2003

Interval mapping of QTL with selective DNA pooling data

Jing Wang
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Genetics Commons, and the Statistics and Probability Commons

Recommended Citation
Wang, Jing, "Interval mapping of QTL with selective DNA pooling data " (2003). Retrospective Theses and Dissertations. 749.
https://lib.dr.iastate.edu/rtd/749

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.
Interval mapping of QTL with selective DNA pooling data

by

Jing Wang

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-majors: Animal Breeding and Genetics (Quantitative Genetics); Statistics

Program of Study Committee:
Jack Dekkers, Co-major Professor
Kenneth Koehler, Co-major Professor
Rohan Fernando
Max Rothschild
Susan Lamont
Dan Nettleton
Jean-Luc Jannink

Iowa State University
Ames, Iowa
2003
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

Jing Wang

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Co-major Program

Signature was redacted for privacy.

For the Co-major Program
DEDICATION

This thesis is dedicated to my mom, husband and daughter, who love and support me all the time. Without their love and support, none of this would have been possible.
# TABLE OF CONTENTS

**ABSTRACT** vi
**CHAPTER 1. INTRODUCTION** 1
  - Background and rationale 1
  - Research objectives 4
  - Thesis organization 5
  - References 7
**CHAPTER 2. LITERATURE REVIEW** 9
  - Literature review 9
  - References 28
**CHAPTER 3. INTERVAL MAPPING OF QTL WITH SELECTIVE DNA POOLING DATA. I. METHODOLOGIES** 39
  - Abstract 39
  - Introduction 40
  - Materials and methods 41
  - Results 60
  - Discussion 61
  - Conclusions 68
  - Literature cited 69
  - Appendix 72
  - Tables 75
  - Figures 79
**CHAPTER 4. INTERVAL MAPPING OF QTL WITH SELECTIVE DNA POOLING DATA. II. SIMULATION STUDIES** 85
  - Abstract 85
  - Introduction 86
  - Materials and methods 87
  - Results 97
  - Discussion 104
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conclusions</td>
<td>113</td>
</tr>
<tr>
<td>Literature cited</td>
<td>114</td>
</tr>
<tr>
<td>Tables</td>
<td>116</td>
</tr>
<tr>
<td>Figures</td>
<td>122</td>
</tr>
<tr>
<td>CHAPTER 5. DETECTION OF QTL AFFECTING ECONOMIC TRAITS IN LAYERS AND COMPARISON OF RESULTS FROM SELECTIVE DNA POOLING AND INDIVIDUAL GENOTYPING</td>
<td>126</td>
</tr>
<tr>
<td>Abstract</td>
<td>126</td>
</tr>
<tr>
<td>Introduction</td>
<td>127</td>
</tr>
<tr>
<td>Materials and methods</td>
<td>129</td>
</tr>
<tr>
<td>Results</td>
<td>141</td>
</tr>
<tr>
<td>Discussion</td>
<td>149</td>
</tr>
<tr>
<td>Conclusions</td>
<td>154</td>
</tr>
<tr>
<td>References</td>
<td>156</td>
</tr>
<tr>
<td>Tables</td>
<td>162</td>
</tr>
<tr>
<td>Figures</td>
<td>165</td>
</tr>
<tr>
<td>CHAPTER 6. GENERAL DISCUSSION</td>
<td>172</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>175</td>
</tr>
</tbody>
</table>
ABSTRACT

Selective DNA pooling is an efficient method to identify chromosomal regions that harbor quantitative trait loci (QTL) by comparing marker allele frequencies in pooled DNA from phenotypically extreme individuals. The currently used single marker analysis can detect linkage of markers to a QTL, however, it does not provide separate estimates of QTL position and effect, nor does it utilize the joint information from multiple linked markers. In this thesis, two interval mapping methods for analysis of selective DNA pooling data were developed. One was based on least squares regression (LS-pool) and the other on approximate maximum likelihood (ML-pool). Both methods simultaneously utilize information from multiple markers and multiple families and both are easily applied to different family structures (half-sib, F$_2$ cross and backcross). Simulation was used to compare these two DNA pooling interval mapping methods with single marker analysis and with selective genotyping analysis of individual genotypes. Results indicated that both LS-pool and ML-pool provided greater power to detect the QTL than the single marker analysis and separate estimates of QTL location and effect. With large family size, both LS-pool and ML-pool provided similar power and estimates of QTL location and effect as selective genotyping. The LS-pool method, however, resulted in severely biased estimates of QTL location with small family size and distal QTL but the bias was reduced with ML-pool. Both interval mapping methods were also applied to two real data sets, a dairy cattle data set from a half-sib design and a layer chicken data set from an F$_2$ cross. In the dairy cattle data application, both LS-pool and ML-pool solved problems of single marker analysis when
missing marker genotypes were present by utilizing joint information from multiple markers. In the chicken data application, both LS-pool and ML-pool provided similar power to detect the QTL and similar estimates of QTL location and effect as analysis of individual genotypes. In conclusion, both LS-pool and ML-pool methods provide powerful tests for QTL detection and accurate estimates of QTL parameters while substantially saving genotyping costs through the use of DNA pooling. In addition, both methods are readily applied to practical situations.
CHAPTER 1. INTRODUCTION

BACKGROUND AND RATIONALE

The ultimate goal of agriculturists is to provide sufficient and healthy food for our society. Improving genetic factors has great potential for improving the quantity and quality of products of livestock and plants but is not an easy task. Most important economic traits are controlled by multiple genes with small or moderate effects and are also modified by environmental factors. Therefore, the distribution of the phenotype of these traits is quantitative (henceforth, named quantitative traits) instead of clear and discrete. Thus, it is difficult to distinguish a few genetic factors that have major effects on the trait and then effectively utilize them in breeding schemes.

For a long time, breeders have directly selected on the collective additive effect of multiple genes contributing to a trait (breeding value), without efforts to identify the specific genes that underlie the trait. Extensive quantitative genetics theories (Henderson, 1984) have been developed to improve such selection and great success has been achieved. However, all information in this conventional selection approach comes from phenotypic measurements and pedigree. Some traits are difficult or expensive to measure (e.g., feed efficiency and carcass traits), some traits are sex-limited (e.g., egg and milk production traits), and some traits can only be observed at a later stage of life (e.g., sexual maturity, survival ability). Since those traits have a limited number of observations or one must wait for a long time before a
selection decision can be made, they are difficult to be efficiently selected by conventional means. However, if one can not only obtain performance information but also can directly observe variation in the genetic material, DNA, the effectiveness of selection can be improved. The DNA can be extracted and detected in both sexes and at the first day of life (or even before birth), and it directly reflects genetic variation among individuals.

In recent years, molecular knowledge and techniques have been rapidly developed, which has enabled us to better understand the genetic mechanisms behind trait expression and provided us powerful tools for further genetic investigation. All of these have helped to cast a light on the black box of breeding values.

Therefore, an important research area in animal and plant breeding is to detect genes (or loci) contributing a significant effect on quantitative traits (henceforth, quantitative trait loci - QTL), with the aid of molecular information, and to estimate their parameters (QTL mapping), such as their location and effects on the trait phenotype. In general, there are two main approaches to search for QTL: the candidate gene approach and the genome scan approach (Rothschild and Soller, 1997).

In the candidate gene approach (Bryne and McMullen, 1996), the biological basis of the trait is studied and several genes with known function that might act as important factors in the physiological pathway for the trait of interest are selected as candidate genes. Polymorphism (variation in genotype) in or near these genes are then tested in the population and those polymorphisms that show a strong association with performance are considered as being
associated with the causative genes. Then, the polymorphic alleles associated with the phenotype have to be examined in other populations to verify their association. The candidate gene approach is efficient and accurate since it can more directly lead to causative genes. However, it requires a good understanding of the biological pathways and functions of discovered genes, which is usually unknown or complex.

In contrast to the candidate gene approach, the genome scan approach (Paterson et al., 1988) does not require previous knowledge about the biological pathways and gene functions. Instead, it requires information about the phenotypic value and a large number of genetic markers with high levels of polymorphism, sufficient to cover the genome. Since the genetic contribution of any specific gene in a quantitative trait may be small, usually several hundreds of markers from several hundreds of individuals are required to achieve a desired power to detect the QTL (Darvasi et al., 1993). Thus, the cost and labor of genotyping is very intensive and often limits QTL mapping research. Thus, a simple and accurate screening for relevant loci among the whole genome will be of great value.

One possible way to reduce the genotyping cost is selective genotyping, in which only individuals with high and low phenotypes are genotyped, thereby reducing genotyping costs (Lander and Botstein, 1989). Since most information about the QTL is provided by individuals that deviate the most from the phenotypic mean, using selective genotyping can provide similar power to detect the QTL as when the whole population is genotyped (Lebowitz et al., 1987; Lander and Botstein, 1989).
Selective DNA pooling goes one step further than selective genotyping, that is, DNA of the selected individuals in the same tail is pooled together, resulting in a high pool and a low pool, and marker allele frequencies are determined in each pool. If a marker is linked with the QTL, its allele frequencies in the pools are expected to differ from 0.5 (Darvasi and Soller, 1994). With this method, no individual genotyping is required, resulting in substantially reduced genotyping cost. The current method to detect QTL with selective DNA pooling data is based on single marker analysis, which was described in Darvasi and Soller (1994). However, single marker analysis can only identify which markers are linked with a QTL but does not provide separate estimates of QTL location and effect, nor does it utilize joint information from multiple markers. For individual genotyping data, interval mapping methods have been developed to simultaneously use data from all markers and to provide separate estimates of QTL position and effect (reviewed by Weller, 2001). Interval mapping methods also have greater power to detect the QTL than single marker analysis. However, interval mapping methods have not been developed for selective DNA pooling data. Dekkers (2000) introduced principles of mapping QTL with selective DNA pooling data, which can be used as a basis for the development of interval mapping methods.

RESEARCH OBJECTIVES

The objectives of the current research were to: (1) develop methods for interval mapping of QTL with selective DNA pooling data, that can efficiently utilize information from selective DNA pooling data and can be readily applied to practical situations, (2) evaluate the DNA pooling interval mapping methods that are developed under the objective 1 by simulation and
compare their statistical power to detect the QTL and the accuracy of estimates of QTL parameters with results from the currently used single marker analysis and results from approaches using individual genotyping, and investigate the effect of different parameters on QTL mapping results from the DNA pooling interval mapping methods, (3) apply the DNA pooling interval mapping methods to real data sets and evaluate them in practical situations.

**THESIS ORGANIZATION**

The present chapter (1) provides readers a general background, an introduction of research issues and their current status, introduces research objectives for the thesis, and provides an outline of the thesis. Chapter 2 provides an extensive review of literature that is relevant to the research objectives. Chapters 3, 4 and 5 are manuscripts in preparation for three individual scientific papers. Each of them includes an abstract, introduction, materials and methods, and results and discussion. References cited within each paper are, therefore, listed at the end of each chapter. Chapters 3 and 4 are intended to be submitted to Genetics and Chapter 5 is intended to be submitted to Genetics, Selection and Evolution. The format of those chapters, therefore, follows the guidelines of the corresponding journals. The last chapter is a general discussion and suggestions for future research.

In Chapter 3, two interval mapping methods with selective DNA pooling data will be introduced. One is based on least squares regression and the other on approximate maximum likelihood. In addition, the two methods are evaluated by simulation with one set of parameters and both of them are applied to a real dairy cattle data set. In Chapter 4, the two
interval mapping methods will be further evaluated by simulation at several combinations of parameters. The QTL mapping results will be summarized by power to detect the QTL, and means and standard errors of estimates of QTL location and QTL effect. Both interval mapping methods will be compared with single marker analysis and with selective genotyping based on individual genotypes. This analysis will demonstrate that the least squares interval mapping method can result in substantial biases in estimates of QTL location for some parameter combinations and methods that attempt to correct the bias will be introduced and evaluated. In Chapter 5, QTL mapping will be conducted using both pooling data and individual genotyping data from an $F_2$ cross in layer chickens. Single marker analysis is first used to scan the whole genome for QTL based on pooling data from multiple grandparent families. Based on results from this analysis, the regular least squares interval mapping method with individual genotyping data will be applied to one of the grandparent families and two chromosomes that are found to have several QTL in the pooling data analysis step. The analysis of selective DNA pooling data using the two DNA pooling interval mapping methods and single marker analysis will be applied to the same grandparent family and QTL mapping results will be compared with those from the analysis of individual genotyping.
REFERENCES


CHAPTER 2. LITERATURE REVIEW

Molecular markers used in the genome scan approach for QTL mapping

The genome scan approach (Paterson et al., 1988) requires information about the phenotypic value and a large number of genetic markers with high polymorphism, sufficiently to cover the genome. The rapid developed molecular genetics has provided research for QTL mapping a large variety of DNA markers that can be used for a genome scan.

Initially, restriction fragment length polymorphisms (RFLPs) were used for QTL mapping (Weller et al., 1988). RFLPs are based on a single nucleotide change at the recognition sites of restriction enzymes, resulting in DNA fragments of various lengths but limited polymorphisms. Another widely used DNA marker in early QTL mapping studies was variable number of tandem repeats (VNTRs or DNA fingerprints) (Plotsky et al., 1993). The VNTRs are much more polymorphic than RFLPs but no genetic location information is available for this type of markers. In recent studies, microsatellites (Litt and Luty, 1989; Tautz, 1989; Weber and May, 1989) and single nucleotide polymorphisms (SNPs) (reviewed by Syvanen, 2001) have been used extensively.

Polymorphism for microsatellite markers is presented as different numbers of repeats of two to four nucleotides, such as (AC)$_{23}$, (ATAC)$_{2}$, etc., where the letters in parentheses are the repeated nucleotides and the subscript refers to the number of repeats. Different genotypes
have different numbers of repeats. An abundance of microsatellite markers exists in the non-functional regions of the genome, such as regions between functional genes and introns within the gene. As a result, microsatellite markers themselves usually have no effect on the trait.

When comparing the DNA sequence of individuals, an abundance of single nucleotide polymorphisms (SNPs) is also observed. In the human genome, an SNP occurs in 1 out of every 1,000 base pairs (Sachidanandam et al., 2001; Venter et al., 2001). The SNPs have a relatively lower degree of polymorphism than microsatellite markers since there are only two possible pairs of nucleotides in DNA. However, they intensively cover the entire genome and may exist in the coding regions of functional genes and, hence, may represent the causative locus. Consequently, SNPs are more and more frequently used in genome research.

**Genome scan approach for QTL mapping with individual genotyping data**

Genetic markers, such as microsatellites and SNPs, can trace how those the chromosomal segments in which that are located are inherited across generations. If some genetic markers are physically located near the QTL, they tend to stay together (linked) over generations because the probability that a recombination occurs between two tightly linked loci is small. Thus, the genetic marker and the QTL will be in linkage disequilibrium (LD). However, there are many other reasons to observe LD between loci: non-random mating, selection, mutation, migration, etc. In the genome scan approach, the challenge for geneticists and
breeders is to distinguish the LD that leads the way to the QTL and to efficiently utilize it to map the QTL.

Different population structures and different levels of LD lead to different statistical analysis methods for QTL mapping studies using the genome scan approach. If the search is for QTL for a complex disease in humans, since the population has a long history, only LD due to tight linkage remains after many generations of meiosis and recombination. Therefore, if population-wide LD is detected between certain markers and the disease, it is likely that the QTL for the disease is located in a close neighborhood to the marker. Thus, utilizing population wide LD over generations is accurate and the statistical test is relatively simple: a t-test for significance of the association between marker genotype and phenotype. However, only a few markers will be tightly linked with the QTL and they are difficult to find. In addition, similar to association tests in the candidate gene approach, significance does not necessarily imply a causative effect and additional examination is required to demonstrate causality.

An alternative is to create population-wide LD by crossing two breeds or inbred lines (Weller et al., 1988; Andersson et al., 1994; Knott et al., 1998). This approach can be used to detect QTL that segregate between breeds or lines, but the estimate of the position of the QTL can be distant from the true position because of the extent of LD in crosses. This method has been extensively used in QTL mapping projects of plants and small animals, such as mouse or chicken, where it is feasible to obtain complete or close to completely inbred lines and it is less expensive than large animals to raise a specific population for a QTL mapping study. In
this approach, two breeds or lines with a large difference in performance are selected as parental lines, the QTL is assumed fixed in the alternative lines, and genetic marker alleles are line specific. The two lines are crossed and both phenotypic data and marker genotype are collected in the F$_2$ or backcross (BC) population. Since the two parental lines are completely or nearly completely inbred, marker alleles can be clearly traced back to their original parental lines and the genetic background for the QTL study is relatively simple. Therefore, the ability to detect QTL responsible for differences between lines is high.

Although this approach works best for a cross between completely inbred lines, there are some successful extensions to crossing outbred lines, where the genetic backgrounds are sufficiently different (Malek et al., 2001a,b). With LD created by breed or line crosses, QTL position estimates are not accurate because there is only one cycle of informative meioses. Thus, markers at 20cM distance from the QTL can still be in high LD with the QTL. The statistical analysis for the QTL can be a direct association test for phenotypic means with genotype at each marker but the interval mapping method is a more powerful and, hence, more popularly used method (Lander and Botstein, 1989). Instead of using one single marker, the interval mapping method utilizes information from flanking markers and examines the probability of harboring QTL at each possible position within the interval flanked by these two markers. In addition to higher power to detect the QTL, the interval mapping method also provides separate estimates for QTL location and QTL effect, whereas the single marker analysis can only identify which markers show close linkage with a QTL. Both maximum likelihood and least squares regression methods can be used to implement the interval mapping method in the line-cross QTL analysis (Lander and Botstein, 1989;
Haley and Knott, 1992). However, since most information about the QTL is captured by the difference in means, the least squares method provides very similar results as the maximum likelihood method (Haley and Knott, 1992). The least squares method is much easier to use and more readily extended to more complicated models.

In general, it is not feasible to create population-wide LD by breed crossing in livestock species such as dairy cattle or beef cattle. In addition, it is also important to detect QTL that segregate within the breed. For these species, QTL mapping must utilize information and population structures that are provided by the current breeding schemes. In populations of these species, there is usually limited population-wide LD, but LD remains within families. A complication with the use of this LD is that it is not consistent across families. That is, a certain marker allele that is associated with the favorable QTL allele in one family may be associated with the unfavorable allele in another family.

For QTL mapping based on within family LD, if the research objective is to estimate QTL effects for only the families of interest, the probability of an individual having a certain QTL genotype can be treated as fixed effects in the model and the QTL effects are separately estimated for each family (Weller et al., 1990). However, if the QTL in the whole population are of interest, the probability of having a certain QTL genotype should be treated as a random effect in the model and the marker information can be used to model covariance between relatives and used to map the QTL (Fernando and Grossman, 1989). The basic principle behind the latter method is that relatives that inherited the same marker allele from the co-ancestor tend to have a higher probability to inherit the same allele at a nearby QTL.
(probability of identity by descent, IBD) than the ones who did not and, therefore, tend to show a higher covariance. For some simple family structures, such as half-sib families, full-sibs families, and the granddaughter design, analysis of variance can be applied to estimate the variance contributed by the QTL, however, if the family structure is more complex, the maximum likelihood method is frequently used. In the likelihood approach, the likelihood of observing the phenotypic data given the observed marker genotype and the pedigree is derived and the QTL parameters that maximize the likelihood are estimated. To compute the covariance of relatives given the marker information, the method of Markov Chain Monte Carlo (MCMC) (Hastings, 1970), can be used to sample genotypes given marker information and phenotypic values. The MCMC method is also heavily used in combination with Bayesian methods, where some prior knowledge about the QTL parameters can be incorporated in the model.

No matter which level of LD exists in the population or which method is used to map the QTL, in general, the genome scan approach results in a wide region in which the QTL might exist. In order to narrow this region, fine mapping research is required. Fine mapping research needs a higher marker density in the region of interest and more recombination events to break the linkage between more distant markers and the QTL. More recombination can be obtained by creating more generations of intercrossing (advanced intercross lines, Darvasi and Soller, 1995), or by utilizing recombination that has accumulated over previous generations. Once the QTL is located to a very small region where a few function-known genes are identified, genes with relevant function can be considered as candidate genes and the candidate gene approach can be used for further analysis.
To identify the QTL by the genome scan approach, there are two challenges with regards to genotyping. First, a sufficient number of markers must be genotyped across the whole genome and, second, a high density of markers is required in the region that shows evidence of QTL. Since the genetic contribution of any specific gene to a quantitative trait may be small, usually several hundreds of markers from several hundreds of individuals are required to achieve a desired power to detect the QTL (Darvasi et al., 1993). Thus, the cost and labor of genotyping is very intensive, often limiting QTL mapping research. Furthermore, only a small number of markers genotyped will show LD with the QTL of interest. The problems of high genotyping cost also exist in saturating marker density. Thus, a simple and accurate screening for relevant loci across the whole genome will be of great value. The purpose of the rest of this literature review will be to describe one method that has been proposed and used to reduce the genotyping costs of genome scan QTL mapping: selective DNA pooling. This review formed the basis for the development of the objectives of the research presented in this thesis, as stated in Chapter 1.

**Principles of selective DNA pooling**

One possible way to reduce the genotyping cost is selective genotyping, in which only individuals with high and low phenotypes are genotyped, thereby reducing genotyping costs (Lander and Botstein, 1989). Since most information about the QTL is provided by individuals that deviate the most from the phenotypic mean, using selective genotyping can provide power to detect the QTL that is similar to genotyping the whole population
(Lebowitz et al., 1987; Lander and Botstein, 1989). This method is very efficient for QTL analysis of a single or a few traits and when the genotyping cost is large relative to the cost of rearing individuals and measuring the phenotypes (Lander and Botstein, 1989; Darvasi and Soller, 1992). However, severe upward biases in estimates of QTL effects were observed if the phenotypic values of the ungenotyped individuals were not included in the analysis (Lander and Botstein, 1989; Darvasi and Soller, 1992; Martinez et al., 1998), which is due to selection of extreme individuals in the population. To correct the bias in the selective genotyping analysis, a maximum likelihood method can be applied and phenotypes of all individuals included, treating the genotype of ungenotyped individuals as missing data (Lander and Botstein, 1989; Muranty and Goffinet, 1997; Muranty and Goffinet, 1997; Johnson et al., 1999). Recently Xu and Vogl (2000) presented another maximum likelihood method using only the phenotypic values of genotyped individuals, accounting for selection in the model. Although Xu and Vogl (2000) indicated that their method was relatively simple and readily implemented in a more complicated analysis, the maximum likelihood method is still intractable when the model becomes more complex and the computations are very intensive. On the other hand, Darvasi and Soller (1992) presented a simple ANOVA approach to correct the biased estimates in a single marker association analysis, which can be easily applied to general cases with reasonable power.

Selective DNA pooling goes one step further than selective genotyping, that is, DNA of the selected individuals in the same tail of the phenotypic distribution is pooled together, resulting in a high pool and a low pool, and marker allele frequencies are determined in both pools (Darvasi and Soller, 1994). The QTL allele frequency is expected to be different
between two pools, that is, the favorable allele should have a higher frequency in the high performance tail and be deficient in the low one. The opposite is expected for the unfavorable allele. If a marker is in LD with the QTL, the marker allele frequency will be "dragged along" by the QTL allele frequency. Therefore, different marker allele frequencies are expected in the two selected pools for linked markers and no difference is expected for unlinked markers. Therefore, by determining whether the marker allele frequency is significantly different from 0.5 (or, the difference of marker allele frequency between two pools is different from zero), one can determine which markers in the genome are associated with the QTL of interested traits (Darvasi and Soller, 1994; Khatib et al., 1994). Thus, selective DNA pooling only requires a few genotypings for each marker, regardless of the sample size in the pool and can substantially reduce the genotyping costs and labor.

The trade-off for substantial savings in genotyping costs is that selective DNA pooling loses some information for accurately detecting or mapping the QTL. First, there is no individual genotype information in the selective DNA pooling data, since all information is contained in the pooled allele frequencies. Second, the observed marker allele frequencies in pools are subject to a certain degree of error (Darvasi and Soller, 1994), which results from making pools, DNA amplification, and estimating the allele frequency in the pool (Barratt et al., 2002). The magnitude and effect of those errors on the QTL mapping results will be discussed later.

It is very critical to accurately estimate marker allele frequencies in the pool. Estimating allele frequencies in the pool is based on the linear relationship between the number of copies
of the allele in the pool and the final allele band intensity on the electrophoresis gel (Pacek et al., 1993). The more copies of a certain allele there are in the selected pool sample, the more intensity the corresponding band should have. Initially, qualitative comparison was used. That is, absence or presence of a certain allele band on the gel was subjectively visualized in the two pools from groups of individuals with different phenotype (Michelmore et al., 1991). However, since effects of most QTL are small to moderate, alleles are not necessarily fixed in the alternative pools. Recently, with the rapid development of molecular technology, it is possible to quantify relative marker allele frequencies in pools (Pacek et al., 1993; Khatib et al., 1994; Breen et al., 2000). The amplified DNA fragments are labeled with probes and alleles with different lengths are separated by electrophoresis. The electrophoresis gel or tube is scanned and the intensity of each band is turned into a quantitative peak height or area by commercial software and then is converted into the relative marker allele frequencies in the DNA pool. This quantitative technique has accelerated the application of selective DNA pooling in QTL mapping research. Detailed aspects of quantifying marker allele frequencies and possible problems will be discussed later.

Selective DNA pooling can detect QTL in a very efficient way but it will provide lower accuracy when compared with the individual genotyping approach, as discussed above. Therefore, selective DNA pooling usually serves as an initial screen for QTL and is followed by individual genotyping in the regions detected by selective DNA pooling (Bansal et al., 2002; Zhou et al., 2003).
Successful examples of selective DNA pooling

Utilizing marker allele frequencies in pooled DNA to efficiently detect associations between markers and responsible genes is not a new idea. As early as 1985, (Arnheim et al., 1985) successfully detected specific alleles (DR in HLA class II loci) associated with the insulin-dependent mellitus disease in the context of a case-control study. In that study, two DNA pools were formed from the diseased and normal individuals, respectively, and the density of bands from Southern blot was subjectively evaluated. Due to technical limitations in quantifying marker density at this early stage, this method was only used to detect genes for complex diseases in some isolated populations, for which the allelic diversity was limited (Puffenberger et al., 1994; Puffenberger et al., 1994; Sheffield et al., 1994; Carmi et al., 1995; Nystuen et al., 1996; Scott et al., 1996). After the technology of quantifying PCR products (Pacek et al., 1993) became mature, this method has been widely used in general case-control studies (Fisher et al., 1999; Plomin and Craig, 2001; Plomin et al., 2001; Williams et al., 2002). In the 1990’s, microsatellites were the most commonly used markers (Barcellos et al., 1997; Hill et al., 1999) but SNPs have become more popular in recent years (Bansal et al., 2002; Herbon et al., 2003). Pareek et al. (2002) provided a detailed list of recent achievements in detecting QTL for human complex diseases by DNA pooling procedures.

