Influence of genomic architecture on the performance of association mapping

Application to ascites syndrome in broiler chickens

JOSEPH E POWELL

ABSTRACT

The availability of high-density genome panels, comprised of SNP data, for the majority of livestock species, has led to considerable growth in the development and investigation of association mapping methodology in recent years. Locating markers that are linked to loci affecting genetic variation is thought to be a promising approach to identifying genetic control for traits of interest. Whilst initial findings were promising, genome-wide association studies have struggled to identify markers explaining high proportions of genetic variation for the majority of complex traits analysed. This thesis has two main objectives that are complementary to one another. The first is to identify QTL associated with susceptibility to ascites syndrome in populations of broiler chickens. The second consists of the investigation and development of alternative association mapping models and a comprehensive evaluation of the influence of genetic architecture on the performance of these models.

Ascites is a metabolic disorder characterised by the accumulation of fluid within the peritoneum and is one of the most common disorders affecting broilers raised in commercial conditions. The thesis begins with a QTL analysis aimed at identifying genetic loci that influence susceptibility to ascites. The constraints and limitations of the methodology are discussed, and alternative methods investigated using real and simulated datasets. Amongst the conclusions, is the identification that model performance is strongly influenced by localised genetic architecture of the causal variants and markers used by the models. Identifying how this architecture influences the performance of models is of considerable importance given its highly variable nature across the genome of all species.

A broad range of genetic conditions were simulated, based on real data, to identify their influence on the ability to provide statistical significance for causal loci, using a range of regression-based mapping models that differ in their use of marker information and parameterization. This work was extended to determine the influence of marker numbers, in multilocus models, on performance under different genomic conditions. Conclusions drawn from this work are applied in the re-analysis for QTL influencing ascites, also identifying the QTL shown in the previous study, as well as additional loci. Finally, the thesis provides a perspective on the future research directions regarding utilisation of observed genetic information on model choice, and an outline of how whole-genome mapping studies can be constructed, in order to maximise the use of information present in the genome panel.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my thanks to my principle supervisor, Chris Haley, for his thoughtful guidance and support over the last three years. I am very grateful that I was allowed the freedom and academic flexibility to explore different aspects of research in quantitative genetics of complex traits throughout the duration of the studentship. I also thank my secondary supervisors at the Roslin Institute, Ian Dunn and Dirk-Jan de Koning, and at the University of Edinburgh, Sara Knott, for their helpful advice throughout the PhD. Many thanks also to Sue Brotherstone, Ian White and Bill Hill at the University of Edinburgh for all their help during my time in Edinburgh. I would also like to send my gratitude to Andreas Kranis, Santiago Avendano and Kellie Watson, for providing support and helping me maintain links with my industrial sponsor, Aviagen Ltd. I also acknowledge financial support from the Biotechnology and Biological Science Research Council, the Genesis Faraday Partnership and Aviagen Ltd.

The Roslin Institute has been a thoroughly enjoyable place to work over the last three years, and that is in no small part due to its members of staff and students. In particular in would like to the following from the 'North wing' for their advice and help over the last few years; Dave Waddington, Suzanne Rowe, Ricardo Pong-Wong, Dave Telford, Caroline Channing, Barry Keenan, Marie-Anne Robertson, Carol Telford, Huw Jones, Chris Warkup, Jamie Floyd, Hans Daetwyler, Steve Bishop, Alex Lam, Ross Houston, Gib Hemani, Wenhua Wei, John Woolliams and Pam Wiener. Thanks also to Liz Archibald, Liz Brown and the HR team for all their help.

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

List of Publications

Refereed:

Lam, A. C., **Powell, J. E.**, Wei, W. H., De Koning, D. J. and Haley, C. S. (2009) A combined strategy for quantitative trait loci detection by genome-wide association analysis. *BMC Proceedings*, **3**, S6.

Powell, J. E., Kranis, A., Dekkers, J. C. M., Knott, S. A. and Haley, C. S. Optimal length of marker windows in multilocus association mapping models. *Submitted: PLoS Genetics*

Powell, J. E., Kranis, A., Floyd, J., Dekkers, J. C. M., Knott, S. A. and Haley, C. S. Influence of localized genomic architecture on the optimal performance of models for analysis of association data. *Submitted: Genetics*

Conference abstracts:

Powell, J. E., Lam, A., Wei, W. H., De Koning, D. J. and Haley, C. S. (2008) A robust strategy for quantitative trait loci detection by genome-wide association. *12th Quantitative Trait Loci and Marker Assisted Selection Workshop*, Uppsala, Sweden.

Powell, J. E., Kranis, A., Avendano, S. and Haley, C. S. (2008) Methods for genome-wide association analysis for ascites susceptability in broiler chickens. *23rd World Poultry Congress*, Brisbane, Australia.

Powell, J. E., Kranis, A. and Haley, C. S. (2009) Factors influencing the optimal multilocus model for analysis of association data. *13th Quantitative Trait Loci and Marker Assisted Selection Workshop*, Wageningen, The Netherlands.

Powell, J. E., Kranis, A. and Haley, C. S. (2009) Influence of genomic architecture on the performance of association mapping models. *EADGENE: Genomics for Animal Health*, Paris, France.

TABLE OF CONTENTS

CHAPTER THREE

Investigation and comparison of multilocus methods for genome-wide association analyses

CHAPTER FIVE

Optimal length of marker windows in multilocus association mapping models

CHAPTER SIX

False positive rates at different heritabilities for single and multilocus models

CHAPTER SEVEN

Mapping for causal variants associated with ascites susceptibility in broiler chickens

LIST OF FIGURES

CHAPTER TWO

CHAPTER THREE

CHAPTER FOUR

CHAPTER FIVE

CHAPTER SIX

CHAPTER SEVEN

LIST OF TABLES

CHAPTER TWO

CHAPTER THREE

CHAPTER SIX

CHAPTER SEVEN

List of abbreviations

LD	linkage disequilibrium
GWAS	genome-wide association analysis
DF	degrees of freedom
cM	centimorgan
kb	kilobase
SNP	single nucleotide polymorphism
QTL	quantitative trait loci
MAS	marker assisted selection
SaO_2	blood oxygen saturation
MAF	minor allele frequency
FCR	feed conversion rate
EBV	estimated breeding value
Mb	megabase
EM	expectation maximisation (algorithm)
HWE	Hardy Weinberg equilibrium
GLM	general linear model
IBD	identical by descent
sQTL	surrogate quantitative trait loci
HST	haplotype score test
SE	standard error

CHAPTER ONE

GENERAL INTRODUCTION

1.1 Association mapping

In recent years, a wealth of genomic information has become available for the majority of livestock and domestic species (e.g. Lindblad-Toh *et al.* 2005; Wong *et al.* 2004) and humans (e.g. Eichler *et al.* 2007; Kidd *et al.* 2008). These resources are indispensible tools for understanding genetic variation across genomes, and its control of phenotypic traits and susceptibility to diseases. Currently, genome-wide association studies (GWAS) are the most widely used contemporary approach used to relate genetic variation to phenotypic diversity (McCarthy *et al.* 2008). Broadly, the standard approach for GWAS studies is to use statistical tools to identify associations between the alleles of genetic markers and the measured phenotypes of a trait. In association studies, LD information is used to identify relationships between these markers and causal loci.

LD refers to the non-random association of alleles at different loci, and for association studies is considered at the population wide level. For example, suppose that allele A at locus 1 and allele B at locus 2 are at frequencies π_A and π_B, respectively, in the population. If the two loci are independent, then we would expect to see the AB haplotype at frequency $\pi_A\pi_B$. If the population frequency of the AB haplotype is either higher or lower than this, implying particular alleles tend to be observed together, then the two loci are said to be in LD. LD can be considered as a way of quantifying the level of informativness between different markers and is an important statistical property in association mapping studies. The extent of LD in a population depends on, amongst other factors, the past effective population size (Haley 1999). Due to domestication and breeding strategies, effective population sizes are small for the majority of livestock species (Hayes *et al.* 2003). As a result, the extent of LD can be considerable, with useful LD (e.g. useful for the detection of QTL) for up to tens of cM in cattle (Farnir *et al.* 2000), pigs (Nsengimana *et al.* 2004), sheep (McRae *et al.* 2002) and chickens (Andreescu *et al.* 2007). This is in sharp contrast to the situation in humans, where recent effective population sizes have been large, and consequently LD extends over much shorter distances, often only a few 100 kb (Pritchard and Przeworski, 2001; Reich *et al.* 2001). This has led to the suggestion that required marker density for mapping can be considerably lower for livestock species (Haley 1999; Hayes *et al.* 2003).

1.2 Association mapping methodology

Nowadays, genetic information available for whole genome mapping studies is typically high density SNP panels. These datasets contain genotypes of SNP markers that are spaced across the genome of the species of interest. The statistical analysis of this information works on the assumption that high levels of LD exists between markers and causal loci that are physically close to one another. Differences in the methodology of association analyses are normally concerned with the statistical procedures applied to information contained by the marker genotypes.

The diversity of statistical approaches used to infer the relationship between marker and trait information is too great a topic to cover in detail here. Instead, I will give a brief introduction to some of the alternative ways to use marker data. In the framework of mixed model machinery advocated by Henderson (1975) the matrix of predictor variables can be populated with information from either a single marker, or a group of markers, leading to the terms single-locus verses multilocus analyses for mapping studies. For single-locus models, if we assume an additive mode of inheritance, the number of copies of "A" alleles are normally fitted as linear covariates, and tested against the null hypothesis in a one DF test. For multilocus models there is a choice in the parameterization of the model, depending on how the information contained between the markers is used. Fitting *N* SNPs in a regression analysis, where coefficients are constrained to fit just main effects is a natural extension to a single-marker test. In the regression model there is now a coefficient $\beta_1, \beta_2 ... \beta_N$ for each SNP, leading to a general test with *N* numerator DF fitted. Using mixed model methodology allows additional covariates, such as sex, age, or environmental conditions to be easily included. With multilocus approaches, the parameterization normally depends on whether genotype or haplotype information is used. Genotype information of markers is observed, however, we are often ignorant about the phase of gametes. That is to say, we do not know which nucleotides reside together on individual chromosomes.

In recent years, the availability of large numbers of markers, and projects such as HapMap in humans (HapMap consortium 2007), has raised interest in using haplotype information in association studies. Haplotype phase is usually inferred using statistical procedures (Browning and Browning 2009; Excoffier and Slatkin 1995; Stephens *et al.* 2001), unless it can be determined using rule based systems from multigenerational family genotype information. Strategies for performing haplotype analyses are still the subject of active debate and research, and are well reviewed in the literature (e.g. Balding 2006; Schaid 2004). A popular strategy, based on the block like structure of haplotype diversity originally observed in the human genome, is to use haplotypes to try and capture the correlation structure of SNPs in regions of low recombination (Gabriel *et al.* 2002). Haplotype blocks are defined as discrete chromosome regions containing SNPs in high LD, and consequently a low diversity of haplotypes (Cardon and Abecasis 2003). However, difficulties in defining LD blocks, the boundaries between them, and choices regarding the inclusion of orphan SNPs, has led to the suggestion that using LD blocks as units for association may not be the most efficient strategy for haplotype analyses (Ding *et al.* 2005; Zhao *et al.* 2003). Recently, the development of methods to evaluate the association of haplotypes with traits has focused on the use of sliding windows, whereby windows of haplotypes from adjacent markers are fitted in a sequential manner, and scanned across a genome panel.

A couple of problems are commonly encountered when using a sliding window approach. The first is how many adjacent SNPs should be included simultaneously in a particular haplotype analysis. This becomes especially apparent when we consider that the potential total number of haplotypes observed is 2^n, where n is the number of makers. At its time of origin, a causal mutation will occur on a single haplotype background, with this haplotype extending over the whole chromosome. As recombination events are accumulated across generations, the length of the original haplotype associated with the mutation shrinks. Thus, long haplotypes may often include alleles not associated with the haplotype background of the causal mutation, especially in the case of old mutations. Moreover, as some haplotypes would only be carried by a few individuals, there could be statistical problems, owing to small sample sizes, in evaluating their effects on the trait (North *et al.* 2006). The second problem occurs when there are large numbers of haplotypes, especially rare ones. This can lead to over parameterisation of the model, and difficulties in accurately estimating effects for haplotypes with low frequencies (Fallin and Schork 2000). A commonly used set of approaches to avoid this problem is to cluster haplotypes, based on some form of similarity, in the hope that clusters will reflect a shared ancestry. Thus, the parameter space can be reduced while, it is hoped, retaining phase information relevant to the causal variant. The concept is that dependency amongst haplotypes will be accounted for, whilst reducing the DF of the test and improving power. Alternatively, statistical tools can be used to identify haplotypes that have an effect on the trait, and shrink or reduce the effects of those that do not. Certain parameter reduction approaches are applied specifically to rare haplotypes in an effort to avoid problems associated with inaccuracies in estimating their effects. To date there has been no full evaluation of the ability of parameter reduction methods to correctly identify between haplotypes with effect, and those without.

A challenge of association mapping is how to identify and optimally parameterize the statistical model, in order to explain the greatest amount of trait variance using the fewest parameters. There has been considerable debate on the relative advantages of single-marker verses haplotypes analyses, and whilst there are a few consensus opinions, this is still an area of active research. For indirect association mapping, haplotype based methods may be more powerful than single-locus tests, as haplotypes may be able to capture ancestral structure between markers and causal variants (Akey *et al.* 2001). However, single-marker tests are expected to outperform haplotype based analyses under scenarios, such as the causal locus being genotyped directly (Zhang *et al.* 2002). The literature on the relative efficiency of analysing haplotypes verses single-markers is complicated by differing assumptions about the number of trait loci, the extent of LD between markers and trait loci and the use of specific simulated datasets. Most reports have compared the maximum single-locus statistic, with bonferroni correction for multiple tests, to a global test of haplotype associations (Schaid 2004). The performance of a model, relative to another, will be dependent on the localised genomic architecture of the markers and trait, and how this interacts with the parameterization of the model. Therefore, given the variation in genomic conditions across the genome, no one model is expected to perform optimally under all conditions. The difficulty is in identifying how the interaction between parameterization and genomic conditions affects model performance.

1.3 QTL mapping in chicken populations

The chicken is an ideal species for QTL mapping studies because of its high reproductive capacity, low generation intervals, breeding structure, and management conditions. Additionally, the chicken is almost unique amongst agricultural species in that a number of highly divergent selection lines are available (Rabie 2004). In the last few years, QTL mapping studies in chickens have identified genetic loci for a wide variety of economically important traits (Abasht *et al.* 2006; Hocking 2005). Identified markers have been used to focus on genomic regions for fine scale mapping, or if they have a proven association with a trait, in MAS breeding programmes (Dekkers 2004). Experimental crosses, such as between breeds, or lines, that are divergent for the trait of interest, are used to increase the probability that the F_2 parents are heterozygous for QTL, thereby increasing power of the experimental design. Whilst this approach has been successful in identifying QTL that explain differences between lines, it provides no insight into whether these QTL are segregating within current commercial lines. For a particular line, it is likely that any QTL associated with a trait under selection, will be fixed for the major alleles, unless there are other mechanisms that maintain variation at these loci (De Koning *et al.* 2004). Consequently, extreme crosses, such as a broiler-layer cross, are analysed under the assumption that the two lines are fixed for alternative alleles at the QTL (Haley *et al.* 1994). Alternatively, analysis within breeds will shift the emphasis towards finding genes that explain differences within a population, offering the opportunity of MAS within the breed (Van Arendonk and Bovenhuis 2003).

1.4 Introduction to the disorder ascites

The metabolic disorder ascites is associated with the accumulation of fluid within the peritoneum, and is one of the most significant metabolic conditions affecting broiler chickens. The disorder is strongly linked to pulmonary hypertension and sudden death syndrome, with these terms often being used interchangeably in the literature (Julian 1998; 2000; Odom 1993; Wideman 1988; 2000). When raised in optimal environmental conditions, incidence of ascites is typically low. However, when broilers are raised in tougher commercial environments, ascites can be more common, leading to considerable economic loses, as mortality usually occurs when birds are close to market weight (Hunton 1998; Maxwell and Robertson 1997). Ascites was originally observed in flocks reared at high altitudes (Maxwell *et al.* 1986; Smith *et al.* 1955), where the partial pressure of oxygen is low, leading to right ventricular hypertrophy and the accumulation of fluid in the abdominal cavity (Julian 1998). Temperature is also well known to have an effect on ascites, with the strong correlation between cold temperature and ascites well documented (Bendheim *et al.* 1992; May and Deaton 1974; Wideman and Tackett 2000). Stolz *et al.* (1992) demonstrated that cold temperatures increase incidences of ascites by raising the metabolic oxygen requirements of the birds, thereby increasing incidences of pulmonary hypertension leading to the development of ascites. Management factors such as feed type, air quality, ventilation, and incubator conditions have all been implicated in the development of ascites (Bendheim *et al.* 1992).

Pulmonary hypertension accounts for the majority of ascites cases in broilers, yet hypertension can originate from numerous causes. The events leading to pulmonary hypertension, resulting in ventricular failure and build up of fluid in the pericardium and abdominal cavity are well understood and have been described in great detail (Decuypere *et al.* 2000; Julian 1990a; 1990b; 2000; Lister 1997; Scheele *et al.* 1991; Wideman 2000). Both the avian respiratory and circulatory systems are important in the susceptibility of broilers to ascites; unlike mammals, avian lungs are relatively rigid, having limited movement during breathing, with air only passing through them on the way to air sacs (Julian 1993). In susceptible birds, an increase in workload to the heart can result in right ventricular failure and consequently ascites. Although a variety of different environmental, disease, diet, and management practices are known to trigger the condition, it is still observed in flocks where these factors are at near optimal levels.

1.5 Genetic control of ascites and SaO_2 as an indicator trait

It is generally accepted that ascites is genetically linked to productions traits (such as growth and breast muscle yield), due to the observation that incidences of ascites have increased together with selection for these traits (Havenstein *et al.* 1994). It has been suggested that ascites in modern broilers, reared in commercial conditions, is related to the high oxygen requirements of rapid growth, and the inability of the heart and lung to deliver sufficient oxygen to respiring tissue (Julian 1993). Continuous selection for body weight and muscle yields, at increasingly younger ages, has affected both growth curves and differential growth of organs in broilers (Dunnington and Siegal 1996). Recognition of physiological constraints on birds led to the implementation of good management practices to reduce oxygen demands, resulting in a significant reduction in ascites mortality in commercial flocks in recent years (Baghbanzadeh and Decuypere 2008; Balog 2003; Julian 2000). However, such approaches are not ideal, as they compromise the efficiency of broiler production. Whilst breeding companies are able to successfully improve growth rates, its full genetic potential is limited in order to avoid mortality from ascites (Druyan *et al.* 2007a).

Genetic variation of ascites has been estimated in several studies, with a range of heritabilities from 0.1 to 0.7 (De Greef *et al.* 2001; Druyan *et al.* 2007a; Lubritz *et al.* 1995; Moghaddam *et al.* 2001; Navarro 2003). These studies indicate the feasibility of selecting against ascites, although this is only possible through the identification of genetically susceptible birds based on phenotypic observations (Druyan *et al.* 2007b). Incidence of ascites is difficult to measure in commercial flocks as an autopsy is required for diagnosis. Furthermore, visible signs of the disorder are not normally seen until the bird is approaching market weight, causing difficulties in selection of birds within breeding programs. Therefore, an indicator trait is required that is highly correlated with the incidence of ascites and is cheap and easy to measure in a non-invasive way. Amongst the candidate indicator traits, blood oxygen saturation (SaO_2) is commonly used, as it can be measured easily and non-invasively using an oximeter – a device that uses spectrophotometer to measure the percentage of haemoglobin saturated with oxygen at the time of measurement. Several studies have demonstrated the link between SaO_2 and ascites, with heritabilities between 0.53-0.63 estimated for SaO_2, and a genetic correlation of -0.5 with ascites susceptibility (Julian and Mirsalimi 1992; Druyan *et al.* 1999). A negative genetic correlation is preferable, as selection resulting in a reduction in ascites would be expected to increase SaO_2 levels, improving its use as an indicator trait.

1.6 Thesis outline

The main objectives of this thesis are to explore methodology used in studies to identify genetic loci for traits, and investigate how performance of alternative models can be affected by genomic architecture observed in genomes. Strategies and tools are compared and developed, and utilised to identify QTL affecting susceptibility to ascites in broiler chicken lines. The following paragraphs outline the contents and main objectives for each of the subsequent chapters.

Chapter 2 details a genome-wide association analysis for QTL associated with ascites susceptibility using six lines of broiler chickens. This initial mapping study uses a standard single-locus mixed model method, to provide a preliminary analysis of QTLs in these populations. Data supplied here is from sire lines used in commercial breeding programs, and is provided by Aviagen Ltd. Relationships between the lines are assessed, and attempts are made to fit joint line analyses where appropriate. The LD properties of the populations are also explored and described here. The chapter concludes with an assessment of the experimental design and challenges facing QTL mapping studies in livestock species.

Chapter 3 investigates and contrasts alternative association analysis approaches using two datasets. One of these datasets is a single line of broilers, whilst the other is a simulated dataset provided as part of a QTL-MAS workshop. Alternative approaches considered here, include two multilocus models that parameterize the information between markers in different ways. The first fits just main effects, whilst the second includes all observed interactions in a haplotype analysis and uses score statistics to evaluate the strength of association with the trait. Both these methods are contrasted against a single-locus approach. A critical assessment of the uses of models is made, along with an evaluation of the likely importance of genomic architecture to model performance.

Chapter 4 investigates how the performance in identifying causal variants, of a range of regression based models, is affected by differences in genetic conditions such as allele frequencies, LD patterns and distance between markers and QTL. Optimal performance of a model is expected to be influenced by the interaction between its parameterization and the genomic information that it is supplied. Thus, models that differ in their parameterization are expected to handle genomic information in different ways. A comprehensive set of conditions is simulated by using the genotype panel from a broiler line dataset, and selecting markers to represent surrogate QTL, and using surrounding markers as predictors in a range of models. Models investigated include single-locus, main effect multilocus and two haplotype-based approaches. All multilocus models are implemented using a three-marker sliding window framework.

Chapter 5 builds on results shown in chapter four, and aims to determine the optimal window length for multilocus approaches under a wide range of genomic conditions. Here I focus on the importance of window length for multilocus models, and how their performance is affected by localised genetic architecture between markers in the windows and QTL. The same approach of using surrogate QTL is utilised here. In this chapter I identify the possible use of observed data in predicting the optimal model to use in a given localised situation.

Chapter 6 provides an analysis of rates of false positives seen in genome-wide associations for the models described in chapters four and five. To provide a comprehensive and full evaluation of model performance for genome-wide association analyses, rates of false positives are assessed for models described in chapters four and five. As before, false positive rates are accessed under a range of different genomic conditions. I also simulate a range of heritabilities, in order to determine if rates of false positives for models relative to one another are constant when the proportion of genetic variance explained by the QTL changes. In this chapter I demonstrate important differences between models in their ability to handle high levels of LD between non-syntenic markers and QTL.

Chapter 7 shows the results from a re-analysis of the six broiler lines for QTL influencing ascites susceptibility, using a haplotype model rather than a single-locus analysis. The choice of re-analysis of this data using a haplotype model reflects results shown in chapters three – six. The importance, and potential uses of QTL identified here are considered in the context of breeding programs that require management of metabolic disorders such as ascites, along with progress in production traits. I demonstrate the advantages of considering and utilizing alternative models when mapping for causal variants affecting ascites susceptibility, and conclude with some cautionary advice on considering model choice for mapping using alternative genome panels and populations.

Chapter 8 features a final summary and concluding remarks of the research involved in producing this thesis. A critical evaluation of the limitations of this research is included to provide a full evaluation of the context of the research presented in this thesis. Some perspectives on the future research directions regarding utilisation of observed genome information on localised model choice are given, along with some thoughts concerning optimal parameterization of models.

CHAPTER TWO

QUANTITATIVE TRAIT LOCI IDENTIFIED FOR ASCITES SUSCEPTIBILITY IN BROILER CHICKENS

SINGLE-MARKER ANALYSIS

2.1 Introduction

Ascites is the end result of a pulmonary vascular system being unable to deliver enough oxygen to feed the metabolic demands of the broiler (Julian 2000). It is a major cause of economic losses to the broiler industry, as it tends to affect birds approaching market weight, in which large amounts of labour and feed have been invested, as well as contamination in the processing sectors (Maxwell and Robertson 1997). The causes of ascites are multifactorial but interactions between diet, environmental and genetic factors play an important role. Over the past 4 decades intense selection for growth, feed conversion rates (FCR) and breast muscle yield in broilers has led to significant physiological changes. One of these changes of particular importance is the increase in physiological disorders, such as ascites (Baghbanzadeh and Decuypere 2008; Dunnington and Siegal 1996). Ascites syndrome cannot be thought of as a contagious or infectious disease, but rather as a progressive disorder starting with pulmonary hypertension and developing to congestive heart failure and death (Lister 1997; Mitchell 1997).

Both the respiratory and circulatory systems are important in influencing broiler susceptibility to ascites. Unlike mammals, avian lungs are relatively rigid, with limited movement during breathing. Instead, air moves through the lungs *en route* to air sacs before moving back through them on expiration. The composition of heart valves is an additional problem, making birds very susceptible to disorders relating to valvular insufficiency (Julian 1993). Ascites becomes particularly apparent under sub-optimal conditions such as high altitude, with low partial pressures of oxygen, or situations requiring high levels of oxygen consumption through increases in metabolic demands (Julian 2000; Scheele *et al.* 1992).

In recent years ascites related mortality in breeding companies' commercial flocks has been reduced considerably, or even avoided altogether, mainly through management practices such as reduced feed intake and growth rate, that consequently lower metabolic rate and demand for oxygen (Julian 2000). Optimisation of the housing temperatures and ventilation in cold weather conditions can be helpful practices to decrease ascites incidence (Baghbanzadeh and Decuypere 2008). These techniques are designed to slow early bird growth, thus not allowing the birds to achieve their full genetic potential (Julian 2000; Balog 2003). Therefore, the full expression of gains in genetic potential made by breeding companies is limited at the farm level to avoid mortality of ascites susceptible birds. This compromise in economic efficiency becomes more of a problem as growth rate increases and broilers are marketed at an earlier age. A better solution would be the identification of genes affecting ascites susceptibility and reducing the frequencies of unfavourable alleles in MAS programmes. Incidences of ascites are considerably higher in commercial environments, where birds are raised in sub-optimal conditions.

The recent development of a genomic map and identification of millions of SNPs for the chicken (Wong *et al.* 2004), combined with advances in statistical methods, have stimulated initiation of mapping experiments for a wide variety of traits (Abasht *et al.* 2006; Hocking 2005). The chicken is an ideal model organism to use for the dissection of complex traits such as ascites, due to a high reproductive rate and relatively short generation intervals when compared to other livestock species. Additionally, the chicken genome is relatively small ($1.2x10^9$ base pairs) compared to the mammalian genome ($3x10^9$ base pairs) which is a big advantage for research aimed at the identification of characterisation of the genetic architecture affecting quantitative traits. The complexity of ascites syndrome, with its effect on a number of organ systems, suggests the potential influence of a large number of genetic loci. Large numbers of genetic loci and their potential to have small effects on the variation in ascites susceptibility make it a trait particularly amenable for a whole genome scan.

2.1.1 Previous mapping studies for ascites

A few studies have looked for QTL for ascites related traits using either candidate gene approaches, or low-density genome-wide scans. Navarro (2003) conducted a linkage study for SaO_2 and other ascites related traits using three candidate regions on a half-sib population. Unfortunately, no significant QTL were detected in any of the linkage groups, aside from a QTL for fleshing score, with suggestive significance ($p < 0.1$) at chromosome wide level. This was followed up with a genome wide scan on an F_2 population derived from a broiler layer cross (Navarro *et al.* 2005). Numerous QTL showing genome-wide suggestive linkage ($p < 0.1$) were identified, with moderate effects for ascites related traits (not SaO_2), although these had wide 95% confidence intervals.

Rabie *et al.* (2005) performed a genome-wide scan with 420 microsatellite markers in 456 birds with phenotypes derived from progeny adjusted means. A number of genome-wide significant QTL were identified for ascites related traits (not SaO_2), although these also had extremely wide confidence intervals, making identification of candidate regions very difficult. A further analysis of these identified regions was conducted using an advanced intercross line (Rabie 2004). To help identify individuals that are susceptible to ascites, the birds were reared in a cold stress environment, designed to induce incidences of the disease. To increase resolution of QTL positions 34 additional microsatellite markers were genotyped in QTL regions identified by Rabie *et al.* (2005). Unfortunately, none of the traits analysed reached the ($p < 0.05$) significance level in these regions and only body weight at five weeks of age reached suggestive significance level ($p < 0.1$).

The studies by Rabie *et al.* (2004; 2005) and Navarro (2003; *et al.* 2005) indicate that some regions of the genome may include loci that affect traits involved in ascites susceptibility. The initial mapping experiments using a genome wide approach identified several significant and suggestive QTL. However, the estimated map positions of these QTL lacked precision, with confidence intervals spanning 20-60 cM, the size of some small chromosomes. Attempts at resolution of these locations did not reveal many interesting results, despite the inclusion of additional markers and the combination of linkage disequilibrium and linkage analysis methodology (Lee and Van der Werf 2004).

In these experiments, microsatellite markers were used due to their higher information content over SNPs. However, subsequent to this work, the availability of the high-density SNP marker map (Wong *et al.* 2004), and the rapidly reducing costs of genotyping, has circumvented the use of the more informative microsatellite markers. A high-density SNP marker map would be expected to dramatically increase the power of a mapping study for ascites related traits (Ardlie *et al.* 2002).

2.1.2 Extent of LD

One requirement for effective use of LD mapping and of LD markers in MAS is that marker density is high enough that at least one marker is in sufficiently high LD with any putative QTL. With the availability of whole-genome sequences and large numbers of SNPs for most livestock species, high-density marker studies have become possible. The cost associated with genotyping, however, leads to an interest in using the smallest required number of markers for LD mapping and MAS. Because the required marker density depends directly on the extent of LD, which varies between populations, an important step prior to any mapping association analysis is to ascertain the extent of LD in the population of interest.

In association analyses combining data across lines or populations is expected to increase power, providing lines are closely related to one another. This required relationship is essentially a consistency of LD across populations, and so it is of interest to ascertain whether patterns of LD in one population extend to related populations

(Dekkers and Hospital 2002). The extent and consistency of LD for mapping and MAS can be assessed by studying marker-marker LD as an estimate for marker-QTL LD.

Research on the extent of LD in populations has been conducted for humans and several livestock species. Although initial findings in humans have shown LD to extend over very short distances (Pritchard and Przeworski 2001), studies in livestock have shown high levels of LD over much longer distances in cattle (Farnir *et al.* 2000; Vallejo *et al.* 2003), pigs (Nsengimana *et al.* 2004) and sheep (McRae *et al.* 2002). This is thought to be caused by the intense artificial selection to which commercial animal breeding populations have been subjected for many generations and the ensuring reduction in the effective population size (Hayes *et al.* 2003). The extent of LD in these livestock species has led to the assumption that haplotype mapping may be straightforward in livestock breeds (Anderson and Georges 2004).

In chickens, Heifetz *et al.* (2005) evaluated LD between microsatellite markers in a number of breeding populations of layer chickens using a standardized chi-square (x^2) measure. Their results showed appreciable LD among markers spaced by up to 5cM. LD within 5cM was strongly conserved across generations, but differed considerably among chromosomal regions. Aerts *et al.* (2007) investigated the extent of LD on chromosomes 10 and 28 in a white layer and two broiler breeds. They found the extent of LD varied dramatically, although this study was only based on 69 SNP markers, across the two chromosomes. The most comprehensive study to date used genotype data for 959 and 393 SNPs on chromosomes 1 and 4 respectively, with 179-244 individuals from nine commercial broiler lines (Andreescu *et al.* 2007). Andreescu *et al.* (2007) showed that whilst there was widespread LD, it only extended over short distances (<1cM), shorter than had previously been reported in other livestock species. This is potentially due to differences in the historical effective population sizes of chickens compared to other livestock species. These results were consistent with those reported by Wong *et al.* (2004), in which SNP haplotype blocks in their populations were rarely as large as 0.3cM. These results indicate that there may be very large differences in patterns of LD between different chicken breeds and different genomic regions.

2.1.3 Measures of LD

Although the principle of LD is fairly simple (i.e. the non-random segregation of markers in close proximity), the complex interplay between confounding factors such as population subdivisions, bottlenecks and expansions, is not yet completely understood and can make interpretation of LD results difficult. As a result, many different statistics have been developed to characterise the amount of LD between markers. Of which Lewontin's D' (Lewontin 1988) and r^2 (Hill and Robertson 1968) are widely used. Both range from 0 (no LD) to 1 (full LD), but differ in the interpretation of the intermediate values. Intermediate values for D' are not clearly interpretable and are known to be biased upwards (Ardlie *et al.* 2002). In addition, D' is affected by the number of individuals used in the calculation (Weiss and Clark 2002). In contrast, intermediate values for r^2 give an indication of the power to detect association. To have the same power to detect association between a disease and marker locus, the sample size must be increased by $1/r^2$ when compared with the sample size for detecting the effect of the susceptibility locus itself (Kruglyak 1999). Therefore, 'useful LD' is often defined as an

r^2 value higher than 0.3, which indicates that the sample size has to be increased 3-fold (Ardlie *et al.* 2002). If, for example, 500 individuals were required for an association study given a 'perfect LD model' ($r^2 = 1.0$), 500/0.3 individuals would be needed if the r^2 statistic has a value of 0.3 in the region under study.

Even though the r^2 statistic gives a good impression of the level of LD between two markers, SNP discovery strategy and demographic history of the population can influence the actual value of the statistic. As a result, two markers that are very close together can exhibit a low level of LD, while markers that are very distant can show a higher than expected level of LD. It is also known that LD between SNPs with a low minor allele frequency is biased upwards (Gaut and Long 2003). In part this can be explained by statistical properties of the LD statistics (Dunning *et al.* 2000), but may also have a biological meaning because SNPs with low minor allele frequencies have a higher probability of having arisen recently (Nordborg and Tavare 2002). A new SNP is in complete LD with all other loci, and therefore, the more recent the SNPs, the less time LD will have had chance to break down.

Compared to other measures of LD such as D', r^2 is the preferred measure for biallelic loci because it is related to the amount of information provided by one locus about another (Ardlie *et al.* 2002), and is less affected by sample size than D'. Consider two biallelic loci on the same chromosome, with alleles A and a at the first locus and with alleles B and b at the second locus, where the labelling is arbitrary. The allelic frequencies can be written as p_A, p_a, p_B and p_b, and the four haplotype frequencies will be written as p_{AB}, p_{Ab}, p_{aB} and p_{ab}. Then, $r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b}$.

2.1.4 Association mapping

Standard approaches for association mapping using high-density genotype data include the sequential analysis of each SNP marker individually. These genotypes based single-marker tests typically fall into least squares or maximum likelihood approaches. With single-marker least squares methods, phenotypes of an individual are regressed onto its genotype, typically in the framework of a mixed model that accounts for variation for fixed and random effects (Fan and Xiong 2002; Long and Langley 1999). Although methods are now available to fit all SNPs for breeding value estimation simultaneously (Meuwissen *et al.* 2001), the properties of these methods for QTL mapping are not well understood, and most association analyses are based on fitting SNPs separately or as small groups in multilocus methods (Dekkers *et al.* 2006). When individuals are related, these analyses must use animal models that include relationship matrices to avoid bias of SNP effects and significance tests (Kennedy *et al.* 1992).

2.2 Materials and methods

2.2.1 Source of data

Genotypic and phenotypic data were provided by Aviagen Ltd. (Newbridge, UK) as part of their genomic initiative project. Data included SNP genotypes, mean progeny performance, and pedigree information for broilers from six sire lines (10, 11, 12, 14, 28 and 29) that are part of a commercial breeding programme. The number of sires in each line ranges from 163-189. The mean, minimum and maximum number of progeny per sire for each of the lines is given in table 2.1. Phenotypic data available for each sire included mean progeny SaO_2 measures, adjusted for fixed effects of age, hatch group, mating group, and the random effects of the dam's permanent environmental effects and one-half of the estimated breeding value (EBV) of the dam, such that the adjusted trait value for a given sire can be shown as;

$$y_{adj} = \frac{\sum_{i=1}^{n}(y_i - a_i - h_i - mg_i - c_i - (0.5 * BV_{dam_i}))}{n},$$

where y_i is the trait record for individual i, a_i, h_i and mg_i are the effects of sex, age, hatch and mating group respectively for individual i, c_i is the random effect corresponding to the permanent environmental effect of the dam for individual $i. BV_{dam_i}$ is the estimated breeding value of the dam for individual i and n is the number of progeny for the sire. Details of the formula used to determine the phenotypic information for each sire are given in appendix one. Using progeny adjusted means based on large numbers of progeny for each sire increases the effective trait heritability by reducing the amount of residual variance in the model, leading to an increase in the power to detect associations (Hassen *et al.* 2009; Ye *et al.* 2006). Details of SaO_2 as an indicator trait for ascites susceptibility are given in chapter 1, section 1.5. Details of the mean and variance of SaO_2 measures for the six lines are given in table 2.1; in all lines the phenotypes were normally distributed (see appendix two for histograms of trait distributions). Sires with phenotypic records outside three standard deviations were removed from analyses; this amounted to the removal of one sire from line 14 and one from 28. Pedigree information included relationships between sires spanning four generations. Sires were genotyped for 12,046 SNP markers across the genome on an Illumina DNA test panel (Illumina, San Diego, CA). Initial SNP assay development for this panel was coordinated by H. Cheng (USDA-ARS, Avian Disease and Oncology Laboratory, East Lansing, MI), which resulted in a 3K SNP panel with genome-wide coverage, with SNP chosen from those identified by a SNP discovery consortium (Wong *et al.* 2004). A file titled "Database of SNP used in the Illumina Corp. Chicken Genotyping Project" that describes the original 3K SNP panel is accessible at http://poultry.mph.msu.edu/resources/resources.htm#SNPs (last accessed April 09, 2009). To complement the 3K panel, another 9,000 SNP across the genome were chosen from the consortium SNP results to fill gaps and to increase the density in some regions. Genotyping and genotype scoring was done by Illumina, utilising a custom designed BeadChip (Gunderson *et al.* 2004). Genotype calls with a GenCall score <0.25 were excluded, which eliminated <0.5% of SNP genotypes. Over 75% of genotypes had a GenCall score > 0.8. Marker positions (given in base or mega base pairs) were those reported for the second draft of the chicken genome (http://genome.ucsc.edu/cgi-bin/hgGateway?org=Chicken&db=0&hgsid=30948908).

		Mean	Min	Max	SaO_2	SaO_2
Line	Sires	offspring	offspring	offspring	Mean	Variance
10	173	21.8	3	96	-0.04	10.3
11	184	22.8	2	77	0.14	5.4
12	163	14.1	4	76	0.31	7.2
14	186	18.7	4	61	0.65	10.2
28	176	20.8	3	112	0.62	11.2
29	189	20.4	4	82	0.76	21.3

Table 2.1

Sire data available for each line. Sire phenotypes were progeny adjusted means, with sires having a wide range in their numbers of progeny. Number of sires represents sample size for the given line. Mean number of offspring per sire, along with the maximum and minimum numbers are provided for each line. The mean and variance for line SaO_2 measures are shown. For distributions of SaO_2 measure see appendix two.

2.2.2 Data analysis

Lines were initially analysed individually, followed by a joint line analysis, with lines combined based on their relatedness to one another. Relationships estimated between lines, based on allele frequencies, showed that no close relationship exists between all six lines (Andreescu *et al.* 2007). From Andreescu *et al.* (2007) two pairs of two lines (12 and 28, 14 and 29) showed relatively close relationships with one another, but not between groups. These two pairs of lines were combined in a joint line analysis with a line by genotype term included in the model. The interaction term was included to account for differences in the direction of genotype effects between lines. An indication of the extent of line divergence is shown by comparison of marker MAF between lines and estimates of F_{ST} statistics. For each pair wise comparison of lines the MAF of markers were regressed against one another. The regression and correlation coefficients for each pair of lines is given in table 2.2. To further clarify between line relationships F_{ST} statistics were also estimated for each combination of line pairs. Wright's *F*-statistic F_{ST} is a measure of population differentiation based on proportions of heterozygosity within a subpopulation relative to the entire population. Here F_{ST} statistics were calculated for each combination of lines using the equation given by Hudson *et al.* (1992), defined as;

$$F_{ST} = (H_T - H_S) / H_T$$

where H_T is the mean expected heterozygosity for the two lines combined, and H_S is the mean expected heterozygosity within a single line assuming Hardy-Weinberg within populations. Between line F_{ST} estimates are given in table 2.3.

	10	**11**	**12**	**14**	**28**	**29**
10	X	0.002	0.003	0.001	0.002	0.002
11	0.041	X	0.003	0.072	0.004	0.054
12	0.058	0.056	X	0.003	0.193	0.003
14	0.031	0.259	0.049	X	0.002	0.410
28	0.051	0.067	0.443	0.047	X	0.003
29	0.038	0.227	0.054	0.625	0.052	X

Table 2.2

Correlation (upper triangle) and regression coefficients (lower triangle) from the comparison of marker minor allele frequencies between pairs of lines.

	10	11	12	14	28	29
10	-					
11	0.119	-				
12	0.014	0.105	-			
14	0.121	0.005	0.105	-		
28	0.085	0.034	0.071	0.035	-	
29	0.125	0.001	0.111	0.001	0.041	-

Table 2.3

F_{ST} statistics for line pairs. Mean heterozygosity was calculated for all markers within a pair of lines and within each line relative to the pair.

Markers not segregating within a particular line were removed. Likewise, for the joint line analysis, only markers segregating in both lines were included. The number of markers remaining for each line, or groups of lines, is given in table 2.3. In all analyses markers with minor allele frequencies below 0.01 were removed to avoid false positives caused by spurious associations between rare genotypes and outlying trait values. Line specific heritabilities (h^2) of sire SaO_2 were provided as part of the standard breeding programme of the broilers, these are given for each line in table 2.3.

2.2.3 Models

At each SNP locus, sire genotypes were assigned values of 0, 1, or 2 based on the number of copies of allele "1" they carried. These values were then included as covariates in a mixed model analysis to estimate the allelic substitution effect for each SNP. These analyses were conducted separately for each SNP. Sires used in this study were from a commercial population that is under selection, and evaluation of pedigree structures showed that sires within each line belonged to complex pedigrees, with a number of half-sib families represented. Therefore, a mixed model that included an average relationship matrix (*A* matrix) among sires was used. The following model was used to evaluate the association of each SNP;

$$Y = 1_n^{'} \mu + Xg + Zu + e$$

where **Y** is the n x 1 vector of SaO_2 adjusted progeny means for n sires; μ is the intercept; ***g*** is the fixed SNP allele substitution effect; ***u*** is the n x 1 vector of random sire polygenetic effects; ***e*** is the n x 1 vector of random residuals. In the individual line

Line	**SaO_2 h^2**	**Segregating markers**	**$p < 0.05$ Threshold**
10	0.044	8107	4.46
11	0.064	9632	4.43
12	0.014	9756	4.45
14	0.214	10352	4.47
28	0.141	9131	4.51
29	0.168	10031	4.59
12 and 28	0.064	8834	4.33
14 and 29	0.183	9707	4.29

Table 2.4

Heritabilities used in the analyses of each line or pair of lines, along with number of segregating markers and genome-wide thresholds of $p < 0.05$. Thresholds are determined independently for each line of pair of lines using a 10,000 cycle permutation analysis (Churchill and Doerge 1994). Heritabilities were estimated using ASReml software (Gilmour *et al.* 1998). For joint lines heritabilities were estimated from pooled data.

analysis ***X*** is the vector of allele copies carried by each sire at the SNP. For the joint line analysis the ***X*** vector also included a line indicator for each sire. For both models the following expectations and variance were assumed:

$$E\begin{bmatrix} Y \\ u \\ e \end{bmatrix} = \begin{bmatrix} \mu 1 + Xg \\ 0 \\ 0 \end{bmatrix}, \text{ and } V\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} A\sigma_u^2 & 0 \\ & \sigma_e^2 D^{-1} \end{bmatrix},$$

where σ_u^2 is the sire variance, σ_e^2 is the residual variance, **D** is the n x n diagonal matrix with the number of progeny for each sire on the diagonal to provide the appropriate residual variance for each progeny mean, and **A** is the additive genetic relationship matrix determined from pedigree information. The algorithm used to determine the relationship coefficients is given in Lange (1997). The residual variance can be expressed as a function of heritability (h^2) given as, $\sigma_e^2 = \left(1 - h^2/4\right)\sigma_p^2$, where σ_p^2 is the phenotypic variance and $\sigma_e^2/\sigma_u^2 = \left[\left(4 - h^2\right)/h^2\right]$. Association was tested against a null hypothesis of $H_0 = \beta = 0$, where β is the effect of the marker, using an F-test with one degree of freedom.

For the analysis of each line a genome-wide significance threshold was determined by permutation analysis (Churchill and Doerge 1994). Permutation procedures are a computationally demanding way of determining significance levels empirically that do not rely on any distribution assumptions. The following steps describe the use of the permutation analyses in determining genome-wide significance thresholds.

1) Phenotype is randomly shuffled with respect to the genotypes of each individual.

2) The randomized phenotype is used as a Y variable in a genome-wide association analysis using the linear model described above. The p-values from this analysis are recorded and ranked to determine the smallest p-value ($p_{\min}$) from the analysis.

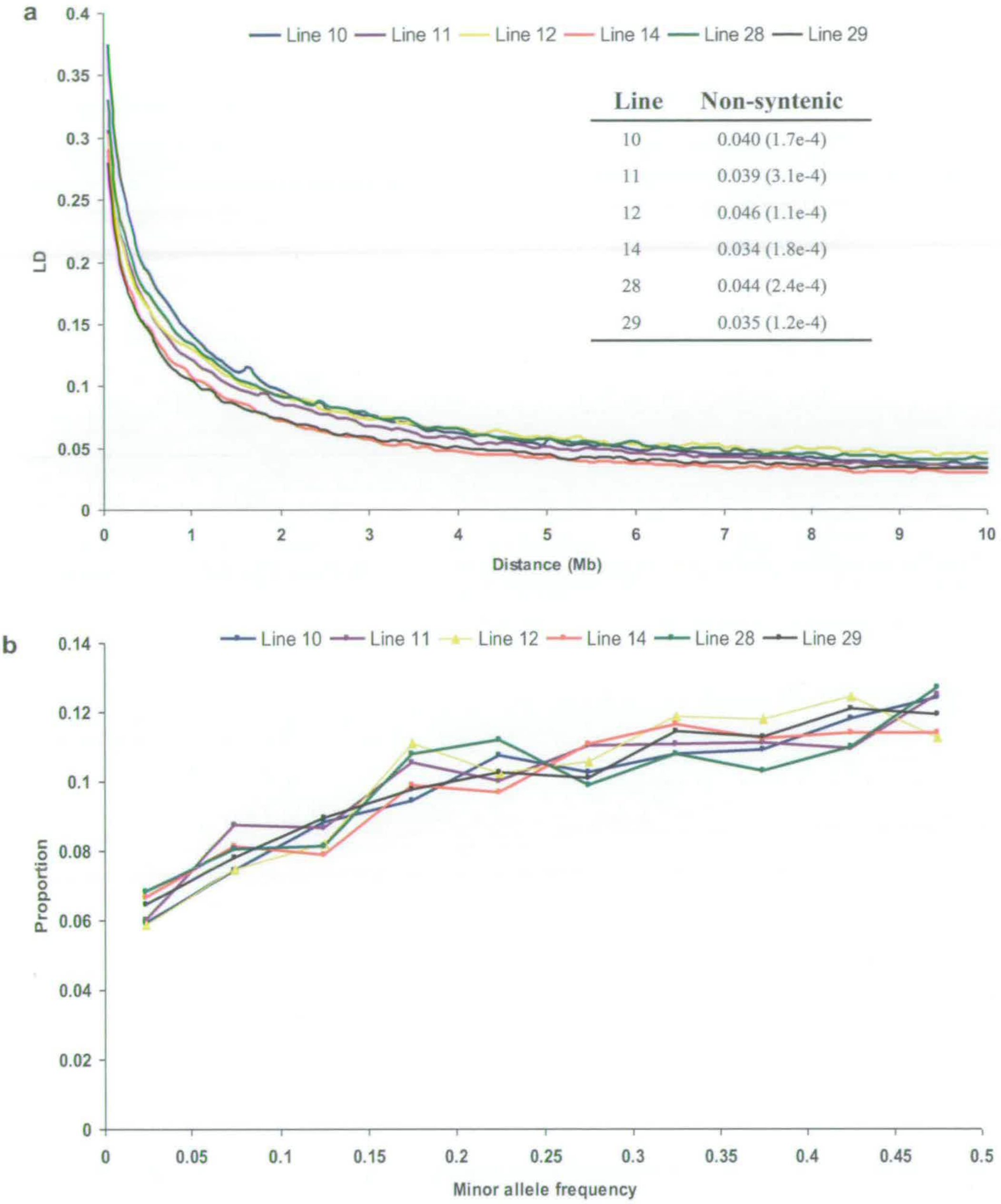
3) Step 2 is repeated 10,000 times. From each run of step 2 the $p_{\min}$ is recorded. After 10,000 iterations the $p_{\min}$ are ranked and the genome-wide significance threshold is determined by taking the value ranked at the 5% level.

During permutation analyses only the phenotype-genotype relationship is destroyed. LD existing between SNPs will remain the same for the original and permuted datasets meaning the genome-wide significance threshold should account for correlations between markers, providing a more realistic value than that from a bonferroni correction.

The permutation method used here randomizes phenotypes amongst all individuals within a line, assuming that individuals are interchangeable under the null. An alternative would have been to shuffle within sire groups to control for within family components of association. However, the majority of sire groups are represented by only a single individual meaning the majority of between family association components would be ignored if phenotypes were randomized only within families (Zou *et al.* 2004).

2.3 Results

2.3.1 Marker allele frequencies and LD distributions



Line	Non-syntenic
10	0.040 (1.7e-4)
11	0.039 (3.1e-4)
12	0.046 (1.1e-4)
14	0.034 (1.8e-4)
28	0.044 (2.4e-4)
29	0.035 (1.2e-4)

Figure 2.1

a) Decline of LD measured by r^2 against distance in Mb for each line. Points shown are mean pairwise LD for syntenic marker pairs in the whole genome against mean distance. The mean and SE of non-syntenic LD measures is also given. **b)** Distributions of MAF for markers in each line, represented as proportions, due to different numbers of markers in each line.

Estimates of LD for each line were calculated for all syntenic pairs of markers. Figure 2.1a illustrates the decline of LD with distance for each of the lines. Both the extent and decline in LD are consistent amongst all lines. Values of mean LD beyond 10 Mb remain constant for all lines. Although the results shown here are the mean for the whole genome, the pattern observed, of high LD at short distances with a steep decline as distance increases, is common for all chromosomes. The extent of LD for non-syntenic marker pairs was also determined by calculating r^2 values for all non-syntenic pairs within a line. The mean pairwise r^2 value for non-syntenic pairs is approximately constant for all lines and is shown on figure 2.1a. For each line non-syntenic LD is equal to the mean of syntenic pairs separated by greater than 10 Mb.

The distribution of minor allele frequencies of markers in each line is given in figure 2.1b. Distributions for all lines follow a similar pattern, showing a roughly uniform distribution, but with an over-representation at intermediate frequencies compared to the assumed neutral U-shaped distribution. These distributions likely reflect the ascertainment bias associated with SNP discovery and marker selection (Solberg *et al.* 2008). Differences between lines are expected, due to population genetic parameters such as random drift, as well as differential selection pressures.

The distribution of maximum r^2 between a SNP and any other syntenic SNP has been used to suggest that SNPs found to be associated with a trait are very likely to be near relevant QTL (De Roos *et al.* 2008). This distribution is examined here for each line, separated into bins on the basis of the distance between the SNP and its maximum r^2 SNP, and shown in figure 2.2. The percentage of SNPs that had a maximum r^2 greater

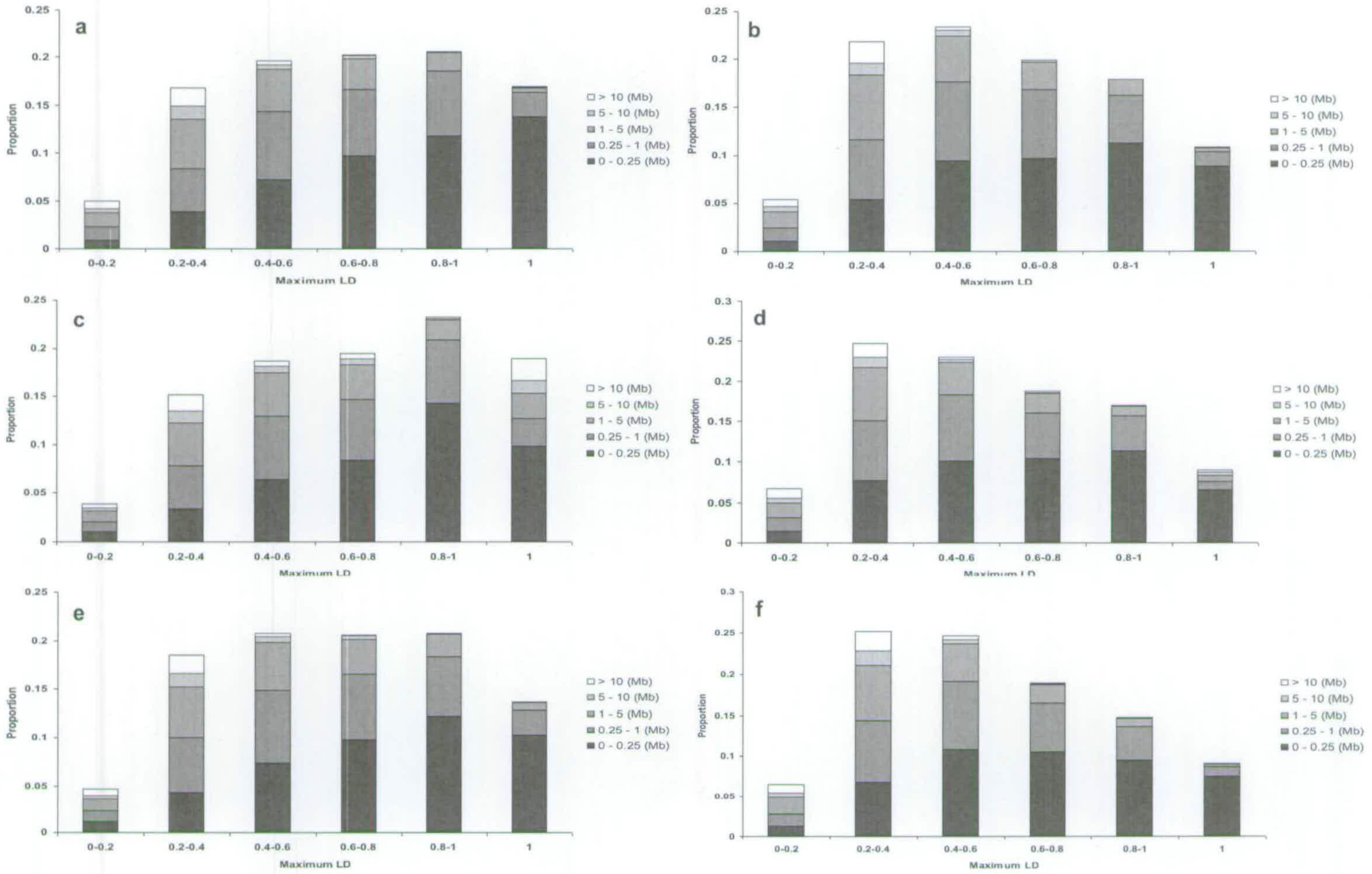


Figure 2.2

Frequency of maximum LD of syntenic SNPs based on r^2. Bins were created on the basis of distance to the SNP for which the maximum LD was registered. **a)** Line 10, **b)** Line 11, **c)** Line 12, **d)** Line 14, **e)** Line 28, **f)** Line 29.

than 0.6 ranged between 42.7-61.6 percent across the lines and between 92.5-95.4 percent of SNPs with a maximum r^2 greater than 0.2. In all lines, for all maximum r^2-value bins greater than 0.2, the shortest-distance bin (< 0.25 Mb) was the most frequent, and the vast majority of maximum distances were less than 1 Mb. For SNPs with a maximum r^2 greater than 0.6, between 2.1-3.4 percent of pairs were separated by distances greater than 1 Mb apart. Whilst this provides promising support that significantly associated markers will be within 1 Mb of the causal variant, it does not exclude the possibility that significant markers may be the results of spurious non-syntenic or long range syntenic high LD, or caused by random associations leading to false positives.

Figure 2.3 summarizes the frequency distribution of r^2 by distance for all syntenic pairs and non-syntenic pairs of markers for each line. Between 11.1-19.6 percent of marker pairs within 0.1 Mb had r^2 values greater than 0.8, across lines. This dropped to between 1.1-2.6 percent for marker pairs between 0.5 and 1 Mb apart. Between 28.6-41.2 percent of markers within 0.1 Mb had r^2 values greater than 0.4, this dropped to between 13.8-20.7 percent for markers 0.25 – 0.5 Mb and between 2.0-1.3 percent for pairs between 1 – 5 Mb. In all lines, markers separated by distances greater than 10 Mb have almost 100 percent of r^2 values less than 0.2. The distribution of r^2 at distances greater than 10 Mb was similar to that of non-syntenic marker pairs with between 99.96-99.99 percent of values less than 0.2.

