UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 973

Two Optimization Problems in Genetics

Multi-dimensional QTL Analysis and Haplotype Inference

CARL NETTELBLAD

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2012

ISSN 1651-6214
ISBN 978-91-554-8473-6
urn:nbn:se:uu:diva-180920

Dissertation presented at Uppsala University to be publicly examined in Room 2446, Polacksbacken, Lägerhyddsvägen 2D, Uppsala, Friday, October 26, 2012 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Nettelblad, C. 2012. Two Optimization Problems in Genetics: Multi-dimensional QTL Analysis and Haplotype Inference. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 973. 57 pp. Uppsala. ISBN 978-91-554-8473-6.

The existence of new technologies, implemented in efficient platforms and workflows has made massive genotyping available to all fields of biology and medicine. Genetic analyses are no longer dominated by experimental work in laboratories, but rather the interpretation of the resulting data. When billions of data points representing thousands of individuals are available, efficient computational tools are required. The focus of this thesis is on developing models, methods and implementations for such tools.

The first theme of the thesis is multi-dimensional scans for quantitative trait loci (QTL) in experimental crosses. By mating individuals from different lines, it is possible to gather data that can be used to pinpoint the genetic variation that influences specific traits to specific genome loci. However, it is natural to expect multiple genes influencing a single trait to interact. The thesis discusses model structure and model selection, giving new insight regarding under what conditions orthogonal models can be devised. The thesis also presents a new optimization method for efficiently and accurately locating QTL, and performing the permuted data searches needed for significance testing. This method has been implemented in a software package that can seamlessly perform the searches on grid computing infrastructures.

The other theme in the thesis is the development of adapted optimization schemes for using hidden Markov models in tracing allele inheritance pathways, and specifically inferring haplotypes. The advances presented form the basis for more accurate and non-biased line origin probabilities in experimental crosses, especially multi-generational ones. We show that the new tools are able to reconstruct haplotypes and even genotypes in founder individuals and offspring alike, based on only unordered offspring genotypes. The tools can also handle larger populations than competing methods, resolving inheritance pathways and phase in much larger and more complex populations. Finally, the methods presented are also applicable to datasets where individual relationships are not known, which is frequently the case in human genetics studies. One immediate application for this would be improved accuracy for imputation of SNP markers within genome-wide association studies (GWAS).

Keywords: quantitative trait loci, genome-wide association studies, hidden Markov models, numerical optimization, linkage analysis, haplotype inference, genotype imputation, high performance computing

Carl Nettelblad, Uppsala University, Department of Information Technology, Division of Scientific Computing, Box 337, SE-751 05 Uppsala, Sweden. Department of Information Technology, Computational Science, Box 337, SE-751 05 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-8473-6
urn:nbn:se:uu:diva-180920 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-180920)

To Whom It May Concern

List of papers

This thesis is based on the following papers, which are referred to in the text by their roman numerals. The papers are ordered in the themes of modelling, model fitting in QTL searches, and Markov methods for tracing inheritance in populations.

I Carl Nettelblad, Örjan Carlborg, Ania Pino-Querido, and José M. Álvarez-Castro. Coherent estimates of genetic effects with missing information. Open Journal of Genetics, 2:31-38, 2012.

II Carl Nettelblad, Behrang Mahjani, and Sverker Holmgren. Fast and Accurate Detection of Multiple QTL. *Submitted.*

III Mahen Jayawardena, Carl Nettelblad, Salman Toor, Per-Olov Östberg, Erik Elmroth, and Sverker Holmgren. A Grid-Enabled Problem Solving Environment for QTL Analysis in R. In Proc. 2nd International Conference on Bioinformatics and Computational Biology (BICoB 2010). ISBN 987-1-880843-76-5.

IV Carl Nettelblad, Sverker Holmgren, Lucy Crooks, and Örjan Carlborg. `cnF2freq`: Efficient Determination of Genotype and Haplotype Probabilities in Outbred Populations using Markov Models. In Proceedings to BICoB 2009, New Orleans, LNBI 462, 307-319, 2009.

V Lucy Crooks, Carl Nettelblad, and Örjan Carlborg. An Improved Method for Estimating Chromosomal Line Origin in QTL Analysis of Crosses Between Outbred Lines. G3 1(1):57-64, 2011.

VI R.M. Nelson, C. Nettelblad, L. Crooks, M.E. Pettersson, F. Besnier, X. Shen, J.M. Álvarez-Castro, L. Rönnegård, W. Ek, Z. Sheng, M. Kierczak, S. Holmgren, and Ö. Carlborg. `MAPfastR`: QTL mapping in outbred line crosses. *Submitted.*

VII Carl Nettelblad. Haplotype inference based on Hidden Markov Models in the QTL-MAS 2010 multi-generational dataset. In BMC Proceedings 5(Suppl 3):S10, 2011.

VIII Carl Nettelblad. Inferring haplotypes and parental genotypes in larger full sib-ships and other pedigrees with missing or erroneous genotype data. Technical Report 2012-026, Department of Information Technology, Uppsala University, 2012. *Submitted.*

IX Carl Nettelblad. Breakdown of methods for phasing and imputation in the presence of double genotype sharing. Technical Report 2012-027, Department of Information Technology, Uppsala University, 2012. *Submitted*.

Reprints were made with permission from the publishers.

Conference attendance: The material of Paper VIII was also presented at the QTL-MAS 2011 workshop in Rennes, France. Further unpublished developments of the haplotyping were presented at the 4th International Conference in Quantitative Genetics in Edinburgh, UK in 2012 as a poster. Another poster relating to Paper II was presented at the same time. A preliminary version of that work was also presented orally at the SIAM Conference on Computational Science and Engineering 2009 in Miami, Florida. All papers appearing in conference proceedings were also presented orally at the respective conferences by the author of this thesis.

Software availability: The software developed during the course of the thesis work comprises an implementation of the updated PruneDIRECT algorithm, available on request, as well as different versions of the `cnF2freq` codebase that was initiated by this author. That code is available as a separate R package, and as a crucial component within the MAPfastR package. The code is available under BSD or LGPL licenses, but proper citation is naturally expected for academic use. The author is offering support in adapting the codebase for new datasets, time permitting. Available at:
`https://r-forge.r-project.org/projects/cnf2freq/`
`http://www.it.uu.se/research/project/ctrait`
`https://r-forge.r-project.org/projects/mapfastr/`

Contributions: Paper I is based on observations made during this author's master's thesis work. The main theory and experiments were developed by him and Dr. Álvarez-Castro, who drafted the main text in the final manuscript. Paper III was coordinated and mainly developed in close collaboration between this author and the first author, Dr. Jayawardena. This author also made all adaptations of the DIRECT code base and the main software design, as well as proofreading and writing some sections of the paper. The transform underlying paper II was suggested by this author, together with the applications described in the paper. The theory for correcting for finite-size populations was developed jointly by this author and Mr. Mahjani.

Paper IV and the underlying software were drafted by this author, with testing and manuscript feedback from the remaining authors. This author provided advice on the underlying methods as well as design and interpretation of experiments for Paper V. Code from this author forms a crucial part in the package presented in Paper VI. This author has also edited relevant sections of the paper. The work in Papers VII, VIII, and IX was designed, developed, and presented independently by this author.

Contents

Introduction

> Le mieux est l'ennemi du bien. [76]

The use of mathematics to treat and analyze scientific problems has brought immense success to a wide variety of fields, both in terms of knowledge and in terms of technological advances. Such a revolution is now also reaching biology.

Some of the problems arising in biology are similar to those traditionally found in physics, where simulating reality will answer the question. If you want to understand how a cell division takes place, simulation can plausibly be used as a versatile time-stepping microscope and the process can be studied, just like the air flow around a wing section or the development of the climate system. Parameters can be changed in a simulation far more easily than reality can be changed. By comparing simulation with reality, the underlying model can also be verified. If the model is expressed in a reasonable form, the model itself can also add to the wealth of human knowledge.

However, not all problems lend themselves to simulation. There can be insufficient information to set up a realistic model, or maybe a relevant simulation similar to physical reality would use too much computing time. Rather than trying to model reality as a set of equations as disciples of Newton, we can instead mathematically express what problem we truly want to solve. Frequently, a simulation is performed with actual intent to investigate a specific aspect of a phenomenon. It can therefore make sense to state the problem in a form focusing on the analysis of that aspect, rather than the full system.

Ideally, our way to state the problem should also, in some sense, lead to a way of attacking it. Stating what an optimal solution should look like is not of practical use, if there is no way to find such an optimal solution (and the total count of possible solutions precludes testing them all). If we can also state what should characterize a solution that is in some sense close to the optimal one, such a description can be used to gradually find a solution, each step bringing us closer to the final goal.

This thesis is concerned with two problems in biology, which have in common that the successful approaches are based on iterative optimization of different kinds. Since the subject is scientific computing, any hurdle relevant to solving the computational problem is of interest. That can mean developing software or improving relevant aspects of theory, as well as designing optimization schemes, including their practical computer implementations. All aspects are present in this work.

The two problems are, briefly: 1. Finding the set of genetic positions that together most fully describe the genetic variation in some trait; 2. Identifying and tracing inheritance of related genetic regions, in face of uncertainty in experimental data. For both of these problems, we purport that descriptions which lend themselves to gradual improvements or refinements of a solution are attractive when one actually wants to solve them using computational resources.

The contributions made have focused on finding sets of multiple interacting genetic positions through attempts to develop theory that would simplify interpretation and automatic evaluation of the models, and infrastructure for performing the actual searches on modern distributed computing resources. We have also developed theory that greatly enhances efficiency while not sacrificing correctness. The work for problem 1 is mainly covered in Papers I, II, III. The remaining papers and otherwise unpublished work mentioned in Chapter 2 cover tracing inheritance using deterministic hidden Markov models including an explicit parametrization of genetic phase.

The outline of the thesis follows the workflow of a genetic experiment. Biological prerequisite knowledge and a general background is presented in Chapter 1. Tracing inheritance in order to determine genetic similarity is treated in Chapter 2. The structure of linear regression models for identifying quantitative trait loci are described in Chapter 3. In Chapter 4, the methods for actually finding a correct set of such loci are presented. These chapters are followed by a review of the included papers, as well as a summary in Swedish.

1. Biological background

In the popular mind, DNA has become the symbol of the secret of biology. Advances in molecular methods mean that processes and phenomena that previously could only be studied externally, as a "black box", can now be controlled and unravelled in great detail. Efficient sequencing means that the full DNA structure, the genome, for sample individuals within species can almost routinely be determined [32, 52, 77, 78]. Another development is that sequencing of large numbers of individuals is also rapidly becoming routine. These sequencing results are then not only scientific pinnacles in themselves, but also form the basis for further work. Through the knowledge of the overall structure of the genome, further analyses such as comparisons between species and specific studies of correlations between genetic data and molecular properties are greatly simplified.

Such advances impose a paradigm shift on experimental workflows. While previously, the prudent method was to first decide a candidate substance, a candidate reaction, or a candidate gene, and then studying its biological role and variations, it is now possible to scan or probe complete systems. The scientific challenge is no longer a matter of designing and performing experiments for extracting enough data of enough quality to devise a model. It is rather a matter of how to analyze the data in a way that can find the networks, reactions and components that are *relevant*, within a cloud of noise.

The field of analysis of quantitative trait loci (QTL) is part of this paradigm shift. A *locus* is a genetic location within the genome. In each locus, different individuals can carry different *alleles* (gene versions). Different alleles can, but will not always, give rise to different observable properties, or *traits*. Sometimes, the expression (result) of an allele will be highly dependent on the environment. These environmental factors can be external, like chemical factors present or absent during the early development of an organism, or the quality of nutrients available. They can also be internal, in the sense of what other alleles are present in other loci. Therefore, to study a quantitative trait, i.e. a phenotype variable of some kind that can be measured for each individual, one wants to find a complete set or network of loci where the allelic variations can explain the variation seen in trait values between individuals.

What loci are found in such a *scan* will be dependent on the experimental setting. Loci that potentially have alleles greatly affecting a trait will not be found – if those alleles are absent or infrequent in the population studied. Likewise, external environmental factors might result in a genetic difference never rendering any observable results. For example, an inability or enhanced

ability to handle a specific nutrient (such as lactose for humans and bacteria), will only be apparent in an environment where that nutrient is available (i.e. in dairy products).

