the peeling also allows extended families to be dealt with efficiently. This approach requires all the possible genotypes of each individual to be considered at each stage. Whilst this may be simple when there are few marker loci, when there are many genotyped loci, this part of the calculation can be computationally expensive (assuming a multipoint analysis of all markers simultaneously is required).

For this reason, the peeling based algorithm works well on large pedigrees but only when there are few marker loci. A popular implementation of the Elston-Stewart algorithm is in the program FASTLINK [44].

An alternative algorithm for likelihood computation is the Lander-Green algorithm [130]. First note that the linkage information of a haplotype can be expressed solely in terms of whether the allele was passed from the parent's paternal or maternal side. In the algorithm, all of the informative gametes in the non-founder individuals are treated simultaneously by specifying a binary digit for the status of each allele (i.e. whether the allele was from the parents maternal or paternal haplotype). The binary digits can be assembled into *inheritance vectors*, summarising the flow of allele transmissions in the pedigree at that marker. In some cases there may be sufficient marker information to unambiguously determine the inheritance vector at a given marker. More likely however, the marker information will only be sufficient to reduce the set of inheritance vectors, to a smaller subset (the *legal* set of vectors [212, 90]) of possible inheritance vectors. The likelihood at this marker is then based upon the probabilities of each of the 2^{2i} inheritance vectors (assuming i non-founders) having occurred, conditional on the observed marker data. For additional markers we treat the vectors along the chromosome as hidden states of a Markov model with the transition probabilities between the state of the vector at one marker and the next determined by the genetic distance between them. Under the Markov model all other vectors beyond the adjacent marker are independent, conditional on the adjacent vector. This means that the algorithm is very efficient for large numbers of loci with computational time only increasing linearly with the number of markers. In contrast, the Elston-Stewart algorithm scales exponentially with the number of markers. However, dealing with large numbers of related individuals is computationally expensive with the Lander-Green algorithm since the inheritance vector becomes very large in large pedigrees (vector has 2^{2i} elements with i non-founders). Implementations of the Lander-Green algorithm include GENEHUNTER [127], Allegro [90] and Merlin [1] (see also section 2.3). The Lander-Green and Elston Stewart algorithm can hence be seen to be complementary, with one dealing with large numbers of markers in small pedigrees and one dealing with large numbers of individuals and a few markers.

For multipoint analyses of many markers in large pedigrees approximations to the exact likelihood are available. These approximations work by sampling inheritance vectors (called descent graphs in [211]) via a Markov Chain Monte Carlo based scheme. Samples from the set of possible (consistent with the observed marker genotypes) inheritance vectors are drawn in proportion to their likelihood, allowing an approximation to the true likelihood based on a large number of simulation iterations. The number of iterations required to provide a good approximation to the true likelihood depends upon factors such as number of pedigree loops, number of markers and inter-marker spacing. This method is implemented in the programs SIMWALK [211] and Loki [99]. The method allows analyses of large numbers of markers in relatively large pedigrees. Inheritance vector based methods also allow simple calculation of identity by descent (IBD) coefficients. Two indi-

viduals are IBD for a given allele if the allele can be traced back through the inheritance vectors to the same ancestor. There are more details on IBD coefficients in section 1.3.3 and section 2.3.

Within the candidate gene paradigm the significance of one single marker (or genomic location) can be assessed by standard asymptotic theory. Given a large enough sample the likelihood ratio test comparing twice the ln-likelihood difference between the likelihood at the likelihood maximum (over $0 \leq \theta \leq 0.5$) and the likelihood evaluated at $\theta = 0.5$ will be asymptotically distributed as a mixture of χ^2_1 and a point mass at 0. This distribution is a mixture because θ under the null distribution is on the boundary of the parameter space ($\theta = 0.5$); see [200] and Chapters 2 and 3. The HLOD statistic does not converge to an asymptotic distribution (since either $\alpha = 0$ or $\theta = 0.5$ specify the null) but it can be approximated by a 50:50 mixture of 0 and the larger of two independent χ^2_1 variables [201].

In contrast to the single marker (candidate gene) case, under the vastly more popular positional cloning paradigm, a large number of genomic locations are considered. The standard statistical method for dealing with such multiple testing issues is the Bonferroni correction. This correction assumes that n independent tests have been done. The corrected p-value, p*, is given by

$$p* = 1 - (1-p)^n = 1 - \sum_{r=0}^{n} \binom{n}{r} (-p)^r \simeq 1 - (1 + n(-p)) = np \text{ for small } p$$

where p is the p-value obtained in a given test. However, the assumption that the tests along the genome are independent is unreasonable for closely linked loci. This multiple testing issue has received considerable attention in the literature with the most cited article on the issue (Lander and Kruglyak, [129]) suggesting an approximation based on the assumption that a genome scan with an infinitely dense marker map has been performed. The Lander and Kruglyak article proposed that a LOD score of 3.3 be deemed sufficient for evidence of genome wide linkage. Assuming the asymptotic mixture distribution described above this corresponds to a p-value of 5×10^{-5}. If there is prior evidence for linkage to a particular region, perhaps because there have been linkages reported to the region previously, then there may be justification for lowering this threshold. Significance thresholds can also be determined by simulation. However, in parametric linkage analysis, the true disease model is unknown, so simulations based upon the assumed disease model (e.g. using SLINK [244]) may not provide a true reflection of the null distribution. Furthermore, methods based upon generation of marker data under the null hypothesis that there is no linkage between any marker and the disease locus may generate data which is not representative of the actual marker data. Sham [201] gives an example of a single, phase known, three generation family (as in figure 1.1 but with 1 third generation offspring). Assume that the available marker linked (with θ=0) to the disease locus was fully informative (i.e. the male parent in figure 1.1 is a heterozygote at both the marker and the inferred disease locus). The LOD score of the family will then be $log_{10}2 \simeq 0.3$.

Assume next that data are simulated so that the marker genotypes are generated on the basis of a set of allele frequencies. The possible LOD values in the simulation replicates will depend entirely on the specified set of allele frequencies. If the allele frequencies are such that one allele is very rare then most replicates will yield a LOD of 0 as a result of the male parent in figure 1.1 being a homozygote (and hence uninformative for linkage). The empirical threshold (for a given significance level) from such a simulation may hence be artificially low.

Power calculation for parametric linkage analysis can be performed using computer simulation. More details are given in chapter 6 study design.

1.3.2 Non-parametric analysis

A popular alternative to parametric linkage analysis considers allele sharing in pairs of individuals with the same phenotype. Allele sharing (alternatively, *non-parametric* or *model free*) techniques aim to avoid the need to specify the set of parameters needed for parametric linkage. Since affected individuals with the disease are thought to be better predictors of disease allele carrier status than unaffected individuals are of wild type allele carrier status, most allele sharing analyses only consider affected individuals [201]. A common design involves pairs of affected sibling pairs or ASPs. The basic idea is to look for deviations from expectation in the proportion of alleles shared identity by descent (IBD). For example, sib pairs are expected to share 1 allele IBD. At any given genomic location they may in fact have 0, 1 or 2 alleles in common. If the genomic location influences disease status, then sibs who share more alleles IBD will be more likely to have the same phenotype. A common test of linkage for ASPs is based on the mean number of alleles shared IBD (the ASP mean test). The properties of this test, including statistical power, are considered in more detail in chapter 6. Alternatives to the ASP mean test include tests for sibs sharing 2 alleles IBD at a marker and a likelihood based formulation [189, 190] which considers the number of ASPs sharing 0, 1 or 2 alleles IBD. Other relative pairs, such as cousins, can be considered in a similar fashion.

Extensions of the relative pair allele sharing approach to general pedigrees have been suggested [127, 246]. However, there are many ways of accounting for the fact that the relative pairs in a single pedigree are correlated with each other. Indeed one reasonable way of structuring the data to account for the correlation between relative pairs would be to fit a parametric model with penetrances et cetera [87, 201]. Given the aim in non-parametric analyses is to not specify a model, some other arbitrary weighting scheme needs to be specified. A variety of weighting schemes based upon IBD configurations within pedigrees (a vector of the allelic state of all of the individuals in a pedigree in terms of the founder alleles [246]) or inheritance vectors (from the Lander-Green algorithm) [211] have been proposed. Sobel and Lange propose five different weighting schemes based on inheritance vectors [211]. Furthermore, with multiple pedigrees, each of different size, there is no single optimal way to weight the contributions from different pedigrees [201]. Inevitably, some of the weighting schemes proposed are more powerful for dominant type

inheritance patterns, whilst others are more powerful when the true mode of inheritance is close to being recessive [246, 211]. Although these weighting schemes are not based on a 'genetic model', clearly there is some choice amongst the possible non-parametric analyses. This makes the claim that non-parametric methods are completely free of model specification untenable. This has also led to comments that whilst parametric methods require specification of parameters that are only approximations to reality, at least the model fitted is more transparent [201]. It has also been shown that the ASP mean test is algebraically equivalent to a parametric analysis with the disease allele frequency set to a very small number (say 10^{-6}), the penetrance of individuals carrying one disease allele set to 0 and the penetrance of individuals carrying two disease alleles set to a very small number (say 10^{-6}); under these conditions only individuals homozygous for the disease allele will have the disease and even then huge numbers will have to be ascertained to find individuals who exhibit the phenotype [122, 87]. This means that whilst the ASP mean test does not explicitly assume a parametric model, in actual fact it is equivalent to a rather unrealistic parametric model.

1.3.3 Quantitative trait analysis

The methods described in sections 1.3.1 and 1.3.2 considered linkage methods for analysis of disease status (qualitative traits). For quantitative traits alternative methods are utilised. One of the simplest methods suitable for QTL mapping is based upon the difference in trait value in sib pairs. If the marker locus of interest is linked to the QTL, as the number of alleles shared IBD at the marker increases (from 0 to 1 to 2), the difference in the trait value between a pair of sibs will be expected to decrease. The Haseman-Elston (HE) test [97] regresses the squared trait value difference on the number of alleles shared IBD between a pair of sibs. A significant slope term in the regression indicates linkage. Univariate normality of the squared sib pair difference in each of the IBD classes is generally assumed, allowing significance to be assessed with results based on asymptotic theory. This approximation has been shown to be robust, even in small samples [256]. Other functions of the sib pair trait values have since been considered [65] and offer additional power under certain circumstances [238, 202].

An obvious limitation of the HE approach is that only sib pairs can be used. Other relative pairings, such as cousin pairs (but not parent-offspring pairs as there is no variation in the number of alleles such pairs share IBD), can be used instead but it is unclear how to include different sets of pairs from the same pedigree in a single analysis. Even in the simple case of a sibship data there have been many proposed corrections for the non-independence of sib pairs within a sibship ([201, 139], see also chapter 6). Extensions of regression based methods to general pedigrees [203] have been proposed but these are only applicable to relatively small pedigrees in practice and require further work to assess their utility.

An alternative to HE regression is variance component (VC) analysis. VC methods are explained in detail in chapter 2. A brief introduction is given here. Variance com-

ponent techniques partition the observed phenotypic variation into different components. With information on individuals' relationships and phenotypes it is possible to estimate a genetic and an environmental part. When there is molecular marker information, the genetic part can be split into a component due to a particular genomic region (the QTL part) and a component attributable to the remainder of the genome (the polygenic part) [83, 8, 6, 79]. The estimation of a QTL specific variance (the variance associated with a genomic region) can be contrasted with the parametric linkage approach (section 1.3.1) where a single gene effect was specified in terms of its frequency and effect. The VC technique utilises a matrix of relationships (the numerator relationship or A matrix) between the pedigreed individuals to allow estimation of polygenic genetic effects and a matrix of marker specific estimated allele sharing (IBD) probabilities between individuals to allow estimation of QTL effects. The covariance between different individuals' trait values will depend upon the degree to which they share polygenic (this depends on the individuals' relationships, i.e. the A matrix) and QTL specific effects (this depends on the alleles at the specified locus present in the individuals, i.e. the set of marker IBD probabilities between individuals). The VC method also has the advantage of allowing the estimation of fixed effects at the same time as the (random) genetic effects are estimated. This means that any measured environmental effects can be appropriately accounted for. By assuming multivariate normality of the phenotypic values it is possible to write down a likelihood based on the phenotypic values and the known covariance structure in the whole pedigree. This likelihood can be maximised, yielding estimates of the proportion of variance attributable to polygenic and QTL effects. The significance of the QTL can be assessed with tests based on likelihood theory. The likelihood ratio test of positive QTL variance versus no QTL variance is asymptotically distributed as $\frac{1}{2}\chi^2_1 : \frac{1}{2}0$. This follows because the variance parameter under the null distribution is on the boundary of the parameter space [200]. If a base 10 logarithm is used in the likelihood ratio the asymptotic distribution is the same as for the parametric linkage LOD. Providing calculation of the A matrix and estimation of the matrix of marker specific IBD probabilities is possible (see chapter 2), the VC method will be suitable for analysis of arbitrary pedigrees. In contrast with relative pair based approaches, where there is a problem with a lack of independence between pairs, the VC approach maximises the likelihood jointly over all individuals, conditional on their relationship in the pedigree. The likelihood formulation is also very flexible, allowing the simple inclusion of the effects of measured covariates. Maximisation of likelihoods is computationally more intensive than regression based procedures such as the HE method but, for all but the simplest pedigree structures, HE type approaches have problems with the non-independence of related relative pairs. Both VC and HE approaches make normality assumptions to allow significance testing based on asymptotic results: in the case of VC analysis the phenotypic data are assumed to be multivariate normal, in the case of HE analysis the distribution of the squared sib pair difference in the IBD sharing classes is assumed to be univariate normal [249]. Note however that the HE approach does not depend upon the normality assumption for estimation but that the VC approach requires

multivariate normality of the phenotypes to hold for unbiased estimation.

Further discussion of the analysis methods suitable for quantitative trait analysis is deferred to chapter 2.

1.4 LD mapping

LD mapping is an important component of the positional cloning paradigm. On the one hand, linkage analyses provide a crucial method for the initial identification of complex disease loci; the limited number of recombinations available in human genetic linkage studies ensures only a relatively sparse map (5-10cM) is required to extract all of the linkage information from a data set ([139], p397). On the other hand, LD measures the association between alleles at a population level and as such utilises the information on all of the recombination events that have occurred since the most recent common ancestor of two 'unrelated' individuals. This means that the mapping resolution is in principle substantially higher, providing appropriate markers are available.

It is important to note that factors such as population stratification and natural selection may generate LD and this should be accounted for in any analysis using LD to map disease genes. If the sample are a collection of unrelated cases and controls, population stratification should be minimised by drawing the control individuals from the same population as the case individuals. Assuming that markers from several chromosomes are available for analysis, methods that use unlinked markers to correct for the effects of stratification have been proposed [55]. Alternatively, samples with 'internal' controls may be gathered by genotyping the parents of affected individuals. This works by genotyping parent-offspring trios and treating the two untransmitted parental alleles as a control sample. This will only work if the parents are heterozygous, otherwise the transmitted and untransmitted alleles are indistinguishable. This approach is used in, for example, the TDT statistic [215] and, assuming the marker of interest is biallelic and denoting the number of times an allele is transmitted as T and the number of times it is not as N, it can be written as

$$TDT = \frac{(T-N)^2}{(T+N)^2}.$$

This statistic is distributed asymptotically as χ^2_1. If the affected individuals are picked at random from the population the TDT offers a test of both linkage and LD. Alternatively, if the affected individuals are all derived from single large pedigree in which there is a single disease allele, derived from a single founder, the TDT is only a test of linkage. In reality, many samples will be a mixture of these two extremes and hence will detect some LD effects alongside the effects of linkage on the marker tested [201]. In practice, the need for heterozygous parents means that TDT approach may be inefficient in terms of data. Furthermore, in the analysis of late onset disease, parental genotypes may not be available. Although it is possible that population stratification caused the sea of false positive LD mapping results observed in studies of complex diseases such as schizophrenia

([242], there have been >50 Web of Science listed journal articles with "no association" and "schizophrenia" *in the title* since 1995, most of these contradicting previous results), low statistical power and excessive multiple testing seems just as likely to have generated these artefacts.

Whole genome association? Some researchers have advocated the use of LD techniques for *initial detection* of complex disease susceptibility genes [192, 182, 75]. Such techniques rely on the huge amounts of single nucleotide polymorphism, or SNP, marker data that are being generated (the SNP Consortium `http://snp.cshl.org/`). However, doubt has been cast upon the efficacy of whole genome association (WGA) techniques for complex disease mapping [245, 241, 39]. For the WGA techniques to be effective the risk alleles must be at appreciable frequencies (>1%, say) in the population of interest and the number of risk alleles must be small [208]. That is, the common disease common variant (CD/CV) hypothesis [184] must be true. The allele frequency spectrum for neutral SNPs is known from theoretical studies [69] to be rather wide. Reich and Lander [184] have suggested that disease susceptibility alleles will in fact be common and have simple spectra but other authors do not share their optimism [178, 179]. Another problem with the population based LD techniques lies in the fact that population derived samples will be very genetically heterogeneous, leading to low correlations between phenotype and any single underlying risk genotype [245, 259].

LD based techniques will be most effective when the observed marker polymorphism is the actual disease variant; indeed this was the hope when Risch and Merikangas published their paper [192] on the future of genetic studies of complex human disease in 1996. More likely however, the observed SNP in WGA studies will not be the actual variant and mapping will depend upon the marker SNP being in significant LD with the disease polymorphism [126]. There is considerable variation in LD levels across the genome [112, 235], making it unclear what density of markers would be required for the WGA strategy to work. Variations in the extent of LD across the genome will mean that initial predictions [126], made based on the assumption of homogeneity in the extent of LD, will have underestimated the number of SNPs required to effectively cover the whole genome. If there is allelic heterogeneity and the marker is not the causal variant it has been demonstrated that the required sample size may reach unattainable levels [208]. To date, the vast majority of genes affecting complex human disease have been initially identified using linkage not LD mapping.

It remains to be seen whether new developments such as the 'Hap-map' project [75, 166], a project aimed at identifying the blocks of highly conserved haplotypes that exist in the human genome, will improve the utility and success of genome wide association studies. Despite controversy over issues such as block definition [172, 37], the Hap-map project seems likely to offer advantages over single marker association analyses due to the identification of tagging SNPs [113, 154] which efficiently summarise the information contained in a set of associated SNPs.

1.5 Applications in non-human populations

The emphasis throughout the thesis is on methods suitable for analysis of general pedigrees. This means that the methods described may have application in non-human populations. Indeed, some of the methods described in chapter 2 are modifications of methods originally proposed for animal breeding applications. The broad application of the variance component methods are well illustrated in chapter 2. In chapter 2 variance components methods are applied to a QTL mapping problem in a natural population of red deer (*Cervus Elaphus*). Since many of the problems encountered, such as identity by descent coefficient estimation in large pedigrees, are common to all (human and non-human) natural populations, progress will be made most rapidly by assimilating methods from different disciplines.

Although the methods applied will sometimes be similar in all (natural) populations, there are differences in the basic properties of the populations studied. In animal breeding and model organism (e.g. mice) applications animals can be bred, allowing (arbitrarily) large sample sizes and efficient experimental design. Obviously this will be impossible in humans and in ecological genetics applications. In samples in which it is not possible to arrange the matings to allow simple assessment of linkage by the counting of recombination events it will be necessary to impose some sort of 'model' for the transmission of genetic effects. As discussed above, this will either take the form of a parametric model or some sort of arbitrary weighting scheme (for example based on inheritance vectors).

The use of relatively small numbers of animals in the initial stages of animal breeding/model organism applications will ensure that the effective population size will be rather smaller than in humans. Effective population size is the size of a hypothetical population that would experience the same loss of genetic diversity due to random (genetic drift) effects as the loss observed in an actual population. Estimates of the effective population size of the human population are in the range 10 000-100 000 [183, 14, 227] whilst the effective population sizes of most livestock or model organism populations are unlikely to exceed 1000 [70]. Similarly, some natural populations of mammals such as deer will often have small effective population size, particularly since in many cases a few males will mate a disproportionate number of times [209]. If a high proportion of males in a population do not mate the effective population size will be much smaller than the census population size [70]. These relatively small populations sizes mean that, compared with human populations, there will be differences in the extent of LD in mammals. Even taking into account substantial variation in LD levels in the human genome (see section 1.4) the orders of magnitude differences in effective population size between humans and many other mammals will mean that LD is likely to extend substantially less far in humans. This will mean that unlike in human populations, LD analyses in mammals may not provide substantially greater mapping resolution than that offered by analyses based on linkage information.

1.6 Summary

Whilst there are a striking number of single genes causing human (Mendelian) disease, such disease alleles are uncommon in human populations and only contribute a small amount to morbidity and mortality. In terms of overall public health, complex diseases such as cardiovascular disease and psychiatric disease are of substantially greater importance. Considerable resources have been devoted to the identification of the genes underlying these complex diseases. In this thesis the main foci are problems in psychiatric disease and cardiovascular disease.

In this introductory chapter I have described various methods for linkage analysis of family data. Providing a small range of disease models are used, parametric linkage analysis offers a robust method for detection of loci affecting qualitative disease traits. There are a number of non-parametric analysis methods available but in complex pedigrees they simply represent alternative ways of structuring the available pedigree to account for the correlation between individuals. The effectiveness of any particular technique will usually depend upon the true, but unknown, genetic model applicable to the disease in question. Becoming overly dogmatic about the choice of method seems unwise; a recent paper compared the discussion to the controversy over the positioning of the table flags in the Panmunjom armistice talks ([87], `http://www.gluckman.com/NKBorder.html`).

For analysis of quantitative traits, the VC approach provides a flexible framework, suitable for analyses of arbitrary pedigrees. In cases in which there is some additional relevant information available in the data set, such as covariates describing pertinent environmental factors, the variance components (VC) approach allows flexible removal of unwanted (environmental or perhaps background genetic) noise. VC approaches can also be used for binary traits (e.g. disease status) via a threshold model (see chapter 5). Furthermore, the VC approach can be extended to multiple traits, with a particular parameterization (see chapter 2) allowing efficient analyses of data sets with multiple trait measures over time.

Summary As indicated in this introductory chapter, this thesis can be regarded in two parts; one addressing gene mapping in cardiovascular disease (CVD) and the other addressing problems in psychiatric disease. The data analysed here are quantitative in the case of CVD and qualitative in the case of psychiatric disease. In chapter 2 analysis methods for quantitative trait locus mapping (quantitative traits) are laid out; the first part of this chapter describes techniques suitable for univariate measures before describing in detail methods suitable for longitudinal data. Incorporated into the discussion of this chapter is a description of an analysis of a quantitative trait in Red Deer. The Red Deer work provides an excellent illustration of the broad application of the VC techniques with the VC method also being applied to human CVD and psychiatric disease data sets. The objective of chapter 3 is to examine some of the properties of the longitudinal QTL analysis techniques described in chapter 2. This simulation chapter allows an assessment

of the appropriateness of these complex multivariate models for analyses of human (or other natural population) data sets and the discussion section takes up again some of the methodological issues discussed in chapter 2. Chapter 4 describes a data analysis of a remarkable set of 330 extended families, measured for a range of CVD risk factors, made available as part of Genetic Analysis Workshop 13 (GAW13, [5]). Analyses of simulated data rarely provide a true reflection of the difficulties encountered in analyses of real data sets and this chapter demonstrates what can be achieved in practice. Chapter 5 describes a genome scan of a set of families collected as part of a European Science Foundation (ESF) project looking at mental illness. An additional analysis utilising additional families is also performed. The ESF chapter illustrates the application of two different methods for linkage analysis of binary (qualitative traits); parametric linkage analysis (described in this chapter) and variance components linkage analysis (described in this chapter, chapter 2, as well as chapter 5). Study design for linkage-based mapping of psychiatric disease has become a hot topic in the past year or so and chapter 6 considers some of the issues involved. In this chapter the effect of locus heterogeneity upon linkage analysis is examined and the results of some recent meta-analyses are discussed. Chapter 7 looks at a particular aspect of parametric linkage analysis, the effect of phenocopies upon disease region identification. The aim of this chapter is to quantify the effect of these phenocopies; this is done both analytically and through computer simulation. Finally, the last chapter gives an overview of the current state of human genetics and discusses future directions. Where possible technical or methodological issues are discussed in the individual chapters, allowing the final chapter to tie together the findings from the different subject areas.

Chapter 2

Analysis of longitudinal quantitative trait data in complex pedigrees: Theory

In this chapter various methods suitable for linkage analysis of quantitative traits in extended families are described. All of the methods are based upon variance component (VC) techniques which partition the phenotypic variance into polygenic, QTL specific and environmental components. The chapter begins with univariate polygenic models. The models are expanded, first to include molecular marker information and then subsequently to allow analyses of multivariate data.

The univariate methods are utilised in the ESF schizophrenia and bipolar disorder data analysis (Chapter 5), the GAW data analysis (Chapter 4) and the Red Deer data analysis (see Discussion section of this chapter). The multivariate methods are used in the simulated data chapter (Chapter 3) and the GAW analyses. To reduce repetition, the methods are not described separately in those chapters.

2.1 Univariate methodology

The basic principle underlying most quantitative genetic methodology is the partitioning of the observed or phenotypic variance into separate components. By examining quantitative trait values in individuals of known relationship it is possible to partition this phenotypic variance into components attributable to genetic factors and components attributable to environmental factors. With the advent of molecular marker technology, it became possible to further partition this genetic component into variance associated with regions of the genome. These regions of the genome that explain some of the observed variation in the trait of interest are known as quantitative trait loci or QTL.

Consider first of all the case where there is no marker information. Assume that there

is one trait measure per individual. The main interest is in the random additive genetic effect, a. The phenotypic value is modelled as

$$y_i = \mu + a_i + e_i \tag{2.1}$$

where μ represents the overall mean, y_i the phenotype of individual i and e the random environmental effect. Cases in which there are fixed effects other than the mean are discussed below. The covariance between each of the y_i depends upon the relationships between the individuals. If the relationships are known this information can be used to determine the variance of the set of a values.

In a non-inbred population the coefficient of coancestry, Θ_{ij}, of individuals i and j is the probability of the same allele (identical by descent) being drawn at random from i and from j. Multiplying this value by 2 will give the average number of alleles shared identical by descent. Using $2\Theta_{ij}$ to describe the degree of relatedness between individuals and assuming no shared environment effects, the covariance between any two relatives can be written as

$$\rho(y_i, y_j) = 2\Theta_{ij}\sigma_a^2. \tag{2.2}$$

The values $2\Theta_{ij}$ are commonly assembled into a matrix, **A**, describing the relationships between all individuals of interest (the numerator relationship matrix).

Assuming the genetic and environmental effects are independent, the covariance matrix for both random effects can be written as

$$\Omega_{polyg} = \mathbf{A}\sigma_a^2 + \mathbf{I}\sigma_e^2.$$

Once σ_a^2 and σ_e^2 are estimated (below) they can be used to calculate parameters of interest such as the heritability (h^2) of the trait. h^2 is defined to be $\frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$.

Instead of dealing with additive genetic effects averaged over the whole genome (*polygenic* effects), the phenotypic value can be decomposed into the effects of Q QTL

$$y_i = \mu + a_i + \sum_{k=1}^{Q} q_{ik} + e_i$$

In matrix form this is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{a} + \sum_{k=1}^{Q} \mathbf{q}_k + \mathbf{e}. \tag{2.3}$$

Assume that for a single QTL, R_{ij} is used to denote the fraction of genes shared identical by descent (IBD) between individuals i and j at the QTL (this fraction is only ≤ 0.5 if the population is non-inbred and there are no monozygotic twins). The values of R_{ij} are estimated using molecular marker information; this is discussed in section 2.3 below. For the Q QTL affecting the trait, the phenotypic covariance from equation 2.2 can be re-expressed as

$$\rho(y_i, y_j) = \sum_{k=1}^{Q} R_{ijk}\sigma^2_{qk} \tag{2.4}$$

where R_{ijk} is the R_{ij} value for the k^{th} QTL and σ^2_{qk} is the variance attributable to the k^{th} QTL. In practice, modelling all Q QTL simultaneously is intractable. Instead a few of the largest QTL (although henceforth one QTL is assumed) are modelled explicitly. The effects of the other smaller QTL are then incorporated into the polygenic term; in effect this term represents the rest of the genome not explicitly modelled. Assuming a single QTL with variance σ^2_q, and assuming no shared environmental effects, the covariance between a pair of relatives can be then expressed (cf. equations 2.2 and 2.4) as

$$\rho(y_i, y_j) = R_{ij}\sigma^2_q + 2\Theta_{ij}\sigma^2_a \tag{2.5}$$

where R_{ij} is the fraction of genes shared identical by descent (IBD) at the putative QTL and $2\Theta_{ij}$ are the entries of the numerator relationship matrix A. When there is no marker information $E(R_{ij}) = 2\Theta_{ij}$. Note that since the model is now includes a QTL effect, the polygenic variance term,σ^2_a, now represents the variance attributable to the rest of the genome (i.e., all genes outside the QTL region). Furthermore, the genetic contribution of the QTL is assumed to be independent of those of other loci. Non-Independence will arise if there is linkage disequilibrium (LD) between the QTL and the other loci ([70], p130). If the QTL is unlinked to the other genes, then LD will not be present from a shared genealogy. Disequilibrium between unlinked loci can, however, be generated by other forces. For example as a result of admixture or migration. Thus the model is most appropriate when the QTL is unlinked to other genes that contribute to the trait and when the population is homogeneous. The QTL specific heritability, h^2_q, is defined to be $\frac{\sigma^2_q}{\sigma^2_q+\sigma^2_a+\sigma^2_e}$.

Assembling the R_{ij} into a matrix R (i.e. $[\mathbf{R}]_{ij} = R_{ij}$), the covariance matrix can then be written as

$$\Omega = \mathbf{R}\sigma^2_q + \mathbf{A}\sigma^2_a + \mathbf{I}\sigma^2_e. \tag{2.6}$$

This expression (equation 2.6) is the basis of the univariate QTL variance components method; the covariance is split into the covariance attributable to the main QTL of interest, a polygenic effect representing the rest of the genome and an error term absorbing all other terms (environment, non-additive genetic variance, epistatic variance).

If multivariate normality of the y_i is assumed and the covariance matrix is as in equation 2.6, parameter estimation can proceed via the log-likelihood of the pedigree(s)

$$\ln L(\mu, \kappa \mid \mathbf{y}) \propto -\frac{1}{2}ln\,|\Omega| - \frac{1}{2}(\mathbf{y} - \mathbf{1}\mu)^T\Omega^{-1}(\mathbf{y} - \mathbf{1}\mu) \tag{2.7}$$

where $\kappa = (\sigma^2_q, \sigma^2_a, \sigma^2_e)$. Estimates of the vector κ can be obtained by forming the score vector (i.e. the first partial derivative of equation 2.7) and equating it to zero for each

variance term of interest [139]. The equations resulting from the score are

$$tr(\widehat{\Omega}^{-1}) = \mathbf{y}^T\widehat{\mathbf{P}}\widehat{\mathbf{P}}\mathbf{y} \text{ for } \sigma_e^2 \quad (2.8)$$

$$tr(\widehat{\Omega}^{-1}\mathbf{A}) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{A}\widehat{\mathbf{P}}\mathbf{y} \text{ for } \sigma_a^2 \quad (2.9)$$

$$tr(\widehat{\Omega}^{-1}\mathbf{R}) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{R}\widehat{\mathbf{P}}\mathbf{y} \text{ for } \sigma_q^2 \quad (2.10)$$

where

$$\mathbf{P} = \Omega^{-1} - \Omega^{-1}\mathbf{1}(\mathbf{1}^T\Omega^{-1}\mathbf{1})^{-1}\mathbf{1}^T\Omega^{-1}. \quad (2.11)$$

Equations 2.8, 2.9 and 2.10 contain the parameters of interest (κ) on both sides of the equations ($\mathbf{P}$ is given a hat in the score equations to denote that it is a function of the variance parameters). This means that in practice numerical methods must be used to obtain estimates of the parameters of κ. Define $\frac{\partial\Omega}{\partial\sigma_i^2} = \Omega_i = \begin{cases} \mathbf{I} \text{ if } \sigma_i^2 = \sigma_e^2 \\ \mathbf{A} \text{ if } \sigma_i^2 = \sigma_a^2 \\ \mathbf{R} \text{ if } \sigma_i^2 = \sigma_q^2 \end{cases}$. Taking the second partial derivatives of equation 2.7 and taking expectations yields the expected information matrix, $\mathbf{F}$, with elements

$$F_{ij} = -E\left(\frac{\partial \ln L}{\partial\sigma_i^2\partial\sigma_j^2}\right) = \frac{1}{2}tr(\Omega^{-1}\Omega_i\Omega^{-1}\Omega_j). \quad (2.12)$$

The inverse of the matrix $\mathbf{F}$ provides sampling variances for the elements of the vector κ. This information matrix can also be used to increase the efficiency of the numerical maximisation needed to estimate κ. The method is based upon a modification of the Newton-Raphson algorithm, Fisher's method of scoring. Given an initial estimate of κ, $\kappa^{(0)}$, an update of κ is

$$\kappa^{(1)} = \kappa^{(0)} + \mathbf{F}^{-1}\mathbf{U}(\kappa^{(0)}) \quad (2.13)$$

where $\mathbf{U}(\kappa^{(0)})$ is the score vector and $\mathbf{F}$ is the expected information matrix, evaluated at $\kappa^{(0)}$ [80]. This updating process continues until stable estimates of κ are obtained. In practice the computation of equation 2.12 can be difficult and modifications are needed. ASREML uses the average information (AI) algorithm [80] to obtain an approximation (the average information matrix) to the required information matrix. This approximation can be used both in the updates in equation 2.13 and to obtain estimates of the variances of the estimates of κ.

In the previous paragraphs it has been assumed that there were no fixed effects other than a mean. This is the case in chapter 5 where there were no covariates available and in chapter 3 (Simulated data) where no fixed effects were simulated. If there *are* other fixed effects, arranged in a design matrix $\mathbf{X}$, the vectors of ones in equation 2.11 should be replaced by $\mathbf{X}$s and there will be a number of other (fixed) effects to estimate alongside the κ parameters discussed above. When there are a large number of fixed effects it will be advantageous to use residual maximum likelihood (REML) based estimation instead of the maximum likelihood (ML) based procedure described here. REML estimates are obtained

by applying a linear transformation to the observed trait values, y_i. This transformation is chosen such that it removes the effects of the fixed terms. ASREML uses REML whilst SOLAR [6] uses standard ML. In most human data sets there are few or no fixed effects; this means the results from a ML analysis will be very similar to those obtained from a REML analysis.

Although the IBD values must also be estimated from the available data it is common to estimate these first [79]. The analyses then proceed assuming that the IBD values are known without error. A likelihood ratio (LR) statistic can be calculated to assess the significance of the putative QTL. The main test of interest in linkage/QTL analysis compares the likelihood fitting the full model (equation 2.7 with $\kappa = (\sigma_q^2, \sigma_a^2, \sigma_e^2)$) with one fitting only the polygenic and environmental terms ($\kappa = (\sigma_a^2, \sigma_e^2)$). In human genetics it is common to take base 10 logarithms of this likelihood ratio; this is referred to as the LOD score. LOD scores can be converted to traditional $2\ln(LR)$ statistics by multiplying them by $2\ln(10) \simeq 4.6$. $2\ln(LR)$ is distributed asymptotically as $\frac{1}{2}\chi_1^2 : \frac{1}{2}0$. This follows because this is a test of a parameter (σ_q^2) on the boundary of its parameter space under the null [200]; if the true value of the additional variance parameter, σ_q^2, is zero then half of the time the likelihood ratio test statistic will be zero (see also section 3.2.3 for multivariate analogues). This is the same asymptotic distribution as the one for the parametric LOD score for discrete traits described in the introduction (Chapter 1). In the case of the parametric LOD the estimated parameter is the recombination fraction; this too has a value on the boundary of the parameter space (since the recombination fraction cannot exceed 0.5) under the null.

2.2 Multivariate methodology

The univariate variance component approach can be extended to deal with multiple trait measures. Equation 2.3 can be re-written as

$$\mathbf{y}^* = \mu^* + \mathbf{a}^* + \mathbf{q}^* + \mathbf{e}^*. \tag{2.14}$$

where $\mu^* = (\mu_1, .., \mu_w, .., \mu_1, .., \mu_w)^T$ is the vector of fixed effects,
$\mathbf{a}^* = (a_{11}^*, .., a_{1w}^*, a_{21}^*, .., a_{2w}^*, .., a_{n1}^*, .., a_{nw}^*)^T$ is the vector of additive genetic effects,
$\mathbf{q}^* = (q_{11}^*, .., q_{1w}^*, q_{21}^*, .., q_{2w}^*, .., q_{n1}^*, .., q_{nw}^*)^T$ is the vector of QTL effects and
$\mathbf{e}^* = (e_{11}^*, .., e_{1w}^*, e_{21}^*, .., e_{2w}^*, .., e_{n1}^*, .., e_{nw}^*)^T$ is the vector of environmental effects for traits 1 to w. The phenotypic data is written $\mathbf{y}^* = (y_{11}, .., y_{1w}, y_{21}, .., y_{2w}, .., y_{n1}, .., y_{nw})^T$, where n is the number of individuals. Let $N = nw$. If $w = 1$ then $\mathbf{a}^* = \mathbf{a}$, et cetera as before.

For many traits there will be a correlation between the different trait measures within an individual. Assuming the genetic and environmental components are uncorrelated we can re-write equation 2.6, accounting for the covariances between relatives and between

multiple trait values as

$$\Omega = \mathbf{A} \otimes \mathbf{K}_A + \mathbf{R} \otimes \mathbf{K}_Q + \mathbf{I}_n \otimes \mathbf{K}_E \tag{2.15}$$

where $\mathbf{K}_A$ is a $w \times w$ matrix of additive genetic covariances between records, $\mathbf{K}_Q$ is a $w \times w$ matrix of additive QTL covariances between records and $\mathbf{K}_E$ is a $w \times w$ matrix of environmental covariances between records. $\otimes$ denotes the direct product of two matrices; for example

$$I_n \otimes \mathbf{K}_E = \begin{pmatrix} \begin{pmatrix} [\mathbf{K}_E]_{11} & \cdots & [\mathbf{K}_E]_{1w} \\ \vdots & \ddots & \vdots \\ [\mathbf{K}_E]_{w1} & \cdots & [\mathbf{K}_E]_{ww} \end{pmatrix} & \cdots & \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} & \cdots & \begin{pmatrix} [\mathbf{K}_E]_{11} & \cdots & [\mathbf{K}_E]_{1w} \\ \vdots & \ddots & \vdots \\ [\mathbf{K}_E]_{w1} & \cdots & [\mathbf{K}_E]_{ww} \end{pmatrix} \end{pmatrix}$$

where $[\mathbf{K}_E]_{ij}$ denotes the ij^{th} entry of $\mathbf{K}_E$. Matrices $\mathbf{K}_A$, $\mathbf{K}_Q$ and $\mathbf{K}_E$ have $w(w+1)/2$ (co)variances to estimate when there are w trait measures. For example, with 5 traits there are 15 (co)variances to estimate for each. This model is referred to as the **full multivariate** model.

Estimation of the random effects of interest proceeds in a similar way to that described for the univariate analyses. The variance covariance matrix in equation 2.15 is used in equation 2.7. The vector κ now contains $w(w+1)/2$ (co)variances for each of the random effects a, p and q. As a result there may be substantially more score vectors to calculate and the information matrix may be very large. The computational demands, when there are more than a few traits, will be considerable and alternative methods for dealing with multiple traits will often be required.

Hypothesis testing under the full multivariate model can be performed by appealing to asymptotic results based on known distributions. The main (null) hypothesis of interest is that the QTL (co)variances are all zero. This is compared with the alternative that some are non-zero. For two traits, the likelihood ratio statistic comparing the two hypotheses is distributed asymptotically as $\frac{1}{4}0 : \frac{1}{2}\chi^2_1 : \frac{1}{4}\chi^2_3$. This statistic has this mixture distribution because there are two variance terms and these are on the boundary of the parameter space under the null. When performing the likelihood ratio test, one quarter of the time both of the variances are estimated to be positive (and their covariance can be non-zero), one half of the time one of the variances is at zero and one quarter of the time all 3 (co)variances are at zero. Generalising to w traits the mixture distribution of primary

interest is

$$\text{mixture}_{r=0,w}\left\{\frac{\binom{w}{r}}{2^w}\chi^2_{\frac{r(r+1)}{2}}\right\} \tag{2.16}$$

where $\binom{w}{r} = \frac{w!}{(w-r)!r!}$ is the binomial coefficient. To see this first consider case 9 in the Self and Liang paper [200]. Case 9 states that with w *independent* terms (trait variances here), all on the boundary of their parameter spaces, the asymptotic distribution of the LR statistic will be based on χ^2 distributions with $0, .., w$ degrees of freedom, with the mixing probabilities for χ^2_k component given by $\frac{\binom{w}{k}}{2^w}$. This would hold if we constrained the covariances between the w variance terms (e.g. to be the square root of the product of the two variance terms, see discussion and [146]). With the covariances unconstrained, the mixture distribution must include additional degrees of freedom for the cases in which there are covariances between the variance terms. With r positive variance terms there are $\frac{r(r-1)}{2}$ non-zero covariance terms to estimate. Adding in the r variance terms gives the degrees of freedom ($r + \frac{r(r-1)}{2} = \frac{r(r+1)}{2}$) specified in equation 2.16.

2.2.1 Repeatability Model

A special case of the full multivariate model where there are multiple measurements of the same trait is often called the **repeatability** model. This model assumes that the polygenic, QTL and environmental correlations (across multiple measures) are 1. In this case the computational demands are considerably lower because a single parameter can be used to model the effect of the QTL and polygenic genetic effects. Since there may be environmental effects which are not constant over time there are two effects fitted alongside the QTL and polygenic effects. The first of these, commonly called the permanent environmental effect, models environmental effects that are present in all of an individual's trait measures. The variance associated with this permanent environmental term is labelled σ^2_p. The second effect models the additional environmental effects that are not constant over time; this is the temporary environmental term, with associated variance term denoted σ^2_e. This second term also serves as an error term for other effects not modelled by the other random effects (such as genetic dominance effects).

Phrasing the repeatability model in terms of the full multivariate model, the covariance matrices, $\mathbf{K}_A$ and $\mathbf{K}_Q$ modelling the relationship between the different trait measures in equation 2.15, are now $\mathbf{1}_w\mathbf{1}_w^T\sigma^2_a$ and $\mathbf{1}_w\mathbf{1}_w^T\sigma^2_q$, respectively. The full multivariate covariance matrix is split into two under the repeatability model; the matrix $\mathbf{K}_E$ becomes $\mathbf{1}_w\mathbf{1}_w^T\sigma^2_p + \mathbf{I}_N\sigma^2_e$. The overall variance covariance matrix is hence

$$\Omega = \mathbf{A}\otimes\left(\mathbf{1}_w\mathbf{1}_w^T\sigma^2_a\right) + \mathbf{R}\otimes\left(\mathbf{1}_w\mathbf{1}_w^T\sigma^2_q\right) + \mathbf{I}_n\otimes\left(\mathbf{1}_w\mathbf{1}_w^T\sigma^2_p\right) + \mathbf{I}_N\sigma^2_e \tag{2.17}$$

Estimation is as in the univariate case but with σ_p^2 added to κ and Ω from equation 2.17. Since only one parameter is added cf. the univariate case, parameter estimation is possible in most practical circumstances. Hypothesis testing of this one additional parameter is as in the univariate case.

The ratio of between individual variance $(\sigma_a^2 + \sigma_q^2 + \sigma_p^2)$ to the total variance is often termed the repeatability. Since the repeatability cannot be smaller than the heritability, h^2, the repeatability offers an upper bound for h^2.

2.2.2 Longitudinal Analysis

Although the repeatability model assumption may be a tenable one for a few traits that have multiple measures over time, in most cases it will not be reasonable. Such **longitudinal** traits are likely to change in composition over the life of the individual and are the subject of the remainder of this chapter. For longitudinal traits it is desirable to explicitly model the relationship between age and the genetic and environmental components of the trait. Doing so will enable the components of the trait to be analysed more reliably than in a repeatability analysis as well as providing an estimate of how the trait composition changes over time. To achieve this a multivariate analysis is performed in which it is assumed that something is known about the covariances between the different trait measurements. The main issue is therefore replacing the unstructured covariance structure from the full multivariate model with one which utilises the natural ordering in time of the trait measurements. Kirkpatrick et al. [121] consider such 'function valued' (varying with time) traits, referring to them as infinite dimensional, with infinitely many possible realisations across time. Since in practice the trait may only be observed at a finite number of time points (i.e. $w(w+1)/2$ distinct (co)variances in a $w \times w$ covariance matrix, $\mathbf{G}$), consider a covariance function (CF) linking the covariances as a function of time. A CF, denoted $\Im$, is a continuous function which describes the covariance between any two time points. For ages $t_0$and $t_1$the CF is

$$\Im(t_0, t_1) = \text{cov}(y_{i0}, y_{i1})$$

where y_{i0} and y_{i1}denote the trait values at times t_0 and t_1. In practice, a separate CF is fitted for the QTL effect, the polygenic effect and the permanent environmental effect, with the effects assumed to be independent of each other. Given the assumption of independence, the overall phenotypic CF is given by summing the component CFs.

To estimate CFs from the available data polynomials of age can be used. Henceforth the term order is used to denote the highest power in a polynomial. For example $x^2 + x + 1$ is of order 2, with (assuming all the coefficients are non-zero) 3 terms in the expression. Whilst an order $w - 1$ polynomial will fit the w-trait data exactly by fitting a line through all the points, in reality a smoother curve which ignores stochastic variation is required. In practice orthogonal polynomials are used. Orthogonal polynomials have the advantage of retaining the values of the lower order coefficients when the order of the polynomial fit

is increased; non-orthogonal polynomials can exhibit large changes in the estimated coefficients when there are small changes in the observed phenotypes. Legendre orthogonal polynomials are used here. Such polynomials are defined on (-1,1) and hence the age values of interest are scaled to have maximum value 1 and minimum value -1. An expression for the CF of interest, $\Im$, can be written in terms of the polynomials chosen, $\phi_i(x)$, and a matrix of coefficients, **C**,

$$\Im(t_0, t_1) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} [\mathbf{C}]_{ij} \phi_i(t_0) \phi_j(t_1) \tag{2.18}$$

where k is the number of terms in the polynomial (i.e. the order of the polynomial, call this k^*, is $k-1$) chosen and t_0 and t_1 are the scaled ages.

Kirkpatrick et al. [121] propose a method whereby one can estimate the matrix of coefficients, **C**, from the observed data. This coefficient matrix can then be inserted into equation 2.18 to obtain an estimate of the CF (as a function of age). This function can be evaluated at any age of interest, allowing the covariance to be estimated at any age (not just at the ages present in the data).

Since this function can be evaluated at any age of interest we are not restricted to considering only the ages present in the data set.

The method proposed by Kirkpatrick et al. [121] to estimate the coefficient matrix involves rearranging expression 2.18. Letting $[\Phi]_{ij} = \phi_j(t_i)$ (numbering the matrix indices 0 to k-1), the elements of $\Im(t_0, t_1)$ can be written in the form of a covariance matrix, **G**,

$$\mathbf{G} = \Phi \mathbf{C} \Phi^T. \tag{2.19}$$

Whilst this equation can be solved for **C** when the number of terms in the polynomial k is equal to the number of trait measures (w), when $w > k$ Φ is not invertible and estimation of **C** is not trivial (Kirkpatrick, [121] Appendix A). An alternative method for estimating **C** is now considered.

Estimation of the coefficient matrix in a general pedigree using Random Regression

Meyer [156] explains how the coefficients of the matrix **C** above can be estimated if one utilises random regression. Now follows a short explanation of Random Regression (RR).

Consider a basic mixed model in which there is a single random factor of interest alongside fixed effects such as sex. The interest is in the deviations of the random effects from the base level of the fixed effects. That is, the distribution of the random effects is the primary focus. Consider the case where the phenotype is known to change with the level of some factor such as age and assume the effect of age on phenotype is linear. The change in age can be accounted for by fitting age as a fixed effect. The interest is now in the deviations about this fixed regression line; consider now deviations about the linear term and

about the constant term. In the following equation, i is the individual and j is the measure. f_0 and f_1 are the fixed effects whilst the random effects of interest are a_{i0}, the usual random effect and a_{i1}, the random effect with linear age dependence. y_{ij} is the phenotype and e_{ij} is the error term.

$$y_{ij} = f_0 + f_1 t_j + a_{i0} + a_{i1} t_j + e_{ij} \tag{2.20}$$

Additional terms allowing for permanent environmental effects, $p_{i\bullet}$, and for QTL effects, $q_{i\bullet}$, may be added. The QTL effects can be estimated by utilising marker information in a similar way to the univariate analyses in section 2.1. The practical problems that arise in the implementation of this are discussed in section 3.2.1. Allowing all three sets of random effects to vary with time yields

$$y_{ij} = f_0 + f_1 t_j + a_{i0} + a_{i1} t_j + p_{i0} + p_{i1} t_j + q_{i0} + q_{i1} t_j + e_{ij}\,. \tag{2.21}$$

Although the two equations above fit mean and a linear terms for the fixed effects and random effects, there is no reason why these two should have the same number of terms. In practice, even if one is only interested in linear deviations from the fixed effects, it may be best to fit a higher order polynomial for the fixed effects to ensure that all systematic effects (e.g. the effects of age on the mean function) are removed before the deviations are considered.

Usefulness of RR Random regression is useful because the covariance between polynomials of age in a random regression can be related to the covariance function coefficients of interest. The random regression model which allows this is

$$y_{ij} = \mu + \sum_{m=0}^{k_a-1} a_{im}\phi_m(t_{ij}) + \sum_{m=0}^{k_p-1} p_{im}\phi_m(t_{ij}) + \sum_{m=0}^{k_q-1} q_{im}\phi_m(t_{ij}) + e_{ij}\,. \tag{2.22}$$

Equation 2.22 is an extension of 2.21 to arbitrary polynomial orders. k_a, k_p and k_q denote the number of terms in each polynomial (=order of the polynomial -1) for the additive genetic, permanent environmental and QTL effects, respectively. t_{ij} is the time at which the measure y_{ij} is taken; this is a generalisation of the t_j in equation 2.21, allowing different individuals to be measured at different ages. Each individual now has w_i measures, where

it is possible that $w_i \neq w$ for some individuals. The covariance structure of such a model is

$$Cov(y_{ij}, y_{ij'}) = \sum_{m=0}^{k_a-1} \sum_{l=0}^{k_a-1} Cov(a_{im}, a_{il})\phi_m(t_{ij})\phi_l(t_{ij'}) + \tag{2.23}$$

$$\sum_{m=0}^{k_p-1} \sum_{l=0}^{k_p-1} Cov(p_{im}, p_{il})\phi_m(t_{ij})\phi_l(t_{ij'}) + \tag{2.24}$$

$$\sum_{m=0}^{k_q-1} \sum_{l=0}^{k_q-1} Cov(q_{im}, q_{il})\phi_m(t_{ij})\phi_l(t_{ij'}) + \tag{2.25}$$

$$+Cov(e_{ij}, e_{ij'}) \tag{2.26}$$

Each of the covariance terms 2.23, 2.24 and 2.25 can now be seen to be of the same form as equation 2.18. If we can estimate these covariance terms in a random regression these can be used directly in equation 2.19 to obtain a covariance matrix for the additive genetic, permanent environment and QTL effects.

To fit the RR model the full multivariate model is re-parameterised. In this re-parameterization we replace the set of trait measures with an order k polynomial for each effect of interest (permanent environment, polygenic, QTL). The full multivariate model is then fitted with these polynomial coefficients regarded as correlated traits. To do this, begin by writing equation 2.22 in matrix notation

$$\mathbf{y}^R = \mu^R + \mathbf{Z}_A\mathbf{a}^R + \mathbf{Z}_Q\mathbf{q}^R + \mathbf{Z}_P\mathbf{p}^R + \mathbf{e}^R$$

where $\mathbf{y}^R = (y_{11}, .., y_{1w_1}, y_{21}, .., y_{2w_2}, .., y_{n1}, .., y_{nw_n})^T$ are the phenotypes,
$\mu^* = (\mu_1, .., \mu_{w_1}, .., \mu_1, .., \mu_{w_n})^T$ is the vector of fixed effects,
$\mathbf{a}^R = (a_{10}, .., a_{1(k_a-1)}, .., a_{n0}, .., a_{n(k_a-1)})^T$ is the $k_a \times n$ vector of polygenic random regression coefficients, $\mathbf{q}^R = (q_{10}, .., q_{1(k_a-1)}, .., q_{n0}, .., q_{n(k_a-1)})^T$ is the $k_q \times n$ vector of QTL random regression coefficients, $\mathbf{p}^R = (p_{10}, .., p_{1(k_a-1)}, .., p_{n0}, .., p_{n(k_a-1)})^T$ is the $k_p \times n$ vector of permanent environmental random regression coefficients and $\mathbf{e}^R$is the $\Sigma_{i=1}^n w_i$ vector of temporary environmental terms (note this is $w \times n$ if all n individuals are measured for all traits, i.e. if $w_i = w$ for all i). $\mathbf{Z}_A$ is a $\Sigma_{i=1}^n w_i$ by nk_a matrix of orthogonal polynomial coefficients,

$$\mathbf{Z}_A = \begin{pmatrix} \phi_0(t_{11}) & \cdots & \phi_{k_a-1}(t_{11}) & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \cdots & 0 & 0 & 0 \\ \phi_0(t_{1w_1}) & \cdots & \phi_{k_a-1}(t_{1w_1}) & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & \phi_0(t_{n1}) & \cdots & \phi_{k_a-1}(t_{n1}) \\ 0 & 0 & 0 & \cdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \phi_0(t_{nw_n}) & \cdots & \phi_{k_a-1}(t_{nw_n}) \end{pmatrix}.$$

$\mathbf{Z}_Q$ and $\mathbf{Z}_P$ are defined similarly, with k_a replaced by k_q or k_p. The covariance terms for the vector $\mathbf{a}^R$ are given in equation 2.23 and, assuming the systematic age effects have been removed by the fixed effects, can be written as $\mathbf{a}^R \sim N(0, \mathbf{A} \otimes \mathbf{K}_A^R)$, where $\mathbf{K}_A^R$ is the $k_a \times k_a$ matrix of CF coefficients (named C above) for the polygenic effects. In a similar fashion $\mathbf{q}^R \sim N(0, \mathbf{R} \otimes \mathbf{K}_Q^R)$ and $\mathbf{p}^R \sim N(0, \mathbf{I}_n \otimes \mathbf{K}_P^R)$. Written as a full variance-covariance matrix,

$$\Omega = \mathbf{Z}_A(\mathbf{A} \otimes \mathbf{K}_A^R)\mathbf{Z}_A^T + \mathbf{Z}_Q(\mathbf{R} \otimes \mathbf{K}_Q^R)\mathbf{Z}_Q^T + \mathbf{Z}_P(\mathbf{I}_n \otimes \mathbf{K}_P^R)\mathbf{Z}_P^T + \sigma_e^2 \mathbf{I}_{\sum_{i=1}^n w_i}$$

where σ_e^2 is the temporary environmental variance term. Estimation is as in the multivariate case. κ now has $k_a(k_a + 1)/2$ entries for the polygenic effect and equivalent terms for the QTL and permanent environmental cases.

Details of the process of hypothesis testing and model selection for longitudinal models is deferred to chapter 3 (Simulation Chapter).

2.3 Discussion

The discussion first looks at issues in univariate analysis before considering issues particular to multivariate methods.

2.3.1 Univariate

This chapter described variance components methodology, suitable for QTL analysis with complex pedigrees. The method allows for straightforward removal of environmental effects through the fitting of fixed effects. The univariate techniques are tractable with all but the smallest data sets. The univariate methods have been shown to yield unbiased estimates of the variance components of interest [6] and are now routinely applied to human, livestock and natural population data sets. However, the estimation procedure relies upon the assumption of normality of the data and under deviations from this the likelihood ratio statistics calculated may be biased [3]. This bias has been shown to depend upon the degree of kurtosis of the data [29]. Furthermore, whilst single point estimates of QTL specific variance may be accurate, if a large number of genomic locations are tested, selecting the highest test statistic will result in upward biases in the QTL specific variance at the test statistic maximum [21]. This is due to the strong correlation between the magnitude of the test statistic for significance of the QTL and the size of the QTL in terms of variance explained. Additionally, the stochastic variation in each individual study leads to some QTL being more readily detectable in that sample than other QTL; the effect size of the QTL that are detected will hence be overestimated. It is not uncommon for the QTL specific variances to be so overestimated at the LOD score peaks from genome scans that the (additive) QTL explains all of the additive genetic variation (i.e. the polygenic variance is estimated as being zero at the LOD peak). This can be seen in the univariate analyses of the GAW13 data (Chapter 4). Categorical data can be analysed

using the described univariate methods through the use of a threshold model (Chapter 25 of [139], [60]). Threshold models assume there is a continuous distribution underlying the observed categories and that parameter estimation can be performed by fitting some function (such as the probit or logit function) which maps the categories to the continuous distribution. A threshold model is used to allow analyses of binary disease outcomes in Chapter 5 and further details are given in that chapter. The model described in section 2.1 can also be extended to deal with dominance and epistatic effects [6] (this can be applied to both polygenic and QTL effects).

