MODEL SELECTION STRATEGIES IN GENOME-WIDE ASSOCIATION STUDIES

DPHIL THESIS AND ABSTRACT

Sarah Keildson

Keble College, University of Oxford

21 June 2011

Abstract

MODEL SELECTION STRATEGIES IN GENOME-WIDE ASSOCIATION STUDIES

Sarah Keildson, Keble College

DPhil Thesis, Trinity Term, 2011

Unravelling the genetic architecture of common diseases is a continuing challenge in human genetics. While genome-wide association studies (GWAS) have proven to be successful in identifying many new disease susceptibility loci, the extension of these studies beyond single-SNP methods of analysis has been limited. The incorporation of multi-locus methods of analysis may, however, increase the power of GWAS to detect genes of smaller effect size, as well as genes that interact with each other and the environment. This investigation carried out large-scale simulations of four multi-locus model selection techniques; namely forward and backward selection, Bayesian model averaging (BMA) and least angle regression with a lasso modification (lasso), in order to compare the type I error rates and power of each method. At a type I error rate of ~5%, lasso showed the highest power across varied effect sizes, disease frequencies and genetic models. Lasso penalized regression was then used to perform three different types of analysis on GWAS data. Firstly, lasso was applied to the Wellcome Trust Case Control Consortium (WTCCC) data and identified many of the WTCCC SNPs that had a moderate-strong association ($p<10^{-5}$) type 2 diabetes (T2D), as well as some of the moderate WTCCC associations ($p<10^{-4}$) that have since been replicated in a large-scale meta-analysis. Secondly, lasso was used to fine-map the 17q21 childhood asthma risk locus and identified putative secondary signals in the 17q21 region, that may further contribute to childhood asthma risk. Finally, lasso identified three potential interaction effects potentially contributing towards coronary artery disease (CAD) risk. While the validity of these findings hinges on their replication in follow-up studies, the results suggest that lasso may provide scientists with exciting new methods of dissecting, and ultimately understanding, the complex genetic framework underlying common human diseases.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my supervisors, Prof. Martin Farrall and Dr. Andrew Morris, for their advice and guidance throughout my DPhil. They have been both patient and supportive and have provided me with helpful advice and excellent insight into the field of statistical genetics. I would also like to thank the GABRIEL asthma consortium, a multidisciplinary study funded through an EC Framework 6 grant, for funding my DPhil and providing me with the opportunity to be a part of their collaboration.

I owe a great deal of thanks to the Department of Cardiovascular Medicine for their financial support, and in particular, a very special thank you to Lynn Clee, for all her help over the years. Being a part of the GGEU group at the Wellcome Trust Centre for Human Genetics has also been a fantastic learning experience and has helped to make my DPhil a much more enjoyable experience.

Another big thanks to John “skelm” Perry for everything he has done for me this year. You have, and always will have, 10% shares in my DPhil. Je’ne, your friendship was a lifesaver on more than one occasion and one of the best things to come out of my DPhil was meeting you!

Most importantly, I’d like to thank my family – Neil and Christina for all the laughs and encouragement, Jake and Carm for being the world’s greatest siblings and the precious Justin, who will hopefully acknowledge me one day when he takes over the world. Without their years of support and unwavering faith in me, I would not be where I am today. And finally, Mom and Dad, I will never be able to fully

express my gratitude to you both for everything that you have done for me. I can only hope that I will be able to inspire people the way you have inspired me – to become *'n doktor of 'n ding*.

Chapter 1: Introduction

One of the major goals in human genetics is the identification of genetic variants that predispose individuals to common, complex diseases such as asthma, coronary artery disease (CAD) and type 2 diabetes (T2D). It is hoped that understanding the genetic basis of common disease susceptibility could lead to the development of potential pharmaceutical targets and public health would benefit from improved diagnosis and better disease prevention (WTCCC, 2007). While genetic linkage studies provide researchers with the tools to discover the rare gene variants that cause monogenic diseases like cystic fibrosis (HARDY and SINGLETON 2009), until recently, the detection of genetic variants that underlie complex diseases has been slow and difficult. Since complex diseases reflect the simultaneous action of multiple genes, together with environmental effects, they do not display the Mendelian patterns of inheritance that are readily identified by linkage studies (MOTULSKY 2006). Human geneticists have therefore turned their attention to new, more powerful statistical approaches that scan representative samples of whole populations for genetic variation that is observed more often in individuals who have a particular disease than in those that do not.

Identifying and understanding the causes of the correlations that exist between molecular and phenotypic variation in humans, could provide a necessary platform for unravelling and understanding this complex disease architecture (CHAKRAVARTI 1999). These correlations are observed in populations because genetic variants either: (a) have a causal role in the disease; (b) co-segregate

with a nearby causal variant in the population; or (c) become associated with the disease because of the underlying genetic structure of the population (CORDELL and CLAYTON 2005). Through international efforts like the Human Genome and HapMap Projects, a wealth of information about human genetic variation has been generated and numerous complex disease – gene associations have already been reported (ALTSHULER *et al.* 2010). While studies that could directly identify causal variants in the genome would be powerful and easy to analyse, teasing out these direct associations from such an enormous amount of human variation is extremely difficult. Statistical genetics has therefore focussed on detecting significant associations between easily identifiable regions of the genome (genetic markers) and the genes responsible for disease phenotypes.

This investigation aims to analyse and compare different statistical methods that are powerful enough to detect such associations and can accurately pinpoint genetic variants that underlie common, complex diseases. Successful methods of analysis, however, begin with a thorough understanding of the complexities of the human genome; more specifically, they begin with the understanding of the genetic variation that underlies common disease phenotypes and how this variation can be pinpointed and understood.

1.1 Linkage

The traditional approach to mapping genetic variants that predispose individuals to disease, is to search for co-segregation of phenotypes and genetic markers in families, by means of linkage analysis. For two loci, A and B, each with two alleles, A_1, A_2 and B_1, B_2, there are four possible gametes, namely A_1B_1, A_1B_2,

A_2B_1, A_2B_2 (HENDRICK 2000a). When the loci are independent of one another, the frequencies of these *haplotypes* are expected to equal the product of their constituent allele frequencies. Alleles, however, do not always assort independently of one another, but may co-vary, such that a haplotype containing the A_1 is more likely to contain B_1 than a randomly chosen allele.

The extent to which alleles at different loci associate in a non-random manner is dependent on the distance, and the consequent degree of recombination (r), that occurs between them. The possible effect of recombination on the transmission of alleles is illustrated in Figure 1.1

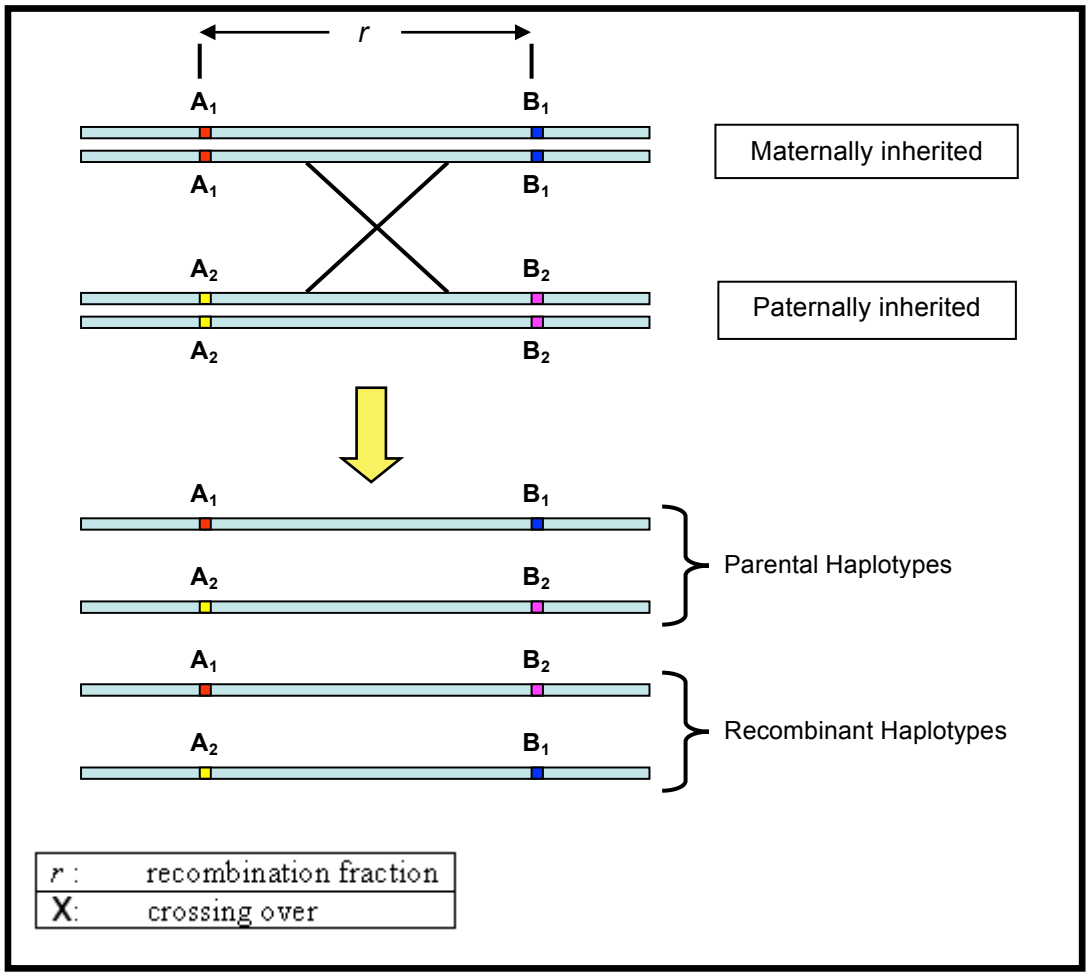


Figure 1.1 The four possible haplotypes produced by a double heterozygous individual.

As is evident in Figure 1.1, the parental configuration has A_1 on the same

chromosome as B_1 and A_2 coupled with B_2. If there is no crossing over, and therefore no recombination, between these two loci, only the two parental haplotypes will result. However, if there is crossing over, then recombination occurs and the two recombinant haplotypes will also result. If two loci are unlinked, r=0.5, then the parental and recombinant haplotypes are equally likely. The closer together the A and B loci are positioned along the chromosome, the less chance there is of crossing over occurring between them and the majority of haplotypes will be of the parental rather than the recombinant type. In this case the A and B loci are said to be linked (HARTL and CLARK 1997a).

In a genome, if A is an observed marker locus and B is an unknown disease locus, it is clear that the marker and the disease locus will co-segregate if the loci are genetically linked. In this way, over many generations, known genetic markers are often co-inherited with disease genes within families and researchers are able to gain information about the genetics of disease by analysing the resulting transmission patterns (BALDING 2006). By pinpointing genetic markers that are found more often in affected individuals in a pedigree than would be expected by chance, it is possible to approximately locate the linked disease gene locus (COLLINS *et al.* 1998). This principle has formed the basis of many family-based linkage studies, which have proven to be successful in identifying highly penetrant single-genes.

The logarithm (to the base 10) of odds or LOD score, is the most common statistical test for linkage analysis in human populations. The score is a ratio of the likelihood that the loci under question are linked (r<0.5), to the likelihood that they are not (r=0.5). Estimates have to be made about the recombination

frequency within an established pedigree and then a LOD score is calculated for each estimate (STRACHAN and READ 2010). A LOD score that is greater than 3 is indicative of linkage, since the likelihood of observing the given pedigree if the two loci are not linked, is less than 1 in 1000.

In the case of complex diseases, however, linkage studies have had limited success. The small sample sizes available from pedigrees, together with the multiple, small-effect genes that play a role in common disease phenotypes, have decreased the power of these studies to detect linkage to any one of the disease genes (BALDING 2006). As a result, attention has shifted away from families and towards populations as a whole, where larger available sample sizes can increase the power to detect alleles with relatively smaller risks (FAN and KNAPP 2003). In population association studies, the transmission of phenotypes over generations cannot be traced, and therefore individuals are treated as 'unrelated' because their relationships are unknown and presumed to be distant (BALDING 2006). Consequently, association studies make use of currently observed correlations between genetic markers and phenotypes of interest.

1.2 Linkage Disequilibrium

While every human genome is made up of unique sequences of DNA, certain allelic combinations, or haplotypes, can also be shared between individuals. The extent of this sharing is known as linkage disequilibrium (LD) and it is this phenomenon that drives the power of association studies to detect important genetic correlations in natural populations.

In a population, when genetic markers and disease causing variants are positioned close enough together on a chromosome such that recombination between them is minimal, they will occur together at a higher frequency than would be expected by chance (FALCONER 1989). In this situation, the two loci are said to be in LD and the presence of the known marker locus can potentially be used to infer the presence of the unknown disease locus in that population (ZONDERVAN and CARDON 2004).

D measures the amount of linkage disequilibrium in a population and is known as the LD coefficient (HARTL and CLARK 1997b). It can be calculated by the formula $D = P_{11}P_{22} - P_{12}P_{21}$, which is the product of the haplotype frequencies in the coupling phase (A_1B_1 / A_2B_2) minus the product of the haplotype frequencies in the repulsion phase (A_2B_1 / A_1B_2) (LEWONTIN 1995). $D = 0$ represents a population in linkage equilibrium and *D* increases or decreases away from 0 as the association between two alleles in a population increases (STRACHAN and READ 2010). Thus, in the case of two loci, strong LD ($D=\pm0.25$) is reached when the allele at one locus is identical to the allele at the other, for all the individuals in the population. A major disadvantage associated with the use of *D*, however, is its dependence on allele frequency and low values of *D* can be obtained in the presence of low frequency variants, even in the presence of strong LD (GAUT and LONG 2003).

A more commonly used measure of LD is D', which normalises *D* by dividing it by its theoretical maximum for the observed allele frequencies in the population (LEWONTIN 1988). Irrespective of allele frequencies, if two loci are in *complete*

LD, the absolute value of D' will be equal to 1. Complete LD occurs when the frequency of at least one out of the four possible haplotypes (Figure 1.1) is equal to 0 in the population, suggesting that the observed pattern of variation is consistent with the occurrence of a mutational event followed by no subsequent recombination (GAUT and LONG 2003). Problems arise, however, with intermediate values of D', since these are difficult to interpret due to its poorly understood sampling properties. Alternatively, r^2, which is equal to D^2 divided by the product of the allele frequencies at the two loci, is most commonly used in human genetics to quantify the amount of LD present between loci in a population.

The values of r^2 range from 0-1, where a value of 1 indicates that the two loci under consideration are in *perfect* LD and a value of 0 represents no correlation between the loci. This value is important as it allows researchers to ascertain whether the information at one locus is a good indicator of the information carried at another. SNPs at two loci in perfect LD, are referred to as genetically identical (LAWRENCE *et al.* 2005). Values of r^2 can be small, even when D'=1, in cases where mutations occur on different branches of chromosome ancestry, and therefore the distinction between complete and perfect LD is important. A useful feature of r^2 is that it is directly related to the power to detect associations between markers and disease loci that are in LD. However, care needs to be taken when using these measures. For example, because r^2 is a function of the allele frequencies at the loci under consideration, to obtain accurate measures of LD, it is important that these two loci have similar frequencies (VANLIERE and ROSENBERG 2008).

Unique patterns of LD are often observed in different populations. This is because evolutionary forces such as mutation, admixture, genetic drift and a lack of recombination can alter allele frequencies over time, to the extent that allele combinations which are rare in some populations, can become common in others (OHTA 1982). Since LD-based association methods offer a promising approach for detecting genetic variations that are responsible for complex human diseases (LIU *et al.* 2008), the patterns of LD and the factors that influence it are of utmost importance.

Current patterns of LD in human populations are generally assumed to be largely due to mutation events that occurred in a shared, relatively recent common ancestor (BALDING 2006). Thus, only closely linked loci will exhibit high levels of correlation, since recombination would have acted to break down any distant associations. The alleles at closely linked loci will therefore be in LD, whereas those at distantly linked loci will tend to be independent (NORDBORG and TAVARE 2002). Mutation-driven linkage disequilibrium is further enhanced by admixture, genetic drift and migration.

Admixture, the mixing of two populations which are genetically different, generates distinct patterns of LD that are population specific (ARDLIE *et al.* 2002). The amount of LD generated by admixture depends strongly on the differences in allele frequencies between the initial populations and the recombination fraction between the two loci (GAUT and LONG 2003). Genetic drift can also rapidly increase the frequency of certain haplotypes in populations, which can then spread to other populations though the migration of its carriers (HARTL and CLARK 1997b). While all these forces act to generate LD that can be utilised by

geneticists in association studies, the correlation between loci in LD is not absolute, and can be broken down by recombination.

The correlation between linked loci is slowly eroded over time until linkage equilibrium is reached and the loci segregate independently of one another (HENDRICK 2000b). Disequilibrium declines by a factor, r, every generation, so that the breakdown of LD is measured by the equation $D_n = (1-r)^n D_0$, where D_n is the amount of disequilibrium in generation n, and D_0 is the amount of disequilibrium in the initial population (GAUT and LONG 2003). Thus, for r>0, as the number of generations (*n*) approaches infinity, the amount of LD approaches 0, and loci will reach a state of linkage equilibrium (GILLESPIE 2004). Recurrent mutations can also break down the amount of LD by generating new alleles and consequently altering the existing frequency distributions of linked loci (TERWILLIGER and WEISS 1998). Thus, some SNPs might have high mutation rates and therefore show little or no LD with nearby markers, even in the absence of historical recombination (ABECASIS *et al.* 2001).

Information about the intricate and population-specific patterns of LD throughout the genome, serves to power association studies that are based on these patterns. The International HapMap Project (www.hapmap.org) has successfully developed a haplotype map of the human genome, describing these common patterns of human variation. This catalogue of variation across different population samples has effectively provided geneticists with the tools and power to perform LD-based association studies, in the search for disease risk factors.

1.3 Genome-Wide Association Studies

Genome-wide association studies (GWAS) are a relatively new method for identifying susceptibility alleles with small relative risks that cannot currently be identified by linkage studies. At a population level, statistical analyses can no longer utilise family-based linkage, since the family-specific polymorphisms and patterns of inheritance are not uniform throughout whole populations. Population-wide studies therefore rely on the current correlations between genetic markers and disease phenotypes that stem from the LD that exists between the respective loci (BALDING 2006).

GWAS involve pinpointing SNPs in the genome that occur more frequently in people with a particular disease (cases) than in people without the disease (controls) (FAN and KNAPP 2003). To achieve this, both case and control individuals are genotyped across their entire genomes at markers on commercially available genotyping chips, and researchers are then able to look for statistical associations between the genotypes and disease status of individuals in the sample. This method signifies an important step up from candidate gene studies, where only a few variants are genotyped based on their suspected role in biological mechanisms, which can often be flawed and lead to irreproducible associations (MANOLIO *et al.* 2009). The strategies employed in GWAS focus on minimising the occurrence of spurious results and increasing the power to detect loci that predispose individuals to complex diseases.

The power of an association study can be defined as the probability of the study to detect a significant signal in a sample of individuals, when the true signal

actually exists in the underlying population (ZONDERVAN and CARDON 2004). The power of GWAS is dependent on several factors, including; (a) the amount of variation captured by the genotyping chip, (b) the extent of LD between the genotyped markers and risk alleles, (c) the frequency and effect size of the risk alleles, (d) the genetic heterogeneity of the sample population and (e) the sample size being analysed.

The availability of dense genotyping chips, containing hundreds of thousands of SNPs that effectively cover much of the common variation (minor allele frequency (MAF) greater than 5%) across human genome, have made it possible to carry out GWAS based on LD between the typed markers and diseases at unknown loci (WTCCC, 2007). In other words, if a causal SNP is not genotyped in a data set, its effects could still potentially be detected through LD with polymorphisms that are typed. Furthermore, the international HapMap Project has served to power GWAS by improving both the design and analysis of these studies by incorporating SNPs that best cover the common variation in the genome (BARRETT and CARDON 2006).

The power of GWAS is also directly dependent on the number and frequency of disease susceptibility alleles (PENG and KIMMEL 2007). SNPs with a low minor allele frequency (MAF) are not suitable for GWAS, because the studies generally do not have adequate power to detect them, unless their effect sizes are large (MIYAGAWA *et al.* 2008). In the GWAS carried out 2000 cases and 3000 controls by the WTCCC (2007), for example, using a p-value threshold of 5×10^{-7}, the power to detect SNPs with a relative risk of 1.5 (MAF > 5%), was 0.798. It is clear that even large studies of this nature, can only detect variants with a

frequency greater than 5% or a relative risk greater than 1.5, with approximately 80% power or more.

Rare and highly penetrant disease alleles, involved in most Mendelian disorders, are kept at low frequencies in the population through strong selection pressure (PRITCHARD and COX 2002) and would thus be unlikely to be identified by GWAS. In contrast, according to the common disease-common variant (CDCV) hypothesis, common diseases are likely to be caused by common variants with relatively small effects that may not have been subject to the same stringent selection pressure (LANDER 1996). As a result, even though their effect sizes are moderate to small, these alleles tend to reach frequencies of 5–10% or more in the population, making it possible for them to be detected by adequately powered association studies (BALDING 2006). Based on this hypothesis, the WTCCC (2007) successfully identified 24 independent association signals across 7 polygenic human diseases, using a significance threshold of $p < 5 \times 10^{-7}$, thereby providing evidence that GWAS provide a powerful method of detecting common disease susceptibility alleles in humans.

Common disease susceptibility alleles may, however, have frequencies lower than 5% or relative risks < 1.5 and therefore, GWA studies have focussed on different methods to improve their power to detect rarer alleles of smaller effect size that may influence diseases of interest.

1.3.1 Genome-wide Tagging

Despite the vast benefits of the International HapMap resource and the

advances in genotyping technology, typing large samples of individuals at every possible marker throughout the genome has, to date, been prohibitively time consuming and expensive. Therefore, scientists have focussed on identifying tag SNPs (small subsets of SNPs that effectively cover the genetic variation of large chromosomal regions) to increase the power of GWAS, while limiting the cost of unnecessary genotyping (HAO 2007).

Certain regions of the genome are known to exhibit high levels of LD between their composite loci (GABRIEL *et al.* 2002). Such 'LD blocks' are subjected to relatively low levels of recombination and therefore diversity within them is mostly driven by rare mutation events (MORRIS 2006). Genotyping individuals at just one of the loci within these blocks could potentially make the other loci redundant for mapping purposes, since the allelic combination is likely to be inherited as a conserved sequence (GOLDSTEIN and WEALE 2001). Indeed, across these portions of extended LD, genotype tagging can sufficiently capture entire haplotype structures, thus reducing the cost and time involved in disease association studies without sacrificing the power (CHAPMAN *et al.* 2003).

It has been observed, however, that LD patterns across the human genome are highly variable; displaying strong LD blocks that are largely unaffected by recombination, interspersed with regions of low LD exhibiting high levels of recombination (CARDON and BELL 2001). In terms of designing a strategy for effective genome-wide tagging, it would therefore be necessary to type a higher density of tag SNPs in regions of low LD, to ensure that these regions are still comprehensively represented in GWAS (HIRSCHHORN and DALY 2005). Furthermore, to maintain efficient coverage of tag SNPs across different

populations, it is important for researchers to use information about the evolution-driven, population-specific patterns of LD (DE BAKKER *et al.* 2006). For example, in recently admixed African populations, a larger number of tag SNPs will be needed to account for the high variability and specific LD structure that is commonly observed (HIRSCHHORN and DALY 2005). While the benefits of genome-wide tagging are clear, association studies would be plagued with low power to detect associations if, when designing the study, care was not taken to account for differences in population LD.

1.3.2 Imputation

Closely linked to tagging SNP-based approaches, is a method known as imputation, which is used in GWAS to predict the genotypes at SNPs that are either missing or not directly typed in a sample.

In one application, imputation can be used to predict the value of missing data points in a study, from the observed data, so that ultimately the imputed genotypes are used as observed values in the sample (MARCHINI *et al.* 2007). This is useful in association studies, since missing data have the potential to decrease the overall power of a study and simply deleting samples with missing data could lead to biased estimates, particularly if the remaining cases do not fairly represent the entire sample. Furthermore, if missing data points are merely replaced with the mean value of the sample, variance estimates of the population parameters could be artificially low and imputing missing values would provide more powerful and accurate tests of association.

Imputation can also be used to predict the genotypes at untyped SNPs in a sample, based on an available reference panel containing a dense set of marker SNPs (MARCHINI and HOWIE 2010). In this scenario, imputation methods work by identifying haplotypes that are common to both the individuals in a study and to the individuals in the reference panel and can then use this shared information to impute the genotypes at SNPs that have not been typed in the study. In this way, imputation complements genome-wide tagging methods, by choosing SNPs for the reference set that are highly correlated with, and hence good predictors of, the untyped SNPs (SERVIN and STEPHENS 2007). As previously mentioned, the international HapMap Project has provided researchers with a high-density reference panel of SNPs that can be used for this type of imputation. Using HapMap to impute SNPs not typed in the initial study, increases the power to detect susceptibility variants, since these imputed genotypes can be used to increase the number of SNPs that are tested for association.

Imputation can either be used to provide a “best guess” genotype, based on the imputed genotype with the highest posterior probability, or it can provide dosage values, where the probabilities for each of the possible genotypes are “averaged”. Taking the best guess imputed genotype as the true value, is the easiest and most computationally efficient method of analysing imputed data, but unless the correlation between the imputed and true genotypes is high, this method may be less powerful than the dosage method (ZHENG *et al.* 2011). Analysis based on the dosage probabilities may be more powerful in these cases, since this method takes account of uncertainty at the imputed genotypes (ZHENG *et al.* 2011).

Generally, in the case of tightly linked SNPs, imputation can be extremely reliable due to the conservation of specific allele sequences. In this way, imputation can potentially extract the maximum amount of information from a dataset, by preserving important characteristics such as means, variances and regression parameters, whose accuracy and precision is likely to have been eroded by missing data. Although the development of cheap, high-throughput genotyping assays have made large-scale association studies a reality, most on-going association studies genotype only a small proportion of SNPs in the region of study (SERVIN and STEPHENS 2007). Therefore, imputation has been used extensively in GWAS, because with a much larger set of SNPs, association with diseases can be tested across a much denser grid of SNP markers throughout the genome (MARCHINI *et al.* 2007).

1.3.3 Population Stratification

Patterns of sequence variation that are currently evident in the human population are non-random. They have been shaped through mutation and the evolutionary forces that act upon them such as migration, genetic drift, admixture and natural selection (HENDRICK 2000b). Population stratification can therefore occur in studies when sub populations that have an over-representation of certain variants are pooled with sub populations of differing allele frequencies. Under these circumstances, important signals could potentially be masked in the sample as a whole or similarly, allele frequency differences between cases and controls can manifest as significant disease associations, when, in reality, they may reflect the results of a number of different population factors, such as evolutionary history, migration, gender differences or other independent

processes (CARDON and BELL 2001).

GWAS test for population stratification by comparing allele frequency differences across relevant geographical regions and constructing quantile-quantile (q-q) plots of the observed versus the expected association statistics (LEE *et al.* 2010). Evidence of stratification can be easily identified from these q-q plots, as an increase in the slope of the line, and corrections for any sub-structure can then be made. Genomic control (measured as a statistic λ) is one method that is used to account for population stratification (FREEDMAN *et al.* 2004), which is calculated by taking the median of the X^2 test statistic and dividing it by 0.456 (DEVLIN *et al.* 2001). Another tool that can be used to alleviate the unwanted effects of population stratification, is principal component analysis (PCA) (ZHU *et al.* 2008). In this method, the first principal component (or axis) accounts for as much of the variability in the data as possible, and each axis thereafter, accounts for as much of the remaining variability as possible. In data sets with ancestry differences between samples, these axes tend to reflect the geographic history of the samples and can be adjusted for in association studies to ensure homogeneity among cases and controls (REICH *et al.* 2008).

1.3.4 Correcting for multiple testing in large data sets

Another challenge facing GWA studies is the problem of multiple testing, which arises when multiple independent tests of association are carried out on the same set of data. Multiple testing increases the chance of spurious associations occurring purely by chance, and the larger the number of SNPs being tested for association with the disease, the more chance these type 1 errors will occur.

One way to account for multiple testing is the use of the Bonferroni correction. This correction states that if an experimenter is testing *n* hypotheses on a set of data, the way in which the type I error rate can be controlled, is to test each individual hypothesis at a statistical significance level of *α/n*, where *α* is the chosen type I error rate (usually 5%) (BLAND and ALTMAN 1995). The Bonferroni correction, however, can become too stringent in cases where there are hundreds of thousands of variables to be tested, since many SNPs across the genome are in LD, and therefore these tests are not independent of one another. In the case of GWAS, using this correction is likely to decrease the power of the study to detect true associations.

The false discovery rate (FDR) is a more powerful way of controlling type I errors than the Bonferroni method. FDR controls the expected proportion of false positive results based on the distribution of p-values in a sample (KANG *et al.* 2009). While FDR has the advantage over the Bonferroni method of having increased power to detect associations, this approach to multiple testing may discourage additional analyses beyond single SNP tests, that could identify important associations (BALDING 2006).

1.3.5 Meta-Analysis

While more accurate genotyping technologies, improvements in the understanding of human sequence variation and statistical developments such as SNP tagging and imputation have reduced the occurrence of false positive GWAS results, the reproducibility of many putative disease variants is yet to be

achieved (ZEGGINI and IOANNIDIS 2009). Generally speaking, this irreproducibility is a product of the small genetic effect sizes in common diseases and studies are often underpowered to detect these effects with confidence due to their small sample sizes. However, by combining samples from different collaborations, in a process known as meta-analysis, this much desired increase in power could be attained since both the sample size and number of variants being examined, will be increased (ZEGGINI and IOANNIDIS 2009).

Meta analysis works by combining summary statistics from each study, for example effect sizes or odds ratios and their error measurements in the form of variance or confidence intervals (ZEGGINI and IOANNIDIS 2009). One of the biggest problems facing any meta-analysis is selection bias. This can be introduced into a meta-analysis through inconsistencies in data collection, phenotype measurement, missing data, imputation accuracy and standardisation of covariates across the different studies, to name a few. Furthermore, publication bias can skew meta-analysis results away from the null hypothesis, since studies that obtain negative results tend not to publish them and therefore these negative findings are omitted from the analysis (NAKAOKA and INOUE 2009). Therefore studies that meta-analyse data from different cohorts need to account for population substructure, the presence of related individuals, study-specific covariates and other ascertainment-related issues, in order to circumvent the biases that can easily be introduced in this type of analysis (WILLER *et al.* 2010).

A major advantage of meta-analysing GWAS results, is that summary statistics can be used instead of individual-level data, which alleviates some of the growing privacy concerns for study participants. Thus, studies can increase their

power to detect associations by pooling data, without the constraint of protected phenotypic information.

Despite the methods undertaken by GWAS to increase power to detect common susceptibility alleles and to decrease the prevalence of spurious association results, these studies have failed to account for a large proportion of the genetic risk influencing complex diseases. Researchers are now looking for novel methods to explain what has become known as the “missing heritability” of common diseases.

1.4 The missing heritability of complex traits

While GWA studies based on the CDCV hypothesis have proven to be successful in pinpointing common loci that contribute to the phenotypic variability of complex diseases, a large portion of the heritability of these traits remains unaccounted for. Even for traits such as human height, with an estimated heritability of 80%, only 5% of the phenotypic variance can be explained by the more than 40 associated loci identified by GWAS (VISSCHER 2008). One possible, albeit contradictory, hypothesis for the elusive genetic component of common disease, is that multiple rare variants each contribute moderate effects to the disease of interest (MORRIS and ZEGGINI 2010), known as the common-disease rare-variant (CDRV) hypothesis. Therefore, instead of common genetic markers being in LD with common, weak effect variants in a population, the markers might be indicative of many strong-effect variants that vary from individual to individual (ROBINSON 2010). The associations between rare higher-impact variants and common markers are known as synthetic associations

(DICKSON *et al.* 2010).

