Bayesian inference for retrospective population genetics models using Markov chain Monte Carlo methods

Matti Pirinen

Department of Mathematics and Statistics
Faculty of Science
University of Helsinki

Academic dissertation

To be presented, with the permission of the Faculty of Science
of the University of Helsinki, for public criticism in
Auditorium A129, Chemicum (A. I. Virtasen aukio 1),
on June 8, 2009, at 12 noon.

HELSINKI 2009

Figures 1 and 2 of the summary part and the figure on the back cover were created by Matti Pirinen. Figure 2(a) and the figure on the back cover were drawn using the program *Pedfiddler*.

Author's email address:
matti.pirinen@iki.fi

ISBN 978-952-92-5619-8 (paperback)
ISBN 978-952-10-5602-4 (PDF)
http://ethesis.helsinki.fi
Helsinki University Printing House
Helsinki 2009

Supervisor	Prof. Elja Arjas
Co-supervisors	Dr. Dario Gasbarra
	Dr. Mikko J. Sillanpää
	Department of Mathematics and Statistics,
	University of Helsinki,
	Finland
Pre-examiners	Prof. Ola Hössjer
	Department of Mathematics,
	Division of Mathematical Statistics,
	Stockholm University,
	Sweden
	Dr. Petter Mostad
	Department of Mathematical Sciences,
	Chalmers University of Technology,
	Göteborg, Sweden
Custos	Prof. Aapo Hyvärinen
	Department of Mathematics and Statistics,
	University of Helsinki,
	Finland
Opponent	Prof. John Whittaker,
	Department of Epidemiology and Population Health,
	London School of Hygiene & Tropical Medicine,
	London, U.K.

Contents

List of original articles

I Pirinen M and Gasbarra D. 2006. Finding consistent gene transmission patterns on large and complex pedigrees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 3:252-262.

II Gasbarra D, Pirinen M, Sillanpää M J, Salmela E and Arjas E. 2007. Estimating genealogies from unlinked marker data: a Bayesian approach. *Theoretical Population Biology* 72:305-322.

III Gasbarra D, Pirinen M, Sillanpää M J, and Arjas E. 2007. Estimating genealogies from linked marker data: a Bayesian approach. *BMC Bioinformatics* 8:411.

IV Gasbarra D, Pirinen M, Sillanpää M J, and Arjas E. 2009. Bayesian QTL mapping based on reconstruction of recent genetic histories. *(submitted)*

V Pirinen M, Kulathinal S, Gasbarra D and Sillanpää M J. 2008. Estimating population haplotype frequencies from pooled DNA samples using PHASE algorithm. *Genetics Research* 90:509-524.

Contributions of the authors of the articles I-V

I MP had a major responsibility for writing the article and was fully responsible for implementing and testing the method. MP and DG designed the algorithm together.

II,III MP took part in the implementation of the MCMC algorithm, tested the program, and conducted the example analyses. DG designed and reported the MCMC algorithm and he had a major responsibility for implementing it. EA initiated and coordinated the study. DG, MP, MJS and EA designed the project jointly and all authors participated in writing the manuscript. ES provided the Finnish data set in article II.

IV MP had a major responsibility for writing the manuscript, he took part in the implementation of the MCMC algorithm, tested the program and conducted the example analyses. DG designed the MCMC algorithm and he had a major responsibility for implementing it. EA initiated and coordinated the study. All authors designed the project jointly and participated in writing the manuscript.

V MP had a major responsibility for writing the article and was fully responsible for implementing and testing the method (except that DG provided the simulated data set). All authors were involved in designing the method and writing the article.

Abstract

Genetics, the science of heredity and variation in living organisms, has a central role in medicine, in breeding crops and livestock, and in studying fundamental topics of biological sciences such as evolution and cell functioning. Currently the field of genetics is under a rapid development because of the recent advances in technologies by which molecular data can be obtained from living organisms. In order that most information from such data can be extracted, the analyses need to be carried out using statistical models that are tailored to take account of the particular genetic processes.

In this thesis we formulate and analyze Bayesian models for genetic marker data of contemporary individuals. The major focus is on the modeling of the unobserved recent ancestry of the sampled individuals (say, for tens of generations or so), which is carried out by using explicit probabilistic reconstructions of the pedigree structures accompanied by the gene flows at the marker loci. For such a recent history, the recombination process is the major genetic force that shapes the genomes of the individuals, and it is included in the model by assuming that the recombination fractions between the adjacent markers are known. The posterior distribution of the unobserved history of the individuals is studied conditionally on the observed marker data by using a Markov chain Monte Carlo algorithm (MCMC). The example analyses consider estimation of the population structure, relatedness structure (both at the level of whole genomes as well as at each marker separately), and haplotype configurations. For situations where the pedigree structure is partially known, an algorithm to create an initial state for the MCMC algorithm is given.

Furthermore, the thesis includes an extension of the model for the recent genetic history to situations where also a quantitative phenotype has been measured from the contemporary individuals. In that case the goal is to identify positions on the genome that affect the observed phenotypic values. This task is carried out within the Bayesian framework, where the number and the relative effects of the quantitative trait loci are treated as random variables whose posterior distribution is studied conditionally on the observed genetic and phenotypic data.

In addition, the thesis contains an extension of a widely-used haplotyping method, the PHASE algorithm, to settings where genetic material from several individuals has been pooled together, and the allele frequencies of each pool are determined in a single genotyping.

1. Introduction

For thousands of years humans have made observations that certain characteristics among individuals, whether animals or plants, are governed by heredity. In its simplest form this is manifested by a tendency of offspring to resemble their parents with respect to some trait like size, color or shape. These vague ideas were brought under the scientific approach about 150 years ago by the Augustinian monk Gregor Mendel (1822-1884), who conducted the famous series of breeding experiments with pea plants [27]. Mendel's experiments suggested that hereditary material was transmitted in discrete units between the generations, and that the hereditary entities of the parents maintained their integrity in the offspring, rather than blended together. Later these units of inheritance were named *genes*; the origin of the term may be traced back to Greek words of *genesis* ("birth") and *genos* ("origin").

Genetics, as the study of heredity is now called, experienced a tremendous progress during the 20th century, most notably because the advances in materials sciences made it possible to reveal the molecular basis of inheritance. In the 1950s the deoxyribonucleic acid (DNA) was already identified as the carrier of the genetic information. In essence, DNA consists of a long sequence of molecules of four different types, conventionally denoted by letters A,C,G and T, whose ordering in the sequence codes the hereditary information of the living organisms. The detection of the genetic code with its seemingly simple structure of four-letter alphabet raised high hopes of discovering the causes of complex traits. So far, however, these goals have been fulfilled only partially.

In the 21th century genomic research is facing a flood of data from rapidly evolving laboratory techniques. The Human Genome Project was completed in 2003, the human variation at over 3 million loci has currently been cataloged by the HapMap project [20] and an even more refined map of the human genome will follow in the next few years as 1000 Genomes Project proceeds [1]. The question no longer is how we can extract data from cells or genomes but rather how can we acquire biologically meaningful knowledge from the available wealth of data. Excepting a few simple Mendelian traits, the biological reality has turned out to be a complex entanglement of the environment and the genome. The urgent need for quantitative methods to discover the relevant pieces of information from a jungle of noise is ever strengthening the role of statistics and computer science in genetics.

Bayesian statistics provides a consistent framework for learning from data. Its roots are in the 18th century works of Thomas Bayes (c. 1702-1761) and Pierre-Simon Laplace (1749-1827). After almost falling into oblivion during the first half of the 20th century, Bayesian statistics experienced a new rise beginning in the late 1980s, as the available computational resources and methods had advanced to the level which enabled analyzing more realistic probability models. In particular, Markov chain Monte Carlo (MCMC) algorithms have had a key role in the resurgence of the Bayesian computation.

This thesis brings together five scientific articles, four of which consider analyses of Bayesian models in genetics by using MCMC methods. The remaining article I is

of an algorithmic nature and introduces a method for building an initial state for the MCMC algorithm that has been used in articles II, III and IV. A unifying theme in these articles is an attempt to more thoroughly utilize our knowledge of the biological processes in modeling genetic data sampled from contemporary individuals. This work can be seen as another product of the Bayesian revolution that has taken over during the 20 years: another step towards more realistic probability models in different fields of science. Here the fundamental question is to estimate how are the individuals related to each other in different parts of the genome, given their genotype data. In addition to the direct applications to relatedness and relationship estimation, the question is essential in gene mapping, where the purpose is to identify such positions from the genome, that are shared among the individuals, who also share certain phenotypic properties.

