ACTA UNIVERSITATIS AGRICULTURAE SUECIAE

VETERINARIA 121

New Methods for Mapping Quantitative Trait Loci

Örjan Carlborg

SWEDISH UNIVERSITY OF AGRICULTURAL SCIENCES

New Methods for Mapping Quantitative Trait Loci

Örjan Carlborg
Department of Animal Breeding and Genetics
Uppsala

Doctoral thesis
Swedish University of Agricultural Sciences
Uppsala 2002

Abstract

Carlborg, Ö. 2002. *New Methods for Mapping Quantitative Trait Loci*.
Doctoral dissertation.
ISSN 1401-6257, ISBN 91-576-6362-9

This thesis presents and discusses the use of various genetic models, high performance computing, global optimization algorithms and statistical methods for mapping Quantitative Trait Loci (QTL). The aim of the work has been to develop statistically powerful and computationally efficient methods to detect genomic loci affecting multifactorial traits, and use the methods use to analyse experimental data.

Imprinting is an epigenetic phenomena which causes differential expression of alleles based on their parental origin. A genetic model handling imprinting was used during QTL mapping in an experimental Wild Boar x Large White intercross. The analyses revealed a paternally imprinted QTL with large effect on the development of muscle mass.

Parallel computing algorithms for interval mapping and randomization testing in QTL mapping are described. New randomization testing schemes are now computationally feasible due to these algorithms. Selection of appropriate kernel algorithms for solving least squares type problems in QTL mapping is discussed. The importance of optimization of QTL mapping software is also illustrated.

A genetic algorithm was shown to be efficient in a multidimensional search for interacting QTL. The genetic algorithm significantly decreases the computational demand when employing simultaneous mapping of multiple QTL, and makes randomization testing based on multidimensional searches computationally feasible. A new randomization testing scheme based on simultaneous mapping of epistatic QTL was also proposed and evaluated. A simulation study showed that the method increases the power to detect epistatic QTL.

A large intercross was derived between Red junglefowl and White Leghorn chickens. A number of QTL affecting growth was revealed using the newly developed method for simultaneous mapping of epistatic QTL pairs. In total, 21 QTL were identified, and eleven of these were only detected by the new simultaneous mapping method. Epistasis was shown to be an important component in the genetic regulation of the growth process.

Keywords: QTL, imprinting, epistasis, high performance computing, randomization testing, genetic algorithm

Author's address: Örjan Carlborg, Department of Animal Breeding and Genetics, SLU. BMC, P.O. Box 597. S-751 24 Uppsala. Email: Orjan.Carlborg@hgen.slu.se.

To my family

Contents

Appendix

Papers I-V

The present thesis is based on the following papers, which will be referred to by their Roman numerals:

I. Jeon J-T., Carlborg Ö, Törnsten A, Giuffra E, Amarger V, Chardon P, Andersson-Eklund L, Andersson K, Hansson I, Lundström K & Andersson L. 1999. A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. Nature Genetics 21, 157-158.

II. Carlborg, Ö., Andersson-Eklund, L., & Andersson, L. 2001. Parallel computing in interval mapping of quantitative trait loci. Journal of Heredity 92, 449-451

III. Carlborg, Ö., Andersson, L. & Kinghorn, B. 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. Genetics 155, 2003-2010.

IV. Carlborg, Ö & Andersson, L. 2002. The use of randomization testing for detection of multiple epistatic QTL. Genetical Research. In Press (Cambridge University Press).

V. Carlborg, Ö., Kerje, S., Schütz, K., Jacobson, L., Jensen, P. & Andersson, L. Detection of single QTL and simultaneously mapped epistatic QTL pairs explaining a large proportion of the difference in growth between the Red junglefowl and White Leghorn chickens. (Submitted)

Paper I-IV is reproduced with the permission of the journals concerned

Introduction

Most traits in plants and animals are affected by both genetic and environmental factors. Differences between individuals for these traits are of degree rather than kind, quantitative rather than qualitative. Quantitative genetics is concerned with the inheritance of these traits. Until recently, this field was mainly focused on studying the aggregate effects of all the genes causing variation. This approach has given estimates of the genetic contribution to the observed phenotypic variation (heritability), as well as knowledge about the genetic correlation between various traits. The knowledge has been successfully used in animal breeding to increase the production of e.g. meat, milk and eggs from our domestic animals.

In some studies, individual genes with direct and measurable effect on quantitative traits (so called major genes) have been detected. A handful of such genes exist, including the Boorola gene (Davis *et al.* 1982), which raises litter size in sheep, and the double muscling gene in cattle, which increases lean meat yield (Grobet *et al.* 1997). The majority of the genes affecting quantitative traits do not have directly measurable effect on the traits and can thus not be detected by segregation analysis. Due to advances in molecular genetic and statistical methodology, it has become possible to map individual genetic factors with smaller effects on the quantitative traits, known as Quantitative Trait Loci or QTL, to specific chromosomal segments in the genome. The methods are also used to infer the mode of inheritance, which gives a better understanding of the genetics underlying quantitative traits. One of the applications for this knowledge is Marker Assisted Selection (MAS), where knowledge about the QTL genotype can help animal breeders to further increase the genetic progress of the domestic animals.

Genetic dissection of quantitative traits

This section briefly introduces the concept of QTL mapping. Further reviews on the topic are given by Darvasi (1998), Broman (2001), Cardon and Bell (2001), Flint and Mott (2001) and Doerge (2002).

Genetic markers and genetic maps

The localization of genes affecting phenotypes to individual chromosome segments requires detection of co-segregation of the phenotype with genetic markers on the chromosome. The genetic markers are chromosomal loci where the genotype can be identified by looking e.g. at the phenotype of the individuals (e.g. coat colour), protein polymorphisms or directly as sequence differences in the DNA molecule (e.g. microsatellites or single nucleotide polymorphisms (SNPs)). Automation has dramatically decreased the time and costs for scoring of genetic markers. This has made it possible to perform studies including typing of hundreds of genetic markers on hundreds of individuals.

A genetic map describes the chromosomal location and the relative order of known markers on each of the chromosome in the genome. Rapid advances in molecular genetics have led to the development of dense genetic maps of linked, polymorphic markers for many species. Detection and localisation of QTL on the genetic map is based on co-segregation between alleles at marker loci and alleles at the QTL. The genetic maps have been used in many gene and QTL mapping studies, which have identified and localised a large number of QTL for various traits (reviewed by Andersson 2001).

Experimental crosses used for QTL detection

When the aim of a QTL mapping study is to identify loci for a particular trait or group of traits it is possible to create a mapping population, which maximizes the chance to have such genes segregating. It is more likely that a given QTL show segregation in a cross between two phenotypically divergent lines than within a population, which has been under strong directional selection.

Crosses between inbred lines are highly efficient for detecting QTL. The crossed lines have a high degree of homozygosity at marker loci and QTL, and their resulting offspring will have high linkage disequilibrium between alleles of all linked loci. Crosses between outbred lines are common in species where inbred lines do not exist, e.g. farm animals. The major disadvantage with this type of cross is that the degree of homozygosity at marker loci is lower than in inbred lines and that it is unknown for the QTL. Since the degree of homozygosity at the QTL is unknown in the divergent breeds in the cross, the parental lines are usually assumed fixed for alternative QTL alleles. If this is not the case in reality, there is a confounding between the allele frequency and the effect of the QTL, which decreases the power of QTL mapping.

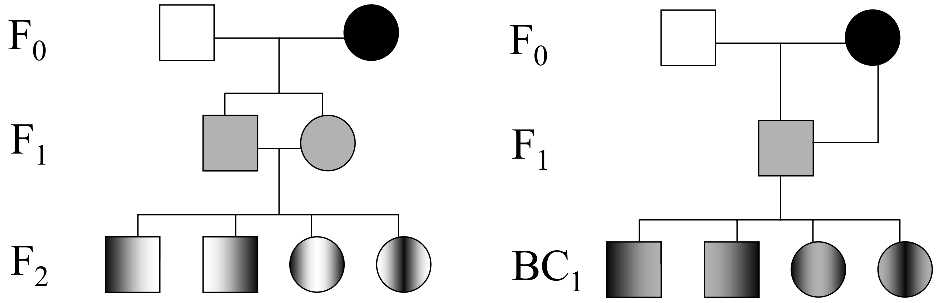

Figure 1. Mating scheme for experimental backcross and F_2 populations used for mapping of QTL.

Several types of populations can be derived from a cross between divergent lines, including F_2, single- or double- backcross and recombinant inbred. Crossing schemes for the single backcross and the F_2 population types are presented in Figure 1. An F_2 is more powerful than either individual backcross for detecting QTL of additive effect, and can also be used to estimate the degree of dominance for detected QTL. In general, several traits are considered in each study and the level and direction of dominance will depend upon the trait. The F_2 or a combination of the two backcrosses may be optimal both in terms of overall power and the ability to estimate the effects of detected QTL.

The F_2 or a combination of the two backcrosses can be used to detect four types of interaction between two loci: additive by additive, additive by dominance, dominance by additive and dominance by dominance. The single backcross can only be used to detect the additive by additive interaction effect. The F_2 thus makes a more thorough investigation of epistasis possible, but a larger population size is needed to obtain the same power to detect epistasis. Varona *et al.* (2001) has used simulations to evaluate the power to detect epistasis in outbred F_2 line crosses. Their studies indicate that the power to detect an interaction effect of size 1-5% of the phenotypic variance ranges from 50 to 80% in populations of 200-400 individuals.

Genetic modeling of QTL

Various ways exist to model the genetic effects of QTL. By estimating the effects of each marker or putative QTL genotype (when interval mapping is used) separately, no genetic model is assumed for the QTL. This analysis can be performed e.g. by regression of the phenotypes of the individuals in a population either on their marker genotypes or on putative QTL genotypes between markers. Simultaneous analysis of multiple loci increases the number of parameters in the model to 3^n (or 4^n if accounting for parental origin), where n is the number of evaluated loci. A disadvantage with this method is that the power of the test decreases for multiple QTL models due to the large number of included parameters.

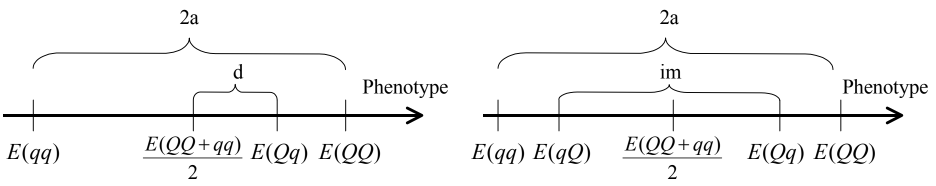

Figure 2. Definition of the additive (*a*), dominance (*d*) and imprinting (*im*) genetic effects used for mapping of QTL.

A QTL can be modelled by an allele substitution effect, which is commonly called an additive effect (Figure 2). The additive model assumes a linear relationship for the three QTL genotype classes, where the heterozygote individuals (*Qq*) have an intermediate phenotype to the homozygotes (QQ and qq). The additive effect is then the effect obtained by replacing the low effect allele (q) for the high effect allele (Q). Sometimes the additive effect is also expressed as the positive and negative deviation of the respective homozygotes from the mean of both homozygotes. In many cases the heterozygotes phenotypes deviates from the mean of the two homozygotes. The situation, where the heterozygote phenotype is closer to either one of the homozygotes, is called dominance, and the allele that mainly influences the phenotype is called dominant. A dominance effect can be modelled by including a parameter for the deviation of the heterozygous phenotype from the mean of the two homozygotes (Figure 2). Both the additive and dominance effects are commonly used in QTL mapping. For some loci in mammals, only one of the two alleles is expressed. The expression is determined by the parental origin of the allele and this phenomenon is known as genetic imprinting. Imprinting can be modelled by treating the heterozygotes obtaining maternal and paternal alleles separately. Further details on imprinting models in QTL mapping will be given later in this text. Genetic models including imprinting have recently been used for QTL mapping (Knott *et al.* 1998; Nezer *et al.* 1999; Jeon *et al.* 1999; de Koning *et al.* 2000; Rattink *et al.* 2000). The additive, dominance and imprinting effects will during the rest of this text be referred to as marginal genetic effects, since these effects are only dependent on the actions of the individual locus and can be detected by considering each locus separately.

