Inferences on the Genetic Control of Quantitative Traits from Selection Experiments.

Simon Charles Heath

BA Zoology (Oxford) 1990

Ph.D.

The University of Edinburgh

1995

Contents

List of Tables

List of Figures

DECLARATION

I declare that this thesis has been composed by me.

Specific contributions of others are acknowledged.

Simon Charles Heath

July 1995

Acknowledgments

I thank Peter Keightley and Bill Hill for supervising my project. Peter, thanks for all your help discussing problems, reading through drafts of papers and this thesis, and, most importantly, preventing me from straying *too* far from the subject area! Bill, thank you for your invaluable comments on various drafts and presentations without which this thesis would have many more flaws. I also thank my third supervisor, Grahame Bulfield of Roslin Institute for his interest in my work and his comments on several draft manuscripts. I thank the Biotechnology and Biological Sciences Research Council for supporting this work with a studentship. I would also like to thank Drs. Robin Thompson and Sara Knott for several very useful discussions on various parts of this thesis, and Drs. Sue Brotherstone and Jon Mercer for many interesting discussions, not necessarily about this thesis!

On a more personal level, I would like to thank the many friends I've made in my three years here. Without you all I would probably have done more work and spent less money, but it would have seemed like twice as long! So thank you (in more or less chronological order!) Andrew C., Jens, Esau, Mike, Jesus, Eimear, Kellie, Said, Angus, Simon, Sharon, Sarah, Tony, Dom, Andrew W., Jennie, Elly, Harold, David, Dimitrios, Victor and not forgetting Ricardo whose appetite for arguing about anything and everything has livened up many an evening! I would also, of course, like to thank my parents. Your continued support for my many odd decisions is much appreciated.

Publications

Papers submitted from this thesis:

- Heath, S.C. (1995). Estimation of the Effects and Map Positions of Linked QTL with Large Complex Pedigrees. *Genetics Selection Evolution* (submitted).

Published papers:

- Heath, S.C., G. Bulfield, R. Thompson, and P.D. Keightley, (1995). Rates of change of genetic parameters of body weight in selected mouse lines. *Genetical Research* **66**:19-25.

- Heath, S.C., (1994). Estimation of linked QTL effects with an animal model using Gibbs Sampling. In *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production* **21**:398-401.

- Keightley, P.D., S.C. Heath, T. Hardge, L. May, A.L. Archibald and G. Bulfield (1994). Changes of genetic variance and frequencies of marker alleles in mouse lines selected on 6 week weight. In *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production* **19**:61-66.

Abstract

The main aim of this thesis is the development of methods for analysing data from selection experiments to make inferences about the genetic control of the selected trait. A series of methods for data analysis are developed and applied to both simulated and experimental datasets under infinitesimal (polygenic) genetic models, discrete locus models and mixed inheritance models (which are a combination of polygenic and discrete locus models). The experimental dataset is from a replicated selection experiment on mice in which an F_2 population formed from an inbred cross was divergently selected on body weight for 20 generations.

The experimental data are initially analysed assuming the infinitesimal model using a Derivative Free Restricted Maximum Likelihood package (Meyer) to produce estimates of genetic parameters. An extension to the package is then developed to allow the variance components to change continuously over time, in effect regressing the variance components on generation number. This method allow for changes to variance components over and above what would be predicted from the infinitesimal model, thereby detecting deviations from the model. When applied to simulated data the method detects no change in additive genetic variance when a polygenic model with a large number of genes (16384) are simulated, but detects significant decreases in the additive variance, as expected, when a smaller number (32) are used. Analysis of the experimental data indicates that the additive and environmental variance components increase over the course of the experiment, significantly so in the Low selected lines. Overall there is an estimated increase in phenotypic variance of 56% in the Low lines and 14% in the

High lines. This increase in variance indicates the presence of non-additive effects within and/or between loci. The exact cause of the observed increase is far from clear, but various genetic models which *could* give such results are discussed.

Two methods are then developed to analyse multi-generation datasets where there is both marker and trait information available throughout the pedigree. The aims of the analyses are to provide estimates of the effects and map positions of putative quantitative trait loci (QTL) linked to the markers. Standard maximum likelihood or regression based mapping techniques are not suitable for datasets of this sort because the calculation of the likelihood becomes intractable. The methods are again applied to both simulated datasets and the experimental dataset, in which information on two coat colour markers was available for all animals in the pedigree. The coat colour markers both show directional changes in frequency under selection for body weight, indicating linkage to QTL affecting the trait. The first, simpler, method uses the decay of the associated effect of a marker over time to infer the effect and map distance of a QTL linked to that marker. The second method makes more use of the available data; the data are analysed using a Gibbs sampling based method in which a mixed inheritance model is fitted to the data. The method produces estimates of QTL effect and map position relative to a number of linked markers, polygenic additive variance, environmental variance, fixed effects and polygenic breeding values. The analysis of the experimental dataset indicates a QTL linked to the coat colour gene *dilute* with an estimated additive effect of 0.4g ($\approx$ 0.2 phenotypic standard deviations) and a recombination frequency between the QTL and *dilute* of 12%. The method has potential for general QTL mapping with complex pedigrees.

Chapter 1

Introduction

A trait which is genetically determined to some degree has contributions from the individual's genotype and its environment which interact to produce the individual's phenotype. Genetic analysis of such a trait typically relies on inferring the genetic contribution from an individual's genotype towards its phenotype, and to do so requires a model of the genetic mechanisms controlling the trait. The simplest model would be for a trait to be controlled by a single gene, and for some situations this model appears to be correct as it fits the observations well. A well known set of examples of such single gene traits are the traits that Mendel studied in peas where plants could be unambiguously assigned into a series of discrete categories such as being tall or short or having round or wrinkled seeds. Many traits of interest, however, are not so amenable to analysis, being more or less continuous in nature. A genetic model for these traits will necessarily be more complex than for the simple traits studied by Mendel. The main assumption underlying quantitative genetics is that continuous traits are still controlled by genes in the same way as simpler traits, the difference being that several genes act together to control continuous traits. As the genetic models for quantitative traits are more complex, analysis of such traits is more difficult than the analysis of simple traits controlled by single genes.

There are two approaches to modelling the genetics of quantitative traits. One

is to treat the genes controlling a trait statistically, considering the properties of the group of genes as a whole. The other approach is to explicitly model the properties of each individual gene controlling a trait. The two approaches require very different techniques and are generally applied to different types of problems. For general quantitative trait analysis the first approach, modelling the behaviour of all genes statistically, is the method normally used. There are two main reasons for this. It is much easier to model genes as a group rather than separately, and the distribution of individual gene effects is not known, making accurate modelling impossible. There are several drawbacks to this approach, however. The many simplifications and assumptions that are necessary to make in order to model the genes as a group can lead to substantial deviations between predictions made from the model and observations from the real world. Another drawback is that the treatment of all genes together makes it difficult to make any inferences about the effects of individual genes. Modelling genes individually, however, is not automatically a better approach; it is very much more difficult and typically cannot be done analytically for more than a couple of genes. Furthermore, to model genes individually requires knowledge of the distribution of gene effects which is not known, although work on the distribution of mutation effects may provide some information on this problem (Caballero and Keightley, 1994). For these reasons, studies using discrete gene models have typically been restricted to theoretical situations.

Several methods are presented in this thesis for analysing quantitative data in such a way as to allow inferences to be made about the genetic control of a quantitative trait. The first method is the analysis of datasets using the statistical approach to model the genes, and to detect and estimate deviations from the model. The approach presented in this thesis was to estimate any changes in variance under selection. These realized estimates of the changes in variance could be compared against the predicted changes under the statistical model. The nature of any deviations provides information about how the model differs from the 'real world'. The second approach is to explicitly estimate the effects

of some of the genes controlling the trait. One way to do so is to assume that the discrete loci considered contribute all of the genetic variance of the trait. Alternatively a hybrid between the group and discrete locus models can be used. In these 'mixed inheritance' models the genetic variance is assumed to derive from two sources, a number of discrete loci which are modelled individually and a large number of polygenic loci, which are modelled together. This allows the explicit estimation of the effects of some of the loci with large effects on the trait, while accounting for the remainder of the loci *en masse*. Using this approach for data analysis presents several problems, mainly because the calculation of the exact likelihood of such models is often intractable. Gibbs sampling is a Markov Chain Monte Carlo based technique which can be used for data analysis under mixed inheritance models, and a technique for doing so is presented in this thesis.

The methods developed are applied to the analysis of simulated datasets and an experimental dataset. The experimental dataset is from a replicated selection experiment on mice, the X Lines, in which an F_2 population was formed by crossing two inbred strains. The F_2 was split into 13 replicate lines, 6 of which were selected for high body weight and 6 for low body weight, and 1 was maintained as an unselected control. Twenty generations of within family selection were carried out, with each animal was recorded for both 6 week body weight and coat colour. There were two coat colour markers segregating in the F_2, *brown* and *dilute*, which could be independently scored.

The genetic model most commonly used in quantitative genetics is the infinitesimal model, in which the genetic variance is assumed to derive from a very large number of unlinked additive loci, which are treated together. A brief overview of the infinitesimal model, and how it can be extended to incorporate individual gene effects, is given in Chapter 2. Chapter 3 is an introduction to Gibbs sampling and its uses for genetic analysis. In Chapter 4, a description of the mouse lines used for the X Lines selection experiment is presented. The X Lines experiment forms the basis for much of the analysis described in this thesis. Also described are the genetic models used and the main computer programs

written for the analyses performed in the thesis. Chapter 5 presents the initial quantitative analysis of the X Lines selection experiment. The responses of body weight, litter size and sexual dimorphism to selection on body weight were described and estimates of variance components for 6-week weight were calculated. Chapter 6 describes a method for estimating changes of variance under selection using restricted maximum likelihood (REML), and its application to the analysis of the X Lines. A method for estimating the effects and positions of linked QTL fitting mixed inheritance models using Gibbs sampling is presented in Chapter 7. The method is illustrated by analysing several simulated datasets. Chapter 8 describes an analysis of the X Lines in order to obtain estimates of the effects and positions of the QTL linked to the coat colour loci *brown* and *dilute* using a simple approximate regression method as well as the Gibbs sampling based method presented in Chapter 7. Finally, a general discussion and conclusions arising from the thesis are presented in Chapter 9.

Chapter 2

The Infinitesimal Model

2.1 Introduction

In early experiments on quantitative characters, attempts were made to resolve the apparent contradiction between the particulate nature of Mendelian theory and the 'blending' inheritance apparently shown by quantitative traits (Emerson, 1910) and this lead to the idea that quantitative traits were controlled by a large number of Mendelian 'factors' producing a quasi-continuous distribution of phenotypes from a series of discrete genotypes. The formulations of models showing how quantitative traits could arise from multiple factors led to the development of the infinitesimal model (Fisher, 1918), which has now become the standard genetic model for quantitative traits. With the infinitesimal model a quantitative trait is assumed to be influenced by an infinite number of unlinked loci, each with an infinitely small additive effect on the trait. The genetic contribution to the trait, G, is produced by summing the contributions of each locus; producing a Gaussian distribution of genotypic values when the number of loci are infinite. The Gaussian distribution of G has advantages for statistical analysis under the infinitesimal model; this fact has contributed a great deal to the predominance of the infinitesimal model in quantitative analysis. Changing the infinitesimal model to make it more realistic often results in a large increase

in the mathematical complexity of analyses.

2.2 Detection of deviations from the infinitesimal model

It is important to be able to test any model in practice. The infinitesimal model can be used to make many predictions which should be tested in biological systems. If there is a divergence between what is observed in practice and the predictions from the infinitesimal model, then information on how the model failed can give clues about the underlying genetic basis of the trait being studied. An important source of information on the performance of the infinitesimal model in practice can be obtained from artificial selection experiments.

Selection experiments have been used for much of this century by researchers as a means of investigating responses and for estimating genetic parameters (Hill and Caballero, 1992). Early selection experiments were designed to investigate whether large, permanent changes to quantitative traits could be made by artificial selection (*e.g.*, Goodale, 1938; MacArthur, 1944a), and to check the validity of quantitative genetic theory by comparing the observed selection responses against those predicted by theory (Falconer, 1953). The analysis of long-term selection experiments can also be used to make inferences on the nature of the genetics underlying a quantitative trait. This last point is of great importance both for theoretical understanding of the genetic control of quantitative traits, and for practical concerns about the best methods for improving commercially important traits in agriculture (Roberts, 1965). Even though the details about the nature of the genes controlling a trait could well differ between species, the general methods for uncovering the control of quantitative characters will be likely to be applicable across species.

Selection experiments can act as checks of quantitative genetic theory in several ways. Genetic parameters can generally be estimated using several

different methods, for example heritability can be estimated from the covariance of collateral relatives, offspring-parent regressions and the response to selection. The different estimates should agree if the infinitesimal model of a large number of genes of equal effect is correct. The changes in genetic variance in populations can also provide information on the performance of theory. The infinitesimal model can be used to predict how the genetic variance should change under selection. If the actual changes in genetic variance can be estimated, then these estimates can be compared against the model predictions. Of course we cannot expect perfect agreement, it is not possible to measure parameters exactly so we end up comparing estimates. The difficulty is in deciding whether any observed discrepancies are statistically significant or not.

2.2.1 Comparing genetic parameter estimates

There have been several studies explicitly designed to test the infinitesimal model. Clayton *et al.* (1957) performed a comparison of the response to short-term selection (5 generations) on abdominal sternital bristles with base population estimates of genetic parameters in *Drosophila melanogaster*. Three estimates of the base population variance were obtained; from the half-sib and full-sib correlations and from the regression of offspring on parent. These three estimates were all very close to each other. The comparison with the mean results of short-term selection response were 'fair', but there was much variation between upwards selected replicates.

A comprehensive review of laboratory and farm animal selection experiments was given by Sheridan (1988) where again comparisons were made between base-population parameter estimates (*i.e.*, from half-sib correlation and offspring-parent regression) and estimates of realized heritability from the response to selection. In general the agreement was poor; a summary of the findings is given in Table 2.1 which shows the number of selection experiments which display a given level of agreement between the two estimates. The experiments are divided

into experiments on laboratory and commercial species, and according to how many generations the experiment was carried out over. The number in the boxes of Table 2.1 show the number and percentage of experiments that fall into each category. For example, 57% (40 studies) of laboratory experiments which were run for 6-10 generations had base population and realized heritability estimates which differed by more than 30%. Over all the experiments, 29% showed an

Table 2.1: Level of agreement between base population estimates and realized genetic parameters (Sheridan, 1988)

Type	Level of agreement	Number of generations			
		1-5	6-10	11-15	16+
Laboratory species	0-10%	6(33%)	16(23%)	24(52%)	-
	10-30%	1(6%)	14(20%)	5(11%)	-
	30%+	11(61%)	40(57%)	17(37%)	4(100%)
Commercial species	0-10%	2(8%)	7(25%)	3(50%)	-
	10-30%	11(42%)	3(11%)	-	-
	30%+	13(50%)	18(64%)	3(50%)	2(100%)
Total	0-10%	8(18%)	23(24%)	27(52%)	-
	10-30%	12(27%)	17(17%)	5(10%)	-
	30%+	24(55%)	58(59%)	20(38%)	6(100%)

agreement between the base population and realized estimates of under 10% and 54% showed a variation between the estimates of over 30%. This lack of agreement may not be as bad as it appears. It was noted by Hill and Caballero (1992) that although agreement would be expected to be highest for short-term experiments, in fact Table 2.1 shows better agreement for the longer term experiments. This may be due to more precise estimates of realized heritability with the longer term experiments. Hill and Caballero (1992) also comment that the sampling errors of the estimates was unlikely to have been known accurately for most experiments, so the significance of many of the discrepancies is unclear. James (1990) pointed out that Sheridan's (1988) study shows that 57% of the realized heritability estimates were smaller than the base population estimates, and 38% were larger. This would

indicate that much of the discrepancy was sampling error since most known factors would act to reduce the realized heritability below its expectation.

2.2.2 Detecting changes in variance

Another way of testing the performance of the infinitesimal model is to estimate the changes in genetic variance under selection. If the genetic variance in a population under selection could be estimated in different generations then the observed changes could be compared against the predictions of the infinitesimal model. Under the infinitesimal model, changes in genetic variance derive from two sources, inbreeding and the the build up of gametic disequilibrium under selection, the 'Bulmer effect' (Bulmer, 1974; 1976). Both of these effects cause the genetic variance to decrease. If the observed changes in variance cannot be accounted for by these effects then this is an indication that the infinitesimal model assumptions are incorrect. A possible cause for large changes in variance than cannot be explained by inbreeding or the Bulmer effect is if there are genes of large effect segregating in the population.

The simplest strategy for estimating the changes in genetic variance is to split the data into blocks of generations and obtain separate estimates of the genetic variance for each generation block. This can be done with or without accounting for the expected changes in variance under the infinitesimal model. The difficulty with using this approach for testing the infinitesimal model is that estimating the genetic variance requires quite large datasets to achieve any degree of accuracy. Splitting the dataset up to obtain separate estimates for different generations reduces the amount of data available for each estimate, producing a decrease in accuracy. It can therefore be difficult to pick up anything other than large deviations from the infinitesimal model using this general method. Since any discrepancies that do occur will be larger the longer the selection is carried out, it is best to carry out this sort of analysis on mid- to long- term selection experiments (> 10 generations).

Variance changes under selection have been estimated in several studies (*e.g.*, Rahnefeld *et al.*, 1963; Meyer and Hill, 1991; Beniwal *et al.*, 1992a), though each study used a different way to estimate the changes. The simplest way to perform the analysis is to consider the generation blocks as completely separate datasets and estimate variances for each block. An early analysis of this type was performed by Rahnefeld *et al.* (1963) on a selection experiment where a line of mice produced from the reciprocal cross of two (unspecified) inbred lines was selected for 17 generations on post-weaning growth. An attempt to quantify changes in additive variance over time was made by estimating the additive variance for each generation using a combined estimate from the sire component of variance and parent-offspring regression. The individual generation estimates were, however, extremely variable and neither linear nor quadratic regressions of the estimates on generation produced a significant trend, though they did indicate a slight increase in additive variance over time. The estimates from several generations were then pooled together to produce more reliable estimates, which are given in Table 2.2. These estimates indicate an increase followed by a decrease, but none

Table 2.2: Estimates of additive genetic variance by six-generation periods (Rahnefeld *et al.*, 1963)

Generations	Estimates for	
	♂	♀
1-6	0.30 ± 0.31	0.08 ± 0.24
7-12	0.75 ± 0.26	0.61 ± 0.20
13-18	0.54 ± 0.34	0.39 ± 0.18

of the differences are significant so no firm conclusions can be drawn from this. This analysis did not take into account the effects of inbreeding and selection on the later generation blocks, *i.e.*, the starting generation of each block was taken to be the base generation which is defined to be a random mating non-inbred population. The expected result under the infinitesimal model would be for the genetic variance to be smaller in the later generations due to the effects of inbreeding and the Bulmer effect. As stated before, there was no evidence for

any drop in additive variance, but the variance of the estimates were too large to conclude that there was a significant deviation from the infinitesimal model predictions.

The method of variance component estimation most commonly used now in animal breeding is REML, the theory of which was developed by Patterson and Thompson (1971). This allows for the estimation of variance components unbiased by selection or inbreeding. Including the numerator relationship matrix (the **A** matrix) accounts for the effect of drift and selection on the additive genetic variance under the infinitesimal model (Sørensen and Kennedy, 1983; 1984). It is possible to use REML methodology to perform essentially the same analysis as Rahnefeld *et al.* (1963) but accounting for the effect of inbreeding in the later generation blocks by including the pedigree back to the base population (Meyer and Hill, 1991). This cannot, however, account for the effects of selection because the data on which the selection decisions were made for the early generations are not in the analysis. The expected results from this analysis under the infinitesimal model would be that the genetic variance estimated in the later generation blocks should be slightly less than the variance in the base population (because of the Bulmer effect). Meyer and Hill (1991) analysed a selection experiment on mice where a population had been divergently selected for 23 generations on appetite. The traits food intake and 6 week body weight were analysed using the above method, and it was found that there was substantial decreases in the additive genetic variance over time. For example, the estimated heritabilities of food intake in generations 2-7, 8-13 and 14-23 were 0.24, 0.10 and 0.07 respectively. Meyer and Hill (1991) concluded that this decrease was too much to be explained by the Bulmer effect, especially since the effect of this should have been to reduce the variance in the early generations, rather than to produce the steady reduction that was observed in this experiment.

A more sophisticated analysis was performed by Beniwal *et al.* (1992a) where the variances in two separate blocks of generations were estimated simultaneously by separating the covariance matrix for the random effects into two containing the

contributions of the animals in each generation block to the additive variance. All of the data are used in the analysis so this method should properly account for both inbreeding and selection. In this case it would be expected under the infinitesimal model that the genetic variance estimated by this method would be the same in the early and late generation blocks. Any significant difference would indicate a failure of the model. The selection experiment analysed by Beniwal *et al.* (1992a) used mice derived from the same base population as that analysed by Meyer and Hill (1991), but in this case selection was on predicted 10 week lean mass in males. The analysis of lean mass, body weight and litter size all showed reductions in heritability in later generations (Beniwal *et al.*, 1992a; Beniwal *et al.*, 1992b). For example in the High lines the additive variance for lean mass decreased from $71g^2$ in generations 0-4 to $12g^2$ in generations 15-20. A similar decrease occurred in the low selected lines. It was concluded by Beniwal *et al.* (1992a) that the results indicate a failure of the infinitesimal model, the most likely reason being that some of the genes influencing lean mass had a non-negligible effect so that their gene frequencies altered under selection, causing the changes in variance.

2.3 Extensions to the basic infinitesimal model

There are several areas where the basic infinitesimal model can be extended to take account of various effects such as non-additive effects, mutation and major genes. The problem of all such modifications is that they make the use of the model for analysis purposes much more difficult.

2.3.1 Non-additive effects

The infinitesimal model can be extended to include non-additive gene action (Fisher, 1918). With directional dominance there can be a problem with the inbreeding depression becoming infinite unless the dominance variance is zero (Robertson and Hill, 1983). Several counter examples to this were demonstrated

by Smith and Mäki-Tanila (1990) who showed that some infinitesimal dominance models were indeed possible, although the examples they gave were somewhat contrived.

To add dominance and interaction effects to the model, G can be partitioned into the additive genetic value and the dominance and interaction (epistatic) deviations. The genotypic variance σ_g^2 can likewise be split into additive, dominance and interaction variances, as shown below:

$$\begin{array}{ccccccccc} G & = & A & + & D & + & I & & \\ \sigma_g^2 & = & \sigma_a^2 & + & \sigma_d^2 & + & \sigma_i^2 & & (2.1) \\ \text{(genotypic)} & & \text{(additive)} & & \text{(dominance)} & & \text{(interaction)} & & \end{array}$$

This model can itself be extended by splitting up the interaction effects into that due to interaction between the additive effects of loci (additive $\times$ additive), that between the dominance effects of loci (dominance $\times$ dominance), that between the additive effects of one loci with the dominance effect of other loci (additive $\times$ dominance) etc. To analyse data with such a model using mixed model methodology, it is necessary to specify the covariance matrices between observations for each effect, and the estimation of the variance components involves using the inverse of these covariance matrices. For the additive effect this is relatively simple to do; the inverse of the additive covariance matrix can be calculated directly (Henderson, 1976; Quaas, 1976) and is normally sparse, reducing storage and handling requirements. In contrast the covariances matrices for non-additive effects are normally much more difficult to calculate and handle; even so it should be possible to analyse such a model using standard mixed model methodology (Henderson, 1985; Smith and Mäki-Tanila, 1990).

2.3.2 Changes in genetic variance due to mutation

The standard infinitesimal model does not allow for the generation of 'new' genetic variance from selection. There have been several studies which report that the amount of new variance arising from mutation can be substantial (*e.g.*,

Lynch, 1988; Keightley and Hill, 1992; Hill *et al.*, 1994) and that this new variation can affect the response to artificial selection (Caballero *et al.*, 1991). Hill (1982) conducted a simulation experiment on the effect of the response to selection of new mutations and concluded that although the effects of mutations were unlikely to have much effect on the selection response in early generations, they should not be ignored in the analysis of long-term selection experiments.

A method to modify the additive genetic (co)variance matrix to model the increases in variance produced by the accumulation of new mutations was described by Wray (1990). It allows the inclusion of mutation effects with standard mixed-model methodology for use in estimating variance components etc. This method was used by Keightley and Hill (1992) to obtain maximum likelihood estimates of the increase in additive genetic variation arising in a inbred line of mice due to new mutations under an animal model. The technique probably has limited use in more general cases (*i.e.*, where the base population is not inbred) because the base population variance and the changes in variance due to causes such as changes in allele frequencies are likely to be much larger than the changes due to new mutations.

2.3.3 Mixed inheritance models

The problems described above with including non-additive effects into the genetic model of quantitative traits are much less than the difficulties encountered in trying to relax the central part of the infinitesimal model - that the genetic effect is caused by the combined effects of a very large number of loci. In producing a model with a small number of genes, the first problem is that there are many more parameters to consider such as the number of loci, the distribution of additive and dominance effects, the distribution of allele frequencies, the distribution of genotype frequencies, the disequilibrium between loci, and how the loci might interact. Such a model is not made any easier by the fact that several of these parameters will change over time depending on a range of factors such as the

method and intensity of selection and the population size and structure. Analysing data with such a model without many simplifications is much harder than using the infinitesimal model. Because of these difficulties, most non-infinitesimal models only involve one or two loci, but these are not appropriate for quantitative traits because they cannot generate the observed continuous distribution of trait values.

Mixed inheritance models can be regarded as a combination of infinitesimal and finite gene models. In the simplest mixed inheritance model an animal's genotypic value would be produced by the sum of the contribution from a single major gene or QTL and a residual infinitesimal (polygenic) effect. Mixed inheritance models were first introduced in human genetics (Elston and Stewart, 1971; Morton and Maclean, 1974) for the analysis of genetically determined diseases. Elston and Stewart (1971) described a general likelihood based method for the analysis of genetic data where the genetic contribution could be from a range of models from single loci up to a few major loci plus an infinitesimal effect. These models are more biologically realistic than either infinitesimal or finite gene models, but are more complex to use for data analysis than the basic infinitesimal model and, up until recently, it was not possible to use such models except with very small pedigrees or nuclear families (*e.g.*, Ott, 1979). The difficulties stem mainly from the use of likelihood based techniques for analysing data.