2.3.2 Genome-wide association results

Whole-genome analyses for each individual line identified a number of markers for association with SaO_2 that showed significant associations above the genome-wide threshold ($p < 0.05$) (figure 2.4). Genome-wide thresholds ($p < 0.05$), as determined from permutation analyses are given for each line in table 2.3. Quantile-Quantile plots (Q-Q plots) for the GWAS of ascites susceptibility in the six lines and two combined line analyses are presented in figure 2.5. All plots indicate that the observed GWAS *p*-values lie close to the expectation and suggest that potential technical and stratification artefacts had negligible impact on the results. Consistent with this interpretation, the genomic inflation factors ranged from $\lambda = 1 - 1.02$ for the six lines indicating little inflation of the p-values. QTL were detected in lines 11, 14, 28 and 29, although, their positions were different between all lines. Details of markers with significant associations are given in table 2.4. In line 11 a single marker with significant association, explaining 1.14 percent of the phenotypic variance, was located on chromosome two. Analysis of line 14 identified a single marker with association above the genome-wide significance on chromosome three that explained 1.56 percent of the phenotypic variation. Both of these single markers identify peaks, comprised of additional adjacent markers that show high levels of association, but do not reach the genome-wide significance threshold. In the line 28 analysis four markers on chromosome one show significant associations, explaining between 2.8 – 5.1 percent of phenotypic variance. These four markers are positioned close to one another, being

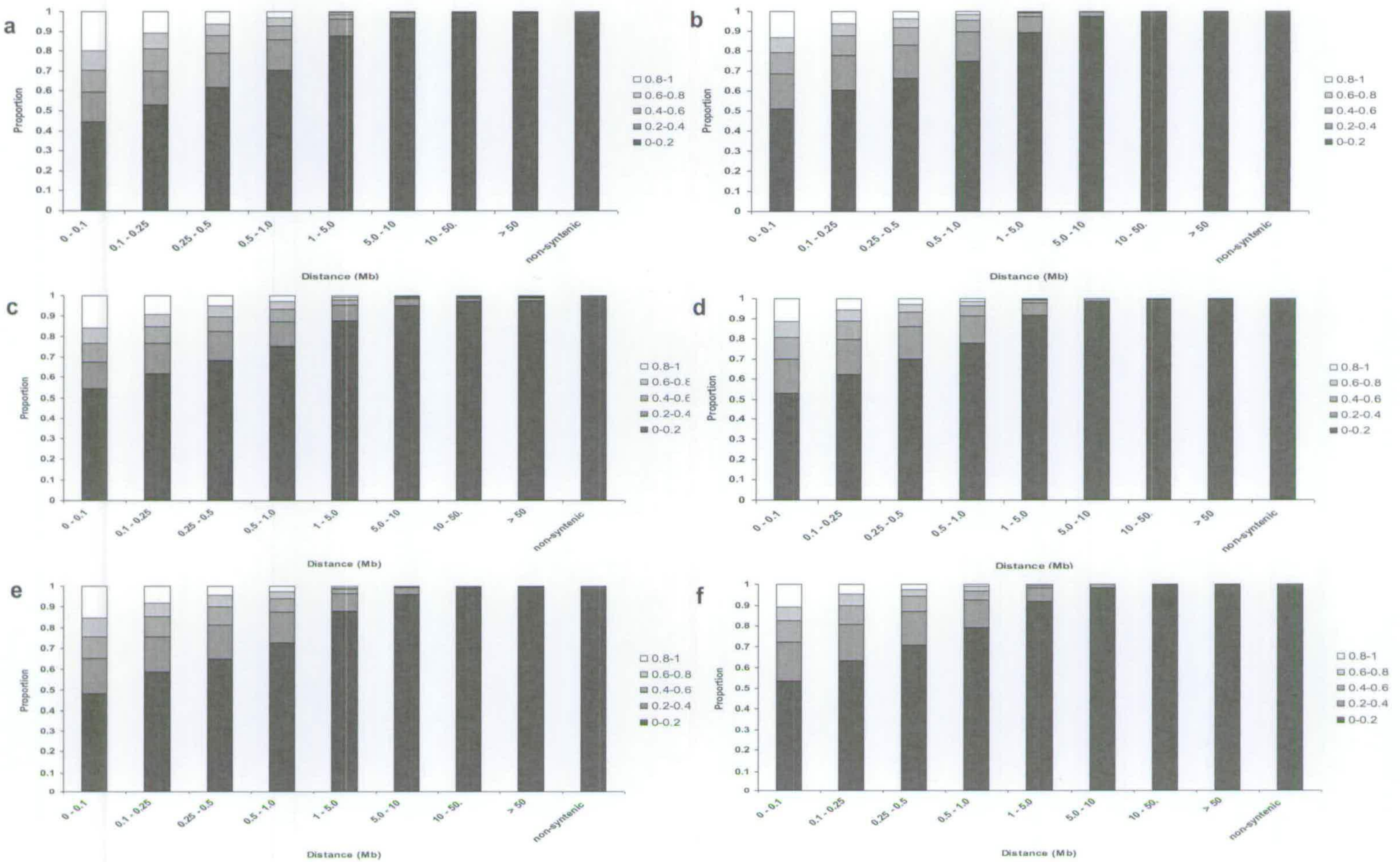
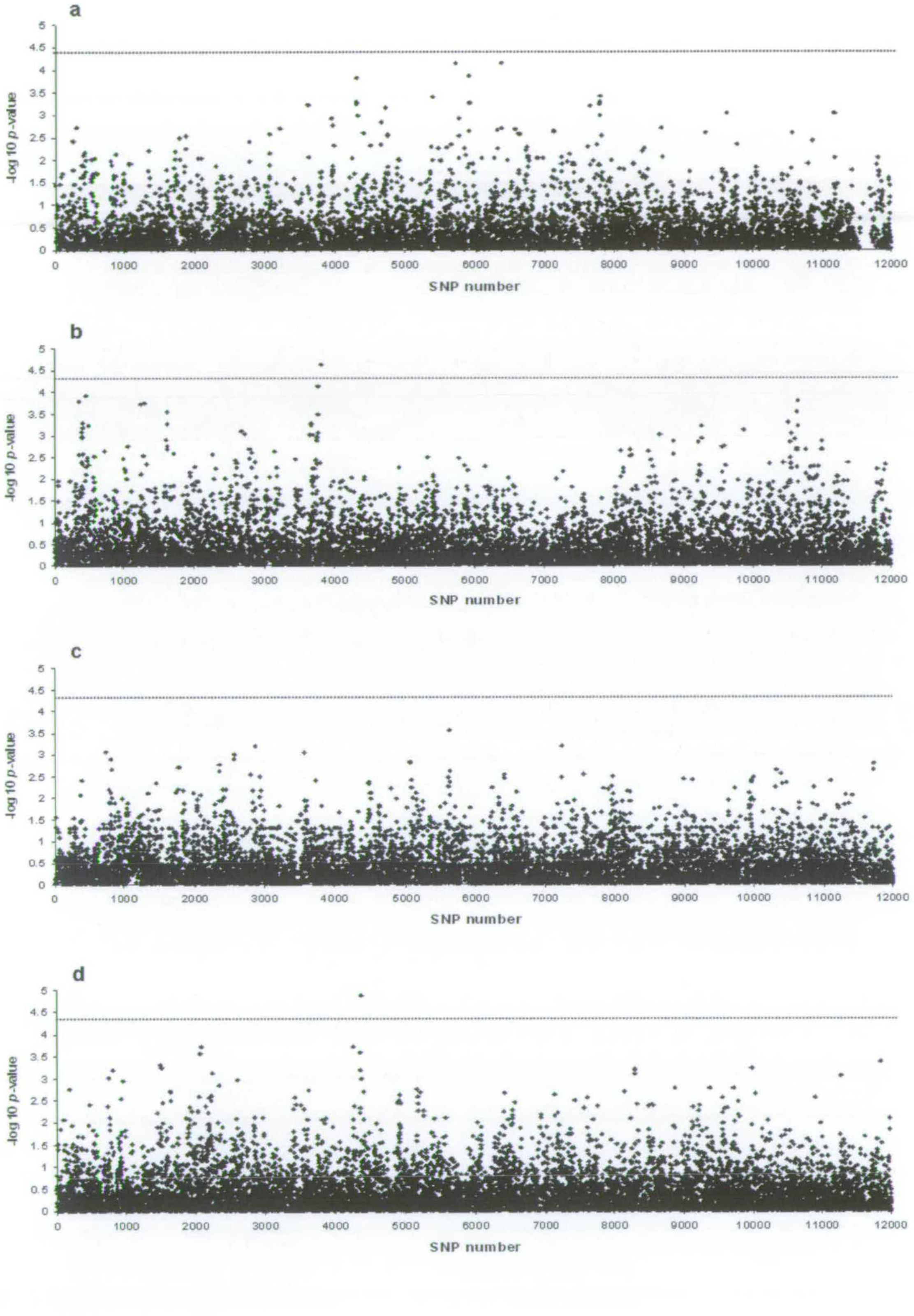


Figure 2.3

Frequency distribution of LD (r^2) for syntenic and non-syntenic marker pairs for each line. Syntenic distribution was computed based on marker distance bins. **a)** Line 10, **b)** Line 11, **c)** Line 12, **d)** Line 14, **e)** Line 28, **f)** Line 29.


a
b
c
d
-log 10 p-value
SNP number
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000

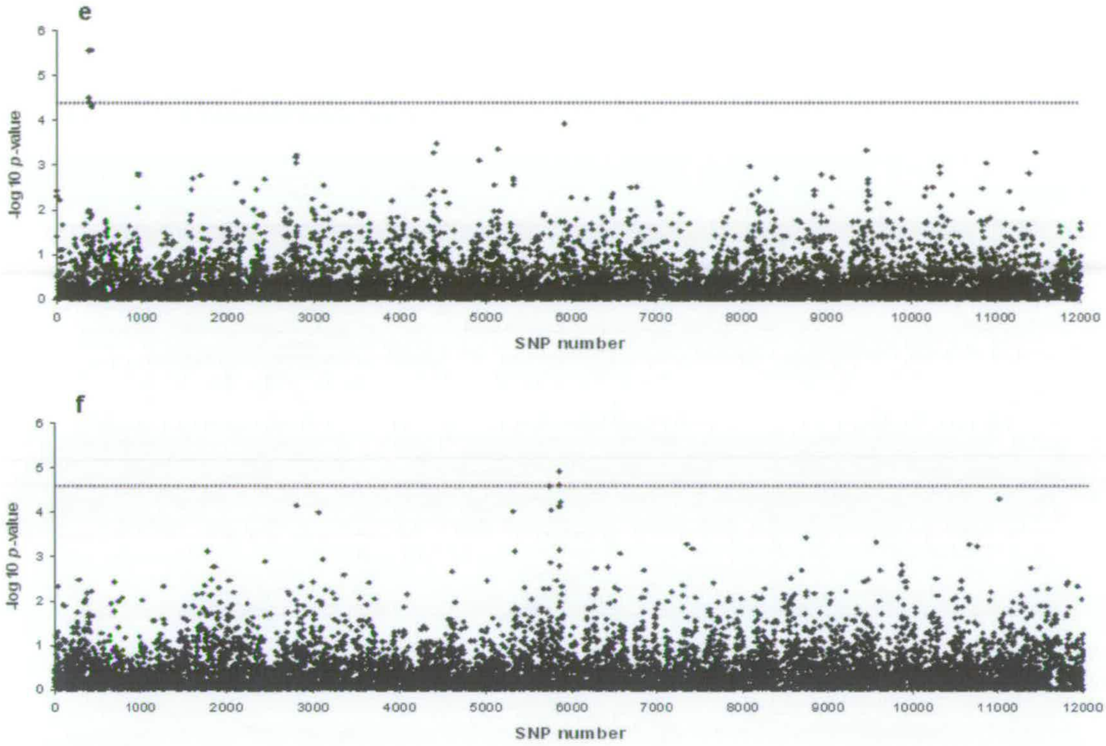
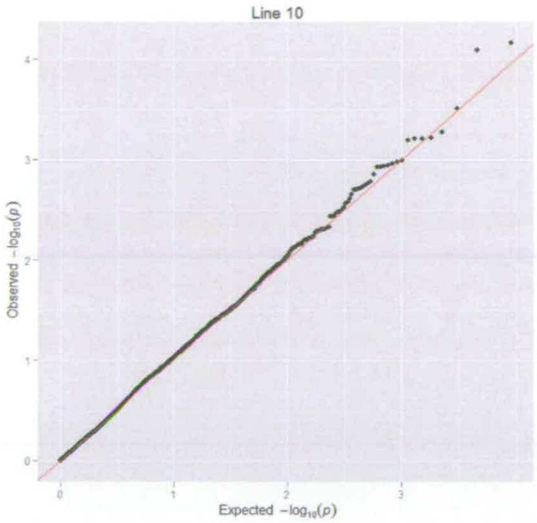
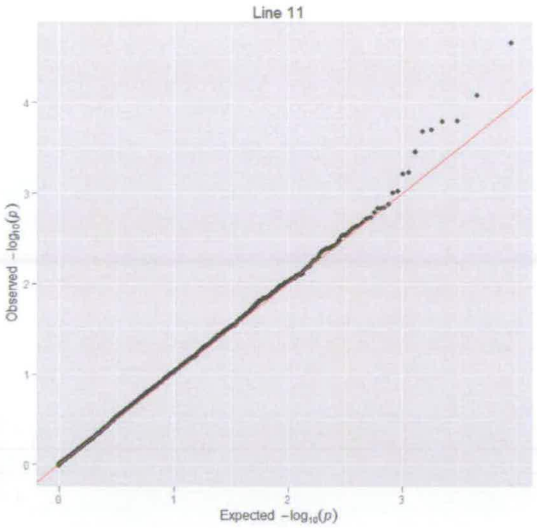


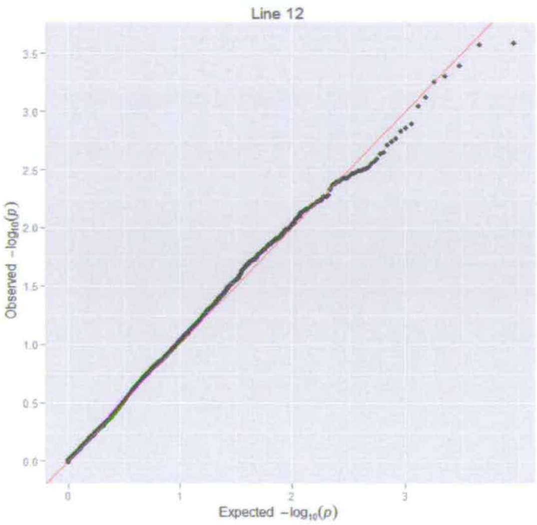
Figure 2.4

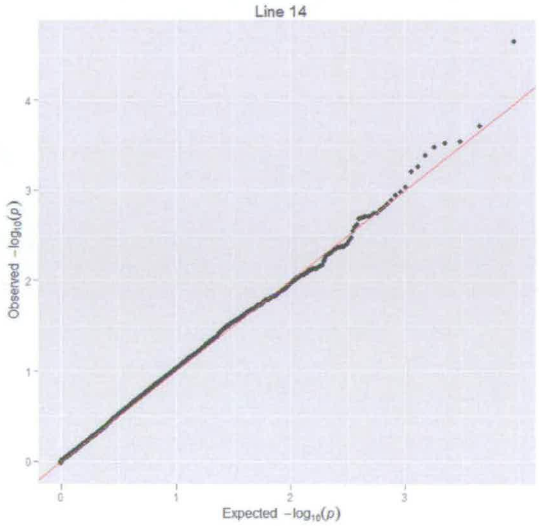
-log 10 p-values for association of each SNP with SaO_2 from genome-wide analysis of each line. **a)** line 10, **b)** line 11, **c)** line 12, **d)** line 14, **e)** line 28, **f)** line 29. For each analysis genome-wide significance was determined using 10,000 cycle permutation analysis. $p < 0.05$ thresholds are shown as the dotted line.

separated by less than 0.7 Mb. Analysis of line 29 revealed two markers showing significant associations, both located on chromosome four. These markers explain 2.37 and 1.36 percent of the phenotypic variance of ascites susceptibility in this line, although they are in high LD ($r^2 = 0.66$) with one another. Association and LD patterns are shown in more detail for the significant region of line 29 (figure 2.6). This figure shows the relationship between marker associations and pairwise LD with one another. This figure also displays the threshold value for suggestive association at a genome-wide level of $p < 0.1$. Within this region, an additional three markers have associations above the suggestive association threshold, providing further support that a locus affecting ascites is located close to this region. –log 10 p-values for SNPs common between lines, were compared, in order to determine whether any regions exhibited high levels of association in more than one line, even if they did not reach genome-wide significance levels. These analyses revealed very few regions that showed promising associations in more than one line (graphical representation of these analyses is not given due to the number of pairwise comparisons produced). This may reflect the high level of divergence between lines and the presence of a number of QTL affecting ascites susceptibility, segregating at different frequencies between lines. Incidentally, the four lines with significant QTL were the four lines with the highest heritability for ascites susceptibility. Combined line analyses included a line by genotype interaction term in the model to account for differences in direction of effects between lines. Results from genome-wide association analyses for joint line analyses are given in figure 2.6. Q-Q plots from the combined line GWAS results are shown in figure 2.5. Plots indicate little stratification artefacts, with corresponding genomic inflation factors of $\lambda = 1.01$ for lines 14 and 29 and $\lambda = 1.006$ for


Line 10
Observed −log10(p)
Expected −log10(p)
0
1
2
3
4
0
1
2
3


Line 11
Observed −log10(p)
Expected −log10(p)
0
1
2
3
4
0
1
2
3


Line 12
Observed −log10(p)
Expected −log10(p)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
1
2
3


Line 14
Observed −log10(p)
Expected −log10(p)
0
1
2
3
4
0
1
2
3

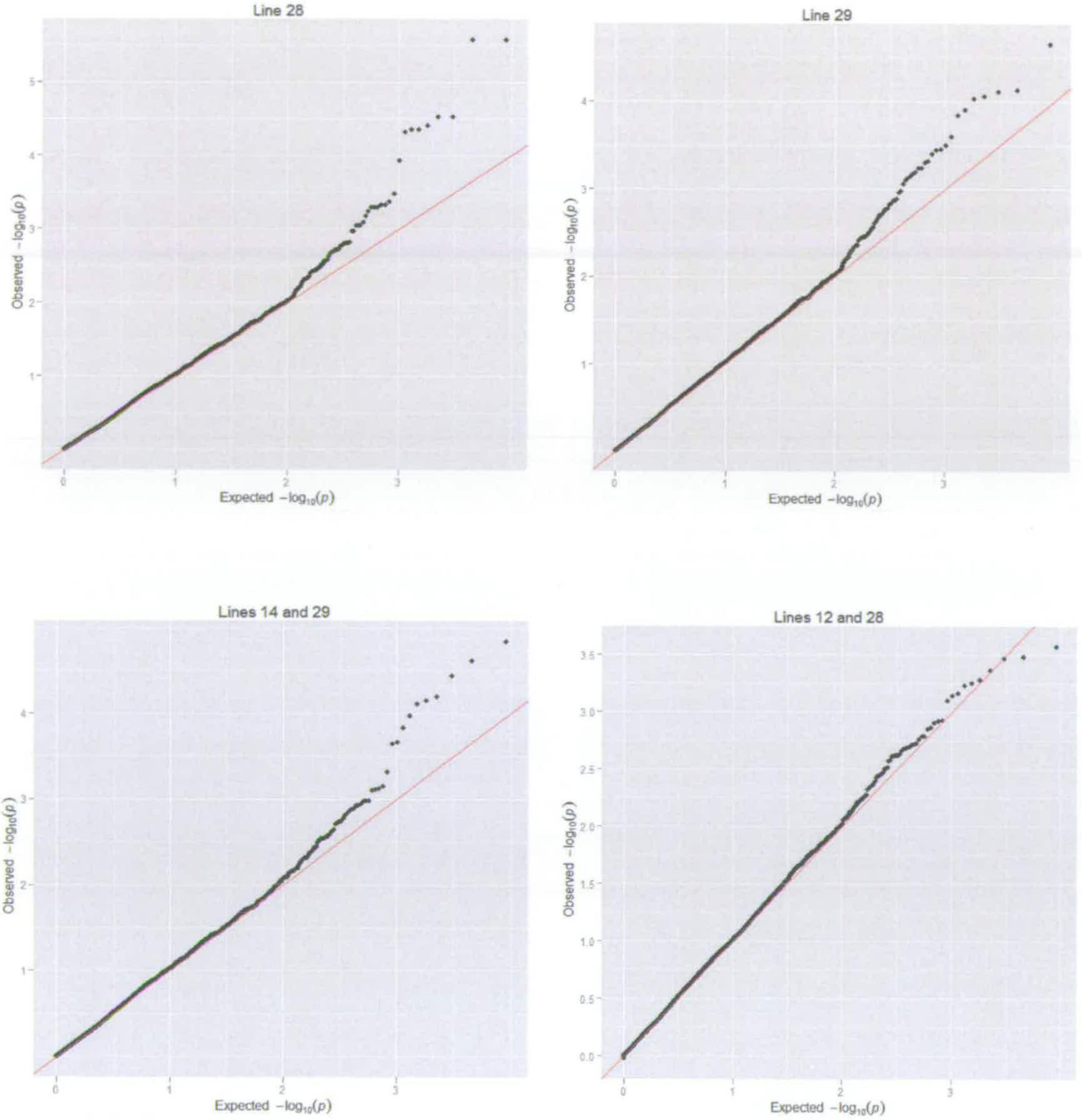


Figure 2.5

Q-Q plots from the GWAS for ascites susceptibility using single marker models. Expected *p*-values under the global null hypothesis of no association are displayed on the x-axis. Observed *p*-values are displayed on the y-axis. The plots show little evidence of stratification.

SNP	Chromosome	Position (Mb)[1]	-log 10 *p*-value	Variance (%)[2]
Line 11				
3819	2	119.8	4.64	1.14
Line 14				
4376	3	13.7	4.87	1.56
Line 28				
370	1	36.5	4.52	1.61
371	1	36.5	5.56	1.31
378	1	36.8	4.52	1.56
409	1	37.2	5.56	1.81
Line 29				
5868	4	24.1	4.93	2.37
5869	4	24.2	4.60	1.36
Line 14 x 29				
5864	4	24.0	4.47	2.47
5871	4	24.3	4.49	1.59
6278	4	73.8	4.82	2.78
6279	4	73.9	4.71	2.49

Table 2.5

Summary of markers identified as significant. [1] is the position of the marker from the start of the chromosome. [2] is the percent phenotypic variance explained by the marker. No redundancy between markers was removed.

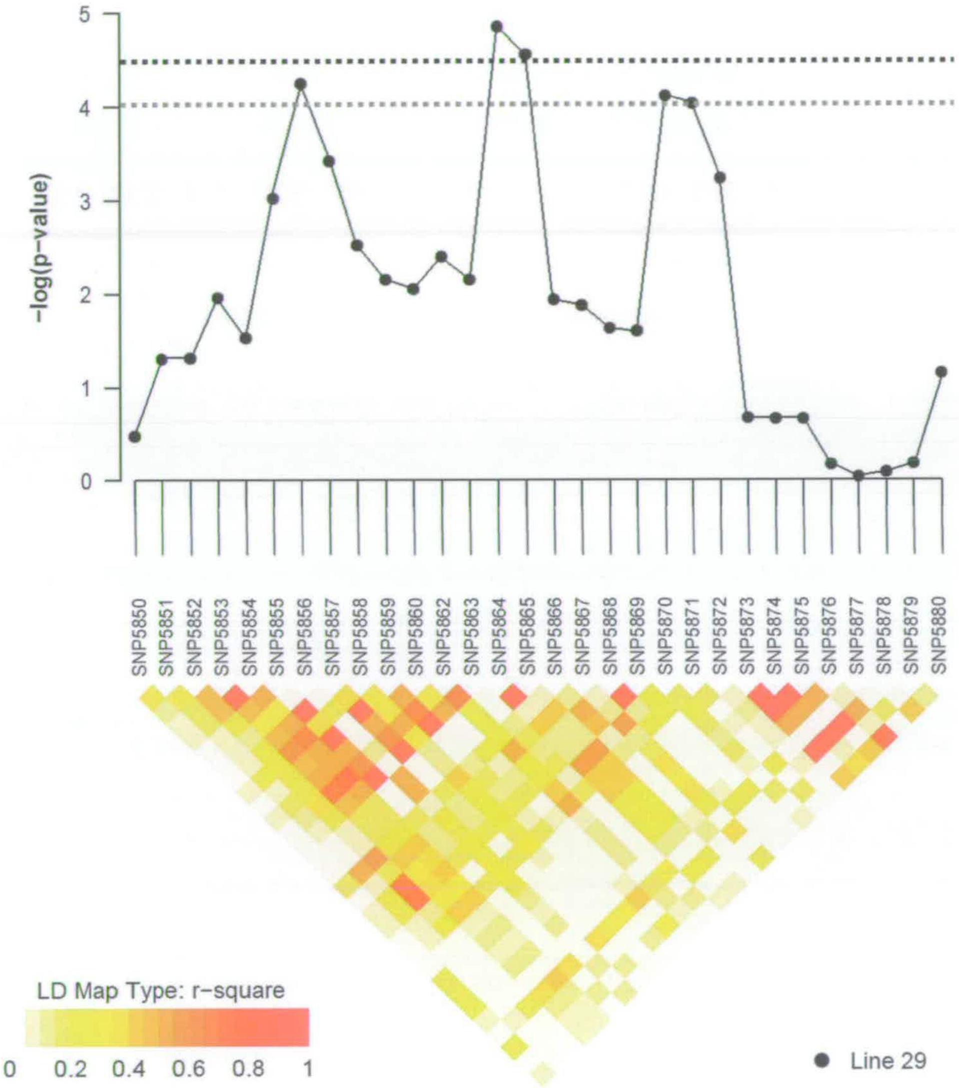


Figure 2.6

Detail of association and LD patterns for the region surrounding markers with significant associations in line 29. LD heatmap is composed of r^2 values for pairs of markers. The black dotted line represents the genome-wide significance threshold of $p < 0.05$, whilst the grey dotted line represents the suggestive association threshold. Both thresholds are determined by a 10,000 cycle permutation analysis.

lines 12 and 28. Analyses of line 14 and 29 identified 4 markers (located in two peaks) that showed significant associations. Both peaks were located on chromosome four, separated by about 50 Mb.

Markers explained between 1.59 and 2.78 percent of genetic variation and showed moderate levels of LD between markers within a peak (0.21-0.24) and low levels of LD between the peaks (0.08). The first of these peaks, located at approximately 24 Mb, was also identified by the line 29 analysis. No significant markers were detected in the analysis of lines 12 and 28. Details of significant markers are given in table 2.4.

2.4 Discussion

Here we report results from genome-wide analyses of six lines of commercial broilers for markers associated with susceptibility of ascites, measured through the indicator trait SaO_2. A total of four QTL were identified by markers that showed significant association above the genome-wide threshold, although each QTL was only identified in a single line. Comparison of results from markers common between lines revealed very poor relationships, with significant markers from one line typically having low associations in other lines. The identification of a number of QTL spread across several chromosomes that explain only a small percentage of phenotypic variance, suggests that numerous loci, with small effects, are responsible for the genetic control of ascites. It is possible that separation of lines within breeding programs and differential selection pressures over many generations may have resulted in QTL segregating at different frequencies between lines, or possibly becoming fixed in some lines (Dekkers

2004). An example of this line divergence can be seen through the comparison of marker MAF between lines, which shows extremely poor relationships for the majority of pairwise comparisons (table 2.2). These potential differences in the genetic architecture of loci affecting SaO_2 values may explain some of the differences in estimated heritability between lines (table 2.3). Difficulty in identifying causal variants common to multiple lines may also be explained by limited power to detect consistent effects. Aside from mutations that occur after line separation, and fixation of loci, QTL that influence susceptibility to ascites are expected to be segregating in all lines. Thus, differences between lines for factors such as allele frequencies, marker effects, trait heritability and sample sizes between lines are expected to influence power to detect associations (Weller 2001; Hassen *et al.* 2009).

Two pairs of lines (12 and 28, 14 and 29) were combined based on their close relationship to one another, reflecting similarities in genetic architecture (tables 2.2 and 2.3). Despite the increased power from a larger population sizes, the combined analyses only identified a single QTL that had also been detected in a single line analysis. Furthermore, a significant marker (SNP 4376) from line 14 failed to reach significance level in the combined analysis of line 14 and 29. This may also reflect differences in segregation frequencies between lines as SNP 4376 has a MAF of 0.44 in line 14 and 0.18 in line 29. However, the four significant markers identified as significant in line 28 had very low levels of association in the analysis of lines 12 and 28 despite having very similar allele frequencies and patterns of marker-marker LD.

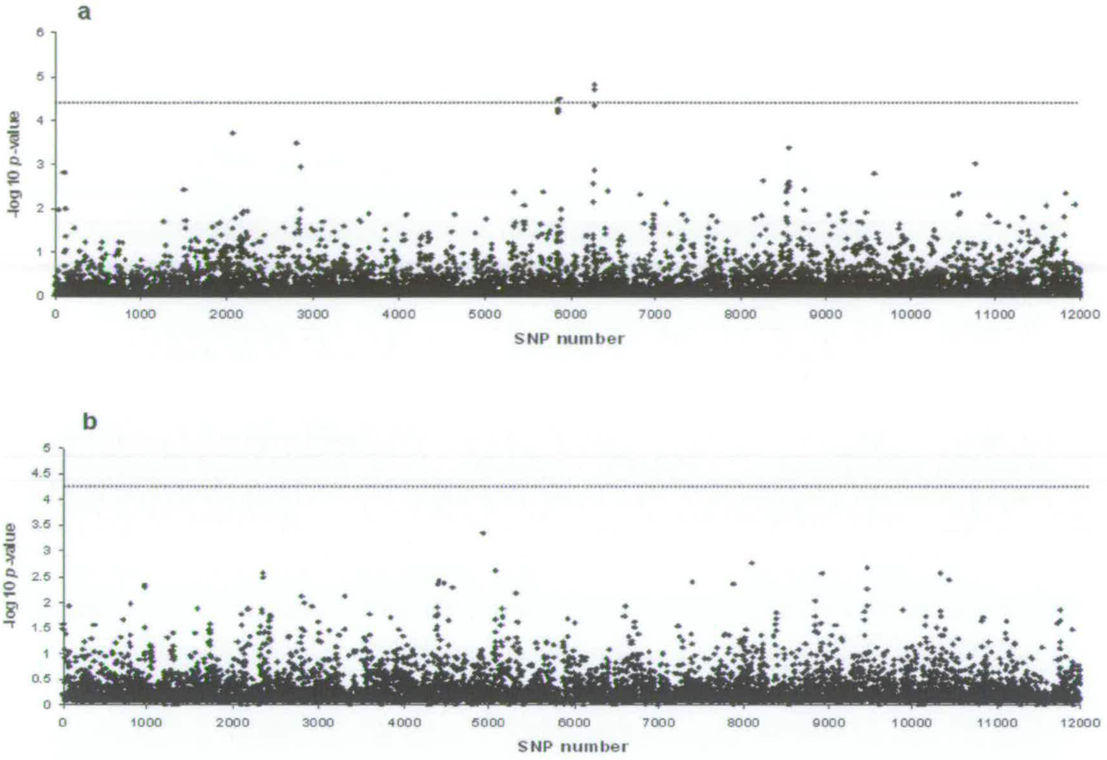


Figure 2.7

-log 10 *p*-values for each SNP from a genome-wide association analysis of line combinations **a)** lines 14 and 29, **b)** lines 12 and 28. For each analysis genome-wide significance was determined using 10,000 cycle permutation analysis. $p <$ 0.05 thresholds are shown as the dotted line.

A number of other studies have identified a few QTL that are related to ascites susceptibility, either through linkage studies or low density genome scans (Navarro *et al.* 2005; Rabie 2004; Rabie *et al.* 2005). Despite these QTL having wide confidence intervals, none of the significant markers identified here are close to significant regions from other studies. Whilst the population and genotype panel studied here represent the most comprehensive dataset analysed to date for ascites related traits, differences in concordance between studies may represent the divergent selection of study populations. This again raises the possibility that ascites susceptibility is influenced by a large number of genetic loci that could be segregating at very different frequencies between populations. Divergence between populations is exacerbated by the current structure of the broiler industry, and their selection for 'product' based lines (Muir *et al.* 2008).

A typical extension to a genome-wide scan is to follow up on identified QTL with either fine mapping tools or candidate gene studies, both of which require identification and genotyping of additional markers in the regions of interest. Given the number and locations of identified QTL this may be an unrealistic continuation of this work, especially if we assume there are potentially many more loci affecting ascites that are undetectable with the methods used here. Exploring alternative approaches to fine map loci and uncover genetic control for ascites, such as inclusion of information from groups of makers either in localised or whole genome situations, may be a more applicable route to take. Using information from a localised set of markers for fine mapping and locus detection can potentially have a power advantage over single-marker analyses as information contained between markers is included as well as information between markers and a QTL. The problem lies in identifying what information to include, as fitting more parameters in the model can potentially lead to an overall reduction in power if they do not explain much additional variance (Akey *et al.* 2001).

2.5 Conclusions

Here we identified four QTL regions on different chromosomes that are associated with susceptibility to ascites, through the analysis of six lines of broilers from a commercial breeding programme using a single-marker regression model. Divergence of lines within commercial breeding programmes, resulting in poor relationships between marker allele frequencies and LD patterns, is thought to have led to a lack of common associations between lines. With the current availability of high density genomic panels, and the potential that ascites susceptibility is influenced by a number of loci, the next stage will be the investigation of alternative genome-wide mapping strategies.

CHAPTER THREE

INVESTIGATION AND COMPARISON OF MULTILOCUS METHODS FOR GENOME-WIDE ASSOCIATION ANALYSIS

APPLICATION TO ASCITES IN BROILER CHICKENS AND SIMULATED DATA

3.1 Introduction

Rapid improvements in high-throughput genotyping technologies have greatly reduced the cost of genome-wide analyses, resulting in a huge range of large scale genetic association studies of quantitative traits and disease variants. These studies typically involve approaches that encompass information spread across the whole genome, using SNP-based LD mapping to systematically evaluate associations with traits of interest. By approaching comprehensive coverage of complex genetic variants, these studies have a statistical power for detecting QTL with moderate effects that is much improved over that of previous studies. However, the density of genotype data has greatly increased the number of variables that need to be tested within a study, adding to computational demands as well as the statistical challenge of accounting for multiple testing.

The majority of association mapping studies use markers individually to detect disease loci. A typical approach was outlined in the previous chapter, whereby genotypes from individual markers are fitted in a mixed model and tested for association with a trait of interest. An alternative approach is to perform the association analysis using information from multiple SNPs that are usually adjacent to one another

(Chapman *et al.* 2003). Broadly, the advantage of multiple-marker based tests is that the LD information contained between markers is included, in addition to any information between markers and potential QTL. However, the advantage is offset by the addition of extra parameters in the models, reducing power to detect associations. Currently there is a huge amount of debate in the literature as to which of these approaches is likely to be more powerful for detecting genetic variation (Akey *et al.* 2001; Chapman *et al.* 2003; Clayton *et al.* 2004). Whilst there is a lack of consensus opinion on which methods will perform best, it is generally expected that optimal performance of models will be heavily influenced by localised genetic architectures and LD patterns (Clayton *et al.* 2004).

When estimating the effects of a single marker, it is well known that the factors that influence power of the test are the size of the effect of the causative genotype on the phenotype, the frequency of the causative allele, and level of LD between the causative locus and marker and how close the allele frequencies match between a causative allele and the marker allele (Weller 2001; Zondervan and Cardon 2004). If we are estimating effects from multiple marker loci, the strength of LD amongst the marker loci will also influence power. The heuristic reasoning, that the inclusion of between marker information conveys an advantage to mapping studies using multiple adjacent markers (Akey *et al.* 2001), has prompted a diverse range of studies that have explored the use of information from multiple markers as opposed to single-marker tests in LD based association studies. If we are interested in fitting multiple markers within a linear model framework then there are a number of considerations on how to parameterize the model in terms of using the information contained between the set of markers. Fitting just main effects is a natural extension to the single-marker model described in chapter two.

Beyond this, interaction parameters can be included, where the number of interactions will be a function of the number of markers fitted. Fitting haplotype information will include all main, interactions as well as phase parameters that are observed between a set of markers.

The majority of studies have focused on comparing single-locus models to those formed with haplotypes. Conclusions drawn from these studies are typically constrained by simulated genetic parameters under which models are tested. Moreover, contradictory results have arisen from empirical data as some studies suggest that haplotype based LD methods improve power over single-marker tests (Akey *et al.* 2001; Calus *et al.* 2009; Guo and Lin 2009; Li *et al.* 2007; Martin *et al.* 2000; Morris and Kaplan 2002; Schaid *et al.* 2002; Yu and Schaid 2007; Zaykin *et al.* 2002), while other studies do not (Fan and Xiong 2002; Grapes *et al.* 2004; Nielsen *et al.* 2004).

3.1.1 Obtaining haplotypes

In genomic studies a traditional method to determine haplotype phase is the collection of genotypic information from multi generational pedigree individuals. However, this information is often very costly to collect or unavailable in many livestock breeding programs and human populations. A solution to this problem is to use a population based statistical algorithm to account for ambiguous haplotypes. Several rule-based and likelihood-based algorithms have been proposed, including a parsimony algorithm (Clark 1990), a Bayesian population genetic model that uses coalescent theory (Stephens *et al.* 2001), and maximum likelihood (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long *et al.* 1995). An advantage of the likelihood approach is that, in addition to the estimated haplotype frequencies, the posterior probabilities of the pairs of haplotypes that are consistent with the observed genotypes can be computed for each subject. This provides an opportunity for phase uncertainly to be accounted for, potentially reducing errors associated with accurately estimating haplotype effects (Morris *et al.* 2004).

3.1.2 Aims

Relative performance of these alternative methods are difficult to discern, and are expected to be influenced by localised genetic architecture and genetic parameters of the causative locus - the effect of these conditions on model performance has been investigated in future chapters. Here we have explored alternative approaches to whole-genome analyses, and their application to map for ascites susceptibility in a line of broiler chickens, as well as simulated data supplied as part of the 12th QTL-MAS workshop held in Uppsala, Sweden, 2008.

3.2 Materials and methods

3.2.1 Data

Three alternative mapping approaches were applied to two different datasets. The first was a single line of broiler chickens, described in detail in chapter two. The second dataset was supplied as part of the 12th QTL-MAS workshop and was analysed in conjunction with a collaborative project (Lam *et al.* 2009). This dataset was supplied and analysed with no knowledge of QTL effects or genetic parameters. After the workshop details of QTL position and effects were released (Crooks *et al.* 2009; Lund *et al.* 2009), providing an opportunity to fully evaluate performance of models.

3.2.1.1 Broiler line

Data included SNP genotypes, mean progeny performance, and pedigree information for a single line (14) that is part of a commercial breeding programme, comprising of 186 individuals. This line was chosen as it had the highest heritability (0.21) for ascites susceptibility. Mean and variance of the progeny adjusted SaO_2 records for sires in line 14 are given in table 2.1. A distribution is also shown in appendix two. Phenotypic data available for each sire included mean progeny SaO_2 measures, the formula used to calculate the adjusted phenotypic records is given in section 2.2.1, along with details in appendix one. Sires were genotyped for the same 12k panel described in more detail in chapter two. Pedigree information included relationships between sires spanning four generations.

3.2.1.2 Simulated data

Simulated data described here was provided by the 12th QTL-MAS workshop, http://www.computationalgenetics.se/QTLMAS08. This comprised of a simulated four-generation pedigree of 4,665 individuals. Phased biallelic marker genotypes were provided at 0.1 cM intervals for six chromosomes, each 100 cM long. The population was simulated with an initial historic population of 50 generations (G_{h1} to G_{h50}), created by 100 founder individuals (50 males and 50 females) in generation G_{h1}. For each subsequent generation, 50 males and 50 females were produced by randomly sampling parents from the previous generation. No phenotype or genotype information was generated for individuals in the historic population. The historic population was followed by four generations of (G_{t1} to G_{t4}), with both genotype and phenotype records. The base generation of the recorded pedigree (G_{t1}) had 15 males and 150 females, the parents of these were sampled randomly from individuals in G_{h50}. Each male was mated to 10 females and each mating pair produced 10 offspring. Generations G_{t2} to G_{t4} were generated by randomly sampling 10 males and 150 females from the previous generation. This created a fullsib-halfsib design, in which each male had 100 progeny and each female had 10 progeny.

3.2.2 Data analysis

Both sets of data were analysed using three different models; a single-locus analysis, a main effects model using three adjacent markers and a haplotype approach also using three adjacent markers. Models utilising information from multiple markers were implemented in a sliding window approach, whereby overlapping windows of three adjacent markers are tested for association, with the window moving forward a single marker after each test. This approach maximises the use of information between adjacent markers and makes comparison with single-markers analyses easier, as an equal number of tests are used. Details of the models fitted are given below.

In both datasets markers with minor allele frequencies below 0.01 were excluded from further analysis. Individuals with phenotypic values outside three standard deviations of the mean in either dataset were also removed from further analysis to avoid false positives caused by rare alleles in individuals with extreme phenotypes. Total variance and heritability were analysed for both traits using ASReml software (Gilmour *et al.* 1998). Heritability of SaO_2 in line 14 was estimated at 0.21 (S.E. = 0.03), and the simulated trait 0.29. Genome-wide significance thresholds were determined for each analysis using a permutation analysis described in detail in section 2.2.3.

3.2.2.1 Single-locus model

The single-locus model used here is the same as that described in detail in chapter two, section 2.2.3. SNP genotypes were fitted as linear covariates in a mixed model analysis to estimate the allelic substitution effect for each SNP. The pedigree structure of the two datasets was evaluated to determine family structure. The broiler line showed that sires belonged to a complex pedigree, with a number of half-sib families represented, whilst the simulated dataset comprised of a number of fullsib-halfsib families covering four generations. Therefore, a pedigree based relationship matrix was also included in the models for both datasets. The same assumptions regarding model variance were made as described in chapter two. Association was tested against a null hypothesis of $H_0 = \beta = 0$, where β is the effect of the marker, using an F-test with one degree of freedom.

3.2.2.2 Main effect model

The main effect model can be considered an extension of the single-locus model above. As before, markers remain coded as 0, 1 or 2, according to the number of copies of the "1" allele that they carry. For single-locus models the design matrix of predictor variable (*X*), has the dimensions 1 x *n*, where *n* is the number of individuals with genotype records, and is composed of marker codes of 0, 1, or 2. Here, this design matrix is extended to fit adjacent markers in a multiple regression framework. As three markers are fitted together, the *X* matrix now has the dimensions 3 x *n*, with each column representing the genotype codes of a single marker. The same model assumptions made for the single-locus model are applied here. This model is implemented as a sliding window, such that a window of three markers is formed, and tested for association with the trait. After each test the window is scrolled forward a single marker and the process is repeated. Association is tested against a null hypothesis of $H_0 = \beta_1 = \beta_2 = \beta_3 = 0$, where β_1, β_2 and β_3 are the effects of the markers, using an *F*-test with 3 degrees of freedom.

Extending this to include more than three markers is simple. In such a case the model can be expressed as;

$$y_i = \mu + \sum_{i=1}^{n} \beta_n x_{n,i} + a_i + e_i$$

Where y_i is the "phenotype" for individual *i*, μ is the phenotype mean, β_n is the substitution effect for SNP *n*, and x_{ni} is the number of copies of "1" allele carried by individual *i* at SNP *n*, a_i is the animal effect for individual *i*. The animal effect was estimated by fitting an A-matrix of kinship coefficients as a random covariate in the mixed model. The A-matrix was determined from the multi-generational pedigree supplied with the broiler chicken and simulated dataset respectively. $e_i \sim N(0,\sigma_e^2)$ is the residual for the i^{th} individual.

3.2.2.3 Haplotype model

As genotype information is only available for a single generation of broilers, haplotype information is not readably available. In these situations a population based haplotyping procedure can be used to statistically infer phase based on observed and expected genotype frequencies of individuals. Under this system haplotype analyses are conducted in a two-stage procedure. Firstly, haplotype information is inferred for the entire dataset using the same principle of overlapping sliding windows. As is used for the main effect model, three marker windows are used. Secondly, inferred haplotypes from a three-marker window are fitted as linear covariates in an H-dimension regression model, where H is the number of haplotypes observed for that set of markers and tested for association with the trait.

Phased genotype information was provided in the simulated dataset, however, for accurate comparison of methods the same haplotyping procedure was used for both datasets. Accuracy of population based haplotyping algorithms is partly dependent on the number of individuals with genotype information (Stephens *et al.* 2001). Therefore, any error introduced by phasing the procedure is expected to be small due to the considerable sample size of the dataset.

Using the sliding window approach, haplotypes were estimated from three adjacent SNP markers using the software "haplo.stats" in R (Sinwell *et al.* 2008). This software utilizes a progressive insertion expectation maximization (EM) algorithm that computes maximum likelihood estimates of haplotype probabilities based on observed genotype frequencies. For individuals with ambiguous phase all haplotype pairs consistent with the observed genotypes are provided, along with the posterior probabilities for each pair. The EM algorithm makes a couple of assumptions; firstly, individuals are considered to be unrelated, and secondly, that marker genotypes are in Hardy-Weinberg equilibrium (HWE). The first of these assumptions is certainly untrue given the data analysed here, whilst the second is likely to be untrue for a number of markers. The implications of deviations from these assumptions are discussed in more detail below.

Inferred haplotypes are evaluated for association with a trait using a score statistic test, which is a computationally efficient alternative to likelihood-ratio or F-test and is implemented within the framework of generalised linear models (GLMs). The score statistic is asymptotically equivalent to the likelihood ratio test, but avoids the need to compute the maximum-likelihood estimates of predictor of haplotype effects, β, making it faster to compute, offsetting some of the time burden associated with inferring haplotype phase. When population based statistical algorithms, such as EM, are used to infer haplotype phase, a posterior probability of each haplotype pair, consistent with the observed genotypes, is provided for each individual with ambiguous phase. As score tests are implemented within a GLM framework, a score statistic can be calculated that accounts for uncertainty in phase by fitting the posterior probabilities of haplotypes. In this situation the score statistic for the effects of haplotypes can be given as $U_\beta = \sum_{i=1}^{N} \frac{(y_i - \tilde{y}_i)}{\sigma^2_{mse}} E_p(X_{gi})$ where X_g is a vector of haplotype codes, β is the regression parameter of X_g, y_i is the measured trait for subject *i*, $\tilde{y}_i$ is the sample mean for the trait, given the assumption of a normal distribution, N is the number of subjects, σ^2_{mse} is the mean squared error and E_p is the expectation over the posterior distribution of genotypes under the null hypothesis, given the observed marker data. That is, $E_p(X) = \sum_{g \in G} X_g Q(g)$ where the posterior probability of a genotype for a subject is $Q(g) = \frac{P(g)}{\sum_{g \in G} P(g)}$. Where *P(g)* are haplotype probabilities obtained from the EM algorithm. If all individuals for a given window have unambiguous haplotype phase then the equation for the score statistic is similar, however, $E_p(X_{gi})$ is replaced by X_{gi}, a design vector of haplotype pairs for each subject. Association of haplotypes with the trait can be tested with a global score statistic against $H_0 = \beta = 0$. The global score statistic is computed according to $S = U_\beta' V_\beta^{-1} U_\beta$ where V_β is the variance matrix of U_β (Schaid *et al.* 2002). The score statistic has a large chi-squared (x^2) distribution with DF equal to *H*-1, where *H* is the number of haplotypes. Whilst score statistics avoid estimating the maximum likelihood of β it still requires the maximum likelihood estimates of haplotype probabilities. To avoid the potential of large numbers of haplotypes (Yu and Schaid 2007), and risk associated with accurately estimating effects of haplotypes that are observed infrequently (Schaid *et al.* 2002), haplotypes with frequencies of less that

5% in the population were pooled into a rare haplotype class before fitting haplotype effects in the hypothesis test. However, the result is that true associations between rare haplotypes and the trait are difficult to pick up, and almost impossible to interpret. The implications of this are discussed in more detail below.

The score statistic framework is fitted in the package "haplo.stats" (Sinwell *et al.* 2008), and is unable to incorporate any relationship matrices. Therefore, analysis of both datasets with the haplotype model assumes individuals are unrelated. The potential error introduced in estimating haplotype effects is covered in more detail in the discussion.

3.3 Results

3.3.1 Broiler line – ascites susceptibility

Genome-wide significance thresholds as determined by permutation decline slightly as model complexity increases due to increases in correlation structure between tests, such that haplotype models have the lowest thresholds, and main effects are lower than single-locus models. Differences in thresholds are due to correlation, or lack of independence, among tests, which are greater amongst multilocus models that are implemented in overlapping sliding windows. Whole-genome analyses using a single-marker test identified one marker with significant association above the genome-wide significance threshold $p > 0.05$, located at 13.7 Mb on chromosome three. Whilst this marker was not included in any significant main effect or haplotype windows (table 3.1), additional markers in close proximately were. Significant windows from the main effect and haplotype models are within 1.3 Mb of the marker found as significant from the single-locus analysis (figure 3.1). In addition the haplotype model detected a second QTL on chromosome one with a single marker showing significant association, locating a clearly defined peak. Q-Q plots for GWAS analyses using the three models are presented in figure 3.2. Plots show little evidence for population stratification impacting on *p*-vales from the analysis. Corresponding genomic inflation factors are all close to 1, with λ = 1, 1.008 and 1.011 for single marker, main effect and haplotype models respectively. The main effect and haplotype models are implemented in a three-marker sliding window based framework, such that each single marker will be represented in three tests for association. Therefore, if a marker shows a high association with a trait it could be expected to produce a series of three high test statistics represented as association of three adjacent marker windows in a whole-genome scan. Details of association results and the LD pattern for the region between 10.4-14.4 Mb on chromosome three is shown in figure 3.3. This region contains markers, and marker windows that are significant for all analyses. The figure shows a complex pattern of LD, with no clearly defined block structure. Furthermore, intermediate values of pairwise r^2 remain for markers separated by considerable distances. Individual markers that show significant levels of association in the single-locus analysis do not show high values of LD with other markers, although LD patterns with causal variants are unknown. Likewise, interpreting multilocus model performance on the basis of local pairwise LD is also difficult, as high order LD patterns can exist between haplotypes and QTL that are not observed in such analyses. Two significant marker windows, adjacent to one another, were identified using the main effects model (SNPs 4354-5). These windows include two markers that are common to both windows, due to the overlapping nature of

SNP	Chromosome	Position (Mb)[1]	-log 10 *p*-value	Variance (%)[2]
Single-Marker				
4376	3	13.7	4.87	1.56
Main effects				
4253	3	4.45	5.36	2.21
4354	3	10.7	4.59	1.74
4355	3	10.8	4.96	1.62
Haplotype				
1484	1	126.8	4.39	2.45
4253	3	4.45	4.40	2.42
4355	3	10.8	4.61	1.82
4360	3	10.9	4.41	1.78

Table 3.1

Summary of markers identified as significant in the three analyses of line 14 data. [1] is the position of the marker in Mb from the start of the chromosome. [2] is the percent phenotypic variance explained by the marker. No redundancy between markers was removed.

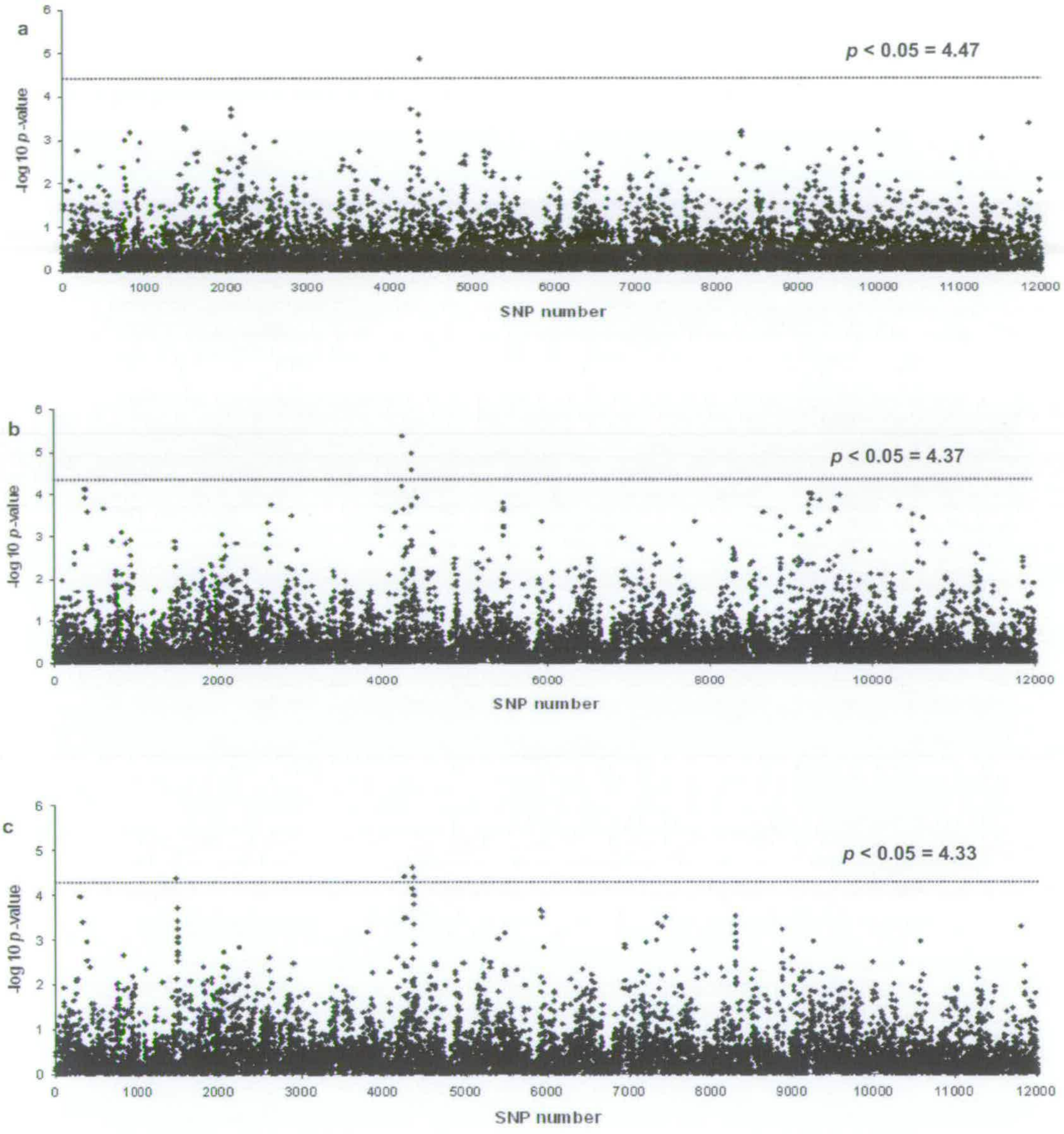


Figure 3.1

-log 10 *p*-values for association of each SNP with SaO_2 from genome-wide analysis of line 14 using **a)** Single-marker, **b)** Main effects (three marker sliding window), **c)** haplotype (three marker sliding window). For each analysis genome-wide significance ($p < 0.05$) was determined using 10,000 cycle permutation analysis (Churchill and Doerge 1994), and shown as the dotted line.

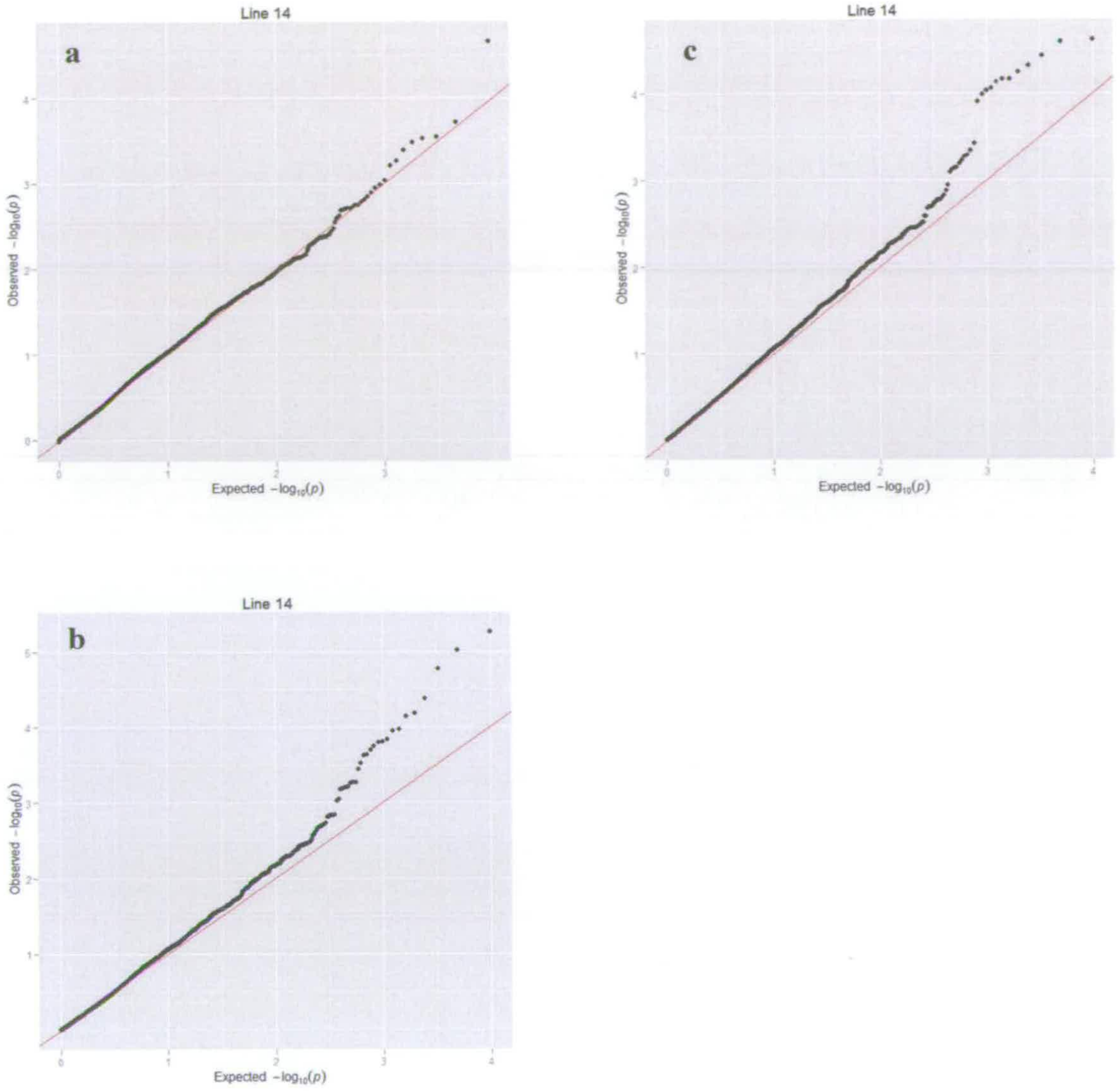


Figure 3.2

Q-Q plots from the GWAS for ascites susceptibility using single marker (a), main effects (b) and haplotype (c) models. Expected *p*-values under the global null hypothesis of no association are displayed on the x-axis. Observed *p*-values are displayed on the y-axis.

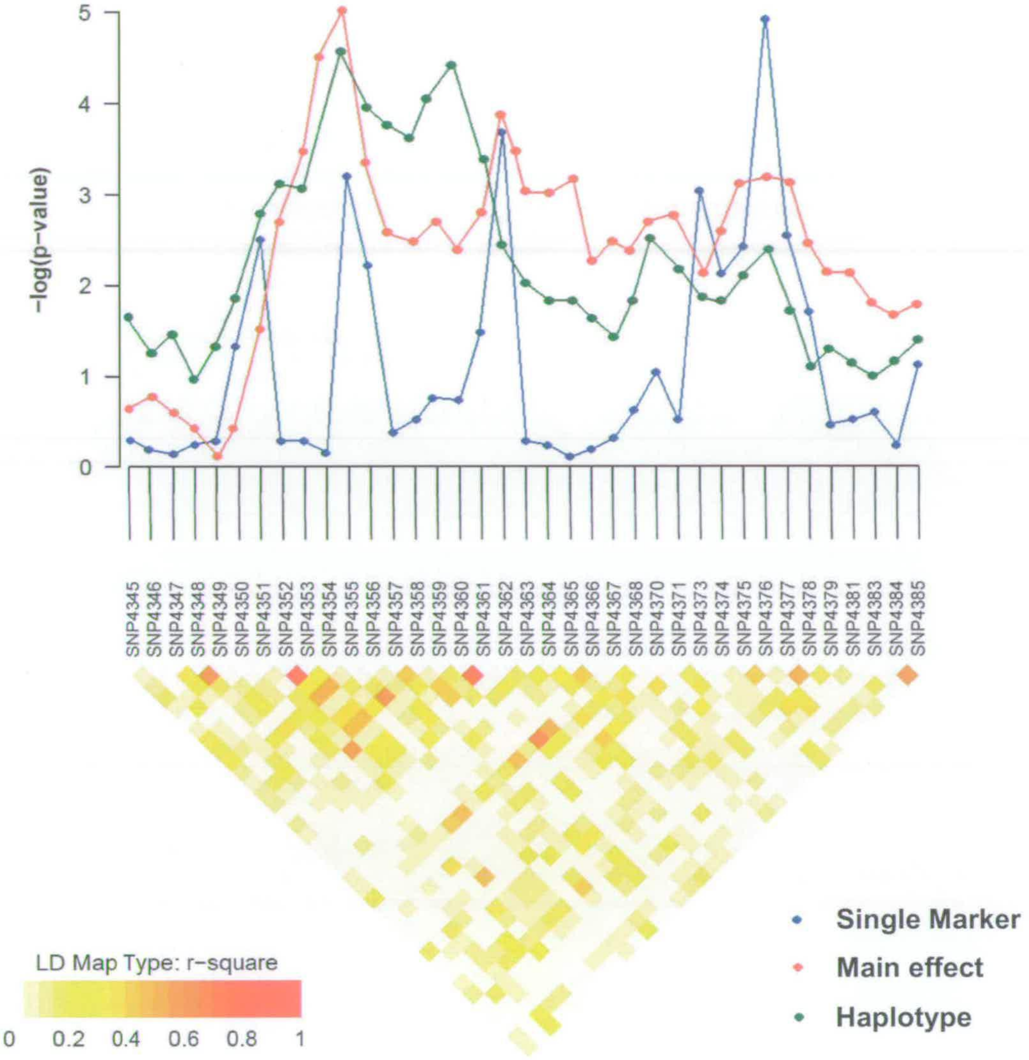


Figure 3.3

Detail of association and LD patterns for the region between 10.4-14.4 Mb on chromosome three. This region includes markers, or marker windows that are significant in all analysis. The lines represent –log 10 *p*-values from each model. Results shown for the main effect and haplotype analysis are shown so as to represent the position of the first marker within the window. The LD heatmap is composed of r^2 values for pairs of markers.

the sliding window approach. However, in the single-locus analysis neither of these markers reaches the genome-wide significance level (-log 10 *p*-values 3.18, 2.21 respectively). Likewise, SNP 4376 is only identified as significant by the single-locus analysis, and is not included in any significant windows from the multilocus model. Such examples highlight the differential use of genomic information contained within and between adjacent markers to provide statistical support for a causative locus.

3.3.2 Simulated dataset

At the time of analysis information on QTL position and effects in the simulated dataset was unknown. One aim of distributing the dataset to participants of the QTL-MAS workshop was to evaluate and compare alternative model performance. Although conclusions drawn from this are limited, the publication of simulated QTL information allowed a more rigorous evaluation of individual model performance and a way of assessing the ability of the models to detect loci. In total 50 indirect QTL were simulated, split into 15 major and 35 secondary QTL. Major QTL were simulated with a pre-defined position and their effects were chosen so that the QTL explained a fixed proportion of the genetic variance. The locations and effects of secondary QTL were randomly sampled. These QTL were spread across chromosomes 1-5, with none simulated on chromosome six (Crooks *et al.* 2009). However, of the 35 secondary QTL simulated only 28 could potentially be detected in the dataset as one was fixed within the population and six displayed no allelic substitution effects.

To evaluate the performance of the models a QTL was considered to be detected if either of the two flanking markers were identified as significant. This could be considered a conservative criterion to determine QTL detection, as only flanking markers are used rather than markers within a certain distance to the QTL location. However, it provides a fair comparison of methods and avoids potential inclusion of false positives. Details of simulated QTL and their detection by the three models are given in tables 3.2 and 3.3 and results from the whole-genome analyses in figure 34. Q-Q plots of the results for the three models are given in figure 3.5. Here the extreme deviations from the expected null likely represent the extent of LD within the dataset and the large effect sizes of the simulated QTL. The single-marker analysis detected flanking markers of 10 major and 5 minor QTL, with the main order-effects model detecting 12 and 7 and the haplotype approach 14 and 12. There were no situations where the single-marker analysis identified a QTL that was not detected by the main effects or the haplotype models, likewise the haplotype approach identified all QTL detected in the single or main effect models.

Total phenotypic variance explained by the models was estimated jointly using the most significant marker within each QTL peak. As a number of markers were detected above the threshold for the majority of QTL, taking the most significant marker removes some redundancy among those SNPs. The remaining significant markers from the single-locus analysis jointly explained 14.4% of phenotypic variance, whilst the main effects model explained 18.9% and haplotype approach 24.7%.

3.3.3 Comparison of single-locus and haplotype models

Results from single-marker and haplotype analyses are compared for the two datasets in figure 3.6. Genotype panels remain the same for both analyses, allowing comparison of results, although p-values from haplotype models represent information from three adjacent markers. The results from haplotype analyses correspond to the single-locus value of the first marker within the window. Whilst this provides an indication of the complementary significant markers between analyses, it does not provide a full picture, as a marker can appear significant from the single-locus analysis, and not from the haplotype analysis. Likewise, a haplotype window that has a significant association includes three markers, but is only correlated with one marker from the single-locus model in figure 3.6. Nevertheless, comparison of the two models provides an indication of the relationship between the two models, along with their ability to identify QTL.

The correlation of p-values between single-locus and haplotype models (figure 3.6) shows a pattern of higher p-values for the haplotype models compared to single-locus methods. It is expected that for some sets of markers this observation is an artefact, caused by the inclusion of a high association level single marker in position two or three in a haplotype window. Nevertheless, it does provide an indication of an advantage in the performance of the haplotype model, compared to a single-locus analysis. A possible explanation for this relationship is that the haplotype model produces a larger number of false positives results than the single-marker analysis. False positive rates are unknown for the broiler dataset and no analyses identified false QTL in the simulated dataset.

QTL	Chromosome	Position (cM)[1]	Effect	MAF	Single	Multi	Haplo
M1	1	20.00	0.62	0.28	●	●	●
M2	1	40.00	0.56	0.07	●	●	●
M3	1	77.23	0.37	0.29	●	●	●
M4	2	27.41	0.35	0.44	●	●	●
M5	2	30.00	0.33	0.21	●	●	●
M6	2	48.62	0.37	0.40	●	●	●
M7	2	74.91	0.50	0.18		●	●
M8	3	14.91	0.30	0.40	●	●	●
M9	3	60.00	0.68	0.07			●
M10	4	3.21	0.61	0.39	●	●	●
M11	4	36.93	0.34	0.24			●
M12	4	76.06	0.58	0.41	●	●	●
M13	4	96.49	0.29	0.19		●	●
M14	5	5.15	0.18	0.21			
M15	5	93.50	0.75	0.26	●	●	●
				Total	**10/15**	**12/15**	**14/15**

Table 3.2

Identification of the major QTL simulated in the Uppsala QTL-MAS 2008 workshop dataset by each of the three methods. If markers flanking either side of the QTL were identified as significant then the model is deemed to have detected the QTL. This is represented by ●. [1] is the position of the marker in cM from the start of the chromosome. MAF is the minor allele frequency of the QTL.

QTL	Chromosome	Position (cM)[1]	Effect	MAF	Single	Multi	Haplo
S1	1	31.87	0.01	0.44	●	●	●
S3	1	50.37	0.06	0.46			●
S4	1	52.50	0.05	0.40			
S6	1	86.68	0.01	0.30			●
S7	1	93.99	0.01	0.47			
S8	2	2.25	0.01	0.39			
S9	2	6.52	0.07	0.38			●
S10	2	32.49	0.04	0.41		●	●
S11	2	45.71	0.01	0.09			
S12	2	48.22	0.04	0.08			●
S13	2	89.04	0.12	0.22			
S14	2	93.54	0.25	0.32		●	●
S15	2	95.66	0.02	0.29			
S16	2	97.83	0.13	0.41			
S18	3	7.89	0.01	0.46			
S19	3	21.07	0.02	0.26			
S20	3	29.81	0.07	0.29			
S21	4	3.44	0.08	0.32	●	●	●
S22	4	3.88	0.02	0.23	●	●	●
S23	4	10.00	0.01	0.04	●	●	●
S25	4	19.84	0.07	0.47			
S26	4	69.56	0.00	0.08			
S27	5	12.98	0.09	0.44			●
S29	5	68.39	0.12	0.44			
S32	5	77.02	0.13	0.25			
S33	5	80.00	0.08	0.11			
S34	5	82.14	0.01	0.36			
S35	5	98.32	0.01	0.45	●	●	●
				Total	**5/28**	**7/28**	**12/28**

Table 3.3

Identification of the secondary QTL simulated in the Uppsala QTL-MAS 2008 workshop dataset by each of the three methods. If markers flanking either side of the QTL were identified as significant then the model is deemed to have detected the QTL. This is represented by ●. [1] The position of the marker in cM from the start of the chromosome. QTL fixed within the population, or with no allelic effect are removed from the table.

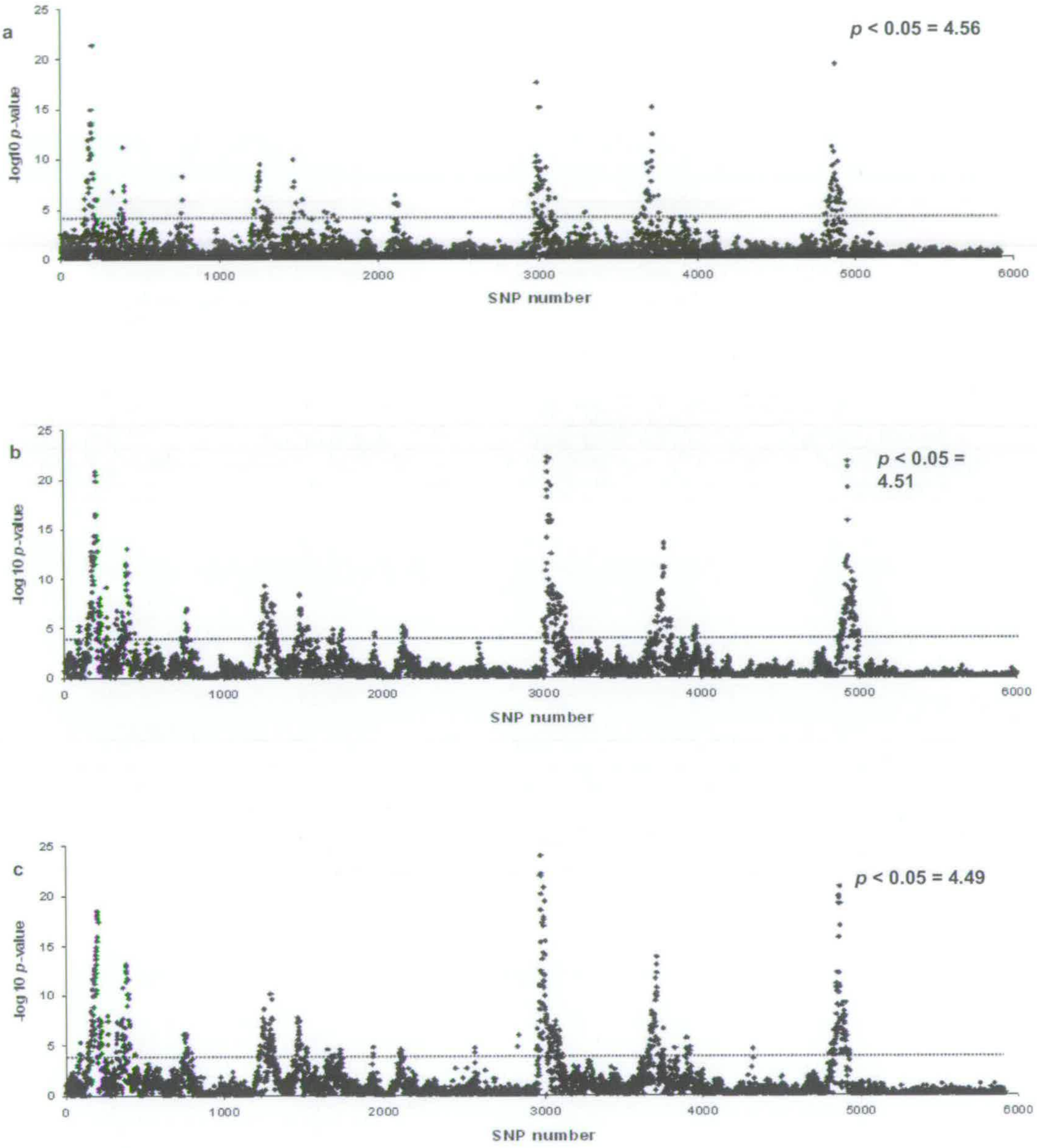


Figure 3.4

-log 10 p-values for association of each SNP with the quantitative trait simulated in the Uppsala QTL-MAS 2008 workshop dataset using **a)** Single-marker, **b)** Main effects (three marker sliding window), **c)** haplotype (three marker sliding window). For each analysis genome-wide significance was determined using 10,000 cycle permutation analysis (Churchill and Doerge 1994). $p < 0.05$ thresholds are shown as the dotted line.

