POPULATION-WIDE LINKAGE DISEQUILIBRIUM AND ITS USES IN QTL MAPPING AND ESTIMATION OF ANCESTRAL POPULATION SIZE

Albert Tenesa-Prunyonosa

Thesis submitted for the degree
of
Doctor of Philosophy

The University of Edinburgh

2003

ABSTRACT

Trait loci mapping methods have traditionally been performed by following the co-segregation of marker loci and trait values within families (i.e. linkage methods). However, these methods have poor resolution and power. Alternative methods based on linkage disequilibrium (LD) at the population level have been advocated to overcome these limitations. The feasibility of LD mapping methods relies on population parameters such as allele frequencies at the trait and marker loci and the extent of LD. The extent of LD was studied in two populations, a dairy cattle population from the United Kingdom (UK) and a human isolated Sardinian population. For the dairy cattle population, data from 50 young bulls were available. These bulls were typed at 6 markers on chromosome 2 and 7 markers on chromosome 6, spanning 38 and 20 cM, respectively. Two different methods, that do not require family information, were used to estimate population haplotype frequencies. LD extended to about 10 cM between pairs of loci in syntenic groups. Given the observed level of LD, mapping methods based upon population-wide association might provide better resolution than linkage methods in the UK dairy cattle population, as well as reduce the required sample sizes of the experiments. For the human population, 381 individuals typed at 22 markers on chromosome 19 were studied. High levels of disequilibrium were found that extended to 8 cM, when based on the LD measure D', and 11 cM when based on the significance level of the allelic association. It was also shown, using bootstrapping, that small sample sizes can overestimate both the mean value of D' and its variance by up to factors of about 3 and 23, respectively, when the sample size decreases from 381 to 25 individuals. Due to the high sampling variance of LD measures, the use of at least 200 unrelated individuals when characterizing the extent of LD is recommended.

Three different strategies and study designs to map quantitative trait loci (QTL) using LD were studied using analytical methods and computer simulation. First, a strategy that involved phenotyping a large number of unrelated individuals and genotyping only selected individuals from the two tails of the trait distribution was considered. Power to detect trait-marker association was derived as a function of the number of QTL and marker alleles. Two patterns of LD were used to assess their influence on power. When the frequency of the QTL allele with the largest effect and that of the marker allele linked in coupling were equal, power was maximum. In this case, increasing the number of QTL alleles reduced the power. The maximum difference in power between the two LD patterns studied was about 30%. For low QTL heritabilities and single trait studies, selecting around 5% of both tails of the trait distribution is recommended. Secondly, two approaches for mapping QTL using LD were compared. In the trait-based (TB) approach, the frequencies of

marker alleles (or genotypes) are compared in individuals selected from the two tails of the trait distribution. In the marker-based (MB) approach, the quantitative trait values for the marker genotypes in the selected individuals are compared. The power of each approach was quantified. It was shown that the power of the MB approach was greater than or equal to that of the TB approach. The advantage of the former is expected to increase with increasing number of traits phenotyped. Thirdly, a design based on collecting concordant sib-pairs for high and low phenotypic values and comparing the allele frequency distribution in both groups was considered. Although the method described was generally less powerful than a regression approach using just one of the sibs in each pair, the collection strategy proposed might still be justified when designing a QTL mapping experiment, because the collected samples would be used in a preliminary linkage analysis followed by a LD study.

Finally, published data from human chromosomes 22 and 19 was used to infer past effective population size in a population of European ancestry. To do so, the extent of LD was first estimated using a multilocus measure of LD, the chromosome segment homozygosity (CSH). Results suggest that this population has had an average effective population size of around 4500 breeding individuals for approximately the last 4500 generations. This population had a relatively constant size (of between 3000 and 5000 individuals) from about 130000 years ago to about 2000 years ago, when it expanded to more than 10000 individuals.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisors Peter Visscher, Sara Knott and Andrew Carothers for their careful supervision and sharing of so many ideas.

I would also like to thank the MRC-Human Genetics Unit for funding me during the duration of my PhD studies and Alan Wright and Nicholas Hastie for making it possible.

This work would not have been possible without the work of many people. David Ward, Deborah Smith and John Williams did the bulls' genotyping and Susan Brotherstone kindly provided the pedigree information for the bulls. Caroline Hayward, Susan Campbell and Isla Campbell did the genotyping of the Talana data. Mario Pirastu's group shared their Talana data. Formulas shown in Chapter 6 were adapted from previous derivations performed by Andrew Carothers. Lon Cardon and Robert Lawrence provided the data for human chromosome 19 analysed in Chapter 7 and Ben Hayes his C++ program for estimating chromosome segment homozygosity.

Ian White also helped me in numerous occasions and probably still will. I thank him for his help and huge patience. I also thank him and the rest of Waverley users (i.e. Sue Brotherstone and Pau Navarro) for letting me over-exploit the computing facilities in times of crisis.

To all those friends (Ana, Valentí, Ximo, Estel, Karla, Miquel, Jorge, Néstor, etc.) that visited Scotland loaded with Spanish goodies as tokens for an uncomfortable bed from a well-known Swedish company, many thanks. Many thanks as well to Beatriz, Marie and Geoff for their friendship and help (note that they also brought us Spanish goodies whenever they went to Spain).

I want to thank my family (parents, parents in law and sister) for their unconditional support and love. Finally and very specially, I would like to express my love and gratitude to Pau who has always been there for me. For all of them, one of those huge group-hugs that Pau loves so much.

PUBLICATIONS

The following publications have resulted from the research described in Chapters 2 and 4 of this thesis:

Tenesa, A., Knott, S. A., Ward, D., Smith, D., Williams, J. L., and Visscher, P. M. 2003. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. J. Anim. Sci. 83: 617-623.

Tenesa, A., Knott, S. A., Carothers, A. D., and Visscher, P. M. 2003. Power of linkage disequilibrium mapping to detect a quantitative trait locus (QTL) in selected samples of unrelated individuals. Ann. Hum. Genet. In press.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Adlo	Adjacent loci in highly significant LD
ANOVA	Analysis of variance
APOE	Apolipoprotein E gene
CEPH	Centre d'Etudes du Polymorphisme Humain
CI	Confidence interval
cM	Centimorgan
CSH	Chromosome segment homozygosity
dbSNP	Public single nucleotide polymorphisms data base
df	Degrees of freedom
DL	Disease locus
DNA	Deoxyribose nucleic acid
EH	Expected heterozygosity
EM	Expectation-maximization
G1	Individuals with children recorded
G2	Individuals with grandchildren recorded
G3	Individuals with great-grandchildren recorded
HWE	Hardy-Weinberg equilibrium
IBD	Identical by descent
IBS	Identical by state
LD	Linkage disequilibrium
LE	Linkage equilibrium
LS	Least squares
MARC97	Linkage map created by the United States of America Meat Animal Research Center
MB	Marker-based
MCMC	Markov-Chain Monte Carlo
ML	Maximum-likelihood
NA	Number of alleles
OH	Observed heterozygosity
QT	Quantitative trait
QTL	Quantitative trait loci
RFLP	Restriction fragment length polymorphism
SD	Standard deviation
SE	Standard error

SNP	Single nucleotide polymorphism
SP	LD mapping method based on sib-pairs described in Chapter 6
STRP	Short tandem repeat poylmorphism
TB	Trait-based
TDT	Transmission disequilibrium test
UK	United Kingdom
VC	Variance component

CHAPTER 1 - GENERAL INTRODUCTION

Traditionally the genetic basis of continuous variation has been described as the cumulative effect of a large (infinite) number of genes with infinitesimal and equal effects under a model known as the "infinitesimal model". Although the model has been widely used and has been shown to perform well under some scenarios (e.g., in plant and animal breeding schemes), it is based on unrealistic assumptions (Falconer and Mackay, 1996) and there are benefits to the use of more realistic models. The use of more realistic genetic models, involving a finite number of genes of different effect, would help us to understand and exploit more efficiently phenotypic and genetic variation. Models including major genes or quantitative trait loci (QTL) (that is genes or portions of the genome that explain a significant part of the genetic variation underlying the traits of interest) as well as polygenic effects (aggregate effect of genes of small effect) are of interest to geneticists for different reasons. For example, population geneticists are interested in knowing how genetic variation is maintained in natural populations and in determining whether loci responsible for variation within a population are the same as those that cause divergence between populations and species (Mackay, 2001). Breeders are interested in applying these models to speed up response to artificial selection through marker-assisted selection and to introgress genes of economical importance through marker-assisted introgression (Haley and Visscher, 1998). Human geneticists are interested in knowing how genetic variation influences individuals' susceptibility to disease and response to drugs, and whether personalised therapies could be used (De la Chapelle and Wright, 1998; Roses, 2000).

The first step in understanding how loci interact and affect phenotypic variation is to identify and map them to the genomic region where the relevant polymorphism is located, the final goal being to identify the molecular variants that produce phenotypic differences between individuals using, for example, a positional cloning or a positional candidate gene approach (Strachan and Read, 1999; Darvasi, 1998; Hugot *et al.*, 2001).

The widespread use of DNA polymorphisms as genetic markers has been possible only in the last two decades. Paterson *et al.* (1988) constructed the first complete linkage map of restriction fragment length polymorphisms (RFLPs) and mapped six QTL in a backcross of tomatoes. Since then, comprehensive linkage maps have been constructed for humans, economically important plant and animal species, and model organisms such as Drosophila (Lynch and Walsh, 1998). These maps have allowed linkage studies to be widely used to map loci affecting a wide number of characters. Such characters include binary or dichotomous traits (e.g. affected/unaffected) and continuous or quantitative traits (those that

can be measured on a continuous scale). Now drafts of the complete DNA sequence are available for a number of model organisms, such as the mouse (Waterston *et al.*, 2002), and humans (Lander *et al.*, 2001). This will greatly facilitate mapping the loci underlying complex traits.

In the following, an overview of several mapping methods used to date is presented, starting with linkage methods and continuing with methods that exploit population-wide linkage disequilibrium.

1.1 OVERVIEW OF LINKAGE MAPPING METHODS

Linkage methods use family information to trace the segregation of marker alleles, test for marker-trait association and estimate the strength of the association. The latter is usually done within a maximum likelihood (ML) or least squares (LS) framework. The strength of the association will depend on the QTL effect and the recombination fraction between marker locus and QTL. Testing one marker at a time has certain limitations (basically, the locus effect and position are confounded), and methods that use information on multiple markers have been proposed (Kruglyak and Lander, 1995; Fulker *et al.*, 1995). Interval mapping uses the individuals' genotype at two flanking markers and the genetic distance between those markers to infer the putative QTL genotype at different locations within the interval spanned by these markers. Thus, locus effect and position can be independently estimated. Although, theoretically QTL effect and position can be separated using ML and single markers (Haseman and Elston, 1972), the estimates of the QTL effect and position obtained are poorer than those from interval mapping. Interval mapping has been used to map QTL in structured populations both within a ML (Lander and Botstein, 1989) and LS (Haley and Knott, 1992) framework. Interval mapping has also been proposed to map QTL in complex pedigrees using ML (Goldgar, 1990; Xu and Atchley, 1995) and regression methods based on LS (Fulker and Cardon, 1994). For outbred populations, a further improvement over interval mapping, multipoint interval mapping, uses information on all available marker loci to infer the putative QTL genotype at given positions on the genome (Kruglyak and Lander, 1995; Fulker *et al.*, 1995). Multipoint interval mapping has more power to detect the QTL than interval mapping and gives a more accurate estimate of the QTL position, especially when there is incomplete marker information. The multipoint interval mapping approach proposed by Kruglyak and Lander (1995) is based on a Hidden Markov Model to infer the probability distribution of the identity by decent (IBD) status at each position in the genome, and ML for estimation of genetic parameters. The multipoint interval mapping approach proposed by Fulker *et al.* (1995) uses multiple regression to

estimate the expected proportion of alleles shared IBD at any position in the genome, and regression for estimation of the QTL position and effect. Fulker and Cherny (1996) showed, using ML, that there were no substantial differences in power when using the probability distribution of the IBD status and when using its expected value. These authors concluded that the use of expected IBD probabilities is more desirable than the full IBD probability distribution because it is easier and faster to implement. If the trait is affected by multiple QTL, one can fit those already found as cofactors and increase the power to detect the remaining QTL (Zeng, 1993; Jansen, 1993).

There are a number of ML and regression methods for estimating QTL effect and position in complex pedigrees. Most of them were initially developed for sib-pairs and later expanded to cope with other type of relatives. Kruglyak and Lander (1995) proposed a ML method based on phenotypic differences between sib-pairs. Under the null hypothesis of no QTL, the variance of the phenotypic difference conditional on the number of alleles shared IBD should be the same for sib-pairs that share 0, 1 or 2 alleles. Under the alternative hypothesis of a QTL at a given position, this variance should increase with decreasing number of alleles shared IBD. However, an alternative parameterisation of the likelihood, which allows for the bivariate structure of the data, has higher power and can readily extended to more complex pedigrees (Goldgar, 1990). This approach, usually known as variance-components (VC), has been extended to cope with large and complex pedigrees (Almasy and Blangero, 1998) and can accommodate more complex genetic models (e.g. epistasis, genotype-environment interactions, etc.). Like ML methods, regression methods were initially proposed for sib-pairs (Haseman and Elston, 1972). They are based on regressing some function of the phenotypes of each sib-pair on the proportion of IBD alleles they share. Regression models are usually solved by LS because it is simpler and faster. Faster computations make feasible the use of resampling techniques. Permutations (i.e. shuffling) of the data and reanalysis can be used to draw empirical distributions of the test statistics (Doerge and Churchill, 1996) and bootstrapping (sampling with replacement) can be used to estimate confidence intervals for the position of the QTL (Visscher *et al.*, 1996). Regression methods have recently been extended to cope with general pedigrees and shown to have similar power to those based on VC (Sham *et al.*, 2002).

For dichotomous characters (e.g. having or not having a disease), both parametric and non-parametric methods have been used to map disease loci (DL) in complex pedigrees (Ott, 1999). Parametric methods are usually more powerful than non-parametric ones if the correct model is assumed (Lander and Schork, 1994). However, they can produce inaccurate results if the model assumed is incorrect (Clerget-Darpoux *et al.*, 1986). Non-parametric

methods are therefore of interest when the genetic model is unknown or complex, e.g. in complex diseases, because they do not need a specified genetic model and are more robust. The basic principle behind non-parametric methods using sib-pair data is that if a marker locus is closely linked to a DL then affected (or unaffected) sib-pairs will share a larger proportion of IBD alleles than would be expected if the marker and DL were unlinked. Similar allele sharing methods to those for sib-pairs have been developed to deal with more complex pedigrees (Kruglyak *et al.*, 1996).

1.1.1 Limitations of linkage methods

Linkage methods have important limitations. First, they have poor resolution in the sense that the confidence intervals (CI) for the position of the QTL are large. One could narrow the CI down by increasing marker density but this strategy has proved quite inefficient (Darvasi and Soller, 1994). For example, Visscher *et al.* (1996) found similar empirical confidence intervals for a 10 cM and a 20 cM map in simulated backcrosses. In human pedigrees of moderate size (5-7 people), denser maps than about 1 cM do not provide any improvement in the location of the QTL (Atwood and Heard-Costa, 2003). This is because linkage methods are based on the co-segregation of marker and trait locus within a pedigree. Relatives share large IBD regions and increasing marker density cannot reduce confidence intervals because there are not enough recombinants. In order to get a higher resolution it would be necessary to identify more recombinant individuals by increasing the number of observable recombinations in the pedigrees under study (i.e. by increasing the number or/and the size of the families). This is difficult to achieve in human studies because pedigree size is limited by the number of relatives that are alive at the time of study. Increasing the number of families is also often ineffective, especially for complex traits where different loci might be segregating within different families. Second, while linkage methods have reasonable power (for affordable sample sizes) when the trait under study has a simple mode of inheritance and is controlled by only one or a small number of genes, it is not so when the trait is genetically complex and controlled by large numbers of loci of small effect. In the latter case, large sample sizes are required to detect loci underlying the trait of interest (Risch and Zhang, 1996; Zhang and Risch, 1996). There are several factors that limit the power of linkage studies when applied to complex traits. Genetic heterogeneity (i.e. multiple loci influencing the trait) effectively reduce the power of the study because some families might be segregating at one locus and others not (MacGregor *et al.*, 2002). In a similar way, high allele frequencies at the trait locus will hinder trait locus detection because a larger proportion of families will be homozygous at the locus. Finally, complex genetic mechanisms such as epistasis might reduce the linkage signal because a locus might be

expressed in a different way depending on the family's genetic background (Kajiwara *et al.*, 1994).

1.2 OVERVIEW OF LINKAGE DISEQUILIBRIUM MAPPING METHODS

An alternative approach to linkage mapping is known as linkage disequilibrium (LD) or association mapping. LD is defined as the non-random assortment of alleles (Falconer and Mackay, 1996) and can be generated by drift, selection, mutation or population stratification. The basic idea behind LD mapping is to treat the whole population as an extended family and make use of the many generations of recombination that have occurred since LD was generated to fine map the QTL or gene (Cardon and Bell, 2001). Although the resolution achievable using LD mapping would depend on the population under study a reasonable expected value would be about 0.001-0.5 cM (Rannala and Slatkin, 2000; Morris *et al.*, 2002), whereas for linkage mapping the resolution achievable in human families of moderate size (5-7 individuals/family and 200 families) for a locus with small effect would be about 10-15 cM (Atwood and Heard-Costa, 2003). LD mapping not only provides a higher resolution than linkage mapping but is also more powerful (Risch and Merikangas, 1996) because the QTL effects are usually expressed as means in a LD mapping framework, but as variances in a linkage mapping framework. Means are estimated with smaller standard errors than variances and therefore QTL effects expressed as means are easier to detect.

1.2.1 Extent of LD and feasibility of LD mapping methods

The efficiency of LD mapping depends on parameters such as the extent of LD, the allelic heterogeneity of the trait locus, the allele frequency difference between marker and trait loci and the heterogeneity of LD across the genome (Reich *et al.*, 2001). In recent years, a number of studies have assessed the extent and structure of LD in human populations. Studies, usually based on very dense maps and small numbers of haplotypes, have revealed that the human genome consists of patches of low LD interspersed by patches of high LD (Daly *et al.*, 2001; Jeffreys *et al.*, 2001; Patil *et al.*, 2001). This would affect the marker density distribution required for LD mapping. Although there is still much controversy about it, these patterns are consistent with recombination happening in localised regions of the genome with much higher frequency (recombination hot-spots) than on random regions of the genome (Jeffreys *et al.*, 2001; Gabriel *et al.*, 2002). Generalisation of these conclusions needs, nevertheless, more research because (i) only one study, to my knowledge, has unequivocally proven that regions where recombination occurs with higher frequency are those that correspond with low LD (Jeffreys *et al.*, 2001) and (ii) simulation studies have shown that it is possible to find a block-like structure when the data is simulated with

recombination events happening at random positions but with varying recombination rate across the genome (Wang *et al.*, 2002; Phillips *et al.*, 2003). Whether haplotype blocks are due to hot-spots or population history has important implications. For example, in the former case, block boundaries would be conserved across populations and in the latter case not. Although not studied in this thesis, the important implications for LD mapping of the genome block-like structure and its origins will be addressed in Chapter 8.

Whereas the extent of LD in human populations has been widely studied, the extent of LD in livestock populations has rarely been. I shall now summarise the level of LD in humans and livestock, but the reader must keep in mind that comparisons among studies with, for example, different sample sizes, densities and types of marker must be interpreted with care. In humans, estimates are very variable across populations and regions of the genome. Generally, African populations show markedly less LD than European or Asian populations. Huttley et al. (1999) found that about 4% of the locus pairs separated by less than 4 cM were in LD for two populations of European ancestry. Dunning et al. (2000) studied the extent of LD in three different regions of the genome in four populations of European origin and found that between 75% and 94% of the locus pairs showed significant LD for distances < 5 kb (~0.005 cM). More than 80% of the locus pairs were in significant LD within a region of ~10 cM in a Sardinian sub-isolate (Zavattari *et al.*, 2000) and in the Saami population (Laan and Paabo, 1997). Reich *et al.* (2001) found that the extent of LD was not longer than about 60-100 kb in Europeans and only about 5 kb in an African population. Similarly, Gabriel *et al.* (2002) found much larger LD in Europeans and Asians, where about half of the genome had blocks larger than 44 kb, than in African-Americans where about half of the genome had blocks larger than 22 kb. In livestock, Farnir *et al.* (2000) found that the Dutch black-and-white dairy cattle population showed significant LD over tens of centimorgans (D' averaged more than 0.33 for marker pairs less than 5 cM apart and was larger than 0.14 for pairs less than 50 cM apart). McRae *et al.* (2002) found similar results for two sheep populations, where about one third of the loci pairs separated by less than 60 cM showed significant LD.

In Chapters 2 and 3, the extent of LD in a cattle and a human population were studied. In Chapter 4, the influence on power to detect a QTL of allelic heterogeneity of the trait locus and of the allele frequency difference between marker and trait loci were studied.

1.2.2 LD mapping: designs and methods

LD or association study designs can be subdivided into those that collect only unrelated individuals and those that also collect family members to be used as controls. For dichotomous traits (e.g., disease status), the first category would involve sampling unrelated

cases (affected individuals) and unrelated controls (unaffected individuals) and comparing marker allele frequencies of the two groups. This design is usually known as a Case-Control design. The second category would involve sampling case individuals and some of their relatives that would be used as controls. The best known of the family-based association tests is the transmission disequilibrium test (TDT) (Spielman *et al.*, 1993), which involves sampling affected individuals and their parents and comparing the number of times a heterozygous parent transmits or does not transmit a marker allele to her/his affected offspring. When parental data is not available other relatives can be used instead (Martin *et al.*, 1997; Curtis, 1997). TDT was initially proposed to test for linkage under the assumption that there was population-wide LD. TDT is, strictly speaking, a test for linkage and LD at the same time, however if either association or linkage have been found by other methods, then its interpretation changes and it can be interpreted as a test for linkage or association, respectively. Family-based association tests, such as TDT, are also available for quantitative traits. Allison (1997) proposed five different ways of using this type of information.

When collecting unrelated individuals and measuring quantitative traits, the simplest test of association is to regress each individual's phenotype on the number of copies of a given allele. If the slope is significantly different from zero, then it is assumed that the marker locus is the QTL or it is in LD with it. One could account for non-additivity by using an analysis of variance (ANOVA) on the genotypes. Schork *et al.* (2000) proposed the use of threshold defined cases and controls (that is, comparing the allele frequencies of individuals selected for high and low trait values) for QTL mapping. This approach was followed in Chapter 4 to study the effect on power to detect a QTL of the number of alleles at the QTL and marker loci. Although it is relatively common among human geneticists to follow this dichotomising approach (e.g., when studying osteoporosis (Langdahl *et al.*, 2003)), it is shown in Chapter 5 that this is not entirely satisfactory unless the selection intensity is high.

Using unrelated controls has two major advantages over family-based controls: (i) it is generally more powerful (Bacanu *et al.*, 2000), and (ii) samples are easier to obtain, especially for late onset traits for which parental data might be impossible to obtain. In this case, one could still use other relatives as controls but this strategy is even less powerful than using parents as controls (Curtis, 1997). On the other hand, family-based association studies are robust to stratification (as discussed below) while case-control studies are not.

Case-control designs must be carefully designed to avoid detection of spurious associations due to population structure. To this aim cases and control samples must be matched with regard to ethnicity, sex, age, etc. Although these covariates can easily be accounted for if known, they pose an important problem when unknown. For example, one

could collect samples of Scottish origin in an Edinburgh hospital under the naïve impression that they are matched for ethnicity, but regional differences in allele frequencies between populations in the Isle of Lewis and, for example, Edinburgh have been shown (Vitart *et al.*, 2003). If these allele frequency differences were coupled with regional differences in disease prevalence, then an improperly matched sample could lead to spurious results. Fortunately, new methods to detect hidden population structure and account for it have been developed in the recent years (Devlin and Roeder, 1999; Pritchard *et al.*, 2000a; Pritchard *et al.*, 2000b). Pritchard and Rosenberg (1999) showed that if stratification is suspected, one could greatly reduce the risk of false positives by using as few as 15-20 unlinked microsatellite markers to test for it.

The case-control tests discussed so far have greater power when it can be assumed that identity by state (IBS) at the marker locus is equivalent to identity by descent (IBD) at the trait locus. However, this is not generally true. The assumption reduces the power to map the trait locus if the marker locus is not the trait locus itself or if the mutation or mutations of interest at the trait locus occurred more than once in the population. For instance, if a marker locus, in close linkage with the trait locus, is typed and the allele of interest at the trait locus appeared only once in the population through mutation (say, $Q_1 \rightarrow Q_2$) in a given chromosomal background containing allele M_2 at the marker locus, then IBS status at the marker locus does not necessarily mean IBD at the trait locus (because since the mutation occurred there have been M_2Q_1 and M_2Q_2 haplotypes in the population). If the same mutation or different mutations at the trait locus appeared at different times in the population history, then not all Q_2 alleles would be IBD and they would be on different chromosomal backgrounds; again IBS status at a linked marker would not be very informative about IBD status at the trait locus. Only if the marker locus typed is the trait locus itself and the allele appeared through mutation just once in history (say, $Q_1 \rightarrow Q_2$) then one could one safely say that IBS is equivalent to IBD.

There have been many methods for DL mapping proposed that have addressed this problem. In a similar way to linkage methods, they attempt to infer IBD status at a given genomic position from given marker information. However, in this case family information is not available and population genetics models have to be used to model the population history and the decay of LD in order to infer IBD status. One can then test if there is an excess of IBD sharing among individuals with the phenotype of interest compared to that among individuals with a different phenotype. The methods developed are varied and differ in many aspects. For instance, they differ in the number of markers used, the estimation procedure or whether the correlation among linked loci or disease haplotypes is accounted

for. Table 1.1 shows a succinct description of some of the methods proposed. As for linkage methods, inference about IBD status and trait locus position benefits from the use of information on more than one marker at a time. Methods that use information on a pair of markers were initially developed followed by methods that use information on a theoretically unlimited number of loci.

Table 1.1. Summary of the main features of LD mapping methods. ML and GLS stand for maximum likelihood and generalised least squares, respectively.

	Number of loci used	Estimation procedure	Accounts for loci correlation	Accounts for stochasticity of evolutionary process	Accounts for shared ancestry disease haplotype
Hastbacka *et al.* (1992)	1	Moments	No	No	No
Kaplan *et al.* (1995)	1-2	ML	Yes	Yes	Yes
Rannala and Slatkin (1998)	1	ML	No	Yes	Yes
Graham and Thompson (1998)	2	ML	Yes	Yes	Yes
Terwilliger (1995)	∞[1]	ML	No	No	No
Xiong and Guo (1997)	∞	ML	No	No	Yes
Collins and Morton (1998)	∞	ML	No	No	No
McPeek and Strahs (1999)	∞	ML	Yes	No	Yes
Lazzeroni (1998)	∞	GLS	Yes	No	No
Morris *et al.* (2000)	∞	Bayesian	Yes	No	Yes
Liu *et al.* (2001)	∞	Bayesian	Yes	No	Yes
Lam *et al.* (2000)	∞	Bayesian	Yes	No	Yes
Rannala and Reeve (2001)	∞	Bayesian	Yes	Yes	Yes
Morris *et al.* (2002)	∞	Bayesian	Yes	Yes	Yes

[1] Here ∞ means that the method can use information on a theoretically unlimited number of loci.

Hastbacka *et al.* (1992) were the first to apply non-equilibrium population genetics models to infer the recombination fraction between marker loci and the unknown DL. They adapted a model developed to infer mutation rates in rapidly (exponentially) growing populations of bacteria (Luria and Delbruck, 1943) to infer recombination frequencies in rapidly growing populations using linkage disequilibrium data. Kaplan *et al.* (1995) showed that the moment estimation (that is, based on equating the expected and observed value of the markers allele frequencies among disease chromosomes) proposed by Hastbacka *et al.* (1992) failed to account for the stochastic nature of the population history and produced unreliable, usually too low, upper bounds of the recombination fraction between individual markers and DL. In order to account for this, Kaplan *et al.* (1995) used forward simulation and Rannala and Slatkin (1998) used coalescent theory to draw multiple simulated samples of the population, compatible with the real sample, for hypothesized DL locations. Then, the probability of the real sample given each realisation of the simulations was computed and averaged over replicates for each trait locus position. The location with the highest likelihood was taken as the trait locus position. Graham and Thompson (1998) proposed a

method similar to that of Rannala and Slatkin (1998), which they extended to interval mapping, and also to cope with multiallelic markers. Terwilliger (1995) and Xiong and Guo (1997) proposed two different multipoint LD mapping methods based on combining single-marker likelihoods as if they were independent (i.e., they ignored the correlation between linked marker loci within a haplotype). The combined likelihood, usually know as the composite likelihood, is obtained simply by multiplying the likelihoods obtained for each of the linked marker loci under study. Moreover, neither the method of Terwilliger (1995) nor the first order approximation to the likelihood of Xiong and Guo (1997) account for the correlation between haplotypes due to population structure. They assume a star-shaped tree, in which each disease haplotype has had an independent history of recombination and mutation since the most recent common ancestor. The star-shaped tree assumption leads to the underestimation of the parameters' variance (for example, the variance of recombination fraction between marker and trait loci) and may produce inaccurate results (Rannala and Slatkin, 2000; Morris *et al.*, 2002). However, Xiong and Guo (1997) accounted for the evolutionary history of the population in their second order approximation to the likelihood, but they concluded that their first order approximation, which does not require modelling the population history, was favoured because it is simpler to model and gave similar empirical results. This suggests that the star-shaped tree might be a good approximation under some scenarios but not under others. Collins and Morton (1998) also proposed a composite likelihood method for the disequilibrium parameter ρ based on the Malecot model but it has the same weaknesses as the method of Terwilliger (1995), and the first order approximation of Xiong and Guo (1997). Lazzeroni (1998) showed that for a broad class of population genetics models the values of the disequilibrium parameters (δ_i) estimated from case-control data for i linked markers could be expressed as a piecewise curve along the chromosome, with the maximum of this curve at the DL position. The method has two steps. In the first step, the variance-covariance matrix that reflects the dependence among the i δ measures is obtained from bootstrapping the data. The bootstrap distribution is also used to check for normality of the data (δ_i). If the distribution of the data is not normal, then a transformation should be used to make it closer to normality. In the second step, the curve is fitted using generalised least squares and the variance-covariance matrix previously estimated. This procedure accounts for the covariance of the data conditional on the realised population, but not across possible realisations of the population. Multipoint LD mapping methods such as those developed by McPeek and Strahs (1999) and Morris *et al.* (2000) represented a substantial improvement over existing multipoint methods (Terwilliger, 1995; Xiong and Guo, 1997; Collins and Morton, 1998) because they accounted both for the population

structure and for the correlation across loci within individual haplotypes. McPeek and Strahs (1999) used a maximum likelihood framework whereas Morris *et al.* (2000) used Markov-chain Monte Carlo (MCMC) methods to draw posterior distributions of the model parameter estimates under a Bayesian framework. McPeek and Strahs (1999) and Morris *et al.* (2000) used a hidden Markov chain to model the ancestral chromosomal region around the disease locus (i.e., to account for correlation among loci). They initially assumed independent recombinational histories (i.e. a star-shaped tree) for each haplotype in the sample to construct the hidden Markov chain and subsequently corrected for this assumption being violated. McPeek and Strahs (1999) accounted for the dependence in the recombinational histories of the disease haplotypes by using a quasi-likelihood score function (Wedderburn, 1974), which is equivalent to the score function used in maximum-likelihood estimation when the data is correlated. The quasi-likelihood estimation process leads to inflated standard errors of the estimates (McPeek and Strahs, 1999). Morris *et al.* (2000) allowed for this dependence by down-weighting the contribution of each haplotype to the total likelihood, which increased the variance of the posterior distribution. It is worth noting here, that when a star-shaped tree is assumed the topology of the tree is fixed and the coalescent times for all the sampled disease haplotype to the most common ancestor is equal (all branch lengths are equal). Given that McPeek and Strahs (1999) and Morris *et al.* (2000) assumed a star-shaped structure of the coalescent tree they could not account for different realisations of the population. Liu *et al.* (2001) and Morris *et al.* (2002) developed multipoint LD mapping methods that accounted for multiple ancestral mutations at the DL and modelled the shared ancestry of the disease and control haplotypes. The main difference between the Liu *et al.* (2001) and Morris *et al.* (2002) approaches is in how they modelled the shared ancestry of the disease haplotypes. Liu *et al.* (2001) assumed a star-shaped genealogy but allowed for multiple ancestral founder haplotypes (clusters) and for different founder ages. In addition, they purged haplotypes that had multiple copies present in the data set (i.e. retained for analysis only a proportion of the haplotypes that had multiple copies in the original data set), which effectively reduces the weight of the haplotypes with shared ancestry in the estimation process. Morris *et al.* (2002) assumed a shattered coalescent model where the shared ancestry of the disease haplotypes is modelled explicitly. Finally, Meuwissen and Goddard (2001) proposed one of the few existing QTL LD mapping methods similar to those just described for DL LD mapping. Meuwissen and Goddard (2001) modelled the length of the chromosome that is inherited IBD by descendants from a common ancestor and proposed to use the estimated IBD probabilities in a variance-component framework. Because they envisaged applying the method for QTL LD mapping, they do not make implicit

assumptions about the number of QTL alleles segregating in the population nor assume that the QTL genotype can be inferred from the phenotype. Neither do they make any assumptions about the marker density, and allow for the model to estimate the probability that a part or parts of the haplotypes are IBD even though the QTL is not IBD.

1.2.3 Population choice

The choice of the population in mapping studies is an important factor, especially when the trait is complex. Different populations have advantages and disadvantages. For example, inferences drawn from populations with large effective population size, such as the UK population, might have greater generality that those obtained from small isolates. However, it might be easier to map genes (provided that enough phenotypes are available) in the latter because they have reduced genetic and environmental variance. Because of founder and drift effects, small isolated populations are expected to show lower levels of locus heterogeneity (a reduced number of segregating loci) and allele heterogeneity (smaller number of alleles at each locus) as well as high levels of LD. In Chapter 3, it is shown that a sub-isolate of the general Sardinian population (the village of Talana) has high levels of LD, and smaller allelic heterogeneity than the CEPH (Centre d'Etudes du Polymorphisme Humain) reference families (assumed to be representative of a population with large effective size). LD mapping methods might benefit (in terms of power to detect a trait locus) more than linkage methods from using populations with high levels of LD and small allelic heterogeneity, because they are much more dependent on them. Nevertheless, fine mapping would require populations with smaller levels of LD.

Domestic animal populations of economic interest such as that studied in Chapter 2 (a Holstein dairy cattle population) might benefit from LD mapping methods to fine map genes of interest. The Holstein dairy cattle population has been intensively selected for traits of economic interest and has experienced a remarkable reduction in effective population size since the advent of artificial insemination. Note that, although the global population size of the Holstein population is more than 25 millions, their estimated effective population size is only about 50 (Farnir *et al.*, 2000). Although in the Holstein dairy cattle population the extent of LD is large, the resolution achievable using LD methods would still be greater than that achieved using linkage methods. It is worth pointing out here, that the resolution of what would be considered fine mapping for human studies is usually smaller than in domestic animal populations.

1.3 ALTERNATIVE DESIGNS FOR MAPPING

Another important factor in mapping studies is the structure of the data collected, which will usually determine the mapping methods to be used subsequently. If one collected unrelated individuals, then one would unavoidably have to use LD mapping methods. If one collected large family cohorts, then linkage methods would need to be used to make the most of the data. One could also collect sib-pairs and their parents, because they are easier to collect than large family cohorts though more difficult than unrelated individuals, and use linkage methods. In Chapter 6, a test is proposed that uses sib-pairs also under a linkage disequilibrium mapping approach. With this design one could use the sib-pairs for a preliminary linkage scan with a coarse linkage map, then saturate with markers the regions of interest pinpointed by the linkage scan, and apply the LD mapping method proposed in Chapter 6 to fine map the QTL.

1.4 OVERVIEW OF IDENTIFIED LOCI INVOLVED IN COMPLEX TRAITS

There are a relative large number of loci that have been identified to be involved in the genetic control of Mendelian traits, such as cystic fibrosis (Kerem *et al.*, 1989) or diastrophic dysplasia (Hastbacka *et al.*, 1992), but the number of loci that have been identified that are involved in the genetic control of complex traits is much smaller. Here, a brief summary of how many loci influencing complex traits have been identified up to date is presented. Ideally, functional tests (that is, substituting one variant by another using knock-in technology) would be the most conclusive evidence for the variant to be causative, but such tests are not always possible. Instead, other lines of evidence, such as gene expression studies, must be used. Results shown here were obtained from two recent reviews (Korstanje and Paigen, 2002; Glazier *et al.*, 2002) where the authors considered that the evidence available was sufficient proof of causality. Table 1.2 shows the number of genes shown to be responsible for variation in traits of interest in mammals. Traits include among others Alzheimer disease, milk yield, blood pressure, type I and type II diabetes. Most of the genes have been identified in the last decade.

Table 1.2. Number of identified loci shown to influence complex traits in mammals (Korstanje and Paigen, 2002; Glazier *et al.*, 2002).

	Humans	Rodents	Cattle	Pig
Number of loci identified	19	22	1	1

1.5 ANOTHER USE OF LINKAGE DISEQUILIBRIUM

Above, it has been pointed out the importance of assessing the extent of LD in a population for designing mapping experiments. In Chapter 7, LD is used to infer past effective population size in a European human population. This is not only interesting as such, but might help us to develop more accurate population genetic models and mapping strategies. Traditionally, estimates of effective population size have been obtained using single locus data, but this has some drawbacks. First, it relies on the infinite alleles mutation model and corrections must be applied when this model does not hold. Second, because mutation rates are lower than recombination rates, estimates of effective population size in the recent past from mutation rates are more difficult to obtain than from recombination rates.

1.6 MAIN OBJECTIVES

The main objectives of this thesis were:

(1) To study the extent of LD in two populations where LD mapping methods might be applied for mapping QTL, and to investigate the effect of different methodological and sampling strategies in the estimation of LD.

(2) To study the power to detect QTL using LD mapping. I considered three different strategies to map QTL and present results of the power of these strategies under a number of assumptions.

(3) To estimate past effective population size using a multilocus measure of LD with known expectation using published data on human chromosomes 19 and 22.

