BAYESIAN ANALYSIS OF MIXTURE MODELS WITH APPLICATION TO GENETIC LINKAGE

FANG FANG

Bayesian Analysis of Mixture Models with Application to Genetic Linkage

by

© Fang Fang

B. Science (Memorial University)

A practicum submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Master of Applied Statistics

Department of Mathematics and Statistics

Memorial University

July 12, 2010

St. John's Newfoundland and Labrador

Abstract

Through an application to genetic linkage analysis, this project describes how the Bayesian approach can be used for the mixture model with an unknown number of components. Genetic linkage analysis based on a complex model can be difficult to manage when a large number of markers loci and/or large pedigrees are involved, due to computation limitations. However, Markov chain Monte Carlo (MCMC) schemes are one alternative, utilizing a reversible jump steps that allow change on the dimension of parameter space. Thus, the MCMC samplers with a different numbers of quantitative trait loci based on complex large pedigrees can be obtained using reversible jump MCMC methodology. The application of the MCMC scheme is illustrated with a case study of genetic linkage to hypercalciuria. This analysis report found strong evidence for linkage of hypercalciuria to calibrated estimates of Bayes factors, the so-called L-Scores. To my knowledge this is the first time that urinary calcium excretion has been clearly linked to a narrow region of the genome. Nevertheless, further study is needed to confirm this finding.

Acknowledgements

I am heartily thankful to Dr. J. C. Loredo-Osti, my supervisor, for his guidance from the initial to final level through all the stages of my Master program. Many thanks to Dr. Wang Hong, my co-supervisor, for his encouragement during both my undergraduate and graduate studies at Memorial University. This report would not have been completed without their support.

It is a pleasure to thank Chen Min and Lester Marshall, who made this thesis possible, for their great help and understanding during the completion of this report.

A special thank you from the bottom of my heart to Dr. Chu-In Charles Lee for developing my interest of Statistics. His encouragement gave me a great sense of uplift.

The Department of Mathematics and Statistics has provided the support I needed to produce and complete my report. I would like to thank Jaide Eustace, and all of the people working in the department.

I appreciate the financial support from the Graduate School of Memorial University as well as CIHR and MITACS through grants to my supervisor. I would also like to thank Professors Alain Bonnardeaux, Kenneth Morgan and Mary Fujiwara for providing and assisting with the French-Canadian family data set.

Finally, I owe my sincere gratitude to my parents for their deep love and endless patience.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statistical analysis methods have been widely employed in genetic linkage studies during the past decade. Genetic linkage refers to the tendency for loci located closely on the same chromosome simultaneously are transmitted to an offspring. The purpose of genetic linkage analysis is to locate disease genes through the modeling of the joint segregation of putative disease loci and genetic markers on each chromosome. Detecting and analyzing the genetic linkage for quantitative trait loci (QTL) is one of the essential tasks in genetic analysis since many diseases can be consider of as continuous traits.

Fundamental to linkage analysis is the computation of the likelihood of the segregation of disease and markers within families. The computation of this likelihood can be very involved and some special algorithms have been developed to perform this task (Elston and Stewart 1971; Cannings et al. 1978; Lander and Green 1987; Ott 1991). For large pedigrees, the Elston-Stewart (1971) algorithm is used to cal-

culate exact likelihoods based on the peeling technique. However, this algorithm only works with a small number of loci because the computing time increases when more markers are involved (Heath 1997; Uimari and Sillanpaa 2001). In contrast, the Lander-Green algorithm is efficient for multipoint linkage analysis with many markers but small pedigrees (Kruglyak et al. 1995; Kruglyak and Lander 1998). Using the Markov chain Monte Carlo (MCMC) method, the likelihood of large pedigree data with a large number of loci can be obtained from the haplotype probabilities and preassigned penetrance (Sobel and Lange 1996; Uimari and Sillanpaa 2001).

One feature that makes the Bayesian methods attractive is that all inference is drawn through the likelihood. The Bayesian methods are suitable for problems that may be analytically intractable because of limitations for computing high-dimension integrals. Many of these methods are based on the Markov chain Monte Carlo (MCMC) method (Metropolis et al. 1953; Hastings 1970; Geman and Geman 1984; Gelfand and Smith 1990) methods, which are stochastic simulation techniques developed in the 20th century to solve this kind of computing problem. In this report, the Bayesian approach, including Bayes factors, is applied to study genetic linkage.

1.1 Some background of the Bayesian approach

The basic idea of the Bayesian approach is to estimate parameters through the Bayes formula,

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta},$$

where the information of both prior distribution with density function $\pi(y)$ and posterior distribution with density $\pi(\theta|y)$ are derived from the joint distribution $f(y|\theta)\pi(\theta)$. However, the computation on the marginal function, $\int f(y|\theta)\pi(\theta)d\theta$, can be very difficult when working with a complex model. As computing power has increased, Bayesian methods have become very popular. Recently, the Bayesian approach has been widely applied as a tool of statistical analysis to numerous fields, such as health science, social sciences, econometrics and physical sciences. There is also a rising interest in the use of the Bayesian method in genetic studies because some genetic problems are built on complex models that cannot be dealt with in a classical setting (Ott 1991; Stephens and Smith 1993).

Bayes factor can be used to test statistical hypotheses. Assume that the data have arisen under one of two mutually exclusive hypotheses H_1 and H_2 according to a probability density $P_r(\boldsymbol{Y}|H_1)$ or $P_r(\boldsymbol{Y}|H_2)$. Given a prior probability $P_r(H_1)$ and $P_r(H_2) = 1 - P_r(H_1)$, the data produces posterior probabilities, $P_r(H_1|\boldsymbol{Y})$ and $P_r(H_2|\boldsymbol{Y})$. Because prior information is transformed into posterior information through consideration of the data, this transformation represents the evidence

provided by the data. Once the evidence is connected to the odds scale, the transformation takes a very simple form. From the Bayes Theorem, we have

$$P(H_m|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|H_m)P(H_m)}{P(\boldsymbol{Y})},$$

where $m = 1, 2$. Thus, we have

$$\frac{p(H_1|\boldsymbol{Y})}{p(H_2|\boldsymbol{Y})} = \frac{p(\boldsymbol{Y}|H_1)}{p(\boldsymbol{Y}|H_2)}\frac{p(H_1)}{p(H_2)},$$

and the transformation is simply the multiplication

$$\text{posterior odds} = \frac{p(\boldsymbol{Y}|H_1)}{p(\boldsymbol{Y}|H_2)} \cdot \text{prior odds ratio.}$$

The ratio,

$$\text{BF} = \frac{p(\boldsymbol{Y}|H_1)}{p(\boldsymbol{Y}|H_2)},$$

is defined as Bayes factor. The Bayes factor is the ratio of the posterior odds of H_1 to its prior odds (Kass and Raferty 1995).

In the simplest case, when the two hypotheses are single distributed with no free parameters, Bayes factor is the likelihood ratio. In other cases, when there are continuous parameters under either hypothesis, the densities $P(\boldsymbol{Y}|H_m)$, $m = 1, 2$, are obtained by integrations over the parameter space. This is

$$P(\boldsymbol{Y}|H_m) = \int P_r(\boldsymbol{Y}|H_m, \boldsymbol{\theta}_m)\pi(\boldsymbol{\theta}_m|H_m)d\boldsymbol{\theta}_m$$

where $\boldsymbol{\theta}_m$ is the parameter under H_m, $\pi(\boldsymbol{\theta}_m|H_m)$, the prior density, and $P_r(\boldsymbol{Y}|H_m, \boldsymbol{\theta}_m)$, the conditional density of $\boldsymbol{Y}$ when $\boldsymbol{\theta}_m$ is given.

1.2 Genetic linkage analysis

The traditional maximum likelihood calculation algorithms including the Elston-Stewart "peeling" algorithm and Lander-Green algorithm are employed in the genetic mapping of complex pedigrees (Elston and Stewart 1971; Cannings et al. 1978; Lander and Green 1987). However, the linkage analysis is difficult to perform using these methods when we have a large number of marker loci and/or large pedigrees. The computation takes a long time to evaluate the likelihoods and there is also potential loss of accuracy due to computer rounding. As an alternative, Markov chain Monte Carlo (MCMC) schemes can be implemented to estimate parameters in the mixture model for large and/or complex problems in genetic linkage (Satagopan et al. 1996). Furthermore, the traditional likelihood methods have limitations for such complex models because of the many parameters involved. In order to obtain an estimate of the parameters, the likelihood has to be maximized over the whole space of parameters. The traditional MCMC sampling scheme applies to the situation in which the parameter space has a fixed dimension. Guo and Thompson (1992) showed how Monte Carlo estimates of likelihoods can be obtained using the MCMC algorithms. Guo (1991) investigated the use of the Gibbs sampler to study quantitative traits in applications such that the space of parameters is fixed and known. When the dimension of the parameter space (k) is a random variable, the traditional MCMC techniques cannot be used without modification. Green (1995) developed the

method, now called the reversible jump MCMC sampler, which allows changing of the dimension of the parameter space. Following this, Richardson and Green (1997) implemented a reversible jump MCMC scheme for the normal mixture model with an unknown number of components. This scheme has been applied to many genetic problems such as genetic segregation and linkage analysis (Heath 1997), construction of genetic linkage maps (Jansen et al. 2001), and hypothesis testing for the existence of genetic linkage. Therefore, MCMC methodology with a reversible jump step has proven to be useful when generating the MCMC sampler with different numbers of quantitative trait loci for a large and complex pedigrees.

