2005

Convergence Analysis of MCMC Method in the Study of Genetic Linkage with Missing Data

Diana Fisher

Convergence Analysis of MCMC Method in the Study of Genetic Linkage with Missing Data

Thesis submitted to
The Graduate College of
Marshall University

In partial fulfillment of the
Requirements for the degree of
Master of Arts
Department of Mathematics

by
Diana Fisher

Dr. Alfred Akinsete, Committee Chairman
Dr. Yulia Dementieva
Dr. Laura Adkins

Abstract

Computational infeasibility of exact methods for solving genetic linkage analysis problems has led to the development of a new collection of stochastic methods, all of which require the use of Markov chains. The purpose of this work is to investigate the complexities of missing data in pedigree analysis using the Monte Carlo Markov Chain (MCMC) method as compared to the exact results. Also, we attempt to determine an association between missing data in a familial pedigree and the convergence to stationarity of a descent graph Markov chain implemented in the stochastic method for parametric linkage analysis.

In particular, we will implement the stochastic method to solve a pedigree problem for a disease trait, in order to look at the associated problems with missing data from the pedigree, and investigate the deviation between the MCMC method and the exact results. Using the method for maximum autocorrelation and bounding of the second largest eigenvalue, we will study the effects of missing data on the convergence rate and the accuracy of the MCMC method in solving the pedigree analysis problem. Finally, we will use the computational implementation of *SimWalk2* to study the convergence rate and accuracy of the MCMC method for the disease Episodic Ataxia.

The implementation of the MCMC method through *SimWalk2* for the disease gene Episodic Ataxia found evidence to suggest that both the efficiency and accuracy of the method may be severely reduced by an increase in missing data in the pedigree. Certain variations of model parameters influenced the ability of the method to produce accurate results, but the most crucial of the variables studied was the level of missing information from the pedigree itself. This can be seen as a detriment to the implementation, as pedigree information is very often missing from the model. Further research in this topic would need to include the implementation of this method on more genetic parameters and differing pedigree variations. Also, it might be of interest to look into possible ways to combat the effects of missing data on the MCMC method.

Acknowledgements

I would like to acknowledge and thank all of the people who made the completion of this thesis possible. In particular, I would like to thank Dr. Dementieva for giving me the opportunity to study a topic that both excites and challenges, and Dr. Akinsete for working with me through the many drafts and complexities that accompanied this work. I would also like to express my great appreciation to Dr. Adkins for her participation on my thesis committee, and to the many other professors who shared their wisdom and expertise as I completed this work. Finally, I wish to thank my family and friends for supporting me through the long hours and difficult times that were a part of my research.

Contents

Chapter			Page
	Abstract		ii
	Acknowledgements		iii
	Contents		iv
	List of Figures		vi
	List of Tables		vii
	Genetic Terminology		ix
1	**Introduction**		1
	1.1	Early Genetic Foundations	1
	1.2	Linkage Analysis	1
	1.3	Components of the Genetic Model	3
	1.4	LOD Score Analysis	3
	1.5	Multipoint Linkage Analysis	5
	1.6	Factors that Can Affect Linkage Analysis	6
2	**The Elston-Stewart Algorithm and Episodic Ataxia**		9
	2.1	The General Problem	9
	2.2	Calculating a Likelihood Using the Elston-Stewart Algorithm	11
	2.3	Episodic Ataxia	12
3	**A MCMC Method for Computing LOD and Location Scores**		14
	3.1	Markov Chain Preliminaries	14
		3.1.1 Defining a Markov Chain	14
		3.1.2 Communication and Classification of States	15
		3.1.3 Periodicity and Ergodicity	16
	3.2	Methodology of the MCMC Process	16
	3.3	Descent Graph Markov Chain Method for Linkage Analysis	17
		3.3.1 MC State Space	17
		3.3.2 The Descent Graph Markov Chain	19
		3.3.3 Likelihood of a Descent Graph	21
		3.3.4 Computing the Location Score	23

3.3.5 Constructing the Markov Chain 24
3.3.6 Detailed Balance and Overall Balance 25
4 Convergence Analysis of the MCMC Method 27
4.1 Transition Matrix of the Markov Chain 27
4.2 Eigen Analysis of the Transition Matrix 27
4.2.1 The Maximal Eigenvalue for the Transition Matrix 30
4.2.2 The Transition Matrix Yielding Stationarity 33
4.3 Rate of Convergence of a Markov Chain 35
4.3.1 Bound on Convergence 35
4.3.2 Convergence Analysis with Maximum Autocorrelation 36
4.3.3 Dirichlet Form 40
4.3.4 The Poincaré Inequality and Results 41
4.4 Bound for Convergence Rate 43
5 Implementation of *SimWalk2* for the Disease Gene Episodic Ataxia 44
5.1 Methodology 44
5.2 Statistical Analysis of *SimWalk2* Results 46
5.2.1 Accuracy of MCMC Method for Missing Data in Pedigree 47
5.2.2 Accuracy of MCMC Method for Variations of Parameter 48
5.2.3 ANOVA Results for MCMC Accuracy 49
5.3 Analysis of Convergence Results 51
5.3.1 Convergence Results 52
5.3.2 Convergence Results by Parameters 53
6 Conclusion 54
References 56
Appendix
A Figures 60
B Input – Output Data Sample Files 61
C Table of Results 67

List of Figures

Figure		Page
2.1	Location Score Curve for Episodic Ataxia Versus 12p Marker	13
3.1(a)	Conventional Pedigree Representation	18
3.1(b)	Descent Graph Describing Gene Flow	18
3.2	Example of Transition Rule T_0	19
3.3	Example of Transition Rule T_1	19
3.4	Example of Transition Rule T_{2a}	20
3.5	Example of Transition Rule T_{2b}	20
3.6	Failure of Descent Graphs A and C to Communicate	21
A1	Pedigree for Episodic Ataxia Analysis	59
A2	Plot of LOD Score for EA vs. Marker S372	59

List of Tables

Table		Page
5.1	Variation of Parameters for Allele Frequency	45
5.2	Variation of Parameters for Penetrances	45
5.3	Trials for *SimWalk2*	46
5.4	Data Summary and ANOVA Results for Location Score	50
5.5	Data Summary and ANOVA Results for Locational Position	51
5.6	Convergence Bounds Results	52
5.7	Convergence Bounds with Variation of Parameters	53
C1	Comparison of Algorithms	67
C2	*SimWalk2* Raw Accuracy Results	68
C3	Pedigree Analysis by Parameter Allele Frequency for Locational Position	69
C4	Pedigree Analysis by Parameter Penetrance for Locational Position	69
C5	Pedigree Analysis by Parameters Allele Frequency and Penetrance for Locational Position	69
C6	Pedigree Analysis by Parameters Allele Frequency and Penetrance (Severe Case) for Locational Position	69
C7	Pedigree Analysis by Parameters Allele Frequency for Locational Score	70
C8	Pedigree Analysis by Parameters Penetrance for Locational Score	70
C9	Pedigree Analysis by Parameters Allele Frequency and Penetrance for Locational Score	70
C10	Pedigree Analysis by Parameters Allele Frequency and Penetrance (Severe Case) for Locational Score	70
C11	Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency for Location Score	71
C12	Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Penetrance for Location Score	71
C13	Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency and Penetrance for Location Score	72

C14 Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency and Penetrance (Severe Case) for Location Score 72
C15 Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency for Locational Position 73
C16 Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Penetrance for Locational Position 73
C17 Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency and Penetrance for Locational Position 74
C18 Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency and Penetrance (Severe Case) for Locational Position 74
C19 *SimWalk2* Raw Convergence Results 75

Genetic Terminology

The following definitions of genetic terminologies are as found in [36].

- **Allele**: one of a pair of alternative forms of a gene that occur at a given locus in a chromosome
- **Autosome**: any chromosome that is not a sex chromosome
- **Codominant Alleles**: alleles that produce independent effects when heterozygous
- **Dominance**: a condition in which one member of an allele pair is manifested to the exclusion of the other
- **Gene**: a unit of inheritance (DNA) located in a fixed position on a chromosome; a hereditary determinant of a specific biological function
- **Genotype**: the genetic constitution (gene makeup) of an organism
- **Genetic Heterogeneity**: similar phenotypes occurring due to different mutations or recombinations
- **Heterozygote**: an organism with unlike members of any given pair or series of alleles that consequently produces unlike gametes
- **Homozygote**: an organism with like members of any given pair or series of alleles that consequently produces like gametes
- **Homologous Chromosomes**: chromosomes that occur in pairs and are generally similar in size and shape, one having come from the male parent and the other from the female parent
- **Linkage**: a relationship among genes in the same chromosome – such genes tend to be inherited together
- **Locus**: a fixed position on a chromosome that is occupied by a given gene or one of its alleles
- **Meiosis**: the process by which the chromosome number of a reproductive cell becomes reduced to half the somatic number, resulting in the formation of gametes in humans
- **Pedigree**: a table, chart, or diagram representing the ancestry of an individual
- **Penetrance**: the percentage of individuals that show a particular phenotype among those capable of showing

- **Phenocopy**: an environmentally-induced, nonhereditary variation in an organism, closely resembling a genetically-determined trait
- **Phenotype**: the observable characteristics of an organism
- **Recessive**: a term applied to one member of an allelic pair lacking the ability to manifest itself when the other or dominant member is present
- **Recombination**: the production of gene combinations not found in the parents by the assortment of nonhomologous chromosomes and crossing over between homologous chromosomes during meiosis. For linked genes, the frequency of recombination can be used to estimate the genetic map distance

1. Introduction

1.1 Early Genetic Foundations

Modern genetics began with Mendel, who postulated his two laws as a probability model, [22]. The following summarizes Mendel's laws:

a. Everyone has two genes controlling a given trait, one from the mother and one from the father.
b. When an individual has an offspring, a copy of a randomly chosen gene from the individual's two genes segregates to the offspring.
c. Gene segregation is independent of the other parent, independent for each child, and independent for each trait (or locus).

Following the rediscovery of Mendel's work in 1900, scientists discovered that the independence of segregation for different loci was not strictly valid. Instead, there are groups of traits that are linked, and the genes controlling them tend to be inherited as a group, not independently. Soon after, geneticists associated this linkage (or dependence) with the actual structure of the chromosome [37].

According to [11], and citing the work of Sturtevant in 1913, patterns of combinations of inherited genes are best explained by a linear arrangement of genes for different traits. The latter created the first *gene ordering* inference on which many linkage studies are still based today. [11], and a host of many others, extended this mathematical model, defining the distance along a chromosome as the expected number of recombination events. Thus, the relationship between physical distance on a chromosome and the statistical measure of linkage between genetic loci was developed.

1.2 Linkage Analysis

In linkage analysis, cosegregation of two or more loci (genes or traits) is examined to determine whether they more frequently segregate independently, according to Mendel's laws, or tend to be inherited together. It is most likely that if the loci segregate together, they reside in close proximity to one another. Thus, the ability to determine

independent or dependent segregation determines the relative location of the loci. Alleles of genes residing on the same chromosome should segregate together at a rate that is related to the distance between them on the chromosome.

The measure of genetic linkage is the recombination fraction (θ), which is the probability that a parent will produce a recombinant offspring. Recombination occurs when homologous chromosomes cross over, and nonrecombination occurs when the parental type remains intact. Since multiple crossovers can occur between two loci, an even number of such crossovers appears the same as a nonrecombination event. Fortunately, multiple crossovers are very rare between closely linked genes, and thus will not impact the final results in such cases.

Recombination fractions range between 0 and 0.5, where $\theta = 0$ indicates "complete" linkage and $\theta = 0.5$ indicates no linkage. It should be noted that the unit of measurement for genetic linkage is the centiMorgan (cM), and that one map unit corresponds to one centiMorgan (or 1% recombination). Small values of θ are equivalent to actual map distances, and so recombination fractions are additive over small distances. For larger distances, this is not the case because multiple crossovers occur with more frequency, and mapping functions must be implemented in order to convert recombination frequencies into actual map distance.

In the most simplistic form, linkage analysis is reduced to counting recombinants and nonrecombinants – as found in the methodology for most experimental animals. However, in humans, this is not feasible. Long generation time, the inability to control matings, the inability to control study participation, and the inability to dictate key exposures and environmental conditions cause linkage analysis to depend extensively on both simple and sophisticated statistical methods [10].

Up to this point, linkage analysis has been described in a general manner, but it is important to recognize that there are several different types of linkage analysis, and the statistical method needed depends upon the type of linkage analysis being completed. For the purpose of this work, the only type of linkage analysis involved is parametric linkage analysis. Given a marker map and any pedigree with partial genotyping of marker loci and a trait, parametric linkage analysis involves the calculation of the

likelihood (LOD) of a trait segregating with a marker. The defining property of parametric linkage analysis is that the genetic model is known.

1.3 Components of the Genetic Model

There are certain components of the genetic model that determine the localization and linkage patterns of a gene. These include the inheritance pattern of the trait locus (whether it is dominant or recessive, sex-linked or autosomal), the trait allele frequency (whether the trait is common or rare), and the trait allele penetrance (the probability that an unaffected individual is unaffected because he is a non-gene carrier or a non-penetrant gene carrier). Other factors that influence linkage analysis studies are the frequency of phenocopies in the population being studied, marker allele frequencies, and mutation rates within the genes being studied [10].

One of the most complicating factors in parametric linkage analysis studies is the phase information of members of the pedigree. In heterozygote individuals, there are two different phases for a particular locus in a genotype: coupling and repulsive. For example, given a locus one with alleles A and B and a locus two with alleles 1 and 2, the individual having genotype AB12 has two possible phases: the coupling phase C = A1 | B2, where A and 1 are inherited together and B and 2 are inherited together, or the repulsive phase R = A2 | B1, where A and 2 are inherited together and B and 1 are inherited together. If it is not known whether the individual has a coupling or repulsive phase, then the individual exhibits *linkage phase unknown status*; otherwise, he/she exhibits *linkage phase known* status. When linkage phase is unknown, the complexity of the statistical calculations increases dramatically [10].

1.4 LOD Score Analysis

LOD score analysis is a likelihood-based parametric linkage approach used to estimate the recombination fraction and significance of the evidence for linkage. The LOD score was first defined by [23] as a logarithmic function of the odds for linkage. The likelihood (L) of observing a particular configuration of a disease and a marker locus in a family is calculated assuming no linkage ($\theta = 0.5$), and this likelihood is then compared to the likelihood of observing the same configuration of the loci within the family, assuming

varying degrees of linkage over a selected range of recombination frequencies ($0 \le \theta < 0.5$). The two-point LOD score involves linkage between only two loci (i.e. a disease locus and a marker locus) and is expressed as

$$z(x) = \log_{10}\left[\frac{L(pedigree \mid \theta = x)}{L(pedigree \mid \theta = 0.5)}\right], \text{ where } 0 \le x < 0.5.$$

In LOD score analysis, the null hypothesis (H_o) assumes that $\theta = 0.5$, or that there is no linkage, while the alternate hypothesis (H_A) assumes that the disease and the marker loci are linked. To demonstrate linkage, there must be evidence of cosegregation that can support the rejection of the null hypothesis. The likelihood function is given by $L = \theta^R (1-\theta)^{NR}$, where R is the number of recombinant offspring, NR is the number of nonrecombinant offspring, and $N = R + NR$ is the total number of offspring. The two-point LOD score for a phase-known pedigree then becomes

$$z(x) = \begin{cases} \log_{10}\left[\dfrac{\theta^R (1-\theta)^{NR}}{(0.5)^R (0.5)^{NR}}\right], & \theta \neq 0 \\ N \cdot \log_{10}(2), & \theta = 0,\ R = 0 \end{cases}$$

and for a phase-unknown pedigree, we have

$$z(x) = \begin{cases} \frac{1}{2} \cdot \log_{10}\left[\dfrac{\theta^R (1-\theta)^{NR}}{(0.5)^R (0.5)^{NR}}\right] + \frac{1}{2} \cdot \log_{10}\left[\dfrac{\theta^{NR} (1-\theta)^{R}}{(0.5)^{NR} (0.5)^{R}}\right]. & \theta \neq 0 \\ N \cdot \log_{10}(2), & \theta = 0,\ R = 0 \end{cases}$$

Note that since recombination fractions may differ between males and females, LOD scores may also be computed using a sex-specific recombination fraction (given that a maximal one is known). Generally, female recombination is greater than male recombination, except in certain telomeric regions of some chromosomes [10].

For a single pedigree, we can define the maximum likelihood estimate for the recombination fraction to be $\theta = \dfrac{R}{NR + R}$, and compute the maximum LOD score. However, in order to sum LOD scores across all families in the study, it is necessary to

compute the LOD scores for varying values of the recombination fraction. This yields the maximum LOD score for *all* families in the study, and allows for a more accurate representation of the likelihood for linkage [23]. Sums of LOD scores of 3.0 or greater are indicative of linkage and scores of -2.0 or less are indicative of no linkage. Values between -2.0 and 3.0 are considered inconclusive and require additional family data [10].

These criteria, originally suggested by Morton, are based on probabilities of type I (α) and type II (β) errors. In this case, α refers to the probability of concluding that there is linkage between the tested loci, when in fact linkage does not exist, leading to the rejection of a true null hypothesis. Also, β refers to concluding that there is no linkage, when in fact linkage does exist, leading to the acceptance of a false null hypothesis. So to guard against false positive evidence for linkage in the data, the conservative value of $z \geq 3.0$ was chosen for concluding that there was significant evidence supporting linkage, [23].

There are several advantages of LOD score analysis over other methods. For example, statistically, it is more powerful than any nonparametric method, utilizing every family member's phenotypic and genotypic information, and provides both an estimate of the recombination fraction and a statistical test for linkage and genetic heterogeneity, [10].