CHAPTER 2 - Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes

2.1 INTRODUCTION

Linkage disequilibrium (LD) mapping methods use LD at the population level to map trait loci. These methods have higher power (Risch and Merikangas, 1996) and higher resolution than traditional linkage methods because they use information based on a larger number of meioses. They exploit all recombination events that have occurred since the LD was generated. The power of LD mapping methods depends on population parameters such as allele frequencies at the marker and trait loci and level of LD. The resolution achievable depends on the extent of disequilibrium between marker and trait loci. For example, the larger the extent of disequilibrium the lower the density of markers required to detect a trait-marker association (e.g. higher power) but the lower the resolution.

Although the extent and patterns of LD have been extensively studied in human populations, (Moffatt *et al.*, (2000); Mohlke *et al.*, (2001); Reich *et al.*, (2001); Jeffreys *et al.*, 2001; Daly *et al.*, 2001) farm animal populations have been rarely studied.

Farnir *et al.* (2000) and McRae *et al.* (2002) studied the extent of LD in the Dutch black-and-white dairy cattle population and in two sheep populations respectively. Both these studies used family information to infer the most likely phase of the dams and used these phased dams to measure the extent of LD in the population. However, family information is not always available and, if available, collecting the additional family members required may be an inefficient use of resources.

In this chapter, the extent of LD in the United Kingdom (UK) dairy cattle population is estimated. This will determine the feasibility of LD mapping methods in this population and the marker density required for LD mapping to be effective. I illustrate the use of statistical methods that do not require family information to infer population haplotype frequencies as an alternative to family-based haplotyping methods. These methods to estimate haplotype frequencies are relatively efficient compared to those that require family information (Hill, 1974; McKeigue, 2001). I applied these methods in a small data set and assessed the extent of LD in two regions of the genome of 50 randomly selected dairy cattle bulls that were being progeny tested. They were assumed to produce a representative sample of the future extent of LD in the UK dairy cattle population.

2.2 MATERIAL AND METHODS

2.2.1 Data

Data comprised genotypes from 50 Holstein bulls that were being progeny tested. The bulls were born between 1988 and 1995. Bulls were genotyped at six marker loci on chromosome 2 and at seven marker loci on chromosome 6. Genotyping was carried out as described by Wiener *et al.* (2000) and marker identities are given in Tables 2.1 and 2.2. Each bull pedigree was known up to three generations. Grandparents were assumed unrelated. Relationships between bulls are shown in Tables 2.3 and 2.4. Genetic distances (Kosambi map function) between markers were obtained from the map MARC97 (Kappes *et al.*, 1997).

Table 2.1. Genetic map, number of alleles at the marker locus, percentage of missing values, observed heterozygosity at the marker loci, expected heterozygosity under Hardy-Weinberg equilibrium (HWE) and significance level (*P*) of the test for departures from HWE for chromosome 2.

Marker	TGLA226	BMS829	BMS2519	BM2113	IDVGA37	IDVGA2
Genetic map (cM)	80	91.5	101.5	106.2	108.2	117.8
Number of alleles	5	5	5	6	3	5
% of missing values	28	28	34	26	18	32
Observed heterozygosity	0.61	0.33	0.58	0.81	0.39	0.59
Expected heterozygosity	0.79	0.40	0.70	0.76	0.39	0.72
Departures from HWE (*P*)	<0.001	0.08	<0.001	0.57	0.75	0.45

Table 2.2. Genetic map, number of alleles at the marker locus, percentage of missing values, observed heterozygosity at the marker loci, expected heterozygosity under Hardy-Weinberg equilibrium (HWE) and significance level (*P*) of the test for departures from HWE for chromosome 6.

Marker	RM28	BM415	CSN3	BM1236	BMS511	AFR227	BM8124
Genetic map (cM)	74.3	76.3	82.6	83.9	89.8	90.4	94.2
Number of alleles	4	7	3	4	5	6	2
% of missing values	18	4	8	14	10	6	0
Observed heterozygosity	0.66	0.67	0.35	0.60	0.78	0.34	0.16
Expected heterozygosity	0.67	0.79	0.40	0.57	0.74	0.74	0.17
Departures from HWE (*P*)	0.61	<0.001	0.26	0.31	0.81	<0.001	0.99

Table 2.3. Number (N) of maternal-grand-sire and half-sib groups in the sample.

n	N of maternal-grand-sire groups with n bulls	N of paternal half-sib groups with n bulls
1	18	23
2	5	6
3	3	3
6	1	1
7	1	0
Total	28	33

Table 2.4. Additive genetic relationships (a) among bulls calculated using the three generation pedigree.

a	Number of relationships with a
0.00000	837
0.01563	15
0.03130	70
0.06250	135
0.07813	2
0.09380	6
0.12500	105
0.15630	8
0.18750	7
0.25000	31
0.31250	3
0.50000	6

2.2.2 Haplotype frequency estimation and Hardy-Weinberg equilibrium proportions

Maximum likelihood (ML) estimates of all 78 (13*(13-1)/2) two-marker loci haplotype frequencies were obtained by employing the expectation-maximization (EM) algorithm (Excoffier and Slatkin, 1995) as implemented in the program *Gold* (Abecasis and Cookson, 2000). Relationships between bulls were ignored when estimating haplotype

frequencies. I tried to obtain maximum likelihood estimates of six loci and seven loci haplotype frequencies for chromosomes 2 and 6 respectively using the program *Arlequin* (Schneider *et al.*, 2000). The algorithm failed to reach a global maximum likelihood estimate of the haplotype frequencies; therefore estimates were not used in this study. Estimating fewer than six and seven loci haplotype frequencies other than two loci haplotype frequencies was not tried.

Bayesian estimates of six and seven loci haplotype frequencies for chromosome 2 and 6 respectively were obtained using the program PHASE (Stephens *et al.*, 2001). No attempt to estimate LD between non-syntenic loci (loci in different linkage group) using the Bayesian approach was made. Haplotypes were reconstructed ten independent times to make sure that the results obtained were robust even if the algorithm was not converging, as suggested by Stephens *et al.* (2001). The algorithm was run for 10^7 iterations after a burn-in period of 10^4 and estimates from every 100th iteration kept. The program PHASE assumes, by default, a stepwise mutation model, however this assumption was relaxed by using a parent-independent mutation model in which each microsatellite allele has the same chance to mutate to any of the other alleles. A stepwise mutation model is more appropriate for microsatellite markers if the length of each microsatellite allele is known, but this was not known and, therefore, this model could not be assumed.

Departures from Hardy-Weinberg equilibrium (HWE) proportions were tested using an exact test as described by Guo and Thompson (1992). This algorithm is implemented in *Arlequin* (Schneider *et al.*, 2000). Hardy-Weinberg equilibrium is an assumption of the EM algorithm and departures from HWE might lead to biased estimates of haplotype frequencies (Excoffier and Slatkin, 1995). In addition, departures from HWE can be an indication of population stratification, selection of the locus or linked locus, different fertility of parents or different allele frequencies in male and female parents, finite population size, etc.

2.2.3 Level of linkage disequilibrium

Hedrick's normalised measure of disequilibrium (Hedrick, 1987) was obtained from the estimates of the two loci haplotype frequencies. Hedrick's normalised measure of disequilibrium is the extension to multiallelic loci of the normalised measure of disequilibrium defined by Lewontin (1964) for biallelic loci. It is defined as follows:

$$D' = \sum_{m=1}^{k} \sum_{n=1}^{l} m_m q_n \left| D'_{mn} \right| \qquad [1]$$

where k and l are the number of alleles at locus M and Q respectively, m_m and q_n are the population allele frequencies of allele m at locus M and allele n at locus Q respectively. $|D'_{mn}|$ is the absolute value of Lewontin's normalised measure:

$$D'_{mn} = \frac{D_{mn}}{D_{mn}^{\max}} = \frac{\left(h_{mn} - m_m q_n\right)}{D_{mn}^{\max}} \qquad [2]$$

where h_{mn} is the estimated population frequency of the haplotype $M_m Q_n$ and $D^{max}{}_{mn}$ is the maximum amount of disequilibrium possible between allele m at locus M and allele n at locus Q that equals:

$$D_{mn}^{\max} = \begin{bmatrix} \min\{m_m q_n, (1-m_m)(1-q_n)\}; D_{mn} < 0 \\ \min\{m_m (1-q_n), (1-m_m) q_n\}; D_{mn} > 0 \end{bmatrix} \qquad [3]$$

In order to test the statistical significance of the allelic association, the statistic $S = 2ln(L_{LD}/L_{LE})$ was compared to a χ^2 distribution with $(k-1)*(l-1)$ degrees of freedom (Slatkin and Excoffier, 1996). Assuming random mating, L_{LD} is the likelihood computed using the haplotype frequencies found by the EM algorithm and L_{LE} is the likelihood under the assumption of linkage equilibrium. It was assumed that the available sample size was large enough for asymptotic assumptions to hold.

A large number of tests (n = 78) were performed and therefore a Bonferroni correction applied to obtain an appropriate significance level (P) for association between each pair of marker loci. The individual test significance level after correction to give a total significance level (γ) of 0.05 was $P = 1 - (1-\gamma)^{1/n} = 0.0007$ where n was the total number of tests performed. Because some tests are likely to be correlated, the stringent threshold applied is expected to be conservative with respect to the type-I error rate.

2.3 RESULTS

2.3.1 Departures from Hardy-Weinberg equilibrium

Thirteen microsatellite markers spanning bovine chromosomes 2 and 6 were genotyped on 50 dairy bulls. Genetic positions of the markers, number of alleles at each locus, percentage of missing values, observed heterozygosities, expected heterozygosities under HWE for the observed population allele frequencies and significance level of the test for departures from HWE proportions are shown in Table 2.1 and Table 2.2. The thirteen markers had an average observed heterozygosity of 0.53 and an average expected heterozygosity of 0.60. The average distance between markers was 5.2 cM across a length of 57.7 cM. The mean number of alleles was 4.6.

Nine of the thirteen markers studied showed a deficiency of heterozygotes. However, only four of these nine showed significant ($P < 0.001$) departures from HWE proportions. Relatedness between individuals in the sample and the small effective population size of the world-wide dairy cattle population could be the cause of the observed deficiency of heterozygotes.

The heterogeneity of departures from HWE for the thirteen markers studied might be due to undetected null alleles. Null alleles would lead to an excess of homozygotes. Family information was not available so potential problems with marker scoring could not be further investigated. Selection or non-random mating might also explain this heterogeneity. If there was, for example, assortative mating or selection, then only those loci influencing the trait (or those closely linked) would show departures from HWE. These departures would be in the direction observed (i.e. an excess of homozygotes). Migration could also explain the heterogeneity of departures from HWE. Importation of semen from the United States of America to Europe might have led to an excess of heterozygotes in the European Holstein population. However, the most likely origin of the parents of the young bulls is the United States of America, so this seems a less plausible explanation.

2.3.2 Linkage disequilibrium between syntenic marker loci using the EM algorithm

Figure 2.1 shows a plot of the extent of disequilibrium (D') versus genetic map distance measured in cM (genetic map distance is hereafter referred to as genetic distance). The average D' was 44%. D' did not appear to vary as a function of the genetic distance (results using another LD estimator, R^2 for multi-allelic loci (Hudson, 1985), did not yield significantly different results (not shown)). The noisy nature of LD estimators (especially for small sample sizes) coupled with the large genetic map distances would most likely explain

why D' (or R^2) did not seem to vary as a function of genetic map distance. A non-linear equation of type $y = a + be^{-cx}$ was fitted using non-linear regression as implemented by the Genstat's FITCURVE directive (Genstat 5 Committee, 1993) where y is D' and x is the genetic distance in cM. Note that y tends to a when x tends to infinity (i.e. for unlinked loci) and y tends to $a+b$ when x tends to zero (i.e. for loci at the same location). Only a was significantly ($P<0.0001$) different from zero. The estimated parameter values are 0.42 ± 0.06 for a, 0.11 ± 0.18 for b and 0.76 ± 0.59 for e^{-c}. The fit of $y = a$ and $y = a + be^{-cx}$ was compared using a likelihood ratio test. The fit of the two curves was not significantly different.

The level of association ($-\log_{10}(P)$) appeared to show a clearer correlation with distance (Figure 2.2) but still highly variable, especially for the smallest distances. In order to test whether there was a trend a line was fitted. The slope of the line was significantly different from zero ($P = 0.049$) but only marginally. The apparent discrepancies might arise because of an upwards bias of D' for chromosome 2 (with smaller average sample size than chromosome 6). If the sample size for chromosome 2 was similar to that for chromosome 6, then one would expect smaller values of D' and larger values of ($-\log_{10}(P)$). In this case, the apparent non-conformity of Figures 2.1. and 2.2. would probably disappear. All $P < 0.01$ ($-\log_{10}(P) > 2$ in Figure 2.2) correspond to genetic distances smaller than 10.3 cM. Only two pairs of markers were in significant linkage disequilibrium after accounting for multiple testing. These were BM1236-BM8124 ($P = 0.0007$; intermarker distance = 10.3 cM) and BMS511-AFR227 ($P = 0.0007$; intermarker distance = 0.6 cM) on chromosome 6. Before correcting for multiple testing there were a total of eight pairs in significant association at the 5% level. Three of these pairs were on chromosome 2 and five on chromosome 6.

Although a high average level of disequilibrium was observed, only two pairs of loci showed a significant association. In order to test whether the mean level of disequilibrium observed was significant I calculated (assuming independence among statistics): (1) the sum of the 36 statistics (6*7/2 and 5*6/2 from chromosome 6 and 2 respectively; $X^2 = 646$) and (2) the sum of the 36 associated degrees of freedom (df = 456). This overall test for average level of LD across all pairs of syntenic loci was highly significant ($P[\chi^2_{456\ df} \geq X^2 = 646] << 10^{-7}$) indicating that the mean level of disequilibrium was significantly different from zero and that there was not enough power when testing individual pairs.

Figure 2.3 shows a plot of $-\log_{10}(P)$ for each pair of marker loci as a function of D'. Significant LD tended to increase with D' though it was very variable. This variance appeared to be dependent on the value of D'. Pairs of loci with larger values of D´ showed more variable levels of significance.

Figure 2.1. Relationship between genetic distance (cM) and level of linkage disequilibrium (D'). The plotted line represents the fitted line. Crosses and diamonds represent comparisons between pairs of loci on chromosome two and chromosome six respectively.

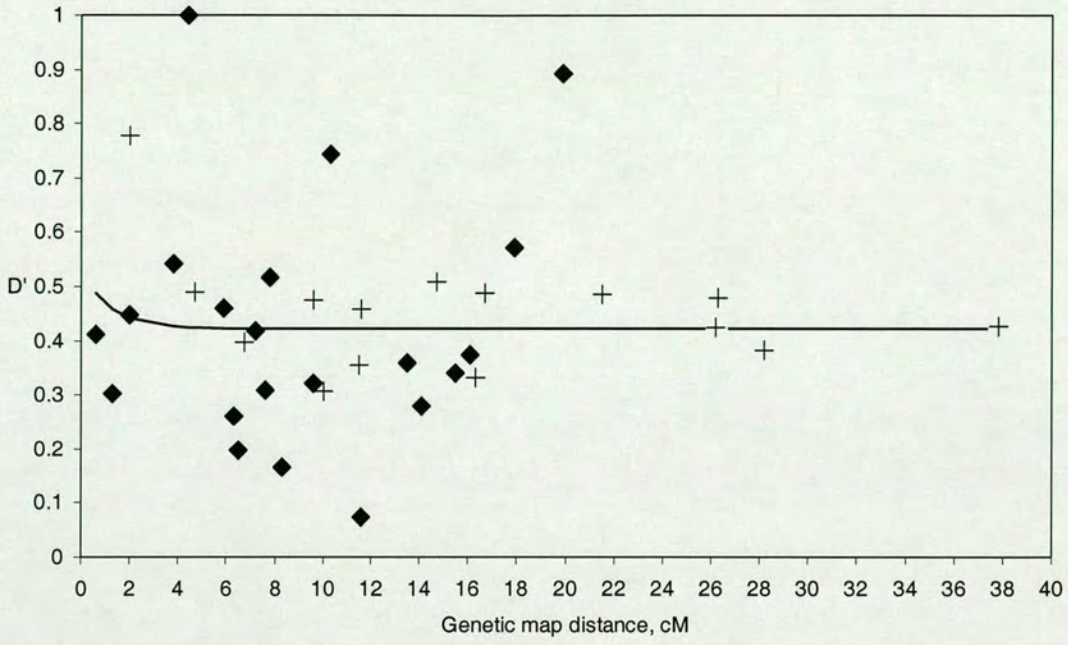


Figure 2.2. Relationship between level of significance ($-\log_{10}$ (P)) and genetic distance (cM) for syntenic loci pairs. Crosses and diamonds represent comparisons between pairs of loci on chromosome two and chromosome six respectively.

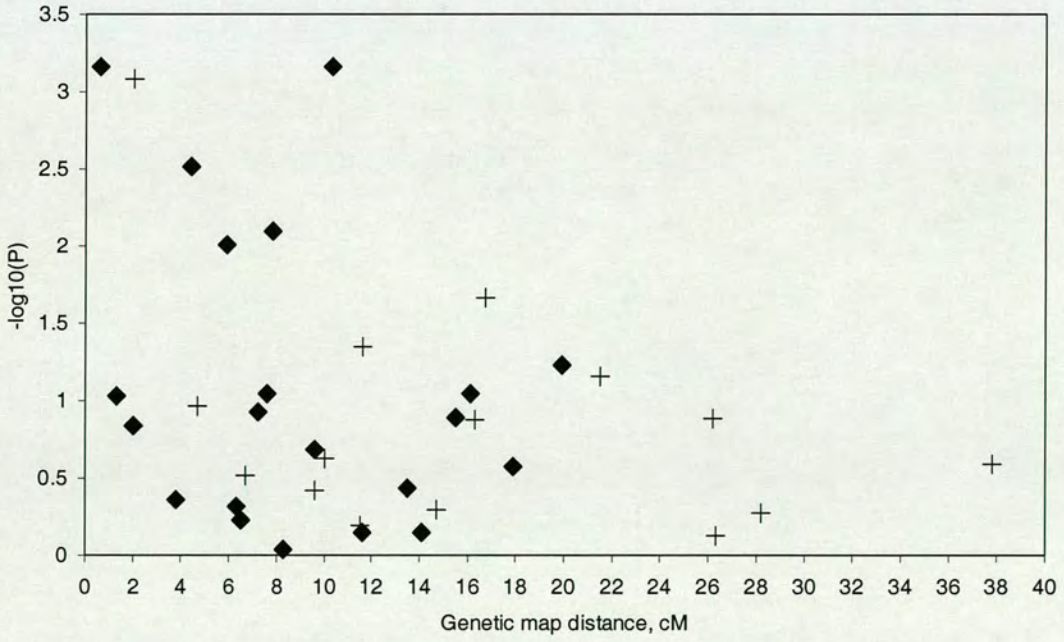
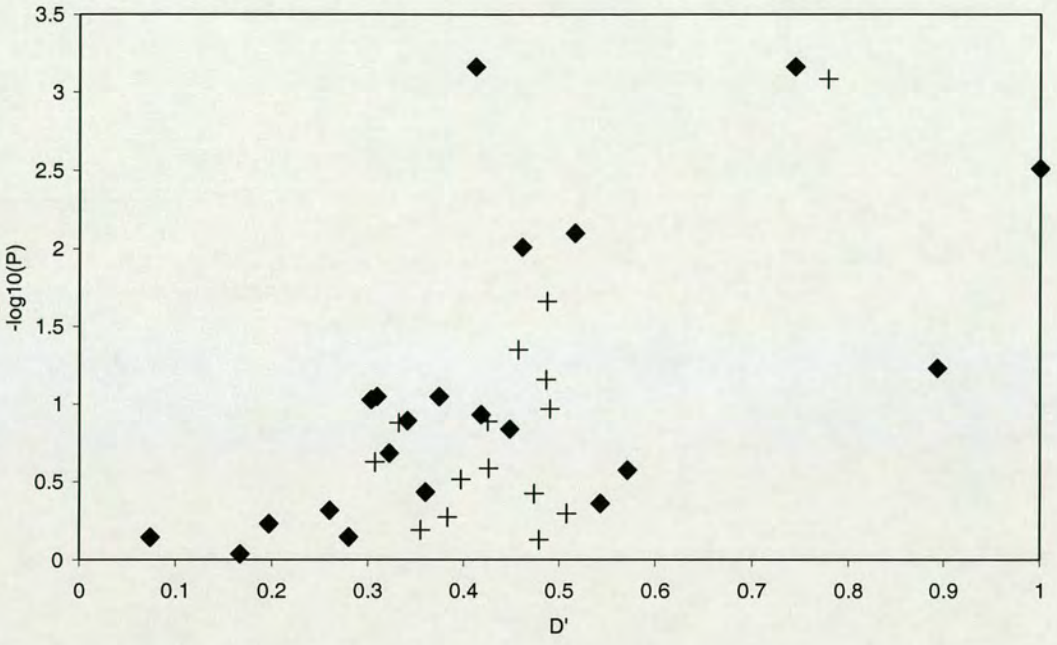


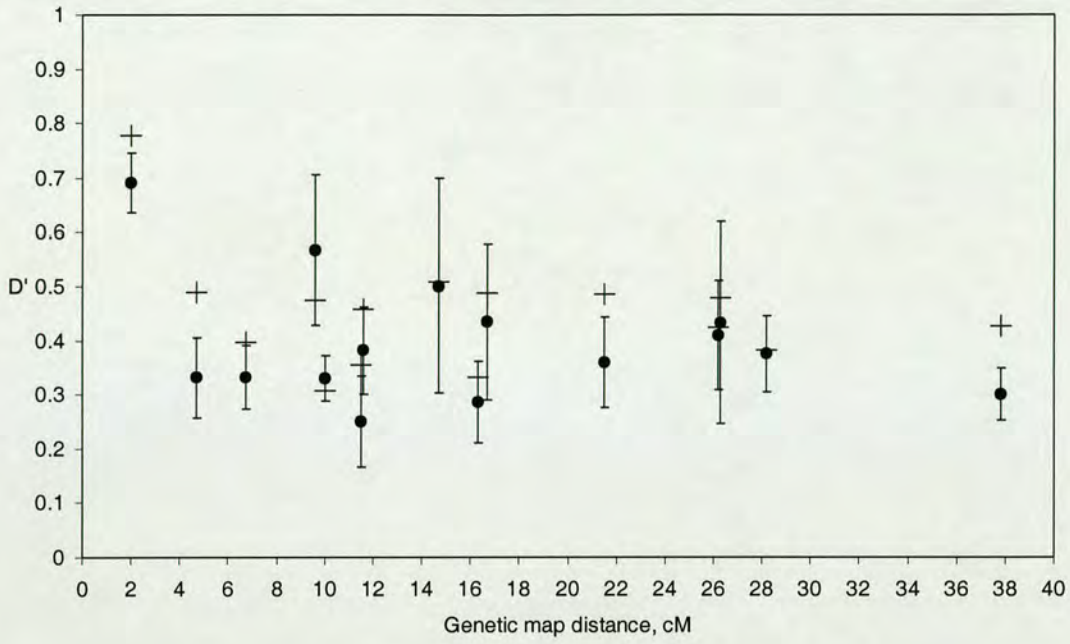
Figure 2.3. Relationship between level of significance ($-\log_{10}(P)$) and level of linkage disequilibrium (D') for syntenic loci pairs. Crosses and diamonds represent comparisons between pairs of loci on chromosome two and chromosome six respectively.



2.3.3 Linkage disequilibrium between syntenic marker loci using the Bayesian algorithm

Figures 2.4 and 2.5 show the comparison in the estimates of D' for chromosome 2 and 6 respectively using the maximum-likelihood (ML) and Bayesian approach to estimate haplotype frequencies. Maximum-likelihood estimates are plotted as single points and Bayesian estimates are plotted as the mean of D' obtained from ten independent estimates of the haplotype frequencies with lines indicating two standard deviations. There is only one estimate of D' when using the EM algorithm therefore formal comparisons between both estimates cannot be performed. However, qualitative comparisons can be done and the general picture is the same regardless of the estimation method used.

Figure 2.4. Comparison of the estimates of D' obtained when using population haplotype frequencies estimated by the maximum-likelihood (crosses) and Bayesian approach (circles) for chromosome 2. Each circle is the mean of ten runs of the program PHASE and the lines are ± 2 SD.



Another important observation is that the variance of D' is highly variable for chromosome 2 but not for chromosome 6 (note that some of the estimates have variance equal to zero). This is probably reflecting more missing values for chromosome 2 than for chromosome 6 (Tables 2.1 and 2.2).

Results using a stepwise mutation model (results not shown) were not significantly different to those from the parent-independent mutation model. This suggests that the algorithm is relatively insensitive to the underlying assumptions about the mutation model.

2.3.4 Linkage disequilibrium between non-syntenic marker loci using the EM algorithm

Figure 2.6 shows the distribution of D' values observed between pairs of non-syntenic loci. The mean level of LD between non-syntenic loci, measured as D', was estimated to be 39%. None of the loci pairs showed significant association between alleles. Indeed, the most significant association was for the pair BM2113-BM1236 (P=0.03; D'= 0.53). The sum of the 42 statistics obtained between non-syntenic loci was 548 and the sum of the 42 associated degrees of freedom was 539. The overall level of association between pairs of non-syntenic loci was not significant ($P[\chi^2_{539\ df} \geq X^2 = 548] = 0.39$). In addition to this overall test, a Fisher's combined probability test (Fisher, 1970) was performed for syntenic and non-syntenic groups that gave similar results (results not shown). Overall, average levels of LD were fairly similar between syntenic and non-syntenic loci. However,

association could be statistically detected between syntenic loci but not between non-syntenic loci, even when the D' values were similar.

Figure 2.5. Comparison of the estimates of D' obtained when using population haplotype frequencies estimated by the maximum-likelihood (diamonds) and Bayesian approach (circles) for chromosome 6. Each circle is the mean of ten runs of the program PHASE and the lines are ± 2 SD.

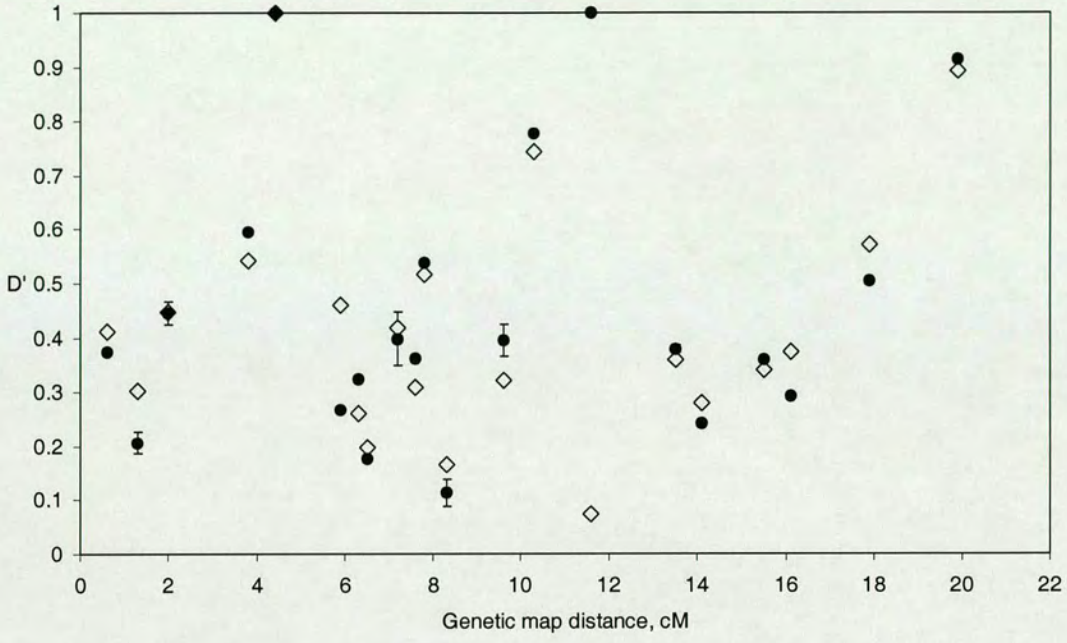
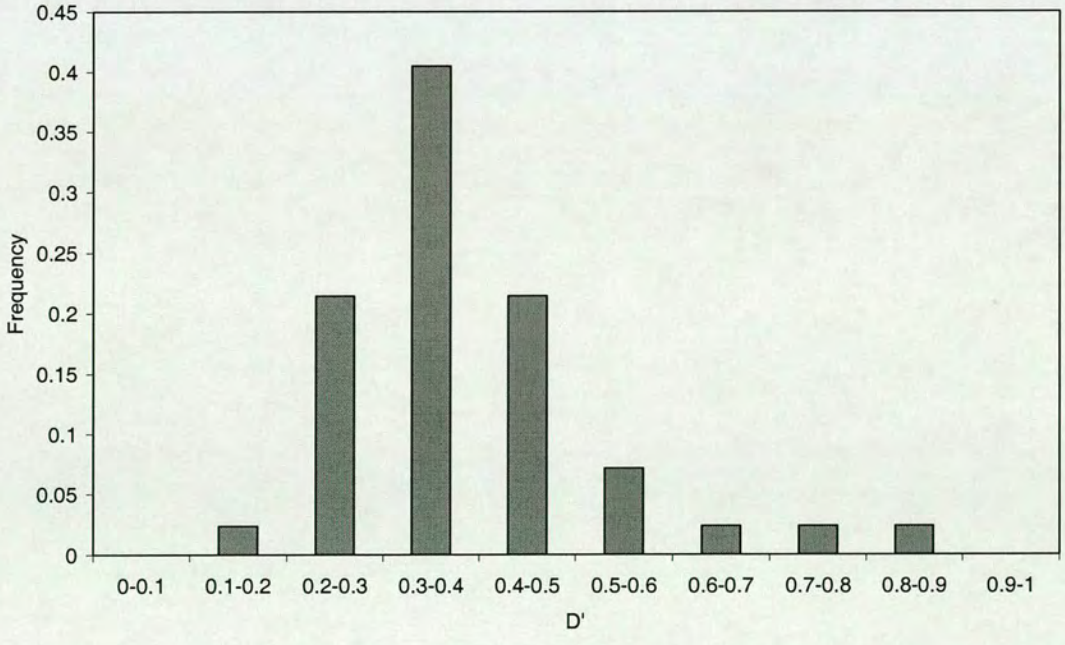


Figure 2.6. Distribution of D' values observed between pairs of non-syntenic loci.



2.4 DISCUSSION

The results show that LD mapping methods could be applied to the future UK dairy cattle population with the available density of microsatellite markers. Significant linkage disequilibrium was found only for genetic distances smaller than about 10 cM, in addition significant association was never found between non-syntenic loci. This would have important implications for LD mapping. Firstly, the mapping resolution achievable with this level of disequilibrium would be finer than with traditional QTL-mapping methods. Secondly, if the lack of significant association found here between loci on chromosomes 2 and 6 were the same across the whole genome, then the number of false positives due to allelic associations between unlinked loci would be small when applying LD methods to map trait loci.

Some aspects of the results presented here differ from those found by Farnir *et al.* (2000). First, they found extensive significant LD between both syntenic and non-syntenic loci. Second, they found average D' values in the same range as those showed here only for genetic distances < 5 cM. Thirdly, they found that only those D' values for the more distant syntenic markers were similar to those between non-syntenic markers. These differences might arise because of two reasons. First, the sample analysed here is more related than theirs, and hence probably shows larger IBD (Identical by descent) regions. They used two different samples for estimating the extent of LD. One sample was composed of bull-dams and the other of cows selected from the general population. Though their first data set might have a level of relatedness as high as that in the data analysed here, it is unlikely that cows in their second data set were as related as the bulls analysed here. Relatedness between individuals can cause an increase in the level of LD even between unlinked loci because larger portions of the genome are identical between related individuals. Second, the sample size of both studies is very different and comparison might be difficult and even inappropriate. The expectation of D' under equilibrium is zero, however its sampling variance depends on the sample size from which it is estimated: the larger the sample size, the smaller the sampling variance. If the sampling variance is large then it is more likely that, just by chance, the estimated value for D' differs from zero. Weir and Hill (1980) derived the variance of R, the correlation of gene frequencies, for biallelic loci. Their arguments about the two sampling processes involved in estimating LD can be extended to a different measure of disequilibrium, say D'. For closely linked loci the variance of R is approximately $1/(1 + 4N_e c) + 1/n$ where N_e is the effective population size, c is the recombination fraction between the two loci and n is the sample size. The variance of R is due to two different

sampling processes, one that reflects the finite size of the population [$1/(1 + 4N_ec)$] and another that reflects that a limited sample of the population [$1/n$] has been drawn (from which disequilibrium and allele frequencies have been estimated). It is worth noting that n is either a sample of n identified chromosomes or n unphased individuals from which disequilibrium and allele frequencies have been estimated. Additionally for D' the difference from its expected value under equilibrium is aggravated by the fact that D' uses the absolute value of D'_{mn}. Even small deviations from equilibrium between pairs of alleles accumulate, leading to an upwards bias in the estimate of D'.

I believe that lack of statistical power, especially after correcting for multiple testing, and an upwards bias (due to the small sample size) in the estimate of D' is the reason why the larger D' values observed did not correspond to more significant allelic associations. It was assumed that all the tests performed were independent, however tests between loci on the same chromosome are correlated, especially if the distance between loci is not large as in these data. The significance thresholds applied after correction are, therefore, very conservative as the number of independent tests really performed was smaller than assumed.

It is unlikely that the departures from HWE expectations observed lead to an important degree of bias in the estimates of haplotype frequencies. The only problem when estimating haplotype frequencies from genotypes comes from individuals that are heterozygous at the loci considered. In this situation, haplotype frequencies cannot be directly counted because it is not possible to distinguish between the two different diplotypes (i.e. an individual with the two loci genotype AaBb could have diplotype Ab/aB or AB/ab). In this case, the EM algorithm iteratively estimates the frequencies of the different haplotypes until the likelihood of the data is maximised and, therefore, maximum likelihood haplotype frequencies obtained. When there is an excess of homozygotes, the number of doubly heterozygous individuals to be resolved is smaller. Consequently, there is little or no bias in the haplotype frequency estimates caused by deviations from HWE due to excess in homozygosity (Osier *et al.*, 1999; Fallin and Schork, 2000).

Six and seven loci maximum likelihood haplotype frequencies for chromosome 2 and 6 respectively could not be obtained. This was because the algorithm failed to reach a global maximum. After each step of the EM algorithm, the likelihood of the data increases (Dempster *et al.*, 1977), however, if the likelihood surface is concave or very flat, then there is no guarantee that a global maximum is reached. Generally, there is no obvious way of knowing if the estimated maximum is just a local or a global maximum. In order to be sure that a global maximum is reached, the algorithm is usually started several times from different starting points and the solution with the maximum likelihood is assumed to be the

global maximum. In the present case, although the likelihood of the data was the same for different runs, different haplotype frequencies were obtained in each of the runs. This suggests that the likelihood surface was very flat due to the insufficient amount of data or dependencies between the data and that the iterative process stopped before reaching the global maximum.

Differences observed between the ML and Bayesian approaches were small and the general conclusions obtained from both estimation procedures were essentially the same. Differences observed between both approaches are slightly larger for chromosome 2, which has more missing values, than for chromosome 6. This might suggest that the amount of data for some loci on chromosome 2 is too small and this is reflected in the slightly larger discrepancies between both approaches. An advantage of the Bayesian over the ML approach is that it provides estimates of the uncertainty associated with each phase, at the cost of a much larger computing time. An advantage of the ML over the Bayesian approach is that implementation of the testing procedure is straightforward in the ML framework. Therefore the decision about the most appropriate method would depend on the intended use of the haplotype frequencies. For example, if one just wanted to test for the presence of LD then the ML approach seems adequate and straightforward but if one wanted to compare haplotype frequencies in a cases/control design then an estimate of the uncertainty of each phase would be necessary.

The fact that the disequilibrium parameter (D') was not dependent on distance (cM) but *P* was (Figures 2.1 and 2.2) and that similar values of D' were observed between syntenic and non-syntenic loci (but significance level was different), suggests that the utility of D' to assess the amount of disequilibrium is limited. This is important if assessment of disequilibrium is done as a preliminary study to determine, for example, the marker density required for a mapping study. In this case, the correlation between *P* and distance will give a clearer "picture" of the marker density required.

The region of chromosome 6 where the most significant LD was detected has been reported to harbour QTLs influencing milk, fat and protein yield in the UK dairy populations (Wiener *et al.*, 2000) and other populations such as the Israeli Holstein population (Ron *et al.*, 2001). This suggests that selection for milk production traits could have generated LD in this region, which was detectable even with the large amount of background LD observed.

Fine mapping of trait loci in outbred populations relies on population-based samples for which linkage disequilibrium between trait and marker loci is expected to occur at smaller distances than in family-based samples. The amount of linkage disequilibrium between marker loci in a population will give us information about the marker density

required to perform the mapping study. In livestock populations, this type of study has always been done using family information to infer phase. However, this procedure requires typing additional family members. Even if possible, typing these extra members might be an inefficient use of resources, especially when statistical methods such as those described in this study are known to perform reasonably well.

CHAPTER 3 - Extent of linkage disequilibrium in a human Sardinian sub-isolate: sampling and methodological considerations

3.1 INTRODUCTION

In recent years, human geneticists have advocated the use of linkage disequilibrium, the non-random association of population allele frequencies at two or more loci, to map genes related to common human complex diseases. Linkage disequilibrium (LD) mapping relies on the assumption that there have not yet been enough generations of recombination to break down the association between a causative locus and nearby markers. The association, generated by, for example, mutation, selection, drift or migration is reduced each generation by a function of the recombination fraction between the loci and the population size (Hill and Robertson, 1966). The closer the marker and trait locus, the larger the number of generations required to break down their association. LD will also be erased faster in large populations than in small ones. However, empirical data has shown conflicting results. Eaves *et al.* (2000) found comparable levels of LD in two genetic isolates (Sardinia and Finland) and two outbred populations in the USA and UK. They argued that the number of founders in each isolate was large enough to have multiple copies of the common alleles represented in the founder population. Therefore, the recombinational history of common alleles for the four populations would date back to the same origin in the general population. Zavattari *et al.* (2000) found similar levels of LD in the general Sardinian population and in the UK population but found increased levels in a sub-isolate of the Sardinian population. Angius *et al.* (2002a) also found increased levels of disequilibrium in the sub-isolate of Talana than on the general Sardinian population using six microsatellites markers on the long arm of chromosome X.