The problem with likelihood base techniques for mixed inheritance models is that usually the genotypes at the discrete loci are not completely known; this often makes the likelihood very difficult to calculate because the likelihood has to be calculated over all possible genotypic configurations of the pedigree. Throughout this thesis, $\mathcal{G}$ is used to represent a particular genotypic configuration. This means that $\mathcal{G}$ completely specifies the haplotypes for each individual at the discrete loci being modelled. Consider a model where there is a vector of observations $\boldsymbol{y}$ on a pedigree with genotypic configuration $\mathcal{G}$, and where $\boldsymbol{\theta}$ is a vector of model parameters (*i.e.*, gene effects, recombination frequencies etc.). The likelihood of

$\boldsymbol{\theta}$ for such a model is of the form:

$$L(\boldsymbol{\theta}) = \sum_{\mathcal{G}} p(\boldsymbol{y}|\mathcal{G}, \boldsymbol{\theta}) p(\mathcal{G}|\boldsymbol{\theta}), \tag{2.2}$$

The sum is over all possible configurations of $\mathcal{G}$ that are consistent with $\boldsymbol{y}$ and $\boldsymbol{\theta}$. $\mathcal{G}$ can refer to either a discrete genotype (*i.e.*, from a one locus genetic model), a continuous Gaussian genotype (*i.e.*, a breeding value from an infinitesimal model) or a combination of both. Obviously for any model the likelihood cannot be evaluated in the form (2.2) except for very small pedigrees. For the infinitesimal model the likelihood can be evaluated using matrices which is the method typically used in animal or plant breeding applications. An alternative method to calculate the likelihood for both infinitesimal and discrete genotype models is to use 'peeling' algorithms (*e.g.*, Cannings *et al.*, 1978). The basic idea behind these methods is to successively 'peel off' individuals from the pedigree, the information in their phenotypes being converted to a function of one or more of the remaining members of the pedigree. With mixed inheritance models, however, an animal's genotype is neither discrete nor Gaussian, but a mixture of both, making the likelihood 'intrinsically unpeelable' (Thompson and Guo, 1991). This means that for the mixed inheritance models the exact evaluation of the likelihood requires the summation over $\mathcal{G}$ (2.2) which is possible only for very small or very simple pedigrees. An example of a very simple pedigree would be an F_2 formed from a cross between two inbred lines. In this case the genotype of each individual is independent so the likelihood can be calculated for each individual separately, and then the likelihoods of all the individuals multiplied together to produce the total likelihood. Methods to approximate the mixed inheritance model likelihood have been suggested (Hasstedt, 1982; Bonney, 1984; Bonney *et al.*, 1988) and, in fact, are widely used in human genetics. It is not known, however, how well these approximate methods work with large or complex pedigrees because there is no exact method to act as a comparison.

An alternative approach to using exact calculation methods would be to use a sampling based technique to estimate the likelihood. The simplest way

this might be done would be to randomly generate $\mathcal{G}$, calculate $p(\boldsymbol{y}|\mathcal{G},\boldsymbol{\theta})$, the probability of the observations given $\mathcal{G}$ and $\boldsymbol{\theta}$, and calculate the average of these probabilities for a large number of random configurations of $\mathcal{G}$. This is, however, very inefficient. There will be a very large number of possible configurations of $\mathcal{G}$, only a subset of which will contribute substantially to the average likelihood. In addition it will often be the case that a large number of configurations of $\mathcal{G}$ will be incompatible with the observations so will have a probability of zero. The sampling technique can, however, be made to work using Markov Chain Monte Carlo (MCMC) sampling methods. With these it is possible to sample only from those configurations that are consistent with the observations, where the probability of sampling a particular configuration is proportional to the likelihood of that configuration; in effect sampling from the distribution of $\mathcal{G}$ conditional on $\boldsymbol{y}$ and $\boldsymbol{\theta}$. An estimate of the pedigree likelihood could then be obtained by averaging $p(\boldsymbol{y}|\mathcal{G},\boldsymbol{\theta})$ for each of the sampled configurations (Thompson and Guo, 1991). This is discussed more fully in Section 3.4.

As well as their contributions towards likelihood calculation, MCMC methods, and in particular Gibbs Sampling, have been used in various rôles in the analysis of mixed inheritance models (*e.g.*, Guo and Thompson, 1992; Janss *et al.*, 1994b; Guo and Thompson, 1994; Janss *et al.*, 1994a), and have been important in making such models practical for use with large datasets.

2.4 Conclusions

The infinitesimal model is widely regarded as an extreme simplification. Despite that, it is the most widely used model for analysing quantitative traits. This is due to two main factors, it is relatively straightforward to use and it provides a reasonable fit in the short-term to the observed data in many situations. The ease of use comes from being able to handle the loci collectively as a group rather than having to consider each locus individually. This is only feasible if the overall

contributions of the loci form some convenient distribution which can be easily manipulated. The assumption of a multivariate normal distribution of additive breeding values is a key feature in making the variety of statistical analyses avaliable for quantitative trait analysis (such as BLUP and REML) possible. It is fortunate that the model appears robust with regard to this assumption in that predictions derived from assuming normality are still quite accurate even when the distribution of breeding values is definitely non-normal due to linkage disequilibria produced by selection (Turelli and Barton, 1994). The conclusions of Turelli and Barton (1994) were that infinitesimal models can fairly accurately predict means and variance changes in the short term; it is only in the longer term that the infinitesimal models 'break down'. It appears that after 10 or more generations, changes in variance are likely to be dominated by changes in allele frequencies, which cannot be accommodated when using infinitesimal models.

The mixed inheritance model has potential for being a useable, more accurate version of the infinitesimal model. If we regard a truly accurate genetic model as explicitly considering all genes affecting the trait, with their individual effects, frequency dynamics and interaction, then the mixed inheritance model is part way between the infinitesimal models and the 'exact' model. Mixed inheritance models provide a means of moving to more accurate models as methods improve and computing power increases by increasing the number of individual loci modelled. Of course how much better mixed inheritance models are than the infinitesimal model depends on what the 'true' underlying genetic model really is. Probably the most important information required about the genes controlling quantitative traits is the distribution of effects. There is, however, very little direct information about this, although inferences can be made. Some selection experiments have shown decreases in variance under selection that are much greater than would be predicted from the infinitesimal model (Meyer and Hill, 1991; Beniwal *et al.*, 1992a; Beniwal *et al.*, 1992b). It is likely that these decreases are caused by changes in gene frequency which indicates that there must be some segregating genes with relatively large effects. There is information

available on the distribution of mutant effects; Mackay *et al.* (1992) obtained direct information on the distribution of mutation effects on bristle number and viability as a results of *P* element insertional mutagenesis. This showed a highly leptokurtic distribution of mutant effects for the traits with a large number of genes of very small effect and a much smaller number with large effects. Evidence for the distribution of mutation effects being highly leptokurtic was also presented in an analysis by Keightley (1994a). This is not the same, however, as the distribution of the effects of *segregating* genes affecting quantitative traits. The mutations of large effect tend to be deleterious, are rapidly eliminated and never reach high frequencies, and therefore contribute little to the genetic variance. An analysis of Mackay *et al.*'s (1992) study concluded that most of the genetic variance could be attributed to mutations with intermediate effects (between about 0.25 and 0.5 phenotypic standard deviations) (Caballero and Keightley, 1994). This class of mutations contributed more than both the larger effect but lower frequency mutants and the high frequency but small effect mutants. A tentative conclusion about the distribution of segregating gene effects from this would be that much of the genetic variance is provided by a relatively few genes of intermediate effect, with a small amount provided by a large number of genes with very small effects. This sort of distribution of gene effects could be modelled quite effectively by a mixed inheritance model, with the genes of intermediate effect being explicitly handled, and the smaller genes being lumped together as a residual infinitesimal effect.

Chapter 3

Gibbs Sampling

3.1 Introduction

The Gibbs Sampler is a class of Markov Chain Monte Carlo (MCMC) methods (Hastings, 1970) developed for the analysis of image-processing models by Geman and Geman (1984). It has become widely used in genetics, being used for example for estimating likelihoods of mixed inheritance (Thompson and Guo, 1991), variance component analysis (Wang *et al.*, 1993; Jensen *et al.*, 1994) and QTL detection and mapping (Hoeschele, 1994). It is a method for generating samples from the joint distribution of several random variables by sampling from the conditional distribution of each variable. It can be used in a wide range of problems where variables take on values from a small discrete set, or have parametric conditional distributions which can easily be sampled from (Neal, 1993). The method is useful when the joint distribution is complex and difficult or impossible to sample from directly and the conditional distributions are known and can be sampled from more easily.

3.2 The Gibbs sampling algorithm

Consider a joint distribution for $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ with distribution given by $\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n)$ then this can be sampled from using Gibbs sampling, provided the conditional distributions of all X_i given all other variables are known. The current values $\{x_1, \ldots, x_n\}$ of the variables are all updated one at a time from their conditional distributions. Therefore the current value x_j of the variable X_j would be updated with a value sampled from $\mathrm{P}\big(X_j = x_j | \{X_i = x_i : i \neq j\}\big)$, the conditional distribution of X_j given all of the other variables. This process produces a series of realizations of X that is a Markov chain because the influence of the realizations $X^{(1)}, \ldots, X^{(t)}$ on the distribution of $X^{(t+1)}$ is solely mediated through $X^{(t)}$. Note that throughout this chapter, subscripts are used to denote a particular variable in X and superscripts in brackets (*i.e.*, $X^{(t)}$) denote a particular realization of X.

The procedure for producing $X^{(t+1)}$, a new realization of X, from $X^{(t)}$ would be as given in the sequence below. Notice that the new value for X_i is immediatly used for the sampling of X_{i+1}.

- Sample $X_1^{(t+1)}$ from $\mathrm{P}(x_1^{(t+1)} | x_2^{(t)}, \ldots, x_n^{(t)})$
- Sample $X_2^{(t+1)}$ from $\mathrm{P}(x_2^{(t+1)} | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_n^{(t)})$
- Sample $X_i^{(t+1)}$ from $\mathrm{P}(x_i^{(t+1)} | x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_n^{(t)})$
- Sample $X_n^{(t+1)}$ from $\mathrm{P}(x_n^{(t+1)} | x_1^{(t+1)}, \ldots, x_{n-1}^{(t+1)})$

The variables do not have to be sampled in any consistent sequence as shown above; the choice of which variable to update could be made at random. It can be shown that under reasonably general conditions as $t \to \infty$, the density of $\{x_1, \ldots, x_n\}$ converges to the joint density of X (Geman and Geman, 1984; Casella and George, 1992), so for any t large enough, a realization may be regarded as a sample from the joint distribution. Moreover, considering any variable X_i individually, the samples of X_i will converge to its marginal distribution, *i.e.*, the

distribution of X_i averaged over $\{X_j : j \neq i\}$. For this to be true, it is necessary that that the Markov chain has an *invariant* (stationary) distribution which is the joint density of X as required, and that the Markov chain is *ergodic*, meaning that it will converge to the required stationary distribution independent of its starting distribution.

Intuitively it is clear that if $X^{(t)}$ is drawn from the joint distribution of X then Gibbs sampling will leave the distribution invariant, *i.e.*, if $X^{(t)}$ is drawn from the desired joint distribution of X then so will $X^{(t+1)}$. Firstly, sampling x_i leaves $\{x_j : j \neq i\}$ unchanged, so the marginal distributions for these will be invariant. Secondly, the conditional distribution of x_i given $\{x_j : j \neq i\}$ is defined to produce the desired marginal distribution for x_i. Together these ensure invariance of the Markov chain (Neal, 1993).

To ensure that the Markov chain is ergodic it is necessary for all transition probabilities to be non-zero (Neal, 1993). This means that in one full iteration of Gibbs sampling the Markov chain should be able to move to any state with a non-zero probability, this property is called *irreducibility*. If it is possible for a Markov chain to become 'stuck' in some subset of the sample space then the chain is not irreducible and the convergence property will not hold. This can cause problems in some situations, and will be discussed in more detail when the use of Gibbs sampling for pedigree analyses is covered in Section 3.4. The fact of convergence does not depend on the initial variable values, as long as the initial values in X are valid (*i.e.*, the joint density > 0), convergence should occur. If, however, the initial distribution is a long way from the equilibrium distribution then convergence could take a long time.

The variables do not have to be sampled individually, several may be sampled simultaneously from their joint conditional distributions. This may be advantageous if certain variables are highly correlated; the Gibbs sampler will work most efficiently when the sampling sub-units are independent (Neal, 1993). When two variables are highly correlated and are sampled independently it can slow down

the movement of the Markov chain through the parameter space since each of the two variables will be able to move only a small amount each time because they are constrained by the other variable.

3.3 Extracting information from the Gibbs sampler

3.3.1 Sampling from the Markov chain

When the Gibbs sampler has reached its equilibrium distribution then samples can be taken from it for various purposes. There are several practical problems involved with the sampling, namely how to tell when convergence to the equilibrium distribution has been achieved, and how often to sample. Although samples taken from each Gibbs cycle (after convergence) are all samples of the equilibrium distribution, there is some auto-correlation between adjacent samples. There are two methods to obtaining independent samples, the first way is to run the Gibbs sampler for t cycles, sample, and then restart the Gibbs sampler from a separate starting point. This is repeated n times. The second way is to run the Gibbs sampler for t times then sample, as before, and then continuing with the same Markov chain run the sampler for a further m cycles and sample. This last stage of m cycles and then sampling is then repeated until n samples have been taken. The advantage of the first method is that it prevents any correlation between subsequent samples, but if t has to be large to ensure convergence of the chain then the first sampling method will be very wasteful. The numbers of cycles required before convergence (t), the total number of samples required (n) and the number of cycles between samples (m) (if the second sampling method is used) all depend on the complexity of the system and the correlation between parameters as these determine how efficiently the chain samples the parameter space.

3.3.2 Parameter and density estimation

Samples of a single variable taken from the Gibbs sampler will converge to samples from the marginal distribution of that variable. If an estimate of the mean of the marginal distribution, for example, is required, then there are two natural ways of obtaining this. If the variable X_j has n samples $x_j^{(k)}, k = 1, \ldots, n$ then the two estimates of μ, the mean of the marginal distribution of X_j would be:

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} x_j^{(k)}, \tag{3.1}$$

$$\tilde{\mu} = \frac{1}{n} \sum_{k=1}^{n} E(x_j | \{x_i^{(k)} : i \neq j\}). \tag{3.2}$$

$\hat{\mu}$ (3.1) is called the empirical estimator and $\tilde{\mu}$ (3.2) is the mixture estimator. $\hat{\mu}$ is the expectation of X_j based on samples from the posterior marginal distribution, whereas $\tilde{\mu}$ is a mixture of complete data posterior means. It was shown by Gelfand and Smith (1990) that for independent samples of x_j then the variance of the mixture estimator will always be less than that of the empirical estimator. Their proof, however, did not apply to the case when the sample of x_j were dependently drawn from the Gibbs sampler, however Liu *et al.* (1994) proved that the superiority of the mixture estimator still applied in that case. Note that the above procedure can be applied to produce estimates of a function of a variable or group of variables.

Other estimates from the marginal posterior distribution of X_j can be produced. The above method could be used to obtain estimates of the mode, median or higher order statistics such as the variance of the marginal distribution. It is also possible to obtain estimates of the actual marginal posterior density itself. For example the mixture estimate of $p(X_j)$, the marginal posterior density of X_j would be obtained by:

$$\tilde{p}(X_j) = \frac{1}{n} \sum_{k=1}^{n} p(x_j | \{x_i^{(k)} : i \neq j\}). \tag{3.3}$$

Once estimates of the marginal densities of a variable or a function of several

variables has been produced, then useful statistics such as confidence limits can be produced by numerical means.

3.4 Applications of Gibbs sampling

It was stated in Section 3.2 that Gibbs sampling allowed sampling from the joint distribution of several random variables and to produce estimates of the marginal posterior densities of each variable. These two properties can be useful for a range of applications as detailed in the following sections.

3.4.1 Likelihood calculations for pedigrees:

As explained in Section 2.3.3, exact calculation of the likelihood for large pedigrees can be very difficult when using complex models such as mixed inheritance models. It is possible to estimate the likelihood using sampling based methods. For example, a large number of random simulations of the genotype structure of the pedigree could be produced. The average of the likelihood of the observed data given each of these would be an estimate of the overall likelihood of the pedigree data. While simulating on pedigrees is simple, simulating on pedigrees conditional on observed data is very difficult (Thompson, 1994). MCMC methods provide feasible ways of simulating genotypes; the use of such techniques was proposed by Lange and Matthysse (1989) and by Lange and Sobel (1991) who outlined a MCMC (but not Gibbs sampling) method of simulating di-allelic loci so that the genotype structure was consistent with observed data.

Gibbs sampling provides a simple means of sampling from all genotype configurations that are consistent with the observed data and model parameters, where the probability of sampling a configuration $\mathcal{G}$ is proportional to $p(\boldsymbol{y}|\mathcal{G},\boldsymbol{\theta})$, the probability of the observed data given $\mathcal{G}$ and $\boldsymbol{\theta}$, the vector of model parameters (such as recombination frequencies, variance components etc.). The use of Gibbs sampling for generating samples of $\mathcal{G}$ is quite straightforward. Consider a dataset

consisting of a number of animals related to various degrees, most of which have a series of observations. These observations, $\boldsymbol{y}$, might be on a continuous trait or genotypic data or a discrete phenotypic trait, or some combination of these. In this case $\mathcal{G}$ refers to the genotypes of the animals which, as discussed in Section 2.3.3 could be either a discrete genotype, a continuous Gaussian genotype or some combination of both. The likelihood of θ would be:

$$L(\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}, \mathcal{G}|\boldsymbol{\theta}) d\mathcal{G} = \int p(\boldsymbol{y}|\mathcal{G}, \boldsymbol{\theta}) p(\mathcal{G}|\boldsymbol{\theta}) d\mathcal{G} \tag{3.4}$$

Note that $\mathcal{G}$ could be discrete so the integration in (3.4) could be replaced by a summation. Gibbs sampling can be used to sample $\mathcal{G}$ from $\mathrm{P}(\mathcal{G}|\boldsymbol{\theta})$ so a Monte Carlo estimate of the $L(\boldsymbol{\theta})$ could be obtained simply by averaging the likelihood of $\mathcal{G}$ for a large number of realizations of $\mathcal{G}$ from the Gibbs sampler.

3.4.2 Sampling of genotypic configurations:

The implementation of the Gibbs sampler for sampling from $\mathcal{G}$ given a vector of observations $\boldsymbol{y}$ and additional parameters $\boldsymbol{\theta}$ requires the ability to sample from the conditional distribution of each individual's genotype given the genotypes of all other individuals, the observations on the individual and $\boldsymbol{\theta}$. This is made simple by the fact that the effect on the conditional distribution of an individual's genotype of the other individuals is mediated solely through the 'neighbours' in the pedigree, *i.e.*, parents, spouses and offspring.

Therefore, using the notation $\mathcal{G}_i$ as the genotype of individual i, $\mathcal{G}_{-i}$ as the genotypes of all other individuals, $\mathcal{G}_{\mathrm{par}}$ as the parental genotypes, $\mathcal{G}_{\mathrm{spouse}_{ij}}$ as the genotype of the jth spouse of i, $\mathcal{G}_{\mathrm{off}_{ijk}}$ as the genotype of the kth offspring of the jth spouse of i, and $\boldsymbol{y}_i$ as the vector of observations on individual i, the conditional distribution of $\mathcal{G}_i$ would be:

$$p(\mathcal{G}_i|\boldsymbol{y}_i, \mathcal{G}_{-i}, \boldsymbol{\theta}) \propto p(\mathcal{G}_i|\mathcal{G}_{\mathrm{par}}, \boldsymbol{\theta}) p(\boldsymbol{y}_i|\mathcal{G}_i, \boldsymbol{\theta}) \prod_{jk} p(\mathcal{G}_{\mathrm{off}_{ijk}}|\mathcal{G}_i, \mathcal{G}_{\mathrm{spouse}_{ij}}, \boldsymbol{\theta}) \tag{3.5}$$

One important point to note is that the Gibbs sampler can easily and 'naturally' handle missing data. If an animal had no observations then the subexpression

$p(\boldsymbol{y}_i|\mathcal{G}_i, \boldsymbol{\theta})$ from the likelihood equation (3.5) would just be set to 1. Similarly if there were no offspring then the subexpression $\prod_{jk} p(\mathcal{G}_{\text{off}_{ijk}}|\mathcal{G}_i, \mathcal{G}_{\text{spouse}_{ij}}, \boldsymbol{\theta})$ would be set to 1. In addition, missing data can easily be estimated from the remaining data, so if some individuals had genotype data but others did not the missing genotypes could be estimated (*e.g.*, Pong-Wong and Woolliams, 1994).

As discussed before in Section 3.2, one of the conditions for the Gibbs sampler to work correctly is that the generated Markov chain should be irreducible, *i.e.*, it should not be possible for the chain to become 'stuck' in a subset of the possible parameter space. This can potentially be a problem with pedigree analysis. It can be shown that with di-allelic loci that the condition holds and the Markov chain is irreducible (Lange and Matthysse, 1989; Lange and Sobel, 1991), but multi-allelic loci can cause the chain to become reducible (Sheehan and Thomas, 1993). It was proposed by Sheehan and Thomas (1993) that this could be avoided by assigning a small positive probability ρ to all zero transmission and zero penetrance probabilities. The resultant Markov chain therefore could generate both 'legal' (*i.e.*, consistent with the observed data) and illegal samples of $\mathcal{G}$. Importance sampling was then used to make inferences about the equilibrium distribution, with legal configurations assigned a weighting of 1 and illegal configurations a weighting of 0. There are practical problems with this method. If ρ is too small then moving from one irreducible step to another requires moving through states with very low probabilities. Conversely if ρ is too large then the augmented set of genotypic configurations with the modified transmission probabilities becomes much larger than the set of valid genotype configurations (Gilks *et al.*, 1993) so most of the sampled configurations are illegal and cannot be used.

Lin *et al.* (1993) showed that it was not necessary to modify *all* zero probabilities to achieve irreducibility, it was only necessary to modify the penetrance probabilities so that $p(d|g) > 0$ for all phenotypes d and all heterozygote genotypes g. Rather than using the importance sampling method of Sheehan and Thomas (1993), they propose using multiple Markov chains. One chain uses the true penetrance probabilities (so is not irreducible) and the other 'companion'

chains use the modified probabilities (and are therefore irreducible). These are set running together and then every kth cycle an attempt is made to switch between the standard Markov chain and one of the companion chains. If $P(\mathcal{G}|\boldsymbol{y})$ and $P'(\mathcal{G}|\boldsymbol{y})$ are the probabilities of genotype configuration $\mathcal{G}$ given the observed data $\boldsymbol{y}$ using the true and modified penetrance probabilities respectively, and if $\mathcal{G}$ and $\mathcal{G}'$ are the current genotype samples from the standard and companion chains respectively, then the chains are switched with probability:

$$r = \min\left\{\frac{P(\mathcal{G}'|\boldsymbol{y})P'(\mathcal{G}|\boldsymbol{y})}{P(\mathcal{G}|\boldsymbol{y})P'(\mathcal{G}'|\boldsymbol{y})}, 1\right\}, \tag{3.6}$$

i.e., the ratio between the product of the probability of the two chains after the potential switch and before the switch. This method is an improvement over that of Sheehan and Thomas (1993), but can still be inefficient with the companion chains spending a large amount of time in the illegal section of the sample space. Lin *et al.* (1993) noted that the sample space of $\mathcal{G}$ can be divided into one or more irreducible sets or islands with the modified penetrance probabilities providing bridges between the islands. Lin *et al.* (1994) proposed a method that explicitly identified all such islands and bridges thus minimizing the number of zero penetrance probability that have to be modified and reducing the size of the augmented sample space, making the previous method more efficient. An improvement in efficiency over the companion chain method was described by Lin (1995). This uses the method of Lin *et al.* (1994) to identify the bridges between noncommunicating islands, and switches between different genotype configurations for the bridging individuals with probability equal to the ratio of the conditional probability of the new configuration to the conditional probability of the old configuration.

3.4.3 Bayesian estimates of parameter distributions:

Statistical analysis can be broadly split into two approaches, 'classical' and Bayesian. In classical analyses, the information used for parameter estimation is

assumed to come solely from the data, whilst Bayesian estimation combines prior information (that was known before the analysis) with information from the data. In some respects there is not too much difference in the approaches; in classical analysis some implicit prior information is generally used, *i.e.*, with regard to parameter distributions - the difference in a Bayesian analysis is that the use of the priors is explicit. It is possible to fit priors that are 'naive' so that the estimation is the same as in a classical analysis in that all the information comes from the data. A further difference in the approaches is that classical methods generally produce point estimates of parameters while Bayesian approaches produce estimates of the posterior distributions of the parameters.

The use of Gibbs sampling for parameter estimation is often done from a Bayesian perspective. One reason for this is because the formulation of the conditional probabilities forces the explicit declarations of the prior distributions used. Another reason is that Bayesian analysis on complex models is often impossible to do analytically, so Gibbs sampling provides a means for performing analyses that otherwise could not be done. This is because the typical Bayesian approach to estimating a parameter would be to formulate an expression for the joint posterior density of all the parameters, and then integrate out all parameters apart from the one of interest. This integration can be unfeasible to perform analytically, but can be performed numerically using Gibbs sampling.

3.4.4 QTL detection:

Gibbs sampling has been used for Bayesian linkage analysis by several authors (*e.g.*, Thompson and Guo, 1991; Hoeschele, 1994; Thompson, 1994). The technique is to use Gibbs sampling to estimate the likelihood ratios between different models *i.e.*, fitting a linked versus and unlinked QTL. In terms of the model discussed in the previous section on pedigree likelihood calculation, this could be done to compare a model with parameters $\boldsymbol{\theta}$ against an alternative model with parameters $\boldsymbol{\theta}_a$ by estimating the ratio $L(\boldsymbol{\theta})/L(\boldsymbol{\theta}_a)$.

3.4.5 Variance component estimation:

There are two ways in which Gibbs sampling has been used in variance component estimation. The first way is in conjunction with more 'traditional' methods of estimation such as the EM algorithm, where Gibbs sampling is used to estimate some of the conditional expectation terms (Thompson and Shaw, 1990; Guo and Thompson, 1991). This is useful because the calculation of these expectations otherwise requires the inversion of large matrices. The second way is to use a more Bayesian approach (*e.g.*, Wang *et al.*, 1993; Sørensen *et al.*, 1994a); most of these are based on the work by Gianola and Foulley (1990) where a Bayesian framework for variance component analysis called VEIL (Variance Estimation by Integrated Likelihoods) was presented. In that paper it was noted that this analysis was very difficult to do exactly and so several approximations were presented. Their method is, however, very well suited for the use of Gibbs sampling because it involves producing estimates of the marginal densities of the variance components by integrating out all of the other model parameters. As discussed earlier, these integrations can be extremely difficult to do analytically, but are easily handled numerically using Gibbs sampling. For datasets smaller than about 100000 animals the Gibbs sampling approach has few practical advantages over more conventional methods such as REML, though there are some theoretical arguments over the relative efficiency of VEIL and REML for variance component estimation (Gianola and Foulley, 1990). The Gibbs sampling method is simpler to program and typically requires less computer memory than an equivalent REML analysis, but can take much longer to obtain parameter estimates. As dataset sizes increase the relative speeds of the two processes are likely to converge since the time required of the Gibbs sampling approach will increase more or less linearly with numbers of animals, whereas for a process such as DFREML (Meyer, 1988) the increase will be more like cubic in the number of animals. An analysis of a large dataset of over 400000 animals by van der Lugt *et al.* (1994) indicates that Gibbs sampling could be very useful for analysing large datasets. The larger the dataset,

the less the speed advantage of DFREML becomes, and the more attractive the Gibbs base approach appears given that it requires much less computer memory for an equivalent analysis.

3.4.6 Mixed inheritance models:

With mixed inheritance models there are the same two approaches to using Gibbs sampling as mentioned for variance component analysis, using Gibbs sampling to estimate conditional expectations for the EM algorithm (Guo and Thompson, 1992; Guo and Thompson, 1994), and using Gibbs sampling to implement a Bayesian analysis (Janss *et al.*, 1994a; 1994b). In either case the basic method is very similar to that used for the equivalent variance component analysis without the QTL. The difference in the model is that the QTL genotype is fitted as an additional fixed effect, with the incidence matrix for the QTL genotype being treated as missing data to be estimated from whatever information is available (*i.e.*, phenotypic information, genetic markers etc.). As explained earlier, the Gibbs sampling lends itself easily to missing data problems. Using these methods it is possible to simultaneously detect QTL, estimate their effect and estimate the recombination frequencies between the QTL and available markers, potentially making these techniques powerful tools for the study of quantitative traits. The main drawback of using Gibbs sampling for these types of analyses is the speed, adding a QTL to a variance component analysis dramatically slows down both the Gibbs cycle time and the convergence time.