The selective DNA pooling procedures are also frequently used in plant genetics research, where it is called bulked segregant analysis (Michelmore et al., 1991). Since pure inbred lines of plants can be obtained, bulked segregant analysis is widely used in F₂ populations.
created by crossing divergent parental lines, either to detect the QTL (Altinkut and Gozukirmizi, 2003; Cai et al., 2003; Shen et al., 2003) or to add markers in an interesting region (Barua et al., 1993; Chague et al., 1999; Kikuchi et al., 2003; Lee et al., 2003). The use of bulk segregant analysis in plant breeding was recently reviewed by Masojc (2002).

Selective DNA pooling has also been successfully used in animals. Lamont et al. (1996) successfully detected QTL for several important egg production traits in layer chicken with VNTRs (DNA fingerprints). Taylor and Phillips (1996; 1997) detected an obesity QTL for mice in an F$_2$ family by selective DNA pooling. Lipkin et al. (1998) and Mosig et al. (2001) detected QTL for milk protein percentage and Tchourzyna et al. (2002) detected multiple QTL for several milk production traits on one chromosome in dairy cattle by selective DNA pooling. Zhou et al. (2003) used selective DNA pooling method as a first step scan for QTL in poultry and confirmed results by individual genotyping.

**Issues in application of selective DNA pooling**

1. Application to different family structures

Here we focus our interests on QTL mapping research in livestock. The selective DNA pooling approach is quite flexible and can be applied to different family structures, such as backcross, F$_2$ cross and half-sib families, by defining two groups with different QTL genotypes (Darvasi and Soller, 1994). In dairy cattle, because artificial insemination is a widely used reproduction technique, large half-sib families are routinely produced. Large
half-sib families provide high power to detect the QTL (Weller et al., 1990). Since selective DNA pooling is independent of sample size, this approach is extremely useful when applied to QTL mapping analysis in large half-sib families. The power to detect the QTL is sufficiently high, while the genotyping costs remain low. Several studies have successfully detected QTL for important milk production traits of dairy cattle by using the selective DNA pooling approach. In pigs and chickens, populations for QTL mapping are typically generated by a line or breed cross and an F_2 is the most frequently used family structure. Several successful examples of detecting QTL with selective DNA pooling in chickens using such a population structure exist (Lipkin et al., 2002; Zhou et al., 2003).

2. Possible errors in DNA pooling procedures

Unlike QTL mapping with individual genotyping, where the genotype information is clear and almost error free, marker allele frequencies in pools are always determined with a certain degree of error. Since all information about the QTL comes from a few observations of marker allele frequency, an accurate, quantitative estimate of marker allele frequencies in the selected pools must be obtained.

Sources of experimental error in data on marker allele frequencies from selective DNA pooling can be classified as: (1) pool formation, which requires equal amounts of DNA from the selected individuals to be included in the pool; (2) DNA amplification by PCR; and (3) estimation of marker allele frequencies in the pool, referred to as frequency estimation (Barratt et al., 2002). Barratt et al. (2002) estimated the error variance involved in different
sources by a hierarchical analysis of variance for SNP markers. Estimates for the above three classes were $1.0 \times 10^{-4}$, $2.9 \times 10^{-4}$ and $5.5 \times 10^{-4}$, respectively. Results from this research suggested that the main errors in the final frequency estimates exist in the PCR and frequency estimation steps.

Technical errors in pool formation mainly depend on technical procedures in the lab and skills of the operator, which can be effectively reduced by setting up appropriate DNA extraction and dilution schemes and by making replications. Although replications at all levels can reduce technical error, replication of pool formation is more effective and economically efficient than replication at later stages, with the same amount of genotyping work (Barratt et al., 2002).

The other two steps for obtaining DNA pooling data, PCR amplification and frequency estimation, are almost automatic procedures and relatively independent of operator skills. In these two steps, two main problems are differential amplification and shadow bands (called stutter bands in human genetics research). Some other factors, such as different fluorescence dyes and sample size, might also cause problems but their effects do not appear to be significant (Shaw et al., 1998; Barratt et al., 2002).

Differential amplification is observed when two alleles with different lengths are amplified. The shorter allele is usually replicated faster than the longer allele during PCR (Demers et al., 1995). Since the longer allele has a higher chance to reanneal, which reduces the efficiency of PCR amplification, the shorter allele tends to show a higher frequency in the
DNA pool, which can bias estimates of marker allele frequencies in the DNA pooling approach.

Shadow bands only appear for microsatellite markers and are prevalent in dinucleotide repeats. During PCR amplification, the Taq polymerase sometimes "slips" on the DNA strand, which results in a product that is one or two repeats smaller than the original allele. Therefore, on the image of the electrophoresis gel, several lighter bands appear around the main allele band, i.e. "shadow bands" (Hauge and Litt, 1993; Litt et al., 1993; Murray et al., 1993). Shadow bands introduce biases in marker allele frequency estimates in the DNA pooling approach if only the main band is scored. Some studies have successfully removed biases due to differential amplification and shadow bands by mathematical adjustments (Barcellos et al., 1997; Daniels et al., 1998; Lipkin et al., 1998).

Several studies have evaluated the accuracy of the estimated allele frequencies from DNA pools for microsatellite markers (Barcellos et al., 1997; Daniels et al., 1998; Shaw et al., 1998) and the more frequently used SNP markers (Breen et al., 2000; Giordano et al., 2001; Norton et al., 2002). All studies showed that, under certain conditions, estimates from DNA pools were accurate when compared with individual genotyping (Mohlke et al., 2002; Shifman et al., 2002) and the QTL detected by the DNA pooling approach were consistent with results from the individual genotype approach (Lipkin et al., 1998; Shaw et al., 1998; Zhou et al., 2003).
Since DNA exists in milk somatic cells and red blood cells of chicken, it is also possible to pool milk or blood instead of pooling isolated DNA from selected individuals, which requires more labor. Research has shown that accurate estimates of allele frequency can be achieved by pooling milk or blood under appropriate procedures (Lipkin et al., 1998; Lipkin et al., 2002).

**Statistical analysis of selective DNA pooling data**

In addition to accurate estimation of marker allele frequencies in pools, development of statistical methods to effectively analyze selective DNA pooling data is also critical. The ideal analysis method should be powerful and accurate. That is, the method should show few false positives and false negatives, and provide information about the QTL location and its effect.

The current method to detect QTL using selective DNA pooling data is based on single marker analysis, which was described in Darvasi and Soller (1994). In this method, linkage to a QTL is independently tested for each marker. Under the null hypothesis of no QTL, marker allele frequencies are expected to be equal in both selected pools since genetic markers themselves have no effect on the phenotypic value. If a single family is analyzed, differences between the observed marker allele frequencies between the pools can be tested by a Z-test for each marker. The standard error of marker allele frequencies for use in this test can either be estimated from the data or derived from sample size and prior knowledge of the variance of technical error. If multiple families are analyzed, the Z statistics can be
squared and summed across families for the same marker. The sum of squares of the $Z$ statistics follows a Chi-square distribution with the number of families as the degrees of freedom. This method is readily extended to different family structures, such as backcross, half-sib, and $F_2$ families. Lipkin et al. (1998) used this method and successfully detected QTL for milk protein percentage in dairy cattle. Mosig et al. (2000) followed the same procedure and population as Lipkin et al. (1998) but extended it to a whole genome scan for the same trait. Marker-QTL linkage was detected by the single marker analysis but the adjusted false discovery rate (Benjamini and Hochberg, 1995) was used as a control for false positives.

Baro et al. (2001) investigated the effects of various factors on the power of the single marker analysis to detect QTL by simulation and by derivation of the exact distribution of marker allele frequencies in pools of a half-sib family. Factors that were investigated included: the magnitude of the additive effect of the QTL, dominance level of the QTL, QTL allele frequencies, marker polymorphism, distance between marker and the QTL, selected proportion, family size, and magnitude of the technical error. Results suggested that the magnitude of the additive QTL effect, dominance level of the QTL, and family size were the most important factors for power and effects of other factors were relatively moderate or small.

Carleos et al. (2002) presented a new single marker method to impute the unobservable individual genotypes in the pool given the observed marker allele frequencies and phenotypic values. In this method, genotypes were linked to phenotypic values for the selected
individuals and it, therefore, can extract more information from the pooling data than the traditional single marker analysis (Darvasi and Soller, 1994). Simulation results indicated that this method provided a more precise estimate of QTL effects. However, this method was presented in an oversimplified framework: no recombination between marker and the QTL and no technical error in estimates of marker allele frequencies. How this method can provide estimates of QTL location and can be extended to a more practical situation is not clear.

The single marker method detects QTL based on information from single markers and does not provide separate estimates of QTL location and QTL effect, nor does it utilize the joint information from multiple markers. Dekkers (2000) developed a method to provide separate estimates of QTL location and effect by using pooling data of flanking markers. The observed marker allele frequencies in the selected pools were modeled as a linear function of QTL allele frequency in the same pool and the recombination between markers and the QTL. Estimates of QTL location and QTL allele frequency were analytically solved from the frequencies of the two flanking markers. The QTL effect was then estimated from the estimates of QTL allele frequencies, phenotypic means of selected tails, and the mean of the whole population. Simulation results suggested that this method provided nearly unbiased estimates when the power was high, but considerable bias when the power was low. In addition, some replicates did not have a solution and some provided estimates outside the parameter space. Also, information of markers outside the flanking markers was not used and this method can not be used for analysis of multiple families.
Carleos et al. (2003) derived the asymptotic variance of the QTL estimators from the Dekkers (2000) method. However, simulation results showed that an impractically large family size was needed to have reasonable estimates of sampling variance and variance because technical errors were not taken into account.


CHAPTER 3. INTERVAL MAPPING OF QTL WITH SELECTIVE DNA POOLING DATA

I. METHODOLOGIES

In preparation for submission to: Genetics

ABSTRACT

Selective DNA pooling is an efficient method to identify chromosomal regions that harbor quantitative trait loci (QTL) by comparing marker allele frequencies in pooled DNA from phenotypically extreme individuals. The currently used single marker analysis can detect linkage of markers to a QTL but does not provide separate estimates of QTL position and effect, nor does it utilize the joint information from multiple markers. In this study, two interval mapping methods for analysis of selective DNA pooling data were developed. One was based on least squares regression (LS-pool) and the other on approximate maximum likelihood (ML-pool). Both methods simultaneously utilize information from multiple markers and multiple families and both are easily applied to different family structures (half-sib, F2 cross and backcross). The QTL mapping results from these two interval mapping methods were compared with results from single marker analysis by simulation and application to a real dairy cattle data set. Simulation results showed that both interval mapping methods resulted in nearly unbiased estimates of QTL location and effect, and slightly greater power than single marker analysis. In the dairy cattle data application, both interval mapping methods solved problems of single marker analysis when missing marker genotypes were present by utilizing joint information from multiple markers.
Detecting genes underlying quantitative variation (quantitative trait loci or QTL) with the aid of molecular genetic markers is an important topic in both animal and plant breeding research. However, for QTL with small or moderate effects, much genotyping is needed to achieve a desired power (DARVASI et al. 1993) and the genotyping cost is usually prohibitive.

Selective DNA pooling is an efficient method to detect linkage between markers and QTL by comparing marker allele frequencies in pooled DNA from phenotypically extreme individuals (DARVASI and SOLLER 1994). Marker allele frequencies can be estimated by quantifying PCR product in the pool (PACEK et al. 1993) and linkage to a QTL can be detected by conducting a significance test at each marker. This approach successfully detected QTL in a half-sib family design of dairy cattle (LIPKIN et al. 1998; MOSIG et al. 2001) and in F2 designs of chickens (LIPKIN et al. 2002; ZHOU et al. 2003). Estimates of marker allele frequencies were accurate and the detected QTL were confirmed by selective genotyping (LIPKIN et al. 1998; ZHOU et al. 2003).

Current analyses of selective DNA pooling data are based on single marker analysis (DARVASI and SOLLER 1994), which cannot provide separate estimates of QTL location and QTL effect, or utilize the joint information from multiple linked markers around a QTL. Interval mapping methods have been developed to get around these problems for individual genotyping data (LANDER and BOTSTEIN 1989) but have not been developed for selective DNA pooling data. DEKKERS (2000), however, showed that pool frequencies for flanking markers contain information to map a QTL within an interval. The observed marker allele frequencies in the selected DNA pools were modeled as a linear function of QTL allele
frequency in the same pool and recombination rates between markers and the QTL. 
DEKKERS (2000) showed that QTL location and QTL allele frequency could then be solved 
for analytically based on observed frequencies at two flanking markers. Simulation showed 
that this method provided nearly unbiased estimates when power was high, but considerable 
bias when power was low (DEKKERS 2000). In addition, estimates did not exist for some 
replicates and some replicates provided estimates outside the parameter space. This method 
is also not suitable for pooled analysis of multiple families and only used data from flanking 
markers and not from markers outside the interval. External markers provide information to 
map QTL in an interval in the case of DNA pooling data because of technical errors that are 
associated with observed frequencies. The objective of the current study, therefore, was to 
develop interval mapping methods to overcome these problems. Two methods that allow 
simultaneous analysis of selective DNA pooling data from multiple markers and multiple 
families were developed. One was based on least squares regression (LS-pool) and the other 
on approximate maximum likelihood (ML-pool). Methods were first evaluated by simulation 
and then applied to a subset of the selective DNA pooling data for protein % in dairy cattle of 
LIPKIN et al. (1998). This subset was previously analyzed using single marker analysis by 
TCHOURZYNA et al. (2002). The population structure used in simulation mimicked the dairy 
cattle data set,

MATERIALS AND METHODS

**Basic principles to detect QTL using selective DNA pooling data:** Using a single half-
sib family as an example, consider a sire that is heterozygous at both a QTL (Qq) and a
nearby marker (Mm). The sire is mated with multiple dams randomly chosen from a population with QTL allele frequency 0.5 and for which the marker is in linkage equilibrium with the QTL. The progeny from this sire can then be classified into two groups: progeny that received the “Q” allele from the sire and progeny that received the “q” allele. Assuming “Q” is the favorable allele for the trait, then the “Q” progeny are expected to have a higher phenotypic mean than the “q” group. The QTL allele from dams, polygenic effects and environmental factors contribute to variation within each group of progeny, and phenotypic values within each group are assumed normally distributed (Figure 1).

All progeny within the half-sib family are ranked according to their phenotypic value and the highest and lowest p% progeny are selected. An equal amount of DNA is extracted from each selected individual and DNA from individuals in the same selected tail is pooled together to form an upper pool and a lower pool. Since the number of progeny that received the “Q” allele from the sire in the upper pool is expected to be greater than the number of progeny that received the “q” allele (Figure 1), the expected frequency of the sire’s "Q" allele in the upper pool \( (P^U_Q) \) will be higher than the expected frequency of the sire’s "q" allele \( (P^U_q) \). The opposite is expected for the lower pool \( (P^L_q < P^L_Q) \).

Frequencies of QTL alleles can not be observed directly. However, if a marker is closely linked to the QTL and the “M” allele is in linkage phase with the “Q” allele in the sire, the expected frequency of allele “M” in the upper pool \( (P^U_M) \) also will be higher than the expected frequency of allele “m” \( (P^U_m) \). On the other hand, if the marker is far away from the QTL, then marker allele frequencies are expected to be equal to 0.5 in both pools. Therefore, marker allele frequencies contain information about linkage with a QTL.
The observed marker allele frequencies in the pools ($f_{M}^{U}$ and $f_{M}^{L}$) will deviate from their expectations ($p_{M}^{U}$ and $p_{M}^{L}$) due to sampling errors ($e_{s}$) and technical errors ($e_{t}$). Sampling errors are due to the finite number of individuals in each pool. The smaller the number of individuals in the pool is, the larger the sampling errors will be. Technical errors are generated during the procedures of creating the pools and of estimating marker allele frequencies in the pools.

Due to the limited number of observations (only two observations per marker, one from each selected pool, if no replicates are considered) and inevitable technical errors, it is necessary to develop statistical methods to powerfully and accurately utilize the information about the QTL in the observed marker allele frequencies.

**Single marker association analysis:** The method to detect linkage between markers and a QTL with selective DNA pooling proposed by DARVASI and SOLLER (1994) is based on a significance test at each marker. Under the null hypothesis, the distribution of marker allele frequencies is not affected by the QTL, so the expected allele frequency is 0.5 in both pools. Therefore, using an approximate normal distribution, the null hypothesis is rejected with type I error $\alpha$ if

$$Z_{i} = \frac{(f_{Mij}^{U} + f_{Mij}^{L}) - 0.5}{\sqrt{\frac{2}{\text{Var}(\frac{f_{Mij}^{U} + f_{Mij}^{L}}{2})}}} > Z_{1-\alpha/2}$$

where $f_{Mij}^{U}$ and $f_{Mij}^{L}$ are the observed frequencies of the alternative alleles ("M" and "m") in the upper (U) and lower (L) pools for the $j^{th}$ marker in the $i^{th}$ family and $Z_{1-\alpha/2}$ is the ordinate of the standard normal distribution such that the area from $-\infty$ to $Z_{1-\alpha/2}$ equals $1 - \alpha/2$. Because
the expectation of $f_{M_{ij}}^U$ is equal to $f_{m_{ij}}^U$, assuming symmetry, the average of two independent observation. $f_{M_{ij}}^U$ and $f_{m_{ij}}^U$ is used as an estimate of the marker allele frequency in the pools.

The variance of the estimated marker frequency under the null hypothesis includes variance of sampling errors and variance of technical errors, and can be derived as (DARVASI and SOLLER 1994):

$$Var\left(\frac{f_{M_{ij}}^U + f_{m_{ij}}^L}{2}\right) = \frac{1}{2} \left( \frac{0.25}{n} + V_{TE} \right)$$

where 0.25 is the expected variance of the marker allele frequency from a binomial distribution under the null hypothesis of no QTL, $n$ is the number of individuals in each pool (assumed the same in both pools), and $V_{TE}$ is the variance of the technical error in observed marker allele frequencies in a pool. In this study, the value of $V_{TE}$ was assumed known and obtained from previous studies by comparing pool estimates of marker allele frequencies with the true frequency, as obtained from individual genotyping, rather than estimated directly from pool data (LIPKIN et al. 1998). Sampling errors were assumed to be independent of technical errors.

If information from $m$ families is available, the Z-test for each family can be extended to a Chi-square test, assuming observations from multiple families are independent (DARVASI and SOLLER 1994). That is, for each marker $j$, the null hypothesis is rejected with type I error $\alpha$ if

$$\sum_{j=1}^{m} Z_{ij}^2 > \chi^2_{\alpha,m} \quad df = m$$
where \( m \) is the number of families, \( \alpha \) is the type I error, and \( \chi^2_{\alpha m} \) is the ordinate of the central Chi-square distribution with degrees of freedom \( m \), such that the area from 0 to \( \chi^2_{\alpha m} \) equals \( 1 - \alpha \).

When several markers are available on a chromosome or within a chromosomal region, the marker with the most significant test statistic is considered as the marker closest to the QTL. However, since multiple tests are involved and information from multiple linked markers are correlated, appropriate chromosome-wise significance thresholds must be derived either by false discovery rate (MOSIG et al. 2001) or by simulation.

**Interval mapping methods to map QTL with selective DNA pooling data:** Single marker analysis identifies markers that are linked to the QTL but provides no estimate of the position of the QTL, nor of its effect. The purpose of this section is to develop interval mapping methods with selective DNA pooling data to overcome the problems of single marker analysis.

**DEKKERS (2000)** presented a method to map the QTL using pooling data from two markers that flank an interval. The expected marker allele frequency in a selected pool (\( p_M^U \) or \( p_M^L \)) was modeled in terms of the expected QTL allele frequency in the same pool (\( p_Q^U \) or \( p_Q^L \)) and the recombination rate (\( r \)) between the marker and the QTL:

\[
p_M^U = (1 - r)p_Q^U + r(1 - p_Q^U)
\]

\[
p_M^L = (1 - r)p_Q^L + r(1 - p_Q^L)
\]

The same principle was employed in here, but extended to multiple markers and multiple families. Two interval mapping methods were developed, one based on least squares
regression (LS-pool) and one based on approximate maximum likelihood (ML-pool). Both methods assumed that sire marker haplotypes were known.

**Least squares interval mapping (LS-pool):** Suppose a chromosome contains $k$ markers and a single QTL, and marker positions are known (Figure 2). The distribution of phenotypic values is assumed to be symmetric, such that the QTL allele frequency in the upper pool, $p_Q^U$, is expected to be equal to one minus the expected frequency in the lower pool, $1-p_Q^L$ (Figure 1). Using these assumptions and including random errors, previous equations were rearranged and extended to multiple markers as follows:

\[
\begin{bmatrix}
    f_{M_{11}}^U - 1/2 \\
    f_{M_{12}}^U - 1/2 \\
    \vdots \\
    f_{M_{k1}}^U - 1/2 \\
    f_{M_{11}}^L - 1/2 \\
    f_{M_{12}}^L - 1/2 \\
    \vdots \\
    f_{M_{k1}}^L - 1/2 \\
\end{bmatrix}
\begin{bmatrix}
    1 - 2r_1 \\
    1 - 2r_2 \\
    \vdots \\
    1 - 2r_k \\
    - (1 - 2r_1) \\
    - (1 - 2r_2) \\
    \vdots \\
    - (1 - 2r_k) \\
\end{bmatrix}
\begin{bmatrix}
    p_{Q_i}^U - 1/2 \\
    e_i^U \\
    e_{i2}^U \\
    \vdots \\
    e_{ik}^U \\
    e_i^L \\
    e_{i2}^L \\
    \vdots \\
    e_{ik}^L \\
\end{bmatrix}
\]

(Model 1)

where $f_{M_{ij}}^U$ and $f_{M_{ij}}^L$ are observed marker allele frequencies in pools, as defined previously, $p_{Q_i}^U$ is the expected QTL allele frequency in the upper pool for the $i^{th}$ family, $r_j$ is the recombination rate between the $j^{th}$ marker and the QTL, and $e$ represent residuals, including sample error ($e_s$) and technical error ($e_t$). For development of the method, no assumptions are required for the distribution of residuals.

For a given putative position of the QTL, the recombination rate between each marker and the QTL ($r_j$) is known, and Model 1 can be written as

\[ Y_i = X_i \beta_i + e_i \]
where \( \mathbf{Y}_i \) is a vector of deviations of observed marker allele frequencies in both pools for the \( i^{th} \) family from the expectation under the null hypothesis of no QTL, \( \frac{1}{2}, \mathbf{X}_i \) is a known vector containing recombination rates of each marker with the putative QTL, which is same for all families, and \( \beta_i \) is the parameter to be estimated for the \( i^{th} \) family and represents the deviation of the expected QTL allele frequency in the upper pool from its expectation under the null hypothesis of \( p_Q^{(i)} - \frac{1}{2} \).

Similar to least squares interval mapping with individual genotyping data (Haley and Knott 1992), at each putative QTL position, ordinary least squares is used to estimate the parameter, \( \beta_i = (p_Q^{(i)} - \frac{1}{2}) \), according to Model 1. Note that with this method, correlation among observations on linked markers is not taken into account. At each position, the following test statistic is calculated to test for significance of the QTL:

\[
\chi^2 = \frac{SS_{regression}}{Var(f_M)_{null}} = \frac{\mathbf{Y}_i' \mathbf{X}_i (\mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{X}_i' \mathbf{Y}_i / 1}{0.25 \left( \frac{1}{n} + V_{TE} \right)}
\]

where \( SS_{regression} \) is the sum of squares of regression.

The position providing the highest value of the test statistic is taken as the estimated QTL position and the estimated QTL allele frequency at this position is used to estimate the QTL substitution effect for sire \( i, \hat{\alpha}_i \), following Dekkers (2000):

\[
\hat{\alpha}_i = \frac{(\bar{x}_{Q_i} - \bar{x}_{q_i})}{i_s}
\]

where \( \bar{x}_{Q_i} \) and \( \bar{x}_{q_i} \) are weighted means of phenotypic values of progeny in the pools that received "Q" and "q" from sire \( i \) (defined below) and \( i_s \) is the selection intensity that
corresponds to the selected fraction \( p \), which can be obtained from \textit{Falconer} and \textit{Mackay} (1996). Means \( \bar{x}_{Q_i} \) and \( \bar{x}_{q_i} \) are computed as (\textit{Dekkers} 2000):

\[
\bar{x}_{Q_i} = \hat{p}_{Q_i}^U (\bar{x}_i^U - \bar{x}_i) + (1 - \hat{p}_{Q_i}^U) (\bar{x}_i^L - \bar{x}_i)
\]

\[
\bar{x}_{q_i} = (1 - \hat{p}_{Q_i}^U) (\bar{x}_i^U - \bar{x}_i) + \hat{p}_{Q_i}^U (\bar{x}_i^L - \bar{x}_i)
\]

where \( \hat{p}_{Q_i}^U \) is the estimated QTL allele frequency in the upper truncated population at the estimated position, \( \bar{x}_i^U \) and \( \bar{x}_i^L \) are the phenotypic means of progeny in the upper and lower pools of the \( i^{th} \) sire family, and \( \bar{x}_i \) is the mean phenotypic value of all progeny of the same family.

With \( m \) independent sire families, the least squares model can be extended to

\[
Y = X\beta + e
\]

where \( Y = [Y_1', Y_2', \ldots, Y_m']' \), \( \beta = [\beta_1, \beta_2, \ldots, \beta_m]' \), \( e = [e_1', e_2', \ldots, e_m']' \), and \( X \) is a matrix with \( X_i \) as blocks on the diagonal and zero in all other places. Matrix \( X_i \) is same for all families but estimates of QTL allele frequency, \( \beta_i \), are allowed to differ by family to allow for different QTL genotypes of sires (QQ, Qq, qQ and qq).

Similar to analysis of a single family, the vector of parameters (\( \beta \)) is estimated at each putative QTL position by ordinary least squares and the following test statistic across all families is calculated at each possible position:

\[
\chi^2 = \sum_{i=1}^{m} \chi_i^2
\]

The position with the highest test statistic provides the estimate of QTL position and the corresponding estimates of QTL allele frequencies in the upper pool of each sire family, \( \hat{\beta}_i \),
are used to estimate the QTL substitution effect for each sire ($\hat{\alpha}_s$), similar to the single family case.