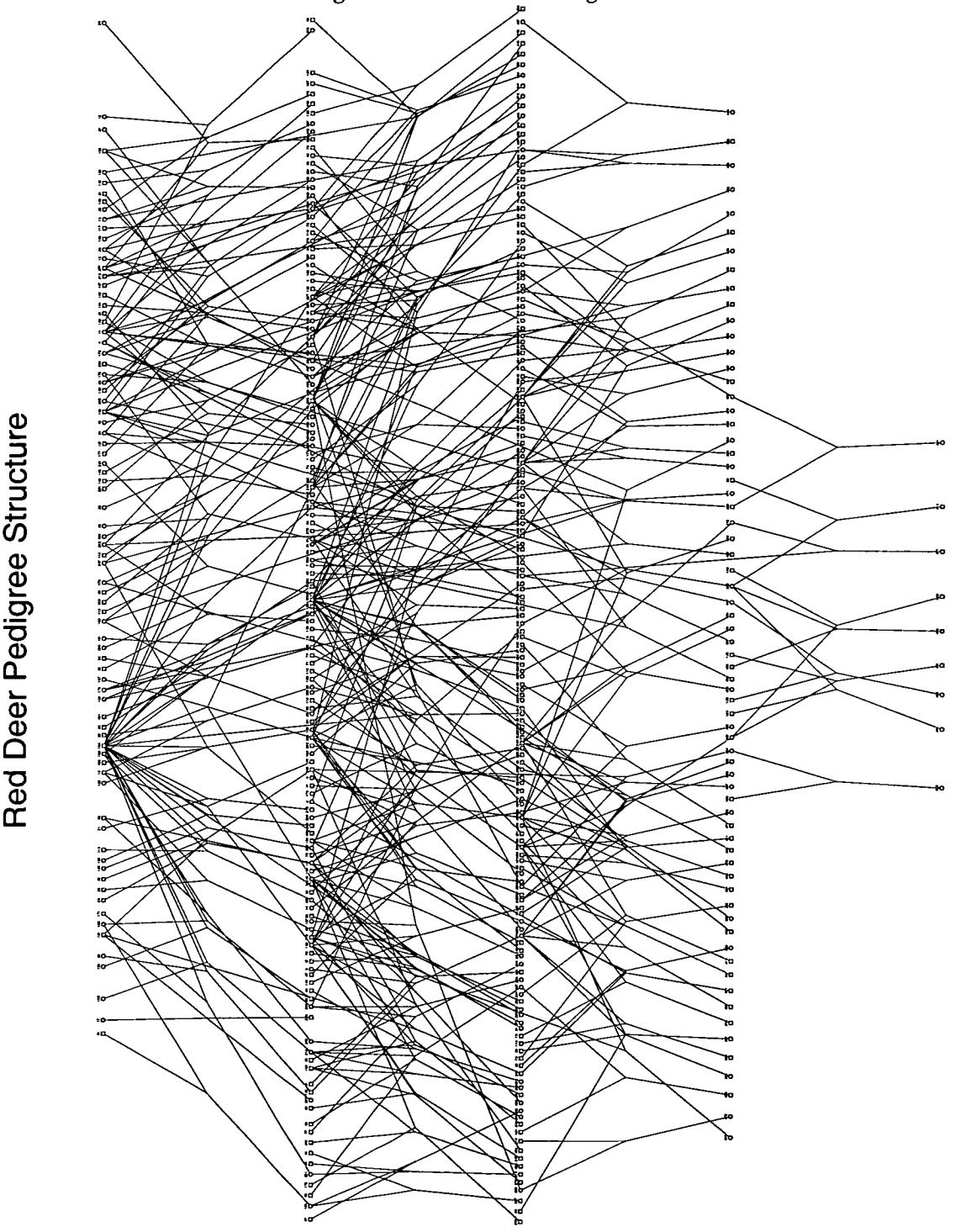
A basic component of all variance component linkage (QTL) analyses is the marker information matrix R. Estimation of R from multiple markers is not trivial in large pedigrees and may constitute a significant part of the computational burden. A number of different methods for the computation of marker specific IBD coefficients have been proposed and subsequently implemented. The methods can be split into exact methods [176, 174, 1] and Markov chain Monte Carlo (MCMC) based approximation methods [211, 99, 6] (see also chapter 1, introduction). Measuring pedigree complexity as twice the number of non-founders minus the number of founders (number of pedigree 'bits', [127]), exact methods can generally be used for pedigrees of less than 30 bits; the major determinants of computational cost for IBD estimation are number of markers, marker spacing, proportion of untyped founder individuals and number of pedigree loops. There have been studies of the relative performances of the available methods. One study [212] compared two of the MCMC approximations (SIMWALK2 [211], SOLAR [6]) with one of the exact methods (Genehunter [176]). Another [174] compared a deterministic approach with the approach implemented in Loki [99]. The main conclusions of these studies were that SOLAR was less accurate than either Genehunter or SIMWALK2. However, Genehunter does not work on large pedigrees and SOLAR is faster than SIMWALK2. Loki was found to give similar results to those from the deterministic approach employed in [174]. There have been no comparisons in the literature between SOLAR/SIMWALK2 and Loki. SOLAR is quicker than some of the other methods because it does not attempt to use multiple marker information simultaneously. Instead, it first calculates single marker IBD coefficients. It uses these single point estimates in a weighted regression, allowing estimation of IBD coefficients at points between the available markers. The weights in the regression are dependent upon a set of formulae which use the relationships between individuals. For this reason SOLAR will not work on arbitrarily large pedigrees.

IBD estimation for the Framingham data (Chapter 4) was performed using SOLAR and took 1 week of computing time on a 700MHz Intel Pentium processor. Although it would be preferable to estimate the variance components and IBD coefficients simultaneously this is likely to take orders of magnitude longer than the two step procedure [79, 6] and hence be untenable for many data sets. With the two stage procedure the IBDs can be calculated and stored for use in analyses. This was crucial in the Framingham analyses as a number of traits and analysis methods were then used.

IBD estimation was investigated as part of a project investigating a quantitative trait,

birth weight, in Red Deer (*Cervus Elaphus*) [209]. My contribution to this paper was to investigate the calculation of IBD estimates in a very complex deer pedigree (a picture of the pedigree is in figure 2.1; this was drawn with the program pedfiddler `http://www.stat.washington.edu/thompson/Genepi/Pedfiddler.shtml`) and to compare two possible analysis methods. The two methods were; a regression based analysis [123] based on splitting the full pedigree into 17 half sib families (program QTL Express [198]) and a VC analysis of the full pedigree, performed as described in section 2.1. For one of the detected QTL there were discrepancies between the results obtained from the two methods (although there was good agreement for the two other QTL detected). These discrepancies may have been due in part to problems with marker specific IBD estimation in such a large pedigree. To investigate, the IBD estimation was performed with SOLAR and with Loki and the results compared. The estimates from SOLAR were single marker based because, as a result of the pedigree having some animals with very distant relationships, the regression based multipoint estimation procedure failed in SOLAR. Note however that the authors of SOLAR have previous reported [254] multipoint linkage analyses of a pedigree of similar complexity to the one in figure 2.1 (i.e. they are likely to have used a newer, not publicly available, version of the program). The IBD estimates from Loki used all of the marker information simultaneously. The IBD estimates from Loki required extensive reformatting before they could be incorporated into the SOLAR routines for likelihood maximisation. Although there were minor differences in the IBD estimates obtained from the two programs, the test statistics based on them were similar, indicating that the discrepancies between the half sib and full pedigree analyses were not caused by problems of IBD estimation in the full pedigree. In a small number of cases the full pedigree likelihoods (parameters estimated under the QTL plus polygenic effects model and under the polygenic effects only model) were maximised in SOLAR and in a maximisation program written by P.M. Visscher (University of Edinburgh). In all cases both programs gave very similar test statistics (this part of the analysis was done by P.M. Visscher). This meant that neither differences in the the maximisation procedures nor IBD computation issues could explain the differences in the results obtained. The regression and VC methods make different assumptions about the underlying model and it seems likely that this explains at least some of the discrepancy. The regression approach assumes a biallelic QTL segregating in the half sib families with the QTL fitted as a fixed effect. This QTL effect is assumed to fit a genetic model with a 'substitution effect', the effect of replacing one allele from a common parent with the other in the offspring. The regression approach does not account for any background (polygenic) variation within the half sib families. In comparison, the VC approach considers the whole pedigree at once, makes no assumptions about the number of QTL alleles and, instead of assuming the QTL is a biallelic fixed effect, the QTL is assumed to be a random effect (i.e. drawn from a distribution of possible effects). In the VC approach the phenotypic values are assumed to be drawn from multivariate normal distribution (see section 2.1); both the polygenic and the QTL effects are assumed to be additive (dominance effects can also be included in the VC approach but

Figure 2.1: Red Deer Pedigree



estimation of the required marker specific matrix of two allele sharing IBD coefficients is not implemented in any programs at present). Since the VC approach was able to handle the whole pedigree at once, the number of phenotyped animals in the VC analysis was rather larger than the number included in the half sib regression analysis (295 animals in the full pedigree, just over 100 in the half sib families). Applying the VC analysis to just the half sib families (i.e. ignoring the known relationships between the half sib-ships and ignoring known full sib relationships) increased the concordance between the results of the two analysis methods but some discrepancies remained. These discrepancies are therefore likely to be attributable to the different assumptions required to apply the two methods.

The power of univariate VC methods to detect QTL has been studied by a number of authors. Analytic power calculations have been performed in [250, 186, 238] whilst simulation studies appeared in [165, 251]; if the data have been ascertained irrespective of individuals' trait values these studies provide a guide to the available power. However, in some cases individuals are selected for analysis on the basis of the trait of interest and this may render such power calculations invalid. The type I error of the test statistics calculated from a VC analysis will be somewhat inflated in highly selected samples [3]. Conditioning on the trait values [204, 32, 49, 103] is the most popular correction for possible ascertainment bias and this is implemented in programs such as SOLAR. Selection schemes for sib pairs have been proposed [193] and may offer additional power if phenotyping is inexpensive relative to genotyping. However, this is generally not the case in human genetic studies. If genotyping is substantially more expensive than phenotyping there may be financial benefit in only genotyping individuals at the extremes of the phenotypic distribution.

The main alternative to VC methods for human quantitative trait data are methods based on functions of sib pair phenotypic measures. The initial method regressed sib pair differences on IBD proportions [97] with subsequent approaches using other functions of sib pair measures [65]. Such methods have the advantage of being considerably simpler computationally than VC based methods but they cannot be readily extended to large pedigree structures. In the case of sib pair only data, VC and regression based sib pair analysis have been shown to be asymptotically equivalent [202], provided the component of variance attributable to the QTL is relatively small. Attempts have been made to extend regression based methods to general pedigrees [203] but such methods are only applicable to relatively small pedigrees in practice and require further work to assess their utility. Regression based approaches have been shown to offer less power than VC approaches [249, 11] in extended pedigrees.

2.3.2 Multivariate

This chapter extended the univariate techniques to consider data sets with multiple trait measures. The multivariate techniques required to effectively analyse such data are more involved than those for single trait measures. This, together with the relative paucity of

suitable data, goes some way toward explaining the lack of research in this area. In this chapter, particular attention has been paid to data with multiple trait measures of the the same trait over time (longitudinal traits). Such traits are often not well described by single, cross sectional, phenotypic measures but, as has been described, the conceptually simple full multivariate model requires the estimation of large numbers of parameters when there are more than a few time points. Since the data sets commonly available for genetic studies in natural populations are small, the full multivariate approach has somewhat limited application. By their very nature, longitudinal traits will be relatively highly correlated across multiple measures of the same trait compared with non-longitudinal multivariate measures (e.g. multivariate analysis of height and weight, say). In Chapter 3 it will be seen that when traits are highly correlated the estimation of large numbers of parameters is difficult. The covariance function based approach may have considerably more utility than the full multivariate model as it can reduce the number of parameters in the model. Fitting a polynomial with order plus one (i.e. k terms in the polynomial) equal to the number of age points in the data is equivalent to a full multivariate model. Fitting lower order polynomials smooths the estimated covariance function, removing individual deviations which are likely to be due to stochastic variation. Testing procedures, suitable for choosing the order of polynomial used for each effect of interest (permanent environment, polygenic, QTL), are discussed in chapter 3.2.4. In reality, it may not be possible to fit a number of different polynomial orders to the data and choosing an particular polynomial order (say linear or quadratic) in advance may be a reasonable procedure (see also the discussion of chapter 3). The covariance function approach will be particularly useful when the data are measured at a large number of ages, perhaps with irregular gaps between measures; this is because the approach fits a polynomial through the set of ages available for each individual. Furthermore, individuals only measured for a few ages can still contribute to the analysis by providing information on the coefficients of the lower order polynomials (information available on constant and linear terms when there are two age measures and so on). These advantages are very well illustrated in chapter 4 where there were 76 different ages in the data set with individuals measured for between 1 and 21 of these ages.

For some traits the repeatability model described above will be suitable for multivariate data. This method is considerably simpler than the other multivariate methods with few parameters to estimate. The information loss in using this model compared with the more complex multivariate models is the subject of Chapter 3.

There have been a number of other methods proposed to allow analyses of multivariate data. In most cases these are for distinct multiple traits (height, weight, etc) rather than longitudinal ones (height at age 20, at age 30,...). The simplest approach involves performing separate univariate analyses for each trait. This approach does not take advantage of the potential power gains inherent in the multivariate structure of the data. Furthermore, it is unclear how to keep the significance level at the desired level when there are multiple tests. A Bonferroni correction can be readily applied but this is almost certain to be overly

conservative. The next simplest alternative is to transform the multiple trait values into a single summary or composite measure, thus allowing a single univariate analysis method to be used. This composite measure can be constructed such that the calculated 'factor score' maximises some parameter of interest, such as the heritability [33]. Furthermore, a multivariate segregation analysis has been proposed for pedigree data [28] and this may allow the construction of a composite measure that is particularly suitable for mapping the major gene affecting a trait. However, even in this second case where there may be more power to detect a particular QTL or locus, neither method is likely to give an optimal composite measure for other QTL or loci [61].

A number of authors have considered extensions of the sib pair regression methods to multivariate data [105, 4, 10, 51, 159]. Such methods offer advantages over the VC based multivariate approaches (introduced in this chapter) in terms of computational ease but, in addition to their unsuitability for extended families, they have been shown to offer less power than VC based approaches (for bivariate data [9]). The power of the sib pair regression methods has also been discussed in [68]. One of the regression based papers [105] suggested a method suitable for bivariate sib pair data in which there is both a quantitative trait and a qualitative trait. Such techniques may be useful for some psychiatric diseases in which there are endophenotypes. For example, P300 measures (a quantitative trait measuring event-related potential amplitude and latency on the scalp) are often found to be higher in individuals affected by schizophrenia than in their unaffected relatives [25] and incorporation of this information into an analysis may improve power compared with methods which only utilise affection status. Methods for joint qualitative-quantitative trait analysis have also been proposed for VC based bivariate analysis [248, 252]; such methods are suitable for extended family data. Extensions to multivariate (more than 2 traits) joint analysis requires further research.

Attempts have been made to fit the full multivariate model (described in section 2.2) to longitudinal data [50, 51]. In both papers the model is fitted to trivariate data but the six parameters (three variances, three covariances) could not be estimated simultaneously for all of the random effects. When the situation was approximated by three bivariate analyses, parameter estimation was possible. Given the data in [50] and [51] only support the estimation of three parameters it would probably be better to fit a first order CF to the full set of three traits than to fit 3 separate full multivariate analyses to three different subsets of the data.

Multivariate linkage analysis related to that described in section 2.2 has been described for sib pair data [61] and applied to developmental dyslexia data [146]. The method used in [146] fitted the polygenic effect as in section 2.2 but the covariance structure of the random effect for the QTL was constrained such that correlation between any two trait measures was equal to one. This is equivalent to the restriction that

$$cov(trait_i, trait_j) = \sqrt{\sigma^2_{trait_i}\sigma^2_{trait_j}}$$

for all traits i and j. This means that there are only k parameters to estimate when there are k traits (compared with $k(k+1)/2$ with an unstructured QTL covariance matrix). Whilst this is highly unlikely to be true for all but the most strongly related traits, this model may allow parameter estimation in cases in which there are limited amounts of data (as in [146]).

Another simplification of the full multivariate model may be possible for longitudinal traits. Consider the case where the primary focus is in whether there is a change in variance over time (e.g. the gene has large effect in early life but its importance with respect to the trait decreases over the life of the individual) and the covariance terms between different ages are of little or no interest. Take for example the full multivariate model applied to data with 5 ages. This would require the estimation of 15 (co)variances (5 variances, 10 covariances). The likelihood of this model can be compared with the likelihood of a (reduced) model in which the diagonal elements are all constrained to be equal (requires the estimation of 10 covariances and 1 variance). Consider the matrix of QTL specific (co)variances between the trait at the different ages, D*, $\begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} \\ d_{41} & d_{42} & d_{43} & d_{44} & d_{45} \\ d_{51} & d_{52} & d_{53} & d_{54} & d_{55} \end{pmatrix}$. The null hypothesis would be $d_{11} = d_{22} = d_{33} = d_{44} = d_{55}$ with the alternative being that the d_{ii}s are unequal. The likelihood ratio test for deviations from the null should be compared with a χ^2_4 distribution. Fitting the full model with ASREML [80] was sometimes computationally possible with the 150 family data set described in chapter 3 but, in practice, achieving convergence was often difficult. Instead of using a likelihood ratio test to test the significance of this model, however, a Score test could be used (see also [109, 237]). This test is based upon the score and the information matrix of the reduced model and hence has the advantage of not requiring the calculation of the maximum likelihood under the full model (this is required for a likelihood ratio test). ASREML generates the Score and Information matrix for this test and should give results that converge asymptotically to those obtained using the likelihood ratio test. Performing a score test in this way reduces the number of parameters requiring estimation from $\frac{w(w+1)}{2}$ (necessary for full model) to $\frac{w^2-w+2}{2}$ (necessary for reduced model) when there are w traits. Although this full multivariate approach allows testing of the hypothesis that there is a change in genetic variance over time it is clearly inefficient in terms of the number of parameters used. CF based models are less flexible in that they prescribe a particular covariance function at the same time as the variance terms but they are more efficient in terms of the number of parameters requiring estimation.

Techniques for longitudinal QTL mapping in experimental crosses using character process [173] modelling have been developed [140]. However the techniques in [140] are not readily extended to the irregular family structures encountered in human genetic studies. In contrast, the CF based techniques can be applied to arbitrary family structures (see

chapter 4).

A multivariate CF model may be possible for a set of longitudinally measured traits, e.g. weight at age 20,.., weight at age 50, height at age 20,.., height at age 50. The longitudinal element of each trait could be fitted with the RR described above, with the covariances between the RR parameters for different traits modelled as in the full multivariate analysis [156]. It seems unlikely however, that there will be many natural population data sets large enough to estimate the large number of parameters in such a model.

Summary In summary, univariate variance component techniques can be applied to general pedigree data. The mixed model framework allows flexible modelling of both the quantitative and discrete trait values and allows covariate information to be included. There have been a number of extensions of the univariate analyses to multiple traits. Here, particular attention has been paid to longitudinally measured traits. Such traits can be efficiently (in terms of number of parameters) modelled using covariance functions, with suitably parameterised random regressions allowing parameter estimation. In chapter 3 the relative power of the methods in section 2.2 are compared using computer simulation. The covariance function based methods are applied to a real data set in chapter 4.

Chapter 3

Analysis of longitudinal quantitative trait data in complex pedigrees: Simulation

3.1 Introduction

To assess the utility of a selection of the techniques available (detailed in chapter 2) for longitudinal data, computer simulations were run. Pedigree data were simulated with individuals assigned genotypes and multiple (longitudinal) trait values. The main interest was in QTL detection and characterisation in samples of moderate size. The samples ascertained for QTL analysis in humans are typically rather small. This can restrict the application of very complex multivariate techniques that involve large numbers of parameters.

Two sets of simulations were run. The first considered a simple model for the assigned trait values with the genetic correlations between different ages equal to one. This simulation was useful because it facilitated simple assessment of the power to detect QTL using univariate, repeatability (denoted Re, see section 2.2.1) and random regression based CF methods (denoted RR, see section 2.2.2). In particular, this first simulation set considered whether it is possible to detect changes in QTL variance over time. The adequacy of the asymptotic likelihood ratio tests was also investigated using these data.

The second set of simulations used a more realistic model for the phenotypic data in which the genetic correlation between QTL effects at different ages was not restricted to be one. Under this model, methods which do not model the covariance between the trait values at different ages (such as repeatability or univariate analysis) were expected to perform poorly and the main comparisons were between RR and full multivariate analyses.

Simulations were also used to investigate different methods for overcoming the practical difficulties arising in the incorporation of IBD information into multivariate analyses.

3.2 Methods

3.2.1 Identity by Descent (IBD) coefficient estimation

IBD matrices were computed for the generated set of pedigrees using SOLAR [6]. SOLAR also performs the maximisation necessary for evaluation of equation 2.7. However SOLAR does not perform multivariate analyses. ASREML [80] was used for the multivariate analyses. The IBD matrices computed in SOLAR required inversion for incorporation into ASREML. In some cases the IBD matrices are singular. For example, for a 2 sib nuclear family with perfect marker information (i.e. parents heterozygous for different alleles) the QTL IBD matrix will be singular if the children share no or both alleles IBD at a marker. However, if one uses multipoint IBDs and evaluates these a small distance from each marker the resultant matrix will have entries that deviate slightly from 0, 0.5 or 1 (in the 2 sib nuclear family example) and will become invertible. Alternatively one may add a small positive number to the diagonal entries of the IBD matrix to ensure the matrix is invertible. If there is more than one family in the data set the IBD matrix will be block diagonal; only the diagonal entries of the families which generate singular sub-matrices at that genomic location need to be modified.

Since SOLAR does not require the inverse of the IBD matrix and allows univariate QTL analyses, the results from SOLAR were compared with the results of univariate QTL analyses performed in ASREML (after manipulation to make the IBD matrices non-singular). It is presumed that if the (singular) IBD matrices can be successfully incorporated into a univariate analysis they will also be suitable for a multivariate QTL analysis.

Univariate analyses were done on two different simulated data structures. The first of these was a set of 200 4 sib nuclear families. The LOD scores and variance components were calculated in SOLAR and in ASREML with the two methods used to render the IBDs suitable for inclusion in ASREML. The second data structure was 200 2 sib nuclear families. In the first case about 90% of families were expected to yield singular IBD matrices when IBDs were calculated at a marker completely linked to the simulated trait locus (see appendix of this chapter). In the second case fewer of the families (~45%) would be expected to have singular IBD matrices (see appendix of this chapter).

Firstly, the IBD matrices were rendered suitable for inclusion in ASREML (made non-singular) by adding 0.001 to the diagonal element of all 200 families. Secondly, the singular sub-matrices of the full 200 family IBD matrix were identified and individually modified to make them non-singular; the non-singular sub-matrices were unchanged. 10 replicates were run in each case.

The effect of adding different values to diagonal entries of singular sub-matrices was investigated by varying the value added from 0.1 to 10^{-5}. Adding too much to the diagonal will cause the matrix to mis-represent the true marker information whilst adding too little will cause computational problems due to the matrix being close to being singular. 10 sets of 100 4 sib nuclear families were analysed. The simulated QTL specific $h^2$was 0.25.

An alternative to modifying the diagonal elements of the IBD matrix (or relevant sub-

matrices) is *bending* [98, 214]. Bending refers to a procedure which modifies a non-positive definite matrix of interest to make it positive definite. In particular, the bending process alters the eigenvalues of the matrix; if the matrix is to be positive definite, all of the eigenvalues must be greater than zero. If one gradually alters the eigenvalues toward their mean until they are all positive and reconstructs the matrix one will (hopefully) obtain a matrix which has similar numerical properties to the original matrix yet is non-singular.

Matrix inversions were done in GNU OCTAVE (`www.octave.org`).

3.2.2 Simulating data

To create data sets for analysis, random effects were drawn from normal distributions. If the random effects are drawn from normal distributions then the overall trait value will have a normal distribution, satisfying a basic assumption of the ML based estimation procedure. Environmental and genetics effects were assumed to be independent and were hence added sequentially.

Environmental effects The permanent environmental effects were generated by adding a single normal variate from N(0,0.5) to each individuals' set of trait measures. Similarly, the temporary environmental effects were generated by adding a normal variate from N(0,0.5) to each separate trait measure in each individual. In simulation 1, the effect of changing these distributions to permanent environment, N(0,0.75) and temporary environment, N(0,0.25), was investigated.

Genetic effects No polygenic effects were simulated. QTL effects were simulated using three methods.

1 Repeatability model For each founder individual two separate allelic effects were drawn from a univariate normal distribution with the required variance. These allelic effects were passed on to the non-founders (descendants) with a completely linked highly informative multi-allelic (20 alleles) marker. Once the first set of non-founders have been allocated genotypes and phenotypes any subsequent descendants can be given values in turn. Arbitrary pedigrees can hence be given simulated values. In contrast with a polygenic genetic effect, where effect transmitted is composed of both a Mendelian sampling component and two parental components (averaged over all genes), the (genotypic) QTL effect comes solely from the two transmitted allelic effects. The total variance attributable to the QTL of interest is the sum of the variances of these two allelic effects.

2 Change in variance but correlations across ages equal to one In this case data were generated as in the repeatability model but a change in variance over time was induced by multiplying the two generated allelic effects in each founder by some function of age (e.g. QTL variance is 0.2 at age 1 and increases linearly to 0.4 at age 5).

3 Change in variance with general correlation structure In this case the two allelic effects for each founder were drawn from a multivariate normal distribution with specified variances and correlations. To generate the required multivariate normal (MVN) variates consider a $p \times p$ correlation matrix **K** and a diagonal matrix **D** of the required standard deviations. Let **S** = **DKD** and re-write **S** using the Cholesky factorisation as **S** = **L'L** (`http://mathworld.wolfram.com/CholeskyDecomposition.html`). For a $n \times p$ matrix **X** of p univariate N(0,1) draws (ages) measured in n individuals, the matrix **B** = **XL'** has multivariate normal distribution with the correlations and variances specified in **K** and **D**.

For the simulated data, no systematic change over time is simulated so whilst a constant (f_0) and age dependent (f_1) overall mean are fitted in the RR analyses they are expected to both yield estimates that are close to 0.

In simulation 1 (below), QTL effects were simulated using methods 1 and 2. In simulations 2a and 2b (below), the QTL effects were generated using the multivariate normal distribution described in method 3. In all simulations 150 4 sib nuclear families (900 individuals) were simulated. All individuals had phenotypic and genotypic information. There were 5 ages, numbered 1 to 5.

3.2.3 Simulation 1

Basic model

In simulation 1 the data were simulated using methods 1 and 2 described above. Under these models the genetic (QTL) correlation between different ages is generated to be one. Although these are unlikely to be realistic models this set up allows a simple test for deviations from the repeatability model (in which the genetic variance does not change over time and the genetic correlations are all one). Testing for deviations from the repeatability model forms the first part of simulation 1. The second part of simulation 1 considers the power of RR based methods to detect QTL (QTL plus polygenic effect versus polygenic only).

Deviations from Repeatability model

The main interest here is in whether the repeatability model (i.e. a model only fitting a single effect, q_{i0}, for the QTL) is appropriate to the generated data or whether adding additional terms to the RR (q_{i1} in this case) significantly improves the fit to the data. The null is therefore $q_{i1} = 0$ and the alternative is $q_{i1} > 0$.

Null model The adequacy of using asymptotic results based on known distributions was evaluated by simulating data under the null hypothesis. The empirical null distribution (repeatability) was evaluated by simulating data using method 1. The QTL variance was set to 0.2 at all ages. A linear RR was fitted to the data and compared with the repeatability model. If one fits the linear term q_{i1} in the RR and constrains the covariance between

q_{i0} and q_{i1} to be zero, twice the log likelihood difference (2logLR, call this **StatDevRe1** (statistic for **dev**iations from **re**peatability model), see table 3.1) between the RR and the Re model is expected to be distributed as $\frac{1}{2}\chi^2_1 : \frac{1}{2}0$ [200]. This follows because under the null the additional variance term is on the boundary of the parameter space. If one fits both the variance and the covariance terms in the RR (subject to the constraint that the coefficient matrix remains positive definite), the 2logLR test statistic (call this **StatDevRe2**) for the RR versus the Re model is a 50:50 mixture of χ^2_2 and a point mass at zero (χ^2_0). Note this test statistic (with the covariance unconstrained) is not a mixture of χ^2_2 and χ^2_1 (this appears to be what is stated in [221, 156]). This is because when the variance term associated with the q_{i1} term is zero the covariance between the q_{i0} and q_{i1} terms must also be zero, resulting in a point mass at zero (χ^2_0) not a χ^2_1. 50:50 mixtures of χ^2_0 and χ^2_{df} (where $df > 0$) are simple to evaluate; to obtain an appropriate p-value one simply halves the p-value obtained from a χ^2_{df} distribution. In practice, the first 2logLR test statistic (**StatDevRe1**) is easier to compute in ASREML. The agreement between these asymptotic results and 1000 simulation replicates was assessed graphically.

Alternative model To assess statistical power when the null hypothesis was false, the genetic variance was altered with increasing age (method 2). The genetic variance attributable to the QTL was simulated to increase linearly from 0.20 at age 1 to either 0.33 at age 5 (case 1) or 0.4 at age 5 (case 2). A further situation was considered in which the ratio of permanent to temporary environmental error variance was altered (case 3): instead of the 50:50 allocation in cases 1 and 2, more of the error variance was allocated to be common to every measurement taken on each individual. In this case the ratio of permanent environmental variance to temporary environmental variance was simulated to be 75:25. The genetic variances in case 3 were the same as those in case 1. 200 replicates were generated in each case. The utility of the two tests for deviations from the repeatability model (**StatDevRe1** and **StatDevRe2**, described above) was assessed by counting the proportion of replicates rejecting the repeatability model in favour of the linear RR model.

Power to detect QTL using RR model

The data from cases 1 to 3 was used to assess the power to detect the simulated QTL. This power was evaluated using 3 analysis methods. Firstly, the repeatability model was fitted to the data. The 2logLR test statistic for the test of (repeatability) QTL versus no QTL (call this **Statdet3**, statistic for **det**ecting QTL effect) is assumed to be distributed as $\frac{1}{2}\chi^2_1 : \frac{1}{2}0$. Secondly, the univariate 2logLR statistic was calculated; the maximum statistic from the 5 single analyses and an analysis of the mean of the 5 trait values was obtained. Ignoring the multiple testing issue, this statistic (call this **Statdet4**) is assumed to be $\frac{1}{2}\chi^2_1 : \frac{1}{2}0$. Bonferroni correction for 6 tests can be applied by reducing the significance level six-fold (call this **Statdet4Bonferroni**). Whilst some correction for multiple testing is in order, the Bonferroni correction is too conservative in this case because the 6 tests are correlated.

Table 3.1: Summary of Statistics, Simulation 1

	QTL RR coefficients in			
Statistic	L_1	L_1	Asymptotic distn. of $2\ln(L_1/L_0)$	Notes
StatDevRe1	q_{i0}	q_{i0}, q_{i1}	$\frac{1}{2}\chi^2_1 : \frac{1}{2}0$	Deviation from Re model test
StatDevRe2	q_{i0}	$q_{i0}, q_{i1}, cov(q_{i0}, q_{i1})$	$\frac{1}{2}\chi^2_2 : \frac{1}{2}0$	Deviation from Re model test
Statdet3	none	q_{i0}	$\frac{1}{2}\chi^2_1 : \frac{1}{2}0$	Power to detect QTL test
Statdet4	n/a	n/a	$\frac{1}{2}\chi^2_1 : \frac{1}{2}0$	Power to detect QTL *univariate* test
Statdet5	none	$q_{i0}, q_{i1}, cov(q_{i0}, q_{i1})$	$\frac{1}{4}\chi^2_3 : \frac{1}{2}\chi^2_1 : \frac{1}{4}0$	Power to detect QTL test

The true power at the given significance level is likely to lie somewhere between the power for **Statdet4** and **Statdet4Bonferroni**. Thirdly, the linear RR was fitted to the data. The 2logLR statistic for the test of QTL (with constant and slope terms) versus no QTL (all 3 (co)variances set to zero; call this **Statdet5**) is assumed to be $\frac{1}{4}\chi^2_3 : \frac{1}{2}\chi^2_1 : \frac{1}{4}0$. This statistic has this mixture distribution because there are two variance terms and these are on the boundary of the parameter space under the null. When performing the likelihood ratio test, one quarter of the time both of the variances are estimated to be positive (and their covariance can be non-zero), one half of the time one of the variances is at zero (together with the covariance, $cov(q_{i0}, q_{i1})$, from equation 2.25) and one quarter of the time all 3 (co)variances are at zero.

Note that in the RR model we compare the QTL model with a model in which there is a polygenic variance term included but, in this study, the polygenic term does not vary with age. In practice, the test for a QTL under the RR model would allow the polygenic term to vary with age (via some polynomial of age). However, to allow a direct comparison between the RR and Re models a constant polygenic term is fitted here.

A further test of significance of the QTL effect could be obtained by fitting a full multivariate model (i.e. 15 (co)variances) or higher order polynomial RRs to the data and comparing this with the univariate, Re and linear RR models. However, fitting these models to the data simulated in Simulation 1 (QTL genetic correlations equal to 1) proved impossible in practice. The estimation of large numbers of parameters is very difficult when the traits of interest are highly correlated. Estimation was more readily achieved in Simulations 2a and 2b where the correlation between the traits was reduced.

The power of **Statdet3**, **Statdet4** and **Statdet5** to detect the simulated QTL was assessed at 3 significance levels: 0.001, 0.0001 (asymptotically equivalent to a univariate base 10 logarithm of odds, or LOD, of 3) and 0.00001.

For reference, the statistics calculated are given in table 3.1.

3.2.4 More complex model with sloping covariance function (Simulation 2a)

A more realistic model of the change in genetic (QTL) effect over time is one which allows the correlation between genetic effects to be below one, particularly for measures

Figure 3.1: Flat CF

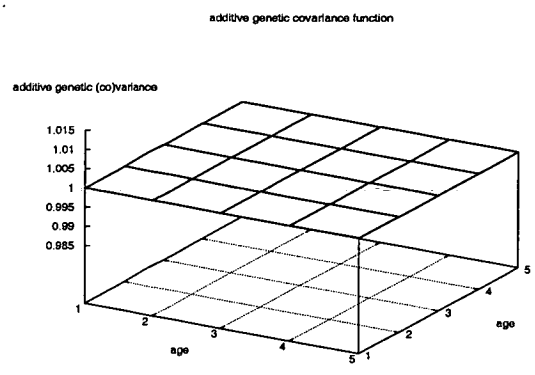
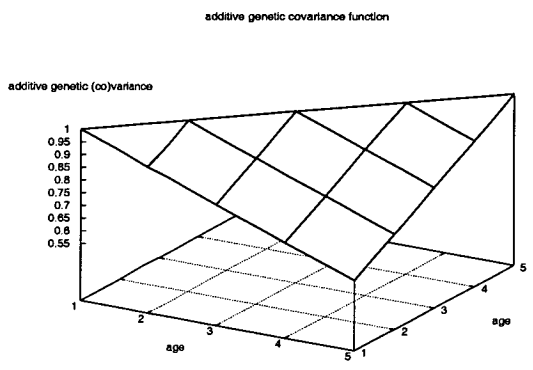


Figure 3.2: Sloping CF



widely separated in time. Graphically, the correlation matrix with all ones results in a genetic (QTL) covariance function (CF) similar to that in figure 3.1. A more realistic situation may be similar to the one shown in figure 3.2 (half of the off-diagonals have been suppressed to make the diagram clearer) where the correlation matrix $\mathbf{K}^*$ is

$$\begin{pmatrix} 1 & 0.9 & 0.8 & 0.7 & 0.6 \\ 0.9 & 1 & 0.9 & 0.8 & 0.7 \\ 0.8 & 0.9 & 1 & 0.9 & 0.8 \\ 0.7 & 0.8 & 0.9 & 1 & 0.9 \\ 0.6 & 0.7 & 0.8 & 0.9 & 1 \end{pmatrix}.$$

In cases such as this the assumption that the genetic correlation is 1 is violated. This means that any model which allows the correlations to be less than one (such as a first or higher order RR) will give a better fit than the repeatability model, even when the genetic variance does not change over time.

Apart from the change in the correlation structure the simulation set up was the same as in Simulation 1. The data are simulated from the MVN distribution (method 3 from

section 3.2.2) with the correlation matrix for the genetic effects specified to be $K*$ (matrix above) with the QTL variance rising linearly from 0.2 to 0.4. 200 replicates were generated.

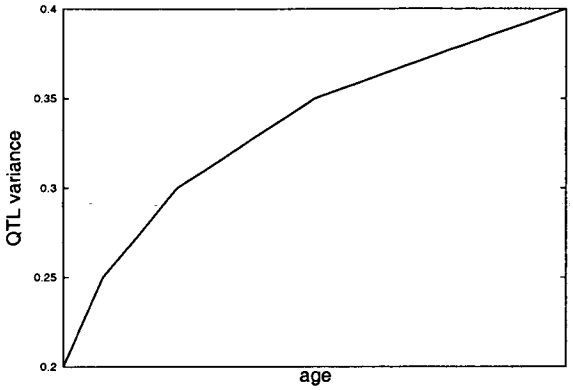
The models fitted to the simulation data were a no QTL (polygenic component only) model, a Re model (equivalent to an order 0 RR model), a series of RR models, with orders 1 to 4 and a full multivariate model. This full multivariate model fits 5 variances for the 5 different ages in the data and attempts to estimate separately all 10 covariances between the effects at different ages. This model should give identical likelihoods to the saturated fourth order RR model. Both fit the same number of parameters for the QTL effect (15 in all). The lower order RRs use polynomials to smooth the covariance function, reducing the number of parameters requiring estimation. Note that the simulated covariance function was not generated from a polynomial. Although the true shape of the CF will not be known in practice, it is highly unlikely to look exactly like that generated from a polynomial.

Comparisons between the fitted models The models listed above are nested and are compared using likelihood ratio tests. Some of the comparisons are the same as those for Simulation 1. The test of Re vs. no QTL is **Statdet3** above. The two tests for the significance of the linear RR are **StatDevRe1** and **StatDevRe2**. Tests for the significance of the higher order polynomial RRs are analogous to **StatDevRe1** and **StatDevRe2**. For the test of the full (i.e. all elements of the CF estimated) order k^*+1 RR versus the order k^* model, twice the logLR was compared with a $\frac{1}{2}\chi^2_{k^*+1} : \frac{1}{2}0$ distribution (rationale for this is given in section 3.2.3, analogous to **StatDevRe2**). Call this **Statkfull**. An alternative test uses the order k^*+1 RR with the correlations between the k^*+1^{th} diagonal term of the CF and the first k^* RR coefficients constrained to zero (analogous to **StatDevRe1** in section 3.2.3, with the correlations between the first k^* coefficients left unconstrained). The logLR statistic comparing this constrained fit to the order k^* RR has a $\frac{1}{2}\chi^2_1 : \frac{1}{2}0$ distribution. Call this **Statkconstrained**. Note that as in section 2.2.2, k^* is the order of the polynomial and k is used for the number of terms in the polynomial; i.e. $k* = k - 1$. As before, the coefficient matrix as a whole was constrained to be positive definite.

The best fitting model was selected by increasing the order of the RR until the additional terms were found to not significantly increase the likelihood. The higher order RR was deemed significantly better than the lower order RR if the p value for the higher order model was below 0.01; for example an order 1 RR was rejected in favour of an order 2 RR if the twice the difference in likelihood exceed 9.84, the 1% level of a $\frac{1}{2}\chi^2_3 : \frac{1}{2}0$ distribution. Once adding an additional term to the RR was found to be non-significant, higher order RRs were not considered.

The RR fitting procedure models the random deviations from a fixed curve for each regression coefficient. In the fourth order RR case fourth order polynomials of age are hence required as a fixed effect. To ensure valid LR tests comparing different polynomial orders for the RRs this same set of fixed effects (i.e., $f_0 + f_1 t_j + f_2 t_j^2 + f_3 t_j^3 + f_4 t_j^4$) were used for all fitted models. If the fixed effects are changed with the order of the RR, the LR test

Figure 3.3: Simulation 2b: Simulated increase in QTL variance with age



is not valid.

3.2.5 Simulation 2b

Simulation 2a was re-run with a non-linear change in the genetic variance over time. To achieve this, the simulated ages were re-scaled so that the increase in genetic (QTL) variance became logarithmic with age. Instead of instructing ASREML to compute orthogonal polynomials based on measures at ages 1 (genetic variance simulated to be 0.2), 2 (genetic variance simulated to be 0.25),..,5 (genetic variance simulated to be 0.4), the ages were specified as being 1 (genetic variance 0.2), 2 (genetic variance 0.25) , 5 (genetic variance 0.3), 11 (genetic variance 0.35) and 21 (genetic variance 0.4). Note that this means that the trait measures are no longer evenly spaced. The simulated increase in QTL variance with age after this re-scaling is shown in figure 3.3. An alternative (not used here) to this re-scaling would be to generate 21 ages initially (with the variance starting at 0.2 at age 1 and rising to 0.4 at age 21) and pick out ages 1, 11, 17, 20 and 21; labelling these 1 to 5 and using these in ASREML would give a curve the same shape as in figure 3.3. Making the increase in variance logarithmic should make it more difficult for the polynomials to model the changes in the CF over time.

The other parameters, models fitted and tests used in simulation 2b were the same as those in simulation 2a.

3.3 Results

3.3.1 IBD results

Selection of families for modification

As predicted, approximately of 45% of the generated 2 sib families and approximately of 90% of the generated 2 sib families had singular IBD matrices.

Table 3.2: IBD Modification: 4 Sib Families

replicate	SOLAR LOD	ASREML LOD	
		Modify all diags	Modify only singular diags
1	0.912	0.910	0.910
2	4.428	4.443	4.443
3	2.387	2.385	2.386
4	2.065	2.060	2.060
5	2.290	2.293	2.294
6	3.299	3.311	3.310
7	2.843	2.855	2.856
8	2.245	2.241	2.242
9	1.101	1.108	1.108
10	3.644	3.646	3.645

Table 3.3: IBD Modification: 2 Sib Families

replicate	SOLAR LOD	ASREML LOD	
		Modify all diags	Modify only singular diags
1	0.003	0.003	0.003
2	0.834	0.840	0.841
3	0.569	0.573	0.574
4	0.784	0.787	0.786
5	0.218	0.219	0.219
6	0.527	0.526	0.524
7	0.875	0.881	0.880
8	0.009	0.009	0.008
9	0.240	0.237	0.238
10	1.964	1.970	1.970

The results from simulations in which the families' IBD matrices were either all modified to make them non-singular or only modified if they were actually non-singular are presented in tables 3.2 and 3.3. Table 3.2 has the results for the case in which most families required modification. Table 3.3 has the results for the case in which a smaller proportion required modification. These results show that it makes very little difference which method is used; one can simply add a small amount to all diagonal IBD matrix entries without biasing the results.

Effects of adding different values to diagonal entries of singular sub-matrices.

Table 3.4 indicates the absolute difference between the h^2/LOD score estimates obtained with ASREML and the modified sub-matrix IBD values and those obtained with SOLAR. The difference between the LOD scores and the QTL specific heritabilities are given. Over the 10 replicates the average estimated QTL specific $h^2$was 0.24 (simulated to be 0.25). Note there are small differences between the results as a consequence of minor differences between the maximisation algorithms used in SOLAR and ASREML (ASREML corrects for the bias introduced by fitting fixed effects but SOLAR does not; here the only fixed

Table 3.4: Difference between ASREML based h^2/LOD and SOLAR h^2/LOD.

	added value									
	1	1	0.1	0.1	0.01	0.01	10^{-3}	10^{-3}	10^{-4}	10^{-4}
rep	h^2	LOD	h^2	LOD	h^2	LOD	h^2	LOD	h^2	LOD
1	0.0154	0.8460	0.0100	0.0160	0.0041	0.0118	0.0035	0.0140	0.0035	0.0140
2	0.1254	1.914	0.0014	0.1629	0.0032	0.0084	0.0032	0.0249	0.0034	0.0267
3	0.0714	0.1118	0.0123	0.0505	0.0043	0.0175	0.0035	0.0140	0.0036	0.0140
4	0.0040	0.3672	0.0056	0.0107	0.0031	0.0080	0.0029	0.0102	0.0029	0.0089
5	0.0048	0.3822	0.0078	0.0296	0.0038	0.0357	0.0034	0.0361	0.0034	0.0361
6	0.1516	1.2043	0.0025	0.0956	0.0002	0.0016	0.0002	0.0070	0.0003	0.0062
7	0.0499	0.1248	0.0057	0.0100	0.0000	0.0008	0.0007	0.0021	0.0007	0.0026
8	0.0996	0.3925	0.0045	0.0334	0.0005	0.0025	0.0003	0.0001	0.0001	0.0009
9	0.0269	0.0143	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.0378	0.7827	0.0221	0.0764	0.0046	0.0208	0.0029	0.0142	0.0028	0.0142

effect is the mean so the difference is negligible). There is little difference between the results obtained provided the added value is in the range (0.01, 0.0001). Adding 10^{-5} to the diagonals resulted in singularity (to machine precision) in all cases.

Since the method of adding small values to the diagonal entries of the non-singular matrices proved perfectly adequate the more complicated procedures based on (matrix) bending were deemed unnecessary.

3.3.2 Simulation 1

Repeatability null model The agreement between the expected asymptotic and simulation based empirical distributions when fitting the RR model to data simulated to fit the repeatability model (no change in variance over time, correlation between effects at different ages equal to one) was excellent. The two statistics of interest, **StatDevRe1** and **StatDevRe2** are expected to follow $\frac{1}{2}\chi^2_1 : \frac{1}{2}0$ and $\frac{1}{2}\chi^2_2 : \frac{1}{2}0$ distributions, respectively. They are shown in figure 3.4. For comparison the $\frac{1}{2}\chi^2_2 : \frac{1}{2}\chi^2_1$ distribution is shown; this shows that neither **StatDevRe1** or **StatDevRe2** converge to this mixture (as suggested in [221, 156]). Note that although the covariance is not constrained to 0 in **StatDevRe2** the overall coefficient matrix (C) is constrained to be positive definite.

Deviations from repeatability model. Three cases were considered. In the first case the genetic (QTL) variance increased moderately (0.2 to 0.33); in the second the increase was larger (0.2 to 0.4). The third case was the same as the first but with the environmental component altered. The power in each case is given in table 3.5.

The results indicate that **StatDevRe2** is more powerful at detecting deviations from the repeatability model. One should note however that in some circumstances **StatDevRe1** is easier to compute than **StatDevRe2**. Reducing the relative amount of temporary environment (ratio of permanent to environmental variance 75:25 instead of 50:50) in case 3 results in the change in genetic variance over time being easier to detect (more power to

Figure 3.4: **StatDevRe1** and **StatDevRe2** Simulation Results: Fit to asymptotic null distributions

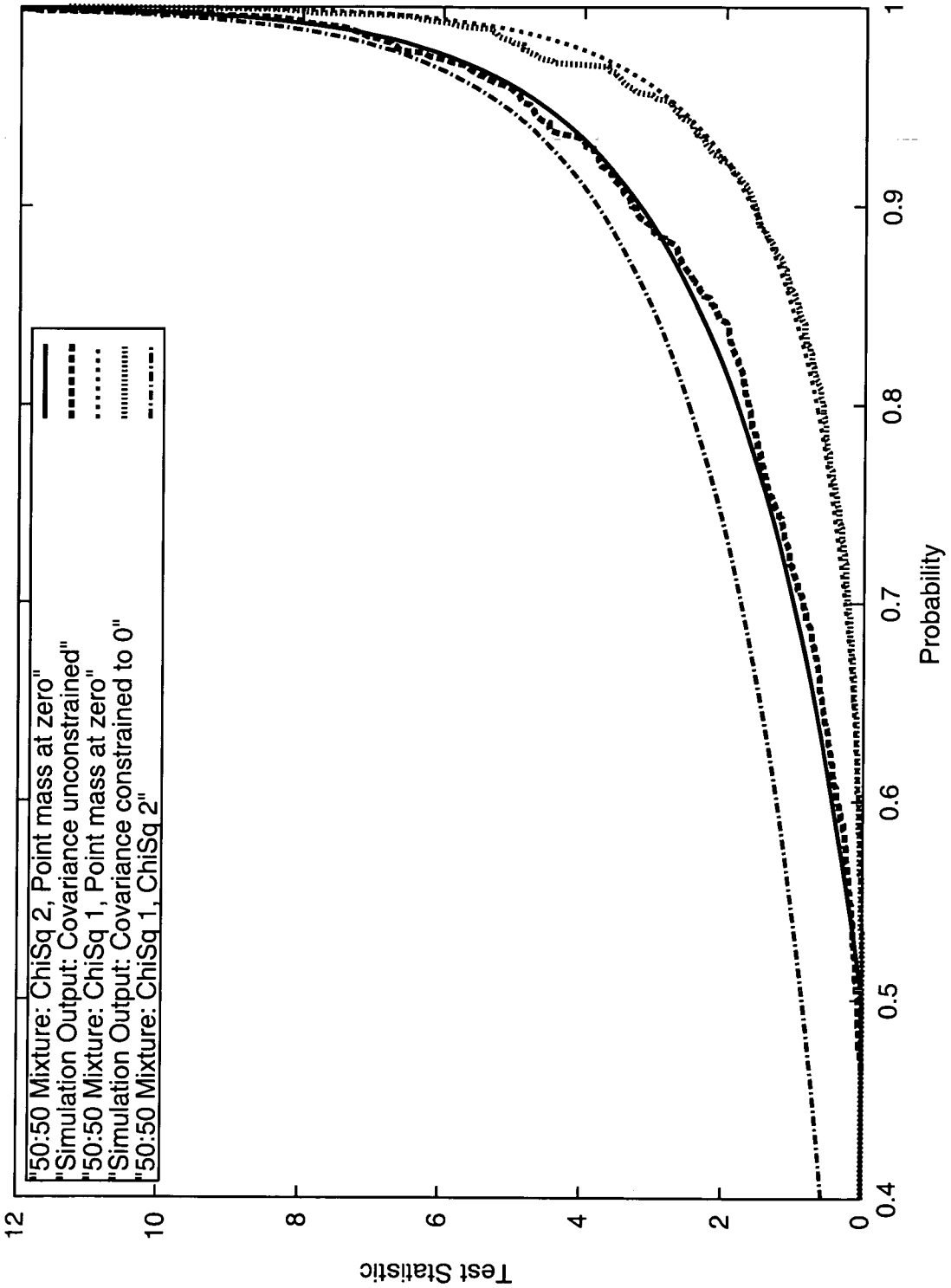


Table 3.5: Simulation 1: Power (at 1% level) to reject the repeatability model

	StatDevRe1	**StatDevRe2**
Case 1	5%	41%
Case 2	12%	76%
Case 3*	16%	75%

*Same as case 1 but with change in ratio of environmental effects

Table 3.6: Simulation 1, Case 1. QTL variance 0.2 (age 1) to 0.33 (age 5)

Statistic\Significance level	10^{-3}	10^{-4}	10^{-5}
Statdet3 (Re QTL vs. no QTL)	54	30	17
Statdet4 (Univariate QTL vs. no QTL)	61	33	18
Statdet4Bonferroni (Corrected Stat4)	38	21	9
Statdet5 (Linear RR QTL vs. no QTL)	64	43	31

detect deviations from repeatability in case 3 compared with case 1).

Power to detect QTL: RR, Re and univariate models The power to detect a simulated QTL was determined using three statistics, **Statdet3**, **Statdet4** and **Statdet5**. The power (proportion of 200 replicates, expressed as a percentage) at different significance levels for case 1 (QTL variance 0.2 to 0.33) is given in table 3.6.

The power (%) for the case 2 (QTL variance 0.2 to 0.4) simulation is given in table 3.7.

The power (%) for the case 3 (QTL variance 0.2 to 0.33, permanent environment variance 0.75, temporary environment variance 0.25) simulation is given in table 3.8.

Looking at the results from **Statdet3** in tables 3.6 and 3.8 we see that much of the power in the repeatability analysis lies in the reduction in temporary environmental noise as a result of averaging over a number of measures; when the temporary environmental effects are small (as in case 3) the repeatability analysis has little power to detect QTL. By contrast, the model allowing for a change in QTL effect over time (**Statdet5** linear RR) gains power when the temporary environmental noise is reduced. This is because the change in genetic variance over time can be more readily detected, increasing the power to detect the QTL when a parameter modelling the change in QTL effect over time is fitted.

Table 3.7: Simulation 1, Case 2. QTL variance 0.2 (age 1) to 0.4 (age 5)

Statistic\Significance level	10^{-3}	10^{-4}	10^{-5}
Statdet3 (Re QTL vs. no QTL)	67	41	21
Statdet4 (Univariate QTL vs. no QTL)	75	47	24
Statdet4Bonferroni (Corrected Stat4)	53	28	13
Statdet5 (Linear RR QTL vs. no QTL)	85	78	65

Table 3.8: Simulation 1, Case 3. QTL variance 0.2 (age 1) to 0.33 (age 5) (0.75 perm, 0.25 temp)

Statistic\Significance level	10^{-3}	10^{-4}	10^{-5}
Statdet3 (Re QTL vs. no QTL)	30	13	5
Statdet4 (Univariate QTL vs. no QTL)	39	18	6
Statdet4Bonferroni (Corrected Stat4)	24	8	4
Statdet5 (Linear RR QTL vs. no QTL)	75	64	46

Note also that a modest increase in the genetic variance at age 5 (from 0.33 in case 1 to 0.4 in case 2) has a relatively large effect upon the power when **Statdet5** is used; the power to detect a LOD of 3 (significance level 10^{-4}) rises from 43% to 78%.

As mentioned earlier, the uncorrected **Statdet4** overestimates the power whilst the **Statdet4Bonferroni** is too conservative. Assuming the true power value at the specified significance levels can be obtained by taking a power estimate between **Statdet4** and **Statdet4Bonferroni** we see that the repeatability and univariate methods have similar power.

3.3.3 Simulation 2a

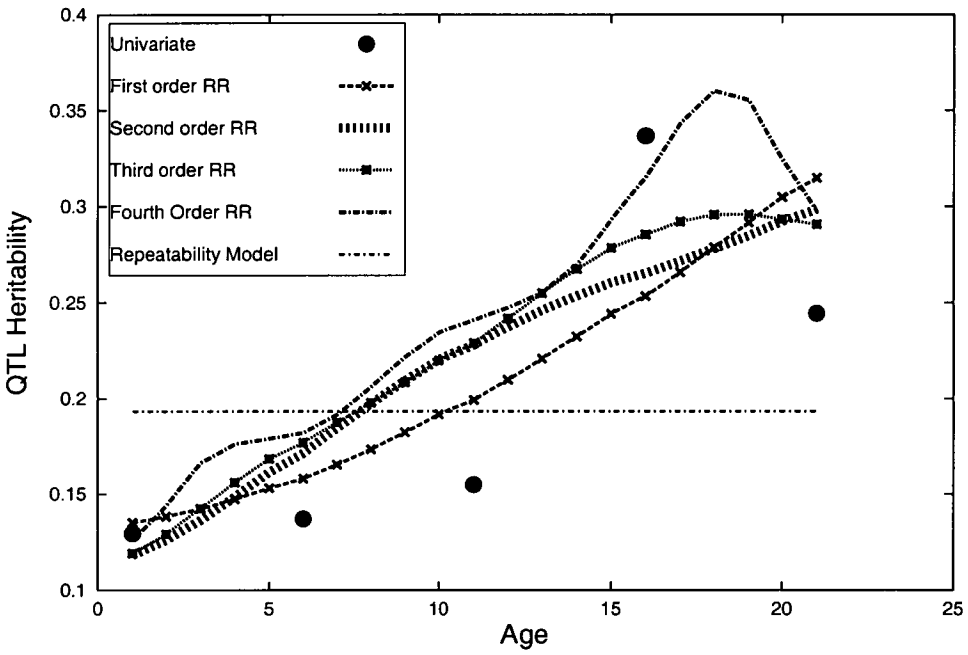
The procedure outlined above was used to determine the best fitting model to the data. 79% of replicates rejected, at the 1% significance level, the no QTL model when the Re model was fitted. However, in all cases (200 replicates) the Re model was rejected in favour of the first order RR model (All p values less than 10^{-5} for **StatDevRe1** and **StatDevRe2**). This was unsurprising since the data were simulated so that the QTL variance changed over time and the genetic (QTL) correlations were <1. 64% of replicates rejected the linear RR in favour of the quadratic RR when **Statkfull** was used to compare the two models. When **Statkconstrained** was used only 23% of replicates provided evidence for the quadratic model. Using **Statkconstrained** for the test for a cubic RR compared with the quadratic fit (for replicates where the quadratic coefficient was significant) resulted in none of the replicates indicating the cubic fit was better. Assessing the further models (unconstrained cubic model and quartic model) proved difficult computationally, with many replicates failing to converge to a likelihood maximum. In the cubic case, roughly one-third of replicates failed to converge when the unconstrained cubic model (i.e. **Statkfull** was calculated) was fitted. Taking the likelihoods as calculated (i.e. one-third of them are underestimates of the true likelihood maximum, biasing the test statistic for the significance of the cubic term down-wards), 7% of replicates rejected the quadratic model in favour of the cubic model. Only in 35% of cases could the full multivariate model be maximised (the number of iterations for the likelihood maximisation was set to 100). Although the fourth order RR should maximise to the same likelihood as the full multivariate case, the RR converges less often than the full multivariate case. Hence, in practice it is some-

Table 3.9: Simulation 2a (QTL variance 0.2 (age 1) to 0.4 (age 5)): Best fitting model (%)

Model	**Statkfull**	**Statkconstrained**
Repeatability	0	0
Linear RR	36	77
Quadratic RR	57	23
Cubic RR	7*	0

*One third of replicates failed to converge so this may be an underestimate

Figure 3.5: Sample results, Simulation 2a



times possible to test for the significance of the fourth order terms but not calculate the coefficients of the 5 by 5 CF. These results are summarised in table 3.9.