In the case of synthetic associations, causal variants may not be identified by current methods, since only few individuals in a population may have the same strong-effect variant. If sequencing is carried out on just a small subset of the population (as is the case when searching for common variants) the effects of multiple rare variants would probably be disregarded as noise in association testing.

Both hypotheses have their place in disease research and in the case of early-onset Alzheimer's disease, for example, both common and rare alleles are responsible for the disease phenotype (MCCLELLAN and KING 2010). Common susceptibility alleles have small effects, but because of their relatively high frequency, they contribute more to the overall risk for a disease. Conversely, rare variants with large effects may give important insight into disease aetiology and have more impact on the development of drug therapies. With current sequencing technologies and genome information databases such as HapMap, it has only been feasible to search for variants with population frequencies greater than 1%. Rare variants are therefore not only too low in frequency to be detected by current genotyping platforms, but their effect sizes are too small to be detected by linkage analysis (MANOLIO *et al.* 2009). The 1000 Genomes Project, however, is hoped to ensure that rare variants can also be assayed on a genome-wide, high-throughput scale.

Since the successful completion of both the HapMap and Human Genome Projects, scientific interest has turned to the 1000 Genomes Project, which aims

to catalogue human genetic variation at an unmatched level of resolution, thereby enabling the potential detection of new, rare functional variants and the improvement of existing methods of imputation (VIA *et al.* 2010). The 1000 Genomes Project is an international, multidisciplinary research effort that aims to sequence the genomes of more than one thousand unidentified participants, at genetic variants with MAFs from as low as 1%, from a number of different ethnic groups (www.1000genomes.org). Faster and more economical technologies have now been developed to enable such an undertaking and the consortium could potentially discover more than 95% of all genetic variants and make these findings publicly accessible. Furthermore, it aims to estimate the population frequencies, historical haplotype structures and linkage disequilibrium patterns of polymorphic alleles. Such a thorough knowledge of genetic variation across the human genome is hoped to provide scientists with a valuable tool to detect both rare and common genetic variants that are associated with all types of human diseases.

Even though the CDCV vs. CDRV debate implicates different strategies for identifying genetic variation that predisposes individuals to complex diseases, attention should be focussed on establishing the degree to which common and rare variants affect a particular phenotype, rather than looking at a right and wrong answer (SCHORK *et al.* 2009). GWAS is likely to remain an efficient way of investigating unexplained heritability of common disease, since GWAS signals may point to regions of the genome harbouring both rare and structural variation (MANOLIO *et al.* 2009). In order to make the most out of existing and future GWA studies, samples sizes could be increased, studies could be expanded to include non-European samples, gene-gene and gene-environment interactions could be

investigated and phenotyping could be improved. What is certain, is that improvements in sequencing technology and projects like 1000 Genomes will integrate structural variation, as well as rare and common SNPs, into an extensive catalogue of human variation. Undeniably, this superior resolution will increase the potential of GWAS to detect variants associated with diseases and other human traits (ALTSHULER *et al.* 2010).

It is imperative, however, that researchers focus not only on new ways to generate information from these studies, but also on the methods they need to analyse it appropriately. In order to capitalise on the enormous amount of available genetic data for mapping complex disease genes, it is necessary to develop model selection methods that are capable of handling large amounts of data and identifying important associations between markers and disease traits (FAN and KNAPP 2003). To date, GWAS has been successful in identifying single SNPs associated with complex diseases, but multi-locus model selection methods could provide more power to detect these associations since they consider the combined effect of more than one locus.

1.5 Thesis aims and outlines

Despite the success of GWAS, explaining the missing heritability in complex diseases continues to challenge researchers. Multiple common alleles of small effect size, gene-gene interactions and structural variation are just some of the factors that are thought to contribute significantly to this unexplained portion of disease variability. Based on these challenges, the overall aim of this thesis was to investigate suitable multi-locus model selection methods that could be applied

to genome-wide association data, as a powerful approach to understanding the genetic variability of common disease.

Initially, this investigation focussed on finding and evaluating multi-locus model selection techniques that could be effectively applied within a GWAS framework to detect variants of varying effect size and population frequency. Chapter 2 of this thesis outlines the multi-locus methods that were used for the purpose of this study and highlights the advantages and disadvantages of each method. In chapter 3, simulations were designed to compare these four multi-locus model selection methods, with the aim of establishing which of these methods had the highest power to detect indirect associations that exist between marker SNPs and disease phenotypes. It was further aimed to establish a not only a powerful model selection method, but one that had a controlled type 1 error rate (5%) and had the computational efficiency necessary to handle the large amounts of data generated by genome-wide association studies.

After the simulations were carried out to establish the most effective and accurate multi-locus selection method, the next aim was to compare the results of traditional single-SNP methods of analysis, with multi-locus analysis. To do this, after the relevant quality control measures were taken, multi-locus model selection was carried out on the T2D data from the Wellcome Trust Case Control Consortium (WTCCC) and the results were compared to those obtained by the single-SNP analysis implemented in the study. Since, theoretically, considering more than one locus at a time provides a more powerful method of detecting putative loci that contribute to the risk of common diseases, it was necessary to see if there were in fact any signals in the data that could be detected by these

methods, which the original study may have been underpowered to identify.

To date, GWA studies have identified many regions across the genome that have been shown to affect disease outcomes. However, unravelling the genetic architecture of these regions and accurately mapping the markers that are most closely associated with disease, is an important next step. The next aim of this investigation was thus to use a multi-locus method of analysis, to fine map a region on chromosome 17 that is associated with the onset of childhood asthma. The 17q21 region is an established risk locus for childhood asthma, but only one strong primary signal is known to reside within this locus. This fine-mapping step was carried out in order to identify putative secondary signals in this region, which may contribute further to the risk of childhood asthma and better explain the genetic contribution of this locus to the disease.

The final aim of this thesis was to establish an accurate, effective and computationally efficient method of detecting putative gene-gene interactions that play a role in CAD. Pair-wise interaction analyses are often limited by computational capacity and plagued by multiple testing problems, due to the exponential increase in the number of data points when all the possible pair-wise interactions are considered, as well as their marginal effects. Therefore, in the final chapter of this investigation, we aimed to carry out a computationally feasible pair-wise interaction analysis, using a multi-locus model selection approach that could circumvent the problems associated with multiple testing in undetermined problems, where the number of predictors far exceeds the number of observations in a data set.

Chapter 2: Comparison of multi-locus model selection methods in genome-wide association studies

Statistical inference in human genetics is widely based on hypothesis testing. In GWAS, the association between genotypes at a single SNP and a disease phenotype can be evaluated by testing the null hypothesis (H_0) that no association exists between the SNP predictor and the disease response (BALDING 2006). Typically, it is assumed that the effect of these putative susceptibility alleles will be additive in nature. In other words, the heterozygote risk is thought to be intermediate between the two homozygote risks (BALDING 2006). The most flexible way to investigate the association between genotype and a quantitative trait or dichotomous disease phenotype is through linear and logistic regression, respectively.

2.1 Linear Regression

In statistical genetics, inferences about populations are generally based upon sample data. Linear regression attempts to model the relationship between an independent predictor variable and a dependent response variable, by fitting the following equation to the observed data:

$$\hat{Y} = b_0 + bx$$

where $\hat{Y}$ is the response variable (in GWAS this would be a disease phenotype),

x is the independent predictor (a marker SNP for example), b is the regression coefficient and b_0 is the intercept term. Under the assumption of an additive model, genotypes are coded as 0,1 or 2, according to the number of minor alleles. The sign and size of the regression coefficient provides valuable information about the effect of x on $\hat{Y}$. Large coefficient values indicate that x has a large effect on $\hat{Y}$ and the sign of the coefficient is indicative of the type of relationship between the two variables. Specifically, a positive coefficient sign results, if $\hat{Y}$ increases as x increases and a negative sign is present if $\hat{Y}$ increases when x decreases.

In genome-wide association studies, large datasets consisting of hundreds of thousands of variables are common place and thus multiple linear regression is generally utilised to identify linear combinations of independent variables that are capable of systematically explaining variation in the dependent variable. The fitted equation is in the form:

$$\hat{Y} = b_0+b_1x_1+ b_2x_2 +...+b_px_p$$

where $x_1,\ldots,x_p$ are the predictor variables, $b_1,\ldots,b_p$ are their respective regression coefficients and b_0 is the intercept term.

The most common mathematical method used to estimate the regression coefficients that yield the best fitting straight line through the data, is ordinary least squares (OLS). This technique seeks to minimise the sum of the squared differences between the observed (Y_i) and predicted ($\hat{Y}_i$) squares of Y. In other words, coefficients are chosen such that

$$\Sigma e_i^2 = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma(Y_i - b_1x_1 - b_2x_2 - \ldots - b_ix_i - b_0)^2$$

is the smallest possible value and e_i is referred to as the "error". In geometric terms, OLS finds the best-fitting line for the observed data, by minimising the sum of the squares of the vertical deviations from each data point to the line. Coefficient estimation is an important step in association testing, since the coefficient value is what is used to estimate the effect size and direction of the variable being measured. Linear regression models are used to test single-SNP associations with continuous outcomes and therefore play an important role in genetic association studies (BALDING 2006).

The use of linear regression and OLS for data analysis can only be justified, however, if certain principal assumptions are met. These are; a linear relationship must exist between the dependent and independent variables, the error terms must be both independent and normally distributed and the variance of the error terms should be constant across the independent variables (known as homoscedasticity). An example of this approach is seen in Chapter 3, where model selection methods that utilise ordinary least squares estimates are used to analyse simulated quantitative response variables as well as looking for associations to lipoprotein(a) levels in the blood.

Generally speaking, however, complex disease status (i.e. case/control) is the response variable of choice in most human genetics studies. Linear regression and ordinary least squares are therefore not appropriate for handling such dichotomous response variables, because the underlying assumptions are not

met. In particular, the assumption of homoskedasticity is violated, since the error variance is directly dependent on the independent variables and thus not uniformly distributed. Secondly, the error is not normally distributed since the response variable can only take on values of 0 and 1. Thus, using linear regression to analyse case/control outcomes could result in errors in the predicted probabilities from the model.

2.2 Logistic Regression

Logistic regression provides a method for modelling a binary response variable, which can take on values of 0 (generally assigned to controls) and 1 (generally assigned to cases) (BEWICK *et al.* 2005). The mathematical framework underlying logistic regression centres around the logit (natural logarithm of the odds ratio), the simplest example of which is derived from a 2x2 contingency table. Contingency table analysis looks to see if a particular genetic variant or genotype occurs more often in cases than controls. For example, a 2x2 contingency table can be used to assess allelic association to a disease, by looking at the frequency distribution of two alleles (A/B) across cases and controls (OHASHI and TOKUNAGA 1999). The odds of an event is defined as the ratio of the probability that an event occurs to the probability that it fails to occur, so the odds ratio (OR) in this case, would be the ratio of specific allele carriers to non carriers, in cases compared to controls (FOULKES 2009). These statistics are calculated to give the approximate increase in disease risk for individuals that carry a particular allele compared to those that do not (LEWIS 2002). Similarly, a 2x3 contingency table is sometimes used to test the association between a bi-allelic SNP with three possible genotypes (AA/AB/BB) and a case-control disease response (LEWIS

2002).

Contingency tables can be further analysed using Pearson's chi squared test, which tests the observed values for departure from the expected values (calculated by assuming the variables are independent of one another) across cells in the table. For example, for genotype-based analyses, a 2x3 table is constructed, and a 2-degree of freedom chi squared test is performed to determine whether the frequencies of any of the genotypes are associated with the phenotype.

Analysis of data in the form of contingency tables usually assumes that there are only a small number of variables that might show an association with the trait of interest. If a large number of variables are considered, methods based on regression modelling are necessary (FORSYTHE *et al.* 1971). Logistic regression can, however, be extended to model the combined effect of independent predictors on a binary response variable, while making no assumption about the distribution of the independent variables.

The logit function is defined as the natural logarithm (ln) of the odds of an event occurring (for example an individual being a disease case). That is,

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

where p is the probability of observing the disease. The probability that a particular disease results from a linear combination of genetic variants can therefore be modelled by the following equation:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_kx_k + e$$

Where b_0 is the intercept term, x_1 to x_k are a set of predictor variables (independent variables), b_1 to b_k are their respective regression coefficient values estimated from the data, and e is the error term that accounts for differences between the observed outcome values and those predicted by the model. These coefficient values are the allelic log odds ratios, which indicate the disease risk carried by one allele, relative to the other. The estimated coefficient values for logistic models are calculated using maximum likelihood, which involves finding the values of the parameters that maximise the likelihood of reproducing the data, given the parameter estimates (HARTZ and ROSENBERG 1975). These equations can generally only be solved by an iterative procedure, easily expressed in computational form (BAGLEY *et al.* 2001).

Regression analysis methods have many strengths in the field of statistical genetics. Some of these advantages include stable parameter estimation algorithms, precise parametric models, easy inclusion of important covariates such as gender and country of origin and the availability of reliable software to carry out such analyses (CANTOR *et al.* 2010). Problems can arise, however, when the number of possible predictors in a data set far exceeds the number of observations. In these cases, regression methods tend to fit the noise in the data rather than any actual effects (termed overfitting), and can falsely identify statistically significant relationships as a result. GWA studies generally analyse data sets that contain many hundreds of thousands of SNPs, where the vast majority of variants have zero effect and only a fraction actually affect the

response. Therefore, model selection strategies can be used to avoid overfitting, by extracting sparse models (that include only those variants with some effect on the response) from a large number of possible predictors.

The goal of model selection is to incorporate scientifically meaningful predictors into an easily interpretable model from which accurate predictions can be made. Accuracy of a model's parameter estimates increases as the mean square error (MSE) of the estimate decreases and thus, minimising MSE is an important aspect of model selection. MSE is affected by two factors, bias and variance, and the relationship can be shown by the following equation:

$$MSE = Variance + Bias^2$$

Bias reflects the ability of a model to accurately predict the data and therefore, the more complex the model, the lower the bias (BURNHAM and ANDERSON 2002). Bias will thus be introduced if predictors with even very small effect sizes are omitted from the final model (a process known as underfitting) and thus increase the MSE of the parameter estimates. In contrast, as more variables are included in the model, so the variance, and consequently the MSE, increases too. High variance is a reflection that the model is possibly overfitting the data and would therefore not be a reliable predictive tool and the parameter estimates may not replicate when encountering new data (WERTHEIM *et al.* 2010). Clearly, a key feature in model selection is finding the optimal trade off between these two measures by attempting to only include predictors with significant effects on the phenotypic outcome up until the point that the amount by which $bias^2$ is decreased by the addition of a new parameter is less than the increase in

variance upon the addition of the same parameter.

In general, model selection is centred around the principle of parsimony, which states that for two models that fit the data equally well, the simpler model with fewer parameters should be chosen (POSADA and BUCKLEY 2004). Variable selection methods use this principle to choose the smallest number of possible predictors that are able to accurately explain the outcome of interest, and thereby select a model that has predictive potential on a wide range of data sets.

2.3 Stepwise Selection

Forward selection and backward elimination are two widely used methods for variable selection. Forward selection begins with a null model and variables are tested for inclusion into the model one at a time. The most significant variable is added, provided its P-value is below a pre-set critical level (BERK 1980). Forward selection, although easy to implement, has drawbacks, in that every addition of a new variable in the final model, may render one or more of the already included variables, non-significant.

An alternate approach, which avoids this problem, is backward elimination. Under this approach, the model begins with fitting all the variables of interest and the least significant variable is dropped one at a time, according to a pre-set critical level (BERK 1980). The reduced models are re-fitted and the process is repeated until all the variables in the final model are statistically significant. In backward selection, however, variables that may have been significant when included in the final model could have been dropped early on in the selection

process, and important information could therefore be lost. This method may also be computationally impractical for large numbers of predictor variables.

Another difficulty that arises when implementing forward and backward selection, particularly when a large number of variables are tested for potential inclusion in the model, is how to specify the predetermined critical level. If this level is too low, the number of false positive events could become unacceptably high, but on the other hand, if the level is set too stringently, power could be lost and important variables may be excluded from the model.

One of the ways in which to overcome the uncertainty of deciding on a feasible critical level, is through the use of information criteria. When performing model selection based on stepwise procedures, functions such as the Akaikie Information Criterion (AIC), can potentially enhance the accuracy of model selection. This criterion includes or excludes variables based on the overall AIC score of a particular model which is measured by the equation:

$$AIC = 2k - 2\ln(L)$$

where k is the number of parameters in the model and L is the maximized value of the likelihood function for the estimated model.

The AIC is a tool for model selection that, given a data set, can rank several competing models according to their AIC, with the one having the lowest AIC being the best (BURNHAM and ANDERSON 2002). While increasing the number of variables included in the model improves the goodness of fit, the AIC also

includes a penalty for increasing the number of predictors in the model, thereby discouraging overfitting.

Similar to the AIC is the Bayesian information criterion (BIC), which can be explained by the following equation:

$$BIC = -2\ln(L) + k.\ln(n)$$

As with the AIC, k is the number of model parameters and L is the maximized value of the likelihood function for the estimated model. In this equation, n is the sample size, and it is clear that the BIC will impose a harsher penalty on free parameters than the AIC, since sample sizes are generally large and the smaller the BIC, the better the model fit.

When the number of predictors in a data set is large, or the predictors themselves are correlated with one another, OLS estimates, while maintaining low bias, often have unacceptably high variance. Prediction accuracy in these situations can sometimes be improved by shrinking the values of the coefficients towards zero because, while bias will increase slightly, the variance of the predicted values will be reduced (TIBSHIRANI 1997). Typical variable selection methods, such as forward selection and backward elimination, often have low prediction accuracy and small changes in the data can result in widely differing models being selected (EFRON *et al.* 2004). Traditional shrinkage methods, such as Ridge regression, are more stable, but produce un-interpretable models because no predictors are actually dropped from the final model (TIBSHIRANI 1996). The least absolute shrinkage and selection operator (Lasso) is able to

shrink coefficients and set others to 0, thereby producing stable and parsimonious models. Lasso has been shown to greatly improve the ordinary least squares estimate in terms of accuracy (LENG *et al.* 2006).

2.4 Least Angle Regression

Lasso is an adaptation of a procedure known as least angle regression (LAR). According to Efron *et al*. (EFRON *et al.* 2004), LAR, as with forward selection, begins with all coefficients equal to zero. It identifies the predictor most correlated with the response, (x_{j1}) and takes the largest step possible in the direction of this predictor until another orthogonal (or uncorrelated) predictor, (x_{j2}), has as much correlation with the current residual. While at this point, forward selection would continue along x_{j1}, LAR chooses to follow the direction equiangular between the two predictors until a third variable (x_{j3}) becomes the most correlated with the residual. LAR then continues equiangularly between x_{j1}, x_{j2} and x_{j3}. This concept is illustrated in the case of two predictors in Figure 2.1.

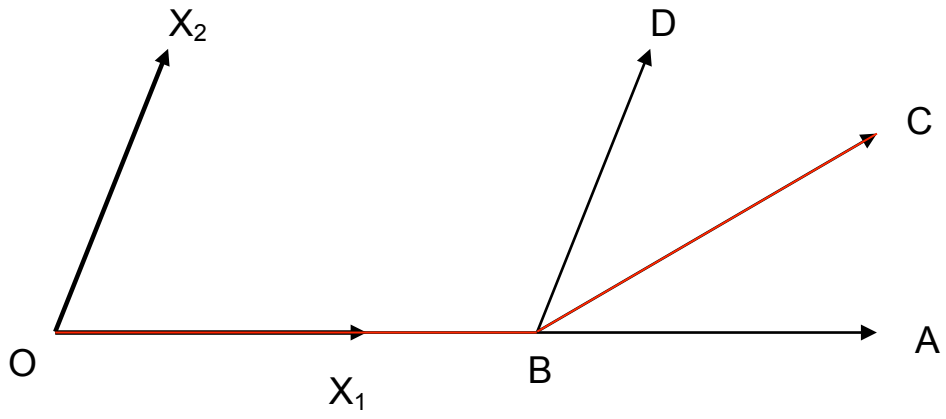


Figure 2.1 The LAR algorithm for two predictors. C is the ordinary least squares fit where C = Y = B_1 X_1 + B_2 X_2. A is the forward selection fit after one step; the second step proceeds to C. Least angle regression jumps from O to B in one step, where B is the point such that BC bisects the angle ABD. At the second step it jumps to C. LASSO follows a path from O to B, then from B to C.

In this way, LAR is considered a less greedy version of the forward selection technique that may have eliminated useful predictors in the second step that may have been correlated with x_{j1} (EFRON *et al.* 2004).

A simple modification of the LAR algorithm implements the Lasso, a version of OLS that constrains the sum of the absolute regression coefficients. LAR with the lasso modification (abbreviated to LARS) calculates all possible Lasso estimates for a given problem. LARS minimises the sum of squared error subject to a bound (*t*) on the sum of the absolute value of the regression coefficients.

In other words, for the vector of regression coefficients $b = (b_1,b_2,\ldots,b_m)$,

minimise $$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} x_{ij} b_j\right)^2$$

subject to $$\sum_{j=1}^{m} | b_j | \leq t$$ ((EFRON *et al.* 2004))

In this way, LARS is referred to as a constrained version of OLS. If t is too large, estimates will approximate their OLS solutions, whereas if t is too small, the coefficients of important variables in the model could be artificially low or even zero. In practice, a cross validation (CV) type approach for selecting the degree of shrinkage, while computationally more expensive, may lead to better predictions (MADIGAN and RIDGEWAY 2004). Cross-validation is a re-sampling method that is used for estimating generalisation error. In cross-validation, data is divided into subsets and a model is fitted to one subset (the training set) and then validated on the other subset (validation set). This process is repeated

multiple times using different subset sizes and the validation results are averaged over all the cycles. In this way, the degree of shrinkage that corresponds to the least amount of error may be evaluated.

Model selection is commonly used to find easily interpretable, parsimonious models, but can sometimes be hindered by the computational burden of searching through a large number of possible models. LARS is a shrinkage and selection method for linear regression that is now being used as a computationally feasible model selection technique (ZHAO and YU 2006). The computational burden for all *m* steps in the LAR algorithms, is of the same order as that required for a least squares solution for the full set of *m* covariates and can therefore be used to handle the large amount of information generated by genome-wide association studies (EFRON *et al.* 2004).

Forward and backward selection methods have often been criticised for providing misleading information, particularly in large samples, because by focussing on any one single model, uncertainty about quantities of interest can be underestimated (RAFTERY 1995). It has been shown, particularly in the linear regression context, that doing this can yield misleading results, often tending to reject null hypotheses more often than the nominal levels would suggest, and to produce confidence intervals that are too narrow (VIALLEFONT *et al.* 2001). Bayesian model averaging (BMA) is a Bayesian approach to model selection and provides a coherent mechanism for accounting for this model uncertainty (HOETING *et al.* 1999).

2.5 Bayesian Model Averaging

While the frequentist approach to probability theory defines the probability of an uncertain event as the frequency of that event based on previous observations, Bayesian probability theory uses evidence or observations from data, combined with prior beliefs, to infer the probability of an event. Rather than simply accepting or rejecting hypotheses, the Bayesian approach calculates the probability of a hypothesis given the data, based on the prior probability and the likelihood of that hypothesis.

If more than one model could fit a given set of data reasonably well then each of these models may still lead to vastly different conclusions about the data at hand (RAFTERY 1995). Bayesian model averaging (BMA) produces a posterior probability for each possible model and a posterior probability for each predictor, based on their respective weighted average over a particular model space. BMA therefore provides a coherent mechanism for accounting for this model uncertainty (HOETING *et al.* 1999).

Raftery *et al*. (RAFTERY 1995) provides the algebraic explanation behind BMA. If Δ is the quantity of interest, an effect size for example, then its posterior distribution given data D is

$$pr(\Delta \mid D) = \sum_{k=1}^{K} pr(\Delta \mid M_k, D) pr(M_k \mid D)$$

This is an average of the posterior distributions under each of the considered

models (M_1,..,M_k) weighted by their posterior model probability. The posterior probability for model M_k is given by

$$pr(M_k \mid D) = \frac{pr(D \mid M_k)\,pr(M_k)}{\sum_{i=1}^{K} pr(D \mid M_i)\,pr(M_i)}$$

where

$pr(D \mid M_k) = \int pr(D \mid M_k)\,pr(b_k \mid M_k)\,db_k$

is the integrated likelihood of model (M_k), b_k is the vector of parameters of M_k, $pr(b_k \mid M_k)$ is the prior density of b_k under M_k, $pr(D \mid b_k, M_k)$ is the likelihood and $pr(M_k)$ is the prior probability that M_k is the true model (given that one of the considered models is true).

Computational efficiency is an important aspect of model selection because the model space increases exponentially with every predictor variable. Since p possible predictors would result in 2^p potential models, averaging over all these models is often not computational feasible. Two approaches were taken by Raftery *et al*. (RAFTERY *et al.* 1997) to reduce the computation burden of model averaging. The first approach was to apply the Occam's Window principle to regression models, to reduce the set of models over which a model average is computed. It does this by disregarding models that predict the data far less accurately than the most favourable model does and also by discarding models that do not fit the data as well as any of their simpler sub-models. Secondly, Raftery *et al*. (RAFTERY *et al.* 1997) implemented a Markov Chain Monte Carlo algorithm (MCMC), which allows the model space to be searched for the most probable models and variables without fitting all possible models (FRIDLEY 2009).

In the presence of model uncertainty, both these model averaging procedures offer better predictive performance than one selected model would have done (MADIGAN and RIDGEWAY 2004).

While stepwise selection and LARS can be applied to data sets consisting of hundreds of thousands of predictor variables, the computational demand of BMA on large data sets remains prohibitively high and is therefore difficult to implement in the field of complex disease GWAS.

In any Bayesian analysis, the choice of the prior distribution of models and parameters is important. If little information is known about the data, no model will be initially favoured over others, and the prior probabilities of each parameter being tested should be set to 0.5/*n*, where *n* is the number of parameters. If, on the other hand, accurate prior knowledge that favours some combination of variables is available, this can and should be incorporated into the prior model probabilities (VIALLEFONT *et al.* 2001). The inclusion of subjective prior probabilities has come under much criticism, but large sample sizes reduce the effect of prior probabilities on the overall result, and will therefore minimise this potential problem.

According to Montgomery *et al*. (MONTGOMERY 2010), it is almost never appropriate to use BMA to conduct theory-free searches of the model space and analysis should begin with the careful development of a model based on theory and previous research. Thus, although BMA accounts for uncertainty in model selection, where typical variable selection procedures do not, it is not, on its own, a panacea in the field of statistical genetics.

Chapter 3: Evaluation of model selection strategies

In Chapter 2, I described a number of alternative multi-locus model selection techniques that could be used to detect association with markers in a GWAS. The aim of this chapter was to compare and evaluate the different methods in order to find the most accurate and efficient approach for selecting the best subset of predictor variables, from a number of candidate markers. In the context of GWAS, these predictor variables would be the genotypes at SNPs and model selection would be used to identify SNPs that may contribute, either independently, or in combination with other SNPs, to the risk of common disease phenotypes. Since there is no definitive method of performing multi-locus model selection on genome-wide association data, the simulations were designed to test these methods under a number of different assumptions, in order to establish which approach would be the most powerful and efficient method of detecting disease susceptibility loci within a GWAS framework.

The initial statistical approach focussed on comparing four model selection strategies; namely forward selection, backward elimination, least angle regression with the lasso modification (lasso) and BMA. These methods assumed additive allelic effects models and the simulation studies were used to compare and evaluate their type I error rates and power. As an example of a practical human genetics application, these methods were then assessed in terms of their ability to detect a known association signal for increased lipoprotein(a) (lp(a)) levels in the blood.

For this investigation, data were simulated using HAPGEN (www.stats.ox.ac.uk/~marchini/software/gwas/**hapgen**.htm), a programme that uses HapMap genotype data and a fine scale recombination map to generate haplotypes in a sample of unrelated individuals. We used this software to simulate a quantitative trait value, based on a single causal SNP, for 1000 individuals. The quantitative trait value was randomly generated from a N(μ,1) distribution, in order to simulate a complex disease phenotype for each individual in the population, where μ depended on the underlying association model. Then, for the genotypic values, 60 SNPs from the Enr112 encode region of chromosome 2 were selected from SNPs on the Affymetrix 500K genotyping chip, and from these 60 SNPs, 10 were randomly chosen and stored for analysis. These 10 SNPs might not necessarily include the causative variant and this process served to thin out the high degree of LD that exists between the tightly linked markers of the encode region, so that the final 10 SNPs were not highly correlated with one another. SNPs with minor allele frequencies less than 1% were not chosen for the simulated data set since GWA studies are designed to detect common, higher frequency variants in the population (generally MAF > 5%) (BALDING 2006).

All the analyses were carried out in the R statistical analysis system, using packages from the Comprehensive R Archive Network (CRAN). To carry out forward and backward selection, the *stepAIC* function was used and the *LARS* and *BMA* packages were used to implement lasso and Bayesian model averaging respectively.

3.1 Type I error

We began by considering the type 1 error rate of each method, at the nominal significance threshold of 5%. Data were simulated as described above, where the quantitative trait for each individual was simulated from a N(0,1) distribution, irrespective of the genetic data. The simulated populations were designed such that no association existed between the quantitative trait and any of the markers, and thus, after carrying out the four model selection methods, any SNP included in the final model would represent a false positive (type I error). Therefore the type I error rate was calculated for each of these methods by adding up the number of simulated data sets in which at least one SNP was included in the final model.

Each model selection method was used to analyse 1000 simulated data sets. The analyses were initially carried out using the R package's default parameters. It was shown that all four methods had high type I error rates and it was therefore necessary to make adjustments to these parameters in order to make the selection criteria more stringent and decrease the type 1 error rate of each strategy.

For forward and backward selection, the final models were selected based on the overall AIC. However, Table 3.1 indicates that the type I error rates based purely on this selection criterion, were 0.735 and 0.811 for forward and backward selection respectively. To make these methods more stringent, only variables in the model with p-values less than 0.005 were considered. This value was chosen based on the Bonferroni correction, such that for 10 SNPs, where 10

independent tests are carried out, the p-value cut-off would be 0.05/10 = 0.005. The simulations were run again, but both methods still had type I error rates greater than 0.05 (Table 3.1) and therefore further adjustments were made.

Within the *stepAIC* function, the parameter, *k*, is the multiple of the number of degrees of freedom used for the penalty, and only when *k*=2, are the models chosen according to the genuine AIC. By increasing the value of *k*, a harsher penalty is imposed on the number of SNPs included in the models. Therefore, the value of *k* was increased to 4 and the simulations were repeated with and without the imposition of the p-value cut-off. The results are laid out in Table 3.1.

Table 3.1 Type I error rates of forward selection and backward elimination based on their performance over 1000 simulated populations. The error rates were measured when *k*, the multiple of the number of degrees of freedom used for the penalty, was set at 2 (the default setting) and 4. The error rates were then measured again at these parameters, but only the SNPs with significant p-values (p<0.005) were considered to be included in the final model.

Technique	k=2	k=4	p<0.005 (k=2)	p<0.005(k=4)
Forward	0.735	0.315	0.086	0.058
Backward	0.811	0.317	0.089	0.058

The type I error rate was calculated to be 0.058 when the value of k was set to 4 and the p-value cut-off was applied to each of the models (Table 3.1). This value is an estimate of the true error rate, 0.050, and falls within the calculated binomial 95% confidence intervals for that estimate (0.0500 $\pm$ 0.0135 or 0.0365 to 0.0635), making the type 1 error rate consistent with 5%. Thus, for any further analyses, it was accepted that forward and backward selection, carried out with these adjustments using the stepAIC package, would yield type 1 error rates of approximately 5%.

