Use of linkage disequilibrium for quantitative trait loci mapping in livestock

by

Honghua Zhao

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics (Computational Molecular Biology)

Program of Study Committee:
Jack C. M. Dekkers, Major Professor
Max F. Rothschild
Rohan L. Fernando
Daniel S. Nettleton
James M. Reecy

Iowa State University

Ames, Iowa

2006

UMI Number: 3243536

UMI Microform 3243536

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

It is a blessing that I met my major professor, Dr. Jack Dekkers, soon after I came to ISU. I deeply appreciate the time and effort he has invested in my graduate career. Without his support and scientific guidance, none of this would have been possible. His wisdom, encouragement, understanding, trust and patience are unforgettable.

Appreciation is also expressed to the other members of my program of study committee. Dr. Rohan Fernando gave me great help in the development of software used in each project; Dr. Dan Nettleton put my meeting time in his busy schedule for several semesters and gave great guidance; Dr. Max Rothschild always replied my email in one minute which is amazing; and Dr. James Reecy gave me excellent lab training during my rotation. All of them are very willing to help whenever I need assistance. Special thanks also go to Dr. Morris Soller for wonderful discussions.

I wish to thank all my fellow graduate students, post-docs, staff and faculties in our department. It’s a pleasure to work with all of them.

This thesis is lovingly dedicated to my parents, husband and daughters. Their understanding, support and everlasting love made the completion of this work possible.

ABSTRACT

The goal of quantitative trait loci (QTL) mapping in livestock is to find genes underlying traits of economic importance for genetic improvement through marker assisted selection (MAS). The studies presented in this thesis address several important issues in QTL detection and fine mapping using candidate gene analysis and linkage disequilibrium (LD) mapping using high density genotyping. Tests for candidate genes in F2 populations for QTL mapping were developed and evaluated. Results show that the extensive between-breed LD that is present in a cross can result in significant associations for candidate genes at considerable distances from the QTL. Tests that removed the impact of between-breed LD were not powerful in detecting candidate genes closely linked to the QTL, unless the candidate gene was the QTL. Therefore, candidate gene tests in QTL mapping populations must be interpreted with caution. Effectiveness of QTL mapping and MAS using LD in outbred populations depends on the extent of LD between markers and QTL which can differ between populations. Nine measures of LD between multi-allelic markers were evaluated as predictors of usable LD when LD is generated by drift. A standardized chi-square statistic (χ^{2}') was found to be the best predictor of usable LD of multi-allelic markers with QTL, while three other measures (χ^{2}_{df}, r^{2} and D^{*}) were found to be good predictors of usable LD of single nucleotide polymorphisms (SNPs) with QTL. The effect of various factors on power and precision of QTL detection was evaluated and power and precision of regression- and identical by descent (IBD)-based LD mapping methods were compared. Power and precision of QTL detection increased with sample size, marker density and QTL effect.

Single marker regression had similar or greater power and precision than other regression models. For IBD methods, fitting a 4-SNP haplotype, in general, resulted in relatively high power and the greatest mapping precision among the haplotype sizes. Single marker regression was comparable to the 4-SNP IBD method. The results for the haplotype regression and the IBD method assume that haplotypes are known, which would not be true in practice. This will obviously reduce power of these methods. Thus, for rapid initial screening, QTL can be detected and mapped by regression on SNP genotypes without recovering haplotypes with adequate sample size. LD mapping using high density genotyping in outbred populations is a promising method for QTL detection and fine mapping, and would result in markers that can immediately be implemented for MAS.

CHAPTER 1. GENERAL INTRODUCTION

INTRODUCTION

Quantitative traits, such as crop yield, weight gain in animals and fat content of meat, are controlled by several genes along with environmental factors, and with their collective effect represented by phenotypes. Although genetic improvement in livestock has been made by selection on phenotypes, it can be further enhanced through marker-assisted selection (MAS) if we could identify some of the genes that affect the trait, so-called quantitative trait loci (QTL) (Dekkers and Hospital 2002). Most QTL can not be observed at the DNA level. However, genetic markers that are linked to QTL can be used to detect QTL by identifying statistical associations between marker genotypes or alleles and phenotype, and to indirectly select for QTL, which is the concept behind MAS (Andersson 2001; Dekkers and Hospital 2002; Weller 2001). The use of markers for this purpose not only requires access of markers that are linked to the QTL, but also population or data structures in which markers are in linkage disequilibrium (LD) with the QTL. Linkage disequilibrium is the condition in which alleles at two loci are not independent. The LD between markers and QTL forms the basis for QTL detection. It can be created in a population by mutation, selection, drift, and crossing, and is broken down by recombination. The nature of LD used in different strategies for QTL mapping in livestock will be discussed below, and forms the basis for the work that will be described in this thesis.

Many statistical methods for QTL mapping have been developed in livestock, including least squares interval mapping in breed crosses and co-segregation analysis, candidate gene analysis, and LD mapping in outbred populations. In livestock, crosses (F2 or

backcrosses) between outbred breeds have been extensively used as the main resource populations for QTL mapping, in particular in poultry and swine, and some in cattle (Andersson *et al*. 1994; De Koning *et al*. 1999; Malek *et al.* 2001a, 2001b; Rohrer and Keele 1998a, 1998b; Zhou *et al*. 2001). One powerful approach to detect QTL in such populations is least squares interval mapping (Haley *et al.* 1994). QTL detected in crosses cannot be directly used for MAS, which is conducted within populations. The breed cross genome scan approach capitalizes on LD that is generated from crossing two breeds that differ in frequencies of marker and QTL alleles. After only one generation of recombination, extensive population-wide LD may still exist between markers and QTL among the progeny and allows identification of QTL regions, but with poor mapping resolution because LD extends over longer distances; *i.e.* confidence intervals of estimates of QTL position tend to be large (20-30cM, Olsen *et al*. 2004). Thus, fine mapping methods are needed to identify the causative gene or closely linked markers.

In populations that have been closed for many generations, alleles at linked loci are expected to be in linkage equilibrium (LE). However, LD always exists within families (Dekkers 2003). This within-family LD can be used to detect QTL by co-segregation analysis using half-sib or full-sib families in outbred populations (Fernando 2004). This approach requires marker effects to be fitted on a within-family basis. Mapping resolution with this strategy is, however, also poor due to the extensive within-family LD. Mapping resolution can be improved by using extended pedigrees.

Although recombination will tend to move outbred populations to LE, even for linked loci, population-wide LD can exist between closely linked loci, which forms the basis for fine mapping of QTL (Dekkers 2003). Thus, when markers that are close enough to the QTL

are analyzed, one can find associations between marker alleles and phenotype across the population. There are two strategies to find markers close enough to QTL: candidate gene analysis and LD mapping (Dekkers 2003; Dekkers *et al*. 2006). The candidate gene approach is usually conducted within a commercial breeding population. Resulting marker-QTL associations are consistent across families and can be readily implemented for MAS in test populations (Rothschild and Soller 1997). However, only a small part of the genome is covered by this approach (Dekkers *et al*. 2006).

Candidate gene analysis requires phenotyping and genotyping large numbers of animals for traits that are often expensive to record. Many studies have evaluated candidate genes in F2 populations developed for genome scans because of the wealth of phenotypic and genotypic data (Ciobanu *et al*. 2001; Li *et al.* 2003; Nguyen *et al*. 2003; Yu *et al*. 1995; Zhou *et al*. 2001). Given the extensive between-breed LD that exists in F2 populations, candidate genes that are found to have significant associations with phenotype in these populations can be at considerable distances from QTL and, therefore, need to be confirmed in one or more closed mating populations.

Recent advances in high density genotyping have enabled QTL detection and fine mapping in outbred populations using historical recombinations, which is called LD mapping. Similar to candidate gene analysis, resulting QTL can immediately be implemented for MAS (Dekkers and Hospital 2002). The success of LD mapping depends on the extent of LD between markers and QTL and how it declines with distance in a population, but this is often not known. Several statistical methods for LD mapping have been developed and compared (*e.g.* Grapes *et al*. 2004, 2006; Meuwissen and Goddard 2000), but a generally simple and efficient method has not been agreed upon.

RESEARCH OBJECTIVES

The work presented in this thesis investigates different strategies of QTL mapping in livestock, focusing on QTL fine mapping using candidate gene analysis and linkage disequilibrium (LD) mapping. The objective of the candidate gene project described in Chapter 2 is to evaluate the potential of positional candidate gene tests for fine mapping of QTL in crosses between outbred lines of livestock that have been used for genome scans. The aim of the LD measure projects discussed in Chapters 3 and 4 is to evaluate alternative measures of LD between multi-allelic markers as predictors of usable LD of markers (microsatellite or SNPs) with QTL when LD is generated by drift. The goal of the LD mapping project presented in Chapter 5 is to evaluate the effect of various factors on power and precision of QTL detection and to compare power and precision of regression- and IBD-based LD mapping methods using high-density SNP genotyping in outbred populations. The universal goal of these projects is to find simple and efficient approaches which lead to rapid and accurate identification of QTL responsible for traits of interest in livestock in order to enhance genetic progress through marker assisted selection.

THESIS ORGANIZATION

The rest of this chapter provides a literature review for further background relevant to the research conducted. The remainder of this thesis is organized into four individual papers which either have been published or will be submitted to scientific journals. The author of this thesis serves as first author for the four papers.

Chapter 2 consists of the paper “Tests of candidate genes in breed cross populations for QTL mapping in livestock”. This paper was published in *Mammalian Genome*14: 472-482

(2003) and was conducted by Honghua Zhao under the direction of Drs. Jack Dekkers, Rohan Fernando and Max Rothschild.

Chapter 3 consists of the paper "Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL". This paper was published in *Genetical Research* 86: 77-87 (2005) and was conducted by Honghua Zhao under the direction of Drs. Jack Dekkers, Dan Nettleton and Morris Soller.

Chapter 4 consists of the paper "Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms and QTL". This paper will be submitted to *Genetical Research* and was conducted by Honghua Zhao under the direction of Drs. Jack Dekkers and Dan Nettleton.

Chapter 5 consists of the paper "Power and precision of alternate methods for linkage disequilibrium mapping of QTL in livestock". This paper will be submitted to *Genetics* and was conducted by Honghua Zhao under the direction of Drs. Jack Dekkers and Rohan Fernando.

Chapter 6 provides general conclusions and discussion based on the projects described in Chapters 2 through 5.

LITERATURE REVIEW

In recent years, several strategies have been developed for QTL mapping in livestock. One powerful approach to detect QTL in crosses like F2 populations is least squares interval mapping (Haley *et al.* 1994). In this approach, genetic markers spread over the genome are used to identify genomic regions that harbor QTL. Phenotypic values are regressed on

additive and dominance coefficients at a putative QTL position, which are calculated by tracing marker alleles in the progeny back to the parental breeds (Haley *et al.* 1994).

QTL mapping can be carried out in outbred populations using co-segregation analysis. In half-sib and full-sib families and extended pedigrees, markers linked to QTL provide co-segregation information, which can be used for QTL detection by modeling covariances between effects of the QTL (Fernando 2004). For example, if two half-sibs receive the same marker allele from their common parent, then they are likely to receive the same allele from a QTL that is closely linked to this marker. This causes them to be more highly correlated than two sibs that receive different marker alleles (Fernando 2004).

Both approaches described above result in wide QTL regions and further research is needed to identify the causative gene or closely linked markers. There are two strategies to find markers close enough to QTL for population-wide linkage disequilibrium (LD). One is candidate gene analysis and the other is LD mapping using a high-density marker map (Dekkers 2003; Dekkers *et al*. 2006). The success of both approaches depends on the extent of LD in a population (Dekkers 2003). The projects reported in this thesis (Chapters 2-5) are related to these research areas. Therefore, this literature review will focus on these topics.

Candidate gene analysis

Candidate gene analyses evaluate markers that are in or close to genes that are thought to be associated with the trait of interest based on their biological role, mutational analysis, location in a QTL region (positional candidate genes), comparative data or gene expression data (Rothschild *et al*. 2003).

Candidate gene analysis is typically conducted through an association study within breeds or lines because LD is expected to extend over short distances. The association of genotype at the candidate gene with phenotype can be estimated using a linear model with the candidate genotype as a fixed effect. Significant results imply that the candidate gene is at least closely linked to the QTL affecting phenotype within the breed or, ideally the candidate gene is the QTL. Associations that are uncovered in a given population must be confirmed in other populations (Rothschild and Soller 1997).

Instead of a dense marker map, the candidate gene approach only needs carefully placed markers (*i.e.* in candidate genes) to detect QTL. Candidate gene analysis can be based on a random sample of individuals from a population and, in principle, does not require pedigree information. However, pedigree would be needed if individuals are related and the model includes a polygenic breeding value effect. Resulting candidate genes can immediately be implemented for MAS in the test populations (Rothschild and Soller 1997). The real beauty of a candidate gene analysis, as pointed out by Rothschild and Soller (1997), is that it requires a researcher's knowledge and intuition for selecting new possible candidate genes of interest, as demonstrated in the extensive candidate gene studies conducted by the Rothschild laboratory (*e.g.* Kim *et al*. 2000; Rothschild *et al*. 1996).

Candidate gene analysis has been a robust method to identify genes involved in reproductive performance of pigs (Rothschild and Soller 1997). A clear example is the study of the estrogen receptor gene for litter size in pigs (Rothschild *et al*. 1996; Short *et al*. 1997). Short *et al*. (1997) showed a significant effect of the estrogen receptor gene on litter size in commercial Large White lines. Other candidate genes that have been found to be associated with litter size in pigs include retinol binding protein 4 (RBP4, Rothschild *et al*. 2000),

prolactin receptor (PRLR, Vincent *et al*. 1998) and the beta subunit of follicle stimulating hormone (FSHB, Li *et al*. 1998).

Kim *et al*. (2000) studied melanocortin-4 receptor gene (MC4R) as a candidate gene in a number of pig lines for the control of growth and performance traits that are important in the pig. Significant associations of MC4R genotypes with backfat, growth rate and feed intake were revealed (Kim *et al*. 2000).

Many candidate gene analyses have been conducted in F2 crosses developed for QTL mapping because of the availability of extensive phenotypes and genotypes (Ciobanu *et al*. 2001; Li *et al.* 2003; Nguyen *et al*. 2003; Yu *et al*. 1995; Zhou *et al*. 2001). Analysis of positional candidate genes in such populations is, however, complicated by the extensive between-breed LD that is created in the cross. While it is essential for QTL interval mapping, such extensive LD may result in a significant association for candidate genes that are at considerable distance from the QTL. Yu *et al*. (1995) conducted a candidate gene analysis for PIT1 in five F2 families from crosses between Chinese and Western breeds. Significant associations of PIT1 with birth weight and backfat traits were identified. Further QTL analysis in these families identified a QTL for birth weight at PIT1, but a QTL for backfat was at least 20 cM from PIT1 (Yu *et al*. 1999). How to remove the impact of between-breed LD on tests of candidate genes in breed crosses, such that significant associations are identified only if the candidate gene marker is closely linked to the QTL, prompted the work reported in Chapter 2.

LD in outbred populations

In outbred populations, the main factors that create LD are mutation, selection and drift. To illustrate generation of LD by mutation, assume a QTL is introduced into a base population as a mutation on a single ancestral haplotype. After many generations of random mating, the original haplotype will remain only for markers close to the QTL because of lack of recombination. Thus, in the current generations, only tightly linked markers will still be in strong LD with the QTL (Meuwissen and Goddard 2000; Olsen *et al*. 2004). Although selection also causes LD (Bulmer 1971), it preferentially generates LD between QTL affecting the selected trait rather than between markers and QTL (Farnir *et al*., 2000).

Random drift plays an important role in generating LD in livestock breeding populations, which are typically of limited size (Flint-Garcia *et al.*, 2003; Terwilliger *et al.* 1998). LD generated by the balance of drift and recombination is expected to equal $1/(1+4N_e c)$ (Sved 1971), where N_e is the effective population size (Falconer and MacKay 1996) and c is the recombination rate. Because of sampling, drift creates a random pattern of LD around the QTL, without distinct haplotype signatures (Dekkers *et al.* 2006; Terwilliger *et al.* 1998). Terwilliger *et al*. (1998), Farnir *et al*. (2002) and Andersson and Georges (2004) suggested that LD mapping in livestock might be more effective than in humans because of the extensive LD that is created by drift as a result of limited N_e of most livestock populations.

A crucial issue in using LD for whole genome scans is the extent of LD between markers and QTL, which is needed to determine the marker density and impacts the power and resolution of LD mapping and effectiveness of MAS (Harmegnies *et al*. 2006). Because QTL cannot be observed directly, LD between markers can be used to predict marker-QTL

LD, in order to evaluate the extent of useful LD in a population (*e.g.* Farnir *et al.* 2000; Harmegnies *et al*. 2006; Pritchard and Przeworski 2001).

The standard measure of LD between two alleles at two different loci is $D_{ij} = p(A_i B_j) - p(A_i)\,p(B_j)$, where $p(A_i)$ is the frequency of allele A_i at locus *A*, $p(B_j)$ the frequency of allele B_j at locus *B*, and $p(A_i B_j)$ the frequency of haplotype $A_i B_j$. For loci with two alleles, D_{ij} completely describes LD between all pairs of alleles. Because D_{ij} depends on gene frequencies, Lewontin (1964) suggested standardizing D_{ij} by the maximum absolute value it can attain, given the allele frequencies: $\left|D'_{ij}\right| = \left|\dfrac{D_{ij}}{D_{ij}^{\max}}\right|$,

where $D_{ij}^{\max} = \min\left[\, p(A_i)\, p(B_j), (1 - p(A_i))(1 - p(B_j))\right]$ when $D_{ij} < 0$,

$D_{ij}^{\max} = \min\left[\, p(A_i)(1 - p(B_j)), (1 - p(A_i))\, p(B_j)\right]$ when $D_{ij} \geq 0$.

Hill and Robertson (1968) suggested using the square of the correlation between A_i and B_j as a standardized measure of LD between biallelic loci:

$$r_{ij}^2 = \frac{D_{ij}^2}{p(A_i)\,(1 - p(A_i))\;p(B_j)\,(1 - p(B_j))}.$$

For biallelic markers, the absolute value of LD is the same between any pair of alleles across two loci. The two most common LD measures used for biallelic markers are *D'* = $\left|D'_{11}\right|$ and $r^2 = r_{11}^2$ (Ardlie *et al.* 2002; Hill and Robertson 1968; Lewontin 1964). Current research appears to prefer r^2 for detecting biallelic markers that might correlate with QTL of interest (Ardlie *et al.* 2002; Flint-Garcia *et al.* 2003).

Compared to biallelic markers, assessing the degree of LD between multi-allelic markers is more complicated, because LD can differ between pairs of alleles and a combined

measure of LD across alleles is needed. A commonly used measure is $D' = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_i)\, p(B_j)\left|D'_{ij}\right|$ (Hedrick 1987), where *k* and *m* are the numbers of alternate alleles at locus *A* and *B*, respectively. However, it is known that LD measured by *D'* tends to be inflated with small sample sizes and/or low allele frequencies (Ardlie *et al.* 2002; Flint-Garcia *et al.* 2003; McRae *et al*. 2002). Using *D'*, extensive LD over a long range was observed in dairy cattle, sheep and pigs (Farnir *et al*. 2000; McRae *et al*. 2002; Nsengimana *et al.* 2004; Tenesa *et al*. 2003), but it is not clear to what extent this was a result of the above artifact.

Although a variety of statistics have been proposed (Hedrick and Thomson 1986; Hedrick 1987; Sabatti and Risch 2002; Yamazaki 1977), a generally satisfactory measure of LD between multi-allelic markers has not been agreed upon. Alternate measures of LD among multi-allelic markers must be compared for their ability to predict the extent of usable LD for QTL mapping or MAS. The work presented in Chapters 3 and 4 addresses these questions.

LD mapping

Recent advances in technology have made large-scale SNP genotyping rapid, accurate, and inexpensive (Kwok 2001). High density SNP maps are now available for both human and livestock. For example, a SNP map of the human genome containing 1.42 million SNPs (International SNP Map Working Group 2001) and a genetic variation map for the chicken genome containing 2.8 million SNPs (International Chicken Polymorphism Map Consortium 2004) have been constructed. High density SNP genotyping has increased the feasibility of

QTL detection and fine-mapping in outbred populations using historical population-wide LD (Grapes *et al.* 2004; Meuwissen and Goddard 2000). Resulting QTL can immediately be implemented for MAS (Dekkers and Hospital 2002).

LD mapping has been used extensively to identify genes for monogenic diseases in humans (Peltonen 2000). Contrary to the situation in human, extensive LD over a long range was observed in dairy cattle, sheep and pigs (Farnir *et al*. 2000; McRae *et al*. 2002; Nsengimana *et al.* 2004; Tenesa *et al*. 2003). Thus, LD mapping in livestock might be effective using marker maps of more limited density than what is required for most human populations because of the extensive LD that is created by drift in livestock as a result of limited effective population sizes compared to humans (Andersson and Georges 2004; Farnir *et al*. 2002; Terwilliger *et al*. 1998).

Several statistical methods for LD mapping have been developed, including random effects methods based on identical by descent (IBD) (Meuwissen and Goddard 2000) and least squares methods based on regression of phenotype on marker genotypes or haplotypes (Grapes *et al*. 2004, 2006). IBD methods model covariances between individuals by deriving IBD probabilities of QTL alleles carried by alternate marker haplotypes under some assumptions about population history (Meuwissen and Goddard 2000). Two individuals with IBD QTL alleles are likely to have higher phenotypic covariance than those without. If the QTL resulted from a single mutation in the founder generation, then, after many generations of recombination, only tightly linked markers will still be in strong LD with the QTL. The IBD probability of a pair of alleles at the putative QTL position increases as the number of markers surrounding the QTL that are consecutively identical in state increases. Meuwissen and Goddard (2001) derived IBD probabilities analytically with assumptions about effective

population size and mutation age. Meuwissen and Goddard (2000), however, showed that mapping precision was robust to these assumptions. Grapes *et al*. (2006) proposed an optimal haplotype size for LD mapping using IBD. They found that using fewer (*e.g.* 4-6) markers in a haplotype to derive IBD resulted in greater mapping precision than using all available markers, because the latter resulted in a flatter likelihood curve that did not discriminate between alternate QTL positions (Grapes *et al*. 2006).

The IBD-based LD mapping methods can be combined with linkage or co-segregation information for fine mapping QTL in livestock (*e.g.* Blott *et al*. 2003; Farnir *et al.* 2002; Meuwissen *et al*. 2002; Olsen *et al*. 2004). Using this combined approach, Meuwissen *et al*. (2002) mapped a QTL within a region <1 cM for twinning rate in large half-sib cattle families.

Compared to IBD methods, regression methods for LD mapping are computationally easier to implement with no assumptions required. Regression on marker genotypes does not need knowledge of marker haplotypes and is, therefore, potentially useful for rapid initial screen for QTL. Grapes *et al*. (2004) showed that regression methods were competitive with IBD methods in terms of accuracy of fine-mapping within a previously identified QTL region. However, they did not exclude SNPs that were fixed in the generation under study from their data analysis, and they did not compare the power of QTL detection between these two approaches (Grapes *et al*. 2004). These limitations prompted the work reported in Chapter 5.

In conclusion, least squares interval mapping in F2 crosses and co-segregation analysis in outbred populations, which are popularly used for QTL mapping in livestock, lead to the identification of wide QTL regions. Further work needs to narrow down QTL regions and to

identify causative genes or closely linked markers. Candidate gene analysis and LD mapping using high-density SNP genotyping in outbred populations are promising QTL fine mapping methods. Some important issues include how to conduct and interpret candidate gene tests in F2 crosses, how to predict the extent of usable LD in a population for QTL mapping, and what are the simple and efficient methods for LD mapping in outbred populations. These questions will be addressed in this thesis.

REFERENCES

Andersson, L. (2001) Genetic dissection of phenotypic diversity in farm animals. Nature Reviews: Genetics 2: 130-138.

Andersson, L. and Georges, M. (2004) Domestic-animal genomics: deciphering the genetics of complex traits. Nature Reviews: Genetics 5: 202-212.

Andersson, L., Haley, C. S., Ellegren, H., Knott, S. A., Johansson, M., et al. (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science **263**: 1771-1774.

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

Blott, S., Kim, J.-J., Moisio, S., Schmidt-Küntzel, A., Cornet, A., et al. (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163: 253-266.

Bulmer, M. G. (1971) The effect of selection on genetic variability. The American Naturalist 105: 201-211.

Ciobanu, D., Bastiaansen, J., Malek, M., Helm, J., Woollard, J., Plastow, G. and Rothschild, M. F. (2001) Evidence for new alleles in the protein kinase adenosine monophosphate-activated γ3-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality. Genetics 159: 1151-1162.

Dekkers, J. C. M., Zhao, H. H. and Fernando, R. L. (2006) Linkage disequilibrium mapping of QTL in livestock. Proceedings of the World Cong. Genet. Appl. Livest. Prod. 8, Brazil (Accepted).

Dekkers, J. C. M. (2003) Principles of QTL mapping. Pre-Proceedings of the 28th Annual National Swine Improvement Federation Conference and Meeting, p23-36.

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

De Koning, D. J., Janss, L. L. G., Rattink, A. P., van Oers, P. A. M., de Vries, B. J., et al. (1999) Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (Sus Scrofa). Genetics 152: 1679-1690.

Falconer, D. S. and Mackay, T. F. C. (1996) Introduction to Quantitative Genetics, 4th edn. Harlow, UK: Addison-Wesley Longman.

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 Research 10: 220-227.

Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P. et al. (2002) Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics 161: 275-287.

Fernando, R. L. (2004) QTL Mapping in Complex Pedigrees. Summer Short Course, Department of Animal Science, Iowa State University.

Flint-Garcia, S. A., Thornsberry, J. M. and Buckler IV, E. S. (2003) Structure of linkage disequilibrium in plants. Annual Review of Plant Biology 54: 357-374.

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

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

Haley, C. S., Knott, S. A. and Elsen, J. M. (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136: 1195-1207. Harmegnies, N., Farnir, F., Davin, F., Buys, N., Georges, M. and Coppieters, W. (2006) Measuring the extent of linkage disequilibrium in commercial pig populations. Animal Genetics 37: 225-231.

Hedrick, P. W. and Thomson, G. (1986) A two-locus neutrality test: application to humans, E. Coli and Lodgepole pine. Genetics 112: 135-156.

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

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

International Chicken Polymorphism Map Consortium (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432: 717-722.

International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.

Kim, K. S., Larsen, N., Short, T., Plastow, G. and Rothschild, M. F. (2000) A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. Mamm. Genome 11: 131-135.

Kwok, P. Y. (2001) Methods for genotyping single nucleotide polymorphisms. Annual Review of Genomics and Human Genetics 2: 235-258.

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

Li, H., Deeb, N., Zhou, H., Mitchell, A. D., Ashwell, C. M. and Lamont, S. J. (2003) Chicken quantitative trait loci for growth and body composition associated with transforming growth factor-β genes. Poult. Sci. 82: 347-356.

Li, N., Zhao, Y. F., Xiao, L., Zhang, F. J., Chen, Y. Z. et al. (1998) Candidate gene analysis for identification of genetic loci controlling litter size in swine. Proceedings of the World Cong. Genet. Appl. Livest. Prod. 6, Armidale, Australia, 26:183-186.

Malek, M., Dekkers, J. C. M., Lee, H. K., Baas, T. J. and Rothschild, M. F. (2001a) A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. I. Growth and body composition. Mamm Genome 12: 630-636.

Malek, M., Dekkers, J. C. M., Lee, H. K., Baas, T. J., Prusa, K., Huff-Lonergan, E. and Rothschild, M. F. (2001b) A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. II. Meat and muscle composition. Mamm Genome 12: 637-645.

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. (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155: 421-430.

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

Meuwissen, T. H. E., Karlsen, A., Lien, S., Olsaker, I. and Goddard, M. E. (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161: 373-379.

Nguyen, N. T., Kim, K. S., Thomsen, H., Helm, J. and Rothschild, M. F. (2003) Investigation of a candidate gene for growth and fatness QTL on the pig chromosome 7. Proceedings of Plant and Animal and Genome XI conference, San Diego, CA, p230.

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

Olsen, H. G., Lien, S., Svendsen, M., Nilsen, H., Roseth, A., Aasland Opsal, M. and Meuwissen, T. H. E. (2004) Fine mapping of milk production QTL on BTA6 by combined linkage and linkage disequilibrium analysis. J. Dairy Sci. 87: 690-698.

Peltonen, L. (2000) Positional cloning of disease genes: advantages of genetic isolates. Hum. Hered. 50: 66-75.

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

Rohrer, G. A. and Keele, J. W. (1998a) Identification of quantitative trait loci affecting carcass composition in swine: I. Fat deposition traits. J Anim Sci 76: 2247-2254.

Rohrer, G. A. and Keele, J. W. (1998b) Identification of quantitative trait loci affecting carcass composition in swine: II. Muscling and wholesale product yield traits. J Anim Sci 76: 2255-2262.

Rothschild, M. F., Jacobson, C., Vaske, D., Tuggle, C. K., Wang, L. et al. (1996) The estrogen receptor locus is associated with a major gene influencing litter size in pigs. Proc. Natl. Acad. Sci. USA 93: 201-205.

Rothschild, M. F. and Soller, M. (1997) Candidate gene analysis to detect genes controlling traits of economic importance in domestic livestock. Probe 8: 13-20.

Rothschild, M. F., Messer, L., Day, A., Wales, R., Short, T., Southwood, O. and Plastow, G. (2000) Investigation of the retinol-binding protein 4 (RBP4) gene as a candidate gene for increased litter size in pigs. Mamm. Genome 11: 75-77.

Rothschild, M. F., Ciobanu, D., Lonergan, S., Dekkers, J. C. M. and Stalder, K. (2003) Identification of genes for carcass merit and meat quality in the pig. Pre-Proceedings of the 28th Annual National Swine Improvement Federation Conference and Meeting, p84-99.

Sabatti, C. and Risch, N. (2002) Homozygosity and linkage disequilibrium. Genetics 160: 1707-1719.

Short, T. H., Rothschild, M. F., Southwood, O. I., McLaren, D. G., de Vries, A. et al. (1997) Effect of the estrogen receptor locus on reproduction and production traits in four commercial pig lines. J. Anim. Sci. 75: 3138-3142.

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

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. Journal of Animal Science 81: 617-623.

Terwilliger, J. D., Zöllner, S., Laan, M. and Pääbo, S. (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. Hum. Hered. 48: 138-154.