For a few of the replicates all models could be maximised, a graphic representation of the results of one replicate is given in figure 3.5. The variance terms from the QTL RR are expressed as a proportion of the total variance (i.e. QTL heritability). For comparison, the univariate and repeatability model results are superimposed on the same graph. This shows that the repeatability model is a poor fit to the simulated model and that the univariate results, while following the simulated model to some degree, are rather noisy. All of the polynomial based RRs follow the simulated model well; the first order model offers an excellent fit with only two extra parameters fitted compared with the repeatability model. The fourth order polynomial follows the univariate results more closely but in this case such variations from the simulated model are simply random variation; the lower order polynomials provide a better fit to the true (simulated) model.

Table 3.10: Simulation 2b (QTL variance 0.2 (age 1) to 0.4 (age 5)): Best fitting model (%)

Model	**Statkfull**	**Statkconstrained**
Repeatability	0	0
Linear RR	16	11
Quadratic RR	67	85
Cubic RR	17*	4**

*Almost one third of replicates failed to converge so this may be an underestimate
** Five percent of replicates failed to converge so this may be a slight underestimate

3.3.4 Simulation 2b

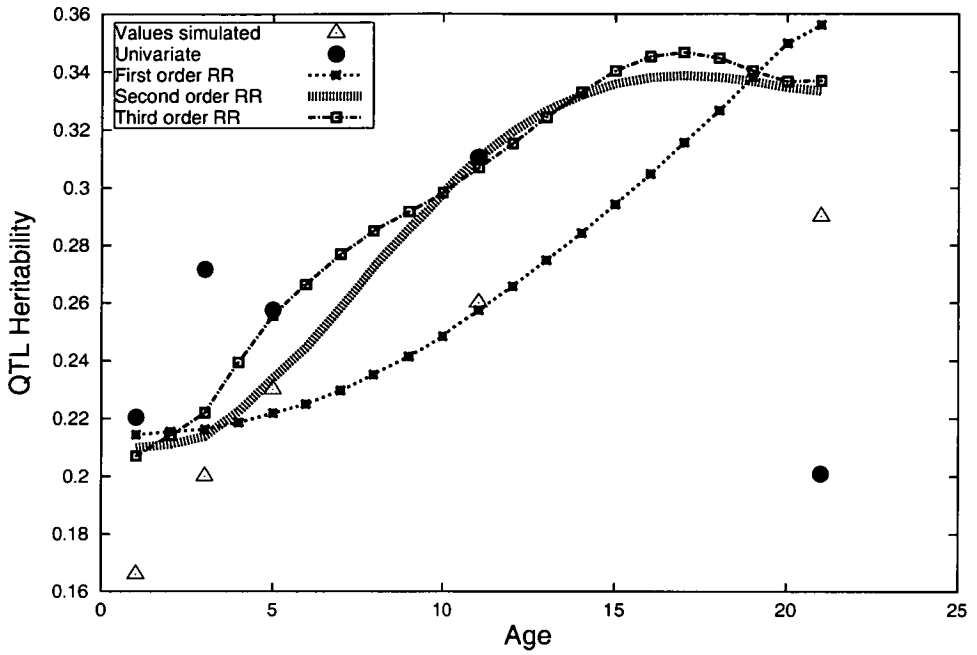
Simulation 2b used the same generating model and analysis methods as simulation 2a but the age scaling was changed so that the increase in QTL variance with age followed the curve in figure 3.3. 71% of replicates rejected (significance level 1%) the no QTL model when the Re model was fitted. In all cases (200 replicates) the Re model was rejected in favour of the first order RR model (all p values less than 10^{-6} for **StatDevRe1** and **StatDevRe2**). 84% of replicates rejected the linear RR in favour of the quadratic RR when Statkfull was used to compare the two models. When the **Statkconstrained** was used 89% of replicates provided evidence for the quadratic model (Note that although the likelihood ratio of **Statkconstrained** is lower than that of **Statkfull**, since the null distributions differ **Statkconstrained** can sometimes give smaller p-values). Using **Statkconstrained** for the test for a cubic RR compared with the quadratic fit resulted in 4% of the replicates indicating the cubic fit was better. This may be a slight underestimate as 5% of the replicates failed to converge to a likelihood maximum. Using the unconstrained cubic model in the test resulted in 17% of replicates rejecting the quadratic model although almost a third failed to converge fully. The quartic and full multivariate models could not be reliably fitted to these data. These results are summarised in table 3.10.

The results of one replicate are given in figure 3.6. The simulated values are superimposed on the graph. As expected, when the change in QTL variance is non-linear the second and higher order RRs have more utility than the first order model. Nonetheless, even the first order RR is substantially better than the repeatability model. Once again the univariate results are rather noisy; univariate methods do not utilise the natural ordering in time of the genetic effects with adjacent measures often yielding very different estimates of QTL heritability.

3.4 Discussion

The results presented show that in a variety of realistic scenarios simple multivariate analyses such as repeatability (Re) analysis (equivalent to an 'average across all measures' univariate analysis when there is regular age spacing) are substantially less powerful than more complex multivariate techniques such as random regression based covariance

Figure 3.6: Sample results, Simulation 2b



function methods (RR). Repeatability models are only useful for traits in which the genetic variance does not change over time and the genetic correlation between repeated trait measures is close to one.

In simulation 1 it was shown that when there was a moderate increase in QTL effect over time fitting a first order RR increased the power to detect the QTL. This increase in power came solely from the RR modelling the change in QTL variance. The increased efficiency of the RR in modelling any decreases in the genetic correlation between trait measures below 1 was ignored by simulating data with no decline in genetic correlation with time. The increase in power was particularly large when the ratio of permanent to temporary environment was high (i.e. when most of the environmental 'noise' affects all of an individuals trait measures).

At the GAW13 meeting [5] the genes that changed in their effect (variance) over time were often referred to as 'slope' genes [77, 181]. Simulation 1 allows a direct test for these slope genes. However, most QTL effects will not be completely correlated across ages and a more realistic simulation model will allow the correlations between QTL effects at different ages to decrease.

In the case where the genetic correlation over time is not 1 all of the RR models (for all polynomial orders > 0) offer substantially more power than the repeatability model. For example, in simulation 1, case 2, **StatDevRel** rejected the repeatability model in 12% of cases (significance level 1%). By comparison, when the data were simulated in simulation 2a with the same parameters apart from a change in the correlation structure, 100% of replicates rejected the repeatability model (significance level 1%, although in fact

all rejected at significance level 0.001%). The univariate results for simulation 2a were similar (data not shown) to those obtained for the repeatability model and were hence substantially less powerful than those obtained from the RR model.

Simulation 2a showed that when the increase in QTL variance was linear the best fitting model was either a linear or quadratic RR (best model quadratic in 23% [constrained case] or 64% [unconstrained] of cases). When the increase was non-linear (figure 3.3, simulation 2b) the quadratic RR was usually the best fit (in ~85% of simulation replicates). Although the simulation 2b showed that polynomial based CFs worked well with the simulated logarithmic increase in QTL variance with age (as shown in figure 3.3), non-monotonic changes in QTL variance with age were not considered here (e.g. an increase in genetic effect at earlier ages, followed by a decline in later life).

It is not possible to know what form real life genetic CFs will take. It was assumed in simulation 2a that the decline in correlation followed a steady decrease with increasing time separation. The correlation was assumed to remain relatively high over the range of ages of interest. This seems likely to be true for QTL effects (whose constituent element is one or more close linked genes) but may be less likely to hold for polygenic effects (whose constituent elements are more heterogeneous and will change over life). The shape of possible CFs for polygenic effects was considered in [108]. The models considered in [108] range from one in which the correlation structure remains high across ages to another in which the correlation becomes negative at widely separated ages. They conclude that RR models are effective for CFs whose correlation remain high across ages but are less effective for CFs with rapidly decreasing correlations [108].

The model selection in simulation 2a/2b was based on differences in likelihood but this may not be the most effective strategy. Exploratory simulation work on larger data sets indicated that even when the data are simulated under a relatively simple model (linear change in genetic variance, correlation structure as in simulation 2a), some of the higher order RRs give the most significant likelihood ratio statistics. In these cases the fitted RR models exploit the stochastic variation present in individual simulation replicates; in reality the true covariance functions are unlikely to follow 'wiggly' high order polynomials and simpler polynomials should be chosen instead. In any event, with most realistic sample sizes high order polynomials are impossible to estimate and this problem will be of little practical consequence. Jaffrezic et al. [110] encountered similar problems when using likelihood as the criterion for model selection in large data sets. Further work on model selection may be useful for studies of particularly large data sets.

Although fitting a model which estimates the full set of (co)variances in the data (there are $w(w+1)/2$ to estimate when there are w trait measures) can capture the change in QTL variance over time, such methods are inefficient in most cases and are difficult to apply in practice. One of the primary aims of this paper was to investigate how much information can be extracted from longitudinal data in realistic scenarios. The work here and other work on human data sets (Chapter 4, [50, 51]) indicate that approaches which do not simplify the covariance structure are unworkable in practice (the relatively small

data sets do not support the estimation of large numbers of parameters). Although some of the data sets simulated here supported the estimation of the full multivariate model parameters, the simulated data did not include age varying polygenic or permanent environmental effects. In addition, all individuals had a full set of phenotypes and genotypes. In practice these complications will make estimating large numbers of parameters more difficult. These simulations also ignore one of the benefits of the RR procedure (compared with longitudinal analyses which do not incorporate age), namely the ability of the RR method to analyse data with phenotypes measured across a wide range of ages. In these simulations all individuals were assumed to be measured at all five ages. In reality human data sets will often feature individuals measured at a variety of different ages; a full multivariate analysis will usually require individuals at proximal ages to be grouped together, discarding information. In chapter 4 the RR method is used to allow an analysis of 76 distinct ages in a single analysis of a real data set.

One disadvantage of the RR techniques is that the method depends on the shape of the full CF (i.e. the off-diagonals as well as the diagonals). A low order polynomial may be adequate to model the change in variance over time but inadequate for approximating the covariance structure or vice-versa. Alternative models which fit separate functions for the change in variance and the change in correlation or covariance have been proposed (Character process models, [173]). The utility of such models in characterising longitudinal QTL is an area which requires further study.

In summary, covariance function techniques have been shown to provide considerably more power for QTL detection than univariate and repeatability techniques. It should be possible to take advantage of this extra power by fitting first order random regressions to most realistic human/natural population data sets. Larger data sets (e.g. animal breeding applications) that support the estimation of higher order polynomials will allow better characterisation of the change in genetic effect over time.

3.5 Appendix

3.5.1 Expected number of singular IBD matrices

For 2 sib nuclear families, half of the time the offspring will share 0 or 2 alleles IBD and half of the time they will share 1 allele IBD. This gives a singular IBD matrix in half of all such families (assuming both parents are heterozygous and there are $\geq$3 alleles, i.e. fully informative).

Proof that for $n \geq 3$, fully informative nuclear families with n siblings will have singular marker-specific IBD matrices A few points to begin with. First of all, since the ordering of the sibs within the sibship is arbitrary, assume that if there is a sib who shares two alleles with any other sib, these two sibs are written as columns 3 and 4 of the IBD matrix; subsequent to this come individuals sharing one allele with the first sib fol-

lowed by individuals sharing no alleles with the first sib. Secondly, if any column (or row) of a (IBD) matrix can be expressed as a linear combination of one or more other columns, the matrix is singular. Thirdly, although, in a sibship with fully informative parents, the entries of the IBD matrix can take values 0, 0.5 or 1, only certain combinations are compatible with the transmission of alleles from two parents. In the case of a 3 sib family, there are 9 possible arrangements of IBD coefficients in the matrix. The IBD matrix is of the following form

$$\mathbf{R}_3 = \begin{pmatrix} 1 & 0 & 0.5 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & x & y \\ 0.5 & 0.5 & x & 1 & z \\ 0.5 & 0.5 & y & z & 1 \end{pmatrix} \tag{3.1}$$

where the values in the x, y and z positions are as follows. Only rows 1 to 4 are possible

	x	y	z
1	**1**	**1**	**1**
2	**1**	**0.5**	**0.5**
3	**1**	**0**	**0**
4	**0.5**	**0.5**	**0**
5	0	0	0
6	0.5	0.5	0.5
7	0	0	0.5
8	0.5	1	1
9	0	0.5	1

Table 3.11: Possible IBD values

given fully informative parents. The possible values are in bold text in table 3.11, impossible values are in non-bold text. Note that since the ordering of the sibs within the sibship is arbitrary the x value can always be made to be one in rows 1, 2 and 3 of table 3.11 and made to be 0.5 in row 4 of the table. In terms of alleles, row 1 corresponds to the case where the parents are AB and CD and the offspring are AC AC AC. Similarly, row 2 is AB CD (parents), AC AC AD (offspring). Row 3 is AB CD (parents), AC AC BD (offspring) and row 4 is AB CD (parents), AC AD BC (offspring).

Now consider the values in the four possible rows of the table.

Rows 1, 2 and 3 For rows 1, 2 and 3 of the table, y must equal z; that is, the third sibling must share the same number of alleles IBD with sib 1 (column 3 of $\mathbf{R}_3$) and sib 2 (column 4 of $\mathbf{R}_3$). Since column 3 of $\mathbf{R}_3$ is then equal to column 4 of $\mathbf{R}_3$, the IBD matrix is singular. Furthermore, any further sibs in the sibship (fourth sib, fifth sib, ...) will have the same relationship to sibs 1 and 2 (since sibs one and two share 2 alleles IBD). This means columns 3 and 4 of the matrix will continue to be equal.

Row 4 For row 4 of the table, x was assumed to be 0.5. Since the allocation of 0.5 and 0 to y and z is also based upon the arbitrary ordering of the sibs (the allocation of the 0 to either x, y or z is dependent upon whether the first and second, first and third, or second and third sibs share 0 alleles IBD, respectively) the only case to consider is where y and z are 0.5 and 0. In this case column 4 plus column 5 of $\mathbf{R}_3$ equals column 1 plus column 2 of $\mathbf{R}_3$. This proves singularity in the 3 sib case.

If there are 4 sibs in the sibship, the only case left to consider is the cases where the fourth sib is added to a 3 sib family with the row 4 of table 3.11 values and this fourth sib does not share 2 alleles IBD with any of the previous 3 sibs (otherwise the sibship could be rearranged to make the argument in the 'Rows 1, 2 and 3' paragraph apply). In this case the IBD matrix is

$$\mathbf{R}_4 = \begin{pmatrix} 1 & 0 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 & 0 \\ 0.5 & 0.5 & 0.5 & 0 & 1 & 0.5 \\ 0.5 & 0.5 & 0 & 0.5 & 0.5 & 1 \end{pmatrix} \tag{3.2}$$

and column 4 plus column 5 equals column 3 plus column 6. For sib-ships of size 5 or greater at least 2 sibs must share 2 alleles IBD and the ordering of the sibs can be set so that columns 3 and 4 of the IBD matrix are the same (as above).

Whilst it is not true that arbitrary pedigrees will give singular IBD matrices given fully informative markers (the 2 sib nuclear family case is a simple counter-example) it seems likely, given the result for 3+ sib nuclear families, that many extended pedigrees will have singular IBD matrices. In practice, extended families tend not to be fully informative at any marker because the founders are usually untyped (deceased).

In the simulations a large number (20) of alleles were generated, resulting in the marker information being close to perfect. This meant that the expected number of singular IBD matrices would be ~90% in the 4 sib nuclear family case and ~45% for the 2 sib nuclear family case.

Chapter 4

Longitudinal Variance Components Analysis of the Framingham Data

4.1 Introduction

This chapter presents the results of analyses performed on the Framingham Heart Study (FHS) data set [142, 76]. This data set was made available as part of Genetic Analysis Workshop 13 (GAW13, [5]). The FHS was established in 1948, with the aim of increasing understanding of the causes of cardiovascular disease (CVD). The FHS has helped establish the relationship between CVD and traits such as blood pressure, obesity and blood cholesterol [115, 117]. These risk factor traits are now a major focus of preventative strategies for the reduction of CVD levels [116]. Traits such as obesity and cholesterol concentration are now known to have a substantial genetic component [144, 125].

The FHS began by recruiting 5209 individuals from the town of Framingham, Massachusetts, U.S.A., and then followed their progress at regular intervals. Individuals were measured for a multitude of traits ranging from blood pressure and cholesterol levels to lifestyle factors such as smoking and drinking rates. A second cohort of individuals was recruited in 1971. Whilst the study was designed as an epidemiological study, mainly interested in the effects of environmental factors upon disease prevalence, many of the individuals were related to each other and in the late 1980s many individuals were grouped into family sets. In the mid 1990s 330 families were typed for markers across the genome. In this chapter a variety of methods are used to interrogate the FHS data. Both univariate and multivariate variance component techniques are used, with particular emphasis on how the genetic factors affecting a number of CVD risk factors change over the life of an individual.

4.2 Data

4692 individuals were available as part of the GAW13 data set. The data were ascertained in two cohorts. The first had up to 21 trait measures for the 40 years following 1948. The second cohort had up to 5 trait measures for the 20 years following 1971. All individuals were measured at examination sessions held in 1948, 1950,...1988 (cohort 1); 1971, 1979, 1983, 1987, 1991 (cohort 2) but were different ages upon entering the study. 1702 individuals had genotype data. The vast majority of individuals in the study had all their measures when they were age 20 or older; measures at younger ages were not analysed. 2885 individuals had phenotype data. In total, 26106 phenotypic records were used in the full multivariate analysis. The traits considered were Body Mass Index (BMI, calculated as weight in kilo-grams divided by height in metres squared), height (measured in inches), fasting high density lipoprotein cholesterol (HDLC, mg/dl) and Total Cholesterol (mg/dl). Other traits available for inclusion as covariates were cigarette consumption, alcohol consumption, sex, hypertension treatment (yes/no) and cohort number.

Manipulation of Data for analysis The data were reorganised to associate a record with an age rather than an examination number. Ages ranged from 20 to 95.

For the repeatability and cross sectional univariate analyses the data were split into 6 age bands; the bandings were trait at ages 20 to 30 (age nearest 30 used), trait at ages 30 to 40 ... 70 to 80. The number of individuals with at least one record in the relevant age band are shown below. When an individual had 2 or more records in a given decade, only the latter of these measures was included.

age	20-30	30-40	40-50	50-60	60-70	70-80
number of individuals	783	1817	2263	1964	1410	879

In addition, one single larger band was considered. This band utilised a single measure on an individual between the ages of 40 and 60 (age nearest 60 used) and is denoted the '40-60' band. This band facilitated a single univariate analyses of most of the individuals (up to 2560).

The longitudinal analyses used all of the data simultaneously (i.e. 76 'bands' for ages 20 to 95). Summary measures for the four traits are given in the table below.

Trait	n	n ignoring repeated measures	mean	standard deviation
height	14929	2358	65.3	2.9
BMI	14910	2357	26.4	4.5
total cholesterol	16130	2219	218.3	42.9
HDLC	8593	1629	49.6	14.3

4.3 Methods

4.3.1 Univariate analyses

For BMI and height, sex, cohort, cigarette consumption and alcohol consumption were screened for use as a covariates. For HDLC and Total Cholesterol, BMI and an indicator variable for hypertension treatment were screened in addition to the covariates used for BMI. Since BMI had a skewed distribution, logBMI was investigated alongside BMI.

Polygenic

The traits were examined for variation across time using Residual Maximum Likelihood (REML) (ASREML program, [80]) to calculate polygenic heritabilities in the six age bands.

QTL

Univariate variance components (VC) analyses (Section 2.1) were done using SOLAR [6] and confirmed using ASREML. Multipoint IBDs were calculated every 1cM using SOLAR. The additional modifications of the IBD matrices to allow them to be used in ASREML were as described in Section 3.2.1; all of the IBD matrix diagonals had 0.001 added to make them suitable for inversion. Random effects were fitted for polygenic, QTL, family environment (household) and environmental noise terms.

4.3.2 Multivariate analyses

Repeatability Model

Repeatability analysis (Section 2.2.1) was performed on the age band data using ASREML. The model included fixed effects described above as well as a linear polynomial of age. Random effects for additive genetic, permanent environment, temporary environment and family (or household) were fitted.

For the repeatability analysis to be meaningful the trait measurements all need to be measures of what is genetically the same character over time (genetic correlation between trait measures equal to one). Furthermore, the variances of the measures should be equal with the environmental components remaining the same over multiple measures. In cases where the composition of a trait is likely to change over time it is desirable to explicitly model the relationship between age and the relevant effects. This was done by fitting a random regression model.

Longitudinal Analysis

Polygenic A random regression based covariance function model (Section 2.2.2) was fitted to the full (up to 26106 records) data set for each trait. This allowed estimation of the coefficient matrices associated with the covariance functions for additive genetic and

permanent environmental random effects. The order of polynomial used in the RR was one; this generated 2×2 matrices of coefficients. The coefficient matrices for the genetic and permanent environment terms are denoted by matrices A and P, respectively. The estimated matrices were used to calculate the covariance matrices for the full set of 76 ages (ages 20-95) using equation 2.19. This gives the (co)variance decomposition for the 76×76 matrix, T, of phenotypic measures as

$$T = XAX^T + XPX^T + eI \tag{4.1}$$

where $X^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ age_1 & age_2 & \dots & age_{76} \end{pmatrix}$, e is the (temporary environment) error variance, age_n is the n^{th} mean corrected age and I is the identity matrix. Fixed effects were included as described for the repeatability model and a random family effect was screened for significance. This family effect was assumed to be constant across measures.

Note that although a linear polynomial of age is fitted to the data, the graph of variances against age are quadratic; this is because, for the polygenic effect (with coefficient matrix terms $a_{11}, a_{21}, a_{12}, a_{22}$) at age x, the variance contribution is

$$a_{11} + 2 \times [x - \overline{x}] \times a_{12} + [x - \overline{x}]^2 \times a_{22}.$$

An analogous term applies in the case of the permanent environmental effect.

Estimates of the phenotypic and component variances (genetic, permanent environment, error) at any age are given by the appropriate diagonals of T, XAX^T, $XPX^T$and eI respectively. Estimates of heritability were obtained from the relevant variances. In cases where a family effect was included an additional term, $f_{var}XX^T$, where f_{var} is the variance term associated with the family effect, was added to equation 4.1. The off-diagonals of the $n \times n$ matrices are the covariances between the ages. These covariances were standardised to give correlations between the different ages.

The variance of the coefficient matrix entries can be estimated by calculating the information matrix associated with the fitted RR. To ease the computational burden the information matrix was replaced by an approximation, the average information matrix (see section 2.1). ASREML calculates the average information matrix and hence allows estimation of the variances of the covariance function coefficient matrix entries and of the variances of the functions of these values. Of particular interest here is the function

$$[A]_{11} + [A]_{12}\,(age(i) + age(j)) + [A]_{22}\,(age(i)age(j))$$

which is the polygenic covariance between ages i and j (the i, j^{th} entry of XAX^T). This procedure was used to obtain standard errors on the estimates of the genetic correlations between ages 30 and 50, 50 and 70 and 30 and 70. Approximate 95% confidence intervals were obtained by calculating values within 2 standard errors of the correlation point estimates.

QTL The above model was extended to include an additional term for an age dependent QTL effect. The CF coefficient matrix was estimated based on marker-specific IBD matrices. The IBD matrices were calculated as described in the univariate QTL analyses (section 4.3.1).

4.4 Results

4.4.1 Univariate analyses

Polygenic

The polygenic heritabilities and variance components for Total Cholesterol and HDLC from the ASREML polygenic analyses are superimposed on the multivariate graphs (figures 4.1, 4.2, 4.3 and 4.4). The results for BMI and height are given in table 4.1. All of these results include a random effect for family environment (although in some cases it was zero). Without the family environment term the height h^2 values were higher (greater than 0.80 for all age bands) than shown in table 4.1.

age	BMI	height
30	0.495	0.684
40	0.405	0.487
50	0.338	0.522
60	0.307	0.526
70	0.271	0.520
80	0.359	0.535

Table 4.1: BMI and height univariate polygenic heritabilities

QTL

A summary of the highest univariate LOD scores and estimates of the proportion of variation explained by the QTL are given in table 4.2. The significance of the LOD scores should be down-weighted as they are the highest values obtained from tests on four different traits. As explained in the discussion of chapter 2, the estimates of the proportion of variation explained by the QTL are likely to be over estimates. This over-estimation is particularly obvious for the estimate of the proportion of variance explained by the QTL for age 70-80 HDLC on chromosome 12 where the estimate exceeds the estimate of the proportion of variance explained under the no QTL model (polygenic only model explained 73% of the variation in age 70-80 HDLC). This trait also demonstrated the limitations of performing univariate tests. The LOD score and estimated proportion of variation explained by the putative QTL for the HDLC trait at the 119cM position was 0 when the trait measure for ages 60 -70 were used. The univariate tests do not take into account the fact that it is highly unlikely for there to be such a substantial change in the genetic effect

Table 4.2: Univariate QTL results

chromosome	Position (cM)	Trait	Age band	LOD	% Variation explained by QTL
16	95	BMI	20-30	3.12	49
5	183	height	60-70	2.61	49
10	23	HDLC	70-80	2.50	44
12	119	HDLC	20-30	2.46	81
14	138	T.Chol	50-60	2.57	38
19	101	T.Chol	50-60	3.11	38
20	24	T.Chol	40-60	3.03	34

Table 4.3: Repeatability Components of Variance

trait	h^2	$\frac{V_{pe}}{V_T}$	$\frac{V_l}{V_T}$	Repeatability
height	0.498	0.171	0.254	0.924
BMI	0.387	0.397	0.000	0.784
Total Cholesterol	0.411	0.159	0.006	0.576
HDLC	0.379	0.188	0.046	0.613

over such a short period of time. Multivariate tests which incorporate age into the model (such as the RRs fitted) are more appropriate as these do not allow such large disparities between proximal ages.

4.4.2 Multivariate analyses

Repeatability Model

Table 4.3 gives details of the various components of variance for the four traits.

The repeatability analysis assumes that the genetic correlation across the repeated measures is 1. If this is true, the repeatability gives an upper bound to the total genetic component of variance. Estimates of the genetic correlations can be obtained from the following longitudinal analyses. logBMI (not shown) is similar to BMI.

Longitudinal Analysis

Polygenic The longitudinal analyses results for Total Cholesterol and HDLC are displayed in figures 4.1, 4.2, 4.3 and 4.4. The results are displayed in two ways. For each trait the variances are shown alongside the variances from the univariate polygenic analyses. Also shown are the heritabilities with the univariate results again superimposed on the same graphs.

The correspondence between the univariate and multivariate results is good, particularly in the middle age range (40 to 60). The curves are significantly less accurate for

Figure 4.1:

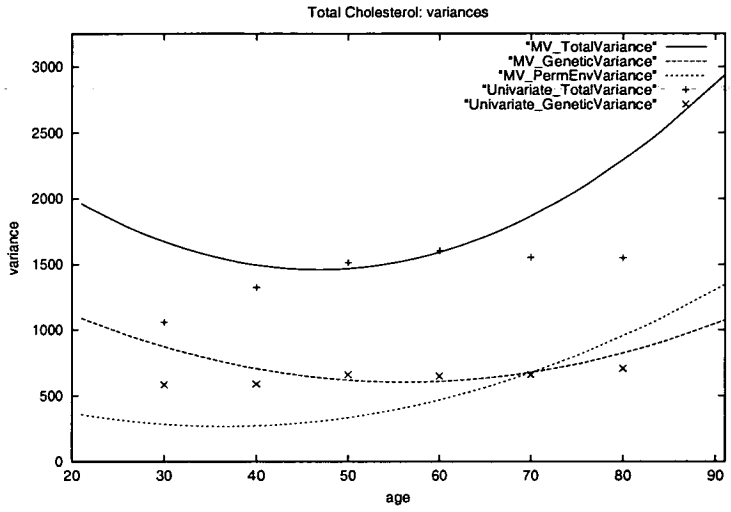


Figure 4.2:

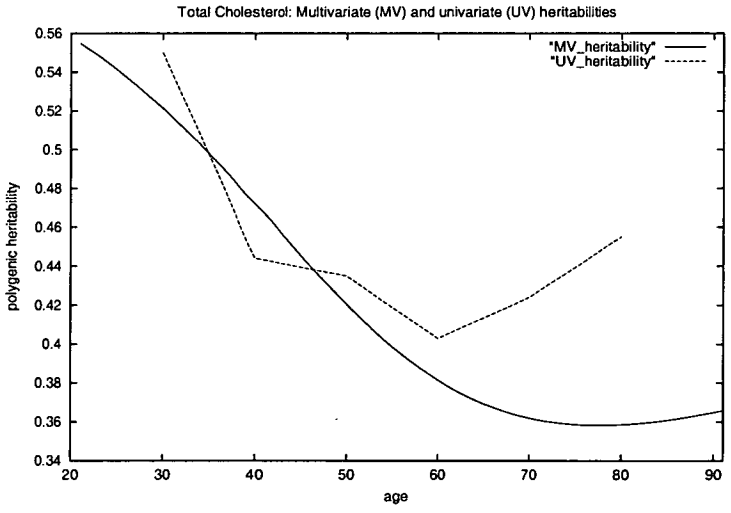


Figure 4.3:

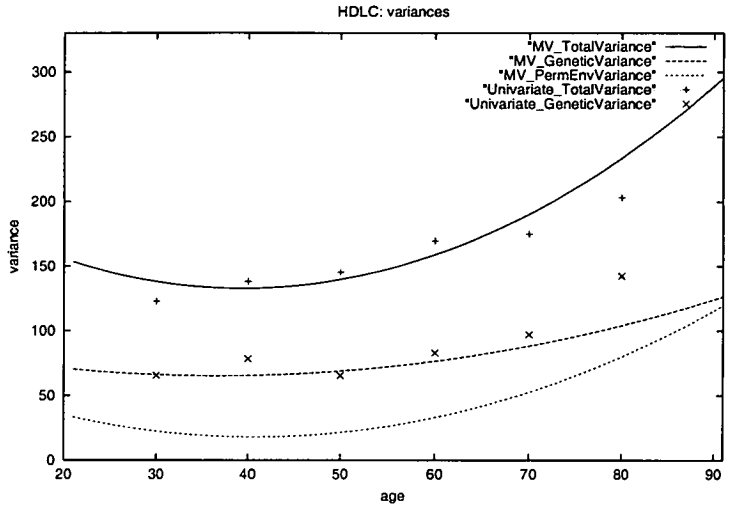


Figure 4.4:

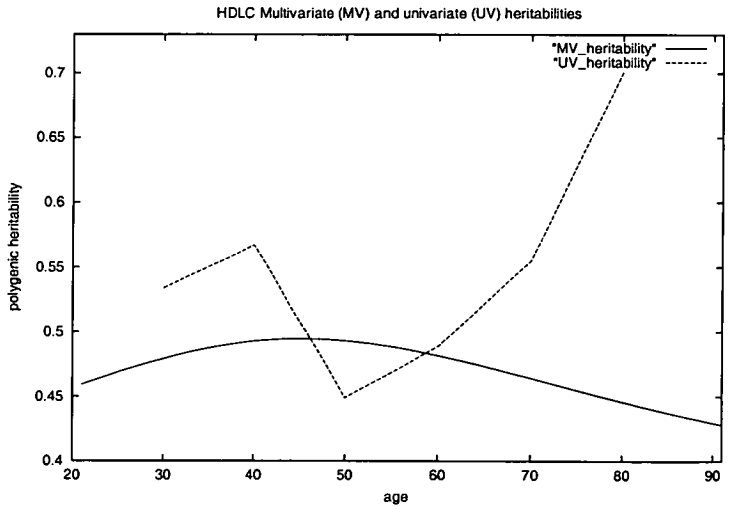


Table 4.4: Phenotypic Correlations

Trait	Phenotypic correlations		
	age 30-70	age 30-50	age 50-70
height	0.79	0.90	0.89
BMI	0.42	0.70	0.84
Total Chol.	0.37	0.57	0.61
HDLC	0.41	0.56	0.64

Table 4.5: Genotypic Correlations

Trait	Genotypic correlations		
	age 30-70	age 30-50	age 50-70
height	0.83 (0.69,0.98)	0.96 (0.81,1.00)	0.95 (0.79,1.00)
BMI	0.42 (0.28,0.57)	0.75 (0.59,0.91)	0.91 (0.74,1.00)
Total Chol.	0.60 (0.44,0.76)	0.90 (0.74,1.00)	0.88 (0.69,1.00)
HDLC	0.80 (0.62,0.98)	0.94 (0.74,1.00)	0.96 (0.77,1.00)

extreme ages since most individuals only have records for ages 35 to 65. Whilst the low order polynomials do not allow the multivariate analyses to closely approximate the univariate heritabilities for traits such as height and HDLC, the true relationship between these traits is likely to be simple, with the univariate results exhibiting stochastic variation about a true smooth curve. Pletcher and Geyer [173] discuss why biological processes are expected to yield reasonably smooth curves.

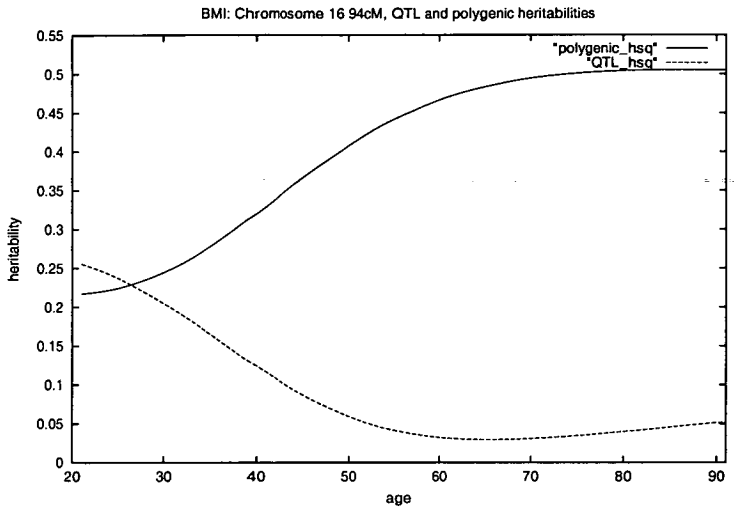
Table 4.4 gives the phenotypic correlations between the traits at different ages. Table 4.5 gives the genotypic correlations, estimated from the model fitting a linear polynomial based covariance function. Approximate 95% confidence intervals are given in brackets after each point estimate. With the exception of BMI, all traits exhibit high genetic correlations across large time periods.

QTL A full genome scan was not performed on the data. Instead a few of the QTL peaks indicated in the univariate analyses were investigated further.

Firstly, the chromosome 16 peak indicated in the univariate analyses for age 20-30 BMI was investigated. Figure 4.5 shows the estimated QTL and polygenic heritability over a range of ages. This QTL is important at lower ages but becomes less so at later ages. The correlation between the QTL heritability at age 30 and at age 50 is high (0.86) but falls away more rapidly when one considers ages 50 and 70 (0.48) and ages 30 and 70 (-0.04). Since the QTL effect is small after the age of 50 it is unsurprising that the correlation is low for later ages.

Secondly, the peak from the 40-60 univariate data for Total Cholesterol on chromosome 20 was examined. Figure 4.6 shows the change in the QTL heritability as one moves along

Figure 4.5:



chromosome 20. The correlation between the QTL effect at different ages was rather higher than for the chromosome 16 QTL, with the correlation between ages 30 and 70 at 24cM being 0.45. This QTL accounted for a sizable proportion of the variance across the range of ages.

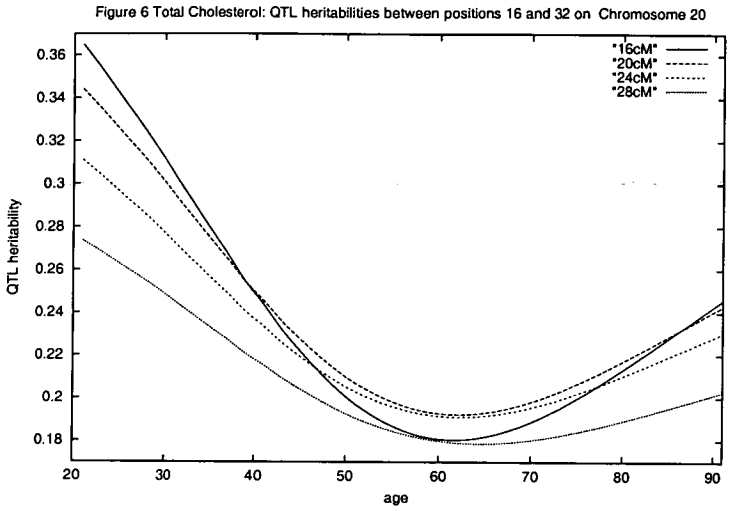
Thirdly, the peak on chromosome 12 (HDLC) was considered. This QTL explained 5% of the total variation at age 30, with the effect rising to 20% at age 80. The correlation between the QTL effect at ages when the effect was largest was high (0.94 between ages 50 and 70), with it decreasing for ages for which there was less of a QTL effect (0.60 between ages 30 and 50).

The other four QTL peaks listed in the univariate results section were considered. However, convergence problems prevented further analyses of these QTL. Similar problems arose when second order polynomials were fitted to the data.

4.5 Discussion

The basic random regression model for polygenic genetic effects, described in [156], has been utilised to allow an analysis of a large number of complex human pedigrees. The covariance function framework has been extended (Chapter 2) to allow the inclusion of marker specific IBD information. This allowed QTL mapping in a large scale genome scan of longitudinal data. The longitudinal method presented efficiently deals with the longitudinal structure of the data (by fitting a polynomial of age), allowing all of the available data to be used in a single, powerful analysis. Given the irregular spacing of the phenotypic measures over time, the fitting of a (covariance) function of age was crucial in allowing multivariate techniques to be utilised.

Figure 4.6:



The results described give a good indication of how the components of variance of these important traits change over time. The multivariate QTL analyses indicated that one of the QTL detected acted across the range of possible ages whilst the other two acted more strongly at the extremes of the age ranges. The agreement between the univariate and multivariate analyses performed was good and some of the larger QTL effects from the univariate analyses were more fully characterised in the longitudinal analyses. Although the repeatability analysis was easier to perform than the longitudinal analysis, the assumptions inherent in such an analysis are likely violated for most of the traits here. In particular, BMI and total cholesterol were shown to have genetic correlations across ages significantly below one, making the repeatability model an inappropriate choice. Modelling the age dependence with a function of age ensured that a sensible pattern of QTL effect changes was prescribed; by contrast, the univariate results implausibly suggested that the QTL effect for HDLC on chromosome 12 underwent dramatic changes in effect as age increased.

For some of the known CVD risk factors there are clear secular trends in trait levels. For BMI, the rise in individuals classified as clinically obese (BMI > 30) doubled in the United Kingdom between 1980 and 1991 with similar increases in the U.S.A. [177]. Whilst BMI is known to have a strong genetic component, with genetic factors explaining 20 to 90% of the observed variation [144], environmental factors must be appropriately accounted for if the genetic component of BMI is to be properly characterised. In the analyses described here, cohort (either 1 - 1948 or 2 - 1971) was fitted as a fixed effect in an attempt to remove the effects of secular trends. Given the rapid change in environment over secular time, a covariate which changes more than just twice in the data set (e.g. a birth data covariate with levels 1900 to 1970 instead of just 1 and 2 for the two cohorts)

may remove more environmental noise. Instead of just fitting year of birth it may be possible to identify some of the factors causing the secular trend. One of the the primary determinants of increased obesity levels over (secular) time appears to be the rise in physical inactivity [177]. There were no measures suitable for assessment of this in the GAW13 data; measures such as 'hours of television watched' mirror obesity levels in recent years [177, 180] and would be suitable covariates. Adult height is highly heritable ($h^2 = 0.8$) and is known to have been subject to secular trends [207]. Both total and HDL cholesterol have shown small decreases with secular time in the late 20^{th} century [175, 170]. However, the trends for height, total and HDL cholesterol are likely to have been less dramatic (in 20^{th} century U.S.A. at least) than those for BMI and are probably fully accounted for by the fitted cohort fixed effect.

The RR based analysis method provided estimates of the genetic correlations between different ages through the estimation of genetic covariance functions. These provided plausible estimates of genetic correlation for all of the traits. The estimate for the genetic correlation between ages 30 and 70 for height is less than 1 (0.83). Although this might be expected to be closer to 1 the confidence interval on the point estimate is close to including 1 (0.69,0.98). Furthermore, the fitted covariance function is based upon phenotypic data predominantly measured in the middle age range (ages 40 to 60). Extrapolation beyond this range will lead to less accurate estimates of the true correlation.

Fitting a higher order polynomial for the relationship between age and the genetic and permanent environmental effects may have resulted in a closer fit between the univariate and multivariate results but there will likely be practical problems fitting such models. As alternative to polynomial based random regression approaches, character process models [108] may be useful for longitudinal data analyses, particularly when the correlation between trait measures at distant ages is low. However, Jaffrezic and Pletcher [108] indicate that when the correlations between trait measures over time are high (as is the case for most of the traits here) polynomial based methods are effective.

It should be borne in mind that for some traits there may be correlations between the trait value and survival. This may lead to biased QTL effects for QTL acting at later ages. It seems unlikely however, that any of the particular QTL considered here accounts for more than a small proportion of the variation in trait value (although the estimates of such variances may be over-estimates, see [21] and section 2.3). Such QTL are hence unlikely to be strongly purged from the population. Furthermore, the maximum likelihood based procedure used here can account for selection (p793, [139]) when the founder individuals are unrelated, unselected and non-inbred and phenotypes are available for all non-founders. Although this property of ML based estimation may not necessary apply here alternative methods of estimation (such as least squares) do not account for selection either.

The longitudinal multivariate QTL analysis using RR presented here enabled the characterisation of QTL effects over time using all the available data. The RR method was seen to be more appropriate for these traits than simple repeatability or univariate methods.

Time constraints prevented a full longitudinal genome scan for QTL. This analysis is now computationally feasible, however, and the results shown here indicate that this could be a very useful/informative approach, possible for other large data sets.

Chapter 5

A Genome Scan and Follow Up Study Identify a Bipolar Disorder Susceptibility Locus on Chromosome 1q42

5.1 Introduction

Bipolar disorder (BPAD) and schizophrenia (SCZ) are severe psychiatric illnesses, with each affecting approximately 1% of most human populations. There is strong evidence for a genetic etiology in such disorders with high heritabilities reported in twin and adoption studies. However, the task of identifying genomic regions conferring susceptibility has yielded inconsistent results, with a large number of candidate regions identified [187].

In recent years, several studies have identified two regions of chromosome 1q (1q21 and 1q42) as important in the etiology of schizophrenia. At 1q21, a study of Canadian families produced a logarithm of odds (LOD) score of 6.5 [34], a study analysing British and Icelandic families generated a LOD of 3.2 [92] and a family based association study considering Spanish origin families reported a p of 0.003 [195]. A meta-analysis of most of the recent schizophrenia genome scans reported the 1q21 region as being amongst the most likely to harbour a schizophrenia susceptibility locus ([137], but see also below for other meta-analysis results). However, this result arose at least in part as a consequence of the inclusion of the Brzustowicz et al. [34] and Gurling et al. [92] data in the meta-analysis. Interest in 1q42 began when the region was implicated by the apparent effects of a chromosomal abnormality on major psychiatric disease in a large Scottish family [216]. The family segregated a balanced t(1;11)(q42;q14.3) translocation, with the presence of the translocation appearing to be linked with disease status. A linkage analysis consider-

ing the translocation as a marker generated a LOD of 3.6 [26] when individuals with SCZ were considered affected, a LOD of 4.5 when individuals with recurrent major depression or BPAD were considered affected and a LOD of 7.1 when individuals with SCZ, BPAD and recurrent major depression were treated as affected. The translocation was inferred to have directly disrupted 2 genes on chromosomes 1 and 11: these have been named DISC1 (OMIM 605210) and DISC2 (OMIM 606271), respectively [158]. Whilst this result shows a striking relationship between the presence of the translocation and psychiatric disease, it was not immediately clear if this result was of relevance to other families in the general population. In the last five years however, a number of studies have reported independent evidence for the role of 1q42 in psychiatric disease susceptibility. Two studies in Finnish families affected by schizophrenia generated LODs of 3.82 and 3.21 [104, 62] for markers close to the translocation break-point. A recent study of Taiwanese families reported nominally significant evidence for linkage to 1q42 for schizophrenia [107]. Since the translocation family also showed linkage between the translocation and recurrent major depression and BPAD, the results of BPAD linkage studies are also of interest. A study of 22 families affected by bipolar disorder reported a LOD of 2.3 to chromosome 1q32, with allele sharing in affected individuals reported to extend across the 30cM region spanning 1q25-q42 [54]. Interestingly, 15 of these 22 families included at least one individual affected by schizophrenia or schizoaffective disorder. A genome scan of 65 North American bipolar families resulted in a LOD of 1.4 for linkage to a marker on chromosome 1q41 [150]. Other positive reports of linkage between markers on chromosome 1q42 and bipolar disorder include a recent study of British and Icelandic families (maximum HLOD 2.0 at D1S251 [46]), a study of North American families (maximum HLOD 1.98 at D1S103 [78]) and a study of Old Order Amish families ($p<0.0001$ under one non-parametric weighting function at D1S103 [128]). Together, these results lend support to the hypothesis that bipolar disorder, recurrent depression and schizophrenia may share causal elements despite clear diagnostic differences ([247, 23, 26], see also chapter 6).

The population wide significance of these loci on 1q has been the subject of recent lively debate [134, 143, 19, 133]. The results reported in a meta-analysis of affected sibling pairs (ASP) [134] are in striking contrast to the strong linkages reported in analyses of extended family samples [26, 34, 92]. It has been previously suggested (e.g., [143], chapter 6) that, in the presence of locus heterogeneity, the power of data sets comprising small family structures such as sib pairs will be poor. Large extended families (which are likely to be more genetically homogeneous) have proved more useful in identifying susceptibility loci on 1q thus far. For this reason the families ascertained for this study were primarily extended (average family size 18, average number of affected individuals per family 7).

An initial genome scan for susceptibility genes was performed on 13 families affected by schizophrenia or bipolar disorder. These families were part of the European Science Foundation (ESF) project on the molecular neurobiology of mental illness (Full results unpublished). Secondary analyses were then performed on an extended superset of the ESF families and on 9 additional families on chromosome 1. All families were Scottish, car-

ried no known chromosomal abnormalities and were unrelated to the previously described translocation family [26]. Robust, multipoint variance components techniques were used to ensure maximal use of the available genotypic information. Additional parametric linkage analyses were also performed.

5.2 Materials and Methods

5.2.1 Study Sample

Sample collection 13 Scottish families (6 BPAD, 7 SCZ) were originally recruited to take part in the ESF project. 132 individuals (64 BPAD, 68 SCZ) were typed for 372 microsatellite markers across the genome to identify regions contributing to psychiatric illness. Family members were interviewed by experienced psychiatrists (Douglas Blackwood and Walter Muir, University of Edinburgh) using the schedule for affective disorders and schizophrenia (SADS-L) [95]. Diagnoses, based on interviews, case note reviews and information from carers and relatives, were based on DSM IV criteria [7]. Families were categorised as either BPAD or SCZ. In the ESF project, families were included where relatives of schizophrenic probands were diagnosed as schizophrenia, schizoaffective disorder or recurrent depressive disorder. Bipolar families included affected individuals with bipolar I, bipolar II, schizoaffective manic or recurrent depressive disorder. Families in which both schizophrenia and bipolar disorder were diagnosed in relatives were not included in the ESF study.

Subsequent to the ESF study additional families were recruited and some families extended. Since the family in which the t(1;11) translocation segregated with major mental illness included relatives with schizophrenia, recurrent major depression and a case of bipolar disorder, the secondary analysis (of the extended sample) included those families classified as "mixed". These "mixed" families had both schizophrenia and bipolar disorder diagnosed in relatives. In all cases the vast majority of individuals in each family were either schizophrenic or bipolar. The families are described in the results as "bipolar" or "schizophrenic" depending upon the predominant diagnosis. In the case of the largest family, a small nuclear sub-branch included a number of schizophrenic sib pairs but the remainder of the family included mainly affective disorder individuals. In this case the small schizophrenia sub-branch was considered a separate schizophrenia family with the rest of the large family considered a bipolar family.

Including the ESF families, 22 families (10 bipolar, 12 schizophrenia) comprising 398 (229 BPAD, 169 SCZ) individuals were considered for analysis. Whilst some families were nuclear (5 families) most were extended (17 families): the structures of two of the bipolar families are shown in figures 5.1 and 5.2. Individuals with bipolar or unipolar disorder are shaded in black, individuals with minor psychiatric illness (e.g. minor depression, anxiety, alcoholism) or unknown phenotype are shaded in grey and unaffected individuals are shown with open symbols. Tables 5.1 and 5.2 indicate the number of individuals affected

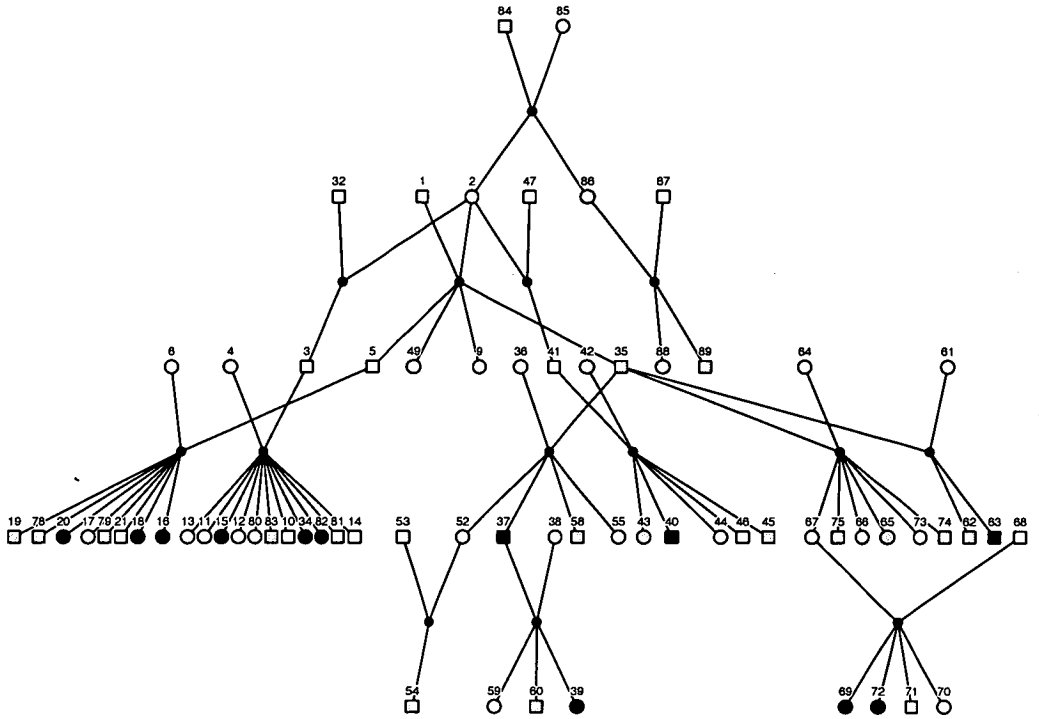


Figure 5.1: The largest bipolar family in the chromosome 1 analysis

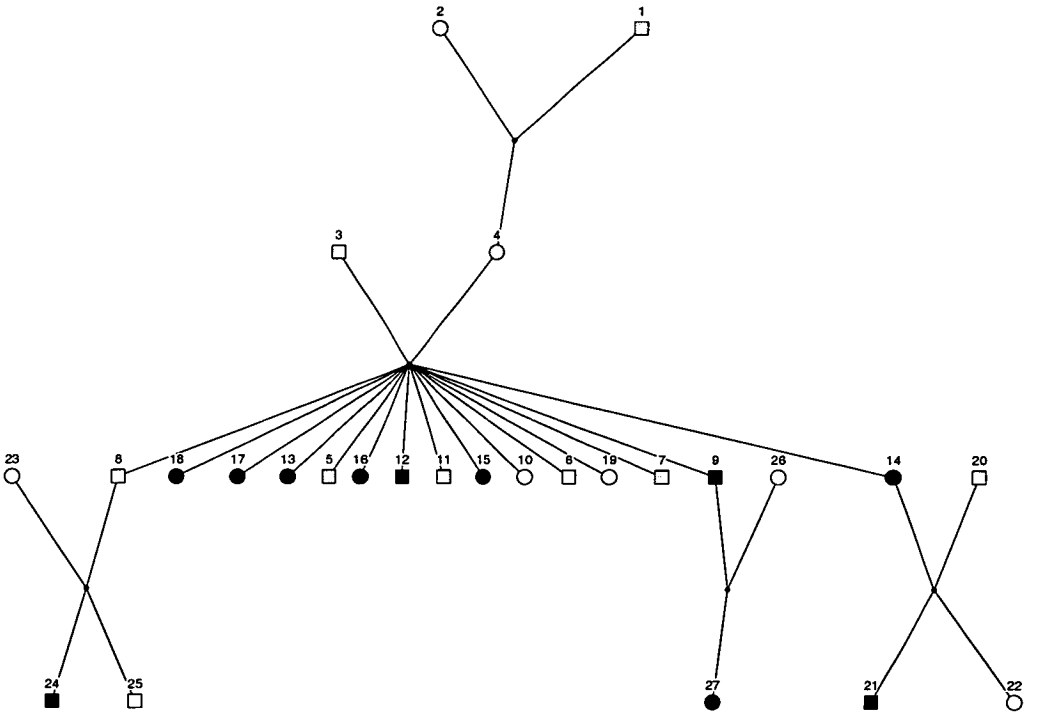


Figure 5.2: One of the densely affected bipolar families

	disease definition	
family number	narrow	broad
1	5	5
2	6	6
6	5	5
9	4	4
10	4	6
11	4	6
20	5	6
19	4	8
29	3	3
33	4	5
36	5	8
500	4	4
total	53	66

Table 5.1: Schizophrenia families summary: number of affected individuals

	disease definition	
family number	narrow	broad
4	8	9
5	6	12
12	5	7
15	6	7
18	3	6
24	8	11
26	5	7
28	4	7
32	5	7
54	5	5
total	55	78

Table 5.2: Bipolar families summary: number of affected individuals

under the narrow and broad definitions (see below for these definitions) of affection for the schizophrenia and the bipolar families.

Genotyping methods The ESF families were typed at the Human Genome Research Centre Genethon. The 372 microsatellite markers were taken from the Genethon reference map [56] and were equally spaced across the genome. Mendelian inconsistencies were resolved before further analysis.

The additional families were typed for 46 markers across chromosome 1. Many of these were single nucleotide polymorphism markers (SNP), typed in house (P. Thompson, University of Edinburgh). Additional microsatellite markers were also typed in some families. The markers in the 1q42 region are displayed in figure 5.3.

The data were scanned to remove unlikely double recombinants (in addition to Mendelian

transmission errors, criteria for removal p<0.05 in MERLIN), using the program MERLIN [1]. Since several of the families were too large for exact analysis using MERLIN, some of the pedigrees had to be split up to perform error checking. The families were analysed whole in the single point parametric and multipoint VC linkage analyses however. Unlike the genome scan data, the additional families were not typed for all markers. Since we were particularly interested in the area around 1q42 all families were typed for marker D1S103, with the vast majority also being typed for D1S459 (266 and 221 individuals after data cleaning, respectively). Families which did not show evidence for linkage were not typed for further markers on the chromosome. Most other markers on chromosome 1 were typed in around 100 individuals. The uneven distribution of marker information is dealt with effectively by the multipoint procedures described below.

Statistical Methods

The same methods were applied to the BPAD sample and the SCZ sample and the methods described below apply in both cases. Two-point (i.e. one molecular marker together with the inferred disease genotype) parametric linkage analysis using FASTLINK [44] was performed across the genome. Two models were fitted to the data; one 'dominant' (labelled model b) and one 'recessive' (labelled model r). Further, under the dominant model, a narrow definition phenotype (labelled model a) was used in addition to the broad definition phenotype . For the schizophrenia families, the narrow definition considered schizophrenic and schizoaffective individuals as affected: the broad definition also considered recurrent depression individuals as affected. For the bipolar families, individuals with bipolar I, bipolar II and schizoaffective (manic) disorder were regarded as affected: the broad definition also considered recurrent depression individuals as affected. For the extended sample, families with both bipolar and schizophrenia were included (mixed families). In the mixed families the narrow definition included the diagnoses schizophrenia, schizoaffective, bipolar I and bipolar II. The broad definition added in recurrent depression. Recurrent depression individuals were regarded as disease status unknown for all narrow definition analyses.