Since the location with the maximum value of test statistic is searched among multiple positions and correlations exist among allele frequencies of linked markers, the test statistic $\chi^2$ does not have a Chi-square distribution. Permutation or bootstrap methods, which are frequently used to derive empirical significance thresholds for QTL interval mapping with individual genotyping data (CHURCHILL and DOERGE 1994), cannot be applied here because of too few observations. Therefore, chromosome-wise significance thresholds at the 5% level were derived by simulation under the null hypothesis of no QTL.

**Approximate maximum likelihood interval mapping method (ML-pool):** Sampling errors in frequencies of multiple markers are correlated due to linkage. The LS-pool method does not account for correlations among residuals, which reduces the efficiency of the method. Thus, an approximate maximum likelihood method, ML-pool, is developed to overcome this problem.

Observed marker allele frequencies can be presented as

$$f_{M_{ij}} = f_{M_{ij}}^{\text{true}} + e_{ij}$$

where $f_{M_{ij}}$ is the observed marker frequency in a selected pool for the $j^{th}$ marker of the $i^{th}$ family and $f_{M_{ij}}^{\text{true}}$ is the true marker allele frequency in that pool from the given sample of individuals.

For each marker, there are only two possible sire alleles in the sample: $M$ or $m$. Therefore, $n^* f_{M_{ij}}^{\text{true}}$ is a random variable following a binomial distribution, where $n$ is the
sample size. The technical error \( e_{\text{te}} \) is assumed to result from multiple technical factors. Thus, by the Central Limit Theorem, \( e_{\text{te}} \) can be assumed normally distributed. In this study, the variance of \( e_{\text{te}} \) \( (V_{\text{TE}}) \) is assumed known. Consequently, the distribution of observed marker allele frequencies is a mixture of a binomial sampling distribution and a normal distribution for technical errors. Since the sample size is usually sufficiently large \( (n>30) \), the mean of this mixed distribution is around 0.5 and since the variance of \( e_{\text{te}} \) is small \( (<0.001) \) (LIPKIN et al. 1998), the distribution of observed marker frequencies can be approximated by a normal distribution and the probability that the modeled frequencies fall outside the parameter space \((0-1)\) can be ignored. Thus the distribution of \( f_{M} \) can be approximated by a normal distribution (assuming \( f_{M_{ij}}^{\text{True}} \) and \( e_{M_{ij}} \) are independent) and by a multivariate normal distribution for multiple marker frequencies. The mean and covariance matrix of this multivariate normal distribution is as presented below.

Using Model 1, the vector of means \( (\mu_{i}) \) of observed allele frequencies of \( k \) markers in both selected pools in the \( i^{th} \) family can be presented as:

\[
\mu_{i} = E \begin{bmatrix} f_{M_{1j}}^{U} \\ f_{M_{12}}^{U} \\ \vdots \\ f_{M_{kj}}^{U} \\ f_{M_{1j}}^{L} \\ f_{M_{12}}^{L} \\ \vdots \\ f_{M_{kj}}^{L} \end{bmatrix} = E \begin{bmatrix} f_{M_{11}}^{\text{True}} \\ f_{M_{12}}^{\text{True}} \\ \vdots \\ f_{M_{kk}}^{\text{True}} \\ f_{M_{11}}^{L} \\ f_{M_{12}}^{L} \\ \vdots \\ f_{M_{kk}}^{L} \end{bmatrix} = \begin{bmatrix} p_{M_{11}}^{U} \\ p_{M_{12}}^{U} \\ \vdots \\ p_{M_{kk}}^{U} \\ p_{M_{11}}^{L} \\ p_{M_{12}}^{L} \\ \vdots \\ p_{M_{kk}}^{L} \end{bmatrix} = \begin{bmatrix} (1-r_{1})p_{Q_{1}}^{U} + r_{1}(1-p_{Q_{1}}^{U}) \\ (1-r_{2})p_{Q_{1}}^{U} + r_{2}(1-p_{Q_{1}}^{U}) \\ \vdots \\ (1-r_{k})p_{Q_{1}}^{U} + r_{k}(1-p_{Q_{1}}^{U}) \\ (1-r_{1})(1-p_{Q_{1}}^{L}) + r_{1}p_{Q_{1}}^{L} \\ (1-r_{2})(1-p_{Q_{1}}^{L}) + r_{2}p_{Q_{1}}^{L} \\ \vdots \\ (1-r_{k})(1-p_{Q_{1}}^{L}) + r_{k}p_{Q_{1}}^{L} \end{bmatrix}
\]
where \( f_{M_{ij}}^{U} \) is the true marker allele frequency for marker \( j \) in the upper pool of the \( i^{th} \) family, which differs from the expected allele frequency in the same pool, \( p_{M_{ij}}^{U} \), by the sampling error \((e_{k})\). Other notations are as defined previously.

The covariance matrix \((\Sigma_{i})\) of observed allele frequencies of \( k \) markers in the two selected pools in the \( i^{th} \) family is:

\[
\Sigma_{i} = \begin{bmatrix}
\Sigma_{i}^{U} & 0 \\
0 & \Sigma_{i}^{L}
\end{bmatrix}
\]

where \( \Sigma_{i}^{U} \) and \( \Sigma_{i}^{L} \) are the covariance matrices for the upper and lower pools, respectively,

\[
\Sigma_{i}^{U} = \Sigma_{i}^{L} = 
\begin{pmatrix}
\text{var}(f_{M_{ij}}^{U}) & \text{cov}(f_{M_{ij}}^{U}, f_{M_{ij2}}^{U}) & \cdots & \text{cov}(f_{M_{ij1}}^{U}, f_{M_{ik}}^{U}) \\
\text{cov}(f_{M_{ij1}}^{U}, f_{M_{ij2}}^{U}) & \text{var}(f_{M_{ij2}}^{U}) & \cdots & \text{cov}(f_{M_{ij2}}^{U}, f_{M_{ik}}^{U}) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(f_{M_{ij1}}^{U}, f_{M_{ik}}^{U}) & \text{cov}(f_{M_{ij2}}^{U}, f_{M_{ik}}^{U}) & \cdots & \text{var}(f_{M_{ik}}^{U})
\end{pmatrix}
\]

where \( \text{var}(f_{M_{ij}}^{U}) \) and \( \text{var}(f_{M_{ij}}^{L}) \) are equal to:

\[
\text{var}(f_{M_{ij}}^{U}) = \text{var}(f_{M_{ij}}^{L}) = \text{var}(f_{M_{ij}}^{\text{true}} + e_{ij}) = \text{var}(f_{M_{ij}}^{\text{true}}) + V_{TE} = \frac{p_{M_{ij}}^{U} (1 - p_{M_{ij}}^{U})}{n} + V_{TE}
\]

If markers \( j \) and \( l \) bracket the QTL \((M_{j}Q-M_{l})\) and the recombination rate between them is \( r_{jl} \), then \( \text{cov}(f_{M_{ij}}^{U}, f_{M_{il}}^{U}) \) is (see Appendix for detailed derivation):

\[
\text{cov}(f_{M_{ij}}^{U}, f_{M_{il}}^{U}) = \text{cov}(f_{M_{ij}}^{L}, f_{M_{il}}^{L}) = \text{cov}(f_{M_{ij}}^{\text{true}} + e_{ij}, f_{M_{il}}^{\text{true}} + e_{il}) = \text{cov}(f_{M_{ij}}^{\text{true}}, f_{M_{il}}^{\text{true}}) = \frac{(1 - 2r_{jl}) p_{Q_{ij}}^{U} (1 - p_{Q_{ij}}^{U})}{n}
\]

If the marker order is \((M_{j}M_{l}Q)\), the \( \text{cov}(f_{M_{ij}}^{U}, f_{M_{il}}^{U}) \) is (see Appendix):
By conditioning on the proportion selected for the upper and lower pool within a family, marker frequencies from the upper and lower pool are uncorrelated.

Both the vector of means \( \mu_i \) and the covariance matrix \( \Sigma_i \) are functions of \( p_{Q_i}^U \), the expected QTL frequency in the upper pool of the \( i \)th family, and of \( r \), the vector of recombination rates between markers and the QTL, which is determined by QTL location. Consequently, for a given QTL location (\( L_Q \)) and a certain value of \( p_{Q_i}^U \), the likelihood function for the vector of observed allele frequencies of \( k \) markers for the \( i \)th family (\( f_{Mi} \)), based on approximation to multivariate normality, is:

\[
L(f_{Mi}) = (2\pi)^{\frac{k}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp(f_{Mi} - \mu_i)^T \Sigma_i^{-1} (f_{Mi} - \mu_i)
\]

Under the null hypothesis of no QTL, \( p_{Q_i}^U = \frac{1}{2} \) and \( L(f_{Mi}) \) is a constant \( L_0(f_{Mi}) \) and does not depend on QTL location. Under the alternative hypothesis, the likelihood function \( (L_A(f_{Mi})) \) can be maximized for \( p_{Q_i}^U \) at a given QTL position (\( L_Q \)) using a golden-section search method (HEATH 2002) and the following log likelihood ratio statistic \( (LR) \) can be calculated at that position

\[
LR(L_Q, p_{Q_i}^U) = \ln\left(\frac{L_a(f_{Mi})}{L_0(f_{Mi})}\right)
\]
Each putative QTL position along the chromosome is tested and the pair of parameters \((L_Q\) and \(p_Q\)) providing the highest \(LR\) gives the estimates of QTL position and QTL allele frequency. The latter is used to estimate the QTL effect for the \(i^{th}\) sire by the method described for LS-pool.

The ML-pool method is also readily extended to multiple families. With \(m\) independent half-sib families, the likelihood function under the null and alternative hypotheses can be written as:

\[
L(f_M) = \prod_{i=1}^{m} L(f_{M_i})
\]

Note that the QTL location parameter \((L_Q)\) is common to all families but QTL allele frequency parameters \((p_Q)\) are allowed to differ by family. The set of parameters providing the highest \(LR\) provides the estimates of QTL parameters. Similar to LS-pool, chromosome-wise significance thresholds are derived by simulation under the null hypothesis of no QTL.

**Simulation model and parameters:** Ten half-sib families with 1,000 progeny per family were simulated. The simulated population structure was designed to mimic a dairy cattle data set used for a selective DNA pooling study by Lipkin *et al.* (1998) and MOSIG *et al.* (2001). For each individual, six fully informative markers were evenly spaced on a 100 cM chromosome (including markers at the ends). All dam alleles were assumed different from sire alleles and in population-wide linkage equilibrium with the QTL. Crossovers were generated according to the Haldane mapping function (see Appendix), which implies independence of recombination events in adjacent intervals on the chromosome. A single bi-allelic QTL with population frequency 0.5 was simulated at 36 cM from one end of the chromosome. The additive effect of the QTL was 0.25 standard deviations of phenotypic
value ($\sigma_p$), and the dominance effect was 0. Phenotype was affected by the QTL along with polygenic effects and environmental factors, which were both normally distributed. The variance explained by the QTL ($\sigma_{QTL}^2$) is 0.03125 $\sigma_p^2$ (FALCONER and MACKAY 1996) and the variance explained by polygenes ($\sigma_u^2$) is $h^2 \sigma_p^2 - \sigma_{QTL}^2$. The variance of phenotype ($\sigma_p^2$) was set to 1 and the heritability ($h^2$) was 0.25. The model for the phenotypic value of each progeny was:

$$y_{ij} = \mu + QTL_{ij} + Usire_i + Udam_{ij} + UMendel_{ij} + \varepsilon_{ij}$$  \hspace{1cm} \text{(Model 2)}$$

where

$$y_{ij} = \text{phenotypic value of progeny } j \text{ of sire } i,$$

$$\mu = \text{overall mean},$$

$$QTL_{ij} = \text{QTL effect of progeny } j \text{ of sire } i, \text{ fixed effect},$$

$$Usire_i = \text{polygenic effect of sire } i \text{ transmitted to an average progeny, assumed to follow } N(0,0.25 \sigma_u^2),$$

$$Udam_{ij} = \text{polygenic effect of dam } j \text{ mating to sire } i \text{ transmitted to its progeny, assumed to follow } N(0,0.25 \sigma_u^2),$$

$$UMendel_{ij} = \text{polygenic effect due to Mendelian sampling for progeny } j \text{ of sire } i,$$

$$\text{assumed to follow } N(0,0.5 \sigma_u^2),$$

$$\varepsilon_{ij} = \text{environmental effect in progeny } j \text{ of sire } i, \text{ assumed to follow } N(0,(1-h^2)\sigma_p^2)$$

Progeny were ranked within each half-sib family by phenotype and the top and bottom 10% contributed to DNA pools. For each marker, the true sire allele frequencies in both
pools were obtained by counting and a normally distributed technical error with mean zero and variance 0.0 or 0.0014 was added. Then, to simulate observed marker allele frequencies in the pools, simulated frequencies were divided by the sum of the frequencies of the two sire alleles to satisfy the condition that frequencies of the two alleles sum to 1. The resulting variance due to technical errors in the observed allele frequencies was either about 0.0 or 0.0007. The value 0.0007 was equal to the estimated technical error variance in LIPKIN et al. (1998). Allele frequencies were observed for each half-sib family and for all markers.

Single marker analysis, LS-pool, and ML-pool were applied to the simulated data. Criteria for comparison of methods were: (1) power to detect the QTL, (2) bias and variance of estimates of QTL location, and (3) bias and variance of estimates of QTL effects. For each set of parameters and each mapping method, 10,000 replicates were simulated under the null hypothesis of no QTL to determine 5% chromosome-wise significant thresholds of the test statistics and 3,000 replicates were simulated under the alternative hypothesis.

Application to selective DNA pooling data from dairy cattle: Selective DNA pooling data was collected on multiple sires from multiple years and on markers on BTA6 to detect QTL for milk protein percentage (protein %). The same data was used in TCHOURZYNA et al. (2002).

Table 1 presents the sires with data available by year, pooling procedures, and the number of daughters in the upper and lower pools for each sire. The number of daughters per selected pool ranged from 150 to 211. The study included 13 sires, of which 4 had data in 1996, 5 had data in 1998, and 7 had data in 2002. Three of the five sires with data in 1998 also had data in 2002, however, fewer than 20% daughters of those sires included in the data from 1998 were also included in 2002, so data were assumed independent. Among the 13
sires evaluated, there were two pairs of fathers and sons and another two sires were half-sibs. Previous QTL mapping studies by Lipkin et al. (1998) and Mosig et al. (2002) used the selective DNA pooling data from all sires from 1996 and 3 sires from 1998.

The DNA pooling procedures differed by year. In 1996, one external and one internal pool was formed in both upper and lower tails for each sire (Lipkin et al. 1998). The external pool included daughters with 10% highest or lowest performance and the internal pool included daughters in the next 10% based on highest or lowest performance. In 1998, only one pool was formed per tail per sire and in 2002, two independent pools were formed per tail per sire. In all years, each pool had two independent replicates of PCR, which usually were prepared at different times and by different technicians (Lipkin et al. 1998). Details of milk sampling and pooling procedures were described in Lipkin et al. (1998). For the present analysis, marker allele frequencies were averaged over pools and PCR replicates.

Pool frequency data for fifteen microsatellite markers spanning BTA6 were used. The average marker distance was 8 cM. Data were not available on all markers for all sires because of missing data, no haplotype information, or the sire was homozygous. Table 2 shows marker positions and the number of usable sires for each marker. Sires with data in two years were considered as two different sires. The number of usable sires for a given marker ranged from two to eight. Estimates of marker allele frequencies were corrected for shadow bands as described in Lipkin et al. (1998) and marker haplotypes were estimated as described in Tchourzyna et al. (1999).

In contrast to the simulation study, the difference in observed marker allele frequencies between the upper and lower pools (D value) was used for analysis. The data used are given
in Table 4. Signs of D values across markers for the same sire were consistent with the sire’s marker haplotypes (TCHOURZYNA et al. 1999).

Single marker analysis, LS-pool, and ML-pool were applied. For simplicity of analysis, in all three methods, relationships between sires were ignored and data of the same sire from different years were treated as independent.

**Single marker analysis:** Single marker analysis was conducted as described by LIIPKIN et al. (1998). Under the null hypothesis of no QTL, the expected D value is zero. To derive the standard error of D values (SE$_D$) under the null hypothesis of no QTL, several complications necessitated modification of the methods of analysis. First, in order to average out technical errors, D values used for analysis were averages of D values of the two alternative sire alleles. If only the sire’s alleles contribute to allele frequencies in the pool and there is no technical error, D values for the two sire alleles ($D_L$ and $D_S$) will be in complete inverse correlation. However, dams also contributed alleles to the pool and technical errors are inevitable. Therefore, D values of the alternative sire alleles were semi-dependent. Covariances between D values (Cov$_{DL}$D$_S$) were calculated from the data set and as described by LIIPKIN et al. (1998) and were used to estimate SE$_D$ as:

$$SE_D^2 = SE_{DL}^2 + SE_{DS}^2 - 2Cov_{DL}D_S$$

Second, the number of daughters in the pool varied across sires and pooling and replication procedures differed by year (Table 1). The number of daughters in the pool will affect the variance of binomial sampling error, the use of two replicates per pool will half the variance of technical errors, and the use of two independent pools will half the total variance of the D value for a given sire allele (L or S) of each marker and each family. Third, dams may have the same marker alleles as the sire. Dam allele frequencies were estimated following Lipkin
et al. (1998) based on the deviation of the sum of marker allele frequencies over the upper and lower pools from 0.5. The binomial variance due to dam allele was included in the \( \text{SED} \) of sire alleles. Further details are in Lipkin et al. (1998).

In the single marker analysis, the null hypothesis was rejected with type I error \( \alpha \) if:

\[
Z_{D_{ij}} = \frac{D_{Mij}}{SE_{D_{ij}}} > Z_{1-\alpha/2}
\]

where \( D_{Mij} \) is the D value of the \( j^{th} \) marker of the \( i^{th} \) sire and \( SE_{D_{ij}} \) is its associated standard error. Similar to the analysis with observed marker allele frequencies, this \( Z \) test was extended to a Chi-squares test when multiple sire families were analyzed.

**LS-pool:** The model used to analyze D values at a given QTL position for the \( k \) markers of the \( i^{th} \) sire was:

\[
\begin{bmatrix}
D_{M_{i1}} \\
D_{M_{i2}} \\
\vdots \\
D_{M_{ik}}
\end{bmatrix}
= \begin{bmatrix}
1 - 2r_1 \\
1 - 2r_2 \\
\vdots \\
1 - 2r_k
\end{bmatrix}
\begin{bmatrix}
D_{Q_i} \\
e_{D_{i1}} \\
e_{D_{i2}} \\
\vdots \\
e_{D_{ik}}
\end{bmatrix}
\]

(Model 3)

where \( D_{Q_i} \) is the expected D value for the QTL allele of the \( i^{th} \) sire, \( r_j \) is the recombination rate between marker \( j \) and the QTL, and \( e_{D_{ij}} \) are residuals.

Since the variance of residuals was not homogenous due to different sample sizes and different pooling procedures, a generalized least squares (SEARLE 1971) method was applied. The test statistic for the \( i^{th} \) family at a given putative QTL position was derived as:

\[
\chi^2_{D_{ij}} = Y_i' V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1} Y_i
\]
where $\mathbf{V}_i$ is a diagonal matrix with variances of D values, as derived in the single marker analysis, on the diagonal.

Similar to the analysis of observed marker allele frequencies, Chi-square statistics were summed over m sires at each position and estimates of QTL effects were from the position with the highest test statistic.

**ML-pool:** The observed D value at marker $j$ for family $i$ can be modeled as:

$$D_{M_{ij}} = D_{M_{ij}}^{\text{true}} + \epsilon_{D_{ij}}$$

where $D_{M_{ij}}^{\text{true}}$ is the true difference of marker allele frequencies between tails and $\epsilon_{D_{ij}}$ is a residual. The vector of means ($\mu_{Di}$) of D values of k markers in the $i^{th}$ family can be modeled as:

$$\mu_{Di} = E \left[ \begin{array}{c} D_{M_{i1}} \\ D_{M_{i2}} \\ \vdots \\ D_{M_{ik}} \end{array} \right] = E \left[ \begin{array}{c} D_{M_{i1}}^{\text{true}} \\ D_{M_{i2}}^{\text{true}} \\ \vdots \\ D_{M_{ik}}^{\text{true}} \end{array} \right] = \left[ \begin{array}{c} (1 - 2r_i)(2p_{Q_i}^U - 1) \\ (1 - r_i)(2p_{Q_i}^U - 1) \\ \vdots \\ (1 - 2r_k)(2p_{Q_i}^U - 1) \end{array} \right]$$

with notations as defined before. Variances of D values were the same as derived in the single marker analysis and covariances were twice those derived for observed marker allele frequencies in one pool.

Similar to analysis with marker allele frequencies, likelihood functions were constructed for each sire family under the null and alternative hypotheses, and the likelihood ratio test for m families was evaluated at each possible position of QTL location and maximized for QTL allele frequencies.
RESULTS

**Simulation:** Figure 3 presents QTL mapping profiles for single marker analysis, LS-pool and ML-pool from a representative replicate of simulation. All three methods provided the highest test statistic in the QTL region. However, single marker analysis was only able to indicate the marker closest to the QTL, while both LS-pool and ML-pool are interval mapping methods and provided estimates of QTL location.

Table 3 presents power to detect the QTL and biases and accuracy of estimates of QTL position and effect. With 10 half-sib families and 1,000 progeny per family, all three methods provided high power (≥ 93%) to detect a QTL with moderate effect. The interval mapping methods provided 2-3% greater power than the single marker analysis and the power of ML-pool was 1% higher than that of LS-pool. Both LS-pool and ML-pool provided nearly unbiased estimates of QTL location (bias < 2 cM). The bias of ML-pool was 0.7 cM smaller than that of LS-pool. Standard errors of estimates QTL location were 8 and 9.5 cM for LS-pool and ML-pool. Both LS-pool and ML-pool provided nearly unbiased estimates of QTL effect and had similar standard errors of effect estimates.

**Real data analysis:** Figure 4 shows results from single marker analysis of the selective DNA pooling data on protein % for BTA6. Across family results from the single marker analysis (Figure 4D) suggested two regions with QTL for protein percentage (44-60 cM and 64-76 cM) on BTA6, separated by a marker at position 60 cM (BM4322). These results are consistent with analysis of the same data by Tchourzyna et al. (2002). Results for individual sires indicated that sires 2 and 4 segregated for a QTL in the region around 50 cM (Figure 4A), sires 5 (data from 2002), 1, and 9 (data from 2002) had a moderate QTL in the region
around 70 cM (Figure 4B), and other sires did not show strong evidence of QTL. The \( \log_{10}(1/p) \) values were very different for closely linked markers for some sires. For example, for sire 4, the \( \log_{10}(1/p) \) value was 3.3 at marker BMS2508 (at 44 cM) and 10.4 at marker BM143, which was only 5 cM from BMS2508. These discrepancies caused curves for individual sires to jump up and down across the chromosome and it was difficult to discern a clear trend from results for individual sires.

Figure 5 presents QTL mapping profiles of LS-pool and ML-pool for the joint analysis of all sires. Profiles from both interval mapping methods showed one wide region (from 8 to 89 cM) with high test statistic values. The highest position in both methods was at 60 cM, which was the lowest position in the single marker analysis (Figure 4). The reason for this discrepancy will be discussed later.

Figure 6 presents likelihood ratio profiles of the ML-pool method for individual sires. Similar to results from the single marker analysis (Figure 4), sires 2 and 4 had high values of test statistics around 50 cM. Sires 9 (data from 1998), 9 (data from 2002), 7 (data from 1998), and 5 (data from 2002) showed moderate evidence of a QTL around 75 cM. Trends in Figure 6 were clear and no large changes in the test statistic occurred at nearby positions, as observed for the single marker analysis (Figure 4). The QTL mapping profiles for LS-pool for individual sires were similar to profiles from ML-pool.

**DISCUSSION**

**Evaluation of LS-pool and ML-pool by simulation:** Both LS-pool and ML-pool are interval mapping methods and, consequently, provide separate estimates of QTL position and
effect, which is not possible for single marker analysis. Results from simulation indicated that both interval mapping methods can provide nearly unbiased estimates of QTL location and effect (Table 3). Standard errors of estimates were reasonable, at least for the example, which had high power. Both methods used algorithms searching the optimum solution within the parameter space, therefore, guarantee estimates within the parameter space, which is not possible with the method of DEKKERS (2000).

Both LS-pool and ML-pool utilize the joint information from multiple linked markers and multiple families. Random technical errors in the observed marker allele frequencies are averaged out by considering multiple markers simultaneously, which is not possible for single marker analysis and the method of DEKKERS (2000). This resulted in slightly higher power for ML-pool and LS-pool than the single marker analysis.

Theoretically, ML-pool is a more appropriate model than LS-pool because it accounts for correlations between sampling errors in allele frequencies of linked markers. This is expected to increase power to detect the QTL and result in better estimates of QTL location and effect. In the simulation analysis, all methods had high power and differences were therefore small. However, ML-pool did provide slightly less biased estimates of QTL location than LS-pool. However, mapping accuracy, as quantified by the standard error of estimates of QTL location, was slightly larger for ML-pool than for LS-pool (Table 3). Estimates of QTL effects were similar for both methods. Therefore, ML-pool might be the preferred method for interval mapping of QTL using selective DNA pooling data over LS-pool. Further comparison of LS-pool and ML-pool for different scenarios and an investigation of the reason why ML-pool provided lower mapping accuracy than LS-pool method will be presented in WANG (2003).
Parameters for the simulation study were chosen to mimic those of the QTL mapping study conducted by LIPKIN et al. (1998) and MOSIG et al. (2000) in dairy cattle. Therefore, simulation results obtained here are relevant to dairy cattle QTL mapping projects. However, family sizes can vary greatly in practice and the magnitude of technical errors can be reduced by replication or advanced technical skills. Thus, QTL mapping results of LS-pool and ML-pool may differ from study to study. In the present analysis, the variance of technical errors was assumed known. This, however, requires previous lab work and extra costs, so it is necessary to develop methods to map QTL without prior studies and evaluate their results. The effect of family size, technical error variance, and prior knowledge about technical error variance on QTL mapping results of LS-pool and ML-pool will be investigated in WANG (2003).

