Molecular, bioinformatic and statistical approaches to identify genes underlying complex traits in livestock

Laura Grapes
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Genetics Commons

Recommended Citation
Grapes, Laura, "Molecular, bioinformatic and statistical approaches to identify genes underlying complex traits in livestock " (2004). Retrospective Theses and Dissertations. 782.
https://lib.dr.iastate.edu/rtd/782

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Molecular, bioinformatic and statistical approaches to identify genes underlying complex traits in livestock

by

Laura Grapes

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Rohan L. Fernando, Co-major Professor
Max F. Rothschild, Co-major Professor
Jack C. M. Dekkers
Susan J. Lamont
Daniel S. Nettleton

Iowa State University
Ames, Iowa

2004

Copyright © Laura Grapes, 2004. All rights reserved.
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
This is to certify that the doctoral dissertation of

Laura Grapes

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Committee Member
Signature was redacted for privacy.

Committee Member
Signature was redacted for privacy.

Committee Member
Signature was redacted for privacy.

Co-major Professor
Signature was redacted for privacy.

Co-major Professor
Signature was redacted for privacy.

For the Major Program
# TABLE OF CONTENTS

**LIST OF FIGURES**

**LIST OF TABLES**

**ACKNOWLEDGEMENTS**

**ABSTRACT**

**CHAPTER 1. GENERAL INTRODUCTION**

**INTRODUCTION** 1

**RESEARCH OBJECTIVES** 5

**THESIS ORGANIZATION** 6

**LITERATURE REVIEW** 8

- Fine mapping QTL using linkage disequilibrium 9
- The candidate gene approach to identifying QTL 12
- Identification of candidate genes for reproductive traits in pigs 13
- *In silico* identification of porcine SNPs and their relationship with human SNP information 16

**REFERENCES** 18

**CHAPTER 2. COMPARING LINKAGE DISEQUILIBRIUM-BASED METHODS FOR FINE MAPPING QUANTITATIVE TRAIT LOCI**

**ABSTRACT** 31

**INTRODUCTION** 32

**METHODS** 33

- Population simulations 33
CHAPTER 6. PROSPECTING FOR PIG SNPs IN THE HUMAN GENOME: HAVE WE STRUCK GOLD?

ABSTRACT

INTRODUCTION

RESULTS

In silicio porcine SNP identification and EST annotation

Comparison of human and pig cSNP density

Comparison of human and mouse cSNP density

DISCUSSION

SUPPLEMENTARY METHODS

EST clustering and annotation

In silicio SNP detection

Correlation of cSNP density

REFERENCES

SUPPLEMENTARY RESULTS

CHAPTER 7. GENERAL CONCLUSIONS AND DISCUSSION

GENERAL CONCLUSIONS

GENERAL DISCUSSION

RECOMMENDATIONS FOR FUTURE RESEARCH

Fine mapping methodology

SNP identification in livestock

REFERENCES
LIST OF FIGURES

CHAPTER 3.  OPTIMAL HAPLOTYPE STRUCTURE FOR LINKAGE DISEQUILIBRIUM-BASED FINE MAPPING OF QUANTITATIVE TRAIT LOCI

FIGURE 1. – The correlation between IBD probabilities for the true QTL position and all other putative QTL positions estimated using haplotypes of one, two, four or ten markers. 87

FIGURE 2. – The loglikelihood for each putative QTL position averaged across 1,000 replicates of the default scenario with marker spacing equal to 1 cM. 88

CHAPTER 4.  PHYSICAL AND LINKAGE MAPPING OF THE PORCINE CONNEXIN 37 (CX37) GENE

FIGURE 1. – MluI PCR-RFLP of the connexin 37 gene. Lane M includes the 1-kb ladder with predicted sizes indicated on the left. Lane 1 includes undigested PCR product using the F2 and R1 primers. 94

CHAPTER 6.  PROSPECTING FOR PIG SNPs IN THE HUMAN GENOME: HAVE WE STRUCK GOLD?

FIGURE 1. – Correlation between porcine in silico cSNP density (# cSNPs per base of coding sequence) and human validated cSNP density (# validated cSNPs per base of coding sequence). 108

FIGURE 2. – Correlation between human and mouse validated cSNP density (# validated cSNPs per base of coding sequence). 109
LIST OF TABLES

CHAPTER 2. COMPARING LINKAGE DISEQUILIBRIUM-BASED METHODS FOR FINE MAPPING QUANTITATIVE TRAIT LOCI

TABLE 1. – Parameters for default and alternative simulated populations 55

TABLE 2. – Least squares mean absolute difference (cM) of QTL position estimates for four mapping methods using 10 or 20 markers under the default scenario 58

TABLE 3. – Least squares mean absolute difference (cM) of QTL position estimate for mapping methods with 1 cM marker spacing in a two-breed cross followed by random mating 59

TABLE 4. – Least squares mean absolute difference (cM) of QTL position estimate and bias (cM) for mapping methods in three alternate scenarios 60

CHAPTER 3. OPTIMAL HAPLOTYPE STRUCTURE FOR LINKAGE DISEQUILIBRIUM-BASED FINE MAPPING OF QUANTITATIVE TRAIT LOCI

TABLE 1. – Default and alternative parameters for the simulated populations 81

TABLE 2. – Least squares mean absolute difference (cM) of QTL position estimates obtained by the IBD method using different haplotype sizes under the default scenario 84

TABLE 3. – Least squares mean absolute difference (cM) of QTL position estimate obtained from the IBD method using different haplotype sizes with 1 cM marker spacing in a two-breed cross followed by random mating 85

TABLE 4. – Least squares mean absolute difference (cM) and bias (cM) of QTL position estimates obtained from the IBD method using different haplotype sizes in six alternate scenarios with 1 cM marker spacing 86
ACKNOWLEDGEMENTS

I would like to acknowledge first and foremost my co-major professors, Rohan Fernando and Max Rothschild. They have played fundamental, while different, roles in my growth as a scientist and as a person. I truly believe that I have benefited from being exposed to two different approaches to science, learning and teaching, and that my experience has been all the more enriched because of that. Sincere thanks are also extended to the members of my program of study committee, Susan Lamont, Jack Dekkers and Dan Nettleton. I wish to thank them for agreeing to help guide me, and my research, in our effort to work at the interface of quantitative and molecular genetics. I am also extremely grateful to my family and closest friends, no matter where in the country they may live, who have provided support and laughter for years. Their encouragement is what made the completion of this work possible. A special note of thanks goes to all of the members of the Rothschild lab, both past and present, for the family-like atmosphere that they helped create. It was never dull working with any of them. The same is also true of the graduate students and post-docs working on “both sides” of Kildee Hall, especially those who started their programs around the same time that I did. I hope that we can continue our relationships, both professional and personal, in the future. Finally, I would like to thank James Mohr. His unflappable attitude towards life was one of the few things I could count on during my time as a graduate student. Hopefully I will be able to rely on it for years to come.
One of the primary goals for molecular geneticists working with livestock species is to identify and characterize genes underlying complex traits, the so-called quantitative trait loci (QTL). The primary strategy for identifying QTL involves several steps, one being fine mapping of a previously defined chromosomal region and another being identification of candidate genetic polymorphisms that may cause differences in phenotype. The studies presented in this dissertation address fine mapping methodologies, use of the candidate gene approach for directly identifying candidate genetic polymorphisms and use of bioinformatic tools for identifying genetic polymorphisms in silico. Results from simulation studies suggest that two linkage disequilibrium-based fine mapping methods, one using haplotype information, the other using single marker information, provide QTL position estimates with comparable accuracy. Additional research is necessary to determine optimal fine mapping methods under experimental research conditions. The candidate gene studies presented, concerning the porcine connexin 37 (CX37) and bone morphogenetic factor 15 (BMP15) genes, highlight use of comparative sequence and biological information for identifying candidate genetic variants. Two synonymous mutations were discovered in the CX37 gene, which was subsequently mapped to SSC6 q24-31. However, these mutations were not significantly associated with fertility traits as hypothesized. Unfortunately, mutations could not be identified in BMP15, which was physically mapped to SSCX p11-13. Bioinformatic tools are shown here to be lucrative for identifying putative single nucleotide polymorphisms (SNPs) from redundant expressed sequence tag (EST) information in the pig. Using computer-identified SNPs, a correlation of 0.77 (p < 0.00001) was found between the frequency of human and porcine SNPs in the coding regions (cSNPs) of 25 genes, while a
correlation of 0.48 ($p < 0.0005$) was found between the frequency of human and mouse cSNPs in 50 genes. This strong human-pig relationship should be verified in a larger sample so that SNP identification in pigs could be expedited by screening porcine genes homologous to human genes known to be SNP-dense in their coding regions. By capitalizing on statistical, bioinformatic and molecular tools in an integrated approach, the rate at which QTL are identified in livestock could be increased.
CHAPTER 1. GENERAL INTRODUCTION

INTRODUCTION

Quantitative trait loci (QTL) have been defined as "loci underlying a quantitative character" (Lynch and Walsh, 1998). As most biological variation is quantitative in nature, identifying and characterizing QTL impacts our understanding of genetic evolutionary history and helps to bridge the gap between genotype and phenotype. The first step towards identifying QTL involves mapping, which identifies the chromosomal regions that contain QTL. The main principle underlying QTL mapping was first noted by Sax (1923), who made the link between phenotypic differences and observable markers, and was later formalized by Thoday (1961). Further attempts to identify QTL were initially few, as they were hindered by a lack of available genetic markers. In the late 1980s, allozyme loci (Edwards et al. 1987) and restriction fragment length polymorphisms (Lander and Botstein 1989) were first used as genetic markers to trace the segregation of phenotypes with genotypes. Since then, there has been a number of QTL mapping studies in human (e.g. Cardon et al. 1994; Davies et al. 1994; Myers et al. 2000), model animals (e.g. Cheverud et al. 1996; Nuzhdin et al. 1997; Shook and Johnson 1999), and livestock populations (e.g. Andersson et al. 1994; Georges et al. 1995), and an entire new area of research, methodology for mapping and identifying QTL, has been born.

Traditionally in livestock populations, improvement in quantitative traits has relied upon selection based upon estimated breeding values. The accuracy of these selection schemes depends upon the level of genetic contribution to the trait and the amount of progeny or relative information available. With knowledge of molecular markers and the underlying mechanisms controlling phenotype, selection can become more accurate. Marker-assisted
selection (MAS) incorporates individuals' genotypes at a marker locus into the selection strategy, usually through either best liner unbiased prediction (Fernando and Grossman 1989) or a selection index (Lande and Thompson 1990). This improves response to selection, especially for traits that are sex-limited, measured late in life or measured on a slaughtered animal (Meuwissen and Goddard 1996).

The accuracy of MAS, however, is limited by the relationship of the genetic marker to the true QTL. If the genetic marker is linked to the QTL but they are in population-wide equilibrium, MAS becomes much more difficult to implement as the linkage phase of the marker and QTL as well as the QTL effect must be determined for every breeding population (Dekkers and Hospital 2001). If the genetic marker is in population-wide linkage disequilibrium (LD) with the QTL or is the functional mutation that underlies the QTL, i.e. the so-called quantitative-trait nucleotide (QTN, Mackay 2001), then implementation of MAS becomes much easier as QTL effect and linkage phase can be estimated for the whole population (Smith and Smith 1993). Thus, a primary goal for molecular geneticists working with livestock is to identify QTN, or markers in population-wide LD with QTN, affecting complex traits.

As mentioned previously, the undertaking of QTL mapping studies has generated an entire area of research that involves proposing and testing methods for mapping and identifying QTL and spans statistics, population and molecular genetics (for a full discussion see Lynch and Walsh 1998). Initial QTL mapping projects focused on specific chromosomal regions, primarily because of the limited supply of genetic markers. With the advent of microsatellite markers, it became possible to scan the entire genome for QTL, typically using the method of interval mapping (e.g. Haley and Knott 1992). However, whole-genome scans in livestock usually can only refine the position of the QTL within a 20 to 30 cM region, with
potentially hundreds of genes existing in these chromosomal regions. To identify the actual
gene(s) and/or QTN responsible for the phenotypic variation, additional tools are needed.

The movement towards using single nucleotide polymorphisms (SNPs) as genetic
markers, because they are more abundant and have a much lower rate of mutation than
microsatellites, enables dense marker maps to be created. With a high density of markers
and sufficient recombination events in the population, QTL can be mapped on a much finer
scale, or in other words, at a higher resolution. Populations such as advanced intercross lines
(AIL) have been proposed for fine mapping QTL in livestock (Darvasi and Soller 1995). The
AILs are created by *inter se* mating of F2 individuals and subsequent generations until
sufficient recombination has occurred to allow mapping resolution on a fine scale (Darvasi
and Soller 1995). However, in most livestock populations, creating AILs is too expensive
and time-consuming to be feasible. Thus, capitalizing on historical recombinations that have
occurred in the population, and the resulting levels of linkage disequilibrium (LD) that exist
between the QTL and closely linked markers, is the most practicable method for fine
mapping QTL. To fine map using only LD, no pedigree information is necessary, and
mapping resolution is limited only by the extent of LD around the QTL and by the marker
density. Typically, QTL regions can be narrowed from the 20 to 30 cM range provided by
whole-genome scans to less than 5 cM in LD-based fine mapping studies (Farnir *et al.* 2002;
Meuwissen *et al.* 2002). In these smaller regions, there are likely only a handful of genes,
possibly 5 to 10, which may contain the underlying QTN and require further study.

Another approach to identify genes responsible for quantitative variation is candidate
gene analysis (e.g. Rothschild *et al.* 1996). There are primarily three ways to choose
candidates: based upon known gene and/or protein function, based upon mutational analysis
and based upon the gene's position on the chromosome (Rothschild and Soller 1997). The
first type of candidate is called a biological candidate because its selection is based upon the function of the gene, which is typically known from studies in other species such as mouse or human. The second is called a mutational candidate because its selection is based upon mutation analysis performed in another species, such as a knock out mutation study in mouse, which causes an altered phenotype, thus suggesting the physiological role of the gene. The third type is called a positional candidate because its selection is based upon a previously identified QTL region. Often, these types of gene selection will be combined such that a candidate gene is selected not only for its known or inferred physiological role, but also because it maps to a region previously associated with a trait or traits. The candidate gene approach, like fine mapping using anonymous markers, also requires the identification of SNPs because this approach directly attempts to identify the QTN or a mutation in population-wide LD with the QTN.

For both approaches that have been discussed, fine mapping and candidate gene analysis, there is a common requirement. That requirement is the identification of SNPs. Due primarily to the large sequencing efforts in humans, there is a vast amount of SNPs that have been identified and made publicly available through resources such as dbSNP (Smigielski et al. 2000). However, no such resource exists presently for livestock species. For example, with no immediate efforts to sequence the porcine genome, large scale SNP identification seems unlikely. Thus, SNP detection proceeds at a fairly slow rate in pigs. Due to the lack of a complete genome sequence, researchers have focused instead on creating large collections of expressed sequence tags (ESTs, Davoli et al. 2002; Fahrenkrug et al. 2002; Caetano et al. 2003; Tuggle et al. 2003) to identify and characterize genes in the pig. Most of these ESTs are publicly available through resources like dbEST (Boguski et al. 1994). With recent advances in the area of bioinformatics, the possibility of identifying SNPs from
redundant EST sequences has emerged. This so-called *in silico* SNP detection has been performed in both human (Buetow *et al.* 1999; Garg *et al.* 1999; Marth *et al.* 1999; Picoult-Newberg *et al.* 1999) and plant species (Kota *et al.* 2003) and seems a feasible option for increasing the rate at which SNPs can be discovered in pigs. The SNP detection rate could also be increased if a link could be made between human and pig SNPs so that researchers could capitalize on the large numbers of SNPs identified in humans.

The work presented here furthers the understanding of fine mapping methodology and its practical use, provides tools to increase the rate of SNP identification in pigs and increases the density of known genes on porcine linkage and physical maps. The ultimate goal of these projects is rapid identification of genes underlying quantitative variation in livestock to assist in genetic improvement programs.

**RESEARCH OBJECTIVES**

Identifying and characterizing QTL in agricultural species furthers our understanding of basic physiology and provides a new type of measurement, in the form of a preferable genotype, that can be selected upon to improve production and/or health traits. The integration of different approaches, *e.g.* performing fine mapping in a previously identified chromosomal region followed by candidate gene analysis, is becoming necessary to accurately identify QTL. Conducting these types of integrated projects is time-consuming, and any effort that can be made to simplify or increase the rate at which this work can be conducted is beneficial to the agricultural and research community. The aim of the fine mapping-related work presented in Chapters 2 and 3 is to provide information about the relative accuracy of two different mapping methodologies so that future studies can use the simplest and most accurate method to refine QTL position. The goal of the candidate gene
projects discussed in Chapters 4 and 5 is to directly identify sequence variation in two porcine genes that are known to be involved in ovulation rate from studies in other species and to use this variation to map the genes. The ultimate goal of the *in silico* SNP identification project described in Chapter 6 is to determine whether a relationship can be established between the frequency of SNPs in human and pig coding regions. If a strong relationship is found, then it could be possible to conduct site-directed screening of porcine coding regions that are homologous to SNP-dense human coding regions to speed the rate of porcine SNP detection. In addition, the *in silico* methods used to identify the porcine SNPs for this project also provide a new approach for identifying SNPs in pigs and potentially other livestock species. Once validated, these computer-derived SNPs can be tested for association with various complex traits in order to identify and characterize QTL, as well as to serve as indicators of an individual’s genetic value for selection purposes in breeding populations. Although the projects described here are seemingly diverse in nature, their universal goal is the identification of genes underlying complex traits in livestock. The fact that molecular, bioinformatic and statistical methods, like the ones described here, can all be utilized in the hunt for QTL stresses the need for an integrated approach to research in complex traits.

THESIS ORGANIZATION

The remainder of this chapter provides a literature review to provide further background on the research conducted. The remainder of this thesis is organized into chapters based upon five individual papers describing the research projects. These articles have all been published, accepted by or submitted to scientific journals. Chapter 2 consists of the paper “Comparing linkage disequilibrium-based methods for fine mapping quantitative
trait loci" which has been accepted by the journal Genetics. This work was conducted by Laura Grapes under the direction of Professor Rohan L. Fernando. Professor Jack Dekkers was involved in many discussions that affected the direction of the research project and, along with Rohan, greatly assisted in the writing of the paper. Professor Max Rothschild contributed suggestions and corrections to the paper.

Chapter 3 consists of the paper "Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci" which has been submitted to the journal Genetics. This work was conducted by Laura Grapes and Mehmet Z. Firat, a postdoctoral research associate in the Department of Animal Science, again under the direction of Rohan L. Fernando as this is a continuation of the work described in Chapter 2. Mehmet Firat was involved in the computational programming for this project that required the Fortran language. Jack Dekkers again was involved in numerous discussions concerning the progress of this work and made significant contributions to the paper. Max Rothschild contributed suggestions and corrections to the paper.

Chapter 4 consists of the paper "Physical and linkage mapping of the porcine connexin 37 (CX37) gene" which is available in the Journal of Animal Science vol. 80(5), pp. 1375-1376 (2002). The work was conducted by Laura Grapes and Yuandan Zhang, a postdoctoral research associate in the Department of Animal Science. Yuandan Zhang provided assistance in linkage mapping analysis. Max Rothschild provided general research guidance and assisted in the writing of the article.

Chapter 5 consists of the paper "BMP15 maps to the X chromosome in swine" which is available in the journal Animal Genetics vol. 33(2), pp. 165-166 (2001). The work was conducted by Laura Grapes, while Max Rothschild provided research guidance and several comments and suggestions for the paper.
Chapter 6 consists of the paper “Prospecting for pig SNPs in the human genome: have we struck gold?” which has been submitted to the journal Nature Genetics. The work was conducted by Laura Grapes, Stephen Rudd, a bioinformatician at the Centre for Biotechnology in Turku, Finland, Dominique Rocha and Karine Megy, both from the Department of Pathology at the University of Cambridge, England. A majority of the bioinformatics analyses for this project was performed by Stephen Rudd. Laboratory experiments and bioinformatics to validate SNPs identified through computational analyses were conducted by Dominique Rocha, Karine Megy and Meena Bagga, who is also from the Department of Pathology at the University of Cambridge. Max Rothschild and Rohan Fernando provided research guidance and ideas and assisted in the writing of the paper.

Chapter 7 summarizes the general conclusions from each of the projects described in chapters 2 through 6. It discusses the relationship between these findings, which are primarily concerned with methodology and new tools, and the contribution of these results to the overall objective of identifying and characterizing QTL. Suggestions for the use of these tools and methods in future research are provided.