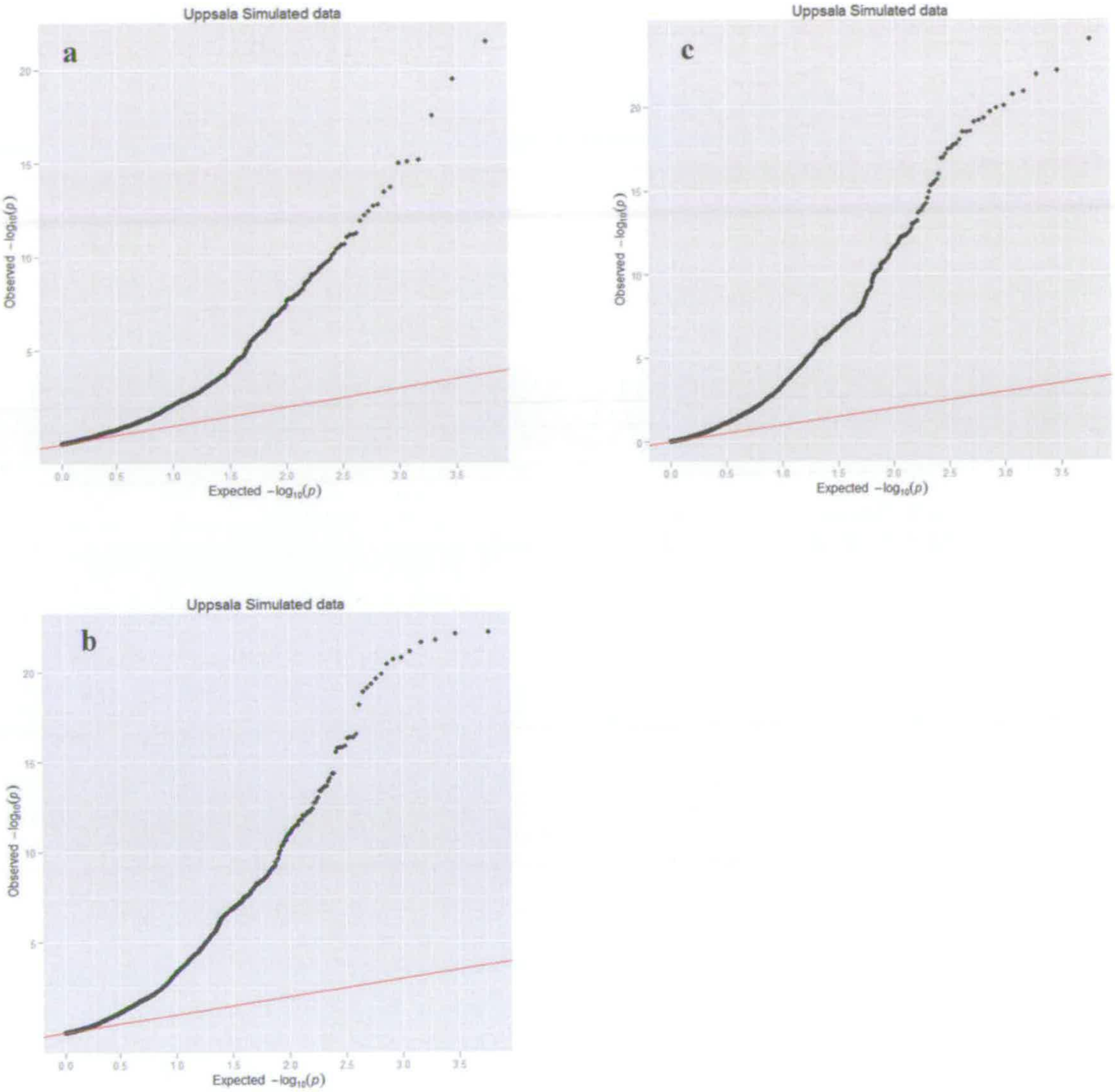


Figure 3.5

Q-Q plots from the GWAS for ascites susceptibility using single marker (a), main effects (b) and haplotype (c) models. Expected *p*-values under the global null hypothesis of no association are displayed on the x-axis. Observed *p*-values are displayed on the y-axis.

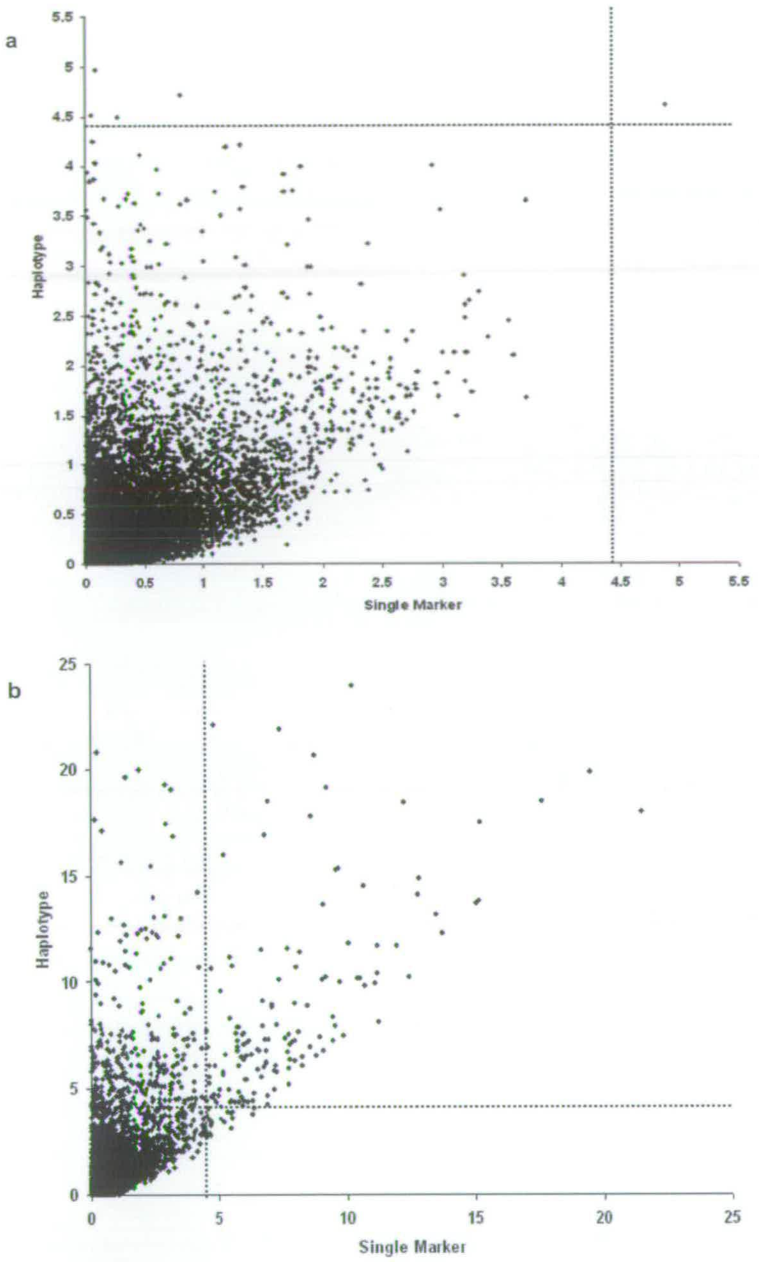


Figure 3.6

Comparison of *p*-values (shown on –log 10 scale) from genome-wide association analyses using single-marker and haplotype models. Results shown for the haplotype analysis are shown so as to represent the position of the first marker within the window. **a)** Broiler data – line 14 **b)** Simulated data. Genome-wide thresholds ($p < 0.05$) are provided as dotted lines for each model.

3.4 Discussion

3.4.1 Broiler line dataset

Results from a genome-wide association analysis of broiler line 14 using a single-marker method were reported in chapter two. Here we investigate the performance of main effects and haplotype models using one of the lines from the same dataset. The QTL identified by the single-marker analysis was also detected by the main effect and haplotype models, strengthening support for the position of a QTL on chromosome 3. The marker with an association above the genome-wide significance threshold (SNP 4376), identified in the single-marker analysis was not included in any of the significant windows of the multilocus models. However, windows that included SNP 4376 typically showed high levels of association that did not quite reach the significance threshold. Instead, the multilocus models identified a number of significant windows that included markers in close linkage with the SNP detected by the single-marker analysis. Between these three analyses a cluster of markers and marker windows, covering a small region of the genome, were identified, lending strong support for the location of a QTL affecting ascites susceptibility. The identification of different significant markers and marker windows by the methods suggests that information provided by markers is utilised differently between models. Support for this conclusion is provided from markers that were identified as significant in a given model in that they typically showed high levels of association in other models, even if they did not reach the genome-wide threshold.

3.4.2 Simulated dataset

Analysis of the simulated dataset provided another opportunity to access differences in performance of the three models, with the additional benefit of the subsequent publication of QTL effects and locations (Crooks *et al.* 2009; Lund *et al.* 2009). Simulated QTL were divided into main and secondary groups, based on their original allelic substitution effects in generation G_0 of the dataset simulation. All three models identified a core set of 15 QTL (10 main, 5 secondary), with the main effects model detecting an additional four (2 main, 2 secondary). The haplotype model identified significant marker windows for all QTL detected by the other two models, as well as an additional seven QTL (2 main, 5 secondary). In terms of identification of QTL position, in this dataset, the haplotype model clearly performs better than the other models, a sharp distinction when performance is contrasted against results from the single-marker analysis. However, broad conclusions need to be cautioned against, given that this is only the analysis of a single dataset that was simulated using simple, idealised conditions. The simulation used is expected to produce overly simplistic patters of IBD, and population stratification, as only 50 generations of random mating, with no selection, were followed by a few generations of population reduction, also with no selection and equal family sizes. Further details of results from analysis of the QTL-MAS workshop data, are given by Lam *et al.* (2009) and Crooks *et al.* (2009).

3.4.3 Fitting a relationship matrix

Both the single-marker and main effects models fit markers as linear covariates in a mixed model with the inclusion of a relationship matrix derived from pedigree information to account for background genetic effects. Currently the R package "haplo.stats" (Sinwell *et al.* 2008) is unable to account for population stratification or relationships among individuals, raising the possibility of error in estimating marker and haplotype effects (Dekkers *et al.* 2006) through confounding with polygenetic effects. If relationship information is not accounted for it can lead to an increase in the number of false positives, a problem that is exacerbated when data from several generations is used (Kennedy *et al.* 1992). Analysis of the simulated dataset using the haplotype model identified very few significant markers that were not close to a reported QTL position, although, as stated before, drawing broad conclusions from this dataset should be cautioned against.

In line 14 assessment of potential errors introduced by ignoring relationship information in the haplotype analysis is much more difficult. Given the complex series of family relationships between the individuals it is likely that some error will be introduced in estimating haplotype effects through confounding with polygenetic effects. Whilst there is strong support for the position of a QTL on chromosome three, the same cannot be said for the QTL identified on chromosome one by the haplotype analysis. In this situation, interpretations of haplotype results are difficult to make, without comparison to results from other methods. To a certain extent, this negates the identification of any benefit of haplotype models when polygenetic effects are not accounted for, although comparison of results shown in figure 3.64 shows a consistent relationship between datasets.

3.4.4 Assumptions of the EM algorithm

Without the availability of parental genotype information for data such as line 14, inferring haplotype requires the use of a statistical algorithm – such as the EM algorithm. The EM algorithm infers haplotype frequencies and posterior probabilities of haplotype pairs for a given individual based on observed genotype frequencies and the assumptions that individuals are unrelated and markers are within HWE (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long *et al.* 1995). The accuracy of haplotype frequency estimation using EM approaches has been considered in relation to haplotype frequency distributions, deviations from HWE, allele frequencies, and LD levels, unaccounted for pedigree (Fallin and Schork 2000; Kirk and Cardon 2002; Osier *et al.* 1999; Tishkoff *et al.* 2000).

Departure from HWE could potentially be a significant source of error in the estimation of haplotype frequencies using EM procedures, simply because the algorithm relies on assumptions of HWE in its "E" step. However, this assumption is only broken for situations of multiple heterozygotes, for other cases the EM algorithm relies on gene-counting, which is not affected by departures from HWE. However, the influence of departures from HWE on estimation accuracy is expected to be dependent on the direction of the disequilibrium. Osier *et al.* (1999) noted that there is a balance between loss and gain of accuracy by divergent directions of departures from equilibrium, and even large deviations make little difference to the estimation of haplotype frequencies. Departures from HWE leading to an excess in homozygosity could decrease the amount of missing phase information in the dataset and, as such, lead to improvements in estimation accuracy. Fallin and Schork (2000) simulated a series of situations with positive (towards homozygosity) and negative (towards heterozygosity) departures from HWE and demonstrated that whilst there were small increases in estimation error for negative departures, positive departures showed no change in estimation error rates.

As part of a study specifically looking at haplotype estimation with genotyping errors Kirk and Cardon (2002) concluded that little error will be introduced when haplotyping a related population using EM procedures, unless the marker map has low density. Using the software package PHASE, that uses a Gibbs sampling algorithm, Stephens *et al.* (2001) showed that haplotype phase can be inferred with a very high level of accuracy using marker moderate density of 10 markers per cM.

3.4.5 Score statistics

The major advantage of score statistics is their computational speed compared to regression-based methods and likelihood ratio tests. Moreover, they are theoretically expected to be relatively robust to deviations from normality of the trait distribution and selected sampling (Bhattacharjee *et al.* 2008). Within the framework of haplotype analyses score statistics can also readily compute a test statistic for each haplotype within the X_g vector which has a standard normal distribution for large samples. Given the small number individuals in the line 14 dataset this may be an unrealistic distribution to use for the analysis of this data. If a single haplotype within the $\boldsymbol{X}_g$ vector has an individual test statistic that is considerably greater than all other haplotypes, then taking this value, maximised over all evaluated haplotypes, is likely to have greater power than the global score statistic (Schaid *et al.* 2002). Except where several haplotypes are associated with the trait, a global score statistic is expected to be a better alternative.

Unfortunately score statistics can suffer when sample sizes of parameters are small, often struggling to obtain reliable estimates of these parameters (Peng and Siegmund 2006). Within the framework of haplotype analysis these problems are encountered when a large number of rare haplotypes are fitted in the model, often leading to over estimates of haplotype effects, resulting in false positives. This situation is analogous to likelihood-ratio and *F* tests, but power for score statistics diminishes more rapidly (Tishkoff *et al.* 2000; Schaid *et al.* 2002). In the package "haplo.stats" one way to avoid reduction in power of the global score test, is to pool rare haplotypes into a single parameter class. Typically this is achieved by setting an arbitrary threshold that defines whether or not a haplotype is considered rare. In the analyses conducted here that threshold was set to 5%. Intuitively, this seems a high threshold, meaning that in line 14 data 18 copies of a haplotype may be observed, and still considered rare. In preliminary analyses lower thresholds were investigated but often produced large numbers of false positive results. Neither of these situations is ideal, prompting investigation into alternative approaches for testing haplotype association. .

3.5 Conclusions

Results from analysis of both line 14 and simulated data show a high degree of concordance between the three models used here. Although it is difficult to quantify advantages, it appears that there is some benefit to analysing data using main effects or haplotype methods. The decision to use multi-marker mapping methodologies in a general linear model framework presents a number of choices in terms of how to use the information contained between a set of markers. Information contained between a set of markers can be defined in terms of main, marginal and interaction effects, and considered as a spectrum of effects, whereby effects can continually be added to a model until all possible parameters contained between a set of markers are included. The current methods used here are the two extremes of this spectrum. Drawing broad conclusions is difficult here, as performance of models is expected to be heavily influenced by localised genetic architecture of the QTL and markers. Determining how these factors influence model performance is explored in more detail in chapters' four to six.

Whilst the haplotype method utilised here has produced promising results, there are some concerns regarding the implementation of the general linear model and score statistics in the "haplo.stats" package. Alternative frameworks need to be explored that allow the incorporation of relationship information, and avoid the pooling of rare haplotypes.

CHAPTER FOUR

FACTORS INFLUENCING THE OPTIMUM MODEL FOR THE ANALYSIS OF ASSOCIATION DATA

4.1 Introduction

Statistical models used for LD mapping can roughly be classified into genotype and haplotype based methods, where haplotype methods are differentiated by their requirement of marker phase (Dekkers *et al.* 2006). Within the closed breeding populations of livestock, LD is generally limited to closely linked loci due to many generations of recombination. Using a measure of LD such as r^2 (Hill and Robertson 1968) will reflect the regression of QTL alleles on marker alleles as the marker associated effect is equal to $r^2(2\alpha)$, where α is the difference between alternative genotypes at the QTL. Therefore, one of the most important factors for designing LD mapping experiments is the degree of LD expected between markers and QTL. However, there is considerable variation in LD among evenly spaced loci. The high variance of the sampling distribution of LD, when based on a single pair of loci, derives from the degrees to which marker genotype groups represent groups that are identical by decent (IBD) with respect to the linked QTL. Clearly, when mapping with a single-locus model, the effective LD between marker and QTL, separated by the same distance, will vary greatly also. Therefore, in some cases using haplotypes will provide stronger differentiation into the IBD groups with respect to the QTL and, hence, be more powerful for LD mapping (Akey *et al.* 2001).

There are broadly two different approaches to inferring haplotype phase for a set of individuals. One approach is to use family data to deterministically resolve phase for genotypes featuring multiple heterozygous loci. However, this information is costly to collect or is sometimes not available. A second approach is to infer haplotypes directly from the population data using a statistical procedure such as an EM algorithm (see, *e.g.* Excoffier and Slatkin 1995; Long *et al.* 1995; Zhao *et al.* 2003). Although this will introduce some uncertainty associated with the inferred haplotypes, Stephens *et al.* (2001) has shown that provided the marker map is sufficiently dense, haplotypes can be constructed with a very high level of accuracy in the absence of family data. Fitting just main effects by regression of SNP genotypes is easier to implement as marker phase does not need to be determined.

A growing number of studies demonstrate that haplotype-based approaches may provide more power and accuracy in locating QTL and causative disease variants than single-locus methods (Akey *et al.* 2001; Martin *et al.* 2000; Morris and Kaplan 2002; Schaid *et al.* 2002; Zaykin *et al.* 2002). However, literature on the relative efficiency of analysing haplotypes verses single markers is complicated by differing assumptions about the number of trait loci, the number of alleles at the trait loci, and the amount of LD between markers and trait loci. If the causative variant is a SNP contained within the marker panel, single-locus tests are expected to be more powerful than haplotype-based tests, when the number of causative SNPs is less than the number of haplotypes (Bader 2001). Morris and Kaplan (2002) suggest that the power advantage for haplotype-based methods is greatest when the marker alleles are not in strong LD with each other, yet haplotypes of those markers are in strong LD with the causative alleles. This situation is likely to occur when the ages of the marker variants are much older than the ages of the alleles at the causative locus, so that the markers have weak LD by the time of origin of the disease susceptibility alleles. How often this occurs is unknown, but the studies suggest that haplotype methods may be more powerful for younger, and hence more rare, causative variants, in contrast to older more common causative variants (Akey *et al.* 2001; Bader 2001; Slager *et al.* 2000). Moreover, another advantage of haplotypes is increased robustness compared to single-marker tests. Evolutionary forces such as random drift, mutation at the marker locus, and varying degrees of initial LD tend to increase the variability of the observed magnitude of LD between any single marker and disease loci leading to complex patterns of association, even with tightly linked markers. In this situation, simultaneous analysis of multiple markers in the form of haplotypes can result in comparatively simpler patterns of LD (Akey *et al.* 2001).

The comparison of the above examples is limited by the fact that they focus on two extremes, either using the maximum single-locus statistic or a global test for all interactions between multiple markers. The choice of using multiple markers to test for association is made more difficult when the information contained between a set of markers to include in a test is considered. Haplotype analyses are the extreme version of multi-marker tests in that they fit all main effects, interactions and phase parameters (Clayton *et al.* 2004). This can lead to large numbers of parameters (haplotypes) being fitted in a model, which can weaken the power to detect associations. Using multiple SNPs in a genotype-based test that simultaneously tests for the main effects of all loci, yet without regard to haplotype phase, is a natural extension to a single-locus analysis. Clayton *et al.* (2004) and Chapman *et al.* (2003) have questioned the value of additional information afforded by haplotyping markers, and suggested that in certain situations using a multilocus genotype based test will be more powerful than analyses based on haplotypes. If a number of adjacent SNPs are interrupted by a recombination event, then some additional information can be gained by looking at the haplotypes rather than markers individually. The extent of recombination between these markers is related to the amount of additional information gained from haplotypes over single markers. When using these markers to test for association, if there is a modest amount of recombination, any gain in variance explained by scoring haplotypes of markers will be more than offset by the additional DF. In this situation a genotype main-effects model is expected to do as well if not better than a haplotype model as the majority of additional variance will be explained by the main-effects (Chapman *et al.* 2003; Clayton *et al.* 2004). However, if we introduce more recombination events into this region, main-effects models will be unable to explain large proportions of variance between markers as higher order interactions become more relevant. Here a haplotype model is expected to have higher power, as the gain in variance explained by enumerating all orders of interactions more than compensates for the extra DF fitted. It has also been suggested that when using the simplified genotype-based multilocus test, slightly more markers are required to fully capture the information of the region (Seltman *et al.* 2001). Unfortunately, empirical values are unknown for these situations, meaning more comprehensive evaluations of the relative efficiencies of single-locus, multilocus genotype and haplotype-based methods are required to clarify these issues.

Grapes *et al.* (2004) and Zhao *et al.* (2007) compared single-marker regression, regression on marker haplotypes and an IBD mapping approach (Meuwissen and

Goddard 2000) for power and precision of QTL mapping using simulated datasets. The conclusion from these papers was that single-marker regression gave greater power and precision than regression on marker haplotypes, and was comparable to the IBD method. However, both Grapes *et al.* (2004) and Zhao *et al.* (2007) were simulating situations where single-markers had very high r^2 values with the QTL; the average r^2 was 0.41 for markers within 0.5 cM and 0.15 for markers within 1.5-2 cM. In situations of high marker-QTL LD, haplotypes are expected to add noise to the estimation of the QTL effect, and reduce power of the test statistic. Furthermore Grapes *et al.* (2004) assumed that a requirement of haplotype analysis would be the collection of additional marker genotypes on relatives, to allow haplotype construction. They therefore compared results from single-marker analysis with twice as many markers as used in the haplotype analysis. These results contradict those of Hayes *et al.* (2007), who found that in real data (9323 SNPs genotyped in Angus cattle) the increase in QTL variance explained from using marker haplotypes more than compensated for the decrease in accuracy of estimating a greater number of haplotype effects. Average marker-marker and marker-QTL r^2 values were low in this Angus population, which is a possible explanation for contradictory results with Grapes *et al.* (2004) and Zhao *et al.* (2007). These studies have not differentiated the ability of models to analyse data with different levels of QTL MAF and marker-QTL LD. This is expected to make a considerable difference to the abilities of models to predict QTL genotypes (Chapman *et al.* 2003; Clayton *et al.* 2004). In the context of genomic selection, Calus *et al.* (2008) demonstrated that the advantage of haplotypes over single markers, to accurately predict QTL effects, decreased as the r^2

between adjacent markers increased. At r^2 = 0.215 between adjacent markers, the haplotype and single-marker approaches gave very similar accuracies.

The main objectives of the analyses in this chapter were to determine the influence of genomic situations such as marker-QTL LD, QTL minor allele frequency and marker to QTL distance on the performance of single-marker, main order effect multilocus and haplotype models to estimate variance and provide statistical support for a QTL. We used the extent of LD in a broiler chicken dataset, consisting of 12046 genome wide SNPs genotyped in 200 individuals to simulate LD between markers and a QTL by selecting a single SNP to act as a surrogate QTL (sQTL), and then determining the ability of models to predict the variance of the sQTL using the surrounding markers. In total 6300 sQTL were generated, representing a range of MAF (0-0.5).

4.2 Materials and methods

4.2.1 Dataset

Aviagen Ltd provided genotype data for a single line of commercial broiler chickens consisting of 200 individuals. Individuals were genotyped for 12046 SNPs that were chosen to cover the genome from the 2.8 million SNPs that were identified in the chicken genome-sequencing project (Wong *et al.* 2004). The SNPs covered the genome, but were not spread evenly, with a mean marker spacing of 0.13 Mb, and a standard deviation of 0.31. Individuals used in this study were from a commercial population under selection, comprising of a complex pedigree structure, with a few small half-sib groups. In these cases, care was taken to ensure that no more than three animals were selected from each sire group to avoid over representation of sire haplotypes. These data were used to simulate the LD between markers and a QTL by selecting markers to act as surrogate QTL (sQTL), and then using the surrounding markers to test for association with four different models: a single-marker regression, a main effects genotype model using three marker windows and two haplotype methods, also using three marker windows. Markers were removed from the dataset if they had any missing genotypes, were represented by less than three genotype classes or had MAF less than 0.01.

To avoid discrete classes representing phenotypes, values randomly drawn from a normal distribution were added to genotype of each individual. The variation of this distribution in relation to the variation of the sQTL genotypes represents the heritability of the "trait". The appropriate heritability of the "trait" to use was determined using a series of permutations with the different models, a range of heritabilities and sQTL with different MAF. High heritability of the trait can result in inflated permutation thresholds for sQTL with low MAF because of the skewed distribution of the phenotype. Therefore, heritability needed to be low enough such that the permutation thresholds remained constant across the range of sQTL MAF for all models, whilst being high enough for models to have a reasonable power to detect the sQTL. A comprehensive series of permutation analyses were run for each of the models, using sQTL with a range of MAF from 0.012-0.5, and a series of heritabilities from 0.01-1. From these permutation analyses a heritability of 0.3 was chosen for the sQTL, as this was the highest value at which the permutation threshold remained constant for all models across the entire range of sQTL MAF. For a given sQTL the variation of the distribution from which values were randomly added to genotypes was determined under the following formula; $\sigma^2_{noise} = (\frac{1}{0.3}\sigma^2_{sQTL\_geno}) - \sigma^2_{sQTL\_geno}$, where σ^2_{noise} is the variation of the normal distribution, and $\sigma^2_{sQTL\_geno}$ is the variation of the sQTL genotypes. Distributions of *p*-values from permutation runs were checked against the expected null distribution for each association model. An example of the Q-Q plots is given in appendix three.

4.2.2 Simulation of sQTL

Each marker in turn was chosen to represent the sQTL, and 25 markers on either side then formed a test region surrounding the sQTL. Test regions of 50 markers were chosen to represent the mean distance at which syntenic LD measures reached background levels (chapter two). To avoid testing markers against sQTL on different chromosomes, the first and last 25 markers on a chromosome were not used as sQTL, meaning that no sQTL were chosen on micro chromosomes with less than 50 markers. After removal of fixed and two-genotype class markers, this left 6300 markers that were chosen to represent sQTL. The four models were tested against each sQTL using the 50 markers in the test region for that sQTL.

The extent of marker-marker LD in the data was used as an indication of the extent of marker-QTL LD we would expect in a typical data set. Using a 'real' rather than a simulated dataset provides a more realistic pattern of structural genetic variation and avoids the difficulty in accurately simulating IBD structures and population genetic parameters (Hoggart *et al.* 2007). The parameter r^2 is an estimate of LD that describes the proportion of QTL variance that would be explained by a marker if one of the markers were actually a QTL (Hill and Robertson 1968). To determine the extent of LD within this population, r^2 was calculated (following Hill and Robertson 1968) for all possible syntenic marker pairs and plotted as a function of distance (figure 4.1 a and b).

4.2.3 Single-locus model

Within a test region, each individual marker was regressed against the sQTL phenotypes using the following model;

$$y_i = \mu + \beta x_i + e_i$$

Where $\boldsymbol{y_i}$ is the "phenotype" of the sQTL for individual *i*, $\boldsymbol{x_i}$ is the number of "1" alleles carried by individual *i* at the SNP; β is the substitution effect for the SNP and $e_i \sim N(0,\sigma_e^2)$ is the residual for the i^{th} individual. Individual markers are fitted in a design matrix coded as 0, 1 or 2 for 1-1, 1-2 and 2-2 allele combinations respectively. In this model, the effect of the marker is fitted as a fixed effect, assuming an additive model of inheritance. Association was tested against a null hypothesis of $H_0 = \beta = 0$, where β is the effect of the marker, using an *F*-test with one degree of freedom.

4.2.4 Main effect model

The linear model fitting a single-marker was extended to fit adjacent markers in a multiple regression framework. A window of three-markers was scrolled across the test region, analysing associations, before moving forward a single marker and repeating the process. Markers representing sQTL were "dropped" from the marker panel, and therefore were not among the markers within a window when the window covered a sQTL position. In the main effects analysis, markers were fitted as linear covariates testing the main effects in the flowing model,

$$y_i = \mu + \sum_{i=1}^{n} \beta_n x_{n,i} + e_i$$

Where β_n is the substitution effect for SNP *n*, and $\boldsymbol{x_{ni}}$ is the number of copies of "1" allele carried by individual *i* at SNP *n*. Association was tested against a null hypothesis of $H_0 = \beta_1 = \beta_2 = \beta_3 = 0$, where β_1, β_2 and β_3 are the effects of the markers, using an *F*-test with 3 degrees of freedom.

4.2.5 Haplotype regression model

Since haplotype data are not readily available for genome-wide LD screens, haplotype analyses are conducted in a two-stage procedure; firstly, haplotypes were inferred for overlapping three-locus windows using an EM approach. For individuals with ambiguous phase all haplotype pairs consistent with the observed genotypes are provided, along with the posterior probabilities for each pair. In order to reduce parameters fitted in the model one commonly used approach is to take just the highest probability pair for an individual and fit those (Niu 2004). An alternative approach is to model for haplotype uncertainty by fitting probabilities in the regression analysis. Secondly, the probabilities of the inferred haplotypes were fitted as linear covariates in an *N*-dimension regression model, where *N* is the number of estimated haplotypes in the three-marker window, given the observed genotypes.

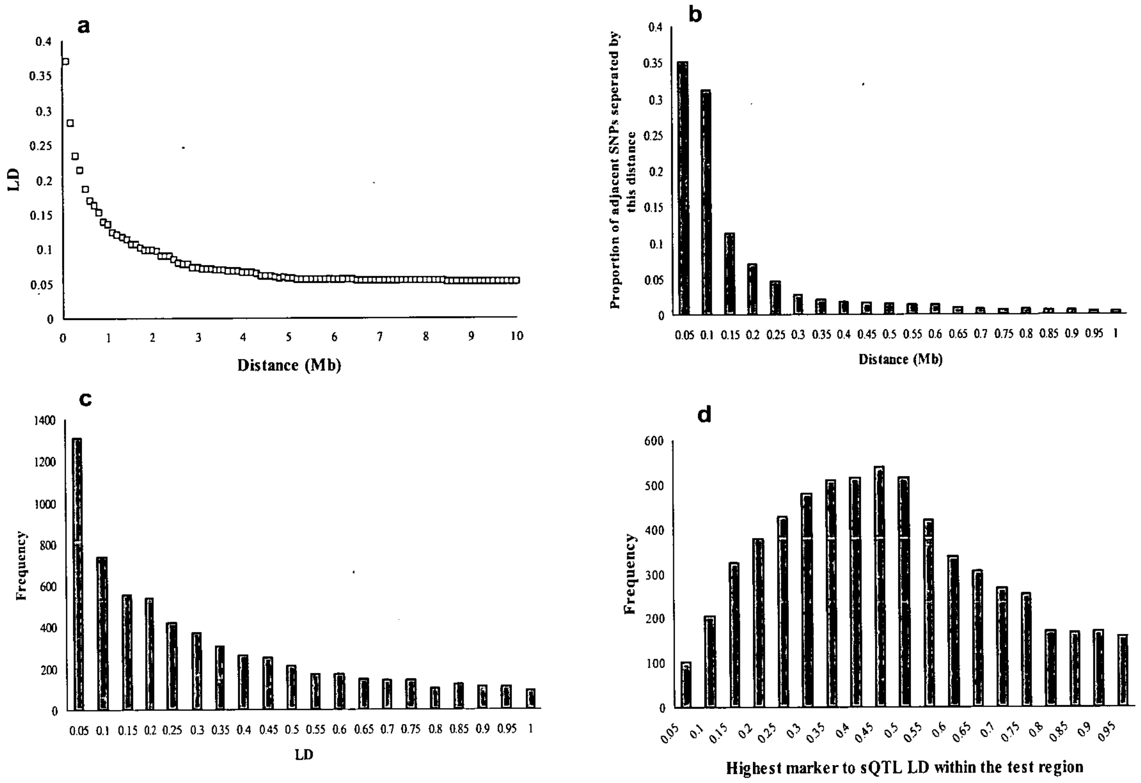


Figure 4.1

a) The decline in mean r^2 for SNP pairs within bins of 100 kb between SNPs. **b)** Distribution of the distances between adjacent SNP pair used as markers and sQTL in the analyses. Distances have been separated into equal sized bins of 0.05 Mb. A small proportion (0.02) of markers are separated by greater than 1Mb. **c)** Distribution of r^2 values between adjacent SNPs. **d)** Distribution of r^2 values of markers in the highest LD with the sQTL in each test region. Values plotted are the proportion of SNP pairs with r^2 values in bins of 0.05.

4.2.5.1 Forming haplotypes

The EM algorithm is used extensively to estimate haplotype frequencies and infer phase for population based studies (Excoffier and Slatkin 1995; Long *et al.* 1995 Stephens *et al.* 2001). EM estimates population haplotype probabilities based on maximum-likelihood given observed genotype frequencies. Across the test region each three-marker overlapping window was haplotyped and then those haplotypes used in regression analyses. sQTL were not used to estimate any haplotypes, having been "dropped" from the marker panel.

The algorithm described below was written and utilized using the programming language R, and is loosely based upon the progressive insertion algorithm implemented in the "snphap" software (Clayton 2009). Original code is given for the 2 haplotyping approaches are given in appendix four. The algorithm attempts to find the value of haplotype frequencies that gives the joint maximum likelihood given observed genotypes for n individuals as

$$L(G \mid F) = \prod_{i=1}^{n} \Pr(G_i \mid F)$$

where G is the observed genotypes, F is the set of population haplotype frequencies and $\Pr(G_i \mid F)$ is the probability of the i^{th} individual's genotypes given haplotype frequencies and assumption of HWE.

$$\Pr(G_i \mid F) = \sum_{u=1}^{h} \sum_{v=1}^{h} c_{uv}^{i} fh_u fh_v$$

where G_i denotes the unphased genotypes for the i^{th} individual, F is the haplotype frequency, fh_u and fh_v are the frequencies of haplotypes u and v respectively and c^i_{uv} is the haplotype-genotype compatibility index for individual i

$$c^i_{uv} = \begin{cases} 1 & (h_u // h_v => G_i) \\ 0 & (h_u // h_v \neq> G_i) \end{cases}$$

.On convergence the probabilities of all pairs of haplotypes for an individual given the observed genotypes can be estimated based on the maximum likelihood estimates of the haplotype frequencies. During the EM iteration stages pairs of haplotypes with probabilities less than the threshold $1e^{-9}$ were dropped from the analysis and the frequencies of the remaining haplotypes were recalculated. The posterior probability for haplotype pair h_u,h_v for the i^{th} individual is given as

$$\Pr(h_u, h_v \mid G_i, F) = \frac{\Pr(h_u, h_v \mid G_i) * \Pr(F)}{\sum_{(u,v)} \Pr(h_u, h_v \mid G_i) * \Pr(F)}$$

where, $\Pr(F) = \Pr(h_u) * \Pr(h_v) = fh_u * fh_v$. The second stage is to take the posterior probabilities for pairs and fit them in an X matrix of a regression model to test for association against the sQTL.

Here two different haplotype models were analyzed, using the probabilities of haplotype pairs differently to populate the X matrix fitted in this model

$$y_i = \mu + \sum_{i=1}^{N} \beta_N x_{N,i} + e_i$$

Where β_N is the substitution effect for haplotype N, and $x_{N,\,i}$ is the probability of carrying haplotype N by individual i. Association was tested against a null hypothesis of

$H_0 = \beta_1 = \beta_2 \ldots \beta_N = 0$, where $\beta_1, \beta_2 \ldots \beta_N$ are the effects of the haplotypes, using an F-test with N-1 degrees of freedom. The design matrix $\boldsymbol{X}$ relates estimated haplotype probabilities for each individual to their value in the $\boldsymbol{Y}$ matrix. The two haplotype approaches differ in their formation of the X matrix for individuals with a haplotype pair of ambiguous phase.

4.2.5.2 Hap_highest_prob

Individuals with ambiguous phase have a number of possible haplotype pairs that are consistent with the observed data, each of which has a posterior probability. For a given individual the pair with the highest probability is considered the most likely and treated as an unambiguous pair. Consider the following example of three individuals:

Individual	Haplotype pairs	Pr	Phase
y_1	h_1 / h_1	1.0	Unambiguous
y_2	h_1 / h_3	1.0	Unambiguous
y_3	h_1 / h_4	0.25	Ambiguous
	h_2 / h_4	0.52	
	h_1 / h_2	0.23	

For this example the X matrix would be;

$$\begin{array}{c|cccc|} & h_1 & h_2 & h_3 & h_4 \\ y_1 & 1 & 0 & 0 & 0 \\ y_2 & 0.5 & 0 & 0.5 & 0 \\ y_3 & 0 & 0.5 & 0 & 0.5 \end{array}$$

for individuals' y_1 and y_2 the probability $\Pr(G_i \mid h_k h_j)$ is either 0 or 1, and so values for the X matrix for haplotypes incompatible with the ith subject genotype are equal to 0, and to 0.5 for heterozygous pairs or 1 for homozygous pairs when haplotype identification is certain. For individuals with ambiguous phased pairs, such as y_3, the pair with the highest probability is taken to be observed as unambiguous, and the X matrix entries are also 0.5 or 1 for heterozygous and homozygous pairs respectively. The remaining haplotypes are given a value of 0 in the X matrix. Haplotype pairs with low probabilities are discarded which can reduce the overall number of haplotypes fitted, and thus DF, but can introduce some error in estimating haplotype effects through not taking into account haplotype uncertainty (Morris *et al.* 2004).

4.2.5.3 Hap_all_prob

This method accounts for the uncertainty of ambiguous haplotype phase by fitting the sum of probabilities of haplotypes in pairs for each individual (Zaykin *et al.* 2002). This will usually lead to an increase in the N dimension of the X matrix over the Hap_highest_prob model through the inclusion of rare haplotypes with lower probabilities. Using the same example as above the X matrix for this dataset would be

$$\begin{array}{c} \\ y_1 \\ y_2 \\ y_3 \end{array} \begin{array}{c} \begin{array}{cccc} h_1 & h_2 & h_3 & h_4 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0.24 & 0.375 & 0 & 0.385 \end{bmatrix} \end{array}$$

For individual $y3$, more than one pair of haplotypes is consistent with the observed genotypes, and values in the X matrix are the posterior probability of pair h_k / h_j. If pair h_k / h_j is heterozygous then its probability is multiplied by 0.5. If an individual has a haplotype that is observed in more than one pair, then its probabilities are summed together in the $\boldsymbol{X}$ matrix. Therefore, the sum of probabilities for all haplotypes observed for the ith subject is equal to one.

4.3 Results

4.3.1 Marker spacing and LD

The average distance between adjacent SNPs was 130 kb; although there was considerable variation in that value, with many SNPs separated by considerably shorter distances (figure 4.1a). The uneven spacing of markers reflects the method of SNP discovery and factors that governed the design of the marker panel. In this population, average LD between adjacent markers is modest and declines slowly with distance (figure 4.1b). This slow decline is likely due to the limited effective population size. The average value of r^2 for adjacent markers is 0.31, and the average of the highest sQTL to marker r^2 within the test region is 0.42. The distribution of r^2 values for adjacent markers and the highest sQTL to marker within the test region is shown in figures 4.1c and d respectively.

4.3.2 Comparison of results under different sQTL MAF

In traditional mapping studies, allelic frequencies of causal variants are unknown, although they are expected to have considerable effect on the power of models to provide statistically significant support for genetic associations. To investigate the effect of sQTL allele frequency on the ability of the models to predict their variance, the sQTL were divided into ten MAF bins, ranging from 0 to 0.5. The distribution of sQTL in each bin is given in figure 4.2. The sQTL MAF bins show a roughly uniform distribution, but with an over-representation at intermediate frequencies compared to the assumed neutral U-shaped distribution, which likely represents the ascertainment bias associated with SNP discovery and marker selection (Solberg *et al.* 2008).

The ability of a model to provide statistical support for a sQTL is related to the interaction between the variance explained and its parameterization. Therefore, *p*-values from the hypothesis test of each analysis are used to evaluate the ability of models to predict the genotypes at the sQTL. The *p*-value provides an accurate means for assessing performance between models that differ in their parameterization, as they account for the expected additional variance explained by a parameter, by drawing from distributions adjusted for specific DF.

Across the test region of a sQTL, the markers that provided the smallest nominal *p*-value for each model were deemed to have performed best at providing statistical support for the sQTL. The markers that provided these values were termed 'model best'. The *p*-values from 'model best' markers are ranked to determine the model that has

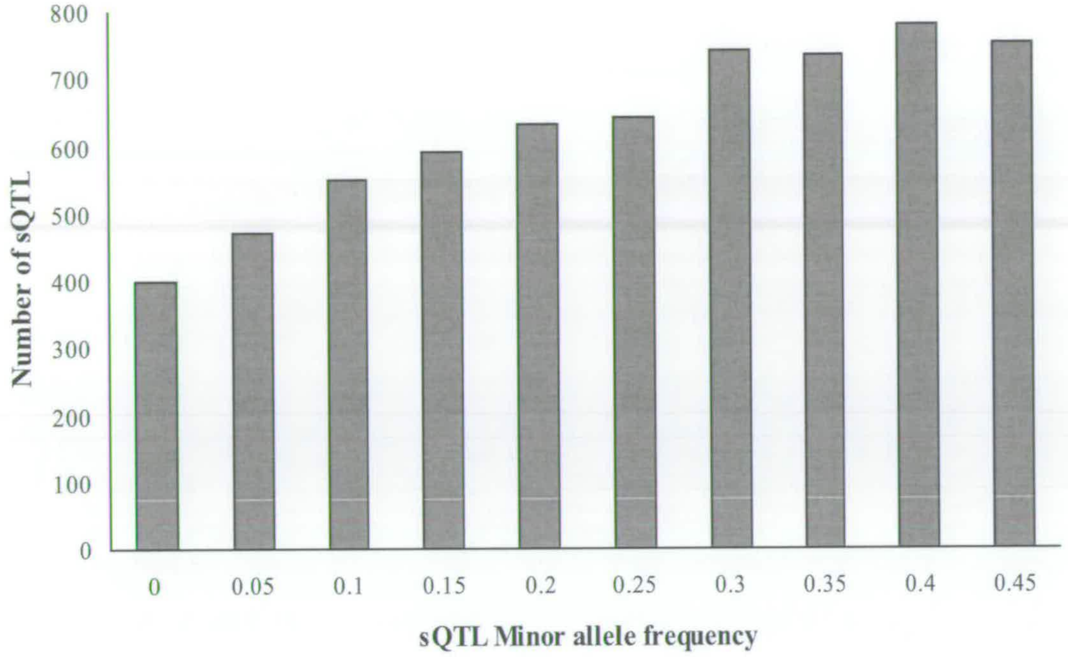


Figure 4.2

The frequency distribution of sQTL within each MAF bin. Each MAF bin had an equal range, with the values shown on the *x*-axis representing the lower bound of this range. The total number of sQTL was 6300.

performed best for that given sQTL, with these termed "best overall". The proportion of times each model is the 'best overall' differs across the sQTL MAF bins (figure 4.3).

Power of the statistical models is related not just to the amount of variance of a QTL that they explain, but also to the number of parameters that they need to explain this variance. The proportion of sQTL variance explained by the markers will increase as additional markers or interaction parameters are included in analyses, although each of these will add DF fitted in the hypothesis test, thereby reducing the power of the test statistic. Performance of a model can therefore be considered along the principals of most variance explained using the fewest DF. As you move from single-locus to main effect and then haplotype models, the extra effects fitted need to explain enough additional variance to compensate for their extra DF. For each of the four models the change in mean proportion of variance explained and DF fitted by the 'model best' markers across the sQTL MAF bins is shown in figures 4.4a and b, respectively. The mean proportion of variance explained by the 'model best' markers increases as the MAF of the sQTL increases, along with the mean DF fitted for the haplotype models. However, the ratio of variance explained to degrees of freedom fitted changes disproportionately for the four models (figure 4.4d). For sQTL in the lowest MAF bin, the Hap_all_prob 'model best' markers explain an average of 0.22 variance with a mean of 5.3 DF, a ratio of 0.041 variance explained for each DF. In the highest MAF bin, the Hap_all_prob variance to DF ratio is 0.037, a decline in the variance that each DF explains. For the Hap_highest_prob model this ratio remains approximately constant. Both the main effect and single-marker models have fixed numbers of DF, therefore,

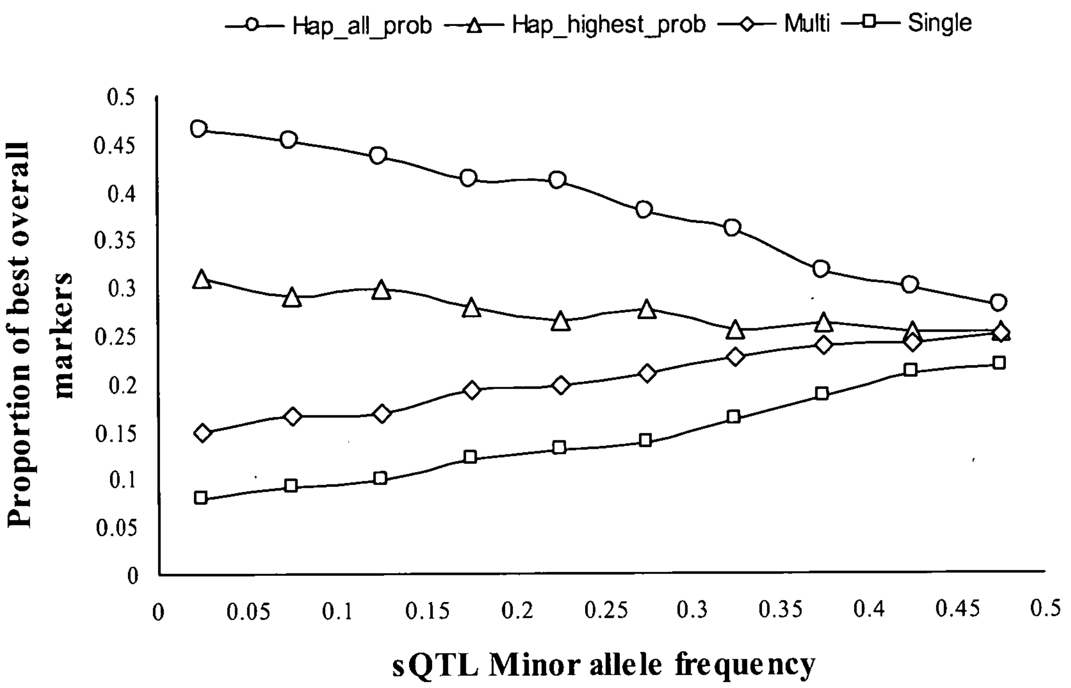


Figure 4.3

Proportion of times each model provided the 'best overall' p-value for sQTL across the MAF bins. Within each MAF bin the proportions for all models will sum to one. Values shown on the x-axis represent the lower bound of each MAF bin, with the proportions aligned to the median position.

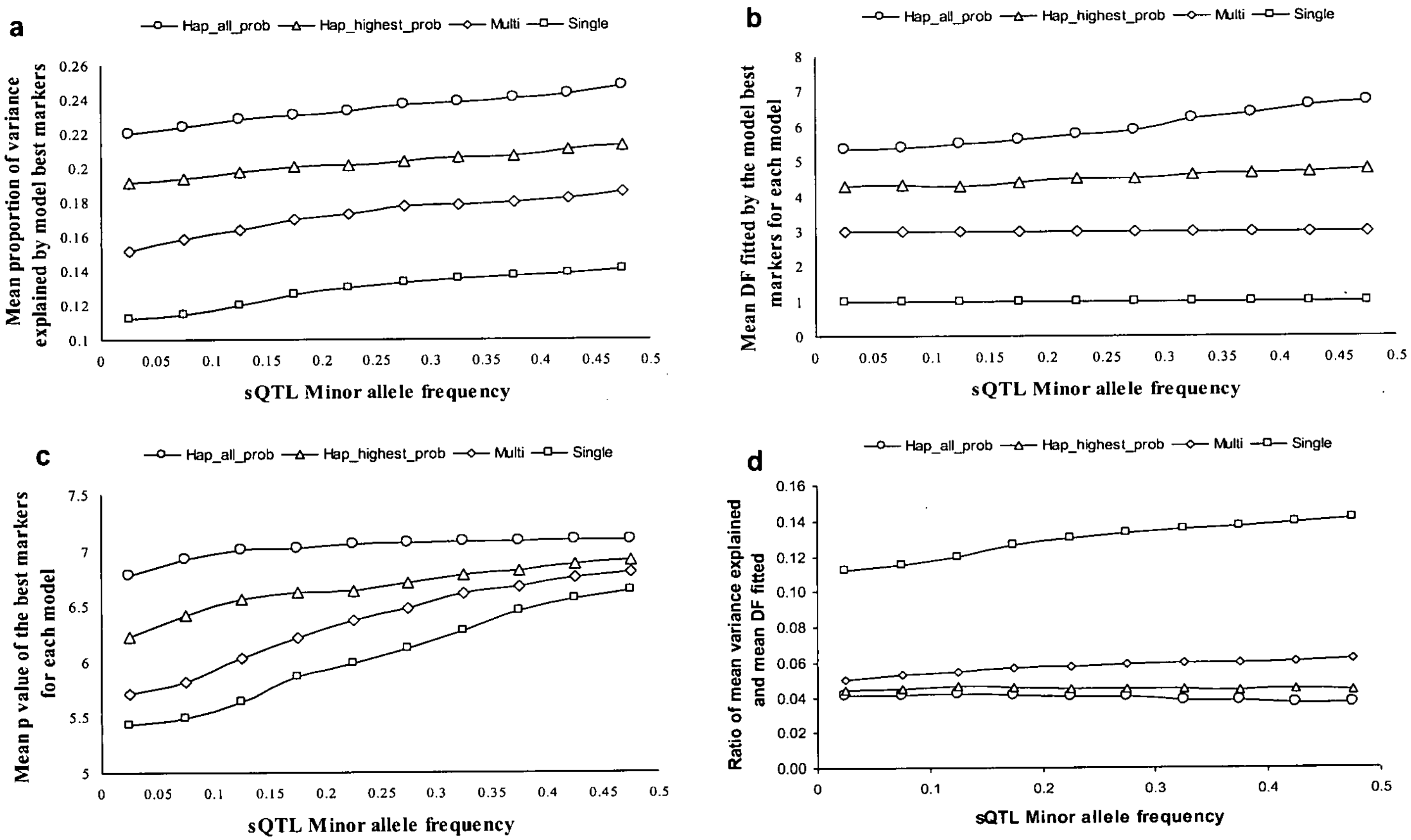


Figure 4.4

a) The mean proportion of sQTL variance explained by the 'model best' markers for each method across the sQTL minor allele frequency bins. **b)** The mean degrees of freedom fitted in the analyses that produce the 'model best', across the sQTL minor allele frequency. **c)** The corresponding mean *p*-values for these analyses. **d)** Ratio of mean variance explained for each DF fitted by the 'model best' markers. This is the ratio of points on figures **a** and **b**.

their ratios of variance explained for DF fitted increase at a proportional rate to the change in mean variance explained at different sQTL MAF (figure 4.4d).

Unequal ratios of variance explained to DF fitted for the four models across the sQTL MAF bins result in mean *p*-values for the 'model best' markers changing relative to one another. The pattern of unequal ratios is reflected in changes to the mean –log10 *p*-values from the 'model best' markers across the range of sQTL MAF (figure 4.4c).

4.3.3 Distance of the 'model best' markers from the sQTL

In traditional QTL mapping studies we assume that the markers that provide the smallest *p*-value are closest to the causal variant, or in the case of multi-marker approaches, surrounding the causal variant. In reality, because of high variation in LD patterns over large distances, our highest test statistic may come from markers some distance away from the causal variant. For single-locus tests, there is the possibility that this situation could occur more commonly than for multilocus tests, as the highest statistic will come from the marker in highest LD with the QTL. For multi-marker approaches this may not necessarily be the case as the smallest *p*-value will be a compromise between the amount of variance explained by parameters and the number of DF fitted in the model. Knowledge of the sQTL position means the ability of the models to provide accurate support for the sQTL location can be assessed by the position of the 'model best' markers relative to the sQTL. For the single-marker analysis, the most accurate mapping of the sQTL will come from identification of markers flanking the sQTL, whilst for the multiple marker methods it will come from windows covering the sQTL (figure 4.5a). Figure 4.5b shows the mean distance of the 'model best' markers for each method from the sQTL across the MAF bins. For the multiple marker methods, this distance is calculated from the centre of the marker window.

4.3.4 Variance explained and the LD structures

Individual markers in high LD with a causative variant will explain a high proportion of their genetic variation; therefore, it is expected that adding additional markers or parameters to a test will not explain much additional variance as a proportion of what has already explained. In these situations, there is the possibility that the additional DF added will undermine the power of the test statistic. When markers are in low LD with a causative variant adding additional markers and parameters to a test has the possibility of dramatically increasing the variance explained compared to the proportion explained by a single marker. In light of this, patterns of LD between markers and sQTL are expected to have a considerable impact on the performance of the four models. This situation is made more complex by the fact that markers in low LD with one another are expected to produce a greater variety of haplotypes than markers in high LD with each other. To investigate these relationships, for each model the LD between markers and the sQTL (for multiple-marker methods the mean of the three pairwise measures was calculated) was taken from every analysis against all sQTL and divided into 100 equal bins. The average variance explained by analyses in each of these bins was calculated and plotted against LD (figure 4.6a). Proportional change in variance explained by models varies dramatically across the range of marker-sQTL LD.

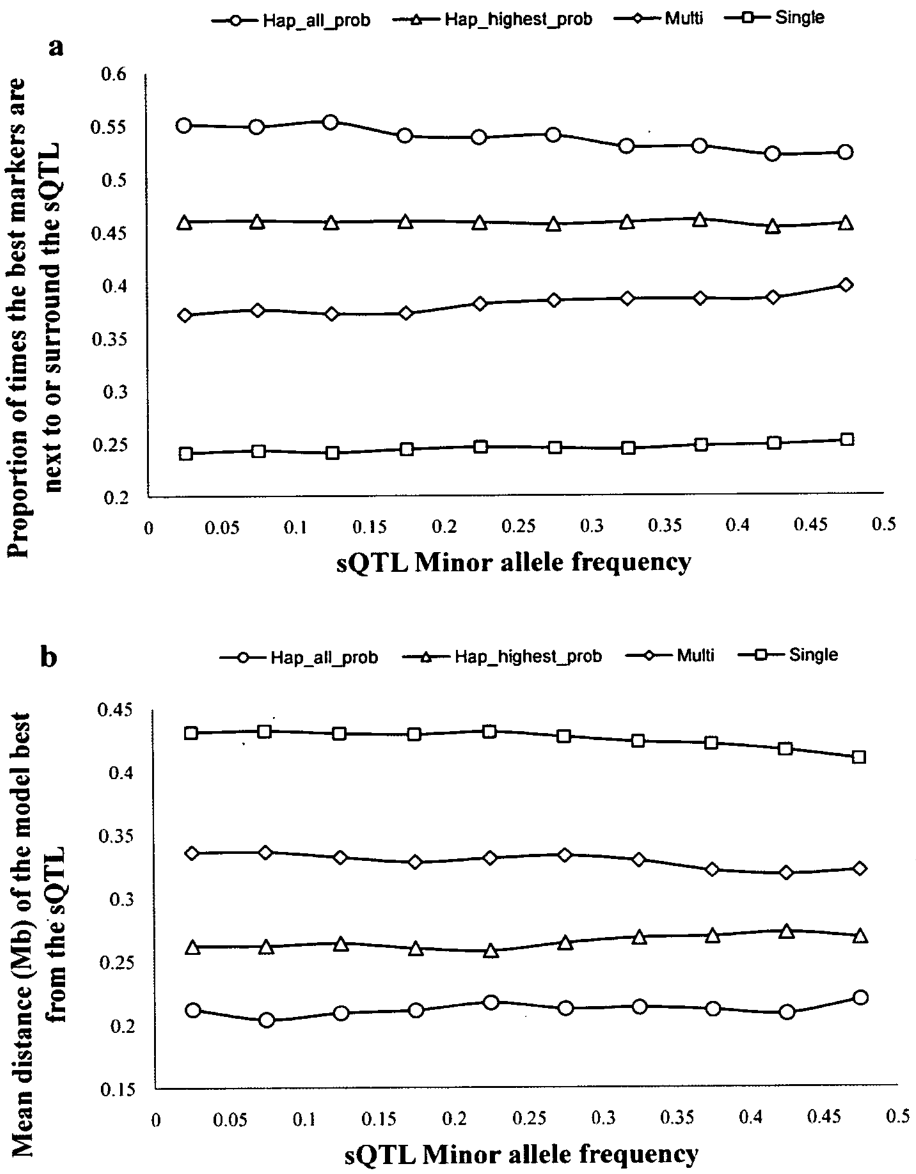


Figure 4.5

a) The proportion of times that the 'model best' markers for each method either surround the sQTL, in the case of multilocus methods, or are the closest marker either side of the sQTL, in the case of the single-locus method. Both situations occur twice across each test region. **b)** The mean distance of the 'model best' markers from the sQTL. Distance for multilocus models is taken from the centre marker in the window.

This is demonstrated when we look at the frequency distribution for LD between markers and sQTL for the 'best overall' markers across all sQTL (figure 4.6b). The haplotype models that performed 'best overall' tend to have low levels of LD between the markers and sQTL, whilst the markers from the 'best overall' single-marker models are in strong LD with the sQTL.

Given the differential performance of the models under different genomic situations, it was hoped that information that can be observed, such as between marker LD, could be used as a predictor of model performance. For each of the multilocus methods the mean pairwise LD of the test markers was regressed against the variance that they explained. For each of the methods this relationship was very poor, with regression coefficients of 0.08, r^2=1e^{-3} (main effect), 0.06, r^2=1.2e^{-3} (Hap_highest_prob) and 0.05, r^2=3e^{-4} (Hap_all_prob) respectively. These results suggest that using just observed LD between pairs of markers cannot be used as an indicator of model performance. However, here LD from the whole test region was used rather than just markers close to the sQTL.

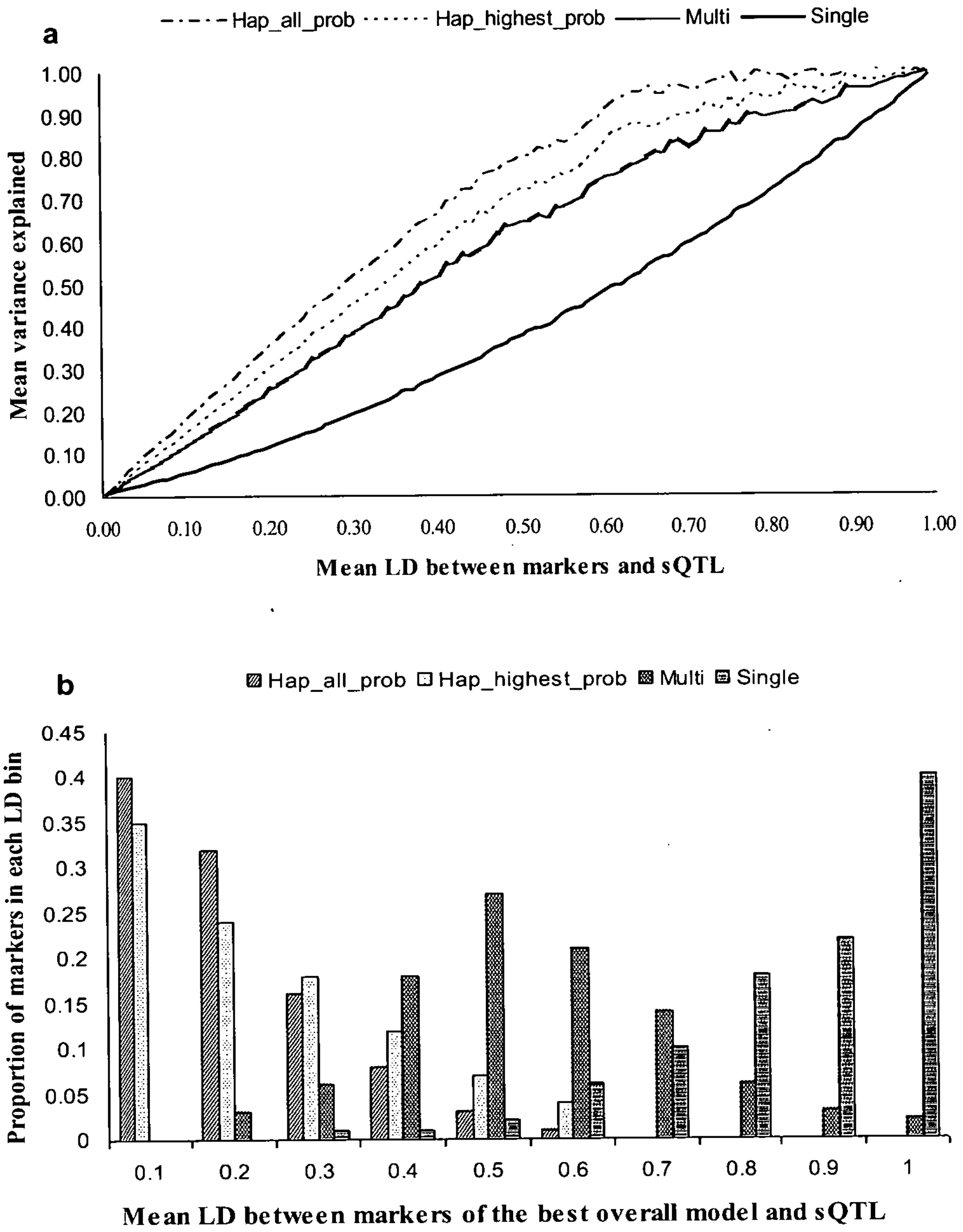


Figure 4.6

a) Mean variance explained by all each model for every analysis, across the range of marker-sQTL LD. For multilocus methods, the measure of LD is the mean from the three pairwise measures. **b)** Proportion of 'best overall' analyses' marker-sQTL LD. For multilocus methods, the measure of LD is the mean from the three pairwise measures.

4.4 Discussion

Our results shown here clearly demonstrate that when mapping for sQTL across the range of MAF the Hap_all_prob model performs best on average, whilst the single-marker approach has the worst performance relative to other models. Performance of the models relative to one another varied considerably when tested against sQTL of differing allele frequencies (figure 4.3). When mapping against sQTL with low MAF there is a clear advantage of using a haplotype model that accounts for uncertainty in phase, such as Hap_all_prob over the other methods tested here. Although this advantage remains across the range of sQTL MAF its magnitude diminishes with more intermediate sQTL MAF. However, care should be taken when interpreting this figure, as performance of the models is relative to one another and does not reflect actual differences in statistical significance. In other words, how much difference is there in *p*-values between models? The 'model best' markers for all models explain higher proportions of variance as the sQTL MAF values increase (figure 4.4a), although, for haplotype analyses, this is coupled with an increase in the mean DF fitted by these models (figure 4.4b). In order to perform well against single-locus models, those with increased parameterization need to explain enough additional variance of the sQTL to compensate for their extra DF. Here we show that the ability to meet this compensation point is not constant across the range of sQTL MAF (figure 4.4c). Taking the ratio of the mean variance explained and DF fitted for the 'model best' markers over the range of sQTL MAF (figure 4.4d), and cross-referencing that with the mean –log10 *p*-values from those analyses, provides an indication of the positions of these compensation points for the various models. When mapping against sQTL with MAF between 0-0.05, the Hap_all_prob 'model best' markers produce considerably smaller *p*-values than other methods with a ratio of variance explained to DF fitted of 0.044. At intermediate sQTL MAF there is little difference between *p*-values provided by the different methods, whilst at low MAF values this difference is very dramatic, suggesting that there is a considerable advantage of using the Hap_all_prob model when mapping for QTL with expected allele frequencies in this range.

For the majority of genome-wide association studies the allele frequencies of causative variants are unknown. However, under assumptions of neutral mutation or stabilizing selection models, the distribution of QTL allele frequencies is expected to resemble a U-shaped distribution, producing a high proportion of QTL with low MAF (Lynch and Hill 1986; Wright 1935). Therefore, unless prior information of QTL allele frequencies is available, the use of a haplotype model that accounts for phase uncertainty is a preferable approach to use for genome-wide association studies.

Detecting genetic variation of causal variants that have uncommon or rare alleles, with sufficient power, is currently a problem for the majority of mapping studies in livestock and humans (Bodmer and Bonilla 2008). Here we have shown that there is a considerable advantage in using a high parameterization model, such as Hap_all_prob, to provide statistical support for QTL with low MAF. The same conclusions have been drawn from a study of similar design, on a non-pedigreed human population (James Floyd, personal communication). This suggests there may be an advantage to revisit analyses that have struggled to identify genetic variation using single-locus methods, and apply multilocus interaction models instead.

For the majority of sQTL the proportional increase in variance explained by haplotype models more than compensates for the additional DF added (figure 4.4c). These results concur with those of Pe'er *et al.* (2006), who used empirical genotype data from the human international Hap-Map project to evaluate the extent to which the sets of SNPs contained on three whole-genome genotyping arrays capture common SNPs across the genome. They concluded that limited inclusion of specific haplotype tests in association analysis can increase the fraction of common variants captured (as evaluated by r^2 between haplotypes and common variants) by 25-100%. However, these specific tests were based on pre-selection of "tagging SNPs" which capture 90% of the variation in SNP genotypes in a defined chromosome region. Use of tagging SNPs reduces the number of effects that need to be estimated compared with haplotypes, increasing the power of the test. De Bakker *et al.* (2005) compared the power of exhaustive haplotype search and single SNP analysis to detect a QTL, where power was a function of r^2 between haplotypes or single marker and the QTL. They found that the use of haplotypes only increased power if the MAF of the sQTL was less than 0.05; otherwise, the use of haplotypes actually decreased power. Although sQTL were only split into two bins, those with MAF below 0.05 and those above 0.05.