The village of Talana was selected as an example of a sub-isolate within the general Sardinian population. Talana is one of the most isolated villages in the Ogliastra region. It was selected because of its documented isolation until 25-40 years ago and the reduced number of founders. Angius *et al.* (2001) estimated that 80% of the ~1300 people that currently live in Talana descend from eight paternal and eleven maternal lineages. Talana has experienced a slow population growth from the beginning of the 17th century to the present. The estimated population size was 200 in the middle of the 17th century, doubling at the end of the 19th century and then tripling at the end of the 20th century. Simulation studies (Slatkin, 1994; Kruglyak, 1999; Wright *et al.*, 1999) showed that populations maintained at constant size or showing slow population growth after their founder event followed by rapid

expansion are more likely to show high levels of LD than those that experience a rapid growth immediately after their founder event. The Talana population meets these requirements and hence, seems more suitable for detecting genes using linkage disequilibrium than other populations, such as the Finnish (Peltonen *et al.*, 1999), that have a larger number of founders and have experienced rapid growth just after their founder event.

In this chapter, results about the extent of LD on chromosome 19 in the Talana population are presented. The effect of the number of generations available to estimate founder haplotypes and the number of founder haplotypes on the measure of disequilibrium D' (Hedrick, 1987) is studied. The effect on D' of estimating population haplotype frequencies without using family information is also studied.

3.2 MATERIAL AND METHODS

3.2.1 Data

A total of 775 individuals distributed in 120 families were available. Table 3.1 shows the distribution of family size in the sample. Founders, those without parents in the pedigree, had different number of documented generations descending through the pedigree. The number of generations (tiers) within each family varied from one to three.

Table 3.1. Distribution of family size for the 120 families available.

Number of families with n members	n	Total number of individuals in families of n members
18	3	54
22	4	88
18	5	90
9	6	54
13	7	91
13	8	104
8	9	72
9	10	90
2	11	22
2	12	24
2	13	26
2	14	28
2	16	32
Total number of families = 120		Total number of individuals = 775

3.2.2 Genetic linkage map

The genetic linkage map for chromosome 19 was constructed using *Cri-Map* (Green *et al.*, 1990) (http://linkage.rockefeller.edu/multimap/crimap/) using Haldane's map function. Genotypes were available for 21 microsatellite markers and the APOE gene. Table 3.2 shows the description of the microsatellite markers, the linkage map and the number of informative meioses at each locus. The constructed map agreed with the published map (ftp://ftp.genethon.fr/pub/Gmap/Nature-1995/) in the order of all the marker loci. The estimated genetic linkage map was considered more appropriate to this study because it was estimated from the population studied and because it is likely to be more accurate than the published map (Dib *et al.*, 1996), which is only based on 186 meioses. The described map was the best-supported one (that is, the one with the largest likelihood [Log-likelihood=-603]). In order to be sure that the likelihood function had reached an absolute maximum and not a local maximum, different orderings of loci were used as starting values, as well as, the

flips2, flips3, flips4 and flips5 options. *Cri-Map*'s flipsX option finds relative Log-likelihoods for all permutations involving X adjacent loci.

Table 3.2. Linkage map and number of informative meioses used to infer it.

Marker	Linkage map (cM)	Informative meioses
D19S886	0.0	332
D19S209	12.4	235
D19S894	17.7	398
D19S216	20.8	344
D19S884	28.3	358
D19S865	31.4	337
D19S221	36.7	337
D19S226	43.1	337
D19S566	53.0	459
D19S931	56.1	378
D19S414	62.5	294
D19S220	70.0	439
APOE	74.2	48
D19S420	74.2	372
D19S903	76.2	409
D19S902	83.8	365
D19S904	92.5	223
D19S888	106.2	321
D19S921	107.2	431
D19S572	109.3	436
D19S418	115.7	252
D19S210	124.4	365

The region studied spanned ~124cM with an average distance between markers of 5.92 cM. The mean number of marker alleles per locus was 8.8. Alleles with only a single copy in the dataset were treated as missing data. The number of locus pairs within 10 cM intervals was roughly the same for all the intervals. There were 30, 32, 24, 34 and 24 pairs for map distances between 0-10cM, 10-20cM, 20-30cM, 30-40cM and 40-50cM respectively.

3.2.3 Hardy-Weinberg equilibrium proportions

Departures from Hardy-Weinberg equilibrium (HWE) proportions were tested using *Arlequin* (Schneider *et al.*, 2000) (http://lgb.unige.ch/arlequin/) which uses a Markov Chain Monte Carlo (MCMC) algorithm (Guo and Thompson, 1992) to estimate the exact probability (Fisher's exact test) of the data based on a hypergeometric distribution of the genotypic counts with the number of classes equal to the number of possible genotypes at the locus. The individual test significance level after correction to give a total significance level

(γ) of 0.05 was $P = 1\text{-}(1\text{-}\gamma)^{1/n} = 1\text{-}(1\text{-}0.05)^{1/22} = 0.002$ where n was the total number of tests performed.

The observed levels of heterozygosity and mean number of alleles in the Talana population (considered here to be an inbred human population because of its small effective population size and low immigration during the last two and a half centuries) was compared with an outbred population (with a larger effective population size and relatively large immigration typical of the large cities). The CEPH (Centre d'Etudes du Polymorphisme Humain) reference families were used as an example of an outbred population. The mean number of alleles and observed heterozygosity at the microsatellite loci used in this study for the CEPH families were obtained from the Genethon web page (ftp://ftp.genethon.fr/pub/Gmap/Nature-1995/). These data are based on 8 families (134 individuals) and 186 meioses.

3.2.4 Haplotype estimation using family information

All individuals' diplotypes (that is, the pair of haplotypes that compose the genotype) for the 22 available loci were estimated using the program *Simwalk2* (http://watson.hgen.pitt.edu/docs/simwalk2.html) (Sobel and Lange, 1996) and founder haplotypes were selected for further analysis. Founder haplotypes were counted and population two-locus haplotype frequencies obtained.

The program *Simwalk2* models genetic descent states of the pedigree, that is, describes how the genes descend through the pedigree and how the alleles from the founders descend down each path. For modelling these genetic states, it uses an MCMC technique known as simulated annealing, in which the Markov Chain steps are accepted with decreasing probability. In the first steps of the chain almost all legal descent graphs are accepted and in the latest states of the chain only those steps that are more likely than the previous one are accepted. By doing so, the space of legal descent graphs is largely sampled at the beginning of the chain and is progressively reduced as the chain advances, making convergence faster. Using diplotypes obtained from two different runs of *Simwalk2* with different random seeds gave virtually the same results (only results for one of them are shown).

A maximum of 381 founder 22-locus diplotypes could be inferred and they were assigned to different categories depending on the number of generations of descent in the pedigree. Those founders that had great-grandchildren in the pedigree were assigned to category G3; founders with grandchildren only to category G2 and founders with children only to G1. G1, G2 and G3 had a total of 187, 168 and 26 diplotypes respectively. Unless otherwise stated the results shown were obtained when using all founders available.

3.2.5 Haplotype frequency estimation without using family information

Maximum likelihood estimates of all 231 (22*(22-1)/2) two-marker locus haplotype frequencies were estimated by employing the expectation-maximization (EM) algorithm (Excoffier and Slatkin, 1995). No attempt to estimate more than two-locus haplotypes using the EM algorithm was made because the computing time of the algorithm increases exponentially with the number of loci and the number of alleles at each locus.

3.2.6 Measuring the amount of linkage disequilibrium

LD was measured using a statistic proposed by Hedrick (1987). This statistic is an extension for multiallelic loci of the normalised measure of disequilibrium defined by Lewontin (1964) for biallelic loci. It is defined as follows:

$$D' = \sum_{m=1}^{k} \sum_{n=1}^{l} m_m q_n \left| D'_{mn} \right| \qquad [1]$$

where k and l are the number of alleles at locus M and Q respectively, m_m and q_n are the population allele frequencies of allele m at locus M and allele n at locus Q respectively. $|D'_{mn}|$ is the absolute value of Lewontin's normalised measure:

$$D'_{mn} = \frac{D_{mn}}{D_{mn}^{\max}} = \frac{\left(h_{mn} - m_m q_n\right)}{D_{mn}^{\max}} \qquad [2]$$

where h_{mn} is the estimated population frequency of the haplotype $M_m Q_n$ and $D^{max}{}_{mn}$ is the maximum amount of disequilibrium possible between allele m at locus M and allele n at locus Q that equals:

$$D_{mn}^{\max} = \begin{bmatrix} \min\{m_m q_n, (1-m_m)(1-q_n)\}; D_{mn} < 0 \\ \min\{m_m (1-q_n), (1-m_m) q_n\}; D_{mn} > 0 \end{bmatrix} \qquad [3]$$

D' was based on estimates of two-locus haplotype frequencies obtained both by counting the 381 22-locus founder diplotypes when using family information (phased founders), and from the two-locus population haplotype frequencies obtained without using family information (unphased founders).

3.2.7 Test for association when using phased individuals

Gold (Abecasis and Cookson, 2000) was used to test the statistical significance of the allelic association between all pairs of loci when using inferred haplotypes. Association is tested by means of a standard chi-square test (based on the observed and expected haplotype frequencies) with *(k-1)*(l-1)* degrees of freedom where *k* and *l* are the number of alleles at the two loci. *Gold* pools low frequency alleles in order to avoid spurious results due to small sample sizes and sparse contingency tables. Two different pooling strategies were used, pooling at 1% and at 7%. Unless otherwise stated the results presented are those corresponding to the 7% pooling. Not all loci from the 381 22-locus diplotypes were scored, therefore the number of two-locus haplotypes on which D' and the significance of the allelic association were based varied from 312 to 606 haplotypes.

3.2.8 Test for association when using unphased individuals

The statistical significance of allelic association was tested comparing the likelihood ratio statistic $S = 2\ln(L_{LD}/L_{LE})$ with a χ^2 distribution with *(k-1)*(l-1)* degrees of freedom (Slatkin and Excoffier, 1996). Assuming random mating, L_{LD} is the likelihood computed using the haplotype frequencies found by the EM algorithm and L_{LE} is the likelihood under the assumption of linkage equilibrium. I used *Gold* at different pooling frequencies of the alleles (1% and 7%) and my own implementation of the EM algorithm, which does not do any pooling. Comparing the statistic S with a χ^2 distribution is, strictly speaking, only valid under asymptotic assumptions, which are likely to hold here due to the large sample size of the data set.

3.3 RESULTS

3.3.1 Hardy-Weinberg equilibrium proportions

Table 3.3 shows the number of genotypes, the observed heterozygosity (OH) and expected heterozygosity (EH) at each locus, and the significance level for the test of HWE proportions. Multiple testing was accounted for by applying a Bonferroni correction as described in the previous section. Five out of twenty-two loci showed departures from HWE proportions after accounting for multiple testing. They showed a deficiency of heterozygotes compared with what one would expect under HWE. The overall difference in observed and expected heterozygosity was tested using a paired t-test. The difference was highly significant ($P<0.0001$). These results are consistent with non-random mating (due to mating of related individuals) or population sub-structure.

The heterogeneity of departures from HWE might be due, for example, to assortative mating or selection if those loci that showed significant departures from HWE or closely linked loci were influencing the traits for which assortative mating or selection is occurring. Alternatively, null alleles might also explain this heterogeneity but this explanation is less likely since the markers used have been extensively tested in other populations where null alleles have not been reported (Alan Wright, personal communication). In addition, null alleles usually lead to Mendelian segregation errors that were not detected when constructing the linkage map or when estimating the individuals' diplotypes. Under all scenarios (non-random mating due to the mating of related individuals, assortative mating, selection and null alleles), one would expect an excess of homozygotes.

Table 3.3. Number of founder genotypes available at each locus, observed and expected heterozygosity given the observed allele frequencies in the founders and the significance level of the test for departures from HWE proportions.

Marker	Number of genotypes	Observed heterozygosity	Expected heterozygosity	Departures from HWE (*P*)
D19S886	254	**0.65**	**0.74**	**0.002**
D19S209	171	0.75	0.77	0.719
D19S894	264	0.75	0.82	0.006
D19S216	271	0.75	0.77	0.230
D19S884	261	**0.67**	**0.79**	**<0.001**
D19S865	258	**0.65**	**0.79**	**0.001**
D19S221	246	**0.74**	**0.85**	**<0.001**
D19S226	267	0.63	0.64	0.794
D19S566	286	0.79	0.85	0.331
D19S931	276	0.66	0.72	0.026
D19S414	270	0.57	0.65	0.011
D19S220	264	0.80	0.81	0.046
APOE	294	0.12	0.12	0.692
D19S420	254	0.70	0.75	0.081
D19S903	267	0.79	0.80	0.085
D19S902	265	0.71	0.72	0.182
D19S904	235	0.54	0.52	0.814
D19S888	270	0.65	0.67	0.735
D19S921	277	0.71	0.76	0.596
D19S572	274	**0.73**	**0.79**	**<0.001**
D19S418	238	0.52	0.60	0.004
D19S210	264	0.69	0.75	0.105
Mean (SD)		0.66 (0.14)	0.71 (0.15)	

Table 3.4 shows the number of alleles (NA) observed in Talana and the CEPH families, as well as, the observed heterozygosity (OH) in Talana and CEPH families.

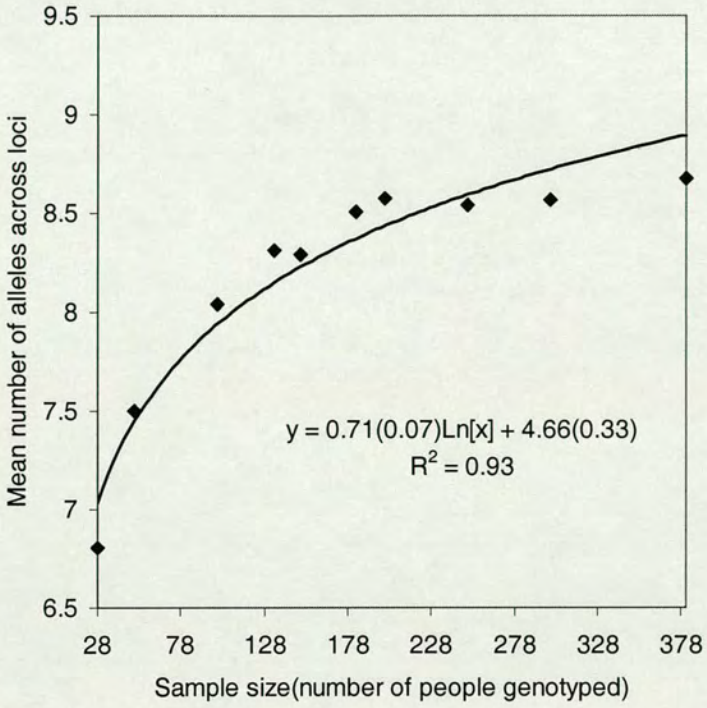
The mean number of alleles in Talana was 8.8 and 7.8 in the CEPH families. The mean difference in the number of alleles at the two populations, tested using a paired t-test, was significant at the 2% level. Also, the mean observed heterozygosity was smaller in Talana (0.66) than in the CEPH families (0.75). The mean difference was highly significant ($P < 0.0001$; paired t-test). The difference observed in the number of alleles could be due to a larger sample size for the Talana population. I tested whether the difference in sample size between the CEPH and Talana samples could explain their difference in the mean number of alleles. For this, 500 samples of different size from the Talana data were bootstrapped. The sample sizes were 28, 50, 100, 134, 183, 150, 200, 250, 300 and 381 diplotypes. For each diplotypes' sample size (28-381) the mean number of alleles across the 500 bootstrapped samples was estimated at each locus and the mean number of alleles across loci for a given sample size estimated. Then a logarithmic curve was fitted to the mean number of alleles across loci obtained from the bootstrapping. Results are shown in Figure 3.1. Substituting the

sample sizes of the CEPH (28 founders) and Talana (381 founders) in the equation shown in Figure 3.1 gave an expected number of alleles of 7.02 and 8.9, respectively. This showed that the difference in the number of alleles observed in the two populations could be attributed to the difference in sample size. These results also showed that the expected number of alleles for a sample of 28 unrelated people from Talana would have an average (over the 22 loci) of ~0.8 alleles less than the sample from the CEPH families gave. Talana showed smaller allele diversity and heterozygosity when corrected for sample size than an outbred population, which is consistent with the hypothesised drift and founder effects.

Table 3.4. Number of alleles (NA) and observed heterozygosity (OH) in Talana and the CEPH families.

Marker	NA in Talana	NA in CEPH families	OH in Talana	OH in CEPH families
D19S886	6	5	0.65	0.66
D19S209	7	7	0.75	0.77
D19S894	12	11	0.75	0.77
D19S216	6	5	0.75	0.75
D19S884	10	10	0.67	0.86
D19S865	8	13	0.65	0.88
D19S221	11	10	0.74	0.86
D19S226	14	12	0.63	0.84
D19S566	10	9	0.79	0.86
D19S931	10	10	0.66	0.77
D19S414	7	7	0.57	0.77
D19S220	12	10	0.80	0.84
APOE	3	3	0.12	0.11
D19S420	8	7	0.70	0.79
D19S903	11	7	0.79	0.78
D19S902	11	9	0.71	0.79
D19S904	7	4	0.54	0.64
D19S888	10	7	0.65	0.81
D19S921	9	8	0.71	0.78
D19S572	10	7	0.73	0.80
D19S418	6	6	0.52	0.65
D19S210	6	6	0.69	0.73
Mean (SD)	8.8 (2.6)	7.8 (2.6)	0.66 (0.14)	0.75 (0.16)

Figure 3.1. Expected relationship between sample size and the mean number of alleles observed for the 22 loci studied, obtained by bootstrapping (see text). The equation of the fitted line and the standard errors (in brackets) of the estimated parameters are shown.



3.3.2 Extent of linkage disequilibrium using phased founders

Figure 3.2 shows how linkage disequilibrium, measured as D', decays with genetic map distance. The mean D' was 0.143 (with a maximum of 0.356 and a minimum of 0.055). An equation of type $y = a + be^{-cx}$ was fitted using non-linear regression as implemented by the Genstat FITCURVE directive (Genstat 5 Committee, 1993) where y is D' and x is the genetic distance in cM. Note that $y \rightarrow a$ when $x \rightarrow \infty$ (a is the mean background level of LD) and $y \rightarrow a+b$ when $x \rightarrow 0$ ($a+b$ is the mean level of disequilibrium for loci at the same location). The fitted curve accounted for 47% of the total variance and the estimated parameters and their standard errors were 0.116 ± 0.004 for a, 0.184 ± 0.016 for b and 0.917 ± 0.012 for e^{-c}. It was considered that a useful level of LD (measured as D') for LD mapping to be effective was half the difference between the fitted maximum (0.300) and minimum (0.116) value (herein, referred to as the half-length). This value was 0.208 and corresponded to a distance of about 8 cM. The half-length was defined as half the difference between the maximum and minimum fitted value following the definition given by Reich *et al.* (2001). They defined it as the distance where D' decays to 0.5, however they observed values of D' between 0 and 1 and therefore this definition was not appropriate for this data set.

Figure 3.2. Decay of D' values observed between marker loci on chromosome 19 as a function of genetic map distance (in cM). Horizontal lines represent the mean of D' values computed at 5 cM intervals. The plotted line represents the fitted line.

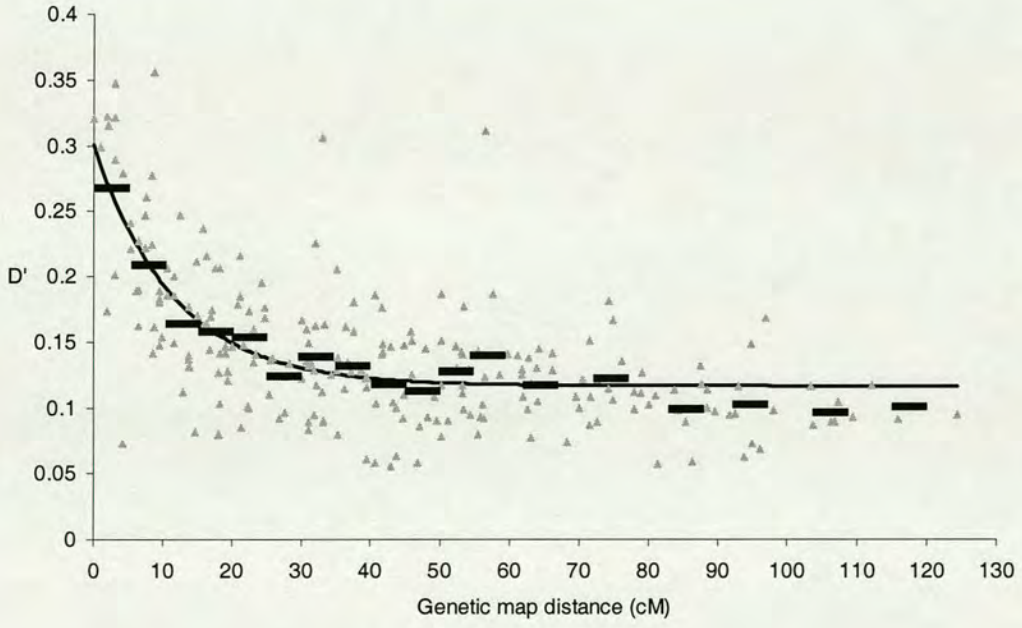
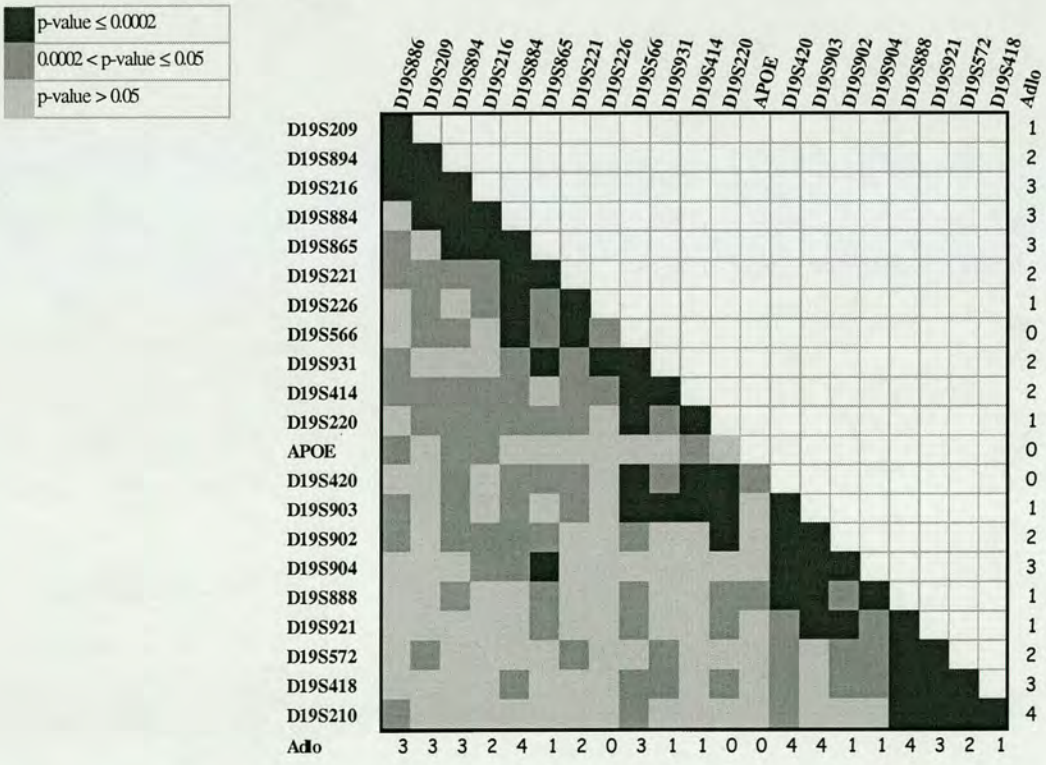


Figure 3.3 shows the statistical significance of the locus pairs. The statistical significance was classified as: highly significant ($P \leq 0.0002$), significant ($0.0002 < P \leq 0.05$) and not significant ($P > 0.05$). The first category accounts for multiple testing using a Bonferroni correction when 231 independent tests were assumed. The numbers of locus pairs in the three classes were 55, 78 and 98, respectively. At each locus the number of adjacent loci in highly significant LD with this locus (abbreviated as Adlo in Figure 3.3) were counted in each direction from the marker locus and the average for all loci obtained. Given a marker locus, the average number of markers adjacent to it that showed highly significant LD was 1.90 (after averaging in both directions) with a variance of 1.60. The average extent of LD from a given marker was estimated by multiplying the average number of markers adjacent to it that showed highly significant LD by the average distance between markers. This was 11.25 cM (1.90*5.92) with a standard deviation of 7.48 (the standard deviation was estimated assuming that 5.92 was constant).

Figure 3.3. Linkage disequilibrium statistical significance between pairs of loci and number of adjacent loci in highly LD (Adlo).



A total of 44 out of the 62 locus pairs (71% of the pairs) that were at a distance ≤ 20cM were in highly significant ($P<0.0002$) LD and only 11 locus pairs out of the 169 (6.5% of the pairs) that were at a distance >20cM were in highly significant ($P<0.0002$) LD. One

must keep in mind that choosing a different distance threshold would yield different results, however for this data set the threshold of 30 cM yielded only slightly different results (60.5% and 2% of the loci in highly significant association for distances less and more than 30 cM, respectively). The proportion of highly significant associations for unlinked loci (distances larger than 30 cM) was smaller than expected by chance, showing that the Bonferroni correction is too conservative.

3.3.3 Effect of the number of generations available to estimate diplotypes

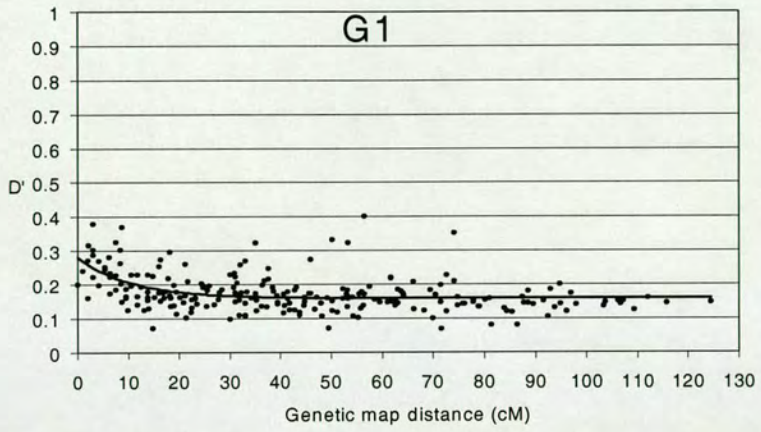
Figure 3.4 shows how LD decayed as a function of genetic map distance for founders of types G1, G2 and G3. For G1, G2, and G3 the mean values of D' were 0.175, 0.189 and 0.456 respectively. A non-linear equation of the type $y = a + be^{-cx}$ was also fitted as described above. The estimated parameter values and their standard errors are shown on Table 3.5.

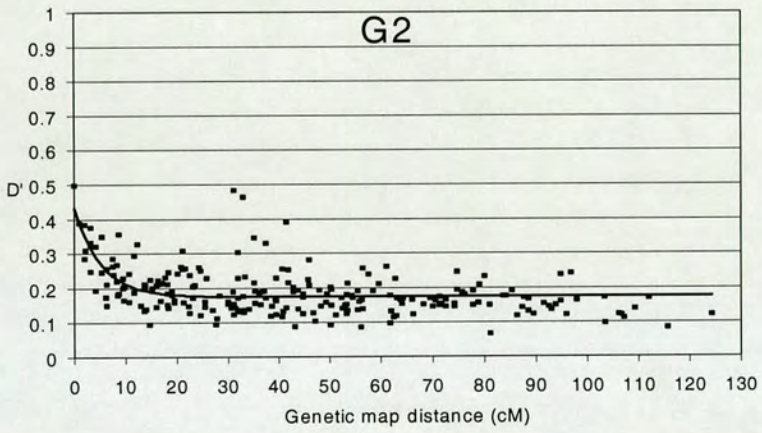
A likelihood ratio test was used to test whether there was a difference in the estimates of D' when using founders G1, G2 and G3. I compared the likelihood of the full model, that is fitting three different a, b, e^{-c} and residual variances for G1, G2 and G3, and that of the reduced model in which I fitted, as with the full model, three different residual variances for G1, G2 and G3 but only one a, b, e^{-c}. To make sure that the algorithm had reach the maximum of the likelihood function the search was started from 20 different starting points and the one with the best ln-likelihood is reported here. The full model had 12 parameters estimated and a ln-likelihood equal to 1458 whereas for the reduced model the number of parameters fitted was 6 and it had a ln-likelihood equal to 1395. Twice the difference in ln-likelihood was compared to a chi-square distribution with 6 degrees of freedom. The full model fitted the data significantly ($P < 10^{-8}$) better than the reduced model. This suggests that the extent of LD differs depending on the number of generations available to infer diplotypes. However, differences in sample size between G1, G2 and G3 might, also, explain those differences.

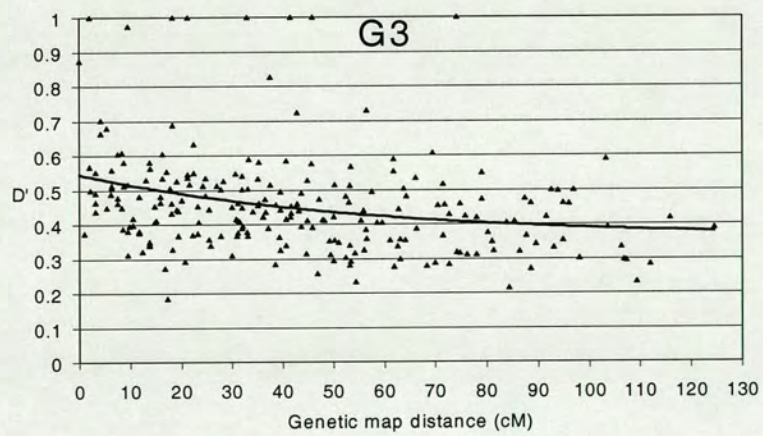
Table 3.5. Estimated parameter values and their standard errors for founders G1, G2 and G3.

Type of founders	a	b	e^{-c}
G1	0.16±0.01	0.12±0.02	0.91±0.02
G2	0.17±0.01	0.26±0.04	0.83±0.03
G3	0.36±0.09	0.19±0.07	0.98±0.02

Figure 3.4. Relationship between genetic distance (cM) and level of linkage disequilibrium (D'). The plotted line represents the fitted line ($y = a + be^{-cx}$). The total number of diplotypes was 187, 168 and 26 for classes of individuals with children only (G1), grandchildren (G2) and great-grandchildren (G3) in the pedigree respectively.



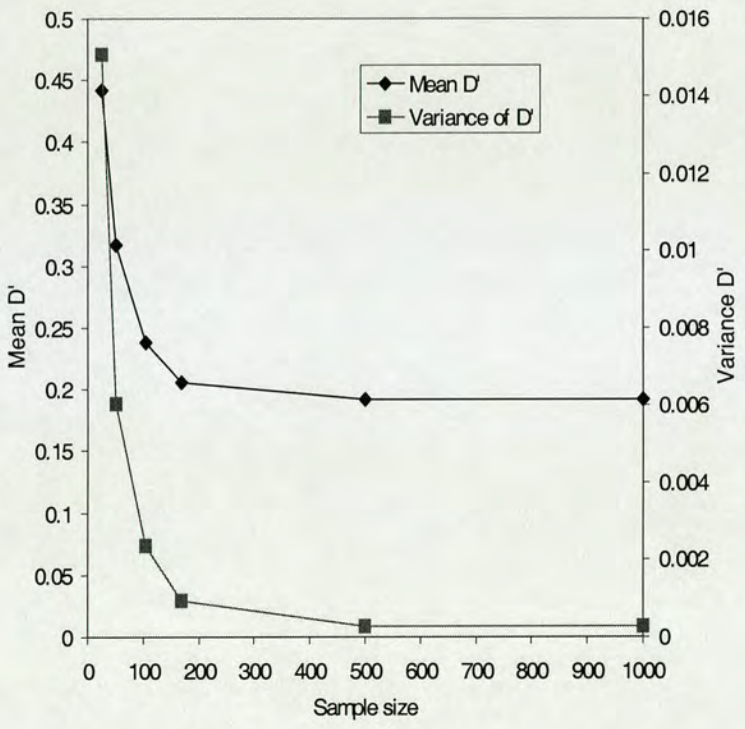




3.3.4 Effect of the sample size

In order to assess the sample size effect on the extent of LD, samples of different sizes of the same data were bootstrapped. Figure 3.5 shows the mean of the means and variances of the 231 locus pairs obtained from bootstrapping samples of 26, 52, 104, 168, 500 and 1000 diplotypes from G2. Note that bootstrapping samples larger that the actual sample size (168), is not going to improve the estimation of the mean or variance of D'. Strictly speaking one should always bootstrap samples of the same size as the original sample. A total of 1000 replicates was obtained, and the mean over replicates and the variance of D' were estimated for all marker pairs, as well as the total mean (that is the mean across loci) of the means and variances obtained.

Figure 3.5. Effect of sample size on the mean of the means and variances of D' for the 231 locus pairs of founders G2. Points represent sample sizes of 26, 54, 104, 168, 500 and 1000 diplotypes.



There was about a 2.1 fold increase in the value of D' when the sample size decreased from 168 to 26. This increase was very similar to that found between G1 or G2 and G3. The estimate of the mean D' was between 2.4-2.6 fold larger in G3 than in G2 or G1. Figure 3.5 shows that the variance of D' was about 16 fold larger for a sample size of 26 than for one of 168 diplotypes. Figure 3.6 shows the mean of the means and variances of the 231 locus pairs obtained from bootstrapping samples of 25, 50, 100, 150, 200, 250, 300, 350

and 381 diplotypes from all the founder diplotypes (G1, G2 and G3). It shows that the mean and variance of D' tend to flatten for sample sizes of about 200 diplotypes.

Figure 3.6. Effect of sample size on the mean of the means and variances of D' for the 231 locus pairs of all the founders available.

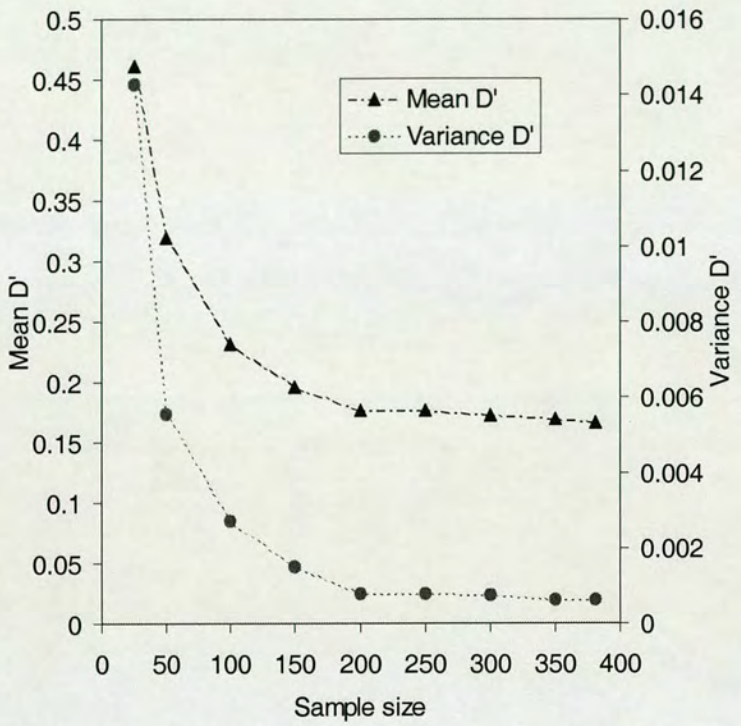
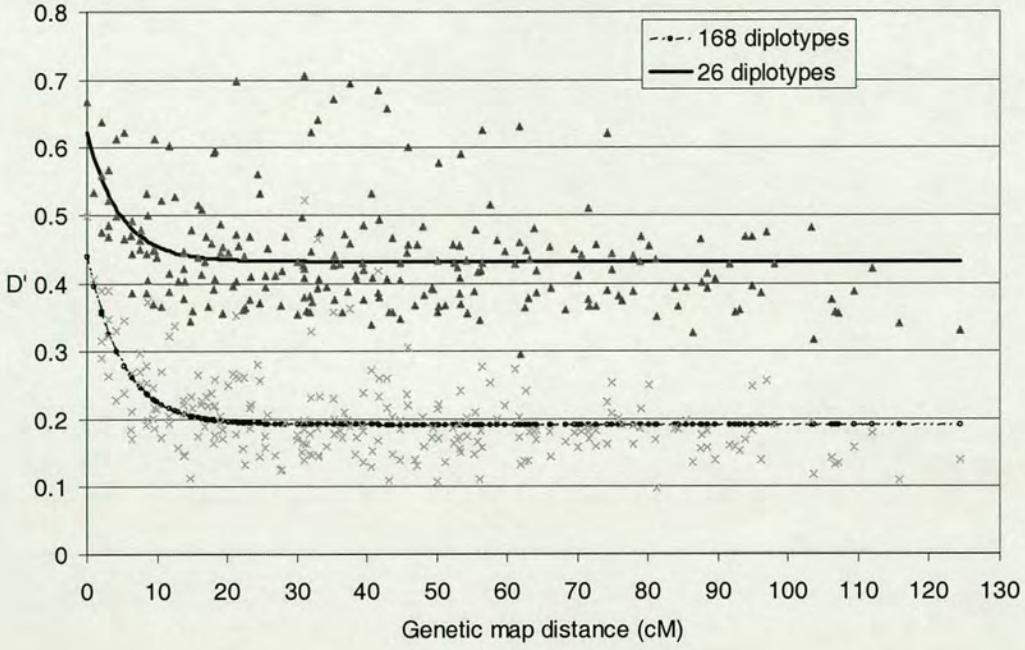


Figure 3.7 shows how small sample sizes tend to flatten the decay of D' with distance (i.e. plateaus sooner). Each point is the average value of D' obtained from bootstrapping 1000 samples of size 168 and 26 diplotypes from G2. The maximum and minimum fitted value for the results shown in Figure 3.7 were 0.44 and 0.192 for a sample size of 168 and 0.623 and 0.432 for a sample size of 26. This makes a difference between maximum and minimum value of 56% ([0.44-0.192]/0.44=0.56) and 30% respectively.