Quantitative traits are by definition affected by the products from genes at multiple loci. It is believed that many of these gene products interact to larger or smaller extent. In the reminder of this text, we will talk about interactions between genes, epistasis, but the biological interpretation of this should be that the products from these genes interact. Genetic effects due to multiple loci have been modeled as independent genetic factors (using the marginal effects for each locus) or treated as polygenes. Interactions between QTL are in most cases ignored, but a number of QTL mapping studies in various species and for various traits have shown that interactions are important (e.g. Fijneman *et al.* 1996;Long *et al.* 1996; Li *et al.* 1997;Shook and Johnson 1999). Epistasis can exist between *2* to *n* loci and the higher the order of the interaction, the more difficult it is to model. Here we will limit ourselves to interactions between pairs of loci. When considering this two-way epistasis, a model needs to consider the marginal effects of both loci in the same way as for a marginal effects model, but also the four possible pairwise interactions between the loci, homozygous-homozygous, homozygous-heterozygous, hererozygous-homozygous and heterozygous-heterozygous. Most models do this by including the parameters *aa*, *ad*, *da* and *dd*, which are biometrical quantities specifying these interactions.

Two alternative ways of partitioning the genetic effects of individuals using these models are given by Cockerham (1954) and Seyffert (1966). The model described by Cockerham (1954) is orthogonal, and it considers epistasis as something additional to the marginal additive and dominance effects of the QTL. This may lead to an underestimation of the biological importance of epistasis. Kao and Zeng (2002) further discuss the modeling of QTL epistasis, using Cockerham's model. The model described by Seyffert (1966) is non-orthogonal in the modeling of the epistatic parameters. In this model, the estimates of the epistatic effects are not modeled as an addition to the additive and dominance effects of the individual loci, and are therefore a more biologically relevant indicator of the importance of epistasis. The interest to use these epistatic models in QTL mapping has increased and several recent papers describe various approaches for this (Broman 1997; Kao *et al*. 1999;this thesis; Jannink and Jansen 2001).

Interval mapping of Quantitative Trait Loci

Statistics is used in QTL mapping to evaluate the significance of putative QTL and to estimate their genomic location and genetic effects on the trait. Regression of phenotype on genotype at a marker location is called a single marker test and can be used to estimate the effects of QTL linked to the marker. In a genome scan composed of multiple single marker tests, the best estimated location for a QTL is taken as the genomic location of the marker with the highest statistical support for an association of the marker genotype with an effect on the phenotype. The major drawback with this method is a confounding between the genetic effect of the QTL and the distance of the QTL from the marker. Is a small effect due to a small QTL close to the marker or to a larger QTL further from the marker?

Lander and Botstein (1989) introduced the concept of interval mapping to disentangle the estimates of the location and genetic effect of a QTL. Interval mapping uses marker brackets instead of individual markers in the analysis and this makes it possible to make independent estimates of location and effect of the QTL. This method has since then become the basis for most QTL mapping methods. Within the concept of interval mapping, various methods have been proposed for significance testing and estimation of location and genetic effects.

Methods based on maximum likelihood estimation of location and genetic effects are widely used in QTL mapping. The advantage of this methodology is that they use the full information from the marker-trait distribution and is thus expected to be powerful. Disadvantages are a high computational demand, difficulties to modify the basic model and the need to construct specific analysis programs to perform the analyses. Construction of the maximum likelihood equations is rather straightforward, but obtaining the maximum likelihood estimates is much more difficult (Lynch and Walsh 1998).

In 1992, a QTL mapping procedure based on ordinary least squares regression was introduced (Haley and Knott 1992; Martinez and Curnow 1992). The basic principle is to estimate the probabilities of unknown QTL genotypes in the marker intervals, and from these calculate regression coefficients to be used for estimation of QTL location and effect. This methodology proved to be a good approximation to the maximum likelihood based methods and greatly reduced the computational demand of QTL mapping. It also simplifies modifications to the basic model and can be performed in standard statistical computer packages.

Haley et al. (1994) later extended the method to analyses of crosses between outbred lines.

Kao (2000) have, analytically and numerically by simulation, investigated the differences between QTL mapping based on maximum likelihood and linear regression. His study indicates that the maximum likelihood based methods can be more accurate, precise and powerful at the cost of an increased computational demand. The properties of the methods in real data where there is likely to be violations of model assumptions, such as unequal variances within QTL genotype classes, segregation distortion and unusual inheritance patterns, were not evaluated. It is therefore difficult to assess the properties of the methods in the analysis of experimental data sets.

Randomization testing in QTL mapping

Since QTL mapping involves multiple statistical tests throughout the genome, the selection of a significance threshold is a key issue of the procedure. Correction for multiple testing is necessary, since the use of a nominal significance threshold will lead to an elevated type I error (large number of falsely detected QTL). Various methods have been suggested to deal with the multiple comparisons (e.g. Lander and Botstein 1989; Kruglyak and Lander 1995; Benjamini and Hochberg 1995; Southey and Fernando 1998). Empirical estimation of overall significance thresholds can be done in a wide range of population designs by resampling techniques, such as randomization testing (Churchill and Doerge 1994). Here the observed trait values are randomly shuffled over individuals (genotypes) generating a sample with the original marker information, but with trait values randomly assigned over genotypes. The test statistic is then computed in the new sample, and the procedure is repeated many times, generating an empirical distribution of the test under the hypothesis of no marker-trait associations. By keeping the marker information for each individual together, the approach accounts for differences in marker densities, missing genotypes and segregation distortion. The major drawback with this method is a 1,000 to 10,000 fold increase in computational demand, which in some cases causes severe restrictions for the use of the method in practice.

Mapping of multiple QTL

Interval mapping, as described by Lander and Botstein (1989), was designed to map single QTL, and does not consider other, linked or unlinked, QTL affecting the trait. This decreases the power and resolution of the procedure when more than one QTL affects the trait. To overcome this, several authors have proposed extensions of interval mapping to mapping of multiple QTL. The basic concept of these methods is to include markers, or previously detected QTL, as cofactors in the model when interval mapping is used to search for QTL. The effects of linked QTL can be reduced by including markers linked to the interval of interest, whereas including unlinked markers can partly account for the segregation variance generated by unlinked QTL (e.g., Jansen, 1992, 1993; Jansen and Stam, 1994; Zeng, 1993a,b). These methods generally increase the power to detect a QTL and improve the precision in the estimates of QTL position

Efficient methods for detecting epistatic QTL are needed to gain a better understanding of the genetics underlying complex traits. Several lines of evidence indicate the importance of epistasis. For instance, approximately 40% of the genes in yeast do not yield an aberrant phenotype when ablated (Wolfe 2000) and the same alleles can cause a strain specific expression of autoimmune disease in mice (Bolland & Ravetch 2000). Epistasis has been reported in QTL mapping studies (Fijneman *et al.* 1996; Long *et al.* 1996; Li *et al.* 1997; Shook & Johnson 1999; Leips & Mackay 2000; Mackay 2001) and in various basic biological processes, which are expected to affect the expression of most traits. Biological processes where epistasis has been shown to be of importance are e.g., signaling pathways in both plants (Beaudoin *et al.* 2000) and animals (Araujo & Bier 2000; Scanga *et al.* 2000; Luschnig *et al.* 2000) and differential crossing-over and segregation (Khazanehdari & Borts 2000).

The first multiple QTL mapping methods assumed that there were no interactions between the QTL included in the model. Ignoring epistasis can decrease power and bias the estimates of QTL effects when epistasis exists. In order to detect interactions, a multidimensional search for QTL have to be performed. A search for QTL in multiple dimensions results in a large number of statistical tests and calls for an appropriate way to correct for multiple testing. If a Bonferroni type correction is used, the threshold would be high and the power to detect QTL decreased. Alternative methods to correct for multiple testing therefore need to be evaluated. Exhaustive searches (i.e. evaluation of all possible combinations of locations) in high dimensions also imposes a high computational demand for the mapping procedure. Various approaches have been described to address these issues.

Evaluating the importance of epistasis between pairs of QTL detected using a marginal effect QTL mapping method, is the most dramatic way to decrease the computational demand and the number of statistical tests performed (e.g. Brockman *et al*. 2000). Testing for epistasis among already detected QTL can be important in for evaluations of the importance of epistasis among already detected QTL, but QTL without significant marginal effects will remain undetected.

The use of multiple one-dimensional searches to detect epistatic QTL has been explored e.g. by Fijneman *et al*. (1996). The number of statistical tests performed are decreased by lowering the dimensionality in the search, by performing repeated one-dimensional searches where significant QTL in each round are added to a total genetic model. By using dimensional searches, the computational demand of the procedure is dramatically decreased. Detection of epistatic QTL is based on that at least one of the QTL has significant marginal effects and the power to detect QTL with small marginal effects, but with large epistatic effects, is thus limited.

Methods for simultaneous, two-dimensional searches for epistatic QTL pairs have been developed despite their computational demand and problems of obtaining appropriate significance thresholds. Epistatic QTL pairs are either searched among all marker combinations (Edwards *et al*. 1987; Damerval *et al*. 1994; Chase *et al*. 1997; Li *et al*. 1997; Holland 1998), or using interval mapping (Haley and Knott 1992).

Several methods have recently been proposed for mapping epistatic QTL. These methods and their properties will be further described later in this thesis.

Computational aspects of QTL mapping

QTL mapping is computationally demanding due to multiple testing and an extensive use of resampling techniques to obtain empirical significance thresholds. Increasing sizes of mapping populations calls for new QTL mapping methods that fully explore the experimental data. If a new improved method is proposed, a relatively low computational demand is desired. Several properties of the mapping procedure make it computationally intense, and in this section we will point out some key issues.

1. The computational kernel

The computational kernel in QTL mapping is the method used for statistical testing and parameter estimation. The calculations involved in this procedure are performed repeatedly, once for each genomic position or positions that are evaluated. Generally, the core problem is based on a minimization of $\|G(X\boldsymbol{b}\text{-}\boldsymbol{y})\|_2$, where G is a matrix that varies depending on the statistical method used. In QTL mapping based on ordinary least squares, as described by Haley and Knott (1992) and Martinez and Curnow (1992), $G=I$, where I is the identity matrix. A further discussion regarding other statistical methods can be found in Ljungberg *et al.* (2002). We have mainly used ordinary least squares in our research, and will therefore only briefly mention the procedures based on maximum likelihood methods.

2. Iterations to obtain maximum likelihood estimates

When maximum likelihood is used to estimate parameters for the QTL, the estimates are normally obtained in an iterative procedure, such as the EM-algorithm. The computational demand is thus increased as compared to the ordinary least squares based methodologies.

3. Global search algorithms used in QTL mapping

QTL mapping involves evaluation of genetic models including one or several QTL. We refer to the selection of genomic locations for these QTL(s) as a global search strategy. The search space in QTL mapping is generally described as a grid, either based on markers or on genetic map positions. In a genomic grid based on genetic markers, each node in the grid is at the location of a genetic marker. The density of the grid depends on the spacing of the markers, and will thus vary across the genome. This mapping procedure is usually called single marker test. For genomic grids based on genetic maps, each node in the grid is a location on the genetic map. The QTL mapping procedure based on genetic map positions is usually called interval mapping. The resolution of the grid can be varied, by choosing the step size through the genetic map, in the interval mapping procedure. We will here briefly introduce the search methods used to search for QTL in these grids

A) Exhaustive searches
A one-dimensional exhaustive search in a grid based on markers is a standard, single marker test. This method has been used since the first genetic markers became available. Multidimensional exhaustive searches have been used in marker grids to detect epistatic QTL (e.g. Edwards *et al*. 1987; Damerval *et al*. 1994; Chase *et al*. 1997; Li *et al*. 1997; Holland *et al.* 1997). Recently, Sugyiama *et al* (2001) proposed to use a randomization test for epistatic QTL pairs based on an exhaustive search in a grid of markers, and this method has later also been applied by others (Kim *et al*. 2001; Shimomura *et al*. 2001).