Lasso was carried out on the same 1000 simulated populations as the stepwise

selection methods. The entire lasso path of solutions was computed initially. The path begins with all the coefficient vectors equal to 0, i.e. the null model, and thus corresponds to an infinitely small shrinkage parameter. As the shrinkage parameter is increased, so the constraint on the sum of the coefficient vectors ($\sum|b_j|$) is relaxed and the coefficient values are allowed to increase. The coefficients of the most significant variables become non-zero first and continue to increase until the coefficients of the variables in the model have reached their ordinary least squares estimates. This is referred to as the saturated model and is equivalent to removing the constraint on the coefficient values (i.e. no shrinkage).

Ten fold cross validation (CV) was used to determine the point along the path at which the predictions regarding the coefficient values were extracted. The cross-validated mean squared prediction error for the lasso model was computed at 100 fractions of the model path. These fractions are the values at which the CV curve was computed, expressed as a fraction of the saturated model, ranging from 0 to 1 and incremented at 0.01. Figure 3.1 is a graphical explanation of this process.

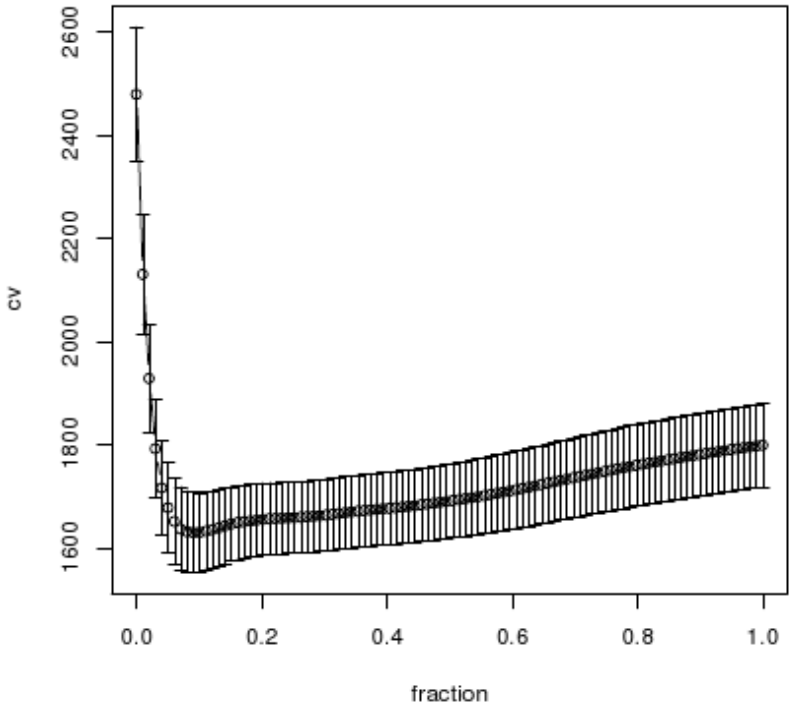


Figure 3.1 Cross validation curve showing the cross-validated mean squared prediction error for the lasso model. The y-axis represents the cross-validated squared prediction error values (CV). The fractions plotted along the x-axis are the abscissa values at which the CV curve is computed, as a fraction of the saturated model. These values range from 0 to 1, (incremented at 0.01) where 0 is the null model and 1 is the unconstrained ordinary least squares estimates of the variable coefficients.

The fraction along the path that produced the lowest cross-validated squared prediction error was 0.08 (Figure 3.1). Any point along the path corresponds to a particular shrinkage estimate, since this shrinkage estimate yields a particular set of coefficient estimates that are used to calculate the prediction error of the model. When the prediction error is at its lowest, the parameter estimates of the lasso model will be the best approximation of the true values and therefore the shrinkage parameter that corresponds to the point along the path with the lowest prediction error, should be used for the lasso analysis.

For the initial lasso analyses, any variable with a non-zero coefficient, was regarded as a variable in the final model (false positive), and yielded a type I error rate of 0.472. Therefore, even though cross validation has been suggested

as a non-subjective and accurate way of selecting the shrinkage parameter, in order to decrease the occurrence of false positive variants included in the final model, a more stringent method was needed. It was evident upon closer examination of these coefficients, that the majority had values so close to 0, that if a cut-off value was applied across the coefficient values of the variables included in the final model, many of these false positive results would be excluded from the model. Therefore, instead of only variables with coefficients equal to 0 being dropped from the model, variables where the absolute value of their coefficients was less than 0.1 SD units, were also dropped from the final model. This coefficient cut-off was chosen, based on empirical evidence, to yield the desired type 1 error rate and has the same effect on lasso model selection as decreasing the lasso shrinkage parameter, and consequently tightening the constraint on the coefficients, to an equivalent degree. The absolute value of the coefficients was used to ensure that the cut-off was applied equally to variables that were negatively or positively correlated with the response. Based on this adjustment, the type I error rate dropped to an acceptable 0.053 (Table 3.2).

Table 3.2 Type I error rate of lasso. The error rates were measured when only variables with non zero coefficient values were included in the final model and when variables with absolute coefficient values (|coefficient|) > 0.1 were included in the final model.

Technique	Non-zero coefficients	|coefficients| > 0.1
Lasso	0.472	0.053

Finally, BMA was carried out on the simulated data sets. The *bicreg* function within the *BMA* package was used to carry out BMA for linear regression. The output of this function cannot be interpreted in the same way as the output from the other methods, since BMA does not only look at one model. Instead, this method outputs a number of competing models as well as the posterior probability of each of these models. In an ideal world, the null model should have

a posterior probability of 1, since the algorithm was searching through data with no underlying signal, but since BMA is designed to account for model uncertainty, the output generally included the null model with the highest posterior probability, as well as models with at least one variable, shown to have much lower posterior probabilities. Type 1 error in this case, was calculated by adding up the number of times out of 1000, an alternative model had a higher posterior probability than the null model. Table 3.3 shows that, based on the aforementioned strategy, the type I error rate was 0.066. Two options were therefore considered to make this method more stringent. Firstly, only models with a posterior probability of greater than 0.5 were considered. Under this consideration, the false positive rate dropped to 0.028 (Table 3.3). Secondly, only models that had a posterior probability that was greater than twice the posterior probability of the null model were considered.

Table 3.3 Type I error rates of Bayesian model averaging for linear regression models. In the first analysis, only the model with the highest posterior probability was considered (BMA). In the second analysis, models with posterior probabilities greater than 0.5 were considered (PP >0.5) and in the final analysis, only models with posterior probabilities greater than twice that of the null model (PP> 2PPnull) were considered.

Technique	BMA	PP>0.5	PP> 2PP null
BMA	0.066	0.028	0.031

The two adjustments ensured that the false positive rate was well below 0.050, but in the interests of this investigation, the *PP>2PP null* adjustment was chosen to be applied to the power simulations. Since the posterior probabilities of two models are taken into consideration, this adjustment is based on the properties of model uncertainty, and therefore takes advantage of this unique property of BMA in model selection.

3.2 Power of model selection methods to detect quantitative disease risk variants

The next aim was to evaluate and compare the power of the different model selection techniques. Power here, is defined as the probability that a statistical significance test will reject a false null hypothesis at a nominal significance threshold of 5%. In other words, it measures the ability of each method to detect an association signal in the data, given that the effect actually exists. The simulated data sets were therefore generated to contain a signal and the post-hoc power of each method to detect the association was measured.

As described in the previous section, HAPGEN was used to simulate 1000 independent data sets for analysis. A quantitative trait value and ten SNPs were selected for analysis and a disease causing SNP was randomly picked along with the other ten and stored as an eleventh SNP in the data set. This SNP, however, was not included in the matrix of marker SNPs analysed by the regression methods. In reality, the actual causative SNP may not have been genotyped in the study and therefore the power simulations were designed to test the ability of these methods to identify the markers in LD with a nearby causative SNP. The adjustments made to the different techniques, in order to decrease their respective type I error rates to approximately 5%, were used in the power analysis. Therefore, the power of each of the methods to detect associations with the quantitative trait, correspond to the type I error rates measured in section 3.1.

Initially, a linear regression analysis between the actual causative SNP and the disease phenotype was performed for each of the 1000 simulations and the p-values from each of these tests were recorded. The number of times the p-values of these regression analyses were below 0.05 was calculated out of 1000 and used as a measure of the maximum amount of power that could potentially be achieved by any of the models selection methods. Therefore, the power was calculated by taking the number of times out of 1000 simulations, each selection technique included at least one SNP in its final model as a proportion of the maximum amount of power that could be achieved under each simulation. This power is shown in the *Causal Variant* column in Table 3.5.

A number of different population parameters were varied across the analyses in order to test the power of the methods under a range of assumptions. Firstly, the percentage of total trait variance accounted for by the disease allele (heritability (h^2)) was varied at 1%, 2%, 5% and 10%. Secondly, four genetic models, namely an additive model, a dominant model, a recessive model and a model that conferred a heterozygous advantage were considered in the analysis. The values of the homozygous dominant, heterozygous and homozygous recessive individuals in the population, under each of the four models, were coded as shown in Table 3.4.

Table 3.4. The mean values given to each of the parameters in the four genetic models used in the power analysis. The four models considered for the analysis were additive, dominant and recessive models and the final model conferred a heterozygous advantage.

Model	Homozygous Dom	Heterozygous	Homozygous Rec
Additive	2	1	0
Dominant	2	2	0
Recessive	0	0	2
Heterozygous	0	2	0

Finally, the disease allele frequency was varied at 0.01, 0.05, 0.10, 0.20 and 0.50. Causal variants were selected from an ENCODE region to have MAFs within 10% of the specified frequency. Therefore, for a disease frequency 0.01, the simulated disease frequency would have had a value between 0.009 and 0.011. The power of each of the model selection methods to detect the signal in the data at the significance levels stipulated in section 3.1 is indicated in Table 3.5.

Table 3.5 Results of the power simulation studies. The simulated data was analysed using four model selection techniques namely forward selection (Forward), backward elimination (Backward), lasso penalized regression (Lasso) and Bayesian model averaging (BMA). The actual causal variant was regressed against the quantitative trait, and the number of times out of 1000 the p-value was less than 0.05, was recorded as a fraction (Causal variant). Disease allele frequency was varied at 0.01, 0.05, 0.10, 0.20 and 0.50 for an additive model (a), a dominant model (d), a model conferring a heterozygous advantage (h) and a recessive model (r). In turn, these combinations were varied for a disease heritability of 1%, 2%, 5% and 10%.

Freq	Model	h^2 (%)	Forward	Backward	Lasso	BMA	Causal Variant
0.01	a	1	0.078	0.085	0.121	0.085	0.399
0.01	a	2	0.085	0.086	0.124	0.083	0.453
0.01	a	5	0.122	0.139	0.185	0.136	0.638
0.01	a	10	0.122	0.139	0.187	0.135	0.829
0.01	d	1	0.080	0.083	0.136	0.093	0.379
0.01	d	2	0.084	0.094	0.164	0.099	0.462
0.01	d	5	0.146	0.156	0.226	0.155	0.627
0.01	d	10	0.151	0.162	0.288	0.167	0.802
0.01	h	1	0.078	0.101	0.189	0.112	0.364
0.01	h	2	0.093	0.106	0.193	0.102	0.416
0.01	h	5	0.141	0.150	0.268	0.137	0.632
0.01	h	10	0.239	0.261	0.335	0.241	0.746
0.01	r	1	0.054	0.060	0.001	0.058	0.179
0.01	r	2	0.076	0.086	0.001	0.080	0.219
0.01	r	5	0.078	0.087	0.000	0.088	0.284
0.01	r	10	0.216	0.226	0.108	0.219	0.382
0.05	a	1	0.145	0.166	0.535	0.138	0.895
0.05	a	2	0.227	0.274	0.625	0.235	0.913
0.05	a	5	0.501	0.544	0.743	0.520	1.000
0.05	a	10	0.730	0.761	0.850	0.730	1.000
0.05	d	1	0.171	0.197	0.601	0.170	0.912
0.05	d	2	0.170	0.201	0.596	0.174	0.679
0.05	d	5	0.550	0.585	0.821	0.567	1.000
0.05	d	10	0.779	0.806	0.918	0.759	1.000

Freq	Model	h^2	Forward	Backward	LASSO	BMA	Causal Variant
0.05	h	1	0.153	0.190	0.579	0.152	0.721
0.05	h	2	0.266	0.316	0.676	0.258	0.798
0.05	h	5	0.768	0.794	0.891	0.759	0.999
0.05	h	10	0.571	0.600	0.840	0.570	1.000
0.05	r	1	0.050	0.059	0.130	0.076	0.150
0.05	r	2	0.077	0.085	0.069	0.097	0.248
0.05	r	5	0.136	0.155	0.039	0.142	0.408
0.05	r	10	0.174	0.192	0.036	0.181	0.576
0.1	a	1	0.240	0.250	0.661	0.258	0.841
0.1	a	2	0.429	0.442	0.755	0.452	0.847
0.1	a	5	0.692	0.707	0.840	0.681	0.932
0.1	a	10	0.813	0.826	0.914	0.800	1.000
0.1	d	1	0.275	0.290	0.691	0.289	0.767
0.1	d	2	0.489	0.499	0.806	0.493	0.968
0.1	d	5	0.857	0.863	0.937	0.847	1.000
0.1	d	10	0.237	0.246	0.662	0.235	1.000
0.1	h	1	0.234	0.246	0.650	0.253	0.690
0.1	h	2	0.436	0.446	0.766	0.447	0.928
0.1	h	5	0.633	0.648	0.781	0.631	0.939
0.1	h	10	0.869	0.878	0.944	0.858	1.000
0.1	r	1	0.097	0.112	0.386	0.102	1.000
0.1	r	2	0.138	0.152	0.375	0.152	0.945
0.1	r	5	0.287	0.301	0.345	0.290	0.965
0.1	r	10	0.453	0.466	0.359	0.468	0.891
0.2	a	1	0.287	0.288	0.601	0.285	0.753
0.2	a	2	0.508	0.516	0.809	0.518	0.948
0.2	a	5	0.755	0.766	0.905	0.759	1.000
0.2	a	10	0.848	0.861	0.976	0.847	1.000
0.2	d	1	0.289	0.307	0.623	0.313	0.886
0.2	d	2	0.533	0.548	0.826	0.544	0.997
0.2	d	5	0.769	0.780	0.901	0.765	1.000
0.2	d	10	0.895	0.913	0.957	0.891	1.000
0.2	h	1	0.195	0.204	0.623	0.211	0.560
0.2	h	2	0.354	0.380	0.749	0.375	0.863
0.2	h	5	0.654	0.668	0.872	0.658	0.997
0.2	h	10	0.812	0.818	0.930	0.800	1.000
0.2	r	1	0.120	0.135	0.493	0.120	0.522
0.2	r	2	0.224	0.233	0.600	0.232	0.610
0.2	r	5	0.467	0.484	0.680	0.482	0.931
0.2	r	10	0.691	0.701	0.782	0.694	0.998
0.5	a	1	0.334	0.348	0.759	0.340	0.883
0.5	a	2	0.546	0.570	0.858	0.556	0.994
0.5	a	5	0.855	0.869	0.962	0.849	1.000
0.5	a	10	0.970	0.975	0.992	0.968	1.000
0.5	d	1	0.290	0.305	0.711	0.294	0.834
0.5	d	2	0.547	0.557	0.751	0.547	0.876
0.5	d	5	0.821	0.838	0.954	0.814	0.993
0.5	d	10	0.949	0.958	0.986	0.947	1.000
0.5	h	1	0.067	0.079	0.389	0.069	0.072
0.5	h	2	0.062	0.068	0.374	0.069	0.074
0.5	h	5	0.061	0.065	0.315	0.064	0.080
0.5	h	10	0.055	0.066	0.356	0.069	0.085
0.5	r	1	0.222	0.238	0.649	0.225	0.677
0.5	r	2	0.377	0.398	0.773	0.379	0.934
0.5	r	5	0.701	0.713	0.904	0.685	1.000
0.5	r	10	0.910	0.923	0.976	0.901	1.000

The results of the power simulations provide a number of platforms from which the power of the regression methods to detect true association signals in the data, can be compared. These different platforms were simulated to investigate the effects that changes in disease allele frequency and heritability, as well as the underlying genetic model, have on the overall power to detect a statistically significant association between a marker SNP and a quantitative trait.

As mentioned, the power was calculated as a proportion of the theoretical maximum and indicated in the p-value column. For example, according to Table 3.5, when the population was simulated to have a disease allele frequency of 50%, a heritability of 1% and an underlying additive genetic model, the relative power of lasso penalized regression to detect the true signal was 759/883 = 0.859.

The results of the simulations laid out in Table 3.5 indicated some general trends in the effect of the underlying genetic architecture of the quantitative trait and marker SNPs, on the power of the model selection methods. Both an increase in disease allele frequency and heritability had a positive effect on relative power. This effect is clearly illustrated by comparing the relative power within the same genetic model for the same model selection method, for varying disease allele frequencies and heritabilities. For example, under an additive model, for a disease allele frequency of 20%, forward selection has a relative power of 287/753=0.381, when the heritability is 1% and a relative power of 848/1000=0.848, when the heritability is 10%. Under the same additive model, when only the disease allele frequency is altered to 50%, forward selection has a

relative power of 334/883=0.378 when heritability is 1% and a relative power of 970/1000=0.970 when heritability is 10%.

The general increase in power of all four selection methods, as both the disease allele frequency and the heritability are increased incrementally, are illustrated in Figure 3.2 and Figure 3.3 respectively.

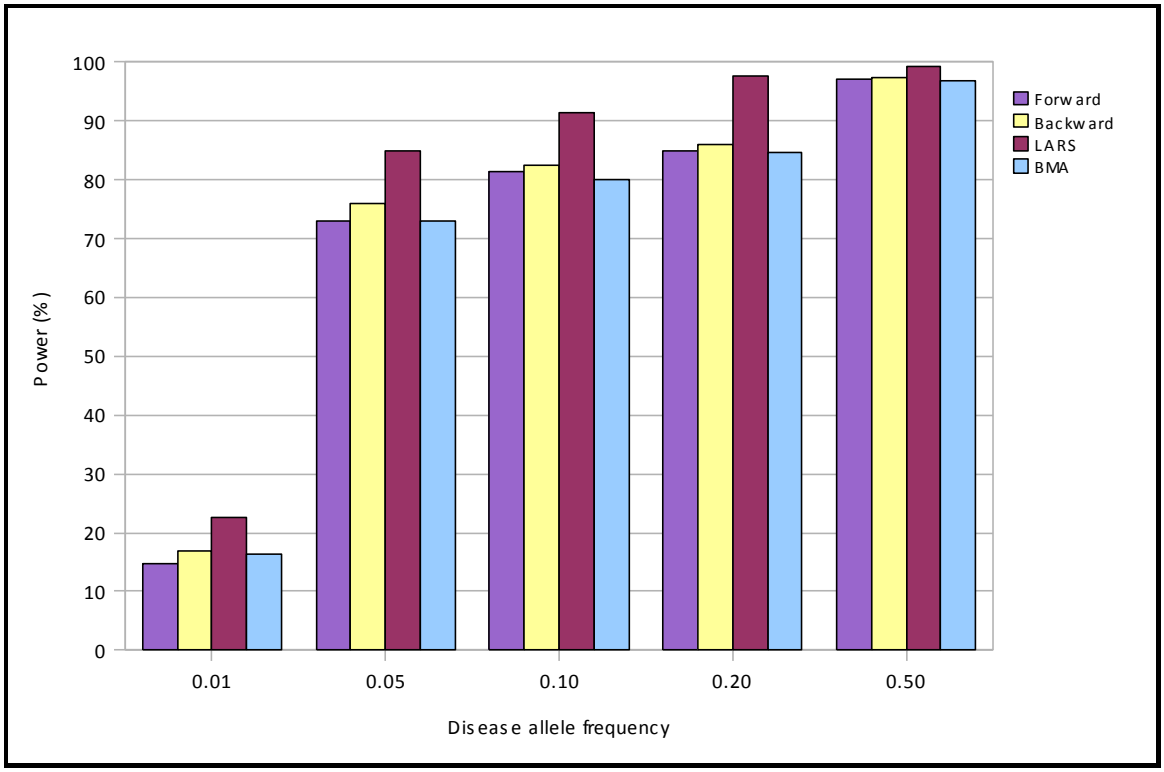


Figure 3.2 Effect of disease allele frequency on power for the four model selection methods assuming an additive model. While the heritability was fixed at 10%, the disease allele frequency was varied at 0.01, 0.05, 0.10, 0.20 and 0.50.

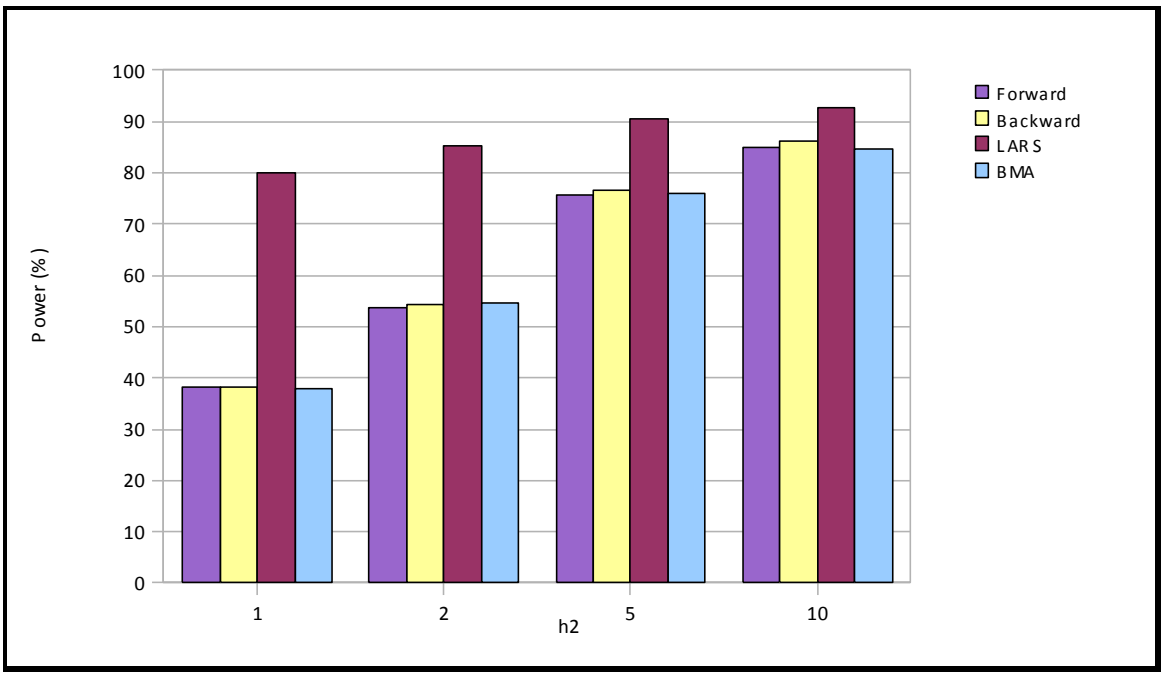


Figure 3.3 Effect of disease heritability (h^2) on power for the four model selection methods assuming an additive model. The heritability was set at 1%, 2%, 5% and 10% and the disease allele frequency was fixed at 0.20.

Between the different model selection methods, it was evident that lasso had the highest power across the various simulated environments, followed by backward elimination, forward selection and finally BMA. It is important to note here, however, that the type I error rate of BMA was lower than both backward and forward selection, i.e. 0.031 (Table 3.3) as opposed to 0.058 (Table 3.1). Thus, prior to the application of these methods to complex disease studies, the preferred method would depend on the compromise a researcher would be willing to make between type I error and power. Alternatively, the BMA method could be further adapted to have a type I error rate of 0.05, which would increase its power to detect associations by relaxing the penalty imposed on the number of parameters included in the final model. Another noticeable trend was that, under the additive model, the model selection methods had the highest power, followed by the dominant model. Generally, under the recessive model, the methods had the lowest power.

A few of the simulation results did not follow the generally observed trends and therefore deserve further attention. Firstly, under the recessive model, for a disease allele frequency of 0.01, the power of each model selection technique is very low. For a recessive trait, an affected individual would need to have two copies of the disease allele and therefore when the disease allele occurs at a frequency of 1/100, the chances of that occurring is 1/10000. Taking into consideration that the population size is only 1000 individuals, it is clear that the power to detect any kind of association would be extremely low. The lasso method appeared to be particularly sensitive to these circumstances, having essentially no power.

Another seemingly inconsistent result was evident under the heterozygous advantage model, when the disease allele frequency was simulated to have a value of 0.50 in the population. It was shown that the regression test between the disease causing SNP and the trait phenotype (the theoretical power maximum) was extraordinarily low. The power of the methods to detect the underlying signal, expressed as a proportion of the theoretical power maximum, were all high, although the actual numbers were also very low. A reason for this could be that the underlying model assigned one parameter to each of the three genotypes and the linear regression test attempted to fit a straight line through the data. Since the disease allele frequency was so large, one of the homozygous group could possibly have had the effect of pulling the fitted line almost horizontal, thereby indicating little to no association. This concept is illustrated by the left hand graph in Figure 3.4, where Ho(2) is the previously mentioned homozygous group. When the disease allele frequency was only

0.10, however, the homozygous group (Ho(2) in the right hand graph) did not have as large a pull on the fitted regression line, and an association was then evident. Indeed, the models that are fitted in this study, assume a linear trend in allelic effects and, since the heterozygous advantage model is not well approximated by this linear tend, the power is expected to be low.

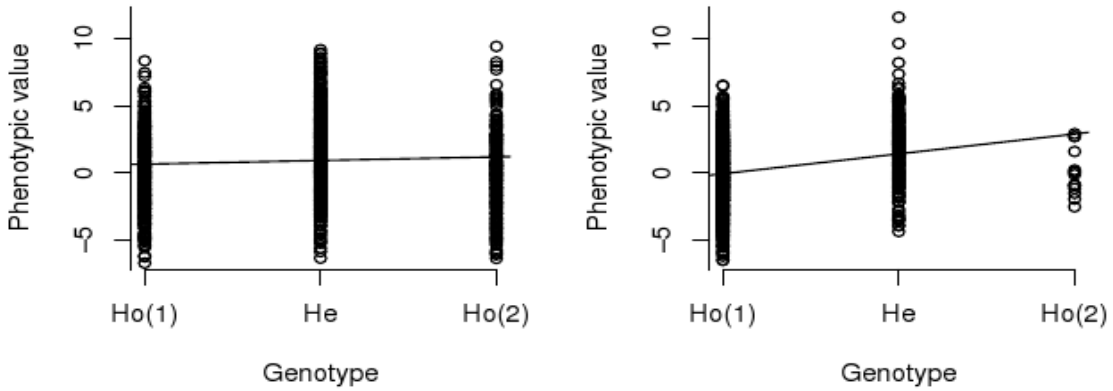


Figure 3.4 Linear regression analysis of the disease causing SNP with the disease phenotype under the assumption of a heterozygous advantage model and a disease heritability of 10%. The left hand graph was produced when the disease allele frequency was simulated to be 0.50 in the population while the right hand graph shows the result when the disease allele frequency was simulated to be 0.10 in the population.

The power of lasso penalized regression under the heterozygous advantage model and a disease allele frequency of 0.50, far exceeded the theoretical maximum under the same conditions. This result is not fully understood, but could possibly reflect the advantage of using shrinkage methods for stable and reliable parameter estimation. The aforementioned heterozygous effect results in very small, insignificant coefficient values (with a regression slope close to 0) for ordinary least squares estimates, but since this effect is uniform across all the SNPs, and the coefficients of the variables included in the lasso model are increased and decreased in relation to one another, it could have less of an effect on lasso penalized regression as the other methods.

The required level of power for experimental designs is widely accepted to be above 80%. Under the assumption of the additive and dominant genetic models, only the lasso method had sufficient power by these standards, to detect an association when disease allele frequency was at least 0.10. It should be noted, however, that the power of the methods in these simulations could be artificially low because there could have been limited LD between the marker SNPs and the causative SNP. Although the simulation programme initially made use of high density haplotype data, a subset of 60 SNPs was first randomly selected from this data and then thinned down to just ten, so linkage disequilibrium between SNPs could have been thinned out by the final set of ten selected SNPs. Additionally, the simulated populations consisted of only 1000 individuals. This would be regarded as a small sample size by GWAS standards and could decrease the power of the methods to detect associations.

Computational efficiency is an important aspect of model selection because the model space increases exponentially with every predictor variable. While stepwise selection and lasso penalized regression could be applied to data sets consisting of hundreds of predictor variables, the computational demand of BMA on large data sets is prohibitively high. Although BMA was successfully applied to these simulated populations, the data consisted of only ten SNPs, which is not a realistic situation in the field of human genetics. It is also worth noting, however, that the simulation study did not fully exploit or give credit to the ability of BMA to account for model uncertainty. Thus the potential advantage of BMA over other multi-locus methods, to produce more accurate and stable predictions because of this ability, should not be ignored. Another advantage of Bayesian

analysis over other methods, which was not demonstrated in this investigation, is the use of prior information to weight probabilities (GAJEWSKI *et al.* 2008). In this way, potentially functional SNPs, for example, may be assigned a higher weight. The subjectivity of this prior information, however, has been highly criticised and therefore care should be taken when incorporating priors into any analysis (SHOEMAKER *et al.* 1999).

In general, the power simulations successfully demonstrated the positive effect of disease allele frequency and heritability on power. Furthermore, when an additive effect was simulated at the disease locus, the power of model selection was at its greatest, which was expected since the linear regression adopted by each of the techniques assumes additivity.

3.3 Extension of simulation study to incorporate multi-locus effects

Section 3.2 focussed on assessing the power of four multi-locus regression techniques to detect the effect of one causal SNP on the phenotype of interest. However, the model selection strategies were evaluated with the purpose of finding a method that could ultimately be used to unravel the genetic architecture of common, complex human diseases. Since multiple loci generally contribute to the risk of complex diseases, the advantages of multi-locus over single-SNP methods of analysis in this field, lie in their ability to consider the effects of many loci simultaneously. It was therefore pertinent that the power of these methods to detect multiple causal loci was investigated and the study was extended to incorporate multi-variant effects.

HAPGEN was used, in a similar way to that described in section 3.1, to simulate both genotype data and a disease phenotype controlled by two variants in the same region. HAPGEN does this by picking two causative SNPs with MAFs that match (as closely as possible) the MAFs that are manually entered into the programme. The genotypes at each of these two loci are coded as 0, 1 and 2 and, since an additive model was assumed in this case, the combined genotype can be attained by adding together the genotypic values at each locus and calculating the mean response for each individual. The mean responses therefore range from 0-4. As before, data was generated for each individual by simulating the genotypes at 10 SNPs, as well as at the two causative loci.

For the multi-locus simulation study, two causative SNPs, of differing minor allele frequencies, were simulated as follows: 0.2/0.05; 0.2/0.1; 0.5/0.1; and 0.5/0.2. Furthermore, the heritability was varied at 1%, 2%, 5% and 10% and an additive genetic model was assumed both within and across causal loci. This particular investigation was designed as an extension of the first set of simulations, to test the power of the model selection methods to detect an association to a quantitative trait controlled by more than one locus. Since, under the assumption of additivity, the results of these multi-locus simulations were consistent with the trends seen in the previous power study (section 3.2), and it is commonly assumed that contributions to disease risk from individual SNPs are roughly additive (BALDING 2006), this study was not extended to other genetic models.

As with the single locus disease model, the adjusted model selection methods were applied to the simulated data sets and the power of each method was calculated by adding up how many times each of the methods was able to detect

at least one SNP in the data associated with the phenotype, out of the 1000 simulated data sets. For each simulation, the p-value of each causative SNP was calculated as well as the significance of the overall model containing both of these SNPs together. Initially, the significance of both causative SNPs was tested one at a time by regressing each SNP against the quantitative trait, and then, the number of times the p-values for each test were below 0.05 was added up out of 1000 simulations. These totals were recorded as P_1 and P_2 (Table 3.6) and represent the maximum power available to detect the effect of each SNPs individually.