1.5 Multipoint Linkage Analysis

Once a disease gene is mapped to a particular region of the chromosome, the goal becomes one of positioning the disease locus on the known marker map. In the method of location scores, it is necessary to evaluate and plot the joint likelihood of the disease and marker genotypes as a function of the position of the disease locus. A location score is an extension of the two-point LOD score, acting as a multipoint LOD score. Instead of testing linkage of a disease trait with a single marker, a location score tests linkage between the disease trait with an entire map of markers. An origin is arbitrarily fixed and the map distance d is now measured relative to that origin. The multipoint LOD score, or location score, is defined to be

$$z(d) = \log_{10} \frac{L(d)}{L(\infty)},$$

where *L(d)* denotes the likelihood that the trait locus is located at a distance *d* on a fixed map consisting of several markers, and L(∞) indicates the likelihood that the trait locus is not on the map (or no linkage).

Location scores have two major advantages over the two-point LOD score analysis that cause them to be the method used in this work. The first is that multipoint linkage analysis results are less sensitive to the uninformative or missing genotype at any single marker, so that this type of analysis can extract more of the total inheritance information from the family. Second, the multipoint LOD score approach can be used to pinpoint a disease-gene location in the mapping of a Mendelian disorder [15]. Usually, this is achieved by the computation of many location scores on a fixed map, as is used in the computational implementation of this work.

1.6 Factors that Can Affect Linkage Analysis

Parametric linkage analysis is based on a prespecified genetic model. The LOD score is a function of both the recombination fraction and the genetic model. If the genetic model is wrong, the true picture of linkage is disguised, leading to either false positive or negative evidence for linkage or for the true location [10].

The impact of misspecified genetic parameters on the LOD score is complicated and depends on several factors, such as the true underlying disease model, the pedigree structures, and the parameters that are misspecified and extent of misspecification. We explain several of these factors in what follows:

- Misspecification of disease allele frequency

  In the case of no linkage, there is little difference in the mean maximum LOD score. In the case of linkage, however, the mean maximum LOD score decreases whenever the disease allele frequencies are underestimated or overestimated. An increased disease allele frequency may have an impact on the LOD score in two ways: either increasing the probability of affected parents being homozygous or increasing the probability that the disease allele is introduced into the pedigree through married-in individuals instead of through a single founder. Nonetheless,

when the frequency is varied at a "reasonable" range, the impact of misspecification of the disease allele frequency on the LOD score and on the estimate of the recombination fraction is usually small [10].

- Misspecification of disease allele penetrance

As long as incomplete penetrance is included in the genetic model, misspecification of disease penetrance has a small impact on the LOD score when there is either linkage or no linkage. As the ratio of penetrances between allele carriers and non-allele carriers decreases, the mean maximum LOD score decreases, since a low ratio decreases the certainty of whether an affected individual is an allele carrier and an unaffected individual a non-allele carrier. However, when there is some degree of phenocopy (phenotypes that are due to some other etiology than the true genotype) and incomplete penetrance, selection of a low ratio is a conservative strategy. Since the clinical phenotype is complex and often confounded by phenocopies, variable expresivity, and penetrance, the complete elimination of phenotypic misspecifications is not feasible [10].

- Misspecification of disease dominance

In general, misspecification of disease dominance has a large impact on the LOD score. This effect is particularly serious when a dominant disease is misspecified as a recessive disease. For example, when a disease that is inherited in a dominant fashioned is analyzed under a recessive model, the random segregation of alleles from the non-disease-allele-carrying parent will be scored (half the time) as a recombination between the disease and marker locus. In addition, the affected parent will be considered homozygous for the disease locus, and thus, uninformative for linkage. A misspecified disease dominance has little impact on LOD score when there is no linkage [10].

- Misspecification of marker allele frequency

Misspecified marker allele frequencies do not always have a large impact on the mean maximum LOD score. When there are many family members without genotype data, incorrect estimates of marker allele frequencies can have a large impact on the LOD score. The sensitivity is directly related to the allele frequency distribution and to the number of ungenotyped founders (those without parents) in the pedigree. The genotype for each ungenotyped founder must be estimated and the population allele frequencies are used. When the parents or grandparents are not available for genotyping, the probability that the allele was present only in the line of descent is calculated from the allele frequencies. If the allele is quite common, there is an increased probability that parents are homozygous for that allele or that married-in family members may have transmitted the allele, and thus there is little evidence for linkage [10].

In this work, we will use the aforementioned biological and statistical information to look at a particular area of the parametric linkage analysis problem. The remainder of this work is divided into the following chapters. In Chapter 2, we introduce an exact method for computing LOD and location scores, the Elston-Stewart algorithm, and discuss the disease Episodic Ataxia (EA) as it relates to analysis by this exact method. We establish in Chapter 3 the MCMC method as discussed in [32], and look at the properties needed for convergence of the Markov chain. Chapter 4 develops the mathematical foundations for the convergence analysis to be applied to the MCMC implemetation of Episodic Ataxia. Chapter 5 discusses the implementation of the MCMC method by *SimWalk2*, a genetics software, for the disease gene Episodic Ataxia, as it relates to location scores and the rate of convergence. We conclude in Chapter 6 with a summary of results and recommendations, and outline possible areas for further research.

2. The Elston-Stewart Algorithm and Episodic Ataxia

The Elston-Stewart algorithm is an exact method used to determine the likelihood scores for a given pedigree [3]. Despite the computational disadvantages of using an exact method in parametric linkage analysis, the Elston-Stewart algorithm is still highly influential and widely used for computing LOD and location scores [37].

2.1 The General Problem

Given a pedigree and phenotypic information about some (or all) of the people in the pedigree, we wish to determine the probability of the observed data, given some probability model for the transmission of alleles. The probability of the observed data is composed of three types of probability functions: founder probabilities, penetrance probabilities, and transmission probabilities.

Founders are those individuals whose parents are not in the pedigree. Thus, probabilities assigned to their genotypes are not based on other members of the pedigree. Instead, these probabilities are computed by assuming the Hardy-Weinberg equilibrium of population genetics [3]. At a biallelic locus with alleles *A* and *a* and corresponding probabilities *p* and *q*, respectively, possible genotypes will exhibit the following frequencies under the Hardy-Weinberg principle:

Genotype	Frequency
AA	p^2
Aa	$2pq$
aa	q^2

These predicted frequencies are the terms in the expansion of the binomial expression $(p+q)^2$ and are referred to as the Hardy-Weinberg genotype frequencies. The key assumption underlying the Hardy-Weinberg principle is that members of the population mate at random with respect to the genes under study. Then with random mating and no differential survival or reproduction among the members of the population, the

Hardy-Weinberg genotype frequencies persist generation to generation, yielding a condition referred to as the Hardy-Weinberg equilibrium. For loci having more than two alleles, this principle, although still valid, now requires a multinomial expansion to determine the genotypic frequencies [36]. Since the loci of a founder are treated as independent, the probability of the multi-locus genotype of founder *K* is

$$P(x_K) = P(x_K^1) \cdot P(x_K^2) \cdot \ldots \cdot P(x_K^n),$$

corresponding to the n loci.

The penetrance probability is the probability of the phenotype given the genotype. Since there is not a one-to-one relationship between genotype and phenotype, this penetrance probability plays an important role in linkage calculations. For example, in a dominant disease with complete penetrance, i.e. for homozygous dominant or heterozygous individuals, the penetrance probability is given by $P(\textit{phenotype} \mid \textit{genotype}) = 1$, and for homozygous recessive individuals,the penetrance probability is given by $P(\textit{phenotype} \mid \textit{genotype}) = 0$. For a recessive disease with incomplete penetrance, a homozygous recessive individual has penetrance probability $0 < P(\textit{phenotype} \mid \textit{genotype}) < 1$. The penetrance probability can be due to sex-dependent, age-dependent, or environment-dependent factors.

Finally, the transmission probability is the probability of a child having a certain genotype given the parents' genotypes. This can be expressed as $P(x_C \mid x_M, x_F)$, where x_C, x_M, and x_F denote the genotypes of a child, male parent, and female parent, respectively. By splitting the ordered genotype x_C into the maternal allele x_{CM} and the paternal allele x_{CF}, the transmission probability becomes

$$P(x_C \mid x_M, x_F) = P(x_{CM} \mid x_M) \cdot P(x_{CF} \mid x_F).$$

Recall that the inheritance from each parent is independent and genotypes of different children are independent given the genotypes of their parents. In this way, transmission probabilities follow Mendel's first law of inheritance [3].

Now given these probabilities, a general formula for the likelihood calculation can be expressed as

$$L(\theta) = \sum_{g_1} \cdots \sum_{g_n} \prod_f P(g_f) \cdot \prod_{\{C,F,M\}} P(g_C \mid g_F, g_M) \cdot \prod_i P(x_i \mid g_i),$$

where $P(g_f)$ is the probability of a founder f having a particular genotypic configuration , $P(g_C \mid g_F, g_M)$ is the transmission probability, and $P(x_i \mid g_i)$ is the penetrance probability for an individual.

2.2 Calculating a Likelihood Using the Elston-Stewart Algorithm

Taking into account all of the information mentioned in section 2.1, the goal of this section is to calculate the likelihood of data in order to determine linkage between two or more loci. As stated previously, the likelihood of the data is the probability of the observed data given certain values for the unknown recombination fractions.

Consider that person *i* has an ordered phenotype $x_i = (x_i^1, x_i^2, \quad \ldots \quad, x_i^n)$ and a multi-locus genotype g_i. Then for a pedigree with *m* people, the likelihood of the data can be expressed as

$$L(\theta) = P(x) = \sum_g P(x, g) = \sum_g P(x \mid g) P(g),$$

where $x = (x_1, x_2 \quad \ldots \quad, x_m)$ and $g = (g_1, g_2, \quad \ldots \quad, g_m)$. Using the Elston-Stewart algorithm, the likelihood function is computed recursively, starting with the most recent generation and working backwards to the most remote. The advantage of the method is that the likelihood for an *individual* can be computed first, and the resulting likelihood then attached as a factor for the computation of his/her parent's likelihood. The individual is no longer needed in the computations that follow. Thus the likelihood of the pedigree with *m* individuals can be expressed as

$$L(\theta) = \sum_{g_i} \cdots \sum_{g_m} \prod_{i=1}^{m} P(x_i \mid g_i) P(g_i \mid \ldots),$$

where x_i denotes the phenotype and g_i denotes the genotype of individual *i* and $P(g_i \mid ...)$ represents the probability of genotype g_i given his parents' genotypes, or the population genotype frequency, in the case that the former is unknown.

2.3 Episodic Ataxia

Episodic Ataxia (EA) is a rare genetic disorder affecting the central nervous system. Affected individuals experience attacks of generalized ataxia brought on by physical or emotional stress, with normal or near-normal neurological function between the attacks. It is considered to be caused by an autosomal dominant disease gene on the 12p13 chromosome. [21] describes the localization of a gene for EA to the chromosome 12p13, where the authors used the methods of molecular biology, and the computational implementation *Fastlink*, which uses the Elston-Stewart algorithm. The study presented evidence that a gene for EA maps to human chromosome 12p, localized to the region between S372 and the KCNA5/S99 cluster, [21]. This was later refined using a pedigree in which all 29 members were available for typing, see Figure A1 in Appendix A. Using the same Elston-Stewart method, the pedigree was analyzed and the LOD scores were computed against a marker map of nine 12p markers. Figure A2 in Appendix A shows a plot of the LOD score between the Episodic Ataxia gene and the marker S372 at various recombination fractions. This pedigree does not prove linkage, though it is strongly suggested.

Finally, the location scores were computed and analysis resulted in the conclusion that the Episodic Ataxia gene resides on the interval from S372 to pY2/1. The figure on the overleaf shows a plot of the location score curve for Episodic Ataxia versus the 12p markers, which illustrates that in the region between S372 and pY2/1, the location scores rise above the level of three [21]. Thus, the implementation of the Elston-Stewart algorithm suggests that the Episodic Ataxia gene resides in the region between the markers S372 and pY2/1, where the largest location score was found to be 3.560.

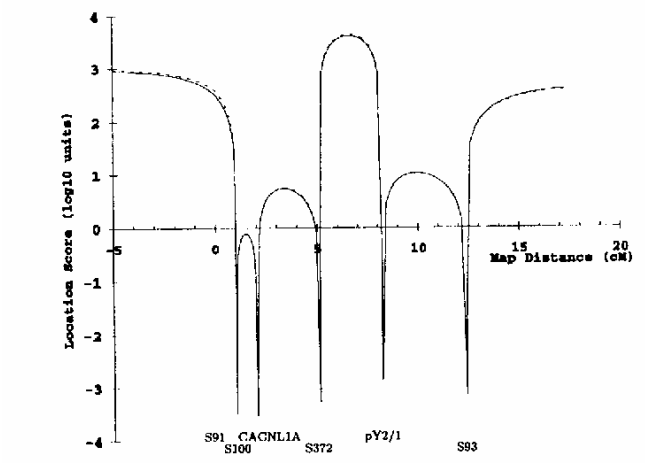


Fig. 2.1 *Location Score Curve for Episodic Ataxia Versus 12p Markers*

Comparison of the commonly-used algorithms (i.e. Elston-Stewart, Lander-Green-Kruglyak, and MCMC algorithms) indicate that large pedigrees considering only a small number of loci can be analyzed easily using the Elston-Stewart algorithm and small pedigrees considering a large number of loci are analyzed easily using the Lander-Green-Kruglyak algorithm (not discussed in this work). See Table C1 in Appendix C. Nonetheless, both of the exact methods have serious limitations when the number of loci and pedigree members are large, and so stochastic (Markov Chain Monte Carlo) methods were developed. The above analysis requires extensive computational time and strength, and thus, the MCMC method for this linkage problem is implemented to improve accuracy and efficiency.

3. A Markov Chain Monte Carlo Method For Computing LOD and Location Scores

Stochastic methods in pedigree analysis have enabled geneticists to handle computations that were previously intractable by standard deterministic methods, [15]. Without any loss of generality, we discuss the basic and relevant principles of Markov chains in what follows.

3.1 Markov Chain Preliminaries

3.1.1 Defining a Markov Chain

A *chain* is a discrete time process in which the random variable X_t undergoes a sequence of changes at a sequence of times or steps. This discrete random variable assumes values $i = 1, 2, ...$ where the actual outcomes are called the *states* of the system, and are denoted by $E_i (i = 1, 2, ...)$. If random variables X_t and X_{t+1} make a transition between the values $X_t = i$ and $X_{t+1} = j$ at the $(t+1)$ th step, then the system has moved from state E_i to state E_j.

A *Markov Chain* has the property that the probability that $X_t = i$ depends *only* on the previous state of the system [4]. That is,

$$P(X_t = j \mid X_{t-1} = i, ..., X_0 = i_0) = P(X_t = j \mid X_{t-1} = i).$$

Note that the probabilities are nonnegative and that since the process must make a transition into some state (which could be itself) at each step, we have a one-step transition probability, denoted by

$$P(X_t = j \mid X_{t-1} = i) = K_{ij}.$$

Let *K* denote the matrix of one-step transition probabilities K_{ij}, so that

$$K = \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1n} \\ K_{21} & K_{22} & \dots & K_{2n} \\ \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & K_{nn} \end{bmatrix}.$$

It should be noted that the transition probability described above has the following property:

$$\sum_{j=1}^{n} K_{ij} = 1, \forall i .$$

Moreover, if there is a distribution π such that if state $X_t \sim \pi$, then $X_{t+1} \sim \pi$, the distribution is called a *stationary, or equilibrium, distribution*. For a Markov process, stationarity is achieved when

$$\sum_{i=1}^{n} \pi_i K_{ij} = \pi_j, \forall j .$$

We can now extend this one-step probability matrix to an *t*-th step probability matrix defining K_{ij}^t to be the probability that a process in state *i* will be in state *j* after *t* transitions.

3.1.2. Communication and Classification of States

State *j* is said to be *accessible* from state *i* if it is possible that the process will ever enter state *j*. In other words, state *j* is accessible from state *i* if and only if $K_{ij}^t > 0$, for some $t \geq 0, i, j \geq 0$. If two states *i* and *j* are accessible from each other, then states *i* and *j* are said to *communicate*, and two states that communicate are said to be in the same *class*. A Markov chain is called *irreducible* if there is only one class, that is, if all states communicate, [28].

Also, we classify each state according to whether reentry into that state is possible once the state has been abandoned. For any state *i*, we let f_i denote the probability that, starting in state *i*, the process will ever reenter state *i*. State *i* is said to be *recurrent* if

$f_i = 1$ and *transient* if $f_i < 1$. Thus, if a process begins its journey in a recurrent state, not only is its return to that state assured, but also, because of Markov properties, that process will return to state *i* infinitely often, [4].

If state *i* is recurrent, then, more specifically, it is called *positive recurrent* if, starting in *i*, the expected time until the process returns to state *i* is finite. According to [4], it can be shown that in a finite-state Markov chain, all recurrent states are positive recurrent.

3.1.3 Periodicity and Ergodicity

State *i* has *period d* if $K_{ij}^t = 0$ whenever *t* is not divisible by *d*, and *d* is the largest integer with this property. For example, if starting in state *i*, a process can reenter state *i* only at times 3, 6, 9, 12, . . . , then state *i* has period 3. Also, a state that has period one is called *aperiodic*, and if positive recurrent, is *ergodic*. If all states in a Markov chain are ergodic, then the chain is described as an *ergodic chain,* [28].

3.2 Methodology of the MCMC Process

The location score, for a location *d,* provides the relative likelihood that a trait locus is at *d*. If you let *T* represent the trait phenotype data and *M* represent the marker genotype data, then for a trait location *d*, the location score is

$$P_d(T \mid M) = \sum_i P_d(T \mid G_i) P(G_i \mid M),$$

where G_i is the complete gene flow configuration of the marker loci. It is obvious that this is an expectation, where

$$E(P_d(T \mid G_i)) = \sum_i P_d(T \mid G_i) P(G_i \mid M).$$

This conditional probability

$$P(G_i \mid M) = \frac{P(G_i \cap M)}{\sum_j P(G_j \cap M)}$$

sums over all underlying complete genetic states that are consistent with the data. The factor $\sum_j P(G_j \cap M)$ found in this conditional probability is the normalizing factor, equivalent to $P(M)$. Deterministic methods must calculate this sum, but a stochastic method is not restricted in this manner. Instead, the stochastic methods are used to find a good estimate of this likelihood, in which one can sample at random from all possible configurations *in proportion* to their individual likelihood and then average the sampled values. In fact, as the number of random samples increases this estimate must come arbitrarily close to the exact value, for any distribution of the $P(G_i \mid M)$. Thus, the location score becomes

$$P_d(T \mid M) = (1/n)\sum_{k=1}^{n} P_d(T \mid G_k),$$

where each of the *n* configurations G_k is sampled in proportion to its likelihood, $P_d(G_k \mid M)$. This strategy takes into account that a relative handful of the configurations might be very likely, and although the others are possible, they are not very likely. In order to sample the underlying complete configurations G_k in proportion to their likelihoods, Markov Chain Monte Carlo procedures are used [32].