In this report, Chapter 2 describes the normal mixture model with an unknown number of components under a Bayesian framework. Then, the method of constructing a Markov chain with the stationary distribution using the reversible jump MCMC algorithm is presented in Chapter 3. In the last chapter, a case study of genetic linkage for hypercalciuria, a condition characterized by a high level of urinary calcium excretion, was conducted. This linkage study was conducted using the Loki 2.4.6, a program developed by Heath (2003). The result of this case study suggested that the genetic linkage underlying gene(s) regulating calcium excretion are possibly located on chromosome 15.

Chapter 2

Mixtures and modeling of quantitative trait loci (QTL)

It is a common practice to use the normal distribution model because of the acceptance of the normal distribution as the "natural" distribution of the errors as well as its connection to the central limit theorem. In particular, the normal mixture model is widely used in statistical literature because it allows for the parametric description of distributions that can not be achieved with conventional probability density functions. Furthermore, Mendelian genetic effects are naturally modeled as mixtures and, when the traits under scrutinizing are quantitative. This has been facilitated by the rapid development of Bayesian inference in conjunction with the reversible jump MCMC methods that include the Hasting-Metropolis algorithm and Gibbs sampler.

In this chapter, we will first introduce the basic Bayesian approach to a mixture from the normal distribution with an unknown number of components, and then

discuss the selection of a prior distributions for the parameters. Following that, the mixed effects model for a quantitative trait will be set up. Also, we will define the distribution types of all the variables, which include number of QTLs, allele frequencies for QTLs and markers, along with the random effects on each QTL in the mixed model.

2.1 The Bayesian analysis for the normal mixture model with an unknown number of components

In this report, let k denote the unknown number of components in the normal mixture model, which will be used to model the number of QTLs when the normal mixture model is applied to genetic linkage analysis. Moreover, the unknown number k is assumed to be larger than the unit ($k > 1$) in the mixture model, since the model would not include any random components if k is less than 1, and the data is a random sample from a univariate normal distribution if $k = 1$ in the mixture model.

2.1.1 Modeling for normal mixture distribution with complete data (y, z)

Suppose $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ are the observed random variables which are independently and identically distributed with given parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma^2}$. Given these parameters, the conditional joint distribution density function is expressed in the form

$$p(\boldsymbol{y}|\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \prod_{j=1}^{n} p(y_j|\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}). \tag{2.1}$$

The conditional density of y_j for any j $(j = 1, \cdots, n)$, $p(y_j|\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$ is given by the mixture of k components expressed as

$$p(y_j|\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \sum_{i=1}^{k} \pi_i \cdot N(y_j|\ \mu_i, \sigma_i^2). \tag{2.2}$$

The elements of the vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_k)$ are the mixing proportions having the following two properties:

1. $0 \leq \pi_i \leq 1$ $(i = 1, 2, \cdots, k)$,

2. $\pi_1 + \pi_2 + \cdots + \pi_k = 1$.

Notice that given $\boldsymbol{\pi} \in \mathbb{R}^k$, it implies that k is also specified. Furthermore, $N(y_j|\ \mu_i, \sigma_i^2)$ is the well-known density function of an observed normally distributed random variable y_j. We assume $\boldsymbol{\theta}$ is the parameter vector of the normal densities in equation (2.2), where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$.

For any given $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, each element in $\boldsymbol{y}$ is treated as a random sample from the set of independent distributions, $\{N(\mu_i, \sigma_i^2)\}$, with corresponding drawing probabilities $\{\pi_i\}$, $i = 1, 2, \cdots, k$. Furthermore, if we know the membership of the population where the y_js was sampled from, then we have $(y_1, z_1), \cdots, (y_n, z_n)$, where z_j $(j = 1, 2, \cdots, n)$ will hold the membership information. Since we cannot observe $z_1,\ z_2, \cdots, z_n$, we can treat them as missing variables. Thus, given $\boldsymbol{\pi}$, $\boldsymbol{z} = (z_1,\ z_2, \cdots, z_n)$ would also be a vector of independently and identically distributed random variables having probability mass function

$$Pr(z_j = i|\ \boldsymbol{\pi}) = \pi_i. \tag{2.3}$$

Therefore, the conditional probability density of the independent variable $\boldsymbol{y} = (\ y_1,\ y_2,\ \cdots, y_n)$, given the values of the z_j, is defined as

$$p(y_j|\ z_j = i, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = N(y_j|\ \mu_i, \sigma_i^2) \qquad (j = 1, 2, \ldots, n). \tag{2.4}$$

From equations (2.3) and (2.4), the density function for the mixture model is shown in equation (2.2) could also be expressed as

$$\begin{aligned} p(y_j|\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) &= \sum_{i=1}^{k} \pi_i \cdot N\ (y_j|\ \mu_i, \sigma_i^2) \\ &= \sum_{i=1}^{k} Pr\ (z_j = i\ |\ \boldsymbol{\pi}) \cdot p\ (y_j|\ z_j = i, \boldsymbol{\mu}, \boldsymbol{\sigma^2}). \end{aligned}$$

We could also obtain the following equation from the conditional distribution function:

$$\begin{aligned} p(y_j|\ z_j = i, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) &= \frac{Pr(y_j,\ z_j = i, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})}{p(z_j = i|\ \boldsymbol{\pi})} \\ &= \frac{Pr(z_j = i|\ y_j, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) \cdot p(y_j|\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})}{p(z_j = i|\ \boldsymbol{\pi})}. \end{aligned}$$

Therefore, substituting both equations (2.3) and (2.4) into the above function, we have

$$N(y_j|\ \mu_i, \sigma_i^2) = \frac{Pr(z_j = i|\ y_j, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) \cdot p\ (y_j|\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})}{\pi_i}. \tag{2.5}$$

Given $\boldsymbol{y}$, the conditional probability function for the missing data has the form

$$\begin{aligned} Pr(z_j = i|\ y_j, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) &= \frac{\pi_i \cdot N(y_j|\ \mu_i, \sigma_i^2)}{p\ (y_j|\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})} \\ &= \frac{\pi_i \cdot N(y_j|\ \mu_i, \sigma_i^2)}{\sum_{l=1}^{k} \pi_l \cdot N(y_j|\ \mu_l, \sigma_l^2)}. \end{aligned}$$

The above equation shows that

$$Pr\ (z_j = i|\ y_j; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) \propto \pi_i \cdot N\ (y_j|\ \mu_i, \sigma_i^2).$$

2.1.2 The Bayesian approach

In genetic studies, maximum likelihood estimation (MLE) is the most common method used to obtain the estimators for parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$. When k is assumed to be known, the likelihood function is given by equation (2.2)

$$\begin{aligned} L(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) &= \prod_{j=1}^{n} p(y_j|\ \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) \\ &= \prod_{j=1}^{n} \sum_{i=1}^{k} \pi_i \cdot N\ (y_j|\ \mu_i, \sigma_i^2), \end{aligned}$$

with π, μ, and σ^2 assumed to be unknown constants, i.e. assumed to be non-random. However, multiple local maxima may exist due to multi-modal aspects of a mixture distribution. Therefore, choosing between several different local maxima can be a non-straightforward issue in the MLE method. Moreover, the high dimensionality leads to other computational challenges.

In contrast, the Bayesian approach, which treats the parameters as random variables, avoids the issue described for the MLE method by integrating over the whole space of parameters. From the Bayesian formula, we have

$$p(\pi, \mu, \sigma^2 | \ y) = \ p(y | \ \pi, \mu, \sigma^2) \cdot p(\ \pi, \mu, \sigma^2)/p(y),$$

where $p(\pi, \mu, \sigma^2)$ and $p(\pi, \mu, \sigma^2 | \ y)$ are known as the *prior* and *posterior* distribution functions of the parameters in the Bayesian approach. Sometimes, $p(y)$ can be ignored because it can be treated as a constant in this formula. The essential idea of the Bayesian approach is to obtain the posterior distribution through the prior distribution, based on the Bayesian formula. The computation challenge still exists, even though more powerful computational methods have been developed to solve the issue.

Diebolt and Robert (1994) discussed the method of applying the Markov Chain Monte Carlo and Gibbs sampler in order to estimate the parameters of the model when the number of components is known. Based on the Dirichlet process, Escobar and West (1995) have demonstrated this process using examples with an unknown

number of components. Moreover, Richardson and Green (1997) solved the problem with an unknown number of components for the normal mixture distribution using the reversible jump (Green 1995) method. The details will be discussed in Chapter 3.