**Evaluation of LS-pool and ML-pool based on real data:** The analysis of the dairy cattle data clearly demonstrated that LS-pool and ML-pool can be applied in practical situations and result in substantial benefits over single marker analysis. The single marker analysis showed two regions with QTL for protein percentage (44-60 cM and 64-76 cM) on BTA6 (Figure 4D), consistent with results from a similar analysis by Tchourzyna et al. (2000). However, results from both LS-pool and ML-pool indicated one wide region for QTL (Figure 5) and the position with greatest evidence of the QTL was at 60 cM, which was the position with lowest significance in the single marker analysis. These differences are caused by missing data, as will be explained in the following.

In this data set, genotypes of some markers of some sires were missing or not informative (homozygous or no haplotype information). Therefore, in single marker analysis, the number of sires included in the analysis differed from marker to marker. For example, for marker
BM4322 (at 60 cM), only three sires - 3, 5 (data from 1998), and 7 (data from 1998) - had observations and all three had low absolute D values. Consequently, the statistic over sires at this marker was low. However, D values from linked markers can provide information about D values at the marker that is missing. Therefore, one can predict D values for missing markers from nearby markers. For example, sire 2 had a missing D value at BM4322 (at 60 cM) but its D values at flanking markers (DIK82 at 55 cM and BMS470 at 64 cM) were 0.233 and 0.195. Therefore, the D value of BM4322 was expected to be around 0.2 because of the tight linkage between markers. However, this information from linked markers was ignored in the single marker analysis since each marker was treated independently. Consequently, a missing genotype of a sire with high D value resulted in a low significance level for the marker, even though the smaller number of sires was taken into account by the lower degrees of freedom. In contrast, information from linked markers is taken into account in both interval mapping methods by modeling the expected allele frequencies at different positions. Figure 6 presents the QTL mapping profiles of ML-pool for each individual sire (profiles of LS-pool were similar) and shows that the likelihood ratio at the marker with missing information was similar that at nearby markers with data. If missing markers were at ends of the chromosome, the likelihood ratio curve was flat, with a value equal to the likelihood ratio at the nearest informative marker. Therefore, for the interval mapping methods, the sum of likelihood ratios across sires at position 60 cM was high, even with fewer usable sires. This result demonstrates that the identification of two QTL in the single marker method may be spurious and caused by missing data, and that both interval mapping methods are able to properly account for missing data.
In addition to dealing with missing data, interval mapping methods can also account for discrepancies in D values between nearby markers. For example, the D value at marker CSN3 of sire 5 (data from 1998) was 0.16 but the D value at BM4311, just 6 cM from CSN3, was substantially lower at 0.06 (Table 4). Interval mapping methods use the joint information of both markers to evaluate the likelihood of the observed data for sire 5 (data from 1998) and the likelihood of observing such data was not high (Figure 6). Such discrepancies cannot be accommodated by single marker analysis, which only utilizes the marginal information at each marker and, therefore, resulted in high significance at CSN3 for sire 5 (data from 1998) in spite of lower D values at markers in the neighborhood. In addition to the magnitude of D values and the consistency of D values at linked markers, results from interval mapping methods also depended on the number of markers with consistent D values. In general, the more consistent the observed data are across linked markers, the higher the value of the test statistics will be with the interval mapping approaches.

Based on the results presented in Figure 6, there might be two QTL for protein % on BTA6. There was one large QTL around 50 cM, but only sires 2 and 4 segregated for this QTL. Another possible QTL was around 75 cM. Its effect was small but many sires - sire 9 (data from 1998), sire 9 (data from 2002), sire 7 (data from 1998) and sire 5 (data from 2002) - segregated for this QTL. The QTL profiles of both interval mapping methods showed high evidence of harboring QTL in a wide region when multiple sires were considered simultaneously.

The large difference in D values for BMS2508 and BM143 for sire 4 was unexpected (Table 4). After excluding possibilities of errors in the data, one possible explanation is that sire 4 is segregating for two QTL in opposite phase in that region. However, our current
interval mapping methods can only handle a single QTL or multiple QTL that are widely separated. Thus, profiles for sire 4 might not be reliable.

**Extension to other family structures:** In this research, LS-pool and ML-pool methods were developed for and applied to a half-sib design. However, the same general procedures can also be applied to backcross or F_2 designs. Similar to single marker analysis (DARVASI and SOLLER 1994), the framework of analysis was based on defining two genotype groups. In a half-sib design, the two genotype groups are defined by receiving alternate QTL alleles from the sire. In a backcross design, the two genotype groups are defined as individuals with QQ and Qq (or qq and Qq) genotypes. In an F_2 design, with the assumption of a co-dominant QTL, the two groups can be defined as individuals with QQ and qq genotypes. Therefore, after these two genotype groups are defined, LS-pool and ML-pool methods can be applied in a similar way as described for the half-sib design.

**Limitations of the current interval mapping methods:** Both LS-pool and ML-pool fit only one QTL but multiple QTL may be present on the same chromosome, as is likely the case for the dairy cattle data analyzed. Methods must be extended to accommodate multiple QTL. Kearsey and Hyne (1994) proposed a marker regression approach to map QTL based on a similar model as LS-pool, except that the dependent variable is a vector of differences between phenotypic means of two marker genotypes (for example, MM and mm) at multiple markers instead of a vector of marker allele frequencies. Since this method easily extends to multiple QTL (KEARSEY and HYNE 1994), the methods of selective DNA pooling analysis presented in this study is expected be extended in a similar way. Research to extend the current methods to multiple QTL is under development.
Both LS-pool and ML-pool require knowledge of marker haplotypes of parents, which is usually not known in practice. Haplotypes can be identified based on progeny, genotyped individually, or based on cosegregant pools (TCHOURZYNA et al. 1999), but requires extra costs. Recent research has, however, shown that haplotypes can be estimated directly from the selective DNA pooling data (WANG et al. 2003).

Another limitation of the interval mapping methods is that there is no easy way to obtain chromosome-wise significant thresholds that account for multiple correlated tests conducted on the chromosome. The only possible way is simulation, in which the phenotypic value and marker information of the progeny are simulated to mimic the real data. However, this depends on assumptions about the model and the phenotypic distribution. When the model or assumptions are invalid, thresholds derived from simulation could result in different rates of false positives or negatives.

The current ML-pool method was based on multivariate normal distributions, which only approximate the true distribution of observed frequencies. Deriving the exact distribution and conducting QTL mapping based on the exact distribution requires further research.

Both LS-pool and ML-pool also assume that the multiple sire families are independent, which may not be true in the practice. Methods must be extended to handle observations from related families.

Both LS-pool and ML-pool simultaneously utilize information from multiple markers, which results in greater power to detect the QTL than single marker analysis by averaging out random technical errors. However, when technical errors are consistent over multiple markers, using joint information from multiple markers may result in biased estimates and increased standard errors.
CONCLUSIONS

Two interval QTL mapping methods were developed for analysis of selective DNA pooling data, LS-pool and ML-pool. The methods simultaneously utilize information from multiple markers and multiple families and provide separate estimates of QTL location and effect. Simulation showed that both methods resulted in nearly unbiased estimates and slightly higher power than the currently used single marker analysis. Both methods are also applicable in practical situations and result in substantial benefits over single marker analysis, especially for data with missing genotypes.
LITERATURE CITED


LIPKIN, E., J. FULTON, H. CHENG, N. YONASH and M. SOLLER, 2002 Quantitative trait locus mapping in chickens by selective DNA pooling with dinucleotide microsatellite
markers by using purified DNA and fresh or frozen red blood cells as applied to marker-assisted selection. Poult Sci 81: 283-292.


APPENDIX

Covariance matrix for observed allele frequencies of multiple linked markers:

Derivation of the covariance between allele frequencies of marker \( j \) and \( l \) in the \( i \)th family, with recombination rate \( r_{ij} \), which is assumed known.

\[
\text{cov}(f_{M_{ij}}^U, f_{M_{il}}^U) = \text{cov}(f_{M_{ij}}^L, f_{M_{il}}^L) = \text{cov}(f_{M_{ij}}^{U\text{true}}, f_{M_{ij}}^{U\text{true}} + e_{ij}) = \text{cov}(f_{M_{il}}^{U\text{true}}, f_{M_{il}}^{U\text{true}})
\]

\[
\text{cov}(f_{M_{ij}}^{U\text{true}}, f_{M_{il}}^{U\text{true}}) = \text{cov}(\frac{\#M^U_{ij}}{n}, \frac{\#M^U_{il}}{n}) = \frac{1}{n^2} \text{cov}(\#M^U_{ij}, \#M^U_{il})
\]

where \( \#M^U_{ij} \) and \( \#M^U_{il} \) are number of “M” alleles from sire \( i \) for marker \( j \) and \( l \) in the upper pool, and

\[
\text{cov}(\#M^U_{ij}, \#M^U_{il}) = E(\#M^U_{ij} \cdot \#M^U_{il}) - E(\#M^U_{ij})E(\#M^U_{il})
\]

Since covariance matrix is same for marker frequencies in the upper and lower pool, the following derivation will not specify upper or lower pools:

\[
E(\#M^U_{ij}) = np_{M_{ij}} = n[(1 - r_j)p_{Q_j} + r_j(1 - p_{Q_j})]
\]

\[
E(\#M^U_{il}) = np_{M_{il}} = n[(1 - r_l)p_{Q_l} + r_l(1 - p_{Q_l})]
\]

\[
E(\#M^U_{ij} \cdot \#M^U_{il}) = E(\sum_{k_1=1}^{n} I_{M_{ijk_1}} \cdot \sum_{k_2=1}^{n} I_{M_{ilk_2}}) = E(\sum_{k_1=1}^{n} \sum_{k_2=1}^{n} I_{M_{ijk_1}} \cdot I_{M_{ilk_2}})
\]

where \( I_{M_{ijk_1}} \) and \( I_{M_{ilk_2}} \) are indicators for markers \( j \) and \( l \) of progeny \( k_1 \) and \( k_2 \). \( I = 1 \) if the progeny gets the “M” allele from sire \( i \), \( I = 0 \) if the progeny gets the “m” allele.

\[
E(\#M^U_{ij} \cdot \#M^U_{il}) = \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} E(I_{M_{ijk_1}} \cdot I_{M_{ilk_2}}) = \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \text{Pr}(I_{M_{ijk_1}} = 1 \text{ and } I_{M_{ilk_2}} = 1)
\]
1. Marker QTL order (Mj-Q-Mi):

If \( k_1 = k_2 \) (two individuals are the same):

\[
\Pr(I_{\text{Mijk}_1} = 1 \text{ and } I_{\text{Mijk}_2} = 1) = (1 - r_j)(1 - r_i) p_{q_i} + r_j r_i (1 - p_{q_i})
\]

If \( k_1 \neq k_2 \) (individuals \( k_1 \) and \( k_2 \) are different):

\[
\Pr(I_{\text{Mijk}_1} = 1 \text{ and } I_{\text{Mijk}_2} = 1) = \Pr(I_{\text{Mijk}_1} = 1) \cdot \Pr(I_{\text{Mijk}_2} = 1)
\]

\[
= [(1 - r_j) p_{q_i} + r_j (1 - p_{q_i})][(1 - r_i) p_{q_i} + r_i (1 - p_{q_i})]
\]

\[
E(#M_{ij}, #M_{ij}) = n \cdot [(1 - r_j)(1 - r_i) p_{q_i} + r_j r_i (1 - p_{q_i})]
\]

\[
+ n(n-1) \cdot [(1 - r_j) p_{q_i} + r_j (1 - p_{q_i})][(1 - r_i) p_{q_i} + r_i (1 - p_{q_i})]
\]

\[
cov(#M_{ij}, #M_{ij}) = n \cdot [(1 - r_j)(1 - r_i) p_{q_i} + r_j r_i (1 - p_{q_i})]
\]

\[
+ n(n-1) \cdot [(1 - r_j) p_{q_i} + r_j (1 - p_{q_i})][(1 - r_i) p_{q_i} + r_i (1 - p_{q_i})]
\]

\[
- n^2 [(1 - r_j) p_{q_i} + r_j (1 - p_{q_i})][(1 - r_i) p_{q_i} + r_i (1 - p_{q_i})]
\]

\[
= n(1 - 2r_i) p_{q_i} (1 - p_{q_i})
\]

where \( r_i = (1 - r_j) \cdot r_i + r_j \cdot (1 - r_i) \), implied by Haldane mapping function.

Thus, \( \text{cov}(f_{Mij}, f_{Mij}) = \text{cov}(f_{Mij}, f_{Mij}) = \frac{1}{n} (1 - 2r_i) p_{q_i} (1 - p_{q_i}) \)

2. Marker QTL order (Mj-Q-Mi-O):

If \( k_1 = k_2 \) (two individuals are the same):

\[
\Pr(I_{\text{Mijk}_1} = 1 \text{ and } I_{\text{Mijk}_2} = 1) = (1 - r_i)(1 - r_j) p_{q_i} + (1 - r_i) r_j (1 - p_{q_i})
\]
If $k_1 \neq k_2$ (individual $k_1$ and $k_2$ are different):

$$\Pr(I_{M_{ij_1}} = 1 \text{ and } I_{M_{ij_2}} = 1) = \Pr(I_{M_{ij_1}} = 1) \cdot \Pr(I_{M_{ij_2}} = 1)$$

$$\Pr(I_{M_{ij_1}} = 1) = [(1 - r_i) p_{Q_i} + r_i (1 - p_{Q_i})] = p_{M_{il}}$$

$$\Pr(I_{M_{ij_2}} = 1) = [(1 - r_j) p_{M_{il}} + r_j (1 - p_{M_{il}})]$$

$$E(#M_{ij} \cdot #M_{ij}) = n \cdot [(1 - r_i) (1 - r_j) p_{Q_i} + r_i r_j (1 - p_{Q_i})]$$

$$+ n(n - 1) \cdot p_{M_{il}} [(1 - r_i) p_{M_{il}} + r_i (1 - p_{M_{il}})]$$

$$E(#M_{ij}) = n p_{M_{il}} = n [(1 - r_i) p_{Q_i} + r_i (1 - p_{Q_i})]$$

$$E(#M_{ij}) = n p_{M_{ij}} = n[(1 - r_i) p_{M_{il}} + r_i (1 - p_{M_{il}})]$$

$$\text{cov}(#M_{ij}, #M_{ij}) = n \cdot (1 - r_i) (1 - r_j) p_{Q_i} + r_i r_j (1 - p_{Q_i})$$

$$+ n(n - 1) \cdot p_{M_{il}} [(1 - r_i) p_{M_{il}} + r_i (1 - p_{M_{il}})]$$

$$- n^2 [(1 - r_i) p_{Q_i} + r_i (1 - p_{Q_i})][(1 - r_i) p_{M_{il}} + r_i (1 - p_{M_{il}})]$$

Thus, $\text{cov}(f_{M_{ij}}^L, f_{M_{ij}}^L) = \text{cov}(f_{M_{ij}}^L, f_{M_{ij}}^L) = \frac{1}{n} (1 - 2 r_i) p_{M_{il}} (1 - p_{M_{il}})$

**Haldane mapping function:** Mapping functions provided a transformation from map distance to recombination rate. Haldane mapping function is derived based on assumptions that the probability of crossover between two loci follows Poisson distribution and that recombination events in two adjacent intervals are independent.

$$x = -\frac{1}{2} \ln(1 - 2 r_i)$$

where $x$ is the map distance and $r_i$ is the recombination rate between two loci.
### TABLE 1

Pooling procedures and numbers of daughters in the upper and lower pools from sires used in different years for the selective DNA pooling data for protein % on BTA6

<table>
<thead>
<tr>
<th>Year</th>
<th>1996</th>
<th>1998</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sire</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1 external pool and 1 internal pool per tail, 2 replicates per pool</td>
<td>1 pool from each tail, 2 replicates per pool&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2 independent pools per tail, 2 replicates per pool&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Number of daughters/pool&lt;sup&gt;b&lt;/sup&gt;</td>
<td>161</td>
<td>153</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>170</td>
<td>166</td>
<td>176</td>
</tr>
</tbody>
</table>

<sup>a</sup> In 1996, the external pools included the top and bottom 10% daughters based on performance and the internal pools included the next 10% of the top and bottom daughters. Pooling procedures were as described by Lipkin et al. (1998). For all years, each pool was prepared in two independent replicates, and each replicate was conducted at a different time and usually by a different person.

<sup>b</sup> Average D values from external and interval pools were used for analysis in 1996. Average D values from the two independent pools were used for analysis in 2002.
TABLE 2

Position of markers and number of sires with usable data for the analysis of selective DNA pooling for protein % on BTA6

<table>
<thead>
<tr>
<th>Markers</th>
<th>Position</th>
<th>Number of usable sires*</th>
</tr>
</thead>
<tbody>
<tr>
<td>INRA133</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>BM1329</td>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>BMS2508</td>
<td>44</td>
<td>8</td>
</tr>
<tr>
<td>BM143</td>
<td>49</td>
<td>8</td>
</tr>
<tr>
<td>DIK82</td>
<td>55</td>
<td>5</td>
</tr>
<tr>
<td>BM4322</td>
<td>60</td>
<td>3</td>
</tr>
<tr>
<td>BMS470</td>
<td>64</td>
<td>7</td>
</tr>
<tr>
<td>ILSTS97</td>
<td>67</td>
<td>4</td>
</tr>
<tr>
<td>RM28</td>
<td>74</td>
<td>7</td>
</tr>
<tr>
<td>BM415</td>
<td>76</td>
<td>6</td>
</tr>
<tr>
<td>CSN3</td>
<td>83</td>
<td>6</td>
</tr>
<tr>
<td>BM4311</td>
<td>89</td>
<td>5</td>
</tr>
<tr>
<td>BM8124</td>
<td>94</td>
<td>3</td>
</tr>
<tr>
<td>BMC4203</td>
<td>113</td>
<td>6</td>
</tr>
<tr>
<td>BM2320</td>
<td>121</td>
<td>7</td>
</tr>
</tbody>
</table>

* The data set included 13 sires, of which 3 had data in both 1998 and 2002. Since fewer than 20% daughters of those sires included in both years, sires with data in two years were treated as two separate sires.
TABLE 3

Power to detect the QTL, bias (SE) of estimates of QTL position and effect from analysis of simulated selective DNA pooling data using LS-pool and ML-pool methods and power from single marker analysis. Ten half-sib families with 1000 progeny were used and a QTL with an effect of 0.25 phenotypic standard deviations was located at 36 cM on a 100 cM chromosome with 6 equidistant markers. The variance of technical errors from pooling was 0.0007 and assumed known. Results were based on 3,000 replicates and 5% chromosome-wise thresholds were determined from 10,000 replicates under the null hypothesis.

<table>
<thead>
<tr>
<th></th>
<th>Single marker</th>
<th>LS-pool</th>
<th>ML-pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>93%</td>
<td>95%</td>
<td>96%</td>
</tr>
<tr>
<td>QTL position - bias (SE) (cM)</td>
<td>-</td>
<td>+1.74 (7.99)</td>
<td>+0.98 (9.50)</td>
</tr>
<tr>
<td>QTL effect - bias (SE) (σp)</td>
<td>-</td>
<td>0.250 (0.095)</td>
<td>0.254 (0.094)</td>
</tr>
</tbody>
</table>

* bias = estimated QTL position - true QTL position. A positive bias means a bias toward the center of the chromosome.
TABLE 4

D value for each marker and each sire for the selective DNA pooling data for protein % on BTA6

<table>
<thead>
<tr>
<th>Sire</th>
<th>1996</th>
<th>1998</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>INRA133</td>
<td>-0.10</td>
<td>0.06</td>
<td>-0.01</td>
</tr>
<tr>
<td>BM1329</td>
<td>-0.09</td>
<td>-0.12</td>
<td></td>
</tr>
<tr>
<td>BMS2508</td>
<td>0.25</td>
<td>0.04</td>
<td>0.11</td>
</tr>
<tr>
<td>BM143</td>
<td>0.00</td>
<td>0.20</td>
<td>-0.04</td>
</tr>
<tr>
<td>DIK82</td>
<td>0.23</td>
<td>0.03</td>
<td>0.09</td>
</tr>
<tr>
<td>BM4322</td>
<td>-0.01</td>
<td>-0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>BMS470</td>
<td>0.04</td>
<td>0.20</td>
<td>0.01</td>
</tr>
<tr>
<td>ILSTS97</td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RM28</td>
<td>-0.12</td>
<td>0.25</td>
<td>0.18</td>
</tr>
<tr>
<td>BM415</td>
<td>-0.10</td>
<td>0.19</td>
<td>0.08</td>
</tr>
<tr>
<td>CSN3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BM4311</td>
<td>0.11</td>
<td>0.10</td>
<td>-0.04</td>
</tr>
<tr>
<td>BM8124</td>
<td>0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMC4203</td>
<td></td>
<td>-0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>BM2320</td>
<td>0.06</td>
<td>-0.03</td>
<td>0.02</td>
</tr>
</tbody>
</table>

* Empty cell indicates either that the sire is homozygous or haplotype is not known for that marker.
* Signs of D values across markers for the same sire were consistent with the sire’s marker haplotypes (TCHOURZYNA et al. 1999).
* Three sires had data in both 1998 and 2002 but fewer than 20% daughters of those sires were common to both years. As a result, data of the same sire from different years was treated as independent data.
FIGURE 1. — Principles of selective DNA pooling in a sire family, showing the phenotypic distribution, observed marker allele frequencies \((f_m^U, f_m^L, f_M^U, f_M^L)\), and expected QTL allele frequencies \((p_Q^U, p_Q^L, p_q^U, p_q^L)\) in selected pools of progeny from a sire that is heterozygous for a QTL and a linked marker.
FIGURE 2. Recombination rates between QTL and markers on a chromosome.
FIGURE 3. — QTL mapping plots from a representative replicate of simulation for single marker analysis (A), LS-pool (B) and ML-pool (C). The position with the highest value of test statistic is the estimated QTL location. The true QTL, with an effect of 0.25 phenotypic standard deviations, was located at 36 cM on a 100 cM chromosome with six equidistant markers (Mj). Other simulation parameters were as in Table 3. Broken lines are 5% chromosome-wise thresholds.
FIGURE 4. — Results from single marker analysis of the selective DNA pooling data for protein % on BTA6. The y-axis presents the $\log_{10}(1/p)$ value, where p is the p-value from a comparison-wise significance test. Signs of $\log_{10}(1/p)$ across markers for the same sire were consistent with the sire’s marker haplotypes. Figures A, B and C are results for individual sires and figure D presents results across sires.
FIGURE 5. — Results from LS-pool (thin, dashed line) and ML-pool (thick, solid line) of the selective DNA pooling data across sires for protein % on BTA6. The left y-axis presents the LR value from the ML-pool method and the right y-axis presents the Chi-square value from the LS-pool method.
FIGURE 6. — Results from ML-pool of the selective DNA pooling data for each sire for protein % on BTA6.
CHAPTER 4. INTERVAL MAPPING OF QTL WITH SELECTIVE DNA POOLING DATA
II. SIMULATION STUDIES

In preparation for submission to: Genetics

ABSTRACT

In this study, two DNA pooling interval mapping methods, one based on least squares regression (LS-pool) and the other on approximate maximum likelihood (ML-pool), were evaluated at several combinations of parameters with simulated data of a half-sib design. Power for QTL detection, means and standard errors of estimates of QTL location and QTL effects from these two methods were compared with results from single marker analysis of pooling data and with selective genotyping analysis of individual genotypes. Effects of different parameters on QTL mapping results were also investigated. Both LS-pool and ML-pool provided greater power to detect the QTL than the single marker analysis and separate estimates of QTL location and effect, which are not available from the single marker analysis. With large family size (2000 progeny), both LS-pool and ML-pool provided similar power and estimates of QTL location and effect as selective genotyping. The LS-pool method resulted in severe bias in estimates of QTL location with small family size (500 progeny) and distal QTL. Biases were in part caused by a heterogenous distribution of the test statistic across the chromosome. Several methods were used to attempt to reduce the bias for LS-pool but none provided ideal results.
INTRODUCTION

Using marker allele frequencies in pooled DNA from individuals with extreme phenotypes, instead of individually genotyping, is a very efficient approach to detect linkage with QTL because the number of genotypings is small and independent of sample size. This method has successfully detected genes responsible for quantitative traits (QTL) in several studies (PLOTSKY et al. 1993; LIPKIN et al. 1998; MOSIG et al. 2001; LIPKIN et al. 2002; ZHOU et al. 2003).

DARVASI and SOLLER (1994) called the use of pools of extreme individuals for QTL detection "selective DNA pooling" and presented a statistical method to identify marker-QTL linkage. A Z significance test was used when a single family was analyzed and a Chi-square test was proposed for analysis of multiple families. BARO et al. (2001) investigated the effect of different parameters, such as size of the QTL effect, family size, and level of technical error, on power to detect the QTL using the DARVASI and SOLLER (1994) method. Results indicated that size of the QTL effect, dominance at the QTL and family size had large effects on power.

The method of DARVASI and SOLLER (1994) is based on single marker information and does not provide separate estimates for QTL position and effect, nor does it utilize the joint information from multiple markers along the chromosome. In addition, when missing marker genotypes are present in some families, which frequently happens in practice, the single marker analysis is problematic, as was demonstrated in WANG (2003).

WANG (2003) introduced two interval mapping methods to overcome the limitations of the single marker analysis. One method is based on least squares regression (LS-pool) and
the other is based on an approximate maximum likelihood (ML-pool). Results from simulation of a large half-sib family design indicated that both interval mapping methods provided separate and nearly unbiased estimates for QTL position and QTL effect, and had slightly higher power than single marker analysis. Results of application to dairy cattle data indicated that both interval mapping methods can overcome the problem of missing genotypes in single marker analysis by utilizing information from nearby markers.

In this study, simulation was used to further compare both interval mapping methods with single marker analysis and selective genotyping analysis of individual genotypes for different combinations of parameters and to investigate effects of different parameters on QTL mapping results. From simulation results that will be presented, biased estimates of QTL location were observed in some cases, especially with the LS-pool method. Reasons for the bias were investigated and methods to correct the bias were introduced and evaluated by simulation. Finally, the impact of the assumption of a symmetric distribution of phenotypic data in relation to marker and QTL genotypes was investigated.