In the case of simple Mendelian dominant disorder the penetrance parameters in the parametric linkage analysis can be simply specified as being $f_0 = 0$ for homozygous wild type disease genotype carriers and $f_1 = f_2 = 1$ for heterozygous or homozygous disease allele disease genotype carriers, where f_x denotes the penetrance parameter for an individual carrying x disease alleles at the putative disease locus. The (conditional) probability of having the disease phenotype given the disease genotype would be $f_0 = 0$, $f_1 = 1$, $f_2 = 1$ for affected individuals and $1 - f_0 = 1$, $1 - f_1 = 0$, $1 - f_2 = 0$ for unaffected individuals. These probabilities are factored into the likelihood for each family in the data. To model the non-Mendelian inheritance pattern in the psychiatric diseases of interest here, a number of different sets of penetrance parameters (sometimes called liability classes) are specified. The penetrance parameters of the unaffected individuals are specified to take into account how long they have lived without being affected by disease. Since un-

	model b			model a			model r		
age	f_0	f_1	f_2	f_0	f_1	f_2	f_0	f_1	f_2
<20	0.0005	0.19	0.19	0.0001	0.15	0.15	0.0003	0.0003	0.15
<30	0.0025	0.77	0.77	0.0005	0.62	0.62	0.0012	0.0012	0.62
>30	0.0025	0.88	0.88	0.0005	0.7	0.7	0.0012	0.0012	0.7

Table 5.3: Penetrances: unaffected individuals

	model b			model a			model r		
definition	f_0	f_1	f_2	f_0	f_1	f_2	f_0	f_1	f_2
narrow	0.0025	0.88	0.88	0.0005	0.7	0.7	0.0012	0.0012	0.7
broad	0.0025	0.88	0.88	0.5	0.5	0.5	0.0012	0.0012	0.7

Table 5.4: Penetrances: affected individuals

affected older persons represent more reliable indicators of affection status three liability classes are specified for persons under the age of 20, under the age of 30 and over the age of 30. Two liability classes were specified for affected individuals: this allowed one analysis with both narrow and broad definition individuals regarded as affected and one analysis with narrow definition individuals regarded as affected but broad definition individuals regarded as having unknown disease status. The penetrance parameters are given in tables 5.3 and 5.4. For the affected individuals f_0, f_1, f_2 are factored into the likelihood whilst for the unaffecteds $1 - f_0$, $1 - f_1$, $1 - f_2$ are used. The penetrance parameters are referred to as probabilities here but strictly speaking they need not lie in $(0, 1)$; inference is based on ratios of the constructed likelihoods and multiplication of the penetrances values by an arbitrary constant will not change the likelihood ratio.

As explained in chapter 1 the specified penetrance parameters are necessarily just educated guesses at appropriate values; single marker parametric analysis is robust to mis-specification of these parameters provided at least a dominant and a recessive disease model are used. Whilst multipoint parametric linkage analysis has greater power to detect loci when the putative locus is not near a fully informative marker, it is not robust to mis-specification of the parameters in the model (Chapter 1, [201]). Explicit modelling of such mis-specification errors within the multipoint parametric framework is possible [85] but not attempted here. A convenient and robust alternative to multipoint parametric linkage analysis is multipoint variance component linkage analysis.

For the extended samples (229 individuals for the BPAD analysis, 169 for the SCZ analysis) two-point parametric linkage analysis was performed for the marker typed in all families, D1S103. Multipoint variance component (VC) linkage analysis was performed with the chromosome 1 markers. A random polygenic effect and a random effect for family were fitted as a basic model. Variance components attributable to quantitative trait loci (QTL) effects were calculated by utilising multipoint identity-by-descent (IBD) coefficients estimated from the marker data. The significance of including a component attributable to

one or more such effects is tested via likelihood ratio tests. Standard VC analysis assumes that the phenotypic data are multivariate normal. Since the data are binary, a threshold model (e.g., [139]) was used within the program. The threshold model maps the binary observed phenotypes to an underlying normal distribution, via a probit transformation. Analysing binary data without the threshold model is known to affect the robustness of the test statistic with samples differing in the proportions of affected individuals yielding different type I errors [3]. The variance components technique was attempted for the ESF data set but, since the sample size was small, the variance components could not be reliably estimated. With the additional families the VC technique had greater utility, giving estimates of disease heritability in addition to measures of QTL significance (LOD scores). To minimise multiple testing only the broad definition phenotype was used for the chromosome 1 analysis. SOLAR [6] was used for the likelihood maximisations and IBD computation.

Since some of the families were preferentially selected for typing at additional markers on chromosome 1q (three of the families which showed no linkage signal to D1S103 were not typed for further markers), the single point LOD score calculated at markers other than D1S103 may be biased up-wards. However, if the markers are analysed within a multipoint framework the region around D1S103 should yield unbiased LOD scores. Since the heterozygosity of the microsatellite D1S103 was 0.8 marker information was high for the majority of individuals around this region. We would expect the information content in all families to remain high enough for multipoint statistics to remain unbiased for at least 10cM either side of D1S103. For this reason multipoint LOD scores are only displayed in the region around 1q42.

Although having more markers available in all families would have enabled more efficient detection of genotyping errors, the small number of families (3) only typed at D1S103 did not contribute to the linkage signal. Genotyping errors in such families would have little impact on results since genotyping errors invariably decrease evidence for linkage (in families segregating the mutation of interest). The large number of markers around 1q42 in the majority of families allowed efficient checks of genotyping errors to be performed in these families.

'Non-parametric' procedures were not utilised since (1) they are no more powerful than VC methods [250] and (2) they can be shown to be equivalent to parametric methods given particular penetrance values ([87]: Goring and Terwilliger, 2000b). Goring and Terwilliger (2000b) detail why the distinction between the two is somewhat arbitrary and explain that one should not select a method simply because it is of a particular type.

In addition to the linkage results we estimated the overall (polygenic) heritability of the traits on the binary (observed) scale. Robertson's equation

$$h^2_{continous} = \frac{h^2_{binary}\Phi_p(1-\Phi_p)}{[p(x_p)]^2}$$

where $p(x_p) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x_p^2}{2}}$ and Φ_p is the incidence

disease	chromosome	model	marker	LOD
bipolar	1q	b	D1S229	1.55
schizophrenia	3p	a	D3S3721	2.00
schizophrenia	8p	b	D8S1989	1.71
bipolar	8q	b	D8S1741	1.53
bipolar	9q	b	D9S175	2.35
schizophrenia	19q	a	D19S220	1.59

Table 5.5: Maximum two-point LOD scores for ESF families

from Dempster and Lerner (1950) [53] was used to convert this binary scale measure to a continuous underlying scale heritability. To ensure there was no upward bias in this estimate due to environmental effects, a random effect for familial environment (household) was fitted.

5.3 Results

5.3.1 ESF data: genome scan

Parametric linkage LOD scores exceeding 1.5 are given in table 5.5. The highest LOD score achieved (chromosome 9) was not at a region previously identified as contributing to psychiatric disease. However, the genomic region identified on chromosome 1 is in close proximity to the DISC1 gene, a candidate gene for schizophrenia identified via a chromosomal translocation [216] and recently replicated in independent samples [62, 104, 107].

5.3.2 Chromosome 1 analyses

The above result prompted our group to type further markers around this region in the ESF families. Furthermore, several more families from a similar geographic location were also available for analysis and some of the ESF families were extended.

Bipolar Results Analysing all bipolar families (229 individuals) together at marker D1S103 with the single point variance components procedure yielded a LOD score of 2.17 (2.15 without familial environment removed with a random effect for family). The two point parametric LOD (broad definition, recessive model, θ= 0.1) was 2.56 at D1S103. The highest single family parametric LOD was 2.16 at marker D1S419. The next highest single family LOD, 2.00, was at D1S103 but this family was only typed at D1S103 and D1S459. Individual family LODs at D1S103 ($\theta = 0.1$) are given in table 5.6. Note that the LOD scores shown in table 5.6 are not strongly negative in the families displaying evidence against linkage because the LOD is evaluated at $\theta = 0.1$ rather than at $\theta = 0$. That is to say, these families are *not* simply uninformative for linkage. At $\theta = 0$ the LODs are higher

Family	LOD
4	-0.14
5	1.75
12	-0.32
15	0.11
18	-0.17
24	1.21
26	0.47
28	-0.21
32	0.09
54	-0.24

Table 5.6: Bipolar families: By family parametric LOD scores at marker D1S103

in the families showing linkage but summed over all families the LOD maximum occurs when $\theta = 0.1$. The evidence for linkage under the narrow definition model was less than under the broad definition, with a maximum parametric LOD of 0.77.

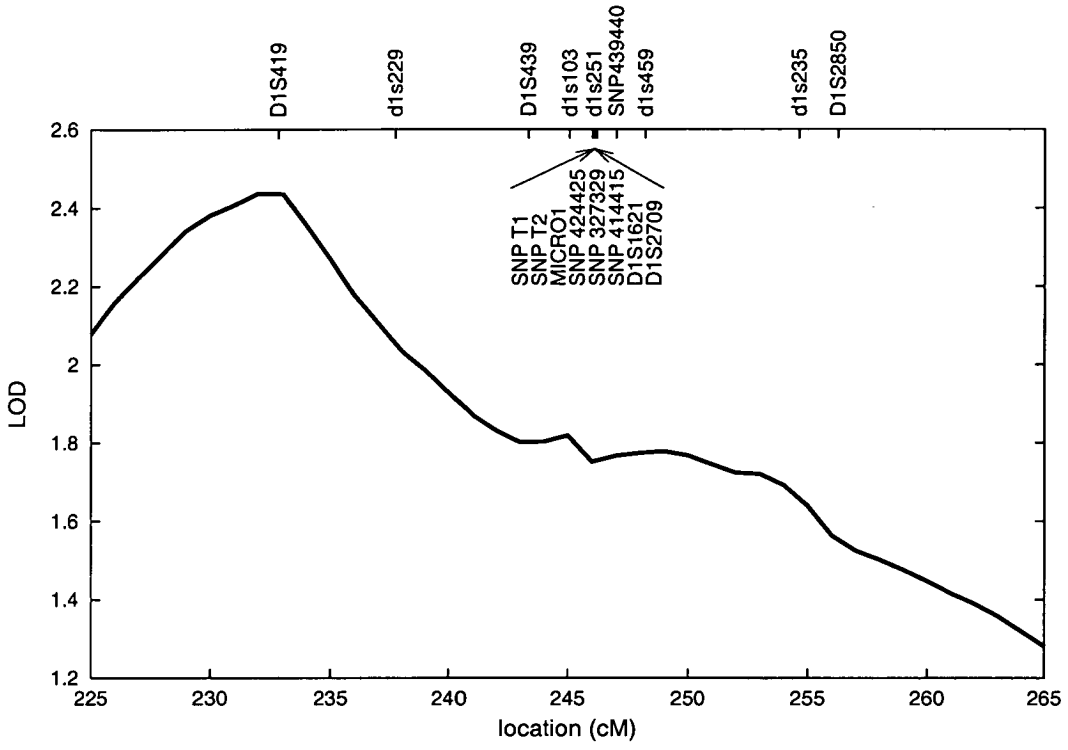
Multipoint variance component LODs are displayed in figure 5.3. The maximum LOD was 2.43 at position 233cM (near marker D1S419, 12cM from D1S103). The estimate of polygenic heritability was 0.71. Without a familial environment term this produced an (upwardly biased) estimate of 0.86.

Schizophrenia Results Analysing all schizophrenia families (169 individuals) together at marker D1S103 with the single point variance components procedure yielded a LOD score of 0. The single point parametric maximum LOD (dominant model, $\theta = 0.3$) was 0.25 at D1S103. Multipoint variance component LODs were less than 0.2 in the 30cM around D1S103. The estimate of polygenic heritability was 0.64. There was no evidence for a family environment term.

5.4 Discussion

This chapter reported the results of a genome scan for psychiatric disease susceptibility loci in 13 Scottish families. In the genome scan, linkage peaks with LOD scores > 1.5 were found on chromosomes 1q (BPAD), 3p (SCZ), 8p (SCZ), 8q (BPAD), 9q (BPAD) and 19q (SCZ). The linkage peak on chromosome 1q was followed up in a substantially larger sample (22 in total, 398 individuals) of families affected by schizophrenia (SCZ) or bipolar affective disorder (BPAD). Adding 9 extended families, together with more individuals from the original (ESF) families, increased the evidence for linkage to bipolar disorder (maximum single marker parametric LOD 2.56), providing further evidence for the importance of the 1q42 region as a risk factor for psychiatric disease. Multipoint variance components linkage gave a maximum LOD of 2.43 12cM from the previously identified schizophrenia susceptibility locus, DISC1 [158].

Figure 5.3: Multipoint Variance Components Linkage: Bipolar Families



To minimise the effect of genetic heterogeneity, large extended families (average family size >18) were ascertained. The families collected were Scottish, carried no chromosomal abnormalities and were unrelated to the large family previously reported as segregating a balanced t(1;11) translocation with major psychiatric disease.

When DISC1 was first identified in a Scottish family which segregated a balanced translocation with major psychiatric disease [216] it was not clear how relevant this locus was to other families or populations. Furthermore, whilst the translocation family which allowed identification of DISC1 had several schizophrenic individuals, the highest LOD score was achieved when a number of unipolar individuals and a bipolar individual were included as affected. This study provides evidence for the effects of a susceptibility locus (or loci) for psychiatric diseases in the 1q42 region in a set of independent Scottish families. Some other studies have reported evidence for linkage of 1q42 to schizophrenia, with two Finnish studies [104, 62] and a Taiwanese study [107] providing evidence for the relevance of the 1q42 region in different populations. The 1q42 region has also been implicated in bipolar disorder susceptibility, with a number of studies, considering a number of different populations reporting evidence for linkage to 1q[54, 78, 128, 150, 46]. The possibility of distinct psychiatric disorders such as bipolar and schizophrenia sharing susceptibility loci has received attention in the literature [23, 62, 26] and, given the main reports of linkage to 1q have been in schizophrenia, the results presented here add weight to this assertion.

There is evidence for an increase in familial risk for one disorder in the presence of the other (e.g., [23], see also chapter 6) and the data presented here suggest that susceptibility loci such as DISC1, may be acting to increase the genetic risk of both. Interestingly, there was negligible evidence for linkage to 1q42 in the schizophrenia families considered here. However, the sample analysed had limited power to detect loci of small effect and, in the event of there being substantial locus heterogeneity, the sample may include families which by chance are affected by psychiatric disease as a result of loci unlinked to 1q42. It is therefore possible that the failure to detect linkage to schizophrenia in these families was a false negative result.

The bipolar multipoint peak was ~12cM from the marker D1S103, mainly as a result of 2 of the families showing linkage to D1S419. It should be stressed that a 95% confidence interval on the peak is likely to be in the tens of centimorgans and that the marker information was only complete across all families at D1S103. The DISC1 gene (MIM 615210, [158]), less than 1cM from D1S103 on 1q42.1, represents the strongest candidate gene and it seems likely that random variation (and/or possible bias due to selective typing of families for markers around D1S103) has moved the linkage peak from this point.

Some 80cM from the DISC1 region, two other groups have reported strong linkage to chromosome 1q21 [92],[34]. These two studies are likely to have found evidence for linkage to a genomic region distinct from 1q42. The 13 family sample analysed here did not show linkage to 1q21 and there was insufficient marker information to adequately assess linkage to 1q21 in the additional families. The bipolar linkage described in [54] is likely to be to 1q42, particularly since the linkage they detected exhibited elevated IBD sharing across some 30cM of 1q, including the DISC1 region. The other linkages on 1q42 described above are clearly to the DISC1 region.

The maximum LOD for the 10 family bipolar data set was obtained when individuals with bipolar I, bipolar II, schizoaffective (manic) disorder or recurrent major depression were regarded as affected (broad definition of affection). The evidence for linkage decreased when individuals with recurrent major depression were regarded as disease status unknown in the analysis (narrow definition of affection). Although the individuals included in the narrow model definition may give a truer reflection of the underlying biology than the grouping including recurrent depression individuals, the number of affected individuals in the analysis is reduced, potentially reducing power. It is worth pointing out that, whilst recurrent depression individuals under study here were included in the broad disease definition for both bipolar and schizophrenia families, the families were ascertained through narrow definition probands. Furthermore, all families had a least three affected individuals using the narrow definition. The inclusion of recurrent major depression individuals in psychiatric genetic studies is not universally agreed upon and many investigators perform at least two separate analyses under different disease definitions (e.g. [150, 199]).

The genome scan of the ESF families generated a number of positive results alongside the peak on 1q. Of most interest amongst these was the LOD of 1.71 on chromosome 8p.

This region has been implicated in a number of independent studies [218, 31, 34, 92] and may merit further follow up in the 9 additional families described here. None of the other regions indicated by the ESF genome scan overlap with any other published reports of strong linkage.

In summary, a genome scan of Scottish families affected by schizophrenia or bipolar disorder provided evidence for linkage to chromosome 1q in bipolar families. In a further analysis of a larger sample of bipolar families a maximum LOD of 2.56 was found. This was close to the previously identified psychiatric disease susceptibility locus DISC1. This finding supports the results of previous studies implicating this locus in a small but significant subset of all families affected by psychiatric disease and suggests that schizophrenia and bipolar disorder may share a common genetic component in this region.

Chapter 6

Study Design for Psychiatric Genetic Linkage Analyses

Study design in psychiatric genetics has been the subject of intense debate recently [134, 143, 19, 133, 155]. Although genetic linkage analysis has had some success in locating disease susceptibility genes for diseases such as schizophrenia [218, 222, 34], considerable resources have been required to reach this stage. Further progress (in linkage studies) will depend upon appropriate study design and analysis methods.

6.1 Background

Attempts at positional cloning of psychiatric diseases such as schizophrenia began around 15 years ago. Since then, two study designs have emerged, with some research groups favouring the collection of extended families whilst others have focused on collecting affected sib pairs (ASPs). First of all some of the studies utilising extended families are reviewed. ASP based studies are then considered and the recent use of meta-analyses in schizophrenia research is discussed. This chapter is based upon a short comment [143] discussing results obtained on chromosome 1q; background details particular to 1q are given below.

6.1.1 Extended families

A number of studies have concentrated on collecting extended families affected by schizophrenia; the main reason for this is that small numbers of large families may be more genetically homogeneous than large numbers of small families. In particular the aim is to reduce locus heterogeneity. It seems highly likely [188, 137, 92, 34, 242] that schizophrenia will have multiple susceptibility loci so reducing locus heterogeneity will be a very important consideration in study design. The extended family design may prove fruitful because (assuming risk alleles are relatively rare) large families are likely to only segregate one

risk allele, reducing the effects of locus heterogeneity compared with large numbers of unrelated ASPs.

Studies analysing extended families affected by schizophrenia have included a study of Canadian families (average family size 14, average number of affecteds 4, highest LOD on chromosome 1q21 [34]), a study of Micronesian families (average number of affected individuals per family 33, highest LOD on chromosome 2p13-14 [35]) and a study of a very large 12 generation Swedish pedigree (highest LOD on chromosome 6q25, [138]). In each of these three cases genome-wide significant linkage (as per the criteria laid out in [129]) was reported, with the maximum LOD occurring on a different chromosome in each case. A subsequent study of extended families (average number of affecteds per family 4, [92]) supported the findings of the Brzustowicz et al. [34] paper in addition to highlighting other chromosomes of interest. In chapter 5 the families considered for the analysis of data from of chromosome 1 were mainly extended with an average family size greater than 18 (average number of affecteds 5). In each of the above cases, the evidence for linkage derives from only a small number of families (sometimes one family) and the chance of the sample including more than one risk allele will be reduced compared with a data set which samples widely from a large number of families. Other studies have had some success when they recruited families from population isolates such as northeastern Finland [167, 63, 62, 104]. In these cases the potential for locus heterogeneity is reduced but it is worthwhile noting that several loci were detected in both the Finnish studies and a number of other studies (e.g. [34, 92, 35]). This suggests that locus heterogeneity may exist even within these isolates and that gathering large numbers of small families from within an isolate might still sample multiple susceptibility loci.

A chromosomal abnormality, segregating with major psychiatric disease in a large Scottish pedigree (mainly schizophrenia, with a few individuals with affective disorders), allowed identification of the susceptibility locus on chromosome 1q42, DISC1 [216, 158]. When all individuals with major psychiatric disease were considered affected, this pedigree generated a LOD of 7.1 [26]. A sample of British and Iceland extended families [206] were used to look for evidence of linkage to another chromosomal region (5q11-13) suggested by cytogenetic abnormalities. This study [206] was one of the earliest linkage studies of schizophrenia. Only 2 markers were genotyped but these were known to be near a chromosomal abnormality found in two Chinese schizophrenics. A LOD of 6.5 (asymptotic p-value 2.3×10^{-8}) was found in the densely affected families studied. A study giving a failed replication of this result was reported [119] at the same time as the Sherrington et al. [206] article and many researchers, including the original authors, subsequently considered the study of Sherrington et al. to be a false positive result ([114, 18], [220], p284). Interestingly, some recent studies have suggested that the 5q11-13 region may harbour a susceptibility gene after all [137].

In addition to the practice of seeking out the largest families possible, some researchers have specifically targeted particular inheritance patterns within the families. Both Brzustowicz et al. [34] and Gurling et al. [92] selected families in which there was a single

'source' of schizophrenia and that the transmission from this individual was unilineal (arising solely as a result of transmissions from this individual), with the mode of transmission appearing dominant. This strategy is similar to that applied successfully to breast cancer (see chapter 1). Schizophrenia clearly does not segregate in this way in most pedigrees but this approach attempts to single out the few families in which the observed disease segregation pattern appears to be mainly caused by a locus with large effect. In reality there may be number of other background genetic effects conferring genetic susceptibility; the hope would be that these families would segregate a risk allele that conferred substantial disease risk, given this genetic background. Obviously it is difficult to reliably assess the disease segregation pattern if one does not have affected individuals spanning multiple generations (i.e. extended families will be suitable but sib pairs will not be).

6.1.2 Affected Sib Pairs

As indicated in the introductory chapter (chapter 1), the ASP design is favoured because it allows collection of large sample sizes. Risch [189] discusses the use of different sets of relative pairs; he frames his results in terms of mapping genes for diseases in which there are different values of λ_s. λ_s is defined as the conditional probability an individual is affected by a disease given its sibling is affected, divided by the population prevalence of that disease. This value reflects the overall increase in incidence in sibs as a result of both environmental and (polygenic) genetic effects. Risch [189] derives the power of different sets of relative pairs for different values of λ_s; he concludes that for larger values of λ_s (>3, say), second or third degree relatives will offer the most power, assuming relatively informative markers are available [190]. As the value of λ_s decreases (particularly below 2) the power of second and third degree relative pairings is not appreciably more than that available for sib pairs. Since sib pairs are arguably easier to collect, Risch [189, 190] recommends that they should be used for diseases with low values of λ_s. This does *not* mean, however, that only sib pairs should be collected, simply that it possible to show analytically that if a particular set of relative pairs is to be used then affected sibs offer similar power to other relative pairings when λ_s is low. In practice, samples may include families with a range of different family structures. In the third paper in a series, Risch [190] showed there were differences in the degree to which marker polymorphism affected the utility of different relative pairs; when the polymorphism information content (PIC, a measure of marker informativity) was low, sib pairs were better than other relative pairings. Families with a variety of relative pairings (i.e. with more than 2 affected people in the family) may be difficult to appropriately analyse with methods that are based on pairs only (the non-independence of pairings is difficult to deal with when there are multiple possible pairings within a family, see discussion of non-parametric methods in chapter 1). Arbitrary family structures can be readily analysed by methods which model the likelihood of all family members simultaneously (e.g. parametric linkage analysis for binary traits and variance components analysis for quantitative traits, see chapter 1). Since many of the diseases with large values of λ_s, that is diseases that have near-Mendelian inheritance patterns

(e.g. Cystic Fibrosis with $\lambda_s \simeq 500$, [220]), have already been mapped, the primary interest is in (complex) diseases with low λ_s values. That is to say, for these complex diseases, all sets of relatives pairings, including sibs (who would be sub-optimal for diseases with large $\lambda_s$values), will be useful for disease gene mapping. It is important to be aware that although diseases such as schizophrenia have λ_s values greater than 2, it seems likely that there are multiple susceptibility loci involved. The appropriate λ_s value for use in predicting the power of a linkage analysis is the risk value attributable to the locus of interest. In this chapter, the symbol $\lambda_{slinked}$ is used to denote the (genetic) effect resulting from the segregation of a susceptibility allele at this locus. For schizophrenia λ_s is 10 but the $\lambda_{slinked}$ value for any single locus is unlikely to exceed 3 ([188], see also discussion in section 6.4.1). Furthermore, the λ_s value may also be inflated due to the effects of familial environment and hence its use may lead to an underestimate of the number of individuals needed for a given power.

A number of groups have taken the recommendations in [189, 190] and used them as a basis for their schizophrenia studies (see also [40]). Cloninger et al. [41] describe the collection of a sample as part of a United States National Institute of Mental Health (NIMH) initiative. The initiative employed a specific data collection scheme, resulting in a sample comprising almost entirely independent sib pairs (average number of affecteds per family < 2.4). The collected sample was likely to be particularly diverse, with the families designated as being half 'European' and half 'African' [41]. No significant linkage (using the definition of genome wide significance in [129]) was detected in the NIMH sample [71, 118]. Other recent genome scans based on predominantly ASPs (number of affecteds per family < 3 in all cases, average family size ~5 in most cases) include [136, 253, 36, 197, 224, 31]. Only one [31] of these scans reported significant [129] linkage to any chromosome. These studies (i.e. [136, 253, 36, 197, 224, 31, 41]) were all included in a meta-analysis published in 2002 [134]. The 2002 meta-analysis focused solely on chromosome 1q, reporting no evidence for linkage [134].

6.1.3 Schizophrenia Meta Analyses

Enough individual schizophrenia studies have now been performed to make meta-analyses possible. An impressive number of groups have taken part in these initiatives, resulting in a number of recent papers [15, 132, 137, 134]. Since there is substantial overlap in the studies included in these analyses only the recent analyses are discussed here. Badner and Gershon [15, 16] based their meta-analysis on the published results of schizophrenia studies, with a literature search used to accumulate data. Their analysis identified susceptibility loci on chromosomes 8p, 13q and 22q. Lewis et al. [137] gathered the test statistics (or p-values) across the genome from the component studies. The test statistics from the component studies were placed in ~30cM long bins and the overall results assessed based on ranks (this method was called the genome scan meta analysis or GSMA method, [135]). The most significant chromosomal region was 2q with susceptibility loci also indicated on chromosomes 5q, 3p, 11q, 6p, 1q, 22q, 8p, 20q, and 14p (listed in de-

creasing order of significance). Despite there being some overlap in the studies used the linkage peaks on 22q and 8p only appear as the 6^{th} and 7^{th} highest results in the Lewis et al. [137] GSMA. More strikingly, the 13q region did not approach significance in the Lewis et al. [137] analysis; it is unclear why this occurred.

The addition of 11 more samples in the Lewis et al. [137] study compared with the Levinson et al. [134] study resulted in a change in the results reported; the Levinson et al. [134] study reported no linkage to chromosome 1q whilst the Lewis et al. [137] study listed 1q as one of the most likely regions to harbour schizophrenia susceptibility loci.

6.1.4 Affective Disorders

For convenience, the main part of the discussion presented here is restricted to schizophrenia and does not consider affective disorders (bipolar disorder, unipolar disorder). It should be pointed out however, that the long standing Kraepelinian dichotomy [13] - the fundamental distinction between schizophrenia and affective disorders - has been challenged [229, 161]. A continuum model, prescribing a gradual change in symptoms from unipolar disorder to bipolar disorder has been proposed, with some researchers extending this continuum from affective disorders to encompass schizophrenia [232, 161]. For the purposes of genetic studies, it is unclear where to include individuals with symptoms associated with both schizophrenia and affected disorders, referred to as schizoaffective individuals. The risk of schizophrenia in an individual closely related to a bipolar individual is higher than the population average and vice-versa [213, 100, 229]. A degree of common biology may also be indicated by the efficacy of common sets of drug treatments in both cases [161]. There has been considerable speculation that there is overlap between the susceptibility loci identified in studies of affective disorders and schizophrenia ([26, 24, 23, 161], see also the results in chapter 5) . However, the results from linkage studies to date are still too inconsistent to make a firm statement about the existence of common susceptibility loci. A meta-analysis [199] similar to that reported for schizophrenia [137] was performed for bipolar disorder. The results of this meta-analysis [199] indicated no evidence for strong linkage to any particular chromosome and were markedly less significant than those reported for schizophrenia [137]. The reasons for this may lie with higher levels of phenotypic heterogeneity (multiple disease definitions), genetic heterogeneity, smaller sample size and/or random variation in results.

6.1.5 Chromosome 1

A recent paper in Science [134] reported the results of a meta-analysis of families showing no major schizophrenia locus on chromosome 1q and questioned the significance of several recent papers reporting susceptibility loci on 1q. The results based on this multi-centre study of affected sib pairs (ASP) are in striking contrast to highly significant findings in extended families. Significant (HLOD 6.5) linkage at 1q21-22 was detected in Canadian families [34] and replicated in European origin families [92, 205]. At 1q42 Blackwood et al.

[26] obtained a LOD of 7.1 in a single Scottish family. Nearby, Ekelund et al. [62] obtained a LOD of 3.2 in Finnish pedigrees. Modest support for the importance of the 1q42 locus was recently reported in Taiwanese families [107]. How can these apparently conflicting results be reconciled? Levinson et al. [134] suggest that their meta-analysis of ~900 ASP would have sufficient power to detect the chromosome 1q loci detected in the above studies. However, Levinson et al. [134] fail to take proper account of locus heterogeneity. In this chapter the effect of locus heterogeneity on the power of linkage analysis is investigated.

6.2 Methods

The effect of heterogeneity is considered in two different ways. The first calculates the power of the ASP based statistics, both with and without heterogeneity. Two ASP statistics are considered, one based on the mean number of alleles shared identical by descent (IBD) and one based on number of pairs sharing 2 alleles IBD. The simple construction of these tests allows algebraic power calculations. Statistical power is also considered by simulating data and using parametric linkage techniques. In this case parametric linkage techniques are utilised, with the heterogeneity parameter α (see chapter 1, introduction) used to model the families not carrying a mutation at the locus of primary interest.

6.2.1 ASP based tests

Firstly, consider the power of the ASP mean test, a test based on the number of alleles shared IBD.

ASP mean test

Consider a sample of ASPs. The mean number of alleles shared identical by descent (IBD) at the genomic location of interest is $p_1 + 2p_2$, where p_j denotes the proportion of pairs sharing j alleles IBD. Let n denote the number of pairs and $\widehat{p}_j$ an estimator of the proportion of pairs sharing j alleles IBD. If there is no linkage then $p_1 + 2p_2$ has expectation 1 and variance $\frac{1}{2n}$. The test statistic

$$T_m = \frac{(\widehat{p}_1 + 2\widehat{p}_2 - 1)}{\sqrt{\frac{1}{2n}}} \tag{6.1}$$

can be used to test for linkage with a sample of ASPs (equation 16.50b, [139]). With large samples T_m is normally distributed. This test has been shown to be more powerful than alternative statistics (such as T_2, see below) providing the effects of dominance are not strong [139]. The null hypothesis of no linkage implies ASPs share 1 allele IBD 50% of the time and 2 alleles IBD 25% of the time. The deviations from such a null are of interest when one wants to assess statistical power. The IBD distribution at the genomic location of interest can be expressed in terms of the null expectations and additional d_j terms,

representing the deviations due to linkage of this region to a risk locus

$$p_0 = 0.25 - d_0 \quad p_1 = 0.5 - d_1 \quad p_2 = 0.25 + d_2$$

(equation 16.52a in [139]). The d_j can be expressed in terms of the recombination fraction of interest and the relevant locus-specific λ_s (risk to sibs of affected individuals relative to the population risk). Assuming no dominance, $d_1 = 0$, and we have

$$d_0 = d_2 = \frac{(1-2\theta)^2}{4}(1 - \frac{1}{\lambda_s}) \tag{6.2}$$

where θ denotes the recombination fraction between the marker used and the putative disease locus [189]. This is one of the basic results in the paper by Risch [189]; that is, the deviations, $d_.$, from the expected IBD distributions can be expressed in terms of λ_s (Risch [189] also considers other relative pairs) and these deviations can be used directly to determine power (see also below). Modifications of equation 6.2 to include the effects of dominance are given in [189].

The relative risk parameter λ_s represents an effect averaged over all families. We are particularly interested here in the increase in risk due to a segregating mutation in a subset of families. As indicated above (section 6.1.2) the symbol $\lambda_{slinked}$ is used to represent the effects of an allele increasing risk only in those families where it its segregating (i.e. not averaged over all families).

Power of ASP mean test

Calculating Power The sample size for a normally distributed test statistic such as T_m, with significance level α, to have power $1 - \beta$ can expressed as

$$n_{T_m}^{homog} = \left(\frac{z_{(1-\beta)} f_1 + z_{(1-\alpha)} f_0}{\mu_1 - \mu_0} \right)^2 \tag{6.3}$$

(equation A5.4b in [139]) where the zs link the α and β to the normal distribution, the μs represent the test statistic means of the null and alternative hypotheses. n_{T_m} is the number of pairs needed when using the statistic T_m. The fs are given by

$$f_i^2 = \sigma_i^2 n$$

where σ^2 is the test statistic variance.

For the ASP test μ_1 is given by

$$\mu_1 = p_1 + 2p_2 = (0.5 - d_1) + 2(0.25 + d_2) = 1 - d_1 + 2d_2$$

whilst the null mean (μ_0) is simply 1 (the expected number of alleles shared IBD when there is no linkage). Details of calculation of the f_i^2 are given in Lynch and Walsh [139].

f_0^2 is 0.5 whilst f_1^2 is given by

$$f_1^2 = 0.5 + d_1 - (2d_2 - d_1)^2. \tag{6.4}$$

For all plausible values of d_j the values of f_1 and f_0 are very similar and number of sib pairs can be well approximated by altering the numerator of equation 6.3 so that $f_1 = f_0$. For example with $\lambda_s = 3$ (a moderate sized effect), recombination fraction 0.05 and 50% of families linked, f_1^2 =0.464. With $\lambda_s = 1.35$ (one of the effect sizes considered in [134]), f_1^2 =0.494.

Applying the approximation $f_1^2 = f_0^2 = 0.5$, the number of sib pairs needed under homogeneity is

$$n_{T_m}^{homog} = \left(\frac{z_{(1-\beta)}\sqrt{0.5} + z_{(1-\alpha)}\sqrt{0.5}}{\mu_1 - \mu_0}\right)^2 = \left(\frac{z_{(1-\beta)}\sqrt{0.5} + z_{(1-\alpha)}\sqrt{0.5}}{1 - d_1 + 2d_2 - 1}\right)^2$$
$$= \left(\frac{z_{(1-\beta)}\sqrt{0.5} + z_{(1-\alpha)}\sqrt{0.5}}{2d_2 - d_1}\right)^2$$

ASP mean test power with heterogeneity

Power with heterogeneity

Under heterogeneity the alternative hypothesis the test statistic mean is given by

$$\mu_{1_{het}} = p\mu_1 + (1-p)\mu_0$$

where p is the proportion of pairs segregating the mutation of interest. Inserting the terms $f_1^2 \simeq f_0^2$ =0.5 into equation 6.3 gives, to a good approximation, the number of sib pairs needed as

$$n_{T_m}^{hetero} = \left(\frac{z_{(1-\beta)}\sqrt{0.5} + z_{(1-\alpha)}\sqrt{0.5}}{p\mu_1 + (1-p)\mu_0 - \mu_0}\right)^2 = \left(\frac{z_{(1-\beta)}\sqrt{0.5} + z_{(1-\alpha)}\sqrt{0.5}}{p(\mu_1 - \mu_0)}\right)^2$$
$$= \left(\frac{z_{(1-\beta)}\sqrt{0.5} + z_{(1-\alpha)}\sqrt{0.5}}{p(2d_2 - d_1)}\right)^2 \propto n_{T_m}^{homog} \times \frac{1}{p^2} \tag{6.5}$$

This shows there is a simple relationship between heterogeneity and required sample size.

Note that the "non-parametric" ASP mean test has been shown to be exactly equivalent to a "parametric" linkage analysis under a recessive model (see chapter 1 and references [122, 87]).

ASP statistic (2 alleles shared)

The calculations given above are based upon the mean number of alleles IBD at the locus. In the presence of dominance a test statistic based on individuals sharing 2 alleles IBD

will provide greater power. Consider a statistic based on the proportion, p_2 (as above), of ASPs that share 2 alleles IBD. The estimator $\widehat{p}_2$ has binomial distribution with mean $\frac{1}{4}$ and variance $\frac{\frac{1}{4}(1-\frac{1}{4})}{n} = \frac{3}{16n}$. The test statistic,

$$T_2 = \frac{\widehat{p}_2 - \frac{1}{4}}{\frac{3}{16n}},$$

is similar to that described for the proportion of alleles shared IBD in equation 6.1. Again expressing deviations from null with a parameter d_j we have

$$p_2 = 0.25 + d_2 \; (= \mu_1 \text{in this case})$$

if the deviations from the null are true for all families. If there is locus heterogeneity, the null and alternative means are

$$\mu_0 = 0.25 \quad \mu_{1_{het}} = 0.25 + pd_2.$$

Again assuming the variances are similar under the null and alternative the number of ASP needed can be calculated using equation 6.3 as

$$n_{T_2}^{hetero} = \left(\frac{z_{(1-\beta)}\sqrt{\frac{3}{16}} + z_{(1-\alpha)}\sqrt{\frac{3}{16}}}{\mu_{1_{het}} - \mu_0} \right)^2 = \left(\frac{z_{(1-\beta)}\sqrt{\frac{3}{16}} + z_{(1-\alpha)}\sqrt{\frac{3}{16}}}{pd_2} \right)^2 \propto n_{T_2}^{homog} \times \frac{1}{p^2}. \tag{6.6}$$

Using this 2 allele sharing test statistic $n_{T_2}^{hetero}$ is the number of pairs need with heterogeneity and $n_{T_2}^{homog}$ is the number of pairs need with homogeneity. The statistic T_2 provides more power than T_m in the presence of dominance [139] but equation 6.6 shows the similar relationship (n needed $\propto \frac{1}{p^2}$) between power and heterogeneity exists.

6.2.2 Parametric linkage techniques

Computer simulations were performed to evaluate the power of linkage analysis to detect disease loci. In particular, these facilitated the evaluation of a parametric model which allowed for locus heterogeneity in the analysis model.

Allowing for locus heterogeneity

Standard parametric linkage analysis (chapter 1, introduction) can be modified to allow for heterogeneity in the recombination fraction [210]. Assume there are two unlinked disease loci and that a molecular marker linked to one of these loci is typed in the families in the data set. A proportion, a, of the families are assumed to be affected by disease because of mutations at a single (primary) locus. The parameter a is traditionally written as α but a is used here to avoid a clash with the symbol for significance level in equations such as 6.6. In these families the recombination fraction between this disease locus and

the molecular marker is less than 0.5. The recombination fraction in the other $(1 - a)$ families is assumed to be 0.5 (these other families are assumed to be affected as a result of other loci in the genome).

Assume there are n families and that the i^{th} family has likelihood function $L_i(\theta)$. For example if the family was the same as that in figure 1.1 in the introductory chapter, then the likelihood for that family would be as in equation 1.1. The likelihood of each family can be rewritten with a used to index the families linked to the locus of interest

$$L_i(\theta, a) = aL_i(\theta) + (1 - a)L_i(0.5).$$

The likelihood of all families in the data set together is therefore

$$L(\theta, a) = \sum_{i=1}^{n} L_i(\theta, a). \tag{6.7}$$

This likelihood (equation 6.7) is maximised over both θ and a simultaneously. Call the likelihood with both parameters unrestricted L1 and the likelihood with either $a = 0$ or $\theta = 0.5$ L0 (either condition is sufficient for the other to hold). The likelihood ratio test (log_{10} version) of L1 versus L0 is often referred to as the HLOD statistic (as in chapter 1, introduction) . The HLOD statistic does not converge to an asymptotic distribution (since either $a = 0$ or $\theta = 0.5$ specify the null) but it can be approximated (assuming conversion to $2 \times log_e$ scale from log_{10} scale) by a 50:50 mixture of 0 and the larger of two independent χ^2_1 variables [201]. As noted in the introductory chapter (chapter 1), the disease allele frequency and penetrances specified in this model are specific to the disease locus of interest (in the $a \times 100\%$ of the families that are segregating mutations at this locus).

Power of parametric linkage technique with heterogeneity

The power of the parametric linkage analysis technique was assessed using SLINK [244]. SLINK is a program which allows ready generation of families for power calculation. Individuals are given genotypes conditional on specified phenotypes and family structures. The first individual is given a marker genotype and a disease genotype; the parameters relating the genotypes at the two loci and the relationship between disease status and inferred disease genotype are given below. For subsequent individuals in the family the conditional distribution of genotypes given phenotypes is

$$P(\mathbf{g}/\mathbf{x}) = P(g_1/x_1)P(g_2/g_1, x_2)P(g_3/g_1, g_2, x_3)... \tag{6.8}$$

where the set of s phenotypes in a pedigree is denoted $\mathbf{x} = (x_1, .., x_s)$ and the set of genotypes is denoted $\mathbf{g} = (g_1, .., g_s)$.

The relationship between genotype and phenotype was specified to follow a specific

parametric model. For the simulation the marker locus was assumed to be linked with recombination fraction 0.05, to a disease locus (disease allele D, wild type d) with a dominant-like effect on the phenotype. There were five equally frequent alleles at the marker and the disease locus parameters were $P(\text{disease}/DD) = P(\text{disease}/Dd) = 0.5$ and $P(\text{no disease}/dd) = 0.01$. The disease allele was assumed to be moderately rare (frequency 0.004). The conditional probabilities of all the possible multi-locus genotypes with phase, $P(g_1/\mathbf{x})$, were calculated for the first individual. Based on these probabilities, marker genotypes were randomly generated. Subsequent individuals were then given marker genotypes in turn (conditional on the genotypes allocated for preceding individuals via equation 6.8).

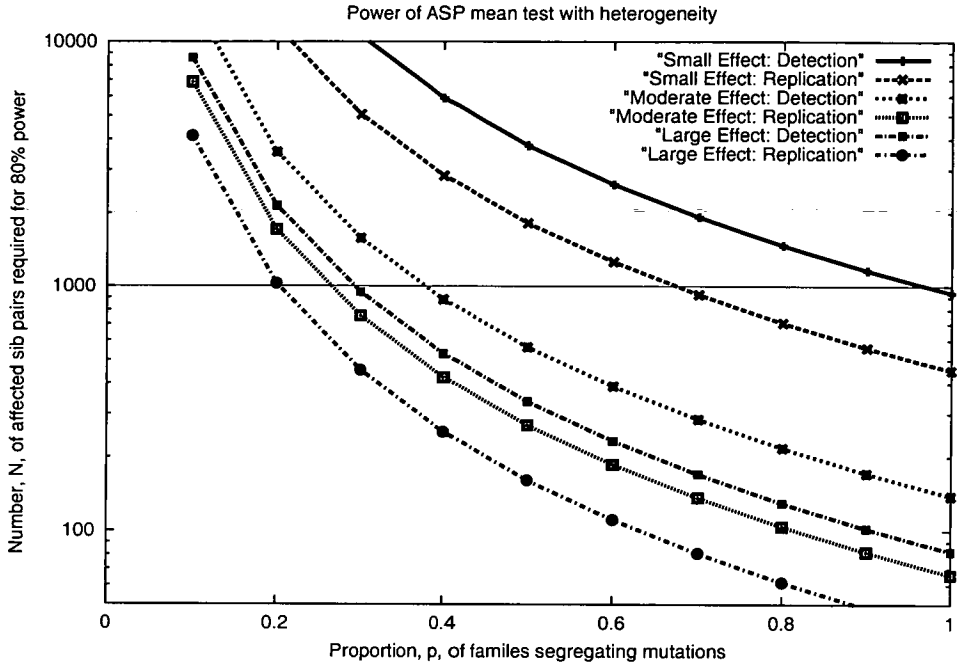
The generated data set consisted of sets of 60 nuclear families. Each family had 6 genotyped affected siblings with untyped, disease status unknown parents. With 6 affected individuals in the sibship there are $\begin{pmatrix} 6 \\ 2 \end{pmatrix} = \frac{6!5!}{2!1!} = 15$ possible pairings per family and 900 in the whole sample of 60 families. The siblings within a family were assumed to be independent. Although this was clearly not true, for simplicity it was assumed that the pairs were independent. It has been shown ([17] chapter 17, [139]) that this should be a reasonable approximation in practice. If all sibs in a sibship are considered together in a likelihood approach then the false positive rates may increase ([17] chapter 17) but proposed weighting schemes (e.g. a weighting of $\frac{2}{n}$ for sib-ships of size n, [226]) to account for this have been found to over-correct for this non-independence ([17] chapter 17). To emulate the effects of heterogeneity, varying proportions of families were assumed to be segregating mutations at the locus of interest. The proportions considered were 75%, 50% and 33%. The data were analysed using the parametric linkage analysis technique described in the section above, with the parameter a fitted to model locus heterogeneity. The proportion of replicates achieving a HLOD of 3 or more were counted over 200 replicates. The analysis model (penetrance parameters and allele frequencies) used in the analysis was the same as that used to generate the families. This should mean that the power of the parametric analysis will be maximised. In practice, a small number of models (e.g. recessive, dominant, see chapter 1, introduction) need to be fitted in a parametric analysis to ensure near optimal power. [17]

6.3 Results

6.3.1 ASP mean test

The number of sib pairs required to detect linkage was shown to be approximately proportional to $\frac{1}{p^2}$, where p is the proportion of pairs segregating the mutation of interest. Figure 6.1 shows the effect of heterogeneity on the power to detect linkage given the effect of an allele, segregating in the linked families, increasing risk to sibs by a given factor (i.e. $\lambda_{slinked}$). The graph shows the results for the ASP mean test. Notice the log_{10} scaling of the y-axis. Three effect sizes are shown, representing a small (factor 1.35), moderate

Figure 6.1: Power of ASP mean test at different heterogeneity levels



(factor 3) and large (factor 7) effect. "Detection" on the graph refers to a significance level assuming a LOD score of 3 is required (asymptotic p-value 0.0001) and "Replication" on the graph refers to a significance level assuming a LOD score of 1.2 (asymptotic p-value 0.01). Historically a LOD of 3 was deemed suitable for a declaration of linkage [201]. Lander and Kruglyak [129] suggest a more stringent criteria of 3.6. Given strong prior evidence of linkage and a very small number of markers typed around the region of interest a nominal p-value (p<0.01, LOD 1.2) may be sufficient for replication of a previous result [129].

Figure 6.1 uses the exact value of f_1^2 (calculated taking into account the change in f_1 for different levels of heterogeneity, effect size and recombination fraction; i.e. using equations 6.3 and 6.4) not the approximation used to show $n \propto \frac{1}{p^2}$.

Figure 6.1 shows that a sample of less than 1000 ASP, as studied in [134], has little power to replicate linkage of schizophrenia to a locus that contributes to risk of illness in less than 20% of families. Even in cases in which there are large increases in risk in some of the families, the value of λ_s over all families may be very low. For example, when the $\lambda_{slinked}$ value is 7 in 20% of families and the remaining 80% of families are linked to other loci the value of λ_s over the whole sample is 1.21. An example in breast cancer illustrates this point well. Genes such as BRCA1/2 have a large effect on risk (10-20 fold) in mutation carriers [102] but, since they are very rare in most populations, they will not be readily detectable in large heterogeneous samples.

Table 6.1: Power to attain a HLOD of 3 with a parametric linkage analysis

Proportion of families segregating mutations	Power
75%	84%
50%	29%
33%	7%

6.3.2 Parametric linkage with heterogeneity

The power to detect a HLOD of 3 decreased rapidly as the proportion of families segregating the relevant mutation decreased. The power for different proportions of families with mutations segregating at the gene of interest are given in table 6.1. These results show that even when the effects of heterogeneity are explicitly modelled (using a within the parametric framework), the power to detect linkage to the locus of interest decreases dramatically as the proportion of linked families decreases.

The parametric linkage analysis results are slightly different to those reported in [143] due to minor changes in the generating model and the number of replicates done.

6.4 Discussion

The results from both the "non-parametric" ASP mean test and the "parametric" linkage analysis of simulated data have shown that locus heterogeneity has a substantial effect on the power to detect linkage. Although the ASP mean test is only one of a number of possible tests (see also the discussion on the variety of possible "non-parametric" statistics in section 1.3.2 in the introduction) suitable for linkage assessment in ASPs it does allow simple assessment of power. The useful relation 'number of ASPs needed is inversely proportional to the heterogeneity proportion squared':

$$n \text{ needed} \propto \frac{1}{p^2}$$

allows one to appreciate the large drop in power when there is heterogeneity. Non-parametric methods such as the ASP mean test do not allow for locus heterogeneity ([17], chapter 17) and may hence be sub-optimal when heterogeneity is present. Nonetheless, the simulation based analysis using SLINK showed that the power of linkage under heterogeneity is still poor even when heterogeneity is allowed for in the analysis. It is therefore crucial to collect samples which are unlikely to have high levels of heterogeneity.

6.4.1 Locus heterogeneity

Locus heterogeneity is an extremely important issue in study design for some relatively well understood Mendelian diseases. A case in point is non-syndromic deafness. To date

over 60 loci have been reported for non-syndromic deafness [171]. For this Mendelian condition the inheritance model is a strict locus heterogeneity model. That is to say, in affected families the inheritance is entirely due to the effects of a single mutation at one of the many distinct disease loci. New loci have been identified with some regularity in recent years with researchers relying on large families: for this form of deafness a study design based on large number of small families would be highly unlikely to identify any disease loci.

The situation in schizophrenia is different to that seen in non-syndromic deafness. It seems highly likely that there are multiple susceptibility loci [188, 137, 92, 34, 242] but the strict locus heterogeneity model observed in non-syndromic deafness is not observed in schizophrenia. Although some schizophrenia families show near Mendelian inheritance patterns (e.g. the Scottish translocation family described in [216]), such families are relatively unusual. Risch [188] considered the decline in λ_R values (where λ_R is the analogue of λ_s for relative pair R) for different pairings of relatives (monozygotic twins, dyzygotic twins, sibs, parent-offspring, grandparent-grandchild, cousins) and showed that it may be possible to use data on a variety of λ_R values to make inferences about the number of loci and their interactions. If there is one locus, affecting the whole population, $\lambda_R - 1$ should half as the degree of the relative decreased. For example, assuming no dominance, $\lambda_s - 1$ should equal $2(\lambda_h - 1)$ when there is a single disease locus, where λ_h is the half sib risk ratio. When there is strict locus heterogeneity (as in non-syndromic deafness), the decline in $\lambda_R - 1$ was shown to be the same as in the single locus case. By contrast, when the loci interacted multiplicatively to determine risk of disease, the decline in λ_R was shown to be sharper than in the single locus or strict heterogeneity case. Risch [188] showed that since the decline of $\lambda_R - 1$ in schizophrenia was greater than twofold with decreasing degree of relationship, there were likely to be a number of susceptibility loci. The results from various schizophrenia genome scans have given support to this assertion [137, 92, 34, 242]. Further, Risch [188] found that the data best fitted a model in which all of the contributory loci had a relatively small effect (λ_s definitely less than 3 and probably less than 2 for all loci). This means that linkage strategies, applied to large diverse samples, will require large sample sizes to detect these loci with small $\lambda_s$values. However, if the families can either be selected so that they are more genetically homogeneous than the general population or selected to segregate alleles which significantly increase risk in particular (extended) families, the effects of individual loci may be easier to detect. The results in section 6.3.1 show that the increase in risk due to such rare alleles in the ascertained families may be substantial and that this may be equivalent to a small λ_s value over a diverse set of families (i.e. $\lambda_{slinked}$ may be large but if the proportion of linked families is low then λ_s will be low).

One cautionary note. Whilst it is impossible to know at this stage what distribution of genetic effects affecting schizophrenia will be it is important not to over-emphasise the significance of the results of linkage studies. A number of studies have suggested that their results show there is locus heterogeneity (based on linkage results) [223, 92, 132, 225,

104]. Whilst this may be true (and selecting large families may well be the best strategy in such cases), when the power to detect linkage is low, the effects detected in any given sample will be dictated largely by chance. For example, say there are several loci of equal additive effect on the phenotype (say for arguments sake pushing up an underlying trait toward a threshold of affection), then by chance only a few of these would show up in the linkage scan and it would not be unusual for each positive linkage to occur in only a subset of the families. Not knowing the true disease model it is tempting to suggest (wrongly in this hypothetical case) that such data is indicative of locus heterogeneity, when in fact no definite conclusions can be drawn.

6.4.2 Particular aspects of Levinson et al. [134] analysis

As indicated at the start of this chapter, this chapter was based on a correspondence discussing a paper reporting a meta-analysis of schizophrenia studies (Levinson et al. [134]). The Levinson et al. [134] paper focused solely on the results from an analysis of markers on chromosome 1q, concluding there was 'no major schizophrenia locus' on 1q. I now discuss the various issues arising from this correspondence.

In contrast to the GSMA used in [137], Levinson et al. [134] analysed the raw data from the 8 constituent studies. They analysed the raw data using parametric linkage under the recessive model which generated the maximum LOD in the paper reporting linkage to 1q21 [34]. However, as discussed in the introduction (chapter 1), the parametric method only has power for complex disease mapping when a small number of different analysis models are considered. In particular, at least a dominant and a recessive model should be fitted to the data set to ensure good power.

Levinson et al. [134] apply the ASP based "non-parametric" test described in Risch et al. [189]. To account for locus heterogeneity they apply a modification of the ASP based test [132]. This uses logistic regression to allow for interstudy heterogeneity. In the regression the logit of the probability of pairs sharing alleles IBD is the dependent variable with indicator variables for the different studies entered as the dependent variable. However, this only allows for heterogeneity between studies and will not be sensitive to heterogeneity within a particular study. In contrast, the parametric analysis using heterogeneity, described in section 6.2.2, allows individual families to segregate different mutations within a particular study. As indicated previously, a number of studies [34, 92, 35, 31] have reported multiple strong linkages within a single study/population. This means it may well be appropriate to allow for heterogeneity within studies as well as between studies.

Results from meta-analyses

The eight studies included in the Levinson et al. [134] study were unusual in some respects. They were almost exclusively based upon constituent studies that gathered ASP data sets. The results of these studies individually were, with the exception of the Blouin et al. paper [31], characterised by the lack of positive findings. In contrast, a more re-

cent meta-analysis paper [137], incorporated a wider range of data sets. In particular, in addition to the ASP data sets included in [134], the more recent paper [137] included the studies by Brzustowicz et al. [34], Gurling et al. [92], Lindholm et al. [138] and DeLisi et al. [52] each of these additional papers found substantial evidence for linkage. The results from the more recent meta-analysis [137] indicated a number of regions of interest for further study: these included chromosome 1q. The addition of the further samples, comprising mainly extended families, goes some way toward explaining the success of the Lewis et al. [137] paper and the inability of the Levinson et al. [134] study to detect the 1q locus. It is still slightly surprising however, that chromosome 1q region, together with a few other regions initially identified in extended family samples (chromosomes 2q, 6q, 11q, [137]), generated a significant result in the Lewis et al. meta-analysis. The GSMA method weighted the studies by the root of the number of affected individuals; this would give greater weight to the studies composed mainly of ASP since these had larger numbers of affected individuals (although less affected individuals per family, typically 2.4 affecteds, as above). Lewis et al. [137] suggested that this meant that these regions (1q, 2q, 6q and 11q) may be of significance to the populations outside of the population in which the putative disease locus was identified.

Since this thesis focuses primarily on linkage studies, this chapter only considers study design for linkage analyses. The issue of study design for LD mapping (association studies) is of substantial importance to future human genetic studies. In chapters 1 and 8 some of the factors determining the success or failure of whole genome association studies are discussed. There is now a substantial body of literature discussing LD study design and I will not attempt to summarise them in any detail here. The two most important considerations are that (i) the disease causing alleles should be common in the population under study (e.g. isolates in which there are risk alleles descended from one or a few founding individuals) [169, 126, 258, 184] and (ii) the increase in risk of disease as a result of individuals bearing a risk allele are not too small [245, 259, 39].

6.4.3 Summary

It has been shown here that, with locus heterogeneity, linkage studies of diseases such as schizophrenia will require extremely large samples. Study designs based on extended families are likely to reduce the degree of heterogeneity encountered, increasing the chances of the study detecting a single locus. Definite identification of even a single schizophrenia susceptibility locus may be of substantial significance, perhaps leading to greater understanding of disease pathogenesis (see also introduction and [96]). Although there may be genes that affect most or all human populations, such genes will have a very low effect on risk in individual families and they will require unfeasibly large samples to detect.

Chapter 7

False disease region identification in the presence of phenocopies

In an attempt to locate disease genes many researchers have applied linkage analysis to identify chromosomal regions which segregate with the disease of interest in a pedigree. In particular, regions unbroken by recombination in affected individuals are sought out. For Mendelian disorders there is usually a single disease region which is completely associated with the disease phenotype. In complex disorders, there are typically multiple disease regions (or multiple ancestral haplotypes), some of which may be the result of mutations at distinct (unlinked) positions along the genome. Such regions may only be partially associated with the disease phenotype (region is neither necessary or sufficient for disease). Complex diseases are commonly modelled as if they were Mendelian, with individuals carrying a disease mutation but not exhibiting the disease phenotype labelled as non-penetrant and affected non-disease mutation carriers labelled as phenocopies. The focus of this report is these *phenocopies*. The phenocopy rate is defined as P(individual in sample is affected/individual does not carry disease mutation of interest) whilst the non-penetrance rate is P(individual in sample is not affected/individual does carry the disease mutation of interest). Many complex diseases are caused by multiple unlinked loci, however, *all individuals not carrying the mutation at the region of primary interest must be regarded as phenocopies*. This will be particularly important for disease in which there is a least some degree of genetic (locus) heterogeneity. Individuals may also be phenocopies if they do not carry the disease gene of interest and are affected as a result of environmental factors. Since phenocopies either have the disease as a result of non-genetic factors or because of other mutations at unlinked chromosomal regions, the set of alleles they have at the putative disease region will be different to that of the other affected individuals.