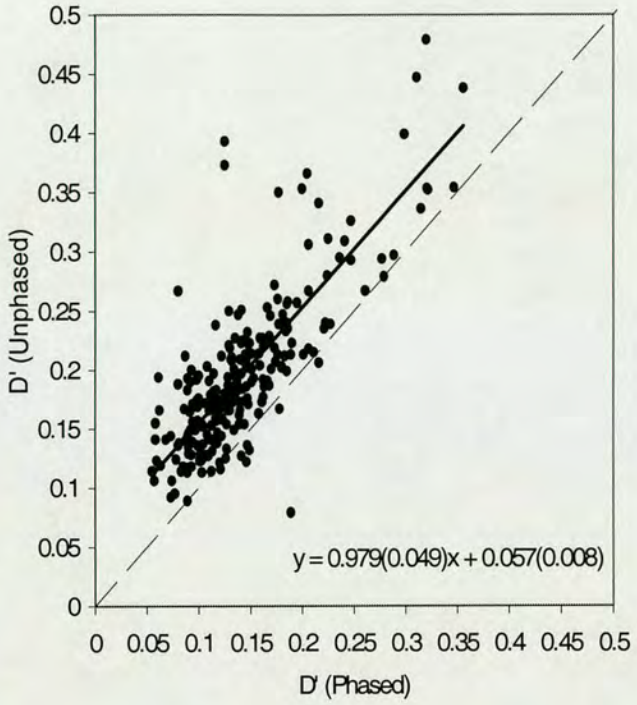
Figure 3.7. Effect of sample size on the decay of LD. Each dot is the average value obtained from bootstrapping 1000 samples of 168 (crosses) and 26 (triangles) diplotypes.



3.3.5 Extent of linkage disequilibrium using unphased founders

Two-locus maximum likelihood estimates of haplotype frequencies for the same founder individuals used in the previous section were obtained using the EM algorithm. D' values were computed. Figure 3.8 shows, for each pair of loci, the estimate of D' obtained from phased individuals (horizontal axis) and unphased individuals (vertical axis). Only 9 out of the 231 pairs showed a smaller D' obtained from unphased individuals than from phased individuals. The regression coefficient of D'(Unphased) on D'(Phased) was not significantly different from one and the intercept was significantly different from zero ($P < 10^{-12}$). The continuous line is the fitted line and the estimated parameters and their standard errors (in brackets) are shown in Figure 3.8.

Figure 3.8. Comparison of D' values obtained from phased and unphased individuals. The fitted line (continuos line) and its equation are shown. In the equation, standard errors (in brackets) follow the parameters estimates.



3.3.6 Effect of the pooling strategy

The effect of different allele pooling strategies is shown in Figure 3.9. Although only results for unphased individuals are shown the results obtained for phased individuals did not qualitatively differ from those shown here. Three different pooling strategies were considered: pooling of alleles with frequency smaller than 7% or 1% and not pooling at all. The pooling of rare alleles tended to reduce the value of D', and is therefore a conservative approach. How the statistical significance of LD changed with pooling was assessed for a range of pooling strategies. Table 3.6 compares the number of times in which the locus pairs were classified as having the same (or different) level of significance (as defined for Figure 3.3) when different levels of pooling were compared with no pooling. For example, if a pair of loci was in significant LD ($P = 0.05$) when not pooling and in highly significant LD ($P = 0.0002$) when pooling at 7%, then this pair added one value to 1 displacement in the 7% column, if it was the other way round, e.g. highly significant when not pooling and not significant when pooling at 10%, then it added one count on to the displacement of –2 in the relevant column.

Figure 3.9. Effect of different pooling strategies on the estimate of D'. Each point represents the average of D' values computed at 5 cM intervals.

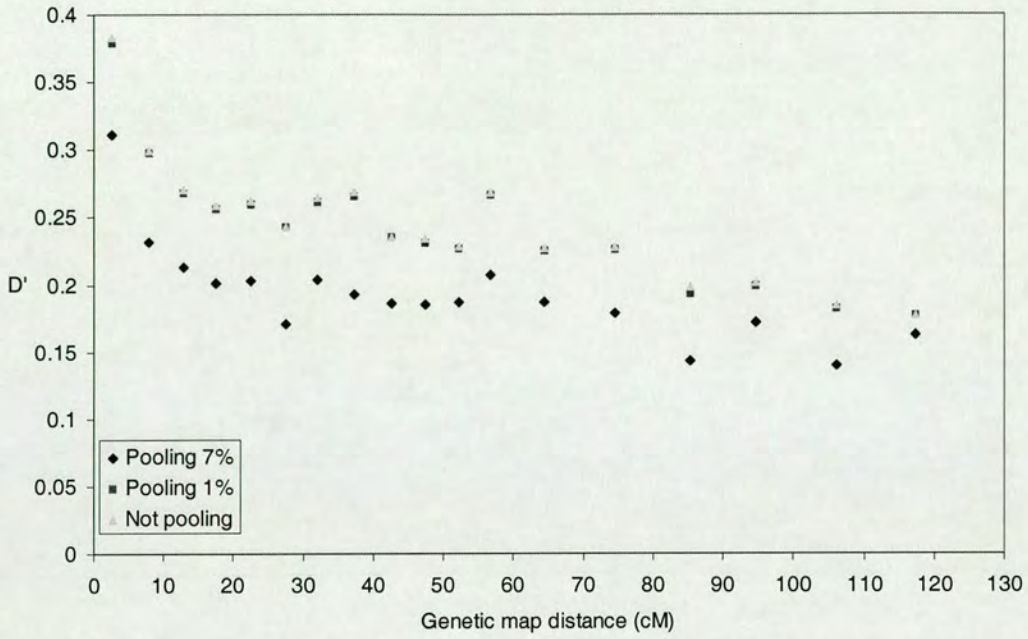


Table 3.6. Number of times the classification of the statistical significance for the 231 locus pairs changed when pooling at different proportions compared with no pooling. The displacement is negative when pooling is less significant than not pooling.

				0%			
Displacement	0.1%	1%	5%	7%	10%	25%	50%
-2	0	0	0	0	1	20	27
-1	3	1	13	19	36	70	77
0	227	201	169	170	162	129	124
1	1	29	48	42	32	12	3
2	0	0	1	0	0	0	0

The test for LD tended to became more conservative when the threshold frequency for pooling increased (i.e. as higher proportions were pooled D' frequently became less significant), although there was some variability for the smallest threshold frequencies for pooling that could not be accounted for.

3.4 DISCUSSION

I have estimated that LD on chromosome 19 in the Talana population extends between 8 and 11 cM. In addition, 71% of the locus pairs showed highly significant association when they were less than 20 cM apart, but only 6.5% when they were more than 20 cM apart. Isolated populations with high levels of LD generated by drift pose the problem of how to distinguish between LD due to close linkage and that generated by drift between unlinked loci (the background level of LD) that would lead to false positives. By chance, one would expect about 5% of the unlinked loci to show significant LD at the $P = 0.05$ level. There was only a slight increase (6.5%) on this, suggesting that the problem associated with high levels of background LD might be of little importance in this population. In order to assess this better, one would require microsatellite markers placed on other chromosomes. However, these data were not available and it was considered that distances of more than 20 cM were an appropriate threshold (that is, two loci 20 cM apart were considered to be effectively unlinked). If a marker locus, more than 20 cM apart from a trait locus, showed a significant association with the trait, then this information would be somehow limited for mapping the position of the trait locus with any confidence.

The conclusions on the extent of LD were based both on the statistic D' and on the significance level of the allelic association. Each criterion has advantages and disadvantages. On the negative side of using summary statistics such as D' are that small values cannot be interpreted as lack of significant association (Slatkin, 1994), that they are difficult to interpret and that their sampling distribution is usually unknown, and changes with parameters such as effective population size, recombination fraction between the two loci, allele frequencies and sample size (Hudson, 1985). On the positive side, their values are usually standardised so that their range of possible values is the same regardless of the allelic frequencies making comparisons easier across pairs, with different numbers and frequencies of alleles. The main advantage of using significance level is that it is easy to interpret. Its main disadvantage is that it depends on the marginals of the contingency table (number and frequency of the alleles and sample size). The fact that conclusions based on both methods are similar suggests that the present estimates are quite robust.

An alternative LD estimator, the multi-allelic R^2 (Hudson, 1985), was estimated (results not shown). The decay of LD with genetic map distance was very similar for both R^2 and D'. The regression of R^2 on D' showed a slope not significantly different from one and an intercept significantly ($P < 0.001$) different from zero. This suggests that both estimators decay at similar rates but differ in their mean values.

This study also suggests that the average level of LD in this population might be greater than in other isolates, as well as in other larger outbred populations. For example, Huttley *et al.* (1999) used 5048 autosomal short tandem repeat poylmorphism (STRP) scatted across the genome and found that about 4% of the locus pairs separated by less than 4 cM were in LD for the European Utah and Amish CEPH families. Zavattari *et al.* (2000) studied the extent of LD in a sub-isolate of Sardinia (the village of Gavoi) on the same region on the long arm of chromosome X studied by Laan and Paabo (1997) in Finns, Estonians, Swedes and Saami. They found similar levels of LD in Gavoi and in the Saami; respectively 19/21 and 17/21 of the pairs were in significant LD within a region of ~10 cM (9-11.5 Mb). In the present study, similar levels of LD to those of the Saami or Gavoi were found, that is, 25/30 pairs in significant LD ($P<0.0002$) for distances ≤ 10 cM. In a previous study of Talana spanning a region of 11.1 cM in X13q3, it was found that 6/15 markers pairs were in significant LD (Angius *et al.*, 2002a). The larger proportion of loci pairs in LD found in this chapter compared to that on the same population are probably due to the larger sample size of the present study. Nevertheless, one must exercise care when comparing measures of LD across studies with, for example, different sample sizes, genome regions, marker informativeness and density or haplotyping methods and interpret them just as a rough estimate of what one might expect if all these factors were the same.

The effect of two different strategies for inferring population haplotype frequencies (i.e. estimating them with and without family information) on the level of LD was studied. I compared D' when estimated from phased and unphased individuals and found that estimates from phased individuals yielded lower estimates than from unphased individuals. This could be due to the fact that *Simwalk2*, used to obtain haplotypes using family information, assumes that the loci are in linkage equilibrium (LE) and the EM algorithm, used to estimate population haplotype frequencies, does not. Moreover, the EM algorithm is expected to work better when the amount of LD increases (Fallin and Schork, 2000) whereas those programs that assume LE are expected to perform worse as the amount of LD increases (Schaid *et al.*, 2002). In the present case, there was a large difference between the average extent of LD (measured as D' or measured as the significance level of the allelic association) when using phased and unphased individuals. I repeated the analysis to find the extent of LD as shown in Figure 3.2 and 3.3 (results not shown) but using haplotype frequencies obtained from unphased individuals and found that the average extent of LD was between 4 cM (using the half-length of D') and 6.6 cM (1.12*5.92, where 1.12 is the mean number of adjacent markers in highly significant LD estimated from haplotype frequencies obtained without using family information and 5.92 is the average distance between markers. The estimated

standard deviation was 4.59 cM.). Although values of D' tend to be larger when using unphased individuals their decay is faster and therefore the estimate of the useful extent of disequilibrium shorter. The differences observed in the estimate of the extent of LD based on the significance level of the allelic association, when using phased and unphased individuals (that is, 11cM versus 6.6 cM), could be due to the methods employed to test it. When using phased individuals one is assuming that one can count the haplotypes and apply a standard chi-squared test. However, what is counted is only an estimate of the haplotypes, not the haplotypes themselves (that is, one is assuming the haplotypes as known but they are only estimated with some degree of confidence that is not incorporated in the testing procedure). On the other hand, when using unphased individuals one compares the likelihood of the data under the assumption of LE and LD.

There have been suggestions (Slatkin *et al.*, 1994; Huttley *et al.*, 1999; Varilo *et al.*, 2003) that the locus heterozygosity might affect the ability to detect LD. In order to test that, the mean heterozygosity of the locus pairs was regressed on their significance level of LD. The slope was found to be significantly different ($P<0.001$) from zero. Mean heterozyosity only accounted for about 6% of the total variance and I therefore considered that its effect on the significance level would be negligible and did not correct for it.

Pooling of rare alleles is usually done just for statistical reasons and alleles are pooled only with regard to their frequency. A more desirable approach would be pooling microsatellites alleles with regard to biological reasons. For example, it might be more appropriate to pool the lower frequency alleles of a microsatellite locus to those alleles with the closest length rather than for allele frequency because, in a step-wise mutation model, alleles are assumed to mutate by increasing or decreasing the length of the repeated motif of base pairs by one.

As shown above, an important factor that influences estimates of LD is the sample size, which in this study varied for the different pairs of loci between 312 and 606 haplotypes. However, Figure 3.6 shows that the mean and variance of D' are similar if the number of diplotypes ranges between 150-300 diplotypes and, therefore in this study, the effect of sample size is expected to be negligible when comparing across pairs of loci.

Estimates of D' from the G1, G2 and G3 founders showed significant differences. These could be explained by the difference in sample size, rather than by the difference in the number of generations available to estimate diplotypes.

In conclusion, Talana has levels of LD similar to those of the Saami and other Sardinian sub-isolates such as Gavoi. The estimated extent of LD and its variance is highly dependent on the sample size from which it has been estimated. Small samples tend to

overestimate the amount of disequilibrium, in this study up to a factor of almost three. Researchers should, therefore, exercise care when planning LD mapping studies based on the amount of LD found from a preliminary study based on a small sample individuals. When studying the extent of LD disequilibrium as a preliminary study for mapping purposes researchers are recommended to use about 200 unrelated individuals and to use both the significance level of the allelic association and D' to interpret their results. It is also recommended when comparing levels of LD to account for differences in sample size using bootstrapping as shown in Figure 3.1 for the number of alleles.

CHAPTER 4 - Power of linkage disequilibrium mapping to detect a quantitative trait locus (QTL) in selected samples of unrelated individuals

4.1 INTRODUCTION

Quantitative traits are those measured on a continuous scale. They are complex because there is not a simple relationship between genotype and phenotype, and may be of interest to human geneticists because they may be easier to collect than binary disease traits and are correlated with disease status. For example, a patient with ischaemic heart disease is generally treated and controlled by his/her blood pressure or cholesterol level, but rarely directly for the heart condition. Linkage disequilibrium (LD) is defined as the non-random association of population allele frequencies at two or more loci (Ayres and Balding, 2001), and is used at the population level to map trait loci. If a marker and trait locus are in LD, then the marker locus will be associated with the phenotype controlled by the trait locus. However, the ability to detect an association between a given allele at a marker locus and a trait depends on the amount of LD between the two loci. Although theoretically one could predict the amount of LD between two loci as a simple function of the physical distance between them (Hartl and Clark, 1997), empirical studies show that this relationship is not simple (Daly *et al.,* 2001; Jeffreys *et al.,* 2001), suggesting that the distribution of LD in the region of interest must be carefully studied before a statistically significant or non-significant association is reported, because the former does not always imply close linkage (e.g. significant LD can arise between non-syntenic loci) and the latter does not always imply a lack of it. Population stratification can generate significant LD between non-syntenic loci and, hence, false positives. Using family data rather than unrelated cases and controls overcomes the problem of population stratification because case and control samples are obtained from the same genetic background and contrasts are done within families and not across families. However, family-based designs are not always possible, especially for late onset traits in which parental data are often unavailable.

In the absence of parental data, the use of unrelated cases and controls is an appealing alternative provided that the possibility of population stratification can be ruled out or the effects of population structure can be eliminated. Pritchard *et al.* (2000a, 2000b) showed that population structure can be inferred using a set of unlinked markers and individuals assigned to different subpopulations. Testing within subpopulations or taking into account the average level of association observed throughout the genome, e.g. by using

a genomic control (Devlin and Roeder, 1999), would make it possible to allow for false positives due to population stratification.

Selective genotyping is the term used when individuals only from the upper and lower tail of the trait phenotypic distribution are genotyped (Lander and Botstein, 1989; Darvasi and Soller, 1992). This strategy is efficient and powerful under some circumstances (Allison *et al.,* 1998) because most of the information resides in individuals with extreme phenotypes (Carey and Williamson, 1991). It is especially useful when the cost of genotyping is much greater than the cost of collecting phenotypes and when a single phenotype is studied.

Schork *et al.* (2000) studied the power to detect a trait-marker association using individuals sampled from the upper and lower tails of the quantitative trait phenotypic distribution. Both marker and QTL were assumed to be biallelic. The aim of this chapter is to investigate and predict power of LD mapping when the QTL and marker loci are multiallelic. In particular, I studied:

1) The effect on power when the QTL is assumed to be multiallelic as opposed to biallelic.

2) Two different and simple patterns of LD to investigate their influence on power.

3) The economically optimal proportion of the quantitative trait (QT) distribution selected for a given power depending on the relative cost of genotyping and phenotyping.

4.2 MATERIAL AND METHODS

Individuals sampled from the tails of the trait distribution are classified as upper or lower tail depending on whether their trait value is respectively greater or less than a given threshold. The study design for a practical case would be: (1) to phenotype a number of individuals for a quantitative trait; (2) to select individuals with extreme phenotypes (e.g. the 10 % upper and lower values for the quantitative trait) to be genotyped; (3) to compare the counts of the different alleles at a locus in the upper and lower tails.

4.2.1 Genetic model

Consider a locus with an arbitrary number of alleles that contributes to the genetic component of a quantitative trait. Alleles at the locus are labelled as Q_i. With n alleles at a locus there are $n(n+1)/2$ possible genotypes and the same number of genotypic values. The population frequency of allele Q_i is labelled q_i. The genotypic value (G_{ij}) for genotype Q_iQ_j is parameterised as:

$$G_{ij} = G_{ji} = G_{ii} + k_{ij} \times (G_{jj} - G_{ii});\ i<j;\ i \in [1,n-1];\ j \in [2,n];\ k_{ij} \in [0,1] \qquad [1]$$

where k_{ij} provides a measure of dominance between alleles Q_j and Q_i. If $k_{ij} = 0$, Q_i is dominant to Q_j; if $k_{ij} = 0.5$, Q_i and Q_j act additively; and if $k_{ij} = 1$, Q_i is recessive to Q_j. The difference between the genotypic value of the Q_jQ_j and Q_1Q_1 genotypes is represented as $2a_j$ (where $G_{11} = 0 = 2a_1$; $j \in [2,n]$) and is expressed in residual standard deviations.

4.2.2 Mixture model

Assuming there are $n(n+1)/2$ genotypes with normally distributed phenotypes, the observed joint phenotypic distribution is a weighted average of the underlying normal distributions. The probability density function for a mixture of normals is:

$$\rho(x) = \sum_{i=1}^{n} \sum_{j=1;j\geq i}^{n} f_{ij} \varphi(x \mid \mu_{ij}, \sigma_{ij}^2) \qquad [2]$$

where f_{ij} is the frequency of genotype Q_iQ_j, μ_{ij} is the mean value for genotype Q_iQ_j, σ_{ij}^2 is the variance in trait values for individuals with genotype Q_iQ_j (within genotype variance) and $\varphi(x/\mu,\sigma^2)$ is the normal probability density function with mean μ and variance σ^2. The locus is assumed to be in Hardy-Weinberg equilibrium with frequencies q_i^2 for

homozygous Q_iQ_i genotypes and $2q_iq_j$ for heterozygous Q_iQ_j genotypes. Without loss of generality, the within-genotype variance (σ^2_E) is assumed to be 1.

When the QTL effect is small, then the observed joint distribution can be approximated by a single normal distribution with mean and variance equal to:

$$\mu_{Pop} = \sum_{i=1}^{n} \sum_{j=1;j\geq i}^{n} \mu_{ij} f_{ij} \quad ; \quad \sigma^2_{Pop} = \sigma^2_G + \sigma^2_E$$

where σ^2_G is the genetic variance due to the QTL and σ^2_E =1 as above. Although all results shown in this work were performed assuming a mixture distribution, the approximation to a normal gave practically the same results for the range of QTL effects considered.

4.2.3 Selecting individuals from the upper and lower tails

It is assumed that one is interested in the QTL allele that is associated with the highest genotypic value and that $2a_1<2a_2<2a_3<...<2a_n$. This seems reasonable when doing selective genotyping because selection of individuals in opposite tails will produce an enrichment of the QTL allele frequencies that cause lower or higher trait values relative to a random sample of individuals. The selected fractions in the upper and lower tails are α_U and α_L respectively with corresponding truncation points τ_U and τ_L. The latter were obtained by solving the following non-linear equations using Newton's method as described in Ducrocq and Quaas (1988):

$$\alpha_U = \sum_{i=1}^{n} \sum_{j=1;j\geq i}^{n} f_{ij} \left[1 - \Phi\{(\tau_U - \mu_{ij})/\sigma_{ij}\}\right] \qquad [3]$$

$$\alpha_L = \sum_{i=1}^{n} \sum_{j=1;j\geq i}^{n} f_{ij} \left[\Phi\{(\tau_L - \mu_{ij})/\sigma_{ij}\}\right] \qquad [4]$$

where $\Phi(\tau)$ is the cumulative standard normal distribution.

Using Bayes' theorem, the conditional probability of sampling a Q_i allele given that individuals have been sampled from the upper α_U percentile of the trait distribution can be written as:

$$P(Q_i \mid x > \tau_U) = \frac{P(x > \tau_U \mid Q_i)P(Q_i)}{P(x > \tau_U)} = \frac{\sum_{j=1}^{n} P(x > \tau_U \mid Q_i Q_j)P(Q_i Q_j \mid Q_i)P(Q_i)}{P(x > \tau_U)}$$

$$= \frac{q_i \sum_{j=1}^{n} q_j \left(1 - \Phi\left(\frac{\tau_U - \mu_{ij}}{\sigma_{ij}}\right)\right)}{\alpha_U} \qquad [5]$$

Equivalent probabilities can be computed for samples from the lower tail.

$$P(Q_i \mid x \leq \tau_L) = \frac{q_i \sum_{j=1}^{n} q_j \left(\Phi\left(\frac{\tau_L - \mu_{ij}}{\sigma_{ij}}\right)\right)}{\alpha_L} \qquad [6]$$

Note that in equations [5] and [6] it is not longer assumed that $i \leq j$ as in equation [2].

4.2.4 LD between trait and marker loci

In most cases genotypic information is obtained on marker loci rather than on the trait locus itself. For instance, one could genotype individuals for a number of marker loci scattered across the whole genome and test for an association between marker status at each locus and phenotype. A statistically significant association between marker status and phenotype would suggest that there is statistically significant LD between marker and trait loci at the population level. This does not always imply linkage between the loci (e.g. significant LD can be found between non-syntenic loci due to stratification, drift, etc.), but it will be assumed in what follows that close linkage is the cause of the LD.

Consider a marker locus with an arbitrary number of alleles, linked to the trait locus and in LD with it. It is assumed that under the null hypothesis the marker locus is in Hardy-Weinberg equilibrium. One requires this, because one is assuming that each of the two marker alleles that constitute the genotype is sampled independently. Under this design, one would sample alleles in pairs, i.e. the pair of alleles that form a genotype. Hence, the sampling of the two alleles could only be considered independent if the assortment of alleles at the marker locus was random, i.e. the marker locus was in Hardy-Weinberg equilibrium. The marker alleles are represented as M_h, with population frequency m_h. The disequilibrium

parameter (δ_{hi}) between marker allele h and QTL allele i is defined as $\delta_{hi} = f_{hi} - m_h q_i$ where f_{hi} is the population frequency of the haplotype $M_h Q_i$. Note also that the following conditions must be fulfilled:

$$\sum_{h=1}^{m} m_h = 1 \quad [7]$$

$$\sum_{i=1}^{n} q_i = 1 \quad [8]$$

$$\sum_{h=1}^{m} \delta_{h1} = \sum_{h=1}^{m} \delta_{h2} = \sum_{h=1}^{m} \delta_{h3} = \ldots = \sum_{h=1}^{m} \delta_{hn} = 0 \quad [9]$$

$$\sum_{i=1}^{n} \delta_{1i} = \sum_{i=1}^{n} \delta_{2i} = \sum_{i=1}^{n} \delta_{3i} = \ldots = \sum_{i=1}^{n} \delta_{mi} = 0 \quad [10]$$

The probability that a haplotype from an individual sampled from the upper tail (α_U) has an allele M_h is given by:

$$P(M_h \mid x > \tau_U) = \sum_{i=1}^{n} P(M_h \mid Q_i) P(Q_i \mid x > \tau_U) = \sum_{i=1}^{n} (m_h + \delta_{hi} / q_i) P(Q_i \mid x > \tau_U)$$

since

$P(M_h|Q_i) = P(M_h Q_i)/P(Q_i) = (m_h q_i + \delta_{hi})/q_i = m_h + \delta_{hi}/q_i$

using $P(Q_i|x > \tau_U)$ from [5].This reduces to:

$$P(M_h \mid x > \tau_U) = m_h + \sum_{i=1}^{n} (\delta_{hi} / q_i) P(Q_i \mid x > \tau_U) \quad [11]$$

and similarly

$$P(M_h \mid x \leq \tau_L) = m_h + \sum_{i=1}^{n} (\delta_{hi} / q_i) P(Q_i \mid x < \tau_L) \quad [12]$$

4.2.5 Linkage disequilibrium distribution patterns

Disequilibrium between the QTL allele with the greatest effect (Q_n) and the marker allele (M_m) is assumed to be positive ($\delta_{mn} > 0$). For convenience, it is assumed that this marker

is the one with the highest suffix (value of m). The disequilibrium parameter is expressed as a fraction of the maximum disequilibrium possible between the two alleles ($D'_{mn}=\delta_{mn}/\delta^{max}_{mn}$) (Lewontin, 1964) where δ^{max}_{mn} is:

$$\delta_{mn}^{\max} = \begin{bmatrix} \min\{m_m q_n, (1-m_m)(1-q_n)\}; \delta_{mn} < 0 \\ \min\{m_m (1-q_n), (1-m_m) q_n\}; \delta_{mn} > 0 \end{bmatrix} \quad [13]$$

In order to explore how the LD distribution affects the power to detect an association between a marker allele and trait status two different ways of fulfilling conditions [9] and [10] were studied as examples. The disequilibrium parameter was first computed as described above for element (m,n) and represented as δ_{mn}. For the first pattern studied, elements in column n were set equal to $-\delta_{mn}/(m-1)$ and elements in row m were all set equal to $-\delta_{mn}/(n-1)$. All other elements were set equal to $\delta_{mn}/(m-1)(n-1)$. This is the model assumed unless otherwise stated. For the second LD pattern, the first element δ_{mn} was computed as above and an element δ_{hi} was selected to be equal to δ_{mn}. Then all elements except δ_{mi} and δ_{hn} were assumed in equilibrium (i.e. $\delta_{hi} = -\delta_{mi} = -\delta_{hn} = \delta_{mn}$).

4.2.6 Calculation of power

Under the null hypothesis (H_0) of no association between a marker locus and a trait, the distributions of the marker alleles in the upper and lower tails are identical. This is tested using a contingency table with m rows and 2 columns, where the entries in the h^{th} row correspond to the numbers of M_h alleles ($h = 1, \ldots, m$), and those in the 1st and 2nd columns correspond to the numbers of alleles in the lower and upper tails respectively. The conventional statistic, X^2, based on this table is distributed under H_0 as chi-squared with $m-1$ degrees of freedom. Under the general alternative hypothesis (H_1), X^2 is asymptotically distributed as non-central chi-squared with $m-1$ degrees of freedom and non-centrality parameter, λ, given by

$$\lambda = N_L \sum_{h=1}^{m} \frac{(p_{Lh1} - p_{Lh0})^2}{p_{Lh0}} + N_U \sum_{h=1}^{m} \frac{(p_{Uh1} - p_{Uh0})^2}{p_{Uh0}} \quad [14]$$

where N_L, N_U denote the numbers of alleles sampled from the lower and upper tails respectively,

$$p_{Lh0} = Pr(M_h \mid x \leq \tau_L, H_0) \quad [15a]$$

$$p_{Uh0} = Pr(M_h \mid x > \tau_U, H_0) \quad [15b]$$

$$p_{Lh1} = Pr(M_h \mid x \leq \tau_L, H_1) \quad [15c]$$

$$p_{Uh1} = Pr(M_h \mid x > \tau_U, H_1) \quad [15d]$$

and the expressions on the right of equations [15a-d] are obtained by substituting appropriate values of δ_{hi} in equations [11] and [12] (Kendall and Stuart, 1961). Power is then defined as the probability that a non-central χ^2 with *m-1* degrees of freedom and non-centrality parameter λ is greater than the critical value defined by a central χ^2 with *m-1* degrees of freedom and significance level α.

4.2.7 Optimal proportion genotyped

The total cost depends on the numbers of individuals phenotyped (S_f) and genotyped ($S_g = (N_U+N_L)/2$) as well as the costs of phenotyping (K_f) and genotyping (K_g) per individual (Darvasi and Soller, 1992). Therefore, for a given power, the ratio S_g/S_f (= p, say) that minimises the cost can be determined. The total proportion selected to genotype (p) is equal to $\alpha_U + \alpha_L$. It is assumed in what follows that $\alpha_U = \alpha_L$ (i.e. that $p = 2\alpha_U = 2\alpha_L$). Although it may not always be optimal to set $\alpha_U = \alpha_L$, it is justified by the absence of prior knowledge concerning the model parameters. If $F(p)$ denotes the total cost, then

$$F(p) = K_g S_g + K_f S_f = K_f S_g \left(K + \frac{1}{p} \right) \quad [16]$$

where

$$K = \frac{K_g}{K_f} \quad [17]$$

Note that in [16] S_g is also a function of p. The value of p that minimises the cost function for a wide range of values of K was obtained numerically for values of p between 0.0001 and 1.

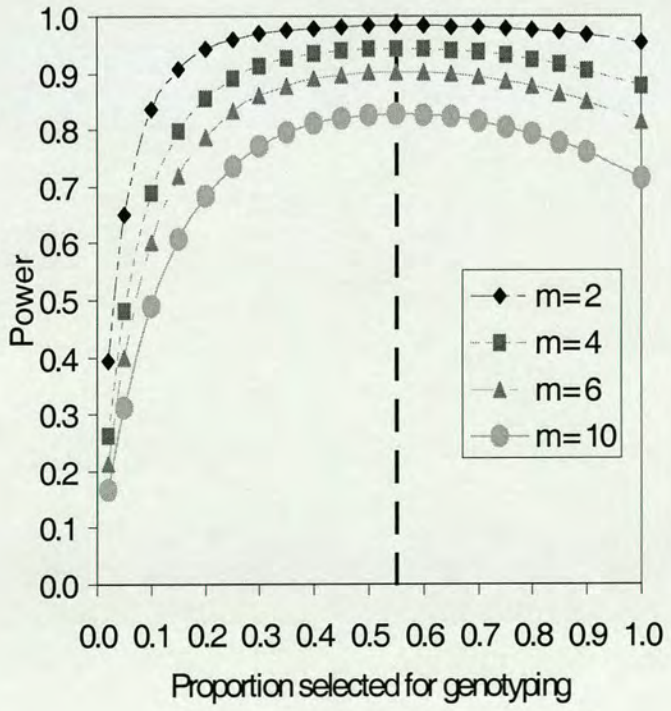
4.3 RESULTS

4.3.1 Effect of the number of individuals genotyped when the number of individuals phenotyped is fixed

Figure 4.1 shows how power increases as a larger proportion of the 2000 individuals phenotyped is genotyped until a maximum is reached (vertical dashed line in Figure 4.1) at 55% for markers with m = 2, 4, 6 and 10 alleles. The frequency of the m^{th} allele at the marker was kept constant in all situations considered; all other alleles were at equal frequencies, $m_h=(1-m_m)/(m-1)$. This assumption leads to the same amount of disequilibrium between M_m and Q_2, regardless of the number of marker alleles (Terwilliger, 1995).

Genotyping more than 55% of the individuals phenotyped leads to a decrease in power when using this type of test and therefore investigations will be restricted to moderate to high intensities of selection. This reduction in power is due to the increased amount of noise added by individuals in the middle of the quantitative trait distribution.

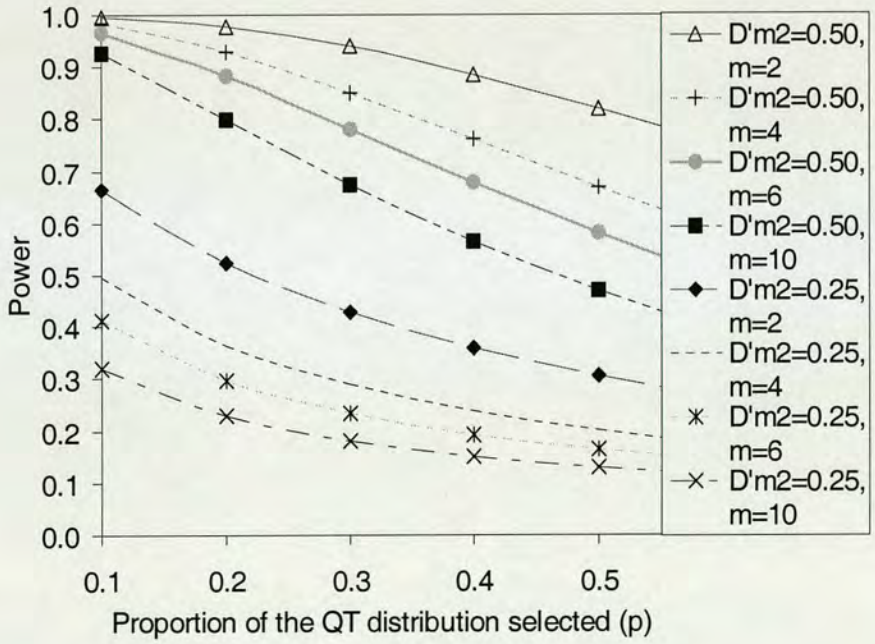
Figure 4.1. Effect of the proportion of individuals genotyped when the number of individuals phenotyped is fixed to 2000. Assumptions: additive model (k_{12}=0.5), biallelic QTL, equal proportions of individuals selected for genotyping from the upper and lower tail, significance level (α) = 0.05, $q_2=m_m$=0.1 and $m_h=(1-m_m)/(m-1)$, where m is the number of marker alleles and $h \in [1,m-1]$, a_2=0.5, h^2_{QTL}=0.043. The vertical dashed line represents the proportion selected that gives the highest power.



4.3.2 Effect of the number of marker alleles and proportion selected on power when the number of individuals genotyped is fixed

Figure 4.2 shows, for a biallelic QTL, how the number of marker alleles influenced power as a function of the proportion of individuals selected to be genotyped and the amount of disequilibrium. Note that in this case (unlike the previous section) the total number phenotypes measured increases as the proportion of the QT distribution decreases. With the total number of individuals genotyped fixed at S_g = 500, power decreased with increasing number of alleles at the marker locus as a result of the increase in the number of degrees of freedom for the X^2 test (df = m-1). Power also decreased as the proportion of the QT distribution genotyped increased. Power was similar (close to 100%) in all cases when the proportion selected was low and the amount of disequilibrium was high.

Figure 4.2. Effect of the proportion of individuals selected, amount of disequilibrium (D') and the number of marker alleles (m) on power. Assumptions: additive model (k_{12}=0.5), biallelic QTL, equal proportions of individuals selected for genotyping from the upper and lower tail, total sample size S_g=500 individuals, significance level (α) = 0.05, q_2=m_m=0.1 and m_h=$(1-m_m)/(m-1)$, where m is the number of marker alleles and $h \in [1,m-1]$, a_2=0.5, h^2_{QTL}=0.043.

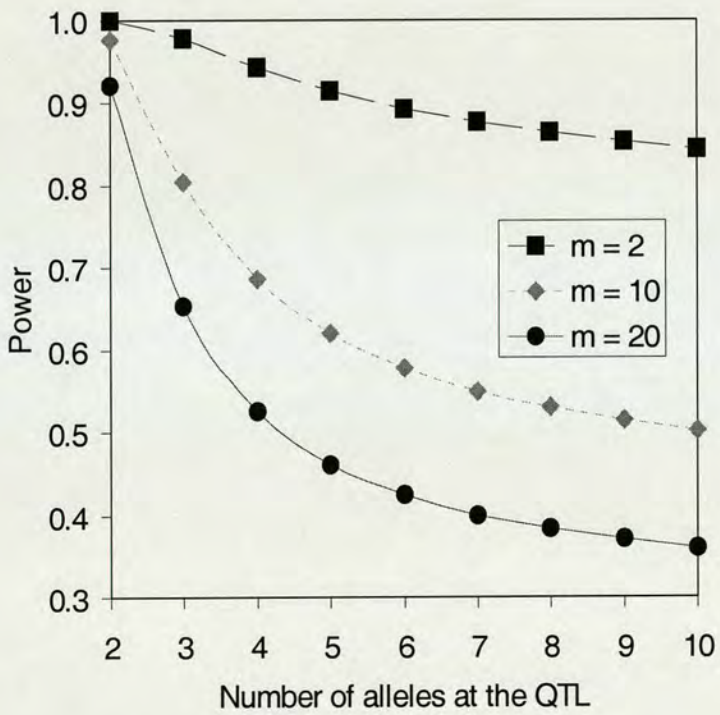


4.3.3 Effect of the number of QTL alleles on power

Figure 4.3 shows how the number of QTL alleles influenced power for a fixed number of individuals genotyped. The number of QTL alleles assumed varied from 2 to 10, and the difference in genotypic values between the two extreme homozygotes (that is, between Q_1Q_1 and Q_nQ_n) remained constant. For the other alleles, the increase in genotypic value of Q_iQ_i with respect to $Q_{i-1}Q_{i-1}$ was equal to $2*a_n/(n-1)$ for $i \in [2,n]$. Note that there are

infinite combinations of genotypic values that would lead to the same QTL heritability for fixed allele frequencies when the QTL is multiallelic. Results expressed in this way (as the difference between the two more extreme homozygous genotypes) have greater generality than a sample of all the possible genotypic combinations with the same QTL heritability. The marker locus was assumed to have 2, 10 or 20 alleles. In all cases considered, when the number of alleles at the QTL increased, power decreased. This reduction in power was larger with a higher number of marker alleles. Note that for the 20-allele marker, almost half of the reduction in power occurred when the number of QTL alleles increased from 2 to 3.