Vincent, A. L., Evans, G., Short, T. H., Southwood, O. I., Plastow, G. S., Tuggle, C. K. and Rothschild, M. F. (1998) The prolactin receptor gene is associated with increased litter size in pigs. Proceedings of the World Cong. Genet. Appl. Livest. Prod. 6, Armidale, Australia, 27:15.

Weller, J. I. (2001) Quantitative Trait Loci Analysis in Animals. CABI publishing.

Yamazaki, T. (1977) The effects of overdominance on linkage in a multilocus system. Genetics 86: 227-236.

Yu, T.-P., Tuggle, C. K., Schmitz, C. B. and Rothschild, M. F. (1995) Association of PIT1 polymorphisms with growth and carcass traits in pigs. J. Anim. Sci. 73: 1282-1288.

Yu, T.-P., Wang, L., Tuggle, C. K. and Rothschild, M. F. (1999) Mapping genes for fatness and growth on pig chromosome 13: a search in the region close to the pig PIT1 gene. J. Anim. Breed. Genet. 116: 269-280.

Zhou, H., Buitenhuis, A. J., Weigend, S. and Lamont, S. J. (2001) Candidate gene promoter polymorphisms and antibody response kinetics in chickens: interferon-γ, interleukin-2, and immunoglobulin light chain. Poultry Science 80: 1679-1689.

CHAPTER 2. TESTS OF CANDIDATE GENES IN BREED CROSS POPULATIONS FOR QTL MAPPING IN LIVESTOCK

A paper published in *Mammalian Genome*[1]

Honghua Zhao, Max F. Rothschild, Rohan L. Fernando, Jack C.M. Dekkers

Department of Animal Science and Center for Integrated Animal Genomics,

Iowa State University

ABSTRACT

In recent years, several F_2 crosses between outbred lines of livestock have been developed to identify quantitative trait loci (QTL). These populations are valuable for further genetic analysis, including of positional candidate gene loci (CGL). Analysis of CGL in F_2 populations is, however, hindered by extensive between-breed linkage disequilibrium (LD). The objectives here were to develop and evaluate three tests for CGL in simulated F_2 breed-cross populations. 1) A standard association test, based on the fixed effect of CGL genotype. This test was significant for CGL at considerable distances from the QTL. 2) A marker-assisted association test, based on a test at the CGL of the fixed effect of CGL genotype in a breed-cross QTL interval mapping model. This removed the impact of between-breed LD, but was not powerful in detecting CGL closely linked to the QTL, unless the CGL was the QTL. 3) An F-drop test, comparing F ratios for a QTL at the CGL with and without the CGL included as fixed effect. It had low power to distinguish close from distant CGL. Power to

[1] Reprinted with permission of *Mammalian Genome* (2003) 14: 472-482.

distinguish two CGL within 10 cM from the QTL was limited and little improved by including QTL effects associated with markers to remove between-breed LD, although power was greater when one of the CGL was the causative mutation. Therefore, while we conclude that candidate gene tests in QTL mapping populations must be interpreted with caution, we now have a clearer picture of the value of candidate gene tests in these populations.

INTRODUCTION

In recent years, several F_2 populations have been developed from crosses between divergent breeds of livestock to identify chromosomal regions that contain quantitative trait loci (QTL) affecting traits of economic importance by breed-cross QTL interval mapping (Andersson et al. 1994; Rohrer and Keele 1998a, 1998b; De Koning et al. 1999; Malek et al. 2001a, 2001b). These studies result in the identification of QTL regions with wide confidence intervals and further research is needed to identify the causative gene or closely linked markers. Fine mapping of QTL can be conducted in outbred populations by identity by descent (Riquet et al. 1999), linkage disequilibrium (LD), or candidate gene analyses (Rothschild et al. 1996). These approaches require phenotyping and genotyping large numbers of animals for traits that are often expensive to record (e.g. meat quality or disease traits). Against this background, the F_2 populations that have been developed for genome scans provide a valuable resource for further genetic analysis, because they have been phenotyped for many traits and genotyped for many genetic markers. Further analyses that can be and have been conducted in these populations include the analysis of positional (comparative) candidate genes within identified QTL regions (Yu et al. 1995; Ciobanu et al. 2001; Zhou et al. 2001; Li et al. 2003; Nguyen et al. 2003). The aim of positional candidate gene analyses is to

determine whether a particular candidate gene is the QTL or, at least, closely linked to the QTL (Rothschild and Soller 1997). Alternatively, if multiple candidate genes are available in the QTL region, the aim is to determine which gene is closest to the QTL for further analysis in outbred populations.

Statistical analysis of candidate genes is typically conducted through an association study, by testing for associations of genotypes at the candidate gene locus (CGL) with phenotypes for the quantitative trait. A significant association is expected if the CGL is the causative mutation or is in population-wide LD with the causative locus. Conclusions about position of the CGL relative to the causative mutation then depend on the extent of LD that exists in the population under study but this is often not known.

Ideally, candidate gene analyses are conducted within breeds or lines because LD is expected to extend over short distances. A clear example is the study of the estrogen receptor gene for litter size in pigs (Rothschild et al. 1996; Short et al. 1997). Short et al. (1997) showed a significant effect of the estrogen receptor gene on litter size in commercial Large White lines.

In contrast, LD in a QTL mapping population, such as a breed cross, extends over long distances. As a result, an association study may show significant effects on phenotype, even for CGL that are at considerable distances from the QTL. For example, Yu et al. (1995) conducted a candidate gene analysis for PIT1 in five F_2 families from crosses between Chinese and Western breeds and identified significant associations of PIT1 with birth weight and backfat traits. Further QTL analysis in these families identified a QTL for birth weight at PIT1, but a QTL for backfat was at least 20 cM from PIT1 (Yu et al. 1999). Because of the extensive LD that existed in these families, the association studies could not exclude the

possibility that the observed associations with PIT1 were due to QTL that may be at some distance from PIT1 (Yu et al. 1999).

Association tests capitalize on the total LD that exists in the population. In crosses between outbred lines, this LD consists of two components: the LD that exists within the, usually outbred, parental breeds and the LD that is created by the cross. The latter component extends over long distances, which complicates the interpretation of association studies in breed crosses. Extensive LD is, however, essential for QTL interval mapping, e.g. using the least squares (Haley et al. 1994), which capitalizes exclusively on the between-breed LD that is created in the cross. Aspects of breed-cross interval mapping could, however, be utilized to remove the impact of between-breed LD on tests of CGL in breed crosses. In this paper, we develop such tests that integrate aspects of a QTL analysis into an association test for CGL in breed cross populations. The specific objective was to evaluate the potential of positional CGL tests for fine mapping of QTL in crosses between outbred lines of livestock that are used for genome scans. Tests were evaluated by simulation.

MATERIALS AND METHODS

Simulation. A chromosome of 100 cM was simulated with 6 markers 20 cM apart and an additive QTL at 36 cM with a substitution effect of 0.25 phenotypic standard deviations. Polymorphisms for a biallelic positional CGL were simulated at alternate positions, with distances from the QTL ranging from 0 to 30 cM (Fig. 1). Markers and CGL were linked to the QTL but had no direct effect on the trait. A three-generation pedigree was simulated by using 10 F_0 grand sires of one breed and 10 F_0 grand dams of another breed. A total of 600 F_2 progeny from 30 matings of F_1 parents were produced. In order to generate LD between the

QTL and CGL within the outbred parental breeds, 20 generations (F_{-1} to F_{-20}) were added to produce F_0 parents, starting with crosses between pairs of ancestral breeds, A x B and C x D (Fig. 2). Markers were simulated with 5 alleles, with frequencies of 0.6, 0.2, 0.1, 0.05, 0.05 in breeds A and B, and 0.05, 0.05, 0.1, 0.2, 0.6 in breeds C and D. Two alleles were simulated for the QTL (Q and q) and for each CGL (C and c). Four cases were considered for allele frequencies of the QTL and CGL in the four ancestral breeds to produce various combinations of within- and between-breed LD (Table 1). Frequencies of alleles Q and C were equal in the F_{-20}. A total of 1,000 replicates were simulated for each case.

Determination of within- and between-breed LD. Association tests capitalize on LD that exists in the population. Therefore, it was important to quantify the extent and nature of the LD in the simulated F_2 populations. This LD originated from the cross of the F_0 parental breeds (between-breed LD), and from the LD that existed in the F_0 parental breeds as a result of the cross of ancestral breeds in the F_{-20} (within-breed LD).

The expected between-breed LD that exists between the QTL and a CGL in the F_2 is equal to (see Appendix):

$$LD_{between} = \frac{1}{4}(1-2r)[\frac{1}{2}(P_Q^A + P_Q^B) - \frac{1}{2}(P_Q^C + P_Q^D)]^2 \qquad [1]$$

where r is the recombination rate between the QTL and the CGL, and P_k^i is the frequency of allele k in ancestral breed i in the F_{-20}. Between-breed LD is maximal when parental breeds in the F_0 are fixed for alternate alleles at both loci (e.g. $P_k^A = P_k^B = 0$ and $P_k^C = P_k^D = 1$) and equal to: $LD_{between} = \frac{1}{4}(1-2r)$.

Within-breed LD between the QTL and a CGL in the simulated F_2 originated from the cross of ancestral breeds in the F_{-20} and is expected to be equal to (see Appendix):

$$LD_{within} = (1-r)^{20}[\frac{1}{4}(1-2r)(P_Q^A - P_Q^B)^2] \quad [2]$$

Within-breed LD is maximal when ancestral breeds A and B are fixed for alternate alleles and equal to: $LD_{within} = (1-r)^{20}\frac{1}{4}(1-2r)$. The total LD that exists in the F_2 population is the sum of LD_{within} and $LD_{between}$. Note that F_2 populations with maximal between-breed LD ($P_k^A = P_k^B = 0$ and $P_k^C = P_k^D = 1$) have no within-breed LD. This antagonism between within- and between-breed LD will be important for interpretation of results of this study.

Although the QTL is in complete LD with itself, this LD can also be partitioned into within- and between-breed LD, depending on allele frequencies. The expected between-breed LD within the QTL in the F_2 is (see Appendix):

$$LD_{between} = \frac{1}{4}[\frac{1}{2}(P_Q^A + P_Q^B) - \frac{1}{2}(P_Q^C + P_Q^D)]^2 \quad [3]$$

The within-breed LD for the QTL with itself in the F_2 originated not only from the cross of ancestral breeds in the F_{-20}, but also from the LD that existed in ancestral breeds. Its expected value is (see Appendix):

$$LD_{within} = \frac{1}{2}[P_Q^A(1-P_Q^A) + P_Q^B(1-P_Q^B)] + \frac{1}{4}(P_Q^A - P_Q^B)^2 \quad [4]$$

Measures of LD can be normalized to range between 0 and 1 following Lewontin (1964) as $D' = \frac{LD}{LD_{MAX}}$, where LD_{MAX} is the maximum numerical value for LD given the

allele frequencies. In this study, LD_{MAX} is 0.25 because the expected frequencies of alleles Q and C in the F_2 are 0.5 in all cases (Table 1).

Statistical tests. Three statistical tests were investigated to test for associations between CGL and phenotypes in the simulated populations and two tests were investigated to discriminate between alternate CGL. These tests are described below.

Standard association test: The standard test that is used for candidate gene analysis is based on fitting the following two models (Short et al. 1997; Yu et al. 1995):

Model 1: $y = \mu + g_{CGL} + e$

Model 2: $y = \mu + e$

where y is the trait value, μ is the overall mean, g_{CGL} is the effect associated with the CGL, and e represents polygenic and environmental effects. Tests for a significant association of the CGL with phenotype were based on an F ratio of residual sums of squares (RSS) for models 1 and 2. Significance thresholds at $\alpha = 0.05$ were obtained from standard F-tables. A significant test implies that the CGL is the QTL or is linked to the QTL.

Marker-assisted association test: In the least squares method for QTL mapping in crosses between outbred lines (Haley et al. 1994), chromosomal regions that contain QTL are identified based on between-breed LD that exists between markers and the QTL by tracing marker alleles in the F_2 progeny back to the parental (F_0) breeds. This marker information can, therefore, also be used to account for between-breed LD in a CGL analysis in an F_2 cross by fitting the following two models at the CGL:

Model 3: $y = \mu + g_{CGL} + C_a a + C_d d + e$

Model 4: $y = \mu + C_a a + C_d d + e$

where a and d are unknown additive and dominance effects of average QTL alleles by breed origin from the two F_0 breeds, respectively, and C_a and C_d are the additive and dominance coefficients, as computed from marker information (Haley et al. 1994). Using these two models, significance of the CGL effect can be tested based on an F ratio of RSS from models 3 and 4. Since this involves only a single test, significance thresholds at $\alpha = 0.05$ were obtained from standard F-tables. Similar to the standard association test, a significant result suggests that the CGL is linked to the QTL or, ideally, that the CGL polymorphism is the causative gene.

F-drop test: The standard test for a QTL in a genome scan breed cross analysis is based on an F ratio of RSS from models 2 and 4. Inclusion of the effect of the CGL in a QTL mapping analysis, as in model 3, is expected to reduce the F ratio for a QTL if the CGL is in LD with the QTL because the CGL effect is expected to absorb part of the between-breed QTL effect. The drop in F ratio for the QTL can be evaluated by comparing the F ratio for models 1 and 3 to the F ratio for models 2 and 4 at the CGL. Empirical significance thresholds for this test were derived from simulation under the null hypothesis that the QTL and the CGL are in linkage equilibrium. In order to generate data under this null hypothesis, the QTL was simulated with the same frequencies as listed in Table 1, but frequencies of CGL alleles were 0.5 in all F_{-20} breeds in all cases. Other parameters in the simulation were the same as before. A total of 1,000 replicates were generated and analyzed to determine significance thresholds for a 5% one-sided test.

Comparison of two candidate gene loci: In most instances, QTL are mapped to rather broad regions that potentially may harbor multiple CGL. One aim of a candidate gene analysis in an F_2 population, therefore, may be to determine which CGL is closest to the QTL. To evaluate the power to distinguish CGL, two CGL, CGL_1 and CGL_2, at 10 and 1 cM from the QTL (Fig. 1), were compared using the following two alternative statistical tests.

A standard association test can be used to determine which CGL is closer to the QTL by fitting the following two models:

$$y = \mu + g_{CGL_1} + e$$

$$y = \mu + g_{CGL_2} + e$$

The following test statistic, which was derived from the likelihood ratio of the two models, can be used to test for a significant difference between the two CGL:

$$T = \frac{RSS(CGL_2)}{RSS(CGL_1)}$$

where $RSS(CGL_i)$ represents the RSS of the model with CGL_i.

To remove the impact of between-breed LD, a marker-assisted association test was also used to compare the two CGL by fitting the following two models at the position of the respective CGL:

$$y = \mu + g_{CGL_1} + C_a a + C_d d + e$$

$$y = \mu + g_{CGL_2} + C_a a + C_d d + e$$

The same test statistic T was used, with $RSS(CGL_i)$ equal to the RSS of the model fitted at the position of CGL_i.

Since CGL_2 is closer to the QTL than CGL_1, the RSS of the model with CGL_2, denoted $RSS(CGL_2)$, is expected to be smaller than $RSS(CGL_1)$. Therefore, the test statistic, $\frac{RSS(CGL_2)}{RSS(CGL_1)}$, is expected to be less than 1. Empirical significance thresholds for a 5% two-sided test were derived from simulation under the null hypothesis that the two CGL are at equal distances from the QTL. To create such a situation for derivation of the critical value for the test statistic, the QTL was simulated central to the two CGL. Other parameters in the simulation were the same as described before. A total of 1,000 replicates were generated under the null hypothesis.

The CGL at 10 cM from the QTL was also compared with the QTL itself by replacing CGL_2 in the models with the true QTL. The thresholds derived above were used to calculate the power to distinguish between the two loci.

RESULTS

Expected between- and within-breed LD. The expected LD between the QTL and CGL in the simulated F_2 populations is illustrated in Fig. 3 for the four cases of Table 1. Total LD in the F_2 (Fig. 3A) depends on distance between the QTL and CGL and on allele frequencies in the ancestral breeds. Ancestral breed allele frequencies also determine the relative contributions of within- (Fig. 3B) and between-breed (Fig. 3C) LD. The rate of decline in total LD with distance is determined by the relative contributions of within- and between-breed LD because within-breed LD declines more rapidly with distance as it is eroded over more generations.

In case I, only between-breed LD contributed to the total LD in the F_2 because ancestral breeds A and B were fixed for the QTL and CGL (Table 1) and, therefore, no LD was generated within the F_0 sire breed (Fig. 3B). Similarly, there was no LD within the F_0 dam breed. However, the between-breed LD was maximized from the cross in the F_0, since F_0 sires and dams were fixed for alternate alleles. The between-breed LD extended over long distances (Fig. 3C). The expected standardized LD (D′) was 0.96 between the QTL and a CGL at 2 cM from the QTL and declined gradually to 0.82 and 0.55 for CGL at 10 and 30 cM (Fig. 3A).

From case II to case IV, gene frequency differences between ancestral breeds A and B and between breeds C and D increased (Table 1). Accordingly, for a given CGL position, within-breed LD in the F_0 breeds increased and was maximal in case IV (Fig. 3B). However, between-breed LD in the F_2 decreased from case II to case IV because the difference in gene frequencies between the F_0 sires and dams decreased. Between-breed LD was minimal for case IV (Fig. 3C).

In case IV, only within-breed LD contributed to total LD in the F_2. Lack of between-breed LD would prevent detection of the QTL using breed-cross interval mapping. The nature of LD in case IV corresponds to what can be expected in closed breeding populations with LD extending over only short distances because LD was eroded for 20 generations since its origin (Fig. 3A); the expected D′ with the QTL was 0.65 for a CGL at 2 cM from the QTL and decayed rapidly to 0.12 and 0.003 for CGL at 10 and 30 cM (Fig. 3A).

Comparison of LD for a CGL that is at the QTL position for the four cases demonstrates the antagonism between within- and between-breed LD. When between-breed LD is high, within-breed LD tends to be low, and vice versa (Fig. 3). Because breed-cross

interval mapping utilizes between-breed LD, between-breed LD within QTL regions is, in general, extensive, which leaves less room for within-breed LD.

Standard association test. Statistical power of the standard association test of CGL is shown in Fig. 4A. Trends in power were in good agreement with the plots of expected total LD (Fig. 3A), indicating that the standard association test relies on total LD, combining between- and within-breed LD.

Case I had very high between-breed LD and no within-breed LD (Fig. 3B, C) and the test showed significant associations with phenotype, even for distant CGL (Fig. 4A). Statistical power was nearly 90% for a CGL that was as much as 30 cM from the QTL. Case II represented high between-breed LD and low within-breed LD (Fig. 3B, C). Power was greater than 85% for CGL within 2 cM from the QTL, but dropped to 53% for a CGL that was 30 cM from the QTL (Fig. 4A). In case III, both between- and within-breed LD were moderately high (Fig. 3B, C). Power was greater than 57% for CGL within 2 cM from the QTL but dropped to 18% for a CGL at 30 cM from the QTL (Fig. 4A). In case IV, only within-breed LD contributed to the total LD (Fig. 3B, C). Power was greater than 84% for CGL within 2 cM from the QTL, and dropped rapidly to 24 and 13% for CGL at 10 and 30 cM from the QTL (Fig. 4A). In all four cases, power was 100% when the CGL polymorphism was the causative gene.

Marker-assisted association test. Statistical power of the marker-assisted association test at CGL is shown in Fig. 4B. Inclusion of QTL effects associated with markers is expected to take out all or part of the between-breed LD that exists at the F_2 level. The resulting test of

the CGL effect is, therefore, expected to rely largely on within-breed LD. Trends in power (Fig. 4B) indeed matched trends in expected within-breed LD (Fig. 3B), except for cases I and II, which had no and low within-breed LD, respectively.

For case I, statistical power of the marker-assisted association test was substantially greater than 5% (the type I error rate) for CGL within 10 cM from the QTL (Fig. 4B), although the expected within-breed LD was zero (Fig. 3B). The actual within-breed LD was also zero in each replicate because the F_0 sires and dams were fixed for alternate alleles at the QTL and, thus, there was no within-breed variation. The observed power suggests that QTL effects associated with markers removed part, but not all, of the between-breed LD and that some between-breed LD was allocated to the CGL.

Within-breed LD was very low for case II (Fig. 3B). Statistical power to detect a significant association was 85% when the CGL was the QTL, but less than 33% for CGL within 2 cM from the QTL, and dropped to 9% for a CGL at 10 cM (Fig. 4B). Within-breed LD was moderately high for case III (Fig. 3B) and power was 99% when the CGL polymorphism was the causative gene, but was less than 54% for CGL within 2 cM from the QTL. Power dropped to 15% for a CGL at 10 cM (Fig. 4B). In case IV, total LD was composed of within-breed LD only (Fig. 3A, B) and including markers had limited effect because between-breed LD is expected to be zero (Fig. 3C), although some may exist by chance. The marker-assisted association test gave similar power as the standard association test for this case (Fig. 4A, B).

F-drop test. In this test, F ratios for a breed-cross QTL effect at the CGL position with and without the CGL included as fixed effect were compared. Its power is shown in Fig. 4C.

With high between-breed LD, a relatively high QTL F ratio is obtained from the model used for breed-cross QTL analysis (Haley et al. 1994). If a CGL that is in LD with the QTL is included as a fixed effect, it is expected to absorb part of the between-breed QTL effect, and the F ratio for the QTL is expected to drop. Observed trends in power of this test (Fig. 4C) were in agreement with the trends for expected between-breed LD (Fig. 3C), except for case IV, indicating that the F-drop test relies on between-breed LD.

In case I, for which between-breed LD is very high, the power to detect a significant association for the CGL with the F-drop test was greater than 98% when the CGL polymorphism was the causative mutation or was within 2 cM from the QTL (Fig. 4C). However, power remained as high as 95 and 49% for CGL at 10 and 30 cM from the QTL, in accord with the slow decline of between-breed LD (Fig. 3C). As between-breed LD decreased in cases II and III (Fig. 3C), power to identify associations for a CGL that is the causative mutation dropped to 83% and 51%, respectively (Fig. 4C). Power for CGL within 2 cM from the QTL was around 70% for case II and 30% for case III and decreased only gradually for CGL further from the QTL. Case IV showed power greater than 5% for CGL within 10 cM from the QTL (Fig. 4C), although between-breed LD is expected to be zero (Fig. 3C). The greater than 5% power for this case is explained by the chance existence of between-breed LD in individual replicates.

Although inclusion of a CGL as a fixed effect in a QTL mapping analysis is expected to reduce the F ratio for a QTL if the CGL is in LD with the QTL, lack of a drop in F ratio does not mean that the CGL is not linked to the QTL. The probability that the F ratio for a QTL remained unchanged or even increased when the CGL was included as fixed effect tended to increase with decreasing between-breed LD (Fig. 5). Probabilities were close to

50% in the absence of between-breed LD (case IV), at least with CGL at some distance from the QTL. This is as expected because inclusion of an uncorrelated fixed effect is expected to result in random changes in the F ratio, i.e. equal probabilities for an increase and a decrease.

Comparison of two candidate gene loci. A standard and a marker-assisted association test were used to distinguish between two CGL at 1 and 10 cM from the QTL. The CGL at 10 cM from the QTL was also compared with the QTL itself. Statistical power for these tests is shown in Table 2. Only cases I, III and IV were examined, representing zero, moderate and maximal within-breed LD, respectively.

In the comparison of two CGL, the standard association test relies not only on the magnitude of total LD, but also on the decline in LD with distance from the QTL, i.e., how steep the decline in total LD is. Similarly, the marker-assisted association test relies on how steep the decline in within-breed LD is.

In case I, the total LD was composed of between-breed LD only, and dropped from 0.98 to 0.82 when the distance between the QTL and the CGL increased from 1 to 10 cM (Fig. 3A). The total LD of the QTL with itself is one. Power to correctly identify the closest CGL was 34.3% based on the standard association test and 25% for the marker-assisted association test (Table 2). Power increased to 42.4% and 32.1% for the standard and the marker-assisted association test when one CGL polymorphism was the causative mutation (Table 2). The power of the marker-assisted association test was greater than 5% because, as described previously, including QTL effects associated with markers takes out part, but not all, of the between-breed LD.

In case III, power to identify the closest CGL was low for both tests but slightly larger for the marker-assisted association test (Table 2). Within-breed LD was present at a moderate level and had a steeper decline with distance than total LD (Fig. 3A, B). As the distance between the QTL and CGL increased from 1 to 10 cM, total LD dropped by 60% (from 0.45 to 0.18), while within-breed LD dropped by 86% (from 0.29 to 0.04). The marker-assisted association test removes between-breed LD, which resulted in the slightly greater power than the standard association test (11.5% vs 10.8%).

When one CGL polymorphism was the causative mutation, power was high for both tests in case III (Table 2). Comparing the LD within the QTL and the LD between the QTL and the CGL (Fig. 3A, B), total LD dropped by 82% (from 1 to 0.18) and within-breed LD dropped by 95% (from 0.84 to 0.04). Due to the steeper decline in within-breed LD, the marker-assisted association test showed a power of 83.3%, slightly higher than the power of 82.9% for the standard association test (Table 2).

In case IV, expected total LD was composed of within-breed LD only. The LD was one for the QTL with itself, and dropped from 0.8 to 0.12 as distance between the QTL and CGL increased from 1 to 10 cM (Fig. 3A, B). Power of the marker-assisted association test to identify the closest CGL was 43.4% and slightly higher than power of the standard association test (42.1%) (Table 2). This is probably due to the between-breed LD that existed in the simulated data by chance. Power increased to 77.5% and 76.6% for the standard and the marker-assisted association test (Table 2) when one of the CGL polymorphisms was the causative mutation.

DISCUSSION AND CONCLUSIONS

Studies to map QTL in livestock populations are expensive because of the cost of animal rearing and phenotyping. Most QTL analyses in poultry and swine, and some in cattle, have utilized crosses (F_2 or backcrosses) between outbred breeds as resource populations because of their statistical power to detect QTL. These populations have been well characterized, both from a phenotypic and a genetic perspective and, in follow-up to a QTL genome scan, have been used for further genetic analysis, including that of positional CGL. Analysis of positional CGL in a QTL mapping population is, however, complicated by the extensive between-breed LD that is created in the cross. In this paper, we developed and evaluated tests for CGL in F_2 resource populations. Four cases (Table 1) were simulated to reflect the degrees of within- and between-breed LD that may exist in QTL mapping populations.

Our results show that the standard association test detects associations for CGL that may be far removed from the QTL, except when no between-breed LD exists between the QTL and the CGL (case IV). The standard association test has limited power to distinguish distant from closely linked CGL. Results for the newly developed marker-assisted association test demonstrate that this test can exclude the confounding effect of the extensive between-breed LD that exists in F_2 populations. However, this does not improve the power to detect CGL that are closely linked to the QTL, unless the CGL is the QTL. In QTL mapping populations, between-breed LD is usually extensive in order to detect a QTL. This leaves limited opportunity for within-breed LD in the identified QTL regions, which explains the limited power of the marker-assisted association test.

The F-drop test is attractive as a test for CGL following QTL analysis and is easily implemented in the breed-cross interval mapping program developed by Haley et al. (1994).

For example, Nguyen et al. (2003) studied the HMGA1 gene in QTL regions for fatness traits on pig chromosome 7 in a Berkshire x Yorkshire F_2 population. A significant reduction in the F ratio for the QTL was observed when HMGA1 genotype was included as a fixed effect in the QTL analysis. This drop in the F statistic suggests direct involvement of HMGA1 or close linkage to the causative mutation (Nguyen et al. 2003). Our results, however, show that results from this test must be interpreted with caution because it will show associations not only for close but also for distant CGL.

In most cases, multiple CGL are present in an identified QTL region and there is a need to identify the most promising CGL for further analyses. Our results demonstrate that power to distinguish two CGL that are within 10 cM from the QTL is limited and little improved by including QTL effects associated with markers to remove between-breed LD. However, the power to distinguish two CGL was much higher (>80%) when one of the CGL was the causative mutation and the amount of within-breed LD was moderate (case III). A significant result for this test does, however, not suggest that one of the CGL is the causative mutation, because a significant test can also result from linked CGL.

Given the extensive between-breed LD that exists in intercrosses of outbred breeds or lines, CGL studies in farm animals cannot rely solely on breed-cross populations and effects uncovered in these crosses must be confirmed in one or more closed mating populations. As an example of such a study, Ciobanu et al. (2001) investigated the PRKAG3 gene as a positional candidate for a QTL for muscle glycogen content and related meat quality traits that was identified by Malek et al. (2001b) on chromosome 15 in a Berkshire x Yorkshire F_2 population of pigs. Initial analysis in the F_2 population using an association study showed significant effects of PRKAG3 mutations on glycogen content and meat quality traits. The

effects of PRKAG3 mutations were further confirmed by association tests in five unrelated commercial pig lines (Ciobanu et al. 2001).

Case IV is virtually identical to a closed breeding population, since only within-breed LD contributed to LD in the F_2. Normalized LD values between the QTL and CGL ranged from the maximum of 1 to 0.34 for CGL at 0 to 5 cM from the QTL, and declined rapidly to 0.12 and 0.003 for CGL at 10 and 30 cM, respectively (Fig. 3B). Farnir et al. (2000) measured genome-wide LD in the Dutch Holstein dairy cattle population and found average normalized LD values of 0.5 for syntenic marker pairs that were less than 5 cM apart. The LD dropped to 0.16 for distances of 50 cM (Farnir et al. 2000). A similar study was performed by McRae et al. (2002) in populations of Coopworth and Romney sheep, where high levels of LD were found to extend for nearly 60 cM. Within-breed LD in case IV in our study, therefore, extended over shorter distances than observed by Farnir et al. (2000) and McRae et al. (2002).

For case IV, the standard association test was robust to detect CGL that are closely linked to the QTL. The power of this test was greater than 84% for CGL within 2 cM from the QTL, and dropped rapidly to 24 and 13% for CGL at 10 and 30 cM from the QTL (Fig. 4A). However, even with only within-breed LD, the power was less than 50% to distinguish a CGL that was 1 cM removed from the QTL from a CGL at 10 cM. Power was higher than 75% if one of the CGL polymorphisms was the causative mutation (Table 2). This suggests that limited power is available for fine mapping, even in outbred populations, unless large sample sizes are available.

Acknowledgments. This work was supported by CSREES IFAFS # 00-52100-9610. The authors thank Dan Nettleton for statistical advice and Radu Totir, Hauke Thomsen, Massoud Malek, Petek Settar, David Casey, Jing Wang, Joseph McElroy and Napapan Chaiwong for helpful discussion and suggestions.