LITERATURE REVIEW

With the completion of several QTL mapping studies in livestock species, regions of chromosomes have been identified that likely contain genes affecting traits of economic importance. The next step is to examine these chromosomal regions and attempt to identify the underlying genes. Presently, two main approaches have been used for identifying the gene or genes underlying the quantitative variation. One is fine mapping and the other is the candidate gene approach.
Fine mapping QTL using linkage disequilibrium

Fine mapping differs from the candidate gene approach in that it is not forced to rely upon comparative information across species that may or may not be useful. However, both approaches require sufficient recombination around the QTL in order to obtain high mapping resolution, and that may be difficult to obtain or observe in a population. One approach to increase the number of recombination events is to consider the historical recombinations in the population. Given enough time and excluding strong selection, an individual's haplotype will break apart, due to repeated crossover events in each generation, such that in the haplotypes of its distant descendants, recombination between any two loci could be observed, as compared to the original haplotype. This breakdown in linkage disequilibrium (LD) between loci in a haplotype was first described by Jennings (1917) and Robbins (1918), and its usefulness for fine mapping was first proposed by Bodmer (1986) for mapping human genes.

Paterson et al. (1990) showed that recombinant tomato plants, identified from an initial generation, could be selectively crossed for several generations in order to obtain chromosomes known to be recombinant at specific chromosomal regions, which could then be associated with a trait. Similarly in mice and plant species, recombinant inbred lines (Bailey 1971; Taylor 1976; Burr et al. 1988), near-isogenic lines (Kaeppler 1997) and congenic strains (Darvasi 1997; Hill 1998) have all been proposed as useful populations for fine mapping QTL. These types of populations utilize recurrent backcrossing or selfing to generate chromosomes that are uniform except for small regions, which can be associated with a trait. For most livestock species, though, it is impossible to create these types of populations due to inbreeding. Darvasi and Soller (1995) proposed the creation of advanced intercross lines (AILs) to generate the recombination events necessary for fine mapping. An
AIL is created by the random intercrossing of individuals in the F$_2$ and in subsequent generations, which originated from inbred lines or lines assumed to be homozygous for alternate QTL alleles. The recurrent generations of intercrossing allow the LD generated by the initial cross to break down to a level amenable for fine mapping. They showed that most of the improvement in mapping accuracy was accrued by the 10$^{th}$ generation of random mating in a population having a constant effective size of 100 (Darvasi and Soller 1995). To create such a population in a species with a generation interval similar to cattle, or even pigs, would take years and incur a high cost. Thus, instead of generating LD and then performing mating or selfing for multiple generations to allow its decay, as in the populations described previously, another method was necessary to utilize historical recombinations for fine mapping in livestock populations.

Recently, fine mapping studies in dairy cattle have utilized the linkage disequilibrium information that exists within a previously identified QTL region after an assumed number of generations of recombination events have already occurred in the population (Farnir et al. 2002; Meuwissen et al. 2002; Blott et al. 2003). Fine mapping results from two of these studies (Farnir et al. 2002; Blott et al. 2003) led to the identification of genes that were strongly associated with the traits (Grisart et al. 2001; Blott et al. 2003) following initial QTL mapping studies (Coppieters et al. 1998; Georges et al. 1995). Fine mapping was a critical step towards the identification of these genes as the initial QTL regions were up to 20 cM (Coppieters et al. 1998; Georges et al. 1995) in size, potentially containing several dozens of genes that could not all have been examined solely by the candidate gene method.

At a given locus, two alleles are said to be identical by descent (IBD) if they can be traced back to one allele in a common ancestor. The probability that alleles at a given locus are IBD can be used for mapping purposes. If two individuals carry IBD QTL alleles, then it
is likely that their phenotypic covariance will be higher than the covariance between two individuals with non-IBD QTL alleles. Typically, IBD probabilities are obtained for a pair of individuals conditional on parental and offspring marker genotypes (Amos and Elston 1989). However, under an assumed population structure, LD information can be also used to obtain IBD probabilities. If a QTL allele is introduced into a population in a manner such that LD is generated between linked markers and the QTL, then, after many generations of random mating, recombination will sufficiently break down that LD such that only the most tightly linked markers will still be in strong LD with the QTL. So, regardless of the relationship between individuals within a generation, the probability that two individuals share IBD QTL alleles can be estimated by considering the number of markers that are identical in state (IIS) between those two individuals. In theory, if two individuals share makers that are IIS within a given chromosomal region, then it is likely that they are IBD for the region spanned by the IIS markers. The larger the number of IIS markers shared between two individuals, the higher the probability that they are IBD for that region. Hence, if a QTL exists within that region of IIS markers, then the QTL alleles can be considered to be IBD with some probability. Again, if two individuals carry IBD QTL alleles, then it is likely that their phenotypic covariance will be higher than the covariance between two individuals with non-IBD QTL alleles.

Meuwissen and Goddard (2000) proposed fine mapping QTL strictly using the LD within a population of animals assumed to be equally related. They assumed LD between the QTL and the marker loci was generated by a mutational event in the founder generation, and IBD probabilities for the QTL were estimated based upon the marker haplotype information of individuals within the most recent generation (Meuwissen and Goddard 2000). They showed that their method was robust to the assumptions about effective population size and
number of generations since the founder population and proposed a multi-stage approach for fine mapping QTL which requires increasing marker density at each stage (Meuwissen and Goddard 2000).

Farnir et al. (2002), Meuwissen et al. (2002) and Blott et al. (2003) utilized both pedigree and linkage disequilibrium information for fine mapping. Meuwissen et al. (2002) and Blott et al. (2003) used LD information to estimate the probabilities that QTL alleles were identical by descent, as described in Meuwissen and Goddard (2001). Based upon coalescence theory (Hudson 1985), Meuwissen and Goddard’s (2001) analytical method predicts the probability that two haplotypes will contain IBD QTL alleles, given their marker alleles are known at a number of genetic markers in a previously identified QTL region. There are some assumptions that must be made to estimate these IBD probabilities, though, such as known effective population size and the number of generations since the base population (Meuwissen and Goddard 2001).

Meuwissen and Goddard (2000) also attempted to prove that their multi-locus LD mapping method had higher accuracy than a single-locus LD mapping method by comparing it to the transmission disequilibrium test (TDT) of Rabinowitz (1997). The TDT, however, utilizes only within-family information, at the cost of reduced power, to avoid false-positive associations stemming from population admixture, while the method of Meuwissen and Goddard (2000) utilizes within- and between-family information, thus making it inherently more powerful and prompting the work reported here in Chapter 2.

**The candidate gene approach to identifying QTL**

While the candidate gene approach is based upon comparative biological, sequence, and map information, which may or may not be accurate, this approach does not require a
dense marker map or sophisticated statistical methods like fine mapping does. Also, the
candidate gene approach does not require pedigree information. The candidate gene
approach does require some luck and intuition for selecting the appropriate gene(s) to study,
but it has proven to be an effective method for identifying genes related to complex traits
(e.g. Fujii et al. 1991; Rothschild et al. 1996; McPherron and Lee 1997). It has been
particularly amenable to traits such as disease risk/resistance (Teale et al. 1996; Meijerink et
al. 1997) and fertility (Grapes and Rothschild 2001), because obtaining informative records
on individuals with pedigree information is more difficult for these than for other traits.

Identification of candidate genes for reproductive traits in pigs

Although reproductive performance is one of the most critical components in swine
production, it has been one of the most difficult areas for geneticists and animal breeders to
improve. Traditional quantitative methods such as selection, inbreeding and crossbreeding
have had limited success in this area (Rothschild 1996). This is primarily due to the low
heritabilities, sex-limited nature and insufficient phenotypic measurements for reproductive
traits.

Molecular genetic approaches, including the use of candidate genes, have allowed an
entirely different method for exploring the differences in fertility between domesticated pigs.
Since the revelation that the estrogen receptor gene (ESR) is significantly associated with
litter size (Rothschild et al. 1996), several other candidate genes for reproductive traits have
been examined. Examples of successful candidate gene studies include retinol binding
protein 4 (RBP4), prolactin receptor (PRLR) and the beta subunit of follicle stimulating
hormone (FSHB). These candidates were selected on the basis of their known role in
reproductive trait physiology.
Retinol binding protein 4 (RBP4) was known to be expressed during the critical elongation phase of pregnancy in the pig and was found to be significantly associated with litter size in six commercial lines of pigs (Rothschild et al. 2000). Prolactin receptor (PRLR), besides its obvious role in the prolactin pathway, was known to have affinity for factors other than prolactin, implicating its diverse function. PRLR has also been shown to be significantly associated with litter size (Vincent et al. 1998), number of fetuses per uterine horn, average fetal weight and total fetal weight (Isler et al. 2000) and number born alive (Drögemuller et al. 2000). The beta subunit of follicle stimulating hormone (FSHB) is responsible for the specificity of the hormone, which promotes the growth of immature follicles to a size large enough for ovulation. Li et al. (1998) reported a large significant effect of FSHB on litter size that was later confirmed by Huang et al. (2000) who showed a significant effect of FSHB on number born alive per parity. Although these studies represent examples of successful results from the candidate gene approach to identify QTL for reproductive traits, there are still several other genes with roles in other pathways that can be considered, some of which will be discussed here.

Ovulation rate in pigs is a highly complex fertility trait affected by many factors, including the mechanisms controlling follicular development. Although it has been reported that a mechanism intrinsic to the oocyte itself is responsible for mammalian follicular development (Eppig et al. 2002), intercellular signaling between the oocyte and the surrounding somatic cells of the follicle is critical for this maturation process to occur (Gilula et al. 1978).

One pathway for intercellular communication is the gap junction. Gap junctions are intercellular channels formed by hexamers of integral membrane proteins called connexins and join cells in nearly all metazoans (Goodenough et al. 1996). They are selectively
permeable to small molecules and allow the passage of ions. Connexins, the proteins that cluster to form these gap junctions, are encoded by a family of at least 13 genes and are named based upon their molecular mass in kiloDaltons (Goodenough et al. 1996). In the developing mammalian follicle, connexin 37 is responsible for forming the gap junctions that exist between the oocyte and granulosa cells (Simon et al. 1997). Knockout mice lacking the connexin 37 gene (Cx37) were healthy and appeared normal when observed grossly but were unable to produce mature follicles and had 5 to 10 times the number of corpora lutea as wild type mice (Simon et al. 1997). Interestingly, while the gap junctions between the oocyte and granulosa cells were absent in the mice lacking connexin 37, the gap junctions between granulosa cells remained present, indicating that direct communication between the oocyte and its surrounding cells is critical for follicular development (Simon et al. 1997). This abundance of functional information pertaining to connexin 37 made it a good biological/mutational candidate gene to investigate in the pig to attempt to identify an association between sequence variation in CX37 and fertility traits.

Bone morphogenetic factor 15 (BMP15) was also amenable to candidate gene analysis due to the sequence and expression information available from other species. It is a member of the transforming growth factor β superfamily and is only known to be expressed in oocytes (Laitinen et al. 1998; Dube et al. 1998; Aaltonen et al. 1999). In mice, Bmp15 was first identified as growth differentiation factor 9 beta (Gdf9B) due to its co-expression with growth differentiation factor 9 (Gdf9) in mouse oocytes (Laitinen et al. 1998). In humans, the gene was independently characterized as GDF9B (Aaltonen et al. 1999) and BMP15 (Dube et al. 1998) independently. While it had been shown that GDF9 affected granulosa cell function and was essential for follicular development (Dong et al. 1996; Hayashi et al. 1999; Elvin et al. 1999), the functional role of BMP15 was not described until
later. Galloway et al. (2000) showed that inactivation mutations in the ovine BMP15 gene were responsible for increased ovulation rate in the heterozygous state and infertility in the homozygous state. This differed from the phenotype of the mouse knockout for Bmp15, which remained fertile despite lacking a functional copy of the gene (Yan et al. 2001). It was later shown that both GDF9 and BMP15 were essential for normal follicular development and luteal function in sheep, likely due to the formation of homodimers by each protein or a heterodimer formed by GDF9 and BMP15 that acts as the functional protein in sheep (Juengel et al. 2002). The current model for initiation of follicular growth assumes a two-step process in which the oocyte begins to grow and secrete GDF9 and BMP15 after the follicle reaches the primary stage (Braw-Tal 2002). The release of GDF9 and BMP15 triggers the accelerated proliferation of granulosa cells (Braw-Tal 2002). In turn, the granulosa cells secrete kit ligand, which promotes oocyte growth, such that the oocyte and granulosa cells function as a self-sufficient unit controlling their own additional growth (Braw-Tal 2002). With its known effects on follicular development, BMP15 was a suitable choice as a biological candidate gene that potentially contributed to variation in ovulation rate in pigs.

In silico identification of porcine SNPs and their relationship with human SNP information

Success in candidate gene studies is dependent upon the ability to identify sequence variation within or across species. Fine mapping studies cannot even be considered unless a sufficiently dense marker map, typically consisting of SNP markers, has been constructed. Thus for both of these approaches the identification of SNPs is vital. In species such as the pig, whose genomes have yet to be sequenced, SNP identification proceeds at a fairly slow rate, usually on a gene-by-gene basis. It had been proposed to increase the number of SNPs identified in humans by performing in silico comparative sequence analyses using the large
number of EST sequences that are publicly available (Gu et al. 1998). With most genes being expressed in several tissue types, the chance that the same gene will be sequenced more than once is high. Also, EST sequences are derived from many individuals representing different genetic backgrounds. Thus, comparing redundant EST sequences is equivalent to screening different individuals for polymorphisms and should be an effective, low-cost method for identifying SNPs.

However, there are certain factors that must be considered when searching for SNPs in this way. First, the quality of EST sequence data is often low, having an estimated error rate of 2%, due to single-pass sequencing (Hillier et al. 1996). Second, ESTs are primarily sequenced from the 3' end of the gene. The 3' untranslated region harbors many polymorphisms but will likely represent the region of poorest quality sequence due to problems associated with sequencing across the poly-A region. Third, the cDNA sequences are generated from mRNA using reverse transcriptase, which has an error rate similar to that of the frequency of human polymorphism (Gu et al. 1998). To address these concerns, methods were proposed that utilized the primary sequencing trace information to help identify SNPs in regions of high sequence quality in humans (Buetow et al. 1999; Garg et al. 1999; Marth et al. 1999; Picoult-Newberg et al. 1999).

The accuracy of SNPs identified using these methods (Buetow et al. 1999; Garg et al. 1999; Marth et al. 1999; Picoult-Newberg et al. 1999) was found to be quite high in humans. Buetow et al. (1999) confirmed the existence of 82% of the candidate polymorphisms identified using their method, while Picoult-Newberg et al. (1999) confirmed 62% of their candidate polymorphisms. However, it was later shown by Cox et al. (2001) that in a sample of 13 genes, only 27% of the true SNPs were identified using the SNPFinder method of Buetow et al. (1999). In other words, the SNPs that were identified in silico were verified at a
high rate, but there were several SNPs existing in the population that were not detected by an \textit{in silico} method. The high number of false-negatives was attributed to the low diversity of subjects from which the EST sequences were derived (primarily of Caucasian origin) and the relatively small number of EST sequences used for computer-based searches compared to the large number of sequences used in laboratory-based studies (Cox \textit{et al.} 2001). Thus, \textit{in silico} methods appear to be biased towards polymorphisms with moderate allele frequencies.

When considering \textit{in silico} approaches for identifying SNPs in livestock, the problems and advantages are nearly opposite to that in humans. The bias of \textit{in silico} methods towards identifying polymorphisms with moderate allele frequencies is a problem for humans where, for example, it may be crucial to identify a rare variant that is associated with a disease. However, this bias towards identifying polymorphisms that have moderate allele frequencies may actually be an advantage when identifying SNPs in livestock. If a SNP is found to be associated with an important trait in a population, then it is preferable that the marker's allele frequency be relatively high in the population so that undesirable levels of inbreeding will not be incurred when selecting for the preferred allele. The advantage of human \textit{in silico} SNP studies is that the sequence trace data for many ESTs are publicly available from the Washington University EST database (Hillier \textit{et al.} 1996). A large-scale public database containing sequence trace data does not exist for livestock species. Therefore, if SNPs are going to be detected in livestock using computer-based methods, they must not rely on trace information to help determine sequence quality.

Recently, a method was proposed for identifying SNPs in barley (\textit{Hordeum vulgare} L.) from EST sequences that does not require raw sequence trace files but instead uses EST sequence information directly from publicly available databases (Kota \textit{et al.} 2003). This method was named SNiPper and was designed to identify variety-specific SNPs (Kota \textit{et al.}
2003) but could also identify intra-varietal SNPs. It utilizes the Sputnik EST pipeline of Rudd et al. (2003) to cluster and annotate EST sequences, which are then analyzed by the SNiPper algorithm. The algorithm assigns a score to a position in the cluster consensus sequence that represents the probability that deviations from the consensus sequence at that position represent a true SNP. The score is based upon the number of similar deviations at a position and the number of other surrounding positions that contain a similar pattern of polymorphism (Kota et al. 2003). This is an attempt to ensure that the putative polymorphisms are less likely to be sequencing errors. Using this method, 54 (86%) of a sample of 63 high-scoring SNPs in barley were confirmed by direct sequencing (Kota et al. 2003). With the large numbers of porcine ESTs that have been made publicly available (Davoli et al. 2002; Fahrenkrug et al. 2002; Caetano et al. 2003; Tuggle et al. 2003), it should be possible to utilize methods similar to the SNiPper algorithm to rapidly identify SNPs in pigs.

Additionally, SNP detection in pigs could be increased if a relationship existed between human and pig polymorphisms such that site-directed screening of porcine genes could be accomplished based upon human information. The frequency of nucleotide difference between two randomly chosen chromosomes is termed nucleotide diversity (Nei and Li 1979). It has been shown that nucleotide diversity in humans varies widely across the genome, with some regions having nucleotide diversity levels less than 0.1% (Nachman et al. 1998) and others having levels of 5% to 10% (Guillaudeau et al. 1998; Horton et al. 1998). The frequency of human polymorphisms within coding regions has also been shown to vary across genes (Cargill et al. 1999; Halushka et al. 1999). An explanation for gene-specific variation in human nucleotide diversity has yet to be provided. However, with the high levels of coding and protein sequence identity between humans and pigs, it may be possible
to observe similar rates of polymorphism in coding regions between humans and pigs. If a relationship could be shown, then it may be possible for researchers working in the swine industry to capitalize on the wealth of human SNP data and perform site-directed screening of areas in pigs that are known to be SNP-dense in humans. Thus, the rate of SNP identification could be advanced in pigs and the goal of a genome-wide SNP map for fine mapping and genome-wide association studies could be achieved more rapidly.

By capitalizing on all available methods, both quantitative and molecular, the genes underlying traits of economic importance can be discovered in livestock. The work presented here focuses on those different methods, spanning topics such as fine mapping methodology for more narrowly defining QTL regions, the candidate gene approach to directly identify QTL and bioinformatic approaches that can speed the rate of SNP detection in pigs and, by extension, in other livestock species.

REFERENCES


related factor is co-expressed with GDF-9 in mouse oocytes during folliculogenesis.

Mech Dev. 78: 135-140.


Meijerink, E., R. Fries, P. Vögeli, J. Masabanda, G. Wigger, C. Stricker, S. Neuenschwander, H.U. Berthsberger and G. Stranzinger. 1997. Two α (1,2) fucosyltransferase genes on porcine chromosome 6q11 are closely linked to the blood group inhibitor (S) and Escherichia coli F18 receptor (ECF18R) loci. Mamm. Genome 8: 736-741.


CHAPTER 2. COMPARING LINKAGE DISEQUILIBRIUM-BASED METHODS FOR FINE MAPPING QUANTITATIVE TRAIT LOCI

A paper accepted by Genetics

L. Grapes\textsuperscript{1}, J. C. M. Dekkers\textsuperscript{1}, M. F. Rothschild\textsuperscript{1}, R. L. Fernando\textsuperscript{1,2}

ABSTRACT
Recently, a method for fine mapping quantitative trait loci (QTL) using linkage disequilibrium was proposed to map QTL by modeling covariance between individuals, due to identical by descent (IBD) QTL alleles, based on the similarity of their marker haplotypes under an assumed population history. In the work presented here, the advantage of using marker haplotype information for fine mapping QTL was studied by comparing the IBD-based method with 10 markers to regression on a single marker, a pair of markers, or a two-locus haplotype under alternative population histories. When 10 markers were genotyped, the IBD-based method more accurately estimated the position of the QTL than single marker regression in all populations. When 20 markers were genotyped for regression, as single marker methods do not require knowledge of haplotypes, regression had similar or greater mapping accuracy in all populations as the IBD-based method using 10 markers. Thus for populations similar to those simulated here, the IBD-based method is comparable to single marker regression analysis for fine mapping QTL.

\textsuperscript{1} Graduate student and Professors, respectively, Department of Animal Science, Iowa State University
\textsuperscript{2} Author for correspondence
INTRODUCTION

The purpose of mapping quantitative trait loci (QTL) in livestock is to identify genes affecting a quantitative trait and ultimately use existing variation in those genes to select superior individuals from a population. One difficulty is that traditional QTL linkage studies identify chromosomal regions, not individual genes, which may affect a trait. Depending on the power of the test and population structure, these regions can range from 20 to 40 cM in size and contain possibly thousands of genes. It is impractical to consider thousands or even hundreds of potential candidate genes to identify the QTL. Therefore, the chromosomal region associated with the trait should be narrowed, i.e. the region should be fine mapped, before attempts to identify the gene are made.