3.3 Descent Graph Markov Chain Method for Linkage Analysis

3.3.1 MC State Space

To apply the MCMC method to the analysis of human pedigree data, we must choose an appropriate state space and a mechanism for moving between neighboring states of the space. Given our goal of computing location scores, the state space must capture gene flow at multiple marker loci. Note that the state space will omit mention of the trait locus; this locus will be handled somewhat differently. The states of our space are rather complicated graphs describing the gene flow in a pedigree at the participating marker

loci. It will suffice to focus on a single pedigree because location scores are computed pedigree by pedigree. We explain the mechanism of this process in what follows.

Suppose that we observe marker phenotypes at ***l*** loci of a pedigree with ***p*** people of whom ***f*** are founders. A genetic descent state completely specifies the gene flow within the pedigree at these loci. Gene flow can be separated into two parts: the paths that the founder genes take as they descend through the pedigree and the allelic form assumed by each of the founder genes. The paths of the gene flow in a genetic descent state constitutes a genetic descent graph, possessing ***2lp*** nodes, each of which is a particular combination of locus, person, and source. If we think of a gene occupying each node, then the source of the node gives whether the gene is maternal or paternal in origin. Rooted at each founder node there is a directed tree incorporating exactly those nodes in the genetic descent graph that inherit the corresponding founder gene. This genetic descent tree forms a connected component of the genetic descent graph, and in all, there are ***2lf*** descent trees. The following figures give the gene flow in a fully typed pedigree.

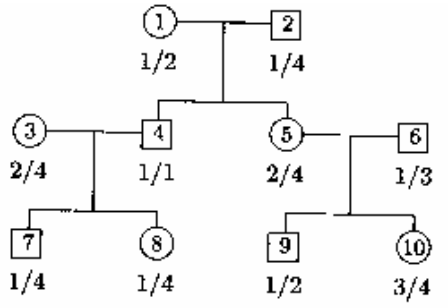


Fig. 3.1(a) *Conventional Pedigree Representation*

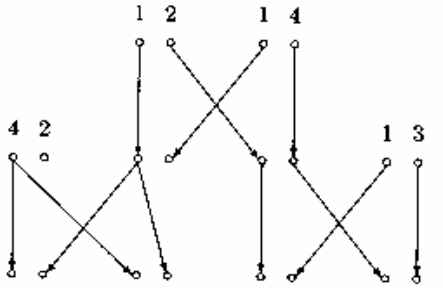


Fig. 3.1(b) *Descent State for Gene Flow*

A descent state at a locus determines an ordered genotype for each and every person in the pedigree. Some descent states are consistent with the observed phenotypes of the pedigree, and some are not. Those that are consistent with the observed phenotype information are considered legal descent states. If a descent graph is consistent with at least one legal descent state, then the descent graph is legal; otherwise, it is illegal. Obviously, the collection of descent graphs is much smaller than the collection of descent states. This is one reason for preferring descent graphs to descent states as points of the state space. The size of the state space is further diminished by allowing only legal descent graphs. One final level of abstraction is that of founder tree graph. The nodes of

the founder tree graph are the descent trees of the descent graph. This level provides a method for keeping track of how founder alleles are constrained in a coupled manner by the observed marker phenotypes in the pedigree. Two nodes of the founder tree graph are connected by an edge if and only if the two corresponding descent trees pass through the same typed locus of some person in the pedigree. Thus, two descent trees associated with different loci cannot be connected in this manner [32].

3.3.2 The descent graph Markov Chain

The set of descent graphs over a pedigree becomes a Markov chain if we incorporate transition rules for moving between the descent graphs. The most basic transition rule, generally known as rule T_0, switches the origin of an arc descending from a parent to a child from the paternal maternal node to the paternal paternal node, or vice versa. The figure shown below depicts an example of this type of transition.

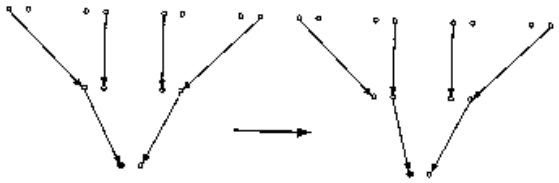


Fig 3.2 *Example of Transition Rule T_0*

From the basic rule T_0, other composite rules are designed to make more radical changes in an existing descent graph, and consequently speed up the circulation of the chain. Transition rule T_1, example shown in the figure below, begins by choosing person ***i*** and locus ***l***. It then performs a T_0 transition at each node determined by a child of ***i***, the locus ***l***, and the sex of ***i***. Thus every child of ***i*** who previously inherited ***i***'s maternal gene now inherits ***i***'s paternal gene and vice versa.

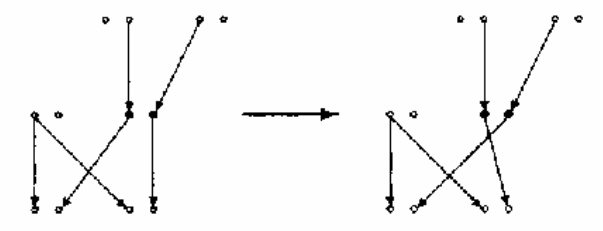


Fig. 3.3 *Example of Transition Rule T_1*

The second composite transition rule has two variations: T_{2a} and T_{2b}, as shown in the examples below. Each variation begins by choosing a locus ***l*** and a couple ***i*** and ***j*** with common children. Four different descent subtrees are rooted at the parents ***i*** and ***j***. Rule T_{2a} exchanges the subtree rooted at the maternal node of ***i*** with the subtree rooted at the maternal node of ***j***; it similarly exchanges the paternally rooted subtrees of ***i*** and ***j***.

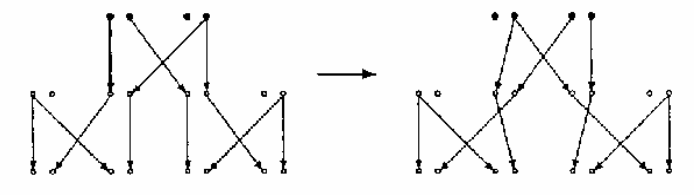


Fig. 3.4 *Example of Transition Rule T_{2a}*

Rule T_{2b} exchanges the maternally rooted subtree of ***i*** with the paternally rooted subtree of ***j*** and vice versa.

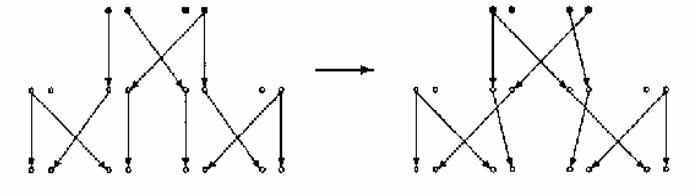


Fig 3.5 *Example of Transition Rule T_{2b}*

It should be noted, that after swapping subtrees, there are paternally derived genes flowing to maternal nodes and vice versa. Adjustments must be made in the children and grandchildren to correct these illegal patterns of gene flow. Also, only the paths descending through the children shared with the chosen spouse are pertinent and not any children coming from additional spouses.

One complication in constructing the Markov chain on legal descent graphs is that two states may not communicate in the presence of three or more alleles per marker. The following figure describes how two descent graphs fail to communicate.

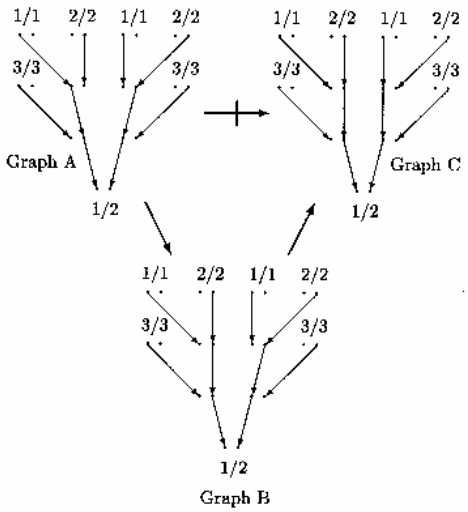


Fig. 3.6 *Failure of Descent Graphs A and C to Communicate*

In this figure, Graph A and Graph C are said to not communicate. Graphs A and C are both legal descent graphs, but you cannot move from A to C because B, the intermediate step, is illegal. To solve this problem, it is necessary to "tunnel through" illegal descent graphs by taking multiple transitions per step of the Markov chain. In applications, a random number of transitions per step of the Markov chain are employed in order to pass through illegal descent graphs on the way between legal descent graphs. To maintain reversibility of the Markov chain, it is necessary that, within a step, independent choices are made for both the sequence of transition rules invoked and the sequence of people and sources to which these transitions are applied [32].

3.3.3 Likelihood of a Descent Graph

The equilibrium distribution π of our Markov chain should match the distribution of legal descent graphs $\hat{G}$ conditioned on the observed marker genotypes M of the pedigree. Because the normalizing factor $P(M)$ is irrelevant in applying the Metropolis algorithm (as we will do later), it suffices to calculate the joint probabilities $P(\hat{G} \cap M)$ as opposed to the conditional probabilities $P(G \mid M)$, in order to calculate the location score.

The joint likelihood of a descent graph $\hat{G}$ and a marker genotype vector *M* can be written as a sum of the joint likelihoods of *M* and all descent states *G* consistent with the gene flow specified by $\hat{G}$. That is,

$$P(\hat{G} \cap M) = \sum_{G \to \hat{G} \cap M} P(G),$$

where $G \to \hat{G} \cap M$ denotes consistency between *G* and both $\hat{G}$ and *M*. Under Hardy-Weinberg and linkage equilibrium for linked markers, the probability *P(G)* reduces to the product of the founder allele frequencies involved in the descent state *G*, and the recombination fractions and their complements for the adjacent intervals separating the markers. Designating the founder allele frequency *Prior(G)* and the transmission probability *Trans(G),* the previous likelihood function becomes

$$P(\hat{G} \cap M) = Trans(\hat{G}) \cdot \sum_{G \to \hat{G} \cap M} \Pr ior(G).$$

This result comes from the fact that all compatible descent states $G \to \hat{G}$ exhibit the same transmission pattern, and *Trans(G)* depends only on the descent graph $\hat{G}$ and not on the particular representative chosen from the set of $\{G : G \to \hat{G} \cap M\}$. To further simplify the formula, the connected components of the founder tree graph are labeled $C_1, C_2, ..., C_m$, and for a given consistent descent state $\{G : G \to \hat{G} \cap M\}$, the vector of alleles assigned to component C_i are labeled a_i. Then since each founder gene is sampled independently,

$$\Pr ior(G) = \prod_{i=1}^{m} P(a_i),$$

where $P(a_i) = \prod_j P(a_{ij})$, a_{ij} being the components of the allele vector a_i. By construction, the founder genes assigned to different components do not impinge on one another. That is, the set of founder genes consistent with $\hat{G}$ and M is drawn from the

product of the sets $S_1,...,S_m$ of legal allele vectors for the components $C_1, C_2,..., C_m$, respectively. So that applying the distributive rule to the above equation yields

$$\sum_{G \to \hat{G} \cap M} \Pr ior(G) = \prod_{i=1}^{m} P(C_i),$$

where $P(C_i) = \sum_{a_i \in S_i} P(a_i)$. Note that a set S_i contains either all, one, two, or no allele vectors. When S_i contains all allele vectors, $\sum_{a_i \in S_i} P(a_i) = 1$, and when S_i contains no allele vectors, $\sum_{a_i \in S_i} P(a_i) = 0$. If S_i contains either one or two allele vectors, the product formula $P(a_i) = \prod_j P(a_{ij})$ is applicable. These results now give the equation

$$P(\hat{G} \cap M) = Trans(\hat{G}) \cdot \prod_{i=1}^{m} P(C_i).$$

and can be used in the computation of the location scores [32].

3.3.4 Computing the Location Score

Previously, we discussed in section 3.2 how location scores are used to position a trait locus relative to an existing set of mapped markers. Without loss of generality, suppose the unknown trait position is denoted by *d,* the trait phenotypes for a pedigree by *T*, and the marker genotypes by *M*. For a trait location, the location score is the expectation

$$P_d(T \mid M) = \sum_i P_d(T \mid G_i) P(G_i \mid M),$$

where G_i is a complete gene flow configuration of the marker loci. The point of this MCMC method is to speed up the calculations by sampling from all possible underlying configurations in proportion to their likelihood and then taking the average of these values. So if we consider the sequence of descent graphs $\hat{G}_1,...,\hat{G}_n$ generated by running

the Markov chain, as the underlying configurations, then the sample average $\frac{1}{n}\sum_{i=1}^{n} P_d(T \mid \hat{G}_i)$ will approximate $P_d(T \mid M)$ well for sufficiently large *n,* [15].

3.3.5 Constructing the Markov Chain

In order to create a Markov chain using the aforementioned transition rules that will have the correct equilibrium distribution, the Metropolis algorithm is implemented. This algorithm consists of two different stages: the proposal stage and the acceptance stage. In the proposal stage, the number and type of transitions, and the pivot nodes are chosen. The probability that $\hat{G}_j$ is the proposed next descent graph given that $\hat{G}_i$ is the current descent graph is expressed as $q_{ij} = P(\hat{G}_j | \hat{G}_i)$, and is called the *proposal probability* for moving from descent graph $\hat{G}_i$ to descent graph $\hat{G}_j$. Because a single step can consist of an unlimited number of transitions, it is possible to move from any legal descent graph to any other legal descent graph in one step. Thus, all entries of the proposal matrix $Q = \{q_{ij}\}$ will be positive. The acceptance stage occurs once a step $\hat{G}_i$ to $\hat{G}_j$ is proposed. At this point, it is accepted with the Metropolis probability

$$a_{ij} = \min\left\{1, \frac{P(\hat{G}_j \mid M)}{P(\hat{G}_i \mid M)}\right\} = \min\left\{1, \frac{P(\hat{G}_j \cap M)}{P(\hat{G}_i \cap M)}\right\}.$$

On the basis of this criterion, a more likely descent graph is always accepted, and an illegal descent graph is always rejected. A less likely but still legal descent graph is sometimes accepted and sometimes rejected. Then the transition probability of moving from $\hat{G}_i$ to $\hat{G}_j$ is

$$K_{ij} = \begin{cases} q_{ij} a_{ij}, & j \neq i \\ 1 - \sum_{k \neq i} K_{ik}, & j = i \end{cases}$$

Since the proposal matrix *Q* has all positive entries, the matrix $K = \{K_{ij}\}$ is also positive on the set of all legal descent graphs. Due to this property, the Markov chain on the legal descent graphs is irreducible and aperiodic. Properties of detailed balance and reversibility will be discussed in what follows [32].

3.3.6 Detailed Balance and Overall Balance

When there is the same probability of arriving at a certain state *j* as departing from it, or

$$\pi_i K_{ij} = \pi_j K_{ji},$$

this is referred to as the detailed balance or reversibility.

Lemma 3.1*: Given that* $t \to \infty$ *and* $x_t \to \pi$, *where* π *is the stationary distribution of the Markov chain, the Metropolis method above guarantees the detailed balance* $\pi_i K_{ij} = \pi_j K_{ji}$, *where* K_{ij} *is a transitional probability from state i to state j.*

Proof*:* The original Metropolis algorithm [42] enables us to write $K_{ij} = \min(1, \frac{\pi_j}{\pi_i})$.

Hence,

$$\begin{aligned}
\pi_i K_{ij} &= \pi_i \min(1, \frac{\pi_j}{\pi_i}) \\
&= \min(\pi_i, \pi_j) \\
&= \min(\pi_j, \pi_i) \\
&= \pi_j \min(1, \frac{\pi_i}{\pi_j}) \\
&= \pi_j K_{ji} \qquad \Delta
\end{aligned}$$

Further, the detailed balance guarantees the overall balance

$$\sum_{j \neq i} \pi_j K_{ji} = \sum_{i \neq j} \pi_i K_{ij},$$

as can be seen in the next proof.

***Lemma 3.2**: Given that $t \to \infty$ and $x_t \to \pi$, where π is the stationary distribution of the Markov chain, the detailed balance $\pi_i K_{ij} = \pi_j K_{ji}$ guarantees the overall balance $\sum_{j \neq i} \pi_j K_{ji} = \sum_{i \neq j} \pi_i K_{ij}$ of the Markov chain.*

Proof:

$$\pi_j = \sum_{i=1}^{n} \pi_i K_{ij} = \sum_{i \neq j} \pi_i K_{ij} + \pi_j K_{jj}$$

$$\pi_j - \pi_j K_{jj} = \sum_{i \neq j} \pi_i K_{ij}$$

$$\pi_j (1 - K_{jj}) = \sum_{i \neq j} \pi_i K_{ij}$$

$$\pi_j \sum_{j \neq i} K_{ji} = \sum_{j \neq i} \pi_j K_{ji} = \sum_{i \neq j} \pi_i K_{ij}$$

Therefore, $\sum_{j \neq i} \pi_j K_{ji} = \sum_{i \neq j} \pi_i K_{ij}$, which gives the overall balance [42]. Δ

The properties of the descent graph Markov chain method, as described above, will be used in the convergence analysis to follow. It should be noted that the MCMC method yields a Markov chain that is irreducible, aperiodic, and symmetric (i.e. $K_{ij} = K_{ji}$), as shown in the previous sections. These properties will be necessary for much of the analysis in the next section.