The set of alleles an individual holds along a chromosome can be referred to as a haplo-

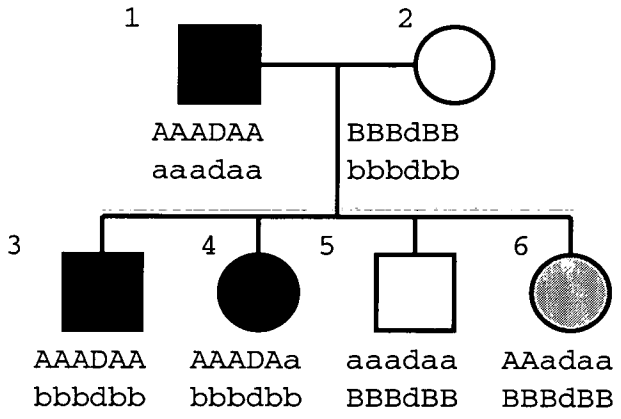
type (chapter 1). Although the word haplotype can be used to describe the set of genotypes with phase passed from parent to offspring (i.e. in the context of linkage analysis), it is also used to refer to ancestral haplotypes. These ancestral haplotypes will have been subject to many generations of recombinations and will be smaller than the haplotypes observed to have been narrowed by recent recombinations in genotyped sets of individuals of known relationship. Since the rest of this chapter is not concerned with ancestral haplotypes, to avoid confusion, the term haplotype is not used.

In this chapter the effect of phenocopies upon disease region identification is considered and it is shown that the regions inferred in the presence of phenocopies may not include the true disease locus. Such errors will impact significantly on subsequent attempts to identify disease causing mutations. The effects of a variety of phenocopy rates are investigated. Calculations are done under the assumption that nuclear families are analysed. However, it is shown that the effect of phenocopies will be similar in larger family structures. To facilitate the evaluation of the effect of phenocopies, the distribution of the disease region length is calculated, both theoretically and empirically. Once the length of the disease region indicated by the affecteds (who carry a mutation at the locus of interest) has been calculated, it will then be possible to determine how likely phenocopies are to interfere with this region. Dominant gene action is assumed but extensions to recessive types are discussed. The effect upon the LOD score profile from a parametric linkage analysis is also considered.

7.1 Phenocopies and disease regions

To identify a disease region using affected individuals in families one uses the marker information to assess where recombination events have occurred. When one or more phenocopies arise within a sample, the recombination events of these phenocopies are erroneously used to narrow the disease region. Consider the nuclear family in figure 7.1. Affected individuals are shaded in black, phenocopies are shaded in grey and unaffected individuals are unshaded. Assume the disease mutation in this region is dominant with alleles D and d. In this family a recombination event in affected individual 4 is used to narrow the disease region on the right of the true disease locus. Call the disease region inferred from the affected individuals carrying the D allele minimal disease region or MDR. Suppose one of the individuals, numbered 6, is a phenocopy. This individual does not carry the disease allele, D, but does carry some of the marker alleles of its affected parent via a recombination. This means that the genomic region shared by all of the affecteds (D allele carriers and phenocopies) spans only the leftmost two markers and does not include the actual disease locus of primary interest. If the phenocopy rate is not low, or there are relatively few affected individuals (with mutations at the locus of interest) available for study, the probability of this happening will be non-negligible. Note that although some phenocopies will occur in families otherwise unaffected by the disease (sporadic cases), linkage analysis samples are typically ascertained to have a large number of affected individuals.

Figure 7.1: Nuclear Family with a phenocopy



The phenocopies ascertained and analysed are therefore likely to occur in circumstances similar to that of individual 6 in figure 7.1.

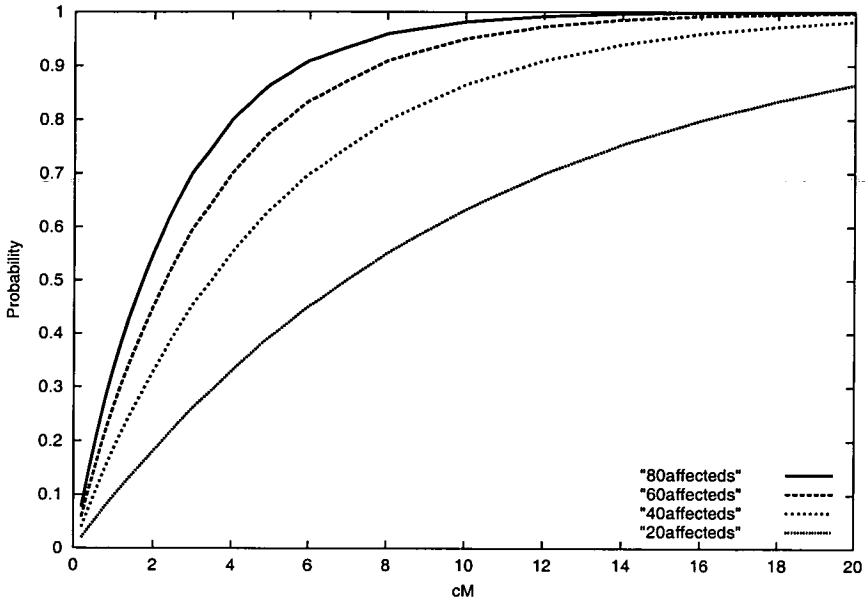
The population prevalence of Psychiatric diseases such as bipolar disorder and schizophrenia rarely exceeds a few percent. Published estimates for the proportion of phenocopies amongst affecteds in the population are high for diseases such as bipolar disorder I (0.67, [149]) and broad definition schizophrenia (0.25, [35]). However, estimating phenocopy rates that are appropriate for the families ascertained for linkage studies is difficult. Firstly, the rate of affection amongst such samples may reach high levels (for example, the proportion of individuals affected was 38% in the data set described in chapter 5). Secondly, ascertained families may be more genetically homogeneous than the general population, with some related individuals affected, at least in part, as a result of the effects of a common genetic background. Below we consider the rate at which phenocopies occur within a sample of non mutation carriers, to be from 0.01 to 0.08.

To assess the probability of phenocopies making the MDR too small one must assess the distribution of disease region lengths and the likely number of phenocopies.

7.2 Distribution of disease region lengths

The length of the MDR can be calculated by considering the distribution of the recombination events. If one ignores linkage interference, the number of recombination events follows a Poisson distribution with parameter equal to 1 per Morgan of genome per meiosis. Consider a putative disease locus on a chromosome. The map distance to the first recombination event to the right is distributed exponentially with parameter 1 Morgan. Given a number of inherited chromosomes n, the distance from the putative disease locus

Figure 7.2: Distribution of region lengths



to the nearest recombination on the right is then distributed as exponential with parameter $\frac{1}{n}$. The distribution of the distance between the first recombination to the left and the first to the right is thus the sum of 2 exponential distributions. This has a gamma distribution with alpha equal to 1 and beta equal to $\frac{2}{n}$. The distribution of region lengths for 20, 40, 60 and 80 affecteds is given in Figure 7.2. The mean region lengths in the four cases are 10cM, 5cM,3.3cM and 2.5cM, respectively.

7.2.1 Quantifying the Effect of Phenocopies

We assume that phenocopies who do not share any of the MDR are removed from the sample. This will usually happen in practice since otherwise it will not be possible to identify a disease region at all. In a nuclear family the probability of a phenocopy interfering with the MDR depends on the average length of the MDR and the probability distribution of the number of phenocopies.

If the likely number of phenocopies is small then one can estimate the probability of at least one phenocopy having a recombination in the MDR (and hence carrying part of it, but not the disease locus of interest) by

$$1 - (1 - L)^w$$

where w is the number of phenocopies and L is the length of the MDR measured in Morgans (assume for simplicity the Morgan map function in which recombination fraction equals map distance).

Assume that a number of pedigreed individuals are ascertained for the analysis and that m of these do not carry the mutation of interest (these will commonly be unaffected individuals; if there is only 1 mutation causing the disease and no environmental factors generating phenocopies then these individuals will definitely be unaffected). If each of these m individuals has probability p of being a phenocopy, the number of phenocopies in the sample will have a binomial distribution with parameters m and p. The probability of at least one phenocopy making the MDR too small (and ruling out the true disease locus) is therefore

$$\sum_{r=1}^{m} \binom{m}{r} p^r (1-p)^{r-m} (1-(1-L)^r) \tag{7.1}$$

This formula will not hold exactly when there exist two or more phenocopies in a sample since in such a case it may be possible for two of the phenocopies to have recombinations in the MDR on opposite sides of the true disease locus. This means that both will agree with parts of the MDR but together they will rule out the whole of the MDR. The probability of this happening is low unless there are a large number of phenocopies in the sample.

The results obtained using equation 7.1 are similar to those obtained using the exact formula (i.e. accounting correctly for multiple phenocopies) in most cases; the exact formula is given in the appendix (section 7.6). All of the results given here use the exact formula.

7.2.2 The effect of varying phenocopy rate and sample size

In table 7.1 the effects of changing the phenocopy rate are shown. The probabilities in the table assume 20 affected (affected assuming the mutation is fully penetrant in its effect on the phenotype) mutation carriers have been included in the analysis. It is assumed that 100 individuals not carrying the mutation have been considered alongside the affected individuals. The proportion of regions which falsely rule out the genomic region where the locus actually resides reaches worryingly high levels (>40%) if the phenocopy rate exceeds a few percent.

In table 7.2 the effects of altering the number of mutation carrying individuals are shown. A phenocopy rate of 0.02 is assumed and again there are assumed to be 100 individuals not segregating the mutation. The proportion of regions which include the actual disease locus is high provided the sample of affected individuals is not too small.

7.3 Computer Simulation

Two sets of computer simulations were performed. The first confirmed the theory above concerning the distribution of MDR lengths after a given number of recombinations had eroded the disease region. The second looked at affected 20 individuals in 5 families and

Table 7.1: Effect of varying phenocopy rate

Phenocopy rate	Probability region does include actual disease locus	Probability region does not include actual disease locus
0.01	0.91	0.09
0.02	0.83	0.17
0.03	0.76	0.24
0.05	0.66	0.35
0.08	0.56	0.44

Table 7.2: Effect of varying number of affected (mutation carrying) individuals

No. of Affecteds	Probability region does include actual disease locus	Probability region does not include actual disease locus
10	0.70	0.30
20	0.83	0.17
30	0.88	0.12
50	0.92	0.08
100	0.96	0.04

assessed the effects of a phenocopy upon the LOD score profile in multipoint linkage analysis.

7.3.1 Simulation 1: Distribution of MDR lengths

A program was written to allow the transmission of gametes, with recombination, from parents to offspring. A single founder parent with marker genotypes at 24 chromosomal positions was generated. This parent was set to be heterozygote for a disease allele at the 12^{th} locus and was mated to individuals set to be homozygote for a wild type allele at this locus. The marker genotypes at all other loci were randomly generated from a pool of 40 equally frequent alleles. Based on the generated genotypes for the parents, sets of either 10, 20 or 40 offspring were generated. All 24 markers were linked, with the recombination fraction between adjacent loci set to be either 0.01 (40 offspring) , 0.02 (20 offspring) or 0.03 (10 offspring). The length of preserved region in each child was counted by starting at position 12 and counting outward (in both directions) until the marker alleles observed in the child differed from that seen in the parent (indicating that a recombination event had occurred). 1000 replicates were used for 10, 20 and 40 offspring. The results obtained are given in table 7.3 and are in good agreement with those derived theoretically above.

7.3.2 Simulation 2: Effect of phenocopies on LOD profile

In this simulation 5 nuclear families, each with 4 affecteds, were generated. Chromosomes with 24 highly polymorphic, 2cM spaced markers were passed from parents to offspring.

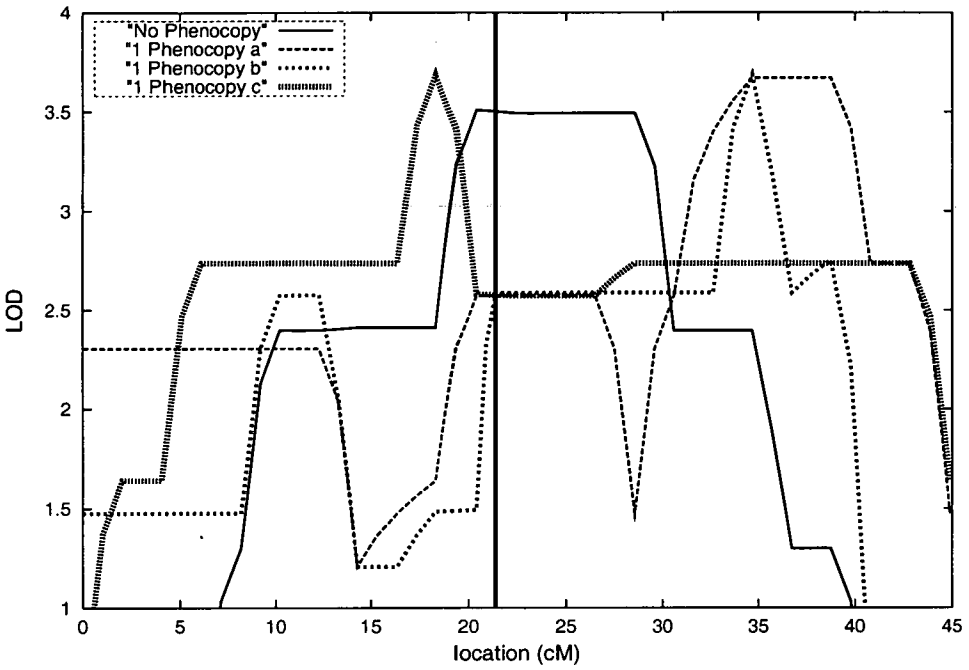
Table 7.3: Simulation 1 results

	Average Length of region (Morgans)	
Number of affecteds	Theory	Simulation
10	0.200	0.183
20	0.100	0.105
40	0.050	0.048

A disease locus with a fully dominant disease allele was placed midway between markers 11 and 12 (21.4 cM). LOD score profiles from multipoint parametric linkage analyses were calculated using the program Allegro ([90]). As expected, the MDR in each case was visible as a plateau (region in which no recombinations occurred in the genotyped individuals, on average 10cM long) in the LOD profile.

To assess the impact of phenocopies, a phenocopy was added to one family and the LOD profile re-calculated. Assuming phenocopies to be binomially distributed 1 phenocopy would arise in this way 37% of the time if 100 'unaffecteds' were ascertained with a phenocopy rate of 0.01. As predicted by the above theory, approximately 10% of these phenocopy individuals shared some of the MDR (based on 300 replicates). Figure 7.3 shows the LOD profiles of three replicates (broken lines) where the MDR was falsely narrowed by recombination(s) in the added phenocopy. For comparison, a replicate in which the added phenocopy had no recombinations (solid line, labelled "no phenocopy") in the MDR is also shown in figure 7.3. When the phenocopy has recombination(s) in the MDR there is a region shared by 21 affected individuals, generating a LOD around 3.6. Conversely, when there are no recombinations around the disease locus in the phenocopy, there are 20 individuals with a common set of alleles and one without this set of alleles. This typically generated a LOD of around 2.8. The addition of a phenocopy increases the maximum LOD score achieved but, crucially, indicates a genomic region which does not include the true location of the disease locus (since the phenocopy cannot actually share the genomic region with the disease gene on it, only a nearby region via a recombination in the affected parent). The discrepancy in location was up to 20cM. Allowing for phenocopies in the analysis (through the phenocopy rate parameter specified in a parametric analysis) *does not improve the situation* since the LOD peak is still at the point where most individuals share the same set of alleles. The only effect of setting the phenocopy rate parameter to 0.1 or 0.2 is to reduce the overall LOD scores achieved; the percentage of replicates with the peak LOD distinct from true disease locus (i.e. there was a region over which all individuals, including the phenocopies shared alleles, and this region did not include the position of the true disease locus) was 11% for both the 0.1 and 0.2 analysis (as before, predicted by theory to be 10%). 300 replicates were generated in each case.

Figure 7.3: Four Simulation replicates



7.4 Extensions from dominant nuclear families

The results in tables 7.1 and 7.2 were obtained by assuming that the phenocopies appeared in nuclear families in which there was dominant disease inheritance. However, similar problems will often arise when larger families and recessive types inheritance patterns are considered. The extension of the above argument to cases other than nuclear families is possible because of two factors:

First Issue Families are generally only ascertained if they have a number of affected individuals. The presence of a single affected individual (perhaps affected as a result of an environmental influence) is unlikely to be enough for researchers to conduct further investigations. More likely, a phenocopy will be included in a disease mapping study alongside a number of other affected individuals (whose affection status is at least in part due to them possessing a particular gene). This will mean that the families used will be relatively densely affected and a number of affecteds will likely have some chromosomal regions in common.

Second Issue As mentioned above, it is common for investigators to remove individuals whose set of alleles are completely distinct from that of the other affecteds.

It is argued in this section that because of the ascertainment procedure and the discarding of incongruous phenocopies (issues 1 and 2), nuclear families with phenocopies

often provide a good approximation to the situation where more general extended pedigrees are analysed.

7.4.1 Extension to larger families (dominant inheritance)

Consider extending a nuclear family through the offspring. There are 3 ways in which the grandchildren of the original founders can be phenocopies. Firstly, these grandchildren may be the offspring of an affected parent and be phenocopies (figure 7.4, case 1). In this case they may still inherit a section of chromosome near to the disease locus via recombination (this is the same situation as in the original nuclear family: unaffecteds and phenocopies can inherit regions of the genome near the disease locus by recombination). Secondly, there may be phenocopies who are the children of an unaffected individual (figure 7.4, case 2). This unaffected individual may possess regions of the genome near the disease locus (via recombination) and will pass this chromosome on to its offspring 50% of the time (any further recombination in the meioses forming the phenocopy will still result in the phenocopy getting at least some of the alleles near the disease locus). The other 50% of the time individual 4 in the pedigree will pass on an chromosomal region unrelated to any of the other affecteds. In this case the phenocopy will often be removed from the group of affecteds since it does not share the MDR with them (issue 2). Thirdly, the grandchildren may be the children of a phenocopy (figure 7.4, case 3). This case is the same as the case where the grand-children's parent is unaffected but, unless the phenocopy rate is rather high, it is unlikely that two such phenocopies will occur.

The family may be further extended to consider great-grandchildren. However, if a branch of the pedigree stems from a second generation individual who is unaffected and who has no affected offspring then the fourth generation is less likely to have been considered for inclusion in the study. In the unlikely event of one being included it will often be excluded because of issue 2, above. Branches with many affected individuals are much more likely to be included (issue 1). Any phenocopies arising in such a branch will hence share much of their genome with the true affecteds. In summary, in many cases the problems caused by phenocopies in nuclear families will also be present when extended families are considered.

7.4.2 Extension to recessive cases

If a disease gene that is recessive in its effect upon the disease is considered, a MDR can be identified where affecteds share two copies of a particular disease segment of the chromosome. In the case of a nuclear family, phenocopy offspring will cause problems similar to those in the dominant case (any recombinations in the transmission of alleles from unaffected carrier parents may falsely narrow the MDR if there are phenocopy offspring). Phenocopy parents will rule out part of the MDR (including the disease locus) obtained from the other affecteds (since they have at most 1 copy of the disease gene) but, in practice they will usually be removed from the group of affecteds (issue 2).

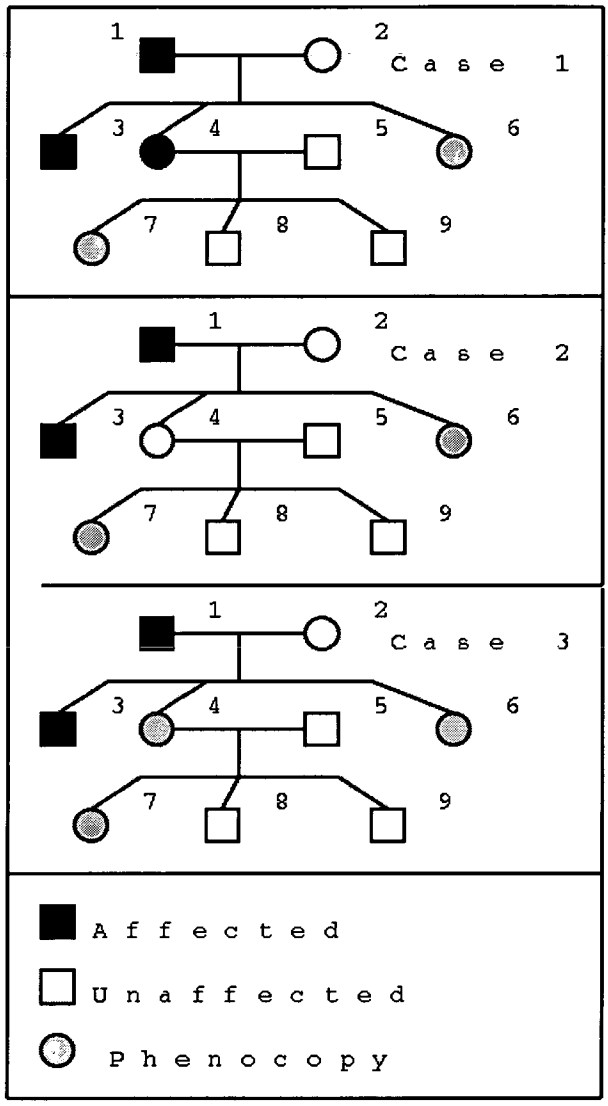


Figure 7.4: Extensions of nuclear families: dominant case

Unlike the dominant case, recessive type families are less likely to extend beyond the offspring generation. Clearly, in the dominant case, the disease will often be transmitted over multiple generations. In the recessive case, a new disease allele must be introduced for the disease to be transferred over more than one generation. Therefore, in the absence of high levels of inbreeding, recessive type extended families will be rarer than nuclear families.

7.5 Discussion

The phenocopy rate for a given disease is often difficult to assess but the above results show that in many cases phenocopies can adversely affect the size of region derived from a linkage or allele sharing analysis. Complex disease phenotypes are affected by both environmental effects and additional genetic effects such as unlinked loci (genetic heterogeneity, epistasis). These results show how to calculate the expected region length as a result of recombination events in a sample of mutation carriers. However, unless it is possible to be sure that all the affected individuals carry a mutation at the disease locus of primary interest (i.e. no phenocopies) researchers should not conclude that the disease locus is in the region obtained.

Parametric linkage techniques are fairly robust to mis-specification of parameters such as penetrance ([201], see also chapter 1, introduction). However, this only applies to the detection of linkage. The simulations described here show that correctly specifying the phenocopy rate in a parametric analysis will not prevent phenocopies from sometimes interfering with disease region identification.

This work was motivated by our group's attempts to identify disease regions for bipolar disorder. A single large family affected by bipolar disorder and recurrent major depression generated a LOD of 4.8 at a marker on chromosome 4p [27]. Although this gives strong evidence for the relevance of this locus to disease susceptibility it is far from clear whether recurrent major depression and bipolar disorder have the same genetic cause(s). This means the phenocopy rate of relevance to this region is unlikely to be zero. In an attempt to narrow the disease region indicated by the initial linkage on 4p, another three families that also showed linkage to this region of 4p were collected. The region identified by the overlap of regions in the 4 families are shown in figure 7.5 (Kathy Evans, University of Edinburgh, `http://www.genetics.med.ed.ac.uk/psygen/4p/`). There was some overlap between the disease regions identified in the families but there was no single region implicated by all four families. This implies that some of the affected individuals considered in these families segregated mutations at loci other than the one of interest here on chromosome 4p. That is, some of the affected individuals were in fact phenocopies (with respect to the 4p locus).

Other researchers have encountered similar problems in identifying a single disease region in all affected individuals. For example, Angius et al. [12] looked at essential hypertension, considering 35 affected individuals. Hypertension is almost certainly caused

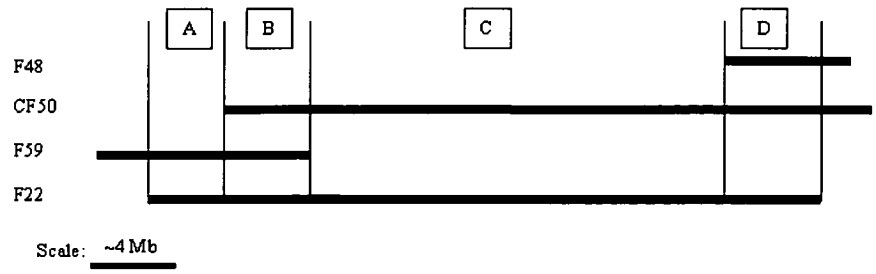


Figure 7.5: 4p regions

by multiple loci [258] and all affected individuals not carrying a mutation at the main locus (2p24) they reported will be phenocopies with respect to this locus. Angius et al. [12] were unable to identify a single set of alleles in this region carried by all the affecteds, indicating the existence of at least one phenocopy. The region they reported may not include the actual disease locus as a result of these phenocopies. In psychiatric disease, [35] performed linkage analyses on a relatively small schizophrenia data set (<50 affecteds) and concluded that, on the basis of traits known to have non-negligible phenocopy rates, the regions of interest for further work were 4.3 and 19.75cM in length. Whilst these are the regions indicated by recombination events in the 'affected' individuals, these regions will not necessarily include the disease locus. Further, investigators performing multiple statistical tests (e.g. fitting a dominant model, a recessive model, a model with broad/narrow disease definition etc.) will normally report the smallest possible 'region of interest' without due regard to the number of tests done.

It may be possible to minimise the number of phenocopies by concentrating on more extreme (more 'genetic') forms of the disease of interest. However, in some cases this may decrease the number of available affecteds unacceptably. For example, in psychiatric disease it is unclear that diseases such as schizophrenia, bipolar disorder and recurrent major depression are genetically distinct (see also chapter 6) and one may acquire a large set of affected individuals if one treats them as a homogeneous group. However, since it has been postulated [24, 26] that there exist some susceptibility loci which affect both and some which do not, it is not clear a priori what effect merging the groups is likely to have on the phenocopy rate. Such 'heterogeneity of phenotype' may cause as many problems as the commonly quoted locus heterogeneity.

It is assumed above that unaffected individuals are not used to help identify the MDR. Including such individuals would decrease the size of the MDR (since there are potentially more recombination events available to reduce the size of the region of interest) and reduce the chance of phenocopies causing problems. However it is rare in the analyses of complex traits for unaffected individuals to be afforded the same significance as affected individuals. For example, unaffecteds may not show disease symptoms because they are still relatively young. For diseases in which it is possible to be sure that unaffecteds are

truly unaffected (perhaps because they are significantly older than the typical age of onset) then it would be advantageous to include them. Most non-parametric and some parametric linkage analysis methods do not include unaffected individuals [127, 189, 201, 2], see also chapter 1.

In the analyses of quantitative traits uncertainty in the position of the trait locus is dealt with by constructing appropriate confidence intervals. However, in the analyses of discrete complex traits some investigators are wont to forget that the affected individuals do not all necessarily carry the mutation of interest, resulting in the reporting of untenably small chromosomal regions. Confidence intervals in discrete trait linkage analyses are considered by [194] but are rarely used in practice. More often researchers ignore the presence of phenocopies and report the smallest region they find.

A more robust MDR may be constructed by removing the effects of certain recombination events. For example, if it is possible to be relatively sure that there exists at most 1 phenocopy in the sample of 'affecteds', a robust MDR can be constructed by removing the two individuals who define the left and rightmost limits of the MDR. If the probability of more than 1 phenocopy is non-negligible then more individuals could be removed, yielding a larger but more robust MDR. Removing such individuals will often only have a small effect upon the size of MDR obtained. In the 20 affecteds example, a phenocopy rate of 0.01 was considered and 100 'unaffecteds' (individuals not carrying the mutation of primary interest) were ascertained. As mentioned in section 7.2.1 (quantifying the effects of...) above, this would mean that the probability distribution of number of phenocopies would be binomial. Therefore the probability of more than 4 affecteds would be small ($Pr(>4\,\text{affecteds}) = \sum_{x=5}^{100} \binom{100}{x} 0.01^x (1-0.01)^{100-x} = 0.0034$). A robust MDR can be construct using this information. The interval would be $\frac{2}{(20-4\times2)} * 100 = 16.7$cM on average and would represent the widest possible disease region, even in the presence of phenocopies. On average this will be 67% larger than the original MDR (10cM) calculated assuming no phenocopies. With more affecteds but a similar number of phenocopies the effect of removing the nearby recombinants would be less severe. For example if there were 40 affecteds, and at most 4 phenocopies the robust MDR would be $\frac{2}{(40-4\times2)} * 100 = 6.25$cM on average. In comparison the original MDR was 5cM. However, if one is less sure of the true phenocopy rate it will not be obvious, a priori, how many individuals to remove to obtain a robust MDR. Clearly, there will come a stage where discarding a large number of affecteds will be counter-productive.

If the available sample is large and/or there are likely to be few phenocopies the disease regions found may be reliable. However, as the number of affected individuals decreases and/or the phenocopy rate rises, investigators must be cognisant of the possibility that the disease region indicated by their linkage result may not necessarily include the true disease locus. Although it would still be recommended to begin fine mapping at the LOD score peak, the results presented here show, particularly in small samples, the LOD score peak may be some way from the true peak when phenocopies are present.

7.6 Appendix

The equation given in section 7.2.1 is not strictly correct. When there are two or more phenocopies in a sample it is possible for more than one to have a recombination in the MDR. These recombinations may indicate different regions of the MDR and hence together rule out the whole region. The probability of this happening can be incorporated into equation 7.1 above. equation 7.1 needs to altered to include 0.5^{n-1} (where n is the number of phenocopies), to ensure that when there are 2 or more phenocopies that their recombination events are on the same side. The full equation for the probability of a phenocopy making the MDR too small is therefore

$$\begin{aligned} Pr(1\text{ phenoc. has rec. in MDR}) + Pr(2\text{ phenocs. have rec. in MDR}) \times 0.5 + \\ Pr(3\text{ phenocs. have rec. in MDR}) \times 0.5^2 + \ldots = \\ \sum_{r=1}^{n} \binom{r}{1} L^1(1-L)^{r-1} Pr(\text{r phenocs. in sample}) + \\ \sum_{r=2}^{n} 0.5 \binom{r}{2} L^2(1-L)^{r-2} Pr(\text{r phenocs. in sample}) + \\ \sum_{r=3}^{n} 0.5^2 \binom{r}{3} L^3(1-L)^{r-3} Pr(\text{r phenocs. in sample}) + \ldots = \\ \sum_{k=1}^{n} \sum_{r=1}^{n} 0.5^{k-1} \binom{r}{k} L^k(1-L)^{r-k} \binom{m}{r} p^r(1-p)^{r-m} \; for\; r \geq k \end{aligned}$$

where m is the number of 'unaffecteds' (individuals not carrying the mutation of primary interest), n is the number of phenocopies, L is the MDR length and p is the phenocopy rate. With 20 affecteds, 100 'unaffecteds' and a phenocopy rate of 0.01 equations 1 and 2 give 0.095 and 0.093 respectively. With a phenocopy rate of 0.05, the difference is more substantial (0.346 cf. 0.393).

Chapter 8

General Discussion

Human genetics has become a very large and active field in recent years and seems set to continue to expand for the foreseeable future. The human genome project, together with other genome projects and related technologies have both driven and been part of this expansion. In terms of one of the major aims, the mapping of genes responsible for disease, only modest progress has been made. However, it seems inevitable that substantial further advances will be made over the next few years. The amount of time, effort and resources invested virtually guarantee success, with the only question being what factors will accelerate this discovery process. In many cases these factors will be based partly in the new technologies recently and continuing to be developed. Efficient (statistical) methodology and study design, together with advances in disciplines such as bioinformatics and relevant clinical practice, will allow these new technologies to be used effectively, ensuring a continuing increase in the understanding of the genetic basis of disease.

At different stages in recent history (since the discovery of the structure of DNA in 1953, say), the emphasis in genetics has changed depending on the differing interests and technologies of the day. This has led to significant advances in fields such as livestock genetics (animal breeding), mouse genetics, fruit-fly (*Drosophila*) genetics and plant (breeding) genetics. Since there are substantial areas of common ground between say human, livestock and ecological (natural non-human population) genetics, success has been and will continue to be accelerated by meaningful cross-talk between the different but related areas of genetical research.

In this chapter the current state of human genetics is reviewed. The impact that new technologies will have on future progress is discussed and predictions are made about the future of human genetic studies.

Successes and Failures Consider first of all some of the successes in gene mapping. Considerable success has been achieved in many rare Mendelian diseases. The genes responsible for diseases such as Huntington's disease [93, 141] and cystic fibrosis (CF) [234, 120] have been identified and lessons can be drawn from these. The most obvious

lesson is that the principal of reverse genetics is sound and approaches such as positional cloning (i.e. linkage followed by linkage disequilibrium) can provide insights, often unavailable otherwise, into the biochemical and development pathways involved. As time has passed it has become clear that even in these simple disorders, the relationship between phenotype and genotype is not necessarily simple. The genetic component of disorders such as non-syndromic deafness [171] and blindness [259] has been shown to be rather complex, with many distinct loci (each individually causing Mendelian or near-Mendelian inheritance) often causing the same end phenotype. Figure 8.1 (from Wright and Hastie [259]) illustrates the diversity of disease loci affecting vision. Even cystic fibrosis (a Mendelian trait where all disease genotype carriers are affected) has been shown to have modifier loci influencing disease development in individuals carrying the same mutation at the major CFTR gene [153]. Although there are a few populations (e.g. Northern Europe; some recent papers have used CF data from Northern Europe as a test-bed for new multiple marker based linkage disequilibrium (LD) techniques [152, 162, 163]) in which a single mutation causes CF, there are in fact hundreds of different allelic variants that cause CF (OMIM 602421). It seems likely that many (genetically) complex diseases will also have substantial allelic (and locus, for that matter) heterogeneity. The genomic region responsible for Huntington's disease (HD) was identified in 1993 [141], some ten years after the initial linkage [93]. From a genetic perspective, discovery of the HD gene, together with a similar result in Fragile-X syndrome (OMIM 309550), was important because it allowed researchers to fully realise the basis for the decrease in the age of onset when the disease was passed from generation to generation (this phenomenon is known as anticipation). Prior to the full dissection of Fragile-X and HD, the anticipation effect was thought to be simply due to ascertainment biases in family collection (i.e. the only families ascertained are those that have particularly severely affected children and relatively mildly affected parents). Anticipation has been reported in complex diseases such as bipolar disorder [84, 240] and schizophrenia [151] and knowledge of the underlying genetic mechanisms in diseases like HD may usefully inform this new research. In Alzheimer's disease (AD), success came as a result of researchers concentrating on individuals with particularly early onset [81, 82, 228]. This success has taught us that definite identification of even a single gene can be of substantial significance; in the case of Alzheimer's this led to greater appreciation of the role that amyloid β peptides (in plaques in the brain) played in the disease [94]. Such insights into disease pathogenesis will ultimately lead to improved diagnosis, therapeutics and pharmaceuticals. Furthermore, the AD studies indicated that selection of individuals with an extreme phenotype could increase the relative importance of any single genetic effect within a sample. In a similar fashion, breast cancer genes were found by examining those rare families who exhibited near-Mendelian inheritance (genes BRCA1; OMIM 113705 [157], BRCA2; OMIM 600185 [257]). A recent study identified a gene responsible for susceptibility to Crohn's disease, a common disease with non-Mendelian inheritance [106]. This success indicated that the techniques applied to Mendelian disease could be successful in common complex diseases. There was how-

Figure 8.1: Loci affecting vision

ever, an element of good fortune in the Crohn's disease success (see chapter 1) and future studies will need to be more efficient (study design, phenotypic definition etc, see below) if the speed of progress is to properly reflect the massive investment of time, money and effort.

Many of the characters under genetic study are quantitative not qualitative in nature (i.e. disease or no disease). Although quantitative traits can be analysed using qualitative methods by truncating the trait distribution, this discards useful information. Quantitative trait locus (QTL) mapping was developed mainly in livestock, plant and model organism (e.g. mouse, *drosophila*) genetic applications, with human genetic applications and techniques only becoming widely used rather recently. There have been some successes in QTL studies with the regions indicated by genome scans beginning to allow identification of the underlying genes. For example, a recent paper showed that a QTL affecting a measure of growth in yeast was composed of 3 linked loci [219]. The 3 identified genes were each neither necessary nor sufficient to cause a discernible difference in trait value. This demonstrated that, whilst initial identification of the chromosomal region on which the QTL resided may be relatively simple, fine scale dissection of the gene or genes involved may be more difficult. It also serves as a reminder that the definition of a QTL is a chromosomal region contributing to the trait value and that in some cases this region will contain more than one gene contributing to the trait. Whilst every QTL will not necessarily be as complex as the 3 locus QTL for growth in yeast, dissection of the actual mutations will be difficult; even in cases in which LD is used to fine map the loci (see also below) it will not necessarily be obvious if the causative mutation has been identified; the identified polymorphism may simply be in strong LD with the disease polymorphism actually causing the disease. Knock outs or functional (expression) studies may help here but these rely on the existence of a suitable model organism (e.g. mouse). Korstanje and Paigen [124] reviewed the (mammalian) QTL literature in 2002, reporting that up to that point 29 genes had been unambiguously identified after being initially implicated by genome scans (i.e. based on linkage, although note that some of the 29 genes were for qualitative not quantitative traits). To gain some impression of the rate of progress, notice that another review of (mammalian) complex trait gene mapping in 2000 [91] reported there were no genes yet identified for quantitative traits. In livestock, genome scans for quantitative traits of commercial importance have become relatively common, leading to the identification of the causative mutation in a few cases (subsequent to the studies reported in [124], positive reports include [89, 73, 30]). These genome scans were based upon linkage initially, with LD mapping facilitating fine-mapping in some cases. In other cases, the post linkage mapping involved transgenic insertions, gene knockouts and/or examinations of functional differences in candidate genes. The multitude of techniques involved, stress that rapid future progress will depend upon the application of sets of complementary techniques.

It may be instructive to look at cases in which genes have not (yet) been identified. Many psychiatric diseases have a strong genetic component (chapter 1) but research to date has produced inconsistent results. Possible reasons for this have been discussed else-

where in the thesis (chapters 1, 6); these include lack of knowledge of which phenotypes best reflect the underlying genotype, appropriate study designs (given the possibility of genetic heterogeneity), population choice (particularly for LD based methods), insufficient sample size, poor matching of case and control populations (again for LD based methods), poor incorporation of essential environmental factors and lack of appreciation of the difficulties inherent in dissecting epistatic genetic effects. In chapter 6 (section 6 study design) the problems with the first major positive linkage to schizophrenia were indicated ([114, 18], [220], p284). Such problems led Owen [164] to ask in 1992, 'will schizophrenia become a graveyard for molecular geneticists?'. A long series of unreplicated association study results (as in chapter 1, there have been >50 Web of Science listed journal articles with "no association" and "schizophrenia" *in the title* since 1995, most of these contradicting previous results) in the years following 1992 only served to reinforce this pessimistic view. Events in bipolar disorder linkage mapping offered little in the way of encouragement either. Guo and Lange [91] describe the initially positive linkage results for bipolar disorder. This positive result was followed in short order by a number of negative results (which in themselves do not necessarily disprove the validity of the initial linkage) and a retraction when the initial families were followed up in more detail. Such false starts are unfortunate in that they undermine funding and public support for future studies, perhaps even suggesting to some of the public that these diseases are not genetic (there is a brief discussion of why they are highly likely to be genetic in chapter 1). These failures are useful in that they allow us to appreciate the need for rigorous application of standards for the declaration of significant linkage; researchers should be encouraged to correct appropriately for the multiple tests invariably applied in (psychiatric) disease (e.g. for multiple parametric models or disease definitions). The failures have also been useful in that they have forced investigators to critically re-evaluate the way in which they design their studies. Although, as discussed in chapter 6, there is no consensus on what constitutes the best way forward in terms of study design, the existence of a debate is healthy and the application of numerous approaches may be beneficial in the long run. The relative lack of progress may also serve as a reminder that, as discussed in chapters 1 and 7, whilst parametric models can be effective in extracting most of the information from non-Mendelian traits the inferences drawn from these should reflect the fact that the Mendelian model is only an approximation. Despite the fact that a first definite susceptibility locus for psychiatric diseases such as schizophrenia remains elusive, recent developments indicate that the field should be optimistic about further progress. The results of studies such as that reported in Iceland [218, 217] and elsewhere indicate that the identification of a demonstrably causal (increasing susceptibility) mutation is unlikely to be much more than a few years away.

Study design and choice of phenotype Since the identification of the susceptibility loci responsible for genetic component of complex disease will be difficult, it will be necessary to choose an appropriate study design and trait. A general principle, relevant to most

diseases/traits relies upon increasing the relative importance of genetic effects. Further, since there are a great many more possible multi-locus models for disease than single locus models, single locus models will likely prove more useful; enriching the sample for a single genetic factor will increase the chance of such single locus models succeeding. There are a number of ways of acquiring a genetically homogeneous sample. As mentioned above in the context of Alzheimer's disease, individuals with more severe forms of the disorder may be selected. Similarly, in bipolar disorder individuals with bipolar I (that is, individuals with severe bipolar disorder; cf. bipolar II the milder form) may be ascertained. Furthermore, families with bipolar I individuals are more likely to contain other individuals with affective disorders (recurrent major depression, bipolar II). Although the underlying biological relationship between these affective disorders is unknown, selecting sets of known relatives (extended families) may help increase the chances of the affecteds being affected as a result of genetic effects acting strongly in those particular families (chapter 6). This strategy was used with good effect in the analyses of families affected by schizophrenia and bipolar disorder in chapter 5. Samples may be selected to include individuals who are at low levels of environmental risk yet are still affected; e.g. non-smokers affected by lung conditions or cardiovascular disease sufferers in populations with low risk diets. Alternatively ethnic groups with high prevalence despite relatively low risk environment may be selected; this may represent one case where the diverse north American population may have an advantage over less diverse populations. One example of this would be studies of type 2 diabetes in Mexican Americans, where prevalence is unusually high (despite their being subject to broadly the same environmental effects as the rest of the United States population)[45, 59]. Furthermore, the incidence of traits such as type 2 diabetes [258] and BMI (see chapter 4, [177]) has changed with secular time, strongly suggesting that both environmental and genetic factors are of substantial importance. Assuming relevant environmental factors can be identified and measured, the challenge is then to apply appropriate methods which recognise the effects of the environment upon the trait. The variance components approach employed in chapters 1, 2 and 6 allows ready incorporation of such information.

In chapter 4 a number of cardiovascular disease risk factors were considered. In retrospect, the traits chosen for this analysis may not have been the most appropriate. On the one hand all of the traits (height, body mass index (BMI), total cholesterol, high density lipoprotein cholesterol) had high heritabilities. On the other hand, whilst the highest univariate LOD score was achieved for BMI, some of the other phenotypic measures (such as the cholesterol measures or other traits measured in the FHS families such as fasting glucose levels) may have a simpler composition (in terms of the underlying genetics) and may be more suitable for genetic studies.

Although many disease outcomes or traits in human genetics have binary clinical endpoints (e.g. schizophrenia, breast cancer), some have a quantitative trait, or *endophenotype*, closely related to the binary trait of interest. For example in Psychiatric disease, measures of certain electrical potentials on the scalp (P300, see chapter 2 and [25]) can

be shown to be related to disease status in schizophrenia. In cardiovascular disease, a number of the traits (e.g. total cholesterol, high density lipoprotein cholesterol, fasting glucose) measured in studies such as the Framingham heart study may be useful for finding genes responsible for traits such as hypertension and stroke. Joint quantitative and qualitative techniques have been described [252, 105] and may be more useful for gene detection than simple binary trait analysis in some complex diseases. Similarly, if there are multiple quantitative trait measures available, general multivariate techniques will allow additional information to be extracted from the available data. In chapters 2, 3 and 6 statistical techniques suitable for the analysis of longitudinal data were considered in detail. Since many of the traits of interest in human genetics change over time, this form of modelling may prove invaluable for data sets such as the Framingham heart study. Other data sets amenable to this form of multivariate analysis include twin studies [144, 233], other studies of CVD risk factors [111, 185, 236], a Framingham based study of children [180] and a number of behavioural genetic and psychiatric studies [160, 64, 67].

For quantitative traits, the selection of individuals with extreme phenotypic values can be seen as the analogue of selecting densely affected families or only using early onset cases. Selective genotyping has long been used to minimise genotyping costs in experimental organisms [139]. Analogous schemes have been proposed for human genetic studies. Risch and Zhang [193] propose selecting extreme discordant sibs, showing large increases in power could be achieved. A nice application of this sampling scheme was reported in a paper analysing data from a questionnaire based study of anxiety [74]. However, collecting sib pairs for studies based on clinical phenotypes may be more difficult. In such cases most individuals must be phenotypically assessed to determine where they lie in the trait distribution. As indicated in chapter 6, data that has already been collected (on additional relatives of the ascertained sib pairs) should not be discarded and should be analysed along with the sib pairs. If this means that additional methodology is required then it should be developed; data should not simply be discarded so that affected sib pair based methods can be applied. After all, whilst the costs of phenotyping and genotyping may vary from disease to disease, the costs involved in the statistical analysis are generally much less than those accumulated in the laboratory and in the clinic or hospital. Care must be taken, however, in applying linkage analyses to highly selected data sets. Simply applying the variance components techniques described in chapter 2 to data selected on the basis of observed phenotypes will lead to test statistics with unpredictable type I error rates (chapter 2), [3]). Techniques such as conditioning on the phenotype of the selected individuals can be applied to remedy this problem [204, 32, 49, 103]. Larger families will also offer benefits in terms of the available information on linkage phase (the haplotype status of the parents will always be unknown in small data structures but may be inferred in larger families). For some late onset diseases, parental information may not necessarily be available but information from other relatives (half sibs, cousins, uncles, aunts, nieces and nephews) can often be included [91]. In human genetic studies the relative advantage gained in selecting individuals will be small because the phenotyping: genotyping

cost ratio in humans is usually different to that in experimental organisms (phenotyping is generally much more expensive in human studies). As marker technologies become cheaper, the relative advantage of selective genotyping will be further diminished. A plausible alternative to selecting individuals on the basis of extreme phenotype is to ascertain prospective samples of whole (extended) families. Ideally the samples will be measured for a number of different phenotypes. Since phenotyping many different traits costs little more than phenotyping one trait (providing they are measured at the same session at the clinic or hospital) this approach may be cost effective. A good example of this would be the Framingham heart study (chapter 4). When traits are measured in this way (i.e. not on the basis of phenotype) no correction is required for variance component analyses (assuming the traits are multivariate normally distributed). The presence of multiple trait measures also opens up possibilities for multivariate analyses (chapters 2, 3 and 4 and above).

Studies based on model organisms are sometimes cited as holding great promise for disease mapping. Some of the advantages (ability to create knockouts, use of mutagenesis et cetera) of this approach were specified above in the discussion of fully dissected QTL. There are also other advantages. In the case of mouse models, large samples can be obtained relatively quickly. Once linkage signals have identified QTL, they can be refined using a far wider range of techniques than those available in unmanipulated populations (recombinant progeny testing, interval specific congenic strains; reviewed in [48]). Furthermore, appropriately annotated genomic information is now readily available for a number of model organisms. In many cases this facilitates the identification of genes with a common evolutionary origin (called homologs). There are also several problems with this approach however. An obvious disadvantage for some traits is the lack of comparable phenotype in animal models for diseases such as depression or schizophrenia. Inventive work by researchers in Oxford have allowed anxiety to be modelled as a quantitative trait in mice (using electric shocks and measures of excrement weight, [72]). However, even in cases in which there is an analogous phenotype, there are no guarantees that the underlying genes have the same effects in different species. Guo and Lange [91] give two pertinent examples; one where a mutation identified in mice has negligible effects in the homologous gene in humans and another in which a mutation, known to adversely affect human carriers, has only minor effects in mice. A general problem, applicable to most model systems, is that the genetic diversity present is usually just a small fraction of that present in unmanipulated populations. The more the experimental organisms are modified to make them simple to dissect, the less they reflect the overall genetic picture in human populations [259].

Technology Advances in technology will offer exciting new approaches for genetic dissection in the future. The recent availability of gene expression data may offer new insights into the genetic structure of many traits of interest. Microarray chips can be used to obtain measures of gene transcript levels in different tissues; simple applications may

involve comparisons of expression levels in diseased and healthy tissue or assessment of the transcript levels through a period of development. The measures of expression levels have been analysed within a QTL framework by treating the expression levels as 'traits' [58]. Although expression levels alone do not implicate the loci in question as causal factors for the disease or trait of interest, they may offer information on which genes constitute good candidates for further study (after initial localisation in a phenotype based genome scan). Wayne et al. [243] combine QTL mapping with expression level analysis by first performing a scan for QTL; this scan indicated over 5000 candidate genes. By subsequently performing analyses of expression data Wayne et al [243] were able to generate a shortlist of only 34 genes. Since microarrays allow assessment of transcript levels in many genes, the joint effect of multiple loci may also be examined. If multivariate QTL analyses (applied to expression levels) can be used to identify interactions between loci then this may provide direct evidence for epistasis, providing insights into the regulatory effects occurring between genes [58]. Furthermore, if expression levels at different stages of development can be assessed, longitudinal analyses (similar to those described for longitudinal QTL mapping in chapters 2 and 3) may offer insights into the changes across time. Although multiple trait analysis of longitudinal traits (see discussion of chapter 2) requires further work before it can be applied, this form analysis may allow characterisation of regulatory genetic effects (at least some of which will switch on and off at different stages of development).

Other advances in technology include laboratory methods for obtaining haplotypes directly and methods for pooling DNA. Whilst these advances are unlikely to substantially change the way we perform genetic studies some advantages may be discerned. In the case of 'direct' haplotypes, new molecular techniques [57] allow one to obtain the section of inherited chromosome; this offers more information than current methods which only allow the two alleles at a locus to be assessed together (it used to be impossible to directly infer whether alleles at nearby loci are part of the same haplotype). However, haplotypes can often be unambiguously reconstructed on the basis of known relationships (as in a multipoint linkage analysis) or estimated on the basis of LD between loci [230, 196]. DNA pooling [131] is a technique where, instead of typing individuals at each locus (assumed to be a single nucleotide polymorphism) individually, the DNA is pooled and the proportion of individuals carrying a particular allele is assessed. This offers no more information than conventional genotyping but the savings in cost may be substantial. A technique to correct for the bias introduced as a result of unequal amplification of the DNA has been proposed [239].

Linkage Disequilibrium mapping Although the main focus of this thesis has been on linkage analysis, the future of genetic studies will rely, to a greater or lesser extent upon association studies. As indicated in the introductory chapter, LD studies are an essential part of the fine mapping component of disease gene identification. Looked at from this point of view, LD studies are a relatively uncontroversial pursuit. The same cannot be

said of LD based whole genome association (WGA) studies. A parallel may be drawn here with the use of Bayes theorem in modern statistics; as a concept in probability theory Bayes theorem is thoroughly uncontroversial. The application of Bayes theorem as a vehicle for the incorporation of prior information into statistical practice however, is fraught with dangers, not least because of the dogmatic stance of some statisticians. Some of the objections to the use of LD for initial detection of disease loci were detailed in the introductory chapter. In short, these are based upon the strong dependence of the WGA technique on some critical assumptions. These are that the distribution of mutations will have to be simple (in terms of the number of alleles) and that the samples collected from population based samples will be reasonably homogeneous. Given the large amount of money invested in WGA methods, it seems important that the validity of these assumptions be rigorously tested. Although there has been considerable debate on this, there remains no consensus in the literature. Reich and Lander suggest that because alleles influencing common disease will not be under much selective pressure they may rise to high frequencies [184]. However, simulations by Pritchard have indicated that, on balance, even small selective effects will often force alleles to become rare [178, 179]. Furthermore, experience with many of the diseases studied thus far indicate that most population based samples will in fact be genetically diverse (implying that any single locus will only have a marginal effect upon risk) [245, 258].

Although the fact that LD only extends small distances (an order of magnitude either side of 10kb depending on the population [126]) makes it ideal for fine-mapping loci, this also means that a massive number of markers are required to cover the human genome; appropriate correction for multiple testing is vital here. It is unclear at the moment to what extent the Hap-Map project (again see introductory chapter) will alleviate this multiple testing problem. A very optimistic view for the future would be that since isolated populations have high levels of LD, initial detection (using WGA) may be possible in these populations; very fine scale mapping could then proceed in populations known to have particularly low levels of LD (for example some African populations [182]). The success of such a strategy depends very heavily upon the assumption that the common disease common variant hypothesis (i.e. that the allelic spectrum is relatively simple) holds for the disease of interest. If there are indeed common loci with small effects in a substantial proportion of individuals one obvious question would be, would knowing about such minor effects be of any practical consequence? It seems unlikely that detection of very small genetic effects would impact significantly upon human disease prevention/health.

Summary Despite some false starts, human genetics seems likely to advance significantly in the next decade or so, offering new insights into the genetic component of human disease. Genetic studies of complex human disease will yield a steady stream of discoveries, with genes likely to be identified for many of the major diseases affecting human populations (although how much impact such discoveries will have on human health from an epidemiological point of view remains to be seen). These discoveries, fuelled by sub-

stantial investment from both governments and private companies, will come as a result of a combination of different approaches; for the immediate future, successes will come as a result of the application of positional cloning (linkage followed by LD) based techniques. The genes found with linkage analysis techniques may not necessarily be those that cause a substantial portion of the overall population disease risk. Nonetheless, knowledge of how these genes act in affected individuals will be invaluable, with the insights provided ensuring further progress is made.

Bibliography

[1] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.*, 30:97–101, 2002.

[2] A. Alcais and L. Abel. Incorporation of covariates in multipoint model-free linkage analysis of binary traits: how important are unaffecteds? *Eur. J. Hum. Genet.*, 9:613–620, 2001.

[3] D. B. Allison, M. C. Neale, R. Zannolli, N. J. Schork, C. I. Amos, and J. Blangero. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am. J. Hum. Genet.*, 65:531–544, 1999.

[4] D. B. Allison, B. Thiel, P. St Jean, R. C. Elston, M. C. Infante, and N. J. Schork. Multiple phenotype modeling in gene-mapping studies of quantitative traits: Power advantages. *Am. J. Hum. Genet.*, 63:1190–1201, 1998.

[5] L. Almasy. GAW13 overall summary. *BMC Genet, in press*, 2003.

[6] L. Almasy and J. Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.*, 62:1198–1211, 1998.

[7] American Psychiatric Association, Washington D.C., United States. *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition, 1994.

[8] C. I. Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.*, 54:535–543, 1994.

[9] C. I. Amos, M. de Andrade, and D. K. Zhu. Comparison of multivariate tests for genetic linkage. *Hum. Hered.*, 51:133–144, 2001.

[10] C. I. Amos, R. C. Elston, G. E. Bonney, B. J. B. Keats, and G. S. Berenson. A multivariate method for detecting genetic-linkage, with application to a pedigree with an adverse lipoprotein phenotype. *Am. J. Hum. Genet.*, 47:247–254, 1990.

[11] C. I. Amos, J. Krushkal, T. J. Thiel, A. Young, D. K. Zhu, E. Boerwinkle, and M. de Andrade. Comparison of model-free linkage mapping strategies for the study of a complex trait. *Genet. Epidemiol.*, 14:743–748, 1997.

[12] A. Angius, E. Petretto, G. B. Maestrale, P. Forabosco, G. Casu, D. Piras, M. Fanciulli, M. Falchi, P. M. Melis, M. Palermo, and M. Pirastu. A new essential hypertension susceptibility locus on chromosome 2p24-p25, detected by genomewide search. *Am. J. Hum. Genet.*, 71:893–905, 2002.

[13] J. Angst. Historical aspects of the dichotomy between manic-depressive disorders and schizophrenia. *Schizophr. Res.*, 57:5–13, 2002.

[14] F. J. Ayala. The myth of eve - molecular-biology and human origins. *Science*, 270:1930–1936, 1995.

[15] J. A. Badner and E. S. Gershon. Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Mol. Psychiatr.*, 7:405–411, 2002.

[16] J. A. Badner and E. S. Gershon. Regional meta-analysis of published data supports linkage of autism with markers on chromosome 7. *Mol. Psychiatr.*, 7:56–66, 2002.

[17] D.J. Balding, M. Bishop, and C. Cannings, editors. *Handbook of Statistical Genetics*. Wiley, Chichester, United Kingdom, 2001.

[18] M. Baron. Genetics of schizophrenia and the new millennium: Progress and pitfalls. *Am. J. Hum. Genet.*, 68:299–312, 2001.

[19] A. S. Bassett, E. W. C. Chow, V. J. Vieland, and L. Brzustowicz. Is schizophrenia linked to chromosome 1q? *Science*, 298:U2–U2, 2002.

[20] A. L. Beaudet, G. L. Feldman, S. D. Fernbach, G. J. Buffone, and W. E. Obrien. Linkage disequilibrium, cystic-fibrosis, and genetic-counseling. *Am. J. Hum. Genet.*, 44:319–326, 1989.

[21] WD Beavis. The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Proceedings of the Corn and Sorghum Industry Research Conference*, pages 250–266, Washington D.C., 1994. American Seed Trade Association.

[22] W. H. Berrettini. Are schizophrenic and bipolar disorders related? a review of family and molecular studies. *Biol. Psychiatry*, 48:531–538, 2000.

[23] W. H. Berrettini. Are schizophrenic and bipolar disorders related? a review of family and molecular studies. *Biol. Psychiatry*, 48:531–538, 2000.