3.5 Conclusions

Gibbs sampling is a powerful data analysis tool. Its statistical properties which make it a simple to use Monte Carlo integration technique give it a large range of applications in data analysis, as outlined in the previous section (3.4). Its main strengths lie in the simplicity of both the concept and the implementation, and

its flexibility in being able to handle a wide range of types of problems. The main drawback with the technique is in its speed; for many problems it is not the fastest or most efficient method (Neal, 1993). There is therefore a tradeoff between the simplicity and flexibility of Gibbs sampling and the increased efficiency and speed of more specialized methods. As computing power continues to increase it seems likely that the advantages of Gibbs sampling will cause it be used in an ever wider spread of applications.

Chapter 4

Materials & Methods

4.1 The X Lines selection experiment

4.1.1 Introduction

There is a long history of selection experiments on mice. They have provided checks on quantitative genetic theory and clues about the underlying genetic control of quantitative traits (Section 2.2). The earliest selection experiment involving mice which included details on the origin of the base population and the selection process was described by MacArthur (1944a), where a genetically diverse stock formed from seven strains of mice was selected for large and small 60 day body size over 23 generations. The foundation stocks were set up to try to incorporate as much genetic variation as possible in a variety of qualitative and quantitative traits. The aim of the experiment was to investigate whether large, permanent changes could be made to quantitative traits using artificial selection and to look for other characters which showed correlated responses to selection on body weight (MacArthur, 1944b). The experiment demonstrated that artificial selection could indeed produce substantial permanent changes to body weight, and also showed that several other traits such as litter size and coat colour displayed correlated responses to selection on body weight.

Falconer (1953) described a later selection experiment in which divergent within family selection was carried out over 11 generations on an foundation population formed from the cross between 4 inbred mouse lines. The within family selection reduced inbreeding and simplified the analysis since there could be no selection for maternal effects. This experiment to a large extent confirmed the results of MacArthur's (1944a) earlier experiment in that substantial and permanent changes to populations could be made by selection. The Q Line experiment of Falconer (1973) was of a similar design to the previous experiment except that the foundation population was made up from five foundation strains in order to incorporate a large amount of genetic variance into the initial population in a similar way to MacArthur (1944a). Another difference between the Q Line experiment and the earlier experiments described here is that the Q Line was replicated with 6 replicate lines being set up in each of three groups. These were six High lines selected upwards on 6 week weight, six Low lines selected downwards on 6 week weight, and 6 non-selected Control lines. The replicate structure made it possible to obtain empirical estimates of the sampling error of variance component estimates by examining the variability amongst estimates obtained from the different replicate lines.

Subsequent to the Q Line experiment, Garnett and Falconer (1975) searched for variation between the High and Low selection lines at 9 loci. They found an indication that one allele, *Hbb*, was associated with body weight since *Hbb* was found to be fixed in all 6 High selected lines. The analysis was made difficult by the fact that the genetic structure of the base population was not known because a large mixture of lines were incorporated into it.

The X Lines selection experiment described in this thesis was set up by G. Bulfield (Roslin Institute, Edinburgh), and followed on from the analysis of Garnett and Falconer (1975) specifically for the identification of QTL affecting body weight in mice. To make the analysis more powerful, the X Lines were established from an inbred cross so that the genetic structure of the base population was known precisely. It could be assumed that any loci differing between the

two founder strains would be at frequency 0.5 in the base population. Associations between any of these loci and body weight could be tested by screening for loci which changed in frequency significantly from 0.5 under selection for weight, making the experiment a potentially very powerful screening tool for QTL.

4.1.2 Inbred strains

The two inbred founder strains, C57BL/6J and DBA/2J, were obtained from the Jackson Laboratory, Maine USA in 1985. Both of these strains had already been established for a long time when the X Lines experiment was started; C57BL was set up in 1921 with substrain 6 being formed in 1937 and DBA was initially set up in 1909 with substrain 2 being formed from crosses between the original substrains in 1929-30 (Festing, 1989). The two strains had therefore been separated at least since 1909, and it is possible that they were in fact derived from separate subspecies (Bonhomme *et al.*, 1987). It would be expected that there would be a large amount of genetic differences between the two strains, and this has been demonstrated by the identification of a wide range of molecular markers which vary between C57BL/6J and DBA/2J (*e.g.*, Frankel *et al.*, 1990). Despite this genetic variation the strains show very little difference in mean body weight.

4.1.3 Selection lines

From the F_1 of the cross between C57BL/6J and DBA/2J, 32 breeding pairs were selected at random. From the resulting F_2 population 13 selection lines were established. These were divergently selected for 6 week body weight for 20 generations with 6 lines being selected upwards, 6 lines downwards, and with 1 unselected control line. In total there were 6503 animals in the Low lines, 8401 in the High lines and 1208 in the Control line. Selection was on a within family basis; each line was maintained with 8 breeding pairs, and 1 from each family of each sex were selected. The mating schedule is shown in Table 4.1; this is the

same schedule that Falconer used for the Q Lines. Second litters were raised in many cases to act as replacements.

Table 4.1: Mating schedule in the X Lines

Family of origin ♀ ♂	New mating number
1 × 2	1
3 × 4	2
5 × 6	3
7 × 8	4
2 × 1	5
4 × 3	6
6 × 5	7
8 × 7	8

Each family was numbered from 1-8. The table shows how the families of the mice to be mated were chosen and the new mating numbers assigned.

All animals from the F_2 population up to generation 20 were recorded for 6 week body weight and coat colour. There were two coat colour markers segregating in the F_2, *brown* and *dilute*. Both markers were recessive so only two marker types could be distinguished for each locus. The two markers acted independently so four marker type combinations were possible producing four distinct coat colours, wildtype, brown, dilute and brown & dilute. In addition, tissue samples were taken from 93 individuals from the Low selected lines and 34 from the High selected lines at the end of the experiment (generation 21). Unfortunately due to a procedural error in the matings at generation 20 the parents of the sampled individuals from the High lines are not known and it is not certain from which replicate each High line sample was drawn, although a reconstruction of the replicate structure of the High lines using information from the marker frequencies was described by Keightley (1994b).

4.2 Genetic models

Throughout this thesis several models are used for analysis, simulations and theoretical calculations. This section describes the main models used so that the details of the models do not need to be explained each time they occur.

- *Monogenic model*: The simplest genetic model of inheritance involves just a single gene. The effect of the gene is described by a and d, the additive and dominance effects of the gene. As populations derived from inbred crosses were analysed in this thesis, only di-allelic models are discussed with the allele with a positive effect being denoted by an uppercase letter, and the the other allele by a lower case letter. In general the letter M is used for marker loci, Q for QTL loci and A for general loci. The frequency of A will be denoted by p, and of a by q, therefore the genotypes, frequencies and genotypic effects relative to the mean are as given in Table 4.2.

Table 4.2: Genotype frequencies and values for simple model

Genotypes	AA	Aa	aa
Frequency	p^2	$2pq$	q^2
Genotypic value	$-a$	d	a

- *Linked QTL model*: This model has two linked loci, a QTL and a marker with no intrinsic effect on the trait. The recombination frequency between the marker and the QTL is given by r. The QTL effect is given by a and d as in the previous model.

- *Polygenic model*: The polygenic models used are of a large number (typically several thousand) unlinked additive loci of equal effect to simulate an infinitesimal model. The additive genetic variance produced is given by σ_a^2.

- *Mixed inheritance model*: The mixed inheritance model is one in which the genetic effect comes form two sources, one or more QTL and a polygenic

background. The QTL and polygenic effects are added together to produce the genetic value of an individual. There is assumed to be no interaction between the polygenic and QTL effects. For more details see Section 2.3.3

4.3 Computer programs

Most of the investigations and analyses detailed in this thesis were computer based involving an array of different programs. These were either commercial *i.e.*, Genstat (Genstat 5 Committee, 1993), free software written by others *i.e.*, DFREML (Meyer, 1988), or written by the author. This section briefly describes two of the main pieces of software written by the author.

4.3.1 Simulation

A lot of testing of the methods developed in this thesis was performed by simulating data structures similar to the X Lines dataset with a variety of different genetic models. For this purpose a general simulation program was written to simulate selection experiments. The genetic variation came from two sources, a polygenic additive effect and one or more QTL. Markers with no intrinsic effect could be positioned at various distances from the major genes. The polygenic effect was produced by a user-specified number of unlinked additive loci of equal effect. In addition a common environment (litter) variance could be simulated. The program could therefore produce datasets similar to that of the X Lines dataset using polygenic, monogenic and mixed inheritance genetic models containing pedigree, trait and marker (but not the actual QTL) information.

4.3.2 Gibbs sampling analysis

For the analyses in Chapters 7 and 8, a suite of programs were written. These were intended to be used in a similar manner to the DFREML suite (Meyer, 1988)

except that they use Gibbs Sampling techniques rather than Mixed Model Methodology, and have the ability to analyse mixed inheritance models.

There are four programs in the suite:

- `recode` handles the recoding and processing of the data into a suitable form for the other programs. Any number of fixed and uncorrelated random effects can be fitted, the only restriction being the amount of memory available on the system. Any number of QTL and markers can be fitted, though in practice the number of QTL should not be too large, and the method has not been thoroughly tested with more than 1 QTL.

- `nrm` generates the inverse of the additive genetic relationship matrix using the methods of Henderson (1976) and Quaas (1976). The covariance matrix of the fixed effects, $\mathbf{X'X}$, is generated and its Cholesky decomposition calculated. This is used to speed up the sampling of the fixed effects during the sampling process. If very large numbers of fixed effect levels were to be fitted it would be better to omit this step since the decomposition of $\mathbf{X'X}$ is non-sparse so can require a lot of storage.

- `gibbs` is the main analysis program which reads in the outputs from the other programs and produces a continuous string of Gibbs realizations and conditional densities which stored. The program has the ability to be restarted if it is interrupted, which is valuable since for complex problems it may have to run for several weeks.

- `density` reads the output from the gibbs program and calculates estimates of the marginal densities of any parameters of interest. Estimates of the mean, mode and median and standard distribution of the distributions are calculated.

Chapter 5

The Initial Quantitative Analysis of the X Line Data

5.1 Introduction

This chapter describes the initial quantitative analysis of the X-Line experiment. The aim of the analysis was to describe the responses of 6-week body weight and litter size to selection on 6-week weight, and to produce estimates of the genetic components of 6-week weight. Estimates of the heritability of 6-week weight were calculated from the regression of cumulative selection differential on response (the 'realized' heritability) and by using the Derivative Free Restricted Maximum Likelihood package (DFREML) of Meyer (1988), allowing a comparison to be made between the different estimates. An additional aim was to describe any unusual or unexpected features of the X Lines data which became apparent during the analysis. One of these features in particular, an apparent acceleration in the rate of response over time in the Low lines, lead to the analysis described in Chapter 6.

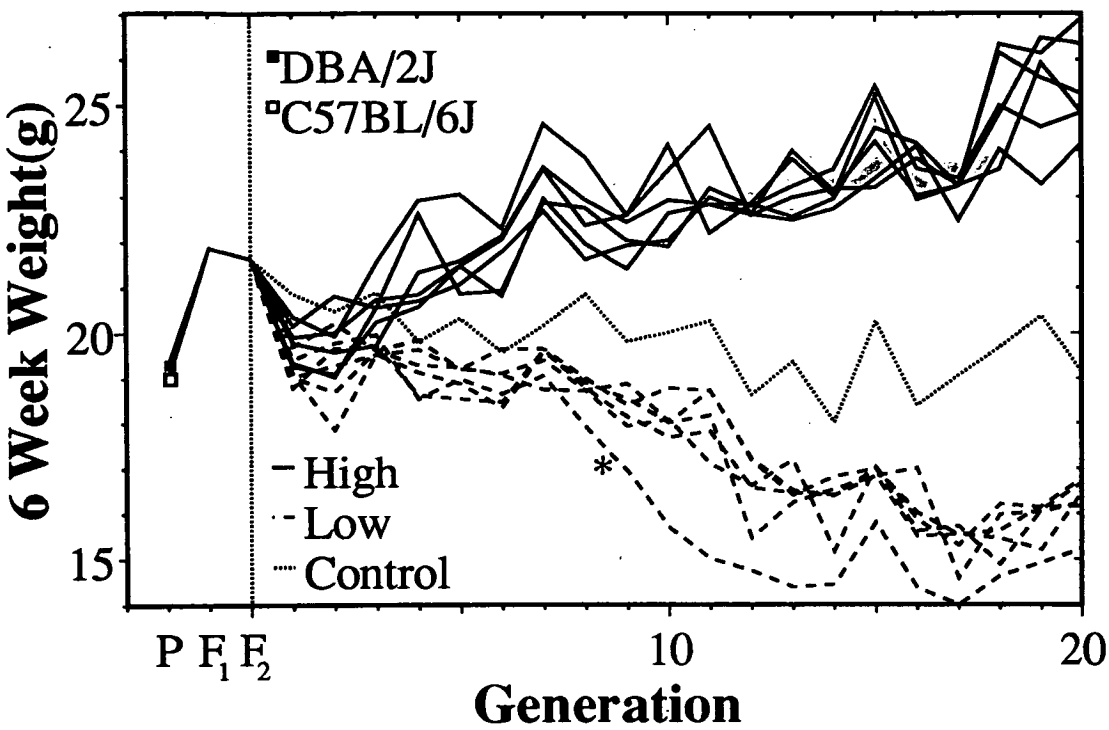


Figure 5.1: Mean 6-week weights (g) averaged over sex of the the 13 replicate lines (6 High, 6 Low and 1 Control). Subline B of the Low lines (referred to in the text) is marked by a *.

5.2 Selection response

The selection responses in 6 week weight are shown in Table 5.1 and Figure 5.1. The results shown are the average weight for each replicate line within generation. There is a strong indication of directional dominance as the F_1 mean weight is well above the average of the parental lines. The response in the High and Low lines is broadly similar, but this does not rule out dominance being present. The phenotypic standard deviation of 6 week body weight is approximately 2g so the total response is about 2 standard deviations in both directions.

Table 5.1: Selection responses

	Rep.	No. Animals	Mean G_1^a	Mean G_{20}^b
Low	1	1199	18.67	15.72
	2	986	18.93	14.69
	3	1059	18.50	16.28
	4	1065	19.21	15.59
	5	1143	18.26	15.87
	6	1051	18.57	16.26
		6503	**18.64**	**15.74**
High	1	1380	19.42	25.83
	2	1548	19.67	23.68
	3	1395	19.86	24.38
	4	1448	18.78	24.72
	5	1368	18.84	24.40
	6	1262	19.29	26.38
		8401	**19.39**	**24.51**
Control		**1208**	**20.34**	**18.70**

†Means calculated from the uncorrected 6 week weights of each mouse averaged over sexes within replicate lines.

[a] G_1 After 1 generation of selection.

[b] G_{20} After 20 generation of selection.

Scale effects, when the variance changes with the mean, are often found in growth data. A quick test for scale effects in the X Lines was carried out by calculating the means, within line standard deviations and coefficients of variance for the last 5 generations of the High and Low lines (Table 5.2). The results are presented before and after $\log_e$ transformation of the data. The coefficients of variance for the untransformed data are around 14%, this reduces to approximately 5% after transformation. There appears to be a difference between the High and Low line standard deviations using the untransformed data. An F-test comparing the High and Low line variances indicates a significant difference for the untransformed data ($F_{2.573,2204,1587} < 0.001$). Log transformation reduces this heterogeneity ($F_{1.07,2204,1587} = 0.07$), so that the difference between the High and Low lines is no longer significant. For this reason, subsequent analysis on

the X Line data in this chapter was carried out on the transformed data, unless otherwise stated.

Table 5.2: Means, within line standard deviations and coefficients of variance for 6-week body weight calculated from the last 5 generations of the X Line data.

	Untransformed		Log_e transformed	
	Low Lines	High Lines	Low Lines	High Lines
Mean(g)	24.3	15.9	3.18	2.75
Standard Deviation(g)	3.56	2.22	0.147	0.142
Coeff. of variation(%)	14.6	14.0	4.6	5.2

Subline B of the Low lines (marked * in Figure 5.1) shows a markedly different response from the other Low lines, being at one point (around generation 10) about 1 s.d. lower than the other Low lines. Although the difference between subline B and the other lines reduced later on, at the end of the experiment there is still a clear difference between this subline and the others. It therefore seems probable that the effect is caused by a rare event which occurred only in that line such as a mutation or a rare recombination.

A further notable feature about Figure 5.1 is that while the response for the High lines appears to be almost linear over time, the Low lines show a sigmoidal response curve with little response until generation 7, increasing rapidly until generation 14 and then slowing down again for the remainder of the experiment. Linear regressions were fitted to the log_e transformed line means separately for generation ranges 0-4, 5-9, 10-14 and 15-20 for the High and Low lines to show how the response changed over time. The results of this are shown in Table 5.3. The High lines respond strongly at the beginning and end of the experiment, but show little response in the middle section. In contrast the Low lines show little response at the beginning and end of the experiment, but respond strongly in the middle stages. It is not clear what caused this change in response, although the changes in variance discussed in section 6 are likely to be connected with it. This will be discussed more fully in Section 6.

Table 5.3: Linear regression coefficients of the log of the selection response on generation, fitted to generations 0-4, 5-9, 10-14, 15-20 separately.

Line	Generation Range	Regression Coefficient	S.E.
High	0-4	0.0269	0.0114
	5-9	0.0099	0.0081
	10-14	0.0023	0.0081
	15-20	0.0136	0.0061
Low	0-4	0.0021	0.0130
	5-9	-0.0102	0.0092
	10-14	-0.0273	0.0092
	15-20	-0.0021	0.0069

Response calculated from the mean body weight averaged within line (High, Low & Control), generation and sex. The regression was calculated using the natural log of the response. S.E. = the standard errors of the regression coefficients.

5.3 Sexual dimorphism

If males and females are considered separately, differences in the selection responses are seen. In the control lines, both sexes show a slight decrease in body weight over the course of the experiment (a further indication of directional dominance). In the selected lines, the males show a roughly equal response in both directions, but the females show significantly less response to downwards selection. The two sexes were therefore converging since the males weighed more than the females. This effect was still apparent with log transformed data. Sexual dimorphism (measured as the ratio of male to female weights) decreased over the course of the experiment in the Low selected lines (Figure 5.2) as shown by the regression of male/female weight on generation, but did not change much in the other two lines (Table 5.4). Note that the regression was conducted assuming that errors between generations were not correlated, which is not the case. The t-ratios quoted in Table 5.4 are therefore likely to biased upwards.

A similar, although less strong, effect was reported by MacArthur (1944a) from a selection experiment on mice where in low selected lines the response in

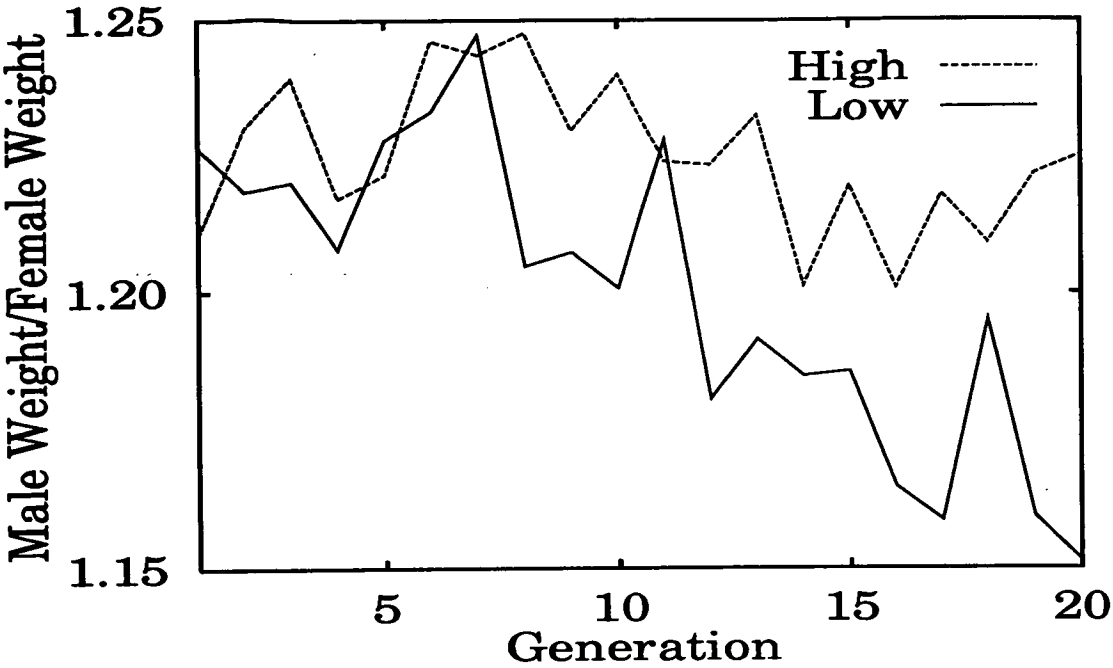


Figure 5.2: Sexual dimorphism measured by the ratio of male mean weight to female mean weight

females was proportionally less than that in males. A possible explanation for this effect is that fertility in females may depend more strongly on body weight than it does in males due to generally higher physiological costs of reproduction in females. This could cause the females to respond less strongly to downwards selection than males if, for example, the smallest females were unable to produce offspring.

Another possible scenario to explain the effect is if the selection intensity in females was less than that in the males. This could arise if there were on average less females in a litter than males. If the selection intensities are examined, however, it can be seen that there is little difference between the males and females for either High or Low selected lines (Table 5.5). The selection intensity for the High lines is larger than that for the Low lines, this is due to the larger litter sizes in the High lines (Section 5.4).

Table 5.4: Linear regression of sexual dimorphism (male weight / female weight)† on generation fitted for High, Low and Control lines separately.

Line	Regression Coefficient	t-ratio
High	-0.000787	−1.58
Low	-0.003370	−6.78*
Control	-0.000009	−0.01

†Weights are the mean body weights for each sex averaged within line (High, Low & Control) and generation. * = significant at the 0.1% level; the remaining t-values were non-significant at the 5% level. The t-ratios were calculated assuming no correlation between errors. This would not be the case, so the t-ratios are likely to be biased upwards.

5.4 Litter Size

The mean litter size at weaning is shown in Figure 5.3. It can be seen that the average litter size in the High lines stays more or less constant at around 8, but in the Low and Control lines it decreases to around 5 and 6 respectively. The response of the Control lines indicates that at least some of the reduction could be due to inbreeding depression, although there is really too little data from the Control lines to allow any firm conclusions to be made. It seems probable that there is a positive correlation between body weight and litter size, but the reduction in litter size due to inbreeding may mask the expected correlated response in litter size to selection on body weight.

Table 5.5: Average selection intensities calculated separately for males and females in the High and Low selected lines.

	♂	♀
High	1.05	1.05
Low	0.86	0.83

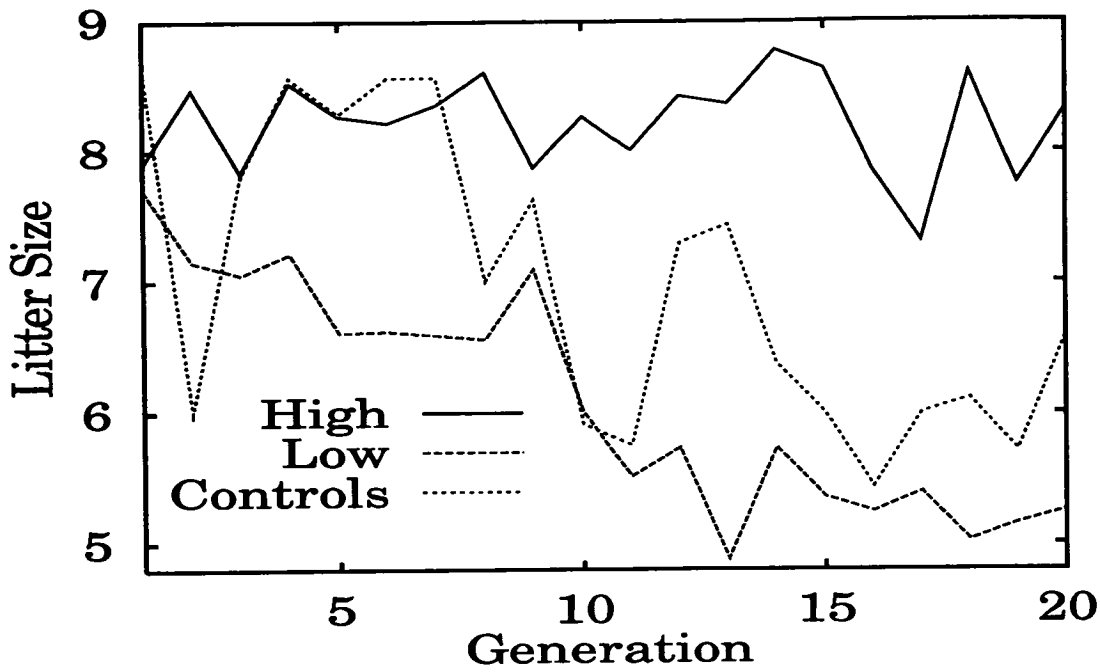


Figure 5.3: Mean litter size at weaning for the High, Low and Control lines

5.5 Realized heritabilities

Estimates of realized within family heritability were calculated from the regression of cumulative selection differential on response using the divergence between pairs of Low and High selected lines. A pooled estimate was also obtained using the means of all replicate lines. This allowed the comparison of the estimated standard error from the pooled regression analysis, and the empirical standard error obtained from the replicate estimates. Standard errors from the regression analyses were estimated assuming independence of the errors of the observations. The selection differentials were calculated from the mean within-sex within-litter deviations using the untransformed data. Realized within family heritabilities (Table 5.6) give a mean estimate from all replicates of 0.2 with an empirical standard error of 0.008. The pooled estimate (achieved by analysing all replicates together) was again 0.2 with the standard error of the regression being 0.007. This is slightly lower than the standard error of the mean estimate, but underestimates

Table 5.6: Realized heritabilities calculated from the regression of response against cumulated selection differentials for the divergence between pairs of lines.

Rep. Pair	*b*	*S.E.*
A	0.191	0.016
B	0.215	0.021
C	0.201	0.018
D	0.230	0.014
E	0.219	0.017
F	0.242	0.014
Pooled	0.220	0.0074
Mean†	**0.216†**	**0.0077†**

b = regression coefficient. S.E. = standard error of regression except where marked †. †Arithmetic mean of regression coefficients among replicates with empirical standard error.

the standard error of the heritability estimate since it assumes independence and homogeneity of the residuals (Hill, 1972).

5.6 REML analysis of heritability of 6-week weight

Further analysis of the heritability of body weight was undertaken using the derivative-free REML packages of Meyer(1988; 1989). This has the advantage over the realized heritability analysis described earlier of using information from the covariance between relatives as well as from the selection response to obtain an estimate of heritability. By including the numerator relationship matrix into the REML equations, the method can also account for the expected loss in additive genetic variance due to inbreeding and selection (Sørensen and Kennedy, 1983; 1984). An animal model was fitted to the data with generation, sex nested within line and generation, parity and litter size as fixed effects, and litter as an additional random effect uncorrelated with the main random effect. Sex was fitted as a nested effect because of the change in sexual dimorphism in the Low

Table 5.7: REML estimates of variance components and genetic parameters using log transformed data.