From a mathematical and statistical standpoint, the problem of QTL analysis is one of model fitting and model selection. The total number of loci included in a final model should not exceed those for which statistical significance is plausible. In other words, a repeated experiment, with nominally identical conditions, should result in the same loci being reported. In practice, linear regression models are frequently used. As complete high-quality genome sequencing is still not a cost-effective operation to be performed on hundreds of individuals in an experimental population, one resorts to analyzing genomic *markers*. The markers are located at known positions. To analyze loci not coinciding with the markers, the probability for a specific allele being present in the locus needs to be assessed based on the marker value. From such probabilities, the correlation between *genotype* and *phenotype* (trait) can be determined. Different sets of loci can be tested, to see what set will render the most faithful description of the observed data.

1.1 Chromosomes, DNA, Genes

The *genome* of an individual is primarily structured in *chromosomes*. Each chromosome contains two interlinked complementary molecules of deoxyribonucleic acid (DNA). These can be used to replicate each other, through the process of base-pairing allowed by the structure first described in [79]. Within a chromosome, the genetic code is a sequence of nucleotides, where each nucleotide is represented by one of the letters A, C, G, T. Pairing is restricted in such a way that A can only pair to T in the complementary strand, while C pairs to G. A single chromosome can consist of a sequence of millions of such *base pairs* (bp). A *mutation* in the genetic code can consist of a deletion of a stretch of base pairs, an insertion of base pairs originating from another region, a duplication of a short genetic segment, or the modification of a single base pair. Therefore, different individuals of a species tend to share the same genome structure on a macroscopic level, but there is not always a one-to-one correspondence for individual base pairs.

A *gene* is a specific region encoding some product, affecting the function of the cell. The *gene product* is frequently a protein, which is produced through transcription into messenger RNA, followed by translation into a protein in a ribosome external to the cell nucleus (in which DNA is preserved). However, the definition of a gene is not completely straightforward. In eukaryotic organisms (organisms that have a cell nucleus), the coding material is frequently "interrupted" by introns, base pair sequences that are not included in the actual mRNA and hence are not translated into a protein sequence. The structure of the introns, as well as surrounding genetic material, both upstream and

downstream relative to the protein-coding sequence, can also influence the practical expression of the gene in different ways.

For these reasons, a quantitative trait locus will frequently coincide with a gene, but it cannot be said beforehand that the genetic cause of trait variation will actually be found within the coding sequence. A gene can also sometimes be moved, *translocated* within the genome, so that it can appear at different loci. Such events will cause problems for the methods described in this thesis. Furthermore, a locus determined using these methods will frequently be wide, so that multiple closely linked genes are indicated as associated with the trait, making other methods necessary to discriminate between them. The process of refining a QTL position down to a specific gene or even specific mutations is called *fine mapping* [8].

1.1.1 Haploid and Diploid

Most of the genetic material in an organism from a species that practices sexual reproduction is duplicated, with one copy originating from each parent. This means that every chromosome exists in pairs. The genome in such a cell is called diploid. During the life cycle of an individual, every cell will preserve that separation, with one copy of maternal origin and another of paternal origin, replicated through each cell division (mitosis). Only in meiosis, when new gametocytes (oocytes and spermatocytes) are created, these two copies will intermingle. From the pairs of chromosomes, a new, recombined, genome is then created for each gametocyte, representing a unique halving of the parental genome. This genome, half the size of the genome in a conventional somatic cell, is called haploid, with one copy from each such pair of *homologous* chromosomes. The term *chromatid* is sometimes used to uniquely refer to one of the two copies within a chromosome pair.

This mixing of genetic material for each chromosome is called *recombination*. An individual instance where the source is switched is known as a *cross-over* event. The frequency of cross-overs is not constant over the genome. It has been shown that some sequences are related to triggering them. There are also macroscopic variations over the genome, recombination being more frequent in some regions compared to others.

In many settings, it is not relevant to distinguish the parental origin for a specific allele. Therefore, the diploid genotype is frequently stated as the two alleles found, but in no specific order with respect to the homologous chromosomes. However, in some cases *genomic imprinting* (chemical modifications on the chromosomes, specific to parental sex) results in the effect of an allele being dependent on from which parent that allele was transmitted. This is a specific case of epigenetic effects, i.e. hereditary effects not directly found in the genetic code. A well-known example of sexual imprinting in humans is the pair of Angelman syndrome and Prader-Willi syndrome, two hereditary

diseases caused by the very same genomic deviation, but with very different symptoms depending on the parental origin of the mutated variant of chromosome 15 [44].

If one wants to trace what regions of DNA have been transmitted unchanged over the course of several generations, it is naturally also of importance to separate the two homologous copies, as they are different and preserved with the exception of individual cross-overs. In such contexts, where the parental origin and linkage across the chromosome are relevant, the two separate genetic sequences are referred to as *haplotypes*, the genotypes that existed in the gametes that fused to give rise to the diploid offspring. Sometimes, a similar concept is also referred to as *phase*, with phase seen as switching at points of recombination.

1.1.2 Genetic Distances

From the perspective of this thesis, and QTL analysis in general, the chromosome can be seen as a continuous object. Distances are not measured in discrete base pairs, but rather in the unit centimorgan, cM. In a one centimorgan interval, the probability of crossing over taking place between the endpoints is 1%. The probability for some crossing to occur over a distance of 2 cM is then naturally 2%. However, there is a definite probability of double recombination, so that the detected origin will be identical to the source. Hence, the probability for a difference in sequence origin for a very large distance will approach 50%.

In the literature, several mapping functions have been suggested for the translation between recombination frequencies and genetic distances. The Morgan mapping function is completely linear by assuming no double recombinations. The Kosambi mapping function [46], on the other hand, includes modelling of the biologically known fact that a single cross-over tends to shadow the neighboring regions, making another recombination event nearby more unlikely than otherwise expected (interference). Kosambi distances are not additive, though.

The Haldane mapping function [27] is more complex than the Morgan function, in that there is no linear correlation between the net parity of recombination events and distance, but it is on the other hand simpler than the Kosambi function. For example, the Haldane mapping function is a true mathematical *distance function*, meaning among other things that the distance between A and C, traversing B, is equal to the sum of the distances AB and BC, a much needed property for modelling purposes. The Haldane function, defined for non-negative distances in cM, is given by

$$p = 0.5 + 0.5e^{-0.02x}, \tag{1.1}$$

where p is the probability of identical genetic phase (chromosome origin) at 0 and x. This relationship can be derived from realizing that the recombina-

tion process itself, assuming no interference, is a Poisson process, and then summing over the probability of an even state (i.e. a total of $0, 2, 4, 6...2k$ recombination events between the two points considered).

1.1.3 Markers

It is generally not feasible to study a haploid genome separately. Rather, the resulting genome in diploid cells in the offspring is studied, as well as the resulting phenotypes in that offspring. Full sequences are generally also not detected, instead specific markers are used. Markers are known locations of variability, which can be based on genetic repeats, including so-called microsatellites. A repeat is a rather short pattern that is repeated in multiple instances adjacent to each other in the genome sequence. Due to molecular details in the process of genome replication, the exact number of such repeats can change due to "slippage" [31]. Therefore, the number of microsatellite repeats in one locus can vary even between closely related individuals.

Another class of markers is that of single-nucleotide polymorphisms (SNPs). These are point-mutations, where a single base-pair in the genetic code has been altered. An SNP, when it has occurred, tends to be preserved. As there are billions of base-pairs in an ordinary genome, the general probability of another mutation in the same position is low (assessments say that each individual carry approximately 175 new SNP mutations [60]). Therefore, all individuals with a specific allele are assumed to share a common ancestor or have a direct ancestor-descendant relationship. The standard situation is therefore that only one *wild-type* allele (the most common) and one mutant version exist, despite the fact that there are theoretically four possible bases in each position. The level of variability in an allele is characterized by the *minor-allele frequency* (MAF), the rate at which the non-wildtype allele appears in the population under study. Different populations can have radically different MAFs for SNPs that occur in both, a situation which has been empirically detected in humans [4].

Genetic testing methods have generally been based on methods where each marker renders an independent signal, e.g. through the locations of bands in electrophoresis or luminescence by hybridization against oligo-nucleotides in a microarray. Since each marker is detected independently, all information on phase is lost.

In recent years, it has become more common to do the equivalent of sequencing the full genome, albeit at "low coverage". Modern sequencing techniques are mostly, despite different technologies involved, based on getting the specific sequence for a short "read" of base pairs. An individual read can vary from tens up to maybe thousands of bps in length, but it is in all cases very short compared to a full chromosome.

"Low coverage" means that there are relatively few copies of the same sequence (overlapping reads). As a consequence, the error rates on the base pair level are relatively high. Despite the fact that a rather complete genome can thus be reconstructed per individual, the analysis is generally confined to "quality controlled" (filtered) genotype data, where again only a set of known markers and known marker alleles are permitted. With SNP chips as well as emerging *next generation sequencing* (NGS) workflows, thousands to millions of unique markers can be genotyped per individual at an affordable cost, making any attempt to manually analyze or even curate data for quality checks futile. At the same time, many existing automatic approaches and algorithms need to be revisited to cope with current data volumes.

Specific tools exist to devise plausible marker maps, including mapping distances between markers, from unordered marker lists [49, 73]. The algorithms used are based on different simplifying assumptions, as it is computationally prohibitive to test for the exponential number of possible marker orderings. From the perspective of this thesis, we consider the determined orderings of markers, as well as their interspersing distances, to be accurate. It should be noted that existing methods for setting up marker maps are conceptually similar to the new algorithms for determining genotype probabilities presented in this thesis.

The problem of determining proper marker maps is also being influenced by the continuing change in total marker counts, and the fact that markers now frequently have a known location within the genome (down to the individual base pair), for fully sequenced species. When the locations are known, the marker ordering does not have to be determined, while the mapping distances still have to be computed based on actual rates of genome recombination.

1.2 Experimental Populations

Theoretically, QTL analysis can be performed in natural as well as experimental populations. In this context, a natural population is a population that would exist, with similar characteristics, even if no QTL experiment would have taken place. A wild population of some mammalian species would be one example of this. In a natural population, the total genetic variability might be great and not completely characterized. Furthermore, environmental conditions are not well-defined, and may even be unknown.

The alternative to a natural population is an experimental population. In many cases, these are based on different lines within a species. A line can be based on inbreeding, where all individuals within the line are genetically identical. A line can also be defined as a specific breed or variant, or based on geographic origin. In these *outbred* conditions, individuals are still expected to be more similar within each line than between lines.

Inbred lines can be preferable, since variations in genetic background between the different founder individuals will not affect the analysis, while such effects can be a confounding factor under outbred conditions. On the other hand, an inbred line might have genetic deficiencies in addition to the intended trait under study, making the experiment results highly significant for the specific population studied, but at the same time resulting in a limited opportunity for generalization.

When the lines demonstrate differences in some trait, e.g. body weight or length of flowering time, crosses between the lines can help elucidate the genetic structure behind the variation in a studied trait. By convention, the founder individuals of such a cross are called the F_0 generation. Individuals from the two lines can be mated, in such a way that each resulting individual has one parent from line 1 and one from line 2. Ideally, parents of both sexes should be represented for both line origins to avoid confusing genetic effects with e.g. epigenetic effects due to imprinting, or womb effects. The resulting individuals form the so-called F_1 generation. All individuals in the F_1 generation show the same genetic background, with one allele from each line in every gene. The phenotype distribution for the F_1 individuals is therefore expected to be identical for all individuals, any differences arising from environmental factors.

From an F_1 generation, two general crossing schemes are common. One is the backcross, where F_1 individuals are crossed with individuals from either founder line. In such a structure, the backcross offspring will in every locus have 1 or 2 alleles from the founder line chosen in the backcross, and 0 or 1 allele from the other line. No effects will be seen in e.g. loci where the founder line used is dominant, as dominance can obscure the presence of another allele. The average genetic contribution of the line used in the backcross is then 75%.

The other common cross structure is the intercross. In an intercross, F_1 individuals are crossed once again, to form an F_2 generation. The average genetic contribution is the same as in the F_1, 50% from each original line in each resulting individual. However, at the locus level, greater variability can be observed, with 0, 1 or 2 alleles from both lines being possible, with the relative proportions $1 : 2 : 1$. An intercross can detect dominant alleles, but the additional degrees of freedom might reduce the effect assessment exactness.

More elaborate crossings are directly conceivable from these basic structures, like repeated crossings of individuals from successive generations, going on from the F_2, resulting in so-called advanced intercross lines and other specific designs. Another example of a multi-generational experimental population is heterogeneous stock, where e.g. 8 founder lines are repeatedly crossed for tens of generations [39, 75]. Repeated crossings make tracing inheritance harder, but at the same time the total number of cross-over events increases. This increase also improves the possible resolution for determining causative mutations, as the region of markers co-segregating with the trait will be limited.

