ANALYSIS OF HIGH-DENSITY SNP DATA FROM COMPLEX POPULATIONS

James A.B. Floyd

PhD Thesis
The University of Edinburgh
2010

Abstract

Data from a Croatian isolate population are analysed in a genome-wide association study (GWAS) for a variety of disease-related quantitative traits. A novel genome-wide approach to analysing pedigree-based association data called GRAMMAR is utilised. One of the significant findings, for uric acid, is followed up in greater detail, and is replicated in another isolate population, from Orkney. The associated SNPs are located in the *SLC2A9* gene, coding for a known glucose transporter, which leads to identification of *SLC2A9* as a urate transporter too (Vitart et al., 2008). These SNPs are later implicated in affecting gout, a disease known to be linked with high serum uric acid levels, in an independent study (Dehghan et al., 2008). Subsequently, investigation into different ways in which to use SNP data to identify quantitative trait loci (QTL) for genome-wide association (GWA) studies is performed. Several multi-marker approaches are compared to single SNP analysis using simulated phenotypes and real genotype data, and results show that for rare variants haplotype analysis is the most effective method of detection. Finally, the multi-marker methods are compared with single SNP analysis on the real uric acid data. Interpretation of real data results was complicated due to low sample size, since only founder and unrelated individuals may be used for population-based haplotype analysis, nonetheless, results of the prior analyses of simulated data indicate that multi-marker methods, in particular haplotypes, may greatly facilitate detection of QTL with low minor allele frequency in GWA studies.

Declaration

I declare that this Thesis was composed by myself and the research presented is my own except where otherwise stated. This work has not been submitted for any other degree or professional qualification except as specified.

James Floyd

24/09/2010

Acknowledgements

First and foremost, I would like to take this opportunity to express my gratitude to my supervisors Prof. Chris Haley and Dr. Sara Knott for all their help and advice throughout this project. In addition to Chris and Sara, there are a number of other people who I would like to thank for their part in helping me while I was working on my Thesis. On the academic side were Chris Franklin and Joseph Powell, two fellow PhD students who I discussed ideas and results with, and Martin Jones, whose help while I was learning Perl was invaluable. On a more personal note, I'd like to thank my parents, Jim and Judy Floyd, and my sister Laura Floyd, for their support during this project. Additionally, I would like to give special thanks to Lydia Saunders, my long suffering girlfriend, who helped keep me going to the end, and who, at long last, will not have to sacrifice any more holiday plans because I'm too busy to go away.

Table of Contents

ANALYSIS OF HIGH-DENSITY SNP DATA FROM COMPLEX POPULATIONS

1. CHAPTER 1 - INTRODUCTION

1.1 Early Molecular Markers and Correlating Genotype with Phenotype

"The search for … linkage will certainly be lengthy, and at first, disappointing." – R. A. Fisher (1935).

The concept of using molecular markers to correlate visible phenotypes with regions of the genome has a rich and diverse history. Indeed, before molecular markers were even known, phenotypes themselves were noted to be linked. The theory underpinning the key concept of linkage has remained broadly constant, yet the methods employed in achieving end results have changed beyond all recognition, particularly in the more recent past. It has been known since the work of Gregor Mendel (or at least, since its rediscovery by Hugo de Vries in 1900) that differences in hereditary material could be responsible for variation in the physical appearances of natural organisms. Before anything was known of genes, or even the structure of genomes other than the existence of chromosomes, Bateson reported "coupling" between traits in the sweet pea (Bateson, 1905), representing the first example of linkage reported in any organism (Edwards, 2005).

This was followed by a conceptual breakthrough in 1911 when Morgan hypothesised that if the "factors" responsible for coupling of traits were on chromosomes, then not only would a linear arrangement of these factors on the same chromosome explain coupling, but also that the strength of linkage would be determined by their mutual proximity (Morgan, 1911). At this stage, all that was lacking was a rigorous statistical framework with which to begin linkage estimation, and this was provided by Ronald Fisher in 1922, with his introduction of the theory of maximum likelihood (Fisher, 1922). Since that time, there has been ever more focus on associating traits of interest first with one another, and later with molecular markers, in an attempt to map the genome. This is evident in the works of, for example, Bell and Haldane (1937) and Lawler and Renwick (1959), who used blood-group systems as molecular markers to infer linkage to certain diseases. These blood-group systems, such as the Lutheran, Rhesus and ABO systems, were the first true molecular markers used to assess linkage – however this is still one step removed from using the actual DNA polymorphisms themselves, as is commonplace today.

It wasn't until after the discovery of the structure of DNA as the unit of heredity in 1953, and the vast amounts of subsequent research into DNA, that linkage analysis was at last applied to human genetic variation and phenotypes. The first type of sequence-based molecular marker to be identified was restriction fragment length polymorphism (RFLP). RFLPs flag polymorphisms in restriction enzyme recognition sites because they result in differential lengths of DNA product, which are then separable when run on a gel. They were particularly useful due to their property of being co-dominant, where heterozygotes are distinguishable from both homozygotes.

RFLPs were first used as a tool for genetic analysis in 1974, and were used to create the first human genetic maps in 1980 (Botstein et al., 1980). They were also used in early attempts at what can be seen as more genome-wide approaches to genetic mapping (Lander and Botstein, 1988).

A number of other DNA-based markers, such as minisatellites, amplified fragment length polymorphisms (AFLPs) and random amplification of polymorphic DNA (RAPDs) became available with the advent of the Polymerase Chain Reaction (PCR) in 1983. AFLPs and RAPDs were favoured fleetingly due to ease of use in the laboratory, and although AFLPs are still used in plant studies (Woodhead et al., 2005), their use in human mapping in particular was compromised by virtue of them being dominant rather than co-dominant markers (Vignal et al., 2002). Therefore, the next major class of marker that was widely utilised in genetic mapping was microsatellites: these are hypervariable, co-dominant markers consisting of differing numbers of short (usually less than 100 bases) tandemly repeated sequences, easily assayed using PCR (Sunnucks, 2000). Microsatellites are also numerous and fairly evenly distributed across the entire genome, therefore it wasn't long before human genetic maps were constructed using this new type of marker (Schlotterer, 2004).

Microsatellites became very popular for assessing linkages of disease with specific regions of the genome. Genome-wide scans consisting of 300-400 microsatellites at an average spacing of around 10cM became the norm, and several statistically significant linkages were found – particularly for Mendelian (monogenic) disorders. However, successes for more complex diseases or traits were far more modest, and

the time and cost required for microsatellite arrays made larger studies unfeasible. At this point attention turned to a new type of molecular marker, a marker which has heralded an unprecedented level of understanding of the human genome and of the genetics behind complex traits: single nucleotide polymorphisms.

1.2 Single Nucleotide Polymorphisms

As the name may imply, single nucleotide polymorphisms (or SNPs) refer to a single base in the genome which exhibits variation from one individual in a population to another. In general, SNPs are classed as common or rare depending on the frequency of the minor allele in the population. In general, the minor allele frequency (MAF) must be greater than 5% for a SNP to be classed as common, although this varies in the literature, and is often relaxed to just 1%. While in theory any one of the four possible bases A, T, C and G could be present at any one locus, in practice there is usually only one of two bases at a SNP locus, hence SNPs are regarded as biallelic markers. This is due both to the very low mutation rate, and the fact that mutations are heavily biased towards transitions (A <-> G or C <-> T) as opposed to transversions (any other single base substitution) (Vignal et al., 2002). While it may seem counterintuitive to replace the highly polymorphic, and therefore vastly more informative (on a per marker basis), microsatellites with biallelic SNPs to assess linkage, there are a number of advantages which SNPs have that more than make up for the loss in information content.

From a practical point of view, SNPs are extremely well suited to automation, and therefore fast high-throughput techniques are possible. This in turn makes genotyping much less expensive and more cost-effective. In a comparison of microsatellites and SNPs, the time taken to perform genome-wide genotyping for analysis decreased from many months to just a few weeks (John et al., 2004). Since this comparison was made, SNP genotyping has improved by many orders of magnitude both in terms of speed and cost-efficiency.

Other than the increase in genotyping speed, there is another major benefit of using SNPs. Even when compared to microsatellites, SNPs are vastly more abundant in the genome. Most recent estimates suggest there are somewhere around seven million SNPs in the human genome with a minor allele frequency of at least 5%, and a further four million or so with minor allele frequencies between 1% and 5% (Frazer et al., 2009). It is this massive increase in the number of SNPs compared to microsatellites that entirely negates the loss of information content when comparing on a per marker basis, because information content incorporates both level of polymorphism at a locus (heterozygosity) and marker density. For this reason, SNPs provide higher overall information content across the genome than do microsatellites.

In a comparison of information content provided by SNPs and microsatellites, it was shown that for realistic densities of each type of marker, SNPs are considerably more informative than microsatellites (Kruglyak, 1997). The information content afforded by microsatellites spaced 10cM apart (the usual distance for most linkage studies) is 0.68 with 10 alleles each of equal frequency (the best possible scenario), but for SNPs

spaced 1cM apart this value is 0.88. Importantly, even at more extreme allele frequencies (where heterozygosity will fall, thus reducing information), the results are not changed drastically for SNPs. In general, to provide the same information content, a SNP marker map would need to be around 2.5 times as dense as a microsatellite map, which means as little as 750 – 1,000 SNP markers.

With an estimated 11 million SNPs in the human genome with MAF >1%, human genetics is almost spoilt for choice regarding which of these polymorphisms to use in searching for disease genes. The majority of studies in recent years have been based upon 300,000 – 500,000 genome-wide SNPs selected either randomly, or as will be seen shortly, by more intelligent SNP selection criteria, but this number is set to increase further to a million or even more. It is armed with this plethora of newly discovered DNA variation to explore that the human genetics community has real ambitions to dissect common complex disease.

1.3 Common Complex Disease and Quantitative Traits

One of the main driving forces behind the development of new markers, and the methods to test them for linkage, is to try to understand the genetics behind common complex diseases in the human population. Most methods were designed explicitly for the detection of disease genes, and consequently in this chapter "disease" is often used where in fact any given trait could be the phenotype under study. The vast majority of diseases studied to date, particularly those studied at the conception of linkage analysis design, essentially comprise a binary presence / absence (or affected

/ unaffected) trait; that is, a qualitative trait composed of discrete categories. One of the earliest and most renowned examples of linkage to a common disease was to cystic fibrosis in 1989 (Kerem et al., 1989), and there has been numerous reports of success in locating genes involved in other diseases since then. While this is encouraging, one thing most of these success stories have in common is that they generally involve diseases which turn out to have only one gene explaining all or nearly all of the cases (i.e., Mendelian disorders), or where there were rare variants segregating within single families which conferred a high risk to the given disease (Botstein and Risch, 2003). There are much fewer successes regarding diseases which appear to have a more complex aetiology, the so-called common complex diseases such as diabetes, cancers, schizophrenia, hypertension and obesity-related diseases.

Much of the difficulty in being able to detect linkages to genes influencing common complex diseases can be attributed to the fact that in reality many of these diseases are not under the control of a single gene. One way in which this could happen would be for a disease to be under the control of many genes that are all fully penetrant, therefore any one of a number of genes would be causal. Another possibility is that many genes are involved in a given disease, but that none are directly responsible and individually contribute only a fraction of the overall disease risk. This idea may be more readily obvious in diseases where diagnosis is based on a threshold for example, where many genes may each increment the value by a small amount. Although the clinical endpoint of disease is the binary outcome of disease status (i.e., affected or unaffected), all common complex diseases are under the control of numerous genes, each predisposing to disease but not necessarily requisite. Additionally, many

diseases can be further divided into the products of numerous risk factors, many of which are quantitative in nature. A quantitative trait is one that is measured on a continuous, i.e., numerical, scale, rather than having categories into which all samples can be assigned, and genes influencing quantitative traits are known as quantitative trait loci (QTL).

When a disease is referred to as "complex" it usually means the disease is thought to be composed of multiple components, a proportion of which will be genetic. In general, complex diseases can be broken down into a genetic component resulting from the action of many genes each of small individual effect, and a non-genetic, or environmental, component. It is also possible for there to be interactions both within and between these components. This means that, excluding any rare high risk alleles, any one risk allele is neither necessary nor sufficient to cause a disease; it is the cumulative effect of many alleles and interacting factors that is most important (Bourgain and Genin, 2005). In addition to this, the set of all possible genes that could affect any given complex disease may include genes involved in numerous different underlying pathways or risk factors. This fact somewhat undermines the attempt to classify and analyse complex diseases in the simple binary manner that has been so successful for Mendelian traits: how can the individual causes of a complex disease be teased apart when there are so many different ways the disease can develop, and furthermore, when a large proportion of these diseases and their potential causes are being analysed as a binary outcome when they are in fact quantitative in nature?

One way to help tackle this issue is to decompose diseases into a set of likely risk factors and analyse these individually, in an attempt to increase the genetic signal to noise ratio. The idea behind this is that while a disease may be extremely heterogeneous in its causes, a genetic risk factor conferring susceptibility to disease is more likely to be caused by the same genes across the population. That is, across a population, a disease may develop as a consequence of some combinations of risk factor genes, however each risk factor itself should be determined by the same set of genes. Since many of these factors – referred to as intermediate phenotypes – are quantitative, analysing them as such can increase the power to detect genes affecting them. Even so, genes influencing quantitative, as opposed to Mendelian, traits are more difficult to detect since there are more of them, and by definition each one must explain a smaller proportion of variation in the trait than if the trait was Mendelian. In general, the more genes there are involved in the control of a trait, the less variation each one individually explains.

There is a current theory building on this idea that common complex diseases are controlled by many QTL. This theory suggests that the majority of QTL involved in controlling common complex disease are common (>5% minor allele frequency) in the population, are numerous and have modest effect sizes. The hypothesis states that because genes conferring only a modest increased risk to a given disease are unlikely to experience strong purifying selection, they act in a largely evolutionarily neutral manner which allows them to reach a moderate frequency in the population. More formally, this theory is known as the common disease / common variant (CD/CV) hypothesis (Reich and Lander, 2001), and while not universally accepted, there is

evidence to suggest common disease risk alleles conferring small and modest effect sizes do segregate in populations (Lohmueller et al., 2003). However, one alternative to this theory is that while there are indeed many alleles of relatively small effect segregating for a given trait, they are very rare at each trait-related locus. This would make detection of such variants much more difficult.

As previously stated, where there have been successes in mapping genes involved in complex disease, the genes uncovered have typically been rare high risk variants. This means that at the population level these alleles explain little of the total disease prevalence. In the context of public health, this can be expressed conveniently as the population attributable fraction (PAF), which can be thought of as the proportion of the disease that would be eliminated if the risk factor were removed. The PAF for rare high risk alleles in complex diseases is normally less than 10%, for example over 150 rare high risk alleles have been identified for Alzheimer's disease, but the combined PAF for all of these is less than 5% (Carlson et al., 2004). It is thought that common modest risk variants may contribute a much greater PAF to common complex disease than rare high risk alleles due to their higher frequency in the population (Risch and Merikangas, 1996). This explains why these genes are so important to find from a public health perspective, and why small to modest effect disease genes are what geneticists are now trying to map. With recent advances in both theoretical and technological fields, the potential for detecting such genes has never been better. Many different methods have been employed in the attempt to do so.

1.4 Linkage Analysis

As previously noted, linkage studies to determine the location of trait-affecting genes are not new. However, after the discovery of microsatellites and SNPs, and the ability to rapidly and efficiently score them, linkage scans became far more commonplace. As a consequence of their increasing popularity, there was also a great increase in the number and type of study design and the testing used to detect possible linkages. Initially, linkage was performed on binary phenotypes using family-based methods designed to trace the inheritance of specific trait-related alleles through a pedigree, and these methods typically required an explicit model of inheritance to be specified. Subsequently, less stringent tests were formulated, and later, tests which searched for linkage with quantitative traits, and which operated at a population level.

The common uniting feature amongst all these tests for linkage is the underlying assumption that the same disease-related allele is physically linked to the same marker allele in all "affected" individuals within the same family (or in the case of quantitative traits, in all related individuals occupying one extreme of the phenotypic distribution). Note that for most methods the specific marker allele linked to the disease predisposing allele may be different across families, so long as within families it is the same. This implicitly assumes that the trait and marker loci are closely linked enough that recombination between them is rare. Indeed, this is a crucial aspect of linkage studies, as no linkages can be detected if there are too many recombination events between the marker and trait loci.

1.4.1 Binary traits and parametric linkage analysis

Linkage analysis was initially only widely used for analysis of binary traits in the form of diseases in families within which the disease was segregating. To test for the presence of linkage, segregation of disease is correlated with markers of known location throughout the genome, hence determining an approximate region containing a disease locus. Classically, evidence for linkage is assessed via the lod-score, which requires specifying a mode of inheritance (i.e., whether the trait is inherited in a dominant or recessive manner) and estimating the recombination fraction between marker and trait loci using recombination events observed in the data. The recombination fraction, θ, is typically small when loci are physically close to one another and there are therefore very few recombination events (tending towards zero as the frequency of recombination drops), and is large (up to a maximum of 0.5) when the loci are a large distance apart. A value of 0.5 indicates the loci are acting statistically independently of one another, which is typically the case for loci many mega-bases (Mb) apart or on different chromosomes.

Since this method requires the definition of mode of inheritance and specification of a number of parameters, it is often referred to as parametric (or model-based) linkage analysis. Other than the mode of inheritance of the trait, some of the other parameters needing specification are the trait and marker allele frequencies in the population and the penetrance of the trait. Penetrance refers to the probability that an individual exhibits the trait phenotype given a trait-causing genotype at the locus. Apparent incomplete penetrance (when the probability of exhibiting a phenotype is less than one despite carrying the susceptible genotype) can occur due to a number of reasons,

although some of these reasons are recognised as unique phenomena themselves. Trait heterogeneity is one of these; if several different combinations of disease genes can all be responsible for causing a disease but no single gene alone, then individuals can display the trait phenotype without having the same genotypes at any given trait locus. Trait heterogeneity can also cause other problems for linkage studies if each gene can itself be entirely responsible for causing disease. In this case the penetrance of each disease gene is in fact one, although not every individual exhibiting the disease will possess the same genotype at any one of the disease loci. There is also the related phenomenon of phenocopies, whereby individuals (even within the same family) may exhibit the same trait due to different, and not necessarily genetic, causes (Forabosco et al., 2005).

After specification of the model the recombination fraction is estimated using maximum likelihood, and the value of the recombination fraction that maximises the likelihood is given as the best estimate of its true value. Significance is determined by the maximum lod-score, which is the $\log_{10}$ ratio of the maximum likelihood score and the likelihood score at $\theta = 0.5$ (corresponding to the null hypothesis of no linkage). Traditionally, a lod-score of 3.3 is used to demonstrate significant linkage at the genome-wide level, corresponding to a p-value more extreme than 0.0001 (Lander and Kruglyak, 1995).

While this type of analysis has proven successful in the past to detect trait loci, particularly in Mendelian disease, there are a number of problems with parametric linkage analysis. The major problem pertaining to this type of linkage analysis relates

to the uncertainty with which the underlying genetic architecture of the disease under study can be predicted. The necessity of parametric linkage analysis to define a given mode of inheritance and set model parameters, while making the test extremely specific, also make it sensitive to errors in these parameters. Although testing a model specified with the correct parameters would have more power to detect linkage, in reality it is hard to know what these correct parameters are, and specifying the wrong mode of inheritance or setting inaccurate parameters can make it extremely difficult to detect linkages. It has been shown that linkage analyses are sensitive to the degree of dominance, although less so to the allele frequency estimates (Clerget-Darpoux et al., 1986). Consequently, there are also a range of other linkage methods that are non-parametric (or model-free) in nature, and these can be much more powerful in situations where parameters are difficult to accurately specify.

1.4.2 Binary traits and non-parametric linkage analysis

Model-free methods of linkage analysis rely on the concept of identity-by-descent (IBD) sharing of alleles amongst relatives. Two alleles are said to be IBD when they both descend from the exact same ancestral allele in a common ancestor when traced back through a pedigree. Two alleles which are identical from a molecular point of view (for example, two "A" bases) but not definitively descended from the same ancestral allele are termed identical-by-state (IBS). Note that alleles that are IBD are always IBS, but the reverse it not necessarily true. Affected relatives will share alleles IBD at disease-causing and surrounding loci more often than expected by chance

alone, and by using markers this information can be used to find the disease-causing alleles.

The first, and still one of the most commonly used, non-parametric tests for linkage is the affected sib pairs (ASP) test (Penrose, 1935). The ASP design tests if there is a significant increase in the average observed IBD sharing between each pair of affected sibs in the sample at each marker in the study. Across a sample of sib pairs the average expected IBD sharing at a given locus is 50%, and a significant increase over this amount indicates linkage between the marker and the gene contributing to the cause of the disease. There are now many other tests based upon the ASP design which allow for more complex situations. For example, some analyses allow a more generalised family structure and are now capable of including complex pedigrees, i.e., pedigrees containing multiple generations and inbreeding loops (where it is possible to trace a route in the pedigree between two individuals, and return back to the first individual by a different route). Other ASP design extensions allow missing data, and some allow data from multiple markers to be used simultaneously (Forabosco et al., 2005).

While non-parametric tests have the advantage of not having to specify large numbers of parameters, they do suffer from having to discard relevant information, for example by leaving out unaffected relatives from the analysis. As a result, there are occasions where a correctly specified parametric test will out-perform non-parametric methods, but in general it is now commonplace for non-parametric methods to be used. One advantage of non-parametric methods is that often large pedigrees with

multiple affected individuals (favouring classical linkage analysis) are hard to find, and therefore collecting many smaller families with two affected sibs is a more feasible way to approach analysing the disease.

1.4.3 Linkage analysis of quantitative traits

As with linkage analysis of binary traits, family data are the focus of methods designed to analyse quantitative traits. The first wide-spread linkage method used for the analysis of quantitative traits was Haseman-Elston regression (Haseman and Elston, 1972). This method is still in use, and has also been the foundation for various extensions and numerous other methods which are based on the same idea. The original Haseman-Elston regression utilises sib-pairs and looks for a correlation between estimated IBD sharing of the sibs and the squared trait difference between the sibs. Under the null hypothesis of no linkage the slope of the regression line is zero; any significant departures from this should be negative in sign since squared trait difference should increase as IBD decreases, thus the test is one-sided (Feingold, 2001).

Extensions to the original Haseman-Elston regression method include allowing more distantly related pairs of individuals to be used in an analysis (Amos and Elston, 1989), and also to allow using many different types of relative pairs in the same analysis (Olson and Wijsman, 1993). Some variations of the test alter the dependent variable; instead of using the squared trait difference between a pair of relatives, an alternative function of the trait values is used. For example, using the mean-corrected

trait product as the dependent variable yields a better estimate of the effect of the locus than simply the squared trait difference (Feingold, 2001). Other alterations of the method adopt a more selective approach in the samples used. Two manifestations of this sort of test use sibling pairs in the study only if together they exhibit either extremely concordant (at either extreme of the distribution) or discordant trait values. By selecting such individuals at the extremes of the distribution, the power of the test can potentially be increased since the difference in the amount of IBD sharing between pairs at different ends of the distribution (in the concordant pair analysis), or between the discordant pairs, will be maximised.

There is also another class of linkage method used primarily for analysing quantitative traits (although applications to binary traits are possible), called variance components (VC) analysis. VC analysis models the phenotype in such a way that it is decomposed into segments accounting for different parts of the total trait variation. Typically these components are a random QTL variance component modelled by an IBD matrix (which is specific to a given location), a variance component pertaining to the remainder of the polygenic effects, and a non-genetic (environmental) variance component. It is usually assumed that these components have no covariance (Amos and de Andrade, 2001). Maximum likelihood is used to evaluate the model under the null hypothesis of no linkage and the alternate hypothesis of linkage, and a likelihood ratio test (LRT) is used to accept or reject the null hypothesis (Blangero et al., 2001). This test is performed at incremental steps along the genome to find the most likely QTL position (Blangero, 2004). VC analysis is more powerful than Haseman-Elston regression when the trait being analysed is normally distributed, and has the

significant advantage of being able to analyse family data from large pedigrees, and not rely on specific relative pairs. However, type I errors (false positives) are inflated when the trait is not normal (Feingold, 2001).

While VC analysis and other linkage techniques have proved extremely useful for detecting high risk variants, linkage tests do not have sufficient power to detect small to moderate effects, certainly not less than 10% of total phenotypic variance (Forabosco et al., 2005). This is because power to detect effects of a given magnitude is a function of sample size, and for smaller effects the sample size required to reach significance is far beyond that generally available to collect within families, or even across multiple families. This is especially the case for late onset diseases such as Alzheimer's, although to some extent this problem can be offset by using an age-of-onset quantitative phenotype instead. Additionally, in most studies there are not enough recombination events within families to be informative for densely packed SNPs, therefore dense maps provide little extra information and sparse marker maps are deemed sufficient. However, using sparse marker maps for linkage has a different consequence - typically the confidence intervals for putative QTL span hundreds of kilobases (Kb), which is problematic with regard to identifying specific candidate genes.

1.5 Association Analyses

After the initial success of linkage studies, it was evident that less had been discovered about the genetics of common complex diseases (and traits in general) than was expected. In an influential piece of work in 1996 it was demonstrated using

binary phenotypes that with realistic effect sizes, linkage studies did not have sufficient power to detect loci using the sorts of sample size that were feasible (Risch and Merikangas, 1996). In the same paper it was also demonstrated that by using a different approach, that of association studies, power was much greater to detect genes of smaller effect size than it was for linkage. Association tests show greater potential for identifying the modest risk genes that are most likely to be involved in controlling common complex human disease.

Association is fundamentally different to linkage in its concept, although both techniques rely on markers in close proximity to QTL in order to detect their presence. Where linkage requires the same marker allele and QTL allele to remain together in related affected individuals, the same is not necessarily true for association. In essence, association hypothesises that the marker being tested is actually the causative variant affecting the trait, and assesses the statistical evidence across the population to support this theory. The technique was originally applied as a direct test for association of variants in a candidate gene type approach, where likely genes were sequenced and markers were tested for association with the disease (Kruglyak, 1999). In reality however, it is extremely unlikely that the causative variant happens to be one of the typed markers being tested.

In order to detect QTL, association studies rely on a phenomenon known as linkage disequilibrium (LD) between the tested marker and a causal variant. While this is also true for linkage studies, for association the LD must be present at a population level, not only a family level (which is trivial for close loci as a consequence of direct

transmission from parents to offspring). Unsurprisingly then, in general association studies are performed population wide, and make no special attempt to sample multiple related individuals. Indeed, sampling related individuals in population-based studies can cause additional complications for analysis, as will be discussed subsequently. Some tests for association do exist in which the basic unit is a nuclear family, although these still depend on LD between the marker and causative locus at a population level. Examples of these are the transmission disequilibrium test (TDT), quantitative transmission disequilibrium test (QTDT) and family-based association test (FBAT). For the remainder of this Thesis, the focus will be study designs which utilise population-based data, therefore these family-based association methods shall not be discussed further.

As will be described more fully in due course, the reliance upon population-wide LD means a denser marker map is required to perform association studies. Consequently, when an abundance of new markers were made available through sequencing projects such as HapMap (The International HapMap Consortium, 2005), and more efficient genotyping methods became available, this caused a shift in study design. Where previously linkage studies were targeted to one or a small number of specified genes, the advent of association sparked a progression to use all of the new information provided, rather than specific parts. Hence association studies became truly genome-wide and inspired a move towards indirect testing, hoping to detect QTL anywhere in the genome by virtue of LD, rather than the traditional direct testing (Collins et al., 1997). Indirect association has the benefit of requiring no prior assumptions about the genomic location of disease-influencing variants, and is therefore innately appealing

since none of the genome is ignored by virtue of having no previously known relevance.

1.5.1 Linkage disequilibrium

Linkage disequilibrium refers to the non-random association of alleles between two loci, causing certain alleles to appear together more often than would be expected by chance alone. There are many population genetic factors that can create LD between loci – genetic drift, selection and mutation for example – but unless sustained by a force such as selection, LD will decay over time. This decay of LD between loci occurs as a consequence of recombination, and since recombination happens more frequently between loci further apart, in general LD persists longer between more closely linked loci. Between very tightly linked loci it is possible for LD to persist for many generations without being eroded, but LD can also exist over much longer regions, and it is not unknown for LD to persist even through recombination hotspots (The International HapMap Consortium, 2005).

Linkage disequilibrium technically exists between any number of loci considered jointly, although usually only pairwise measures of LD are considered. Given that the majority of SNP loci have only two alleles, a single LD measure is enough to capture all LD information present at those loci, although for loci with more alleles there is an LD statistic for each combination of alleles. LD is measured using the coefficient of linkage disequilibrium, known as D. For two biallelic loci, A (with alleles A and a) and B (with alleles B and b);

$$D = p_{AB} - p_A p_B$$

where p_{AB} is the frequency of the AB haplotype in the population, and p_A and p_B are the population frequencies of the A and B alleles. In this case, where the loci are biallelic, the coefficient of linkage disequilibrium, D, between all alleles is the same (i.e., $D_{AB} = D_{ab} = -D_{Ab} = -D_{aB}$), therefore no subscript is required. If D is equal to zero then no LD exists between those loci, and the alleles are therefore behaving statistically independently of one another (Slatkin, 2008).

The coefficient of linkage disequilibrium, D, is not the best descriptor for the level of LD between two alleles, although the two other measures of LD more frequently used are both derived from D. The first of these is D', which is the ratio of the observed value of D to the maximum value of D possible given the allele frequencies. D' ranges from minus one to one, tending away from zero as the level of LD between alleles increases. The sign attached to a D' value reflects whether the alleles are in coupling or repulsion (i.e., more often found on the same chromosome together, or on opposite chromosomes), and it is the absolute magnitude of D' that reflects the level of LD. D' is particularly useful to detect recombination, since it only has the value one if at least one of the four possible haplotypes is not present, indicating that recombination has not occurred. The second other measure of LD is called r^2 and this has a value between zero and one. R^2 is the squared statistical correlation between two alleles, and this property of r^2 is one reason why in general it is used more often than D', since it provides an indication of which SNPs are highly correlated and can therefore be used as proxies for each other (often called tag SNPs). r^2 is also more sensitive to the allele frequencies at the two loci than D', as r^2 can only be one when

both alleles have the same frequency. The importance of the ability of SNPs to serve as proxies for each other shall be discussed in greater detail subsequently.

LD is essential for successful association studies because when testing markers for association, even those from within candidate genes, it is highly probable that none will be a causative locus. However, the estimated effect size of a tested marker is a function of both the true effect of the causative QTL and the level of pairwise LD that exists between the marker and this QTL (Blangero, 2004). With this being the case and due to the fact that LD decays with distance, a much denser marker map is required for association compared to linkage. As stated previously, the HapMap project was responsible for cataloguing a large proportion of the common human genetic variation world-wide (four distinct populations were analysed), and discovered a large number of novel SNPs which could be exploited in association studies.

The International HapMap Consortium genotyped over one million SNPs in 269 samples in the first pass (The International HapMap Consortium, 2005), and at the second pass over 3.1 million SNPs in 270 samples were genotyped (The International HapMap Consortium, 2007). Interestingly, for the European samples in the study, the common SNPs (minor allele frequency ≥ 0.05) had a mean maximum r^2 – that is the mean of all SNPs' maximum LD value with another SNP – of 0.96 to any other typed SNP, and even the rare SNPs (minor allele frequency < 0.05) had a mean maximum r^2 of 0.79. This illustrates how each SNP in the genome is generally well tagged by at least one other SNP, and it is likely that many others do so almost as well. It was

estimated that to capture the variation from all common SNPs in Phase II of the HapMap project with an r^2 of ≥0.8, only slightly more than 500,000 SNPs would be needed. If HapMap has done a sufficient job of capturing most of the genome-wide SNP variation present, this means that almost all the common SNP variation in humans is tagged with on average 80% efficiency in just half a million SNPs (The International HapMap Consortium, 2007).

1.5.2 Association studies of binary traits

With the availability of a larger number of SNPs to analyse and their greater power compared to linkage, association studies began to be utilised for binary traits. The degree of precision afforded by the much greater number of markers used in association enabled studies to fine-map QTL to regions of 5-10Kb or less once a putative QTL had been initially detected using linkage. However, as opinions on the likely nature of QTL began to change, association studies became more attractive as an alternative, not just supplementary, way to analyse data. Just as with linkage, there are now association methods for analysing both categorical (in most circumstances with human data this means binary) and quantitative traits.

Population-based association of binary case-control phenotypes requires careful selection of an appropriate dataset. Unlike in linkage and family-based association, a single or multiple affected probands across several families are not sufficient for an analysis. Ideally, hundreds or even thousands of unrelated affected individuals and an equal number of carefully selected matched controls are required to perform an

analysis. Usually the number of affected individuals with data available is the limiting factor on the sample size of the study, but having as many as possible is crucial since power is partly determined by sample size. Controls should be matched as closely as possible to cases for factors such as age, sex, ethnicity, socio-economic status and other relevant variables (for example smoking) for the study. Failure to select appropriate controls may results in spurious associations.

The classical way of assessing evidence for association with a binary trait is by using case-control data in a contingency table, where rows and columns are tested against the null hypothesis of statistical independence. This can be performed using a genotype model which produces a 2x3 contingency table, or as is more usually the case, by using a 2x2 table and assuming an additive model for allele action. It is also possible to specify a fully dominant / recessive mode of inheritance and test this in a 2x2 contingency table. The additive test is the one most frequently used, and essentially states that the heterozygote risk is expected to be intermediate between the two homozygote risks. A Pearson one degree of freedom (df) test can then be used to test for association. One problem with this test is that it assumes Hardy-Weinberg equilibrium (HWE) in cases and controls combined, which may not hold. However, a way around this is to use the Cochran-Armitage test instead, which also assumes an additive mode of action for the locus but does not assume HWE (Balding, 2006).

A slightly more complex way to analyse case-control data is using logistic regression. Logistic regression uses the logit transformation to assess the log odds of disease, and can perform either a 2df genotype test or a 1df additive test. Significance is tested via

a likelihood ratio test to compare the full model (with a SNP effect) with the reduced model not containing a SNP effect. One benefit of logistic regression is that it is extremely flexible and models can be specified in a way that tests much more specific alternative hypotheses (for example, for dominance or for an interaction). Logistic regression can also be extended in order to fit covariates in the model which may not be of direct interest themselves but that need to be accounted for, such as age or sex for example. Similarly, multiple SNPs may be fitted in the model simultaneously, and a general test of their overall association can be performed. In the simplest case, logistic regression models are equivalent to the simpler contingency table counterpart, however the additional benefits afforded by logistic regression makes it a very useful tool for detecting genetic loci with effects on disease.

1.5.3 Association studies of quantitative traits

In order to dissect complex traits, researchers are turning more and more to the analysis of quantitative phenotypes. With the advent of association analyses, techniques have had to be developed to deal with this demand. Typically, the method of analysis used on quantitative data is simple linear regression, where the SNP marker can be specified as either a linear covariate (i.e., a 1df test where the effect is additive), or as a fixed effect where each genotype has its own unique effect on the trait. In this case, each genotype can be tested individually in a 1df T-test, or a global F-test for association of all genotypes can be performed. Just as with Haseman-Elston regression and logistic regression, additional "nuisance" covariates and fixed effects can be fitted in the model to better fit the data. A two-sided T-test can be used to test

the effect of the marker in the additive model, or an F-test for the genotypic model, a significant result rejecting the null hypothesis of no association.

An alternate but related method of analysis to regression is analysis of variance (ANOVA). This partitions the variance due to each factor in the model and performs an F-test to assess the null hypothesis that the amount of variance attributable to any given factor is equal to zero. The F-statistic produced for a genotype effect using this test should be the square of the T-statistic produced for the corresponding genotype using linear regression. As with regression, additional factors not of primary interest may be included. Both linear regression and ANOVA require that trait residuals are normally distributed, and usually means the trait is approximately normally distributed itself. If a trait is not normally distributed, or there is difficulty obtaining normal residuals, a transformation of the data can be performed in an attempt to solve this problem (for example, a log-transformation or square-root transformation).

As with the analysis of binary traits by association, the norm nowadays is to interrogate the whole genome by testing a vast array of SNP markers. While this method is exhaustive in its search for association, it also introduces something which had never been a serious problem until now; multiple testing. Multiple testing is not the only issue pertaining to genome-wide association studies (GWAS) however, as there are a number of other matters which also require consideration. Some of these issues relate initially to study design, but also have ramifications for both the analysis of data and subsequent interpretation of the results. As this thesis is concerned with

using association methods to detect QTL in complex traits, many of these issues are discussed in greater detail below.

1.6 Issues with GWAS

Issues concerning the design and analysis of GWA studies can be decomposed into matters concerning the study population, marker panel selection, method of analysis, interpreting the results and replication of any positive findings. There are very few definitive solutions to these issues however, and the literature is divided on the best way to perform a GWAS. Nevertheless, a number of standard procedures have arisen, and while not necessarily ideal, they are broadly recognised as the most acceptable way to conduct GWA studies.

1.6.1 The study population

"Study population" refers more specifically to the population from which the samples for the study were taken, since it is rare that data from the entire population will be collected. There are a number of considerations regarding the study population, and one of the most important of these when designing a GWAS is sample size. This problem is clearly not unique to association studies, but it is a crucial consideration nonetheless. In general, the larger the study the more power there will be to detect genes of small to modest effect, and while the cost of a study scales linearly with the number of samples, this is not the case for power. Having too few samples will severely under-power a study to the extent that any putative associations will be

suspect unless the effect size is extremely large, because power to detect small to moderate effects will be negligible. Consequently, it is important to reach an appropriate trade-off between the cost of the study and an acceptable power to detect effects of a given magnitude.

Another important consideration for population-based association studies is ensuring that samples are a randomly selected subset of the population as a whole (to the extent possible - cases for disease studies are obviously selected based on being affected). Unless specified by adding terms into the model, samples are considered independent, so selecting individuals that are related may introduce covariances to both phenotype and genotype that cause spurious results. It is also important that the demographic history of the population is known. In modern society, particularly in the developed world, it is common for many different populations to become admixed. Hidden population admixture can lead to false positives in GWA studies if both the allele frequencies and incidence of disease differ between recently admixed populations.

The study population also affects the choice of SNP marker panel. Although choice of marker panel will be discussed further in the next section, it is important to emphasise here that it is critical the panel used to analyse a study population was designed with that population in mind. For example, using a panel in which SNPs were selected to tag an Asian population for analysis of data from a European population would not be optimal. This is because SNPs can vary in minor allele frequency and patterns of LD between populations. If SNPs selected on the panel

were identified in a population with a different history to the study population, then it is likely that some will not even segregate in the study population. Additionally, some SNPs missing from the panel may no longer be captured by LD for the study population by those that are on the panel.

1.6.2 The marker panel

A major reason for the increasing prevalence of association studies is the increase in availability of relatively cheap and fast-throughput SNP genotyping platforms. However, the choice of marker panel for a GWAS is very important. One reason for this has already been described, however there are several more. New high-throughput technologies are enabling ever more SNPs to be genotyped on a single panel, but even so the cost of GWA still scales with the number of SNPs that are typed. Unlike with sample size where in general more is better, and money is usually the limiting factor, for SNPs there is a different trade-off. Due to the reliance of association on LD, a relatively dense marker map is required. This is because LD decays quickly over large distances and the aim is to capture as much genetic variation in the genome as possible. Preferably, good genomic coverage would be achieved with as few markers as possible however, to avoid generating an excess of redundant information which would exacerbate the problem of multiple testing. When r^2 between neighbouring SNPs is high, one SNP captures almost all variation at the other, therefore only one of them needs to be typed. However, LD is not uniform across the genome and in some places the variation can be reliably captured with a fraction of the number of markers of other places. The ideal SNP panel would

therefore have an enrichment of SNPs in regions where LD is typically quite low, and be slightly more sparse where LD is high.

Even so, there is still a somewhat arbitrary choice of what level of LD on average is acceptable, in order to capture as much genotypic variation as possible. For example, requiring that at least one SNP on the panel has a minimal r^2 of 0.7 with each SNP not included will not provide as much power as if an r^2 of 0.8 was used, because power is partly a function of LD between the marker and causative locus. However in the latter case, many more SNPs would have to be included. As already mentioned (and will be discussed in more detail shortly), the major drawback of simply including all or most segregating SNPs in the analysis is the multiple testing burden this imposes on the study.

There is a variety of options of SNP panel to choose from. Two of the leading manufacturers of SNP chips are Illumina (www.illumina.com) and Affymetrix (www.affymetrix.com), both of which offer several different panels depending on the level of SNP coverage required. A typical panel includes upwards of 300,000 SNPs, although there are now chips which contain 500,000 and even up to a million SNPs. The companies also offer different panels depending on the population under study; for example there are different SNP chips for use in European and African populations. Of the two companies, only Illumina selects its SNPs based on maximising the overall variation captured by the SNPs and reducing redundancy. Affymetrix initially produced panels based on enzyme restriction sites, and

subsequently filled these panels with an essentially random selection of SNPs spaced evenly throughout the genome.

1.6.3 Population stratification

The majority of association studies use population-based data rather than family data, or at least operate under the assumption that each individual is independent, i.e., unrelated. In reality this is not always possible, and also ignores the fact that if ancestors are traced back far enough, any two people are related. In studies where known family members are included, or some other sort of population structure or relatedness is expected, there are ways to account for this in the analysis. If unaccounted for, hidden population stratification (or cryptic relatedness as it is also known) can cause spurious associations, just as with using admixed populations.

There are a number of ways in which either known relationship information or cryptic relatedness can be accounted for. Where known relationships exist, the typical method is to use a relationship matrix to fit a random polygenic term into a regression model. This accounts for the genome-wide covariances between relatives and allows a test to be performed for association at a specific marker. The relationship matrix for this method is based upon expected values of IBD sharing for each type of relative pair; for example, full sibs are expected to share on average 50% of their genome IBD, and this value decreases for more distantly related pairs of individuals. At any given locus however, the actual proportion of alleles shared IBD between full sibs can be zero, one or two with probabilities 0.25, 0.5 and 0.25. This means that the

estimated amount of IBD sharing over multiple loci will be wrong 50% of the time. While using a relationship matrix derived in this manner is on average correct, purely by chance the amount of alleles actually shared IBD by two full sibs across the genome may vary considerably, leading to inaccurate estimates of genome-wide IBD.

Inaccurate estimates of genetic relatedness can lead to incorrect results that generate false positives and obscure true positives, both of which want be minimised in GWA studies. There is an alternative, more accurate, way to estimate genetic relatedness between individuals that may be more suitable. Genome-wide marker data can be used to give an estimate of relatedness across the genome by averaging marker similarity at each individual locus. This is not entirely straightforward however, since only IBS between individuals is actually observed, and IBD must be inferred. This impacts upon accuracy, but while these methods inevitably introduce some degree of error, programs do exist that make reasonable estimates of true IBD sharing from dense marker data.

Another way of dealing with cryptic population relatedness is to use something called genomic control (Devlin and Roeder, 1999). Genomic control makes no attempt to account for relatedness between individuals in the tests, but instead induces a post-hoc adjustment to p-values produced from the tests. This is because the effect of population substructure on the test statistic should be constant throughout the genome, hence test statistic inflation due to substructure is the same for all markers (Bourgain and Genin, 2005). To account for this, observed p-values from a genome-wide scan are plotted against the expected p-values from the null hypothesis on a Q-Q

plot. If there is a general inflation of all observed p-values from the genome-wide scan this indicates that cryptic relatedness exists in the population. To adjust for this, a deflation factor known as λ is calculated as the ratio of the median expected p-value to the median observed p-value. All p-values are then multiplied by this deflation factor, thus adjusting the results for the presence of population structure.

One other technique for removing problems associated with either population- or pedigree-based relatedness is to identify groups of individuals using clustering approaches. A similar approach is principle components analysis (PCA), which is a data reduction technique that attempts to assign individuals into a number of groups based on covariances at marker loci. The principle components (PCs) can then be used as factors in a regression analysis so that phenotypic covariance caused by these PCs is not unaccounted for.

1.6.4 Method of analysis

Although some methods of analysing data are more widely used than others, there is no consensus upon which is best. In all likelihood, the best method for any given dataset will vary depending on a number of unknown factors such as the number of underlying QTL present and their frequency and effect sizes. However, one aspect that remains constant in the majority of GWA studies is that SNP data are interrogated one SNP at a time. Generally, the same test is performed on each SNP in a stepwise fashion along the genome, until the entire SNP panel has been used. This is undoubtedly the simplest way of analysing GWAS data, and has been moderately

successful to date. However, this may not necessarily be the most effective way of using the data, and there are a number of interesting alternatives.

All alternatives of using a single SNP to analyse GWAS data naturally involve using more than one SNP at a time to test for association. The theory behind performing association with multiple SNPs is that more SNPs may better tag a causative variant which exhibits low LD to typed SNPs. If no single SNP has high r^2 with a causative variant then single SNP analyses will fail to detect the QTL unless power is very high due to other factors such as sample size. However, QTL may well be in high LD with certain combinations of alleles at multiple nearby loci. Specific combinations of alleles inherited together on the same chromosome are known as a haplotype, and the existence of LD causes some haplotypes to become more frequent than would be expected based solely on the frequencies of their SNPs. Fundamentally, the idea behind using multiple markers in association studies is to attempt to locate QTL by virtue of identifying haplotypes conferring higher or lower risk to disease.

While all alternate methods for GWA studies use multiple SNPs, the way in which the multiple SNPs are used determines the differences between them. The simplest way of using multiple SNPs is to incorporate more SNPs into the regression model, analogous to the way extra covariates and fixed effects are fitted for example. The only difference is that unlike the covariates these extra SNPs are of interest, and can be tested individually in 1df or 2df tests (depending whether the SNPs are parameterised for an allelic or genotypic test), or globally in a F-test of all SNPs in the model. Adding large numbers of SNPs into multiple regression models can cause

problems however, since each SNP added takes up an additional degree of freedom for performing a global test of association. It is possible to avoid over-parameterisation of the model by implementing a step-wise procedure that determines which SNPs should be kept and which left out. Step-wise procedures are defined as either forwards or backwards depending on whether SNPs are added (forwards) or removed (backwards) from an initial model. SNPs are accepted or rejected from the model based on a model comparison metric such as the AIC (Akaike's Information Criterion), and in this way the optimal model can be found.

Fitting multiple SNPs in models like this only tests the marginal effects of each of SNP, that is, the main effect each SNP has individually on the trait. Most SNP effects are expected to be main effects, however, some genetic effects are epistatic in nature. "Epistatic effects" (or epistasis) refers to effects which are the consequence of an interaction between two (or more) loci. To detect epistasis using multiple regression, interaction terms must be explicitly added to the model and tested. While there is only one interaction term for two SNPs (using an additive model), the number of possible interactions increases rapidly as the number of SNPs increases. In GWA studies epistatic effects are usually ignored at least initially, since the number of two-way interactions possible given the number of SNPs in a typical GWAS is prohibitively large, and even more so for interactions involving larger numbers of SNPs.

Another more sophisticated method of using multiple SNPs to test for association is haplotype analysis. This is no easy task however, and there are many things to think about when considering using haplotypes for a GWAS. Not least of these is obtaining

haplotypes in the first place, since only single-locus genotypes are generated from SNP panels, not multilocus haplotypes. While there are technologies available to obtain haplotypes directly from DNA samples, these are extremely expensive and therefore generally not a favoured choice for most GWA studies. An alternative way to obtain haplotypes is to use a statistical algorithm which estimates haplotypes and their frequencies in the study population from the genotypes. Although these algorithms inevitably introduce some level of error into the data, some programs are able to generate very accurate estimates of the haplotype structure of the population. One suite of such programs that was originally based on the Expectation Maximisation (EM) algorithm (Excoffier and Slatkin, 1995) is PHASE, fastPHASE and warpPHASE (Stephens et al., 2001).

Once haplotypes are obtained, a decision regarding the method of analysis must be made. To some extent this may also be reliant upon the algorithm that was used to generate the haplotypes, since different algorithms provide different outputs. Some algorithms give only the most likely haplotype pair for each individual for example, while others give the probability of each possible haplotype for each individual. Both the type of analysis performed and the interpretation of results are dependent on the data that are available from the haplotyping algorithm. Regardless of the type of data provided, there is a wealth of different options for how this data can be used. One particularly pertinent issue is how to deal with rare haplotype classes for example. From a statistical standpoint it is preferable to remove rare classes prior to analysis, but it is far from clear that this is the optimal way of dealing with these classes.

Haplotypes provide a very interesting and potentially powerful new way to look at GWAS data, yet thus far they are little used and questions remain regarding their application. Where haplotypes have been used in the literature it is sparingly, often as a last resort, and with sub-optimal analysis methods. One of the aims of this project was to investigate genome-wide data using haplotype approaches, and as a result a much more detailed discussion of haplotypes and their application to GWA studies will be provided in subsequent chapters.

1.6.5 Multiple testing

The explosion in availability of genome-wide SNP data has undoubtedly been beneficial for attempting to identify loci influencing quantitative traits. However, it has also brought with it a problem of its own. By testing many more SNPs for association, the nominal 5% significance threshold is no longer appropriate to ensure an acceptable type I error rate at the genome-wide level. Where a single test might be considered significant if it exceeds a nominal p-value threshold of 0.05, performing more than one test requires this threshold to be altered. If no adjustment is made, then as the number of tests performed increases, it is increasingly probable that one will report a significant association by chance alone. For example, when performing 100 tests with a nominal p-value of 0.05, the expectation is that a total of five tests would reject the null hypothesis by chance even if no association was present.

The solution frequently used in the literature to the problem of multiple testing is to adopt a Bonferroni correction to the nominal p-value. This involves dividing the level

of significance required by the total number of independent tests performed, and yields a point-wise threshold appropriate for genome-wide significance. So, in the example above with 100 tests requiring an overall p-value of 0.05 for significance, the p-value required to show significance for an individual test would be $5x10^{-4}$. The Bonferroni correction is a useful tool to assess significance of genome-wide results, however it is not perfect. The problem is that it assumes all tests are independent, however this assumption is clearly violated in the case of SNPs due to the presence of LD. This results in the correction being too stringent, i.e., the p-value required for significance being too small, and therefore the number of type II errors (false negatives) in the results increases. As power is defined as one minus the false negative rate, clearly power is reduced by inflated type II error.

In practice, while the Bonferroni correction is useful as a guideline for the level of significance required to indicate true association, the significance threshold is often relaxed somewhat when interpreting GWAS results. However, this produces uncertainty regarding exactly how small a p-value must be in order to be considered a potential true association. Small QTL effect sizes may cause genuine signals to have non-significant p-values, and contrastingly, highly significant results may be entirely spurious, therefore unequivocal conclusions about which associations are real can be hard to draw.

There is another way to determine significance for genome-wide analyses, although this also has disadvantages. Permutation analysis can be used to construct an empirical distribution of the test statistic under the null hypothesis of no association.

This is achieved by permuting phenotypes with respect to genotypes in such a way that the LD structure remains intact, but any association between genotype and phenotype is broken. A genome-wide scan is then performed on the permuted data, and the most significant result is stored. Data is permuted a large number of times so that a distribution of the most extreme result is created, which can then be used to set a genome-wide significance level for the real data. While this method is very robust, it is extremely computationally intensive due to performing what amounts to thousands of genome-wide scans. Therefore it is more usual for the imperfect but far more practical Bonferroni correction factor to be used.

1.6.6 The next step: replication and follow up

One of the major problems of any linkage or association study is knowing whether results exceeding the significance threshold are genuine associations. Rarely is an associated SNP located in the intron of a relevant gene for the trait under study, or code for a non-synonymous mutation. Most putative associations are to SNPs not in genes or even necessarily close to genes, and even if they are it is not immediately clear why the gene would affect the trait under study. To make results such as these more reliable, it is necessary to replicate them in an independent study. This is not always trivial, since there is not always a study population available to replicate the finding in. If no alternative study population currently exists, then the options are to either wait for a suitable study to commence, or spend more money genotyping a whole new study population.

Even when there is a population in which to attempt replication, results are not always positive. Generally a much reduced significance threshold (typically 0.05) is acceptable to act as replication of a putative association, but often even then the replication fails. If this is the case then, assuming the replication study was well-powered enough to detect the effect, it is probable that the finding was a false positive. One other possible explanation for failed replication is that the effect is real, but that it is unique to the population it was discovered in. This could be due to either a founder effect, or random genetic drift causing the frequency of the mutation to increase in the discovery cohort. Either of these would result in there being little chance of replication, and is therefore likely to be of limited relevance to other populations, since the variant will be extremely rare or not present at all.

If replication is successful, there is a much greater chance that the finding is a real genetic effect. However, in many respects this is just the beginning; the next task is to identify the causal variant and discover how this mutation mediates an effect on the phenotype. This task is made easier if the association is discovered in or near a candidate gene or gene rich region, although even so it is unlikely to be fast or cheap. To identify the causative mutation requires sequencing individuals in the study at the associated locus to ascertain what variation is segregating in the population within the region of interest. As sequencing is relatively expensive, it is of paramount importance to select which regions are of most interest to examine more closely. However, if the original finding is located in a gene desert it is hard to know how to follow the association up since there are no good candidate regions for sequencing. A possible exception to this is if the variant falls in what appears to be a regulatory

region, although in this case it can be hard to determine which gene or genes are being regulated.

After identification of the causal variant, the next task is to elucidate how the mutation is affecting the phenotype. This can involve complex pathways that take years to fully understand, highlighting the fact that identification of QTL via association studies is only the first step, and there remains much work to be done subsequent to the discovery. The manner in which a project proceeds at this point is largely dependent upon the funding and resources available to the project, and whether functional studies of the putative QTL are therefore feasible.

1.7 The Current Project

This Thesis is concerned with detecting QTL using genome-wide association studies in human populations. A typical GWAS now entails hundreds or thousands of individuals and several hundred thousand SNP markers, and usually the traits of interest are disease endpoints, or intermediate phenotypes thought to play major roles in disease aetiology. The next chapter of this Thesis describes the analysis and results of such a study performed in a Croatian isolate population. The most interesting of the findings in this initial scan, a highly significant association between a region of chromosome 4 and uric acid, is then taken forward and investigated further in the third chapter.

The remaining two analysis chapters in this Thesis move away from using traditional methods to analyse GWAS data. Instead the focus moves towards looking at alternative methods of using genotype data in a GWAS, and in the fourth chapter several different multiple marker methods are used to analyse simulated data. The performance of these methods is compared and contrasted with that of traditional single SNP analysis. Subsequently, in chapter five, all these methods are applied to a subset of the data that were analysed in chapter two, to see how the methods compare when used on real data.

2. CHAPTER 2

2.1 INTRODUCTION

Common complex diseases currently represent the primary health burden in the Western world (Zondervan and Cardon, 2004). Consequently, there is much interest in identifying novel genes involved in controlling these diseases. Early successes in mapping disease genes were typically achieved using highly-structured family based linkage designs (Forabosco et al., 2005), however more recently genome-wide association (GWA) studies have become the method of choice for determining the genetics behind multifactorial disease. This is for a number of reasons, the foremost of which is the recent improvement in SNP typing technology and the development of methods to quickly analyse genome-wide data. Additionally, a recent change in philosophy regarding study design means that quantitative intermediate phenotypes conferring increased disease risk to disease, as opposed to clinical disease endpoints themselves, are now more often the focus of study than was initially the case. This is not to say that direct disease association is not still popular however (see for example The Wellcome Trust Case Control Consortium, 2007).

Intermediate phenotypes affecting disease generally have higher heritabilities and are less complex than the disease they are a risk factor for, therefore it is thought that disease loci will be detected more readily through analysis of these intermediate phenotypes than the disease. The change of emphasis towards analysing intermediate phenotypes reflects the shifting opinion regarding the likely nature of genes involved

in common complex disease - instead of being determined by a single dominant high-penetrance gene, these diseases are the result of combinations and interactions of numerous genetic and non-genetic factors.

This chapter describes the analysis and results from a typical genome-wide association study (GWAS). Data for this project were supplied by a large European collaboration involving a number of different study populations. The population used in this study is from a Croatian island in the Adriatic Sea off the Dalmation coast. A variety of phenotypes were measured and analysed, ranging from those contributing to disease risk, to those of a more purely academic interest, and to those necessary for accurate analysis of the data (i.e., covariates that may required to adjust the phenotype, such as alcohol intake). Data were analysed using a novel software package, GenABEL, in the statistical program R, which was designed to analyse GWA data in the presence of pedigree structure within the population. There were a number of interesting results produced in this study, at least one of which was worthy of taking forward to replication.

2.1.1 Genetic isolates

It is becoming relatively common for association studies to use genetic isolates as the study population. Genetic isolation simply means that the population is closed (there is very little or no migration into the population), and has been so since its founding some number of generations in the past. As a consequence, isolated populations generally have longer stretches of LD present in the genome (Shifman and Darvasi,

2001), since there are a limited number of founder chromosomes. Also, isolated populations often have fewer generations since founding (leading to less recombination), and have no new genetic variation introduced other than by mutation, which itself is a very rare event. Long stretches of LD such as those exhibited by individuals from isolated populations are very useful for association mapping because they make it possible to detect causative variants with more distant SNPs - although it can make fine-scale localisation of the causative variant more difficult.

Assuming panmixis (i.e., random mating within the population), genetic isolation also eliminates the problem of population stratification, which is one great advantage over using populations which are potentially admixed for GWA studies. Another advantage of isolated populations is that there is expected to be much less trait heterogeneity, i.e., a specific mutation influencing a particular trait is likely to be present in all individuals showing increased (or decreased) trait value (Bourgain and Genin, 2005). This is due to the fact that only a subset of all loci affecting any given quantitative trait (QT) will be segregating in the population founders, and if by genetic drift one of these loci reaches an appreciable frequency in the descendant population, all descendents showing an increase / decrease in the trait are likely to be identical by descent (IBD) for the same QTL, by virtue of common descent from the same ancestor. The probability of genetic drift raising the frequency of a QTL to a level where it is easier to detect is higher for isolated populations due to the relatively small size of the founder population (Kruglyak, 1999). Instead, if the disease was a consequence of a mutation within the population since the founding, genetic drift could still cause the allele to become frequent, however in this situation it may be

difficult to replicate the association as the same mutation may not be present in other populations. Nevertheless, it may still highlight an involvement of previously unknown pathways or genes in the trait.

One further advantage of using genetic isolates is that individuals from these populations often have shared environmental exposures as a consequence of similar natural environments, diet and lifestyle choices influenced by their culture. Shared environments cause a reduction in the residual environmental variance usually associated with quantitative traits in more urban, admixed populations, and also reduce the proportion of trait variance attributable to gene-by-environment interactions. This can improve the signal-to-noise ratio for genetic effects by effectively increasing the heritability of traits in these populations (Heutink and Oostra, 2002). There have been many examples of successful studies using isolated populations to date (see for example Stefansson et al., 2002; Gianfrancesco et al., 2003).

2.1.2 EUROSPAN and CROAS

The collaborative project of which the dataset analysed here is part is called EUROSPAN, and is comprised of five isolated study populations; these came from Croatia, Orkney, Sweden, Italy and the Netherlands. In addition to performing association analyses for these populations, EUROSPAN aimed to compare properties of the isolated populations. EUROSPAN was funded by the EU, and a portion of this money went to fund the genotyping of the Croatian part of the project, called

CROAS. The Medical Research Council (MRC), the University of Zagreb and the Institute for Anthropological Sciences (Zagreb) all provided additional funding. Also involved in analysis and processing of the data were the University of Edinburgh, the MRC Human Genetics Unit (Edinburgh), Public Health Sciences (Edinburgh), the University of Zagreb, the Institute for Anthropological Sciences (Zagreb) and genotyping labs in Germany.

Croatia has become a focal point for genetic studies into QTL mapping and inbreeding in recent times as a result of harbouring many genetically isolated populations. There are a total of 1,185 islands along Croatia's Adriatic coastline, although only 67 are inhabited. The inhabited islands have been isolated for many generations, and have very little migration between themselves or the mainland, making them perfect for association analyses. One island in particular was identified as a good candidate on which to perform genetic analyses; the island of Vis. The main aim of the CROAS project was to identify QTL involved in susceptibility to common complex disease through analysis of disease risk factors.

2.2. MATERIALS AND METHODS

2.2.1 Study population

Participants in this study were mostly residents of two large villages on the Croatian island of Vis, although some participants also came from smaller villages on the island. The two main villages on this island are called Komiza and Vis, and

population sizes of these villages at the time the project commenced were 1,523 and 1,776 respectively. All participants in the study were volunteers, and had given informed consent. Volunteers were a mixture of families and single individuals with no known relatives (referred to as singletons henceforth). In this way, a total of 1,062 of the adult population of Komiza (N = 584) and Vis (N = 478) – approximately 65 - 70% of the total adult population – were recruited to take part. These individuals ranged from 17 to 90 years of age.

Detailed pedigrees were constructed for the study population using both historical family records collected during the fieldwork, and parish registers containing genealogical records dating back to 1838 and covering the period up to 1950 (the earliest records detail a couple born in 1838 and 1851 respectively). The largest pedigree joined 481 participants in a pedigree encompassing 7,242 individuals, with another 138 individuals comprising a further 58 small pedigrees of between three and 13 individuals (the larger pedigree was subsequently split as some links were un-informative for analysis). Pedigree reconstruction was performed before this PhD project commenced. Dr. Ozran Polasek and Dr. Ivana Kolcic of the University of Zagreb undertook reconstruction of the large pedigree. A total of 412 individuals could not be connected to any other individuals, hence remaining as singletons.

2.2.2 Phenotyping

Phenotyping was performed before this PhD project commenced. Fieldwork was carried out over two summers, May 2003 for Vis, and May 2004 for Komiza. During

this time, fasting blood samples were taken (20ml EDTA either for DNA extraction or liquid nitrogen storage of 2x0.5ml aliquots for future transformation; 4.5ml citrate for clotting factors and 10ml clotted blood for serum biochemistry), and plasma and serum were rapidly frozen and stored at -70˚C in 200µl aliquots using standardised sample handling procedures. A number of biochemical traits were recorded from these samples. Volunteers were also asked for their clinical histories and given questionnaires, in addition to having anthropometric and physiological measures taken. All traits measured, along with a brief description and an abbreviation used to refer to each trait, are shown in Table 2.1. Of the original 1,062 individuals, 1,031 had data successfully recorded.

TRAIT	DESCRIPTION / UNIT	ABBREV.
Ankle Brachial Pressure Index	Lowest of ankle pressure measure divided by highest of brachial pressure measure	ABPI
Albumin	Serum albumin Measured in g/l	Albumin
Body mass index	Weight (kg) divided by the square of height (m)	BMI
Brachial circumference	Measured in Mm	Brach Cir
Brachial width	Measured in Mm	Brach Wid
Systolic pressure at brachialis left	Measured in mm Hg	Bra L
Systolic pressure at brachialis right	Measured in mm Hg	Bra R
Calcium concentration	Measured in mmol/l	Calcium
Cholesterol	Measured in mmol/l	Cholesterol
Cholesterol ratio	Total cholesterol divided by HDL cholesterol	Chol ratio
Cortisol concentration	Measured in nmol/l	Cortisol
Creatinine	Measured in mmol/l	Creat
Ddimer	Measured in ng/ml	Ddimer
Diastolic blood pressure	Measure in mm Hg	Diast
Systolic pressure at dorsalis pedis left	Measured in mm Hg	Dor L

Systolic pressure at dorsalis pedis right	Measured in mm Hg	Dor R
Eysenk personality questionnaire-Neurotism component	Extracted from personality questionnaire using factor analysis. Described in Ivkovic et al, 2007. Personality and individual differences 42: 123-133.	Epqe
Eysenk personality questionnaire-Extravertion component	Extracted from personality questionnaire using factor analysis. Described in Ivkovic et al, 2007. Personality and individual differences 42: 123-133.	Epqn
Eysenk personality questionnaire- Psychoticism component	Extracted from personality questionnaire using factor analysis. Described in Ivkovic et al, 2007. Personality and individual differences 42: 123-133.	Fev
Fibrinogen 1	Measured in Zagreb (antibody method) g/l	Fib1
Fibrinogen 2	Measured in Glasgow (clotting factor assay) g/l	Fib 2
General health questionnaire		GHQ
Glucose	Measured in mmol/l	Glucose
High density lipoprotein	Measured in mmol/l	HDL
Height	Measured in m	Height
Hip circumference	Measured in mm	Hip Cir
Hip-waist ratio	Hip circumference (mm) divided by waist circumference (mm)	Hip-Waist
Low density lipoprotein	Measured in mmol/l	LDL
Pulse pressure	Difference between systolic and diastolic blood pressure	PulseP
Reactance	From bioelectrical impedance analysis	React
Resistance	From bioelectrical impedance analysis	Resist
Biceps skinfold thickness	Measured in $1/10^{th}$ mm	Skin B
Triceps skinfold thickness	Measured in $1/10^{th}$ mm	Skin T
Subscapular skinfold thickness	$1/10^{th}$ mm	Subscap
Suprailiac skinfold thickness	$1/10^{th}$ mm	Suprail
Systolic blood pressure	Measured in mm Hg	Syst
Systolic pressure at tibialis posterior left	Measured in mm Hg	Tib L
Systolic pressure at tibialis posterior right	Measured in mm Hg	Tib R
Tissue plasminogen activator	Measured in ng/ml	Tpa

Triglycerides	Measured in mmol/l	Trigly
Uric acid	Measured in mmol/l	UA
von-Willebrand factor	Measured in IU/dl	vWF
Waist circumference	Measured in mm	Waist Cir
Weight	Measured in kg	Weight

Table 2.1 List of traits analysed, with the units they were measured in and a brief description where necessary, and their abbreviation.