Var. Components $\times 10^{-3}$					
$\hat{\sigma}_a^2$ (s.e.)	$\hat{\sigma}_c^2$ (s.e.)	$\hat{\sigma}_e^2$ (s.e.)	$\hat{h}^2$ (s.e.)	$\hat{c}^2$ (s.e.)	$\hat{e}^2$ (s.e.)
3.31 (0.15)	4.38 (0.18)	4.49 -	0.27 (0.01)	0.36 (0.01)	0.37 -

lines over the course of the experiment discussed previously (section 5.3). The model was

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\lambda} + \boldsymbol{e} \tag{5.1}$$

where $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{\alpha}$ is the vector of fixed effects, $\boldsymbol{\beta}$ is the vector of additive genetic values, $\boldsymbol{\lambda}$ is the vector of litter effects, and $\boldsymbol{e}$ is the vector of environmental effects; $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$, and $\boldsymbol{e}$ have mean zero and are uncorrelated. $\text{Var}(\boldsymbol{\beta}) = \mathbf{A}\sigma_a^2$ where $\mathbf{A}$ is the numerator relationship matrix and σ_a^2 is the initial additive genetic variance, $\text{Var}(\boldsymbol{\lambda}) = \mathbf{I}\sigma_c^2$ where $\mathbf{I}$ is the identity matrix and σ_c^2 is the litter variance, and $\text{Var}(\boldsymbol{e}) = \mathbf{I}\sigma_e^2$ where σ_e^2 is the environmental variance. $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}$ are incidence matrices. Phenotypic variance $= \sigma_p^2 = \sigma_a^2 + \sigma_c^2 + \sigma_e^2$, so the heritability $= h^2 = \sigma_a^2/\sigma_p^2$, and the litter or 'c-squared' coefficient $= c^2 = \sigma_c^2/\sigma_p^2$. All REML analyses used $\log_e$ transformed data to account for changes in variance due to differences in means between Low and High selected lines as discussed earlier. Standard errors of σ_a^2 and σ_e^2 were estimated using the second derivatives of a polynomial approximation to the joint likelihood function of σ_a^2 and σ_e^2.

The results of the REML analysis are shown in Table 5.7. The heritability estimate is larger than the mean realized estimate and the standard error is slightly higher (0.012 compared to 0.008). The more complex model used for the REML analysis may account for the lower precision of the heritability estimate. The two heritability estimates are not of the same quantity since the realized heritability is an estimate of within family heritability (h_w^2) whilst the REML estimate is of the individual heritability (h^2). To compare the estimates the REML heritability

estimate must be converted into the within family estimate using the following formula (Falconer, 1989):

$$h^2 = h_w^2(1-t)/(1-r) \tag{5.2}$$

where r = the relationship between family members (1/2 in this case) and t = the intra-class correlation of family members = $\frac{1}{2}h^2 + c^2$. As $t \approx 0.5$ using the estimates obtained from the analysis, h^2 and h_w^2 are almost identical.

5.7 Discussion

The main aim of this initial analysis was to discover whether there was substantial genetic variance present in the population and compare the results against earlier selection experiments. There were grounds for expecting there to be little genetic variation present in the X Lines because they were started from the cross between only two lines. Falconer (1973), however, used a large number of lines to form his original crosses so his selection lines would have had a much 'broader' genetic base. In fact a substantial selection response was achieved for the X Lines, and the estimate for h^2 of 0.27 that was obtained, although lower than Falconer's (1973) estimates of realized heritability (0.37), was still quite high. Since the two founder lines for the X Lines were so distinct (Section 4.1.2), however, the cross between them would contain a substantial proportion of the variance of the whole population, so the result is not as suprising as it at first might appear.

There were several unusual aspects to the results discussed in this chapter, the first of these being the odd behaviour of subline B of the Low lines which for most of the experiment is almost 1 s.d. below the other Low lines. There is no information available yet to determine the cause of this; possible causes would be a rare recombination event or a new mutation. It is not possible to distinguish between these possibilities simply on the basis of the data used here. It would theoretically be feasible to search for differences in molecular markers between subline B and the other lines to see if there was a section of chromosome associated

with this difference. Unfortunately there are likely to be many differences between the sublines so finding one associated with the weight difference would be very difficult, especially since the sublines are no longer available for further study.

The second unusual feature is the change in sexual dimorphism seen in the low lines, with the males responding more rapidly to downwards selection than the females so that the male and female means become more similar. Possible causes for this have been discussed in this chapter, but no further work on this phenomenon has been carried out. The main concern was how to account for this effect when performing more sophisticated analyses, this being done by fitting sex as a nested effect within generation and line, so allowing the effect of sex to differ between generations and selection lines.

The third peculiarity is the non-linear response curve shown by the Low lines, indicating that the response to selection accelerated over the middle portion of the experiment. This could indicate an increase in genetic variance for some reason over that period. There is, however, a problem with this dataset in that it is difficult to separate selection response from environmental change. In theory, the Control line should allow this separation, but the relatively small number of animals in the Control line means that not too much weight can be placed on this information. It is not really possible, therefore, to discount environmental change as a factor in the changing rate of response.

5.8 Conclusions

The anomalous behaviour of subline B of the Low lines, and the apparent changes in the rate of selection response, both indicate a divergence from the predictions of the infinitesimal model. Chapter 6 presents methods to modify the infinitesimal model allowing changes in variance can be estimated. If estimates of variance change are obtained that are significantly different from zero, then this will provide firmer evidence of the discrepancies between the X Lines dataset and the

infinitesimal model predictions. It is not possible to make firm conclusions about the causes of the discrepancies. It is likely that there are QTL which individually have a medium to large effect on body weight segregating in the population, and it is due to this fact that the infinitesimal model can not adequately explain the data. Chapters 7 and 8 describe methods for the estimation of the effects and positions of some of these QTL.

Chapter 6

The Analysis of Changes in Variance Over Time Using REML

6.1 Introduction

In Chapter 2 the estimation of changes in variance components under selection was discussed as a means of detecting deviations from the infinitesimal model. This chapter describes a series of modifications to Beniwal *et al.*'s (1992a) method (Section 2.2.2) to produce estimates of the changes in variance over time. The basis of Beniwal *et al.*'s (1992a) method was that the covariance matrices for the random effects were split into two blocks allowing separate variance components to be fitted to each block simultaneously. The first change to this was that the analysis was extended to allow the splitting of the data into an arbitrary number of blocks. The second modification was to allow all of the variance components within a block to change continuously over time by, in effect, fitting a linear regression on generation to all variance components. This is simpler to use than Wray's (1990) method (Section 2.3.2). With Wray's (1990) method a continuous change in the additive genetic variance per generation can be modelled, but the amount of change has to be set before generating the additive genetic covariance matrix. To estimate the degree of variance change, therefore, requires multiple evaluations of the covariance matrix. With the method described here

the covariance matrices have to be evaluated only once, making the procedure more efficient. The other drawback of Wray's (1990) method is that it is only applicable to the additive random genetic effect, whereas the method described here can be used for any random effect.

The method allows nesting of the regression within genetic groups so, for example, changes in variance can be estimated separately for High and Low selection lines. This is important because in many cases where the infinitesimal model does not hold (*i.e.*, if there was non-additive genetic variance present), the size and direction of any changes in variance would be affected by the direction and strength of selection and so could differ between the selection lines.

The data from the X Lines were analysed using these methods. Changes of variance were estimated and compared against the predictions from the infinitesimal model and the observed variance changes then used to make inferences about the effects of the genes controlling the trait.

6.2 Method

The method described here is an extension to the animal model allowing (a) the fitting of separate variance components to blocks of animals and (b) variance components to change continuously over the course of the experiment (in effect fitting the variances as regressions on generation). Parts (a) and (b) can easily be combined so that the variance component regressions are nested within blocks, allowing the variance to change separately in each block. For the analysis of the X Lines the blocks refer to the different selection directions, *i.e.*, variance components were fitted separately to High and Low selected lines.

Beniwal *et al.* (1992a) described fitting separate variance components to two blocks of animals. It is straightforward to extend this so that a larger number of blocks can be fitted. The general method for this is to split the (co)variance matrices for the random effects that are to be fitted separately into

the contributions from the different blocks. For example, let $\mathbf{V}$ be the covariance matrix for a random effect so the Cholesky decomposition of $\mathbf{V}$ can be written as $\mathbf{V} = \mathbf{TDT}'$ where $\mathbf{D}$ is diagonal and $\mathbf{T}$ is lower triangular. Each element of $\mathbf{D}$ then corresponds to a level of the random effect, so $\mathbf{D}$ can be partitioned into a set of sub-matrices according to which block each random effect level is in. If there were n blocks then $\mathbf{V}$ can be written as follows:

$$\mathbf{V} = \mathbf{T} \begin{pmatrix} \mathbf{D}_1 & 0 & \dots & 0 \\ 0 & \mathbf{D}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{D}_n \end{pmatrix} \mathbf{T}' \tag{6.1}$$

Then if $\mathbf{W}_i$ is an incidence matrix such that the element $w_{jj} = 1$ if random effect level j is in block i and 0 otherwise and $\sigma^2_{v_i}$ is the variance component for the ith element:

$$\text{Variance of random effect } = \sum_{i=1}^{n} \mathbf{T}(\mathbf{W}_i\mathbf{D})\mathbf{T}'\sigma^2_{v_i} = \sum_{i=1}^{n} \mathbf{V}_i\sigma^2_{v_i} \tag{6.2}$$

When this procedure is applied to the main animal effect, the separate covariance matrices for each block can easily be calculated using a slight modification of the 'normal' method for calculating the $\mathbf{A}$ matrix. In this, the diagonal elements of $\mathbf{A}$ are found using a procedure developed by Quaas (1976) and used to calculate the contributions of each animal to the off-diagonal elements of $\mathbf{A}$ using the method of Henderson (1976). In the modified method, a separate covariance matrix is calculated for each block in turn with the calculation being carried out as in the original method except that an animal's contributions are added to a given matrix only if it belongs to the relevant block.

The method can also be applied to other random effects if the covariance matrix can easily be split up. If it is assumed that there is no correlation between levels of the random effect, *i.e.*, the covariance matrix is proportional to the identity matrix, then the matrix for each block is simply a diagonal matrix which is only non-zero where the corresponding levels of the random effect belong to a given group. Applying (6.2) to this produces a modified covariance matrix for the

random effect which is diagonal with each element being equal to the variance component for the block corresponding to that level of the effect. For example the covariance matrix might look something like:

$$\mathbf{V} = \begin{pmatrix} \sigma^2_{v_1} & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2_{v_1} & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma^2_{v_2} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2_{v_n} & 0 \\ 0 & 0 & 0 & \dots & 0 & \sigma^2_{v_n} \end{pmatrix} \tag{6.3}$$

This method allows the comparison of variance components in different lines and/or different generations. The data were scored as being from Low or High lines, and from Generations 0-4, 5-9, 10-14, or 15-20. Analyses were then carried out to investigate differences in variance components between Low and High selected lines, different generation intervals, and the interaction between these. For each variance 'block', separate values for the additive, litter, and environmental variances were fitted.

The interpretation of the different variance estimates is straightforward when considering a random effect which is proportional to **I**. In this case, the information for the estimates come from records on animals within a given block. The situation is more complicated when the random effect has a non-diagonal covariance matrix, as is the case with the additive genetic effect. When the additive genetic matrix is not split up then the estimate of σ^2_a obtained is an estimate of the variance in the base population. When the covariance matrix is split up then the same principle holds, the estimates obtained are still of the base population variance. If separate variance components are fitted to generations 0-9 and 10-20, for example, then both components will be estimates of the base population variance, *i.e.*, the variance component for the block containing generations 10-20 is an estimate *not* of the variance in generation 10 but the variance in generation 0. Analysing a dataset which behaves completely as predicted by the infinitesimal model with this method would be expected to produce the same variance

component estimates for each block.

Part (b), where the variance is allowed to change continuously, was suggested by R. Thompson (Roslin Institute, Edinburgh). It requires the calculation of the standard covariance matrix plus an additional matrix for each order of regression fitted, for example, $\mathbf{V} = \mathbf{V_0} + b\mathbf{V_1}$ for a linear regression. In this case, $\mathbf{V_0}$ is equal to the normal additive covariance matrix and $\mathbf{V_1}$ is equal to $\mathbf{TDRT'}$ where $\mathbf{V_0} = \mathbf{TDT'}$ and $\mathbf{R}$ is a diagonal matrix with the ith element equal to $1/r_i$, r_i being the regression variable for level i of the random effect.

When applied to the main animal effect, this method can again be simply incorporated into the methods of Henderson (1976) and Quaas (1976). When calculating the 'regression' covariance matrix the normal procedure is followed except that the diagonal elements of $\mathbf{A}$ are divided by r_i (or r_i^n for the nth order regression matrix) before calculating the contributions of an animal to the matrix. As before, the procedure is simpler when the random effect in question is assumed to be proportional to the identity matrix. The linear regression covariance matrix for such a case would look like:

$$\mathbf{V} = \begin{pmatrix} \sigma_v^2/r_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & \sigma_v^2/r_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \sigma_v^2/r_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_v^2/r_n & 0 \\ 0 & 0 & 0 & \dots & 0 & \sigma_v^2/r_n \end{pmatrix} \tag{6.4}$$

The interpretation, again, is simplest when the covariance matrix is diagonal. In this case it is assumed that the variance of the effect changes linearly with some x variable. With the additive genetic effect the principle remains the same, the additive genetic variance *after accounting for changes in variance due to inbreeding and the Bulmer effect* is assumed to change linearly with the x variable. If the infinitesimal model holds for a given dataset, then the expected value of the regression coefficient is zero.

The two methods described above are complementary and can be applied together. In the case of the X Lines analysis this allowed the fitting of separate variance changes to the High and Low lines. A further refinement was to fit a common 'intercept' variance to all lines with separate 'gradient' variances being fitted to the High and Low lines. This is appropriate since the lines started from a common base population so the initial variances for all lines must be the same. The analysis was performed only fitting a linear regression to the variance components. This was due to computational requirements and because the regression coefficients are highly confounded making higher order regression analysis difficult. For each random effect thus treated, therefore, three variance components were fitted, a common intercept variance at generation zero (the F_2) and separate variances for the High and Low lines at generation 20. The additive, common environment, and environmental variances were all treated in this way.

The analysis was implemented by modifying the DFREML code of Meyer (1988). This involved modifying three main areas of the original code:

- The section which calculated the inverse of the **A** matrix was extended to allow the generation of the additional additive covariance matrices required.
- The setting up of the mixed model equations was modified to allow the fitting of multiple equations per level of random effect and to handle the multiple variance components per random effect.
- The calculation of the likelihood was extended to account for the additional variance components.

The analysis involved fitting a large number of variance components so it was not feasible to estimate standard errors for all terms simultaneously. For the regression model the standard errors for the variance components were estimated using quadratic approximations to the individual profile likelihoods (Meyer and Hill, 1992).

Another problem of fitting so many terms into the analysis was that the maximization procedure used (the downhill simplex method of Nelder and Mead (1965)) became somewhat unreliable, displaying a tendency to 'stick' at sub-optimal solutions. This could be resolved by restarting the process from the best point found until this resulted in no further improvement to the likelihood. Note that if the likelihood surface is truly multimodal then this method can only ensure that a local maximum has been reached. This is a general problem in maximization and there is no simple solution; the easiest procedure is to start off the maximization procedure from many different starting points and pick the best solution chosen, but this is very time consuming and not guaranteed to work (Fletcher, 1987).

A simulation test of the regression method was performed by analysing simulated data produced by the program described in 4.3.1 using (a) a large number (16384[1]) of unlinked genes of equal effect (*i.e.*, as an approximation to the infinitesimal model) and (b) a smaller number of genes (32). In both cases the initial gene frequency was set to 0.5. For case (a), the simulated datasets should produce zero estimates for the additive genetic variance regression coefficient, since the dataset should closely follow the predictions of the infinitesimal model. For case (b), however, the limited number of genes simulated should produce to a larger reduction in σ_a^2 over time (due to changes in gene frequency) that can be accounted for under the infinitesimal model, resulting in a negative estimate for the regression coefficient.

The simulated data structure closely followed the real experiment except that a litter effect was not simulated or estimated to reduce the computing costs. The base population for the simulation was an F_2 formed from a cross between two inbred strains. The F_2 was then split into 12 selection lines, 6 which were selected upwards and 6 selected downwards for 20 generations. Within each line

[1]The number of genes are in multiples of 32 because the simulation program uses 1 bit to represent each locus and can process 32 loci simultaneously on 32-bit word machines. It is therefore more efficient to set the number of genes to a 'round' number.

there were 8 full-sib families/generation and 8 individuals/family. The best 2 individuals from each family were selected. For all the simulations, the additive variance σ_a^2 was set to 1.0 and the environmental variance σ_e^2 was set to 3.0 (these being close to the actual values derived from the data). For both cases (a) and (b) the simulation tests were replicated 10 times and the mean and empirical standard error of the replicate tests were calculated.

6.3 Results

6.3.1 Simulated data

The means and empirical standard errors from 10 replicates of the simulation analysis using (a) 16384 genes and (b) 32 genes are given in Table 6.1. For both cases the estimated initial values for the variance components were close to the

Table 6.1: REML Estimates of the additive and environmental variance components from simulated data using (a) 16384 additive genes and (b) 32 additive genes fitting linear regressions to both variance components nested within lines. The results given are the mean of ten replicates along with the empirical standard errors.

	Initial Values		Low Line Increments		High Line Increments	
	$\hat{\sigma}_a^2$ (s.e)	$\hat{\sigma}_e^2$ (s.e)	$\hat{\sigma}_a^2$ (s.e)	$\hat{\sigma}_e^2$ (s.e)	$\hat{\sigma}_a^2$ (s.e)	$\hat{\sigma}_e^2$ (s.e)
16384	1.02 (0.02)	2.96 (0.02)	0.00 (0.04)	0.04 (0.03)	0.01 (0.04)	0.02 (0.05)
32	1.01 (0.03)	3.02 (0.02)	−0.27 (0.05)	−0.04 (0.04)	−0.23 (0.07)	0.02 (0.04)

The simulated values for the components are $\sigma_a^2 = 1.00$ and $\sigma_e^2 = 3.00$. σ_e^2 should not change over time but σ_a^2 should show a decrease when only a few genes are simulated due to changes in gene frequency. Note that changes in σ_a^2 due to inbreeding and the Bulmer effect are accounted for by inclusion of the **A** matrix into REML.

simulated values and the changes in σ_e^2 were not significantly different from zero. There is a difference between the two cases, however, with regard to the change in σ_a^2 which was not significantly different from zero for case (a) but was significantly less than zero for case (b), indicating a reduction in σ_a^2 greater than would be predicted under the infinitesimal model. This is to be expected with a small number of additive genes affecting the trait because there will be changes in gene frequencies away from 0.5 under selection leading to a reduction in σ_a^2, which is at a maximum when gene frequencies are at 0.5.

6.3.2 Experimental data - fitting variance 'blocks'

The estimates obtained when the data are divided into 2 blocks for the Low & High selected lines are shown in Table 6.2. There are differences between the variance estimates for the different lines. The additive variance estimate in the Low lines is over twice that in the High lines, whereas the litter variance in the High lines is $\sim$ 16% higher than that in the Low lines. The environmental variance is over 30% higher in the High lines than the Low lines.

Table 6.2: Estimates of variance components and genetic parameters for High and Low lines using log transformed data.

Line.	Variance Components $\times 10^{-3}$				$\hat{h}^2$	$\hat{c}^2$
	$\hat{\sigma}_a^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_p^2$		
Low	3.50	3.94	5.90	13.34	0.252	0.295
High	1.56	4.59	4.39	10.55	0.148	0.435

Variances estimated using the DFREML program fitting separate additive, litter, & environmental variances to the Low and High lines.

The estimates for the analysis with separate variance components fitted to Generations 0-4, 5-9, 10-14 and 15-20, taking the High and Low lines together are shown in Table 6.3. It shows differences in all the variance component estimates

between the different generation ranges. The additive variance estimates differ between the ranges, with an initial decrease from the original (generation 0-4) estimate shown in generations 5-9, followed by a large increase to over twice the original estimate in generation 10-15, followed by a decrease almost back to the original estimate. The litter variance shows a decline over the course of the experiment with the estimate for generations 15-20 being 86% that of the generation 0-4 estimate. The environmental variance does not change much until the last quarter of the experiment when it increases by 45% .

Table 6.3: Estimates of variance components and genetic parameters for different generation ranges using $\log_e$ transformed data

Gen.	Variance Components $\times 10^{-3}$				$\hat{h}^2$	$\hat{c}^2$
	$\hat{\sigma}_a^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_p^2$		
0-4	2.97	4.74	3.96	11.67	0.255	0.406
5-9	2.02	4.74	3.96	11.67	0.210	0.377
10-14	6.67	4.48	3.98	15.11	0.441	0.297
15-20	3.19	4.12	5.76	13.08	0.244	0.315

Variances estimated using the DFREML program fitting separate additive, litter, & environmental variances to generations 0-4, 5-9, 10-14, & 15-20

The results of splitting the data into 8 blocks so that each of the High and Low lines are split into an 4 blocks of generations 0-4, 5-9, 10-14 and 15-20 are shown in Table 6.4. The Low lines show a large increase in σ_a^2 over the first three quarters of the experiment, after which it appears to decrease. This change in σ_a^2 can also be inferred from Table 5.3, which shows that the response of the Low lines follows a similar pattern. The High lines also show a change in σ_a^2 which again follows a similar pattern to the changes in rate of response noted in Table 5.3, with σ_a^2 being highest at the beginning and end of the experiment, and lowest during the middle sections. The other large change in the variance component estimates occurs in the Low lines where σ_e^2 shows a large increase at the end of the experiment, with the estimate for generation 15-20 almost double those for

the other blocks. The High lines show only a slight increase in σ_e^2. Although there are differences between the σ_c^2 estimates between lines and generation ranges, a pattern to the differences is not readily apparent.

Table 6.4: Estimates of variance components and genetic parameters for different lines and different generation ranges using $\log_e$ transformed data.

Line.	*Gen.*	Variance Components $\times 10^{-3}$				$\hat{h}^2$	$\hat{c}^2$
		$\hat{\sigma}_a^2$	$\hat{\sigma}_c^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_p^2$		
Low	0-4	3.68	4.40	3.92	12.01	0.307	0.367
	5-9	4.27	3.31	3.92	11.50	0.371	0.288
	10-14	11.84	4.82	3.95	20.62	0.574	0.234
	15-20	2.71	3.50	7.85	14.06	0.193	0.249
Low	0-4	2.52	5.02	3.92	11.47	0.220	0.438
	5-9	0.30	3.86	3.92	8.08	0.037	0.478
	10-14	3.30	4.46	3.92	11.69	0.283	0.382
	15-20	4.35	4.38	4.10	12.83	0.339	0.341

Variances estimated using the DFREML program fitting separate additive, litter, & environmental variances to generations 0-4, 5-9, 10-14, & 15-20 in the High & Low lines

6.3.3 Experimental data - fitting continuous variance changes

The results from the analysis of the experimental data fitting linear regression coefficients to all variance components (additive, common environmental and environmental) nested within the High and Low lines are shown in Table 6.5a. The standard errors presented are calculated using a quadratic approximation to the profile likelihood for each component. The main changes are found in the Low lines where there is a substantial change in all variance components over the course of the experiment. The phenotypic variance in the Low lines increases from 10.6×10^{-3} to 16.5×10^{-3}; this increase is due to increases in both the additive and

Table 6.5: REML estimates of variance components and genetic parameters using log transformed data fitting (a) linear regressions to all variance components nested within lines and (b) as previous analysis but omitting subline B of the Low line.

	Var. Components $\times 10^{-3}$					
	$\hat{\sigma}_a^2$ (s.e.)	$\hat{\sigma}_c^2$ (s.e.)	$\hat{\sigma}_e^2$ (s.e.)	$\hat{h}^2$ (s.e.)	$\hat{c}^2$ (s.e.)	$\hat{e}^2$ (s.e.)
(a) Heterogenous variance analysis - all lines						
Initial Values	2.69 (0.26)	4.55 (0.13)	3.33 -	0.25 (0.02)	0.43 (0.00)	0.32 -
Low Line Increments†	2.66 (1.15)	-1.26 (0.72)	4.53 (0.53)	0.07 (0.07)	-0.23 (0.04)	0.16 (0.05)
High Line Increments	0.96 (0.84)	-0.26 (0.72)	0.76 (0.37)	0.05 (0.06)	-0.07 (0.04)	0.02 (0.04)
(b) Heterogenous variance analysis - omitting Low Subline B						
Initial Values	2.49	4.65	3.53	0.23	0.44	0.33
Low Inc.†	1.01	-1.08	3.98	0.01	-0.19	0.18
High Inc.	1.94	-0.48	0.20	0.13	-0.10	-0.03

†Increments are the estimated differences between components at the start and end of the experiment. Variances are assumed to change linearly between their starting and finishing values.

environmental variance with the litter variance, by contrast, decreasing. When the variances are considered as proportions of the total variance at the beginning and end of the experiment, h^2 increases from 0.25 to 0.32 while c^2 decreases from 0.43 to 0.20. These results indicate that the response to selection in the Low lines should increase over time due to the increases in h^2 and σ_p^2, and the observed response (Figure 5.1) does support this, with the Low lines showing an acceleration of response over the middle section of the experiment. The analysis was repeated omitting subline B of the Low lines (which showed a very different response from the other sublines; Table 6.5b). The Low lines show a smaller

increase in σ_a^2 than in the previous analysis while the High lines show a larger increase, although the differences between the analyses are not significant.

6.4 Discussion

A significant increase in additive and environmental variance under selection was detected in the low selected lines by both the discrete and continuous variance change methods, a result contrary to the predictions of the infinitesimal model. There are several possible causes for this increase. It has been noted before that inbreeding can result in a reduction in the capacity of organisms to regulate developmental processes. This can lead to inbred lines being more variable than the outbred parental lines (Maynard Smith, 1989). Most of the increase in variance was 'attributed' to an increase in environmental variance. This does not, however, necessarily mean that the increase is mainly non-*genetic* but rather that it is non-*additive* genetic in nature. Genetic variance changes that do not fit the model of a linearly increasing additive variance may be erroneously partitioned into the environmental or litter components. Increases in genetic variance could be caused by new mutations, non-additive gene action (*i.e.*, dominance or epistasis) and the breakdown of linkage disequilibrium. These possibilities are discussed below.

Increases in additive genetic variance have been reported in small populations undergoing random drift where the infinitesimal model would predict a decrease (Bryant and Combs, 1986). A relevant example of this is a study (Rahnefeld *et al.*, 1963) of a selection experiment in mice using a cross between two unspecified inbred lines as the foundation population. When the additive variance in individual generations was estimated using the average value obtained from the sire component and parent-offspring regressions they found a slight (but non-significant) increase over the course of the experiment.

The analysis presented here shows that the infinitesimal model cannot adequately explain the results of the X-Line experiment. One possible alternative

model would be a trait which is controlled by a relatively few genes. Simply reducing the number of genes in the model, however, leads to a poorer fit since if gene action is assumed to be exclusively additive then such a model predicts that the additive variance should decrease under selection due to changes in gene frequency away from 0.5. This is shown by the simulations using 32 genes described earlier. Several studies have reported decreases in additive variance under selection (Meyer and Hill, 1991; Beniwal *et al.*, 1992a), which is more in line with what would be expected if the trait was largely under the control of a few additive genes.

If there was some directional dominance in gene action, as indicated by the hybrid vigour shown in the F_1 generation, then this could lead to an increase in variance under selection because the maximum genetic variance is no longer when the gene frequency is at 0.5. Under this model, however, whilst selection in one direction would produce a rise in genetic variance, selection in the opposite direction would yield a decrease in variance faster than that under a purely additive model, a pattern of variance changes not seen in this study.