4. Convergence Analysis of the MCMC Method

Although the application of a MCMC algorithm can prove most useful in linkage analysis, it is still necessary to assess the validity of this stochastic technique – most importantly, perhaps, the convergence of the proposed estimators to the true limit. In order to do this, we need to analyze the transition matrix of our Markov chain.

4.1 Transition Matrix of the Markov Chain

The transition probabilities $K_{ij} = P(X_t = j \mid X_{t-1} = i)$ developed in section 3.3.5 control the progress of the Markov chain. We define the marginal distribution of X_t to be $P_i^t = P(X_t = i)$, where

$$P_i^t = \sum_{j=1}^{n} P(X_{t-1} = j) P(X_t = i \mid X_{t-1} = j) = \sum_{j=1}^{n} P_j^{t-1} K_{ji},$$

or as expressed in matrix notation, $P^t = P^{t-1}K$, where $P^t = (P_1^t, ..., P_n^t)$. It is easy to see that

$$P^t = P^{t-1}K = P^{t-2}K^2 = ... = P^0 K^t,$$

where $P_i^0 = P(X_0 = i)$ and $K_{ij}^t = P(X^t = j \mid X_0 = i)$. If we prove that $K_{ij}^t \to \pi_j$ as $t \to \infty$, then we will have that $K^t = (\pi, \pi, ..., \pi)'$ as $t \to \infty$, and consequently, that the Markov chain has reached stationarity [43].

4.2 Eigen Analysis of the Transition Matrix

Suppose that we define $diag(\pi)^{1/2} \equiv \pi^{1/2} I$, where I is the identity matrix. From the reversibility property, we have $\pi_i K_{ij} = \pi_j K_{ji}$ and multiplying both sides by $\frac{1}{\sqrt{\pi_i \pi_j}}$, we have $\sqrt{\frac{\pi_i}{\pi_j}} K_{ij} = \sqrt{\frac{\pi_j}{\pi_i}} K_{ji}$.

***Proposition 4.1**: Let matrix M be such that $M \equiv diag(\pi)^{1/2} \cdot K \cdot diag(\pi)^{-1/2}$, then M is also symmetric.*

Proof:

Let $M \equiv diag(\pi)^{1/2} \cdot K \cdot diag(\pi)^{-1/2}$, then

$$M_{ij} = diag(\pi_i)^{1/2} \cdot K_{ij} \cdot diag(\pi_j)^{-1/2} = \pi_i^{1/2} \cdot I \cdot K_{ij} \cdot \pi_j^{-1/2} \cdot I = \frac{\pi_i^{1/2}}{\pi_j^{1/2}} \cdot K_{ij} = \sqrt{\frac{\pi_i}{\pi_j}} \cdot K_{ij}$$

On the other hand,

$$M_{ji} = diag(\pi_j)^{1/2} \cdot K_{ji} \cdot diag(\pi_i)^{-1/2} = \pi_j^{1/2} \cdot I \cdot K_{ji} \cdot \pi_i^{-1/2} \cdot I = \frac{\pi_j^{1/2}}{\pi_i^{1/2}} \cdot K_{ji} = \sqrt{\frac{\pi_j}{\pi_i}} \cdot K_{ji}$$

Since $\sqrt{\frac{\pi_i}{\pi_j}} K_{ij} = \sqrt{\frac{\pi_j}{\pi_i}} K_{ji}$, we have $M_{ij} = M_{ji}$, and M is symmetric. Δ

Finally, note that

$$M^t = \left(diag(\pi)^{1/2} \cdot K \cdot diag(\pi)^{-1/2}\right) \cdots \left(diag(\pi)^{1/2} \cdot K \cdot diag(\pi)^{-1/2}\right)$$
$$= diag(\pi)^{1/2} \cdot K^t \cdot diag(\pi)^{-1/2}$$

Now assume that $M = Q \wedge Q' = \sum_{i=1}^{n} \lambda_i \bar{q}_i \bar{q}'$, and $M^t = Q \wedge^t Q' = \sum_{i=1}^{n} \lambda_i^t \bar{q}_i \bar{q}_i'$, where Q is an arbitrary matrix such that $Q = (\bar{q}_1, ..., \bar{q}_n)$, $Q' = (\bar{q}_1', ..., \bar{q}_n')$, and $QQ' = I = Q'Q$. Generally, the set $(\bar{q}_1, ..., \bar{q}_n)$ is defined to be a basis, and as such for any vector $\bar{v} = (v_1, ..., v_n)$,

$\bar{v} = \sum_{i=1}^{n} \langle \bar{v}, \bar{q}_i \rangle \bar{q}_i$.

***Proposition 4.2**: For a basis $(\bar{q}_1, ..., \bar{q}_n)$ and any vector $\bar{v} = (v_1, \ldots, v_n)$, $\bar{v} = \sum_{i=1}^{n} v_i \bar{q}_i^2$.*

***Proof**:*

$$\bar{v} = \sum_{i=1}^{n} \langle \bar{v}, \bar{q}_i \rangle \bar{q}_i = \langle (v_1, ... v_n), \bar{q}_1 \rangle \cdot \bar{q}_1 + \langle (v_1, ..., v_n), \bar{q}_2 \rangle \cdot \bar{q}_2 + ... + \langle (v_1, ..., v_n), \bar{q}_n \rangle \cdot \bar{q}_n$$

$$= (v_1 \bar{q}_1 + v_2 \bar{q}_1 + ... + v_n \bar{q}_1)\bar{q}_1 + (v_1 \bar{q}_2 + v_2 \bar{q}_2 + ... + v_n \bar{q}_2)\bar{q}_2 + ... + (v_1 \bar{q}_n + v_2 \bar{q}_n + ... + v_n \bar{q}_n)\bar{q}$$
$$= (v_1 + v_2 + ... + v_n)\bar{q}_1^2 + (v_1 + v_2 + ... + v_n)\bar{q}_2^2 + ... + (v_1 + v_2 + ... + v_n)\bar{q}_n^2$$

$$= \sum_{i=1}^{n} v_i \bar{q}_i^2 \qquad \Delta$$

If we define $u_i = v_i \bar{q}_i$, or $\bar{u} = (\langle \bar{v}, \bar{q}_1 \rangle, ..., \langle \bar{v}, \bar{q}_n \rangle)$, then $\bar{v} = Q\bar{u}$, or $\bar{u} = Q^{-1}\bar{v}$.

Proposition 4.3: *Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ be the eigenvalues of the transition matrix K, and $\bar{u} = (\bar{u}_1, ..., \bar{u}_n)'$ be defined as above, then the maximal eigenvalue λ_1 is achieved when $\bar{u} = (1,0,...,0)'$ and $Q\bar{u} = (\bar{q}_1, ..., \bar{q}_n) \cdot (1,0,...,0)' = \bar{q}_1$.*

***Proof*:**

Let $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ be the eigenvalues of the transition matrix *K*, and define the ratio R to be

$$R \equiv \frac{\bar{v}' M \bar{v}}{\bar{v}' \bar{v}}.$$

Then,

$$R \equiv \frac{\bar{v}' M \bar{v}}{\bar{v}' \bar{v}} = \frac{(Q\bar{u})' M (Q\bar{u})}{(Q\bar{u})'(Q\bar{u})} = \frac{(Q\bar{u})'(Q \wedge Q')(Q\bar{u})}{(Q\bar{u})'(Q\bar{u})} = \frac{(\bar{u}'Q')(Q \wedge Q')(Q\bar{u})}{(\bar{u}'Q')(Q\bar{u})} = \frac{\bar{u}' \wedge \bar{u}}{\bar{u}'\bar{u}}.$$

Note that $\bar{u} = (\bar{u}_1, ..., \bar{u}_n)'$, so that $\bar{u}'\bar{u} = \sum_{i=1}^{n} \bar{u}_i^2$, and as before,

$$\bar{u}' \wedge \bar{u} = \sum_{i=1}^{n} \lambda_i \bar{u}_j' \bar{u}_i = \sum_{i=1}^{n} \lambda_i \bar{u}_i^2 .$$

Then, $$R = \frac{\sum_{i=1}^{n} \lambda_i \bar{u}_i^2}{\sum_{i=1}^{n} \bar{u}_i^2}.$$

By the assumption that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$,

$$R = \frac{\sum_{i=1}^{n} \lambda_i \bar{u}_i^2}{\sum_{i=1}^{n} \bar{u}_i^2} \leq \frac{\sum_{i=1}^{n} \lambda_1 \bar{u}_i^2}{\sum_{i=1}^{n} \bar{u}_i^2} = \lambda_1 \cdot \frac{\sum_{i=1}^{n} \bar{u}_i^2}{\sum_{i=1}^{n} \bar{u}_i^2} = \lambda_1 .$$

If we normalize $\bar{u}$ to a unit length so that, $|\bar{u}| = \sqrt{\sum_{i=1}^{n} \bar{u}_i^2} = 1$, than all $\bar{u}_i \le 1$, and the maximal eigenvalue λ_1 is achieved when $\bar{u} = (1,0,...,0)'$ and $Q\bar{u} = (\bar{q}_1,...,\bar{q}_n) \cdot (1,0,...,0)' = \bar{q}_1$. Thus at this point, $\bar{v} = Q\bar{u} = \bar{q}_1$. Δ

Similarly, the ratio achieves its second largest eigenvalue λ_2 when $\bar{u} = (0,1,0,...0)$ and $\bar{v} = Q\bar{u} = \bar{q}_2$.

4.2.1 The Maximal Eigenvalue for the Transition Matrix

Previously, $\bar{v}$ was defined to be any vector. Let us now define it as

$$\bar{v} = (\sqrt{\pi_1}h_1, \sqrt{\pi_2}h_2, ..., \sqrt{\pi_n}h_n)', \text{ or } v_i = \sqrt{\pi_i}h_i,$$

where h_i is an arbitrary function, and $h = \{h_i\}$.

***Lemma 4.1**: Let $v_i = \sqrt{\pi_i}h_i$. At steady state, the ratio R can be defined as*

$$R = \frac{E[2 \cdot h(X_t)h(X_{t+1})]}{E[(h(X_t))^2 + (h(X_{t+1}))^2]} \le 1$$

implying that the maximal eigenvalue assumes a unit value.

Proof:

Consider the ratio $R \equiv \frac{\bar{v}'M\bar{v}}{\bar{v}'\bar{v}}$, where

$$\bar{v}'M\bar{v} = \sum_{j=1}^{n}\sum_{i=1}^{n} v_i M_{ij} v_j = \sum_{ij} M_{ij} v_i v_j,$$

and recall that, $M_{ij} = \sqrt{\frac{\pi_i}{\pi_j}} K_{ij}$.

Then the ratio becomes

$$R \equiv \frac{\bar{v}' M \bar{v}}{\bar{v}' \bar{v}}$$

$$= \frac{\sum_{ij} M_{ij} v_i v_j}{\bar{v}' \bar{v}} = \frac{\sum_{ij} M_{ij} v_i v_j}{\sum_i v_i^2} = \frac{\sum_{ij} \sqrt{\frac{\pi_i}{\pi_j}} K_{ij} v_i v_j}{\sum_i v_i^2} = \frac{\sum_{ij} \sqrt{\frac{\pi_i}{\pi_j}} K_{ij} \sqrt{\pi_i} h_i \sqrt{\pi_j} h_j}{\sum_i (\sqrt{\pi_i} h_i)^2}$$

$$= \frac{\sum_{ij} \pi_i K_{ij} h_i h_j}{\sum_i \pi_i (h_i^2)}$$

Now consider the denominator and numerator separately. By definition, $E(Y) = \sum_y y \cdot p(y)$, and $P(X_t = j \mid X_{t-1} = i) = K_{ij}$.

Also by conditioning, namely, $E[X] = E[E[X \mid Y]]$, we have

$$\sum_{ij} \pi_i K_{ij} h_i h_j = \sum_j \sum_i \pi_i K_{ij} h_i h_j$$
$$= \sum_j \sum_i P(X_{t-1} = i) P(X_t = j \mid X_{t-1} = i) h(X_{t-1}) h(X_t)$$
$$= E[E[h(X_{t-1}) h(X_t) \mid h(X_{t-1})]]$$
$$= E[h(X_{t-1}) h(X_t)]$$
$$= E[h(X_t) h(X_{t+1})], \text{ at stationarity.}$$

And similarly,

$$\sum_i \pi_i h_i^2 = \sum_i P(X_t = i)(h(X_t))^2$$
$$= E[(h(X_t))^2]$$

Hence at steady state, the ratio becomes:

$$R = \frac{\sum_{ij} \pi_i K_{ij} h_i h_j}{\sum_i \pi_i h_i^2} = \frac{E[h(X_t) h(X_{t+1})]}{E[(h(X_t))^2]}$$

$$= \frac{2}{2}\left(\frac{E[h(X_t)h(X_{t+1})]}{E[(h(X_t))^2]} \right) = \frac{E[2 \cdot h(X_t)h(X_{t+1})]}{E[2 \cdot (h(X_t))^2]} = \frac{E[2 \cdot h(X_t)h(X_{t+1})]}{E[(h(X_t))^2 + (h(X_t))^2]}$$
$$= \frac{E[2 \cdot h(X_t)h(X_{t+1})]}{E[(h(X_t))^2 + (h(X_{t+1}))^2]}$$

And as desired at steady state, $R = \frac{E[2 \cdot h(X_t)h(X_{t+1})]}{E[(h(X_t))^2 + (h(X_{t+1}))^2]}$.

We know that by Cauchy's Inequality, $\left(\sum_k a_k b_k \right)^2 \le \left(\sum_k a_k^2 \right) \cdot \left(\sum_k b_k^2 \right)$.

Let $a_k = X_k$, $b_k = 1$, and $1 \le k \le 2$, then

$$\left(\sum_{k=1}^{2} a_k^2 \right) \cdot \left(\sum_{k=1}^{2} b_k^2 \right) = \left(\sum_{k=1}^{2} X_k^2 \right)(2) = 2X_1^2 + 2X_2^2$$

and

$$\left(\sum_{k=1}^{2} a_k b_k \right)^2 = (X_1 + X_2)^2 = X_1^2 + 2X_1X_2 + X_2^2.$$

The inequality implies that $X_1^2 + 2X_1X_2 + X_2^2 \le 2X_1^2 + 2X_2^2$, or $2X_1X_2 \le X_1^2 + X_2^2$. Now applying this result to the ratio above, we have

$$E[2 \cdot h(X_t)h(X_{t+1})] \le E[(h(X_t))^2 + (h(X_{t+1}))^2],$$

and therefore, $R \le 1$.

Recall that,

$$R = \frac{\sum_{i=1}^{n} \lambda_i \bar{u}_i^2}{\sum_{i=1}^{n} \bar{u}_i^2},$$

and that when the maximal eigenvalue λ_1 is realized, then $\bar{u} = (1,0,...,0)$ and the ratio $R = \lambda_1$. Since the maximum value of the ratio is one, we have λ_1=1. Δ

4.2.2 The Transition Matrix Yielding Stationarity

Using Cauchy's Inequality, it has been realized that $R \le 1$. And $R = 1$ only when $h(X_t) = h(X_{t+1})$, as in that case

$$R = \frac{E[2 \cdot h(X_t)h(X_{t+1})]}{E[(h(X_t))^2 + (h(X_{t+1}))^2]} = \frac{E[2 \cdot (h(X_t))^2]}{E[2 \cdot (h(X_t))^2]} = 1.$$

Notice that this will occur only when h_i is constant. Then for

$$v_i = \sqrt{\pi_i} h_i = \sqrt{\pi_i} C,$$

where C is a constant, and with $\bar{v}$ normalized to one,

$$\sqrt{\sum_{i=1}^{n} v_i^2} = \sqrt{\sum_{i=1}^{n} \pi_i C^2} = C\sqrt{\sum_{i}^{n} \pi_i} = 1.$$

But since π is a probability distribution, the normalizing condition $\sum_{i=1}^{n} \pi_i = 1$ yields $C = 1$.

Thus, we know that when λ_1 is maximized,

$$\bar{q}_1 = \bar{v} = (v_1, ..., v_n)' = (\sqrt{\pi_1}, ..., \sqrt{\pi_n})'.$$

***Theorem 4.1**: As $t \to \infty$, $K^t = (\pi, \pi, ..., \pi)'$, or the Markov chain reaches equilibrium.*

***Proof**:*

Recall from proposition 4.1 that,

$$M^t = diag(\pi)^{1/2} K^t diag(\pi)^{-1/2},$$

and therefore,

$$K^t = diag(\pi)^{-1/2} M^t diag(\pi)^{1/2}.$$

By definition of M,

$$K^t = \left(diag(\pi)^{-1/2}\right)\left(\sum_{i=1}^{n} \lambda_i^t \overline{q}_i \overline{q}_i'\right)\left(diag(\pi)^{1/2}\right).$$

Since $\lambda_1 = 1$, and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$,

$$\sum_{i=1}^{n} \lambda_i^t \to 1, \text{ as } t \to \infty .$$

Hence, $\sum_{i=1}^{n} \lambda_i^t \overline{q}_i \overline{q}_i' \to \overline{q}_1 \overline{q}_1'$,

and therefore, $K^t \to \left(diag(\pi)^{-1/2}\right) \cdot I \cdot \left(\overline{q}_1 \overline{q}_1'\right) \cdot \left(diag(\pi)^{1/2}\right) \cdot I$.

Now,

$$\left(diag(\pi)^{-1/2}\right) \cdot I \cdot \left(\overline{q}_1 \overline{q}_1'\right) \cdot \left(diag(\pi)^{1/2}\right) \cdot I$$
$$= (diag(\pi)^{-1/2}) \cdot I \cdot (\sqrt{\pi_1},...,\sqrt{\pi_n})' \cdot (\sqrt{\pi_1},...,\sqrt{\pi_n}) \cdot (diag(\pi)^{1/2}) \cdot I$$

Let $L = (diag(\pi)^{-1/2}) \cdot I \cdot (\sqrt{\pi_1},...,\sqrt{\pi_n})' = (1, ... ,1)'$,

then

$$L \cdot (\sqrt{\pi_1},...,\sqrt{\pi_n}) = \begin{bmatrix} \sqrt{\pi_1} & \sqrt{\pi_2} & . & . & \sqrt{\pi_n} \\ \sqrt{\pi_1} & \sqrt{\pi_2} & . & . & \sqrt{\pi_n} \\ . & . & . & . & . \\ . & . & . & . & . \\ \sqrt{\pi_1} & \sqrt{\pi_2} & . & . & \sqrt{\pi_n} \end{bmatrix} = M ,$$

and $M \cdot (diag(\pi)^{1/2}) = (\pi,...,\pi)'$.