2.2.3 Genotyping

Genotyping was also performed before this PhD project commenced. Genotypes of each of the study participants were taken using the Illumina genotyping platform HumanHap300-Duo Genotyping BeadChip. This chip contains 317,503 tag SNPs, with a high density of SNPs in areas of the genome within 10Kb of a gene or evolutionarily conserved region, and approximately 7,300 non-synonymous SNPs (www.illumina.org). Of the original 1,031 individuals successfully phenotyped, 986 had their genotypes successfully recorded.

2.2.4 Quality control

Genotyping rates for both SNPs and individuals were calculated, and SNPs or individuals with <90% call rate were removed from analysis. This was to ensure poor quality or contaminated DNA (in the case of low genotype call rate per individual), and unreliable SNPs (in the case of low SNP call rate) were not used in the analyses. There was a total of 17 individuals removed from the study due to a poor genotyping rate, and the number of SNPs removed was 9,552.

A detailed investigation into the genetic structure of the study population was performed to ascertain whether any genetic stratification existed between inhabitants of the two villages involved in the study. The program STRUCTURE (Pritchard et al., 2000), was used to perform this analysis, which was carried out by Dr. Caroline Hayward at the MRC HGU in Edinburgh. STRUCTURE allows the user to determine the number of sub-populations to be formed from the data using the parameter "k". By gradually increasing the number of sub-populations it is possible to establish which individuals / groups are more genetically distinct. Results from this analysis indicated that genetic stratification was present within the study population, and this broadly corresponded to the individuals of Komiza and Vis clustering into groups representing their respective villages. In addition to this, it was also discovered that there were three individuals who failed to cluster with any other samples. These three successively emerged as singletons as k was increased until the two groups corresponding to villages resolved themselves at k = 5. As these three individuals were unable to cluster they were treated as genetic outliers, and were removed from further analysis. Two of these individuals were also present in those found to have poor call rate, therefore a total of 966 individuals remained in the analyses after quality control. The total number of SNPs retained for the analysis was 307,951.

One further alteration was also made to the genotypic data before analysis, and this was to remove very rare genotypes for each SNP. With a total of 966 individuals in the analysis it was decided to remove (i.e., set to missing) genotypes with less than 10 occurrences in the data (a frequency of ~0.01). This was performed because rare genotypes can erroneously inflate the magnitude of an effect estimate (particularly for

a genotypic test) if coupled with an extreme phenotype, due to the small sample sizes of such genotypes.

2.2.5 Preliminary trait analysis

A total of 44 traits were analysed, all of which were quantitative. Before analysis, basic checking of the phenotypic data was performed. Distributions of the traits analysed were investigated, and where they were found to be non-normal, transformation was performed in an attempt to produce trait normality. Table 2.2 lists each trait analysed and shows which transformation was used, if any.

TRAIT	TRANSFORMATION	FIXED EFFECTS / COVARIATES	RANDOM EFFECTS
ABPI	Rank transformation to normality	Age Sex Smoking	
Albumin	Natural logarithm	Age*Sex	
BMI	Natural logarithm	Age*Sex	
Brach Cir	N/A	Age*Sex BMI Healthy food consumption SES Carbohydrate index	
Brach Wid	3 x Natural logarithm	Age*Sex BMI SES ci	
Bra L	N/A	Age*Sex BMI	
Bra R	N/A	Age*Sex BMI	
Calcium	Rank transformation to normality	Sex Age-by-Sex interaction Years schooling	
Cholesterol	Natural logarithm	Age*Sex BMI	

Chol ratio	Rank transformation to normality	Age*Sex Alcohol	
Cortisol	N/A	Age BMI Years schooling Physical activity Carbohydrate index	
Creatinine	N/A	Age Sex	
Ddimer	Natural logarithm	Age*Sex BMI	
Diast	Square root	Sex BMI	
Dor L	N/A	Age*Sex BMI Smoking	
Dor R	N/A	Age*Sex BMI Smoking	
Epqe	N/A	Age*Sex	
Epqn	N/A	Age Sex	
Fev	Square root	Age Sex	
Fib	N/A	Age*Sex BMI Smoking Alcohol	
Fib 2	Rank transformation to normality	Age*Sex BMI Smoking Alcohol	
GHQ	N/A	Age*Sex SES vi	
Glucose	Rank transformation to normality	Age Sex BMI	
HDL	N/A	Age Sex	
Height	N/A	Age Sex	
Hip Cir	N/A	Age Sex	
Hip-Waist	Rank transformation to normality	Age*Sex	

LDL	N/A	Age Sex	
PulseP	RTN	Age*Sex BMI Years schooling	
React	Rank transformation to normality	Age*Sex	
Resist	N/A	Sex BMI	
Skin B	N/A	Age Sex	
Skin T	N/A	Age Sex	
Subscap	N/A	Sex BMI	
Suprail	N/A	Age BMI Years schooling Carbohydrate index	
Syst	Natural logarithm	Age*Sex	
Tib L	N/A	Age*Sex BMI Smoking	
Tib R	N/A	Age*Sex BMI Smoking	
TPA	Rank transformation to normality	Age Sex BMI Alcohol	
Triglycerides	Natural logarithm	Sex BMI Smoking	Maternal ID
Uric acid	Square root	Age*Sex BMI	
vWF	Natural logarithm	Age	
Waist Cir	Natural logarithm	Age*Sex	Maternal ID
Weight	Natural logarithm	Age Sex	Maternal ID

Table 2.2 Table showing all fixed effects, covariates and random effects fitted for each of the traits analysed. Transformations performed for traits are also indicated. A random polygenic effect accounting for relatedness, and a fixed effect accounting for population stratification are fitted for all traits, and are therefore omitted from the table. Age*Sex indicates that age, sex and an age-by-sex interaction were fitted.

2.2.6 Genome-wide association

2.2.6.1 GRAMMAR

The method of analysis for this study was a new technique called Genome-wide Rapid Association using Mixed Model And Regression (GRAMMAR – Aulchenko et al., 2007). This is a three-step technique designed to greatly reduce the computational time required to perform genome-wide association analyses in the presence of relatedness in the population. The first step involves analysing the data under a model including all relevant fixed effects and covariates, and a term for the polygenic variance that is modelled using a relationship matrix. The relationship matrix contains information on expected identity-by-descent sharing estimated from the pedigree. The vector of residuals calculated from this stage of the analysis is then used in the next step as the dependant variable in simple linear regression. Each SNP is tested as a fixed effect in a two degrees of freedom (df) genotypic test, or a covariate in a 1df additive allelic test, and p-values are recorded. In the final step, all SNPs with p-values exceeding some pre-determined threshold are selected to perform the full model analysis on, which consists of the model in step one with the additional SNP effect term.

2.2.6.2 GRAMMAR – Step one

Step one of GRAMMAR was carried out using ASReml (Gilmour et al., 2002). Traits were analysed under the following mixed model;

$$y_i = \mu + \Sigma_j \beta_j c_{ji} + G_i + e_i^*$$

where y_i is the phenotype of the i^{th} individual, μ is the mean, β_j are the effects associated with the covariates and each level of the fixed effects, c_{ji} is the value of the j^{th} fixed effect or covariate for the i^{th} individual, G_i is a random polygenic effect for the i^{th} individual, and e_i^* is the residual error term for the i^{th} individual. It is the vector of residual error terms that is used in the second step of the analysis. The specific fixed effects and covariates included for each trait was dependent upon which were significant at the 5% level. Some traits also fitted an additional maternal environmental random effect since the shared environment induced by a similar upbringing in full-sib families can cause additional trait covariance. Details of the covariates and fixed / random effects fitted for each trait are in Table 2.2. Heritability estimates for each trait were also obtained from this stage of the analysis.

2.2.6.3 GRAMMAR – Step two

The vector of residual errors from the first stage of the analysis was used as the dependant variable in simple linear regression with the following model;

$$e_i^* = \mu + kg_i + e_i$$

where e_i^* is the residual of the i^{th} individual obtained from the previous step, μ is the mean, k is the additive effect of an allele (additive allelic model), or fixed effect of a

genotype (genotypic model), g_i is the genotype of the i^{th} individual, and e_i is the residual error term for the i^{th} individual. This part of the analysis can be carried out rapidly using the library GenABEL (Aulchenko, 2007) in the statistical software R (R Development Core Team, 2006).

The score test performed for each SNP is as follows;

$$T_G^2 = \frac{((g - E[g])^T \cdot \hat{e})^2}{(g - E[g]^T \cdot (g - E[g])},$$

where g is the vector containing allelic or genotypic values (depending on whether it is the 1df or 2df test), E[g] is the expectation, superscript T represents the transpose of the vector and ê is a normally distributed error term. GenABEL calculates significance based on determining the empirical distribution of the score test through permutation (Aulchenko et al., 2007).

Due to the use of residuals corrected for polygenic variation from the first stage of the analysis, the test statistic for each SNP is deflated below what it should be (i.e., made less significant). This is a consequence of factoring out some of the SNP effect with the random polygenic effect, however the magnitude of this decrease can be estimated using genomic control, which is also implemented in GenABEL. Observed p-values from the GWAS are plotted on a Q-Q plot against expected values from under the null hypothesis. The deflation factor, λ, is taken as the median value of the observed distribution divided by the median value of the null distribution, and corrected p-values are calculated as the product of the original p-values and λ.

Normally, where genomic control is used to correct for the presence of cryptic relatedness, λ takes on a value greater than one since the observed p-values are too small (i.e., more significant). However in this case there is a general inflation of observed p-values, therefore λ is less than one, and the p-values are adjusted downwards. Once the factoring out has been accounted for, a threshold is needed to determine which SNPs are significant enough to follow up with the more time-consuming full model in GRAMMAR step three. Although the Bonferroni-corrected significance threshold for 307,951 SNPs is 1.62×10^{-7}, for this study an initial threshold was set at 1×10^{-6} since in practice the Bonferroni correction is likely to be too stringent due to the dependence between tests caused by SNPs in LD.

2.2.6.4 GRAMMAR – Step three

For the third stage of GRAMMAR, SNPs exceeding an arbitrary threshold (in these analyses set to 1×10^{-6}) are analysed using the full model. This consists of all the appropriate covariates and fixed effects for that trait along with the random polygenic effect, but also includes the SNP effect. This is equivalent to the "gold standard" measured genotype (MG) approach. The model is;

$$y_i = \mu + kg_i + \Sigma_j \beta_j c_{ji} + G_i + e_i$$

where all terms in the model are as previously described. This part of the analysis can again be performed using ASReml, or in R independently of GenABEL.

2.3. RESULTS

2.3.1 Trait models

The most appropriate model for each trait was determined by finding the fixed effects, covariates and random effects that were significant at the 5% level. For each trait, only those covariates with a biologically plausible explanation affecting the trait were tested. Table 2.2 shows the model fitted for each trait, in addition to any trait transformations that were performed. Table 2.3 provides a brief summary of how each of the fixed effects and covariates was calculated, and what unit they were measured in.

EFFECT NAME	FIXED EFFECT / COVARIATE	DESCRIPTION
Alcohol	Fixed effect	Three categories; - No alcohol or less than RDA - Drinks over RDA in one type of alcohol - Drinks over RDA in two types of alcohol *RDA is intake of 0.5l beer, 0.2l of wine/bevanda, or 0.3l of hard liquor
Smoking	Fixed effect	Five categories; - Never smoked - Former smoker, stopped for over five years - Former smoker, stopped for less than five years - Current smoker for less than 10 years - Current smoker for over 10 years

Carbohydrate index	Covariate	Number (5-25) calculated from five questionnaire questions
Healthy food consumption	Covariate	Number (6-30) calculated from six questionnaire questions
Physical activity	Covariate	Number (1-4, at 0.5 intervals) calculated from two questionnaire questions
Socio-economic status	Covariate	Number (1-16) from one questionnaire question
Years schooling	Covariate	Number (2-22) from one questionnaire question

Table 2.3. Fixed effects and covariates, and how they were calculated for use in analyses.

2.3.2 Heritabilities

Heritability estimates were calculated during the first stage of GRAMMAR. A total of eight traits had a heritability that was not significantly greater than zero: ABPI, Brach Wid, Bra R, Calcium, Cortisol, Creat, GHQ and HDL. Significant heritabilities ranged from 0.179 (standard error 0.101) for Diast1, to 0.970 (standard error 0.067) for height. All heritability estimates with their standard errors and p-values can be found in Table 2.4.

TRAIT	HERITABILITY	S. E.	P-VALUE
ABPI	0.109	0.112	0.166
Albumin	0.662	0.091	2.98×10^{-8}
BMI	0.459	0.108	1.12×10^{-4}
Brach Cir	0.770	0.094	$1.11e^{-12}$
Brach Wid	0.166	0.100	0.051
Bra L	0.296	0.108	3.68×10^{-3}
Bra R	0.166	0.106	0.059
Calcium	0.101	0.106	0.169
Cholesterol	0.187	0.103	0.037
Chol ratio	0.203	0.102	0.024
Cortisol	0.080	0.108	0.233
Creat	0.047	0.093	0.309

Ddimer	0.290	0.097	1.64×10^{-3}
Diast	0.179	0.101	0.040
Dor L	0.250	0.112	0.014
Dor R	0.354	0.113	1.26×10^{-3}
Epqe	0.390	0.103	9.04×10^{-4}
Epqn	0.263	0.112	0.022
Fev	0.305	0.106	1.69×10^{-3}
Fib	0.314	0.122	6.04×10^{-3}
Fib 2	0.311	0.118	4.94×10^{-3}
GHQ	0.113	0.102	0.136
Glucose	0.594	0.108	2.45×10^{-7}
HDL	0.090	0.097	0.176
Height	0.970	0.067	~0
Hip Cir	0.503	0.106	3.73×10^{-6}
Hip-Waist	0.335	0.106	1.11×10^{-3}
LDL	0.209	0.106	0.020
PulseP	0.383	0.107	3.02×10^{-4}
React	0.273	0.110	7.55×10^{-3}
Resist	0.265	0.096	7.55×10^{-3}
Skin B	0.915	0.072	1.27×10^{-14}
Skin T	0.614	0.097	1.89×10^{-7}
Subscap	0.848	0.075	~0
Suprail	0.671	0.094	2.19×10^{-10}
Syst	0.237	0.107	0.026
Tib L	0.328	0.116	3.07×10^{-3}
Tib R	0.284	0.115	7.76×10^{-3}
TPA	0.323	0.107	1.59×10^{-3}
Trigly	0.275	0.134	4.15×10^{-3}
Uric acid	0.389	0.105	3.62×10^{-4}
vWF	0.621	0.099	7.14×10^{-9}
Waist Cir	0.230	0.131	0.020
Weight	0.487	0.127	8.42×10^{-3}

Table 2.4.Table showing heritabilities and standard errors for each trait.

2.3.3 Genome wide analysis results – additive model

Vast quantities of results were produced in this GWAS since each trait was tested at 307,951 SNPs, for each of two models (additive and genotypic). Consequently, only results exceeding the significance level of 1×10^{-6} are presented. This is slightly more relaxed than the Bonferroni corrected threshold of $p \leq 1.62\times10^{-7}$, although as already

noted, Bonferroni corrections tend to be too stringent due to the assumption of independence of SNPs. For an initial look at the results, it is justifiable to consider slightly less significant results since there may be a strong signal just short of reaching Bonferroni significance. The results from step two of GRAMMAR are shown in Table 2.5. Results from the additive model have been corrected using genomic control as discussed previously. A number of results from the additive model provide evidence for putative QTL for a variety of traits. In particular, there are strong signals of associations with uric acid on chromosome 4, vWF on chromosome 9, and creatinine on chromosome 5. At this stage, these associations are the most convincing due to very high significance for at least one SNP, and the presence of multiple other SNPs supporting the associations at these loci.

TRAIT	ALLELIC MODEL			GENOTYPIC MODEL		
	SNP name	Chr	P-value	SNP name	Chr	P-value
Brach Cir	rs6974152	7	1.29×10^{-7}	rs6448326	4	1.01×10^{-13}
	rs2899046	4	1.62×10^{-7}	rs10502942	18	1.37×10^{-13}
				rs7649544	3	2.54×10^{-10}
				rs1159851	1	4.71×10^{-10}
				rs9873442	3	4.96×10^{-10}
				* List truncated by 34 *		
Brach Wid	rs904554	11	~0	rs6656902	1	~0
	rs17067136	8	4.23×10^{-12}	rs6534405	4	~0
	rs2131002	4	3.27×10^{-11}	rs10484941	6	~0
	rs7950298	11	1.41×10^{-8}	rs10795659	10	~0
	rs613836	1	2.50×10^{-8}	rs10905409	10	1.11×10^{-16}
	* List truncated by 11 *			* List truncated by 521 *		
Bra R	rs7786279	7	5.28×10^{-7}			
Chol ratio	rs10875171	1	1.68×10^{-7}	rs10875171	1	1.00×10^{-6}
Cortisol	rs10826151	10	1.57×10^{-7}	rs10826151	10	8.18×10^{-7}
	rs3734061	5	3.18×10^{-14}	rs13103146	4	~0
	rs11959439	5	3.24×10^{-9}	rs11959439	5	~0
	rs1911216	14	9.80×10^{-8}	rs3799884	6	~0

Creat	rs2153527	1	1.88×10^{-7}	rs9472810	6	~0
	rs2301472	1	2.34×10^{-7}	rs7378011	4	1.11×10^{-16}
	* List truncated by 8 *			* List truncated by 212 *		
Ddimer	rs2022309	1	5.63×10^{-7}			
	rs6585454	10	9.60×10^{-7}			
Dor L				rs4126472	10	7.71×10^{-7}
Fev	rs1925324	1	9.53×10^{-7}	rs11652164	17	6.07×10^{-7}
				rs238342	13	8.90×10^{-7}
Fib 2	rs11695082	2	7.98×10^{-7}			
HDL				rs7315833	12	4.62×10^{-7}
				rs1441541	5	6.51×10^{-7}
Hip Cir				rs17135557	7	8.64×10^{-10}
				rs11766744	7	4.91×10^{-7}
				rs677214	1	4.93×10^{-7}
Hip-Waist				rs3843354	3	1.13×10^{-7}
Resist	rs2319188	3	5.57×10^{-7}	rs2180621	6	2.01×10^{-7}
				rs2319188	3	7.79×10^{-7}
Tib L				rs9842344	3	1.77×10^{-8}
				rs10011689	4	6.53×10^{-8}
				rs13013285	2	8.29×10^{-8}
				rs405970	20	8.75×10^{-8}
				rs6017819	20	9.37×10^{-8}
				* List truncated by 8 *		
Tib R				rs9379722	6	2.93×10^{-8}
				rs12583158	13	3.20×10^{-8}
				rs1405040	10	3.56×10^{-8}
Uric acid	rs737267	4	1.22×10^{-9}	rs737267	4	2.21×10^{-8}
	rs13129697	4	6.06×10^{-9}	rs13129697	4	9.54×10^{-8}
	rs6449213	4	1.36×10^{-8}	rs6449213	4	1.69×10^{-7}
	rs1014290	4	2.02×10^{-8}	rs1014290	4	2.50×10^{-7}
	rs13131257	4	2.42×10^{-8}	rs13131257	4	3.56×10^{-7}
	* List truncated by 1 *					
vWF	rs657152	9	~ 0	rs657152	9	~ 0
	rs505922	9	~ 0	rs505922	9	~ 0
	rs630014	9	6.95×10^{-11}	rs630014	9	3.90×10^{-10}
	rs8176749	9	3.98×10^{-9}	rs8176749	9	8.08×10^{-8}
	rs8176746	9	5.62×10^{-9}	rs8176746	9	1.09×10^{-7}
	* List truncated by 4 *			* List truncated by 1 *		
Waist Cir	rs4954723	2	3.97×10^{-7}	rs4954723	2	6.90×10^{-7}
				rs4347759	2	7.64×10^{-7}
Weight	rs2633254	2	3.07×10^{-7}			

Table 2.5 Table showing p-value and chromosome from step two of GRAMMAR for all SNPs more significant than 1×10^{-6}. Where "~ 0" is displayed, the actual figure was too small for GenABEL to report. Where the list is longer than five SNPs for any trait/model combination, the list is truncated, and the number of missing SNPs is indicated.

Figures 2.1 - 2.3 show results for the specific trait / chromosome combinations mentioned above. Physical distance (Mb) is plotted on the x-axis and $-\log_{10}$ p-value on the y-axis. Figure 2.1 shows results from analysis of uric acid on chromosome 4. There is a clear signal for association around 11Mb into the chromosome, as five SNPs exceed Bonferroni significance, the most significant of which (rs737267) with a p-value of 1.22×10^{-9}. There are also a number of supporting SNPs in the region that do not quite reach this level of significance, but lend credence to the more significant results. Figure 2.2 shows the vWF results on chromosome 9. The most obvious peak is at around 135Mb, and is supported by a total of six SNPs exceeding Bonferroni significance (two marked in red on the figure due to p-values of approximately zero from R), and there are also a couple of smaller peaks at 6Mb and 81Mb that may be worth examining closer. Results of chromosome 5 for creatinine are displayed in Figure 2.3. The most significant hit is an isolated SNP around 151Mb into the chromosome. There is also a cluster of three SNPs falling to either side of 120Mb, two of which are close to Bonferroni significance and the third of which exceeds it. This peak at 120Mb looks promising, although the evidence is less conclusive than for either UA or vWF.

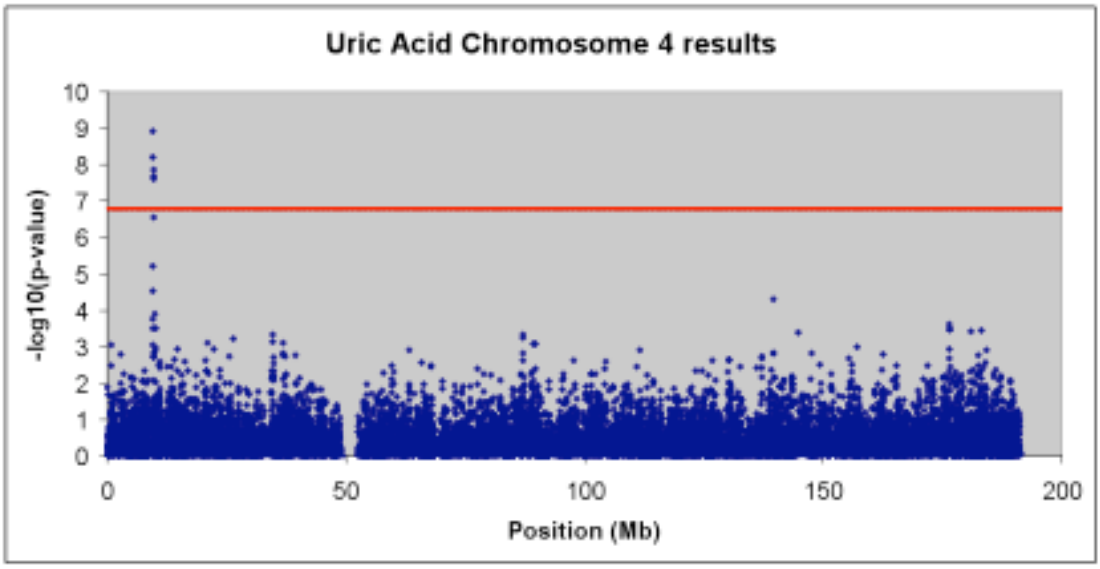


Figure 2.1 Results for chromosome 4 for uric acid. Results are from a 1df allelic test using step two of GRAMMAR. The red line indicates Bonferroni significance.

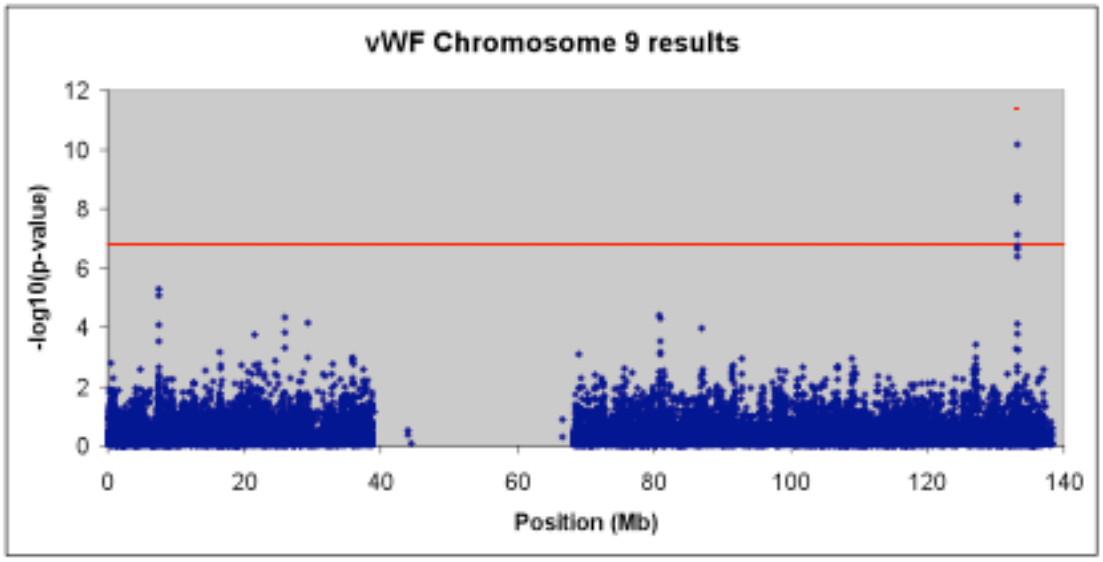


Figure 2.2 Results for chromosome 9 for vWF. Results are from a 1df allelic test using step two of GRAMMAR. The red line indicates Bonferroni significance. The two red points at 133.17Mb and 133.18Mb correspond to p-values set to zero by R.

It is often difficult to interpret associations consisting of only a single SNP, such as the one to rs3734061 at 151Mb on chromosome 5 mentioned above (Figure 2.3). While it is possible they represent true signals, there is a greater chance of the finding

being either spurious or artefactual than for loci where multiple SNPs show association. This is because common SNPs like those on the panel used for this study usually have at least moderate LD to other nearby SNPs, and consequently two typed SNPs in moderate to high LD are expected to tag a third variant (i.e., a QTL) almost as well as each other. An association where no flanking SNPs show at least suggestive significance thus has a higher chance of being false. It should be noted that multiple close false positives or a single true positive may occur depending on the local LD structure, however the probability of this occurring is smaller. The most likely cause of an artefactual association is poor genotype calling, whereby the calling algorithm used to assign genotypes to individuals from allelic intensity data performs badly. Intensity data for this study were unavailable, therefore it was not possible to verify that this was the case for the association to creatinine on chromosome 5. However, in this case, LD to SNPs nearby rs3734061 was low (highest r^2 was 0.152 to rs2278370), meaning that there is a greater chance of this particular association being genuine.

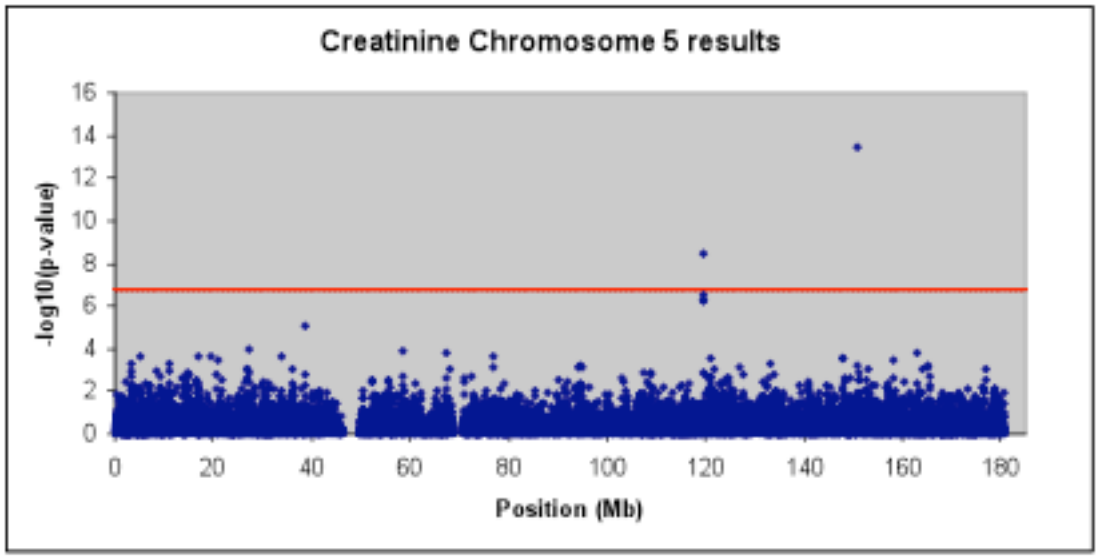


Figure 2.3 Results for chromosome 5 for creatinine. Results are from a 1df allelic test using step two of GRAMMAR. The red line indicates Bonferroni significance.

In addition to peaks consisting of at least one significant SNP and a number of suggestive SNPs such as those mentioned above, there are numerous cases of well-supported associations where none of the individual hits reach Bonferroni significance. In these situations it could be worth investigating further as the associations may still be real, however these associations would not be a top priority. Accordingly, the most promising results from the additive tests show evidence for a QTL on chromosome 4 affecting uric acid levels, a QTL on chromosome 9 affecting vWF levels, and a QTL on chromosome 5 affecting creatinine levels.

2.3.4 Genome-wide analysis results – genotypic model

Results for the genotypic model are from 2df tests for overall association of each SNP. Initially, it appears that there are numerous positive results from the genotypic model. However, for two traits in particular – Brach Wid and Creat – the number of significant SNPs for the genotypic test begins to look suspicious. In the most extreme case, that of brachial width, there are a total of 526 SNPs with a p-value of less than 1×10^{-6}. This is clearly inflated in comparison to the number of hits for the vast majority of traits. The only other trait to have such an extreme number of significant results is creatinine, totalling 217, which is also remarkably high. Figure 2.3 shows the additive model results for creatinine on chromosome 5, therefore the same trait and chromosome combination was selected to illustrate a typical example of the results obtained for creatinine and brachial width using the genotypic model. This is

shown in Figure 2.4. Given that the only difference in Figures 2.3 and 2.4 is the model of analysis, the difference in these figures is dramatic.

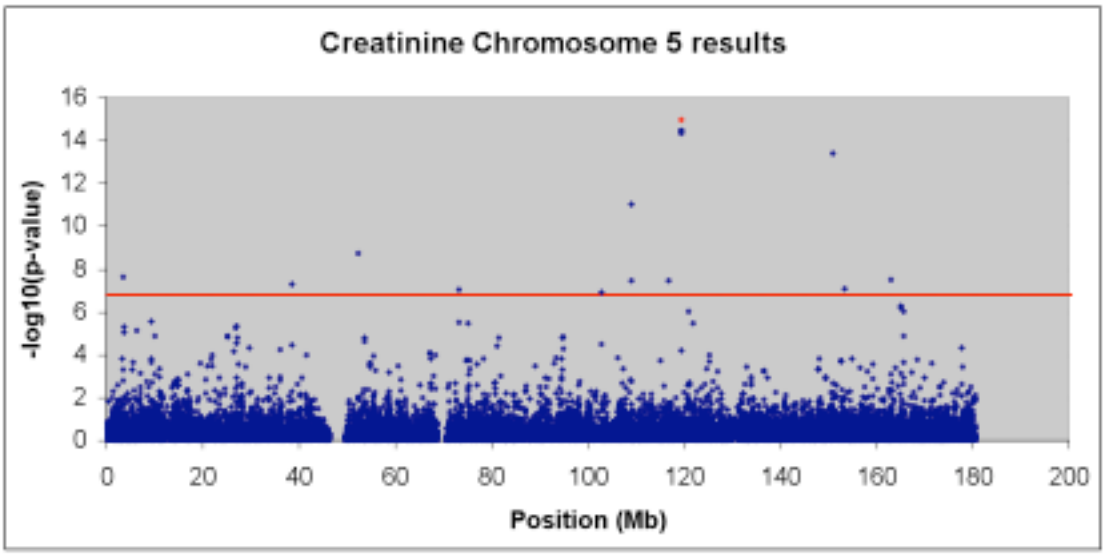


Figure 2.4 Results for chromosome 5 for creatinine. Results are from a 2df genotypic test using step two of GRAMMAR. The red line indicates Bonferroni significance. The red point at 119.43Mb corresponds to a p-value set to zero by R.

In addition to the number of associations for these traits, another factor decreasing the likelihood of their validity is how scattered the hits are across chromosomes. In Figure 2.4, there are several places where Bonferroni significance is surpassed, however there are no well-supported peaks. The strongest convincing signals are at around 100-125Mb and 150-170Mb, as these do have multiple suggestive SNPs. There is also a SNP at 119Mb represented by a red point (again, due to a limitation of R) on Figure 2.4, further validating that peak. However, since the reliability of these results is questionable no strong conclusions can be based upon them.

The strange results mentioned above are restricted to just creatinine and brachial width, suggesting that there is some phenomenon affecting only these two traits. For

example, the genotypic model had the same top five SNPs (in the same order) on chromosome 4 for uric acid, and also the same associated SNPs on chromosome 9 for vWF, as the additive model, therefore there is reason to believe that the genotypic test can perform well. There are also a number of other traits for which the genotypic test looks to have identified interesting associations. There is evidence for associations to Hip Cir, Tib L and Brach Cir for example, as seen in Figures 2.5 - 2.7. The hip circumference results from chromosome 7 (Figure 2.5) show a putative association 100Mb into the chromosome. There is a single SNP above Bonferroni significance, and this SNP has a p-value three orders of magnitude beyond the threshold. In addition, there are a number of other SNPs above background noise at the same location.

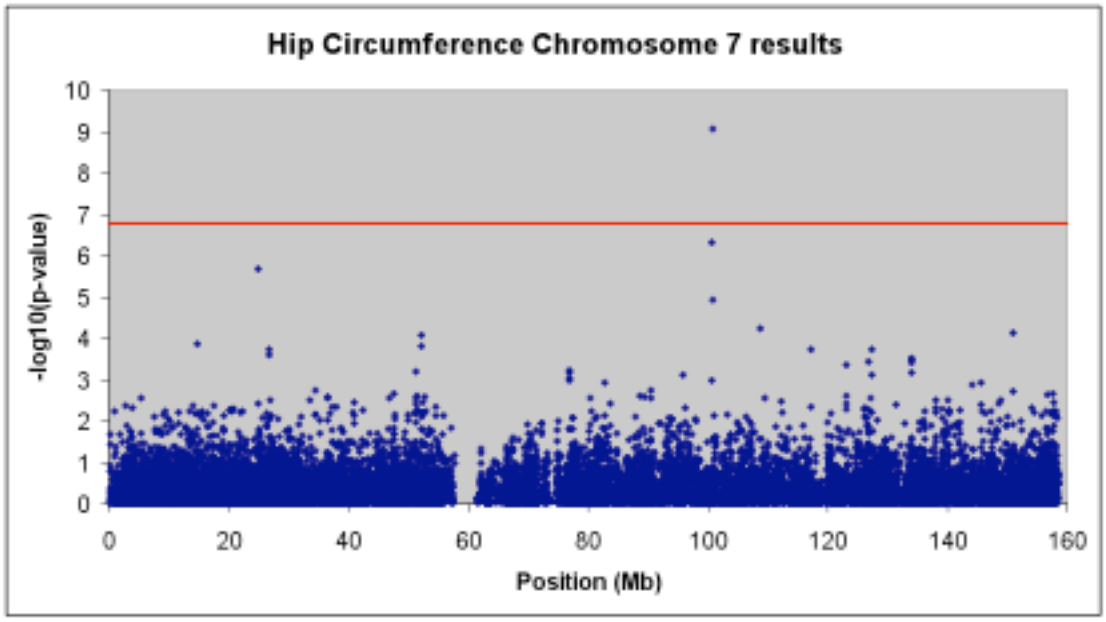


Figure 2.5 Results for chromosome 7 for hip circumference. Results are from a 2df genotypic test using step two of GRAMMAR. The red line indicates Bonferroni significance.

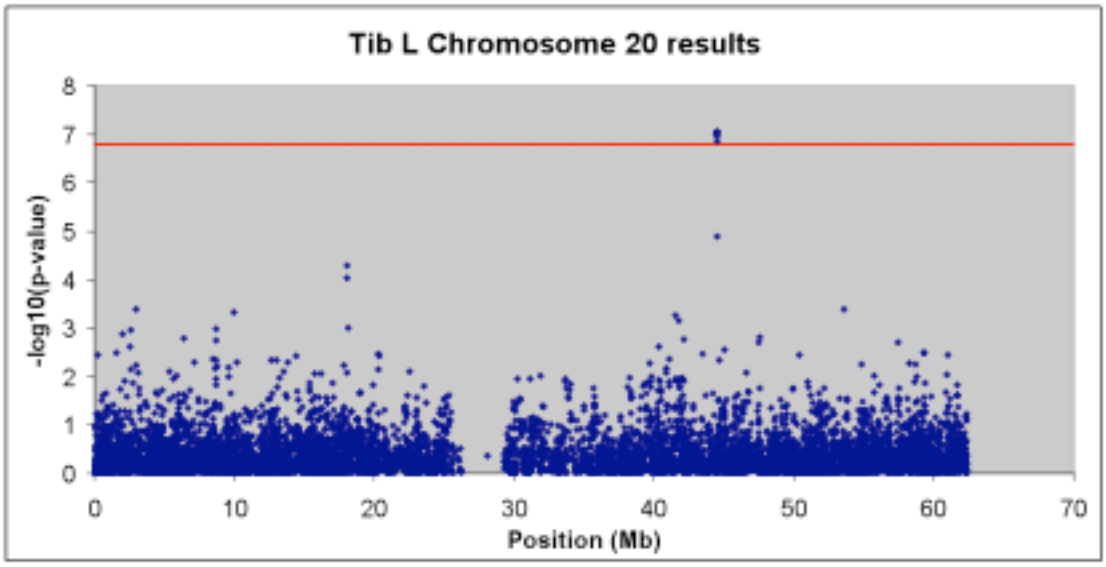


Figure 2.6 Results for chromosome 20 for systolic pressure at tibialis posterior left. Results are from a 2df genotypic test using step two of GRAMMAR. The red line indicates Bonferroni significance.

Figure 2.6 shows the Tib L results for chromosome 20. There is a peak with three SNPs exceeding Bonferroni significance standing above the background noise around 44Mb into the chromosome. There are also a number of peaks above background noise on the short arm of chromosome 20, although these fall well short of significance. The final graph (Figure 2.7) shows the Brach Cir results from chromosome 1. Here, the most prominent peak is at 67Mb, with four hits more extreme than the Bonferroni threshold. There are no additional suggestive SNPs in the flanking region, but the four hits themselves provide good evidence for association. There is also a suggestive peak consisting of three SNPs at around 98Mb, and another one at 202Mb where one SNP almost reaches Bonferroni significance.

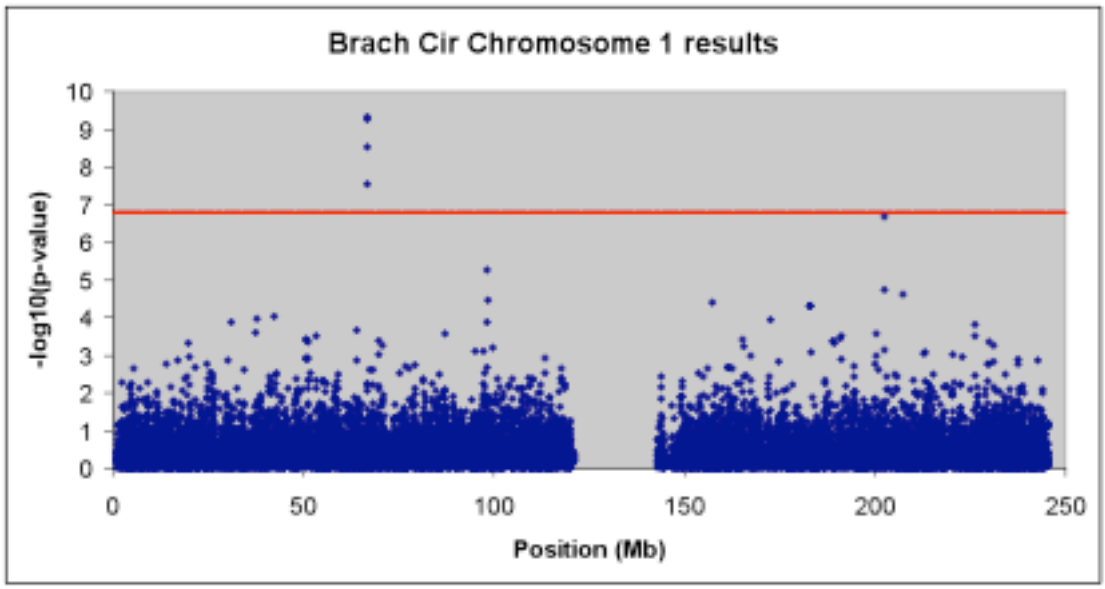


Figure 2.7 Results for chromosome 1 for brachial circumference. Results are from a 2df genotypic test using step two of GRAMMAR. The red line indicates Bonferroni significance.

2.3.5 Brach width and creatinine

2.3.5.1 Analysis revisited

A mini follow-up on brachial width and creatinine was performed to investigate the strange results from the genotypic model more closely. Genotypic analysis for these traits was repeated, however on this occasion a more stringent limit on the proportion of observations in each genotype class for each SNP was imposed. This was to eliminate the possibility that a large proportion of significant SNPs were produced due to sensitivity of the score test to SNPs with an extremely rare genotype class. Where previously SNPs with a rare genotype count of less than 10 were set to missing, a new threshold of 50 was set (a frequency of approximately 0.05), and genotypic analysis was then performed as described in the methods.

Figure 2.8 shows the results for creatinine on chromosome 5 from the new analysis. On this occasion, unlike the previous genotypic analysis for this trait / chromosome, there are no results reaching Bonferroni significance. Results from the new analysis are very different from those in Figure 2.4 for example, with the most extreme $-\log_{10}$ p-value reaching just under five, and the results now appear similar to the genotypic analysis results for most other traits. No significant associations for either creatinine or brachial width were found in the new analysis of these traits however. Interestingly, there is also no evidence in Figure 2.8 to support the significant associations on chromosome 5 for creatinine with the additive model (Figure 2.3). This does not necessarily mean the significant results from the additive model are false positives however, since the genotypic test will have reduced power compared to the additive test for loci acting in an additive manner.

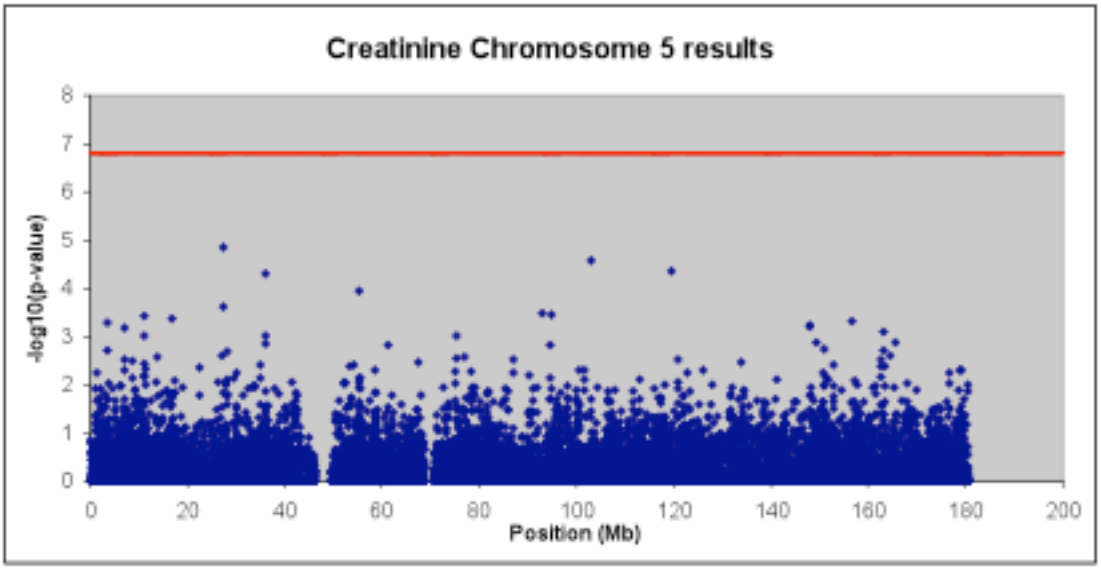
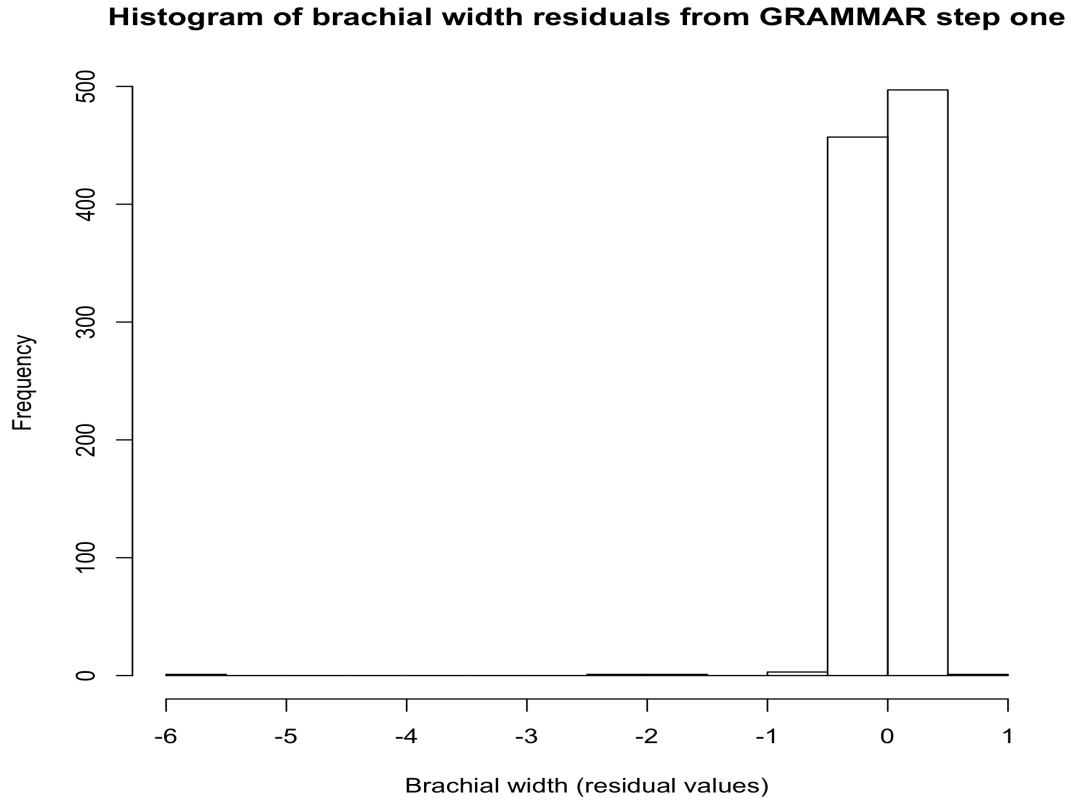
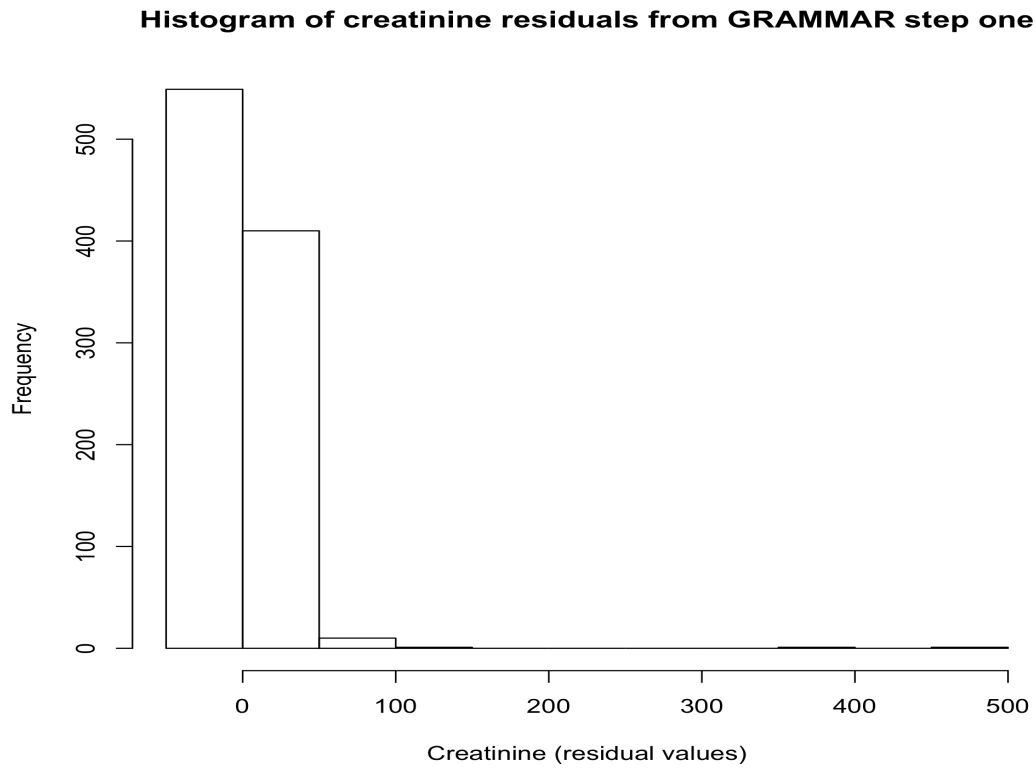


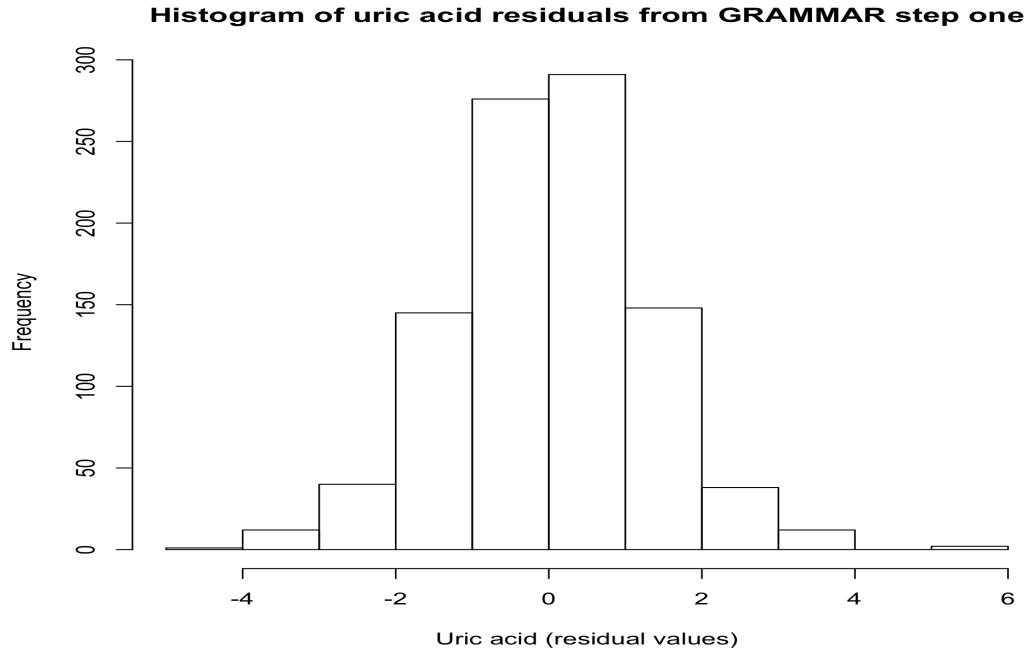
Figure 2.8 Results for chromosome 5 for creatinine. SNPs with one genotype class with less than 50 occurrences are removed from the dataset (set to missing). Results are from a 2df genotypic test using step two of GRAMMAR. The red line indicates Bonferroni significance.

2.3.5.2 Trait investigation

Brachial width and creatinine were both traits whose residuals (i.e., trait value after correcting for polygenic effects and other covariates) were not normally distributed, and for which it was difficult to find a transformation leading to normalisation of the residuals. Figures 2.9 and 2.10 illustrate the distribution of the residuals produced from GRAMMAR step one for Brach Wid and Creat respectively. Figure 2.11 shows the residuals produced for uric acid, as a comparison. Although regression is thought to be fairly robust concerning trait distribution, normality is nevertheless an assumption of the regression model fitted in step two of GRAMMAR. Therefore it would be interesting to see if a combination of rare genotype classes and poor trait distribution were causing the numerous false positives observed for Brach Wid and Creat.







Figures 2.9 - 2.11 Distributions of the brachial width, creatinine and uric acid residuals from GRAMMAR step one.

The frequency of the rare genotype class for each of the top five hits (genome-wide, not only chromosome 5) from the Creat genotypic tests was recorded. All five SNPs had a rare genotype class with frequency less than 0.05 (the maximum was 0.018). Five SNPs matched for chromosome were then selected from each of three different categories of rare genotype class; 0.05 - 0.1, 0.1 - 0.2, and over 0.2. For each of the 20 SNPs in total, permutation analyses were performed with respect to two traits; the non-normally distributed creatinine residuals, and the normally distributed uric acid residuals. In each case, SNP genotypes were permuted with respect to phenotypes 10,000 times. Permutations across SNPs were performed independently to avoid correlations between SNPs due to LD (since some SNPs are located on the same chromosome). Simple linear regression was performed for each SNP in R for each of the permutations, and the p-value resulting from the F-test for an overall SNP effect was recorded.

Table 2.6 shows results from these analyses, along with basic descriptors of the SNPs selected. The categories the SNPs were divided into are defined in the third column in the table. The fourth and fifth columns show the means of the F-statistic over the 10,000 permutations for each SNP for creatinine and uric acid respectively. The overall mean for a given category of SNPs is in bold beneath the individual SNP means for that category.

SNP	MAF	FREQ RARE GENOTYPE CLASS	MEAN F-STAT CREATININE	MEAN F-STAT URIC ACID
rs13103146	0.1057	0.01708	1.043	0.990
rs7378011	0.1066	0.01751	1.013	1.003
rs11959439	0.0928	0.01116	1.066	0.990
rs3799884	0.1325	0.01421	1.031	1.015
rs9472810	0.1210	0.01124	1.057	0.988
			1.042	**0.997**
rs11727494	0.2650	0.07005	0.995	1.006
rs6828802	0.2670	0.07411	1.027	1.004
rs10076494	0.2819	0.08722	0.979	1.014
rs7756332	0.3063	0.09635	1.007	0.995
rs6596860	0.2444	0.05386	1.007	1.001
			1.003	**1.004**
rs7682616	0.3373	0.12255	1.007	1.013
rs7666919	0.2969	0.10480	1.001	0.996
rs7446851	0.2915	0.10153	1.006	1.004
rs6932895	0.3316	0.11055	1.001	1.005
rs727056	0.4190	0.17591	0.997	1.009
			1.002	**1.005**
rs10027536	0.4815	0.24820	1.007	1.018
rs7665922	0.4770	0.28208	1.000	1.003
rs4956987	0.4959	0.25922	0.997	1.023
rs1540771	0.4613	0.21261	0.998	1.009
rs6937708	0.4505	0.22914	0.996	1.012
			1.000	**1.013**

Table 2.6. Results for the 20 SNPs used for permutation analysis to assess effects of trait distribution and genotype class on regression analysis. The table shows SNP name, MAF, rare genotype frequency and the means of 10,000 permutations with respect to each of creatinine and uric acid. Numbers in bold represent the overall mean for that genotype class (i.e., a mean of 50,000 permutations). The four groups of five SNPs correspond to different categories of rare genotype class frequency, and within each five SNPs the first two are from chromosome 4, the third is from chromosome 5 and the last two are from chromosome 6.

The expectation of an F-statistic with two and (effectively) infinite degrees of freedom is one. The most striking result is how high the overall F-statistic mean is for SNPs with a rare genotype frequency (<5%) when analysed for creatinine. At 1.042, it is by far the highest overall mean, and the discrepancy between this overall mean and the overall mean for the same five SNPs analysed for uric acid is the biggest discrepancy of all four categories. A T-test indicates that the mean of 1.042 is

significantly different from one ($p = 1.97\times10^{-5}$). The only other overall mean which was noticeably higher was for the >20% rare genotype frequency category analysed with uric acid, at 1.013. This is the only other overall mean that is also significantly different from the expected F-statistic of one ($p = 0.003$).

Two-sample T-tests were used to determine whether the overall means were significantly different from one another between categories. When considering creatinine, three of the six possible comparisons yielded significant differences, and these were all when comparing the <5% rare genotype class to the others. The p-values for comparison of means of the <5% class with each ascending frequency class are 3.9×10^{-4}, 2.2×10^{-4} and 7.6×10^{-5} respectively, showing that the mean for this class is significantly higher than the means of all other classes. None of the other comparisons for creatinine are significant. The same tests for the uric acid results produced one significant result at the 5% level, for the comparison of the <5% and >20% classes (p-value 0.0114).

2.3.6 Putative genes

Ensembl (Hubbard et al., 2007) was used in an attempt to identify candidate genes for each trait in the regions suggesting associations in the results. All hits for uric acid fell within a gene called *SLC2A9* (also known as *GLUT9*). This gene is a member of the solute carrier family 2, which is a facilitative glucose transporter family. There are two transcripts of *SLC2A9*, one having a shorter N-terminus and different start codon. This shorter transcript is known as GLUT9deltaN.

As was the case with uric acid, all vWF significant hits fell within the same region. This region is located very close to the ABO blood group gene. It is already well documented that the *ABO* gene has a significant effect on vWF levels (Souto et al., 2000), therefore our results replicate that association. The SNP panel used in this study does not include the polymorphism known to cause the difference in blood group, although it is likely that this is where the signal originates, and that the significantly associated SNPs in this study are in high LD with it.

Interestingly, the top hit for creatinine on chromosome 5 (the lone SNP around 151Mb into the chromosome) codes for a non-synonymous mutation in coding sequence. The change occurs in the *FAT2* gene, the second human homolog of the Drosophila fat gene, which encodes a tumour suppressor essential for controlling cell proliferation during Drosophila development. The mutation is a transition from G to A, which causes the amino acid change of proline (P) to serine (S) (a conservative change, as both amino acids have neutral side chain charges). The protein is a member of the cadherin superfamily, and most likely functions as a cell adhesion molecule, controlling cell proliferation and playing an important role in cerebellum development. At this stage it is hard to see a link between creatinine and the *FAT2* gene, since creatinine is largely an indicator of renal function. The other significant result for creatinine on chromosome 5 involve SNPs not in the vicinity of any genes. There are no genes for over 400Kb in either direction of the hit SNP, effectively ruling out the possibility of *cis* effects between these SNPs and genes that may affect creatinine. The possibility of these SNPs tagging a variant involved in long-distance

gene regulation remains, however this would be hard to confirm without extensive research.

Both SNPs for the hip circumference significant result fall within the gene *EMID2*. This gene encodes a collagen alpha-1 chain precursor protein thought to be involved in both biological processes and cellular component activities. How this would relate to hip circumference is hard to see at this stage. Similarly, for Tib L the top hits appear to be located in a gene called Zinc Finger Protein 663, and there is limited information on the proposed function of this gene, therefore it is unclear whether an effect on Tib L is likely.

2.4. DISCUSSION

2.4.1 Heritabilities

Of the 44 traits analysed, 36 had heritabilities significantly different from zero. Significant heritabilities are important because they confirm that a significant proportion of trait variation is controlled by additive genetic effects, meaning it should be possible to identify the underlying genetic factors. Higher heritabilities indicate more genetic control of the trait relative to environmental (or non-genetic) sources of variation, which could mean either more genes involved than in traits with a lower heritability, or that the genes involved have larger effect sizes. Either way, appropriately powered GWA studies have a better chance of detecting QTL for traits with high heritability. It is also encouraging to note that many heritabilities reported

here are similar to those cited in the relevant literature, for example height (Silventoinen et al., 2003) and BMI (Mora et al., 2005).

2.4.2 GWA Results

There were not many SNPs reaching the Bonferroni corrected genome-wide significance level after analysis at step two, i.e., analysing residuals. This may partly be a consequence of "factoring out" some of the SNP effect due to the method of analysis, but genomic control should largely correct for this. It is more likely that failure to detect many associations is a consequence of low power for this study. Analysis was performed out on approximately 966 individuals (the exact number varies across traits due to differing patterns of missing data), however, considering that around 500 of these are in pedigrees the effective sample size is considerably less, since related individuals cannot be classed as independent. As a result, it seems likely that power is limited to detect very small effects in this study.

This study did detect some significant results however, of which the uric acid and vWF results were the most exciting. After analysis using GRAMMAR step two, the most significant hit for vWF was assigned a p-value of approximately zero by R, and that for uric acid was 1.22×10^{-9}. Both these associations were supported by numerous flanking SNPs also exceeding Bonferroni significance. This outcome is expected due to higher levels of LD between physically close SNPs, since SNPs in high LD will generally tag the same causative variant. For example, two of the SNPs showing strongest support of the top SNP for uric acid, rs13131257 and rs6449213, had

pairwise r^2 of 0.83 and 0.79 with rs737267 in this dataset. If there were no SNPs in high LD with either the causal variant or the primary tagging SNP, then no supporting flanking SNPs would be present, however this is less likely for physically close common SNPs. Single significant SNPs can also be a consequence of artefacts such as poorly called genotypes, and these are likely to explain some instances of individual SNPs exhibiting a low p-value.

The fact that the vWF hits were almost certainly due to the already known association between vWF and the ABO blood group system meant that these results were no longer of primary concern for further investigation. Although there is a slight possibility that the effect on vWF is enhanced by another variant located very close to the *ABO* gene, this is unlikely. Contrastingly, the gene that the UA hits fall within has no previous links to uric acid metabolism, but does look like a suitable candidate due to its transporter-like function. Combined with a very strong association signal, this is a strong candidate to follow up in a replication study.

Both results above came from the additive model, however there was also a promising result for hip circumference from the genotypic test. The top hit is three orders of magnitude smaller (i.e., more significant) than the required Bonferroni significance threshold of 1.62×10^{-7}, and two other SNPs at this locus are also suggestive. However, these results are more difficult to interpret in light of the nature of hip circumference as a trait, since it can be thought of as a composite measure of two other underlying traits – bone structure and fat around the hips. It is possible that bone structure around the hips is under evolutionary pressure in women due to child-

birth for example, and both composite measures are correlated with hip circumference. Hip fat is likely to be under both genetic and environmental control, and gene-by-environment interactions are also likely to be involved. Consequently, hip circumference is likely to be under similar genetic control to other weight / fat measures such as waist circumference, weight, biceps and triceps skin fold thickness and BMI, however there was no concordance in the results for any of these traits. There were no SNPs near significance for these traits on chromosome 7. As a result, although the hip circumference results would be followed up more extensively in due course, it was decided that the other results would be more thoroughly investigated first.