1.1. **Main questions.** Here is a short description of the specific questions that are studied in this thesis. The rest of the summary part provides an introduction to the concepts and terminology that are used below.

Article I introduces an algorithm that extends partially observed genotype data at a single marker locus to the whole pedigree in accordance with the Mendelian inheritance. The algorithm can be used to verify the consistency between the observed pedigree and partially observed marker data as well as to create initial states for MCMC algorithms on pedigrees.

Article II introduces a model for the unobserved recent history of the sampled contemporary individuals. The model can be used to estimate the relatedness between the individuals both in terms of pedigree relationships and of identical-by-descent sharing of marker alleles. Furthermore, the model simultaneously captures the relatedness structure in different scales (e.g. weaker population structure and stronger family structure). The model is analyzed conditionally on the observed marker data at unlinked loci.

Article III extends the model of article II to linked marker loci, which makes it possible to estimate also haplotype configurations and allele sharing along a chromosome.

Article IV extends the model of article III to settings where also a quantitative phenotype has been measured from the sampled individuals. The model can be used for gene mapping, i.e., to find locations and relative effects of genetic variants that affect the observed phenotypic values.

Article V extends the widely-used haplotyping software PHASE [42] to settings where the genetic material of the sampled individuals is divided into pools, and the allele frequencies within each pool are determined in a single genotyping. The goal is to estimate the haplotype frequencies of the sampled individuals based on the observed pooled data.

2. Bayesian probability modeling

The fundamental question in the field of statistics is that of inference. Some data are observed and we wish to make statements about the unknown process or system that gave rise to these data. Only in rare occasions are we able to gain the complete certainty about the underlying circumstances. Instead, in most cases we remain *uncertain* to some degree and our knowledge seems to be best described by using statements that involve *probabilities*. Thus the modern statistics has become the science of formulating, evaluating, updating and interpreting such probabilities.

The axiomatic probability theory laid down by Andrey Kolmogorov (1903-1987) in 1933 has become the established mathematical description of the concept of probability. Despite large unanimity with respect to the logical structure of the theory, there exist several interpretations of probability when it is applied to the real world phenomena.

2.1. **Bayesian approach.**

The Bayesian interpretation considers probability as a means to quantify one's beliefs about any phenomenon that involves uncertainty. Thus the role of the individual, the one whose beliefs are quantified, is decisive in Bayesian modeling. Different individuals may have different prior knowledge on the subject and hence their conclusions from the same data may also differ from each other. Also the knowledge of any particular individual evolves in time as it becomes updated by new information.

A fundamental idea in Bayesian statistics is to treat all unknown quantities in the model equally as *random variables*, independently of the particular roles that they have in the model, for example, whether they are parameters or latent or unobserved quantities. This results in a universal framework where the information about any unknown quantity is captured by a probability distribution. Bayesian statistics provides the rules for updating these distributions as new data are observed. Before the observations are made, the knowledge of the modeler is represented by his/her *prior* distribution. The observations then transform the prior to a *posterior* distribution, according to the rules of probability calculus. This learning procedure can be continued in a natural way by always considering the achieved posterior distribution as a new prior for subsequent observations.

The link between prior and posterior is *Bayes' formula*. Its name comes from Thomas Bayes (c. 1702-1761), a British mathematician and Presbyterian minister who formulated a special case of it [4]. In order to express Bayes' formula let us divide the considered random variables into two sets Y and Θ, where the former represents the variables for which we have observed some estimates and the latter contains the unobserved variables in which our interest lies. The (density of the) joint probability model $p(Y, \Theta)$ is usually specified in parts by using the chain rule of probabilities

$$p(Y, \Theta) = p(Y|\Theta)p(\Theta),$$

where $p(\Theta)$ is the prior distribution of Θ and the *likelihood* $p(Y|\Theta)$ describes our conception of the structure of the process giving rise to the data Y in terms of the

unknown variables Θ. If our observation is $Y = y$, then Bayes' formula states that

$$p(\Theta|Y = y) = \frac{p(Y = y|\Theta)p(\Theta)}{p(Y = y)}. \tag{2.1}$$

In the Bayesian approach this is interpreted as a way to update the knowledge about Θ. It tells how the observations $Y = y$ transform the prior knowledge about Θ, expressed by the probability distribution $p(\Theta)$, into the posterior distribution $p(\Theta|Y = y)$. The additional term $p(Y = y)$ is the marginal probability (density) of the data and is thus a constant once $Y = y$ has been observed. Comprehensive references on Bayesian statistics are e.g. [32, 6].

2.2. **Frequentist approach.** An alternative to the Bayesian approach is *frequentist statistics*. A fundamental difference between the two is that in frequentist statistics the unknown quantities Θ are not treated as random variables, but instead it is postulated that there exist some true and fixed values for them that need to be estimated from the observed data. As a consequence there is no prior distribution for Θ and the inference will be based completely on the likelihood function $p(Y|\Theta)$, where Y represents the data. It also follows that, strictly speaking, frequentist statistics cannot make any probability statements about Θ, but instead the concept of randomness is attached to the data. The inference about Θ thus proceeds indirectly through questions like: What is the probability (here meaning the hypothetical relative frequency in a long series of repeated experiments) that if the true value of Θ is in some set A, then we would observe data Y in some set B?

At times the controversies between the supporters of Bayesian and frequentist viewpoints have been quite fierce. But as better understanding of the statistical methodology is spreading, a more fruitful discussion is also possible, as illustrated recently by Gelman [11].

Important merits for the Bayesian approach are that it seems to be fundamentally sound, consistent and unified and that it gives a direct answer to the question we are interested in: How does our personal degree of uncertainty change as we observe new data? Also the formulation of modern complex hierarchical models is well suited for the Bayesian approach.

A source of criticism towards Bayesian statistics has been the use of prior distributions which, by definition, are subjective. By some this has been thought to be contrary to the doctrine of objective science that would ideally have only one truth and no room for subjectivity. However, any kind of probability modeling requires subjective choices and usually quite stringent assumptions regarding, for example, the form of the sampling distribution of the data $p(Y|\Theta)$. Thus it can also be seen as a merit for Bayesian statistics that the subjectivity involved in the modeling is clearly stated, admitted and understood. Furthermore, in cases where substantial prior information exists it is essential to be able to include that into the model.

Another complication in adapting Bayesian statistics use to be the lack of computational methods and resources, which made it impossible to analyze more complex models in practice. But this issue has changed dramatically during the last few

decades due to the development of both the theory of statistical computing and the computer hardware. Some of these advances are considered next.

3. Analyzing Bayesian models

The posterior distribution $p(\Theta|Y)$ is the basis for Bayesian inference. In high dimensional cases the posterior must be summarized using, for example, moments of some numerical functions with respect to the posterior distribution. In case of complex models the analytic integration with respect to the posterior is often impossible. Important ways to overcome this problem are importance sampling and Markov chain Monte Carlo algorithms [35, 9]. In this thesis we consider only the latter methods.

3.1. **Monte Carlo.** Monte Carlo (MC) methods apply randomness to explore properties of functions, for example, to compute integrals. These methods are named after the famous casino, because they utilize repeatable random sampling which resembles games of chance. MC methods trace back to 1930s and 1940s when several physicists and mathematicians (E. Fermi, N. Metropolis, S. Ulam, J. von Neumann, among others) who worked in Los Alamos were looking for ways to utilize new computing devices in their physical calculations [3].

If, for example, we are about to estimate the expectation of function f under distribution π, an MC method would be to sample a sequence $x_1, \ldots, x_n$ independently from π and approximate

$$E_\pi[f] \approx \frac{1}{n} \sum_{t=1}^{n} f(x_t). \tag{3.1}$$

Theoretical justification for the approximation comes from the law of large numbers, which states that the right hand side of equation (3.1) converges to $E_\pi[f]$ almost surely as $n \to \infty$, (given that the expectation is finite).

Monte Carlo integration turns out to be useful especially in high dimensional spaces where numerical methods using integration grids become inefficient. On the other hand, a necessary requirement for a successful application of an MC method is a procedure to sample (efficiently) from the target distribution.

3.2. **Markov chain Monte Carlo.** In more complex cases it may not be possible to sample efficiently an independent sequence from the target distribution π. Fortunately the requirement of independence may be relaxed to *Markov dependence* of the sequence. A stochastic process is said to be *Markov*, if the distribution of the future states is conditionally independent of the past states, given the present state. The goal of the MCMC methods is to produce a Markov chain $(X_t)_{t=0}^{\infty}$ that is *ergodic* and whose *stationary distribution* is π. Then the theory assures that the chain converges to its stationary distribution and that (3.1) becomes a good approximation as $n \to \infty$ for any π-integrable function f. In this thesis Markov chains on finite spaces have a central role and below we will make the above terminology

more precise in that setting. A comprehensive reference on Markov chains is e.g. [29]; extensions to infinite spaces can also be found e.g. in [35, 9].