Interaction between rather than within loci can also increase additive variance as frequencies shift from 0.5. If a population experiences a bottle neck and is then maintained with a small population size so that gene frequencies alter under drift, the additive variance can increase substantially for many generations (Goodnight, 1988). This can be explained by an epistatic model of gene action since epistatic variance is at a maximum at intermediate gene frequencies. As genes become fixed by drift or selection, the epistatic variance is converted into additive variance. If this is enough to compensate for the loss of initial additive variance caused by genes approaching extreme frequencies then the additive variance could increase under selection in both directions.

An increase in genetic variance can also be caused by a breakdown of linkage disequilibrium between pairs of loci of opposite effect. When the genes show complete association (i.e. the same alleles at both loci always occur together),

the variance due to the gene pair will be proportional to the square of the sum of the effects of the two genes (assuming additivity), whereas if the genes are not associated then the variance due to the pair will be proportional to the sum of the squares of the effects of the two genes. If the effects of the genes oppose each other then the variance due to the genes with total linkage disequilibrium will be less than that with no disequilibrium. Since the F_1 population is in total linkage disequilibrium, as the experiment proceeds this should break down, potentially leading to an increase in genetic variance. The size and duration of any increase are dependent on the degree of linkage; if the genes are tightly linked then a small, gradual increase will result and if the genes are loosely linked or unlinked then a large, but short lived, increase will occur.

The main problem with this model is that adjacent genes must be in repulsion (having opposing effects). For example, in the simulation described earlier with 16384 unlinked additive genes, the parental lines were set up with alleles assigned randomly to each parent. In this situation there was no change in additive variance apart from that predicted by the infinitesimal model, so it is not enough for '+' and '−' alleles to be assigned randomly, rather they must be arranged as '+ − + − + − +−' etc. A possible mechanism for achieving this is stabilizing selection in the parental lines because if an allele becomes fixed at one locus then there will be selection for a 'compensatory' allele at another locus to 'balance out' the effect of the first allele so that the overall effect of the chromosome is minimized (Mather, 1941; Lewontin, 1964). The genes either have to be tightly linked, however, or the selection very strong for the gene combinations to depart much from a random arrangement (Wright, 1969).

A further problem is that with this experimental design, most of the initial linkage disequilibrium will disappear by the F_2. Because the F_1 is not produced by random mating but by crossing between the two lines only, the disequilibrium between any genes that are unlinked will be zero in the F_2 rather than decaying at a rate of 0.5 per generation. This means that, again, the + and − alleles discussed in the above paragraph must be tightly linked for there to be more

than a very transient effect on the variance produced by the breakdown of linkage disequilibrium.

Rather than the increase in variance being due to an unlocking of existing genetic variance through changes in gene frequency or loss of disequilibrium, mutation could lead to new genetic variance being generated. The anomalous response of subline B of the Low lines could be due to a new mutation arising in this line during the experiment. Since mutations would enter the population at low frequencies, favourable mutations could cause an increase in genetic variance as their frequency moved towards 0.5, although this increase would be offset by losses due to inbreeding.

6.5 Conclusions

It is clear that the infinitesimal model cannot adequately explain the behaviour of the X Lines. Phenotypic variance has increased significantly under selection in both directions, and several models are presented as possible candidates for this increase, although determining which is closest to the actual model is difficult. The directional dominance model does not predict the observed pattern of variance changes so is unlikely to be the main cause of the increase in variance. It is, however, likely to play a role since it is evident that some directional dominance is present. The linkage model seems unlikely given the stringent conditions that must be met for it to produce the effect seen here. An epistatic model has a more plausible explanation for how it could occur. Epistatic variance will be highest at intermediate gene frequencies because it is caused by the interactions between loci, and as loci become fixed there is less chance of interactions. In the inbred founder lines there should be no genetic variance of any type, however, crossing the lines restores some of the epistatic variance present in the ancestral population from which the lines were originally derived. Drift then acts to convert this epistatic variance into additive variance as described earlier. Mutation could

also be an important source of new genetic variation, although this should have resulted in much divergence between the replicate lines and, apart from subline B of the Low lines, this divergence is not apparent.

There is potential for further analysis that could shed light on the causes of the increase in variance seen here. The model of analysis could be extended to explicitly include non-additive and interaction terms into the estimation process. Also, marker frequencies measured at the end of the experiment could be used to obtain estimates of gene effects linked to the markers (Keightley and Bulfield, 1993) and interactions between genes, eventually producing a distribution of gene effects and interactions which may explain the variance increase. A Gibbs sampling based approach to estimating QTL effects using trait and marker data is described in the following chapter.

Chapter 7

The Estimation of Linked QTL Effects with Mixed Inheritance Models using Gibbs Sampling

7.1 Introduction

This chapter presents a Markov Chain Monte Carlo (MCMC) method for the detection and mapping of one or more QTL in a population derived from an inbred cross. The population can be large, contain many generations of animals and have almost any structure (apart from the restriction that the initial population must be formed from a inbred cross). It is not assumed that the QTL contribute all of the genetic variance of the trait, instead a mixed inheritance model (2.3.3) is fitted which partitions the genetic variance into that due to the QTL and a residual additive polygenic effect. The method is suitable for analysing datasets where animals are recorded both for a quantitative trait and for one or more genetic markers. It is not necessary for all animals to have records, and the method can be applied to situations where, for example, there is marker information available only on a subset of animals. The genetic model is of one or more linkage groups each consisting of a number of linked QTL and markers (multiple linkage groups can be fitted with little difficulty since they can be treated independently).

Estimates of the effects and positions of the QTL with respect to the observed markers in the linkage group are obtained, as well as estimates of the polygenic additive and environmental variance components. The method is developed for several different discrete gene models, which can be distinguished by the numbers of linked genes in a linkage group. The simplest genetic model would be a single QTL with no markers linked to it. This is essentially a segregation analysis in which the position of the QTL is irrelevant, so the only quantities estimated are the QTL effects. The next level of complexity would be a linkage group with two genes, a single QTL and a linked marker. In this case it is not possible to map the QTL with respect to the marker (because it is not possible to tell on which side of the marker the QTL is), so instead the recombination frequency between the QTL and marker is estimated. The most complex discrete gene model that could be fitted would have multiple linked QTL and markers, however for the analyses described here only a single QTL was fitted. The marker positions were assumed to be known, and the QTL was allowed to 'float', so it could be anywhere along the chromosome. The effectiveness of the method is demonstrated by analysing simulated datasets, and in the next chapter the method is used to analyse the X-Line data to produce estimates of the effects of putative QTL for body weight linked to the coat colour loci *brown* and *dilute*.

7.2 Method

The method of analysing the data is to fit a univariate linear mixed animal model with the animals polygenic value fitted as a random effect and the QTL being fitted as fixed effects. Since, however, the QTL genotypes are unknown and must be inferred from the trait and marker data, the incidence matrix for the QTL effects is itself a parameter to be estimated. The model is therefore

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{Q}\boldsymbol{\lambda} + \boldsymbol{e} \tag{7.1}$$

where $\boldsymbol{y}$ is an $(n \times 1)$ data vector; $\mathbf{X}$ is the known incidence matrix of the fixed effects apart from the QTL ($n \times l$ where l is the number of levels of fixed effects); $\mathbf{Z}$ is the known incidence matrix for the polygenic (residual) additive effects ($n \times u$ where u is the number of animals); $\boldsymbol{\alpha}$ is the vector of fixed effects apart from the QTL ($l \times 1$); $\boldsymbol{\beta}$ is the vector of polygenic additive effects ($u \times 1$); $\boldsymbol{e}$ is the vector of random residual effects ($n \times 1$); $\mathbf{Q}$ is the unknown incidence matrix of the QTL effects ($n \times 2m$ where m is the number of QTL with each QTL having an additive and a dominance effect); and $\boldsymbol{\lambda}$ is the vector of QTL additive and dominance effects ($2m \times 1$). The model can easily be extended to allow additional random effects such as common environmental effects.

The difficulty with using this model arises from $\mathbf{Q}$ being unknown. The maximum likelihood method to estimating the model parameters would be to find the configuration of all unknown parameters in the model (including $\mathbf{Q}$) that maximized the likelihood of the observed data (both trait and marker data). Maximizing the likelihood with respect to $\mathbf{Q}$, however, is extremely difficult due to the very large number of parameters (the number of genotypes to be estimated) and the large degree of dependency between parameters. As discussed in Sections 2.3.3 and 3.4, Gibbs sampling provides a simple method of sampling from the distribution of $\mathbf{Q}$ conditional on the observed data and on the other model parameters. The samples of $\mathbf{Q}$ from the Gibbs sampling process can then be used in a conventional mixed model analysis to obtain estimates of the other model parameters (Guo and Thompson, 1992; Guo and Thompson, 1994). An alternative approach is to use Gibbs sampling for the whole analysis (Janss *et al.*, 1994b; Janss *et al.*, 1994a), producing Bayesian estimates of the posterior marginal distributions of the parameters of interest. The method described here follows the second approach.

Gibbs sampling is used to sample the 'missing data' (polygenic breeding values, the complete genotypic configuration $\mathcal{G}$(which is described in Section 2.3.3 and from which $\mathbf{Q}$ can be derived), gene positions, fixed effects, QTL effects and variance components) from their joint distribution conditional on the data, so that

estimates of any or all of these can be obtained. Note that the complete ordered genotype of each animal is estimated, *i.e.*, the maternal and paternal alleles are estimated separately, so the source of each allele can be traced. This means that linkage phase is known, thereby simplifying the estimation of recombination frequencies.

Gibbs sampling requires that the conditional distributions of all parameters are known and can be sampled from. It is therefore necessary to specify these distributions and describe the sampling process. The conditional distributions can be obtained from the joint posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ (where $\boldsymbol{\theta}$ is the vector of model parameters) by regarding all but the parameter in question as known. The joint posterior distribution itself is proportional to the product of the prior distributions of the parameters and the likelihood function.

7.2.1 Prior distributions

The prior distributions reflect what is known about the parameters before the analysis. For the method described here, the prior distributions for all effects were defined so that little or no prior information was used in the estimation *i.e.*, 'naive' prior distributions were used (Wang *et al.*, 1993). The fixed effects (including the QTL effects) were assumed to have a uniform prior distribution, so that all values were equally likely.

$$p(\boldsymbol{\alpha}) \propto \text{constant} \tag{7.2}$$

$$p(\boldsymbol{\lambda}) \propto \text{constant} \tag{7.3}$$

The polygenic values were assumed to be normally distributed about zero with variance $\mathbf{A}\sigma_a^2$ where $\mathbf{A}$ is the numerator relationship matrix and σ_a^2 is the polygenic additive variance. The residual effects were likewise assumed to be normally distributed about zero with variance $\mathbf{I}\sigma_e^2$ where σ_e^2 is the environmental variance. These are the same assumptions that are made in a 'classical' statistical analysis. The priors for the variance components are also assumed to be uniform

distributions.

$$p(\sigma_e^2) \propto \begin{cases} 0 & \text{if } \sigma_e^2 < 0, \\ \text{constant} & \text{if } \sigma_e^2 \geq 0. \end{cases} \tag{7.4}$$

$$p(\sigma_a^2) \propto \begin{cases} 0 & \text{if } \sigma_a^2 < 0, \\ \text{constant} & \text{if } \sigma_a^2 \geq 0. \end{cases} \tag{7.5}$$

For linkage groups with just two linked loci (a single QTL linked to one marker) the prior distribution assumed for the recombination frequency was uniform between 0 and 0.5. For linkage groups with more than two linked loci, the position of the QTL relative to any linked markers was assumed to be uniformly distributed. Since information on map positions derives from the observed recombination rate between loci, it was necessary to specify a mapping function to convert from recombination rates to map distances. Haldane's mapping function was used which assumes no chiasma interference. With this the expected recombination rate between two genes located d Morgans apart will be:

$$r = \frac{1}{2}(1 - e^{-2d}), \tag{7.6}$$

The prior distribution of genotypes is uniform so that any genotype is equally likely, except in the case of F_2 individuals which are taken to be heterozygous at all loci, with all the '1' alleles for all F_2 individuals coming from one parent, and all the '0' alleles coming from the other parent. This comes from the definition of the F_2 as being derived from a cross between two inbreds. This definition does *not* assume that the alleles with positive effects on the trait all come from one parent; the '1' alleles can be associated with either positive or negative effects.

7.2.2 Posterior distributions

The joint posterior distribution, as mentioned before, is proportional to the product of the joint prior distribution and the likelihood function and, following

Sørensen *et al.* (1994b), can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto (\sigma_e^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma_e^2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\alpha} - \mathbf{Z}\boldsymbol{\beta} - \mathbf{Q}\boldsymbol{\lambda})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\alpha} - \mathbf{Z}\boldsymbol{\beta} - \mathbf{Q}\boldsymbol{\lambda})\Big)$$
$$\cdot (\sigma_a^2)^{-u/2} \exp\Big(-\frac{1}{2\sigma_a^2}\boldsymbol{\beta}'\mathbf{A}^{-1}\boldsymbol{\beta}\Big) p(\mathcal{G}|.) p(\boldsymbol{\kappa}|.), \tag{7.7}$$

where $\mathcal{G}$ is the ordered genotype of all animals in the pedigree and $\boldsymbol{\kappa}$ is the vector of map positions of all QTL and markers. $p(X|.)$ refers to the full conditional distribution of X. The full conditional posterior distributions for a given parameter can be derived from (7.7) by regarding all other parameters as fixed.

From (7.7), the conditional posterior distribution for $\boldsymbol{\alpha}$ will be a multivariate normal distribution with a mean vector equal to:

$$\tilde{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\boldsymbol{y} - \mathbf{Q}\boldsymbol{\lambda} - \mathbf{Z}\boldsymbol{\beta}), \tag{7.8}$$

with variance

$$V(\tilde{\boldsymbol{\alpha}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2. \tag{7.9}$$

The conditional posterior distribution for the QTL effects will also be a normal distribution with a mean vector given by:

$$\tilde{\boldsymbol{\lambda}} = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\alpha} - \mathbf{Z}\boldsymbol{\beta}), \tag{7.10}$$

with variance:

$$V(\tilde{\boldsymbol{\lambda}}) = (\mathbf{Q}'\mathbf{Q})^{-1}\sigma_e^2. \tag{7.11}$$

The conditional posterior distribution for the polygenic additive random effects will be a multivariate normal with, following the results of Sørensen *et al.* (1994b), a mean vector given by:

$$\tilde{\beta}_i = \Big(\mathbf{Z}_i'\mathbf{Z}_i + a_{ii}\frac{\sigma_e^2}{\sigma_a^2}\Big)^{-1}\Big(\mathbf{Z}'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\alpha} - \mathbf{Q}\boldsymbol{\lambda}) - a_i\boldsymbol{\beta}_{-i}\frac{\sigma_e^2}{\sigma_a^2}\Big), \tag{7.12}$$

with variance:

$$V(\tilde{\beta}_i) = \left(\mathbf{Z}_i'\mathbf{Z}_i + a_{ii}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\sigma_e^2, \tag{7.13}$$

where β_i is the polygenic additive genetic effect for animal i, $\boldsymbol{\beta}_{-1}$ is the vector of polygenic effects for all other animals, $\boldsymbol{a}_i$ is the row of $\mathbf{A}^{-1}$ corresponding to animal i and a_{ii} is the i^{th} diagonal element of $\mathbf{A}^{-1}$.

The conditional posterior distribution for σ_a^2 will be:

$$p(\sigma_a^2|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{Q},\boldsymbol{y}) \propto (\sigma_a^2)^{-u/2}\exp\left(-\frac{1}{2\sigma_a^2}\boldsymbol{\beta}'\mathbf{A}^{-1}\boldsymbol{\beta}\right) \tag{7.14}$$

which is an inverted gamma with parameters:

$$\nu_a = u - 2, \quad s_a^2 = \boldsymbol{\beta}'\mathbf{A}^{-1}\boldsymbol{\beta}/\nu_a \tag{7.15}$$

Similarly the conditional posterior distribution for σ_e^2 will be:

$$p(\sigma_e^2|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{Q},\boldsymbol{y}) \propto$$
$$(\sigma_e^2)^{-n/2}\exp\left(-\frac{1}{2\sigma_e^2}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\alpha}-\mathbf{Z}\boldsymbol{\beta}-\mathbf{Q}\boldsymbol{\lambda})'(\boldsymbol{y}-\mathbf{X}\boldsymbol{\alpha}-\mathbf{Z}\boldsymbol{\beta}-\mathbf{Q}\boldsymbol{\lambda})\right) \tag{7.16}$$

which is an inverted gamma with parameters:

$$\nu_e = n - 2, \quad s_e^2 = \left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\alpha}-\mathbf{Z}\boldsymbol{\beta}-\mathbf{Q}\boldsymbol{\lambda}\right)'\left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\alpha}-\mathbf{Z}\boldsymbol{\beta}-\mathbf{Q}\boldsymbol{\lambda}\right)/\nu_e \tag{7.17}$$

The conditional distribution of an individual's genotype depends solely on the genotypes of the immediate neighbours in the pedigree and the animals own observations (Sheehan and Thomas, 1993), so to calculate this distribution it is only necessary to consider the genotypes of the individual's parents and any spouses and offspring, as well as the individual's own records. The full conditional probability of the genotype for animal i at the QTL and marker loci, $\mathcal{G}_i$ where $\mathbf{Q}_i$ is the QTL genotype corresponding to $\mathcal{G}_i$, $\mathcal{G}_{b_i}$ represent the genotype of the neighbours of animal i in the pedigree, $\mathcal{G}_{\text{par}}$ is the parental genotypes, $\mathcal{G}_{\text{spouse}_j}$ is the genotype of the jth spouse, $\mathcal{G}_{\text{off}_{j,k}}$ is the genotype of the kth offspring of the jth spouse, $\boldsymbol{\kappa}$ is the vector of gene positions, $\boldsymbol{y}_i$ is the vector of observations on

animal i (including trait and marker data) and $\mathbf{Z}_i$ and $\mathbf{X}_i$ are the rows of $\mathbf{Z}$ and $\mathbf{X}$ respectively relating to animal i, will therefore be given by:

$$p(\mathcal{G}_i|\sigma_e^2, \mathcal{G}_{b_i}, \boldsymbol{\kappa}, \boldsymbol{y}_i) = p(\mathcal{G}_i|\mathcal{G}_{b_i}, \boldsymbol{\kappa})p(\mathcal{G}_i|\boldsymbol{y}_i, \sigma_e^2) \tag{7.18}$$

where:

$$p(\mathcal{G}_i|\mathcal{G}_{b_i}, \boldsymbol{\kappa}) \propto p(\mathcal{G}_i|\mathcal{G}_{\text{par}_i}, \boldsymbol{\kappa}) \prod_{j,k} p(\mathcal{G}_{\text{off}_{j,k}}|\mathcal{G}_i, \mathcal{G}_{\text{spouse}_j}, \boldsymbol{\kappa}), \tag{7.19}$$

$$p(\mathcal{G}_i|\boldsymbol{y}_i, \sigma_e^2) \propto \exp\Big(-\frac{1}{\sigma_e^2}(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\alpha} - \mathbf{Q}_i\boldsymbol{\lambda} - \mathbf{Z}_i\boldsymbol{\beta})'(\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\alpha} - \mathbf{Q}_i\boldsymbol{\lambda} - \mathbf{Z}_i\boldsymbol{\beta})\Big), \tag{7.20}$$

For linkage groups with just two loci, a marker and a QTL, an estimate is made of the recombination frequency between the two loci. Treating all recombination events as independent, then the number of observed recombinations (R) will be distributed as a binomial variable with parameters r and $R + R'$, where r is the recombination frequency between the marker and the QTL and R' is the number of observed non-recombination events. The sampling procedure for $\mathcal{G}$reconstructs the ordered genotype of all animals, *i.e.*, it is known from which parent a given chromosome derives. It is only possible to observe a recombination between two loci if a parent is heterozygous for both loci, so $R+R'$ is equal to the total number of parents heterozygous for both the marker and QTL, and R is the number of recombinant chromosomes in the offspring of these parents. The conditional probability of R given r and R' will be a binomial probability:

$$p(R|R', r) = \frac{\Gamma(R + R' + 1)}{\Gamma(R + 1)\Gamma(R' + 1)} r^R (1 - r)^{R'}, \tag{7.21}$$

and the conditional probability of r given R and R' for the interval $0 \leq r \leq 0.5$ will be:

$$p(r|R, R') = \frac{p(R|R', r)p(r)}{\int_0^{0.5} p(R|R', r)p(r)dr}, \tag{7.22}$$

and 0 everywhere else. The integral in (7.22) evaluates to $1/(R + R' + 1)$ if taken over the interval 0 to 1, however in this case we are only integrating from 0 to 0.5.

If $R + R'$ is large then this still evaluates to very close to $1/(R + R' + 1)$ except when $R/(R + R')$ is close to 0.5. The conditional probability (7.22) can therefore be approximated by a beta distribution with parameters $R + 1$ and $R' + 1$:

$$p(r|R, R') \approx \frac{\Gamma(R + R' + 2)}{\Gamma(R + 1)\Gamma(R' + 1)} r^R (1 - r)^{R'}. \tag{7.23}$$

Note that the approximation does not affect the accuracy of the Gibbs sampling. Sampling is carried out on the distribution conditional on R and R', and for any given set of values for R and R' the integral in (7.22) is a constant. It is only necessary to use the exact formula (7.22) when estimating the marginal density (Section 7.2.4).

For linkage groups with more than two loci then the likelihood of gene positions, $\boldsymbol{\kappa}$, is determined by the number of observed recombinations between pairs of adjacent informative genes. These are independent if it is assumed that there is no interference, so the probability of observing a given number of recombinations conditional on the total number of observable recombinations and the distance between all such gene pairs would be the product of the probability for each gene pair. Equivalently, the probability can be obtained from the product of the probability of each individual's ordered genotype given the parental genotypes and $\boldsymbol{\kappa}$. Considering each QTL separately, then the chromosome containing the QTL can be split up into a number of intervals by the loci linked to the QTL. For any given interval, the probability in terms of the position of the QTL within that interval can be calculated. This is done for each interval in turn to give the probability of the QTL position along the whole chromosome. The probability for an individual chromosome C conditional on the parental genotypes $\mathcal{G}_{par}$ and $\boldsymbol{\kappa}$ is given by:

$$p(C|\mathcal{G}_{par}, \boldsymbol{\kappa}) \propto \prod \Big(\delta_{ij} r_{ij} + (1 - \delta_{ij})(1 - r_{ij}) \Big) \tag{7.24}$$

where the product is taken over all pairs of adjacent informative loci, r_{ij} is the recombination frequency between a pair of loci (given by Haldane's mapping function) and δ_{ij} is 1 if a recombination has occurred between the loci, and

0 otherwise. δ_{ij} is determined by examining the parental chromosomes. The probability (7.24) conditional on the complete genotype configuration ($\mathcal{G}$) and $\boldsymbol{\kappa}$, is independent for every haploid genome in the pedigree. An overall probability for $\mathcal{G}$ can therefore be obtained from the product of (7.24) taken over both chromosomes for all animals in the pedigree:

$$p(\mathcal{G}|\kappa) \propto \prod_i p(C_i|\mathcal{G}_{par_i}, \boldsymbol{\kappa}) \tag{7.25}$$

The probability (7.24) and, correspondingly, (7.25) can be split into two components, one of which involves the QTL and its flanking informative markers, and the other which considers the remaining gene pairs. This second part of the probability is not dependent on the QTL position *within an interval*, but obviously is affected by which interval the QTL is in. For a given interval, therefore, (7.25) can be expressed as the product of a term dependent on the QTL position and a constant. The constant has to be included so that the information content of the different intervals is the same and the probabilities *across* intervals can be compared. The conditional probability of the QTL position will be given by:

$$p(\kappa_j|\mathcal{G}, \boldsymbol{\kappa}_{-j}) = \frac{\prod_i p(C_i|\mathcal{G}_{par_i}, \kappa_j, \boldsymbol{\kappa}_{-j})}{\int \prod_i p(C_i|\mathcal{G}_{par_i}, \kappa_j, \boldsymbol{\kappa}_{-j}) d\kappa_j} \tag{7.26}$$

where κ_j is the position of QTL j and $\boldsymbol{\kappa}_{-j}$ is the vector of positions of the other loci. The integral in the denominator of (7.26) is constant conditional on $\boldsymbol{\kappa}_{-j}$ and $\mathcal{G}$, and can therefore be ignored for sampling purposes, but must be considered when estimating the marginal density of κ_j. For sampling, therefore, the following was used for the conditional probability of κ_j:

$$p(\kappa_j|\mathcal{G}, \boldsymbol{\kappa}_{-j}) \propto \prod_i p(C_i|\mathcal{G}_{par_i}, \kappa_j, \boldsymbol{\kappa}_{-j}) \tag{7.27}$$

Writing K_i as the constant for interval i, R_{i_k} as the number of recombinations between the QTL and locus k when the QTL is in interval i (only counting if k is a flanking informative marker of the QTL), R'_{i_k} similarly as the number of non-recombinations and r_k as the recombination frequency between the QTL and

locus k, then the conditional probability (7.27) can be written as:

$$p(\kappa_j|\mathcal{G}, \boldsymbol{\kappa}_{-j}) \propto K_i \prod_k r_k^{R_{i_k}} (1 - r_k)^{R'_{i_k}} \tag{7.28}$$

where

$$r_k = \frac{1}{2}(1 - e^{-2|\kappa_j - \kappa_k|}), \tag{7.29}$$

the recombination between the QTL and gene k.

7.2.3 Sampling scheme

The basis of the sampling scheme is as follows: the parameters are initialized to arbitrary starting values. The only restriction on the starting values is that they should be valid, *i.e.*, the likelihood should be non-zero. If, however, the starting values are a long way from the equilibrium joint distribution, then it can take a long time to achieve convergence of the Markov chain. Producing valid initial samples for most of the parameters is not difficult, for example the vectors of effects can simply be initialized to zero. Generating an initial valid genotype configuration $\mathcal{G}$ can be more complex. The method used here was that any animal whose genotype at a given locus was not completely known (*i.e.*, all the QTL loci and heterozygous marker loci), was assigned an initial heterozygous genotype for that locus. This will produce an initial valid *unordered* configuration of $\mathcal{G}$, *i.e.*, the genotypes of all animals will be valid, but the haplotypes may not be. For this reason, the first time the genotypes are sampled, the calculation of the conditional probabilities of offspring is done without taking account of the source of each allele. This allows the production of a consistent ordered configuration of $\mathcal{G}$.

After initialization, each effect is sampled in turn from its full conditional posterior distribution. After all effects have been sampled the cycle begins again. One cycle of sampling would therefore entail:

- Sample $\boldsymbol{\lambda}$ and update $\boldsymbol{\lambda}$.

- Sample α and update.
- For each individual i in turn, sample β_i and update.
- Sample σ_a^2 from (7.14) and update.
- Sample σ_e^2 from (7.16) and update.
- For each individual i in turn, sample $\mathcal{G}_i$ from (7.18) and update.
- For each linkage group with two genes (a marker and a QTL) count the numbers of informative parents and observed recombinations between the two loci. Sample r from (7.22) and update.
- For each linkage group with > 2 genes, count the numbers of informative parents and observed recombinations between a locus and all other loci linked to it for each 'floating' locus in turn. Sample κ from (7.28) and update.