2.4.3 Notes on distributions

The investigation into the effects of non-normal trait distribution and different rare genotype classes was very instructive in interpreting certain results from this GWAS. There was statistically significant inflation in the mean of the test statistic for SNPs whose rare genotype frequency was under 0.05 when analysed for a trait that was non-normally distributed. The majority of the data for the Creat distribution (Figure 2.10) falls below a certain threshold, with a relatively small number of higher outliers that exceed this. With this distribution, it is possible that over a large number of SNPs in which one of the three genotypes is rare, by chance all or most of the rare genotypes may be present in the high outliers. This would cause a spurious association between that genotype and the trait. The same SNP analysed under an additive model would not necessarily be found to be associated because of the large

number of alleles in heterozygotes that are present in individuals with lower trait values.

Results of the permutation appear to validate this hypothesis, with the test statistic on average being approximately 0.04 higher for SNPs with the lowest rare genotype frequency class. If these 20 SNPs comprised the entire dataset (as might be the case if they were part of a candidate gene study for example), a Bonferroni significance threshold of 0.0025 would be set. Performing 200,000 permutations for the trait (10,000 each for 20 SNPs) would mean that an expected number of 500 tests would be found significant purely by chance. Of the 200,000 permutations for creatinine, there were 1,881 tests that exceeded the 0.0025 threshold, almost 1,400 more than expected, of which 1,350 were SNPs in the <5% rare genotype category. In each of the three remaining categories (in ascending order of rare genotype frequency) the number of hits exceeding the threshold is 296, 171 and 64 respectively. Interestingly, applying the Bonferroni corrected threshold used in this GWA study (1.62×10^{-7}), there would be 221 SNPs found significant, with only one of them not being in the <5% rare genotype frequency category (this one is in the 5-10% category).

In comparison, results from performing permutations on uric acid are much more similar to the expectation. Of the 200,000 tests, only 523 are found significant, where by chance 500 are expected, and of those the breakdown is 116, 132, 142 and 133 in ascending rare genotype frequency class order. There appears to be a slight reduction in the number of hits reaching 0.0025 in the rarest genotype frequency class compared to the other three categories. With creatinine, the number of hits exceeding

0.0025 gets smaller as the frequency of the rare genotype increases, corresponding to the decreased chance that most of the rare genotypes will be paired with a high trait value. Compared with the 221 hits found with creatinine, there would be no results significant at the Bonferroni level when analysing uric acid.

These results indicate that when there is one rare genotype class the test for association of genotypes fitted in a regression model is not reliable, although this is also clearly tied in with having a trait that is not normally distributed. Trait normality is one of the assumptions made when using regression models, however regression is generally considered fairly robust to this assumption. Nevertheless, there is also a slight increase in the number of significant hits found in Creat and Brach Wid even using the additive model compared to other traits, although this is nothing like that seen for the genotypic model. Clearly the genotypic test is far more sensitive to the distribution of the trait than the additive model. With this in mind, it may not be worth using the genotypic model for poorly distributed traits. Other potential solutions are removing SNPs with very extreme rare genotype frequencies (as opposed to setting only the rare genotype to missing), or treating traits distributed in this way slightly differently. For example, it may be better to consider creatinine and brachial width as binary traits above and below some threshold instead of as quantitative traits.

2.4.4 Method of analysis

Data for this GWAS were analysed using a new technique for analysing genetic data in the presence of a pedigree, called GRAMMAR. This involves adjusting raw trait values for all relevant covariates, fixed effects and random effects before testing each SNP in a simple linear regression model. GRAMMAR is many orders of magnitude faster on a genome-wide scale than the gold standard measured genotype approach; the estimated time taken to complete a GWA scan using the measured genotype approach is in the region of 2.5 years for each trait (Aulchenko et al., 2007). In this study, computational time for analysis (up to step two) of each trait did not exceed ten minutes. The compromise made to enable this increase in speed was a slight decrease in power to detect QTL effects, due to removing some of this effect along with the polygenic variation in step one. This is largely corrected for by using the method of genomic control however, which accounts for cryptic relatedness and population substructure (Devlin and Roeder, 1999). Since effect of population substructure on the test statistic is constant throughout the genome, test statistic inflation due to substructure is the same for all markers (Bourgain and Genin, 2005). The principle is the same when removing polygenic variation, only in this case the test statistic experiences deflation rather than inflation. After correction, SNPs can be selected for analysis under the full measured genotype approach which will not suffer from test-statistic deflation. Given the time penalty for analysing every SNP with MG, GRAMMAR is a much more feasible alternative.

Considering the relatively few associations found during this GWAS so far, it may be valid to question the suitability of these data to association analyses as opposed to

linkage. The limited sample size and the fact that many participants are related means that these data are not ideally suited to an association study, however it should be noted that a recent study suggests that loss of power due to the latter is minimal (Visscher et al., 2008). Additionally, the data are even less suited for linkage analyses, since there are even less individuals (than the 966 available for association) that form part of a pedigree. Considering that power to detect small to moderate effect sizes is far greater in an association framework (Risch and Merikangas, 1996), GWA is thus justified as the best method of analysing this dataset. This project would clearly benefit from an increase in the number of participants however, and this was achieved over the summer of 2007, when a further 1,000 individuals from the island of Vis were recruited.

One aspect of this GWAS that proved successful was the use of intermediate quantitative phenotypes to analyse instead of clinical disease endpoints. This is reflected in the uric acid findings, as high uric acid levels, or hyperuricemia, is a common feature in individuals suffering from gout (although not all people with hyperuricemia go on to develop gout), therefore it is possible that the putative QTL identified in this study will also increase risk of gout. This shows how SNPs identified by analysis of an intermediate phenotype may provide information that is relevant to disease, and indicates that prospects for detection of QTL underlying common complex disease in future GWA studies look good.

3. CHAPTER 3

3.1. INTRODUCTION

3.1.1 Background

In the previous chapter a genome-wide association study (GWAS) finding multiple significant hits on chromosome 4 affecting uric acid levels was reported. In total five SNPs exceeded the Bonferroni-corrected nominal p-value threshold of 1.62×10^{-7}. These SNPs were identified using step two of the GRAMMAR procedure outlined in the second chapter (section 2.2.6.3), and all fell within introns of the same gene, *SLC2A9*. These results were particularly appealing to follow up not only due to their genome-wide significance, but also because *SLC2A9* is part of a known solute carrier family of genes. Although originally reported as a transporter of glucose (Phay et al., 2000), *SLC2A9* is nonetheless a gene that provides a biologically plausible reason for containing SNPs affecting uric acid level. Uric acid is considered to be of great public health significance due to the link it has with many prevalent population diseases (see for example, Heinig and Johnson, 2006), further enhancing this finding as one worth investigating.

The first stage in following up these SNPs is to perform a full measured genotype (MG) analysis for each of the significant SNPs, as implemented in step three of GRAMMAR. Assuming these results validate the original findings, other considerations may then be taken into account to investigate further. This could

include a closer inspection of appropriate covariates and fixed effects to fit in the model for example, or extrapolating the findings to other related traits. Performing a replication study in an independent cohort to confirm the validity of the results is also required.

3.1.2 Uric acid

Uric acid is the end product of purine metabolism in humans and the great apes. It is produced via oxidation of oxypurines by xanthine oxidase, which in other mammals is then further oxidised by the enzyme uricase into allantoin. Loss of hepatic uricase activity by humans and the great apes has therefore lead to uniquely high uric acid levels compared to other mammals, being around 200-500μmol/L as opposed to 3-120μmol/L. The American Medical Association considers a range of between 3.6 mg/dL (~214μmol/L) and 8.3 mg/dL (~494μmol/L) to be normal for humans.

Uric acid is important to study because it has been implicated in numerous afflictions of public health significance. Low uric acid levels have been associated with multiple sclerosis (Toncev et al., 2002), while high levels have been implicated in kidney stones (Heinig and Johnson, 2006), Lesch-Nyhan syndrome (Luo et al., 2006; Nyhan, 2005) and cardiovascular disease (Heinig and Johnson, 2006). Gout is the condition most often associated with uric acid however (Seegmiller et al., 1963), where high uric acid levels, also known as hyperuricemia, greatly facilitate its onset. Gout is a type of arthritis that most commonly affects the big toe, but can also affect other joints - such as the ankle, heel, instep, knee, wrist, elbow, fingers and spine - causing

excruciating pain. Approximately 10% of individuals with hyperuricemia go on to develop gout, and >90% of subjects with gout have diminished fractional excretion of uric acid (FEUA) (Graessler et al., 2006). Given that 5-25% of the human population have impaired renal excretion leading to hyperuricemia (Becker and Jolly, 2006), it is clear that identifying QTL underlying uric acid levels is of immense importance to public health.

3.1.3 *SLC2A9*

All SNPs implicated in affecting uric acid levels from the initial genome-wide association (GWA) scan are found either within introns of the *SLC2A9* gene, or slightly 5' of it. *SLC2A9*, also called *GLUT9*, has previously been documented as a glucose transporter, and belongs to a family of genes known to be carriers of glucose and other solutes. Another gene within the solute carrier gene superfamily, *SLC22A12* (encoding URAT1), is a known urate transporter (Enomoto et al., 2002). There have also been reports of *SLC2A9* being expressed in human (Augustin et al., 2004) and mouse (Keembiyehetty et al., 2006) kidney tubules, further strengthening the link between the hit SNPs, uric acid and *SLC2A9*.

The *SLC2A9* gene codes for two transcripts; the longer version of the gene is approximately 214Kb long, (9,718,140 – 9,504,117; 4p16-p15.3) with 13 exons, whereas the shorter version, known as GLUT9deltaN, has a shorter N-terminus and only 12 exons, spanning 195.3Kb (9,699,382 – 9,504,117; 4p16-p15.3). Both forms of the gene are transcribed in a 5' – 3' direction. Of the five SNPs exceeding

Bonferroni significance for uric acid in our study, all were found within introns 3-7 of the *SLC2A9* gene. One of the aims of performing further studies into the region of interest is to narrow the search and draw more definite conclusions about the location of the quantitative trait nucleotide (QTN) responsible for this variation in uric acid level.

3.1.4 Follow-up research

The first step in following up the significant uric acid results is to use the MG approach (as implemented in GRAMMAR step three), as this provides more accurate estimates of significance and effect size for the polymorphisms tested (Aulchenko et al., 2007). In this step, SNPs are analysed with uric acid as the dependent variable, as opposed to residuals from a polygenic model as performed previously. Greater consideration of appropriate fixed effects and covariates was taken at this stage, to ensure that the best model to fit the data was used.

Assuming the MG results were significant, the next stage would be to look more closely at the associated SNPs in the context of the gene in question, and attempt to fine map the location of the causative polymorphism. Viewing the LD structure among the associated SNPs, using software such as HaploView (Barrett et al., 2005) for example, is one way of interrogating the locus to identify the QTN. Another way to help define the associated region is to test multiple SNPs. Accounting for the variation explained by one SNP may entirely remove the signal of association at other nearby SNPs if they all tag the same causal variant. However, due to the stochastic

nature of LD it is possible that two SNPs not in LD with each other may each be in partial LD with the causative QTN. Using the strength of signal of numerous secondary SNPs may therefore help to refine the likely location of a QTN. Accounting for significant SNPs may also detect secondary, independent associations at the locus, if any exist.

Also of interest is whether the same SNPs are associated with other medical conditions or diseases that may be connected in some way to urate metabolism. For example, uric acid may have a role to play in the epidemiology of another more complex disorder, therefore SNPs explaining variation in uric acid would also have a diluted effect on this other trait. One obvious possibility in the case of uric acid is gout, however another possibility is the metabolic syndrome. Evidence for association to more complex phenotypes would be harder to obtain however, since the effect of these SNPs is small even on the uric acid scale, and could conceivably be orders of magnitude smaller when considering more complex diseases, due to the number of other risk factors likely to be contributing.

3.1.5 Replication studies

Any positive GWAS results must be replicated in order to ensure that they are not spurious. Without successful replication it is impossible to know whether significant results are true or false positives, and this impacts upon the credibility of the results. It may be the case, for example, that interesting results pertain to effects unique to that particular study population, and are absent or greatly diminished in other

populations. In this regard, one of the great strengths of population isolates becomes a potential weakness, since alleles that have risen in frequency in an isolate may be at lower frequency and hence harder to replicate in other populations. However, since a replication study is not burdened a by stringent multiple testing correction, the relaxed significance threshold generally facilitates successful replication of true positive results.

The EUROSPAN collaboration, of which this project was part, has five GWA studies running in tandem, and one of these was used as a population in which to replicate the uric acid findings. These population data were collected on the island of Orkney, off the north coast of Scotland, and are therefore also from an isolated population. In addition to this, a German population collected as part of a case / control study of FEUA was used as another replication cohort.

3.2 MATERIALS AND METHODS

3.2.1 GRAMMAR step three

GRAMMAR step three (i.e., the measured genotype approach) was used for the follow-up analyses. The exact model for analysis at this stage is shown below - see chapter two sections 2.2.6.1 - 2.2.6.4 for description of terms.

$$y_i = \mu + kg_i + \Sigma_j \beta_j c_{ji} + G_i + e_i$$

Covariates and fixed effects included in the model remain unchanged from those used in step one of GRAMMAR, which were age, sex, BMI and an age-by-sex interaction. In total, 26 SNPs in *SLC2A9* were analysed using this model. Of these 26 SNPs, five were the SNPs exceeding Bonferroni significance from the original analyses, and two were SNPs that were located within *SLC2A9* but fell marginally short of the this significance level. These seven SNPs are located on a stretch of DNA encompassing 17 SNPs on the Illumina platform, therefore the remaining SNPs followed up were the intervening SNPs plus some extending past the outermost of the original seven in either direction. A likelihood ratio test (LRT) was used for each SNP to test whether an additive or genotypic model fitted the data better (twice the difference in log likelihood is chi-square distributed with d.f. equal to the difference in parameters estimated). All SNPs were found to act in an additive manner, as in the initial analyses, therefore the SNPs were parameterised in the model as linear covariates.

3.2.2 The *SLC2A9* gene and LD patterns

HapMap data were used to obtain the level of LD surrounding *SLC2A9* in the European (CEU) population. These data were also used in conjunction with the HaploView software (Barrett et al., 2005) to visualise regions of LD within and around the gene. HaploView allows the choice of either r^2 or D' to present the LD structure, since they are indictors of slightly different things. R^2 is a correlation measure that takes into account factors such as the age of mutation and the frequency of mutation and recombination, whereas D' is primarily a measure of whether

recombination has occurred since it is only ever less than one if all four haplotypes between two (biallelic) markers are present (Slatkin, 2008). Both measures range between zero and one in HaploView, as only the absolute magnitude of D' is given.

3.2.3 Using multiple markers

The use of multiple markers should elucidate a clearer picture of the genetic architecture of the effect on uric acid. It is possible that the effect is caused by a single QTL, and this is the hypothesis that has been tested so far in the analyses. However, it is also possible that the effect is a consequence of more than one QTL, and is instead due to small cumulative effects from multiple causal loci in the *SLC2A9* gene, either all found on one "high risk" haplotype, or on numerous genetic backgrounds. One way to test whether there is an effect over and above that of a single SNP is to fit multiple SNPs in a multiple regression model, thus determining whether any variation exists in uric acid that is not fully explained by the first SNP in the model. Using this approach could also help to fine-map the location of a QTL, as previously explained.

Multiple regression was performed for numerous combinations of the five SNPs exceeding Bonferroni significance in the original analyses, in order to determine whether any effects near the top SNPs were due to a secondary, independent, mutation at the locus, and to help fine-map the location of any QTN. Should any secondary QTN with low r^2 to the hit SNP exist, these independent effects would be detected (assuming the presence of tagging SNPs and requisite power to do so)

regardless of whether the top SNP was adjusted for, since the top SNP does not correlate with these other markers. However, after adjusting nearby SNPs for the variance explained by the hit SNP, any strong signals remaining may indicate the presence of a secondary QTN, or help to better define the location of a single QTN. Using only five SNPs also helped to control the number of tests to some degree. For example, if all seven highly significant SNPs were considered, there are 42 possible ways in which any two of the seven SNPs could be fitted (accounting for both orientations of the two SNPs in order to test each SNP in the presence of the other), therefore allowing greater numbers of SNPs in the model would result in an extremely large number of tests.

3.2.4 Alterations to the model

Due to the known link between uric acid and gout, individuals in the dataset suffering from this disease were identified from the records collected initially. This is because individuals suffering from gout are likely to be on medication that would include urate-lowering drugs. In addition to treatment for gout, other types of medication also have an effect on uric acid concentration, including those taken for cardiovascular disease for example, therefore individuals using these were also identified. The major uric acid lowering drug found to be taken by individuals in the study was Allopurinol, which is used as a treatment for gout.

Including individuals with environmentally (i.e., drug) altered uric acid levels in the analysis would not be optimal, since no term in the model accounted for this. This

would affect the analysis by reducing power and hence making any association harder to find, since those who would have high uric acid levels due to genetics instead have a much lower trait value. It is possible to fit a fixed effect for treatment / non-treatment, but this would not capture the situation fully since there is no way to know whether there was a different reaction to the drugs among different people (i.e., gene-by-drug interactions), or any dose-dependent effects, and this method would also class all uric acid lowering drugs as the same. Assigning each drug a different level of fixed effect is an alternative, although this solution still fails to account for any gene-by-drug interactions or dose effects. The only way in which this information can be modelled is with time-series data that record of the change in uric acid level of individuals pre and post drug taking.

In total 174 of the original 1,031 individuals with phenotypic information were removed from the analyses due to using medication known to affect uric acid levels, since no more preferable way of dealing with them could be found. Most of these individuals had also been excluded on the basis of the quality control described in the previous chapter however, therefore only 20 additional individuals were removed. After removing the additional individuals that were on urate-altering medication there was a total of 946 individuals (400 males and 546 females) to perform the analyses on, although the exact number varied between SNPs due to missing genotypes.

One interesting thing that came to light in the follow up analyses was that there were drastically different effect sizes estimated for the top SNPs for men and women.

Consequently, the model was changed to include a sex-by-SNP interaction term for all follow-up SNPs.

3.2.5 *SLC2A9* SNPs and alternative traits

The SNPs associated with uric acid were tested for association with other traits potentially linked to it. Some traits analysed at this stage were not quantitative, and others that were recorded as quantitative had well-defined thresholds in medical literature with which to dichotomise them, therefore these traits were analysed as both quantitative and binary phenotypes. For example, diastolic blood pressure is dichotomised at 90mmHg, and systolic blood pressure is dichotomised at 140mmHg in the medical literature. Binary traits included gout and the metabolic syndrome.

For traits that were binary, the method of analysis used was Fisher's Exact Test, or where numbers were too large, a 2x2 contingency table (assuming additivity) and chi-square approximation. Only those individuals who are "unrelated", namely founders and singletons (individuals for whom no information exists to connect them to other members of the dataset) were used. This was to prevent potential inflation of the test statistic due to phenotypic correlations between family members that may be present due to polygenes or shared environmental factors. The quantitative phenotypes were analysed in the same way as for the rest of the quantitative traits (QTs) during the GWAS, i.e., tests were performed using step three of GRAMMAR on the raw trait data, including a random effect accounting for polygenic variation in the model.

Details of traits analysed, transformations used and covariates / fixed effects fitted can be found in Table 3.1.

TRAIT	TRANSFORMATION	COVARIATES / FIXED EFFECTS
Gout (Binary)	N/A	None
Metabolic Syndrome – IDF (Binary)	N/A	None
Metabolic Syndrome – ATP (Binary)	N/A	None
Creatinine (QT)	None	None
Diastolic BP (Binary)	N/A	None
Diastolic BP (QT)	Square Root	Sex BMI
Systolic BP (Binary)	N/A	None
Systolic BP (QT)	Natural Logarithm	Sex*Age
Brachial – left (QT)	Natural Logarithm	Sex*Age LogBMI
Brachial – right (QT)	Natural Logarithm	Sex*Age LogBMI
Dorsal – left (QT)	Natural Logarithm	Sex*Age LogBMI Smoke
Dorsal – right (QT)	Natural Logarithm	Sex*Age LogBMI Smoke
Tibial – left (QT)	Natural Logarithm	Sex*Age LogBMI Smoke
Tibial – right (QT)	Natural Logarithm	Sex*Age LogBMI Smoke

Table 3.1 Additional traits analysed for association to the *SLC2A9* SNPs. Transformations used, and all covariates and fixed effects are also given.

3.2.6 Replication studies

The Orkney population used as a replication cohort was also collected as part of the large EUROSPAN collaboration to which the CROAS project belongs (see chapter

two, section 2.1.2). Orkney Complex Disease Study (ORCADES) is an ongoing family-based, cross-sectional study with data collected from the isolated Scottish archipelago of Orkney, off the north coast of Scotland, and was supported by the Scottish Executive Health Department and the Royal Society. Genetic diversity in this population is decreased compared to Mainland Scotland, consistent with the historically high levels of endogamy. Data for participants aged 18-100 years, from a subgroup of ten islands, were taken for this study. Fasting blood samples were collected and over 200 health-related phenotypes and environmental exposures were measured in each individual, including uric acid. In total the study genotyped 758 individuals and phenotyped 1,048 individuals, however after genotypic quality control (individual call rate threshold of 98%), there were 719 individuals with both genotypic and phenotypic information recorded. All participants gave informed consent and the study was approved by Research Ethics Committees in Orkney and Aberdeen.

The model used for the Orkney population was the same as that for the Croatia population. Individuals taking medication affecting uric acid were similarly removed, and the sex-by-SNP interaction was included. Six of the top seven SNPs from the original study had been genotyped in the ORCADES project (rs733175 was not genotyped in this population), therefore these were selected as replication SNPs. In total there were 719 individuals in the analyses; 317 males and 402 females.

One further population used as a replication was a German cohort initially sampled for use in research into FEUA. FEUA is the main risk factor for hyperuricemia,

therefore this dataset should be well suited for positively replicating the *SLC2A9* SNPs. Additionally, since 90% of subjects with gout have below normal FEUA (Graessler et al., 2006), this phenotype may be one step closer to gout than uric acid is, and replicating the association may more strongly implicate these SNPs in affecting gout. The dataset consists of 349 German subjects with low FEUA (≤6.6%), and 255 controls matched for ethnicity with normal FEUA (≥7.4%), therefore the trait is binary, unlike uric acid. These data were analysed for the same six SNPs as were replicated in the ORCADES dataset, using logistic regression with age, sex and BMI as covariates and fixed effects.

3.3 RESULTS

3.3.1 Measured genotype method results

Results from the full model can be found in Table 3.2, which also shows the minor allele frequency, effect size and direction of effect of the SNPs. Effect estimates given in the table indicate the effect of substituting a minor allele for a major allele. There is clearly evidence for a QTL affecting uric acid at this locus. The ordering of SNPs in terms of significance changes slightly from the reduced model (i.e., step two of GRAMMAR), but these changes are minimal and all five SNPs originally exceeding genome-wide significance are still highly significant. The SNPs are slightly more significant than they were in the reduced model; for example the p-value for the most significant SNP, rs727367, decreases from 1.22×10^{-9} to 1.12×10^{-10}. This effect is expected as a result of the "factoring out" effect mentioned earlier,

whereby some of the SNP effect is lost in the original analysis due to modelling of the polygenic effects. Many SNPs failing to reach genome-wide significance still show moderate association; of the 19 extra SNPs tested, six of the non genome-wide significant SNPs reach significance at the 0.1% level, and four more do at the 5% level.

SNP	MAF	EFFECT (S.E)	P-VALUE
rs4697693	0.319	0.21 (0.115)	0.067
rs2280204	0.115	0.21 (0.170)	0.197
rs2280205	0.488	0.23 (0.105)	0.026
rs4697695	0.326	-0.41 (0.112)	2.73×10^{-4}
rs10805346	0.473	0.47 (0.107)	1.03×10^{-5}
rs3733591	0.119	-0.39 (0.162)	0.016
****rs13129697****	**0.369**	**0.64 (0.104)**	**6.02×10^{-10}**
rs881971	0.453	-0.44 (0.105)	2.34×10^{-5}
****rs737267****	**0.297**	**0.70 (0.109)**	**1.12×10^{-10}**
rs4447863	0.435	-0.49 (0.102)	1.00×10^{-6}
rs12498956	0.490	0.44 (0.104)	2.6×10^{-5}
rs4505821	0.203	-0.17 (0.133)	0.194
****rs13131257****	**0.269**	**0.66 (0.113)**	**3.86×10^{-9}**
rs6849736	0.069	0.32 (0.225)	0.153
rs4502681	0.060	0.30 (0.238)	0.201
****rs6449213****	**0.246**	**0.69 (0.114)**	**1.54×10^{-9}**
****rs1014290****	**0.316**	**0.64 (0.107)**	**1.97×10^{-9}**
rs6845554	0.468	0.36 (0.106)	7.79×10^{-4}
rs6827754	0.466	0.35 (0.106)	8.38×10^{-4}
rs6820230	0.254	-0.15 (0.125)	0.238
rs10939663	0.277	-0.37 (0.118)	1.58×10^{-3}
rs9291645	0.238	-0.20 (0.127)	0.112
****rs733175****	**0.285**	**0.61 (0.110)**	**4.02×10^{-8}**
rs7683832	0.081	0.14 (0.209)	0.490
rs2241469	0.196	-0.13 (0.134)	0.322
rs10516200	0.340	-0.27 (0.112)	0.017

Table 3.2 Results for each of the 26 follow up SNPs in *SLC2A9*. SNPs are ordered as they would be found moving along the chromosome in a 3' to 5' direction. Minor allele frequency, effect size and p-value are given for each SNP. Analysis was performed using step three of GRAMMAR. P-values are from two-sided T-tests, and effects are on square root of uric acid scale. SNPs exceeding suggestive significance (1×10^{-6}) from step two of GRAMMAR are shown in bold, and are marked by asterisks.

In general, the minor allele is associated with a decrease in uric acid levels, however for one of the SNPs, rs4447863, the minor allele associated is with an increase in uric acid levels, opposite to the direction of the effect for all other SNPs. The minor allele frequency for this SNP is slightly higher than that of the rest. The average minor allele frequency of the main seven follow-up SNPs is 0.316, and that of the remaining SNPs is 0.287. A t-test reveals that there is no significant difference in minor allele frequency (MAF) of the seven top SNPs compared to the remaining 19 ($p = 0.491$), although with such a small number of SNPs a significant difference would be hard to detect. A significant difference may have indicated that SNPs at a higher MAF better tag the QTN in general, suggesting the QTN may also have higher MAF.

The MAF and effect sizes of the SNPs that were followed up are displayed graphically in Figure 3.1. Initially, it does not appear that there is a narrow minor allele frequency band unique to the significant hits - while the six SNPs with effects in the same direction all have reasonably similar MAF, other nearby SNPs also with similar MAF do not reach genome-wide significance. For example, five of the top seven hits have a MAF of between 0.269 and 0.316, however the three SNPs just 3' of rs733175 have MAF within or only slightly below this range, and do not have particularly significant results. However, for these three SNPs the effect is in the opposite direction, meaning the allele associated with decreasing uric acid has frequency one minus the MAF. Therefore in general, it would seem that alleles associated with decreasing uric acid and that have a frequency of approximately 0.3 are most highly associated. Interestingly, for a stretch of 11 consecutive SNPs starting at the 3' end, including four genome-wide significant SNPs, the effects of

neighbouring SNPs are in opposing directions, i.e., the effect direction of the minor allele for each consecutive SNP changes. It is around this stretch of 11 SNPs that three of the top four significant hits can be found, and many intervening SNPs also have very low p-values without quite reaching genome-wide significance.

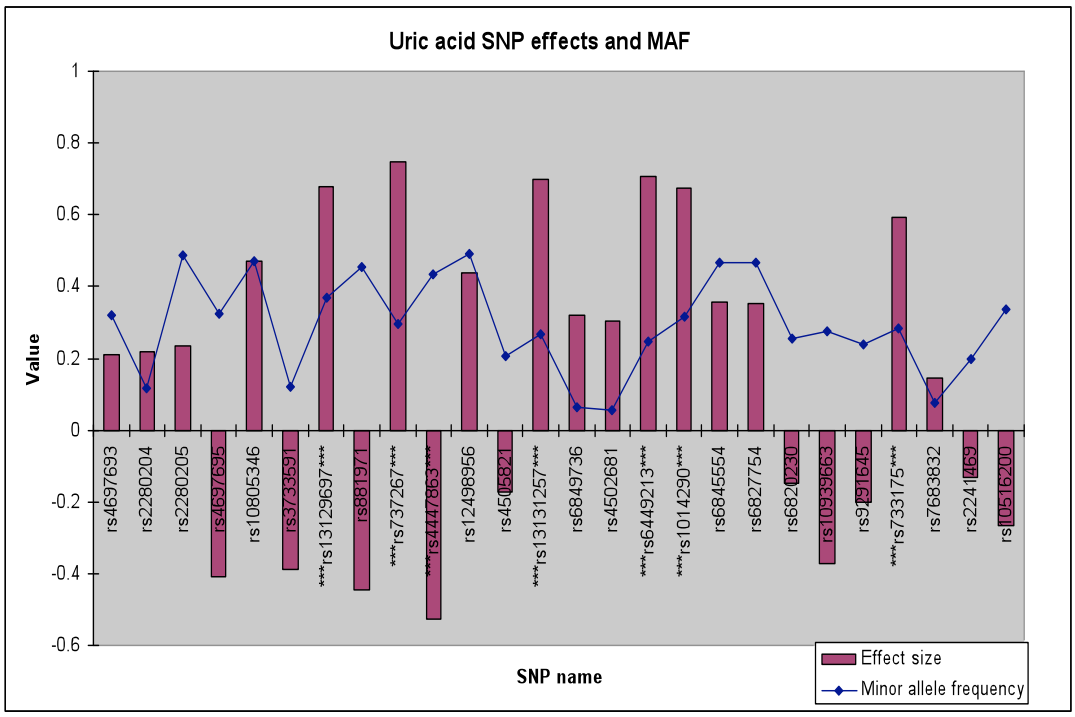


Figure 3.1 Effect size (in square root uric acid in mmol/L) and minor allele frequency of significant and non-significant SNPs along the *SLC2A9* gene on chromosome 4. Effect size is for substitution of one minor allele for one major allele. Asterisks indicate SNPs exceeding 1×10^{-6} in the follow-up analysis using step three of GRAMMAR.

3.3.2 *SLC2A9* LD plots

Figure 3.2 shows the patterns of pairwise linkage disequilibrium (LD) between SNPs located in and around the *SLC2A9* gene, taken from the CEU HapMap data using HaploView. Bright red colouration represents a D' of one, meaning that no recombination has been observed between the two markers. Above the LD plot in Figure 3.2 is a display of the locations of all typed SNPs on our panel, and above that is shown the short and long isoforms of *SLC2A9* from the Ensembl genome browser

(Hubbard et al., 2007). There are 13 and 12 exons respectively in the long and short forms of the gene, the difference between the two being an alternative transcription start site.

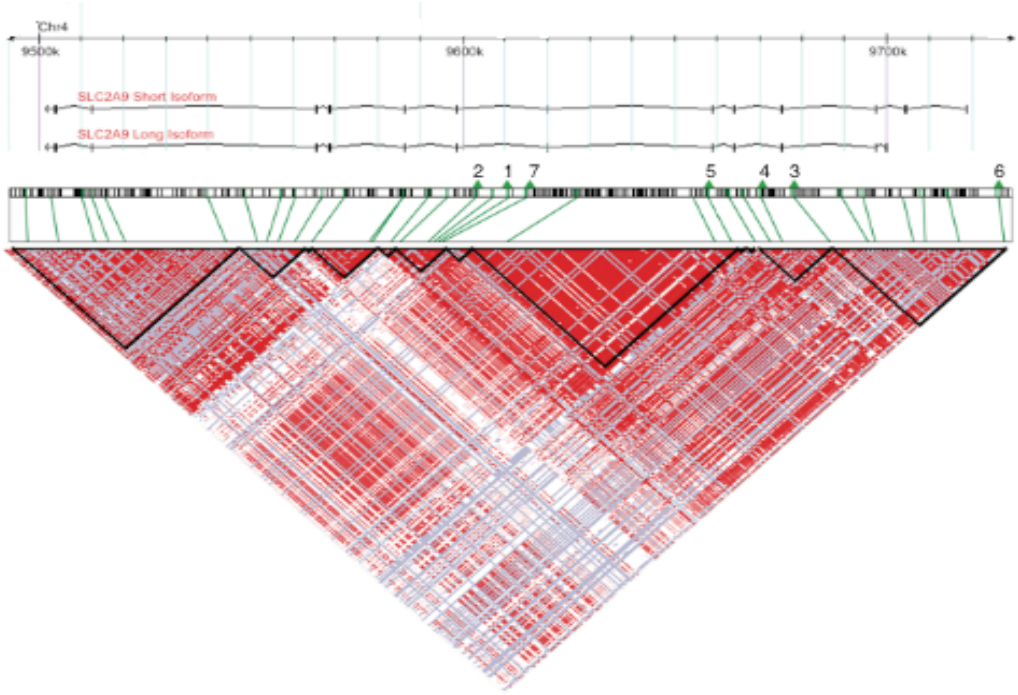


Figure 3.2 LD blocks in the *SLC2A9* gene. Red indicates high D', white indicates D' of zero. All typed SNPs are displayed immediately above the LD plot, and the main seven SNPs followed up are labelled in order of significance. SNPs are labelled as follows; (1) rs737267, (2) rs13129697, (3) rs1014290, (4) rs6449213, (5) rs13131257, (6) rs733175, (7) rs4447863. The two transcripts of the *SLC2A9* gene are shown above the SNP locations.

All significant SNPs fall within introns 3-7 of the long isoform (2-6 of the short isoform), with the exception of rs733175 that is in the 5' UTR. The seven most significant SNPs are labelled numerically on the graph in order of significance, and SNP names are provided in the figure legend. The two most significant SNPs are found within the same LD block, and a highly significant third SNP (rs4447863) is immediately 5' of it. This is shown in close up in Figure 3.3, where coloration represents D' (reddest being D' = 1) and the number represents r^2. The r^2 value for the CEU HapMap population is 0.7, where in the Croatian dataset the value is 0.73.

Similarly, r^2 between rs737267 and rs4447863 is 0.34 in HapMap and 0.36 in the Croatian dataset, indicating that r^2 is consistent between these populations for this region.

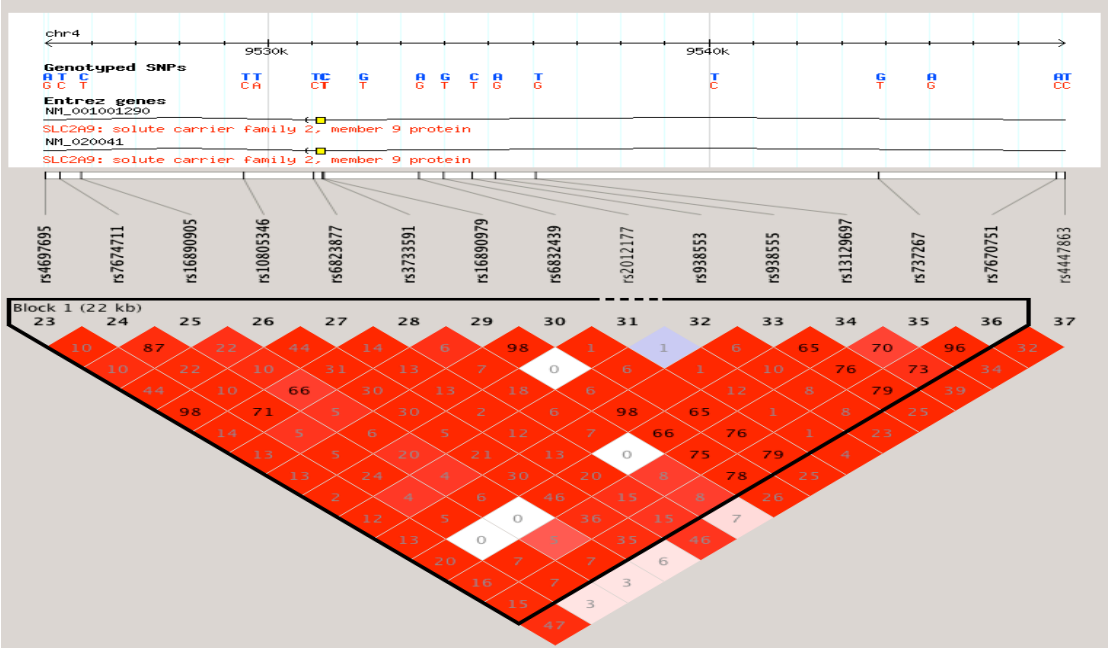


Figure 3.3 Plot from HaploView showing LD in D' (colour) and r^2 (number) for the LD block containing the most significant SNP. Two other of the top seven SNPs are also shown in the plot.

Figure 3.4 shows the associated region and -$\log_{10}$ p-values for all typed markers using the SNP Annotation and Proxy Search tool (Johnson et al., 2008). This also uses CEU HapMap data to display pairwise r^2 between all SNPs all the top SNP, represented by the shade of red. This figure shows that the top seven SNPs are all in moderate to high LD, but that the SNPs flanking these seven show reduced association with uric acid. Overlaid on the plot is the recombination rate in this region of the genome, which clearly indicates that to either side of the main association (at around 9.3Mb and 9.8Mb into chromosome 4) there is an increase in recombination. No SNPs reach a -$\log_{10}$ p-value of three past these recombination hot-spots, which strongly suggests that the causative variant is located somewhere within these bounds. Two other genes

flanking *SLC2A9* and within the boundary defined by the recombination hot-spots are *WDR1* (WD repeat-containing protein 1) in the 5' direction, and *DRD5* (dopamine receptor D5) past the 3' end. SNPs in both these genes show weak association with uric acid.

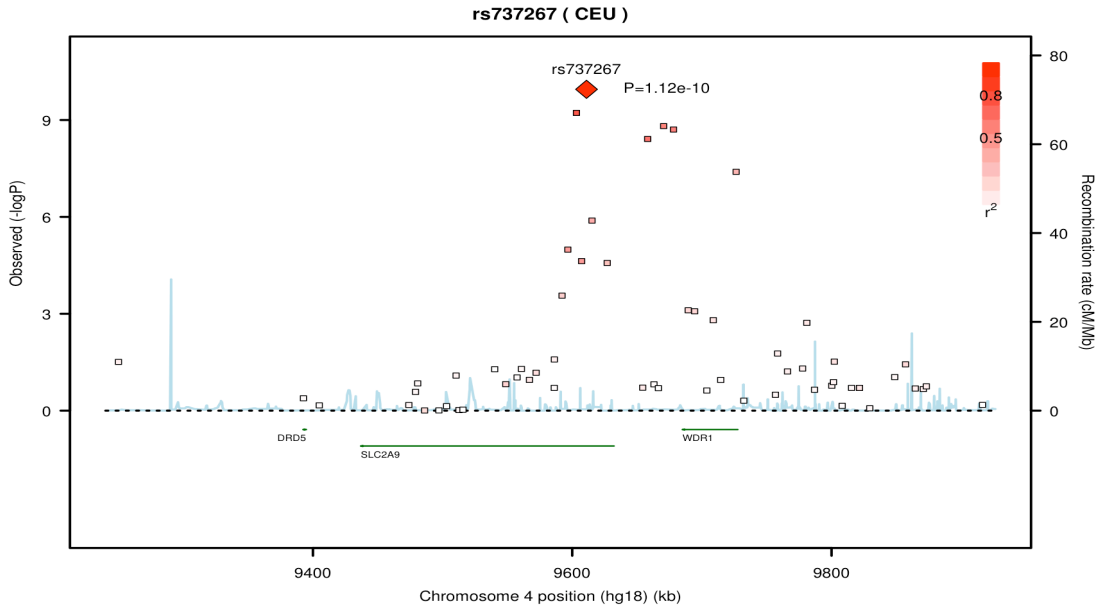


Figure 3.4 Plot showing the association within the *SLC2A9* gene using the SNP Annotation and Proxy Search tool. All typed SNPs within this region shown. Colour of each SNP indicates r^2 with the top SNP rs737267. Overlaid on the graph is an estimate of the recombination rate. The location of *SLC2A9* is shown below the graph.

In order to determine how the strength of association changed across the follow-up SNPs relative to the top hit rs737267, $-\log_{10}$ p-values were plotted against r^2 to rs737267 (calculated for this dataset). This is shown in Figure 3.5. The figure shows that there is a clear increase in the strength of association as r^2 to rs737267 increases, and the correlation between r^2 and $-\log_{10}$ p-value was 0.93. This suggests that there is a single effect being detected at this locus, since none of the other p-values are particularly larger than expected, given their r^2 to the top SNP. This evidence therefore implies the presence of a single QTN, and that rs737267 is the SNP that best captures the variation explained by this QTN.

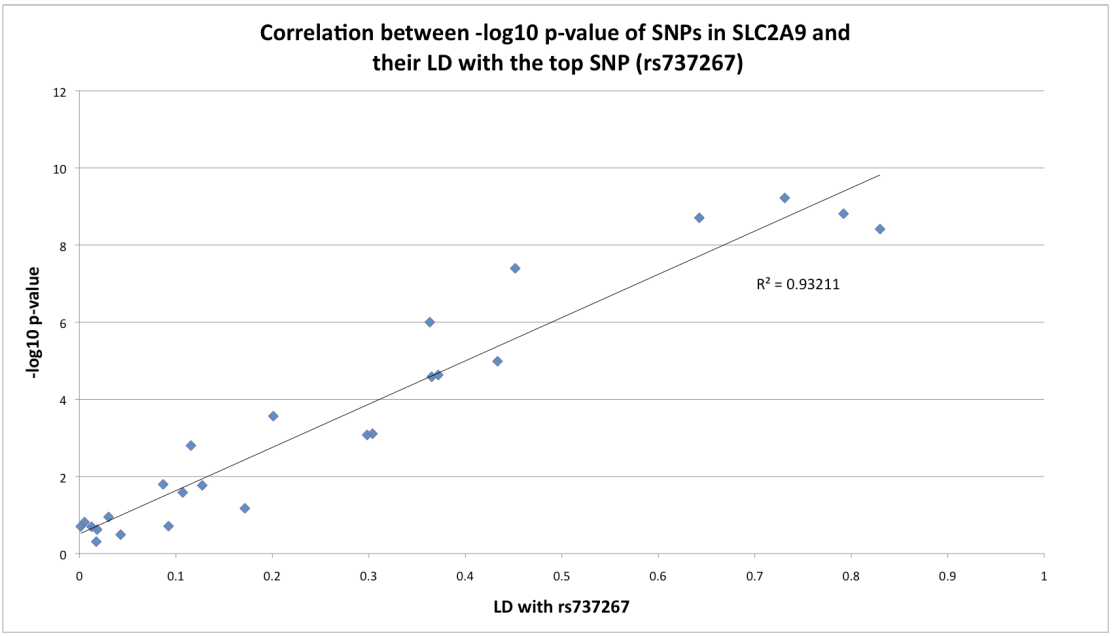


Figure 3.5 Correlation between the -$\log_{10}$ p-value of SNPs within *SLC2A9* and their LD with the top SNP rs737267.

3.3.3 Multiple marker results

Table 3.3 shows the results from fitting two of the top SNPs in a multiple regression framework. For each two-SNP model, both orientations of SNPs were tested, because after correcting for the effect of one SNP, secondary, independent associations may be identified for the other. Orientation 1 shows results where the more significant of the two SNPs from the original analyses was fitted first in the model. Eight of the 10 tests in orientation 1 have non-significant p-values for the second SNP, and the other two have p-values significant at the 5% level. The p-values for the first SNP are not identical to those for that SNP when analysed alone since slight differences are produced due to differing patterns of missing data when combining data from more than one SNP. P-values for the second SNP are largest (i.e., least significant) when the top SNP, rs737267, is fitted first in the model.

SNPs FITTED IN MODEL		ORIENTATION 1		ORIENTATION 2	
		SNP1	SNP2	SNP1	SNP2
rs737267	rs13129697	$1.14e^{-10}$	0.174	$4.01e^{-10}$	0.038
rs737267	rs6449213	$2.00e^{-10}$	0.313	$2.83e^{-9}$	0.013
rs737267	rs1014290	$1.10e^{-10}$	0.137	$3.81e^{-9}$	0.003
rs737267	rs13131257	$1.15e^{-10}$	0.777	$3.58e^{-9}$	0.009
rs13129697	rs6449213	$1.05e^{-9}$	0.031	$2.84e^{-9}$	0.010
rs13129697	rs1014290	$6.22e^{-10}$	0.271	$3.42e^{-9}$	0.033
rs13129697	rs13131257	$5.97e^{-10}$	0.099	$5.26e^{-9}$	0.008
rs6449213	rs1014290	$1.53e^{-9}$	0.145	$3.62e^{-9}$	0.052
rs6449213	rs13131257	$1.57e^{-9}$	0.325	$5.57e^{-9}$	0.064
rs1014290	rs13131257	$1.81e^{-9}$	0.037	$3.49e^{-9}$	0.018

Table 3.3 Results from fitting two markers in a multiple regression. The two SNPs fitted in each case are given in the table, and results are shown for both orders the SNPs could be arranged in for that pair. Orientation 1 refers to the order given in the first column of the table, and orientation 2 gives the reverse order. P-values are from an F-test with 1 and effectively infinite degrees of freedom.

For orientation 2, where the order of the SNPs has been reversed and the more significant SNP is fitted second, in eight of the 10 tests the second SNP is significant at the 5% level at least. The top SNP, rs737267, is always at least significant at the 5% level when fitted second, and for all tests where the SNPs are arranged in orientation 2, the second SNP is more significant than when those SNPs were in orientation 1. This suggests that all SNPs are tagging the same variant, but that one is doing so more effectively (i.e., rs737267). Since there was little suggestion of multiple QTN from these results, only two-SNP combinations were considered.

3.3.4 Final model results

Table 3.4 shows results from using the final model with which the seven follow-up SNPs were analysed. The sex-by-SNP interaction was significant at the 5% level (at

least) for the five most significant SNPs. For rs733175 and rs4447863 the interaction term is not significantly different from zero. A 5% significance level is appropriate to test for the interaction because it is a specific hypothesis test performed on seven SNPs only. The effect sizes for substituting a minor allele for a major allele are also shown in Table 3.4. Effect sizes for all SNPs are much larger for females than males, and the p-values are also far more significant. The largest effect size estimate for the females was 0.962, for rs737267. Interestingly, the effects in females are so striking that the SNPs would not have been found significant in a GWA scan if only the males had been used – for three SNPs the p-value doesn't exceed 0.05, and even the most significant only reaches 0.014. In contrast, the most significant of the p-values for the women was 9.76×10^{-12}. For all SNPs, results from the females alone produces p-values more significant than with the sexes combined.

SNP	N	EFFECT (S.E)	P-VALUE	SNP*SEX INTERACTION
rs737267	390	0.358 (0.162)	0.027	0.004
	521	0.962 (0.141)	$9.76e^{-12}$	
rs13129697	399	0.314 (0.154)	0.041	0.004
	542	0.880 (0.133)	$3.84e^{-11}$	
rs6449213	394	0.332 (0.174)	0.056	0.007
	545	0.938 (0.146)	$1.46e^{-10}$	
rs1014290	400	0.360 (0.162)	0.026	0.020
	546	0.844 (0.137)	$8.25e^{-10}$	
rs13131257	400	0.317 (0.166)	0.056	0.005
	546	0.930 (0.147)	$2.28e^{-10}$	
rs733175	400	0.409 (0.166)	0.014	0.110
	544	0.751 (0.143)	$1.36e^{-7}$	
rs4447863	399	-0.303 (0.154)	0.050	0.101
	539	-0.625 (0.130)	$1.51e^{-6}$	

Table 3.4 Results showing individual sex effects, p-values and the sex-by-SNP interaction p-values for the SLC2A9 SNPs analysed for uric acid in the Croatian population. The top line for each SNP corresponds to the males and the bottom one to the females. Effects are for substitution of one minor allele for one major allele.

3.3.5 Alternative traits results

Results for each SNP analysed for the alternate traits are in Table 3.5. With 14 traits and seven tests for each trait, if independence over all SNPs and traits was assumed then the appropriate significance level derived through a simple Bonferroni correction would be approximately 5×10^{-4}. There are no p-values that reach this level of significance, therefore it would appear that there is no association between these SNPs and any of the uric acid related traits. This is notable in the case of gout, given the known links between uric acid and the disease.

TRAIT	rs737267	rs13129697	rs6449213	rs1014290	rs13131257	rs733175	rs4447863
Gout	0.176	0.122	0.256	0.292	0.166	0.320	0.663
MS (IDF)	0.806	0.146	0.301	0.066	0.215	0.237	0.560
MS (ATP)	**0.018**	**0.009**	**0.039**	**0.012**	**0.019**	**0.020**	0.265
Creatinine	0.497	0.688	0.597	0.934	0.566	0.649	0.904
Diastolic BP (binary)	0.671	0.348	0.791	0.632	0.357	0.920	0.493
Diastolic BP (QT)	0.694	0.300	0.854	0.536	0.270	0.355	0.666
Systolic BP (binary)	1.000	0.888	0.420	0.549	0.354	0.254	0.597
Systolic BP (QT)	0.867	0.712	0.968	0.888	0.767	0.778	0.879
Brachial (left)	0.308	0.075	0.418	0.220	0.249	0.146	0.955
Brachial (right)	0.483	0.186	0.409	0.275	0.301	0.097	0.670
Dorsal (left)	0.111	0.052	0.230	0.291	0.053	0.919	0.367
Dorsal (right)	0.094	0.064	0.171	0.239	**0.035**	0.240	0.710
Tibial (left)	0.154	0.073	0.285	0.271	**0.048**	0.504	0.373
Tibial (right)	0.151	**0.047**	0.236	0.228	0.063	0.263	0.480

Table 3.5 Table showing p-values for each SNP tested for each of the extra traits. P-values significant at the 5% level are shown in bold.

Interestingly however, six SNPs reach the 5% significance level when analysed for the ATP measure of the Metabolic Syndrome. At a 5% significance level, the expectation for 98 tests would be that five were significant purely by chance, however it is unlikely that these would be concentrated on the same trait. The fact that there is an enrichment of SNPs reaching 5% significance within this trait (six of the nine p-values that are below 0.05 are for ATP Metabolic Syndrome) may be suggestive of a true association there.

3.3.6 Replication studies results

Results from analysis of the *SLC2A9* SNPs in the Orkney population are shown in Tables 3.6 and 3.7. Table 3.6 shows the overall test for association, without the sex-by-SNP interaction term fitted in the model, and Table 3.7 shows the results for when this interaction is included.

SNP	EFFECT (S.E)	P-VALUE
rs737267	0.175 (0.050)	$5.00e^{-4}$
rs13129697	0.193 (0.045)	$1.72e^{-5}$
rs6449213	0.211 (0.048)	$1.33e^{-5}$
rs1014290	0.204 (0.052)	$7.74e^{-5}$
rs13131257	0.262 (0.057)	$4.50e^{-6}$
rs4447863	0.155 (0.041)	$1.60e^{-4}$

Table 3.6 Table showing effect sizes and p-values for the SLC2A9 Uric Acid hit SNPs in the Orkney population. Effects are for substituting one minor allele for a major one.

These tests positively replicate the results found for uric acid in the Croatia population, as all SNPs are highly significant for the overall test, albeit in a different order of significance. The tables show the effects produced by substitution of a minor

allele for a major allele, therefore in all cases except that of rs4447863, the effects of the SNPs in both populations are in the same direction (the major allele in both populations is associated with the higher uric acid level). With rs4447863, it is the minor allele in the Croatian dataset, but the major allele in the Orkney dataset that causes the increase in uric acid. This can be explained by the fact that the same nucleotide is the minor allele in one population, but the major allele in the other. This is a consequence of the allele frequency for this particular SNP being much closer to 0.5 than the other significant SNPs.

SNP	N	EFFECT	P-VALUE	SNP*SEX INTERACTION
rs737267	313	-0.006 (0.072)	0.936	5×10^{-4}
	390	0.324 (0.066)	8.22×10^{-7}	
rs13129697	315	0.064 (0.070)	0.358	0.016
	392	0.273 (0.056)	9.58×10^{-7}	
rs6449213	313	0.024 (0.070)	0.726	2×10^{-4}
	393	0.356 (0.062)	1.13×10^{-8}	
rs1014290	317	0.085 (0.076)	0.267	0.034
	397	0.292 (0.061)	1.0×10^{-5}	
rs13131257	315	0.050 (0.080)	0.529	2×10^{-4}
	395	0.454 (0.077)	2.86×10^{-9}	
rs4447863	313	0.056 (0.061)	0.352	0.030
	393	0.226 (0.053)	1.71×10^{-5}	

Table 3.7 Results showing individual sex effects, p-values and the sex-by-SNP interaction p-values for the *SLC2A9* SNPs analysed for uric acid in the Orkney population. The top line for each SNP corresponds to the males and the bottom one to the females. Effects are for substitution of one minor allele for one major allele.

The sex-by-SNP interaction in the Orkney population is significant at the 5% level (at least) for all SNPs replicated, so this term was included in the model. Three of the sex-by-SNP interactions are more significant than for the Croatian population, exceeding the 0.1% threshold. Similarly to the Croatian population, fitting the sex-by-SNP term for the Orkney population enhances the size of effect in women and

reduces it in men. However, for those SNPs where the interaction is significant in the Orkney population, the difference in p-value between males and females is far more extreme, and none of the p-values for association in the males is significant even at the 5% level. The effect sizes of the association in absolute terms for the women are much smaller than the effect sizes found in the Croatian population, the largest being 0.454 (0.077) for rs13131257 compared with 0.962 (0.141) for rs737267.

Results of the replication analysis in the German cohort of individuals with reduced FEUA are presented in Table 3.8. A sex-by-SNP interaction was not significant at the 5% level for any of the SNPs and was therefore omitted. Again there is positive replication of the SNPs implicated in the initial GWAS, as all SNPs are significant at the 5% significance level. Unlike for the Orkney replication, in this dataset the major allele of rs4447863 was found to decrease uric acid, therefore all SNPs have direction of effect in agreement with the initial findings.

SNP	EFFECT (S.E)	P-VALUE
rs737267	0.480 (0.158)	0.002
rs13129697	0.515 (0.154)	0.001
rs6449213	0.425 (0.170)	0.012
rs1014290	0.448 (0.160)	0.005
rs13131257	0.457 (0.163)	0.005
rs4447863	-0.347 (0.132)	0.009

Table 3.8 Effect sizes and p-values for the *SLC2A9* uric acid hit SNPs in the German cohort, collected as part of a study into fractional excretion of uric acid (FEUA). Effects are for substituting one minor allele for one major allele.

3.4 DISCUSSION

3.4.1 Overview

The full model results for uric acid clearly provide strong evidence for an association on chromosome 4, as a total of six SNPs from the full measured genotype approach exceed the Bonferroni threshold for genome-wide significance. The association is strongly supported by replication in two independent European populations, as six SNPs in both of these studies exceeded the 5% significance level at least. Direction of effect was in agreement for all SNPs with the German cohort, but differed for one SNP with the Orkney population. However, this SNP has a minor allele frequency approaching 0.5, therefore it is likely the difference represents the major allele in the Orkney study being the minor allele in the remaining two studies.

The causal variant (or variants) is probably located within or very close to the *SLC2A9* gene. Previous studies have identified the *SLC2A9* gene product as belonging to a solute carrier family of proteins, and it has been reported as acting as a glucose transporter (Phay et al., 2000). Levels of linkage disequilibrium within and around *SLC2A9* are consistent with the hypothesis that a causative variant may be tagged by one of the SNPs found significant in this study. One other interesting finding to report from this follow up is a significant sex-by-SNP interaction for the top uric acid hits, and considerably larger effect sizes in the women compared to in men.

3.4.2 Croatia uric acid follow-up

Results for the seven main uric acid SNPs followed up became slightly more significant once analysed under the full model. This is expected due to the partial factoring out of SNP effects by steps one and two of GRAMMAR. "Factoring out" refers to the loss of some of the variation pertaining to the SNPs tested in step two due to removing polygenic variation associated with the pedigree in step one. Genomic control can correct for this to some degree as it detects systematic inflation or deflation of the test statistic, however it has been suggested that the correction is conservative when used for this purpose. Given that all top hits experienced a decrease in p-value despite the application of genomic control after step two, it appears this is likely to be the case, although in absolute terms the changes are very small.

The fact that there were multiple significant hits (five at Bonferroni significance, and in total 17 of the 26 followed up were under the 5% level) is highly suggestive of a real effect. Analogous to the way in which causative QTN are detected by the presence of genotyped SNPs in strong LD with it, correlations between marker genotypes may result in similar test statistics among these markers. Since many of the supporting hits for uric acid are at SNPs that have an r^2 measure in the region of 0.7 with the most significant SNPs, it is likely that they also tag the same causal variant reasonably well, thus reaching suggestive significance. The level of significance these SNPs reach ultimately depends on their own level of LD with the causative variant, which explains why there is variation in the strength of significance of the intervening and surrounding SNPs not identified initially in step two.

Further support in favour of the association being real is that of the seven SNPs brought forward from the initial analysis, six produced an effect on uric acid going in the same direction (i.e., the minor allele decreases uric acid). If many of the top hits had effects of the same magnitude but opposing direction then interpretation of the results would be more difficult. This is because for two markers in high LD, it is highly probable that the minor alleles are in coupling, not repulsion (this becomes more likely as MAF decreases, since rare alleles in repulsion never produce high r^2), and should therefore both tag the same effect from the causative locus. If this is the case, it is therefore impossible for the two markers to disagree on the direction of effect. Given that the top SNPs associated with uric acid are in relatively high LD, with only moderate MAF (the mean of the top six hits is 0.297), it is to be expected that the effect directions are consistent.

Despite the top six hits having the same effect direction however, the direction of effect does fluctuate for some consecutive SNPs, as seen in Figure 3.1. It follows from the previous paragraph that for this to be the case, SNPs with one direction of effect should either have low r^2 with nearby markers of opposing direction estimates, or have a higher MAF. As can be seen using Figures 3.1 and 3.3, this is indeed the case. The most significant of the SNPs where the minor allele increases uric acid is rs4447863, which has a MAF of 0.435, and r^2 with the three nearby SNPs with opposite direction (rs10805346, rs13129697 and rs737267 respectively) of 0.06, 0.38 and 0.34. Similarly, for two other SNPs where the minor allele increases uric acid (rs4697695 and rs3733591) r^2 values are also low; rs4697695 has r^2 of 0.44, 0.20 and

0.16, and rs3733591 has r^2 of 0.31, 0 and 0.08 to the three SNPs mentioned above. This means that the changes in direction of effect are not inconsistent with what may be expected and therefore do not bring the validity of the association into doubt.

In addition to fluctuations in the direction of effect, the effect sizes also vary considerably between nearby SNPs. Again this is consistent with theory and is a consequence of changing allele frequencies at the typed markers, since estimated effect size is proportional to the level of r^2 between the QTN and marker, and r^2 itself is highly dependent upon both the frequency of each SNP and how close these frequencies are to one another. For example, the consecutive SNPs rs6820230, rs10939663 and rs9291645 all have negative effect estimates for the minor allele – i.e., the minor allele here is the uric acid increasing allele – and have a minor allele frequency of around 0.25 - 0.3 (Figure 3.1). Given that it appears the uric acid increasing allele has a much higher frequency than 0.3 (frequencies at the most significant SNPs suggest an allele frequency of ~0.7), r^2 would be low between these three markers and the QTN. Consequently, effect sizes at these loci are smaller in magnitude. Correspondingly, for SNPs where the major allele increases uric acid, there is a clear relationship between the size of the effect and the allele frequency. SNPs with either a higher or lower allele frequency than the top six SNPs generally have smaller effect sizes. For example, SNPs rs6849736, rs4502681, rs6845554 and rs6827754 all have smaller effect sizes, and have minor allele frequencies different to the top SNPs (Figure 3.1), just as predicted.

Considering the genetic structure surrounding the *SLC2A9* gene, the most likely location of the causative QTN is between the recombination hotspots flanking *SLC2A9* in both directions (Figure 3.4). The presence of the hotspots causes a reduction in the level of LD between SNPs on opposing sides of the hotspot. Given that all significant SNPs are found within the recombination hotspots it is most probable that this is where the true causative variant lies. More specifically, Figures 3.1 and 3.4, and Table 3.2 show that the strength of association tails off to either side of the main association (i.e., rs737367), suggesting that the causal mutation is most likely to be somewhere between the region just 5' of *SLC2A9* and intron seven of the *SLC2A9* gene.

3.4.3 Multiple markers

Results from the follow-up are suggestive of at least one QTL affecting uric acid within or near the *SLC2A9* gene. Discussion has so far assumed the existence of only a single QTN, however it is plausible that more than one QTL is present. One way in which this would be possible is if the QTL were all in high LD with each other, as then the association would be detecting the effect of all of them. At one extreme it may be the case that a specific haplotype increased uric acid level due to a collection of QTL, meaning that results shown here tag the entire haplotype rather than an individual QTN. It is not necessary that the multiple QTN be on the same haplotype however. One way to help tease out individual effects from multiple QTL is by using multiple regression, fitting SNPs already found significant in the model as further

covariates and then looking for the significant effect of another SNP, and this was performed in the follow up investigation.

In orientation 1, where the more significant of the two SNPs from the original analyses was fitted first in the model, there was little evidence to suggest that any additional SNPs contribute an independent effect to uric acid. Eight of the 10 tests have non-significant p-values for the second SNP, and the other two have p-values significant at the 5% level. It is not clear that these represent secondary loci however, as the level of significance is fairly low, and due to the nature of pairwise LD, the secondary-SNP associations are likely to be a consequence of LD between the second SNP and the causative variant that is left unexplained by LD between the first SNP and the causative variant. This is possible due to the complex nature of three-way LD, and is more probable where pairwise LD between the first and second SNPs is not so high, but the two pairwise LD measures between each SNP and the causative variant are both high. It is not possible to entirely rule out a second QTN, however a greater level of significance for the secondary SNP would be more convincing. In orientation 2, where the order of the SNPs is reversed, most tests have the second SNP significant to at least the 5% level. This conforms to expectation if only one causative SNP exists however, given the fact that the SNP now second in the model has a larger effect on uric acid than the SNP fitted first.

An additional use of the multi-SNP regression analysis was to help fine-map the location of a putative QTL (since in this case it appears there is only one at the locus). Results for secondary SNPs corrected for the top hit (rs737267) indicate that there is

very little of the effect on uric acid not explained by this SNP, as none of the p-values for the second SNPs in this scenario exceed 0.05. This is also the case for the second most significant SNP, which is located very close to rs737267 within intron seven of *SLC2A9*. The best interpretation of these results is that all significant SNPs are tagging the same causative variant, indicating that the causal QTN is in LD with all SNPs identified during the initial GWA scan, and rs737267 most of all. Consequently, close to intron seven is the most likely location of the QTL affecting uric acid.

While there is no solid evidence to imply that there is more than one QTL influencing uric acid in this region, there does remain the possibility that multiple QTN exist in almost perfect LD with one another (and hence with the same LD with our markers). Without complete sequence information, this is extremely difficult to detect however. Likewise, it is not straightforward to identify the location of the QTN should the association consist of a single causative mutation. Sequencing would be the most effective way of getting down to the true causal variant, but this process is both time consuming and expensive to perform.

One other potential use of multiple marker data that is not performed here is the construction of haplotypes to test for association. It would be interesting be see whether any high risk haplotypes exist in this population that better explain the variation seen at this locus than any of the individuals SNPs. A single QTL is likely to be found more frequently on one genetic background than others in regions of the genome exhibiting high LD, such as the uric acid associated region. This is especially

likely in a closed population such as the one analysed here, since there is no introduction of new genetic information through migration, and founder haplotypes can exist largely intact for many generations. Characterising any such haplotypes that exist in these data, and determining whether they would better tag the variation in uric acid than any of the individual SNPs is a large undertaking in itself however.