Suppose that the target distribution π is defined on a finite state space $\mathcal{X}$. A Markov chain $(X_t)_{t=0}^{\infty}$ on $\mathcal{X}$ is a countable sequence of random variables taking values in $\mathcal{X}$ and satisfying the Markov property for all t:

$$P(X_{t+1} = y | X_t = x_t, \ldots, X_0 = x_0) = P(X_{t+1} = y | X_t = x_t).$$

Furthermore, we consider only chains that are *time homogeneous*, that is, the transition probabilities $P(X_{t+1} = y | X_t = x)$ between any states $x, y \in \mathcal{X}$ are independent of t. Thus the distribution of our Markov chain $(X_t)_{t=0}^{\infty}$ is completely defined by its *transition matrix* K, with entries $K_{xy} = P(X_1 = y | X_0 = x)$, together with the initial state x_0 (or initial distribution) of the chain. The rules of matrix algebra and probabilities match in such a way that the transition matrix for n sequential transitions of the chain, denoted by $K^{(n)}$, is given by the matrix power K^n.

A distribution μ (here a row vector) on space $\mathcal{X}$ is said to be a *stationary distribution* of the Markov chain, if $\mu K = \mu$, i.e., if the Markov chain remains distributed as μ ever since it has reached μ for the first time. Every finite state space Markov chain has at least one stationary distribution and the uniqueness of the stationary distribution is guaranteed, if the chain is *irreducible*, i.e., if for all pairs of states x, y there is a positive integer n for which $K_{xy}^{(n)} > 0$ (Thm 2.7. in [15]). In words irreducible chains are those that are able to explore the whole space independently of their initial values.

For finite state spaces the irreducibility already guarantees the following form of the Law of large numbers for Markov chains. Suppose that $(X_t)_{t=0}^{\infty}$ is an irreducible Markov chain with stationary distribution μ and that $f : \mathcal{X} \to \mathbf{R}$. Then

$$\frac{1}{n+1} \sum_{t=0}^{n} f(X_t) \xrightarrow[n \to \infty]{} E_\mu[f], \tag{3.2}$$

regardless of the initial distribution of the chain (Thm 2.11. in [15]). This result is known as the Ergodic theorem for Markov chains.

An irreducible chain may not converge towards its stationary distribution if the chain exhibits certain cyclic behavior. To rule out the periodic chains we say that an irreducible Markov chain (with transition matrix K) is *aperiodic* if for all $x \in \mathcal{X}$

$$\gcd\{n \geq 1 : K_{xx}^{(n)} > 0\} = 1,$$

where gcd denotes the greatest common divisor of numbers. Thus, for an aperiodic chain the waiting times between consecutive visits to any state are not restricted to be multiples of any basic period (larger than 1). An irreducible aperiodic chain with stationary distribution μ converges towards μ in the sense that

$$K_{xy}^{(t)} \to \mu_y \text{ as } t \to \infty,$$

for all $x, y \in \mathcal{X}$ (Corollary of Thm 2.9. in [15]). Thus, independently of the initial state x_0, the state X_t of the chain will eventually (for large enough t) be distributed as closely according to μ as required.

The above properties of the Markov chains suggest that if we could generate an irreducible and aperiodic Markov chain which has a given target distribution as its stationary distribution, then we would have a means to compute (approximately) expectations with respect to the target distribution, as well as to sample (approximately) from it. A sufficient condition for the chain (with transition matrix K) to have π as its stationary distribution is that the chain satisfies the *detailed balance condition* with respect to π, that is,

$$\pi_x K_{xy} = \pi_y K_{yx}, \text{ for all } x, y \in \mathcal{X}.$$

This is because by summing both sides of this equation with respect to y yields $\pi_x = \sum_{y \in \mathcal{X}} \pi_y K_{yx}$, which is equivalent to the matrix equation $\pi = \pi K$, thus showing that π is the stationary distribution of the chain.

For simplicity we have so far formulated the properties and theorems only for the Markov chains on finite state spaces, but they generalize also to infinite state spaces, the difference being mainly in the need for measure theoretic concepts (see e.g. [35, 9]). Next we look at a way to generate a suitable Markov chain for given target distribution π and we no longer restrict the considerations to finite state spaces.

3.2.1. *Metropolis-Hastings algorithm.* Suppose that our target distribution π is defined on a state space $\mathcal{X}$ and that the distribution has a density function p (with respect to some underlying measure), where we interpret $p = \pi$ as the probability mass function in discrete cases. Given that we can compute p pointwise up to a normalizing constant, there exists a very general scheme to sample $x_1, \ldots, x_n$ from a Markov chain which has π as its stationary distribution. The procedure is called Metropolis-Hastings (MH) algorithm (Metropolis et al. [28] and Hastings [16]) and it requires a specification of a *proposal density* $q(\cdot|x)$ that defines a probability distribution on $\mathcal{X}$ for any given $x \in \mathcal{X}$. Metropolis-Hastings algorithm samples the next state x_{t+1}, based on the current state x_t, by the following procedure:

(1) Sample $y_t \sim q(\cdot|x_t)$.
(2) Set

$$x_{t+1} = \begin{cases} y_t & \text{with probability } a(x_t, y_t), \\ x_t & \text{with probability } 1 - a(x_t, y_t), \end{cases}$$

where

$$a(x_t, y_t) = \min \left\{ \frac{p(y_t) q(x_t|y_t)}{p(x_t) q(y_t|x_t)}, 1 \right\}.$$

The idea of the algorithm is to perturb the proposal distribution in such a way that the chain satisfies the detailed balance condition with π as its stationary distribution. This is achieved by modifying the sequence of the proposed states by an *acceptance probability* $a(\cdot, \cdot)$, which in practice works by allowing the chain to maintain its current position for one or more steps.

An advantage of Metropolis-Hastings algorithm is that the proposal distribution can be chosen quite freely. The empirical average converges (with probability 1) towards the expectation as in (3.2), if the chain is π-irreducible, and the convergence

(in total variation norm) of the distribution of X_t to π is guaranteed, if the chain is also aperiodic (Thm. 6.2.5 in [35]). However, the theoretical convergence results are valid only in the limit as the length of the chain approaches infinity. In practice the approximations must always be based on a finite subchain. Thus the design of the applicable MCMC algorithms has become a craft of formulating rapidly mixing proposal distributions which are able to explore the target distribution reasonably well in some given finite time. This is also a central issue in this thesis.

3.3. **Reversible jump MCMC.** An extension of Metropolis-Hastings algorithm called *Reversible jump MCMC* (RJMCMC) was introduced by Green [13]. It allows the Markov chain to move between the spaces of different dimensions. This is necessary in cases where the model consists of submodels that are defined with different numbers of continuous parameters. Such problems are encountered, for example, in mixture modeling, changepoint analysis, and model choice applications.

Suppose that we are studying a target distribution (a probability measure) on space $\mathcal{X} = \prod_{m=1}^{M} \{m\} \times \mathbf{R}^{n_m}$ that includes M different models, and that the distribution can be decomposed into a discrete model probability $p(m)$, as well as to densities for parameters $p(\theta_m|m)$ for all $m = 1, \ldots, M$ (with respect to the underlying Lebesgue measure of each space $\mathbf{R}^{n_m}$). RJMCMC operates through a collection of proposal distributions $q_l(\cdot|\cdot)$, indexed by l, each of which proposes transitions between two particular subspaces in either direction. The subspaces need not be different and usually proposal distributions operating within a single model are also necessary for the mixing of the chain. In line with the original Metropolis-Hastings algorithm, an acceptance probability is then defined in such a way that each move type satisfies the detailed balance condition with respect to the target distribution.

To see how this is usually done, let us consider a single move of type l that operates between subspaces m_1 and m_2. Starting from a current state (m_1, θ_{m_1}), the parameter vector θ_{m_2} of the proposal state (m_2, θ_{m_2}) is determined by first sampling a random variable $u \in \mathbf{R}^{d_1}$ from a density $q_1^l(\cdot)$ and then applying a deterministic differentiable bijection $t_l : \mathbf{R}^{n_l} \to \mathbf{R}^{n_l}$ to have $(\theta_{m_2}, u_2) = t_l(\theta_{m_1}, u_1)$. Here $u_2 \in \mathbf{R}^{d_2}$ and $n_l = n_{m_1} + d_1 = n_{m_2} + d_2$. The move of type l also defines a density $q_2^l(\cdot)$ on $\mathbf{R}^{d_2}$, which allows us to go back from the state (m_2, θ_{m_2}) to the state (m_1, θ_{m_1}) by reversing the process, now sampling u_2 from $q_2^l(\cdot)$ and setting $(\theta_{m_1}, u_1) = t_l^{-1}(\theta_{m_2}, u_2)$.