2.1.3 Choosing the prior distribution and hyperparameters

Using the Bayesian approach, both the observed data $\boldsymbol{y}$ and the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are treated as random variables. The distributions of parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are known as prior distributions, and the parameters of those distributions are the hyperparameters. The choice of the prior distribution is not free of controversy, although the Bayesian methodology has been well-developed through extensive research by many statisticians. Two of the more popular choices are the conjugate prior family and the closed by sampling family (Heath 1997; Richardson and Green 1997; Robert and Casella 1999). In this report, we use the prior distributions suggested by Richardson and Green (1997):

$$\begin{aligned}
\boldsymbol{\pi}|k &\sim \mathrm{D}_i(\delta,\ \delta, \cdots,\ \delta) \\
\mu_i &\sim \mathrm{N}(\xi,\ \kappa^{-1}) \\
\sigma_i^{-2}|\omega &\sim \Gamma(\varsigma,\ \omega) \\
\omega &\sim \Gamma(g,\ h) \\
k &\sim \mathrm{U}[1, k_{max}]
\end{aligned}$$

Quantities (δ, ξ, κ, ς, g, h, ω) are the hyperparameters. Parameter ω follows a Gamma distribution with parameters g and h. Notice that $\Gamma(g,\ h)$ is a hyperprior distribution. In addition, the mixing proportions, $\boldsymbol{\pi} = (\pi_1,\ \pi_2, \cdots,\ \pi_k)$, are assigned to be the symmetric Dirichlet distribution with parameter $\delta\mathbb{1}$. The joint prior distribution of the parameters is given by

$$p(\boldsymbol{\theta}') = p(k)\ p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}|k),$$

where $\boldsymbol{\theta}' = (k, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$ are the parameters of the univariate normal mixture distribution. The estimated number of components (k) depends on a marginal posterior distribution (Green 1995; Stephens 2000). Assuming $p(\theta'|\boldsymbol{y})$ is the stationary distribution and $(\theta')^{(1)}, (\theta')^{(2)}, \cdots, (\theta')^{(N)}$ are the realizations of the Markov chain, then k can be estimated according to the average:

$$\begin{aligned} Pr\ (k = i|\ \boldsymbol{y}) \quad &= E(\ I(k = i)|\ \boldsymbol{y}) \\ &\approx \frac{1}{N}\ \sum_{t=1}^{N} I(k^{(t)} = i) \\ &= \frac{1}{N}\ \sharp\{t : k^{(t)} = i\}. \end{aligned}$$

2.2 Modeling of quantitative trait loci

In genetic linkage analysis, the modeling of a quantitative trait is more difficult when the exact number of genes which control the trait is unknown. Two different cases are considered under different assumptions. The first case is under the assump-

tion that the quantitative trait is controlled by an infinite number of genes. In such a case, ordinary linear model techniques can be used to carry out the inference. There is no single gene to be mapped according to its infinitesimal nature. The second case arises when few genes, each having a large effect, are assumed to be responsible for the trait. Not knowing the exact number of genes involved implies an unknown number of terms (k) in the mixture model. This issue is addressed in the context of the Monte Carlo Markov Chain technique with reversible jumping (Green 1995; Green and Richardson 1997; Heath 1997). In this report we are fundamentally interested in the second case, the oligogenic model, under the following assumptions:

1. There are no interactions between the QTLs and the environmental convariates nor between any QTLs.

2. The information about quantitative traits, covariates, and marker data are the observations included in the data set $\boldsymbol{Y}$.

3. The marker positions are known and marker data are correct.

4. There is an equal prior probability for each QTL, and equal probability for each QTL on one chromosome.

5. The map distances are the same for both males and females.

2.2.1 Mixed effect model for quantitative trait

The mixed model is a valuable statistical tool which can be applied to genetic linkage analysis. For the pedigree data, the mixed effect model for the quantitative trait y is defined in the form of

$$\boldsymbol{y} = \mu\mathbb{1} + \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{k_q} \boldsymbol{Q_i}\boldsymbol{\alpha_i} + \boldsymbol{e}, \tag{2.6}$$

where $\mu\mathbb{1}$ is the overall mean; $\boldsymbol{X}$ is an $(n \times m)$ incidence matrix for covariates; $\boldsymbol{\beta}$ is an $(m \times 1)$ vector of covariate effects; k_q is the number of diallelic QTLs (of course, $k_q = k - 1$, where k is the number of components in the mixture); $\boldsymbol{Q_i}$ is an $(n \times 2)$ incidence matrix that denotes the effect of the i^{th} QTL, and $\boldsymbol{\alpha_i}$ is a (2×1) vector of random effects of such a QTL. At the i^{th} position, G_i and the QTL genotypes A_1A_1, A_1A_2, and A_2A_2 have corresponding effects a_i, d_i, and $-a_i$. Thus, if the vector of random effects is expressed as

$$\boldsymbol{\alpha_i} = \begin{pmatrix} a_i \\ d_i \end{pmatrix},$$

then the j^{th} row $(j = 1, 2, \cdots, n)$ of the incidence matrix $\boldsymbol{Q_i}$ for the i^{th} QTL effects will be one of

$$\begin{pmatrix} 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \end{pmatrix} \text{ or } \begin{pmatrix} -1 & 0 \end{pmatrix},$$

corresponding to the specific genotypes, $(A_1A_1, A_1A_2$, or A_2A_2, respectively) at the i^{th} location for the j^{th} individual.

Finally, the error term $\boldsymbol{e}$ that indicates the residual effect in model (2.6) is a vector of dimension n that follows a normal distribution, with zero mean and diagonal variance matrix.

2.2.2 Joint distribution and prior distribution

Heath (1997) indicated that the joint distribution for a large pedigree data set can be expressed as

$$p(k,\ \boldsymbol{G},\ \boldsymbol{M},\ \boldsymbol{\beta},\ \boldsymbol{\lambda},\ \boldsymbol{\varphi},\ \boldsymbol{\eta},\ \boldsymbol{\alpha},\ \boldsymbol{\sigma_e^2},\ \boldsymbol{\mu},\ \boldsymbol{Y}).$$

The complete genotype of all QTLs is denoted as $\boldsymbol{G}$; $\boldsymbol{M}$ represents the complete genotype of all markers; vector $\boldsymbol{\lambda}$ indicates the QTL map positions for linked QTLs; $\boldsymbol{\varphi}$ represents the currently linked QTL; vector $\boldsymbol{\eta}$ denotes allele frequencies for all QTLs and markers; and $\boldsymbol{\sigma_e^2}$ is the residual variance.

The prior distribution for each parameter in the joint distribution function is designated in the following way:

$$\begin{aligned}
a_i &\sim N\ (0,\ \tau^2) \\
d_i &\sim N\ (0,\ \tau^2) \\
\boldsymbol{\eta} &\sim Di\ (1,\ 1,\ 1,\ \cdots) \\
k &\sim U\ (0,\ k_{max}) \\
\lambda &\sim U\ (0,\ L).
\end{aligned}$$

In particular, a_i and d_i are the effects of the i^{th} QTL which have normal prior distributions with both variances equal to τ^2. In this report, τ^2 is assumed to be a constant which is estimated from phenotypic variation in the data. Furthermore, $\boldsymbol{\eta}$ has a Dirichlet prior distribution with parameter vector $\mathbb{1}$. The number of components, k, is uniformly distributed on $[1,\ k_{max}]$ where k_{max} is assumed to be 10 in this report.

2.2.3 L-Score

Through MCMC, we can generate "complete data" samples conditional on the observed data proportional to their probability given the model assumptions. Complete data samples have values for every unknown variable in the model, from complete ordered genotype information on all pedigree members to information about QTL positions and effect sizes. Estimates of quantities of interest such as posterior means for QTL effects or positions can simply be obtained by averaging across samples. Also we can measure the evidence in favor of a particular hypothesis, H_1 against another, H_2 by using Bayes factors, which in our case are based on the integrated marginal distributions of the data given the hypotheses. However, since a priori the number of linked loci and their location is also a random variable, if one integrates over all the space of possible positions (the whole genome), traditional Bayes factors will not be of great use to identify the position where a putative locus may lie. To estimate the most likely chromosomal regions for linked loci, the following procedure has been

proposed. Chromosomes are divided into equal-length bins (1 cM). The prior probability of linkage of at least one trait locus to a particular bin (expected under a random uniform distribution over the whole genome of size L), denote as p, can be calculated as $p = 1-(1-t/L)^n$, where t is bin size and L is total genome map length for n trait loci. Given a set of complete data samples, the posterior probability q of linkage to that region in a given sample is 1 if at least one QTL is located in the bin, and 0 otherwise. The value q/p is averaged over all sampling iterations to obtain an L-score. The L-score is a conservative estimate of the Bayes factor (ratio of posterior to prior odds), where the null hypothesis is that QTLs are evenly distributed along the genome, and the alternative hypothesis is that QTLs are more likely to be linked to a given bin (Wijsman et al. 2004).

One must be cautious in the use of L-scores to evaluate the evidence, because they can vary in repeated analyses because of sampling, mixing, or lack of convergence. Furthermore, the magnitude of the L-score at a real locus can differ with model changes (Shmulewitz and Heath, 2001; Snow and Wijsman, 1998).

Chapter 3

Markov Chain Monte Carlo

The Bayesian paradigm is used to obtain estimators of the posterior average and the posterior mode based on the posterior distribution. However, analytic solutions can be obtained for simple posterior distribution densities such as the uniform distribution. The estimation can become computationally demanding when the model is high dimensional and/or contains many latent variables or missing data. The models that are used in this report belonging to this class have complex and intensive computations.

In genetic linkage analysis, most statisticians concentrate their research on calculating the likelihoods. The Expectation-Maximization (EM) algorithm (Dempster et al. 1977) is the most popular statistical method that has been used to estimate the maximum likelihood parameters when latent variables are involved in models. The Lander-Green algorithm (Lander and Green 1987; Kruglyak et al. 1995) can be applied to situations involving small pedigree containing large numbers of loci.