1.3 Genetic Studies in Complex Populations

In order to discern how a genotype affects the phenotype of an individual, a method to bring together sets of individuals with similar genotypes is needed. Before genetic testing based on molecular methods was common, the foremost way to accomplish this was consequently to ignore genotype at individual loci. Instead, only the population structure was used. If the correlation found between siblings is stronger than the one found between cousins, then there is (supposedly) a genetic basis for the trait in question. Out of a total of 2 alleles, the expected count of allele sharing between two siblings is exactly 1. Between cousins, it is 0.25, and similar fractions can be presented for any relationship. This is more commonly expressed in terms of the *coefficient of relationship*, which is half the expected number of shared alleles.

More generally, a matrix $\mathbf{A}$ can be constructed representing the coefficients of relationship between all individuals that are studied. This matrix is generally called an identity-by-descent (IBD) matrix, as it identifies the proportion of alleles that are expected to be shared due to common heritage from a single ancestor. Using random effect linear models, the phenotype in each individual is considered a random value, but with a specified correlation structure. Just like in simpler designs comparing different levels of relatives, a metric called *heritability* can be derived, identifying the portion of variance explained by genotype.

However, such heritability estimates only reflect the inheritable variance arising from the genome as a whole. All loci are averaged together. Knowing that a trait is inheritable in nature, rather than purely due to environmental factors, can be relevant, but pinpointing locations associated to the trait opens up very different prospects in areas such as breeding, basic research, and medicine.

With the advent of dense molecular testing, it is now possible to map heritability to individual loci. In recent years, the concept of *genome-wide association studies* (GWAS) has become common. A GWAS is frequently based on associating variation in individual markers to a trait. If one of the markers tracked is actually causative of trait variation, then no further analysis is needed to trace relationships.

One popular approach to increase statistical power in GWAS analyses has been to *impute* genotypes for study individuals based on denser marker maps. The process of imputation consists of inferring the genotype phase for all markers, then finding the best corresponding genotypes in the reference population, for markers missing from the experimental data. A study population can be genotyped using e.g. SNP chips covering hundreds of thousands of markers. Similar regions are then identified from databases of known haplotypes for millions of SNPs. Short stretches, on the range of a couple of markers, can be preserved over many generations, so reference individuals do not have to be closely related for filling in regions between SNP chip markers.

2. Tracing Inheritance using Hidden Markov Models

In quantitative genetic studies for experimental populations, the most common way to analyze the mapping between genotypes and phenotypes has been to express the offspring genotypes in terms of from which founders alleles were inherited.

If one is only interested in marker positions and the founders are differing in all markers (e.g. the separate founder lines are inbred and the marker set is limited to a set where the founders are contrasting), then no specific analysis is needed to determine genetic relations. In other cases, probabilities of specific relations need to be determined through some algorithm. Within QTL analysis, this was introduced with *interval mapping* [49].

The case of computing line origin probabilities for QTL analysis in outbred populations was explicitly treated by [29], but the suggested algorithm was inefficient, and would not scale to current marker counts. Our initial work in the field of tracing inheritance was motivated by the need to handle hundreds of markers with partial information. This resulted in the original `cnF2freq` tool presented in Paper IV. All our later work, expect for that in Paper IX, has been gradual refinements and extensions of the original codebase for that tool.

In experimental as well as natural populations, causative mutations for many traits of interest can be rare [36]. This means that they are not necessarily covered in the marker set used. Therefore, we propose that there is a need to represent genetic identity between study individuals for short regions in a form that goes beyond the immediate marker data. When interval mapping is used, this is automatically done. However, for natural populations the most popular tool of late has been the GWAS concept, which is based on directly identifying associations to markers.

Imputed genotypes extend the marker set used in analysis and they therefore increase the probability of detecting a causative allele, or a marker allele that is strongly linked to the causative allele. However, it is possible that a causative mutation for a trait effect is missing from the marker set, even if imputed alleles are used. This can especially be true if the marker set is confined in some way, e.g. to only simple SNPs, as the trait-affecting mutation can be of another type, e.g. a copy-number variation (CNV). As the genetics of CNV regions also allow far faster mutation rates, the SNP background can appear identical, unless the overall region similarity for a longer stretch is taken into account.

Although some approaches have been suggested to better truly track shared ancestry beyond single markers, e.g. creating multi-marker windows out of several adjacent markers [26], most GWAS methodologies are still based on regressing directly on (imputed) single markers and the possible losses are not much discussed. There is, though, a general appreciation of the fact that the "missing heritability" needs to be properly explained [57]. Missing heritability refers to the smaller heritability estimates derived from GWAS results compared to classical genetics results for genome-wide heritability for the same traits. In some cases, lack of ability to uniquely discern shared ancestry for affected individuals might be part of the explanation.

The following sections will give an overview of the types of models we are using to trace and reconstruct inheritance in different settings. These methods are focused on explicitly reconstructing haplotypes. The models can be used for comparing shared ancestry between any pair of individuals directly, or as a tool for inferring phase in its own right. Inferred phase data is a pre-requisite for popular GWAS imputation methods. These advances are also important in experimental populations, since computing line origin probabilities in a multi-generational cross without taking true phase into account will introduce a bias, especially if the number of individuals in some generations is small. Therefore, any analysis trying to ascertain line origin on the marker level will benefit from inferred phases.

2.1 Hidden Markov Models

A Markov chain is a stochastic construct where each element in a sequence (chain) only depends on the element immediately preceding it. One example is Brownian motion, where the next location of a particle is considered to be a random small movement from the current position. This model is used and effective, despite the fact that according to Newton's law momentum should also be taken into account. However, on practical time and length scales, momentum becomes irrelevant due to viscosity for particles of suitable size in different media. In a similar manner, we know that recombination is truly a rather complex stochastic process, but we can approximate it to a Markov process:

If the chromatid transmitted at location x_0 is A, then the probability is high that the chromatid transmitted at location $x_0 + x$ is also A. Given the chromatid identity at x_0 and assuming Haldane's function, the identity at $x_0 - x$ will not influence the identity at $x_0 + x$. Although we are aware of cross-over interference and other phenomena, the Markov chain is still a useful simplified model.

Tracing inheritance is a matter of tracking what regions of which sister chromatid were transmitted during meiosis from one parent, to a gamete, to resulting offspring. Even if the method of choice will not track individual

parent-offspring relationships the process of recombination is the main source of variation in the population, bringing together genetic material with different ancestries. If the chromatid identity could be directly observed, the Markovian nature of the transmission process would be of theoretical interest, but of lesser importance for methods development with the goal of performing genetic analyses based on that identity. What we generally have is instead only unphased genotype data at markers, i.e. two recorded allele values, where it is known which one belongs to which copy of the chromosome. Hence, we are observing a "shadow" of a Markov process. The true state, the chromatid identity, is the variable that is considered to be Markovian, but that variable is only observed indirectly. In order to perform an analysis where genotypes are mapped to phenotypes, we would prefer to know the hidden variable and need some method to estimate it.

The setting with an indirectly observed Markov process is generally called a hidden Markov process and models based on such processes are called hidden Markov models, or HMMs. This model structure has been used to accurately describe many different processes with a goal of pattern identification, such as protein motif-finding in bioinformatics [21], speech recognition [41], and part-of-speech tagging in linguistics [10]. They can also be used in a way where the actual quantities of interest are not the states visited, but rather some set of parameter values. One such example is the early bioinformatical tool MAPMAKER [49], which was one of the first applications of HMMs to model recombination. In that application, the models were initially used to determine the mapping distances (and the overall marker map).

General notation and an algorithm overview for HMMs can be found in two classical papers [63, 64], that introduced the model concepts to a wider audience. Briefly, the hidden process is seen as *states*. Given a state at one discrete location in the process, there is a set of transition probabilities, determining the next state. For each state, there is also a set of *emission probabilities*, making the connection between the hidden process and the observed variables. The processes are frequently described in this generative manner, considering each element in the sequence as a time step, with the observed variable being "emitted" based on the state.

Different standard algorithms exist for performing different operations on HMMs. The two most important algorithms for the scope of this thesis are the Baum-Welch algorithm that can iteratively determine suitable values for unknown parameters in the model (e.g. transition or emission probabilities), and the forward-backward algorithm, on which the Baum-Welch algorithm is based. The forward-backward algorithm can determine the posterior probability distribution for the different states in each time step, given the observed variable.

2.2 Applying HMM-based Inheritance Pathway Models

The HMM state should indicate the chromatid origin of the genetic material transmitted in a locus, in all meioses that are tracked. If the genotype of a single individual is traced, the state would amount to two binary flags, corresponding to the two parents, making for four possible states in total. Several applications of HMMs for tracing inheritance in pedigrees, including the package Merlin [2], view the allele transmission process as one single process for all individuals that are connected in the pedigree. This could mean that tens or hundreds of meioses need to be traced. Since the chromatid identity in each meiosis in the state is independent, the total number of states increases exponentially and thus precludes evaluating the model in practice.

The work in Papers IV-VIII on determining inheritance has been characterized by attempts to explicitly reconstruct the phasing, and doing so using hidden Markov models while maintaining a relatively small state space. In the case of known pedigrees, the latter is achieved by using what we call *focus pedigrees*. Although the individuals in an extended family are interconnected by shared meiosis events, the Markov process studied is reduced in each focus pedigree to a single offspring individual and a few generations of direct ancestors to that individual. With two generations of ancestors, 6 meioses are tracked in total, resulting in 64 unique states (compared to 4 states when only including the parents).

Not tracking the full pedigree results in some loss of information in the model. For this reason, we have proposed and implemented an explicit parametrization of phase, that can be iteratively optimized. Each genotype consists of two marker values, and we introduce a parameter that we have named *skewness*, indicating the probability for *AB* to truly be ordered *AB* (skewness 0) or *BA* (skewness 1), or somewhere in between, when information is not clear. The latter case, where uncertain information is represented by an intermediate value, is unique to our approach.

All focus pedigrees where an individual is included can contribute information on what is a likely assignment based on the information found in that specific pedigree. Through repeated iterations, a consistent view of phase in all individuals can be determined. When phase is available, the probabilities for different states change, as some inheritance pathways that are possible given only the unphased data turn out to require an excessive number of recombinations to be possible, given the actual phases.

In Paper VIII we have also added a *sureness* parameter, which is not found in other models of this kind. The intent is to consider all recorded genotypes as uncertain. Even with good experimental protocols, it is frequent to find error rates on the range of 1 in 100 or 1 in 1000 [66]. It is also common that genotype data is missing for some or all markers in some individuals, especially with current low-coverage resequencing methods for genotype determination. In these cases, our model can not only allow for the chance of incorrectly

determined genotypes, but also infer the correct one based on the rest of the pedigree, as a component of the same iterative optimization that determines the skewness values.

2.3 Models for Individuals with Unknown Relationships

Although not included in the papers in this thesis and otherwise only covered as a poster presentation at the 4th International Conference in Quantitative Genetics in Edinburgh, 2012, significant work in the preparation of this thesis has been spent on extending the parametrization using sureness and skewness to cases with unknown relations between individuals. Such data appear in run-of-the mill GWAS data from humans, but also in e.g. collected wild animal samples without pedigrees. In this regime, the state cannot consist of tracking individual meiosis, as the specific meioses resulting in specific offspring are not known.

One way would be to try to reconstruct the pedigree, fully or partially. This approach has shown some success [45]. However, we have chosen to instead base our approach on models similar to those in [34, 51]. In these, a Markov chain is established identifying chromatid identity for the two chromatids in each individual, similar to our work for pedigreed data. However, since the four parental haplotypes are not known, all chromatids or haplotypes found in the remaining population are used as possible states, in the most straightforward presentation of the approach. As two haplotypes are tracked, and there are $2n$ haplotypes (for n individuals) $4n^2$ states are possible.

While $4n^2$ states for data without pedigrees is far less than an exponential increase in number of states relative to the number of individuals, which is found in e.g. Merlin for pedigreed data, work has been done by several authors to reduce the resulting computational complexity further. Among strategies tried are random resampling a subset of individuals for each iteration [51], as well as more targeted subsampling strategies [34], or division of the genome into smaller regions within which all haplotypes present can be tracked [19]. Most of these modifications could be applicable to our implementation as well, although this has not been a focus. Furthermore, if the transition probability matrix has a simple off-band structure, the total resulting complexity order of the forward-backward algorithm can be brought down from $O(n^5)$ to $O(n^3)$, even when the $4n^2$ state count is kept. Basically, all transitions to haplotypes from another individual can be merged into one meta-state during some of the calculations.

Most existing implementations of this family of Markov models tend to consider the haplotypes of the state-defining individuals as fixed in each iteration, allowing no intermediates in phasing. Approaches related to Markov chain Monte Carlo (MCMC) methods are then used to resample the haplotypes per iteration, based on the sequence probabilities predicted by the chain.