3.4.4 Sex-by-SNP interaction

The results in Tables 3.4 and 3.7 are conclusive in demonstrating that a sex-by-SNP interaction exists between the significant *SLC2A9* SNPs and uric acid in both the Croatian and Orkney populations. There have been very few reports about such interactions in the literature, therefore this finding has potentially important lessons concerning the way future GWA studies are approached. For example, had the study initially been performed using the Orkney population, it is possible the association with uric acid would have been missed entirely, as no SNPs reached genome-wide significance when males and females were analysed jointly. Only after including the interaction in the model does the association become clear in the females. It is possible that past studies have missed genuine associations as a consequence of sex specific effects that are diluted when the sexes are analysed jointly. However, it does not make sense to include an interaction term in the model to begin with, as it reduces power to detect effects in the male and female subsets by effectively halving population size. Analysing each SNP twice, once with and once without the interaction term, is also unfeasible since the number of tests would double, further

compounding the multiple testing burden. One possible solution is to take SNPs at suggestive significance and fit those in a model with a sex-by-SNP interaction.

This phenomenon of a sex specific effect is in itself very worthy of note, and is particularly interesting with regard to the known distribution of uric acid and gout among males and females. It is known that uric acid is higher in men (mean uric acid levels in the Croatia population for men and women were 362.74 and 273.22μmol/L respectively), as is the prevalence of gout. It is interesting to speculate therefore that the overall effect of *SLC2A9* on gout may be small, given that high uric acid is a risk factor for gout and both high uric acid and gout incidence is higher in males, yet *SLC2A9* has little or no effect on uric acid in males. This theory is supported by the fact that no association between gout and the *SLC2A9* SNPs was found in the follow-up analysis.

More generally, it would be interesting to see how many traits were controlled (either partially or entirely) by genes with a sex-limited effect. It is possible this phenomenon is a feature of conditions in which there is an unequal distribution among the sexes, given it is known that gout, of which uric acid is a good indicator, has a higher prevalence in males than females. If this were the case, then intermediate phenotypes that are risk factors for sex-biased genetic conditions could be tested with a sex-by-SNP interaction to elucidate which genes are affecting them. This may be particularly useful for candidate gene approaches, where the interaction could be tested without causing a huge multiple testing problem.

3.4.5 Alternative traits

The vast majority of the traits connected to uric acid proved to have no significant associations with the *SLC2A9* hit SNPs. This is not unexpected because the suggested links between uric acid and some of these traits are only considered tentative. Additionally, some of the extra traits were themselves a disease endpoint, and for these traits the effect sizes of the SNPs (assuming they do have an effect on the related trait) will be far smaller than any this study is powered to detect. This is because rs737267 explains only around 5% of the total variance in uric acid, and any effect on other related phenotypes or diseases is thus likely to constitute a smaller proportion of variance for these traits.

One instance where the relationship with disease is not so tenuous is with gout, which has been strongly linked to uric acid for many years (Seegmiller et al., 1963). Both the initial uric acid association, and the subsequent replication in an independent cohort analysing FEUA (a stronger predictor of risk to gout) indicate that it is likely *SLC2A9* also affects gout risk. The *SLC2A9* SNPs clearly show no evidence of an association to gout in the Croatian population however, as the most significant p-value was only 0.122 (Table 3.5). An association of this gene to gout was subsequently discovered in a study of 11,024 white individuals. A non-synonymous SNP within the gene, rs16890979, was associated with gout risk with a p-value of 7×10^{-14}, and an odds ratio of 0.59 per minor allele (Dehghan et al., 2008).

The one exception where there was some evidence in the Croatian population to suggest the *SLC2A9* SNPs were associated with one of the additional traits was with

the ATP measure of the Metabolic Syndrome. Here, six of the seven main *SLC2A9* SNPs were found associated at the 5% significance level. This would represent a very exciting finding if it were true, considering that the Metabolic Syndrome has been a difficult disease to characterise both genetically and epidemiologically, and even a concrete definition of the condition is hard to find. It must be noted however that none of the p-values exceeded the Bonferroni significance threshold, therefore the evidence for association is not strong in this case.

4. CHAPTER 4

4.1 INTRODUCTION

Since the conception of the genome-wide association study (GWAS), it has been the norm to analyse genetic data using each single SNP separately. This approach has clearly had some degree of success (Hirschhorn et al., 2002), and remains a robust way in which to interrogate the genome. However, recently there has been a greater focus on using SNP information in more complex ways in an attempt to increase power to detect QTL. The justification for this is that while analysing SNPs one at a time is currently the favoured method of analysis, it may not necessarily the best. Consequently, there is an increasing interest in using alternative methods to analyse genome-wide association (GWA) data which may have a better chance of detecting QTL, and these methods typically involve utilising more than one marker at a time. There are two broad categories that these methods fall into; those that use un-phased multi-locus genotypes, and those that use haplotypes.

4.1.1 Haplotypes

4.1.1.1 What are haplotypes?

A haplotype can be defined as an ordered set of multi-locus genotypes such that the arrangement of alleles on each chromosome is known. The arrangement of alleles on a given chromosome is called its phase. For a haploid organism therefore, it is trivial

to determine the phase and haplotype (simply a matter of genotyping), but for diploid organisms such as humans, phase can be ambiguous with only genotype data if more than a single locus is heterozygous. The haplotype pair belonging to each diploid individual is collectively known as their diplotype.

Haplotypes are the name given to specific combinations of alleles at a particular locus. While haplotypes may be present at equilibrium frequency (i.e., no more frequent than would be expected by chance, given the individual allele frequencies at markers comprising the haplotype), for nearby alleles it is more common that the alleles forming haplotypes display non-random association. This causes specific alleles to be found in combination more or less frequently than would be expected given their frequencies. As described earlier in this Thesis, non-random association of alleles is known as linkage disequilibrium (LD). Note that although LD between a pair of loci can exist over large distances on a chromosome, and that LD can even exist between loci on different chromosomes, this does not comprise a haplotype. Here we will consider haplotypes that are stretches of adjacent SNPs that collectively exhibit linkage disequilibrium. To better illustrate, consider three loci with genotypes A/C, C/T and A/G. There is a total of four ways these multi-locus genotypes could be arranged into haplotypes pairs, and a total of eight different haplotypes. The four combinations are:

A C A	A C G	A T A	A T G
C T G	C T A	C C G	C C A

Note that with these three heterozygous loci there are 2^2 possible diplotypes, and 2^3 possible haplotypes. In general, where *n* denotes the number of biallelic polymorphic loci, the number of possible diplotypes and haplotypes can be expressed as $2^{(n-1)}$ and

2^n respectively. Non-polymorphic loci have no effect on the number of possible haplotypes, since these will be consistent across all haplotypes in the population. The number of possible haplotypes in a population increases rapidly with the length of the sequence in question; by the time the number of polymorphic loci reaches 20, the number of possible haplotypes has grown to more than a million. In practice, although this number of haplotypes is possible, there are only a limited number that reach an appreciable frequency in a given population.

4.1.1.2 Why use haplotypes?

There are two distinct ways in which haplotypes may act to help detect QTL. The first of these is when a collection of SNP markers, through LD with causal variants, define a set of mutations that together create a "super-allele" that has a large effect on an observed phenotype (Schaid, 2004). For example, consider the situation below (which for simplicity makes the assumption that all individuals are homozygous at these SNPs):

Individual 1	A	G	C	Case
Individual 2	A	G	T	Control
Individual 3	A	G	C	Case
Individual 4	A	A	C	Control
Individual 5	A	G	C	Case
Individual 6	T	G	C	Control

By analysing these three SNPs one at a time, no association would be found with the trait (which in this case is a simple binary case/control phenotype). However, looking at all three SNPs simultaneously reveals that the cases all share the same AGC haplotype, which none of the controls have.

This sort of situation may arise, for example, due to mutations in protein coding DNA regions - protein function often depends upon how the protein is folded, which in turn depends on the specific amino acid sequence (Clark, 2004). Multiple specific mutations within exons coding for a protein may therefore act jointly to affect protein structure, but have no effect on structure individually. By using only single SNP analysis on GWA data, it is possible that this type of association will be missed. One example of exactly this situation in humans is a gene affecting intestinal lactase activity (Hollox et al., 2001), therefore there are clear biological reasons why it may be important to study the effect of haplotypes (Schaid, 2004).

The other way in which haplotypes may aid in GWA studies is by increasing statistical power to detect marginal effects of single causative QTL. The reason for this is that where a single marker may only capture a certain proportion of the variance of a nearby causative SNP, other nearby markers may capture variance not already explained by this first marker. A haplotype may capture all of this variation simultaneously by virtue of accounting for all first order, second order and any higher order interaction terms present between markers within the haplotype (Schaid, 2004). It has been shown that using two and three SNP haplotypes can increase the number of common SNPs tagged at an r^2 of at least 0.8 by 25-100% (Pe'er et al., 2006). The

concept is similar to the example above, except that here the haplotype tags a single SNP variant contributing to the disease / trait rather than several SNPs working together to contribute. It is easier to envisage this situation by imagining a relatively young causative mutation being introduced to a population on some previously established haplotype background. Due to disparate allele frequencies, there will be little r^2 at the population level between the causative SNP and any nearby SNPs, however the new mutation will be in strong LD with the specific set of alleles (i.e., haplotype) it arose on.

The type of situation described above is more likely to be found when a causative mutation has been introduced by only one or a limited number of founders, and also where fairly few generations have passed since the introduction of the mutation. This is because fewer founders introducing the mutation means that a greater proportion of those carrying the mutation in the current generation will be identical by descent (IBD). Consequently, nearby regions surrounding the mutation in these individuals will also be IBD, and the individuals are therefore more likely to share common haplotypes. Similarly, the fewer generations back to the introduction of the mutation there are, the less chance there is for recombination to break down LD around the mutation. An isolated population is most likely to fulfil these criteria and for this reason may be a particularly suitable population for use of haplotype-based association methods (Bourgain and Genin, 2005).

The argument about young variants influencing traits is one of the reasons why the common disease / rare variant (CD/RV) hypothesis is becoming more popular, since

all young mutations will also be rare without the presence of strong positive selection. There is also increasing evidence to suggest that rare variants contribute to common complex disease (see for example Crawford et al., 2004), and that common SNPs on current genome-wide SNP panels are not good at tagging them (Bodmer and Bonilla, 2008). This is one area in which haplotypes should perform favourably when compared to single SNPs, since joint LD of SNPs in a haplotype background are more adept at capturing rare variation. If the CD/RV hypothesis is correct, then GWA studies would greatly benefit from using haplotypes-based analyses.

4.1.2 Issues with using haplotypes

While there are valid reasons to expect that analysing haplotypes may greatly help detect QTL, there are also situations in which haplotypes will never perform as well as a single SNP. In addition to this, there are a number of considerations and unresolved issues which can make haplotypes difficult to implement and therefore less attractive, thus explaining why haplotype analysis is not currently favoured as the method of choice in GWA studies. Some of the issues pertaining to haplotype analyses are introduced and discussed below.

4.1.2.1 Factors affecting haplotype analysis

There are a number of factors that determine the performance of haplotype-based analyses, and thus how well haplotype methods compare in efficiency and power to single SNP analysis. Foremost of these are the number of loci affecting a trait, the

number of alleles at loci affecting the trait, the density of markers, LD between markers, LD between markers and QTL, and finally the allele frequencies of both the markers and the QTL (Schaid, 2005). These factors are discussed with reference to the recent literature in this section. It is important to note however, that the vast majority of relevant literature (and indeed this analysis) assumes that all trait loci are biallelic, but this may not always be the case.

It has been shown that haplotype regression analysis is more powerful than single SNP regression when the number of haplotypes present is less than the number of (non-interacting) trait loci (Bader, 2001). Results from this study also imply however, that where there are only one or two QTL within a region, as expected most of the time, single SNP analysis will be more powerful. However, this study was performed under optimal conditions for single SNP regression, as the QTL themselves were included as markers for the analyses. Where this is the case (or if a SNP has r^2 of one with a QTL), single SNP regression cannot be improved upon by adding extra markers because these add parameters to the model without adding new information. This highlights the importance of the role that marker-QTL LD plays in the relative efficiencies of single- and multi-marker methods. As pairwise LD between markers and the QTL drops, and tends towards the extreme of linkage equilibrium, haplotype analyses will become more strongly favoured (Bader, 2001). As already noted, one way for low LD to exist between markers and QTL is if the QTL is recent, although there is currently not enough evidence to say how often young mutations are disease predisposing.

Another important factor for haplotype analysis is the level of LD present between markers. While r^2 between SNPs is loosely correlated to their distance apart, patterns of LD can be complex, and the variance of pairwise r^2 statistics is large for pairs of SNPs of equal distance apart. Ideally for haplotype analysis there would be moderate to high LD between markers comprising haplotypes, since the number of haplotypes increases as LD between markers decreases, until at linkage equilibrium there is the theoretical maximum number of haplotypes (2^n for biallelic loci where *n* is the number of polymorphic loci). At this extreme, haplotypes only introduce noise into association tests at the cost of additional degrees of freedom. At the other extreme, with complete LD, there would only ever be two haplotypes and any association test would be equivalent to testing individual SNPs. More haplotypes means fewer degrees of freedom with which to test for association, and herein lies the trade-off with haplotype analysis. Extra variation captured by a model by using haplotypes must more than compensate for the extra degrees of freedom the haplotypes take up. This was illustrated in a power study of haplotype analysis under three different haplotypic distributions. In order to maintain power for the flat distribution (i.e., where each possible haplotype had equal frequency), sample size had to be dramatically increased (Schaid, 2004), illustrating that the distribution of haplotype frequencies is key to the power of the haplotype-based tests, and LD between SNPs is key to this haplotypic distribution.

Allele frequency is another important issue for association studies, whether for single SNP- or haplotype-based analyses. One obvious reason for this is that statistically speaking it is not ideal to include low frequency classes in analyses (hence why rare

alleles are often removed from GWA studies by quality control), but additionally, allele frequency is closely linked to the r^2 measure of LD. Loci with markedly different allele frequencies will have a smaller pairwise r^2, and r^2 can never be one unless allele frequencies are identical. Following from this, loci with different allele frequencies will have lower pairwise LD, and consequently there will be a greater number of underlying haplotypes. In regard to QTL detection, disparate frequencies of marker and QTL alleles has a direct impact on the ability to detect the QTL since marker-QTL r^2 is correlated to power (Blangero, 2004). Frequently in the literature on simulated haplotype analyses, only very common QTL alleles are tested (for example Meuwissen and Goddard, 2000), which implicitly favours single SNP analysis, since SNP panels are generally designed to tag common SNP variants by inclusion of almost exclusively common SNPs.

While there has been much work in elucidating how the above factors affect the relative efficiencies of single marker and haplotype methods, uncertainty still remains as to which methods perform best for GWA studies. This has lead to a tendency to overlook more sophisticated analysis methods, since the advantages are not well characterised, and it is generally faster and simpler to use single SNP regression. The above factors are not the only considerations involved in haplotype analyses however, and some more of these are discussed below.

4.1.2.2 Haplotyping algorithms

The most obvious disadvantage of using haplotype methods to analyse genome-wide association data is that usually the haplotypes are all unknown, and must therefore be estimated from the genotype data first. While molecular methods to experimentally determine haplotypes from diploid DNA do exist (for example the long-range allele-specific PCR method of Michalatos-Beloin et al., 1996), these methods are not widely utilised since they are expensive and generally low-throughput (Niu, 2004). Consequently, haplotypes are usually estimated using statistical methods, and there are a variety of these for estimating haplotype frequencies in either family-based or population-based studies. Family-based approaches, for example Merlin (Abecasis et al., 2002), are deterministic and can work well for small pedigrees and few SNPs, however larger and/or sparse pedigrees with high numbers of SNPs can cause problems both in terms of information to accurately phase all individuals (e.g., if key individuals are heterozygous and hence not informative), and also in terms of computational feasibility. Of particular concern for family-based haplotype reconstruction methods is that most make the assumption of linkage equilibrium between all markers, meaning that whenever phase is ambiguous, all potential haplotypes consistent with the genotypes are classed as equally likely (Niu, 2004).

The simplest, and one of the most popular algorithms now used for population-based haplotype estimation, is the Expectation-Maximisation (EM) algorithm, developed in the 1970s (Dempster et al., 1977), and first applied to estimating population haplotype frequencies in 1995 (Excoffier and Slatkin, 1995). The EM algorithm entails a two-stage iterative procedure that gradually converges on the most likely

population frequencies for haplotypes observed in the study population. The haplotype frequency estimates are population-based, and therefore based on the assumption that all individuals in the sample are independent, i.e., unrelated. Related individuals can cause an upwards bias in the frequency of some haplotypes since haplotypes of related individuals are not independent. The original version of the EM algorithm had stringent limitations regarding both the number of individuals and the number of loci that could be phased simultaneously. This was because the algorithm started by enumerating all possible haplotypes from the data before dropping those that were unobserved, or observed at very low frequency. As already noted, only 20 SNPs are required to take the number of possible haplotypes above one million, so the computation involved to enumerate all possibilities quickly becomes unfeasible. Later versions of the EM algorithm have been improved upon so that enumerating all possible haplotypes is no longer necessary, and frequencies are only estimated for haplotypes consistent with the observed data.

There is also a large number of population-based haplotype estimation algorithms based on coalescent theory. One of the most popular implementations in this bracket of methods is a Bayesian program called PHASE (Stephens et al., 2001) which arrives at haplotype frequency estimates using Markov-chain Monte Carlo (MCMC) iterations. Later versions of this algorithm incorporate a useful "divide and conquer" feature that allows phasing of longer tracts of DNA than could previously be handled (Stephens and Donnelly, 2003). This technique is called partition ligation (PL), and works by subdividing the data into smaller segments and phasing these individually before re-joining the segments back to original length. Partition ligation was initially

introduced by Niu et al. (Niu et al., 2002), and was subsequently incorporated into many other methods. There is now a whole suite of Bayesian PL methods, and indeed a number of EM methods which also utilise PL (for example the "PLEM" method of Qin et al., 2002).

The main drawback of any haplotyping algorithm is that inevitably some level of error will be introduced into the data. However, most of the more widely used algorithms manage to limit errors to a very small proportion of results. For example, Haplotyper (Niu et al., 2002) and PLEM (Qin et al., 2002), two of the recent partition ligation EM methods were compared in a simulation study, and were shown to have almost 100% accuracy in six different populations representing different haplotype distributions (Niu, 2004).

4.1.2.3 Rare haplotype classes

Obtaining haplotypes is not the only consideration when using haplotypes for analysis of GWA data. For example, it is usually the case that some haplotype frequency estimates are extremely low (well below 0.05 for example), and when this is the case, it is not obvious how these rare classes should be treated, and nor is it clear what the most appropriate threshold is to declare a haplotype rare. From a statistical standpoint, it can be problematic to include rare classes in an analysis for two reasons. Firstly, rare data classes can lead to unreliable effect estimates and spurious positive associations, and secondly, a large number of rare haplotype classes use up a correspondingly large number of degrees of freedom with which to test for

association, thereby having a detrimental effect on the power of the haplotype test (Schaid, 2004). However, it is not clear that other ways of dealing with rare haplotype classes are superior.

One of the alternative ways of dealing with rare haplotypes is to pool all rare classes and analyse these together, often as the base of comparison for all other haplotype classes. This approach is not optimal because potentially very large numbers of disparate haplotypes are being grouped together, and this may swamp any true effects of the mixed classes. Also, it makes sound interpretation extremely difficult if this "junk" bin is found to be associated with the trait. Another way to handle rare classes is to exclude them from analysis entirely, but this is unappealing because it may mean losing valuable information from the analyses. A slightly more sophisticated way to incorporate rare haplotype classes is to group all related haplotypes using a clustering algorithm (e.g., Morris, 2005; Li et al., 2006), although this may still homogenise important differences between some haplotype groups.

4.1.2.4 Method of analysis

The most variable aspect of haplotype-based tests is the method of analysis. Many different ways to analyse haplotype data have been proposed, however there is little consensus upon which is best. A typical method of analysis would involve a score test or some sort of regression of the trait onto an X-matrix consisting of the haplotypes, but even then parameterisation of the test can vary. The method will also vary depending on the exact nature of the haplotype data. Many haplotyping algorithms

produce both a best estimate of an individual's haplotype pair (a "best-pair estimate"), and probabilities for each of the possible haplotype pairs per individual (Stephens and Scheet, 2005). Details of the analysis will vary depending upon which of these outputs is used.

When best-pair estimates are used, there are two main ways in which this information can be utilised for testing the data. The first of these is to test each haplotype separately against one joined pool of all the remaining haplotypes. One problem with this is that the joined pool will consist of many heterogeneous haplotypes, and some of these may be similar to the tested haplotype, making an association both hard to detect and difficult to interpret. Not only that, but depending on the number of distinct haplotypes in the data, there may be a large number of tests performed, which must be corrected for in order to account for type I errors. Again, it would be possible to reduce the total number of tests performed in this situation by using a haplotype clustering algorithm. This use of best-pair estimates is best suited as a direct approach for when a particular haplotype is already suspected to have an effect on the trait.

The second way in which best-pair haplotype estimates can be used is to fit all separate haplotypes individually in the model as fixed or random effects, and simultaneously estimate effects for all of them. Wald's F-statistic, for example, could then be used as an overall test for association of all haplotypes on the mean trait value. The numerator degrees of freedom for this test is the number of haplotypes, therefore this technique also suffers from the problem of reduced power when the total number of haplotypes is high, as there may no longer be sufficient degrees of

freedom left to reject the null hypothesis. Another problem with this type of test is that as the number of haplotypes increases the test becomes rather general, since no single haplotype is implicated as being the cause of association.

The strategies mentioned above are all based on using best-pair haplotype prediction for each individual. However, this makes the assumption that the best-pair prediction for each individual is known to be the correct diplotype, which may introduce substantial error into the data if the prediction is wrong. For example, given a situation where an individual has two potential diplotypes with probabilities 0.49 and 0.51, the individual will always be assigned the diplotype that has 0.51 probability. Or, to consider an even more extreme example, if an individual has ten possible diplotypes each of approximately equal probability, the diplotype with the highest probability will be assigned to the individual with certainty, and no record of the uncertainty with which the diplotype was assigned is retained.

Where information exists about the probability of each possible pair of haplotypes an individual might possess, a more complex regression model can be fitted allowing each individual to be a mix of many probabilities for different haplotypes. Haplotypes are fitted as linear covariates in this parameterisation, each individual having a vector of length n (n = number of haplotypes) giving the probability of possessing each haplotype. Again, overall significance of the haplotypes on the trait can be tested using Wald's F-statistic. One disadvantage of this parameterisation is that no dominance or other deviation from additivity can be tested for however.

4.1.3 Lessons from the Literature

There has been extensive work over the last few years in an attempt to determine how haplotype methods perform compared to single SNP and multi-SNP genotype methods, and many novel haplotypic tests have consequently been proposed. Many of these methods were designed to detect loci affecting binary phenotypes in case-control studies, however they can still provide valuable insights into what may be expected for quantitative phenotypes. The typical way to test for differential haplotype frequencies between case and controls is to use the EM algorithm to estimate haplotype frequencies, then calculate sample likelihoods in cases, controls and the pooled sample (L_1, L_2 and L_3). The chi-square distributed likelihood ratio test statistic is then $\chi^2 = -2 \ln[L_1/(L_2L_3)]$, with degrees of freedom given by the number of haplotypes minus one (Zaykin et al., 2002). Numerous alternative methods have now been suggested, and debate concerning the relative merits of using haplotypes compared to single marker analysis continues. Many of the methods that have been proposed for testing haplotypic effects in case-control studies are described in detail in Cordell, 2006, although this is not in direct comparison to how single SNP analyses perform.

Considerable evidence from the literature exists to suggest that haplotypes can increase power to detect associations for binary phenotypes (e.g., Akey et al., 2001; Morris and Kaplan, 2002; Li and Jiang, 2005). Akey et al. point out that haplotype tests are more robust than single marker analyses because, where evolutionary forces (such as random drift and mutation) and varying degrees of initial marker-QTL LD act to increase the variability of observed pairwise LD between marker and disease

loci, simultaneous analysis of multiple markers as haplotypes can result in comparatively simpler patterns of LD (Akey et al., 2001). In the case of Morris and Kaplan, it was demonstrated that haplotype-based analyses are optimal for detecting disease susceptibility loci when multiple alleles are present at a single disease susceptibility locus (Morris and Kaplan, 2002), even though power will be below that possible if the locus was biallelic (Slager et al., 2000; Longmate, 2001). In this multi-allelic situation, haplotypes do best when there is less LD between markers, therefore increasing the number of potential haplotypes, since this allows each disease locus allele to be present on a unique haplotype.

Other case-control based studies have found that relative efficiencies of the contrasting methods are dependent upon the level of LD between markers, such that the optimal method varies (Nielsen et al., 2004), and yet others suggest that single SNP analyses, or analyses based on multiple SNP genotypes are always more powerful than using haplotypes (Chapman et al., 2003; Clayton et al., 2004). It is undoubtedly be the case that single SNP analysis will perform best when a single causative QTL is in extremely high LD with a typed SNP, for example.

In other situations, It is also conceivable that multiple different SNPs each in partial LD with a given QTL with no strong tagging SNP will pick up extra marginal effects and hence perform better than any other method. In this latter case, it is unlikely that haplotype analysis would perform better, since many more degrees of freedom would be used to explain only a little more variance. However, it is also likely that in some situations a haplotype will be better able to characterise the variation present at a QTL

and hence capture the effect more efficiently than any single marker or set of unphased markers. The frequency with which this might be the case is one thing the literature is unsure of however, and furthermore is currently unable to address, given that haplotypes are relatively unused in GWA studies to date.

More recently, there has been a plethora of association methods designed to use haplotypes to analyse quantitative phenotypes as opposed to binary ones. Once again there are no conclusions as to which method is definitively best, but there is yet more evidence to suggest that haplotypes can out-perform single SNP analysis. Often, where results suggest that single SNP analysis is uniformly best, it is not unexpected given that the parameters of the analysis strongly favour this outcome, for example by including the QTL as part of the tested marker set (Jannot et al., 2003), having high QTL minor allele frequency (Zhao et al., 2007), or by having extremely dense markers likely to be in high LD with the QTL (Grapes et al., 2004). These are all situations which would be ideal for QTL detection regardless of the method used, and therefore it seems reasonable to surmise that the majority of QTL falling into this category have already been found – the challenge remains to find QTL which have so far proven elusive however.

There are many papers which have reported haplotypes increasing the power of an association test. One of the key factors in developing a good test for association is being able to take into account phase uncertainty from the haplotype prediction (in situations where phase is not already known). Only selecting the most probable haplotype pairs can cause substantial loss of information, particularly if LD is not

strong between the marker SNPs, since then confidence in the most probable pair is lower (Schaid et al., 2002). This loss of information introduces errors into the X-matrix (haplotype scoring matrix) which in turn can lead to biased estimates of haplotypic effects (Zhao et al., 2003), and possibly inflate the error of estimated parameters (Tanck et al., 2003). It has been suggested that haplotype uncertainty is not as bad in family-based studies since there is often enough information to be accurate (Lee and Van Der Werf, 2005), however it has been found in population-based association studies that simply assuming the most likely diplotype is true can have detrimental effects (Morris et al., 2004).

Due to the importance of accounting for haplotype ambiguity appropriately, recent methods now explicitly model the probabilities of all possible haplotype pairs for each individual (Schaid, 2004). Many of these models have been shown to out-perform single SNP analyses even when there is only a single QTL influencing a trait (e.g., Zaykin et al., 2002; Becker and Herold, 2009), and the relative performance of haplotypes compared to single SNP analysis should improve as the number of QTL loci increases (Bader, 2001). The vast wealth of existing data on which method of analysis is likely to work best for GWA studies only acts to highlight the truly complex nature of the underlying genetics, and emphasises the fact that as yet there are still very few concrete conclusions. Much work clearly remains to be done in order to clarify when and where the contrasting models will do best.

4.1.4 Current analysis

In this chapter, several different methods for analysing GWAS data are compared; single SNP regression, multiple regression using three, five and seven SNPs, and haplotype analysis using windows of three, five and seven SNPs. These analyses should add to the current literature and help elucidate under which circumstances the respective methods perform best. In particular, a key aspect of these analyses was to explore how the different methods perform for QTL with rare minor allele frequency (MAF), something not investigated in the literature to date. In addition to assessing how the overall methods perform, these analyses should also give some indication about how differing window lengths for the multi-marker methods affects their power to detect QTL.

In these analyses a single causative polymorphism at one locus was considered, which should be the most beneficial situation for single SNP analysis (Bader, 2001; Schaid, 2004). To determine which methods were performing best, "pseudo-traits" were created using real genotype data. This was achieved by removing one SNP at a time from the marker genotype data, and adding to the genotype value of the removed SNP a random number drawn from a normal distribution. The additional noise reduced the heritability of the "trait" and also made it more truly quantitative. Local SNPs were used in each of the statistical models in an attempt to detect the surrogate QTL (sQTL), and this was performed for each SNP along a whole chromosome. Comparing results from each of the different methods should help clarify the role haplotype analyses can play in GWA studies.

4.2. MATERIALS AND METHODS

4.2.1 Genotype data

The data used in this chapter were genotypes from chromosome 4 of the CROAS dataset introduced in chapter one. Using real genotype data meant that realistic patterns of genomic architecture were used (LD patterns, allele frequencies and haplotypes) rather than simulated IBD patterns and population genetic parameters (Hoggart et al., 2007). Quality control was performed such that SNPs or individuals with low call rate (<95%) were removed. Rare SNPs, i.e., those with a minor allele frequency below 0.05, were retained because these were of intrinsic interest to the study. Only founders and singleton individuals were used due to the fact that the haplotyping algorithm used for these analyses (PHASE) assumes all individuals are independent (i.e., unrelated). Including related individuals would therefore upwardly bias the population frequency of certain haplotypes due to sharing between family members. After quality control, the final dataset consisted of 17,022 SNPs and 453 individuals, down from 19,113 SNPs and 986 individuals.

4.2.2 Phenotype data

No "real" phenotypic data were used in these analyses. Instead, phenotypes were generated using the genotypes of SNPs that comprised the genotypic dataset. Each SNP used as a phenotype took the raw genotype score for each individual (i.e., 0, 1 or 2 corresponding to number of a given allele), and added to this a random number

drawn from a normal distribution, in order to produce 70% noise. The added noise was calculated in the following way:

$$\sigma^2_{noise} = (1/0.3)\sigma^2_{sQTL} - \sigma^2_{sQTL}$$

where σ^2_{noise} is the variance added corresponding to noise, and σ^2_{sQTL} is the variance of the sQTL genotypes. In this way, the heritability of the "trait" was reduced from 1 to 0.3. A heritability of 0.3 was decided upon for reasons explained subsequently.

4.2.3 Performing the analysis

4.2.3.1 Outline

The aim of this study was to compare the performance of several methods for analysing GWAS data in terms of their power and ability to capture trait variation. The methods compared were single SNP regression, multiple regression using three, five, and seven SNPs, and haplotype analysis using windows of three, five and seven SNPs. With 17,022 SNPs used as pseudo-traits, performing a genome-wide scan with all methods for each one was not feasible, and any significant results arising from SNPs on different chromosomes would only represent false positives anyway. Therefore to cope with the computational and time constraints, a test region of 100 SNPs (50 in each direction) surrounding each sQTL was selected to perform analyses on. Since each sQTL required 50 SNPs to either side of it, this meant that the first sQTL was the 51st SNP of the dataset, and the last was the 16,972nd SNP, and

therefore the total number of sQTL was 16,922. Note that the total number of tests performed on the test region of each sQTL varied depending on the method used. For single SNP regression there was always exactly 100 tests, but with three-SNP multiple regression for example, only 98 tests were possible because that is the highest number of three-SNP windows possible in 100 SNPs.

4.2.3.2 Single SNP regression

For each sQTL, simple linear regression was performed on each SNP within the 100-SNP test region. The SNPs were coded as 0, 1, 2 or NA for missing, and were fitted as linear covariates, therefore the test was additive. No covariates or fixed / random effects were fitted. The sQTL was dropped from the set of test SNPs and therefore not tested. The model is as follows:

$$y_i = \mu + kg_i + e_i$$

where y_i is the phenotype for the i^{th} individual, μ is the mean, k is the additive effect of an allele, g_i is the genotype of the i^{th} individual, and e_i is the residual error term for the i^{th} individual. A two-sided F-test was used to test the effect of the SNP, with one and n-2 degrees of freedom (df), where n is the number of individuals in the test. This method is referred to as SSR (single SNP regression).

4.2.3.3 Multiple regression

Multiple regression was performed using a sliding window across the 100-SNP test region of each sQTL, moving forward a single SNP at a time. Regression models using windows of three, five and seven SNPs were used. This meant that different numbers of tests were performed for the sQTL over different window lengths, since the number of *n*-SNP windows possible within 100 SNPs decreases as *n* increases. The total number of tests performed for each sQTL was 98, 96 and 94 for multiple regression window lengths of three, five and seven respectively. As for SSR, sQTL were always dropped from the test region, and therefore never used in any of the windows.

Again, SNPs were coded as 0, 1, 2 or NA for missing. SNPs were fitted as linear covariates, and no other covariates or fixed / random effects were fitted. The model is:

$$y_i = \mu + \Sigma_j k_j g_{ji} + e_i$$

where k_j now represents the effects of the j^{th} SNP and g_{ji} is the genotype of the i^{th} individual for the j^{th} SNP. The test performed for each model was Wald's F-statistic test, which tests for an overall association of SNPs with the sQTL. The F-test has *n* and *N* degrees of freedom, where *n* is the number of SNPs in the regression model, and *N* is the number of non-NA individuals minus *n* minus one. The multiple

regression methods are referred to as MR3, MR5 and MR7 respectively for window sizes of three, five and seven.

4.2.3.4 Haplotype analysis

As with the multiple regression methods, haplotype analyses were conducted using three, five and seven SNPs. To perform haplotype analyses, haplotype estimates must first be obtained. This was carried out using the statistical software PHASE, which calculates not only the most likely diplotype for each individual, but also diplotype probabilities for all possible haplotype pairs. While the analyses themselves comprised of haplotypes consisting of three, five and seven SNPs, in order to ensure haplotypes were as accurate as possible, 11 SNPs at a time were phased, and haplotypes for the middle *n* were extracted for analysis. Each window of 11 consecutive SNPs from the beginning to the end of the chromosome was haplotyped. In addition, extra windows were also haplotyped to account for situations where the window spanned an sQTL position. For each sQTL ten windows required dropping of the sQTL to haplotype. Consequently, the number of windows haplotyped was increased approximately ten-fold.

In determining that 11 was the optimal number of SNPs to phase, a trade-off was made between the time taken to phase a single window of *n* SNPs and the accuracy of the estimates produced. 11 SNPs represented the best compromise as there was little difference in the frequency of the central seven-SNP haplotypes when using 11 SNPs to estimate them than from haplotypes estimated using longer windows. The

difference was even less so for haplotypes of length three and five SNPs. The increase in time taken to phase windows was not linear with number of SNPs, and became prohibitive above 11 SNPs. PHASE took approximately 41 seconds to run with 11 SNPs, therefore performing all the phasing required for these analyses would take 90 days for a single computer. However, increasing the number of SNPs phased each time to 13 would have increased this figure to over 195 days. Haplotyping 11-SNP windows meant that for all haplotype methods there were 90 tests for each sQTL test region. This is because the last full 11-SNP window available within any 100-SNP test region will encompass SNPs 90-100 (inclusive).

Once haplotype estimates were obtained from PHASE, diplotype probabilities for each individual were used to calculate probabilities of each three, five and seven SNP haplotype for that individual. The probabilities for each individual / haplotype combination were calculated by summing the halved diplotype probabilities from a given individuals' possible diplotypes. For example, consider the simplified situation below:

Diplotype: 0 1 1 0 **1 1 1** 1 1 1 1 / 1 0 0 1 **0 0 0** 0 0 0 0 probability of 0.6

Diplotype: 1 0 1 0 **1 1 1** 0 0 1 1 / 0 1 0 1 **0 0 0** 1 1 0 0 probability of 0.3

Diplotype: 0 0 0 1 **0 0 1** 1 1 1 1 / 1 1 1 0 **1 1 0** 0 0 0 0 probability of 0.1

This individual would have the following three-SNP haplotype probabilities:

Haplotype: 1 1 1 probability of 0.45

Haplotype: 0 0 0	probability of 0.45
Haplotype: 0 0 1	probability of 0.05
Haplotype: 1 1 0	probability of 0.05
All other population haplotypes:	probability of 0

These probabilities would then be used to create a haplotype vector for each individual, where for each possible haplotype there was a non-zero probability, and for each haplotype not possible the probability was zero (so the length of the vector is equal to the number of haplotypes in the population for this window). In this way, for every 11-SNP window phased, each individual had one vector for each haplotype length of lengths "n_3", "n_5" and "n_7", corresponding to the number of haplotypes in the population for haplotype lengths three, five and seven respectively, with each vector summing to unity. All haplotypes were kept, regardless of frequency.

The haplotype probabilities for each haplotype were fitted in a regression model as linear covariates. No other covariates or fixed/random effects were fitted in the model. The model was identical to that shown for multiple regression, except that "j" now refers to haplotypes, not SNPs. The test performed on these data was Wald's F-statistic with numerator degrees of freedom equal to the number of haplotypes in the model, and denominator degrees of freedom equal to *N* (number of non-NA individuals) minus numerator degrees of freedom minus one. As already mentioned, this is only one of a number of ways that such haplotype data can be analysed, and is similar to the haplotype trend regression (HTR) method of Zaykin et al. (Zaykin et

al., 2002). The haplotype analyses are referred to as HL3, HL5 and HL7 respectively for haplotype lengths of three, five and seven SNPs.

4.2.3.5 The analyses

Analyses were conducted using in-house scripts written in the computer programming language Perl and the statistical software R programming language. Haplotyping via PHASE was called from Perl scripts, and relevant parts of the output from PHASE were then extracted. To perform the statistical tests, the scripts called R and used the regression function lm(). The proportion of variance explained from each model, along with the overall p-value for association, were recorded for each test performed for each sQTL. For example, for three-SNP multiple regression there were 98 tests per test region, and 16,922 test regions in total, therefore there were 1,658,356 lines of results for this method.

Both the p-value and the proportion of variance explained by the model were extracted, as these are both indicators of how well a method has performed. As the basis of comparison across methods, the p-value is the better statistic of the two, as this takes into account the fact that the methods have different numbers of parameters. In general, the more parameters there are in a model, the more variance will be explained, however it is not always the case that the model explaining most variance is the best (i.e., most significant) model. This is because a highly parameterised model may explain only slightly higher proportion of variance than a model with less parameters, but it will have less degrees of freedom to test with. This

is important, particularly in the haplotype analyses, because the total number of explanatory variables may be large; for example the theoretical maximum number of haplotypes with seven SNPs is 128. For this reason, using the p-value (or $-\log_{10}$ p-value) provides a more appropriate comparison across methods than the amount of variance explained since it accounts for the differing degrees of freedom of the tests.

4.2.4 Determining the phenotype heritability

A heritability of 0.3 was decided upon based on permutation analyses to calculate the empirical distribution of the test statistic under the null hypothesis. Due to the computationally intensive nature of permutation analysis, and therefore the inherent time constraints, four methods were chosen on which to perform the permutations; SSR, MR3, MR7 and HL7. Of the haplotype methods, HL7 was selected because this is the one that should exhibit the most extreme test statistic inflation. This is because HL7 will have a large number of classes (i.e., haplotypes), some of which will be rare, and there is a greater probability that by chance a permutation creates spurious association between one of these haplotype classes and the sQTL. This is especially true for sQTL with low minor allele frequency, therefore of particular interest was how the test statistic of sQTL with low MAF behaved. Consequently, while six sQTL with differing MAF were permuted, four of these were sQTL with MAF < 0.05. The MAF of the six sQTL permuted were 0.018, 0.025, 0.037, 0.046, 0.252 and 0.457. For each method and sQTL combination, 1,000 permutations were performed.

The sQTL phenotypes were permuted over individuals, therefore breaking correlations between sQTL and SNPs, but retaining patterns of LD between SNPs. For each permutation, the whole of chromosome 4 was used to test for association (which lead to marginally different numbers of tests over methods, due to the differing window lengths). The most significant p-value from each permutation was retained, and a distribution of most significant p-values of length 1,000 was created for each method and sQTL combination. From this distribution of most significant p-values, the 95th percentile was used to determine the true -$\log_{10}$ p-value required to reject the null hypothesis at a chromosome-wide significance level.

Results for sQTL where no noise was added to the trait (i.e., the phenotype is the raw genotypic value 0, 1 or 2) are shown in Figure 4.1. Assuming a Bonferroni correction, the -$\log_{10}$ p-value for chromosome-wide significance for all methods is 5.53. As can be seen from Figure 4.1, the three regression methods have slightly elevated -$\log_{10}$ p-values for the four SNPs with MAF below 0.05, but in comparison, the test statistics for HL7 are vastly inflated. Interestingly, -$\log_{10}$ p-values are all slightly below 5.53 for the common sQTL. This is likely to be because for each method, tests for neighbouring SNPs / windows are correlated due to LD between markers, therefore the number of independent tests is less than the total number of tests performed, and the Bonferroni correction is too stringent. Tests for methods using more SNPs are more highly correlated, and correspondingly the 95% threshold is smallest for HL7 and MR7.

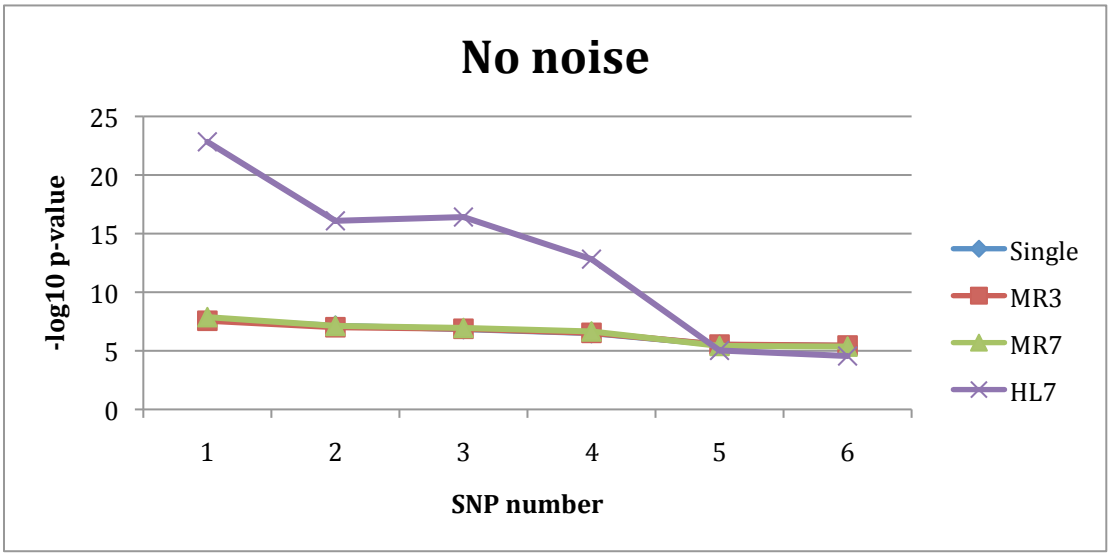


Figure 4.1 Empirical threshold for chromosome-wide significance for SSR, ML3, ML7 and HL7 for a variety of sQTL with different MAF. SNP numbers correspond to the six SNPs mentioned in the text to be used as sQTL. MAF for these sQTL are 0.018, 0.025, 0.037, 0.046, 0.252 and 0.457 respectively.

Results of permutations for sQTL with 50% and 70% noise are shown in Figures 4.2 and 4.3. These show that test statistic inflation for rare sQTL is greatly reduced by introducing noise to the raw sQTL value. The largest test statistic inflation is still seen for the sQTL with lowest MAF, however with 70% noise (Figure 4.3), the 95th percentile of the test statistic distribution for HL7 is just over seven for the rarest sQTL, and this drops to just over six for the rare sQTL with highest MAF (of the four rare sQTL). This indicates that for rare sQTL, test statistic inflation for HL7 will be minimal. Test statistics for the other methods remain roughly constant across all sQTL for 50% and 70% noise. Note the greater reduction in test statistic for common SNPs with HL7 once again.

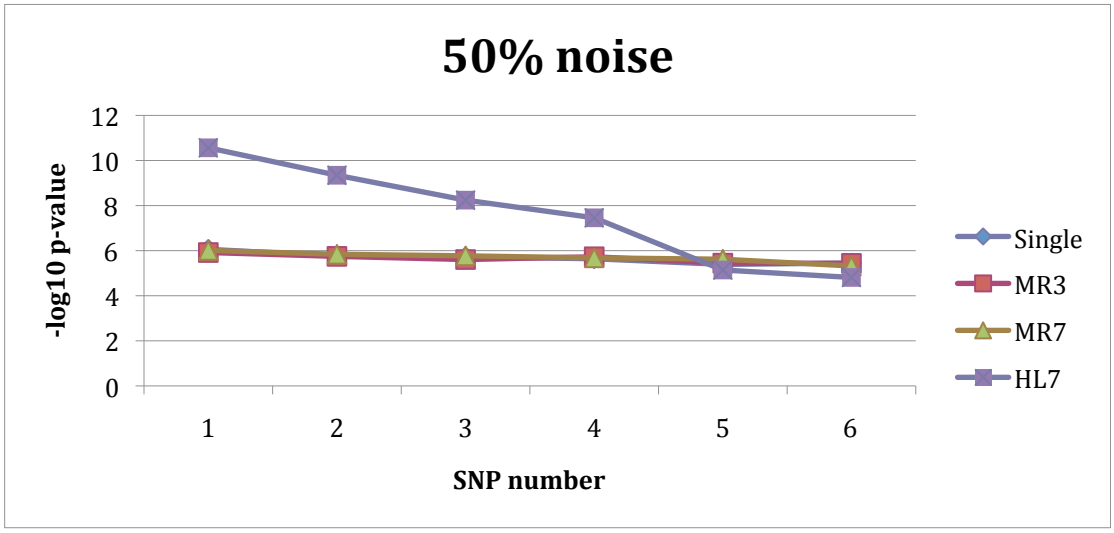


Figure 4.2 Empirical threshold for chromosome-wide significance for SSR, ML3, ML7 and HL7 for a variety of sQTL with different MAF. 50% noise is added to each of the sQTL. SNP numbers correspond to the six SNPs mentioned in the text to be used as sQTL. MAF for these sQTL are 0.018, 0.025, 0.037, 0.046, 0.252 and 0.457 respectively.

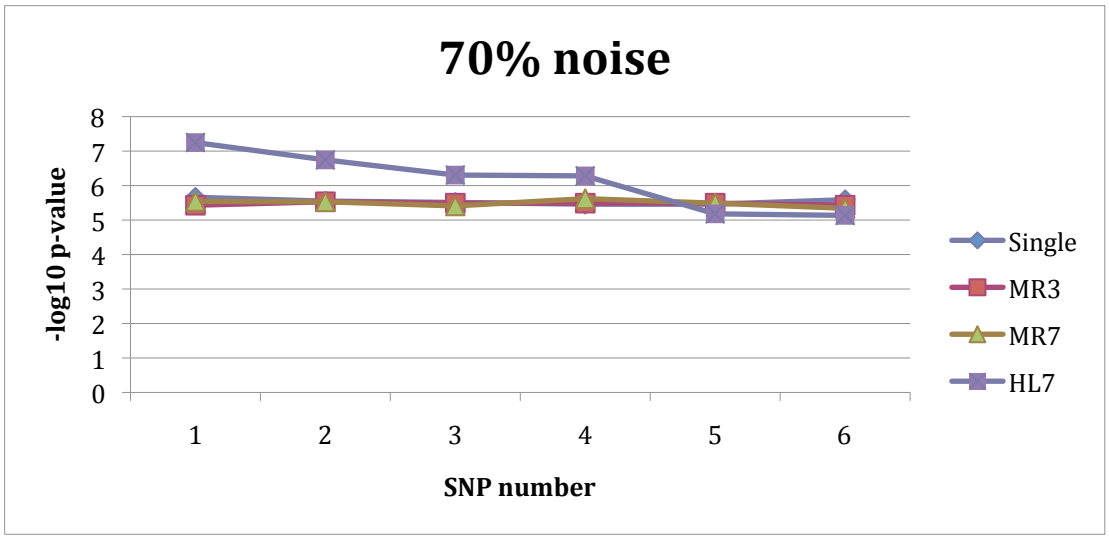


Figure 4.3 Empirical threshold for chromosome-wide significance for SSR, ML3, ML7 and HL7 for a variety of sQTL with different MAF. 70% noise is added to each of the sQTL. SNP numbers correspond to the six SNPs mentioned in the text to be used as sQTL. MAF for these sQTL are 0.018, 0.025, 0.037, 0.046, 0.252 and 0.457 respectively.

Results from these permutations indicate that introducing noise to the sQTL is essential to prevent large inflations of the test statistic for sQTL with low MAF. At 70% noise, there is still minor inflation of the test statistic for HL7 however. This

means that p-value (or $-\log_{10}$ p-value) comparisons across methods for sQTL with MAF this low will be slightly biased in favour of HL7, since this method has higher test statistics under the null hypothesis. However, it is likely that the inflation for HL7 still present at 70% noise is largely a problem unique to this method, since the number of classes for both HL3 and HL5 will be less (and the frequencies of these classes less likely to be very rare). Indeed, the average degrees of freedom fitted for each of HL3, HL5 and HL7 was 6, 12 and 22 respectively. Consequently, test statistic inflation should be much less, and perhaps not evident at all, for HL3 and HL5. Increasing noise further would greatly affect power for all methods, thus decreasing both the ability to distinguish between them, and the ability to draw meaningful conclusions about their relative performances. Therefore, a noise comprising 70% of sQTL variation was selected for the main analyses.

4.3. RESULTS

4.3.1 Diagnostics

To characterise the minor allele frequency distribution of the sQTL in this study, the MAF for each SNP used as a sQTL was calculated, and the sQTL placed MAF bins accordingly. Figure 4.4 shows the total number of sQTL in each MAF bin. The nine bins with MAF above 0.05 each contain more than 1,500 sQTL, and are comparable in sQTL number. The MAF bin with the most sQTL is 0.1 – 0.15, containing 2,121 sQTL. The bin with the fewest sQTL is the one containing sQTL that would be classed as rare SNPs, the 0 – 0.05 category. There are only 579 sQTL in this category

- almost a third less than the next smallest group. This is almost certainly a consequence of ascertainment bias in the type of SNPs selected for inclusion on the Illumina SNP panel, since their aim was to provide common SNPs.

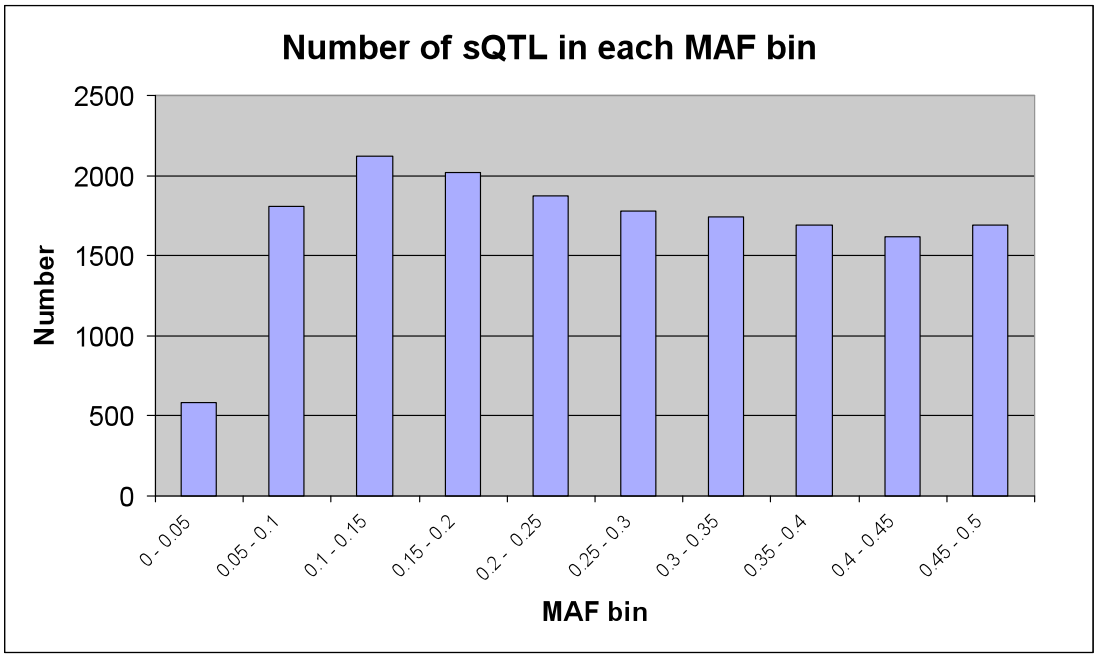


Figure 4.4 The number of sQTL that have minor allele frequencies within each of ten different 0.05-wide MAF bins.

The frequency distribution of the distance between adjacent SNPs is shown in Figure 4.5. The plot shows the expected pattern, with a high proportion of adjacent SNPs being close to one another, and increasingly smaller proportions as distance becomes larger. Figure 4.6 shows the sizes of the 100-SNP test regions in Kb. Changes in density of the SNPs on chromosome 4 would cause the test region size to vary, which would make comparisons across sQTL less meaningful. In general the window sizes are fairly uniform, except for a large increase at around the 5000th sQTL. This is caused by a 3.05Mb gap between SNPs 5005 and 5006, and represents the location of

the centromere on chromosome 4. There are exactly 100 windows comprising this spike on the graph, as is consistent with a single large gap. The mean test region size is 1.12Mb while including the 100 centromeric test regions, and 1.10Mb without. It also appears that the sizes of the final thousand windows or so are slightly smaller. This is presumably because LD between neighbouring SNPs is relatively low for this region and therefore more are required for the appropriate level of coverage, hence decreasing the distance between the sets of 100 SNPs. It should be noted that the difference for these windows is very small however.

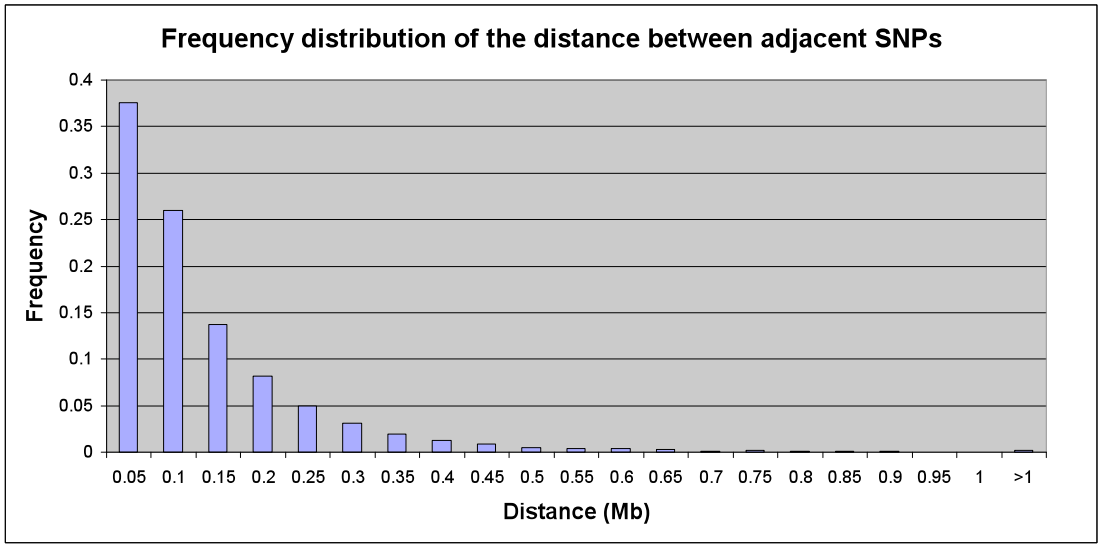


Figure 4.5 Distribution of distances between each adjacent pair of SNPs (in a 3' to 5' direction) used in the analysis. Groups are separated by 50,000bp (0.05Mb). Distances over 1Mb are grouped into a single group.

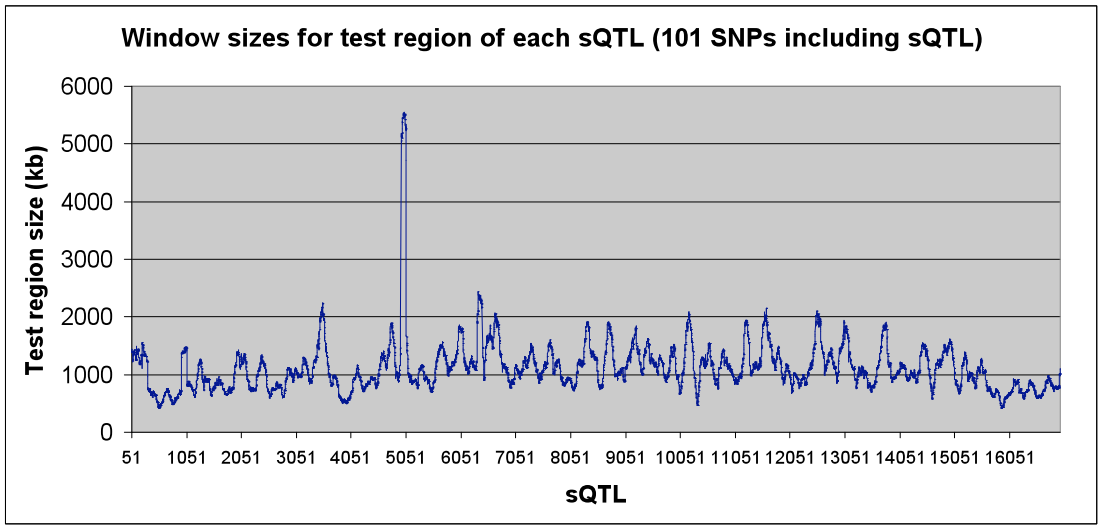


Figure 4.6 Test region sizes in kilobases for each of the 16922 sQTL.

Table 4.1 shows the mean, minimum and maximum test region sizes for sQTL separated into MAF bins. Interestingly, the mean test region size for the lowest MAF bin is higher than the rest of the means, and to a lesser extent so is the mean for the 0.05 – 0.1 MAF bin. This phenomenon may again reflect the way SNPs were selected for the Illumina panel. Adjacent SNPs often have similar allele frequencies, and for these particular SNPs, MAF is very small. Low MAF SNPs were not selected by Illumina, therefore there is more likely to be a less densely packed region on the SNP panel near SNPs exhibiting low MAF. Low MAF SNPs present on the panel may just happen to have low MAF in this specific population, yet be above the cut-off MAF for the panel in the original discovery dataset. The pattern seen in Figure 4.6 is also produced by plotting the 11-SNP haplotype windows used in PHASE. Once again there is a spike corresponding to the centromere, and as expected there are ten windows comprising this spike (graph not shown). The average window size is 112.26Kb including the ten centromeric windows and 110.12Kb not including them.

MAF BIN	MEAN	MINIMUM	MAXIMUM
0 – 0.05 (bin 1)	1188041	446989	5523241
0.05 – 0.1 (bin 2)	1147076	415616	5474898
0.1 – 0.15 (bin 3)	1134994	423244	5527773
0.15 – 0.2 (bin 4)	1125144	416672	5544199
0.2 – 0.25 (bin 5)	1112178	415275	5523025
0.25 – 0.3 (bin 6)	1103402	416916	5513875
0.3 – 0.35 (bin 7)	1113161	426085	5521348
0.35 – 0.4 (bin 8)	1084275	415127	5494813
0.4 – 0.45 (bin 9)	1124262	429845	5512279
0.45 – 0.5 (bin 10)	1119036	412296	5501764

Table 4.1 Mean, minimum and maximum test region size for sQTL within the ten MAF bins. Bin numbers in parentheses denote how MAF bins are referred to in the text.

Another important aspect to consider is the pairwise LD between SNPs. With 17,022 SNPs, the full pairwise matrix would be extremely computationally demanding to calculate. Instead, overlapping windows of 5,000 SNPs were analysed, moving on 3,000 SNPs each time. This meant there were at least 5,000 pairwise estimates for any given SNP, and at most there were 8,000. The distribution of pairwise r^2 estimates between all adjacent markers is shown in Figure 4.7. There is a large proportion of SNPs with low (0 - 0.05) r^2 to adjacent SNPs, almost a quarter of all SNPs. This was expected as Illumina explicitly aims to select independent SNPs where possible. Interestingly, there are also a reasonable proportion of adjacent SNPs in high LD (15% have r^2 of at least 0.7), and over 5% have an r^2 of one. This may be a consequence of the fact that the sample size is relatively low, and that the data are from an isolated population where LD is expected to be higher.

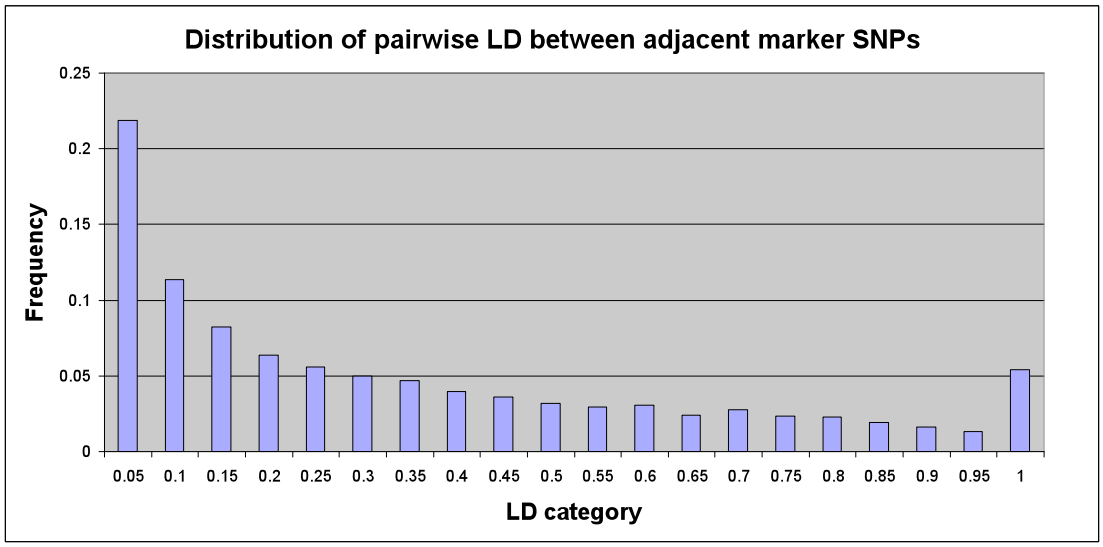


Figure 4.7 Distribution of pairwise LD between adjacent marker SNPs (in a 3' to 5' direction), separated by 0.05-wide LD bins.

Pairwise r^2 estimates were also used to calculate the mean sQTL-marker LD for each SNP position within the test region, and this was performed separately for each MAF bin. Averaging was over the marker-sQTL LD estimate for each sQTL within a MAF bin at that given position in the test region. A graph of this is shown in Figure 4.8. LD is greatest between SNPs which are closest to the sQTL, as would be expected, and tails off as distance increases. At either end of the test region all MAF bins have converged at a base level of average LD, of around 0.02. The differences in the peaks of LD for the lowest four MAF categories are quite pronounced, however the remainder of the MAF bins are fairly clustered. Even so, they follow the same general pattern that sQTL with larger MAF have greater average pairwise LD with SNPs at a given position within the test region. The maximum average level of LD for the highest MAF bin was 0.40, but just 0.13 for the lowest MAF bin.