The exhaustive search is also commonly used for one-dimensional genome scans based on genetic maps (interval mapping). When multidimensional grids are considered, the computational demand of the exhaustive search increases exponentially, and therefore several alternative search methods have been proposed for this task.

B) Search methods reducing the dimensionality of the search
Forward selection can be used to decrease the number of positions to evaluate when searching for multiple QTL. By performing sequential one-dimensional exhaustive genome scans, where the most significant QTL in each scan is added to the genetic model, one can obtain an approximation to the exhaustive multidimensional search. This significantly reduces the computational complexity of the search (Jansen 1992, 1993; Jansen and Stam 1994; Broman 1997). The method leaves large portions of the genome unexplored, and thus will have limited power to detect QTL with small marginal effects.

C) Multidimensional search algorithms
Stochastic multidimensional searches based on e.g. Markov Chain Monte Carlo (MCMC) can be used to simultaneously map multiple QTL. MCMC has been used in Bayesian QTL mapping to map multiple QTL (Satagopan *et al*. 1996; Hoeschele *et al*. 1997; Sillanpää and Arjas 1999). The methodology is powerful, but its implementation and use require substantial knowledge and experience of the technique.

In paper III, we propose the use of an evolutionary algorithm (the genetic algorithm) for simultaneous multidimensional searches in interval mapping for QTL. The major strength of evolutionary algorithms, in the context of QTL mapping, is their robustness. We have shown that the search algorithm works properly for mapping multiple QTL.

4. Randomization testing to derive empirical significance thresholds for QTL

Randomization testing is extensively used to obtain significance thresholds in QTL mapping. The major drawback of the method is that it increases the computational demand by performing repetitive (in practise 1,000 to 10,000) analyses on permuted data sets. In many practical studies, where multiple traits are analysed, randomization tests are performed on a smaller number of the traits and the same thresholds are then used for the other traits in the study (Malek *et al.* 2001). This might be reasonable for many traits, but it would be desirable to use a unique threshold for each trait. One- and multidimensional randomization tests can be made computationally feasible by improving the computational efficiency of task 1-3 described above. Further decreases in analysis times can also be obtained by parallel computing, and this will be further discussed later in this text.

General code optimization

To increase the performance of the QTL mapping software, the appropriate algorithms and software libraries should be selected to perform the desired tasks, the code should be optimized and the appropriate compiler options selected. This procedure can lead to substantial decreases in analysis times. Available software tools simplifies the profiling and optimization process, by identifying bottlenecks in the software.

Objectives of the present studies

The most commonly used model for QTL mapping includes marginal additive and dominance genetic effects. When dealing with imprinted or epistatic QTL, this model is not appropriate. To obtain maximum power to detect QTL and obtain correct estimates of genetic effects, other genetic models should also be evaluated.

Simultaneous mapping of multiple QTL and randomization testing for significance testing increases the power to map QTL and help to avoid detection of false positive QTL. Both techniques, however, dramatically increase the computational demand. These techniques can be more widely used by implementing efficient algorithms and designing optimized software utilizing high performance computing. This will lead to a more in-depth analysis of collected datasets.

The objectives of the present thesis were:

- to examine and extend the use of alternative genetic models in QTL mapping
- to evaluate whether high performance computing is useful in QTL mapping and to develop highly optimized software for mapping QTL in outbred line-crosses
- to develop and evaluate strategies and algorithms for simultaneous mapping of and statistical testing for multiple interacting QTL
- to apply the developed methods to data collected from divergent farm animal crosses

Material and Methods

Supercomputers have been used throughout this thesis for simulation studies and mapping of QTL in experimental line crosses. The following section will describe the computers and experimental populations that have been used.

Computer resources and techniques

High performance computer resources were throughout this project provided by the National Supercomputing Centre (NSC) at Linköping University. Simulation studies and analysis of experimental data were performed in parallel on a 272 processor distributed memory supercomputer (Cray T3E-600). A minor part of the work was also performed on a 32 processor PC Cluster at NSC. Profiling and code optimization was done on the Cray T3E, using the Apprentice software package (Anderson *et al.* 1997).

The original version of the QTL mapping software used for paper I and II, was kindly provided by Drs Sarah Knott and Chris Haley. The software was rewritten for parallel computing and for testing an imprinted QTL model. The computer software used in paper III-V is new, entirely written in Fortran 90 and has been extensively optimized and involves routines for parallel execution of interval mapping and randomization testing. It can perform all analyses described in this thesis. The software is currently ported to a Linux workstation (PII-400 MHz, NAG f95 compiler) and has been optimized for high performance on a Cray T3E (Alpha 600 MHz, Cray f90 compiler) and PC clusters (AMD Athlon, 900 MHz, Portland Group f90 compiler) at NSC.

Analysed data

Study I and II used data from an F_2 intercross between the European Wild Boar and the Large White breed. The pedigree consists of 191 F_2 animals in 26 full sib families, which have been genotyped for 248 genetic markers on 18 autosomes and the X chromosome. The cross was initiated in 1989 and numerous QTL have been detected for various traits (e.g. Andersson *et al.* 1994; Edfors-Lilja *et al.* 1998; Andersson-Eklund *et al.* 2000).

Study III and IV were based on simulated data for two inbred F_2 intercrosses. The simulated pedigrees consisted of 520 individuals in 26 full sib families. All individuals had full marker information every 10 cM on 20 chromosomes each of length 100 cM.

Study V used data from an F_2 intercross between the Red junglefowl and a White Leghorn selection line. The cross was initiated in 1998 and consists of 852 F_2 individuals in 38 full sib families. The current genetic map includes 104 genetic markers on 24 autosomal linkage groups with an average marker spacing of 25.4 cM. The cross was developed to study the genetics of domestication of production and behavioural traits in the chicken and was set up at the Department of Animal Environment and Health at SLU in Skara. Several QTL mapping studies have been performed using this material including analysis of behavioural traits (Schütz *et al.* submitted), production traits (Kerje *et al.* manuscript), body composition and bone density (Kindmark *et al.* in preparation) and meat- and egg quality (Babol *et al.* in preparation).

Results and Discussion

My work has focused on developing more comprehensive methods for QTL detection and improving the computational efficiency of QTL mapping. Key topics have been simultaneous mapping of multiple QTL and randomization testing in multiple QTL models, since these techniques have great potential to improve the efficiency in QTL mapping. We have proposed how the techniques can be used and their computational demand can be decreased. Finally, the proposed strategies have been evaluated using simulated and experimental data.

Mapping of imprinted QTL in an experimental linecross (I)

The first report of an imprinted locus affecting body composition in farm animals was the callipyge (*CPLG*) locus in sheep (Cocket *et al.* 1996). Later, we (study I) and Nezer *et al.* (1999) reported imprinting of an *IGF-2* linked QTL in pigs. A more recent study in pigs indicates that imprinting might be a more common phenomenon, as four out of five detected QTL for body composition in pigs were reported to be imprinted (de Koning *et al.* 2000; Rattink *et al.* 2000). This high incidence of imprinting is surprising, since genetic imprinting has only been reported for a small number of genes in mammals (Reik and Walter 2001).

Material and Methods

The QTL mapping method used in our study was developed to detect QTL in a three generation pedigree from a cross between divergent lines using an additive/dominance genetic model (Haley *et al.* 1994). Knott *et al.* (1998) later described how this method could be extended to map imprinted QTL. The analysis is performed in two stages, where first marker genotypes are used to estimate the probability of an F_2 offspring being each of four possible QTL genotypes (accounting for parental origin) at fixed 1-cM intervals across the genome. The estimation of origin is simpler for loci with multiple (>2) alleles, and is impossible for biallelic markers, where the parental lines are fixed for alternative alleles. Estimation of parental origin is thus impossible in inbred linecrosses where all parental F_1 animals have the same heterozygous genotype. Under the assumption that the grandparental breeds are fixed for alternative alleles, the probabilities are used in a least squares framework to investigate two alternative genetic models, one including additive and dominance components of a QTL and the other including an additional imprinting component of the QTL.

Genetic models

By denoting the effect of *QQ* as *a*, the effect of (*Qq*+*qQ*)/2 as *d* and the effect of *qq* as *–a* (Falconer and Mackay 1996), the following linear model describe the expected value of an offspring in terms of the additive and dominance contributions at a QTL (Knott *et al.* 1998):

$$y_i = \mu + c_{ai}a + c_{di}d + \varepsilon_i$$

where μ is the mean, c_{ai} is the regression indicator variable for the additive component for individual *i* at the given location which, denoting the probability of an individual of being genotype *XX* as P_{XX}, is equal to P_{QQ}-P_{qq} and c_{di} is the regression indicator variable for the dominance component for individual *i* at the given location, which is equal to P_{Qq}+P_{qQ}, and ε_i~$(0,\sigma^2)$.

When taking account for the grandparental origin of the alleles, there are four possible genotypes in the F_2 generation. This makes it possible to fit three effects, the third being the difference between heterozygote classes. The difference between the two classes of heterozygotes is the parental origin of the two alleles. In the first heterozygote class, the wild boar allele is inherited through the F_1 mother and in the other through the F_1 father, and the reverse for the Large White alleles. For the autosomes, the difference between the heterozygotes should indicate whether imprinting is an important effect (Knott *et al.* 1998). By denoting the difference in mean effects of *Qq* and *qQ* individuals as *im*, the QTL model that fits an imprinting effect additional to the additive and dominance QTL effects can be described as:

$$y_i = \mu + c_{ai}a + c_{di}d + c_{imi}im + \varepsilon_i$$

where μ, c_{ai} and c_{di} and ε_i are defined as before and c_{imi} is the regression indicator variable for the imprinting component (*im*) for individual *i* at the given location, which is equal to P_{Qq}-P_{qQ}.

QTL analysis and significance levels

A one dimensional genome scan was performed using both the additive/dominance model and the additive/dominance/imprinting model. The grid size was 1 cM. An F-ratio test was used to compare the non-imprinted QTL model and the imprinted QTL model at each location to a reduced model without the QTL. The best estimate for the position of the QTL was taken to be the location giving the highest F-ratio. If the QTL model fitting imprinting was significant, it was compared with the QTL model without imprinting (one degree of freedom) to see whether the imprinting effect was significant. The significance levels were obtained empirically by a permutation test (Churchill and Doerge 1994) according to Knott *et al.* (1998). In addition to the regressions on the additive, dominance and imprinting coefficients, statistical models included the effects of other relevant phenotypic observations which had significant effect on the respective trait (Andersson-Eklund *et al.* 1998). Family and significant unlinked QTL were also included in the analyses to account for background genetic effects.

Summary of results

In study I, we report the presence of an imprinted QTL on the p-arm of pig chromosome 2. The QTL had large effects on lean meat content in ham and explained 30% of the residual phenotypic variation in the F_2 population. Effects were also detected for the area of the longissimus dorsi muscle, heart weight and backfat thickness. The Large White allele was associated with larger muscle mass and reduced backfat thickness and a clear paternal expression was shown.

Discussion

Imprinting is in general considered to be a relatively rare occurrence. Due to this, relatively few attempts have been made to design QTL mapping methods to detect imprinted QTL. In this section the proposed methods will be briefly discussed in relation to the method we have used.