By extending the parameter vectors θ_1 and θ_2 with u_1 and u_2 in such a way that the dimensions match, we are able to express the density of the joint equilibrium-proposal distribution with respect to a certain measure. This density can now be used to define an acceptance probability that assures the detailed balance condition for this move type with respect to the target distribution. The common form of the acceptance probability is given by

$$a[(m_1, \theta_{m_1}, u_1), (m_2, \theta_{m_2}, u_2)] = \min \left\{ \frac{p(m_2, \theta_{m_2}) q_2^l(u_2) j(l|m_2, \theta_{m_2})}{p(m_1, \theta_{m_1}) q_1^l(u_1) j(l|m_1, \theta_{m_1})} \left| \frac{\partial(\theta_{m_2}, u_2)}{\partial(\theta_{m_1}, u_1)} \right|, 1 \right\},$$

where $j(l|m_i, \theta_{m_i})$ is the probability of attempting a move of type l from state (m_i, θ_{m_i}) and

$$\left| \frac{\partial(\theta_{m_2}, u_2)}{\partial(\theta_{m_1}, u_1)} \right|$$

is the absolute value of the Jacobian of the transformation $(\theta_{m_2}, u_2) = t_l(m_1, \theta_{m_1})$.

In addition to the works of Green [13, 14], more details and some examples of the algorithm can be found e.g. in [39] and [32].

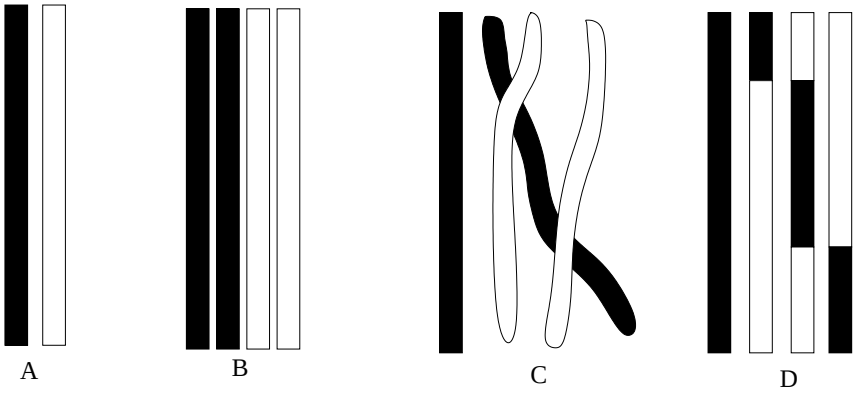
4. Genetic data

In this thesis we consider *diploid*, sexually reproducing species. The *genome* of the species is divided into n *chromosomes*, separate entities that contain the DNA in doubled strand form. (E.g. in humans $n = 23$.) Each individual carries two copies of the genome, i.e., $2 \times n$ chromosomes, half of which are inherited from each of the two parents. Genetic material is transmitted from parents to offspring through *meiosis*, genesis of germ cells. As a result of a meiosis, each germ cell contains only a *haploid* genotype, i.e., only a single copy of each chromosome. In meiosis chromosomes are susceptible to several physical processes, whose outcome is usually that descendants do not carry exactly similar genetic material as their parents. The two prominent causes of this variation are *recombination* and *mutation*. A recombination happens when the two homologous chromosomes of the parent mix, (or recombine), during the meiosis, whence the offspring will inherit a mosaic of the two parental chromosomes. Mutation refers to the processes where the content of the inherited material is different from the source chromosome either because of a point mutation (change in one nucleotide of the DNA), deletion, insertion, duplication, or translocation of the genetic material. Other known processes, such as gene conversions, are not considered in this thesis.

Let us take a closer look at how the genetic data is transmitted from a parent to a descendant. According to Mendel's first law, at any single position (*locus*) on the genome, the descendant is equally likely to inherit the material from each of the parent's two chromosomes. The updated version of Mendel's second law states that the segregations at two loci residing on different chromosomes are independent of each other. Thus, what remains to be specified, is the behavior of the loci that are located on the same chromosome.

Figure 1 gives a simplified illustration of the meiosis process for a single chromosome. First the parent's two copies of the same chromosome (A) duplicate and line up so that the homologous positions are next to each other (B). At this stage the chromosomes may physically cross over (C), cut from the crossing over positions, and recombine as four novel chromosomes that contain mixtures of segments from the two original chromosomes (D). Each of the four new chromosomes ends up in a separate gamete that may then be passed on to the next generation. If the material at two particular loci on a newly formed chromosome originate from the different source chromosomes, we say that there has been a recombination event between the loci in the meiosis. Thus, for example, the leftmost chromosome in Figure 1D is

FIGURE 1.



non-recombinant, whereas in the other three a recombination has occurred between any two loci that are colored differently. The fraction of the recombinant gametes with respect to the two loci in a large number of meioses defines a proximity measure between the loci, taking values from 0 (complete linkage) to $\frac{1}{2}$ (independent segregation). Two loci having recombination fraction less than $\frac{1}{2}$ are said to be *genetically linked*.

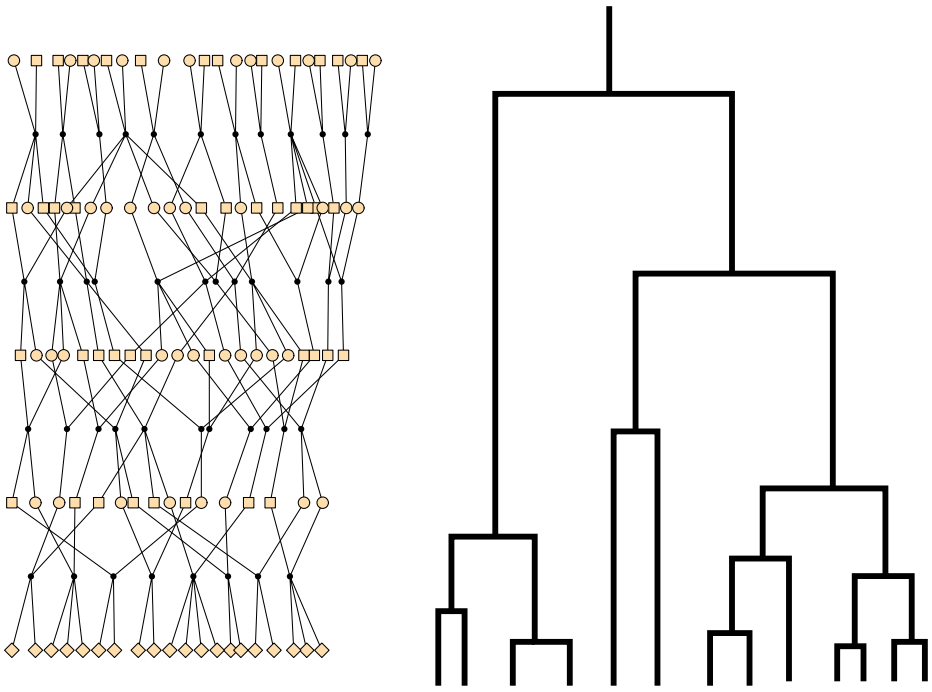
The laboratory techniques that are considered in this thesis observe genetic data at certain fixed positions on the genome (*marker loci*) that are known to be polymorphic in the population. Two types of markers considered in this work are microsatellites and single nucleotide polymorphisms (SNPs). Microsatellite loci consist of varying numbers of repeatable units of short DNA sequences and may exhibit tens of different variants, *alleles*, in a population. SNPs are usually diallelic (only two variants in the population) and are formed by a difference in a single nucleotide of the DNA.

5. Modeling genetic data with ancestry process

Population genetics studies how the genomes in the population change throughout time as a consequence of natural selection, migration, stochastic drift, mutation, recombination and other such forces. Foundations of the field were established by S. Wright (1889-1988), J. B. S. Haldane (1892-1964) and R. A. Fisher (1890-1962) during the 1920s and 1930s. For several decades the theory was developed according to the prospective framework: How does the future look like, given the current state of the population. In the 1980s the retrospective framework gained popularity, especially through the seminal works of J. F. C. Kingman, who formulated the *coalescent* process [22]. The population models in this thesis are built on the retrospective ideas. Our interest lies in the contemporary individuals whose genetic composition we model by taking into account their common past.