With pedigrees of large size, the peeling technique is efficient for small numbers of loci (Cannings et al. 1978). However, both approaches have computing limitations, especially when there are a large number of marker loci and large pedigrees. In this report, the Markov chain Monte Carlo (MCMC) sampling scheme is employed to overcome these issues related to complex calculations.

Traditional the MCMC sampling scheme applies to the situation in which the parameter space has fixed dimension. However, when the parameter space is variable MCMC with a reversible jump step the methodology is used to obtain a sampler generated under different numbers of quantitative trait loci (Heath 1997). This is in general terms the Bayesian methodology used to carry on the inference for the quantitative trait loci mode of inheritance, allele frequency, map position, number of loci affecting the trait and effect size of these trait loci.

3.1 Classical Monte Carlo integration and importance sampling

3.1.1 Markov chain

According to the definition, a Markov chain is a sequence of random variables with the property that the future state only depends on the present state and is independent of the past states. In other words, if a sequence of random variables

$\{X^{(0)},\ X^{(1)},\ X^{(2)}, \cdots\}$ can be expressed in the form

$$Pr\ (X^{(t+1)} = x^{(t+1)} |\ X^{(0)} = x^{(0)},\ X^{(1)} = x^{(1)}, \cdots, X^{(t)} = x^{(t)})$$

$$= Pr\ (X^{(t+1)} = x^{(t+1)} |\ X^{(t)} = x^{(t)}),$$

then this sequence is a Markov chain. At any state t $(t \geq 0)$, $x^{(t+1)}$, which is the value of the next state $(t+1)$, can be obtained from the conditional distribution $P\ (x|X^{(t)})$.

3.1.2 Monte Carlo simulation

Suppose that a sample $x_1, x_2, \cdots, x_T$ generated from density function $\pi(x)$ is used to approximate the integration,

$$E_\pi[f(x)] = \int_x f(x)\pi(x)dx < \infty \tag{3.1}$$

with $f(x) \geq 0$ for all x, by using the empirical average $\widehat{f_T}$, where

$$\widehat{f_T} = \frac{1}{T} \sum_{t=1}^{T} f(x_t).$$

According to the law of large numbers, when $x_1, x_2, \cdots, x_T$ are independently sampled from $\pi(\cdot)$, $\widehat{f_T}$ converges almost surely to $E_\pi[f(x)]$. That is

$$\widehat{f_T} \xrightarrow{a.s.} E_\pi[f(x)], \qquad T \to \infty. \tag{3.2}$$

This procedure is known as the Monte Carlo method (Metropolis and Ulam 1949).

3.1.3 Importance sampling

In most cases, it is not straightforward to directly obtain the sample from $\pi(x)$ and an alternative approach has to be used. A method used to approximate the defined integral in equation (3.1) without sampling from the distribution of $\pi(x)$ is the so-called importance sampling.

Considering a density function $h(x)$ which has the same support as $\pi(x)$, we can rewrite the integral (3.1) in order to obtain the expectation I with respect to h. Equation 3.1 can be expressed in the form

$$\begin{aligned} I = E_h\Big[\frac{f(x)\pi(x)}{h(x)}\Big] &= \int_x \frac{f(x)\pi(x)}{h(x)} h(x)dx \\ &= \int_x f(x)\pi(x)dx \\ &= E_\pi f(x). \end{aligned}$$

Assume that $x_1, x_2, \cdots, x_T$ are independently drawn from the density function $h(x)$, where

$$\widehat{I} = \frac{1}{T}\sum_{j=1}^{T} \frac{f(x_j)\pi(x_j)}{h(x_j)} \tag{3.3}$$

is an unbiased estimator for I. This estimator $\widehat{I}$ is also a strongly consistent estimator of I when T approaches infinity with probability 1 (Feller 1968).

$$\widehat{I} \xrightarrow{a.s.} I \qquad \text{as} \quad T \to \infty.$$

Moreover, Geweke (1989) proved a particular version of the Central Limit Theorem

which states that

$$\sqrt{T}\,\frac{\widehat{I}-I}{\sigma} \longrightarrow N(0,1) \qquad \text{as} \quad T \to \infty,$$

with $\sigma^2 = \int \left[\frac{f^2(x)\pi^2(x)}{h(x)}\right] dx - I^2$. The constant σ^2 depends on the density function $h(x)$. Even though there is no restriction on choosing the importance density, $h(x)$, it should be selected so that it has the "same shape" as $\pi(x)$ with the easy sampling conditions. However, it may not be possible to have easy sampling adapted density in genetic problems that are built on complicated models.

3.2 Markov chain Monte Carlo (MCMC) sampling methodology

Under an MCMC scheme for a large t, the distribution of $X^{(t)}$ is independent of $X^{(0)}$. Also, when t is sufficiently large, the distribution of $X^{(t)}$ converges to the stationary distribution. The essential idea for the MCMC algorithm is to have a procedure in which the stationary distribution has density $\pi(x)$ in order to obtain a sample from such a density.

We consider the Markov chain $\{X^{(t)}\}_{t\geq 0}$ where the initial transition probability function for a move from x to x' can be expressed as

$$p(x,\ x') \ = \ P(x \to x') = \ P(X^{(t+1)} = x' \mid X^{(t)} = x).$$

If one assumes that the transition kernel $p(\cdot\,,\cdot\,)$ is independent of state t, then the

transition probability function at the t^{th} step is

$$p(t;\ x, x^{'}) \ = \ P(X^{(t+r)} = x^{'} |\ X^{(r)} = x)$$

for r= 0, 1, 2,$\cdots$. If the distribution function of $X^{(0)}$ is

$$m(x) = P(X^{(0)} = x),$$

after t steps, the marginal distribution of $X^{(t)}$ is

$$m^{(t)}\ (x) = P(X^{(t)} = x).$$

If $\pi(x)$ is the stationary distribution density of transition kernel $p\ (\cdot\ ,\ \cdot)$, then $\pi(x)$ satisfies the condition

$$\int p(x,\ x^{'})\ \pi(x)dx \ = \ \pi(x^{'}).$$

In fact, the marginal distribution of $X^{(t)}$ in any state t $(t \gg 0)$ is arbitrarily close to the stationary distribution $\pi(x)$ without considering the initial value of $X^{(0)}$. Under the MCMC framework, the major difference between the Metropolis-Hastings algorithms and Gibbs sampling method is determined by the variety of ways setting up the transition kernel.

3.2.1 Metropolis-Hastings algorithm

The Metropolis-Hastings method is a common algorithm used to generate samples from a complicated and/or high-dimensional probability distribution. The main

idea of Metropolis-Hastings algorithm is to compare the acceptance ratio with the acceptance probability, in order to decide whether either if each chain should move to the next state or remain at the current state.

On a Markov chain $\{X^{(t)}\}_{t\geq 0}$, we have $X^{(t)} = x$ at state t. If a move from the current state x to a new state $x^{'}$ is proposed with proposal probability $q(x, x^{'})$, we need to choose a transition kernel $p(x, x^{'})$ such that

$$p(x, x^{'}) = q(x, x^{'})a(x, x^{'}), \tag{3.4}$$

where $a(x, x^{'})$ is known as the acceptance probability ($0 < a(x, x^{'}) \leq 1$). Therefore, $x^{'}$ will be accepted as the value for state $(t+1)$, that is $X^{(t+1)} = x^{'}$, with probability $a(x, x^{'})$. Otherwise, $x^{'}$ will remain at state t, $X^{(t)} = x^{'}$, with probability $(1-a(x, x^{'}))$. To implement this, we randomly selected u from a uniform distribution between 0 and 1, then we have

$$X^{(t+1)} = \begin{cases} x^{'}, & u \leq a(x, x^{'}) \\ x, & u > a(x, x^{'}). \end{cases}$$

Remember that the goal is to choose a probability function $a(x, x^{'})$ such that the stationary distribution of $p(x, x^{'})$ is the same as the posterior distribution $\pi(x)$. The most common acceptance probability function was given by Hastings (1970) which is shown in the form of

$$a(x, x^{'}) = \min\{1,\ A\}, \tag{3.5}$$

where A is the acceptance test ratio which is indicated as

$$A = \frac{\pi(x^{'})\ q(x^{'}, x)}{\pi(x)\ q(x, x^{'})}. \tag{3.6}$$

In order to increase ratio A, the best method is to choose a proposal distribution function $q(x)$ such that this $q(x)$ is proportional to the stationary distribution function $\pi(x)$.