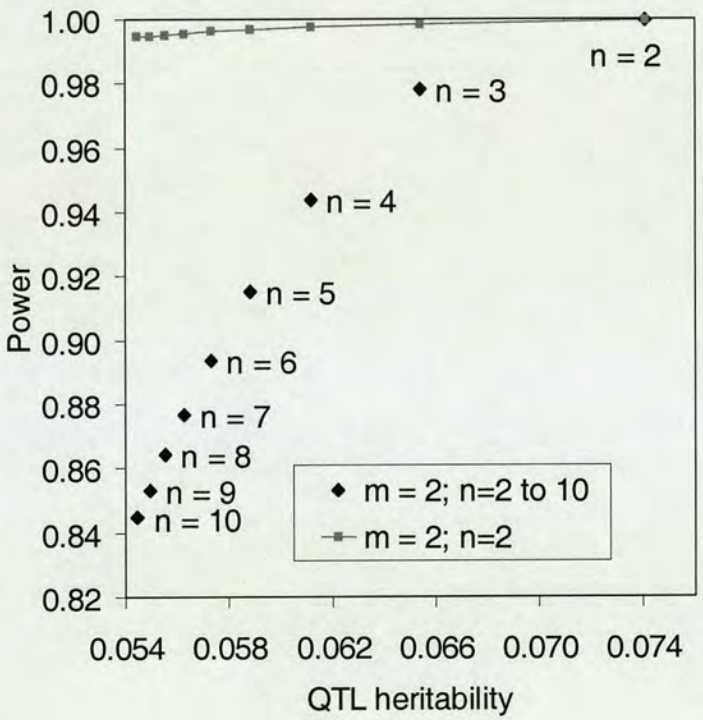
Figure 4.3. Effect of the number of QTL alleles on power. a_i is defined as $(i-1)a_n/(n-1)$, q_i is defined as $(1-q_n)/(n-1)$ where n is the number of QTL alleles, $i \in [1,n-1]$, a_n=0.5 and q_n=0.2. Marker allele frequencies were set to m_m=0.2 and $m_h = (1-m_m)/(m-1)$ where m (=2, 10 or 20) is the number of marker alleles and $h \in [1,m-1]$. A total of 500 individuals (S_g) were selected for genotyping from the upper and lower 10% QT distribution (N_U=N_L). The genetic model was assumed additive (k_{ij}=0.5), D'_{mn}=0.5, with the significance level (α) = 0.05.



The QTL heritability under the conditions assumed in Figure 4.3 varied with the number of alleles at the QTL. It showed a slight decrease with the increase in the number of alleles (h^2_{QTL}=0.074 for a biallelic QTL and h^2_{QTL}=0.054 for a 10-allele QTL). In order to check whether the difference in power was due to the increasing number of alleles or to the reduction in the locus heritability, the case of a QTL locus with two versus 3-10 alleles (keeping the same heritability as for the biallelic locus) was studied. Figure 4.4 shows how

power varied with heritability. The continuous line shows how power varied with heritability when the QTL was assumed biallelic. For a given heritability, the individual dots represent the power obtained for a QTL with different number of alleles. In all cases the marker was assumed biallelic, with $m_2=q_n=0.2$. The reduction in power with heritability was much larger when accompanied by an increase in the number of alleles at the QTL, showing that it was mainly due to the increase in the number of QTL alleles rather than to the reduction in heritability.

Figure 4.4. Effect of the number of QTL alleles on power. Comparison between a QTL with 2 alleles and a QTL with n alleles when the locus has the same heritability. a_i is defined as $(i-1)a_n/(n-1)$, q_i is defined as $(1-q_n)/(n-1)$ where n is the number of alleles of the QTL, $i \in [1,n-1]$, $a_n=0.5$ and $q_n=0.2$. Marker was assumed biallelic and allele frequency was set to $m_2=0.2$. A total of 500 individuals (S_g) was selected for genotyping from the upper and lower 10% QT distribution ($N_U=N_L$). The genetic model was assumed additive ($k_{ij}=0.5$), $D'_{2n}=0.5$, with the significance level (α) = 0.05.

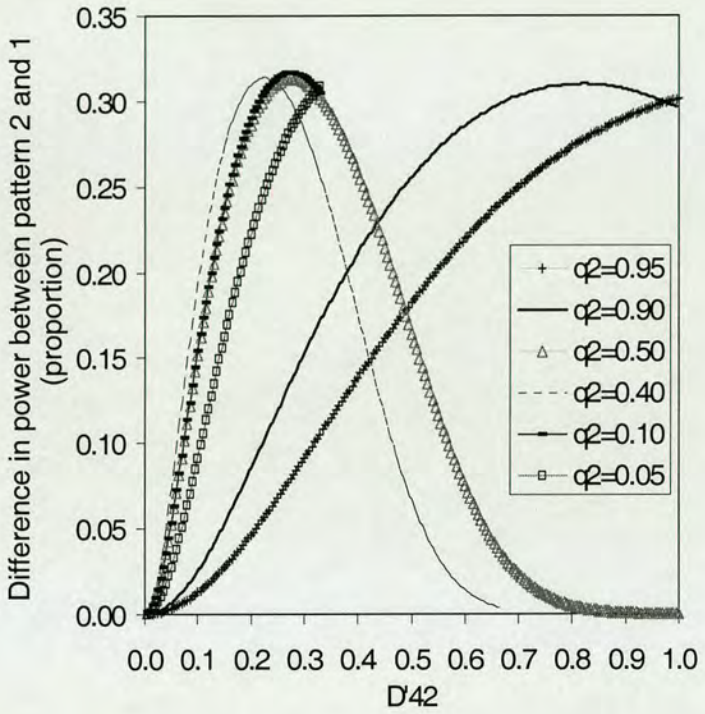


4.3.4 Difference in power between patterns of LD

Figure 4.5 shows the difference in power between LD patterns 1 and 2 for a biallelic QTL and a marker with 4 alleles, as a proportion of the power obtained with pattern 2 (the more powerful of the two). In this case both LD patterns had an equal total amount of disequilibrium as measured by Hedrick's D' (Hedrick, 1987). The maximum difference between patterns was about 30%, regardless of the QTL frequency. Differences in power increased with D'_{42} if q_2 was high or low, but if q_2 was intermediate differences in power

were maximum when D'_{42} had values that were intermediate for the range of possible values given the allele frequencies (q_2 and m_4).

Figure 4.5. Difference in power between patterns of LD 1 and 2 as a function of the amount of LD. The difference in power is expressed as a proportion of the power obtained for pattern 2 ($\delta_{11} = -\delta_{41} = -\delta_{12} = \delta_{42}$).). A total of 2000 individuals were selected for genotyping (S_g) from the upper and lower 10% QT distribution (N_U=N_L). A biallelic QTL was assumed with locus h^2_{QTL}=0.02. The genetic model was additive (k_{12}=0.5), significance level (α) = 0.05. Marker locus assumed to have four equally frequent alleles.

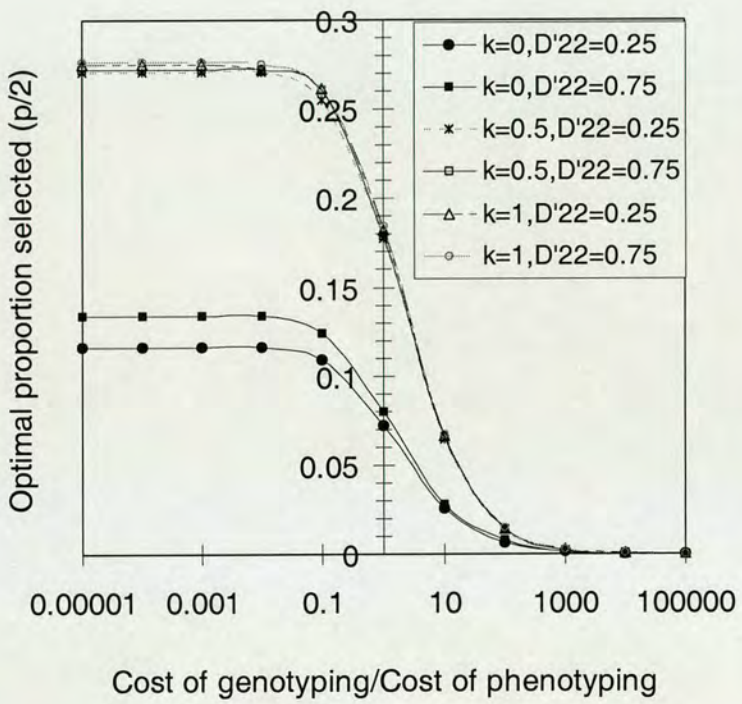


4.3.5 Optimum selected proportion

Figure 4.6 shows how the relative cost of genotyping and phenotyping influenced the proportion of individuals selected to be genotyped in order to achieve cost-effectiveness. Two levels of LD and three genetic models were studied for a biallelic QTL and a biallelic marker. The power was fixed at 80% and the QTL heritability was kept constant for all the models. Figure 4.6 illustrates how for the cases studied it would not be worthwhile to genotype more than the upper and lower 27.5% of the individuals phenotyped if the genetic model were additive or dominant, and 12.5% of the individuals when the genetic model were recessive, even when the cost of genotyping these individuals was 100-100000 times less than that of phenotyping ($K < 0.01$). This is because most of the information comes from individuals with extreme phenotypes, so that genotyping less informative individuals

produces no increase in power. For example, 80% power for the parameters shown in Figure 4.6 and D'_{22} equal to 0.75 could be obtained by genotyping and phenotyping 1100 individuals (p=1) or phenotyping 872 individuals and genotyping the upper and lower 218 individuals (p=0.5).

Figure 4.6. Optimum selection proportion for a power of 80 % as a function of the relative costs of genotyping and phenotyping. Different genetic models [recessive (k_{12}=0), additive (k_{12}=0.5) and dominant (k_{12}=1)] and amount of disequilibrium (D'_{22}) were assumed. The same proportions and the same number of individuals were selected for genotyping from the upper and lower tails ($p/2 = \alpha_L = \alpha_U$). The QTL and the marker were assumed biallelic ($q_2 = m_2 = 0.2$), locus $h^2_{QTL} = 0.02$, significance level (α) = 0.05. The horizontal axis is on the logarithmic scale.



The optimal proportion selected was always largest for the dominant model and smallest for the recessive model when the favourable allele (Q_2) was the less frequent one. This suggests that, for the recessive model and when the frequency of the Q_2 allele is small, the most extreme individuals of the trait distribution must be genotyped in order to increase the frequency of Q_2 alleles in the upper tail. By doing so, the relative frequencies of the individuals Q_1Q_1 and Q_1Q_2 with a positive deviation from their genotypic mean are reduced and the relative frequency of Q_2Q_2 with positive deviations are increased in the upper tail. The level of LD affects the optimal proportion selected. Increasing amounts of LD produced an increase in the optimal proportion selected for all the models of inheritance, and this increase was much more apparent in the recessive model than in the others.

When the less frequent allele was dominant or additive, then the optimal proportions selected were relatively insensitive to variations in heritability. For example, for an additive model and D'_{22}=0.5 with h^2_{QTL} values ranging from 0.01 to 0.1, the optimal proportion of individuals to be genotyped varied from p=0.538 to p=0.566 for K<0.01 (results not shown). When the less frequent allele was recessive, then the optimal proportion selected decreased with increasing heritability (results not shown).

4.4 DISCUSSION

Quantitative genetics theory is commonly applied under the simplified assumption that loci are biallelic. In this study, power to detect an association between a marker and a trait has been explored and quantified for multiallelic QTL and markers. Although others have previously noted that there may be loss of insight when the assumption that loci are biallelic is made (Nielsen and Weir, 1999), to my knowledge this has not been quantified. Conclusions are restricted to moderate-high intensities of selection because, when selecting individuals from the upper and lower tail of the trait distribution, one is dichotomizing the quantitative trait and therefore ignoring the information within each tail. This loss of information decreases as the selection intensity increases. The loss of information is quantified in Chapter 5.

Results shown here are based on the assumption that asymptotic conditions hold, i.e. that sample sizes are sufficiently large. Spurious results can arise if the sample size and/or some of the marker allele frequencies are small. However, it was found that relatively large sample sizes are necessary to obtain a reasonable power and these are expected to be large enough for asymptotic assumptions to hold. Significance thresholds used in this study are insufficient for a whole genome scan, which would require greater stringency. However, this is merely a scaling factor that does not change the general conclusions.

For a given QTL heritability, there is a large difference in power depending on the number of QTL alleles, with the power decreasing with increasing numbers of alleles. This is important because it is usually assumed that the QTL is biallelic, whereas a number of empirical studies have shown that disease loci may have multiple alleles (Hugot *et al.,* 2001; Ogura *et al.,* 2001; Wright and Hastie, 2001). Therefore, calculations performed assuming a biallelic QTL can seriously overestimate the power.

Two patterns of LD were investigated. Although these were just examples and did not correspond to any particular population genetics model, they illustrate the differences in power that can be seen as a result of the pattern of LD rather than of the amount of disequilibrium as measured by D' (which was identical for the two patterns studied in Figure 4.5). LD patterns would probably differ from one population to another (and from one pair of markers to another) and depend on the population history. The present approach would not be more or less general than one assuming a given population genetics model.

Bader *et al.* (2001) obtained the optimal proportions selected for DNA pooling when the objective was to minimise the number of individuals to be phenotyped. Their results were similar to those shown here for the lowest cost ratios (cost genotyping/ cost phenotyping). If

the cost ratio approximates zero, then what is basically minimised is the amount of phenotyping required.

The optimal proportion of individuals selected to be genotyped decreases to very small values under some circumstances. This is more striking for the most realistic relative costs of genotyping and phenotyping (that is, $K>1$). As discussed by Lander and Botstein (1989) it is probably unwise to select less than the 5% tails of the trait distribution because very extreme phenotypes can be the result of inaccurate observation (outliers). For the recessive model, the optimum proportion to select (p) was always lower than this suggested threshold (that is $p=0.1$, in Figure 4.6) for a locus with $h^2_{QTL}=0.02$ when genotyping was 10 times more expensive than phenotyping. For the additive and dominant models the genotyping costs could be up to 10 to 50 times greater than the phenotyping costs for the optimum proportion to be greater than the suggested threshold ($p=0.1$). Therefore the most cost-effective proportion of individuals genotyped obtained from the present study for small QTL heritabilities and realistic cost ratios should be used with caution if the amount of phenotyping done is not large. For practical purposes, it is recommended to select about the 5% of both tails of the quantitative trait distribution, which corresponds to reasonable genotyping/phenotyping cost ratios for most quantitative traits.

CHAPTER 5 - Mapping quantitative trait loci using linkage disequilibrium: marker- versus trait- based methods

5.1 INTRODUCTION

In recent years, human geneticists have advocated the use of linkage disequilibrium (LD) at the population level to fine-map genes associated with complex diseases. The reasons are that traditional linkage methods offer poor resolution of the trait locus position (due to the small number of recombination events available in most human pedigrees), that have low power to detect associations between genes of small effect and complex traits, and that technological advances, such as high-throughput genotyping methods, are now available for typing large numbers of genetic markers (e.g. Single nucleotide polymorphisms (SNPs)) in large numbers of individuals which makes mapping methods that use population-wide LD feasible. At the same time there has been an increased interest in quantitative traits that are genetically correlated with disease status because they are generally more easily and more objectively measured than are binary traits (such as, disease status). However, geneticists sometimes dichotomise continuous traits in an attempt to classify individuals as affected or unaffected. Osteoporosis is a good example of this (Langdahl *et al.*, 2003). According to World Health Organization criteria, a person has osteoporosis if they have bone mineral density less than two and a half standard deviations below the young population mean. In this case, people in the lower tail of the trait distribution would be treated as cases and the rest of the population as controls. This dichotomizing effect is sometimes taken to the extent that only individuals with very extreme phenotypes are used (Little *et al.*, 2002; Angius *et al.*, 2002b). These are then treated as disease phenotypes, and the data analyzed using the appropriate linkage approach. This "dichotomizing" approach may be favoured because it mimics traditional disease mapping methods, allowing similar interpretation of results an the use of readily-available software. However, the price paid (as shown below), in terms of loss of statistical power, might in some cases, be too high. Although the "dichotomizing" approach might be justified from a practitioner point of view, allowing a decision to be made on whether a patient needs treatment, it is not always justified when trying to map the relevant trait loci.

Following Lebowitz *et al.* (1987), the "dichotomizing" approach for mapping quantitative trait loci (QTL) using LD at the population level will be referred to as the trait-based (TB) approach, and the "non-dichotomizing" approach as the marker-based (MB) approach. TB methods dichotomise individuals into two classes and, therefore, information

contained within each class is lost. In this chapter, this loss is quantified in terms of statistical power and it is shown that MB methods either outperform or, at worst, are equivalent to TB methods.

The TB approach compares allele (or genotype) frequencies in individuals selected from the two extremes of the trait distribution, whereas the MB approach compares the mean phenotypic value for each marker allele (or genotype) in the same individuals. For each of the TB and MB approaches, two tests were performed, based on an additive and a dominant model of analysis, respectively.

For the additive model of analysis, the TB and MB tests were respectively a χ^2 test and an F-test from the regression of phenotype on the number of a given marker allele, respectively. The χ^2 test (based on a *2x2* contingency table; 2 marker alleles and 2 tails) compared the allele counts in individuals selected from the two tails of the trait distribution. If the distribution of allele frequencies differed significantly in the two tails, then this suggested that the marker was in LD with the locus affecting the trait. In the regression analysis, the presence of a QTL in LD with the marker locus would lead to a non-zero slope of the regression line.

For the dominant model of analysis, the TB and MB tests were respectively a χ^2 test (based on a *3x2* contingency table; 3 possible marker genotypes and 2 tails) and an F-test from an ANOVA, respectively. The χ^2 test was based on the comparison of genotype counts in individuals selected from the two tails of the trait distribution. A significant difference in the frequency distribution of the marker genotypes between the upper and lower tails suggested the presence of a QTL influencing the trait in LD with the marker locus. The ANOVA tested whether the quantitative trait values for the marker genotypes of the selected individuals were different. Under the null hypothesis of no QTL in LD with the marker locus, the phenotypic values for the different marker genotypes would not differ significantly. In order to make a fair comparison of the two approaches (MB vs. TB), comparisons should be made only for tests with the same degrees of freedom (df). Comparisons were made between tests with two degrees of freedom (the ANOVA F-test and the χ^2 test based on genotype counts) and between tests with one degree of freedom (the regression F-test and the χ^2 test based on allele counts).

5.2 MATERIAL AND METHODS

It is supposed that the trait is influenced by a bi-allelic QTL with alleles Q_1 and Q_2, having frequencies q_1 and q_2 $(=1-q_1)$ respectively. It is also assumed that the QTL is in Hardy-Weinberg equilibrium and that phenotypic values for the three genotypes Q_1Q_1, Q_1Q_2 and Q_2Q_2 are normally distributed about mean values of $\mu_{11}(=-a)$, μ_{12} $(=ad)$ and μ_{22} $(=a)$ respectively, and with equal variances (taken to be 1). The QTL genotype Q_1Q_2 is considered to be phenotypically identical to Q_2Q_1 $(\mu_{12}=\mu_{21})$. However, they are distinguished to clarify the mathematical expressions below. The QTL heritability is defined as h^2_{QTL} $=V_A/(V_A+V_D+1)$ where V_A $(=2q_1q_2a^2[1+\{q_1-q_2\}d]^2)$ and $V_D(=[2q_1q_2da]^2)$ are respectively the additive and dominant variances for the QTL. If x denotes the phenotypic value of an individual, then the probability density function of x is

$$\rho(x)=\sum_{i=1}^{2}\sum_{j=1}^{2} q_i q_j \varphi(\mu_{ij},\sigma_{ij}^2)$$

where $\varphi(\mu,\sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2 (assumed, without loss of generality, to be 1).

The upper and lower tails of the distribution are defined, respectively, as the proportions α_U and α_L of the individuals phenotyped that are to be genotyped. The upper and lower cut-offs τ_U and τ_L were determined by solving:

$$\alpha_U = \int_{\tau_U}^{\infty} \rho(x)\,dx \; ; \quad \alpha_L = \int_{-\infty}^{\tau_L} \rho(x)\,dx$$

If $\Phi(\tau-\mu_{ij})=\int_{-\infty}^{\tau}\varphi(x\,|\,\mu=\mu_{ij},\sigma^2=1)dx$, then the probability that an individual selected from one of the tails of the trait distribution has a given QTL genotype is:

$$P(Q_iQ_j\,|\,x>\tau_U)=\frac{q_iq_j(1-\Phi(\tau_U-\mu_{ij}))}{\alpha_U} \; ; \; P(Q_iQ_j\,|\,x<\tau_L)=\frac{q_iq_j\Phi(\tau_L-\mu_{ij})}{\alpha_L} \; ; \; i,j\in[1,2]$$

and the frequency of the QTL alleles in the two tails of the trait distribution is:

$$P(Q_i \mid x > \tau_U) = \frac{q_i \sum_{j=1}^{2} q_j \left(1 - \Phi(\tau_U - \mu_{ij})\right)}{\alpha_U} \quad ; \quad P(Q_i \mid x < \tau_L) = \frac{q_i \sum_{j=1}^{2} q_j \left(\Phi(\tau_L - \mu_{ij})\right)}{\alpha_L} \quad ; \quad i \in [1,2]$$

where *i, j* denote the possible QTL alleles.

The expected values for QTL genotype Q_iQ_j, in the upper and lower tails are η_{ij}^{U} and η_{ij}^{L} respectively, where

$$\eta_{ij}^{U} = \frac{1}{\left(1 - \Phi(\tau_U - \mu_{ij})\right)} \int_{\tau_U}^{\infty} \frac{x}{\sqrt{2\pi}} e^{-\frac{(x-\mu_{ij})^2}{2}} dx = \frac{z_{ij}^{U}}{\left(1 - \Phi(\tau_U - \mu_{ij})\right)} + \mu_{ij} = i_{ij}^{U} + \mu_{ij}$$

$$\eta_{ij}^{L} = \frac{1}{\Phi(\tau_L - \mu_{ij})} \int_{-\infty}^{\tau_L} \frac{x}{\sqrt{2\pi}} e^{-\frac{(x-\mu_{ij})^2}{2}} dx = \frac{-z_{ij}^{L}}{\Phi(\tau_L - \mu_{ij})} + \mu_{ij} = i_{ij}^{L} + \mu_{ij}$$

where z_{ij}^{U} and z_{ij}^{L} are the ordinates of the appropriate Gaussian distribution $[\varphi(\mu_{ij}, 1)]$ at the cut-offs τ_U and τ_L respectively.

Since one does not usually genotype the QTL itself, it is assumed that there is a linked bi-allelic marker locus in LD with the trait locus, that the marker locus is in HWE under the null hypothesis, and that the marker locus does not have an independent effect on the trait. The marker locus has alleles M_1 and M_2 with frequencies m_1 and m_2 $(=1-m_1)$ respectively.

The disequilibrium parameter (D) between marker allele M_2 and QTL allele Q_2 is defined as $D = f_{Q_2M_2} - q_2m_2$ where $f_{Q_2M_2}$ is the population frequency of the haplotype Q_2M_2. Results were expressed as a function of Lewontin's normalised measure of disequilibrium D' (Lewontin, 1964). D' is the value of D expressed as a fraction of its maximum possible value, that is, D' $(=D/D_{max})$ where D_{max} is the minimum value of q_2m_1 or q_1m_2 (as D is assumed, without loss of generality, to be positive between alleles M_2 and Q_2).

The four possible QTL-marker haplotype frequencies are $P(Q_1M_1)=q_1m_1+D'\times D_{max}$, $P(Q_1M_2)=q_1m_2-D'\times D_{max}$, $P(Q_2M_1)=q_2m_1-D'\times D_{max}$ and $P(Q_2M_2)= q_2m_2+D'\times D_{max}$. Assuming random mating, the probability of each of the possible QTL-marker diplotypes is equal to the product of their component haplotypes (e.g. $P(Q_iM_l,Q_jM_n)=P(Q_iM_l)\times P(Q_jM_n)$). The probabilities of occurrence of each marker genotype in the upper and lower tails were obtained using Bayes' theorem. After some algebra:

$$P(M_lM_n \mid x > \tau_U) = \sum_{i=1}^{2}\sum_{j=1}^{2} P(Q_iM_l, Q_jM_n)\left(1-\Phi(\tau_U - \mu_{ij})\right)/\alpha_U; l,n \in [1,2] \quad [1]$$

$$P(M_lM_n \mid x < \tau_L) = \sum_{i=1}^{2}\sum_{j=1}^{2} P(Q_iM_l, Q_jM_n)\Phi(\tau_L - \mu_{ij})/\alpha_L; l,n \in [1,2] \quad [2]$$

where l, n denote the possible marker alleles. From equations [1] and [2] one can obtain the probabilities of occurrence of the k^{th} marker allele in the two tails.

$$P(M_k \mid x > \tau_U) = \sum_{l=1}^{2}\sum_{n=1}^{2} P(M_k \mid M_lM_n)P(M_lM_n \mid x > \tau_U)$$

$$P(M_k \mid x < \tau_L) = \sum_{l=1}^{2}\sum_{n=1}^{2} P(M_k \mid M_lM_n)P(M_lM_n \mid x < \tau_L)$$

where $P(M_k|M_lM_n)$ is 1,1/2 and 0 for $k=l=n$, $k=l\neq n$ or $k=n\neq l$ and $k\neq l=n$, respectively. The expected quantitative trait value for marker genotype M_lM_n in the selected sample is equal to:

$$E(x \mid M_lM_n) = \frac{\alpha_U \times P(M_lM_n \mid x > \tau_U) \times E(x \mid M_lM_n)^U + \alpha_L \times P(M_lM_n \mid x < \tau_L) \times E(x \mid M_lM_n)^L}{\alpha_U \times P(M_lM_n \mid x > \tau_U) + \alpha_L \times P(M_lM_n \mid x < \tau_L)}; \quad l,n \in [1,2]$$

where $E(x \mid M_lM_n)^U$ and $E(x \mid M_lM_n)^L$ are the expected quantitative trait values for marker genotype M_lM_n in the upper and lower tails (see Appendix).

Table 5.1 shows the non-centrality parameters for the four tests studied. Derivations are shown in the Appendix. The MB tests were based on an F_{n_1,n_2} distribution. However, when the denominator degrees of freedom are large ($n_2 \rightarrow \infty$) this distribution can be approximated to n_1^{-1} times a chi-squared distribution with n_1 degrees of freedom. The sample sizes required to detect a QTL with small effect (as considered here) are large, and I therefore considered the approximation to be valid, and referred all the results to a chi-squared distribution. This makes the comparison of the two approaches easier. Simulations were carried out to check that the approximations, shown in Table 5.1, were very close (results not shown).

Table 5.1. Non-centrality parameters for the tests studied. N_T is the total number of individuals genotyped.

	Test	df	Non-centrality parameter
Trait-Based	χ^2 alleles	1	$\frac{2\alpha_U\alpha_L N_T}{(\alpha_U+\alpha_L)^2}\times\sum_{k=1}^{2}\frac{\sum_{l=1}^{2}\sum_{n=1}^{2}\left\{\left(\sum_{i=1}^{2}\sum_{j=1}^{2}P(Q_iM_l,Q_jM_n)\times[(1-\Phi(\tau_U-\mu_{ij}))/\alpha_U-\Phi(\tau_L-\mu_{ij})/\alpha_L]\right)\times P(M_k\mid M_lM_n)\right\}^2}{\sum_{l=1}^{2}\sum_{n=1}^{2}P(M_lM_n)\times P(M_k\mid M_lM_n)}$
	χ^2 genotypes	2	$\frac{\alpha_U\alpha_L N_T}{(\alpha_U+\alpha_L)^2}\times\sum_{l=1}^{2}\sum_{n=1}^{2}\frac{\left\{\sum_{i=1}^{2}\sum_{j=1}^{2}P(Q_iM_l,Q_jM_n)\times[(1-\Phi(\tau_U-\mu_{ij}))/\alpha_U-\Phi(\tau_L-\mu_{ij})/\alpha_L]\right\}^2}{P(M_lM_n)}$
Marker-Based	Regression	1	$\frac{N_T}{\sigma_W^2}\times\frac{\left\{\sum_{l=1}^{2}\sum_{n=1}^{2}[x_{ln}-\bar{x}]\times[E(x\mid M_lM_n)-\mu]\times P(M_lM_n)\right\}^2}{\sum_{l=1}^{2}\sum_{n=1}^{2}[x_{ln}-\bar{x}]^2\times P(M_lM_n)}$
	ANOVA	2	$\frac{N_T}{\sigma_W^2}\times\sum_{l=1}^{2}\sum_{n=1}^{2}[E(x\mid M_lM_n)-\mu]^2\times P(M_lM_n)$

Φ_{ij}	$1-\Phi(\tau_U-\mu_{ij})+\Phi(\tau_L-\mu_{ij})$
$\bar{x}$	$\frac{1}{\alpha_U+\alpha_L}\sum_{i=1}^{2}\sum_{j=1}^{2}\{P(Q_iM_{2,}Q_jM_2)-P(Q_iM_{1,}Q_jM_1)\}\times\Phi_{ij}$
x_{ln}	$x_{11}=-1,\ x_{12}=x_{21}=0, x_{22}=1$
σ^2_W	$\frac{1}{\alpha_U+\alpha_L}\sum_{l=1}^{2}\sum_{n=1}^{2}\left\langle P(Q_iM_l,Q_jM_n)\times\{(1+\mu_{ij}^2)\Phi_{ij}+z_{ij}^U(\tau_U+\mu_{ij})-z_{ij}^L(\tau_L+\mu_{ij})\}-\left(\sum_{i=1}^{2}\sum_{j=1}^{2}[P(Q_iM_l,Q_jM_n)\times(\mu_{ij}\Phi_{ij}+z_{ij}^U-z_{ij}^L)]\right)^2\Big/\left(\sum_{i=1}^{2}\sum_{j=1}^{2}[P(Q_iM_l,Q_jM_n)\times\Phi_{ij}]\right)\right\rangle$
μ	$\sum_{l=1}^{2}\sum_{n=1}^{2}\frac{q_iq_j\times(z_{ij}^U-z_{ij}^L+\mu_{ij}\Phi_{ij})}{\alpha_U+\alpha_L}$
$P(M_lM_n)$	$\sum_{i=1}^{2}\sum_{j=1}^{2}[P(Q_iM_{l,}Q_jM_n)\times\Phi_{ij}]\Big/(\alpha_U+\alpha_L)$
$E(x\mid M_lM_n)$	$\sum_{i=1}^{2}\sum_{j=1}^{2}[P(Q_iM_{l,}Q_jM_n)\times(\mu_{ij}\Phi_{ij}+z_{ij}^U-z_{ij}^L)]\Big/\sum_{i=1}^{2}\sum_{j=1}^{2}[P(Q_iM_{l,}Q_jM_n)\times\Phi_{ij}]$

5.3 RESULTS

All the results shown assumed that selection was symmetric, so that $2\alpha_U = 2\alpha_L = P$. Figures 5.1,5.2 and 5.3 show that with equal degrees of freedom MB methods always performed better than TB methods. This was so regardless of the genetic model considered for the generation of the data. Power was practically the same for both approaches when selection was sufficiently intense. However, differences in power were important when the whole population was genotyped as shown in Table 5.2.

Figure 5.1. Comparison of the power obtained when using the TB and MB approach for different proportions selected (P). The marker is assumed to be the QTL (D'=1) with allele frequency 0.3, h^2_{QTL}=0.05 and d=1 (dominant model). The significance level was 10^{-8}, the total number of individuals genotyped was fixed to 200 and number phenotyped was 200/P.

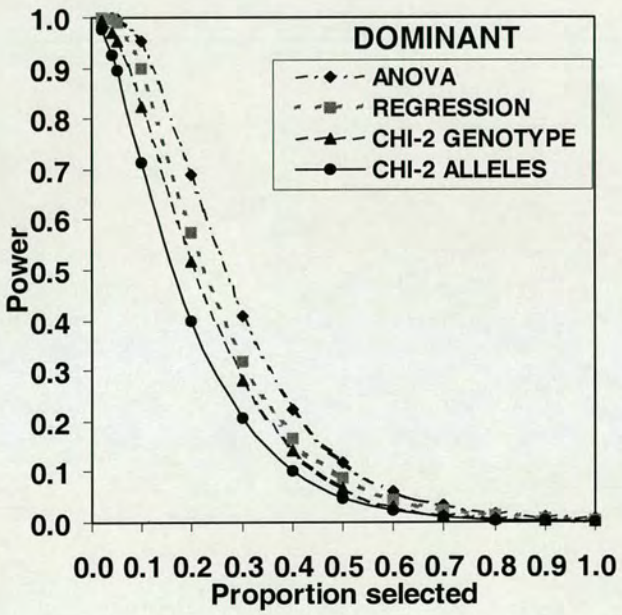


Figure 5.2. Comparison of the power obtained when using the TB and MB approach for different proportions selected (P). The marker is assumed to be the QTL (D'=1) with allele frequency 0.3, h^2_{QTL}=0.05 and d=0 (additive model). The significance level was 10^{-8}, the total number of individuals genotyped was fixed to 200 and number phenotyped was 200/P.

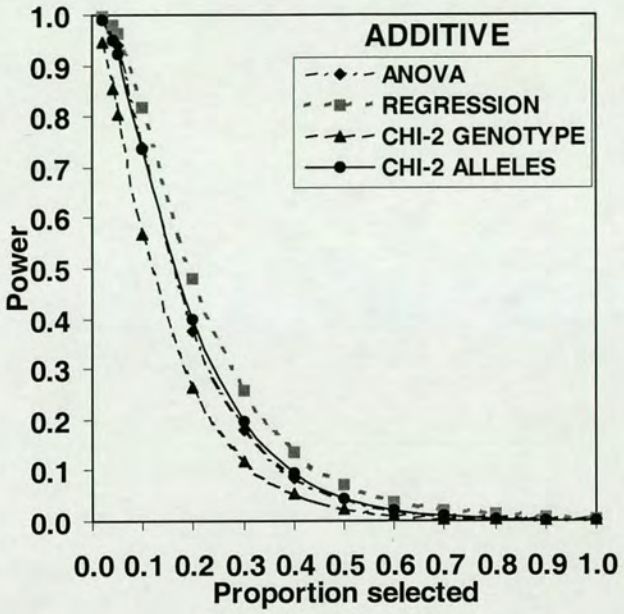


Figure 5.3. Comparison of the power obtained when using the TB and MB approach for different proportions selected (P). The marker is assumed to be the QTL (D'=1) with allele frequency 0.3, h^2_{QTL}=0.05 and d=-1 (recessive model). The significance level was 10^{-8}, the total number of individuals genotyped was fixed to 200 and number phenotyped was 200/P.

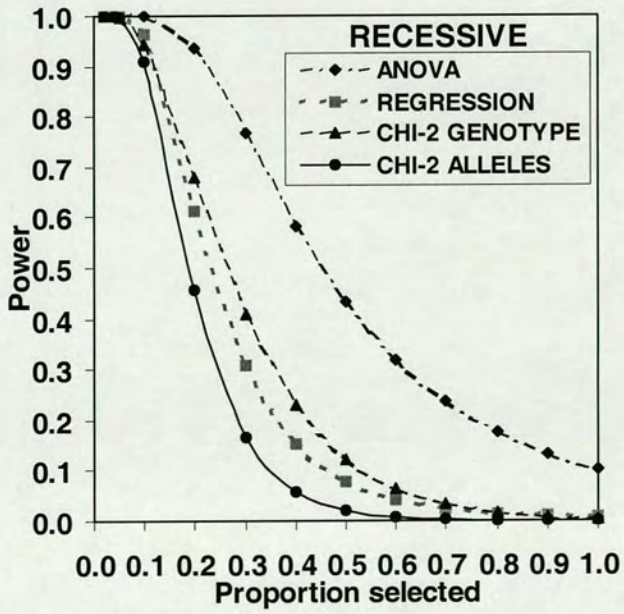


Table 5.2. Comparison of the power obtained with the TB or MB approach for different levels of disequilibrium when the whole population is genotyped. The marker and QTL were assumed to be in varying levels of disequilibrium (D'), $m_2=q_2=0.3$ and $h^2_{QTL}=0.05$. The significance level was 10^{-8} and the total number of individuals genotyped was 1000.

		Marker-based		Trait-based	
Model	D'	ANOVA (2 df)	Regression (1 df)	χ^2 Genotype (2 df)	χ^2 Alleles (1 df)
Dominant (d=1)	1.00	0.98	0.93	0.64	0.51
	0.75	0.37	0.35	0.08	0.08
	0.50	0.01	0.01	<0.01	<0.01
Additive (d=0)	1.00	0.88	0.93	0.38	0.48
	0.75	0.26	0.35	0.04	0.07
	0.50	0.01	0.01	<0.01	<0.01
Recessive (d=-1)	1.00	>0.99	0.96	0.91	0.23
	0.75	0.82	0.37	0.12	0.02
	0.50	0.03	<0.01	<0.01	<0.01

Figure 5.4 shows the effect of the amount of LD on power for three different intensities of selection. For extreme selection the power curves for the MB and TB approach almost overlapped regardless of the amount of LD. The difference between the two approaches was largest when no selection was applied.

Figure 5.4. Effect of the amount of LD on power of the MB and TB approach. The marker and the QTL both have allele frequency equal to 0.3, $h^2_{QTL}=0.05$ and the model is additive. The significance level was 10^{-8} and the total number of individuals genotyped was 1000. The proportion selected (P) is shown in the Figure.