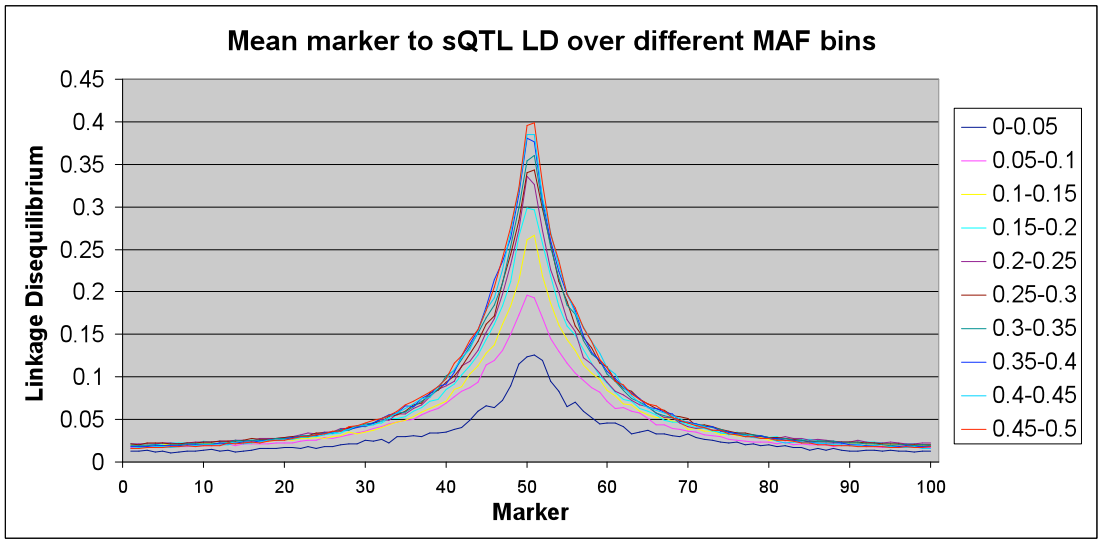


Figure 4.8 Mean pairwise LD between each SNP position and the sQTL, for each sQTL within a given MAF bin. All ten MAF bins are shown. The sQTL is located between the 50th and 51st SNPs.

4.3.2 Single SNP regression results

Figure 4.9.1 shows the mean proportion of sQTL variance explained by each SNP position within the test region, for each of the separate MAF bins. Along the x-axis is the position within the test region of the tested SNP, and the y-axis shows the mean proportion of variance explained. For all MAF bins, the SNPs closest to the sQTL position (i.e., the one on each side) explain on average the highest proportion of variance. For the lowest MAF bin the mean proportion of variance explained reaches just over 0.03 at SNPs adjacent to the sQTL. For the next two MAF bins the proportion of variance explained increases by around 0.017 each time, then the increase between neighbouring MAF bins tapers off. MAF bin 0.45 - 0.5 has the highest mean at just over 0.101 (10.1% of the total variance being explained). Note that while on average only 10.1% of the total variance is explained, this amounts to

just over 33% of the heritable variation, since each sQTL has a heritability of approximately 30%. As this is the average amount of variance explained, the maximum proportion of variance explained by SNPs in each bin would be higher. At only around 20 SNPs to either direction of the sQTL there is no appreciable difference in variance explained for different MAF of the sQTL.

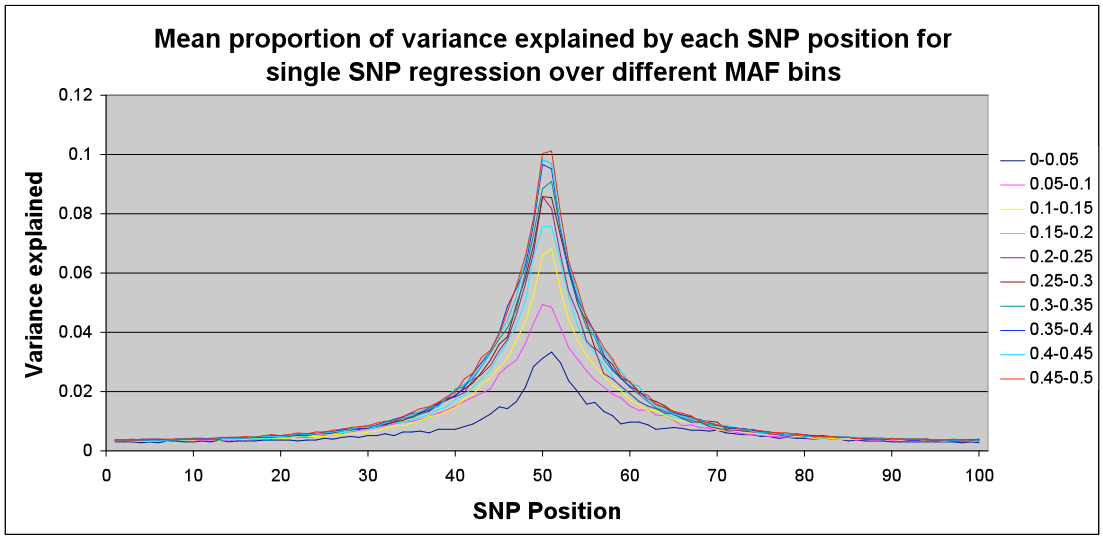


Figure 4.9.1. Mean proportion of variance explained by each SNP position for single SNP regression. Averages are calculated separately for different sQTL MAF bins. SNP position within the 100-SNP test region is represented on the x-axis, the sQTL position is between the 50th and 51st SNPs.

Results in Figure 4.9.2 show the percentage of times each SNP position explains the largest proportion of variance. As expected, the closest SNPs to the sQTL are more often found to explain most variation, and this amount is greatest when MAF is large. As MAF of the sQTL drops, there is a greater chance that the SNP explaining most variation will be located further away. Even so, except for the lowest two MAF bins, the chance that a SNP will explain most variation is very small from around ten SNPs in either direction of the sQTL. For the lowest MAF bin, the SNP explaining the highest proportion of variance is located more than ten SNPs away in either direction

from the sQTL 26.4% of the time, and for MAF bin 0.05 – 0.1 this value is just 13.2%.

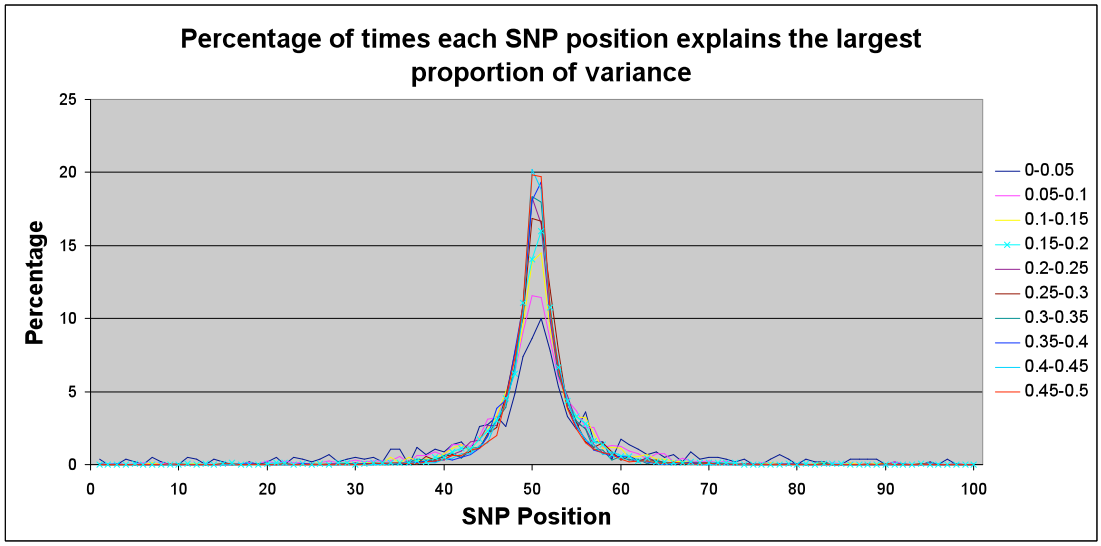


Figure 4.9.2. Percentage of times each SNP position explains the highest proportion of variance, calculated separately for different sQTL MAF bins. SNP position within the 100-SNP test region is represented on the x-axis, the sQTL position is between the 50th and 51st SNPs.

Figure 4.9.3 shows the mean $-\log_{10}$ p-values for each SNP position in the test region, once again separated by MAF bin. As in Figures 4.9.1 and 4.9.2, the MAF bins are ordered such that sQTL with the lowest MAF perform worst (i.e., have low $-\log_{10}$ p-values), and sQTL with the highest MAF are performing best (have high $-\log_{10}$ p-values). For the highest MAF bin, the most significant SNPs reach a $-\log_{10}$ p-value just under 12, and for the lowest MAF bin they are just over four. There is little separating the top few MAF bins; just 0.5 between the top six MAF bins, whereas there is almost 1.5 $-\log_{10}$ p-value units between each pair of adjacent MAF bins from the first to the fourth.

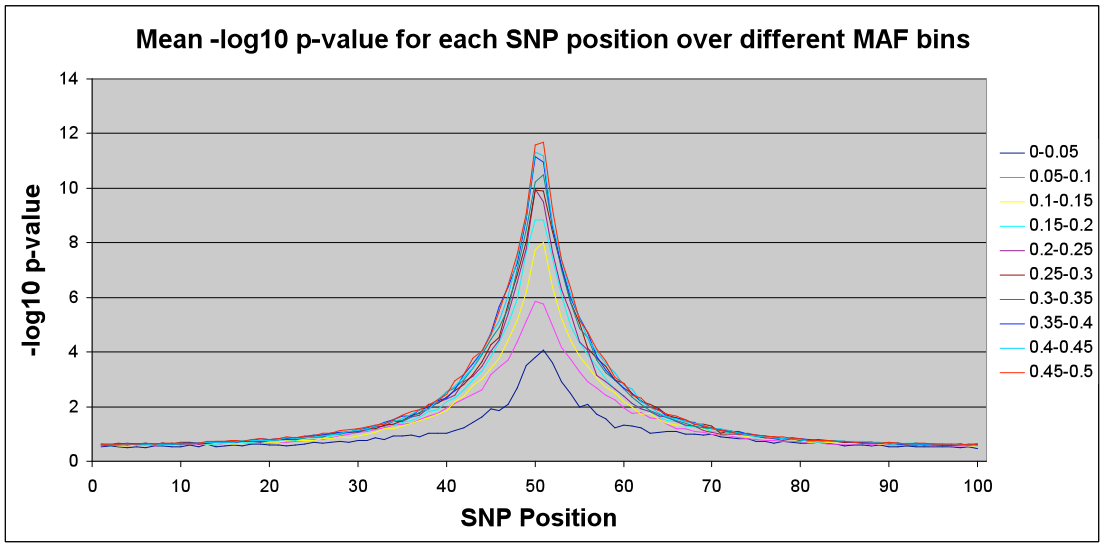


Figure 4.9.3 Mean $-\log_{10}$ p-value for each SNP position for single SNP regression. Averages are calculated separately for different sQTL MAF bins. SNP position within the 100-SNP test region is represented on the x-axis, the sQTL position is between the 50th and 51st SNPs.

4.3.3 Multiple regression results

4.3.3.1 Overall method comparison

In real GWA studies, the frequency of a causative QTL is unknown, therefore it may be instructive to compare overall results from each method averaged over all MAF bins, to see on average how well each method performs. Results from single SNP regression are also shown in this comparison. Figure 4.10.1 compares the overall mean variance explained by each SNP position for each of the four methods. Where the method involves multiple SNPs, the graph is aligned such that the middle SNP in each regression window is constant.

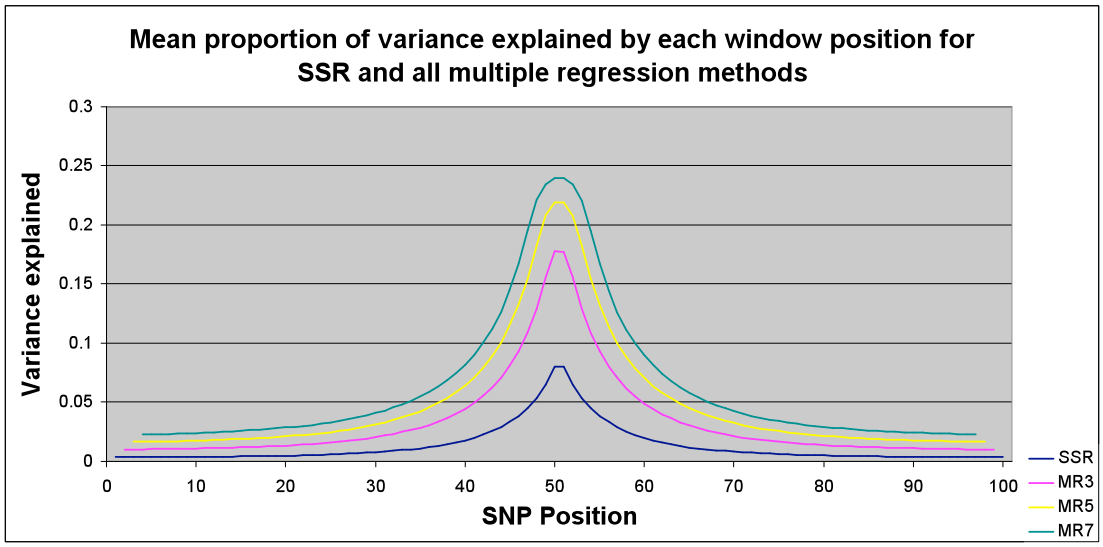


Figure 4.10.1 Mean proportion of variance explained over all MAF bins by each window position for single SNP regression and multiple regression using three, five and seven SNPs. Points are centred on the middle SNP of the regression window for the multi-SNP methods.

In all instances the SNP or group of SNPs located closest to the sQTL explain the highest proportion of variance for a given method. Since the number of SNPs in a window is always odd, the sQTL is never positioned centrally, therefore just as with SSR there are always two windows which are expected to do equally well, and better than all others (one to either side of the sQTL). Looking at Figure 4.10.1 this does appear to be the case. The averages for the methods peak at around 0.08, 0.18, 0.22 and 0.24 respectively for SSR, MR3 MR5 and MR7. These values correspond to approximately 26.7%, 60%, 73.3% and 80% of the sQTL variation. There is clearly an increase in the proportion of sQTL variance explained that is attributable to increasing the number of SNPs in the model. There is an initial fairly large step up in the proportion of variance explained by going from SSR to MR3, then the increases become smaller. Indeed, with each subsequent increase in the number of SNPs included, on average the amount of extra variance explained decreases.

The comparison in Figure 4.10.2 is of the overall percentage of times each window position explains the largest proportion of variance. The general trend is that as the number of SNPs in the window increases, there are more windows which regularly explain the most variance, therefore the peak of the curve is wider. Consequently, the peak for methods with larger numbers of SNPs is lower; MR3 tops out at around 18%, MR5 at just under 16% and MR7 at just under 13%. The SNPs most frequently explaining the highest proportion of variance for SSR do so approximately 16% of the time. Looking at the distributions as a whole also reveals a telling pattern. The cumulative frequency that a group of window positions explain the highest proportion of variance was calculated as distance from the sQTL was increased (by moving outwards one SNP at a time in each direction from the centre). This reveals that methods using larger numbers of SNPs are more precise; the number of SNPs / windows required in order to ensure a cumulative frequency of over 0.9 was 16 for SSR, but just 12 for each of MR3, MR5 and MR7.

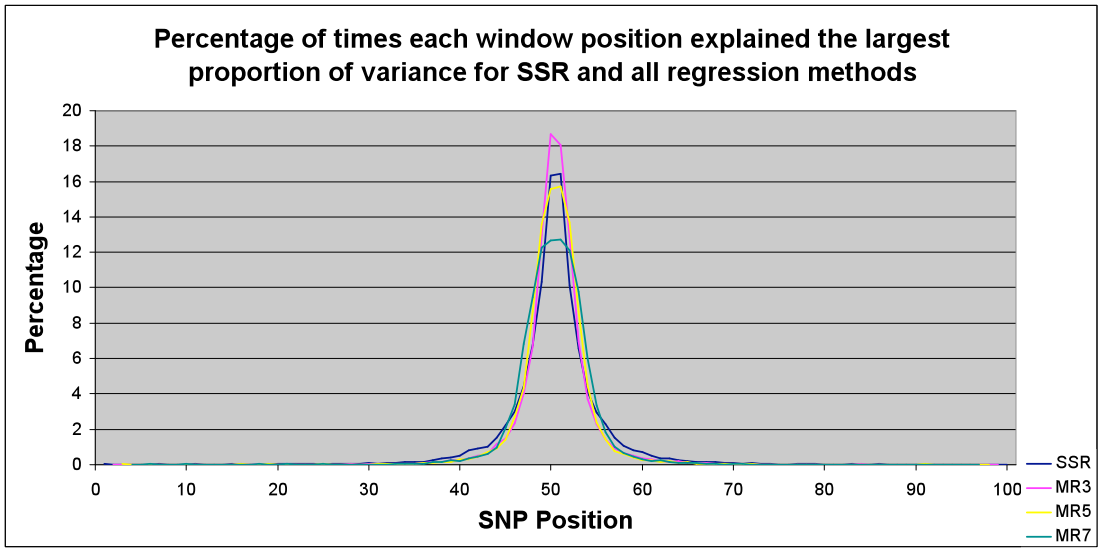


Figure 4.10.2 Percentage of times each window position explains the largest proportion of variance over all MAF bins for single SNP regression and multiple regression using three, five and seven SNPs. Points are centred on the middle SNP of the regression window for the multi-SNP methods.

The last overall comparison is of $-\log_{10}$ p-values, shown in Figure 4.10.3. There are considerable differences between the methods, the greatest of these being between single SNP regression and the remaining six. Each successive method performs better than the previous one for windows closest to the sQTL – further away from the sQTL all methods perform equally. The overall means at the sQTL-adjacent SNPs are 9.33, 18.6, 21.7 and 22.4 respectively for SSR, MR3, MR5 and MR7. As previously noted with the mean proportion of variance explained, the increase from five to seven SNPs is not as great as that from three to five SNPs, which itself is not as great as the increase for three SNPs over a single SNP. The mean $-\log_{10}$ p-value centred at SNPs 50 and 51 of the test region increases by 140% from SSR to MR7. At the tails of the graph, the $-\log_{10}$ p-values converge to approximately the same value. This is in contrast to Figure 4.10.1 (mean variance explained), where even at the tails there are consistent differences between methods.

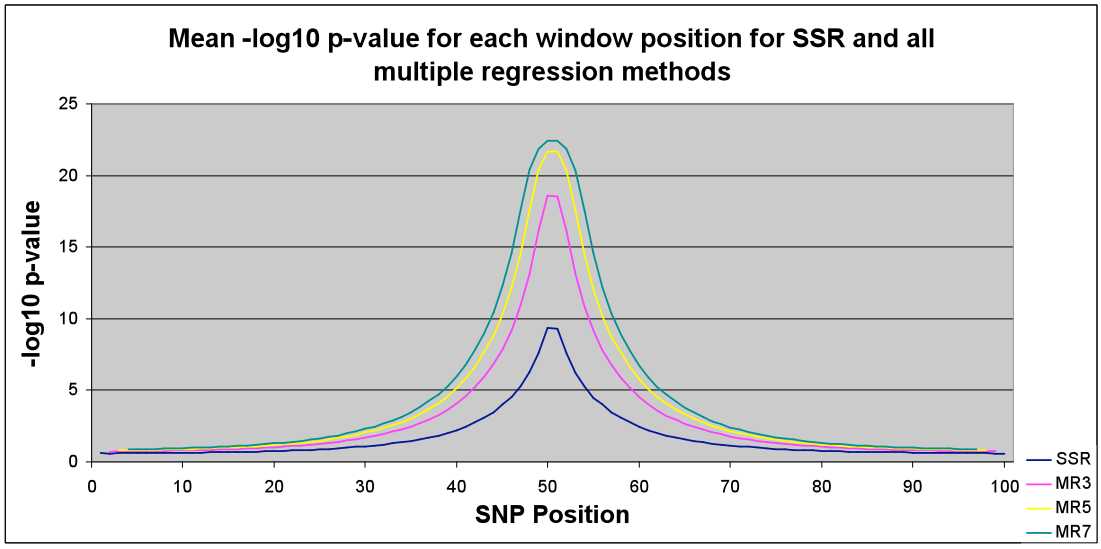


Figure 4.10.3 Graph showing the mean $-\log_{10}$ p-value over all MAF bins for each window position for single SNP regression and multiple regression using three, five and seven SNPs. For the multiple regression methods, the points are centred on the middle SNP of the regression window.

4.3.3.2 Proportion of variance explained

For the remainder of the multiple regression method results, comparisons are made across MAF bins. Results for the proportion of variance explained by MR3 are shown in Figure 4.11.1. As with the equivalent single SNP regression results, there are clear distinctions between MAF bins. Once again, the pattern is such that as the sQTL MAF increases, a larger proportion of variance is explained. There is a noticeable increase in the amount of variance explained between each of the first four sQTL MAF bins, but only a minimal increase thereafter.

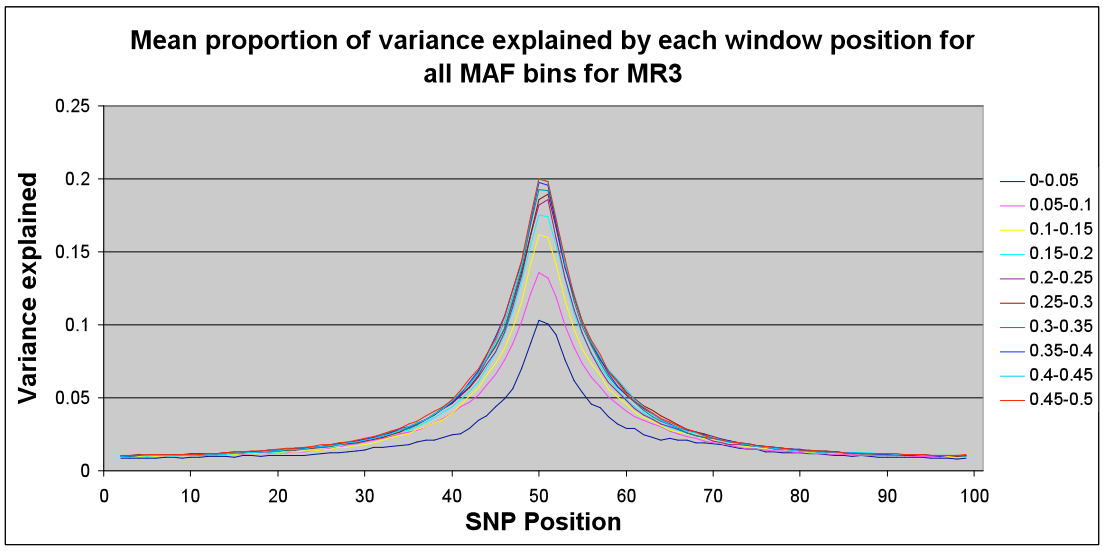


Figure 4.11.1 Mean proportion of variance explained by each window position for all MAF bins for three-SNP multiple regression. Points are centred on the middle SNP of the regression window.

As expected, the peak for the sQTL MAF bin explaining most variance (the 0.45 - 0.5 MAF bin) is higher than the overall peak (i.e., averaged over all MAF bins) for MR3 seen in Figure 4.10.1. The peak here is 0.20, which is 11% higher than the 0.18 corresponding to the overall mean. The overall mean has been dragged down by the

poor ability of windows to explain variation for sQTL with low MAF. The lowest MAF bin at most only explains just over 10% of the total variation (and 33% of heritable variation). The amount of variance explained by all MAF bins converges at around 20 SNPs in either direction from the sQTL, therefore regardless of the minor allele frequency of the sQTL, at this distance (or further away) from the tested SNP it will be almost impossible to detect since the proportion of variance explained is so low.

Figure 4.11.2 displays the proportion of variance explained for MR5 over the different MAF bins. Again, the best results (on this criterion) belong to the two SNPs closest to the sQTL, although these are not in the highest MAF bin this time, but the 0.4 - 0.45 bin instead. The peak for these windows is 0.24, which is 0.02 higher than the overall mean for this window length, and also 0.04 higher than the analogous point in the MR3 results.

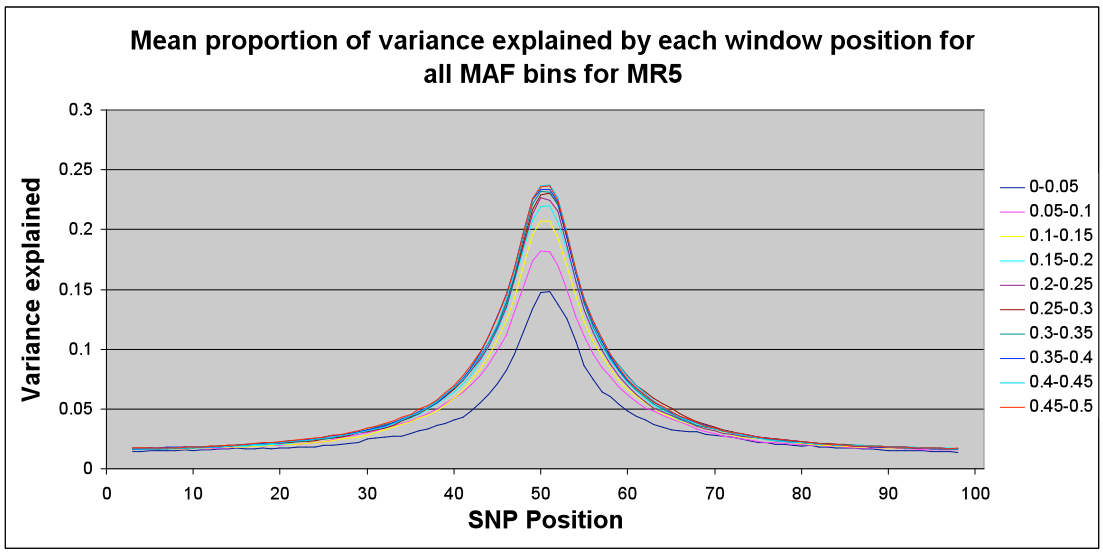


Figure 4.11.2 Mean proportion of variance explained by each window position for all MAF bins for five-SNP multiple regression. Points are centred on the middle SNP of the regression window.

Again the lowest MAF bin does worst, peaking at 0.15. This is only a 0.05 increase over the 0.1 for MR3. This makes it appear that the lowest MAF bin has experienced a smaller increase in the amount of variance explained than the highest MAF bin compared with the MR3 method. In actual fact however, the proportional increases are 19% for the highest MAF bin and 44% for the lowest, suggesting that the lowest MAF bin benefits more.

Results for MR7 are shown in Figure 4.11.3. The 0.40 – 0.45 MAF category marginally produces the highest mean proportion of variance explained by any of the MAF bins. The highest value is 0.26, a 0.02 increase from the overall MR7 average, representing 85% of the total heritable variation. The most that the lowest MAF bin explains is just under 58% of the variance of the sQTL. The peaks of the curves are wider than is found for either of the other two multiple regression methods, or the single SNP regression method. This appears to be a gradual increase associated with having a larger number of SNPs in the model.

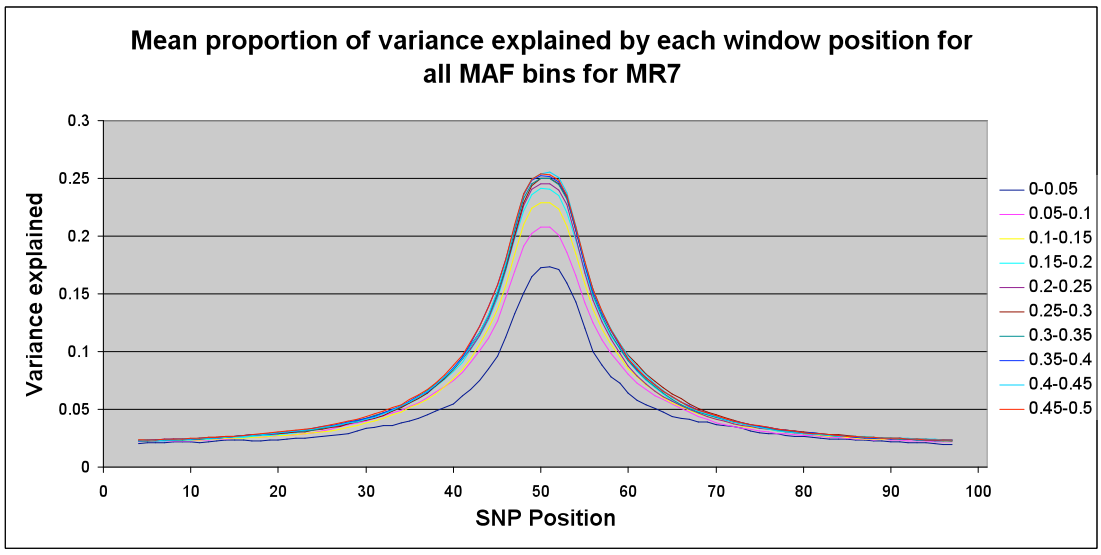
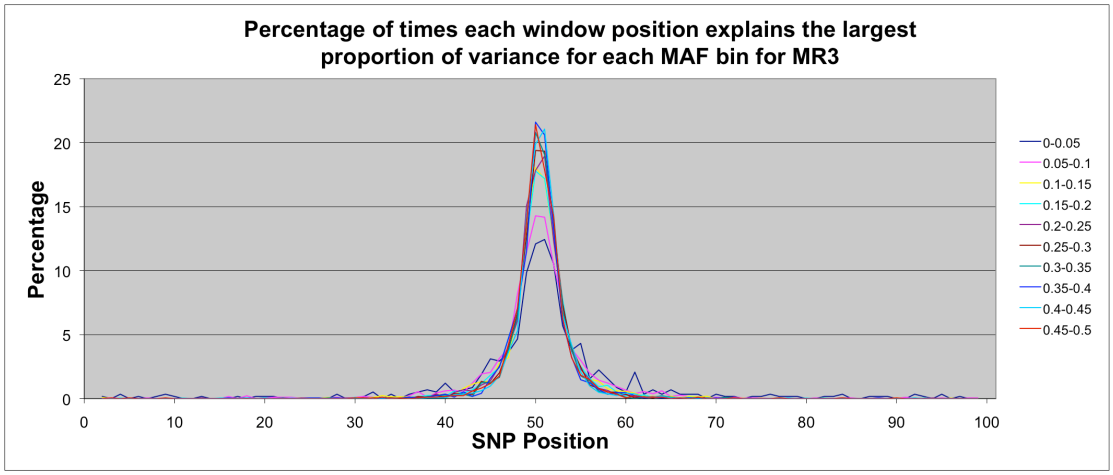


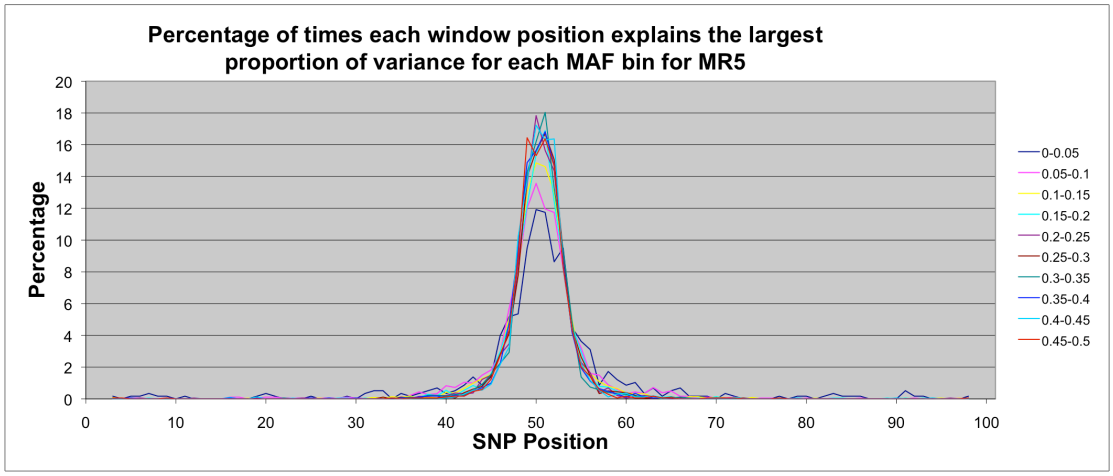
Figure 4.11.3 Mean proportion of variance explained by each window position for all MAF bins for seven-SNP multiple regression. Points are centred on the middle SNP of the regression window.

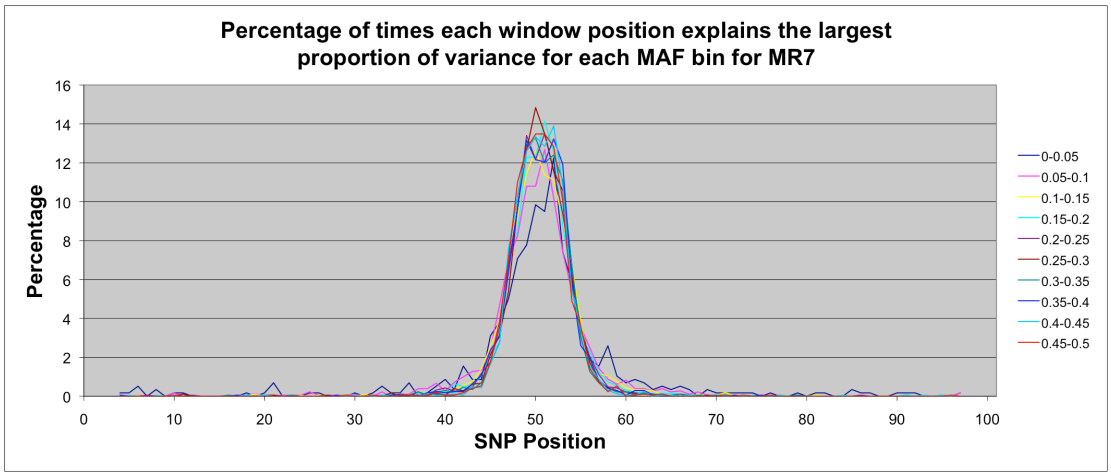
The differences between the best results for this criterion for MR7 compared to those for SSR, MR3 and MR5 are 0.152, 0.053 and 0.017 respectively. This means that the increase in the proportion of variance explained over the differing window sizes for the highest MAF bin changes from 0.099 for SSR to MR3 (98% increase), to 0.036 for MR3 to MR5 (18% increase), and to 0.017 for MR5 to MR7 (7% increase). For sQTL with a MAF of 0 – 0.05, the differences are 0.070 from SSR to MR3 (210% increase), 0.045 for MR3 to MR5 (44% increase), and 0.025 from MR5 to MR7 (17% increase). Thus, the benefit of using more SNPs is greater the smaller the number of SNPs there is to begin with. While the absolute gain from using more markers in the regression is greater for sQTL with a higher MAF (0.140 total increase in variance explained for lowest MAF bin compared to 0.152 for the highest), proportionally the increase is much better for the low MAF bins. The total increase for the lowest MAF bin over the four methods corresponds to a 422% increase, whereas for the highest MAF bin the increase is only 150%.

4.3.3.3 Most predictive window

Figures 4.12.1, 4.12.2 and 4.12.3 show, for each of the three multiple SNP regression methods, the percentage of times that each window position explains the largest proportion of variance for each sQTL MAF bin. These all show the familiar pattern (from the overall comparison) of curves becoming wider and lower as the number of SNPs in the model increases. Consequently, there is also a difference in how clustered the lines representing different MAF bins are across the three graphs.





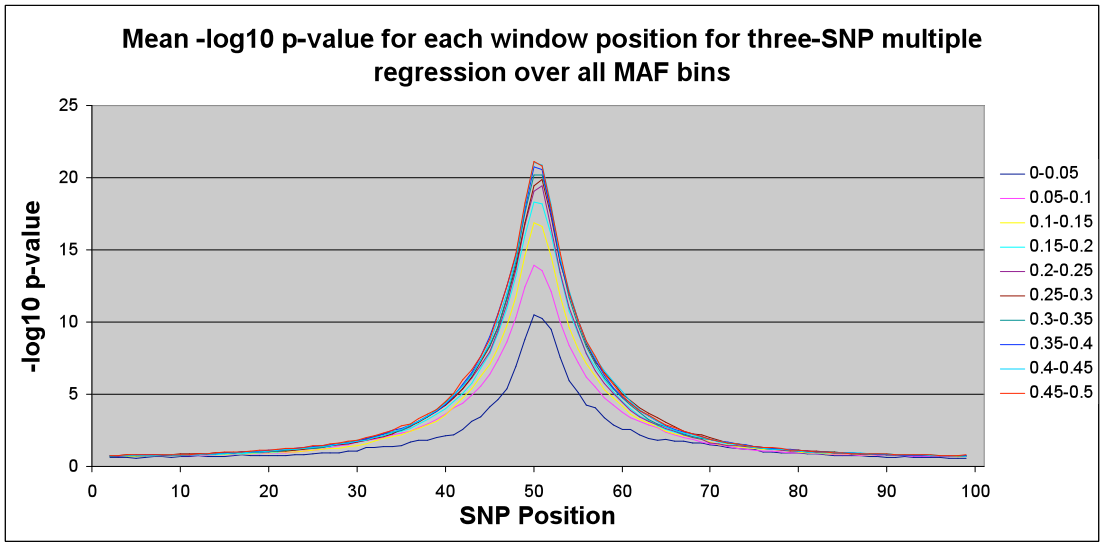


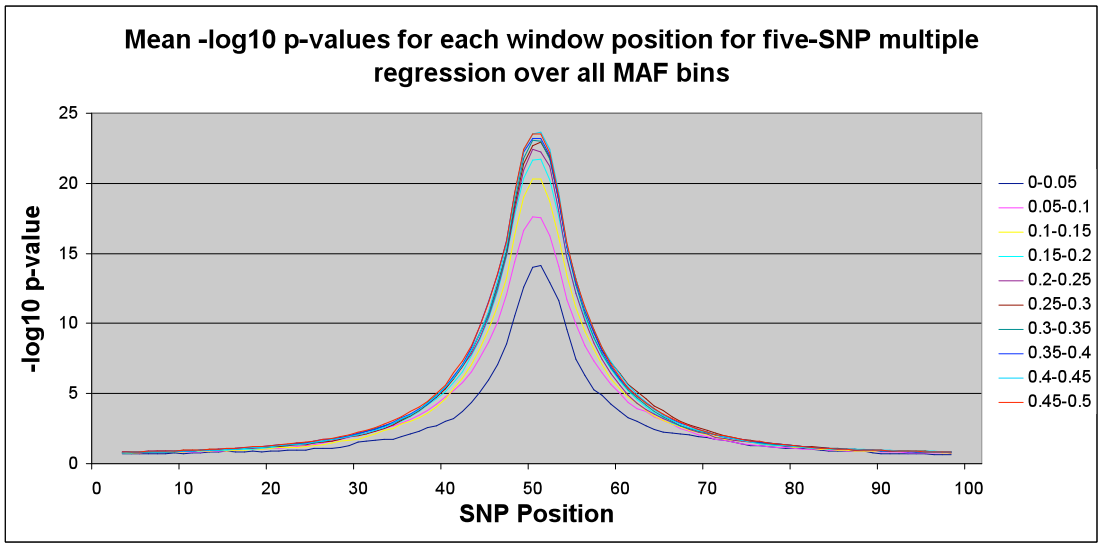
Figures 4.12.1 - 4.12.3 Percentage of times each SNP position explains the highest proportion of variance for each MAF bin for three-, five- and seven-SNP multiple regression. Points are centred on the middle SNP of the regression window.

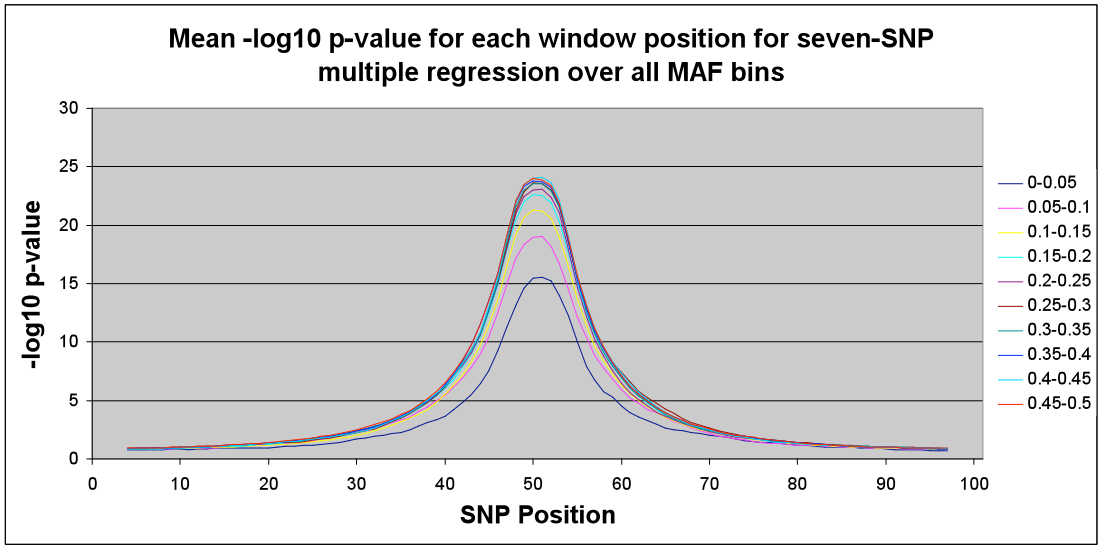
With MR7, there is little change over different MAF compared with the MR3 results. In addition to the curves becoming lower as the number of SNPs increases, the low MAF bins have a wider spread (and consequently lower peak) than the higher MAF bins, as seen for SSR (Figure 4.9.2). This again implies that sQTL with lower MAF are more frequently explained best by windows other than those adjacent to them than sQTL with high MAF. For MR3 (Figure 4.12.1) the highest percentage a window position has is 21.6%; this drops to 18.0% for five SNPs, and then to 14.8% for MR7. For none of these methods is the 0.45 – 0.5 MAF bin the one with the highest percentage, although differences between results for any of the top six MAF bins are minimal.

4.3.3.4 P-values

The $-\log_{10}$ p-values for MR3 are shown in Figure 4.13.1, and those for MR5 and MR7 are in Figures 4.13.2 and 4.13.3. The three common observations from all results so far also hold for these results; as window length increases the performance improves (in this case $-\log_{10}$ p-value increases), as window length increases more windows within the test region perform better (i.e., the curves become wider), and the low MAF bins perform worse than high MAF bins, although generally little separates the top few.







Figures 4.13.1 - 4.13.3 Mean $-\log_{10}$ p-value of each window position for each MAF bin, for multiple regression of three, five and seven SNPs. Points are centred on the middle SNP of the regression window.

For MAF bin 0 – 0.05 the peaks for MR3, MR5 and MR7 are 10.5, 14.1 and 15.5 respectively. This represents an increase of 48% in mean $-\log_{10}$ p-value from MR3 to MR7. The three to five SNP and five to seven SNP jumps represent 34% and 10% increases. Again, there is little difference between the top MAF bins for any of the methods, and these differences diminish as the number of SNPs in the model increases. The maximum $-\log_{10}$ p-value for each method are all from the 0.40 – 0.45 or 0.45 – 0.5 MAF bin, and the values are 21.1, 23.6 and 24.1 respectively. The overall increase is 13%, and the increases between neighbouring methods are 12% and 1% respectively. The increases are therefore larger for three to five rather than five to seven SNPs, and there is a greater benefit for sQTL in lower MAF bins.

The larger increase for lower MAF bins is also reflected in the fact that the height of the lowest MAF bin peak proportional to the highest result increases as window size increases, i.e., the $-\log_{10}$ p-values for the top MAF bins and the 0 – 0.05 MAF bin are proportionally closer together for larger window sizes. For SSR the lowest MAF bin peak was 35% of the maximum, and this increased to 50%, 60% and 64% of the maximum value for MR3, MR5 and MR7 respectively.

In summary, all multiple regression methods on average have greater power than SSR (see for example, Figure 4.10.3), and the amount by which they are superior varies depending on both the number of SNPs in the model, and the minor allele frequency of the sQTL. Methods with a larger number of SNPs are more powerful on average than those with fewer SNPs, although the difference between models with higher numbers of SNPs (i.e., MR5 and MR7) and those with lower numbers of SNPs (i.e.,

SSR and MR3) are much smaller. For all method comparisons, the greatest increase in power afforded by a superior model is for the lowest MAF bin. Even so, all results show that methods have greatest power when the MAF is large (see for example Figure 4.13.1).

4.3.4 Haplotype analysis results

4.3.4.1 Overall comparison

The overall mean proportion of variance explained by each window position for the three haplotype methods is shown in Figure 4.14.1. Note again that haplotypes were phased with 11 SNPs, and the relevant *n*-SNP haplotypes for each individual were then extracted, therefore there are only 90 windows within each test region for all haplotype methods. As before, each window on the graph is represented by its central SNP.

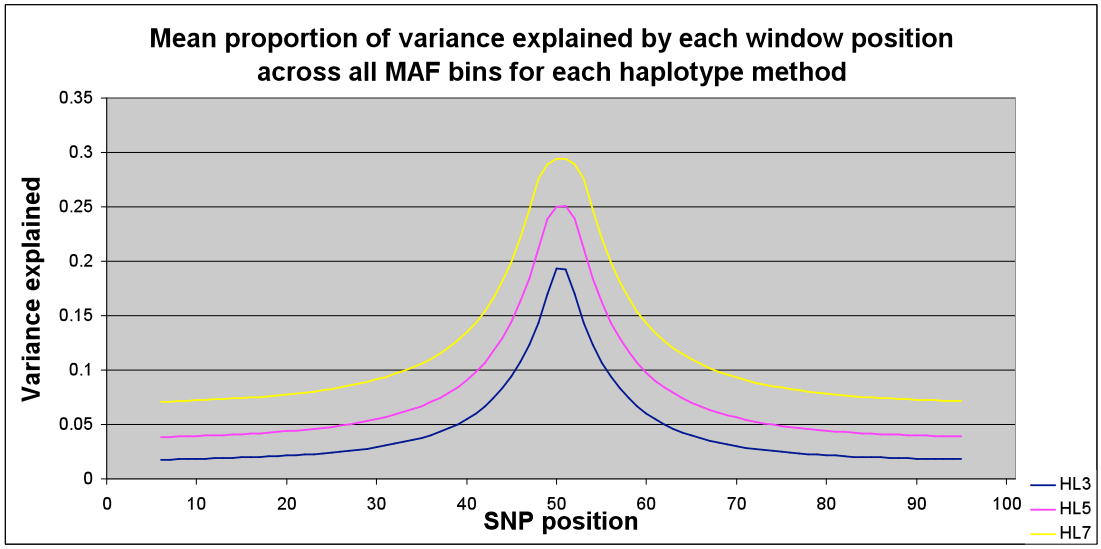


Figure 4.14.1 Mean proportion of variance explained for the three, five and seven-SNP haplotype methods over all MAF bins. Points are centred on the middle SNP of the regression window.

The two window positions expected to perform best have mean proportion of variance explained reaching 0.193, 0.250 and 0.294 for HL3, HL5 and HL7 respectively. This corresponds to on average explaining 64%, 83% and 98% of the total heritable variation, therefore methods with a higher number of SNPs once again do best. This means that over all methods, the amount of sQTL variation explained is lowest using SSR, increases with each multiple regression method (i.e., as the number of SNPs in the model increases, see Figure 4.10.1), and except for HL3, is highest for the haplotype methods. Both MR5 and MR7 explain more variance than HL3 on average, although HL3 explains more than MR3. The overall change in the amount of heritable variation explained by the best window is around 71% (~27% for SSR and 98% for HL7).

Figure 4.14.2 shows for each haplotype length, the window positions that explained the largest proportion of variance most frequently. The percentages for the window positions doing best (which also happened to be those closest to the sQTL position) were around 20%, 17.5% and 12.5% respectively. Compared to the multiple regression results, each peak is slightly higher, meaning that the windows closest to the sQTL explained most variance more often. Consequently, each curve is slightly narrower too. There is a much larger gap between these three methods than observed for the three multiple regression methods. As previously seen, as haplotype length increases so does the width of the curve, although the height decreases. This represents a greater number of windows near the sQTL position being best more often for haplotype methods with a larger number of SNPs.

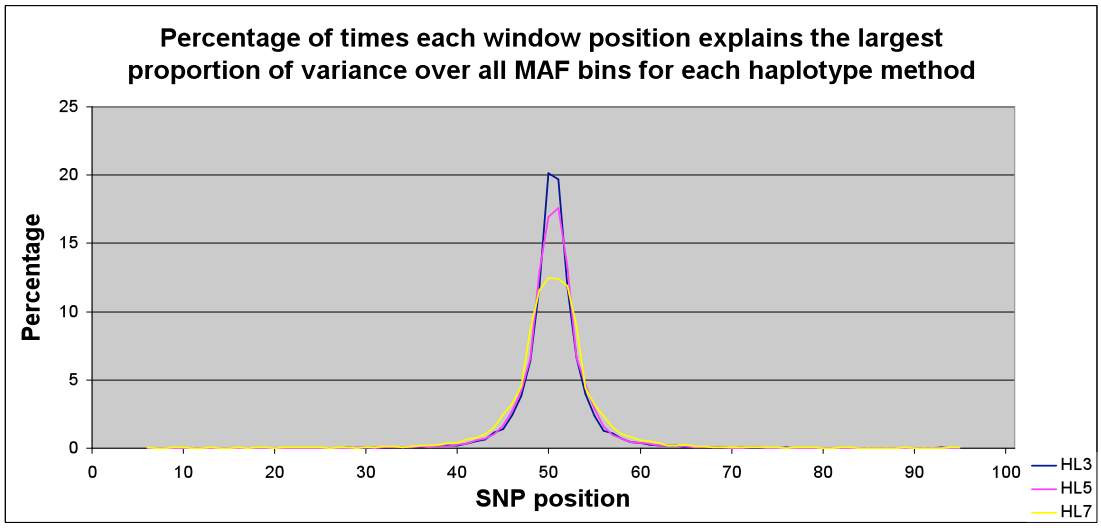


Figure 4.14.2 Percentage of times for haplotype method with three, five and seven SNPs that each window position explained the highest proportion of variance. Points are centred on the middle SNP of the regression window.

The overall mean $-\log_{10}$ p-values for each of the haplotype methods are displayed in Figure 4.14.3. Interestingly, this is the first instance where there hasn't been a direct trend of increasing $-\log_{10}$ p-value through the methods with increasing numbers of SNPs. The maximal overall mean $-\log_{10}$ p-values are from HL5, although HL7 still does better than HL3. The superiority of HL5 over HL7 occurs only for the central few windows however, and for the remainder HL7 is marginally superior. Maximum $-\log_{10}$ p-values for HL3, HL5 and HL7 respectively are 18.4, 20.2 and, 18.9, therefore HL7 is closer to HL3 than it is to HL5.

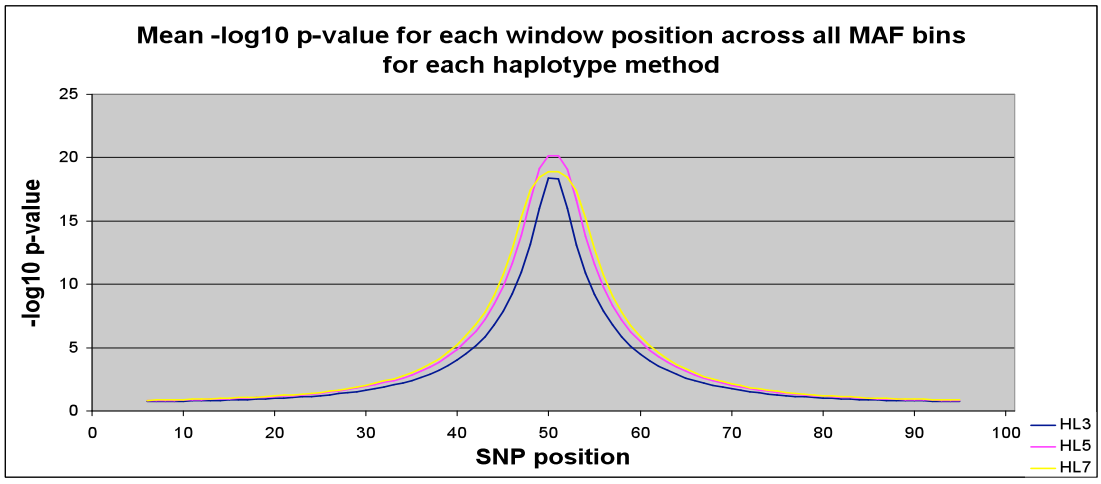
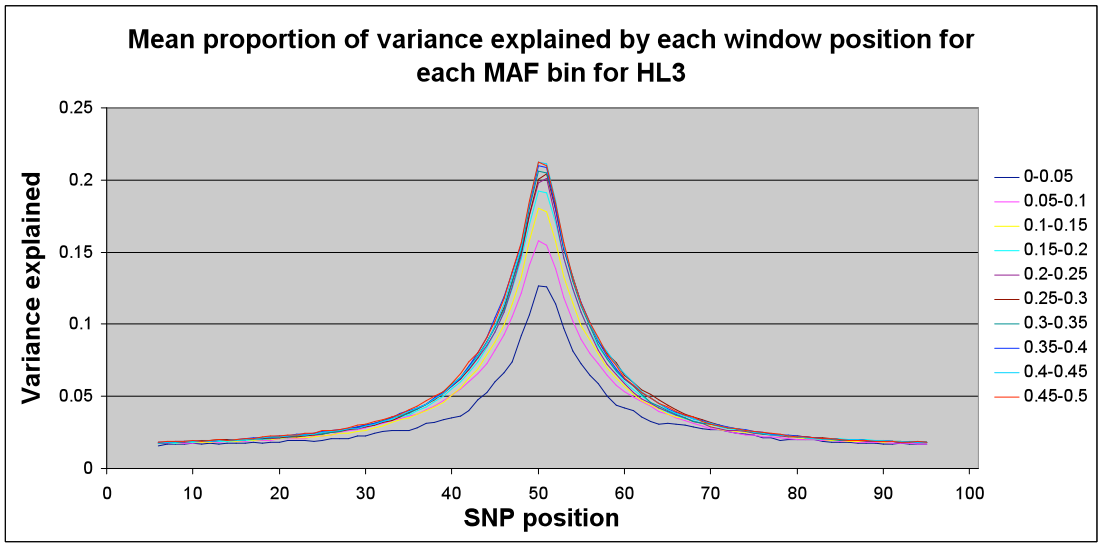


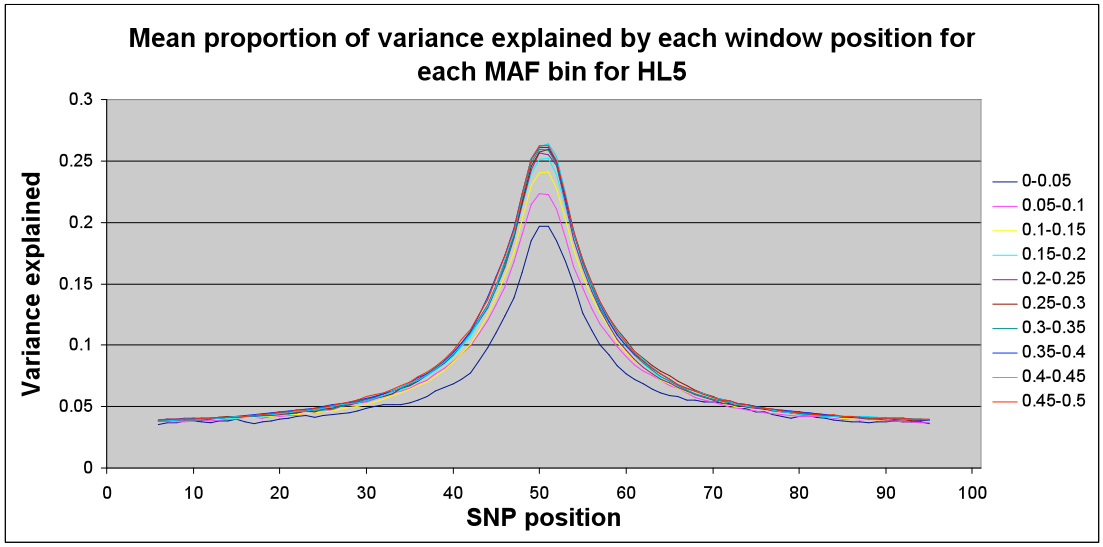
Figure 4.14.3 Mean $-\log_{10}$ p-value over all MAF bins for each window position for the three-, five- and seven-SNP haplotype methods. Points are centred on the middle SNP of the regression window.

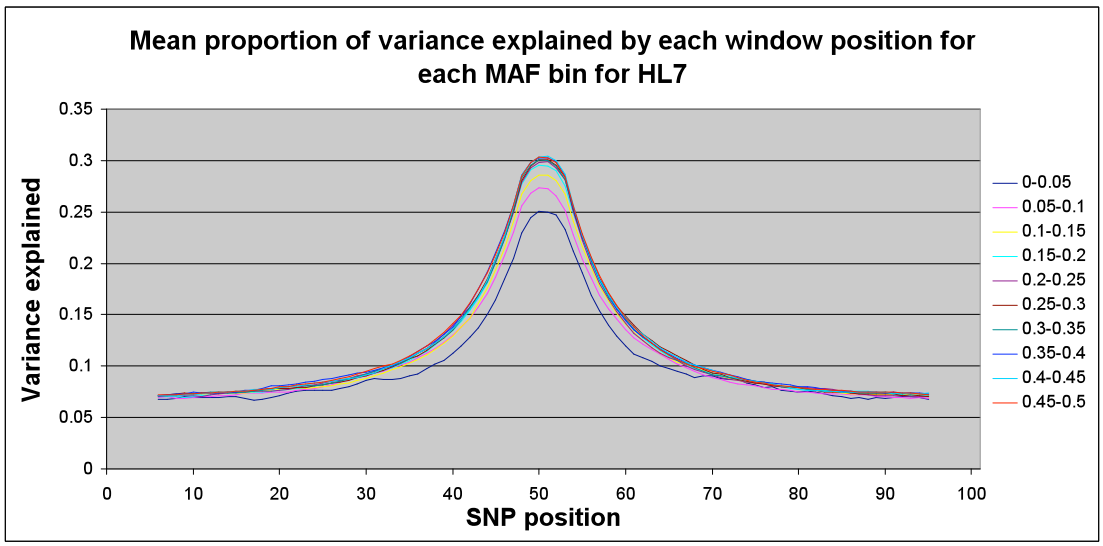
Regardless of the internal ordering, all haplotype methods represent a large increase over single SNP regression, but not over the multiple regression methods. In ascending $-\log_{10}$ p-value, the average values at the most significant SNP / window are 9.33 (SSR), 18.4 (HL3), 18.6 (MR3), 18.9 (HL7), 20.2 (HL5), 21.7 (MR5) and 22.4 (MR7). A haplotype method never does better on average than the method with equivalent number of SNPs in multiple regression. Both five-SNP methods do better than HL7. However, even the worst of the haplotype methods (HL3) has on average almost double the overall mean for SSR.

4.3.4.2 Proportion of variance explained

Results for the proportion of variance explained by HL3, HL5 and HL7 are shown in Figures 4.15.1, 4.15.2 and 4.15.3. There is an across-the-board increase in the proportion of variance explained for all MAF bins as the number of SNPs in the haplotype window increases. Within each method the highest results marginally belong to the 0.45 – 0.5 MAF bin, although there is no appreciable difference between the top five MAF bins. The maximum proportions of variance explained for the three methods respectively are 0.212, 0.264 and 0.304. On average the top five MAF bins for HL7 explain more heritable variance than is really present in the "trait", which is a consequence of the haplotype windows happening, by chance, to explain a very small proportion of variance which is not in fact heritable. HL7 explains almost all, if not all, of the heritable variation for sQTL with MAF above 0.25.







Figures 4.15.1 - 4.15.3 Mean proportion of variance explained by each window position for all MAF bins for the three-, five- and seven-SNP haplotype methods. Points are centred on the middle SNP of the regression window.

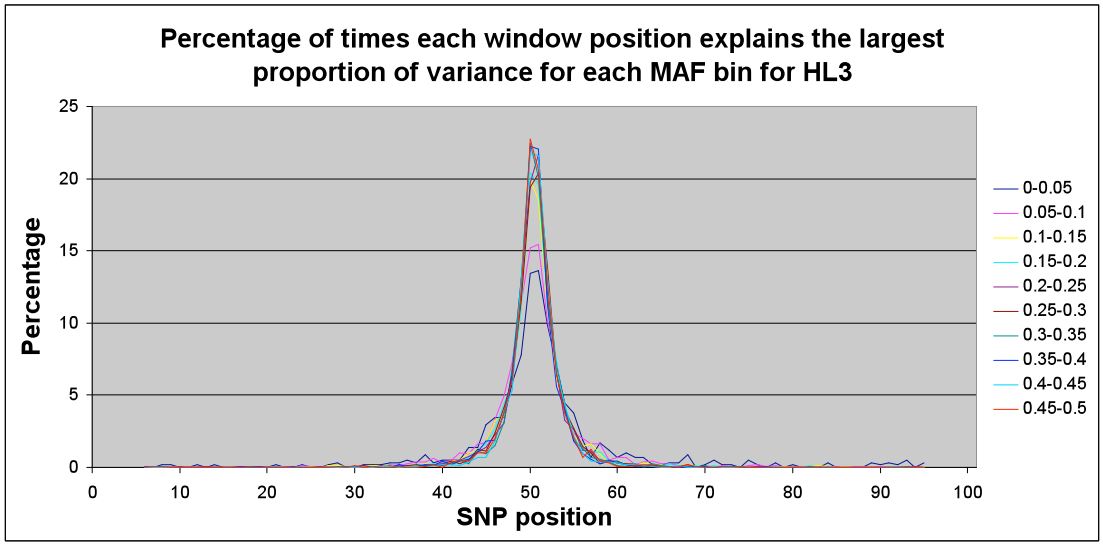
In comparison to multiple regression, the haplotype-based methods are able to explain more variance than the equivalent multiple regression method (i.e., the method with the same number of SNPs). All methods are far superior to single SNP regression. The best results for each method categorically show this (percentage of heritable variation shown in parentheses): 0.101 for SSR (34%), 0.200 for MR3 (67%), 0.236 for MR5 (79%), 0.255 for MR7 (85%), 0.212 for HL3 (71%), 0.264 for HL5 (88%) and 0.304 for HL7 (~100%). For the lowest sQTL MAF bin, the overall improvements look as follows; 0.033 for SSR (11%), 0.103 for MR3 (34%), 0.148 for MR5 (49%), 0.173 for MR7 (58%), 0.127 for HL3 (42%), 0.197 for HL5 (66%) and finally 0.251 for HL7 (84%). Once again, it is interesting that for the same number of SNPs, haplotype methods are able to explain much more variation than multiple regression methods. No method is able to explain as much variance for sQTL with low MAF as they do for high MAF, although incredibly, haplotype analysis using seven SNPs captures 84% of the heritable variance in a sQTL with a minor allele frequency under 0.05.