In addition, the transition kernel $p(x, x^{'})$ with acceptance probability $a(x, x^{'})$ is given as

$$p(x, x^{'}) = \begin{cases} q(x, x^{'}), & \pi(x^{'})q(x^{'}, x) \geq \pi(x)q(x, x^{'}) \\ q(x^{'}, x)\dfrac{\pi(x^{'})}{\pi(x)}, & \pi(x^{'})q(x^{'}, x) < \pi(x)q(x, x^{'}) \end{cases}. \tag{3.7}$$

The Markov chain which is built up on the transition kernel described in equation (3.7) is reversible. Moreover, this form of $p(x, x^{'})$ guarantees that $\pi(x)$ is the stationary distribution of this Markov chain. That is

$$\pi(x)p(x, x^{'}) = \pi(x^{'})p(x^{'}, x).$$

Proof: Let $x \neq x^{'}$, then

$$\begin{aligned} \pi(x)p(x, x^{'}) &= \pi(x)\ q(x, x^{'})\ \min\left\{\ 1,\ \frac{\pi(x^{'})\ q(x^{'}, x)}{\pi(x)\ q(x, x^{'})}\ \right\} \\ &= \min\left\{\ \pi(x)q(x, x^{'}), \pi(x^{'})q(x^{'}, x)\ \right\} \\ &= \pi(x^{'})\ q(x^{'}, x)\ \min\left\{\ 1,\ \frac{\pi(x)\ q(x, x^{'})}{\pi(x^{'})\ q(x^{'}, x)}\ \right\} \\ &= \pi(x^{'})\ p(x^{'}, x) \end{aligned}$$

Therefore,

$$\begin{aligned}\pi(x^{'}) &= \int \pi(x)\ p(x, x^{'})dx \\ &= \int \pi(x^{'})\ p(x^{'}, x)dx \\ &= \pi(x).\end{aligned}$$

3.2.1.1 Independent chain

If the proposal distribution $q(x, x^{'})$ is independent of the current state x on the chain, then according to formula (3.4), we have $q(x, x^{'}) = q(x^{'})$. Hence, the acceptance test ratio A in equation (3.6) becomes

$$A = \frac{\pi(x^{'})/\ q(x^{'}, x)}{\pi(x)/\ q(x, x^{'})},$$

and the acceptance probability function is

$$a(x, x^{'}) = \min\left\{\ 1,\ \frac{\pi(x^{'})/\ q(x^{'})}{\pi(x)/\ q(x)}\ \right\}.$$

The prior density function is usually used as the proposal distribution function $q()$ for each independent chain. The complete information about $\pi(x)$ is not necessary; however, multiplicative constants will be needed.

3.2.2 Gibbs sampler

The Gibbs sampling method, which is a special case of the Metropolis-Hastings algorithm, was introduced by Geman and Geman (1984). Even though limitations

exist on choosing instrumental distributions, there are several advantages of the Gibbs sampler over other sampling methods. The most extraordinary advantage of the Gibbs sampling method in practice, especially for high dimensional problems, is that the Gibbs sampler can be simulated from only the full conditional density $\pi(\boldsymbol{x}_S|\boldsymbol{x}_{-S})$, where $\boldsymbol{x}_S = \{x_i, i \in S\}$, $\boldsymbol{x}_{-S} = \{x_i, i \notin S\}$, and $S \subset \{1, 2, \cdots, n\}$, i.e. $\boldsymbol{x}_S \in R^{|S|}$ and $\boldsymbol{x}_{-S} \in R^{n-|S|}$.

Given that S and $\boldsymbol{X}_{-S} = \boldsymbol{x}_{-S}$, $\pi(x)$ is the distribution density function of $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$, a collection of random variables $\boldsymbol{X}' = (X_1', X_2', \cdots, X_n')$ with property $\boldsymbol{X}'_{-S} = \boldsymbol{X}_{-S}$ have density function $\pi(\boldsymbol{x}_S'|\ \boldsymbol{x}_{-S})$. Therefore, for any B, we have

$$\begin{aligned} P(X' \in B) &= \int_B \pi(\boldsymbol{x}'_{-S})\ \pi(\boldsymbol{x}_S'|\ \boldsymbol{x}'_{-S})\ d\boldsymbol{x}' \\ &= \int_B \pi(\boldsymbol{x}')\ d\boldsymbol{x}' \\ &= \pi(B). \end{aligned}$$

Thus, $\pi(\boldsymbol{x})$ is also the distribution density of $\boldsymbol{X}'$ which implies that $\pi(\boldsymbol{x})$ is indeed the stationary distribution function. Given a proposed move from $\boldsymbol{x}$ to $\boldsymbol{x}'$, the transition kernel, which is formed by the full conditional distribution and defined by the above equation, can be expressed as

$$P_S(\boldsymbol{x} \to B) = I[\boldsymbol{x}_S' \in B_{-S}] \cdot \int_{B_S} \pi(\boldsymbol{x}_S'|\boldsymbol{x}'_{-S}) d\boldsymbol{x}_S'.$$

In particular, when S contains only one element, the process is called a single-site Gibbs sampling method. For example, given that $S = \{i\}$ and $x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$,

the single-site Gibbs sampler which is based on the full conditional distribution of $(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$, has an acceptance test ratio A represented as

$$\begin{aligned} A &= \frac{\pi(\boldsymbol{x}_S^{'}, \boldsymbol{x}_{-S}) \cdot \pi(\boldsymbol{x}_S | \boldsymbol{x}_{-S})}{\pi(\boldsymbol{x}_S, \boldsymbol{x}_{-S}) \cdot \pi(\boldsymbol{x}_S^{'} | \boldsymbol{x}_{-S})} \\ &= \frac{\pi(\boldsymbol{x}_S^{'}, \boldsymbol{x}_{-S})}{\pi(\boldsymbol{x}_S, \boldsymbol{x}_{-S})} \cdot \frac{\pi(\boldsymbol{x}_S, \boldsymbol{x}_{-S})}{\pi(\boldsymbol{x}_{-S})} \cdot \frac{\pi(\boldsymbol{x}_{-S})}{\pi(\boldsymbol{x}_S^{'}, \boldsymbol{x}_{-S})} \\ &= 1. \end{aligned}$$

Thus, the acceptance probability, $a(\boldsymbol{x}, \boldsymbol{x}^{'}) = \min\{1,\ A\}$, is always 1. Therefore, the transition kernel becomes

$$P_i(\boldsymbol{x} \rightarrow B) = I[\boldsymbol{x}_{-i} \in B_{-i}] \cdot \int_{B_i} \pi(x_i | \boldsymbol{x}_{-i}) d(x_i).$$

The distribution density can be obtained by repeating the Gibbs sampling algorithm and it can be shown that the density function converges to $\pi(\boldsymbol{x})$.

Given the initial point $\boldsymbol{x}^{(0)} = (x_1^{(0)}, \cdots, x_i^{(0)}, \cdots, x_n^{(0)})$, if the starting point is $x^{(t-1)}$ for the t^{th} step, then the Gibbs sampling method for the t^{th} step is described as follows:

(1) $x_1^{(t)}$ from the full conditional distribution $\pi(x_1 |\ x_2^{(t-1)}, \cdots, x_n^{(t-1)})$;

...

(i) $x_i^{(t)}$ from the full conditional distribution $\pi(x_i |\ x_1^{(t)}, \cdots, x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, \cdots, x_n^{(t-1)})$;

...

(n) $x_n^{(t)}$ from the full conditional distribution $\pi(x_n|\ x_1^{(t)}, \cdots, x_{n-1}^{(t)})$.

As a result, $x^{(1)}, x^{(2)}, \cdots, x^{(t)}, \cdots$ are the realization of a Markov chain where $\boldsymbol{x}^{(t)} = (x_1^{(t)}, \cdots, x_n^{(t)})$. The transition probability distribution function for a move from $\boldsymbol{x}$ to $\boldsymbol{x}^{'}$ is

$$p\ (\boldsymbol{x}, \boldsymbol{x}^{'}) = \pi(x_1|\ x_2, \cdots, x_n)\ \pi(x_2|\ x_1^{'}, x_3 \cdots, x_n) \cdots\ \pi(x_n|\ x_1^{'}, \cdots, x_{n-1}^{'}),$$

and $\pi(\boldsymbol{x})$ is the stationary distribution.

3.2.3 Full conditional distribution and partial conditional distribution

3.2.3.1 Full conditional distribution

Previous subsections indicated that the single-site Gibbs sampler is based on the full conditional distribution. Let $\boldsymbol{x} = (x_1, \cdots, x_n)$, then we have

$$\pi(\boldsymbol{x}) = \prod_{i=1}^{n} \pi(x_i|\boldsymbol{x}_{<i}), \tag{3.8}$$

where $\boldsymbol{x}_{<i} = \{x_j, j < i\}$. According to equation (3.8), the MCMC algorithm would not be necessary if the density function $\pi(x_i|\boldsymbol{x}_{<i})$ of $(x_1, \cdots, x_n)$ was known. However, this simple case cannot be applied to complicated situations such as genetic linkage analysis.

For any $S \subset N$, we have

$$\pi(\boldsymbol{x}_S|\boldsymbol{x}_{-S}) = \frac{\pi(\boldsymbol{x})}{\int \pi(\boldsymbol{x}) d\boldsymbol{x}_S} \propto \pi(\boldsymbol{x}). \tag{3.9}$$

Equation (3.9) shows that the inference could be only made from term $\boldsymbol{x}_S$. Similarly, given $\boldsymbol{x}_{-S} = \boldsymbol{x}'_{-S}$, we have

$$\frac{\pi(\boldsymbol{x}'_S|\boldsymbol{x}'_{-S})}{\pi(\boldsymbol{x}_S|\boldsymbol{x}_{-S})} = \frac{\pi(\boldsymbol{x}')}{\pi(\boldsymbol{x})}. \tag{3.10}$$

3.2.3.2 Partial conditional distribution

An update of the MCMC algorithm using the partial conditional distribution was discussed under special conditions (Besag et al. 1995). Indeed, $\boldsymbol{x}$ is an invalid value if it was updated on the unconditional distribution of $\boldsymbol{x}_{-i}$. If an invalid $\boldsymbol{x}$ value occurs, this invalid $\boldsymbol{x}$ can be ignored. Then, we can update it using the Gibbs sampler since the Gibbs sampling algorithm is independent of the current value.