Cocket *et al.* (1996) and Nezer *et al.* (1999) used interval mapping based on maximum likelihood to test for imprinting. Three alternative hypothesis were compared to the null hypothesis (H_0) of no QTL affecting the trait:

H_1: a Mendelian QTL segregates in the population
H_2: a paternally expressed imprinted QTL segregates in the population
H_3: a maternally expressed imprinted QTL segregates in the population

To infer the genetic model for the QTL they compare the best imprinting QTL model (paternal (H_2) or maternal imprinting (H_3)) to the model of a Mendelian QTL using a LOD score test ($\log_{10}(H_{2(3)}/H_1)$) at the best location for the QTL. This approach to map imprinted QTL is rather similar to that described by Knott *et al.* (1998). Instead of including all parameters in the same model, three alternative genetic models are evaluated and later compared. The advantage of this procedure is that only appropriate genetic effects are included in each genetic model, and therefore a high power can be obtained. The disadvantage is that multiple genome scans with several alternative genetic models needs to be performed, which limits the use of this approach as a standard QTL mapping procedure.

de Koning *et al*. (2000) suggest a rather different genetic model for mapping imprinted QTL. Rather than using the additive, dominance and imprinting coefficients described above, they suggest that the following model is used

$$y_i = \mu + c_{impi}imp + c_{immi}imm + c_{di}d + \varepsilon_i$$

where μ is the mean, c_{impi} is the regression indicator variable for paternal imprinting for individual *i* at the given location which, denoting the probability of an individual of being genotype *XX* as P_{XX}, is equal to $[P_{QQ}+P_{Qq}]$- $[P_{qQ}+P_{qq}]$, c_{immi} is the regression indicator variable for maternal imprinting for individual *i* at the given location which is equal to $[P_{QQ}+P_{qQ}]$- $[P_{Qq}+P_{qq}]$ and c_{di} is the coefficient for the dominance component calculated as described above. The use of the paternal and maternal imprinting coefficients represents contrasts for paternal and maternal imprinting, and will give the estimates of the imprinting effects for imprinted QTL. The model was proposed to give estimates of the imprinting effect, and whether the imprinting is paternal or maternal, directly in the genome-scan. The benefit of using the model is questionable. First, imprinting is an exception rather than the rule, and therefore it should not replace the additive effect as the main effect in the model. Second, the proposed model is still not a true imprinting model, since the dominance effect is included. Dominance and imprinting are mutually exclusive genetic effects and if only imprinting is sought, the dominance effect should be removed from the model. Third, if there is Mendelian segregation, both the paternal and maternal imprinting parameters will have significant, non-zero, effects, which is confusing in the interpretation of the results. Based on this, we argue that the imprinting model proposed by Knott *et al*. (1998), should be used to map non-imprinted and imprinted QTL.

Conclusions

The results from this study clearly indicates the presence of an imprinted QTL and that the power to map imprinted QTL increases when using an appropriate genetic model. Subsequent studies have been performed to map imprinted QTL in pigs (de König *et al*. 2000; Rattink *et al*. 2000), and these studies indicate an unexpectedly high incidence of imprinted QTL. Even though it needs to be confirmed that imprinting is such a widespread phenomena, it might be worthwhile to include an imprinting effect in the genetic model to acchieve maximal power for mapping QTL in those species where imprinting occurs (e.g. mammals).

QTL mapping benefits from high performance computing (II)

To improve the throughput in QTL analyses, the computational aspects of the procedure needs to be considered. We will here describe several ways to improve the performance of software used for QTL mapping.

Optimization of QTL mapping software for single processor machines

We have based our work on the least squares interval mapping method to map QTL (Haley and Knott 1992; Haley *et al.* 1994). The method solves a standard least squares problem for each (or each combination of) proposed genomic location(s). To improve the efficiency of the computational kernel of the mapping procedure, we have evaluated the following options to solve the least squares problem:

1. A general least squares solver (g02DAF) from the commercial NAG software library.
2. Our own code to solve the normal equations $b = (X'X)^{-1}X'y$ by explicitly inverting the matrix *X'X*.
3. Code from Numerical Recipes (Press *et al.* 1999) to solve the normal equations by LU- or QR- factorization or by singular value decomposition
4. General numerical libraries, BLAS and LAPACK (Anderson *et. al.* 1999) freely available from Netlib (http://www.netlib.org), to solve the normal equations by LU decomposition.
5. Specialised solver for solving the series of least squares problems in QTL mapping (Ljungberg *et al.* 2002)

The general least squares solver g02DAF is robust and designed to solve a wide range of least squares problems. The long computational time of the routine when used for QTL mapping is mainly due to computations, which are not required for computing the quantity used in mapping of QTL.

We initially replaced the NAG library routine with our own code to solve the normal equations by matrix inversion. This code was more efficient, since unnecessary computations were removed. Later, we also evaluated LU- and QR-factorization and singular value decomposition as other alternatives to solve the normal equations. LU-factorization of the normal equations was most efficient, and was selected as the standard method in our analysis software. The theoretical predicted speedup from using LU-factorization (four times), instead of solving the normal equations by explicitly inverting the matrix *X'X*, in practise turned into a 1.6 times speedup of our analysis software. We compared the LU-factorization implementations in the Numerical Recipes and the computer platform optimized BLAS/LAPACK libraries. The freeware LAPACK libraries include efficient routines for a wide range of numerical problems, and exist in several versions optimized for a many computer platforms. The LAPACK routines were more efficient when the design matrix (*X*) in the least squares problem contained more than 9 columns. For smaller matrices the code provided by Numerical Recipes implementation was more efficient, which most likely is a reflection of the data control statements included in the LAPACK library routines. This will only be seen where the matrixes are small, and a very small number of computations are performed. For general problems, the LAPACK libraries are recommended.

Several fixed effects and covariates are often included in the analysis when the least squares method for QTL mapping is used. Many of the columns in the design matrix (*X*) will thus remain unchanged during consecutive computations in genome scans and randomization tests. QR factorization with updating is an efficient algorithm for solving these problems (Björck 1996). A slightly modified version of this algorithm was evaluated by Ljungberg *et al.* (2002). A significantly improved computational efficiency was shown for sizes of the least squares problem corresponding to realistic QTL mapping scenarios. Decreases in computational times in the order of 10 and 100 times were obtained when this strategy was compared to the LAPACK implementation of LU-factorization and the g02DAF routine from NAG. The advantage of using this method increases with the number of individuals in the cross and the number of unchanged columns in the design matrix. In Figure 3, we show the results for two different population sizes and number of fixed columns representing mapping one QTL by two marginal genetic effects and two QTL using an epistatic QTL model with eight genetic effects. The number of individuals and fixed effects used for the evaluations equals those in the previously described Wild Boar x Large White and White Leghorn x Red junglefowl intercrosses.

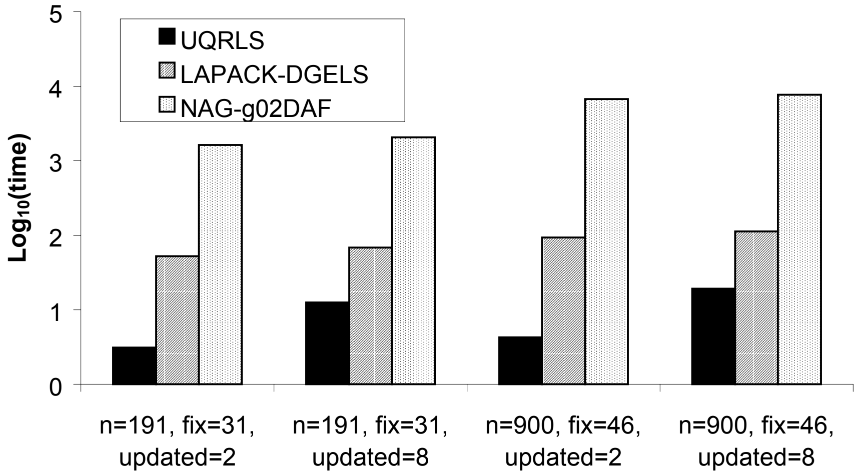

Figure 3. The computational performance of three different solvers for least squares problems. The NAG library routine g02DAF, the LAPACK routine DGELS and the custom written routine UQRLS. n: number of individuals in the population, fix: number of fixed columns in the design matrix (*X*), updated: number of updated columns in the design matrix (*X*).

Our analysis software has been profiled and optimized using the Apprentice software on the Cray T3E at the National Supercomputing Centre. Figure 4 shows the distribution of tasks, the ratio between memory access and computations as well as the performance of the code in Megaflops (Mflops) before and after optimization.

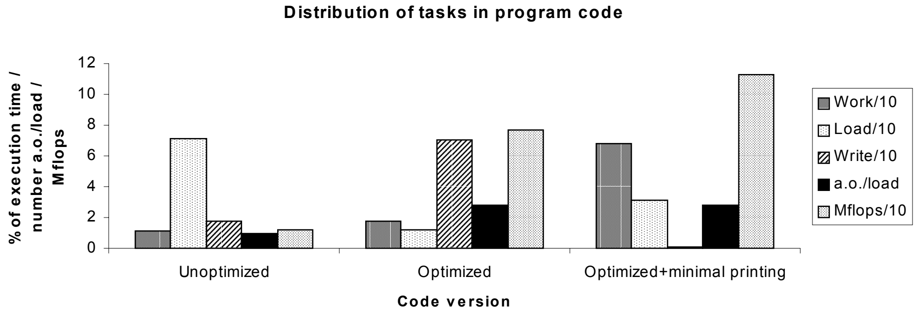

Figure 4. Profiling of program used for QTL mapping before and after optimization and after removing a majority of the output statements. a.o.: arithmetic operations.

By improved memory access patterns, and decreasing the number of Input/Output operations, the performance of the analysis code increased from 11.7 Mflops to 115 Mflops on the Cray T3E. The Alpha 300 MHz processors on the Cray have a theoretical peak performance of 600 Mflops, and a performance of 115 Mflops is considered to make efficient use of the computer. High peak performance can thus be obtained in software for QTL mapping type problems. For a population size of 520 individuals without fixed effects or cofactors, a randomization test for a single QTL takes about 8 minutes using the optimized version of the software, which means that randomization tests are computationally feasible for single processor machines. The analysis times for the conditional and additional randomization tests proposed in paper IV, are about 25 and 425 minutes respectively using the optimized software based on LAPACK LU factorization to solve the normal equations.

Parallel algorithms in QTL mapping and randomization testing for QTL

A large number of statistical tests are performed during a genome-scan for QTL. The tests can be performed in any order, without knowledge of the result from any of the other tests. Thus, the tests can be solved on separate computers or in parallel on a multiprocessor computer. In paper II, we present a simple algorithm for parallelising existing QTL mapping software by allocating the computations for each chromosome to alternative processors. We show that this simple partitioning of the problem gives reasonable speedups of the QTL mapping procedure.

To obtain a better load balance between the processors, the computations should be separated independent on their chromosomal location. This procedure is used in our software, and gives large speedups for exhaustive simultaneous searches for epistatic QTL pairs. Figure 5 shows the speedups obtained during the simultaneous mapping step used in study V. The speedup is linear up to 32 processor (31.5 times speedup). Further speedups are obtained up to 64 processors (50 times) after which no further improvements are made. The analysis times decreases from 95 minutes on one processor to two minutes on 64 processors.

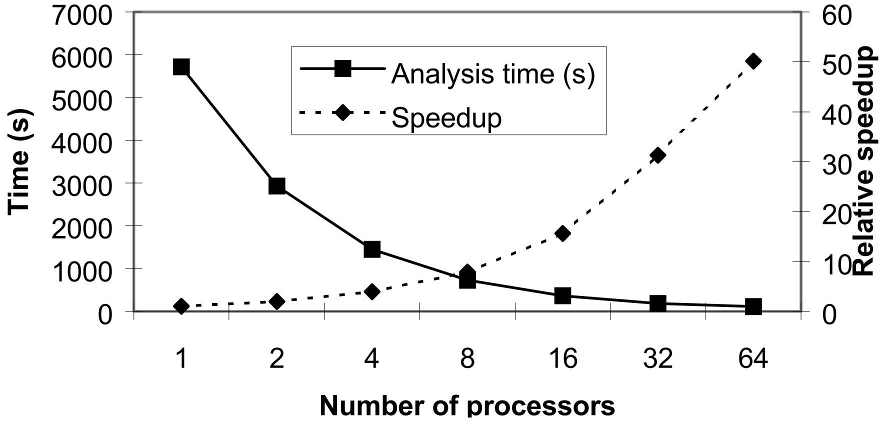

Figure 5. The decrease in analysis time and speedup of an exhaustive search for two epistatic QTL on one to 64 processors on a Cray T3E computer.

In randomization testing, several independent genome-scans are performed on permuted data sets. Distributing the analyses of the permuted datasets to different processors is a simple and efficient parallel algorithm for this problem. The load balance is good, provided that the number of data sets is a multiple of the number of used processors. Figure 6 shows the results from parallel execution of the three randomization tests proposed in study IV.