Thus $K^t = (\pi, \pi,..., \pi)'$, as $t \to \infty$. Δ

The proof above shows that after a sufficiently long time, K reaches the equilibrium state; that is $\pi \cdot K = \pi$. Thus, under the Metropolis algorithm, the Markov chain will eventually reach its stationary distribution [43].

4.3 Rate of Convergence of a Markov Chain

Now that we have shown that the Markov Chain will reach its stationary distribution after a given amount of time, we are really interested in how much time is necessary to attain stationarity. In other words, what is the rate of convergence? From the previous analysis, when λ_1=1 and $\lambda_2 \le 1$, $K^t \to (\pi,...,\pi)'$ as $t \to \infty$. How fast this result is attained depends on the value of λ_2. If $\lambda_2 \approx 1$, then the convergence rate will be very slow.

4.3.1 Bound on Convergence

From the Cauchy-inequality, we obtain that $2X_1X_2 \le (X_1^2 + X_2^2)$, and similarly, that $-2X_1X_2 \le (X_1^2 + X_2^2)$. Using this result, we know that

$$R = \frac{E[2 \cdot h(X_t)h(X_{t+1})]}{E[(h(X_t))^2 + (h(X_{t+1}))^2]} \ge -1.$$

But if we implement the previous definition of *R*,

$$R = \frac{\sum_{i=1}^{n} \lambda_i \bar{u}_i^2}{\sum_{i=1}^{n} \bar{u}_i^2}$$

we get the result

$$R = \frac{\sum_{i=1}^{n} \lambda_i \bar{u}_i^2}{\sum_{i=1}^{n} \bar{u}_i^2} \ge \frac{\sum_{i=1}^{n} \lambda_n \bar{u}_i^2}{\sum_{i=1}^{n} \bar{u}_i^2} = \frac{\lambda_n \sum_{i=1}^{n} \bar{u}_i^2}{\sum_{i=1}^{n} \bar{u}_i^2} = \lambda_n .$$

Combining these results, we now have $\lambda_n \ge -1$.

The lower bound $R = -1$ occurs only when $h(X_t) = -h(X_{t+1})$, as then

$$R = \frac{E[2 \cdot h(X_t)h(X_{t+1})]}{E[(h(X_t))^2 + (h(X_{t+1}))^2]} = \frac{E[2 \cdot (-h(X_{t+1}))h(X_{t+1})]}{E[(-h(X_{t+1}))^2 + (h(X_{t+1}))^2]} = -1$$

4.3.2 Convergence Analysis with Maximum Autocorrelation

We are primarily interested in the value of λ_2 in order to determine the rate of convergence. Recall that the ratio (from our previous analysis) achieves its second largest value λ_2 when $\bar{v} = \bar{q}_2$. In this case, $\bar{v} \perp \bar{q}_1$, meaning that $\langle \bar{v}, \bar{q}_1 \rangle = 0$. That is,

$$\langle \bar{v}, \bar{q}_1 \rangle = (\sqrt{\pi_1} h_1, ..., \sqrt{\pi_n} h_n)(\pi_1, ..., \pi_n)' = \sum_{i=1}^{n} \pi_i h_i = E(h) = 0$$

***Proposition 4.4**: Given the existence of a stationary distribution for a Markov chain, the second largest eigenvalue can be denoted by the correlation coefficient for $f(X_t)$ and $f(X_{t+1})$, where f is an arbitrary function of the state space. That is,*

$$\lambda_2 = \max_{f \neq const.} \{Corr(f(X_t), f(X_{t+1}))\}.$$

***Proof**:*

Assume a stationary distribution where, $X_t \sim \pi$ and $X_{t+1} \mid X_t \sim K$. Consider the ratio

$$R = \frac{E[h(X_t)h(X_{t+1})]}{E[(h(X_t))^2]},$$

and

$$\lambda_2 = \max_{E[h(X_t)]=0} \left\{ \frac{E[h(X_t)h(X_{t+1})]}{E[(h(X_t))^2]} \right\}$$

Now define a function $\bar{f} \equiv E[f(X)]$, and $h(X) = f(X) - \bar{f}$. Then

$$E[h(X)] = E[f(X) - \bar{f}] = E[f(X) - E[f(X)]]$$
$$= E[f(X)] - E[E[f(X)] = E[f(X)] - E[f(X)] = 0$$

The above now enables us to express λ_2 as

$$\lambda_2 = \max_{f \neq const.} \left\{ \frac{E[(f(X_t) - \bar{f})(f(X_{t+1}) - \bar{f})]}{E[(f(X_t) - \bar{f})^2]} \right\}$$

$$= \max_{f \neq const.} \left\{ \frac{E[(f(X_t) - E[f(X_t)])(f(X_{t+1}) - E[f(X_{t+1})])]}{E[(f(X_t) - E[f(X_t)])^2]} \right\}$$

By definition, if Y_1 and Y_2 are random variables with means μ_1 and μ_2, respectively, the covariance of Y_1 and Y_2 is expressed as $Cov(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$. Hence,

$$\lambda_2 = \max_{f \neq const.} \left\{ \frac{E[(f(X_t) - E[f(X_t)])(f(X_{t+1}) - E[f(X_{t+1})])]}{E[(f(X_t) - E[f(X_t)])^2]} \right\}$$
$$= \max_{f \neq const.} \left\{ \frac{Cov(f(X_t), f(X_{t+1}))}{E[(f(X_t) - E[f(X_t)])^2]} \right\}$$
$$= \max_{f \neq const.} \left\{ \frac{Cov(f(X_t), f(X_{t+1}))}{V(f(X_t))} \right\},$$

since $V(X) = E[(X - E[X])^2]$. The above may also be written as

$$\lambda_2 = \max_{f \neq const.} \left\{ \frac{Cov(f(X_t), f(X_{t+1}))}{\sqrt{V(f(X_t))} \cdot \sqrt{V(f(X_{t+1}))}} \right\},$$

since $V(f(X_t)) = V(f(X_{t+1}))$, which now reduces to

$$\lambda_2 = \max_{f \neq const.} \{Corr(f(X_t), f(X_{t+1}))\},$$

being the correlation between $f(X_t)$ and $f(X_{t+1})$. Δ

According to [12], on the spectral analysis of Markov chains, an irreducible Markov chain has an eigenvalue $\lambda_1 = 1$ with multiplicity 1. Moreover, if $\lambda_n = -1$, the Markov chain is periodic, and aperiodic if $\lambda_n > -1$. Therefore, the convergence rate will be determined strictly from $\lambda^* = \max(\lambda_2, |\lambda_n|)$ [43]. Since the MCMC method of [32] yields an irreducible, aperiodic, and symmetric Markov chain, these conditions concerning λ_2 and λ_n are the focus of the analysis for rate of convergence. Note that a correlation coefficient measures the *linear* relationship between two random variables. Usually, the stronger the correlation, the less randomness there is, and the slower the convergence.

***Proposition 4.5**: For an irreducible Markov chain the second largest eigenvalue satisifies*

$$\lambda_2 = 1 - \min_{f \neq const.} \left\{ \frac{\sum_{ij} (f_i - f_j)^2 \pi_i K_{ij}}{\sum_{ij} (f_i - f_j)^2 \pi_i \pi_j} \right\}.$$

***Proof**:*

Consider the equation $\lambda_2 = \max_{f \neq const.} \{Corr(f(X_t), f(X_{t+1}))\}$.

Using the relationship, $V(X - Y) = V(X) + V(Y) - 2Cov(X, Y)$, this expression may be written as

$$\lambda_2 = \max_{f \neq const.} \left\{ \frac{\frac{1}{2}[V(f(X_t)) + V(f(X_{t+1})) - V(f(X_t) - f(X_{t+1}))]}{V(f(X_t))} \right\}$$
$$= \max_{f \neq const.} \left\{ \frac{\frac{1}{2}[2V(f(X_t)) - V(f(X_t) - f(X_{t+1}))]}{V(f(X_t))} \right\}$$
$$= \max_{f \neq const} \left\{ 1 - \frac{\frac{1}{2}(V(f(X_t) - f(X_{t+1})))}{V(f(X_t))} \right\}$$

It is obvious that the smaller $\frac{V(f(X_t) - f(X_{t+1}))}{V(f(X_t))}$ is, the larger λ_2 becomes. Hence,

$$\lambda_2 = 1 - \frac{1}{2} \cdot \min_{f \neq const.} \left\{ \frac{V(f(X_t) - f(X_{t+1}))}{V(f(X_t))} \right\}$$

Now since,

$$V[f(X_t) - f(X_{t+1})] = E[(f(X_t) - f(X_{t+1}))^2] - (E(f(X_t) - f(X_{t+1})))^2$$
$$= E[(f(X_t) - f(X_{t+1}))^2] - (E(f(X_t)) - E(f(X_{t+1})))^2$$
$$= E[(f(X_t) - f(X_{t+1}))^2]$$

our equation becomes:

$$\lambda_2 = 1 - \frac{1}{2} \cdot \min_{f \neq const.} \left\{ \frac{E[(f(X_t) - f(X_{t+1}))^2]}{V(f(X_t))} \right\}$$
$$= 1 - \min_{f \neq const.} \left\{ \frac{E[(f(X_t) - f(X_{t+1}))^2]}{2V(f(X_t))} \right\}$$

Consider a change of variables such that

$$\lambda_2 = 1 - \min_{f \neq const.} \left\{ \frac{E[(f(X_t) - f(X_{t+1}))^2]}{2V(f(X_t))} \right\}$$

becomes

$$\lambda_2 = 1 - \min_{f \neq const.} \left\{ \frac{E[(f(Y_0) - f(Y_1))^2]}{2V(f(Y_0))} \right\},$$

and assume that we have reached stationarity so that $Y_0, Y_1, ..., Y_\infty \sim \pi$. Since Y_0 and Y_∞ are so "far" apart, they are considered independent (as will be $f(Y_0)$ and $f(Y_\infty)$), and by implication, $V(f(Y_0)) = V(f(Y_\infty))$. Then,

$$\begin{aligned}
&2 \cdot V(f(Y_0)) = V(f(Y_0)) + V(f(Y_\infty)) \\
&= E[(f(Y_0))^2] - (E(f(Y_0)))^2 + E[(f(Y_\infty))^2] - (E(f(Y_\infty)))^2 \\
&= E[(f(Y_0))^2] - 2 \cdot (E(f(Y_0)))^2 + E[(f(Y_\infty))^2] \\
&= E[(f(Y_0))^2] - 2 \cdot (E(f(Y_0))E(f(Y_\infty))) + E[(f(Y_\infty))^2] \\
&= E[(f(Y_0) - f(Y_\infty))^2]
\end{aligned}$$

This now enables us to write

$$\begin{aligned}
\lambda_2 &= 1 - \min_{f \neq const.} \left\{ \frac{E[(f(Y_0) - f(Y_1))^2]}{2V(f(Y_0))} \right\} \\
&= 1 - \min_{f \neq const.} \left\{ \frac{E[(f(Y_0) - f(Y_1))^2]}{E[(f(Y_0) - f(Y_\infty))^2]} \right\} \\
&= 1 - \min_{f \neq const.} \left\{ \frac{\sum_{ij} (f_i - f_j)^2 \pi_i K_{ij}}{\sum_{ij} (f_i - f_j)^2 \pi_i \pi_j} \right\} \qquad \Delta
\end{aligned}$$

This is the convergence rate expressed in terms of the ratio between the first step size and the total length [43]. We already know that $\lambda_2 < 1$, but if we can show that

$$\sum_{ij} (f_i - f_j)^2 \pi_i \pi_j < B \cdot \sum_{ij} (f_i - f_j)^2 \pi_i K_{ij},$$

where B is a constant, then we will be able to write out a general format for the rate of convergence [43].

4.3.3 Dirichlet Form

Assume that $Y_0 \sim \pi$ and $Y_i \mid Y_0 \sim K_{ij}$. Now consider the form of the equation

$$\lambda_2 = 1 - \min_{f \neq const.} \left\{ \frac{E[(f(Y_0) - f(Y_1))^2]}{2V(f(Y_0))} \right\}.$$

A Dirichlet form as defined in [2] is expressed as

$$\varepsilon_\pi(\varphi, \varphi) = \langle (I - P_\pi)\varphi, \varphi \rangle = \mid G \mid \cdot E[(\varphi(Y_0) - \varphi(Y_1))\varphi(Y_0)],$$

where G is a finite group, φ is an arbitrary function, *Y* is chosen from *G* according to uniform measure, π is the probability distribution on *G*, and P_π is the transition kernel of the random walk $\{Y_n\}$. When Y_0 and Y_1 are real-valued random variables with the same distribution,

$$E[(Y_0 - Y_1)Y_0] = \tfrac{1}{2} \cdot E[(Y_0 - Y_1)^2].$$

This yields the alternate Dirichlet form

$$\varepsilon_\pi(\varphi, \varphi) = \tfrac{|G|}{2} \cdot E[(\varphi(Y_0) - \varphi(Y_1))^2.$$

In our case, we have $\frac{1}{2} \cdot E[(f(Y_0) - f(Y_1))^2]$, or $\frac{1}{2} \cdot \sum_{ij} (f_i - f_j)^2 \pi_i K_{ij}$, that can be written in the Dirichlet form $\varepsilon_\pi(f, f)$, so that our equation for λ_2, as seen in [43], becomes

$$\lambda_2 = 1 - \min_{f \neq const.} \left\{ \frac{\varepsilon_\pi(f, f)}{V(f(Y_0))} \right\}.$$

4.3.4 The Poincaré Inequality and Results

Consider the graph with vertex set *X* defined to be the set of legal descent graphs, where {i, j} is an edge if and only if $Q(i, j) \equiv \pi_i K_{ij} = \pi_j K_{ji} > 0$. For each pair of descent graphs $i, j \in X$, define γ_{ij} to be a path from descent graph *i* to descent graph *j*. Paths may have repeated vertices but a given edge may occur at most once in a given path. Finally, denote the collection of paths (one for each ordered pair *i, j*) by Γ. Irreducibility guarantees that such paths will exist, but the quality of the estimate depends on the selection of Γ. Path length is then defined, for each $\gamma_{ij} \in \Gamma$, to be $|\gamma_{ij}| = \sum_{e \in \gamma_{ij}} (Q(e))^{-1}$, where the sum is over the edges in the path and $Q(e) = Q(z, w)$ if $e = \{z, w\}$ [2].

Proposition (Poincaré Inequality) 4.6*: For an irreducible Markov chain, the second largest eigenvalue satisifies* $\lambda_2 \le 1 - \frac{1}{B}$, *where* $B = \max_e \sum_{e \in \gamma_{ij}} |\gamma_{ij}| \pi_i \pi_j$.

Proof:

Consider
$$V(f) = \tfrac{1}{2} \sum_{i,j \in X} (f_i - f_j)^2 \pi_i \pi_j$$
$$= \tfrac{1}{2} \sum_{ij} \left(\sum_{e \in \gamma_{ij}} \left(\frac{Q(e)}{Q(e)} \right)^{1/2} f(e) \right)^2 \pi_i \pi_j$$
$$\le \tfrac{1}{2} \sum_{ij} |\gamma_{ij}| \pi_i \pi_j \sum_{e \in \gamma_{ij}} Q(e) f(e)^2$$
$$= \tfrac{1}{2} \sum_{e} Q(e) f(e)^2 \sum_{e \in \gamma_{ij}} |\gamma_{ij}| \pi_i \pi_j$$

Here, $f(e) = f(e^+) - f(e^-)$, where e is the oriented edge in a path from e^- to e^+. The inequality is Cauchy-Schwarz, and the final sum is over all oriented edges in the graph. Then it is now obvious that

$$\tfrac{1}{2} \sum_{e} Q(e) f(e)^2 \sum_{e \in \gamma_{ij}} |\gamma_{ij}| \pi_i \pi_j$$
$$\le \tfrac{1}{2} \sum_{e} Q(e) f(e)^2 \cdot \max_e \sum_{e \in \gamma_{ij}} |\gamma_{ij}| \pi_i \pi_j$$
$$= \varepsilon_\pi(f, f) \cdot B$$

giving the desired result that $V(f) \le \varepsilon_\pi(f, f) \cdot B$.

Now using,

$$\lambda_2 = 1 - \min_{f \neq const.} \left\{ \frac{\sum_{ij} (f_i - f_j)^2 \pi_i K_{ij}}{\sum_{ij} (f_i - f_j)^2 \pi_i \pi_j} \right\}$$

$$\lambda_2 = 1 - \min_{f \neq const..} \left\{ \frac{\varepsilon_\pi(f, f)}{V(f(Y_0))} \right\}$$

we have

$$\sum_{ij} (f_i - f_j)^2 \pi_i \pi_j = V(f(Y_0)) = V(f),$$

and

$$\sum_{ij} (f_i - f_j)^2 \pi_i K_{ij} = \varepsilon_\pi(f, f)$$

Since $V(f) \le B \cdot \varepsilon_\pi(f, f)$, we have the inequality

$$\sum_{ij} (f_i - f_j)^2 \pi_i \pi_j \le B \cdot \sum_{ij} (f_i - f_j)^2 \pi_i K_{ij}$$

Consequently, we can now show that

$$1 - \frac{\sum_{ij} (f_i - f_j)^2 \pi_i K_{ij}}{\sum_{ij} (f_i - f_j)^2 \pi_i \pi_j} \le 1 - \frac{1}{B},$$

which is equivalent to

$$1 - \min_{f \neq const.} \frac{\sum_{ij} (f_i - f_j)^2 \pi_i K_{ij}}{\sum_{ij} (f_i - f_j)^2 \pi_i \pi_j} \le 1 - \frac{1}{B}$$

and $\lambda_2 \le 1 - \frac{1}{B}$, where $B = \max_e \sum_{e \in \gamma_{ij}} | \gamma_{ij} | \pi_i \pi_j$ and $\frac{1}{2} < B < \infty$ Δ

4.4 Bound for Convergence Rate

The proposition above is a discrete analog of the classical method of Poincaré for estimating the spectral gap of the Laplacian on a domain, [2]. According to [29], this proposition is revised as follows

For an irreducible Markov chain, the second largest eigenvalue satisifies $\lambda_2 \le 1 - \frac{1}{B}$, *where* $B = \max_e Q(e)^{-1} \sum_{e \in \gamma_{ij}} |\gamma_{ij}| \pi_i \pi_j$.