[24] W. H. Berrettini. Susceptibility loci for bipolar disorder: Overlap with inherited vulnerability to schizophrenia. *Biol. Psychiatry*, 47:245–251, 2000.

[25] D. Blackwood. P300, a state and a trait marker in schizophrenia. *Lancet*, 355:771–772, 2000.

[26] D. H. R. Blackwood, A. Fordyce, M. T. Walker, D. M. St Clair, D. J. Porteous, and W. J. Muir. Schizophrenia and affective disorders - cosegregation with a translocation at chromosome 1q42 that directly disrupts brainexpressed genes: Clinical and P300 findings in a family. *Am. J. Hum. Genet.*, 69:428–433, 2001.

[27] D. H. R. Blackwood, L. He, S. W. Morris, A. McLean, C. Whitton, M. Thomson, M. T. Walker, K. Woodburn, C. M. Sharp, A. F. Wright, Y. Shibasaki, D. M. StClair, D. J. Porteous, and W. J. Muir. A locus for bipolar affective disorder on chromosome 4p. *Nature Genet.*, 12:427–430, 1996.

[28] J. Blangero and L. W. Konigsberg. Multivariate segregation analysis using the mixed model. *Genet. Epidemiol.*, 8:299–316, 1991.

[29] J. Blangero, J. T. Williams, and L. Almasy. Robust LOD scores for variance component-based linkage analysis. *Genet. Epidemiol.*, 19:S8–S14, 2000.

[30] S. Blott, J. J. Kim, S. Moisio, A. Schmidt-Kuntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart, D. Johnson, L. Karim, P. Simon, R. Snell, R. Spelman, J. Wong, J. Vilkki, M. Georges, F. Farnir, and W. Coppieters. Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics*, 163:253–266, 2003.

[31] J. L. Blouin, B. A. Dombroski, S. K. Nath, V. K. Lasseter, P. S. Wolyniec, G. Nestadt, M. Thornquist, G. Ullrich, J. McGrath, L. Kasch, M. Lamacz, M. G. Thomas, C. Gehrig, U. Radhakrishna, S. E. Snyder, K. G. Balk, K. Neufeld, K. L. Swartz, N. DeMarchi, G. N. Papadimitriou, D. G. Dikeos, C. N. Stefanis, A. Chakravarti, B. Childs, D. E. Housman, H. H. Kazazian, S. E. Antonarakis, and A. E. Pulver. Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21. *Nature Genet.*, 20:70–73, 1998.

[32] M. Boehnke and D. A. Greenberg. The effects of conditioning on probands to correct for multiple ascertainment. *Am. J. Hum. Genet.*, 36:1298–1308, 1984.

[33] D. I. Boomsma and C. V. Dolan. A comparison of power to detect a QTL in sib-pair data using multivariate phenotypes, mean phenotypes, and factor scores. *Behav. Genet.*, 28:329–340, 1998.

[34] L. M. Brzustowicz, K. A. Hodgkinson, E. W. C. Chow, W. G. Honer, and A. S. Bassett. Location of a major susceptibility locus for familiar schizophrenia on chromosome 1q21-q22. *Science*, 288:678–682, 2000.

[35] N. J. Camp, S. L. Neuhausen, J. Tiobech, A. Polloi, H. Coon, and M. Myles-Worsley. Genomewide multipoint linkage analysis of seven extended Palauan pedigrees with schizophrenia, by a markov-chain monte carlo method. *Am. J. Hum. Genet.*, 69:1278–1289, 2001.

[36] Q. H. Cao, M. Martinez, J. Zhang, A. R. Sanders, J. A. Badner, A. Cravchik, C. J. Markey, E. Beshah, J. J. Guroff, M. E. Maxwell, D. M. Kazuba, R. Whiten, L. R. Goldin, E. S. Gershon, and P. V. Gejman. Suggestive evidence for a schizophrenia susceptibility locus on chromosome 6q and a confirmation in an independent series of pedigrees. *Genomics*, 43:1–8, 1997.

[37] L. R. Cardon and G. R. Abecasis. Using haplotype blocks to map human complex trait loci. *Trends Genet.*, 19:135–140, 2003.

[38] L. L. Cavalli-Sforza and W. F. Bodmer. *The Genetics of Human Populations*. Freeman, San Francisco, U.S.A., 1971.

[39] A. G. Clark. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr. Opin. Genet. Dev.*, 13:296–302, 2003.

[40] C. R. Cloninger. Turning-point in the design of linkage studies of schizophrenia. *Am. J. Med. Genet.*, 54:83–92, 1994.

[41] C. R. Cloninger, C. A. Kaufmann, S. V. Faraone, D. Malaspina, D. M. Svrakic, J. Harkavy-Friedman, B. K. Suarez, T. C. Matise, D. Shore, H. Lee, C. L. Hampe, D. Wynne, C. Drain, P. D. Markel, C. T. Zambuto, K. Schmitt, and M. T. Tsuang. Genome-wide search for schizophrenia susceptibility loci: The NIMH genetics initiative and millennium consortium. *Am. J. Med. Genet.*, 81:275–281, 1998.

[42] F. S. Collins. Positional cloning - lets not call it reverse anymore. *Nature Genet.*, 1:3–6, 1992.

[43] F. S. Collins. Positional cloning moves from perditional to traditional. *Nature Genet.*, 9:347–350, 1995.

[44] R. W. Cottingham, R. M. Idury, and A. A. Schaffer. Faster sequential genetic-linkage computations. *Am. J. Hum. Genet.*, 53:252–263, 1993.

[45] N. J. Cox, M. Frigge, D. L. Nicolae, P. Concannon, C. L. Hanis, G. I. Bell, and A. Kong. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in mexican americans. *Nature Genet.*, 21:213–215, 1999.

[46] D. Curtis, G. Kalsi, J. Brynjolfsson, M. McInnis, J. O'Neill, C. Smyth, E. Moloney, P. Murphy, A. McQuillin, H. Petursson, and H. Gurling. 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.*, 13:77–84, 2003.

[47] D. Curtis and P. C. Sham. Model-free linkage analysis using likelihoods. *Am. J. Hum. Genet.*, 57:703–716, 1995.

[48] A. Darvasi. Experimental strategies for the genetic dissection of complex traits in animal models. *Nature Genet.*, 18:19–24, 1998.

[49] M. de Andrade and C. I. Amos. Ascertainment issues in variance components models. *Genet. Epidemiol.*, 19:333–344, 2000.

[50] M. de Andrade, R. Gueguen, S. Visvikis, C. Sass, G. Siest, and C. I. Amos. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. *Genet. Epidemiol.*, 22:221–232, 2002.

[51] M. de Andrade and C. Olswold. Comparison of longitudinal variance components and regression based approaches for linkage detection on chromsome 17 for systolic blood pressure. *BMC Genet, in press*, 2003.

[52] L. E. DeLisi, S. H. Shaw, T. J. Crow, G. Shields, A. B. Smith, V. W. Larach, N. Wellman, J. Loftus, B. Nanthakumar, K. Razi, J. Stewart, M. Comazzi, A. Vita, T. Heffner, and R. Sherrington. A genome-wide scan for linkage to chromosomal regions in 382 sibling pairs with schizophrenia or schizoaffective disorder. *Am. J. Psychiat.*, 159:803–812, 2002.

[53] E.R. Dempster and I.M. Lerner. Heritability of threshold characters. *Genetics*, 35:212–236, 1950.

[54] S. D. Detera-Wadleigh, J. A. Badner, W. H. Berrettini, T. Yoshikawa, L. R. Goldin, G. Turner, D. Y. Rollins, T. Moses, A. R. Sanders, J. D. Karkera, L. E. Esterling, J. Zeng, T. N. Ferraro, J. J. Guroff, D. Kazuba, M. E. Maxwell, J. I. Nurnberger, and E. S. Gershon. A high-density genome scan detects evidence for a bipolardisorder susceptibility locus on 13q32 and other potential loci on 1q32 and 18p11.2. *Proc. Natl. Acad. Sci. U. S. A.*, 96:5604–5609, 1999.

[55] B. Devlin, K. Roeder, and L. Wasserman. Genomic control, a new approach to genetic-based association studies. *Theor. Popul. Biol.*, 60:155–166, 2001.

[56] C. Dib, S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, 380:152–154, 1996.

[57] C. M. Ding and C. R. Cantor. Direct molecular haplotyping of long-range genomic DNA with M1PCR. *Proc. Natl. Acad. Sci. U. S. A.*, 100:7449–7453, 2003.

[58] R. W. Doerge. Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.*, 3:43–52, 2002.

[59] R. Duggirala, J. Blangero, L. Almasy, R. Arya, T. D. Dyer, K. L. Williams, R. J. Leach, P. O'Connell, and M. P. Stern. A major locus for fasting insulin concentrations and insulin resistance on chromosome 6q with strong pleiotropic effects on obesity-related phenotypes in nondiabetic mexican americans. *Am. J. Hum. Genet.*, 68:1149–1164, 2001.

[60] R. Duggirala, J. T. Williams, S. Williams-Blangero, and J. Blangero. A variance component approach to dichotomous trait linkage analysis using a threshold model. *Genet. Epidemiol.*, 14:987–992, 1997.

[61] L. J. Eaves, M. C. Neale, and H. Maes. Multivariate multipoint linkage analysis of quantitative trait loci. *Behav. Genet.*, 26:519–525, 1996.

[62] J. Ekelund, I. Hovatta, A. Parker, T. Paunio, T. Varilo, R. Martin, J. Suhonen, P. Ellonen, G. Y. Chan, J. S. Sinsheimer, E. Sobel, H. Juvonen, R. Arajarvi, T. Partonen, J. Suvisaari, J. Lonnqvist, J. Meyer, and L. Peltonen. Chromosome 1 loci in Finnish schizophrenia families. *Hum. Mol. Genet.*, 10:1611–1617, 2001.

[63] J. Ekelund, D. Lichtermann, L. Hovatta, P. Ellonen, J. Suvisaari, J. D. Terwilliger, H. Juvonen, T. Varilo, R. Arajarvi, M. L. Kokko-Sahin, J. Lonnqvist, and L. Peltonen. Genome-wide scan for schizophrenia in the Finnish population: evidence for a locus on chromosome 7q22. *Hum. Mol. Genet.*, 9:1049–1057, 2000.

[64] T. C. Eley, P. Lichtenstein, and T. E. Moffitt. A longitudinal behavioral genetic analysis of the etiology of aggressive and nonaggressive antisocial behavior. *Dev. Psychopathol.*, 15:383–402, 2003.

[65] R. C. Elston, S. Buxbaum, K. B. Jacobs, and J. M. Olson. Haseman and Elston revisited. *Genet. Epidemiol.*, 19:1–17, 2000.

[66] R. C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Hum. Hered.*, 21:523–542, 1971.

[67] J. Endrass, J. Angst, A. Gamma, W. T. Gallo, V. Ajdacic-Gross, D. Eich, and W. Rossler. Premorbid personality in bipolar ii disorders, with reference to family genetics. results of a prospective epidemiological study. *Neurol. Psychiatr. Brain Res.*, 10:121–124, 2003.

[68] D. M. Evans. The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between variables. *Am. J. Hum. Genet.*, 70:1599–1602, 2002.

[69] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.*, 3(1):87–112, 1972.

[70] D. S. Falconer and T. F. C. Mackay. *Introduction to Quantitative Genetics*. Longman, Essex, United Kingdom, fourth edition, 1996.

[71] S. V. Faraone, T. Matise, D. Svrakic, J. Pepple, D. Malaspina, B. Suarez, C. Hampe, C. T. Zambuto, K. Schmitt, J. Meyer, P. Markel, H. Lee, J. Harkavy-Friedman, C. Kaufmann, C. R. Cloninger, and M. T. Tsuang. Genome scan of European-American schizophrenia pedigrees: Results of the NIMH genetics initiative and millennium consortium. *Am. J. Med. Genet.*, 81:290–295, 1998.

[72] J. Flint. Animal models of anxiety and their molecular dissection. *Semin. Cell Dev. Biol.*, 14:37–42, 2003.

[73] B. A. Freking, S. K. Murphy, A. A. Wylie, S. J. Rhodes, J. W. Keele, K. A. Leymaster, R. L. Jirtle, and T. P. L. Smith. Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. *Genome Res.*, 12:1496–1506, 2002.

[74] J. Fullerton, M. Cubin, H. Tiwari, C. Wang, A. Bomhra, S. Davidson, S. Miller, C. Fairburn, G. Goodwin, M. C. Neale, S. Fiddy, R. Mott, D. B. Allison, and J. Flint. Linkage analysis of extremely discordant and concordant sibling pairs identifies quantitative-trait loci that influence variation in the human personality trait neuroticism. *Am. J. Hum. Genet.*, 72:879–890, 2003.

[75] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

[76] W.J. Gauderman, S. Macgregor, L. Briollais, K. Scurrah, M. Tobin, T. Park, D. Wang, S. Rao, S. John, and S. Bull. Longitudinal data analysis in pedigree studies. *BMC Genet, in press*, 2003.

[77] C. Gee, T. Kang, J.L. Morrison, D.C. Thomas, and W.J. Gauderman. Hierarchical modeling of longitudinal data in segregation and linkage analysis. *BMC Genet, in press*, 2003.

[78] P. V. Gejman, M. Martinez, Q. H. Cao, E. Friedman, W. H. Berrettini, L. R. Goldin, P. Koroulakis, C. Ames, M. A. Lerman, and E. S. Gershon. Linkage analysis of 57 microsatellite loci to bipolar disorder. *Neuropsychopharmacology*, 9:31–40, 1993.

[79] A. W. George, P. M. Visscher, and C. S. Haley. Mapping quantitative trait loci in complex pedigrees: A twostep variance component approach. *Genetics*, 156:2081–2092, 2000.

[80] A. R. Gilmour, B. J. Gogel, B. R. Cullis, S. J. Welham, and R. Thompson. *ASREML User Guide Release 1.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, 2002.

[81] A. Goate, M. C. Chartierharlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, L. Giuffra, A. Haynes, N. Irving, L. James, R. Mant, P. Newton, K. Rooke, P. Roques, C. Talbot, M. Pericakvance, A. Roses, R. Williamson, M. Rossor, M. Owen, and J. Hardy. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimers-disease. *Nature*, 349:704–706, 1991.

[82] D. Goldgaber, M. I. Lerman, O. W. McBride, U. Saffiotti, and D. C. Gajdusek. Characterization and chromosomal localization of a cdnaencoding brain amyloid of Alzheimers-disease. *Science*, 235:877–880, 1987.

[83] D. E. Goldgar. Multipoint analysis of human quantitative genetic-variation. *Am. J. Hum. Genet.*, 47:957–967, 1990.

[84] D. Goossens, J. Del-Favero, and C. Van Broeckhoven. Trinucleotide repeat expansions: Do they contribute to bipolar disorder? *Brain Res. Bull.*, 56:243–257, 2001.

[85] H. H. H. Goring and J. D. Terwilliger. Linkage analysis in the presence of errors i: Complex-valued recombination fractions and complex phenotypes. *Am. J. Hum. Genet.*, 66:1095–1106, 2000.

[86] H. H. H. Goring and J. D. Terwilliger. Linkage analysis in the presence of errors iii: Marker loci and their map as nuisance parameters. *Am. J. Hum. Genet.*, 66:1298–1309, 2000.

[87] H. H. H. Goring and J. D. Terwilliger. Linkage analysis in the presence of errors iv: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am. J. Hum. Genet.*, 66:1310–1327, 2000.

[88] D. A. Greenberg, P. Abreu, and S. E. Hodge. The power to detect linkage in complex disease by means of simple LOD-score analyses. *Am. J. Hum. Genet.*, 63:870–879, 1998.

[89] B. Grisart, W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. Positional candidate cloning of a qtl in dairy cattle: Identification of a missense mutation in the bovine dgat1 gene with major effect on milk yield and composition. *Genome Res.*, 12:222–231, 2002.

[90] D. F. Gudbjartsson, K. Jonasson, M. L. Frigge, and A. Kong. Allegro, a new computer program for multipoint linkage analysis. *Nature Genet.*, 25:12–13, 2000.

[91] S. W. Guo and K. Lange. Genetic mapping of complex traits: Promises, problems, and prospects. *Theor. Popul. Biol.*, 57:1–11, 2000.

[92] H. M. D. Gurling, G. Kalsi, J. Brynjolfson, T. Sigmundsson, R. Sherrington, B. S. Mankoo, T. Read, P. Murphy, E. Blaveri, A. McQuillin, H. Petursson, and D. Curtis. Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21-22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3-24 and 20q12.1-11.23. *Am. J. Hum. Genet.*, 68:661–673, 2001.

[93] J. F. Gusella, N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, A. Y. Sakaguchi, A. B. Young, I. Shoulson, E. Bonilla, and J. B. Martin. A polymorphic DNA marker genetically linked to Huntingtons disease. *Nature*, 306:234–238, 1983.

[94] J. Hardy and D. J. Selkoe. Medicine - the amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics. *Science*, 297:353–356, 2002.

[95] R. Harrington, J. Hill, M. Rutter, K. John, H. Fudge, M. Zoccolillo, and M. Weissman. The assessment of lifetime psychopathology - a comparison of 2 interviewing styles. *Psychol. Med.*, 18:487–493, 1988.

[96] P. J. Harrison and M. J. Owen. Genes for schizophrenia? recent findings and their pathophysiological implications. *Lancet*, 361:417–419, 2003.

[97] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genetics*, 1:11–19, 1972.

[98] J. F. Hayes and W. G. Hill. Modification of estimates of parameters in the construction of genetic selection indexes (bending). *Biometrics*, 37:483–493, 1981.

[99] S. C. Heath. Markov chain monte carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.*, 61:748–760, 1997.

[100] S. Henn, N. Bass, G. Shields, T. J. Crow, and L. E. DeLisi. Affective illness and schizophrenia in families with multiple schizophrenic members: Independent illnesses or variant gene(s)? *Eur. Neuropsychopharmacol.*, 5:31–36, 1995.

[101] J. M. Hettema, M. C. Neale, and K. S. Kendler. A review and meta-analysis of the genetic epidemiology of anxiety disorders. *Am. J. Psychiat.*, 158:1568–1578, 2001.

[102] J. L. Hopper. Genetic epidemiology of female breast cancer. *Semin. Cancer Biol.*, 11:367–374, 2001.

[103] J. L. Hopper and J. D. Mathews. Extensions to multivariate normal-models for pedigree analysis. *Ann. Hum. Genet.*, 46:373–383, 1982.

[104] I. Hovatta, T. Varilo, J. Suvisaari, J. D. Terwilliger, V. Ollikainen, R. Arajarvi, H. Juvonen, M. L. Kokko-Sahin, L. Vaisanen, H. Mannila, J. Lonnqvist, and L. Peltonen. A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *Am. J. Hum. Genet.*, 65:1114–1124, 1999.

[105] J. A. Huang and Y. M. Jiang. Genetic linkage analysis of a dichotomous trait incorporating a tightly linked quantitative trait in affected sib pairs. *Am. J. Hum. Genet.*, 72:949–960, 2003.

[106] J. P. Hugot, M. Chamaillard, H. Zouali, S. Lesage, J. P. Cezard, J. Belaiche, S. Almer, C. Tysk, C. A. O'Morain, M. Gassull, V. Binder, Y. Finkel, A. Cortot, R. Modigliani, P. Laurent-Puig, C. Gower-Rousseau, J. Macry, J. F. Colombel, M. Sahbatou, and G. Thomas. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, 411:599–603, 2001.

[107] H-G. Hwu, C-M. Liu, CS-J. Fann, W-C. Ou-Yang, and SF-C. Lee. Linkage of schizophrenia with chromosome 1q loci in Taiwanese families. *Mol. Psychiatr.*, 8:445–452, 2003.

[108] F. Jaffrezic and S. D. Pletcher. Statistical models for estimating the genetic basis of repeated measures and other function-valued traits. *Genetics*, 156:913–922, 2000.

[109] F. Jaffrezic, I. M. S. White, and R. Thompson. Use of the score test as a goodness-of-fit measure of the covariance structure in genetic analysis of longitudinal data. *Genet. Sel. Evol.*, 35:185–198, 2003.

[110] F. Jaffrezic, I. M. S. White, R. Thompson, and P. M. Visscher. Contrasting models for lactation curve analysis. *J. Dairy Sci.*, 85:968–975, 2002.

[111] G. P. Jarvik, M. A. Austin, R. R. Fabsitz, J. Auwerx, T. Reed, J. C. Christian, and S. Deeb. Genetic influences on age-related change in total cholesterol, low-density lipoprotein-cholesterol, and triglyceride levels - longitudinal apolipoprotein-E genotype effects. *Genet. Epidemiol.*, 11:375–384, 1994.

[112] A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.*, 29:217–222, 2001.

[113] G. C. L. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. L. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nature Genet.*, 29:233–237, 2001.

[114] G. Kalsi, B. Mankoo, D. Curtis, R. Sherrington, G. Melmer, J. Brynjolfsson, T. Sigmundsson, T. Read, P. Murphy, H. Petersson, and H. Gurling. New dna markers with increased informativeness show diminished support for a chromosome 5q11-13 schizophrenia susceptibility locus and exclude linkage in two new cohorts of british and icelandic families. *Ann. Hum. Genet.*, 63:235–247, 1999.

[115] W. B. Kannel. Blood pressure as a cardiovascular risk factor - prevention and treatment. *JAMA-J. Am. Med. Assoc.*, 275:1571–1576, 1996.

[116] W. B. Kannel and R. B. DAgostino. Update of old coronary risk factors. *Cardiovasc. Risk Factors*, 6:244–253, 1996.

[117] W. B. Kannel, R. B. DAgostino, and J. L. Cobb. Effect of weight on cardiovascular disease. *Am. J. Clin. Nutr.*, 63:S419–S422, 1996.

[118] C. A. Kaufmann, B. Suarez, D. Malaspina, J. Pepple, D. Svrakic, P. D. Markel, J. Meyer, C. T. Zambuto, K. Schmitt, T. C. Matise, J. M. H. Friedman, C. Hampe, H. Lee, D. Shore, D. Wynne, S. V. Faraone, M. T. Tsuang, and C. R. Cloninger. NIMH genetics initiative millennium schizophrenia consortium: Linkage analysis of African-American pedigrees. *Am. J. Med. Genet.*, 81:282–289, 1998.

[119] J. L. Kennedy, L. A. Giuffra, H. W. Moises, L. L. Cavallisforza, A. J. Pakstis, J. R. Kidd, C. M. Castiglione, B. Sjogren, L. Wetterberg, and K. K. Kidd. Evidence against linkage of schizophrenia to markers on chromosome-5 in a northern swedish pedigree. *Nature*, 336:167–170, 1988.

[120] B. S. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L. C. Tsui. Identification of the cystic-fibrosis gene - genetic-analysis. *Science*, 245:1073–1080, 1989.

[121] M. Kirkpatrick, D. Lofsvold, and M. Bulmer. Analysis of the inheritance, selection and evolution of growth trajectories. *Genetics*, 124:979–993, 1990.

[122] M. Knapp, S. A. Seuchter, and M. P. Baur. Linkage analysis in nuclear families .2. relationship between affected sib-pair tests and lod score analysis. *Hum. Hered.*, 44:44–51, 1994.

[123] S. A. Knott, J. M. Elsen, and C. S. Haley. Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theor. Appl. Genet.*, 93:71–80, 1996.

[124] R. Korstanje and B. Paigen. From QTL to gene: the harvest begins. *Nature Genet.*, 31:235–236, 2002.

[125] F. Kronenberg, H. Coon, R. C. Ellison, I. Borecki, D. K. Arnett, M. A. Province, J. H. Eckfeldt, P. N. Hopkins, and S. C. Hunt. Segregation analysis of HDL cholesterol in the NHLBI family heart study and in utah pedigrees. *Eur. J. Hum. Genet.*, 10:367–374, 2002.

[126] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.*, 22:139–144, 1999.

[127] L. Kruglyak, M. J. Daly, M. P. ReeveDaly, and E. S. Lander. Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am. J. Hum. Genet.*, 58:1347–1363, 1996.

[128] M. C. LaBuda, M. Maldonado, D. Marshall, K. Otten, and D. S. Gerhard. A follow-up report of a genome search for affective disorder predisposition loci in the old order Amish. *Am. J. Hum. Genet.*, 59:1343–1362, 1996.

[129] E. Lander and L. Kruglyak. Genetic dissection of complex traits - guidelines for interpreting and reporting linkage results. *Nature Genet.*, 11:241–247, 1995.

[130] E. S. Lander and P. Green. Construction of multilocus genetic-linkage maps in humans. *Proc. Natl. Acad. Sci. U. S. A.*, 84:2363–2367, 1987.

[131] S. Le Hellard, S.J. Ballereau, P.M. Visscher, H.S. Torrance, J. Pinson, S.W. Morris, M.T. Thomson, C.A.M. Semple, W.J. Muir, D.H.R. Blackwood, D.J Porteous, and K.L. Evans. SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage. *Nucleic Acids Research*, pages 30–74, 2002.

[132] D. F. Levinson, P. Holmans, R. E. Straub, M. J. Owen, D. B. Wildenauer, P. V. Gejman, A. E. Pulver, C. Laurent, K. S. Kendler, D. Walsh, N. Norton, N. M. Williams, S. G. Schwab, B. Lerer, B. J. Mowry, A. R. Sanders, S. E. Antonarakis, J. L. Blouin, J. F. DeLeuze, and J. Mallet. Multicenter linkage study of schizophrenia candidate regions on chromosomes 5q, 6q, 10p, and 13q: Schizophrenia linkage collaborative group III. *Am. J. Hum. Genet.*, 67:652–663, 2000.

[133] D. F. Levinson, P. A. Holmans, C. Laurent, J. Mallet, B. Riley, K. S. Kendler, A. E. Pulver, P. V. Gejman, A. R. Sanders, S. G. Schwab, D. B. Wildenauer, M. J. Owen, and B. J. Mowry. Is schizophrenia linked to chromosome 1q? Response. *Science*, 298:U2–U4, 2002.

[134] D. F. Levinson, P. A. Holmans, C. Laurent, B. Riley, A. E. Pulver, P. V. Gejman, S. G. Schwab, N. M. Williams, M. J. Owen, D. B. Wildenauer, A. R. Sanders, G. Nestadt, B. J. Mowry, B. Wormley, S. Bauche, S. Soubigou, R. Ribble, D. A. Nertney, K. Y. Liang, L. Martinolich, W. Maier, N. Norton, H. Williams, M. Albus, E. B. Carpenter, N. deMarchi, K. R. Ewen-White, D. Walsh, M. Jay, J. F. Deleuze, F. A. O'Neill, G. Papadimitriou, A. Weilbaecher, B. Lerer, M. C. O'Donovan, D. Dikeos, J. M. Silverman, K. S. Kendler, J. Mallet, R. R. Crowe, and M. Walters. No major schizophrenia locus detected on chromosome 1q in a large multicenter sample. *Science*, 296:739–741, 2002.

[135] D. F. Levinson, M. D. Levinson, R. Segurado, and C. M. Lewis. Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: Methods and power analysis. *Am. J. Hum. Genet.*, 73:17–33, 2003.

[136] D. F. Levinson, M. M. Mahtani, D. J. Nancarrow, D. M. Brown, L. Kruglyak, A. Kirby, N. K. Hayward, R. R. Crowe, N. C. Andreasen, D. W. Black, J. M. Silverman, J. Endicott, L. Sharpe, R. C. Mohs, L. J. Siever, M. K. Walters, D. P. Lennon, H. L. Jones, D. A. Nertney, M. J. Daly, M. Gladis, and B. J. Mowry. Genome scan of schizophrenia. *Am. J. Psychiat.*, 155:741–750, 1998.

[137] C. M. Lewis, D. F. Levinson, L. H. Wise, L. E. DeLisi, R. E. Straub, I. Hovatta, N. M. Williams, S. G. Schwab, A. E. Pulver, S. V. Faraone, L. M. Brzustowicz, C. A.

Kaufmann, D. L. Garver, H. M. D. Gurling, E. Lindholm, H. Coon, H. W. Moises, W. Byerley, S. H. Shaw, A. Mesen, R. Sherrington, F. A. O'Neill, D. Walsh, K. S. Kendler, J. Ekelund, T. Paunio, J. Lonnqvist, L. Peltonen, M. C. O'Donovan, M. J. Owen, D. B. Wildenauer, W. Maier, G. Nestadt, J. L. Blouin, S. E. Antonarakis, B. J. Mowry, J. M. Silverman, R. R. Crowe, C. R. Cloninger, M. T. Tsuang, D. Malaspina, J. M. Harkavy-Friedman, D. M. Svrakic, A. S. Bassett, J. Holcomb, G. Kalsi, A. McQuillin, J. Brynjolfson, T. Sigmundsson, H. Petursson, E. Jazin, T. Zoega, and T. Helgason. Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *Am. J. Hum. Genet.*, 73:34–48, 2003.

[138] E. Lindholm, B. Ekholm, S. Shaw, P. Jalonen, G. Johansson, U. Pettersson, R. Sherrington, R. Adolfsson, and E. Jazin. A schizophrenia-susceptibility locus at 6q25, in one of the world's largest reported pedigrees. *Am. J. Hum. Genet.*, 69:96–105, 2001.

[139] M. Lynch and B Walsh. *Genetics and analysis of Quantitative Traits*. Sineaur Associates, Sunderland, USA, 1998.

[140] C. X. Ma, G. Casella, and R. L. Wu. Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics*, 161:1751–1762, 2002.

[141] M. E. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, C. Lin, L. Srinidhi, G. Barnes, S. A. Taylor, M. James, N. Groot, H. MacFarlane, B. Jenkins, M. A. Anderson, N. S. Wexler, J. F. Gusella, G. P. Bates, S. Baxendale, H. Hummerich, S. Kirby, M. North, S. Youngman, R. Mott, G. Zehetner, Z. Sedlacek, A. Poustka, A. M. Frischauf, H. Lehrach, A. J. Buckler, D. Church, L. Doucettestamm, M. C. Odonovan, L. Ribaramirez, M. Shah, V. P. Stanton, S. A. Strobel, K. M. Draths, J. L. Wales, P. Dervan, D. E. Housman, M. Altherr, R. Shiang, L. Thompson, T. Fielder, J. J. Wasmuth, D. Tagle, J. Valdes, L. Elmer, M. Allard, L. Castilla, M. Swaroop, K. Blanchard, F. S. Collins, R. Snell, T. Holloway, K. Gillespie, N. Datson, D. Shaw, and P. S. Harper. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntingtons-disease chromosomes. *Cell*, 72:971–983, 1993.

[142] S. Macgregor, S.A. Knott, I. White, and P.M. Visscher. Longitudinal analysis of the Framingham data. *BMC Genet, in press*, 2003.

[143] S. Macgregor, P. M. Visscher, S. Knott, D. Porteous, W. Muir, K. Millar, and D. Blackwood. Is schizophrenia linked to chromosome 1q? *Science*, 298:2277a, 2002.

[144] H. H. M. Maes, M. C. Neale, and L. J. Eaves. Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.*, 27:325–351, 1997.

[145] M. E. Marenberg, N. Risch, L. F. Berkman, B. Floderus, and U. Defaire. Genetic susceptibility to death from coronary heart-disease in a study of twins. *N. Engl. J. Med.*, 330:1041–1046, 1994.

[146] A. J. Marlow, S. E. Fisher, C. Francks, I. L. MacPhie, S. S. Cherny, A. J. Richardson, J. B. Talcott, J. F. Stein, A. P. Monaco, and L. R. Cardon. Use of multivariate linkage analysis for dissection of a complex cognitive trait. *Am. J. Hum. Genet.*, 72:561–570, 2003.

[147] P. McGuffin, R. Katz, S. Watkins, and J. Rutherford. A hospital-based twin register of the heritability of DSM-IV unipolar depression. *Arch. Gen. Psychiatry*, 53:129–136, 1996.

[148] P. McGuffin, F. Rijsdijk, M. Andrew, P. Sham, R. Katz, and A. Cardno. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch. Gen. Psychiatry*, 60:497–502, 2003.

[149] L. A. McInnes, M. A. Escamilla, S. K. Service, V. I. Reus, P. Leon, S. Silva, E. Rojas, M. Spesny, S. Baharloo, K. Blankenship, A. Peterson, D. Tyler, N. Shimayoshi, C. Tobey, S. Batki, S. Vinogradov, L. Meza, A. Gallegos, E. Fournier, L. B. Smith, S. H. Barondes, L. A. Sandkuijl, and N. B. Freimer. A complete genome screen for genes predisposing to severe bipolar disorder in two costa rican pedigrees. *Proc. Natl. Acad. Sci. U. S. A.*, 93:13060–13065, 1996.

[150] M. G. McInnis, T. H. Lan, V. L. Willour, F. J. McMahon, S. G. Simpson, A. M. Addington, D. F. MacKinnon, J. B. Potash, A. T. Mahony, J. Chellis, Y. Huo, T. Swift-Scanlan, H. Chen, R. Koskela, O. C. Stine, K. R. Jamison, P. Holmans, S. E. Folstein, K. Ranade, C. Friddle, D. Botstein, T. Marr, T. H. Beaty, P. Zandi, and J. R. DePaulo. Genome-wide scan of bipolar disorder in 65 pedigrees: supportative evidence for linkage at 8q24, 18q22, 4q32, 2p12, and 13q12. *Mol. Psychiatr.*, 8:288–298, 2003.

[151] M. G. McInnis, F. J. McMahon, T. Crow, C. A. Ross, and L. E. DeLisi. Anticipation in schizophrenia: A review and reconsideration. *Am. J. Med. Genet.*, 88:686–693, 1999.

[152] M. S. McPeek and A. Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.*, 65:858–875, 1999.

[153] F. Mekus, U. Laabs, H. Veeze, and B. Tummler. Genes in the vicinity of CFTR modulate the cystic fibrosis phenotype in highly concordant or discordant DeltaF508del homozygous sib pairs. *Hum. Genet.*, 112:1–11, 2003.

[154] Z. L. Meng, D. V. Zaykin, C. F. Xu, M. Wagner, and M. G. Ehm. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am. J. Hum. Genet.*, 73:115–130, 2003.

[155] K. R. Merikangas and N. Risch. Will the genomics revolution revolutionize psychiatry.? *Am. J. Psychiat.*, 160:625–635, 2003.

[156] K. Meyer. Estimating covariance functions for longitudinal data using a random regression model. *Genet. Sel. Evol.*, 30:221–240, 1998.

[157] Y. Miki, J. Swensen, D. Shattuckeidens, P. A. Futreal, K. Harshman, S. Tavtigian, Q. Y. Liu, C. Cochran, L. M. Bennett, W. Ding, R. Bell, J. Rosenthal, C. Hussey, T. Tran, M. McClure, C. Frye, T. Hattier, R. Phelps, A. Haugenstrano, H. Katcher, K. Yakumo, Z. Gholami, D. Shaffer, S. Stone, S. Bayer, C. Wray, R. Bogden, P. Dayananth, J. Ward, P. Tonin, S. Narod, P. K. Bristow, F. H. Norris, L. Helvering, P. Morrison, P. Rosteck, M. Lai, J. C. Barrett, C. Lewis, S. Neuhausen, L. Cannonalbright, D. Goldgar, R. Wiseman, A. Kamb, and M. H. Skolnick. A strong candidate for the breast and ovarian-cancer susceptibility gene BRRA1. *Science*, 266:66–71, 1994.

[158] J. K. Millar, J. C. Wilson-Annan, S. Anderson, S. Christie, M. S. Taylor, C. A. M. Semple, R. S. Devon, D. M. St Clair, W. J. Muir, D. H. R. Blackwood, and D. J. Porteous. Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.*, 9:1415–1423, 2000.

[159] L. Mirea, S. Bull, A. Paterson, J. Stafford, and L. Briollais. Comparison of Haseman-Elston regression analysis using single, summary and longitudinal measures of systolic blood pressure from GAW13 simulated data. *BMC Genet, in press*, 2003.

[160] T. E. M. Moffitt, A. Caspi, M. Rutter, and P. A. Silva. *Sex Differences in Antisocial Behaviour: Conduct Disorder, Delinquency and Violence in the Dunedin Longitudinal Study*. Cambridge University Press, Cambridge, 2001.

[161] H. J. Moller. Bipolar disorder and schizophrenia: Distinct illnesses or a continuum? *J. Clin. Psychiatry*, 64:23–27, 2003.

[162] A. P. Morris, J. C. Whittaker, and D. J. Balding. Bayesian fine-scale mapping of disease loci, by hidden markov models. *Am. J. Hum. Genet.*, 67:155–169, 2000.

[163] A. P. Morris, J. C. Whittaker, and D. J. Balding. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am. J. Hum. Genet.*, 70:686–707, 2002.

[164] M. J. Owen. Will schizophrenia become a graveyard for molecular geneticists. *Psychol. Med.*, 22:289–293, 1992.

[165] G. P. Page, C. I. Amos, and E. Boerwinkle. The quantitative LOD score: Test statistic and sample size for exclusion and linkage of quantitative traits in human sibships. *Am. J. Hum. Genet.*, 62:962–968, 1998.

[166] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. N. Nguyen, M. C. Norris, J. B. Sheehan, N. P. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O.

Trulson, K. R. Vyas, K. A. Frazer, S. P. A. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science*, 294:1719–1723, 2001.

[167] T. Paunio, J. Ekelund, T. Varilo, A. Parker, I. Hovatta, J. A. Turunen, K. Rinard, A. Foti, J. D. Terwilliger, H. Juvonen, J. Suvisaari, R. Arajarvi, J. Suokas, T. Partonen, J. Lonnqvist, J. Meyer, and L. Peltonen. Genome-wide scan in a nationwide study sample of schizophrenia families in finland reveals susceptibility loci on chromosomes 2q and 5q. *Hum. Mol. Genet.*, 10:3037–3048, 2001.

[168] L. Peltonen and V. A. McKusick. Genomics and medicine - dissecting human disease in the postgenomic era. *Science*, 291:1224–1229, 2001.

[169] L. Peltonen, A. Palotie, and K. Lange. Use of population isolates for mapping complex traits. *Nat. Rev. Genet.*, 1:182–190, 2000.

[170] M. Peltonen, F. Huhtasaari, B. Stegmayr, V. Lundberg, and K. Asplund. Secular trends in social patterning of cardiovascular risk factor levels in Sweden. the northern Sweden MONICA study 1986-1994. *J. Intern. Med.*, 244:1–9, 1998.

[171] C. Petit, J. Levilliers, and J. P. Hardelin. Molecular genetics of hearing loss. *Annu. Rev. Genet.*, 35:589–646, 2001.

[172] M. S. Phillips, R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding, M. A. Donaldson, J. F. Studebaker, W. M. Ankener, S. V. Alfisi, F. S. Kuo, A. L. Camisa, V. Pazorov, K. E. Scott, B. J. Carey, J. Faith, G. Katari, H. A. Bhatti, J. M. Cyr, V. Derohannessian, C. Elosua, A. M. Forman, N. M. Grecco, C. R. Hock, J. M. Kuebler, J. A. Lathrop, M. A. Mockler, E. P. Nachtman, S. L. Restine, S. A. Varde, M. J. Hozza, C. A. Gelfand, J. Broxholme, G. R. Abecasis, M. T. Boyce-Jacino, and L. R. Cardon. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genet.*, 33:382–387, 2003.

[173] S. D. Pletcher and C. J. Geyer. The genetic analysis of age-dependent traits: Modeling the character process. *Genetics*, 153:825–835, 1999.

[174] R. Pong-Wong, A. W. George, J. A. Woolliams, and C. S. Haley. A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genet. Sel. Evol.*, 33:453–471, 2001.

[175] B. M. Posner, M. M. Franz, P. A. Quatromoni, D. R. Gagnon, P. A. Sytkowski, R. B. Dagostino, and L. A. Cupples. Secular trends in diet and risk-factors for cardiovascular disease - the framingham-study. *J. Am. Diet. Assoc.*, 95:171–&, 1995.

[176] S. C. Pratt, M. J. Daly, and L. Kruglyak. Exact multipoint quantitative-trait linkage analysis in pedigrees by variance components. *Am. J. Hum. Genet.*, 66:1153–1157, 2000.

[177] A. M. Prentice and S. A. Jebb. Obesity in Britain - gluttony or sloth. *Br. Med. J.*, 311:437–439, 1995.

[178] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, 69:124–137, 2001.

[179] J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease - common variant ... or not? *Hum. Mol. Genet.*, 11:2417–2423, 2002.

[180] M. H. Proctor, L. L. Moore, D. Gao, L. A. Cupples, M. L. Bradlee, M. Y. Hood, and R. C. Ellison. Television viewing and change in body fat from preschool to early adolescence: The Framingham children's study. *Int. J. Obes.*, 27:827–833, 2003.

[181] S. Rao, L. Li, X. Li, K.L. Moser, Z. Guo, G. Shen, R. Cannata, E. Zirzow, E.J. Topol, and Q. Wang. Comparison of different summary measures (mean, slope and principal components) for genetic linkage analysis of longitudinal hypertensive phenotypes. *BMC Genet, in press*, 2003.

[182] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. Linkage disequilibrium in the human genome. *Nature*, 411:199–204, 2001.

[183] D. E. Reich and D. B. Goldstein. Genetic evidence for a Paleolithic human population expansion in africa. *Proc. Natl. Acad. Sci. U. S. A.*, 95:8119–8123, 1998.

[184] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends Genet.*, 17:502–510, 2001.

[185] K. E. Remsberg and R. M. Siervogel. A life span approach to cardiovascular disease risk and aging: The fels longitudinal study. *Mech. Ageing Dev.*, 124:249–257, 2003.

[186] F. V. Rijsdijk, J. K. Hewitt, and P. C. Sham. Analytic power calculation for QTL linkage analysis of small pedigrees. *Eur. J. Hum. Genet.*, 9:335–340, 2001.

[187] B. P. Riley and P. McGuffin. Linkage and associated studies of schizophrenia. *Am. J. Med. Genet.*, 97:23–44, 2000.

[188] N. Risch. Linkage strategies for genetically complex traits .1. multilocus models. *Am. J. Hum. Genet.*, 46:222–228, 1990.

[189] N. Risch. Linkage strategies for genetically complex traits .2. the power of affected relative pairs. *Am. J. Hum. Genet.*, 46:229–241, 1990.

[190] N. Risch. Linkage strategies for genetically complex traits .3. the effect of marker polymorphism on analysis of affected relative pairs. *Am. J. Hum. Genet.*, 46:242–253, 1990.

[191] N. Risch. The genetic epidemiology of cancer: Interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol. Biomarkers Prev.*, 10:733–741, 2001.

[192] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517, 1996.

[193] N. Risch and H. P. Zhang. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, 268:1584–1589, 1995.

[194] S. B. Roberts, C. J. MacLean, M. C. Neale, L. J. Eaves, and K. S. Kendler. Replication of linkage studies of complex traits: An examination of variation in location estimates. *Am. J. Hum. Genet.*, 65:876–884, 1999.

[195] A. Rosa, L. Fananas, M. J. Cuesta, V. Peralta, and P. Sham. 1q21-q22 locus is associated with susceptibility to the reality-distortion syndrome of schizophrenia spectrum disorders. *Am. J. Med. Genet.*, 114:516–518, 2002.

[196] D. J. Schaid. Relative efficiency of ambiguous vs. directly measured haplotype frequencies. *Genet. Epidemiol.*, 23:426–443, 2002.

[197] S. G. Schwab, J. Hallmayer, M. Albus, B. Lerer, G. N. Eckstein, M. Borrmann, R. H. Segman, C. Hanses, J. Freymann, A. Yakir, M. Trixler, P. Falkai, M. Rietschel, W. Maier, and D. B. Wildenauer. A genome-wide autosomal screen for schizophrenia susceptibility loci in 71 families with affected siblings: support for loci on chromosome 10p and 6. *Mol. Psychiatr.*, 5:638–649, 2000.

[198] G. Seaton, C. S. Haley, S. A. Knott, M. Kearsey, and P. M. Visscher. QTL express: mapping quantitative trait loci in of simple and complex pedigrees. *Bioinformatics*, 18:339–340, 2002.

[199] R. Segurado, S. D. Detera-Wadleigh, D. F. Levinson, C. M. Lewis, M. Gill, J. I. Nurnberg, N. Craddock, J. R. DePaulo, M. Baron, E. S. Gershon, J. Ekholm, S. Cichon, G. Turecki, S. Claes, J. R. Kelsoe, P. R. Schofield, R. F. Badenhop, J. Morissette, H. Coon, D. Blackwood, L. A. McInnes, T. Foroud, H. J. Edenberg, T. Reich, J. P. Rice, A. Goate, M. G. McInnis, F. J. McMahon, J. A. Badner, L. R. Goldin, P. Bennett, V. L. Willour, P. P. Zandi, J. J. Liu, C. Gilliam, S. H. Juo, W. H. Berrettini, T. Yoshikawa, L. Peltonen, J. Lonnqvist, M. M. Nothen, J. Schumacher, C. Windemuth, M. Rietschel, P. Propping, W. Maier, M. Alda, P. Grof, G. A. Rouleau, J. DelFavero, C. Van Broeckhoven, J. Mendlewicz, R. Adolfsson, M. A. Spence, H. Luebbert, L. J. Adams, J. A. Donald, P. B. Mitchell, N. Barden, E. Shink, W. Byerley, W. Muir, P. M. Visscher, S. Macgregor, H. Gurling, G. Kalsi, A. McQuillin, M. A. Escamilla, V. I. Reus, P. Leon, N. B. Freimer, H. Ewald, T. A. Kruse, O. Mors, U. Radhakrishna, J. L. Blouin, S. E. Antonarakis, and N. Akarsu. Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. *Am. J. Hum. Genet.*, 73:49–62, 2003.

[200] S. G. Self and K. Y. Liang. Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, 82:605–610, 1987.

[201] P. Sham. *Statistics in Human Genetics*. Arnold, London, UK, 1998.

[202] P. C. Sham and S. Purcell. Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am. J. Hum. Genet.*, 68:1527–1532, 2001.

[203] P. C. Sham, S. Purcell, S. S. Cherny, and G. R. Abecasis. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am. J. Hum. Genet.*, 71:238–253, 2002.

[204] P. C. Sham, J. H. Zhao, S. S. Cherny, and J. K. Hewitt. Variance-components QTL linkage analysis of selected and nonnormal samples: Conditioning on trait values. *Genet. Epidemiol.*, 19:S22–S28, 2000.

[205] S. H. Shaw, M. Kelly, A. B. Smith, G. Shields, P. J. Hopkins, J. Loftus, S. H. Laval, A. Vita, M. De Hert, L. R. Cardon, T. J. Crow, R. Sherrington, and L. E. DeLisi. A genome-wide search for schizophrenia susceptibility genes. *Am. J. Med. Genet.*, 81:364–376, 1998.

[206] R. Sherrington, J. Brynjolfsson, H. Petursson, M. Potter, K. Dudleston, B. Barraclough, J. Wasmuth, M. Dobbs, and H. Gurling. Localization of a susceptibility locus for schizophrenia on chromosome-5. *Nature*, 336:164–167, 1988.

[207] K. Silventoinen. Determinants of variation in adult body height. *J. Biosoc. Sci.*, 35:263–285, 2003.

[208] S. L. Slager, J. Huang, and V. J. Vieland. Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genet. Epidemiol.*, 18:143–156, 2000.

[209] J. Slate, P. M. Visscher, S. MacGregor, D. Stevens, M. L. Tate, and J. M. Pemberton. A genome scan for quantitative trait loci in a wild population of red deer (*Cervus elaphus*). *Genetics*, 162:1863–1873, 2002.

[210] C. A. B. Smith. Testing for heterogeneity of recombination fraction values in human genetics. *Ann. Hum. Gen*, page 175, 1963.

[211] E. Sobel and K. Lange. Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.*, 58:1323–1337, 1996.

[212] E. Sobel, H. Sengul, and D. E. Weeks. Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum. Hered.*, 52:121–131, 2001.

[213] C. P. Somnath, Y. C. J. Reddy, and S. Jain. Is there a familial overlap between schizophrenia and bipolar disorder? *J. Affect. Disord.*, 72:243–247, 2002.

[214] A. C. Sorensen, R. Pong-Wong, J. J. Windig, and J. A. Woolliams. Precision of methods for calculating identity-by-descent matrices using multiple markers. *Genet. Sel. Evol.*, 34:557–579, 2002.

[215] R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium - the insulin gene region and insulin-dependent diabetes-mellitus (IDDM). *Am. J. Hum. Genet.*, 52:506–516, 1993.

[216] D. Stclair, D. Blackwood, W. Muir, A. Carothers, M. Walker, G. Spowart, C. Gosden, and H. J. Evans. Association within a family of a balanced autosomal translocation with major mental-illness. *Lancet*, 336:13–16, 1990.

[217] H. Stefansson, J. Sarginson, A. Kong, P. Yates, V. Steinthorsdottir, E. Gudfinnsson, S. Gunnarsdottir, N. Walker, H. Petursson, C. Crombie, A. Ingason, J. R. Gulcher, K. Stefansson, and D. St Clair. Association of neuregulin 1 with schizophrenia confirmed in a Scottish population. *Am. J. Hum. Genet.*, 72:83–87, 2003.

[218] H. Stefansson, E. Sigurdsson, V. Steinthorsdottir, S. Bjornsdottir, T. Sigmundsson, S. Ghosh, J. Brynjolfsson, S. Gunnarsdottir, O. Ivarsson, T. T. Chou, O. Hjaltason, B. Birgisdottir, H. Jonsson, V. G. Gudnadottir, E. Gudmundsdottir, A. Bjornsson, B. Ingvarsson, A. Ingason, S. Sigfusson, H. Hardardottir, R. P. Harvey, D. Lai, M. D. Zhou, D. Brunner, V. Mutel, A. Gonzalo, G. Lemke, J. Sainz, G. Johannesson, T. Andresson, D. Gudbjartsson, A. Manolescu, M. L. Frigge, M. E. Gurney, A. Kong, J. R. Gulcher, H. Petursson, and K. Stefansson. Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet.*, 71:877–892, 2002.

[219] L. M. Steinmetz, H. Sinha, D. R. Richards, J. I. Spiegelman, P. J. Oefner, J. H. McCusker, and R. W. Davis. Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, 416:326–330, 2002.

[220] T. Strachan and A. P. Read. *Human Molecular Genetics*. BIOS, Magdalen Road, Oxford, U.K., second edition, 1999.

[221] D. O. Stram and J. W. Lee. Variance-components testing in the longitudinal mixed effects model. *Biometrics*, 50:1171–1177, 1994.

[222] R. E. Straub, Y. X. Jiang, C. J. MacLean, Y. L. Ma, B. T. Webb, M. V. Myakishev, C. Harris-Kerr, B. Wormley, H. Sadek, B. Kadambi, A. J. Cesare, A. Gibberman, X. Wang, F. A. O'Neill, D. Walsh, and K. S. Kendler. Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am. J. Hum. Genet.*, 71:337–348, 2002.

[223] R. E. Straub, C. J. MacLean, Y. Ma, B. T. Webb, M. V. Myakishev, C. Harris-Kerr, B. Wormley, H. Sadek, B. Kadambi, F. A. O'Neill, D. Walsh, and K. S. Kendler. Genome-wide scans of three independent sets of 90 irish multiplex schizophrenia families and follow-up of selected regions in all families provides evidence for multiple susceptibility genes. *Mol. Psychiatr.*, 7:542–559, 2002.

[224] R. E. Straub, C. J. MacLean, R. B. Martin, Y. L. Ma, M. V. Myakishev, C. Harris-Kerr, B. T. Webb, F. A. O'Neill, D. Walsh, and K. S. Kendler. A schizophrenia locus may be located in region 10p15-p11. *Am. J. Med. Genet.*, 81:296–301, 1998.

[225] R. E. Straub, C. J. MacLean, F. A. Oneill, J. Burke, B. Murphy, F. Duke, R. Shinkwin, B. T. Webb, J. Zhang, D. Walsh, and K. S. Kendler. A potential vulnerability locus for schizophrenia on chromosome 6p24-22 - evidence for genetic-heterogeneity. *Nature Genet.*, 11:287–293, 1995.

[226] B.K. Suarez and S.E. Hodge. A simple method to detect linkage for rare recessive disease: an application to juvenile diabetes. *Clinical Genetics*, 15:126–136, 1979.

[227] N. Takahata. Allelic genealogy and human-evolution. *Mol. Biol. Evol.*, 10:2–22, 1993.

[228] R. E. Tanzi, J. F. Gusella, P. C. Watkins, G. A. P. Bruns, P. Stgeorgehyslop, M. L. Vankeuren, D. Patterson, S. Pagan, D. M. Kurnit, and R. L. Neve. Amyloid beta-protein gene - cDNA, messenger-RNA distribution, and genetic-linkage near the Alzheimer locus. *Science*, 235:880–884, 1987.

[229] M. A. Taylor. Are schizophrenia and affective-disorder related - a selective literature-review. *Am. J. Psychiat.*, 149:22–32, 1992.

[230] S. Thomas and P.M. Visscher. Efficiency of direct haplotyping versus genotyping in association studies. *Genet. Epidemio., in press*, 2003.

[231] J. A. Todd. Human genetics - tackling common disease. *Nature*, 411:537–539, 2001.

[232] E. F. Torrey. Epidemiological comparison of schizophrenia and bipolar disorder. *Schizophr. Res.*, 39:101–106, 1999.

[233] A. Trouton, F. M. Spinath, and R. Plomin. Twins early development study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems in childhood. *Twin Res.*, 5:444–448, 2002.

[234] L. C. Tsui, M. Buchwald, D. Barker, J. C. Braman, R. Knowlton, J. W. Schumm, H. Eiberg, J. Mohr, D. Kennedy, N. Plavsic, M. Zsiga, D. Markiewicz, G. Akots, V. Brown, C. Helms, T. Gravius, C. Parker, K. Rediker, and H. Doniskeller. Cystic-fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science*, 230:1054–1057, 1985.

[235] R. C. J. Twells, C. A. Mein, M. S. Philips, J. F. Hess, R. Veijola, M. Gilbey, M. Bright, M. Metzker, B. A. Lie, A. Kingsnorth, E. Gregory, Y. Nakagawa, H. Snook, W. Y. S. Wang, J. Masters, G. Johnson, I. Eaves, J. M. M. Howson, D. Clayton, H. J. Cordell, S. Nutland, H. Rance, P. Carr, and J. A. Todd. Haplotype structure, LD blocks, and uneven recombination within the lrps gene. *Genome Res.*, 13:845–855, 2003.

[236] H. Ulmer, C. Kelleher, G. Diem, and H. Concin. Long-term tracking of cardiovascular risk factors among men and women in a large population-based health system - the Vorarlberg health monitoring & promotion programme. *Eur. Heart J.*, 24:1004–1013, 2003.

[237] G. Verbeke and G. Molenberghs. The use of score tests for inference on variance components. *Biometrics*, 59:254–262, 2003.

[238] P. M. Visscher and J. L. Hopper. Power of regression and maximum likelihood methods to map QTL from sib-pair and dz twin data. *Ann. Hum. Genet.*, 65:583–601, 2001.

[239] P. M. Visscher and S. Le Hellard. Simple method to analyze SNP-based association studies using DNA pools. *Genet. Epidemiol.*, 24:291–296, 2003.

[240] P. M. Visscher, M. H. Yazdi, A. D. Jackson, M. Schalling, K. Lindblad, Q. P. Yuan, D. Porteous, W. J. Muir, and D. H. R. Blackwood. Genetic survival analysis of age-at-onset of bipolar disorder: evidence for anticipation or cohort effect in families. *Psychiatr. Genet.*, 11:129–137, 2001.

[241] W. Y. S. Wang, H. J. Cordell, and J. A. Todd. Association mapping of complex diseases in linked regions: Estimation of genetic effects and feasibility of testing rare variants. *Genet. Epidemiol.*, 24:36–43, 2003.

[242] D. M. Waterworth, A. S. Bassett, and L. M. Brzustowicz. Recent advances in the genetics of schizophrenia. *Cell. Mol. Life Sci.*, 59:331–348, 2002.

[243] M. L. Wayne and L. M. McIntyre. Combining mapping and arraying: An approach to candidate gene identification. *Proc. Natl. Acad. Sci. U. S. A.*, 99:14903–14906, 2002.

[244] D. E. Weeks, J. Ott, and G. E. Lathrop. SLINK: a general simulation program for linkage analysis. *Am. J. Hum. Genet.*, 47:A204, 1990.

[245] K. M. Weiss and J. D. Terwilliger. How many diseases does it take to map a gene with SNPs? *Nature Genet.*, 26:151–157, 2000.

[246] A. S. Whittemore and J. Halpern. A class of tests for linkage using affected pedigree members. *Biometrics*, 50:118–127, 1994.

[247] D. B. Wildenauer, S. G. Schwab, W. Maier, and S. D. Detera-Wadleigh. Do schizophrenia and affective disorder share susceptibility genes? *Schizophr. Res.*, 39:107–111, 1999.

[248] J. T. Williams, H. Begleiter, B. Porjesz, H. J. Edenberg, T. Foroud, T. Reich, A. Goate, P. Van Eerdewegh, L. Almasy, and J. Blangero. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. ii. alcoholism and event-related potentials. *Am. J. Hum. Genet.*, 65:1148–1160, 1999.

[249] J. T. Williams and J. Blangero. Comparison of variance components and sibpair-based approaches to quantitative trait linkage analysis in unselected samples. *Genet. Epidemiol.*, 16:113–134, 1999.

[250] J. T. Williams and J. Blangero. Power of variance component linkage analysis to detect quantitative trait loci. *Ann. Hum. Genet.*, 63:545–563, 1999.