REFERENCES

Andersson L, Haley CS, Ellegren H, Knott SA, Johansson M et al. (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science 263, 1771-1774

Ciobanu D, Bastiaansen J, Malek M, Helm J, Woollard J, Plastow G, Rothschild MF (2001) Evidence for new alleles in the protein kinase adenosine monophosphate-activated γ_3-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality. Genetics 159, 1151-1162

De Koning DJ, Janss LLG, Rattink AP, van Oers PAM, de Vries BJ et al. (1999) Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs *(Sus Scrofa).* Genetics 152, 1679-1690

Farnir F, Coppieters W, Arranz J-J, Berzi P, Cambisano N et al. (2000) Extensive genome-wide linkage disequilibrium in cattle. Genome research 10, 220-227

Haley CS, Knott SA, Elsen JM (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136, 1195-1207

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49, 49-67

Li H, Deeb N, Zhou H, Mitchell AD, Ashwell CM, Lamont SJ (2003) Chicken quantitative trait loci for growth and body composition associated with transforming growth factor-β genes. Poultry Science 82, 347-356

Lo LL, Fernando RL, Grossman M (1993) Covariance between relatives in multibreed populations: additive model. Theor Appl Genet 87, 423-430

Malek M, Dekkers JCM, Lee HK, Baas TJ, Rothschild MF (2001a) A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. I. Growth and body composition. Mamm Genome 12, 630-636

Malek M, Dekkers JCM, Lee HK, Baas TJ, Prusa K, Huff-Lonergan E, Rothschild MF (2001b) A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. II. Meat and muscle composition. Mamm Genome 12, 637-645

McRae AF, McEwan JC, Dodds KG, Wilson T, Crawford AM, Slate J (2002) Linkage disequilibrium in domestic sheep. Genetics 160, 1113-1122

Nguyen NT, Kim KS, Thomsen H, Helm J, Rothschild MF (2003) Investigation of a candidate gene for growth and fatness QTL on the pig chromosome 7. Proceedings of Plant and Animal Genome XI Conference, San Diego, CA, p230

Riquet J, Coppieters W, Cambisano N, Arranz JJ, Berzi P et al. (1999) Fine-mapping of quantitative trait loci by identity by descent in outbred populations: Application to milk production in dairy cattle. Proc Natl Acad Sci USA 96, 9252-9257

Rohrer GA, Keele JW (1998a) Identification of quantitative trait loci affecting carcass composition in swine: I. Fat deposition traits. J Anim Sci 76, 2247-2254

Rohrer GA, Keele JW (1998b) Identification of quantitative trait loci affecting carcass composition in swine: II. Muscling and wholesale product yield traits. J Anim Sci 76, 2255-2262

Rothschild MF, Soller M (1997) Candidate gene analysis to detect genes controlling traits of economic importance in domestic livestock. Probe 8, 13-20

Rothschild MF, Jacobson C, Vaske D, Tuggle CK, Wang L et al. (1996) The estrogen receptor locus is associated with a major gene influencing litter size in pigs. Proc Natl Acad Sci USA 93, 201-205

Short TH, Rothschild MF, Southwood OI, McLaren DG, de Vries A et al. (1997) Effect of the estrogen receptor locus on reproduction and production traits in four commercial pig lines. J Anim Sci 75, 3138-3142

Yu TP, Tuggle CK, Schmitz CB, Rothschild MF (1995) Association of PIT1 polymorphisms with growth and carcass traits in pigs. J Anim Sci 73, 1282-1288

Yu TP, Wang L, Tuggle CK, Rothschild MF (1999) Mapping genes for fatness and growth on pig chromosome 13: a search in the region close to the pig PIT1 gene. J Anim Breed Genet 116, 269-280

Zhou H, Buitenhuis AJ, Weigend S, Lamont SJ (2001) Candidate gene promoter polymorphisms and antibody response kinetics in chickens: interferon-γ, interleukin-2, and immunoglobulin light chain. Poultry Science 80, 1679-1689

APPENDIX

The between- and within-breed LD in the simulated F_2 populations is derived. Following Lo et al. (1993), the expected between-breed LD between the QTL and a candidate gene locus

(CGL) in the F_2 is $\frac{1}{4}(1-2r)(P_Q^{Sire}-P_Q^{Dam})(P_C^{Sire}-P_C^{Dam})$, where r is the recombination rate between the QTL and the CGL, and P_k^{Sire} and P_k^{Dam} are the frequencies of allele k in the F_0 sires and dams, respectively. Because alleles Q and C were simulated with the same frequencies in the F_{-20}, their frequencies in the F_0 are expected to be the same, that is, $P_Q^{Sire}=P_C^{Sire}=\frac{1}{2}(P_Q^A+P_Q^B)$ and $P_Q^{Dam}=P_C^{Dam}=\frac{1}{2}(P_Q^C+P_Q^D)$, where P_k^i is the frequency of allele k in ancestral breed *i* in the F_{-20}. Therefore, between-breed LD in the F_2 was calculated as:

$$LD_{between}=\frac{1}{4}(1-2r)[\frac{1}{2}(P_Q^A+P_Q^B)-\frac{1}{2}(P_Q^C+P_Q^D)]^2 \qquad [1]$$

The LD that existed in the F_{-18} (Fig. 2) from the cross A x B in the F_{-20} is $\frac{1}{4}(1-2r)(P_Q^A-P_Q^B)(P_C^A-P_C^B)$ (Lo et al. 1993), which equals $\frac{1}{4}(1-2r)(P_Q^A-P_Q^B)^2$ since frequencies of alleles Q and C were equal in the F_{-20}. This LD decayed each generation by a factor $(1-r)$. Therefore, the LD that existed in the F_0 sires is $(1-r)^{18}[\frac{1}{4}(1-2r)(P_Q^A-P_Q^B)^2]$. Similarly, the LD that existed in the F_0 dams is $(1-r)^{18}[\frac{1}{4}(1-2r)(P_Q^C-P_Q^D)^2]$, which equals $(1-r)^{18}[\frac{1}{4}(1-2r)(P_Q^A-P_Q^B)^2]$ because $P_Q^A-P_Q^B = P_Q^D-P_Q^C$ (Table 1). From the F_0 to the F_2, this LD was further reduced by $(1-r)^2$, therefore, the expected within-breed LD between the QTL and a CGL in the F_2 is:

$$LD_{within}=(1-r)^{20}[\frac{1}{4}(1-2r)(P_Q^A-P_Q^B)^2] \qquad [2]$$

Setting $r=0$ in Eq [1], the expected between-breed LD within the QTL in the F_2 is:

$$LD_{between} = \frac{1}{4}[\frac{1}{2}(P_Q^A + P_Q^B) - \frac{1}{2}(P_Q^C + P_Q^D)]^2 \quad [3]$$

The within-breed LD for the QTL with itself in the F_2 originated not only from the cross of ancestral breeds in the F_{-20}, but also from the LD that existed in ancestral breed *i*, which is equal to $P_Q^i(1-P_Q^i)$. Setting $r=0$ in Eq [2] and adding the average LD within breeds A and B in the $F_{-20,}$ the expected within-breed LD for the QTL in the F_2 is:

$$LD_{within} = \frac{1}{2}[P_Q^A(1-P_Q^A) + P_Q^B(1-P_Q^B)] + \frac{1}{4}(P_Q^A - P_Q^B)^2 \quad [4]$$

Table 1. Frequencies of alleles Q and C of the QTL and candidate gene loci in the four ancestral breeds in the F_{-20} of the simulated pedigree for four alternative cases. Two alleles were simulated for the QTL (Q and q) and for each candidate gene locus (C and c). Alleles Q and C were simulated with equal frequencies in the F_{-20} in all four cases.

	Ancestral breed				Extent of linkage disequilibrium	
Case	A	B	C	D	Between-breed	Within-breed
I	1	1	0	0	high	0
II	1	0.8	0	0.2	medium	low
III	1	0.4	0	0.6	low	medium
IV	1	0	0	1	0	high

Table 2. Power to distinguish a candidate gene, CGL_1, at 10 cM from the QTL from a candidate gene, CGL_2, at 1 cM from the QTL or from the causative mutation for the QTL, based on standard and marker-assisted association tests for three cases of combinations of between- and within-breed linkage disequilibrium (see Table 1). Power (%) to determine that CGL_2 explains more variance than CGL_1 is based on 1,000 replicates at the 5% significance level for a two-sided test.

	CGL_1 vs. CGL_2			CGL_1 vs. QTL		
Statistical test	Case I	Case III	Case IV	Case I	Case III	Case IV
Standard association test	34.3	10.8	42.1	42.4	82.9	77.5
Marker-assisted association test	25.0	11.5	43.4	32.1	83.3	76.6

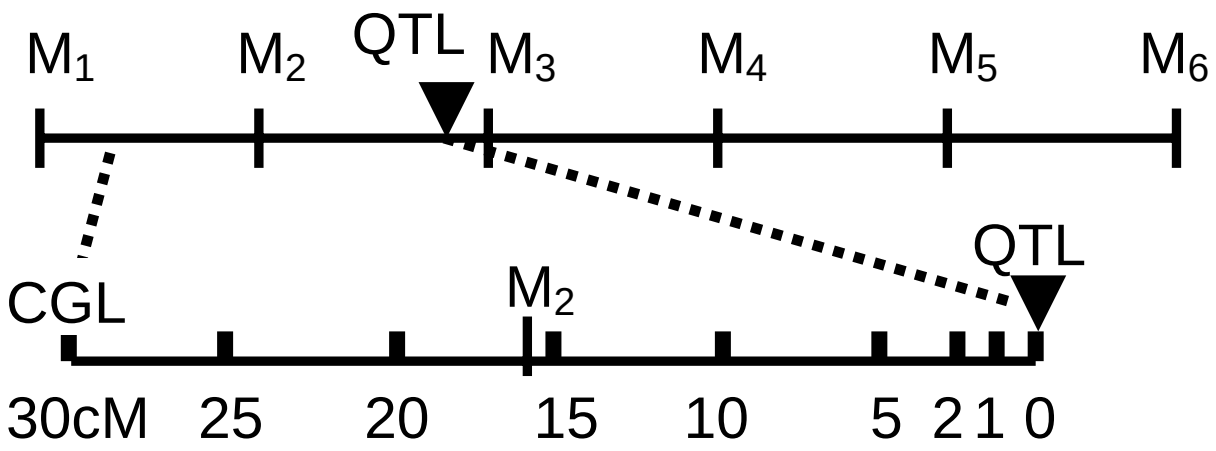


Fig. 1. Linkage map used for simulation of a chromosome of 100 cM with 6 markers (M_1 to M_6) 20 cM apart, an additive QTL at 36 cM from M_1 with a substitution effect of 0.25 phenotypic standard deviations, and 9 candidate gene loci (CGL) at 0, 1, 2, 5, 10, 15, 20, 25 and 30 cM from the QTL.

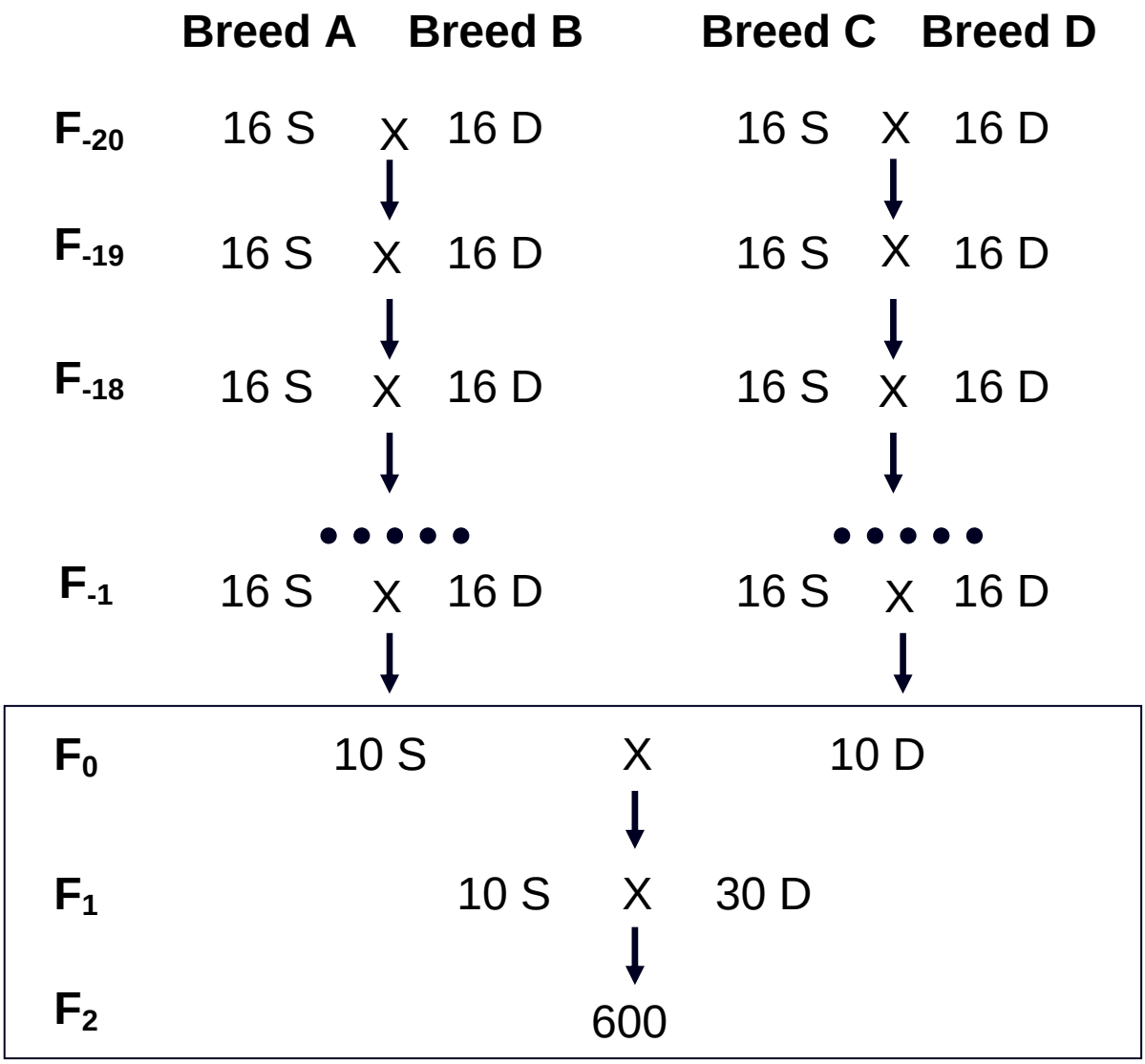


Fig. 2. Pedigree used for simulation of the F_2 population, starting with crosses between two pairs of ancestral breeds, A x B and C x D, followed by 20 generations of random selection and mating of 16 sires (S) and 16 dams (D) within each cross to generate within-breed linkage disequilibrium within each F_0 breed. Data from generations F_0, F_1 and F_2 were used for analysis.

Fig. 3A

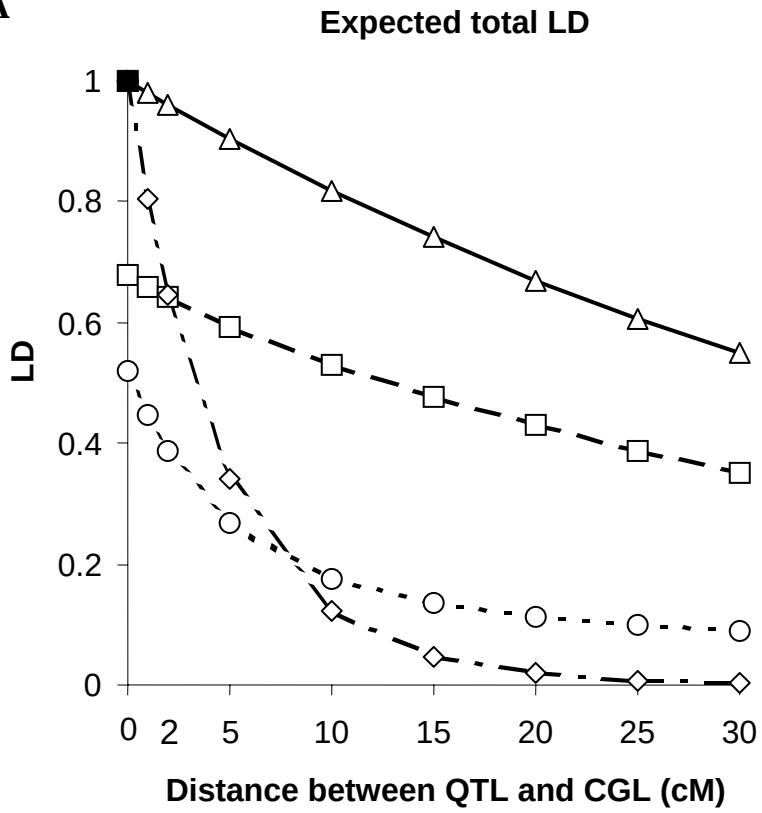

Expected total LD
LD
1
0.8
0.6
0.4
0.2
0
0 2 5 10 15 20 25 30
Distance between QTL and CGL (cM)

Fig. 3B

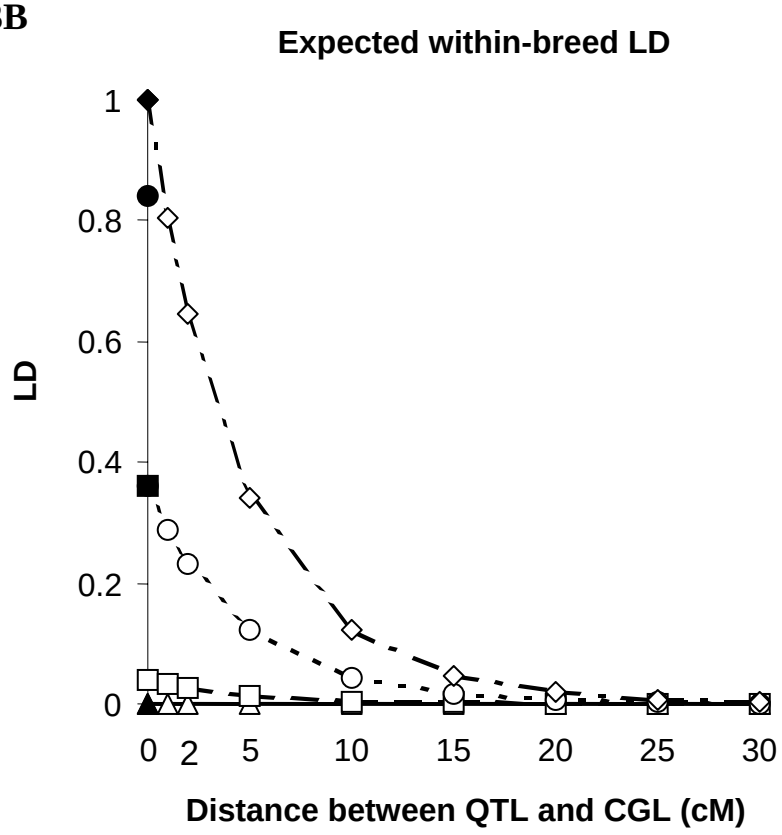

Expected within-breed LD
LD
1
0.8
0.6
0.4
0.2
0
0 2 5 10 15 20 25 30
Distance between QTL and CGL (cM)

Fig. 3C

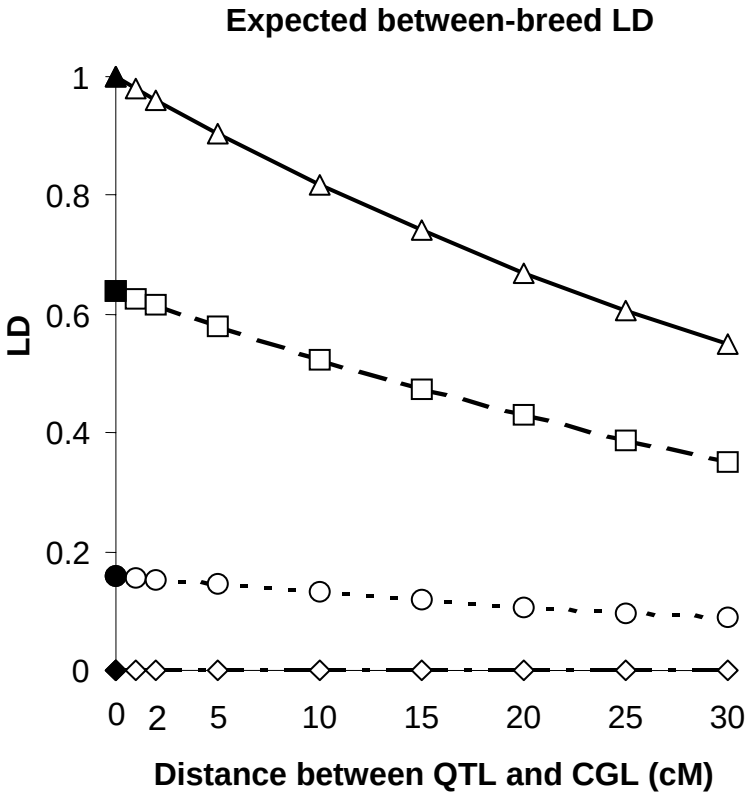


Fig. 3. The expected total (A), within-breed (B), and between-breed (C) linkage disequilibrium (LD) in the F_2 between the QTL and candidate gene loci (CGL) at alternate positions for four different cases (see Table 1): case I (—△—), case II (- -□- -), case III (····○····) and case IV (– ·◇· -). Open symbols indicate the expected LD between the QTL and a CGL. Closed symbols indicate the expected LD of the QTL with itself.

Fig. 4A

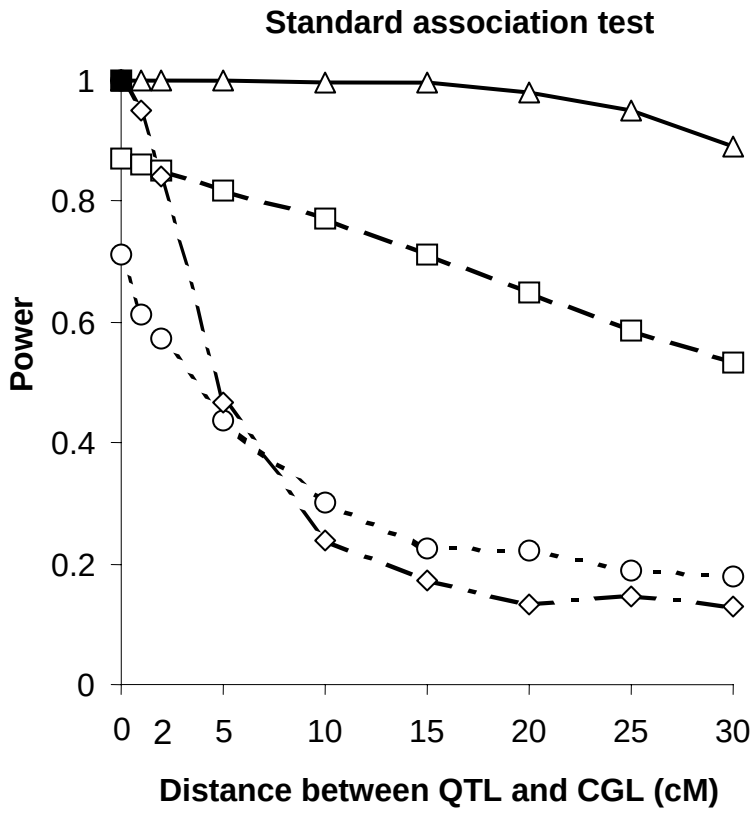

Standard association test
Power
1
0.8
0.6
0.4
0.2
0
0 2 5 10 15 20 25 30
Distance between QTL and CGL (cM)

Fig. 4B

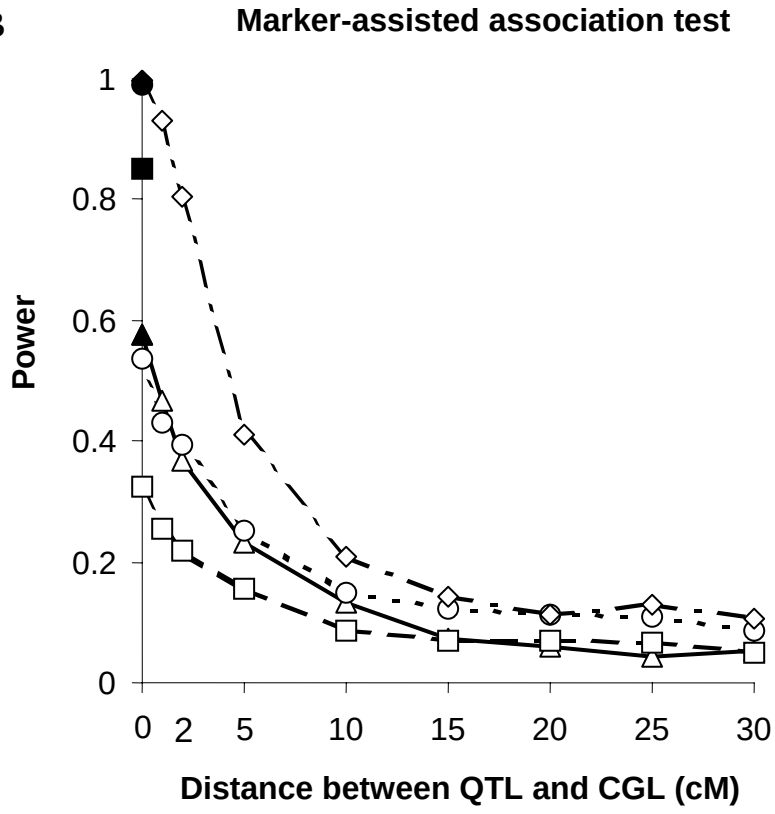

Marker-assisted association test
Power
1
0.8
0.6
0.4
0.2
0
0 2 5 10 15 20 25 30
Distance between QTL and CGL (cM)

Fig. 4C

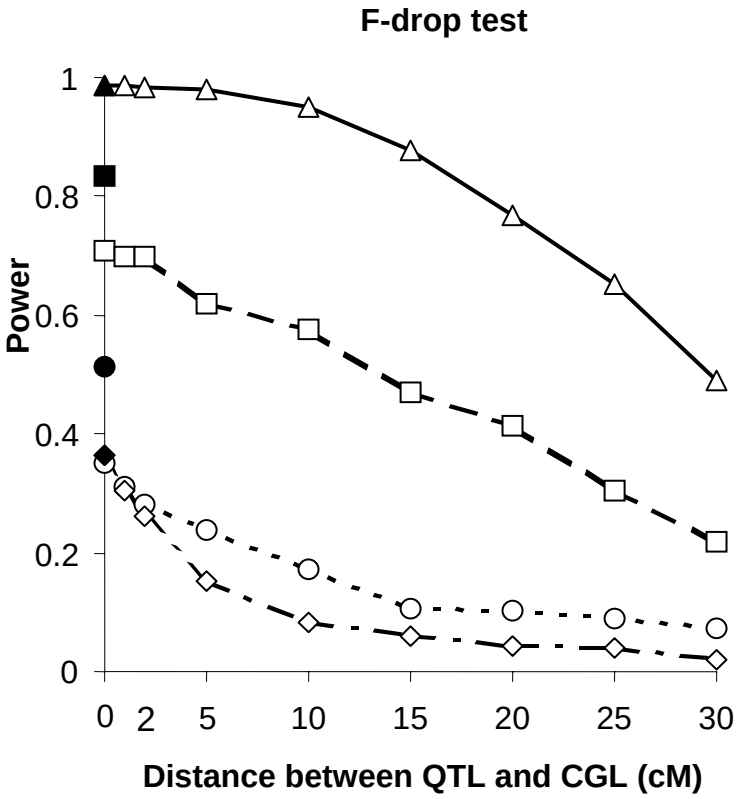


Fig. 4. Power of three statistical tests to identify candidate gene loci (CGL) that are associated with the QTL for four cases (see Table 1): case I (—△—), case II (- -□- -), case III (····○····) and case IV (- -◇· -). (A) standard association test, (B) marker-assisted association test, (C) F-drop test. Open symbols represent CGL linked to the QTL. Closed symbols refer to a CGL polymorphism that is the causative gene for the QTL. Power is evaluated at the 5% significance level and is based on 1,000 replicates.

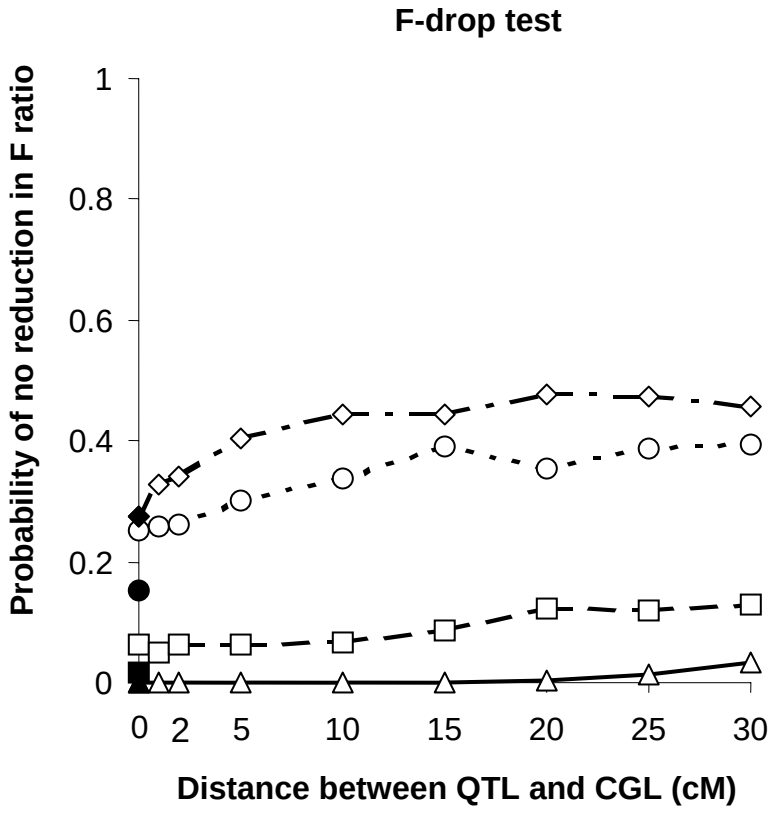


Fig. 5. The probability that the breed-cross interval mapping F ratio for a QTL at the candidate gene locus (CGL) position remained unchanged or increased when including the CGL as fixed effect in the F-drop test for four different cases (see Table 1): case I (—△—), case II (– -□- –), case III (·····○·····) and case IV (– ·◇· -). Open symbols refer to tests for CGL linked to the QTL. Closed symbols refer to tests for the QTL itself. Probabilities are based on 1,000 replicates.