Differences between sQTL MAF thresholds for performance of haplotype models shown here and by De Bakker *et al.* (2005) can possibly be explained by two main factors. Firstly, the approaches taken by De Bakker *et al.* (2005) focus on case-control situations and the identification of haplotypes that tag SNPs. To determine the haplotype that tags a SNP best, haplotypes are sequentially tested individually and their performance is set using an empirical threshold. Our multilocus models test for the global effects of markers or haplotypes, which, whilst adding DF, are expected to explain more real variance of the sQTL without imposing an empirical threshold from multiple testing, as would be done with a series of 1-DF tests (Schaid 2004). However, if a single haplotype is strongly associated with the trait, testing each haplotype individually can potentially be more powerful than a global test that spreads association across multiple haplotypes (Schaid 2004). Secondly, in human data the density of SNPs is very much higher than in our data. Even accounting for the increased effective population size of humans relative to QTLs, the average level of LD between adjacent SNPs is very much greater in humans.

4.4.1 Marker density and LD

Marker density, or more accurately, the extent of LD between markers, is of considerable importance to the performance of different association models. We can consider this in terms of the proportional increase in variance explained by models across different values of LD between markers and the QTL. Calus *et al.* (2008) suggested that the advantage to using haplotypes, derived using an IBD haplotyping method (Meuwissen and Goddard 2001), increased at lower marker densities, although when r^2 values between adjacent markers were above 0.2 there was little advantage in using haplotypes. We have shown an advantage of using haplotypes persists when using data with a mean r^2 between adjacent markers of 0.31, although the magnitude of the advantage is not consistent across the whole range of QTL MAF. Our results are consistent with those of Hayes *et al.* (2007), who also concluded that there was an advantage of mapping with haplotypes composed of four markers, using an approach similar to the Hap_highest_prob, in a population with average r^2 between adjacent markers of 0.1. Zhao *et al.* (2007) compared several methods for LD-based QTL fine mapping: regression of SNP genotypes, regression of SNP haplotypes, and an IBD method of Meuwissen and Goddard (2000), across a range of SNP densities. They concluded that as marker density is increased, the advantage of haplotypes over single-markers would be reduced. However, they only investigated power and precision of methods to map QTL with MAF ranging between 0.3-0.5.

The relationship between LD, variance explained and DF fitted in a model is complex, especially when considering haplotype based models, where DF varies depending on a number of conditions related to procedures used with inferred haplotypes. More work is clearly needed to untangle the relationships between marker-marker LD, number of haplotypes and variance explained. Here, we found little relationship between marker-marker LD and proportion of variance explained for all multilocus models, indicating that observed LD would be a poor indicator of model performance. With all models, higher LD between the markers and sQTL results in greater proportions of variance explained. However, proportional gain in variance of models relative to one another will not be constant for levels of marker-sQTL LD (figure 4.6a). Here we have shown that when LD between markers and the sQTL is low the proportional gain in variance from haplotypes more than compensates for the additional DF fitted (figure 4.6b). These relationships are expected to change as marker density and LD increases.

4.4.2 Haplotype uncertainty

The Hap_highest_prob method makes the assumption that haplotypes were known, something that is untrue in practice. This haplotype approach has been commonly used in other studies that compared the ability of models to map QTL (Grapes *et al.* 2004; 2006; Hayes *et al.* 2007; Zhao *et al.* 2007). Discarding low probability haplotype pairs is expected to reduce power and precision of these methods, although the degree to which this happens will be dependent on the level of uncertainty in haplotyping. Morris *et al.* (2004) showed that assuming the most likely haplotype pair to be true for phase ambiguous individuals, results in substantial loss of information compared with modeling for the uncertainty by fitting probabilities in the model. Using only the most likely haplotypes introduces measurement error into the ***X*** matrix, resulting in biased estimates of haplotype effects (Zhao *et al.* 2003). The possibility of large numbers of DF fitted by haplotype based models is an often cited criticism, because of the potential effect on power. Fitting just highest probability pairs in a regression analysis is a useful way of reducing the number of haplotypes fitted in a model. Here we have shown that the performance of the two haplotype methods differs when mapping across a range of sQTL allele frequencies. Performance of the models is a product of the amount of variance they explain for a given number of DF and unequal ratios of these properties across the range of sQTL allele frequencies leads to this pattern of performance. The slight advantage of Hap_all_prob over Hap_highest_prob when mapping for sQTL of intermediate allele frequencies (figure 4.4c), and the potential inclusion of errors in estimating haplotype effects with the Hap_highest_prob model (Morris *et al.* 2004), suggests that approaches such as Hap_highest_prob be avoided for genome-wide association studies in the future.

4.4.3 Accuracy of 'model best' markers

In genome-wide association studies, accurately predicting the location of a QTL is important for both the inclusion of markers in MAS program and fine mapping studies. Typically, markers showing the highest statistical support are assumed to be the ones closest to the causal variant. Because ancestral recombination events usually weaken associations, informative haplotypes normally cover small regions, although this distance is dependent on levels of local LD and effective population sizes. Haplotype fine mapping methods take advantage of the fact that in the close vicinity of a causative locus, haplotypes tend to share close ancestry with causative alleles, with the extent of sharing decreasing with distance from the causative locus (Schaid 2004). If haplotypes are too long, being composed of many distant loci that have recombined with the causative locus, then associations with the trait can be diluted through the inclusion of too many random alleles. Here we have shown that multilocus methods, and in particular haplotype models, are able to accurately identify the position of sQTL considerably more often than a single-locus model.

4.5 Conclusions

The potential benefit in the use of haplotypes for genome-wide association analysis is still widely debated, and further work is clearly required to unravel the interactions between genomic factors and formation of optimal haplotype structures to use in mapping studies. However, here we have demonstrated a clear advantage for using haplotype models that take into account phase uncertainly, in mapping for causal variants with rare or uncommon alleles. Many genome-wide association studies, particularly in humans, have used single-locus approaches and struggled to explain high levels of genetic variation, with the significantly identified markers, despite high-density panels (e.g. Bodmer and Bonilla 2008; Frazer *et al.* 2009; Visscher 2008). Given the findings shown here, there may be some advantage in revisiting these datasets and re-applying haplotype based approaches.

Given the differing performance of models across the range of genomic situations, the challenge will be to find adaptive mapping strategies that optimally use genomic information that is either observed or inferred in order to detect the position of QTL and explain genetic variation for important traits. The variety of models explored here, provide a convenient spectrum of parameterization choices, from simplistic single-locus additive model, to including all interaction terms, as is done in Hap_all_prob. The regression-based framework allows for easy extension to include additional covariates, and relationship matrices for all models, with the only difference consisting of the choice of X matrix. The complementary nature of these models means that parameterization decisions based on observed local conditions could easily be incorporated into a strategy for whole genome mapping. The key will be identifying how well observed local marker information is able to predict the optimal parameterization structure of the model. The natural extension to the results shown here is to consider how many markers should be included in multilocus models, and how their parameterization influences performance under the range of genomic conditions.

CHAPTER FIVE

OPTIMAL LENGTH OF MARKER WINDOWS IN MULTILOCUS ASSOCIATION MAPPING MODELS

5.1 Introduction

5.1.1 Adjacent markers

Strategies for performing multilocus association mapping analyses are still the subject of active debate and research. One important issue is how many adjacent SNPs should be included simultaneously in a particular model. If we are using haplotype-based analyses, this issue is of considerable importance when we consider the relationship between the number of markers included and the potential number of parameters to fit in a model. Early suggestions were to perform haplotype analysis within regions of high LD, typically referred to as "LD blocks", where most of the genetic variation can be captured by a small number of haplotypes (Gabriel *et al.* 2002). However, difficulties in defining LD blocks, the boundaries between them, and choices regarding the inclusion of orphan SNPs, has led to the suggestion that using LD blocks as units for association may not be the most efficient strategy for haplotype analyses (Zhao *et al.* 2003).

Choosing the most appropriate set of markers for haplotype analyses is essential to improve their power (Yu and Schaid 2007). It is impractical to analyse haplotypes constructed from a large number of markers spanning a wide genomic region because the core associated haplotype might be short, yet a longer haplotype region could contain a large number of haplotypes, especially in the presence of weak LD. One approach for using multilocus information is based on a sliding-window framework, in which a number of adjacent SNPs are grouped together in a window to test for association. This sliding-window approach is described in chapters three and four, where it is utilised in a step-wise manner, whereby windows move forward a single marker after each test. Using a method such as this makes maximum use of the information contained between markers and allows easy comparison with single-locus approaches (Zaykin *et al.* 2002). It also provides a choice in how information contained between markers is used, such as fitting just main effects, or inferring all interactions - as is done for haplotype analyses. Currently, we have investigated windows that fit three adjacent markers. However, in theory, the optimal window size should be one that results in the maximum amount of trait variance explained using the fewest parameters. Therefore, the optimal window size is expected to be influenced by the genetic architecture of the trait and local markers. The decision to use a single size of window over the whole genome is always a compromise, as LD patterns vary considerably between across regions. Therefore, it is impossible to predefine a single optimal window size for a whole-genome sliding window analysis.

The effects of the number of markers included in a window on haplotype model performance for QTL mapping (Abdallah *et al.* 2004; Calus *et al.* 2009; Grapes *et al.* 2006; Zhao *et al.* 2007), marker assisted selection (Hayes *et al.* 2007) and genomic selection (Calus *et al.* 2009) has been reported, with optimal haplotype length determined by criteria specific to the type of study. However, in all cases optimal performance was based on the best performance on average, and did not differentiate between genomic conditions. For QTL fine mapping Grapes *et al.* (2006) used the IBD haplotype method developed by Meuwissen and Goddard (2000) to test haplotype performance in a series of simulated populations, and concluded that haplotypes comprised of four to six markers represented the optimal compromise between power and discrimination between successive tested positions. The premise was that whilst using large numbers of markers provided the most accurate estimate of IBD probabilities, it limited the mapping precision of the model. However, Grapes *et al.* (2006) did not compare IBD to regression based methods in terms of power to detect QTL, and also used a simulated dataset that comprised of a small set of evenly spaced markers with high average LD to the simulated QTL. Abdallah *et al.* (2004) showed that a haplotype comprised of two markers resulted in more precise estimates of QTL position than a haplotype of six markers, using an LD based maximum likelihood method.

Calus *et al.* (2009) compared the effects of haplotype definition on the precision of QTL mapping and accuracy of predicting genomic breeding values using an IBD haplotype method similar to that used in Grapes *et al.* (2006) in a multi-QTL simulated dataset. The IBD haplotype method is based on an IBD probability matrix, which is used to cluster haplotypes if they share an IBD probability above a certain threshold. This meant haplotype definition could be controlled by both the number of markers included and the threshold probability. The simulated dataset comprised of 383 SNP markers spread across three Morgan with an average r^2 value of 0.14 between adjacent markers. Calus *et al.* (2009) concluded that window size has a considerable impact of precision of QTL mapping, with windows of six and 12 markers providing the best results, although it made little difference to accuracy of predicting breeding values. This is perhaps not surprising when we consider that genomic selection aims to predict total breeding values with high accuracy, whilst QTL mapping aims at correctly identifying QTL position through contrast with surrounding markers. Therefore, optimal conditions for QTL mapping require a trade off between maximum variance explained with the fewest parameters, while genomic selection aims at capturing the maximum genetic variance explained by models, regardless of localised parameterisation conditions.

5.1.2 Local LD and haplotype diversity

The pattern of LD within a population is determined not only by the distribution of recombination events but also by demographic factors such as the extent of random drift, effective population size, and in localized situations, selection for genetic loci. The idea that resultant variation in localized patterns of LD across the genome could lead to differences in observed haplotype diversity within and between populations, is commonly used as a leading concept in the characterization and identification of haplotype blocks (Gabriel *et al.* 2002). Whilst there is a connection between the statistical concept of LD and the biological reality of haplotypes, a distinction is difficult as a variety of different haplotype structures can be reflected as a single LD pattern (Sawyer *et al.* 2005). Although the relationships between LD and haplotype diversity is complex, a generality of rules exist, such as those applied in the concept of low diversity haplotype blocks defined by strong associations between markers (Gabriel *et al.* 2002).

In this chapter I have attempted to show the relationship between haplotype diversity, defined as the number of haplotypes observed within a given marker window, and the extent of LD between the set of markers for the two haplotype methods. Getting an accurate measure of multilocus LD is difficult without the use of phase information (Bill Hill, personal communication; Weir 1996). Whilst there are numerous methods to calculate multilocus LD, such as chromosome segment homozygosity (Hayes *et al.* 2003) or methods based on entropy (Liu and Lin 2005; Nothnagel *et al.* 2002), these require knowledge of marker phase and are therefore counterintuitive for use in determining the relationship between LD and haplotype diversity. An alternative is to take the mean from the series of pair-wise r^2 measures that exist between the set of markers.

5.1.3 Sliding window framework

Sliding windows-based multilocus methods can also be performed without a fixed length window. In this strategy variable sized windows can be used, where the marker length at a given position can be determined on a set of conditions. Lin *et al.* (2004) presented an approach that exhaustively exploits haplotype information in a TDT test from sliding windows of all sizes. Such exhaustive searches, whilst being computationally feasible, do not necessarily make the most efficient use of local information and will typically be constrained by very conservative multiple testing corrections that will inevitably cause a loss of power. Li *et al.* (2007) introduced a method that fits a variable length sliding window haplotype model where the length of the window is determined by local levels of LD and haplotype diversity. Within the determined window a combined analysis of all the haplotypes of different lengths (up to the maximum of the window) is performed using a regularised regression procedure that adjusts for dependency and complementariness amongst haplotypes. Whilst this is an attractive approach it suffers from the exhaustive fitting of all haplotype lengths within the window, even if there is some reduction in parameters through regularised regression. Yu and Schaid (2007) presented a sequential haplotype scan method based on the concept of a reduction in redundancy between markers within a window. The goal of the method is to choose appropriate markers to include in haplotypes by adding markers sequentially in to a window if they contribute to the association, conditional on current haplotypes. Whilst remaining an exhaustive test, it ensures that the power is not compromised by increases in DF. Many of the current methods designed to allow flexibility in window length have interesting and potentially advantageous uses in association mapping (Bahlo *et al.* 2006; Cheng *et al.* 2005; Li *et al.* 2007; Lin *et al.* 2004; Yu and Schaid 2007) but, are constrained by the requirement of imposing multiple testing corrections caused by exhaustive searches to find the best group of markers or haplotypes. Whilst these methods provide some flexibility in marker choice for multilocus mapping, they are unable to incorporate any knowledge of model performance under different genomic conditions into the choice of window length.

To date, investigation of optimal haplotype lengths have focused on identifying the number of markers that work best the majority of the time. Whilst this is clearly of interest, they do not consider that performance of multilocus models will be strongly influenced by the genetic architecture of the QTL and local markers. Additionally, these studies have typically used simulated datasets that showed unrealistic patterns of LD compared to those found in humans (Pritchard and Przeworski 2001) and the majority of livestock species (Farnir *et al.* 2000; McRae *et al.* 2002; Nsengimana *et al.* 2004; Vallejo *et al.* 2003), making conclusions difficult to draw beyond those of the simulated conditions.

Our previous work (chapter four) identified that certain genomic conditions, such as QTL allele frequencies and patterns of LD influence the performance of different regression-based models. We compared single-marker models to multilocus models that fitted either main effects or haplotypes, using only three-marker windows for the multilocus models, which was essentially an arbitrary choice in the window length. If we consider that performance of a model is dependent on the variance explained by a given number of predictor parameters, then optimal performance of a multilocus model will be affected by localised genetic architecture and how well parameterisation of a model is able to capture available genetic information. Therefore, the optimal number of markers to include in a window will be affected by the combination of how the marker information is parameterised and the genetic architecture of the QTL and markers. Although certain combinations of model and window size will perform best in the greatest number of circumstances, no one combination will be optimal across the whole genome.

5.1.4 Aims

Our aims were to determine the influence of localized LD architecture on haplotype diversity and how this relates to the utilization of between marker information in a variety of multilocus models with different window sizes. We show how the optimal performance in providing statistical support for QTL under different model parameterization and marker window length is dependent on genomic situations such as marker-QTL LD, marker-marker LD, QTL MAF and marker to QTL distance. We go on to demonstrate the use of observed marker information in predicting optimal model performance. We used the extent of LD in a broiler chicken dataset, consisting of 12046 genome wide SNPs genotyped in 200 individuals to simulate LD between markers and a QTL by selecting a single SNP to act as a surrogate QTL (sQTL), and then determining the ability of models to predict the variance of the sQTL using the surrounding markers. Three multilocus models with window lengths between three and nine markers were tested, as well as a single-locus model as a reference comparison. In total 6300 sQTL were generated, representing a range of MAF (0-0.5).

5.2 Materials and methods

5.2.1 Dataset

The dataset used here was the same as was described in detail in chapter four, comprising of genotypic information from a single line of broiler chickens provided by Aviagen Ltd. The same controls, such as removal of missing information and markers represented by fewer than three genotype classes, were placed on the data as described in the previous chapter, allowing it to be used in the manner to simulate LD between markers and a QTL by selecting markers to act as sQTL. The surrounding markers were then used to test for association using a single-marker regression method and three multilocus approaches, with the multilocus models adapted to allow them fit a series of different window sizes of three, five, seven and nine adjacent markers.

Random variation drawn from a normal distribution was added to the genotype classes of the sQTL producing the sQTL "phenotype". The variation of the distribution was chosen such that the sQTL "phenotype" had a heritability of 0.3. This provides high enough power to detect associations in this dataset, whilst avoiding problems associated with regressing against discrete classes of sQTL genotypes when the MAF was low (Chapter four).

sQTL were chosen using the same procedures outlined in Chapter four, whereby each sQTL was surrounded with a test region of 50 markers which were used in the association models. In total the same 6300 sQTL were chosen and analysed with the models. As this was the same dataset as previously reported the descriptive statistics of patterns of LD and sQTL MAF remain as those described in detail in chapter four (figures 4.1 and 4.2).

5.2.2 Models

In chapter four we looked at a variety of models that used the information from SNPs and between SNPs in different ways, and how their performance was influenced by genomic conditions. Three of the models were multilocus models that were implemented in a sliding window framework. One of the difficulties associated with using marker windows is the definition of how many markers should be included within a window. In the previous chapter only windows of three markers were tested, here we have extended this to fit the same models with different length marker windows. To act as a comparison we also included the single-marker analysis.

The following four models were used; single-marker regression, main effects regression, Hap_highest_prob and Hap_all_prob. The later three models were fitted using window lengths of three, five, seven and nine markers. The models are described in detail in chapter four and are implemented in the same sliding window manner. For all models the sQTL was "dropped" from the genotype panel so that it was not included as either a marker or in the formation of haplotypes. The parameter r^2 was used as a measure of LD (Hill and Robertson 1968).

The first stage of the haplotype analyses is to infer the phase of markers in the various size sliding windows which are scanned across the genome panel. This information is used to build the $\boldsymbol{X}$ matrices, which differ depending on the haplotype model used. For a given model haplotype diversity within a window can be determined by the number of unique haplotypes that populate the $\boldsymbol{X}$ matrix. Therefore, the relationship between haplotype diversity and genomic conditions such as marker-marker LD and marker allele frequencies can be assessed.

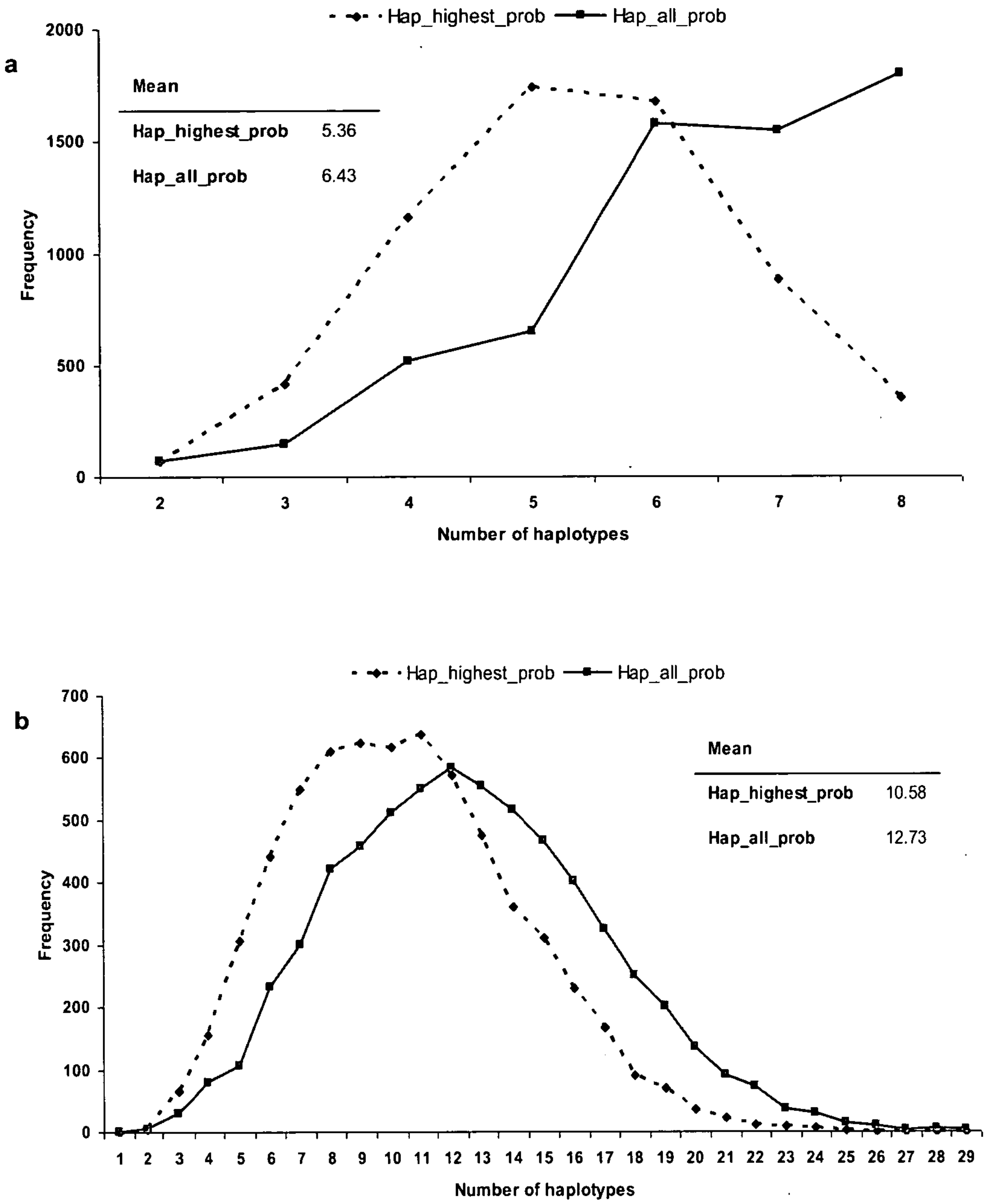
5.3 Results

Information on the average distance between adjacent SNPs, marker spacing and patterns of LD across the dataset is provided in chapter four (figure 4.1).

5.3.1 Haplotype diversity

Localized haplotype diversity is typically given as the number of haplotypes identified in the study population for a given set of markers. It is a product of local LD patterns, or LD complexity, and sample size. Although sample size does not directly influence haplotype diversity, it provides an upper bound to the maximum number of haplotypes that can be observed, given as $2N$, where N is the number of diploid individuals with genotype data provided. High levels of LD between markers results in low levels of haplotype diversity through the complementary relationship between marker alleles. The length of haplotypes, in number of markers, also provides an upper bound on the maximum number of haplotypes observed, as there are only a given number of allele combinations and interactions that can exist between biallelic loci. This upper bound is given as 2^n, where n is the number of markers in the haplotype. For a given set of individuals and group of markers the lower of the two upper boundaries represents the maximum number of haplotypes that can be observed. For example, in a dataset of 200 individuals and nine-marker haplotypes, the maximum number of haplotypes observed can be 400, whilst for three-marker haplotypes in the same data set the maximum is eight. Here haplotype diversity has been investigated for the models that differ in their window size and how haplotypes are inferred. For the haplotype models used here, diversity is an important measure as it defines the parameterization of the models, i.e. the number of DF fitted in the *X* matrix.

The frequency distribution of haplotype diversity in the entire dataset is shown in figure 5.1, for the two models with different window lengths. Across all window sizes the Hap_all_prob method has a distribution shifted to the right of Hap_highest_prob, which naturally reflects the different process of using inferred haplotype information. The relative difference in mean haplotype frequency between the two methods reduces as window size increases, although absolute difference increases. This is expected to be due, in part, to increases in the upper limit of haplotype diversity and the consequential increase in frequency classes. The relative difference in haplotype diversity provides an indication the extent of similarity between models in how information contained between markers is used.


a
Hap_highest_prob
Hap_all_prob
Mean
Hap_highest_prob 5.36
Hap_all_prob 6.43
Frequency
0
500
1000
1500
2000
2
3
4
5
6
7
8
Number of haplotypes
b
Hap_highest_prob
Hap_all_prob
Mean
Hap_highest_prob 10.58
Hap_all_prob 12.73
Frequency
0
100
200
300
400
500
600
700
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Number of haplotypes

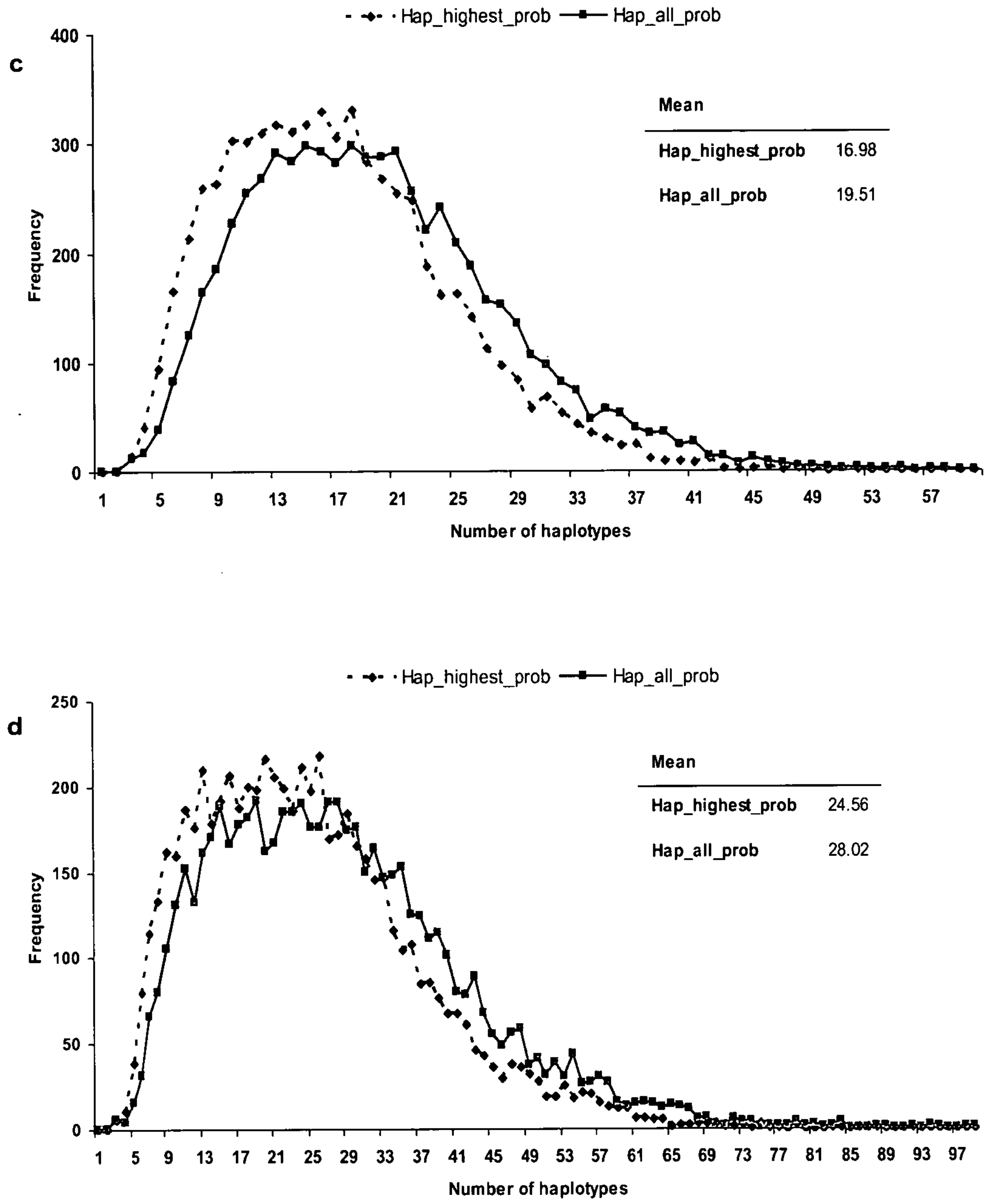


Figure 5.1

Frequency distribution of haplotype diversity found across the entire dataset. Diversity is measured as the number of haplotypes that occur in a given marker window, as determined by the methods Hap_highest_prob and Hap_all_prob. Figures are split by window length: **a)** three markers, **b)** five markers, **c)** seven marker, and **d)** nine markers. The mean of the distributions is given in the tables on each figure.

5.3.2 Relationship between local LD and haplotype diversity

The mean of pairwise r^2 values for markers within a window (termed 'window LD') is used to show the relationship between the extent of LD and haplotype diversity. Figure 5.2 shows mean haplotype diversity from the two models plotted against the mean and variance of 'window LD'. For all window lengths, 'window LD' provides a good predictor of haplotype diversity, showing an exponential relationship connecting high levels of LD with low diversity and low LD with high diversity. The rate of decline in mean 'window LD' with haplotype diversity increases as the length of the haplotype windows gets longer. There is a considerable decrease in variance of the "window LD" measures as the length of windows increases, suggesting that predicting haplotype diversity from 'window LD' becomes more practical as window length improves. This may possibly be due to a greater differentiation in haplotype diversity across the range of LD and improvements in inferring haplotypes correctly as window length increases.

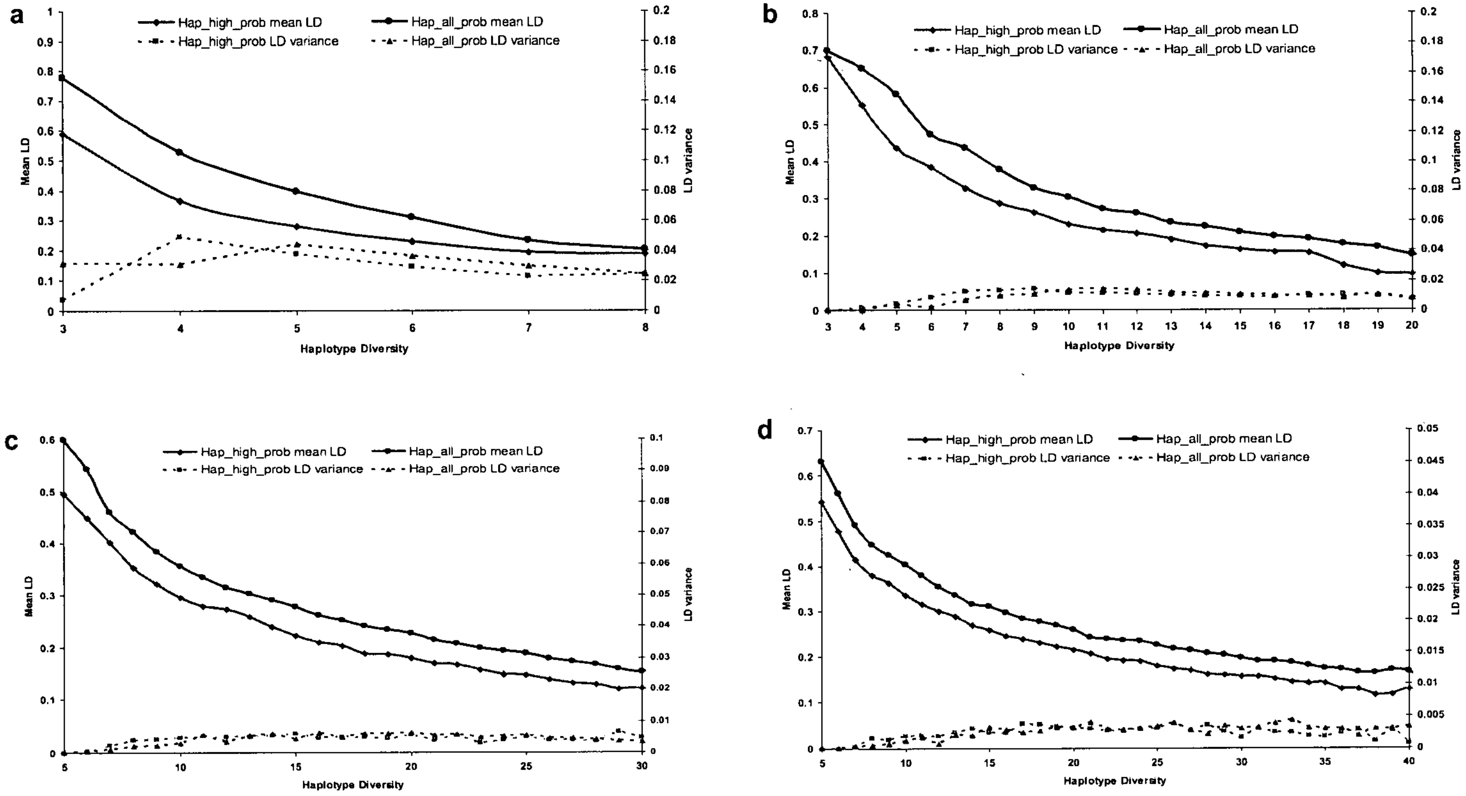
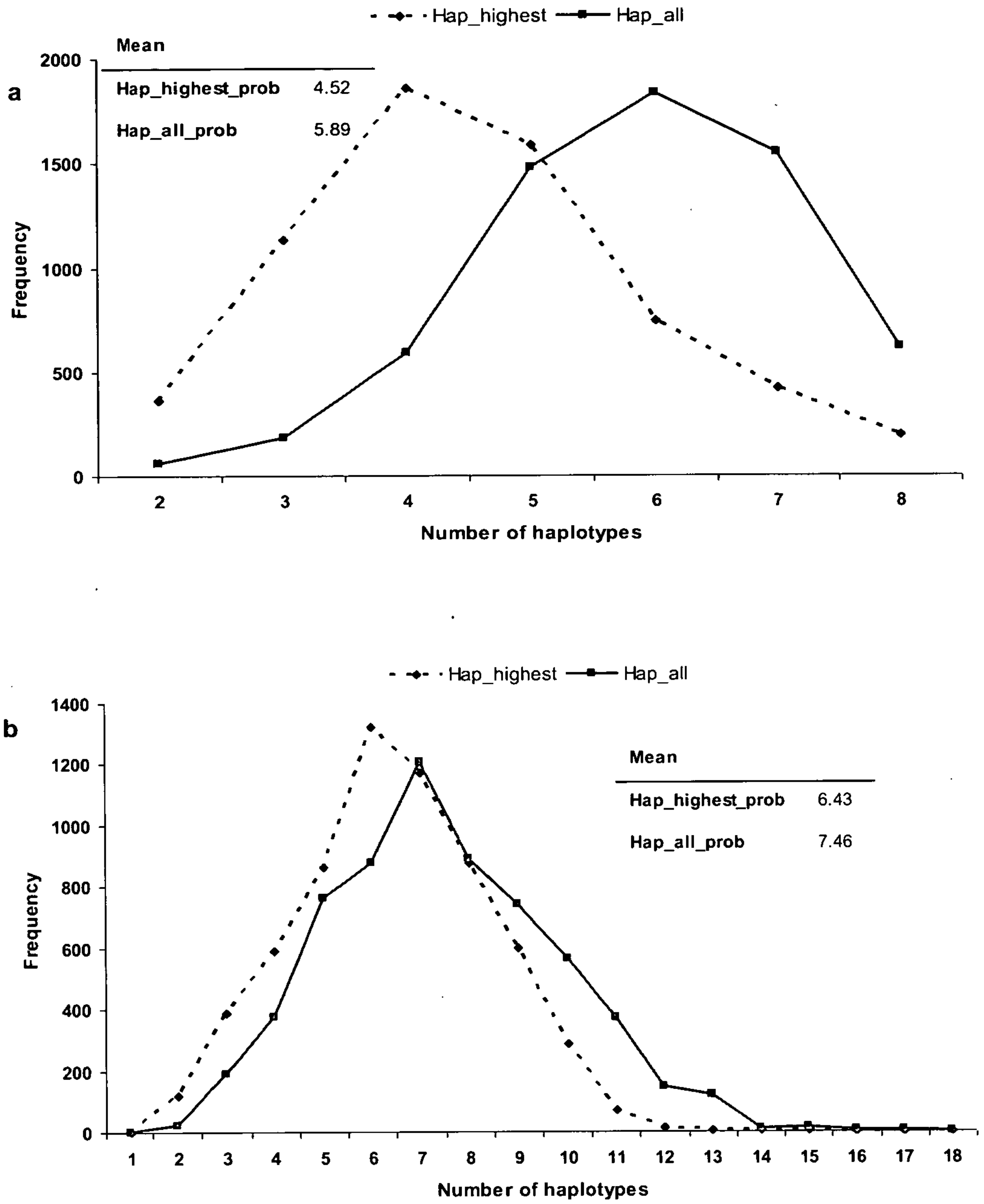


Figure 5.2

Relationship between marker LD and haplotype diversity across the whole dataset. The mean 'window LD' is shown for each level of haplotype diversity, as determined by the methods Hap_highest_prob and Hap_all_prob. The second y-axis shows the variance of this mean. Figures are split by window length: **a)** three markers, **b)** five markers, **c)** seven marker, and **d)** nine markers. Note: differences in the scales of the axes and Hap_highest_prob has been abbreviated to Hap_high_prob.

5.3.3 'Model best' window haplotype diversity

For each model, there is a marker, or marker window, that provided the highest –log 10 *p*-value within each test region. For each model this marker or window is referred to as the 'model best'. For haplotype models, diversity within the 'model best' windows is shown in figure 5.3. These distributions represent haplotype diversity that provides the best statistical explanation of sQTL variance. Compared to distributions for the whole dataset (figure 5.1), haplotype diversity of the 'model best' windows is considerably lower (figure 5.3). It is clear that certain haplotype windows are 'model best' for a number of sQTL, as can be shown by the frequency of certain haplotype diversities in 'model best' windows compared to their frequency in the dataset as a whole. This is not surprising given the replicated overlapping nature of the experimental design. Differences in the distributions of haplotype diversity between those observed in the whole dataset and 'model best' reflects the trade off in variance explained by a model and its parameterization. In the majority of situations, highly parameterized models are unable to explain enough additional variance from the complex interaction parameters to account for the DF added, resulting in lower –log 10 *p*-values.


Hap_highest
Hap_all
a
Mean
Hap_highest_prob 4.52
Hap_all_prob 5.89
Frequency
2000
1500
1000
500
0
2
3
4
5
6
7
8
Number of haplotypes
Hap_highest
Hap_all
b
Mean
Hap_highest_prob 6.43
Hap_all_prob 7.46
Frequency
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Number of haplotypes

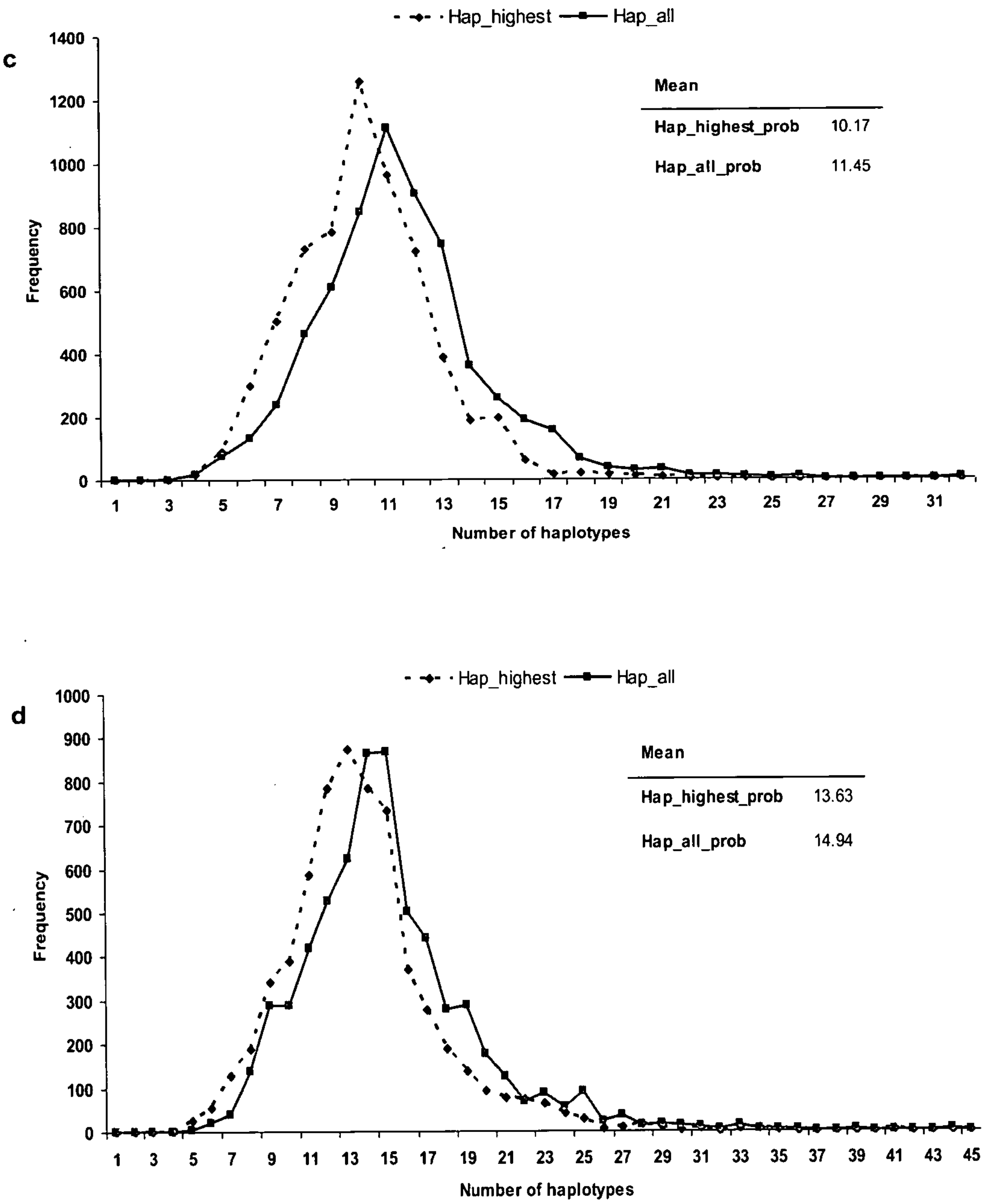


Figure 5.3

Frequency distribution of haplotype diversity seen in the 'model best' windows for the Hap_highest_prob and Hap_all_prob models. The mean of these distributions is shown in the table on each figure. Figures are split by window length: **a)** three markers, **b)** five markers, **c)** seven marker, and **d)** nine markers.

5.3.4 Influence of sQTL MAF on model performance

Results presented in chapter four showed the effect sQTL MAF had on the performance of models. Here we follow similar procedures to investigate the relationship between the causal variant allele frequency and model performance at different window lengths. As before, the sQTL were divided into ten MAF bins, ranging from 0 to 0.5. The distribution of sQTL in each bin is given in chapter four (figure 4.2). The *p*-values from the 'model best' markers are used to evaluate the ability of models to provide statistical support for the sQTL. Naturally, each parameter added will explain some additional variance, although there will be a trade off with the *p*-value, if this parameter is unable to explain enough additional variance to compensate for the DF it adds. *p*-values provide an accurate means for assessing performance between models that differ in their parameterization, as they account for the expected additional variance explained by a parameter, by drawing from distributions adjusted for specific DF.

The mean number of DF fitted by 'model best' markers across each sQTL MAF bin is shown in figure 5.4, with the corresponding mean variance explained and *p*-values in figures 5.5 and 5.6 respectively. There is a ranking of mean DF fitted across the methods based on the number of markers fitted in a window and the complexity of the model parameterization. The mean DF fitted by the haplotype-based models reflects the mean diversity of 'model best' haplotypes shown in figure 5.3. For all models there is an increase in the mean variance explained as the sQTL MAF increases. A similar pattern of model ranking for the mean variance explained is seen for mean DF fitted. The similarity in ranking is expected, being based on the mean number of parameters fitted

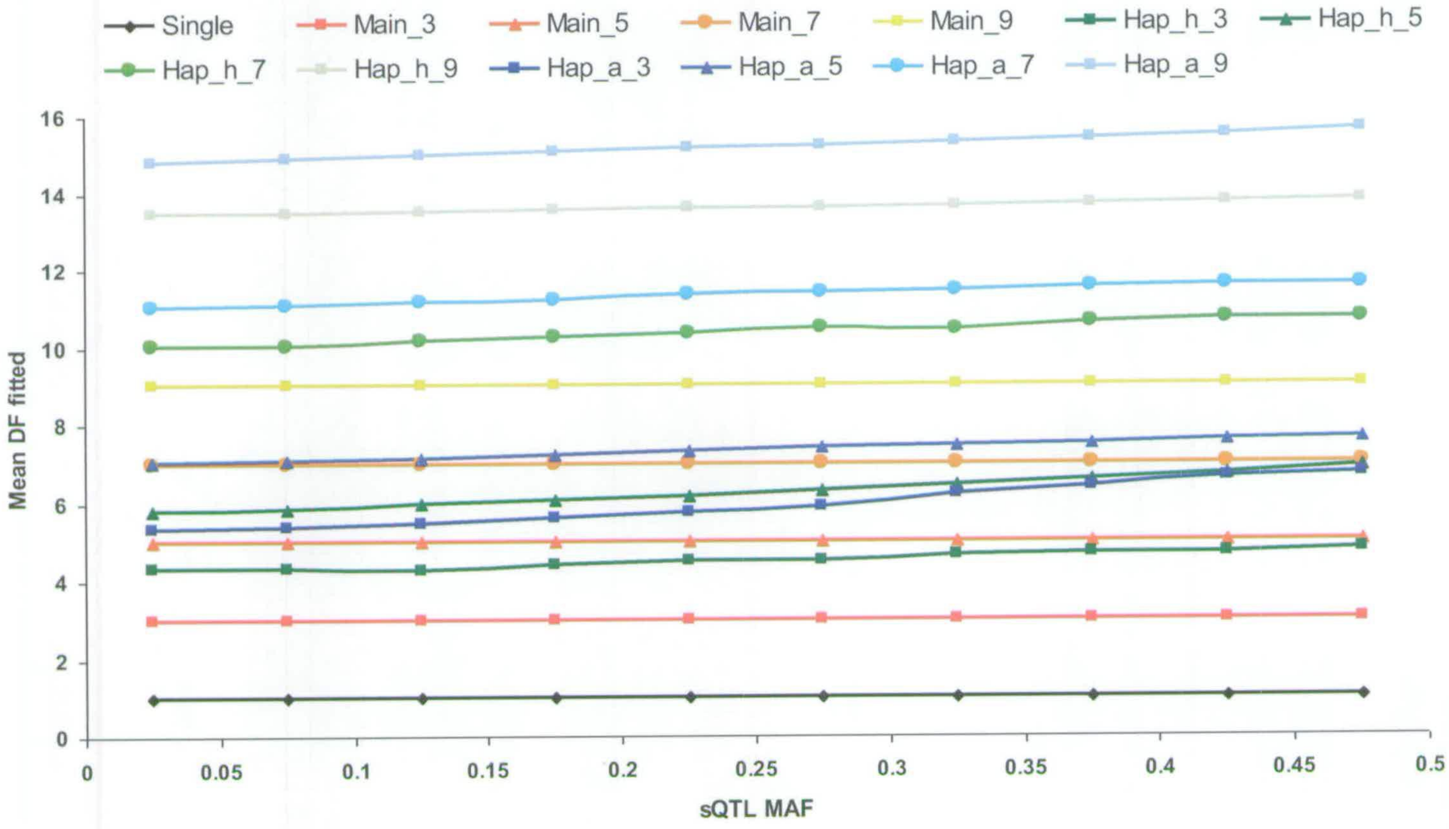


Figure 5.4

Mean DF fitted by the 'model best' windows for each sQTL MAF bin. Figures are split to clarify data points into the three major groups of models and finally all models together. Note: model names have been abbreviated to fit in the keys.

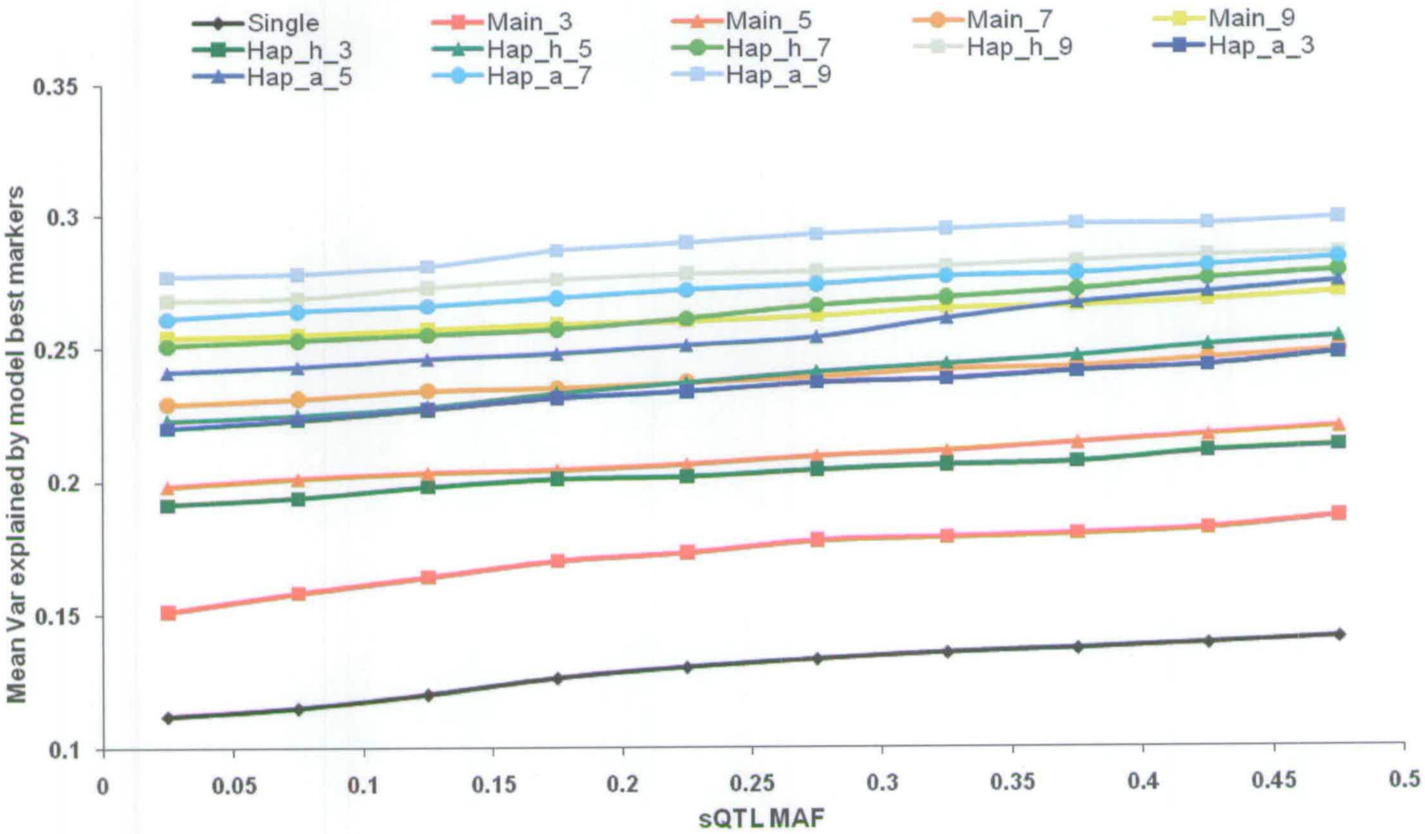


Figure 5.5

Mean variance explained by the 'model best' windows for each sQTL MAF bin. Figures are split to clarify data points into the three major groups of models and finally all models together Note: model names have been abbreviated to fit in the keys.

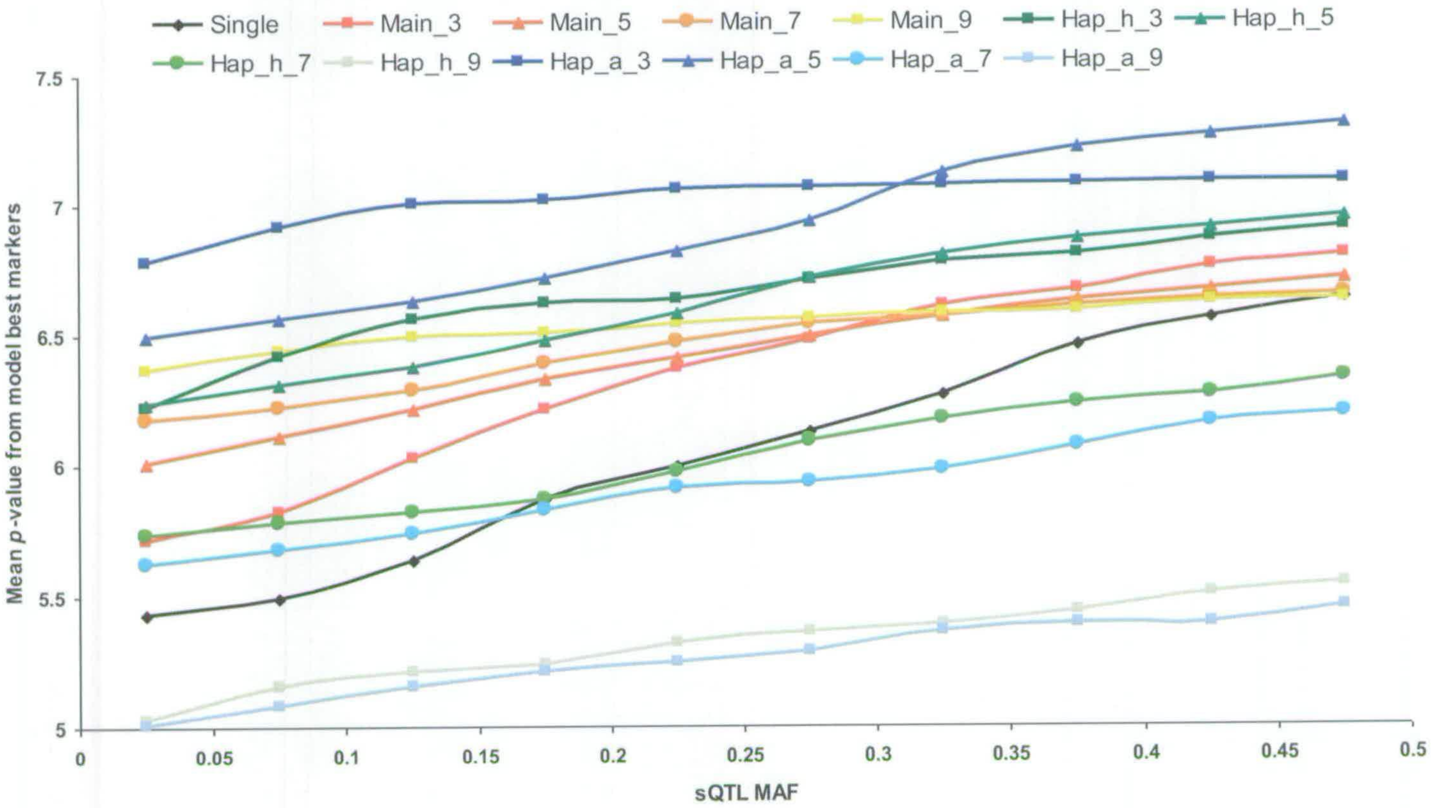


Figure 5.6

Mean –log10 *p*-values from the 'model best' windows for each sQTL MAF bin. Figures are split to clarify data points into the three major groups of models and finally all models together. Note: model names have been abbreviated to fit in the keys.

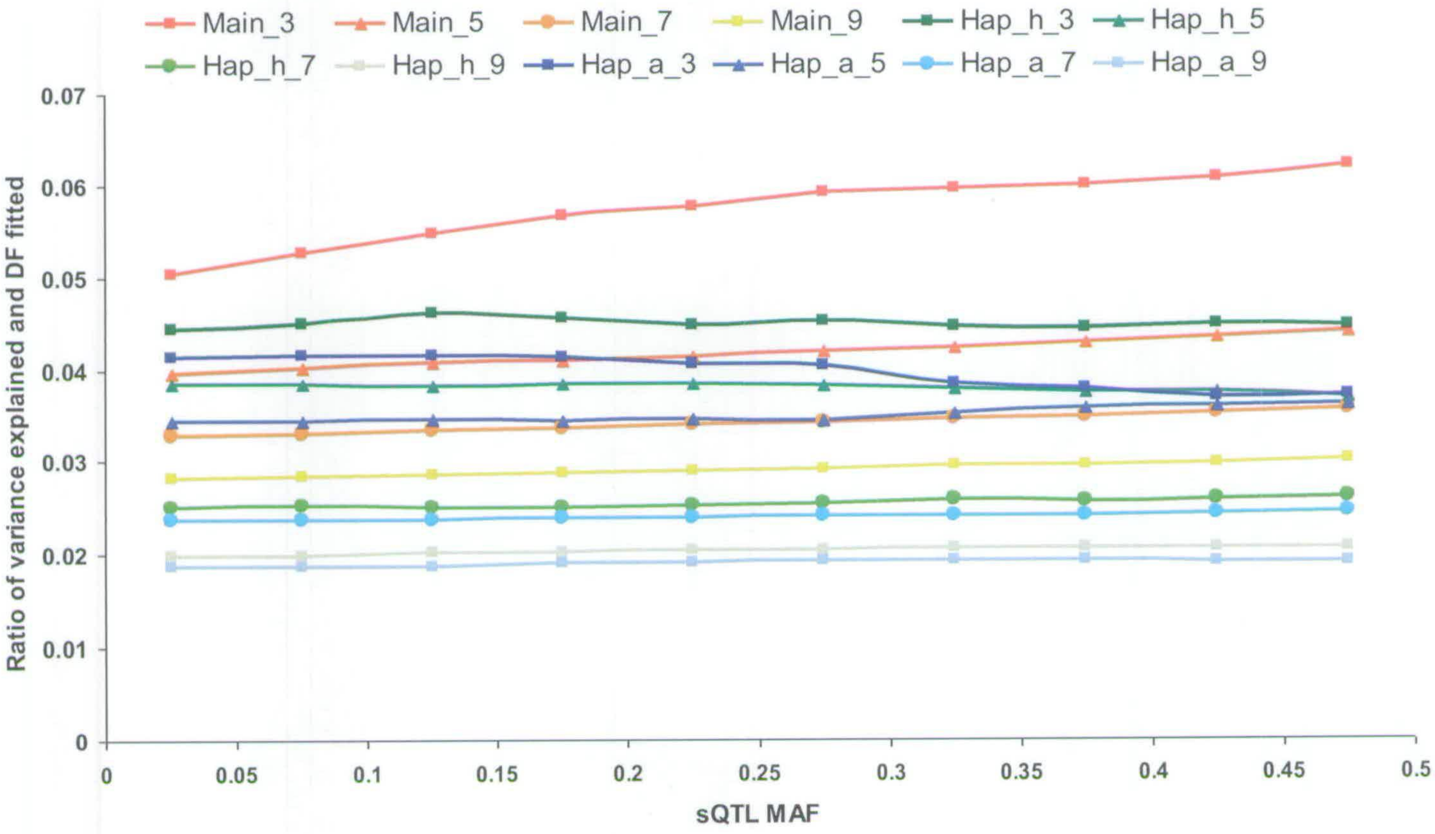


Figure 5.7

Ratio of the variance explained and DF fitted by the 'model best' windows for each sQTL MAF bin. This provides a measure of the amount of variance explained by each DF on average for a given model. Results from single-marker models are not shown, as they have only a single DF.

by the models. Differences in the proportion of variance explained for each DF fitted across the models result in a complex pattern of mean *p*-values, shown on the –log10 scale in figure 5.6. The ratio of variance explained to DF fitted is shown in figure 5.7. This can be considered the amount of variance explained by each parameter in a model on average. For all models there is an improvement in the mean *p*-value as the sQTL MAF increases, indicating that a given model will have higher power to detect common variants. Single-marker and main effect models have the greatest rate of improvement in *p*-values across the sQTL MAF compared to haplotype models, suggesting their performance is more sensitive to the allele frequency of the causal variant.

Haplotype models with window sizes of seven and nine-markers have a poor performance compared other models, and especially compared to shorter window length haplotype models. The advantage of three to five-marker windows for the haplotype models reflects the optimal use of localized information with their parameterization. Whilst mean diversity of seven and nine marker haplotypes from 'model best' markers is considerably less than the whole genome average, their increases in DF are unable to explain enough additional variance, with the result of low –log10 *p*-values. This is reflected by their very low ratios of variance explained to DF fitted (figure 5.7). There is very little difference in performance of the two haplotype models when the window lengths are seven and nine markers, reflecting similar uses in parameterization and utilization of information. For low to intermediate sQTL MAF the Hap_all_prob_3 model performs best, with Hap_all_prob_5 best for intermediate to high sQTL MAF.

5.3.5 LD patterns and model performance

The relationship between marker-QTL LD and model performance is critical, with all models expected to perform better when the markers are in high LD with the causal variant. However, performance of models relative to one another is expected to change disproportionally through differences in their abilities to optimally utilize marker-QTL information contained in LD structures. During the analyses the mean of the series of pairwise LD measures between the markers in a window and the sQTL was calculated for each test. This mean value is termed 'window-sQTL_LD'. Making comparisons between different length windows is difficult, due to different numbers of pairwise combinations being used to construct the 'window-sQTL_LD' measure. However, we are able to compare model performance for equal length windows, as the different models use the same set of markers, just using the information contained between them in different ways. The entire dataset analysis consisted of a total of 6300 sQTL tested against over a 50 marker test region. For the three-marker windows this equates to a total of 302,400 individual windows used by each model, whilst for five, seven and nine-marker window lengths the totals are 289,800, 277,500 and 264,600 respectively. For each of these individual windows, the performance of different models was compared based on their *p*-values. For each window a model was deemed to have performed best if it produced the smallest *p*-value from the set of models. The proportion of times each of these models performed best across the range of 'window-sQTL_LD' is shown in figure 5.8.

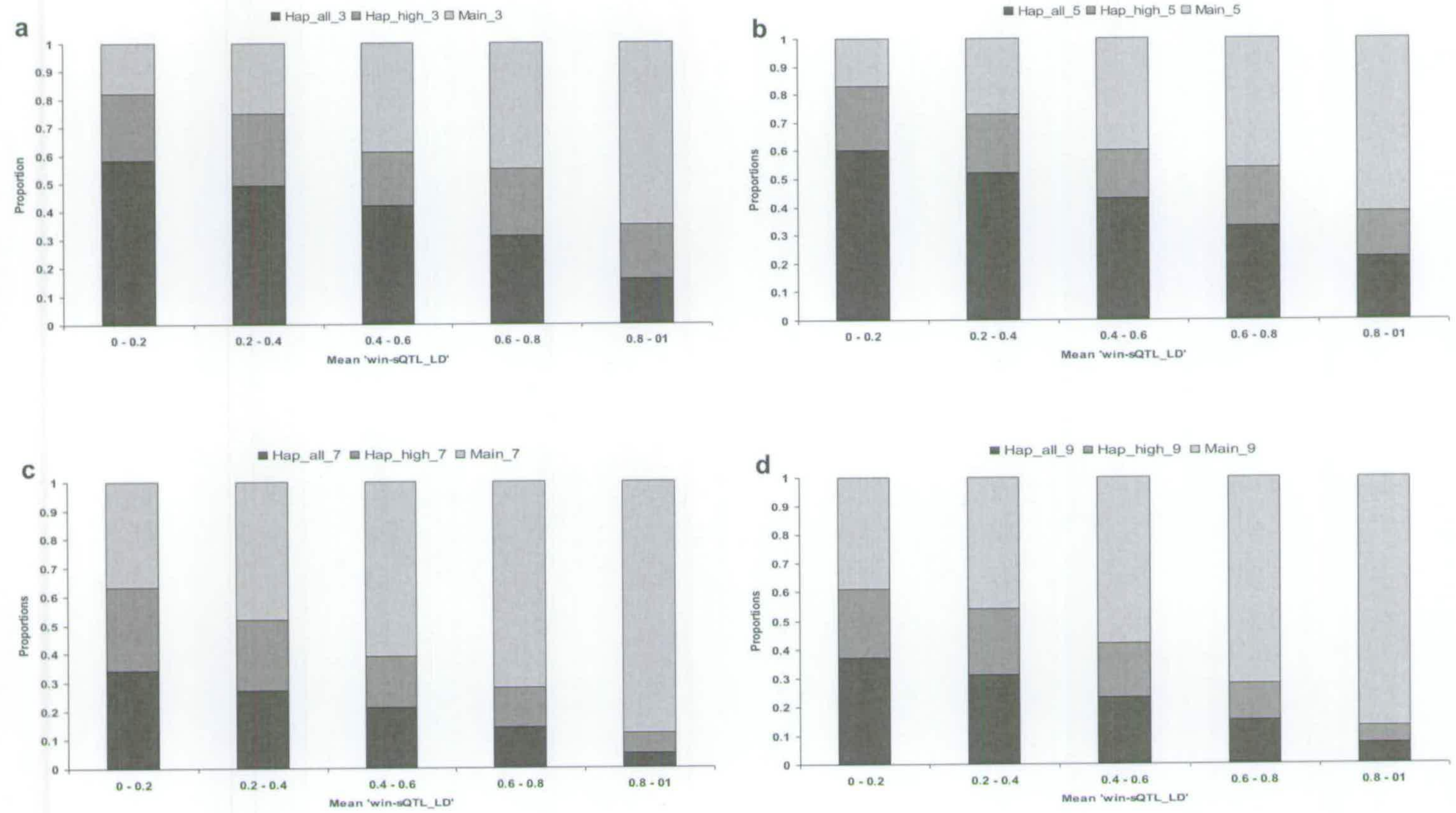


Figure 5.8

Proportions of times models perform best for each window of markers in the entire set of analyses, split into 'window-sQTL_LD' bins. The figure is divided by models with equal window lengths; **a)** models with 3 marker windows, **b)** models with 5 marker windows, **c)** models with 7 marker windows **d)** models with 9 marker windows.