3.3 Reversible jump Markov chain Monte Carlo

The MCMC sampling scheme is constructed with a fixed dimension of parameter space. However, the reversible jump technique (Green 1995) allows the movement of the sampler between different parameter spaces with unequal dimensions.

Suppose that the current state $\boldsymbol{x}$ and future state $\boldsymbol{x}'$ have different dimensions denoted as d_1 and d_2 respectively, where d_1 is smaller than d_2. There exists a vector $\boldsymbol{u}$ such that the length of vector $\boldsymbol{u}$ is the difference between d_1 and d_2. The extra

elements are discarded during the reversible jump step. In such a case, the acceptance test ratio A has the form

$$A = \frac{\pi(\boldsymbol{x}')\ q(d_1; d_2)}{\pi(\boldsymbol{x})\ q(d_2; d_1)\ q(\boldsymbol{u})} \cdot J,$$

where J is the Jacobian term for the transformation between $(\boldsymbol{x}, \boldsymbol{u})$ and $\boldsymbol{x}'$, represented as $J = \left| \frac{\partial \boldsymbol{x}'}{\partial(\boldsymbol{x}, \boldsymbol{u})} \right|$, $\pi(\boldsymbol{x})$ denotes the posterior distribution function, and $q(\cdot)$ is the proposal probability density function (Green 1995; Richardson and Green 1997; Heath 1997).

3.4 Sampling scheme for quantitative trait loci

In Chapter 2, the joint distribution function of a large pedigree data set has been defined as

$$p\ (k,\ \boldsymbol{G},\ \boldsymbol{M},\ \boldsymbol{\beta},\ \boldsymbol{\lambda},\ \boldsymbol{\varphi},\ \boldsymbol{\eta},\ \boldsymbol{\alpha},\ \boldsymbol{\sigma}_e^2,\ \boldsymbol{\mu},\ \boldsymbol{Y}).$$

Based on the full conditional distribution, overall mean $(\boldsymbol{\mu})$, covariate effects $(\boldsymbol{\beta})$, residual variance $(\boldsymbol{\sigma}_e^2)$, and marker frequencies $(\boldsymbol{\eta})$ can be updated first by using Gibbs sampling method (Heath 1994, 1997).

Guo and Thompson (1992) investigated a study in which they combined segregation and linkage analysis on complex large pedigrees, using the MCMC approach to simulate samples. The genotype can be updated individually from any given locus. Alternatively, the Gibbs sampler with the peeling technique can be used to update all marker genotypes $(\boldsymbol{M})$ and QTL genotypes $(\boldsymbol{G})$ simultaneously, with all individuals

at a given locus, within a complex pedigree (Ott 1989; Kong 1991; Heath 1997). This approach is called the reverse peeling method (Heath 1997). Even though the applied pedigrees have to be peelable at a single locus, the reverse peeling method avoids the irreducibility problems better than individual-by-individual updating steps (Lin et al. 1993; Heath 1997).

As described in Chapter 2 and 3, the reversible jump MCMC algorithm allows one to collect samples from the posterior distribution within the spaces with different dimensions (Green 1995). Therefore, given any QTL genotype G_i, both reverse peeling and a reversible jump step are required in order to update the information on the QTL map position (λ_i) and linkage status (φ_i) using the Gibbs algorithm because it leads to a change in model dimensions. The movement is between either marker intervals or chromosomes that depend on the partial conditional distribution (Besag et al. 1995; Heath 1997).

With a successful birth step, the QTL effects, frequency, linkage status, map position, and genotypes for pedigree members are generated using the reverse peeling method for new QTLs. For a death step, a random selected QTL is discarded (Heath 1997). When a move is proposed, the acceptance probability is not affected because the peeling method has been applied. In contrast, split/combine steps contain the changing of two QTLs. In a split step, an existing QTL is randomly selected and the effect is separately distributed for two QTLs. For a combine step, a reverse process, we combine two selected QTLs' effects in order to obtain a new QTL. More details

are given by Heath (1997).

The complete updating steps for one iteration based on the reversible jump MCMC method was suggested by Heath (1997) are listed as follows:

Updating Procedure

1. Update complete marker genotypes $\boldsymbol{M}$ for each locus in turn;
2. For each QTL i:
   (a) Update QTL effects α_i;
   (b) Update QTL position λ_i and linkage status φ_i;
   (c) Update QTL genotypes G_i;
3. Update QTL and marker frequencies $\boldsymbol{\eta}$;
4. Update covariate effects $\boldsymbol{\beta}$ and overall mean μ;
5. Update residual variance $\boldsymbol{\sigma}_e^2$;
6. Birth or death of a QTL;
7. Split one QTL into two, combine two QTLs into one.

3.4.1 Acceptance Probability

Heath (1997) stated that the acceptance probability for the change from position λ_i to λ_i' is $\min(1, A)$, where A is in the product form including the likelihood ratio, the prior ratio, and the proposal ratio for the i^{th} QTL at such positions of λ_i' and λ_i that can be expressed as

$$A = \frac{p(\boldsymbol{Y}|k, \boldsymbol{G}_{-i}, \boldsymbol{M}, \boldsymbol{\beta}, \lambda_i', \boldsymbol{\lambda}_{-i}, \boldsymbol{\varphi}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}_e^2, \mu)p(\lambda_i')q(\lambda_i; \lambda_i')}{p(\boldsymbol{Y}|k, \boldsymbol{G}_{-i}, \boldsymbol{M}, \boldsymbol{\beta}, \lambda_i, \boldsymbol{\lambda}_{-i}, \boldsymbol{\varphi}, \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}_e^2, \mu)p(\lambda_i)q(\lambda_i'; \lambda_i)},$$

when a linked QTL with no change in φ_i is proposed.

In contrast, the only major difference when using a reversible jump step is that the move leads to a change in φ_i. Heath (1997) pionted out that when a QTL moves to a linked state from an unlinked state, the map position for such a QTL has to be proposed, and the corresponding acceptance probability is $\min(1, A)$, where

$$A = \frac{p(\boldsymbol{Y}|k, \boldsymbol{G}_{-i}, \boldsymbol{M}, \boldsymbol{\beta}, \lambda_i', \boldsymbol{\lambda}_{-i}, \varphi_i', \boldsymbol{\varphi}_{-i}, , \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}_e^2, \mu)p(\varphi_i')p(\lambda_i')q(\varphi_i; \varphi_i')}{p(\boldsymbol{Y}|k, \boldsymbol{G}_{-i}, \boldsymbol{M}, \boldsymbol{\beta}, \boldsymbol{\lambda}_{-i}, \varphi_i', \boldsymbol{\varphi}_{-i}, , \boldsymbol{\eta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}_e^2, \mu)p(\varphi_i)q(\lambda_i', \varphi_i'; \varphi_i)}.$$

On the other hand, if a reversible jump step is applied, then the map position is discarded.

Chapter 4

Case Study: Hypercalciuria

Hypercalciuria is defined as an elevated level of urinary calcium excretion. Although by itself hypercalciuria may not be considered as a medical condition, quite often, it is found in patients who have been diagnosed with calcium nephrolithiasis or kidney stones in a clinical setting. Briefly, hypercalciuria is diagnosed using the criteria that the amount of urine calcium found over a 24-hour period exceeds 250 mg for females and 300 mg for males. Furthermore, patients with hypercalciuria appear to be more vulnerable to calcium nephrolithiasis. Because of the significant correlation between hypercalciuria and calcium nephrolithiasis, it is suspected that both traits share a common genetic background (Coe et al. 1979; Pak et al. 1981; Petrucci et al. 2000; Polito et al. 2000).

In a study of French-Canadian families, Tessier et al. (2001) indicated that stone formation is likely regulated by a metabolic phenotype delineating substantial calcium excretion in the urine. Loredo-Osti et al. (2005) performed a complex segregation

analysis of urine calcium excretion using the relatives of patients from 221 nuclear French-Canadian families in which nephrolithiasis was identified. They concluded that there is most likely a major gene with a polygenic background determining the calcium excretion trait.

Here a genetic linkage analysis for hypercalciuria was conducted by using the Bayesian methodology on the same aforementioned French-Canadian data set. Bayes factors, more specifically, L-Score profiles were used to identify the locations of putative linked loci. The results for this case study show two L-Score peaks: one with value 16.20 at position 41.55cM on chromosome 15 and the other with value 11.09 at position 92.75cM on chromosome 19. According to the Bayes factor criterion, both peaks' positions are strongly linked to hypercalciuria, suggesting that the position of the gene(s) that regulate urinary calcium excretion may be in the neighborhood of these peaks.