Advanced intercross lines (Darvasi and Soller 1995) and recombinant inbred lines (Taylor 1978) have been proposed as resource populations to be used for fine mapping. In these populations, due to repeated recombination, the linkage disequilibrium (LD) generated by the initial cross is limited to closely linked loci. However, these types of populations are nearly impossible to create for most livestock species, as well as humans, because of time, ethical and financial constraints, as well as inbreeding depression. To overcome this, it has been proposed to use the existing LD from historical recombinations for fine mapping (e.g. Bodmer 1986; Xiong and Guo 1997).

Meuwissen and Goddard (2000) proposed a method to fine map a QTL using LD within a haplotype of closely linked markers. In their work, they showed that haplotype-based LD mapping was more accurate than single marker-based LD mapping by comparing their method to the transmission-disequilibrium test (TDT) of Rabinowitz (1997). The TDT is, however, restricted to within-family information, unlike the method of Meuwissen and Goddard (2000). The TDT has an advantage in that it is not affected by breed or line
differences (population admixture), but this advantage comes at the expense of the power of the test. The method of Meuwissen and Goddard (2000) is affected by population admixture, but it is an inherently more powerful test because it uses across-family information. A simple and more appropriate comparison would be to test the haplotype-based method of Meuwissen and Goddard (2000) against least squares regression on single markers because both these approaches use within and between-family information, and both are subject to admixture. Thus, the purpose of this work was to compare the haplotype-based method of Meuwissen and Goddard (2000) to single marker-based regression methods in order to determine if haplotypes provide additional information for fine mapping QTL.

The method of Meuwissen and Goddard (2000) maps QTL by modeling the covariance between individuals based upon the similarity of their haplotypes. Individuals with similar marker haplotypes will likely share QTL alleles that are identical by descent (IBD) and so will have a higher covariance. Assumptions about the population history are made to model the covariance. Meuwissen and Goddard (2000) showed that their IBD method is quite robust to departures from these assumptions, but it is unclear whether these assumptions affect comparisons with least squares regression methods. So, determining the impact of population history on comparisons between the methods was the second objective in this study.

METHODS

Population Simulations

Following Meuwissen and Goddard (2000) it was assumed that a previous linkage analysis study had mapped a QTL to a region of 2.25 to 9 cM in size, and within that region 10 bi-allelic markers were available. Thus in all simulations, individuals were generated
with 10 evenly spaced, bi-allelic markers, a QTL centered between two adjacent markers, and a trait phenotypic value according to their QTL genotype.

**Default population:** The IBD method is based upon modeling the covariance between individuals under the following assumptions: 1) variation in a QTL is due to a mutation that occurred 100 generations ago, 2) during the last 100 generations the effective population size was 100 and 3) each marker locus has two alleles with equal frequencies in the founder population. It was known which markers were maternally and paternally inherited so that haplotypes could be constructed. The data under the default simulation were generated under these assumptions with the QTL placed in the middle of the marker haplotype.

Phenotypic values for individuals in the final generation were generated similar to Meuwissen and Goddard (2000). In all simulated populations, except for a crossbred population that will be described later, the QTL alleles were uniquely numbered in the founders. So with an effective population size of 100, the initial frequency of each QTL allele is 0.005. In all simulations, one QTL allele with a frequency higher than 0.1 in the final generation was randomly selected to be the mutant QTL allele. This mutant allele was given an additive genetic value of 1, and the value of all other QTL alleles was set to 0. The phenotypic value for each individual in the final generation was calculated by adding the QTL allele effects to an environmental effect sampled from \( N(0, 1) \).

As explained below, additional resources would be necessary to complete an experiment that uses haplotypes as compared to single markers. To determine the haplotypes of an individual, the genotypes of both parents may be required. Assuming all individuals in the final generation have different parents, up to three times as many genotypes would be required for an experiment that uses a haplotype-based analysis as compared to a single marker-based analysis. Thus, given the same resources, single-marker based analyses would
permit a higher marker density. So, the regression analyses were also simulated with a higher density of twenty markers to compare the methods under more equitable resources.

**Alternative populations:** To test robustness of the methods to population history assumptions, several populations were created that differed from the default for one or more conditions. In the first, the population was created by crossing two breeds with divergent allele frequencies for two QTL alleles (see Table 1). After crossing, the population was randomly mated for 1, 5, 10, 20 or 100 generation(s). In the second population, the QTL was fixed at a position other than the center of the haplotype. In the third population, marker allele frequencies were assigned at random in the founder generation within a range of 0.2 to 0.8. In the last population, a “worst case scenario” was created that differed from the default for all three conditions listed above. Details of all simulations are summarized in Table 1.

**Maximum-likelihood estimation (IBD method):** To fine-map the QTL, phenotypic data in the final generation for a single trait, assuming one record per individual, were modeled following the method of Meuwissen and Goddard (2000) by

\[
y = Xb + a + e, \tag{1}
\]

where \(y\) is a vector of phenotypic values, \(b\) is a vector of fixed effects, which here reduces to the overall mean, \(X\) is an incidence matrix for \(b\), which reduces to a vector of ones, \(a\) is the vector of random genotypic values at the QTL, and \(e\) is the vector of residuals. The variance-covariance matrix of residuals is \(\text{Var}(e) = R\sigma_e^2\), where \(R\) is an identity matrix. The variance of the vector of genotypic values is \(\text{Var}(a) = G_p\sigma_a^2\), where \(G_p\) is the additive relationship matrix for the QTL conditional on marker information, when the QTL is at position \(p\). In the model used by Meuwissen and Goddard (2000) in place of \(a\) in equation (1) they fitted \(Zh\), where \(h\) is a vector of random haplotype effects, and \(Z\) is an incidence matrix for \(h\). The size of \(h\) is \(q \times 1\), where \(q\) is the number of unique marker haplotypes in
the final generation. Their model assumed that identical marker haplotypes contain the same QTL allele. However, it is theoretically possible for two identical marker haplotypes to contain different QTL alleles. Model 1 does not make this assumption. Thus the covariance is modeled more accurately using equation (1) than by the model used by Meuwissen and Goddard (2000), which likely overestimates the covariance between individuals in some cases.

The additive relationship coefficient between two individuals is twice the probability that a random allele from one individual is identical by descent to a random allele from the other individual. Matrix $G_p$ contains these relationship coefficients for a QTL at position $p$ given the marker haplotypes. To determine IBD probabilities for the QTL based upon marker haplotypes, the gene drop method described in Meuwissen and Goddard (2000) was used. This method compares a pair of haplotypes from the final generation by counting the number of markers to the left ($N_l$) and to the right ($N_r$) of the QTL that are consecutively identical in state (IIS). This assigns a haplotype pair to a distinct $(N_l, N_r)$ category. The purpose of the $(N_l, N_r)$ category is two-fold. First, the category defines a region around the QTL of size $(N_l, N_r)$ that may be IBD. Second, the number of IBD probabilities that must be estimated is reduced because multiple haplotype comparisons fall into the same $(N_l, N_r)$ category. After assigning a haplotype pair to a $(N_l, N_r)$ category, it is then determined whether the haplotype pair shares QTL alleles that are IBD. The QTL alleles are all uniquely numbered in the founder generation. So, individuals with QTL alleles that are IIS must also be IBD. Each pair of haplotypes from the final generation is categorized by its $(N_l, N_r)$, and the IBD state of its QTL alleles is determined. To obtain estimates of IBD probabilities for each $(N_l, N_r)$ category, the number of times the QTL alleles were IBD for that category was divided by the number of times the $(N_l, N_r)$ category was observed across one hundred
thousand replicates of the default simulation. These probabilities were calculated for each position that the QTL could take. Meuwissen and Goddard (2000) presented these IBD probabilities as approximations to the IBD probabilities that would be calculated if every possible haplotype pair was considered. However, as will be demonstrated in the discussion, these IBD probabilities are in fact not approximations to IBD probabilities for individual haplotypes.

By assuming multivariate normality, the residual loglikelihood of model (1) is

\[
L(G, \sigma_a^2, \sigma_e^2) = -0.5 \ln \left( |V| \right) + \ln \left( |X'V^{-1}X| \right) + (y - X\hat{b})'V^{-1}(y - X\hat{b}),
\]

where \( V = \text{Var}(y) = [G\sigma_a^2 + R\sigma_e^2] \) and \( \hat{b} \) is the generalized least-squares estimate of \( b \). For every central position of a marker bracket, \( p \), that was considered for the QTL, the likelihood was maximized with respect to the variance components \( \sigma_a^2 \) and \( \sigma_e^2 \). The position with the highest loglikelihood was the estimated position of the QTL. Simulations using the IBD method for mapping were replicated 1,000 times.

**Single-locus regression models:** For fine mapping using marker regression methods, the phenotypic data for the final generation were modeled by

\[
y = Xb + e. 
\]

In the first single locus model (SL), \( y \) is a vector of phenotypic data, \( b \) is a 2x1 vector \((\mu_0, \mu_1)\) that contains the intercept and the regression coefficient for a single marker locus, and \( X \) is an incidence matrix for \( b \). The hypothesis \( H_0: \mu_{ij} = 0 \) vs. \( H_A: \mu_{ij} \neq 0 \) was tested for each effect \( i \) at every marker locus \( j \). The position of the marker locus with the largest F statistic was the estimated position of the QTL. Simulations using any regression-based method for mapping were replicated 10,000 times as they were much less computationally intensive than the IBD method.
For the second single locus model (SL2), two adjacent loci were tested for association with the QTL. This model was included to determine if regression on two flanking markers could perform better than regression on a single marker or the IBD method, which also attempts to position the QTL between two flanking markers. Phenotypic data for the final generation were modeled as in equation (2) except that \( b \) is a \( 4 \times 1 \) vector of allelic effects (\( \mu_{0i}, \mu_{1i}, \mu_{0j}, \mu_{1j} \)) for alleles 0 and 1 at two adjacent marker loci \((i, j)\). The hypothesis \( H_0: \mu_{0i} = \mu_{1i} \) and \( \mu_{0j} = \mu_{1j} \) vs. \( H_A: \mu_{0i} \neq \mu_{1i} \) or \( \mu_{0j} \neq \mu_{1j} \) was tested for every pair of adjacent marker loci (marker bracket). The center of the marker bracket with the largest F statistic was the estimated position of the QTL.

**Two-locus haplotype regression model:** In this model (HAP), a haplotype was constructed from two adjacent marker loci. This model was included to examine the ability of regression to utilize flanking marker information, but in this case the markers were fit as a haplotype to more closely resemble the IBD method. Phenotypic data for the final generation were modeled as in equation (2), except that \( b \) is a \( 5 \times 1 \) vector including the intercept and haplotype effects (\( \mu, \mu_{00}, \mu_{01}, \mu_{10}, \mu_{11} \)) for alleles 0 and 1 at two adjacent marker loci. The hypothesis \( H_0: \mu_{00} = \mu_{01} = \mu_{10} = \mu_{11} \) vs. \( H_A: \mu_{00} \neq \mu_{01} \) or \( \mu_{00} \neq \mu_{10} \) or \( \mu_{00} \neq \mu_{11} \) was tested for every marker bracket. The center of the two-locus haplotype (marker bracket) with the largest F statistic was the estimated position of the QTL.

**Comparison of methods:** To evaluate the ability of the methods to estimate the QTL position, the absolute differences between the estimated QTL position and the true QTL position were obtained for each method from each replicate of a simulation as

\[
\text{absolute difference} = |\hat{\Theta}_i - \Theta|
\]

where \( \hat{\Theta}_i \) is the estimated QTL position in cM for replicate \( i \) and \( \Theta \) is the true position of the QTL in cM.
Bias of each method was estimated by

\[ \text{bias} = \frac{\sum_{i=1}^{n} \hat{\Theta}_i}{n} - \Theta, \]

where \( n \) is the number of replicates performed for a method.

To test for differences in mapping accuracies between methods, absolute differences for all replicates of a simulation were analyzed using ANOVA (JMP ver. 5.0, SAS Institute, Inc.) with method fit as a fixed effect. Although absolute differences are not normally distributed, ANOVA is known to be robust when the sample size is large as in this study. The least-squares mean of absolute differences (LSMD) was obtained for each method. The LSMD is a measure of a method's ability to estimate the position of the QTL, and a method with a smaller LSMD is preferable.

RESULTS

Comparison under the default population: The IBD method with 10 markers was compared to the regression methods SL, SL2 and HAP, each with 10 markers. The LSMD for each method using three different marker spacings is presented in Table 2.

The average LSMD across methods using 10 markers was 1.41 cM when the marker spacing was 1 cM, indicating that the mapping resolution of all methods was fairly good. At this marker spacing, an average QTL position estimate could be expected to deviate from the true QTL position by less than two markers or marker brackets from the QTL. Additionally, average mapping resolution increased proportionately as the marker spacing decreased. The average LSMD across methods using 10 markers was 0.74 and 0.42 cM for marker spacings of 0.5 and 0.25 cM, respectively. In both cases, an average QTL position estimate could be expected to deviate from the true QTL position by less than two markers or marker brackets.
The bias of all four methods under the default simulation was approximately zero. The mean QTL position estimate for each regression method differed from the true QTL position by ±0.05 cM or less, regardless of marker spacing. The IBD method's mean QTL position estimate differed from the true QTL position by 0.1 cM when the marker spacing was 1 cM and differed by ~ 0.02 cM when the markers were spaced 0.5 and 0.25 cM apart. A bias of zero was expected because the QTL was positioned in the center of the marker haplotype.

Comparing LSMD across methods, the IBD method was significantly better at estimating position of the QTL than the SL method with 10 markers (SL-10) for all three marker spacings (Table 2). The SL-10 method was significantly better than the SL2 method with 10 markers (SL2-10) when the marker spacing was 1 and 0.5 cM. Interestingly, fitting a two-locus haplotype in regression (HAP method) using 10 markers performed similarly to the IBD method regardless of marker spacing.

Next the regression methods, excluding HAP, were allowed to have 20 markers genotyped and were then compared to the IBD method in an attempt to evaluate the approaches with more equitable genotyping costs, considering that the IBD method requires knowledge of haplotypes. The SL method using 20 markers (SL-20) was significantly better at positioning the QTL in its true location than all other methods when markers were spaced either 0.5 cM or 0.25 cM apart (Table 2). However, when markers were spaced 0.125 cM apart (0.25 cM for IBD), SL-20 was not significantly better than IBD. With 20 markers, SL2 was significantly poorer at positioning the QTL than SL-20 and IBD. This regression method, SL2, may perform consistently worse than SL because there are more degrees of freedom associated with the markers for this model (2 degrees of freedom) as compared to SL (1 degree of freedom).
Again, biases of the regression-based methods were small (less than ±0.04 cM) except for the SL2 method with 20 markers at 0.5 cM marker spacing. Its mean position estimate differed from the true position by -0.12 cM. However, at smaller marker spacings, bias of the SL2 method was less than -0.04 cM.

In general, LSMD of the SL method was smaller when 20 markers were used as compared to 10 for all marker spacings (Table 2). Interestingly, in the case of SL2, LSMD changed very little when 20 markers were used as compared to 10 for all marker spacings (Table 2). So the ability to utilize extra information from additional markers appears to be dependent upon the method of analysis.

**Two-breed cross followed by random mating:** Two breeds were simulated, each of effective size 100, that had the same two QTL alleles but at different frequencies (see Table 1). The number of generations of random mating that occurred after the initial cross of the two breeds ranged between 100 and one. The LSMDs for the IBD method and the SL method with 10 (20) markers for each of the different numbers of generations of random mating are shown in Table 3. Marker spacing was set to 1 (0.5) cM, and the QTL was located at the center of the marker haplotype. Due to the poor performance of the SL2 in the default population, it was not tested in any of the alternative populations. The HAP method was not tested in any of the alternative populations in order to focus on the comparison between single marker-based analysis and the IBD method.

Population admixture affected the accuracy of all methods negatively (Table 3). Even with 100 generations of random mating, LSMD was greater than in the default population for both methods (Table 2). In fact, the LSMD of the IBD and regression methods was often greater than the LSMD of a randomly selected QTL position, which is 2 cM for the 10 marker case (1 cM spacing) and 2.25 cM for the 20 marker case (0.5 cM spacing).
spacing) with a centrally located QTL. Note, however, that a centrally located QTL is most favorable for a random estimator of QTL position, i.e. the LSMD of a randomly selected QTL position will be smallest when the true QTL is located in the center of the chromosome. All of the simulated populations, except for the non-central QTL and “worse case scenario”, included a centrally located QTL. So, the accuracy of the methods is being compared to the most accurate random QTL position estimate. Bias of the methods remained small, ranging from −0.17 to 0.16 cM. As the number of generations of random mating decreased, LSMD tended to increase. However, when the number of generations of random mating decreased from 100 to 20, LSMD decreased for all methods. This may be due to the fact that there were initially only two QTL alleles in this population and after 100 generations of mating the QTL alleles attained extreme frequencies or became fixed in many replicates, resulting in lower mapping resolution.

In nearly all cases, the IBD method was significantly better than the SL-10 method but not significantly different from the SL-20 method (Table 3). With 100 generations of random mating, however, the SL-20 method was significantly better and there was no difference between the IBD and SL-10 methods. When only one generation of random mating occurred after the cross, a situation comparable to an F2 population, the SL-20 and IBD methods were better than the SL-10 method. A basic assumption of the IBD method was violated in this population, i.e. the event that created linkage disequilibrium. It was expected that the mapping accuracy of the IBD method would be more negatively affected than the mapping accuracy of regression methods because they make no assumptions about population history. However, both methods had similar mapping accuracies. So, violating this assumption had no impact on the comparison of the methods.
Non-central QTL position: In this population, the QTL was positioned halfway between markers 3 and 4 (or markers 6 and 7 when 20 markers were genotyped) and the IBD method was compared to the SL method with 10 (20) markers. The LSMD for each method with marker spacing of 1 (0.5) cM is presented in Table 4.

Both the SL-10 method and the IBD method had larger LSMDs when the QTL was positioned towards the beginning of the marker haplotype instead of at the center. However, LSMD of the SL-20 method did not change when the QTL was positioned towards the beginning of the marker haplotype. For this population, the SL-20 method was best able to estimate the position of the QTL while the SL-10 method was worst. However, all methods had much greater mapping accuracy than that of a randomly selected QTL position. The LSMD for a randomly chosen QTL position is 2.4 cM when 10 markers (1 cM spacing) are used and the QTL is between markers 3 and 4 and 2.58 cM when 20 markers (0.5 cM spacing) are used and the QTL is located between markers 6 and 7.

Bias was observed in all methods, as expected, due to the non-central position of the QTL. Bias was smallest for the SL-20 method, at 0.36 cM, followed by the IBD method at 0.51 cM, and the SL-10 method at 0.63 cM (Table 4). Although bias of the SL-20 method increased from 0.02 cM to 0.36 cM with a non-central position of the QTL, LSMD of the SL-20 method did not change (Table 4). Unlike the SL-20 method, the SL-10 and IBD methods showed an increase in both bias and LSMD for a non-central QTL. The bias of all three methods remained relatively small though, as the bias for a randomly selected QTL position is 2 cM for both the 10 and 20 marker case.

Variable marker allele frequencies: In all previous populations, initial frequency of the marker alleles was 0.5. Here marker allele frequencies in the founders were randomly set at each marker locus within a range of 0.2 and 0.8 and then the IBD method was compared to
the SL method using 10 (20) markers. The LSMDs for these methods at a marker spacing of 1 (0.5) cM are shown in Table 4.

The performance of all methods in this population was similar to their performance in the default population (Tables 2 and 4). The LSMDs of all methods increased by 0.04 cM or less from their LSMDs in the default. Additionally, the bias for all three methods remained close to zero, ranging from 0.03 to -0.09 cM (Table 4). Comparing methods, LSMD of the SL-20 method was smallest, while LSMD of the SL-10 method was highest. This ranking of methods is the same as for the default population. So, it appears that SL and IBD methods were not sensitive to marker allele frequencies.

"Worst case scenario": The previous alternative populations differed from the default by only one condition. Here, several conditions were changed from the default population to create a "worst case scenario". First, the two breeds described previously were crossed, followed by 10 generations of random mating. Second, the QTL was positioned between marker loci 3 and 4 when 10 markers were genotyped and between marker loci 6 and 7 when 20 markers were genotyped. Third, marker frequencies of the founders were set at random, as described previously.

The IBD method and the SL method using 10 (20) markers were tested for this "worst case scenario" with a marker spacing of 1 (0.5) cM and their LSMDs are shown in Table 4. The LSMD of all methods increased drastically compared to the default population. The average LSMD for the SL-10, SL-20 and IBD methods increased from 1.33 cM under the default conditions to 2.52 cM in this population. The LSMD of the three methods were similar to the LSMD of a randomly selected QTL position, which is 2.4 cM when 10 markers (1 cM spacing) are used and 2.58 cM when 20 markers (0.5 cM spacing) are used and the QTL is in a non-central location as mentioned previously. Biases also increased markedly,
from a range of -0.04 to 0.1 cM in the default scenario, to a range of 1.49 to 1.76 cM in the “worst case scenario” (Table 4). These values are similar to the bias of a randomly selected QTL position, which is 2 cM as described previously. Bias was towards the center of the chromosome for all methods. The large positive bias and the near doubling of the LSMD when compared to the default are unique to this population. However, when comparing LSMD across methods, the results are not unique. Here the SL-20 method was not significantly different from the IBD method, and both were significantly better than the SL-10 method. This result is similar to the results from the two-breed cross in which, in nearly all cases, the SL-20 method and the IBD method were similar and significantly better than SL-10 (Table 3).