When the 'window-sQTL_LD' is low haplotype based models perform best the greatest proportion of the time. From the three marker length window models, Hap_all_prob performs best 58 percent of the time when the 'window-sQTL_LD' is between 0-0.2. This falls to 16 percent of the time when 'window-sQTL_LD' is between 0.8-1. Main effect models have an opposite pattern of performance to haplotype models, with the three-marker main effect model performing best for 18 percent of tests with 'window-sQTL_LD' between 0-0.2, and 65 percent with 'window-sQTL_LD' between 0.8-1. Across almost all combinations of window size, and 'window-sQTL_LD' bin, the Hap_all_prob model performs best in a higher proportion of tests than the Hap_highest_prob model. Across the whole range of 'window-sQTL_LD' the main effect model performs best an average of 38.5 percent of the time when the window length is three-markers, with the same average percent for five-marker window tests. When the window lengths are seven and nine-markers this average is 61.2 and 60.3 percent respectively.

5.3.6 Predicting model performance from observed information

One of our aims here was to be able to use observable marker data as a means of predicting model performance. Naturally, in order to use this for traditional mapping studies, assumptions regarding the position of markers relative to the QTL are required. Therefore, information from a set of markers that surround the sQTL position was used as predictors of their relationship with the sQTL. The performance of models within that set of markers was then assessed.

For each sQTL the mean of pairwise LD measures amongst the four markers either side of its position was calculated, along with the mean pairwise LD between those eight markers and the sQTL. In this way, observed marker information (from a window of eight markers) can be used as a predictor of their relationship with an underlying causal variant. Amongst the set of eight markers for each sQTL the model that performed best, deemed as having the smallest *p*-value, was recorded. For this prediction stage only the performance of models using markers from this set of eight were considered, not the performance across the whole test region. This allowed the relationship between predictive ability of the markers and model performance to be determined. To reduce the number of comparisons amongst models the performance of Hap_highest_prob models was not considered due to their poor performance relative to Hap_all_prob models of comparable length. Likewise, main effect and Hap_all_prob models using nine-marker windows were also not considered.

Figure 5.9 shows the relationships between mean pairwise LD amongst the eight markers, and the mean LD between the markers and the sQTL. The points for the mean LD measures are first shown as a scatter plot for the 6300 sQTL. This relationship is then represented in a hexagon binning histogram, where the size of the hexagons represents the density of underlying points (Carr 1991). The colour of the hexagons represents the model that had the best performance for the majority of points summarized by that hexagon. Using a hexagon plot provides a clear visualization of the ability of using observed marker information, in this case LD, and optimal model performance. In a traditional mapping study only marker LD information is available to use as a predictor for model performance. Therefore, the proportion of times each model

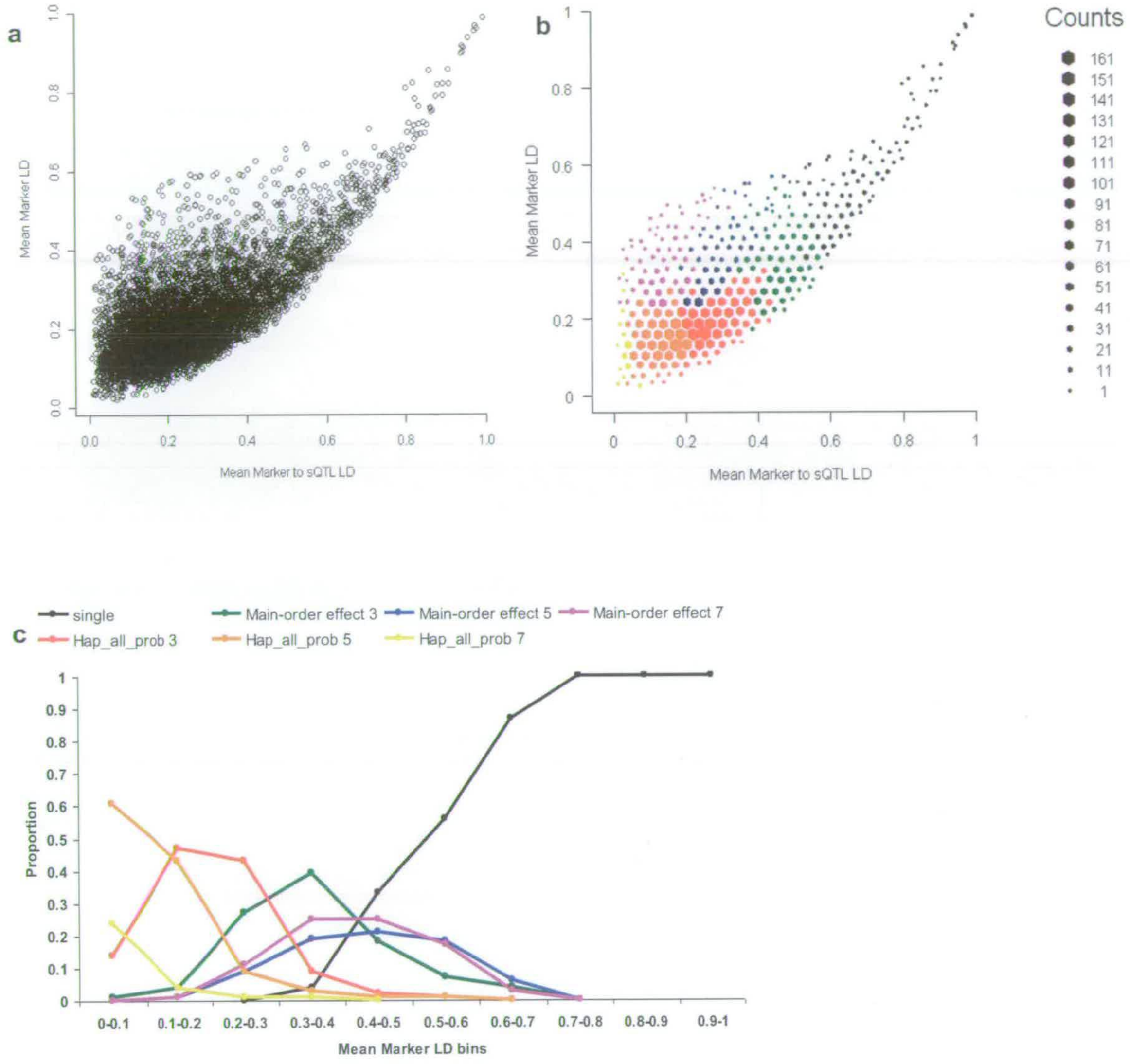


Figure 5.9

Total set of 6300 sQTL - a) Relationship between the mean pairwise LD from a set of eight markers surrounding a sQTL (four either side) and the mean of pairwise LD measures of these markers and the sQTL. **b)** The relationship shown in a) represented as hexagon plots, where the size of the hexagon represents the density of underlying points. Hexagons are coloured based on the model that has the highest proportion of "best" results amongst the points within the hexagon. The model colour key is shown in part c). **c)** Proportion of times each model performs "best" across bins of the mean LD between the set of eight markers (observable information).

performs best is shown for the range of mean marker LD (values shown on the *y*-axis of the scatter and hexagon plots) divided into ten equal sized bins. Proportions are taken from the sQTL within the bin, thus, will represent different numbers of sQTL tests. The hexagon plot shows a clear relationship indicating the optimal model performance given certain patterns of LD between the markers and between the markers and the sQTL (figure 5.9b). However, model choice becomes more difficult to discern, when we consider that only marker LD is observable. The patterns of proportion of times each model performs best across observable marker LD (figure 5.9c) indicates that model decision would be difficult unless either low or high mean LD is observed.

Given the difference in the performance of models across the range of sQTL MAF, results from a subset of sQTL with MAF < 0.1, and sQTL with MAF > 0.4 are also shown in figures 5.10 and 5.11 respectively. The pattern of LD observed between markers and between markers and the sQTL is less clear when the sQTL has a low MAF (figure 5.10a) compared to the high MAF sQTL (figure 5.11a). This likely represents the mathematical properties of r^2 values when one locus has a low MAF (VanLiere and Rosenberg 2008), as here the mean of a series of pairwise measures is calculated. If we take just observable information, then the proportions for best model performance are markedly different for sQTL at either end of the MAF scale. When the sQTL MAF is low, single-marker models very rarely perform best and only when mean marker LD is at intermediate levels. For sQTL with low MAF, high levels of marker LD are infrequently observed. The overall proportion of times the Hap_all_prob 3 model performs best is higher when the sQTL has low MAF compared to the proportion from high MAF sQTL.

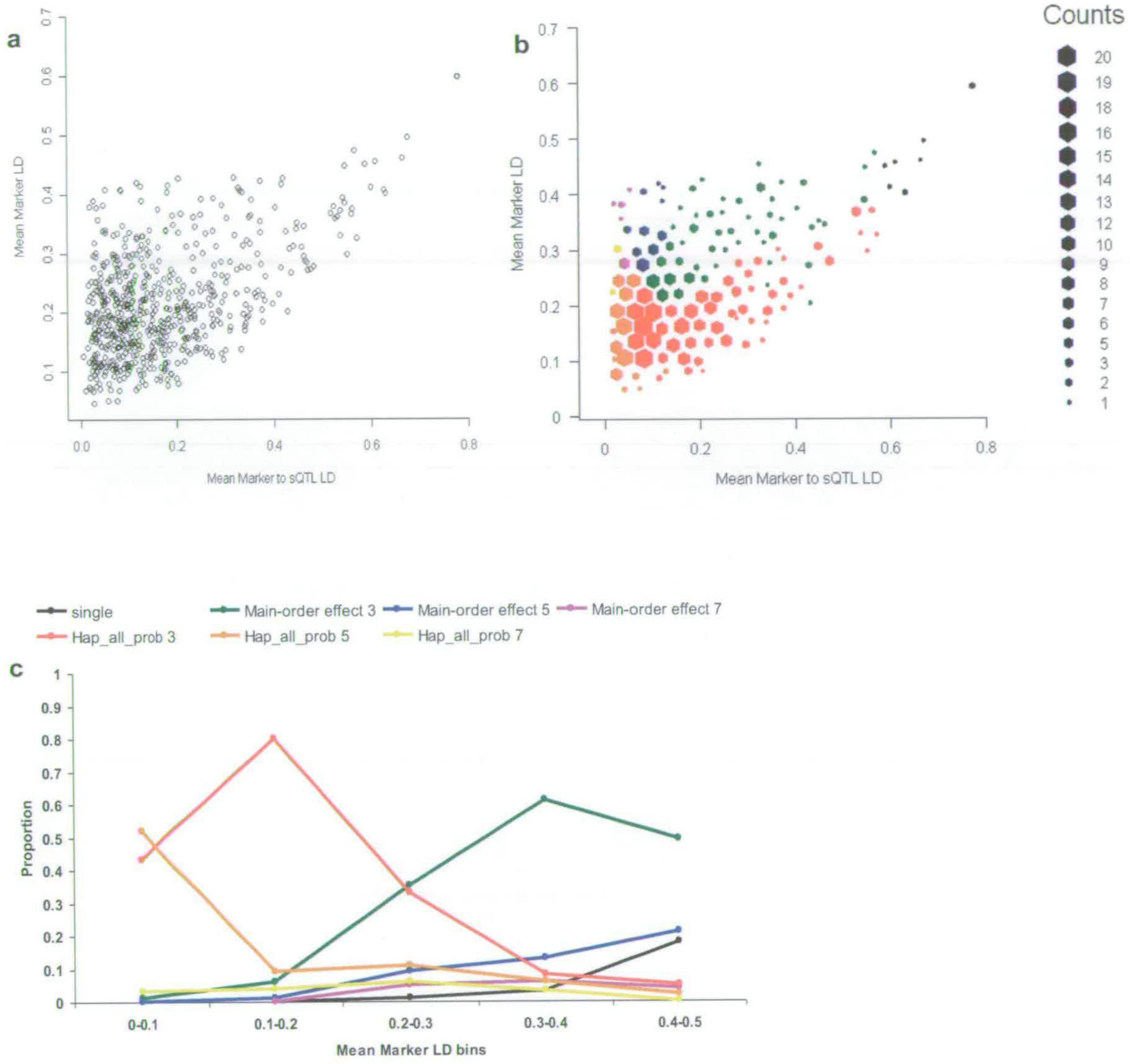


Figure 5.10

sQTL with MAF < 0.1 - a) Relationship between the mean pairwise LD from a set of eight markers surrounding a sQTL (four either side) and the mean of pairwise LD measures of these markers and the sQTL. **b)** The relationship shown in a) represented as hexagon plots, where the size of the hexagon represents the density of underlying points. Hexagons are coloured based on the model that has the highest proportion of "best" results amongst the points within the hexagon. The model colour key is shown in part c). **c)** Proportion of times each model performs "best" across bins of the mean LD between the set of eight markers (observable information).

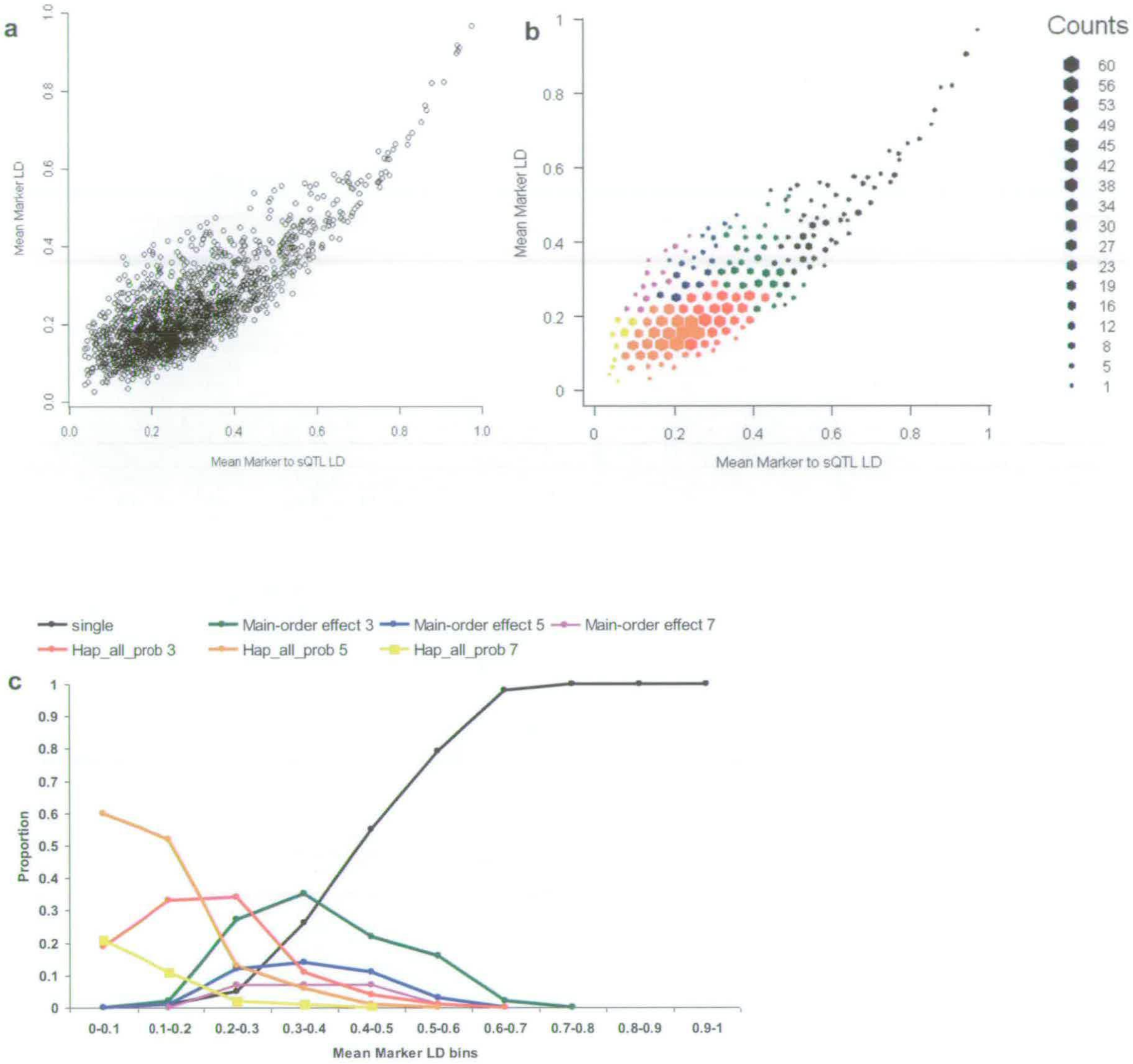


Figure 5.11

sQTL with MAF > 0.4 - a) Relationship between the mean pairwise LD from a set of eight markers surrounding a sQTL (four either side) and the mean of pairwise LD measures of these markers and the sQTL. **b)** The relationship shown in a) represented as hexagon plots, where the size of the hexagon represents the density of underlying points. Hexagons are coloured based on the model that has the highest proportion of "best" results amongst the points within the hexagon. The model colour key is shown in part c). **c)** Proportion of times each model performs "best" across bins of the mean LD between the set of eight markers (observable information).

5.4 Discussion

Here we have shown how performance of multilocus models is strongly influenced by the number of markers fitted in the analysis, and choices made regarding the use of information contained between markers. We have also demonstrated the diversity of haplotypes as a product of their marker length and localized LD structures. Given the extent of LD observed in this dataset, there is a considerable advantage to using a haplotype model that accounts for uncertainty in phase, such as Hap_all_prob over all other methods tested here. The choice in the number of markers to include in a window is of critical importance, with optimal window length also affected by the allele frequencies of the causal variant, marker LD structures and marker-sQTL LD.

5.4.1 Haplotype diversity

Diversity of haplotypes within a dataset is a product to local LD variation, sample size and the defined length of haplotypes, with considerable variation in diversity across the genome. There is a clear relationship between the length of haplotype window and range of observed haplotype diversity, as determined under the two different models (figure 5.1). Data analysed here is supplied with unknown phase and individuals from a single generation, resulting in some uncertainty in phase after the inferring procedure. The two haplotype methods handle this uncertainty differently, leading to different measures of haplotype diversity for each specific model. In this situation, haplotype diversity can be considered a measure of model parameterization, allowing haplotype diversity to be viewed in terms of model performance.

Accurately measuring and producing a meaningful statistic to explain the pattern of LD between a set of multilocus markers is difficult without the use of haplotype information (Bill Hill: Personal communication; Hayes *et al.* 2003). Using the arithmetic mean, we summarize the series of pairwise r^2 measures between the set of markers within a window to provide an indication of the extent of multilocus LD. Here it is shown that using the mean of pairwise LD measures provides a useful indicator of haplotype diversity, and consequently model parameterization, for the two haplotype models (figure 5.2).

Naturally, given the importance of marker length to haplotype diversity, factors such as marker density and effective population size need to be considered given their relationship to localized LD patterns. Thus, conclusions drawn here regarding haplotype diversity and model performance need to be considered in terms of marker spacing and localized LD. Using a measure such as 'window LD' to provide an indication of the extent of LD between a set of markers, provides a useful indication of the expected haplotypes diversity under the two haplotype models. Variance in haplotype diversity for a given level of 'window LD' is likely due, in part, to inaccuracies in estimating multilocus LD using a measure such as 'window LD', and the fact that different haplotype structures can be reflected by a single LD pattern (Sawyer *et al.* 2005). As marker window length increases, there is a reduction in the variance of haplotype diversity. Whilst the reason for this is not fully known, it suggests that more accurate predictions of haplotype diversity can be made using larger sets of markers.

5.4.2 Identification of optimal window lengths

The extent of LD between the set of markers within a window clearly has an important effect on the ability of a model to provided statistical support for a QTL. For haplotype–based models, diversity is increased for lower levels of LD (figure 5.2), and hence, the number of parameters fitted in a model is greater. Here we have shown that for a specific multilocus model the optimal window length is influenced by the allele frequencies of the sQTL (figure 5.6) as well as extent of LD between the markers and the sQTL (figure 5.8). As expected there is a ranking of mean variance explained by the 'model best' markers, based on the average DF fitted amongst the models (figures 5.4 and 5.5). Differences in the performance of models to provide statistical support from the 'model best' markers is reflected in the critical values of the *F*-tests. The main effect model consistently outperforms single-marker analysis across all sQTL MAF and window sizes. The greatest difference in performance is when mapping for rare causal variants. In this situation a three-marker model performs best on average, although the difference between the models narrows as the MAF of the sQTL increases. When the sQTL MAF is greater than 0.3 the five-marker window model performs best on average. There is very little difference in the performance of alternative window lengths and single-marker analysis when the sQTL has intermediate allele frequencies.

In theory, optimal window size should be one that results in haplotypes that maintain the highest LD with the causal variant. Whilst this is true in principle, situations exist whereby a single or small group of haplotypes have a strong association with a causal variant, but are located in regions of high diversity. Fitting all haplotypes seen at this locus can weaken the power of the test (Akey *et al.* 2001; Schaid 2004). A similar problem occurs if haplotypes are too long, composed of many distinct loci that have recombined with the causative locus. The consequence of large numbers of haplotypes, composed of many random alleles, is a weakening of the association with the causal loci. Whilst recombination events break up associations between haplotypes and QTL, problems associated with long haplotypes are also related to the density of the marker panel in the study. Therefore, the term 'long haplotypes' needs to be thought of in relation to marker density and mean levels of LD between adjacent markers. In lower density panels, such as that used here, including seven or nine markers in a haplotype window drastically reduces the performance of the model relative to main effect models of comparable length. At these lengths, haplotypes are covering sets of markers that have likely accumulated a large numbers of recombination events, destroying associations with surrounding markers and producing high diversity that is weakly powered in a global test for association. In this situation the maximum test statistic from the set of single-marker tests is expected to perform better (Schaid 2004).

Haplotypes for both Hap_highest_prob and Hap_all_prob models are coded as additive effects in a fixed effects model, with the number of numerator DF in the global *F*-statistic equal to the number of observed haplotypes. When haplotype diversity is high, power can weaken due to an inability of parameters to explain enough additional variance to compensate for their DF and a more stringent critical value. Igo *et al.* (2009) showed that overall power of both a haplotype score test (HST) and cluster based test (Tzeng *et al.* 2006) dropped uniformly with increasing haplotype diversity, suggesting that the genetic architecture in regions of high diversity contains complexity that is neither explained by HST or adequately simplified by the clustering scheme. This observation is not surprising, as a causal mutation, whose susceptibility allele has a low frequency, in a region of high complexity is expected to occur on several different haplotype backgrounds (Igo *et al.* 2009).

High levels of parameterization in longer window haplotype models could result in an inability to accurately estimate haplotype effects in diverse sequence regions, due to small sample sizes of haplotypes (Bardel *et al.* 2006; Li *et al.* 2007). One strategy is to compute an individual test statistic for each haplotype, and then use the maximum of these to test for association, with a correction for multiple testing. Seltman *et al.* (2003) showed that while this approach may be most powerful when only one haplotype is strongly associated with the trait, its power is weakened when the association is spread across multiple haplotypes. In this situation, a global test, as used here, is expected to be more powerful (Schaid 2004; Seltman *et al.* 2003). An alternative approach could be to use a variance component model that uses a likelihood ratio statistic that is not penalized by high levels of parameterization. If haplotypes are coded as additive effects in a fixed effects model with the number of distinct haplotypes denoted as *K*, the DF for the global *F*-test is *K* for the numerator and *N-K* for the denominator, where *N* is the number of individuals. Whilst high values of *K* can lead to low power of the test statistic because of more stringent critical values, no matter the value of *K*, the variance component model tests the null hypothesis of no association of any haplotypes with the trait, $H_o : \sigma_\beta^2 = 0$ verses the one-sided alternative $H_o : \sigma_\beta^2 > 0$ (Schaid 2004). Because of this, the variance component model is likely to be more powerful than the fixed effects model when there are many haplotypes.

Whilst it initially appears counter intuitive, using all available information for QTL mapping is not always optimal, given the considerations of model parameterization. Abdallah *et al.* (2004) showed that a haplotype composed of two markers resulted in more precise estimates of QTL position than a haplotype of six markers for an LD-based maximum-likelihood model that is a generalized method of Terwilliger (1995). Methods fitting haplotypes clustered on the basis of IBD probabilities (Meuwissen and Goddard 2001) have shown that window size has a strong effect on QTL mapping. Using a simulated dataset, Grapes *et al.* (2006) showed that using haplotypes of four or six markers in a sliding window framework resulted in the greatest mapping accuracy. Using this IBD approach, fitting single markers resulted in a worse mapping accuracy than all haplotype length models, although this is likely due to a very poor accuracy in estimating the IBD probabilities from just a single marker (Grapes *et al.* 2006). Using the same IBD approach, Calus *et al.* (2009) compared the effects of haplotype definition and length on the precision of QTL mapping, using a range of window sizes and clustering of related haplotypes based on different thresholds of IBD probabilities. Under these criteria, window length and probability thresholds had a considerable impact on mapping precision, with windows of six and 12 markers providing the best results. The discrepancy in optimal window size between these studies is expected when we consider the differences in population genetic parameters such as the extent of LD simulated in their datasets, and a lack of consideration to differences in model performance under the range of genomic architecture.

5.4.3 Predicting model performance from observed information

As we have shown in this, as well as the previous chapter, performance of regression-based models is strongly influenced by the localized genomic architecture of the markers and QTL. Given the variable nature of this architecture across the genome, the ability to predict the optimal choice of model using observable information is a considerable advantage for mapping studies. Using observed information as a predictor of model performance requires the assumption that the causal variant is located close to the markers, a usual assumption in traditional mapping studies. In this chapter, LD information from a set of eight markers surrounding a sQTL position was shown to be a useful predictor of model choice (figure 5.9). Using this marker information allows us to make an informed choice of the type of model to use in a given situation. If we make the assumption that the QTL has either a low or high MAF then we gain a more informed set of choices. Naturally, knowledge of LD between markers and the sQTL would be the ideal situation.

In genome-wide association studies information available for the prediction of model parameterization is typically observed from markers, rather than any knowledge of marker interactions with causal loci. As is shown in figures 5.10 – 11 the choice of model, based on observed marker LD, can be clarified if prior knowledge of the QTL allele frequencies is available. In the absence of prior knowledge assumptions regarding the likely distribution of QTL allele frequencies could be applied. For example, under assumptions of neutral mutation or stabilizing selection models, the distribution of QTL allele frequencies is expected to resemble a U-shaped distribution, with a high proportion of loci with low MAF (Lynch and Hill 1986; Wright 1935). Applying these assumptions in a genome-wide mapping strategy that uses observed marker information to influence model choice could lead to more efficient use of genomic information and consequently improve power of the study, particularly in respect to mapping for rare variants.

Events such as selection criteria and population bottlenecks, that are a reality in animal breeding systems, will influence the distribution of QTL allele frequencies away from the expected distribution of the neutral model (Zhang *et al.* 2004), making assumptions harder to accept. Likewise, the genetic architecture of an individual trait may not adhere to an assumed distribution, especially if it is influenced by a few loci. Nevertheless, applying the model choice criteria observed for low MAF sQTL (figure 5.10b) is expected to improve the ability of a genome-wide mapping study to detect rare variants, which is an often cited struggle in association mapping (Andersson and Georges 2004; Bodmer and Bonilla 2008; Wang *et al.* 2005; Zondervan and Cardon 2004).

5.4.4 Comparison of haplotype models

As is shown in chapter four, there are considerable differences in performance of the two haplotype based models. Although, there is a reduction in the differences of their performance as window size increases (figure 5.6). This reflects the increasing similarity in the distributions of haplotype diversity of the two methods as window length increases, suggesting that there is little difference in their parameterization. At lower window lengths the Hap_all_prob model consistently outperforms Hap_highest_prob, across the range of sQTL MAF, providing consensus results for five-marker windows to those shown in chapter four for three-marker windows.

To account for unphased haplotypes, some investigators have used statistical methods, such as the EM algorithm, to infer the most likely haplotype pair per subject, and then use these inferred haplotypes as if they are observed (Grapes *et al.* 2004; 2006; Hayes *et al.* 2007; Zhao *et al.* 2007), as is done for the Hap_highest_prob method. Whilst this reduces the number of parameters slightly, it makes the assumption that haplotype pairs are known, something that is untrue in practice. This approach does not account for the discarded haplotypes pairs that are possible, and if LD is not strong, there can be substantial loss of information (Schaid 2002). Using only the most likely haplotypes can introduce measurement error into the X matrix, resulting in biased estimates of haplotypes effects (Zhao *et al.* 2003), and a possible reduction in the accuracy of estimating parameters (Tanck *et al.* 2003). Discarding low probability haplotype pairs is expected to reduce power and precision of methods such as Hap_highest_prob, although the degree to which this happens will be dependent on the level of uncertainty in haplotyping (Morris *et al.* 2004).

The occurrence of large numbers of haplotypes, especially rare ones, is an often cited criticism of the reduced power of haplotype over genotype based models (Clayton *et al.* 2004; Schaid 2004). Using a model such as Hap_highest_prob is a useful way to reduce parameters, although the relative reduction compared to Hap_all_prob becomes smaller as window size increase, as can be seen from the distributions of haplotype diversity (figures 5.1 and 5.3). Here we have shown that the performance of the two haplotype methods differs when mapping across a range of sQTL allele frequencies, and that optimal performance is also based on window size. As performance of models is a product of variance they explain with a given number of DF, the unequal ratio of these properties, across the range of sQTL MAF, leads to the pattern in performance see in figure 5.6. The advantage of Hap_all_prob over Hap_highest_prob for window lengths of three and five markers (figure 5.6), and the knowledge of a potential errors in estimating haplotype effects with the Hap_highest_prob model (Morris *et al.* 2004; Schaid 2002), suggests that approaches such as Hap_highest_prob be avoided for genome-wide association studies in the future. This suggestion is supported by the results shown in chapter four.

5.4.5 Optimal parameterization

The models investigated here have clearly defined parameterization choices. For example, main effect models fit just the main effect parameters, whilst Hap_all_prob models fit all main effect and interaction parameters contained within a set of markers. For a set of multiple markers, these choices represent the two extremes of a spectrum of model parameterization. For each given set of markers there would be a clear advantage in identifying the individual set of parameters that explains the maximum proportion of variance using the fewest DF, rather than relying on the constraints of the two extremes. Several approaches have been proposed to tackle this problem, normally referred to as parameter reduction methods.

In situations of high haplotype diversity, it is typical to observe large numbers of rare haplotypes. When this occurs, standard errors of frequency estimates for rare haplotypes can be large due to sampling variation. The problems this creates are compounded when phase uncertainty is not accounted for (Fallin and Schork 2000; Zhao *et al.* 2003). Rare haplotypes also suffer from a reduction in the ability to accurately estimate their regression parameters, often leading to model instability (Schaid 2004). A common approach is to remove rare haplotypes from the $\boldsymbol{X}$ matrix, yet this implicitly groups them into a baseline category and ignores any information that they contain. An alternative strategy is to group rare haplotypes into a single haplotype class. Whilst this is an attractive strategy, it requires a frequency threshold to be defined, under which a haplotype is considered rare. This procedure drastically reduces power when rare haplotypes are associated with the causal variant, and makes interpretation of the regression coefficient for this group nearly impossible. A more appealing approach is to shrink the estimated effects of each of the rare haplotypes. This shrinkage can be towards a common mean, with the effects of rare haplotypes shrunk to a similar degree as those haplotypes with which they are most similar (Tzeng *et al.* 2003). Alternatively, the effects of rare haplotypes can be shrunk toward the effects of haplotypes that are most similar to the rare one (Tanck *et al.* 2003). Beyond rare haplotypes statistical procedures have been considered to reduce parameterization through the evaluation of haplotype dependency and complementariness (Conti and Gauderman 2004; Guo and Lin 2009; Li *et al.* 2007).

An alternative approach to circumvent the problem of many haplotypes is by grouping similar haplotypes in the hope that such similarity will reflect a shared ancestry. Thus the parameter space can be reduced while, it is hoped, retaining that phase information relevant to the causal variant. The concept is that dependency amongst haplotypes will be accounted for, whilst reducing the DF of the test and improving power. The most common form of clustering is based on cladistic analysis which uses the concepts of a coalescent history amongst haplotypes (Durrant *et al.* 2004; Seltman *et al* 2003). Haplotype clustering methods that use cladistic approaches, or similarity measures, may encounter problems when the region under study exhibits a complex pattern of LD, produced by a series of recombination events. These models assume haplotypes are caused by mutation events alone, therefore, if recombination is considered these cladistic trees become a network of connected haplotypes, which are difficult to model, even using sampling algorithms (Larribe *et al.* 2002; Nordberg and Tavare 2002).

5.5 Conclusions

A common finding in genome-wide association studies is that only a small number of SNPs exceed the specific significance threshold, with these markers typically only explaining a small proportion of the trait variance (Bodmer and Bonilla 2008; Frazer *et al.* 2009; Maher 2008; Visscher 2008). Given the difficulty in explaining genetic variance, this raises the possibility that large numbers of markers are linked to causal variants but with lower levels of association, unable to reach stringent genome-wide thresholds. In most studies these markers are generally ignored because of their lack of statistical significance. As shown here, and in chapter four, optimal performance of a model is strongly influenced by the genomic architecture of the trait and markers, suggesting that associations are being missed through a failure to use the optimal model for a given marker or set of markers.

Optimal performance of haplotype models, such as Hap_all_prob, is strongly influenced by window length. Compared to main effect models, 'model best' marker windows from haplotype models typically have lower levels of LD between themselves and the sQTL. Whilst this is advantageous for three or five-marker windows, with their comparatively low levels of diversity, it leads to poor performance when window size is seven or nine markers. This suggests that haplotype models are able to utilize information contained in low LD regions, provided overall model parameterization is considered.

CHAPTER SIX

FALSE POSITIVE RATES AT DIFFERENT HERITABILITIES FOR SINGLE AND MULTILOCUS MODELS

6.1 Introduction

6.1.1 Multiple testing

Whole genome association mapping experiments involve large numbers of multiple hypothesis tests. Evaluation of results from such studies now have a strong prior assumption that a certain proportion of statistically "true" tests will in fact be caused by statistical chance and have no biological meaning. Such examples are called false positives, or type I errors. Recognition of this problem has lead to almost all studies imposing some form of correction factor, or threshold adjustment, to account for the difficulty in identifying true positive results. Ideally, these methods need to identify true associations and account for non-independence, or correlations between tests. These correlations can be viewed as a product of LD between markers within a dataset, and therefore, are difficult to account for, except in circumstances of extremely high, or extremely low LD (Lander and Kruglyak 1995).

The bonferroni correction is a popular and easy to implement method to correct a significance threshold to account for multiple testing (Hochberg and Tamhane 1987). It is popular because of its simplicity, as it only requires the adjustment of the significance threshold α_{nom} by α_{nom} / m, where m is the number of tests. However, Bonferroni's method is conservative when tests have some positive correlation, as is the case for markers in LD with one another. More recently, the false discovery rates (FDR) introduced by Benjamini and Hochberg (1995) and variations (Efron and Tibshirani 2002) have gained support, as we may tolerate some type I errors, provided they are a small proportion of the rejected hypotheses. FDR is defined as the proportion of false positives among the claimed positives, and involves identifying the top *r* ranking tests as true positives, where *r* is chosen based on an acceptable level of false positives.

Re-sampling approaches, such as the permutation based method introduced by Churchill and Doerge (1994), estimate a significance threshold value that accounts for correlations amongst tests. The quantitative trait data are permuted with respect to the marker data a large number of times to effectively sample from the distribution of the test statistic under a null hypothesis of no association between the phenotypes and genotypes. The great advantage of this approach is that it is intuitive, does not rely on distributional assumptions regarding the quantitative trait, and its general applicability in different population structures (Piepho 2001). A commonly cited criticism of permutation analyses is that they are computationally demanding. For example, to compute a significance threshold for a genome wide type one error rate of 0.05, at least 1000 permutations are required to obtain a reasonably accurate estimate of the threshold (Churchill and Doerge 1994).

6.1.2 Power and false positives

Association studies are based on the fundamental assumption that the genetic variants underlying a phenotypic trait will co-segregate with the trait of interest in a given population. The statistical analyses are thus aimed at identifying the markers whose genotypes correlate best with the trait values across a population of individuals. Clearly, factors affecting the characteristics of either or both the phenotypic or genotypic data can severely affect the power and accuracy of detection.

The type I error, defined as α, is the probability that the null hypothesis will be rejected, even though it is correct. The probability of not rejecting the null hypothesis when the alternative hypothesis is true is called the type II error, β. Type I and II errors are typically called false positives and false negatives respectively. Statistical power of an experiment is the probability that the null hypothesis is rejected when it is incorrect, hence, is equal to $1-\beta$. Whilst power is influenced by genetic and phenotypic properties specific to a study, it also depends on the rate of type I errors that are accepted. In other words power will be influenced by the constraint placed on the false positive rate (Weller 2001). To make an estimate of the power of a study design the rate of type I and II errors at different significance thresholds need to be determined or estimated. Both type I and II error rates will be influenced by the model fitted in the hypothesis test and its interaction with characteristics of genetic architecture, such as heritability and allele frequencies of the causal variant, although, not necessarily in the same way (Van der Beek *et al.* 1995; Weller 2001).

If genetic variance explained by a causal variant is small, models will have a lower power for their detection, due to increased rates of both false negatives and positives. Hassen *et al.* (2009) showed that heritability of the trait had a limited impact on SNP effect estimates, using a single-marker regression model and fitting SNPs as covariates in a mixed model. This is consistent with statistical theory, which shows that, under normality, estimates of fixed effects are not biased by the use of an incorrect variance-covariance matrix for residuals (Searle, 1987). Sorensen and Kennedy (1986) noted that when true variances were replaced by values estimated from data, estimates of genetic and environmental effects were unbiased. Heritability is expected to have a large impact on the SE of SNP effects (Hassen *et al.* 2009), with the SE of SNP effect estimates decreasing as heritability gets larger. As heritability has little influence on SNP effect estimates, but a substantial impact on the SE of the estimates, this can result in considerable impact on p-values of the test statistic. High values of heritability can lead to conservative evaluations of significance of SNP effects, whilst low levels of heritability may lead to a high proportion of false positives.

The examples above have described properties of power using hypothesis tests incorporating single-marker regression models. When multilocus tests are used, discerning the effect of genetic conditions on power becomes more difficult, partly due to alternative uses of information contained between markers, and consequently the parameterization of the hypothesis test. It has been suggested that haplotype methods would be more powerful than single-marker tests due to their simultaneous use of information contained between markers and between markers and a causal variant (Akey *et al.* 2001). However, genetic conditions and parameterization of the models are expected to considerably influence the performance. For example, in situations where the causal locus is genotyped directly (or in very high LD with a local marker), single-locus tests are expected to outperform the haplotype-based analyses (Zhang *et al.* 2002).

There are a number of studies that have compared power of haplotype-based tests to those of single-marker models, although these typically focus on comparisons using case-control phenotypes rather than continuous data (Akey *et al.* 2001; Morris and Kaplan 2002; Pe'er *et al.* 2006; Zaykin *et al.* 2002). Whilst these studies have demonstrated an advantage in the power of haplotype based tests under a range of population genetic parameters, they differ considerably in their methods used to infer haplotypes, hypothesis tests, and simulated data. Using regression based haplotype models, applied to case-control data, Zaykin *et al.* (2002) showed that the gain in power over single-marker models was greatest in regions of low LD and when the causal variant had low MAF. Whilst this supports the results shown in chapters four and five, it is difficult to know how the haplotype models would have performed using continuous data, given that there is a clearly demonstrated advantage to using haplotypes for case-control phenotypes (Schaid 2004). Other studies have suggested there is an advantage in power for single-marker tests (Fan and Xiong 2002; Grapes *et al.* 2004; Nielsen *et al.* 2004), although conclusions from these are also constrained by the use of specific simulated datasets, such as those exhibiting very high levels of LD between markers (Grapes *et al.* 2004).

6.1.3 Aims

The previous two chapters have demonstrated how genomic factors such as LD patterns and sQTL allele frequencies influence the performance of models that use information contained by marker in different ways. We have shown a clear advantage to using models such as Hap_all_prob when mapping for causal variants that are expected to have low to intermediate allele frequencies, and that optimal length of a fixed multilocus window is dependent on genomic conditions. The use of a localised 50 marker test region and a relatively high heritability meant that rates of type I errors were not investigated for these studies. Here our aim is to determine rates of false positives using the models described in detail in chapters four and five, using a variety of population genetic parameters, such as allele frequency of the causal variant, LD patterns and heritability of the sQTL. A total of 300 sQTL are chosen, assigned a range of heritabilities and mapped for in a genome-wide association analysis with a series of models representing the full spectrum of model parameterization and length of sliding windows.

6.2 Materials and methods

6.2.1 Dataset

The dataset used here is described in detail in chapter four, and comprises of genotypes for 12046 SNP markers in 200 individuals in a single line of broiler chickens supplied by Aviagen Ltd. Chromosome and positional information were supplied for each marker. Markers were spread across the genome, although there is uneven spacing between markers. Individuals used in this study were from a commercial population under selection, comprising of a complex pedigree structure, with a few small half-sib groups. In these cases, care was taken to ensure that no more than three animals were selected from each sire group to avoid over representation of sire haplotypes. Markers with missing values and those represented by less than three genotypes were removed from the dataset, leaving a total of 7910 markers. These remaining markers constituted the genome panel from which a single marker was chosen to represent a QTL (sQTL), which was then tested for in a genome-wide association analysis using a series of models. In total 300 sQTL were chosen, split into three groups of one hundred based on a range of MAF, assigned a range of heritabilities and included in a genome-wide analysis with each of the models. This allowed us to evaluate the rates of false positives under a range of simulated sQTL heritabilities.

LD between markers in this genome panel was used to simulate the marker-QTL LD we would expect in a typical dataset, as markers were removed from the panel and used to represent sQTL. LD properties of the genome panel were measured using r^2 (Hill and Robertson 1968). Whilst this dataset is the same as that used in chapters four and five, LD properties are expected to vary slightly as markers at the start and end of each chromosome were included along with those on the micro chromosomes. To determine the extent of LD within this population, r^2 was calculated (following Hill and Robertson 1968) for all possible syntenic and non-syntenic marker pairs. Measures of LD were also calculated for markers or windows that provided false positive associations.

6.2.2 Formation of the test panel

In total 300 sQTL were chosen split into three distinct MAF bins of 100 sQTL. The bins consisted of sQTL with MAF in the ranges of 0-0.05, 0.05-0.1, and 0.45-0.5, in other words the two lowest and the highest bins from analyses in chapters four and five.

This distribution of MAF was chosen to represent the extremes of the range, with greater number of low MAF based on the higher level of differentiation in model performance seen in chapters four and five. sQTL were chosen by randomly selecting markers from the index of markers within the given MAF range. Therefore, sQTL were distributed randomly across the genome panel. The entire genome panel consisted of 7910 SNP makers, with at any given time a single marker dropped from it to represent the sQTL. As before sQTL markers were not included in the genome panel, or used to infer haplotypes.

6.2.3 Heritability of the sQTL

For each sQTL a range of heritabilities were simulated by adding a value randomly drawn from a Gaussian distribution to the genotype of each individual. The variation of this distribution in relation to the variation of the sQTL genotypes represents the heritability of the "trait". All sQTL were biallelic, with an assumed additive effect a (therefore the difference between alternative homozygotes is $2a$) and allele frequencies p and q (= 1 - p), the variance of the sQTL ($\sigma^2_{sQTL\_geno}$) is $2pqa^2$. For a given sQTL the variation of the distribution from which values were randomly added to genotypes was determined under the following formula; $\sigma^2_{noise} = (\frac{1}{h^2}\sigma^2_{sQTL\_geno}) - \sigma^2_{sQTL\_geno}$, where h^2 is the heritability value chosen for the sQTL, and $\sigma^2_{sQTL\_geno}$ is the variation of the sQTL genotypes. Therefore heritability of the sQTL 'phenotype' is $h^2_{sQTL} = \sigma^2_{aQTL\_geno} / \sigma^2_p$, where σ^2_p is the phenotypic variance given as $\sigma^2_{sQTL\_geno} + \sigma^2_{noise}$.

Five different heritabilities were simulated for each sQTL, 0.3, 0.2, 0.1, 0.025 and 0. Five different heritabilities were simulated for each sQTL, 0.3, 0.2, 0.1, 0.025 and 0. The final heritability of 0, assumes no genetic component to variation, therefore, only the randomly generated phenotype was included in the y vector. This provides a base-line to estimate the 'true' rate of false positives across the genome for each model. The upper bound heritability of 0.3 was chosen based on the permutation threshold results described in chapter four. A high heritability of the 'trait' can result in inflated permutation thresholds for sQTL with low MAF due to the skewed distribution of the 'phenotype'. High heritability of the sQTL can result in inflated p-values for sQTL with low MAF, due to the skewed distribution of the phenotype.

6.2.4 Models

Models used in chapters four and five were chosen to investigate the performance of methods that represent the spectrum of parameterization from single to multiple markers, and then the optimal use of alternative sizes of marker windows. These models are implemented here to test their rate of false positive across a range of heritabilities. To recap; all models are regression based, with major differences in the composition of the $\boldsymbol{X}$ matrix, and the consequential parameterization of the model. Four main models were used, a single-marker regression, a main effect multi-marker regression and two haplotype based regression models. The main effects and haplotype models fit information from a number of adjacent markers implemented in a fixed size sliding window design. Window sizes of three, five and seven markers are tested here. Nine marker windows were not included in this analysis due to their poor performance across a range of genomic conditions (chapter five, figure 5.6), and to reduce the number of comparisons between models. The haplotype analyses require marker phase to be inferred prior to testing their effects in the hypothesis test. Phase was inferred under the two model assumptions (for details see chapter four) for the entire genome using the same sized sliding windows as used by the multi-marker models. Initially all markers, including potential sQTL, were included. When the genome-wide association analyses were conducted, any haplotype windows that included the sQTL marker were re-phased with the sQTL removed. This increased computation efficiency of the analysis as it did not require an entire genome haplotype panel for each sQTL, only the re-analysis of a small sub-set of windows per sQTL. Summary details for each model are given in table 6.1.

6.2.5 False positive rates

Empirical genome-wide significance thresholds were determined for each combination of model, sQTL heritability, and MAF bin using a permutation analysis. A single sQTL in each MAF bin was randomly selected and included in a 10,000 cycle permutation analysis for each combination of model and heritability. The permutation analysis is described in detail in section 2.2.3. Due to computational demands, we were unable to run a permutation analysis for each combination of sQTL, model and heritability. Although sQTL were chosen from each MAF bin, differences in their allele frequencies are not expected to affect threshold values for the heritabilities used here. In chapter four we established that for heritabilities of 0.3 and below, the threshold values from permutation analyses remained constant across the entire range of sQTL MAF for all models. A number of sQTL were chosen to provide a consensus on the threshold value for a given model and heritability. In all cases the genome-wide empirical thresholds (p < 0.05) were identical within 2 decimal places for all sets of models with heritabilities. Permutation analyses (see section 2.2.3) were used to provide accurate estimations of empirical genome-wide thresholds ($p < 0.05$) (Churchill and Doerge 1994) for each model and sQTL heritability. These thresholds are expected to change slightly between models due differences in the correlation between individual tests. For example, the use of overlapping sliding windows in multilocus approaches means that individual tests are expected to have a higher degree of correlation than a single-marker analysis. Genome-wide threshold values for each model and heritability are given in table 6.1.

For each analysis, any marker or window located on a different chromosome to the sQTL with an association above the genome-wide threshold ($p < 0.05$) was deemed a false positive. sQTL simulated with a heritability of 0, have no genetic component or marker genotypes included, and therefore no positional information. In these analyses all markers and windows in the genome panel that showed association levels above the thresholds were identified as false positives. The rate of false positives for a given analysis is the sum of the number of such false positives divided by the total number of markers either not on the sQTL chromosome (for sQTL with $h^2 > 0$), or the total number of markers in the test panel (for sQTL with $h^2 = 0$). The level of LD between false positive markers and the sQTL was estimated using the r^2 statistic.

			$p < 0.05$ Thresholds				
Model	**Markers**	**Parameterization**	$h^2 = 0.3$	$h^2 = 0.2$	$h^2 = 0.1$	$h^2 = 0.025$	$h^2 = 0$
Single	1	One marker. Genotypes fitted as linear covariates	4.82	4.81	4.79	4.75	4.74
Main_order_3	3	Three markers. Just main effects fitted	4.78	4.77	4.76	4.73	4.71
Main_order_5	5	Five markers. Just main effects fitted	4.74	4.71	4.70	4.68	4.67
Main_order_7	7	Seven markers. Just main effects fitted	4.71	4.70	4.69	4.67	4.66
Hap_highest_prob_3	3	Three markers. High probability haplotypes treated as observed	4.73	4.71	4.68	4.66	4.65
Hap_highest_prob_5	5	Five markers. High probability haplotypes treated as observed	4.71	4.70	4.68	4.65	4.63
Hap_highest_prob_7	7	Seven markers. High probability haplotypes treated as observed	4.70	4.69	4.67	4.65	4.62
Hap_all_prob_3	3	Three markers. Haplotypes modeled as probabilities	4.72	4.71	4.68	4.65	4.64
Hap_all_prob_5	5	Five markers. Haplotypes modeled as probabilities	4.70	4.69	4.67	4.63	4.62
Hap_all_prob_7	7	Seven markers. Haplotypes modeled as probabilities	4.68	4.66	4.63	4.59	4.57

Table 6.1

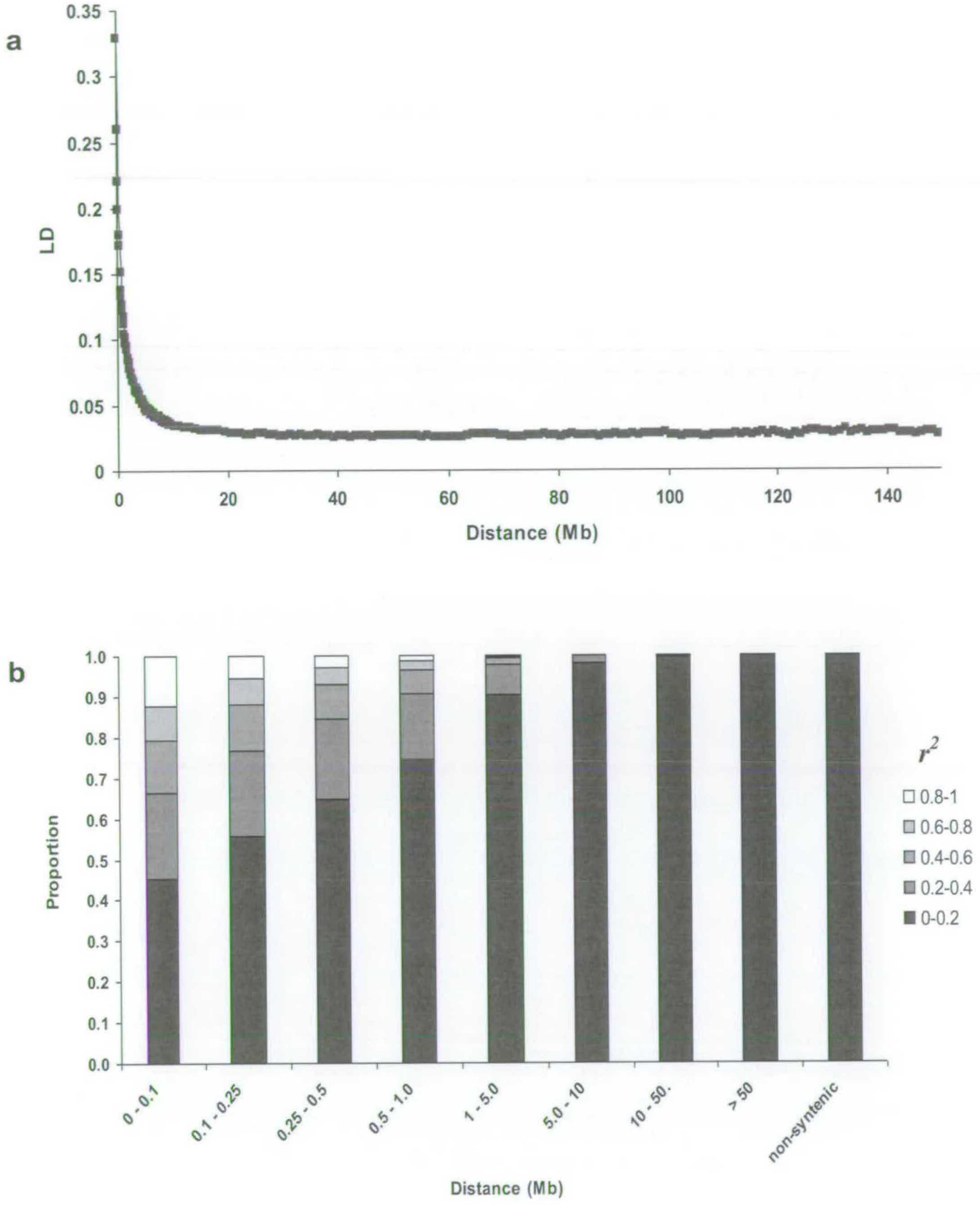
Summary of models and their parameterization along with the empirical genome-wide thresholds derived using a 10,000 cycle permutation analysis for each model and heritability of the sQTL.

Naturally, no estimate of LD can be obtained for false positive markers in analyses were sQTL heritability is 0.

6.3 Results

6.3.1 Patterns of LD

Estimates of LD, using r^2, were calculated for all syntenic pairs of markers. Figure 6.1a illustrates the decline of LD with distance. Results shown here are for the whole genome, although the pattern of high LD at short distances with a steep decline as distance increases was common for all chromosomes. The extent of LD for non-syntenic marker pairs was also determined by calculating r^2 values for all non-syntenic pairs in the dataset. The mean pairwise r^2 value for non-syntenic pairs was 0.035 (SE = 0.00012), equal to the mean for syntenic pairs separated by greater than 20 Mb. Figure 6.1b summarizes the frequency distribution of r^2 by distance for all syntenic pairs and non-syntenic pairs. About 11 percent of marker pairs within 0.1 Mb had r^2 values greater than 0.8 and this dropped to 1 percent for marker pairs between 0.5 and 1 Mb apart. 91 percent of markers between 1-5 Mb had r^2 values less than 0.2, with this raising to 97 percent for markers 5-10 Mb and nearly 100 percent for markers greater than 10 Mb apart. About 33 percent of markers within 0.1 Mb had r^2 values greater than 0.4, this dropped to about 15 percent for markers 0.25 – 0.5 Mb and about 2 percent for pairs between 1 – 5 Mb. The distribution of r^2 at distances greater than 10 Mb was similar to that of non-syntenic marker pairs with 99.96 percent of values less than 0.2. The distribution of maximum r^2 of a SNP with all other SNPs (De Roos *et al.* 2008) suggests that SNPs found to be associated with a trait in association studies are very likely to be near relevant QTL.


a
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
LD
0
20
40
60
80
100
120
140
Distance (Mb)
b
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Proportion
0 - 0.1
0.1 - 0.25
0.25 - 0.5
0.5 - 1.0
1 - 5.0
5.0 - 10
10 - 50.
> 50
non-syntenic
Distance (Mb)
r^2
□ 0.8-1
▨ 0.6-0.8
▨ 0.4-0.6
▨ 0.2-0.4
■ 0-0.2

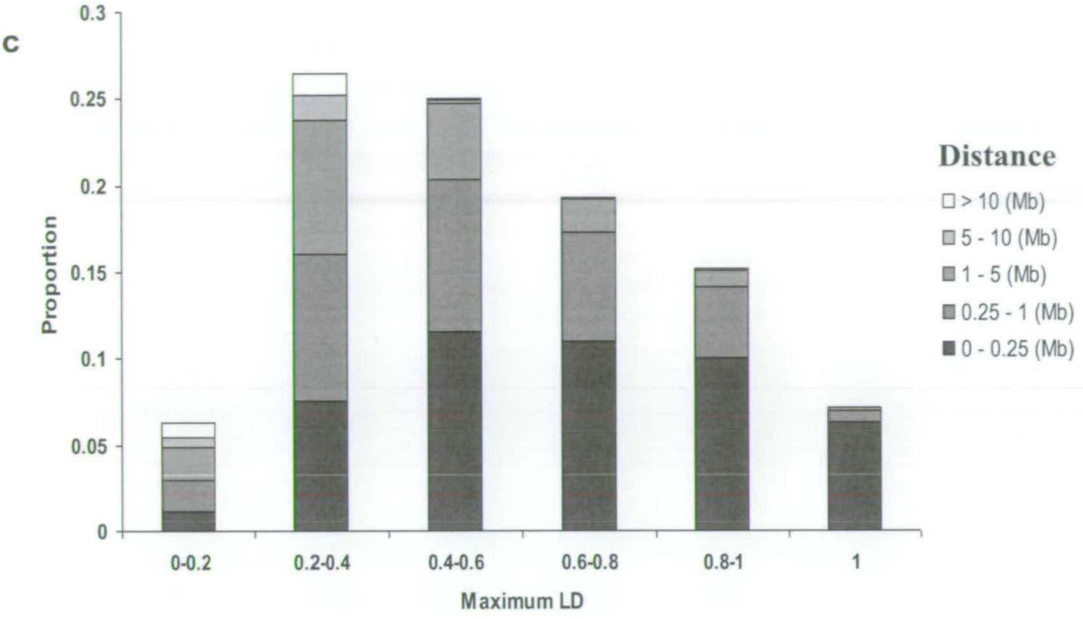


Figure 6.1

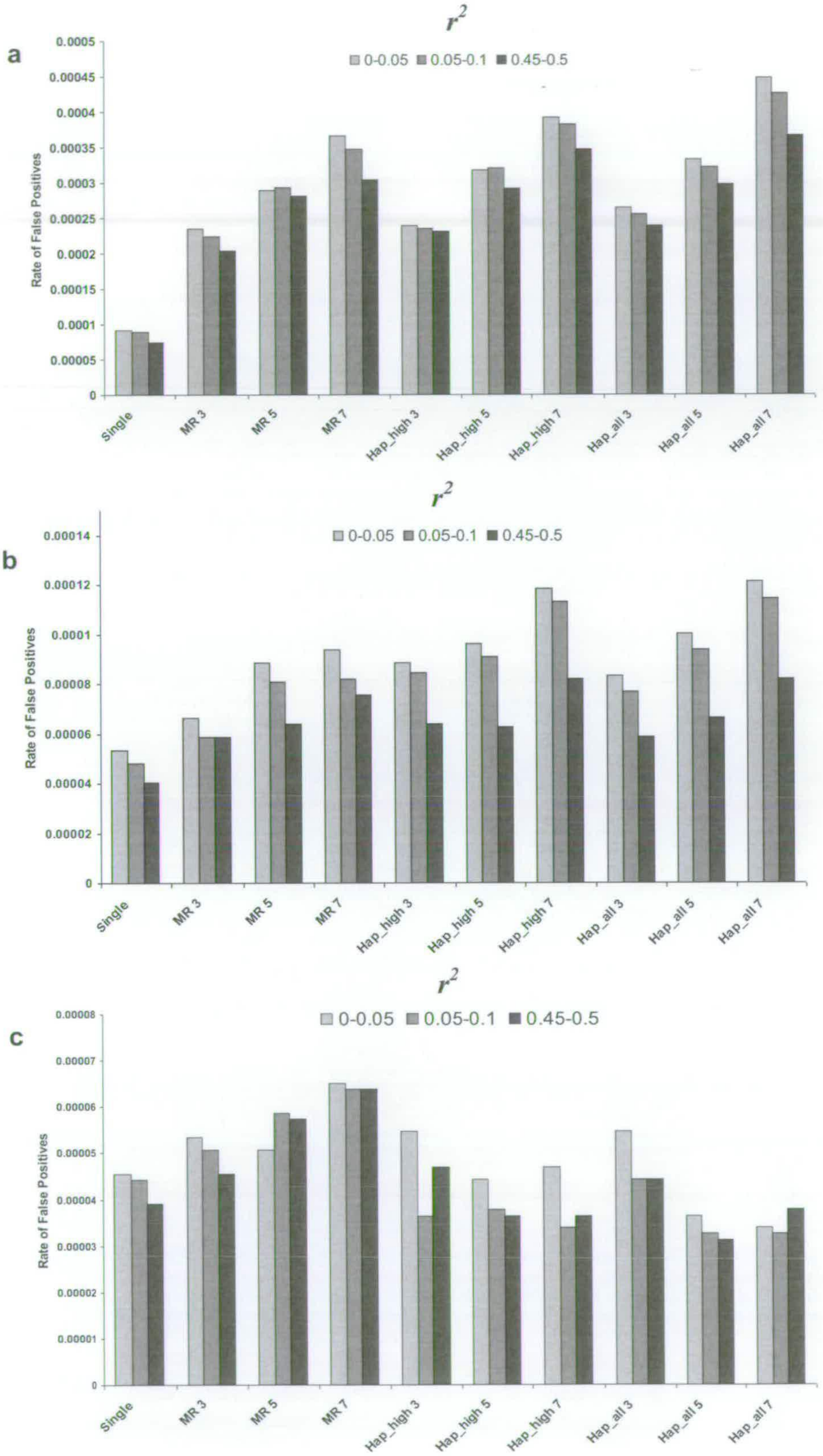
a) Decline of LD measured by r^2 against distance in Mb. Points shown are mean pairwise LD for syntenic marker pairs in the whole genome against mean distance. **b)** Frequency distribution of LD (r^2) for syntenic and non-syntenic marker pairs for each line. LD proportions are shown for syntenic markers dived based on marker distance bins. **c)** The frequencies of maximum LD of syntenic SNPs based on r^2. Bins were created on the basis of distance to the SNP for which the maximum LD was registered.

This distribution is shown graphically in figure 6.1c, and separated into bins on the basis of the distance between the SNP and its maximum r^2 SNP. About 40 percent of SNPs had a maximum r^2 greater than 0.6 and 93 percent of SNPs had a maximum r^2 greater than 0.2. For all maximum r^2 value bins greater than 0.2, the shortest-distance bin (< 0.25 Mb) was the most frequent, and the vast majority of maximum distances were less than 1 Mb. For SNPs with a maximum r^2 greater than 0.6, only 3 percent were greater than 1 Mb apart.

6.3.2 Rate of false positives

The rate of false positives, shown as the chance that a single test will be a false positive, are given for each combination of model, heritability and sQTL MAF bin in figure 6.2. The rate of false positives seen for sQTL within the high MAF bin is consistently lower than the rate observed for sQTL in the lowest bin, for all models and heritability combinations. The average false positive rate from the highest MAF bin is 80 percent of the rate for the lowest bin. No trend, between models or heritability, appears to exist in the value of this difference.