[251] J. T. Williams, R. Duggirala, and J. Blangero. Statistical properties of a variance components method for quantitative trait linkage analysis in nuclear families and extended pedigrees. *Genet. Epidemiol.*, 14:1065–1070, 1997.

[252] J. T. Williams, P. Van Eerdewegh, L. Almasy, and J. Blangero. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. i. likelihood formulation and simulation results. *Am. J. Hum. Genet.*, 65:1134–1147, 1999.

[253] N. M. Williams, M. I. Rees, P. Holmans, N. Norton, A. G. Cardno, L. A. Jones, K. C. Murphy, R. D. Sanders, G. McCarthy, M. Y. Gray, I. Fenton, P. McGuffin, and M. J. Owen. A two-stage genome scan for schizophrenia susceptibility genes in 196 affected sibling pairs. *Hum. Mol. Genet.*, 8:1729–1739, 1999.

[254] S. Williams-Blangero, J. L. VandeBerg, J. Subedi, M. J. Aivaliotis, D. R. Rai, R. P. Upadhayay, B. Jha, and J. Blangero. Genes on chromosomes 1 and 13 have significant effects on Ascaris infection. *Proc. Natl. Acad. Sci. U. S. A.*, 99:5533–5538, 2002.

[255] J. A. Williamson and C. I. Amos. On the asymptotic-behavior of the estimate of the recombination fraction under the null hypothesis of no linkage when the model is misspecified. *Genet. Epidemiol.*, 7:309–318, 1990.

[256] A. F. Wilson and R. C. Elston. Statistical validity of the Haseman-Elston sib-pair test in small samples. *Genet. Epidemiol.*, 10:593–598, 1993.

[257] R. Wooster, G. Bignell, J. Lancaster, S. Swift, S. Seal, J. Mangion, N. Collins, S. Gregory, C. Gumbs, G. Micklem, R. Barfoot, R. Hamoudi, S. Patel, C. Rice, P. Biggs, Y. Hashim, A. Smith, F. Connor, A. Arason, J. Gudmundsson, D. Ficenec, D. Kelsell, D. Ford, P. Tonin, D. T. Bishop, N. K. Spurr, B. A. J. Ponder, R. Eeles, J. Peto, P. Devilee, C. Cornelisse, H. Lynch, S. Narod, G. Lenoir, V. Egilsson, R. B. Barkadottir, D. F.

Easton, D. R. Bentley, P. A. Futreal, A. Ashworth, and M. R. Stratton. Identification of the breast-cancer susceptibility gene BRCA2. *Nature*, 378:789–792, 1995.

[258] A. F. Wright, A. D. Carothers, and M. Pirastu. Population choice in mapping genes for complex diseases. *Nature Genet.*, 23:397–404, 1999.

[259] A. F. Wright and N. D. Hastie. Complex genetic diseases: controversy over the Croesus code. *Genome Biology*, 2:1–8, 2001.

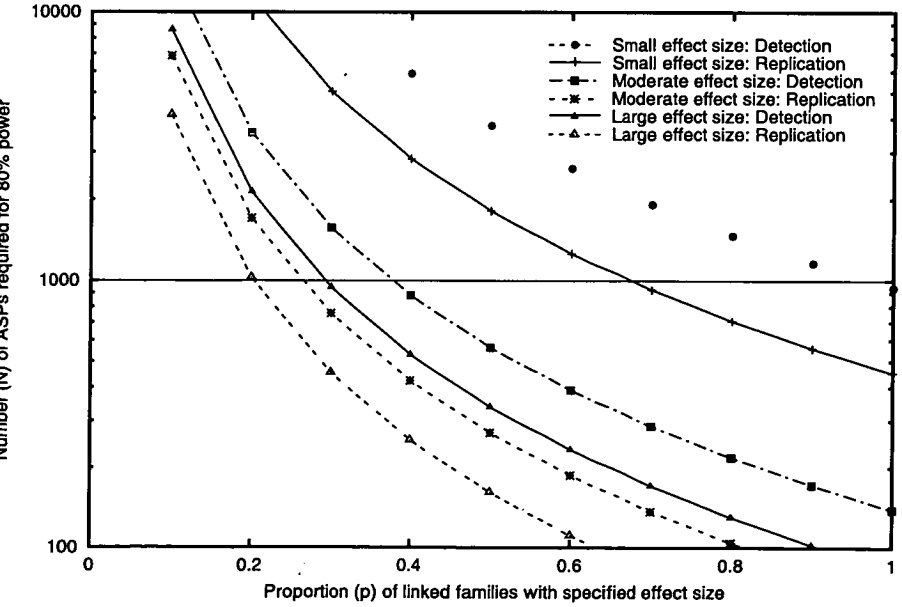
Is Schizophrenia Linked to Chromosome 1q?

Levinson *et al.* (*1*) reported the results of a meta-analysis of families showing no major schizophrenia locus on chromosome 1q. These results, based on a multicenter study of affected sibling pairs (ASPs), are in striking contrast to findings of several recent papers reporting susceptibility loci on 1q in extended families. Significant linkage (LOD = 6.5) at 1q21-22 was detected in Canadian families (*2*) and replicated in European origin families (*3, 4*). At 1q42, Blackwood *et al.* (*5*) obtained a LOD of 7.1 in a single Scottish family, while Ekelund *et al.* (*6*) obtained a LOD of 3.2 in Finnish pedigrees. How can these apparently conflicting results be reconciled? We suggest that locus heterogeneity adequately explains the failure of an ASP study with any reasonable sample size to replicate results from large extended families, and we have strong reservations about the limited interpretation of the results in (*1*).

We considered the effect of heterogeneity in two ways. First, we evaluated the power of the ASP mean test under heterogeneity. The number of sib pairs required to detect linkage is inversely proportional to the square of the proportion of linked families (*7*). Fig. 1 shows the effect of heterogeneity on the power to detect linkage given the effect of an allele segregating in the linked families, which increases risk to sibs by a given factor. Three effect sizes—small (factor 1.35), moderate (factor 3), and large (factor 7)—were considered. As shown, a sample of less than 1000 ASPs, as studied in (*1*), has little power to replicate linkage of schizophrenia to a locus that contributes to risk of illness in less than 20% of families. Note that Levinson *et al.* used the relative risk to siblings of affected individuals across the whole sample [λ_{sibs} in (*1*)] to determine power. Our interest is in showing how large a part heterogeneity plays in determining power. In the case of breast cancer, for example, the BRCA1 and BRCA2 genes have a large effect on risk (10- to 20-fold) in mutation carriers (*8*) but, because they are very rare in most populations, they are not readily detectable in large heterogeneous samples.

We also considered the power of nuclear families using SLINK software (*9*). Sixty families (each with 6 individuals in the sibship, equivalent to 15 ASPs) were simulated under a partially penetrant model and analyzed allowing for heterogeneity (*10*). The power to detect a LOD of 3 decreased rapidly; power for 75%, 50%, and 33% of families with mutations segregating at the gene of interest was 80%, 40%, and 5%, respectively.

In concluding that there is no locus of major effect on chromosome 1q, Levinson *et al.* have not appropriately considered locus heterogeneity. The logistic regression used in (*1*) ignores within-sample heterogeneity. Parametric linkage analysis incorporating heterogeneity is used but only with a recessive model. To ensure good power one must also fit a dominant model (*11*).



ig. 1. Power of the ASP mean test at different heterogeneity levels. The power to detect linkage LOD = 3) and replicaton of linkage (LOD = 1.2) were determined for three effect sizes: small factor 1.35), moderate (factor 3), and large (factor 7).

Though the results in initial genome scans are likely to be overestimates of effect size, the effects found in the studies reporting linkage to chromosome 1q21-22 and 1q42 are unlikely to be small in magnitude. Such effects will account for a sizable proportion of the variance in liability in particular families. The distribution of risk to schizophrenia can be well described by a model that incorporates genes of major effect and substantial locus heterogeneity. Under heterogeneity, ASP studies will require extremely large samples. Linkage analyses with large families and identification of cytogenetic variants associated with schizophrenia are appropriate strategies when heterogeneity is expected.

Stuart Macgregor
Peter M. Visscher
Sara Knott
Institute of Cell, Animal and Population Biology
Ashworth Laboratory
University of Edinburgh
West Mains Road
Edinburgh, EH9 3JT, UK
E-mail: stuart.macgregor@ed.ac.uk

David Porteous
Department of Medical Genetics
Molecular Medicine Centre
University of Edinburgh
Crewe Road
Edinburgh, EH4 2XU, UK

Walter Muir
Department of Psychiatry
Kennedy Tower
University of Edinburgh
Morningside Park
Edinburgh, EH10 5HF, UK

Kirsty Millar
Department of Medical Genetics, Molecular Medicine Centre

Douglas Blackwood
Department of Psychiatry, Kennedy Tower

References and Notes

1. D. Levinson *et al.*, *Science* **296**, 739 (2002).
2. L. M. Brustowicz, K. A. Hodgkinson, E. W. C. Chow, W. G. Honer, A. S. Bassett, *Science* **288**, 678 (2000).
3. H. Gurling *et al.*, *Am. J. Hum. Genet.* **68**, 661 (2001).
4. S. Shaw *et al.*, *Am. J. Med. Genet.* **81**, 364 (1998).
5. D. H. R. Blackwood *et al.*, *Am. J. Hum. Genet.* **69**, 428 (2001).
6. J. Ekelund *et al.*, *Hum. Mol. Genet.* **10**, 1611 (2001).
7. Derivation is available upon request. E-mail requests to S. Macgregor at stuart.macgregor@ed.ac.uk.
8. J. L. Hopper, *Semin. Cancer Biol.*, **11** (no. 5), 367 (2001).
9. D. Weeks *et al.*, *Am. J. Hum. Genet.* **47**, A204 (1990); ftp://linkage.rockefeller.edu/software/slink/
10. C. A. B. Smith, *Ann. Hum. Genet.* **27**, 175 (1963).
11. P. C. Sham, Statistics in Human Genetics (Arnold, London, 1998).
12. One or more of the authors are supported by Azko Nobel Organon, Medical Research Council, Biotechnology and Biological Sciences Research Council, the Scottish Executive, the Royal Society, and Caledonian Research Foundation.

21 June 2002; accepted 11 October 2002

A Genome Scan for Quantitative Trait Loci in a Wild Population of Red Deer (*Cervus elaphus*)

J. Slate,*,†,[1] P. M. Visscher,† S. MacGregor,† D. Stevens,* M. L. Tate* and J. M. Pemberton†

**AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand and †Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, Scotland, United Kingdom*

Manuscript received August 27, 2002
Accepted for publication September 20, 2002

ABSTRACT

Recent empirical evidence indicates that although fitness and fitness components tend to have low heritability in natural populations, they may nonetheless have relatively large components of additive genetic variance. The molecular basis of additive genetic variation has been investigated in model organisms but never in the wild. In this article we describe an attempt to map quantitative trait loci (QTL) for birth weight (a trait positively associated with overall fitness) in an unmanipulated, wild population of red deer (*Cervus elaphus*). Two approaches were used: interval mapping by linear regression within half-sib families and a variance components analysis of a six-generation pedigree of >350 animals. Evidence for segregating QTL was found on three linkage groups, one of which was significant at the genome-wide suggestive linkage threshold. To our knowledge this is the first time that a QTL for any trait has been mapped in a wild mammal population. It is hoped that this study will stimulate further investigations of the genetic architecture of fitness traits in the wild.

A common interpretation of Fisher's fundamental theorem of natural selection (Fisher 1958) is that selection will deplete additive genetic variance fastest for traits related to lifetime fitness (see also Frank and Slatkin 1992). By extension, fitness traits should be less heritable than other traits. Empirical studies provide some support for the theorem as there appears to be a negative relationship between a trait's heritability and its association with lifetime fitness (Kruuk *et al.* 2000; Merilä and Sheldon 2000), and life history traits tend to be less heritable than morphometric traits (Mousseau and Roff 1987; Roff and Mousseau 1987). However, the low heritability of fitness traits appears to be attributable to high levels of residual variance (*e.g.*, environmental variance, maternal effects, nonadditive genetic variance) rather than to low levels of additive genetic variance (Kruuk *et al.* 2000; Merilä and Sheldon 2000), and some studies suggest that traits closely related to fitness actually have the greatest additive genetic variance (Houle 1992; Merilä and Sheldon 2000).

The apparent maintenance of additive genetic variance for fitness-related traits raises several key questions that must be addressed to understand the mechanisms of natural selection (Barton and Keightley 2002). For example, can additive variation be attributed to many genes of small effect (polygenes) or relatively few of larger effect (oligogenes)? Are epistasis and pleiotropy important forces in the maintenance of genetic variation? One approach that can be used to address these questions is quantitative trait locus (QTL) mapping (Mitchell-Olds 1995). Over the last decade QTL mapping has been used to investigate the molecular basis of quantitative traits in disciplines such as medicine (Risch 2000), animal and plant breeding (Kearsey and Farquhar 1998; Andersson 2001), and evolutionary genetics (Lynch and Walsh 1998).

QTL studies in evolutionary genetics can be broadly broken down into two areas. First, considerable progress has been made in understanding the genetic basis of reproductive isolation (*e.g.*, Bradshaw *et al.* 1995) and species differences (Orr 2001), by producing experimental crosses between related species. A second major area of focus is the genetic architecture of quantitative traits within model species such as Drosophila (Mackay 2001). Using mapping resources such as recombinant inbred lines, a number of well-studied traits such as abdominal bristle number have been dissected so that their molecular basis is increasingly well understood. Recently, QTL have been detected for fitness components in Drosophila (Nuzhdin *et al.* 1997; Wayne *et al.* 2001) and *Caenorhabditis elegans* (Shook *et al.* 1996). However, these experiments have all been conducted within specially created crosses, which invariably have elevated levels of phenotypic and genetic variation relative to the parental lines. No study to date has been conducted in an unmanipulated, wild population unless one regards humans as wild mammals. The extent to which the genetic architecture of fitness traits in the laboratory mirrors the situation in the wild is controversial and unclear (Hoffmann 2000). Quite clearly, data

[1]*Corresponding author:* Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK. E-mail: j.slate@sheffield.ac.uk

are needed to assess the magnitude of QTL effects in the wild.

Despite previous suggestions that QTL for fitness traits could be detected within natural populations (Mitchell-Olds 1995), obtaining the necessary resources is not trivial. First, phenotypic data for traits known to influence lifetime fitness must be collected from a large sample of individuals—a notoriously difficult undertaking in wild populations (Endler 1986). Second, a panel of mapped, variable markers is required. Third, the relationship between the phenotyped individuals must be established to follow the segregation of marker alleles. Only when all of these criteria are met, can a genome-wide QTL scan can be conducted.

The vast majority of QTL experiments involve specially created populations, such as an F_2 generation or backcross created from different parental strains. These crosses offer a powerful approach to detecting QTL, but cannot be created in an unmanipulated, wild population. Similar limitations hinder complex disease gene mapping in human populations. To maximize the power of available pedigrees, sophisticated gene mapping algorithms and methodologies have been developed (Almasy and Blangero 1998; George *et al.* 2000). In particular, it has been suggested that complex multigenerational pedigrees offer greater power than the half-sib or full-sib families nested within them (Williams *et al.* 1997; Slate *et al.* 1999). The main drawback to complex pedigree methods is that they are computationally demanding, especially when pedigrees contain loops due to inbreeding. However, their use is becoming increasingly widespread, particularly in human populations. In natural populations where large sibships are generally uncommon, mapping in complex pedigrees may be the only available option. A two-step method to map QTL in complex pedigrees was recently described by George *et al.* (2000). First the number of genes identical-by-descent (IBD) between all individuals in the pedigree at any given chromosomal location is estimated using a Markov chain Monte Carlo (MCMC) sampler (Heath 1997). Once this IBD matrix is calculated, the contribution of the chromosomal location to the trait's variance is assessed using restricted maximum likelihood (REML). This methodology has been used to map a locus influencing bipolar disorder in a complex human pedigree (Visscher *et al.* 1999) and has been shown to be capable of detecting QTL in simulated livestock pedigrees, even when some marker genotypes are absent (George *et al.* 2000). Using this approach, it should be possible to map QTL in pedigreed wild populations, provided the necessary phenotypic and life history data are available.

Here we describe an attempt to map QTL for birth weight in a wild population of red deer (*Cervus elaphus*) on the Isle of Rum, Inner Hebrides, Scotland. The study population is well suited to QTL mapping for several reasons. Detailed life histories have been collected (Clutton-Brock *et al.* 1982; Kruuk *et al.* 2000), extensive pedigrees have been determined (Marshall *et al.* 1998; Kruuk *et al.* 2000), and the deer genome is mapped (Slate *et al.* 2002). Furthermore, a previous quantitative genetic analysis estimated the heritability, additive genetic variance, and relationship to lifetime fitness of a number of traits (Kruuk *et al.* 2000). Birth weight is a suitable trait for QTL analysis as it is known to have an additive genetic variance component (Kruuk *et al.* 2000), does not have a skewed distribution (unlike many life history traits), is positively associated with several fitness components (Clutton-Brock *et al.* 1987; Coulson *et al.* 1998; Kruuk *et al.* 1999), and, perhaps most importantly, is recorded in more individuals than any other trait.

MATERIALS AND METHODS

Study population: Historically red deer were known to be resident on the 10,600-ha island of Rum (57°0′ N, 6°20′ W), but they had been hunted to extinction by 1787. In 1845 the island was restocked for stalking purposes, and further reintroductions were made during the nineteenth and twentieth centuries. Introduced animals originated from at least five British deer parks or estates. The most recent introduction to the population is of greatest relevance to this article. In 1970 a hummel (antlerless stag) was crossed to Rum hinds to investigate the inheritance of hummellism. All male offspring developed normal antlers and were released on Rum following vasectomy operations. However, one of these male offspring, MAXI, subsequently achieved considerable reproductive success in the study area, siring over 30 offspring and having an estimated 400 descendants to date.

Since 1971, the North Block population has been intensively monitored with all resident animals individually recognizable (Clutton-Brock *et al.* 1982). In 1973 culling ceased in the study area and the population has remained stable at ~270 adult animals since 1982. Calves are routinely captured for marking and weighing and since 1982 have been sampled for genetic analysis. Other individuals born prior to 1982 were sampled postmortem or by chemical immobilization. Using nine microsatellite markers and three proteins, a detailed paternity analysis has been made (Marshall *et al.* 1998) with fathers assigned to 475 calves born between 1982 and 1996. Maternity is inferred from behavioral data and has never been contradicted by molecular data. A previous analysis concluded that the pedigree of animals descended from MAXI provides sufficient power to detect QTL (Slate *et al.* 1999). We chose this pedigree for several reasons. First, the fact that MAXI was sired by an immigrant animal may aid QTL detection, due to the probable introduction of novel additive genetic variation and by virtue of the fact that MAXI is the most heterozygous animal in the study population (Slate *et al.* 1999). Second, the MAXI pedigree contains many of the largest half-sibships documented in the study population, increasing the power to detect QTL (Figure 1). Finally, the reproductive success of MAXI and his descendants is such that it would have been impossible to construct a similarly sized pedigree of animals unrelated to MAXI.

Genotyping: The MAXI pedigree contained 364 individuals, of which 221 were known descendants of MAXI, and the remainder were "married-ins." The pedigree was typed for 90 microsatellite loci, the majority of which were originally characterized in cattle or sheep and mapped in their species of origin. The remaining loci were derived in other ruminants:

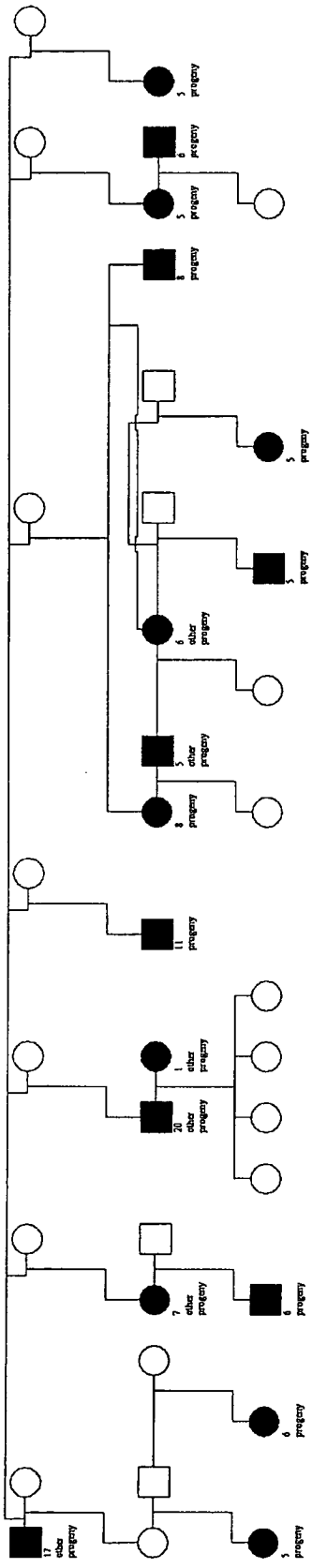


FIGURE 1.—Overview of the MAXI pedigree. Males are represented as squares and females as circles. MAXI is the male in the top left corner. The 17 parents with five or greater progeny are indicated by shading. All 17 half-sibships are interconnected. For clarity, genotyped and/or phenotyped individuals within the pedigree are omitted unless they are parents of five or greater offspring or connect 2 parents with five or greater progeny.

red deer, caribou (*Rangifer tarandus*), gazelle (*Gazella gazelle*), and wapiti (*Cervus elaphus canadensis*). Briefly, microsatellites were amplified by PCR using direct incorporation of [α-^{32}P]dCTP and products were run out on 6% polyacrylamide gels prior to visualization on X-ray film. Detailed amplification and electrophoresis protocols are described elsewhere (SLATE *et al.* 1998, 2000).

Pedigree checking: Paternity assignment in the population was initially declared with 80 or 95% confidence, using a battery of nine microsatellite and three protein loci (MARSHALL *et al.* 1998). Thus, a proportion of the paternities in the MAXI pedigree were likely to be wrong. By employing likelihood and multilocus genotypes at 84 loci, 44 of the 172 paternities initially included in the MAXI pedigree were identified as erroneous (SLATE *et al.* 2000). All maternal relationships inferred from behavioral data were confirmed by molecular evidence. The corrected pedigree is summarized in Figure 1.

Map construction: A deer genetic linkage map was constructed from the genotyped MAXI pedigree with the program CRI-MAP v2.4 (GREEN *et al.* 1990). Linked markers were initially identified using a two-point threshold of LOD = 3.0. Markers were also assumed to be linked if they were supported by LOD > 1.0 and there was an *a priori* reason for expecting linkage: *i.e.*, they were known to be linked in deer (SLATE *et al.* 2002) or in other ruminants (BARENDSE *et al.* 1997; MADDOX *et al.* 2001). Marker order and distances were determined using the BUILD and ALL commands. Any double-recombinant individuals were identified using the CHROMPIC command, and genotypes were reexamined. All genotypes found to be misscored were corrected.

In addition to the 90 microsatellite markers, the three protein loci screened by PEMBERTON *et al.* (1991) were included in the CRI-MAP analysis and in subsequent QTL mapping analyses. To compare the location and order of markers with their location on other ruminant maps the following sources were used:

Cattle: Reference was made to three published cattle linkage maps (MA *et al.* 1996; BARENDSE *et al.* 1997; BAND *et al.* 2000). Information from the maps can be accessed at the following web addresses:

The cattle genome database: http://spinal.tag.csiro.au/
The U.S. Meat Animal Research Center cattle genome mapping project: http://www.marc.usda.gov/genome/genome.html
The ARK database maintained by the Roslin Institute (Roslin, UK): http://www.thearkdb.org/browser?species=cow

Sheep: Linkage information on sheep was obtained from the third-generation map (MADDOX *et al.* 2001). Data from this map can be obtained at the following addresses:

Third-generation sheep map: http://rubens.its.unimelb.edu.au/~jillm/pages/gr_fig.htm
The ARK database: http://www.thearkdb.org/browser? species=sheep

Deer: A deer linkage map of >700 markers has now been completed (SLATE *et al.* 2002). An abbreviated version of this map can be viewed at:

The ARK database: http://www.thearkdb.org/browser?species=deer

Birth weight: Since 1982, >80% of calves have been weighed within 14 days of birth. Birth weight was estimated by backdating from capture weight, assuming a gain of 0.015 kg/hr (CLUTTON-BROCK *et al.* 1982), and was available for 295 individuals in the MAXI pedigree. To maximize the chances of detecting birth weight QTL, attempts were made to control

for potentially confounding environmental effects. A general linear model (GLM) identified four terms that explained 22% of the variation in birth weight: mean spring temperature in the April and May prior to birth, birth date (the number of days after May 1 that the calf was born), mother's reproductive status (a five-level categorical term describing whether the mother had produced a calf the previous year and how long she had reared it), and subdivision of study area (five-level categorical term). Residuals from the full model were used in subsequent QTL analyses. Model fitting was implemented in SPLUS v4.5 (MathSoft, Cambridge, MA).

QTL analysis: Two methods were used to detect QTL.

Interval mapping by linear regression of half-sib families: The revised MAXI pedigree contained a number of moderately sized half-sib families (Figure 1). A total of 17 parents (8 male and 9 female) with ≥5 genotyped and phenotyped offspring were identified (total number of offspring is 140). Seven parents (4 male and 3 female) had 8 or more offspring. Two individuals (MAXI and his son, RED7) sired >20 progeny each. An interval-mapping by linear regression method, based on Knott *et al.* (1996), was implemented in the web-based software package QTL Express (Seaton *et al.* 2002). Briefly, the phenotype is regressed on the conditional probability (inferred from the marker genotype) that a particular QTL allele was inherited from the sire. The analysis is nested within families and the test statistic is an F ratio with numerator degrees of freedom equal to the number of families and denominator degrees of freedom equal to $n - k - 1$, where n is the total number of progeny and k is the number of families. The process is repeated at 1-cM intervals along the chromosome. Analyses were performed on sibships of ≥8 informative progeny and on sibships of ≥5 informative progeny. Progeny were regarded as informative if typed for at least one marker on the linkage group and they were weighed at birth. Note that the inclusion of families with ≥5 progeny results in a greater number of progeny being analyzed, but may also result in a lower test statistic than when sibships of ≥8 are analyzed, as the test statistic has numerator degrees of freedom equal to the number of families. For this reason, half-sib families with <5 progeny were not analyzed. Interval mapping by linear regression is computationally undemanding, but does not utilize the full power of the MAXI pedigree (Slate *et al.* 1999). However, the empirical significance of possible QTL can be determined by permutation testing (Churchill and Doerge 1994).

The magnitude of QTL effects was calculated in two ways. First, the weighted mean of the *absolute values* of QTL allelic substitutions was calculated from only those families that appeared to be segregating for a QTL (nominally significant at $P < 0.05$). Second, QTL effects were calculated by taking the weighted mean of the absolute values of QTL allelic substitutions in *every* half-sibship with eight or more progeny. Weights were $1/\sigma^2$, where σ is the standard error of the estimated allelic substitution. Both approaches have their limitations. Under the first approach an upward bias is introduced as those families in which the QTL effect is overestimated by chance sampling are the most likely to achieve statistical significance (Beavis 1994). The second approach introduces a downward bias (without correcting the upward bias) because the assumption that every sire is segregating for QTL is unlikely to be correct. An additional upward bias is introduced by this approach. Because absolute effect sizes are used to estimate the mean effect size, every location in the genome will yield a positive effect size, even in cases where no QTL is present; *i.e.*, the true effect size is zero. However, given the very small number of progeny involved in each half-sib family it seems likely that the latter estimate will provide more biologically realistic estimates and, despite the known downward bias, may still produce overestimates of QTL effect. The latter estimate of QTL effect is the focus of discussion in the remainder of this article.

Two-step variance components analysis: At every marker location and at 5-cM intervals IBD coefficients were determined between all individuals in the revised MAXI pedigree, using the software LOKI v2.3 (http://www.stat.washington.edu/thompson/Genepi/Loki.shtml). IBD coefficients obtained after 1000 and 10,000 iterations of the program showed good concordance, and so we chose 1000 iterations as the default setting for subsequent analyses. IBD coefficients were estimated at 2-cM intervals for any chromosomal regions that were suggestive of a QTL. Variance components (VC) analysis was performed as described in George *et al.* (2000): First, a mixed linear model was fitted, under the assumption that birth weight was controlled by a number of unknown loci, acting additively and each of small effect. This model is termed the polygenic model and under matrix notation can be written as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{e}, \quad (1)$$

where $\mathbf{y}$ is an $(m \times 1)$ vector of phenotypes, $\mathbf{X}$ is an $(m \times s)$ design matrix, $\boldsymbol{\beta}$ is a $(s \times 1)$ vector of fixed effects, $\mathbf{Z}$ is an $(m \times q)$ incidence matrix relating animals to phenotypes, $\mathbf{a}$ is a $(q \times 1)$ vector of additive polygenic effects, and $\mathbf{e}$ is a residual vector.

The model provides an estimate of the trait's heritability, in addition to a likelihood value (L_0) for the REML solution. Essentially this model is the "animal model" used to estimate heritability and breeding values in animal breeding (Lynch and Walsh 1998) and more recently in evolutionary genetics (Kruuk *et al.* 2000).

A second linear model was fitted, which included all polygenic model terms *plus* a putative QTL effect at the location of interest. This model, termed the "polygenic + QTL model," may be written as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{Zq} + \mathbf{e}, \quad (2)$$

where $\mathbf{q}$ is a $(q \times 1)$ vector of additive QTL effects.

Estimates of the polygenic heritability (h^2) and the variance explained by the QTL (q^2) are obtained, in addition to a likelihood value (L_1).

Comparison of the likelihoods from the two models provides a test of the statistical significance of a possible QTL. For a single chromosomal location, the likelihood-ratio test statistic,

$$\text{LRT} = -2\ln(L_0 - L_1)$$

follows a 50:50 mixture distribution, where one component is a point of mass 0 and the other mixture component is a χ_1^2 distribution (Self and Liang 1987; Almasy and Blangero 1998; George *et al.* 2000). The distribution of the chromosome-wide test statistic is dependent on a number of factors such as pedigree structure, chromosome length, and missing marker data. However, under a variety of conditions it approximates a χ_1^2 distribution under the null hypothesis of no QTL segregating (George *et al.* 2000).

Significance thresholds: Any genome scan for QTL involves a large number of statistical tests, and the use of stringent significance thresholds before declaring linkage is well established (Churchill and Doerge 1994; Lander and Kruglyak 1995; Lynch and Walsh 1998). Permutation testing was used to assess statistical significance in the linear regression analysis because missing genotypes, differences in marker density, and segregation distortion are all accounted for (Churchill and Doerge 1994; Lynch and Walsh 1998). Chromosome-wide statistical significance was determined using 10,000 permuta-

TABLE 1

Summary of the markers typed in the MAXI pedigree

Linkage group	Markers (position)	Linkage group	Markers (position)
1	BR3510; FSHB (40.5); RM4 (50.3)	17	ILSTS93; BM1329 (35.0); JP27 (44.2)
2	JP15; TGLA86 (40.9)	18	RM188; OarCP26 (49.7); MGTG4B (85.4)
3	FCB5; AGLA293 (0)	19	OarMAF109; BM6506 (27.6); INRA11 (59.5); RT6 (73.5); TF (91.0); CSSM19 (94.8)
4	RT25; INRA121 (17.8); IDVGA55 (64.3); JP23 (78.0)	20	INRA6; HUJI177 (24.5); TGLA127 (65.2)
5	TGLA322; OarVH98 (30); TGLA94 (49.0); IDVGA46 (63.0); OarFCB193 (77.5); IOBT965 (82.0)	21	CSSM66; BM4513 (0); BM2934 (8)
6	ILSTS87	23	BMS1669; C217 (21.6); BL1071 (28.8); OarMAF18 (43.6); BMS2319 (49.2); AGLA232 (56.3)
7	BM1815; BM1258 (13.9); BM1818 (36.4); PRL (50.7)	24	HUJ175; CSSM41 (33.6); OarFCB304 (52.0); HIS-H1 (71.2)
8	IDVGA37; IDH (16.4); TGLA226 (32.7)	26	RT1; BM4208 (8.5); MM12 (18.9)
9	RM12; ILSTS6	27	JP38; OarMAF35 (15.6)
10	TGLA40	28	BM757; ETH225 (8.2)
11	ILSTS12; INRA131 (9.1); CSSM16 (16.5)	29	TGLA10
12	SPS113; TGLA378 (11.2); RM90 (17.9); BM888 (23.8); CSRM60 (42.0); CSSM39 (76.9)	30	ILSTS33
13	OarVH54; MCM527 (21.9); MPI (34.3); TGLA337 (34.3)	31	RM95
14	INRA35; BM1706 (3.4); TGLA334 (16.6); JP14 (46.2)	32	CSSM43; BM203 (33.7)
15	RT5; IRBP (34.7); ABS12 (40.4); IDVGA8 (42.2); PGAZac2 (60.8)	33	INRA40

Ninety-three markers (90 microsatellites and 3 allozymes) were typed and mapped to 30 linkage groups. Linkage groups 16, 22, and 25 were not typed for any marker. The position of each marker (in Kosambi centimorgans) is indicated in parentheses, with the first marker given position 0 cM. Linkage groups are orientated in the same direction as reported in Slate *et al.* (2002).

tions of the data. A threshold for genome-wide significance can be obtained by correcting the chromosome-wide significance threshold for the number of chromosomes analyzed. If it is assumed that 30 chromosomes were analyzed (see RESULTS), then a threshold of $P < 0.0017$ represents genome-wide significance. However, only 24 chromosomes were typed for two or more markers (Table 1), making a threshold of $P < 0.002$ appropriate. Confidence intervals for the location of possible QTL were determined by bootstrapping the data 1000 times (Visscher *et al.* 1996).

Permutation testing is problematic for the VC approach as it is unclear how to permute the data while retaining the association between polygenic variation and marker information (George *et al.* 2000). An alternative approach to permutation testing is to describe QTL as "suggestive" if they exceed a threshold expected to be observed once by chance in a genome scan and "significant" if exceeding a threshold expected to be observed by chance in only 5% of genome scans (Lander and Kruglyak 1995). Solving the formula given in Lander and Kruglyak (1995), and assuming a map length of 1548 cM covering 30 chromosomes (see RESULTS), the suggestive and significant thresholds are equivalent to likelihood-ratio test statistics of 7.02 and 13.64, respectively. However, these values assume an infinitely dense map of informative markers and it is suggested that significance thresholds are dropped by 20% for a map with 10-cM intervals (Lander and Kruglyak 1995). In this study the average marker interval was >15 cM, but to be conservative we assumed a mean interval of 10 cM giving thresholds of 5.62 and 10.91.

All regions of the genome that provided support for segregating QTL at the nominal $P < 0.05$ significance level are reported. While it is probable that some of these possible QTL are false positives, it is generally regarded as informative to the mapping community to report all regions that offer any evidence of linkage (Lander and Kruglyak 1995). Here we use the notation "possible QTL" to describe regions nominally significant at $P < 0.05$, while recognizing that QTL need to exceed a genome-wide threshold of 0.05 and be identified in a separate, independent sample of individuals or another population to be confirmed.

RESULTS

Genetic map: Ninety microsatellites and 3 allozyme loci were typed in the MAXI pedigree. Among the 93 loci, 53 were linked to another locus with support of LOD > 3.0. A further 25 loci were mapped on the basis of a LOD >1.0 and an *a priori* expectation of assignment to that linkage group (on the basis of marker location on other ruminant maps). Of the remaining 15 loci, 6 were expected to be singletons by inference from their location on other ruminant maps. The other 9 loci could not be placed on the expected (or any other) linkage group, presumably because they were relatively uninformative (observed heterozygosity <0.35) or their predicted location was >35 cM from the nearest mapped marker. One locus, McM527, mapped to deer linkage group 13, homologous to sheep chromosome 18, yet is mapped on chromosome 5 in sheep. The location of McM527 had reasonably high support (LOD = 9.55), so the location in deer was treated as genuine. It is assumed that the chromosomal segment containing McM527, underwent a translocation during ruminant karyotype evolu-

TABLE 2

Summary of chromosome-wide significant QTL

	Linear regression						Variance components				
Linkage group	Position (cM)	F	d.f.	P	Allelic effect(1) (kg)	Allelic effect(2) (kg)	Position (cM)	LRT	P	q^2	h^2
12	75	3.92	5, 61	0.004*	1.68	1.06	76	0.01	0.5	0.00	0.25
14	47	2.92	4, 55	0.029	2.11	0.82	32	4.36	0.018*	0.30	0.00
21	0	1.67	9, 54	0.119	3.38	0.80	0	6.27	0.006*,**	0.29	0.00

Possible QTL were detected using linear regression within half-sib families with eight or more progeny and by a VC analysis of the entire MAXI pedigree. Results for linkage groups 12, 14, and 21 are reported. For each methodology the location of the position (and associated nominal significance, P) giving the highest test statistic is reported. The linear regression yields an F ratio and the VC method yields a log-likelihood-ratio test statistic (LRT). The linear regression estimate of QTL magnitude is summarized as an allelic substitution effect in kilograms estimated from (1) families providing significant evidence for a segregating QTL or (2) all families of eight or more progeny. The VC estimate of QTL magnitude is summarized as the proportion of variance in residual birth weight explained by the QTL (q^2). For the VC method variance components are separated into the proportion of residual birth weight explained by the QTL (q^2) and by polygenic effects at other loci (h^2). *Significant at the chromosome-wide $P < 0.05$ level; **significant at the genome-wide suggestive linkage level.

tion, but the ancestral state is unknown. All other markers mapped to locations consistent with their position on other ruminant maps.

The total length of the map inferred from the MAXI pedigree was 978 cM. However, we considered any unlinked marker as potentially capable of detecting QTL up to 10 cM away in either direction. If the marker was predicted (from comparative location) to be at the end of a chromosome, then that marker was treated as capable of detecting QTL within 10 cM in one direction only. Using this somewhat arbitrary rule of thumb, it was predicted that the panel of 93 markers covered 1548 cM. The deer genome is estimated to be 2500 cM long (SLATE *et al.* 2002); thus the entire panel of markers gives ~62% genome coverage. Red deer have 33 autosomes of which 30 were typed for at least 1 marker and 24 were typed for two or more loci (Table 1). No markers were mapped to the sex chromosomes.

QTL analysis: In accordance with previous analyses (KRUUK *et al.* 2000), residual birth weight had a heritability significantly greater than zero in the MAXI pedigree ($h^2 = 0.24$, LRT = 9.99, $P < 0.002$). Statistical significance of polygenic heritability was determined by assuming that the likelihood-ratio test statistic obtained from the polygenic model and a residuals-only model (*i.e.*, a model without the polygenic component fitted) follows a χ_1^2 distribution (LYNCH and WALSH 1998).

Four linkage groups (LG8, -12, -14, and -21) provided evidence for birth weight QTL at the nominal $P < 0.05$ significance level, of which three exceeded the chromosome-wide significance level (Table 2; Figure 2). One region (LG21) was significant at the genome-wide suggestive linkage threshold.

Linkage group 12: Linear regression within half-sib families provided evidence for a birth weight QTL at the chromosome-wide significance level whether families of eight or more progeny ($F_{5,61} = 3.92$, nominal $P = 0.004$, chromosome-wide $P < 0.05$) or five or more progeny ($F_{16,103} = 2.36$, nominal $P = 0.005$, chromosome-wide $P < 0.01$) were considered. The effect of an allelic substitution at the possible QTL was estimated to be 1.06 kg. The QTL peak was at marker CSSM39 located at 76 cM (Figure 2), although the 95% confidence interval covered the entire linkage group. In fact, all possible QTL identified in this study had 95% confidence intervals that spanned the length of their linkage group. In contrast to linear regression, the VC analysis of the entire pedigree provided no evidence for a QTL on linkage group 12 (see DISCUSSION).

Linkage group 14: Linear regression of half-sibships with eight or more progeny provided evidence for a birth weight QTL ($F_{4,55} = 2.92$, nominal $P = 0.029$), but the test statistic was significant only at the chromosome-wide level at $P < 0.10$. When families with five or more progeny were analyzed the test statistic was not significant at the nominal level ($F_{14,94} = 1.61$, nominal $P = 0.090$) and did not exceed the threshold of $F = 1.93$ required for chromosome-wide significance. The possible QTL was at 47 cM (at marker JP14), with an allelic substitution equivalent to 0.82 kg.

The VC analysis of the full pedigree provided evidence for QTL at the chromosome-wide level at two locations (Figure 2). The first location (3.4 cM) is the map position of marker BM1706 and the second (34 cM) is flanked by markers TGLA334 and JP14. The second location provided a marginally higher test statistic (LRT = 4.36, nominal $P = 0.018$) and was estimated to explain 30% of the variation in residual birth weight. Given the wide confidence intervals of each QTL it cannot be assumed that the two peaks represent different QTL. The test-statistic profiles along the linkage group for the two methods are reasonably similar.

Linkage group 21: Linear regression of half-sibships with eight or more progeny ($F_{4,55} = 0.59$, nominal $P =$

0.67) or with five or more progeny ($F_{10,83} = 1.57$, nominal $P = 0.13$) did not provide evidence for a QTL segregating on LG21. However, a closer inspection of the data suggested that half-sibships in which the common parent was a female MAXI descendant inheriting allele 96 at marker BM2934 and allele 128 at marker BM4513 were segregating for a QTL. Nine half-sibships (six maternal and three paternal) where the common parent had inherited the "96-128" haplotype from MAXI were identified. Analysis of all nine sibships did not provide evidence that a QTL was segregating ($F_{9,54} = 1.67$, nominal $P = 0.119$, chromosome-wide $P = 0.103$). However, the possibility of a parent-of-origin effect (*i.e.*, paternal silencing) was further investigated by use of reduced linear regression where the sire QTL effects were set to zero in a reduced model (SEARLE 1971). This model provided evidence of a QTL in the maternal half-sibships ($F_{3,54} = 4.81$, $P = 0.005$), but not in the paternal half-sibships ($F_{6,54} = 0.30$, $P = 0.93$). Ideally, a larger number of sibships are required before a paternally silenced QTL can be confirmed.

The VC analysis of the entire pedigree provided evidence for a QTL that was significant at the suggestive experiment-wide level (LRT = 6.27, nominal $P = 0.006$, chromosome-wide $P = 0.013$). This possible QTL was located at marker BM2934 (0 cM) and explained 29% of the variation in residual birth weight. Note that the test statistic exceeded the chromosome-wide significance threshold at every location between markers BM2934 and BM4513 (Figure 2).

Linkage group 8: In addition to the previously mentioned linkage groups, LG8 provided very limited evidence for a birth weight QTL. Linear regression in half-sibships of eight or more progeny gave a nominally significant test statistic ($F_{4,55} = 2.54$, nominal $P = 0.050$), below the threshold required for chromosome-wide significance ($F = 3.00$). In families of five or more progeny the test statistic approached nominal significance ($F_{6,64} = 2.21$, nominal $P = 0.053$) but did not exceed the chromosome-wide significance threshold of $F = 2.47$. An allelic substitution at the possible QTL had an effect of 0.76 kg.

The VC method also provided weak evidence for a QTL at the nominally significant level ($P = 0.05$) but the test statistic did not exceed the chromosome-wide level. The possible QTL was estimated to explain 14% of variance in residual birth weight. The test-statistic profiles were similar for both methods, with the QTL peak located at marker IDVGA37. At present LG8 cannot be regarded as the location of a birth weight QTL

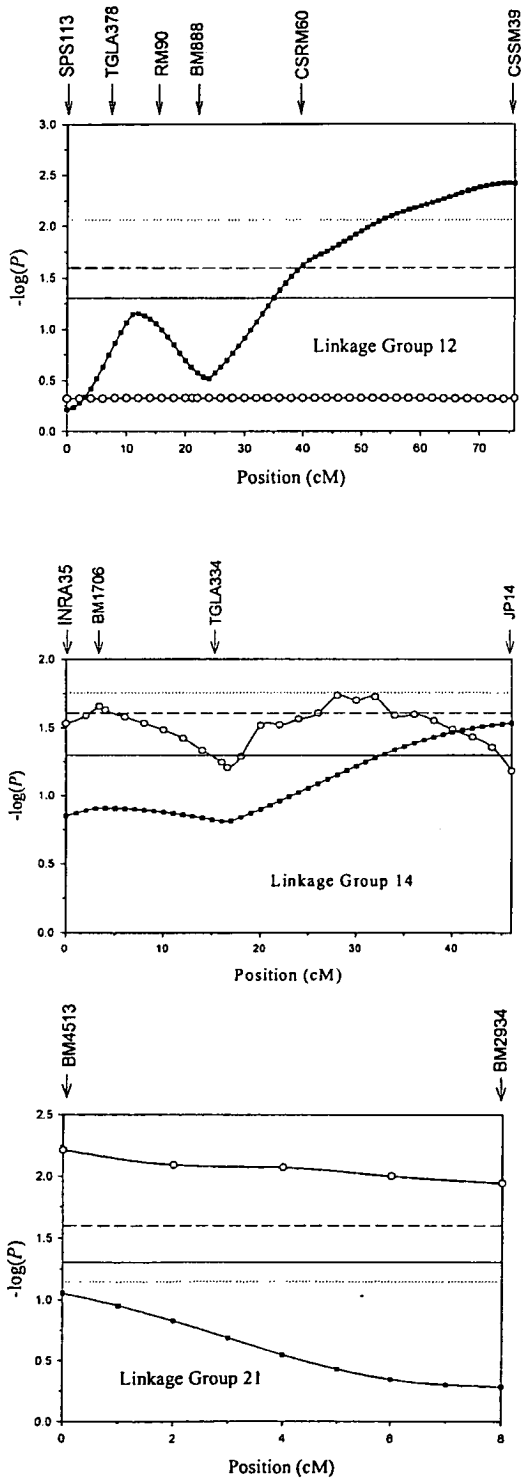


FIGURE 2.—Evidence for possible QTL on linkage groups 12, 14, and 21. Results from linear regression in half-sib families with eight or more progeny (■) and from VC analysis (○) of the entire MAXI pedigree are shown. The *y*-axis shows the statistic $-\log(P)$, where P is the nominal significance for a QTL at that location. Horizontal lines represent nominal significance at $P < 0.05$ (—), chromosome-wide significance at $P < 0.05$ for the linear regression approach (· · ·), and chromosome-wide significance at $P < 0.05$ for the VC approach (- - -). Vertical arrows indicate marker location. Note that the test statistic for the VC method on linkage group 21 also exceeds the threshold for suggestive linkage at the experiment-wide level. The profile for linear regression analysis on linkage group 21 represents the nine families that inherited the "96-128" haplotype from MAXI (see RESULTS).

although this region is worthy of investigation in follow-up studies.

DISCUSSION

Using two alternative methodologies, possible QTL for birth weight were identified on three separate linkage groups in a wild population of red deer. One possible QTL (on LG21) exceeded the threshold for genome-wide suggestive linkage, while two others (on LG12 and LG14) were significant at the chromosome-wide level. Two of the possible QTL were detected using both linear regression in half-sib families and VC in the entire pedigree, while the QTL on LG12 was detected by linear regression only. All of the possible QTL were estimated to be of large effect whether measured as an allelic substitution effect (in kilograms) or in terms of the proportion of variation in birth weight explained. Thus, questions arising from this analysis are: (1) Are the possible QTL genuine?, (2) how inflated are estimates of QTL magnitude?, and (3) why do the two methodologies provide different results for LG12?

Are the possible QTL genuine? Any genome-wide QTL mapping experiment is liable to generate false-positive QTL at the nominally significant $P < 0.05$ threshold, due to the large number of tests that are conducted (Churchill and Doerge 1994; Lander and Kruglyak 1995). We report all nominally significant chromosomal regions, but with a cautionary note that some of them may be artifactual. However, there is evidence to suggest that these QTL are real. First, two of the three QTL were detected by two approaches that make different assumptions in the underlying model. Linear regression in half-sib families assumes a QTL is a fixed effect with two alleles segregating in each family; the analysis takes place within families, background polygenic variation is disregarded, and the conditional probability of inheriting a particular QTL allele is estimated by the algorithm described in Knott *et al.* (1996). In contrast, VC makes no assumption about the number of QTL alleles segregating—rather, it assumes that the trait is described by a multivariate normal (MVN) distribution; the entire pedigree is considered simultaneously; *i.e.*, within- and between-family variances are utilized, background polygenic variance is included in the model, and the probability of two individuals sharing a QTL allele identically by descent is derived by a MCMC estimator. Of course, the two methods were applied in data sets with a number of common animals and so cannot be regarded as two wholly independent tests.

Further (admittedly weak) evidence that the QTL are genuine is provided by the location of birth weight QTL identified in related species. The only previous attempt to map birth weight QTL in deer identified loci on linkage groups 4 and 23 (Goosen 1997). There was little evidence for birth weight QTL in these regions in the Rum study population, although both linkage groups were reasonably well mapped (four and six markers, respectively). However, we are aware of three publications reporting birth weight QTL in cattle (Davis *et al.* 1998; Stone *et al.* 1999; Grosz and MacNeil 2001), located on bovine chromosomes 1, 2, 5, 6, 14, 18, and 21. Bovine chromosomes 2 and 14 are homologous to deer linkage groups 8 and 21—two regions where we found possible birth weight QTL. The QTL on bovine chromosome 2 was flanked by markers BM2113 and FCB11 (Grosz and MacNeil 2001), which also flank IDVGA37, the marker yielding a nominally significant QTL on LG8 in this study. Marker order appears to be conserved between cattle and deer in this region (Slate *et al.* 2002). The study of bovine chromosome 14 (equivalent to deer linkage group 21) indicated that two birth weight QTL may be segregating in cattle (Davis *et al.* 1998), although the closest markers were not reported, making cross-species comparisons problematic. For the time being we simply note the overlap in the location of cattle and deer birth weight QTL. It is tempting to ascribe this concordance to conserved QTL, but we prefer to reserve judgment until the causative mutations are identified or, at the very least, until a formal test of the similarity of across-experiment genome-wide test statistics is conducted (*e.g.*, Keightley and Knott 1999). Ultimately, it will be necessary to confirm the Rum birth weight QTL in a follow-up study. Since 1996 ~300 calves have been born and weighed, making this a feasible goal once these cohorts are pedigreed.

How inflated are estimates of QTL magnitude? Falconer and Mackay (1996) define a major gene as one that has an allelic substitution effect of 0.5 of a phenotypic standard deviation. The standard deviation of residual birth weight in the MAXI pedigree was 1.06 kg, and so QTL effects ranged from 0.75 to 1.0 phenotypic standard deviations (see Table 2). These estimates are at the upper end of the distribution of QTL effects described in domestic pig and dairy cattle QTL experiments (Hayes and Goddard 2001). Note that we estimated these QTL effects from all half-sibships of eight or more progeny, weighting each estimate by its standard error. However, this conservative approach does have some limitations: In particular, the assumption that all sires are segregating for a biallelic QTL may be erroneous, while a mean effect size estimated from absolute values must, by definition, yield an effect size greater than zero. An alternative methodology to calculate QTL effect size is to estimate the proportion of overall variation explained by each QTL, using the mean squares from the reduced and full linear regression models (see Knott *et al.* 1996 for a detailed description). Using this approach the possible QTL on LG12, -14, and -21 explained ~58, 27, and 25% of variation in birth weight, respectively. The VC method also estimated the QTL to be of large effect (each explaining ~30% of the variation in residual birth weight; Table 2). Given that the heritability of residual birth weight was estimated as only 0.24, these QTL estimates must be inflated. It is well known that estimates of QTL magnitude can

be upwardly biased, especially when sample sizes are relatively small (Beavis 1994). The so-called "Beavis effect" is an issue in all QTL mapping experiments, and it has been suggested that 500 or more phenotype records are required to minimize any bias (Beavis 1994; Orr 2001). Given that the linear regression analysis relied on little more than 100 phenotyped progeny (in some cases fewer) while the VC analysis relied on 295 phenotypes, it is accepted that both methods, particularly the former, would have provided upwardly biased estimates of QTL magnitude. In simulations involving 500 individuals and some missing marker data, the VC method overestimated QTL magnitude more than twofold (George *et al.* 2000). Given the obvious problems associated with small samples, it would be preferable to estimate QTL magnitude from an additional data set of study area animals. In the meantime we hypothesize that the QTL effects described here are upwardly biased, although they are likely to be of moderate-to-large effect or they would not have been identified. It is worth noting that the detection of QTL of smaller effect would have required sample sizes far larger than those available to us. In fact, it would probably require several centuries of intensive sampling of the study population to generate a suitably large data set. For example, if 100 half-sib families, each with 40 progeny, were sampled, the power to detect an allelic substitution of effect 0.2 of a phenotypic standard deviation (at the relaxed threshold of $\alpha = 0.05$) would be only 0.40. This power calculation applies to least-squares linear regression in half-sib families assuming a heritability of 0.25 and was calculated using the approach described in Slate *et al.* (1999).

An important issue when measuring the magnitude of QTL in complex pedigrees is distinguishing between a relatively rare QTL allele of large magnitude and the scenario of more common alleles of smaller effect. This problem of confounding between one and several QTL alleles is likely to be an issue in all studies that aim to map QTL in complex pedigrees. One possible solution to this problem is to investigate the magnitude of QTL in both the overall pedigree and the constituent families. This approach is reliant on the complex pedigree containing sufficiently large families to conduct the within-constituent family analysis. The MAXI pedigree probably represents a marginal case as only seven families contained eight or more progeny. A related problem involves distinguishing between a single QTL of large effect and several tightly linked QTL of smaller effect. Here we have assumed that each possible QTL represents a single locus, although this assumption can be confirmed only by finer mapping using larger sample sizes and/or molecular cloning of the loci responsible.

Comparison between the linear regression and VC methods: In general the two approaches yielded similar results, with possible QTL on LG14 and -21 detected by both methods. However, the VC method did not detect a QTL on linkage group 12. One possible explanation for this discrepancy is that the significant test statistic obtained from the linear regression approach was due to type I error (*i.e.*, a false-positive result). However, the test statistic was robust to permutation testing, and at least five sires appeared to be heterozygous for the QTL. Thus, we conducted a number of diagnostics to attempt to determine the cause of this discrepancy, using the software SOLAR 1.7.3 (http://www.sfbr.org/sfbr/public/software/solar/index.html; Almasy and Blangero 1998). SOLAR is similar to the approach we employed in that it uses IBD coefficients to perform QTL analysis by VC in a general pedigree framework, although a different algorithm is used. Although SOLAR was able to calculate only single IBD coefficients at marker locations rather than multipoint IBD coefficients at all positions, it was in agreement with our VC analysis in that no LG12 QTL was found in the MAXI pedigree. Points to note are that (i) LOKI and SOLAR provided similar IBD estimates at the marker locations and (ii) SOLAR provided the same maximum-likelihood solutions (yielding a test statistic of zero) as the REML software we used, even when handling IBD coefficients derived from LOKI. Thus, it seems unlikely that the failure of the VC method to find a QTL on LG12 can be attributed to problems associated with LOKI or with the REML program that provided the VC estimates. Both LOKI and SOLAR were subsequently used to conduct a VC analysis within the half-sibships where the linear regression approach had found evidence for segregating QTL. The VC methods found evidence (sometimes highly significant) for segregating QTL within these families, but generally with higher P values (*i.e.*, less significant) than those obtained by linear regression. Given the different assumptions underlying the linear regression and VC methods, it is perhaps not surprising that the two approaches yielded some inconsistencies. The VC method assumes that QTL effects are additive and could be confounded by maternal effects or QTL acting in a nonadditive fashion (*e.g.*, dominance). Reassuringly, the diagnostics suggested that the IBD coefficients estimated with LOKI were robust and accurate.

Intuitively, the VC method might be expected to have greater power than the linear regression approach as more phenotypic records are used. However, we note that in a simulated four-generation sheep pedigree containing 500 individuals, no inbreeding, and with highly informative markers (mean heterozygosity 0.88), the power of the VC method to detect a QTL that explained 10% of trait variation was only 0.48 (George *et al.* 2000). Power declined to ~0.30 when missing marker data were introduced into the simulations. Thus, the VC method may simply have failed to detect a genuine QTL on linkage group 12 (type II error).

QTL for traits associated with fitness: Ideally it would have been desirable to perform a linkage analysis on traits more intimately related to lifetime fitness. As adult males and females in the study population have a mean longevity of 10.5 and 11.5 years, respectively (Kruuk *et al.* 2000), estimates of lifetime reproductive success were

not available for surviving individuals (a large proportion of animals in the data set were still alive). However, this constraint is likely to be remedied within the next few years, and male lifetime reproductive success, which is known to have considerable levels of additive genetic variance (KRUUK *et al.* 2000), would be an interesting trait to investigate further. Given the highly skewed nature of traits such as male reproductive success, it will be necessary to minimize the risk of type I error. However, a combination of permutation testing and perhaps nonparametric QTL detection methods should overcome these difficulties.

The observation that additive genetic variation for a trait related to fitness is at least partially explained by major genes is contrary to predictions made from Fisher's theorem. Birth weight may be under directional selection, as only positive associations between birth weight and fitness components have been reported in the study population (CLUTTON-BROCK *et al.* 1987; COULSON *et al.* 1998; KRUUK *et al.* 1999). Alternatively, birth weight may be under stabilizing selection as very large calves may result in dystocia (calving difficulty). Although major genes may persist longer under stabilizing than directional selection, it is nonetheless expected that QTL of large effect will be selected to fixation under equilibrium conditions. Of direct relevance to this study is the observation that QTL of moderate to large effect on Drosophila bristle number—a trait subject to stabilizing selection—appear to be segregating at intermediate frequency in wild populations (LAI *et al.* 1994; LONG *et al.* 1998). A number of not necessarily exclusive mechanisms could result in the persistence of QTL of medium to large effect. Any wild population is likely to experience environmental heterogeneity and mutational input—forces that can maintain and generate additive genetic variation (HOULE *et al.* 1996; BARTON and KEIGHTLEY 2002). The role of additional forces that could serve to maintain variation, such as epistasis and antagonistic pleiotropy, is unknown in this population, although there is evidence for the latter (PEMBERTON *et al.* 1991). The question of whether birth weight has a negative genetic correlation with other fitness-related traits is worthy of further investigation.