**DISCUSSION**

**Comparing performance of mapping methods:** Results from this work show that least squares regression on a single marker is an effective method for LD-based fine mapping of QTL if a dense marker map is available. In situations that were both ideal and non-ideal for the IBD method of Meuwissen and Goddard (2000), mapping precision of the IBD method was greater than that of the SL method given an equal number of markers. Mapping precision of the SL method using 20 markers was similar to or greater than that of the IBD method with 10 markers. It should be pointed out, however, that mapping precision of the SL method was underestimated in the populations simulated here, because the SL method estimates the position of the QTL at a marker locus, but the true position of the QTL was always simulated at the center between two marker loci. Thus, the most accurate QTL position estimate the SL method can have is at one of the markers flanking the true QTL, which introduces an inherent level of error for the simulations performed here. In contrast,
the IBD method estimates the position of the QTL at the center of a marker bracket, which is where the QTL is simulated, so it does not have an inherent error.

The comparable performance of the IBD and SL methods is contradictory to the generally held expectation that using more information (i.e. a haplotype) results in better estimates. One possible explanation is that IBD probability matrices were similar for adjoining positions of the QTL. In other words, IBD probability matrices were not sensitive to position of the QTL. Thus, for adjoining positions of the QTL the likelihoods were also similar, possibly resulting in decreased mapping precision. Further studies will examine how the number of markers considered in the haplotype affects the sensitivity of the IBD probability matrices and mapping precision.

Another possible explanation for this contradictory result may stem from the fact that the regression-based methods model the disequilibrium using location parameters (mean effects of marker alleles), while the IBD method models the disequilibrium using dispersion parameters (variance of genotypic values and error variance). It is well known that location parameters are easier to estimate than dispersion parameters. Thus, single marker regression-based methods may have an inherent advantage over the IBD method.

**Effects of alternative populations:** Several alternative populations were considered in this study to test robustness of the fine mapping methods and to determine if any methods were particularly sensitive to deviations from the default population.

First, in the default, it was assumed that a mutation on a founder chromosome was responsible for creating the linkage disequilibrium in the population. The IBD probabilities were generated under the assumption that 100 generations of random mating in a population of effective size 100 had elapsed since the mutation occurred. Meuwissen and Goddard (2000) showed that the mapping accuracy of their method was not affected by violations of
these assumptions such as altering effective population size and the number of generations of random mating since the mutation occurred. However, they did not consider an alternative event to create the initial linkage disequilibrium.

In two alternative populations in this study, the two-breed cross and the “worst case scenario”, a cross between two breeds created initial disequilibrium. It may be that these two breeds diverged from a common population several generations ago and were re-introduced. Sabry et al. (2001) tested the IBD method in a population similar to this in which four populations diverged from a founder population, were reintroduced after 90 generations and allowed to randomly mate for six generations. Sabry et al. (2001) found the IBD method to be robust to this population structure, in contrast to our result, which found that performance of the IBD method in the two-breed cross and the “worst case scenario” was much worse than in the default population. However, the regression methods also performed much worse in these two alternative populations than in the default population (Tables 2, 3 and 4). In fact, the mapping accuracy of all methods was similar to, or even less than, the accuracy of a randomly selected QTL position for both alternative populations. The “worse case scenario” does include a non-central QTL and randomly set marker allele frequencies, which the two-breed cross does not, but these were shown to have little effect on mapping ability. So the decrease in mapping accuracy for all methods is apparently due to the introduction of population admixture. Other population events such as recent bottlenecks or recurrent mutation at the QTL may also decrease the ability of the methods to fine map a QTL. Further research is needed to compare methods under these scenarios.

Second, any or all methods may be affected if the QTL is not located in the center of the chromosomal region evaluated. If the QTL is closer to either end of a chromosomal region, then there will be fewer markers on one side of the QTL than on the other. Thus
there is no longer a symmetric distribution of information across the chromosomal region. The fact that LSMD of the SL-20 method did not change when the QTL position was shifted towards the beginning of the chromosome (Table 4) supports this idea. The SL-20 method maintained six markers to the left of the alternative QTL position while the IBD and SL-10 methods maintained only three markers. The additional marker information may have allowed the SL-20 method to map the QTL equally well at both QTL positions. Also, additional marker information may have allowed SL-20 to maintain smaller bias than SL-10 or IBD with a non-central QTL (Table 4). The finite parameter space considered for the non-central QTL introduced bias for all methods. Bias of SL-10 was largest (Table 4) indicating that the additional markers, and possibly the decreased marker spacing, of SL-20 greatly improved its mapping accuracy.

Third, IBD probabilities were calculated under the assumption that initial frequencies of all marker alleles were 0.5, and violating this assumption may have an effect on the IBD method. A marker is most informative when its frequency is 0.5 so marker allele frequencies that deviate from 0.5 should also affect any fine mapping method. However, results from this study showed that the IBD method and the regression-based methods perform as well in this alternative population as in the default population. Thus, the deviation of marker frequencies from 0.5 had essentially no impact on the ability of the methods to map the QTL. This is an important result because it seems unlikely that in an actual population the frequencies of all marker alleles would be 0.5. Markers with more extreme allele frequencies were not considered because they would not be utilized in an experimental situation. So the range of founder allele frequencies used in this population is reasonable because it does not cause marker alleles to have extreme frequencies or to reach fixation in generation 100 such that mapping precision is decreased. Although all methods were robust to this alternative
population, the SL-20 method was again best able to estimate the position of the QTL and thus would be the preferred method for a fine mapping experiment if the markers were available.

**Estimation of IBD probabilities:** As noted earlier, IBD probabilities were not obtained for every possible haplotype pair but instead were estimated for groups of haplotype pairs that shared a similar distribution of IIS marker alleles around the QTL. Meuwissen and Goddard (2000) presented the IBD probabilities derived from the gene drop method as approximations to those based on individual haplotype comparisons. In fact, the IBD probabilities based on haplotype pairs are identical to IBD probabilities based on \((N_l, N_r)\) categories. This is because the IBD state of two QTL alleles is only dependent upon the number of consecutive marker alleles flanking the QTL that are IIS. The first pair of non-IIS alleles that is reached indicates a recombination event in the population simulated here. Thus, marker alleles beyond this locus are no longer informative for determining the IBD state of the QTL alleles. This was confirmed by simulating a default population with four markers instead of ten and calculating an IBD probability for each haplotype pair. The IBD probability of each haplotype pair was the same as the IBD probability of the appropriate \((N_l, N_r)\) category for the haplotype pair. This is an important result because if IBD probabilities are based on individual haplotype pairs, the number of IBD probabilities that must be estimated increases exponentially as the number of markers increases. The ability to group haplotype pairs into \((N_l, N_r)\) categories is essential for the efficient use of the IBD method.

**Current use of fine mapping methodology:** The application of fine mapping methods for positional cloning of a QTL in livestock has appeared only recently (Grisart et al. 2001; Blott et al. 2003). These studies showed that fine mapping of a previously identified chromosomal region was an important step towards identification of the gene and its causative mutation(s).
Using a maximum-likelihood approach that simultaneously mined linkage and LD information in outbred half-sib pedigrees from five different dairy cattle populations, Farnir et al. (2002) were able to refine the position of a previously identified QTL on BTA 14. This eventually led to the positional cloning of the DGAT1 gene (Grisart et al. 2001). Blott et al. (2003) modified the method of Farnir et al. (2002) to consider IBD probabilities for sires’ haplotypes so that a hierarchical clustering algorithm could be used to group haplotypes to fine map a QTL on BTA 20 affecting milk yield and composition. The bovine growth hormone receptor gene (GHR) was identified as a positional candidate gene and mutation in GHR was found to be associated with milk yield and composition (Blott et al. 2003). Meuwissen et al. (2002) extended the IBD method of Meuwissen and Goddard (2000) to also include pedigree information and fine mapped a QTL for twinning rate in dairy cattle to a region less than 1 cM. Each of these experiments took advantage of both linkage and LD information for the purposes of fine mapping, so results from this study cannot directly be extrapolated to form a comparison between regression-based fine mapping methods and the fine mapping methods used in Grisart et al. (2001), Meuwissen et al. (2002) or Blott et al. (2003).

However, it can be stated that if a fine mapping experiment was to be conducted using a sample of individuals assumed unrelated, regression-based LD mapping methods would be expected to perform as well as IBD-based LD mapping methods. If individuals were related, given the same number of individuals, the expected number of informative markers and haplotypes would decrease which could decrease mapping precision. Meuwissen and Goddard (2000) showed that mapping precision of their IBD method decreased when phenotypic records from 100 individuals in a population of effective size 50 were used as compared to records from the default population of effective size 100.
However, the decrease in mapping precision was not large (Meuwissen and Goddard 2000). Further research is necessary to examine whether population size and relation between individuals will impact LD-based mapping methods.

Evidence to support our result that single marker-based analysis is comparable to haplotype-based analysis was presented in a recent study by Zhang et al. (2003) where a variance components analysis (Abecasis et al. 2000) was used to detect association between markers and immunoglobulin E concentration in humans. The association results that were obtained using a three-, four- or five-marker haplotype as a sliding window across the region were not different from the association results obtained using single markers (Zhang et al. 2003). Future studies using experimental data rather than simulated data should also examine haplotype- and single marker-based analyses to determine their mapping precision under experimental conditions.

**Mapping under equitable resources:** Justification for the use of 20 markers in regression analysis comes from the need to compare methods as they could be used in an experimental situation. For the population described here, resources required to conduct an experiment using information from a 10-locus haplotype are more comparable to resources required to conduct an experiment using information from 20 marker genotypes rather than 10. In practice, it is possible to estimate haplotype information without knowing parental genotypes, or to infer the haplotypes when half-sib family information is available, but the IBD method as presented by Meuwissen and Goddard (2000) requires known haplotypes from equally unrelated individuals with no pedigree information. The effect of using estimated haplotype information in the IBD method has not been studied, but it is expected that this will reduce mapping accuracy. It is debatable whether it is statistically fair to
compare the SL-20 method to the IBD method with 10 markers but for experimental purposes described here it was considered fair.

The benefit of using twenty instead of ten markers was most evident in the default population (Table 2) and in the following two alternative populations (Table 4): 1) for a non-central QTL and 2) when marker allele frequencies were random. So genotyping additional markers can improve the SL method's ability to fine map a QTL by making it more robust. Of course, depending on the extent of the LD, there will be a limit to the extra information that can be obtained simply by genotyping additional markers. It may be possible that an optimum number of markers spaced an optimum distance apart exists for fine mapping. Further work is being conducted to examine this theory and to examine additional properties of haplotype-based LD mapping.

ACKNOWLEDGMENTS

This work was supported in part by funding from the United States Department of Agriculture-National Research Initiative, Sygen International, the Iowa Agriculture and Home Economics Experiment Station and by Hatch Act and State of Iowa funds. Laura Grapes was supported by a United States Department of Agriculture National Needs fellowship in quantitative and molecular genetics. The authors also wish to thank Dan Nettleton for his comments and contribution to this work.

LITERATURE CITED


Blott, S., J. Kim, S. Moisio, A. Schmidt-Küntzel, A. Cornet 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: 253–256.


<table>
<thead>
<tr>
<th><strong>Default population</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective population size</td>
<td>100</td>
</tr>
<tr>
<td>Number of generations of random mating since QTL mutation occurred</td>
<td>100</td>
</tr>
<tr>
<td>Number of markers genotyped</td>
<td>10, 20</td>
</tr>
<tr>
<td>Number of alleles per marker in founder population</td>
<td>2</td>
</tr>
<tr>
<td>Initial marker / QTL allele frequencies in founder population</td>
<td>0.5 / 0.005</td>
</tr>
<tr>
<td>Distance (cM) between adjacent markers</td>
<td></td>
</tr>
<tr>
<td>10 markers</td>
<td>1, 0.5, 0.25</td>
</tr>
<tr>
<td>20 markers</td>
<td>0.5, 0.25, 0.125</td>
</tr>
<tr>
<td>Position of QTL</td>
<td></td>
</tr>
<tr>
<td>10 markers</td>
<td>Halfway between markers 5 and 6</td>
</tr>
<tr>
<td>20 markers</td>
<td>Halfway between markers 10 and 11</td>
</tr>
<tr>
<td>Additive effect of QTL allele mutation</td>
<td>1</td>
</tr>
<tr>
<td>Residual standard deviation</td>
<td>1</td>
</tr>
<tr>
<td>Number of individuals (records) in final generation</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Two-breed cross</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of generations of random mating following the initial cross</td>
<td>1, 5, 10, 20, 100</td>
</tr>
<tr>
<td>Initial marker / QTL allele frequencies in founder population</td>
<td></td>
</tr>
<tr>
<td>Breed 1</td>
<td>0.5 / 0.1, 0.9</td>
</tr>
<tr>
<td>Breed 2</td>
<td>0.5 / 0.9, 0.1</td>
</tr>
<tr>
<td>Distance (cM) between adjacent markers</td>
<td></td>
</tr>
</tbody>
</table>
10 markers 1
20 markers 0.5

**Non-central QTL position**

Distance (cM) between adjacent markers

10 markers 1
20 markers 0.5

Position of QTL

10 markers Halfway between markers 3 and 4
20 markers Halfway between markers 6 and 7

**Random founder allele frequencies**

Initial marker / QTL allele frequencies in founder range from 0.2 – 0.8 / 0.005

Distance (cM) between adjacent markers

10 markers 1
20 markers 0.5

**“Worst case scenario”**

Number of generations of random mating following the initial cross 10

Initial marker / QTL allele frequencies in founder population range from 0.2 – 0.8 / 0.1,

Breed 1 0.9
Breed 2 range from 0.2 – 0.8 / 0.9,

0.1

Distance (cM) between adjacent markers

10 markers 1
20 markers
Position of QTL
10 markers
Halfway between markers 3 and 4
20 markers
Halfway between markers 6 and 7

*Parameters for alternative populations are the same as the default except for those specified here*
### TABLE 2

**Least squares mean absolute difference (cM) of QTL position estimates**

for four mapping methods using 10 or 20 markers under the default scenario

<table>
<thead>
<tr>
<th>Marker spacing (cM)</th>
<th>Method</th>
<th>10*</th>
<th>20</th>
<th>10</th>
<th>20</th>
<th>10</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (0.5)**</td>
<td>SL</td>
<td>1.48</td>
<td>1.14</td>
<td>1.57</td>
<td>1.58</td>
<td>1.35</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>SL2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5 (0.25)</td>
<td>SL</td>
<td>0.78</td>
<td>0.63</td>
<td>0.83</td>
<td>0.81</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>SL2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25 (0.125)</td>
<td>SL</td>
<td>0.45</td>
<td>0.38</td>
<td>0.45</td>
<td>0.44</td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>SL2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HAP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IBD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The mean absolute difference of the QTL position estimate from its true position for each mapping method (SL - regression on a single marker, SL2 - regression on two markers, HAP - regression on a two-locus haplotype, IBD - likelihood based on haplotypes) used in populations created under the default scenario. The QTL is located in the center of the haplotype.

* Indicates the number of markers genotyped and used in the model.

** Distances without parentheses are for methods with 10 markers, while those inside parentheses are for methods with 20 markers.

*a, b, c, d, e* For a given marker spacing, least squares means with different superscripts are significantly different ($p < 0.05$)
TABLE 3
Least squares mean absolute difference (cM) of QTL position estimate for mapping methods with 1 cM marker spacing in a two-breed cross followed by random mating

<table>
<thead>
<tr>
<th>Generations of random mating</th>
<th>Method</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SL</td>
<td>SL</td>
</tr>
<tr>
<td>10*</td>
<td>2.34&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.11&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>20</td>
<td>2.27&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.97&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>10</td>
<td>2.35&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.16&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>5</td>
<td>2.48&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.28&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>1</td>
<td>2.51&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.47&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

The mean absolute difference of the QTL position estimate from its true position for each mapping method (SL – regression on a single marker, IBD – likelihood based on haplotypes) used in populations created under the crossbred scenario. The position of the QTL is the center of the haplotype, and the effective population size is 100.

* Indicates the number of markers genotyped and used in the model.

<sup>a, b</sup> For a given number of generations, least squares means with different lettered superscripts are significantly different ( p < 0.05 )
### TABLE 4

Least squares mean absolute difference (cM) of QTL position estimate and bias (cM) for mapping methods in three alternate scenarios

<table>
<thead>
<tr>
<th>Alternate scenario</th>
<th>Marker spacing</th>
<th>Method</th>
<th>SL (cM)</th>
<th>SL (cM)</th>
<th>IBD (cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-central QTL position</td>
<td>1 (0.5)**</td>
<td>LSMD</td>
<td>1.54</td>
<td>1.14</td>
<td>1.38</td>
</tr>
<tr>
<td>Random founder</td>
<td>1 (0.5)</td>
<td>bias</td>
<td>0.63</td>
<td>0.36</td>
<td>0.51</td>
</tr>
<tr>
<td>allele frequencies</td>
<td>1 (0.5)</td>
<td>LSMD</td>
<td>1.44</td>
<td>1.18</td>
<td>1.36</td>
</tr>
<tr>
<td>“Worst case scenario”</td>
<td>1 (0.5)</td>
<td>bias</td>
<td>-0.09</td>
<td>0.02</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSMD</td>
<td>2.67</td>
<td>2.43</td>
<td>2.45</td>
</tr>
</tbody>
</table>

The mean absolute difference of the QTL position estimate from its true position and bias for each mapping method (SL – regression on a single marker, IBD – likelihood based on haplotypes) used in populations created under three alternate scenarios.

* Indicates the number of markers genotyped and used in the model.

** Distances without parentheses are for IBD with 10 markers, while those inside parentheses are for models with 20 markers.

\( a, b, c \) For a given alternate scenario, least squares means with different lettered superscripts are significantly different (\( p < 0.05 \) ).
CHAPTER 3. OPTIMAL HAPLOTYPE STRUCTURE FOR LINKAGE DISEQUILIBRIUM-BASED FINE MAPPING OF QUANTITATIVE TRAIT LOCIS

A paper submitted to Genetics

L. Grapes¹, M. Z. Firat², J. C. M. Dekkers¹, M. F. Rothschild¹, R. L. Fernando¹,³

ABSTRACT

A linkage disequilibrium-based method for fine mapping quantitative trait loci (QTL) has been described that uses similarity between individuals’ marker haplotypes to determine if QTL alleles are identical by descent (IBD) in order to model covariance. It was found that incorporating all markers into the haplotype did not result in the highest mapping accuracy. To determine an optimal haplotype structure for this IBD-based method, haplotypes consisting of one, two, four, six or all ten available markers were fit in ideal and non-ideal simulated population conditions. It was found that using a haplotype of four or six markers as a sliding “window” resulted in the greatest mapping accuracy in nearly all conditions. Fitting one marker as the haplotype resulted in the worst mapping accuracy in all conditions. In conclusion, for fine mapping, marker information must be used in a manner that results in sensitivity of IBD probabilities to the putative position of the QTL while maintaining power to detect the QTL. A haplotype of four markers best fits these criteria. Thus for populations

¹ Graduate student and Professors, respectively, Department of Animal Science, Iowa State University
² Postdoctoral associate, Department of Animal Science, Iowa State University and Faculty of Agriculture, Department of Animal Science, Akdeniz University, Antalya, Turkey
³ Author for correspondence
similar to those simulated here there is an optimal haplotype size for this IBD-based fine mapping method.

INTRODUCTION

In a previous study of methods for fine mapping a quantitative trait locus (QTL) using linkage disequilibrium (LD), a haplotype-based method was compared to a single marker-based method (Grapes et al. 2003). The haplotype-based method was derived from the fine mapping method proposed by Meuwissen and Goddard (2000), which maps QTL to the center of a marker bracket by modeling the covariance between individuals based upon the similarity of their haplotypes. Individuals with similar marker haplotypes will likely share QTL alleles that are identical by descent (IBD) and so will have a higher covariance. Grapes et al. (2003), however, showed that regression on a single, bi-allelic marker had similar or greater mapping precision as this IBD method. The ability of single marker-based analysis to perform as well as haplotype-based analysis was also shown recently by Zhang et al. (2003), who found that results from a variance components analysis (Abecasis et al. 2000) for association with immunoglobulin E levels were similar regardless of whether single markers or a haplotype of three, four or five markers was used.