Then, a regression analysis was performed where the combined effect of both causative SNPs was regressed against the quantitative trait, and the number of times the overall p-values for the combined model was below 0.05, was added up out of 1000 simulations and displayed as a fraction in the P_{model} column in Table 3.6. The latter test indicates the maximum power to detect the combined effect of both loci. Since two SNPs were simulated to each have an independent effect on the quantitative trait, it was expected that the number of times the overall p-value was significant would be higher than the number of times the p-value for each individual SNP was significant. Indeed, this result is shown in Table 3.6.

A combination method using both lasso penalized regression and BMA was introduced into this study and evaluated along with the other four methods. The motivation behind this combinatorial method stemmed from the practical considerations of applying either lasso or BMA to real data. Firstly, the cut-off applied to the lasso coefficients (|coefficients| > 0.1) to decrease the type 1 error

rate, was empirically based on the results of the simulation study and would not be appropriate when used on different data. This is because coefficient estimation in lasso is data dependent, since the coefficients of variables are increased and decreased in relation to one another, based on a particular shrinkage penalty. Using cross validation to establish the best shrinkage parameter, however, yielded a high type 1 error rate and thus could not be used in isolation. Furthermore, the cross validation method, used to determine the lasso shrinkage parameter, is computationally demanding, and this could potentially be computationally infeasible when carried out on large-scale GWA data.

On the other hand, BMA was shown to have low type 1 error rates and good power, but had the disadvantage of being computationally inefficient. In fact, BMA was unable to analyse more than 60 SNPs for a population of 1000 individuals. Even for sample sizes limited to 1000 individuals, the increase in the number of predictors being analysed significantly increases the computational demand of the analysis. For this reason, it was decided to combine these methods and test the power of this combined method (*Comb*), along with the other four (Table 3.6). Firstly, lasso penalized regression was carried out on the simulated data, using cross validation to estimate the shrinkage parameter. After this, BMA was carried out on all the SNPs in the final lasso model with non-zero coefficients to establish the final model from the combined method. No adjustments were made to either the lasso or BMA methods of selection for this combined strategy.

Table 3.6 Results of the multi-locus power simulations. Power to detect associations to a quantitative trait controlled by two loci, was compared between five model selection techniques; namely forward selection (Forward), backward elimination (Backward), lasso penalized regression (Lasso), Bayesian model averaging (BMA) and a combination of lasso and BMA (Comb). Each causative SNP was regressed against the quantitative trait and the number of times out of 1000 the respective p-values were below 0.05, was recorded (P_1 and P_2). The number of times the P value for the overall model containing both causative SNPs was below 0.05 was also recorded (P_{model}). The minor allele frequencies at the two causative loci (MAF1/2) were simulated as follows: 0.2/0.05; 0.2/0.1; 0.5/0.1; and 0.5/0.2. Furthermore, the heritability (h^2) was varied at 1%, 2%, 5% and 10% and an additive genetic model was assumed.

h^2;MAF1/2	Forward	Backward	Lasso	BMA	Comb	P_{model}	P_1	P_2
1; 0.2/0.05	0.215	0.213	0.690	0.280	0.273	0.904	0.899	0.491
2; 0.2/0.05	0.546	0.545	0.835	0.546	0.542	0.965	0.961	0.732
5; 0.2/0.05	0.795	0.792	0.922	0.781	0.781	1.000	1.000	0.916
10;0.2/0.05	0.908	0.908	0.960	0.904	0.901	1.000	1.000	0.980
1; 0.2/0.1	0.287	0.284	0.703	0.265	0.257	0.875	0.821	0.504
2; 0.2/0.1	0.548	0.548	0.865	0.530	0.527	0.985	0.932	0.801
5; 0.2/0.1	0.764	0.762	0.929	0.791	0.782	1.000	1.000	0.930
10;0.2/0.1	0.913	0.911	0.969	0.913	0.912	1.000	1.000	0.990
1; 0.5/0.1	0.244	0.242	0.715	0.230	0.226	0.911	0.798	0.492
2; 0.5/0.1	0.468	0.469	0.808	0.443	0.441	0.954	0.939	0.781
5; 0.5/0.1	0.789	0.785	0.938	0.765	0.762	1.000	1.000	0.933
10;0.5/0.1	0.921	0.918	0.969	0.919	0.916	1.000	1.000	1.000
1; 0.5/0.2	0.262	0.261	0.725	0.237	0.232	0.817	0.703	0.479
2; 0.5/0.2	0.598	0.595	0.887	0.534	0.528	0.919	0.903	0.797
5; 0.5/0.2	0.845	0.845	0.961	0.838	0.831	1.000	1.000	0.989
10;0.5/0.2	0.936	0.935	0.984	0.932	0.930	1.000	1.000	1.000

The results of this analysis show similar trends to the results of the analysis laid out in section 3.2. As expected, the power of each method increases with an increase in heritability and lasso generally showed the highest power across the different heritability and differing MAF estimates.

Table 3.6 shows that when the heritability was set at 1% and the MAFs of the two SNPs were set at 0.5/0.2 (highlighted blue), P_{model} is 0.817 while P_1 and P_2 are 0.703 and 0.479 respectively. What is interesting to note in this case, is that the lasso was able to detect an effect 725/1000 times, which is higher than the maximum power available to detect any one of the SNPs individually. In practical terms, this would mean that lasso has the potential to increase the power to detect multi-locus effects, compared to single-SNP methods of analysis, since

the combined effect of both SNPs is greater than any one on its own and lasso is able to detect this combined effect.

It was assumed that because HAPGEN used the same method to generate data for both the single and multi locus simulations, the type 1 error rates of each method would remain the same and therefore provide comparable power estimates. To be certain, however, the heritability estimates in the multi-locus study were set to 0 (i.e. the SNPs contribute no additive genetic variance to the overall variation of the trait) and the methods were re-run for the same combination of MAFs. Furthermore, it was necessary to check the type 1 error rate of the lasso/BMA combined method (*Comb*), which had not been done in the previous study.

Table 3.7 Type 1 error rate of five model selection methods; namely forward selection (Forward), backward elimination (Backward), lasso penalized regression (Lasso), Bayesian model averaging (BMA) and a combination of lasso and BMA (Comb). The disease heritability (h^2) was set at 0 so that there was no simulated genetic association in the data.

h^2 (%)	Forward	Backward	Lasso	BMA	Comb
0	0.054	0.053	0.048	0.039	0.030

The results, shown in table 3.7, indicate that the type 1 error rates of each method are approximately the same as those calculated in section 3.1 and therefore the assumptions for the power comparisons have not been violated. The type 1 error rates of the lasso/BMA combined method were lower than 0.05 and lower than the type 1 error rates of all the other methods. Consequently, the power of the combined method was lower than that of the other four methods (Table 3.6), but has the advantage of not relying on any parameter adjustments before or after the analysis is run.

While the results of the simulations studies showed that these model selection

methods were able to detect multiple disease risk factors across varying allele frequencies and disease heritabilities, the next step was to assess the performance of these model selection methods on actual data.

3.4 Application of multi-locus models of selection

Forward and backward selection, as well as a lasso/BMA combination, were used to search for the genetic variants that contribute to the risk of increased Lipoprotein(a) (Lp(a)) levels in the blood, which has been shown to increase the risk of developing coronary artery disease (BERG *et al.* 2002). This quantitative trait was analysed for 1189 individuals with CAD from the PROCARDIS study. Sample genotypes were extracted from a GWAS experiment using a 1M Illumina genotyping chip and the individual genotypes across 17825 SNPs on chromosome 6 were extracted. SNPs with low genotyping call rates (<95%), low minor alleles frequency (<1%) or Hardy-Weinberg disequilibrium in controls ($P=1x10^{-6}$) were excluded from the study. A strong signal of association was known to exist within this region.

For the analysis, the model selection methods required a complete dataset, with no missing genotypes, and therefore imputation was performed on the raw data using MACH 1.0. MACH 1.0 is a Markov chain-based haplotyper that can be used to infer missing genotypes in samples of unrelated individuals using the "best guess" genotype. In this imputation method, individuals with the most complete genotype information, were used as templates to infer the genotypes at missing markers for individuals with more missing data. Parameters were set so that the 200 haplotypes with the most complete genotype information in the

sample were used as templates to infer the genotypes at any missing markers.

In order to ensure that poorly imputed SNPs were not included in the analysis, any imputed SNPs with an R-squared value of less than 0.3 in the MACH output, were removed from the analysis. This cut off value is an estimate of the squared correlation between the estimated genotype scores and the true genotypes. In regions of strong LD, these R-squared values will be high, since genotyped markers will provide more accurate information about the missing genotypes. The genotype with the highest posterior probability was taken as the “best guess” genotype.

STATA software was then used to do a preliminary single-SNP analysis so that the specific region of the genome containing the signal could be targeted for analysis. A matrix containing the lp(a) response variable and 400 marker SNPs containing the signal of association (Figure 3.5) was created in R.

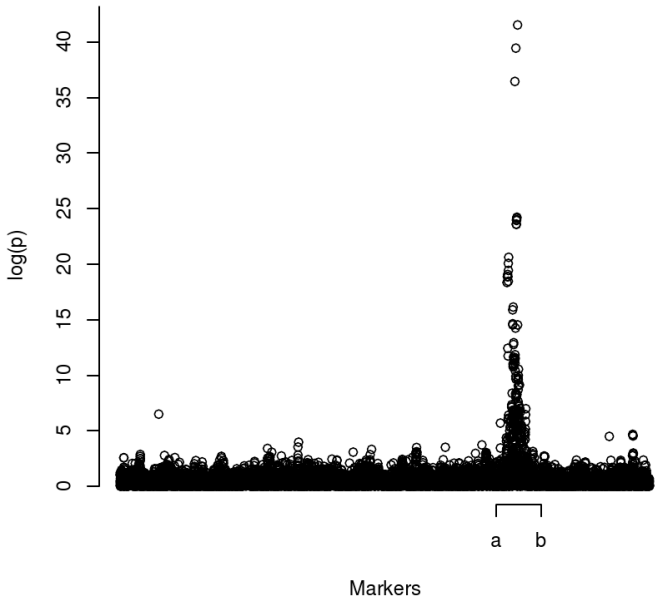


Figure 3.5 Single-marker analysis to test for the association between 17325 marker SNPs on chromosome 6 and Lp(a) levels in 1189 individuals. The signal peak is between **a** and **b** on the graph.

Forward and backward selection, as well as a lasso/BMA combination, were applied to these 400 extracted SNPs in order to identify the variants associated with Lp(a) levels. The results of this analysis are shown in Table 3.8.

Table 3.8 Models chosen by forward and backward selection and a lasso/BMA combination method when applied to an association region on chromosome 6 associated with lipoprotein(a) levels.

SNP ID	Position(bp)	**Forward/Backward Selection**			**lasso/BMA**	
		Regression estimates	Standard error	P-value	Regression estimates	Standard error
rs7757997	160602887	-93.261	23.301	6.68e-05	-	-
rs492315	160612684	99.415	22.700	1.30e-05	-	-
rs614564	160612709	97.365	22.793	2.10e-05	-	-
rs3125056	160655271	25.253	6.081	3.53e-05	11.659	3.483
rs3918291	160748132	44.200	8.739	4.93e-07	42.889	8.350
rs4708867	160762715	15.201	4.498	1.20e-4	15.282	4.742
rs2457012	160778961	-35.529	10.143	1.40e-4	-	-
rs2941383	160790466	-	-	-	12.120	2.780
rs7744658	160796346	-48.402	9.937	1.27e-06	-23.609	4.191
rs6919346	160880349	14.946	3.536	2.56e-05	17.715	2.773
rs3798220	160881127				67.803	7.602
rs9355296	160937983	119.233	28.401	5.93e-06	29.114	2.876
rs7770628	160938164	26.420	2.684	<2e-16	-	-
rs1321196	161001832	-	-	-	-24.789	2.454
rs9457997	161027008	-56.897	17.214	1.30e-4	-21.430	3.276
rs1652508	161048937	-	-	-	-10.278	3.262
rs4252121	161063624	38.027	10.355	1.30e-4	-	-
rs783166	161097227	-	-	-	-14.909	6.417
rs2064712	161136598	-	-	-	-10.409	4.696
rs783156	161151167	44.574	11.602	1.31e-4	-	-

Forward selection and backward elimination were carried out on the reduced matrix, using the parameters that corresponded to a type I error rate of 0.058 (section 3.1). The p-value cut-off was adjusted using the Bonferroni correction to $0.058/400 = 1.45 \times 10^{-4}$. Forward and backward selection produced identical final models consisting of 14 SNPs that were below the p-value threshold of 1.45×10^{-4} (Table 3.8).

In the case of lasso, the coefficient cut-off used in section 3.1 could not be applied since the lp(a) data was structured differently to the simulated data and consequently the distribution of coefficient values and the type I error estimates would be different to those discussed in section 3.1. In addition to this, it was not possible to carry out BMA using this R package on any more than 60 SNPs, due to its high computational demand. Thus, the lasso/BMA combination described in section 3.3 was used.

For the first lasso step of the combined method, a CV-estimated shrinkage parameter was used and this yielded 53 SNPs with non-zero coefficients.

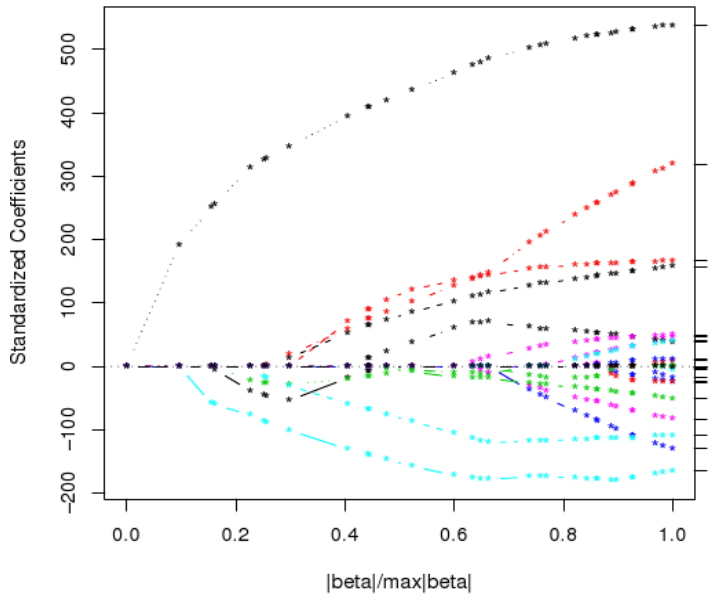


Figure 3.6 Coefficient pathway of the 30 most significant variables chosen by the lasso algorithm. The x-axis represents the absolute value of the constrained β-coefficient for each variable as a proportion of its maximum value for this particular pathway (|beta|/max|beta|). The y-axis measures the standardised coefficient value. Only the variables with non-zero coefficients are included in the model. For relatively low values of |beta|/max|beta|, only a few variables are included into the model with low coefficient values. As the algorithm progresses, more variables are included into the model as their coefficients become non-zero and approach their ordinary least squares estimates at |beta|/max|beta| =1. Each variable is represented by a different colour. The most significant SNP variant is the first to become non zero (first to enter the model) and consistently has the highest coefficient value (shown in black).

Figure 3.6 shows the initial lasso path for the 30 most significant SNPs. Including

all 53 variables in this figure resulted in a highly cluttered plot and therefore 30 were chosen as a way of visually explaining the lasso selection path. The most significant SNP is indicated by the black points and consistently has the highest standardised coefficient.

In the second step, BMA was carried out on the 53 lasso predictor variables with non-zero coefficients. This method assigned the highest posterior probability (0.332) to a model comprising of 13 SNPs. The results are shown in Table 3.8.

The comparison of the different multi-locus methods, when applied to the lp(a) data, could not be directly quantified since there was no known 'true model' to available to provide a basis for comparison. The lasso/BMA combination method did however, select rs3798220 to be the most significantly associated SNP with elevated lp(a) levels in the blood. This minor allele variant has already been reported to be associated with elevated plasma lp(a) levels (CHASMAN *et al.* 2008) and this association has since been confirmed (CLARKE *et al.* 2009). The final models of all the multi-locus selection methods included between 13 and 14 disease-associated loci, and a study carried out by Ober *et al*. (OBER *et al.* 2009), suggested that the genetic architecture underlying lp(a) levels was complex and involved multiple loci within the analysed region. Therefore, even though rs3798220 was not chosen by forward and backward selection, this method was able to detect multiple variants associated with lp(a) within the same region. The fact that the results obtained by these methods reflect the results confirmed by the literature is encouraging, since it is an indication that these techniques can be used to pinpoint genetic risk factors involved in complex human diseases.

The fact that forward and backward selection did not include the variable rs3798220 into the final model is an interesting observation. Once a variable is included into these models, any other correlated variables are unlikely to be included after that, since variables that enter or leave the model are forced to remain that way. Thus, if a SNP in high LD with rs3798220 was initially included in the stepwise model, it is unlikely that rs3798220 would be considered for inclusion in the model, even though it may be more significantly associated with the outcome. This shortfall of greedy algorithms can be overcome by methods such as lasso (since significant variants can enter the model later on and cause less significant variants to be dropped) and BMA (which accounts for model uncertainty and looks at competing models that identify different combinations of SNPs that account for variation in the quantitative trait). The lasso/BMA combination method appeared to produce stable models, where the coefficients of the variables selected by lasso penalized regression, approximated the coefficients of the variables included into the models after BMA was performed.

3.4 Chapter summary

While this lasso/BMA combination takes advantage of the properties of both of the selection methods of which it comprises, it is cumbersome to carry out two separate model selection methods for one analysis and, although the type 1 error rate is well under 5%, the power to detect significant associations is not as high as either method on its own and it may therefore not have sufficient power to detect low frequency variants or variants of small effect size.

Based on the evidence from these simulation studies, lasso penalized regression shows the highest power to detect complex disease risk factors, at a computational speed that is comparable to forward and backward selection methods. The disadvantage of this method, however, is choosing a shrinkage parameter that will produce the most accurate model selection. The computation cost of using cross validation to estimate the shrinkage parameter is high and, as shown in section 3.1, the type 1 error rate based on this cross-validated estimation is also too high. One solution to this problem has been provided by Wu *et al*. (Wu *et al.* 2009b), who have developed a lasso method that circumvents the estimation of a shrinkage parameter and can be applied to case/control data within a GWAS context. This method is discussed further in the following chapters and a number of its different applications are investigated.

Chapter 4: Analysis of type 2 diabetes genome-wide association data using lasso penalized regression

Type 2 diabetes (T2D) is a common metabolic disorder that usually develops later in life, after the age of 40 (TAYLOR 2006). The disease is characterized by hyperglycaemia, which generally occurs when pancreatic beta-cells can no longer meet the progressive demand for insulin (KRUSZYNSKA *et al.* 1995). In some cases, T2D results when target tissues such as fat and muscle become insulin resistant and do not respond to the circulating levels of insulin in the blood. While T2D is known to be influenced by environmental factors such as increased age, increased body weight and decreased physical activity, the underlying genetic factors that contribute to T2D risk are complex and not well understood. Identifying the underlying genetic variants contributing to T2D aetiology could help elucidate the biological mechanisms involved in the disease and pave the way to better disease management and the development of therapeutic targets.

In 2007, the WTCCC carried out the genome-wide analysis of approximately 1924 T2D cases and 2938 population controls (WTCCC, 2007) in order to identify common variants associated with the disease. Single-SNP methods of analysis were used to identify these associations and quantify effect sizes and

levels of significance. The results of the WTCCC study identified 3 variants that were associated with T2D with genome-wide significance ($p<5x10^{-8}$) and a number of other SNPs throughout the genome that were associated with the disease, but with p-values greater than $5x10^{-8}$. These SNPs mapped to candidate genes which had been implicated for their role in T2D and demonstrate the success of GWAS to detect common variants involved in complex disease aetiology.

The aim of this chapter is to carry out model selection on the WTCCC T2D data, using lasso penalized regression, in order to identify a set of SNPs that systematically explain the variation in the T2D phenotype. These results can then be compared to the results obtained by the WTCCC using single-SNP methods of analysis, to establish the effectiveness of lasso as a model selection technique in the context of GWA studies.

4.1 Implementation of lasso penalized logistic regression using Mendel software

Since multivariate regression can only accurately be applied to data sets where the number of predictor variables (SNPs) does not far exceed the number of individuals in the study, large genome-wide association studies with millions of potential predictors, generally rely on univariate methods of analysis. Lasso penalization provides a solution to this problem, by assigning coefficient values of 0 to the majority of variables, effectively eliminating them from the model and providing sparse data sets that can be readily analysed.

In this chapter, Mendel software (www.genetics.ucla.edu/software/Mendel) was used to implement lasso penalized logistic regression in order to select the combinations of variants throughout the genome that best explain the variation of the T2D phenotype. Mendel implements a modification of the lasso algorithm (previously discussed in Chapter 2) so that it can be applied within a logistic regression framework, which is necessary in this case, since the response variable is either 0 (control) or 1 (case). This is done, by replacing the sum of squares with the negative loglikelihood, such that:

$$f(\theta) = l(\theta) - \lambda ||\beta||_1$$

where $l(\theta)$ is the loglikelihood and $\lambda ||\beta||_1$ is the lasso penalty (Wu *et al.* 2009b). Since the object is now to maximize the likelihood, rather than minimizing the sum of squares as in the case of continuous response variables, the lasso penalty is subtracted from the negative loglikelihood.

An attractive feature of using Mendel to carry out lasso regression, is the process by which the tuning parameter is selected. It has been suggested that cross validation is an accurate method of assessing the tuning parameter, but while this may be feasible for small data sets, the computational time to establish a cross validated tuning parameter in data sets containing hundreds of thousands of variables, would be prohibitively high. Upon implementation of lasso to the Lp(a) data in R (described in section 2.5), it was found that for data sets containing more than 500 variables, the lasso penalty could no longer efficiently be calculated by cross validation. In these cases, a fault would occur within the LARS package in R and the analysis had to be re-done using fewer variables.

Even if the algorithm used in the LARS package could be adapted so that this error would not occur when cross-validation was applied to more than 500 variables, the time it would take to run this analysis, would be hugely impractical. In light of this, Wu *et al*. (Wu *et al.* 2009b) proposed a way to circumvent this problem, by allowing a pre-specified number of predictors to be included into the final model. This means that the value of the tuning parameter is dependent on the number of predictors that are selected to be included in the final model, rather than the other way around. Due to the limited power to detect common disease associated loci in a study of this size, it is unlikely that more than a few independent signals will be found to be associated with T2D. Thus, a lasso model incorporating 20 or 30 variables in the final model, should in theory, capture all the potential associations in the data.

After the analysis is carried out, and Mendel has chosen a model consisting of the a pre-determined number of predictors that best explain the variation in the phenotype, a number of statistics that can be used to assess the significance of these variables are output. Firstly, Mendel outputs a univariate p-value for each of the variables in the final model. This is calculated using a 1-degree of freedom test that tests the null hypothesis that $\beta=0$, where β is the coefficient of a variable in the model. If there are 30 variables in the final model, this 1-degree of freedom test is carried out for each of the 30 variables in the model, to calculate their univariate p-values. Thus, univariate p-values are not adjusted for other variables in the final lasso model, and therefore, do not give any indication of the degree to which variables in the model may be correlated with one another. It is important to note that lasso is a model selection method designed to select a predictive subset of SNPs, rather than a method used to assess the significance of SNP-

phenotype associations and therefore these univariate p-values are not part of the lasso model selection analysis, but are calculated in a subsequent step.

The second parameter that is calculated is the leave-one-out (LOO) index. In contrast to the univariate p-value, this value does take into account the other variables in the final lasso model. In order to produce LOO indices, Mendel takes only the variables in the final lasso model (30 for example) and sets the lasso penalty to 0, so that there is no constraint on the coefficient values. In other words, there is no parameter shrinkage and the coefficients values of the variables in the model will be the same as their least-squares estimates. Then, one predictor is omitted from the model and likelihood ratio tests can be conducted on the remaining set of 29 predictors. The LOO index therefore provides a way of assessing the degree to which other predictors in the model account for the same proportion of variation in the phenotype, as the variable being excluded from the model. In practical terms, if predictors in the model have some degree of correlation between them, it should manifest as high LOO indices for the correlated variables. There is no particular cut-off for LOO indices that can tell us whether variables are independent of one another or not, but in a study that applied lasso regression to both simulated and real genome-wide data, the authors used a LOO index of 0.01 to indicate low correlation between SNPs in the model (Wu *et al.* 2009b). According to Wu *et al*. (Wu *et al.* 2009b) extensive linkage disequilibrium between SNPs in the data set, can result in multiple correlated variables being included in the lasso model. In these cases, larger models can be pre-specified in order to ensure that signals of smaller effect size or less significant associations, are not masked by a set of highly correlated SNPs each tagging the same significant association.

In terms of the regression coefficients, these are also estimated after the lasso analysis has been carried out and are estimated by setting the lasso penalty to 0 (as described above for the LOO indices). These values, however, are based on the multivariate model and therefore differ from the univariate regression estimates because they account for the other variables in the model. The lasso algorithm is supposed to pick predictors most strongly associated with the response and drop any other correlated predictors from the final model (SUNG *et al.* 2009), by setting their coefficient values to 0. According to other studies (WU *et al.* 2009b), and based on the evidence seen in Chapter 3 of this investigation (as well as on the evidence from subsequent lasso analyses that are documented in chapters 5 and 6), this is not the case. In cases where extensive correlation is present in the data, lasso tends to act in a similar way to ridge regression, where correlated predictors borrow strength from one another and ultimately the coefficient value that would be assigned to any one of the SNPs, will be shared amongst the group of correlated SNPs in the model (FRIEDMAN *et al.* 2010). Therefore, their regression estimates are likely to be lower than the univariate estimates in cases such as these.

4.2 Data preparation

For the WTCCC analysis, only European descent samples were used and individuals were excluded from the data set on the basis of false identities, relatedness and non-Caucasian ancestry. Duplicates and samples with outlying heterozygosity (>30% or <23% across all SNPs) were also removed. The 2938 study control samples were made up of 1,480 individuals from the 1958 Birth

Cohort and 1,458 individuals from UK blood donors. Cases were made up of 1924 individuals diagnosed with T2D. Individuals with more than 3% missing data across all SNPs were excluded from the sample.

The WTCCC samples were genotyped across 500,568 SNPs using the Affymetrix GeneChip 500K Mapping Array Set (WTCCC, 2007). SNPs, which had a missing data rate higher than 5% (or more than 1% for SNPs with a MAF less than 5%) across all the samples, were excluded from the data set, as well as SNPs with Hardy-Weinberg exact P-value < 5.7×10^{-7}. A total of 61,461 SNPs, with a MAF less than 1%, were excluded from the analysis on the basis that there was not sufficient power to detect association with these variants. For the purpose of this investigation, PLINK was used to carry out the aforementioned exclusions, prior to the implementation of lasso.

Despite the removal of individuals exceeding 3% missing-ness across all genotyped SNPs, the genotyping rate for the whole data set was 0,9964. Since lasso requires each individual to have values for all the predictors being analysed (i.e. a genotyping rate of 1), imputation was carried out on the data. The internal impute function within Mendel was used to perform the imputation, because the relevant files were already in Mendel format in preparation for the lasso analysis.

Mendel imputes missing values by making inferences about the genotypes at missing SNPs, based on the LD between markers in the neighbouring genomic region within the data set and does not make use of an external reference panel. A posterior probability is assigned to each possible genotype at the current SNP

and then the best genotype (i.e. the genotype with the highest posterior probability) is selected for inclusion in the data set. While, as discussed in Chapter 2, this imputation method may not be the most accurate method predicting the true genotypes, for the purpose of this study, it was the most time and computationally efficient method for the purpose of this investigation.

4.3 Analysis of WTCCC T2D data using lasso penalized regression and comparison to the WTCCC single-SNP results

After the quality control measures were successfully carried out and imputation was complete, the final data set consisted of 396,787 SNPs across 1924 T2D cases and 2938 controls. For the initial lasso analysis, the number of markers included in the final model was pre-specified to be 30. Table 4.1 shows the results of this analysis.

The adjusted p-value column in Table 4.1 was calculated by multi-locus logistic regression in STATA. The entire model of 30 SNPs was regressed against the T2D phenotype and therefore the p-value of each SNP is adjusted for the remaining 29 SNPs in the model. It is interesting to compare these adjusted p-values to the univariate p-values and the LOO indices. When correlated variables are included in the model, their univariate p-values remain low, while the adjusted p-values and LOO indices are far less significant, because the correlated variables are explaining the same portion of variance in the T2D phenotype. This can be seen for the SNPs on chromosome 10 that all reside within the 114.71-114.81million base pair region and are not independent of one another. SNP rs4506565 has a univariate p-value of 9.7×10^{-13}, and an adjusted

p-value and LOO index of 0.0549 and 0.0003, respectively (Table 4.1).

Table 4.1 The results of the lasso analysis showing the top 30 predictors associated with type 2 diabetes using the data from the WTCCC.

SNP	Chr	Position (bp)	Univariate P-value	Adjusted P-value	MAF	LOO index	Regression estimate
rs4506565	10	114746031	9.70E-13	0.0549	0.3519	0.0003	-0.4625
rs7077039	10	114779067	1.88E-12	0.0340	0.4838	0.1332	-0.1953
rs8050136	16	52373776	2.94E-08	2.51E-08	0.4209	0.0019	-0.2669
rs7917983	10	114722872	5.91E-08	0.0324	0.4866	0.0575	-0.1000
rs1495377	12	69863368	8.09E-07	1.71E-06	0.4831	1.64E-05	0.1942
rs9465871	6	20825234	1.98E-06	0.0069	0.1941	0.0041	-0.2122
rs11693602	2	161050165	4.21E-06	1.44E-06	0.2136	0.0001	0.2247
rs2930291	15	72391887	6.13E-06	1.19E-05	0.3614	0.0003	0.3234
rs9326506	10	43388564	6.41E-06	5.63E-06	0.4905	0.0007	0.2859
rs2903265	15	78200439	8.87E-06	1.79E-05	0.2678	0.0001	0.3566
rs17248501	2	205864551	1.60E-05	1.49E-05	0.3753	0.0005	0.3022
rs2099106	16	9288209	1.62E-05	7.46E-06	0.2522	0.0001	-0.1973
rs1665901	6	107540093	1.84E-05	0.0001	0.3416	0.0041	0.2591
rs11688935	2	189001411	2.40E-05	7.82E-05	0.3118	0.0538	-0.1711
rs10946398	6	20769013	3.31E-05	0.2301	0.3356	0.3216	-0.0613
rs1481279*	4	104335542	3.87E-05	3.96E-05	0.3753	0.0005	0.3024
rs10748582*	10	94467199	4.03E-05	0.0003	0.2522	0.0002	0.1686
rs10817674	9	114602134	4.05E-05	0.0011	0.2723	2.14E-05	0.1946
rs440646	3	11266171	4.72E-05	6.33E-05	0.3358	4.62E-05	-0.1805
rs1340430	1	73976149	5.75E-05	0.0003	0.3118	0.0004	0.1563
rs4600815	3	134503666	6.21E-05	0.0002	0.3416	0.0001	-0.1762
rs6846031	4	178394297	6.46E-05	1.13E-05	0.1790	3.82E-05	0.1804
rs7583600	2	30783715	6.60E-05	2.55E-05	0.2206	2.78E-05	-0.2053
rs243018	2	60498358	6.84E-05	1.18E-05	0.3652	9.96E-06	-0.1988
rs10120268*	9	108627659	7.55E-05	5.12E-05	0.3389	8.83E-05	0.1860
rs2296947	9	114510078	7.61E-05	0.0058	0.3955	0.0029	-0.1413
rs897445*	8	2283162	7.62E-05	8.15E-05	0.4079	5.10E-05	-0.1820
rs11679606*	2	16717046	9.99E-05	0.0010	0.3996	0.0001	0.2247
rs4958711	5	153564556	0.0001	1.13E-05	0.4395	4.57E-05	-0.1865
rs4343209	14	98286712	0.0001	0.0001	0.3604	0.0004	0.1614

* SNPs that were included in the final lasso model that did not map to regions identified by the WTCCC to be associated with T2D.