MATERIALS AND METHODS

QTL mapping methods

WANG (2003) introduced two interval mapping methods for analysis of selective DNA pooling data, LS-pool and ML-pool. Both methods utilize multiple markers and require known parental marker haplotypes. These same methods will be used in the present study and will be briefly described in what follows. Further details can be found in WANG (2003).
**LS-pool:** The observed allele frequencies in the selected pools among half-sib progeny of sire $i$ for $k$ markers on a chromosome are modeled in terms of the expected QTL allele frequency and the recombination rate between the marker and the QTL (the position of markers and the marker haplotypes of the sires are assumed known):

$$
\begin{bmatrix}
  f_{M_{i1}}^U - 1/2 \\
  f_{M_{i2}}^U - 1/2 \\
  \vdots \\
  f_{M_{ik}}^U - 1/2 \\
  f_{M_{i1}}^L - 1/2 \\
  f_{M_{i2}}^L - 1/2 \\
  \vdots \\
  f_{M_{ik}}^L - 1/2 
\end{bmatrix} =
\begin{bmatrix}
  1 - 2r_i \\
  1 - 2r_2 \\
  \vdots \\
  1 - 2r_k \\
  -(1 - 2r_i) \\
  -(1 - 2r_2) \\
  \vdots \\
  -(1 - 2r_k) 
\end{bmatrix} \begin{bmatrix}
  p_{Q_i}^U - 1/2 
\end{bmatrix} +
\begin{bmatrix}
  e_{i1}^U \\
  e_{i2}^U \\
  \vdots \\
  e_{ik}^U \\
  e_{i1}^L \\
  e_{i2}^L \\
  \vdots \\
  e_{ik}^L 
\end{bmatrix}
$$

(Model 1)

where $f_{M_{ij}}^U$ and $f_{M_{ij}}^L$ are observed marker allele frequencies in the upper (U) and lower pools (L) for the $j^{th}$ marker of the $i^{th}$ family, $p_{Q_i}^U$ is the expected QTL allele frequency in the upper pool of the $i^{th}$ family, $r_j$ is the recombination rate between the $j^{th}$ marker and the QTL, and $e$ are residuals. Multiple families are accommodated by extending the dimensions of the vector and matrix. Least squares regression is used at each possible QTL position along the chromosome to fit this model. The position that provides the highest value of the test statistic is taken as the estimate of QTL position and the associated estimate of QTL allele frequencies is used to estimate QTL effects, following to DARVASI and SOLLER (1994). The test statistic that is calculated at each QTL position depends on whether the technical error variance ($V_{TE}$) is known from previous research, or unknown.

**Known technical error variance:** At each QTL position, the following Chi-square type statistic is calculated:
where \( SS_{\text{regression}} \) is the sum squares of regression, 0.25 is the variance of the marker allele frequency from a binomial distribution under the null hypothesis of no QTL and \( n \) is the number of individuals in the selected pool (assumed the same in both pools).

**Unknown technical error variance:** At each QTL position, the following F statistic is calculated:

\[
F = \frac{SS_{\text{regression}} / m}{SS_{\text{error}} / m \cdot (2k - 1)}
\]

where \( SS_{\text{regression}} \) is the sum squares of regression, \( SS_{\text{error}} \) is the sum squares of residuals, and \( k \) is the total number of markers. The degrees of freedom for regression is \( m \) and the degrees of freedom for residuals is \( m^* (2k - 1) \), where \( m \) is the number of families.

**ML-pool:** If technical errors in marker allele frequencies are assumed independent and follow identical normal distributions with mean zero, the distribution of the \( k \) observed marker allele frequencies in the selected pools is a mixture of a binomial frequency and a normal distribution, which can be approximated by a multivariate normal distribution. Means of this normal distribution are the same as in the LS-pool method and the variance-covariance matrix was described in WANG (2003). Based on the likelihood function of the multivariate normal distribution, a likelihood ratio test is constructed, depending on whether \( V_{TE} \) is known or unknown.

**Known technical error variance:** The likelihood under the null hypothesis of no QTL is a constant. The likelihood function is evaluated at all possible values of QTL location and QTL allele frequencies and values that maximize the likelihood ratio provide estimates of QTL
location and frequencies. The latter are then used to estimate a QTL effect for each family, similar to LS-pool, based on Darvasi and Soller (1994).

*Unknown technical error variance:* The maximum value of the likelihood under the null hypothesis depends on QTL location, QTL allele frequencies in the pools and the magnitude of $V_{TE}$. Under the alternate hypothesis, the likelihood function is evaluated at all possible values of QTL location, QTL allele frequency in the upper pool, and $V_{TE}$. Parameter values that maximize the likelihood ratio provide estimates of QTL location, QTL allele frequencies, and $V_{TE}$.

*Single marker analysis:* The usual method to detect linkage between markers and the QTL is based on a significance test for each single marker (Darvasi and Soller 1994). With approximation to the normal distribution, the null hypothesis is rejected with type I error $\alpha$ if

$$Z_{ij} = \frac{2(f_{M_{ij}}^U + f_{M_{ij}}^L) - 0.5}{\sqrt{\text{Var}(\frac{f_{M_{ij}}^U + f_{M_{ij}}^L}{2})}} > Z_{1-\alpha/2}$$

The average of $f_{M_{ij}}^U$ and $f_{M_{ij}}^L$ is expected to be 0.5 under the null hypothesis.

*Known technical error variance:* The variance of the estimated marker frequency under the null hypothesis is:

$$\text{Var}(\frac{f_{M_{ij}}^U + f_{M_{ij}}^L}{2}) = \frac{1}{2} \left( \frac{0.25}{n} + V_{TE} \right)$$

with notations similar to that used for LS-pool.
Unknown technical error variance: The variance of the observed marker frequencies under the null hypothesis can be estimated as in LIPKIN et al. (1998). Assuming symmetry, \( f_{m_{ij}}^U \) and \( f_{m_{ij}}^L \) are expected to be equal and the only reason for a difference between them is binomial sampling error and technical error. Consequently,

\[
\hat{\text{Var}}\left(\frac{f_{m_{ij}}^U + f_{m_{ij}}^L}{2}\right) = \frac{1}{4}\hat{\text{Var}}(f_{m_{ij}}^U - f_{m_{ij}}^L) = \frac{1}{4(mk - 1)} \sum_{j=1}^{m} \sum_{k=1}^{l} (f_{m_{ij}}^U - f_{m_{ij}}^L)^2
\]

If information from \( m \) families is available, for each marker \( j \), the null hypothesis is rejected with type I error \( \alpha \) if

\[
\sum_{i=1}^{m} Z_{ij}^2 > \chi_{\alpha}^2 \quad \text{df} = m
\]

Analysis of selective genotyping data: With selective genotyping, individuals with high and low phenotypes are individually genotyped. In this study, a regular least squares interval mapping method (Haley and Knott, 1992) was applied to the selected individuals. Marker haplotypes of sires were assumed known. Within each half-sib family, the phenotypic value of each progeny was regressed on the probability of getting the 'Q' allele from the sire, which was calculated from the genotype of flanking markers (Weller et al. 1990). Families were assumed independent and the F test statistic over all families was obtained at each possible QTL position. The position providing the highest F value was the estimated QTL position. Estimates of QTL effects at this position were adjusted based on selection intensity following Davarshi and Soller (1994). Chromosome-wise significance thresholds at the 5% level were derived by simulation under the null hypothesis of no QTL.
**Simulation model and parameters**

Ten half-sib families with 2000 or 500 progeny per family were simulated. The simulated population structure was designed to mimic the dairy cattle data used for a selective DNA pooling study by Lipkin et al. (1998) and Mosig et al. (2001). For each individual, six fully informative markers were evenly spaced on a 100 cM chromosome (including markers at the ends). All dam alleles were assumed different from sire alleles and in linkage equilibrium with the QTL. Crossover probabilities were generated according to the Haldane mapping function, which implies independence of recombination events in adjacent intervals on the chromosome. A single bi-allele QTL with frequency 0.5 was either at 11 or at 46 cM from one end of the chromosome. The QTL genotype of each sire, QQ, Qq, qQ, or qq, was randomly assigned, with equal probability for each. The additive effect of the QTL was 0.25 phenotypic standard deviations ($\sigma_p$), and the dominance effect was 0. The phenotype was affected by the QTL along with polygenes and environmental factors, which were normally distributed. The variance explained by the QTL ($\sigma^2_{QTL}$) was 0.03125 $\sigma_p^2$ (Falconer and Mackay 1996) and the variance explained by polygenes ($\sigma^2_{G}$) was $h^2\sigma_p^2 - \sigma^2_{QTL}$. The variance of the possible phenotypic value ($\sigma_p^2$) was set to 1 and the heritability ($h^2$) was set to 0.25. The model for the phenotypic value of a progeny was:

$$ y_{ij} = \mu + QTL_{ij} + Usire_i + Udam_{ij} + UMendel_{ij} + \epsilon_{ij} $$

(Model 2)

where

- $y_{ij}$ = phenotypic value of progeny $j$ of sire $i$,
- $\mu$ = overall mean,
\( QTL_{ij} \) = QTL effect of progeny \( j \) of sire \( i \), fixed effect,

\( U_{sire_i} \) = polygenic effect of sire \( i \) transmitted to an average progeny, assumed to follow

\[ N(0,0.25 \sigma^2_u) \],

\( U_{dam_j} \) = polygenic effect of dam \( j \) mating to sire \( i \) transmitted to its progeny, assumed to follow

\[ N(0,0.25 \sigma^2_u) \],

\( U_{Mendel_i} \) = polygenic effect due to Mendelian sampling for progeny \( j \) of sire \( i \), assumed to follow

\[ N(0,0.5 \sigma^2_u) \],

\( \varepsilon_{ij} \) = environmental effect in progeny \( j \) of sire \( i \), assumed to follow

\[ N(0,(1-h^2)\sigma^2_p) \]

Progeny were ranked by phenotype within each half-sib family and the top and bottom 10\% contributed to DNA pools. For each marker, the true sire allele frequencies in both pools were obtained by counting and a normally distributed technical error with mean zero and variance 0.0 or 0.0014 was added. Then, to simulate observed marker allele frequencies in the pools, simulated frequencies were divided by the sum of the frequencies of the two sire alleles to satisfy the condition that frequencies of the two alleles sum to 1. The resulting variance due to technical errors in the observed allele frequencies was either about 0.0 or 0.0007. The value 0.0007 was equal to the estimated technical error variance in Lipkin et al. (1998). Allele frequencies were observed for each half-sib family and for all markers.

Single marker analysis, LS-pool and ML-pool were applied to the simulated selective DNA pooling data and least squares interval mapping was applied to the simulated selective genotyping data. For each set of parameters and each mapping method, 10,000 replicates were simulated under the null hypothesis of no QTL to determine the 5\% chromosome-wise
significant thresholds of test statistics and 3,000 replicates were simulated under the alternative hypothesis. Criteria for comparison of methods were: (1) power to detect the QTL, (2) bias and variance of estimates of QTL location, and (3) bias and variance of estimates of QTL effects. Results from different methods were compared based on all replicates and based on only significant replicates.

Methods to correct the bias in estimates of QTL location from the LS-pool method

Substantial bias was observed in estimates of QTL location from the LS-pool method for some cases. Details about the bias and reasons for the bias are discussed in the results section. Here we present several methods that were used to attempt to correct the bias.

Use of flanking markers: One reason for the bias in estimates from the LS-pool method was the heterogeneous distribution of the test statistic across the chromosome because multiple markers were used simultaneously. Therefore, one method that was attempted to reduce the bias was to map the QTL in each marker interval based on only markers that flank the interval, similar to regular interval mapping method with fully informative individual genotyping data (Haley and Knott, 1992).

Standardization of the test statistic: Using multiple markers simultaneously causes a greater mean and variance of the test statistic at central positions under the null hypothesis. Therefore, another method attempted to reduce the bias was to standardize the test statistic before mapping the QTL. The mean and variance of test statistic under the null hypothesis were obtained for each position by simulation and then used to standardize the test statistic under the alternative hypothesis. The standardized test statistic was used for QTL mapping.
**Parametric bootstrap:** The concept behind this method to correct for the bias was to establish a "correction" table that provides the average estimated location for each true QTL position. Once this table is determined, one can easily obtain a "true", unbiased estimate from the table by referring to the biased estimates from LS-pool. A parametric bootstrap method (CHERNICK, 1999) was used to obtain this table. Parametric bootstrap simulates data according to an assumed model and distribution. Phenotypic values for each individual were simulated following model 1. The estimate of the QTL effect obtained from the original data by LS-pool was used as the true QTL effect when simulating the data, because effect estimates were found to be nearly unbiased (see results section). For a given QTL position, after phenotypic values were simulated by parametric bootstrap, progeny were selected, marker allele frequencies were created by the same simulation procedure as before, and LS-pool was applied to estimate the QTL position. This procedure was repeated 500 times and a distribution of estimates of QTL position from LS-pool was obtained for a given QTL position. The mean of this distribution was taken as the estimate of location corresponding to the true QTL position in the correction table. The same procedure was used to obtain estimates for all other true QTL positions.

**Validation of the symmetry assumption**

One important assumption in both LS-pool and ML-pool is that the distribution of phenotypic values within the group of progeny receiving the "Q" or "q" allele from the sire is same and symmetric (symmetry assumption). Under this assumption, the favorable QTL allele frequency in the upper pool (\( p^U_{q_1} \)) is expected to be equal to the unfavorable QTL allele
frequency in the lower pool \( (p_{\bar{q}_1}^L) \) and, therefore, there is only one parameter for QTL allele frequency that needs to be estimated. This symmetry assumption will be invalid if the QTL is dominant or the QTL allele frequency among dams is not 0.5. Under these situations, individuals with heterozygous QTL (Qq) will not be equally distributed across the upper and lower pools. Therefore, for such situations, it may be more appropriate to fit two QTL allele frequency parameters in the model, one for each selected pool. Since ML-pool is more computationally complex and the difference between LS-pool and ML-pool is not expected to be large, only LS-pool was investigated here. The symmetry assumption was evaluated and results from least squares models that fitted one (LS-pool-1) or two QTL frequencies (LS-pool-2), one for the upper and one for the lower pool, were compared for different combinations of QTL dominance and QTL allele frequencies among dams.

If two QTL allele frequency parameters are used in LS-pool, the model (LS-pool-2) becomes:

\[
\begin{bmatrix}
    f_{M_{11}}^U & f_{M_{12}}^U & \cdots & f_{M_{ik}}^U \\
    f_{M_{11}}^L & f_{M_{12}}^L & \cdots & f_{M_{ik}}^L \\
    \vdots & \vdots & \ddots & \vdots \\
    f_{M_{11}}^L & f_{M_{12}}^L & \cdots & f_{M_{ik}}^L
\end{bmatrix}
\begin{bmatrix}
    1 - 2r_i \\
    1 - 2r_2 \\
    \vdots \\
    1 - 2r_k
\end{bmatrix}
\begin{bmatrix}
    p_{\bar{q}_i}^U - 1/2 \\
    p_{\bar{q}_i}^L - 1/2
\end{bmatrix}
+ 
\begin{bmatrix}
    e_{11}^U \\
    e_{12}^U \\
    \vdots \\
    e_{ik}^U
\end{bmatrix}
\]

(Model 3)

where \( p_{\bar{q}_1}^U \) and \( p_{\bar{q}_1}^L \) are the expected QTL allele frequencies in the upper and lower pools of the \( i^{th} \) family and other terms are same as the model with one QTL frequency parameter (LS-pool-1).
RESULTS

Comparison of QTL mapping results from alternate methods

The LS-pool, ML-pool and single marker analysis methods utilize information from pooled DNA, while the selective genotyping method utilizes individual genotypes. The three methods using DNA pooling data were applied with or without previous knowledge about technical error variance. The LS-pool, ML-pool and selective genotyping methods are interval mapping methods and provide separate estimates of QTL location and QTL effect. The single marker analysis only provides the marker closest to the QTL and its position was assumed to be the estimate of QTL position for those analyses. Methods were applied to the simulated data and compared based on power and bias and accuracy of estimates.

Power: Table 1 shows power for the alternate methods. Ten half-sib families with large (2000 progeny) or small (500 progeny) family size were used, the true QTL location was central (46 cM from one end of the 100 cM chromosome) or distal (11 cM), and the true $V_{TE}$ was 0 or 0.007. Results from selective genotyping are independent of $V_{TE}$. All four methods resulted in high and similar power ($\geq 97\%$) for large family size and moderate power (51 to 80%) with small family size (Table 1). For the small family size, which is what will be discussed in the remainder, selective genotyping provided 2-27% greater power than the three selective DNA pooling methods, with the exception that LS-pool had slightly higher or similar power when the QTL was at 46 cM and $V_{TE}$ was known. Among the three methods using selective DNA pooling data, for most situations, both LS-pool and ML-pool provided higher power than the single marker analysis and ML-pool had the highest power, which was
only 2-6% less than that of selective genotyping. The only exception was that, when \( V_{TE} \) was known, LS-pool provided the highest power with a central QTL (even higher than selective genotyping) but lowest power with a distal QTL (even lower than single marker analysis). This discrepancy will be investigated further in the discussion section.

Incorporating previous knowledge of \( V_{TE} \) in the analysis resulted in 16-21% greater power for single marker analysis, 8-13% greater power for LS-pool and no increment for ML-pool, when compared with strategies that did not include this information (Table 1, small family size). Power of LS-pool was 10-14% higher for a central QTL than for a distal QTL. This difference was 2-5% for single marker analysis and only 1-2% for ML-pool. Presence of technical errors only slightly decreased power (≤4%) for all methods and in all situations, except that single marker analysis with known \( V_{TE} \) and a distal QTL had 7% greater power when no technical error was present.

**Estimates of QTL location:** Table 2 shows means and standard errors of estimates of QTL location obtained from the different methods. Means and accuracy of estimates of QTL location with known or unknown technical error variance were similar for all three methods for analysis of selective DNA pooling data. With central QTL or large family size, all four methods resulted in nearly unbiased estimates of QTL location (bias ≤ 4.5 cM) but with distal QTL and small family size, all four methods resulted in some bias toward the center of the chromosome. Biases were smallest for selective genotyping (< 5 cM) and greatest for LS-pool (9-11 cM). Estimates from ML-pool had similar biases as single marker analysis (6-8 cM). For all four methods and at all situations, presence of technical errors only slightly increased biases (<2 cM). For all four methods, standard errors (SE) of estimates of QTL location were reasonable with large family size (<12 cM) but large (11 to 21 cM) with small
family size. Single marker analysis had location estimates with the largest SE. With large family size, selective genotyping had smaller SE of location estimates than other methods. But with small family size, LS-pool had the smallest SE, even smaller than selective genotyping, except for distal QTL and presence of technical errors (see discussion). With small family size or central QTL, ML-pool had a larger SE of location estimates than LS-pool but smaller than single marker analysis. However, for the large family size and distal QTL, ML-pool estimates had slightly smaller SE than LS-pool estimates. Standard errors were up to 4.6 cM larger for distal than central QTL and presence of technical errors increased SE by 1 to 3.6 cM.

**Estimates of QTL effects:** Only interval mapping methods (LS-pool, ML-pool and selective genotyping methods) provided estimates of QTL effects. Single marker analysis did provide estimates of marker-associated effects but these were not evaluated. All methods gave unbiased or nearly unbiased estimates of QTL effects and similar SE of estimates (results not shown). Means and accuracy of estimates of QTL effects with known or unknown technical errors were essentially the same for LS-pool and ML-pool. Standard errors were small (0.06-0.07 \( \sigma_p \)) for large families (2000 progeny) but were doubled (0.13-0.14 \( \sigma_p \)) for small families (500 progeny). The ratio of SE of estimates of QTL effects was proportional to the square root of the ratio family size, as expected for estimates from regular linear regression. True QTL location and presence of technical error had little effect on estimates of QTL effects.

**Comparison of methods based on significant replicates:** Generally, only significant QTL mapping results are reported from actual experiments. Thus, it is also necessary to evaluate methods based on significant replicates only. Table 3 shows means and standard
errors of estimates of QTL location based on only significant replicates for the small family size. All methods had high power with large family size, so results were almost unchanged when comparing results from all versus only significant replicates and are not shown. Only QTL location estimates with known $V_{TE}$ are presented. Results with unknown $V_{TE}$ were similar to those with known $V_{TE}$.

Similar to results from all replicates (Table 2), biases in estimates of QTL position for significant QTL were negligible with central QTL (Table 3). When the QTL was distal, biases were reduced from 4.8 to 2.6 cM for selective genotyping, from 6-7 cM to 3-4 cM for single marker analysis and ML-pool, but from 10 to only 9 cM for LS-pool. Therefore, biases toward the center of estimates of location were nearly halved for the selective genotyping method, single marker analysis, and ML-pool, when considering only significant replicates, but a large bias remained for LS-pool with distal QTL. For ML-pool, single marker analysis, and selective genotyping, standard errors of estimates of QTL location were reduced by about 3 cM with central QTL and by 5-6 cM with distal QTL. But for LS-pool, standard errors were reduced only by 0-2 cM with central QTL and by about 3 cM with distal QTL. For all methods, the QTL effect was overestimated when selecting only significant results (mean estimates were 0.27 $\sigma_p$ while the true effect was 0.25 $\sigma_p$) but the SE of estimates was almost unchanged (results not shown). Comparisons of estimates of QTL location and effect among the four methods were similar when considering only significant instead of all replicates.
Bias in estimates of QTL location with LS-pool

**Reasons for bias:** Although all methods resulted in some biases of estimates of QTL location (Table 2), LS-pool resulted in the greatest biases, in particular when family size was small and the QTL was distal (Table 2). Biases were not reduced when considering only significant replicates. There are two reasons for bias when the QTL is distal, as will be demonstrated below: heterogeneous distribution of the test statistic across the chromosome and the asymmetric position of the QTL in relation to the parameter space when the QTL is distal.

Figure 1 shows the mean and variance of the test statistic under the null hypothesis at each putative QTL position for LS-pool, ML-pool, and selective genotyping, with small family size and unknown $V_{TE}$ of 0.0007. Both mean and variance of the F statistic were greater at positions around the center of the chromosome for LS-pool (Figure 1A), but similar across positions for ML-pool and selective genotyping methods (Figures 1B and 1C). The heterogeneous distribution of the test statistic, unique to the LS-pool method, is caused by the fact the LS-pool uses information from all markers simultaneously but does not account for correlations in frequencies between linked markers. This results in a greater mean and variance of the test statistic at central positions under the null hypothesis for LS-pool, where more markers are in the neighborhood of the evaluated position than at the ends of the chromosome.

The second reason for the biases toward center with LS-pool, the asymmetric position of the QTL, is common to all QTL mapping methods but causes larger bias for methods with lower power. Figure 2 shows the distribution of estimates of QTL location across the
chromosome from selective genotyping when the true QTL was at 11 or 46 cM. For a central QTL, estimates of position were symmetrically distributed around the true position and unbiased estimates were observed (Figure 2A). For a distal QTL, however, due to the asymmetric position of the QTL relative to the chromosome boundaries, errors in position estimates were fewer and smaller in the direction of the shorter region than toward the longer region. Therefore, position estimates were biased toward the center of the chromosome (Figure 2B). This bias was more prominent for 500 progeny than for 2000 progeny because, if power is high, estimates of QTL location are accurate and the chance to locate the QTL far away from the true position is more rare.

**Correcting for bias:** Based on reasons for biases in estimates of QTL location in LS-pool, different methods for correcting the bias were developed and evaluated:

*Using flanking markers only:* Table 4 shows power for QTL detection, means and standard errors of estimates of QTL location and effect when using all markers simultaneously or only the pair of flanking markers for each interval. Only results with small family size are presented because biases were negligible with large family size (Table 1). Results with unknown $V_{TE}$ are presented as an example. Using only flanking markers resulted in lower power, larger bias and SE for QTL position estimates, and larger bias but similar SE for QTL effect estimates. Therefore, using only flanking markers to map the QTL with selective DNA pooling data is not a good method compared to using all markers simultaneously.

*Standardization of the test statistic:* Table 5 shows power for QTL detection and means and standard errors of estimates of QTL location and effect with and without standardization of the test statistic. Results for small family size (500 progeny), a distal QTL and a $V_{TE}$ of
0.0007 or 0.0 are presented as an example. Standardization greatly reduced the bias in estimates of QTL position and increased power by 6-15%. However, it also increased the SE of estimates of QTL position by 4-7 cM. In addition, about 50% of the replicates located the QTL at one end of the chromosome and that proportion remained high when only significant replicates were considered (Figure 3 B). This indicates that standardization of the test statistic does not result in good estimates of QTL position for the LS-pool method. Estimates of QTL effect were not affected by standardization.

Parametric bootstrap: Figure 4 shows the distribution of estimates of QTL location with and without correction using the parametric bootstrap method based on 100 replicates. For each replicate, a correction table was generated from 500 bootstrap samples at each possible QTL location and used to correct the estimate of QTL location. Correction by parametric bootstrap reduced bias but increased the SE of estimates of QTL position, and, similar to results from standardization of the test statistic, a high proportion of replicates (>20%) tended to locate the QTL at one end of the chromosome.

Validation of the symmetry assumption

Table 6 shows the sum of true QTL allele frequencies over selected pools, power, and estimates of QTL location and of QTL substitution effects from LS-pool-1 (one parameter for QTL allele frequency) and LS-pool-2 (two parameters for QTL allele frequency, one for each pool), with no and complete dominance at the QTL and different QTL allele frequencies in the dam population. Results in Table 6 indicate that the sum of the true QTL allele frequencies over both selected pools was very close to one, which suggests that the symmetry
assumption was valid even if the QTL was dominant or the QTL frequency among dams deviated from 0.5. Method LS-pool-1 consistently had greater power to detect the QTL, and lower bias and standard errors of estimates of QTL location than LS-pool-2, except with complete dominance and high frequency (0.9) of the dominant QTL allele in the dam population, for which both methods had very low power and poor estimates. Estimates of QTL effects were similar and unbiased for both methods. The difference in power between LS-pool-1 and LS-pool-2 was about 20% when the QTL was co-dominant or when the frequency of the dominant QTL allele in the dam population was 0.5 or lower. Frequency of the QTL among dams had little effect on power and estimates of QTL location when the QTL was co-dominant but had a large impact with complete dominance. Low frequency of a dominant QTL allele in the dam population greatly increased power and precision of estimates of QTL location, while a high frequency decreased both power and precision of estimates of location. Estimates of QTL effect were similar for LS-pool-1 and LS-pool-2, were nearly unbiased, and had similar standard errors for all situations.

DISCUSSION

Comparison of methods

LS and ML pool versus single marker analysis: Simulation of a non-distal QTL for a large half-sib family design in WANG (2003) showed that both LS-pool and ML-pool provided nearly unbiased estimates of QTL location and effect and slightly greater power to detect the QTL than single marker analysis. This conclusion was further investigated for
several additional scenarios (small or large family size, central or distal QTL, and with or without technical error) in the current study.