In the IBD method, the covariance between individuals is proportional to the probability that the individuals are IBD at a QTL. These IBD probabilities can be obtained using the gene drop method described by Meuwissen and Goddard (2000) under a set of assumptions about the population history. Grapes et al. (2003) observed that the IBD probabilities were not sensitive to the putative position of the QTL, such that probabilities for adjoining QTL positions (i.e. marker brackets) were similar across a 10-locus haplotype. Thus, considering all available markers simultaneously in a haplotype may not be optimal for fine mapping a QTL. The objective of this study was to determine the optimal haplotype size
for fine mapping. In this work, haplotype size will refer to the number of markers in the haplotype, which will be used as a sliding “window” across a previously identified QTL region containing 10 total markers.

Meuwissen and Goddard (2000) showed that the IBD method is quite robust to departures from assumptions about the population history when all available markers are considered in the haplotype. However, these assumptions may affect mapping precision when a smaller haplotype is considered. Thus the IBD method using marker haplotypes of various sizes was also evaluated for alternative population structures.

METHODS

Population Simulations

Following Meuwissen and Goddard (2000), it was assumed that a previous linkage analysis had mapped a QTL to a region of 2.25 to 9 cM, and 10 bi-allelic markers were available in that region. Thus in all simulations, individuals were generated with 10 evenly spaced, bi-allelic markers, a QTL centered between two adjacent markers, and a trait phenotypic value, which was the sum of the QTL genotypic value and an independent random normal error.

Default population: The IBD method of Meuwissen and Goddard (2000) is based upon modeling the covariance between individuals under the following assumptions: 1) each marker locus has two alleles with equal frequencies in the founder population, 2) variation in a QTL is due to a mutation that occurred 100 generations ago and 3) effective population size was 100 during the last 100 generations. Data in the default population were generated under these assumptions with the QTL placed in the middle of the 10-marker haplotype. It was assumed known which markers were maternally and paternally inherited so that marker haplotypes could be constructed.
Phenotypic values for individuals in the final generation were generated similar to Meuwissen and Goddard (2000). In all simulated populations, except for a crossbred population that will be described later, QTL alleles were uniquely numbered in the founders. So with an effective population size of 100, the initial frequency of each QTL allele was 0.005. In all simulations, including the crossbred, one QTL allele with a frequency higher than 0.1 in the final generation was randomly selected to be the mutant allele. This mutant allele was given an additive genetic value of 1, and the value of all other QTL alleles was set to 0. The phenotypic value for each individual in the final generation was calculated by summing the QTL allele effects with a random error sampled from $N(0, 1)$.

**Alternative populations:** To test robustness of the method to population history assumptions, six alternative populations were created that differed from the default for one or more conditions. In the first population, two breeds with divergent allele frequencies for two QTL alleles were crossed (see Table 1). After crossing, the population was randomly mated for 1, 5, 10, 20 or 100 generation(s). In the second population, the QTL was placed at a non-central position in the haplotype. In the third population, marker allele frequencies were assigned at random in the founder generation within a range of 0.2 to 0.8. In the fourth population, a "worst case scenario" was created that incorporated all three alternative conditions listed above. In the fifth population, effective population size was 50 or 200, instead of the default size of 100. In the last population, size of the QTL effect was decreased from 1 to 0.5. Details of all population parameters are summarized in Table 1.

**Maximum-likelihood estimation:** For haplotypes with one, two, four or six markers, phenotypic data of the final generation for a single trait were modeled as

$$y = Xb + Zh + e,$$

(1)
where $y$ is a vector of phenotypic values, $b$ is a vector of fixed effects, which here reduces to the overall mean, $X$ is the incidence matrix for $b$, which reduces to a vector of ones, $h$ is the vector of random haplotype effects, $Z$ is the incidence matrix for $h$, and $e$ is the vector of residuals. The variance-covariance matrix of residuals is $\text{Var}(e) = I \sigma_e^2$, where $I$ is the identity matrix. The variance of the vector of haplotype effects is $\text{Var}(h) = H_p \sigma_h^2$, where $H_p$ contains the IBD probabilities for the QTL at position $p$. For a haplotype containing 10 markers, there are 1,024 possible haplotypes, requiring $Z$ and $h$ to be large. So for 10 markers, phenotypic data were analyzed by an equivalent model as described in Grapes et al. (2003).

The IBD probabilities for the QTL based upon marker haplotypes with one, two, four or six markers were calculated using the analytical method of Meuwissen and Goddard (2001). The IBD probabilities for the QTL based upon a 10-marker haplotype were available from a previous study (Grapes et al. 2003), which used the gene drop method. Meuwissen and Goddard (2001) showed that the analytical and the gene drop methods give nearly identical results. All IBD probabilities were calculated assuming population history conditions of the default scenario for each position that the QTL could take within the haplotype window. These positions were the center of each marker bracket within the haplotype window.

Under multivariate normality, the residual loglikelihood for the model given by equation (1) is

$$L(p, \sigma_a^2, \sigma_e^2) \propto -0.5 \left( N \left( \ln \sigma_e^2 \right) - \ln |H_p \sigma_h^2| - \ln |C| -\left( y' R^{-1} y - \hat{\Theta}' X' R^{-1} y \right) \right)$$

(Searle, 1979) where $N$ is the number of phenotypic observations, $C$ is the coefficient matrix of the mixed model equations, $y$ is the vector of phenotypic values, $R = I \sigma_e^2$, and $\hat{\Theta}$ is the vector of solutions to the mixed model equations. For every putative QTL position, $p$, in the
haplotype window, the loglikelihood was maximized with respect to the variance components $\sigma_a^2$ and $\sigma_e^2$. When all 10 markers were fitted as the haplotype, the residual loglikelihood was obtained as described in Grapes et al. (2003). When one marker was fitted as the haplotype window, the QTL position was estimated at the marker locus, but for all other haplotype sizes, the QTL position was estimated at the center of a marker bracket. With haplotype windows of 4 or 6 markers, likelihoods were calculated multiple times for most marker brackets as a result of sliding the window across the 10-marker region. For these marker brackets, the highest likelihood was kept. Regardless of the haplotype size considered, the position with the highest loglikelihood overall was the estimated position of the QTL. Each scenario was replicated 1,000 times for each marker haplotype window size.

**Comparison of methods:** To evaluate the ability of the IBD method to estimate the QTL position using marker haplotypes of various sizes, absolute differences between the estimated and true QTL positions were obtained for each replicate of a scenario as

$$\text{absolute difference} = |\hat{\Theta}_i - \Theta|$$

where $\hat{\Theta}_i$ and $\Theta$ are the estimated and true QTL positions in cM for replicate $i$.

Bias of estimates of QTL position was estimated by

$$\text{bias} = \frac{\sum_{i=1}^{n} \hat{\Theta}_i}{n} - \Theta,$$

where $n$ is the number of replicates performed.

Absolute differences for all replicates of a simulation were analyzed using ANOVA (JMP ver. 5.0, SAS Institute, Inc.) with haplotype size fit as a fixed effect. Although absolute differences are not normally distributed, ANOVA is known to be robust when the sample size is large, as in this study. The least-squares mean of the absolute differences
(LSMD) was obtained for each haplotype size and was used as a measure of a method's ability to estimate the position of the QTL. A method with a smaller LSMD is preferable.

RESULTS

Comparison under the default scenario: After 100 generations of random mating, marker informativeness was similar across the chromosomal region. The polymorphism information content (Botstein et al. 1980) of each marker, averaged across 1,000 replicates of the default simulation, ranged between 0.23 and 0.25 for all marker spacings, which is 61 to 67% of the maximum for bi-allelic markers. Also, across 1,000 replicates of the default population, the probability of a marker locus being fixed ranged between 0.13 and 0.16 regardless of marker spacing. Thus, variability of marker informativeness likely had little impact on mapping accuracy.

The IBD method was used to fine map a QTL using a window of one, two, four, six or ten markers in a haplotype. The average LSMD across haplotype sizes was 1.32 cM when marker spacing was 1 cM (Table 2). At this marker spacing, an average estimate deviated from the true QTL position by less than two markers or marker brackets from the QTL regardless of haplotype size. Mapping resolution increased proportionately as the marker spacing decreased; average LSMD across haplotype sizes was 1.32, 0.69, and 0.39 cM for marker spacings of 1, 0.5, and 0.25 cM, respectively. Bias of QTL position estimates was close to zero (-0.4 to 0.1 cM) for all haplotype sizes under the default conditions. A bias of zero was expected because the QTL was positioned in the center of the chromosomal region.

Depending on marker spacing, a haplotype of four or six markers resulted in the greatest mapping precision (Table 2). When markers were 1 cM apart, a haplotype of six markers had the smallest LSMD, but it was not significantly different from the LSMD of a haplotype with four markers (Table 2). When markers were 0.5 cM apart, a four-marker
haplotype had the smallest LSMD, but it was not significantly different from a haplotype with all 10 markers. However, when marker spacing was reduced to 0.25 cM, the LSMD for the four-marker haplotype was significantly smaller than the LSMD for any other haplotype size. Interestingly, LSMD was largest when one marker was fitted regardless of marker spacing (Table 2).

**Two-breed cross followed by random mating:** Two breeds were simulated, each of effective size 100, that had the same two QTL alleles but at different frequencies (see Table 1). The number of generations of random mating that occurred after the initial cross of the two breeds ranged from 1 to 100. Marker spacing was 1 cM and the QTL was at the center of the 10-marker haplotype. Results are in Table 3.

Mapping precision using any haplotype size was negatively affected by the introduction of population admixture. Even when 100 generations of random mating followed the cross, the LSMDs were all higher than in the default population (Tables 2 and 3). Bias remained small, though, ranging from -0.19 to 0.15 cM. As the number of generations of random mating decreased, LSMD tended to increase for all haplotype sizes. However, when the number of generations decreased from 100 to 20, LSMD decreased for all haplotype sizes. This may be due to the fact that there were initially only two QTL alleles in this population, and after 100 generations of mating the QTL alleles attained extreme frequencies. In 36% of the replicates with 100 generations of random mating following the cross, QTL allele frequencies were greater than 0.85 or between 0.15 and 0.1, compared to 2% of the replicates with 20 generations of random mating, resulting in lower mapping resolution. QTL alleles never became extreme following 1 or 5 generations of random mating and became extreme in only 0.4% of the replicates having 10 generations of random mating.
Regardless of the number of generations of random mating that followed the cross, haplotypes with four and six markers had numerically smaller LSMDs than other haplotype sizes (Table 3). The LSMDs of four and six-marker haplotypes were significantly smaller than LSMD for a haplotype of ten markers when 1, 5 or 100 generations of random mating followed the cross (Table 3). Also, similar to the default, use of one marker consistently resulted in significantly larger LSMD than any other haplotype size.

**Non-central QTL position:** In this population, the QTL was positioned halfway between markers three and four instead of its central position between markers five and six as in the default scenario. Results for each haplotype size with marker spacing of 1 cM are presented in Table 4.

The LSMDs for nearly all haplotype sizes increased when the QTL was positioned towards the beginning of the chromosomal region instead of at the center (Tables 2 and 4). A haplotype with 2 markers was least affected by a non-central QTL, as its LSMD increased by only 0.01 cM, while a single marker haplotype was most affected, with an increase in LSMD of 0.07 cM (Tables 2 and 4). Interestingly, the LSMD for a four-marker haplotype decreased by 0.06 cM as compared to the default, while the LSMD for six and ten-marker haplotypes increased slightly (0.03 and 0.02 cM, respectively) (Tables 2 and 4). While the LSMD of the four-marker haplotype was not significantly different from that of six markers under this alternate scenario, it was numerically the smallest for all haplotype sizes.

Bias was observed in all cases, due to the non-central position of the QTL. Although the four-marker haplotype had the smallest LSMD, its bias was intermediate (Table 4). Also, as the size of the haplotype increased, bias consistently decreased.

**Variable marker allele frequencies:** In all previous populations, initial frequency of marker alleles was 0.5. Here marker allele frequencies in the founders were randomly set at
each marker locus within a range of 0.2 and 0.8. The LSMDs for each haplotype size, for a marker spacing of 1 cM, are in Table 4.

In this population, haplotype size was a determining factor in the effect of variable marker allele frequencies on LSMD. The LSMDs for all haplotype sizes increased as compared to the default, except for a 10-marker haplotype (Tables 2 and 4). The LSMDs of haplotypes with one and six markers increased by more than 0.1 cM, while LSMDs of haplotypes with two and four markers increased by ~ 0.5 cM with variable marker allele frequencies. Although LSMD increased in nearly all cases, the bias for all haplotype sizes remained close to zero, ranging from 0.038 to -0.034 cM (Table 4). Comparing haplotype sizes, the four-marker haplotype had numerically smallest LSMD, followed by six markers, but both were not significantly different from two or ten marker haplotypes. This result differs from the default population in which both four and six markers were significantly better than using all 10 markers (Table 2). So, it appears that sensitivity of the IBD method to marker allele frequencies depends on the size of the haplotype that is considered.

"Worst case scenario": The previous alternative populations differed from the default by only one condition. Here, several conditions were changed from the default population to create a “worst case scenario”. First, the two breeds described previously were crossed, followed by ten generations of random mating. Second, the QTL was positioned between marker loci three and four. Third, marker frequencies of the founders were set at random, as described previously.

The LSMD of all methods increased drastically for this “worst case scenario” compared to the default (Tables 2 and 4). The average LSMD increased from 1.32 cM under the default population to 2.50 cM. Biases also increased markedly, from a range of -0.04 to 0.1 cM in the default, to a range of 1.55 to 1.91 cM in the “worst case scenario” (Table 4).
Direction of the bias was towards the center of the chromosome for all haplotype sizes. The large bias and near doubling of LSMD when compared to the default are unique to this alternative population. However, when comparing LSMD across haplotype sizes, the ranking of haplotype sizes was not unique. Here haplotypes of size four and six had the smallest LSMDs, though not significantly different from a haplotype size of ten (Table 4). Again, using one marker resulted in the largest LSMD, although it was not significantly different from the LSMD of two markers.

**Alternative effective population size and smaller QTL effect:** To test the general power of the IBD method using different haplotype sizes, the effective population size and the size of the QTL allele effect were changed from the default population (Table 1). With effective population size reduced to 50, the LSMD increased for each haplotype size as compared to the default (Tables 2 and 4). However, a haplotype size of four once again had smaller LSMD than other haplotype sizes, although it was not significantly different from the LSMD of a haplotype of six markers. When effective population size was increased to 200 individuals, the LSMD decreased for each haplotype size, compared to the default (Tables 2 and 4). The LSMD for the one-locus haplotype was significantly worse than the LSMDs for the other haplotype sizes, which were not significantly different from each other (Table 4)

When the QTL effect was reduced by half, the LSMD increased for each haplotype size, compared to the default (Tables 2 and 4). In this case, haplotypes with four and six markers were significantly better than the three other haplotype sizes, which were all similar (Table 4). In these alternative populations involving effective population size and QTL effect, bias remained close to zero as expected (Table 4).
DISCUSSION

Identification of optimal haplotype size: As the ultimate goal of most QTL mapping studies is the identification of the gene and/or mutation underlying observed quantitative variation, it is important to identify mapping methods that provide the greatest resolution. By comparing mapping precision of the IBD method using various haplotype sizes, an apparent optimal haplotype size of four or six markers was identified for the populations simulated here. In populations that both adhered to assumptions about the population history for derivation of IBD probabilities and that violated those assumptions, using a haplotype of four or six markers resulted in the greatest mapping resolution, as evaluated by LSMD, although differences were not always significant. Only for the finest level of mapping, when markers were spaced 0.25 cM apart, was the four-marker haplotype significantly more accurate than the six-marker haplotype (Table 2). This marker spacing was, however, only evaluated for the default population; all alternative populations considered a marker spacing of 1 cM. Thus, it may be possible that a four-marker haplotype would also have been most accurate in the alternative populations if a marker spacing of 0.25 cM had been used.

In all populations, mapping resolution was consistently largest for the one-marker haplotype (Tables 2, 3 and 4). However, use of a single marker created inherent error that did not occur with other haplotype sizes because the estimated position of the QTL was at a marker locus, while the true position was between two markers. Thus, a single-marker haplotype is forced to have error in its position estimate and consequently its mapping resolution was reduced.

To estimate the amount by which mapping resolution was reduced for the single-marker haplotype, the QTL was fit at marker locus 6. Across 1,000 replicates of the default population, the LSMD for the one-marker haplotype was 1.26, 0.7 and 0.4 cM for 1, 0.5 and
0.25 cM marker spacing, respectively. These were significantly lower \((p < 0.05)\) than the LSMD of the one-marker haplotype that was obtained with the QTL at the center of the marker bracket for the three marker spacings (Table 2). With the adjusted QTL position, mapping resolution of the one-marker haplotype was not significantly different from any other haplotype size when marker spacing was 1 cM. At both 0.5 and 0.25 cM marker spacing, though, mapping resolution of the single-marker haplotype was significantly less \((p < 0.05)\) than that of the four-marker haplotype but similar to other haplotype sizes. Thus, the assumption that the QTL was centered between two markers significantly impacted the mapping resolution of the single-marker haplotype.

Robustness of the IBD method: Under ideal conditions and in four alternative populations (a non-central QTL, randomly assigned founder marker allele frequencies, variable effective population size and reduced QTL effect) the IBD method estimated the position of the QTL within two markers or marker brackets from its true position (Tables 2 and 4). This was true regardless of the haplotype size used. Only when LD was generated by crossing of two breeds was the mapping precision of the IBD method greatly reduced regardless of haplotype size (Tables 3 and 4). Thus, the IBD method in general is robust to violations of most assumptions about population history, except for recent crossing or migration.

Sensitivity of IBD probabilities: The reduced accuracy of position estimates when using all ten markers compared to smaller haplotype sizes is counterintuitive to the general notion that use of more information should result in better estimates. However, as already noted by Grapes et al. (2003), IBD probabilities are not sensitive to QTL position when all markers are used, which reduces mapping precision (Grapes et al. 2003). To demonstrate this, IBD probabilities were obtained for every putative QTL position within the 10-locus haplotype for all haplotype pair comparisons across 1,000 replicates of the default population, using a
haplotype window of one, two, four, or ten markers. When the haplotype size was four, IBD probabilities were obtained assuming the putative QTL position was at the center of the haplotype window, except for the endmost marker brackets, where the QTL was assumed to be either the first or last marker bracket in the haplotype window. For each haplotype size, correlations between the IBD probabilities for the true QTL position and all other putative QTL positions were estimated. With a haplotype size of one, the true QTL position was assumed to be at marker locus 5, otherwise the true QTL was centered between markers 5 and 6. The correlation between IBD probabilities at the true QTL position and QTL positions in flanking marker brackets was highest (~0.86) when all ten markers were fitted as the haplotype and lowest (~0.12) when only one marker was fitted (Figure 1). Thus, the true QTL position is best distinguished from its surrounding positions by fitting only one marker. With only one marker fitted as the haplotype, there are only two possible IBD probabilities that can be assigned to a QTL position, and there is no overlap in marker information when moving from one putative QTL position to the next. In contrast, when fitting all ten markers in the haplotype, the same set of markers is used for every putative QTL position. Also, there are cases where different haplotype states have similar or identical IBD probabilities associated with them. The high correlation between IBD probabilities results in high correlations between the estimated covariance of a pair of individuals at different QTL positions, which results in similar likelihoods across positions. This makes it more difficult to distinguish between QTL positions and obtain accurate position estimates.

Following the logic that less marker information allows greater distinction between putative QTL positions, fitting a single marker in the IBD method should result in greatest mapping accuracy, which is contrary to what was observed (Tables 2, 3, and 4). The ability to map a QTL is, however, also affected by the ability of the method to detect the QTL, i.e.
the power. The latter can be discerned from the alternative populations involving a smaller QTL effect and different effective population sizes (Table 4). As expected, when effective population size was increased to 200, thus increasing power, the performance of nearly all haplotype sizes was similar (Table 4). However, when power was reduced by either decreasing effective population size or by decreasing the size of the QTL effect, mapping accuracy was greatest for four and six-marker haplotypes (Table 4). This was unexpected, as the extra marker information in a six or ten-marker haplotype should have provided greater mapping resolution when power was lowered. Thus, it appears that a balance is required, such that the marker haplotype window must provide enough information to detect the QTL but also that the information be as variable as possible as the haplotype window is moved across the QTL region, such that an accurate position estimate can be obtained. The nature of the relationship between these two factors can be seen in Figure 2. The average likelihood across the chromosomal region is high when all ten markers are fitted as the haplotype, indicating that using ten markers results in high power to detect the QTL. However, there is not much difference in the average likelihood between the true QTL position at the center of the chromosomal region and the outermost positions (Figure 2). The shape of the mean likelihood curve of the six-marker haplotype size is similar to that of the ten-marker haplotype, although it is slightly more peaked (Figure 2). Now compare the likelihood curve for ten markers to that of two markers in Figure 2. The average likelihood when using two markers was much less than that of the ten or six-marker haplotypes, but there was a much greater distinction between the likelihood at the true QTL position and the outermost positions. Thus, a two-marker haplotype may have less power overall to detect the QTL, but it has a greater ability to provide an accurate position estimate as compared to using all marker information. By considering the relationship between detection and distinction of the
QTL position and then examining Figure 2, it seems that a four-marker haplotype is most favorable among the haplotype sizes considered here. By fitting four markers, enough information is provided such that the average value of the likelihood is high while the difference between the likelihood at the true QTL position and the outermost positions is greatest among all haplotype sizes. In fact, at the true QTL position, the average likelihood of a four-marker haplotype was nearly identical to that of six and ten-marker haplotypes. Although the results presented in Tables 2, 3 and 4 show that haplotypes of size four and six performed similarly, it may be that a four-marker haplotype is optimal for fine mapping using the IBD method, especially when markers are closely spaced.