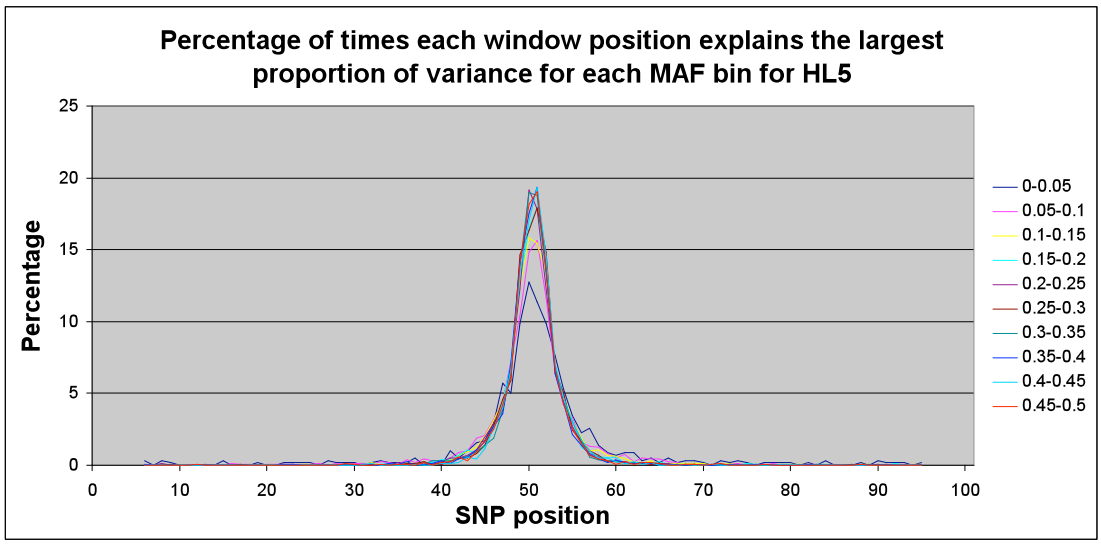
As seen with multiple regression it appears that the most beneficial transition between haplotype methods is from three to five SNPs. The percentage increase from HL3 to HL5 for the lowest MAF bin was 55.4%, and was 27.0% for HL5 to HL7. The increases for the MAF bin with the highest results were 23.8% for HL3 to HL5 and 15.3% for HL5 to HL7. Not only is the jump from three to five SNPs again better than five to seven, but also the increase is once again proportionally greater for the lower MAF bins (for both three to five, and five to seven SNPs). Again, one visible

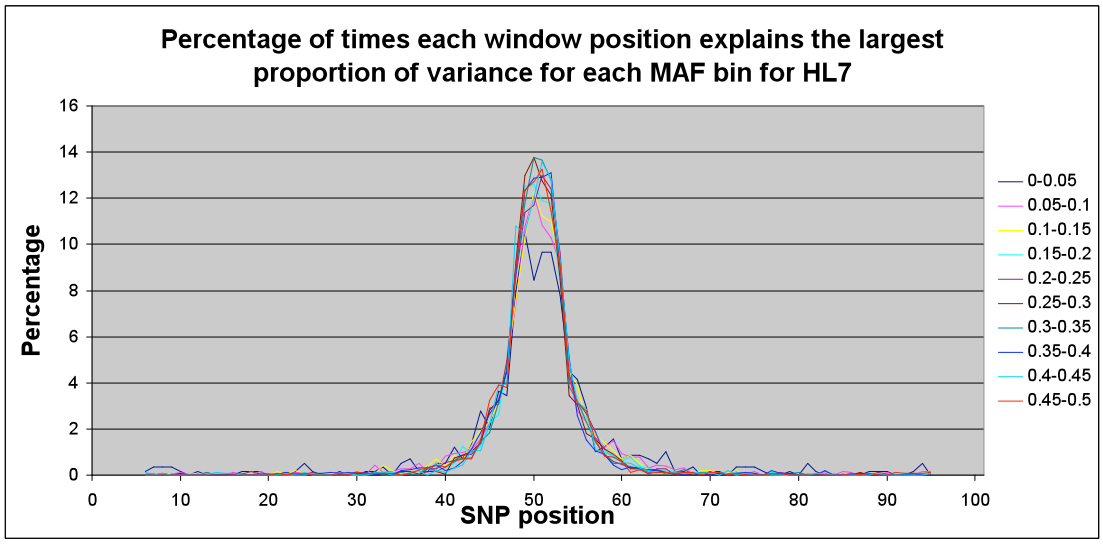
consequence of this in the figures is that the lines for different MAF bins become closer as the number of SNPs in the model increases (Figures 4.15.1 - 4.15.3).

4.3.4.3 Most predictive window

Figures 4.16.1, 4.16.2 and 4.16.3 display the percentage of times that each window does best at explaining sQTL variance. The overall patterns are the same as those for the equivalent graphs for multiple regression. The major difference between the two sets of graphs however is that in the three pertaining to haplotype analyses, the percentages at the peaks are all higher, meaning there is a greater proportion of times in which the window explaining most variation is located in close proximity to the sQTL. Consequently, the distribution for each MAF bin is also slightly narrower. For HL3, the closest two windows to the sQTL position for the highest sQTL MAF bin together explain the largest proportion of variance almost 45% of the time.







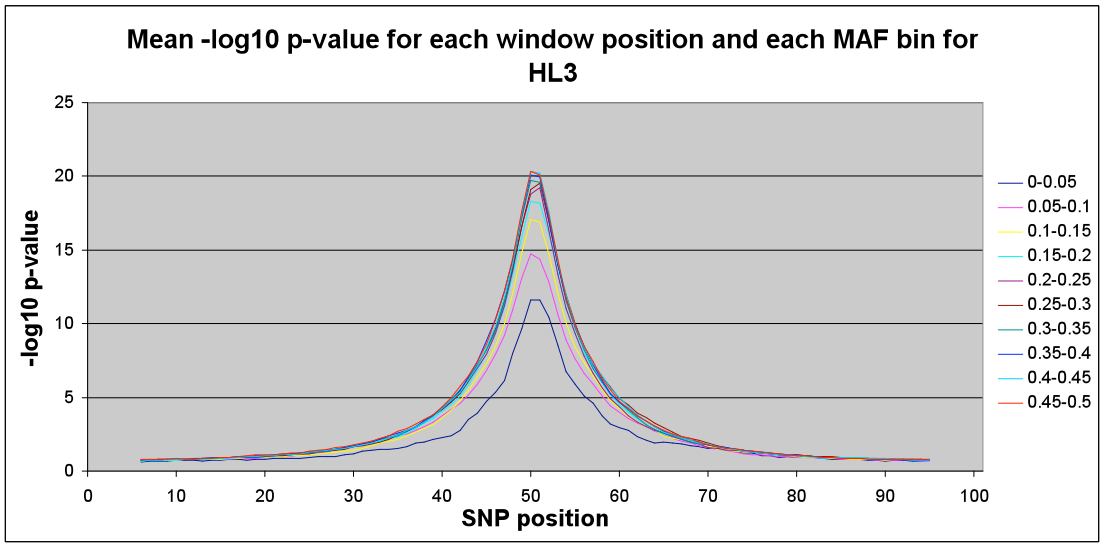
Figures 4.16.1 - 4.16.3 Percentage of times each window position explains the largest proportion of variance for all MAF bins for the three-, five- and seven-SNP haplotype methods. Points are centred on the middle SNP of the regression window.

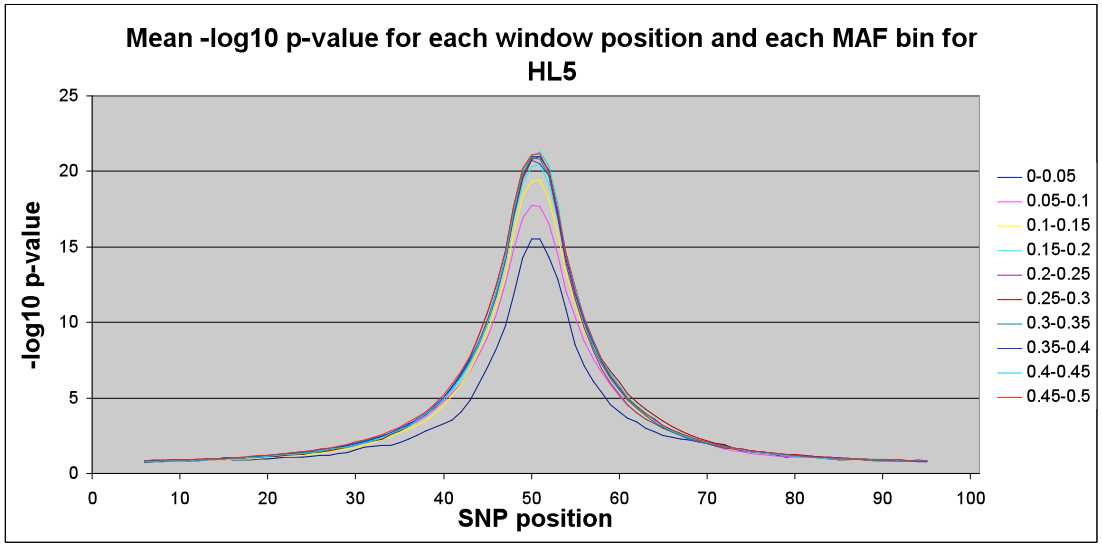
4.3.4.4 P-values

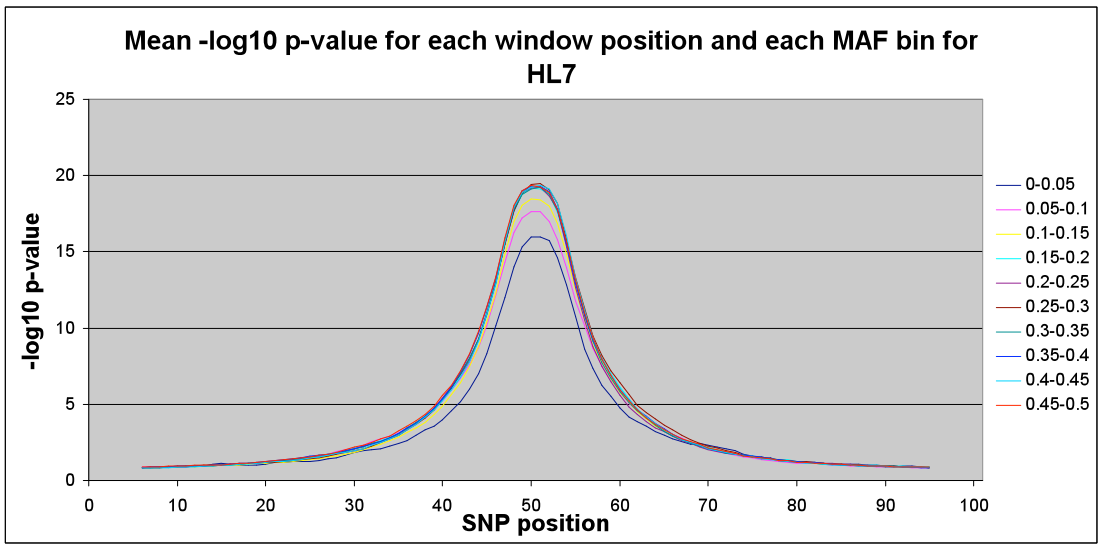
The MAF-separated mean $-\log_{10}$ p-values for the three haplotype methods are shown in Figures 4.17.1, 4.17.2 and 4.17.3. The maximum $-\log_{10}$ p-values for each of the haplotype lengths are 20.3, 21.3 and 19.5, and for MAF bin 0 – 0.05 they are 11.6,

15.5 and 16.0. The MAF bins producing the highest $-\log_{10}$ p-values were 0.45 – 0.5 for HL3, 0.4 – 0.45 for HL5 and 0.25 – 0.3 for HL7. This last one in particular illustrates how similar the results are for all sQTL with a minor allele frequency of at least 0.25. The most striking part of these results is that, as was suggested by the results averaged over all MAF bins, the trend for more SNPs to give better performance no longer holds. The only MAF bin this does still hold for is the lowest (i.e., 0 - 0.05) – the intermediate bins see HL7 falling gradually further behind HL5.

While it may no longer necessarily hold that the model with most SNPs in is superior, all haplotype methods are clearly far superior to SSR. The ordering of the haplotype and multiple regression methods is more complex however, and depends on the MAF of the sQTL. To make a consistent comparison instead of switching reference MAF bin, results from the highest MAF bin (0.45 – 0.5) are used despite not necessarily being the best $-\log_{10}$ p-values, since differences were very small in any case. The ordered maximum $-\log_{10}$ p-values are as follows; 11.7 (SSR), 19.2 (HL7), 20.3 (HL3), 21.1 (MR3), 21.2 (HL5), 23.5 (MR5) and 23.9 (MR7). The equivalent results for the lowest MAF bin were 4.06 (SSR), 10.5 (MR3), 11.6 (HL3), 14.1 (MR5), 15.53 (MR7), 15.54 (HL5) and 16.00 (HL7). For lower MAF, haplotype methods start outperforming multiple regression, whereas for high MAF the converse is true.







Figures 4.17.1 - 4.17.3 Mean $-\log_{10}$ p-value produced from each window position for each MAF bin for the three-, five- and seven-SNP haplotype methods. Points are centred on the middle SNP of the regression window.

When comparing different methods with the same number of SNPs (e.g., MR3 and HL3), the haplotype analysis results increased the $-\log_{10}$ p-values for sQTL with low MAF, but decreased them for sQTL with high MAF. For the 0 – 0.05 MAF bin, the increases in $-\log_{10}$ p-values of the haplotype methods amounted to 10.7% for the three-SNP methods, 9.9% for the five-SNP methods and 3.0% for the seven-SNP methods. For the highest MAF bin, the decreases in $-\log_{10}$ p-value were 3.6%, 10.1% and 19.6% respectively. Therefore, haplotypes improve analysis for sQTL with small MAF. Where minor allele frequency is moderate to large, the gain in $-\log_{10}$ p-value from haplotype analysis observed at lower MAF becomes a loss, and multiple regression does better. The numbers above also show there is a larger proportional increase (or smaller proportion decrease) in $-\log_{10}$ p-value for changing from multiple regression to haplotype analysis for methods with fewer SNPs.

4.3.5 Method comparison

Where previously it was how methods performed on average across the test region that was examined, in this section it is how the best result for a method changes (on average) over the MAF bins that is considered. Figure 4.18 shows the means of the best $-\log_{10}$ p-values selected for each sQTL separated by MAF bins, within each method. For example, take the 0 – 0.05 bin for SSR, for each of the 579 sQTL there is one SNP that produced the best $-\log_{10}$ p-value for that sQTL – so the mean displayed is the mean of those 579 $-\log_{10}$ p-values. SSR always has the lowest average best – $\log_{10}$ p-value, although it does increase substantially as MAF increases, up to a point where SSR is nearly as good as the next best method (HL7). For the lowest MAF bin,

the three haplotype methods all have the highest averages, although from the third MAF bin onwards HL7 is only better than SSR. HL3 is the best method most often for this criterion, having the highest best p-value average for five of the MAF bins (from 0.10 - 0.35), before MR5 becomes marginally better for the final three MAF bins. From a sQTL MAF of around 0.15, all methods except SSR and HL7 are very clustered however; for the 0.45 – 0.5 bin all have an average best $-\log_{10}$ p-value between 24 and 25.

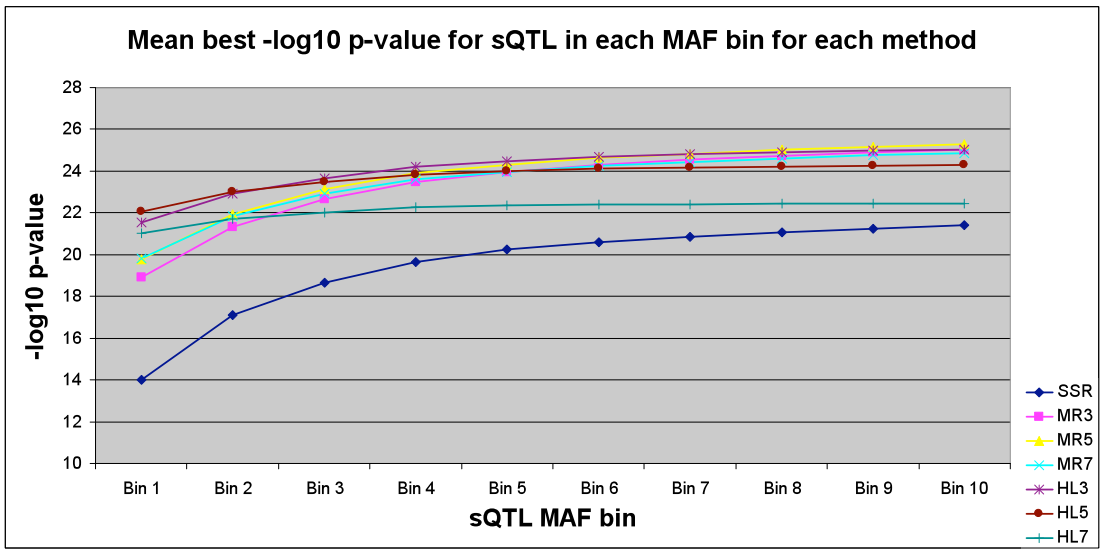


Figure 4.18 Means of the best $-\log_{10}$ p-values for each sQTL in each MAF bin for all methods. Note the y-axis begins at 10 for better resolution.

One further thing worth examining is how accurately each method is able to locate the position of the sQTL. One way to estimate how well each method performs at this is to see how far the best window position is from the known position of the sQTL. For SSR, it is easy to count the number of SNPs in-between the most significant results and the sQTL; for the multiple SNP methods, the central SNP in the window is used. Figure 4.19 shows on average how far away (in SNPs) the true sQTL position

is from the most significant window position for each method. This reveals a similar grading of methods to that already seen. SSR does worst (i.e., is least accurate), but also shows the greatest improvement as MAF increases. SSR is followed by a cluster of the remaining methods, although the two seven-SNP methods appear to be slightly less accurate than the rest. All methods are better at defining the correct location of the sQTL when the minor allele frequency of the sQTL is high.

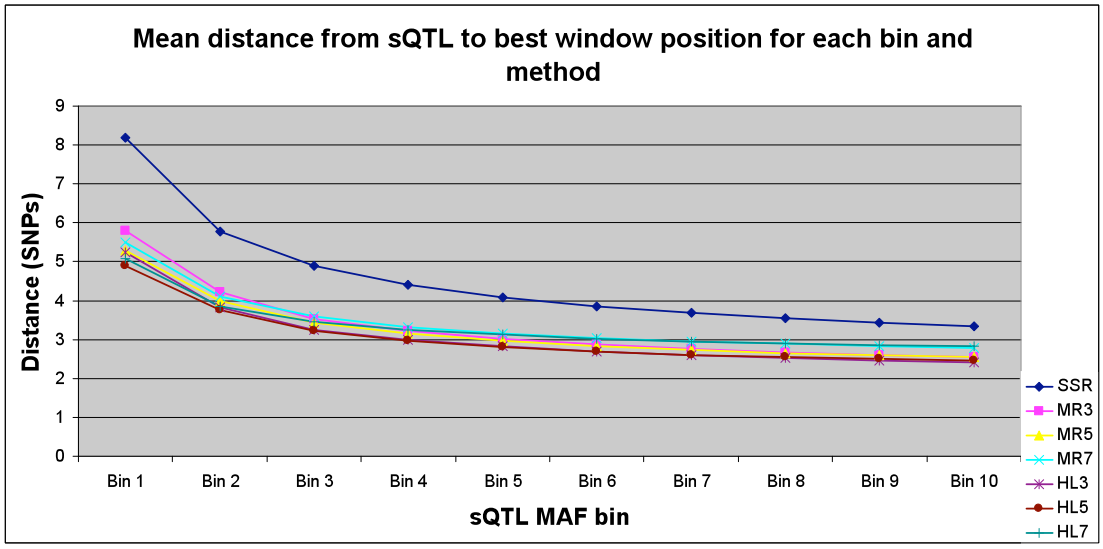


Figure 4.19 Average distance (in SNPs) between the sQTL and the most significant SNP / window position for each method. For window based methods (multiple regression and haplotypes method), the central SNP in the window is used.

A similar statistic to the accuracy of the methods (shown in Figure 4.19) is the proportion of times the best results for a method (i.e., highest $-\log_{10}$ p-value) are those which either flank (for single SNP analysis) or overlap with (for multi-marker methods) the position of the sQTL. This is shown in Figure 4.20. There is a slightly different pattern to that seen in Figure 4.19; as the number of SNPs included in the model increases (and therefore the number of windows that overlap with the sQTL also increases), the greater the proportion of times that the most significant analysis

overlaps or flanks the sQTL. For example, for the two seven-SNP methods, the windows that overlap the sQTL produce the best result for almost 70% of sQTL (Figure 4.20). The seven-SNP models are followed in turn by the five-SNP methods, then HL3, MR3 and SSR produce the least amount of best results flanking the sQTL. Interestingly, for bins 4-10, MR7 is marginally ahead of HL7, becoming the method that most often produces a best result overlapping the sQTL.

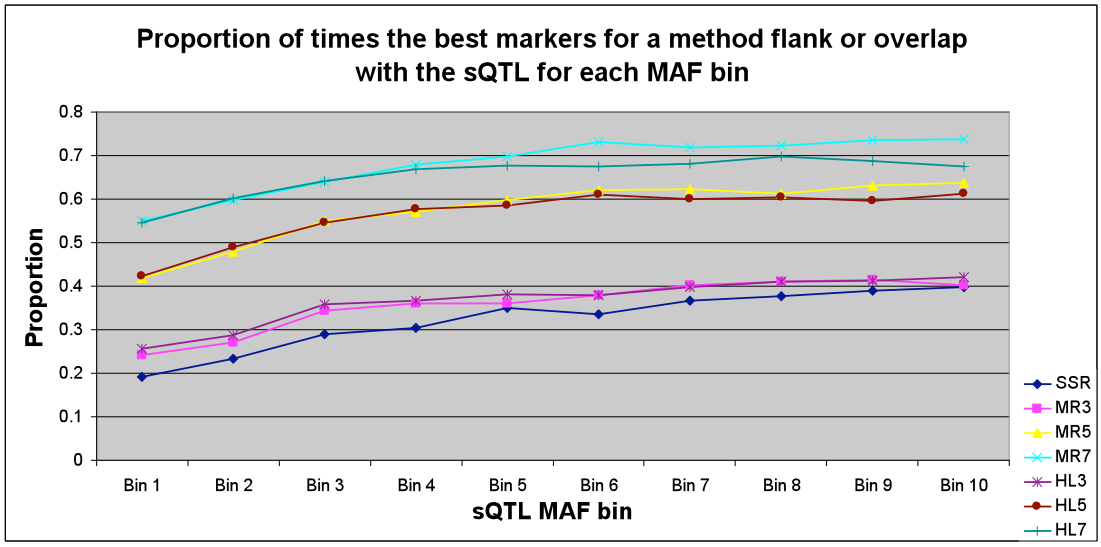


Figure 4.20 Proportion of times the best SNP / window for a model is flanking / surrounding the position of the sQTL, shown for each MAF bin.

In addition to determining how well each method performed on average, the method performing best most often was calculated. For each sQTL one specific window position / method combination produced the best result for that sQTL. On this occasion, best is meant in terms of most significant p-value, since this takes into account the differences in degrees of freedom, as already explained. Below, the percentage of times each method provides the best p-value for a sQTL is shown:

SSR:	28.13%	HL3:	15.50%
MR3:	22.49%	HL5:	7.83%
MR5:	12.55%	HL7:	3.70%
MR7:	9.80%		

Interestingly, single SNP regression is best for over a quarter of the sQTL, despite the much higher (i.e., less significant) p-values on average. The two three SNP methods are the next most frequently best methods. The five- and seven-SNP haplotype methods are best least of all.

These results have been dissected to see how the distribution of the best results changes over methods across the different MAF bins, and are shown in Figure 4.21. The figure illustrates that the relative performance of the methods is dynamic, and heavily dependent upon the minor allele frequency of the sQTL. Similar to the overall results, the method which is best most often for the majority of the MAF bins is SSR. This is best most often for eight of the bins, although it is overtaken by MR3 for the 0.40 – 0.45 MAF bin. Most interesting are the changes to the graph over the first three MAF bins. For the lowest bin, HL3 is best most often, followed by HL5, and then by SSR and HL7 which have very similar percentages. The three multiple regression methods do worst for this MAF bin. As MAF increases, the haplotype methods fall off, and the multiple regression methods and SSR in particular, begin taking over. For the highest MAF bin, the haplotype methods have become the three

worst performing methods, and the remainder of the methods are ordered in terms of the number of SNPs in the model.

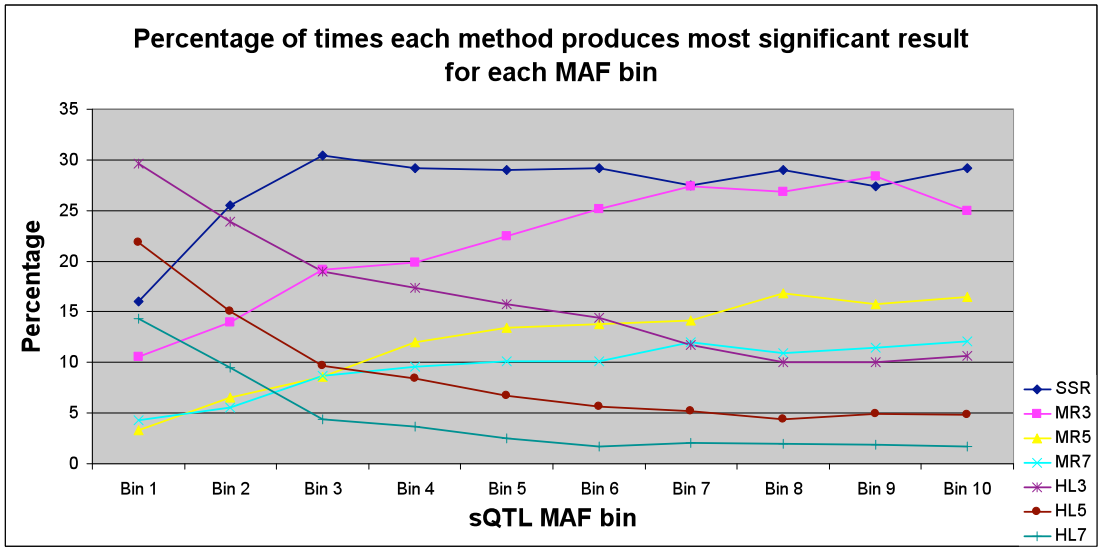


Figure 4.21 Percentage of times for each MAF bin that each method produced the most significant p-value.

The comparisons above do not tell the whole story however. When comparing all seven methods SSR generally does best most often, but this may be because when SSR is not best, there is no one other method that is consistently better. Put another way, on the occasions where SSR is not best, the method which is best may vary among all other methods, meaning that no one other method is best overall more than SSR. One way of testing this theory is to plot the percentage of times SSR does best in a series of pairwise comparisons with all other methods. This is shown in Figure 4.22 for all MAF bins. All methods are uniformly better than SSR when compared on a pairwise basis. The only method which is not better more often than SSR for all MAF bins is HL7, although it is still better more often for the lowest four MAF bins. All haplotype methods do very well for the lowest MAF bin, being better than SSR around 75% of the time, however from the third MAF bin onwards the multiple

regression methods are better than SSR more often than the haplotype methods. Even so, HL3 remains better than SSR around 60% of the time from a sQTL MAF of 0.25 onwards. The best of the multiple regression methods (MR5) is better than SSR around 65% of the time for these MAF bins.

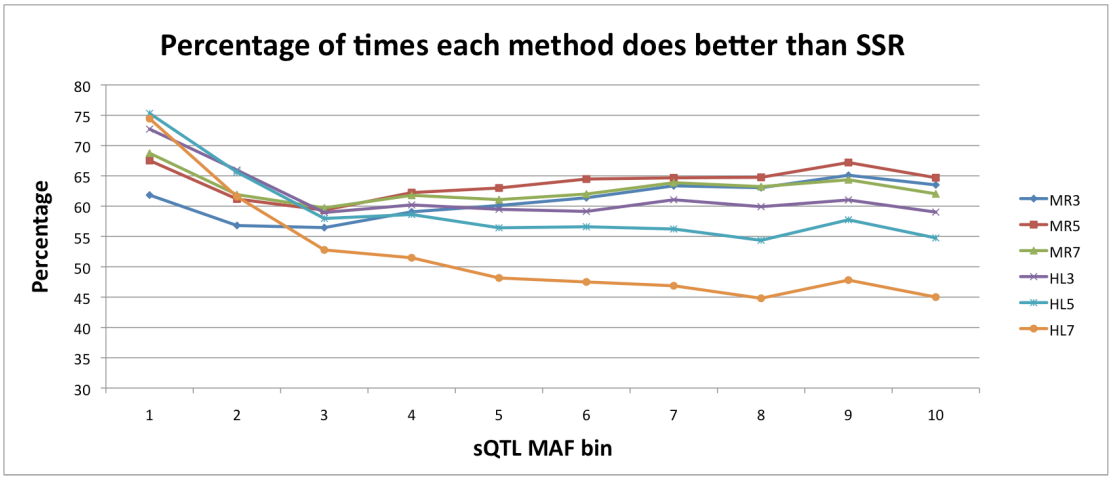


Figure 4.22 Pairwise comparisons between single SNP regression and each of the other six methods. For each bin, the percentage of times the alternate method has a more significant best p-value than SSR is shown.

To further clarify the difference in performance between single SNP regression and the best of the haplotype methods (HL3), the best p-values for these methods for sQTL in each of the lowest and highest MAF bins are plotted in Figures 4.23 and 4.24. On these figures is a line representing the event that p-values for SSR and HL3 are equal, and the density of points is illustrated by shading. Points above / to the left of the line indicate that the p-value is larger (i.e., less significant) for HL3, and the converse is true for points below / to the right of the line. Figure 4.23 shows that for sQTL for which SSR does better than HL3 (those above the line), the difference in best p-value between the methods is generally small. Contrastingly, for sQTL where HL3 does best, often the difference is very large - for the most extreme of these, SSR

has a best p-value greater than 1×10^{-10} when HL3 reaches significance in the region of 1×10^{-40}. The figure also shows that there is a greater density of points where $p > 1\times10^{-10}$, which is expected given that power is low for rare sQTL. Additionally, in concordance with Figure 4.22, there are more points below the line, i.e., where HL3 is more significant.

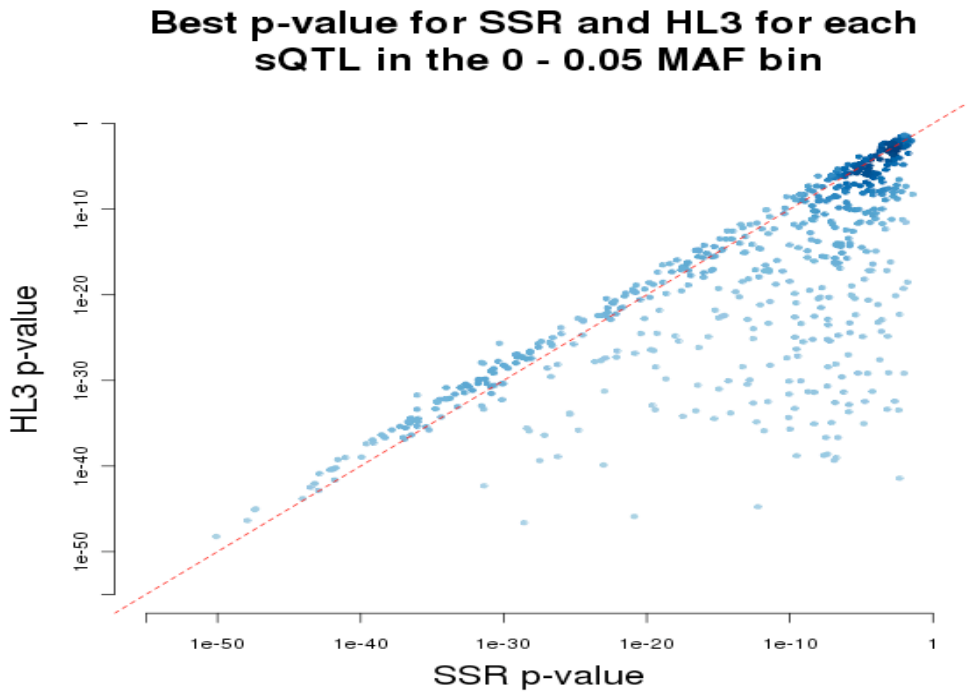


Figure 4.23 P-values for the best SNP / window for SSR and HL3 plotted against one another for all sQTL in the 0 - 0.05 MAF bin. The central SNP of the HL3 windows is used for the comparison. The dotted red line represents x = y.

In comparison to Figure 4.23, results for the highest MAF bin (Figure 4.24) show that there are slightly more points above the line, i.e., where SSR has smaller best p-value, and the greatest density of point is at a much smaller p-value. This verifies again that for common sQTL it is difficult to do better than SSR since there is a high probability

that at least one marker is in high LD with the causal variant. However, just as in Figure 4.23, where HL3 does do better than SSR, the difference in best p-value for the methods is large.

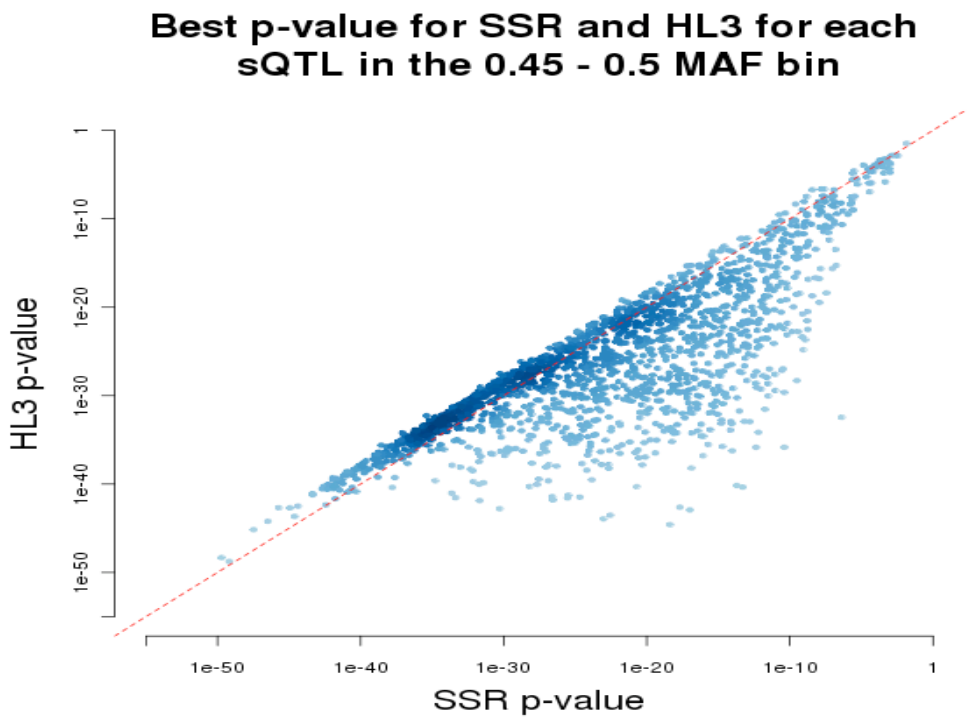


Figure 4.24 P-values for the best SNP / window for SSR and HL3 plotted against one another for all sQTL in the 0.45 - 0.5 MAF bin. The central SNP of the HL3 windows is used for the comparison. The dotted red line represents x = y.

4.4. DISCUSSION

4.4.1 Overall model differences

The results indicate there is no overwhelming distinction between model types in terms of their performance (i.e., irrespective of number of SNPs used). For example,

while SSR is the model which most often produces the single best p-value for a given sQTL, when categorised into separate MAF bins, for sQTL with lowest MAF HL3 is best most often (Figure 4.21). Interestingly, even though SSR is most frequently the model with the best p-value, SSR also has the highest average p-values (i.e., lowest -$\log_{10}$ p-values - see for example Figure 4.18). This may come about because for those sQTL which have at least one SNP in high LD – over 15% of adjacent SNPs had r^2 of at least 0.7 – very little extra variation can be explained by extra parameters, but there are fewer degrees of freedom to test with when they are present. All multi-marker methods will have smaller $-\log_{10}$ p-values in this situation. When LD drops between the sQTL and individual markers, multi-marker methods can perform much better than SSR, thus boosting the average -$\log_{10}$ p-value, however it is likely that fewer sQTL have no tagging SNP than have at least one.

Using the criterion of the proportion of variance explained by a model (and assuming that the larger the proportion of variance explained, the better the model is), the methods are ordered as follows for the overall results; SSR (0.08), MR3 (0.18), HL3 (0.19), MR5 (0.22), MR7 (0.24), HL5 (0.25) and HL7 (0.29). This indicates that although haplotype methods always explain more variation than the equivalent multiple regression method (i.e., the method with the same number of SNPs), haplotypes are not consistently better than multiple regression irrespective of window length. A better way to compare models is using -$\log_{10}$ p-value, and overall results for these are; SSR (9.33), HL3 (18.4), MR3 (18.6), HL7 (18.9), HL5 (20.2), MR5 (21.7) and MR7 (22.4). Once again neither multiple regression nor haplotype analysis is definitively better irrespective of SNPs / window length, although multiple regression

does better for the same number of SNPs. This pattern changes depending on the minor allele frequency of the sQTL however, as will be discussed in more detail subsequently.

4.4.2 Effect of number of SNPs

Within the multiple regression and haplotype analysis methods, window lengths of three, five and seven SNPs were tested. While there were only subtle differences in performance across different types of method irrespective of SNP number, there were more clear differences in how alternative window lengths performed both within and across method types. This is especially clear using the example of the proportion of variance explained by the models, which as expected increases as the number of parameters increases. Haplotype analysis is able to explain more variance than the equivalent number of SNPs in multiple regression, because not only marginal effects of the SNPs are captured, but also any interaction effects. More importantly however, the order is slightly different for the overall $-\log_{10}$ p-values. In general, significance still increases on average as the number of SNPs increases, but this is not true for HL7, which is only marginally better than the two three-SNP methods. This is because the trade-off between the proportion of variance explained and the number of degrees of freedom left to test with has become too great. The average degrees of freedom for HL7 was 22, which is ten more than the next highest of 12 (on average) for HL5. HL3 had just under six degrees of freedom on average, while SSR, MR3, MR5 and MR7 had one, three, five and seven degrees of freedom respectively.

Another consistent finding from within-model comparisons was that the overall increase in model performance in terms of significance (where there is an increase) becomes smaller as the number of SNPs gets bigger. The only times there is not an increase is in the case of the HL5-HL7 comparison; HL7 does worse than HL5 for all situations except when the sQTL MAF is under 0.05. However, for the rest of the results this finding remains true. For example, on average the proportional increases in $-\log_{10}$ p-values from SSR to MR3, MR3 to MR5 and from MR5 to MR7 are 99%, 17% and 4% respectively. This means that the change from SSR to MR3 is greater than the change from MR3 to MR5, which in turn is larger than from MR5 to MR7. Analogous conclusions can be drawn from individual MAF bin results with both the multiple regression and haplotype method results - with the exception of those already mentioned above. This suggests that although MR7 may produce more significant p-values on average than MR3 and MR5, a step up from one SNP to three SNPs will already account for a large proportion of the benefit of jumping straight to seven SNPs.

One other finding was that as the number of SNPs in the model increased, there were a greater number of windows performing well (manifested by an increase in the width of the curves). As seen in Figure 4.20, this is partly a consequence of an increase in the number of windows overlapping the true position of the sQTL, since these will do better and there are more of them for larger window sizes. MR3 and HL3 each have two windows in which the position of the sQTL is surrounded, and this increases to four and six windows for the five- and seven-SNP methods. Even so, it is important

to note that this improvement in accuracy would still be valid in real GWA studies searching for QTL.

4.4.3 Effect of sQTL MAF

The conclusions drawn thus far could largely have been made independently of results from individual MAF bins (although supporting evidence from these results back the conclusions up). However, one consistent finding throughout all results is that the ability to predict sQTL varies considerably depending on their minor allele frequency. Each method exhibited a gradual improvement in performance as sQTL minor allele frequency increased; the lowest MAF bin always performed worst, and in general the highest MAF bin performed best. At the higher end of the MAF range results became homogeneous for some methods, although this tended to happen more for methods explaining nearly all heritable variance. For example, results for SSR and MR3 are distinguishable for all MAF bins, however for HL5 and HL7 the top five MAF bins are almost inseparable. This is likely to be a consequence of the fact that the models were limited to explaining only the 30% of variance which was heritable (unless by chance some random noise was also captured). Once the 30% heritability was accounted for, different MAF bins had no room to continue improving over one another and therefore appear close in terms of performance.

One of the more interesting findings from these results is that across different methods some MAF bins experience larger proportional increases than others, in otherwise identical comparisons. The lowest MAF bin experiences the greatest

increase, and the highest MAF bin experiences the smallest increase (at least until higher MAF bins become inseparable). All intermediate bins form a sliding scale between these two extremes. For example, the percentage increases in proportion of variance explained from one method to the next for the lowest MAF bin are as follows; 210% increase from SSR to MR3, 44% increase from MR3 to MR5, 17% increase from MR5 to MR7, 56% increase from HL3 to HL5 and 27% increase from HL5 to HL7. The percentage increases become lower for the next MAF bin (0.05 – 0.1), and so on up to the highest MAF bin where the figures are at their lowest. This is also the case for the increases between methods when comparing $-\log_{10}$ p-values. Of particular note is that only the lowest MAF bin still experiences an improvement at all in the average $-\log_{10}$ p-value comparing HL5 and HL7.

It might be expected that lower MAF bins experience a larger increase than the higher MAF bins for methods where most heritable variation is being explained, since in these methods only small improvements are possible for the higher MAF bins. Indeed, this is probably a contributing factor to the fact that only MAF bin 0 – 0.05 experiences an improvement in mean $-\log_{10}$ p-value between HL5 and HL7. However, as shown above, the bigger improvement of low MAF bins is present between all methods, not just those performing best overall, so the phenomenon is not unique to only the best models. This may have implications with regard to the optimal GWAS analysis method, given that more complex methods take longer to perform. If the majority of the benefit (in terms of variance explained, if not power) of using seven-SNP models is also captured with only three SNPs, then in terms of efficiency

of time, and negating the issue of power loss from over-parameterisation, then three-SNP analyses may be the best compromise.

4.4.4 Comparison to single SNP regression

In the seven-way comparison of which method most often produced the most significant p-value across MAF bins (Figure 4.21), for eight of the MAF bins this was SSR. For the 0 – 0.05 bin, HL3 produced the most significant results for nearly twice as many sQTL as SSR, and HL5 also does better than SSR. However, for all remaining MAF bins except 0.4 - 0.45 (when MR3 is best), SSR is the best method. This result is quite striking, and seems to imply that SSR is the best method to analyse GWAS data. However, as suggested by Figure 4.22, this conclusion is not based on the optimal comparisons. Given that single SNP regression is currently the favoured method for analysing GWAS data, it is pertinent to compare results from each of the methods directly with the SSR results. These comparisons reveal that on a pairwise basis, with the exception of HL7 at moderate to high sQTL MAF, SSR is best less often than all other methods.

Initially this appears to contradict the results shown in Figure 4.21, however the reason for this apparent reversal in performance of SSR is that in general the sQTL can be grouped into two categories. The first of these contains sQTL which have one or more markers in high r^2 (note that these sQTL are thus more likely to be common), and the second category contains sQTL where there are no markers in high LD. SSR will have greater power than the alternate methods to detect sQTL in group one,

however crucially, any of the other methods might be best for detection of sQTL in the second group. In essence, for occasions where SSR is not best, the best method is shared out among the remaining methods, and this will depend on the local pattern of genetic architecture (i.e., allele frequencies and LD). The evidence to support this is present in Figure 4.22, since not only do all methods other than HL7 consistently outperform SSR on a one-to-one basis, but the relative difference in performance is greatest for the lower MAF bins (as predicted because higher MAF sQTL are more likely to have markers in high LD).

Given that it appears fewer common alleles affecting complex disease remain to be found and the CD/RV hypothesis is gaining popularity, it is encouraging that the alternate methods analysed here seem to have greater power to detect rare loci. Figure 4.22 shows that HL3 is the method which outperforms SSR most often for sQTL where MAF is under 0.05, and while other methods outperform SSR more than HL3 does at higher MAF bins, HL3 is still best more often than SSR throughout the whole MAF range. More encouraging still however, is that even for the situations where SSR is the optimal method of analysis to use, HL3 does not do much worse than SSR, i.e., HL3 still has appreciable power. Results in Figures 4.23 and 4.24 indicate that although SSR does better than HL3 around 27% and 41% of the time for the lowest and highest MAF bins (see Figure 4.22), on these occasions the best p-value is only marginally better than that produced by HL3. However, when HL3 is the better method, frequently it outperforms SSR in terms of p-value by many orders of magnitude. This suggests that although minimal power maybe lost with haplotype methods for the detection of common variants, they are not markedly inferior to

single SNP regression. Contrastingly, when it comes to the detection of rare QTL, haplotype methods are vastly superior. It should also be noted that the conclusions drawn here are based on analyses where the MAF distribution of the "disease-causing" allele (i.e., sQTL) is biased towards intermediate frequencies due to SNP selection. If the real MAF distribution of variants involved in common complex disease follows the CD/RV hypothesis, then haplotype methods will be even more favourable.

4.4.5 Further directions

One of the main drawbacks of these analyses is that there are only limited estimates of the type I error rate of the methods. Although some permutations were performed to ensure test statistics were the appropriate size for all sQTL MAF bins under the null hypothesis. a more detailed investigation involving more sQTL would have allowed a better look at the empirical distribution of the test statistic for each method, and therefore determination of type I error rates. False positive rates are just as important as the power of the test, since even with high power, it is impossible to tell whether significant results are real if the type I error rate is also high. A more detailed permutation investigation of false positive rates for methods in this study was not performed due to time constraints, given the computationally demanding nature of these analyses.

Another aspect of these analyses which could have been improved upon without time and computational constraints was the amount of heritability simulated for the sQTL.

In these analyses a heritability of 30% was used, however, individual complex disease loci identified to date rarely explain even a tenth of this. Simulating smaller effect sizes may reflect true complex disease loci more accurately, however in this study, power would have been reduced to a point where the ability to make meaningful comparisons between methods was compromised. Ideally, larger sample sizes and smaller sQTL effects would have been investigated, however this would have increased the duration of the analyses substantially, and moreover, should produce similar overall conclusions. Therefore, while not ideal, the use of high heritability is more of a means to clarify discrepancies between the methods, and is not of intrinsic importance to the results.

There are a number of other things which would have been interesting to investigate with these data. For example, it would be interesting to compare how well haplotypes would perform if only the most likely haplotype pair was used for each individual. The literature implies that this technique would not perform as well as the haplotype method employed in this study (Morris et al., 2004), however it would be good to verify this and quantify any differences there were. Assuming that using the most likely diplotype per individual as if it were known did indeed perform worse than utilising all possible haplotypes for each individual, it would be interesting to see whether this method still managed to out-perform single SNP or multiple SNP regression.

One further aspect of this work that could have been explored was how alternate ways of dealing with rare haplotypes affected the conclusions. In these analyses, no

haplotype classes were removed, regardless of how rare, since the best way of dealing with rare haplotype classes is not clear, and discarding information is innately unappealing. However, this approach may not be optimal, as frequency estimates for rare haplotypes can have large variances due to sampling variation (Fallin and Schork, 2000), and parameter estimates when using rare haplotypes will also have large variances (Schaid, 2004). In addition, using rare haplotypes is likely to have contributed to the fact that HL7 is unable to keep up with the performance of HL5, since on average there were around 10 extra degrees of freedom taken up in the HL7 analyses. Pooling rare haplotype classes or using a clustering algorithm are both methods that could have been compared, along with variable selection type methods such as lasso regression that drop parameters adding little to the model.

4.4.6 Implications

Results of this study have direct implications for detection of QTL in GWA studies. The most widely (and often only) analysis method used in GWA studies is analysis of a single SNP. Frequently, where alternative forms of analysis are used, it is only in regions already implicated by prior results, therefore it is rare for an exhaustive, genome-wide scan to be performed using any of the multiple SNP methods compared in this study. However, results from this chapter support those already in the literature suggesting that analysing GWA data using multi-marker methods can be beneficial. For example MR3 is the best method for detecting sQTL with moderate to high MAF almost as often as SSR (Figure 4.21), indicating that multiple regression may have greater power to detect some associations than single SNP regression, even when the

frequency of the sQTL is high. The pairwise method comparison results offer more compelling evidence in favour of multi-marker methods however. From Figure 4.22, MAF bin 0.1 - 0.15 is the one where SSR has the most favourable comparisons to all other methods (i.e., it is better more frequently in this MAF bin than for other MAF bins), but even here HL3, MR3 and MR5 have smaller best p-values for 60% of the sQTL.

It is for very low sQTL MAF that most benefit from using haplotype analysis is seen however. In the seven-way comparison, haplotype methods together produce the most significant p-value for around 67% of sQTL when MAF is 0 – 0.05, and both HL3 and HL5 perform better than SSR. In the pairwise comparison, HL3 is better than SSR almost 75% of the time for this MAF bin however, and even HL7, which in general does worse than SSR (is better less than 50% of the time) is also better than SSR 75% of the time. Together with results from Figures 4.23 and 4.24, this suggests that for detection of rare QTL, the gain in power from haplotype methods is considerable. The fact that HL7 also does well for low MAF sQTL also gives some impression of how the optimal length of haplotype window is dynamic, and will vary according to local genetic architecture. Where there are dense SNPs and longer stretches of LD, haplotype windows can be longer without risk of there being too many haplotypes in the population and causing over-parameterisation of the model. Conversely, having short haplotypes in regions of high LD will not add much information to a model, since in this case haplotypes tend towards biallelic markers. Therefore the optimal choice of haplotype window length is a delicate balance that depends on marker density and local patterns of LD. There are some methods which

have attempted to alter the size of the window tested by using local LD patterns, but these are currently not widely used (e.g., Browning, 2006; Li et al., 2007).

The results for rare sQTL are of particular relevance to current literature. As will be discussed in greater detail in the final chapter, there is currently concern that GWA studies are failing to detect QTL accounting for the majority of heritable variation for human complex diseases, and complex traits in general (Maher, 2008). For this reason the common disease / rare variant (CD/RV) hypothesis is gaining popularity as a potential explanation of where some of this heritable variation may be hiding (Reich and Lander, 2001). It certainly seems the case that, even with larger and larger GWA studies, not enough common QTL are being identified to account for the heritable variation present in these traits, and there must come a time where no more can be found. If the CD/RV hypothesis is correct, then the analyses performed here suggest that haplotype methods may be even more valuable, since detecting rare variants is where the advantage of using haplotypes is most keenly seen. While it is easier to implement single- and multiple-SNP regression since it can be achieved without haplotype estimation, haplotypes are far more likely to detect rare QTL. Given that there are fast and accurate haplotyping algorithms freely available to the scientific community, and that haplotypes may provide the boost in power required for rare QTL to be detected, this study strongly advocates the use of haplotypes for future GWA studies.

5. CHAPTER 5

5.1 INTRODUCTION

In the previous chapter, a number of multi-marker methods for analysing genome-wide association study (GWAS) data were compared to single SNP regression (SSR), currently the most popular way of analysing such data. Results suggested that single SNP regression was not always likely to be the most effective way to perform genome-wide association (GWA), and that both multiple regression and haplotype analysis may be preferable in certain situations. The phenotypes used in these analyses were not real however, instead consisting of SNP genotypes with added noise to make the trait quantitative, and to reduce heritability from one to 0.3 (on average). Simulating phenotypes in this way was useful because it allowed comparisons to be made where the real "answers" were known, therefore it was possible to identify which methods were performing best. However, it would also be interesting to apply these methods to real GWAS phenotype data to see how results differ from those when using single SNP regression.

In this chapter, phenotypic data was selected from the CROAS dataset on which to perform the analysis methods from the previous chapter. Performing GWA using all seven methods on the entire set of phenotypes collected as part of CROAS would be prohibitively time-consuming, therefore a single trait was chosen; uric acid. Although the choice of phenotype was largely arbitrary, uric acid was chosen due to the

identification of *SLC2A9* as a urate transporter in the initial GWAS, since it would be interesting to see how well this association was picked up by the alternative methods.

The aim of using multiple regression and haplotype analysis methods on real phenotype data is to identify some of the so-called "missing heritability" of traits involved in common complex disease. Rare variants (<5% frequency) are currently postulated as one of the reasons for the failure to explain all genotypic variance displayed by multifactorial diseases (Maher, 2008). The haplotype method in particular has more chance of detecting rare variants affecting complex disease than both single- and multiple-SNP regression due to the fact that haplotypes are better able to tag rare SNPs (as described in chapter four).

Assuming rare variants are responsible for an appreciable proportion of unexplained heritability in common complex disease, it is reasonable to expect that this is because they are relatively young rather than due to selection; variants with a strong enough effect on disease to be under purifying selection are likely have been discovered already. As a consequence of their young age, undiscovered rare disease-influencing variants are likely to be on one or a small number of haplotype backgrounds, and their r^2 with any given pre-existing SNP is unlikely to be high due to allele frequency differences, particularly as the allele frequency difference between the new QTL and pre-existing SNPs becomes more extreme. The local haplotype background may define the new variant rather well however. In these situations haplotype analyses should perform better than the more widely used approach of SSR, and potentially uncover associations that would have otherwise been missed.

5.2 MATERIALS AND METHODS

5.2.1 Phenotypic and genotypic data

Both phenotypic and genotypic data for these analyses came from the CROAS project described in chapter two. Rather than using all traits, the only trait analysed here was uric acid, although unlike chapter four, all chromosomes were analysed. To be consistent across methods, not all individuals from the dataset were used. This was because the study population is a mixture of unrelated and related individuals, causing haplotyping of all individuals jointly to become more complex, since most haplotyping algorithms assume populations are either entirely related or entirely unrelated. The haplotype estimation software used for these analyses, PHASE (Stephens et al., 2001), uses an algorithm assuming independence of each individual. Therefore to avoid the difficulties associated with haplotyping a study population with mixed levels of relatedness, only unrelated and founder individuals were used. Eight founder / unrelated individuals did not have uric acid phenotypic records and were also excluded. Quality control was performed for this reduced dataset, since exclusion of extra samples may result in a different set of SNPs not reaching the call rate threshold. The total numbers of individuals and SNPs remaining were 445 and 291,253 respectively.

5.2.2 Performing the analyses

Analyses were performed as described in chapter four, with the exception that relevant fixed effects and covariates (as decided from analysis in chapter two) were also fitted in the model. These were BMI, an age-by-sex interaction and structure (i.e., which group the individual belonged to as determined by the program STRUCTURE (Pritchard et al., 2000) - roughly corresponding to village for this dataset). The random polygenic effect accounting for genetic covariances caused by family structure was omitted, since only founder and unrelated individuals were used in these analyses. This model was used to test all SNPs in a single SNP regression framework. For multiple regression, a sliding window of n SNPs (where n equals three, five, and seven respectively) was used in a global test for association (model described in chapter four, section 4.2.3.3), such that for a chromosome of length N SNPs there were N-n+1 tests. No windows spanned two chromosomes. Multiple regression methods are referred to as MR3, MR5 and MR7, depending on the number of SNPs used (n).

For the haplotype analyses, as in chapter four every possible consecutive 11-SNP window was phased, regardless of the window length tested in the model (i.e., three, five or seven SNPs), although this was performed for all chromosomes in this chapter, not only chromosome 4. Haplotypes used to perform the test were extracted from the centre of the 11-SNP windows. As previously described, 11 SNP-windows were phased to ensure that the estimated haplotypes were as accurate as possible, without causing the time taken to perform the phasing to become unfeasible. Phasing

of windows on chromosome 4 was already complete since it had been done for the analyses in chapter four. For a chromosome of *N* SNPs there were *N*-11+1 tests, since each test was performed on haplotypes from the centre of an 11-SNP window that did not span more than one chromosome. Haplotype methods are referred to HL3, HL5 and HL7, depending on the length of window used.

A Bonferroni correction was used independently for each of the methods in order to correct for multiple testing. This was to better reflect a true GWAS, where only one method would be used to initially scan the data. As the number of tests performed was approximately equal for each method, the Bonferroni correction was identical to two decimal places: 1.72×10^{-7} (corresponding to a $-\log_{10}$ p-value of 6.76).

5.3 RESULTS

5.3.1 Manhattan plots

Manhattan plots showing genome-wide results for each of the methods are presented in Figures 5.1 - 5.7. There is no line indicating Bonferroni significance in the figures as no results exceeded this threshold, although some came close. It should be noted however that the Bonferroni correction is very stringent due to correlations present between SNPs in the form of linkage disequilibrium (LD). Additionally, for the multiple SNP methods there is an extra source of correlation introduced by re-using SNPs in numerous tests, which acts to make the correction factor even more stringent. Therefore "suggestive" results close to Bonferroni significance may still be worthy of

closer examination. Accordingly, the top 20 results for each method are examined more closely and compared across methods, and Figure 5.8 (shown later in the chapter) displays the top 50 results for each method for clarification and ease of reference.

Genome-wide results for single SNP regression can be seen in Figure 5.1. The closest to obtaining genome-wide significance for SSR was rs1323771 on chromosome 9, with a p-value of 3.38×10^{-7}. This association consists of only a single SNP, as no other nearby SNPs show suggestive association. There are no SNPs in the dataset that are in particularly high LD with rs1323771 however, the highest two being rs10114830 ($r^2 = 0.52$) and rs10117817 ($r^2 = 0.47$). The p-values for these two SNPs are 0.006 and 0.001 respectively, therefore these SNPs do stand above the background noise. The fact that these two SNPs have the highest LD with rs1323771 of any in this dataset may explain why no other SNPs reach a similar level of significance as rs1323771.

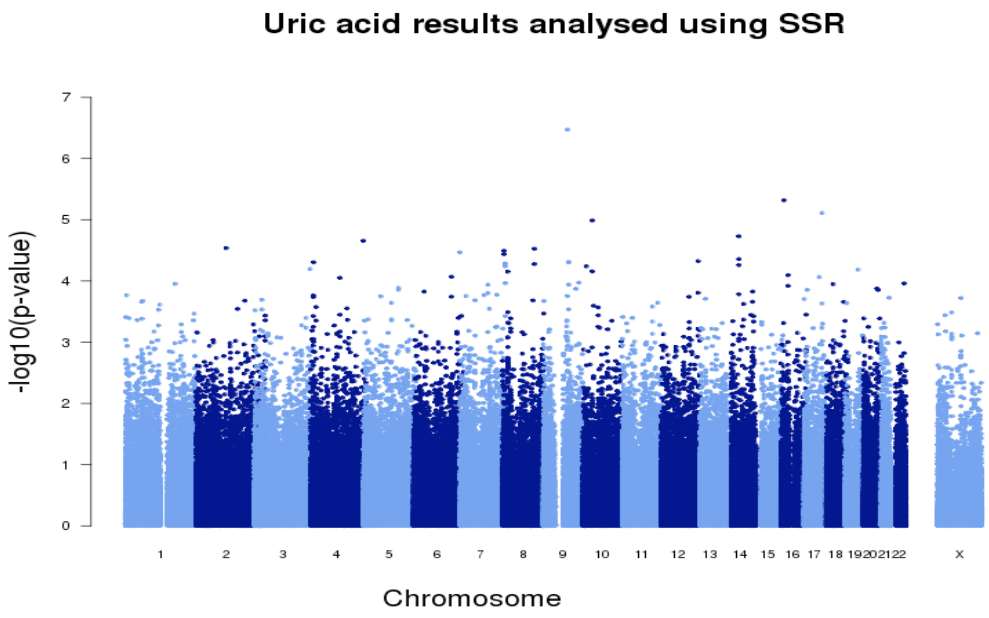


Figure 5.1 Genome-wide results for uric acid analysed using single SNP regression.

Also of interest for SSR are three SNPs in the top 20 results spanning a 165Kb region on chromosome 14 (two of which are consecutive), the most significant of which, rs858893, with a p-value of 1.87×10^{-5}. There are other occurrences of consecutive suggestively significant SNPs on chromosomes 7 (at 155Mb), 8 (at 4.6Mb), and 9 (at 87Mb). Also worthy of note is that a SNP tagging the *SLC2A9* finding on chromosome 4, rs7659670, is among the top 20 hits.

Results for the multiple regression methods are displayed in Figures 5.2 - 5.4. No test has a p-value more extreme than 1×10^{-6}, and only six have a p-value more extreme than 1×10^{-5} for the three methods. However, three of these windows are positioned very close to rs1323771 on chromosome 9, which produced the most significant result for SSR, and one of the remaining three tags *SLC2A9*. The remaining two windows were found with MR5 and are located 52Mb into chromosome 17.

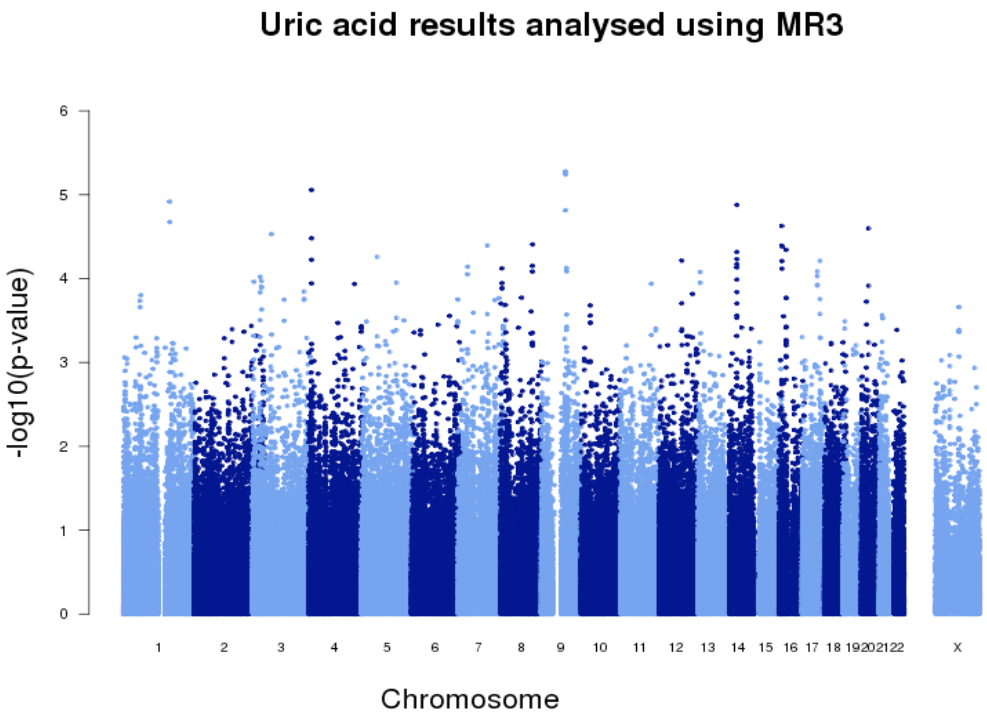


Figure 5.2 Genome-wide results for uric acid analysed using three-SNP multiple regression.

On chromosome 14, in total 12 results (amounting to seven different SNPs / windows) over the regression methods (i.e., SSR, MR3, MR5 and MR7) are located near to one another at 46Mb. There are also regions shared in common in the top 20 hits for the four regression methods on chromosomes 7 (at 155Mb), 8 (at 109Mb) and 16 (at 10Mb). Within each multiple regression method almost all top results have at least one other result in the top 20 from a nearby window; there are five regions taking up 14 of the top 20 results for MR3, six regions encompassing 14 results for MR5 and five regions taking up 15 of the top results for MR7. This reflects the extra source of correlation between tests mentioned earlier.

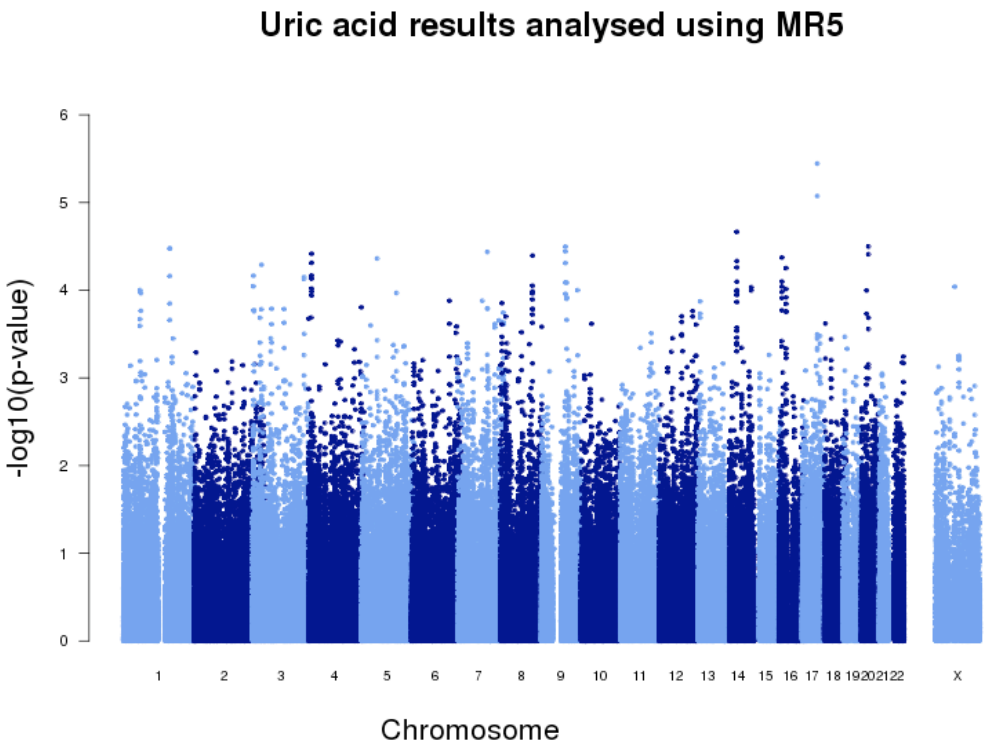


Figure 5.3 Genome-wide results for uric acid analysed using five-SNP multiple regression.