In this study, positions of the marker nearest to the QTL were taken as the estimate of QTL position for the single marker analysis and compared with results from both interval mapping methods. Simulation results showed that, for almost all scenarios, both LS-pool and ML-pool provided greater power to detect the QTL, estimated QTL location with less bias and greater precision than the single marker analysis (Tables 1 and 2), and provided unbiased or nearly unbiased estimates of QTL effects, which are not available from single marker analysis. Both LS-pool and ML-pool provide separate estimates of QTL position and effect and utilize information from multiple linked markers, therefore, resulting in more accurate QTL mapping results and greater power to detect the QTL than the single marker analysis. The only exception was that, when family size was small, the QTL was distal, and technical error variance was known, LS-pool had slightly lower power and more biased but more accurate estimates of QTL location than the single marker analysis. The reason for this will be discussed later.

**LS and ML pool versus selective genotyping:** with selective genotyping, each individual in the pools is genotyped and no errors in genotyping are present. Note that the latter might not be true in reality but the error proportion is expected to be small. Therefore, power and accuracy of estimates of QTL location and effect from selective genotyping provide an upper bound for results from selective DNA pooling data. Although the best QTL mapping method with selective genotyping data is a maximum likelihood method (LANDER and BOTSTEIN 1989; Xu and VOGL 2000) instead of least squares regression, it requires
intensive computation and is not expected to give much better results than least squares for the balanced data sets simulated here.

Although simulation results showed that selective genotyping methods did indeed provide greater power to detect the QTL and more accurate estimates of QTL location than LS-pool and ML-pool, differences were small with large family size or central QTL (Tables 1 and 2). Differences were even smaller if only significant records were considered (Table 3). Method ML-pool provided only 3-6% lower power than selective genotyping, even with small family size and distal QTL. The LS-pool method provided higher power and greater accuracy of estimates of QTL location than the selective genotyping method when family size was small, the QTL was central, and $V_{TE}$ was known. This was not expected and reasons for this result will be discussed in the next section. Methods LS-pool, ML-pool, and selective genotyping provided unbiased or nearly unbiased estimates of QTL effect and estimates had similar accuracy for the three methods. These results suggest that most QTL information from selective genotyping data is contained the allele frequencies in the phenotypic extremes and both LS-pool and ML-pool can efficiently retrieve this information, even if a certain level of error is present in estimates of marker allele frequencies (DARVASI and SOLLER 1994).

**LS-pool versus ML-pool:** Methods LS-pool and ML-pool provided similar QTL mapping results for the large family size (Tables 1 and 2). When family size was small, ML-pool had greater power to detect the QTL, with the exception of a central QTL and known technical error variance (Table 1). Method LS-pool resulted in severe biases in estimates of QTL location with small family size and distal QTL. These biases were reduced with ML-
pool but estimates from ML-pool were less accurate than estimates from LS-pool (Table 2). Estimates of QTL effects from LS-pool and ML-pool were similar for all situations.

**Effect of family size:** Power for QTL detection and estimates of QTL location and effect from LS-pool and ML-pool methods were substantially affected by family size. With a large family size, both LS-pool and ML-pool provided almost 100% power to detect the QTL, nearly unbiased estimates of QTL location, and unbiased estimates of QTL effect (Table 1 and 2). This is an encouraging result. Genotyping cost is usually prohibitively high with large size families and our results indicate that, with large families, analysis of selective DNA pooling data by LS-pool or ML-pool provides similar QTL mapping results as selective genotyping, while reducing genotyping costs a hundred-fold or more. Since artificial insemination has been a widely used technique in dairy cattle, half-sib families with more than 1,000 progeny are frequently observed. For these types of families, both LS-pool and ML-pool can be directly used to map QTL and individual genotyping is not expected to provide much additional information. With small family size, however, both interval mapping methods resulted in moderate power and biased estimates of QTL location, especially with LS-pool. In addition, accuracy of estimates of QTL location and effect were low, so 95% confidence intervals might cover the whole chromosome. Although similar problems exist for selective genotyping with small family size, problems were greater with pooling methods. Therefore, it might be worthwhile to use individual genotyping methods to map the QTL for small and moderate size families, where the genotyping cost is usually acceptable. However, selective DNA pooling methods are still recommended as a first step scan to identify QTL regions, followed by individual genotyping for a finer map. This two-step genotyping strategy can also substantially reduce genotyping costs.
Effect of true QTL position: Power and estimates of QTL location from LS-pool were(117,181),(905,977) also heavily affected by the true QTL position when the family size was small. For a central QTL, LS-pool had high power (even higher than selective genotyping) and unbiased estimates of QTL location with small SE (even smaller than selective genotyping). For a distal QTL, however, LS-pool had low power (even lower than single marker analysis) and severely biased estimates of QTL location. Some of these results were unexpected and reasons will be discussed later. Results from ML-pool were less affected by QTL position, although slightly higher power, less bias, and smaller SE were also observed with a central compared to a distal QTL, but biases were smaller than for LS-pool. Since one does not know where the QTL is on the chromosome, ML-pool would be the preferred method from this point of view.

Effect of prior knowledge of $V_{TE}$: Utilizing prior knowledge of the amount of technical error ($V_{TE}$) did not affect estimates of QTL location and QTL effect but resulted in greater power to detect the QTL. The increment in power was greater than 8% for LS-pool but small ($\leq 2\%$) for ML-pool (Table 1). The small increment for ML-pool is probably due to more accurate estimates of $V_{TE}$ for ML-pool than LS-pool when $V_{TE}$ is unknown. Therefore, it is useful to conduct initial studies to estimate $V_{TE}$ if LS-pool will be used but this may not be necessary for ML-pool. However, it is always very important to conduct preliminary studies to eliminate markers with problems of differential amplification and to ensure that estimates of marker allele frequencies are reasonably accurate because all information comes from allele frequencies in the selective DNA pooling approach.

Effect of magnitude of $V_{TE}$: Simulation results showed that the magnitude of $V_{TE}$ only had a small effect on QTL mapping results for all three pool analysis methods (Tables 1 and
The effect of $V_{TE}$ was found to be larger in previous research (DARVASI and SOLLER 1994; BARO et al. 2001) but they evaluated a very wide range for $V_{TE}$ (0-0.1). Interval mapping methods that simultaneously use multiple markers should theoretically be more robust to technical error than the single marker analysis because large technical errors in some marker allele frequencies will be averaged out by considering information from linked markers. However, this trend was not very clear in the current study (Tables 1 and 2).

In the simulation, technical errors were assumed to be independent across markers. In practice, however, allele frequencies on linked markers are usually estimated from the same batch, by the same machine, and laboratory analysis are conducted by the same person. In addition, there will be variation in the amount of DNA that is present in the pool from each individual. All these factors cause correlations between technical errors at linked markers, which were ignored in the current study. Ignoring correlations among technical errors might also result in biases in estimates of QTL location, for the same reason as ignoring correlations among sampling errors in the LS-pool.

**General discussion on LS-pool and ML-pool:** Based on our simulation results, ML-pool is a better method than LS-pool to map the QTL with selective DNA pooling data. In general, ML-pool had greater power to detect the QTL, smaller biases in QTL location estimates than LS-pool, and results from ML-pool were more stable for central or distal QTL and with or without prior information about the magnitude of technical errors. These advantages of ML-pool over LS-pool exist because ML-pool accounts for correlations in allele frequencies between linked markers and is, therefore, based on a more appropriate model than the LS-pool method. However, ML-pool may not give better QTL mapping results than LS-pool for all situations. For example, when there is a limited number of
individuals in the pool, the approximate normality assumption of marker allele frequencies will not hold, the derived covariance matrix might deviate from the true matrix, and LS-pool may provide better results. In addition, ML-pool is computationally intensive, while LS-pool can be readily applied with standard statistical software. The LS-pool method also provided high power and nearly unbiased estimates of QTL location and effects in most cases, except for the situation of small family size and distal QTL. Therefore, LS-pool is a very useful method to map QTL with selective DNA pooling data and it is important to investigate reasons why LS-pool resulted in severe biases in QTL location estimates for some cases and find ways to correct them. In addition, since estimates of QTL location from all methods resulted in some biases in estimates of QTL location, methods to successfully correct biases for LS-pool may also be applied to correct biases from other methods.

Correcting the bias in estimates of QTL location from the LS-pool method

Reasons for bias: Biases of estimates of QTL location toward the center existed for all QTL mapping methods for distal QTL but were severe for the LS-pool method because of the heterogeneous distribution of the test statistic across the chromosome (Figure 1). This heterogeneous distribution of the test statistic also caused power to detect the QTL to be overestimated for central QTL and underestimated for distal QTL, since a uniform significance threshold was applied while the test statistic tended to be greater around the center and lower close to the end. This is why power was greater for LS-pool than for selective genotyping for central QTL (Table 1). Also, for the same reason, standard errors of
estimates of QTL location were lower with LS-pool than for selective genotyping, since estimates are regressed toward the center with LS-pool when power is limited (Table 2).

**Using flanking markers:** This method did not successfully correct biases in estimates of QTL location from LS-pool (Table 4) but, instead, increased biases, reduced power to detect the QTL, reduced accuracy of location estimates, and increased bias in QTL effect estimates. The poorer QTL mapping results are caused by the limited information that is contained in only a pair of observations from the interval flanking markers. Regular interval mapping also uses only flanking markers but individual genotype information is available. With selective DNA pooling, all information comes from marker allele frequencies. In addition, with regular interval mapping, markers outside informative flanking markers do not provide additional information but outside markers do provide information in the selective DNA pooling method in relation to technical errors in estimated allele frequencies.

**Standardization of the test statistic:** This method helped to reduce biases in estimates of QTL position and also increased power to detect the QTL. But this was at the cost of reduced accuracy of estimates of QTL position (Table 5). In addition, a large proportion of replicates located the QTL at one end of the chromosome (Figure 3). This method also required much computing to obtain empirical means and variances of the test statistic under the null hypothesis, which makes it difficult to apply.

**Parametric bootstrap:** Similar to the standardization method, this method helped to reduce biases but also reduced accuracy of estimates of QTL position and a high proportion of replicates was located the QTL at one end of the chromosome (Figure 3). This method was also computationally demanding and, therefore, difficult to apply.
In summary, the three methods that were evaluated to attempt to reduce the bias in estimates of QTL location did, or have the potential to, reduce the bias but each also generated other problems (Tables 4 and 5, Figures 3 and 4). Further research is needed to effectively correct biases in estimates of QTL location.

**Validation of the symmetry assumption**

Both LS-pool and ML-pool methods were robust to deviations from the symmetry assumption due to dominance of QTL and different QTL allele frequencies among dams. Thus, QTL allele frequency in the upper pool is highly correlated with the frequency in the lower pool in most situations and it is redundant to include two frequency parameters in the model. In addition, LS-pool-1 has fewer parameters to estimate than LS-pool-2, which explains why LS-pool-1 always had greater power, smaller bias, and greater accuracy of estimates of QTL location than LS-pool-2 (Table 6).

When the QTL is dominant and the dominant allele is rare in the dam population, the ability to detect the QTL is large but when the QTL is dominant but the frequency of the dominant allele is greater than 0.5 in dam population, it was almost not possible to detect a QTL of moderate effect (Table 6). A similar result was also found in the single marker analysis (BARO et al. 2001). Dominance of the QTL and the frequency of that allele in the dam population affects the true substitution effect of the QTL (FALCONER and MACKAY 1996), which determines the power to detect the QTL and, thereby, affects the bias and accuracy of estimates of QTL location and effect. When the QTL is co-dominant, the substitution effect is the same regardless of the frequency in the dam population; when the
QTL is completely dominant, the QTL effect is larger if the dominant QTL allele has a lower frequency in the dam population but it is smaller with a higher frequency. Therefore, the magnitude of the QTL substitution effect, instead of the additive or dominance effect alone, has an important effect on QTL mapping results with selective DNA pooling data from a half-sib design.

CONCLUSIONS

Simulation results from the current study show that both LS-pool and ML-pool provide greater power to detect the QTL than the currently used single marker analysis and provide separate estimates of QTL location and effect, which are not available from the single marker analysis. With large family size, both LS-pool and ML-pool provide similar power and estimates of QTL location as selective genotyping analysis. Method LS-pool, however, results in severe biases in estimates of QTL location with small family size and distal QTL. Biases are in part caused by a heterogenous distribution of the test statistic and were reduced with ML-pool. Having prior knowledge of the variance of technical error improves power for LS-pool but the improvement is limited for ML-pool. Presence of technical errors has a small effect on power and estimates of QTL parameters. Several methods were evaluated to correct biases in QTL location estimates from LS-pool. Using flanking markers did not correct the bias. Standardization of the test statistic and parametric bootstrap provided results with less bias but a greater proportion of estimates at a distal marker.
LITERATURE CITED


LIPKIN, E., J. FULTON, H. CHENG, N. YONASH and M. SOLLER, 2002 Quantitative trait locus mapping in chickens by selective DNA pooling with dinucleotide microsatellite markers by using purified DNA and fresh or frozen red blood cells as applied to marker-assisted selection. Poult Sci 81: 283-292.


TABLE 1

Power (%) to detect the QTL with selective DNA pooling data using least squares (LS-pool), maximum likelihood (ML-pool) and single marker analysis and of least squares analysis with selective genotyping data. Variance of technical errors from pooling ($V_{TE}$) was unknown or known. There were 10 half-sib families with 500 or 2000 progeny and the QTL effect was 0.25 phenotypic standard deviations at 11 or 46 cM on a 100 cM chromosome with 6 equidistant fully informative markers. The selected proportion was 10% in each pool and the $V_{TE}$ was 0.0007 or 0. Results of selective genotyping were independent of $V_{TE}$ and are presented twice. Results were based on 3,000 replicates and the 5% chromosome-wise thresholds were obtained from 10,000 replicate of simulation under the null hypothesis.

<table>
<thead>
<tr>
<th>Family Size</th>
<th>$V_{TE}$ (x10^4)</th>
<th>QTL location</th>
<th>Selective DNA Pooling</th>
<th>Selective genotyping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS-pool $V_{TE}$ un/known</td>
<td>ML-pool $V_{TE}$ un/known</td>
</tr>
<tr>
<td>500</td>
<td>7</td>
<td>11</td>
<td>56 / 67</td>
<td>72 / 72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46</td>
<td>70 / 78</td>
<td>73 / 73</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>11</td>
<td>57 / 70</td>
<td>74 / 75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46</td>
<td>70 / 80</td>
<td>77 / 77</td>
</tr>
<tr>
<td>2000</td>
<td>7</td>
<td>11</td>
<td>97 / 98</td>
<td>99 / 99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46</td>
<td>99 / 99</td>
<td>99 / 99</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>11</td>
<td>98 / 99</td>
<td>99 / 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46</td>
<td>99 / 99</td>
<td>99 / 100</td>
</tr>
</tbody>
</table>
TABLE 2

Mean and standard error (in brackets) of estimates of QTL location with selective DNA pooling data using least squares (LS-pool), maximum likelihood (ML-pool) and single marker analysis and of least squares analysis with selective genotyping data. QTL location estimates were essentially same with unknown or known $V_{TE}$. Results of selective genotyping were independent of $V_{TE}$ and are presented twice. Results were based on 3,000 replicates. Other simulation parameters were the same as Table 1.

<table>
<thead>
<tr>
<th>Family Size</th>
<th>$V_{TE}$ (x10^4)</th>
<th>QTL location (cM)</th>
<th>Selective DNA Pooling</th>
<th>Selective genotyping (cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS-pool (cM)</td>
<td>ML-pool (cM)</td>
</tr>
<tr>
<td>500</td>
<td>7</td>
<td>11</td>
<td>21.1 (16.7)</td>
<td>18.0 (19.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>46.6 (12.1)</td>
<td>45.6 (16.2)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>11</td>
<td>20.4 (15.1)</td>
<td>16.8 (17.6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>46.5 (11.1)</td>
<td>45.2 (14.8)</td>
</tr>
<tr>
<td>2000</td>
<td>7</td>
<td>11</td>
<td>13.4 (7.7)</td>
<td>12.0 (7.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>45.8 (5.7)</td>
<td>45.3 (6.7)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>11</td>
<td>12.8 (5.9)</td>
<td>11.3 (4.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>46.9 (4.5)</td>
<td>45.4 (4.6)</td>
</tr>
</tbody>
</table>
Mean and standard error (in brackets) of estimates of QTL location with selective DNA pooling data using least squares (LS-pool), maximum likelihood (ML-pool) and single marker analysis and of least squares analysis with selective genotyping data based on only significant replicates. Only QTL location estimates with known $V_{TE}$ were presented as an example. Results of selective genotyping were independent of $V_{TE}$ and are presented twice. Other simulation parameters were the same as Table 1, except that only results with 500 progeny were presented.

<table>
<thead>
<tr>
<th>Family Size</th>
<th>$V_{TE}$ (x10^4)</th>
<th>QTL location (cM)</th>
<th>Selective DNA Pooling</th>
<th>Selective genotyping (cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>7</td>
<td>11</td>
<td>20.1 (13.2)</td>
<td>14.7 (13.9)</td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
<td>46.6 (12.1)</td>
<td>45.5 (13.1)</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>11</td>
<td>19.7 (11.9)</td>
<td>14.0 (11.6)</td>
</tr>
<tr>
<td>46</td>
<td></td>
<td></td>
<td>46.1 (9.4)</td>
<td>45.1 (11.9)</td>
</tr>
</tbody>
</table>
TABLE 4

Power to detect the QTL, means and standard errors of estimates of QTL location and effect using all markers simultaneously or only flanking markers in LS-pool. Ten half-sib families with 500 progeny were used. Results with unknown variance of technical errors are presented as an example. Other simulation parameters were the same as Table 1.

<table>
<thead>
<tr>
<th>$V_{TE}$ (x10^2)</th>
<th>QTL location</th>
<th>Estimates</th>
<th>All markers</th>
<th>Flanking markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Power</td>
<td>56%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Position – mean (SE)</td>
<td>21.1 (16.7)</td>
<td>23.3 (23.3)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QTL effect – mean (SE)</td>
<td>0.24 (0.14)</td>
<td>0.23 (0.14)</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>Power</td>
<td>70%</td>
<td>54%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Position – mean (SE)</td>
<td>46.6 (12.1)</td>
<td>47.3 (18.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QTL effect – mean (SE)</td>
<td>0.25 (0.14)</td>
<td>0.24 (0.14)</td>
</tr>
<tr>
<td>0</td>
<td>11</td>
<td>Power</td>
<td>57%</td>
<td>51%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Position – mean (SE)</td>
<td>20.4 (15.1)</td>
<td>22.2 (21.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QTL effect – mean (SE)</td>
<td>0.24 (0.13)</td>
<td>0.23 (0.13)</td>
</tr>
<tr>
<td>46</td>
<td>11</td>
<td>Power</td>
<td>70%</td>
<td>56%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Position – mean (SE)</td>
<td>46.5 (11.1)</td>
<td>46.9 (17.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QTL effect – mean (SE)</td>
<td>0.25 (0.13)</td>
<td>0.23 (0.13)</td>
</tr>
</tbody>
</table>
TABLE 5

Power to detect the QTL, means and standard errors of estimates of QTL location and effect from LS-pool, with and without standardization. Ten half-sib families with 500 progeny were used. Results with unknown $V_{TE} = 0.0007$ are presented as an example. Other simulation parameters were the same as in Table 1.

<table>
<thead>
<tr>
<th>QTL location</th>
<th>$V_{TE}$ (x10^4)</th>
<th>Power</th>
<th>LS-pool</th>
<th>Without standardization</th>
<th>With standardization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>56%</td>
<td>62%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21.1 (16.7)</td>
<td>11.7 (20.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.24 (0.14)</td>
<td>0.25 (0.14)</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>57%</td>
<td>72%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20.4 (15.1)</td>
<td>11.0 (18.7)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.24 (0.13)</td>
<td>0.25 (0.13)</td>
</tr>
</tbody>
</table>
Sum of true QTL allele frequency over both selected tails and power to detect the QTL, estimates of QTL location and effect from least squares methods with QTL allele frequency in the upper tail or frequencies in both tails as parameters (LS-pool-1 and LS-pool-2) with no and complete dominance level and different QTL allele frequencies in the dam population. Ten half-sib families with 500 progeny were used and the true QTL was at 11 cM. Results with unknown technical error variance equal to 0.0 are presented as an example. Other simulation parameters were the same as Table 1.

<table>
<thead>
<tr>
<th>QTL dominance</th>
<th>Dam QTL frequency</th>
<th>Sum of true QTL allele frequency over both tails</th>
<th>Power (%)</th>
<th>QTL location</th>
<th>QTL effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-dominance (d=0)</td>
<td>0.3</td>
<td>1.00</td>
<td>56</td>
<td>34</td>
<td>20.3 (14.8)</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>1.00</td>
<td>56</td>
<td>35</td>
<td>20.4 (15.1)</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>1.00</td>
<td>56</td>
<td>34</td>
<td>20.1 (14.6)</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>1.00</td>
<td>57</td>
<td>35</td>
<td>20.3 (15.1)</td>
</tr>
<tr>
<td>Complete dominance</td>
<td>0.3</td>
<td>0.97</td>
<td>85</td>
<td>67</td>
<td>15.7 (9.6)</td>
</tr>
<tr>
<td>(d=a)</td>
<td>0.5</td>
<td>0.97</td>
<td>53</td>
<td>34</td>
<td>20.4 (15.5)</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.97</td>
<td>18</td>
<td>12</td>
<td>32.0 (24.0)</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.99</td>
<td>5</td>
<td>6</td>
<td>47.0 (27.9)</td>
</tr>
</tbody>
</table>
FIGURE 1. — Mean and variance of the test statistic at each possible QTL position for the LS-pool, ML-pool and selective genotyping methods under the null hypothesis. Ten half-sib families with 500 progeny were used. The variance of technical errors from pooling ($V_{TE}$) was 0.0007 and assumed unknown. Results were based on 100,000 replicates for LS-pool and ML-pool and 10,000 replicates for selective genotyping method. Other simulation parameters were the same as in Table 1.

Plot A. F statistic from LS-pool

Plot B. LR statistic from ML-pool

Plot C. F statistic from selective genotyping
FIGURE 2. — Frequencies of QTL location estimates for the selective genotyping method under two alternatives. Ten half-sib families with 500 progeny were used. The substitution QTL effect was 0.25 phenotypic standard deviation. The selected proportion was 10% in each pool. Results were based on 3,000 replicates.

**Plot A. True QTL = 46 cM**

Average estimates of QTL location = 45.6

**Plot B. True QTL = 11 cM**

Average estimate of QTL location = 15.8 cM
FIGURE 3. — Distribution of estimates of QTL location from LS-pool with and without standardization. Ten half-sib families with 500 progeny were used and the true QTL was at 11 cM. Results with unknown technical variance equal to 0.0007 are shown as an example. Other simulation parameters were the same as Table 1.

Plot A. Distribution without standardization

Plot B. Distribution with standardization
FIGURE 4. — Distribution of QTL location with and without correction by parametric
bootstrap from LS-pool. Ten half-sib families with 500 progeny were used and the true QTL
is at 11 cM and the substitution effect of QTL was 0.25 phenotypic standard deviation.
Results with unknown variance of technical error of 0.0 were presented as an example. For
each replicate, a correction table was generated based on 500 bootstrap samples and results
below were from 100 replicates.

**Plot A. Before correction**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.01</td>
<td>15.12</td>
</tr>
</tbody>
</table>

**Plot B. After correction**

<table>
<thead>
<tr>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.66</td>
<td>17.70</td>
</tr>
</tbody>
</table>
CHAPTER 5. DETECTION OF QTL AFFECTING ECONOMIC TRAITS IN LAYERS AND COMPARISON OF RESULTS FROM SELECTIVE DNA POOLING AND INDIVIDUAL GENOTYPING

In preparation for submission to: Genetics Selection Evolution

Abstract - Egg quality traits are important for the layer chicken industry but few studies have been conducted to detect quantitative trait loci (QTL) for these traits. Selective DNA pooling is known as an efficient method to detect QTL. In this study, an F₂ cross between two partially inbred commercial lines was used to detect QTL for egg quality and production traits. First, selective DNA pooling data from five grandparent families were used to detect QTL across the whole genome by single marker analysis. Second, individual genotyping data for one grandparent family for two chromosomes of interest were analyzed by least squares interval mapping. For comparison, single marker analysis and two interval mapping methods, LS-pool and ML-pool, were applied to the selective DNA pooling data from the same family and chromosomes. Individual genotyping analysis identified a region on chromosome 4 with significant QTL for body weight, egg weight, puncture score, albumen height and early production rate and a region on chromosome Z with significant QTL for sexual maturity and early puncture score. Both LS-pool and ML-pool provided similar power for QTL detection and estimates of QTL location and QTL effect as interval mapping analysis of individual genotyping data. Therefore, the two DNA pooling interval mapping methods successfully captured most information to detect the QTL, while substantially reducing genotyping costs.

QTL mapping/selective DNA pooling/layer chickens
1. INTRODUCTION

Egg quality and production traits are important economic traits for the layer chicken industry. However, these traits are sex-limited, cannot be measured until after sexual maturity, and some are difficult and expensive to measure. In addition, heritabilities of these traits are small to moderate [39]. Therefore, it is difficult to make efficient improvement in these traits by traditional selection and marker-assisted selection (MAS) could be a good alternative approach.

For MAS to be effective, it is critical to estimate the position and effect of loci that control the traits (quantitative trait loci, QTL) or identify markers associated with QTL. Recent molecular research has provided powerful tools for accurate QTL mapping. The most recent consensus genetic linkage map for the chicken contains more than 800 microsatellite markers [13] and the first draft sequence of the chicken genome is expected in 2004 [1]. However, not many genome wide QTL scans have been conducted in poultry and most focus on QTL for growth and carcass traits in broilers [8, 16, 26, 30, 31, 34-36] or QTL for resistance to disease [18, 20, 25, 33, 40-45].

Tuiskula-Haavisto et al. [32] conducted a genome scan for QTL controlling several egg quality and production traits in a cross of two layer breeds (Rhode Island Red and White Leghorn). They identified significant QTL for body weight, egg weight, and feed intake in the same region on chromosome 4. They also found QTL for eggshell strength, age at first egg, egg weight, and number of eggs on chromosome Z, and for albumen height on chromosome 2. Wardecka et al. [38] identified significant markers associated with the same traits in a cross between the Rhode Island Red and a native Polish breed, Green-legged
Parthenogenetic. Results from this research confirmed associations between some microsatellite markers and egg quality and production traits that were identified by Tuiskula-Haavisto et al. [32].