CHAPTER 3. EVALUATION OF LINKAGE DISEQUILIBRIUM MEASURES BETWEEN MULTI-ALLELIC MARKERS AS PREDICTORS OF LINKAGE DISEQUILIBRIUM BETWEEN MARKERS AND QTL

A paper published in *Genetical Research*[*]

H. Zhao[1], D. Nettleton[2], M. Soller[3] and J. C. M. Dekkers[1]

[1] Department of Animal Science and Center for Integrated Animal Genomics, 239 Kildee Hall, Iowa State University, Ames, Iowa, 50011, USA

[2] Department of Statistics, 124 Snedecor Hall, Iowa State University, Ames, Iowa, 50011, USA

[3] Hebrew University of Jerusalem, Israel

Summary

Effectiveness of marker-assisted selection (MAS) and quantitative trait loci (QTL) mapping using population-wide linkage disequilibrium (LD) between markers and QTL depends on the extent of LD and how it declines with distance in a population. Because marker-QTL LD cannot be observed directly, the objective of this study was to evaluate alternative measures of observable LD between multi-allelic markers as predictors of usable LD of multi-allelic markers with presumed biallelic QTL. Observable LD between marker pairs was evaluated using eight existing and one new measure. These included two pooled and standardized measures of LD between pairs of alleles at two markers based on Lewontin's LD measure,

[*] Reprinted with permission of *Genitical Research* (2005) 86: 77-78.

two pooled measures of squared correlations between alleles, one standardized measure using Hardy-Weinberg heterozygosities, and four measures based on the chi-square statistic for testing for association between alleles at two loci. In simulated populations with a range of LD generated by drift and a range of marker polymorphism, marker-marker LD measured by a standardized chi-square statistic (denoted $\chi^{2}{}'$) was found to be the best predictor of usable marker-QTL LD for a group of multi-allelic markers. Estimates of the level and decline of marker-marker LD with distance obtained from $\chi^{2}{}'$ were linearly and highly correlated with usable LD of those markers with QTL across population structures and marker polymorphism. Corresponding relationships were poorer for the other marker-marker LD measures. Therefore, when LD is generated by drift, $\chi^{2}{}'$ is recommended to quantify the amount and extent of usable LD in a population for QTL mapping and MAS based on multi-allelic markers.

1. Introduction

Linkage disequilibrium (LD) is the condition in which alleles at two loci are not independent. The extent of LD is a topic of great interest in both humans and livestock. Effectiveness of marker-assisted selection (MAS) and fine mapping of quantitative trait loci (QTL) using population-wide LD between markers and QTL depends on the extent of LD and how it declines with distance (Lande & Thompson, 1990; Terwilliger & Weiss, 1998; Dekkers & Hospital, 2002). Although population-wide LD can be created by crossing lines or breeds, here we focus on LD within outbreeding populations. Because QTL cannot be observed directly, LD between markers can be used to predict marker-QTL LD, in order to evaluate

the extent of useful LD in a population (e.g. Pritchard & Przeworski, 2001; Farnir *et al.*, 2000).

The two most common LD measures for biallelic markers are D' and r^2 (Lewontin, 1964; Hill & Robertson, 1968; Ardlie *et al.*, 2002), although other measures have been used (Devlin & Risch, 1995; Morton *et al*., 2001). Current research prefers the square of the correlation coefficient between markers, r^2, to detect markers that might correlate with the QTL of interest, because r^2 quantifies the amount of information about one locus provided by the other (Ardlie *et al.*, 2002; Flint-Garcia *et al.*, 2003), although other optimal measures have been proposed (Devlin & Risch, 1995; Morton *et al*., 2001). For biallelic markers, the absolute value of LD is the same between any pair of alleles across two loci. However, this is not true when one or both markers have more than two alleles, as is the case for the still frequently used microsatellite markers. This makes assessing the degree of LD between multi-allelic markers more complicated.

A variety of statistics have been proposed to measure LD between multi-allelic markers (Yamazaki, 1977; Hedrick & Thomson, 1986; Hedrick, 1987; Sabatti & Risch, 2002). Hedrick's (1987) multi-allelic extension of Lewontin's (1964) normalized LD measure, D', is commonly used. Using D', extensive LD over a long range was observed in dairy cattle, sheep and pigs (Farnir *et al*., 2000; Tenesa *et al*., 2003; McRae *et al*., 2002; Nsengimana *et al.,* 2004). However, it is known that LD measured by D' tends to be inflated with small sample sizes and/or low allele frequencies (Ardlie *et al.*, 2002; Flint-Garcia *et al.*, 2003; McRae *et al*., 2002). A generally satisfactory measure of LD between multi-allelic markers has not been agreed upon, nor have alternate measures of LD among multi-allelic markers been compared for their ability to predict the extent of usable LD for QTL mapping or MAS

(see, however, Devlin and Risch (1995), where disease and marker loci were both assumed to have two alleles).

Random drift plays an important role in generating LD in livestock breeding populations, which are typically of limited size (Flint-Garcia *et al.*, 2003). The objective of this study was, therefore, to evaluate, by simulation, alternative measures of LD between multi-allelic markers as predictors of usable LD of multi-allelic markers with QTL and, more generally, as predictors of LD of multi-allelic markers with biallelic single nucleotide polymorphisms (SNPs), when LD is generated by drift. The ability to use LD between multi-allelic markers to predict LD among SNPs or usable LD of SNPs with QTL will be addressed in a subsequent paper.

2. Materials and methods

(i) *Measures of marker-marker LD*

The standard measure of LD between two alleles at two different loci is

$$D_{ij} = p(A_iB_j) - p(A_i)p(B_j),$$

where $p(A_i)$ is the frequency of allele A_i at locus A, $p(B_j)$ the frequency of allele B_j at locus B, and $p(A_iB_j)$ the frequency of haplotype A_iB_j. For loci with two alleles, D_{ij} completely describes LD between all pairs of alleles. Because D_{ij} depends on gene frequencies, Lewontin (1964) suggested standardizing D_{ij} by the maximum absolute value it can attain, given the allele frequencies: $D'_{ij} = \frac{D_{ij}}{D_{ij}^{\max}}$,

where $D_{ij}^{\max} = \min[\, p(A_i)\,p(B_j), (1-p(A_i))(1-p(B_j))\,]$ when $D_{ij} < 0$,

$$D_{ij}^{\max} = \min\left[\, p(A_i)(1-p(B_j)), (1-p(A_i))\, p(B_j)\,\right] \quad \text{when} \quad D_{ij} \geq 0\,.$$

Hill and Robertson (1968) suggested using the square of the correlation between A_i and B_j, denoted by r_{ij}^2, as a standardized measure of LD between biallelic loci. This measure can be computed from D_{ij} and allele frequencies as follows:

$$r_{ij}^2 = \frac{D_{ij}^2}{p(A_i)\,(1-p(A_i))\;p(B_j)\,(1-p(B_j))}\,.$$

Measures $\left|D'_{ij}\right|$ and r_{ij}^2 range from 0 to 1 but $\left|D'_{ij}\right|$ is strongly inflated if some haplotypes are not observed, which can occur for haplotypes of low frequency alleles in small samples (Flint-Garcia *et al.*, 2003). Compared to $\left|D'_{ij}\right|$, r_{ij}^2 is less inflated in small samples (Ardlie *et al.*, 2002) and quantifies the information one locus provides about the other. Current research appears to prefer r_{ij}^2 for finding biallelic markers that might correlate with QTL of interest (Ardlie *et al.*, 2002; Flint-Garcia *et al.*, 2003), although there are other viewpoints (Devlin & Risch, 1995; Morton *et al*., 2001) .

As noted above, when markers have more than two alleles, LD can differ between pairs of alleles and a combined measure of LD across alleles is needed. Several such measures have been proposed (Yamazaki, 1977; Hedrick & Thomson, 1986; Hedrick, 1987; Sabatti & Risch, 2002). In this study, we compared eight existing measures and one new measure of LD between multi-allelic markers. The first two measures are based on pooling and standardizing D_{ij} across loci based on allele frequencies, following Hedrick (1987):

$$D' = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_i)\, p(B_j) \left| \frac{D_{ij}}{D_{ij}^{\max}} \right|, \tag{1}$$

or based on haplotype frequencies, following Karlin and Piazza (1981):

$$D_{hap} = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_i B_j) \left| \frac{D_{ij}}{D_{ij}^{\max}} \right|, \tag{2}$$

where k and m are the numbers of alternate alleles at locus A and B, respectively.

The next two measures are based on pooling r_{ij}^2 based on allele frequencies:

$$r^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_i)\, p(B_j)\, r_{ij}^2 , \tag{3}$$

or based on haplotype frequencies:

$$r_{hap}^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_i B_j)\, r_{ij}^2 . \tag{4}$$

Using Hardy-Weinberg heterozygosities at two loci, the fifth measure is

$$D^* = \frac{D^2}{H_A H_B} \tag{5}$$

(Maruyama, 1982; Hedrick & Thomson, 1986; Hedrick, 1987), where $D^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} D_{ij}^2$,

$H_A = 1 - \sum_{i=1}^{k} p^2(A_i)$ and $H_B = 1 - \sum_{j=1}^{m} p^2(B_j)$.

The final four measures are related to the chi-square statistic to test for independence between alleles at two loci. The chi-square statistic has been discussed by Hedrick (1987) and Hill (1975) as a measure of LD and is defined as

$$\chi^2 = 2N \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{D_{ij}^2}{p(A_i)\, p(B_j)}, \tag{6}$$

where *N* is the sample size and *2N* is the number of haplotypes that occurs in the sample. Two standardized measures of χ^2 have been proposed to quantify LD with values between 0 and 1:

$$\chi^2_{df} = \frac{\chi^2}{2N\,(k-1)\,(m-1)} \tag{7}$$

(Hedrick & Thomson, 1986; Hedrick, 1987), where (*k*-1)(*m*-1) is equal to the degrees of freedom of χ^2, and

$$\chi^{2\prime} = \frac{\chi^2}{2N\,(l\text{-}1)} \tag{8}$$

(Yamazaki, 1977), where $l = \min(k, m)$. The quantity $2N(l-1)$ gives an upper bound for the maximum of χ^2 with given marginals (i.e. given allele frequencies) in a classical χ^2 contingency table. In most cases, however, $2N(l-1)$ is much higher than the true maximum of χ^2 (Kalantari, 1993).

To standardize χ^2 by an upper bound closer to the maximum of χ^2 than $2N(l-1)$, we developed the ninth measure by casting maximization of χ^2 conditional on marginal frequencies as a transportation problem (see Appendix 1; Winston, 1991). The optimal solution to the transportation problem provides a sharper bound, $\chi^2_{\max}$, for the maximum of χ^2 (see Appendix 1; Kalantari, 1993) and is used to standardize χ^2:

$$\chi^2_{tr} = \frac{\chi^2}{\chi^2_{\max}}. \tag{9}$$

Note that for biallelic markers, these nine measures reduce to four because $D' = D_{hap}$ and $r^2 = r^2_{hap} = D* = \chi^2_{df} = \chi^{2\prime}$.

(ii) *Simulation*

The nine measures of marker-marker LD were evaluated for their ability to quantify LD of multi-allelic markers with biallelic QTL in simulated populations. The following criteria were used to determine the most appropriate measure of LD between markers: (1) the measure should have easy interpretation with values between 0 and 1; (2) for a given population, the measure should give a trend of marker-marker LD across distance that is similar to that of marker-QTL LD; and (3) estimates of the level and decline of LD with distance obtained from marker-marker LD should be linearly and highly correlated with the level and decline of marker-QTL LD across population structures and degrees of marker polymorphism.

To allow generation of multiple comparisons between pairs of markers and between markers and QTL at different distances, multiple markers and QTL were simulated on a 100 cM chromosome. In generation zero, markers with 2, 4, 6, 8, or 10 equi-frequent alleles were simulated at 0, 2, …, 100 cM, and QTL with two equi-frequent alleles (*Q* and *q*) were simulated at 1, 3, …, 99 cM. A total of 2*N* haplotypes were randomly sampled by independently selecting alleles at each locus. Thus, all markers and QTL were in Hardy Weinberg and linkage equilibrium in generation zero. Subsequent generations were produced by randomly selecting and mating *N* parents, allowing selfing. Recombination between loci was simulated using the Haldane mapping function (Haldane, 1919).

To generate populations with varying levels of LD, data were generated for 20 combinations of population size (N = 50, 100, 150 or 200) and number of marker alleles (2, 4, 6, 8 or 10) in generation zero. Population size was constant across generations and data on segregating loci in generation 100 were used for analysis. Each population was replicated 100 times. Sved (1971) showed that when the number of generations is large, the expected value of LD becomes steady as a function of the product of effective population size and distance between loci. We verified that LD had reached a 'steady state' condition in generation 100 by comparing the average amount of LD for combinations that resulted in the same product of effective population size and distance between loci. These were found to be similar.

(iii) *Quantification of marker-QTL LD*

Marker-QTL LD at a given distance d in the final generation of each simulated replicate was quantified based on the ability to predict the allele at a biallelic QTL from the observed allele at a linked marker at distance d cM. To measure marker-QTL LD, presence or absence of allele Q in a haplotype consisting of a marker and QTL was treated as a Bernoulli random variable with probability $p(Q)$ of "success" (i.e. presence of Q), and usable marker-QTL LD was quantified as the R^2 of the regression of Q on alleles (A_i) at a single marker. An expression for this R^2 (derived in the Appendix 2) is:

$$R^2 = \sum_{i=1}^{k} p(A_i) \frac{[\, p(Q \mid A_i) - p(Q)\,]^2}{p(Q)\,(1 - p(Q))}, \qquad (10)$$

where $p(Q| A_i)$ is the frequency of Q given A_i. If marker A and the QTL are in linkage equilibrium, then $p(Q \mid A_i) = p(Q)$ and $R^2 = 0$. If $p(Q \mid A_i) \neq p(Q)$, the marker allele

contains information about the QTL allele and $R^2 > 0$. Measure R^2 was used as the standard to evaluate the various LD measures between markers described in (i) and quantified in the simulated populations by regressing each QTL allele separately on each marker. Note that algebraically, $R^2 = \chi^{2}{}'$ for LD between a multi-allelic and a biallelic locus, and $R^2 = \chi^{2}{}' = r^2$ when both loci are biallelic.

(iv) *Comparison of LD curves predicted from marker-QTL and marker-marker LD*

To assess and compare the decline in LD with distance (≤ 20 cM) for marker-QTL LD and marker-marker LD, the function

$$LD_d = 1/(1 + 4\beta d) \qquad (11)$$

(Sved, 1971; Hayes *et al.*, 2003) was fitted to the LD data that were generated for each replicate, where LD_d is LD at distance d Morgans, as measured by the marker-QTL R^2 or by a marker-marker LD measure, and β is a parameter that is related to effective population size (N_e = actual population size for the idealized populations that were simulated (Falconer & Mackay, 1996)). Because the variance of LD tends to decline with distance, a weighted least squares regression, which took heterogeneity of variance of LD into account, was used to estimate β for each simulated data set. The LD data for loci separated by 20 cM or less was used for this purpose. At a given distance (≤ 20 cM), the weight used was the inverse of the LD variance, which was estimated from the LD data for each replicate by using the lowess function in R software (Cleveland, 1979) to fit a smooth curve through the scatterplot of the absolute difference fo the observed LD from the median LD at a given distance. The fraction of data used for smoothing at each distance point was 0.6.

Two criteria were used to compare LD curves estimated from marker-marker LD to those estimated from marker-QTL LD. The first was a measure of the correlation of estimates of β obtained from marker-QTL LD ($\hat{\beta}_{MQ}$) with those from marker-marker LD ($\hat{\beta}_{MM}$) for the various simulation conditions. To evaluate whether this relationship was consistent across population sizes and number of marker alleles, estimates $\hat{\beta}_{MQ}(i,j,k)$ obtained for population size *i* (*i* = 50, 100, 150 or 200), number of marker alleles *j* (*j* = 2, 4, 6, 8 or 10) and replicate *k* (*k* = 1, 2, …, 100) were analyzed using a model that included $\hat{\beta}_{MM}(i,j,k)$ as a covariate, population size *i* and number of marker alleles *j* as class variables, and all interactions among these three variables. The second criterion used to compare estimated LD curves was the mean of the squared difference between LD predicted based on marker-QTL LD and LD predicted using marker-marker LD over distances of 1, 2, ..., 20 cM:

$$MSE = \frac{\sum_{i=1}^{20}\left(L\hat{D}_{MQ(i)} - L\hat{D}_{MM(i)}\right)^2}{20}, \qquad (12)$$

where $L\hat{D}_{MQ(i)}$ and $L\hat{D}_{MM(i)}$ are LD predicted at *i* cM (*i* = 1, 2, …, 20) using $\hat{\beta}_{MQ}$ and $\hat{\beta}_{MM}$, respectively, in equation (11).

(v) *Relationship of marker-QTL LD with local marker-marker LD*

The previous comparisons quantify the extent of LD in a population, as measured by marker-marker LD in relation to marker-QTL LD, as a function of distance. This quantifies the general magnitude and extent of LD within a population. It is, however, well known that the extent of LD within a population can differ from region to region, even if variability of LD is quantified against map distance (Heifetz *et al*., 2005) rather than physical distance (Taillon-

Miller *et al*., 2000; Nordborg & Tavar, 2002). It is, therefore, of interest to determine whether local marker-marker LD can be used to identify genomic regions with high marker-QTL LD. To assess this, LD between two linked markers was compared to the LD of these same markers with a QTL that is bracketed by these markers. For this purpose, usable LD between a pair of markers and a bracketed QTL was quantified by regressing each QTL allele on the haplotype of its two flanking markers. The R^2 of regression of Q on flanking marker haplotype A_iB_j was calculated as:

$$R^2_{hap} = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_iB_j)\frac{[\,p(Q \mid A_iB_j) - p(Q)\,]^2}{p(Q)\,(1-p(Q))}, \qquad (13)$$

where $p(Q \mid A_iB_j)$ is the frequency of Q given A_iB_j. The correlation of marker-marker LD measures with marker-QTL LD was used to indicate whether marker-QTL LD was greater in marker intervals that showed strong marker-marker LD. This was done for various levels of effective population size.

3. Results

(i) *Decline of LD with distance*

Fig. 1 illustrates observed relationships of several LD measures with distance for a representative replicate with a population size of 100 and 4 alleles per marker. Extensive LD between markers and QTL existed at short distances but declined rapidly with distance (Fig. 1A). Similar declines were observed when using r^2, r^2_{hap}, D^*, χ^2, χ^2_{df}, $\chi^{2\prime}$ (Fig. 1C) and χ^2_{tr}. Marker-marker LD measured by D' (Fig. 1B) and D_{hap} was strongly inflated relative to

marker-QTL LD (Fig. 1A), and high values were obtained even for markers in near equilibrium.

To assess the decline of LD with distance, equation (11) was fitted to the sample data for the replicate pictured in Fig. 1. Estimates were $\hat{\beta}$ = 53.3 for marker-QTL LD, and 5.4, 5.4, 92.0, 89.8, 93.5, 110.4, 42.6 and 24.1 for D', D_{hap}, r^2, r^2_{hap}, $D*$, χ^2_{df}, $\chi^{2}{}'$ and χ^2_{tr}, respectively. Measure χ^2 was not used to estimate β because of its non-standardized scale. Estimate $\hat{\beta}$ obtained from $\chi^{2}{}'$ was most similar to $\hat{\beta}$ obtained from marker-QTL LD (42.6 *vs*. 53.3) and resulted in very similar estimated LD curves (Fig. 1C). Based on mean LD at a given distance, the estimated curves appeared to provide a good fit to the data for marker-QTL LD (Fig. 1A) and for all marker-marker LD measures except for D' (Fig. 1B) and D_{hap} due to their inflated values at larger distances.

(ii) *Comparison of LD curves predicted from marker-QTL and marker-marker LD*

Results in this section are based on analyzing 100 replicates for each of the 20 combinations of population size and number of marker alleles. All LD measures were evaluated except χ^2.

Table 1 shows the mean $\hat{\beta}$ across 100 replicates obtained from marker-QTL and marker-marker LD for each simulated scenario. Comparing simulations with 2 and 10 alleles per marker in generation zero, the average number of marker alleles still segregating in generation 100 for *N*=50 were 2 and 2.4, respectively, and corresponding mean estimates of $\hat{\beta}$ for usable marker-QTL LD (R^2) decreased from 52.2 to 34.5 (Table 1). As population size increased, LD due to drift decreased. However, less drift also increased the number of alleles

per marker that remained at segregating loci, e.g., to 2.4 for N=50 and to 5.7 for N=200 when starting with 10 alleles (Table 1), which increased LD by providing more information about the amount of association between alleles at different loci. The combination of these two processes resulted in a decline in mean estimates of $\hat{\beta}$ for R^2 for a given population size with an increase in the number of marker alleles that remained (Table 1). This phenomenon was more obvious for larger population sizes (Table 1). These changes were best captured by mean estimates of $\hat{\beta}$ obtained from marker-marker $\chi^{2\prime}$, which was very close to the mean $\hat{\beta}$ for R^2 (Table 1).

For biallelic markers, r^2 provided good estimates of N_e (recall that the population size is equal to N_e in our simulation) (Table 1). For multi-allelic markers, neither R^2 nor $\chi^{2\prime}$ provided good estimates of N_e (Table 1). Instead, mean $\hat{\beta}$ for χ^2_{df} was closest to the true N_e for most cases, with an upward bias of less than 12% from the true N_e (Table 1). Although slightly worse than χ^2_{df}, mean $\hat{\beta}$ for r^2 and $D*$ were also good estimates of N_e, but were biased downward (Table 1). Mean $\hat{\beta}$ for D' and D_{hap} were very low and did not reflect N_e (Table 1).

To get a better understanding of the relationship of marker-marker LD with marker-QTL LD for a given population, estimates $\hat{\beta}$ obtained from each replicate were analyzed and results are shown in Fig. 2 and Table 2. Fig. 2 illustrates the relationship of $\hat{\beta}$ for marker-QTL LD ($\hat{\beta}_{MQ}$) with $\hat{\beta}$ for marker-marker LD ($\hat{\beta}_{MM}$) across the 20 simulated cases with varying population size and number of marker alleles. Results for biallelic markers were distinctly different from those for multi-allelic markers for D' (Fig. 2A). The same was true

for D_{hap}, r^2, r^2_{hap}, $D*$, χ^2_{df} and χ^2_{tr} (results not shown). Fig. 2B shows a good linear relationship of $\hat{\beta}_{MM}$ for $\chi^{2}{}'$ with $\hat{\beta}_{MQ}$, and the regression lines for bi- and multi-allelic markers were almost overlapping.

Table 2 shows the correlation and slope of the regression of $\hat{\beta}_{MQ}$ on $\hat{\beta}_{MM}$ pictured in Fig. 2 for biallelic, multi-allelic and all markers. Correlations and slopes differed greatly between biallelic and multi-allelic markers for all LD measures except for $\chi^{2}{}'$ (Table 2). For $\chi^{2}{}'$, the correlation of $\hat{\beta}_{MM}$ with $\hat{\beta}_{MQ}$ was consistently high (≥ 0.95) and the slope was close to one, regardless of number of marker alleles (Table 2). Using all markers, the regression line for $\chi^{2}{}'$ in Fig. 2B was $\hat{\beta}_{MQ} = 6.22 + 0.98\hat{\beta}_{MM}$, with a correlation of 0.98, showing good correspondence of this measure of marker-marker LD with marker-QTL LD.

The effects of population size and number of marker alleles on the relationship between $\hat{\beta}_{MQ}$ and $\hat{\beta}_{MM}$ pictured in Fig. 2 were tested using analysis of variance. The proportion of variance in $\hat{\beta}_{MQ}$ that was explained by simple regression on $\hat{\beta}_{MM}$ across the 20 simulated cases was 0.96 for $\chi^{2}{}'$ and ranged from 0.07 to 0.49 for the other LD measures. After including effects of population size, number of marker alleles, and all interactions among them and $\hat{\beta}_{MM}$, these proportions increased slightly for $\chi^{2}{}'$ (from 0.96 to 0.98) but greatly (from as low as 0.07 to 0.97) for other measures. Although population size and number of marker alleles explained significant ($p < 0.001$) amounts of variance in $\hat{\beta}_{MQ}$ for all LD measures (including $\chi^{2}{}'$), the relationship between $\hat{\beta}_{MQ}$ and $\hat{\beta}_{MM}$ was relatively independent of the effects of population size and number of marker alleles for $\chi^{2}{}'$.

Table 3 shows the average MSE (*1000) over 100 replicates for various marker-marker LD measures. The MSE was largest for D' and smallest for $\chi^{2}{}'$ for all 20 simulated cases (Table 3). This implies that, regardless of population size and number of marker alleles in the ranges we considered, LD curves predicted from $\chi^{2}{}'$ were very close to LD curves predicted from marker-QTL LD.

(iii) *Relationship of marker-QTL LD with local marker-marker LD*

Relationship of marker-QTL LD with local marker-marker LD was tested for different combinations of population size (25, 50, 75 or 100) and marker-QTL distance (0.5, 1 or 2 cM). The correlation of $\chi^{2}{}'$ between two biallelic markers with LD of these same markers with a bracketed QTL increased as population size decreased (Table 4). For a population size of 50, correlations were 0.06, 0.11 and 0.10 for marker-QTL distances of 0.5, 1 and 2 cM, respectively (Table 4). The low correlation implies that, in a population with LD generated by drift alone, LD between markers and QTL will be determined by the overall degree of LD in the population, but will not necessarily be greater in marker intervals that show strong LD between markers.

4. Discussion and conclusions

Various measures of LD between multi-allelic markers were evaluated as predictors of usable LD of multi-allelic markers with QTL for the purpose of QTL detection and MAS. The R^2 of the regression of QTL allele on alleles at a single marker was used as the standard for evaluation of the various LD measures between markers because it quantifies the ability to

predict the allele at a linked biallelic QTL based on the observed marker allele. Although biallelic QTL were simulated in this study, the results are expected to hold for multi-allelic QTL as well, because QTL alleles can always be grouped into favorable and unfavorable alleles. Although the focus was on predicting marker-QTL LD, our conclusions also apply to relating multi-allelic marker LD to LD of multi-allelic markers with SNPs. However, results do not apply to predicting LD among SNPs or between SNPs and biallelic QTL, which will be addressed in a subsequent paper.

Our study showed that $\chi^{2}{}'$ is the best measure of LD among multi-allelic markers to predict the extent of LD of those markers with QTL across population sizes and number of marker alleles. Estimates of the decline of LD with distance (β) based on $\chi^{2}{}'$ were highly and linearly related to those obtained for marker-QTL LD across population structures and number of marker alleles, and resulted in very similar LD curves. Corresponding relationships were poorer for the other marker-marker LD measures.

In the simulated populations, extensive marker-QTL LD existed at short distances but declined rapidly with distance. Similar declines were observed for all LD measures between markers, except for D' and D_{hap}. Due to haplotypes of low or zero frequencies in small samples, these measures gave rise to LD estimates that were strongly inflated relative to marker-QTL LD and could be high for markers that were in near equilibrium. Therefore, D' and D_{hap} are not good for high resolution LD mapping of QTL. Measure D' was used to study the extent of LD in the Dutch black-and-white dairy cattle population by Farnir *et al*. (2000), in Coopworth and Romney sheep populations by McRae *et al*. (2002), in the U. K. dairy cattle population by Tenesa *et al*. (2003), and in five populations of commercial pigs by

Nsengimana *et al*. (2004). Using this measure, substantial LD was observed over a long range in all four studies, but it is not clear to what extent this may be a result of the above artifact.

Although β is related to N_e, estimates of β obtained from $\chi^{2\prime}$ and marker-QTL LD (R^2) were not useful estimates of N_e, because they reflect not only N_e, but also the number of marker alleles that remained in the generation under consideration. Sved (1971) showed that for biallelic markers, the decline in LD measured by r^2 estimates N_e, which was also observed in our study (Table 1). For multi-allelic markers, r^2, $D*$ and χ^2_{df} all provided good estimates of N_e.

The upper bound for the maximum of χ^2 used in our new measure χ^2_{tr} is sharper than the upper bound used in $\chi^{2\prime}$. Nevertheless, χ^2_{tr} was a poorer predictor of usable marker-QTL LD than $\chi^{2\prime}$. The reason for this is that marker-QTL LD measured by R^2 attains 1.0 if and only if there is a perfect dependence of QTL alleles on marker alleles. This can only occur when each QTL allele frequency is equal to the sum of the frequencies of one or more alleles at the marker. When this condition is not satisfied, the maximum possible R^2 is less than 1.0, yet the maximum of χ^2_{tr} will be close to 1.0 because χ^2 is standardized by a relatively sharp upper bound to χ^2, conditional on marker allele frequencies. Thus, χ^2_{tr} over-standardizes χ^2 in predicting marker-QTL LD. Nevertheless, χ^2_{tr} might be of interest for other circumstances where the χ^2-metric is used.

In summary, $\chi^{2\prime}$ is recommended to quantify the amount and extent of usable LD in a population for QTL mapping and MAS for a group of multi-allelic markers when LD is

generated by drift alone. However, it must be noted that, while marker-marker LD enables assessment of the general extent of usable LD in populations, high marker-marker LD in specific regions may not necessarily identify regions with high marker-QTL LD; in the simulated data, with LD generated by drift alone, observed LD between two markers was not correlated to LD of these same markers with a bracketed QTL. This implies that, for a given population and when quantified against map distance rather than physical distance, LD between markers and QTL will not necessarily be greater in marker intervals that show strong LD between markers.

The populations under study were simulated with maximum QTL segregation in the founder generation and LD generated by drift alone. Under these circumstances, the effect of mutation on marker-QTL LD should not change our conclusions because mutation rates are generally very low (Falconer & Mackay, 1996). Although selection also causes LD (Bulmer, 1971), it preferentially generates LD between QTL affecting the selected trait rather than between markers and QTL (Farnir *et al*., 2000). Selection decreases N_e, which accordingly increases LD through the effect of drift. Therefore, our conclusions are expected to hold for populations that are under selection or mutation. Selection can, however, result in differences in LD between genomic regions on the linkage map scale because of selective sweeps (Kim & Nielsen, 2004). This would result in some ability of local marker-marker LD to predict the extent of marker-QTL LD relative to other regions in the genome, unlike what was observed here for LD generated by drift alone.