MCMC approaches work well if they are effectively exploring the full configuration space. Like deterministic methods, they can get trapped in local minima in practice. The work in Paper IX has been devoted to this problem of ensuring *chain mixing*, that the new haplotypes sampled in each iteration are different enough to actually explore the full genotype space.

The other approach we suggest is to use a deterministic optimization, similar to the one introduced for the case where the pedigree is known. By starting out with all phases (and possibly unknown genotypes) assigned an intermediate value and gradually updating them, finer details can be resolved, effectively exploring a higher number of possible interpretations than what is the case for the random binary optimization schemes.

2.3.1 Relativizing Skewness

In our work with pedigreed data, it makes sense to assign a single heterozygotic marker the role of an absolute reference for skewness, letting all other values relate to that marker. For the non-pedigreed case, this seemed to introduce severe bias, when some regions would converge quickly, and others would not. Rather than this type of absolute skewness, we suggest instead a relative skewness definition for analyzing sets of unrelated individuals. This means that rather than having skewness act on emission probabilities, it should act on transition probabilities. The states map to the individual alleles recorded in the unphased data, and the skewness tells at what points that arbitrary ordering needs to be inverted in order to reconstruct the true phases. This parametrization is then local, in the gaps between individual markers.

However, this change to a relative skewness definition is no panacea. For example, it is not immediately clear how to handle homozygous markers. Likewise, if all individuals are heterozygous in some marker, there is no point of reference. With the exception for that marker, phase might still be determinable with absolute skewness, but a single ambiguous marker could break linkage when relative skewness is used. If marker genotypes are absolute (i.e. sureness values and no expected genotyping error), it is also foreseeable to implement diagnostics that would adjust the relative skewness values in relevant cases, but detection of those cases becomes much harder to accomplish if the genotypes themselves are not fully defined.

One tantalizing prospect of a relative skewness definition is that it becomes straightforward to include information on extreme short-range linkage from so-called next-generation sequencing (NGS) genotyping. For these methods, the marker genotypes are actually detected by short sequencing reads. Within reads, phase information *is* indeed known, meaning that two alleles detected in the same read are physically residing on the same sister chromatid in the original genome. Including such short-range information into existing hap-

lotyping schemes could greatly enhance their performance in highly heterozygous and marker-rich regions.

2.4 Advances in Optimization Techniques

The Baum-Welch algorithm [7] is the predominant method for fitting (optimizing) HMM parameters. Compared to many other HMM optimization problems, our models contain a very high degree of parametrization, as there are skewness and sureness values for every marker in every individual, and there are approximately the same number of data points. However, one fact reducing the total space of possible final models is that it is known that the true value (corresponding to the physical reality) for each parameter is in fact 0 or 1. Therefore, any optimization approach can be designed with the goal that each value should converge to either of those extremes. In many other HMM applications, probabilities converging to extreme values is rather a signal of model breakdown, such as severe overfitting.

In pedigreed data, our main efforts have been focused on adding dampening in the updates, which is a known technique to improve the local optimum found by the Baum-Welch algorithm. We have also chosen to do a logit transform ($\log(p/1-p)$) of the parameter values, since all terms in a likelihood involving a specific individual will either include the skewness factor p or its complement $1-p$ (one of the two possible allele orderings).

More importantly, we have also made the observation that the absolute skewness parametrization can result in local bubbles of inverted skewness values, so that the optimal value would be $1-p$, where p is the current value. The Baum-Welch algorithm iteratively updates each parameter conditioned on the current values for the complete sequence and each update is thus local relative to a fixed environment. If the skewness values for a stretch of markers have been incorrectly assigned, the optimal local updates will not converge to the correct values.

We have implemented efficient ways to use the variables computed in the forward-backward algorithm to check explicitly whether an inversion of skewness values at all markers downstream of some position would improve likelihood, i.e. changing all values from p_i to $1-p_i$. If this is found to be the case (and not conflicting with some other such inversion already detected), the skewness values are changed to that new configuration. This added check greatly improves the quality of the optima found, as the inversions performed result in a global likelihood improvement, while the possible local improvements would be minimal or totally absent. The technique, as well as the general observation of the importance of paying attention to the nature of the HMM parametrization used, could be relevant to other applications.

The specific details in the operation of the optimization for pedigreed data have been gradually refined, from a coarse implementation (but already

including simple inversion testing and some dampening) in Paper IV, to a full optimization including sureness with a more generalized inversion support in Paper VIII. The latter contains the most full description of our model.

When considering the relative skewness definition we are proposing for non-pedigreed data, explicit inversion testing becomes irrelevant. One could devise a similar scheme for detecting point gaps, as discussed earlier. Even if that might improve the result, the overall change should be less pronounced compared to the absolute skewness case. Instead, our focus has been to ensure convergence for skewness everywhere to one of the extreme values of 0 or 1.

Since the skewness values are not only defining the transition probabilities for the individual analyzed, but also the transitions defining state haplotypes from the other individuals, the likelihood becomes a product of multiple skewness values (which can be in a feedback loop influencing each other for each iteration), so the likelihood is in fact in some places a polynomial in terms of the same skewness value. The optimal value might therefore be somewhere between 0 and 1, rather than at the end points, and this is the value the Baum-Welch algorithm will converge to.

To avoid feedback behavior stopping convergence, our method is choosing updates based on the assumption that the likelihood function will be linear between 0 and 1, scaling the logit change in likelihood to the logit of the relative difference in likelihood for hypothetical $p = 0$ and $p = 1$ scenarios. This issue could be worth noting for other HMMs where emissions or parameters are somehow reused for defining the states in the following iteration (which is not the norm).

2.5 Comparison of Results for Analysis without Pedigrees

In Paper VII, haplotypes were perfectly reconstructed in founders for a simulated dataset using a multi-generational pedigree with thousands of descendants to 20 founders. In Paper IX, we improve the MaCH [51] tool by ensuring proper chain mixing and test the method on a set of the 30 first trio parents (i.e. 15 parental pairs where the offspring was also genotyped in the original data, ensuring high-quality phase information) from the Hapmap project subpopulation of Utah residents with Northern and Western European ancestry from the CEPH collection (CEU). To illustrate the benefits of applying our methods with a relative scalar skewness parameter for phasing in populations where pedigrees are not known, we applied this method, the original MaCH tool and our modified version of MaCH presented in Paper IX on these two datasets, both with only the founders present, and with at least one offspring individual per founder. This expanded the simulated dataset to 35 individuals in total, and the trios went from 2 individuals from each of the 15 families, to 3. The total number of phase switches in the inferred haplotypes are presented

in the table below. More information on the datasets is found in the respective papers.

	`cnF2freq`	Original MaCH	Modified MaCH
30 CEU trio parents	6 051	14 770	6 597
Trio parents + offspring	1 537	10 692	2 150
14th QTLMAS simulated founders	347	985	386
14th QTLMAS founders + offspring	33	1 779	89

Table 2.1. *Number of switches when using our method, the original MaCH [51] tool, and our modified version presented in Paper IX, on a simulated [74] and a real dataset based on Hapmap CEU trio parents. Both datasets were tested with all three tools, with only the parent/founder individuals, as well as when including at least one offspring individual per founder. The number of switches needed to restore the true phase for all individuals was counted. Our method stays on top in all cases, but clearly excels when additional long-range phase overlap is present, even when the relationship was not explicitly encoded.*

For the CEU dataset with no offspring, the recently presented Shape-IT [19] tool was also tested, resulting in 6 189 flips after 8 000 iterations. The resulting improvements from adding further iterations seem to be minimal for both this tool and MaCH.

If there is reason to believe that some individuals in a dataset without known pedigrees are only a few generations removed from each other, there is clear reason to investigate the use of phasing methods that go beyond simple MCMC, as those will generally not be sensitive enough to capture long-range phase information. Our method can therefore be used as a step for pre-phasing before imputation in e.g. isolated human populations, as well as a phasing tool of choice for small unstructured animal or plant populations. For larger and non-related populations, the benefits are less clear, but our methods are still competitive.

2.6 Summary

Finding genetically similar regions is a crucial first step in elucidating the mapping between genotype and phenotype, whether it is in terms of simple line origin probabilities or more complex models. The former case will be discussed in more detail in later chapters. Determining genetic phase can

also be a tool in other genetic analyses, and also used as a step for quality control and improvement on genotype data, e.g. for doing imputation based on a reference panel or inferring missing genotypes based on recorded data for relatives.

Although the code version included in the package described in Paper VI does not cover haplotyping support, all our work in the field has been based on variants of the same codebase. As the MAPfastR package provides an easy-to-use interface for biologists experienced in using the R statistical environment [1], it would be attractive to extend it with the most recent version of the haplotyping engine used in Paper VIII as a step in making that work accessible to a wider audience.

3. Models for QTL Analysis

In the two previous chapters, we have treated the biological foundation and methods for tracing inheritance in populations. When probabilities of state or line origin have been defined for any putative set of candidate loci, a proper model is needed to analyze the quality of a fit. Several approaches are available in the literature. Many are related to interval mapping (IM), which is a specific approach for handling non-definite genotypes. Earlier approaches only analyzed single markers, discarding those samples where genotype information was not conclusive. Such an "available case analysis" is well-known in the statistical literature, but known to be flawed, with limited power as well as distorted results as possible consequences [53].

What constitutes a model is not always evident, especially as different methods and frameworks in some simple and illustrative cases give identical or directly equivalent results, while the outcome can vary in more complex cases. One crucial aspect in the modelling is what assumptions are made on the phenotype distribution for single individuals, given their genetic information. This results in a statistical framework for the analysis. Another is what phenotype values are expected for different genotypes, defining the parametrization of the model. The framework will primarily affect the detection of a proper QTL location, while the parametrization is critical for making the estimated genetic effects possible to interpret and analyze from a biological standpoint, including generalizing them to possible other populations. These two aspects tend to be intertwined, something that becomes even more clear when *model selection* is considered. Beforehand, it is in general not known what number of loci and genetic mechanisms to expect, so different parametrizations need to be tested in a common framework.

The work in this thesis has focused on linear regression models, due to their simplicity and computational advantages. In Paper I we have suggested a specific imputation scheme similar to what has been called the multi-QTL model, but with another theoretical motivation related to the concept of orthogonality. In the following sections of this chapter, we put that work in context by also briefly introducing other model approaches and their relation to linear models, as well as the issue of selecting a proper model size through model selection. Our advances in optimization methods and software infrastructure presented in Papers II, III are relevant for these issues, as well.

3.1 Preliminaries

The main hypothesis of all QTL models described here is that the total variance in a trait can be decomposed into multiple independent components, $V_{tot} = V_g + V_e$. Here, the component V_g is the genetic variance, while V_e is an environmental component. These components are assumed to be separate, or statistically independent, i.e. with zero covariance. If a proper separation of the genetic variance is constructed, the within-class variance for each genotype identified would be V_e.

If a genetic property is in fact interacting with the environment, or with another locus not covered by the model, the within-class variance will increase. The explainable variance from the model will probably be smaller than V_g, as multiple loci are exerting a limited influence on the trait and some of them will not be included in the model. An effect counteracting this underestimation is the problem of overfitting where, by chance, the recombinations between an analyzed position and the true position can redistribute individuals between classes in such a way that the within class variances are reduced. Even at the correct locus, a model with a high number of parameters can give a better-than-true fit, as the parameter estimates are fitted against the specific individuals sampled and can describe the random variations within that sample as part of the model.

Assume that the genotype class for each individual i is described by a vector $z_i = (00\cdots010\cdots0)$, where the location of a single non-zero element 1 represents the true genotype. The distribution for observed trait values for i can then be expressed as:

$$\mathscr{L}_i = \sum_j z_{ij} N\left(\mu_j, V_e\sigma^2)\right), \tag{3.1}$$

where σ^2 is ideally expected to approach V_e. This interpretation is common to all the models presented here, when z indeed is a binary vector. The maximum likelihood (ML) can in this case be computed using standard methods for least-squares linear regression, or using the much faster PERF algorithm [54]. The variables μ_j are fitted directly, and σ computed from those variables.

3.2 Partial Information

In the case of partial information, several interpretations are possible. The original presentation of IM reuses (3.1), but defines z to be a general probability vector, i.e. only imposing the conditions $\sum_j z_{ij} = 1, 0 \leq z_{ij} \leq 1$. The result is a normal mixture, a linear combination of multiple normal distributions describing the expected phenotype for the individual i [47]. It should be noted that this distribution is not equivalent to the distribution of the sum of two normally distributed variables. Rather, the shape of the single distribution is

the sum of multiple identical normal distributions, weighted by probability, and with identical standard deviations, determined by σ.