**Accuracy of genetic prediction:** Although using ten markers was not favorable for discriminating between QTL positions, IBD probabilities were most accurate when all markers were considered, i.e. the correlation between the true IBD state of two QTL and the IBD probabilities obtained given the individuals' marker haplotype information was highest; 0.52 when ten markers were used, 0.5 when four markers were used, and 0.34 when one marker was used. Interestingly, the higher accuracy of IBD probabilities given ten marker loci did not result in more accurate estimates of the genetic value of an individual. In fact, at the true QTL position, the correlation between individuals' true genetic value and the best linear unbiased predictor (BLUP, Henderson 1973) of the genetic value was the same (~0.36) for haplotype sizes with four and ten marker, but lowest for a haplotype size of one (~0.27). Thus, although IBD probabilities were most accurate when all marker information was considered, this greater accuracy had little effect on the accuracy of BLUP estimates in this population.

**Comparison to regression-based fine mapping:** For populations similar to those considered here, Grapes *et al.* (2003) showed that regression on a single marker was as
effective for fine mapping QTL as the IBD method using ten markers. Since ten markers are not optimal for the IBD method, it is worth comparing the IBD method using other haplotype sizes to the regression method. For the regression method, the number of markers genotyped was doubled from 10 to 20 because the regression method does not require haplotype information and, therefore, does not require the genotype of parents (Grapes et al. 2003). Under the default population, none of the haplotype sizes had mapping precision that was significantly better than a single marker-based regression method with 20 markers genotyped (SL-20) at 1 and 0.5 cM marker spacing. However, when marker spacing was smallest, the IBD method using a four-marker haplotype was significantly better than all other haplotype sizes as well as SL-20 (data not shown). Under the alternative population involving a two-breed cross, mapping precision of SL-20 was similar to IBD with a four or six-marker haplotype when either 20 or 100 generations of random mating followed the cross. As the number of generations of random mating decreased, mapping precision of SL-20 decreased such that the four and six-marker haplotypes were both significantly better when one and five generations of random mating followed the cross (data not shown). In the population involving a non-central QTL, SL-20 had the greatest mapping precision but it was not significantly different from a four-marker haplotype (data not shown). When founder marker allele frequencies were set randomly, SL-20 did have significantly greatest mapping precision (data not shown). Finally, in the “worse case scenario”, mapping precision of SL-20 was not significantly different from a four, six or ten-marker haplotype but was greater than the mapping precision of a one or two-marker haplotype (data not shown). So, SL remained comparable to the IBD method even when the more favorable haplotype sizes were considered.
It is also worth comparing the IBD method using one marker to single marker regression when 10 markers are available (SL-10), as they used similar information for mapping. These methods use the same information but differ in the way the effects are modeled. The IBD method models the haplotype effect as random, and SL models the allelic effect of the marker as fixed. Under the default population, SL-10 and IBD with a single marker were not significantly different (data not shown). Under the two-breed cross, SL-10 and IBD with one marker were also not significantly different, except when 100 generations of random mating followed the cross, for which SL-10 had significantly higher mapping precision (data not shown). In the population having a non-central QTL and in the “worst case scenario”, SL-10 and IBD again were not significantly different. However, when founder marker allele frequencies were set randomly, SL-10 had significantly greater mapping precision (data not shown). Thus, the way in which information was modeled had little impact on mapping precision when a single marker was used, except in two specific cases. Once again, SL was comparable to the IBD method.

Optimal methods for fine mapping: Of the methods examined here and in Grapes et al. (2003), fitting a four-marker haplotype in the IBD method and SL-20 were the optimal methods for fine mapping a previously identified QTL. There is an advantage to using SL-20, though, as it does not require knowledge of haplotypes. However, if the IBD method is used for fine mapping, it would be preferable to fit a smaller haplotype instead of all available markers in the region. As seen here, using a haplotype with four or six markers as a sliding window across the region resulted in the greatest mapping accuracy compared to the other haplotype sizes tested.

The only published uses of the IBD method for fine mapping a QTL are Meuwissen et al. (2002) and Blott et al. (2003), which both incorporated the IBD method into analyses.
that combine linkage and LD information. Although 29 total markers were available, Blott et al. (2003) used a haplotype window of 16 markers and the method of Meuwissen and Goddard (2001) to estimate IBD probabilities. Meuwissen et al. (2002) utilized all 15 available markers to obtain IBD probabilities, also using the IBD method of Meuwissen and Goddard (2001). Further research is needed to determine if using smaller haplotype windows can improve the mapping accuracy of fine mapping methods that utilize IBD and combined linkage and LD information.

ACKNOWLEDGMENTS
This work was supported in part by funding from the United States Department of Agriculture-National Research Initiative, Sygen International, Monsanto Co., the Iowa Agriculture and Home Economics Experiment Station and by Hatch Act and State of Iowa funds. Laura Grapes was supported by a United States Department of Agriculture National Needs fellowship in quantitative and molecular genetics. The authors also wish to thank Dan Nettleton for his comments and contribution to this work.

LITERATURE CITED
Blott, S., J, Kim, S. Moisio, A. Schmidt-Kültzel, A. Cornet 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: 253–256.


TABLE 1

Parameters for default and alternative* simulated populations

<table>
<thead>
<tr>
<th>Default population</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective population size</td>
<td>100</td>
</tr>
<tr>
<td>Number of generations of random mating since QTL mutation occurred</td>
<td>100</td>
</tr>
<tr>
<td>Number of markers genotyped</td>
<td>10, 20</td>
</tr>
<tr>
<td>Number of alleles per marker in founder population</td>
<td>2</td>
</tr>
<tr>
<td>Initial marker / QTL allele frequencies in founder population</td>
<td>0.5 / 0.005</td>
</tr>
<tr>
<td>Distance (cM) between adjacent markers</td>
<td></td>
</tr>
<tr>
<td>10 markers</td>
<td>1, 0.5, 0.25</td>
</tr>
<tr>
<td>20 markers</td>
<td>0.5, 0.25, 0.125</td>
</tr>
<tr>
<td>Position of QTL</td>
<td></td>
</tr>
<tr>
<td>10 markers</td>
<td>Halfway between markers 5 and 6</td>
</tr>
<tr>
<td>20 markers</td>
<td>Halfway between markers 10 and 11</td>
</tr>
<tr>
<td>Additive effect of QTL allele mutation</td>
<td>1</td>
</tr>
<tr>
<td>Residual standard deviation</td>
<td>1</td>
</tr>
<tr>
<td>Number of individuals (records) in final generation</td>
<td>100</td>
</tr>
</tbody>
</table>

<p>| Two-breed cross |         |
| Number of generations of random mating following the initial cross | 1, 5, 10, 20, 100 |
| Initial marker / QTL allele frequencies in founder population |         |
| Breed 1 | 0.5 / 0.1, 0.9 |
| Breed 2 | 0.5 / 0.9, 0.1 |
| Distance (cM) between adjacent markers |         |</p>
<table>
<thead>
<tr>
<th>Distance (cM) between adjacent markers</th>
<th>10 markers</th>
<th>20 markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-central QTL position</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position of QTL</th>
<th>10 markers</th>
<th>20 markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position of QTL</td>
<td>Halfway between markers 3 and 4</td>
<td>Halfway between markers 6 and 7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random founder allele frequencies</th>
<th>Initial marker / QTL allele frequencies in founder</th>
<th>range from 0.2 – 0.8 / 0.005</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Distance (cM) between adjacent markers</th>
<th>10 markers</th>
<th>20 markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance (cM) between adjacent markers</td>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Worst case scenario</th>
<th>Number of generations of random mating following the initial cross</th>
<th>10</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Breed 1</th>
<th>Breed 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial marker / QTL allele frequencies in founder</td>
<td>range from 0.2 – 0.8 / 0.1,</td>
</tr>
<tr>
<td>population</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Distance (cM) between adjacent markers

10 markers 1
20 markers 0.5

Position of QTL

10 markers Halfway between markers 3 and 4
20 markers Halfway between markers 6 and 7

**Alternative effective population size**

Effective population size 50, 200

**Smaller QTL effect**

Additive effect of QTL allele mutation 0.5

*Parameters for alternative populations are the same as the default except for those specified here.*
TABLE 2

Least squares mean absolute difference (cM) of QTL position estimates obtained by the IBD method using different haplotype sizes under the default scenario

<table>
<thead>
<tr>
<th>Marker spacing (cM)</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.46 (^a)</td>
<td>1.32 (^{b,c})</td>
<td>1.25 (^{c,d})</td>
<td>1.20 (^d)</td>
<td>1.36 (^{a,b})</td>
</tr>
<tr>
<td>0.5</td>
<td>0.76 (^a)</td>
<td>0.70 (^b)</td>
<td>0.63 (^c)</td>
<td>0.70 (^b)</td>
<td>0.68 (^{b,c})</td>
</tr>
<tr>
<td>0.25</td>
<td>0.44 (^a)</td>
<td>0.38 (^b)</td>
<td>0.34 (^c)</td>
<td>0.40 (^b)</td>
<td>0.40 (^b)</td>
</tr>
</tbody>
</table>

The mean absolute difference of the QTL position estimate from its true position for the IBD mapping method used in populations created under the default scenario. The QTL is located in the center of the 10-marker haplotype.

\(^a, b, c, d\) For a given marker spacing, least squares means with different superscripts are significantly different (\( p < 0.05 \))
TABLE 3

Least squares mean absolute difference (cM) of QTL position estimate obtained from the IBD method using different haplotype sizes with 1 cM marker spacing in a two-breed cross followed by random mating

<table>
<thead>
<tr>
<th>Generations of random mating</th>
<th>Number of markers used in IBD method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>2.53&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>20</td>
<td>2.30&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>10</td>
<td>2.35&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>5</td>
<td>2.43&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>1</td>
<td>2.52&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

The mean absolute difference of the QTL position estimate from its true position for the IBD mapping method used in populations created under the crossbred scenario. The position of the QTL is the center of the 10-marker haplotype, and the effective population size is 100.<sup>1,2,3</sup> For a given number of generations, least squares means with different lettered superscripts are significantly different (<i>p < 0.05</i>)
<table>
<thead>
<tr>
<th>Alternate scenario</th>
<th>Number of markers used in IBD method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Non-central QTL position</td>
<td>LSMD</td>
</tr>
<tr>
<td>bias</td>
<td>0.62</td>
</tr>
<tr>
<td>Random founder allele frequencies</td>
<td>LSMD</td>
</tr>
<tr>
<td>bias</td>
<td>0.016</td>
</tr>
<tr>
<td>“Worst case” scenario</td>
<td>LSMD</td>
</tr>
<tr>
<td>bias</td>
<td>1.75</td>
</tr>
<tr>
<td>Effective population size = 50</td>
<td>LSMD</td>
</tr>
<tr>
<td>bias</td>
<td>-0.182</td>
</tr>
<tr>
<td>Effective population size = 200</td>
<td>LSMD</td>
</tr>
<tr>
<td>bias</td>
<td>0.017</td>
</tr>
<tr>
<td>Smaller QTL effect</td>
<td>LSMD</td>
</tr>
<tr>
<td>bias</td>
<td>-0.016</td>
</tr>
</tbody>
</table>

The mean absolute difference of the QTL position estimate from its true position and bias for the IBD mapping method used in populations created under six alternate scenarios with 1 cM marker spacing.

For a given alternate scenario, least squares means with different lettered superscripts are significantly different (p < 0.05)
FIGURE 1. – The correlation between IBD probabilities for the true QTL position and all other putative QTL positions estimated using haplotypes of one, two, four or ten markers. With a haplotype size of one, the true QTL position was at position 5. Results are based on 1,000 replicates of the default scenario with a marker spacing of 1 cM.

(▲) IBD-1, (x) IBD-2, (○) IBD-4, (■) IBD-10, (▼) QTL position
FIGURE 2. – The loglikelihood for each putative QTL position averaged across 1,000 replicates of the default scenario with marker spacing equal to 1 cM.

(▲) IBD-1, (x) IBD-2, (○) IBD-4, (●) IBD-6, (□) IBD-10, (▽) QTL position
CHAPTER 4. PHYSICAL AND LINKAGE MAPPING OF THE PORCINE

CONNEXIN 37 (CX37) GENE

A paper published in the Journal of Animal Science\textsuperscript{1}

L. Grapes\textsuperscript{2}, Y. Zhang\textsuperscript{3}, and M. F. Rothschild\textsuperscript{2,4}


Source and Description of Primers. A set of primers (F1, R1) was designed from the porcine connexin 37 (CX37) mRNA sequence (GenBank Accession no. X86024) as well as an additional reverse primer (R2) from human (GenBank Accession no. 6093424) and murine (GenBank Accession no. NM_008120) CX37 consensus sequence. The F1 and R2 primers were used to amplify porcine CX37 from genomic DNA. Using sequence obtained from the amplified product, an additional pig-specific forward primer (F2) was designed.

Primer Sequences. F1: 5'-TTC CTG GAG AAG CTG CTG GA-3'; R1: 5'-CGA GAT CTT GGC CAT CTG TC-3'; F2: 5'-ACT CGA CCG TGG TGG GCA A-3'; R2: 5'-GTG GTC AGG TTG GCC CAG TT-3'.

Method of Detection. A polymerase chain reaction (PCR) of 10 \( \mu \)L volume containing 1 \( \mu \)L PCR buffer, 1 \( \mu \)L MgCl\(_2\) [15 mM], 1 \( \mu \)L dNTPs [2 mM], 0.25 \( \mu \)L of each PCR primer (F1 and R2) [10 pM], 0.07 \( \mu \)L Promega Taq Polymerase (Madison, WI) and 5.43 \( \mu \)L H\(_2\)O was

\textsuperscript{1} Reprinted with permission of \textit{J. of Anim. Sci.}, 2002, 80(5), 1375-1376.
\textsuperscript{2} Graduate student and Professor, respectively, Department of Animal Science, Iowa State University
\textsuperscript{3} Postdoctoral associate, Present address - Animal Genetics and Breeding Unit, The University of New England, Armidale, New South Wales, Australia
\textsuperscript{4} Author for correspondence
used to assay 12.5 ng of genomic DNA from 4 individuals for each of five swine breeds (Landrace, Hampshire, Yorkshire, Berkshire and Meishan). An 872 bp fragment from within the single CX37 exon was amplified using primers F1 and R2 in a Robocycler (Stratagene, La Jolla, CA) under the following thermocycling conditions: initial denaturation at 94°C for 4 min, 35 cycles of 94°C for 45 sec, 62°C for 1 min, 72°C for 1 min 20 sec, and a final extension time of 9 min at 72°C. For each breed, the PCR products from the four individuals were pooled. These pools were then directly sequenced using dye terminators and an ABI 377 sequencer (Perkin-Elmer, Foster City, CA) at the Iowa State University DNA Sequencing and Synthesis Facility. The F1 and R1 primers produced a 399 bp fragment that was used for physical mapping. The F2 and R1 primers produced a 385 bp fragment that was used for linkage mapping. All mapping was completed using the thermocycling conditions described above. A PCR-RFLP marker was confirmed using the MluI enzyme, and this marker was used for linkage mapping and determining allele frequencies in a commercial population of pigs (see below). For the PCR-RFLP assays, 3 μL of PCR products were digested with 4 U MluI in 1x buffer solution with 1x BSA added. Total volume for the digestion reaction was 10 μL. Following digestion for 3 hours at 37°C, digested products were loaded into a 4% agarose gel containing ethidium bromide, electrophoresed and photographed.

Sequencing and Polymorphisms. The 872 bp exonic fragment amplified using F1 and R2 primers was sequenced and showed 97% homology to the 413 bp of pig CX37 mRNA sequence available from GenBank (GenBank Accession no. 6093424). Comparison of this fragment from the five breed pools revealed two single nucleotide polymorphisms, neither of which changed an amino acid. A single T to C base substitution was identified at position 366 from the ATG site based on human CX37 sequence (GenBank Accession no. AF132674)
and resulted in the formation of an MluI restriction site. A PCR-RFLP test was designed using this restriction site and products from the F2 and R1 primers. Resulting allelic fragment sizes from this test were 385 bp (allele 1) and 317 bp and 68 bp (allele 2) (Figure 1). The smaller fragment for allele 2 (68 bp) is not visible in this figure. The second polymorphism, a G to A base change, was located at position 345 from the human CX37 ATG site listed above but was not used for any mapping purposes.

*Inheritance Pattern.* In the five PiGMaP families (Archibald et al., 1995) the MluI PCR-RFLP segregated in accordance with autosomal Mendelian inheritance.

*Allele Frequencies.* Individuals (n=844) from four commercial populations consisting of Landrace, Large White, Duroc and Pietrain backgrounds were genotyped using the MluI PCR-RFLP assay. Allele 2 was the rarer allele with an average frequency of 0.14 (range 0.07-0.20).

*Chromosomal Location.* Physical mapping of CX37 was completed using the French pig/rodent somatic cell hybrid panel (Yerle et al., 1996) and products resulting from PCR using primers F1 and R1. Analysis of the PCR results was completed as previously described (http://www.toulouse.inra.fr/lgc/pig/pcr/pcr.htm) and placed CX37 on SSC6 q24-31 with probability of 0.87. Linkage mapping was performed using CRI-MAP (Green et al., 1990) analysis of genotypes for the PiGMaP families. Using two-point linkage analysis, four markers were found to be significantly linked to CX37. The markers were (LOD score and recombination frequencies in parentheses) S0031 (6.21, 0.10), S0059 (6.66, 0.09), S0228 (3.52, 0.08) and SW71 (7.50, 0.02). Using CRI-MAP, a multipoint map of SSC6 including CX37 was constructed and placed its location centromeric of S0059 by 6.3 centimorgans. These results confirmed the position obtained from physical mapping.
Comments. Gap junctions, channels formed between adjacent cell membranes, facilitate cellular communication by allowing ions and small molecules to transfer from cell to cell. Oocyte maturation and ovulation are thought to be regulated by metabolic cooperation between the oocyte and surrounding granulosa cells. Gap junctions form between the oocyte and granulosa cell processes (Anderson and Albertini, 1976) as well as between granulosa cells of the follicle (Gilula et al., 1978). It has been shown that connexin 37 is expressed in these gap junctions, and mice lacking CX37 were unable to produce mature follicles and developed false corpora lutea (Simon et al., 1997). Thus, proper function of CX37 appears to be critical for oocyte maturation and ovulation. Previously, CX37 was mapped to HSA1p35.1 in humans (Camp et al., 1995). Comparative mapping aligns this region to SSC6 q22-26 and q31-35. Our results are in agreement with this location.

Acknowledgements. We thank Jeannine Helm and Daniel Ciobanu for their advice and suggestions, Alan Archibald for use of the PiGMaP DNA and the ResPig database, and Martine Yerle for use of the French SCH panel. Support was provided by an USDA National Needs Fellowship in Animal Biotechnology and Quantitative Genetics, PIC International Group and the Iowa Agriculture and Home Economics Experimental Station, Ames, paper no. J-19292, project no. 3600, as well as by Hatch Act and State of Iowa funds.

Literature Cited


Geldermann, R. Davoli, D. Ruyter, E. Verstege, M. Groenen, W. Davies, B.
Hoyheim, A. Keiserud, L. Andersson, H. Ellegren, M. Johansson, L. Marklund, J.
Miller, D. Anderson Dear, E. Signer, E., A. Jeffreys, C. Moran, P. Le Tissier,
Muladno, M. Rothschild, C. Tuggle, D. Vaske, J. Helm, H. C. Liu, A. Rahman, T.P.
Yu, R. G. Larson, and C. Schmitz. 1995. The PiGMaP consortium linkage map of the
Willems. 1995. The gene for human gap junction protein connexin37 (GJA4) maps to
chromosome lp35.1, in the vicinity of D1S195. Genomics 30:402-203.
Washington University School of Medicine, St. Louis, MO.
Yerle, M., G. Echard, A. Robic, A. Mairal, C. Dubut-Fontana, J. Riquet, P. Pinton, D. Milan,
Y. Lahbib-Mansais, and J. Gellin. 1996. A somatic cell hybrid panel for pig regional

Key Words: Gap Junctions, Gene Mapping, Ovulation Rate, Pigs
Figure 1. *MluI* PCR-RFLP of the connexin 37 gene. Lane M includes the 1-kb ladder with predicted sizes indicated on the left. Lane 1 includes undigested PCR product using the F2 and R1 primers. Lane 2 indicates the 1,1 genotype, lane 3 indicates the 1, 2 genotype and lane 4 indicates the 2, 2 genotype. Fragment sizes of alleles 1 and 2 are listed on the right.
CHAPTER 5. BMP15 MAPS TO THE X CHROMOSOME IN SWINE

A paper published in Animal Genetics

L. Grapes and M. F. Rothschild

Source/description: Bone morphogenetic factor 15 (BMP15), also known as GDF-9B (growth differentiation factor-9B), is a member of the transforming growth factor-β (TGF-β) family and is expressed in oocytes during follicular development. Its exact function is still unclear, but follicular growth after the primary stage is arrested in sheep homozygous for mutations in BMP15, while those with only one affected copy have increased ovulation rate. Thus, BMP15 appears to be essential for proper follicular development.