4.1 Data description

This study investigated 1219 individuals from French-Canadian families who were identified as having nephrolithiasis. However, only 985 participants provided blood samples and urine samples over a 24-hour period. At the time of examination, data regarding age, gender, weight, height, and thiazide drug use were also obtained for these participants. Based on a biochemical analysis of urine and serum samples, the urine creatinine level for each individual was predicted using the Cockcroft-Gault

Table 4.1: Summary statistics for the study sample (N=897)

Variables	**Males** (N=455)	**Females** (N=442)
Number of observations	455	442
Age at examination *years*	48.6 ± 12.1	49.2 ± 12.2
Weight *kg*	78.6 ± 13.4	65.4 ± 12.5
Height *cm*	172.4 ± 6.2	159.4 ± 6.5
Body mass index kg/m^2	26.5 ± 4.4	25.8 ± 4.8
Serum creatinine $\mu mol/L$	99.7 ± 13.0	83.2 ± 13.3
Urine calcium *mmol/24 hours*	6.2 ± 2.9	4.9 ± 2.3

(Source: Loredo-Osti et al (2005))

formula. The well-known Cockcroft-Gault formula is defined as

$$\frac{(140 - \text{Age}) \times \text{Mass (in } kg) \times \text{Constant}}{\text{Serum Creatinine(in } \mu\text{mol/L)}},$$

where the constant is 1.23 for males and 1.04 for females. To control for any over- and/or under-collection of a urine sample within any 24-hour period, an individual's actual urine creatinine level must lie within a predicted range in order for the sample to be considered suitable for inclusion in the study. If a collected sample for a given individual is found to be greater or less than 20% of predicted level, as calculated by the Cockcroft-Gault formula, then another urine sample and blood sample for an additional 24-hour period would need to be collected. Of the 985 people examined, 897 individuals representing 154 two-, three- or four-generation pedigrees were selected.

The Cockcroft-Gault formula takes into consideration the fact that males have a higher urine creatinine clearance than females at the same level of serum creatinine.

As expected, the test data set contained more males than females (455 men versus 442 women; see Table 4.1). There was no significant difference in the average age of males and females. It was found that 455 male participants had higher levels of both serum creatinine and urine calcium compared with females (99.7 $\pm$ 13.0 versus 83.2 $\pm$ 13.3 μmol/L; 6.2 $\pm$ 2.9 versus 4.9 $\pm$ 2.3 mmol/24 hours, respectively). Even though the mean level of serum creatinine of both males and females in these French-Canadian families were in the normal range (60-110 μmol/L for men, and 45-90 μmol/L for women), the averages approached the upper limit. On average, the male group had an apparently higher index of urine calcium excretion than did the female group (6.2 $\pm$ 2.9 versus 4.9 $\pm$ 2.3, respectively). As shown in Table 4.1, means and standard deviations are listed for each variable corresponding to the different genders. It is evident that all the parameters were higher in males than in females except for the age at which the disease was ascertained.

4.2 Statistical analysis

In previous chapters, we have discussed the Bayesian inference for a mixture model with an unknown number of components, as well as the application of the Bayesian method to genetic linkage analysis under a model in which the location of the QTL is treated as a variable. The objective of this analysis was to determine the existence of the linkage between a QTL and the gene(s) controlling calcium excretion. The null and the alternative hypotheses are stated as

H_0: QTL_i is linked with the gene controlling calcium excretion.

H_1: No such genetic linkage exists.

In this report, the L-Scores with 1 cM bins were used as the estimates of the Bayes factors (ratio of posterior to prior odds) for the 22 autosomes under the premise that a larger peak L-Score at a given position provides evidence of genetic linkage at that position.

4.2.1 Results

The linkage analysis can be done by the study of the recombination patterns. The theory indicates that the recombination fraction (r) would be 50% if the transmission of two loci were independent, i.e. in absence of linkage. On the other hand, a recombination fraction smaller than 50% provides evidence for linkage: the smaller that the estimate of the recombination fraction is, the stronger the evidence that the estimate will provides. One of the most popular methods of linkage analysis is the LOD score, which is defined as negative logarithm base 10 of the likelihood ratio. In contrast to the LOD score, L-Scores are the estimates of the Bayes factors, which were discuessed in Chapter 2.

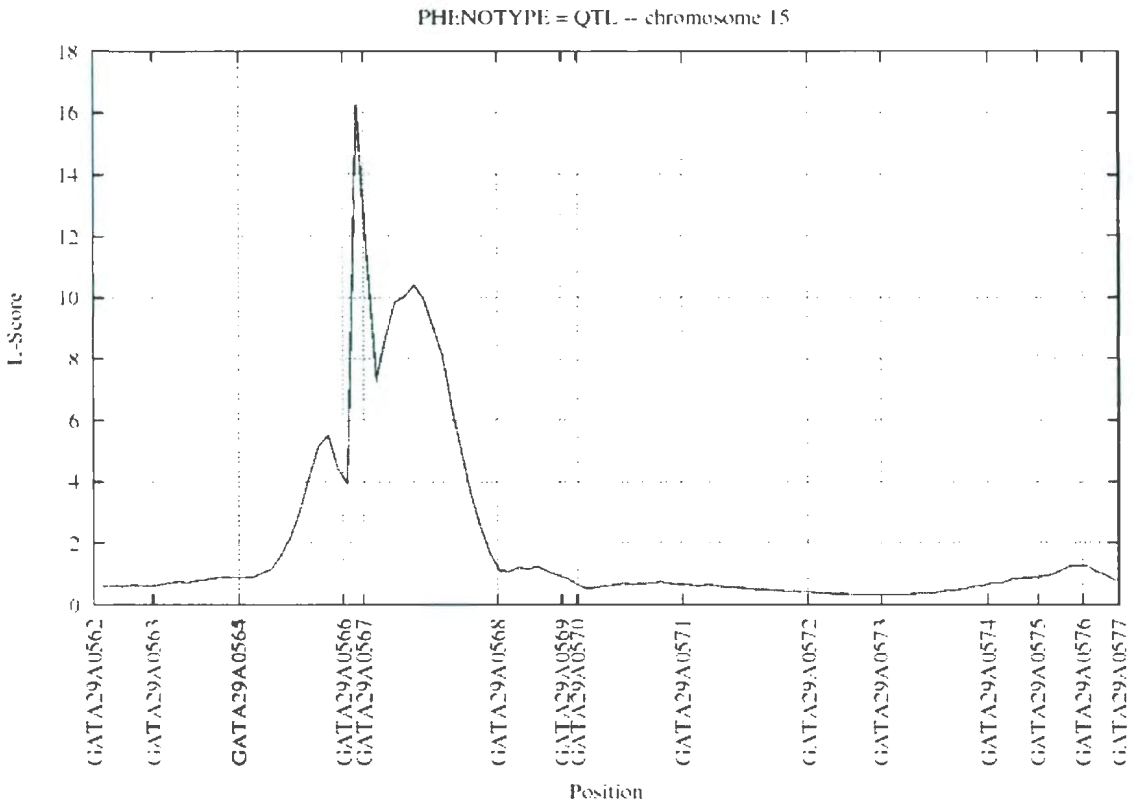

Figure 4.1: Estimates of the L-score, when the model is fitted for at least one QTL, which is being linked to a chromosomal region. The peak represents the possible position of the linkage on chromosome 15. The total length of chromosome 15 is 122.42 cM.

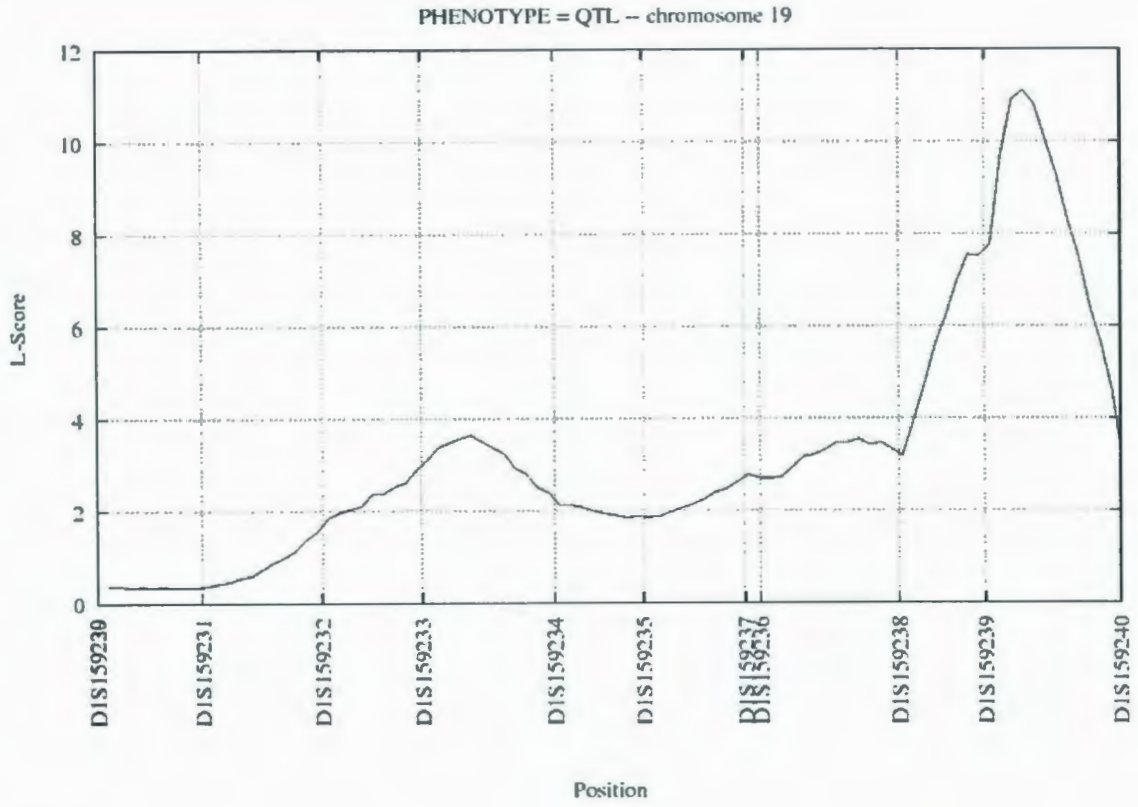

Figure 4.2: Estimates of the L-score, when model is fitted for at least one QTL, which is being linked to a chromosomal region. The peak represents the possible position of the linkage on chromosome 19. The total length of chromosome 19 is 101.98 cM.