Appendix 1. Derivation of sharp bounds for the maximum of χ^2

Consider *2N* haplotypes with two loci: locus A with *k* alleles and locus B with *m* alleles. The frequency of allele A_i at locus A is a_i ($i = 1, \ldots, k$), the frequency of allele B_j at locus B is b_j ($j = 1, \ldots, m$), and $\sum_{i=1}^{k} a_i = \sum_{j=1}^{m} b_j = 2N$. The frequency of haplotype A_iB_j is x_{ij} such that $\sum_{j=1}^{m} x_{ij} = a_i$, $\sum_{i=1}^{k} x_{ij} = b_j$, $x_{ij} \geq 0$. The classical χ^2 contingency table is:

	B_1	…	…	B_m	
A_1	x_{11}			x_{1m}	a_1
…	…	…	…	…	…
…	…	…	…	…	…
A_k	x_{k1}	…	…	x_{km}	a_k
	b_1	…	…	b_m	$2N$

The chi-square statistic for testing for association between alleles is:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(x_{ij} - a_i b_j/(2N)\right)^2}{a_i b_j/(2N)} = 2N\left(\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{x_{ij}^2}{a_i b_j} - 1\right) = 2N\left(g(x) - 1\right),$$

where $g(x) = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{x_{ij}^2}{a_i b_j}$.

In order to standardize χ^2, we want to find the set of $x = (x_{11}, \ldots, x_{km})$ that can maximize $g(x)$ under the constraints:

$$\sum_{j=1}^{m} x_{ij} = a_i,\ \sum_{i=1}^{k} x_{ij} = b_j \text{ and } x_{ij} \geq 0 \ (i = 1, \ldots, k;\ j = 1, \ldots, m). \qquad \text{(A1)}$$

However, this is computationally hard (Kalantari, 1993). The idea that Kalantari (1993) introduced is to replace $g(x)$ by an upper plane $h(x)$ such that $h(x) \geq g(x)$:

$$h(x) = \sum_{i=1}^{k}\sum_{j=1}^{m} h_{ij}(x_{ij}) = \sum_{i=1}^{k}\sum_{j=1}^{m}\left\{ g_{ij}(l_{ij}) + \frac{g_{ij}(u_{ij}) - g_{ij}(l_{ij})}{u_{ij} - l_{ij}}(x_{ij} - l_{ij})\right\},$$

where $g_{ij}(x_{ij}) = \dfrac{x_{ij}^2}{a_i b_j}$, $l_{ij} = \max(0, a_i + b_j - 2N)$, $u_{ij} = \min(a_i, b_j)$ and $l_{ij} \leq x_{ij} \leq u_{ij}$.

Now the question is how to find the set of $x = (x_{11}, \ldots, x_{km})$ that can maximize $h(x)$ under the constraints in (A1). Maximizing $h(x)$ is equivalent to maximizing

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\left\{ \frac{g_{ij}(u_{ij}) - g_{ij}(l_{ij})}{u_{ij} - l_{ij}} x_{ij}\right\} = \sum_{i=1}^{k}\sum_{j=1}^{m}\{c_{ij}x_{ij}\}, \text{ where } c_{ij} = \frac{g_{ij}(u_{ij}) - g_{ij}(l_{ij})}{u_{ij} - l_{ij}}.$$

Maximizing $\sum_{i=1}^{k}\sum_{j=1}^{m}\{c_{ij}x_{ij}\}$ under the constraints in (A1) is an ordinary linear transportation problem (Winston, 1991) where c_{ij} can be considered as the "cost" for cell (*i*, *j*) in the χ^2 contingency table. It can be solved by the transportation simplex method (Winston, 1991).

If $\hat{x}$ is the optimal solution to this transportation problem, then $\chi^2_{\max} = 2N(h(\hat{x}) - 1)$ is an upper bound for the maximum of χ^2. Kalantari (1993) proved that this upper bound is never worse than the upper bound used in $\chi^{2\prime}$, that is $\chi^2_{\max} \leq (2N)\min(k-1, m-1)$.

A C++ program was developed to solve the transportation problem and to get $\chi^2_{\max}$ given the allele frequencies at two loci.

Appendix 2. Derivation of regression R^2 for marker-QTL LD

Consider *2N* haplotypes with two loci: marker A with *k* alleles and a QTL with two alleles (*Q* and *q*). The estimated frequency of allele A_i (i = 1, 2, …, k) is $p(A_i)$, and the estimated frequency of allele *Q* is $p(Q)$.

Let $Y_j = 1$ (j = 1, 2, …, *2N*) if the allele *Q* is present in the j^{th} haplotype and $Y_j = 0$ otherwise.

Let $X_{ij} = 1$ (i = 1, 2, …, k ; j = 1, 2, …, *2N*) if the allele A_i is present in the j^{th} haplotype and $X_{ij} = 0$ otherwise.

Let $Y = [Y_1,\ldots,Y_{2N}]'$, $X_i = [X_{i1},\ldots,X_{i(2N)}]'$ and $X = [X_1,\ldots,X_k]$.

Then the R^2 for the regression of *Y* on *X* (i.e. the proportion of QTL variance explained by marker A) is:

$$R^2 = \frac{(\hat{Y} - 1\bar{Y})'(\hat{Y} - 1\bar{Y})}{(Y - 1\bar{Y})'(Y - 1\bar{Y})}, \qquad \text{(A2)}$$

where 1 denotes a vector of *2N* ones, $\hat{Y} = X(X'X)^{-1}X'Y$, and $\bar{Y} = \frac{1}{2N}\sum_{j=1}^{2N} Y_j$.

First, we calculate the numerator in (A2). Because $X'X = (2N)diag[p(A_1),\ldots,p(A_k)]$ and $X'Y = 2N[p(Q\,A_1),\ldots,p(Q\,A_k)]'$ where $p(Q\,A_i)$ is the estimated frequency of haplotype QA_i (i = 1, 2, …, k), we get $(X'X)^{-1}X'Y = [p(Q \mid A_1),\ldots,p(Q \mid A_k)]'$ and $\hat{Y}_j = \sum_{i=1}^{k}[X_{ij}\,p(Q \mid A_i)]$ where $p(Q|\,A_i)$ is the estimated conditional probability of allele *Q* given A_i (i = 1, 2, …, k). Therefore,

$$\begin{aligned}\left(\hat{Y}-1\bar{Y}\right)'\left(\hat{Y}-1\bar{Y}\right) &= \sum_{j=1}^{2N}\left[\sum_{i=1}^{k}\left[X_{ij}\,p(Q\mid A_i)\right]-p(Q)\right]^2 \\ &= 2N\sum_{i=1}^{k}p(A_i)\left[p(Q\mid A_i)-p(Q)\right]^2.\end{aligned} \tag{A3}$$

Second, we calculate the denominator in (A2):

$$\left(Y-1\bar{Y}\right)'\left(Y-1\bar{Y}\right)=\sum_{j=1}^{2N}\left[Y_j-p(Q)\right]^2=(2N)\,p(Q)\left[1-p(Q)\right]. \tag{A4}$$

From equations (A2), (A3) and (A4), we get $R^2=\dfrac{\sum_{i=1}^{k}p(A_i)\left[p(Q\mid A_i)-p(Q)\right]^2}{p(Q)\left[1-p(Q)\right]}$.

Acknowledgments

We are very grateful to Rohan Fernando for providing us with programs used in our simulations. We also thank Laura Grapes, Radu Totir, Eli Heifetz and Janet Fulton for their help and valuable discussion. The editor and reviewers are acknowledged for suggestions that resulted in substantial improvements in the final manuscript. This research was supported by State of Iowa Hatch and Multi-state Research Funds.

References

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

Bulmer, M. G. (1971). The effect of selection on genetic variability. *The American Naturalist* **105**, 201-211.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829-836.

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

Devlin, B. & Risch, N. (1995). A Comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311-322.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. Edinburgh Gate, Harlow, Essex, England: Addison Wesley Longman.

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. & Georges, M. (2000). Extensive genome-wide LD in cattle. *Genome Research* **10**, 220-227.

Flint-Garcia, S. A., Thornsberry, J. M. & Buckler IV, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**, 357-374.

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 299-309.

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

Hedrick, P. W. & Thomson, G. (1986). A two-locus neutrality test: application to humans, *E. Coli* and Lodgepole pine. *Genetics* **112**, 135-156.

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

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

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

Hill, W. G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**, 117-126.

Kalantari, B. (1993). Sharp bounds for the maximum of the chi-square index in a class of contingency tables with given marginals. *Computational Statistics & Data Analysis* **16**, 19-34.

Karlin, S. & Piazza, A. (1981). Statistical methods for assessing linkage disequilibrium at the HLA-A, B, C loci. *Annals of Human Genetics* **45**, 79-94.

Kim, Y. & Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513-1524.

Lande, R. & Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743-756.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49-67.

Maruyama, T. (1982). Stochastic integrals and their application to population genetics. In: *Molecular Evolution, Protein Polymorphism and the Neutral Theory*, pp. 151-166. Tokyo: Japan Scientific Societies Press.

McRae, A. F., McEwan, J. C., Dodds, K. G., Wilson, T., Grawford, A. M. & Slate, J. (2002). Linkage disequilibrium in domestic sheep. *Genetics* **160**, 1113-1122.

Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P.-Y. & Collins, A. (2001). The optimal measure of allelic association. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5217-5221.

Nordborg, M. & Tavar, S. (2002). Linkage disequilibrium: what histoty has to tell us. *Trends in Genetics* **18**, 83-90.

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

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

Sabatti, C. & Risch, N. (2002). Homozygosity and linkage disequilibrium. *Genetic*s **160**, 1707-1719.

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

Taillon-Miller, P., Bauer-Sardiña, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P. & Kwok, P.-Y. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genetics* **25**, 324-328.

Tenesa, A., Knott, S. A., Ward, D., Smith, D., Williams, J. L. & Visscher, P. M. (2003). Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *Journal of Animal Science* **81**, 617-623.

Terwilliger, J. D. & Weiss, K. M. (1998). Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* **9**, 578-594.

Winston, W. L. (1991). *Operations Research: Applications and Algorithms*, 2nd edn. Boston: PWS-Kent.

Yamazaki, T. (1977). The effects of overdominance on linkage in a multilocus system. *Genetics* **86**, 227-236.

Table 1. *Mean estimates of the decline of LD with distance (β) over 100 replicates based on a measure of marker-QTL LD (R^2) and eight measures of marker-marker LD for simulated data based on different combinations of number of marker alleles in generation zero (g_0) and population size. The number of marker alleles in generation 100 (g_{100}) is the average of the mean number of alleles across markers still segregating in g_{100} over 100 replicates*

# marker alleles (g_0)	Population size	# marker alleles (g_{100})	R^2	D'	D_{hap}	r^2	r^2_{hap}	D^*	χ^2_{df}	$\chi^{2'}$	χ^2_{tr}
2	50	2	52.2	2.5	2.5	55.7	55.7	55.7	55.7	55.7	13.1
	100	2	102.0	4.2	4.2	101.5	101.5	101.5	101.5	101.5	24.6
	150	2	152.4	6.6	6.6	148.6	148.6	148.6	148.6	148.6	40.5
	200	2	199.4	9.9	9.9	193.0	193.0	193.0	193.0	193.0	62.9
4	50	2.2	39.9	2.7	2.7	49.1	48.8	49.3	51.1	35.5	12.8
	100	2.8	55.0	5.5	5.6	92.6	89.7	93.8	104.6	46.2	25.6
	150	3.3	64.2	8.7	9.1	130.3	123.4	133.0	156.1	56.0	37.0
	200	3.6	75.6	11.3	12.2	176.4	164.6	178.9	207.0	69.0	48.5
6	50	2.3	36.8	2.8	2.8	47.8	47.2	48.0	50.5	31.6	12.8
	100	3.2	44.4	5.7	5.9	89.4	83.9	91.3	106.3	37.4	23.7
	150	4.0	48.1	8.0	8.7	130.1	115.4	133.4	161.4	42.5	31.0
	200	4.6	53.7	9.3	10.7	174.1	148.2	178.5	218.1	49.0	37.3
8	50	2.4	35.8	2.7	2.7	46.9	46.3	47.2	49.8	29.8	12.4
	100	3.5	39.5	5.7	6.0	87.3	79.6	89.6	107.5	33.1	22.5
	150	4.4	42.2	7.4	8.3	128.3	109.1	131.7	162.6	37.3	28.0
	200	5.2	45.1	8.3	10.0	170.8	137.9	175.5	219.0	41.5	32.4
10	50	2.4	34.5	2.9	2.9	47.1	46.3	47.7	51.4	29.4	13.0
	100	3.6	37.4	5.7	6.0	87.9	79.1	90.4	107.8	31.5	21.8
	150	4.7	38.4	7.0	8.1	127.1	105.0	131.0	165.5	34.3	26.2
	200	5.7	40.2	7.4	9.3	170.4	130.8	175.3	222.7	37.3	29.4

Table 2. *Correlation and slope of the regression of the decline of LD with distance (β) estimated from marker-QTL LD on β estimated from different measures of marker-marker LD for biallelic, multi-allelic (4, 6, 8 or 10) and all markers across four population sizes (50, 100, 150 or 200), with 100 replicates for each combination of population size and number of marker alleles*

		D'	D_{hap}	r^2	r^2_{hap}	D^*	χ^2_{df}	$\chi^{2\prime}$	χ^2_{tr}
Biallelic markers	Correlation	0.87	0.87	0.95	0.95	0.95	0.95	0.95	0.87
	Slope	16.05	16.05	1.01	1.01	1.01	1.01	1.01	2.37
Multi-allelic markers	Correlation	0.69	0.61	0.57	0.69	0.56	0.49	0.95	0.79
	Slope	3.15	2.40	0.14	0.22	0.14	0.09	1.02	0.93
All markers	Correlation	0.36	0.26	0.50	0.65	0.47	0.29	0.98	0.70
	Slope	5.55	3.53	0.43	0.64	0.40	0.20	0.98	2.19

Table 3. *The mean of the squared difference (MSE) between LD predicted based on marker-marker and marker-QTL LD at 1, 2, …, 20 cM for simulated data generated from different combinations of population size and number of marker alleles in generation zero (g_0). Values are the average MSE over 100 replicates multiplied by 1000 for each combination. Results for D_{hap} (not shown) were similar to those for D', and results for r^2_{hap}, D^* and χ^2_{df} were similar to those for r^2*

# marker alleles (g_0)	Population size	D'	r^2	$\chi^2{}'$	$\chi^2{}_{tr}$
2	50	242.4	0.3	0.3	30.5
	100	168.6	0.1	0.1	15.9
	150	115.4	0.0	0.0	7.9
	200	75.5	0.0	0.0	3.6
4	50	197.7	0.5	0.3	22.0
	100	103.0	1.2	0.3	5.9
	150	59.0	1.6	0.1	2.2
	200	43.5	1.6	0.0	1.1
6	50	183.1	0.8	0.4	19.4
	100	86.2	2.5	0.3	4.5
	150	55.6	3.8	0.1	1.8
	200	47.1	4.0	0.1	1.0
8	50	186.2	0.8	0.6	20.0
	100	81.7	3.6	0.3	3.9
	150	56.6	5.2	0.2	1.7
	200	50.1	6.0	0.1	1.0
10	50	173.5	1.0	0.4	17.2
	100	78.0	4.3	0.3	3.7
	150	57.6	6.5	0.1	1.6
	200	54.6	7.6	0.1	1.0

Table 4. *The correlation of observable LD between two markers using χ^2' with LD of these same markers with a bracketed QTL. Markers and QTL were bi-allelic and segregating in generation 100. Population size was 25, 50, 75 or 100. Marker-QTL distance was 0.5, 1 or 2 cM. Results are based on 10,000 replicates*

Population size	Marker-QTL distance (cM)		
	0.5	1	2
25	0.19	0.21	0.19
50	0.06	0.11	0.10
75	0.02	0.07	0.07
100	0.02	0.06	0.06

Fig. 1

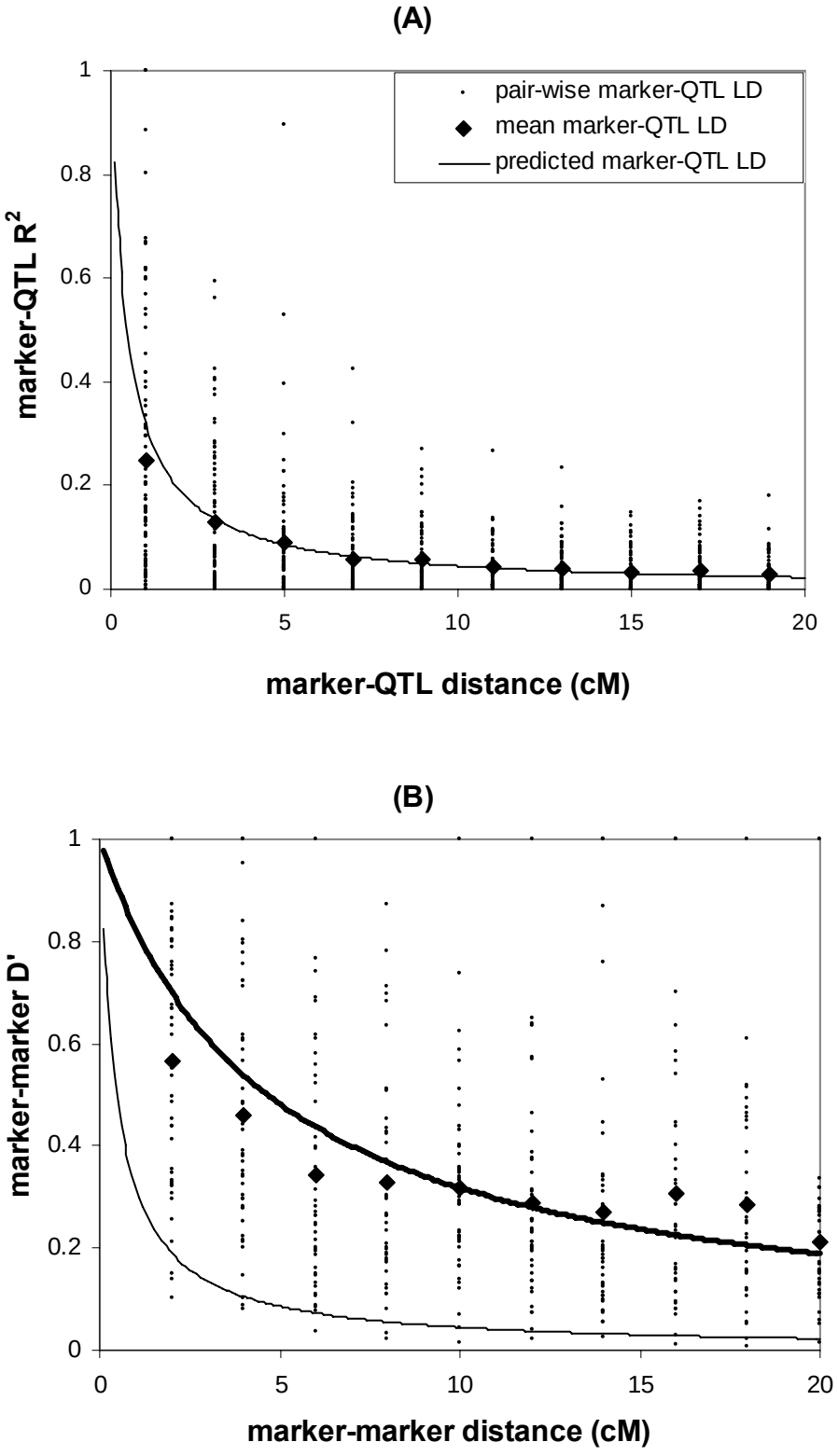

(A)
pair-wise marker-QTL LD
mean marker-QTL LD
predicted marker-QTL LD
marker-QTL R2
marker-QTL distance (cM)
(B)
marker-marker D'
marker-marker distance (cM)

Fig. 1 (continued)

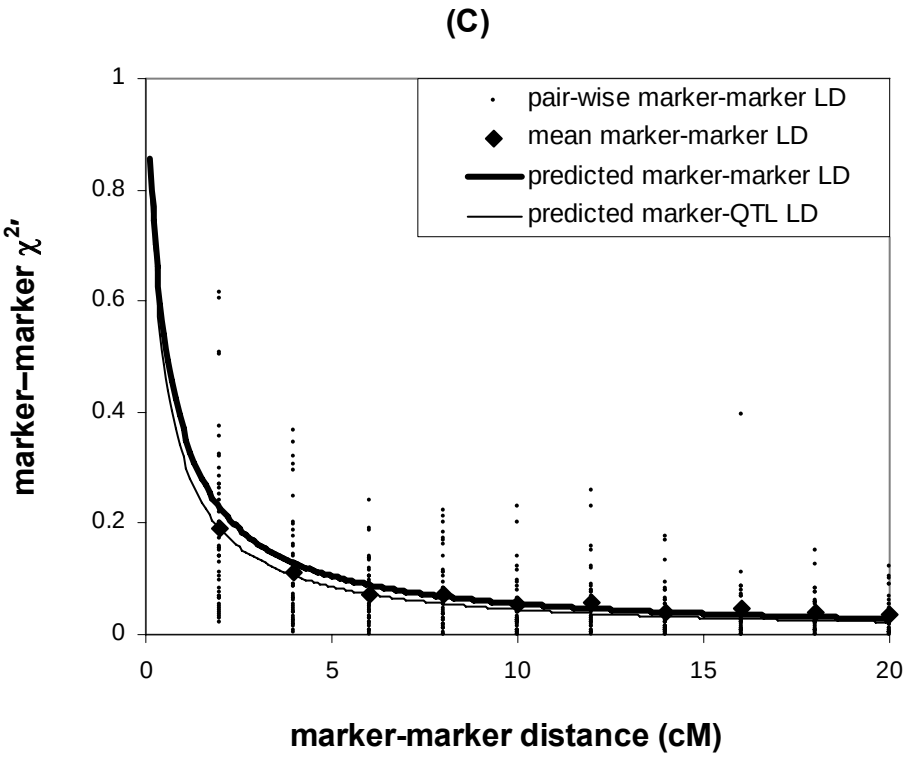


Fig. 1. Observed relationships of marker-QTL LD (A) and marker-marker LD measured by D′ (B) and $\chi^{2\prime}$ (C) against map distance for a representative replicate with a population size of 100 and 4 alleles per marker. Legend for (B) is the same as legend for (C). LD at distance *d* Morgans was predicted from $\hat{LD}_d = 1/\left(1+4\hat{\beta}\,d\right)$, where $\hat{\beta}$ was obtained from the simulated data for each LD measure.

Fig. 2

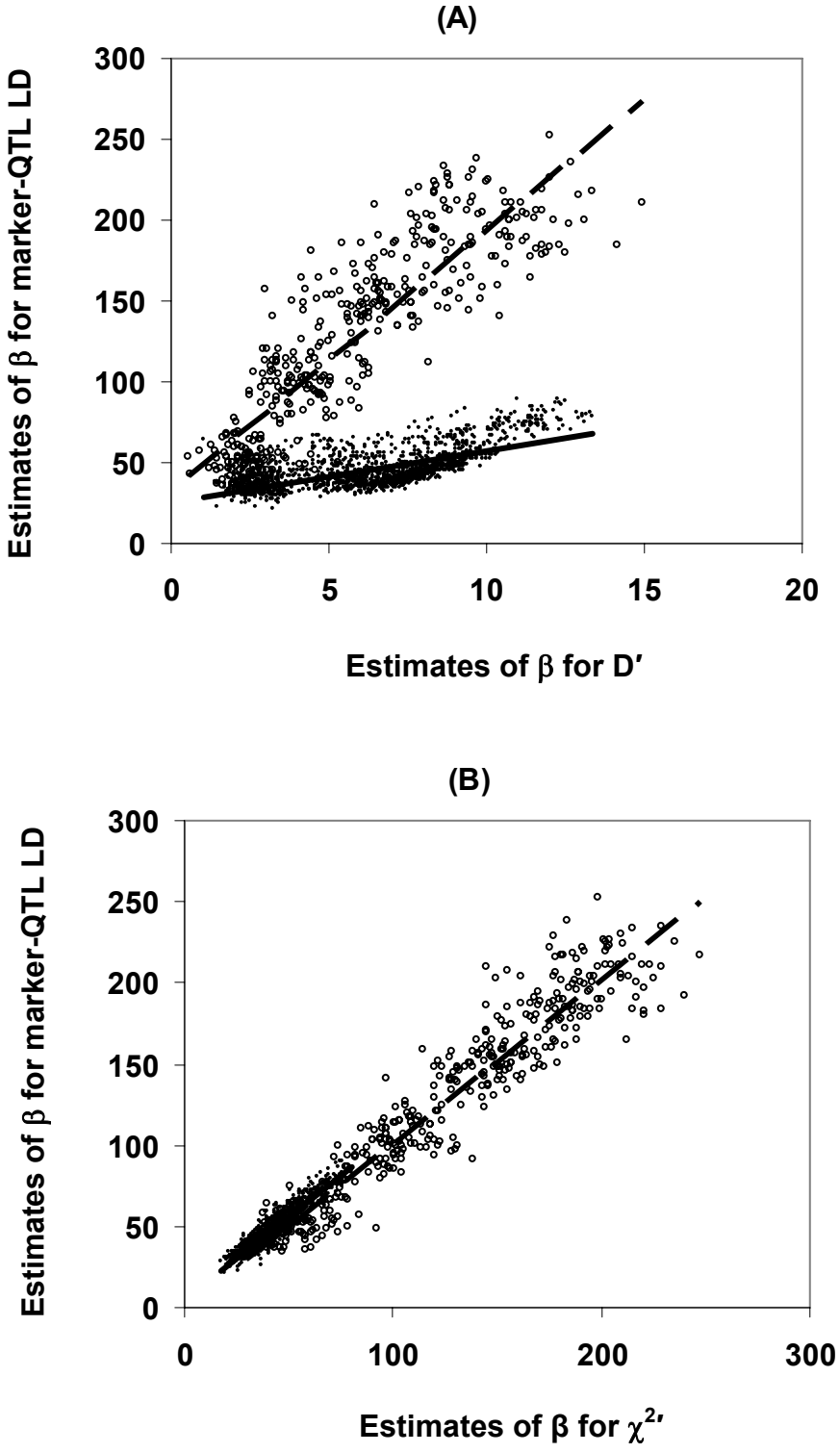

(A)
Estimates of β for marker-QTL LD
300
250
200
150
100
50
0
0
5
10
15
20
Estimates of β for D′
(B)
Estimates of β for marker-QTL LD
300
250
200
150
100
50
0
0
100
200
300
Estimates of β for χ2′

Fig. 2. Regression of estimates of the decline of LD with distance (β) obtained from each replicate for marker-QTL LD on estimates of β for marker-marker LD measured by D' (A) and $\chi^2{}'$ (B) for biallelic (open circle) and multi-allelic (dot) markers. Dashed and solid lines indicate the regression lines for bi- and multi-allelic markers, respectively. Data are based on 100 replicates simulated for each of the 20 combinations of population size (50, 100, 150 or 200) and number of marker alleles (2, 4, 6, 8 or 10).

CHAPTER 4. EVALUATION OF LINKAGE DISEQUILIBRIUM MEASURES BETWEEN MULTI-ALLELIC MARKERS AS PREDICTORS OF LINKAGE DISEQUILIBRIUM BETWEEN SINGLE NUCLEOTIDE POLYMORPHISMS AND QTL

A paper to be submitted to *Genetical Research*

H. Zhao[1], D. Nettleton[2] and J. C. M. Dekkers[1]

[1] Department of Animal Science and Center for Integrated Animal Genomics, 239 Kildee Hall, Iowa State University, Ames, Iowa, 50011, USA

[2] Department of Statistics, 124 Snedecor Hall, Iowa State University, Ames, Iowa, 50011, USA

Summary

Effectiveness of marker-assisted selection and quantitative trait loci (QTL) mapping using population-wide linkage disequilibrium (LD) between markers and QTL depends on the extent of LD and how it declines with distance between markers and QTL in a population. Marker-QTL LD can be predicted from LD between markers. In our previous work, observable LD between multi-allelic markers measured by a standardized chi-square statistic (χ^{2}') was found to be the best predictor of usable LD of multi-allelic markers with QTL. Since single nucleotide polymorphisms (SNPs) are the current marker of choice for high density genotyping and LD–mapping of QTL, the objective of this study was to use LD between multi-allelic markers to predict usable LD between biallelic SNPs and QTL.

Observable LD between marker pairs was evaluated using nine measures. These included two pooled and standardized measures of LD between pairs of alleles at two markers based on Lewontin's LD measure, two pooled measures of squared correlations between alleles, one standardized measure using Hardy-Weinberg heterozygosities, and four measures based on the chi-square statistic for testing for association between alleles at two loci. Although $\chi^{2}{}'$ is a good predictor of LD of multi-allelic markers with biallelic QTL, it over-estimated usable SNP-QTL LD. Measures χ^{2}_{df}, r^{2} and $D*$ were found to be good predictors of usable SNP-QTL LD when LD is generated by drift. Measures D' and D_{hap} between multi-allelic markers, although not recommended for measuring LD due to their inflated LD estimates, can be used to predict D' of SNP-QTL and SNP-SNP LD.

1. Introduction

Effectiveness of marker-assisted selection (MAS) and quantitative trait loci (QTL) mapping using population-wide linkage disequilibrium (LD) between markers and QTL depends on the extent of LD and how it declines with distance in a population. Although marker-QTL LD cannot be observed directly, it can be predicted from LD between markers. Zhao *et al*. (2005) evaluated nine LD measures between multi-allelic markers as predictors of usable LD between the same group of markers and biallelic QTL. When LD is generated by drift, a standardized chi-square statistic ($\chi^{2}{}'$) was recommended to quantify the amount and extent of usable LD in a population for QTL mapping and MAS based on multi-allelic markers (Zhao *et al*., 2005).

While highly polymorphic microsatellite markers are still popularly used in genome-wide linkage analysis to track inheritance of chromosome regions, biallelic single nucleotide polymorphism (SNP) markers have been receiving more and more attention in genetics research. In addition to their abundance in the genome, recent advances in technology have made large-scale SNP genotyping rapid, accurate, and inexpensive (Kwok, 2001). High density SNP maps are now available for both human and livestock. For example, a SNP map of the human genome containing 1.42 million SNPs (International SNP Map Working Group, 2001) and a genetic variation map for the chicken genome containing 2.8 million SNPs (International Chicken Polymorphism Map Consortium, 2004) have been constructed.