All three randomization tests scale very well to 10 processors (relative speedup 8.1/9.4 /9.7), and utilize up to 25 processors efficiently, where they reach relative speedups of 17.7, 20.5, and 23.3. The randomization tests for epistatic QTL pairs can utilize up to 32 processors, where they reach speedups of 21.6 and 24.2 times respectively. By using 32 Alpha 300 MHz processors on a Cray T3E, a randomization test for a single QTL takes about 30s, a randomization test for a single QTL plus a second epistatic QTL takes about 75s and for a single QTL plus two additional QTL takes about 22 minutes. A similar evaluation has been performed on a 32 processor (AMD Athlon 900 MHz) PC Cluster, and it shows that the scalability of the randomization testing procedure equals that achieved on the Cray T3E. The computation times on the PC Cluster are generally decreased by a factor 1.2 compared to those on the Cray T3E. Larger data sets including more individuals and several cofactors will increase the analysis times.

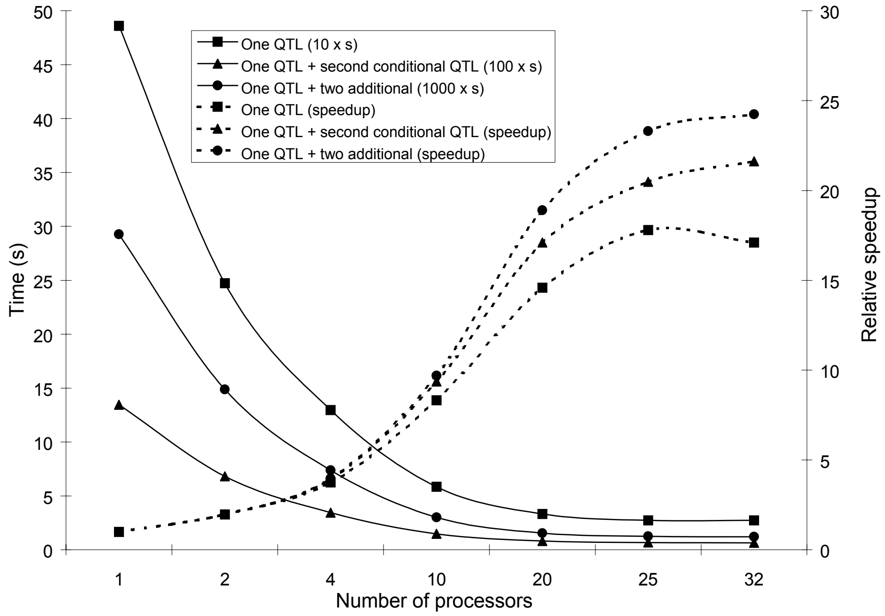

Figure 6. Serial and parallel performance of the analysis software to perform randomization tests for one-, one plus two conditional and one plus two additional QTL as described in study IV, on one to 32 Alpha 300MHz processors on a Cray T3E computer. The analyses were performed in a data set with 520 individuals and two or eight genetic parameters in the model.

In Table 1, we present the relative increases in computational time as the number of columns in the design matrix of the least squares problem (*X*) increases. This has to be considered when mapping QTL in experimental data sets, but can to some extent be accounted for by the method described by Ljungberg *et al.* (2002).

Table 1. The relative increase in the computational demand as the number of columns increases in the design matrix (*X*) in the ordinary least squares problem $\boldsymbol{y}=X\boldsymbol{b}+\boldsymbol{e}$.

	Columns in design matrix				
	3	21	41	61	81
Relative computational demand	1	3	7.8	15.8	26.6

An overview of recent approaches to map epistatic QTL

There is a great interest among geneticists to find interacting QTL and understand the genetic basis for these interactions. It is feasible to design QTL mapping populations, which are large enough to detect epistasis. Therefore it is necessary to develop and evaluate methodologies to efficiently analyse such experimental data. Several methods for mapping of epistatic QTL have been proposed during the course of the thesis work. In this section, these will be briefly introduced, to provide an introduction to the current status of the field.

Multiple Interval Mapping (MIM)

Zeng and co-authors have proposed an interval mapping method to simultaneously map multiple QTL and their interactions in experimental crosses between inbred lines. The method is called Multiple Interval Mapping (MIM) (Kao and Zeng 1997; Kao *et al.* 1999; Zeng *et al.* 1999). MIM models two-locus epistasis between all pairs of QTL using an orthogonal genetic model (Cockerham 1954). It uses maximum likelihood for parameter estimation in a statistical framework based on simultaneous modeling of multiple interacting QTL. The implemented method is not a true simultaneous mapping method, since the search algorithm is limited to map QTL in pre-selected genomic regions. The full potential of the method has thus not been reached in its current implementation (Zeng *et al.* 2000). The authors have proposed several alternative strategies for significance testing and model selection, but there is still no obvious way to handle these issues for this method. The method has not been evaluated by simulation, but QTL mapping for high heritability traits in experimental *Drosophila* crosses suggests a high power to detect interacting QTL.

A mixed model approach to mapping of epistatic QTL

It is desirable to take account of genetic background effects in the QTL mapping procedure. Wang *et al.* (1999) proposed a mixed model approach to map epistatic pairs of QTL and to detect QTL by environment interactions. To take account for the genetic background in the mapping procedure, the method uses a model including a random polygenic effect besides the fixed effects for an epistatic QTL pair. The method, however, does not simultaneously search for multiple QTL and multiple interactions.

Simultaneous mapping and randomization testing for QTL at marker locations

Methods based on exhaustive simultaneous searches for epistatic QTL pairs have been developed despite their computational demand and problems of obtaining appropriate significance thresholds. Recently, a method was proposed for randomization testing and simultaneous mapping of epistatic QTL pairs(Sugiyama *et al.* 2001). The proposed method is based on exhaustive searches in genomic grids based on markers, and introduces randomization tests for the null hypothesis of no QTL versus the alternative hypotheses of two interacting QTL. An F-test is then used to discriminate whether the significance was due to strong main effects of the first QTL or truly due to the QTL pair. An F-test is also used to select the final model (additive/dominance or interaction) for the QTL pair. In a simulation study performed by Fridlyand (2001), the method has problems to discriminate between large main effects of single QTL and the effect of an epistatic QTL pair.

Mapping of epistatic loci using QTL by genetic background interactions

Jannink and Jansen (2001) proposed a method for mapping epistatic QTL in large populations derived from multiple related inbred line crosses. These types of crosses are typically available in plants. The method is based on one-dimensional genome scans and maps epistatic QTL by identifying loci with high QTL by genetic background interaction. The advantage of the method is that it can handle higher order epistasis. The power to detect epistasis is large according to the simulation study performed. This method is not applicable in the type of line crosses available in animals.

Non-parametric methods

A non-parametric method for mapping of multiple interacting QTL has been described by Fridlyand (2001). This method attempts to detect QTL regions only and makes no attempt to make inferences about the underlying genetic model. The method is based on regression trees, forward selection and stratification of data based on genotypes at marker loci. The advantage of the method is that makes no assumptions about the genetic effects of each detected QTL, which makes it possible to detect higher order epistasis. Disadvantages with the method are i) the forward selection approach, which limits the power to detect QTL with small major effects, ii) the need for very large population sizes to map more than three QTL and iii) the lack of provided information about the underlying genetic model for the trait.

Bayesian methods

Sen and Churchill (2001) have described a general statistical framework that can accommodate multiple interacting QTL. The method is based on a Monte Carlo algorithm to implement QTL analysis and allows the analyst to focus on modeling and model comparisons. In their implementation, single and pairwise genome scans are performed to identify those regions that exceed randomization testing thresholds for no QTL versus one QTL and no QTL versus two QTL. Multiple gene models are then fitted for the regions that are significant in the genome scans. Model comparisons are made using Bayes factors or likelihood-ratio tests. The method is an extension of that described by Sugiyama *et al*. (2001)

A new strategy for simultaneous mapping of epistatic QTL

In paper III and IV, we describe an interval mapping strategy to simultaneously search and statistically test for epistatic QTL pairs. The method uses i) a genetic algorithm to simultaneously search for epistatic QTL pairs, ii) a least squares regression based method for evaluating the fit of a two locus epistatic model for the pairs proposed by the genetic algorithm and iii) significance testing based on randomization testing.

Simultaneous mapping of epistatic QTL using a genetic algorithm (III)

When QTL are mapped by their marginal genetic effects, as in a standard one-dimensional search, epistatic QTL pairs can only be detected when both QTL have significant marginal effects. In this case it is possible to test if any of all the possible pairs of QTL interacts. When only one or none of the QTL are significant, an alternative mapping approach is needed.

Forward selection can be used for a one-dimensional search for QTL that interact with QTL that can be detected by their marginal effects. With forward selection it is possible to perform one-dimensional genome scans with an epistatic model to detect additional epistatic QTL. This will increase the power to find QTL pairs, where one of the QTL has significant marginal effects. However, it is not possible to detect QTL pairs where none of the QTL has significant marginal effects.

A simultaneous search for QTL pairs will increase the power to detect QTL without significant marginal effects, but with significant epistatic effects (study III+IV). An exhaustive search for QTL pairs is computationally demanding, and an exhaustive search for more than two QTL or the use of exhaustive multidimensional searches in randomization testing is computationally infeasible. An efficient search algorithm is needed to make the simultaneous search feasible. The genetic algorithm (Goldberg 1989) is a robust search method with good computational performance. In study III, we compared a genetic algorithm to a forward selection strategy to search for epistatic QTL pairs. The one-dimensional, forward selection based, search algorithm had the least computational demand, but the lowest efficiency in the search. The genetic algorithm based two-dimensional, simultaneous search performed well and only increases the computational demand fourfold when compared to forward selection. The computational demand of the exhaustive search was considerably higher, a 200-fold increase, as compared to the genetic algorithm. In summary, the genetic algorithm proved to be an efficient search algorithm for mapping interacting QTL pairs, and the ability of the method to detect the QTL is almost as good as for an exhaustive search. The genetic algorithm also clearly outperformed the forward selection based method for almost all evaluated types of epistasis.

Randomization testing for multiple epistatic QTL (IV)

Detection of epistasis can be considered as a model selection problem, i.e. is a model including marginal and epistatic effects more suitable for the data than a model that only includes the marginal effects. The properties of several model selection methods have been evaluated in the context of QTL mapping for marginal effects (Broman 1997) and epistatic QTL models (Zeng *et al*. 2000). The genetic models used in these studies were orthogonal, and the individual estimates of the parameters in the model thus remain unchanged when other effects are added or withdrawn from the model. Study III was based on a non-orthogonal genetic model and we wanted to evaluate randomization testing for model selection in this context.

In order to obtain correct estimates of the genetic effects in the non-orthogonal model, model selection means selection of either the marginal effects model:

$$y_i = \mu + \sum_{j=1}^{2} \left(c_{aij} a_j + c_{dij} d_j \right) + \varepsilon_i$$

where μ is the mean, c_{aij} is the regression indicator variable for the additive component for individual *i* and QTL *j* at the given location which, denoting the probability of an individual of being genotype *XX* as P_{XX}, is equal to P_{QQ}-P_{qq} and c_{dij} is the regression indicator variable for the dominance component for individual *i* and QTL *j* at the given location which is equal to P_{Qq}+P_{qQ}, and ε_i~$(0,\sigma^2)$.

or the full epistatic model

$$y_i = \mu + \sum_{i=1}^{2} (c_{aij} a_j + c_{dij} d_j) + c_{aai} aa + c_{adi} ad + c_{dai} da + c_{ddi} dd + \varepsilon_i$$

where μ, c_{aij}, c_{dij} and ε_j are the same as above, and c_{aai}, c_{adi}, c_{dai} and c_{ddi}, are the regression indicator variables for the additive by additive, additive by dominance, dominance by additive and dominance by dominance components of the model and are calculated as c_{aai}= c_{ai1}* c_{ai2}, c_{adi}= c_{ai1}* c_{di2}, c_{dai}= c_{di1}* c_{ai2}, c_{ddi}= c_{di1}* c_{di2} (Haley and Knott 1992).