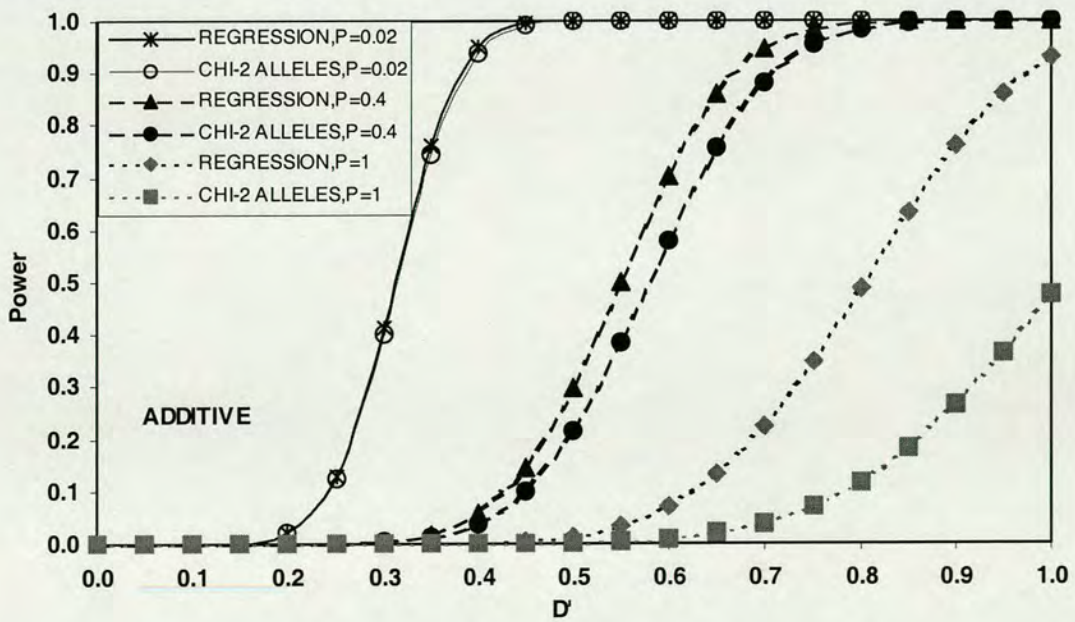
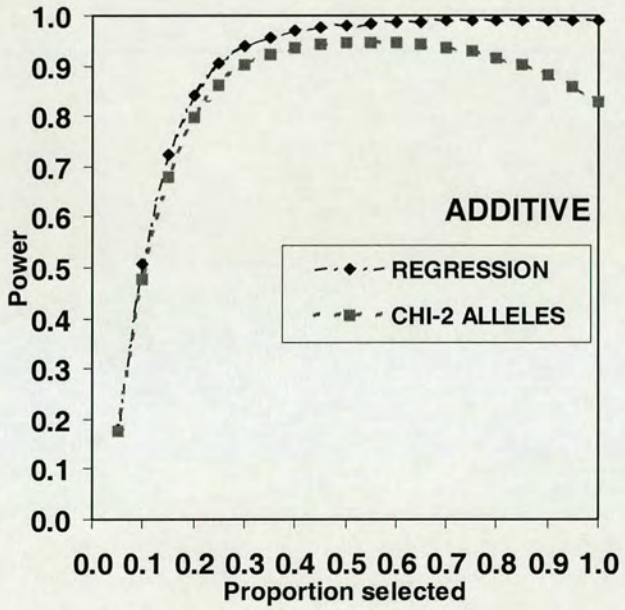


Figure 5.5 shows the power obtained for the MB and TB approach with 1df when the number of individuals phenotyped was fixed. The TB approach provides maximum power when about 27.5% of individuals are genotyped in each tail, and is less powerful than the MB approach. This level of selection (P=55%) provides the maximum power for the TB approach (in accordance with Lebowitz *et al.*(1987) and Bader *et al.*(2001)) but not for the MB approach, for which power increases monotonically with the number of individuals genotyped and included in the analysis.

Figure 5.5. Comparison of the power obtained when using the TB and MB approach with 1 df for different proportions selected (P) when the total number of phenotyped individuals is fixed. The total number of phenotypes is 4500 and the total number of genotypes is 4500×P. The marker is assumed to be the QTL (D'=1) with allele frequency 0.3, h^2_{QTL}=0.01 and the model is additive. The significance level was 10^{-5}.



5.4 DISCUSSION

Quantitative traits are of interest for human genetics because they are often correlated with disease traits. Schork *et al.*(2000) proposed the use of threshold-defined case/controls (that is, the TB approach with 1 df) for mapping loci influencing quantitative traits using LD at the population level. The objective of the present study was to assess how much information is lost when analyzing a quantitative trait as a threshold-defined binary trait as opposed to analyzing it using all the information available. The information lost in the former case is clearly reflected in a loss of statistical power to detect an association between marker genotype and phenotype. The proposed method of analysis was more powerful, except under very extreme selection, when both methods performed similarly. The method can be implemented with standard statistical packages or spreadsheets.

The results obtained from the TB approach are valid only under asymptotic assumptions, that is, for large sample sizes. Small sample sizes and low frequency alleles might lead to sparse contingency tables and hence spurious results. On the other hand, the MB approach does not assume large samples and is quite robust to departures of normality (simulation results not shown).

Currently, large numbers of SNPs are available, and researchers are interested in exploiting them by using multi-locus haplotypes instead of single marker alleles. Using multi-locus haplotypes might have higher power than using single marker alleles for detecting an association. However, the relative efficiency of the MB and TB approach would be the same, regardless. On the other hand, in the final stage of a study researchers would like to know which variant or variants, within haplotypes, are causing the phenotypic differences, and this would involve testing each variant independently (as assumed here).

Finally, the results demonstrate the advantage of the MB over the TB approach. Although both approaches would be similarly efficient when just one trait was phenotyped and high selection performed for just this trait, this would not be so when the selection intensity was low. The most realistic scenario would be one in which a number of individuals had been phenotyped for a number of traits. If selection had to be applied for a large number of traits, then all or almost all of the individuals phenotyped would eventually be genotyped. The MB approach would then clearly be the most powerful.

5.5 APPENDIX

The expected values for marker genotype M_lM_n in the upper tail and lower tails are $E(x|M_lM_n)^U$ and $E(x|M_lM_n)^L$ respectively.

$$E(x \mid M_l M_n)^U = \sum_{i=1}^{2}\sum_{j=1}^{2} \eta_{ij}^U \times \frac{P(x > \tau_U \mid Q_i Q_j) \times P(Q_i M_l, Q_j M_n)}{P(M_l M_n \mid x > \tau_U) \times \alpha_U}; \quad l, n \in [1,2]$$

$$E(x \mid M_l M_n)^L = \sum_{i=1}^{2}\sum_{j=1}^{2} \eta_{ij}^L \times \frac{P(x < \tau_L \mid Q_i Q_j) \times P(Q_i M_l, Q_j M_n)}{P(M_l M_n \mid x < \tau_L) \times \alpha_L}; \quad l, n \in [1,2]$$

The within-genotype variance for the marker genotypes in both tails combined is:

$$\mathrm{var}(x \mid M_l M_n) = E(x^2 \mid M_l M_n) - E(x \mid M_l M_n)^2 =$$

$$\frac{\sum_{i=1}^{2}\sum_{j=1}^{2}\left[P(Q_i M_l, Q_j M_n) \times \left((1+\mu_{ij}^2)(\Phi_{ij}) + z_{ij}^U(\tau_U + \mu_{ij}) - z_{ij}^L(\tau_L + \mu_{ij})\right)\right]}{\sum_{i=1}^{2}\sum_{j=1}^{2}\left[P(Q_i M_l, Q_j M_n) \times (\Phi_{ij})\right]} -$$

$$\left\{\frac{\sum_{i=1}^{2}\sum_{j=1}^{2}\left[P(Q_i M_l, Q_j M_n) \times \left(z_{ij}^U - z_{ij}^L + \mu_{ij} \times \Phi_{ij}\right)\right]}{\sum_{i=1}^{2}\sum_{j=1}^{2}\left[P(Q_i M_l, Q_j M_n) \times \Phi_{ij}\right]}\right\}^2$$

The X^2 statistics, obtained from contingency tables of *3x2* and *2x2* for the counts of genotypes and alleles respectively, are distributed under the null hypothesis (H_0) of no association as chi-squared with *2* and *1* degrees of freedom respectively. Under the alternative hypothesis (H_1), X^2 is asymptotically distributed as non-central chi-squared with respectively *2* and *1* degrees of freedom and non-centrality parameter $\lambda_{Genotypes}$ and $\lambda_{Alleles}$ given by

$$\lambda_{Genotypes} = N_T \times \left[\sum_{l=1}^{2}\sum_{n=1}^{2} \frac{\left(P(M_l M_n, x < \tau_L \mid H_1) - P(M_l M_n, x < \tau_L \mid H_0)\right)^2}{P(M_l M_n, x < \tau_L \mid H_0)}\right] +$$

$$N_T \times \left[\sum_{l=1}^{2}\sum_{n=1}^{2} \frac{\left(P(M_l M_n, x > \tau_U \mid H_1) - P(M_l M_n, x > \tau_U \mid H_0)\right)^2}{P(M_l M_n, x > \tau_U \mid H_0)}\right] =$$

which after some algebra reduces to

$$\frac{\alpha_U \alpha_L N_T}{(\alpha_U + \alpha_L)^2} \times \sum_{l=1}^{2} \sum_{n=1}^{2} \frac{\left(\sum_{i=1}^{2} \sum_{j=1}^{2} P(Q_i M_l, Q_j M_n) \times [(1 - \Phi(\tau_U - \mu_{ij}))/\alpha_U - \Phi(\tau_L - \mu_{ij})/\alpha_L] \right)^2}{P(M_l M_n)}$$

and

$$\lambda_{Alleles} = 2N_T \times \left[\sum_{l=1}^{2} \frac{(P(M_l, x < \tau_L \mid H_l) - P(M_l, x < \tau_L \mid H_0))^2}{P(M_l, x < \tau_L \mid H_0)} \right] +$$

$$2N_T \times \left[\sum_{l=1}^{2} \frac{(P(M_l, x > \tau_U \mid H_l) - P(M_l, x > \tau_U \mid H_0))^2}{P(M_l, x > \tau_U \mid H_0)} \right] =$$

which after some algebra reduces to

$$\frac{2\alpha_U \alpha_L N_T}{(\alpha_U + \alpha_L)^2} \times \sum_{k=1}^{2} \frac{\sum_{l=1}^{2} \sum_{n=1}^{2} \left\{ \left(\sum_{i=1}^{2} \sum_{j=1}^{2} P(Q_i M_l, Q_j M_n) \times [(1 - \Phi(\tau_U - \mu_{ij}))/\alpha_U - \Phi(\tau_L - \mu_{ij})/\alpha_L] \right) \times P(M_k \mid M_l M_n) \right\}^2}{\sum_{l=1}^{2} \sum_{n=1}^{2} P(M_l M_n) \times P(M_k \mid M_l M_n)}$$

where N_T denotes the number of individuals genotyped (Kendall and Stuart, 1961). Power is then defined as the probability that a non-central χ^2 with respectively *2* and *1* degrees of freedom and non-centrality parameters $\lambda_{Genotypes}$ and $\lambda_{Alleles}$ is greater than the critical value defined by a central χ^2 with *2* and *1* degrees of freedom and significance level α.

Testing for association between marker genotype and phenotype using ANOVA requires specifying the model. The model is $y_{gz} = \mu + \tau_g + e_{gz}$ where y_{gz} is the phenotype for individual z with marker genotype g (*=1 if M_1M_1, =2 if M_1M_2, =3 if M_2M_2)*; μ is the mean of the selected individuals from both tails; τ_g is the g^{th} marker genotype effect taken to be fixed (its effect is constrained so that $\sum_{g=1}^{3} n_g \tau_g = 0$); and e_{gz} is the residual effect for individual z with genotype g. The total number of individuals sampled is $N_T = \sum_{g=1}^{3} n_g$ where n_1, n_2, n_3 are respectively the numbers with genotypes M_1M_1, M_1M_2 and M_2M_2 (strictly speaking, n_g are random variables but here they are treated as fixed and equal to their expected values). The between marker genotype sum of squares is

$SS_B = \sum_{g=1}^{3} n_g \times (\bar{y}_{g\bullet} - \bar{y}_{\bullet\bullet})^2$ and the within marker genotype sum of squares is $SS_W = \sum_{g=1}^{3}\sum_{z=1}^{n_g} (y_{gz} - \bar{y}_{g\bullet})^2$ where $\bar{y}_{g\bullet}$ and $\bar{y}_{\bullet\bullet}$ are the mean phenotypic values for the g marker genotype of the selected individuals and for the selected individuals respectively. When selection is applied, the within genotypic variance (σ_W^2) is not equal for all genotypes. Therefore, the weighted average is used (the weights being n_g).

Under H_0, the statistic $F = \frac{MS_B}{MS_W} \sim F_{2,N_T-3} \sim \frac{\chi_2^2}{2}$ for large N_T.

Under H_1, $E(SS_B/\sigma_W^2) = 2 + \lambda_{ANOVA} = 2 + \frac{\sum_{x=1}^{3} n_x \tau_x^2}{\sigma_W^2}$ where λ_{ANOVA} is the non-centrality parameter and $F \sim F_{2,N_T-3,\lambda_{ANOVA}}$ (Kendall and Stuart, 1961). The non-centrality parameter is expressed as:

$$\lambda_{ANOVA} = \sum_{l=1}^{2}\sum_{n=1}^{2} \frac{N_T \times P(M_l M_n) \times (E(x \mid M_l M_n) - \mu)^2}{\sigma_W^2}$$

where $\mu = \sum_{i=1}^{2}\sum_{j=1}^{2} E(x \mid Q_i Q_j, x > \tau_U \text{ or } x < \tau_L) \times P(Q_i Q_j, x > \tau_U \text{ or } x < \tau_L) =$

$$\sum_{l=1}^{2}\sum_{n=1}^{2} \frac{q_i q_j \times (z_{ij}^U - z_{ij}^L + \mu_{ij}\Phi_{ij})}{\alpha_U + \alpha_L}$$

Power is then defined as the probability that a non-central $F_{2,N_T-3,\lambda_{ANOVA}}$ with *2* and N_T-*3* degrees of freedom and non-centrality parameter λ_{ANOVA} is greater than the critical value defined by an F_{2,N_T-3} with *2* and N_T-*3* degrees of freedom and significance level α.

Regression of phenotype on marker genotype is the last type of analysis considered here. The model is $y_z = a + bx_z + e_z$ where y_z is the phenotype for individual z, a is the intercept, b is the slope, x_z is a dummy variable for individual z (taking values *–1, 0* or *1* depending on whether the individual's genotype is M_1M_1, M_1M_2 (=M_2M_1) or M_2M_2 respectively) and e_z is the residual for individual z. Regression tests for marker-trait association with one degree of freedom (i.e. ignores non-additivity), while the ANOVA based test has *2* degrees of freedom. The expected value for the estimate of b equals:

$$E(\hat{b}) = E\left(\frac{SS_{xy}}{SS_{xx}}\right) = \frac{\sum_{l=1}^{2}\sum_{n=1}^{2} P(M_l M_n)\times\left(x_{ln} - \bar{x}\right)\times\left(E(x \mid M_l M_n) - \mu\right)}{\sum_{l=1}^{2}\sum_{n=1}^{2} P(M_l M_n)\times\left(x_{ln} - \bar{x}\right)^2}$$

where SS_{xx}, SS_{xy} are respectively the sum of squares and the sum of products, x_{11}=-*1*, x_{12}=x_{21}=*0*,x_{22}=*1* and

$$\bar{x} = \left(P(M_2 M_2 \mid x > \tau_U) + P(M_2 M_2 \mid x < \tau_L)\right) - \left(P(M_1 M_1 \mid x > \tau_U) + P(M_1 M_1 \mid x < \tau_L)\right) =$$

$$\frac{1}{\alpha_U + \alpha_L}\sum_{i=1}^{2}\sum_{j=1}^{2}\left\{P\left(Q_i M_{2,} Q_j M_2\right) - P\left(Q_i M_{1,} Q_j M_1\right)\right\}\times\left(\Phi_{ij}\right).$$

Under H_0 the expected value of $\hat{b}$, *b*, is zero and the statistic $T = \hat{b}^2 / \hat{\text{var}}(\hat{b})$ is $F_{1,N_T-2} \sim \chi_1^2$ (for large N_T) distributed. Under H_1 the statistic T is non-central *F* distributed ($F_{1,N_T-2,\lambda_{Regression}}$), where $\lambda_{Regression} = SS_{xx}\times b^2 / \sigma_w^2$ (Lynch and Walsh, 1998).

Power is then defined as the probability that a non-central $F_{1,N_T-2,\lambda_{Regression}}$ with *1* and N_T-*2* degrees of freedom and non-centrality parameter $\lambda_{Regression}$ is greater than the critical value defined by an F_{1,N_T-2} with *1* and N_T-*2* degrees of freedom and significance level α.

CHAPTER 6 - Verifying the presence of a quantitative trait locus (QTL) by comparing concordant-high with concordant-low sib-pairs

6.1 INTRODUCTION

In recent years, there has been controversy about the best strategy to map and characterise the genetic variants that influence complex traits. Some argue that researchers should focus on collecting large family cohorts (Weiss and Terwilleger, 2000) and apply linkage methods. They argue that methods based on linkage disequilibrium at the population level are doomed because the stochastic processes that influence linkage disequilibrium will hinder the detection of a clear genotype-phenotype signal. Others argue that human families are generally too small to have enough power to detect the genes with small effect assumed to be involved in complex traits (e.g. complex diseases or quantitative traits) (Chakravarti, 1999) and therefore researchers should focus on collecting large cohorts of unrelated individuals. In most of the cases, researchers use a combination of both approaches because they both have pros and cons, and their success or failure depends on parameters that at the moment are unknown (e.g. the allelic architecture of complex traits). Unfortunately, the allelic architecture of complex traits will be unravelled only once the loci that influence them are characterised and researchers' positions on this matter are sometimes based more on faith than on facts. In this chapter, a study design based on collecting family information that could be used both in a linkage and linkage disequilibrium approach is proposed. I propose to collect samples of sib-pairs and use them in a two-stage approach. In the first stage, the samples would be used in a typical sib-pair linkage analysis of quantitative traits (QT) (Haseman and Elston, 1972) and in the second stage they would be used to confirm a candidate locus in a region, found in the first stage, as the causative one. The idea is that in a prospective study such as, for example, the BioBank UK study (Wright *et al.*, 2002), in which 500000 unrelated individuals would be ascertained, it might be justified to collect family information and quantitative traits to maximise the study possibilities of success. If one could collect phenotypes on, say 50000-60000 sib-pairs and their parents, and genotype only those pairs with extreme high and low phenotypes, then one would stand a good change of finding a QTL of moderate effect through linkage, and reduce to some extent the limitations of an association-alone approach in the case of, for example, allelic heterogeneity. This first stage would pinpoint interesting regions of the genome that would then be densely typed for further analysis using the test proposed here. This strategy would

reduce the costs of genotyping when compared to an association-alone approach in which all the genome would be densely typed.

A strategy based on using concordant as opposed to discordant sib-pairs was investigated. The reason for concentrating on concordant sib-pairs is that they are easier to ascertain than discordant sib-pairs. This is so because sibs share environmental and genetic factors and, therefore they show similar phenotypic values more frequently than dissimilar phenotypic values. For an equal number of discordant and concordant sib-pairs the former might have larger power. Discordant sib-pairs may be better than concordant sib-pairs because it is less likely that they both are outliers because of shared environment. However, they are also more difficult to collect and a larger number of sib-pairs would need to be phenotyped. The best strategy would therefore depend on the total cost of the experiment required to achieve a given power (e.g. the total cost of genotyping and phenotyping). A study based on discordant sib-pairs would probably have lower genotyping costs but higher phenotyping costs than a study based on concordant sib-pairs. Given that genotyping costs are decreasing and phenotyping costs increasing it seems likely that in the immediate future a strategy based on concordant sib-pairs would be the more cost-effective.

The statistical problem considered here was that of comparing gene frequencies in a sample of sib-pairs concordant for high phenotypic values of a particular trait, with those in a sample concordant for low phenotypic values. If an allele at a particular locus influenced the phenotype then this would cause a difference in allele frequency between the two samples, which should be detectable given a large enough sample. The assumption was that a candidate region found by linkage using the same data already existed, and that one wished to confirm, at minimum cost, that it did indeed affect the phenotype.

The power of the design proposed here using sib-pairs is shown and compared to a design in which only one of the sibs from each pair is used (this design would be equivalent, in terms of power, to using unrelated individuals).

6.2 MATERIAL AND METHODS

6.2.1 Model and notation

Assume a bi-allelic locus with alleles Q_1 and Q_2, having frequencies p and q (= *1* - p) respectively. Assume that the locus is Hardy-Weinberg equilibrium, and that the phenotypic values for the 3 genotypes $G_1 = Q_1Q_1$, $G_2 = Q_1Q_2$ and $G_3 = Q_2Q_2$ are normally distributed about mean values of μ_1 *(=a),* μ_2 *(= da)* and μ_3 *(=-a)* respectively, and with equal variances (V_E). The genetic model is considered additive if $d = 0$, dominant if $d = 1$ and recessive if $d = -1$. The QTL narrow-sense heritability is defined as $h^2_{QTL} = V_A/V_P = V_A/(V_A+V_D+V_E)$ where V_A *(=2pqa*^{2}*[1+{q−p}d]*^{2}*)* and V_D*(=[2pqda]*^{2} *)* are, respectively, the additive and dominance variances for the QTL and V_E is the within QTL genotype residual variance taken (without loss of generality) to be 1. If x and y denote the phenotypic values of a pair of sibs, then the siblings covariance is cov $(x, y) = \frac{1}{2} V_A + \frac{1}{4} V_D + w V_E = t V_P$ where w is the intraclass correlation for the 'background factor' (i.e., deviations within the QTL genotype), which will be a function of background genetic factors other than the QTL and the environment; and t is the siblings intraclass correlation coefficient for the phenotypic values.

Then, the joint distribution of the sibs' phenotypic values is

$$p(x,y) = \sum_{i=1}^{3}\sum_{j=1}^{3} \varphi(x,y \mid G_i, G_j) p(G_i, G_j) \qquad [1]$$

where $\varphi(x,y \mid G_i, G_j)$ denotes the bivariate normal distribution with means of μ_i and μ_j, variances of one and correlation t (the full-sib intraclass correlation); and $p(G_i,G_j)$ is the joint distribution of sib-pairs QTL genotypes. If one considers all the possible parental genotypes that can produce a pair of sibs' genotypes and their population frequency, then one has:

$$p(G_1,G_1) = p^4 + p^3q + p^2q^2/4; \quad p(G_1,G_2) = p^3q + p^2q^2/2; \quad p(G_1,G_3) = p^2q^2/4$$
$$p(G_2,G_1) = p^3q + p^2q^2/2; \quad p(G_2,G_2) = p^3q + 3p^2q^2 + q^3p; \quad p(G_2,G_3) = p^2q^2/2 + q^3p$$
$$p(G_3,G_1) = p^2q^2/4; \quad p(G_3,G_2) = p^2q^2/2 + q^3p; \quad p(G_3,G_3) = q^4 + q^3p + p^2q^2/4$$

If one specifies the "high-concordant" (i.e. both sibs have a trait value above a given threshold) sub-group as a proportion α_U of the total number of sib-pairs, then one requires determining the cut-off threshold τ_U from:

$$\int_{x=\tau_U}^{\infty}\int_{y=\tau_U}^{\infty} p(x,y)dxdy = \alpha_U \qquad [2]$$

Similarly for the "low-concordant" group of sibs,

$$\int_{-\infty}^{x=\tau_L}\int_{-\infty}^{y=\tau_L} p(x,y)dxdy = \alpha_L \qquad [3]$$

6.2.2 Allele frequencies in the "high-concordant" and "low-concordant" groups

Following Bayes' rules, the probability that any particular pair of genotypes occurs within the "high-concordant" group can be represented as

$$p(G_i,G_j \mid x>\tau_U, y>\tau_U) = \frac{p(x>\tau_U, y>\tau_U \mid G_i,G_j)\times p(G_i,G_j)}{p(x>\tau_U, y>\tau_U)} = \frac{p(x>\tau_U, y>\tau_U \mid G_i,G_j)\times p(G_i,G_j)}{\alpha_U} \qquad [4]$$

where $p(x>\tau_U, y>\tau_U \mid G_i,G_j) = \int_{x=\tau_U}^{\infty}\int_{y=\tau_U}^{\infty}\varphi(x,y \mid G_i,G_j)dxdy$.

Similarly, for the "low-concordant" group of sibs the probability that any pair of genotypes occurs is:

$$p(G_i,G_j \mid x<\tau_L, y<\tau_L) = \frac{p(x<\tau_L, y<\tau_L \mid G_i,G_j)\times p(G_i,G_j)}{p(x<\tau_L, y<\tau_L)} = \frac{p(x<\tau_L, y<\tau_L \mid G_i,G_j)\times p(G_i,G_j)}{\alpha_L} \qquad [5]$$

where $p(x<\tau_L, y<\tau_L \mid G_i,G_j) = \int_{-\infty}^{x=\tau_L}\int_{-\infty}^{y=\tau_L}\varphi(x,y \mid G_i,G_j)dxdy$.

If one denotes the number of Q_1-alleles in each pair of sibs in the "high-concordant" group by $n_U(Q_1)$ and in "low-concordant" group by $n_L(Q_1)$, then

$$p_U(4) = p\{n_U(Q_1) = 4\} = p(G_1, G_1 | x > \tau_U, y > \tau_U)$$

$$p_U(3) = p\{n_U(Q_1) = 3\} = p(G_1, G_2 | x > \tau_U, y > \tau_U) + p(G_2, G_1 | x > \tau_U, y > \tau_U)$$

$$p_U(2) = p\{n_U(Q_1) = 2\} = p(G_2, G_2 | x > \tau_U, y > \tau_U) + p(G_1, G_3 | x > \tau_U, y > \tau_U) + p(G_3, G_1 | x > \tau_U, y > \tau_U)$$

$$p_U(1) = p\{n_U(Q_1) = 1\} = p(G_2, G_3 | x > \tau_U, y > \tau_U) + p(G_3, G_2 | x > \tau_U, y > \tau_U)$$

$$p_U(0) = p\{n_U(Q_1) = 0\} = p(G_3, G_3 | x > \tau_U, y > \tau_U)$$

and

$$p_L(4) = p\{n_L(Q_1) = 4\} = p(G_1, G_1 | x < \tau_L, y < \tau_L)$$

$$p_L(3) = p\{n_L(Q_1) = 3\} = p(G_1, G_2 | x < \tau_L, y < \tau_L) + p(G_2, G_1 | x < \tau_L, y < \tau_L)$$

$$p_L(2) = p\{n_L(Q_1) = 2\} = p(G_2, G_2 | x < \tau_L, y < \tau_L) + p(G_1, G_3 | x < \tau_L, y < \tau_L) + p(G_3, G_1 | x < \tau_L, y < \tau_L)$$

$$p_L(1) = p\{n_L(Q_1) = 1\} = p(G_2, G_3 | x < \tau_L, y < \tau_L) + p(G_3, G_2 | x < \tau_L, y < \tau_L)$$

$$p_L(0) = p\{n_L(Q_1) = 0\} = p(G_3, G_3 | x < \tau_L, y < \tau_L)$$

Then the expected numbers of Q_1-alleles in each pair of sibs from the "high-concordant" and from the "low-concordant" groups are:

$$\mu_U(Q_1) = \sum_{k=0}^{4} k \times p_U(k)$$

$$\mu_L(Q_1) = \sum_{k=0}^{4} k \times p_L(k)$$

Also,

$$E[n_U(Q_1)]^2 = \sum_{k=0}^{4} k^2 \times p_U(k)$$

$$E[n_L(Q_1)]^2 = \sum_{k=0}^{4} k^2 \times p_L(k)$$

Hence the variances of $n_U(Q_1)$ and of $n_L(Q_1)$ are respectively,

$$\sigma_U^2(Q_1) = E[n_U(Q_1)]^2 - [\mu_U(Q_1)]^2$$

$$\sigma_L^2(Q_1) = E[n_L(Q_1)]^2 - [\mu_L(Q_1)]^2$$

6.2.3 Testing for allele frequency differences between high- and low-concordance groups

From now on, the approach proposed here is referred to as SP. I propose to test the null hypothesis that the locus has no effect on the phenotype i.e. that $a = 0$. Suppose the numbers of sib-pairs in the high- and low- concordance groups are N_U and N_L respectively. These are determined by the proportions α_U ($N_U = N\alpha_U$) and α_L ($N_L = N\alpha_L$), where N is the total number of pairs of sibs phenotyped in the study. Strictly speaking, if α_U and α_L are pre-specified then N_U and N_L are random variables, but here they were assumed fixed. Under the null hypothesis, the proportions $\sum n_U(Q_I)/N_U$ and $\sum n_L(Q_I)/N_L$ (where summation is over all sib-pairs in the high- and low- concordance groups respectively) both have expectation $4p$ and respective variances of $6pq/N_U$ and $6pq/N_L$ (see appendix for details). Hence, under the null hypothesis the statistic

$$S = \frac{\dfrac{\sum n_U(Q_I)}{N_U} - \dfrac{\sum n_L(Q_I)}{N_L}}{\sqrt{6\,pq\left(\dfrac{1}{N_U} + \dfrac{1}{N_L}\right)}} \qquad [6]$$

can be considered approximately Normally distributed with mean 0 and variance 1. Under the alternative hypothesis, one has

$$E[S] \approx \frac{\mu_U(Q_I) - \mu_L(Q_I)}{\sqrt{6\,pq\left(\dfrac{1}{N_U} + \dfrac{1}{N_L}\right)}} = \mu_S, \text{ say} \qquad [7]$$

and

$$Var[S] \approx \frac{\sigma_U^2(Q_I)/N_U + \sigma_L^2(Q_I)/N_L}{6\,pq\left(\dfrac{1}{N_U} + \dfrac{1}{N_L}\right)} = \sigma_S^2, \text{ say} \qquad [8]$$

6.2.4 Power calculations

If the Type I error (i.e. the probability of rejecting the null hypothesis when true) is set to γ for a two-tailed test, then this determines a constant k_γ such that $\frac{1}{2}\gamma = \Phi(-k_\gamma)$ since the test statistic has a standard Normal(0,1) distribution under the null hypothesis. The power (i.e. the probability of accepting the alternative hypothesis when true) is then

$$\Pi = 1 - \Phi\left(\frac{k_\gamma - \mu_S}{\sigma_S}\right) + \Phi\left(-\frac{k_\gamma + \mu_S}{\sigma_S}\right) \qquad [9]$$

6.2.5 Parameterisation

The problem can be specified in terms of the 9 parameters N, p, h^2_{QTL}, d, t, α_U, α_L, *Type I error, Power.*

6.3 RESULTS

6.3.1 Effect of the QTL allele frequency and the selection intensity on power when the number of sib-pairs phenotyped is fixed

Figures 6.1, 6.2 and 6.3 show the effect of the QTL allele frequency on the power to detect a locus-phenotype association for a dominant, additive and recessive model, respectively. It was assumed that there were a fixed number (5000) of sib-pairs phenotyped and that only a proportion ($\alpha_U + \alpha_L$; $\alpha_U = \alpha_L$) of them would be selected to be genotyped.

For all genetic models and intermediate QTL allele frequencies, power was maximal when about 50% of the pairs phenotyped were genotyped (25% low-concordant + 25% high-concordant). Power was essentially the same and close to one for selection intensities ranging between 0.3 and 0.8 and QTL allele frequencies (p) between 0.3 and 0.7, regardless of the genetic model considered.

Figure 6.1. Effect of the QTL allele frequency and the selection intensity on power when using SP. It was assumed that there were 5000 sib-pairs phenotyped and that a proportion of them ($\alpha_U + \alpha_L$; $\alpha_U = \alpha_L$) were selected for genotyping. $h^2_{QTL} = 0.01$, $t = 0.25$, $\gamma = 10^{-5}$. The genetic model was assumed dominant ($d = 1$).

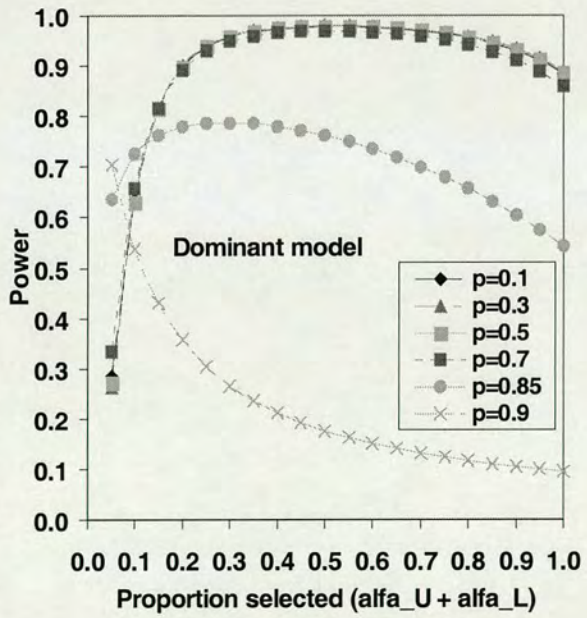


Figure 6.2. Effect of the QTL allele frequency and the selection intensity on power when using SP. It was assumed that there were 5000 sib-pairs phenotyped and that a proportion of them ($\alpha_U + \alpha_L$; $\alpha_U = \alpha_L$) were selected for genotyping. $h^2_{QTL} = 0.01$, $t = 0.25$, $\gamma = 10^{-5}$. The genetic model was assumed additive (d=0).

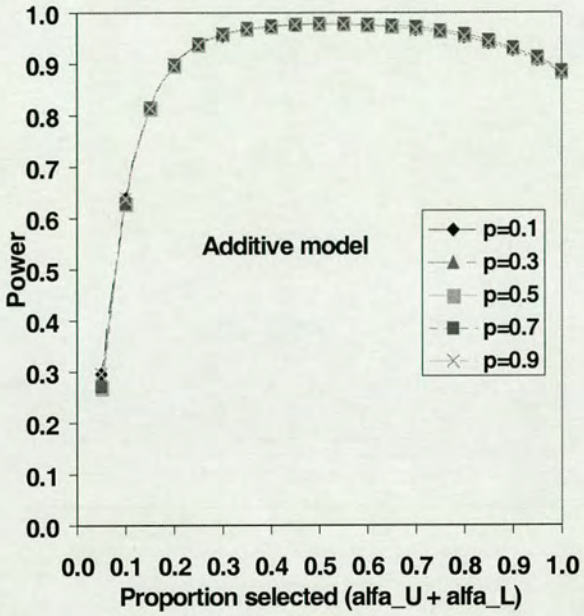
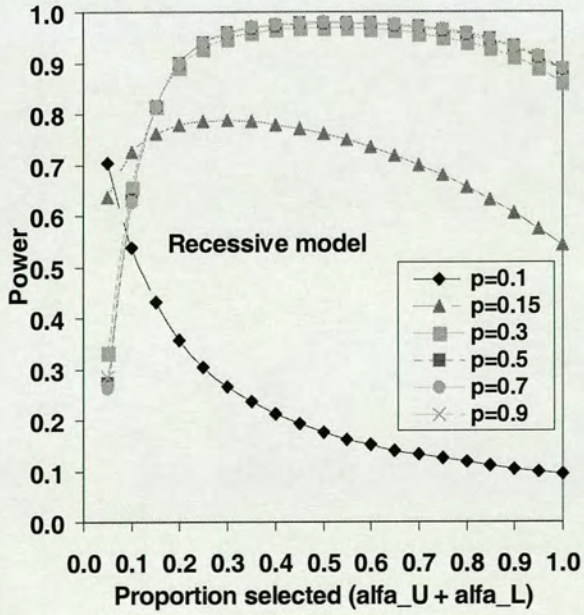


Figure 6.3. Effect of the QTL allele frequency and the selection intensity on power when using SP. It was assumed that there were 5000 sib-pairs phenotyped and that a proportion of them ($\alpha_U + \alpha_L$; $\alpha_U = \alpha_L$) were selected for genotyping. $h^2_{QTL} = 0.01$, $t = 0.25$, $\gamma = 10^{-5}$. The genetic model was assumed recessive (d=-1).



If the QTL allele frequency was either low (say, $p \leq 0.1$) or high (say, $p \geq 0.9$) and dominance was in the direction of the commonest allele, then power decreased as the proportion selected increased (Figure 6.1, dominant model and p=0.9, and Figure 6.3, recessive model and p=0.1). This is so because the enrichment of the low frequency alleles in the tails is smaller when less extreme selection is practised and therefore the test becomes less powerful.

6.3.2 Comparison of the selection strategy: both sibs versus one sib of each pair

Table 6.1 shows the sample size required to detect a QTL with 80% power and significance level equal to 10^{-5} using the SP selection strategy and a selection strategy based on sampling unrelated individuals (labelled as *One sib*, in Table 6.1). Note that numbers for SP are given in pairs of sibs whereas those for unrelated individuals are given in number of individuals. The power obtained for unrelated individuals was based on the regression of the phenotype of the selected individuals on their marker genotype (see Chapter 5). For comparison between the SP and regression approaches, it was assumed that all the sib-pairs were unrelated to each other, that only one sib from each pair was selected, at random, and that the selection of the individuals to be genotyped was based on those sibs from each pair selected at random.

For the SP design, the numbers of sib-pairs required followed the pattern shown in Figures 6.1-6.3. That is, the sample size required to detect the locus-phenotype association was very similar if the model was additive, the model was recessive and QTL allele frequency was intermediate to high or the model was dominant and the QTL allele frequency was low to intermediate.

Table 6.1 shows the effect of the full-sib intraclass correlation (t) on the sample sizes required to detect the association. As the full-sib intraclass correlation increased the sample sizes required increased (though not by very much).

Table 6.1 also shows that if $t \geq 0.3$, selecting one sib from each pair at random, applying selection on the phenotype of those sibs selected at random and genotyping those selected would have, in the majority of cases, a power larger than or equal to that obtained when selecting and genotyping both sibs from a pair (with the same selection intensity). That is, one would obtain the same or larger power if one applied selection on only one sib from each pair (selected at random) and applied a regression approach than if one used SP on both sibs. For example, for p=0.1, $\alpha_U = \alpha_L$=0.05 and the recessive model one would need 628 sib-pairs for obtaining 80% power using SP but using one sib from each pair (628 individuals) and the regression approach power would be larger than 80% (note that for 80% power and

the regression approach only 143 individuals are required). Given a h^2_{QTL}=0.02 and the allele frequencies shown in Table 6.1, the minimum t possible that complies ($t \geq ½ V_A/V_P + ¼ V_D/V_P + w V_E/V_P$) for all cases shown is $t = 0.0325$. This corresponds to the recessive model and p=0.1. If $t = 0.0325$ genotyping just one sib from each pair would not have, in more than half of the cases, as much power as genotyping both sibs but the cost of genotyping would double for only a small increase in power. Table 6.2 shows this clearer. It shows that using only one of the sibs from each pair of sibs in a regression approach would have similar power (80%) to that shown in Table 6.1 for $t = 0.0325$ using SP but the total cost of genotyping would be halved.

Table 6.1. Comparison of the numbers of sib-pairs needed to be genotyped ($2N_U$=$2N_L$) for the SP approach and the numbers of unrelated individuals (one sib) needed to be genotyped if h^2_{QTL} = 0.02, γ = $10^{-5}$and Π = 0.8. Three different intensities of selection were considered. Note that numbers for SP are given in pairs of sibs whereas those for unrelated individuals (one sib) are given in number of individuals. For formatting reasons, t=0* means t=0.0325.