5.1. **Pedigrees.** When considering only a few generations backwards in time, the appropriate structure for describing relationships is a pedigree, a family tree, which

defines the parental relationships between the individuals (Figure 2a). A pedigree does not specify the exact routes of genetic material, but it constrains the possibilities. Usually pedigrees are not known reliably for many generations, excepting some bred species. And even more rare are the cases where also genotype measurements are available for ancestral individuals in pedigrees that extend several generations backwards in time. The approach taken in this thesis extends the traditional framework of pedigree analyses to settings where pedigrees may not be known a priori. Instead, they are considered as latent, unobserved variables.



(a) Pedigree. Squares are males, circles are females, the sexes are not specified at the youngest generation.

(b) Coalescent Tree. Time runs from top to bottom. Two lineages join at their most recent common ancestor.

FIGURE 2.

5.1.1. *Pedigree model.* Our pedigree model [10] considers nonoverlapping generations $t = 0, \ldots, T$, where 0 refers to the contemporary generation and T to the founder generation, i.e., to the most distant generation that is included in the model. The model assumes that a fixed number of n_0 individuals are sampled from the generation 0. The idea is to include explicitly only those individuals in the pedigree who have descendants among the sampled individuals. The pedigree is embedded in a population by specifying the number of males ($N_t^{'}$) and of females ($N_t^{''}$) in the population in each generation t, as well as two parameters α_t and β_t that govern the mating behavior in the population.

The model applies the Polya's urn scheme to the assignment of parents to children. The pedigree grows one generation at a time from the present to the founder level, and given the parameters and the number of individuals included in generation t, the family structures between generations t and $t+1$ are assumed to be independent of the structures at the younger generations. Thus the model is specified by giving the assignment probabilities of different choices of parents to individuals at a single generation t.

First, the individuals belonging to generation t choose fathers from among the N'_{t+1} males available in generation $t+1$. After the first k individuals have chosen the fathers, the next individual will choose male i as his/her father with a probability proportional to $\alpha_{t+1} + C_i^{(k)}$, where α_{t+1} is a model parameter and $C_i^{(k)}$ tells how many of the first k individuals are already assigned to male i.

After the fathers have been assigned, the choice of mothers is such that the $(k+1)^{th}$ individual in generation t will choose female j in generation $t+1$ with a probability proportional to $\beta_{t+1} + C_{i,j}^{(k)}$, where i is the (already assigned) father of this individual, $C_{i,j}^{(k)}$ is the number of common children of couple i, j among the k individuals who had already chosen their mothers, and β_{t+1} is a model parameter.

The ordering of the individuals was introduced here only to simplify the description of the model and it does not affect the probability of the pedigree.

By adjusting parameters α_t and β_t together with the population size different mating behaviors from monogamy to random mating can be modeled. It is also possible to constrain the model on the pedigrees that contain certain fixed parts, or that conform to certain rules, such as those preventing close relatives from having common offspring.

5.1.2. *Model for gene flow.* In this thesis the flow of genetic material on the pedigree is completely determined by the recombination process. Mutations are not included in the model, since the time scale of the pedigree based ancestry process is at most tens of generations. The marker map, i.e., estimates of the recombination fractions between the consecutive marker loci, is assumed to be given in advance. The genetic material of each individual is divided into paternal and maternal haplotypes and the meioses are modeled by choosing at each locus either the paternal or the maternal allele according to the origins of the neighboring loci and the known recombination fractions.

The final part of the joint probability model on the pedigree structures and gene flows is given by the probability distribution of the founder alleles. There, the simplifying assumptions of Hardy-Weinberg and linkage equilibria are made, meaning, respectively, that the two chromosomes of a founder are assumed to be independent of each other, and that a founder's alleles at different loci are assumed to be independent of each other. However, if some information on deviations from these assumptions were available, that could easily be included into the model.

5.2. **Model for phenotype.** In pedigree based *gene mapping* the idea is to follow the segregation of the marker alleles through consecutive generations and search for

positions on the genome at which the segregation of the genetic material shows similarities to the patterns observed in the distribution of the values of some phenotype. The key point is to estimate the inheritance process between the markers, which is governed by the recombination process conditioned on the marker information.

In article IV we add a model for a univariate quantitative phenotype to the model for the genealogical history. The phenotype model is based on variance components. This means that the phenotypic value is decomposed into a random number of *quantitative trait loci* (QTLs), the polygene component, and the residual error. The model specifies accurate locations of the QTLs with respect to the flanking markers. There is a natural covariance structure for genetic components arising from the pedigree and from the gene flow at each particular QTL. These covariance structures then let us estimate the relative contribution that each genetic component has on the total phenotypic variance.

The model described above extends the traditional one used in variance component linkage analysis by modeling also the unknown part of the pedigree. It has become a custom in linkage analysis first to analyze a set of small pedigrees separately and then to combine the results by assuming that the subpedigrees are independent of each other. Some work has recently been done also on the modeling of the shared ancestry of different subpedigrees [19]. The novelty in our model is that it more explicitly mimics the process as we know it: augmenting the available data by unknown parts of the pedigree as well as by an unobserved gene flow. Our approach can be expected to be powerful especially when the studied data come from a setting where the subpedigrees are indeed related to each other through common ancestors within the recent history (some ten generations or so), as is shown in article IV.

5.3. **Gene trees and recombination graphs.** When the time scale is shifted from tens to hundreds or thousands of generations, explicit pedigrees are no longer feasible descriptions of the relatedness structure of DNA segments. Instead, one may trace the ancestry of contemporary alleles backwards in continuous time ignoring the individuals. This results in a tree structure for any particular locus (Figure 2b) and in a graph structure for longer recombining segments.

The coalescent theory [22, 18, 31] is a mathematical description of the ancestry process in continuous time. To give an idea of the theory, let us first consider the coalescent process of a single locus in a Wright-Fisher population (constant size N, random mating, non-overlapping generations), with no selection. The forward-time dynamics of the model are simply described: the next generation is always formed by choosing with replacement $2N$ copies of the alleles from the preceding generation. Retrospectively the model corresponds to letting each offspring allele choose its parent among the alleles of the parents' generation.

The probability that two alleles from the current generation maintain distinct parents at least T generations backwards in time is then

$$\left(1-\frac{1}{2N}\right)^{T}=\left(1-\frac{1}{2N}\right)^{2Nt}\underset{N\to\infty}{\longrightarrow}e^{-t},$$

where $t=T/(2N)$ is the time scaled to the units of population size. Thus the coalescing time for the pair of alleles becomes an Exp(1)-distributed random variable, when the population is large and time is scaled in the units of $2N$.

These ideas can be extended to the genealogy of n recombining sequences. One starts with n separate sequences and proceeds into the past by drawing the waiting time and the identity (coalescence, mutation or recombination) of the next event from competing exponential distributions. If the event is a coalescence a randomly chosen pair of lineages unite, in case of a mutation a randomly chosen sequence experiences a mutation, and if the event turns out to be a recombination then a randomly chosen segment splits into two subsequences.

The coalescent theory provides an efficient way to simulate genealogies for n DNA segments, since it considers only that part of the history of the population that is relevant to the sample. If the same simulation task were attempted using a forward-time model, one would need to keep track of the history of the whole population, which for the usual cases where $n \ll N$ would lead to an enormous waste of resources compared to the retrospective approach.

In this thesis the continuous time coalescent theory comes into play in article V which considers an extension of the haplotyping algorithm PHASE [42]. The original PHASE algorithm models the ancestry of a population sample of contemporary haplotypes by utilizing approximations to the coalescent theory.

6. Analyzing the models

It was relatively simple to formulate the models for genetic data in the previous section. Such models can easily be used to derive null distributions of some summary statistics under different evolutionary scenarios, but direct statistical inference on the model parameters given the observed data (present day state of the process) is a very challenging task [40].

6.1. **Observed data.**

In this thesis the model for pedigrees and gene flows is analyzed conditionally on the genotype data at marker loci at the youngest generation. Such data consist of unordered pairs of alleles at each locus and, in particular, they do not include haplotype information. In addition to the marker data, we also fix the values of the population parameters, the allele frequencies, and the recombination fractions in the model. Article II considers the case where the markers are assumed to be unlinked, whereas in article III the recombination fractions between the adjacent markers are about 0.05 per meiosis.