Sampling from these distributions is mostly straightforward; routines to sample from uniform, normal, gamma and beta distributions are readily available. The exception to this is sampling κ from (7.28), because this distribution is non-standard being a product of several transformed beta distributions, with the transformation function (from recombination frequency to a map position) changing depending on which genes are flanking the current position. A further problem with sampling from (7.28) is that it is defined separately for each interval and is generally multimodal, with a mode in each interval. The sampling procedure used samples from each interval in turn and then picks one of the samples x with a probability proportional to $p(x)$ (7.28). The method developed for sampling from (7.28) is described in more detail in the Appendix.

Note that some of these sampling sub-units could be further split up. For example, if several loci are modelled then sampling all loci simultaneously to update $\mathcal{G}_i$ will be time consuming. It is much easier and quicker to sample each gene individually using the distribution conditional on the current states

of any other linked genes. This makes it much easier to 'scale' the procedure to handle multiple linked loci. There is the drawback that if there is a high degree of dependency between sampling sub-units (which will be the case with tightly linked loci), then the movement of the Markov chain through the available sample space will be slowed down, making the Gibbs sampler less efficient (Neal, 1993).

The order of sampling is not rigid and can, in fact, be determined at random for each sampling cycle. The order can, however, be very important with regards the speed that the Markov chain moves through the sample space. For example, when sampling multiple linked loci, this could be done in two ways, either sampling individuals within loci or sampling loci within individuals. Either sampling scheme is valid, but the second way appears to produce much quicker convergence of haplotype frequencies with the datasets used in this study, and was the scheme used for the analyses described here.

After an initial 'burn in' period of m cycles, current values for parameters of interest and their predicted values are stored every k rounds of the process. For example, if we were interested in estimating the QTL effect $\boldsymbol{\lambda}$ then at every k rounds of sampling the current realization of $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\lambda}}$ (7.10) would be stored. These values are used in estimating the posterior density of $\boldsymbol{\lambda}$ (Section 7.2.4). The process is continued until N samples have been stored. The values of m, k and N depend on the data structure and model, in general the larger the dataset and the more complex the model, the larger all these parameters will be. If m is too small then the samples may still be influenced by the starting values used, if k is too small then subsequent samples will be correlated and if N is too small then the estimated distribution will be inaccurate. For all the analyses used here, k was set to 10 but m and N were determined on a *post-hoc* basis for each dataset.

7.2.4 Density estimation

After the sampling is completed, there will be N samples of the parameters of interest. These are samples from the marginal distribution of that parameter,

i.e., the distribution irrespective of all other model parameters. An estimate of the parameter can be made simply by calculating the mean, mode or median of these samples. More information can be obtained by estimating the entire posterior marginal density of the parameter (see Section 3.3.2). This can be done by calculating the density for a value of the parameter averaged over all of the sampled values. Alternatively instead of the sampled values themselves, samples of the conditional distribution of the parameter can be used. For example estimates of the density of the QTL effects $\boldsymbol{\lambda}$ could either be obtained using $\boldsymbol{\lambda}^{(i)}, i = 1, \ldots N$, or

$$p(\boldsymbol{\lambda}|\boldsymbol{\alpha}^{(i)}, \boldsymbol{\beta}^{(i)}, \mathbf{Q}^{(i)}, (\sigma_e^2)^{(i)}, \boldsymbol{y}^{(i)}), i = 1, \ldots N. \tag{7.30}$$

Using (7.30) will always result in an estimate with a lower variance (Gelfand and Smith, 1990; Liu *et al.*, 1994). The estimate of the marginal density is obtained by averaging (7.30) for a range of values of the parameter. For example, an estimate of the marginal density of the QTL effect $\boldsymbol{\lambda}$ would be:

$$\hat{p}(\boldsymbol{\lambda}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi\tilde{\sigma}_{e_i}^2}} \exp\big((\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda})'(\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda})/2\tilde{\sigma}_{e_i}^2\big) \tag{7.31}$$

while an estimate of the marginal density of σ_e^2 would be:

$$\hat{p}(\sigma_e^2) = \frac{(\nu_e/2)^{\nu_e/2}}{\Gamma(\nu_e/2)}(\sigma_e^2)^{(-\frac{\nu_e}{2}-1)}\frac{1}{N}\sum_{i=1}^{N}(\tilde{\sigma}_{e_i}^2)^{\nu_e/2}\exp\Big(-\frac{\nu_e\tilde{\sigma}_{e_i}^2}{2\sigma_e^2}\Big) \tag{7.32}$$

where $\nu_e = n - 2$.

A check on the accuracy of the density estimation can be obtained by fitting a spline-smoothed curve through the estimated points of the marginal density, and numerically integrating across the whole range to verify that the total density is close to 1. Numerical integration can also be used to yield approximate confidence limits for the parameters.

7.2.5 Simulations

To illustrate the operation of the method, several simulated datasets were generated with the program described in Section 4.3.1 using a range of data structures

and parameter values, and these were subsequently analysed. The data structures simulated were similar to that of the mouse X Line experiment (described in Section 4.1.3) so that it could be seen how well the method performed with this type of dataset. The base population for all of the simulations was an F_2 formed from a cross between two inbreds. The F_2 was then split into a number of selection lines, half of which were selected upwards and the other half selected downwards for n generations. Within each line there were 8 full-sib families/generation and 8 individuals/family. The best 2 individuals from each family were selected. For all the simulations, the polygenic variance (σ_a^2) and the environmental variance (σ_e^2) were set to 50.

The first set of simulations demonstrates the analysis for the simplest genetic model, where there is a single marker linked to a single QTL and an unlinked polygenic effect. Several datasets were generated, varying the QTL effect, recombination frequency between the QTL and the marker and the mode of action of the marker (recessive/dominant). For all of these simulations 10 generations of data and 6 lines (3 in each selection direction) were simulated. Estimates were obtained of the additive and environmental variance components, QTL effects (a and d) and the recombination frequency between the QTL and the marker.

The second set of simulations demonstrates the more complex genetic models where there are several markers linked to the QTL. For this set, a single linkage group of 4 linked markers spaced at 10 cM intervals was simulated, with the first marker positioned 10cM and the last 40cM from the end of the chromosome. One QTL was simulated at position 15cM. For these examples only 5 generations of data and 2 lines (1 in each selection direction) were generated. The QTL effect was always additive with an effect (half the distance between the homozygotes) of 10 or 5 units. The analyses were carried out using different subsets of the available marker information, as detailed below:

1. Information from all 4 markers available for all individuals. QTL effect set to 10 units.

2. Information from all 4 markers available for all individuals. QTL effect set to 5 units.

3. Information from 1 marker only (at 20cM) used.

4. Information from all 4 markers available for individuals in generation 5, but information available for only one marker (at 20cM) for the rest of the individuals.

Each of the cases listed above was replicated 6 times. Estimates were obtained of the same quantities as for the first set of simulations, except that the QTL map position was estimated rather than recombination frequencies. The analyses using only 1 marker (case 3) was carried out to (a) compare the standard errors of the estimates when single *vs.* multiple markers are used and (b) to demonstrate that the posterior density of map position using a single marker is symmetrical about the marker position.

7.3 Tests on simulated data

7.3.1 Single marker analyses

The results from the simulation analyses with single marker-QTL systems (Table 7.1) show the means and standard deviations from 6 separate simulation runs using a range of parameters for the simulations. Each line of results in Table 7.1 is therefore from a single simulation run. It can be seen that the method performs well in estimating the parameters with all of the estimates being within 2 standard errors of the simulated values.

7.3.2 Multiple marker analyses

The multiple marker analyses estimated QTL map position rather than recombination frequency relative to a number of linked markers. As described earlier in

Table 7.1: Analysis of simulated data of a QTL linked to a single marker showing simulated parameter values and the means and approx. standard errors of the parameters.

Simulated Parameters					Estimated Parameters				
σ_a^2	σ_e^2	a	d	r	$\hat{\sigma}_a^2$	$\hat{\sigma}_e^2$	$\hat{a}$	$\hat{d}$	$\hat{r}$
(a) Additive Marker									
50	50	10	0	0.1	50.2 (2.4)	49.9 (1.5)	9.6 (0.5)	1.3 (0.7)	0.11 (0.02)
50	50	10	10	0.1	55.0 (2.6)	49.8 (1.6)	9.9 (0.4)	9.3 (0.5)	0.09 (0.01)
50	50	0	10	0.1	51.8 (2.5)	49.2 (2.2)	0.5 (0.3)	9.5 (0.4)	0.10 (0.01)
50	50	10	0	0.02	49.2 (2.3)	49.2 (1.5)	9.5 (0.4)	0.3 (0.5)	0.03 (0.01)
(b) Recessive Marker									
50	50	10	0	0.1	49.9 (2.3)	46.2 (1.5)	10.6 (0.5)	0.4 (0.8)	0.14 (0.02)
50	50	10	10	0.1	50.5 (2.4)	48.2 (1.6)	10.1 (0.5)	9.8 (0.6)	0.11 (0.01)

the method, 4 separate dataset types with 6 replicates for each type were used. Type 1 datasets had information on 4 markers available for every animal in the pedigree. Type 2 datasets were the same, except that the gene effect was half that for type 1 datasets. Type 3 datasets had information available only for 1 marker and type 4 datasets had information available for 1 marker throughout the pedigree, and the other 3 markers for animals in the last generation only. The results from the analyses of these datasets are presented in Table 7.2, which gives the average mean estimates and average standard deviation of the estimates for each dataset type. Estimated posterior densities of QTL position are shown for one each of the replicate datasets of type 1 and 3 in Figure 7.1. The distribution from dataset type 3, the single marker case, is perfectly symmetrical - this comes from the definition of the posterior density and is not a reflection of the methods

Table 7.2: Analysis of 4 types of simulated datasets with a single QTL and 4 linked markers spaced at 10cM showing the simulated parameter values, means and approximate standard errors of the parameters. The results for each dataset type are the averaged results from 6 replicates.

Type	Parameter	True Value	Estimate	s.e.
(1)	σ_a^2	50	53.8	7.5
(2)			52.0	7.4
(3)			48.3	7.4
(4)			56.4	8.2
(1)	σ_e^2	50	52.2	4.6
(2)			49.6	4.4
(3)			48.7	4.4
(4)			48.3	4.7
(1)	a	10	9.39	0.97
(2)		5	5.31	0.89
(3)		10	10.53	1.03
(4)		10	9.27	1.01
(1)	d	0	-0.01	1.13
(2)			0.19	1.05
(3)			-0.22	1.24
(4)			1.19	1.21
(1)	x†	15	14.6	1.5
(2)			14.7	1.5
(3)			13.2	2.8
(4)			15.0	1.5

†The mean and s.e. are only shown for the major mode. In most of the multiple marker datasets the major mode accounted for $> 95\%$ of the posterior density.

ability to switch between the two modes! The distribution from dataset type 1, the multiple marker case, also has two modes, though most of the density (96%) is concentrated in the mode between the markers at 10 and 20 cM. As with the first set of simulations, the method appears to be able to produce satisfactory parameter estimates, with the estimates all being within 2 standard deviations of the simulated values.

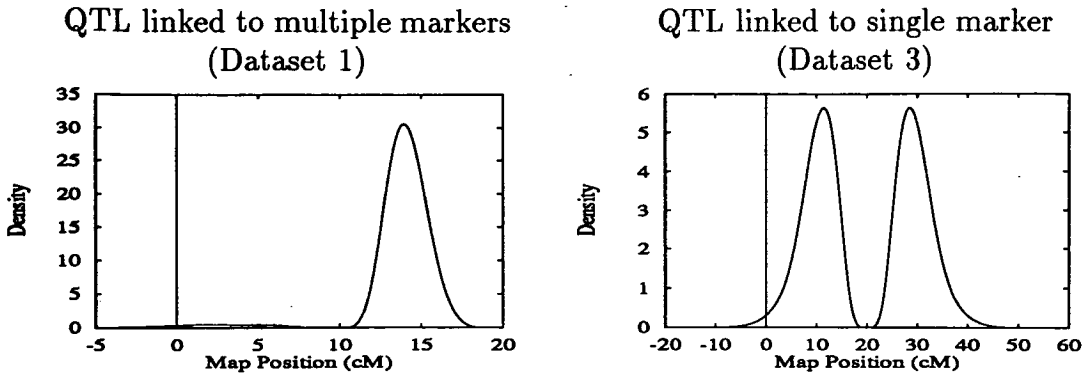


Figure 7.1: Estimated posterior distribution of map position of a QTL from single replicates from Datasets 1 and 3.

7.4 Discussion

The method described in this chapter allows the analysis of large pedigreed datasets using a mixed inheritance model producing estimates of the effect and position of a QTL linked to genetic markers. The method presented here is shown to work well at estimating the effects of QTL linked to single and multiple markers for large complex pedigrees. With single markers, despite the confounding of the polygenic additive effect, QTL effect and recombination frequency, the simulation results show that the method can successfully disentangle these parameters when given the correct model. The ability to do this depends on having multiple generations; in an F_2 population, for example, QTL effect and position are completely confounded when only a single marker is used. It is interesting that there were no large differences detected in the estimated standard errors between the different analyses, except for σ_e^2 which has a larger standard error when $a = 0$ than when $a = 10$. Even when using recessive rather than fully informative markers there does not seem to be a noticeable drop in the precision of estimation. These results are, however, all from single simulation runs; it is not possible to draw firm conclusions about the relative sizes of the standard errors without replicating the simulation experiment.

The multiple marker analyses also show good agreement between the estimates

and the simulated values. Unlike the previous set of examples, the multiple marker analyses were replicated, so some inferences about the standard errors of the estimates can be made. For the QTL position, all datasets gave very similar standard errors except for dataset type 3 when data from only a single marker was used. In this case the standard error of the estimated position was almost twice that when multiple markers were used. Note that when multiple modes were present, the standard error was calculated *within* the largest mode. The larger error when a single marker was used is not, therefore, simply a reflection of the posterior distribution being bimodal.

The standard errors for the variance component estimates appear to be fairly similar, though there seems to be an indication that they might be slightly higher with dataset type 4, which has information on a single marker available throughout the pedigree and multiple marker information available only at the end of the pedigree. It is peculiar that analyses on this dataset type appear to perform less well than with dataset type 3, which has only a single marker so has less information. More replicates would, however, need to be done to test whether this difference was significant.

The standard errors for the QTL effects show a peculiar pattern with the ranking between the database types being type 2 < type 1 < type 4 < type 3. The fact that the standard errors for dataset types 1 and 2 are less than for type 4 which is less than for type 3 is not suprising given the differing amounts of marker information that is present in each dataset type. It *is* surprising that the standard errors for type 2 are smaller than for type 1 datasets, given that type 1 datasets have a larger QTL effect. It seems counter-intuitive that a smaller QTL effect should be estimated more precisely than a larger one, given that the smaller the effect the harder it is to estimate an animal's genotype. The differences are not very large, however, so more replicates are required to see if the effect is real.

Further work that should be carried out would be to assess the method's robustness, in particular to fitting the 'wrong' model. QTL mapping procedures

are prone to giving spurious answers when, for example, two QTL are present on a chromosome and only one is fitted. Convergence is another area on which more work could be done. Convergence detection at the moment is carried out on a *post hoc* basis. To produce a program for general use, it would be better if the number of rounds to convergence could be determined automatically. A method to do this is described by Raftery and Lewis (1994). Decreasing convergence times could also be a profitable area for future work. Datasets with a large number of animals with missing marker data tended to require many more rounds of sampling before convergence was achieved. Large differences in convergence speed can be obtained by using different starting configurations for the genotype structure. It is possible that much improvement in convergence times could be made by using better starting configurations.

7.5 Conclusions

Although the type of dataset analysed here is quite unusual since marker typing is normally expensive, the method could be employed in situations where only a proportion of animals in a pedigree are typed. This would allow the use of marker information to, for example, increase the accuracy of genetic evaluations of animals as well as to help estimate the size and position of QTL as in the present study. The method is fairly general, the major assumption being that the base population should be formed from an inbred cross. It would be simple to modify the procedure to allow other types of base population, although obviously the less that is known about the structure of the base population the less information can be obtained from the analysis. The main problem that would have to be resolved if the base population structure was altered would be the handling of multi-allelic loci. Multi-allelic markers can be very useful in mapping studies because a high proportion of families will be informative. The drawback is that, as discussed in Section 3.2, with more than two alleles the Markov chain becomes 'reducible'. That is, the genotype sampling procedure can become 'stuck' in a

subset of the parameter space because to move from one area to another requires changing more than one individual at once. There have been methods proposed to handle this situation, but it would complicate the sampling procedure. The method could also be easily extended to handle multiple linked QTL and 'floating' markers. In this case it would be necessary to impose some order restrictions on the floating loci to prevent equivalent loci from exchanging positions along the chromosome. Theoretically it should be possible to combine prior information about the marker positions with information from the data, however calculating the weighting to be given to the two data sources would be very difficult. The next chapter describes the application of the single marker method presented in this chapter to the analysis of the QTL associated with the coat colour loci *brown* and *dilute* using the X Lines dataset.

Chapter 8

The Analysis of Putative QTL Linked to the *brown* and *dilute* Coat Colour Loci of Mice

8.1 Introduction

It has long been noted that body weight in mice appears to be associated with coat colour (Green, 1931; Feldman, 1935; Castle, 1941a; Castle, 1941b; MacArthur, 1944b; Butler, 1954; Hedrick and Comstock, 1968). This could be explained either by the coat colour genes having a direct effect on body weight, or by the coat colour genes being linked to QTL for weight. The original founder strains for the X Lines, C57BL/6J and DBA/2J, differed at the two coat colour loci, *brown* and *dilute*, so both loci were segregating in the F_2. The X Lines data could therefore be used as a test of any association between *brown* and *dilute* and body weight.

This chapter presents two methods to estimate the effects and position of putative QTL linked to *brown* and *dilute*, firstly using information from the regression of the estimated effects of the coat colour markers against generation, and secondly using the Gibbs sampling method described in Chapter 7. The regression analysis estimates the decay of the associated effect of the markers over time. This decay is assumed to be caused by the breakdown of linkage

disequilibrium between the marker and a QTL affecting the trait. The rate of decay provides information about the recombination frequency between the marker and the putative QTL, and the intercept of the decay curve with the y-axis gives information about the QTL effect. The Gibbs sampling analysis fits a mixed inheritance model to the X Lines data thereby partitioning the genetic variance in body weight in the X Lines into two components, that due to putative QTL linked to the coat colour loci, and that due to a residual polygenic effect. The basic strategy is to estimate the QTL and marker genotype of each animal, and then use this information to estimate the QTL effect and position. Unlike the simple regression analysis, the mixed inheritance model analysis accounts for the effects of inbreeding and selection on the genotype frequencies of the QTL, as well as allowing for the stochastic changes of QTL and marker frequency within each replicate.

8.2 Estimation of allele frequencies for *brown* and *dilute*

The two coat colour genes under investigation, *brown* and *dilute*, are both recessive, each having two distinct phenotypes. The genes act independently on the coat colour phenotype producing 4 distinct phenotypic classes, wildtype, brown, dilute and brown dilute. All the animals in the X Lines dataset were scored for coat colour and 6-week weight. Any association between the coat colour genes and body weight would be expected to lead to a change of frequency of the colour alleles under selection for body weight. The frequency of both *brown* and *dilute* in the F_2 can be taken as 0.5 because the alleles were both fixed in the original founder strains. Because the genes are both recessive it is not possible to calculate the gene frequencies without making some assumptions about the distribution of genotypes. The simplest estimates of gene frequency can be obtained by assuming that the genotypes throughout the experiment are

Table 8.1: Estimated frequency of *brown* and *dilute* at generation 20

	brown		*dilute*	
Rep.	High	Low	High	Low
1	0.74	0.52	0.20	1.00
2	0.98	0.00	0.25	0.72
3	0.66	0.24	0.00	0.66
4	0.27	0.00	0.00	0.84
5	0.62	0.00	0.49	1.00
6	0.39	0.00	0.39	0.70

in Hardy-Weinberg equilibrium. An estimate of the gene frequency of a particular colour allele can then be made by treating the frequency of animals displaying that colour phenotype as an estimate of p^2 where p is the frequency of the gene in question.

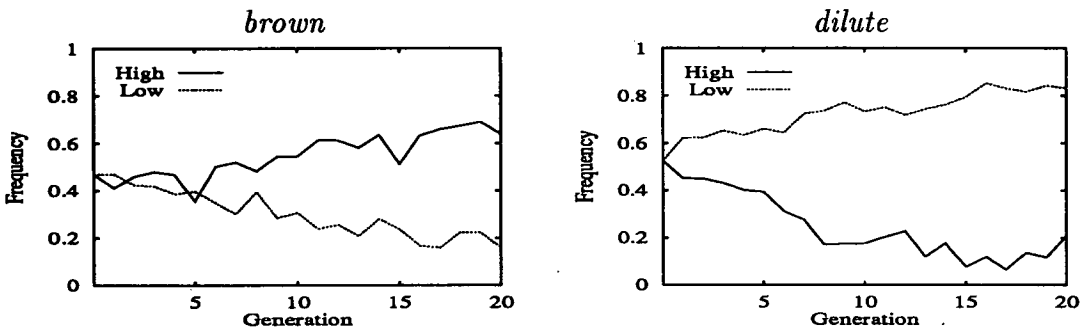
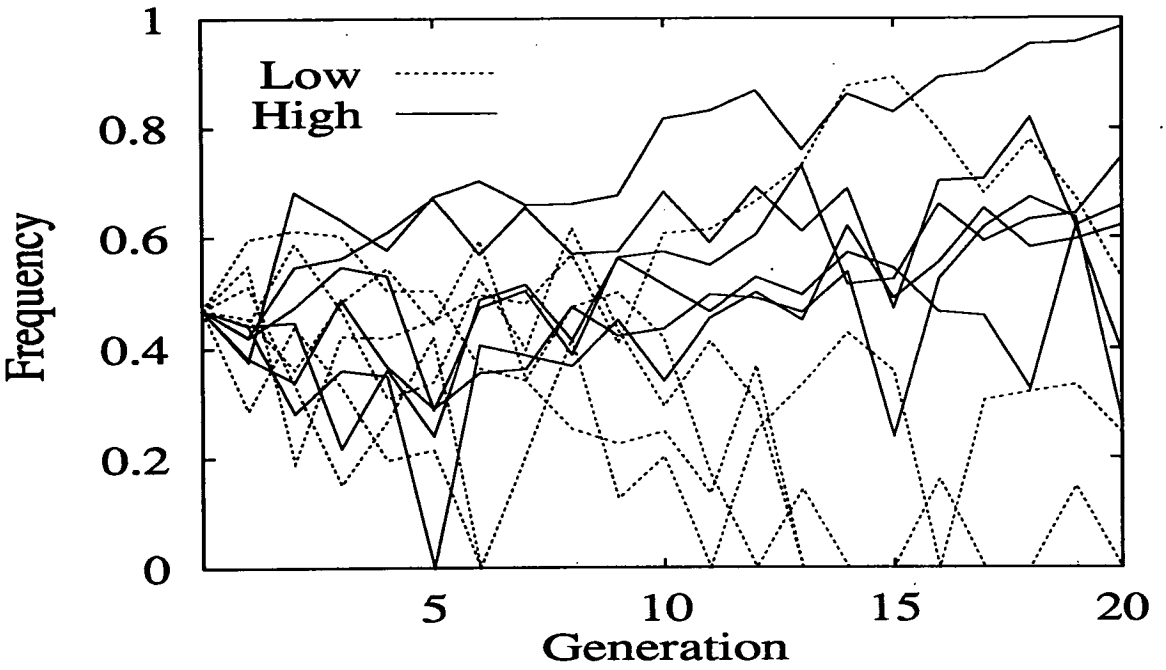


Figure 8.1: Average estimated frequency of *brown* and *dilute*

The assumption of the genotypes being in Hardy-Weinberg equilibrium is unrealistic due to the effects of inbreeding and selection. The gene frequency estimates should take account of the expected effects of inbreeding on the genotype frequencies, but accounting for selection would be much harder since the effect of selection on genotype frequencies depends on a range of factors such as population size, selection intensity, size of the effect associated with the locus etc, therefore neither correction was done here.

The estimated frequencies of *brown* and *dilute* averaged over replicates for the High and Low selection directions are shown in Figure 8.1. When the individual

brown



dilute

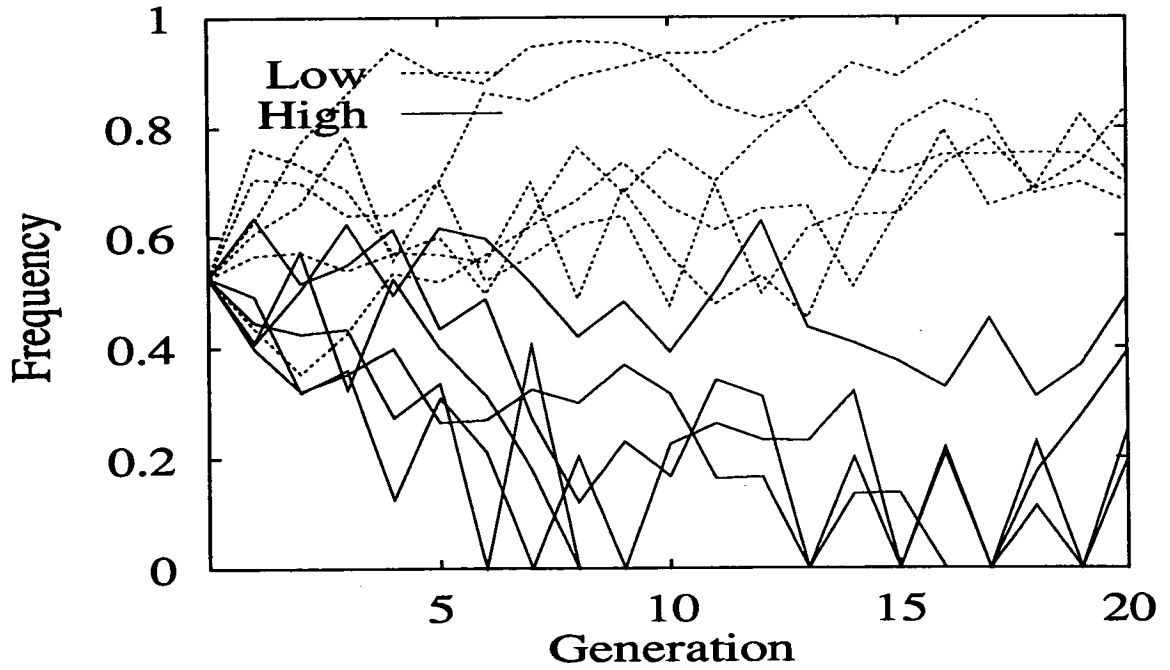


Figure 8.2: Replicate frequency of *brown* and *dilute*

lines are examined it is found that whilst the changes in frequency for *dilute* are fairly consistent across all replicates, the same is not true for *brown* where several replicates show frequency changes in the opposite direction to the other lines being selected in the same direction (Figure 8.2 and Table 8.1). The most likely explanation for this behaviour is that *brown* is only loosely linked to a region affecting body weight so that recombinations between *brown* and the region occur reasonably frequently.

8.3 Estimation of the effects of *brown* and *dilute*

The simplest way to obtain estimates for the effects of *brown* and *dilute* would be to assume that the coat colour alleles themselves have an effect on body weight and analyse them as fixed effects. The problem with this method is that if the model is incorrect, and in fact the marker has no direct effect but is linked to a QTL affecting body weight, then the associated effect of the marker will change from generation to generation as the linkage disequilibrium between the marker and QTL present at the beginning of the experiment breaks down. Because of these considerations, a model was fitted with *brown* and *dilute* as nested effects within generation so that separate estimates were produced for each generation. The analysis was done using DFREML fitting a similar model to that used in the variance component analysis (Section 5.6), except for the addition of the marker effects already discussed.