The latter proposition uses the same notation as does the former proposition, with the only exception that $|\gamma_{ij}|$ denotes the number of edges in the path γ_{ij}, [2]. This bound is sometimes easier to use, and can be more effective than the previous bound. For random walks on graphs, as we have in our case, the bounds presented previously coincide, and it is the revised form that will be implemented in this work.

5. Implementation of *SimWalk2* for the Disease Gene Episodic Ataxia

SimWalk2 is a statistical genetics computer application for haplotype, parametric linkage, non-parametric linkage, identity by descent, and mistyping analyses on any size of pedigree. *SimWalk2* uses the MCMC method as found in [32]. See also [34, 35]. The software for parametric linkage analysis requires five input files, namely, the *map*, *locus*, *pedigree*, and *penetrance* data files, and the *input control* file.

5.1 Methodology

For the purposes of this work, the map data file, consisting of the marker loci and the recombination fractions between them, and the input control file, which contains the instructions for the software, will remain the same in each of the trials. Alterations will be made to the locus data file, concerning disease allele frequency, the penetrance data file, and the pedigree data file. Input data files B1, B2, and B3 in Appendix B show the data files used in this analysis. The analysis considered the following four different variations of the pedigree.

a. Original Pedigree (No Missing Individuals)
b. Pedigree A (Four Missing Individuals – Two Founders (F) and Two Nonfounders (NF))
   Missing Individuals: 2001(F), 103(F), 1000(NF), 115(NF)
c. Pedigree B (Ten Missing Individuals – Five Founders and Five Nonfounders)
   Missing Individuals: 2001(F), 103(F), 1002(F), 199(F), 1011(F), 1000(NF), 115(NF), 9003(NF), 1010(NF), 1(NF)
d. Pedigree C (Thirteen Missing Individuals – Five Founders and Eight Nonfounders)
   Missing Individuals: 2001(F), 103(F), 1002(F), 199(F), 1011(F), 1000(NF), 115(NF), 9003(NF), 1010(NF), 1(NF), 113 (NF), 9097(NF), 9006(NF)

The analysis also considered the following variations of parameters: Disease Allele Frequency and Penetrances:

Variation Name	Allele 1	Allele 2
Locus 1 (Orig.)	0.99990	0.00010
Locus 2	0.90000	0.10000
Locus 3	0.85000	0.15000
Locus 4	0.70000	0.30000
Locus 5	0.60000	0.40000

Table 5.1 *Variation of Parameters for Allele Frequency*

	Affection	Genotypes		
Variation Name		1/1	1/2	2/2
Pen 1 (Orig.)	101	0.99900	0.01000	0.01000
	201	0.00100	0.99000	0.99000
Pen 2	101	0.90000	0.10000	0.10000
	201	0.10000	0.90000	0.90000
Pen 3	101	0.80000	0.10000	0.10000
	201	0.20000	0.90000	0.90000
Pen 4	101	0.90000	0.15000	0.01000
	201	0.10000	0.85000	0.99000
Pen 5	101	0.80000	0.20000	0.20000
	201	0.20000	0.80000	0.80000

Table 5.2 *Variation of Parameters for Penetrance*

Affection is used in this file to denote normal (101) and affected alleles (201) in an individual. Similarly, the genotypic alleles 1 and 2 correspond to the normal/wild type and affected alleles, respectively.

For each pedigree above, the following trials were completed using *SimWalk2*,

Case	Parameters	Case	Parameters	Case	Parameters
1	Locus 1, Pen 1	6	Locus 1, Pen 2	11	Locus 3, Pen 4
2	Locus 2, Pen 1	7	Locus 1, Pen 3	12	Locus 3, Pen 5
3	Locus 3, Pen 1	8	Locus 1, Pen 4	13	Locus 4, Pen 2
4	Locus 4, Pen 1	9	Locus 1, Pen 5	14	Locus 4, Pen 4
5	Locus 5, Pen 1	10	Locus 3, Pen 2	15	Locus 4, Pen 5

Table 5.3 *Trials for SimWalk2*

Cases 1 through 5 investigate the effects of varying only the allele frequencies, while assuming accurate penetrance values. Cases 1 and 6 through 9 vary the penetrance values, while assuming accurate allelic frequencies. Finally, cases 10 through 15 investigate the most severe examples of variations in both allelic frequencies and penetrance values.

The computational implementation provides results in the form of a location corresponding to the largest location score, and the bound for the rate of convergence. Other results are available corresponding to information such as the location scores for all other points on the marker map and log likelihoods for the trait and markers only, but are not used in this study. See Tables C2 and C19 in Appendix C for the results.

5.2 Statistical Analysis of *SimWalk2* Results

In this work, we are interested in two different aspects of the MCMC implementation. First, we want to investigate the accuracy of the MCMC method when faced with missing data in the model – particularly, individual missing data from the pedigree, disease allele frequency, and penetrance. Since we know the exact results for Episodic Ataxia from the Elston-Stewart implementation, it is possible to compare the results from the MCMC method. For this study, we will consider five different variables: pedigree only, pedigree and disease allele frequency, pedigree and penetrance, and finally, pedigree and two combinations of disease allele frequency and penetrance. For the case in which only the data missing from each pedigree is considered, a single factor analysis of variance was performed to determine whether the null hypothesis that all of the pedigrees have the same mean *location position* on the chromosome is rejected or not. For the same data

parameters, a second single factor analysis of variance was performed to determine whether the null hypothesis that all of the pedigrees have the same mean *location score* is rejected or not. Since this analysis includes the exact results, missing data pedigrees are compared to the true result in an effort to determine the accuracy of the MCMC method for pedigrees missing data.

The second case considered missing data in the pedigrees as well as variations in the disease allele frequency, deviating from the known disease allele frequency. For this case, a two-factor analysis of variance was performed to determine whether the null hypothesis that all of the pedigrees have the same mean location position is rejected or not, while considering the effects of the disease allele frequency on the method. Similarly, this analysis was performed for the location score.

The third case considered missing data in the pedigrees as well as variations in the penetrance values. Case four and case five considered missing data in the pedigrees as well as variations in both the allele frequency and the penetrance values. As with the second case, two-factor analysis of variances were performed for both location position and location scores.

5.2.1 Accuracy of MCMC Method for Missing Data in Pedigree

The pedigrees range from no missing individual data to a great amount of missing data (original pedigree → pedigree A → pedigree B → pedigree C). By inspection of the raw output, it is clear that when using the original pedigree, the method most consistently chose the true location position for the EA gene. With the exception of the last case, case 15, all of the positions fall within one centiMorgan of the true position. See Table C2 in the appendix. Further inspection reveals that eight of the trials resulted in location scores greater than the cut-off value of three, and four others suggest strong evidence that the gene is at the corresponding location position. Inspection of pedigrees A and B both reveal that the most likely location position is within a one cM region surrounding the position 6.6609 (between S372 and pY2/1). Analysis of pedigree A reveals that five of the trials resulted in location scores greater than three, and four others suggest strong evidence that the gene is at the corresponding location position.

The results of pedigree A show only slight inaccuracies as compared to the results of the original pedigree. Pedigree B, which has a greater level of missing data, has zero location scores above the cut-off value of three and only one location score suggests strong evidence that the gene is at the corresponding location position. Thus, while the MCMC method seems to find the true region of localization of the EA gene, there is no supporting evidence (i.e. location score to support linkage) that suggests that the region is correct. Thus, this analysis returns inconclusive results. Finally, inspection of pedigree C reveals that the location position of the EA gene is within a 1 cM region between the pY2/1 and pY21/1 markers. Thus, in the most severe case of missing data in the pedigree, the method finds the wrong location for the EA gene. Nonetheless, analysis of pedigree C results in a location position within 3cMs (or 3% recombination) of the true position. Inspection of the marker map places that position in between the subsequent markers. Thus, for this pedigree, the result is inaccurate by less than 3%. Further, the location scores for pedigree exceed the cut-off level of three only once, but there are three other location scores that provide strong evidence that the gene is at the corresponding location position.

The results above suggest that as missing data in the pedigree increases the level of accuracy decreases. Nonetheless, it is not a dramatic decrease in accuracy – the location position remains accurate and consistent throughout the first three pedigrees, while only the significance of the support for this position decreases. At extreme levels of missing data, the location position becomes inaccurate, and the significance of support for this incorrect result increases. While the method is unable to pinpoint the exact location when there is a great amount of missing data, it is still able to get within range of the true value and with a significant level of support.

5.2.2 Accuracy of MCMC Method for Variations of Parameters

Inspection of the location scores by variation in disease allele frequency, in Table C7 in the appendix, suggests that there is little influence from variation in disease allele frequency. All of the location scores where disease allele frequency was varied in both the original pedigree and pedigree A remained above the level of three. Also, many of

the location scores in pedigree C suggest strong evidence of the location corresponding to these location scores.

With variation in penetrance values only, there are only five out of twenty location scores above the cutoff value of three, as seen in Table C8 of the appendix. Many of the remaining location scores do not suggest evidence of localization for the EA gene. Moreover, when there are combinations of disease allele frequency and penetrance value variations, location scores are greatly reduced. The final most severe combination of variations in allele frequency and penetrance values resulted in zero location scores exceeding the cut-off level of three. In fact, many of the location scores did not even exceed two.

In conclusion, this inspection of the location scores for varying parameters provides strong evidence to suggest that the variation in disease allele frequency had very little influence on the accuracy of the MCMC method, whereas, the variation in penetrance had significant influence on the accuracy of the method. Most importantly, however, is that a combination of variation in disease allele frequency and penetrance can cause location scores to decrease dramatically to the point in which no conclusion may be reached about the location of the gene. All results for location position and location score classified by the variation of parameters are shown in Appendix C Tables C3 through C10.

5.2.3 ANOVA Results for MCMC Accuracy

The analysis of variance results for both location scores and location positions provide further evidence to suggest that missing data does have a significant effect on the accuracy of the MCMC method. Note that all ANOVA were completed with $\alpha = 0.05$; that is, we rejected the null hypothesis when $p < 0.05$. In the case of location scores, all of the cases discussed previously resulted in the rejection of the null hypothesis. This means that the location scores found through the implementation of *SimWalk2* for the different pedigrees are not equivalent, regardless of the variation in parameters or lack thereof. The conclusion, then, is that missing data in the pedigree, variations in allele frequency or penetrance values, or any such combination, leads to dramatic deviations in the location score. The table found below gives the data summary and ANOVA results

for the single-factor analysis of location score for data missing only from the pedigree. Recall that the EXACT group refers to analysis of this pedigree using the Elston-Stewart algorithm and not by *SimWalk2*. All other two – factor ANOVA results for varying parameters is found in Appendix C, Tables C11 through C14.

Groups	*Count*	*Sum*	*Average*	*Variance*
EXACT	15	53.4	3.56	0
ORIGINAL	15	45.656	3.0437333	0.2431305
PEDIGREE A	15	39.544	2.6362667	0.3504515
PEDIGREE B	15	24.778	1.6518667	0.5232488
PEDIGREE C	15	29.226	1.9484	0.8389665

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Pedigrees	36.5737646	4	9.1434411	23.375226	2.678E-12	2.5026594
Within Pedigrees	27.3811632	70	0.3911595			
Total	63.954928	74				

Table 5.4 *Data Summary and ANOVA Results for Location Score*

In the case of location positioning, there are three cases in which we do not reject the null hypothesis that the mean location positions are equal. For the variables disease allele frequency only and both combinations of variations in disease allele frequency and penetrance values, the results suggest that we should not reject the null hypothesis. This indicates that the MCMC method's ability to determine location position is not nearly as dependent on the missing parameters. The missing data in the pedigree still had a dramatic effect on the accuracy of the location positioning. Therefore, the conclusion is that missing data in the pedigree or variation in the penetrance values alone, leads to dramatic deviations in the location positioning, while variation in the disease allele frequency or combinations of variations in the disease allele frequency and penetrance values do not lead to *dramatic* deviations. The table found on the overleaf gives the data summary and ANOVA results for the single-factor analysis of location position for data missing only from the pedigree. All other two – factor ANOVA results for varying parameters is found in Appendix C, Tables C15 through C18.

Groups	*Count*	*Sum*	*Average*	*Variance*
EXACT	15	99.9135	6.6609	3.25E-14
ORIGINAL	15	105.8738	7.0582533	0.4140553
A	15	104.2451	6.9496733	0.0884348
B	15	103.9357	6.9290467	0.0802295
C	15	128.9748	8.59832	0.2155355

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Pedigrees	35.916008	4	8.979002	56.24143	3.79E-21	2.5026594
Within Pedigrees	11.175572	70	0.159651			
Total	47.09158	74				

Table 5.5 *Data Summary and ANOVA Results for Locational Position*

The analysis completed by *SimWalk2* provides evidence of several important aspects of parametric linkage analysis with missing data in the model. First, the degree of inaccuracy seems to be directly correlated to the level of missing pedigree data. As the individual pedigree data decreases, the accuracy of both the location score and the location positioning decreases. Also, variations of the variables in the genetic model lead to inaccuracies within the results. Finally, the ANOVA results yield evidence that missing data causes dramatic decreases in accuracy with respect to the localization of the disease gene (EA) and its support for linkage.

5.3 Analysis of Convergence Results

In Section 4, a theoretical analysis of convergence for a Markov chain was presented, utilizing a bound on the second largest eigenvalue of the transition matrix. We restate here without proof, the proposition as in Section 4.4.

Proposition*: For an irreducible Markov chain, the second largest eigenvalue satisfies* $\lambda_2 \le 1 - \frac{1}{B}$, *where*

$$B = \max_e Q(e)^{-1} \sum_{e \in \gamma_{ij}} |\gamma_{ij}| \pi_i \pi_j .$$

Given the convergence parameter *B*, the bound on the second largest eigenvalue can then be computed, and a comparison between the rates of convergence can be made.

As λ_2 approaches a unit value, the rate of convergence becomes increasingly slower. Thus, larger values of *B* indicate a smaller ratio of one-step length to total length (of the chain), and a decrease in the rate of convergence.

5.3.1 Convergence Results

Several different aspects of the convergence bound were considered in this work. First, the range and mean for the bound on convergence for each of the pedigrees was looked at, and their values were as follows:

Pedigree	Range	Mean
Original Pedigree	0.2427 – 0.4704	0.3355
Pedigree A	0.2612 – 0.4525	0.3644
Pedigree B	0.3998 – 0.5026	0.4658
Pedigree C	0.5604 – 0.6155	0.5957

Table 5.6 ***Convergence Bounds Results***

Inspection of these values indicates that missing data in the pedigree does cause a decrease in the rate of convergence. The original pedigree had the smallest convergence bound, indicating that no missing data allows the Markov chain to reach stationarity more efficiently than when missing data is introduced into the pedigree. Pedigree A, which contained the least amount of missing data, had a very similar convergence result, having almost an identical range of values. The mean bound for convergence was only slightly greater for pedigree A. This suggests that a small amount of missing data does not have a significant impact on the rate of convergence of the Markov chain.

As expected, however, pedigrees B and C both had elevated bounds on their rates of convergence. The drastic increase in the convergence rate for pedigree B suggests that the introduction of variations in parameters induced a decrease in the convergence rate of the Markov chain. Pedigree C, which contained the greatest amount of missing information, had the most severe decrease in the convergence rate. The mean bound on convergence for pedigree C was nearly twice that of the original pedigree, and its range

fell entirely outside of the other pedigrees' ranges for convergence bounds. This analysis provides evidence that as the missing data from the pedigree increases, the rate of convergence dramatically decreases.

5.3.2 Convergence Results by Parameters

Next, we looked at the impact of the variation in parameters (i.e. disease allele frequency and penetrance values) on the rate of convergence for these pedigrees.

Original Pedigree			A			B			C		
	Avg. CP	Avg. CR		Avg. CP	Avg. CR		Avg. CP	Avg. CR		Avg. CP	Avg. CR
Pen	1.35	0.2589	Pen	1.4918	0.3275	Pen	1.7818	0.4371	Pen	2.3925	0.5859
Locus	1.5287	0.3367	Locus	1.5585	0.3562	Locus	1.8846	0.4688	Locus	2.507	0.601
Both 1	1.5585	0.3565	Both 1	1.6603	0.3963	Both 1	1.9867	0.4966	Both 1	2.3975	0.5822
Both 2	1.7047	0.4128	Both 2	1.6966	0.4076	Both 2	1.9034	0.474	Both 2	2.5897	0.6139

Table 5.7 *Convergence Bounds with Variation of Parameters*

The results shown give the average convergence parameter (Avg. CP) and bound on convergence (Avg. CR) for each of the pedigrees broken down by a particular variation in parameters. As can be seen from the table above, the variation in disease allele frequency only had the least effect on the convergence bounds, and thus, the least effect on the convergence rate of the Markov chain. Variation in the penetrance values alone and the first combination of variation in disease allele frequency and penetrance values had a very similar effect on the convergence rates. This suggests that the variation in disease allele frequency in the first combination had little to no effect on the convergence rate. As such, the variation in penetrance values combined with missing data in the pedigree has a significant effect on the rate of convergence of the Markov chain. Finally, the last combination of variations of allele frequency and penetrance values, which was the most dramatic deviation from the original parameters, caused the greatest decrease in the convergence rate of the MCMC process. This analysis provides evidence to suggest that as the variations of parameters increase, the rate of convergence decreases.