PCR conditions: Polymerase chain reactions (PCR) were performed using 1X PCR buffer, 1.5 mM MgCl₂, 200 μM each dNTP, 3 pmol each PCR primer, and 0.35 units of Taq DNA polymerase (Promega). Using this protocol, both exons 1 and 2 of porcine BMP15 were amplified. The PCR cycling conditions to amplify exon 1 included an initial denaturation of 4 min at 94 °C, followed by 40 cycles of 94 °C for 45 s, 55 °C for 30 s, 72 °C for 45 s, and a final extension of 72 °C for 3 min. The cycling conditions to amplify exon 2 included an initial denaturation of 4 min at 94 °C, followed by 35 cycles of 94 °C for 45 s, 55 °C for 1

---

1 Reprinted with permission of Animal Genetics, 2002, 33(2), 165-166.
2 Graduate student and Professor, respectively, Department of Animal Science, Iowa State University
3 Author for correspondence
min, 72 °C for 1 min, and a final extension of 72 °C for 5 min. Both fragments were amplified using a PTC-200 Peltier Thermal Cycler (MJ Research, Inc.).

**PCR primers:** Preliminary genomic sequence was obtained for porcine BMP15 using primers described in Galloway et al. (2000) which amplify a portion of sheep exon 2 (F: 5’-AGAGCCACTGTGTTTACCAGC-3’; R: 5’-TTCCTGGGAACCTTGAGATAGC-3’). From this sequence, pig-specific primers were designed that amplified a 614 bp fragment of exon 2 (GenBank accession no. AY050270) (F: 5’-AAAGCCTTCCCTGCTGCC-3’; R: 5’-TCCCATTTGCCTCAATCA-3’). Primers were also designed to obtain porcine sequence for exon 1 (GenBank accession no. AY050269) using consensus sequence from BMP15 exon 1 of human (GenBank accession no. AF082349) and sheep (GenBank accession no. AF236078) (F: 5’-AAACATAGGACCTGCCTGCC-3’; R: 5’-ATGGTGCCTTTCCCTA-3’).

**Sequence Analysis:** PCR products consisting of approximately 1125 bp of the porcine BMP15 coding region and approximately 150 bp of the 5’ UTR were amplified from genomic DNA for individuals from several swine breeds. Individual PCR products were pooled by breed and sequenced using dye terminators and an ABI 377 sequencer (Perkin-Elmer). Porcine exon 1 showed 88% sequence identity in a 420 bp overlap to sheep exon 1 sequence, as well as 81% to human in a 344 bp overlap. Porcine exon 2 had 90% sequence identity to sheep exon 2 (GenBank accession no. AF236079) in an 855 bp overlap and 82% sequence identity in an 824 bp overlap with human exon 2 (GenBank accession no. AF082350). None of the sequences obtained for exons 1 and 2 were found to be
polymorphic. Linkage mapping was therefore not possible. The lack of polymorphisms in multiple breeds sequenced for exons 1 and 2 suggests that the coding region of \textit{BMP15} is highly conserved in the pig.

\textit{Chromosomal assignment:} Since linkage mapping was not possible, a chromosomal location could only be obtained utilizing the pig-rodent somatic cell hybrid panel\textsuperscript{5}. Since amplification of PCR products similar in size to the porcine exon 2 fragment was seen using both mouse and hamster genomic DNA, a restriction enzyme test was designed to discriminate porcine PCR products from rodent. A specific \textit{StuI} site was identified in the porcine exon 2 PCR product resulting in two bands of sizes 374 bp and 240 bp. Only clones with digested products identical to the pig results were considered positive. Analysis of the PCR/digest results was completed as described (http://www.toulouse.inra.fr/lgc/pig/pcr/pcr.htm) and placed \textit{BMP15} on SSC X with chromosomal probability 1.0 and in region \textit{p11-13} with probability 0.995 and error risk less than 0.1\%.

\textit{Comparative mapping:} While the number of comparative loci between HSA X and SSC X are quite small, it seems that few chromosomal rearrangements have occurred during evolution\textsuperscript{6}. The physical map positions for many of the genes on SSC X are nearly identical to their positions on HSA X\textsuperscript{7}. In humans, \textit{BMP15} maps to Xp11.2, which corresponds directly with the porcine map location reported here. This further supports the idea of a highly conserved X chromosome.

Interestingly, the human \textit{BMP15} physical position is in one of the key Turner Syndrome intervals (Xp11.2-11.4), a disease known to cause a reduction in female fertility,
and it has been regarded as a good candidate gene based on the effect in sheep. However, there have been a number of studies in human populations that have failed to find any modification of the gene in women with POF (premature ovarian failure) or reduced fertility.

Acknowledgements: We thank Jeannine Helm, Carole Sargent and Graham Plastow for their advice and suggestions and Martine Yerle for use of the French somatic cell hybrid panel. Support was provided by an USDA National Needs Fellowship in Animal Biotechnology and Quantitative Genetics, PIC International Group and the Iowa Agriculture and Home Economics Experimental Station, Ames, paper no. J-19509, project no. 3600, as well as by Hatch Act and State of Iowa funds.

References
CHAPTER 6. PROSPECTING FOR PIG SNPS IN THE HUMAN GENOME: HAVE WE STRUCK GOLD?

A paper submitted to Nature Genetics

Laura Grapes¹, Stephen Rudd²,³, Rohan L. Fernando¹, Karine Megy⁴, Dominique Rocha⁴ & Max F. Rothschild¹,⁵

ABSTRACT

Gene-to-gene variation in the frequency of single nucleotide polymorphisms (SNPs) has been observed in humans, mice, primates and pigs, but a relationship across species in this variation has not been described. Here, the gene-specific frequencies of porcine coding SNPs (cSNPs) identified by in silicio methods and murine cSNPs were compared to the frequency of human cSNPs across homologous genes, resulting in a human-pig correlation of 0.77 (p < 0.00001) and a human-mouse correlation of 0.48 (p < 0.0005). This is the first evidence of conserved gene-to-gene variability in cSNP frequency across species and indicates that site-directed screening of porcine genes homologous to cSNP-rich human genes may rapidly advance SNP discovery in pigs.

¹ Graduate student and Professors, respectively, Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University
² Centre for Biotechnology, Tykistökatu 6, FIN-20521, Turku, Finland
³ Institute for Bioinformatics, GSF-National Research Center for Environment and Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
⁴ Sygen International plc, University of Cambridge, Dept. of Pathology, Cambridge, CB2 1QP, UK
⁵ Author for correspondence
INTRODUCTION

Nucleotide diversity is thought to be affected by variation in local mutation rates and recombination events\(^1,2\), and natural selection forces\(^3\), and has been found to be highly variable across the human genome\(^4\), even when comparing coding regions of genes\(^5,6\). Similar observations about variation in nucleotide diversity have been made in mouse\(^7\) and, on a smaller scale, in chimpanzee\(^8\) and pig\(^9\). With the high degree of coding and protein sequence identity between humans and pigs, the forces affecting nucleotide diversity in humans may affect pigs similarly. Thus, the cSNP densities of humans and pigs may be similar when compared on an individual gene basis. Conversely, human and mouse cSNP densities should be less similar than that of human and pig, as humans and mice have a lower level of sequence identity on average.

RESULTS

In silicio porcine SNP identification and EST annotation

Unlike humans and mice, pigs do not have a large repository of identified SNPs and access to such information in the near future is unlikely. Sequencing the porcine genome would allow large-scale SNP detection, and there is a joint sequencing project between Denmark and China, however it is unclear when results from this project will be made public (Fredholm, personal communication). So, to obtain the necessary SNPs to perform a comparison of nucleotide diversity, in silicio methods were utilized for identifying putative SNPs from pig sequences. All porcine ESTs (~150,000) were downloaded from EMBL (ftp://ftp.ebi.ac.uk/pub/databases/dbEST/) and assembled into strict unigenes using the HarvESTer software (Biomax informatics, Martinsried). For specific assembly information, refer to the Supplementary Methods online. The resulting EST-derived unigenes were
annotated within the Sputnik database\textsuperscript{10} by transfer of information from heterologous human, mouse and other sequence resources.

For SNP detection, the SNiPper algorithm\textsuperscript{11} was applied to all multi-member unigenes. To minimize the detection of false-positive SNPs at least 8 residues were required to establish a consensus sequence, and for a given position within a unigene, 30\% of nucleotides must represent a consistent deviation relative to the consensus sequence to score a putative SNP.

A total of 452 SNP-containing EST clusters were found, totaling 1,394 SNPs. All of the clustered EST annotations and SNP data have been made publicly available at http://sputnik.btk.fi/sus. This represents the first \textit{in silicio} SNP database for pigs, similar to the Interactive Bovine In Silico SNP (IBISS) database recently released for cattle by CSIRO Livestock Industries (http://www.livestockgenomics.csiro.au/ibiss).

\textbf{Comparison of human and pig cSNP density}

Using the BLASTX annotations\textsuperscript{12} against a non-redundant protein database to determine homology and likely coding sequence position within the cluster, 231 SNPs were found to locate to the coding regions of 80 different genes or hypothetical proteins. Validation studies were performed for a sample of 9 SNPs from clusters having significant matches to mammalian genes or hypothetical proteins. Of these, 6 (67\%) were experimentally validated, indicating that the stringent conditions of the \textit{in silicio} methods produced reliable, high-specificity data. Sequence information for the validated SNPs is available online as Supplementary Results. The number of cSNPs per base of coding sequence in the consensus sequence defined the density of porcine cSNPs. Human and mouse cSNP density was determined for each gene by the number of validated and total cSNPs available from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) per base of coding sequence as annotated in GenBank (http://www.ncbi.nlm.nih.gov/). Ignoring hypothetical proteins and alternatively spliced
genes, and requiring that a human gene have at least one validated cSNP, the correlation between human validated and pig cSNP densities was found to be 0.77 (p < 0.00001) in a sample of 25 genes (Figure 1). The density of all human cSNPs was also determined for these 25 genes and had a correlation of 0.39 (p < 0.06) with the pig cSNP density (data not shown). The average pig cSNP density of these 25 genes was 1.9 cSNPs per 1,000 bp, with the average human validated cSNP density being 1.6 cSNPs per 1,000 bp and average total human cSNP density being 3.7 cSNPs per 1,000 bp. By including unvalidated human cSNPs, the cSNP density of many human genes became several-fold larger than the corresponding pig cSNP density. It is probable that the true number of porcine cSNPs in these 25 genes is underrepresented by the sample used here, in part because of our stringent requirement that a minor allele account for at least 30% of sequence reads. Computer-based methods for deriving SNPs in humans have additionally been shown to have low sensitivity with not more than 27% of true SNPs detected\textsuperscript{13}.

Using stringent parameters in the SNiPper algorithm likely resulted in a high number of false-negative SNPs, which limited the size of the cSNP data set. To obtain a larger data set, the parameters were relaxed to allow just 6 instead of 8 sequence reads to form a consensus with the minimum allele frequency remaining at 0.3, which allows for a minimum of two consistent deviations relative to a consensus sequence to declare a putative SNP. The number of false-positive SNPs that were detected likely increased by relaxing the parameters and resulted in a larger, though lower-specificity, data set. The new data set contained 21 times more cSNPs than the original data set. Ignoring hypothetical proteins, genes in families, and alternatively spliced genes, a total of 158 genes containing 669 porcine cSNPs were compared between human and pigs, and a correlation of 0.31 (p < 0.0001) was found between the porcine and human validated cSNP densities (data not shown). The average pig
cSNP density increased from 1.9 cSNPs per 1,000 bp in the original data set to 7.1 cSNPs per 1,000 bp, indicating that a large number of false-positive porcine cSNPs were likely included in this new data set. The average human validated cSNP density increased marginally from the previous sample (1.6 cSNPs) to 2.2 cSNPs per 1,000 bp, while the average total cSNP density remained the same at 3.7 cSNPs per 1,000 bp. Thus, in 52% of the genes, the pig cSNP density was more than double that of the human validated cSNP density, compared to only 7% of human genes having validated cSNP density double that of pigs. Although the primary data set was limited, it is a more reliable indicator of the frequency of SNPs in porcine coding regions. However, further investigation is necessary to obtain an accurate estimate of the true density of pig cSNPs.

**Comparison of human and mouse cSNP density**

If the high level of human and pig coding sequence identity leads to a strong correlation between their cSNP densities, then the correlation between human and mouse cSNP densities should be lower because their sequence identity is generally less. The validated cSNP frequencies from a primarily random sample of 50 homologous human and mouse genes were compared and showed a correlation of 0.48 ($p < 0.0005$) (Figure 2), while the correlation between human and mouse total cSNP frequencies was 0.36 ($p < 0.01$) for this sample of genes (data not shown). Unfortunately, only 7 of the 25 genes from the initial pig-human data set contained cSNPs in the mouse, according to dbSNP. These were included in the set of 50 genes used for the human-mouse comparison. Although a more extensive comparison of cSNP frequency across all three species was not possible, the correlation between the mouse and human cSNP densities for these 7 genes was zero, while the pig and human cSNP densities had a correlation of 0.73 ($p < 0.07$) (data not shown). The decreased correlation between human and mouse cSNP frequencies, as compared to that of human and
pig, supports the idea that forces regulating nucleotide diversity in coding regions will affect closely related species in a similar manner.

DISCUSSION

The apparent relationship between human and pig cSNP density is directly applicable to pig genomics research, as it will allow site-directed screening of porcine genes for cSNPs, resulting in their increased discovery. Future use of EST-based *in silicio* SNP detection methods in pigs is dependent upon the amount of available sequence data. Results from a large-scale porcine EST project are currently awaited to allow creation of large data sets for bioinformatic analyses. Validation of *in silicio*-derived SNPs, along with those derived from human comparative studies, will contribute to the supply of markers suitable for performing genome-wide association studies in pigs to determine an animal’s total genetic value\(^\text{14}\) and to promote genetic improvement in traits such as reproduction, disease resistance, production and longevity.

SUPPLEMENTARY METHODS

**EST clustering and annotation**

All porcine ESTs were obtained from the EMBL sequence database using the BioRS tool ([http://biors.gsf.de:8111](http://biors.gsf.de:8111)) and were loaded into the Sputnik sequence analysis database structure\(^\text{10}\). Sequence clustering and assembly were performed using the HarvESTer software (Biomax informatics, Martinsried, Germany). The Hashed Position Tree (HPT) clustering method that was employed used a similarity link threshold of 0.7, and a maximum distance of six steps was required to define a cluster from the similarity network, thus encouraging the separation of likely paralogs. Assembly of the EST clusters was performed using default CAP3 settings. A complete sequence annotation was performed as described previously\(^\text{10}\), but mammalian annotation references were used instead of plant genome sequences. Peptide
predictions were performed using the framefinder method from the ESTATE package and derived *Sus scrofa* hexanucleotide frequency tables. The resulting peptide sequences were annotated for Interpro domains, functional role using the MIPS funcat\textsuperscript{15} and for the presence of likely transmembrane domains. All annotations were performed within Sputnik and are retained within the *Sus scrofa* Sputnik project database (http://sputnik.btk.fi/sus).

**In silicio SNP detection**

SNPs were predicted using the SNiPper method as described\textsuperscript{11}. From the lists of predicted SNPs, a minimal SNP score of 4 was used to select a subset of self-validating polymorphisms – this subset was further limited by imposing a restriction on the minimum cluster size. Polymorphisms were examined within the context of the EST assembly to validate the number of ESTs present at the specified base and to score the relative frequency of each nucleotide. By imposing a requirement for the minimum cluster size at a minimal allowable allele frequency we could select for polymorphisms in a neighborhood-score free approach. Polymorphisms satisfying these requirements were labeled as candidate SNPs.

cSNPs were identified from the list of candidate SNPs by anchoring the SNP residue to BLASTX matches against the human proteome that had been filtered using an arbitrary expectation value of 10e-15. Individual nucleotides can be assigned to coding sequence or to non-coding sequence and cSNPs can be labeled as candidate synonymous or non-synonymous polymorphisms. The full list of porcine SNPs, their annotation and description are available as supplementary information.

**Correlation of cSNP density**

For each porcine cluster consensus sequence, the size of the coding region contained within the consensus was determined as described previously. The cSNP density was calculated as

\[ D = \frac{N}{L}, \]
where D was the cSNP density for a consensus sequence, N was the number of cSNPs identified in silico, and L was the length of the coding sequence contained within the consensus sequence. Human and mouse cSNP densities were calculated in a similar manner, except N equaled either the number of validated cSNPs or the total number of cSNPs as listed by dbSNP, and L was the length of the coding sequence as listed by GenBank.

REFERENCES


Figure 1 Correlation between porcine *in silico* cSNP density (# cSNPs per base of coding sequence) and human validated cSNP density (# validated cSNPs per base of coding sequence).
Figure 2 Correlation between human and mouse validated cSNP density (# validated cSNPs per base of coding sequence).
SUPPLEMENTARY RESULTS

Sequence information* for validated SNPs**

Cluster consensus 1

Most similar to: Sus scrofa MHC class I antigen (SLA-1), SLA-1*wxd allele

TCAGCTTCTCCCGACCCGGAGATGCGGCTATGGGGCCTCGNAGCCTCTTCC
CTGCTGCTGTCGGGNRCCTGCCCTGNACCAGGNNACCCRGCCGGGCCTGCCAC
TCCCTGNAAGCTATTCTACAACCGCGGCTGTCCCAGCAGCGCCAGCGGGNASYCCG
CTTCATCGCCGTCGGCTACGTGAGCAACGCACAGTTCGCTGAGTTGCACAGCGA
CGCCCGCASATCCGGGATGAGGACCCGGCCGGCCGCGCNCTGCAGTSAAGMGAGGAGG
CAGGNNNNNNNAGTATTTGGGATGRGGAGACGCGGAACGTCAWGGGRMAASC
CACAGACCTANCCAGGTAACCTGAAASACCTGCGGCTACTACAACCCAGNA
CGAGGCGCGGTCTCACAACCCTCCAGAGCATGATACGCTACTKGTTGAGACCA
CGGGCTCCTCCTCCGGGAGGTACAGTCAGGAGCGCTACAGCAGCGGCAACCGATTCA
TCGCCCTGAAACGAGGACCTGGCTCTGGACCGCGGACACGGCGGCTCAGA
TCACCAAGCGCAAGTNGGGAGGCGGCCGGATGAGGGAGCATAGGAGGAGCTAA
CCTGCAGGGCCTGTGTGAGTCTCGCTCCGCAAATACCTGGAGATNGGGGAAGG
ACACGCTGCAGCGCAGCGCAGCAGGCTCACAAGACACATGTGACCACCCGCAAAAA

Cluster consensus 2

Most similar to: Homo sapiens mitogen-activated protein-binding protein-interacting protein (MAPBIP)

GCACAGAAATTATGCATAATAGGAATTTTGTTCTTTATTTTTTAAATTACTGAATTGGAGATTTGCAAG
CCTGACAAAGTGATAAGGATTTTGTCTTTTTTTTTTAATTACTGAATCTTGAGAATTGGTCAAG
CAGATATGACTGCTCGGGCGAGGGGTTAGTTAAATGTCTCCTCCACGTAGCTGCCATAT
CCTTTTACTCCCGCTCCTTTCCCCCAACCCGCTGAGTTCTCACTGGAAAGGCGAA
CTGTAAGTCCAGGGGTGTCCGCTGCAGAACTCAGTCAATCCCAAGCCGCCC
CCCGCGTGGAGGCGGGTGGTGAGTCAAGAAAACCAGCACGTCAGCCAGGAAACTAC
AACTCCAGGAGTTCTGAGCGCCAGGCGAGCTGGAACAGCGGAGCGGCAGGGCGG

* EST cluster consensus sequence with gene annotation according to highest BLAST matching score
** Validated SNPs are indicated by underline and boldface font. W = T/A, Y = T/C, R = A/G, M = C/A
Cluster consensus 3

Most similar to: Homo sapiens S100 calcium binding protein A14 (S100A14)
Cluster consensus 4

Most similar to: *Homo sapiens* oxidase (cytochrome c) assembly 1-like (OXA1L)

```
TTTTTTTTTTTTTTTAATCTAGTTAGGTGCAAAACTGAGCATCAAGTTCT
GACCTTCAGTGGAGCAAAATTTATAGTCCAGTGCACCCCTGGCAGTCCTGCCC
TAAGCTTTCCATTCCTTCATAGTGTTATTGCTCTCTCCACACATGTTC
TCCCTGGGCAAGTGCTTTGGCCTCAGTATTTCCAGGCTGCTGACGAGT
TCCCTGGGCAAGTGCTTTGGCCTCAGTATTTCCAGGCTGCTGACGAGT
CCATAGCTTTACACCCCATCGCTACAGCCCTGGAAAGAATGAGCGGCTCTCCA
```