8.3.1 Inference of the effect and position of a linked QTL

With this experimental design, if a marker were linked to a QTL affecting body weight then initially we would expect to see an effect associated with that marker, because the two loci would be in linkage disequilibrium. Over time, however, this effect would be expected to decay as recombination breaks up the association, with the rate of decay being determined largely by the recombination frequency

between the QTL and the marker.

By fitting an exponential curve to the estimated effects and making a number of simplifying assumptions it is possible to use the rate of decay of the curve to make a rough estimate of the recombination frequency, and the intercept of the curve with the y-axis can give information on the size of the effect. Assuming a model with a marker linked to a single QTL (see Section 4.2 for details), then several conclusions can be made. In this model, the marker-QTL genotype frequencies are determined solely by the frequencies of the marker and QTL alleles and the disequilibrium between them (*i.e.*, ignoring the effect of inbreeding and non-random mating) and the gene frequencies are assumed to remain constant at 0.5. Linkage disequilibrium (D) is defined as the difference between the equilibrium and actual gametic frequencies and can be estimated by $f(\mathrm{MQ})f(\mathrm{mq}) - f(\mathrm{Mq})f(\mathrm{mQ})$ where $f(\mathrm{xx})$ is the frequency of gametic type xx, M & m refer to the two alleles at the marker locus and Q & q refer to the two alleles at the QTL locus. The disequilibrium at generation t, D_t, is given by $D_t = D_0(1-r)^t$ where D_0 is the disequilibrium in the F_2 and r is the recombination frequency. To estimate D_0 requires having estimates of the gametic frequencies in the F_2. For the X Lines these are $(1-r)/2$ for the non recombinant gamete types (MQ and mq), and $r/2$ for the recombinant types (Mq and mQ). The expected value of D_0 will therefore be $(1-r)^2/4 - r^2/4 = (1-2r)/4$.

Table 8.2: Expected gametic frequencies under simple model

Gametic types	MQ	Mq	mQ	mq
Expected freq.	$\frac{1}{4} + D_t$	$\frac{1}{4} - D_t$	$\frac{1}{4} - D_t$	$\frac{1}{4} + D_t$

The expected marker-QTL gametic frequencies at generation t are given in Table 8.2. The expected frequencies are calculated assuming gene frequencies remain constant at 0.5. There are two marker phenotype classes M and m with frequencies 0.75 and 0.25, and the expectations of each marker class in terms

of a and d can be obtained by calculating the expected genotype frequencies at the QTL within each marker class. The expected frequencies and expectations in terms of a and d of each marker-QTL genotype and for each marker class are given in Tables 8.3 and 8.4 respectively.

Table 8.3: Expectations of genotype frequencies

Marker Type	Genotype	frequency	a	d
M	MQ/MQ	$(\frac{1}{4}+D_t)^2$	1	0
	MQ/Mq	$2(\frac{1}{4}+D_t)(\frac{1}{4}-D_t)$	0	1
	MQ/mQ	$2(\frac{1}{4}+D_t)(\frac{1}{4}-D_t)$	1	0
	MQ/mq	$2(\frac{1}{4}+D_t)^2$	0	1
	Mq/Mq	$(\frac{1}{4}-D_t)^2$	-1	0
	Mq/mQ	$2(\frac{1}{4}-D_t)^2$	0	1
	Mq/mq	$2(\frac{1}{4}+D_t)(\frac{1}{4}-D_t)$	-1	0
m	mQ/mQ	$(\frac{1}{4}-D_t)^2$	1	0
	mQ/mq	$2(\frac{1}{4}+D_t)(\frac{1}{4}-D_t)$	0	1
	mq/mq	$(\frac{1}{4}+D_t)^2$	-1	0

Table 8.4: Expectations of marker classes for recessive markers

Marker class	frequency	a	d
M	$\frac{3}{4}$	$\frac{4}{3}D_t$	$\frac{1}{2}+\frac{8}{3}D_t^2$
m	$\frac{1}{4}$	$-4D_t$	$\frac{1}{2}-8D_t^2$

By assuming that the QTL acts additively, *i.e.*, that $d = 0$, the expected difference between the two marker classes in terms of a will therefore be:

$$\text{diff} = (\frac{4}{3}D_t + 4D_t)a = \frac{16}{3}D_t a = \frac{16}{3}D_0(1-r)^t a \qquad (8.1)$$

Estimates of r and a can be obtained by fitting an exponential curve of the form $y = AB^t$ to the REML estimates of the effects of the markers in each generation

(*i.e.*, the difference between the two marker classes). The curve parameter B will give an estimate of $(1 - r)$ and the parameter A will give an estimate of $\frac{16}{3}D_0a$. An estimate of the standard errors could be made using the standard errors of the regression, but this is likely to be an underestimate because of the many simplifications made in deriving (8.1), and because the regression analysis assumes that the errors between generations are uncorrelated and remain constant which would not be the case. The error variance from the regression would be expected to increase over time for several reasons. The variance of gene frequency increases with the inbreeding coefficient, and so would increase over time as the population became more inbred. The variance of D_t would also increase over time producing another source of increasing error variance. This means that the early generations contain more information about the gene effect and position than the later generations, so the regression should have been weighted to take more account of the data from the early generations.

The estimates of the associated effects of *brown* and *dilute* in each generation obtained from the DFREML program are shown in Figure 8.3. In the case of *dilute* it can be seen that the effect appears to decay over time, while this pattern is not readily apparent for *brown*.

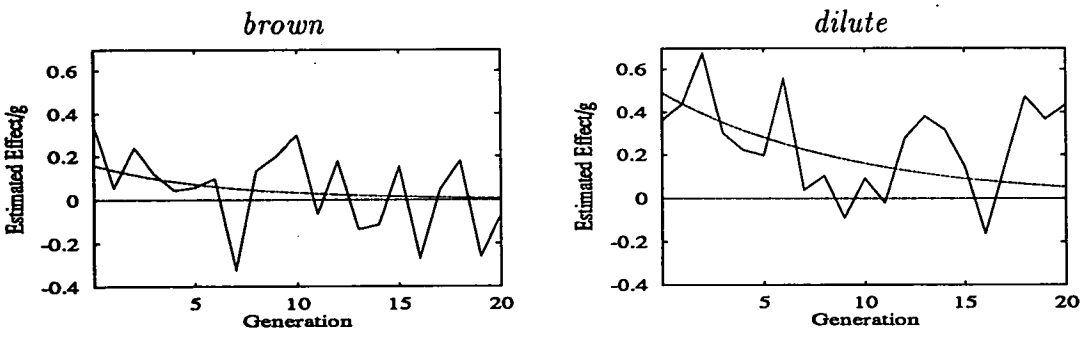


Figure 8.3: REML estimates of the effects *brown* and *dilute* along with the fitted exponential decay curve

To obtain a more realistic idea of the standard errors that could be expected applying this analysis to a dataset of the size and structure of the X Lines data a series of simulated datasets were generated using the program described in Section

4.3.1. The data structure simulated was similar to that of the actual X Lines with the main difference being that there was no variation in litter size with all litters being of size 8. The model simulated was

$$y = \mu + \lambda + \beta + e \tag{8.2}$$

where μ is the overall mean, λ is the QTL effect, β is the polygenic additive effect and e is an error effect. A marker was simulated as being linked to the QTL with a recombination frequency r. Twenty replicate datasets were generated and then analysed with the above procedure (analysing with DFREML to obtain the marker effects in each generation, fitting an exponential curve to the effects and estimating a and r from the curve parameters) to get estimates of the QTL effect and recombination frequency between the QTL and the marker. The curve was fitted using the non-linear regression function `FITCURVE` in Genstat 5.3 (Genstat 5 Committee, 1993). Each effect was weighted by the inverse of the variance of the REML estimate. The same genetic parameters were used for each simulation; $\sigma_a^2 = 1.0$, $\sigma_e^2 = 3.0$, $a = 0.5$ and $r = 0.1$, these being similar to the values estimated from the X Lines dataset. The results of the analysis of the simulated data are given in Table 8.5. There are several points to note

Table 8.5: Regression analysis of simulated data. The average estimates of the effect (a) of a QTL linked to a marker with recombination frequency (r).

Parameter	Simulated Value	Average Estimate	Mean Replicate Standard Error	Empirical Standard Error
a	0.5	0.59	0.11	0.26
r(%)	10	15	5	7

about the results. The mean results are within one standard deviation from the simulated values, however the empirical standard errors are large with respect to the effects, indicating that this test would not be very powerful at detecting effects of this size in practice. The mean standard errors of the individual replicate estimates were calculated from the variance of the regression coefficients and are very approximate because of the many simplifications made in the model detailed

Table 8.6: Regression estimates of the additive effect (a) in grams of a putative QTL linked to *brown* and *dilute* and the recombination frequency (r) between marker and QTL.

	A†	B	a(g)	s.e.(g)‡	r(%)	s.e.
brown	0.158	0.852	0.17	0.14	15	17
dilute	0.492	0.894	0.47	0.10	11	4

† A & B are the estimated parameters of the curve $y = AB^t$ where y is the estimated effect of the markers and t is the generation.
‡ Calculated from the standard errors of the regression coefficients.

earlier, and because the variances of the regression coefficients were calculated assuming that the errors were uncorrelated and constant. Assuming that the empirical standard errors are more likely to reflect the 'true' uncertainty about the parameter estimates, it appears that, not surprisingly, the replicate standard errors are an underestimate of the actual standard errors. More seriously, several of the individual replicates produced answers which are outwith two standard errors of the simulated values; this again indicates that this analysis would not be reliable as a means of estimating linked QTL effects of the size simulated here.

The parameters of the fitted curves and the estimated values of a and r (with the standard errors from the regression) for the X Lines data are given in Table 8.6. Following the results of the analysis of the simulated data (Section 8.3.1), the standard errors are likely to be underestimates. Given the probable size, therefore, of the standard errors of the estimates, few conclusions can be drawn about any effect linked with *brown*. It is possible, however, to make some tentative conclusions about *dilute* in that it does appear to be linked to a region affecting bodyweight with an estimated additive effect of about 0.5g situated roughly 10cM from *dilute*, though the confidence limits for these estimates would be large.

8.4 Gibbs sampling analysis

The initial analysis undertaken using the Gibbs sampling method was to compare the results described in Chapter 7 in a standard variance components analysis (without fitting any QTL) against a conventional REML analysis, fitting the same model for both methods. This also allowed the comparison of the variance components with and without fitting QTL to the model allowing an estimate to be made of the contribution of the QTL to the observed variance. The REML analysis used the derivative-free REML packages of Meyer (1988; 1989). The model was the same as that used for the initial homogenous variance component analysis described in Section 5.6 with the additive genetic animal effect and litter fitted as random effects and generation, parity, litter size and sex nested within line and generation fitted as fixed effects. Sex was fitted as a nested effect because a significant change in sexual dimorphism was noted in the Low lines over the course of the experiment (Section 5.3). For the analyses, the untransformed data were used because there was little difference found between estimates of genetic parameters using log transformed or untransformed data, and it is easier to interpret the estimated QTL effects using untransformed data. The animal and litter effects were uncorrelated; the covariance matrix for the animal effect was the numerator relationship matrix **A** and for the litter effect was the identity matrix **I**.

The model was then extended to fit a QTL linked to *brown* and *dilute*. The markers *brown* and *dilute* have been mapped to different chromosomes (*brown* on chromosome 4 and *dilute* on 9) so it was possible to analyse them without considering the possibility either of linkage between the two markers or that they were both linked to the same QTL. The QTL linked to *brown* and *dilute* were analysed separately due to computing limitations.

The results of the analysis of the X-Line data are given in Table 8.7. These show the results of a conventional REML analysis of the data and a Gibbs sampling analysis using the same model as the REML analysis. The REML and Gibbs

Table 8.7: Analysis of 6 week body weight data from the X Line selection experiment using (a) a REML variance component analysis and (b) a Gibbs sampling based variance component analysis. The means and approx. standard errors of the parameters are given.

Analysis	$\sigma_a^2(g^2)$	s.e.(g^2)	$\sigma_c^2(g^2)$	s.e.(g^2)	$\sigma_e^2(g^2)$	s.e.(g^2)
REML	1.35	0.06	1.75	0.07	1.68	0.03
Gibbs	1.31	0.06	1.81	0.08	1.74	0.04

sampling variance component analyses give very similar results, as expected since the same model was used in both cases. The estimates are not expected to be exactly the same since the REML estimates are from the joint mode of the posterior density of the variance components (assuming flat priors for the fixed effects and variance components), while the Gibbs estimates are from the posterior marginal means (Gianola and Foulley, 1990).

The results of the analyses including a single QTL linked to either of the coat colour loci *brown* or *dilute* are given in Table 8.8. The estimated marginal posterior densities for the QTL effects and the recombination frequency between the QTL and the marker are shown in Figure 8.4. It can be seen from Tables 8.7 and 8.8 that fitting the QTL causes a reduction in the additive and environmental variance components, but seems to have no effect on the common environmental (litter) variance component.

Table 8.8: Analysis of putative QTL linked to *brown* and *dilute* fitting a mixed inheritance model. The means and 95% confidence intervals of the model parameters are given.

Marker Locus	σ_a^2 (g^2)	σ_c^2 (g^2)	σ_e^2 (g^2)	a (g)	d (g)	r (%)
brown	1.11 (1.01,1.27)	1.80 (1.65,2.01)	1.34 (1.27,1.44)	-1.70 (-2.03,-1.38)	1.91 (1.52,2.31)	45 (38,49)
dilute	1.24 (1.14,1.38)	1.80 (1.66,2.01)	1.67 (1.58,1.78)	0.40 (0.11,0.66)	0.65 (0.20,1.01)	12 (9,15)

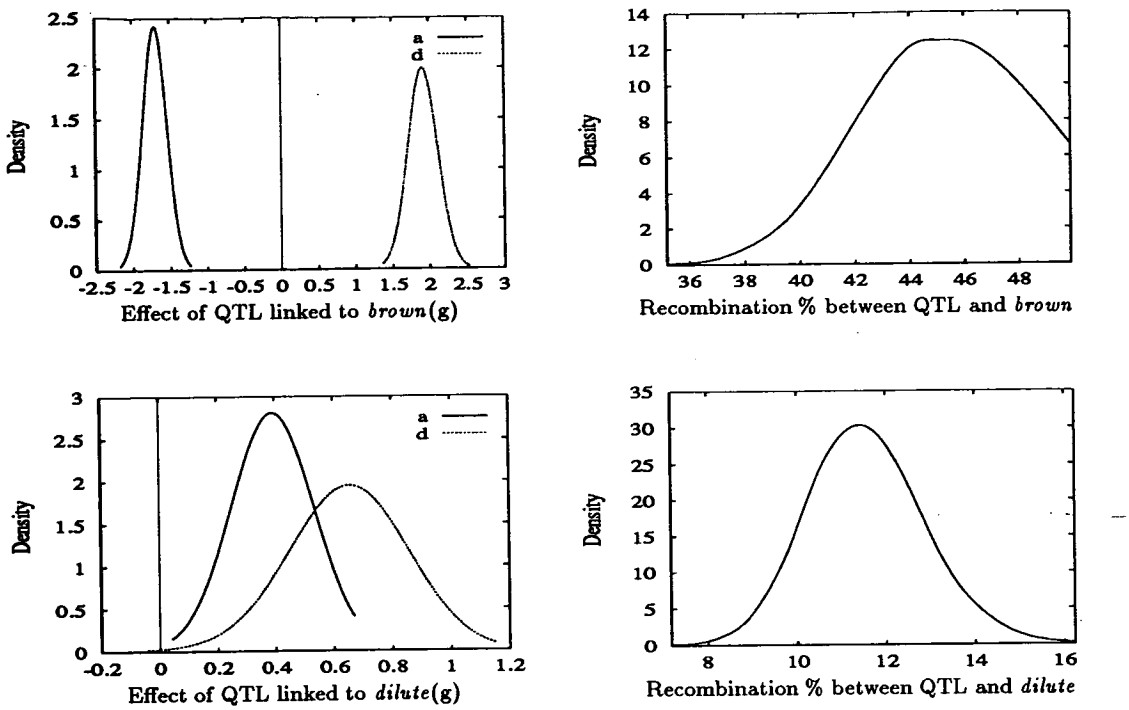


Figure 8.4: Estimated posterior marginal densities for QTL linked to *brown* and *dilute* for QTL effect and recombination frequency

8.5 Discussion

It is noted that body weight and coat colour have been reported to be associated by several previous authors, and the X Lines data are analysed with a view to investigating whether any association can be found between two coat colour genes, *brown* and *dilute*, and body weight. Both alleles show frequency changes under selection for body weight although the frequency changes for *brown* are inconsistent and differ between replicate lines. This could either mean that the alleles had a direct effect on body weight or that they were linked to a region of chromosome that affected weight. A third possibility would be if the animals were being selected on colour rather than on the desired selection criteria. The proportion of times the 'best' animal in a litter (with respect to 6 week weight) was not selected, however, was less than 4%, indicating that the selection had been carried out correctly. A small number of mis-selections is to be expected

because it may not be possible to mate from the 'best' animal due to infertility or health problems.

A simple method to estimate the effect and position of a putative QTL linked to a single marker based on the modelling the decay of the marker's effect due to the breakdown of disequilibrium is described and tested with a replicated simulation experiment. This appears to show that the method could work but the estimates have very large standard errors, so it would probably only be useful when investigating QTL that have a large effect or are tightly linked to the marker locus. The method is used to analyse the *brown* and *dilute* data. The results are highly inconclusive in the case of *brown*, but do give an indication of an effect linked to *dilute*. The method could be made more efficient by taking account of the increase in the error variance in the later generations, and weighting the regression accordingly.

The Gibbs sampling based approach described in Chapter 7 was applied to the same dataset. The analysis was initially performed with and without fitting the QTL. As commented in the results section, it can be seen that the estimates of σ_a^2 and σ_e^2 are reduced when QTL linked to *brown* and *dilute* are fitted. The expected contributions to the additive and dominance variances in the F_2 from a QTL are given by $a^2/2$ and $d^2/4$ respectively (Falconer, 1989). Assuming that the majority of σ_d^2 in the original analysis was attributed to σ_e^2, we can predict that fitting the QTL linked to *dilute* should lead to reductions in σ_a^2 and σ_e^2 of around $0.08g^2$ and $0.11g^2$ respectively. These predictions compare well quite with the observed reduction of $0.07g^2$ in both variance components. For *brown*, however, the reduction in variance when fitting a linked QTL is much less than the predicted contribution to the variances from the QTL, which are $1.44g^2$ for σ_a^2 and $0.91g^2$ for σ_d^2. Note that the predicted contribution to σ_a^2 from the linked QTL is greater than the total additive variance when no QTL is fitted. An explanation for this is that a QTL of large effect becomes fixed very rapidly under selection, so only contributes to the variance in the early generations. When analysing multi-generation data under the infinitesimal model, information from all generations is used to estimate

Table 8.9: Analysis of simulated data; simulating a QTL and analysing under the infinitesimal model.

Parameter	Simulated Value	Average Estimate	Empirical Standard Error
σ_a^2 (g^2)	1.11	1.54	0.10
σ_e^2 (g^2)	1.34	1.54	0.09
a (g)	-1.70	-	-
d (g)	1.91	-	-

the variance in the F_2. The pattern of change in σ_a^2 when a QTL of large effect is segregating is very different from that predicted from the infinitesimal model. Information from the early generations would suggest a large amount of σ_a^2, whereas the later generations would show very little σ_a^2. Analysing such data under the infinitesimal model would therefore tend to produce underestimates of the variance in the F_2. Table 8.9 shows the results of a simulation experiment where a population consisting of 2 replicate lines (1 selected in each direction) was selected for 20 generation. The size and structure of the lines were the same as in the previous simulation experiments. A single QTL, a polygenic additive component and an environmental component were simulated, with the simulated values for these effects being taken from the analysis of the effect associated with *brown* presented in Table 8.8. The experiment was replicated 100 times, with the simulated datasets all being analysed under a purely infinitesimal model. The results show that the estimate of the additive variance in the F_2 is indeed greatly underestimated using the infinitesimal model (the predicted value for σ_a^2 in the F_2 would be $1.11 + 1.70^2/2 = 2.56$g^2). The average results from the simulation analysis are, in fact, not that different from the actual estimates produced from the analysis of the X Line dataset under a purely infinitesimal model (Table 8.7).

8.6 Conclusions

Differences between the estimates produced by the regression and Gibbs sampling methods would be expected due largely to the many simplifications made in the derivation of the regression method. One point to note is that the estimates from the Gibbs sampling method indicated a dominant mode of action for both QTL. The regression analysis assumed that the QTL was additive; this would lead to biased estimates. The results for *dilute* are quite similar for both methods. The regression analysis gives an estimate for the additive effect of the QTL linked to *dilute* of 0.47g with a recombination frequency between the two loci of 11%, whereas the Gibbs sampling analysis produced an estimate of 0.40g and 12%. The estimates from the Gibbs sampling analysis are, however, more precise (remembering that the estimated standard errors from the regression analysis are underestimates). The results from the two methods do differ greatly for *brown*, with the regression analysis producing estimates of the QTL effect and recombination frequency of 0.17g and 15% and the Gibbs sampling analysis producing 1.70g and 45%. The results from the regression analysis had very large standard errors, however, with the gene effect being not significantly different from 0g, and the recombination frequency being not significantly different from either 0% or 50%. It is, therefore, difficult to draw any firm conclusions from these analyses about the QTL linked to *brown*. There is potentially a source of more information which could be used in mapping the QTL linked to *brown* and *dilute*. As discussed in Section 4.1.3, a group of animals from generation 21 of the High and Low selection lines have been typed for a range of genetic markers, several of which are linked to *brown* and *dilute*. These markers also show differences in frequency between the High and Low selected lines. A further analysis could be performed utilizing this extra marker information. Chapter 7 describes the analysis of a simulated experiment where a QTL linked to several markers was simulated. Marker information was available on one marker throughout the pedigree, but information on the other markers was available only at the end of

the pedigree. This is essentially the same situation that occurs in the X Lines, so it should be possible to perform this type of analysis on the X Line data without too much difficulty. The major drawback is likely to be computer time; the simulated datasets, which were much smaller than the X Lines dataset, required several days to produce stable estimates. Analysis of the X Lines would be likely to take very much longer. Another potential problem is that with the simulated data, the proportion of animals with complete marker information is much higher (24%) than in the X Lines dataset (1%). The potential additional information content from the additional markers is reduced still further because the High line samples have no pedigree data due to a procedural error. These factors lead to the conclusion that the additional information from the extra markers is probably small compared to the information from the coat colour markers and the trait data.

Chapter 9

General Discussion and Conclusions

In this thesis are developed several analytical methods which are applied to an experimental dataset from a selection experiment on mice in which an F_2 population is divergently selected on 6-week body weight for 20 generations. The main aim of the analyses was to use the selection experiment data to make inferences about the genetic control of body weight in mice. This is done by (a) detecting deviations from the predictions of the infinitesimal model and (b) explicitly estimating the effects and positions of some of the genes affecting body weight. The methods are also applied to simulated datasets with the same general structure as the experimental dataset, so that the performance of the methods could be assessed.

Chapter 2 presents a review of the infinitesimal model as applied to genetic analysis. The infinitesimal model, for which it is assumed that the genetic variance in a trait derives from an infinite number of unlinked additive genes each of infinitely small effect, is the standard model for quantitative genetic analysis. The reasons for the predominance of the infinitesimal model are ease of use and lack of knowledge about the distribution of gene effects. There are several reasons why real data might diverge from predictions made using the infinitesimal model. The infinitesimal model can be made to account for changes in genetic variance due to inbreeding and linkage disequilibrium caused by selection (the Bulmer

effect), but it cannot account for the changes in variance caused by changes in gene frequency. If there are genes with a significant effect on the trait segregating in the population, then selection and/or drift will cause stochastic changes in gene frequency which will create changes in genetic variance. Deviations from the infinitesimal model brought about by changes in gene frequency are likely to become greater the more generations are considered. This expectation was not borne out by a literature review of selection experiments (Sheridan, 1988) which looked for discrepancies between base population and realized estimates of genetic parameters. The results showed a *better* fit between the two estimates with long-term experiments. This might, however, have been due to better estimation of realized heritability with the longer experiments (Hill and Caballero, 1992).

One way in which the infinitesimal model could be extended to make it more realistic could be to model the genes with larger effects discretely, while modelling the remainder as a residual infinitesimal effect. These mixed inheritance models are arguably more accurate representations of the actual distribution of effects of segregating genes than the basic infinitesimal model. There is evidence that the distribution of the effects of mutations on quantitative traits is highly leptokurtic with a few genes of large effect and many genes of small effect (Keightley, 1994a; Caballero and Keightley, 1994). If the distribution of *segregating* genes was also leptokurtic then a mixed inheritance model would be more accurate than a model of equal gene effects (i.e. the infinitesimal model).

The analysis of data under a mixed inheritance model is more difficult than under the infinitesimal model. The main difficulty arising from the use of mixed inheritance models is that the calculation of the likelihood is intractable for all but very simple pedigree structures (i.e. an F_2 population). Approximations to the mixed inheritance likelihood can be obtained (Hasstedt, 1982; Bonney, 1984), but how well they perform when used to analyse large complex pedigrees is not known (Guo and Thompson, 1994). Mixed inheritance models can, however, be analysed using Monte Carlo sampling techniques.

Chapter 3 presents a brief introduction to Gibbs sampling, a Markov Chain Monte Carlo method, and its applications for genetic analysis. Gibbs sampling is a powerful and flexible technique for data analysis. It is simple to implement on a computer and it can handle complex models with ease. If Gibbs sampling is compared to more conventional approaches, several advantages and drawbacks become apparent. For example Gibbs sampling can be used for variance component analyses, an area where the standard analytical method in animal breeding is REML. Comparing the Gibbs sampling to REML leads to the following conclusions:
Advantages of Gibbs sampling over REML:

- Gibbs sampling is computationally simpler which makes it easier to implement, and it typically requires less memory than a REML analysis.
- More complex models can easily be analysed using Gibbs sampling, including discrete gene and mixed inheritance models.

Disadvantages of Gibbs sampling:

- A Gibbs sampling analysis typically requires more CPU time than an equivalent REML analysis. This time penalty may become less apparent when larger datasets are analysed, and of course Gibbs sampling is most useful where there *is* no equivalent REML analysis.
- It is difficult to predict convergence of the Markov chain. The number of sampling cycles required varies greatly between datasets and models. This makes it very difficult to write a fully 'automatic' Gibbs sampling algorithm.

The initial quantitative analysis of the X Lines experiment is presented in Chapter 8. The analysis shows that there was substantial genetic variance present in the population resulting in sustained selection response over the 20 generations of selection. This is despite the X Lines being derived from a cross of only two inbred strains, although the two strains are thought to be genetically very

different and are possibly from separate sub-species. The cross between them would therefore include much of the genetic variance in the whole population.

There were several unusual features of the X Lines which were noticed during this initial analysis. The first is that subline B of the Low lines showed a markedly different response from the other Low lines, being at one point about 1 s.d. below the other lines (Figure 5.1). This could be due to a mutation or a rare recombination event arising in that subline. It should be possible to search for a new mutation in subline B by looking for molecular differences between subline B and the other Low lines. Such a search would be very difficult because there are likely to be very many genetic differences between the sublines due to drift alone. Samples from all sublines at generation 20 have, however, been typed for a wide range of genetic markers, and so far no allele has been found that is just restricted to subline B (P. Keightley, *pers. com.*).