These exciting developments of dense SNP maps present tremendous opportunities for high resolution LD mapping of QTL. Within a closed breeding population in livestock, LD is limited to closely linked loci due to many generations of recombination. Therefore, high density SNP genotyping enables detection and fine-mapping of QTL in outbred populations using historical LD, and resulting QTL can immediately be implemented for MAS (Dekkers & Hospital, 2002; Grapes *et al.*, 2004; Meuwissen & Goddard, 2000). A crucial issue in using high density SNP maps is the extent of LD among SNPs or between SNPs and QTL, which impacts the power of LD mapping and effectiveness of MAS and is needed to determine the SNP density that is required to obtain a given power to detect QTL. Harmegnies *et al*. (2006) evaluated the extent of LD in two commercial pig populations using microsatellite markers and found extensive LD in both populations. Since microsatellite markers are still frequently used, it is of interest to predict the extent of LD that exists in a population among SNPs or between SNPs and QTL based on LD between available microsatellite markers, which is the objective of this study. This research has

practical implications. Before collecting data on SNPs, it is important to know how many SNPs and what sort of density will be needed. The data we do have on microsatellites can help us address this prior to collecting any SNP data.

2. Materials and methods

The methods in this paper are the same as in Zhao *et al*. (2005). Briefly, observable LD between marker pairs was evaluated using nine alternate measures: (1) D' is based on Lewontin's normalized LD measure weighted by the product of allele frequencies; (2) D_{hap} is similar to D' but weighted by haplotype frequencies; (3) r^2 is pooled square of correlations between alleles weighted by the product of allele frequencies; (4) r^2_{hap} is similar to r^2 but weighted by haplotype frequencies; (5) $D*$ uses the Hardy-Weinberg heterozygosities to normalize LD; (6) χ^2 is the chi-square statistic for testing for association between alleles at two loci; (7) $\chi^2_{df} = \chi^2/(2N*df)$, where N is the population size and df denotes degrees of freedom; (8) $\chi^2{}' = \chi^2/[2N(l-1)]$, where l is the smallest number of alleles across the two markers, and $2N(l-1)$ provides an upper bound for the maximum of χ^2; and (9) $\chi^2_{tr} = \chi^2/\chi^2_{\max}$, where $\chi^2_{\max}$ is an upper bound for the maximum of χ^2 which is sharper than $2N(l-1)$. For LD between biallelic markers, $D' = D_{hap}$ and $r^2 = r^2_{hap} = D* = \chi^2_{df} = \chi^2{}'$.

These nine measures of marker-marker LD were evaluated for their ability to quantify LD between biallelic SNPs and QTL or among SNPs in simulated populations. On a 100 cM chromosome, markers with 2, 4, 6, 8 or 10 alleles in generation zero were simulated at 0, 2, …, 100 cM and biallelic SNPs at 1, 3, …, 99 cM. All loci were in Hardy Weinberg and

linkage equilibrium in generation zero. LD was generated by drift by 100 generations of random mating of N parents (N = 50, 100, 150 or 200). Data on segregating loci in generation 100 were used for analysis.

Estimates of SNP-QTL LD (or SNP-SNP LD) were obtained from LD between a pair of biallelic SNPs in our simulation and measured by r^2 and D'. LD measured by r^2 is equivalent to usable LD for biallelic markers (Zhao *et al*., 2005). Because many studies have used D' to evaluate multi-allelic marker LD (Farnir *et al*., 2000; McRae *et al*., 2002; Nsengimana *et al.,* 2004; Tenesa *et al*., 2003;), we evaluated the ability of multi-allelic D' to predict biallelic D'.

To assess and compare the decline in LD with distance (≤ 20 cM) for marker-QTL LD and marker-marker LD, the function $LD_d = 1/(1+4\beta d)$ (Hayes *et al.*, 2003; Sved, 1971) was fitted to the LD data that were generated for each replicate, where LD_d is LD at distance d Morgans, as measured by the SNP-QTL (or SNP-SNP) r^2 or D' or by a marker-marker LD measure, and β is a parameter that is related to effective population size (N_e = actual population size for the idealized populations that were simulated (Falconer & Mackay, 1996)). A weighted least squares regression was used to estimate β for each simulated data set, as described in Zhao *et al*. (2005).

Following the same criteria as described in Zhao *et al*. (2005), LD curves predicted from different measures of multi-allelic marker-marker LD were compared to SNP-QTL (or SNP-SNP) LD measured by: (1) r^2 to find which multi-allelic marker measure best predicts usable SNP-QTL LD, and (2) D' to find which multi-allelic marker measure best predicts SNP-QTL LD and SNP-SNP LD based on D'.

3. Results

(i) *Decline of LD with distance*

The observed relationships of several LD measures with distance for a representative replicate with a population size of 100 and 4 alleles per marker are illustrated in Fig. 1. Usable SNP-QTL LD measured by r^2 was relatively high at short distances and declined rapidly with distance (Fig. 1A). Similar declines were observed when r^2, r^2_{hap}, D^*, χ^2, χ^2_{df} (Fig. 1D), $\chi^{2\prime}$ and χ^2_{tr} were used to measure marker-marker LD. The SNP-QTL LD measured by D' was strongly inflated relative to SNP-QTL r^2 (compare Fig. 1A *vs*. 1B), and high LD values were obtained even for markers that approached equilibrium. The same was true for marker-marker LD measured by D' (Fig. 1C) and D_{hap}.

To assess the decline of LD with distance, equation $LD_d = 1/(1+4\beta d)$ was fitted to the sample data for the replicate pictured in Fig. 1. Estimates were $\hat{\beta}$ = 86.8 for SNP-QTL r^2 and 3.3 for SNP-QTL D', and 5.4, 5.4, 92.0, 89.8, 93.5, 110.4, 42.6 and 24.1 for marker-marker LD measured by D', D_{hap}, r^2, r^2_{hap}, D^*, χ^2_{df}, $\chi^{2\prime}$ and χ^2_{tr}, respectively. Measure χ^2 was not used to estimate β because of its non-standardized scale. The LD curve predicted from SNP-QTL r^2 was very close to LD curves predicted from marker-marker LD measured by r^2, r^2_{hap}, D^* and χ^2_{df} (Fig. 1D). The LD curve predicted from SNP-QTL D' was close to LD curves predicted from marker-marker LD measured by D' (Fig. 1C) and D_{hap}. Based on mean LD at a given distance (Fig. 1), the estimated curves appeared to provide a good fit to the data for all LD measures except for D' (Fig. 1B, 1C) and D_{hap} due to their inflated values at larger distances.

(ii) *Comparison of LD curves predicted from SNP-QTL and marker-marker LD*

Results in this section are based on analyzing 100 replicates for each of the 20 combinations of population size and number of marker alleles. All LD measures were evaluated except χ^2.

The mean $\hat{\beta}$ across 100 replicates obtained from SNP-QTL and marker-marker LD for each simulated scenario is shown in Table 1. The mean $\hat{\beta}$ for r^2, $D*$ and χ^2_{df} between markers were very close to the mean $\hat{\beta}$ for SNP-QTL r^2, and they all provided good estimates of N_e (Table 1). With more than 2 alleles per marker in generation zero, the mean estimates of $\hat{\beta}$ obtained from marker-marker $\chi^{2\prime}$ were much lower than the mean $\hat{\beta}$ for SNP-QTL r^2 (Table 1). Because of the relationship between LD at a given distance and β based on equation $LD_d = 1/(1+4\beta d)$, this implies that measure $\chi^{2\prime}$ over-estimated usable SNP-QTL LD. The mean $\hat{\beta}$ for D' and D_{hap} between markers were very close to the mean $\hat{\beta}$ for SNP-QTL D' (Table 1).

The relationship of marker-marker LD with SNP-QTL LD for a given population was further analyzed using estimates $\hat{\beta}$ obtained from each replicate. Fig. 2A, 2B and Table 2A illustrate the relationship of $\hat{\beta}$ for SNP-QTL LD measured by r^2 with $\hat{\beta}$ for marker-marker LD across the 20 simulated cases with varying population sizes and numbers of marker alleles. Using all markers, a good linear relationship was observed for marker-marker LD measured by r^2 (Fig. 2A), $D*$ and χ^2_{df}, with a correlation of 0.93, 0.93, 0.94 and slope of 1.0, 1.0, 0.8, respectively (Table 2A). Corresponding relationships were poorer for $\chi^{2\prime}$ (Fig. 2B, Table 2A) and for the other marker-marker LD measures (Table 2A). The mean of the

squared difference (MSE) averaged over 100 replicates between LD predicted based on SNP-QTL r^2 and marker-marker LD measured by r^2, r^2_{hap}, $D*$ and χ^2_{df} was low for all 20 simulated cases (Table 3A). Therefore, usable LD between SNPs and QTL can be best predicted from LD between multi-allelic markers measured by r^2, $D*$ and χ^2_{df}, but not by $\chi^{2'}$.

Corresponding relationship of $\hat{\beta}$ for SNP-QTL LD measured by D' with $\hat{\beta}$ for marker-marker LD is shown in Fig. 2C and Table 2B. The relationship appeared to be linear for marker-marker LD measured by D' (Fig. 2C) and D_{hap}. Using all markers, correlations were 0.79 and 0.83 and slopes were 0.90 and 0.83 for D' and D_{hap}, respectively (Table 2B), while slopes ranged from 0.02 to 0.16 for the other marker-marker LD measures (Table 2B). The MSE between LD predicted based on SNP-QTL D' and marker-marker LD measured by D' and D_{hap} was much lower compared to the other marker-marker LD measures (Table 3B) for all 20 simulated cases. This implies that, D' and D_{hap} between multi-allelic markers, although not recommended for measuring LD, can predict SNP-QTL LD based on D'.

4. Discussion and conclusions

Zhao *et al*. (2005) evaluated various measures of LD between multi-allelic markers as predictors of usable LD of multi-allelic markers with QTL for the purpose of QTL detection and MAS. Their study showed that $\chi^{2'}$ is the best measure of LD among multi-allelic markers to predict the extent of LD of those markers with QTL, across population sizes and numbers of marker alleles (Zhao *et al*., 2005). Since SNPs are the current marker of choice

for high density genotyping and LD–mapping of QTL, it is also of interest to predict the extent of LD that exists in a population among biallelic SNPs or between SNPs and QTL based on LD between available microsatellite markers.

Our study shows that LD between multi-allelic markers measured by r^2, $D*$ and χ^2_{df} are good predictors of usable SNP-QTL LD when LD is generated by drift. Although $\chi^{2}{}'$ is the best predictor of marker-QTL LD based on the same group of multi-allelic markers (Zhao *et al*., 2005), it over-estimated usable LD between SNPs and QTL. The decline of LD with distance estimated from marker-marker r^2, $D*$ and χ^2_{df} and SNP-QTL r^2 all provide good estimates of N_e; but for $\chi^{2}{}'$, the decline of LD reflects not only N_e but also the number of marker alleles that remained in the generation under consideration (Zhao *et al*., 2005). Therefore, the LD measure between multi-allelic markers that is best for predicting usable LD in a population depends on the type of markers that will eventually be used for QTL mapping or MAS (*i.e.* multi-allelic or biallelic).

Because many previous studies have used D' to evaluate multi-allelic marker LD (Farnir *et al*., 2000; McRae *et al*., 2002; Nsengimana *et al.,* 2004; Tenesa *et al*., 2003), we also evaluated its ability to predict D' for biallelic loci. We found that SNP-QTL LD and SNP-SNP LD based on D' can be predicted from LD between multi-allelic markers measured by D' and D_{hap}. However, they are not recommended to quantify LD due to their inflated LD estimates (Zhao *et al*., 2005).

Although our study is based on simulated populations where LD was generated by drift alone, the conclusions are expected to hold for populations that are under selection or mutation, as reasoned by Zhao *et al*. (2005).

Acknowledgments

We are very grateful to Rohan Fernando for providing us with programs used in our simulations. We also thank Morris Soller, Eli Heifetz and Janet Fulton for their help and valuable discussion. This research was supported by State of Iowa Hatch and Multi-state Research Funds.

References

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

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. Harlow, UK: Addison-Wesley Longman.

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. & Georges, M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10**, 220-227.

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

Harmegnies, N., Farnir, F., Davin, F., Buys, N., Georges, M. & Coppieters, W. (2006). Measuring the extent of linkage disequilibrium in commercial pig populations. *Animal Genetics* **37**, 225-231.

International Chicken Polymorphism Map Consortium (2004). A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* **432**, 717-722.

International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-933.

Kwok, P. Y. (2001). Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics* **2**, 235-258.

McRae, A. F., McEwan, J. C., Dodds, K. G., Wilson, T., Grawford, A. M. & Slate, J. (2002). Linkage disequilibrium in domestic sheep. *Genetics* **160**, 1113-1122.

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

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

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

Tenesa, A., Knott, S. A., Ward, D., Smith, D., Williams, J. L. & Visscher, P. M. (2003). Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *Journal of Animal Science* **81**, 617-623.

Zhao, H., Nettleton, D., Soller, M. & Dekkers, J. C. M. (2005). Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genetical Research* **86**, 77-87.

Table 1. *Mean estimates of the decline of LD with distance (β) over 100 replicates based on SNP-QTL LD and marker-marker LD for simulated data based on different combinations of number of marker alleles in generation zero (g_0) and population size. Results for marker-marker LD measured by* D_{hap}, r^2_{hap} and χ^2_{tr} *(not shown) can be found in Zhao et al. (2005)*

No. of marker alleles (g_0)	Population size	SNP-QTL LD		*Marker-marker LD*				
		D′	r^2	*D′*	r^2	*D**	χ^2_{df}	$\chi^{2'}$
2	50	2.6	55.0	2.5	55.7	55.7	55.7	55.7
	100	4.1	101.9	4.2	101.5	101.5	101.5	101.5
	150	6.9	146.4	6.6	148.6	148.6	148.6	148.6
	200	10.1	190.0	9.9	193.0	193.0	193.0	193.0
4	50	2.6	53.7	2.7	49.1	49.3	51.1	35.5
	100	4.2	102.2	5.5	92.6	93.8	104.6	46.2
	150	6.3	146.5	8.7	130.3	133.0	156.1	56.0
	200	9.9	190.6	11.3	176.4	178.9	207.0	69.0
6	50	2.5	55.0	2.8	47.8	48.0	50.5	31.6
	100	4.1	101.5	5.7	89.4	91.3	106.3	37.4
	150	6.5	146.6	8.0	130.1	133.4	161.4	42.5
	200	9.7	192.1	9.3	174.1	178.5	218.1	49.0
8	50	2.4	56.5	2.7	46.9	47.2	49.8	29.8
	100	4.2	103.5	5.7	87.3	89.6	107.5	33.1
	150	6.3	146.6	7.4	128.3	131.7	162.6	37.3
	200	10.0	190.6	8.3	170.8	175.5	219.0	41.5
10	50	2.5	55.6	2.9	47.1	47.7	51.4	29.4
	100	4.2	106.6	5.7	87.9	90.4	107.8	31.5
	150	6.5	145.3	7.0	127.1	131.0	165.5	34.3
	200	10.2	193.2	7.4	170.4	175.3	222.7	37.3

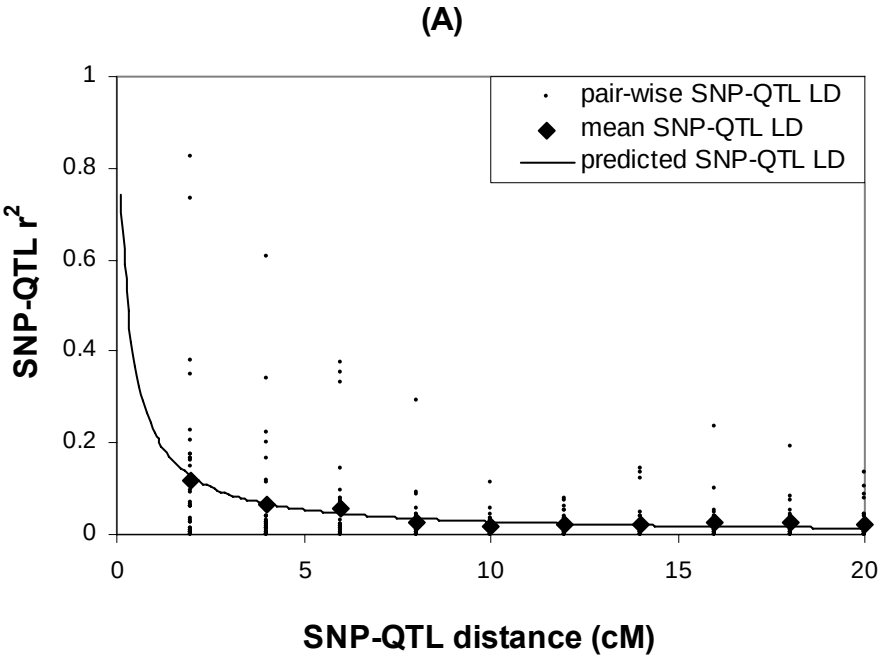
Table 2. *Correlation and slope of the regression of the decline of LD with distance (β) estimated from (A) SNP-QTL r^2 and (B) SNP-QTL D′ on β estimated from different measures of marker-marker LD for biallelic, multi-allelic (4, 6, 8 or 10) and all markers across four population sizes (50, 100, 150 or 200), with 100 replicates for each combination of population size and number of marker alleles*

		D'	D_{hap}	r^2	r^2_{hap}	D^*	χ^2_{df}	$\chi^2{}'$	χ^2_{tr}
(A) SNP-QTL r^2									
Biallelic markers	Correlation	0.84	0.84	0.92	0.92	0.92	0.92	0.92	0.84
	Slope	14.61	14.61	0.91	0.91	0.91	0.91	0.91	2.16
Multi-allelic markers	Correlation	0.87	0.91	0.94	0.92	0.95	0.94	0.57	0.84
	Slope	17.70	15.80	1.06	1.28	1.04	0.80	2.72	4.35
All markers	Correlation	0.86	0.89	0.93	0.88	0.93	0.94	0.35	0.76
	Slope	16.82	15.25	1.01	1.08	1.00	0.82	0.43	2.99
(B) SNP-QTL D'									
Biallelic markers	Correlation	0.83	0.83	0.85	0.85	0.85	0.85	0.85	0.82
	Slope	0.85	0.85	0.05	0.05	0.05	0.05	0.05	0.13
Multi-allelic markers	Correlation	0.79	0.84	0.88	0.85	0.88	0.88	0.52	0.76
	Slope	0.93	0.84	0.06	0.07	0.06	0.04	0.14	0.23
All markers	Correlation	0.79	0.83	0.87	0.82	0.87	0.87	0.34	0.72
	Slope	0.90	0.83	0.05	0.06	0.05	0.04	0.02	0.16

Table 3. *The mean of the squared difference (MSE) between LD predicted based on marker-marker LD and (A) SNP-QTL r^2, (B) SNP-QTL D′ at 1, 2, …, 20 cM for simulated data generated from different combinations of population size and number of marker alleles in generation zero (g_0). Values are the average MSE over 100 replicates multiplied by 1000 for each combination. Results for D_{hap} (not shown) were similar to those for D′, and results for r^2_{hap}, D^* and χ^2_{df} were similar to those for r^2*

No. of marker alleles (g_0)	Population size	*(A) SNP-QTL r^2*				*(B) SNP-QTL D′*			
		D′	*r^2*	*χ^2′*	*χ^2_{tr}*	*D′*	*r^2*	*χ^2′*	*χ^2_{tr}*
2	50	244.2	0.6	0.6	31.4	11.3	230.7	230.7	106.6
	100	168.2	0.1	0.1	15.8	3.5	172.5	172.5	92.8
	150	114.4	0.0	0.0	7.6	3.3	109.7	109.7	63.9
	200	74.7	0.0	0.0	3.5	1.7	73.8	73.8	47.6
4	50	216.5	0.4	1.8	29.3	6.4	220.3	196.3	103.8
	100	127.0	0.1	3.0	13.2	6.1	164.3	131.1	93.5
	150	80.5	0.0	2.9	8.1	6.2	117.9	89.1	70.1
	200	61.8	0.0	2.3	5.5	1.9	74.4	52.9	42.0
6	50	209.3	0.5	3.0	29.6	4.5	225.8	193.8	108.1
	100	119.5	0.1	5.3	15.0	8.0	168.3	123.5	93.1
	150	88.7	0.0	5.9	11.5	3.3	112.8	72.7	57.7
	200	78.4	0.0	5.5	9.6	1.1	75.4	43.2	34.0
8	50	216.0	0.5	3.8	31.7	7.8	239.8	203.2	116.9
	100	121.7	0.1	7.4	16.8	6.4	162.2	111.4	85.2
	150	96.6	0.0	7.9	14.0	3.5	117.9	71.0	56.8
	200	89.9	0.0	7.7	12.5	2.1	72.3	35.7	27.5
10	50	204.3	0.5	3.8	28.9	6.6	226.3	189.1	111.0
	100	121.3	0.2	8.6	18.1	7.5	164.2	109.7	85.0
	150	102.7	0.0	9.4	15.7	2.3	113.1	63.4	50.2
	200	102.0	0.0	9.7	15.0	4.1	71.9	32.0	24.3

Fig. 1


(A)
pair-wise SNP-QTL LD
mean SNP-QTL LD
predicted SNP-QTL LD
SNP-QTL r2
SNP-QTL distance (cM)
0
0.2
0.4
0.6
0.8
1
0
5
10
15
20

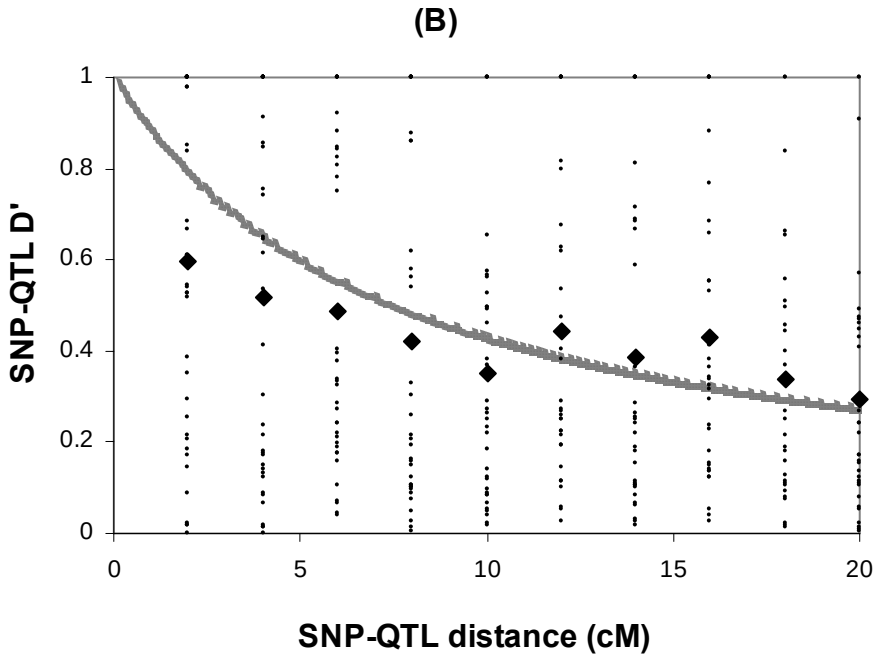

(B)
SNP-QTL D'
SNP-QTL distance (cM)
0
0.2
0.4
0.6
0.8
1
0
5
10
15
20

Fig. 1 (continued)

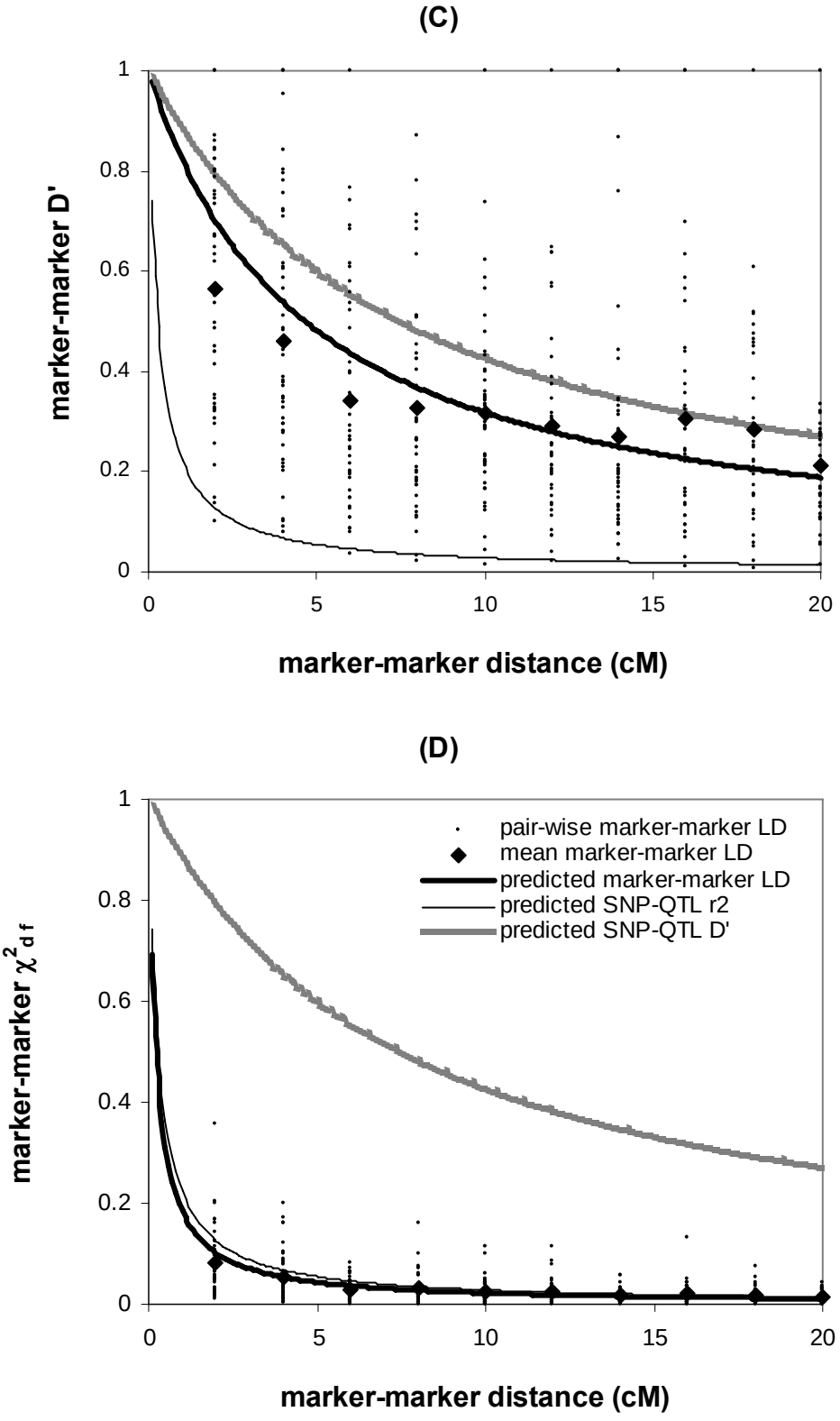


Fig. 1. Observed relationships of SNP-QTL LD measured by r^2 (A) and D' (B) with marker-marker LD measured by D' (C) and χ^2_{df} (D) against map distance for a representative replicate with a population size of 100 and 4 alleles per marker. Legend for (B) is the same as legend for (A), and legend for (C) is the same as legend for (D). LD at distance *d* Morgans was predicted from $L\hat{D}_d = 1/\left(1 + 4\hat{\beta} d\right)$, where $\hat{\beta}$ was obtained from the simulated data for each LD measure.

Fig. 2

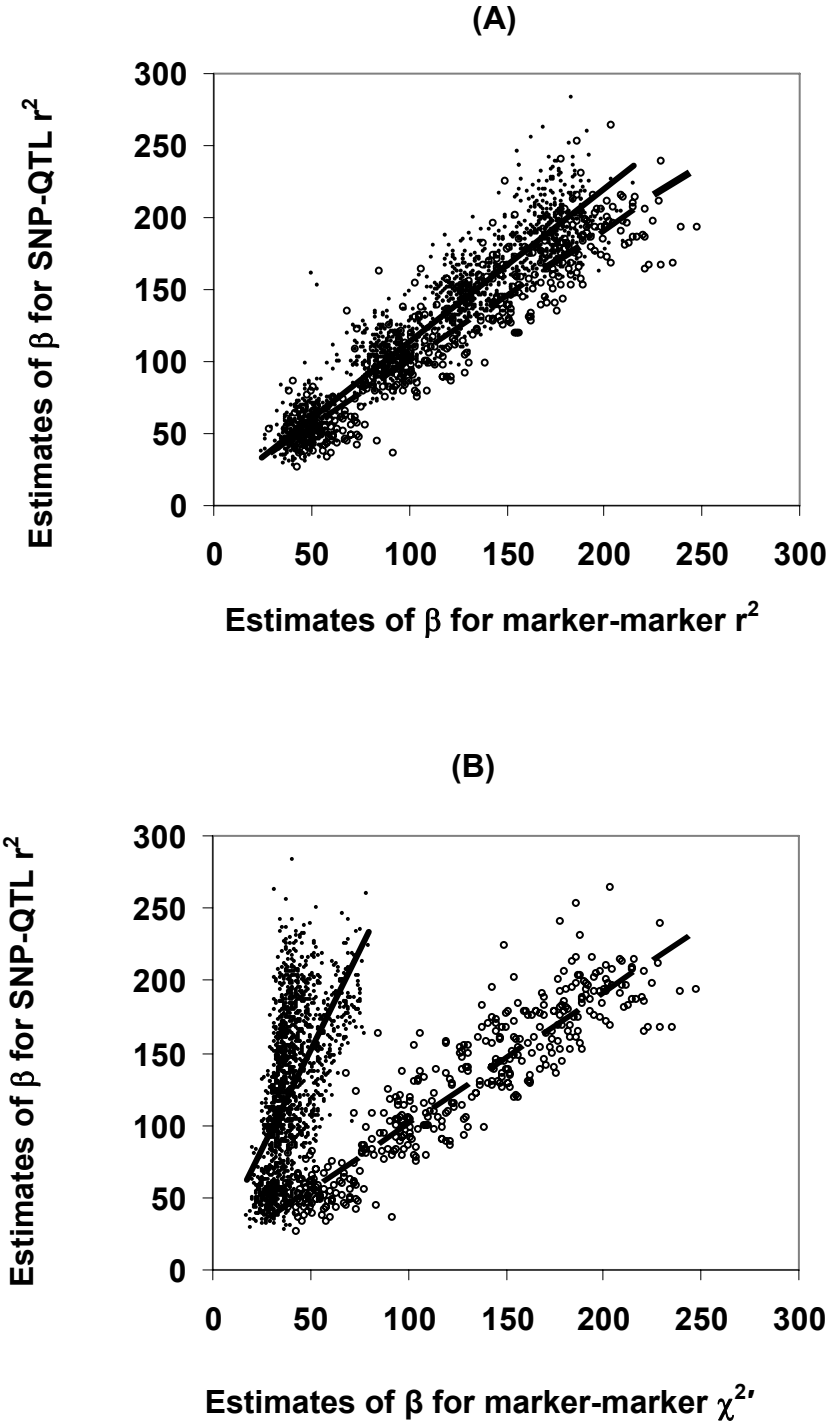

(A)
Estimates of β for SNP-QTL r²
Estimates of β for marker-marker r²
0
50
100
150
200
250
300
(B)
Estimates of β for SNP-QTL r²
Estimates of β for marker-marker χ²′

Fig. 2 (continued)

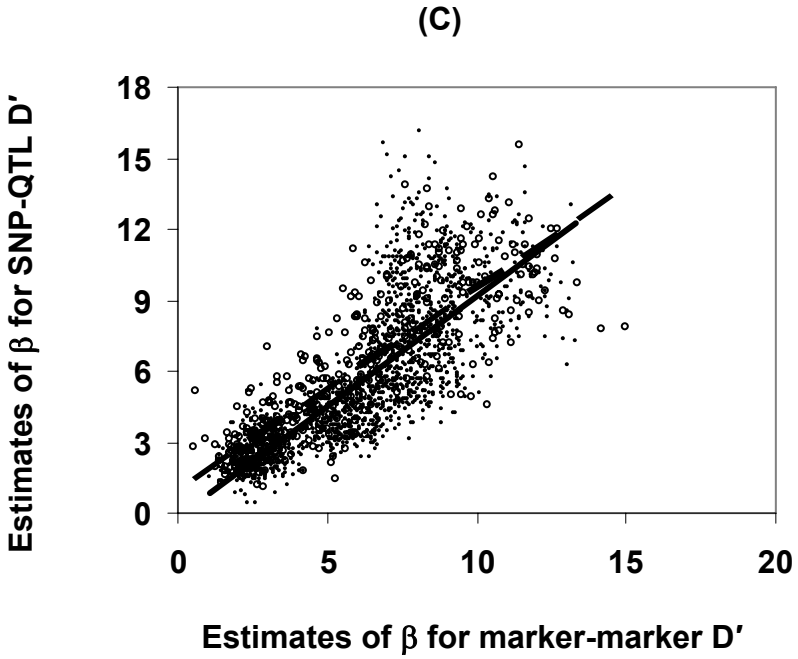


Fig. 2. Regression of estimates of the decline of LD with distance (β) obtained from each replicate for SNP-QTL r^2 (A, B) and D' (C) on estimates of β for marker-marker LD measured by r^2 (A), $\chi^2{}'$ (B) and D' (C) for biallelic (open circle) and multi-allelic (dot) markers. Dashed and solid lines indicate the regression lines for bi- and multi-allelic markers, respectively. Data are based on 100 replicates simulated for each of the 20 combinations of population size (50, 100, 150 or 200) and number of marker alleles (2, 4, 6, 8 or 10).