6. Conclusion

This work investigates the complexities of missing data in pedigree analysis using the Markov Chain Monte Carlo (MCMC) method as compared to the exact results. We developed the biological and mathematical problems to be analyzed, and in particular, described the MCMC method for parametric linkage analysis as created by [32]. The theoretical results of the MCMC method and convergence of a Markov chain using a maximum autocorrelation procedure and eigenvalue analysis were described in order that they could be implemented in a latter section of this work. Finally, a computational application *SimWalk2* was used to examine the concrete properties of convergence and accuracy of the MCMC method as seen for the disease gene Episodic Ataxia (EA).

In the case in which everything about the disease gene within a pedigree is known, the MCMC method can be expected to be a statistically significant estimate of the exact likelihood for linkage between the disease locus and marker loci. Our implementation of *SimWalk2* involved three different missing data parameters, namely, disease allele frequency, penetrance, and pedigree members. We have looked at how misspecification of disease allele frequency, misspecification of disease penetrance, and missing members of the observed pedigree, influenced both the accuracy of the MCMC method and the convergence of the Markov chain being implemented. The output variables of interest were the location position on the chromosome of the EA gene, the corresponding location score, and the bound on convergence of the MCMC method.

As expected, the analysis suggests that the more severe the missing data, the greater the inaccuracy of the method, and the slower the convergence of the Markov chain. Considering only missing data from the pedigrees (i.e. assuming correct disease allele frequency and penetrance values), larger numbers of members missing from the pedigree caused the harshest cases of inaccuracy and decreased convergence rates. In fact, this missing data variable was the most influential in determining the correct location position and location score. Nonetheless, the MCMC method performed better than would be expected of an exact method under similar missing data conditions. The most dramatic inaccuracy was attained at a rate of only 3% deviation from the true value. The analysis of the misspecified parameters, that is, disease allele frequency and penetrance, indicated

that misspecification of disease allele frequency had very little influence on the accuracy of the MCMC method, while penetrance had a significant influence, particularly in combination with data missing from the pedigree or severe misspecification of disease allele frequency.

Convergence analysis gave a similar result. Missing data from the pedigree had a great impact on the convergence rate of the Markov chain being implemented in this method; the more severe the missing data, the slower the convergence. When the misspecification of disease allele frequency and penetrance values are considered, misspecification of penetrance appeared to have the least impact on the rate of convergence. Combinations of misspecification of disease allele frequency and penetrance had the greatest impact on the rate of convergence. Thus, it is expected that as these parameters deviate further from their true values, the rate of convergence decreases dramatically as well.

Further research into this area of study should include the implementation of this method on more genetic parameters and differing pedigree variations. Along these lines, a comparison of an exact method to the MCMC method, using the corresponding genetic software, at each point on the chromosome would be a useful tool in determining the accuracy of each method when faced with missing data in the model. Also, it would be of interest to look into possible ways to combat the effects of missing data on the MCMC method.

References

[1] M.K. Cowles and B. Carlin, *Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review,* Journal of the American Statistical Association 91, 883 – 904, (1996).

[2] P. Diaconis and D. Stroock, *Geometric Bounds for Eigenvalues of Markov Chains,* The Annals of Applied Probability 1, 36-61, (1991).

[3] R.C. Elston and J. Stewart, *A General Model for the Genetic Analysis of Pedigree Data,* Human Heredity 21, 523 – 542, (1971).

[4] W. Feller, *An Introduction to Probability Theory and Its Applications,* John Wiley, New York, (1957).

[5] M. Fishelson and D. Geiger, *Exact Genetic Linkage Computations for General Pedigrees,* Bioinformatics 18, 189 – 198, (2002).

[6] J. Fulman and E.L. Wilmer, *Comparing Eigenvalue Bounds for Markov Chains: When Does Poincare Beat Cheeger?,* The Annals of Applied Probability 9, 1 – 13, (1999).

[7] R.A. Gatti, E. Lange, E. Sobel, and K. Lange, *Localization of the Ataxia – Telangiectasia Gene(s) to a 3cm Interval on Chromosome 11q23 by Linkage Analysis,* Cytogenetics and Cell Genetics 58, 1959 – 1960, (1991).

[8] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice,* Chapman and Hall, Boca Raton, Florida, (1998).

[9] S.W. Guo and E.A. Thompson, *A Monte Carlo Method for Combined Segregation and Linkage Analysis,* American Journal of Human Genetics 51, 1111 – 1126, (1992).

[10] J.L. Haines, and M.A. Pericak-Vance, *Approaches to Gene Mapping in Complex Human Diseases,* Wiley-Liss, Inc., New York, (1998).

[11] J.B.S. Haldane, *The Combination of Linkage Values and the Calculation of Diastances between the Loci of Linked Factors,* Journal of Genetics 8, 299 – 309, (1919).

[12] S. Karlin, *A Second Course in Stochastic Processes,* Academic Press, London, (1991).

[13] L. Kruglyak, M.J. Daly, M.P. Reeve – Daly, and E. Lander, *Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach,* American Journal of Human Genetics 58, 1347 – 1363, (1996).

[14] E. Lange, et al, *How Many Ataxia – Telangiectasia Genes?,* Ataxia – Telangiectasia (NATO – ASI Series) H77, 37 – 54, (1993).

[15] K. Lange, *Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.,* Springer, New York, (2002).

[16] K. Lange and R.C. Elston, *Extensions to Pedigree Analysis. I. Likelihood Calculations for Simple and Complex Pedigrees,* Human Heredity 25, 95 – 105, (1975).

[17] K. Lange and S. Matthysse, *Simulations of Pedigree Genotypes by Random Walks,* American Journal of Human Genetics 45, 959 – 970, (1989).

[18] K. Lange and E. Sobel, *A Random Walk Method for Computing Genetic Location Scores,* American Journal of Human Genetics 49, 1320-1334, (1991).

[19] G.M. Lathrop, J.M. Lalouel, C. Julier, and J. Ott, *Multilocus Linkage Analysis in Humans: Detection of Linkage and Estimation of Recombination,* American Journal of Human Genetics 37, 482 – 98, (1985).

[20] S. Lin, *A Scheme for Constructing an Irreducible Markov Chain for Pedigree Data,* Biometrics 51, 318 – 322, (1995).

[21] M. Litt, et. al., *A Gene for Episodic Ataxia/ Myokymia Maps to Chromosome 12p13,* American Journal of Human Genetics 55, 702 – 709, (1994).

[22] G. Mendel, *Experiments in Plant Hybridisation,* [Mendel's original paper in English translation, with a commentary by R.A. Fisher, published by Oliver and Boyd, Edinburgh, 1965.]

[23] N.E. Morton, *Sequential Tests for the Detection of Linkage*, American Journal of Human Genetics 7, 277 – 318, (1955).

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

[25] J. Ott, *Estimation of the Recombination Fraction in Human Pedigree – Efficient Computation of the Likelihood for Human Linkage Analysis*, American Journal of Human Genetics 26, 588 – 597, (1974).

[26] C.P. Robert, *Convergence Control Methods for Markov Chain Monte Carlo Algorithms,* Statistical Science 10, 231 – 253, (1995).

[27] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods,* Springer – Verlag, New York, (1999).

[28] S. Ross, *Introduction to Probability Models, 8th ed.,* Elsevier, San Diego, (2003).

[29] A. Sinclair, *Improved Bounds for Mixing Rates of Markov Chains on Combinatorial Structures,* Technical Report, Department of Computer Science, University of Edinburgh, (1990).

[30] E. Sobel, et al, *Ataxia –Telangiectasia: Linkage Evidence for Genetic Heterogeneity,* American Journal of Human Genetics 50, 1343 – 1348, (1992).

[31] E. Sobel, et al, *Localization of an Ataxia – Telangiectasia Gene to a 500 kb Interval on Chromosome 11q23.1: Linkage Analysis of 176 Families by an International Consortium,* American Journal of Human Genetics 57, 112 – 119, (1995).

[32] E. Sobel and K. Lange, *Descent Graphs in Pedigree Analysis: Applications to Haplotyping, Location Scores, and Marker Sharing Statistics*, American Journal of Human Genetics 58, 1323-1337, (1996).

[33] E. Sobel and K. Lange, *Metropolis Sampling in Pedigree Analysis*, Statistical Methods in Medical Research 2, 263 – 282, (1993).

[34] E. Sobel, H. Sengul, and D.E. Weeks, *Multipoint Estimation of Identity-by-Descent Probabilities at Arbitrary Positions among Marker Loci on General Pedigrees,* Human Heredity 52, 121-131, (2001).

[35] E. Sobel, J.C. Papp, and K. Lange, *Detection and Integration of Genotyping Errors in Statistical Genetics,* American Journal of Human Genetics 70, 496-508, (2002).

[36] D.P. Snustad, and M.J. Simmons, *Principles of Genetics, 2nd ed.,* John Wiley & Sons, Inc., New York, (1997).

[37] E.A. Thompson, *Likelihood and Linkage: From Fisher to the Future,* The Annals of Statistics 24, 449 – 465, (1996).

[38] E.A. Thompson, *Monte Carlo Analysis on a Large Pedigree*, Genetic Epidemiology 10, 677 – 682, (1993).

[39] E.A. Thompson, *Monte Carlo Likelihood in Genetic Mapping*, Statistical Science 9, 355 – 366, (1994).

[40] E.A. Thompson and S.W. Guo, *Evaluation of Likelihood Ratios for Complex Genetic Models,* IMA Journal of Mathematics Applied in Medicine and Biology 8, 149 – 169, (1991).

[41] E.A. Thompson and S.W. Guo, *Monte Carlo Estimation of Mixed Models for Large Complex Pedigree,* Biometrics 50, 417 – 432, (1994).

[42] B. Walsh, *Markov Chain Monte Carlo and Gibbs Sampling,* Lecture notes for EEB 581, Arizona University, April 2004.

[43] Y. Wang and Y. Wu, *Metropolis Algorithm and the Rate of Convergence of Markov Chains,* Lecture notes for Stochastic Processes course, Dept. of Mathematics, University of Pennsylvania, February 2003.

APPENDIX A

Fig. A1
Pedigree for Episodic Ataxia Analysis

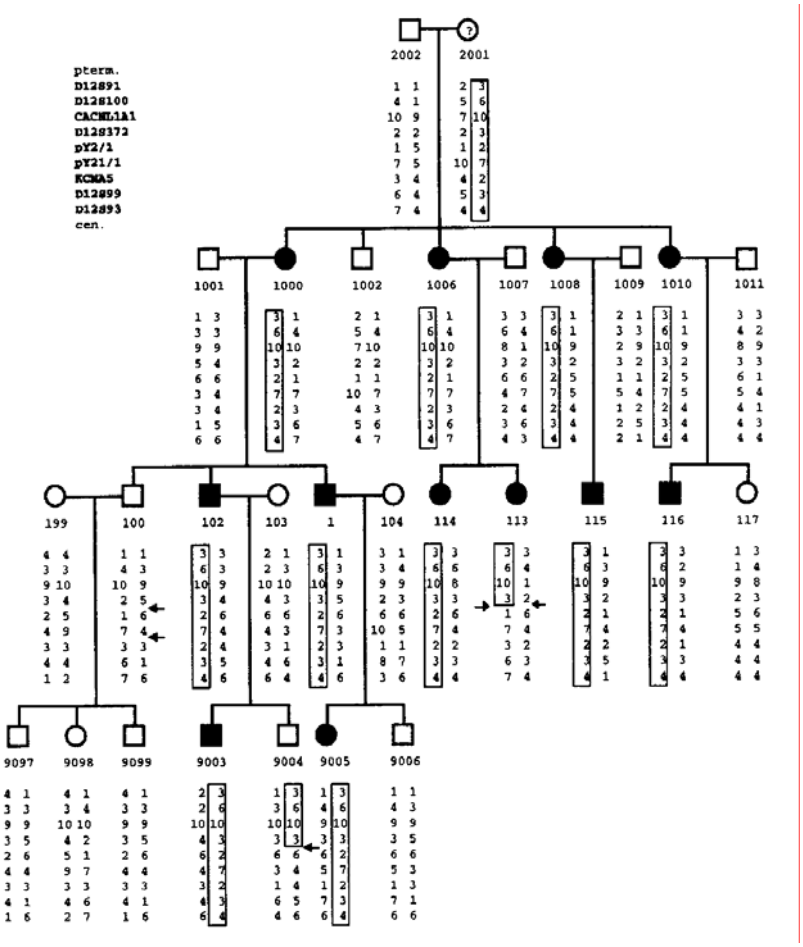
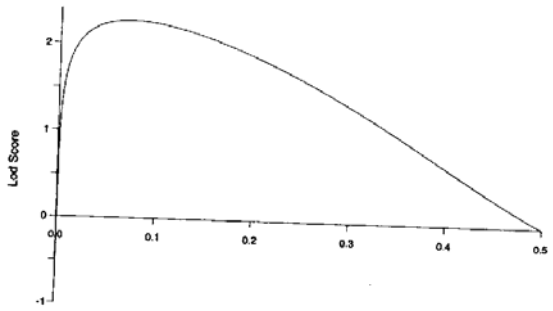


Fig. A2
Plot of LOD Score for EA vs. Marker S372



APPENDIX B

Input Data B1
Marker Map Data File for EA

```
EA
         0.5000
S91
        0.0100
S100
        0.0100
CACNL1A1
        0.0300
S372
        0.0300
pY2/1
        0.0100
pY21/1
        0.0100
KCNA5
        0.0100
S99
        0.0100
S93
```

Input Data B2
Input Control Data File for EA

```
01
2

02
22

03
Episodic Ataxia (EA) pedigree: LOD score

09
MAP.DAT

10
LOCUS.DAT

11
PEDIGREE.DAT

12
F
M

13
Y

18
1

19
PEN.DAT
```

Input Data B3

Original Pedigree (No Missing Data)

```
(I5,2X,A8)
(3A5,1X,2A1,(T19,3(A8),:))
   29       20
    1 1001 1000 M   2       3/  1   3/  6
                    9/ 10   5/  3   6/  2
                    3/  7   3/  2   1/  3
                    6/  4   201
 1001           M   1       3/  1   3/  3
                    9/  9   5/  4   6/  6
                    4/  3   4/  3   5/  1
                    6/  6   101
 1000 2002 2001 F   2       1/  3   4/  6
                   10/ 10   2/  3   1/  2
                    7/  7   3/  2   6/  3
                    4/  7   201
  100 1001 1000 M   1       1/  1   3/  4
                    9/ 10   5/  2   6/  1
                    4/  7   3/  3   1/  6
                    6/  7   101
 2002           M   1       1/  1   1/  4
                    9/ 10   2/  2   1/  5
                    5/  7   3/  4   4/  6
                    4/  7   101
 2001           F           3/  2   6/  5
                   10/ 7    3/  2   2/  1
                    7/ 10   2/  4   3/  5
                    4/  4
 1002 2002 2001 M   1       1/  2   4/  5
                   10/  7   2/  2   1/  1
                    7/ 10   3/  4   6/  5
                    4/  7   101
 1006 2002 2001 F   2       1/  3   4/  6
                   10/ 10   2/  3   1/  2
                    7/  7   3/  2   6/  3
                    4/  7   201
 1007           M   1       3/  3   4/  6
                    1/  8   2/  3   6/  6
                    4/  7   2/  4   3/  6
                    3/  4   101
 1008 2002 2001 F   2       1/  3   1/  6
                    9/ 10   2/  3   5/  2
                    5/  7   4/  2   4/  3
                    4/  4   201
 1009           M   1       1/  2   3/  3
                    2/  9   2/  3   1/  1
                    4/  5   1/  2   2/  5
                    1/  2   101
  199           F   1       4/  4   3/  3
                    9/ 10   3/  4   2/  5
                    4/  9   3/  3   4/  4
                    1/  2   101
 1010 2002 2001 F   2       1/  3   1/  6
                    9/ 10   2/  3   5/  2
                    5/  7   4/  2   4/  3
                    4/  4   201
 1011           M   1       3/  3   2/  4
                    9/  8   3/  3   1/  6
                    4/  5   1/  4   3/  4
                    4/  4   101
  102 1001 1000 M   2       3/  3   3/  6
                    9/ 10   4/  3   6/  2
                    4/  7   4/  2   5/  3
                    6/  4   201
  103           F   1       2/  1   3/  2
                   10/ 10   4/  3   6/  6
                    4/  3   3/  1   6/  4
                    4/  6   101
  104           F   1       3/  1   4/  3
                    9/  9   3/  2   6/  6
                   10/  5   1/  1   8/  7
                    6/  3   101
 113 1007 1006 F   2       3/  3   6/  4
                    1/ 10   3/  2   6/  1
                    4/  7   2/  3   6/  3
                    4/  7   201
 114 1007 1006 F   2       3/  3   6/  6
                    8/ 10   3/  3   6/  2
                    4/  7   2/  2   3/  3
                    4/  4   201
 115 1009 1008 M   2       1/  3   3/  6
                    9/ 10   3/  2   1/  2
                    4/  7   2/  2   5/  3
                    1/  4   201
 116 1011 1010 M   2       3/  3   2/  6
                    9/ 10   3/  3   1/  2
                    4/  7   1/  2   3/  3
                    4/  4   201
 117 1011 1010 F   1       1/  3   4/  1
                    8/  9   2/  3   6/  5
                    5/  5   4/  4   4/  4
                    4/  4   101
9099  100  199 M   1       1/  4   3/  3
                    9/  9   5/  3   6/  2
                    4/  4   3/  3   1/  4
                    6/  1   101
9098  100  199 F           1/  4   4/  3
                   10/ 10   2/  4   1/  5
                    7/  9   3/  3   6/  4
                    7/  2
9097  100  199 M   1       1/  4   3/  3
                    9/  9   5/  3   6/  2
                    4/  4   3/  3   1/  4
                    6/  1   101
9003  102  103 M   2       3/  2   6/  2
                   10/ 10   4/  3   2/  6
                    7/  4   2/  3   3/  4
                    4/  6   201
9004  102  103 M   1       3/  1   6/  3
                   10/ 10   3/  3   6/  6
                    4/  3   4/  1   5/  6
                    4/  6   101
9005    1  104 F   2       3/  1   6/  4
                   10/  9   3/  3   2/  6
                    7/  5   2/  1   3/  7
                    4/  6   201
9006    1  104 M           1/  1   3/  4
                    9/  9   5/  3   6/  6
                    3/  5   3/  1   1/  7
                    6/  6
```