For a given model, heritability of the sQTL plays an important role in the rate of false positives. For all models the highest rates of false positives are observed when the sQTL has a $h^2 = 0.3$. As heritability falls, so does the rate of false positives, reflecting the decrease in power of the models to detect associations from markers with spurious levels of LD with the sQTL. When the sQTL has a $h^2 = 0.3$ there is a considerable


a
r^2
0-0.05
0.05-0.1
0.45-0.5
Rate of False Positives
0.0005
0.00045
0.0004
0.00035
0.0003
0.00025
0.0002
0.00015
0.0001
0.00005
0
Single
MR 3
MR 5
MR 7
Hap_high 3
Hap_high 5
Hap_high 7
Hap_all 3
Hap_all 5
Hap_all 7
b
r^2
0-0.05
0.05-0.1
0.45-0.5
Rate of False Positives
0.00014
0.00012
0.0001
0.00008
0.00006
0.00004
0.00002
0
Single
MR 3
MR 5
MR 7
Hap_high 3
Hap_high 5
Hap_high 7
Hap_all 3
Hap_all 5
Hap_all 7
c
r^2
0-0.05
0.05-0.1
0.45-0.5
Rate of False Positives
0.00008
0.00007
0.00006
0.00005
0.00004
0.00003
0.00002
0.00001
0
Single
MR 3
MR 5
MR 7
Hap_high 3
Hap_high 5
Hap_high 7
Hap_all 3
Hap_all 5
Hap_all 7

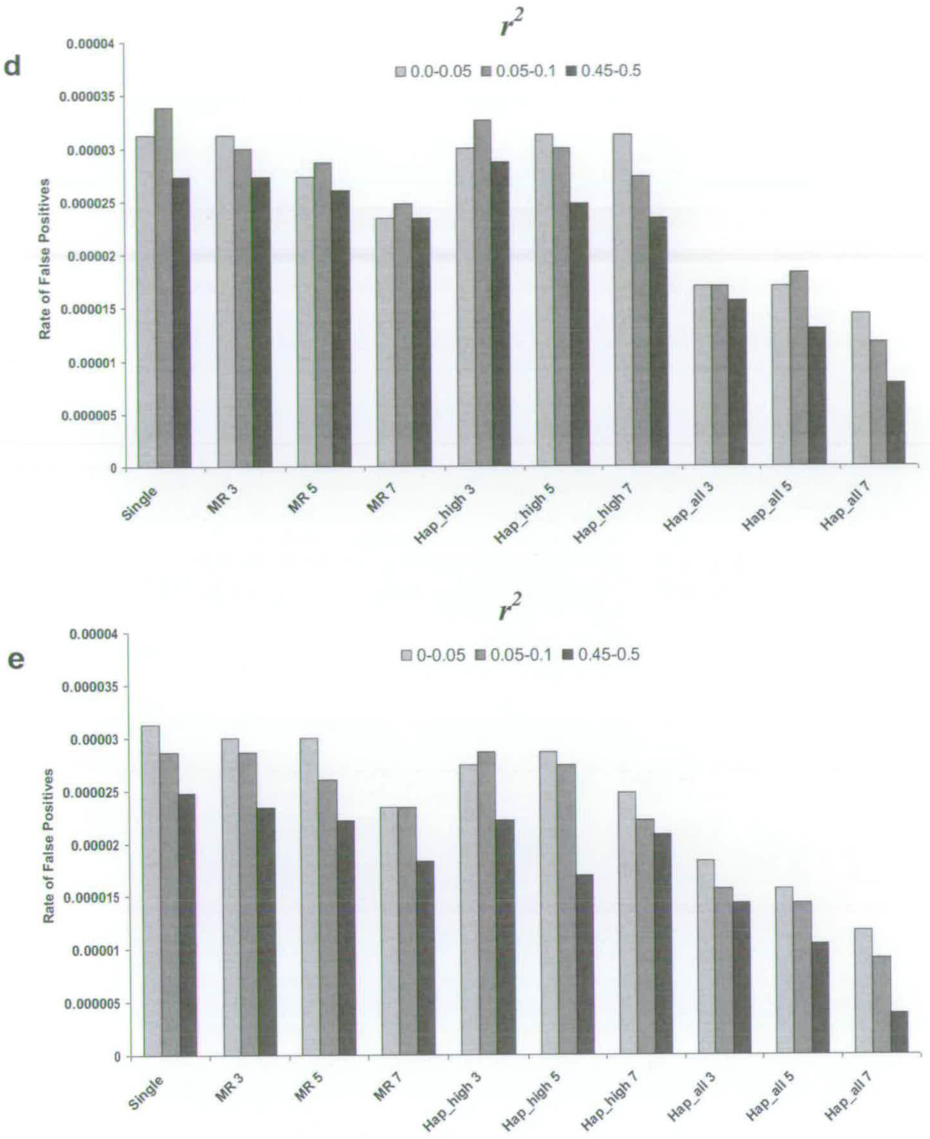


Figure 6.2

The rate of false positive tests, shown as a mean proportion of all non-syntenic tests. Markers or marker-windows located on a different chromosome to the sQTL that show a significant ($p < 0.05$) associations are termed false positives. The proportions of false positives are shown for each model, averaged across sQTL in each of the MAF bins, and heritability. Results from are shown for sQTL with **a)** $h^2 = 0.3$, **b)** $h^2 = 0.2$, **c)** $h^2 = 0.1$, **d)** $h^2 = 0.025$, **e)** $h^2 = 0$. Note: Differences in the scale of the *y*-axes.

difference in the rate of false positives between models. There is an increase in false positive rate for multilocus methods as markers are added to the windows. The majority of this difference between models is a product of using a sliding window approach for multilocus methods. False positive windows often occur next to one another in a string of false positive windows, indicating that a single marker showing a high level of association with the sQTL is causing several adjacent windows to produce significant associations. Table 6.2 shows the proportion of false positive tests that include markers found in at least one additional false positive window. For the high heritability sQTL analysis, a very large proportion of tests occur in a string of windows that contain at least one marker in common. The random nature of non-syntenic LD means that there is a low probability that a marker will be in strong LD with the sQTL, and that this marker is unlikely to be surrounded by other high LD markers. Thus, whilst in a single-marker analysis this marker will only lead to one false positive association, in a multi-marker analysis, this marker is included in a number of adjacent over-lapping windows, and produces several false positive associations. The number of windows it is included in is a same as the length of the window, resulting in an increase in apparent false positive rate as window length increases. As the heritability of the sQTL drops this artifact disappears due to the lack of power. For lower heritabilities a single marker in high LD with the sQTL is able to produce a significant association, but this significance disappears when additional markers are included and the parameterization of the model reduces the level of association. When the heritability sQTL is high, multilocus models still show a significant association despite the burden of their higher parameterization. Under these situations the majority of the variance explained by the model is from the main effects of

	Heritability				
Model	**0.3**	**0.2**	**0.1**	**0.025**	**0**
Main_order_3	0.92	0.73	0.21	0	0
Main_order_5	0.93	0.74	0.24	0	0
Main_order_7	0.98	0.74	0.21	0.01	0
Hap_highest_prob_3	0.93	0.71	0.14	0	0
Hap_highest_prob_5	0.94	0.69	0.13	0	0
Hap_highest_prob_7	0.89	0.72	0.13	0	0
Hap_all_prob_3	0.91	0.68	0.18	0.01	0
Hap_all_prob_5	0.92	0.69	0.14	0	0
Hap_all_prob_7	0.89	0.74	0.11	0	0

Table 6.2

The proportion of false positive tests that have a marker within the window that occurs in at least one other false positive window. The proportion is that of all false positive tests for a given model and heritability.

the marker in high LD with the sQTL, and little if any is added from other parameters in the model. As heritability of the sQTL falls, so does the level of significance, to the point where the window no longer produces a significant association.

The trend in different rates of false positives between models is reversed when the sQTL have low or zero heritability. In these situations the single-marker model has the highest rates of false positives, with the rate decreasing as parameterization of the models increases. The rate of false positives for the zero heritability sQTL can be viewed as the true baseline rate of false positives, as significant associations are wholly due to random associations rather than spurious non-syntenic LD. Given the sample size in this study, power to detect a significant association when the sQTL has an $h^2 = 0.025$ will be very low, unless the causal variant is in very high LD with a marker. Therefore, the similar pattern of false positive rates for sQTL with h^2 of 0.025 to that of analyses with no genetic component is also expected to be caused by random associations rather than markers showing strong non-syntenic LD with the sQTL.

The difference in the rate of false positives between the two haplotype models, across heritabilities, for windows of comparable length is shown in table 6.3. When the sQTL has heritability between 0.3 – 0.1 the rate of false positive between the two haplotype models is approximately equal. However, when the heritability is very low, or when there is no genetic component to the variance, on average the Hap_all_prob model has half the rate of false positives of Hap_highest_prob models of comparable length windows. The difference between the two models is greatest for seven-marker windows.

In an attempt to account for the overlapping false positive windows seen for multilocus models, significant associations occurring in adjacent windows were only

	Heritability				
Window length	**0.3**	**0.2**	**0.1**	**0.025**	**0**
3	0.93	1.08	0.96	1.84	1.62
5	0.98	0.96	1.18	1.78	1.81
7	0.91	0.99	1.12	2.42	2.74
Mean	**0.94**	**1.01**	**1.09**	**2.01**	**2.06**

Table 6.3

The proportional difference in the rates of false positives between the Hap_highest_prob and Hap_all_prob models, across the range of heritabilities and window lengths. Values shown are the ratio of the false positive rate for Hap_highest_prob compared to Hap_all_prob.

counted once, along with all significant independent windows. The rate of false positives, adjusted for overlaps, is shown in figure 6.3 for sQTL with heritabilities of 0.3, 0.2, and 0.1. Adjusted rates are not shown for sQTL with heritabilities of 0.025 and 0, as the proportions of overlapping windows are extremely low, or non existent (table 6.2). Whilst this could be considered a conservative estimate, as it assumes each set of overlapping false positive windows is caused by a single marker, it provides an indication of underlying rates of false positives amongst models when sQTL h^2 is high. With this adjustment, false positive rates do not show the dramatic increases with window length for multilocus models, although, there are slight differences when the sQTL has $h^2 = 0.3$. The adjusted rate for sQTL with $h^2 = 0.1$ is similar to the unadjusted rate, likely due to the low occurrence of overlapping false positive windows at this heritability (table 6.2).

6.3.3 Level of LD between false positive markers and the sQTL

Pairwise LD between the sQTL and each false positive marker was calculated. This included all markers within a window. For each combination of model and heritability the proportion of false positives with different ranges of marker-sQTL LD is shown in figure 6.4. For a false positive window, the maximum pairwise LD between a marker within the window and the sQTL is given. With all models, the distribution of LD between false positive markers and sQTL with $h^2 = 0.025$ closely resembles the distribution of background non-syntenic LD (figures 6.2d-e). As heritability of the sQTL increases the proportions of false positives with higher levels of LD between the markers

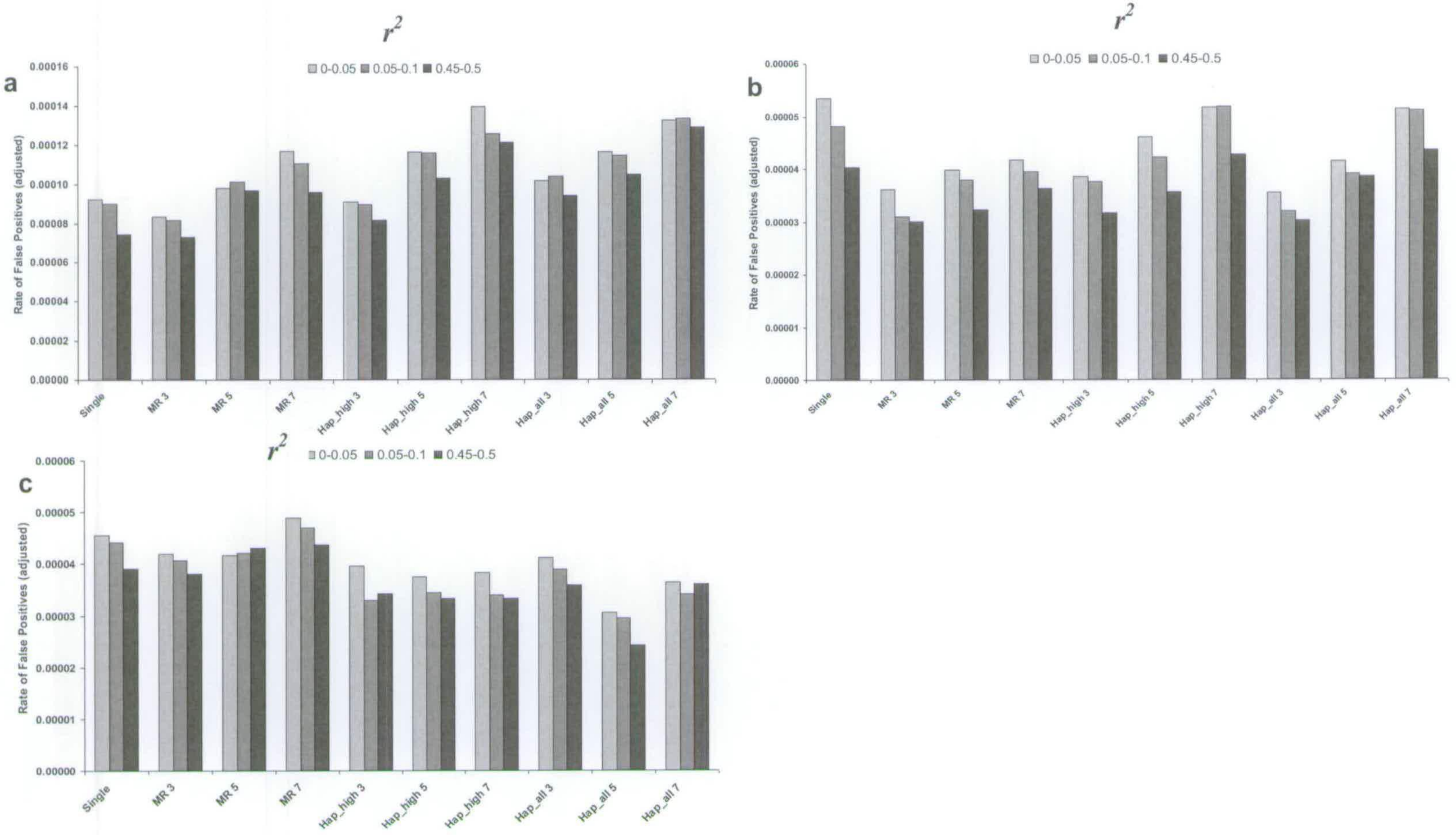
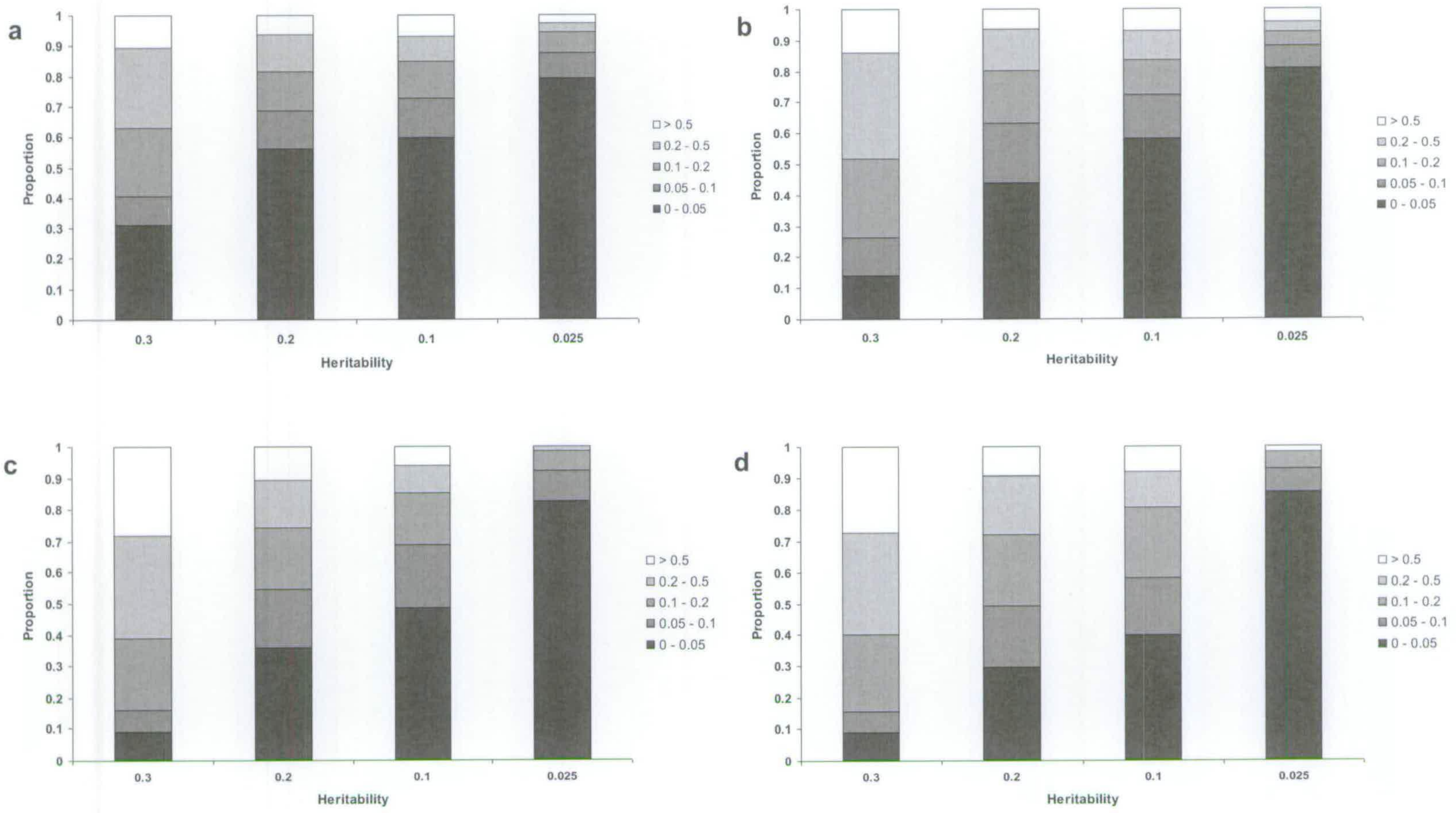
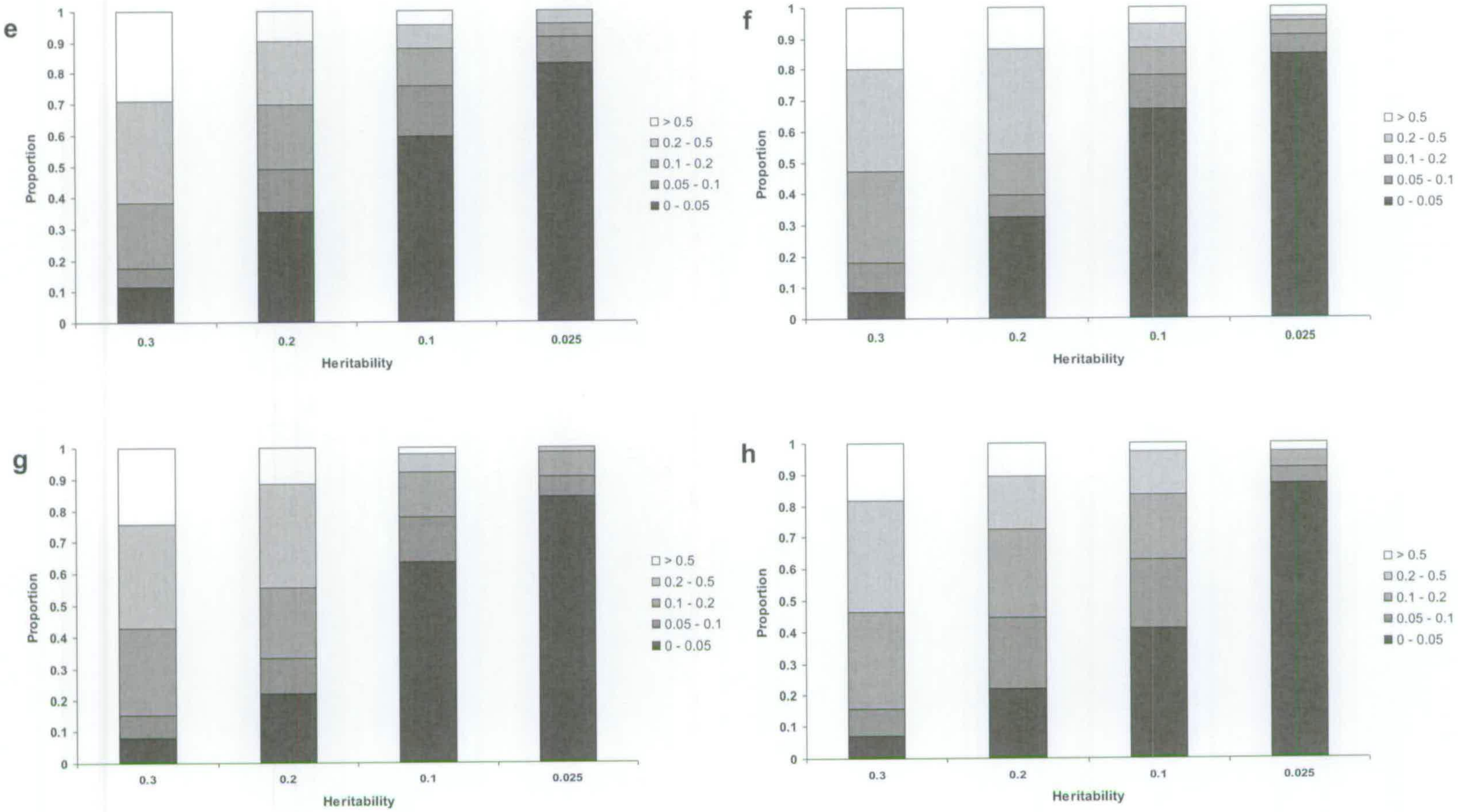


Figure 6.3

Rate of false positive tests shown as a mean proportion of all non-syntenic tests. The rate has been adjusted to remove associations amongst adjacent, overlapping significant windows. The rate of false positives is shown for each model, averaged across sQTL in each MAF bins, and heritability. Results from are shown for sQTL with **a)** $h^2 = 0.3$, **b)** $h^2 = 0.2$, **c)** $h^2 = 0$.


a
b
c
d
Proportion
Heritability
0.3
0.2
0.1
0.025
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
□ > 0.5
▨ 0.2 - 0.5
▨ 0.1 - 0.2
▨ 0.05 - 0.1
■ 0 - 0.05


e
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.3
0.2
0.1
0.025
Heritability
□ > 0.5
▨ 0.2 - 0.5
▨ 0.1 - 0.2
▨ 0.05 - 0.1
■ 0 - 0.05
f
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.3
0.2
0.1
0.025
Heritability
□ > 0.5
▨ 0.2 - 0.5
▨ 0.1 - 0.2
▨ 0.05 - 0.1
■ 0 - 0.05
g
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.3
0.2
0.1
0.025
Heritability
□ > 0.5
▨ 0.2 - 0.5
▨ 0.1 - 0.2
▨ 0.05 - 0.1
■ 0 - 0.05
h
Proportion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.3
0.2
0.1
0.025
Heritability
□ > 0.5
▨ 0.2 - 0.5
▨ 0.1 - 0.2
▨ 0.05 - 0.1
■ 0 - 0.05

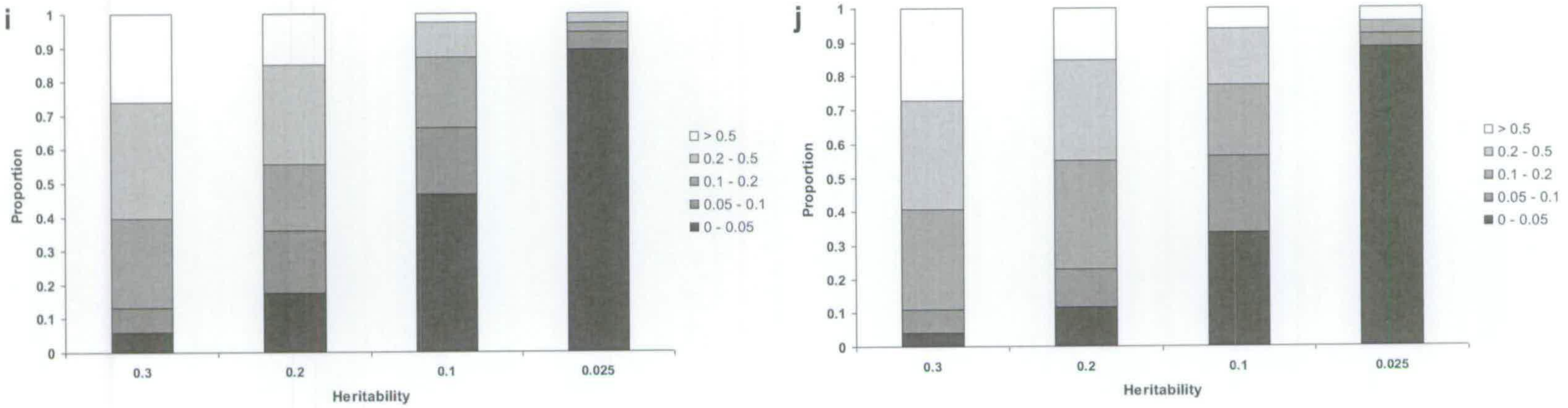


Figure 6.4

LD values between false positive markers and sQTL are shown as a proportion of all false positives. For each model proportions are shown for all heritabilities where there was a genetic component. Distributions of LD proportions are shown for sQTL in all MAF bins, as there was very little difference in distributions between MAF bins. LD is divided into a series of bins. The maximum marker – sQTL LD was taken for windows that showed a false positive association. The proportion of LD is shown for each individual model; **a)** Single-marker analysis, **b)** Main effects – 3 marker window, **c)** Main effects – 5 marker window, **d)** Main effects – d marker window, **e)** Hap_highest_prob – 3 marker window, **f)** Hap_highest_prob – 5 marker window, **g)** Hap_highest_prob – 7 marker window, **h)** Hap_all_prob – 3 marker window, **i)** Hap_all_prob – 5 marker window, **j)** Hap_all_prob – 7 marker window.

and sQTL dramatically increases, such that when models are tested against sQTL with $h^2 = 0.3$ between 10.1 – 28.2 percent of markers had LD values greater than 0.5. When the sQTL has $h^2 = 0.2$ the range of the proportion of LD greater than 0.5 is 6.4 – 15.1 percent, for $h^2 = 0.1$ it is between 2.2 – 7.7 percent and finally when $h^2 = 0.025$ it is between 0.0 – 4.4 percent. The difference in the proportion of false positive LD with values greater than 0.5 between $h^2 = 0.3$ and $h^2 = 0.025$ is smallest for the single-marker test. For multiple-marker methods, there is a trend that larger windows have a higher proportion of markers in high LD with the sQTL. This trend disappears as the heritability of the sQTL reduces. This also likely reflects that the causation of the increase in false positive rates as window length increases for high heritability sQTL is due to the inclusion of a single marker in high LD with the sQTL in a number of adjacent windows.

The proportion of false positive tests with a maximum LD between the markers and the sQTL between 0 – 0.05 increases considerably as the heritability of the sQTL decreases. For a single-marker analysis with sQTL of $h^2 = 0.3$ the proportion of false positive to sQTL LD between 0 – 0.05 is 31.6 percent. When the sQTL has $h^2 = 0.025$ this proportion is 72.1 percent, a difference of 40.5 percent. For the other models this difference ranges from 66.2 to 78.1 percent. For all multiple-marker models this difference in proportions between high and low h^2 increases as window length increases.

6.4 Discussion

Quantitative geneticists have struggled to identify genetic markers that explain large proportions of estimated genetic variance (Bodmer and Bonilla 2008; Frazer *et al.* 2009; Maher 2008; Visscher 2008), despite the availability of high density genome panels. An explanation in part, may be the difficulty in identifying the optimal model to use in mapping studies, especially given the variation in performance of a model across the genomic landscape of the genome (chapters four and five). Clearly, it is important to choose powerful and appropriate statistical methods that are designed to relate genotype or haplotype information to the phenotypes of interest. The identification of the importance of genetic architecture on performance of a variety of regression-based models (chapters four and five) leads to the question of false positive rates of the models in genome-wide studies. Difference in the use of LD information contained between markers, and between non-syntenic markers and QTL, by the models, is expected to result in differences in their rates of false positives.

Here we have shown that rates of false positives among models are strongly affected by the proportion of genetic variance explained by the sQTL. When the heritability of the sQTL is high, rates of false positives amongst the multilocus models are considerably inflated, through the inclusion of a spurious high LD marker in multiple overlapping test windows (figure 6.2). However, removal of overlapping significant windows produces a more even pattern of false positive rates amongst models, when heritability of the sQTL is high (figure 6.3). As heritability falls the true pattern of false positive rates are revealed as even high LD non-syntenic markers do not have the power to provide a significant association with the sQTL. When heritability of the sQTL is 0.025 the lowest false positive rates are observed for the Hap_all_prob method, with the rate declining as the window length increases. Across the range of heritabilities, and for all models, there is a lower rate of false positives when the sQTL has a high MAF. Rates of false positives for the two haplotype models are approximately equal when the heritability of the sQTL is high, although as this falls the Hap_highest_prob method shows an average rate of twice that of the Hap_all_prob model.

6.4.1 Effect of sQTL heritability and window length

False positives, as defined here, are significant associations between non-syntenic markers or marker windows and the sQTL. In this situation false positives can be caused by two main factors. The first is caused by a true random association between sQTL phenotypes and information contained by the marker or set of markers. Statistical models differ in their susceptibility to such associations, and their robustness in providing statistical support when a true association exists. The second is caused by high levels of non-syntenic LD. Whilst these are true associations in the sense that there is a statistical link (LD) between the markers and sQTL genotypes, the cause of this LD is expected to be random genetic drift.

In many livestock species the extent of LD can be considerable, due to recent small effective population sizes (Hayes *et al.* 2003). For syntenic markers, the typical pattern of LD observed is high levels that decline with distance. However, the relationship between marker distance and LD can be highly variable, and high levels of LD can occur between markers separated by large distances. For non-syntenic markers, LD is expected to resemble background levels caused by random drift alone. However, spurious, high levels of LD can occur, although these are usually isolated and random incidences that are not observed in clusters or high LD regions. In some situations selection can cause LD between unlinked loci that contribute to phenotypes undergoing selection (Ardlie *et al.* 2002), although this is not expected to figure in our analyses due to the use of randomly selected markers to represent the sQTL, and repetition of the number of sQTL included in each class of results. The mean background non-syntenic LD in this dataset is low, although some higher pairwise measures of LD exist.

When the sQTL has a high heritability, false positives are predominantly caused by spurious high levels of non-syntenic LD (figure 6.4). Under these conditions false positives are essentially a reflection of the ability of a model to provided statistical support for an association, as there is an observed genetic correlation between the markers and sQTL loci. For single markers, the effect of LD on the level of association is well known, although this becomes more difficult to characterize for multilocus models, especially haplotypes, where LD structure between additional markers and model parameterization are more complex (Zondervan and Cardon 2004). The majority of the proportional increase in false positive rates with marker window size is caused by single high non-syntenic LD marker in multiple windows. When the genetic variance explained by the sQTL is low, a single marker in high LD with the sQTL is able to produce a significant association, but this significance disappears when additional markers are included and the parameterization of the model reduces the level of association. When the heritability sQTL is high multilocus models still show a significant association despite the burden of their higher parameterization. Under these situations the majority of the variance explained by the model is from the main effects of the marker in high LD with the sQTL, and little if any is added from other parameters in the model. The fact that this increase in rate is not directly proportional to the difference in window sizes likely reflects more conservative critical values of the test regions with the inclusion of additional parameters. The false positive rate for all models changes dramatically as the heritability of the sQTL falls. When the sQTL has heritabilities of 0.2 and 0.1 the proportional increase in false positive rate with window length reduces further. Here, high levels of non-syntenic LD still exist, although the power of multilocus models to detect these associations is further compromised by reduced amounts of genetic variance at the sQTL. When testing for true associations, multilocus models benefit from high order LD between haplotype structures located in regions physically linked to causal variants. However, this structure of LD does not exist for non-syntenic markers, where spurious associations are typically affected by random genetic drift (McKay *et al.* 2007).

Results are discussed in terms of proportional genetic variance explained by the sQTL, although they can also be considered as a function of power, influenced by the sample size of the dataset. Regardless of the heritability of the causal variant, an increase in sample size will lead to improvements in power. In the context of this study, a larger sample size would lead to a greater number of associations detected between spurious non-syntenic markers in high LD with the causal variant. Sample size influences ability of markers to detect associations in two, interconnected, ways. For a QTL with a given heritability, an increase in sample size means that a lower range of LD between markers and the QTL will yield significant associations. Connected to this, is the improvement in the ability of markers to detect associations with QTL that explain smaller proportions of genetic variance. As the proportion of genetic variance explained by a marker is a product of the genetic variance explained by the causal variant and the extent of LD with the marker, increasing sample size could lead to greater numbers of false positives amongst non-syntenic markers for all sQTL heritabilities. It could also lead to the pattern of false positive markers being picked up in multiple adjacent windows being observed for sQTL with lower heritabilities. Nevertheless, greater sample size would also improve power to detect true associations, amongst markers close to QTL.

The occurrence of false positives for sQTL where $h^2 = 0$ represent the rate caused by random associations between traits values and markers, as no genetic competent, and thus LD, exists for these associations. In a sense, these reflect the robustness of the models in avoiding significant associations in the absence of any genetic component of trait variation. The rate of false positives for all models is low, with levels translating into approximately 0.2 false positive associations per genome scan for single-marker tests, and 0.04 for Hap_all_prob_7, given the number of markers used here. It is difficult to identify the causes of differences between models in their ability to avoid significant associations with randomly generated traits. A possible cause may be differences in the likelihood of marker alleles mimicking the trait distribution by chance. The residual variance used in this analysis was drawn from a Gaussian distribution, with the ability of random associations to occur if the genotypes or haplotypes of individuals correlate with the trait values. This situation is more likely to occur when genotype information is used from a marker with low MAF. The random occurrence of value from the tails of the distribution with the rare genotype class for an individual could lead to a significant association, partly caused by the inability of the model to accurately estimate the effect of the marker. The more complex pattern of genotype and haplotype structures for the higher parameterized models may make this less likely to occur, especially when we also penalize with a large number of DF.

6.4.2 Comparison of haplotype models

The difference in the rate of false positives observed for the two haplotype models may be due to the use of information from ambiguous haplotypes. As has been discussed in previous chapters, the assumption made for Hap_highest_prob models, that haplotypes are known without error, can lead to increases in errors for accurately estimating haplotype effects (Morris *et al.* 2004). Here we have shown that the rate of false positives for the two haplotype models is approximately constant when heritability is high, although the rates are considerably different when the sQTL has very little or no genetic variance. When the sQTL $h^2 = 0$ no haplotypes are expected to have any effect of the trait. Thus, it is difficult to understand how any errors potentially introduced by not accounting for haplotype uncertainly could lead to an increased rate of false positives. It has been shown that a loss of power can occur due to haplotype phase uncertainty, although this is due to a decrease in the ability of models to accurately identify true associations, hence and increase type II rather than type I errors (Schaid 2005; Zaykin *et al.* 2002).

6.4.3 Implications for mapping studies

For analyses shown here and in the previous two chapters, an upper bound of sQTL heritability of 0.3 was used to provide a sensible level of power for the models, given that sample sizes of only 200 individuals were available. In a typical mapping study for complex traits, genetic heterogeneity is expected to be extensive, with the genetic effects shared across multiple, possibly independent, loci. In this situation we expect the genetic variance explained by a single causal variant to be low. Thus, comparison of false positive rates amongst models, in relation to the implications for mapping studies, should be made from results of sQTL $h^2 = 0.025$. Whilst this still represents a high level of genetic variance explained by a single causal variant, it is a more realistic value than other heritabilities used here, especially when we consider the sample size. There has been some suggestion that variation in the levels of non-syntenic LD is due to a partial dependence on marker information content (Farnir *et al.* 2000). Although the extent of the effect is unknown, increasing sample size is expected to decrease the dependence of non-syntenic LD measures on marker heterozygosity (Farnir *et al.* 2000; Khatkar *et al.* 2008; McRae *et al.* 2002). A complication, however, is that these studies used D' as a measure of LD, which is known to be upwardly biased for small sample sizes (Lewontin 1988).

False positives caused by significant gametic associations were also shown to be very common between non-syntenic loci. With this sample size this is not a problem when the genetic variance of the causal variant is low, however, this may become a problem if larger sample sizes are used. In such situations, the common occurrence of high levels of non-syntenic LD raises serious concerns about the generation of false positive results, especially for the multilocus models. Despite the computational demands it would be recommended to evaluate the level of such LD patterns prior to mapping. Alternatively, mapping methods that combine both linkage and LD information could be applied, although this may reduce the chances of finding true positives associations (George *et al.* 2000; Lee and Van der Werf 2002; Marchini *et al.* 2004 Zhao *et al.* 2007).

6.5 Conclusion

The observed differences in model performance, under a range of genomic conditions, as shown in chapters four and five, raised the question of differences in the rate of false positives between the models. Differences were expected given the alternative use of information contained between the markers and sQTL. An evaluation of these rates across a range of sQTL heritabilities was deemed necessary to fully evaluate the likely performance of models in identifying causal variants in traditional mapping studies. The higher levels of false positives seen here when mapping for sQTL with h^2 = 0.3 reflect differences in the ability of models to provide statistical associations, given the majority of these are due to spurious high levels of non-syntenic LD. On the other hand, false positives produced by models when mapping for sQTL that have no genetic component to their variation (h^2 = 0) represent the true nature of false positives caused by random associations between the trait and marker information.

We have seen here that the proportion of genetic variance explained by the sQTL strongly influences the rates of false positives amongst the models, with the high parameterized and longer window length models showing higher rates than the single-marker model. However, as this increase is caused by high levels of non-syntenic LD, these rates provide further support for the ability of these models to identify significant associations when there is a genetic link between markers and the causal variant. When heritability of the sQTL is 0 the lowest false positive rates are observed for the Hap_all_prob method, with the rate declining as the window length increases. Whilst little can be done about high levels of non-syntenic LD, the results shown here provide comfort that models such as Hap_all_prob, are able to correctly identify markers genetically linked to QTL, and minimize the levels of significant results caused by random associations between trait phenotypes and haplotypes. In typical mapping studies, genetic variance of an individual causal variant is expected to be low. Here we have shown that under such conditions, a haplotype model such as Hap_all_prob has a lower rate of false positives than other models tested here, with this rate decreasing as window size increases. Whilst other factors, such as the ability to identify true QTL, need to be taken into account, the results shown here validate the use of haplotype methods over single-marker and main effect models.

CHAPTER SEVEN

MAPPING FOR CAUSAL VARIANTS ASSOCIATED WITH ASCITES SUSCEPTIBILITY IN BROILER CHICKENS

HAPLOTYPE ANALYSIS

7.1 Introduction

The last few years have seen extensive efforts to identify genetic variation and explain its effect on phenotypic differences in populations. GWAS studies are the most widely used approach to locate causal genetic variants, with some major successes in livestock species (for review see: Abasht *et al.* 2006; Ron and Weller 2007) and humans (for review see: Altshuler *et al.* 2008). Often numerous QTL, spread across the genome, have been identified for complex traits, suggesting that their genetic control is influenced by a number of loci. However, these studies have struggled to replicate results in different populations, and to explain large proportions of the genetic variation from the identified markers. A well known example of this is height in humans (Visscher 2008). Three large scale studies mapping QTL for height identified a total of 54 loci using whole-genome approaches. Yet, the proportion of variance explained by the QTL represents a small fraction of estimated genetic variance of height. Whilst there are many possible components of missing genetic variance, a major one is thought to be large numbers of rare loci that have a small effect on the trait (Bodmer and Bonilla 2008). Naturally, identifying such genetic variants is difficult due to their effect sizes, and frequency in standard study populations.

Many of the factors affecting the inability of a given study to identify rare variants are beyond the control of researchers, such as allele frequencies and the genetic effect of the causal variant. Increasing sample size is typically cited as the easiest way of improving our ability to explain higher proportions of genetic variation for a given trait. With the decreasing costs of genotyping this is a reality for many studies, especially in humans. However, often this is impractical, or unfeasible, in livestock mapping studies where we are constrained by the practicalities of the breeding industries. To maximize the use of the information available alternative approaches should be considered when the trait under study is expected to exhibit complex patterns of genetic control.

Results shown in chapters four and five have highlighted the importance of considering how variation in the genetic architecture of causal variants and markers will influence the performance of a model. We demonstrated that, on average, there is a distinct advantage to using haplotype models, such as Hap_all_prob over single-marker tests. Window length is an important criterion to consider, however, with the optimal number of markers depending on, amongst other things, the expected MAF of the QTL (chapters four and five). Localized levels of haplotype diversity can be estimated from observed pairwise LD measures and have a strong influence on model performance. However, it is important to note that no one model will have optimal performance across the entire genome.

Ascites syndrome in broiler chickens is characterized by the accumulation of ascitic fluid in the peritoneum, normally caused by metabolic demands of the growing birds (Wideman 2000). Pulmonary hypertension accounts for the majority of ascites cases in broilers, yet hypertension can originate from numerous causes and affects multi-organ systems (Julian 2000). Our previous mapping study, using a single-marker model, identified a number of QTL markers in individual lines, but these did not replicate between lines. This is possibly due to the divergent nature of the lines, having been separated for numerous generations with different breeding objectives. The complex phenotypic nature of ascites, along with the identification of numerous loci spread across the genome by our previous study and others (Navarro 2003; *et al.* 2005; *et al.* 2006; Rabie *et al.* 2004; 2005; chapter two), suggests that genetic control may differ between different populations. Genetic loci affecting ascites may remain segregating within these populations, but differences in their effects and frequencies caused by genetic parameters such as drift, selection, and founder effects could lead to difficulties in identifying QTL common between populations.

7.1.1 Combined line analysis

In chapter two, lines were combined based on their relatedness to one another, in joint line analyses. This increases sample size and thus power of the tests, although if differences exist in the genetic architecture of causal loci then it can lead to a loss of power as effects from one line can become undetectable with the inclusion of individuals from other lines. For single-marker models, data from related lines are combined into a single dataset and a line effect term is added to the mixed model. Using a similar approach for joint line analyses using haplotype models is made difficult when we consider that haplotype frequencies and probabilities need to be estimated, using statistical algorithms, before the association analysis. Combining line genotype information prior to estimating haplotypes can lead to increased errors in accurately estimating haplotype frequencies and probabilities, due to differences in allele frequencies, deviations from HWE, and LD patterns of local markers between lines (Fallin and Schork 2000; Tishkoff *et al.* 2000). Alternatively lines could be haplotyped individually, and then the information combined in the design matrix. Under this scenario differences in haplotype frequencies and frequency distribution between lines could result in inflated test statistics due to accumulated divergent genetic properties between a set of markers (Liu *et al.* 2004; Sawyer *et al.* 2004). Difference in haplotype properties between lines could be accounted for in a linear model by including a line by haplotype interaction effect; although this has the potential to under power the test through the additional inclusion of a large number of DF. Nevertheless, inclusion of the interaction term would circumvent any problems regarding differences in the direction of haplotype effects between lines.

An alternative to parametric approaches outlined above is to perform a meta-analysis on results from individual lines. Fisher's combined *p* method (Fisher 1925; 1948) is a commonly used approach for combining the results from independent statistical tests that have the same overall null hypothesis. Here, the alternative hypothesis being tested is that a QTL exists in a particular region, but the nature of the association with particular haplotypes differs between populations. The method combines *p*-values from different studies, in this case the individual lines, into a single test statistic that has a chi-squared distribution with 2*k* DF, where *k* is the number of tests combined.

7.1.2 Aims

Here, we re-analyse data from the six broiler lines, using haplotype methods developed and described in chapters' four to six. Ours aims are to attempt to replicate the results seen using a single-marker model, and identify any additional QTL using an alternative method, shown to perform better under a range of genomic conditions. Based on results shown in chapters four to six, the haplotype model described as Hap_all_prob was used, with three-marker windows. Whilst this model will not be optimal across the entire genome, it has been shown to have performed best, relative to other models tested, across a range of genomic conditions. Compared to models such as single-marker analysis, it has a considerable advantage when the MAF of the causal variant is low.

7.2 Materials and methods

7.2.1 Source of data

Data analysed here consists of genotype, phenotype, and pedigree information from six commercial broiler sire lines, described in detail in chapter two and appendix two. The number of sires in each line ranges from 163-189, and for each of these individuals a phenotypic record is supplied determined from progeny adjusted mean values. The indicator trait SaO_2 is measured in progeny of each sire, and these are used to calculate a progeny adjusted mean values for each sire. The formula used to calculate the progeny adjusted means is given in section 2.2.1, with details in appendix one.

In all lines the phenotypes were normally distributed. Sires with phenotypic records outside three standard deviations were removed from analyses; this amounted to the removal of one sire from line 14 and one from 28. Pedigree information included relationships between sires spanning four generations. For each sire genotype information was supplied for ~12,000 SNP markers from a genome-wide panel. Details of the formation of the SNP panel are given in chapter 2 (section 2.2.1). Likewise, information on ranges of progeny numbers for sires in each line is given in table 2.1.

7.2.2 Haplotype analysis

Lines were initially analysed individually, followed by a meta-analysis of multiple lines. A parametric joint line model was investigated that fitted a line by haplotype interaction in the model. This interaction should account for differences in haplotype effects between lines. Markers not segregating within a given line were removed. The number of markers remaining for each line is given in table 2.2. In all analyses markers with minor allele frequencies below 0.01 were removed to avoid false positives caused by spurious associations between rare genotypes and outlying trait values. Details of heritabilities of sire SaO_2 and markers remaining for specific lines are provided in table 2.2.

Each line was analysed using a haplotype model described as Hap_all_prob in previous chapters, using a three-marker sliding window. The haplotype analysis is a two stage procedure. First, haplotypes were inferred in three-marker sliding windows for each line individually. Lines were haplotyped individually rather than pooled to avoid errors in the estimated haplotype frequencies due to differences in the allele frequencies and deviations from Hardy-Weinberg of markers in different lines (Liu *et al.* 2004; Sawyer *et al.* 2004). Second, inferred haplotypes were fitted in a regression-based model to test for their association with SaO_2.

7.2.2.1 Inferring haplotypes

Pedigree information supplied here consists of animal records rather than genotype information. Therefore, the population based, EM algorithm was used to infer probabilities of haplotype pairs for each individual (Excoffier and Slatkin 1995; Long *et al.* 1995). The EM algorithm estimates population haplotype probabilities based on maximum-likelihood given observed genotype frequencies. The algorithm is described in detail in chapter four, section 4.2.5.1. When phase of the marker genotypes are uncertain for an individual, all haplotype pairs, consistent with the observed data are provided, along with their posterior probability. For the genome panel of each line, three-marker overlapping windows were haplotyped, and the probabilities of each pair of haplotypes for each individual were stored. These probabilities are used to form the $\boldsymbol{X}$ matrix for the model, such that probabilities of haplotypes are used as predictor variables. During the EM iteration stages, pairs of haplotypes with probabilities less than $1e^{-9}$ were removed from the analysis and the frequencies of the remaining haplotypes recalculated. This avoids the problem of large numbers of very low probability haplotypes occupying the $\boldsymbol{X}$ matrix and reducing power of the test. The number of haplotypes observed in each window was recorded for each line, and represents diversity of haplotypes.

7.2.2.2 Model fitting

The analysis was conducted using a sliding-window approach, using haplotype information derived in the same manner. For a given window, a matrix of probabilities for haplotypes observed in haplotype pairs of each individual is formed. This matrix constitutes the design matrix of predictor variables, with haplotypes fitted as linear covariates in a mixed model analysis. Sires used in this study were from a commercial breeding population under selection. Evaluation of the pedigree structure showed that sires belonged to a complex pedigree, with a number of small half-sib families. Thus, an average relationship matrix, derived from pedigree information, was included in the model and used in fitting a polygenetic variance component. The following model was used to evaluate the association of haplotypes within a window with SaO_2;

$$Y = 1_n^{'} \mu + Xg + Zu + e$$

Where $\boldsymbol{Y}$ is the n x 1 vector of SaO_2 adjusted progeny means for n sires and μ is the intercept. $\boldsymbol{X}$ represents the matrix of haplotype probabilities and $\boldsymbol{g}$ is a vector of haplotype effects of length N, where N is the number of haplotypes observed in the window. $\boldsymbol{Z}$ is an n x n average relationship matrix of the sires, $\boldsymbol{u}$ a 1 x n vector of random sire polygenetic effects, and $\boldsymbol{e}$ is the 1 x n vector of random residuals. The following expectations and variance were assumed;

$$E\begin{bmatrix} Y \\ u \\ e \end{bmatrix} = \begin{bmatrix} \mu 1 + Xg \\ 0 \\ 0 \end{bmatrix}, \text{ and } V\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} Z\sigma_u^2 & 0 \\ 0 & \sigma_e^2 D^{-1} \end{bmatrix}$$

Where σ_u^2 is the sire variance, σ_e^2 is the residual variance, and ***D*** is an *n* x *n* diagonal matrix of progeny numbers for each sire. Thus, the variance of the residuals is adjusted to account for differences in the numbers of progeny that comprise the adjusted trait values for each sire. Expressed as a function of heritability, the residual variance is given as, $\sigma_e^e = \left(1 - h^2/4\right)\sigma_p^2$, hence, $\sigma_e^2/\sigma_u^2 = \left[(4-h^2)/h^2\right]$, where σ_p^2 is the phenotypic variance and h^2 is the heritability of SaO_2 for a given line. Association is tested in a global analysis against the null hypothesis of $H_0 = \beta_1, \beta_2 ... \beta_N$, where β_i is the effect of haplotype *i*, using an *F*-test with the corresponding *p*-values drawn from tabulated asymptotic *F*-distributions with *N* DF (Weir and Cockerham 1977; Zaykin *et al.* 2002). For each individual line, a genome-wide significance threshold of $p < 0.05$ was determined using a permutation analysis with 1,000 cycles (Churchill and Doerge 1994). Details of the permutation analysis are given in section 2.2.3.

7.2.3 Joint line analysis

7.2.3.1 Meta-analysis

For joint line analyses, *p*-values for haplotype windows are require from all lines. Thus only markers segregating in all lines can be used to infer haplotype windows and test for association. Therefore, for joint line analyses, genotype panels were adjusted to remove markers fixed in any one line, prior to being analysed using the haplotype model described above. Results from individual line analyses were combined using Fisher's combined p method for joint line meta-analysis (Fisher 1925). The p-values for a given window are combined using the following formula; $p_{conb} = \sum_{i=1}^{k} -2\log(p_i)$, where p_i is the p-value for haplotype association in a given window for line i. Under H_0, $p_{comb} \sim \chi^2_{2k}$, thus, the test statistic that has a chi-squared distribution with $2k$ DF, where k is the number of tests combined. For each joint analysis the p-value associated with p_{comb} is calculated for each haplotype window.

As in chapter two, data from line pairs 12 – 28, and 14 – 29 were combined based on the relationships estimated between lines (Andreescu *et al.* 2007; table 2.2). In addition, a whole dataset meta-analysis was conducted using combined results from all six lines. A total of 8834 markers remain segregating commonly in line 12 – 28, whilst 9707 remain for 14 – 29, and 6258 for all lines.

7.3.2.2 Parametric analysis

Haplotype information determined for each line individually was combined in the design matrix rather than estimating haplotypes for the whole set of lines together. Line by haplotype interaction terms were then included in the linear model.

The following model was used;

$$Y = 1_n^{'}\mu + Xg + LHv + Zu + e$$

where ***LH*** is an indicator matrix for line by haplotype interactions and v is a vector of haplotype by line effects. The number of interactions, hence DF fitted, is equal to six times the number of haplotypes fitted in the $\boldsymbol{X}$ matrix. For the joint line analysis $\mathbf{Z}$ is a block diagonal matrix composed of sub matrices of n_k x n_k average relationship matrix of the sires within line k. Sub matrices are inverted independently. The same model expectations and variance as described in section 7.2.2.2 were assumed.

7.3 Results

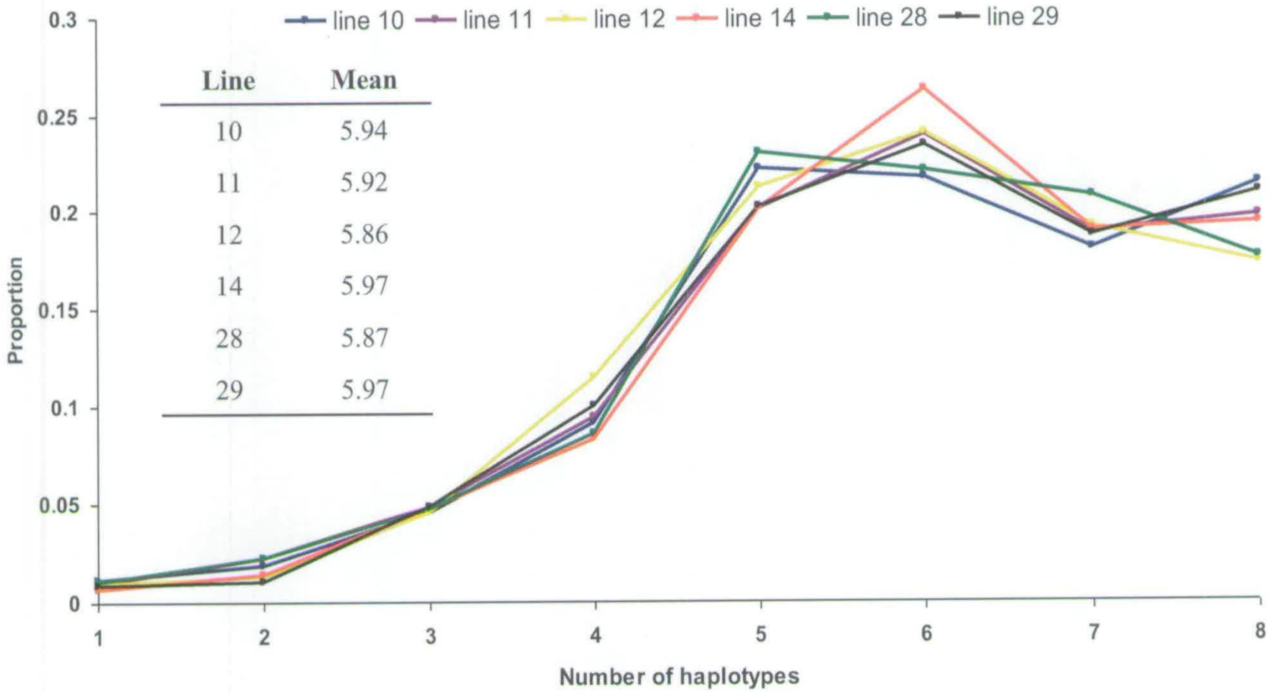
Patterns of LD and distributions of marker allele frequencies are given in detail in chapter two for the six lines analysed here. Results are given below on the distributions of haplotype diversity across the genomes for the six lines, along with summary of results from the genome-wide analyses. Results are also shown for the parametric and meta joint line analyses.

7.3.1 Haplotype diversity

Here, haplotype diversity within each three-marker window is determined by the observed number of haplotypes. It is a product of local LD complexity, allele frequencies of markers, and sample size, and thus, can vary considerably across the genome and between populations. The distributions of haplotype diversity, for the six lines are shown in figure 7.1. The distributions are given as proportions due to differences in the number of markers between lines. All lines show similar distributions of haplotype diversity, particularly amongst proportions of low numbers of haplotypes. Figure 7.1 shows the proportions of haplotypes numbers seen across the entire genome for each of the six lines. However, it does not give any indication of the relationship in diversity seen amongst common markers between lines. Table 7.1 shows the correlation and regression coefficients of observed haplotype diversity between the six lines. Differences in population genetic parameters, selection procedures, and accumulation of recombination events between lines contribute to variation in allele frequency and LD patterns between markers observed in different lines. These factors are expected to influence observed haplotype diversity, leading to the poor relationships between lines as is shown in table 7.1. This highlights the considerable extent of differences in genetic architecture across the genomes of the broiler populations.

7.3.2 Genome-wide association results

Whole-genome association analyses were conducted for each line using the haplotype method described above. *p*-values obtained for each sliding window are shown for each line in figure 7.2. The genome-wide significance threshold ($p < 0.05$) was determined for the analysis of each line individually and represented as dotted lines in figure 7.2. The actual value of the threshold for each line is also provided on figure 7.2. Q-Q plots for the analyses are presented in figure 7.3. These plots indicate that the observed GWAS *p*-values lie close to the expectation with deviations representing associations of haplotype windows with SaO_2 measures. The Q-Q plots suggest that population stratification artefacts had a negligible impact on the results. This is supported by the genomic inflation factors for each analysis which range from $\lambda = 1 - 1.013$ for the six lines analysed. This analysis is designed as a follow up study to the single-marker analysis of the same datasets, described in chapter two.



Line	Mean
10	5.94
11	5.92
12	5.86
14	5.97
28	5.87
29	5.97

Figure 7.1

Distributions of haplotype diversity inferred in three-marker sliding windows for the whole genome. The distributions are provided as a proportion, due to differences in the number of markers between lines. The mean number of haplotypes inferred in each window is also given for each line.

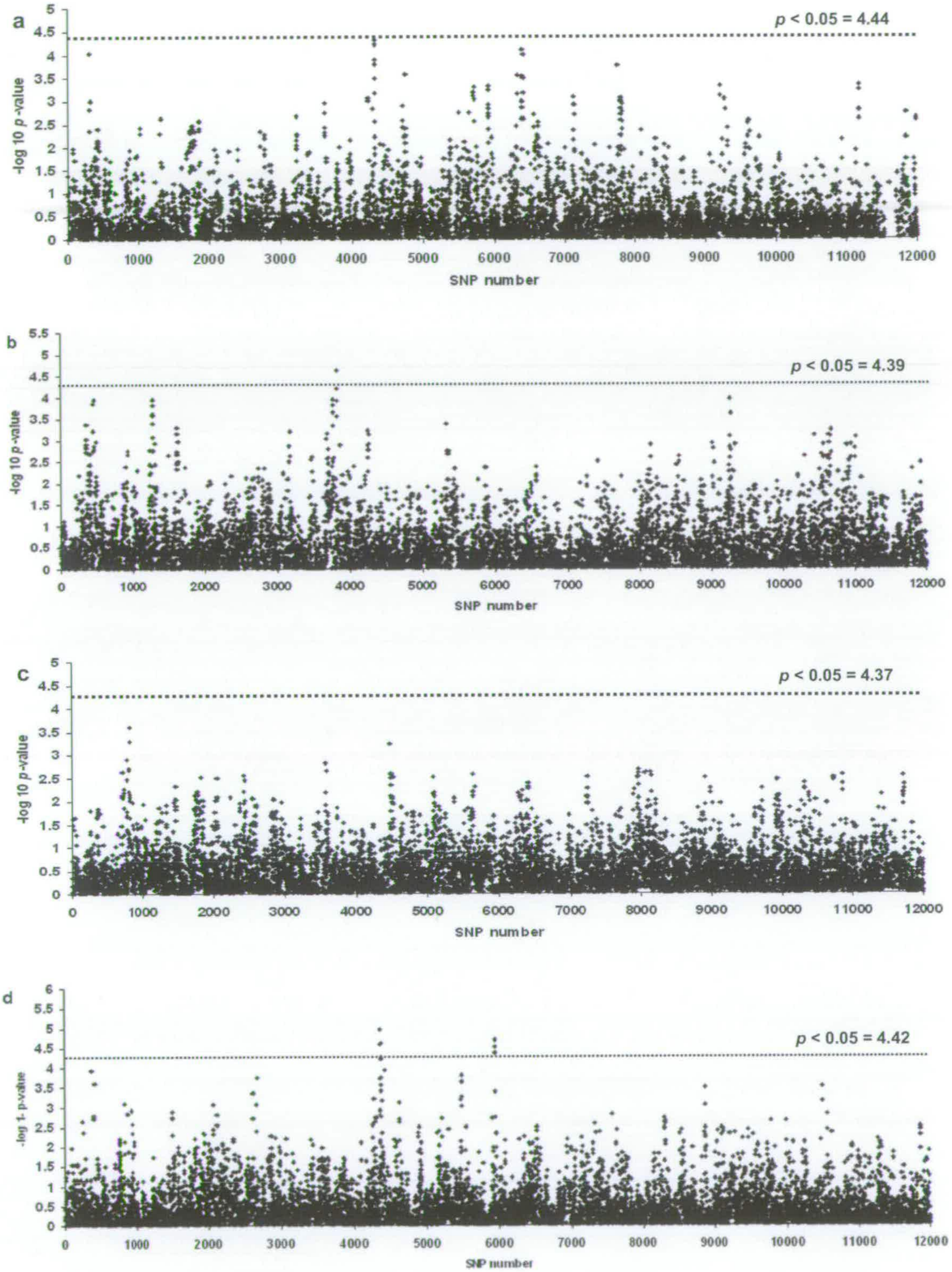
	10	11	12	14	28	29
10	X	$1.2e^{-4}$	$1.4e^{-3}$	$3.2e^{-5}$	$4.6e^{-4}$	$3.2e^{-3}$
11	0.001	X	$6.7e^{-5}$	$4.2e^{-4}$	$3.3e^{-5}$	$7.7e^{-3}$
12	0.007	0.008	X	$5.6e^{-5}$	$1.1e^{-2}$	$8.3e^{-4}$
14	0.009	0.023	0.011	X	$3.5e^{-5}$	$2.3e^{-3}$
28	0.001	0.007	0.025	0.007	X	$5.9e^{-5}$
29	0.003	0.005	0.004	0.003	0.002	X

Table 7.1

Correlation (upper triangle) and regression coefficients (lower triangle) of observed haplotype diversity between pairs of lines. Only markers segregating in the pair of lines were included.

Thus, comparisons with results from the previous analysis are made. The haplotype analysis identified a total of seven QTL regions in lines 11, 14, 28 and 29. Details of windows that show significant associations are given in table 7.2.

For line 11 a single window (start SNP 3818) had an association level greater than the genome-wide threshold. This window is located between 119.7-119.8 Mb on chromosome two, and includes the marker identified in the single-locus analysis, shown in chapter two. In line 14 two QTL regions are identified. The first consists of two windows located on chromosome three (start SNPs 4372 and 4376). These windows are very close together, but do not include any common markers. One of these windows also contains a marker that was identified as significant in the single-marker analysis. The second QTL consists of three significant windows (start SNPS 5946-8), each located next to one another and in total covering 38.8-39.7 Mb on chromosome four. No markers were identified, or approached significance, in this region using the single-marker model. Two QTL regions were also identified in line 28. The first consists of a total of 12 significant windows (start SNPs 369-71, 377-9, 402-4, 406, 408-9), some of which contain markers common to other windows, located between 36.4-37.3 Mb on chromosome one. This region has a high marker density, including a total of 40 markers spaced within 1 Mb. This region contains a clearly defined peak, consisting of additional


a
p < 0.05 = 4.44
-log 10 p-value
SNP number
b
p < 0.05 = 4.39
-log 10 p-value
SNP number
c
p < 0.05 = 4.37
-log 10 p-value
SNP number
d
p < 0.05 = 4.42
-log 1- p-value
SNP number
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000

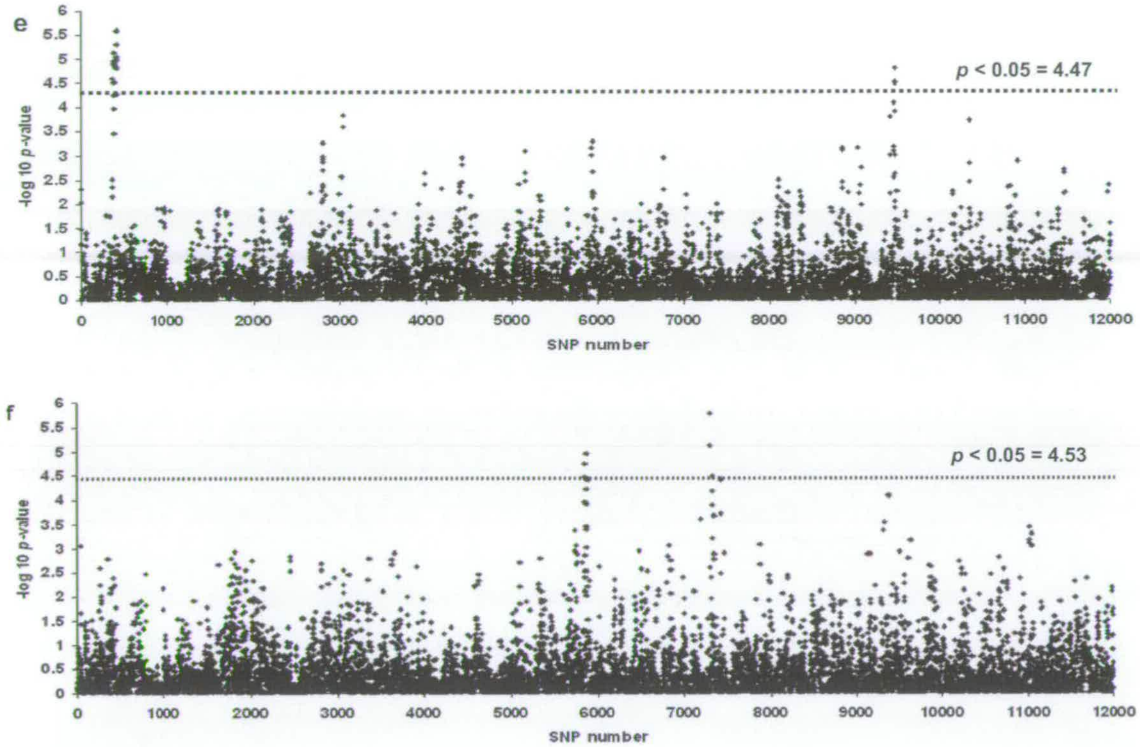


Figure 7.2

-log 10 p-values for the association of each haplotype window with SaO_2 measures, across the whole genome for each line. **a)** line 10, **b)** line 11, **c)** line 12, **d)** line 14, **e)** line 28, **f)** line 29. For each analysis genome-wide significance was determined using 10,000 cycle permutation analysis. $p < 0.05$ thresholds are represented as the dotted line.

Start SNP[1]	Chromosome	Window Distance (Kb)[2]	Start Position (Mb)[3]	-log 10 *p*-value	Number of haplotypes[4]
Line 11					
3818	2	100	119.7	4.62	4
Line 14					
4372	3	159	13.3	4.96	2
4376	3	124	13.7	4.60	5
5946	4	98	38.8	4.47	5
5947	4	96	38.9	4.51	6
5948	4	77	39.0	4.68	6
Line 28					
369	1	81	36.4	5.14	3
370	1	79	36.4	4.95	5
371	1	81	36.5	4.50	4
377	1	74	36.7	4.88	5
378	1	73	36.8	4.91	5
379	1	64	36.8	4.52	7
402	1	47	37.0	4.81	5
403	1	26	37.1	4.87	7
404	1	35	37.1	4.98	4
406	1	59	37.1	5.58	5
408	1	103	37.2	5.05	3
409	1	95	37.2	5.31	6
9467	12	98	6.8	4.52	5
9468	12	168	6.9	4.81	6
9475	12	74	7.3	4.48	5
Line 29					
5869	4	122	24.2	4.95	6
5870	4	85	24.3	4.74	7
7295	5	101	57.9	5.12	5
7296	5	89	58.0	5.78	3

Table 7.2

Summary details of windows identified as significant. [1] represents the first SNP within the three marker window. [2] is the distance that is covered between the first and last marker within the window. [3] is the position of the first marker in the window, from the start of the chromosome. [4] the number of haplotypes inferred and fitted for that window.