Both above studies involved crossing two different breeds and did not provide insight into whether the identified QTL are segregating in current commercial lines that have been selected for more than 40 generations. For practical MAS, identified QTL must be confirmed in current commercial populations.

To achieve reasonable power to detect small to moderate QTL, a substantial number of individuals must be genotyped [5]. Prohibitive costs of genotyping severely impede whole genome scans for QTL. Selective DNA pooling has been proposed to reduce genotyping costs by comparing marker allele frequencies in pooled DNA from phenotypically extreme individuals [4] and has been successfully used to detect markers linked to QTL in dairy cattle [24, 27] and poultry [22, 23, 44]. Results indicated that pooling data can effectively detect QTL that would also be found by individual genotyping of individuals in the phenotypic extremes (selective genotyping). Therefore, selective DNA pooling data can be used as a first step scanning for QTL, followed by selective or whole population genotyping. Zhou et al. [44] recently employed this strategy in poultry and found it to be efficient and successful.

The current method to analyze selective DNA pooling data is based on single marker analysis [4]. This method cannot provide separate estimates of QTL location and QTL effect, nor can it utilize the joint information from multiple markers. Wang [37] developed two interval mapping methods - one based on least squares regression (LS-pool) and one based on maximum likelihood (ML-pool) - to overcome those problems. Simulation results showed
that both interval mapping methods provided higher power to detect the QTL than single marker analysis and separate estimates of QTL effects.

The first objective of the current study was to detect QTL for several important egg quality and production traits segregating in an F₂ cross of two partially inbred commercial layer lines. The second objective was to verify the success of the selective DNA pooling approach and to evaluate the new interval mapping methods for pooling data by comparing results from DNA pooling data with results from individual genotyping analysis.

In the research described here, a two-step genotyping strategy was used to detect QTL. First, selective DNA pooling data of F₂ hens derived from five grandparent families were used to detect QTL across the whole genome by single marker analysis. Second, individual genotyping data from one of the five grandparent families on two chromosomes that showed QTL in the first step, 4 and Z, were analyzed by least squares interval mapping. For comparison purposes, single marker analysis, LS-pool and ML-pool were applied to the selective DNA pooling data from the same family.

2. MATERIALS AND METHODS

2.1. Family structure and management.

Two elite commercial layer lines of White Leghorns were used to produce the experimental cross. Both lines had been maintained as distinct pure-breeding populations with extensive selection for layer production traits for over 40 generations. Microsatellite screening showed that the two lines had 80% and 60% homogeneity of alleles. This
homogeneity resulted from the inbreeding program that was used to develop these lines prior to the 1970's.

Five males and five females were randomly selected from each line and mated in pairs. From each of the five matings, 5 to 7 $F_1$ males were randomly chosen and each mated to two full sisters. For each of the five grandparent families, 127 to 203 $F_2$ hens were produced, resulting in a total population size of 824 hens for early measured traits and 705 for late measured traits. All birds were reared and maintained at the Hy-Line International research farm under common commercial management conditions.

The single grandparent family used for individual genotyping analysis and analyzed in this study, consisted of 11 full-sib families derived from 7 $F_1$ males. Each full-sib family had 12 to 20 $F_2$ hens produced in 8 hatches. The total family size was 191 for early measured traits and 164 for late measured traits (Table I).

### 2.2. Traits measured

Traits measured included six egg quality traits (first egg weight, egg weight, egg deviation, eggshell color, puncture score, and albumen height) and three production traits (body weight, sexual maturity, and egg production rate). Most traits were measured twice, the early measurement was at 28 weeks of age and the late measurement was at 42 or 43 weeks of age. Early and late measurements were considered as different traits, thus a total of 15 traits were evaluated. More detailed trait descriptions are in Table I.

Sample sizes, means, and standard deviation for each trait are listed in Table I. For each trait, potential outliers were checked by SAS, PROC UNIVARIATE (SAS Institute, Cary,
Most traits had 2 to 6 potential outliers but excluding them did not affect QTL mapping results and, therefore, results with all observations will be shown. The normality assumption was also checked for each trait by normal probability plot (SAS) but no severe deviations from normality were observed.

All traits listed in Table I were evaluated by individual genotyping analysis. Body weight and eggshell color were not evaluated by selective DNA pooling.

2.3. Genotyping procedures

Selective DNA pooling was employed as a first-step search for QTL across the genome. Pool data was collected for all five grandparent families. Two chromosomes that showed evidence of QTL in this step (4 and Z) were confirmed by individual genotyping of one of the five grandparent families. The QTL mapping results from pooling and individual genotyping were compared based on this family.

2.3.1. DNA isolating, marker selection and genotyping.

For each F₂ bird, DNA was isolated using a modified salt-precipitation procedure. Briefly, 50μl of packed RBC (either frozen or fresh) was resuspended in 3 ml of Tris buffer (10mM Tris-HCL, pH 8); 400mM NaCl; 2mM EDTA), lysed with SDS, and then digested with proteinase K overnight at 37°C. The DNA was precipitated with NaCl and ethanol and resuspended in TE (10mM Tris-HCL, pH 8 and 2mM EDTA). Spectrophotometric readings
showed that all DNA had 260/280 ratios between 1.8 and 2.0, with yields ranging from 200-600μg. The DNA was diluted to 25ng/μl prior to use.

A total of 184 microsatellite markers were selected for DNA pool genotyping based on prior screening of the two parental lines. All markers were fully informative between the two parental lines, or between mating pairs. Markers were chosen to cover the genome, but some chromosomes lacked informative markers for some regions.

2.3.2. DNA pooling.

Within each grandparent family, F₂ hens were ranked within each full-sib family by phenotype for a given trait and the top and bottom 20% chicks were selected. Purified DNA from individuals in the same selected tail was pooled across full-sib families within the same grandparent family.

Pools of DNA were constructed by combining equal volumes of quantified DNA (25ng/μl) from each selected bird. Marker allele frequencies were determined for each pool based on peak height following separation of PCR products on an ABI 377 instrument. Corrections were performed for internal repeat number, overlapping shadow bands, and differential amplification, using procedures described by Lipkin et al. [23]. The difference in marker allele frequencies between the upper and lower pools (D value) was calculated for each marker within each grandparent family and used for analysis. This procedure was repeated for each trait.
2.3.3 Individual genotyping

Two chromosomes (4 and Z) that were found to harbor QTL by selective DNA pooling were further analyzed using individual genotyping of the grandparental family with large D values for egg weight. In this step, 4 fully informative markers were added to chromosome Z, resulting in a total of 13 markers, and 18 fully informative markers were added to chromosome 4, for a total of 33 markers.

2.3.4 Linkage analysis.

Linkage maps of chromosomes 4 and Z were estimated from individual genotypes of F2 hens from one grandparent family using CRI-MAP version 2.4 [12]. Option BUILD with LOD score 3 was used for configuration of a linkage group. Then, option FLIPS(2) was used to check optional orders between adjacent markers and option FIXED was used to get final map distances. The sex-averaged map was used for autosomes and the male specific map was used for chromosome Z in the QTL mapping analysis.

2.4. QTL mapping

2.4.1 QTL mapping with selective DNA pooling data

Single marker analysis was used to detect QTL across the genome with selective DNA pooling data from the five grandparent families. For chromosomes 4 and Z, single marker
analysis and two interval mapping methods, LS-pool and ML-pool, were then applied to the pooling data of the grandparent family that had individual genotyping data.

2.4.1.1. Single marker analysis.

Under the null hypothesis of no QTL, the difference in marker allele frequencies between the high and the low pools (D) is expected to be zero. This hypothesis is rejected with type I error $\alpha$ if:

$$Z_{D_{M_{ij}}} = \frac{D_{M_{ij}}}{SE_{D_{M_{ij}}}} > Z_{1-\alpha/2}$$

where $D_{M_{ij}}$ is the D value of the $j^{th}$ marker of the $i^{th}$ grandparent family and $SE_{D_{M_{ij}}}$ is its associated standard error [4]. Taking account of binomial sampling of marker alleles and technical errors in creating DNA pools and measuring the allele frequency, standard errors of D values ($SE_{D_{M_{ij}}}$) were derived as:

$$SE_{D_{M_{ij}}} = \sqrt{2 \times \left( \frac{0.25}{2n \times V_{TE}} \right)}$$

where 0.25 is the expected variance of the marker allele frequency from a binomial distribution under the null hypothesis, $n$ is the number of individuals in each pool (assumed the same in both pools) and $V_{TE}$ is the variance of technical errors in observed marker allele frequencies in the DNA pool. A value of 0.0016 was used for $V_{TE}$, which was estimated from preliminary studies by comparing estimates of marker allele frequencies in pools with the true frequency, as obtained from individual genotyping.
For each marker, evidence of linkage with a QTL was tested based on a Chi-square test for joint analysis of the five grandparent families [4]. That is, for marker $j$, the null hypothesis was rejected with type I error $\alpha$ if

$$\sum_{i=1}^{k} Z_{D_{M_i}}^2 > \chi_{\alpha,5}^2 \quad df = 5$$

2.4.1.2. Least squares interval mapping of pool data (LS-pool).

Details of the LS-pool method applied to D values were described in [37]. In this study, LS-pool was applied to one single grandparental family with D values for $k$ markers ($D_{M_1}, D_{M_2}, \ldots, D_{M_k}$). The D values were modeled as:

$$\begin{bmatrix} D_{M_1} \\ D_{M_2} \\ \vdots \\ D_{M_k} \end{bmatrix} = \begin{bmatrix} 1 - 2r_1 \\ 1 - 2r_2 \\ \vdots \\ 1 - 2r_k \end{bmatrix} D_Q + \begin{bmatrix} e_{D_{M_1}} \\ e_{D_{M_2}} \\ \vdots \\ e_{D_{M_k}} \end{bmatrix}$$

(Model 1)

where $D_{M_j}$ is the observed D value of the $j^{th}$ marker, $D_Q$ is the expected D value for the QTL allele in the family, $r_j$ is the recombination rate between marker $j$ and the QTL, and $e_{D_{M_j}}$ are residuals. Model 1 can be written in the format of matrices as:

$$Y = XD_Q + e$$

where $Y$ is a vector of observed marker D values, $X$ is a vector with recombination rates and $e$ is a vector of residuals.
For a given putative QTL position, ordinary least squares estimation was used to fit the model, and the following statistic was calculated:

\[ \chi^2 = \frac{Y'X(XX')^{-1}XY}{2n + V_{TE}} \]

The position providing the highest value of the test statistic was taken as the estimate of QTL position and the estimate of the QTL allele frequency at this position was used to estimate the QTL substitution effect, as described by Dekkers [9].

2.4.1.3. Maximum likelihood interval mapping of pool data (ML-pool)

The ML-pool method was also applied to data from one grandparent family. Details of the method were described in [37]. The distribution of observed D values of multiple markers was approximated by a multivariate normal distribution with mean \( \mu_D \):

\[
\mu_D = E \begin{bmatrix}
D_{M_1} \\
D_{M_2} \\
\vdots \\
D_{M_k}
\end{bmatrix} = \begin{bmatrix}
2p_{M_1}^{u} - 1 \\
2p_{M_2}^{u} - 1 \\
\vdots \\
2p_{M_k}^{u} - 1
\end{bmatrix} = \begin{bmatrix}
2^*[(1-r_1)p_{Q}^u + r_1(1-p_{Q}^u)] - 1 \\
2^*[(1-r_2)p_{Q}^u + r_2(1-p_{Q}^u)] - 1 \\
\vdots \\
2^*[(1-r_k)p_{Q}^u + r_k(1-p_{Q}^u)] - 1
\end{bmatrix}
\]

where \( p_{M_j}^u \) is the expected allele frequency for marker \( j \) in the upper pool, which was assumed to be equal to \( 1 - p_{M_j}^L \), with the assumption of a symmetric distribution of phenotypic values, and \( p_{Q}^u \) is the expected QTL allele frequency in the upper pool.

The variance covariance matrix of D-values, \( \Sigma_D \), is:
where

\[
\text{var}(D_{M_j}) = 2 \times \text{var}(f_{M_j}^{\prime\prime}) = 2 \times \left[ \frac{p_{M_j}^{U} (1 - p_{M_j}^{U})}{2n} + V_{TE} \right]
\]

If markers \( j \) and \( l \) bracket the QTL (M\(_j\)-Q-M\(_l\)):

\[
\text{cov}(D_{M_j}, D_{M_l}) = 2 \times \frac{(1 - 2r_{jli}) p_{Q}^{U} (1 - p_{Q}^{U})}{2n}
\]

where \( r_{jli} \) is the known recombination rate between markers \( j \) and \( l \).

If the marker order is (M\(_j\)-M\(_l\)-Q):

\[
\text{cov}(D_{M_j}, D_{M_l}) = 2 \times \frac{(1 - 2r_{jli}) p_{M_j}^{U} (1 - p_{M_j}^{U})}{2n}
\]

For a given QTL location (\( L_Q \)) and a certain value of \( p_{Q}^{U} \), the likelihood function of the vector of observed D values of \( k \) markers (\( D_M \)), based on approximation to multivariate normality, is:

\[
L(D_M) = (2\pi)^{-k/2} |\Sigma_D|^{-1/2} \exp(D_M - \mu_D)' \Sigma_D^{-1} (D_M - \mu_D)
\]

Under the null hypothesis of no QTL, \( p_{Q}^{U} \) is \( \frac{1}{2} \) and \( L(D_M) \) is a constant \( L_0(D_M) \), which does not depend on QTL location. Under the alternative hypothesis, the likelihood function \( (L_A(D_M)) \) was maximized at a given QTL position \( L_Q \) for \( p_{Q}^{U} \) using a golden-section search method [15] and the following log likelihood ratio (LR) was calculated:
All possible QTL positions along the chromosome were evaluated and the pair of parameters \((L_Q \text{ and } p_Q^v)\) providing the highest LR gave the estimates of QTL position and QTL allele frequency, which was used to estimate the QTL effect [9].

### 2.4.2. Individual genotyping interval mapping analysis

The QTL analysis in the single grandparental family was performed using QTLexpress [29] based on the line cross least squares regression interval mapping method [14]. The analysis assumed at a chromosomal region contained one putative QTL for the trait of interest, which was fixed for alternate alleles in the two parental lines (\(Q_1Q_1\) for line 1 and \(Q_2Q_2\) for line 2). The probability to obtain none, one, or two QTL alleles from a given line for a putative QTL \((P_{11}, P_{12}, P_{21}\text{ and } P_{22})\) was calculated at each 1 cM position along the chromosome based on the marker genotype for each F\(_2\) hen and the distance between the putative QTL and markers.

Estimates of additive and dominance effects at the QTL, \(a\) and \(d\), were obtained by regressing the phenotypic value on additive and dominance coefficients derived from the QTL genotype probabilities: \(P_a = P_{11}P_{22}\text{ and } P_d = P_{12}P_{21}\). At every centi-Morgan position, the following model was fitted:

\[
y_{ijk} = \mu + f_i + h_j + aP_{ak} + dP_{bk} + e_{ijk}
\]

where \(y_{ijk}\) is the phenotypic value of the \(k^{th}\) hen in family \(i\) and hatch \(j\), \(\mu\) is the overall mean, \(f_i\) is the polygenic effect of F\(_2\) full-sib family \(i\), \(h_j\) is the effect of hatch \(j\), \(a\) and \(d\) are the to be
estimated additive and dominance effects for a given putative QTL position, $P_{ajk}$ and $P_{dk}$ are the additive and dominance coefficients for hen $k$, as described above, and $e_{ijk}$ is the residual. All terms in the model were fitted as fixed effects except for $e_{ijk}$. An F statistic was calculated at each position and the position with the highest F statistic across the chromosome was used to estimate QTL position and the parameters, $a$ and $d$. Further details are described in Haley et al. [14].

To map QTL on the Z chromosome, the method developed by Knott et al. [21] that allows for different sex recombination rates was used as described in [32]. Recombination rate was set to 0 for females and a FORTRAN program developed by Haley et al. [14] was employed instead of QTLexpress. Contrasts between hens that inherited the Z allele from line 1 versus from line 2 were estimated.

2.4.3. Significance thresholds.

To make QTL mapping results comparable, equivalent thresholds for significance were used for the analysis of pooling data analysis and individual genotyping data.

For the individual genotyping data, average 1% and 5% chromosome-wise significance thresholds across traits were derived based on 1,000 permutations [3] for each trait, for a total of 15,000 permutations. Genome-wise thresholds were derived from the chromosome-wise thresholds [6] as,

$$P_{\text{genomewise}} = 1 - (1 - P_{\text{chromosomewise}})^{1/r}$$

where $r$ is the contribution of the current chromosome to the total genome length, which was calculated based on the consensus map [13]. For the single marker pool analysis,
comparison-wise $p$ values corresponding to significance thresholds obtained for the individual genotyping data analysis were used as the significance thresholds of tests for each marker.

For LS-pool and ML-pool methods, simulation was used to obtain chromosome-wise thresholds. For two traits with different heritabilities (0.03 and 0.28, corresponding to the low and high estimates for the evaluated traits), the following model was used to generate phenotypic values with a null QTL effect:

$$y = u + QTL + \text{polygene} + \epsilon$$  \hspace{1cm} (Model 3)

where polygenic and environmental effects ($\epsilon$) were drawn from normal distributions with variances derived from the phenotypic variance in the F$_2$ population and assumed heritability. The simulated family structure, population size, selected proportion, and marker information mimicked the real data and the marker data for the two chromosomes, and technical errors drawn from a normal distribution with mean zero and variance 0.0016 added to the true marker allele frequencies. Both LS-pool and ML-pool were applied to 10,000 replicates of the simulated data and the maximum statistics, $\chi^2$ for LS-pool, $LR$ for ML-pool, across all positions of the chromosome in each replicate were sorted to obtain the 5% chromosome-wise thresholds for both traits and both chromosomes. Since thresholds were little affected by heritability, the final chromosome-wise thresholds for each chromosome were based on the average across the two simulated heritabilities.
3. RESULTS

3.1 Linkage maps.

Linkage maps for chromosomes 4 and Z are presented in Figure 1. In addition to the 24 informative markers used for selective DNA pooling analysis, 22 informative microsatellite markers were added to these two chromosomes for individual genotyping (Figure 1). Estimated map lengths were 196 cM and 115 cM for chromosomes 4 and Z, respectively. The average space between markers was less than 7 cM. Among the total 46 markers used in this analysis, 9 markers on chromosome 4 and 6 markers on chromosome Z were also found in the consensus map [13] (Figure 1). Map comparisons were made based on the shared markers. Our map showed the same order and similar marker spacing as the consensus map for both chromosomes, except for the interval between markers ROS0072 and ADL0273 on chromosome Z. Distance between these two markers in our map was 10 cM, compared to 35 cM in the consensus map.

3.2 QTL mapping analysis using individual genotyping data.

3.2.1. General QTL mapping results.

Plots of F statistics for each trait for chromosomes 4 and Z are in Figures 2 and 3, respectively. Some traits were highly correlated [39] and might be controlled by the same
QTL but will be presented here as separate QTL. Some traits showed multiple significant QTL over the chromosome, although only a one QTL model was tested (Figures 2 A and B).

A total of 14 QTL were detected at the 5% chromosome-wise significance level for the 15 traits on chromosomes 4 and Z (Table II). Among these, seven were significant at the 1% genome-wise level, another three were significant at the 1% chromosome-wise level and the rest were significant at the 5% chromosome-wise level. A 20 cM region on chromosome 4, between 165 to 185 cM, harbored QTL significant at the 1% genome-wise level for early and late body weight, and for first, early, and late egg weight. The same region also harbored QTL significant at the 1% chromosome-wise level for late puncture score and late albumen height and a QTL significant at the 5% chromosome-wise level for early production rate (Figure 2 and Table II). Another region on chromosome 4, between 74 to 88 cM, contained QTL significant at the 1% chromosome-wise level for sexual maturity and QTL significant at the 5% chromosome-wise level for first egg weight and late albumen height (Figure 2 and Table II). A region around 35 cM on chromosome Z contained QTL significant at the 1% genome-wise level for sexual maturity and early puncture score, and a QTL significant at the 5% chromosome-wise level for early body weight (Figure 3 and Table II). The F curves for several other traits also showed peaks in the above regions but they did not reach significance (Figures 2 and 3). The QTL mapping results will be described in further detail by trait in the following sections.
3.2.2. Egg weight

Highly significant QTL (F > 1% genome-wise threshold) for first egg weight and for early and late measurements of egg weight were detected around 170 cM on chromosome 4 (Figure 2 A and Table II). These QTL accounted for 12 to 17% of the F$_2$ variance. Since these traits are highly correlated, they may be controlled by the same QTL. The QTL allele from line 1 tended to be associated with heavier egg weight and the heterozygotes tended to have lower weight than the average of two homozygotes (Table II). A second significant QTL (F > 5% chromosome-wise threshold) for first egg weight was detected at 74 cM on chromosome 4 and accounted for 7% of the F$_2$ variance. The allele from line 2 for this QTL tended to be associated with heavier egg weight and the heterozygotes tended to be slightly heavier than the average of two homozygotes (Table II). No QTL was found for egg weight deviation on chromosome 4 and no QTL was found for any egg weight trait on chromosome Z (Figures 2 A and 3 A).

3.2.3. Egg quality traits

A highly significant QTL (F > 1% genome-wise threshold) for early puncture score was detected around 33 cM on chromosome Z, which accounted for 8% of the F$_2$ variance (Figure 3A and Table II). The allele from line 1 tended to be associated with higher early puncture score than the allele from line 2 (Table II). However, no QTL was significant for late puncture score on the same chromosome (Figure 3 B). A significant QTL (F > 1% chromosome-wise threshold) for late puncture score was detected around 181 cM on
chromosome 4 and accounted for 12% of F\textsubscript{2} variance (Figure 2 B and Table II). The allele from line 1 tended to be associated with greater late puncture score and heterozygotes tended to have lower puncture score than the average of the two homozygotes (Table II). No significant QTL for early puncture score was detected on chromosome 4 (Figure 2 B).

A significant QTL (F > 1% chromosome-wise threshold) for late albumen height was also detected on chromosome 4, around 174 cM (Figure 2 B and Table II). The allele from line 1 tended to be associated with larger late albumen height and heterozygotes tended to have the lowest height among the three possible genotypes (Table II). A second significant QTL (F > 5% chromosome-wise threshold) for late albumen height was detected at 88 cM on chromosome 4. Heterozygotes tended to have the highest albumen height among the three genotypes and the allele from line 1 tended to be associated with larger height (Figure 2 B and Table II). The curve of F statistics for early albumen height on chromosome 4 also showed peaks around 88 and 174 cM but they did not reach the 5% chromosome-wise threshold (Figure 2 B). No significant QTL for early or late albumen height was detected on chromosome Z. (Figure 3 B). Neither chromosome 4 nor Z showed significant QTL for early and late eggshell color (Figures 2 B and 3 B).

### 3.2.4. Body weight

Highly significant QTL (F > 1% genome-wise threshold) were detected at 165 cM on chromosome 4 for early and late body weight, which accounted for 25% of the F\textsubscript{2} variance (Figure 2 C and Table II). The allele from line 1 tended to be associated with heavier body weight and heterozygotes had slightly lighter body weight than the average of two
homozygotes (Table II). A significant QTL (F > 5% chromosome-wise threshold) for early body weight was detected at 108 cM on chromosome Z that accounted for 13% of the F\textsuperscript{2} variance (Figure 3 C and Table II). The allele from line 2 tended to be associated with heavier body weight (Table II). The F curve for late body weight on chromosome Z also showed peaks around 108 cM, but they did not reach the 5% chromosome-wise threshold (Figure 3 C).

3.2.5. Sexual maturity and egg production rate

A very significant QTL (F > 1% genome-wise threshold) for sexual maturity was detected in the region around 37 cM on chromosome Z. This QTL accounted for 22% of the F\textsuperscript{2} variance (Figure 3 D and Table II). The allele from line 1 tended to be associated with later sexual maturity (Table II). A significant QTL (F > 1% genome-wise threshold) for sexual maturity was also detected on chromosome 4, around position 78 cM, which accounted for 8% of F\textsuperscript{2} variance (Figure 2 D and Table II). For this QTL, the allele from line 1 tended to be associated with earlier sexual maturity and heterozygotes tended to be slightly earlier matured than the average of two homozygotes (Table II). Another peak was shown in the F curve for sexual maturity on chromosome 4, around position 190 cM, but it was not beyond the 5% chromosome-wise threshold (Figure 2 D).

A significant QTL (F > 5% chromosome-wise threshold) for early production rate was detected on chromosome 4, around position 177 cM, which accounted for 9% of the F\textsuperscript{2} variance (Figure 2 D and Table II). For this QTL, heterozygotes tended to have the highest early production rate and the allele from line 1 tended to be associated with lower early
production rate (Table II). Early production rate showed another peak on chromosome 4, around position 70 cM, but this was not significant at the 5% chromosome-wise level (Figure 2 D). The curve of late production rate on chromosome 4 was similar to the curve the early production rate, but not beyond the threshold (Figure 2 D). No QTL were detected significant for early or late production rate on chromosome Z (Figure 3 D).

3.3 Comparison of selective DNA pooling vs. individual genotyping

3.3.1. QTL mapping profiles.

Profiles of statistics for chromosome 4 for early egg weight, late albumen height, and late production rate are illustrated in Figures 4, 5 and 6 as representative examples. In each figure, profiles of statistics for least squares interval mapping analysis of individual genotyping data (A), and for single marker analysis (B), LS-pool (C), and ML-pool (D) analyses of selective DNA pooling data are presented. Results from individual genotyping were based on genotype information from the entire family and almost twice as many markers as used in the pooling analyses. Thus, profiles in figures A were considered as a basis for comparison to profiles from the DNA pooling analyses.

Figure 4 shows QTL mapping profiles for early egg weight on chromosome 4. Profiles from all three selective DNA pooling analyses were very similar to the profile based on individual genotyping (Figure 4 A). All profiles showed a single significant QTL around 160 cM. The profile from ML-pool was slightly sharper than the profile from LS-pool, suggesting greater accuracy of the position estimate. Figure 5 shows profiles for late albumen height on
chromosome 4. For this trait, individual genotyping analysis detected two significant QTL along the chromosome, one around 170 cM and a smaller one around 90 cM. The profiles from single marker analysis and from ML-pool also showed two peaks, but neither reached significance for single marker analysis and only the QTL around 170 cM reached significance for ML-pool. The profile from LS-pool showed significance over a wide region that extended across both regions detected with individual genotyping. Figure 6 shows profiles for the trait late production rate on chromosome 4. None of the analysis methods showed significant QTL for this trait.