Immigration from mainland populations has probably resulted in novel additive genetic variation being introduced to the study population. Despite being a descendant of the most recently introduced stag, MAXI does not appear to be heterozygous for the possible QTL on linkage groups 14 or 21, suggesting that polymorphism at these loci was already a feature of the study population. However, the role of gene flow in the maintenance of genetic variation in the wild is receiving increasing attention (SMITH *et al.* 1997). It is noteworthy that several other longitudinal studies of wild populations document introgression due to both conspecific and interspecific hybridization (GRANT and GRANT 2000; KELLER *et al.* 2001; VEEN *et al.* 2001). Confirmation and fine mapping of QTL in the study population will provide an opportunity to estimate the intensity of selection on recently introduced genes.

In conclusion, the presence of QTL of moderate to large effect in this population is consistent with findings in Drosophila (MACKAY 2001), plants (KEARSEY and FARQUHAR 1998), livestock (ANDERSSON 2001), and in crosses between reproductively isolated species (ORR 2001). Whether this consistency between experimental and wild populations will turn out to be a generalization remains to be seen. Clearly one of the major challenges awaiting evolutionary geneticists is to determine the molecular basis of additive genetic variation for fitness traits in the wild. It is hoped that this study will stimulate further attempts to address this crucial gap in the literature.

We thank Scottish Natural Heritage for permission to work on Rum; Tim Clutton-Brock, Fiona Guinness, and Steve Albon for their long-term contributions to the project; Angela Alexander, Ailsa Curnow, Sean Morris, and numerous volunteers for field data collection; John Williams for the donation of bovine primer sets; and Nick Barton and Terry Burke for helpful discussion. The manuscript was improved by the astute comments of two anonymous reviewers and the associate editor. The work was funded by the Biotechnology and Biological Sciences Research Council, the Natural Environment Research Council, and The Royal Society.

LITERATURE CITED

ALMASY, L., and J. BLANGERO, 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. **62:** 1198–1211.

ANDERSSON, L., 2001 Genetic dissection of phenotypic diversity in farm animals. Nat. Rev. Genet. **2:** 130–138.

BAND, M. R., J. H. LARSON, M. REBEIZ, C. A. GREEN, D. W. HEYEN *et al.*, 2000 An ordered comparative map of the cattle and human genomes. Genome Res. **10:** 1359–1368.

BARENDSE, W., D. VAIMAN, S. J. KEMP, Y. SUGIMOTO, S. M. ARMITAGE *et al.*, 1997 A medium-density genetic linkage map of the bovine genome. Mamm. Genome **8:** 21–28.

BARTON, N. H., and P. D. KEIGHTLEY, 2002 Understanding quantitative genetic variation. Nat. Rev. Genet. **3:** 11–21.

BEAVIS, W. D., 1994 The power and deceit of QTL experiments: lessons from comparative QTL studies, pp. 250–266 in *Proceedings of the Forty-Ninth Annual Corn and Sorghum Industry Research Conference.* American Seed Trade Association, Washington, DC.

BRADSHAW, JR., H. D., S. M. WILBERT, K. G. OTTO and D. W. SCHEMSKE, 1995 Genetic mapping of floral traits associated with reproductive isolation in monkeyflowers (*Mimulus*). Nature **376:** 762–765.

CHURCHILL, G. A., and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

CLUTTON-BROCK, T. H., F. E. GUINNESS and S. D. ALBON, 1982 *Red Deer—Behavior and Ecology of Two Sexes.* Edinburgh University Press, Edinburgh.

CLUTTON-BROCK, T. H., M. MAJOR, S. D. ALBON and F. E. GUINNESS, 1987 Early development and population dynamics in red deer. I. Density-dependent effects on juvenile survival. J. Anim. Ecol. **56:** 53–67.

COULSON, T. N., J. M. PEMBERTON, S. D. ALBON, M. A. BEAUMONT, T. C. MARSHALL *et al.*, 1998 Microsatellites measure inbreeding depression and heterosis in red deer. Proc. R. Soc. Lond. Ser. B **265:** 489–495.

DAVIS, G. P., D. J. S. HETZEL, N. J. CORBET, S. SCACHERI, S. LOWDEN *et al.*, 1998 The mapping of quantitative trait loci for birth weight in a tropical beef herd. Proceedings of the 6th World Congress

on Genetics Applied to Livestock Production, Armidale, Australia, pp. 441–444.
ENDLER, J. A., 1986 *Natural Selection in the Wild.* Princeton University Press, Princeton.
FALCONER, D. S., and T. F. C. MACKAY, 1996 *Introduction to Quantitative Genetics.* Longman, New York.
FISHER, R. A., 1958 *The Genetical Theory of Natural Selection.* Dover Publications, New York.
FRANK, S. A., and M. SLATKIN, 1992 Fisher's fundamental theorem of natural selection. Trends Ecol. Evol. **7:** 92–95.
GEORGE, A. W., P. M. VISSCHER and C. S. HALEY, 2000 Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. Genetics **156:** 2081–2092.
GOOSEN, G. J. C., 1997 An interspecies hybrid in deer. Ph.D. Thesis, University of New England, Armidale, Australia.
GRANT, P. R., and B. R. GRANT, 2000 Quantitative genetic variation in populations of Darwin's finches, pp. 3–40 in *Adaptive Genetic Variation in the Wild,* edited by T. A. MOUSSEAU, B. SINERVO and J. ENDLER. Oxford University Press, Oxford.
GREEN, P., K. FALLS and S. CROOKS, 1990 Documentation for CRIMAP. Washington University, St. Louis.
GROSZ, M. D., and M. D. MACNEIL, 2001 Putative quantitative trait locus affecting birth weight on bovine chromosome 2. J. Anim. Sci. **79:** 68–72.
HAYES, B., and M. E. GODDARD, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. **33:** 209–229.
HEATH, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61:** 748–760.
HOFFMANN, A. A., 2000 Laboratory and field heritabilities: some lessons from *Drosophila,* pp. 200–218 in *Adaptive Genetic Variation in the Wild,* edited by T. A. MOUSSEAU, B. SINERVO and J. A. ENDLER. Oxford University Press, Oxford.
HOULE, D., 1992 Comparing evolvability and variability of quantitative traits. Genetics **130:** 195–204.
HOULE, D., B. MORIKAWA and M. LYNCH, 1996 Comparing mutational variabilities. Genetics **143:** 1467–1483.
KEARSEY, M. J., and A. G. L. FARQUHAR, 1998 QTL analysis in plants: Where are we now? Heredity **80:** 137–142.
KEIGHTLEY, P. D., and S. A. KNOTT, 1999 Testing the correspondence between map positions of quantitative trait loci. Genet. Res. **74:** 323–328.
KELLER, L. F., K. J. JEFFERY, P. ARCESE, M. A. BEAUMONT, W. M. HOCHACHKA *et al.,* 2001 Immigration and the ephemerality of a natural population bottleneck: evidence from molecular markers. Proc. R. Soc. Lond. Ser. B **268:** 1387–1394.
KNOTT, S. A., J. M. ELSEN and C. S. HALEY, 1996 Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. Theor. Appl. Genet. **93:** 71–80.
KRUUK, L. E. B., T. H. CLUTTON-BROCK, K. E. ROSE and F. E. GUINNESS, 1999 Early determinants of lifetime reproductive success differ between the sexes in red deer. Proc. R. Soc. Lond. Ser. B **266:** 1655–1661.
KRUUK, L. E. B., T. H. CLUTTON-BROCK, J. SLATE, J. M. PEMBERTON, S. BROTHERSTONE *et al.,* 2000 Heritability of fitness in a wild mammal population. Proc. Natl. Acad. Sci. USA **97:** 698–703.
LAI, C., R. F. LYMAN, A. D. LONG, C. H. LANGLEY and T. F. MACKAY, 1994 Naturally occurring variation in bristle number and DNA polymorphisms at the scabrous locus of *Drosophila melanogaster.* Science **266:** 1697–1702.
LANDER, E., and L. KRUGLYAK, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. **11:** 241–247.
LONG, A. D., R. F. LYMAN, C. H. LANGLEY and T. F. C. MACKAY, 1998 Two sites in the *Delta* gene region contributing to naturally occurring variation in bristle number in *Drosophila melanogaster.* Genetics **149:** 999–1017.
LYNCH, M., and B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer, Sunderland, MA.
MA, R. Z., J. E. BEEVER, Y. DA, C. A. GREEN, I. RUSS *et al.,* 1996 A male linkage map of the cattle (*Bos taurus*) genome. J. Hered. **87:** 261–277.
MACKAY, T. F. C., 2001 Quantitative trait loci in *Drosophila.* Nat. Rev. Genet. **2:** 11–20.
MADDOX, J. F., K. P. DAVIES, A. M. CRAWFORD, D. J. HULME, D. VAIMAN *et al.,* 2001 An enhanced linkage map of the sheep genome comprising more than 1000 loci. Genome Res. **11:** 1275–1289.
MARSHALL, T. C., J. SLATE, L. E. B. KRUUK and J. M. PEMBERTON, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. **7:** 639–655.
MERILA, J., and B. C. SHELDON, 2000 Lifetime reproductive success and heritability in nature. Am. Nat. **155:** 301–310.
MITCHELL-OLDS, T., 1995 The molecular basis of quantitative genetic variation in natural populations. Trends Ecol. Evol. **10:** 324–328.
MOUSSEAU, T. A., and D. A. ROFF, 1987 Natural selection and the heritability of fitness components. Heredity **59:** 181–197.
NUZHDIN, S. V., E. G. PASYUKOVA, C. L. DILDA, Z-B. ZENG and T. F. MACKAY, 1997 Sex-specific quantitative trait loci affecting longevity in *Drosophila melanogaster.* Proc. Natl. Acad. Sci. USA **94:** 9734–9739.
ORR, H. A., 2001 The genetics of species differences. Trends Ecol. Evol. **16:** 343–350.
PEMBERTON, J. M., S. D. ALBON, F. E. GUINNESS and T. H. CLUTTON-BROCK, 1991 Countervailing selection in different fitness components in female red deer. Evolution **45:** 93–103.
RISCH, N. J., 2000 Searching for genetic determinants in the new millennium. Nature **405:** 847–856.
ROFF, D. A., and T. A. MOUSSEAU, 1987 Quantitative genetics and fitness: lessons from *Drosophila.* Heredity **58:** 103–118.
SEARLE, S. R., 1971 *Linear Models.* John Wiley, New York.
SEATON, G., C. S. HALEY, S. A. KNOTT, M. KEARSEY and P. M. VISSCHER, 2002 QTL Express: user-friendly software to map quantitative trait loci in outbred populations. Bioinformatics **18:** 339–340.
SELF, S. G., and K. Y. LIANG, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. **82:** 605–610.
SHOOK, D. R., A. BROOKS and T. E. JOHNSON, 1996 Mapping quantitative trait loci affecting life history traits in the nematode *Caenorhabditis elegans.* Genetics **142:** 801–817.
SLATE, J., D. W. COLTMAN, S. J. GOODMAN, I. MACLEAN, J. M. PEMBERTON *et al.,* 1998 Microsatellite loci are highly conserved in red deer (*Cervus elaphus*), sika deer (*Cervus nippon*), and Soay sheep (*Ovis aries*). Anim. Genet. **29:** 307–315.
SLATE, J., J. M. PEMBERTON and P. M. VISSCHER, 1999 Power to detect QTL in a free-living polygynous population. Heredity **83:** 327–336.
SLATE, J., T. MARSHALL and J. PEMBERTON, 2000 A retrospective assessment of the paternity inference program CERVUS. Mol. Ecol. **9:** 801–808.
SLATE, J., T. C. VAN STIJN, R. M. ANDERSON, K. M. MCEWAN, N. J. MAQBOOL *et al.,* 2002 A deer (subfamily *Cervinae*) genetic linkage map and the evolution of ruminant genomes. Genetics **160:** 1587–1597.
SMITH, T. B., R. K. WAYNE, D. J. GIRMAN and M. W. BRUFORD, 1997 A role for ecotones in generating rainforest biodiversity. Science **276:** 1855–1857.
STONE, R. T., J. W. KEELE, S. D. SHACKELFORD, S. M. KAPPES and M. KOOHMARIE, 1999 A primary screen of the bovine genome for quantitative trait loci affecting carcass and growth traits. J. Anim. Sci. **77:** 1379–1384.
VEEN, T., T. BORGE, S. GRIFFITH, G. SAETRE, S. BURES *et al.,* 2001 Hybridization and adaptive mate choice in flycatchers. Nature **411:** 45–50.
VISSCHER, P. M., R. THOMPSON and C. S. HALEY, 1996 Confidence intervals in QTL mapping by bootstrapping. Genetics **143:** 1013–1020.
VISSCHER, P. M., C. S. HALEY, S. C. HEATH, W. J. MUIR and D. H. R. BLACKWOOD, 1999 Detecting QTLs for uni- and bipolar disorder using a variance component method. Psychiatr. Genet. **9:** 75–84.
WAYNE, M. L., J. B. HACKETT, C. L. DILDA, S. V. NUZHDIN, E. G. PASYUKOVA *et al.,* 2001 Quantitative trait locus mapping of fitness-related traits in *Drosophila melanogaster.* Genet. Res. **77:** 107–116.
WILLIAMS, J. T., R. DUGGIRALA and J. BLANGERO, 1997 Statistical properties of a variance components method for quantitative trait linkage analysis in nuclear families and extended pedigrees. Genet. Epidemiol. **14:** 1065–1070.

Communicating editor: T. F. C. MACKAY

Genetic Epidemiology 25 (Supplement 1): S18–S28 (2003)

Longitudinal Data Analysis in Pedigree Studies

W. James Gauderman,[1]* Stuart Macgregor,[2] Laurent Briollais,[3] Katrina Scurrah,[4] Martin Tobin,[4] Taesung Park,[5] Dai Wang,[6] Shaoqi Rao,[7] Sally John,[8] and Shelley Bull[3]

[1]*Department of Preventive Medicine, University of Southern California, Los Angeles, California*
[2]*Institute of Cell, Animal and Population Biology, Ashworth Laboratories, University of Edingurgh, Scotland, UK*
[3]*Division of Epidemiology and Biostatistics, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada, and Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada*
[4]*Institute of Genetics and Department of Epidemiology and Public Health, University of Leicester, Leicester, England, UK*
[5]*Department of Statistics, Seoul National University, Seoul, Korea and Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania*
[6]*Division of Medical Genetics, Medical Genetics Birth Defect Center, Cedars-Sinai Research Institute, Cedars-Sinai Medical Center, Los Angeles, California*
[7]*Center for Cardiovascular Genetics, Department of Cardiovascular Medicine and Department of Molecular Cardiology, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, Ohio*
[8]*Centre for Integrated Genomic Medical Research, University of Manchester, Manchester, England, UK*

Longitudinal family studies provide a valuable resource for investigating genetic and environmental factors that influence long-term averages and changes over time in a complex trait. This paper summarizes 13 contributions to Genetic Analysis Workshop 13, which include a wide range of methods for genetic analysis of longitudinal data in families. The methods can be grouped into two basic approaches: 1) two-step modeling, in which repeated observations are first reduced to one summary statistic per subject (e.g., a mean or slope), after which this statistic is used in a standard genetic analysis, or 2) joint modeling, in which genetic and longitudinal model parameters are estimated simultaneously in a single analysis. In applications to Framingham Heart Study data, contributors collectively reported evidence for genes that affected trait mean on chromosomes 1, 2, 3, 5, 8, 9, 10, 13, and 17, but most did not find genes affecting slope. Applications to simulated data suggested that even for a gene that only affected slope, use of a mean-type statistic could provide greater power than a slope-type statistic for detecting that gene. We report on the results of a small experiment that sheds some light on this apparently paradoxical finding, and indicate how one might form a more powerful test for finding a slope-affecting gene. Several areas for future research are discussed. *Genet Epidemiol* 25 (Suppl. 1):S18–S28, 2003. © 2003 Wiley-Liss, Inc.

Key words: linkage; heritability; segregation; mixed models; correlation; variance components; Framingham Heart Study; simulation

Grant sponsor: NIH; Grant numbers: ES-10421, 5P30-ES07048.
*Correspondence to: W. James Gauderman, Ph.D., Department of Preventive Medicine, University of Southern California, 1540 Alcazar St., Suite 220, Los Angeles, CA 90089. E-mail: jimg@usc.edu
Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.10280

INTRODUCTION

Longitudinal studies provide a valuable resource for investigating factors that affect long-term averages and changes over time in a complex trait. Statistical methods that assume independence across observations (e.g., standard linear or logistic regression) are not applicable to longitudinal data, due to the correlation among multiple measurements per subject. More advanced methods were developed to handle this intrasubject correlation [summarized in Diggle et al., 1995], including generalized estimating equations and hierarchical mixed models. These models have enjoyed wide application in epidemiological studies.

Family studies are a valuable resource for investigating genetic factors that influence an outcome. As with longitudinal data, standard statistical models will be inadequate due to the nonindependence in outcomes, in this case among related individuals. In fact, methods of genetic analysis rely on the correlation among family members' outcomes to infer genetic effects. Depending on the study goals and types of data available, the analyst will utilize methods appropriate for analysis of aggregation (e.g., heritability), segregation, linkage, and/or association.

Methods for each of these types of analysis have typically been developed assuming that only one outcome value has been measured on each subject.

The Framingham Heart Study (FHS) represents a marriage of longitudinal and family study designs. The FHS data provided to the Genetic Analysis Workshop 13 (GAW13) participants include repeated measurements of several clinical outcomes (e.g., blood pressure, cholesterol) on 2,885 individuals from 330 pedigrees. Recruitment occurred in two waves, producing two cohorts of individuals within the data set. The original cohort was initiated in 1948. Clinical measurements on this cohort's subjects were scheduled every 2 years to the present, yielding as many as 21 repeated observations on some subjects. The second cohort was initiated in 1971, and included the offspring of original-cohort members. Clinical measurements on these subjects were scheduled every 4 years, yielding up to five repeated observations per subject. The FHS has been a landmark study for advancing our understanding of factors, including diet and lifestyle, that affect coronary outcomes.

Attention recently focused on the analysis of genetic factors that influence coronary outcomes in this data set. Levy et al. [2000] performed a linkage analysis of systolic blood pressure (SBP), using a panel of 399 markers spaced across the genome. They found significant evidence of linkage to a region on chromosome 17, and suggestive linkage signals on chromosomes 5 and 10. In their analysis, Levy et al. [2000] first computed a person-specific residual SBP from a model that included age and other effects, and then utilized these residuals in the program SOLAR [Almasy and Blangero, 1998] to perform a variance-components linkage analysis. The residual used in the linkage analysis for a given subject represented their long-term average SBP, after adjustment for covariates. Their paper did not consider linkage analysis of change (slope) in SBP over time.

There is a relative paucity of methods for genetic analysis of longitudinal data in families. Contributors to GAW13 have developed a wide range of approaches to help fill this gap. Included in the Group 2 contributions are aggregation, segregation, linkage, and association analysis approaches to unraveling genetic effects on both long-term averages and changes over time. Methods were applied to the FHS and to similarly structured simulated data. This paper will describe and compare methods proposed by Group 2 contributors, summarize results of applications to FHS and simulated data, and synthesize the general lessons that were learned and issues that remain.

METHODS

OVERVIEW

Thirteen papers were contributed by Group 2 participants (Table I). Seven contributors applied their methods to the FHS data, with five focusing their primary analysis on SBP and two on body mass index (BMI). Six papers analyzed the simulated data, with four focusing on SBP and two on cholesterol. Additional traits were considered in some papers. All contributions except one included some form of linkage analysis. The analytic approaches are described in some detail below.

NOTATION

We let Y_{ij} denote the measurement of trait Y obtained on subject i at calendar time j, and let T_{ij} denote the corresponding age of the subject at that time. We let X denote one or more covariates, with subscripts included as necessary to indicate whether X represents time-dependent (e.g., BMI) or time-independent (e.g., sex) variables. The methods used by Group 2 contributors can be categorized into one of two general types: a two-step approach, or a joint model approach. These are described below.

TWO-STEP MODELS

Several contributors utilized a two-step approach, consisting of a longitudinal model in the first step, followed by a second-step linkage analysis of one or more statistics derived from the first-step model.

The first-step models had the general form

$$Y_{ij}=a_i+b_i\,T_{ij}+\gamma' X_{ij}+e_{ij} \qquad (1)$$

where a_i and b_i are the subject-specific intercept and slope, respectively, and e_{ij} is a residual, assumed to be normally distributed with mean zero and variance σ^2. The slope b_i has the interpretation as the change in Y per increase of 1 year in age. The intercept a_i in this model can be interpreted as the mean of Y when T=0 (i.e., at birth) for a subject with all covariates X_{ij} equal to zero. Transformations to T or X (e.g., centering them on their means) are useful and will not affect b_i or e_{ij}, but will change the estimates and

TABLE I. Summary of data sets and analytic approaches used by Group 2

Data set	Lead author	Cohorts	Reps.	Trait[a]	Markers	Analysis approach[b]	Software[c]
Framingham	de Andrade	1 and 2	N/A	SBP	Ch. 17	L1: longitudinal VC linkage	ACT
	Barnholtz-Sloan	1	N/A	SBP	Ch. 10, 17	L1: linear mixed model; association	SAS
	Briollais	1 and 2	N/A	SPB	All	L2: linear mixed model, VC linkage	SAS/SOLAR
	Cheng	1 and 2	N/A	BMI	All	C2: linear model, VC linkage	SAS/SOLAR
	Gee	1 and 2	N/A	SBP	All	L2: linear model, segregation and linkage	SAS/GAP
	Macgregor	1 and 2	N/A	BMI	All	L1: heritability, VC linkage	SOLAR/ASREML
	Rao	2	N/A	SBP	Ch. 10	L2: principal components, HE linkage	SAS/SAGE
Simulated	Mirea	2	34	SBP	All selected	L2: linear mixed model, HE linkage	SAS/SAGE
						L1: multivariate HE linkage by GEE	SAS
	Scurrah	1 and 2	1	SBP	All	L2: linear mixed model, VC linkage	WinBUGS/Merlin
	Shephard	1 and 2	4, 10, 21	SBP	All	C2: heritability, VC linkage	Stata/SOLAR/GH
	Suh	1 and 2	10	SBP	Selected	L2: linear model, HE linkage	SAS
	Wang	1 and 2	All	Chol	Selected	L2: linear model, HE linkage	SAS/SAGE/GH
	Yang	1 and 2	8	Chol	All	L1: heritability, VC linkage	SOLAR/SAS

[a]Primary trait analyzed. In some contributions, additional traits were considered.
[b]L1, longitudinal one-step approach, with a single model that combines longitudinal and genetic analysis; L2, longitudinal two-step approach, with a first step longitudinal model and separate second step genetic analysis; C2, cross-sectional two-step approach, with a first step model of a selected time point and second step genetic analysis.
[c]GH, GENEHUNTER; GAP, Genetic Analysis Package; Ch, chromosome; Chol, cholesterol; NA, not applicable; see individual papers for descriptions of software programs and references.

interpretation of the a_i. The goal of the first-step analysis was to reduce the data to one observation per subject.

The second-step model was a genetic analysis of a person-specific statistic obtained from the first model. Since there was only one value per subject, standard modern genetic analyses were possible. These included analysis of heritability, segregation, model-free and model-based linkage, and association. For those conducting linkage analysis, most used either the variance- components (VC) approach described by Almasy and Blangero [1998] or the revised Haseman-Elston (HE) approach described by Elston et al. [2000].

Below is a brief summary of the specific approach used by each contributor of a two-step method, highlighting the differences and similarities among contributions.

Briollais et al. [2003]. This contribution expanded the first-step model in Equation (1) to include subject- and family-level models. Letting subscript f denote family, $\bar{T}_{fi}$ be the mean of observed ages for subject i, and $\bar{T}$ be the overall mean age in the sample, they used a three-level model of the form:

Level 1: $Y_{fij}=a_{fi}+b_{fi}\,(T_{fij}-\bar{T}_{fi})+c_{fi}\,(T_{fij}-\bar{T}_{fi})^2 +\gamma' X_{fij}+e_{fij}.$

Intercepts

Level 2: $a_{fi}=a_f+\phi\,(\bar{T}_{fi}-\bar{T})+\eta' X_{fi}+e_{fi}.$

Level 3: $a_f=\alpha+e_f.$

Slopes

Level 2: $b_{fi}=b_f+\omega' X_{fi}+h_{fi}.$

Level 3: $b_f=\beta+h_f.$

The intercept and slope residuals *e* and *h* at each level were assumed to have mean zero and unstructured covariance matrix. The second step was a VC linkage analysis conducted on the adjusted mean, using the sum of intercept residuals $e_{fi}+e_f$, and on the slope, using sum of slope residuals $h_{fi}+h_f$. Analyses were conducted on SBP in the FHS data set.

Gee et al. [2003]. This paper utilized the first-step regression model shown in Equation 1, applied to analysis of SBP in the FHS. In addition to the intercepts and slopes, they also derived person-specific standard errors of the intercepts (s_{ai}) and slopes (s_{bi}) from the first-stage model. For a given subject, the magnitude of these standard errors was a function of the length of follow-up, the number and age distribution of measurements during follow-up, and the intraindividual variation in measurements over time. Subjects with longer follow-up tended to have lower estimated standard errors. The second step consisted of a formal segregation analysis of the intercepts and slopes, followed by parametric (LOD score) linkage analysis. The genetic analyses of the intercepts a_i (or slopes b_i) incorporated weights based on s_{ai} (or s_{bi}). Use of these weights allowed subjects with more precise first-step regression parameter estimates to contribute more

information to second-step segregation and linkage parameter estimates.

Mirea et al. [2003]. This paper evaluated the ability to detect linkage in a genome screen using HE analysis applied to several first-step statistics, including the first SBP, last SBP, mean SBP, time-adjusted change between first and last SBP, and linear regression slope of SBP on age. Phenotypic data on Cohort 2 subjects in one replicate of simulated data were utilized, with multiple sibships extracted from the pedigrees. An alternative joint-model analysis was also considered; this approach is described later.

Scurrah et al. [2003]. These authors extended earlier work on generalized linear mixed models [Scurrah et al., 2000] to the longitudinal data setting. The approach utilized a more complex first-step model than that shown in Equation (1), including parameters for polygenic, common family environment, and common sibling environment effects on both the intercepts and slopes. The Markov chain Monte Carlo technique of Gibbs sampling was utilized to fit this model. The subject-specific polygenic residuals for intercepts and slopes were derived from their first step and used in a VC linkage analysis. The method was applied to both cohorts in Replicate 1 of the simulated data.

Wang et al. [2003]. This paper utilized all replicates of the simulated data to perform an analysis of the power to detect linkage using a variety of first-step statistics. They analyzed cholesterol and considered first-visit level, mean level, and slope (the b_i values). Both two-point and multipoint linkage analyses were conducted. They analyzed markers near true trait-causing genes (to evaluate power), and markers on a chromosome not containing any trait-causing genes (to evaluate type I error rates).

Rao et al. [2003]. This contribution focused on Cohort 2 of the FHS and performed three different types of first-step models, each followed by a second-step HE linkage analysis. The first approach repeated the analysis of Levy et al. [2000] and thus focused the second-step genetic analysis on a measure of average SBP. The second approach utilized the model in Equation (1), focusing on slopes. The third was a principal components analysis, in which five separate components estimated in the first step were each utilized in the second-step linkage analysis. The first two components corresponded roughly to the overall mean and slope of SBP and explained most of the variation in the trait, while the remaining three components captured various nonlinear trends.

Shephard et al. [2003]. This contribution utilized the first-step model in Equation 1, but with the slope on age treated as a fixed effect. In other words, the subject-specific slopes b_i were replaced by a single slope parameter β common to all subjects. Subject-specific intercepts a_i were utilized in a second-step VC linkage analysis. Using simulated data, they compared the consistency of linkage results across three separate replicates, and also compared the results to results based on simply using the first-visit value.

Cheng et al. [2003]. This paper analyzed repeated cross-sectional data, and attempted to infer trends in genetic effects across age. Measurements of BMI obtained from FHS participants in 1970, 1978, and 1986 were utilized in three separate VC linkage analyses. The first-step model was analogous to that in Equation (1) without the person-specific slope (b_i) terms. These results were compared to similar analysis using the mean BMI from these three time points.

Suh et al. [2003]. This paper used a first-step model similar to that of Levy et al. [2000], and utilized residuals from this model in an HE linkage analysis. In the linkage analysis, mixed models were used to incorporate a range of correlation structures. In the simplest model, they assumed independence for each pair, as in the standard HE approach. As alternatives, they compared two types of correlation: correlation among sib pairs sharing a common individual, and correlation among all sibs within the same family. The method was applied to SBP in the simulated data.

Collectively, these contributions demonstrated many different approaches for reducing longitudinal data to obtain person-specific statistics for genetic analysis.

JOINT MODELS

In contrast to the two-step methods described above, the goal of these contributors was to simultaneously estimate genetic and longitudinal model parameters. A joint approach is appealing because estimates of genetic and longitudinal parameters will be mutually adjusted for one

another. Additionally, effects that cross models (e.g., interactions between genetic and longitudinal parameters) are more naturally included in a joint model framework. A current limitation of joint models is the increase in computational demands relative to a two-step approach, which can limit the types of analyses that can be considered. Following is a brief summary of the work by joint-model contributors, highlighting these issues in the context of their specific approach.

de Andrade and Olswold [2003]. This paper utilizes the method described by de Andrade et al. [2002], applied to SBP in the FHS. Their mean model had the form

$$Y=\alpha+\gamma'X+a+g+s+e.$$

Here the terms a, g, s, and e represent matrices of additive polygenic, additive major gene, shared-environment, and random-environment effects, respectively. The covariance between pairs of observations was specified using variances of these random effects, with specific contributions depending on the relationship between subjects and the time at which measurements were recorded.

For example, the covariance of observations from two relatives was modeled as a function of π, the observed proportion of alleles shared identical by descent (IBD) at some marker locus, and terms that depend on whether measurements were recorded at the same or different times. No structure with respect to age or calendar year was assumed for the covariance matrix. While such an unstructured covariance matrix is appealing, the number of variance/covariance terms to be estimated grows rapidly as the number of visits increases. Because of this, they restricted each analysis to two time points and focused on chromosome 17 markers.

Yang et al. [2003] and Macgregor et al. [2003]. These two contributions used very similar approaches and will be summarized together. Both papers focused on estimating age-specific heritability across predefined intervals of age. They attempted to solve the computational difficulties alluded to above by modeling the covariance matrix as a smooth function of age. The rationale was based on the fact that repeated observations of a trait are ordered in time, and thus one might expect the variances and covariances of proximal measures to be more similar than measures widely separated in time. Both groups assumed a trait model of the form

$$Y=\alpha+\gamma'X+(a_1+a_2T+a_3T^2+\ldots)+(g_1+g_2T+g_3T^2+\ldots)+(p_1+p_2T+p_3T^2+\ldots)+f+e.$$

The random effects a, g, and e are as described above, while p and f represent permanent environmental effects and time-constant family-specific effects, respectively. Legendre polynomials were used to model the a, g, and p random effects as a function of age. Yang et al. [2003] used a mixture of cubic and linear polynomials applied to age arranged into five age bands, averaging phenotypic and covariate values for subjects with more than one measurement within an age band. Macgregor et al. [2003] used linear polynomials applied to actual adult ages (i.e., 76 bands, one for each age between the ages 20–95). Both papers estimated age-specific total heritability and age-specific heritability attributable to a specific quantitative trait locus (QTL). For the latter, they performed preliminary linkage analysis using a two-step VC approach to identify linked markers. The covariance for the major gene effects (g) was then was modeled as a function of π at a given marker to estimate QTL-specific heritability. Yang et al. [2003] analyzed total cholesterol (TC) in the simulated data, while Macgregor et al. [2003] analyzed BMI, TC, HDLC, and height in the FHS.

Mirea et al. [2003]. In contrast to the above joint-model methods, the unit of analysis in this approach was the sib pair. Focusing on selected loci, they developed an HE-type joint linkage analysis of repeated longitudinal measurements and compared this to their two-step HE approach described above. The joint analysis involved using generalized estimating equations (GEE) to account for serial correlation in repeated measures of the sib-pair trait cross-product over time, ignoring residual correlation among sib pairs within the same family. An advantage of this approach was that, once IBD estimates were obtained, the analysis was possible using standard statistical software, and gene × time or gene × age interactions were easily incorporated.

Barnholtz-Sloan et al. [2003]. This contribution was unique as it did not perform linkage analysis, but rather focused on association analysis. A preliminary association analysis was conducted using the binary trait "high SBP," defined as SBP above 140 on two consecutive visits, or reported use of hypertension treatment. A genome scan in the FHS revealed three markers showing

association to this trait. These markers were then analyzed in a joint model for SBP (in its continuous form), using mixed linear regression with random effects to account for family, sibship, and repeated measures.

RESULTS

FRAMINGHAM DATA

Not surprisingly, the various analytic approaches produced many different types of results. Rather than cover each result in detail, we summarize some of the key findings and focus on comparisons/contrasts among findings. We refer to specific marker loci by their chromosome and location in centimorgans (cM), rather than using their specific locus names.

There was much interest in chromosome 17 for SBP, given the LOD score of 4.7 (at 67 cM) observed previously by Levy et al. [2000]. de Andrade and Olswold [2002] were unable to detect any significant linkage to markers on chromosome 17 using their longitudinal VC approach. However, they also repeated the analysis of Levy et al. [2000] on these GAW data, and found a LOD score of 3.0 at position 68 cM on chromosome 17, but only when the sample was restricted to ages 25–75. Briollais et al. [2003] found evidence of linkage on chromosome 17 (62 cM), using intercept residuals in both an unselected (LOD=2.1) and selected (LOD=3.5) sample. Gee et al. [2003] did not find evidence of linkage in this specific region, but reported a modest linkage signal for intercepts to chromosome 17 (100 cM, LOD=1.5).

Evidence of genes on other chromosomes was also detected for those who analyzed SBP. Using first-step model intercepts, Gee et al. [2003] found LOD scores above 2.0 on chromosomes 1 (202 cM and 212 cM), 9 (32 cM), and 10 (125 cM). Their LOD scores were generally larger in analyses that utilized weights based on first-step standard errors, compared with not using weights. Rao et al. [2003] also found linkage evidence at position 125 cM on chromosome 10, using either mean SBP, principal components, or selected cross-sectional observations. Interestingly, Barnholtz-Sloan et al. [2003] found evidence of association (P=0.02) to a marker in this region of chromosome 10 (at 135 cM). Briollais et al. [2003] did not find linkage evidence to any marker on chromosome 10, but did report LOD scores above 2.0 for intercept residuals on chromosomes 2 (38 cM), 3 (79 cM), 8 (37 cM), and 13 (64 cM). This was the only group to also find linkage support for genes that affect SBP slope, on chromosomes 1 (212 cM), 3 (153 cM), and 11 (33 cM). Briollais et al. [2003] reported that the magnitude of their LOD scores at all these markers was quite sensitive to whether they adjusted for BMI in their first-step model. They obtained lower LOD scores in models that did not include BMI.

In two-step analyses of cross-sectional BMI observations, linkage to markers on chromosome 16 was detected by both Cheng et al. [2003] (75 cM, LOD=2.4) and Macgregor et al. [2003] (95 cM, LOD=3.1). Based on subsequent joint-model analysis, Macgregor et al. [2003] reported that the heritability attributable to a gene linked to this 95-cM marker varied substantially across the age range. Specifically, they estimated that 25% of the total variation in BMI could be attributed to this locus at age 20, but this declined to less than 5% for ages greater than 60. On the other hand, they found that a locus linked to total cholesterol (chromosome 20, 24 cM) accounted for a large proportion of variation in cholesterol across all age intervals. Cheng et al. [2003] also found linkage evidence for BMI on chromosomes 3 (181 cM), 6 (146 cM), and 9 (88 cM).

In summary, there was some agreement for genes affecting SBP on chromosomes 1, 10, and 17, and for a gene affecting BMI on chromosome 16. Linkage signals were generally higher for level-type statistics (intercepts, means, and intercept residuals), and most contributors found no evidence for genes affecting slopes. Two questions are suggested from analyses of the FHS data: 1) When are longitudinal data superior to cross-sectional data for genetic analysis? and 2) Do we have adequate power to detect slope genes? With these questions in mind, we turn to results from analyses of the simulated data.

SIMULATED DATA

There were six genes simulated to have direct effects on SBP, three with effects on baseline SBP (b34–b36), and three on slope over age (s10–s12). Slope genes s10 and s12 were simulated to be on the same chromosome.

Performing their analysis without knowledge of the answers, Mirea et al. [2003], Scurrah et al. [2003], and Shephard et al. [2003] were all able to successfully detect some of these genes by linkage analysis, each using different first-step approaches to modeling the longitudinal data. The

performance of the various methods cannot be directly compared, because each paper analyzed different replicates of the simulated data. However, some interesting trends emerged across these contributions with respect to the types of first-step statistics that showed the most significant linkage evidence.

Mirea et al. [2003] found that linkage evidence for baseline genes b34 and b35 was much more significant using visit 1 SBP than using last SBP, mean SBP, or slope of SBP. This is not surprising, given that these genes were simulated to have their effect early in follow-up. What was surprising in their results, however, was that all three slope genes were detected with greater significance using a first-step level-type statistic (e.g., mean SBP or last visit SBP) than by using a first-step slope statistic. Scurrah et al. [2003] reported analogous results. Their most significant linkage evidence was for a marker near slope genes s10 and s12, but the LOD score for this locus was much greater using a first-step intercept residual (LOD=12.9) than using a first-step slope residual (LOD=5.1). They also found suggestive linkage evidence for a marker near slope gene s11, here again using their intercept rather than slope statistic. Shephard et al. [2003] also found strong evidence of linkage near s10 and s12, using the intercepts from their first-step longitudinal model. They reported greater LOD scores using longitudinal data in the first-step model, compared to simply using first-visit SBP, even for detecting baseline genes. In analyses conducted unblinded to the answers, Suh et al. [2003] were also able to detect linkage to slope genes using level-type statistics.

Two contributors analyzed total cholesterol, which was simulated to depend on four baseline genes (b30–b33) and three slope genes (s7–s9). Without knowledge of the answers, Yang et al. [2003] were able to detect linkage to b30, b31, and b32 using visit 1 cholesterol. They were also able to detect slope gene s7, with a slightly higher LOD score using first-step mean (LOD=10.6) than first-step slope (LOD=10.3). In their joint model analysis, they found that heritability was relatively flat across age for baseline genes b30 and b32, but showed a marked increasing trend with age for s7. Wang et al. [2003] were the only contributors to analyze all 100 replicates in a true simulation study. They reported greater power for detecting the baseline genes using exam 1 cholesterol, compared to using mean or slope of cholesterol. They reported the greatest power for detecting slope gene s7 using first-step slope (80%), although first-step mean also provided reasonable power (62%) for detecting this gene. Power was low with any statistic to detect slope genes s8 and s9. Wang et al. [2003] also analyzed several unlinked markers and reported acceptable type I error rates.

Collectively, these simulated-data contributions shed some light on the questions raised by the FHS analyses. Well-selected cross-sectional data (e.g., first or last visit) provided good power for detecting some genes. However, summaries of longitudinal data (e.g., means, slopes) were generally most effective for finding genes, particularly those that affected trends in outcome over time. Somewhat paradoxical was the general finding that level-type statistics (e.g., intercept, mean) provided greater power for detecting slope genes than did slope-type statistics. We now explore this finding further.

A SMALL EXPERIMENT

We performed a small experiment to investigate the use of intercept and slope statistics for detecting a slope gene. We simulated a sample of 1,000 independent individuals. Each individual was randomly assigned a genotype (G) at a slope-affecting locus, with probability 50% each of carrying a normal (G=0) or variant (G=1) genotype. Age (T) was also randomly generated for each subject from a uniform distribution on the range 0–50 years. Conditional on G and T, the trait Y was randomly sampled from a normal distribution with mean $100+\beta^*G^*T$ and variance σ^2. It is clear that under this model, the gene G has no baseline effect, but rather only affects slope.

Can we detect this slope gene with more power using a test based on slope or mean statistics? We first fit a linear model of the form

$$Y=\alpha+\beta_1 G+\beta_2 T+\beta_3 G^*T+e \qquad (2)$$

where e was assumed to be normally distributed with mean 0 and variance α^2. The parameter β_3 quantifies the difference in slopes between G=0 and G=1, and the estimated slope β_3 can be used to form a slope-based test of the form $t=\beta_3/\mathrm{se}(\beta_3)$. We then considered a model of the form

$$Y=\alpha+\beta_1 G+e$$

where the parameter β_1 measures the difference in mean Y between genotype groups, with corresponding mean-based test $t=\beta_1/\mathrm{se}(\beta_1)$.

For three different settings of σ (1, 8, and 32, respectively), Figure 1 plots simulated Y vs. T

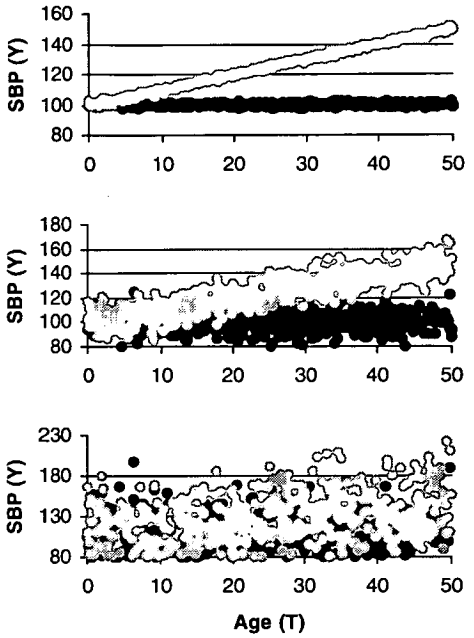


Fig. 1. Simulated SBP, based on model in equation (2), assuming β=1, with σ=1.0 (top), σ=8.0 (center), or σ=32 (bottom). Gray dots have variant genotype (G=1); black dots have G=0.

by genotype G for 1,000 subjects, when β is set to 1.0. The difference in slope of Y on T between G=0 and G=1 is clear when σ=1, but becomes less obvious as σ increases.

Table II gives the expected *t*-statistic for the slope and mean-based tests when the true β=1.0 (as in Fig. 1), and for a larger slope effect (β=3.0). When σ=1, the *t*-statistic (and thus power) is much larger for the slope test than for the mean test. However, as the variance increases, the power of the slope test is dramatically reduced, while the power of the mean test is much less affected. When σ=32, the mean test is more powerful than the slope test both for β=1 and β=3.

The conclusion one can draw from this experiment is that when the residual variance is large, as it is for traits in both the FHS and simulated data, a test based on means can provide greater power to detect a slope-affecting gene than a test based on slopes alone. In practice, many additional factors will determine the relative power of a mean-based to slope-based test, including not only the underlying true effect sizes, but the number of repeated observations and the length of follow-up. Also important will be the relative magnitude of the within- and between-subjects variance of Y.

TABLE II. Expected *t*-statistics for mean- and slope-based tests

		Expected *t*-statistics	
σ	β	Mean test	Slope test
1	1	38.2	229.3
	3	38.4	687.7
8	1	29.9	28.7
	3	37.0	86.0
32	1	11.0	7.2
	3	26.2	21.5

DISCUSSION

Complex traits such as SBP and cholesterol vary with age and likely depend on both genetic and environmental determinants. For such traits, longitudinal data allow one to disentangle genetic and environmental effects, both on the rate of change of the phenotype over time (e.g., slope) and on trait level (e.g., mean). Unlike the FHS, most family studies collect a single cross-sectional measurement on each subject. While this type of data can also be used to analyze mean and slope effects, estimates will be more prone to bias from confounding and more affected by measurement error.

How does the current value of an age-dependent trait depend on genotype? For simplicity, consider two groups of subjects, carriers (C) and noncarriers (N), respectively, of a variant allele at a particular locus. Differences in expected trait value between C and N groups at age T will be a function of their difference at birth plus any difference that accrues between birth and T. There are four possible scenarios: 1) G has no affect at birth or thereafter, 2) G only affects level at birth, 3) G has no affect at birth, but affects development, and 4) G affects both level at birth and development. Without knowing the truth, the analyst is faced with choosing the test statistic that provides the greatest power to detect G.

Although the best statistic to choose will depend on the true situation, these GAW contributions shed some light on the relative robustness of different alternatives. Obviously, any statistic needs to have the correct test size when situation 1 holds. For situation 2, equivalent to a baseline gene in the GAW simulation, only level-type statistics (e.g., mean or cross-sectional value)

provide power. This makes sense, since there is no difference between the C and N groups in slope. When the gene does affect slope (situation 3 or 4), statistics based on slope or change in level over time can be used. However, several contributions and the small experiment indicated that a mean-based statistic can often provide greater power for finding a slope gene than a slope-based statistic.

The reason that a mean-based statistic has any power to detect a slope gene is that a slope gene will typically lead to a difference in the mean of the trait by genotype. This can be seen in Figure 1, for example, where the difference in means is approximately the difference in genotype-specific linear predictions at the midpoint of age (T=25). A notable exception will occur if genotype-specific baseline means are different *and* one slope is positive while the other is negative (graphically, an X-shape rather than the sideways V-shape shown in Fig. 1). However, such an X-shaped relationship is unlikely for most biological systems.

When a slope gene does affect both slope and mean, neither the mean- nor slope-based statistics used by many contributors will be optimal for finding genes that affect rate of change. On the basis of the model in Equation (2), the null hypothesis we should be interested in for such a slope gene is $\beta_1=0$ (no level effect) *and* $\beta_3=0$ (no slope effect). A two-degree-of-freedom likelihood ratio test comparing the likelihood at the joint maximum likelihood estimate (MLE) of β_1 and β_3 to the likelihood with both fixed to zero would be appropriate. This type of test is analogous to previously proposed joint tests in the context of using gene × covariate interaction information to improve power for detecting linkage [Greenwood and Bull, 1999; Olson, 1999; Gauderman and Siegmund, 2000; Gauderman et al., 2001]. In fact, one can think of the slope parameter β_3 as a measure of gene × covariate interaction, in this case with age being the covariate.

Careful consideration of covariates will be essential for understanding both environmental and genetic (through G × E interaction) determinants of complex traits. The current value of an age-dependent trait will likely depend on both current and previous values of environmental covariates. There are several ways covariate information can be included in a model. One can include time-varying covariate values, e.g., smoking status at each visit, directly into a multilevel or joint model. An alternative approach is to incorporate cumulative exposure through a single covariate, e.g., total number of pack years of smoking. One may choose to focus on exposure during a critical period of life, e.g., in utero or early-life exposure to parental smoking. More complicated covariates can also be constructed, e.g., allowing current covariate effects to be modified by previous exposure levels or by genotype. Of course, all these methods depend on the availability of reliable covariate data, which is more likely to derive from longitudinal rather than cross-sectional studies.

In terms of modeling approaches, contributors to this group adopted either a two-step or joint model for the genetic analysis of longitudinal data. In general, a joint model should be preferable for two main reasons. First, parameter estimates in the longitudinal and genetic models are mutually adjusted for one another. Second, a joint model correctly accounts for within-individual and between-individual variability, so that uncertainty in the estimated phenotype (e.g., person-specific intercept or slope) is accounted for during the linkage analysis. While one can weight first-step summary statistics to account for the relative degree of within- and between-subject variance [Gee et al., 2003], such weighting comes about naturally in a joint model.

While a joint modeling approach has theoretical advantages, the two-step approach is attractive for practical (computational) reasons. First-step longitudinal models can be fit using standard statistical software packages (e.g., SAS, SPLUS, and STATA). Once subject-specific summary statistics are abstracted from this first step, a number of available programs can be used for linkage, heritability, or segregation analysis. Commonly used genetic software programs are not designed for longitudinal data analysis, and there is a clear need to develop integrated programs. Regardless of whether a two-step or joint approach is adopted, the analyst should always carefully consider model assumptions, e.g., normality and homoscedascity, since violations can lead to invalid conclusions.

Multilevel modeling, which can take into account the hierarchical structure of the data, may help disentangle the proportion of the trait variability explained by fundamental variation in the mean trait and in the trait slope from the proportion explained by random within-individual variability. Joint modeling in the multilevel model framework is theoretically possible. As an example, the multilevel model of Briollais et al.

[2003] can be expressed as a single mixed model, with the form

$$Y_{fij}=\alpha+\phi(\bar{T}_{fi}-\bar{T})+\beta(T_{fij}-\bar{T}_{fi})+\gamma'X_{fij}+\eta'X_{fi}+\omega'X_{fi} \times (T_{fij}-\bar{T}_{fi})+e_f+e_{fi}+e_{fij}+(h_f+h_{fi})(T_{fij}-\bar{T}_{fi}).$$

This model is easily extended to include additional levels (e.g., sibships within family), with corresponding covariates and random effects. The variance-covariance matrix of the random effects (*e* and *h* values) can be expressed as a function of marker-IBD sharing probabilities among relatives, thus facilitating a test of linkage on intercepts and/or slopes. One could also include a marker genotype as a covariate in the above model, thus also providing tests and estimates of association on trait level and/or slope. This type of model generalizes the hierarchical modeling structure described by Fulker et al. [1999] in the context of cross-sectional data.

In population studies of blood pressure, a significant proportion of blood pressure observations will be affected by hypertensive treatment (HRX). Levy et al. [2000] reported that 15.3% of observations reflected HRX in the FHS. In such observations, measured SBP will be lower than the "true" untreated SBP, which will impact estimates of genetic and environmental effects. Members of this group utilized various methods of accounting for this problem. These include ignoring the problem completely, excluding individuals on treatment, including HRX as a covariate, replacing the phenotypes of all individuals on HRX with a single high value, adding a constant (an average HRX effect) to observations on treatment, and imputing post-HRX SBP based on pre-HRX measurements and/or or the SBP of other family members. Some of these approaches will produce biased results, and the extent of the bias is likely to depend on the proportion of individuals on treatment and the actual effects of treatment on those individuals. The advantages and disadvantages of each approach will not be discussed here, and we do not aim to recommend a single best approach, as the problem is still being researched [e.g., Cui et al., 2003]. However, the results of any linkage analysis for such phenotypes will depend on the way in which treatment has been accounted for, and it is an issue that should be considered in population-based studies such as the FHS.

Another important issue in longitudinal studies is that of missing data. All of the contributions in this group ignored the problem of missing data, focusing their analyses on observations with complete outcome and covariate data. It is well-known that the elimination of missing observations can lead to bias if data are not missing completely at random (MCAR), and particularly if there is informative missingness [Little and Rubin, 2002]. An example of informative missingness is cohort dilution, e.g., the elimination of subjects at later ages from the cohort in a nonrandom way with respect to trait genotype. In some situations, one may need to specify a joint model of both the phenotype and the missingness process. This type of analysis was used to model survival and quality-of-life data in cancer patients, when quality of life was not missing at random [Billingham et al., 2001]. Some approaches to dealing with missing data in the FHS have been developed [Badzioch et al., 2003], but this important topic needs further statistical attention.

Understanding the magnitude of within- and between-subject variability in a trait is important in designing a longitudinal family study. When intrasubject variability in a trait is high (as was observed for SBP in the FHS), precision will be increased by having many repeated measurements per subject. On the other hand, when intrasubject variability is low, power will be greater by increasing the number of individuals rather than by increasing the number of measurements. This adds a level of complexity to the design of family studies, for which one also has to consider within- and between-family trait variability. In addition, practical considerations (e.g., stability of the population over time, cost of obtaining measurements) will play heavily in the design of a longitudinal family study.

In conclusion, this group proposed, applied, and evaluated several approaches to the analysis of longitudinal family data. Collectively, our findings confirmed some of those previously reported by Levy et al. [2000], and indicated some additional chromosomal locations that may warrant further investigation. From a methodological standpoint, we described several variations of two-step and joint modeling approaches. Across many different approaches, we found that the use of a mean-based statistic is likely to provide more power for detecting a slope-affecting gene than a slope-based statistic. This finding warrants further study. Also an important topic for future research is the development of models that integrate the estimation of genetic and longitudinal parameters, along with associated software for fitting the models. We encourage readers to see the individual contributions to learn more about each specific method.

REFERENCES

Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211.

Badzioch MD, Thomas DC, Jarvik GP. 2003. Summary report: missing data and pedigree and genotyping errors. Genet Epidemiol 25 (Suppl. 1):S36–S42 (this issue).

Barnholtz-Sloan JS, Poisson LM, Coon SW, Chase GA, Rybicki BA. 2003. Analysis of gene × environment interaction in sibships using mixed models. BMC Genet [Suppl] 4:18.

Billingham LJ, Abrams KR, Jones DR. 2001. Simultaneous assessment of quality of life and survival data. In: Stevens A, Abrams KR, Brazier JE, Fitzpatrick R, Lilford R, editors. Advanced handbook of methods in evidence based healthcare. London: Sage Publications. Chap 20, p 352–366.

Briollais L, Tzontcheva A, Bull S. 2003. Multilevel modeling for the analysis of longitudinal blood pressure data in the Framingham Heart Study pedigrees. BMC Genet [Suppl] 4:19.

Cheng R, Park N, Hodge SE, Juo S-HH. 2003. Comparison of the linkage results of two phenotypic constructs from longitudinal data in the Framingham Heart Study: analyses on data measured at three time points and on the average of three measurements. BMC Genet [Suppl] 4:20.

Cui JS, Hopper JL, Harrap SB. 2003. Antihypertensive treatments obscure familial contributions to blood pressure variation. Hypertension 41:207–210.

de Andrade M, Olswold C. 2003. Comparison of longitudinal variance components and regression-based approach for linkage detection on chromosome 17 for systolic blood pressure. BMC Genet [Suppl] 4:17.

de Andrade M, Gueguen R, Visvikis S, Sass C, Siest C, Amos CI. 2002. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. Genet Epidemiol 22:221–232.

Diggle PJ, Liang KY, Zeger SL. 1995. Analysis of longitudinal data. Oxford: Clarendon Press.

Elston RC, Buxbaum S, Jacobs KB, Olson JM. 2000. Haseman and Elston revisited. Genet Epidemiol 19:1–17.

Fulker D, Cherny S, Sham P, Hewitt J. 1999. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259–267.

Gauderman W, Siegmund K. 2000. Gene-environment interaction and affected sib pair linkage analysis. Hum Hered 52:34–46.

Gauderman W, Morrison J, Siegmund K. 2001. Should we consider gene × environment interaction in the hunt for quantitative trait loci? Genet Epidemiol 21: 831–836.

Gee C, Morrison JL, Thomas DC, Gauderman WJ. 2003. Segregation and linkage analysis for longitudinal measurements of a quantitative trait. BMC Genet [Suppl] 4:21.

Greenwood CMT, Bull SB. 1999. Analysis of affected sib pairs, with covariates—with and without constraints. Am J Hum Genet 64:871–885.

Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH. 2000. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. Hypertension 36:477–483.

Little RJA, Rubin DB. 2002. Statistical analysis with missing data. New York: John Wiley & Sons, Inc.

Macgregor S, Knott S, White I, Visscher P. 2003. Longitudinal variance-components analysis of the Framingham Heart Study data. BMC Genet [Suppl] 4:22.

Mirea L, Bull SB, Stafford J. 2003. Comparison of Haseman-Elston regression analyses using single, summary, and longitudinal measures of systolic blood pressure. BMC Genet [Suppl] 4:23.

Olson JM. 1999. A general conditional-logistic model for affected-relative-pair linkage. Am J Hum Genet 65:1760–1769.

Rao S, Li L, Li X, Moser KL, Guo Z, Shen G, Cannata R, Zirzow E, Topol EJ, Wang Q. 2003. Genetic linkage analysis of longitudinal hypertension phenotypes using three summary measures. BMC Genet [Suppl] 4:24.

Scurrah KJ, Palmer L, Burton P. 2000. Variance components analysis for pedigree-based censored survival data using general linear mixed models (GLMMs) and Gibbs sampling in BUGS. Genet Epidemiol 19:127–148.

Scurrah KJ, Tobin MD, Burton PR. 2003. Longitudinal variance-components models for systolic blood pressure, fitted using Gibbs sampling. BMC Genet [Suppl] 4:25.

Shephard N, Falcaro M, Zeggini E, Chapman P, Hinks A, Barton A, Worthington J, Pickles A, John S. 2003. Linkage analysis of cross-sectional and longitudinally derived phenotypic measures to identify loci influencing blood pressure. BMC Genet [Suppl] 4:26.

Suh YJ, Park T, Cheong SY. 2003. Linkage analysis of longitudinal data. BMC Genet [Suppl] 4:27.

Wang D, Li X, Lin Y-C, Yang K, Guo X, Yang H. 2003. Power of linkage analysis using traits generated from the simulated longitudinal data of the Framingham Heart Study. BMC Genet [Suppl] 4:28.

Yang Q, Chazaro I, Cui J, Guo C-Y, Demissie S, Larson M, Atwood LD, Cupples LA, DeStefano AL. 2003. Genetic analyses of longitudinal phenotype data: a comparison of univariate methods and a multivariate approach. BMC Genet [Suppl] 4:29.