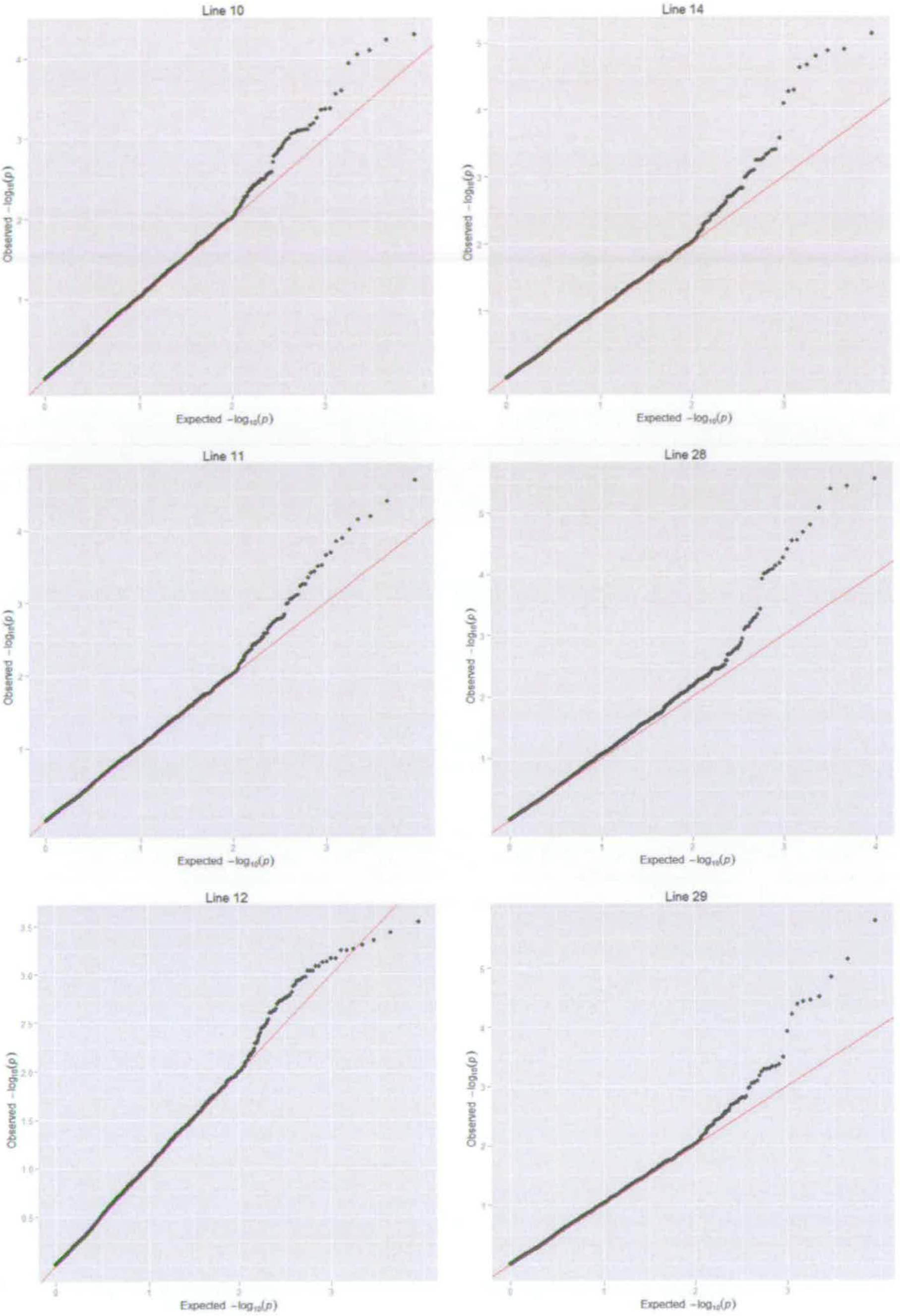


Figure 7.3

Q-Q Plots for analysis of SaO2 using a haplotype model. Expected chi-square value under the global null hypothesis of no association is displayed on the x-axis. Observed chi-square value is displayed on the y-axis. The plots show little evidence of stratification windows that show strong levels of association that do not quite reach the genome-wide significance threshold. The four significant markers identified from the single-marker analysis are included in significant windows from the haplotype analysis. Details of the association and LD patterns for this region are shown in figure 7.4. This figure also shows the results from the single-marker analysis (chapter two) of the same region to act as a comparison. This figure clearly identifies two distinct peaks, separated by a region of low association. Given the close marker spacing, the pattern of peaks could possibly represent associations with a single causal variant, with the association patterns reflecting the complex nature of the relationship between localised genetic information and model performance. Alternatively, this pattern could be due to a single locus with two different causative variants segregating without LD between them. Interpreting the association pattern of haplotype results in terms of the underlying pairwise LD pattern is difficult, as this reflects the relationship between pairs of markers, rather than the complex higher order LD amongst haplotypes. However, it provides an understanding of how differences in performance occur when compared to less flexible approaches, such as single-marker models. The second region identified for this line is located between 6.8-7.3 Mb on chromosome 12 (start SNPs 9467-8 and 9475), and is composed of three significant windows. This region was not identified as significant in any lines when analysed with the single-marker model. A further two QTL regions were identified from the analysis of line 29. One of these is also located on chromosome four (start SNPs 5869-70), between 24.2-24.4 Mb. These windows include the markers identified as significant from the single-marker analysis. There is a possibility that these haplotypes are associated with the same causal variant as the haplotypes in the

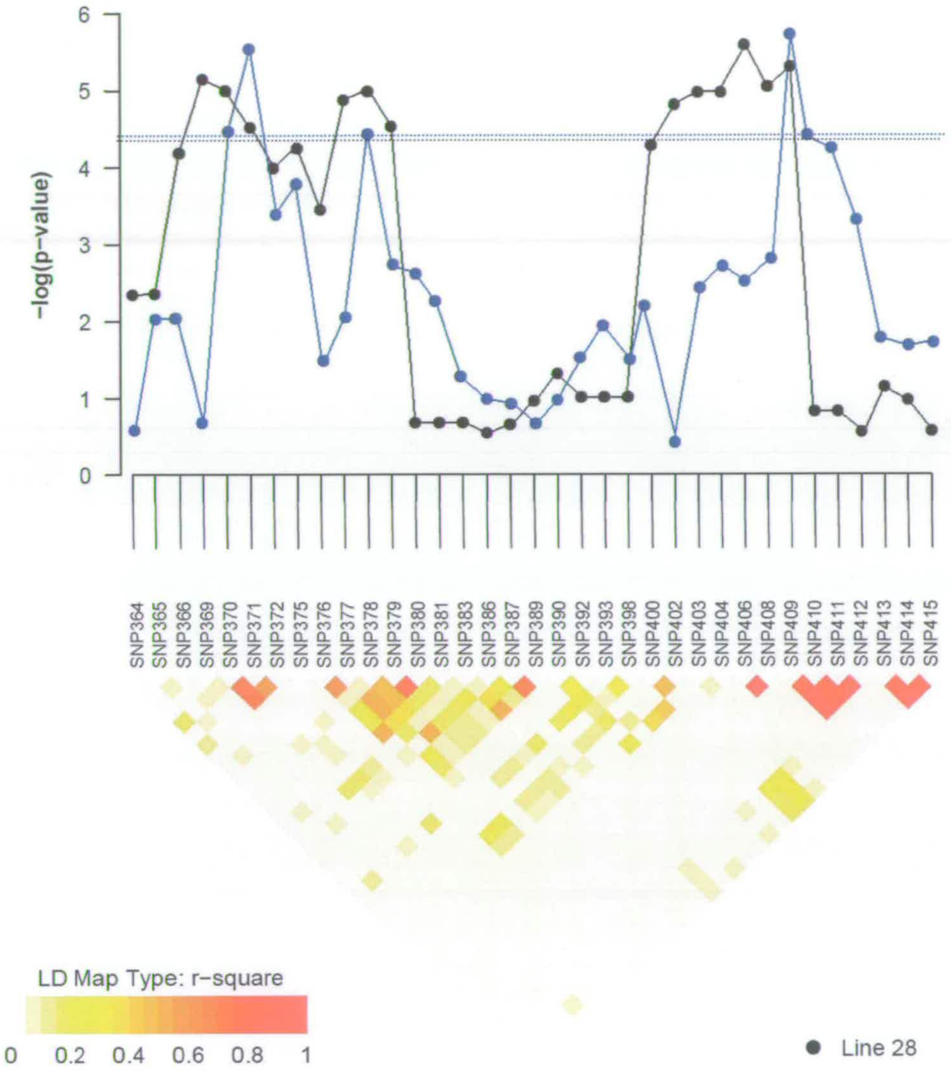


Figure 7.4

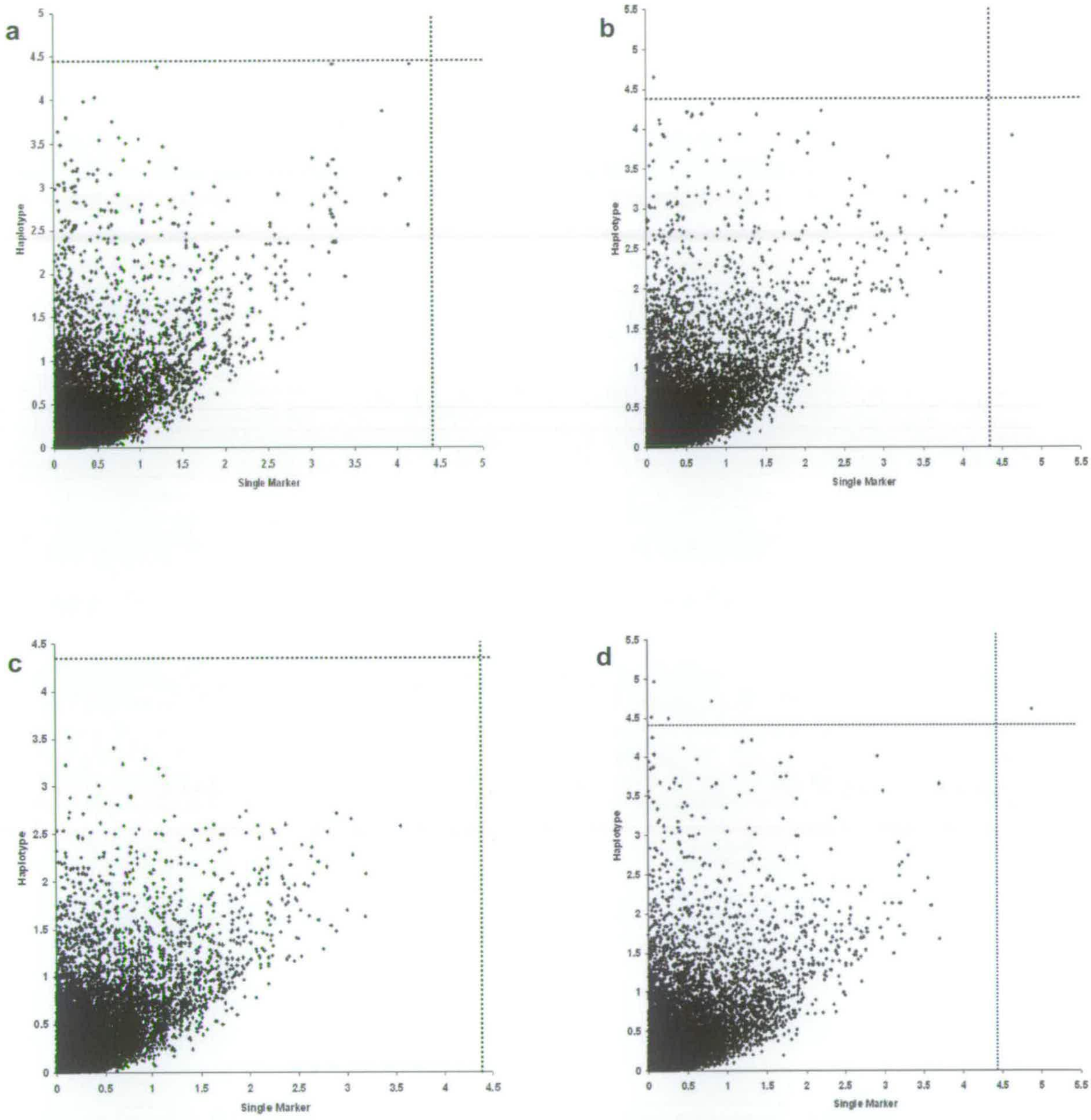
Detail of association and LD patterns for the region surrounding windows with significant associations in line 28. This region is located between 36.1-37.5 Mb on chromosome one. The black dots represent the haplotype windows, as defined by the first marker within that window. The blue dots represent the results from single-marker analysis (chapter two). LD heatmap is composed of r^2 values for pairs of markers. The dotted lines represent the genome-wide significance threshold of $p < 0.05$ of the haplotype and single-marker models. The threshold is determined by a 10,000 cycle permutation analysis.

significant windows identified in line 14, although the regions are separated by 14.6 Mb. This region contains 68 markers segregating within line 14 and 71 within line 29. Line 14 and 29 are the most closely related pairs of lines (chapter two), suggesting the possibility that a QTL is located between the two significant regions. The second QTL regions found for line 29 consists of two adjacent windows (start SNPs 7295-6), comprised of markers covering 57.9-58.1 Mb on chromosome five. No significant markers were identified in this region in the single-locus analysis.

To recap – all QTL regions identified as significant from the single-locus analysis were included in significant haplotype windows in this set of analyses. In addition, three more QTL regions were located on chromosomes four, five and 12. On chromosome four, two regions have been identified by significant windows, one is located at 24.2 Mb and is found only in line 29, whilst the other begins at 38.8 Mb, and found only in line 14. This raises the possibility that both regions are linked to a single causal variant.

7.3.3 Comparison with single-marker analysis

Results from haplotype analyses were compared to those obtained from the single-marker analysis shown in chapter two. Figure 7.5 shows the scatter plots of –log 10 *p*-values from single-marker analysis against those of the haplotype model for each line. Genotype panels remain the same for both analyses, although *p*-values from haplotype models represent information from three adjacent markers. The results are aligned such that the value from the haplotype window corresponds to the single-locus value of the first marker within the window. Genome-wide ($p < 0.05$) thresholds for each analysis are


a
b
c
d
Haplotype
Single Marker
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5

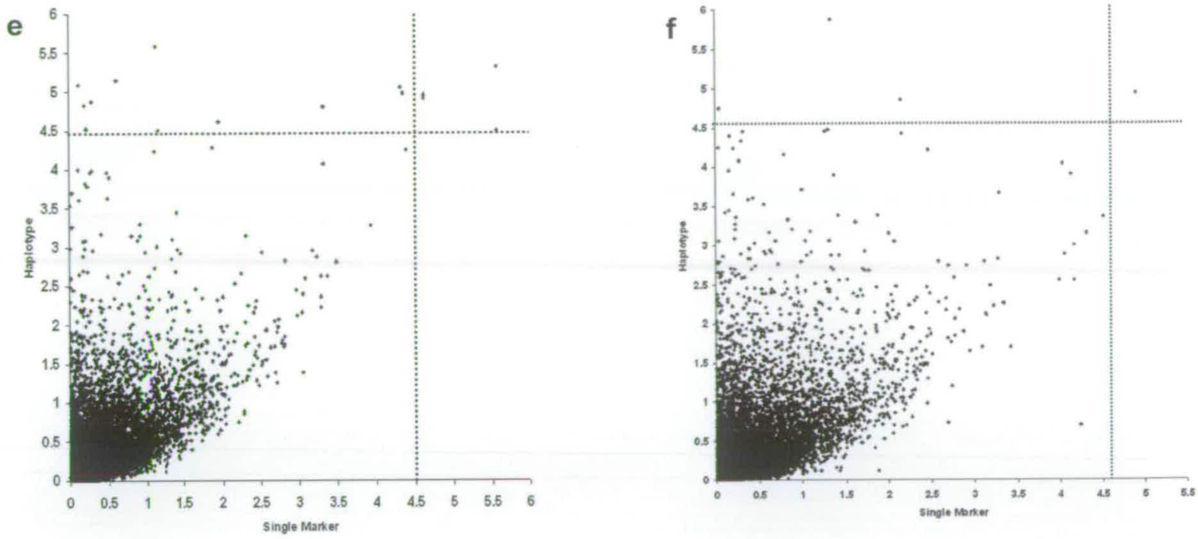


Figure 7.5

Comparison of *p*-values (shown on –log 10 scale) from genome-wide association analyses using single-marker and haplotype models. Results from single-marker tests are shown in chapter two. Results shown for the haplotype analysis are shown so as to represent the position of the first marker within the window. Lines are; **a)** line 10, **b)** line 11, **c)** line 12, **d)** line 14, **e)** line 28, **f)** line 29. Genome-wide ($p < 0.05$) thresholds are provided as dotted lines for each model.

represented by dotted lines on the figure. Whilst this provides an indication of complementary significant markers between analyses, it does not provide a full picture, as a marker that is significant from a single-locus test, may be included in a significant haplotype window at marker position two or three. Thus, on this graph it is possible that a marker can appear significant from the single-locus analysis, and not from the haplotype analysis. Likewise, a haplotype window that has a significant association includes three markers, but is only correlated against one marker from the single-locus model in figure 7.5. In these situations both models can identify the same significant QTL region, but the identified positions may appear to not match with each other as haplotype models are plotted based on their start SNP. This problem would remain if other SNP positions within the haplotype windows were used. Nevertheless, comparison of the two models provides an indication of the relationship between the two models, along with their ability to identify QTL.

In all but one case, a single-marker that is shown as significant is also identified as a significant window from the haplotype analysis. The one situation where this does occur (figure 7.5b) is an artefact of comparing single-locus model results against a haplotype test comprised of three markers, as described above. The marker identified as significant from the single-locus model is in position two of a significant haplotype window. Therefore, any markers identified as significant from a single-locus analysis are also picked up in significant windows by the haplotype approach. The regression and correlation coefficients from the relationship between the two models are shown in table 7.3.

Line	Regression Coefficient	Correlation Coefficient
10	0.603 (0.009)	0.339
11	0.610 (0.011)	0.265
12	0.552 (0.009)	0.262
14	0.522 (0.012)	0.187
28	0.622 (0.008)	0.318
29	0.586 (0.010)	0.235

Table 7.3

Regression and correlation coefficients from the regression of p-values from a single-marker analysis against those of a haplotype model. Standard errors are shown in brackets.

7.3.4 Joint line analysis

7.3.4.1 Meta analysis

Results from the meta-analysis of lines 12 – 28, 14 – 29, and all lines are shown in figure 7.6. For all sets of meta-analyses no marker windows were shown to have a significant association above the genome-wide level ($p < 0.05$). Permutation thresholds are not shown on the figures, as they are all above a –log10 p-value of four. It is possible that the lack of significant results from the meta-analyses is due to markers within the QTL regions of a given line not segregating amongst all lines for the meta-analysis. Thus, QTL regions are not fully represented in the meta-analysis genome panels. To investigate this, we determined whether markers from QTL regions for any line within a meta-analysis were also represented in the genome panels of the common segregating markers. This was done for each set of lines that comprise the three meta-analyses, and summarised in table 7.4. For the meta-analyses composed of the pairs of lines, markers from all QTL regions were represented in the common genome panel, and all but one region was represented for the all line analysis.

7.3.4.2 Parametric analysis

Results are shown from the parametric analysis of all six lines in figure 7.7. No haplotype windows showed genome-wide significant association. These poor results possibly reflect the decrease in statistical power caused by the large number of additional DF included in the haplotype by line interaction term. This term corresponded to between 18-48 DF depending on the number of inferred haplotypes within a window. A model fitting just line effects was also investigated, although this produced a large number of highly significant false positives caused by extreme differences in haplotype frequencies between lines.

7.4 Discussion

Here we report results from genome-wide analysis, for loci affecting ascites syndrome, of six lines of a commercial broiler population using a haplotype method implemented in a three-marker sliding window. This study consisted of the re-analysis of data presented in chapter two, using a choice of model based on the results shown in chapters four – six. Genomic conditions that influence model performance include those that are observed, such as LD patterns and marker allele frequencies, as well as those that are unobserved, such as QTL MAF and proportion of genetic variance explained by the causal variant. Identification of numerous QTL regions using single-marker (chapter two) and haplotype models individual to a given population, suggest a complex pattern of genetic control for ascites, with effects differing amongst populations. Given the unknown properties of the QTL frequencies and surrounding LD patters, using a single model for the genome-wide analysis will always involve a compromise. The Hap_all_prob model, implemented using a three-marker window, provides the best choice in terms of overall performance across a range of genomic conditions (chapter five). Out of the models investigated in chapters four and five, Hap_all_prob_3 was shown to have performed best on average when the MAF of the causal variant is low,

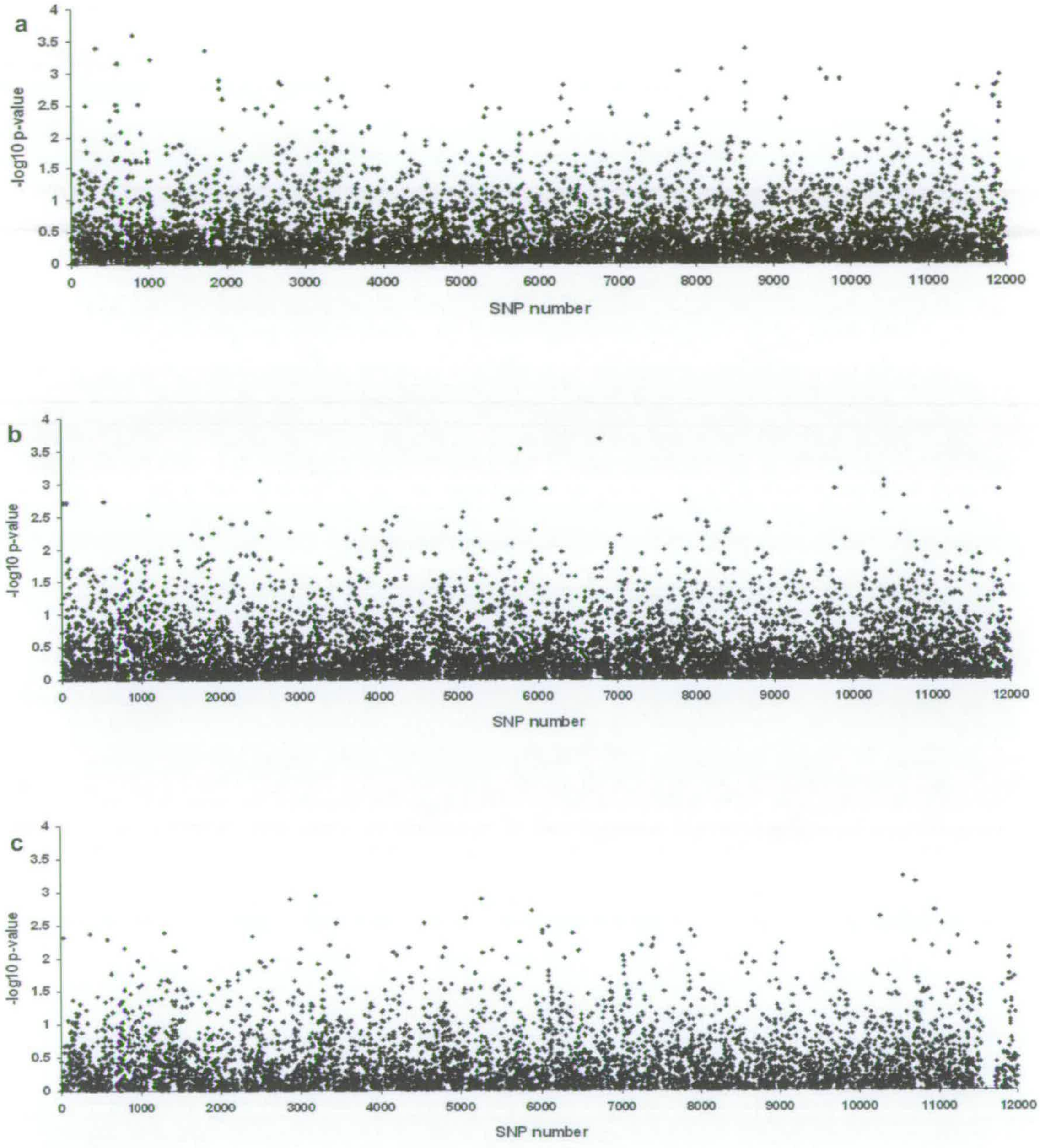


Figure 7.6

-log 10 *p*-values from fisher's combined *p* method for joint line meta-analysis. Markers segregating in all combined lines are analysed for each line individually using the haplotype method described in section 7.2.2, and *p*-values from each window are combined for sets of lines. **a)** lines 12 and 28, **b)** lines 14 and 29, **c)** all six lines.

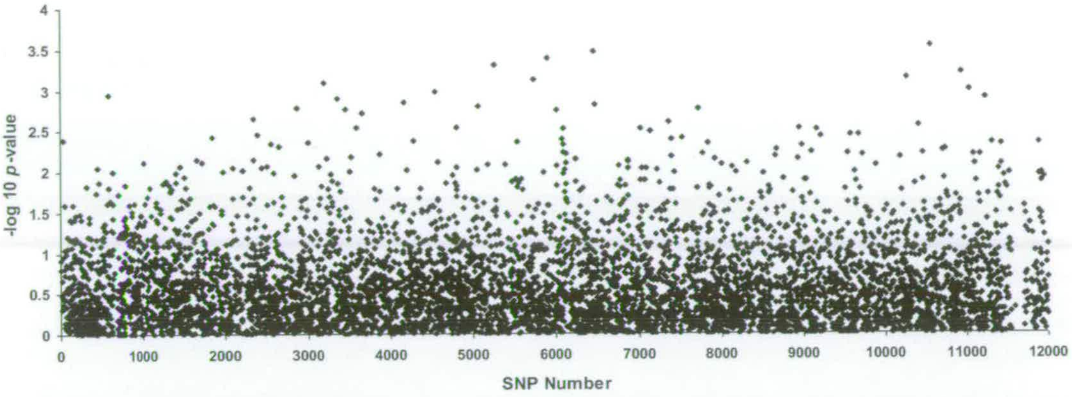


Figure 7.7

-log 10 *p*-values for the association of each haplotype window with SaO_2 measures, across the whole genome using parametric joint line analyses. Joint line analysis of all six lines with a line by haplotype term fitted.

QTL Region	Segregating in Line 12 and 28	Segregating in Line 14 and 29	Segregating in All Lines
Line 11			
Ch 2. 119.7-119.8 Mb	--	--	Yes
Line 14			
Ch 3. 13.3-13.8 Mb	--	Yes	Yes
Ch 4. 38.8-39.1 Mb	--	Yes	Yes
Line 28			
Ch 1. 36.4-37.3 Mb	Yes	--	Yes
Ch 12 6.8-7.4 Mb	Yes	--	Yes
Line 29			
Ch 4 24.2-24.4 Mb	--	Yes	Yes
Ch 5 57.9-58.1 Mb	--	Yes	No

Table 7.4

Summarization of markers from individual line QTL regions represented in meta-analysis tests. Meta-analysis tests only use markers that are segregating in all lines represented within the meta-analysis. "Yes" signifies that markers identified in QTL region from an individual line analysis are also included in the respective meta-analysis panel, whilst "No" means they were removed due to markers not segregating in all lines. -- means the QTL peaks are not applicable for that meta-analysis.

and almost as well as Hap_all_prob_5 for intermediate to high MAF. The Hap_all_prob_3 model has the lowest rate of false positives amongst models with marker windows of comparable length when the causal variant has a low heritability (chapter six). Compared to the single-locus test used in chapter two, Hap_all_prob models perform better on average across the range of QTL MAF, as well as having a lower rate of false positives when the heritability of causal variants are low (figure 6.2).

Between all lines, a total of seven QTL regions were identified by marker windows with association levels above genome-wide significance ($p < 0.05$) levels, although, each QTL was identified in a single line only. Two QTL regions on chromosome four, separated by 14.6 Mb, were identified in lines 14 and 29. Details of the LD patterns for markers in the region between the two QTL positions are shown in figure 7.7. The position of these significant windows raises the possibility that a single QTL, common to both lines, is being detected by different sets of markers in the two lines, particularly when we consider that the two closest related lines are 14 and 29 (Andreescu *et al.* 2007; table 2.2). However, the smallest distance a causal locus could be from identified QTL regions is 7.3 Mb, at which distances mean pairwise LD is expected to be very low, reaching almost background levels. Whilst high levels of LD are occasionally observed between markers separated by long distances, here, we are concerned with the association of haplotype structures at marker loci with the causal variant, which is characterised by complex high order LD patterns, not readily observed in pairwise comparisons (Schaid 2004; Zhao *et al.* 2007). Haplotypes that have a close ancestry with a causal mutation typically occur in close vicinity, as haplotype associations decay rapidly with the accumulation of recombination events (Akey *et al.* 2001).

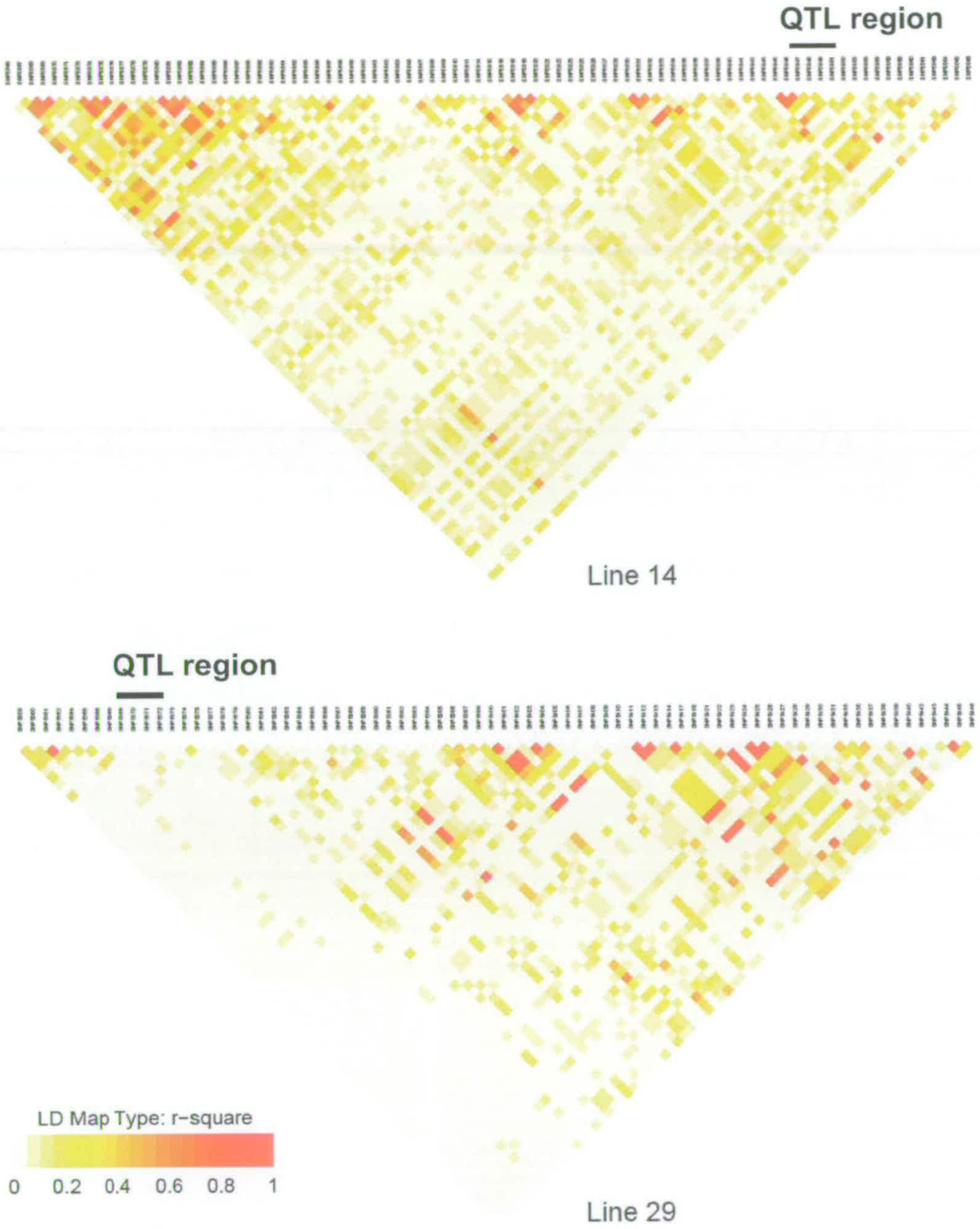


Figure 7.8

Patterns of LD for markers between the two QTL regions identified as significant in lines 14 and 29. Only markers segregating within a given line are shown. The QTL regions for each line are indicated with a black line.

At the time of mutation, the haplotype background of the causal variant encompasses the entire chromosome. Over generations, this association is broken up by accumulation of recombination and mutation events, reducing the distance that the haplotype background extends. Therefore, the haplotype background length will be a function of the recombination rate for that chromosome and the number of generations that have occurred since mutation (Hudson 1983; Sved 1971) suggesting that strong associations between marker haplotypes and the causal mutation are unlikely to occur at the distances observed here, unless the mutation was very recent. This conclusion is supported when we consider differences observed between the two lines in the patterns of LD for the markers encompassing the QTL regions (figure 7.8).

7.4.1 Comparison with single-marker analysis

The seven QTL regions reported here included those identified by the single-marker analysis (chapter two). Thus, the haplotype model identified the same QTL regions as the single-marker model, as well as an additional three regions. The total number of significant haplotype windows, across all lines, was 25, against a total of eight markers from the individual line analysis using single-marker model. This difference in the number of significant results possibly reflects the inclusion of significant markers in multiple adjacent windows. Under this situation, a marker in high LD with a causal variant could lead to a single significant result when analysed using a single-locus approach, although when analysed using overlapping windows could result in multiple significant results. Although, it is worth noting that this situation does not apply in the case of the three new regions identified by the haplotype model here. However, it is also likely that haplotype models are able to utilize information contained between a set of markers that is not available to single-locus models (Akey *et al.* 2001; Clayton *et al.* 2004), leading to a greater number of significant results.

Figure 7.5 shows the comparison of p-values (shown on the –log10 scale) from genome-wide association analyses using the haplotype and single-marker (chapter two) models. This figure shows a relationship whereby low –log10 p-values from single-marker analyses can be represented by high values from haplotype analysis, although low –log 10 p-values from haplotype model do not reach high values with single-marker models. The regression coefficients of these comparisons range between 0.522 – 0.622, and the correlation coefficients between 0.187 – 0.339 (table 7.3). A possible explanation for this pattern is that haplotype models result in a higher rate of false positive associations, leading to an increase in the –log 10 p-values. However, as we have shown in chapter six, when the heritability of a causal variant is low, as is expected here, the rate of false positive association caused by haplotype models are considerably lower than the rate observed for single-locus tests (figure 6.2). The alternative is that haplotype models are better at providing statistical support, given the conditions of sample size, LD patterns and marker density in this dataset, with the result, that they are able to replicate the findings of single-locus models, as well as identify additional QTL regions.

7.4.2 Joint line analysis

To investigate the identification of common effects amongst lines, a meta-analysis using Fisher's combined *p* method (Fisher 1925) was conducted for two pairs of lines as well as for the set of six lines. Additionally, a parametric approach was investigated that fitted a line by haplotype interaction term. Unfortunately this produced poor results, with no windows approaching genome-wide significance, possibly caused by the large number of denominator DF fitted in the model. Results for the meta-analyses (figure 7.6) reveal no marker windows approaching the genome-wide significance level in any of the combinations of lines. The two pairs of lines were chosen for their relationship to one another, based on the allele frequencies of common markers (Andreescu *et al.* 2007; table 2.2). Whilst these pairs were the closest related pairs amongst the possible combinations of lines, they still represented populations separated from one another for numerous generations in individual breeding programs. It is likely that causal variants influencing ascites arose in populations prior to the separation of breeding lines. Thus, unless these variants have become fixed within a given line, they are expected to be segregating within all lines. A possible explanation for the lack of significant results from combined line meta-analyses, is that population genetic parameters such as drift, selection, and line specific founder events, have led to markedly different genetic architecture between lines. This is reflected by the comparison of allele frequencies from markers common between lines (Andreescu *et al.* 2007; table 2.2). Combining the populations of line 14 and 29 for a joint line analysis with the single-marker model revealed two QTL regions on chromosome four (table 2.4; figure 2.6). The inability to identify these QTL using a haplotype meta-analysis of the same populations is possibly due to the complex nature of haplotype divergence between lines, caused by the accumulation of divergent genetic properties between the set of markers. This relationship is expressed in the correlation coefficients of haplotype diversity between pairs of lines (table 7.1).

7.4.3 Implications and further study

The identification of a number of QTL regions, spread across different chromosomes, and located in individual lines, suggests a complex pattern of genetic control for ascites, with effects differing amongst populations. These results suggest that, in broilers in general, ascites is affected by large numbers of causal loci. Given the sample size, and the relatively low density genotype panel used here, it is likely that there are large numbers of causal variants with small effects on susceptibility to ascites that are undetectable using in the current datasets. The continued presence of ascites in commercial populations that are controlled by complex breeding programmes suggests an interaction or linkage with loci that affect production traits. Within the poultry industry, the parents of the next generation are selected using breeding values estimated from mixed model analysis of phenotypic records, along with pedigree information. The fact that selection programs have been able to sustain rapid genetic progress for growth and feed efficiency during the past decades suggests that the traits under selection are also affected by many genes (Havenstein *et al.* 2003; McKay *et al.* 2000; Pakdel *et al.* 2005).

Genomic architecture for ascites appears complex, with the heterogeneous nature indicating control by a number of loci. The complexity is confounded by line differences and the sample sizes available in standard commercial breeding programmes. To further investigate the genetic control of ascites, improvements in tools, such as density of markers in genome panels, and increases in sample sizes are required. The likely occurrence of a number of genetic loci, combined with the expected small effects, will limit the potential of approaches such as MAS to combat incidences of ascites (Dekkers 2004). Alternative techniques such as whole-genome approaches may offer a potential solution to further understanding the genetic control of ascites (Goddard and Hayes 2009).

7.5 Conclusions

Here we identified seven QTL regions across six chromosomes that are associated with susceptibility to ascites, through the analysis of six lines of broiler chickens using a haplotype-based sliding window approach. This study comprised of a re-analysis of data previously investigated using a single-marker model that located four QTL regions between the six lines. The choice of this haplotype model was made based on results shown in chapters four – six, showing it to the best performance on average across a range of genomic conditions compared to other models investigated. In this chapter, using the haplotype model, the four QTL from the single-marker analysis were identified along with an additional three. These results support evidence from previous chapters of using this haplotype model for an improved ability to provide statistical support for QTL.

Results combined from individual lines in a meta-analysis yielded poor results, with no haplotype windows identified as significant at the genome-wide association level of $p < 0.05$. This possibly reflects the divergence amongst lines separated within a commercial breeding programme, and the effects of population genetic parameters, such as drift, selection and line specific founder effects, on haplotype associations.

CHAPTER EIGHT

GENERAL DISCUSSION

This chapter features a summary and evaluation of the main contributions to this thesis, along with some perspectives on future research directions regarding utilisation of observed genomic information for model choice in genome-wide association mapping studies. Issues concerning model parameterisation for association mapping models are addressed, and finally, the implications of the findings are discussed in relation to association mapping for complex traits in livestock species and humans.

8.1 Summary

This thesis has two main research components whose objectives interact with one another to provide an overview of the understanding of how statistical tools can be optimally applied to identify genetic loci affecting complex traits. The thesis began with the application of standard statistical tools to locate QTL influencing ascites susceptibility in broiler chickens, and addressed the need to understand how association model performance needs to be considered in terms of the localised genomic information that they use. Lessons learnt from a comprehensive evaluation of the interaction between model parameterization and localised marker and QTL information were applied in the re-analysis of broiler data to optimise the use of information provided across the genome. An overall aim was to be able to improve our ability to identify and explain genetic variation for complex traits by considering association mapping strategies that use observed marker information to choose optimal model parameterisation in a given localised situation. The following paragraphs outline and summarise some of the main conclusions from chapters two – seven.

The four QTL regions on different chromosomes shown in chapter two were previously unidentified in other mapping studies looking at ascites, or ascites related traits in broilers. However, the lack of consensus results between the six lines, and with other studies, suggested that genetic control for ascites susceptibility is likely complex, and spread throughout the genome. The identification of the QTL regions are in themselves a pleasing result, although it highlights questions regarding the inability it identify loci that explain the remainder of genetic variation as well as the lack statistical support for regions between lines. If we expect loci affecting ascites to be segregating in multiple lines, then the immediate assumption is that the single-marker regression model is unable to handle the differences in the localised genetic architecture that exists between lines. Likewise, for complex traits, with a heterogeneous genetic control, loci are expected to be located throughout the genome, existing across a broad range genetic architecture. To maximise the ability to provide statistical support for these loci both within, and between, populations, considerations need to be made regarding variation in genomic conditions across the genome, as well as how this information is utilization by different association models.

When measuring a single genetic marker, it is well known that the factors that influence power are size of the effect of the causative locus, the frequency of the causal allele, the extent of LD between the QTL and the marker and how close the allele frequencies match between the causal locus and marker (Zondervan and Cardon 2004).

When measuring multiple marker loci, the strength of LD among the marker loci will additionally influence power. Although the benefit of haplotype analyses versus multi-marker tests for association that ignore phase have been widely discussed and debated in the literature (Bardel *et al.* 2006; Chapman *et al.* 2003; Clayton *et al.* 2004; Humphreys and Iles 2005; North *et al.* 2006), it appears that the greatest gain in power provided by haplotype analyses occurs when linkage disequilibria exist among the marker loci at orders higher than pair-wise LD (Nielsen *et al.* 2004).

Within a flexible, regression based framework, considering different uses of marker information is easy through extension of the $\boldsymbol{X}$ matrix to include information from multiple markers. It also provides some flexibility in model parameterization, as a choice can be made regarding which main and interaction effects should be fitted in the model. Information contained between a set of adjacent markers is broken down into main and interaction effects. These effects can be considered as a spectrum of parameters of increasing complexity (Clayton *et al.* 2004), such that effects can be continually added to a model until all possible parameters contained between the set of markers are included. Likewise, the number of SNPs that are considered jointly in models is also an important choice. As has been shown in chapter five, this is especially the case for haplotype models due to the rapid increase in the potential number of parameters as additional markers are included. In the chapters exploring the use of multilocus models, this choice was constrained to the two extremes of parameterization, the 'locus scoring model' which includes a main effect for each locus, and the 'haplotype scoring model' which models an effect for each marker haplotype. Locus-scoring methods are often considered appealing because they do not require haplotype phase resolution.

The concerns regarding the implementation and strategy for fitting multilocus, specifically haplotype, models have been discussed throughout this thesis. When the choice of model moves beyond using a single-marker test, a number of additional decisions present themselves in relation to how multilocus information should be utilized within a genome-wide mapping study. Most of these issues are well discussed in the literature, and have been addressed and investigated as they have arisen through the course of the research covered in the previous chapters. The first main choice when using multilocus model is whether to use a block design or a sliding window approach (Gabriel *et al.* 2002; Li *et al.* 2007). Overlapping sliding windows were applied here based on their ease of implementation, lack of requirement in defining blocks, and their flexibility in terms of marker length and parameterization of the regression models. An often cited problem when using haplotype models is how to handle rare haplotypes. A large number of approaches have been proposed that either, pool, cluster or remove the effects of rare haplotypes from the hypothesis test with the aim of reducing the number of parameters and avoiding errors in estimating their effects. Many of these approaches were investigated during the course of the research; although they all rely on the concept that rare haplotypes are unlikely to be associated with a causal variant. Given that a major strength of haplotype models is the rare allele from causal variants with low MAF which are likely to exist in a single haplotype background, discarding effects of rare haplotypes is expected to reduce the ability to identify associations with rare alleles.

Thus, instead of ignoring rare haplotypes, a hypothesis test that was more robust to low frequency parameters was used rather than the score test investigated in chapter three.

In recent years there have been a number of theoretical and empirical studies that identify a particular set of genomic conditions that leads to the optimal performance of a given statistical model (Akey *et al.* 2001; Calus *et al.* 2008; 2009; Chapman *et al.* 2003; Clayton *et al.* 2004; De Bakker *et al.* 2005; Grapes *et al.* 2004; 2006; Guo and Lin 2009; Hayes *et al.* 2007; Morris and Kaplan 2002; North *et al.* 2006; Pe'er *et al.* 2006; Schaid 2004). These studies provide a huge range of information regarding model choice for a given set of genetic and trait architecture, although identifying these conditions in traditional mapping studies is normally extremely difficult, or impossible. None of these studies investigated how the performance of different models is affected by the range of local genomic architecture of the markers and causal variants.

The comprehensive set of evaluations presented in chapters four to six, provided strong evidence of the impact of localised genomic architecture on the performance of regression-based models that differed in their use of marker information. The ability to provide statistical support for the presence of the sQTL for all models was influenced by genetic conditions such as between marker LD, LD between markers and the causal loci, and the allele frequencies of the sQTL. These factors affect the performance of models relative to one another due to differences in a model's ability to use the range of complexities seen in genomic data. Turning this around, we can consider that a model's parameterisation is best suited to certain types, or complexities, of information, and that there is a reduction in its performance when presented with different forms of genetic information. In this sense, a model would be termed 'robust' if its performance was constant across the range of genomic conditions.

When using multilocus methods for genome-wide association studies the number of markers included within a window affects performance. However, the effect on performance differs between models due to the influence of marker number on parameterization space. For main effect models this relationship is simple, as the effect of each marker is only described using a single parameter. For haplotype models the parameters include between marker interaction terms, and so, each additional marker can raise the number of fitted effects by a power. Unfortunately, this situation is not so straight forward, as many of the potential allelic combinations are not observed in practice, meaning the actual number of fitted parameters is typically far fewer than the theoretical maximum. In chapter five the optimal numbers of markers to fit in a window was shown to be specific to each model, as well as being influenced by the localised genetic conditions.

The results from chapters four and five show that, relative to alternative models, Hap_all_prob with a three-marker window, has the best performance when mapping for causal variants with low to intermediate MAF. When the MAF is higher than that, using the same model with a five-marker window will provide the best performance on average. These results are presented in terms of relative performance between models, although all models do better when the MAF of the causal variant is high. The role of LD between markers was also shown to affect model performance. As is expected, when high LD exists between markers and QTL using a simple model such as the single-marker regression is best. In situations of low level LD between a set of markers as well as situations of low LD between markers and the sQTL, haplotype models perform best relative to other models, although window length has a strong influence on performance under these conditions. Such examples highlight the considerations needed to identify how models use the information they are provided in different ways. We also need to consider the difference between model performance relative to one another, as well as the optimal conditions for a given model. Often the optimal conditions for models are similar to one another, for example, high LD between markers and the causal variant and high QTL MAF. In a genome-wide association study, the genomic landscape seen in the genome panel for a given population are fixed within that dataset. Thus, we must think in terms of the optimal performance of models relative to one another, and consider that the optimal model will change along with the variation in the underlying genomic architecture.

In most population based association mapping studies the absence of family genotype data means statistical procedures are required to determine marker phase information that provide probabilities for the possible haplotype consistence with the observed information. A number of studies using haplotypes have accounted for the phase uncertainty by accepting the highest probability haplotype pair from each individual as though it were known (Grapes *et al.* 2004; 2006; Hayes *et al.* 2007; Zhao *et al.* 2007). The justification for using this approach is usually to reduce the potential number of fitted parameters. As has been shown in this thesis, on average fitting a model that accounts for the uncertainty in phase by modelling haplotype probabilities in the X matrix, leads to better performance. Other studies have shown that ignoring this uncertainty can increase errors associated with estimating haplotype effects (Morris *et* *al.* 2004; Tanck *et al.* 2003). Therefore, it is strongly recommended that methods such as Hap_all_prob be used over Hap_highest_prob for future genome-wide association studies.

Many large scale mapping studies, particularly in humans, have struggled to identify significant markers that are able to explain the estimated genetic variance for a given trait, using single-locus models (Bodmer and Bonilla 2008; Frazer *et al.* 2009; Visscher 2008). The occurrence of large numbers of rare loci affecting these traits is often cited as an explanation for a proportion of this missing heritability (Maher 2008). Although it will not be optimal across all genomic situations, applying a model such as Hap_all_prob with three-marker windows is expected to improve power, particularly in detecting rare variants. This method has also been shown to have a lower false positive rate than single-marker models when the heritability of the causal variant is low (chapter six).

For genome-wide mapping studies in general, there is a need to consider that the genetic architecture that influences model performance is highly variable across the genome and between different complex traits. This suggests that associations are being missed through a failure to use the optimal model for a given set of markers when only a single approach is used for genome-wide mapping. No one method will be ideal across the genome and so using observable information as a predictor of model performance may provide a useful strategy to make maximum use of the information available in a given dataset. In previous chapters we have seen the use of marker information in predicting haplotype diversity and providing support for model choices. Clearly more work is required in order to make full use of observed or inferred information, and some of these considerations are discussed below.

8.2 Use of observed genomic information

As has been demonstrated in previous chapters, genetic information used by association models is highly variable across the genome, and influences the performance of the regression-based models. Based on the results shown in chapter three to six, the choice of model used to identify causal variants associated with ascites was revised from a single-marker to Hap_all_prob_3 model. Whilst this improved the performance of the mapping study, applying just a single model is not expected to be an optimal mapping strategy. We can demonstrate that the Hap_all_prob_3 method is expected to be best on average, although situations will exist throughout the genome where alternative models will have better performance. This is particularly the case when we expect the trait to be affected by large numbers of loci. With this premise, considering a flexible strategy, that fits the optimal model given the local genetic architecture, will improve power of the association study.

In many ways applying a flexible model approach is one of the ultimate goals in genome-wide association studies. There are two main components that need to be understood before this can be implemented. The first is to determine the model that has optimal performance for a given set of genomic architecture, such as marker LD, QTL MAF, and LD between markers and QTL. This has been one of the main aims of this thesis, and we now have a comprehensive understanding of the conditions that lead to optimal model performance. The second, and more difficult component, is to be able to correctly identify the underlying set of genetic architecture using observed or inferred information. Such a strategy would require assumptions regarding the position of causal loci, although these assumptions are routinely used in traditional association mapping studies. The complexity lies in the variable nature of the relationship between genetic parameters between markers such as LD and allele frequencies. This makes it difficult to use observed marker information to accurately predict the state of the local genetic parameters of loci that are not genotyped. Thus, incorporating the mathematical relationships to the genetic parameters may allow a flexible approach to be modeled using the available observed information as well as inferred information for the unknown parameters.

As was shown in chapter five (figures 5.9 - 5.11) the optimal performance of a given model occupies a clearly defined genetic parameter space. The relationship between the mean LD measure of a set of eight markers and the markers and the sQTL can be considered as a landscape of parameters over which certain models perform better than others. The hexagon plot given in figure 5.9 shows that models occupy a visibly distinct position within this landscape based on their optimal performance. Genetic factors other than marker and QTL LD patterns affect model performance, as is easily seen when the hexagon plots are shown for sQTL with high and low MAF (figures 5.10-5.11). Likewise, chapter five showed that a relationship exists between model performance and haplotypes observed, and is specific to the length of the marker window used. In this situation, haplotype diversity can be estimated using local levels of LD between markers, and that information used in the decision process for the length of window to use. Unfortunately information such as QTL allele frequencies and LD between markers and QTL needs to be inferred, or alternatively, assumptions of the likely distributions of parameters are required. In this thesis we have identified how the landscapes for genetic parameters are populated by the optimal performance of models. In other words, if the localized genetic architecture is known then the ideal model to use in that situation can be identified. It has also been shown that in some cases, observed information can be used as a predictor for the underlying genetic parameters, although variation still exists within these relationships. Clearly, identifying a useful predictor for model choice will need to incorporate numerous sources of information. Future research building on the work presented here should aim at identifying how observable and inferred genetic parameters interact to produce a meaningful indicator of the optimal model parameterization for a given localized situation. Incorporating this information into a flexible association mapping strategy would undoubtedly improve power over current single model approaches, and make maximum use of genomic information within a dataset.

8.3 Mapping for complex traits

Identifying and understanding the influence of genetic variants on complex traits and disease susceptibility is considered the main goal in contemporary genetic research. Quantitative genetics has proven to be a useful approach to locating QTL and explaining the genetic component of phenotypic variation in both livestock species and humans. However, whilst there are numerous success stories, many studies have struggled to identify the genetic variants required to explain large proportions of genetic variation – the so-called 'missing heritability' problem (Maher 2008). In this thesis association mapping approaches have been applied to identify QTL for ascites susceptibility in broiler chickens, although the conclusions ought to be considered in a wider context regarding mapping for complex traits in other livestock species and humans.

Typically, no information is available about the underlying complexity of the relationship between the alleles of the causal variant and the phenotype, as can be detected by linked markers. This complexity is characterised by the parameterisation of models, which include effects normally termed main and interaction effects. Interaction effects build in their complexity as they include all combinations amongst a set of markers (Clayton *et al.* 2004). The haplotype models described in previous chapters are parameterized to include all main and interaction effects between a set of markers, although in reality, it is rare to observe all possible combinations, unless the sample size is very large. As their name suggests, main effect models fit just the main effects of markers. Therefore, these two groups of models represent the two extremes of multilocus parameterization. For a given pattern of LD and marker density, the ability to provide statistical support for a causal variant will depend on the nature of the effects as is seen in the associations with markers (North *et al.* 2006). It is often the case that fitting all available information between a set of markers is not always optimal given considerations of model parameterization.

Taking the main and haplotype models to be the extremes of the parameterization range, we can envisage a spectrum of effects that represent the intermediate ground between these two models. For example, a four-marker model could include just the main effects and the pairwise interaction terms. Such a model will have a lower complexity than a full haplotype model, and hence fewer DF in the hypothesis test. As has been shown in previous chapters, situations exist whereby either the main effect or haplotype model have optimal performance, it is a natural extension to consider that situations occur where optimal performance is supplied by an intermediately parameterized model. Chapman *et al.* (2003) have proposed that the optimal model for an association test may have complexity between locus-scoring and haplotype scoring. This is certainly expected to be the case in some situations, although this raises the question of how to identify the combination of parameters that has the optimal performance for a given situation. An approach similar to that outlined in chapters four and five could applied, with models representing each combination of parameterization being tested to identify the underlying genetic conditions that lead to their optimal performance over other models. However, the potential number of parameter combinations between a set of markers makes this an unrealistic strategy to follow. Clustering approaches are often applied to haplotype analyses as a way of reducing the total number of parameters fitted by the model. However, as has been previously discussed, clustering approaches often struggle to identify differences in statistical effects between haplotypes that are deemed similar to one another (Durrant *et al.* 2004).

An alternative approach is to use statistical tools to reduce the parameter space such that only effects that contribute to the overall association remain. The difficulty here is to correctly identify the contributing effects, and removing or shrinking those that do not, without penalising the statistical support of the model. There is a large number of parameter reduction and model selection approaches available, although their ability to correctly identify the optimal parameterisation of a model is not fully understood with respect to association testing amongst multilocus models. Investigating the ability of these approaches to correctly identify genetic effects and provide an accurate model choice across the diverse range of genomic architecture would form a natural extension to the research provided within this thesis.

REFERENCES

Abasht, B., Dekkers, J. C. M. and Lamont S. (2006) Review of quantitative trait loci identified in the chicken. *Poultry Science*, **85**, 2079-2096.

Abdallah, J. M., Mangin, B., Goffinet, B., Cierco-Ayrolles, C. and Perez-Enciso, M. (2004) A comparison between methods for linkage disequilibrium fine mapping. *Genetical Research*, **83**, 41-47.

Aerts, J., Megens, H. J., Veenendaal, T., Ovcharenko, I., Crooijmans, R., Gordon, L., Stubbs, L. and Groenen M. (2007) Extent of linkage disequilibrium in chicken. *Cytogenetic and Genome Research*, **117**, 338:345.

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

Altshuler, D., Daly, M. J. and Lander, E. S. (2008) Genetic mapping in human disease. *Science*, **322**, 881-888.

Andersson, L. and Georges, M. (2004) Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews Genetics*, **5**, 202-213.

Andreescu, C., Avendano, S., Brown. S., Hassen, A., Lamont, S. and Dekkers, J. C. M. (2007) Linkage disequilibrium in related breeding lines of chickens. *Genetics*, **177**, 2161-2169.

Ardlie, K. G., Kruglyak, L. and Seielstad M. (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, **3**, 299-309.

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

Baghbanzadeh A. and Decuypere E. (2008) Ascites syndrome in broilers: physiological and nutritional perspectives. *Avian Pathology*, **37**, 117-126.

Bahlo, M., Stankovich, J., Speed, T. P., Rubio, J. P., Burfoot, R. K. and Foote, S. J. (2006) Detecting genome-wide haplotype sharing using SNP or microsatellite haplotype data. *Human Genetics*, **119**, 38-50.

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

Balog, J. M. (2003) Ascites syndrome (pulmonary hypertension syndrome) in broiler chickens: Are we seeing the light at the end of the tunnel? *Avian Poultry Biological Reviews*, **14**, 99-126.

Bardel, C., Darlu, P. and Genin, E. (2006) Clustering of haplotypes based on phylogeny: how good a strategy for association testing? *European Journal of Human Genetics*, **14**, 202-206.

Bendheim, H., Berman, E., Zadikov, I. and Sholsberg, A. (1992) The effects of poor ventilation, low temperatures, type of feed and sex of bird on development of ascites in broilers. Production parameters. *Avian Diseases*, **39**, 285-291.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289-300.

Bhattacharjee, S., Kuo, C. L., Mukhopadhyay, N., Brock, G. N., Weeks, D. E. and Feingold, E. (2008) Robust score statistics for QTL linkage analysis. *American Journal of Human Genetics*, **82**, 567-582.

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

Browning, B. L. and Browning, S. R. (2009) A unified approach to genotype imputation and haplotype phase inference for large datasets of trios and unrelated individuals. *The American Journal of Human Genetics*, **84**, 210-223.

Calus, M. P. L., Meuwissen T. H. E., De Roos, A. P. W. and Veerkamp, R. F. (2008) Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, **178**, 553-561.

Calus, M. P. L., Meuwissen, T. H. E., Windig, J. J., Knol, E. F., Schrooten, C., Vereijken, A. L. J. and Veerkamp, R. F. (2009) Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genetics Selection Evolution*, **41**, 11.

Cardon, L. R. and Abecasis, G. R. (2003) Using haplotype blocks to map human complex traits loci. *Trends in Genetics*, **19**, 135-140.

Carr, D. B. (1991) *Looking at large data sets using binned data plots*. Computing and Graphics in Statistics. Springer-Verlag, New York, pp. 7-39.

Chapman, J., Cooper, J., Todd, J. and Clayton, D. (2003) Detecting disease associations due to linkage disequilibria using haplotype tags: a class of tests and the determinants of statistical power. *Human Heredity*, **56**, 18-31.

Cheng, R., Ma, J. Z., Elston, R. C. and Li, M. D. (2005) Fine mapping functional sites or regions from case control data using haplotypes of multiple linked SNPs. *American Journal of Human Genetics*, **69**, 102-112.

Churchill, G. A. and Doerge, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963-971.

Clark, A. G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, **7**, 111-122.

Clayton, D., Chapman, J. and Cooper J. (2004) Use of unphased multilocus genotype data in indirect association studies. *Genetic Epidemiology*, **27**, 415-428.

Clayton, D. Personal web page (2009). Last accessed 29.05.2009. http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt

Conti, D. and Gauderman, J. (2004) SNPs, haplotypes and model selection in a candidate gene region: The SIMPle analysis for multilocus data. *Genetic Epidemiology*, **27**, 429-441.

Crooks, L., Sahana, G., de Koning, D. J., Lund, M. S. and Carlborg O. (2009) Comparison of analyses of the QTLMAS XII common dataset. II: genome-wide association and fine mapping. *BMC Proceedings*, **3** (Supplement 1): S2.

De Bakker, P. I., Yelensky, R., Pe'er I., Gabriel, S. B., Dalym, M. J. and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nature Genetics*, **37**, 1217-1223.

De Greef, K. H., Janss L. L., Vereijken A. L., Pit, R. and Gerritsen C. L. (2001) Disease-induced variability of genetic correlations: Ascites in broilers as a case study. *Journal of Animal Science*, **79**, 1723-1733.

De Koning, D. J., Haley, C., Windsor, D., Hocking, P., Griffen, H., Morris, A., Vincent, J. and Burt, D. (2004) Segregation of QTL for production traits in commercial meat-type chickens. *Genetical Research*, **83**, 211-220.

De Roos, A. P. W., Hayes, B. J., Spelman, R. J. and Goddard, M. E. (2008) Linkage disequilibrium and persistence of phase in Holstein-Friesian Jersey and Angus cattle. *Genetics*, **179**, 1503-1512.

Decuypere, E., Buyse, J. and Buys, N. (2000) Ascites in broiler chickens: exogenous and endogenous structural and functional causal factors. *World's Poultry Science*, **56**, 367-377.

Dekkers, J. C. M. (2004) Commercial application of marker and gene assisted selection in livestock: Strategies and lessons. *Journal of Animal Science*, **82**, 313-328.

Dekkers, J. C. M. and Hospital, F. (2002) The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics*, **3**, 22-32.

Dekkers, J. C. M., Zhao, H. and Fernando, R. L. (2006) Linkage disequilibrium mapping of QTL in livestock. 8th World Congress on Genetic Applications for Livestock Production. Belo Horizonte, Brazil.

Ding, K., Zhou, K., Zhang, J., Knight, J., Zhang, X. and Shen, Y. (2005) The effect of haplotype block definitions on inference of haplotype block structure and htSNP selection. *Molecular Biology and Evolution*, **22**, 148-159.

Druyan, S., Cahaner, A., Bellaiche, M. and Shlosberg, A. (1999) Genetic evaluation of blood oxygenation, heart rate, electrocardiographic (ECG) waveforms and their associations with ascites in broilers. Presented at the 1999 European poultry breeders Roundtable, Wiesensee (Germany).

Druyan, S., Ben-David, A. and Cahaner, A. (2007a) Development of ascites resistant and ascites susceptible broiler lines. *Poultry Science*, **86**, 811-822.

Druyan, S., Shlosberg, A. and Cahaner, A. (2007b) The association between the ascites syndrome and body weight, heart rate and blood measurements in young chicks and in ascitic verses healthy broilers. *Poultry Science*, **86**, 621-629.

Dunning, A. M., Durocher, F., *et al.* (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics*, **67**, 1544-1554.

Dunnington, E. A. and Siegal, P. B. (1996) Long term divergent selection for eight week body weight in White Plymouth Rock chickens. *Poultry Science*, **75**, 1168-1179.

Durrant, C., Zondervan, K. T., Cardon, L. R., Deloukas, P. and Morris, A. (2004) Linkage disequilibrium mapping via Cladistic analysis of SNP haplotypes. *American Journal of Human Genetics*, **75**, 35-43.

Efron, B. and Tibshirani, R. (2002) Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**, 70-86.

Eichler, E. E., Nickerson, D. A., *et al.* (2007) Completing the map of human genetic variation. *Nature*, **447**, 161-165.

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

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

Fan, R. and Xiong, M. (2002) High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. *European Journal of Human Genetics*, **10**, 607-615.

Farnir, F., Coppieters, W., *et al.* (2000) Extensive genome-wide linkage disequilibrium in cattle. *Genome Research*, **10**, 220-227.

Fisher, R. A. (1925) *Statistical methods for research workers*. London: Oliver and Lloyd.

Fisher, R. A. (1948) Combining independent tests of significance. American Statistician, 2, 30.

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

Gabriel, S. B., Schaffner, S. F., *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225-2229.

Gaut, B. S. and Long, A. D. (2003) The lowdown on linkage disequilibrium. *Plant Cell*, **15**, 1502-1506.

George, A. W., Visscher, P. M. and Haley, C. S. (2000) Mapping quantitative trait loci in complex pedigrees: A two-step variance component approach. *Genetics*, **156**, 2081-2092.

Gilmour, A., Cullis, B., Welham, S., Thompson, R. (1998) ASREML User's Manual New South Wales Agriculture Institute, Orange, NSW Australia.

Goddard, M. E. and Hayes, B. J. (2009) Mapping genes for complex traits in domestic animals and their uses in breeding programmes. *Nature Reviews Genetics*, **10**, 381-391.

Grapes, L., Dekkers, J. C. M., Rothschild, M. and Fernando, R. L. (2004) Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics*, **166**, 1561-1570.

Grapes, L., Firat, M. Z., Dekkers, J. C. M., Rothschild, M. and Fernando, R. L. (2006) Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. *Genetics*, **172**, 1955-1965.

Gunderson, K. L., Kruglyak, S. *et al.* (2004) Decoding randomly ordered DNA arrays. *Genome Research*, **14**, 870-877.

Guo, W. and Lin, S. (2009) Generalized linear modelling with regularization for detecting common disease rare haplotype association. *Genetic Epidemiology*, **33**, 308-316.

Haley, C. S. (1999) Advances in QTL mapping, pp. 47-59 in *Proceedings of From Lush to Genomics: Visions for animal breeding and Genetics*, edited by J. C. M. Dekkers, S. J. Lamont and M. F. Rothschild. Iowa State University, Ames, IA.

Haley, C. S., Knott, S. A. and Elsen J. M. (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, **136**, 1195-1207.

HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851-862.

Hassen, A., Avendano, S., Hill, W. G., Fernando, R. L., Lamont, S. J. and Dekkers, J. C. M. (2009) The effect of heritability estimates on high-density single nucleotide polymorphism analyses with related animals. *Journal of Animal Science*, **87**, 868-875.

Havenstein, G. B., Ferket, P. R., Scheideler, S. E. and Larson, B. T. (1994) Growth, liveability, and feed conversion of 1957 versus 1991 broilers when fed 'typical' 1957 and 1991 broiler diets. *Poultry Science*, **73**, 1785-1794.

Havenstein, G. B., Ferket, P. R. and Qureshi, M. A. (2003) Growth, liveability, and feed conversion of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. *Poultry Science*, **82**, 1500-1508.

Hawley, M. E. and Kidd, K. K. (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, **86**, 409-411.

Hayes, B. J., Visscher, P. M., McPartlan, H. C. and Goddard, M. E. (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*, **13**, 635-643.

Hayes, B. J., Chamberlain, A., McPartlan, H., Macleod, I., Sethuraman, L. and Goddard, M. (2007) Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genetic Research*, **89**, 215-220.

Heifetz, E. M., Fulton, J. E., Sullivan, N. O., Zhao, H., Dekkers, J. C. M. and Soller, M. (2005) Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics*, **171**, 1173-1181.

Henderson, C. R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **32**, 423-447.

Hill, W. G. and Robertson A. (1968) Linkage disequilibrium in finite populations. *Theoretical Applied Genetics*. **38**, 226-231.

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

Hochberg, Y. and Tamhane, A. C. (1987) *Multiple comparison procedures*. Wiley, New York.