3.2.2. Power to detect QTL.

Using 5% chromosome-wise thresholds, 11 QTL were detected using individual genotyping data for the 11 traits that were analyzed by both individual genotyping and selective DNA pooling methods (Table III). Of these, only 6 were detected by single marker analysis (results not shown) and 9 by both LS-pool and ML-pool (Table III). All QTL detected by selective DNA pooling were also identified in the individual genotyping analysis. Both first egg weight and late albumen height showed two significant QTL on chromosome 4 in the analysis with individual genotyping data. None of the analyses with selective DNA pooling data detected the smaller QTL for these traits (Table II), although profiles from single marker analysis and ML-pool did show peaks in the same region (Figure 5). Profiles from LS-pool showed a wide significance region and did not identify the separate QTL found by the individual analysis (Figure 5).
3.3.3. Estimates of QTL location and effect

Single marker analysis does not provide separate estimates of QTL location and effect and, therefore, comparisons of estimates of QTL location and effect were only between the analysis of individual genotyping data and LS-pool and ML-pool analysis of selective DNA pooling data. Only estimates of the nine QTL found by both individual and pooling data were compared and are summarized in Table III. All estimates of QTL effect were standardized by dividing by the phenotypic standard deviation of the trait.

Among the nine QTL, LS-pool located five QTL within a 5 cM region from estimates based on individual genotyping, two within 6-10 cM and two at more than 10 cM (Table III). The ML-pool method located seven QTL to within a 5 cM, one within 6-10 cM, and one more than 10 cM away (Table III). Correlations of estimates of QTL location from individual genotyping with estimates from LS-pool and ML-pool were 0.99.

Estimates of additive QTL effects from LS-pool and ML-pool were nearly the same as estimates from individual genotyping (Table III). Linear regression of standardized estimates of the additive effects of the nine QTL from individual genotyping on estimates from LS-pool and ML-pool resulted in near zero intercepts (0.01 and -0.037), slopes near one (1.086 and 0.97), and $R^2$'s near one (0.98 and 0.97).
4. DISCUSSION

4.1. Linkage maps

Linkage maps derived for chromosomes 4 and Z were consistent with the published consensus map, except a much shorter distance was estimated between markers ROS0072 and ADL0273 on chromosome Z (10 cM in our map versus 35 cM in the consensus map). However, the distance of this interval in the study by Tuiskula-Haavisto et al. [32] estimated an even shorter distance between these two markers (less than 2 cM). Reasons for such big differences between maps are unclear.

4.2. QTL detected

In this study, a region on chromosome 4 around 170 cM was found to harbor QTL significant at the 1% genome-wise level for early and late body weight, first egg weight, and early and late egg weight, 1% chromosome-wise significant QTL for late puncture score and late albumen height, and 5% chromosome-wise significant QTL for early production rate. A secondary region of interest on chromosome 4 was around 80 cM, which had QTL significant at the 1% chromosome-wise significant for sexual maturity, at the 5% chromosome-wise level for first egg weight and late albumen height, and suggestive QTL (showed peaks but not beyond the significant threshold) for early and late production rate, and for early albumen height. A region on chromosome Z around 35 cM showed QTL significant at the 1% genome-wise level for sexual maturity and early puncture score, and
another region around 108 cM was significant at the 5% chromosome-wise level for early body weight and suggestive for late body weight.

Our results were highly consistent with those of Tuiskulat-Haavisto et al. [32]. They also found that the same region on chromosome 4 contained QTL significant at the genome-wise level for body weight and egg weight and at the chromosome-wise level for albumen height and for number of eggs at 18-40 wk, which is correlated with early egg production rate. Wardecka et al. [38] also found significant QTL on chromosome 4 for body weight and albumen height, but QTL positions were difficult to compare since different markers were used. They also found that chromosome 4 was significant for eggshell color, which was not confirmed in our study. Significant QTL on chromosome 4 have also been identified for body weight in broilers in several studies [7, 30, 36]. Tuiskulat-Haavisto et al. [32] also found genome-wise significant QTL for eggshell strength and age at first egg on chromosome Z, but in a different region than what was detected in our study. Some DNA fingerprint markers also showed association with traits of interest in this study [19, 22] but results are not comparable with our research because no location information is available for DNA fingerprint markers.

Alleles from line 1 of the QTL around 170 cM on chromosome 4 were associated with heavier egg weight, stronger eggshell and thicker egg white, while alleles from line 2 for the same QTL were associated with lighter body weight and higher egg production rate. Alleles from line 1 of the QTL around 35 cM on chromosome Z were associated with stronger eggshell, while alleles from line 2 of the same QTL were associated with earlier sexual maturity. These differences between lines are mainly due to the different genetic
backgrounds of two lines instead of different selection directions, because similar egg quality and production traits were intensively selected in both lines.

Results from our study will be very useful to apply MAS in the layer chicken breeding industry. Strong QTL for body weight, egg weight, puncture score, albumen height and production rate were located in the same region of chromosome 4, and explained 25%, 12-17%, 12%, 10% and 9% variance in the F$_2$ population, respectively. Strong QTL for sexual maturity and puncture score were located in the same region of chromosome Z and explained 22% and 8.4% of the F$_2$ variance. Associated marker genotypes or haplotypes can be selected for in both male and female chicks, from the first day of life. However, some QTL showed antagonistic effects. For example, a heavier body weight will accompany the better egg quality and heavier egg weight if the allele from line 1 for the strong QTL on chromosome 4 is selected. These results might be explained by pleiotropic effects or tight linkage [7] and make selection difficult. In addition, most significant QTL showed different degrees of dominance and sometimes over dominance (Table II, early production rate and late albumen height). This may open opportunities for employing crossbreeding strategies.

Results from our study are also useful for candidate gene analysis because very strong QTL controlling multiple interested traits were located in relatively narrow chromosomal regions. Noakes et al. [28] mapped the CLOCK gene, affecting the circadian rhythm, to the region of interest on chromosome 4. It is logical that the circadian rhythm directly or indirectly affects egg production traits, however, it is not clear how this gene affects egg quality traits. The growth hormone receptor gene has been mapped to chromosome Z and has been found to be associated with body weight and egg production [10, 11]. This gene, however, maps to a different region than the region that was significant for body weight in
our study. The feathering $K$ gene has also been reported to affect growth and egg production [2] but in our study, this gene was located about 14 cM away from the QTL that was significant for sexual maturity (Figures 1 and 3).

Some studies used BLUP EBV instead of phenotypic values for QTL mapping analysis [7, 34-36]. The BLUP EBV combines information from pedigree, the animal itself and progeny and removes systematic environmental effects and can, therefore, provide more accurate QTL estimates. However, in the current study, no progeny information was available and the only environmental factor was hatch. Therefore, using phenotypes of $F_2$ individuals and including families as fixed effects in the model was equivalent to using BLUP EBV in the current study because pedigree information is the same for full-sibs. Analysis of BLUP EBV of $F_2$ individuals in the individual genotyping analysis (results not shown), resulted in identical values of F statistics and QTL mapping profiles as using phenotypic data but in smaller estimates of QTL effects. The QTL estimates obtained from an analysis of BLUP EBV are regressed toward zero, as has been demonstrated by several [17]. The magnitude of regression mainly depended on heritability of the trait (results not shown).

**4.3. Evaluation of QTL mapping methods using selective DNA pooling data**

Both interval mapping methods (LS-pool and ML-pool) provided greater power to detect the QTL than single marker analysis, and accurate estimates of QTL location and effect (Table III and Figure 7). The number of QTL detected and estimates of QTL effects were similar for the LS-pool and ML-pool methods (Table III), but ML-pool appeared to provide
sharper mapping profiles than LS-pool (Figure 4 and 5), which might imply a smaller confidence interval of estimates of QTL location with ML-pool. The smaller confidence interval for QTL location estimates with ML-pool might result from the fact that ML-pool accounted for correlations between sampling errors in allele frequencies of linked markers. The profiles of ML-pool also reflected multiple QTL, while the profile of LS-pool showed a wide and flat region of significance (e.g. Figures 5 C and D).

In this study, individual genotyping data included genotype information on 46 markers (33 on chromosome 4 and 13 on chromosome Z) for each hen in the F2 family (189 to 191 individuals for early measurements and 158 to 163 individuals for late measurements). Selective DNA pooling data included pool information for only 24 markers (15 on chromosome 4 and 9 on chromosome Z) from pools of the 20% top and 20% low individuals for each trait. Consequently, the total number of individual genotypes was 46 (markers) * 158 to 191 (individuals) = 7268 to 8786, while the number of PCR reactions conducted for selective DNA pooling for the 11 traits was 2 (selected pools) * 24 (markers) * 11 (traits) = 528. Therefore, in this particular example, using selective DNA pooling reduced genotyping costs over 12 fold. Even if the same number of markers were used in both individual genotyping and selective DNA pooling approaches, the genotyping cost would still be reduced over 6 fold. Even greater efficiencies can be obtained with larger size families because the number of PCR reactions required for the pooling strategy does not depend on the number of individuals in each pool. In addition, simulation studies by Wang [37] indicated that larger size families also increased the accuracy of QTL estimates with selective DNA pooling. On the other hand, the efficiency of DNA pooling will decrease if multiple traits are considered because pools must be created separately for each trait. In the current
study. 11 traits were measured but the savings in genotyping cost were still substantial. This indicates that the selective DNA pooling strategy can be applied to research investigating multiple traits with considerable reductions in genotyping costs.

In this study, significance thresholds for selective DNA pooling analyses were derived by simulation. The limited number of observations prevented the application of permutation or bootstrap methods to derive thresholds directly from the data. In addition, observations were not independent due to linkage. Use of simulation may result in erroneous thresholds if the model is inappropriate or the normality assumption is not valid. False positives or negatives could result. However, in this particular study, no false positive was observed compared to results from individual genotyping and the proportion of false negatives compared to individual genotyping was small (Table III).

Although both LS-pool and ML-pool required the assumption of co-dominant QTL in the F2 population for a symmetric distribution of phenotypic data in relation to marker and QTL genotypes, results indicated that estimates of additive QTL effects from LS- and ML-pool were accurate for QTL with substantial or even over dominance (Table III). This suggests that the effect of dominance on marker allele frequencies in the selected pools was limited, which was also indicated by the simulation study of Wang [37] and by Darvasi and Soller [4].

5. CONCLUSIONS

With interval mapping analysis of individual genotyping data in an F2 population derived by crossing two partially inbred commercial white leghorn lines, a region on chromosome 4
around 170 cM was found to harbor significant QTL for early and late body weight, first egg weight, early and late egg weight, late puncture score, late albumen height and early production rate, and a region on chromosome Z around 35 cM was significant for sexual maturity and early puncture score. Both LS-pool and ML-pool provided similar power for QTL detection and estimates of QTL location and QTL effect as results from interval mapping analysis of individual genotyping data but substantially reduced the genotyping cost. Both LS-pool and ML-pool provided economical methods for capturing most of the information for QTL detection.
REFERENCES


[38] Wardecka B., Olszewski R., Jaszczak K., Zieba G., Pierzchala M., Wicinska K., Relationship between microsatellite marker alleles on chromosomes 1-5 originating
from the Rhode Island Red and Green-legged Partrigenous breeds and egg production and quality traits in F(2) mapping population, J Appl Genet 43 (2002) 319-29.


Table I. Means, standard deviations and number of hens evaluated for traits measured in the \( F_2 \) family with individual genotyping data

<table>
<thead>
<tr>
<th>Trait</th>
<th>Unit</th>
<th>Description</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>First egg weight</td>
<td>g</td>
<td>Mean of first 3 eggs</td>
<td>40.4</td>
<td>3.7</td>
<td>189</td>
</tr>
<tr>
<td>Egg weight</td>
<td>g</td>
<td>Mean of 5 eggs</td>
<td>52.6</td>
<td>3.8</td>
<td>191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>early*</td>
<td>62.0</td>
<td>4.7</td>
<td>164</td>
</tr>
<tr>
<td>Egg deviation</td>
<td>g</td>
<td>Late minus early egg weight</td>
<td>9.4</td>
<td>3.9</td>
<td>164</td>
</tr>
<tr>
<td>Eggshell color</td>
<td>lab</td>
<td>Measured by Minolta</td>
<td>early</td>
<td>8.6</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>late</td>
<td>9.5</td>
<td>1.9</td>
<td>164</td>
</tr>
<tr>
<td>Puncture score</td>
<td>kg</td>
<td>Pin point pressure to the eggshell for puncture, mean of 5 eggs</td>
<td>early</td>
<td>1700</td>
<td>87.0</td>
</tr>
<tr>
<td>Albumen height</td>
<td>mm</td>
<td>Height of thick albumen</td>
<td>early</td>
<td>6.9</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>late</td>
<td>5.4</td>
<td>1.0</td>
<td>164</td>
</tr>
<tr>
<td>Body weight</td>
<td>kg</td>
<td>early</td>
<td>1.5</td>
<td>0.2</td>
<td>191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>late</td>
<td>1.7</td>
<td>0.3</td>
<td>164</td>
</tr>
<tr>
<td>Sexual maturity</td>
<td>day</td>
<td>Age at first egg</td>
<td>143</td>
<td>10.7</td>
<td>191</td>
</tr>
<tr>
<td>Egg production rate</td>
<td>%</td>
<td>(Number of eggs)/(number of production days)</td>
<td>early</td>
<td>77.3</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>late</td>
<td>76.1</td>
<td>16.0</td>
<td>158</td>
</tr>
</tbody>
</table>

* Early measurement was at 28 week of age and the late measurement was at 42 or 43 week of age.
Table II. QTL significant at the 5% chromosome-wise level for different egg quality and production traits by chromosome using the individual genotyping approach.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Trait</th>
<th>F value</th>
<th>Location (cM)</th>
<th>Additive effect</th>
<th>Dominance effect</th>
<th>% var</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>First egg wt. (g)</td>
<td>15.15++</td>
<td>173</td>
<td>1.75</td>
<td>-1.10</td>
<td>13.6</td>
</tr>
<tr>
<td></td>
<td>First egg wt. (g)</td>
<td>6.74*</td>
<td>74</td>
<td>-1.42</td>
<td>0.23</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>Early egg wt. (g)</td>
<td>20.41++</td>
<td>161</td>
<td>2.18</td>
<td>-0.63</td>
<td>17.1</td>
</tr>
<tr>
<td></td>
<td>Late egg wt. (g)</td>
<td>12.48++</td>
<td>175</td>
<td>2.18</td>
<td>-1.26</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td>Late puncture score</td>
<td>7.97**</td>
<td>181</td>
<td>42.66</td>
<td>-31.1</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>Late albumen ht.</td>
<td>8.04**</td>
<td>174</td>
<td>0.35</td>
<td>-0.40</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>Late albumen ht.</td>
<td>6.31*</td>
<td>88</td>
<td>0.27</td>
<td>0.44</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>Early body wt. (Kg)</td>
<td>38.81**</td>
<td>165</td>
<td>0.144</td>
<td>-0.08</td>
<td>25.5</td>
</tr>
<tr>
<td></td>
<td>Late body wt. (Kg)</td>
<td>30.89++</td>
<td>165</td>
<td>0.18</td>
<td>-0.11</td>
<td>24.7</td>
</tr>
<tr>
<td></td>
<td>Sexual maturity (day)</td>
<td>7.68**</td>
<td>78</td>
<td>-4.15</td>
<td>1.54</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td>Early prod. rate (%)</td>
<td>6.96*</td>
<td>177</td>
<td>-4.21</td>
<td>4.85</td>
<td>9.0</td>
</tr>
<tr>
<td>Z</td>
<td>Early puncture score</td>
<td>18.33++</td>
<td>33</td>
<td>35.68</td>
<td>-</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>Sexual maturity (day)</td>
<td>26.15++</td>
<td>37</td>
<td>7.10</td>
<td>-</td>
<td>22.0</td>
</tr>
<tr>
<td></td>
<td>early body wt. (Kg)</td>
<td>10.15*</td>
<td>108</td>
<td>-0.11</td>
<td>-</td>
<td>12.9</td>
</tr>
</tbody>
</table>

a. ++ significant at 1% genome-wide level (F>10.92 for chromosome 4 and F>18.05 for chromosome Z)
   + significant at 5% genome-wise level (F>8.84 for chromosome 4 and F>14.01 for chromosome Z)
   ** significant at 1% chromosome-wise level (F>7.48 for chromosome 4 and F>10.66 for chromosome Z)
   * significant at 5% chromosome-wise level (F>5.71 for chromosome 4 and F>7.07 for chromosome Z).

b. For chromosome 4, both additive (a) and dominance (d) QTL effects can be estimated. Estimates of +a, d and -a correspond to genotype values for, respectively, individuals homozygous for line 1 alleles, heterozygotes and individuals homozygous for line 2 alleles. For chromosome Z, only an additive effect can be estimated, which corresponds to the difference between line 1 and line 2 alleles. A positive additive effect indicates the allele from line 1 increases the trait.

c. % var. is the estimated genetic variance for the QTL (additive plus dominance variance for chromosome 4 and additive variance for chromosome Z, as a percentage of the total phenotypic variance in the F2 population.
Table III. QTL significant at the 5% chromosome-wise level for individual genotyping analysis and LS-pool and ML-pool analysis of selective genotyping data from chromosomes 4 and Z, along with their estimates of location and additive effect (in units of $\sigma_p$).

<table>
<thead>
<tr>
<th>Chr</th>
<th>Trait</th>
<th>Individual genotyping</th>
<th>Selective DNA pooling</th>
<th>LS-pool$^a$</th>
<th>ML-pool$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>loc.</td>
<td>effect</td>
<td>loc.</td>
<td>effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>((\sigma_p))$^b$</td>
<td>((\sigma_p))$^b$</td>
<td>loc.</td>
<td>effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>loc.</td>
<td>effect</td>
</tr>
<tr>
<td>4</td>
<td>First egg wt. (g)</td>
<td>173</td>
<td>0.48**</td>
<td>170</td>
<td>0.45**</td>
</tr>
<tr>
<td></td>
<td>First egg wt. (g)</td>
<td>74</td>
<td>-0.39*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Early egg wt. (g)</td>
<td>161</td>
<td>0.57**</td>
<td>172</td>
<td>0.43**</td>
</tr>
<tr>
<td></td>
<td>Late egg wt. (g)</td>
<td>175</td>
<td>0.46**</td>
<td>175</td>
<td>0.39**</td>
</tr>
<tr>
<td></td>
<td>Late puncture score (Kg)</td>
<td>181</td>
<td>0.43**</td>
<td>181</td>
<td>0.40**</td>
</tr>
<tr>
<td></td>
<td>Late albumen ht. (mm)</td>
<td>174</td>
<td>0.34**</td>
<td>174</td>
<td>0.37**</td>
</tr>
<tr>
<td></td>
<td>Late albumen ht. (mm)</td>
<td>88</td>
<td>0.26*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Sexual maturity (day)</td>
<td>78</td>
<td>-0.39**</td>
<td>100</td>
<td>-0.36**</td>
</tr>
<tr>
<td></td>
<td>Early prod. rate (%)</td>
<td>177</td>
<td>-0.33*</td>
<td>168</td>
<td>-0.33*</td>
</tr>
<tr>
<td>Z</td>
<td>Early puncture score (Kg)</td>
<td>33</td>
<td>0.41**</td>
<td>43</td>
<td>0.44**</td>
</tr>
<tr>
<td></td>
<td>Sexual maturity (day)</td>
<td>37</td>
<td>0.66**</td>
<td>33</td>
<td>0.57**</td>
</tr>
</tbody>
</table>

a. Only additive effects can be estimated by LS-pool and ML-pool.
b. ++ significant at 1% genome-wise level, + significant at 5% genome-wise level, ** significant at 1% chromosome-wise level, * significant at 5% chromosome-wise level.
c. The 5% and 1% chromosome-wise thresholds for LS-pool for chromosome 4 were 37.25 and 57.52, thresholds for chromosome Z were 39.68 and 64.84. The 5% and 1% chromosome-wise threshold for ML-pool for chromosome 4 were 2.175 and 3.09, thresholds for chromosome Z were 3.54 and 5.39. Thresholds for individual genotyping analysis were given in Table II. Significance at the genome-wise level was not tested for LS-pool and ML-pool.
Figure 1. Marker linkage maps for chromosomes 4 and Z for markers used in the individual genotyping analysis. Distance is in cM, relative to the position of the first marker on each chromosome. Markers with • are also shown in the 2000 consensus map [13]. Markers with ↔ were also used in the selective DNA pooling analysis.
Figure 2. F curves for evidence of QTL on chromosome 4 based on individual genotyping analysis. Horizontal lines represent the 5% chromosome-wise threshold (-----), the 1% chromosome-wise threshold (-----), the 5% genome-wise threshold (-----) and the 1% genome-wise threshold (-----).
Figure 3. F curves for evidence of QTL on chromosome Z based on individual genotyping analysis. Horizontal lines represent the 5% chromosome-wise threshold (-----), the 1% chromosome-wise threshold (--------), the 5% genome-wise threshold (-----) and the 1% genome-wise threshold (--------).
Figure 4. QTL mapping profiles for early egg weight on chromosome 4 using individual genotyping (A), and single marker (B), LS-pool (C), and ML-pool (D) analyses of selective DNA pooling data. The broken line indicates the 5% chromosome-wise significance level.
**Figure 5.** QTL mapping profiles for late albumen height on chromosome 4 using individual genotyping (A), and single marker (B), LS-pool (C), and ML-pool (D) analyses of selective DNA pooling data. The broken line indicates the 5% chromosome-wise significance level.
Figure 6. QTL mapping profiles for late production rate on chromosome 4 using individual genotyping (A), and single marker (B), LS-pool (C), and ML-pool (D) analyses of selective DNA pooling data. The broken line indicates the 5% chromosome-wise significance level.
Figure 7. Estimates of QTL effects from individual genotyping for significant QTL regressed on estimates of QTL effects from LS-pool and ML-pool. Estimates were standardized by phenotypic standard deviations.
CHAPTER 6. GENERAL DISCUSSION

In this thesis, two interval mapping methods for analysis of selective DNA pooling data were developed. One was based on least squares regression (LS-pool) and the other on approximate maximum likelihood (ML-pool). Both methods simultaneously utilize information from multiple markers and multiple families and are easily applied to different family structures (half-sib, F2 cross, backcross).

Using simulation at several combinations of parameters, QTL mapping results from these two DNA pooling interval mapping methods were compared with results from the currently used single marker analysis and with selective genotyping analysis of individual genotypes. Results indicated that both LS-pool and ML-pool provided greater power to detect the QTL than the single marker analysis and separate estimates of QTL location and effect were obtained, which is not possible with single marker analysis. With large family sizes (2,000 progeny in our study), both LS-pool and ML-pool provided similar power and estimates of QTL location and effect as selective genotyping. The ML-pool method provided slightly greater power to detect the QTL and less biased estimates of QTL location than the LS-pool method. This could be explained by the fact that the ML-pool method takes account of correlations between allele frequencies of linked markers, which is ignored in the LS-pool method. Effects of different parameters on QTL mapping results of the two DNA pooling interval mapping methods were also investigated. Based on the ranges of parameters evaluated, results indicated that family size had a large effect on power and on precision and
accuracy of QTL mapping results. The true position of the QTL and prior knowledge of the magnitude of technical error had moderate effects on QTL mapping results and the magnitude of the variance of technical errors had relatively small effects.

Both DNA pooling interval mapping methods were also evaluated by analysis of two real data sets, a dairy cattle data set from a half-sib design and a layer chicken data set from an F₂ cross. In the dairy cattle data application, both LS-pool and ML-pool solved the problem of single marker analysis when data is missing for some markers and sires by utilizing joint information from multiple markers. In the chicken data application, both LS-pool and ML-pool provided similar power to detect the QTL and similar estimates of QTL location and effect as interval mapping analysis of individual genotypes.

In conclusion, both LS-pool and ML-pool methods provide powerful tests for QTL detection and accurate estimates of QTL parameters, while substantially saving genotyping costs through the use of DNA pooling. In addition, both methods can be readily applied to practical situations.

However, there are some limitations to the LS-pool and ML-pool methods. First, both methods are based on the assumption that only one QTL is segregating on the chromosome, along polygenes affecting the trait. Further research is needed to extend this assumption to multiple QTL. Second, both methods require knowledge about marker haplotypes of parents, which is usually unknown with pooling data, and requires extra costs to determine. Third, there is no good way to provide chromosome-wise significance thresholds for these DNA
pooling interval mapping methods. Finally, a toward-center bias was observed in estimates of QTL location for LS-pool with small family size and distal QTL. Biases were in part caused by a heterogeneous distribution of the test statistic across the chromosome. Although several methods were developed and evaluated in this thesis, in an attempt to reduce the bias, they were not completely successful or caused other problems while correcting the bias. Further research is needed to develop better correction methods or alternative analysis methods. Simulated adjustment of p-values, for example, may help to resolve some of these problems. In addition, the ML-pool method was based on an approximate normal distribution for frequency data and the distribution of technical errors was simplified by assuming independence and normality. Further research is needed to develop more accurate models for smaller family sizes.
ACKNOWLEDGEMENTS

This thesis is the result of my research efforts whereby I have been accompanied and supported by many people. I would like to take the opportunity to express my gratitude to all of them.

I would like to express my sincere gratitude to my co-major professors, Dr. Jack Dekkers and Dr. Ken Koehler for their scientific guidance, continued encouragement, wisdom for life, and great patience during my academic training at ISU. What I have learned from you will benefit whole my life. I am looking forward to learning even more from you in the future.

I also want to extend my thanks to my committee member Dr. Rohan Fernando, Dr. Max Rothschild, Dr. Susan Lamont, Dr. Dan Nettleton and Dr. Jean-Luc Jannink for their research and study guidance.

Special thanks are given to Dr. Janet Fulton, Dr. Petek Settar, Dr. Jim Arthur and Dr. Neil O'Sullivan from Hy-Line International, for giving me the opportunity to work and interact with industry and helpful discussions on my research.

Special thanks are also given to Dr. Morris Soller and Dr. Ehud Lipkin from Hebrew University of Jerusalem for research guidance and wonderful discussions.

My thanks also go to all my fellow graduate students, all faculty and staff in both the Animal Science the Department and Statistics Department for their great help, valuable discussions and encouragement.

Finally, I would give very special thanks to my family. Thanks for their understanding, sacrifice, and everlasting love and support that made completion of this work possible.