Optimizing this likelihood as a function of the vector $\theta = \{\sigma, \mu_j\}$ is a non-linear problem which is usually solved using the expectation-maximization (EM) algorithm. This algorithm, first presented in its general-purpose form with that name in [20], is relatively robust, but sometimes converges slowly. The algorithm consists of two main steps within each iteration, explaining the name:

1. Expectation: Compute the expected value of the full (log-)likelihood function for the population, given a current parameter vector θ.
2. Maximization: Compute the optimal value of θ, given the log-likelihood function.

When the optimization is complete, marginalization can be done to determine a posterior probability vector where the prior genotype probabilities p_{ij} are replaced by posterior probabilities π_{ij}.

3.2.1 Methods of Imputation

In the IM methodology, a single mixture realization of the population is considered. The prior probabilities for the individual genotypes are used over all iterations. There are also a set of competing schemes, called imputation methods, that use multiple realizations of the population. One form of this is the multi-QTL model (MQM) method [37], which encompasses e.g. the inclusion of covariates for markers at non-modelled positions to handle genetic background, and as the name indicates, the modelling of multiple loci within a single model.

However, in the context of treatment of partial information, MQM differs from IM by its use of repeated weighted linear regression. The individuals are not realized as a single normal mixture phenotype distribution, but rather multiple separate distributions, where the total influence of each onto the composite likelihood for the population is defined as $L_j^{\pi_{ij}}$. The distributions resulting from the different states for a single individual are multiplied by each other, rather than merged into a single distribution by addition. In the likelihood space, a choice has to be made between defining the composite likelihood as an arithmetic mean or a geometric mean of the likelihoods for each genotype.

Furthermore, MQM is adjusting the probabilities π_{ij} at each step, using these within the model. The genotype probability is then based on a combination of prior information and the resulting phenotype probabilities. This two-sided scheme is in fact a variation of the general location model, for which an excellent description of the general case and different specializations can be found in [53]. It should be noted that the general location model is not

equivalent to generalized linear models, although specific realizations of one can fall within the other.

The linear regression case for full information, the partial information model in IM, and the partial information model in MQM share a central property of posing a single realization of the population. MQM introduces multiple rows for each individual, but they are all part of a common parameter fitting step, which is iterated until the effects, as well as the π_i values, have converged together. When the individuals are separated in the way done in MQM, as well as in the new interval mapping by imputations (IMI) method we propose in Paper I, the individual observations entering the regression can uniquely be assigned to a single "genotype class", creating a virtual case of full information.

There are also several methods employing Bayesian and Monte Carlo methodology for QTL analysis. Some are "true" Markov Chain Monte Carlo approaches, like the ones described in [72], but there are also other methods employing the methodology towards the analysis of a single set of candidate loci, i.e. a comparable setting to the cases for the models already discussed. The input for these algorithms consists of genotype probabilities and phenotype values, the output of a likelihood of fit, and a set of model parameters focusing on the genetic effects and the residual variance. In [70] one such approach is described, where multiple full realizations of the genotype at evenly spaced "pseudomarkers" are constructed.

In the context of a single position, creating a set of realizations through pseudomarkers is equivalent to sampling from the individual genotype probabilities already seen in the other methods. Each realization will be a case of full information, which can be handled through the first method described above. These realizations are then weighted together, based on the residual variance determined. The residual variance has an immediate algebraic relationship to the likelihood for the model. In effect, the method is approximating sampling from the posterior distribution by taking samples from the (much simpler) prior distribution, and then applying this weighting. The authors of [70] note that there is a close similarity to general IM, the critical difference being that IM finds a single optimum parameter vector, whereas imputation methods marginalize over all vectors.

3.2.2 Linear Models for Partial Information

Performing IM using the EM algorithm is computationally expensive. The model computations in each iteration requires evaluation of the Gaussian probability distribution function in different points for each individual. In addition, the EM method has only linear convergence, normally resulting in a relatively large number of iterations. For these reasons, the authors of [28, 58] independently suggested the use of linear regression even for partial

information, where the z_i vectors are not binary. The result is a probability distribution with σ^2 variance, centered not on either genotype mean, but on a point between them, as if the continuum of probabilities directly corresponded to a continuity of genetic effects on phenotype.

The simplicity, both in computation time and regarding simple use of existing software tools, has made this form of linear regression highly popular [14, 33, 67]. However, multiple limitations have been pointed out, including a confusion of residual variance with unexplainable variance due to lack of information [80], and other aberrations found in systematic simulation studies [42]. Suggested remedies have included iteratively re-weighted regression [81], and estimation equations (EE) [22]. The latter of these can also be implemented using Fisher scoring [30]. We have also noted related problems in Paper I in this thesis.

3.3 Orthogonal Models

From a biological perspective, handling arbitrary mappings of full genotype realizations for all loci modelled into mean phenotype values is cumbersome. The total number of parameters also becomes very large for multiple loci, as the total set of phenotype effects would have size n^d, where n is the number of genotypes per locus (generally 3 for F_2 populations), and d is the number of loci. In addition, a number of covariates can be included in the model, e.g. sex and other factors not captured directly by the markers, while these factors are still expected to have a significant influence on the trait under study. By correcting for such effects, the actual correlation between the genotypes and the trait can be made more clear [56].

In basic Mendelian genetics, the traits are binary. The inheritance pattern for a trait can introduce an asymmetry between alleles, with one dominant allele, meaning that the presence of that specific allele results in the trait being "active". Other alleles are recessive, meaning that the trait effect associated to the allele will only be seen if the dominant allele is not present, i.e. both copies in a diploid genome being copies of the recessive allele. The dominant and recessive effects can frequently be understood as a matter of functional versus silent (non-functional) alleles. If one copy of the allele is damaged in some way, it might never be used, with the working copy being used instead, or supplanting the damaged gene product. The damaged, or silent, allele is then recessive, as no phenotype change is seen unless both copies are silent. Against a background of silent alleles, a single functional allele will be considered dominant, as its presence will modify the phenotype. Two copies of the dominant allele will not affect the trait further than a single copy, as the working piece of genetic code is present in both cases.

3.3.1 Historical Models

Based on Mendelian thinking, a continuous trait with two alleles can be dissected into an additive and a dominant portion. The additive portion is linear to the allele count, while the dominant portion is contrasting heterozygotes against homozygotes. For a F_2 population, a design matrix $\mathbf{S}$ can be defined as

$$\mathbf{S}_{F_2} = \begin{pmatrix} 1 & -1 & -0.5 \\ 1 & 0 & 0.5 \\ 1 & 1 & -0.5 \end{pmatrix}, \tag{3.2}$$

where the columns represent base vectors for the variables μ, α, δ. This design matrix can be multiplied with a $\mathbf{Z}$ matrix with full-information individuals to build the model for linear regression. Similar approaches can also be devised for the partial information methods, using the parameters to compute the per-genotype class means. The original F_2 model was due to Fisher, in [24]. Further extensions were made in the 1950s for more properly handling epistatic interactions in loci showing F_2 frequency proportions.

3.3.2 Current Models

The renewed interest in quantitative genetics due to the molecular developments presented in Chapter 1 has also spurred new efforts regarding the basic modelling and parametrization issues. These efforts have been focused on achieving *orthogonality* in more cases. In mathematical terms, the property of an orthogonal model is equivalent to the full model design matrix $\mathbf{X} = \mathbf{ZS}$ being orthogonal, for which $\mathbf{X^T X}$ being diagonal is a sufficient and necessary condition.

Orthogonality, in this context, is generally analyzed in the case of individuals of full information, independently of the regression model and parametrization. The intent is to ensure that estimates of the different variables (model parameters) are independent. This independence ensures that the estimates of the remaining parameters are unaffected if a parameter is removed. Removing parameters is a natural step when attempting to reduce a model to the most crucial loci affecting a trait. The explained variance from the model can also be clearly and uniquely attributed to the parameters if independence is ensured. The original F_2 model and other models were conceived to be orthogonal for populations exactly matching that structure. No population of finite size will be a perfect F_2, as the allele and genotype frequencies will not match the expectation values exactly. Therefore, no matter the problem structure, a slight deviation from orthogonality will be present for actual data. Subsequent developments implement support for other frequency distributions, first handling deviations where Hardy-Weinberg equilibrium (genotype frequencies being products of allele frequencies) still holds [84] and finally handling any genotype frequencies for a single locus [5]. The parameter estimates for

one specific population can also be translated into corresponding functional estimates (estimates describing gene function). Functional estimates derived from different populations can then be accurately compared.

In Paper I we have shown that a linear regression as normally performed using a linear model designed for orthogonality in cases of full information, will not be orthogonal in cases of partial information. Since we know from Chapter 2 that there is almost always some level of uncertainty regarding line origin, information is almost always partial to some extent. Therefore, any approach relying on strict orthogonality will fail. In order to remedy this, we suggest the IMI method in the paper.

3.4 Model Selection

Selecting the appropriate model is a matter of finding the proper set of loci, with the proper parameter set best describing the external factors. There is an inherent trade-off between specifying multiple loci and several parameters, and avoiding the risk for fitting the specific sample, rather than capturing the true biological properties of the underlying population.

One common approach is to separate the selection of model size (or rather a specific model structure, including size), and the selection of the optimum locus set and eventually parameter values for this model configuration. This has also been the view used in the papers in this thesis. The other option would be a more general Markov chain Monte Carlo walk, with transitions including variations in model configuration as well as the specific loci included [72].

3.4.1 Model Choice between Parameter Sets

The traditional likelihood for a model is, in essence, the likelihood for the observed data, given the model, i.e. $P(y|M)$. If we want to choose the single most likely model, we should maximize $P(M|y)$. However, assuming the prior probability for any model M to be a constant $C = P(M)$, we can see that, according to Bayes' law, $P(M|y) \propto P(y|M)$.

The assumption of a constant $P(M)$ will break down if models of varying size are analyzed. A larger number of parameters increases the model space, although not all models are relevant. Several criteria have been suggested for introducing a correction to the likelihood to account for model size differences. Two general approaches (not restricted to QTL analysis) are the Akaike information criterion (AIC) [3], and the Bayesian information criterion (BIC) [68]. These have been studied for QTL applications [11], including simple modifications of the BIC. In recent work, a more context-sensitive approach is proposed, handling main effect and interaction parameters with separate weights [9, 82]. A review of these methods can be found in [61]. For reference, the expressions for AIC, BIC are given below:

$$AIC = n \ln\left(\frac{RSS}{n}\right) + 2K \quad (3.3)$$

$$BIC = n \ln\left(\frac{RSS}{n}\right) + K \ln n \quad (3.4)$$

Here, n is the number of individuals, K is the number of parameters, and RSS is the residual sum of squares (which relates to the model likelihood in the linear regression case).

3.4.2 Orthogonality for Model Selection

If the modelling approach is not orthogonal, all models in the model selection set must be evaluated to select the most suitable model. The parameter estimates and variance components per parameter cannot directly be used to estimate the total explainable variance in a model where that parameter would be excluded. For K parameters, 2^K evaluations need to be made. This is prohibitive even for limited values of K.

If the modelling approach is fully orthogonal, both in model design and evaluation method, only the full model would need to be fitted. The variance components for specific parameters will be unchanged when the total parameter set is modified. Therefore, an optimal combination of variance components can be selected with no new model fitting. This optimum can generally be found using a simple dynamic programming algorithm, depending on what constraints are imposed on the parameter set, e.g. requiring main effects to be present before allowing corresponding interaction effects, as frequently advocated [70].

In practice, our work in Paper I demonstrates the grave issues involved in designing a truly orthogonal model and statistical framework for multiple loci, especially when genotype information is not fully complete.

3.4.3 Model Selection and Significance Testing

The concept of model selection is closely related to significance testing. For model selection, as treated here, the interest is to determine which model is the most likely one to be true for the underlying population. The goal of significance testing is to assess whether the effects found have significantly stronger support compared to the null hypothesis, i.e. compared to the hypothesis that no loci in the data can explain the genotype. In a way, this is a very specific case of model selection, the only models specifying 0 and n loci, respectively, with the significance level indicating to what extent the larger model should be preferred.

3.4.4 Permutation Tests

Deriving theoretical expressions for the distribution of the likelihood function is a complex task. Some attempts have been made [47, 65], but since the QTL analysis setting exhibits non-symmetric population structures, non-normal phenotype distributions and non-homogeneous marker maps and marker information patterns, the efforts based on conventional statistical theory by necessity tend to become only crude approximations.