Most analytically based model selection techniques, e.g. Akaike's Information Criterion (AIC) (Akaike 1969), Mallows Cp (Mallows 1973), or Bayes Information Criterion (BIC) (Schwartz 1978; Hannan and Quinn 1979), treat each additional model parameter equally. In order to add an additional parameter, the parameter needs to explain a predefined proportion of the residual variance. When comparing the marginal effects model with the epistatic model for a QTL pair, the number of genetic parameters is doubled. In the epistatic genetic models, the variance explained by epistasis will be the amount of additional variance that can not be explained by the parameters for the marginal genetic effects. For some types of epistasis, this additional variance will be relatively small, although epistasis is rather important for the trait. It is therefore important to have a sensitive method for model selection to be able to detect these types of epistasis. Instead of using an analytically derived model selection method, a model selection method based on the properties of the genetic models in QTL mapping experiments would be desirable. In study IV, we propose randomization testing to derive an empirical distribution of the additional variance explained by the epistasis parameters. The properties of the method were evaluated using data simulated under various additive and epistatic genetic models. We were able to show that the proposed randomization testing strategy for model selection was powerful and gives reasonable type I errors.

The computational demand of this model selection method is modest, since it only performs a randomization test for one pair at a time. It is thus possible to evaluate numerous combinations of QTL pairs on a single processor machine. The number of permutations can also be decreased by including a check for the sufficient number of permutations in the testing procedure (Nettleton and Doerge 2000).

A simultaneous search for epistatic QTL pairs involves many statistical tests, and an appropriate significance level needs to be selected for the proposed QTL pairs. In study IV, we describe a randomization testing strategy procedure for simultaneously mapped epistatic QTL pairs. The strategy is an add-on to a standard one-dimensional search, and it uses the information about QTL that are significant by their marginal effects in significance testing for QTL pairs. The strategy introduces consecutive steps with forward selection of simultaneously mapped epistatic QTL pairs. In the forward selection procedure, epistatic QTL are simultaneously mapped using a genetic algorithm, and statistically tested for depending on whether the QTL in the pair are significant by their marginal effects or not (see points 1-3 below). If the proposed QTL pair is significant, the QTL are added to the genetic model and the procedure continues, otherwise the forward selection procedure is terminated. The significance testing for the QTL pairs is based on the following three alternatives:

1. Both QTL are significant by their marginal effects

When both of the QTL in the proposed pair are significant by their marginal effects, we conclude that the QTL in the proposed epistatic QTL pair exist. In this case, a model selection randomization test is performed to evaluate whether epistasis is present or not.

2. None of the QTL are significant by their marginal effects

When none of the QTL in the proposed pair can be detected by their marginal effects, we propose to test the alternative hypothesis that an epistatic QTL pair exists versus the null hypothesis that no QTL exists. This is done by the additional randomization test proposed in study IV, where the model including two epistatic QTL is compared to the model including no QTL. By using the genetic algorithm for the simultaneous searches in the permuted data sets, the computational demand for this test is reasonable for a single processor computer.

3. One of the QTL in the proposed pair is significant by its marginal effects

When one of the QTL in the pair is significant by its marginal effects, the test of no QTL versus two epistatic QTL can not be used, since it is already known that one of the QTL exists. The additional randomization test could become significant due to existence of an epistatic QTL pair, or due to a strong main effect of the significant QTL. We propose that the conditional randomization test (study IV) is used instead. This test compares the model including two epistatic QTL to the model with the marginal effects of the significant QTL.

By simulations we were able to show that the proposed strategy was powerful in detecting epistatic QTL pairs, and that it gave reasonable type I errors. Some concerns were raised with regard to the second dimension of multiple testing, where a QTL can be detected in either one of three consecutive tests (0 versus 1, 0 versus 2, and 1 versus 2 QTL). It needs to be evaluated whether it is necessary to correct the significance thresholds for the subsequent tests in order to avoid an elevated type I error.

Detection of epistatic QTL in an experimental line cross (V)

To further evaluate the properties of the procedure proposed in paper IV, we decided to use it for mapping of QTL in an F_2 intercross between the Red junglefowl and White Leghorn. The application of the method to experimental data, where the true genetic background of the traits is unknown, will indicate the practical importance of the method and help to identify topics that needs to be addressed for further improvents.

The analysed data set

An optimal data set for evaluation of a method to map epistatic QTL pairs would include large numbers of individuals in each of the genotype classes for a QTL pair, phenotypic data on high heritability traits, for which the lines used to generate the cross differ substantially, and a dense map of fully informative genetic markers. We have created an F_2 intercross between the Red junglefowl and a White Leghorn selection line consisting of 852 F_2 individuals. The individuals were genotyped for 104 genetic markers on 25 linkage groups. Phenotypic data has been collected for various traits, for example growth traits, for which the parental lines differ substantially. We expect to have more than 50 individuals in each two-locus genotype class and data for traits of medium to high heritability. The genetic map is currently relatively sparse (average marker distance 25 cM). Even though the data is not ideal, we decided to perform an analysis in this data set, based on the following. i) In the analysis of an experimental cross, it is common that a relatively sparse genetic map is used for the initial analysis, and that the map later will be filled in interesting regions. By performing analyses in a data set with a sparse map and later perform a follow-up study using a denser map would mimic how the method would be used in most practical situations. ii) By evaluating the genomic regions where the significant QTL were located with regard to inconsistencies (such as segregation distortion) and low information content could indicate if there is a risk of false positives due to deviations from the underlying assumptions.

Since the genetic map is rather sparse, the power to detect QTL in the middle of a marker bracket is decreased. One of the incentives of using this method is to identify QTL regions, which should be further characterised by molecular work. By lowering the significance threshold, the power to detect QTL in regions with low information content is increased. In this initial study we thus wanted to study both those QTL which were significant at a 5% and at a 20% genome-wide significance level. By including new markers in regions with low information content that contains suggestive QTL, these regions can be further examined using the new marker information and a more stringent significance threshold.

The computational demand of the mapping strategy

The randomization testing strategy proposed in study IV include multiple randomization tests for each trait. Before this strategy is used, the need for computational resources should be assessed. The calculations are based on i) the number of QTL identified by their marginal effects in the initial genome scan and the estimated number of additional epistatic QTL, ii) the number of individuals in the population, and iii) the number of cofactors to be included in the analysis

Based on our data set and previous analyses we have made the following assumptions and calculations. We would, on average, use 850 individuals in the analysis of each trait. On average four QTL were detected by their marginal effects for the growth traits (Kerje *et al.* manuscript). These QTL would contribute to 4 QTL x 2 marginal genetic effects = 8 columns in the design matrix, *X.* The four QTL detected by their marginal effects can form

$$\binom{n}{2} = \frac{n*(n-1)}{2} = \frac{4*3}{2} = 6$$

QTL pairs. If we assume that three of the six possible pairs interact, the additional contribution of these to the design matrix would be 3 pairs x 4 epistatic effects = 12 columns. The mean effect would be included together with the fixed effects of sex (2 levels) and batch (6 levels) in the analysis. The inclusion of these effects contributes with an additional 1 + (2-1) + (6-1) = 7 columns to the design matrix. We assume that 20 forward selection rounds are performed before convergence, and that the conditional and additional randomization tests are used in equal proportions in these steps. If we assume that 8 new QTL regions will be identified and that 10 detected QTL pairs interact, the total contribution to the design matrix would be: 8 new QTL x 2 marginal genetic effects = 16 columns with marginal genetic effects and 10 QTL pairs x 4 epistatic effects = 40 columns with epistatic effects. The average contribution to the design matrix would then be (40+16)/2=28 columns.

Based on this, we can predict the total computational demand of the testing procedure. Most of the computational time is spent in the additional randomization tests, and we therefore base the calculations entirely on the computational demand of this test. The average size of the design matrix used in the randomization tests would be 850 rows times 8 + 12 + 7 + 28 = 55 columns. This size of the design matrix would according to table 1 on average increase the computational demand by approximately 15 times that of a matrix with size 850 rows times 3 columns (which is the size used to map single QTL without cofactors and fixed effects). Each randomization test for an additional QTL pair would then based on the previously described performance of our analysis program take approximately 425 minutes for each test x 15 times increase in computing time = 6,375 minutes. We would analyse 9 traits and perform 10 randomization tests each of these traits, which would lead to an expected requirement of CPU time of: 9 traits x 10 randomization tests x 6,375 minutes per test =573,750 CPU minutes = 9,563 CPU hours.

The computational demand corresponds to approximately one year of computations on a single processor machine or 20 days on 20 processors on a parallel machine. This initial study does not justify such an extensive use of computational resources, and we thus decided to use a slightly modified version of the proposed strategy.

A modified mapping strategy

The strategy we describe in study IV detects epistatic QTL by forward selection of simultaneously mapped epistatic QTL pairs. To decrease the computational demand, we instead used a modified version of the strategy. A comparison of the original and the modified strategies are given in Figure 7. The principle of the modified strategy is to first map QTL by their marginal additive and dominance effects using forward selection. Significance testing is performed by a randomization test for each step in the forward selection procedure. Secondly, use an exhaustive search to evaluate the model fit of all possible epistatic QTL pairs. Significance levels for model selection are derived from the conditional and additional randomization tests described in study IV. Thirdly, select the model for each of the significant QTL pairs using the model selection randomization test described in study IV.

The updated strategy uses one conditional randomization test for each QTL that was significant as a single QTL and one additional randomization to test for epistatic QTL pairs where none of the QTL had significant marginal effects. This significantly decreases the computational demand of the strategy. On the other hand, the updated strategy has a lower power to detect epistatic QTL, since it does not correct for significant background genetic effects. By using a 20% genome-wide threshold, we hope to identify most QTL regions that would be significant by the original strategy, and thus we could include new markers in these regions. Further studies using the original strategy in the same experimental cross will help to address the importance of using the full strategy to obtain maximal power of the mapping procedure.

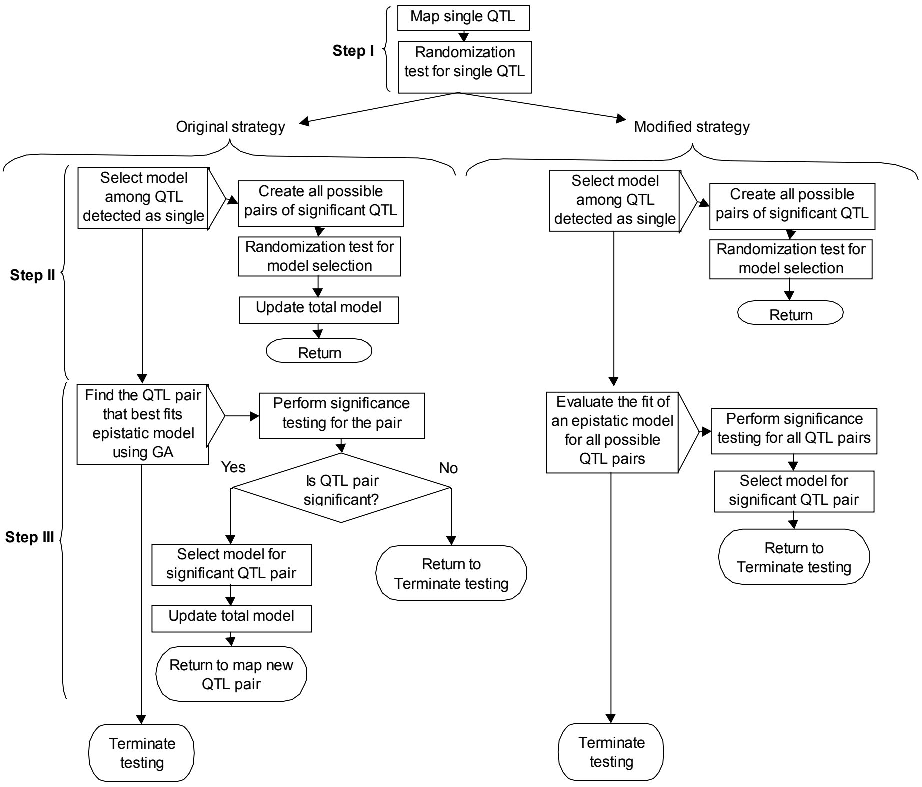
Step I
Map single QTL
Randomization test for single QTL
Original strategy
Modified strategy
Step II
Select model among QTL detected as single
Create all possible pairs of significant QTL
Randomization test for model selection
Update total model
Return
Select model among QTL detected as single
Create all possible pairs of significant QTL
Randomization test for model selection
Return
Step III
Find the QTL pair that best fits epistatic model using GA
Perform significance testing for the pair
Yes
No
Is QTL pair significant?
Select model for significant QTL pair
Update total model
Return to map new QTL pair
Return to Terminate testing
Terminate testing
Evaluate the fit of an epistatic model for all possible QTL pairs
Perform significance testing for all QTL pairs
Select model for significant QTL pair
Return to Terminate testing
Terminate testing

Figure 7. A comparison of the original and modified randomization testing strategies.