Cluster consensus 5

Most similar to: *Sus scrofa* ribophorin I (RPN1)

```
ACAAAGACGGGAGATCCCATGTTAGGAGAGCTAGGTAAATAATATATAATCAT
ATGCCCATTTAGTCAGCAGAAAATAAAGCTAGCAGAAGACTGGACG
CACAGGCTCAAGGAGCTGACTGAAGTGCTAGCGTGAGGCCAGAGCTTNGTG
GCTGGGACAGCTCGAAAGAACANCATACNATTGAGAATGAGAAGCTTACCTCG
GAAGCAGCGAGGAGCTGCTACAGCAGTACCAGACATCCTTGGATGCMCTGAGC
CATTGACTTACCTTTAAGAAGGGCAGAATGGGGGTGTGGAACCCAGGCAGCAGA
ATGCTGTCTCTGTCTGAGAGACCTCTAAAGGAGAGCAGAACGCCAGGCC
TGCTCAGGGCACAAGCAGAAGAGCTTATTGTGTGCTCCGAGATTTCTTTCCNTT
TTTTTCCAAAAACACTACCCACTTTAAATCCCTTAATACAAAAATATTTTGTG
TTTTTGGGAAAAAGAAATTTTTAGTCTGCTTTGGTGTGTGTAAGCCTAGGATATT
TTTTGCTGCTAAAGTGCTCCCTGCTTATTTGCTGTGTTTGAGTATT
```

TTATAAGCATTGGTGGTTTTTTAAGTGTGTGTGCGCAAATGAAATAGTGGG
ACTGTGAAACNNNNNNNNNNNNNNNNNNNNN
Cluster consensus 6

Most similar to: *Homo sapiens* protein kinase C, delta binding protein (PRKCDBP)

TTTTTAGAGGTGACACTATTTTATTGATGGTGAAAGGAAGCAAGGTTGCGTGTTG
GGGTGCCTCCTTCAGAGGACAAGGGGGGTGCCGGCTGAAGAAGCTGGGAAGGGG
AGGGCCCGCCCTGGGCTGTGGGTGCAGGATTCACTCCTTATTTGGGTGAGAATG
CTGATTTCTGAGAGGAGATTATTATTGTGGGACTGAGCAAGCACCACCGGGGAGGA
GAAGCACCAGCCCGTCCAGCCGCACTCCTTATTGGAGACACGCGCGCCGCTCC
GCTGCTCCTCGGTCTCCCGGGAGCTTTCTCGGGTCTGGGAGAAGTTCTGGCTCCA
GGCTTGAACTCCAGCGCATGCTCCGCTTGCGGCTGCCCTGCCTCGCTCCCGCCTG
ACCATCGCGAGGCGGCCTTCACAGGGCCTGCCTGGCTGGCCTGATGCCCTTCCG
GCCGAAAGGCGCCTTTGCCAGGTCTCGAGCCTCTGTGAAGCAGCGGTGGCGGCA
CGCCTCGCCMTGCACTCCACGGGCTCCCTCGCCAGCTCCTCCCCACCTTCGCCC
TCCAGMTGCTCCGGCNCCGGGTTCGGNCTCGGTCTGCTNCCCGGGCMTAAGGGGT
CCGCTTTCTGGAAGGCGCTGGCGGATTTCAGCCTCCTTGAAAGCAGAAA
CGTGAGCTTTCCCGCCGCGCAACAGCCCGTCCGCTCCAGCCGCTTGAGCGGCA
CGCMTGGGGCGCGCGCGCCACCGGCGCTCGGCGCATCCGGGCCTGGCGAGGCC
ACCGGGCTCGGCCCTTGCCAGACCTGACGCGGCTTGTGGCCTCCAGCCGCTGGA
CGCCCGCTGCTCGCCCGAGCGAAGCCCGCCGAGGCGCCCTCGCCAGCGCTCGA
TGGCCAGCTCTCCACAGCAGGTCACCCAGGCTGGCCGCCCTTCCGATCTCCCGCC
GANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN


CHAPTER 7. GENERAL CONCLUSIONS AND DISCUSSION

GENERAL CONCLUSIONS

The research objectives of this thesis were to compare statistical methods for fine mapping of a previously detected QTL, to detect and map candidate gene polymorphisms, and to use bioinformatic methods for the \textit{in silico} identification of SNPs. The ultimate goal of all of these methods is more rapid and efficient QTL identification and characterization in livestock. Chapters 2 and 3 address fine mapping methodology, chapters 4 and 5 describe candidate gene studies in pigs and chapter 6 discusses bioinformatic tools for the \textit{in silico} identification of SNPs in pigs.

The main objective of the fine mapping work presented in Chapter 2 "Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci" was to compare mapping accuracy of the haplotype-based, identity by descent (IBD) mapping method of Meuwissen and Goddard (2000) to a single marker-based regression method when fine mapping a previously detected QTL. The main objective of the fine mapping work in Chapter 3 "Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci" was to determine the optimal number of markers to consider as the haplotype when using the IBD method of Meuwissen and Goddard (2000) for fine mapping. The IBD and regression methods were tested in populations that both adhered to and deviated from the assumptions made about population history in the haplotype-based IBD method. The findings of this work were:

- Given an equal number of markers genotyped within a region, the IBD method had significantly greater mapping accuracy than the single marker regression method for every population and data structure evaluated.
• Assuming that twice the number of markers can be genotyped for the regression method, because haplotype-based methods such as the IBD method will likely require additional genotypes to determine haplotypes, single marker-based regression had significantly greater or comparable mapping accuracy to the IBD method for every population and data structure evaluated.

• Presence of linkage disequilibrium generated by population admixture, rather than a historical mutation event, had a significant impact on mapping accuracies of both methods.

• Deviations off marker allele frequencies from 0.5 within the range of 0.2 to 0.8 had little or no effect on the mapping accuracy of either the IBD or regression method.

• When all available markers within the region were fitted as the haplotype for the IBD method, the method was not sensitive to position of the QTL, which decreased mapping accuracy.

• Using a haplotype of four or six markers as a sliding "window" across the chromosomal region resulted in the greatest mapping accuracy in all populations and data structures evaluated, although differences were not always significant.

• For fine mapping with the IBD method, marker information must be used in such a way that the IBD method is sensitive to the putative position of the QTL while maintaining power to detect the QTL. A haplotype of four markers best meets these requirements.

The main objective of Chapter 4 “Physical and linkage mapping of the porcine connexin 37 (CX37) gene” was to identify polymorphisms within porcine CX37 and to map
the gene. Similarly, the main objective of Chapter 5 “BMP15 maps to the X chromosome in swine” was to identify polymorphisms within porcine BMP15 and to map the gene. Findings from these studies were:

- Two synonymous single nucleotide polymorphisms (SNPs) were discovered in the coding sequence of the CX37 gene. One was a single T to C transition located at position 366 from the human ATG site for CX37 (GenBank Accession no. AF132674). The other was a G to A transition located at position 345 from the human ATG site in the CX37 sequence listed previously.

- The T to C SNP was used to link the porcine CX37 gene to SSC 6, centromeric of S0059 by 6.3 centimorgans.

- Physical mapping placed the porcine CX37 gene on SSC 6 q24-31 with a probability of 0.87, which agrees with the linkage mapping results.

- No polymorphisms were identified in approximately 1125 bp of the porcine BMP15 coding region and 150 bp of the 5’ untranslated region among several animals from several swine breeds, yet no polymorphisms were identified. This indicates that the coding region of BMP15 is highly conserved in the pig.

- Physical mapping placed the porcine BMP15 gene on SSC X, p11-13 with a probability of 0.995. This corresponds directly to the human position of BMP15, which is HSA X p11.2.

The main objective of Chapter 6, “Prospecting for pig SNPs in the human genome: have we struck gold?” was to determine if a relationship could be shown between the nucleotide diversity in the coding regions of humans and pigs on a gene-specific basis. A
secondary objective was to determine the efficiency of *in silico* methods for identifying SNPs from redundant EST sequences in pigs. Findings from this study were:

- A total of 452 SNP-containing EST clusters were identified, containing a total of 1,394 putative SNPs.
- All of the EST annotations and SNP data have been made available by Dr. Stephen Rudd at [http://sputnik.btk.fi/sus](http://sputnik.btk.fi/sus).
- A total of 231 putative SNPs of the 1,394 identified SNPs were located to the coding regions of 80 different porcine genes or hypothetical proteins.
- Using computer-identified coding SNPs (cSNPs) within 25 porcine genes and validated cSNPs from the 25 homologous human genes available from dbSNP ([http://www.ncbi.nlm.nih.gov/SNP/](http://www.ncbi.nlm.nih.gov/SNP/)), the correlation between the cSNP frequency in humans and pigs was high (0.77; *p* < 0.00001) given that stringent parameters were used to minimize false-positive SNP detection in pigs.
- Using the frequency of human and mouse validated cSNPs from dbSNP for a primarily random sample of 50 genes, the correlation between the cSNP frequency in humans and mice was moderate (0.48; *p* < 0.0005) and lower than that of humans and pigs.
- From a sample of 9 putative pig SNPs identified *in silico*, 6 (67%) have been experimentally validated by PCR-RFLP tests.
- Given a large supply of publicly available EST information, *in silico* methods will be efficient for identifying putative SNPs on a large-scale in pigs.
• The rate of cSNP discovery in pigs will be increased by site-directed screening of the coding regions of homologous human genes known to be cSNP-dense in humans.

GENERAL DISCUSSION

Despite the completion of numerous whole-genome scans that attempted to identify chromosomal regions harboring genes affecting complex traits in livestock, very few QTL, or more specifically, quantitative trait nucleotides (QTN, Mackay 2001), have actually been characterized (e.g. Van Laere et al. 2003; Kim et al. 2004). A set of working guidelines has been proposed by Glazier et al. (2002) for the confirmation of gene discovery in the area of complex traits, regardless of species. According to Glazier et al. (2002), the first step is to establish statistically significant evidence that genetic markers in a chromosomal region are linked to gene(s) that affect a trait, usually by conducting whole-genome linkage studies. The second step is to reduce the size of a significant chromosomal region by examining a population amenable to high-resolution mapping, such as congenic strains or near-isogenic lines, or by performing linkage disequilibrium (LD) mapping in experimental crosses or case-control studies (Glazier et al. 2002). Third, sequence analysis of the refined region should be performed to identify candidate nucleotide polymorphisms, which should then be considered individually and in all possible combinations and prioritized for further testing (Glazier et al. 2002). In the fourth step, functional analyses of the candidate variants should be performed, either by knock-out and knock-in technology or by gene-targeted deficiency and transgenic complementation, to show that replacing one genotype, or genotypes, with another, alters the phenotype accordingly (Glazier et al. 2002). Admittedly, there are complicating factors at the stage of functional confirmation, especially in livestock, such as
possible dependence of the nucleotide variant(s) on genetic background, that may make these functional tests uninformative (Glazier et al. 2002).

Fortunately for those working in livestock genomics, it is often sufficient for the purposes of marker assisted selection to identify genetic markers that show strong association with a trait(s) across a large sample of individuals, and it may not be necessary to expend effort to functionally characterize those variants. A genetic marker that is in population-wide LD with the QTN, if it is not actually the true QTN itself, will be useful in marker-assisted selection programs. While it is important to determine the QTN to further basic biological knowledge, the ultimate goal of characterizing QTL in livestock is often centered around a more practical purpose - the pursuit of more efficient and accurate selection methods to improve health, welfare and production. Thus, when the objectives for identifying QTL are practical in nature, it should only be necessary to complete the first three steps recommended by Glazier et al. (2002), which are whole-genome linkage analyses, fine mapping and sequencing to identify candidate genetic variants.

As mentioned previously, many whole-genome scans have already been performed in livestock for a variety of complex traits. For example, at least 40 QTL have been identified in pigs that are related to reproductive traits alone (Cassady et al. 2000). Thus, it seems that future research should progress to steps 2 and 3 as described by Glazier et al. (2000) in order to identify the genes underlying complex traits.

Step 2 involves fine mapping of a previously identified QTL region, and for most livestock species LD mapping will likely be the preferred approach, as creation of congenic or near-isogenic lines is not practicable. Results from the fine mapping methodological research presented in chapters 2 and 3 suggest that single marker-based LD mapping has equal or greater accuracy than haplotype-based LD mapping, given an equitable amount of
experimental resources. Single marker-based mapping also was conducted using a simple linear model with marker-QTL associations determined by the marker locus with the highest model sums of squares. The haplotype-based approach used a more complicated mixed linear model with marker-QTL association determined by maximum likelihood methods and required estimation of probabilities that QTL were identical by descent (IBD). Thus, when considering the experimental requirements of each method, performing fine mapping using a single marker-based approach may be more efficient.

Step 3 in the process of identifying and confirming QTL according to Glazier et al. (2002) involves sequencing DNA within the narrowed QTL region in order to identify candidate genetic variants. Although specific chromosomal regions were not pre-determined for study by fine mapping in the candidate gene work presented in chapters 4 and 5, the focus of these projects was to identify sequence polymorphisms that may affect ovulation rate in pigs. Generally candidate gene studies rely heavily upon comparative sequence and map information, as well as assumptions about comparative biological function, taken primarily from the abundance of human and mouse information. While this information often can be used successfully, it is not always completely reliable. The initial impetus for examining the BMP15 gene in pigs stemmed from the report of mutations found in the ovine form of the gene affecting ovulation rate in a dosage-sensitive manner (Galloway et al. 2000). It was hypothesized that mutation(s) in BMP15 may exist in pigs and have smaller or even similar effects to those observed in sheep. However, after sequencing nearly the entire coding region of BMP15 in more than 60 individuals from 6 pure breeds and 1 synthetic line of pigs, no polymorphisms were identified. Later, three polymorphisms located in exon 2 of the porcine BMP15 gene were discovered; however, in a sample of more than 380 animals, the frequencies of their rare alleles were only 0.14, 0.02 and 0.01, respectively, for the three
SNPs (Wang et al. 2003). As of late, no studies have been presented that test whether these mutations are associated with, or directly influence, ovulation rate in pigs. In the case of the CX37 gene, there was strong evidence from knock-out studies in mice that CX37 was critical for follicular maturation, and hence, ovulation. Examination of this gene did yield two synonymous coding SNPs, one of which was used for mapping purposes. There have been examples of the candidate gene approach successfully identifying genes that are strongly associated with complex traits, without the assistance of fine mapping, e.g. Kim et al. (2000) and Ciobanu et al. (2001). However, when screening large QTL regions that have not been refined by additional mapping, it is critical to consider as many genes as are feasible, within the bounds of the available resources, in order to identify sufficient candidate polymorphisms, as sometimes the seemingly obvious candidate genes yield little useful results.

An alternative approach to identifying candidate polymorphisms in QTL regions involves the use of bioinformatic tools, such as the ones described in chapter 6. If putative SNPs can be discovered in silico and subsequently validated at a reasonable rate, say at least 60%, the rate at which SNPs are discovered and mapped could dramatically increase. Results from chapter 6 indicate that in silico SNP detection from redundant EST sequences is a reliable method for obtaining new SNPs. Because only species-specific EST sequences are used in the analysis to identify sequence variants, oligonucleotide primers can easily be designed from the consensus cDNA sequence surrounding the putative SNP for performing validation studies, such as repeated sequencing or PCR-RFLP tests. This differs from the way primers are often designed in candidate gene studies, where usually little or no species-specific sequence is available, and so primer sequences are derived from regions of high sequence identity across multiple species. Obviously, there is great potential for in silico
methods to assist in not only the identification of candidate SNPs that may play a functional role, but also in the detection of SNPs for use as anonymous markers. Validated SNPs can be used to create a dense-marker map for fine mapping in a previously identified QTL region and potentially contribute to a genome-wide SNP map, similar to the one created for humans (The International SNP Map Working Group 2001).

However, there is a limit to the number of putative SNPs that can be identified through this type of in silico process, and that is dependent upon the initial amount and quality of EST information available for a species. Of the livestock species, chickens, specifically Gallus gallus, currently have the largest number of ESTs deposited in dbEST at 451,655, followed by cattle (Bos Taurus) at 331,140 and then pigs (Sus scrofa) at 259,350. While those numbers seem large, none of them is even one-tenth as large as the number of human ESTs in dbEST. So, it may be possible that an in silico analysis of porcine ESTs may not identify putative SNPs in a chromosomal region of interest because there was no EST information available in the database to begin with for the genes located in that chromosomal region. Also consider that nearly all of the SNPs on the human SNP map were identified by shot-gun sequencing of genomic fragments or from regions of overlap between large-insert clones such as bacterial artificial chromosomes (BACs) (The International SNP Map Working Group 2001). As there are no immediate plans to release information from the Sino-Danish sequencing project of the porcine genome, unlike those of chicken and cattle, and considering the relatively limited amount of porcine EST information available for in silico analyses, it may be highly beneficial in the short term to capitalize on the large amount of human SNP information to aid in the discovery of porcine SNPs. As shown in chapter 6, a strong relationship may in fact exist between the frequency of human and porcine coding SNPs (cSNPs). So, site-directed screening of porcine genes that are homologous to cSNP-
dense human genes may lead to increased discovery of porcine SNPs. As these will be cSNPs that are identified, they could not only serve as anonymous markers for refining a QTL region, but also directly as candidate variants that could affect the trait of interest. By combining this human-pig comparative approach with that of bioinformatic tools, the rate of SNP detection in pigs could be greatly increased.

The results from the entirety of the work described in this thesis are primarily involved with the completion of steps 2 and 3, as proposed by Glazier et al. (2002), in the search for genes that underlie complex traits. While there is still a hefty burden of proof for cloning a QTL, in livestock it is often sufficient to show significant association between DNA sequence and phenotypic variation and ignore functional studies. However, identifying sequence variants in pigs is a tedious process, often proceeding on a gene-by-gene basis. Even when variants are identified, they are not always found to be associated with an economic trait, even though comparative biology suggests that they should be. Hopefully, the tools and approaches described in this work will be used to speed the rate at which genes underlying complex traits in all livestock species are characterized.

RECOMMENDATIONS FOR FUTURE RESEARCH

Fine mapping methodology

In the process of examining the IBD-based LD mapping method of Meuwissen and Goddard (2000), an apparently novel phenomenon was observed, in that the optimal mapping accuracy of this haplotype-based method occurred when less than all available marker information was utilized in the model. Discussions concerning this issue were held with faculty from the Department of Statistics at Iowa State University, but no conclusions were reached and as of yet no explanation has been found. It may be that this result is unique to
the method used for haplotype-based fine mapping in the studies presented here, i.e. an IBD-based method, and another method that utilizes haplotype information differently will not behave in such a manner. This should be explored more fully to determine if the optimum number of markers to consider in a haplotype is always less than all available markers, regardless of the chosen LD mapping method. Also to consider is whether the optimal number of markers is the same for alternative LD mapping methods and/or experimental parameters such as effective population size. It may also be true that this "less is more" result in mapping accuracy is unique to fine mapping studies, where it is already known that a QTL is located in the chromosomal region being fine mapped, and so power to detect the QTL becomes less important than sensitivity in estimating the QTL position. Future research should examine all of these possibilities in order to create LD-based method(s) of fine mapping that can narrow QTL regions as much as possible for further study.

SNP identification in livestock

With a patent already granted to Malcom J. Simons titled "Intron Sequence Analysis Method for Detection of Adjacent and Remote Locus Alleles as Haplotypes" (U.S. Patent No. 5,612,179; http://www.uspto.gov) which claims rights to all intronic SNPs discovered in all eukaryotic organisms, it has now become even more important to identify SNPs in coding and regulatory regions if animals' genotypic information is to be used in commercial livestock production. Due to the small sample size used to show a relationship between human and porcine cSNP frequency in chapter 6, future work should attempt to demonstrate this relationship on a larger scale. If this relationship is found to be strong in a larger study, then human SNP information can be directly applied to searches for cSNPs in pigs. Also, if EST-based in silico methods are to be applied successfully in pigs and result in a large number of SNPs that span the genome, then additional EST projects will have to be
undertaken so as to increase the amount of data available for bioinformatic analyses. This will make it more likely that putative SNPs will be identified in any chromosomal region previously shown to be associated with complex trait(s).

Methods could also be designed to analyze overlapping sequences from contigs created from large-insert clone sequences, such as BAC contigs, for the livestock species. It is likely that these methods will differ only slightly from the EST-based analyses already available. However, as whole genomic regions will be available for analysis in these cases, rather than just cDNA sequences, a larger, more widespread number of putative SNPs could be identified, compared to that found from EST-based analyses alone. This likely will make the completion of genome-wide SNP maps more feasible for livestock species, as it is unlikely that SNPs derived from ESTs alone will occur at a high enough frequency across the genome to create a sufficiently dense map. If the ultimate goal in livestock is identifying all of the genes underlying complex traits, for both the purposes of improving production and extending the use of livestock as model organisms for human disease, then obtaining SNPs for fine mapping as well as for candidate study is one of the most critical needs.

REFERENCES