Hocking, P. (2005) Review of QTL mapping results in chickens. *World's Poultry Science Journal*, **61**, 215-226.

Hudson, R. R., Slatkin, M. and Maddison, W. P. (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583-589.

Hudson, R. R. (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**, 183-201.

Humphreys, K. and Iles, M. M. (2005) Fine-scale mapping in case-control samples using locus scoring and haplotype-sharing methods. *BMC Genetics*, **6**(Suppl 1):S74.

Hunton, P. (1998) The potential of genetics to combat ascites. *World Poultry*, **14**, 64-66.

Igo, R. P. Jr., Li, J. and Goddard, K. A. B. (2009) Association mapping by generalized linear regression with density-based haplotype clustering. *Genetic Epidemiology*, **33**, 16-26.

Julian, R. J. (1993) Ascites in poultry. *Avian Pathology*, **22**, 419-454.

Julian, R. J. (1998) Pulmonary hypertension as a cause of right ventricular failure and ascites in broilers. *Zootechnica International*, **11**, 58-62.

Julian, R. J. (1990a) Cardiovascular disease, In Poultry Diseases, 3rd edition, Bailliere Tindell, London, England, pp. 330-353.

Julian, R. J. (1990b) Pulmonary hypertension: A cause of right heart failure, ascites in meat-type chickens. *Feedstuffs*, January, **78**, 19-22.

Julian, R. J. (2000) Physiological, management and environmental triggers of ascites syndrome: a review. *Avian Pathology*, **29**, 519-527.

Julian, R. J. and Mirsalimi, S. M. (1992) Blood oxygen concentration of fast growing and slow growing broiler chickens, and chickens with ascites from right ventricular failure. *Avian Diseases*, **36**, 730-732.

Kennedy, B. W., Quinton, M. and van Arendonk, J. A. M. (1992) Estimation of effects of single genes on quantitative traits. *Journal of Animal Science*, **70**, 2000-2012.

Khatkar, M. S., Nicholas, F. W., *et al.* (2008) Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC genomics*, **9**, 187.

Kidd, J. M., Cooper, G. M., *et al.* (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56-64.

Kirk, K. M. and Cardon, L. R. (2002) The impact of genotyping error on haplotype reconstruction and frequency estimation. *European Journal of Human Genetics*, **10**, 616-622.

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

Lam, A. C., Powell, J. E., Wei, W. H., De Koning D. J. and Haley C. S. (2009) A combined strategy for quantitative trait loci detected by genome-wide association. *BMC Proceedings*, **3** (Supplement 1): s8.

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

Lange, K. (1997) *Mathematical and statistical methods for genetic analysis*. Springer-Verlag, New York

Larribe, F., Lessard, S and Schork, N. J. (2002) Gene mapping via the ancestral recombination graph. *Theoretical Population Biology*, **62**, 215-229.

Lee, S. H. and Van der Werf, J. H. (2004) The efficiency of designs for fine mapping of quantitative trait loci using combined linkage disequilibrium and linkage. *Genetics, Selection, and Evolution*, **36**, 145-161.

Lewontin, R. C. (1988) On measures of gametic disequilibrium. *Genetics*, **120**, 849-852.

Li, Y., Sung, W. K. and Liu, J. J. (2007) Association mapping via regularised regression analysis of single-nucleotide-polymorphism haplotypes in variable-sized sliding windows. *American Journal of Human Genetics*, **80**, 705-715.

Lin, S., Chakravarti, A. and Cutler, D. J. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nature Genetics*, **36**, 1181-1188.

Lindblad-Toh, K., Wade, C. M., *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803-819.

Lister, S. (1997) Broiler ascites: a veterinary viewpoint. *World's Poultry Science*, **12**, 499-510.

Liu, N., Sawyer, S. L., Mukherjee, N., Pakstis, A. J., Kidd, J. R., Kidd, K. K., Brookes, A. J. and Zhao, H. (2004) Haplotype block structures show significant variation among populations. *Genetic Epidemiology*, **27**, 385-400.

Liu, Z. and Lin, S. (2005) Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genetic Epidemiology*, **29**, 353-364.

Long, J. C., Williams, R. C. and Urbanek, M. (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, **56**, 799-810.

Long, A. D. and Langley, C. H. (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research*, **9**, 720-731.

Lubritz, D. L., Smith, J. L. and McPherson, B. (1995) Heritability of ascites and the ratio of right to total ventricle weight in broiler breeder male lines. *Poultry Science*, **74**, 1237-1241.

Lund, M. S., Sahana, G., De Koning, D. J., Su, G. and Carlborg, O. (2009) Comparison of analyses of the QTLMAS XII common dataset I: Genomic selection. *BMC Proceedings*, **3** (Supplement 1): s1.

Lynch, M. and Hill, W. G. (1986) Phenotypic evolution and neutral mutation. *Evolution*, **40**, 915-935.

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

Marchini, J., Cardon, L. R., Phillips, M. S. and Donnelly, P. (2004) The effects of human population structure on large genetic association studies. *Nature Genetics*, **36**, 512-517.

Martin, E. R., Lai, E. H., *et al.* (2000) SNPing away at complex diseases: Analysis of Single-Nucleotide polymorphisms around APOE in Alzheimer disease. *American Journal of Human Genetics*, **67**, 383-394.

Maxwell, M. H., Robertson, G. W. and Spence, S (1986) Studies on a ascitic syndrome in young broiler. 1. Haematology and pathology. *Avian Pathology*, **15**, 511-524.

Maxwell M. H. and Robertson G. W. (1997) World broiler ascites survey. *Poultry International* **36**, 16-30.

May, J. D. and Deaton, J. W. (1974) Environmental temperature effect on heart weight of chickens. *International Journal of Biometeorology*, **15**, 295-300.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A. and Hirschhorn, J. N. (2008) Genome-wide association studies for complex traits: consensus uncertainty and challenges. *Nature Reviews Genetics*, **9**, 356-369.

McKay, J. C., Barton, N. F., Koerhuis, A. N. M. and McAdam, J. (2000) The challenge of genetic change in broiler chicken. *BSAS occasional publication number 27: The Challenge of Genetic Change in Animal Production*, pp. 1-7.

McKay, S. D., Schnabel, R. D., *et al.* (2007) Whole genome linkage disequilibrium maps in cattle. *BMC genetics*, **8**, 74.

McRae, A. F., McEwan, J. C., *et al.* (2002) Linkage disequilibrium in domestic sheep. *Genetics*, **160**, 1113-1122.

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

Meuwissen, T. H. E. and Goddard, M. (2001) Prediction of identity by descent by descent probabilities from marker haplotypes. *Genetics, Selection, Evolution*, **33**, 605-634.

Meuwissen, T. H. E., Hayes, B. J. and Goddard, M. E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819-1829.

Mitchell M. A. (1997) Ascites syndrome: A physiological and biochemical perspective. *World's Poultry Science* **53**, 61-64.

Moghaddam, H. K., McMillan I., Chambers J. R. and Julian R. (2001) Estimation of genetic parameters for ascites syndrome in broiler chickens. *Poultry Science*, **80**, 844-848.

Morris, R. and Kaplan, N. (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *American Journal of Human Genetics*, **23**, 221-233.

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

Muir, W. M. Wong, G. K. S., *et al.* (2008) Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proceedings of the National Academy of Science*, **105**, 17312-17317.

Navarro, P. (2003) Genetic study of ascites in broiler populations. PhD thesis. University of Edinburgh.

Navarro, P., Visscher, P., Knott, S. A., Burt, D., Hocking, P. M. and Haley, C. S. (2005) Mapping of quantitative trait loci affecting organ weights and blood variables in a broiler layer cross. *British Poultry Science*, **4**, 430-442.

Navarro, P., Visscher, P., Chatziplis D., Koerhuis A. N. K. and Haley, C. (2006) Segregation analysis of blood oxygen saturation in broilers suggests a major gene influence on ascites. *British Poultry Science*, **47**, 671-684.

Nielsen, D. M., Ehm, M. G., Zaykin, D. V. and Weir, B. S. (2004) Effect of two and three locus linkage disequilibrium on the power to detect marker / phenotype associations. *Genetics*, **168**, 1029-1040.

Niu, T. (2004) Algorithms for inferring haplotypes. *Genetic Epidemiology*, **27**, 334-347.

Nordborg, N. and Tavare, S. (2002) Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, **18**, 83-90.

North, B. V., Sham, P. C., Knight, J., Martin, E. R. and Curtis, D. (2006) Investigation of the ability of haplotype association and logistic regression to identify associated susceptibility loci. *Annals of Human Genetics*, **70**, 893-906.

Nothnagel, M., Frst, R. and Rohde, K. (2002) Entropy as a measure of linkage disequilibrium over multilocus haplotype blocks. *Human Heredity*, **54**, 186-198.

Nsengimana, J., Baret, P., Haley C. S. and Visscher P. M. (2004) Linkage disequilibrium in domesticated pig. *Genetics*, **166**, 1395-1404.

Odom, T. W. (1993) Ascites syndrome: Overview and update. *Poultry Digest*, **52**, 14-22.

Osier, M., Pakstis, A., Kidd, J. R., Lee, J. F., Yin, S. J., Ko, H. C., Edenberg, H., Lu, R. B. and Kidd, K. K. (1999) Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. *American Journal of Human Genetics*, **64**, 1147-1157.

Pakdel, A., Van Arendonk, J., Vereijken, A. and Bovenhuis, H. (2005) Genetic and phenotypic correlations for ascites related traits in broilers measured under cold and normal conditions. *British Poultry Science*, **46**, 35-42.

Pe'er, I., De Bakker, P. I., Maller, J., Yelensky, R., Altshuler, D. and Daley, M. J. (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics*, **38**, 663-667.

Peng, J. and Siegmund, D. (2006) QTL mapping under ascertainment. *Annuals of Human Genetics*, **70**, 867-881.

Piepho, H. (2001) A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics*, **157**, 425-432.

Pritchard, J. K. and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, **69**, 1-14.

Rabie, T. (2004) Pulmonary hypertension syndrome in chicken: peeking under QTL peaks. PhD Thesis, Wageningen University.

Rabie, T., Crooijmans, R., Bovenhuis, H., Vereijken, A., Veenendaal, T., Poel, J., Van Arendonk, J., Pakdel, A. and Groenen. (2005) Genetic mapping of quantitative trait loci affecting susceptability in chicken to develop pulmonary hypertension syndrome. *Animal Genetics*, **36**, 468-476.

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

Ron, M. and Weller, J. I. (2007) From QTL to QTN identification in livestock – winning by points rather than knock-out: a review. *Animal Genetics*, **38**, 429-439.

Sawyer, S. L., Mukherjee, N., Pakstis, A. J., Feuk, L., Kidd, J. R., Brookes, A. J. and Kidd, K. K. (2005) Linkage disequilibrium patterns vary substantially among populations. *European Journal of Human Genetics*, **13**, 677-686.

Schaid, D. (2004) Evaluating associations of haplotypes with traits. *Genetic Epidemiology*, **27**, 348-364.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. and Poland, G. A. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics*, **70**, 425-434.

Scheele, C. W., De Wit, W., Frankenhuis, M. T. and Vereijken, P. F. G. (1991) Ascites in broilers. 1. Experimental factors evoking symptoms related to ascites. *Poultry Science*, **70**, 1069-1093.

Searle, S. R. (1987) *Linear models for unbalanced data.* John Wiley and Sons, New York, NY.

Seltman, H., Roeder, K. and Devin, B. (2001) Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *American Journal of Human Genetics*, **70**, 425-434.

Seltman, H., Roeder, K. and Devin, B. (2003) Evolutionary-based association analysis using haplotype data. *Genetic Epidemiology*, **25**, 48-58.

Sinwell, J. P., Schaid, D. J., Rowland, C. M. and Yu, Z. (2008) haplo.stats: statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. R package version 1.3.4. Ref type: Computer program. http://cran.r-project.org/web/packages/haplo.stats/index.html

Slager, S., Huang, J. and Vieland, V. (2000) Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genetic Epidemiology*, **18**, 143-156.

Smith, A. H., Wilson, W. O. and Pace, N. (1955) Growth and reproduction in domestic birds at high altitudes. *Poultry Science*, **35** (Suppl.1), 1175.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A. and Meuwissen, T. H. E. (2008) Genomic selection using different marker types and densities. *Journal of Animal Science*, **86**, 2447-2454.

Sorenson, D. A. and Kennedy, B. W. (1986) Analysis of selection experiments using mixed model methodology. *Journal of Animal Sciences*, **63**, 245-258.

Stephens, M., Smith, N. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978-989.

Stolz, J. L, Rosenbaum, L. M., Jeong, D. and Odom, T. W. (1992) Ascites syndrome, mortality and cardiological responses of broiler chickens subjected to cold exposure. *Poultry Science*, **71** (Suppl. 1), 4.

Sved, J. A. (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, **2**, 125-141.

Tanck, M., Klerkx, A., Jukema, J., DeKnijff, P., Kastelein, J. and Zwinderman, A. (2003) Estimation of multilocus haplotype effects using weighted penalised log-likelihood: analysis of five sequence variations at the cholesterol ester transfer protein gene locus. *Annals of Human Genetics*, **67**, 175-184.

Tishkoff, S. A., Pakstis, A. J., Ruano, G. and Kidd, K. K. (2000) The accuracy of statistical methods for estimation of haplotype frequencies: An example from CD4 locus. *American Journal of Human Genetics*, **67**, 518-522.

Terwilliger, J. D. (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics*, **56**, 777-787.

Tzeng, J. Y., Devlin, B., Wasserman, L. and Roeder, K. (2003) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *American Journal of Human Genetics*, **72**, 891-902.

Tzeng, J. Y., Wang, C. H., Kao, J. T. and Hsiao C. K. (2006) Regression-based analysis with clustered haplotypes through use of genotypes. *American Journal of Human Genetics*, **78**, 231-242.

Vallejo, R. L., Li, L. Y., Rogers, G. W. and Ashwell, M. S. (2003) Genetic diversity and background linkage disequilibrium in the North American Holstein cattle population. *Journal of Dairy Science*, **86**, 4137-4147.

Van Arendonk, J. A. M. and Bovenhuis, H. (2003) Designs and methods to detect QTL for production traits based on mapped genetic markers. In *Poultry genetics, breeding and technology*. Cambridge International, pp 439-464.

Van der Beek, S., Van Arendonk, J. A. M. and Groen, A. F. (1995) Power of two-generation and three-generation QTL mapping experiments in an outbred population containing full-sib or half-sib families. *Theoretical and Applied Genetics*, **91**, 1115-1124.

VanLiere, J. M. and Rosenberg, N. A. (2008) Mathematical properties of the r^2 measure of linkage disequilibrium. *Theoretical Population Biology*, **74**, 130-137.

Visscher, P. M. (2008) Sizing up human height variation. *Nature Genetics*, **40**, 489-490.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G. and Todd, J. A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nature Review Genetics*, **6**, 109-118.

Weller, J. I. *Quantitative trait loci analysis*, CABI Publishing, Wallingford, Oxon, UK, 2001.

Weir, B. *Genetic Data Analysis II.* Sinauer Associates, Inc; Sunderland, Massachusetts. 1996.

Weir, B. S. and Cockerham, C. C. (1977) *Two-locus theory in quantitative genetics*: Proceedings of the International Conference on Quantitative Genetics. Ames, Iowa State University Press, pp. 247-269.

Weiss, K. M. and Clark, A.G. (2002) Linkage disequilibrium and mapping complex traits. *Trends in Genetics*, **18**, 19-24.

Wideman, R. F. Jr. (1988) Ascites in poultry. *Monsanto Nutrition Update*, **6(2),** 1-7.

Wideman, R. F. Jr. (2000) Cardio-pulmonary hemodynamics and ascites in broiler chickens. *Avian and Poultry Biological Review*, **11**, 21-43.

Wideman, R. F. Jr. and Tackett, C. D. (2000) Cardio-pulmonary function in broiler reared at warm and cool temperatures: effect of acute inhalation of 100% oxygen. *Poultry Science*, **79**, 257-364.

Wong, G. K. Lui, B. *et al.* (2004) A genetic variation map for chicken with 2.8 million single nucleotide polymorphisms. *Nature*, **432**, 717-722.

Wright, S. (1935) Evolution in populations in approximate equilibrium. *Journal of Genetics*, **30**, 257-266.

Ye, X., Avendano, S., Dekkers, J. C. M. and Lamont, S. J. (2006) Association of twelve immune-related genes with performance of three broiler lines in two different hygiene environments. *Poultry Science*, **85**, 1555-1569.

Yu, Z. and Schaid, D. J. (2007) Sequential haplotype scan methods for association analysis. *Genetic Epidemiology*, **31**, 551-564.

Wright, S. (1935) Evolution in populations in approximate equilibrium. *Journal of Genetics*, **30**, 257-266.

Zaykin, D. V., Westfall, P., Young, S. S., Karnoub, M. A., Wagner, M. J. and Ehm, M. G. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity*, **53**, 79-91.

Zhang, K., Calabrese, P., Nordberg, M. and Sun, F. (2002) Haplotype block structure and its application to association studies: Power and study design. *American Journal of Human Genetics*, **71**, 1386-1394.

Zhang, X. S., Wang, J. and Hill, W. G. (2004) Redistribution of gene frequency and changes of genetic variation following a bottleneck in population size. *Genetics*, **167**, 1475-1492.

Zhao, H., Pfeiffer, R. and Gail, M. H. (2003) Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, **4**, 171-178.

Zhao, H., Fernando, R. L. and Dekkers, J. C. M. (2007) Power and precision of alternative methods for linkage disequilibrium mapping of quantitative trait loci. *Genetics*, **175**, 1975-1986.

Zhao, L., Li, S. and Khalid, N. (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics*, **72**, 1231-1250.

Zondervan, K. T. and Cardon, L. R. (2004) The complex interplay among factors that influence allelic association. *Nature Reviews Genetics*, 5, 89-100.

Zou, F., Fine, J. P,, Hu, J. and Lin, D. Y. (2004) An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics* 168, 2307–2316.

Appendix one

Details of the formula to calculate sire SaO_2

The formula given on page 39, section 2.2.1, describes the method used to determine the adjusted phenotypic records for each sire. Recorded trait values are adjusted to account for variance due to fixed and random effects not associated with the breeding value of the sire. The model was tested and the significance of factors explored. Since SaO_2 was analysed in a multivariate analysis along with early and late mortality in pedigree and stress environments, these effects described below were important for all the traits.

As is shown in section 2.2.1, the adjusted trait value for a given sire is calculated according to;

$$y_{adj} = \frac{\sum_{i=1}^{n}(y_i - a_i - h_i - mg_i - c_i - (0.5 * BV_{dam_i}))}{n},$$

where y_i is the trait record for individual i, a_i, h_i and mg_i are the effects of sex, age, hatch and mating group respectively for individual i, c_i is the random effect corresponding to the permanent environmental effect of the dam for individual i. BV_{dam_i} is the estimated breeding value of the dam for individual i and n is the number of progeny for the sire. Details of the components of the formula are given below.

Age: This is the age of the dam for individual *i*. Age of the dam is positively correlated with egg size, and therefore weight of chick at hatching.

Mating Group: Refers to the parents of individual *i*. It denotes that the parents were located on a specific farm and are a contemporary group from the same generation, with laying starting at a specific point in time (i.e. hatch week). The mating group captures all environmental factors specific to the parents, such as when they were hatched, farm and when they started and when they started laying.

Hatch: This is the hatch week of individual *i*. It captures factors associated with hatching and growing of the individual.

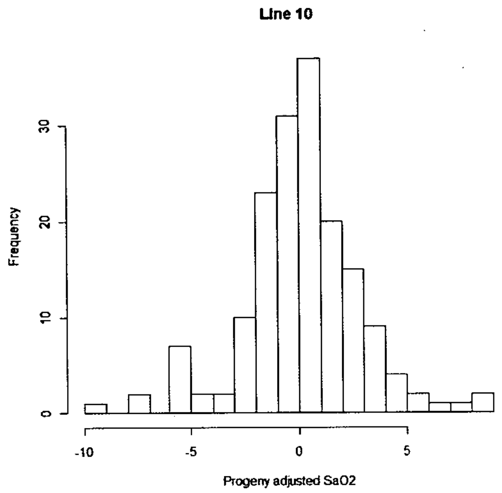
Dam permanent environment: Is a random effect corresponding to the permanent environment of the dam for individual *i*. Among the six lines the ratio of c_i over phenotypic variance varied a lot, yet log-likelihood tests suggested that it needed to be included in the model.

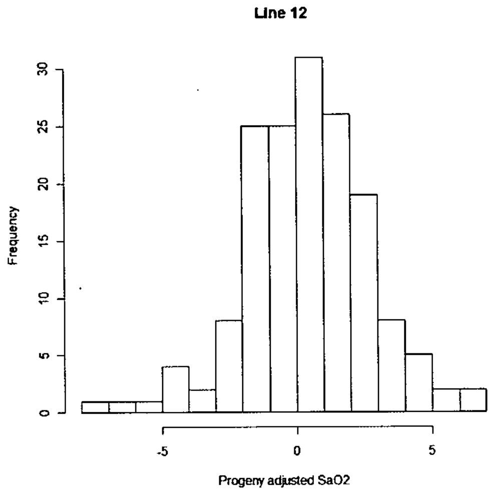
Breeding value of dam: The breeding value of the dam was also included within the model. This breeding value was calculated using a normal multivariate BLUP analyses, fitting a full pedigree, so breeding values were calculated having accounted for all relationships. Since the breeding value of the dam is determined using a BLUP analysis, no genetic effects of sires should be included, since the A-matrix should have accounted for those.

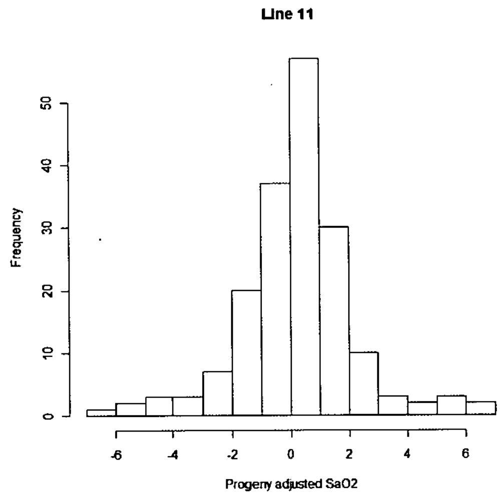
Appendix two

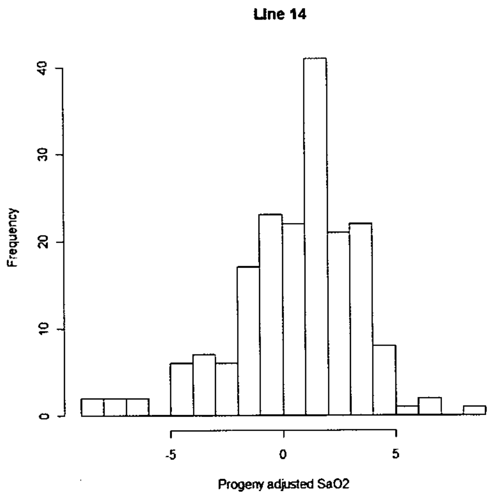
Distributions of phenotypic records for each line

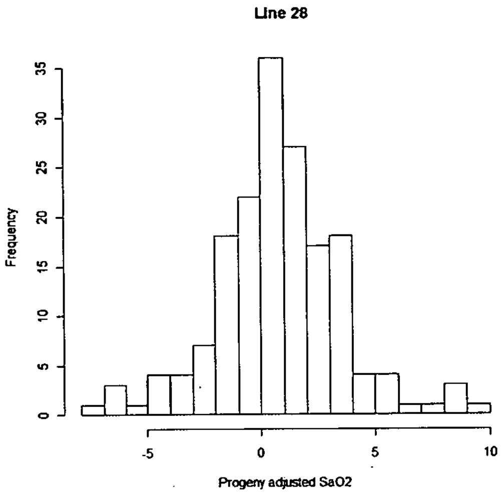
The distribution of the progeny adjusted SaO_2 measures for each line are shown in figures below. Individuals with a trait value greater than 3 standard deviations from the mean of the line were removed from dataset and not shown on the distributions.

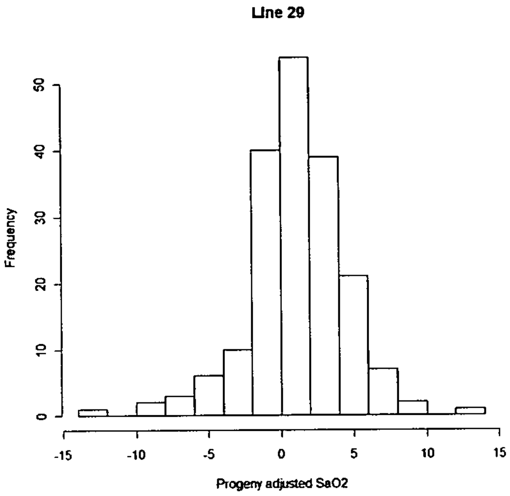






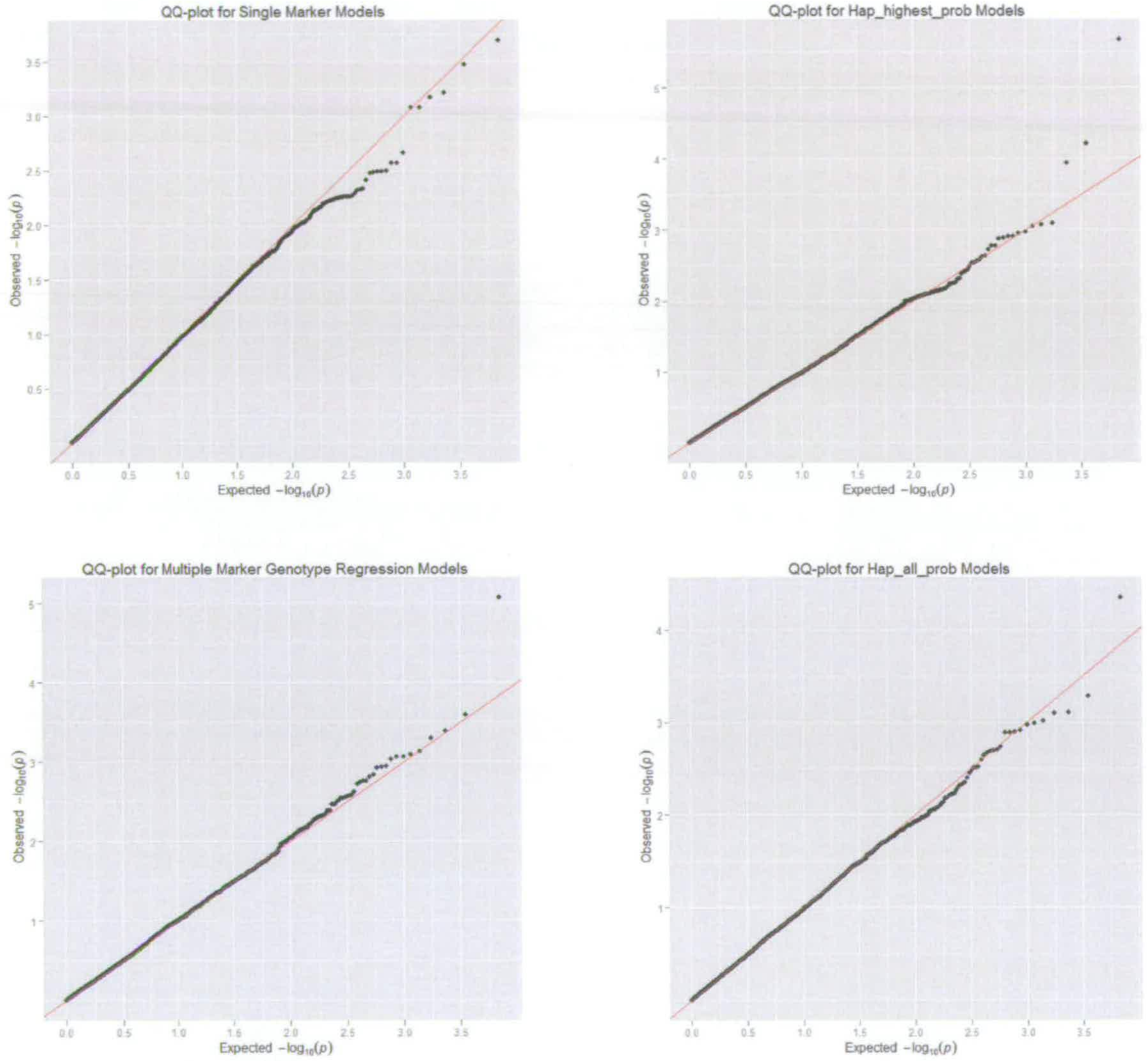



Line 28
Frequency
0
5
10
15
20
25
30
35
-5
0
5
10
Progeny adjusted SaO2


Line 29
Frequency
0
10
20
30
40
50
-15
-10
-5
0
5
10
15
Progeny adjusted SaO2

Appendix three

Q-Q plots from permutation analyses



Examples of Q-Q plots from a single permutation run for each of the models. Black dots represent the *p*-values from the permutation analysis whilst the red line represents the expected null distribution of *p*-values. Plots show little evidence for population stratification impacting on *p*-vales from the analysis. Corresponding genomic inflation factors (λ) are all close to 1.

Appendix four

Code to determine haplotypes (EM – algorithm) and fit linear models

```
#=================================================================#
#                                                                 #
#     Syntax:     R                                               #
#                                                                 #
#     Description:  Script to read in pedigree, match with the    #
#     data file, invert the pedigree and use the output to run    #
#     a hap_highest_prob and hap_all_prob models                  #
#                                                                 #
#     Require: Use of haplotype data - EM_algo.R                  #
#                                                                 #
#     Author:     Joseph Powell                                   #
#                 joseph.powell@qimr.edu.au                       #
#                                                                 #
#=================================================================#

#----------#
#     Section 1: Read in relevent data files and sort
#----------#

library(GenABEL)
genabel.data <- load.gwaa.data(phenofile="genabel.pheno.line29.dat",
genofile="GenABEL.input.line29.raw", force=F, makemap=T)
binary.geno <- as.numeric(genabel.data@gtdata)

maf.info <- read.csv("maf_adj.csv", header=T)
attach(maf.info)

dim(maf.info)[1]==12046
dim(maf.info)[1]==dim(binary.geno)[2]

pedigree <- read.table("ped_line29_rtools_edit.txt", skip=1)
dim(pedigree)

animal.id <- read.table("animal.id.txt")
head(animal.id)
dim(animal.id)

#----------#
#     Section 2: Remove fixed SNPs + low call rate markers and
individuals
#----------#

#     Fixed markers

fixture <- 0
fixed.snps <- NULL
for (i in 1:(dim(maf.info)[1])) {
      fixed <- maf.info[i, 10] > fixture
      fixed.snps <- rbind(fixed.snps, fixed)
```

```
}

fixed.snps <- as.data.frame(fixed.snps)
head(fixed.snps)
table(fixed.snps)

binary.geno.qc.removed <- binary.geno[, fixed.snps$V1]
genabel.data.qc.removed <- genabel.data[, fixed.snps$V1]
maf.info.qc.removed <- maf.info[fixed.snps$V1, ]

#     Low call rate markers

marker.call.rate <- 0.95
bad.markers.tot <- NULL
for (i in 1:(dim(maf.info.qc.removed)[1])) {
      snp <- binary.geno.qc.removed[, i]
      bad.markers <- (sum(table(snp)))/(dim(binary.geno)[1])
      bad.markers <- bad.markers > marker.call.rate
      bad.markers.tot <- rbind(bad.markers.tot, bad.markers)
}

bad.markers.tot <- as.data.frame(bad.markers.tot)
head(bad.markers.tot)
table(bad.markers.tot)

binary.geno.qc.removed <- binary.geno.qc.removed[, bad.markers.tot$V1]
genabel.data.qc.removed <- genabel.data.qc.removed[,
bad.markers.tot$V1]
maf.info.qc.removed <- maf.info.qc.removed[bad.markers.tot$V1, ]

#     Low call rate individuals

indi.call.rate <- 0.99
bad.indi.tot <- NULL
for (i in 1:(dim(binary.geno.qc.removed)[1])) {
      bad.indi <- (sum(table(binary.geno.qc.removed[i,
])))/(dim(binary.geno.qc.removed)[2])
      bad.indi <- bad.indi > indi.call.rate
      bad.indi.tot <- rbind(bad.indi.tot, bad.indi)
}

bad.indi.tot <- as.data.frame(bad.indi.tot)
head(bad.indi.tot)
table(bad.indi.tot)

binary.geno.qc.removed <- binary.geno.qc.removed[bad.indi.tot$V1, ]
genabel.data.qc.removed <- genabel.data.qc.removed[bad.indi.tot$V1, ]
animal.id <- animal.id[bad.indi.tot$V1, ]

# Remove all markers with a missing value left

missing.tot <- NULL
for(i in 1:(dim(binary.geno.qc.removed)[2])) {
      marker <- binary.geno.qc.removed[,i]
      num.genotyped <- sum(table(marker))
      missing <- num.genotyped==(dim(binary.geno.qc.removed)[1])
      missing.tot <- rbind(missing.tot, missing)
}

missing.tot <- as.data.frame(missing.tot)
head(missing.tot)
table(missing.tot)

binary.geno.qc.removed <- binary.geno.qc.removed[, missing.tot$V1]
genabel.data.qc.removed <- genabel.data.qc.removed[, missing.tot$V1]
maf.info.qc.removed <- maf.info.qc.removed[missing.tot$V1, ]

# Finally remove markers with less than 3 genotypes

geno.len.tot <- NULL
for (i in 1:(dim(binary.geno.qc.removed)[2])) {
      geno <- length(unique(binary.geno.qc.removed[,i]))
      geno.len <- geno==3
      geno.len.tot <- rbind(geno.len.tot, geno.len)
}

geno.len.tot <- as.data.frame(geno.len.tot)
head(geno.len.tot)
table(geno.len.tot)

binary.geno.qc.removed <- binary.geno.qc.removed[, geno.len.tot$V1]
genabel.data.qc.removed <- genabel.data.qc.removed[, geno.len.tot$V1]
maf.info.qc.removed <- maf.info.qc.removed[geno.len.tot$V1, ]

#----------#
#     Section 3:  Create haplotype data
#----------#
#     Note: this stage can take a while so save the output and read it
in for future analysis
#geno.haplo.qc.removed <- as.hsgeno(genabel.data.qc.removed)
#write.table(geno.haplo.qc.removed, "geno.haplo.qc.removed.txt",
quote=F, row.names=F, col.names=F, sep="\t")

#     Read haplotype genotype in from here
geno.haplo.qc.removed <- read.table("geno.haplo.qc.removed.txt",
header=F)
attach(geno.haplo.qc.removed)
dim(geno.haplo.qc.removed)

#----------#
# Section 3.1 EM - algorithm.
# See Dave Clayton's progressive insertion also
# + haplo.stats
#----------#

EM_algo.fun <- function (geno, locus.label = NA, miss.val = c(0, NA),
weight = NULL,
    control = haplo.em.control())
```

```
{
    n.loci <- ncol(geno)/2
    n.subject <- nrow(geno)
    subj.id <- 1:n.subject

    # Determine data is correct
    if (n.loci < 2)
        stop("Must have at least 2 loci for haplotype estimation!")
    if (any(is.null(weight))) {
        weight <- rep(1, n.subject)
    }
    if (any(weight < 0)) {
        stop("negative weights not allowed")
    }
    if (length(weight) != n.subject) {
        stop("Length of weight != number of subjects (nrow of geno)")
    }
    if (all(is.na(locus.label)))
        locus.label <- paste("loc-", 1:n.loci, sep = "")
    if (length(locus.label) != n.loci) {
        stop("length of locus.label != n.loci")
    }

    temp.geno <- loci(geno, locus.names = locus.label, miss.val =
miss.val)
    max.pairs <- geno.count.pairs(temp.geno)
    max.haps <- 2 * sum(max.pairs)

    if (max.haps > control$max.haps.limit)
        max.haps <- control$max.haps.limit
    rows.rem <- numeric(0)
    geno.vec <- as.vector(temp.geno)
    geno.vec <- ifelse(is.na(geno.vec), 0, geno.vec)
    allele.labels <- attr(temp.geno, "unique.alleles")

    if (length(allele.labels) != n.loci)
        stop("Number of loci in alleles list != n.loci")
    n.alleles <- numeric(n.loci)
    a.freq <- vector("list", n.loci)

    for (i in 1:n.loci) {
        n.alleles[i] <- length(allele.labels[[i]])
        j <- (i - 1) * 2 + 1
        p <- table(temp.geno[, c(j, (j + 1))], exclude = NA)
        p <- p/sum(p)
        a.freq[[i]] <- list(p = p)
    }

    if (is.null(control$loci.insert.order)) {
        control$loci.insert.order <- 1:n.loci
    }
    loci.insert.order <- (control$loci.insert.order - 1)
    if (length(loci.insert.order) != n.loci) {
        stop("length of loci.insert.order != n.loci")
    }
    if (sum(abs(sort(loci.insert.order) - (0:(n.loci - 1)))) >
        0) {
        stop("All loci are not accounted for in  loci.insert.order")
    }
    if (control$insert.batch.size > n.loci) {
        control$insert.batch.size <- n.loci
    }
    if (!is.null(control$iseed)) {
        set.seed(control$iseed)
    }
    else {
        runif(1)
        control$iseed <- .Random.seed
    }
    # Set seeds
    seed.array <- runif(3)
    iseed1 = 10000 + 20000 * seed.array[1]
    iseed2 = 10000 + 20000 * seed.array[2]
    iseed3 = 10000 + 20000 * seed.array[3]
    fit <- haplo.em.fitter(n.loci, n.subject, weight, geno.vec,
        n.alleles, max.haps, max.iter = control$max.iter,
loci.insert.order,
        min.posterior = control$min.posterior, tol = control$tol,
        insert.batch.size = control$insert.batch.size, random.start =
control$random.start,
        iseed1 = iseed1, iseed2 = iseed2, iseed3 = iseed3, verbose =
control$verbose)
    if (control$n.try > 1) {
        for (i in 2:control$n.try) {
            seed.array <- runif(3)
            iseed1 = 10000 + 20000 * seed.array[1]
            iseed2 = 10000 + 20000 * seed.array[2]
            iseed3 = 10000 + 20000 * seed.array[3]
            fit.new <- haplo.em.fitter(n.loci, n.subject, weight,
                geno.vec, n.alleles, max.haps, max.iter =
control$max.iter,
                loci.insert.order, min.posterior =
control$min.posterior,
                tol = control$tol, insert.batch.size =
control$insert.batch.size,
                random.start = 1, iseed1 = iseed1, iseed2 = iseed2,
                iseed3 = iseed3, verbose = control$verbose)
            if (fit.new$tmp1$lnlike > fit$tmp1$lnlike) {
                fit <- fit.new
            }
        }
    }

    tmp1 <- fit$tmp1
    tmp2 <- fit$tmp2
    u.hap <- matrix(tmp2$u.hap, nrow = tmp2$n.u.hap, byrow = TRUE)
    haplotype <- data.frame(I(allele.labels[[1]][u.hap[, 1]]))
```

```
  for (j in 2:n.loci) {
      haplotype <- cbind(haplotype, I(allele.labels[[j]][u.hap[,
          j]]))
  }

  names(haplotype) <- locus.label
  hap1code <- tmp2$hap1code + 1
  hap2code <- tmp2$hap2code + 1
  uhapcode <- tmp2$u.hap.code + 1
  n1 <- length(uhapcode)
  n2 <- length(hap1code)
  tmp <- as.numeric(factor(c(uhapcode, hap1code, hap2code)))
  uhapcode <- tmp[1:n1]
  hap1code <- tmp[(n1 + 1):(n1 + n2)]
  hap2code <- tmp[(n1 + n2 + 1):(n1 + 2 * n2)]
  uhap.df <- data.frame(uhapcode, tmp2$hap.prob, u.hap)
  names(uhap.df) <- c("hap.code", "hap.prob", locus.label)
  indx.subj = tmp2$indx.subj + 1

 if (length(unique(tmp2$indx.subj)) < n.subject) {
      unique.subj <- unique(indx.subj)
      rows.rem <- c(rows.rem, which(is.na(match(1:n.subject,
          unique.subj))))
      warning("Subject(s) ", paste(rows.rem, sep = ","), " removed in
trimming steps.\n Try decreasing min.posterior control parameter to
reduce trimming.\n")
  }

  subj.used.id <- subj.id[indx.subj]
  hap.prob.noLD <- a.freq[[1]]$p[u.hap[, 1]]
  df.noLD <- length(a.freq[[1]]$p) - 1

  for (j in 2:n.loci) {
      hap.prob.noLD <- hap.prob.noLD * a.freq[[j]]$p[u.hap[,
          j]]
      df.noLD <- df.noLD + length(a.freq[[j]]$p) - 1
  }

  hap.prob.noLD <- hap.prob.noLD/sum(hap.prob.noLD)
  prior.noLD <- hap.prob.noLD[hap1code] * hap.prob.noLD[hap2code]
  prior.noLD <- ifelse(hap1code != hap2code, 2 * prior.noLD,
      prior.noLD)
  ppheno.noLD <- tapply(prior.noLD, indx.subj, sum)
  lnlike.noLD <- sum(log(ppheno.noLD))
  lr = 2 * (tmp1$lnlike - lnlike.noLD)
  df.LD <- sum(tmp2$hap.prob > 1e-07) - 1
  df.lr <- df.LD - df.noLD

  obj <- list(lnlike = tmp1$lnlike, lr = lr, df.lr = df.lr,
      hap.prob = tmp2$hap.prob, hap.prob.noLD = hap.prob.noLD,
      converge = tmp1$converge, locus.label = locus.label,
      indx.subj = indx.subj, subj.id = subj.used.id, post =
tmp2$post,
        hap1code = hap1code, hap2code = hap2code, haplotype =
haplotype,
        nreps = table(indx.subj), rows.rem = rows.rem, max.pairs =
max.pairs,
        control = control)

    if (exists("is.R") && is.function(is.R) && is.R()) {
        class(obj) <- "haplo.em"
    }
    else {
        oldClass(obj) <- "haplo.em"
    }
    return(obj)
}

#---------#
#     Section 4:  em analysis and output packaging
#---------#

#---------#
# Section 4.1 - Hap_highest_prob
#---------#

win.size <- 3
num.haplo.tot <- NULL
haplo.array.final <- NULL

# Determine window and run em

for (s in seq(1, dim(geno.haplo.qc.removed)[2]-((win.size*2)-1), 2)) {
      em.geno <- geno.haplo.qc.removed[, s:(s+((win.size*2)-1))]
      em.out <- EM_algo.fun(em.geno)
      em.out.info <- as.data.frame(cbind(em.out$subj.id,
em.out$hap1code, em.out$hap2code, em.out$post))
      num.haplo <- dim(em.out$haplotype)[1]
      haplo.array.tot <- array(0, c(1, num.haplo))

# Pick all pairs for individual t

            for(t in 1:(dim(geno.haplo.qc.removed)[1])) {
                  indi <- em.out.info[em.out.info$V1==t, ]
                  haplo.array <- array(0, c(1,num.haplo))

# Pick haplotype pair j for individual t

                        for (j in 1:(dim(indi)[1])) {
                              indi.line <- indi[j, ]


# Pick haplotype k from pair j for individual t

                                    for (k in 2:3) {
                                          haplo <- indi.line[1, k]
# Determine Pr of haplotype k from pair j for individual t

                                                  for (i in 1:num.haplo)
{
                                                       if(i==haplo) {
                                                       haplo.array[1, i]
<- (haplo.array[1,i]+((indi[j,4]*0.5)))
                                                  }
                                        }
                              }
                    }
# Packed into the arrays

     haplo.array.tot <- rbind(haplo.array.tot, haplo.array)
     }

# Final array of the haplotype Pr

haplo.array.final <- cbind(haplo.array.final, haplo.array.tot)

# Final vector of number of colunms in haplo.array.final for each
window

num.haplo.tot <- rbind(num.haplo.tot, num.haplo)
}

#----------#
# Section 4.2 - Hap_all_prob
#----------#

#---------#
#     Section 4:  em analysis and output packaging
#---------#

win.size <- 3
num.haplo.tot <- NULL
haplo.array.final <- NULL

# Determine window and run em

for (s in seq(1, dim(geno.haplo.qc.removed)[2]-((win.size*2)-1), 2)) {
     em.geno <- geno.haplo.qc.removed[, s:(s+((win.size*2)-1))]
     em.out <- haplo.em(em.geno)
     em.out.info <- as.data.frame(cbind(em.out$subj.id,
em.out$hap1code, em.out$hap2code, em.out$post))
     num.haplo <- dim(em.out$haplotype)[1]
     haplo.array.tot <- array(0, c(1, num.haplo))

# Pick all pairs for individual t

          for(t in 1:(dim(geno.haplo.qc.removed)[1])) {
               indi <- em.out.info[em.out.info$V1==t, ]
               haplo.array <- array(0, c(1,num.haplo))
```

```
# Pick haplotype pair j for individual t

                        for (j in 1:(dim(indi)[1])) {
                              indi.line <- indi[j, ]


# Pick haplotype k from pair j for individual t

                                    for (k in 2:3) {
                                          haplo <- indi.line[1, k]

# Determine Pr of haplotype k from pair j for individual t

                                                for (i in 1:num.haplo)
{
                                                      if(i==haplo) {
                                                      haplo.array[1, i]
<- (haplo.array[1,i]+((indi[j,4]*0.5)))
                                                }
                                    }
                        }
            }
# Packed into the arrays

      haplo.array.tot <- rbind(haplo.array.tot, haplo.array)
      }

# Final array of the haplotype Pr

haplo.array.final <- cbind(haplo.array.final, haplo.array.tot)

# Final vector of number of colunms in haplo.array.final for each
window

num.haplo.tot <- rbind(num.haplo.tot, num.haplo)
}


#----------#
#     Section 5:  Apply a level row to the first row
#----------#

level.tot <- NULL

for (p in 1:dim(num.haplo.tot)[1]) {
      num.haplo <- num.haplo.tot[p,1]
      level <- rep(p, num.haplo)
      level.tot <- c(level.tot, level)
}
length(level.tot)==dim(haplo.array.final)[2]

haplo.array.final[1, ] <- level.tot
```

```
#---
# Read in the haplo.array.final
#---

haplo.array.final <- read.table("haplo.array.final.txt", header=F)
dim(haplo.array.final)

#----------#
#     Section 6: Function to calculate the Relationship matrix
#----------#


A_matrix  <-  function ( pedigree )

{

  nanim  <-  nrow(pedigree)

  if (nanim == 1)
  {
    new <- 1
  }
  else {
#----------#
#  calculating the inverse of A
#----------#

    new <- matrix(0,nrow=nanim,ncol=nanim)

    for (id in 1:nanim)
    {

      dad <- pedigree[id,1]
      mum <- pedigree[id,2]
#      a <- [id,dad,mum]
#      a
      if(dad < 0 || mum < 0 || dad > id || mum > id)
      {
        error ("Error problem with pedigree")
      }

      if(dad == 0 && mum == 0)
      {
#      both parents unknown
        new[id,id] <- 1
#         for (otherid in 1:id-1)
#         {
#           new[id     ,otherid] <- 0.0
#           new[otherid,id]      <- 0.0
#         }
      }
#----------#
#     sire known dam unknown
#----------#
```

```
      else if (dad > 0 && mum == 0) {
        new[id ,id ] <- 1
        cont_dam <- 0.0
        for (otherid in 1:id-1)
        {
          cont_sire <- new[dad,otherid]/2.0
          new[id     ,otherid] <- cont_sire+cont_dam
          new[otherid,id]      <- cont_sire+cont_dam
        }
      }
#----------#
#      sire unknown dam known
#----------#
      else if (dad == 0 && mum > 0) {
        new[id ,id ] <- 1
        cont_sire <- 0.0
        for (otherid in 1:id-1)
        {
          cont_dam <- new[mum,otherid]/2.0
          new[id     ,otherid]  <-  cont_sire+cont_dam
          new[otherid,id]       <-  cont_sire+cont_dam
        }

      }
#----------#
#     both parents known
#----------#
      else {
        new[id ,id ] <- 1+(new[dad,mum]/2.0)
        for (otherid in 1:id-1)
        {
          cont_sire <- new[dad,otherid]/2.0
          cont_dam  <- new[mum,otherid]/2.0
          new[id     ,otherid] <- cont_sire+cont_dam
          new[otherid,id]      <- cont_sire+cont_dam
        }
      }
#----------#
    }
  }
  new
}

#----------#
#     Section 7: Apply function, edit the output
#----------#

# Order by the pedigree

animal_in_pedigree.index <- which(pedigree$V4 %in% animal.id)
animal_in_pedigree <- pedigree[animal_in_pedigree.index, 4]

data_sorted.index <- order(animal_in_pedigree)
```

```
binary.geno.qc.removed.sorted <-
binary.geno.qc.removed[data_sorted.index, ]
genabel.data.qc.removed.sorted <-
genabel.data.qc.removed[data_sorted.index, ]

haplo.array.final.sorted <- haplo.array.final[2:200,
][data_sorted.index, ]
haplo.array.final.sorted <- rbind(haplo.array.final[1,],
haplo.array.final.sorted)

# A matrix forming

pedigree_A_matrix <- A_matrix(pedigree[,2:3])

animal_A_matrix <- pedigree_A_matrix[animal_in_pedigree.index,
animal_in_pedigree.index]
dim(animal_A_matrix)[1]==dim(binary.geno.qc.removed.sorted)[1] &
dim(animal_A_matrix)[2]==dim(binary.geno.qc.removed.sorted)[1]


# Calculate the inverse

animal_A_matrix_invert <- solve(animal_A_matrix)

#----------#
#     Section 8: Parameters to use in the analysis
#----------#

nrecords <- dim(binary.geno.qc.removed.sorted)[1]
lambda <- 0.00001

ones_mat <- array(1, c(nrecords))
z_mat <- diag(nrecords)
a_mat <- animal_A_matrix_invert

win.size <- 3
up.stream <- 25
down.stream <- 25
p.value.tot <- NULL
f.value.tot <- NULL
mean.within.ld.tot <- NULL
mean.between.ld.tot <- NULL
sQTL.cumulative.tot <- NULL
sQTL.chromosome.tot <- NULL
sQTL.mycode.tot <- NULL
window.start.cumulative.tot <- NULL
window.start.chromosome.tot <- NULL
window.start.mycode.tot <- NULL
window.distance.tot <- NULL
window.to.sQTL.distance.tot <- NULL
sQTL.maf.tot <- NULL

x.size.tot <- NULL
```

```
#----------#
#     Section 9:  Analysis
#----------#

for (y in
(down.stream+1):((length(unique(as.numeric(haplo.array.final.sorted[1,]
)))-up.stream))) {
      y_mat <- as.matrix(binary.geno.qc.removed.sorted[,y])

# determine test space about the sQTL

      level.index <- c((y-(down.stream-1)):(y+up.stream))
      test.space.index <- (haplo.array.final.sorted[1,] %in%
level.index)
      haplo.data.region <- haplo.array.final.sorted[, test.space.index]

# For a given sQTL run model across the test space

            for (k in level.index) {
                  haplo.index <- (haplo.data.region[1,] %in% k)
                  haplo.data <- haplo.data.region[,haplo.index]
                  x_mat <- as.matrix(haplo.data[2:(dim(haplo.data)[1]),
])

# Weed out singularities

                  fit <- summary(lm(y_mat~x_mat))
                  singular.index <-
as.matrix(fit$aliased[2:(dim(x_mat)[2]+1)])
                  singular.index <- singular.index==F
                  x_mat <- as.matrix(x_mat[ ,singular.index])

                  x.size <- dim(x_mat)[2]

                  coeff <- array(0, c((nrecords+2+(x.size-1)),
(nrecords+2+(x.size-1))))
                  rhs <- array(0, c((nrecords+2+(x.size-1))))

# Fill coefficient matrix

                  coeff[1:1, 1:1] <- t(ones_mat)%*%ones_mat
                  coeff[1:1, 2:(2+(x.size-1))] <- t(ones_mat)%*%x_mat
                  coeff[1:1, ((2+(x.size-1))+1):(nrecords+2+(x.size-
1))] <- t(ones_mat)%*%z_mat
                  coeff[2:(x.size+1), 1:1] <- t(x_mat)%*%ones_mat
                  coeff[2:(x.size+1), 2:(x.size+1)] <- t(x_mat)%*%x_mat
                  coeff[2:(x.size+1), (x.size+2):(nrecords+2+(x.size-
1))] <- t(x_mat)%*%z_mat
                  coeff[(x.size+2):(nrecords+2+(x.size-1)), 1:1] <-
t(z_mat)%*%ones_mat
                  coeff[(x.size+2):(nrecords+2+(x.size-1)),
2:(x.size+1)] <- t(z_mat)%*%x_mat
```

```
                  coeff[(x.size+2):(nrecords+2+(x.size-1)),
(x.size+2):(nrecords+2+(x.size-1))] <-
(t(z_mat)%*%z_mat)+(a_mat*lambda)

# Fill rhs

                  rhs[1:1] <- t(ones_mat)%*%y_mat
                  rhs[2:(x.size+1)] <- t(x_mat)%*%y_mat
                  rhs[(x.size+2):(nrecords+2+(x.size-1))] <-
t(z_mat)%*%y_mat

# Solve equation

                  solution_vec <- solve(coeff, rhs)
                  mu_hat <- solution_vec[1]

# Get test statistics
                        if(x.size==1) {
                        g_hats <- solution_vec[2]
                        df1 <- dim(x_mat)[2]
                        df2 <- dim(x_mat)[1]-(df1+1)
                        y_mean <- mean(y_mat)
                        y_hat <- mu_hat + (g_hats[1] * x_mat[,1])
                        ssq1 <- sum((y_hat-y_mean)^2)
                        ssq2 <- sum((y_mat-y_hat)^2)

                        msq1 <- ssq1/df1
                        msq2 <- ssq2/df2
                        f_value <- msq1/msq2
                        p_value <- 1-pf(f_value, df1, df2)
                        }

                        if(x.size==2) {
                        g_hats <- solution_vec[2:3]
                        df1 <- dim(x_mat)[2]
                        df2 <- dim(x_mat)[1]-(df1+1)
                        y_mean <- mean(y_mat)
                        y_hat <- mu_hat + (g_hats[1] * x_mat[,1] +
(g_hats[2] * x_mat[,2]))
                        ssq1 <- sum((y_hat-y_mean)^2)
                        ssq2 <- sum((y_mat-y_hat)^2)

                        msq1 <- ssq1/df1
                        msq2 <- ssq2/df2
                        f_value <- msq1/msq2
                        p_value <- 1-pf(f_value, df1, df2)
                        }


                        if(x.size==3) {
                        g_hats <- solution_vec[2:4]
                        df1 <- dim(x_mat)[2]
                        df2 <- dim(x_mat)[1]-(df1+1)
                        y_mean <- mean(y_mat)
                        y_hat <- mu_hat + (g_hats[1] * x_mat[,1] +
(g_hats[2] * x_mat[,2]) + (g_hats[3] * x_mat[,3]))
                        ssq1 <- sum((y_hat-y_mean)^2)
                        ssq2 <- sum((y_mat-y_hat)^2)

                        msq1 <- ssq1/df1
                        msq2 <- ssq2/df2
                        f_value <- msq1/msq2
                        p_value <- 1-pf(f_value, df1, df2)
                        }

                        if(x.size==4) {
                        g_hats <- solution_vec[2:5]
                        df1 <- dim(x_mat)[2]
                        df2 <- dim(x_mat)[1]-(df1+1)
                        y_mean <- mean(y_mat)
                        y_hat <- mu_hat + (g_hats[1] * x_mat[,1] +
(g_hats[2] * x_mat[,2]) + (g_hats[3] * x_mat[,3]) + (g_hats[4] *
x_mat[,4]))
                        ssq1 <- sum((y_hat-y_mean)^2)
                        ssq2 <- sum((y_mat-y_hat)^2)

                        msq1 <- ssq1/df1
                        msq2 <- ssq2/df2
                        f_value <- msq1/msq2
                        p_value <- 1-pf(f_value, df1, df2)
                        }

                        if(x.size==5) {
                        g_hats <- solution_vec[2:6]
                        df1 <- dim(x_mat)[2]
                        df2 <- dim(x_mat)[1]-(df1+1)
                        y_mean <- mean(y_mat)
                        y_hat <- mu_hat + (g_hats[1] * x_mat[,1] +
(g_hats[2] * x_mat[,2]) + (g_hats[3] * x_mat[,3]) + (g_hats[4] *
x_mat[,4]) + (g_hats[5] * x_mat[,5]))
                        ssq1 <- sum((y_hat-y_mean)^2)
                        ssq2 <- sum((y_mat-y_hat)^2)

                        msq1 <- ssq1/df1
                        msq2 <- ssq2/df2
                        f_value <- msq1/msq2
                        p_value <- 1-pf(f_value, df1, df2)
                        }

                        if(x.size==6) {
                        g_hats <- solution_vec[2:7]
                        df1 <- dim(x_mat)[2]
                        df2 <- dim(x_mat)[1]-(df1+1)
                        y_mean <- mean(y_mat)
                        y_hat <- mu_hat + (g_hats[1] * x_mat[,1] +
(g_hats[2] * x_mat[,2]) + (g_hats[3] * x_mat[,3]) + (g_hats[4] *
x_mat[,4]) + (g_hats[5] * x_mat[,5]) + (g_hats[6] * x_mat[,6]))
                        ssq1 <- sum((y_hat-y_mean)^2)
                        ssq2 <- sum((y_mat-y_hat)^2)

                        msq1 <- ssq1/df1
                        msq2 <- ssq2/df2
                        f_value <- msq1/msq2
                        p_value <- 1-pf(f_value, df1, df2)
                        }

                        if(x.size==7) {
                        g_hats <- solution_vec[2:8]
                        df1 <- dim(x_mat)[2]
                        df2 <- dim(x_mat)[1]-(df1+1)
                        y_mean <- mean(y_mat)
                        y_hat <- mu_hat + (g_hats[1] * x_mat[,1] +
(g_hats[2] * x_mat[,2]) + (g_hats[3] * x_mat[,3]) + (g_hats[4] *
x_mat[,4]) + (g_hats[5] * x_mat[,5]) + (g_hats[6] * x_mat[,6]) +
(g_hats[7] * x_mat[,7]))
                        ssq1 <- sum((y_hat-y_mean)^2)
                        ssq2 <- sum((y_mat-y_hat)^2)

                        msq1 <- ssq1/df1
                        msq2 <- ssq2/df2
                        f_value <- msq1/msq2
                        p_value <- 1-pf(f_value, df1, df2)
                        }

                        if(x.size==8) {
                        g_hats <- solution_vec[2:9]
                        df1 <- dim(x_mat)[2]
                        df2 <- dim(x_mat)[1]-(df1+1)
                        y_mean <- mean(y_mat)
                        y_hat <- mu_hat + (g_hats[1] * x_mat[,1] +
(g_hats[2] * x_mat[,2]) + (g_hats[3] * x_mat[,3]) + (g_hats[4] *
x_mat[,4]) + (g_hats[5] * x_mat[,5]) + (g_hats[6] * x_mat[,6]) +
(g_hats[7] * x_mat[,7]) + (g_hats[8] * x_mat[,8]))
                        ssq1 <- sum((y_hat-y_mean)^2)
                        ssq2 <- sum((y_mat-y_hat)^2)

                        msq1 <- ssq1/df1
                        msq2 <- ssq2/df2
                        f_value <- msq1/msq2
                        p_value <- 1-pf(f_value, df1, df2)
                        }

                  p.value.tot <- rbind(p.value.tot, p_value)
                  f.value.tot <- rbind(f.value.tot, f_value)

# Take LD information

                  window.start <- k
                  window.geno <- genabel.data.qc.removed[,
c(window.start:(window.start+2))]
                  within.ld <- r2fast(window.geno)
                  mean.within.ld <- mean(within.ld[c(4,7,8)])
                  between.geno <- genabel.data.qc.removed[,
c((window.start:(window.start+2)), y)]
                  between.ld <- r2fast(between.geno)
                  mean.between.ld <- mean(between.ld[c(13,14,15)])

                  mean.within.ld.tot <- rbind(mean.within.ld.tot,
mean.within.ld)
                  mean.between.ld.tot <- rbind(mean.between.ld.tot,
mean.between.ld)

# Positon and distance info

                  window.start.cumulative <-
maf.info.qc.removed[window.start, 11]
                  window.start.chromosome <-
maf.info.qc.removed[window.start, 3]
                  window.start.mycode <-
maf.info.qc.removed[window.start, 1]

                  window.finish <- k+2
                  window.distance <- maf.info.qc.removed[window.finish,
11]-window.start.cumulative

                  window.to.sQTL.distance <- maf.info.qc.removed[y,
11]-window.start.cumulative

                  window.start.cumulative.tot <-
rbind(window.start.cumulative.tot, window.start.cumulative)
                  window.start.chromosome.tot <-
rbind(window.start.chromosome.tot, window.start.chromosome)
                  window.start.mycode.tot <-
rbind(window.start.mycode.tot, window.start.mycode)

                  window.distance.tot <- rbind(window.distance.tot,
window.distance)
                  window.to.sQTL.distance.tot <-
rbind(window.to.sQTL.distance.tot, window.to.sQTL.distance)

                  x.size.tot <- rbind(x.size.tot, x.size)

                  }

# sQTL information

            sQTL.maf <- maf.info.qc.removed[y, 10]
            sQTL.cumulative <- maf.info.qc.removed[y, 11]
            sQTL.chromosome <- maf.info.qc.removed[y, 3]
            sQTL.mycode <- maf.info.qc.removed[y, 1]

            sQTL.cumulative.tot <- rbind(sQTL.cumulative.tot,
sQTL.cumulative)
            sQTL.chromosome.tot <- rbind(sQTL.chromosome.tot,
sQTL.chromosome)
```

```
            sQTL.mycode.tot <- rbind(sQTL.mycode.tot, sQTL.mycode)
            sQTL.maf.tot <- rbind(sQTL.maf.tot, sQTL.maf)

            print(y)

}


#----------#
#     Section 10: Turn outputs into matrices
#----------#

p.value.tot <- matrix(p.value.tot, nrow=length(level.index),
ncol=length(unique(as.numeric(haplo.array.final.sorted[1,])))-
(up.stream+down.stream), byrow=F)
f.value.tot <- matrix(f.value.tot, nrow=length(level.index),
ncol=length(unique(as.numeric(haplo.array.final.sorted[1,])))-
(up.stream+down.stream), byrow=F)

mean.within.ld.tot <- matrix(mean.within.ld.tot,
nrow=length(level.index),
ncol=length(unique(as.numeric(haplo.array.final.sorted[1,])))-
(up.stream+down.stream), byrow=F)
mean.between.ld.tot <- matrix(mean.between.ld.tot,
nrow=length(level.index),
ncol=length(unique(as.numeric(haplo.array.final.sorted[1,])))-
(up.stream+down.stream), byrow=F)

window.start.cumulative.tot <- matrix(window.start.cumulative.tot,
nrow=length(level.index),
ncol=length(unique(as.numeric(haplo.array.final.sorted[1,])))-
(up.stream+down.stream), byrow=F)
window.start.chromosome.tot <- matrix(window.start.chromosome.tot,
nrow=length(level.index),
ncol=length(unique(as.numeric(haplo.array.final.sorted[1,])))-
(up.stream+down.stream), byrow=F)
window.start.mycode.tot <- matrix(window.start.mycode.tot,
nrow=length(level.index),
ncol=length(unique(as.numeric(haplo.array.final.sorted[1,])))-
(up.stream+down.stream), byrow=F)

window.to.sQTL.distance.tot <- matrix(window.to.sQTL.distance.tot,
nrow=length(level.index),
ncol=length(unique(as.numeric(haplo.array.final.sorted[1,])))-
(up.stream+down.stream), byrow=F)
window.distance.tot <- matrix(window.distance.tot,
nrow=length(level.index),
ncol=length(unique(as.numeric(haplo.array.final.sorted[1,])))-
(up.stream+down.stream), byrow=F)

sQTL.details <- as.data.frame(cbind(sQTL.mycode.tot,
sQTL.chromosome.tot, sQTL.cumulative.tot, sQTL.maf.tot))
names(sQTL.details)[1] <- "sQTL.mySNPcode"
names(sQTL.details)[2] <- "sQTL.chromosome"
names(sQTL.details)[3] <- "sQTL.cumulative"
names(sQTL.details)[4] <- "sQTL.maf"

#----------#
#     Section 11: Save output files
#----------#

write.csv(sQTL.details, "sQTL.details.haplo1.pedigree.csv", quote=F,
row.names=F)
write.table(window.start.cumulative.tot,
"window.start.cumulative.tot.haplo1.pedigree.txt", quote=F,
row.names=F, sep="\t", col.names=F)
write.table(window.start.chromosome.tot,
"window.start.chromosome.tot.haplo1.pedigree.txt", quote=F,
row.names=F, sep="\t", col.names=F)
write.table(window.start.mycode.tot,
"window.start.mycode.tot.haplo1.pedigree.txt", quote=F, row.names=F,
sep="\t", col.names=F)
write.table(p.value.tot, "p.value.tot.haplo1.pedigree.txt", quote=F,
row.names=F, sep="\t", col.names=F)
write.table(f.value.tot, "f.value.tot.haplo1.pedigree.txt", quote=F,
row.names=F, sep="\t", col.names=F)
write.table(window.distance.tot,
"window.distance.tot.haplo1.pedigree.txt", quote=F, row.names=F,
sep="\t", col.names=F)
write.table(mean.within.ld.tot,
"mean.within.ld.tot.haplo1.pedigree.txt", quote=F, row.names=F,
sep="\t", col.names=F)
write.table(mean.between.ld.tot,
"mean.between.ld.tot.haplo1.pedigree.txt", quote=F, row.names=F,
sep="\t", col.names=F)
write.table(window.to.sQTL.distance.tot,
"window.to.sQTL.distance.tot.haplo1.pedigree.txt", quote=F,
row.names=F, sep="\t", col.names=F)
```