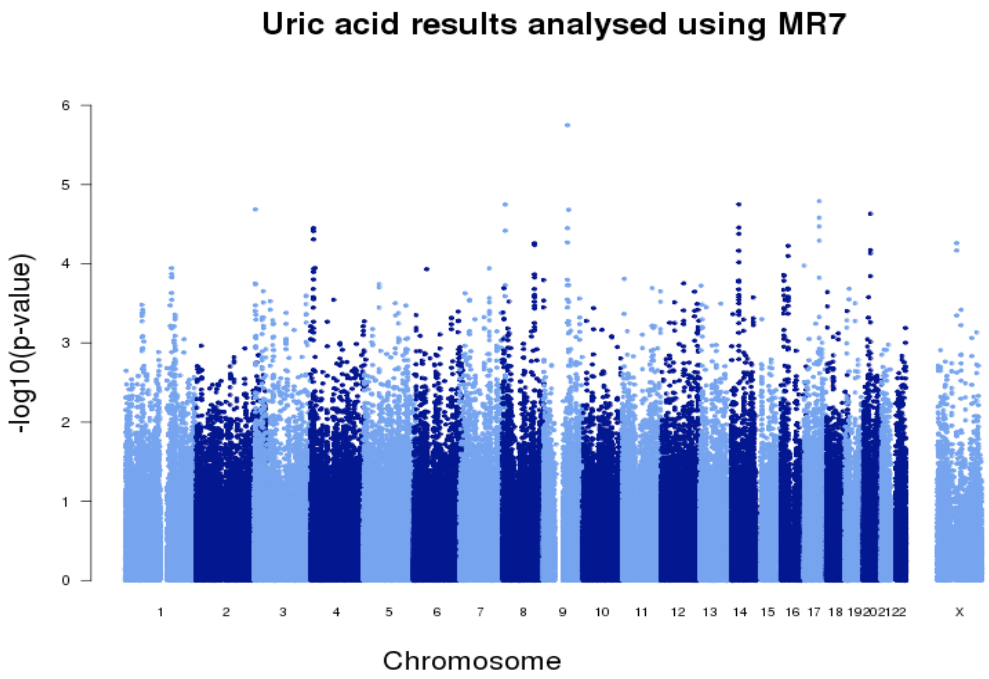
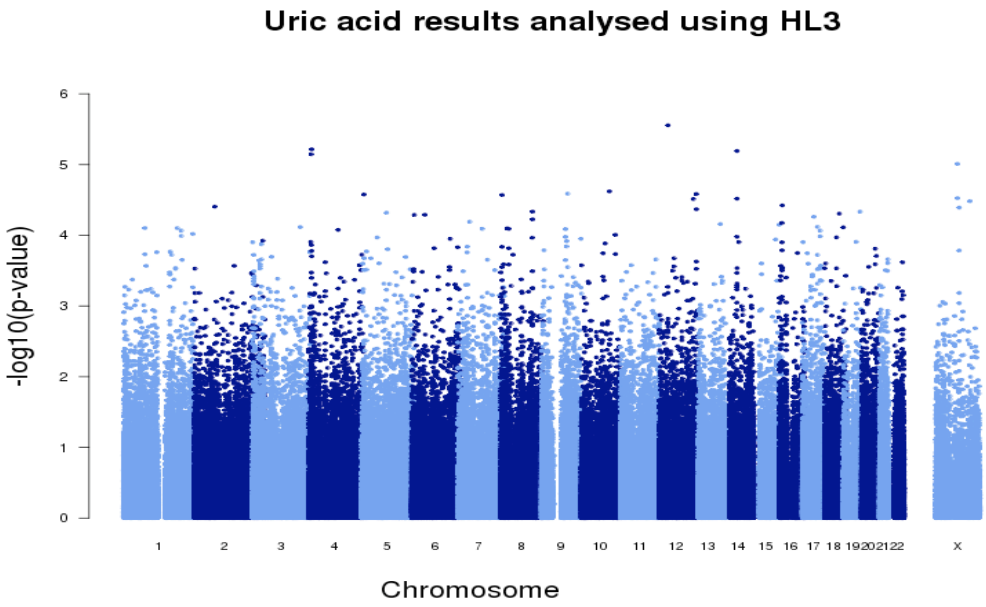
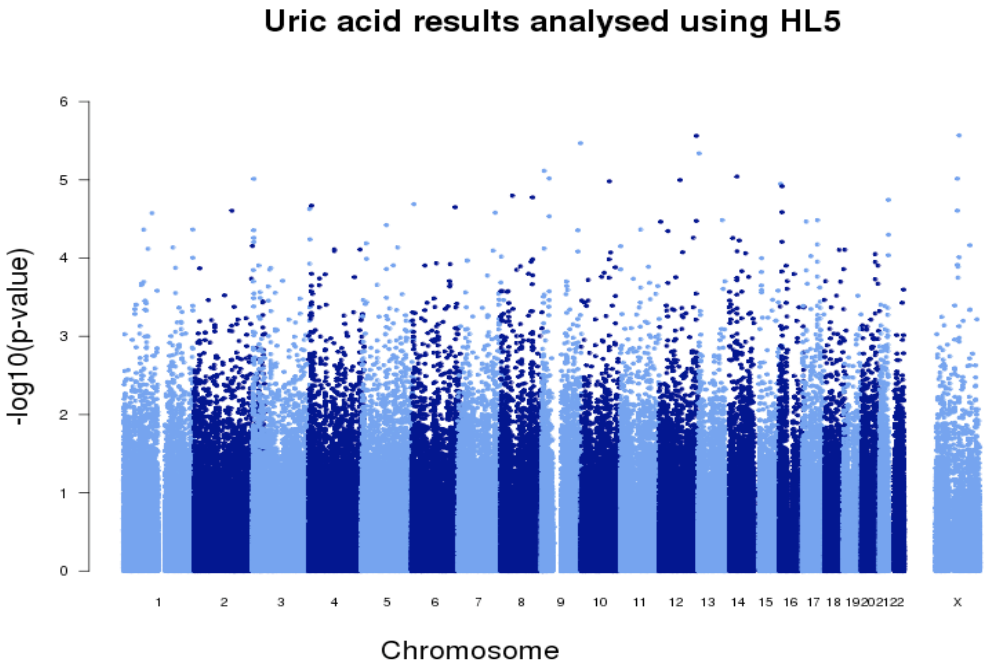


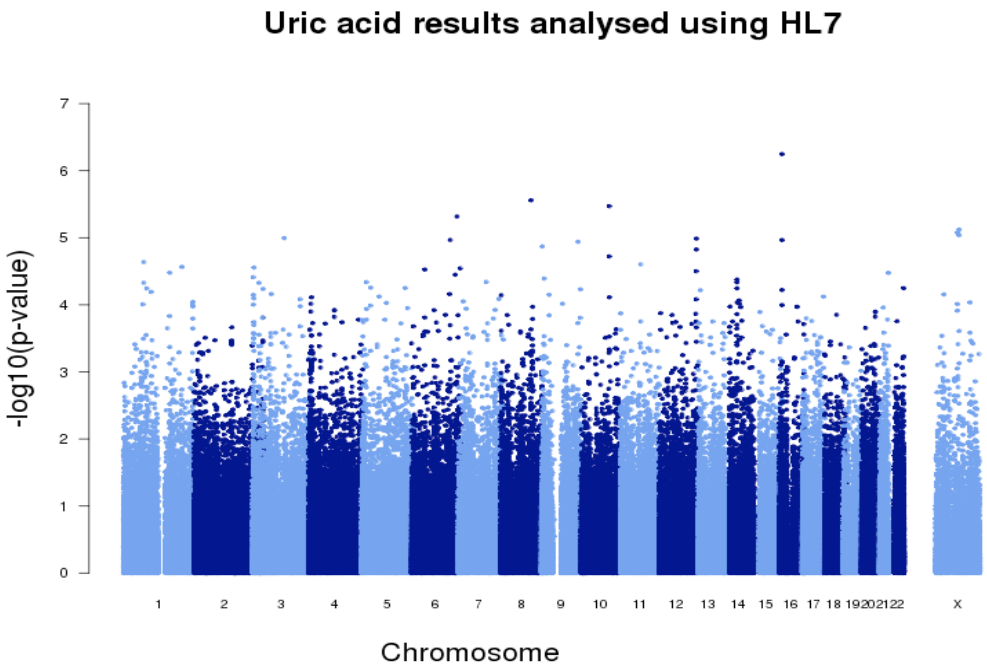
Figure 5.4 Genome-wide results for uric acid analysed using seven-SNP multiple regression.

Figures 5.5 - 5.7 show the Manhattan plots for the haplotype methods. Again no windows exceed Bonferroni significance, and only one exceeded 1×10^{-6}. This window centred on rs2033188 on chromosome 16 for HL7, with a p-value of 5.67×10^{-7}. No window for any haplotype method shows strong support for the significant results for SSR and the multiple regression methods on chromosome 9. There are four occurrences for both HL3 and HL7 where two nearby windows are in the top 20 results, but none of the 20 most significant windows for HL5 are located close to one another. This is marked contrast to the multiple regression methods where most of the top results had another significant window close by. Two of the top 20 windows for HL3 are on chromosome 14, as is one of the top 20 for HL5, supporting the associations found using the other methods at this location. These windows differ from those previously identified, but are all located nearby. Unlike

the other methods, HL3 has a top hit located inside the *SLC2A9* gene. The four regression methods and HL5 all have windows tagging the association further upstream than the gene itself, although interestingly, there are no windows tagging *SLC2A9* in the top 20 results for HL7.







Figures 5.5 - 5.7 Genome-wide results for uric acid analysed using three-, five- and seven-SNP haplotype analysis.

Figure 5.8 plots the top 50 hits for each method on the same graph, with different colours representing the alternate methods. The majority of points represent p-values greater than 1×10^{-5}, highlighting the general lack of highly significant results. Two of the three most significant points are on chromosome 9, one corresponding to rs1323771 identified using SSR, and the other corresponding to a MR7 window centred on rs7041080. The string of results at the beginning of chromosome 4 is at the *SLC2A9* locus; the most significant window (centred on rs4697674) has p-value 6.08×10^{-6}, and was produced using the HL3 method. There are a number of other occurrences of a string of results at a particular locus, and these may be indicative of additional real associations. These have all been mentioned above, and are examined more closely subsequently.

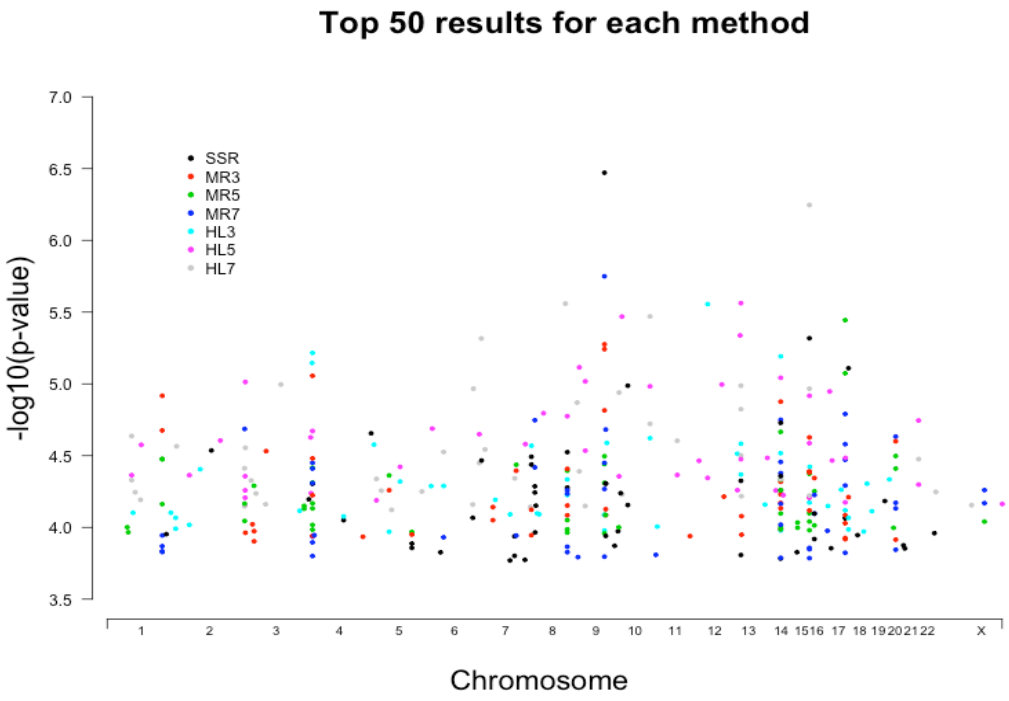


Figure 5.8 Top 50 results for each of the methods. Colours distinguish the alternate methods. Note that the y-axis starts at 3.5.

5.3.2 Examining the plots and identifying further associations

A plot of each method and chromosome combination was examined in addition to the genome-wide Manhattan plots. This allows easier identification of signals that look promising by virtue of multiple close associations, and also shows where there is a lack of support. Viewing the results in this manner also ensures any suggestive regions are still found should none of the individual SNPs / windows be in the top 20 results of any method. However, the only region not already identified that was showing slightly elevated significance for a cluster of windows was on chromosome 12, the most significant of which was from HL5 (see Figure 5.8). Numerous SNPs / windows at 79Mb stand above the background noise, however the p-value of the most

significant test is only marginally less than 1×10^{-5}, therefore the overall evidence for association is not very strong. Several method and chromosome combinations for the results discussed previously are presented below, however the entire set of graphs is not shown since there are too many (23 for each of the seven methods).

The single most significant result from these analyses was on chromosome 9 from the SSR method, although other methods also produced significant results at this locus. The association is shown for SSR in Figure 5.9. There is a single SNP reaching a $-\log_{10}$ p-value of around 6.5, however the next most significant result in the same peak has a $-\log_{10}$ p-value just under 4. As explained earlier, it not surprising that there are no more supporting SNPs for this association, as the most significant SNP (rs1323771) is not in high LD with any other SNPs.

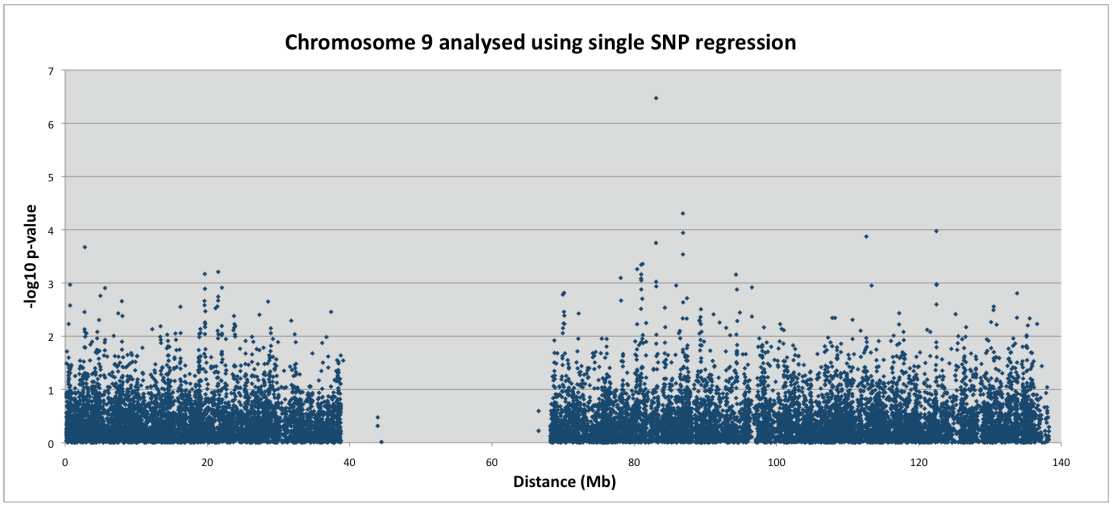


Figure 5.9 Results from single SNP regression analysis on chromosome 9 for uric acid.

Another of the most strongly supported associations identified from the top results was that on chromosome 14. Figure 5.10, showing results for this chromosome analysed using MR7, illustrates this. There are many windows in the peak at around

46Mb into the chromosome, however even the most significant of these only reaches 1.78×10^{-5}. It should be noted for this particular method however, that correlations between tests may inflate the number that show suggestive significance.

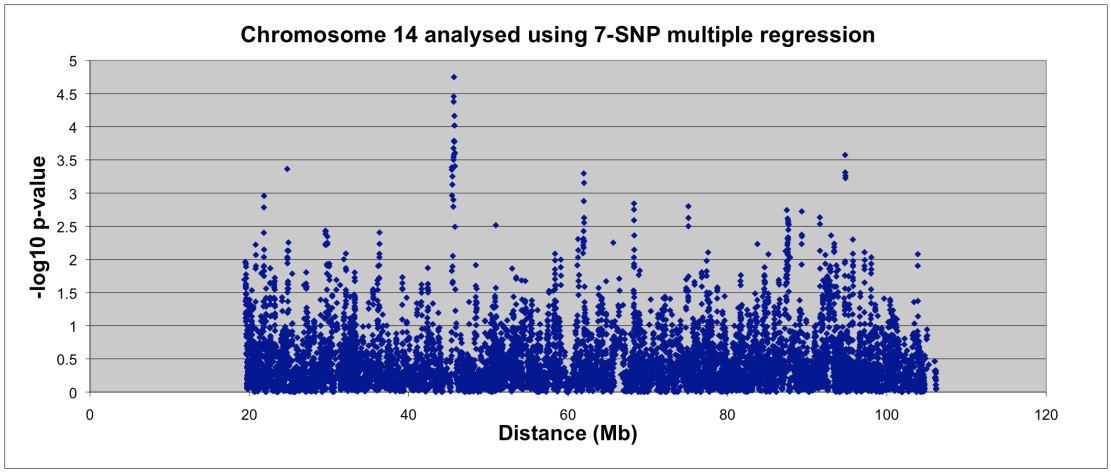


Figure 5.10 Results from seven-SNP multiple regression analysis on chromosome 14 for uric acid.

Figure 5.11 shows chromosome 1 analysed using HL7. Five SNPs / windows across all methods showed an association 71Mb into the chromosome, but from this figure the evidence for association is poor. In addition to lacking a genome-wide significant p-value, there is a large amount of background noise from which the "significant" results barely stand out. Based on this, it would appear that no true association is picked up on chromosome 1.

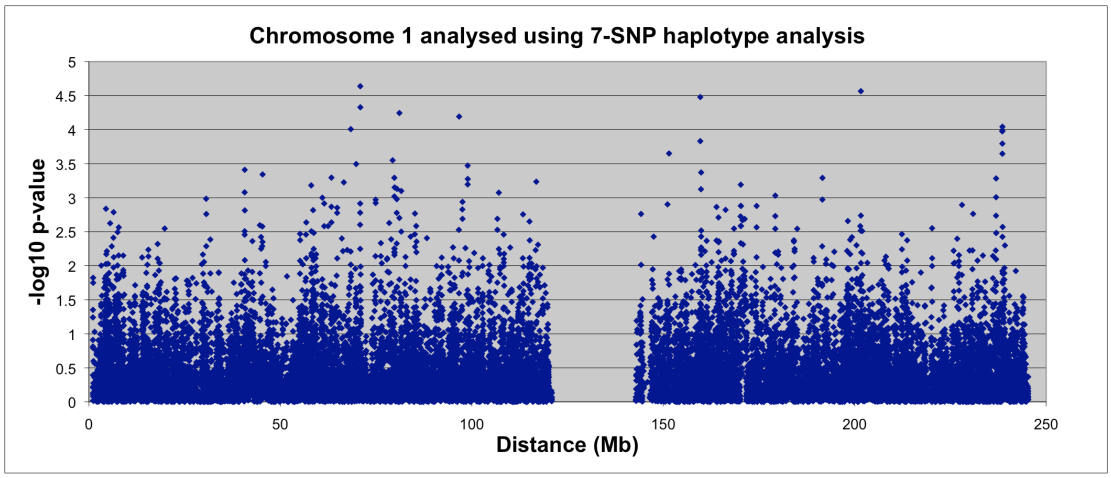


Figure 5.11 Results from seven-SNP haplotype analysis on chromosome 1 for uric acid.

Figure 5.12 shows a putative association at 103Mb on chromosome 2 for SSR. This association is another consisting of only a single SNP (rs11123953) that does not exceed Bonferroni significance, but is suggestive. The SNP with highest r^2 to this SNP in the dataset is rs10176694, which has an r^2 of only 0.26 with rs11123953, and a p-value of 0.006. As with rs1323771 on chromosome 9, the lack of SNPs in high LD means that having no further support for this association is not unexpected.

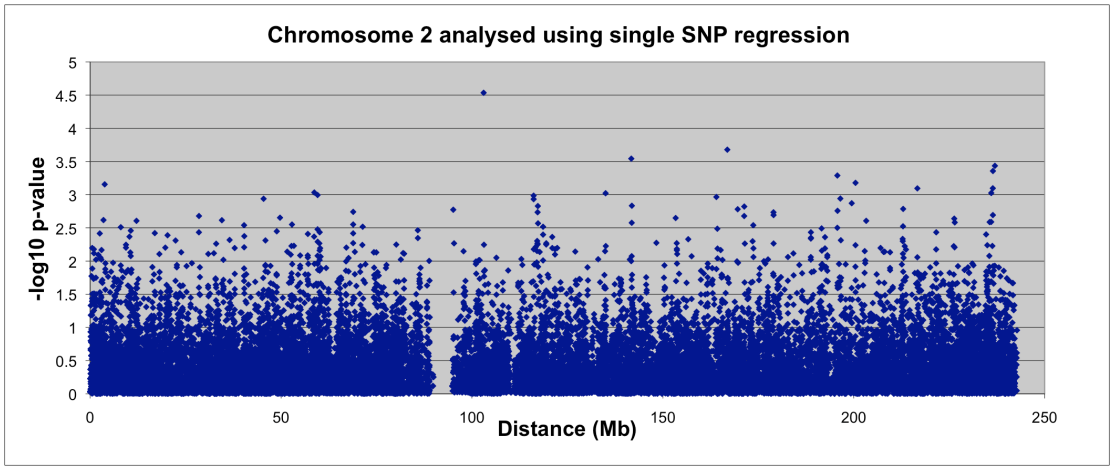


Figure 5.12 Results from single SNP regression analysis on chromosome 2 for uric acid.

Figures 5.13 and 5.14 show three other potential associations on chromosomes 16 and 17 respectively. The associations on chromosome 16 (Figure 5.13, showing MR3) are at 10Mb and 25Mb into the chromosome, and on this occasion there are more convincing strings of less significant results which also implicate the regions, while the level of background noise appears to be slightly less. The final association, shown in Figure 5.14, is around 52Mb into chromosome 17. Again, while significance is low, numerous windows support this association, although as in Figure 5.10 these may be due to the correlations between tests, since the method producing these results is MR7.

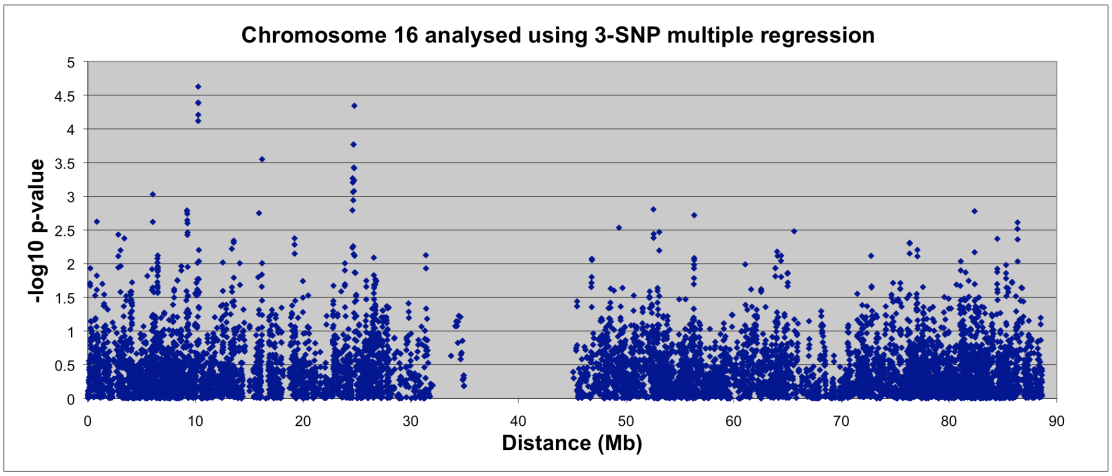


Figure 5.13 Results from three-SNP multiple regression analysis on chromosome 16 for uric acid.

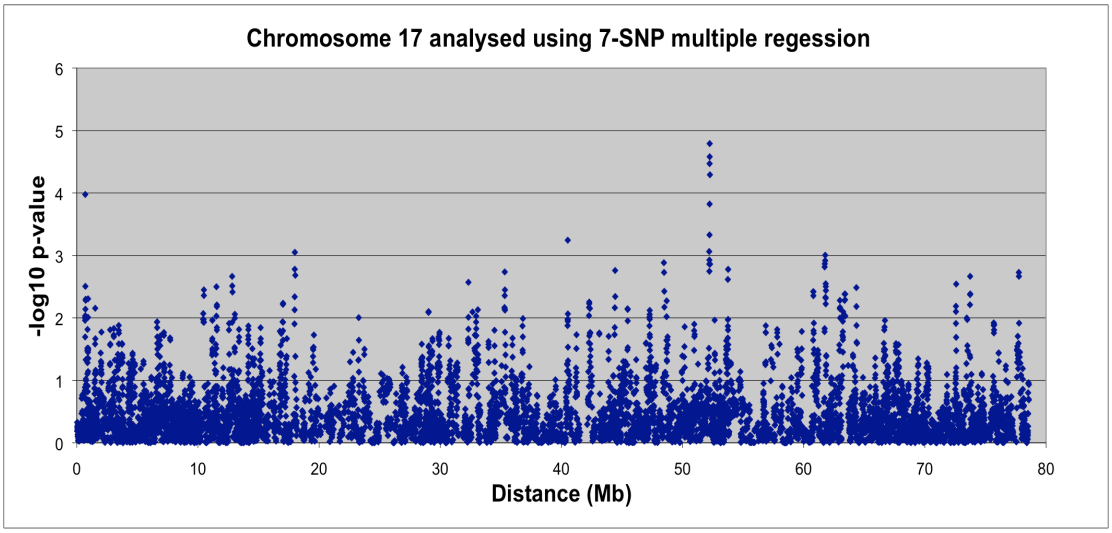
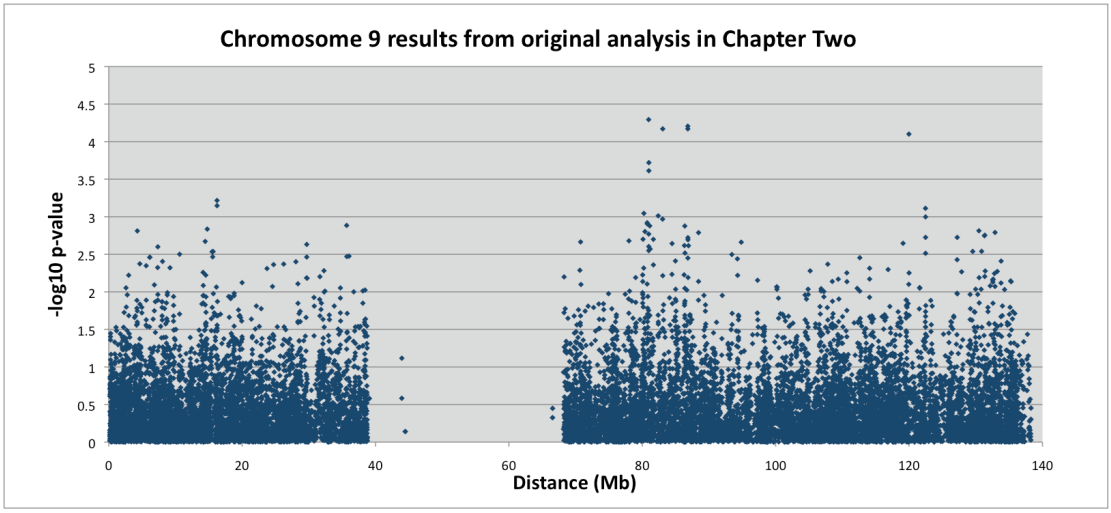


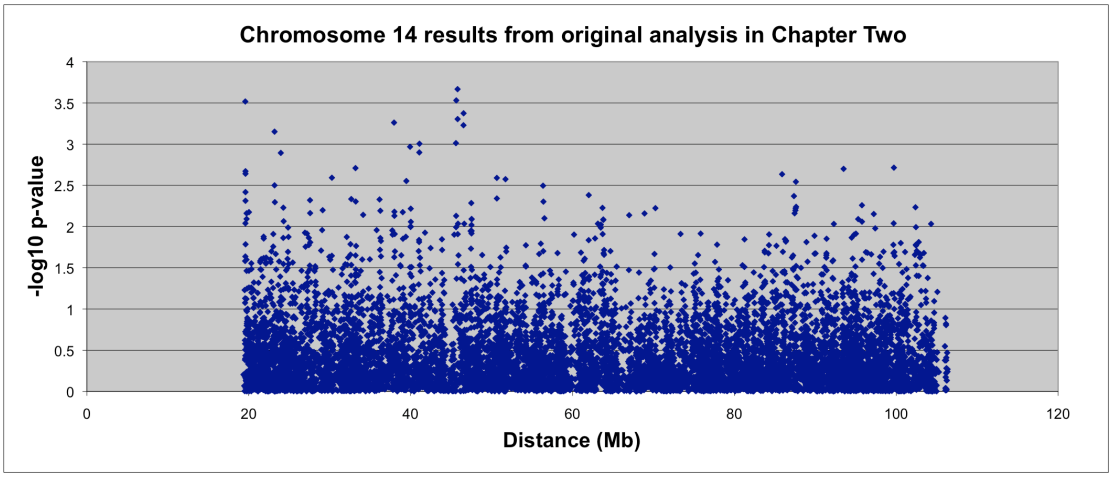
Figure 5.14 Results from seven-SNP multiple regression analysis on chromosome 17 for uric acid.

5.3.3 Cross-referencing with the original analysis

Results from this chapter were compared with the original uric acid analysis results from chapter two to determine whether the most significant results were also showing elevated significance in the original study. Since it was not restricted to only founder and unrelated individuals, the original study therefore had increased power in this

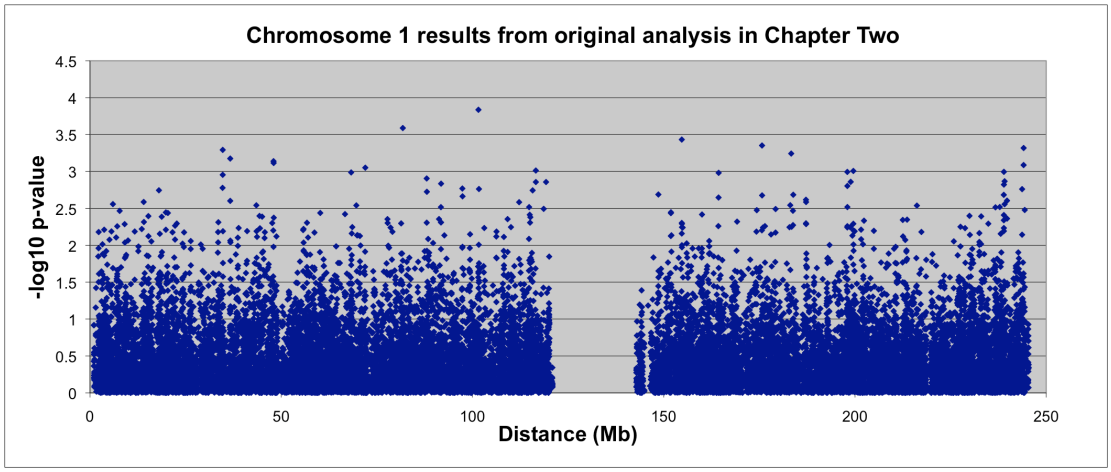
respect. However, results from the original scan do not provide overwhelming supporting evidence for an association at the most significant locus identified in these results - 85Mb into chromosome 9 (see Figure 5.15). There are a cluster of SNPs at around 80 - 87Mb into the chromosome showing elevated significance, however the most significant has a $-\log_{10}$ p-value of less than 4.5. Original analysis results are similar for the association 46Mb into chromosome 14 (see Figure 5.16) Although the most significant result for this chromosome is at the correct place, and there is also a small cluster of nearby SNPs with the same level of significance, the absolute significance level is too low to suggest association.

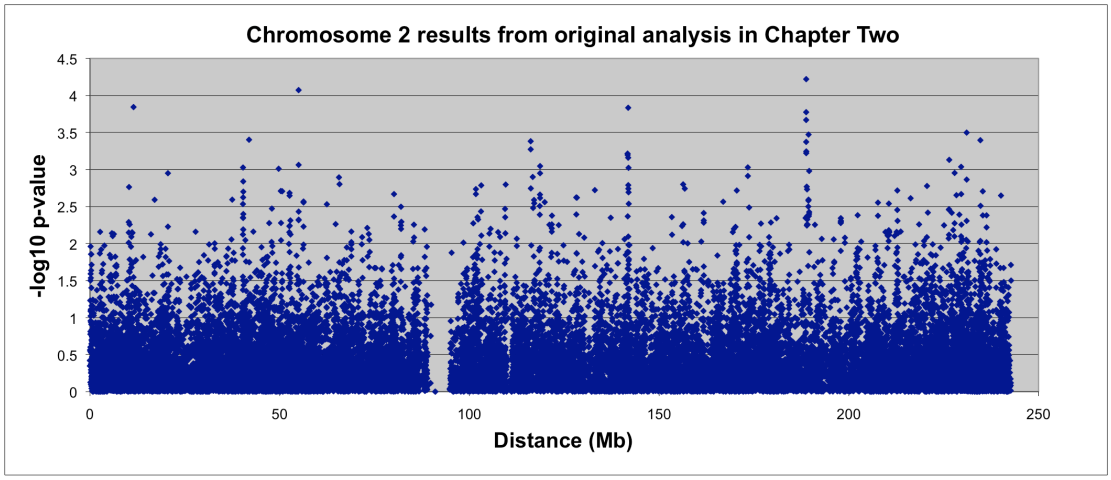




Figures 5.15 - 5.16 Results for chromosomes 9 and 14 from the original uric acid analysis performed in chapter two. The test used in this analysis was the additive (1df) test.

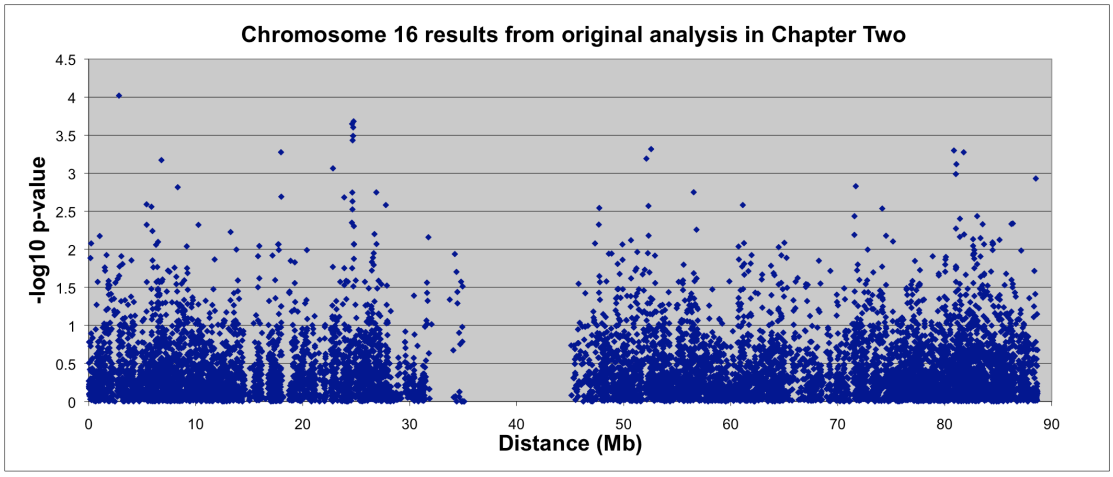
Figure 5.17 shows the results for chromosome 1 from the original uric acid analysis, and there is clearly no supporting evidence to suggest the presence of a QTL on this chromosome. Likewise, Figure 5.18 shows original analysis results for chromosome 2, where another putative QTL is located. This is the association shown in Figure 5.12 that consists of a single SNP from the SSR method, which, similar to the association on chromosome 9, has no other SNPs in high LD. There is no support for this association from any of the other methods, and also no indication of an association in the original, full dataset.

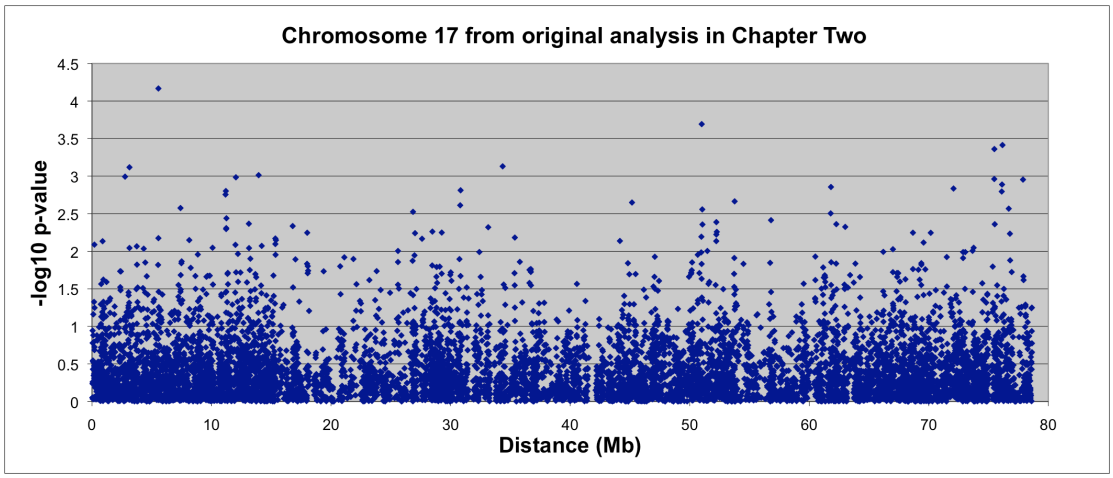




Figures 5.17 - 5.18 Results for chromosomes 1 and 2 from the original uric acid analysis performed in chapter two. The test used in this analysis was the additive (1df) test.

The remaining three putative associations are those on chromosomes 16 and 17, displayed in Figures 5.13 and 5.14. Corresponding graphs for the original analysis are shown in Figures 5.19 and 5.20, and overwhelmingly suggest that there are no true associations with uric acid as no SNPs from the original analysis attain anywhere near genome-wide significance. While both chromosomes 16 and 17 have one or a number of SNPs that stand out above background noise at the locations specified, the level of significance is too small.





Figures 5.19 - 5.20 Results for chromosomes 16 and 17 from the original uric acid analysis performed in chapter two. The test used in this analysis was the additive (1df) test.

5.3.4 Genes as putative QTL

Ensembl genome browser (Hubbard et al., 2007) was used to determine which genes the most significant hits fell in, or which genes were closest if the SNPs were not intragenic. While no results reached genome-wide significance, it may still be instructive to look at the genes underlying regions showing suggestive significance, since the Bonferroni threshold used was very stringent, particularly for the multi-marker methods. The majority of genes that contained or were near suggestive SNPs had no known function to implicate them in uric acid regulation however. Only results implicating plausible genes for uric acid regulation are subsequently discussed.

Table 5.1 shows five genes whose current known functionality does not preclude an effect on uric acid levels. The first of these genes is *PTGER3* on chromosome 1. The product of *PTGER3* is known to be an inhibitor of sodium and water reabsorption in the kidney tubules. Most uric acid reabsorption occurs in the proximal tubules of the kidney, and it has been shown that the product of *SLC22A12* (URAT1) - thought to be responsible for around 50% of urate transport - is driven by sodium-anion transporters (MyPhuong at al, 2008). An earlier study also associates high serum uric acid with an increase in proximal tubular sodium reabsorption, although only in men (Cappuccio et al., 1993). *PTGER3* may therefore have an effect on uric acid concentration indirectly via its effect on sodium concentration in kidney tubules.

GENE NAME	CHR	PUTATIVE FUNCTION
PTGER3	1p31	Inhibition of sodium and water reabsorption in kidney tubules
SLC9A2	2q12	Sodium ion transport; pH regulation
SLC9A4	2q12	Sodium ion transport; pH regulation; Rectifying cell volume in kidney cells
SLC5A11	16p12	Co-transport of D-glucose and D-xylose; Regulation of myo-inositol concentration in serum via reabsorption in proximal tubule of kidney
C17orf67	17q22	Uncharacterised protein

Table 5.1 Gene names, locations and current known functions of genes identified as putative QTL affecting uric acid.

Two more genes with a plausible effect on uric acid levels lie adjacent to one another on chromosome 2. These genes, *SLC9A4* and *SLC9A2*, are approximately 300Kb upstream of the associated SNP rs11123953, and are members of the solute carrier family 9. Both genes have a role in sodium ion transport in the kidney and are involved in pH regulation, and *SLC9A4* has also been implicated in rectifying cell volume in response to hyperosmolar-stimulated cell shrinkage. The importance of sodium has already been stated above, but pH also plays a key role in the properties of uric acid. Conversion of uric acid to urate is highly dependent upon pH, and uric acid solubility is also greatly affected by pH; the more acidic uric acid is, the less soluble it becomes. Furthermore, urate excretion is affected by the pH of kidney tubules (Fahlen and Agraharkar, 2009). It is entirely possible therefore, that these two genes can facilitate an effect on uric acid in this manner. *SLC9A4* may further influence uric acid levels through its effect on cell volume, since "extracellular volume expansion or contraction, respectively, enhances or reduces uric acid excretion though the paired movement of sodium" (Fahlen and Agraharkar, 2008).

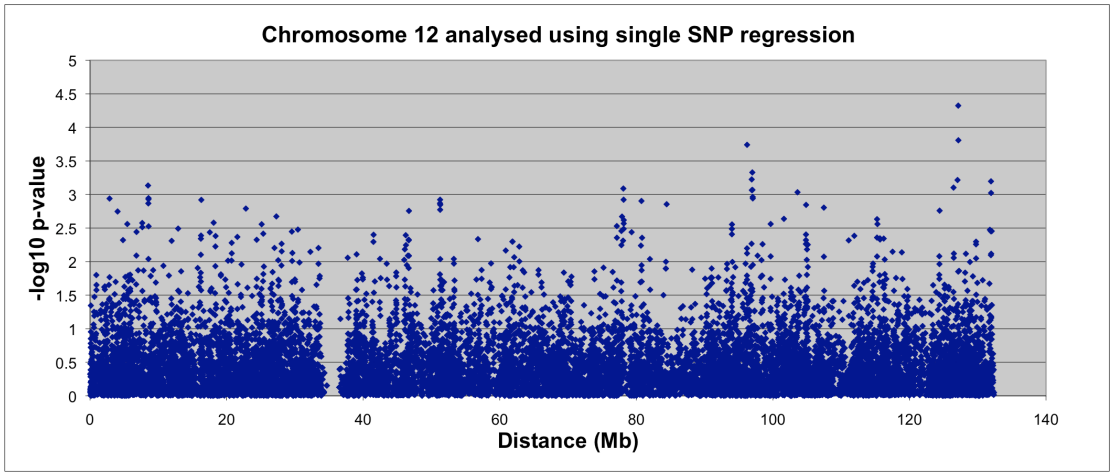
SLC5A11 on chromosome 16 (the association at 25Mb, Figure 5.12) may also affect uric acid regulation. *SLC5A11* is another gene of the solute carrier families; solute carrier family 5, member 11. The known functions of this gene include co-transport of D-glucose and D-xylose, and also a role in the regulation of myo-inositol concentration in serum, which involves reabsorption in the proximal tubule of the kidney. An effect on uric acid may be mediated indirectly as an effect of this myo-inositol regulation. Phytic acid, itself composed of inositol (inositol hexakisphosphate), is a known inhibitor of xanthine oxidase (XO), the enzyme that produces uric acid (Muraoka and Miura, 2004). If *SLC5A11* regulates inositol in the serum, this may in turn regulate the amount of phytic acid which is present, and therefore affect the amount of XO which is inhibited. It should also be noted that *SLC5A11* is a glucose transporter, which was the only role known of the *SLC2A9* gene product before its role as a uric acid transporter was discovered (Vitart et al., 2008).

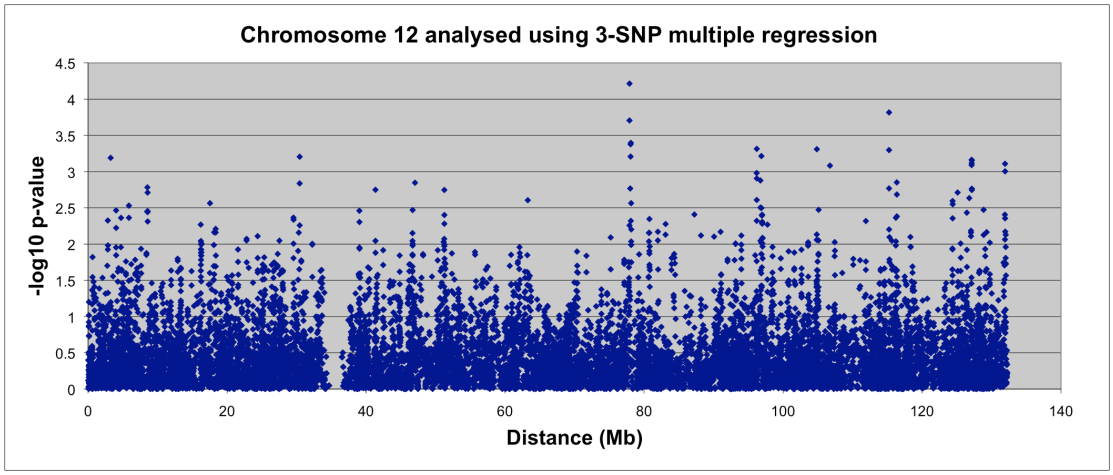
The final gene in Table 5.1 is *C17orf67* of chromosome 17, which codes for an uncharacterised protein. This does not provide any explicit support for an effect on uric acid, but it does raise the possibility of an as yet unknown function in uric acid regulation.

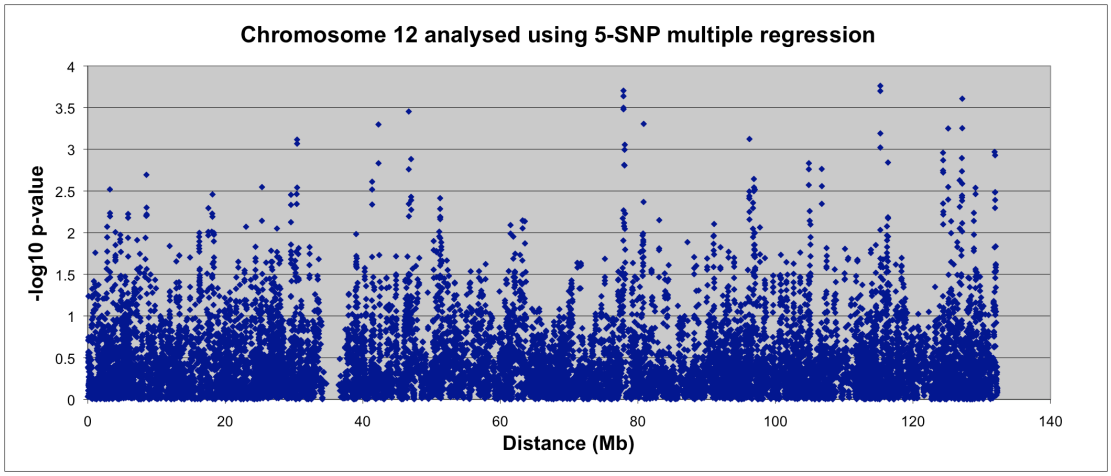
5.3.5 Comparing the methods

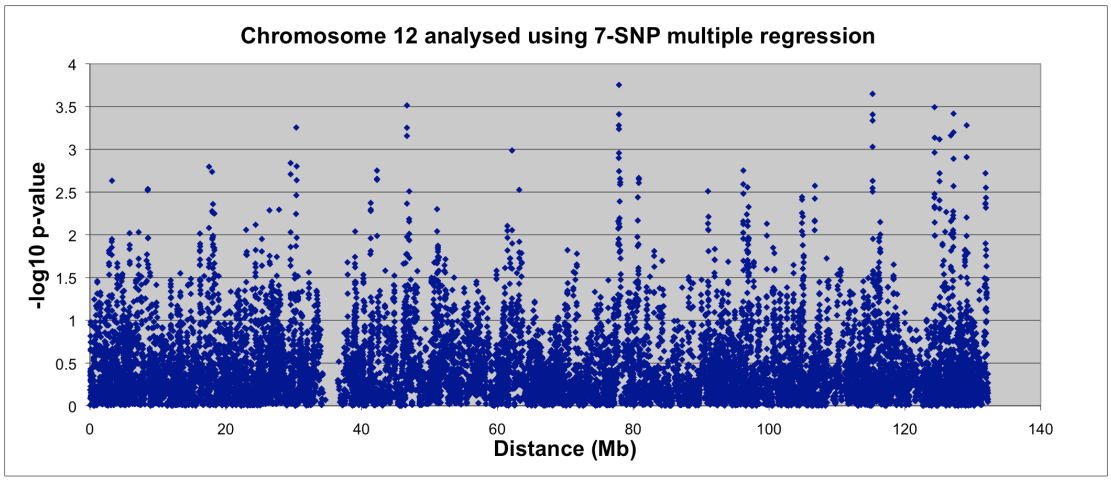
In addition to searching for putative QTL, these analyses were also performed as a basis for comparison between the seven methods used. The results indicate that the

four regression methods are highly correlated; that is, there are similar patterns of association across the methods, which are more obvious where strength of association is greater. However, as the number of SNPs in the model increases so too does the number of results exhibiting higher significance, and in addition to this the level of background noise decreases. This is illustrated in Figures 5.21.1 - 5.21.4, which show the results briefly mentioned earlier for chromosome 12 for each of the four regression methods. There are numerous (moderate) associations which become supported by a greater number of tests as the number of SNPs in the method increases. In Figure 5.21.1 there are SNPs showing weak association at regions around 80Mb, 100Mb and both before and after 120Mb into the chromosome. These associations are each found again with greater numbers of tests using MR3, MR5 and MR7, and are also made more prominent since the base level of significance for tests showing no association decreases. This latter phenomenon is likely to be a consequence of the extra degrees of freedom used in tests involving more SNPs at loci showing no association.









Figures 5.21.1 - 5.21.4 Results from single SNP regression and three-, five- and seven-SNP multiple regression on chromosome 12 for uric acid.

The pattern described above is not always the case for the haplotype analyses. For example, all associations on chromosome 12 (Figures 5.21.1 - 5.21.4) except one (just after 120Mb) disappear for the haplotype methods. Some associations are present across all methods however – the one on chromosome 14 for example. In addition to some association signals disappearing from multiple regression to haplotype analyses, other associations are found where previously there were none. For instance, there is a reasonably strong signal on chromosome 10 for HL5 and HL7 (there is a single test showing association for HL3 too), that is not present for any of the other methods. There is also a difference between the background noise levels of the methods. The base level of significance for windows showing no association appears to be less for both HL5 and HL7 compared with all other methods; closer to a value of one instead of 1.5.

5.4 DISCUSSION

5.4.1 GWA results

There were no genome-wide significant results produced in this study, however numerous tests provided suggestive significance for association. The most significant result was produced using SSR for a SNP on chromosome 9, although there was supporting evidence from a number of other methods. The SNP identified using SSR was rs1323771, and as described earlier, this SNP has no other SNPs (in this dataset) in high LD. When this is the case it can be hard to know whether the lack of LD is a consequence of local genetic architecture at the locus, or due to some other reason.

For example, it is possible for low LD to exist as a consequence of poorly clustered SNPs for which the calling algorithm was unable to accurately assign genotypes, thus breaking SNP correlations.

One way to identify SNPs such as these is to test for Hardy-Weinberg equilibrium (HWE), as SNPs where genotypes have been badly called often have a significantly different distribution of genotypes than would be expected given their allele frequencies. Another way is to examine the cluster plots of intensity data from which the genotypes were called, using software such as Evoker (Morris et al., 2010). This can be particularly important for rare SNPs, which are typically harder to call than common SNPs due to the difficulty of assigning points to three unique clusters, where one cluster may be absent or consist of very few points. Finally, it is also possible to check for poorly called SNPs by comparing levels of LD between SNPs in the study population to those in the publicly available data such as the HapMap.

A HWE test for rs1323771 on chromosome 9 had a p-value of 0.82, thus indicating that the genotype calling for this SNP may be good. Additionally, r^2 between rs1323771 and the two SNPs in highest LD with rs1323771 in this dataset is similar to that in the European (CEU) HapMap data (rs10114830 has $r^2 = 0.65$ in HapMap and 0.52 in this dataset; rs10117817 has $r^2 = 0.46$ in HapMap and 0.47 in this dataset). It should be noted that for this population HapMap may not be ideally suited, given that this population is a genetic isolate and therefore may differ in allele frequency from the HapMap population; nevertheless, the r^2 values are in general agreement. To conclusively demonstrate that rs1323771 was called accurately in this

dataset the intensities would need to be examined, however intensity data for the genotypes analysed here were not available to this study. Although poor clustering cannot be ruled out with certainty, it seems probable based on the HWE test and LD comparison with HapMap that this association is not due to an artefact.

The case is not quite so clear for the other association consisting of a single SNP association from the SSR method (rs11123953 on chromosome 2). The HWE p-value for this SNP is 0.0015, suggesting that either the calling is imperfect or a deviation from HWE exists due to population genetic factors such as selection. The MAF of this SNP is 0.047 however, and as already stated, rare SNPs are harder to accurately assign genotypes to. Additionally, r^2 between rs11123953 and rs10176694 (the SNP in highest LD with rs11123953 in this dataset) is 0.26 in this dataset and 0.25 in HapMap, suggesting that this association is also not artefactual. It should be noted that while the above associations do not appear to be due to an artefact this does not necessarily mean they represent true associations, since the associations may still be spurious. Nevertheless, the possibility also remains that these two SNPs tag true causal variants that are not in LD with any other SNPs on the panel.

The association to rs11123953 is one of the five results that implicate genes in the regulation of uric acid levels (Table 5.1). However, the fact that these results do not reach genome-wide significance means that any subsequent study would not constitute a replication in the statistical sense. In order to truly implicate these genes in affecting uric acid, it is likely that two additional studies would need to find the association; one to obtain genome-wide significance, and the other to replicate this

again. One thing to note however is that the *SLC2A9* finding originally reported in chapter two does not reach genome-wide significance here either. This means the lack of genome-wide significance for suggestive loci identified here does not preclude them being real associations. Given the small sample size for these analyses, it is doubtful the study was powered to detect many effects at a genome-wide level. However, there is also no evidence from the original analyses to back up the new putative QTL identified in this chapter. In some regards this is not too problematic; the benefit of the haplotype methods is that they should be able to detect QTL that SSR cannot (namely rare QTL) - it just leaves the problem of determining the validity of these novel associations.

One way in which the validity can be tested is to compare the results against those already published. In a recent meta-analysis of over 28,000 individuals for uric acid five novel QTL were identified (Kolz et al., 2009), adding to the four loci already known; *SLC2A9* (4p16), *ABCG2* (4q22), *SLC17A1* (6p23-21.3) and *SLC22A12* (coding for URAT1 - 11q13) (Vitart et al., 2008; Dehghan et al., 2008; Enomoto et al., 2002). Examination of results in this study from SNPs within loci known to affect uric acid reveals that none of the previously identified QTL show even suggestive significance, with the exception of *SLC2A9*. Moreover, comparisons with the largest uric acid GWA study to date (Kolz et al., 2009) indicate that none of the putative genes identified here represent true associations. Given the difference in power between the Kolz et al. meta-analysis and the study presented here, it is highly probable that if the findings were real they would have been identified in the larger study also.

5.4.2 Determining genome-wide significance

While it is tempting to speculate about the potential involvement of the genes in Table 5.1 with uric acid regulation, it is important to note that there was not enough statistical evidence to reject the null hypotheses of no association. None of the p-values for association exceeded Bonferroni significance, and even the closest were over an order of magnitude away. However, the Bonferroni correction is too stringent since it does not take into account that many tests are correlated due to the presence of LD. Methods involving multiple SNPs suffer worse from the Bonferroni correction than SSR, and more so as the number of SNPs in the method increases. This is because there is an extra source of correlation for multi-SNP methods in addition to that caused by the correlation between tests on single SNPs in LD. With three-SNP multiple regression for example, not only are tests correlated because the starting SNP for each window is directly next to (and therefore in moderate LD with) the previous starting SNP, but also because two of the SNPs within any given test were also part of the previous test. It is probable therefore that the true, empirical, distribution of the test statistic for multi-marker methods would result in a far less stringent p-value for genome-wide significance than for single SNP regression. It should be noted that this is less of a problem for the study presented here than it may be for other studies using a different genotyping platform however, since this study uses a panel designed to reduce redundancy between SNPs.

It is not possible to determine what the appropriate level of genome-wide significance for the multi-marker methods is without performing permutation analyses. Given that seven methods were used in these analyses, and that permutation is time-consuming and computer intensive for determining the appropriate threshold even for a single analysis, permutation analyses were not performed here. Consequently, some of the results presented may be a lot closer to true genome-wide significance than it would first appear. While the putative associations from these analyses cannot yet be regarded as real, it may be premature to rule them out entirely, particularly where plausible genes are implicated.

5.4.3 Comparing the methods

As already stated, directly comparing the methods is difficult when it is uncertain how much there is to find, and particularly when power for all methods is low, thus obscuring any differences. Comparisons would be made easier if the type I error rate and power for each of the analyses were known more accurately and for all methods, however determining these would have required extensive permutation, which for seven methods was clearly not feasible. Power for haplotype analyses is particularly difficult to assess since it depends on a great many parameters (many of which are unknown), not least the actual model specified to analyse the data. It is also dependent on the distribution of haplotypes and the amount of variance that the model will explain (Schaid, 2005).

Based on literature using a similar haplotype trend regression (HTR) method to the one used here, there is reason to believe that the type I error rate of the global F-test for association may be the correct size (Zaykin et al., 2002). This study used simulated quantitative phenotypes and five-SNP haplotypes to assess the type I error rate in a number of LD scenarios, and always found the size of the test did not exceed 0.05. Similar findings about the type I error of the global F-statistic were also reported in a more recent paper (Schaid, 2005). In the same study, Schaid also calculated the sample sizes required for a certain level of power, depending on other parameters such as the haplotypic distribution and amount of variance the method is able to explain (dependent upon the QTL heritability and level of LD between the QTL and marker loci). With a sample size similar to that of this study, along with the likely haplotypic distribution, it was shown that theoretical power to detect an effect explaining 5% of the phenotypic variance (estimated effect of the marker, therefore QTL effect would be greater) is less than 80% (Schaid, 2005). Effect sizes of this magnitude for common complex disease are not common in the literature to date, therefore it is unlikely many exist and remain to be found. Power to detect variants as small or smaller than 0.01 of the total phenotypic variance, however, will be negligible. It is also worth noting that for rare variants, to achieve an effect size explaining 5% of the phenotypic variance or less the effect must be extremely large on the trait mean scale.

In the absence of type I error rate and / or power calculations to guide conclusions, it is left to the known results to inform about the relative performance of the methods. The only previously known QTL for this trait that was identified at a suggestive level

was *SLC2A9* on chromosome 4. All seven methods picked up this association, although some methods did better than others. For example, HL3 was the only method to associate a window from within the *SLC2A9* gene itself with a moderate level of significance. Ordering the methods by significance of the most significant SNP / window within the *SLC2A9* gene produced the following; HL3 (6.08×10^{-6}), SSR (1.72×10^{-4}), MR5 (2.05×10^{-4}), MR7 (2.08×10^{-4}), HL5 (2.26×10^{-4}), HL7 (6.86×10^{-4}) and MR3 (9.94×10^{-4}). HL3 also produced the two most significant p-values of any method in the vicinity of the QTL location (i.e., the *SLC2A9* gene plus 100Kb in either direction), although MR3 had a result almost as significant (Figure 5.8).

It is interesting that the most significant result for HL3 within *SLC2A9* is two orders of magnitude more significant than the top SNP for SSR, because in the initial study only SSR was used (Vitart et al., 2008). This suggests that the haplotype analysis is capturing extra information included in the interactions between SNPs. In the initial analysis of uric acid, no boost in significance was required to detect *SLC2A9*, however this may not always be the case, and for other QTL using haplotypes may be crucial for detection. HL5 and HL7 did not do as well as HL3 in detecting *SLC2A9*, therefore it would appear that the extra information provided by HL5 and HL7 did not compensate for the additional degrees of freedom used in the model. Results from chapter three indicate that the causal variant at this locus may have a moderate allele frequency (thereby allowing shorter haplotypes to capture all relevant variation more easily), which would support this suggestion.

An interesting phenomenon from these analyses is how the number of results at a given significance level increases as the number of SNPs in the method being used increases. As already discussed, this is due to correlations induced between adjacent tests for the multi-marker methods since they only move on one SNP at a time. For example, a SNP showing significance when analysed singly, if incorporated into three separate tests using three-SNP multiple regression, is likely to elevate all those windows to a comparable significance level, even if the other SNPs add little or no extra information. This phenomenon is indiscriminate with regard to whether an association is real or spurious, therefore care must be exercised in interpreting results of a sliding-window technique that analyses each possible window of consecutive SNPs. This is analogous to the expectation that single SNPs in high LD show similar results, but on a larger scale; with multi-marker methods, two associated SNPs in LD would elevate the test statistic of many more windows - regardless of whether the association was real or false. In light of this, allowing multiple regression to advance more SNPs at a time, or potentially even the entire window of used SNPs, may be preferable. Alternatively, calculating an appropriate threshold empirically that accounts for the correlation would be optimal.

5.4.4 Final remarks and conclusions

Analyses in this chapter yielded no results that exceeded Bonferroni significance, therefore no putative QTL were strongly implicated. There were a number of suggestive associations, however these were not validated by the original uric acid analysis, and must therefore remain doubtful. Three-SNP haplotype analysis was able

to detect the *SLC2A9* association best, both in terms of significance and proximity to the gene, suggesting that haplotype methods may have an important role to play in the detection of QTL, and that single SNP regression is not always necessarily the most effective way of analysing GWAS data.

These analyses provide an interesting insight into the similarities and differences of a number of ways to analyse GWAS data. In order to make better comparisons of these methods, permutations to determine the empirical distribution of the test statistic under the null hypothesis would need to be performed. Power for these methods was low as a consequence of small sample size, but again, extensive simulations would be required to estimate power empirically. Theoretical power for haplotype analyses is complex to determine, and furthermore depends on the distribution of haplotypes, which will vary from one window to the next. Clearly much work is still required to determine how useful multi-marker approaches will be for GWA studies in the future, and real datasets will be an important part of this. One useful strategy may be to use large datasets including associations already known to exist, in order to see how well multi-marker methods perform.