Therefore the extent to which the adjusted p-values and LOO indices differ from the univariate p-value (that does not consider other SNPs in the model) provides us with information about the correlation between variables in the model.

The top lasso hit is the same SNP (rs4506565 mapping to *TCF7L2,* odds ratio

(OR) 1.36 (1.2-1.54)) that was found to have the strongest association with T2D in the WTCCC analysis. The lasso analysis included two further SNPs tagging the 10q25 region - rs7077039 and rs7917983. As previously mentioned, the correlation between rs7077039 and rs4506565 is evident from the high adjusted p-values and LOO indices. In fact, rs4506565 and rs7077039 have an r^2 and D' of 1 between them. However, rs7917983 does not seem to be in such high LD with these two SNPs (r^2=0.152, D'=0.491 and r^2=0.294, D'=0.561 with 7077039 and rs4506565 respectively), even though the LOO index and adjusted p-value are far less significant than the univariate p-value, suggesting that they are, in fact, highly correlated. The third lasso hit was SNP rs8050136, which lies within the 16q12 region. This is the *FTO* locus was identified in the WTCCC study, which was tagged by SNP rs9939609 (OR = 1.34 (1.17-1.52)). In this case, the adjusted p-value and the univariate p-value are similar, suggesting that there are no variables correlated with rs8050136 in this model. In agreement with this, the LOO index for this SNP is also low (0.0019, Table 4.1).

In the case of the signal on chromosome 6p22 (tagged by SNP rs9465871, OR = 1.18 (1.04-1.34)) that was identified in the WTCCC study to be significantly associated with T2D, lasso included SNP rs10946398, which is poorly correlated with the WTCCC SNP in terms of r^2 between the SNPs (0.264) but also maps to the *CDKAL1* gene, in the final model. The D' between these SNPs, however, is 0.890, which indicates that the two SNPs could be on the same haplotype and indeed tagging the same gene. One of the reasons that lasso failed to identify the WTCCC SNP could be that this SNP shows a strong departure from additivity. In fact, in the WTCCC study, it was established that the best fitting model at this locus is a 2-degree of freedom model that does not assume

additivity. Since lasso assumes additivity at all SNPs being analysed, this could have resulted in a lower power of lasso to detect the WTCCC SNP.

The SNP rs10946398 that lasso did identify in this region, however, had univariate p-value of 0.183×10^{-4} (Table 4.1), and therefore, while lasso failed to identify the WTCCC SNP, it may have identified a secondary signal within the same chromosomal region. This highlights a potential advantage of lasso over single-SNP methods of analysis, in that lasso is able to consider more than one SNP at a time, and therefore has the ability to detect putative independent signals within a single locus.

In the WTCCC, SNPs that failed to meet genome-wide significance levels, but still showed moderate or strong associations with T2D, were also pinpointed and tabulated. These results are important, since these signals, although not genome-wide significant, could represent important disease loci, that due to the inadequate sample size (and therefore power) failed to pass the genome-wide significance threshold.

The SNPs showing moderate-strong association ($5 \times 10^{-7} < p < 1 \times 10^{-5}$) with T2D, identified by the WTCCC single-SNP analysis is shown in table 4.2. In cases where lasso penalization selected the same SNP, or a SNP in a nearby region, the SNP is indicated in the “Lasso” column. If the SNP selected by lasso regression (lasso SNP) was the same SNP that was indicated in the WTCCC study (WTCCC SNP), then no r^2 or D’ value was entered in the 6^{th} column. For instances where lasso selected a different SNP, in the same or nearby region to the WTCCC SNP, the LD between these two SNPs (measured by r^2 and D’) is

indicated in the final column. For most of these SNP-pairs, the r^2 between them is greater than 0.8, indicating that both variants tag the same signal.

Table 4.2 Comparison of SNPs included in the final lasso model with those SNPs shown in the WTCCC, to have a moderate – strong association ($5 \times 10^{-7} < p < 1 \times 10^{-5}$) with type 2 diabetes. Position for the WTCCC SNPs are measured in million base pairs (Mil bp).

Strong - moderate association					
Chr	WTCCC		Lasso		LD between SNPs (r^2/D')
	Position (Mil bp)	SNP	Position (bp)	SNP	
1	66.04-66.36	rs4655595			
2	160.90-161.17	rs6718526	161050165	rs11693602	0.832/1
2	205.87	rs7587983	205864551	rs17248501	1/1
3	55.24-55.32	rs358806			
4	122.92-123.02	rs7659604			
5	65.87	rs4583845			
6	20.63-20.84	rs9465871	20769013	rs9465871	Match
10	43.43-43.63	rs9326506	43388564	rs9326506	Match
10	81.90-81.91	rs2789686			
10	114.71-114.81	rs4506565	114746031	rs4506565	Match
12	49.50-49.87	rs12304921			
12	69.58-69.96	rs1495377	69863368	rs1495377	Match
15	72.24-72.50	rs2930291	72391887	rs2930291	Match
15	78.12-78.36	rs2903265	78200439	rs2903265	Match
16	9.29	rs2099106	9288209	rs2099106	Match
16	52.36-52.41	rs7193144	52373776	rs8050136	0.967/1
16	"	rs9939609	52373776	rs8050136	1/1

Table 4.2 shows that lasso penalized regression successfully identified a number of regions that were found to be strongly associated with T2D in the WTCCC. In all cases where lasso selected a different SNP in the same genomic region as the WTCCC SNP, the correlation between these SNPs was greater than 0.8. The lasso analysis also identified SNPs that were moderately associated ($5 \times 10^{-7} < p < 1 \times 10^{-5}$) with T2D in the WTCCC analysis. These results are seen in Table 4.3, which demonstrates that lasso successfully identified a number of moderate association signals established in the WTCCC study.

There are, however, a few exceptions. Table 4.3 shows that on chromosome 2, the lasso SNP nearest the association signal in the region 60.46 (tagged by

rs9309324) is rs243018.

Table 4.3 Comparison of SNPs included in the final lasso model with those SNPs shown in the WTCCC, to have a moderate association ($5 \times 10^{-7} < p < 1 \times 10^{-5}$) with type 2 diabetes. Position for the WTCCC SNPs are measured in million base pairs (Mil bp).

$1 \times 10^{-7} <$ p-value $< 1 \times 10^{-5}$					
	WTCCC		Lasso		LD between SNPs (r^2 /D')
Chr	Position (Mil bp)	SNP	Position (Mil bp)	SNP	
1	48.93	rs12086219			
1	73.98	rs1340430	73976149	rs1340430	Match
1	205.98	rs6691406			
1	219.68	rs1341987			
2	30.78	rs7583600	30783715	rs7583600	Match
2	60.46	rs9309324	60498358	rs243018	0.407/0.649
2	160.97	rs1020731			
2	189	rs11688935	189001411	rs11688935	Match
2	205.86	rs17248501	205864551	rs17248501	Match
3	11.27	rs440646	11266171	rs440646	Match
3	134.46	rs769097	134503666	rs4600815	0.961/1
3	150.03	rs16861027			
3	154.52	rs10513440			
4	17.12	rs1852749			
4	123.02	rs6815973			
4	161.74				
4	178.39	rs6846031	178394297	rs6846031	Match
5	72.74	rs4292434			
5	122.49	rs6872465			
5	153.56	rs4958711	153564556	rs4958711	Match
6	2.4	rs9391949			
6	55.33	rs7452656			
6	107.54	rs1665901	107540093	rs1665901	Match
8	15.75	rs2736010			
8	98.43	rs2679765			
9	88.03	rs7019589			
9	114.58	rs2185935	114510078	rs2296947	0.08/0.426
9	135.59	rs2590504			
10	7.78	rs7474871			
10	53.47	rs11000542			
10	104.07	rs17780667			
10	130.48	rs10829494			
11	94.53	rs11021059			
11	119.32	rs657317			
12	12.55	rs16908188			
12	18.47	rs12581163			
12	49.61	rs17125088			
12	69.69	rs11178531			
13	71.99	rs4053550			
14	83.09	rs1007383			
14	98.08	rs8012854			
14	98.29	rs4343209	98286712	rs4343209	Match
16	9.29	rs2099106	9288209	rs2099106	Match
18	62.34	rs508987			
18	75.56	rs70198			
20	40.25	rs7262414			

22	45.01	rs739164		

The r^2 and D' between these two SNPs is 0.407 and 0.649 respectively, suggesting that lasso could be identifying a secondary signal. While these SNPs are close in proximity to one another, there is a lot of recombination in the region and hence the low LD across variants. Both of these variants, although uncorrelated, lie downstream of the *BCL11A* gene and are likely to map to the same gene. This gene has now been established as a T2D risk locus ($p=2.9x10^{-15}$;(VOIGHT *et al.* 2010))

On chromosome 9, the WTCCC SNP and the lasso SNP have an r^2 and D' of 0.08 and 0.426 between them respectively, despite being located within 70 kb of one another (Table 4.3). While the WTCCC SNP is likely to be tagging the *TNFSF15* gene, the lasso SNP is situated nearest *C9orf191*. These two SNPs are separated by a region with a high recombination rate and it would appear that the two analyses have identified independent association signals.

Overall, there were another 5 SNPs that were selected by the lasso analysis that were not identified by the WTCCC study to be moderately associated with T2D ($5 \times 10^{-7}<p<1x10^{-5}$). Firstly, rs1481279 (chromosome 4), appears to map to the *NHEDC2* gene, encodes sodium hydrogen antiporters that control the electroneutral exchange of protons for Na(+) and Li(+) across the inner mitochondrial membrane. This SNP was put forward for replication in DIAGRAM (Diabetes Genetics Replication And Meta-analysis consortium), but failed to replicate (VOIGHT *et al.* 2010). A second unique lasso SNP was rs10748582 on chromosome 10, which lies near the *KIF11* and *HHEX* genes. *HHEX* encodes a member of the homeobox family of transcription factors, many of which are

involved in developmental processes, and this locus has now also been established as a T2D risk locus (VOIGHT *et al.* 2010). *KIF11* encodes a motor protein that has been implicated in human diseases such as Alzheimer's (*KIF5*), polycystic kidney disease (*KIF3A / KIF3B*) and diabetes (*KIF5B*). Lasso also uniquely identified SNP rs10120268 (on chromosome 9) near the *ACTL7A* and *ACTL7B* genes, both involved in a wide variety of cellular processes. These genes are expressed in a numerous adult tissues, but their exact function remains unknown. Finally, lasso also identified rs897445 on chromosome 8, which does not map near any genes and rs11679606 on chromosome 2, which appears to tag the *FAM49A* gene of unknown function.

It is unclear whether the *C9orf191* locus, that was not established or confirmed by the large-scale meta-analysis carried out by Voight *et al*. (VOIGHT *et al.* 2010) have real biological associations to T2D. This could represent false positive association identified by lasso, or they may be risk loci with small effects that even large meta-analysis efforts are underpowered to detect.

4.4 Conclusion

The main goal of this chapter was to explore the performance of LASSO in the context of genome-wide association analysis. In order to do this, the SNPs included in the lasso model including 30 predictors were compared to the type 2 diabetes-associated SNPs identified by the WTCCC single-marker analysis.

Lasso successfully identified 2 of the 3 WTCCC association signals that reached genome-wide significance, as well as including many of the WTCCC SNPS with

a moderate-strong association to T2D, in the final model. In cases where lasso did not select the same SNP as the WTCCC analysis, it generally selected a SNP that was highly correlated (r^2 >0.8) with the WTCCC SNP. This demonstrates the capability of lasso to detect the underlying genes that might be influencing the disease phenotype.

A very important outcome of this study, was that lasso included two SNPs in the final model, that were not pinpointed by the WTCCC as having more than a moderate association with T2D. In a recent, highly powered meta-analysis (Voight) the loci tagged by these SNPs (*NHEDC2* and *HHEX*) were found to be associated with T2D with genome-wide significance. This demonstrates the importance of using model selection methods that do not select variables exclusively according to their p-values, but rather look for combinations of variables that best explain the observed variation in the phenotype at hand. If p-value cut-offs are used exclusively, important risk loci could be missed.

Lasso did, however, fail to select one of the WTCCC SNPs significantly associated with T2D. Instead, a less significantly associated SNP, mapping to the same gene, was included in the final lasso model. It is proposed that the departure from additivity at this locus is a likely cause for this anomaly and highlights the fact that the relative advantage of multi-marker analyses over single-marker methods, appears to depend heavily on the underlying disease model.

Lasso penalized regression is a model selection technique that can be easily applied to the genome-wide association analysis of case/control data. While

there is some possibility of lasso penalized regression identifying false positive associations, the addition of prior information about the regions of interest and information about candidate genes, may help to distinguish true underlying associations from spurious ones. Indeed, the ability of lasso to cope with large undetermined data sets will become more and more desirable, as sequencing studies that are already underway, produce data sets consisting of increasingly large numbers of common variants, as well as newly discovered rare variants.

Another advantage of lasso is that it tends to select independent variables to include in the final model. The results in this chapter demonstrated the potential of lasso to identify independent signals within the same locus, where variants that map to the same gene could be contributing independently to the disease phenotype. Therefore, lasso can be used within a fine-mapping framework, to search for secondary signals in a well-established disease locus, which could have been overlooked by single-SNP methods of analysis. In the next chapter, we use lasso penalized regression to identify novel secondary signals within a known childhood asthma risk locus.

Chapter 5: Fine mapping of the childhood asthma susceptibility locus on chromosome 17 using lasso penalized regression

Asthma is an inflammatory disease of the small airways of the lungs, characterised by intermittent airway narrowing, as well as symptoms of wheeze and shortness of breath (COOKSON 1999). The prevalence of asthma in westernised societies is a growing health problem, affecting one in every eleven children in the UK, and it is estimated that the NHS spends £1 billion a year treating and caring for people with asthma (www.asthma.org.uk). Previous GWAS have indicated that multiple SNPs on chromosome 17q21 are significantly associated with the onset of childhood asthma and, furthermore, that these SNPs were strongly associated with transcript levels of *ORMDL3* (BOUZIGON *et al.* 2008; MOFFATT *et al.* 2007). While these findings have been replicated and confirmed in follow up studies in different populations, they account for only a small amount of the genetic variation of childhood asthma and a large portion of the heritability is yet to be explained.

To date, it remains to be confirmed whether there are multiple risk variants in the chromosome 17q21 region that work either independently, or together with, *ORMDL3* in the aetiology of childhood asthma. This region harbours a number of genes whose functions are not well understood and it is not clear which, if any, of these genes are childhood asthma risk factors (MICHEL *et al.* 2010). While it is known that *ORMDL3* is a member of a family of genes responsible for encoding transmembrane proteins of the endoplasmic reticulum, the exact mechanism by

which *ORMDL3* confers childhood asthma susceptibility remains unclear (GALANTER *et al.* 2008). Interestingly, the variant most significantly associated with *ORMDL3* expression (rs7216389), is located within the first intron of the neighbouring *GSDML* (gasdermin-like) gene (TAVENDALE *et al.* 2008). Recently, this SNP was found to be associated with the expression of both *ORMDL3* and *GSDML (COOKSON et al. 2009)*, suggesting that *GSDML* SNPs may either modify *ORMDL3* expression or contribute directly to asthma susceptibility (WU *et al.* 2009a).

Since GWAS are designed to detect common variants with moderate to high effect sizes (WTCCC, 2007), these studies may not be sufficiently powered to detect secondary signals of smaller effect size, especially if these signals are masked by a strong primary association signal in the same region. In this case, if additional susceptibility variants that contribute independently to disease risk remain undetected, the amount of genetic variance attributed to this region will be underestimated if only the primary signal is identified (VOIGHT *et al.* 2010). Thus, fine mapping methods are used as a more powerful way of localising all the variants within a previously identified region of association, that are significantly and independently associated with a phenotype of interest, with the ultimate goal of identifying the true functional variant(s) (HARDY and SINGLETON 2009). This study therefore focuses on the fine mapping of the 17q21 region, using lasso regression, to search for secondary signals that may account for a portion of the unexplained variation contributing to childhood asthma.

5.1 Initial fine mapping analysis

Data was combined from 17 cohorts within GABRIEL, a multi-disciplinary consortium aimed at identifying both the genetic and environmental causes of asthma in the European community and used for the analysis. The data was divided into a discovery set, on which the lasso analysis aimed to identify any potential association signals, and a replication set, that was used to confirm any significant, independent associations. The division of data was dictated by the individual privacy clauses from each of the cohorts. Cohorts that were not permitted to share their phenotypic data for the purpose of this fine mapping study, were included in the replication set. Table 5.1 demonstrates the way in which the data was divided.

Table 5.1 Table showing the number of individuals in each of the GABRIEL cohorts that were used for the discovery and replication sets in the initial fine mapping analysis.

Discovery Set		
Cohort	**Case**	**Control**
BAMSE	239	246
BUSSEL	188	390
CNS	88	182
FIN	33	36
KAB	841	851
PIAMA	172	187
POKOV	112	116
SAPAL	237	356
SLSJ	373	390
TOM	197	91
UFA	269	209
UK/GER	861	1034
Total	**3610**	**4088**
Replication Set		
ALSPAC	607	609
EGEA	482	598
WJST	279	620
GSK	462	1576
B58C	213	200
Total	**2043**	**3603**

For fine mapping studies, it is important that the maximum amount of information

is extracted from the region being analysed, in order to increase the probability of detecting all putative genetic risk factors. Imputation is one way of doing this, since it provides a much denser set of markers that can be analysed within an associated region and increases the probability of directly identifying a causal SNP (MARCHINI and HOWIE 2010).

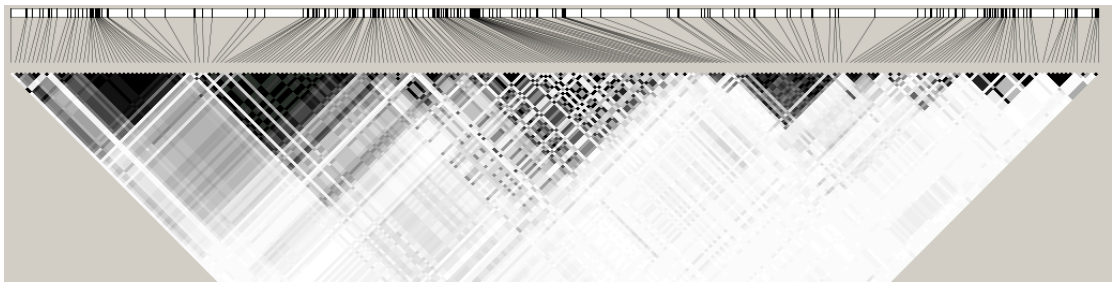
To maximise the power of our fine mapping analysis across this region, the data was imputed using MACH 1.0 software. In the first phase of the GABRIEL project, all the child samples from the UK and German cohorts were genotyped with the Illumina Sentrix HumanHap300 BeadChip (MOFFATT *et al.* 2007). Samples obtained subsequent to this first analysis, were genotyped using the Illumina Human610 quad array (www.illumina.com) and therefore different cohorts were genotyped across different sets of markers. MACH was therefore used to impute those SNPs that were present in the Illumina 610K array, but not in the Hap300 array used in the first study (MOFFATT *et al.* 2010). In this way, samples that had been genotyped using different genotyping platforms, could be combined and analysed together in one data set. Imputation was used to provide a “best guess” genotype that could be used as the true genotype for this analysis and imputation was carried out internally and no external reference panel was utilised. Therefore, only SNPs from the 610K panel were analysed in this study.

In terms of imputation quality control, only SNPs that were imputed with an R-squared value greater than 0.3 were considered for analysis. This R-squared value is an estimate of the squared correlation between the true and imputed genotypes and by implementing the cut-off of 0.3, most of the poorly imputed SNPs and less than 1% of well-imputed SNPs would be excluded from the final

data set (http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html). Furthermore, both genotyped and imputed SNPs with minor allele frequencies less than 1% were excluded from the analysis.

A single-SNP test of association, assuming additive effects across all loci, for childhood asthma was performed in Mendel across a 3 Megabase (Mb) region encompassing the childhood asthma susceptibility locus and 57 SNPs were found to be significantly associated with childhood asthma ($p<5x10^{-8}$). Due to the dense set of imputed markers being analysed, the majority of these significantly associated SNPs are likely to be in high LD with one another and this high degree of correlation between markers, is what makes the dissection of independent signals in the region so difficult. Figure 5.1 shows the high degree of LD that exists between the markers within this region before imputation was carried out.

A



B

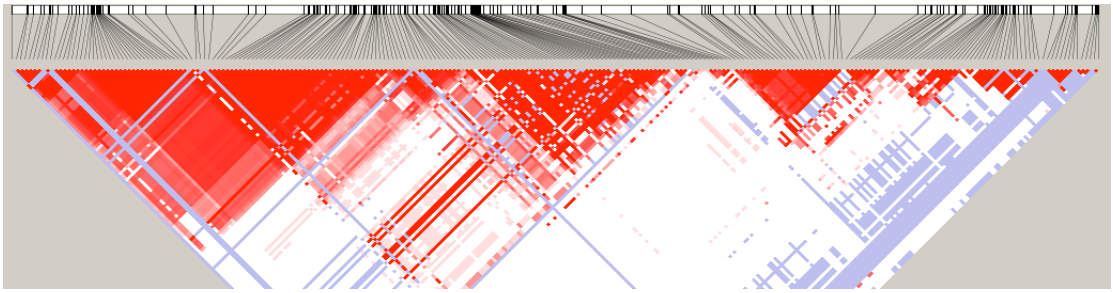


Figure 5.1 LD plot of the chromosome 17 region from 35milbp – 36milbp, generated by Haploview, using data from HapMap samples. Figure A shows the plot of the r2 values, while figure B plots the D' values. The high degree of correlation between the markers in plots A and B, is indicated by the blocks shaded in black ($r^2>0.8$) and red (D' > 0.8), respectively.

The extensive LD between the imputed SNPs suggests that the majority of the SNPs that reached genome-wide significance in the single-SNP analysis, will be tagging the same signal. The aim of this study was therefore to use lasso penalized regression to analyse the region on chromosome 17 (SNPs within the physical base pair position 35000000 – 38000000), in order to identify genetic variants that were significantly and independently associated with the onset of childhood asthma.

Lasso regression analysis was carried out on the discovery set, which consisted of 3610 cases and 4088 controls, and was set to include 20 predictors in the final model. This analysis pinpointed 6 SNPs that were associated with childhood asthma with genome-wide significance (P-value < $5x10^{-8}$) (Table 5.2).

Table 5.2 Results of the lasso analysis indicating the SNPs associated with childhood onset asthma with genome-wide significance (p-value < 5×10^{-8}).

Top Lasso Predictors						
Predictor name	Position (BP)	Univariate P-value	LOO index	MAF	Regression estimate	Single SNP reg. est.
rs4794820	35342870	6.83E-16	0.79872	0.3975	-0.02363	-0.27022
rs8069176	35310723	1.07E-15	0.10587	0.4141	-0.15871	0.25976
rs7219923	35328044	3.94E-15	0.97656	0.4706	0.00576	0.25268
rs9303280	35327557	3.94E-15	0.80595	0.4606	0.04615	0.25349
rs7212938	35376206	2.65E-14	0.27143	0.4705	0.0652	-0.24868
rs3859192	35382174	2.27E-12	0.05073	0.4754	-0.09874	-0.22807

Theoretically speaking, lasso penalization is intended to incorporate variants that are independent of one another into the final model (ZHAO and YU 2006), and thus provides a way of selecting independent variants. However, when performing this method on highly correlated, imputed data sets, as can be seen in this study, the final lasso model does include SNPs that show high levels of correlation between one another. Evidence of this can be seen in the values of

the LOO index (Table 5.2). These indices are all above 0.01 and, as explained in the previous chapter, indicate that there are other variables in the model that are explaining the same variation in the phenotype. In other words, some of these SNPs are not entirely independent of one another in their association with the phenotype.

Therefore, it was necessary to examine the extent of LD between these 6 SNPs, as the first step towards identifying possible independent, secondary signals. Two programmes were used to do this; Haploview and SNAP. Firstly, Haploview allows the user to upload a set of data, from which several pairwise measures of LD are calculated and an LD plot pertaining to the uploaded data is generated (Barrett et al., 2004). Figure 5.2 is an LD plot generated by Haploview, which graphically explains the extent of linkage disequilibrium between the 6 SNPs laid out in Table 5.2.

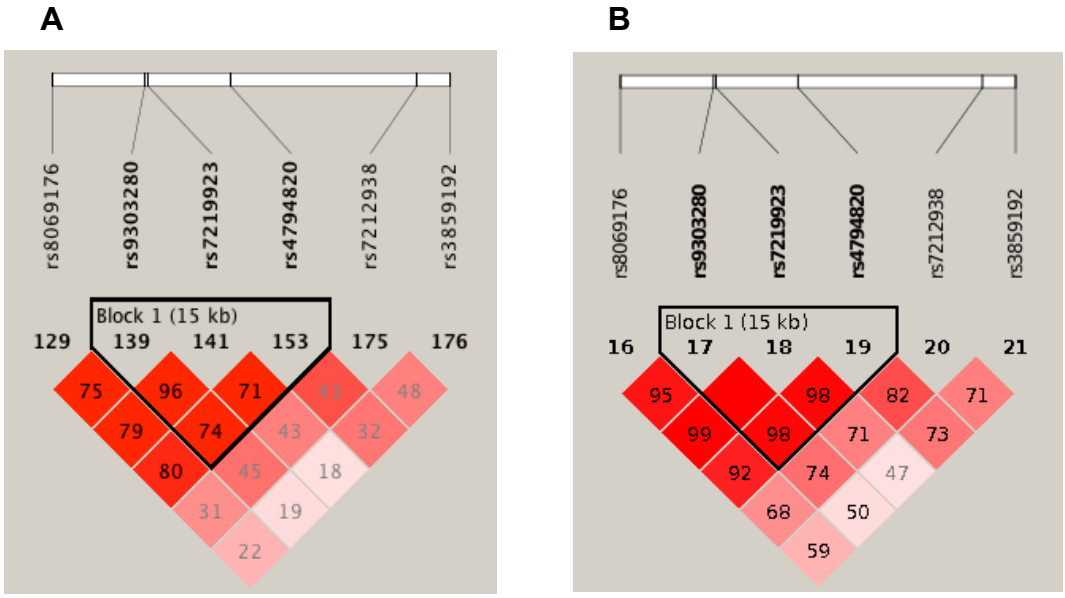


Figure 5.2 Plot showing the amount of linkage disequilibrium that exists between the top 6 SNPs identified by lasso penalized regression. Plot A and plot B show the r^2 and D' values between the markers, respectively. The darker shade of red indicates higher values of r^2 and D' and the actual values are indicated in the blocks as a percentage.

The three SNPs that form the haplotype block in Figure 5.2, rs4794820, rs7219923 and rs9303280, have high LOO indices of 0.79872, 0.97656 and 0.80595 respectively (Table 5.2). This demonstrates how LOO indices can be used as an indication of the extent of correlation that exists between the SNPs in the model. SNP rs8069176 is in high LD with these three SNPs (r^2 = 0.95), and the LOO index for this SNP is 0.10587 (Table 5.2), suggesting that this variant is likely to account for the same portion of the phenotypic variation.

In contrast to Haploview, SNAP generates LD plots based on genotype data from the HapMap Project (http://www.broadinstitute.org/mpg/snap). While the information provided by this programme does not pertain to the data set being analysed, it does show the rate of recombination (thereby identifying recombination hot spots across the genome) in the location of the variants under investigation and provides further insight into the relationship between the variants of interest.

Figure 5.3 is an LD plot, constructed using SNAP, based on data from European samples (CEU) from HapMap and shows the genomic position of rs8069176 (green vertical line), rs7212938 (purple vertical line) and rs3895192 (red vertical line). All the SNPs in the region that are highly correlated (r^2 > 0.8) with rs8069176, are shown between the dotted vertical lines. This region includes the SNPs rs4794820, rs7219923 and rs9303280 that are all highly correlated with one another and have pair-wise LD estimates that exceed 0.9, as seen in Figure 5.2.

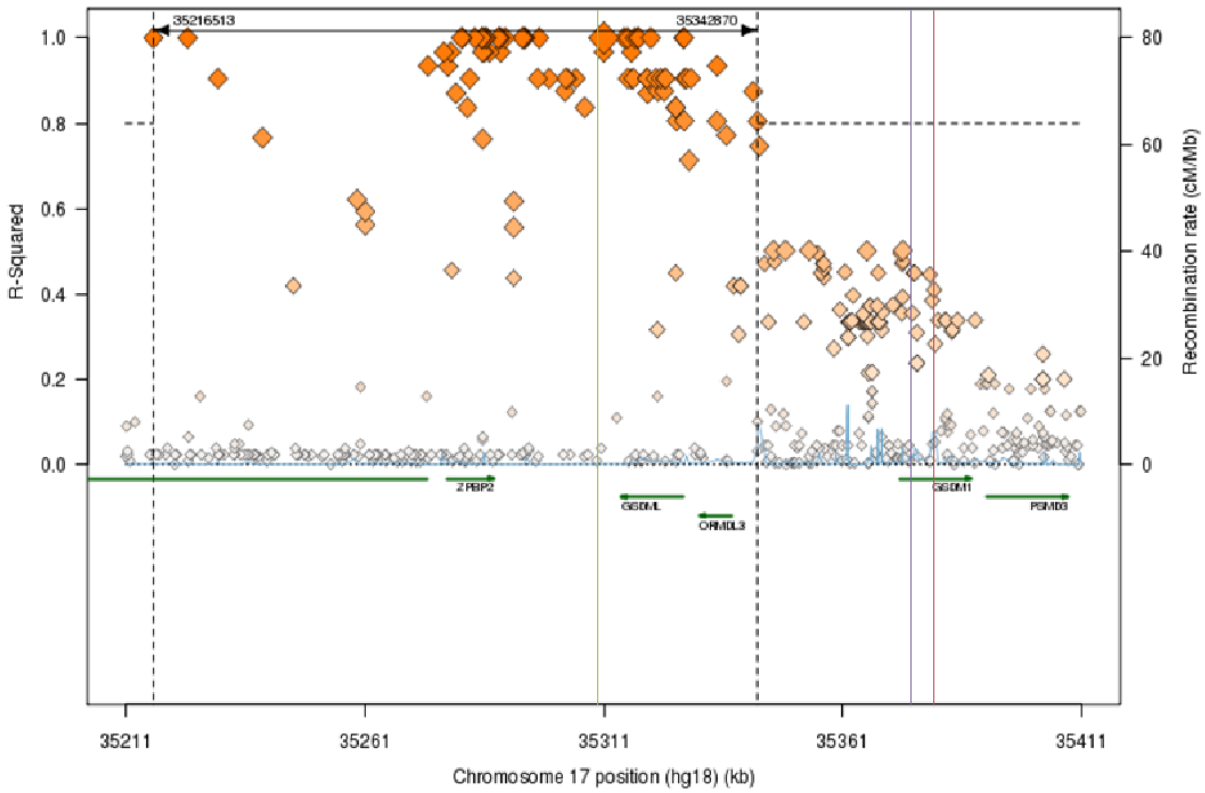


Figure 5.3 Plot generated in SNAP showing the genomic position of rs8069176 (green vertical line), rs7212938 (purple vertical line) and rs3895192 (red vertical line). The SNPs that have an r2 value of more than 0.8 with the primary signal rs8069176, are shown between the dotted vertical lines. The recombination rate is indicated by the blue line and triangles represent the SNPs at their relative r2 value (y axis) and physical position (x axis) on chromosome 17.

The results of the Haploview analysis showed that three of the 6 top SNPs in the lasso analysis, rs8069176, rs7212938 and rs3895192, showed pair-wise LD estimates less than 0.8 between them. Further examination of these polymorphisms in SNAP suggested that there were three potential independent signals, which, while residing in close proximity to one another within this region on chromosome 17, were separated by regions where the recombination rates were moderately high (so called “warm” recombination spots) (seen by the light blue line in Figure 5.3). Between rs8069176 and the other two SNPs, there are three peaks in the recombination rates, the highest of which is 10cM/Mb. Between rs7212938 and rs3895192, the recombination rates are slightly

elevated, although do not exceed 5cM/Mb.