Article IV considers a situation where, in addition to the marker data, we also have measurements of a univariate quantitative phenotype on the sample of individuals belonging to the youngest generation of the population. In the examples, the number

of markers varies between 100 and 120 and they are either all located in the same chromosome or divided into 4 distinct chromosomes. The recombination fractions between the adjacent markers in the same chromosome are about 0.04 per meiosis.

In article V the data are gathered from DNA pools, where genetic material from several (about 2-5) individuals are mixed and the pool contents are measured in a single genotyping. Thus the data contain the pool-specific frequencies of different alleles at each locus, but do not specify which pairs of the alleles belong to the same individual. The purpose of this kind of data is to reduce the overall genotyping costs of the genetic study.

6.2. **Goals of inference.** The models of ancestry are harnessed to estimate how the sampled contemporary individuals are related to each other. In that task the concept of identity-by-descent (IBD) between the observed marker alleles has a central role. Two alleles are said to be IBD if they descend from a common ancestral allele within some specified time frame. Here the natural time frame is T, the number of generations that are included in the pedigree model. However, we can also estimate more accurate IBD-probabilities with respect to any generation that is covered by the model. As the model includes the complete description of the inheritance process within the given framework, the IBD-probabilities could be estimated jointly for any group of individuals. In our examples IBD-distributions have been estimated mainly for pairs of individuals, but also for a larger group of the carriers of the same ancestral mutation (article III).

For unlinked marker data (article II) the IBD-estimates yield information on the overall relatedness between the individuals. Because the pedigree model extends over several generations, the relatedness can be studied simultaneously in different time scales. Starting from the familiar relationship categories, such as siblings and cousins, the model can also capture a weaker and more general population structure.

For linked data IBD-estimates can reveal information on varying degrees of relatedness that the same pair/group of individuals may possess at different regions of the genome (article III). Such an IBD-distribution, when augmented with some phenotype data, can serve as an input for several available gene mapping methods. We have applied this idea in article IV, where the goal is to locate such positions from the available marker map, that affect the observed phenotype values.

For linked marker data another goal is to estimate the haplotype configurations of the observed genotype data. The haplotyping problem – to resolve the unordered diploid genotype data into two haplotypes – is an important one in statistical genetics, and it has gathered interest as well in the situations where the pedigree is known as in the cases where the genotyped individuals are simply sampled from a population [30]. In article III the data come without a pedigree record, but it would be straightforward to condition the model on some fixed parts of the pedigree, if such were available. In article V haplotyping is studied in a different context: the time scale is extended by shifting from pedigrees to the continuous time coalescent theory, and thereby the length of the considered chromosomal segment is decreased.

Naturally these Bayesian models could also be utilized to study any other variable that is included in the models. As an example, one could estimate the properties of the unknown pedigree by, for instance, identifying pairs of full or half siblings from the contemporary individuals.

6.3. **Problems with simulation approach.** Given the population parameters and marker spacing, it is straightforward to simulate sample configurations from these retrospective models. For the pedigree-based model one first simulates a pedigree, then the founder alleles and finally the meioses on the pedigree, whence the genetic state is completely determined. Continuous time recombination graphs are sampled by drawing waiting times for the events (backwards in time) from competing exponential distributions corresponding to the different possible types of the events, (e.g. mutations, recombinations and coalescences). Unfortunately these simulation procedures are useless in evaluating the probabilities of different ancestral configurations *given the observed data*. This is because only a tiny fraction of all possible configurations are consistent with the observed marker data at the youngest generations, and therefore by mere simulation we would almost never reach a single one of them.

6.4. **MCMC on pedigrees and gene flows.** To overcome the computational problem related to the direct simulation, we have designed an MCMC algorithm that explores the space of the ancestral configurations that are consistent with the data. The obvious advantage over the direct simulation approach is that every state of the chain will respect the observed data. This is achieved by first generating a consistent initial state for the chain and then applying a cycle of proposal distributions that maintain the consistency between the proposed states and the data.

Several earlier applications of MCMC algorithms for analyzing genetic data on pedigrees have been published [44, 37, 38]. A general challenge for the design of the proposal distributions is posed by the structure of the space of the ancestral configurations which includes the pedigrees together with the gene flows at the marker loci. The variables are discrete and form highly dependent blocks of closely related individuals and tightly linked markers. It follows that a large number of variables need to be updated simultaneously in order to guarantee the irreducibility of the Markov chain and the mixing of the sampler. This requires computationally demanding block-updates.

Compared to the earlier works, the most notable differences in our approach are the large block-updates that consider the chromosomes of several closely related individuals simultaneously, and the ability to model also unknown parts of the pedigree. The fact that we are not working with a fixed pedigree structure also guarantees the irreducibility of our sampler, at least in the situations where the population size is large enough (proof is given below). However, when the model includes several generations and a large number of markers, the practical mixing of the sampler remains a challenging issue and an application of computational techniques like MCMCMC [12] could be a topic for further studies.

For the MCMC applications it is possible to combine several proposal distributions in order to enhance the mixing properties of the algorithm. This can be done either by defining the transition kernel as a mixture of individual transition kernels or by combining the kernels into a cycle that is run through in the same order at every iteration [32]. In our work both of these approaches are utilized.

A summary of the ideas behind our proposal distributions is given next. More details can be found in the appendices of article III and in article IV.

6.5. **Proposal distributions for the pedigree and the allelic paths.**

6.5.1. *Block Update 1: Children choosing parents.* In this update a random group of individuals is chosen from the pedigree and their parents will be resampled. Because the novel parents can be chosen also from outside of the current pedigree, the initial phase of the update extends the pedigree structure by adding to it some individuals from the population. Technically this update scheme is divided into many separate proposal distributions, one for each combination of the choices of the (ordered) group of children who are changing their parents and of the sampled pedigree structure on the individuals outside of the current pedigree. Thus this update is implemented as a mixture of separate transition kernels.

The construction of the proposal state is commenced by withdrawing such alleles from the current parents that have been transmitted to only those of their children, who are currently chosen to resample their parents. Then the children choose their new parents sequentially according to the prior distribution on the pedigree graphs and the (unlinked) transmission probabilities of alleles. Finally the phases of the new parents are sampled by a forward-backward algorithm that also accounts for the linkage between the markers.

While constructing the proposal state, we also calculate the probability of the reverse move, which is needed in defining the acceptance probability of the proposal. Because of that, the withdrawal of the alleles from the parents is done in the reverse order compared to the one by which the children choose their new parents.

6.5.2. *Block Update 2: Half siblings changing their parent.* Block Update 1 transmits the alleles from the children to their new parents sequentially and may thus be unable to create large families. This block update provides and improvement on that aspect by choosing one parent from the pedigree and by letting his/her children to resample their other parent. Again we allow the parents to be chosen from outside of the currently existing pedigree.

First the children will sequentially choose their other parent according to the pedigree prior and the unlinked allele transmission probabilities. However, all allele transmissions from the children to their parents occur jointly at the end of the update. By creating the locuswise joint distributions of the children's phases, parent's genotypes and parent's phases, we can use a forward-backward algorithm to sample a configuration that takes into account the recombinations that have taken place in the parents.

This update is computationally demanding and has been used only for up to 7 children in our examples.

6.5.3. *Block Update 3: Switching sex.* The fixed sexes of the parents restrict the assignments of the children to the parents. Sometimes this may limit the mixing of the algorithm. For instance, using only the two above mentioned block updates, it is difficult to replace two genetically similar individuals in the parents' generation by a single individual, if the original individuals happen to be of the opposite sexes. To enhance the mixing we introduce an additional updating procedure, where the sexes of the parents belonging to the same connected component are proposed to be switched. The acceptance probability of this proposal depends only on the prior probabilities of the corresponding pedigree configurations.

6.5.4. *Irreducibility of the sampler.* Theoretically the irreducibility of the sampler can be proven, for instance, by considering only Block Update 1 (BU1). To see that this results in the irreducibility of the whole chain, we may assume that the other updates do not change the configuration between successive applications of BU1.

Let us proceed by induction with respect to the number of individuals in the youngest generation. This proof assumes that in each generation there are at least n males and n females available outside of the current pedigree, where n is the number of individuals in the youngest generation.

If there is only one individual in the sample, then by applying BU1 to this individual, the pedigree and the allelic paths are sampled completely anew. Thus it is possible to reach any configuration consistent with the population size and the mating parameters (called *legal* configuration in the sequel) in just one transition. Thus the chain is irreducible.

For the induction assumption, suppose that for some $n \geq 1$ it is possible to shift between any pairs of legal configurations in at most $2n - 1$ BU1-transitions, when there are n individuals in the youngest generation, and when there are at least n males and n females available outside of the current pedigree in each generation.