Instead, the standard approach is to perform permutation testing [17]. From the original dataset, permuted versions are created. Here, the individuals and their genetic makeup are kept unchanged, while the phenotypes are permuted. Several of the sources for specific deviations in the objective function will be kept by this method. However, if covariates of different kinds are included in the model, the permutation method needs to take these into account, possibly by doing permutes within separate classes defined by the covariates [18]. The intent is to permute individuals considered to be identical, except for the actual genotypes modelled. Choosing suitable individuals to exchange gets exceedingly hard when more background factors are to be taken into account. The details of choosing good candidates for permutation in more complex settings show similarities to approaches and methods for designing balanced case-control studies.

Assuming suitable permutes can be created, the objective function distribution for the null hypothesis can be found empirically. If 99% of the values are inferior to the determined objective function value for the real model, the probability of a Type-I error is 1%, resulting in a practical implementation of a significance test.

Although this has not been fully evaluated, the software presented in Paper III and the methodology presented in Paper II can together be used to perform a sort of model selection based on permutation tests. The specific quantile of a QTL candidate within the null distribution can be determined and different model configurations can then be ranked based on their quantile. This approach is demonstrated in the experiments in Paper III.

4. Optimization over the Genome Landscape

In order to find a proper set of QTL, using the types of data, genotype probabilities, and models that have been discussed in the previous chapters, the genome is scanned in some manner to find the optimum model in terms of the set of loci included. This optimum is then considered to be the true QTL, although the literature is frequently reporting multiple "peaks" of models fitted separately, assuming independence between them [14, 43]. However, for such cases, a multi-dimensional model with only main effects would be preferable over a single-dimensional scan, to avoid confounding effects from other QTL to be included in the respective main effect estimates.

4.1 Exhaustive Search

The most straightforward, almost naïve, approach to explore the model fit landscape over the genome (in whatever dimensionality d), is to loop over all genome positions in a closely spaced grid, say every 1 cM, evaluating the model at each of these positions. The position resulting in the highest likelihood will be stored and reported in the end. The goal in the work in this thesis has been to perfectly replicate the results from an exhaustive search, while making the computational demand feasible. This is done by analyzing the linear regression model in further detail in Paper II, and providing simple to use software for doing distributed QTL searches on grid computational resources in Paper III.

4.2 Forward Selection and Backward Elimination

Exhaustive searches for QTL can be computationally intractable for any dimensionality $d > 2$, unless extensive high-performance computing resources are used. A common approach is then forward selection [9, 12, 82, 83]. The basis for *forward selection* is to find a single QTL with the highest possible likelihood, for inclusion of that QTL in the model. When that locus has been selected, a new scan is performed, keeping the first locus fixed and adding a second locus. This process can continue to an arbitrary dimensionality d. Unless full independence of parameters per locus can be guaranteed, this approach will not necessarily find the optimal set of loci. If interaction effects are included, the forward selection approach is also inappropriate, since the

selection of the first locus will not take into account what level of explanative power will be possible when the locus selected is allowed to interact with further loci. One possible approach is then to use pairs of loci as the basic forward selection unit. However, this solution will not handle networks of gene interaction, where the same locus is part of multiple pairs. The best first locus to include would be a "hub" in such a network, but the hub will not necessarily be the first selected locus in a pair-based forward selection method, if there are also other interactions. It is also possible that e.g. selecting a single locus between two linked QTL on the same chromosome might be the best single-locus model, while such an intermediate locus is not at all present in the best two-locus model. This is true even if the two linked QTL display no interaction in the two-locus model.

The suboptimality of forward selection is also acknowledged by authors advocating use of the method [12]. This can be partially compensated by a later step of *backward elimination*. A model of a size greater than the optimal final model is developed through forward selection. A backward iteration process is then initiated, systematically testing the removal of the included loci. In each iteration, a new model is created, removing the locus where the removal proved to have the lowest reduction of the likelihood. The resulting models of size $[1,d]$ will most likely have a higher likelihood than the original models of equivalent size during the forward selection phase, since the removal step has a form of hindsight with a larger total set of relevant loci, where the non-independence and interaction contributions that make forward selection sub-optimal can better be taken into account.

The results are, despite the use of backward elimination, not guaranteed to be optimal. The total set of models to test per iteration is much smaller for backward elimination compared to the forward selection process, and therefore computationally cheaper, since the space only consists of those QTL already identified in the forward selection phase. This small set can be contrasted to the dense grid of all locations in the genome, or even all pairs of such locations, that form the basis for the exhaustive search in that phase. Therefore, using forward selection without backward elimination can rarely be warranted. It is also possible to extend the single "cycle" of forward selection and backward elimination by adding additional and more flexible paths of extension and elimination from a model. Optimality of the resulting solution still relies on the problem structure being overly simple, or the signals from individual loci being strong.

4.3 Genetic Algorithms

Genetic algorithms (GA) form a class of methods for general global optimization, frequently used when details about the analytical properties of the problem at hand are scarce. Inspired by natural evolution, a population of

"individuals" are evolved. Each individual carries some form of "genetic code" determining the properties of the individual. Each individual represents an object that can fulfil a defined purpose or objective, i.e. a solution to the optimization problem in our case. Between generations, reproduction is taking place, mixing the genes between the individuals, followed by selection, removing individuals with inferior performance relative to the objective. Many different schemes exist with different specific methods for reproduction, representation of genes, mixing of genetic material, but the main ideas are shared by all concepts. A summary of GA approaches in general can be found in [6].

For QTL problems, genetic algorithms were pioneered in [15], where each locus was treated as a separate GA gene. As the effects from the loci are frequently almost independent, separate searches would tend to find good candidate loci, or small networks. By mixing the GA genes, these individual genes or networks are merged into a total optimum.

4.4 DIRECT

In global optimization theory, different mathematical properties of the objective function in the model space can be exploited. One example of such a property is that of Lipschitz continuity, which can be defined as

$$|f(x+v)-f(x)| \leq rK, |v| \leq r, \tag{4.1}$$

where $r, K \in \mathbf{R}, > 0$, v is an arbitrary vector of maximum length r, and x is some vector within the model space. In other words, within any environment, there is a limit on the total variation relative to its center point. When we seek for an optimum of $f(x)$, this property can be used to reduce the space explored. If there is a known f_{min} candidate, and f has been evaluated in another point x', where:

$$f_{min} < f(x'), \tag{4.2}$$

no point x'' where

$$|x'-x''| \leq \frac{f_{min}-f(x')}{K} \tag{4.3}$$

can result in a value

$$f(x'') < f_{min}. \tag{4.4}$$

Exploiting (4.1) for Lipschitz optimization has been frequently described in the literature [25, 59, 62, 71], and even more frequently as an implicit assumption in different heuristics within other methods.

4.4.1 General Presentation of DIRECT

A straightforward application of Lipschitz optimization requires K to be known. However, if K is unknown, the algorithm DIRECT, for DIviding

RECTangles, can be used [40]. In DIRECT, concurrent hypotheses are examined for K. This is done by dividing the search space in so-called boxes. For each box, the objective function is evaluated in the center. The first box will cover the entire search space, and this box is split into three boxes along some dimension. The centroid of the center box will coincide with the centroid of the original box, while the objective function has to be evaluated in the two new boxes adjacent to the center box. From this point on, a convex hull is determined in a space consisting of box radius on one axis, and function value on the other. All boxes represented by vertices included in this hull are split in one iteration of the DIRECT algorithm.

A box will be part of the convex hull if there is no box with higher radius that presents a smaller value for f in the centroid. Among the active boxes (the original box in a splitting is always removed, replaced by the three children), one, out of potentially many, with maximum radius will always be split. This is due to the fact that the global minimum of the function might actually be found within that box, no matter what the value of f is in the centroid, assuming some arbitrarily high value of K. Likewise, for all boxes in the hull, there will be some value of K which would allow a value of f superior to the currently known f_{min} to exist within the box radius from the centroid. If a larger box b_1 has a smaller value of f than some smaller box b_2, the minimum possible f value within b_1 would always be lower than that of b_2, no matter the value of K. Hence, b_2 is not a suitable candidate for splitting.

During the iterations of DIRECT, a possible result is that the environment of b_1 is shown not to contain any value close to the theoretical Lipschitz bounds possible with free assignment of K, and thus b_2 might be considered at a later iteration, at the very least when the descendant boxes after splitting of b_1 all have a smaller radius than b_2. As the largest box of the active set will always be split, there is no natural termination condition for DIRECT. The objective function can be studied in ever-increasing detail, but assuming no upper bound on K, there is in theory always a possibility that there will be a new optimum to be found within the vicinity of the centroid of any box, no matter the evaluated function value at the centroid.

Standard termination criteria for the iteration process can be based on simply running a fixed number of iterations. Other options include termination after a certain number of iterations with no new global optimum found, or a certain minimum maximum radius (i.e. a maximum size on the largest remaining box), or similar heuristic variations shown to be very effective in practice [23]. If a maximum radius is chosen equivalent to that of the lattice spacing in an exhaustive search, a DIRECT search will require at least the same number of function evaluations as an exhaustive search over the same model space, thus resulting in no net benefit.

4.4.2 DIRECT for QTL Analysis

The QTL objective function, whether it is based on the linear regression residual or the log-likelihood from some other model, is a candidate for optimization using DIRECT. There is no known Lipschitz constant, but on the other hand there is clearly a continuous quality to the shape of the function in the limit of an infinite-size population with recombination taking place in a continuous manner. DIRECT has also been successfully used for QTL analysis, with the minor modification of introducing so-called chromosome boxes [55] briefly described below. We have also built on this work in Paper II.

When DIRECT is used for QTL analysis, the initial configuration used represents any combination of QTL (the total number determined by a chosen dimensionality d) located on the chromosomes included, rather than a single box representing the full search space. This is done since even though an arbitrary concatenation of the chromosomes of the genome could be constructed, no continuity in the objective function is expected over chromosome boundaries. Although DIRECT supports values of K arbitrarily large in theory, the method also becomes arbitrarily inefficient in actually finding the optimum under such conditions. The division into chromosome boxes also presents an obvious choice for loosely connected parallelization of the DIRECT search, executing a search for part of the search space, defined as a subset of the initial chromosome boxes, on each node in e.g. a computational grid environment [38].

Our contribution in Paper II originates in a new objective function, LogVar (the logarithm of the proportion of explainable variance), which has a demonstrated well-defined Lipschitz constant in an infinite-size population. We also suggest a bound that is close to a Lipschitz structure for actual experimental populations, allowing some boxes to be permanently ignored from the DIRECT process, thus providing a new form of termination condition. We call the modified algorithm PruneDIRECT, since some parts of the search space are permanently excluded from consideration based on the bound. This approach also enables efficient determination of the quantile of a QTL candidate relative to the null distribution through permutation testing, since the permutation searches can be done against the minimum found in the true search, greatly accelerating the search.

5. Summary of Attached Papers

The papers included in this thesis cover many aspects needed for QTL analysis in line crosses, as well as more generical issues of analyzing genetic data. The papers can be considered to cover three areas: models (Paper I), the QTL search problem in line crosses in theory and practice, including applications to model selection (Papers II, III), and methods for analyzing inheritance patterns in terms of haplotype reconstruction, genotype inference, and line origin probabilities (Papers IV, V, VI, VII, VIII, IX).

The results in this thesis form the basis for actual experiments unravelling the secrets of biology, as well as realizing possible economical gains for important crops and animal breeds. The thesis is not focused directly on the genetic applications, but the efforts have been directed by an awareness of the needs arising, including the need to describe gene networks and epistasis, aspects critical for more complex traits [16]. Another limitation in scope is that the results are directly applicable only to applications with a clear, known, population structure, with the exception of our latest work not covered in the papers, but summarized within Chapter 2.

Modelling

Knowledge of the behavior, structure, and interpretation of QTL models from a biological, computational, and mathematical perspective has been critical to all work in this thesis. However, only Paper I is directly concerned with modelling as such.

Paper I

This article deals with the question of whether orthogonal estimates are feasible in a relevant form in spite of only having partial genotype information. Even though genotyping quality has improved over the years, genetic analyses are also more likely to encompass markers of lower quality, where the information content can be limited, making this work relevant. Recent contributions to the field [5, 84] have attempted to achieve orthogonality in more general configurations, but only in the case of full information. We describe how the residual sum of squares is influenced by the lack of full information, and also suggest an imputation method where the parameter estimates are independent

(and thus orthogonal). We show in practical small examples how a traditional Haley-Knott regression might misjudge the effects and even overestimate the phenotype for a genotype class beyond the range of phenotype values actually observed.

Multi-dimensional QTL Searches

The practice of optimization in the sense of identifying the set of optimal QTL has been the original main motivation for this project. This thesis built on the foundation laid using the DIRECT algorithm [55], making it more accessible to non-technical users, and improving the theoretical foundation for its use for QTL searches. The subproject on tracing inheritance was initiated when it became clear that existing standard tools produced disjoint sequences of line-origin probabilities that destroyed the results on the Lipschitz bound found in Paper II.