Based on both the pairwise measures of LD between the 6 top SNPs identified in the lasso analysis and the elevated rates of recombination separating these SNPs into three distinct signals, further model selection was carried out using STATA to analyse the independence of these signals.

A manual method of model selection was carried out in STATA, which entailed systematically comparing all the possible models incorporating any one of the top lasso SNPs. Logistic regression models were fitted and each of the models were compared to identify which of them provided the best possible fit to the data at hand. The model likelihood values, the AIC, BIC and the R-squared values were used to assess the fit of each of the models.

Practically speaking, if two SNPs were both explaining identical, or even similar, portions of the variation of the phenotype (i.e. tagging the same signal), a model containing both of these SNPs would be shown by these three model selection criteria, to fit the data less well than a model containing just one of the SNPs. Therefore, only SNPs that are independently associated with the phenotype will improve the model fit without introducing too much variance. The AIC and BIC both assess model fit by rewarding models that fit the data the best, while simultaneously penalizing them for each extra variable included in the model (GHOSH *et al.* 2006). Similarly, the higher the value of the negative log-likelihood of the models [*ll(model)*], the more likely the data are to have occurred under the model. The *R-squared* value increases as the amount of phenotypic variation explained by the SNPs in the model increases and will almost certainly increase

upon the addition of new variables into the model and therefore measures such as AIC and BIC, that account for the bias/variance trade-off, were focussed on in this model comparison.

The results of this analysis in STATA can be seen below in Table 5.3

Table 5.3 Results of the model selection carried out in STATA.

Model SNPs	ll(null)	ll(model)	df	AIC	BIC	R-squared
1 rs8069176	-5319.731	-5287.581	2	10579.16	10593.06	0.0060
2 rs8069176; rs7212938	**-5319.731**	**-5280.412**	**3**	**10566.82**	**10587.67**	**0.0074**
3 rs8069176; rs7212938; rs3859192	-5319.731	-5278.938	4	10565.88	10593.67	0.0077
4 rs4794820; rs7212938	-5319.731	-5282.362	3	10570.72	10591.57	0.0070
5 rs9303280; rs7212938	-5319.731	-5284.02	3	10574.04	10594.89	0.0067
6 rs7219923; rs7212938	-5319.731	-5283.621	3	10573.24	10594.09	0.0068
7 rs8069176; rs3859192	-5319.731	-5281.006	3	10568.01	10588.86	0.0073

Model 2 and model 3 are the two best models in terms of model fit versus the variance introduced by the inclusion of extra variables, but the AIC and BIC differ in terms of the model which they choose to be the best fit to the data. Therefore a likelihood ratio test was carried out to further test whether the addition of rs3859192 significantly improved the fit of model 2 or not.

The likelihood ratio (lr) test is commonly used to evaluate the difference between nested models. A model can be nested in another, if the first model can be generated by imposing restrictions on the parameters of the second. In this case, the restriction involves constraining the parameter estimate to zero - equivalent to omitting the predictor variables from the model. Thus it tests whether constraining the extra parameters in the second model to zero will significantly reduce the fit of the model. The results of the likelihood ratio test indicated that the addition of rs3859192 did not significantly improve the fit of the model since

the test for the assumption that model 2 was nested in model 3 had a non-significant p-value of 0.0859. In other words, the hypothesis that the coefficient of rs3859192 is 0 was not rejected.

The BIC and likelihood ratio test suggest that a model incorporating two SNPs, rs8069176 and rs7212938, maximises the fit of the model with the fewest number of parameters and this supports the notion that these two variants could potentially be contributing independently to childhood asthma susceptibility. Of these SNPs, rs8069176 has the highest association with the phenotype ($p=0.10683x10^{-14}$, Table 5.2), and is highly correlated (r2 = 0.99) with the variant rs7216389, previously confirmed by Moffatt *et al*. (MOFFATT *et al.* 2007) to regulate the expression of the *ORMDL3* gene. Both SNPs lie in the first intron of the neighbouring *GSDML* gene. Models 4,5 and 6 were included to demonstrate that out of the top four lasso SNPs that were all in high LD with each other ($r^2>0.92$), the combination of rs8069176 and the putative secondary signal provided the best overall model fit.

The next step was to assess the fit of the most significant model from the discovery data to the replication set consisting of 2043 childhood asthma cases and 3603 controls (Table 5.1). STATA was again used to fit logistic regression models, containing rs8069176 and rs7212938, to each of the studies in the replication set. For each study, the resulting coefficient and standard error values for the SNPs in the model were scaled as a log odds ratio and 95% confidence interval and meta-analysed in STATA using the *metan* command. STATA was used to carry out a fixed-effects meta-analysis, whereby the true effect size (θ) is assumed to be normally distributed with $N(\theta; \sigma_\theta)$, where σ_θ is the within study

variance of the effect size. Since fixed effect meta-analysis assumes that all studies have the same underlying effect size (θ), it does not allow for between-study variance. Thus if heterogeneity does exist across studies, estimates of this effect may be biased. It is important to note here, that while each SNP was meta-analysed individually, the coefficient and standard error values that were used in the analysis, were parameters taken from the model where the combined effect of both SNPs on childhood onset asthma were being modelled. The pooled effect sizes of each SNP, therefore take the effects of the other SNP in the model, into account.

STATA converts the input coefficient and standard error values from each study and outputs the odds ratio and their respective 95% confidence intervals. The output shows, for each study, the effect size and the corresponding 95% confidence interval, as well as the percentage weight contributed to the overall meta-analysis by each cohort and outputs a summary statistic for the overall odds ratio (*pooled effect*) and the 95% confidence intervals for this effect.

STATA also produces an I^2 statistic to quantify the amount of heterogeneity in the studies. This statistic measures the percentage of between-study heterogeneity that is attributable to variability in the true effect, rather than to sampling variation (HIGGINS *et al.* 2003). Values of I^2 lie between 0% and 100% and Higgins *et al*. (HIGGINS *et al.* 2003) suggest that low, moderate and high levels of heterogeneity are indicated by I^2 values between 25%-50%, 50%-75% and greater than 75% respectively.

The results of the meta-analyses for both SNPs in the best-fit model are shown in Table 5.4 and Table 5.5.

Table 5.4 Meta-analysis of rs8069176 carried out in STATA, using data from the replication cohorts.

Meta-analysis: rs8069176				
Study	Odds ratio	95% Confidence Interval		% Weight
ALSPAC	0.803	0.656	0.984	24.56
B58C	0.936	0.654	1.339	7.89
GSK	0.736	0.611	0.887	29.14
EGEA	0.772	0.624	0.957	22.07
WJST	0.903	0.704	1.158	16.34
Pooled Effect	0.801	0.724	0.886	100

I^2 = 0.0%
Test of ES=1: p = 0.0005

Table 5.5 Meta-analysis of rs7212938 carried out in STATA, using data from the replication cohorts.

Meta-analysis: rs7212938				
Study	Odds ratio	95% Confidence Interval		% Weight
ALSPAC	1.090	0.879	1.352	23.68
B58C	1.312	0.899	1.913	7.74
GSK	0.939	0.777	1.134	30.76
EGEA	1.057	0.846	1.321	22.19
WJST	1.271	0.975	1.658	15.62
Pooled Effect	1.074	0.967	1.193	100

I^2 = 13.3%
Test of ES=1: p = 0.180

The results of the STATA meta-analysis for SNP rs8069176, show a significant pooled effect (p=0.0005) (Table 5.4). The 95% confidence intervals do not include a value of 1, indicative of no effect, and the effects of this SNP throughout the cohorts show directional consistency. Furthermore, the I^2 value for this SNP is 0.0%, indicating that there is no influence of heterogeneity amongst the different studies on the odds ratio estimate.

Table 5.5 shows the results of the meta-analysis carried out on SNP rs7212938. These results indicate a much smaller, positive pooled effect of 1.074. This effect

estimate is not significant, however, since a pooled effect of 1 is included in the 95% confidence intervals and the p-value of 0.180 suggests that the odds ratio is insignificantly different from 1. While the GSK cohort shows an odds ratio that is less than 1, the odds ratios in the other four studies are directionally consistent. The I^2 test of heterogeneity indicates that 13.3% of the variation in the effect size is attributable to heterogeneity. Even though this is considered to be a low value of the statistic, both a fixed and random meta-analysis were carried out and compared to investigate the possible presence heterogeneity in the analysis.

Both fixed and random effects analyses can be performed using the same command. Under the fixed effect model, it is assumed, a priori, that one true effect size exists and is shared by all the studies (FIELD 2001). In contrast, random effects model allow for the possibility that the true effect may vary between studies (SCHMIDT *et al.* 2009). Therefore, the studies included in the meta-analysis are assumed to be a random sample of the relevant distribution of effects, and the combined effect estimates the mean effect in this distribution (HEDGES and VEVEA 1998).

STATA was used to carry out fixed- and random-effects meta-analysis of rs7212938, and the effect size was re-estimated under both of these models.

Table 5.6 Comparison of fixed and random effects meta-analysis of SNP rs7212938.

Model	Odds ratio	Lower	Upper	p-value
Random	1.080	0.964	1.211	0.186
Fixed	1.074	0.967	1.193	0.180
Test for heterogeneity: Q= 4.613 on 4 degrees of freedom (p= 0.329)				
Der Simonian and Laird estimate of between studies variance = 0.002				

The results indicate that the modest heterogeneity between samples does not significantly affect the overall results of the analysis. While the odds ratio estimate is slightly higher in the random effects model, the 95% confidence intervals are also wider and the p-value remains insignificant at 0.186. Therefore, we chose to continue only with fixed effects models.

Overall, the results of the meta-analyses showed that the size and direction of the coefficient values were consistent across the replication sets. While the p-value for the primary signal showed significance, the secondary signal was not replicated with any significance. It is possible, however, given the size of the replication set, that there simply was not enough power to replicate this secondary signal of smaller effect size.

5.2 Fine mapping analysis using equal size discovery and replication sets

The meta-analysis laid out in section 5.1 failed to replicate the putative secondary signal identified by lasso penalized regression. To see whether a larger replication set would increase the power to confirm both the primary and putative secondary signal, the sizes of the discovery and replication sets were adjusted so that they were approximately equal and the same method, described in section 5.1, was carried out on the new discovery and replication data sets. The data was divided as shown in Table 5.7

Table 5.7 Table showing the number of individuals in each of the GABRIEL cohorts that were used for the discovery and replication sets in the second fine mapping analysis. In this analysis, approximately equal sample numbers were used in both the discovery and replication analysis, in an attempt to increase the power to replicate a putative secondary effect.

Discovery Set		
Cohort	**Case**	**Control**
BAMSE	239	246
FIN	33	36
KAB	841	851
POKOV	112	116
SAPAL	237	356
SLSJ	373	390
GSK	462	1576
UFA	269	209
TOM	197	91
Total	**2763**	**3871**
Replication Set		
ALSPAC	607	609
EGEA	482	598
WJST	279	620
BUSSEL	188	390
PIAMA	172	187
UK/GER	861	1034
CNS	88	182
B58C	213	200
Total	**2890**	**3820**

Lasso penalized regression was carried out on the initial discovery set in Mendel and out of the top 20 SNPs that were included in the final lasso model, 4 SNPs reached genome-wide significance levels (p-value < 5×10^{-8}).

Table 5.8 Results of the lasso analysis indicating the SNPs associated with childhood onset asthma with genome-wide significance (p-value < 5×10^{-8}).

Top Lasso Predictors					
Predictor name	Position (BP)	Univariate P-value	LOO index	Regression estimate	Single SNP reg. est.
rs1054609	35342870	0.75366E-15	0.6743	0.21096	0.25002
rs4795397	35310723	0.76553E-15	0.41552	-0.38243	-0.28417
rs11078926	35328044	0.12230E-14	0.82831	-0.04234	-0.28417
rs7212938	35376206	0.40687E-09	0.1109	0.07539	0.2211

These SNPs were again analysed in Haploview to assess the LD between them in an attempt to differentiate between independent signals in the data. The lasso did not identify the same primary signal in this analysis, as it did in section 5.1, but the primary signal identified was rs11078926 (p=0.1223E-14), which Haploview showed to be almost perfectly correlated (r^2 = 0.99) with rs8069176 (p=0.1068E-14), identified in the first analysis.

A meta-analysis of rs11078926 and the putative secondary signal, rs7212938, was then carried out in STATA using the same technique described in the previous section. The results of the meta-analysis in STATA for the two SNPs are shown in Tables 5.9 and 5.10.

Table 5.9 Meta-analysis of rs11078926 using samples from the replication cohorts, consisting of 2890 childhood asthma cases and 3820 controls.

Meta-analysis: rs11078926				
Study	Odds ratio	95% Confidence Interval		% Weight
ALSPAC	0.792	0.646	0.971	23.52
EGEA	0.773	0.624	0.959	21.28
WJST	0.914	0.712	1.174	15.68
BUSSEL	1.138	0.827	1.567	9.59
PIAMA	0.786	0.550	1.123	7.69
UK/GER	0.886	0.646	1.216	9.80
CNS	1.089	0.693	1.712	4.79
B58C	0.937	0.655	1.340	7.65
Pooled Effect	0.867	0.785	0.957	100

I^2 = 0.0%
Test of ES=1: p = 0.005

The association of rs1107926 with the phenotype showed both a decreased odds ratio and level of significance in the second analysis (pooled effect = 0.867, p=0.005) as compared to rs8069176 in the first (pooled effect = 0.801; p=0.0005).

Table 5.10 Meta-analysis of rs7212938 using samples from the replication cohorts, consisting of 2890 childhood asthma cases and 3820 controls.

Meta-analysis: rs7212938				
Study	Odds ratio	95% Confidence Interval		% Weight
ALSPAC	1.078	0.868	1.340	22.81
EGEA	1.068	0.855	1.335	21.61
WJST	1.281	0.981	1.674	15.05
BUSSEL	1.497	1.074	2.087	9.73
PIAMA	0.927	0.644	1.334	8.11
UK/GER	1.284	0.934	1.764	10.63
CNS	1.234	0.757	2.010	4.51
B58C	1.312	0.899	1.913	7.55
Pooled Effect	1.172	1.056	1.300	100

I^2 = 0.0%
Test of ES=1: p = 0.002

The effect size of rs7212938 is 1.074 (p-value = 0.180) and upon the second meta-analysis, this increased to 1.172 with a p-value of 0.002. This increase in significance could be because more samples were added to the replication set, which provides more power to detect an effect of smaller size. The I^2 value in this analysis is 0% (as opposed to the 13.3% heterogeneity-driven variation in the effect size seen in the first meta-analysis of this SNP) (Table 5.5).

Since neither of the discovery/replication sets managed to convincingly replicate any secondary signals, in the third step of the analysis, it was decided to combine the discovery and replication sets, in an overall meta-analysis, for the SNPs meta-analysed in section 5.1.

5.3 Meta-analysis of discovery and replication sets

Table 5.11 and 5.12 show the results of the meta-analysis between the replication cohorts and discovery set for the two SNPs (rs8069176 and

rs7212938) shown in section 5.1 to contribute independently to the childhood asthma phenotype. As with the previous meta-analyses, the coefficient values and standard errors for each of the SNPs (that are converted into odds ratios and 95% confidence intervals by STATA) are those values that were obtained when two SNPs were included in the model. Therefore, the parameters of the one SNP are adjusted for the other SNP in the model

Table 5.11 Meta-analysis of rs8069176 using samples from both the discovery and individual replication cohorts. In this analysis, rs8069176 is adjusted for SNP rs7212938.

Meta-analysis: rs8069176				
Study	Odds ratio	95% Confidence Interval		% Weight
ALSPAC	0.803	0.656	0.984	9.11
B58C	0.936	0.654	1.339	2.93
GSK	0.736	0.611	0.887	10.81
EGEA	0.772	0.624	0.957	8.19
WJST	0.903	0.704	1.158	6.06
Discovery	0.510	0.342	0.762	62.91
Pooled Effect	0.824	0.775	0.876	100

I^2 =0.0%
Test of ES=1; p = 2.65×10^{-10}

Table 5.12 Meta-analysis of rs7212938, adjusted for rs8069176, using discovery and replication sets.

Meta-analysis: rs7212938				
Study	Odds ratio	95% Confidence Interval		% Weight
ALSPAC	1.090	0.879	1.352	8.32
B58C	1.312	0.899	1.913	2.72
GSK	0.939	0.777	1.134	10.80
EGEA	1.057	0.846	1.321	7.79
WJST	1.271	0.975	1.658	5.49
Discovery	1.161	1.075	1.254	64.88
Pooled Effect	1.130	1.062	1.202	100

I^2 = 16.2%
Test of ES=1; p = 6.15×10^{-5}

The results of these meta-analyses are promising. The primary signal had an odds ratio of 0.824 (p-value=2.65×10^{-10}) (Table 5.11) and the secondary signal,

adjusted for this SNP, had an odds ratio of 1.130 (p-value=6.15×10^{-5}) (Table 5.12). While the secondary signal failed to reach genome-wide significance, it is possible that this was due to inadequate sample sizes, rather than a spurious signal.

Overall, although the putative secondary signal (rs7212938) failed to replicate with genome-wide significance in the overall meta-analysis of the discovery and replication cohorts, the possibility of more than one signal contributing to the variation in the childhood asthma susceptibility locus on chromosome 17, cannot be ruled out. It is encouraging that there is directional consistency across the discovery and replication sets and, as previously mentioned, the failure to replicate could be due to insufficient power to detect signals of relatively smaller effect sizes.

One of the risks involved in the methods of analysis described in both section 5.1 and 5.2, is that re-analysing the same set of data using different variable selection techniques (i.e. lasso penalized regression, followed by further model selection methods on the same set of data) introduces the possibility of increased bias and false positive rates in the study. For this reason, it is pertinent that stringent multiple testing controls are adhered to, which in the case of this investigation, should be genome-wide significant p-values of less than 5×10^{-8}. Since these cut-off values were not reached by the models in this study, no clear cut conclusions can be made about the nature of the possible secondary signals that may or may not reside in the susceptibility locus.

5.4 Re-analysis of the chromosome 17 region using 1000 genomes

imputation

The MACH software that was initially used to carry out imputation in this study uses reference data from the International HapMap Project to infer the genotypes at all untyped markers in the data set. While this strategy provides a successful method of imputation, when this analysis was carried out, MACH was limited to using reference panels from HapMap. However, imputation methods that are able to incorporate information from dense genome-wide haplotypes from the 1000 Genomes Project, are likely to provide more accurate imputation results (HOWIE *et al.* 2009) and may provide more power and resolution to fine mapping studies.

Thus, Impute v2.1.0 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) was used to carry out imputation across the region on chromosome 17 encompassing the childhood asthma susceptibility locus. This programme combines the reference haplotypes from the 1,000 Genomes Project (containing many more SNPs than the HapMap) and HapMap Phase 3 (consisting of a larger sample of human chromosomes) and integrates these wide and deep panels into a single analysis framework. We imputed based on the CEU individuals from 1000 Genomes, as well as TSI and additional CEU individuals from HapMap3, since the individuals from this investigation were of European decent and therefore these reference panel shared the most similar ancestry. Imputation was carried out across a 3megabase region on chromosome 17 (physical base pair position 35 000,000 – 38 000,000) and lasso was again applied to the newly imputed data set. The “best guess” genotypes were taken as the true genotypes for this study.

In order to achieve the maximum amount of power, the samples that were used in this analysis were the same as the samples used in section 5.1 (see Table 5.1), consisting of 3610 childhood asthma cases and 4088 controls. Impute outputs an "r2_typeX" file that contains the squared correlations between the imputed and real genotypes. SNPs that had a value of less than 0.3 were excluded from the analysis to ensure that poorly impute SNPs were not analysed. In total, 8067 SNPs were included in the final data set. The results of the lasso analysis are shown in Table 5.13.

Table 5.13 Results of the lasso analysis carried out on chromosome 17 (physical base pair position 35 000,000 – 38 000,000) that has been imputed using impute v2.1.0. The SNPs that were included in the final model and significantly associated with childhood onset asthma (p-value < 5×10^{-8}) are shown below.

SNP ID	Position (bp)	Univariate p-value	LOO Index	Regression estimate	MAF	Univariate estimate
rs2872516	35326253	0.45308E-15	0.89506	-0.02043	0.4239	0.26255
rs4795399	35314965	0.10683E-14	0.51558	0.08835	0.4141	0.25976
rs3902920	35328542	0.11073E-14	0.70151	0.03601	0.4858	0.25688
rs7223717	35368859	0.21788E-14	0.81725	0.01657	0.4153	0.26098
rs3894193	35378872	0.11042E-13	0.33517	-0.07342	0.4705	-0.25201
rs60134943	35387318	0.47158E-12	0.58926	-0.03276	0.4721	-0.23538
rs72832941	35296965	0.21637E-09	0.05132	0.09801	0.2097	0.25274
rs8078692	35468643	0.12440E-08	0.17562	0.06140	0.4432	0.20038

The lasso regression estimates in Table 5.13 are all much lower than the univariate regression estimates for each of the SNPs. This is likely due to the presence of correlated markers in the final lasso model, because lasso tends to share the coefficient value of the overall signal amongst all the correlated variables that explain it in the model. It can also be seen from the high LOO indices that correlated variables have been included in the final lasso model. These outcomes are not surprising, since the high density of SNPs in the imputed region of chromosome 17 would have shown extensive LD. To assess

the extent to which the significant SNPs in the final lasso model are correlated with each other, a plot showing the measures of pair-wise LD between these 8 SNPs was generated in Haploview (Figure 5.4).

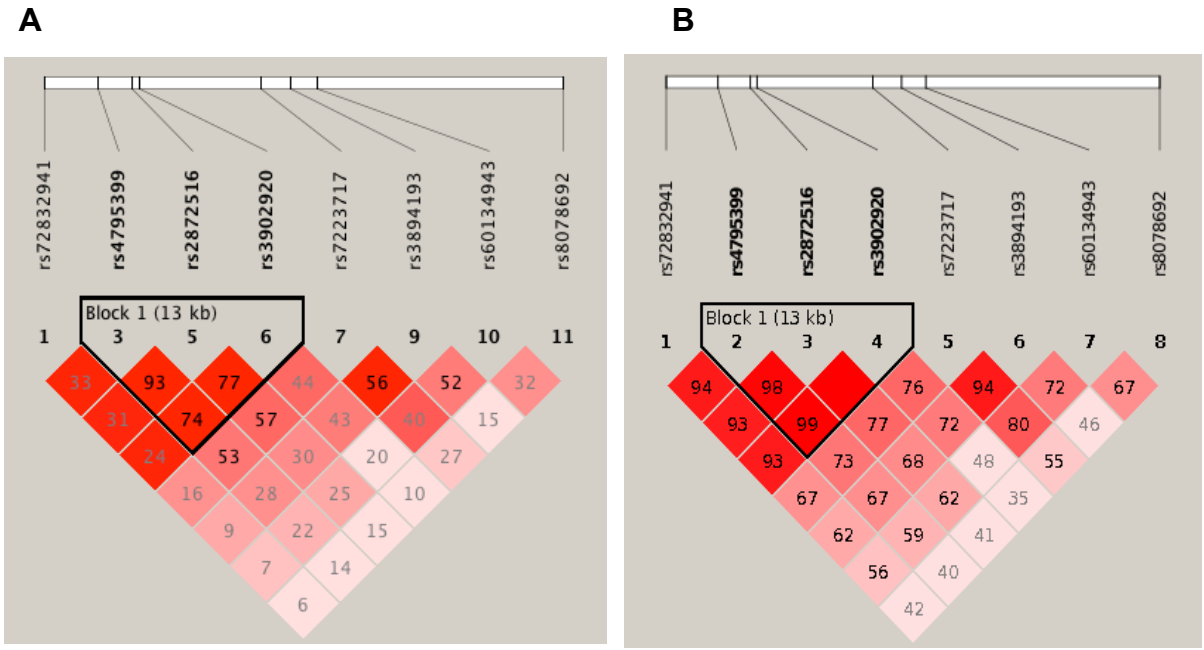


Figure 5.4 Linkage Disequilibrium plot showing the LD pattern across the 8 top lasso hits identified from analysis done on data that had been imputed using 1000 genomes and HapMap 3 samples. Plot A and B show the r^2 and D' values between these SNPs, respectively,

The most significantly associated SNPs in this analysis are in high LD with the most significant SNPs seen in the previous two analyses (rs8069176 and rs1107926). The primary signals from all three analyses reside within the linkage disequilibrium block, *Block 1,* in Figure 5.4. The secondary signal in the first two analyses (rs7212938) is in the same haplotype block as rs3894193 in Figure 5.4.

What is interesting from these results is the inclusion of SNPs rs60134943 and rs8078692, which have a significant association with childhood asthma in this analysis, but not in the previous two, as they are newly imputed SNPs. SNP rs60134943 (bp position 35387318) is positioned in the *GSDM1* gene, encoding the gasdermin 1 protein that resides in the same 17q21 region as *ORMDL3*.

SNP rs8078692 is situated in between the *MED24* and *THRA* genes and, as seen in Figure 5.4, does not appear to be in strong LD with any other SNP laid in Table 5.13. The *MED24* gene encodes a transcriptional co-activator complex that is thought to be required for the expression of almost all genes, while *THRA* encodes a nuclear hormone receptor for triiodothyronine and has been shown to mediate the biological activities of thyroid hormone. While the possible role of these genes in childhood asthma is unclear, these results provide statistical evidence that independent signals may reside within the 17q21 locus and contribute to the variation of the childhood asthma phenotype.

STATA was again used to carry out the model selection steps described in the previous two analyses, to assess whether the inclusion of any of the possible secondary signals from the lasso analysis (Table 5.13) would improve the fit of the model containing only the primary signal. The five best fitting models are shown in Table 5.14 below.

Table 5.14 Results of the model selection carried out in STATA for the data imputed using 1000 genomes and HapMap3 imputation.

Model SNPs	ll(null)	ll(model)	df	AIC	BIC
1 rs3902920	-5319.731	-5287.616	2	10579.23	10593.13
2 rs3902920; rs60134943	-5319.731	-5279.394	3	10564.79	10585.63
3 rs3902920;rs60134943;rs8078692	-5319.731	-5279.155	4	10566.31	10594.10
4 rs3902920;rs60134943;rs3894193	-5319.731	-5277.668	4	10563.34	10591.13
5 rs3902920; rs8078692	-5319.731	-5280.939	3	10567.88	10588.72

It can be seen from Table 5.14 that model 2 is the best model according to the BIC, while model 4 is the best model according to the AIC. A likelihood ratio test (lr test) was performed in STATA to check the significance of adding rs3894193 to model 2 and the results indicate that the addition of rs3894193 does not significantly improve the fit of the model (p-value of likelihood ratio test assuming

model 2 is nested in model 4 = 0.0632). This suggests that the addition of rs3894193 does not significantly improve the fit of the model and it is likely that model 2 is the most likely model.

Results from the model selection in STATA indicate that it is likely that rs60134943 could be a significant secondary childhood asthma susceptibility signal. It would be necessary to replicate this result using the remaining GABRIEL cohorts, but in order for this to be done, the cohorts would have to carry out the same imputation step using IMPUTE v2.0 and fit this model to the imputed data to see if the results replicate or not.

5.4 Conclusion

Fine mapping is an important step in the elucidation of the genetic factors that contribute to common disease susceptibility. Numerous studies indicate that variants in the 17q21 region are strongly associated with ORMDL3 expression and the onset of childhood asthma, and efforts are underway to not only explain the mechanism underlying this association, but to search for smaller secondary signals that may influence the overall contribution of this susceptibility locus to the disease.

Lasso penalized regression provides an effective and computationally efficient method of carrying out fine mapping analysis. The model selection property of lasso allows independent signals to be teased apart far more readily than in single SNP methods of analysis, where large numbers of highly correlated SNPs, significantly associated with the disease of interest, make it difficult to distinguish

between one primary signal and putative independent signals of smaller effect size in the same region. Since lasso generally includes one variable in the model and then drops any remaining variables that are highly correlated with this variable, the top lasso hits are, to an extent, independent of one another. In other words, lasso regression includes new variables into the final model, while conditioning on all the variables already in the model. In this way, both strong primary signals and moderate-small effect secondary signals can be identified in a computationally efficient way.

When large amounts of LD are present across the data being analysed, however, it was observed that groups of correlated SNPs can be included in the final model. While this is not ideal when looking specifically for independent association signals, it is important to note that even when the data set had been imputed using 1000 genomes and HapMap 2 imputation, and therefore contained many highly correlated marker variants, the final model included no more that 6 variables that were in strong LD with one another (Figure 5.4). Practically speaking, this means that setting lasso to select a model with an increased number of variables (for example, the top 30 or 50 lasso hits, instead of the top 20) should ensure that smaller effect sizes are not masked by strong primary signals.

Lasso penalized regression focuses on the selection of subsets of variables that systematically explain portions of the phenotypic variability being observed in populations. Therefore, lasso should be seen as a step towards unravelling the complex relationship between underlying genetic risk factors, rather than as an alternative to frequentist methods of analysis that look specifically at establishing

effect sizes and their respective significance levels. The lasso penalized regression and model selection criteria used in this chapter, have the potential to detect individual variants that reside in close genomic proximity to one another and contribute independently towards the overall aetiology of the disease of interest.

Fine-mapping approaches are necessary to advance our understanding of the genetic architecture underlying specific disease loci. Indeed, the discovery of multiple independent signals within a locus, that may have gone undetected in studies with insufficient power to detect their potentially small effect sizes, may explain a portion of the “missing heritability” of a disease. However, this portion is likely to be small, and therefore it remains crucial to search genome-wide, across all loci, to enable the identification of complex gene-gene interactions that may further contribute towards the unexplained heritability of common, human diseases, which will be considered in Chapter 6.

Chapter 6: Detecting gene-gene interactions in coronary artery disease using lasso penalized regression

Coronary artery disease (CAD) occurs when the arteries leading to the heart become hardened and narrow due to the build up of plaque (known as atherosclerosis). This restricts the heart's supply of blood and can often result in chest pain and ultimately heart failure. CAD is the most common type of heart disease and is the leading cause of death worldwide (MURRAY and LOPEZ 1997). Factors such as age, sex, hypertension, hypercholesterolemia, diabetes and smoking are known to play an important role in the pathogenesis of the disease (KHOT *et al.* 2003), but it is also widely accepted that there is a strong genetic component that controls CAD susceptibility. Further evidence for the genetic contribution to CAD risk is shown in twin studies (MARENBERG *et al.* 1994), as well as in epidemiological studies, which demonstrate that parental (LLOYD-JONES *et al.* 2004) and sibling (MURABITO *et al.* 2005) cardiovascular disease is an independent predictor of cardiovascular disease risk in their offspring.

CAD is a complex trait, controlled by multiple genes and environmental risk factors that may be further complicated by gene-gene and gene-environment interactions (OGAWA *et al.* 2010). Therefore, understanding the genetic basis of CAD is an important first step to tackling the dire health consequences of the disease – a step that would potentially enable scientists to improve current risk assessments and provide better measures for prevention and treatment.

While genome-wide association studies (GWAS) have proven to be successful in identifying many new disease susceptibility loci, these studies have focused mainly on the identification of common, marginal genetic risk factors. The estimated heritability of CAD is approximately 40% and, until now, common variants identified by GWA and meta-analysis studies, explain only 10% of the additive genetic variance of CAD (SCHUNKERT *et al.* 2011). This suggests that other common small-effect variants, rare and structural variants and the interactions between genes, may account for the unexplained portion of CAD heritability.