		Genetic model											
		Recessive				*Additive*				*Dominant*			
	α_U = α_L	*SP*			*One Sib*	*SP*			*One Sib*	*SP*			*One Sib*
		t=0*	t=0.1	t=0.3		t=0*	t=0.1	t=0.3		t=0*	t=0.1	t=0.3	
p=0.1													
	0.05	524	549	628	143	263	276	313	332	264	277	314	321
	0.15	2554	2645	2950	421	490	511	573	545	485	506	567	534
	0.25	5145	5284	5746	688	756	786	873	756	746	774	862	745
p=0.5													
	0.05	266	278	315	298	266	279	316	313	266	279	316	299
	0.15	482	503	563	507	480	503	564	525	482	503	564	511
	0.25	737	767	852	717	736	766	852	738	737	766	852	724
p=0.8													
	0.05	267	280	316	306	265	278	315	318	230	244	287	240
	0.15	480	501	561	515	484	505	566	527	631	656	736	503
	0.25	732	762	847	725	742	772	859	739	1103	1138	1245	736

Table 6.2. Power obtained when using a regression approach and genotyping only one of the sibs from each pair. The sample size numbers correspond to those shown in Table 6.1 for obtaining 80 % power when using the SP approach and t=0.0325.

		Genetic model					
		Recessive		*Additive*		*Dominant*	
	$\alpha_U = \alpha_L$	*One sib from each pair*		*One sib from each pair*		*One sib from each pair*	
		Sample size	*Power*	*Sample size*	*Power*	*Sample size*	*Power*
p=0.1							
	0.05	524	>0.99	263	0.60	264	0.63
	0.15	2554	>0.99	490	0.72	485	0.73
	0.25	5145	>0.99	756	0.80	746	0.80
p=0.5							
	0.05	266	0.70	266	0.66	266	0.70
	0.15	482	0.76	480	0.73	482	0.76
	0.25	737	0.82	736	0.80	737	0.82
p=0.8							
	0.05	267	0.69	265	0.64	230	0.77
	0.15	480	0.75	484	0.73	631	0.93
	0.25	732	0.81	742	0.80	1103	0.98

6.4 DISCUSSION

A method based on collecting family information (sib-pairs) and analysing it in the context of association mapping has been described. Although the method described (SP) was generally less powerful than a regression approach using just one of the sibs in each pair, the collection strategy proposed might still be justified when designing a QTL mapping experiment. Several reasons argue in favour of the selection scheme proposed here. First, our understanding of the allelic architecture of complex traits is limited (Reich and Lander, 2001; Pritchard, 2001), but it is unlikely that all complex traits would have either a simple or complex allelic architecture. Some diseases or traits will have a simple allelic architecture, others a complex one and others will be complex in some contexts and simple in others. In the absence of such knowledge, it therefore seems sensible not to invest all the resources on a design that will in principle work only in a particular context. Second, there could be important savings in phenotyping if schemes designed for linkage analysis using sib-pairs could be also used in an association or linkage disequilibrium-mapping framework. Third, if the association study (preferably using just one sib from each pair) was done only on regions previously identified by linkage, then the possibility of false positives due to population stratification would be smaller than in an association-alone study because evidence for linkage on the region already existed. Fourth, savings in genotyping could be made (with respect to an association-alone approach) if only genomic regions of interest found by linkage were to be densely covered by markers (this might be more important for research groups with limited funding).

There are also some drawbacks about the design proposed here that ought to be mentioned. First, the collection of sib-pairs would be more difficult than collecting unrelated individuals especially for late onset traits where parental data might be impossible to collect (although parental data is not required for SP, it would make IBD estimation for the linkage approach much accurate and therefore increase power). Second, it requires that the first linkage scan be powerful enough to detect at least the QTL with the largest effect on the trait.

An interesting comparison would be that of a two-stage approach (a linkage genome scan + association studies on regions of interest) and a one-stage approach (genome-wide association study). In this case, one would require combining the Type I and Type II errors of the two parts of a two-stage approach (linkage + association) in such a way that the final Type I and Type II errors were the same as in the one-stage approach. This was not

addressed here and it is left for future research. Alternatively, one could fix the total number of genotyping and phenotyping and compare both approaches.

The general idea proposed here has previously been outlined by Risch and Merikangas (1996) who argued that researchers could use the samples collected for linkage analysis in association studies. However, some researchers have ignored this possibility, and have concentrated their efforts on the collection of samples of unrelated individuals. Although in some cases, this might have been due to technical problems, such as lack of DNA to do further genotyping in the initially collected sib-pairs, there might still be groups with enough DNA available to do further genotyping. For those that plan to start collecting sib-pairs in the future this might be a reminder that they should aim to get enough DNA for further association studies.

The results shown above are based on the optimal assumption that the genetic variant tested was the causative one. Generally, the situation will not be so optimal and the genetic variant tested would be one in linkage disequilibrium with the causative one. This would, of course, reduce the power of the study but the relative power of the SP and regression would not be affected greatly. Moreover, if the regions of interest were very densely typed, in such a way that even if the causal polymorphism is not typed at least one typed polymorphism is in complete LD with it and with similar allele frequencies, then the assumptions would still be valid.

I have shown that the effect of the full-sib correlation on the power of SP might be important but this alone does not explain the big differences in power between SP and regression even when the full-sib correlation was almost zero. Given that selection (when applying SP) is restricted to concordant sibs, important information might be lost by not using discordant sibs. For example, if one had a pair of discordant sibs (one sib with very extreme phenotype and the other sib with an average phenotype), then the most extreme sib could be used in the regression approach whereas the sib-pair would not be used at all in the SP approach. In addition, there might be important information contained within the tails of the quantitative trait that the regression approach is exploiting and the SP approach not (see Chapter 5). Note that the relative differences in power were smaller when the selection intensities increased. Differences in power between the SP and regression would most likely be smaller if one regressed, for example, the sib-pair phenotypic mean on the number of Q_1-alleles in each pair. In this case, a slope significantly different from zero would suggest an effect of the locus on the phenotype. An alternative approach would be to regress each individual's phenotype on the number of, say, Q_1-alleles and include in the model the effect of each sib-pair as a random effect.

6.5 APPENDIX

Under the null hypothesis, the proportions $\sum n_U(Q_1)/N_U$ and $\sum n_L(Q_1)/N_L$ (where summation is over all sib-pairs in the high- and low- concordance groups respectively) both have expectation $4p$ and respective variances of $6pq/N_U$ and $6pq/N_L$.

Under the null hypothesis, formula [4] and [5] reduce to $p(G_i, G_j)$. Then,

$$p_U(4) = p(G_1,G_1) = p_L(4)$$
$$p_U(3) = p(G_1,G_2) + p(G_2,G_1) = p_L(3)$$
$$p_U(2) = p(G_2,G_2) + p(G_1,G_3) + p(G_3,G_1) = p_L(2)$$
$$p_U(1) = p(G_2,G_3) + p(G_3,G_2) = p_L(1)$$
$$p_U(0) = p(G_3,G_3) = p_L(0)$$

Substituting the values of $p(G_i, G_j)$ defined above:

$$\mu_U(Q_1) = \sum_{k=0}^{4} k \times p_U(k) = \mu_L(Q_1) = \sum_{k=0}^{4} k \times p_L(k) = 4p(p^3 + 3p^2q + 3pq^2 + q^3) = 4p$$

and

$$\sigma_U^2(Q_1) = E[n_U(Q_1)]^2 - [\mu_U(Q_1)]^2 = \sigma_L^2(Q_1) = E[n_L(Q_1)]^2 - [\mu_L(Q_1)]^2 =$$
$$[16p^4 + 38p^3q + 28p^2q^2 + 6pq^3] - [4p]^2$$

which after some algebra reduces to $6pq$.

CHAPTER 7 - Estimation of effective population size in humans

7.1 INTRODUCTION

Estimation of past effective population size in humans is of interest to biologists for several reasons. From an evolutionary point of view, it might help to understand how humans evolved and how and when they expanded from some small region, probably in Africa, to the entire planet. Geneticists interested in gene mapping could use this information to improve their understanding and modelling of the genetic architecture underlying complex traits (Reich and Lander, 2001; Pritchard, 2001).

Traditionally, effective population size (N_e) has been estimated by comparing DNA sequences. For a population of constant size under mutation-drift equilibrium and with no recombination, the expected number of nucleotide differences between two sequences is $2N_e\mu$, with μ the per nucleotide mutation rate multiplied by the number of nucleotides in the sequence. Knowing μ, one can estimate the effective population size. If the population size has changed over time, then the estimated N_e reflects some kind of average population size between the present and the time of coalescence of the sample (Harpending *et al.*, 1998). This approach relies on the infinite alleles mutation model; therefore corrections must be applied when this model does not hold (Watterson, 1975). Kuhner *et al.* (1998) and Slatkin and Bertorelle (2001) proposed methods to estimate the exponential growth rate of the population using mutational data when the locus can be assumed selectively neutral. Knowing the current effective population and the exponential growth rate one should be able to estimate N_e at different times in the past. Although these methods can accommodate different exponential growth rates since the most common recent ancestor it might be difficult to accommodate a model in which there are continuous fluctuations in effective population size.

In this chapter, a method to estimate N_e based on linkage disequilibrium (LD) between multiple adjacent markers is used. Hill (1981) estimated N_e based on the extent of disequilibrium between genetic markers using the statistic R^2. Frisse *et al.* (2001) estimated (under an infinite-sites Wright-Fisher model with both recombination and gene conversion) the parameter $4N_ec$ (where c is the recombination rate between two sites (c can be measured in Morgans if c is small)) from which they estimated effective population size. Their parameter estimation was based on a composite-likelihood approach (e.g. they considered that there was no correlation between pairs of sites when constructing the likelihood function). Hayes *et al.* (2003) proposed a multilocus measure of disequilibrium, chromosome

segment homozygosity (CSH), which has less sampling variation and less dependence on allele frequencies than does R^2. Because LD is eroded by recombination, it is easy to control (i.e. is flexible) at what time in the past N_e is estimated. N_e at different times in the past can be estimated simply by changing the length of the chromosome segments for which CSH is estimated. Human genetic data from chromosome 19 and 22 was used to estimate past effective population size in a population of European ancestry. Results suggest that this population has had an average effective population size of ~4500 breeding individuals for the last ~4500 generations. This population had a relatively constant size (~3000-5000 individuals) from about 130000 years ago (assuming 25 years/generation) to about 1000-2000 years ago when it expanded to more than 10000 individuals.

7.2 MATERIAL AND METHODS

7.2.1 Data

7.2.1.1 Chromosome 19

Published data on human chromosome 19 (Phillips *et al.*, 2003) was used to infer past effective population size (N_e). Dr L. Cardon and Dr R. Lawrence from the Wellcome Trust Centre for Human Genetics in Oxford kindly provided these data. Data consisted of 80 unrelated haplotypes from ten CEPH reference families of European origin (Utah families). The total number of biallelic markers typed was 3697. They spanned a region of 63.46 Mb with an average spacing of 17.17 kb. The average relationship between physical and genetic distance for chromosome 19 was obtained by comparing the physical and genetic map positions for markers TSC0224540 and TSC0148348 (data available at http://bioinformatics.well.ox.ac.uk/~lon/chr19/chr19index.html) to give a figure of 1.20 cM/Mb. To check the assumed linear relationship between physical and genetic distance I plotted physical (horizontal axis) versus genetic (vertical axis) and fitted a line. The line had a slope close to 1.2 and explained 99% of the variation (results not shown). See Phillips *et al.* (2003) for a more comprehensive description of these data.

7.2.1.2 Chromosome 22

As for chromosome 19, published data on human chromosome 22 (Dawson *et al.*, 2002) was used to estimate past effective population size (N_e). The data can be downloaded from: http://sanger.ac.uk/HGP/Chr22/.

Although a full description of the data can be found in Dawson *et al.* (2002), a short description is presented here. Data consisted of 59 unrelated haplotypes from seven CEPH (Centre d'Etudes du Polymorphisme Humain) reference families of European origin (Utah families). Four of these reference families were the same as for chromosome 19. The total number of biallelic markers typed was 1504. They spanned a region of 34.49 Mb with an average spacing of 22.95 kb. The average relationship between physical and genetic distance was 2.46 cM/Mb (Dawson *et al.*, 2002).

7.2.1.3 Marker ascertainment process for the two chromosomes

The two chromosomes studied had very different marker ascertainment processes that determined how far into the most distant past N_e could be estimated. Phillips *et al.* (2003) selected markers on chromosome 19 to reflect public data bases, that is, the distribution of marker spacing is L-shaped, with a much larger proportion of closely spaced

markers than distantly spaced markers whereas Dawson *et al.* (2002) selected markers on chromosome 22 to be approximately evenly spaced every 15kb. N_e could be estimated between about 130000-1350 and 67500-760 years ago for chromosome 19 and 22, respectively.

7.2.2 Linkage disequilibrium and past effective population size estimation

The extent of linkage disequilibrium (LD) for each of the two chromosomes was estimated using the chromosome segment homozygosity (CSH) defined by Hayes *et al.* (2003). Dr B. Hayes kindly provided a C++ program that estimates CSH from haplotyped data. I shall briefly explain how CSH is defined and its estimation procedure (further details can be found in Hayes *et al.* (2003)). CSH is the probability that two chromosome segments of the same length and position drawn at random from a population are identical by descent (IBD). CSH is a multi-locus measure of LD that takes into account the linear nature of chromosomes and recombination. It, therefore, has more desirable properties than other commonly used two-locus measures such as R^2 (Hill and Robertson, 1968) or D' (Lewontin, 1964). Hayes *et al.* (2003) showed that CSH is less variable and dependent on allele frequencies than is R^2. The expectation of CSH for a population of constant size is, as for R^2 (Sved, 1971), $1/(4N_ec+1)$. Assuming that N_e varies linearly with time ($N_{e[t]}$ is the effective population size at time t), Hayes *et al.* (2003) showed that the expectation of CSH for a random mating population under a mutation-drift model is $1/(4N_{e[t]}c+1)$. $N_{e[t]}$ can be estimated at time t ($=1/2c$ generations ago), provided an estimate of CSH for a chromosomal section of length c Morgans is available. By varying this length (that is, estimating CSH for different chromosomal lengths), estimates of $N_{e[t]}$ at different times in the past can be obtained. CSH over long chromosomal segments reflects the most recent past whereas CSH over short distances reflects the most distant past.

CSH can be estimated for increasingly large chromosomal regions using information on 2 to n marker loci. For example, for a chromosomal region spanned by, say, 3 marker loci one could estimate CSH for the regions spanned by markers 1-2, 2-3, 1-3. The last estimate would use information based on three markers whereas the rest would use information based on only two markers.

CSH cannot be directly observed; therefore estimates of CSH are based on the observed haplotype homozygosity. The observed haplotype homozygosity for the population under study is defined as:

$$HH = \frac{\sum_{i=1}^{n} p_i^2 - 1/n}{1 - 1/n} \qquad [1]$$

where n is the total number of observed haplotypes and p_i is the observed frequency for haplotype i.

For the simplest case, that is CSH for 2 markers, CSH can be estimated by solving the following equation:

$$HH = CSH + \frac{(OH_1 - CSH)(OH_2 - CSH)}{1 - CSH} \quad [2]$$

where OH_i is the observed homozygosity for locus i.

After some algebra, CSH equals:

$$CSH = \frac{HH - OH_1 \times OH_2}{1 + HH - OH_1 - OH_2} \quad [3]$$

The algorithm for estimating CSH when more than two loci are involved is described by Hayes *et al.* (2003).

7.2.3 Variability of CSH and Ne

In order to account for CSH variability due to historical sampling processes and to obtain confidence intervals (CI) for the estimates of N_e, I proceeded as follows. CSH measures were binned according to the length of the chromosome on which they were based. For both chromosomes, CSH measures were binned at 0.025 cM intervals, which corresponded to about 20.8 kb and 10.1 kb for chromosome 19 and 22, respectively. The mean and the variance for each bin were estimated, as well as the mean physical distance for each interval. The mean physical distance was converted to genetic distance using the conversion rates shown above. Assuming CSH is approximately normally distributed, the 95% CI for CSH within each bin was obtained as the mean ± 2 SE (standard errors). The corresponding N_e values obtained from the upper and lower limits of the 95% CI for CSH were taken to be the 95% CI for N_e. Where the 95% CI for CSH contained the value zero, the lower CI was truncated and set to 10^{-5} (the program's output precision for zero).

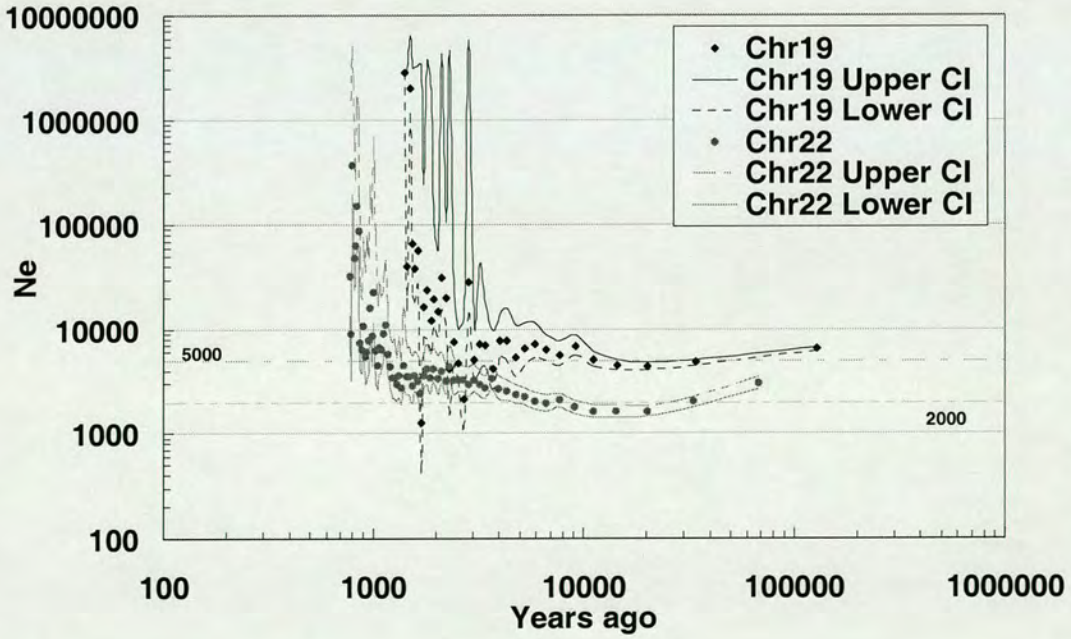
Due to computing limitations, CSH measures were obtained only for up to 23 marker loci. Nevertheless, at this point there were already too few complete haplotypes including all 23 markers for precise estimation of CSH.

7.2.4 Past effective population size based chromosome19 or chromosome 22

In this section, estimates of past effective population size based on CSH measures from chromosome 19 or chromosome 22 are shown. In the following section, estimates of N_e based on pooling CSH measures from both chromosomes are shown. Although results shown are based on binning CSH measures at 0.025 cM intervals, larger binning intervals did not yield qualitatively different results. Pooling different regions of the genome accounts for variation in the sampling of gametes from one generation to the next.

Figure 7.1 shows the effective population size estimated using chromosome 19 or chromosome 22 data. For each chromosome, lines represent the CI and the points represent the estimates obtained at a given point in the past. All the figures shown herein assumed a fixed generation time of 25 years.

Figure 7.1. Change of effective human population size (N_e) with time and its 95% confidence interval (CI). The lines with the 5000 and 2000 flags are shown to ease interpretation. Axes are on the logarithmic scale.



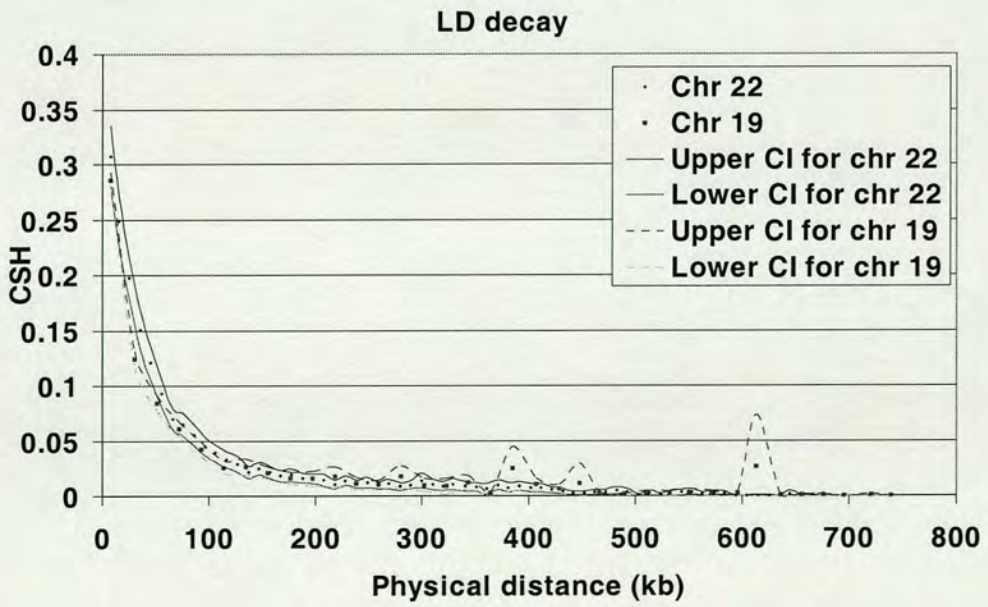
Chromosome 19 shows that N_e was about 5000 individuals, between 130000 and 10000 years ago, when it started to increase. N_e was larger than 10000 individuals from about 2000 years ago.

Chromosome 22 shows a similar, but displaced, picture. N_e was fairly constant, around 2000 individuals, between 67000-4000 years ago, and had increased by more than a factor of five 1000 years ago.

CI for both chromosomes are smaller for the distant past than for the most recent past. This, already noted by Hill (1981) and Hayes *et al.* (2003), probably reflects more complex recombinational histories over long distances than over short ones. Notably, the 95% CI for chromosome 19 and 22 hardly overlap over the whole period of time studied.

To investigate the difference in result between the two chromosomes further, the decay of LD, measured as CSH, with physical distance was plotted. Results are shown in Figure 7.2. The decay of CSH with physical distance is very similar for both chromosomes, with the 95% CI overlapping for most of the distances studied. The use of a different conversion factor from physical to genetic distance for the two chromosomes produces the shift shown in Figure 7.1. If the same conversion factor were used, then the population size estimates obtained from the two chromosomes would be practically identical (results not shown).

Figure 7.2. Decay of LD, measured as CSH, with physical distance for chromosomes 19 and 22. Lines are the 95% confidence intervals (CI).



A constant population size would give a different LD pattern to that observed in Figure 7.2, in particular at very short and very long distances, because these reflect N_e from ancestral and recent populations (Hill, 1981). Figures 7.3 and 7.4 show that the expected value of CSH ($1/(4N_ec+1)$) for a given constant N_e does not agree with the observed CSH for all physical distances. Small values of N_e (5000 and 2000 in Figure 7.3 and 7.4, respectively) approximate the data better than large values of N_e (10000 in Figures 7.3 and 7.4) for small physical distances. For large physical distances, large values of N_e = 10000 approximate the data better than small values of N_e (5000 and 2000 in Figure 7.3 and 7.4, respectively). This suggests that a model with varying N_e would fit the data better. Although the assumption of a linear relationship between N_e and time (as suggested by Hayes *et al.* (2003)) might not be the most realistic one it produces a simple expectation for CSH. Simulations performed by Hayes *et al.* (2003) showed that under a wide range of population size expansion and contraction models the approximation works well.

Figure 7.3. Observed and expected (CSH = $1/(4N_ec+1)$) decay of LD with physical distance for chromosome 19.

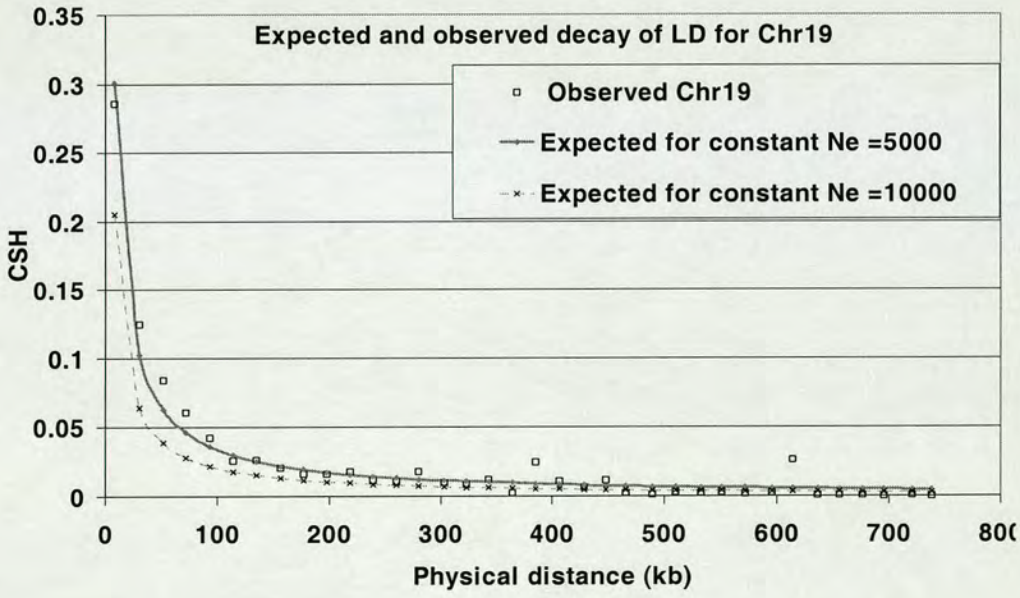
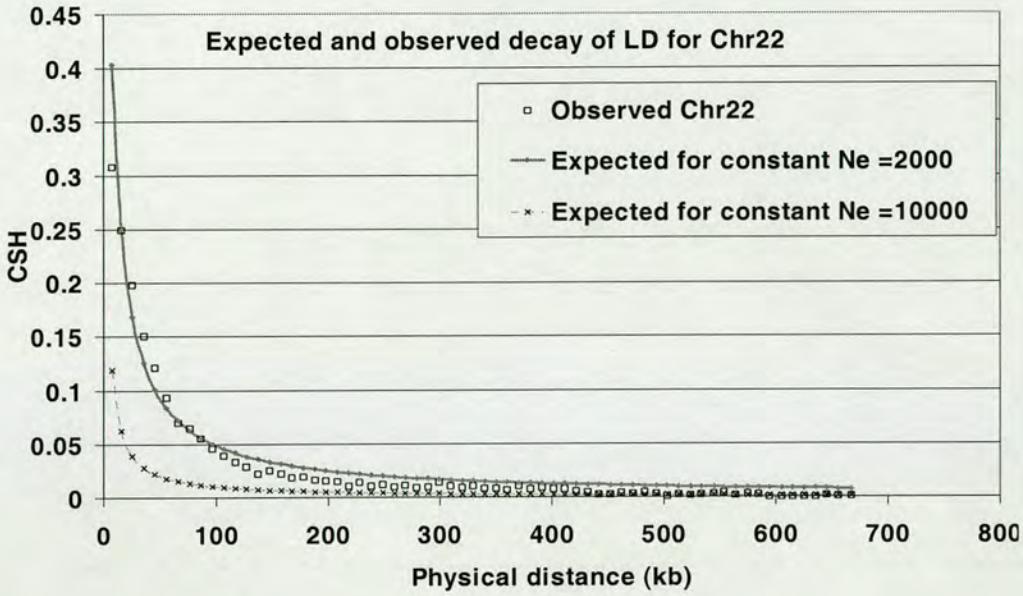


Figure 7.4. Observed and expected (CSH = $1/(4N_e c+1)$) decay of LD with physical distance for chromosome 22.

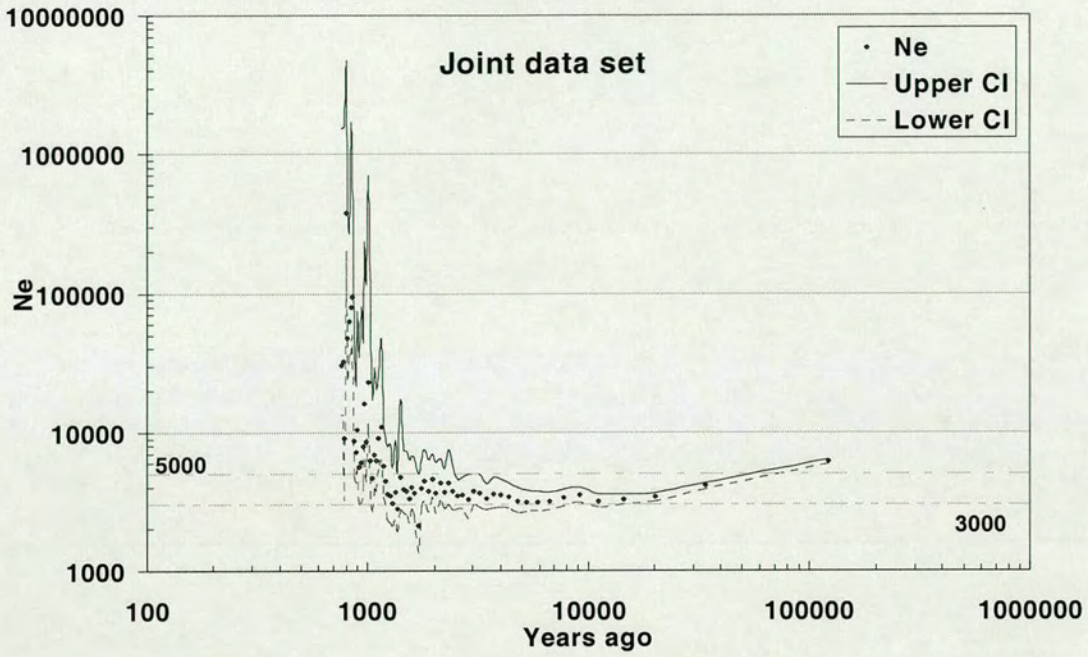


7.2.5 Past effective population size based on the joint analysis of chromosome 19 and chromosome 22

Finally, N_e was estimated by combining data for chromosome 19 and 22. Figure 7.5 shows how N_e changes with time in the past. As in the previous section, CSH measures were binned in 0.025 cM intervals. For the combined data, estimates in the past range from about 130000 to 760 years ago. The population seems to be declining slowly but steadily from 130000 years ago to about 10000 years ago when N_e is slightly larger than 3000 individuals, then it remains fairly constant for about 7000 years and starts a rapid increase between 3000 to 760 years ago.

In order to estimate an average N_e for approximately the last 130000 years the harmonic mean of all the data points (66 points) shown in Figure 7.5 was obtained. This gave an average value of 4670 individuals for 4884 generations. Following Reich *et al.* (2001), I estimated the mean inbreeding coefficient ($F = 1-(1-1/2N_e)^t$) that a population of constant effective population size 4670 individuals without mutation, selection or migration would have after 4884 generations of random mating. The estimated value of F was 0.4, in agreement with the conclusion of Reich *et al.* (2001) that, in order to explain the large extent of LD they observed in their data most of the current European population had to be descended from a population that had reached an inbreeding coefficient of at least 0.2.

Figure 7.5. Change of effective human population size (N_e) with time and its 95% confidence interval (CI) for the combined data. The lines with the 5000 and 3000 flags are only shown to ease interpretation. Both axes are on the logarithmic scale.



7.3 DISCUSSION

I have estimated effective population size in a human population of European ancestry using a large number of marker loci on two chromosomes and a recently-proposed measure of linkage disequilibrium. The estimates show that N_e in this population remained relatively constant for more than 120000 years and that only in the last 1000-2000 years did its size increase in a substantial way from about 3000 individuals to more than 10000 individuals. The estimates for N_e are smaller than those obtained from mutational data, that is, about 10000 individuals (Takahata, 1994). However, others have found results similar to those reported here. Frisse *et al.* (2001) estimated that N_e in a European population was between 2700-5600 individuals when they accounted for recombination and gene conversion in their estimation procedure. Reich *et al.* (2001) found by simulations that in order to explain their European data, the population had to go through a bottleneck with a size substantially smaller than 10000 individuals. Hayes *et al.* (2003) found similar values using data from chromosome 14.

The two data sets produced different estimates, which might be for a number of reasons. First, the sample size from chromosome 22 was substantially smaller than for chromosome 19 (59 vs. 80 haplotypes). If, as shown in Chapter 3 for D', CSH values are also overestimated for smaller sample sizes, then CSH for chromosome 22 would have been on average smaller if the sample size had been 80 rather than 59. Smaller CSH values, would have produced larger estimates of N_e. Another possible explanation might be the difference in ascertainment procedure and the effect of gene conversion. If one considers that the effect of gene conversion at short distances is equivalent to recombination, then the conversion rate from physical to genetic distance would be underestimated for the shortest genomic regions considered for chromosome 19. This is so because the conversion rate is obtained from comparing physical and genetic distances over much larger genomic regions where gene conversion effects might be hidden. If the average conversion tract length is between 500-1000bp as considered by Frisse *et al.* (2001), then conversion rates from physical to genetic distance obtained from larger distances would not account for the effect of gene conversion. Underestimation of the relationship between physical and genetic distance at short distances would lead to an overestimation of the effective population size in the more distant past. For chromosome 22, this effect would remain hidden because marker spacing is much larger and the shuffling of alleles due to gene conversion would not be observed. In this case the conversion factor between physical and genetic distance obtained from large genomic regions would be appropriate. Further research would be required to test these hypotheses.

As well as giving interesting insight into the population dynamics in the past, the estimates in this chapter could have useful implications for gene mapping. Given the relatively small effective size of the population studied, it may show drift and founder effects that would be relevant to the search for genes controlling complex traits.

CHAPTER 8 - GENERAL DISCUSSION

Chapter 8 is subdivided in two sections. The first section is a summary of the main features and results for each of the six research chapters of this thesis. The second section is a discussion of them and some other general matters.

8.1 SUMMARY OF RESULTS

In Chapter 2, the association between genetic marker alleles was estimated for two regions of the bovine genome from a random sample of 50 young dairy bulls born in the United Kingdom between 1988 and 1995. Microsatellite marker genotypes were obtained for 6 markers on chromosome 2 and 7 markers on chromosome 6, spanning 38 and 20 cM, respectively. Two different methods, that do not require family information, were used to estimate population haplotype frequencies. Haplotype frequencies were estimated for pairs of loci using the expectation-maximization algorithm, and for all linked loci using a Bayesian approach via a MCMC algorithm. Significant (P = 0.0007) linkage disequilibrium was detected between pairs of loci in syntenic groups (that is, loci in the same linkage group), extending to about 10 cM. No significant linkage disequilibrium was detected between markers in non-syntenic regions. Given the observed level of linkage disequilibrium, mapping methods based upon population-wide association may provide better resolution than traditional QTL-mapping methods in the United Kingdom dairy cattle population, as well as reduce the required sample sizes of the experiments.

In Chapter 3, the extent of LD in a sub-isolate of the general Sardinian population (775 members of Talana village) was assessed using 22 polymorphic markers on chromosome 19. High levels of disequilibrium were found that extended to 8 cM, when based on the measure of linkage disequilibrium D', and 11 cM when based on the significance level of the allelic association. The fact that conclusions based on both methods were similar suggests that the estimates are quite robust. It was also shown, through a simple resampling technique, that small sample sizes can overestimate both the mean value of D' and its variance by up to a factor of about 3 and 23, respectively, when the number of diplotypes (the pair of haplotypes that compose the genotype) decreases from 381 to 25. I evaluated the effect on D' of the depth of the pedigree available when using phased founders, and compared the estimates to those obtained when using unphased founders; and also the effect of grouping alleles on the value of D' and on the significance level. Due to the high sampling variance of LD measures, the use of at least 200 unrelated individuals when characterizing the extent of LD is recommended.

In Chapter 4, a strategy to map QTL using LD when the QTL and marker locus were multiallelic was considered. The strategy involved phenotyping a large number of unrelated individuals and genotyping only selected individuals from the two tails of the trait distribution. Power to detect trait-marker association was assessed as a function of the number of QTL and marker alleles. Two patterns of LD were used to study their influence on power. When the frequency of the QTL allele with the largest effect and that of the marker allele linked in coupling were equal, power was maximum. In this case, increasing the number of QTL alleles reduced the power. The maximum difference in power between the two LD patterns studied was ~30%. For low QTL heritabilities ($h^2_{QTL} < 0.1$) and single trait studies selecting around 5% of the upper and lower tails of the trait distribution is recommended.

In Chapter 5, two approaches for mapping QTL using LD at the population level were investigated. In the trait-based (TB) approach, the frequencies of marker alleles (or genotypes) were compared in individuals selected from the two tails of the trait distribution. The TB approach uses phenotypic information only in the selection step. In the marker-based (MB) approach, the quantitative trait values for the marker genotypes in the selected individuals were compared. The MB approach uses both the difference in marker allele (or genotype) frequencies and the phenotypic values of each marker genotype in the selected samples. I quantified the power of each approach and showed that the power of the MB approach was greater than or equal to that of the TB approach. The advantage of the former is expected to increase with increasing number of traits phenotyped. The approximations were validated by simulation.

In Chapter 6, a design based on collecting concordant sib-pairs for high and low phenotypic values and comparing the allele frequency distribution in both groups was considered. Although the method described was generally less powerful than a regression approach using just one of the sibs in each pair, the collection strategy proposed might still be justified when designing a QTL mapping experiment, because the collected samples would be used in a preliminary linkage analysis followed by a LD study.

In Chapter 7, human genetic data from chromosomes 19 and 22 was used to estimate past effective population size in a population of European ancestry. Estimates were based on a multilocus measure of LD, the chromosome segment homozygosity, which has known expectation for a random mating population of constant size under a mutation-drift model. Results suggested that this population had an average effective population size of ~4500 breeding individuals for the last ~4500 generations. This population had a relatively constant

size (~3000-5000 individuals) from about 130000 years ago (assuming 25 years/generation) to about 1000-2000 years ago when it expanded to more than 10000 individuals.