Summary of results

We found 21 significant loci contributing to the nine evaluated growth traits using a 5% genome wide significance threshold. Eleven of these were only found by the simultaneous search. When using a 20% genome wide significance level, 14 additional QTL were detected, eight of those could only be detected by simultaneous mapping. The relative contribution of epistasis was more pronounced for early growth (prior to 46 days of age), whereas additive genetic effects explained the major portion of the genetic variance later in life. Several of the detected loci affected either early or late growth but not both. Very few loci affected the entire growth process, which points out that early and late growth, at least to some extent, have different genetic regulation. The amount of residual variation explained was also increased considerably for most of the growth traits. Partitioning of the genetic variance showed that epistasis was considerably more important for early growth, than for late growth.

Once applied, the mapping strategy appears to have high power to detect epistatic QTL. The number of significant QTL regions that influence growth in the Red junglefowl x White Leghorn cross was more than doubled when simultaneous mapping and randomization testing were used.

Conclusions

Our work has shown that high performance computing is a valuable tool in QTL mapping. The use of modern search algorithms and parallel computing have made simultaneous searches and randomization tests for interacting QTL computationally feasible. Development of efficient software for QTL mapping is advantageous both for making new mapping methodologies available to geneticists as well as for shortening the time from data collection to the completion of the statistical evaluation. This gives users of this technology a competitive edge in the field of genetical research.

We have also shown that the use of genetic models, including imprinting or epistasis, can increase the power in QTL mapping experiments. The use of these methods also gives new insights into the genetics underlying important traits in farm animals and other organisms.

Future prospects

In progress

Implementation of the least squares solver described by Ljungberg *et al*. (2002), should make it possible to use the original strategy described in study IV to map QTL in the Red junglefowl x White Leghorn chicken intercross. The importance of using forward selection of epistatic QTL and including genetic background effects in the genetic model can thus be addressed.

We have in study V started to evaluate the sensitivity of epistatic and non-epistatic models to inconsistencies in the data and low information content in the genome. By performing new analyses in an updated data set, where new genetic markers are typed in putative QTL regions with low information content, we will be able to evaluate the sensitivity of the epistatic QTL mapping method to inconsistencies in the analysed data sets.

Immediate future

We have in study III and IV proposed that an epistatic model should be used to simultaneously search and statistically test for QTL pairs. In this procedure we have used a two locus non-orthogonal genetic model. It would be interesting to evaluate whether the use of alternative epistatic genetic models (e.g. the orthogonal model described by Cockerham 1954) would give similar results in the mapping procedure we propose in study IV.

The use of an epistatic genetic model in significance testing for proposed QTL pairs decreases the power to detect non-interacting QTL, since several unnecessary parameters are included in the testing procedure. If a marginal effects model would be used instead, the power to detect non-interacting QTL should increase. Further studies could be performed to evaluate how much additional power that could be gained by simultaneous mapping of non-interacting QTL. It would also be interesting to evaluate a strategy where QTL pairs are sought using an epistatic model, and that model selection is used to select an appropriate model for each QTL pair prior to significance testing. Based on the selected model, randomization testing would then be performed using either a marginal effects or an epistatic QTL model. This strategy should increase the power to detect QTL by using an appropriate model when testing for the QTL, but has the drawback of introducing another level of multiple testing.

In study III, we proposed the use of a genetic algorithm for simultaneous mapping of epistatic QTL pairs. The search was limited to QTL pairs, since we believe that a simultaneous search is most important when epistatic QTL are sought, and the currently available data sets are not large enough to handle genetic models including third order epistasis. To fully explore future larger data sets it is necessary to evaluate genetic algorithms, or other search algorithms, in simultaneous searches for epistatic QTL in higher dimensions. Simultaneous mapping might also increase the power to map multiple QTL, even if only two-way epistasis is included in the genetic model. To make this computationally feasible, the evaluation of alternative search algorithms is necessary.

One important topic in QTL mapping is to infer the dimension of the mapping problem, i.e. deciding how many QTL that affect the trait. We have proposed randomization testing, and other authors have proposed analytical methods (Broman 1997; Zeng *et al.* 2000) for this. Comparative studies of the proposed methods under various genetic models will be necessary to design powerful methods for comparing models with different numbers of QTL and genetic parameters for the QTL.

Most analyses of outbred line crosses are today based on the method described by Haley *et al.* (1994), which assumes that alternative QTL alleles are fixed within lines. If QTL alleles segregate within the lines, the QTL effects will be confounded with the allele frequencies in the parental lines. Is a small QTL effect caused by a small QTL which is fixed in the parental lines, or a large QTL that is still segregating? A further complication when mapping QTL in outbred lines is that there might exist multiple alleles for the same QTL. This will decrease the power to detect the QTL, but also cause serious misinterpretations of the results from the QTL mapping study. When denser genetic maps becomes available, it will be possible to identify individual haplotypes which segregate within populations, and thus increase the need for genetic models that can handle this situation.

Looking ahead

QTL mapping is performed with various statistical methods in a wide range of populations. The mapping population and statistical method selected for a specific study depends on the aim of the study. Is the aim to increase the understanding of the genetic background of a quantitative trait, or to detect QTL affecting a population specific phenotype? When considering the analysis methodology, three major approaches can be identified and I will here briefly introduce them.

First, there are the association-based tests, where the effects of individual candidate genes or genetic markers are evaluated in population samples. The design of association studies and the methodology used for statistical analysis is rather general and can therefore be used in most species and populations. Today, the power of these tests is limited, since the currently used density of the markers in these studies is too low to cover the genome, since linkage disequilibrium only span short genomic segments in most outbreeding populations. The advantage of the approach is mainly that the detected QTL have a direct effect on the expression of the quantitative phenotype in the studied population. This is important in medicine or agriculture to find loci affecting disease phenotypes or various production traits.

Secondly, there are the linkage-based tests, where analyses are generally performed among related individuals in the population. More closely related individuals share larger linked segments of the genome, and this linkage disequilibrium can be used to detect QTL. A relatively sparse genetic map of linked genetic markers can cover large portions of the genome and increase the power to detect QTL. Various methods have been proposed for linkage mapping in different population designs. The power to detect QTL within the analysed populations is higher than for the association-based tests. Major disadvantages with the method are that the relative importance of the QTL on a population level remains unknown and that the resolution of the QTL position is low.

Linkage disequilibrium (LD) mapping combines association and linkage mapping. The method aims to detect associations between haplotypes of linked genetic markers and the phenotype. Linkage is used to indicate location, and association tests are used to indicate the strength of the association. The method uses the linkage disequilibrium that exists in short genomic segments, even in outbred population samples. A very dense map of genetic markers is necessary to utilize this linkage, and based on the population history, the method can provide a very high resolution of the QTL location.

The human genome sequence has recently become available to the public. One of the current efforts of the human genome project is to sequence individuals from several different ethnic groups to identify the major genome wide haplotypes in the human population. The sequencing of the genome of several farm animal and experimental animal species is also in progress, and we can expect to obtain the same haplotype information for these species as well. The genome projects will also develop dense genetic maps, with a genetic marker every 100 kb or closer, throughout the genome. By using the dense genetic maps, it will be possible to decide the genome-wide haplotype for each individual in a population. In ten years from now, the cost for genotyping will have decreased to levels where it is possible to perform genome-wide genotyping of all individuals in large populations. It will then be possible in practise to perform QTL mapping experiments in populations containing all individuals in a species that have phenotypic observations using LD mapping. These studies will have very high power to map QTL with high resolution. One could e.g. consider mapping of QTL for growth or fat deposition in the entire pig breeding population of the world!

The ability to collect genetic data from dense genetic maps in large populations calls for QTL mapping methods that can handle genome wide linkage disequilibrium mapping of QTL. Today, methods exist for detection of single QTL using LD mapping. In order to fully utilize the information in the large population, several issues needs to be addressed. For a better understanding of the genetics behind multifactorial traits, methods for genome wide LD screens for multiple QTL are needed. Genetic models needs to be developed, which can simultaneously include multiple alleles at multiple QTL. These models will help improve the understanding about the contribution of genetic polymorphism at individual or groups of loci to the total genetic variance. These analyses will be possible based on the knowledge about the haplotypes that exist in the population. Model selection and significance testing will be key issues, due to the multiple testing that inevitably will be performed. High performance computing will be a necessity for the analyses of the huge data sets. Once the methods are available, they will be of general interest since they would be applicable to a wide range of existing populations in various species. It is therefore of general interest to the whole QTL mapping community to make a joint effort in improving these methods. Based on the wide range of competence's needed in this effort, it is an absolute necessity to establish inter-disciplinary collaborations to develop the QTL mapping methods of the future.

References

Akaike, H. 1969. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* 21, 243-247.

Anderson, E., Bai, D., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. & Sorensen, D. 1999. LAPACK users' guide, Third edition. SIAM Publications, Philadelphia, PA.

Anderson, E., Brooks, J. & Hewitt, T. 1997 The Benchmarker's Guide to Single-processor Optimization for CRAY T3E Systems. Cray Research. Available at http://www.cray.com/products/systems/crayt3e/benchmark.ps.

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

Andersson, L., Haley, C.S., Ellegren, H., Knott, S.A., Johansson, M., Andersson, K., Andersson-Eklund, L., Edfors-Lilja, I., Fredholm, M., Hansson, I., *et al.* 1994. Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science* 263, 1771-1774.

Andersson-Eklund, L., Marklund, L., Lundstrom, K., Haley, C.S., Andersson, K., Hansson, I., Moller, M. & Andersson, L. 1998. Mapping quantitative trait loci for carcass and meat quality traits in a wild boar x Large White intercross. *Journal of Animal Science* 76, 694-700.

Andersson-Eklund, L., Uhlhorn, H., Lundeheim, N., Dalin, G. & Andersson, L. 2000. Mapping quantitative trait loci for principal components of bone measurements and osteochondrosis scores in a wild boar x large white intercross. *Genetical Research* 75, 223-230.

Araujo, H. & Bier, E. 2000. *sog* and *dpp* exert opposing maternal functions to modify *toll* signaling and pattern the dorsoventral axis of the *Drosophila* embryo. *Development* 127, 3631-3644.

Beaudoin, N., Serizet, C., Gosti, F. & Giraudat, J. 2000. Interactions between abscisic acid and ethylene signaling cascades. *Plant Cell* 12, 1103-1116.

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

Björck, Å. 1996. Numerical methods for liast squares problems. SIAM Publications, Philadelphia, PA.

Bolland, S. & Ravetch, J.V. 2000. Spontaneous autoimmune disease in Fc(gamma)RIIB-deficient mice results from strain-specific epistasis. *Immunity* 13, 277-285.

Broman, K. W. 1997. Identifying quantitative trait loci in experimental crosses. Ph.D. thesis, Department of Statistics, University of California, Berkeley.

Broman, K.W. 2001. Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal* 20, 44-52.

Cardon, L.R. & Bell, J.I. 2001. Association study designs for complex diseases. *Nature Reviews Genetics* 2, 91-99.

Carlborg, Ö., Andersson, L. & Kinghorn, B. 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting Quantitative Trait Loci. *Genetics* 155, 2003-2010.

Chase, K., Adler, F.R. & Lark, K.G. 1997. Epistat: a computer program for identifying and testing interactions between pairs of quantitative trait loci. *Theoretical and Applied Genetics* 94, 724-730.

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

Cockerham, C.C. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics* 39, 859-882

Damerval, C., Maurice, A., Josse, J.M. & de Vienne, D. 1994. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. *Genetics* 137, 289-301.