Paper II

Paper II improves upon the previous use of the DIRECT Lipschitz optimization method [55], by imposing a branch-and-bound criterion. The results are based on a novel transformation of the residual sum of squares, which is the objective function previously used for DIRECT. The transformed objective function that we introduce under the name of LogVar, is shown to be Lipschitz continuous with a well-defined Lipschitz constant in many settings, given a theoretical infinite-size population and perfect recombination frequencies.

Actual real populations are not infinite-size. Therefore, the original DIRECT iteration approach is retained, but it is extended by an additional pruning condition based on a quantile of the objective function distribution, bounding the possible values in the vicinity of an optimum (with some residual probability ε). Using simulations, we show how this condition gives rise to a termination condition resulting in the correct optimum being found with very high probability. Using this condition, extremely high performance can be achieved in permutation tests. When starting from a set of candidate models with a specific objective function value f_{min}, DIRECT searches on permuted sets can be performed, with the goal of finding an optimum superior to f_{min}. The termination condition will allow most such searches to terminate early, if the candidate f_{min} is significant. If it is not, the permutation test run itself can instead be cancelled, since the candidate is no longer of interest. We use the name PruneDIRECT for this augmented version of the DIRECT algorithm using the new termination condition.

Based on these advances, we demonstrate how it becomes easily feasible to complete three-dimensional QTL searches including permutation testing in a matter of hours on a single CPU core, a process that has until now been

prohibitively demanding for users not having access to high-performance computing resources and know-how.

Paper III

The larger datasets and the interest in epistasis and accompanying multidimensional models increase the computational needs for QTL analysis radically. Therefore, there is a clear motivation to utilize modern high-performance computing resources, such as computational grids. In this paper, we extend on previous work by building an interface for the R statistical environment [1] to a DIRECT-based implementation for QTL searches in multiple dimensions. Efficient random permutation tests will also require substantial computational resources, especially in order to determine relevant thresholds for high significance levels. We also use the resulting package for performing simulated model selection based on a permutation testing criterion.

Tracing Inheritance

Most of the work in the thesis has been related to hidden Markov models for tracing inheritance. This is also the part of the work that can be applied to more general model structures, including cases with known, but complex, pedigrees (i.e. not only line crosses). In addition to the papers listed here, progress has also been made on transferring the model methodology for data with known pedigrees in Papers IV-VIII to the setting of unknown pedigrees and individuals without close genetic relationships. Details on that work are found in Chapter 2.

Paper IV

In this conference paper, hidden Markov models are used to determine genotype probabilities. The use of HMMs for genetics was first introduced over 20 years ago [48, 49]. The approach has also been used in specific cases for QTL analysis [13, 70]. In Paper IV, we present a general tool replicating the results for outbred populations in [29], but now exploiting HMMs, reducing the algorithm runtime complexity from exponential to log-linear. This development is critical in modern applications, as the total number of markers possible to genotype has increased tremendously. Furthermore, while it is simple to inspect a larger number of markers, the general data quality will decrease. Values will be missing for some individuals, or there will be ambiguities. These properties were handled in [29], but with an algorithm requiring an exponential increase in time and memory use, making the problem intractable in practice.

In this paper, we also introduce our methods for using the same HMMs for haplotyping, inferring the phasing for linked markers in especially the F_0 and F_1 generations. This is a critical development for making the genotype probabilities in the F_2 generation independent. If the probabilities are not independent, a proper analysis of the population should take the dependence structure into account, something current models will not do. Hence, the probability adjustments due to haplotyping will not only decrease the uncertainty of the F_2 probabilities, but also remove a bias in the estimates.

Paper V

This journal article uses the non-haplotyping code and model developed in Paper IV, exploring in detail its performance and accuracy compared to the then-existing genotype probability implementation in GridQTL [69]. Accuracy is found to be clearly improved, and so is the identification of the specific regions for crossover events. Some short stretches of crossing overs were completely missing in the GridQTL results. Furthermore, the scalability benefits are illustrated by a difference in runtime in the range of orders of magnitude, and the fact that our full intended dataset could not even be reliably analyzed using GridQTL.

Paper VI

This application note presents the code and model (without haplotyping) presented and analyzed in Papers IV, V, in a user-friendly package within the R statistical environment [1], combined with other tools for preparing and analyzing primarily outbred linecross QTL data. The hidden Markov models developed within the scope for this thesis underlie several of the main use cases of the resulting MAPfastR package.

Paper VII

This paper details the use of our haplotyping approach for data with known pedigrees, using a multi-generational simulated dataset. Since our focus pedigrees at most go two generations back, it was relevant to verify performance for thousands of individuals spanning five generations. We were able to fully reconstruct the phase, based on skewness values, without any deviations in all 20 founder individuals, despite the fact that most heuristics-based methods would find founders the hardest to phase, since there would be no parental genotypes to use for constraining the possible options.

Paper VIII

After introducing skewness and demonstrating its use in Paper VII, we here propose to introduce the sureness parameter as well. The resulting method is described in greater detail than before, making this paper the authoritative description of our haplotyping approach. The effectiveness in inferring missing genotypes is demonstrated by artificially removing all founder genotypes from a two-generation simulated dataset, reconstructing them only based on offspring genotypes. Our approach is successful, even when we add an artificial 2% error rate to the offspring data. Comparisons are made against the leading Merlin package. That package is also based on Markov models, but with no explicit parametrization of phase, as it is instead tracking all meioses in a single Markov chain. The comparison demonstrates better scaling for our approach, allowing far larger datasets to be analyzed, as well as superior results for datasets of all sizes.

Paper IX

Several methods for phasing populations with unknown individual relationships rely on sampling possible haplotype distributions in a Markov chain Monte Carlo-like scheme. In this paper, we discuss one way in which the chain sampling process might be distorted and locked in a small region (based on random initialization values), rather than exploring the possible set of haplotype resolutions more in full. The case arises when long regions of identical genotype pairs exist between different individuals. Such situations are more likely to occur if siblings or other close relatives of certain configurations are included in the dataset, whether that is done consciously or not. We demonstrate that our slight change to the popular MaCH [50, 51] phasing and genotype imputation tool improves results in a real-world dataset, even though the individuals extracted from that dataset were not supposed to be close relatives.

6. Svensk sammanfattning

Utvecklingen av laboratorieteknik och automatisering under de senaste decennierna har inneburit att det i dagsläget är möjligt att läsa hela eller delar av den genetiska koden för enskilda individer av olika arter på detaljnivå. Utifrån detta blir det relevant att utforma studier och experiment där man också kan tolka och dra nytta av detta potentiellt överväldigande material.

En typ av sådana studier är att leta efter genetiska positioner (loci) som på något sätt styr eller påverkar specifika kvantitativa egenskaper, så kallade *quantitative trait loci* (QTL). QTL-analys kan göras med syftet att sedan avla på de genvarianter som ger önskat resultat, men det kan minst lika gärna vara ett värdefullt verktyg i medicinsk eller biologisk grundforskning. Genom att identifiera var i genomet det finns genvarianter som påverkar viktiga egenskaper blir det sedan möjligt att studera dessa regioner mer i detalj, genom till exempel transgena organismer med förändringar i de regionerna, nya korsningsexperiment eller jämförelser med liknande gener i genetiska databaser. På det viset kan en gen som visar sig vara relaterad till en egenskap i hund eller mus kopplas till motsvarande gen i människa.

Arbetet i den här avhandlingen innehåller nya metoder för att effektivt identifiera nätverk av QTL:er. Det är ganska naturligt att tänka sig att en variant av en viss gen kan påverka egenskaper på olika sätt, beroende på vilka varianter som finns av andra gener. När man då jämför de egenskaper man kan observera med det genetiska materialet blir det nödvändigt att ta hänsyn till variationen i flera positioner samtidigt. Om man vill testa 100 000 genetiska positioner finns det i storleksordningen 10 miljarder möjliga par och 1 000 biljoner möjliga uppsättningar av tre positioner. Det blir omöjligt att undersöka var och en av dessa möjligheter. Vi har därför försökt göra de statistiska modellerna ortogonala, vilket skulle tillåta en mer flexibel hantering av flera genetiska positioner samtidigt. Vi har även visat hur man med mycket stor säkerhet kan skära ned det antal kombinationer som testas och ändå hitta de kombinationer som faktiskt är relevanta på ett automatiserat sätt, motiverat med statistisk teori.

QTL-analys är enklast att tillämpa på besläktade individer. Ett grundläggande exempel är när det finns två sorter av samma art som korsas med varandra så man får olika hybrider och i den avkomman kan jämföra vilka positioner som ger vilka egenskaper. Information om släktskap mellan individer måste emellertid kompletteras med information om faktiskt genetiskt släktskap i olika delar av genomet. I varje generation blandas det material som erhållits från respektive förälder och för att göra en korrekt analys bör man för varje

lokus i möjligaste mån spåra arvsmönstret, ibland över flera generationer. Det är nödvändigt för att avgöra vilka individer som faktiskt bär på samma genvariant. Problemet kompliceras av att experimentellt bestämda gendata kan vara ofullständiga och innehålla fel, särskilt med moderna experimentflöden där processen är mer automatiserad och inte innehåller lika mycket manuell kontroll som tidigare.

Vi har därför arbetat med statistiska modeller och effektiva datorimplementationer av dessa för att spåra arv. Det viktigaste bidraget består i att aktivt rekonstruera varje individs två kopior av genomet (som erhållits från respektive förälder), medan somliga andra metoder enbart gör detta implicit och därmed tillåter arvsmönster som i praktiken vore omöjliga. Våra metoder kan hantera fler genetiska positioner och fler individer än tidigare spridd metodik och ger dessutom bättre resultat i de fall metoderna är direkt jämförbara. Huvuddelen av detta arbete berör spårning av arv mellan individer med känt släktskap, men vi har även gjort vissa bidrag rörande obesläktade individer, vilka kan vara mer direkt tillämpliga i humangenetiska studier.

Acknowledgments

Life might sometimes be like a theme park ride. And you keep wondering who certified it.

First of all, I would like to thank my family. Some parts of it are newcomers that came along during the ride. Thank you Jessica for finding me (or allowing yourself to be found) and marrying me. There is nothing local about this optimum and my love for you is boundless. I am also grateful that our first true breeding experiment, Casper, has been an immense success. Despite everything else that has happened during these five years, the two of you are what I truly want to keep with me and treasure for the rest of my life.

I would also like to thank my parents, Folke A. Nettelblad and Karin Nettelblad, who have been there all the ride. Some patches have been rough during these years, but you have always listened to my thoughts, professional and personal. I would specifically like to thank you for the extraordinary amount of proofreading on the side that I have got from you at various stages in the life of manuscripts, as well as for patiently listening to me talking about exceedingly esoteric "skewness inversions" and whatnot when we met. When half the terms your son is talking about are words he more or less made up himself, you know that he is a scientist or a crank, or just possibly a superposition of both.

Robert Rosén, you entered your PhD studies a bit later than me, but it has been a relief to discuss different aspects of life with someone in a similar situation in so many ways, and in addition someone as wise as you. I am honored to be your friend. Let us hope that we are awarded a patent in the end for that weird discussion idea.

Without Örjan Carlborg, the scientific interest within the community for epistatic QTL would probably be far less than what it is today, and I would most likely not have entered this field. Had it not been for a bulletin board note (not a poster, a mere sheet of paper) about a summer project job (which I skipped, in the end), I would maybe have become a PhD student, but in another field. I would also like to thank my assistant supervisor José M. Álvarez-Castro, first and foremost for giving me a unique insight in the issues and structure of linear regression models for QTL analysis. Thank you also for bearing with me when I have been too busy with coursework or other projects to keep properly in touch, and for sharing your views on the broader scientific community. Finally, I would also thank you for giving me the opportunity to learn more about turbot sex determination than I had ever expected, even

though that work never became a publication. (Turbot is a species of flatfish which is also actively farmed.)

Sverker Holmgren has been my main supervisor and also stayed along for the full ride. You have instilled the importance of understanding the organizational and political context of the scientific community. You have also shown understanding when I accepted positions for the student union and PhD student council. I know that not all advisors have shared that attitude. I would also like to thank you for letting me develop the aspects and ideas I found most promising. Although the reception has been limited, I would consider Papers IV-IX to be the most important part of this work.

During my years, I have collaborated with fellow PhD students Mahen Jayawardena, Salman Toor and Behrang Mahjani here at the department. Although my main interests may have been different, collaborating and discussing with you has been a worthwhile and rewarding experience, especially the work with Behrang that led to us jointly developing the proper statistical analysis in Paper II.