Statistical methods used in GWAS do not typically consider the effects of gene-gene interactions and therefore the genetic contribution of these effects to common diseases like CAD, are likely to go undetected, unless they also have strong marginal effects. The incorporation of multi-locus methods of analysis such as lasso, may therefore have the potential to increase the power of GWAS by allowing for such gene-gene interactions. Furthermore, the identification of interaction effects between genes may provide us with valuable information about the biological mechanisms underlying CAD. Therefore, the aim of this investigation was to assess the ability of lasso penalized regression to detect potential gene-gene interactions that may play a role in the aetiology of the disease.

6.1 Detecting gene-gene interactions in GWAS

It is broadly accepted that gene-gene interactions play an important role in common disease susceptibility (CORDELL 2002). There is, however, a great deal

of uncertainty about how common these interactions are and how important they are likely to be, relative to independent main effects (MOORE 2003). Despite this uncertainty, many studies have focused on the potential gain in power when allowing for interactions in association testing. Marchini *et al*. (MARCHINI *et al.* 2005) concluded that, in the presence of gene-gene interactions, fitting models that allow for interaction effects within a GWAS context, can substantially improve the power of single-locus approaches. Furthermore, Evans *et al*. (EVANS *et al.* 2006) found that, in situations where the majority of the genetic variance is explained by the interaction variance component, two-locus methods that allow for interactions, outperformed single-SNP methods of analysis that were unable to identify the loci involved in the interaction. These studies highlight the importance of developing powerful and efficient test that allow for gene-gene interactions.

The biological interaction between genes was first termed ‘epistasis’ by Bateson in 1909, and describes the effect whereby a variant at one locus prevents the variant at another locus from manifesting its effect (MOORE 2003). In contrast, statistical interactions can be defined as the departure from additivity in a model (CORDELL 2009) and it is these interactions that will be the focus of this chapter.

In a multiple logistic regression model (described in detail in Chapter 2), testing for an interaction would entail a one-degree of freedom test that tests the hypothesis that the regression coefficient of the interaction term is 0. For example, assuming an additive-additive interaction when fitting a logistic regression model for two loci, A and B, where A and B are coded as 0,1 or 2 according to the number of minor alleles in the genotype, the interaction between

them can be modeled as follows:

$$\ln(p/1-p)) = \mu + \beta_1 A + \beta_2 B + \beta_3 A*B$$

Thus, a test to assess the significance of the association between the interaction term (A*B) and the phenotype of interest, would involve testing the hypothesis that $\beta_3=0$ (CORDELL 2009). While this concept is simple, data sets with large numbers of predictor variables increase exponentially when considering pair-wise interactions (and even more so for higher order interactions between three or more variants), and thus the computational burden of examining all the possible pair-wise interactions, as well as multiple testing problems, is a severely limiting factor in this type of analysis.

One way to circumvent the computational intensity of performing an exhaustive search of all the possible two-way interactions, is to focus on the potential interactions between variants that show strong marginal effects. While this method reduces the dimensionality of the data, it completely excludes factors that display interaction effects without displaying marginal effects. According to Cordell (CORDELL 2009), many statistical models that include an interaction term in the absence of main effects, still show genotype-phenotype correlations that can be detected in single-locus analysis, but focusing purely on the interactions between main-effects only, could exclude potentially important gene-gene interactions. In agreement with this, Evans *et al*. (EVANS *et al.* 2006) found that two-stage methods that rely on selecting a subset of predictors with strong marginal effects, will only succeed if a portion of the genetic variance is explained by one of the loci's variance components. Therefore, the best method

chosen to investigate the role of gene-gene interactions in complex human diseases will depend heavily on the resources available to run analyses and the initial hypothesis about the underlying genetic architecture of the disease under study.

It is clear that in large, undetermined problems, where the number of observations far exceeds the number of samples in the data set, algorithms designed to efficiently reduce data dimensionality, will be key analytical tools. For this reason, lasso penalized regression was applied to coronary artery disease case/control data from the PROCARDIS study, to carry out a genome-wide analysis of gene-gene interactions.

6.2 Description and preparation of the CAD data

As part of the PROCARDIS study, 3146 CAD cases and 3352 controls were recruited from United Kingdom, Italy, Sweden, and Germany. Only individuals diagnosed with CAD before the age of 66, with a sibling who had also been diagnosed with CAD before the age of 66, were included as case subjects and control samples comprised of individuals with no personal or sibling history of CAD before the age of 66.

Samples were genotyped using the HumanCVD BeadChip (Illumina). This platform uses 48,742 SNP markers to capture the variation across approximately 2100 candidate genes, as well as pathways for cardiovascular, inflammatory and metabolic phenotypes (KEATING *et al.* 2008). CAD susceptibility variants previously identified in GWA studies were also included on the chip. This chip

was designed to maximise the coverage at loci that have been previously implicated in cardiovascular related diseases, across all HapMap populations. SNPs with low call rates (<95%), MAFs less than 1% and Hardy–Weinberg p-values less than $1x10^{-6}$ were excluded from the analysis.

Missing data was imputed across the samples, as part of the PROCARDIS study, using the MACH software. This imputation analysis did not rely on an external reference set of phased haplotypes, and therefore variants with 1%<MAF<5% could be sensibly included in the data set and analysed. After imputation, 47155 SNPs with complete genotype data were available for analysis. It was necessary to remove poorly imputed SNPs and therefore 9168 SNPs with r-squared values less than 0.3, were excluded from the study. A further 3843 imputed SNPs with MAF less than 1% were also excluded and, as in the previous chapters, the “best guess” genotypes from the imputation analysis were used as the true genotypes in this study.

After these quality control steps were carried out using PLINK, the cleaned, imputed data was loaded into Mendel. In total, 34,144 SNPs were available to test associations with CAD across 3146 CAD cases and 3352 controls.

6.3 Interaction analysis in Mendel

In this investigation, pair-wise (or two-way) gene-gene interaction analysis was carried using lasso penalized regression and implemented using Mendel software. Even for computationally efficient model selection methods, such as lasso, the computational capacity needed to search for all the possible pair-wise

interactions potentially contributing to CAD, could not be met. Thus, a two-step method was used in order to ease the computational burden of the analysis.

In the first step, lasso was used to perform a marginal analysis and select the 2000 most significant predictors associated with CAD. While marginal lasso analyses would generally focus on a smaller subset of top hits (because the likelihood of having large numbers of independent signals in one data set is low), when running the interaction analysis in Mendel, this initial group of main effects was pre-set to include as many predictors as was computationally feasible. Thus, this investigation focussed computational time on increasing the pool of lasso main effects, to minimise the chance of excluding predictors that have non-significant marginal effects, but may be involved in important interactions.

In the second step of this analysis, lasso was again used to examine the 2000 marginal SNPs identified in the first step, as well as all the possible pair-wise interactions between these 2000 SNPs. The pair-wise interaction terms are modelled in Mendel by taking the genotypic products of the marginal predictors involved in the interaction and individual predictors are treated as one-way interactions. Therefore, in the second stage of analysis, lasso was used to search for the 30 one-way or pair-wise interactions that best explained the variation in the CAD phenotype. This step was limited to only include 30 terms in the final model, so as to ease the overall computational demand.

Marginal p-values, LOO indices and regression estimates for both the marginal analysis (stage 1) and the interaction analysis (stage 2) are calculated in Mendel as described in Chapter 4.

6.3.1 Results of the two-stage lasso interaction analysis

As mentioned, the first step of the lasso analysis, carried out on the imputed data set, selected the top 2000 marginal predictors associated with CAD. Table 6.1 shows the 9 lasso SNPs that reached genome-wide significance in this first step.

Table 6.1 The 9 marginal SNPs associated with coronary artery disease, with genome-wide significance, selected by lasso in the first step of the interaction analysis.

SNP	Chr	Position (Bp)	Univariate P-value	LOO Index	MAF	Regression Estimate
rs10455872	6	1690930108	6.64E-18	1.75E-20	0.0930	-0.8640
rs10757272	9	22078260	2.80E-13	2.36E-05	0.4876	0.3330
rs1333049	9	22115503	1.12E-11	0.9641	0.4963	1.8463
rs10965224	9	22057276	6.69E-11	1.0000	0.3805	1.8211
rs4970834	1	109616403	9.49E-10	0.9436	0.1625	9.0853
rs602633	1	109623034	3.03E-09	0.2507	0.1926	0.1455
rs1333040	9	22073404	2.03E-08	0.9969	0.1926	0.7574
rs1412832	9	22067543	7.85E-08	1.0000	0.2932	-0.0161
rs2075650	19	50087459	2.40E-07	0.9819	0.1547	-1.1865

Lasso identified rs10455872 as the SNP most significantly associated with CAD (p-value = 6.64×10^{-18}; Table 6.1). This SNP tags the 6q27 chromosomal region in which the *LPA* locus (encoding LP(a) lipoprotein) resides and this region has a well-established association with CAD (CLARKE *et al.* 2009; SCHUNKERT *et al.* 2011). Of the 9 significant lasso SNPs shown in Table 6.1, five reside within the 9p21 region, which has also been shown to be significantly associated with CAD in genome-wide association studies (CLARKE *et al.* 2009; SAMANI *et al.* 2007; SCHUNKERT *et al.* 2011). Two of the SNPs in Table 6.1 map to the 1p13 chromosomal region, which too has been established as a genome-wide significant CAD risk locus (CLARKE *et al.* 2009; SAMANI *et al.* 2007). Finally, a CAD-associated signal was identified on chromosome 19, near *TOMM40* and *APOE*. This region harbours risk alleles for carotid artery disease (CAAD), and it

has been suggested that rs2075650 has an effect on low-density lipoprotein (LDL) buoyancy. Since smaller, denser LDL particles make up the more atherogenic LDL pattern, it can be seen that this region also has important implications for CAD (RONALD *et al.* 2009).

Table 6.1 highlights the nine genome-wide significant marginal effects identified by lasso in the first stage of the interaction analysis, but all 2000 SNPs selected in this step were carried into the second stage of the analysis.

In step 2, lasso identified the 30 interactions (one-way or pair-wise) that best explained the variation in the CAD phenotype. This step of the analysis took 71 hours and 43 minutes to run (on a Sun Fire V40z server, 2.6 GHz with a 4-dual core 64-bit AMD processor and 28GB RAM) and highlights the intense computational cost of this type of analysis, even when the interaction step is limited to pair-wise interactions between 2000 variants. The results of this step are shown in Table 6.2.

It was evident upon the examination of these results, that many of these interactions were likely to be false positive results. Out of the 30 possible one-way or two-way interactions, only one one-way interaction (marginal SNP rs10757272) was included in the final lasso model. Some of the SNPs with significant marginal effects were included in a number of different interactions with other SNPs and, in some cases, the p-value of the interaction term was the same as the marginal p-value of one of the SNPs making up that interaction. Furthermore, some of the interaction terms had high LOO indices and it was therefore likely that not all of these interactions were independently associated

with CAD.

Table 6.2 The thirty marginal effects/pair-wise interactions, selected in the second step of the lasso interaction analysis, that best explain the variation in the CAD phenotype.

P-value	LOO Index	Regression estimate	Interaction members	
1.22E-19	0.0802	-0.1399	rs10455872	rs1256063
1.92E-18	0.2166	-0.1109	rs10455872	rs2288817
1.42E-17	0.0005	-0.1711	rs10455872	rs2075650
1.10E-16	0.3453	-0.0770	rs10455872	rs3746619
4.62E-14	0.0170	0.2132	rs10757272	rs6974518
4.87E-14	0.1067	-0.0854	rs10455872	rs7077817
5.61E-14	0.0078	-0.1342	rs10455872	rs7110197
6.83E-14	0.0474	-0.1011	rs10455872	rs4148872
7.35E-14	0.0046	0.1630	rs10757272	rs10491051
7.39E-14	**0.0152**	**0.1160**	**rs10757272**	**rs2425634**
7.72E-14	0.0455	0.1176	rs10455872	rs10757272
7.73E-14	0.0169	0.1009	rs602633	rs12679834
1.72E-13	0.1399	0.0859	rs4970834	rs6837735
2.02E-13	**0.2622**	**0.1400**	**rs10757272**	**rs9594782**
2.80E-13	0.0005	-0.4079	rs10757272	
8.32E-13	0.2706	0.0694	rs10965224	rs1367413
1.22E-12	0.5969	0.0408	rs4970834	rs6065
1.23E-12	0.3367	-0.0626	rs10965224	rs6974518
1.25E-12	0.0029	-0.1390	rs10455872	rs9939224
1.47E-12	1.0000	0.0140	rs10965224	rs17047757
1.60E-12	0.1088	0.0974	rs10965224	rs3015656
1.69E-12	1.0000	0.0812	rs602633	rs6837735
1.21E-11	0.0036	0.1561	rs1412832	rs3917245
1.54E-11	0.0606	0.1105	rs4970834	rs37602
2.80E-11	0.0022	0.1579	rs4970834	rs6967267
4.56E-11	0.3855	0.0296	rs1412832	rs36097961
1.69E-10	0.0085	0.1333	rs602633	rs6703796
1.44E-09	0.0001	-0.1749	rs10455872	rs2285428
4.43E-09	**0.9382**	**0.0077**	**rs3780909**	**rs3777142**
4.54E-09	0.8496	0.2171	rs3780909	rs2303070

For example, the last two interaction terms in Table 6.2 (rs3780909*rs3777142 and rs3780909*rs2303070) had LOO indices of 0.9382 and 0.8496 respectively. Upon examination of rs3777142 and rs2303070 in SNAP, it was found that these two SNPs were highly correlated with one another (r^2 = 0.901) and both interaction terms were modelling the same underlying statistical interaction.

Once the initial two-step interaction analysis had been implemented in Mendel, a further model selection step was carried out in STATA, in order to assess the validity of these putative gene-gene interactions.

6.4 Model selection in STATA

Firstly, STATA was used to fit main-effects logistic regression models for each of the 29 pair-wise interactions laid out in Table 6.2, with and without the interaction term, to assess whether the addition of the interaction term improved the overall fit of the model. The interaction terms were generated in STATA by taking the product of the genotypic values of the two relevant SNPs.

It is important to note that data used for this part of the analysis, was the original, un-imputed genotype data described in section 6.2. We used this data set for the model selection step, because, unlike the lasso function in Mendel, STATA is capable of fitting logistic regression models to data that contains missing values. However, samples with missing values at a particular SNP are excluded from the regression analysis and this could cause potential problems when generating interaction terms between two SNPs with differing sample numbers. In light of this, any individuals that were excluded from the analysis for missing values, were stored in a variable called *e(sample).* This variable was then included as a command statement in the logistic regression analysis, in order to ensure that the same individuals were examined in main-effects models and the models containing interaction effects.

The data set also included some related samples. In any analysis where related

individuals are treated as independent samples, the amount of information in that dataset is overestimated. Since error is inversely proportional to the amount of available information (error = 1/information), ignoring the relatedness in the sample would underestimate the standard errors of the regression coefficients – resulting in narrower standard error bands. Therefore, all logistic regression models fitted in STATA allowed for familial clustering (using the *cluster* command), to ensure a robust estimation of the covariance matrix.

As mentioned earlier, the samples were made up of individuals from United Kingdom, Italy, Sweden and Germany and, for each individual, their country of origin was recorded in a covariate called *group*. To ensure that population structure had minimal confounding effects on this analysis, the *group* covariate was included as a factor in the logistic regression models, thereby controlling the possibility of over-dispersion in the case and control groups.

Model selection steps were then carried out to compare the fit of main-effects only models with models that included an interaction term. The fit of each of the models was compared using the BIC, AIC and the Wald test. While likelihood ratio tests are generally used to compare the fit of two models, the asymptotic distribution of the likelihood ratio statistic is approximated by a chi squared distribution, under the assumption that samples are independent of one another (KENT 1982) and therefore, cannot be used for this data. Thus, because related individuals were included in this study, the Wald test was used in place of the likelihood ratio test.

The Wald test is a parametric statistical test used to test the significance of a

predictor in a statistical model and does not assume independence of the samples. This test works by testing whether the parameters associated with a particular variable (or a group of variables) in a model, is zero. If the Wald test is significant, it strongly suggests that removing these parameters from the model will significantly reduce the model fit and should therefore remain in the model. In contrast, if the Wald test is not significant, it suggests that dropping this variable will not substantially reduce the fit of that model.

6.4.1 Results of the model selection carried out in STATA

From the 29 pair-wise interactions shown in Table 6.2, the inclusion of four of these interactions was found to improve the fit of the main effects models, based on the AIC and BIC. However, as explained in section 6.3, two of these possible interactions (rs3780909*rs3777142 and rs3780909*rs2303070) were essentially modelling the same statistical interaction and therefore the most significant of these interactions was considered for further analysis. The results of the model selection analysis for each of these three cases are shown below.

Table 6.3 Results of the model selection carried out in STATA to investigate whether the inclusion of the interaction between rs10757272 and rs2425634 (int1) improved the fit of the model consisting of main effects only.

Model terms	No. Obs	AIC	BIC	Df	P-value
1 rs10757272; rs2425634	6494	8932.100	8973.772	6	5.76E-12
2 rs10757272;rs2425634; int1	6494	8925.926	8972.376	7	7.73E-13
3 rs10757272; int1	6494	8925.124	8965.796	6	3.33E-13
4 rs2425634; int1	6494	8986.949	9027.621	6	0.0033
5 int1	6494	8991.341	9025.234	5	0.0143

Based on the information provided by the AIC and BIC, model 3, consisting of rs10757272 and the interaction term (int1), is the model that fits the data the best. A Wald test was then carried out in order to assess the significance of each

term on the overall model and the results are shown in Table 6.4

Table 6.4 Results of the Wald test carried out in STATA, to test whether each term, or combination of terms, in the model has a significant effect on the phenotype. The full model consists of the marginal SNPs rs10757272 and rs2425634, as well as the interaction between them (int1).

Wald test	
Test	**P-value**
int1 = 0	0.0073
rs2425634 = 0	0.3120
rs10757272 = 0	7.53E-13
rs2425634 = int1 = 0	0.0052
rs10757272=int1=0	3.96E-12

The results of the Wald test show that show that we are not able to reject the null hypothesis that coefficient of rs2425634 = 0 (based on the p-value of 0.3120). This means that removing this SNP from the model would not significantly decrease the overall fit of that model, and it should be dropped. In contrast, this test indicates that the coefficient for rs10757272 is not equal to zero ($p=7.53x10^{-13}$) and the coefficients for rs10757272 and int1 are not simultaneously equal to zero ($p=3.96x10^{-12}$), suggesting that these variables create a statistically significant improvement in the fit of the model. The results of the Wald test are thus consistent with the information provided by the AIC and BIC and, according to all three criteria, model 3 (Table 6.3) is the most probable model.

The SNP with the most significant marginal effect, rs10757272, maps to the 9p21.3 region and was one of the SNPs reaching genome-wide significance in the first step of the lasso analysis (Table 6.1). This region spans a 58 kb LD block on chromosome 9 that does not contain any known protein-coding genes (VISEL *et al.* 2010), but is located nearby *CDKN2A/B*. Visel *et al*. (VISEL *et al.* 2010) suggest that this region affects CAD progression by altering the dynamics of vascular cell proliferation. Alternatively, SNP rs2425634 maps to the *HNF4A*

gene, which encodes a nuclear transcription factor that controls the expression of several genes; including hepatocyte nuclear factor 1 alpha (*HNF4A*). This gene regulates the expression of several hepatic genes and common *HNF4A* variants have been shown to be associated with high serum lipid levels and the metabolic syndrome (WEISSGLAS-VOLKOV *et al.* 2006).

The results of the model selection step for SNPs rs10757272, rs9594782 and the interaction between them (int2) are shown in Table 6.5 and Table 6.6 below.

Table 6.5 Results of the model selection carried out in STATA to investigate whether the inclusion of the interaction between rs10757272 and rs9594782 (int2) improved the fit of the main effects only model.

Model terms	No. Obs	AIC	BIC	Df	P-value
1 rs10757272; rs9594782	6275	8620.452	8660.918	6	4.52E-13
2 rs10757272;rs9594782; int2	6275	8620.320	8667.530	7	5.45E-13
3 rs10757272; int2	6275	8619.436	8657.901	6	1.77E-13
4 rs9594782; int2	6275	8674.916	8715.382	6	0.0002
5 int2	6275	8681.410	8715.132	5	0.0011

Table 6.6 Results of the Wald test carried out in STATA, to test whether each term, or combination of terms, in the model, has a significant effect on the phenotype. The full model consists of rs10757272, rs9594782 and int2 - the interaction between them.

Wald test	
Test	**P-value**
Int2 = 0	0.1701
rs9594782 = 0	0.3860
rs10757272 = 0	5.56E-12
rs9594782 = int2 = 0	0.0015
rs10757272=int2=0	4.23E-11

The AIC, BIC from this analysis suggest that model 3 is the model that best fits the data. The Wald test shows that while dropping rs9594782 from the model would not decrease the significance of this model, simultaneously dropping rs10757272 and int2, would ($p=4.23x10^{-11}$; Table 6.6). SNP rs10757272 is again included as part of this putative interaction, as well as rs9594782. This

polymorphism resides in the *RANKL* gene, which has been demonstrated to be associated with aortic calcification in females (RHEE *et al.* 2010). Aortic calcification has been known to have serious clinical consequences, such as coronary artery disease, stroke and cardiovascular mortality.

The results of the model selection for SNPs rs3780909 and rs3777142 and the interaction between them (int3) are laid out in Table 6.7 and Table 6.8.

Table 6.7 Results of the model selection carried out in STATA to investigate whether the inclusion of the interaction between rs3780909 and rs3777142 (int3) improved the fit of the main effects only model.

Model terms	No. Obs	AIC	BIC	Df	P-value
1 rs3780909; rs3777142	6280	8660.638	8701.104	6	8.08E-07
2 rs3780909;rs377714; int3	6280	8645.551	8692.559	7	4.85E-09
3 rs3780909; int3	6280	8645.349	8686.017	6	3.39E-09
4 rs3777145; int3	6280	8674.174	8714.640	6	0.0001
5 int3	6280	8672.229	8705.951	5	4.54E-05

Table 6.8 Results of the Wald test carried out in STATA, to test whether each term, or combination of terms, in the model, has a significant effect on the phenotype. The full model is made up of the marginal SNPs rs3780909 and rs377714, as well as int3 (the interaction between the two SNPs).

Wald test	
Test	**P-value**
Int3 = 0	3.28E-05
rs3780909 = 0	2.22E-07
rs3777142 = 0	0.0767
rs3780909 = int3 = 0	8.55E-08
rs3777142=int3=0	7.69E-06

Model 3 (Table 6.7) has the lowest values of AIC and BIC, suggesting that this is the most probably model. In agreement with these criteria, the Wald test shows that removing rs3780909 and int3 from the full model, would result in a decreased model fit and therefore these variables are significant when considered together in the model.

The SNP with the highest marginal effect in this analysis, rs3780909, is situated in the chromosomal region 10q11.21, which was recently identified by Schunkert *et al.* (SCHUNKERT *et al.* 2011) in a large-scale analysis, to be significantly associated with CAD. The relatively smaller sample size of this investigation is responsible for the decreased power to detect associations with genome-wide significance. Nevertheless, this region on chromosome 10 lies 100kb downstream from the *CXCL12* gene. CXCL12 plays a role in the recruitment of leucocytes in response to vascular injuries and has been implicated in atherosclerosis in rodent models (MEHTA *et al.* 2011). rs2303070 maps to the *SPINK5* gene, encoding a serine protease inhibitor that may play a role in anti-inflammatory and/or antimicrobial protection of mucous epithelia. Although no studies have published significant associations between variants in the *SPINK5* gene and CAD, research has shown that inflammation is an important factor in the development of CAD (HANSSON 2005).

Interestingly, we noted from all three putative interactions that one (or both) of the SNPs making up the interaction had non-significant p-values in the first stage of the lasso analysis that selected the pool of marginal SNPs only. Besides the main-effect SNP, rs10757272, the four remaining SNPs involved in these 3 potential gene-gene interactions do not have significant marginal effects. The results of these four SNPs from the first lasso step are shown in Table 6.9.

Table 6.9 Results of the first step of the lasso analysis for the SNPs involved in the three putative interactions that were not laid out in Table 6.1.

SNP	Chr	Position (Bp)	Marginal P-value	LOO Index	Regression Estimate
rs2425634	20	42454870	0.03256	0.9480	-3.2441
rs9594782	13	42049186	0.0012	0.9856	-0.6762
rs3780909	10	45245221	1.07E-05	1.0000	0.9730
rs3777142	5	147467712	0.0050	0.9998	-0.0003

These terms would have been excluded if the interaction analysis had focussed on significant marginal effects only and thus highlights the benefit of using computational power to maximise the pool of SNPs selected in the first step of the lasso analysis.

6.6 Conclusion

Finding ways to overcome the computational demand of examining very large data sets, that expand exponentially when considering all statistical interactions, is a huge challenge. Computationally efficient algorithms, such as lasso, can be used in these circumstances to reduce the dimensionality of data and overcome some of the problems associated with the simultaneous analysis of millions of predictors.

In this investigation the analysis of pair-wise interaction effects using a two-step lasso procedure, identified three putative interaction effects that improved the fit of the main-effects, model according to three independent model selection criteria. Furthermore, the p-value of these interaction terms was more significant than the p-values of the marginal SNPs involved in the interaction. In particular, the interaction between rs3780909 and rs3777142 had a p-value of 4.43×10^{-9}, whereas the p-values of the two SNPs in the marginal analysis were 1.07×10^{-5} and 0.005 respectively. This potential interaction highlights the shortcomings of searching for interactions between significant main-effects only. Loci that may exert no significance when considered individually, may interact with other loci in a way that could greatly contribute to the phenotype of interest. It also emphasises the importance of multi-locus methods of analysis, which focus on

the simultaneous effects of several loci on human diseases.

There are, however, inherent problems with the lasso interaction analysis in this study. Firstly, only the top 2000 lasso hits were examined for possible two-way interactions and therefore it is possible that variants that were not included in the initial pool of 2000 SNPs, were not considered for the pair-wise interaction analysis. However, since lasso tends to drop correlated variables from the model, the marginal p-value of the final SNP included in the lasso model, was 0.533 and therefore it is clear that even variables with no significant marginal effects were analysed for possible interaction effects.

Another limitation was that only the top 30 interaction effects were considered in the second step of the analysis. Thus it is possible that interaction effects could have been missed in this step. Having said that, the final interaction term included in the second step of the lasso model, did have a marginal p-value of 4.54×10^{-9} (Table6.2) and therefore further potential interaction terms may have been excluded anyway, on the basis of the genome-wide significance threshold.

An increase in computational capacity would allow this method to search through more than 2000 marginal predictors for larger numbers of pair-wise, and possibly higher order, interactions and this could increase the power to detect statistical interactions. For the purpose of this investigation, however, this method of analysis provided the best trade off between computational feasibility and testing for interactions among the largest possible subset of marginal SNPs.

Finally, it was observed in the second step of the lasso interaction analysis, that

the inclusion of an interaction term was favoured over the inclusion of the two SNPs making up that interaction. One reason for this could be that, by including an interaction term between two significant variables, the variance explained by both of these variables can be accounted for by the model at the expense of using one degree of freedom for the interaction term, as opposed to two for the main effects model. This emphasises the importance of the model selection step in STATA, which provides information about plausible interactions that improve the overall fit of the main-effect only models.

It is important to remember that discovering statistical interactions is only the first step towards understanding the role of these interactions in human disease. The difficulty lies in having to make inferences about the biological explanations underpinning the statistical interactions. Certainly, the incomplete LD between genetic markers and the actual genes involved in the interaction, could lead to significantly reduced power to detect the effects of these genes. Furthermore, the difficulties with mapping marker SNPs to causal genes will be increased exponentially with each order of interaction being investigated and, as demonstrated in this study and others (COFFEY *et al.* 2004; HARRELL 2001), gene-gene interaction analysis may be prone to high type 1 error rates. Therefore, as with all novel genetic associations, formal statistical replication in independent data sets is a crucial step in this type of analysis and putative statistical interactions need to be tested using functional approaches in order to confirm their biological role in complex diseases.

Chapter 7: General discussion of results and implications for future research

7.1 Thesis overview

The rise in the prevalence of many complex human diseases, such as T2D, asthma and CAD, is often attributed to rapidly changing environmental factors that range from poor lifestyle choices to an increase in global pollution. Twin and epidemiological studies, however, have emphasised the important contribution of genetic effects to complex disease pathogenesis and researchers continue to focus their efforts on the identification of these genetic factors.

Over the past few years, GWAS have successfully identified many common variants that play a role in the aetiology of complex traits and diseases. In fact, it has been shown across several studies, encompassing a wide array of phenotypes, that the combined effect of moderate-small effect variants can increase common disease risk by about 10-50% (LANDER 2011). While this contribution to the overall disease phenotype is relatively small compared to the highly penetrant genes involved in Mendelian disorders, the identification of these variants has readily advanced our understanding of the genetic architecture underlying common human disease. GWAS have also pinpointed plausible biological pathways that may be involved in disease aetiology (BALLARD *et al.* 2010; ELKS *et al.* 2010). Generally speaking, GWA studies represent the first step towards pinpointing specific disease mechanisms, and ultimately the development of target therapeutics.

Interestingly, even though common variants have small individual effects on disease, in some cases, these variants have provided targets for therapeutic intervention. The *HMGCR* locus, for example, which houses a common variant (MAF ~ 40%) with a small effect on low-density lipoprotein (LDL), encodes the protein that is targeted by statin drugs (LANDER 2011). Statins are a class of drug that are widely used to lower cholesterol levels and the risk of myocardial infarction, so this example highlights the potentially important biological implications of detecting common variants associated with disease. This gives further impetus to the development of more highly powered GWA studies, capable of detecting novel variants with smaller effect sizes, relative to the effects of loci currently detected by GWAS.

As discussed throughout this thesis, the single-SNP methods of analysis commonly used in GWAS, may not be powerful enough to detect both these small-effect variants, as well as loci that may be involved in gene-gene interactions thought to play an important role in complex traits. Multi-locus model selection methods, which consider the joint effect of many loci on a phenotype, may increase the power of GWAS to detect these effects, since they allow for the complex genetic relationships that are believed to underpin disease. However, multi-locus methods of analysis are plagued with large, undetermined problems, where the number of predictors in a data set far exceeds the number of samples.

Model selection methods circumvent both these problems by effectively reducing the dimensionality of GWAS data and modelling the simultaneous effect of multiple loci on a phenotype of interest. Forward and backward selection models

have been traditionally preferred as model selection methods because of their speed at searching through large numbers of variables for the final set of predictors that best explain the variation in the phenotype. These methods have also been criticised for being “greedy”, as variables are either included or excluded at a particular step in the model selection path, with no regard for the effects that the variables in the final model may have their correlation with the phenotype. Chapter 3 of this investigation showed that for a type 1 error rate of approximately 5%, these stepwise regression methods had a lower power to detect disease risk loci, relative to lasso penalized regression, which performs model selection with the same computational speed as stepwise selection, but allows variables to re-enter the mode at any stage during the selection path. The power of lasso to detect both single and multi-locus effects on the phenotype, was consistently higher than these stepwise methods across varying effect sizes, allele frequencies and disease prevalence. Bayesian model averaging showed comparable power, but the computational complexity of this method proved to be a severely limiting factor and therefore its application to GWAS would not be feasible. Thus, based on evidence from the simulation studies, lasso penalized regression was chosen as the most powerful and computationally efficient method for carrying out multi-locus model selection within the context of GWA studies.

In Chapter 4, lasso was successfully used to identify many of the variants associated with T2D in the WTCCC. One of the benefits of model selection over single-SNP methods of analysis was shown in this study, where the lasso analysis identified a variant tagging the *HHEX* locus, which did not reach genome-wide significance in the WTCCC, but has since been established as a

risk locus for T2D (VOIGHT *et al.* 2010). Furthermore, a putative independent signal within the *BCL11A* locus, only shown subsequent to the WTCCC study to be associated with T2D ($p<5x10^{-8}$) (VOIGHT *et al.* 2010), was also identified by the lasso analysis. Although the p-values of these variants were not genome-wide significant based on the WTCCC data, the lasso analysis was able to pinpoint them as two of the top thirty loci affecting T2D status. This highlights one of the potential advantages of using model selection methods in GWAS, rather than univariate analysis, where important loci may not necessarily be the loci with the most significant statistical associations.