Darvasi, A. 1998. Experimental strategies for the genetic dissection of complex traits in animal models. *Nature Genetics* 18, 19-24.

de Koning, D.J., Rattink, A.P., Harlizius, B., van Arendonk, J.A., Brascamp, E.W. & Groenen, M.A. 2000. Genome-wide scan for body composition in pigs reveals important role of imprinting. *Proceedings National Academy of Sciences USA* 97, 7947-7950

Doerge, R.W. 2002. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3, 43-52.

Edfors-Lilja, I., Wattrang, E., Marklund, L., Moller, M., Andersson-Eklund, L., Andersson, L. & Fossum, C. 1998. Mapping quantitative trait loci for immune capacity in the pig. *Journal of Immunology.* 161, 829-835.

Edwards, M.D., Stuber, C.W. & Wendel, J.F. 1987. Molecular-marker-facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics* 116, 113-125.

Falconer, D.S. & Mackay, T.F.C. 1996. Introduction to quantitative genetics 4th edition. Longman Group Limited, Essex, England.

Fijneman, R.J., De Vries, S.S., Jansen, R.C. & Dermant, P. 1996. Complex interactions of new quantitative trait loci, *Sluc1*, *Sluc2*, *Sluc3*, and *Sluc4*, that influence the susceptibility to lung cancer in the mouse. *Nature Genetics* 14, 465-467.

Flint, J. & Mott, R. 2001. Finding the molecular basis of quantitative traits: successes and pitfalls. *Nature Reviews Genetics* 2, 437-445.

Fridlyand, Y.J.M. 2001. Resampling methods for variable selection and classification: application to genomics. Ph.D. thesis, Department of Statistics, University of California, Berkeley.

Goldberg, D.E. 1989. Genetic algorithms in search, optimization and machine learning. Addison and Wesley, Reading, MA.

Grobet, L., Martin, L.J., Poncelet, D., Pirottin, D., Brouwers, B., Riquet, J., Schoeberlein, A., Dunner, S., Menissier, F., Massabanda, J., Fries, R., Hanset, R. & Georges, M. 1997. A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nature Genetics* 17, 71-74.

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

Haley, C.S., Knott, S.A. & Elsen, J.M. 1994. Mapping Quantitative Trait loci in crosses between outbred lines using least squares. *Genetics* 136, 1195—1207.

Hannan, E.J. & Quinn, B.G. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B.* 41, 190-195.

Hoeschele, I., Uimari, P., Grignola, F.E., Zhang, Q. & Gage, K.M. 1997. Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics* 147, 1445-1457.

Holland, J.B. 1998. EPISTACY: a SAS program for detecting two-locus epistatic interactions using genetic marker information. *Journal of Heredity* 89, 374-375.

Holland, J.B., Moser, L.S., O'Donoughue, L.S. & Lee, M. 1997. QTLs and epistasis associated with vernalization responses in oat. *Crop Science* 37, 1306-1316.

Jannink J.L. & Jansen R. 2001. Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* 157, 445-454.

Jansen, R. C. 1992. A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* 85, 252-260.

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

Jansen, R. C. & Stam, P. 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* 136, 1447-1455.

Kao, C.-H. 2000. On the differences between maximum likelihood and regression interval mapping in the analysis of Quantitative Trait Loci. *Genetics* 156, 855-865.

Kao, C.–H. & Zeng, Z.-B. 1997. General formulae for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM-algorithm. *Biometrics* 53, 653-665.

Kao, C.-H., Zeng, Z.-B. & Teasdale, R. 1999 Multiple interval mapping for quantitative trait loci. *Genetics* 152, 1203-1216.

Kao, C.-H. & Zeng, Z-.B. 2002. Modeling epistasis of Quantitative Trait Loci using Cockerham's model. *Genetics* 160, in press.

Khazanehdari, K.A. & Borts, R.H. 2000. EXO1 and MSH4 differentially affect crossing-over and segregation. *Chromosoma* 109, 94-102.

Kim, J.H., Sen, S., Avery, C.S., Simpson, E., Chandler, P., Nishina, P.M., Churchill, G.A. & Naggert, J.K. 2001. Genetic analysis of a new mouse model for non-insulin-dependent diabetes. *Genomics* 74, 273-286.

Knott, S.A., Marklund, L., Haley, C.S., Andersson, K., Davie, W., Ellegren, H., Fredholm, M., Hansson, I., Hoyheim, B., Lundström, K., Moller, M. & Andersson, L. 1998. Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* 149, 1069-1080.

Kruglyak, L. & Lander, E.S. 1995. High-resolution genetic mapping of complex traits. *American Journal of Human Genetics* 56, 1212-1223.

Lander, E.S. & Botstein, D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121, 185-199.

Leips J. & Mackay, T.F. 2000. Quantitative trait loci for life span in *Drosophila melanogaster:* interactions with genetic background and larval density. *Genetics* 155, 1773-1788.

Li, Z., Pinson, S.R., Park, W.D., Paterson, A.H. & Stansel, J.W. 1997. Epistasis for three grain yield components in rice (*Oryza sativa L.*). *Genetics* 145, 453-65.

Ljungberg, K., Holmgren, S. & Carlborg, Ö. 2002. Efficient kernel algorithms for QTL mapping problems. Technical Report No. 2002-005, Department of Information Technology, Uppsala University.

Long, A.D., Mullaney, S.L., Mackay, T.F. & Langley, C.H. 1996. Genetic interactions between naturally occurring alleles at quantitative trait loci and mutant alleles at candidate loci affecting bristle number in *Drosophila melanogaster*. *Genetics* 144**,** 1497-1510.

Luschnig, S., Krauss, J., Bohmann, K., Desjeux, I. & Nusslein-Volhard, C. 2000. The *Drosophila* SHC adaptor protein is required for signaling by a subset of receptor tyrosine kinases. *Molecular Cell* 5, 231-241.

Lynch, M. & Walsh, B. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA, USA.

Mackay, T. 2001. Quantitative Trait Loci in *Drosophila*. *Nature Reviews Genetics* 2, 11-21.

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

Mallows, C.L. 1973. Some comments on *Cp*. *Technometrics* 15, 661-675.

Martinez, O. & Curnow, R.N. 1992. Estimating the locations and the sizes of effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* 85, 480-488.

Nettleton, D. & Doerge, R.W. 2000. Accounting for variability in the use of permutation testing to detect quantitative trait loci. *Biometrics* 56, 52-58.

Nezer, C., Moreau, L., Brouwers, B., Coppieters, W., Detilleux, J., Hanset, R., Karim, L., Kvasz, A., Leroy, P. & Georges, M. 1999. An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nature Genetics* 21, 155-156.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. 1999. Numerical recipes in Fortran: The art of scientific computing 2nd edition. Cambridge University Press, New York.

Rattink, A.P., de Koning, D.J., Faivre, M., Harlizius, B., van Arendonk, J.A. & Groenen, M.A. 2000. Fine mapping and imprinting analysis for fatness trait QTLs in pigs. *Mammalian Genome* 11, 656-661.

Reik, W. and Walter, J. 2001. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics* 2, 21-32.

Satagopan, J.M., Yandell, B.S., Newton, M.A. & Osborn, T.C. 1996. A bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* 144, 805-816.

Scanga S.E., Ruel, L., Binari, R.C., Snow, B., Stambolic, V., Bourchard, D., Peters, M., Calvieri, B., Mak, T.W., Woodgett, J.R. & Manoukian, A.S. 2000. The conserved PI3'K/PTEN/Akt signaling pathway regulates both cell size and survival in *Drosophila*. *Oncogene* 19, 3971-3977.

Schwartz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.

Sen, S. & Churchill, G.A. 2001. A statistical framework for quantitative trait mapping. *Genetics* 159, 371-387.

Shimomura, K., Low-Zeddies, S.S., King, D.P., Steeves, T.D., Whiteley, A., Kushla, J., Zemenides, P.D., Lin, A., Vitaterna, M.H., Churchill, G.A. and Takahashi, J.S. 2001. Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Research* 11, 959-980.

Shook, D.R. & Johnson, T.E. 1999. Quantitative trait loci affecting survival and fertility-related traits in *Caenorhabditis elegans* show genotype-environment interactions, pleiotropy and epistasis. *Genetics* 153, 1233-43.

Sillanpää, M.J. & Arjas, E., 1999. Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* 151, 1605-1619.

Southey, B.R. & Fernando, R.L. 1998. Controlling the proportion of false positives among significant results in QTL detection. *Proceedings of the 6th World Congress of Genetics Applied to Livestock Production*. Armidale, NSW, Australia. 26, 341-244.

Sugiyama, F., Churchill, G.A., Higgins, D.C., Johns, C., Makaritsis, K.P., Gavras, H. & Paigen, B. 2001. Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics* 71, 70-77.

Seyffert W. 1966. Die Simulation quantitativer Merkmale durch Gene mit biochemisch definierbarer Wirkung. I. Ein Einfaches modell. *Züchter* 36, 159-163.

Varona, L., Raya, L.G., Rauw, W.M. & Noguera, J.L. 2001. Can F_2 mapping experiments be used to detect epistatic interactions? 7th Quantitative trait locus mapping and marker-assisted selection workshop. Universidad Politécnica de Valencia, Spain.

Wang, D.L., Zhu, J., Li, Z.K. & Paterson, A.H. 1999. Mapping QTLs with epistatic effects and QTL x environment interactions by mixed linear model approaches. *Theoretical an Applied Genetics* 99, 1255-1264.

Wolfe, K. 2000. Robustness-it's not where you think it is. *Nature Genetics* 25, 3-4.

Zeng, Z.-B. 1993a. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Science of the USA* 90, 10972-10976.

Zeng, Z.-B. 1993b. Precision mapping of quantitative trait loci. *Genetics* 136, 1457-1468.

Zeng, Z.-B., Kao, C.-H. & Basten, C.J. 1999. Estimating the genetic architecture of quantitative traits. *Genetical Research* 74, 279-289.

Zeng, Z.-B., Liu, J., Stam, L., Kao, C.-H., Mercer, J.M. & Laurie, C.C. 2000. Genetic architecture of a morphological difference between two *Drosophila* species. *Genetics* 154, 299-310.

Acknowledgements

The studies presented in this thesis were performed at the Department of Animal Breeding and Genetics at the Swedish University of Agricultural Sciences. The studies were supported by the National Graduate School in Scientific Computing (NGSSC), the Food21 (Mat21) project, the Medical Research Council (MFR) and the Swedish University of Agricultural Sciences, whose grants are gratefully acknowledged. Computer time and excellent technical support has been provided by the National Supercomputing Centre (NSC) at Linköping University. I am also grateful to the Uppsala Graduate School in Biomedical Sciences (UGSBR) for giving me an excellent introduction to post-graduate studies.

I would like to thank all my past and current colleagues at the Department of Animal Breeding and Genetics, SLU and the Department of Animal Science, University of New England, Australia for providing a creative work environment and for all kinds of discussions. Special thanks are due to my office mates, my closest collaborators in the chicken project, as well as Dr W. I would also like to thank all our other collaborators and people I have met during courses and conferences in Sweden and around the world for all fruitful discussions. In addition to this, I would like to express my special gratitude to

My main supervisor Prof. Leif Andersson for being an excellent guide in the world of science. Your sincere interest, constructive criticism, open mind, pragmatism, enthusiasm, ability to give free rains and so much more, has stimulated me to grow as a person and made our scientific collaboration fruitful.

My co-advisor Ass. Prof. Lena Andersson-Eklund for invaluable supervision and guidance as well as helping to plan and pursue my PhD education. Neither this thesis nor I would be the same without you!

My Australian supervisors Professor Brian Kinghorn and Dr Julius van der Werf, for supporting me during my first stumbling, autonomous scientific steps.

Dr Sverker Holmgren for helping me to get started in the field of scientific computing, and to write the application that became the foundation of the scientific work in this thesis.

Dan Fjellström for giving me an outstanding introduction to science, a motivation to pursue science, and for all our other valuable discussions.

My parents, Per-Erik and Agneta, for all their support and for always stressing that knowledge is important.

Last, but not least, I would like to thank Anna and Sara for being the joy of my life, the light in my tunnel, my past, my present and my future.