At TDB, I would also want to mention former head of division Tom Smedsaas and professor Michael Thuné for acting supportive in times of need. Tom has also had a crucial role in creating a welcoming and generous climate at the division. Another specific person I want to acknowledge here is Elisabeth Larsson, whom I consider a role model. I would also like to thank you for providing a mental smorgasbord through the Mathematical and Computational Consulting (MC^2) course, which I attended early on.

Uppsala Learning Lab has been my second professional home during these years, with different forms of technical work on the Uppsala University Student Portal. It has sometimes been a relief to do something down-to-earth with real applications and real users. Specifically, I would like to thank the lab head during most of this time, Mia Lindegren, now the university IT Director, for giving me this opportunity.

Finally, there are numerous acquaintances with varied subject background that I have drafted more or less willingly as random proofreaders over the years. Some of them are mentioned in different capacities above, but I would specifically like to mention David Ekstrand, Ylva Lindahl, Carin Wesslau, Olov Winstrand, and Siavosh Zarrasvand, who helped me in the final stages. You have reduced the amount of abruptly completed sentences or trains of thought, as well as the number of awkward hyphenations, congruential inconsistencies, and "innovative" spellings and expressions. Any remaining occurrences of this sort are the sole responsible of this author, and should be considered a sign of personal flavor.

This work was initially supported by the Graduate School in Mathematics and Computing (FMB), although the school did not last for the ride, but was disbanded halfway along the way.

Je suis tombé par terre, c'est la faute à Voltaire. [35]

Almost as an afterthought, I should also thank Janos Hajdu for waving with the ticket for the next ride eagerly enough to convince me that the pastures at BMC, on the other side of those grazed by sheep at Polacksbacken, might in fact be greener. It was actually time for me to get off in the middle of a loop, with formally one year of financing remaining.

References

[1] The R project for statistical computing. http://www.r-project.org.

[2] G. R. Abecasis, S. S. Cherny, W. O. Cookson, and L. R. Cardon. Merlin – Rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1):97–101, 2001.

[3] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[4] D. M. Altshuler, R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, F. Yu, P. E. Bonnen, PI. de Bakker, P. Deloukas, S. B. Gabriel, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010.

[5] J. M. Alvarez-Castro and Ö. Carlborg. A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics*, 176(2):1151–1167, 2007.

[6] W. Banzhaf, F. D. Francone, R. E. Keller, and P. Nordin. *Genetic programming: an introduction: on the automatic evolution of computer programs and its applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[7] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[8] J. Bergelson and F. Roux. Towards identifying genes underlying ecologically relevant traits in *arabidopsis thaliana*. *Nature Reviews Genetics*, 11(12):867–879, 2010.

[9] M. Bogdan, J. K. Ghosh, and R. W. Doerge. Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. *Genetics*, 167(2):989–999, 2004.

[10] T. Brants. TnT: a statistical part-of-speech tagger. In *Proceedings of the 6th conference on Applied natural language processing*, pages 224–231. Association for Computational Linguistics, 2000.

[11] K. W. Broman. *Identifying quantitative trait loci in experimental crosses*. PhD thesis, Department of Statistics, University of California, Berkeley, 1997.

[12] K. W. Broman and T. P. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(4):641–656, 2002.

[13] K. W. Broman, H. Wu, S. Sen, and G. A. Churchill. R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7):889–890, 2003.

[14] G. R. Brown, D. L. Bassoni, G. P. Gill, J. R. Fontana, N. C. Wheeler, R. A. Megraw, M. F. Davis, M. M. Sewell, G: A. Tuskan, and D. B. Neale. Identification of quantitative trait loci influencing wood property traits in

Loblolly pine (*pinus taeda l.*). iii. QTL verification and candidate gene mapping. *Genetics*, 164(4):1537–1546, 2003.
[15] Ö. Carlborg, L. Andersson, and B. Kinghorn. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics*, 155:2003–2010, 2000.
[16] Ö. Carlborg and C. S. Haley. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, 2004.
[17] G. A. Churchill and R. W. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138:963–971, 1994.
[18] G. A. Churchill and R. W. Doerge. Naive application of permutation testing leads to inflated Type I error rates. *Genetics*, 178(1):609–610, 2008.
[19] O. Delaneau, C. Coulonges, and J.-F. Zagury. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, 9(1):540, 2008.
[20] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
[21] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.
[22] B. Feenstra, I. M. Skovgaard, and K. W. Broman. Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimation equations. *Genetics*, 173:2269–2282, 2006.
[23] D. E. Finkel. *DIRECT Optimization Algorithm User Guide*. North Carolina State University, March 2003.
[24] R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, (52):399–433, 1918.
[25] E. Galerpin. The cubic algorithm. *Journal of Mathematical Analysis and Applications*, pages 635–640, 1985.
[26] Y. Guo, J. Li, A. J. Bonham, Y. Wang, and H. Deng. Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies. *European Journal of Human Genetics*, 17(6):785–792, 2008.
[27] J. B. S. Haldane. The combination of linkage values, and the calculation of distance between the loci of linked factors. *Journal of Genetics*, 8:299–309, 1919.
[28] C. S. Haley and S. A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4):315–24, 1992.
[29] C. S. Haley, S. A. Knott, and J. M. Elsen. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, 136(3):1195–1207, 1994.
[30] L. Han and S. Xu. A Fisher scoring algorithm for the weighted regression method of QTL mapping. *Heredity*, 101:453–464, Nov 2008.
[31] P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564, Aug 2009.
[32] L. D. W. Hillier et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*,

432(7018):695–716, December 2004.
[33] D. D. Houston, C. S. Haley, A. L. Archibald, and K. A. Rance. A QTL affecting daily feed intake maps to chromosome 2 in pigs. *Mammalian Genome*, 16:464–470, 2005.
[34] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, 06 2009.
[35] V. Hugo. *Les Misérables*, volume 5:1, chapter 15. A. Lacroix, Verboeckhoven & C:o, Brussels, 1862.
[36] M. M. Iles. What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genetics*, 4(2):e33, 02 2008.
[37] R. C. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135(1):205–211, 1993.
[38] M. Jayawardena and S. Holmgren. Grid-enabling an efficient algorithm for demanding global optimization problems in genetic analysis. In *3rd IEEE International Conference on e-Science and Grid Computing, IEEE Conference Proceedings 10.1109*, pages 205–212. IEEE, 2007.
[39] M. Johannesson, R. Lopez-Aumatell, P. Stridh, M. Diez, J. Tuncel, G. Blazquez, E. Martinez-Membrives, T. Canete, E. Vicens-Costa, D. Graham, et al. A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: The NIH heterogeneous stock. *Genome Research*, 19(1):150–158, Oct 2008.
[40] D. Jones, C. Perttunen, and B. Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Application*, 79:157–181, 1993.
[41] B. H. Juang and L. R. Rabiner. Hidden Markov models for speech recognition. *Technometrics*, pages 251–272, 1991.
[42] C. H. Kao. On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics*, 156:855–865, 2000.
[43] S. Kerje, Ö. Carlborg, K. Schütz, C. Hartmann, P. Jensen, and L. Andersson. The twofold difference in adult size between the red junglefowl and white leghorn chickens is largely explained by a limited number of QTLs. *Animal Genetics*, 34(4):264–274, 2003.
[44] J. H. M. Knoll, R. D. Nicholls, R. E. Magenis, J. M. Graham, M. Lalande, S. A. Latt, John M. Opitz, and J. F. Reynolds. Angelman and Prader-Willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. *American Journal of Medical Genetics*, 32(2):285–290, 1989.
[45] A. Kong, G. Masson, M. L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P. I. Olason, A. Ingason, S. Steinberg, T. Rafnar, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068–1075, 2008.
[46] D. Kosambi. The estimation of map distances from recombination values. *Annals of Eugenics*, pages 172–175, 1944.
[47] E. S. Lander and D. Botstein. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199, 1989.
[48] E. S. Lander and P Green. Construction of multilocus genetic linkage maps in

humans. *Proceedings of the National Academy of Sciences of the United States of America*, 84(8):2363–2367, 1987.

[49] E. S. Lander, P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newburg. Mapmaker: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, 1(2):174 – 181, 1987.

[50] Y. Li, C. Willer, S. Sanna, and G. Abecasis. Genotype imputation. *Annual review of genomics and human genetics*, 10:387, 2009.

[51] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8):816–834, 2010.

[52] K. Lindblad-Toh et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438:803–819, Dec 2005.

[53] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, New York, 2nd edition, 1987.

[54] K. Ljungberg. Efficient evaluation of the residual sum of squares for quantitative trait locus models in the case of complete marker genotype information. Technical Report 2005-033, Division of Scientific Computing, Department of Information Technology, Uppsala University, 2005.

[55] K. Ljungberg, S. Holmgren, and Ö. Carlborg. Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. *Bioinformatics*, 20(12):1887–1895, 2004.

[56] M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, 1 edition, January 1998.

[57] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[58] O. Martinez and R. Curnow. Estimating the location and the sizes of effects of quantitative trait loci flanking markers. *Theoretical and Applied Genetics*, 85:480–488, 1992.

[59] R. Mladineo. An algorithm for finding the global maximum of a multimodal, multivariate function. *Mathematical Programming*, pages 253–271, 1987.

[60] M. W. Nachman and S. L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, 2000.

[61] C. Nettelblad. Model selection criteria in the NOIA framework for gene interaction. Master's thesis, School of Engineering, Uppsala University, Sweden, October 2007.

[62] J. Pinter. Globally convergent methods for n-dimensional multiextremal optimization. *Optimization*, 17:187–202, 1986.

[63] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[64] L. R. Rabiner and B. Juang. An introduction to hidden Markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[65] A. Rebai, B. Goffinet, and B. Mangin. Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, 138(1):235–240, 1994.

[66] I. W. Saunders, J. Brohede, and G. N. Hannan. Estimating genotyping error

rates from Mendelian errors in SNP array genotypes and their impact on inference. *Genomics*, 90(3):291–296, Sep 2007.

[67] K. E. Schütz, S. Kerje, L. Jacobsson, B. Forkman, Ö Carlborg, L. Andersson, and P. Jensen. Major growth QTLs in fowl are related to fearful behavior: possible genetic links between fear responses and production traits in a red junglefowl x white leghorn intercross. *Behavior Genetics*, 34:121–130, 2004.

[68] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[69] G. Seaton, J. Hernandez, J. A. Grunchec, I. White, J. Allen, DJ. De Koning, W. Wei, D. Berry, C. Haley, and S. Knott. GridQTL: A grid portal for QTL mapping of compute intensive datasets. In *Proceedings of the 8th world congress on genetics applied to livestock production*, pages 13–18, 2006.

[70] Ś. Sen and G. A. Churchill. A Statistical Framework for Quantitative Trait Mapping. *Genetics*, 159:371–387, 2001.

[71] B. Shubert. A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, pages 379–388, 1972.

[72] M. J. Sillanpää and J. Corander. Model choice in gene mapping: what and why. *Trends in Genetics*, 18(6):301–307, 2002.

[73] J. Slate. Quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Molecular Ecology*, 14:363–379(17), February 2005.

[74] M. Szydlowski and P. Paczyńska. QTLMAS 2010: simulated dataset. In *BMC Proceedings*, volume 5 (Suppl 3), page S3. BioMed Central Ltd, 2011.

[75] W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. Nicholas, P. Rawlins, R. Mott, and J. Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, 38(8):879–887, Jul 2006.

[76] Voltaire. *La Bégueule: Conte Moral*. Genève, 1772.

[77] C. M. Wade et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, 326:865–867, Nov 2009.

[78] R. H. Waterston et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, Dec 2002.

[79] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[80] S. Xu. A comment on the simple regression method for interval mapping. *Genetics*, 141:1657–1659, 1995.

[81] S. Xu. Further investigation on the regression method of mapping quantitative trait loci. *Heredity*, 80:364–373, 1998.

[82] M. Zak, A. Baierl, M. Bogdan, and A. Futschik. Locating multiple interacting quantitative trait loci using rank-based model selection. *Genetics*, 176(3):1845–1854, 2007.

[83] Z-B. Zeng, J. Liu, L. F. Stam, C-H. Kao, J. M. Mercer, and C. C. Laurie. Genetic architecture of a morphological shape difference between two drosophila species. *Genetics*, 154(1):299–310, 2000.

[84] Z-B. Zeng, T. Wang, and W. Zou. Modeling quantitative trait loci and interpretation of models. *Genetics*, 169(3):1711–1725, 2005.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 973

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology.

Distribution: publications.uu.se
urn:nbn:se:uu:diva-180920

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2012