Following on from this analysis, in Chapter 5, lasso was used to fine-map the 17q21 childhood asthma risk locus in an attempt to identify putative secondary signals that may reside within this region. Model selection techniques, such as lasso, are particularly beneficial in analyses that look for independent signals, since each variable that is included in the final model, is conditioned on any other variables already in that model. While the highly correlated imputed data in this analysis resulted in lasso including some correlated variables in the final model, we were still able to identify putative secondary signals within the same locus, using further model selection techniques. These potentially novel secondary signals may contribute to the risk of childhood asthma, independently of the well-established primary signal within the 17q21 locus.

In the final chapter, a lasso gene-gene interaction analysis was carried out. This investigation set out to identify loci with marginal effects, as well as pair-wise interaction effects, that play a role in CAD aetiology. In a two-step lasso approach, we analysed all the possible pair-wise interactions between the 2000

marginal effects that were identified in the first model selection step of the lasso analysis. While many of the interactions identified by this method proved to be false positive results (where two loci were modelled as one significant interaction term, rather than two significant marginal effects), there were three plausible interactions identified by lasso. After further model selection methods were carried out on these three possible interactions, it was demonstrated that the addition of the interaction term significantly improved the fit of the main effects model and the p-value of the interaction term was far more significant than the p-value of the marginal effects making up that interaction.

The advantages of lasso as a model selection method were demonstrated throughout this thesis. Firstly, we were able to apply lasso to both quantitative traits and case-control studies, making it a flexible framework in which to study disease associations. Secondly, it is computationally efficient and able to handle the massive amount of data generated by genome-wide association studies. Thirdly, lasso is able to perform model selection on undetermined problems, where the number of predictors exceeds the number of individuals in a study, and finally, lasso has the potential to identify both pair-wise interactions and independent genetic effects within the same locus.

7.2 Future work and method development

Despite the advantages of lasso that were demonstrated throughout this investigation, it remains crucial that the results of the fine-mapping and interaction analyses are replicated in independent cohorts. This step not only serves to validate the potentially important findings laid out in Chapters 5 and 6,

but will further allow the accuracy of these methods to be fully evaluated.

There are a number of ways in which the methods outlined in this investigation could be improved and developed, in order to produce a more accurate and powerful approach to the analysis carried out in each chapter. Firstly, it is important to note that for the purpose of these studies, environmental factors were not included in the analyses. The original version of Mendel that was used was unable to model covariates and including these factors may have led to the identification of loci involved in gene-environment interactions, that would not have been identified when these loci were considered on their own. Indeed, environmental factors are established risk factors for common disease such as T2D, asthma and CAD and therefore the addition of these factors as covariates in any analysis, may increase the power to detect novel loci. Recent versions of Mendel that enable covariates to be included in the model, could therefore be used for subsequent analyses.

Secondly, since the inclusion of relevant prior information can be used to increase the power of association testing (Li *et al.* 2010), prioritizing certain variables in the same way may increase the power of lasso to select biologically important loci. Furthermore, for interaction analyses, the incorporation of prior information could enable studies to focus on SNPs in genes from the same pathway, that are more likely to interact. This prior information could be based on candidate gene studies, previous knowledge of the underlying genetic architecture of a trait, or imputation and genotyping accuracy. Of course, on the other hand, use of inappropriate priors may increase the level of bias introduced into a study, so it is important that large, highly powered studies are undertaken

in order to minimise the chances of this occurring. The incorporation of accurate prior information into model selection can be achieved by varying the degree to which different variables are penalized. SNPs with biologically plausible functions, for example, could be penalized less than other SNPs and are therefore more likely to enter the final lasso model.

Thirdly, the lasso interaction analysis outlined in Chapter 6 could greatly benefit from advances in technology and increased computational capacity. The interaction analysis in this study focussed on pair-wise interactions between the 2000 predictors that lasso selected to best explain the variation in the coronary artery disease phenotype. Increasing computational capacity to enable lasso to analyse all the possible pair-wise interactions between every SNP in the data set, could potentially identify novel interactions that may have been excluded from this analysis.

Finally, different methods of imputation may be more accurate than the methods employed in this study, and improving imputation accuracy may improve the power of these methods to detect common disease loci. Although different imputation software was used in each of the chapters, we consistently used the "best guess" genotype, whereby the unmeasured genotype with the highest posterior probability was used as the true genotype. According to Zheng *et* al. (ZHENG *et al.* 2011), this method does not account for uncertainty and only performs well when imputation accuracy is high. If imputation accuracy is low, however, this may decrease the power to detect disease associations. In contrast, regressing the phenotype on the estimated allelic dosage has been found to be a more accurate imputation method over a range of allele

frequencies, sample sizes and imputation accuracies (Zheng, 2011). Therefore, using this dosage method on data imputed from 1000 genomes, would maximize the amount of information in the available data, and could improve the power of the methods outlined in this study to detect novel disease loci.

7.3 Uncovering the missing heritability of common diseases

Despite the identification of many common disease risk variants in GWA studies, there is still a long way to go to fully understanding the mechanisms that contribute to disease phenotypes. Identifying statistical associations is merely the first step. In the study of complex diseases, many of the risk loci that have been discovered map to gene desserts and non-coding regions, and we know very little about their underlying function. Thus, some of the candidate genes that are positioned nearby GWAS risk-loci, that have been hypothesised to play a role in the disease aetiology, may not be the actual causative genes. Furthermore, even with the advances in genotyping technology and the rapid increase in the number of common genetic variants that have found to be associated with many human diseases, the effect sizes of these variants are often small and GWAS has failed to explain a large portion of the heritability of complex diseases.

Scientists have suggested that this “missing heritability” may be explained by rare variants that exist across the human genome. Since GWAS relies on the linkage disequilibrium between common genetic markers and common disease-risk loci, these studies are not able to detect rare variants that are only weakly correlated with the common markers (LI and LEAL 2008). In this regard,

sequencing technologies have emerged as a new method of detecting the rare variation that many believe to account for a large portion of missing disease heritability (CIRULLI and GOLDSTEIN 2010). It is thus evident, that statistical methods that are able to analyse both rare and common variants would be ideally suited to the next generation of sequencing data.

In addition to this, the huge amount of data that will be generated by sequencing technologies will challenge traditional single-SNP methods of analysis that are prone to multiple testing problems and consequent type 1 errors (HUTCHISON 2007). Model selection methods that are able to reduce the dimensionality of the data, such as lasso penalized regression, may prove to play a pivotal role in these types of analyses and therefore the next step would be to find ways in which lasso can be used to search through sequence data for rare variants.

Since rare disease variants may only be present in a few samples, methods of analysis that incorporate marginal effects only, will be hugely underpowered. To overcome this challenge, collapsing methods that search for an accumulation of minor alleles within a gene or pathway (LI and LEAL 2008; MORRIS and ZEGGINI 2010) have been developed. These methods aim to enrich the association signal while decreasing the overall degrees of freedom (i.e. testing for association with a combined effect across a gene versus testing multiple rare variants individually) (LI and LEAL 2008). However, signals with opposing directional effects in the same functional unit may cancel each other out and this may decrease power to detect these effects.

Lasso offers an alternative approach to rare variant analysis, by grouping

variants in a similar way, but in a penalized regression context. Rare variants are grouped together into biologically related units and then penalized as a group, rather than on an individual level. This circumvents the possibility of opposing signal directions cancelling each other out and allows group penalties for clusters of rare variants and individual penalties for common variants to be incorporated into a single analysis. Therefore, lasso may provide a valuable method for the simultaneous analysis of both rare and common variants (ZHOU *et al.* 2011).

In order to fully elucidate the role of genetic factors in complex human diseases, the statistical methods that are used to identify these factors need to be continually developed and refined in order to match the rapid improvements in genotyping technology and computational efficiency. At present, the identification of disease risk loci is still in its infancy and scientific efforts continue to further increase the power of existing GWA studies, as well as focussing on new rare variant and sequencing studies. Accurate and efficient statistical methods will be needed to capitalize on the vast wealth of information provided by these efforts. Therefore, the flexibility of lasso penalized regression within the context of GWA studies, its computational efficiency, and its potential application to sequencing data to search for rare variants, will be pivotal to future research into the genetics of common human disease.

References

WTCCC, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447:** 661-678.

ABECASIS, G. R., E. NOGUCHI, A. HEINZMANN, J. A. TRAHERNE, S. BHATTACHARYYA *et al.*, 2001 Extent and distribution of linkage disequilibrium in three genomic regions. American journal of human genetics **68:** 191-197.

ALTSHULER, D. M., R. A. GIBBS, L. PELTONEN, E. DERMITZAKIS, S. F. SCHAFFNER *et al.*, 2010 Integrating common and rare genetic variation in diverse human populations. Nature **467:** 52-58.

ARDLIE, K. G., L. KRUGLYAK and M. SEIELSTAD, 2002 Patterns of linkage disequilibrium in the human genome. Nature reviews. Genetics **3:** 299-309.

BAGLEY, S. C., H. WHITE and B. A. GOLOMB, 2001 Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. Journal of clinical epidemiology **54:** 979-985.

BALDING, D. J., 2006 A tutorial on statistical methods for population association studies. Nature reviews. Genetics **7:** 781-791.

BALLARD, D., C. ABRAHAM, J. CHO and H. ZHAO, 2010 Pathway analysis comparison using Crohn's disease genome wide association studies. BMC Med Genomics **3:** 25.

BARRETT, J. C., and L. R. CARDON, 2006 Evaluating coverage of genome-wide association studies. Nature genetics **38:** 659-662.

BERG, K., A. SVINDLAND, A. J. SMITH, R. M. LAWN, S. DJUROVIC *et al.*, 2002 Spontaneous atherosclerosis in the proximal aorta of LPA transgenic mice on a normal diet. Atherosclerosis **163:** 99-104.

BERK, K. N., 1980 Forward and backward stepping in variable selection. Journal of statistical computation and simulation **10:** 177-185.

BEWICK, V., L. CHEEK and J. BALL, 2005 Statistics review 14: Logistic regression. Critical care **9:** 112-118.

BLAND, J. M., and D. G. ALTMAN, 1995 Multiple significance tests: the Bonferroni method. BMJ **310:** 170.

BOUZIGON, E., E. CORDA, H. ASCHARD, M. H. DIZIER, A. BOLAND *et al.*, 2008 Effect of 17q21 variants and smoking exposure in early-onset asthma. The New England journal of medicine **359:** 1985-1994.

BURNHAM, K. P., and D. R. ANDERSON, 2002 *Model Selection and multimodel inference: A practical information- theoretic approach*. Springer, New York.

CANTOR, R. M., K. LANGE and J. S. SINSHEIMER, 2010 Prioritizing GWAS results: A review of statistical methods and recommendations for their application. American journal of human genetics **86:** 6-22.

CARDON, L. R., and J. I. BELL, 2001 Association study designs for complex diseases. Nature reviews. Genetics **2:** 91-99.

CHAKRAVARTI, A., 1999 Population genetics--making sense out of sequence. Nature genetics **21:** 56-60.

CHAPMAN, J. M., J. D. COOPER, J. A. TODD and D. G. CLAYTON, 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.

CHASMAN, D. I., G. PARE, R. Y. ZEE, A. N. PARKER, N. R. COOK *et al.*, 2008 Genetic loci associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein A1, and Apolipoprotein B among 6382 white women in genome-wide analysis with replication. Circulation. Cardiovascular genetics **1:** 21-30.

CIRULLI, E. T., and D. B. GOLDSTEIN, 2010 Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature reviews. Genetics **11:** 415-425.

CLARKE, R., J. F. PEDEN, J. C. HOPEWELL, T. KYRIAKOU, A. GOEL *et al.*, 2009 Genetic variants associated with Lp(a) lipoprotein level and coronary disease. The New England journal of medicine **361:** 2518-2528.

COFFEY, C. S., P. R. HEBERT, M. D. RITCHIE, H. M. KRUMHOLZ, J. M. GAZIANO *et al.*, 2004 An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. BMC bioinformatics **5:** 49.

COLLINS, F. S., L. D. BROOKS and A. CHAKRAVARTI, 1998 A DNA polymorphism discovery resource for research on human genetic variation. Genome research **8:** 1229-1231.

COOKSON, W., 1999 The alliance of genes and environment in asthma and allergy. Nature **402:** B5-11.

COOKSON, W., L. LIANG, G. ABECASIS, M. MOFFATT and M. LATHROP, 2009 Mapping complex disease traits with global gene expression. Nature reviews. Genetics **10:** 184-194.

CORDELL, H. J., 2002 Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Human molecular genetics **11:** 2463-2468.

CORDELL, H. J., 2009 Detecting gene-gene interactions that underlie human diseases. Nature reviews. Genetics **10:** 392-404.

CORDELL, H. J., and D. G. CLAYTON, 2005 Genetic association studies. Lancet **366:** 1121-1131.

DE BAKKER, P. I., R. R. GRAHAM, D. ALTSHULER, B. E. HENDERSON and C. A. HAIMAN, 2006 Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing**:** 478-486.

DEKKERS, J. C., and F. HOSPITAL, 2002 The use of molecular genetics in the improvement of agricultural populations. Nature reviews. Genetics **3:** 22-32.

DEVLIN, B., K. ROEDER and L. WASSERMAN, 2001 Genomic control, a new approach to genetic-based association studies. Theoretical population biology **60:** 155-166.

DICKSON, S. P., K. WANG, I. KRANTZ, H. HAKONARSON and D. B. GOLDSTEIN, 2010 Rare variants create synthetic genome-wide associations. PLoS biology **8:** e1000294.

EFRON, D. T., T. HASTIE and R. TIBSHIRANI, 2004 Least angle regression. Annals of Statistics **32:** 407-499.

ELKS, C. E., J. R. PERRY, P. SULEM, D. I. CHASMAN, N. FRANCESCHINI *et al.*, 2010 Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. Nature genetics **42:** 1077-1085.

EVANS, D. M., J. MARCHINI, A. P. MORRIS and L. R. CARDON, 2006 Two-stage two-locus models in genome-wide association. PLoS genetics **2:** e157.
FALCONER, D. S., 1989 *Introduction to quantitative genetics*. John Wiley & Sons, New York.
FAN, R., and M. KNAPP, 2003 Genome association studies of complex diseases by case-control designs. American journal of human genetics **72:** 850-868.
FIELD, A. P., 2001 Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. Psychological methods **6:** 161-180.
FORSYTHE, A. B., P. R. MAY and L. ENGELMAN, 1971 Prediction by multiple regression how many variables to enter? Journal of psychiatric research **8:** 119-126.
FOULKES, A. S., 2009 *Applied statistical genetics with R: for population based association studies*. Springer, New York.
FREEDMAN, M. L., D. REICH, K. L. PENNEY, G. J. MCDONALD, A. A. MIGNAULT *et al.*, 2004 Assessing the impact of population stratification on genetic association studies. Nature genetics **36:** 388-393.
FRIDLEY, B. L., 2009 Bayesian variable and model selection methods for genetic association studies. Genetic epidemiology **33:** 27-37.
FRIEDMAN, J., T. HASTIE and R. TIBSHIRANI, 2010 Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw **33:** 1-22.
GABRIEL, S. B., S. F. SCHAFFNER, H. NGUYEN, J. M. MOORE, J. ROY *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225-2229.
GAJEWSKI, B. J., S. D. SIMON and S. E. CARLSON, 2008 Predicting accrual in clinical trials with Bayesian posterior predictive distributions. Statistics in medicine **27:** 2328-2340.
GALANTER, J., S. CHOUDHRY, C. ENG, S. NAZARIO, J. R. RODRIGUEZ-SANTANA *et al.*, 2008 ORMDL3 gene is associated with asthma in three ethnically diverse populations. Am J Respir Crit Care Med **177:** 1194-1200.
GAUT, B. S., and A. D. LONG, 2003 The lowdown on linkage disequilibrium. The Plant cell **15:** 1502-1506.
GHOSH, J. K., M. DELAMPADY and T. SAMANTA, 2006 And introduction to Bayesian analysis: theory and methods, pp. Springer, New York.
GILLESPIE, J. H., 2004 *Population genetics: A concise guide.* Johns Hopkins University Press, Baltimore.
GOLDSTEIN, D. B., and M. E. WEALE, 2001 Population genomics: linkage disequilibrium holds the key. Current biology : CB **11:** R576-579.
HANSSON, G. K., 2005 Inflammation, atherosclerosis, and coronary artery disease. The New England journal of medicine **352:** 1685-1695.
HAO, K., 2007 Genome-wide selection of tag SNPs using multiple-marker correlation. Bioinformatics **23:** 3178-3184.
HARDY, J., and A. SINGLETON, 2009 Genomewide association studies and human disease. The New England journal of medicine **360:** 1759-1768.
HARRELL, F. E., 2001 *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis.* Springer, New York.
HARTL, D. L., and A. CLARK, 1997a *Principles of population genetics*. Sinauer Associates, Sutherland.
HARTL, D. L., and A. CLARK, 1997b *Principles of population genetics.* Sinauer Associates, Sunderland.
HARTZ, S. C., and L. A. ROSENBERG, 1975 The computation of maximum

likelihood estimation for the multiple logistic risk function for use with categorical data. Journal of chronic diseases **28:** 421-429.
HEDGES, L. B., and J. L. VEVEA, 1998 Fixed- and random-effects models in meta-analysis. Psychological methods **3:** 486-504.
HENDRICK, P. W., 2000a *Genetics of populations*. Bartlett Publishers, London.
HENDRICK, P. W., 2000b *Principles of population genetics*. Bartlett Publishers Inc., London.
HIGGINS, J. P., S. G. THOMPSON, J. J. DEEKS and D. G. ALTMAN, 2003 Measuring inconsistency in meta-analyses. BMJ **327:** 557-560.
HIRSCHHORN, J. N., and M. J. DALY, 2005 Genome-wide association studies for common diseases and complex traits. Nature reviews. Genetics **6:** 95-108.
HOETING, J., D. MADIGAN, A. E. RAFTERY and C. VOLINSKY, 1999 Bayesian model averaging. Statistical science. **14:** 382-410.
HOWIE, B. N., P. DONNELLY and J. MARCHINI, 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics **5:** e1000529.
HUTCHISON, C. A., 3RD, 2007 DNA sequencing: bench to bedside and beyond. Nucleic acids research **35:** 6227-6237.
KANG, G., K. YE, N. LIU, D. B. ALLISON and G. GAO, 2009 Weighted multiple hypothesis testing procedures. Statistical applications in genetics and molecular biology **8:** Article23.
KEATING, B. J., S. TISCHFIELD, S. S. MURRAY, T. BHANGALE, T. S. PRICE *et al.*, 2008 Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PloS one **3:** e3583.
KENT, T., 1982 Robust Properties of Likelihood Ratio Tests. Biometrika **69:** 19-27.
KHOT, U. N., M. B. KHOT, C. T. BAJZER, S. K. SAPP, E. M. OHMAN *et al.*, 2003 Prevalence of conventional risk factors in patients with coronary heart disease. JAMA : the journal of the American Medical Association **290:** 898-904.
KRUSZYNSKA, Y. T., M. A. GHATEI, S. R. BLOOM and N. MCINTYRE, 1995 Insulin secretion and plasma levels of glucose-dependent insulinotropic peptide and glucagon-like peptide 1 [7-36 amide] after oral glucose in cirrhosis. Hepatology **21:** 933-941.
LANDER, E. S., 1996 The new genomics: global views of biology. Science **274:** 536-539.
LANDER, E. S., 2011 Initial impact of the sequencing of the human genome. Nature **470:** 187-197.
LAWRENCE, R. W., D. M. EVANS and L. R. CARDON, 2005 Prospects and pitfalls in whole genome association studies. Philosophical transactions of the Royal Society of London. Series B, Biological sciences **360:** 1589-1595.
LEE, S., F. A. WRIGHT and F. ZOU, 2010 Control of Population Stratification by Correlation-Selected Principal Components. Biometrics.
LENG, C., Y. LIN and G. WAHBA, 2006 A note on lasso and related procedures in model selection. Statistica Sinica **16:** 1273-1284.
LEWIS, C. M., 2002 Genetic association studies: design, analysis and interpretation. Briefings in bioinformatics **3:** 146-153.
LEWONTIN, R. C., 1988 On measures of gametic disequilibrium. Genetics **120:** 849-852.
LEWONTIN, R. C., 1995 The detection of linkage disequilibrium in molecular

sequence data. Genetics **140:** 377-388.
LI, B., and S. M. LEAL, 2008 Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American journal of human genetics **83:** 311-321.
LI, H., Z. WEI and J. MARIS, 2010 A hidden Markov random field model for genome-wide association studies. Biostatistics **11:** 139-150.
LIU, X. G., Y. J. LIU, J. LIU, Y. PEI, D. H. XIONG *et al.*, 2008 A bivariate whole genome linkage study identified genomic regions influencing both BMD and bone structure. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research **23:** 1806-1814.
LLOYD-JONES, D. M., B. H. NAM, R. B. D'AGOSTINO, SR., D. LEVY, J. M. MURABITO *et al.*, 2004 Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. JAMA : the journal of the American Medical Association **291:** 2204-2211.
MADIGAN, D., and G. RIDGEWAY, 2004 Discussion of least angle regression. Annals of Statistics **32:** 465-469.
MANOLIO, T. A., F. S. COLLINS, N. J. COX, D. B. GOLDSTEIN, L. A. HINDORFF *et al.*, 2009 Finding the missing heritability of complex diseases. Nature **461:** 747-753.
MARCHINI, J., P. DONNELLY and L. R. CARDON, 2005 Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics **37:** 413-417.
MARCHINI, J., and B. HOWIE, 2010 Genotype imputation for genome-wide association studies. Nature reviews. Genetics **11:** 499-511.
MARCHINI, J., B. HOWIE, S. MYERS, G. MCVEAN and P. DONNELLY, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics **39:** 906-913.
MARENBERG, M. E., N. RISCH, L. F. BERKMAN, B. FLODERUS and U. DE FAIRE, 1994 Genetic susceptibility to death from coronary heart disease in a study of twins. The New England journal of medicine **330:** 1041-1046.
MCCLELLAN, J., and M. C. KING, 2010 Genetic heterogeneity in human disease. Cell **141:** 210-217.
MEHTA, N. N., M. LI, D. WILLIAM, A. V. KHERA, S. DEROHANNESSIAN *et al.*, 2011 The novel atherosclerosis locus at 10q11 regulates plasma CXCL12 levels. European heart journal **32:** 963-971.
MICHEL, S., L. LIANG, M. DEPNER, N. KLOPP, A. RUETHER *et al.*, 2010 Unifying candidate gene and GWAS Approaches in Asthma. PloS one **5:** e13894.
MIYAGAWA, T., N. NISHIDA, J. OHASHI, R. KIMURA, A. FUJIMOTO *et al.*, 2008 Appropriate data cleaning methods for genome-wide association study. Journal of human genetics **53:** 886-893.
MOFFATT, M. F., I. G. GUT, F. DEMENAIS, D. P. STRACHAN, E. BOUZIGON *et al.*, 2010 A large-scale, consortium-based genomewide association study of asthma. The New England journal of medicine **363:** 1211-1221.
MOFFATT, M. F., M. KABESCH, L. LIANG, A. L. DIXON, D. STRACHAN *et al.*, 2007 Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature **448:** 470-473.
MONTGOMERY, J. M., 2010 Bayesian Model Averaging: Theoretical developments and practical applications. Political Analysis **18:** 245-270.
MOORE, J. H., 2003 The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human heredity **56:** 73-82.

MORRIS, A. P., 2006 A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. American journal of human genetics **79:** 679-694.

MORRIS, A. P., and E. ZEGGINI, 2010 An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology **34:** 188-193.

MOTULSKY, A. G., 2006 Genetics of complex diseases. Journal of Zhejiang University. Science. B **7:** 167-168.

MURABITO, J. M., M. J. PENCINA, B. H. NAM, R. B. D'AGOSTINO, SR., T. J. WANG *et al.*, 2005 Sibling cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults. JAMA : the journal of the American Medical Association **294:** 3117-3123.

MURRAY, C. J., and A. D. LOPEZ, 1997 Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. Lancet **349:** 1436-1442.

NAKAOKA, H., and I. INOUE, 2009 Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse. Journal of human genetics **54:** 615-623.

NORDBORG, M., and S. TAVARE, 2002 Linkage disequilibrium: what history has to tell us. Trends in genetics : TIG **18:** 83-90.

OBER, C., A. S. NORD, E. E. THOMPSON, L. PAN, Z. TAN *et al.*, 2009 Genome-wide association study of plasma lipoprotein(a) levels identifies multiple genes on chromosome 6q. J Lipid Res **50:** 798-806.

OGAWA, N., Y. IMAI, H. MORITA and R. NAGAI, 2010 Genome-wide association study of coronary artery disease. International journal of hypertension **2010:** 790539.

OHASHI, J., and K. TOKUNAGA, 1999 Selecting a contingency table in a population-based association study: allele frequency or positivity? Journal of human genetics **44:** 246-248.

OHTA, T., 1982 Linkage disequilibrium with the island model. Genetics **101:** 139-155.

PENG, B., and M. KIMMEL, 2007 Simulations provide support for the common disease-common variant hypothesis. Genetics **175:** 763-776.

POSADA, D., and T. R. BUCKLEY, 2004 Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Systematic biology **53:** 793-808.

PRITCHARD, J. K., and N. J. COX, 2002 The allelic architecture of human disease genes: common disease-common variant...or not? Human molecular genetics **11:** 2417-2423.

RAFTERY, A. E., 1995 Bayesian model averaging in social research. Sociological Methodology **25:** 111-196.

RAFTERY, A. E., D. MADIGAN and J. HOETING, 1997 Bayesian model averaging for linear regression models. Journal of American Statistical Association **92:** 179-191.

REICH, D., A. L. PRICE and N. PATTERSON, 2008 Principal component analysis of genetic data. Nature genetics **40:** 491-492.

RHEE, E. J., E. J. YUN, K. W. OH, S. E. PARK, C. Y. PARK *et al.*, 2010 The relationship between Receptor Activator of Nuclear Factor-kappaB Ligand (RANKL) gene polymorphism and aortic calcification in Korean women. Endocr J **57:** 541-549.

ROBINSON, R., 2010 Common disease, multiple rare (and distant) variants. PLoS biology **8:** e1000293.

RONALD, J., R. RAJAGOPALAN, J. E. RANCHALIS, J. K. MARSHALL, T. S. HATSUKAMI *et al.*, 2009 Analysis of recently identified dyslipidemia alleles reveals two loci that contribute to risk for carotid artery disease. Lipids Health Dis **8:** 52.

SAMANI, N. J., J. ERDMANN, A. S. HALL, C. HENGSTENBERG, M. MANGINO *et al.*, 2007 Genomewide association analysis of coronary artery disease. The New England journal of medicine **357:** 443-453.

SCHMIDT, F. L., I. S. OH and T. L. HAYES, 2009 Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. The British journal of mathematical and statistical psychology **62:** 97-128.

SCHORK, N. J., S. S. MURRAY, K. A. FRAZER and E. J. TOPOL, 2009 Common vs. rare allele hypotheses for complex diseases. Current opinion in genetics & development **19:** 212-219.

SCHUNKERT, H., I. R. KONIG, S. KATHIRESAN, M. P. REILLY, T. L. ASSIMES *et al.*, 2011 Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature genetics.

SERVIN, B., and M. STEPHENS, 2007 Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS genetics **3:** e114.

SHOEMAKER, J. S., I. S. PAINTER and B. S. WEIR, 1999 Bayesian statistics in genetics: a guide for the uninitiated. Trends in genetics : TIG **15:** 354-358.

STRACHAN, T., and A. READ, 2010 *Human Molecular Genetics*. Garland Science Publishing, New York.

SUNG, Y. J., T. K. RICE, G. SHI, C. C. GU and D. RAO, 2009 Comparison between single-marker analysis using Merlin and multi-marker analysis using LASSO for Framingham simulated data. BMC proceedings **3 Suppl 7:** S27.

TAVENDALE, R., D. F. MACGREGOR, S. MUKHOPADHYAY and C. N. PALMER, 2008 A polymorphism controlling ORMDL3 expression is associated with asthma that is poorly controlled by current medications. The Journal of allergy and clinical immunology **121:** 860-863.

TAYLOR, A., 2006 The genetics of type 2 diabetes: A review. International journal of Diabetes and Metabolism **14:** 76-81.

TERWILLIGER, J. D., and K. M. WEISS, 1998 Linkage disequilibrium mapping of complex disease: fantasy or reality? Current opinion in biotechnology **9:** 578-594.

TIBSHIRANI, R., 1996 Regression shrinkage and selection via the lasso. Journal of the royal statistical society **58:** 267-288.

TIBSHIRANI, R., 1997 The lasso method for variable selection in the Cox model. Statistics in medicine **16:** 385-395.

VANLIERE, J. M., and N. A. ROSENBERG, 2008 Mathematical properties of the r2 measure of linkage disequilibrium. Theoretical population biology **74:** 130-137.

VIA, M., C. GIGNOUX and E. G. BURCHARD, 2010 The 1000 Genomes Project: new opportunities for research and social challenges. Genome medicine **2:** 3.

VIALLEFONT, V., A. E. RAFTERY and S. RICHARDSON, 2001 Variable selection and Bayesian model averaging in case-control studies. Statistics in medicine **20:** 3215-3230.

VISEL, A., Y. ZHU, D. MAY, V. AFZAL, E. GONG *et al.*, 2010 Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Nature **464:** 409-412.

VISSCHER, P. M., 2008 Sizing up human height variation. Nature genetics **40:**

489-490.
VOIGHT, B. F., L. J. SCOTT, V. STEINTHORSDOTTIR, A. P. MORRIS, C. DINA *et al.*, 2010 Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nature genetics **42:** 579-589.
WEISSGLAS-VOLKOV, D., A. HUERTAS-VAZQUEZ, E. SUVIOLAHTI, J. LEE, C. PLAISIER *et al.*, 2006 Common hepatic nuclear factor-4alpha variants are associated with high serum lipid levels and the metabolic syndrome. Diabetes **55:** 1970-1977.
WERTHEIM, J. O., M. J. SANDERSON, M. WOROBEY and A. BJORK, 2010 Relaxed molecular clocks, the bias-variance trade-off, and the quality of phylogenetic inference. Systematic biology **59:** 1-8.
WILLER, C. J., Y. LI and G. R. ABECASIS, 2010 METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics **26:** 2190-2191.
WU, H., I. ROMIEU, J. J. SIENRA-MONGE, H. LI, B. E. DEL RIO-NAVARRO *et al.*, 2009a Genetic variation in ORM1-like 3 (ORMDL3) and gasdermin-like (GSDML) and childhood asthma. Allergy **64:** 629-635.
WU, T. T., Y. F. CHEN, T. HASTIE, E. SOBEL and K. LANGE, 2009b Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics **25:** 714-721.
ZEGGINI, E., and J. P. IOANNIDIS, 2009 Meta-analysis in genome-wide association studies. Pharmacogenomics **10:** 191-201.
ZHAO, P., and B. YU, 2006 On model selection consistency of Lasso. Journal of machine learning research **7:** 2541-2563.
ZHENG, J., Y. LI, G. R. ABECASIS and P. SCHEET, 2011 A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genetic epidemiology **35:** 102-110.
ZHOU, H., D. H. ALEXANDER, M. E. SEHL, J. S. SINSHEIMER, E. M. SOBEL *et al.*, 2011 Penalized regression for genome-wide association screening of sequence data. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing**:** 106-117.
ZHU, X., S. LI, R. S. COOPER and R. C. ELSTON, 2008 A unified association analysis approach for family and unrelated samples correcting for stratification. American journal of human genetics **82:** 352-365.
ZONDERVAN, K. T., and L. R. CARDON, 2004 The complex interplay among factors that influence allelic association. Nature reviews. Genetics **5:** 89-100.