Consider the case where there are $n + 1$ individuals in the sample and where there are at least $n + 1$ representatives of both sexes outside of the current pedigree in each generation. It is possible to separate any particular sampled individual i from the rest of the current pedigree in one BU1-transition in such a way that there remain at least n representatives of both sexes outside of the current pedigree in each generation. According to the induction assumption, $2n-1$ transitions are enough to make the configuration on the remaining n individuals match any legal configuration on those individuals. Finally it is possible to attach individual i to the rest of the pedigree in any possible way in a single transition. Thus $2(n + 1) - 1$ transitions are enough for updating any initial configuration to any other legal configuration, when the sample size is $n + 1$. This completes the proof by induction.

Note that in practice it is likely that there are possible transitions between different legal configurations under much weaker conditions on the number of individuals

outside of the current pedigree. Here the assumption of n males and n females was utilized in order to make the proof simple and general.

6.6. **MCMC for phenotype model.** The MCMC updates for phenotype parameters are much simpler to implement than the pedigree updates. The variance parameters related to the phenotype model are updated by multiplying with a log-normally distributed random variables. The QTL positions are updated by a (normally distributed) random walk proposal and the number of QTLs is updated with the reversible jump methodology. (Details are in article IV.)

6.7. **Haplotyping with PHASE algorithm.** PHASE is a widely-used haplotyping method for population samples [43, 41, 42]. As an input it requires the unphased diploid genotype data $(G_i)_{i\in\mathcal{I}}$, and its goal is to estimate the probability distribution of the haplotype configuration of each individual $(H_i)_{i\in\mathcal{I}}$, as well as population haplotype frequencies. PHASE can also be applied to estimate the recombination probabilities for the marker intervals.

PHASE is based on the ideas from MCMC computing and in particular it applies a Gibbs' update scheme to sample sequentially each H_i given G_i and the current state of the other haplotypes $(H_j)_{j\neq i}$. Informally, the Gibbs' sampling step favors the haplotype configurations that are either similar to those in the remaining set $(H_j)_{j\neq i}$, or that can be formed from them by mutations and recombinations [42].

Article V of this thesis extends PHASE to the setting where the individuals are not genotyped individually, but instead their genetic material is mixed into a pool of DNA and analyzed with a single genotyping. By using several such pools, the goal is to estimate the population frequencies of the haplotypes, instead of the individual haplotype configurations. This requires an additional Gibbs' update step in the MCMC algorithm. The novel step pairs by a uniformly chosen random permutation the current haplotypes within the pool, and in this way forms artificial genotypes G_i on which the original PHASE algorithm can be run. This step allows the algorithm to explore the whole space of the possible haplotype configurations given the pooled observations.

7. Results

Article I. An algorithm (APE) that extends partially known genotype data to the whole pedigree in line with the Mendelian inheritance was introduced. APE was compared with the program START [25] that tackles exactly the same problem. In the examples APE was found to outperform START, when the performance was measured in running time.

Article II. The pedigree based gene flow model was analyzed with four data sets containing unlinked microsatellite loci.

In example I the data contained individuals from nuclear families, which themselves were collected from three different populations. The relatedness estimates visualized by a dendrogram and a multidimensional scaling plot revealed both the

structures (families and populations) simultaneously. This was compared with the program STRUCTURE [34], which classifies the individuals to a fixed number of groups according to the genotype data. From STRUCTURE's results it was difficult to identify both levels of relatedness at the same time.

In examples II and III two real human data sets were analyzed. In example II the results were similar to those reported earlier by Rosenberg et al. [36]. In example III individuals from Eastern and Western Finland were analyzed, but no clear distinction between the groups was found. (Neither STRUCTURE nor previous analyses with other methods had found any geographical structure from these data.)

In example IV the relatedness estimates obtained by our method were compared with three moment-estimators [23, 26, 45] on a simulated data set. The moment estimators do not answer exactly the same question as our method, since their frame of reference for IBD-calculations cannot be specified at the level of generations. However, we were not aware of other methods that would have been more suitable for that task. As a result, our method gave smaller average errors than the moment estimators, when the reference generation was specified similarly in our reconstruction as in the original simulated data from which the true values were computed.

Article III. The pedigree based gene flow model was analyzed with two data sets containing linked microsatellite loci with recombination fractions of about 0.05 between the adjacent markers.

In example I simulated data from a 10-generation pedigree was analyzed. Our method gave good results in haplotyping when compared with the program PHASE [42], and good accuracy in IBD-estimation when compared with the three moment-estimators [23, 26, 45]. Note again that these moment-estimators were unable to take into account the linkage and the exact reference generation for the IBD-computations. The advantage of the linkage model on linked data over the unlinked version of the algorithm was also confirmed.

In example II the simulated data set from a 20-generation pedigree were analyzed. All carriers of one particular founder allele were sampled from the current generation and their joint ancestry was estimated. The highest IBD-sharing among the sampled individuals was correctly estimated to be near the trait locus. No similar phenomenon was observed when the similarity between the alleles was defined by the states of the alleles. However, no clear sign of the excess haplotype sharing was observed near the trait locus, even though in the original simulation such excess sharing was present. A reason for this may be that the algorithm was started with $T = 9$ generations right from the beginning. In article IV we found that it is better to create an initial state for the algorithm by a sequential approach, where one generation at a time is added to the model, always conditionally on the already existing generations of the pedigree.

Article IV. The pedigree and gene flow model was extended to include also a model for a quantitative phenotype.

The model was tested on two simulated data sets. In example I the children of 50 three-child nuclear families formed the sample in the youngest generation of the population. The marker data were simulated on 4 chromosomes, each containing 30 markers, and spaced in such a way that the recombination fractions between the adjacent markers were 0.04 per meiosis. Furthermore, a univariate phenotype was simulated by an additive model which included two QTLs and a residual effect. The results were compared to a variance component linkage analysis program SOLAR [2] and to an association analysis program TASSEL [7]. Our method and SOLAR produced qualitatively similar results that correctly indicated the locations of the two QTLs, whereas the association analysis was not able to separate the true signals from false positives.

In example II a more challenging situation was considered. This time a pedigree was simulated for 50 two-child nuclear families and the variances of the two QTLs relative to the total phenotypic variance were decreased compared to example I. The population parameters were such that the pedigree of the sampled individuals had experienced a bottleneck in the recent history (within 5 of the most recent generations). In the results it was clearly seen how the QTL signals were captured when the model included 4 ancestral generations, while 3 or less ancestral generations were not enough to produce accurate estimates. Also comparisons with SOLAR and TASSEL suggested that if the recent ancestry could not be taken into account for more than one generation backwards in time, it is difficult to capture the true QTL signals.

Article V. An extension of the PHASE algorithm [42] for pooled genotype data was compared with a deterministic greedy algorithm and a previously available program LDPooled [21]. Both simulated data and real human data extracted from the HapMap database were used. In simulated examples and in the majority of the real data sets the proposed method outperformed two others in the accuracy of the frequency estimates.

It was also shown that pooling DNA from 2-3 individuals before genotyping the samples may be advantageous when estimating the population haplotype frequencies with fixed number of genotypings. Thus, even though pooling results in some loss of the haplotype information, that loss can be well compensated by the increase in the overall sample size provided by pooling.

8. Conclusion

The main goal of this thesis has been an accurate, application-driven modeling of some population genetic phenomena, and the design of computational methods by which the models can be analyzed. This approach has required considerable efforts, especially in the implementation issues, but they were considered worth taking so that most information from the data could be extracted. The results summarized above show that a more accurate modeling of the genetic processes has indeed proven advantageous in several settings.

In the field of genetics the development of the laboratory technologies is accelerating in an enormous pace. For instance the 1000 Genomes Project launched in January 2008 aims to sequence the genomes of at least a thousand people from around the world. During its three-year course the 1000 Genomes Project will generate 60-fold more sequence data than have been deposited into public DNA databases over the past 25 years [1]. Such a project would have been impossible just a few years ago.

It may thus be that in the near future the focus of (human) genetics will shift from marker data towards sequence data, not only in the huge international projects, but also in the smaller scale studies of individual research groups around the world. It is not computationally feasible to fully exploit such data by the methods introduced in this thesis, and there is an evident need for computationally new approaches to handle the next generation of genomic data. At the same time, however, studies on wild animal and plant populations continue to be carried out with only a dozen microsatellite loci, simply because the resources for sequencing such species have not been available. Furthermore, pedigree and relatedness estimation from marker data is a timely topic in those fields of research [8, 33]. Naturally the methods developed in this thesis will remain readily applicable to those settings.