CHAPTER 5. POWER AND PRECISION OF ALTERNATE METHODS FOR LINKAGE DISEQUILIBRIUM MAPPING OF QTL IN LIVESTOCK

A paper to be submitted to *Genetics*

H. H. Zhao, R. L. Fernando and J. C. M. Dekkers

Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, 239D Kildee Hall, Ames, IA, 50011, USA

ABSTRACT

Linkage disequilibrium (LD) analysis in a closed outbred population uses historical recombinations and is useful to detect and fine map quantitative trait loci (QTL). The objective here was to evaluate the effect of various factors on power and precision of QTL detection using different LD mapping methods. An 11 cM region with 6 to 38 segregating single nucleotide polymorphisms (SNPs) and a central QTL was simulated. After 100 generations of random mating with effective population size (N_e) of 50, 100 or 200, SNP genotypes and phenotypes were generated on 200, 500 or 1000 individuals with effects of the biallelic QTL set to explain 2 or 5% of phenotypic variance. To detect and map the QTL, phenotypes were regressed on genotypes or haplotypes for 1, 2 or 4 linked SNPs. Empirical 1% thresholds were derived assuming no QTL effect. Based on 10,000 replicates, power to detect QTL increased with sample size, marker density and QTL effect, decreased with N_e, and was similar between methods. Sample size had a greater effect on power than density. For significant replicates, precision of QTL position estimates increased with sample size,

marker density and QTL effect, but was little affected by N_e. Single marker regression had similar or better precision than other models. Regression-based methods were compared to the identical by descent (IBD) method which modeled covariances between QTL alleles carried by alternate marker haplotypes consisting of 1, 2, 4, 6 or 8 adjacent SNPs. Among the haplotype sizes, fitting a 4-SNP haplotype in the IBD method, in general, resulted in relatively high power and the greatest mapping precision. Single marker regression was found to be comparable to the 4-SNP IBD method. The results for the haplotype regression and the IBD method assume that haplotypes are known, which would not be true in practice. This will obviously reduce power of these methods. Thus, for rapid initial screening, QTL can be detected and mapped by regression on SNP genotypes without recovering haplotypes with adequate sample size.

INTRODUCTION

Recent advances in technology, such as high density single nucleotide polymorphism (SNP) genotyping, have increased the feasibility of quantitative trait loci (QTL) detection and fine-mapping in outbred populations using historical population-wide linkage disequilibrium (LD). Goals for LD mapping include both QTL detection and fine-mapping of a previously detected QTL, although most studies and methods developed for LD mapping may only deal with one of these (ZÖLLNER and PRITCHARD 2005). LD mapping has been used extensively to identify genes for monogenic diseases in humans (PELTONEN 2000). Contrary to the situation in human, extensive LD over a long range was observed in dairy cattle, sheep and pigs (FARNIR *et al*. 2000; MCRAE *et al*. 2002; NSENGIMANA *et al*. 2004; TENESA *et al*. 2003). LD mapping in livestock might be effective using marker maps of more

limited density than what is required for most human populations because of the extensive LD that is created by drift in livestock as a result of limited effective population sizes compared to humans (FARNIR *et al*. 2002; TERWILLIGER *et al*. 1998). Because LD mapping can be implemented in outbred populations, resulting QTL can immediately be implemented for marker assisted selection (DEKKERS and HOSPITAL 2002).

Several statistical methods for LD mapping have been developed, including random effects methods based on identical by descent (IBD) (MEUWISSEN and GODDARD 2000) and least squares methods based on regression of phenotype on marker genotypes or haplotypes (GRAPES *et al*. 2004, 2006). IBD methods model covariances between individuals by deriving IBD probabilities of QTL alleles carried by alternate marker haplotypes under some assumptions about population history (MEUWISSEN and GODDARD 2000). By combining IBD-based LD mapping methods and linkage analysis, MEUWISSEN *et al*. (2002) mapped a QTL within a region <1 cM for twinning rate in large half-sib cattle families.

Regression methods for LD mapping have been shown to be competitive with IBD methods in terms of accuracy of fine-mapping within a previously identified QTL region (GRAPES *et al*. 2004). For the IBD approach, GRAPES *et al*. (2006) found that derivation of IBD based on haplotypes of 4-6 markers around the postulated QTL position resulted in greater mapping precision than IBD derived using all 10 markers, because the latter resulted in a flatter likelihood curve that did not discriminate between alternate QTL positions.

However, GRAPES *et al.* (2004 and 2006) did not compare the power of QTL detection of IBD versus regression methods. In addition, they simulated 10 or 20 SNPs evenly spaced in the base population and used them for QTL fine mapping in the final generation although some of them became fixed after 100 generations of random mating. In practice, SNPs that

are not informative will not be used for analysis. To aid in design and analysis of LD mapping studies, our objective was to use well-spaced and segregating SNPs to evaluate the effect of various factors on power and precision of QTL detection using regression- and IBD-based LD mapping methods.

METHODS

Simulations: A QTL region of 11 cM with 1000 single nucleotide polymorphisms (SNPs) with frequencies ½ and in equilibrium, and a central QTL with uniquely numbered alleles was simulated. LD was created by g (g = 50, 100, 200) generations of random mating in a population of effective size N_e (N_e = 50, 100, 200). In generation g, N (N = 200, 500, 1000) individuals were generated by randomly sampling N mating pairs and the QTL was converted to be biallelic by making the allele with frequency closest to 0.5 and between 0.3 and 0.7 the favorable allele, with an additive effect that explained x% (x = 2, 5) of phenotypic variance. The remainder was generated as a random standard normal deviate. Genotypes with known linkage phase for k (k = 6, 10, 20, 38) SNPs that were still segregating in generation g (minor allele frequency ≥ 0.2) and that were well-spaced over the QTL region were identified by k-medoids clustering (SPEED 2003) of segregating SNPs. The SNPs at the median of each cluster were used for analysis.

Regression-based LD mapping: Following GRAPES *et al.* (2004), QTL detection and fine-mapping was by regression of phenotypes in the final generation on genotypes or haplotypes of m neighboring SNPs (m = 1, 2 or 4) for each window of m SNPs within the k-SNP interval. The model of regression on genotypes for the window starting with SNP j (j =

1 to k-m+1) was: $y_i = \mu + \sum_{l=j}^{j+m-1} b_l g_{l\,i} + e_i$, where y_i = phenotype of individual i, $g_{l\,i}$ = number of copies of allele 1 at SNP l for individual i, b_l = substitution effect, and e_i = residual. For each window of m SNPs (m = 2 or 4), there are $n = 2^m$ possible haplotypes. The model of regression on haplotypes for each window was: $y_i = \mu + \sum_{l=1}^{n} b_l g_{l\,i} + e_i$, where $g_{l\,i}$ = number of copies of haplotype l for individual i, b_l = haplotype effect. The window with the most significant F-value was chosen as the best model and the center of that window as the estimate of QTL position. Significance thresholds at a 1% region-wide level were determined empirically by simulating 10,000 replicates where the QTL had no effect on phenotype (x = 0).

IBD-based LD mapping: Following MEUWISSEN and GODDARD (2000), phenotypic records in the last generation were modeled by $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zh} + \boldsymbol{e}$, where $\boldsymbol{y}$ is the vector of N records, $\boldsymbol{\beta}$ is a vector of fixed effects, which reduces to the overall mean here, $\boldsymbol{X}$ is an incidence matrix for $\boldsymbol{\beta}$, which reduces to a vector of N ones, $\boldsymbol{h}$ is a (q x 1) vector of random effects of q unique marker haplotypes present in the final generation, $\boldsymbol{Z}$ is known incidence matrix for $\boldsymbol{h}$, and $\boldsymbol{e}$ is the vector of residuals. The variance of the residuals is $\boldsymbol{R} = \mathbf{I}\sigma_e^2$, where σ_e^2 is the residual variance and $\mathbf{I}$ is an identity matrix. The variance of the haplotype effects is $\sum_{\boldsymbol{h}} = \sigma_h^2 \boldsymbol{H}_{\mathbf{p}}$, where σ_h^2 is the variance of QTL effect, and $\boldsymbol{H}_{\mathbf{p}}$ is a (q x q) matrix of the probabilities that QTL alleles at the assumed position p are IBD given a pair of marker haplotypes. This method models covariances between QTL alleles carried by alternate marker haplotypes using IBD. The IBD probability of a pair of alleles at the putative QTL position increases as the number of markers surrounding the QTL that are consecutively

identical in state increases (MEUWISSEN and GODDARD 2000). It was derived with assumptions about historical population structure, such as N_e and mutation age, using the method of MEUWISSEN and GODDARD (2001).

For each window of m neighboring SNPs (m = 1, 2, 4, 6 or 8), the full model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zh} + \boldsymbol{e}$ was fitted sequentially in the m-1 marker intervals by assuming that the QTL was at the center of each interval. The corresponding mixed model equation (MME) is

$$\begin{bmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z} \\ \boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z} + \sum_{\boldsymbol{h}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{h}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{Z'R^{-1}y} \end{bmatrix}.$$

The residual log likelihood under multivariate normality was obtained from MEYER and SMITH (1996):

$$\log L\left(\boldsymbol{H}_{\boldsymbol{p}}, \sigma_h^2, \sigma_e^2\right) \propto -0.5\left(\log|\boldsymbol{R}| + \log|\Sigma_{\boldsymbol{h}}| + \log|\boldsymbol{C}| + \boldsymbol{y'R^{-1}y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'R^{-1}y} - \hat{\boldsymbol{h}}'\boldsymbol{Z'R^{-1}y}\right)$$

where $\boldsymbol{C}$ is the coefficient matrix of the MME.

Given a QTL position, p, *i.e.*, given $\boldsymbol{H}_{\boldsymbol{p}}$, this log L was maximized to obtain estimates of the variance components $\hat{\sigma}_h^2$ and $\hat{\sigma}_e^2$ using the Newton-Raphson algorithm. The $\log L\left(\boldsymbol{H}_{\boldsymbol{p}}, \hat{\sigma}_h^2, \hat{\sigma}_e^2\right)$ was calculated for every putative QTL position and the position with the highest $\log L\left(\boldsymbol{H}_{\boldsymbol{p}}, \hat{\sigma}_h^2, \hat{\sigma}_e^2\right)$ was identified as the most likely QTL position. To test the significance of the QTL, a reduced model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ was also fitted. Under the reduced model,

$$\log L\left(\sigma_e^2\right) \propto -0.5\left(\log|\boldsymbol{R}| + \log\left|\boldsymbol{X'R^{-1}X}\right| + \boldsymbol{y'R^{-1}y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'R^{-1}y}\right)$$

and was maximized with respect to σ_e^2 using Newton-Raphson, regardless of QTL position.

In the Newton-Raphson algorithm, object oriented programming was used to implement automatic differentiation (TSUKANOV and HALL 2003) in the calculation of the first and

second derivatives of log *L*. We constrained σ_h^2 and σ_e^2 to their parameter spaces by using *xh* and *xe* in the Newton-Raphson algorithm and taking $\sigma_h^2 = \frac{e^{xh}}{1+e^{xh}} + 0.001$ and $\sigma_e^2 = 2\left(\frac{e^{xe}}{1+e^{xe}}\right) + 0.001$. A grid of starting values was used to improve convergence of the Newton-Raphson algorithm to the global maximum.

The log-likelihood ratio was the difference between the maximum of $\log L\left(\boldsymbol{H}_p, \hat{\sigma}_h^2, \hat{\sigma}_e^2\right)$ under the full model and the maximum of $\log L\left(\hat{\sigma}_e^2\right)$ under the reduced model. Significance thresholds at a 1% region-wide level were determined empirically by simulating 10,000 replicates where the QTL had no effect on phenotype ($x = 0$).

RESULTS

Power of regression methods: The impact on power to detect a QTL of sample size (N = 200, 500, 1000), SNP density (k = 6, 10, 20, 38), QTL effect (x = 2, 5%), effective population size (N_e = 50, 100, 200), number of generations since mutation (g = 50, 100, 200) and model of regression analysis is shown in Figure 1. Regardless of model of analysis, power increased with sample size, SNP density and QTL effect (Figure 1, A and B), but decreased with increasing N_e (Figure 1C), and increased a little with the number of generations since mutation (Figure 1D). In general, sample size had a greater effect on power than marker density. For example, doubling the number of genotypes by increasing density from 10 to 20 for N = 500 resulted in a smaller increase in power than doubling N from 500 to 1000 (Figure 1, A and B). This may, however, depend on the extent of LD; with N_e = 200, an increase in density had a substantial impact on power (Figure 1C).

To test the effect of QTL position on power to detect a QTL, a non-central QTL was simulated at 3.25 cM from the left end of the 11 cM chromosome region. The power was very similar for central and non-central QTL (results not shown). The effect of allele frequencies of SNPs in the base population was tested by simulating SNPs with allele frequencies randomly chosen from 0.2 to 0.8 in generation zero and the power was little affected (results not shown).

Differences between models of analysis were generally small; 1-SNP regression had very similar power as regression on 2 or 4 SNPs (Figure 1). Regression on haplotypes of 2 SNPs had similar power as genotype regression, but 4-SNP haplotype regression generally had lower power (Figure 1).

Precision of regression methods: Mapping precision was quantified as the mean absolute error of position estimates and summarized in Figure 2 for significant replicates. Similar to power, precision increased with sample size, SNP density and QTL effect (Figure 2, A and B), but was little affected by N_e (Figure 2C), and increased with the number of generations since mutation (Figure 2D). Sample size and density had similar effects on precision (Figure 2, A and B). QTL effect had less impact on precision (Figure 2, A and B) than on power (Figure 1, A and B).

Similar to power, precision was very similar for central and non-central QTL except that when there were 6 SNPs in the 11 cM region, the precision of 4-SNP genotype or haplotype regression was poorer when QTL was non-central (results not shown). Precision was little affected by allele frequencies of SNPs in the base population (results not shown).

Figure 3 shows mapping precision of regression methods for all replicates. Average precision was poorer when considering all (Figure 3, A and B) versus only significant

replicates (Figure 2, A and B). In contrast to considering significant replicates only (Figure 2), precision across all replicates decreased with increasing N_e (Figure 3C) and QTL effect had much greater effect on precision (Figure 3, A and B); other results were the same.

Single marker regression resulted in similar or better precision than all other methods, which was the same when considering all and only significant replicates (Figures 2 and 3). When considering multiple markers, regression on genotypes resulted in similar precision as regression on haplotypes (Figures 2 and 3).

Power of IBD methods: The power of IBD methods is presented in Tables 1 and 2 for QTL effects of 5 and 2% of the phenotypic variance, respectively, in comparison with regression methods. Power of IBD methods increased with SNP density and QTL effect, although this increase was not obvious when SNP density increased from 10 to 20 per 11 cM (Tables 1 and 2). Single marker regression had higher power than the 1-SNP IBD method (Tables 1 and 2). With 2 or 4 SNPs in the model, IBD resulted in similar or higher power than regression (Tables 1 and 2), except for a SNP density of 20 and a QTL effect of 5% (Table 1). In that case, regression on genotypes had better power than IBD (Table 1). For SNP densities ranging from 6 to 10 within the region, power of the IBD method increased with the number of SNPs included in the model up to 4 or 6, and IBD using more than 1 SNP had similar or higher power than single marker regression, although this pattern was not clear for a SNP density of 20 (Tables 1 and 2). IBD using 6 and 8 SNPs had similar power for a SNP density of 10 (Table 1).

Precision of IBD methods: Tables 1 and 2 also show the precision of IBD in comparison with regression methods. Precision of IBD methods increased with SNP density and QTL effect (Tables 1 and 2). Single marker regression, in general, had similar or higher

precision than the 1-SNP IBD method, which was more obvious for significant replicates (Tables 1 and 2). With 2 or 4 SNPs in the model, the IBD method was better than regression, which was true when considering all and only significant replicates (Tables 1 and 2), except for a SNP density of 6 and a QTL effect of 2% of the phenotypic variance where regression on 4-SNP haplotypes gave the best precision for all replicates (Table 2). Comparing mapping precision of IBD methods for both significant and all replicates using different haplotype sizes, 4-SNP IBD, in general, resulted in the best precision (Tables 1 and 2); IBD using 8 SNPs resulted in precision as poor as IBD using 1 SNP (Table 1). Comparing 4-SNP IBD with single marker regression, which gave the best precision among regression methods, single marker regression was in general better than 4-SNP IBD when considering only significant replicates (Tables 1 and 2); however, 4-SNP IBD had better precision when considering all replicates (Tables 1 and 2).

DISCUSSION AND CONCLUSIONS

Power of QTL detection: This study compared two methods for LD-based QTL fine mapping: regression on SNP genotypes or haplotypes and the IBD method. Regression on SNP genotypes does not require knowledge of SNP haplotypes and is therefore easier to implement. Our study showed that single marker regression provided similar or higher power than other regression-based methods for SNP densities ranging from 6 to 38 per 11 cM, while fitting haplotypes of 4 markers generally had low power (Figure 1). Part of the limited extra or lower power of multi-marker and, in particular, haplotype methods over single marker regression may be caused by the additional parameters fitted, which would be avoided when using IBD methods. Using haplotypes of more SNPs in IBD methods is expected to improve

power of QTL detection because it improves the accuracy of IBD probabilities without fitting additional parameters. Such a trend was observed for a SNP density of 10 (Tables 1 and 2). Among IBD methods using 1, 2, 4, 6 or 8 SNPs to derive IBD probabilities, the 1-SNP IBD resulted in the lowest power (Tables 1 and 2) because of the poor accuracy of IBD probabilities (GRAPES *et al*. 2006), and the power kept increasing until the number of SNPs in the IBD model reached 6 (Tables 1 and 2). However, IBD with 8 SNPs had similar power as IBD with 6 SNPs (Table 1). It appears that the power of IBD method can only be improved to a certain point by using more SNPs to derive IBD probabilities. Compared to SNPs close to the true QTL position, distant SNPs may provide less information to determine if QTL alleles are IBD. This is consistent with GRAPES *et al*. (2006), who showed that the likelihood at the true QTL position increased greatly when the haplotype size used in the IBD model increased from 1 to 4 SNPs, but only slightly increased from 4 to 10 SNPs.

Our study only used consecutive SNPs when multiple SNPs were included in a model. Because of the rather random nature of LD generated by drift, it seems reasonable to fit all possible combinations of SNPs within the chromosome region, which was used by BONNEN *et al*. (2006). This strategy did, however, not improve power of QTL detection because of more stringent significance thresholds (results not shown).

The results for the haplotype regression and IBD methods assume that haplotypes are known, which would not be true in practice. This will obviously reduce power of these methods. Although it is unclear how much the power will be reduced, this will nevertheless make the genotype regression methods look even better. Several other studies also found that single marker tests provide as much or greater power than haplotype-based test (LONG and LANGLEY 1999; NIELSEN *et al*. 2004). It appears that, at least for SNP maps with medium

density, haplotype information may not be essential for QTL detection, which is consistent with the more random pattern of LD expected from drift (ZHAO *et al*. 2005), and that rather simple regression methods can provide sufficient power to detect QTL in data of reasonable size (LONG and LANGLEY 1999).

Precision of QTL detection: Precision was evaluated for all and only significant replicates. Significant replicate results should be considered if there is no prior information on QTL position (*i.e.* if QTL detection is part of the experiment). Results for all replicates would apply when the objective is to fine-map a QTL in an already identified region, like studies in GRAPES *et al*. (2004, 2006).

Our greater mapping precision for 1- *vs* multiple-SNP regression is in contrast to GRAPES *et al*. (2004). They found that 2-SNP haplotype regression performed better at estimating the position of the QTL than single marker regression under the same marker spacings (GRAPES *et al*. 2004). They, however, simulated the QTL at the center of a SNP interval, which advantaged 2-SNP regression. Here, SNP positions varied, which resulted in an average distance of the QTL to the closest SNP of 0.52 cM for 6 SNPs and 0.27 cM for 10 SNPs, compared an average distance to the center of the flanking SNP interval of 0.41 cM for 6 SNPs and 0.28 cM for 10 SNPs, resulting in no inherent bias of using 1 vs. 2 SNPs with 10 (or more) SNPs and a slight disadvantage to the 1 SNP method with 6 SNPs. GRAPES *et al*. (2004) also found greater precision for 2-SNP haplotype than 2-SNP genotype regression, while we showed no benefit to using haplotypes, which may be due to their much larger QTL effects (x>15%).

GRAPES *et al.* (2004 and 2006) also compared regression to IBD methods and found that single marker regression was not significantly different in precision from IBD with a single

SNP (GRAPES *et al.* 2006), while we found that single marker regression had similar or higher mapping precision than 1-SNP IBD (Tables 1 and 2). Based on GRAPES *et al.* (2004 and 2006), 2-SNP haplotype regression gave similar precision as the IBD method that used all 10 SNPs in the region to determine IBD. We observed similar precision for 2-SNP haplotype regression and 8-SNP IBD (1.05 vs. 1.01 for all replicates, Table 1). GRAPES *et al.* (2004 and 2006) also showed that single marker regression (with the number of SNPs genotyped doubled) gave similar precision as the IBD method using only 4 or 6 SNPs, except when marker spacing was small (0.125 cM for 1-SNP regression and 0.25 cM for IBD); however, the IBD method using 4 or 6 SNPs had better precision than single marker regression if the same number of SNPs were genotyped in the two approaches (GRAPES *et al.* 2004, 2006), which was also observed in our study for all replicates (Tables 1 and 2). Based on our study, even with the same number of SNPs genotyped, single marker regression was in general better than 4-SNP IBD when considering only significant replicates (Tables 1 and 2).

The most interesting finding in GRAPES *et al.* (2006) was that for the IBD approach, derivation of IBD based on haplotypes of 4-6 markers around the postulated QTL position resulted in greater mapping precision than IBD derived using all 10 markers. This is in good agreement with our study. Although the QTL effect in our study (x=5%) was much smaller than that in GRAPES *et al.* (2006) (x>15%), we found that fitting a 4-SNP haplotype in the IBD method, in general, resulted in the best precision compared to other haplotype sizes (Tables 1 and 2) and IBD using 8 SNPs gave precision as poor as IBD using 1 SNP (Table 1). As explained by GRAPES *et al.* (2006), the use of 4-SNPs provides enough information to accurately derive IBD probabilities while allowing for discrimination between alternate QTL

positions. Fitting 8 SNPs in our case may reduce the sensitivity of IBD probabilities to the QTL position and therefore reduce mapping precision.

It should be noted that GRAPES *et al.* (2004 and 2006) simulated 10 or 20 SNPs evenly spaced in the base population and used them for QTL fine mapping in the final generation although some of them became fixed after 100 generations of random mating. In practice, SNPs that are not informative will not be used for analysis. By simulating 1000 SNPs initially and identifying 6-38 SNPs that were still segregating in the last generation (minor allele frequency ≥ 0.2) and that were well-spaced over the QTL region, our study reflects the real situation better.

Impact of the nature of LD on QTL detection: In our study, LD was generated by drift and mutation. To evaluate the impact of LD generated by mutation versus drift on power to detect QTL, populations described in the footnote of Table 1 were simulated. Allele Q was either simulated to be unique in the base population, representing complete LD, or with frequency ½ and in linkage equilibrium (LE). QTL detection by single SNP regression showed limited difference in power between the LE and LD scenarios (0.7 for LE vs. 0.77 for LD with 10 SNPs and 0.82 vs. 0.85 with 20 SNPs). The average absolute error of QTL position was also only slightly increased for LE (0.88 cM for LE vs. 0.79 cM for LD with 10 SNPs and 0.73 vs. 0.64 with 20 SNPs). With LE, all Q alleles traced back to a single ancestral allele, which makes it equivalent to LD, for only 21% of all and 24% of significant replicates. The number of common ancestors of the Q allele was 2, 3 and 4 or greater for 56, 33 and 11% of the other replicates, for which the most frequent common ancestor accounted for only 67% of all Q alleles. These results demonstrate that mutation is not essential for sufficient LD to detect QTL and that QTL can be detected even if substantial heterogeneity

exists with regard to ancestral origin of the Q alleles. ABDALLAH *et al*. (2003) found that power to detect QTL by single marker regression was even greater with LE than LD when using multi-allelic markers and similar when using SNPs.

Conclusions: With adequate sample size, and levels of LD expected based on limited N_e, most livestock populations lend themselves to QTL detection by LD with SNPs at medium density (1-2/cM). Because of the rather random nature of LD generated by drift and when using marker maps of limited density, use of haplotype information may not increase power to detect QTL. For rapid initial screening, QTL can be detected and mapped by regression on SNP genotypes without recovering haplotypes. In addition to computational speed, regression offers flexibility to include dominance and epistatic effects. To account for relationships, a random polygenic effect should be added.

ACKNOWLEDGEMENTS

We thank Laura Grapes for her previous work and Long Qu for his valuable advice and discussion. This work was funded by Monsanto Co. and Genus Plc.

LITERATURE CITED

ABDALLAH, J. M., B. GOFFINET, C. CIERCO-AYROLLES and M. PÉREZ-ENCISO, 2003 Linkage disequilibrium fine mapping of quantitative trait loci: a simulation study. Genet. Sel. Evol. **35**: 513-532.

BONNEN, P. E., I. PE'ER, R. M. PLENGE, J. SALIT, J. K. LOWE *et al.*, 2006 Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. Nat. Genet. **38**: 214-217.

DEKKERS, J. C. M., and F. HOSPITAL, 2002 The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. **3**: 22-32.

FARNIR, F., W. COPPIETERS, J.-J. ARRANZ, P. BERZI, N. CAMBISANO *et al*., 2000 Extensive genome-wide linkage disequilibrium in cattle. Genome Res. **10**: 220-227.

FARNIR, F., B. GRISART, W. COPPIETERS, J. RIQUET, P. BERZI *et al*., 2002 Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics **161**: 275-287.

GRAPES, L., J. C. M. DEKKERS, M. F. ROTHSCHILD and R. L. FERNANDO, 2004 Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. Genetics **166**:1561-1570.

GRAPES, L., M. Z. FIRAT, J. C. M. DEKKERS, M. F. ROTHSCHILD and R. L. FERNANDO, 2006 Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. Genetics **172**: 1955-1965.

LONG, A. D., and C. H. LANGLEY, 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. **9**: 720-731.

MCRAE, A. F., J. C. MCEWAN, K. G. DODDS, T. WILSON, A. M. CRAWFORD and J. SLATE, 2002 Linkage disequilibrium in domestic sheep. Genetics **160**: 1113-1122.

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

MEUWISSEN, T. H. E., and M. E. GODDARD, 2001 Prediction of identity by descent probabilities from marker-haplotypes. Genet. Sel. Evol. **33**: 605-634.

MEUWISSEN, T. H. E., A. KARLSEN, S. LIEN, I. OLSAKER and M. E. GODDARD, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics **161**: 373-379.

MEYER, K., and S. P. SMITH, 1996 Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. Genet. Sel. Evol. **28**: 23-49.

NIELSEN, D. M., M. G. EHM, D. V. ZAYKIN and B. S. WEIR, 2004 Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics **168**: 1029-1040.

NSENGIMANA, J., P. BARET, C. S. HALEY and P. M. VISSCHER, 2004 Linkage disequilibrium in the domesticated pig. Genetics **166**: 1395-1404.

PELTONEN, L., 2000 Positional cloning of disease genes: advantages of genetic isolates. Hum. Hered. **50**: 66-75.

SPEED, T., 2003 Statistical analysis of gene expression microarray data, pp. 172-173, Chapman and Hall/CRC.

TENESA, A., S. A. KNOTT, D. WARD, D. SMITH, J. L. WILLIAMS and P. M. VISSCHER, 2003 Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. J. Anim. Sci. **81**: 617-623.