8.2 DISCUSSION

Geneticists have an interest in finding loci that underlie phenotypic variation. This interest stems from the fact that there are only a limited number of genes that control phenotypic variation, and that it has been shown that at least some of them explain a significant proportion of this variability. The world's human population is estimated to be about 6 billion. Apart from monozygotic twins, each of these people has their own unique genotype, expressed in a set of unique phenotypes, some unique to the human species. It is both surprising (at first glance) and stimulating that 30000 protein-encoding genes control such a huge phenotypic diversity and complexity (Lander, *et al.*, 2001). Since one tends to think in Mendelian terms and therefore (wrongly) associate one gene genotype with one phenotype, 30000 genes seem surprisingly few. It is also surprising that complex organisms such as humans have only twice as many genes as worms and three times as many as flies (Baltimore, 2001). Nevertheless, it might not be so surprising that only 30000 genes control such a huge phenotypic diversity if one takes into account that there are epistatic interactions (note that with 30000 genes there might be about 450 million two-way additive by additive gene interactions and that the numbers scale very fast with higher order interactions and dominance) and that human genes are regulated and expressed in a much more complex way that genes in worms or flies (Lander, *et al.*, 2001). It is stimulating because this relatively small number of genes makes us think that we can deal with this level of complexity, especially after the initial successes in mapping and cloning some of the loci underlying Mendelian human diseases such as diastrophic dysplasia (Hastbacka *et al.*, 1992). Nevertheless, the task ahead seems to be the hardest one. It is true that the number of genes in the human genome is not so large, but it is also true that these genes interact, have mainly small effects, and that they are expressed in different ways in different tissues, developmental stages and environments. In addition, the phenotypes researchers tend to measure or categorise arise from complex and almost always unknown pathways and thus tend to be a poor and imprecise clue to the underlying biology. This complexity is probably the main reason why so called complex traits are hardest to map and why there has been, to date, limited success (see Table 1.2 in Chapter 1). Although associations between marker genotypes and complex phenotypes are usually reported, the findings are also usually not replicated. This is so regardless of whether linkage or linkage disequilibrium methods are used. The reasons for this are diverse and probably different for each unreplicated finding. In

my opinion, the three main reasons are (i) lack of statistical power, (ii) false positives and (iii) true biological differences between populations.

Lack of statistical power is probably the main reason for lack of replication across studies. Two of the most common and probably important causes for lack of statistical power will be considered. The first is a consequence of our limited understanding of the basic biology of most of the traits studied. For example, two people might have the same phenotype (say, high glucose level, which may lead to serious health problems) but if one looked closer then one might find that they have in fact very different phenotypes for insulin levels, as is the case for type I and type II diabetes. Here, knowledge of the underlying biology helps to distinguish two very similar phenotypes for glucose levels but very different for insulin levels, and therefore to concentrate efforts on the latter. The second reason is related to the assumptions made when studies are planned. In power calculations, it is usually assumed that loci responsible for variation are biallelic. In Chapter 4, it was shown that when this assumption is not met, the power of LD mapping studies assuming a biallelic QTL could be seriously overestimated, and that power to detect a QTL with a given heritability decreases with increasing number of QTL alleles. Although the number of alleles involved cannot be known *a priori*, empirical data (Hugot *et al.*, 2001; Ogura *et al.*, 2001; Grisart *et al.*, 2002) shows that multiallelic trait loci exist and this should be taken into account when designing studies. Designs planned under the assumption that one locus will explain a given proportion of the phenotypic variance (estimated from an initial report) will be also underpowered, since the effects reported in original reports are usually biased upwards (Goring *et al.*, 2001). This might be further aggravated if there were, for example, two closely linked QTL and alleles with effects in the same direction were on the same haplotypes. Then, the original reported effect would have included the joint effect of both. If the samples collected for the replication study did not show the same linkage phase as the original report, then the estimated power of the study would be overestimated.

The second main reason for lack of replication is that the initial report was in fact a false positive. Stratification or hidden population structure has probably been the commonest reason given (although in few cases clearly demonstrated) for explaining lack of replication in LD mapping studies. Although it is obviously something that researchers should be aware of and try to avoid, it is probably not as likely to be the cause for lack of replication as are underpowered studies. For stratification to be important, it requires both a difference in allele frequencies between populations and a difference in disease incidence. Although the first might be true even for relatively uniform populations, the incidence of the most common

human diseases does not vary greatly between geographically and genetically close populations (which are the most likely to be admixed).

Lastly, apart from type I and type II errors other reasons might account for the lack of replication. Because human populations are naturally stratified, one locus might be segregating in one population but not in another. Even if it was segregating in both populations allele differences between the two populations may explain lack of replication, because not enough informative families or individuals can be found in one population whereas it is relatively easy to find them in another. Also, epistatic interactions and gene by environment interactions might explain differences among populations.

Quantitative traits are of interest to human geneticists because they are risk factors for disease and they have a number of advantages over disease status records. They are less heterogeneous and more easily scored, selection strategies can be easier to apply, and each individual *per se* provides better phenotypic information. For example, pulmonary capacity can be measured in all individuals, and any change in this trait between multiple measurements is expected to be smoother than, for example, presence or absence of asthma. This reduces the effect of misclassification. For several reasons (discussed in Chapter 5) some researchers still dichotomise quantitative traits. This is not always the best strategy because as shown in Chapter 5 power can be significantly reduced in some cases. However, I used this dichotomising approach in Chapters 4 and 6, and the reasons why I did so should be discussed here. In Chapters 4 and 6, I wanted to study the effect of certain parameters on power, for example, allele frequencies, level of disequilibrium, etc. The effect of these does not depend on whether the trait is dichotomised or not. Given that, I opted for using a dichotomising approach because power calculations are simpler. In addition, Chapter 4 was prompted by the method proposed by Schork *et al.* (2000), and it made sense to use the similar methods. Besides, given high enough selection intensities for both tails of the trait distribution, both approaches have roughly the same power (Chapter 5), and results from Chapters 4 and 6 were obtained by applying high selection intensities. Given that making full use of the information contained in quantitative traits would provide higher power, the results shown in Chapters 4 and 6 can be taken as the minimum power achievable.

An important parameter for LD mapping methods is the extent of LD in the population of interest. This will determine, for example, the density of markers required to achieve a given power, or the resolution achievable. In Chapters 2 and 3, the extent of LD in two populations of different species was studied. There are important differences, apart from species, in the two chapters. In Chapter 2, pedigree information was not available and population haplotype frequencies had to be estimated from unphased individuals. In addition,

sample size was small. In Chapter 3, pedigree information was available, allowing me to infer founder haplotypes and hence to estimate population haplotype frequencies (by counting founder haplotypes), and sample size was large (about seven times larger than in Chapter 2). Like many published studies about the extent of LD (Reich *et al.*, 2001; Patil *et al.*, 2001), Chapter 2 had a very small sample size leading to an upwards bias in the estimate of the extent of LD measured as D'. It was shown in Chapter 3 that reducing the sample size from 381 diplotypes to 25 diplotypes increased almost three-fold the estimate of D'. This has important implications if a small sample is used to estimate the density of markers required for a LD mapping study, and suggests that researchers should aim to do these preliminary studies with at least 200 individuals; otherwise information should be interpreted with caution. In addition, it was shown in Chapter 3 that estimates of D' from phased individuals tended to be smaller than from unphased individuals, suggesting a further upwards bias of D' when using unrelated individuals. However, estimates of the useful level of LD (as defined in Chapter 3) from unphased individuals extended over shorter map distances because the decay of LD with distance was faster than with phased individuals.

The use of alternative two-locus measures of LD, such as R^2 (Hill and Robertson, 1968), would probably be more appropriate if the aim was to estimate, for example, the density of markers required for a LD mapping study, because there is a clearer relationship between R^2 and power. Nevertheless, D' was used in Chapters 2-5 because there are a number of studies published with this measure, making comparisons across studies easier. Multilocus measures of LD, such as CSH (Hayes *et al.*, 2003) or τ (McPeek and Strahs, 1999), could also have been used because they are more informative about the average length of IBD segments that two randomly selected individuals share, but this would have made comparisons with other studies complicated. Moreover, these multilocus measures tend to work better for small genomic regions with a high density of markers (where the assumption that there is a correlation between marker IBD and IBS status is more likely to hold) and are computationally more demanding. In Chapter 7, CSH was used for estimating past effective human population size. The results will help to improve models of human expansion and facilitate more accurate estimates of those LD mapping methods that model the population history.

Recently, there have been a number of studies that claim a block-like structure of the human genome, but it is yet not clear whether the blocks are a consequence of biological processes (i.e. hot-spots), population history or a combination of both. Even more importantly, it is not clear (at least to me) how well *in silico* methods of block detection (Patil *et al.*, 2001; Zhang *et al.*, 2002) would reflect real biological processes, such as

recombination hot-spots and whether they would pick up the same recombination hot-spots in different populations (e.g., in African and European populations). For instance, it would be interesting to test (in different populations) the *in silico* methods of block detection proposed to date on data such as that presented by Jeffreys *et al.* (2001), where recombination hot-spot positions are known, and to see how accurately they pinpoint the hot-spot positions. Data presented to date seem consistent with the hypothesis that haplotype blocks are a consequence of both recombination hot-spots and population history. For example, Gabriel *et al.* (2002) studied the haplotype structure in different populations and concluded that block boundaries were conserved in European and African populations but also that block lengths were twice as large in Europeans as in Africans (which, at first glance, seems somehow contradictory). Since the boundaries across populations were the same, this would certainly favour the hot-spots hypothesis, but because block lengths were different, this suggests that population history or marker ascertainment also has an important role in shaping haplotype blocks (Phillips *et al.*, 2003). All these matters need clarification before information on haplotype blocks can be really useful for LD mapping studies and projects such as the HapMap (Couzin, 2002) will provide important information.

The implications that haplotype blocks would have for LD mapping studies are now discussed. Because studies that describe the block-like structure are based on common polymorphisms (say, a minor allele frequency larger than 5-10%), the discussion here will be centred on such polymorphisms and the assumption that variants that predispose to disease or affect phenotypic variation of quantitative traits are common. If these polymorphisms happened to be less common, then selection of SNPs that reflect common haplotype blocks would probably not be a good strategy since most of these rare alleles would be missed when defining the haplotype blocks. The first consequence of a block-like structure of the genome would be a considerable overall reduction, compared to previous estimates (Kruglyak, 1999), in the number of SNPs required for LD mapping. Given that typing all SNPs within blocks would provide redundant information, one would only need to type those that capture most of the haplotype diversity within each block. The required distribution of SNPs across the genome for LD mapping experiments would also be affected. In order to characterise most of the haplotype diversity, regions of low LD would require a high density of SNPs whereas regions of high LD would require a much lower density. The second consequence is that association studies would be based on comparison of haplotypes within blocks, as opposed to haplotypes of arbitrary length or single polymorphisms. Once a block-phenotype association was found, one would try to distinguish if one of the multiple SNPs that define a haplotype within the relevant block is associated with the trait. Since all SNPs within blocks

will be in strong LD this may prove difficult and typing more SNPs or sequencing the whole region might be necessary. To reduce costs, one might first try to reduce further the length of the associated block by looking in other populations where recombination might have reduced its length. Then, one would use biological information to pinpoint candidates in exonic regions (in the simplest case, if there were mutations that might cause a change in the final protein product). If none of the polymorphisms that change the protein structure were causative of the phenotype studied, then one would need to look for less obvious candidates in regulatory and untranslated regions. All this, of course, assumes that LD across blocks is negligible and therefore the causative variant is not in another block than the one showing strongest association. If LD across blocks was not negligible, then a SNP selection strategy based on block structure might not be the best or cheapest one. Although initial results suggest that inter-block LD may be important (Gabriel *et al.*, 2002) more empirical research is needed to assess this issue properly. A third implication of the existence of haplotype blocks would be that the number of statistical tests performed would be reduced to the number of blocks present in the study, and this would reduce to some extent the problems associated with multiple testing, especially for whole-genome scans.

The allelic architecture of complex traits is another highly debated issue, because the level of complexity will determine whether it will be possible to map trait loci with existing mapping methods and designs. Above, it has been discussed what I think will be the future for LD mapping methods under the assumption that the causative allelic variant is common in the population. While this will be the case for some variants and traits, it will not be so in all cases, and researchers will have to work out how to deal with this problem. If there is allelic heterogeneity, it will be hard to map polymorphisms affecting these traits using LD mapping methods (because none of the alleles would be common), and probably linkage methods will yield better results, provided that locus heterogeneity is not too high (so that enough families segregating for a given locus can be ascertained). If both allelic and locus heterogeneity are fairly common, then prospects for QTL mapping are not very encouraging and researchers will have to rely on alternative methods such as comparative mapping across species or gene expression studies that might provide good candidate genes for further studies and biological insight of the trait of interest. In this case, population isolates might also help since they are expected too show much less locus and allelic heterogeneity (Wright *et al.*, 1999). Nevertheless, it seems a reasonable approach to start working with the easiest part (that is, finding out the common variants that affect variation) and, with the experience gained from this, follow with the most difficult problem of finding rare variants. This, of course, would mean that groups working with diseases or traits where variation is only due

to rare variants might have little or no success but one does not know that *a priori*, and for the general advance of knowledge and personalised medicine this would probably be the best strategy.

In the current state of knowledge, it is not yet clear if researchers applying LD mapping methods should focus on carefully selected candidate loci or on genome-wide scans. Selection of good candidate loci might be difficult for poorly understood traits (e.g. those for which we do not know much about the underlying biology) but for those for which we know their underlying biology or where we have good candidates identified in another species (for example, from knock-out mice), they might prove the method of choice. Since causality is easier to prove, exonic regions should be screened first. Failing those, promoter and regulatory regions should be screened but here, proving causality will be harder and will involve for example gene expression analysis for which not all laboratories might have the expertise required. Genome-wide scans pose important economical and methodological issues. High-throughout genotyping technology is still very expensive (and not affordable for small research groups), especially if one takes into account that in many cases public SNP databases (such as, dbSNP (http://www.ncbi.nlm.nih.gov/SNP/)) are of limited utility and that sequencing has to be done on the available samples to discover SNPs segregating in the population under study. For example, Johnson *et al.* (2001) screened 135 kb from nine genes and found 122 SNPs. Fewer than 25% of the SNPs they identified at any one gene were present in dbSNP and for three genes none of the SNPs they identified were in dbSNP. The most important methodological issue concerns multiple testing and the associated loss of power after correcting for multiple testing. For example, let us assume for simplicity that the number of blocks in each of the 23 human autosomes is equal to the number (4135) found in chromosome 21 (the third shortest one) by Patil *et al.* (2001). This gives a total of about 100000 (~ 23x4135) blocks. How to perform 100000 tests (even if they are not independent) with relatively modest sample sizes (say, 2000 individuals) without losing statistical power and keeping the number of false positives at reasonable levels surely will require further challenging research before researchers attempt to do genome-wide LD mapping scans.

Finally, there will come a time (probably sooner than expected, maybe 10-20 years) when all the technological and biological unknowns discussed above will not longer be unknowns. Then, gene by gene and gene by environment interactions will have predictable consequences and personalised risk assessments for disease and therapies will be available. It might well be that these personalised risk assessments and therapies will come first for quantitative traits such as blood pressure, cholesterol levels or body weight which probably have a simpler underlying biology than disease traits. Nevertheless, the final consequences

for an individual with his/her risk assessment in his/her hands would be very similar. Having this information will be important for several reasons.

First, it will give the individual the option to choose whether or not to change his/her habits. The impact of this is difficult to predict. For a long time medical doctors and epidemiologists have warned the general population about the risks on health of smoking and drinking but yet a considerable proportion of the population smokes and drinks in excess in most European countries. Knowing this information might help change the general population habits but only if accompanied with education. In my opinion, the lack of population response to health hazards such as smoking or drinking is due to the enormous variability in risk assessments and poor understanding of what they mean by the general population. For example, there is large variability in the onset of disease such as lung cancer due to smoking and it is difficult to convince a healthy smoker to stop smoking just by giving him/her a probability of developing the disease if he/she keeps smoking when he/she sees 90 year old smokers perfectly healthy. Individual risk assessments would need to be a precise and show to work if the general population is to take them seriously. Also, education of the general population in a simple and comprehensive manner has to be in line, if not ahead, with technological and biological advances. Only then they will be useful for and accepted by society.

Second (and probably more importantly), it will provide more efficient and less toxic medicines. Since the underlying biological pathways will be known, drugs will only have to target those routes where it is necessary to do so, hence reducing toxicity and increasing efficiency.

Nevertheless, there are also dangers that society will have to avoid when this information is available. Discrimination for genetic reasons might be one of them. Insurance companies might not want to insurance people at risk, employees with high-risk assessments might lose their jobs or people with rare diseases might not benefit from targeted drugs since it would not be cost-effective to produce them.

I think that the benefits of the genomic revolution for science and health care will be immense and that society will have to understand and learn how to use them in an altruistic way so they can benefit all of us.

REFERENCES

Abecasis, G. R., and Cookson, W. O. C. 2000. Gold - Graphical overview of linkage disequilibrium. Bioinformatics. 16: 182-183.

Allison, D. B. 1997. Transmission-disequilibrium tests for quantitative traits. Am. J. Hum. Genet. 60: 676-690.

Allison, D. B., Moonseong, H., Schork, N. J., Wong, S. L., and Elston, R. C. 1998. Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. Hum. Hered. 48: 97-107.

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

Angius, A., Melis, P. M., Morelli, L., Petretto, E., Casu, G., Maestrale, G. B., Fraumene, C., Bebbere, D., Forabosco, P., and Pirastu, M. 2001. Archival, demographic and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits. Hum. Genet. 109: 198-209.

Angius, A., Bebbere, D., Petretto, E., Falchi, M., Forabosco, P., Maestrale, B., Casu, G., Persico, I., Melis, P.M., and Pirastu, M. 2002a. Not all isolates are equal: linkage disequilibrium analysis on Xq13.3 reveals different patterns in Sardinian sub-populations. Hum. Genet. 111: 9-15.

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

Atwood, L. D., and Heard-Costa, N. L. 2003. Limits of fine-mapping a quantitative trait. Genet. Epidemiol. 24: 99-106.

Ayres, K. L., and Balding, D. J. 2001. Measuring gametic disequilibrium from multilocus data. Genetics. 157: 413-423.

Bader, J. S., Bansal, A., and Sham, P. 2001. Efficient SNP-based tests of association for quantitative phenotypes using pooled DNA. Genescreen. 1: 143-150.

Bacanu, S., Devlin, B., and Roeder, K. 2000. The power of genomic control. Am. J. Hum. Genet. 66: 1933-1944.

Baltimore, D. 2001. Our genome unveiled. Nature. 409: 814-816.

Cardon, L. R., and Bell, J. I. 2001. Association study designs for complex diseases. Nat. Rev. Genet. 2: 91-99.

Carey, G., and Williamson, J. 1991. Linkage analysis of quantitative traits - increased power by using selected samples. Am. J. Hum. Genet. 49: 786-796.

Chakravarti, A. 1999. Population genetics-making sense out of sequence. Nat. Genet. 21: 56-60.

Clergetdarpoux, F., Bonaitipellie, C., and Hochez, J. 1986. Effects of mis-specifying genetic-parameters in lod score analysis. Biometrics. 42: 393-399.

Collins, A., and Morton, N.E. 1998. Mapping a disease locus by allelic association. Proc. Natl. Acad. Sci. USA. 95: 1741-1745.

Couzin, J. 2002. New mapping project splits the community. Science. 296: 1391-1393.

Curtis, D. 1997. Use of siblings as controls in case-control association studies. Ann. Hum. Genet. 61: 319-333.

Daly, M. J., Rioux, J. D., Schaffner, S. E., Hudson, T. J., and Lander, E. S. 2001. High-resolution haplotype structure in the human genome. Nat. Genet. 29: 229-232.

Darvasi, A., and Soller, M. 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet. 85: 353-359.

Darvasi, A., and Soller, M. 1994. Optimum spacing of genetic-markers for determining linkage between marker loci and quantitative trait loci. Theor. Appl. Genet. 89: 351-357.

Darvasi, A. 1998. Experimental strategies for the genetic dissection of complex traits in animal models. Nat. Genet. 18: 19-24.

Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R., and Dunham, I. 2002. A first-generation linkage disequilibrium map of human cromosome 22. Nature. 418: 44-548.

De La Chapelle, A., and Wright, F. A. 1998. Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc. Natl. Acad. Sci. USA. 95: 12416-12423.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B. 39: 1-38.

Devlin, B., and Roeder, K. 1999. Genomic control for association studies. Am. J. Hum. Genet. 65: 437.

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

Doerge, R. W., and Churchill, G. A. 1996. Permutation tests for multiple loci affecting a quantitative character. Genetics. 142: 285-294.

Ducrocq, V., and Quaas, R. L. 1988. Prediction of genetic response to truncation selection across generations. J. Dairy Sci. 71: 2543-2553.

Dunning, A. M., Durocher, F., Healey, C.S., Teare, M.D., McBride, S.E., Carlomagno, F., Xu, C.F., Dawson, E., Rhodes, S., Ueda, S., Lai, E., Luben, R. N., Van Rensburg, E. J., Mannermaa, A., Kataja, V., Rennart, G., Dunham, I., Purvis, I., Easton, D., and Ponder, B. A. J. 2000. The extent of linkage disequilibrium in four populations with distinct demographic histories. Am. J. Hum. Genet. 67: 1544-1554.

Eaves, I. A., Merriman, T.R., Barber, R. A., Nutland, S., Tuomilehto-Wolf, E., Tuomilehto, J., Cucca, F., and Todd, J. A. 2000. The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nat. Genet. 25: 320-323.

Excoffier, L., and Slatkin, M. 1995. Maximum-likelihood-estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. 12: 921-927.

Falconer, D.S, and Mackay, T.F.C. 1996. Introduction to Quantitative Genetics. Longman, 4th edition, England.

Fallin, D., and Schork, N. J. 2000. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am. J. Hum. Genet. 67: 947-959.

Farnir, F., Coppieters, W., Arranz, J. J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D., and Georges, M. 2000. Extensive genome-wide linkage disequilibrium in cattle. Genome Res. 10: 220-227.

Fisher, R. A. 1970. Statistical methods for research workers. Oliver and Boyd, 14th edition, Edinburgh, UK.

Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., and Di Rienzo, A. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. 69: 831-843.

Fulker, D. W., and Cardon, L. R. 1994. A sib-pair approach to interval mapping of quantitative trait loci. Am. J. Hum. Genet. 54: 1092-1103.

Fulker, D. W., Cherny, S. S., and Cardon, L. R. 1995. Multipoint interval mapping of quantitative trait loci, using sib pairs. Am. J. Hum. Genet. 56: 1224-1233.

Fulker, D. W., and Cherny, S. S. 1996. An improved multipoint sib-pair analysis of quantitative traits. Behav. Genet. 26: 527-532.

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

Genstat 5 Committee. 1993. Genstat 5 reference manual. Clarendon Press, Oxford, UK.

Glazier, A. M., Nadeau, J. H., and Aitman, T. J. 2002. Finding genes that underlie complex traits. Science. 298: 2345-2349.

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

Goring, H. H. H., Terwilliger, J. D., and Blangero, J. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. Am. J. Hum. Genet. 69: 1357-1369.

Graham, J., and Thompson, E. A. 1998. Disequilibrium likelihoods for fine-scale mapping of a rare allele. Am. J. Hum. Genet. 63: 1517-1530.

Green, P., Falls, K., and Crooks, S. 1990. Cri-Map version 2.4. Washington University School of Medicine, St. Louis.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M., and Snell, R. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12: 222-231.

Guo, S. W., and Thompson, E. A. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics. 48: 361-372.

Haley, C. S., and Knott, S. A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity. 69: 315-324.

Haley, C. S., and Visscher, P. M. 1998. Strategies to utilize marker-quantitative trait loci associations. J. Dairy Sci. 81: 85-97.

Harpending, H. C., Batzer, M. A., Gurven, M., Jorde, L. B., Rogers, A. R., and Sherry, S. T. 1998. Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA. 95: 1961-1967.

Hartl, D. L., and Clark, A. G. 1997. Principles of population genetics. Sinauer, 3th edition, Massachusetts.

Haseman, J. K., and Elston, R. C. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. 2: 3-19.

Hastbacka, J., Delachapelle, A., Kaitila, I., Sistonen, P., Weaver, A., and Lander, E. 1992. Linkage disequilibrium mapping in isolated founder populations - diastrophic dysplasia in Finland. Nat. Genet. 2: 204-211.

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 Res. 13: 635-643.

Hedrick, P. W. 1987. Gametic disequilibrium measures - proceed with caution. Genetics. 117: 331-341.

Hill, W.G., and Robertson, A. 1966. The effect of linkage on limits to artificial selection. Genet. Res. 8: 269-294.

Hill, W. G., and Robertson, A. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38: 226-231.

Hill, W. G. 1974. Estimation of linkage disequilibrium in randomly mating populations. Heredity. 33: 229-239.

Hill, W.G. 1981. Estimation of effective population size from data on linkage disequilibrium. Genet. Res. 38: 209-216.

Hudson, R. R. 1985. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics. 109: 611-631.

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

Huttley, G. A., Smith, M. W., Carrington, M., and O'Brien, S. J. 1999. A scan for linkage disequilibrium across the human genome. Genetics. 152: 1711-1722.

Jansen, R. C. 1993. Interval mapping of multiple quantitative trait loci. Genetics. 135: 205-211.

Jeffreys, A. J., Kauppi, L., and Neumann, R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. 29: 217-222.

Kajiwara, K., Berson, E. L., and Dryja, T. P. 1994. Digenic retinitis-pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. Science. 264: 1604-1608.

Kaplan, N. L., Hill, W. G., and Weir, B. S. 1995. Likelihood methods for locating disease genes in nonequilibrium populations. Am. J. Hum. Genet. 56: 18-32.

Kappes, S. M., Keele, J. W., Stone, R. T., Sonstegard, T. S., Smith, T. P. L., McGraw, R. A., LopezCorrales, N. L., and Beattie, C. W. 1997. A second-generation linkage map of the bovine genome. Genome Res. 7: 235-249. Available: http://www.ri.bbsrc.ac.uk/cgi-bin/mapviewer?species=cattle. Accessed Oct. 12, 2001.

Kendall, M. G., and Stuart, A. 1961. The advanced theory of statistics. Volume 2. Inference and relationship. Charles Griffin and Company Limited, 4th edition, Great Britain.

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

Korstanje, R., and Paigen, B. 2002. From QTL to gene: the harvest begins. Nat. Genet. 31: 235-236.

Kruglyak, L., and Lander, E. S. 1995. Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am. J. Hum. Genet. 57: 439-454.

Kruglyak, L., Daly, M. J., Reevedaly, M. P., and Lander, E. S. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. 58: 1347-1363.

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. 22: 139-144.

Kuhner, M. K., Yamato, J., and Felsenstein, J. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics. 149: 429-434.

Laan, M., and Paabo, S. 1997. Demographic history and linkage disequilibrium in human populations. Nat. Genet. 17: 435-438.

Lam, J. C., Roeder, K., and Devlin, B. 2000. Haplotype fine mapping by evolutionary trees. Am. J. Hum. Genet. 66: 659-673.

Lander, E. S. and Botstein, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 121: 185-199.

Lander, E. S., and Schork, N. J. 1994. Genetic dissection of complex traits. Science. 265: 2037-2048.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C.,

Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H. M., Yu, J., Wang, J., Huang, G. Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S. Z., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H. Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W. H., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J. R., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh,

R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., and Morgan, M. J. 2001. Initial sequencing and analysis of the human genome. Nature. 409: 860-921.

Langdahl, B. L., Carstens, M., Stenkjaer, L., and Eriksem, E. F. 2003. Polymorphisms in the transforming growth factor Beta 1 gene and osteoporosis. Bone. 32: 297-310.

Lazzeroni, L. C. 1998. Linkage disequilibrium and gene mapping: an empirical least-squares approach. Am. J. Hum. Genet. 62: 159-170.

Lebowitz, R. J., Soller, M., and Beckmann, J. S. 1987. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor. Appl. Genet. 73: 556-562.

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; Heterotic models. Genetics. 49: 49-67.

Little, R. D., Carulli, J. P., Del Mastro, R. G., Dupuis, J., Osborne, M., Folz, C., Manning, S. P., Swain, P. M., Zhao, S., Eustace, B., Lappe, M. M., Spitzer, L., Zweier, S., Braunschweiger, K., Benchekroun, Y., Hu, X., Adair, R., Chee, L., FitzGerald, M. G., Tulig, C., Caruso, A., Tzellas, N., Bawa, A., Franklin, B., McGuire, S., Nogues, X., Gong, G., Allen, K. M., Anisowicz, A., Morales, A. J., Lomedico, P. T., Recker, S. M., Van Eerdewegh, P., Recker, R. R., and Johnson, M. L. 2002. A mutation in the LDL receptor-related protein 5 gene results in the autosomal dominant high-bone-mass trait. Am. J. Hum. Genet. 70: 11-19.

Liu, J. S., Sabatti, C., Teng, J., Keats, B. J. B., and Risch, N. 2001. Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res. 11: 1716-1724.

Luria, S. E., and Delbruck, M. 1943. Mutations of bacteria from virus sensitivity to virus resistance. Genetics. 28: 491-511.

Lynch, M., and Walsh, B. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, 1st edition, USA.

Macgregor, S., Visscher, P. M., Knott, S., Porteous, D., Muir, W., Millar, K., and Blackwood, D. 2002. Is schizophrenia linked to chromosome 1q?. Science. 298:U1.

Mackay, T. F. C. 2001. Quantitative trait loci in Drosophila. Nat. Rev. Genet. 2: 11-20.

Martin, E. R., Kaplan, N. L., and Weir, B. S. 1997. Tests for linkage and association in nuclear families. Am. J. Hum. Genet. 61: 439-448.

McKeigue, P. M. 2001. Efficiency of estimation of haplotype frequencies: Use of marker phenotypes of unrelated individuals versus counting of phase-known gametes. Am. J. Hum. Genet. 67: 1626-1627.

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

McRae, A.F., McEwan, J.C., Dodds, K.G., Wilson, T., Crawford, A.M., and Slate, J. 2002. Linkage disequilibrium in domestic sheep. Genetics. 160: 1113-1122.

Meuwissen, T. H. E., and Goddard, M. E. 2001. Prediction of identity-by-decent probabilities from marker haplotypes. Genet. Select. Evol. 33: 605-634.

Moffatt, M. F., Traherne, J. A., Abecasis, G. R., and Cookson, W. O. C. M. 2000 . Single nucleotide polymorphism and linkage disequilibrium within the TCR Alpha/Delta locus. Hum. Mol. Gen. 9: 1011-1019.

Mohlke, K. L., Lange, E. M., Valle, T. T., Ghosh, S., Magnuson, V. L., Silander, K., Watanabe, R. M., Chines, P. S., Bergman, R. N., Tuomilehto, A., Collins, F. S., and Boehnke, M. 2001 . Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. Genome Res. 11: 1221-1226.

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

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

Nielsen, D. M., and Weir, B. S. 1999. A classical setting for associations between markers and loci affecting quantitative traits. Genet. Res. 74: 271-277.

Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., Britton, H,, Moran, T., Karaliuskas, R., Duerr, R. H., Achkar, J. P., Brant, S. R., Bayless, T. M., Kirschner, B. S., Hanauer, S. B., Nunez, G., and Cho, J. H. 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature. 411: 603-606.

Osier, M., Pakstis, A. J., Kidd, J. R., Lee, J. F., Yin, S. J., Ko, H. C., Edenberg, H. J., Lu, R. B., and Kidd, K. K. 1999. Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. Am. J. Hum. Genet. 64: 1147-1157.

Ott, J. 1999. Analysis of Human Genetic Linkage. The Johns Hopkins University Press, 3rd edition, Baltimore.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., and Tanksley, S. D. 1988. Resolution of quantitative traits into mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature. 335: 721-726.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., Mcdonough, D. P., Nguyen, B. T. N., Norris, M.

C., Sheehan, J. B., Shen, N. P., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. A., and Cox, D. R. 2001. Blocks of limited haplotype diversity revealed by high- resolution scanning of human chromosome 21. Science. 294: 1719-1723.

Peltonen, L., Jalanko, A., and Varilo, T. 1999. Molecular genetics of the Finnish disease heritage. Hum. Mol. Gen. 8: 1913-1923.

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

Pritchard, J. K., and Rosenberg, N. A. 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. 65: 220-228.

Pritchard, J. K., Stephens, M., and Donnelly, P. 2000a. Inference of population structure using multilocus genotype data. Genetics. 155: 945-959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. 2000b. Association mapping in structured populations. Am. J. Hum. Genet. 67: 170-181.

Pritchard, J. K. 2001. Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. 69: 124-137.

Rannala, B. and Slatkin, M. 1998. Likelihood analysis of disequilibrium mapping, and related problems. Am. J. Hum. Genet. 62: 459-473.

Rannala, B., and Reeve, J. P. 2001. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am. J. Hum. Genet. 69: 159-178.

Rannala, B., and Slatkin, M. 2000. Methods for multipoint disease mapping using linkage disequilibrium. Genet. Epidemiol. 19: S71-S77.

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.

Reich, D. E., and Lander, E. S. 2001. On the allelic spectrum of human disease. Trends Genet. 17: 502-510.

Risch, N., and Merikangas, K. 1996. The future of genetic studies of complex human diseases. Science. 273: 1516-1517.

Risch, N. J., and Zhang, H. 1996. Mapping quantitative trait loci with extreme discordant sib pairs: Sampling considerations. Am. J. Hum. Genet. 58: 836-843.

Ron, M., Kliger, D., Feldmesser, E., Seroussi, E., Ezra, E., and Weller, J. I. 2001. Multiple quantitative trait locus analysis of bovine chromosome 6 in the Israeli Holstein population by a daughter design. Genetics. 159: 727-735.

Roses, A. D. 2000. Pharmacogenetics and the practice of medicine. Nature. 405: 857-865.

Schaid, D. J., Mcdonnell, S. K., Wang, L., Cunningham, J. M., and Thibodeau, S. N. 2002. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. Am. J. Hum. Genet. 71: 992-995.

Schneider, S., Roessli, D., and Excoffier, L. 2000. Arlequin: A software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva. Available: http://lgb.unige.ch/arlequin/. Accessed Nov. 21, 2001.

Schork, N. J., Nath, S. K., Fallin, D., and Chakravarti, A. 2000. Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects. Am. J. Hum. Genet. 67: 1208-1218.

Sham, P. C., Purcell, S., Cherny, S. S., and Abecasis, G. R. 2002. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am. J. Hum. Genet. 71: 238-253.

Slatkin, M. 1994. Linkage disequilibrium in growing and stable populations. Genetics. 137: 331-336.

Slatkin, M., and Excoffier, L. 1996. Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. Heredity. 76: 377-383.

Slatkin, M., and Bertorelle, G. 2001. The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics. 158: 865-874.

Sobel, E., and Lange, K. 1996. Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Genet. 58: 1323-1337.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. 1993. Transmission test for linkage disequilibrium - the insulin gene region and insulin-dependent diabetes-mellitus (IDDM). Am. J. Hum. Genet. 52: 506-516.

Stephens, M., Smith, N. J., and Donnelly, P. 2001. A new method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68: 978-989.

Strachan, T., and Read, A.P. 1999. Human molecular genetics. BIOS Scientific Publishers Ltd, 2nd edition, Oxford. UK.

Sved, J. A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Pop. Biol. 2: 125-141.

Takahata, N. 1994. Repeated failures that led to the eventual success in human- evolution. Mol. Biol. Evol. 11: 803-805.

Terwilliger J. D. 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am. J. Hum. Genet. 56: 777-787.

Varilo, T., Paunio, T., Parker, A., Perola, M., Meyer, J., Terwilliger, J.D., and Peltonen, L. 2003. The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. Hum. Mol. Gen. 12: 51-59.

Visscher, P. M., Thompson, R., and Haley, C. S. 1996. Confidence intervals in QTL mapping by bootstrapping. Genetics. 143: 1013-1020.

Vitart, V., Suffolk, R., Teague, P., Carothers, A., Campbell H., and Wright, A. 2003. Genetic characterisation of Scottish regional subpopulations. Book of abstracts of the conference: Genetics of complex diseases and isolated populations. Poster 72. Accessed at http://www.genosconference.it/poster72.htm.

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., and Jin, L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am. J. Hum. Genet. 71: 1227-1234.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N.,

Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., Mccarthy, M., Mccombie, W. R., Mclaren, S., Mclay, K., Mcpherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'connor, M. J., Okazaki, Y., Oliver, K., Larty, E. O., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Strange-Thomann, N. S., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Vidal, A. U., Vinson, J. P., Von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S. P., Zdobnov, E. M., Zody, M. C., and Lander, E. S. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420: 520-562.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. 7: 256-276.

Weir, B. S., and Hill, W.G. 1980. Effect of mating structure on variation in linkage disequilibrium. Genetics. 95: 477-488.

Weiss, K. M., and Terwilliger, J. D. 2000. How many diseases does it take to map a gene with SNPs?. Nat. Genet. 26: 151-157.

Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika. 61: 439-447.

Wiener, P., Maclean, I., Williams, J. L., and Woolliams, J. A. 2000. Testing for the presence of previously identified QTL for milk production traits in new populations. Anim. Genetics. 31: 385-395.

Wright, A. F., Carothers, A. D., and Pirastu, M. 1999. Population choice in mapping genes for complex diseases. Nat. Genet. 23: 397-404.

Wright, A. F., and Hastie, N. D. 2001. Complex genetic diseases: controversy over the Croesus code. Genome Biol. 2: 2007.1-2007.8.

Wright, A. F., Carothers, A. D., and Campbell, H. 2002. Gene-environment interactions-the BioBank UK study. Pharmacogenomics J. 2: 75-82.

Xiong, M. M., and Guo, S. W. 1997. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am. J. Hum. Genet. 60: 1513-1531.

Xu, S. H., and Atchley, W. R. 1995. A random model approach to interval mapping of quantitative trait loci. Genetics. 141: 1189-1197.

Zavattari, P., Deidda, E., Whalen, M., Lampis, R., Mulargia, A., Loddo, M., Eaves, I., Mastio, G., Todd, J. A., and Cucca, F. 2000. Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography chromosome recombination frequency and selection. Hum. Mol. Gen. 9: 2947-2957.

Zeng, Z. B. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA. 90: 10972-10976.

Zhang, H. P., and Risch, N. 1996. Mapping quantitative-trait loci in humans by use of extreme concordant sib pairs: selected sampling by parental phenotypes. Am. J. Hum. Genet. 59: 951-957.

Zhang, K., Deng, M. H., Chen, T., Waterman, M. S., and Sun, F. Z. 2002. A dynamic programming algorithm for haplotype block partitioning. Proc. Natl. Acad. Sci. USA. 99: 7335-7339.