There are several topics for further study where the pedigree and gene flow estimation algorithm presented here can be utilized. For example, one may fix some parts of the pedigree and consider the model as a way to build bridges between the known pieces of the pedigree. In the special case where the whole pedigree is considered known, our method becomes comparable to some other MCMC methods like SimWalk2 [38] (IBD-estimation) and Loki [17] (QTL mapping). Since our MCMC updating scheme is different from the other available methods, comparisons in such settings would be of interest. On the other hand, when the pedigree structure is fixed, the questions of possible reducibility of the MCMC samplers must be considered carefully [37].

Another question of interest would be the reconstruction of the families from linked SNP data. As SNPs usually possess only two alleles, they are not very informative about the family structure unless the linkage can be taken into account or unless there are very many SNPs available. The method introduced in this thesis takes linkage into account and might thus be advantageous in certain settings compared to the approaches that assume unlinked markers (e.g. COLONY2 [46]). On the other hand, if/when hundreds of thousands of SNPs become routinely available also with other species than human, then the recent history between the individuals may be accurately revealed already by more straightforward methods than the ones presented in this thesis.

Also the limits of the method are of interest, both in computational and theoretical terms. The computational limits may be extended by switching to a tempering version of the algorithm, where parallel chains are run on different processors, and the chains are allowed to communicate and switch states at certain time points. This idea called Metropolis-coupled Markov chain Monte Carlo (MCMCMC) is expressed

by Geyer [12]. Interesting theoretical questions relate to the time scale within which the reconstruction of the pedigrees is a reasonable task, given the amount of marker data at hand.

Outside of this thesis we have extended the methods for haplotyping pooled genetic data to situations, where some prior information about the haplotypes in the population is available, for example, from a database such as HapMap. An interesting task would be to combine such prior knowledge with a realistic model for the population haplotype distribution (e.g. [24]).

Under the massive flood of genetic data, the data-specific models and software are deemed to have a short lifespan. There are, however, the principles of good modeling that are forever. As both theoretical and experimental knowledge on the genetic processes keep accumulating, there remains an important role in genetics for Bayesian modeling as a coherent and consistent way to combine the already known to newly observed [5].

Acknowledgments

I would like to thank my principal supervisor Elja Arjas for giving me an opportunity to work with statistical genetics even though I did not have any prior knowledge on the subject. Working under Elja's guidance has been a very pleasant and instructive experience, as I have had much freedom to study those issues that I have found interesting, while at the same time I have had a possibility to discuss with Elja virtually any time. It is an honor to be included in the long and impressive list of Elja's former PhD students.

The completion of this thesis owes a lot to my senior colleague and assistant supervisor Dario Gasbarra, who has been very patient and willing to help both with the program codes and with the theoretical questions. Dario's knowledge on stochastics and sampling methods like MCMC has been an endless resource of ideas for our joint work.

The third central figure during my PhD studies has been my assistant supervisor Mikko J Sillanpää, whom I thank especially for many discussions about statistical genetics. Mikko has provided me with a great number of references and a huge amount of experience on the subjects that we have been working on in these articles.

I also thank my fellow PhD students at the Biometry group for a nice atmosphere: Jukka K, Pekka M, Jukka S, Jing, Crispin, Rashi, Sarish and Rossana. It has been vital change for the every day work to have some nice trips with you every now and then: Oulanka, Warwick, Benidorm, Bergen, Tübingen, Bielefeld and Hamilton Island are the first in my mind.

Also our Biometry seminar has provided me with many interesting ideas and pieces of information. So in addition to the previously mentioned seminar people I also want to thank Siru and Anders. I also thank Sangita Kulathinal for our joint work on the haplotyping problem on pooled data and Elina Salmela for collaboration in article II.

My PhD studies have been funded by the Academy of Finland (Centre of Excellence in Population Genetic Analyses) 2004-2006 and by the ComBi graduate school 2005-2009.

Lastly I am thankful for my parents Ulla and Mikko and brother Antti who have always let me go on my own way and choose the topics I am interested in, providing me with continuous support in all possible ways that one can imagine.

References

[1] 1000 Genomes Project. International consortium announces the 1000 Genomes Project. *www.1000genomes.org*, Jan 22, 2008.

[2] L. Almasy and J. Blangero. Multipoint quantitative trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, 62:1198–1211, 1998.

[3] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

[4] T. Bayes. Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

[5] M. A. Beaumont and B. Rannala. The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5:251–261, 2004.

[6] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, 1994.

[7] P. J. Bradbury, Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23:2633–2635, 2007.

[8] F. D. Frentiu, S. M. Clegg, J. Chittock, T. Burke, M. W. Blows, and I. P. F. Owens. Pedigree-free animal models: the relatedness matrix reloaded. *Proceedings of the Royal Society B*, 275:639–647, 2008.

[9] D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, second edition, 2006.

[10] D. Gasbarra, M. J. Sillanpää, and E. Arjas. Backward simulation of ancestors of sampled individuals. *Theoretical Population Biology*, 67:75–83, 2005.

[11] A. Gelman. Objections to Bayesian statistics. *Bayesian Analysis*, 3:445–450, 2008.

[12] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. 1991.

[13] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[14] P. J. Green. Transdimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*. Oxford University Press, 2003.

[15] P. Guttorp. *Stochastic Modeling of Scientific Data*. Chapman & Hall/CRC, 1995.

[16] W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[17] S. C. Heath. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, 61:748–760, 1997.

[18] J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, 2005.

[19] O. Hössjer. Modeling the effect of inbreeding among founders in linkage analysis. *Theoretical Population Biology*, 70:146–163, 2006.

[20] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861, 2007.

[21] T. Ito, S. Chiku, E. Inoue, M. Tomita, T. Morisaki, H. Morisaki, and N. Kamatani. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype

copies in each pool by use of pooled DNA data. *American Journal of Human Genetics*, 72:384–398, 2003.

[22] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[23] C. C. Li, D. E. Weeks, and A. Chakravarti. Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity*, 43:45–52, 1993.

[24] N. Li and M. Stephens. Modeling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics*, 165:2213–2233, 2003.

[25] Y. Luo and S. Lin. Finding starting points for Markov chain Monte Carlo analysis of genetic data from large and complex pedigrees. *Genetic Epidemiology*, 25:14–24, 2003.

[26] M. Lynch and K. Ritland. Estimation of pairwise relatedness with molecular markers. *Genetics*, 152:1753–1766, 1999.

[27] G. J. Mendel. Versuche über Plflanzenhybriden. *Verhandlungen des Naturforschenden Vereines in Brünn*, Bd. IV für das Jahr, 1865 Abhandlungen:3–47, 1866.

[28] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. N. Teller, and E. Teller. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21:1097–1091, 1953.

[29] S. P. Meyn and R. L. Tweedie. *Markov Chains for Stochastic Stability*. Springer-Verlag, 1993.

[30] T. Niu. Algorithms for inferring haplotypes. *Genetic Epidemiology*, 27:334–347, 2004.

[31] M. Nordborg. Coalescent theory. In B. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 179–212. John Wiley & Sons, 2001.

[32] A. O'Hagan and J. J. Forster. *Vol. 2B: Bayesian Inference*. Kendall's Advanced Theory of Statistics. Arnold, second edition, 2004.

[33] J. M. Pemberton. Wild pedigrees: the way forward. *Proceedings of the Royal Society B*, 275:613–621, 2008.

[34] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[35] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 1999.

[36] N. Rosenberg, E. Woolf, J. Pritchard, T. Schaap, D. Gefel, I. Shpirer, U. Lavi, B. Bonné-Tamir, J. Hillel, and M. W. Feldman. Distinctive genetic signatures in the Libyan Jews. *Proceedings of the National Academy of Sciences USA*, 98:858–863, 2001.

[37] N. A. Sheehan. On the application of Markov chain Monte Carlo methods to genetic analyses on complex pedigrees. *International Statistical Review*, 68:83–108, 2000.

[38] E. Sobel and K. Lange. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, 58:1323–1337, 1996.

[39] D. Sorensen and D. Gianola. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag, 2002.

[40] M. Stephens. Inference under the coalescent. In B. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 213–238. John Wiley & Sons, 2001.

[41] M. Stephens and P. Donnelly. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, 73:1162–1169, 2003.

[42] M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, 76:449–462, 2005.

[43] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989, 2001.

[44] E. A. Thompson. *Statistical Inference from Genetic Data on Pedigrees*. Institute of Mathematical Statistics, 2000.

[45] J. Wang. An estimator for pairwise relatedness using molecular markers. *Genetics*, 160:1203–1215, 2002.

[46] J. Wang and A. W. Santure. Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, 181:1579–1594, 2009.