In our data set, the L-Score peaks on chromosomes 15 and 19 with much larger peak values than on the other 20 autosomes. Fig. 4.1 and Fig. 4.2 show the plots of the L-Scores for each locus on chromosomes 15 and 19, respectively. Summary statistics for the 22 autosomes are given in Table 4.2.

An inspection of the generated complete data sample set seems to indicate good mixing. Nevertheless, the identified loci in this analytic represent candidate regions that need to be further evaluated by using a denser set of markers as well as other analytical approaches.

Table 4.2: The estimates of Bayes factors (L-Scores) from the linkage analysis

Linkage group	Count	Prop. linked	Ave. L-Score	Peak L-Score	Location
Unlinked	930864	0.46543	1.27518		
Chromosome 1	904448	0.45222	1.03488	3.96	52.95 cM
Chromosome 2	642322	0.32116	0.73559	1.68	238.96 cM
Chromosome 3	695256	0.34763	0.88716	2.77	102.41 cM
Chromosome 4	657705	0.32885	0.92284	2.81	13.32 cM
Chromosome 5	471848	0.23592	0.65118	1.91	168.93 cM
Chromosome 6	505210	0.25261	0.76467	1.27	88.89 cM
Chromosome 7	469143	0.23457	0.74524	2.00	42.72 cM
Chromosome 8	826041	0.41302	1.39525	4.30	74.89 cM
Chromosome 9	422940	0.21147	0.72367	1.18	14.42 cM
Chromosome 10	912216	0.45611	1.54213	8.11	70.51 cM
Chromosome 11	605605	0.30280	1.08709	3.25	50.33 cM
Chromosome 12	498038	0.24902	0.84211	1.90	163.69 cM
Chromosome 13	437588	0.21879	1.02289	2.82	51.52 cM
Chromosome 14	271167	0.13558	0.60671	1.03	84.55 cM
Chromosome 15	758818	0.37941	1.90601	16.20	41.55 cM
Chromosome 16	246915	0.12346	0.51354	0.92	112.47 cM
Chromosome 17	312095	0.15605	0.58970	0.86	99.13 cM
Chromosome 18	487481	0.24374	1.02018	2.28	53.50 cM
Chromosome 19	1052489	0.52624	3.00155	11.09	92.75 cM
Chromosome 20	415320	0.20766	1.16595	1.99	31.36 cM
Chromosome 21	244651	0.12233	1.12247	2.01	57.66 cM
Chromosome 22	140717	0.07036	0.52246	0.69	6.25 cM

4.2.2 Conclusion

In preparation for later application to a data set, we described the use of Bayesian methodology to carry out a genetic linkage analysis. The case study comprised data on urinary calcium excretion of families ascertained through a proband diagnosed with nephrolithiasis. The analysis identified two condidate regions of the genome (one on chromosome 15 and the other on chromosome 19) strongly linked to hypercalciuria. Although the nature of this work is preliminary since more studies are needed to confirm the finding, the nature of the L-Score profile suggests that the identified linked regions are promising.

Bibliography

Besag J, Green P, Higdon D, Mengersen K (1995) Bayesian computation and stochastic systems. Stat Sci 10:3-66

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Adv Appl Prob 10:26-61

Coe FL, Parks JH, Moore ES (1979) Familial isiopathic hypercalciuria. N Eng J Med 300:337-340

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Roy Statist Soc B 39:1-38

Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. J Roy Statist Soc B 56:363-375

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523-542

Escobar MD, West M (1995) Bayesian density estimation and inference using mix-

tures. J Am Statist Ass 90:577-588

Feller W (1968) An Introduction to probability Theory and Its Applications, Volume I. (3rd edition) Wiley, New York

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Amer Statist Assoc 85:398-409

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal Mach Intell 6:721-741

Geweke J (1989) Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57:1317-1339

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82:711-732

Guo SW (1991) Monte Carlo method on quantitative genetics. PhD. thesis, Department of Biostatistics, University of Washington, Seattle, Washington

Guo SW, Thompson EA (1992) A Monte Carlo method for combined segregation and linkage analysis. Am J Hum Genet 51:1111-1126

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1):97-109

Heath SC (1994) Estimation of linked QTL effects with an animal model using Gibbs sampling. In‘: Smith C, Gavora JS, Benkel B, Chesnais J, Fairfull W, Gibson JP, Kennedy BW, Burnside EB (eds) Proceedings of the Fifth World Congress in Genetics Applied to Livestock Production. Vol 18. University of Guelph, Guelph, Ontario, 398-401

Heath SC (1997) Markov Chain Monte Carlo Segregation and Linkage Analysis for Oligogenic Models. AM J Hum Genet 61:748-760

Heath SC (2003) A package for multipoint linkage analysis on large pedigrees using Reversible jump Markov chain Monte Carlo

Jansen J, de Jong AG, van Ooijen JW (2001) Constructing dense genetic linkage maps. Theor Appl Genet 102(7):1113-1122 [DOI]

Kass RE, Raftery AE (1995) Bayes Factors. J Amer Statist Assoc 90:773-795

Kong A (1991) Analysis of pedigree data using methods combining peeling and Gibbs sampling. In: Keramidas EM, Kaufman SM (eds) Computer Science and Statistics Proceedings of the 23rd Symposium on the Interface. Interface Foundation, Fairfax Station, VA 379-385

Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56:519-527

Kruglyak L, Lander ES (1998) Faster multipoint linkage analysis using Fourier transforms. J Comput Biol 5:1-7

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363-2367

Lin S, Thompson EA, and Wijsman E (1993) Achieving Irreducibility of the Markov chain monte carlo method applied to pedgree data. IMA J Math Appl Med Biol 10:1-17

Loredo-Osti JC, Roslin NM, Tessier J, Fujiwara TM, Morgan K, Bonnardeaux A (2005) Segregation of urine calcium excretion in families ascertained for nephrolithiasis: Evidence for a major gene. Kidney Int 68:966-971

Metropolis N, Ulam S (1949) The Monte Carlo method. J Amer Statist Assoc 44:335-341

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087-1091

Ott J (1989) Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 86:4175-4178

Ott J (1991) Analysis of Human Genetic Linkage. Johns Hopkins University Press, Baltimore

Pak CY, McGuire J, Peterson R, Britton F, Harrod MJ (1981) Familial absorptive hypercalciuria in a large kindred. J Urol 126:717-719

Petrucci M, Scott P, Ouimet D, Trouve ML, Proulx Y, Valiquette L, Guay G, Bonnardeaux A (2000) Evaluation of the calcium-sensing receptor gene in idiopathic hypercalciuria and calcium nephrolithiasis. Kidney Int 58:38-42

Polito C, La Manna A, Nappi B, Villani J, Di Toro R (2000) Idiopathic hypercalciuria and hyperuricosuria: family prevalence of hephrolithiasis. Pediatr Nephrol 14:1102-1104

Richardson S, Green PJ (1997) On Bayesian Analysis of Mixtures with an Unknown Number of Components. J R Statist Soc 59:731-792

Robert CP, Casella G (1999) Monte Carlo Statistical Methods. Springer-Verlag, New York

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144:805-816

Shmulewitz D, Heath SC (2001) Linkage analysis of quantitative traits for obesity, diabetes, hypertension, and dyslipidemia on the island of Kosrae, Federated States of Micronesia. Genet Epidemiol (Suppl) 21:S686-S691

Snow GL, Wijsman EM (1998) Pedigree analysis package (PAP) vs. MORGAN: model selection and hypothesis testing on a large pedigree. Genet Epidemiol 15:355-369

Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. Am J Hum Genet 58:1323-1337

Stephens DA, Smith AFM (1993) Bayesian inference in multipoint gene mapping. Ann Hum Genet 57:65-82

Stephens M (2000) Bayesian analysis of mixture models with an unknown number of components- an alternative to reversible jump methods. Ann Statist 28:40-74

Tessier J, Petrucci M, Trouve ML, Valiquette L, Guay G, Ouimet D, Bonnardeaux A (2001) A Familt-based study of metabolic phenotypes in calcium urolithiasis. Kidney Int 60:1141-1147

Uimari P, Sillanpaa MJ (2001) Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. Genetic Epidemiology 21:224-242

Wijsman EM, Daw EW, Yu CE, Payami H, Steinbart EJ, Nochlin D, Conlon EM, Bird TD, Schellenberg GD (2004) Evidence for a novel late-onset Alzheimer disease locus on chromosome 19p13.2. Am J Hum Genet 75:398-409