6. CHAPTER 6 - DISCUSSION

6.1 GWA Studies: Past Successes and Current Status

Genome-wide association (GWA) studies are currently the most popular way of searching for quantitative trait loci (QTL) involved in common complex disease. One of the main reasons for this is the ability of GWA studies to exploit the mass of human genetic variation data made available through projects such as HapMap (The International HapMap Consortium, 2005; The International HapMap Consortium, 2007), and consequently the ability to use indirect mapping approaches lacking pre-defined and often constrained hypotheses. While candidate gene studies can still play an important role in QTL identification, the agnostic philosophy of GWA studies makes them very powerful since hypotheses are no longer dependent upon prior knowledge. Also, due to their genome-wide nature, these studies are no longer specific to any one trait (by virtue of candidate genes), meaning that analysis of vast numbers of traits in one study is possible, acting to further enhance their utility.

Recent literature is filled with GWA studies reporting discovery of novel QTL affecting complex diseases and their risk factors. As of March 2010 there were a total of 779 published genome-wide association study (GWAS) hits exceeding a genome-wide significance threshold of 5×10^{-8} across 148 traits (Hindorff et al., accessed 04/06/2010), and the rate of progress since the early days of GWA studies is dramatic. For example, in 2008 there were 20 QTL known to be involved in each of Crohn's disease and type 2 diabetes, and 40 for height. In 2006 these numbers were

just two, three and zero respectively (Donnelly, 2008). This highlights just how far the field has moved, and how quickly the great leaps forward in QTL detection for genetic traits have been achieved.

While intellectually rewarding to understand the genetic mechanisms affecting any heritable trait (not least because this may in turn lead to better understanding of other traits), of more consequence from a public health perspective is to determine the genetics underlying common complex diseases. Detecting disease-related QTL is of great importance to public health both in terms of prediction and prevention of disease, and also to potentially treat the disease after diagnosis. As a result, QTL identified as affecting intermediate phenotypes are frequently treated as candidate loci to analyse with respect to related disease phenotypes. This is a useful strategy, however it does not always lead to successful association with disease since the QTL may have an extremely small odds ratio for disease if it is only one contributing factor of many, and sample size thus is too low to confer significance.

Proof of principle regarding the utility of GWA studies, particularly in conjunction with intermediate phenotypes, was provided by analyses in chapters two and three of this Thesis. Complex diseases were not analysed directly, instead underlying quantitative traits (QTs) were used, which provided greater power for the analyses. Although it is likely this study was underpowered to detect very small QTL as a consequence of a small effective sample size (both due to low number of samples and because many of these samples were related), one novel QTL was nevertheless identified with this population. Prior to this study, *SLC2A9* was a known glucose and

fructose transporter, but has now been found to preferentially transport uric acid (Vitart et al., 2008). Crucially, *SLC2A9* was then also found to have an effect on gout, a disease associated with (although not necessarily a consequence of) high uric acid levels, in a much larger study (Dehghan et al., 2008). Not only does this demonstrate that QTL of moderate effect can be detected using studies with limited power, but also that the strategy of detecting QTL affecting disease through the use of intermediate phenotypes is valid. The lower heterogeneity and a greater signal to noise ratio for QTL underlying intermediate phenotypes allows easier identification of these loci.

6.2 Missing Heritability

Despite the early success of linkage, and the more recent success of GWA studies, in detecting QTL involved in common complex disease, it is becoming increasingly evident that there is a problem: large proportions of the genetic contribution to common disease remain hidden. As more QTL are discovered it is becoming clear that for the majority of complex traits, irrespective of the total number of QTL found to date, the total proportion of trait variation that the known QTL explain is much less than the trait heritability. This fact is particularly worrying because for a reasonable number of traits, all large or moderate effect size genes are likely to have already been identified by virtue of well-powered GWA studies with large sample sizes. A much-used example is that of height, a trait with 80 – 90% heritability, where a recent study took the number of known QTL past 40, yet altogether these loci only account for just over 5% of the total genetic variation (Gudbjartsson et al., 2008).

This situation is now becoming the rule rather than the exception, and as a result there is much debate over where the remainder of the genetic portion of phenotypic variance is hiding, and why current GWA studies are unable to account for it (Maher, 2008). There will inevitably be a number of variants of such small effect that studies of any reasonable sample size would still fail to detect them. However, given that the effects would need to be small, for these variants to be the whole explanation for the missing heritability, the number of such variants would need to be immensely large. There are a variety of alternate explanations for where parts of the missing heritability may be found however.

6.2.1 Rare variants

One of the current hypotheses concerning the missing heritable variation of common complex disease suggests that rare variants are responsible (i.e., causative variants with a MAF under 5% - or sometimes even more stringently, under 1%). Variants that are rare are not only difficult to discover in the first place, but they are also problematical to genotype, and furthermore, provide statistical problems when it comes to analysis. Historically, these variants have therefore not been well represented in either candidate gene or GWA studies. However, suggesting that rare variants are responsible for the remaining phenotypic variation is contradictory to the theory typically thought to describe the underlying architecture of human complex disease, the so-called common disease / common variant (CD/CV) hypothesis.

The CD/CV hypothesis posits that there are a moderate number of genes affecting any given common complex disease, and that these have an appreciable population minor allele frequency (Reich and Lander, 2001). One particularly attractive property of the CD/CV hypothesis is that under this model of disease common disease variants are able to explain a large proportion of disease prevalence even with only a small effect on disease risk (Gorlov et al., 2008). There is some theoretical evidence (Reich and Lander, 2001), and also a vast body of empirical evidence in the literature to support the CD/CV hypothesis, however this is unsurprising given that most discoveries to date are from studies only powered to detect QTL of moderate minor allele frequency (Hirschhorn et al., 2002).

As implied above, there are many reasons why the literature may be enriched with associations to common QTL. For instance, the majority of genotyping panels that GWA studies are based on explicitly only include common SNPs. Therefore, any association detected using these panels by definition has to have been identified by a SNP that was common in the initial discovery dataset (thus also likely to be common in the study population), and therefore is highly likely to be tagging a causal variant that is also common. While it is possible for a SNP to have low frequency only in the study population, this situation would be rare since most GWA studies remove all SNPs below a certain threshold (usually 2%) in their population to begin with due to the afore-mentioned genotype-calling and statistical analysis problems associated with rare SNPs (Gorlov et al., 2008).

It would also be possible for the causative variant driving a SNP association to be rare itself, however few rare QTL would be found in this way since power is dependent on allele frequency and the LD between marker and QTL alleles, which in general will be low between common SNPs on marker panels and rare QTL. This explains why GWA studies are generally only sufficiently powered to detect variants that are not rare, and therefore why the literature is biased towards reporting common QTL. Given the loss of power due to poor tagging of rare (untyped) SNPs by common SNPs on genotyping panels, the sample size required to detect slightly rare variants (i.e., those between 2-5% MAF) would need to be considerably greater than those of most current studies, while variants even more rare would never be detected using single SNPs (Bodmer and Bonilla, 2008).

The CD/CV hypothesis is not the only theory pertaining to the allelic spectrum of common complex disease. One of the original alternatives to the CD/CV hypothesis was the genetic heterogeneity model, which states that for any given disease, multiple rare alleles exist at numerous loci, all of which are fully penetrant and therefore disease-causing if present (Smith and Lusis, 2002). This model fits well with the breast cancer genes *BRCA1* and *BRCA2* for example (Wang and Pike, 2004), however it is quite different to the CD/CV hypothesis in that each variant essentially acts in a Mendelian fashion with respect to disease, which is certainly not the case for all QTL. Another suggestion is that the allelic spectra of variation involved in complex disease is no different to the allelic spectrum of the entire genome, i.e., the MAF frequency distribution of disease-affecting QTL over the genome as a whole is

U-shaped and has large numbers of rare alleles but much fewer alleles at intermediate MAF (Wang and Pike, 2004).

A similar version of this latter theory is currently gaining much popularity however – the common disease / rare variant (CD/RV) hypothesis. In contrast to suggesting that the MAF distribution of disease-influencing variants is similar to that of the genome as a whole, the CD/RV hypothesis posits that a much larger proportion of the inherited susceptibility to common disease is due to the effects of independent low frequency variants (Bodmer and Bonilla, 2008). This theory is less extreme than the genetic heterogeneity theory however, since in the CD/RV hypothesis each rare variant confers only a small to moderate increase to disease risk, instead of being fully penetrant.

There is theoretical evidence to support the CD/RV hypothesis. Using stochastic modelling, one study found that for common complex disease "it is unlikely that any single mutation will constitute a large fraction of the susceptible class" when the mutation rate was at the upper end of its predicted range (Pritchard, 2001). Another study concluded that if common diseases were caused by multiple loci then a diverse allelic spectrum with rare causal alleles is expected (Peng and Kimmel, 2007). Additionally, there is an increasing volume of empirical evidence supporting the CD/RV hypothesis; see for example LpL gene function (Nickerson et al., 1998), numerous phenotypes in a follow up study (Crawford et al., 2004), and cystic fibrosis (Bobadilla et al., 2002). One recent paper also points out that GWA studies appear to be reaching their limit in identification of common variants affecting complex human

disease, and that rare variation is one likely source of new QTL that is as yet largely unexplored (Schork et al., 2009).

Another recent study in particular greatly enhances the argument for a considerably larger role in the development of complex diseases for rare variants than previously assumed. Using all SNP data, therefore including rare (<5% MAF) SNPs, from both ENCODE (ENCODE Project Consortium, 2004) and HapMap (Thorisson et al., 2005), the study analysed the proportion that were predicted to be functional in each of 20 equally-sized MAF bins (Gorlov et al., 2008). The study reported three findings; there was a higher prevalence of non-synonymous SNPs in low MAF bins, there was evidence for purifying selection at the lower MAF bins, and there was an inverse correlation between the proportion of predicted protein damaging SNPs and MAF (Gorlov et al., 2008). This last conclusion was in agreement with two other studies reporting the same negative correlation (Cargill et al., 1999; Wong et al., 2003).

It should also be noted that, as predicted from the expected U-shaped distribution, rare SNPs constitute proportionally by far the larger of the two groups (common or rare); around 50% of the SNPs in ENCODE were rare, and 38% of the HapMap SNPs, despite the space for being declared rare encompassing only 10% of the total. Furthermore, ENCODE sequencing was based on a limited number of individuals, meaning that 50% is likely an underestimate of the total proportion of rare SNPs in the genome (since large numbers of rare SNPs, by virtue of being rare, will not be present in limited samples of individuals). Consequently, this may have resulted in

observing a lower prevalence of non-synonymous SNPs in low MAF bins than really exists.

Gorlov et al. also suggest that there may be as many as two or three "slightly deleterious" SNPs per gene in the genome, even ignoring regulatory regions, and that their effects on fitness in the population are low enough to allow them to persist due to mutation-selection balance. The proportion of disease in the population caused by any given variant (which can be expressed as the population attributable fraction (PAF)) is directly related to frequency of the variant; as MAF increases so does PAF. For example, it has been estimated that a MAF of 1% is around the upper limit attainable by clearly deleterious alleles sustained by mutation-selection balance (Bodmer and Bonilla, 2008). This may suggest that rare variants (in the 1-5% range) constitute a reasonable proportion of these slightly deleterious SNPs, since they can act in a largely evolutionarily neutral manner by virtue of their small PAF (Gorlov et al., 2008). The fact that each variant contributes little at the population level would be offset by the fact that so many of these QTL exist in the genome, allowing rare variants to cumulatively have an appreciable effect on disease prevalence. Needless to say, if this is truly the case, there is much work still to be done before we are in a position to detect most of these variants.

One way in which rare variants may influence disease risk is for some number of individually rare, non-synonymous, mutations within a gene to each disrupt gene function, as opposed to a single more common mutation (Schork et al., 2009). This gene perturbation effect has been witnessed in “driver” cancer mutations for example

(Wood et al., 2007). Techniques for detecting rare variants involved in complex disease however are currently very few and in their infancy, due to the much greater focus on detecting common variants. Presently, the most popular method of detecting rare variants is based on the candidate gene case-control approach, since it requires the cohort to be fully sequenced in the regions of interest. This identifies all variants (rare or otherwise) within the candidate region in the study cohort, and tests for significant allele frequency differences between cases and controls. This can be performed either for each rare SNP singly, or for groups of rare SNPs within a given region (Schork et al., 2009). Potential functional consequences of a mutation can be used to help identify disease-related loci, for example by virtue of location within a conserved sequence, by causing a non-synonymous mutation, or by causing an amino acid charge change (Bodmer and Bonilla, 2008).

Candidate gene studies currently need to sequence regions of interest to capture rare variation, which highlights one of the greatest challenges to our ability to uncover rare disease-related genetic variation. The dominance of the CD/CV hypothesis until recently inevitably led to genotyping companies, and hence GWA studies, focussing almost entirely on common variation, and this has been at the expense of rare SNPs (Gorlov et al., 2008). Even the rare genetic variation which has so far been uncovered from projects such as ENCODE and HapMap only tell part of the story, since the population sizes are fairly small, and, in the case of ENCODE, do not cover the entire genome. In order to detect rare variants influencing disease, first there must be further attempts to catalogue the extent of rare variation present in the human genome. At present very few sources of information provide the level of coverage sufficient to

find large amounts of genome-wide rare variation. Studies sequencing the whole genome of specific individuals are useful (see for example Levy et al., 2007), but this would need to be performed on a much larger scale to be truly informative. One such study has already discovered *de novo* mutations which are functional (Ng et al., 2008), therefore prospects for identifying more functional *de novo* and rare variants seem good. The 1,000 genomes project, which aims to characterise sequence variation in 1,000 individuals should greatly facilitate the discovery and utility of rare SNPs in disease mapping in the near future (www.1000genomes.org/).

In the absence of detailed information about the extent of rare human genetic variation and also the absence of methods to effectively use this information to detect QTL, at the current time knowledge of many rare disease-influencing loci may be unobtainable. Results in chapter four of this Thesis strongly suggest that better use of current information and readily available tools would greatly enhance the chances of detecting rare disease-influencing loci however, since haplotypic analysis was shown to perform vastly better than either single SNP or multiple SNP regression when the sQTL in question had low MAF. There are also other situations in which haplotypes will vastly outperform single SNP regression however, for example where multiple rare variants affecting disease exist in a single gene. Long haplotypes each associated with a single rare variant could thus capture all variation at the locus, providing power to associate the gene with disease. This would require testing in a variance components style analysis as opposed to fitting each haplotype as a fixed effect as in chapter four, to avoid the number of degrees of freedom used in the test becoming prohibitively large. Given the mounting evidence to indicate an important role for

rare variants in common disease, and that rare variants are so ubiquitous in the genome, haplotypes may thus have a crucial part to play in understanding the genetic architecture of complex disease.

6.2.2 Structural variation

Rare SNPs are just one of the possible ways in which the unexplained variation in heritable traits may be hidden. Another potential source of missing heritable variation concerns a different type of genetic variation altogether, indeed variation on a totally different scale from that of a single base pair. This variation is known collectively as structural variation, although the name encompasses many distinct types of genetic variation. The presence of very large structural variants in the genome has been known since chromosomes were first examined under the microscope, but these observable variations were originally limited to phenomena such as aneuploidy (the presence of an abnormal number of chromosomes – such as in Downs' Syndrome for example) or large unbalanced mutations (resulting in a net loss or gain of DNA, unlike translocations) such as insertions, deletions or duplications.

Microscopic structural variation of this sort generally involves stretches of DNA upwards of 3Mb in size. At the other extreme, much is known about genetic variation at the single nucleotide level through SNPs, and also variation up to around 1Kb through mini- and microsatellites. However, until recently very little was known about the genetic variants occupying the middle ground of variation encompassing between 1Kb and 3Mb of DNA, and consequently any possible effects of these on

disease susceptibility. Recent technological advances have enabled detection of these sub-microscopic variants, and early studies have discovered a surprising amount of structural variation that was previously unknown (Iafrate et al., 2004; Eichler et al., 2007; Kidd et al., 2008). Sub-microscopic variation consists of all the same types of variation that exist on a larger scale; insertions, deletions, translocations, transversions and duplications among others. Another type of sub-microscopic variation has also been found however, and is surprisingly abundant in the genome. These new variants are analogous to microsatellites, only much larger, and have been called copy number variants (CNVs).

Copy number variants are defined as segments of DNA of 1Kb or larger that are present at variable copy number in comparison to a reference genome (Feuk et al., 2006), therefore classes of CNV can also include insertions, deletions and duplications. Early studies based on fairly small segments of the genome found an average of 12 CNVs per genome (Iafrate et al., 2004), however since then research has suggested there may be over 100 CNVs per genome of at least 50Kb in size, and many more of a smaller size (Feuk et al., 2006). Another study, using the HapMap data, found almost 1,500 large CNV regions covering 12% of the genome (Redon et al., 2006). One of the largest studies looking at smaller-scale copy number variation to date, also using the HapMap data, found that a comparison of any two genomes yielded almost 1,100 CNVs of less than 1Kb. Unsurprisingly then, studies also report that many novel CNVs are found when comparing these genomes to the human reference assembly, NCBI36. For example there are 23 novel CNVs in the genome of James Watson (Wheeler et al., 2008), and 34% of the structural variants in a

sequenced Asian individual were also novel (Kidd et al., 2008). This suggests that large numbers of individuals would need to be analysed to fully capture the diversity of structural variation in the human genome (Henrichsen et al., 2009).

There are already reports that CNVs and other sub-microscopic structural variants can affect gene expression, Mendelian disorders, and common complex disease (Conrad et al., 2009). For example, lower than average copy number of the chemokine receptor gene *CCL3L1* is associated with markedly higher susceptibility to HIV infection (Gonzalez et al., 2005), and numerous associations of CNVs and other structural variants to various related disorders such as schizophrenia, autism and mental retardation also exist (Henrichsen, 2009). One study found that genes harbouring CNVs with an effect on gene dosage (i.e., expression level) were enriched for genes involved in immune response and responses to biotic stimuli (Feuk et al., 2006), and another found that the CNVs they identified overlapped with 13.4% of genes, and altered the structure of 12.5% of transcripts (Conrad et al., 2009). Studies have also found evidence for selection on CNVs (Perry et al., 2007), and also on other types of sub-microscopic variation, for example a 900Kb inversion on chromosome 17 (Stefansson et al., 2005).

It appears likely that more QTL consisting of CNVs and other types of structural variation will exist, and as more about the distribution of these variants becomes known it will be important to investigate whether this is the case. Indeed, on a nucleotides-per-genome basis CNVs alone encompass more DNA than SNPs, highlighting their potential importance in understanding the genetics behind complex

disease (Redon et al., 2006). Recent work also suggests the possible existence of "CNV hotspots", which may provide a starting point to look for association to disease (Lee et al., 2008).

In the short term it may be possible to test a subset of CNVs for association with disease by virtue of proxy SNPs in high LD, but this approach may not be optimal. There was initially much debate as to what extent SNPs and CNVs co-locate; initial studies in the mouse found that CNVs are significantly enriched among sequences with low and moderate SNP coverage (Cutler et al., 2007), however the more extensive study performed by Conrad et al. suggests that common (>5%) CNVs of around 1Kb are particularly well tagged by the set of HapMap SNPs (77% of common CNVs tagged at r^2 of at least 0.8). An even more recent and comprehensive study conclusively indicates that 2- and 3-class CNVs (i.e., CNVs that are biallelic) are tagged well by SNPs on the Affymetrix 500k, Affymetrix 6.0 and Illumina 1.2M arrays; 79% with MAF > 0.2 have $r^2 > 0.8$ with at least one SNP, and 22% with MAF < 0.05 have $r^2 > 0.8$ with at least one SNP (The Wellcome Trust Case Control Consortium, 2010). The study also suggests that these biallelic CNVs are the predominant type of CNV in the genome, comprising 88% of their dataset after quality control, although it should be noted that this may represent an upwards bias due the ability of existing CNV-discovery methods to reliably identify this type of CNV above others.

It would thus appear that a substantial proportion of copy number variation is captured indirectly through LD to SNPs. However, even if the majority of CNVs

present in the genome are essentially biallelic markers and are tagged to the same extent as untyped SNPs by existing SNP panels, this means that CNVs will also suffer from the same shortfalls as SNPs. Rare CNVs would still not be particularly well tagged, and ideally a method of testing each variant for association with disease, whether directly or indirectly, is required. Typing rare SNP variation may help capture a greater proportion of total copy number variation, but ultimately there will also be CNVs missed by relying on SNPs. Contrastingly, if the estimated proportion of biallelic CNVs has been vastly inflated by the study above, then there is the even greater challenge of designing new methods for both discovery and testing of these multi-allelic markers.

6.2.3 Epistasis and gene-environment interactions

While both rare SNPs and sub-microscopic structural variants may account for some of the missing heritable variation of common complex disease, it is still unlikely they will explain it all. One of the likely alternative sources of missing variation is interactions, and these can be of a gene-by-gene (i.e., epistatic) or gene-by-environment nature. Knowledge of the existence of gene-by-gene interactions is not new; the term "epistasis" was first used by William Bateson to describe deviations from Mendelian inheritance over 100 years ago (Phillips, 1998). At the time, Bateson was referring to what is now known more generally as biological epistasis (i.e., masking of an allele at one locus by expression of alleles at another locus), although more recently the focus has turned to identifying statistical epistasis between polymorphisms (Moore and Williams, 2009). Statistical epistasis was defined by

Fisher as an explanation for departure from additivity in linear models (Fisher, 1918), and the distinction between the biological and statistical definitions of epistasis is important with regard to interpretation of the recent GWA studies into gene-by-gene interactions. One recent review has defined terms for three different types of gene-by-gene interaction in an attempt to distinguish concepts that are often confused under the ambiguous title of epistasis (Phillips, 2008).

Statistical interactions between markers can be explicitly modelled and tested in GWA studies. From a practical point of view however, GWA studies are generally unable to consider all potential two-way marker interactions (much less higher order interactions) as this causes a massive multiple testing problem in addition to the extra computational demand (Moore and Williams, 2009). The approach is more feasible in candidate gene studies however, where the number of markers is significantly less and multiple testing becomes less of an issue. There are also others ways to go about detecting (statistical) epistasis that reduce the search space for gene-by-gene interactions, and thereby avoid exhaustive genome-wide tests.

One of the ways to detect gene-by-gene interactions in GWA studies without incurring a prohibitive multiple testing burden is to only test for interactions between SNPs already exceeding significance in a test of marginal effects. It should be noted however, that for QTL exhibiting significant gene-by-gene interactions any independent marginal effects may not necessarily be large, therefore the level of significance required to take forward loci for interaction analysis is typically less than genome-wide. This approach vastly reduces the number of tests performed, however

it will miss epistatic effects where one or both of the genes involved do not have marginal effects (Cordell, 2009). It is currently far from clear that all statistically interacting loci do have marginal effects, for example, genes interacting with no marginal effects were discovered for susceptibility to cancer in mice (Fijneman et al., 1996). As is often the case with in human genetics however, the problem of detecting epistatic effects is exacerbated by the fact that more complex study designs cannot be used; the study of cancer in mice cited above was performed using a study design that could not have been applied to humans (Jannot et al., 2003).

It is currently unknown how important the role of gene-by-gene interactions may be in the control of complex traits. This is one area of research that holds promise, but at present may be difficult to get to grips with due to the high dimensionality of current GWA datasets, and the lack of an appropriate way to systematically test for these interactions (Moore and Williams, 2009). Consequently, there are few reports of statistical interactions in the literature to date, although some do exist. For example, epistasis has been discovered at QTL affecting coronary artery disease (Tsai et al., 2007), diabetes (Wiltshire et al., 2006), and multiple sclerosis (Gregersen et al., 2006). For many of the QTL with interaction effects that have been detected however, the functional basis of the interaction has not been identified, highlighting the challenge of interpreting statistical interaction, and relating it to the underlying biology (Phillips, 2008).

While the biological interpretation of epistatic interactions can often by obscure, this is not always the case with gene-by-environment (G×E) interactions. Similar to

epistasis, G×E interactions represent the joint effect of a gene and an environmental factor above that of the marginal effect of the gene individually, and were originally tested for primarily in candidate gene based studies. Testing for G×E interactions may be important for a number of reasons, the foremost of which is to establish a greater understanding of the biological mechanisms and pathways that underlie common complex disease. Other potential uses for knowledge of G×E interactions are to increase understanding of heterogeneous results across different studies of the same trait (resulting, for example, in lack of replication), better characterisation of the individual components constituting any given environmental effect (for example, air pollution - Hunter, 2005), and identifying environmental factors that may adversely affect subgroups of individuals (Thomas, 2010). For these reasons, more interest has been shown in detecting G×E interactions in recent GWA studies. One of the major difficulties of searching for these interactions however, is that very often GWA studies do not collect extensive data on environmental factors.

In addition to the frequent lack of appropriate information with which to test for G×E interactions, the best study designs for detection of these effects are different to those of a normal GWAS (Le Marchand and Wilkens, 2008). For example, sample-size requirements to have reasonable power to detect G×E interactions are suggested to be at least four times that needed to detect main effects of a similar magnitude (Smith and Day, 1984). This means tens of thousands of individuals would be required to detect G×E interactions using the typical study design that is pre-eminent in GWA studies at the present time, and would thus necessitate large consortia dedicated to finding them (Thomas, 2010). Using large consortia raises issues regarding

consistency across each of the study participants however, as otherwise there is the potential to introduce sources of heterogeneity. An additional problem regarding G×E effect detection in general is that the majority of environmental factors have exposures that vary over time, and most will also be dependent upon features such as duration of exposure or age at exposure (Thomas, 1988). Also, some environmental factors are comprised of multiple components, each of which could be measured and analysed individually for better interpretation and understanding of the underlying pathways. One such example is air pollution, as mentioned above, which is comprised of a mixture of gases and particles (Hunter, 2005).

These difficulties have lead some to suggest that searching for G×E interactions may not be worthwhile (Clayton and McKeigue, 2001). Nevertheless, examples in the literature do exist to indicate that G×E interactions can have important enough effects on the prevalence of common complex disease to be of value for study. For example, there is the effect of smoking and the *NQO1* gene on risk of lung cancer (Xu et al., 2001), folate status and the *MTHFR* gene on colorectal cancer (Le Marchand et al., 2005) and multiple genes showing environmental interactions for posttraumatic stress disorder (Koenen et al., 2009). Clearly there is an important role of G×E interactions for some common complex diseases, and this is one direction that should not be ignored. Future challenges for the advancement of G×E effect detection include finding more robust methods of testing for G×E interactions in genetic studies that do not compound the problem of multiple testing, and more detailed exposure assessment of the environmental factors that are studied.

6.2.4 Expression QTL

Another source of variation contributing to complex traits is what are known collectively as expression QTL (eQTL). eQTL are genomic loci that regulate gene expression, therefore mutations at these loci lead to altered transcript levels of the genes they regulate. These changes can be detected by measuring transcript levels (using DNA microarrays, for example), and subsequently treating gene expression as a quantitative trait to analyse in an identical way to conventional QTs. eQTL can be either *cis*- or *trans*-acting, reflecting their proximity to the genes they regulate, and are also classified as "static" or "dynamic" depending on whether they are consistently active, or cell-type-specific (Gerrits et al., 2009).

There is evidence to show that eQTL play a role in the control of complex human disease. For example, eQTL affecting childhood asthma (Moffatt et al., 2007), age-related macular degeneration (Coleman et al., 2008) and Crohn's disease (Barrett et al., 2008) have already been identified. Another study has even suggested that many of the trait-associated SNPs already identified through GWA studies are likely to be eQTL (Nicolae et al., 2010). In their study, 625 of 1,598 confirmed trait-associated SNPs taken from the GWA study catalogue (Hindorff et al., accessed 29/06/09) would have been classified as eQTLs with a p-value of 1×10^{-4}. Thus, current challenges of considerable importance are to assess the extent of genes acting in an expression-related manner, and detecting the genetic determinants of this expression variation. Recent studies have begun gathering genome-wide expression data and providing publicly available database resources for the research community to interrogate with regard to trait-associated SNPs identified through GWA studies (see,

for example, Dixon et al., 2007). Projects like these are essential to begin addressing the challenges mentioned above, and herald the next wave of trait-associated variants that advance our understanding of the genetic architecture of complex traits.

6.2.5 Lost in Diagnosis

One other possible explanation for not finding all heritable components of phenotypic variation is phenotypic misclassification. This is more problematic for some traits as opposed than others, but can greatly hinder attempts to detect genes involved in such traits. Inclusion of individuals suffering from slightly different diseases, or not suffering from disease at all would add noise to a study and decrease power. Such misclassifications are possible for example in threshold traits where diagnosis is dependent upon instruments introducing a degree of error, or in cases where diagnosis is a result of the subjective decision of a doctor. Also, due to the very large sample sizes of modern GWA studies, it is not uncommon that samples for a single study are collected from different centres, thus relying on the opinions of different doctors and / or measuring apparatus.

In the case of misclassification of one disease for another, power is decreased because the genes causing disease in some individuals will not be the same as those causing disease in others, and this will obscure any association (Maher, 2008). This effect is similar to genetic heterogeneity with multiple loci exhibiting low penetrance, except that in the case of genetic heterogeneity at least all disease alleles are consistent across the sample. Note that while misclassification would also affect heritability

estimates in the given study (therefore meaning the heritability was never actually present to be found missing in that study), poorly collected study data may nevertheless be one reason that some of the true genetic proportion of complex disease cannot be found. Ironically however, in order to better diagnose samples for GWA studies, it may require an understanding of the genetic mechanisms and pathways behind complex disease in order to categorise them better.

6.3 The Future of Complex Disease Mapping

Even with ever-increasing sample sizes and new study populations, the continuing success of GWA studies can only last so long. As already described, there is still much to be learned about the genetics of complex disease, and it appears that many of the upcoming challenges cannot be tackled using existing methods. While rare SNP detection continues in projects such as the 1,000 genomes, and other studies elucidate the full extent of CNV and other sub-microscopic structural variation in the human genome, it is important to make the best use possible of data already available. There is much debate over the benefit of using SNPs in new ways in an attempt to glean extra information from what already exists, and results presented in this Thesis add to that currently in the literature suggesting that utilising multi-SNP methods, in particular haplotype analysis, may do just that. Results from chapter four show strong evidence that haplotypes greatly assist the detection of rare variants, and these are looking to have a more prominent role in disease than previously thought.

For the longer term, the most important aim must be the construction of more detailed maps to characterise rare SNP variation, CNVs and other sub-microscopic variation present in the genome. This is going to require the in-depth sequencing of large numbers of individuals, assuming that results from the first attempts at cataloguing this information are accurate, and may be a particularly important challenge with respect to CNVs, given that at the present time much work is focussed on CNV discovery rather than genotyping (Conrad et al., 2009). However, once this information has been gathered the prospects of taking our understanding of complex human disease to a new level are excellent. As our knowledge of individual diseases increases, it is possible that ever widening networks of genes will be discovered to be linked, and that many underlying genetic pathways may be shared between diseases that now seem disparate. Integral to this broadening of knowledge will be expanding our understanding of the ways genes interact both with each other and the environment, and also how other phenomena such as epigenetics (through mechanisms such as DNA methylation and chromatin remodelling, for example) act to affect disease aetiology.

Nearly 75 years on from his prophetic statement, it seems unlikely that Ronald Fisher could have imagined just how long the search for "linkage" would indeed turn out to be. However, if the search for linkage was at first disappointing, the advancement in understanding the genetic architecture of both Mendelian and common complex disease in recent years, and the promise of greater discoveries to come, are anything but that.

References

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

Affymetrix website; www.affymetrix.com

Akey, J., Jin, L. and Xiong, M. (2001) Haplotypes *vs* single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics,* **9**:291-300

Amos, C. I. and de Andrade, M. (2001) Genetic linkage methods for quantitative traits. *Statistical Methods in Medical Research* **10**:3-25

Amos, C. I. and Elston, R. C. (1989) Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genet. Epidemiol.* **6**:349-360

Augustin, R., Carayannopoulos, M. O., Dowd, L. O., Phay, J. E., Moley, J. F. and Moley, K. H. (2004) Identification and characterisation of human glucose transporter-like protein-9 (GLUT9): alternative splicing alters trafficking. *The Journal of Biological Chemistry,* **279(16)**:16229-36

Aulchenko, Y. (2007) GenABEL: genome-wide SNP association analysis. R package version 1.1-8.

Aulchenko, Y., de Koning, D-J. and Haley, C. (2007) Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. *Genetics,* **177**:577-585

Aulchenko, Y., Ripke, S., Isaacs, A. and van Duijn, C. (2007) GenABEL: an R library for genome-wide association studies. *Bioinformatics*, **23(10)**:1294-1296

Bader, J. S. (2001) The relative power of SNPs and haplotypes as genetic markers for association tests. *Pharmacogenomics,* **2(1)**:11-24

Balding, D. J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics,* **7**:781-791

Barrett, J. C., Fry, B., Maller, J. and Daly, M. J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21(2)**:263-265

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitten, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M.

T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J.; NIDDK IBD Genetics Consortium, Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J. P., de Vos, M., Vermeire, S., Louis, E.; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium, Cardon L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. J., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M. and Daly, M. J. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, **40**:955-962

Bateson, W., Saunders, E. R., and Punnett, R. C. (1905) Experimental studies in the physiology of heredity. *Rep Evol Com R Soc* **II**:1-131

Becker, T. and Herold, C. (2009) Joint analysis of tightly linked SNPs in screening step of genome-wide association studies leads to increased power. *European Journal of Human Genetics,* 1-7

Becker, M. A. and Jolly, M. (2006) Hyperuricemia and associated diseases. *Rheumatoid Disease Clinics of North America* **32**:275-293

Bell, J., and Haldane, J. B. S. (1937) The linkage between the genes for colour-blindness and haemophilia in man. *Proceedings of the Royal Society B* **123**:119-150

Blangero, J. (2004) Localization and identification of human quantitative trait loci: King Harvest has surely come. *Current Opinion in Genetics and Development,* **14**:233-240

Blangero, J., Williams, J.T. and Almasy, L. (2001) Variance Component Methods for Detecting Complex Trait Loci. *Advances in Genetics* **42**:151-181

Bobadilla, J. L., Macek, M. Jr., Fine, J. P. and Farrell, P. M. (2002) Cystic fibrosis: a worldwide analysis of CFTR mutations-correlation with incidence data and application to screening. *Human Mutation,* **19**:575-606

Bodmer, W. and Bonilla, C. (2008) Common and rare variants in multifactorial susceptibility to common disease. *Nature Genetics,* **40(6)**:695-701

Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics Supplement,* **33**:228-237

Botstein, D., White, R. L., Skolnick, M. and Davis, R. W. (1980) Construction of a genetic map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**:314-331

Bourgain, C. and Genin, E. (2005) Complex trait mapping in isolated populations: Are specific statistical methods required? *European Journal of Human Genetics,* **13**:698-706

Browning, S. R. (2006) Multilocus Association Mapping Using Variable-Length Markov Chains. *American Journal of Human Genetics,* **78**:903-913

Cappuccio, F. P., Strazzullo, P., Farinaro, E. and Trevisan, M. (1993) Uric acid metabolism and tubular sodium handling. Results from a population-based study. *Journal of the American Medical Association,* **270(3)**:354-359

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q. and Lander, E. S. (1999) Characterisation of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics,* **22**:231-238

Carlson, C. S., Eberle, M. A., Kruglyak, L. and Nickerson, D. A. (2004) Mapping complex disease loci in whole-genome association studies. *Nature,* **429**:446-452

Chapman, J. M., Cooper, J. D., Todd, J. A. and Clayton, D. G. (2003) Detecting Disease Associations due to Linkage Disequilibrium Using Haplotype Tags: A Class of Tests and the Determinants of Statistical Power. *Human Heredity,* **56**:18-31

Clark, A. G. (2004) The role of haplotypes in candidate gene studies. *Genetic Epidemiology,* **27**:321-333

Clayton, D., Chapman, J. and Cooper J. (2004) Use of Unphased Multilocus Genotype Data in Indirect Association Studies. *Genetic Epidemiology,* **27**:415-428

Clayton, D. and McKeigue, P. M. (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*, **358**, 1356-1360

Clerget-Darpoux, F., Bonaiti-Pellie, C. and Hochez, J. (1986) Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* **42**:393-399

Coleman, H. R., Chan, C. C., Ferris, F. L. 3rd and Chew, E. Y. (2008) Age-related macular degeneration. *Lancet*, **372**:1835-1845

Collins, F. S., Guyer, M. S. and Chakravarti, A. (1997) Variations on a theme: Cataloguing Human DNA Sequence Variation. *Science,* **278**:1580-1581

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., MacArthur, D. G., MacDonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups. K., Valsesia, A., Walter, K., Wei, J., Wellcome Trust Case Control Consortium, Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W. and

Hurles, M. E. (2009) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**:704-712

Cordell, H. J. (2006) Estimation and Testing of Genotype and Haplotype Effects in Case-Control Studies: Comparison of Weighted Regression and Multiple Imputation Procedures. *Genetic Epidemiology*, **30**:259-275

Cordell, H. J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, **10**:392-404

Crawford, D. C., Carlson, C. S., Rieder, M. J., Carrington, D. P., Yi, Q., Smith, J. D., Eberle, M. A., Kruglyak, L. and Nickerson, D. A. (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *American Journal of Human Genetics,* **74**:610-622

Cutler, G., Marshall, L. A., Chin, N., Baribault, H. and Kassner, P. D. (2007) Significant gene content variation characterises the genomes inbred mouse strains. *Genome Research,* **17**:1743-1754

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B),* **39**:1-38

Dehghan, A., Kottgen, A., Yang, Q., Hwang, S. J., Kao, W. L., Rivadeneira, F., Boerwinkle, E., Levy, D., Hofman, A., Astor, B. C., Benjamin, E. J., van Duijn, C. M., Witteman, J. C., Coresh, J. and Fox, C. S. (2008) Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *Lancet*, **372**:1953-1961

Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics,* **55**:997-1004

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R. and Cookson, W. O. C. (2007) A genome-wide association study of global gene expression. *Nature Genetics*, **39**:1202-1207

Donnelly, P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, **456**:728-731

Edwards, A. W. F. (2005) Linkage methods in human genetics before the computer. *Human Genetics,* **118**:515-530

Eichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., Carter, N. P., Church, D. M., Felsenfeld, A., Guyer, M., Lee, C., Lupski, J. R., Mullikin, , J. C., Pritchard, J. K., Sebat, J., Sherry, S. T., Smith, D., Valle, D. and Waterston, R. H. (2007) Completing the map of human genetic variation. *Nature,* **447**:161-165

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**:636-640

Enomoto, A., Kimura, H., Chairoungdua, A., Shigeta, Y., Jutabha, P., Cha, S. H., Hosoyamada, M., Takeda, M., Sekine, T., Igarashi, T., Matsuo, H., Kikuchi, Y., Oda, T., Ichida, K., Hosoya, T., Shimokata, K., Niwa, T., Kanai, Y. and Endou, H. (2002) Molecular identification of a renal urate-anion exchanger that regulates blood urate levels. *Nature,* **417**:447-452

Excoffier, L., and Slatkin, M. (1995) Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution,* **12**:921-927

Fahlen, M. T. and Agraharkar, M. (2009) Uric Acid Nephropathy. Accessed from emedicine.medscape.com

Fallin, D. and Schork, N. (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the Expectation-Maximisation algorithm for unphased diploid genotype data. *American Journal of Human Genetics,* **67**:947-959

Feingold, E. (2001) Methods for Linkage Analysis of Quantitative Trait Loci in Humans. *Theoretical Population Biology* **60**:167-180

Feuk, L., Carson, A. R. and Scherer, S. W. (2006) Structural variation in the human genome. *Nature Reviews Genetics,* **7**:85-97

Fijneman, R. J., de Vries, S. S., Jansen, R. C. and Demant, P. (1996) Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse. *Nature Genetics,* **14**:465-467

Fisher, R. A. (1918) The correlations between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, **52**:399-433

Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A* **222**:309-368

Fisher, R. A. (1935) Eugenics, academic and practical. *Eugenics Review* **27**:95-100

Forabosco, P., Falchi, M. and Devoto, M. (2005) Statistical tools for linkage analysis and genetic association studies. *Expert Review of Molecular Diagnostics.* **5(5)**:781-795

Frazer, K. A., Murray, S. S., Schork, N. J. and Topol, E. J. (2009) Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**:241-251

Gerrits, A., Li, Y., Tesson, B. M., Bystrykh, L. V., Weersing, E., Ausema, A., Dontje, B., Wang, X., Greitling, R., Jansen, R. and de Haan, G. (2009) Expression

Quantitative Trait Loci Are Highly Sensitive to Cellular Differentiation State. *PLoS Genetics*, **5(10)**:e1000692

Gianfrancesco, F., Esposito, T., Ombra, M. N., Forabosco, P., Maninchedda, G., Fattorini, M., Casula, S., Vaccargiu, S., Casu, G., Cardia, F., Deiana, I., Melis, P., Falchi, M. and Pirastu, M. (2003) Identification of a Novel Gene and a Common Variant Associated with Uric Acid Nephrolithiasis in a Sardinian Genetic Isolate. *American Journal of Human Genetics,* **72**:1479-1491

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., Murthy, K. K., Rovin, B. H., Bradley, W., Clark, R. A., Anderson, S. A., O'Connell, R. J., Agan, B. K., Ahuja, S. S., Bologna, R., Sen, L., Dolan, M. J. and Ahuja, S. K. (2005) The influence of the CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science,* **307**:1434-1440

Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. and Amos, C. I. (2008) Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. *American Journal of Human Genetics,* **82**:100-112

Graessler, J., Graessler, A., Unger, S., Kopprasch, S., Tausche, A. K., Kuhlisch, E. and Schroeder, H. E. (2006) Association of the human urate transporter 1 with reduced renal uric acid excretion and hyperuricemia in a German Caucasian population. *Arthritis and Rheumatism*, **54(1)**:292-300

Grapes, L., Dekkers, J. C. M., Rothschild, M. F. and Fernando, R. L. (2004) Comparing Linkage Disequilibrium-Based Methods for Fine Mapping Quantitative Trait Loci. *Genetics,* **166**:1561-1570

Gregersen, J. W., Kranc, K. R., Ke, X., Svendsen, P., Madsen, L. S., Thomsen, A. R., Cardon, L. R., Bell, J. I. and Fugger, L. (2006) Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature*, **443**:574-577

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., Helgadottir, A., Ingason, A., Steinthorsdottir, V., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Pedersen, O., Aben, K. K., Witjes, J. A., Swinkels, D. W., den Heijer, M., Franke, B., Verbeek, A. L., Becker, D. M., Yanek, L. R., Becker, L. C., Tryggvadottir, L., Rafnar, T., Gulcher, J., Kiemeney, L. A., Kong, A., Thorsteinsdottir, U. and Stefansson, K. (2008) Many sequence variants affecting diversity of adult human height. *Nature Genetics,* **40(5)**:609-615

Haseman, J. K. and Elston, R. C. (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behavioural Genetics,* **2**:3-19

Heinig, M. and Johnson, R. J. (2006) Role of uric acid in hypertension, renal disease, and metabolic syndrome. *Cleveland Clinic Journal of Medicine,* **73(12)**:1059-1064

Henrichsen, C. N., Chaignat, E. and Reymond, A. (2009) Copy number variants, diseases and gene expression. *Human Molecular Genetics,* **18**:R1-R8

Heutink, P. and Oostra, B. (2002) Gene finding in genetically isolated populations. *Human Molecular Genetics,* **11(20)**:2507-2515

Hindorff, L. A., Junkins, H. A., Hall, P. N., Mehta, J. P. and Manolio, T. A. A Catalogue of Published Genome-Wide Association Studies. Available at www.genome.gov/gwastudies

Hirschhorn, J. N., Lohmueller, K., Byrne, E. and Hirschhorn, K. (2002) A comprehensive review of genetic association studies. *Genetics in Medicine,* **4(2)**:45-61

Hoggart, C. J., Chadeau-Hyam, M., Clark, T. G., Lampariello, R., Whittaker, J. C., (2007) Sequence-level population simulations over large genomic regions. *Genetics,* **177**:1725-1731

Hollox, E. J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A. I. and Swallow, D. M. (2001) Lactose haplotype diversity in the old world. *American Journal of Human Genetics,* **68**:160-172

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, C., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A. and Birney, E. (2007) Ensembl 2007. *Nucleic Acids Research*, **35(database issue)**:D610-D617

Hunter, D. J. (2005) Gene-environment interactions in human diseases. *Nature Reviews Genetics*, **6**:287-298

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nature Genetics,* **36(9)**:949-951

Illumina website; www.illumina.com

Ivkovic,V., Vitart,V., Rudan, I., Janicijevic, B., Smolej-Narancic, N., Skaric-Juric, T., Barbalic, M, Polasek, O, Kolcic, I., Biloglav, Z., Visscher, P. M., Hayward, C., Hastie, N. D., Anderson, N., Campbell, H., Wright, A. F., Rudan, P. and Deary, I. J. (2007) The Eysenck personality factors: psychometric structure, reliability, heritability and phenotypic and genetic correlations with psychological distress in an

isolated Croatian population. *Personality and Individual Differences*, ***42***:123-133.

Jannot, A.-S., Essioux, L., Reese, M. G. and Clerget-Darpoux, F. (2003) Improved Use of SNP Information to Detect the Role of Genes. *Genetic Epidemiology,* **25**:158-167

Jing, L. and Jiang, T. (2005) Haplotype-based linkage disequilibrium mapping via direct data mining. *Bioinformatics,* **21(24)**:4384-4393

John, S., Shephard, N., Liu, G., Zeggini, E., Cao, M., Chen, W., Vasavda, N., Mills, T., Barton, A., Hinks, A., Eyre, S., Jones, K. W., Ollier, W., Silman, A., Gibson, N., Worthington, J. and Kennedy, G. C. (2004) Whole-Genome Scan, in a Complex Disease, Using 11,245 Single-Nucleotide Polymorphisms: Comparison with Microsatellites. *American Journal of Human Genetics* **75**:54-64

Johnson, A. D., Handsaker, R. E., Pulit, S., Nizzari, M. M., O'Donnell, C. J. and de Bakker, P. I. W. (2008) SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, **24(24)**:2938-2939

Keembiyehetty, C., Augustin, R., Carayannopoulos, M. O., Steer, S., Manolescu, A., Cheeseman, C. I. and Moley K. H. (2006) Mouse glucose transporter 9 splice variants are expressed in adult liver and kidney and are up-regulated in diabetes. *Molecular Endocrinology,* **20(3)**:686-697

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tuzun, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoluos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R. and Eichler, E. E. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature,* **453**:56-64

Koenen, K. C., Amstadter, A. B. and Nugent, N. R. (2009) Gene-environment interaction in posttraumatic stress disorder: An update. *Journal of Traumatic Stress*, **22(5)**:416-426

Kolz, M., Johnson, T., Sanna, S., Teumer, A., Vitart, V., Perola, M., Mangino, M., Albrecht, E., Wallace, C., Farrall, M., Johansson, A., Nyholt, D. R., Aulchenko, Y., Beckmann, J. S., Bergmann, S., Bochud, M., Brown, B., Campbell, H.; EUROSPAN Consortium, Connell, J., Dominiczak, A., Homuth, G., Lamina, C., McCarthy, M. I.; ENGAGE Consortium, Meitinger, T., Mooser, V., Munroe, P., Nauck, M., Peden, J., Prokisch, H., Salo, P., Salomaa, V., Samani, N. J., Schlessinger, D., Uda, M., Volker, U., Waeber, G., Waterworth, D., Wang-Sattler, R., Wright, A. F., Adamski, J., Whitfield, J. B., Gyllensten, U., Wilson, J. F., Rudan, I., Pramstaller, P., Watkins, H.; PROCARDIS Consortium, Doering, A., Wichmann, H. E.; KORA Study, Spector, T.

D., Peltonen, L., Volzke, H., Nagaraja, R., Vollenweider, P., Caulfield, M.; WTCCC, Illig, T. and Gieger, C. (2009) Meta-analysis of 28,141 Individuals Identified Common Variants within Five New Loci That Influence Uric Acid Concentrations. *PLoS Genetics*, **5(6)**:e1000504

Kruglyak, L. (1997) The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics* **17**:21-24

Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics,* **22**:139-144

Lander, E. S. and Botstein, D. (1988) Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics* **121**:185-199

Lander, E. and Kruglyak, L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**:241-247

Lawler, S. D., and Renwick, J. H. (1959) Blood groups and genetic linkage. *British Medical Bulletin* **15**:145-149

Lee, A. S., Gutierrez-Arcelus, M., Perry, G. H., Vallender, E. J., Johnson, W. E., Miller, G. M., Korbel, J. O. and Lee, C. (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Human Molecular Genetics,* **17**:1127-1136

Lee, S. H. and Van Der Werf, J. H. J. (2005) The role of pedigree information in combined linkage disequilibrium and linkage mapping of quantitative trait loci in a general complex pedigree. *Genetics,* **169**:455-466

Le Marchand, L., Wilkens, L. R., Kolonel, L. N., and Henderson, B. E. (2005) The *MTHFR* C667T polymorphism and colorectal cancer: the Multiethnic Cohort Study. *Cancer Epidemiology Biomarkers, and Prevention,* **14**:1198-1203

Le Marchand, L. and Wilkens, L. R. (2008) Design Considerations for Genomic Association Studies: Importance of Gene-Environment Interactions. *Cancer Epidemiology, Biomarkers and Prevention*, **17(2)**:263-267

Levy, S. Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., Busam, D. A., Beeson, K. Y., McIntosh, B. C., Remington, K. A., Abril, J. F., Gill, J., Borman, J., Rogers, Y. H., Frazier, M. E., Scherer, S. W., Strausberg, R. L. and Venter, J. C. (2007) The diploid genome sequence of an individual human. *PLoS Biology,* **5**:e254

Li, J., Zhou, Y. and Elston, R. C. (2006) Haplotype-based quantitative trait mapping using a clustering algorithm. *BMC Bioinformatics,* **7**:258-268

Li, Y., Sung, W-K. and Liu, J. J. (2007) Association Mapping via Regularized Regression Analysis of Single-Nucleotide-Polymorphism Haplotypes in Variable-Sized Sliding Windows. *American Journal of Human Genetics,* **80**:705-715

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. and Hirschhorn, J. N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics,* **33**:177-182

Longmate, J. A. (2001) Complexity and power in case-control association studies. *American Journal of Human Genetics,* **68**:1229-1237

Luo, Y. C., Do, J. S. and Liu, C. C. (2006) An amperometric uric acid biosensor based on modified Ir-C electrode. *Biosensors and Bioelectronics,* **22(4)**:482-488

Maher, B. (2008) The case of the missing heritability. *Nature,* **456**:18-21

Meuwissen, T. H .E., and Goddard, M. (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics,* **155**:421-430

Michalatos-Beloin, S., Tishkoff, S. A., Bentley, K. L., Kidd, K. K. and Ruano, G. (1996) Molecular haplotyping of genetic markers 10kb apart by allele-specific long-range PCR. *Nucleic Acids Research,* **24**:4841-4843

Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E., Heinzmann, A., Simma, B., Frischer, T., Willis-Owen, S. A. G., Wong, K. C. C., Illig, T., Vogelberg, C., Weiland, S. K., von Mutius, E., Abecasis, G. R., Farrall, M., Gut, I. G., Lathrop, G. M. and Cookson, W. O. C. (2007) Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature*, **448**:470-473

Moore, J. H. and Williams, S. M (2009) Epistasis and Its Implications for Personal Genetics. *American Journal of Human Genetics*, **85**:309-320

Mora, S., Yanek, L. R., Moy, T. F., Fallin, M. D., Becker, L. C. and Becker, D. M. (2005) Interaction of body mass index and framingham risk score in predicting incident coronary disease in families. *Circulation,* **111(15)**:1871-1876

Morgan, T. H. (1911) Random segregation versus coupling in Mendelian inheritance. *Science* ns 34:384

Morris, A. P. (2005) Direct Analysis of Unphased SNP Genotype Data in Population-Based Association Studies Via Bayesian Partition Modelling of Haplotypes. *Genetic Epidemiology,* **29**:91-107

Morris, A. P., Whittaker, J. C. and Balding, D. J. (2004) Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *American Journal of Human Genetics,* **74**:945-953

Morris, J. A., Randall, J. C., Maller, J. B. and Barrett, J. C. (2010) Evoker: a visualisation tool for genotype intensity data. *Bioinformatics*, **26(14)**:1786-1787

Morris, R. W. and Kaplan, N. L. (2002) On the Advantage of Haplotype Analysis in the Presence of Multiple Disease Susceptibility Alleles. *Genetic Epidemiology,* **23**:221-233

Muraoka, S. and Miura, T. (2004) Inhibition of xanthine oxidase by phytic acid and its antioxidant action. *Life Sciences,* **74**:1691-1700

MyPhunong, T. L., Shafiu, M., Mu, W. and Johnson, R. J. (2008) SLC2A9 – a fructose transporter identified as a novel uric acid transporter. *Nephrology Dialysis Transplantation,* **23(9)**:2746-2749

Ng, P. C., Levy, S., Huang, J., Stockwell, T. B., Walenz, B. P., Li, K., Axelrod, N., Busam, D. A., Strausberg, R. L. and Venter, J. C. (2008) Genetic variation in an individual human exome. *PLoS Genetics,* **4**:e1000160

Nickerson, D. A., Taylor, S. L., Weiss, K. M., Clark, A. G., Hutchinson, R. G., Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E. and Sing, C. F. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genetics,* **19**:233-240

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010) Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genetics*, **6(4)**:e1000888

Nielsen, D. M., Ehm, M. G., Zaykin, D. V. and Weir, B. S. (2004) Effect of Two- and Three-Locus Linkage Disequilibrium on the Power to Detect Marker/Phenotype Associations. *Genetics,* **168**:1029-1040

Niu, T. (2004) Algorithms for Inferring Haplotypes. *Genetic Epidemiology,* **27**:334-347

Niu, T., Qin, Z. S., Xu, X. and Liu, J. S. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics,* **70**:157-169

Nyhan, W. L. (2005) Lesch-Nyhan disease. *Journal of the History of the Neurosciences,* **14(1)**:1-10

Olson, J. M. and Wijsman, E. M. (1993) Linkage between quantitative trait and marker loci: Methods using all relative pairs. *Genetic Epidemiology* **10**:87-102

Pe'er, I., de Bakker, P. I. W., Maller, J., Yelensky, R., Altshuler, D. and Daly, M. J. (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics,* **38**:663-667

Peng, B. and Kimmel, M. (2007) Simulations provide support for the common disease-common variant hypothesis. *Genetics,* **175**:763-776

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villenea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee. C. and Stone, A. C. (2007) Diet and the evolution of human amylase gene copy number variation. *Nature Genetics,* **39**:1256-1260

Phay, J. E., Hussian, H. B. and Moley, J. F. (2000) Cloning and Expression Analysis of a Novel Member of the Facilitative Glucose Transporter Family, SLC2A9 (GLUT9). *Genomics,* **66(2)**:217-220

Phillips, P. C. (1998) The language of gene interaction. *Genetics*, **149**:1167-1171

Phillips, P. C. (2008) Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9(11)**:855-867

Pritchard, J. K. (2001) Are Rare Variants Responsible for Susceptibility to Complex Diseases? *American Journal of Human Genetics,* **69**:124-137

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics,* **155**:945-959

Qin, Z. S., Niu, T. and Liu, J. S. (2002) Partition-ligation-expectation-maximisation algorithm for haplotype inference with single-nucleotide polymorphisms. *American Journal of Human Genetics,* **71**:1242-1247

R Development Core Team. (2006) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W. and Hurles, M. E. (2006) Global variation in copy number in the human genome. *Nature,* **444**:444-454

Reich, D. E. and Lander, E. S. (2001) On the allelic spectrum of human disease. *Trends in Genetics,* **17(9)**:502-510

Risch, N. and Merikangas, K. (1996) The Future of Genetic Studies of Complex Human Diseases. *Science,* **273**:1516-1517

Schaid, D. J. (2004) Evaluating Associations of Haplotypes With Traits. *Genetic Epidemiology,* **27**:348-364

Schaid, D. J. (2005) Power and Sample Size for Testing Associations of Haplotypes with Complex Traits. *Annals of Human Genetics,* **70**:116-130

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. and Poland, G. A. (2002) Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous. *American Journal of Human Genetics,* **70**:425-434

Schlotterer, C. (2004) The evolution of molecular markers – just a matter of fashion? *Nature Reviews Genetics.* **5**:63-69

Schork, N. J., Murray, S. S., Frazer, K. A. and Topol, E. J. (2009) Common vs. rare allele hypothesis for complex diseases. *Current Opinion in Genetics and Development,* **19**:212-219

Seegmiller, J. E., Laster, L. and Howell, R. R. (1963) Biochemistry of uric acid and its relation to gout. *New England Journal of Medicine,* **4**:764-773

Shifman, S. and Darvasi, A. (2001) The value of isolated populations. *Nature Genetics,* **28**:309-310

Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., De Lange, M., Harris, J. R., Hjelmborg, J. V., Luciano, M., Martin, N. G., Mortensen, J., Nistico, L., Pedersen, N. L., Skytthe, A., Spector, T. D., Stazi, M. A., Willemsen, G. and Kaprio, J. (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Research,* **6(5)**:399-408

Slager, S. L., Huang, J. and Vieland, V. J. (2000) Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genetic Epidemiology,* **18**:143-156

Slatkin, M. (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**:477-485

Smith, D. J. and Lusis, A. J. (2002) The allelic structure of common disease. *Human Molecular Genetics,* **11(20)**:2455-2461

Smith, P. G. and Day, N. E. (1984) The design of case-control studies: the influence of confounding and interaction effects. *International Journal of Epidemiology*, **13**:356-365

Sobel, E. and Lange, K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *American Journal of Human Genetics*, 58:1323-1337

Souto, J. C., Almasy, L., Muniz-Diaz, E., Soria, J. M., Borrell, M., Bayen, L., Mateo, J., Madoz, P., Stone, W., Blangero, J. and Fontcuberta, J. (2000) Polymorphism on

Plasma Levels of von Willebrand Factor, Factor VIII, and Activated Partial Thromboplastin Time. *Arteriosclerosis, Thrombosis, and Vascular Biology,* **20**:2024-2025

Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., Gunnarsdottir, S., Ivarsson, O., Chou, T. T., Hjaltason, O., Birgisdottir, B., Jonsson, H., Gudnadottir, V. G., Gudmundsdottir, E., Bjornsson, A., Ingvarsson, B., Ingason, A., Sigfusson, S., Hardardottir, H., Harvey, R. P., Lai, D., Zhou, M., Brunner, D., Mutel, V., Gonzalo, A., Lemke, G., Sainz, J., Johannesson, G., Andresson, T., Gudbjartsson, D., Manolescu, A., Frigge, M. L., Gurney, M. E., Kong, A., Gulcher, J, R., Petursson, H. and Stefansson, K. (2002) *Neuregulin 1* and Susceptibility to Schizophrenia. *American Journal of Human Genetics,* **71**:877-892

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., Olafsdottir, A., Cazier, J-B., Kristjansson, K., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., Kong, A. and Stefansson, K. (2005) A common inversion under selection in Europeans. *Nature Genetics,* **37**:129-137

Stephens, M. and Donnelly, P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics,* **73**:1162-1169

Stephens, M. and Scheet, P. (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation. *American Journal of Human Genetics,* **76**:449-462

Stephens, M., Smith, N. J. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics,* **68**:978-989

Sunnucks, P. (2000) Efficient genetic markers for population biology. *TREE* **15**:199-203

Tanck, M., Klerkx, A., Jukema, J., DeKnijff, P., Kastelein, J. and Zwinderman, A. (2003) Estimation of multilocus haplotype effects using weighted penalized log-likelihood: analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. *Annals of Human Genetics,* **67**:175-184

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437**:1299-1320

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature,* **449**:851-861

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* **447**:661-678

The Wellcome Trust Case Control Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature,* **464**:713-720

Thomas, D. C. (1988) Exposure-time-response relationships with applications to cancer epidemiology. *Annual Review of Public Health*, **9**:451-482

Thomas, D. (2010) Gene-environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, **11**:259-272

Thorisson, G. A., Smith, A. V., Krishnan, L. and Stein, L. D. (2005) The International HapMap Project Web site. *Genome Res.*, **15**:1592-1593

Toncev, G., Milicic, B., Toncev, S. and Samardzic, G. (2002) Serum uric acid levels in multiple sclerosis patients correlate with activity of disease and blood-brain barrier dysfunction. *European Journal of Neurology,* **9(3)**:221-226

Tsai, C. T., Hwang, J. J., Ritchie, M. D., Moore, J. H., Chiang, F. T., Lai, L. P., Hsu, K. L., Tseng, C. D., Lin, J. L. and Tseng, Y. Z. (2007) Renin-angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene-gene interaction. *Atherosclerosis*, **195**:172-180

Vignal, A., Milan, D., SanCristobal, M. and Eggen, A. (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution* **34**:275-305

Visscher, P. M., Andrew, T. and Nyholt, D. R. (2008) Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. *European Journal of Human Genetics,* **16**:387-390

Vitart, V., Rudan, I., Hayward, C., Gray, N. K., Floyd, J., Palmer, C. N., Knott, S. A., Kolcic, I., Polasek, O., Graessler, J., Wilson, J. F., Marinaki, A., Riches, P. L., Shu, X., Janicijevic, B., Smolej-Narancic, N., Gorgoni, B., Morgan, J., Campbell, S., Biloglav, Z., Barac-Lauc, L., Pericic, M., Klaric, I. M., Zgaga, L., Skaric-Juric, T., Wild, S. H., Richardson, W. A., Hohenstein, P., Kimber, C. H., Tenesa, A., Donnelly, L. A., Fairbanks, L. D., Aringer, M., McKeigue, P. M., Ralston,, S. H., Morris, A. D., Rudan, P., Hastie, N. D, Campbell, H. and Wright, A. F. (2008) SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nature Genetics,* **40(4)**:437-442

Wang, W. Y. S. and Pike, N. (2004) The allelic spectra of common diseases may resemble the allelic spectrum of the full genome. *Medical Hypotheses,* **63**:748-751

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., Gomes. X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A. and Rothberg, J. M. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature,* **452**:53-59

Wiltshire, S. Bell, J. T., Groves, C. J., Dina, C., Hattersley, A. T., Frayling, T. M., Walker, M., Hitman, G. A., Vaxillaire, M., Farrall, M., Froguel, P. and McCarthy, M. I. (2006) Epistasis between type 2 diabetes susceptibility Loci on chromosomes 1q21-25 and 10q23-26 in northern Europeans. *Annals of Human Genetics*, **70**:726-737

Wong, G. K., Yang, Z., Passey, D. A., Kibukawa, M., Paddock, M., Liu, C. R., Bolund, L. and Yu, J. (2003) A population threshold for functional polymorphisms. *Genome Research,* **13**:1873-1879

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K. V., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V. K., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculesco, V. E. and Vogelstein, B. (2007) The genomic landscapes of human breast and colorectal cancers. *Science,* **318**:1108-1113

Woodhead, M., Russell, J., Squirrell, J., Hollingsworth, P. M., MacKenzie, K., Gibby, M. and Powell, W. (2005) Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. *Molecular Ecology* **14**:1681–1695

Xu, L. L., Wain, J. C., Miller, D. P., Thurston, S. W., Su, L., Lynch, T. J. and Christiani, D. C. (2001) The NAD(P)H:quinone Oxidoreductase 1 Gene Polymorphism and Lung Cancer: Differential Susceptibility Based on Smoking Behaviour. *Cancer Epidemiology, Biomarkers and Prevention*, **10**:303-309

Zaykin, D V., Westfall, P. H., Young, S. S., Karnoub, M. A., Wagner, M. J. and Ehm, M. G. (2002) Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals. *Human Heredity,* **53**:79-91

Zhao, H. H., Fernando, R. L. and Dekkers, J. C. M. (2007) Power and Precision of Alternate Methods for Linkage Disequilibrium Mapping of Quantitative Trait Loci. *Genetics,* **175**:1975-1986

Zhao, L., Li, S. and Khalid, N. (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental

variables in case-control studies. *American Journal of Human Genetics,* **72**:1231-1250

Zondervan, K. T., and Cardon, L. R. (2004) The complex interplay among factors that influence allelic association. *Nature Reviews, Genetics,* **5:**89-101