TERWILLIGER, J. D., S. ZÖLLNER, M. LAAN and S. PÄÄBO, 1998 Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. Hum. Hered. **48**: 138-154.

TSUKANOV, I., and M. HALL, 2003 Data structure and algorithms for fast automatic differentiation. Int. J. Numer. Meth. Engng. **56**: 1949-1972.

ZHAO, H., D. NETTLETON, M. SOLLER and J. C. M. DEKKERS, 2005 Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genet. Res. **86**: 77-87.

ZÖLLNER, S., and J. K. PRITCHARD, 2005 Coalescent-based association mapping and fine mapping of complex trait loci. Genetics **169**: 1071-1092.

TABLE 1

Comparison of regression-based LD mapping methods with identical by descent (IBD) methods when the QTL explains 5% of the phenotypic variance

No. SNPs included in model	Marker density (no. SNPs in 11 cM region)								
	6			10			20		
	Geno	Haplo	IBD	Geno	Haplo	IBD	Geno	Haplo	IBD
	Power to detect QTL (%)								
1	67	-	48	77	-	52	85	-	53
2	69	69	69	78	79	78	85	84	81
4	68	59	76	79	70	82	84	74	77
6	-	-	75	-	-	85	-	-	83
8	-	-	-	-	-	84	-	-	-
	Mean absolute error of position (cM) for significant QTL								
1	1.05	-	1.14	0.79	-	0.90	0.64	-	0.81
2	1.20	1.21	1.17	0.92	0.92	0.87	0.67	0.64	0.64
4	1.38	1.34	1.11	1.31	1.30	0.85	0.88	0.91	0.62
6	-	-	1.21	-	-	0.86	-	-	0.62
8	-	-	-	-	-	0.93	-	-	-
	Mean absolute error of position (cM) for all QTL								
1	1.34	-	1.33	0.96	-	1.00	0.74	-	0.82
2	1.40	1.36	1.32	1.05	1.05	0.98	0.76	0.74	0.69
4	1.38	1.36	1.22	1.37	1.37	0.93	0.94	1.01	0.66
6	-	-	1.31	-	-	0.93	-	-	0.66
8	-	-	-	-	-	1.01	-	-	-

Power (detection at 1% region-wise level) and precision for each LD mapping method (Geno, regression on genotypes at 1, 2 or 4 adjacent SNPs; Haplo, regression on assumed known haplotypes of 2 or 4 adjacent SNPs; IBD, identical by descent methods using single SNP genotype or assumed known haplotypes of 2, 4, 6 or 8 adjacent SNPs) is shown. The other parameters are $N_e = 100$, no. of generations since mutation = 100 and sample size in generation 100 = 500. SNPs were simulated with allele frequency of 0.5 and in linkage equilibrium in the base population and QTL at the center of the 11 cM region. Results are based on 10,000 replicates.

TABLE 2

Comparison of regression-based LD mapping methods with identical by descent (IBD) methods when the QTL explains 2% of the phenotypic variance

No. SNPs included in model	Marker density (no. SNPs in 11 cM region)								
	6			10			20		
	Geno	Haplo	IBD	Geno	Haplo	IBD	Geno	Haplo	IBD
	Power to detect QTL (%)								
1	26	-	18	31	-	21	34	-	22
2	25	23	25	28	27	30	31	28	34
4	24	15	28	28	18	32	30	19	31
6	-	-	27	-	-	34	-	-	32
	Mean absolute error of position (cM) for significant QTL								
1	1.13	-	1.26	0.93	-	1.16	0.85	-	1.03
2	1.33	1.31	1.27	1.10	1.13	1.06	0.96	0.94	0.95
4	1.39	1.36	1.23	1.42	1.48	1.06	1.15	1.25	0.99
6	-	-	1.36	-	-	1.10	-	-	0.96
	Mean absolute error of position (cM) for all QTL								
1	1.71	-	1.67	1.41	-	1.45	1.28	-	1.28
2	1.71	1.69	1.64	1.51	1.55	1.44	1.37	1.41	1.25
4	1.41	1.38	1.54	1.64	1.66	1.38	1.50	1.64	1.27
6	-	-	1.69	-	-	1.41	-	-	1.25

Power (detection at 1% region-wise level) and precision for each LD mapping method (Geno, regression on genotypes at 1, 2 or 4 adjacent SNPs; Haplo, regression on assumed known haplotypes of 2 or 4 adjacent SNPs; IBD, identical by descent methods using single SNP genotype or assumed known haplotypes of 2, 4 or 6 adjacent SNPs) is shown. The other parameters are $N_e = 100$, no. of generations since mutation = 100 and sample size in generation 100 = 500. SNPs were simulated with allele frequency of 0.5 and in linkage equilibrium in the base population and QTL at the center of the 11 cM region. Results are based on 10,000 replicates.

FIGURE 1

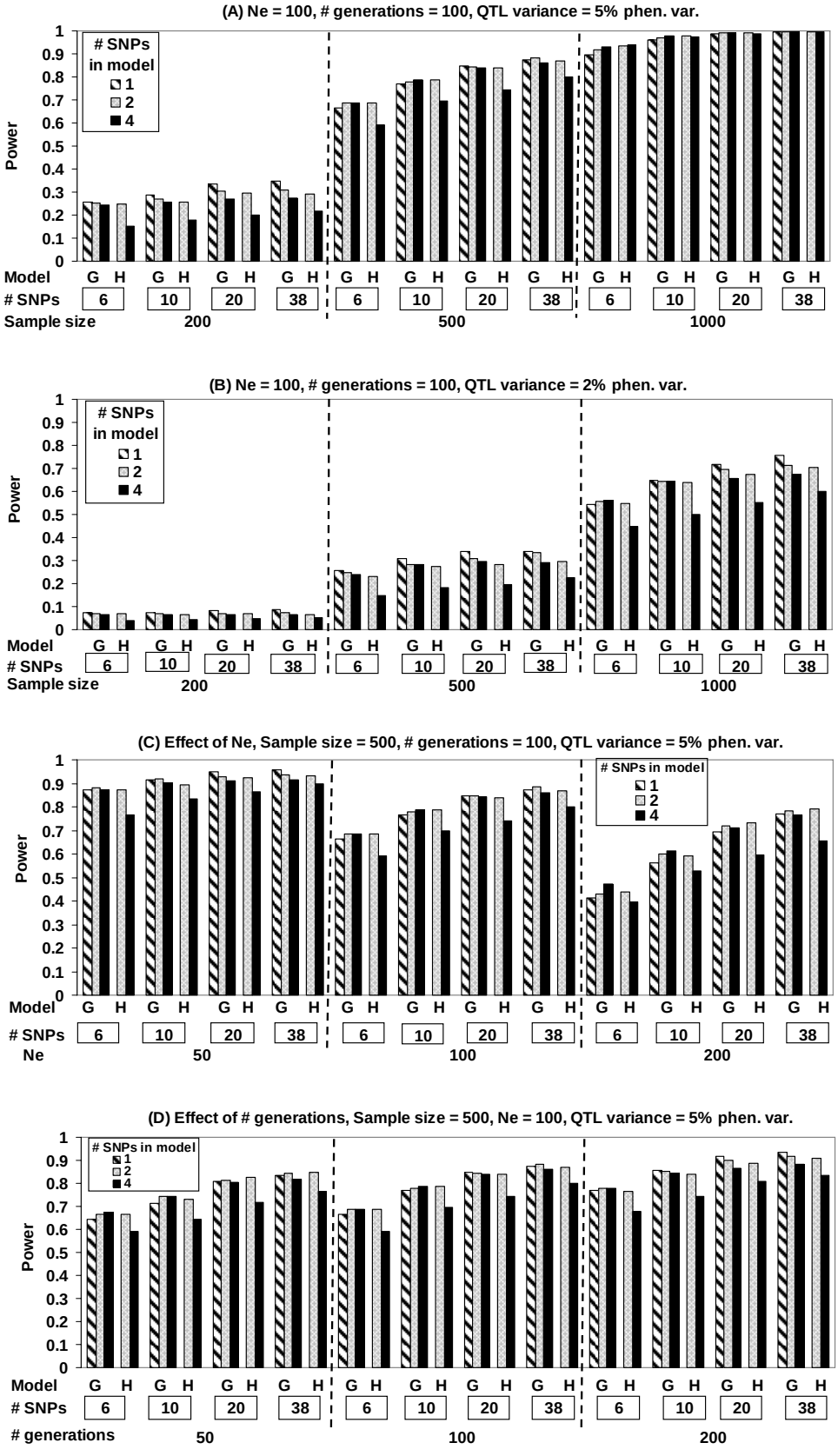

(A) Ne = 100, # generations = 100, QTL variance = 5% phen. var.
SNPs in model
1
2
4
Power
Model
SNPs
Sample size
200
500
1000
(B) Ne = 100, # generations = 100, QTL variance = 2% phen. var.
(C) Effect of Ne, Sample size = 500, # generations = 100, QTL variance = 5% phen. var.
Ne
50
100
200
(D) Effect of # generations, Sample size = 500, Ne = 100, QTL variance = 5% phen. var.
generations

FIGURE 1.—Effects of sample size, marker density (# SNPs), QTL effect, effective population size (Ne), no. of generations since mutation (# generations) and model of analysis (regression on genotype (G) for 1, 2 or 4 SNPs or on haplotype (H) for 2 or 4 SNPs) on power to detect QTL. Based on 10,000 replicates.

FIGURE 2

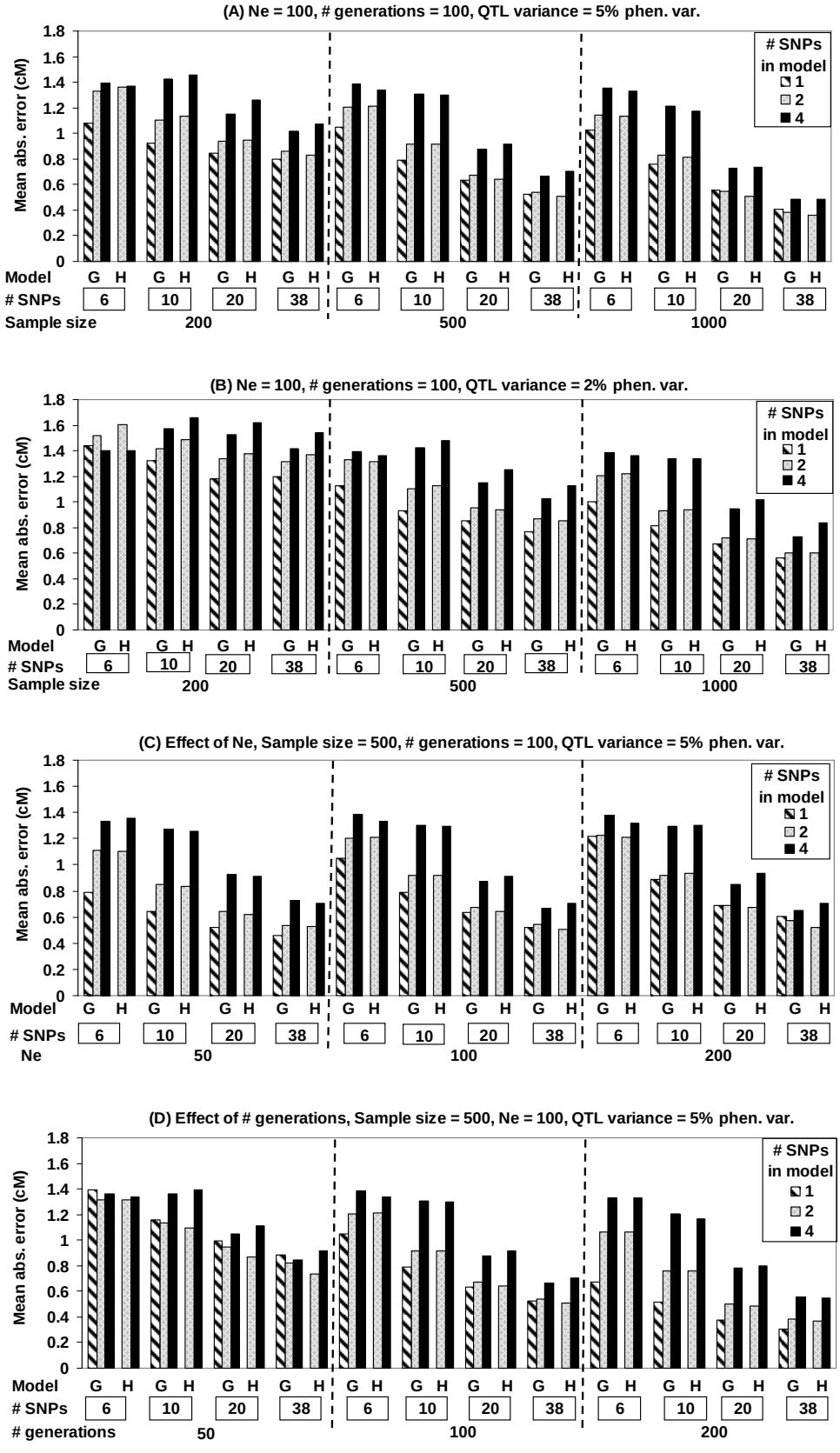

(A) Ne = 100, # generations = 100, QTL variance = 5% phen. var.
Mean abs. error (cM)
SNPs in model
1
2
4
Model
SNPs
Sample size
200
500
1000
(B) Ne = 100, # generations = 100, QTL variance = 2% phen. var.
(C) Effect of Ne, Sample size = 500, # generations = 100, QTL variance = 5% phen. var.
Ne
50
100
200
(D) Effect of # generations, Sample size = 500, Ne = 100, QTL variance = 5% phen. var.
generations

FIGURE 2.—Effects of sample size, marker density (# SNPs), QTL effect, effective population size (Ne), no. of generations since mutation (# generations) and model of analysis (regression on genotype (G) for 1, 2 or 4 SNPs or on haplotype (H) for 2 or 4 SNPs) on precision of estimates of position for significant QTL. Based on 10,000 replicates.

FIGURE 3

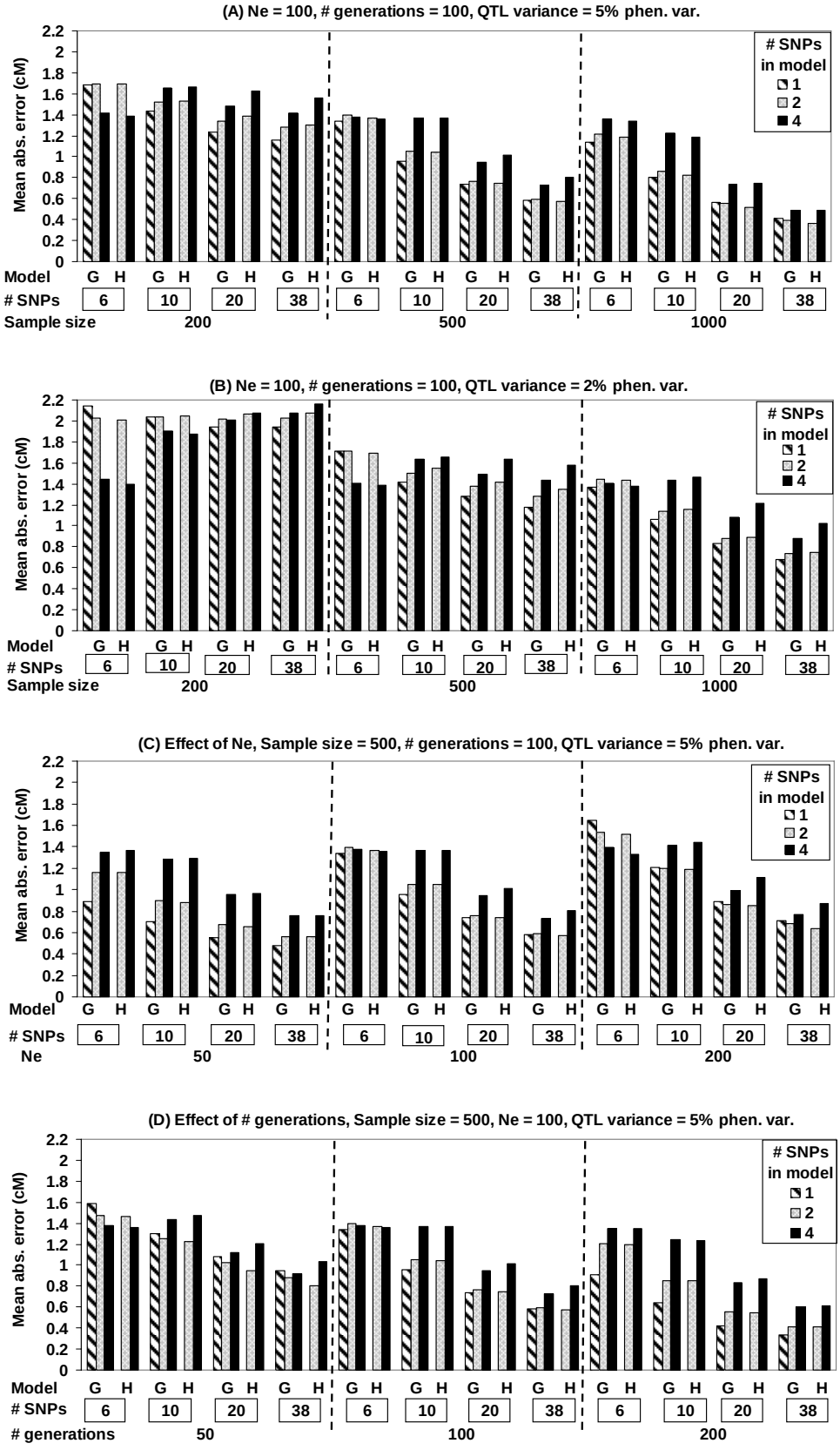

(A) Ne = 100, # generations = 100, QTL variance = 5% phen. var.
Mean abs. error (cM)
SNPs in model
1
2
4
Model
SNPs
Sample size
200
500
1000
(B) Ne = 100, # generations = 100, QTL variance = 2% phen. var.
(C) Effect of Ne, Sample size = 500, # generations = 100, QTL variance = 5% phen. var.
Ne
50
100
200
(D) Effect of # generations, Sample size = 500, Ne = 100, QTL variance = 5% phen. var.
generations

FIGURE 3.—Effects of sample size, marker density (# SNPs), QTL effect, effective population size (Ne), no. of generations since mutation (# generations) and model of analysis (regression on genotype (G) for 1, 2 or 4 SNPs or on haplotype (H) for 2 or 4 SNPs) on precision of estimates of position for all QTL. Based on 10,000 replicates.

CHAPTER 6. GENERAL CONCLUSIONS AND DISCUSSION

GENERAL CONCLUSIONS

The research conducted in this thesis addresses several important issues in QTL fine mapping using candidate gene analysis and LD mapping using high density genotyping. In Chapter 2, we developed and evaluated three tests for candidate genes in F2 resource populations for QTL mapping. The findings of this work were:

- Use of the standard association test for candidate genes based on the fixed effect of candidate gene genotype in F2 populations can result in significant effects for candidate genes that are at considerable distances from the QTL because of the extensive between-breed LD that exists in these populations.
- A marker-assisted association test was developed that was based on a test at the candidate gene of the fixed effect of candidate gene genotype in a breed-cross QTL interval mapping model. This test removed the impact of between-breed LD, but was not powerful in detecting candidate genes closely linked to the QTL, unless the candidate gene was the QTL.
- An F-drop test that compares F ratios for a QTL at the candidate gene with and without the candidate gene included as fixed effect had low power to distinguish close from distant candidate genes.
- Power to distinguish two candidate genes within 10 cM from the QTL was limited and little improved by including QTL effects associated with markers to remove between-breed LD, although power was greater when one of the candidate genes was the causative mutation.

- Candidate gene tests in QTL mapping populations must be interpreted with caution. Given the extensive between-breed LD that exists in intercrosses of outbred breeds or lines, candidate gene studies in farm animals cannot rely solely on breed-cross populations and effects uncovered in these crosses need to be confirmed in one or more closed mating populations.

Chapters 3 and 4 evaluated nine measures of LD between multi-allelic markers as predictors of usable LD of markers (microsatellite or SNPs) with QTL in outbred populations when LD is generated by drift. Findings from these studies were:

- Usable LD can be quantified based on evaluation of LD between multi-allelic or biallelic markers. The LD measure between markers that is best for predicting usable LD in a population depends on the type of markers that will eventually be used for QTL mapping or MAS (*i.e.* multi-allelic or biallelic).
- The $\chi^{2'}$ measure developed by Yamazaki (1977) based on multi-allelic markers is recommended to quantify the amount and extent of usable LD for the same group of markers in a population for QTL mapping and MAS.
- $\chi^{2'}$ based on multi-allelic markers does not give a good assessment of usable LD when biallelic SNPs are used for QTL mapping or MAS. It over-estimates usable LD between SNP and QTL.
- χ^2_{df}, r^2 and D^* based on multi-allelic markers are good predictors of usable SNP-QTL LD.

- D' and D_{hap} give rise to LD estimates that are strongly inflated relative to usable marker-QTL LD. However, D' and D_{hap} between multi-allelic markers can predict SNP-QTL and SNP-SNP LD based on D'.
- For biallelic markers, the decline in LD measured by r^2 estimates effective population size. For multi-allelic markers, r^2, $D*$ and χ^2_{df} all provide good estimates of effective population size.
- For a given population and when quantified against map distance rather than physical distance, LD between markers and QTL will not necessarily be greater in marker intervals that show strong LD between markers. Therefore, although marker-marker LD enables assessment of the general extent of usable LD in populations, high marker-marker LD in specific regions may not necessarily identify genomic regions with high marker-QTL LD.

Chapter 5 evaluated the effect of various factors on power and precision of QTL detection and compared power and precision of regression- and IBD-based LD mapping methods. Findings from this study were:

- LD mapping in livestock is effective using marker maps of more limited density than what is required for most human populations because of the extensive LD that is created by drift in livestock as a result of limited effective population sizes compared to humans.
- For regression methods, power to detect QTL increased with sample size, marker density and QTL effect, decreased with N_e, and increased a little with the number of

generations since mutation. Sample size had a greater effect on power than marker density.

- For regression methods, precision of QTL position estimates increased with sample size, marker density, QTL effect and mutation age, which was true when considering all and only significant replicates. Precision was little affected by N_e for significant replicates, but decreased with N_e across all replicates.
- Power and precision of IBD methods increased with SNP density and QTL effect.
- Single marker regression had similar or greater power and precision than other regression models.
- For IBD methods, fitting a 4-SNP haplotype, in general, resulted in relatively high power and the greatest mapping precision among the haplotype sizes.
- Single marker regression was comparable to the 4-SNP IBD method.
- The results for the haplotype regression and the IBD method assume that haplotypes are known, which would not be true in practice. This will obviously reduce power of these methods. Thus, for rapid initial screening, QTL can be detected and mapped by regression on SNP genotypes without recovering haplotypes with adequate sample size.

GENERAL DISCUSSION

In the literature, different strategies have been developed and integrated to identify genes underlying traits of economic importance in livestock (Andersson and Georges 2004). Least squares interval mapping in breed crosses and co-segregation analysis in ourbred populations, which rely on family data, are appropriate for low resolution genetic mapping.

These approaches can be conducted as a first step in QTL mapping to localize trait loci to broad chromosome regions using sparse marker maps. LD mapping in outbred populations, which rely on both family and population data, uses historical recombinations for high resolution genetic mapping. It can be used as a genome scan to detect QTL or as a follow-up to fine-map a QTL in an already identified region using high density marker maps. The refined region can be further examined by candidate gene analysis in order to find markers that are in or close to genes that are thought to be associated with the trait of interest.

The research conducted in this thesis focuses on fine mapping of QTL using candidate gene analysis and LD mapping. The findings have some important practical implications. First, the research in Chapter 2 provides a clear assessment of the value of candidate gene tests in breed cross populations developed for QTL mapping. As many candidate gene analyses have been conducted in F2 crosses because of the wealth of phenotypic and genotypic data (Ciobanu *et al.* 2001; Li *et al*. 2003; Nguyen *et al*. 2003; Yu *et al*. 1995; Zhou *et al*. 2001), it is important to keep in mind that the extensive between-breed LD that is created in the cross may result in significant associations for candidate genes at considerable distance from the QTL. Therefore, significant associations found in these crosses must be interpreted with caution and need to be confirmed in one or more closed mating populations. Kim *et al*. (2005) shows the usefulness of F2 populations to detect and characterize QTL. In their research, least squares line-cross and half-sib models were combined to analyze data from an F2 cross of two breeds of pigs (Kim *et al*. 2005). The combined model was shown to be able to increase power and precision of QTL detection and to characterize QTL that segregate within the parental breeds (Kim *et al*. 2005).

Second, studies in Chapters 3 and 4 identified LD measures between multi-allelic markers that are appropriate for predicting the extent of usable LD in a population for QTL mapping and MAS. In recent years, several studies have measured the extent of LD in livestock populations using D' or r2 between microsatellite markers without justifying their ability of predicting usable LD in those populations (Farnir et al. 2000; Harmegnies et al. 2006; McRae et al. 2002; Nsengimana et al. 2004; Tenesa et al. 2003). The research presented in Chapters 3 and 4 shows that the LD measure between markers that is best for predicting usable LD in a population depends on the type of markers that will eventually be used for QTL mapping or MAS (i.e. multi-allelic or biallelic). This research has important implications for more accurate prediction of the extent of LD between markers and QTL, which is needed to determine the marker density and impacts the power and resolution of LD mapping and effectiveness of MAS. Heifetz *et al*. (2005) applied this research to commercial chicken populations. Using the $\chi^{2}{}'$ measure which is shown in Chapter 3 to be good for predicting the extent of usable LD in a population based on the same group of multi-allelic markers, Heifetz *et al*. (2005) found extensive LD among markers within 5 cM in commercial chicken populations. This short-range LD declined rapidly with distance, differed among chromosome regions and was strongly conserved across generations (Heifetz *et al*. 2005).

Third, the research in Chapter 5 shows that, using SNP maps with medium density (1-2/cM), single marker regression can provide sufficient power and precision to detect and fine map QTL in outbred populations of reasonable size, and haplotype information may not be essential. These results can make LD mapping simple and computationally fast to implement in industry, especially when the scan is used as a first screen for QTL.

Further work is needed to fully explore the potential of LD mapping using high density genotyping. Current research in Chapter 5 is based on a random sample of unrelated individuals. Genetic relationships can be accounted for by including a polygenic effect in a mixed model (Dekkers *et al*. 2006; Goddard and Meuwissen 2005). When pedigree information is available, a combined linkage and LD analysis can be conducted (Blott *et al*. 2003; Farnir *et al.* 2002; Meuwissen *et al*. 2002; Olsen *et al*. 2004). Evidence for a QTL is declared only if linkage and LD results are consistent (Dekkers *et al*. 2006; Goddard and Meuwissen 2005). This combined approach can increase the power and precision of QTL mapping and avoid false associations for markers that are not linked to QTL (Dekkers *et al*. 2006; Goddard and Meuwissen 2005). Using this approach, Meuwissen *et al*. (2002) mapped a QTL within a region <1 cM for twinning rate in large half-sib cattle families.

In conclusion, alternate QTL mapping methods can be integrated in livestock to increase power and mapping accuracy. In F2 populations, the combined least squares line-cross and half-sib model can characterize QTL that segregate within breeds. In outbred populations, the combined linkage and LD analysis using candidate gene markers or high density genotyping is promising for QTL fine mapping, and would result in markers that can immediately be implemented for MAS.

REFERENCES

Andersson, L. and Georges, M. (2004) Domestic-animal genomics: deciphering the genetics of complex traits. Nature Reviews: Genetics 5: 202-212.

Blott, S., Kim, J.-J., Moisio, S., Schmidt-Küntzel, A., Cornet, A., et al. (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163: 253-266.

Ciobanu, D., Bastiaansen, J., Malek, M., Helm, J., Woollard, J., Plastow, G. and Rothschild, M. F. (2001) Evidence for new alleles in the protein kinase adenosine monophosphate-activated γ3-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality. Genetics 159: 1151-1162.

Dekkers, J. C. M., Zhao, H. H. and Fernando, R. L. (2006) Linkage disequilibrium mapping of QTL in livestock. Proceedings of the World Cong. Genet. Appl. Livest. Prod. 8, Brazil (Accepted).

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 Research 10: 220-227.

Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P. et al. (2002) Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics 161: 275-287.

Goddard, M. E. and Meuwissen, T. H. E. (2005) The use of linkage disequilibrium to map quantitative trait loci. Australian Journal of Experimental Agriculture 45: 837-845.

Harmegnies, N., Farnir, F., Davin, F., Buys, N., Georges, M. and Coppieters, W. (2006) Measuring the extent of linkage disequilibrium in commercial pig populations. Animal Genetics 37: 225-231.

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

Kim, J.-J., Zhao, H., Thomsen, H., Rothschild, M. F. and Dekkers, J. C. M. (2005) Combined line-cross and half-sib QTL analysis of crosses between outbred lines. Genet. Res. 85: 235-248.

Li, H., Deeb, N., Zhou, H., Mitchell, A. D., Ashwell, C. M. and Lamont, S. J. (2003) Chicken quantitative trait loci for growth and body composition associated with transforming growth factor-β genes. Poult. Sci. 82: 347-356.

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., Karlsen, A., Lien, S., Olsaker, I. and Goddard, M. E. (2002) Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161: 373-379.

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

Nguyen, N. T., Kim, K. S., Thomsen, H., Helm, J. and Rothschild, M. F. (2003) Investigation of a candidate gene for growth and fatness QTL on the pig chromosome 7. Proceedings of Plant and Animal and Genome XI conference, San Diego, CA, p230.

Olsen, H. G., Lien, S., Svendsen, M., Nilsen, H., Roseth, A., Aasland Opsal, M. and Meuwissen, T. H. E. (2004) Fine mapping of milk production QTL on BTA6 by combined linkage and linkage disequilibrium analysis. J. Dairy Sci. 87: 690-698.

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. Journal of Animal Science 81: 617-623.

Yamazaki, T. (1977) The effects of overdominance on linkage in a multilocus system. Genetics 86: 227-236.

Yu, T.-P., Tuggle, C. K., Schmitz, C. B. and Rothschild, M. F. (1995) Association of PIT1 polymorphisms with growth and carcass traits in pigs. J. Anim. Sci. 73: 1282-1288.

Zhou, H., Buitenhuis, A. J., Weigend, S. and Lamont, S. J. (2001) Candidate gene promoter polymorphisms and antibody response kinetics in chickens: interferon-γ, interleukin-2, and immunoglobulin light chain. Poultry Science 80: 1679-1689.