Another unusual feature was the marked decrease in sexual dimorphism over time which was noticed in the Low lines. There are several possible explanations for the decrease, which are discussed in Chapter 5, but no work was carried out to distinguish between these possibilities. It would be interesting to see how general this effect is. A literature search only uncovered one other reported instance of a correlated change in sexual dimorphism when selecting for body weight (MacArthur, 1944a), and in that case the effect was nonsignificant. Unpublished data from another selection experiment conducted on mice at Edinburgh University also shows apparent reductions in sexual dimorphism over time, but the effect is not as clear cut as with the X Lines (S. Mbaga, *pers. com.*).

The third unusual feature of the X Lines noticed in the initial analysis was the apparent non-linear response of the Low lines, which showed an acceleration in response over the middle portion of the experiment. The infinitesimal model predicts a linear response to selection. A decline in response over time can be caused by the loss of genetic variance due to the fixation of alleles. An increase in the selection response is more complicated to explain, but indicates an increase

in genetic variance for some reason. Chapter 6 presents a method for detecting and estimating changes in variance components under selection over and above what would be predicted using the infinitesimal model. The method was tested on simulated datasets of the same general size and structure as the X Line dataset. This simulation analysis demonstrated that the method could detect decreases in the additive variance produced by changes in gene frequency when a small number (32) of genes controlling the trait were simulated. When the method was applied to the X Lines, it indicated increases in the additive and environmental variance components (significantly so in the Low lines) and decreases in the common environmental component. The pattern of changes was the same in both the High and Low lines, but the magnitude of the changes was much greater in the Low lines. The large increase in variance in the Low lines corresponds to the apparent acceleration in the Low line selection response (although this could be due to environmental change). Several possible models that *could* give rise to the observed variance changes are discussed in Chapter 6, but it is not possible to choose between the models given the available data. One way forward would be to attempt to estimate the direct and interaction effects of individual QTL. If the effects and positions of all of the genes affecting a trait, or at least the genes contributing a large proportion of the genetic variance, are known then this should provide an accurate genetic model of the trait. The current situation is still a very long way from this, but the large amount of work on QTL mapping means that the information required to produce more accurate genetic models is becoming available. Utilizing such information for data analysis, however, will still be very difficult since it would require the modelling of many discrete loci (including their interactions).

The Gibbs sampling approach developed in Chapter 7 for solving mixed inheritance models has potential both as a means of estimating QTL effects and for fitting quantitative genetic models which are more realistic than the infinitesimal model. The method described in Chapter 7 is, of course, a long way from being able to fit a general genetic model with a large number of interacting discrete loci.

It is theoretically quite simple to extend the method to handle several discrete loci and a residual infinitesimal effect taking account of the other loci. In this way, a large proportion of the genetic variance in the trait could be taken account of by the discrete loci. It is likely, however, that this approach would have low power when fitting several loci.

There are several areas in which further work is required on the method. The method could be generalized by relaxing the restriction on the makeup of the initial population. As discussed in Chapter 7, this leads to several problems. With the initial population being formed from an inbred cross, the genotypes of the base animals are completely known. With any other type of base population the amount of information is obviously reduced. A more serious drawback is that with a more general base population, there is the possibility of multi-allelic loci which can cause the Markov Chain to 'stick' in subsets of the parameter space. There are methods that have been proposed to deal with multiple loci (Sheehan and Thomas, 1993; Lin *et al.*, 1993; 1994; Lin, 1995), and these could be incorporated within the existing method. It is not clear how well these methods, which have been developed for human pedigrees, would work with the pedigrees typically found in animal breeding analyses. It appears from the work by Lin *et al.* (1994) that one of the pedigree structures that can lead to irreducibility is half sib families, which are very common in, for example, dairy cattle data, but are less common in human pedigrees.

Other areas where further work could be profitable are looking at ways of increasing the speed of convergence. This can be done by improving the initialization of the unknown parameters; if the initial configuration is a long way from the equilibrium distribution then this can delay convergence considerably. The problem is most acute when setting up the initial genotype configuration when there are animals with missing or incomplete marker information. Another way of improving convergence would be to decrease the autocorrelation so the Markov Chain samples the parameter space more efficiently. This can be done by simultaneous sampling of highly correlated parameters, but in many cases this

is not practical. For example the QTL and polygenic genotypes and the QTL effect are all highly correlated, but simultaneous sampling of all these effects is not practical. The most immediate requirement for the method is to test it more thoroughly, in particular when data are analysed on purpose with the wrong model. Most QTL estimation techniques are vulnerable to bias if there are a different number of QTL on a chromosome than are being modelled, and it is unlikely that this method is any different.

In chapter 8 is described the analysis of the associated effect of the coat colour loci *brown* and *dilute* on 6 week body weight from the X Lines experiment. Both loci show directional changes in frequency over the course of the X Lines experiment indicating linkage with one or more regions of chromosome that affect body weight. The aim of the analyses in this chapter was to model these associated effects as single QTL linked to the coat colour loci, and produce estimates of the QTL effects and the recombination frequencies between the QTL and the marker loci. This was done in two ways. The first analysis used the log regression of the associated effect of the markers against generation to estimate the parameters. The associated effect is created by linkage disequilibrium between the marker loci and the QTL. There is complete disequilibrium in the F_1, as this breaks down due to recombination the associated effect of the markers will decrease. The rate of decay of the effect depends in part on the recombination frequency between the QTL and the marker. An estimate for the QTL effect can be obtained from the associated effect of the marker in the F_2. This analysis relies on many assumptions about the marker effect (additive), gene frequencies (constant), genotype frequencies (in Hardy-Weinberg equilibrium) and the error variance from the regression (homogenous). It is consequently not very accurate, but could be useful as a quick and simple test for a QTL. The analysis indicated a QTL linked to *dilute* with an additive effect of 0.5g and a recombination frequency between the QTL and *dilute* of 11%. The analysis of *brown* did not yield a significant estimate for the QTL effect.

The second analysis used the Gibbs sampling based method previously dis-

cussed. Only single marker analyses were carried out despite there being multiple marker information available on some animals from generation 21. This was for several reasons: the multiple marker analyses take longer to converge than the single marker analyses, and even the single marker runs on the X Line data took over a week to converge. Secondly, there is a much smaller proportion of animals with multiple marker data available in the X Lines than in the simulation studies where multiple marker information was only available at the end of the pedigree (1% *vs.* 24%). This is likely to have an adverse effect on the convergence time. Lastly, for the High lines the parentage of the animals with multiple marker data is unknown, so it becomes very difficult to integrate the information from these animals with the information from the rest of the pedigree.

The single marker analysis of *dilute* produced similar results to the regression method, with an estimated additive effect for the QTL of 0.4g ($\approx 0.2\sigma_p$) and a recombination frequency between the QTL and *dilute* of 12%. There were less clear results from *brown* using both the regression and and the Gibbs sampling analysis. The regression analysis estimated the QTL as having a small (non-significant) effect whereas the Gibbs analysis estimated the QTL as having a large effect but being only loosely linked to the marker. It is difficult to draw any conclusions about the QTL linked to *brown*. The analyses presented here appear to suggest it is either a small QTL situated fairly close to *brown*, or a large QTL located a long way from *brown*. Information on the marker frequencies from the end of the experiment indicate that several markers located very close to *brown* are associated with a large effect on body weight (Keightley *et al.*, 1995). This would indicate that the QTL is in fact close to *brown*. The contradictory results may be due to there being more than one QTL linked to *brown*. It is not possible to test between 1 or 2 QTL using just a single marker. Using the information from the linked markers available from generation 21, it might be possible to distinguish between these two scenarios. The Gibbs sampling method described in Chapter 7, however, has not been tested with multiple QTL yet, and its ability to distinguish between rival hypotheses is another area which has not yet been

touched upon.

Two general approaches are followed in this thesis for the investigation of the genetic control of body weight: detecting deviations from the infinitesimal model, and explicitly estimating the effects of individual genes on the trait. The conclusions from the first approach are that the experimental dataset displays significant deviations from the predictions of the infinitesimal model. The rates of response to selection appear to change over the course of the experiment, one of the the selection lines displays a markedly different response to the others, indicating a mutation or a rare recombination event, and significant changes in the additive and environmental variances over the course of the experiment are detected. It is not possible to make firm conclusions about the genetics underlying body weight, but is is possible to conclude that (a) there are significant non-additive genetic effects present and (b) there are individual genes segregating with a detectable effect on body weight. It would appear from the analysis of the X Lines dataset that the infinitesimal model cannot provide an adequate description of the behaviour of datasets with more than a small number of generations. The largest single cause of the discrepancies between the model predictions and real data is likely to be changes in the frequencies of genes affecting the trait. This effect can not be accounted for by using infinitesimal models. It seems likely that as more information becomes available on the genetic control of quantitative traits, such as the distribution of gene effects, then it will become advantageous to use more realistic genetic models which can account for gene frequency changes. The methods described in this thesis for the analysis of data under mixed inheritance model may be regarded as a first step, albeit a limited one, towards the goal of such realistic models.

References

Beniwal, B. K., I. M. Hastings, R. Thompson, and W. G. Hill, (1992a). Estimation of changes in selected lines of mice using REML with an animal model. 1. Lean mass. *Heredity* **69**:352–360.

Beniwal, B. K., I. M. Hastings, R. Thompson, and W. G. Hill, (1992b). Estimation of changes in selected lines of mice using REML with an animal model. 2. Body weight, body composition and litter size. *Heredity* **69**:361–371.

Bonhomme, F., J. Guenet, B. Dod, K. Moriwaki, and G. Bulfield, (1987). The polyphyletic origin of laboratory inbred mice and their rate of evolution. *Biological Journal of the Linnean Society* **30**(1):51–58.

Bonney, G. E., (1984). On the statistical determination of major gene mechanisms in continuous human traits: regressive models. *American Journal of Medical Genetics* **18**:731–749.

Bonney, G. E., G. Lathrop, and J.-M. Lalouel, (1988). Combined linkage and segregation analysis using regressive models. *American Journal of Human Genetics* **43**:29–37.

Bryant, E. H. and L. M. Combs, (1986). The effect of an experimental bottleneck upon quantitative variation in the housefly. *Genetics* **114**:1191–1211.

Bulmer, M. G., (1974). Linkage disequilibrium and genetic variability. *Genetical Research* **23**:199–203.

Bulmer, M. G., (1976). The effect of selection on genetic variability: a simulation study. *Genetical Research* **28**:101–117.

Butler, L., (1954). The effect of the coat colour dilution gene on body size in the mouse. *Heredity* **8**:275–278.

Caballero, A. and P. D. Keightley, (1994). A pleiotropic nonadditive model of variation in quantitative traits. *Genetics* **138**:883–900.

Caballero, A., M. A. Toro, and C. López-Fanjul, (1991). The response to artificial selection from new mutations in *Drosophila melanogaster*. *Genetics* **128**:89–102.

Cannings, C., E. A. Thompson, and M. H. Skolnick, (1978). Probability functions on complex pedigrees. *Advances in Applied Probability* **10**:26–61.

Casella, G. and E. I. George, (1992). Explaining the Gibbs sampler. *The American Statistician* **46**(3):167–174.

Castle, W. E., (1941a). Influence of certain colour mutations on body size in mice, rats and rabbits. *Genetics* **26**:117–191.

Castle, W. E., (1941b). Size inheritance. *American Naturalist* **75**:488–498.

Clayton, G. A., J. A. Morris, and A. Robertson, (1957). An experimental check of quantitative genetical theory. I. Short-term response to selection. *Journal of Genetics* **55**:152–170.

Devroye, L., (1986). *Non-uniform random variate generation.* Springer-Verlag, New York.

Elston, R. C. and J. Stewart, (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**:523–542.

Emerson, R. A., (1910). The inheritance of sizes and shapes in plants. *American Naturalist* **44**:739–746.

Falconer, D. S., (1953). Selection for large and small size in mice. *Journal of Genetics* **51**:470–501.

Falconer, D. S., (1973). Replicated selection for body weight in mice. *Genetical Research* **22**:291–321.

Falconer, D. S., (1989). *Introduction to quantitative genetics.* Longman, 3rd edition.

Feldman, H. W., (1935). The brown variation and growth of the house mouse. *American Naturalist* **69**:370–374.

Festing, M. F. W., (1989). *Inbred Strains of Mice*, volume 1, chapter 15, pages 636–648. New York: Oxford University Press, 2nd edition.

Fisher, R. A., (1918). The correlation between relatives under the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**:399–433.

Fletcher, R., (1987). *Practical methods of optimization.* Wiley-Interscience, 2nd edition.

Frankel, W., J. Stoye, B. Taylor, and J. Coffin, (1990). A linkage map of endogenous murine leukemia proviruses. *Genetics* **124**:221–236.

Garnett, I. and D. Falconer, (1975). Protein variation in strains of mice differing in body size. *Genetical Research* **25**:45–57.

Gelfand, A. E. and A. F. Smith, (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**:398–409.

Geman, S. and D. Geman, (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**:721–741.

Genstat 5 Committee, (1993). *Genstat 5 Release 3 Reference Manual.* Oxford: Clarendon Press.

Gianola, D. and J. L. Foulley, (1990). Variance estimation from integrated likelihoods (VEIL). *Genetics, Selection, Evolution* **22**:403–417.

Gilks, W. R., D. G. Clayton, D. J. Spiegelhalter, N. G. Best, A. J. McBeil, L. D. Sharples, and A. J. Kirby, (1993). Modelling complexity: Applications of the Gibbs sampler in medicine (with discussion). *Journal of the Royal Statistical Society, Series B* **55**:39–52.

Goodale, H. D., (1938). A study of the inheritance of body weight in the albino mouse by selection. *Journal of Heredity* **29**:101–112.

Goodnight, C. J., (1988). Epistasis and the effect of founder events on the additive genetic variance. *Evolution* **42**:441–454.

Green, C. V., (1931). Linkage in size inheritance. *American Naturalist* **65**:502–511.

Guo, S. W. and E. A. Thompson, (1991). Monte Carlo estimation of variance component models for large complex pedigrees. *IMA Journal of Mathematics Applied in Medicine and Biology* **8**:171–189.

Guo, S. W. and E. A. Thompson, (1992). A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics* **51**:1111–1126.

Guo, S. W. and E. A. Thompson, (1994). Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics* **50**:417–432.

Hasstedt, J. K., (1982). A mixed-model likelihood approximation on large pedigrees. *Computers and Biomedical Research* **15**:295–307.

Hastings, W. K., (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**:97–109.

Hedrick, P. W. and R. E. Comstock, (1968). Role of linkage in gene frequency change of coat colour alleles in mice. *Genetics* **58**:297–303.

Henderson, C. R., (1976). A simple method of computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**:69–83.

Henderson, C. R., (1985). Best linear unbiased prediction of non-additive merits in non-inbred populations. *Journal of Animal Science* **60**:111–117.

Hill, W. G., (1972). Estimation of realised heritabilities from selection experiments. I. divergent selection. *Biometrics* **28**:747–765.

Hill, W. G., (1982). Predictions of response to selection from new mutations. *Genetical Research* **40**:255–278.

Hill, W. G. and A. Caballero, (1992). Artificial selection experiments. *Annual Review of Ecology and Systematics* **23**:287–310.

Hill, W. G., A. Caballero, and P. D. Keightley, (1994). Variation from spontaneous mutation for body size in the mouse. In *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production* **19**:67–70. University of Guelph, Guelph.

Hoeschele, I., (1994). Bayesian QTL mapping via the Gibbs Sampler. In *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production* **21**:241–244. University of Guelph, Guelph.

James, J. W., (1990). Selection theory versus selection results; a comparison. In *Proceedings of the Fourth World Congress on Genetics Applied to Livestock Production* **13**:195–204. University of Edinburgh, Edinburgh.

Janss, L. L. G., R. Thompson, and J. A. M. van Arendonk, (1994a). Bayesian inference in a mixed inheritance model by Gibbs sampling. *Theoretical and Applied Genetics* (Submitted).

Janss, L. L. G., J. A. M. van Arendonk, and E. W. Brascamp, (1994b). Identification of a single gene affecting intramuscular fat in Meishan crossbreds

using Gibbs sampling. In *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production* **18**:361–364. University of Guelph, Guelph.

Jensen, J., C. Wang, D. Sørensen, and D. Gianola, (1994). Bayesian-inference on variance and covariance components for traits influenced by maternal and direct genetic effects, using the Gibbs sampler. *Acta Agriculturae Scandinavica* **44**(4):193–201.

Keightley, P. D., (1994a). The distribution of mutation effects on viability in *Drosophila melanogaster*. *Genetics* **138**:1315–1322.

Keightley, P. D., (1994b). Recovery of phylogenetic information with microsatellite markers. *Mouse Genome* **92**:683–684.

Keightley, P. D. and G. Bulfield, (1993). Detection of quantitative trait loci from frequency changes of marker alleles under selection. *Genetical Research* **62**:195–203.

Keightley, P. D., T. Hardge, L. May, and G. Bulfield, (1995). A genetic map of quantitative trait loci controlling body weight in the mouse based on frequency changes of marker alleles under artificial selection. *Genetics* (submitted).

Keightley, P. D. and W. Hill, (1992). Quantitative genetic variation in body size of mice from new mutations. *Genetics* **131**:693–700.

Lange, K. and S. Matthysse, (1989). Simulation of pedigree genotypes by random walks. *American Journal of Human Genetics* **45**:959–970.

Lange, K. and E. Sobel, (1991). A random walk method for computing genetic location scores. *American Journal of Human Genetics* **49**:1320–1334.

Lewontin, R. C., (1964). The interaction of selection and linkage. II. Optimum models. *Genetics* **50**:757–82.

Lin, S., (1995). A scheme for constructing an irreducible Markov chain for pedigree data. *Biometrics* **51**:318–322.

Lin, S., E. A. Thompson, and E. Wijsman, (1993). Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. *IMA Journal of Mathematics Applied in Medicine and Biology* **10**:1–17.

Lin, S., E. A. Thompson, and E. Wijsman, (1994). Finding noncommunicating sets for Markov Chain Monte Carlo estimations on pedigrees. *American Journal of Human Genetics* **54**:695–704.

Liu, J. S., W. H. Wong, and A. Kong, (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**(1):27–40.

Lynch, M., (1988). The rate of polygenic mutation. *Genetical Research* **51**:33–43.

MacArthur, J. W., (1944a). Genetics of body size and related characters. I. selecting small and large races of the laboratory mouse. *American Naturalist* **78**:142–157.

MacArthur, J. W., (1944b). Genetics of body size and related characters. II. satellite characters associated with body size in mice. *American Naturalist* **78**:224–237.

Mackay, T. F. C., R. Lyman, and M. S. Jackson, (1992). Effects of P elements on quantitative traits in *Drosophila melanogaster*. *Genetics* **130**:315–332.

Mather, K., (1941). Variation and selection of polygenic characters. *Journal of Genetics* **41**:159–93.

Maynard Smith, J., (1989). *Evolutionary Genetics*. Oxford: Oxford University Press.

Meyer, K., (1988). DFREML - A set of programs to estimate variance components under an individual animal model. *Journal of Dairy Science* **71**:33.

Meyer, K., (1989). Restricted maximum likelihood to estimate variance components under an individual animal model with several random effects using a derivative-free algorithm. *Génétique, Sélection, Évolution* **23**:317–340.

Meyer, K. and W. G. Hill, (1991). Mixed model analysis of a selection experiment for food intake in mice. *Genetical Research* **57**:71–81.

Meyer, K. and W. G. Hill, (1992). Approximation of sampling variances and confidence intervals for maximum likelihood estimates of variance components. *Journal of Animal Breeding and Genetics* **109**(4):264–280.

Morton, N. E. and C. J. Maclean, (1974). Analysis of family resemblance. III. complex segregation of quantitative traits. *American Journal of Human Genetics* **26**:489–503.

Neal, R. E., (1993). Probabilistic inference using Markov Chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, Toronto, Canada.

Nelder, J. A. and R. Mead, (1965). A simplex method for function minimization. *Computer Journal* **7**:308–313.

Ott, J., (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *American Journal of Human Genetics* **31**:161–175.

Patterson, H. D. and R. Thompson, (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**:545–554.

Pong-Wong, R. and J. A. Woolliams, (1994). Recovery of information on major gene effects using Gibbs sampling when genotypes are known for a subset of the population. In *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production* **21**:256–259. University of Guelph, Guelph.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, (1992). *Numerical Recipes in C:The art of scientific computing.* Cambridge University Press, Cambridge, 2nd edition.

Quaas, R. L., (1976). Computing the diagonal elements of a large numerator relationship matrix. *Biometrics* **32**:949–953.

Raftery, A. E. and S. M. Lewis, (1994). The number of iterations, convergence diagnostics and generic Metropolis algorithms. Technical report, Department of Statistics, University of Washington, Seattle, Washington.

Rahnefeld, G. W., W. J. Boylan, R. E. Comstock, and M. Singh, (1963). Mass selection for post-weaning growth in mice. *Genetics* **48**:1567–1583.

Roberts, R. C., (1965). Some contributions of the laboratory mouse to animal breeding research. part I. *Animal Breeding Abstracts* **33**:339–353.

Robertson, A. and W. G. Hill, (1983). Population and quantitative genetics of many linked loci in finite populations. *Proceedings of the Royal Society of London Series B* **219**:263–264.

Sheehan, N. and A. Thomas, (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* **49**:163–75.

Sheridan, A. K., (1988). Agreement between estimated and realised genetic parameters. *Animal Breeding Abstracts* **56**:877–89.

Smith, S. P. and A. Mäki-Tanila, (1990). Genotypic covariance matrices and their inverses for models allowing dominance and inbreeding. *Genetics, Selection, Evolution* **22**.

Sørensen, D. A., S. Andersen, J. Jensen, C. S. Wang, and D. Gianola, (1994a). Inferences about genetic parameters using the Gibbs sampler. In *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production* **18**:321–328. University of Guelph, Guelph.

Sørensen, D. A. and B. W. Kennedy, (1983). The use of the relationship matrix to account for genetic drift variance in the analysis of genetic experiments. *Theoretical and Applied Genetics* **66**:217–220.

Sørensen, D. A. and B. W. Kennedy, (1984). Estimation of genetic variances from selected and unselected populations. *Journal of Animal Science* **59**:1213–1223.

Sørensen, D. A., C. S. Wang, J. Jensen, and D. Gianola, (1994b). Bayesian-analysis of genetic change due to selection using Gibbs sampling. *Genetics, Selection, Evolution* **26**:333–360.

Thompson, E. A., (1994). Monte Carlo likelihood in the genetic mapping of complex traits. *Philisophical Transactions of the Royal Society of London, Series B* **344**:345–351.

Thompson, E. A. and S. W. Guo, (1991). Estimation of likelihood ratios for complex genetic models. *IMA Journal of Mathematics Applied in Medicine and Biology* **8**:149–169.

Thompson, E. A. and R. G. Shaw, (1990). Pedigree analysis for quantitative traits: variance components without matrix inversion. *Biometrics* **46**:399–413.

Turelli, M. and N. H. Barton, (1994). Genetic and statistical analyses of strong selection on polygenic traits: what, me normal? *Genetics* **138**:913–941.

van der Lugt, A., L. L. G. Janss, and J. A. M. van Arendonk, (1994). Estimation of variance components in large animal models using Gibbs sampling. In *Proceedings of the Fifth World Congress on Genetics Applied to Livestock Production* **18**:329–332. University of Guelph, Guelph.

Wang, C. S., J. J. Rutledge, and D. Gianola, (1993). Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics, Selection, Evolution* **25**:41–62.

Wray, N. R., (1990). Accounting for mutation effects in the additive genetic variance covariance-matrix and its inverse. *Biometrics* **46**(1):177–186.

Wright, S., (1969). *Evolution and the Genetics of Populations. Vol. 2, The theory of Gene Frequencies.* Univ. of Chicago, Chicago.

Appendix

Sampling gene positions

This Appendix describes the method used in Section 7.2.3 to sample the map position of a gene conditional on the positions of all other linked genes, and on the observed recombinations between the gene and all the other genes. The conditional distribution of the map position is given by:

$$p(\kappa_j|\mathcal{G}, \boldsymbol{\kappa}_{-j}) \propto K_i \prod_k r_k^{R_{i_k}} (1 - r_k)^{R'_{i_k}} \tag{A.1}$$

where K_i, R_{i_k} and R'_{i_k} are as defined in Section 7.2.2, κ_j is the locus of the jth gene, $\boldsymbol{\kappa}_{-j}$ is the vector of the positions of all other genes on the same chromosome and r_k is the recombination rate between the QTL and the kth gene given by Haldane's mapping function:

$$r_k = \frac{1}{2}(1 - e^{-2|(\kappa_j - \kappa_k)|}). \tag{A.2}$$

A suitable method for sampling from (A.1) is the rejection method. This relies on being able to find a function $f(x)$ which can be readily sampled from and which is similar to the desired distribution p, but with $p(x)$ being less than $f(x)$ for all points within the desired range of x. A random deviate from p can be generated by sampling x from f and accepting or rejecting the sample with respective probabilities $p(x)/f(x)$ and $1 - p(x)/f(x)$ (Devroye, 1986). That is, the sample x is accepted with a probability proportional to the ratio of the heights

of the two curves at x. The efficiency of this method depends on how often a sample is rejected, which in turns depends on how well the generator density f matches the required density p. For the method described here, the efficiency of the sampling is not that important because relatively few samples are required for each Gibbs sampling cycle. It was therefore decided to concentrate on finding a simple working algorithm for sampling rather than attempt to optimize the sampling for speed. The conditional distribution (A.1) is very 'lumpy', typically having a sharp peak in each gene interval[1]. Rather than find a function that emulated this, it was decided to sample from each gene interval separately, and then pick a sample from one interval with probability proportional to the value of (A.1) at the sample point. This allows the Markov Chain to 'jump' between gene intervals.

A useful function for rejection sampling is derived from the *Lorentzian distribution*:

$$p(y)dy = \frac{1}{\pi}\Big(\frac{1}{1+y^2}\Big)dy \tag{A.3}$$

which has the tangent function as its inverse indefinite integral. A general version of the function (A.3) is given by:

$$f(x) = \frac{c_0}{1+(x-x_0)^2/a_0^2} \tag{A.4}$$

It follows that the x-coordinate of an area-uniform random point under (A.4) can be generated for any values of the constants c_0, a_0 and x_0 by:

$$x = a_0 \tan(\pi U) + x_0, \tag{A.5}$$

where U is a uniform deviate between 0 and 1 (Press *et al.*, 1992). This produces a bell-shaped distribution with maximum height c_0 when $x = x_0$, with a width determined by a_0. This can be made to fit (A.1) reasonably well, with the major differences being that (A.4) is symmetrical with long tails while (A.1) has short tails and can be asymmetrical (particularly if the mode of the distribution is close

[1]A gene interval is simply a locus flanked by two genes or by one gene and the end of the chromosome.

to a marker). The procedure used for finding values for the parameters c_0, a_0 and x_0 and sampling from (A.1) was as follows:

For each gene interval i:

1. Find the maximum[2] of (A.1) in the interval. Set x_0 to the location of the maximum and set c_0 marginally above this maximum value.
2. Find the minimum value of a_0 required to ensure that $f(x) > p(x)$ at all points within the interval.
3. Sample x_i from (A.4) using (A.5), truncating the distribution so all points lie within the desired gene interval.
4. Sample a uniform deviate z between 0 and 1.
5. Accept the sample x_i if $z < p(x_i)/f(x_i)$, otherwise repeat from 3.

One of the samples x_i is then selected with probability proportional to $p(x_i)$.

[2]The maximization/minimization routines used were the line search routines `mnbrak()` and `brent` from Numerical Recipes (Press *et al.*, 1992).