Output Data File B4

Original Pedigree (No Missing Data)

```
Episodic Ataxia (EA) pedigree: LOD score

Overall Location Scores from SimWalk2
MENDEL version 3.35 by Ken Lange is used in the computation of these results


This file contains the location scores calculated using ALL pedigrees.
These location scores are directly comparable to multipoint LOD scores.
All location scores and log-likelihoods are in log10 units.

All the parameters used to set up this run, including the marker map
and penetrance model, if any, are listed at the end of this file.

The following tables are comma-delimited to facilitate graphing these results.

The first table below lists the largest location score found in the
analysis of each pedigree individually. (Only the pedigrees within
this run are listed.) Also listed for each pedigree is the first
position that had this score, given in distance (in sex-average Haldane cM)
from the first marker in this study.

        ,            ,                  ,      LARGEST      ,
PEDIGREE,  PEDIGREE  ,     POSITION     ,  LOCATION SCORE   ,
NUMBER ,    NAME     ,   in Haldane cM  ,  when alpha=1.00  ,

 001    ,  20        ,6.6609    , 3.609       ,


The remaining results in this file are calculated by combining
ALL the pedigrees in this study (including any previous runs that
were 'continued' into this run).

The overall log-likelihood considering ONLY THE TRAIT is:       -8.554

The overall log-likelihood considering ONLY THE MARKERS is:   -342.330

The following table lists the overall location scores for positions
(measured in sex-average Haldane cM) throughout the specified markers.

 MARKER ,  POSITION  , LOCATION SCORE ,                , MAX HETEROGEN  ,  ALPHA
  NAME  , Haldane cM ,   alpha=1.00   ,                , LOCATION SCORE ,  VALUE

        ,   -49.9999 ,       1.546    ,                ,       1.546    ,   1.00
        ,   -45.0000 ,       1.698    ,                ,       1.698    ,   1.00
        ,   -40.0000 ,       1.859    ,                ,       1.859    ,   1.00
        ,   -35.0000 ,       2.030    ,                ,       2.030    ,   1.00
        ,   -30.0000 ,       2.208    ,                ,       2.208    ,   1.00
        ,   -25.0000 ,       2.390    ,                ,       2.390    ,   1.00
        ,   -20.0000 ,       2.574    ,                ,       2.574    ,   1.00
        ,   -15.0000 ,       2.751    ,                ,       2.751    ,   1.00
        ,   -10.0000 ,       2.905    ,                ,       2.905    ,   1.00
        ,    -5.0000 ,       2.993    ,                ,       2.993    ,   1.00
        ,    -0.0001 ,       2.790    ,                ,       2.790    ,   1.00
S91
        ,     0.0001 ,       2.790    ,                ,       2.790    ,   1.00
        ,     0.1010 ,       2.768    ,                ,       2.768    ,   1.00
        ,     0.2020 ,       2.745    ,                ,       2.745    ,   1.00
        ,     0.3030 ,       2.721    ,                ,       2.721    ,   1.00
        ,     0.4041 ,       2.695    ,                ,       2.695    ,   1.00
        ,     0.5051 ,       2.667    ,                ,       2.667    ,   1.00
        ,     0.6061 ,       2.638    ,                ,       2.638    ,   1.00
        ,     0.7071 ,       2.606    ,                ,       2.606    ,   1.00
        ,     0.8081 ,       2.572    ,                ,       2.572    ,   1.00
        ,     0.9091 ,       2.534    ,                ,       2.534    ,   1.00
        ,     1.0100 ,       2.494    ,                ,       2.494    ,   1.00
S100
        ,     1.0102 ,       2.494    ,                ,       2.494    ,   1.00
```

```
        ,     1.1111 ,       2.489    ,                ,       2.489    ,   1.00
        ,     1.2122 ,       2.485    ,                ,       2.485    ,   1.00
        ,     1.3132 ,       2.480    ,                ,       2.480    ,   1.00
        ,     1.4142 ,       2.475    ,                ,       2.475    ,   1.00
        ,     1.5152 ,       2.471    ,                ,       2.471    ,   1.00
        ,     1.6162 ,       2.466    ,                ,       2.466    ,   1.00
        ,     1.7172 ,       2.460    ,                ,       2.460    ,   1.00
        ,     1.8182 ,       2.455    ,                ,       2.455    ,   1.00
        ,     1.9193 ,       2.450    ,                ,       2.450    ,   1.00
        ,     2.0202 ,       2.444    ,                ,       2.444    ,   1.00
CACNL1A1
        ,     2.0204 ,       2.444    ,                ,       2.444    ,   1.00
        ,     2.3296 ,       2.429    ,                ,       2.429    ,   1.00
        ,     2.6390 ,       2.412    ,                ,       2.412    ,   1.00
        ,     2.9484 ,       2.393    ,                ,       2.393    ,   1.00
        ,     3.2578 ,       2.373    ,                ,       2.373    ,   1.00
        ,     3.5672 ,       2.351    ,                ,       2.351    ,   1.00
        ,     3.8765 ,       2.327    ,                ,       2.327    ,   1.00
        ,     4.1859 ,       2.300    ,                ,       2.300    ,   1.00
        ,     4.4953 ,       2.272    ,                ,       2.272    ,   1.00
        ,     4.8047 ,       2.241    ,                ,       2.241    ,   1.00
        ,     5.1139 ,       2.208    ,                ,       2.208    ,   1.00
S372
        ,     5.1141 ,       2.209    ,                ,       2.209    ,   1.00
        ,     5.4234 ,       3.199    ,                ,       3.199    ,   1.00
        ,     5.7328 ,       3.428    ,                ,       3.428    ,   1.00
        ,     6.0422 ,       3.539    ,                ,       3.539    ,   1.00
        ,     6.3515 ,       3.593    ,                ,       3.593    ,   1.00
        ,     6.6609 ,       3.609    ,                ,       3.609    ,   1.00
        ,     6.9703 ,       3.591    ,                ,       3.591    ,   1.00
        ,     7.2797 ,       3.533    ,                ,       3.533    ,   1.00
        ,     7.5891 ,       3.415    ,                ,       3.415    ,   1.00
        ,     7.8984 ,       3.170    ,                ,       3.170    ,   1.00
        ,     8.2077 ,       1.506    ,                ,       1.506    ,   1.00
pY2/1
        ,     8.2079 ,       1.499    ,                ,       1.499    ,   1.00
        ,     8.3088 ,       1.503    ,                ,       1.503    ,   1.00
        ,     8.4098 ,       1.506    ,                ,       1.506    ,   1.00
        ,     8.5109 ,       1.508    ,                ,       1.508    ,   1.00
        ,     8.6119 ,       1.509    ,                ,       1.509    ,   1.00
        ,     8.7129 ,       1.510    ,                ,       1.510    ,   1.00
        ,     8.8139 ,       1.509    ,                ,       1.509    ,   1.00
        ,     8.9149 ,       1.508    ,                ,       1.508    ,   1.00
        ,     9.0159 ,       1.506    ,                ,       1.506    ,   1.00
        ,     9.1169 ,       1.503    ,                ,       1.503    ,   1.00
        ,     9.2178 ,       1.499    ,                ,       1.499    ,   1.00
pY21/1
        ,     9.2180 ,       1.499    ,                ,       1.499    ,   1.00
        ,     9.3190 ,       1.503    ,                ,       1.503    ,   1.00
        ,     9.4200 ,       1.506    ,                ,       1.506    ,   1.00
        ,     9.5210 ,       1.508    ,                ,       1.508    ,   1.00
        ,     9.6220 ,       1.509    ,                ,       1.509    ,   1.00
        ,     9.7230 ,       1.510    ,                ,       1.510    ,   1.00
        ,     9.8240 ,       1.509    ,                ,       1.509    ,   1.00
        ,     9.9250 ,       1.508    ,                ,       1.508    ,   1.00
        ,    10.0261 ,       1.506    ,                ,       1.506    ,   1.00
        ,    10.1271 ,       1.503    ,                ,       1.503    ,   1.00
        ,    10.2280 ,       1.499    ,                ,       1.499    ,   1.00
KCNA5
        ,    10.2282 ,       1.499    ,                ,       1.499    ,   1.00
        ,    10.3291 ,       1.503    ,                ,       1.503    ,   1.00
        ,    10.4301 ,       1.506    ,                ,       1.506    ,   1.00
        ,    10.5311 ,       1.508    ,                ,       1.508    ,   1.00
        ,    10.6321 ,       1.509    ,                ,       1.509    ,   1.00
        ,    10.7331 ,       1.510    ,                ,       1.510    ,   1.00
        ,    10.8342 ,       1.509    ,                ,       1.509    ,   1.00
        ,    10.9352 ,       1.508    ,                ,       1.508    ,   1.00
        ,    11.0362 ,       1.506    ,                ,       1.506    ,   1.00
        ,    11.1372 ,       1.503    ,                ,       1.503    ,   1.00
        ,    11.2381 ,       1.499    ,                ,       1.499    ,   1.00
S99
```

```
        ,    11.2383 ,       1.499    ,                ,       1.499    ,   1.00
        ,    11.3392 ,       1.500    ,                ,       1.500    ,   1.00
        ,    11.4402 ,       1.499    ,                ,       1.499    ,   1.00
        ,    11.5413 ,       1.498    ,                ,       1.498    ,   1.00
        ,    11.6423 ,       1.496    ,                ,       1.496    ,   1.00
        ,    11.7433 ,       1.493    ,                ,       1.493    ,   1.00
        ,    11.8443 ,       1.489    ,                ,       1.489    ,   1.00
        ,    11.9453 ,       1.484    ,                ,       1.484    ,   1.00
        ,    12.0463 ,       1.479    ,                ,       1.479    ,   1.00
        ,    12.1473 ,       1.473    ,                ,       1.473    ,   1.00
        ,    12.2483 ,       1.466    ,                ,       1.466    ,   1.00
S93
        ,    12.2485 ,       1.466    ,                ,       1.466    ,   1.00
        ,    17.2484 ,       2.832    ,                ,       2.832    ,   1.00
        ,    22.2484 ,       2.809    ,                ,       2.809    ,   1.00
        ,    27.2484 ,       2.680    ,                ,       2.680    ,   1.00
        ,    32.2484 ,       2.517    ,                ,       2.517    ,   1.00
        ,    37.2484 ,       2.342    ,                ,       2.342    ,   1.00
        ,    42.2484 ,       2.166    ,                ,       2.166    ,   1.00
        ,    47.2484 ,       1.993    ,                ,       1.993    ,   1.00
        ,    52.2484 ,       1.827    ,                ,       1.827    ,   1.00
        ,    57.2484 ,       1.669    ,                ,       1.669    ,   1.00
        ,    62.2483 ,       1.520    ,                ,       1.520    ,   1.00


 For alpha = 1.00,  the largest overall location score is:      3.609
                              at position (in Haldane cM):      6.6609

 On the alpha grid, the largest overall location score is:      3.609
                              at position (in Haldane cM):      6.6609
                       obtained by using an alpha value of:      1.000
```

Episodic Ataxia (EA) pedigree: LOD score

Recombination Statistics from SimWalk2 2.91

Here, for each marker interval, is the number of recombinations observed
and the number expected, given the total number of meioses & the interval size.

The 'observed value' listed below is the average over all pedigrees that were
sampled during the MCMC phase. Thus, this value is an estimate of the average
over all possible configurations each weighted by its likelihood, i.e.,
given the data, this is our best estimate of the true value.

We also estimate the p-value for these observations, i.e., the probability that
one would observe this many recombination events OR MORE within this interval,
again given the total number of meioses & the interval size. A very small
p-value indicates that the number of observed recombinations would be more
consistent with a larger recombination distance for this marker interval.
Similarly, a p-value close to 1 indicates that the user-specified recombination
distance for this interval should be re-evaluated and perhaps made smaller.
Such extreme p-values are flagged when they appear.

```
The total number of pedigrees actually analyzed here:             1
The total number of meioses contained in these pedigrees:        40
```

Since the user-specified recombination distances are sex-independent,
we report the combined female and male recombination statistics.

OVERALL RECOMBINATION STATISTICS

POSITION Haldane cM	MARKER NAME	RECOMB. FRACTION	RECOMBINATION EVENTS OBSERVED	& EXPECTED	SIGNIFICANCE (P-VALUE)		INTERVAL NUMBER
0.000	S91						
		0.01000	0.173	0.400	0.87444		1
1.010	S100						
		0.01000	0.158	0.400	0.88566		2
2.020	CACNL1A1						
		0.03000	0.917	1.200	0.73621		3
5.114	S372						
		0.03000	2.716	1.200	0.16373		4
8.208	pY2/1						
		0.01000	0.418	0.400	0.69082		5
9.218	pY21/1						
		0.01000	1.030	0.400	0.31689		6
10.228	KCNA5						
		0.01000	0.006	0.400	0.99585	##!	7
11.238	S99						
		0.01000	0.137	0.400	0.90131		8
12.248	S93						

##! indicates a p-value which is so large that one should reconsider whether the specified recombination fraction for this interval is too large!

Model at the trait locus named: EA

ALLELE	FREQUENCY
1	0.99990
2	0.00010

PHENOTYPE	COMPATIBLE ORDERED GENOTYPES		
1	1/1	1/2	2/2
1		2/1	
2	1/1	1/2	2/2
2		2/1	

PHENOTYPE/LIABILITY CLASS	PENETRANCES FOR UNORDERED GENOTYPES: 1/1	1/2	2/2
101	0.999000	0.010000	0.010000
201	0.001000	0.990000	0.990000

Marker Map:

POSITION (Haldane cM) FEMALE	& MALE	MARKER ORDER	RECOMBINATION RATES FEMALE	& MALE	INTERVAL NUMBER
0.000	0.000	S91			
			0.01000	0.01000	1
1.010	1.010	S100			
			0.01000	0.01000	2
2.020	2.020	CACNL1A1			
			0.03000	0.03000	3
5.114	5.114	S372			
			0.03000	0.03000	4
8.208	8.208	pY2/1			
			0.01000	0.01000	5
9.218	9.218	pY21/1			
			0.01000	0.01000	6
10.228	10.228	KCNA5			
			0.01000	0.01000	7
11.238	11.238	S99			
			0.01000	0.01000	8
12.248	12.248	S93			

APPENDIX C

Table C1

Comparison of Algorithms

ALGORITHM	PROGRAMS	SOLUTION	SIZE RESTRICTIONS
ELSTON – STEWART	FASTLINK	EXACT	VARIES: ABT 8 LOCI
	LINKAGE		(LESS WITH LOOPS)
	MENDEL		
	VITESSE		
LANDER-GREEN-KRUGLYAK	ALLEGRO	EXACT	ABT 20 PEOPLE
	GENEHUNTER		(2n - f) < 20
	MENDEL		
	MERLIN		
MARKOV CHAIN MONTE CARLO	LOKI	ESTIMATE	MUCH LARGER
	SIMWALK2		> 1000 INDIVIDUALS
			> 1000 LOCI

ALGORITHM	INCREASE IN COMPUTATIONAL TIME WITH INCREASE IN:		
	PEOPLE	MARKERS	MISSING DATA
ELSTON – STEWART	LINEAR	EXPONENTIAL	SEVERE
LANDER-GREEN-KRUGLYAK	EXPONENTIAL	LINEAR	MODEST
MARKOV CHAIN MONTE CARLO	LINEAR	LINEAR	MILD

Table C2
SimWalk2 Raw Accuracy Results

	Case	Position	Lgst Location Score		Case	Position	Lgst Location Score
Original Pedigree				**Pedigree A**			
	1	**6.6609**	**3.609**		1	**6.6609**	**3.291**
	2	**6.6609**	**3.597**		2	**6.6609**	**3.258**
	3	**6.6609**	**3.59**		3	**6.6609**	**3.241**
	4	**6.6609**	**3.563**		4	**6.6609**	**3.185**
	5	**6.6609**	**3.537**		5	**6.6609**	**3.143**
	6	**6.9703**	**3.265**		6	6.9703	2.954
	7	7.2797	2.783		7	7.2797	1.86
	8	**6.6609**	**3.176**		8	6.6609	2.863
	9	7.2797	2.397		9	7.2797	1.436
	10	**6.9703**	**3.084**		10	6.9703	2.76
	11	6.9703	2.939		11	6.9703	2.634
	12	7.2797	2.414		12	7.2797	2.126
	13	6.9703	2.876		13	6.9703	2.543
	14	6.9703	2.714		14	6.9703	2.412
	15	9.2178	2.112		15	7.5891	1.838
	AVG	7.058253333	3.043733333		AVG	6.949673333	2.636266667
Pedigree B				**Pedigree C**			
	1	6.6609	2.718		1	**8.2079**	**3.291**
	2	6.6609	2.442		2	8.2077	2.992
	3	6.6609	2.353		3	8.2077	2.892
	4	6.6609	2.168		4	8.2077	2.669
	5	6.6609	2.079		5	8.5109	2.553
	6	6.9703	1.745		6	8.2077	1.677
	7	7.2797	0.282		7	8.2077	0.259
	8	6.6609	1.63		8	8.2077	1.537
	9	6.9703	0.161		9	8.2077	0.146
	10	6.9703	1.843		10	9.2178	2.255
	11	6.9703	1.781		11	8.9149	2.175
	12	7.2797	1.28		12	9.2178	1.537
	13	6.9703	1.623		13	9.1169	1.992
	14	6.9703	1.586		14	9.1169	1.948
	15	7.5891	1.087		15	9.2178	1.303
	AVG	6.929046667	1.651866667		AVG	8.59832	1.9484

Table C3
Pedigree Analysis by Parameter Allele Frequency for Locational Position

LOCUS	EXACT	ORIGINAL	A	B	C
LOCUS 1	6.6609	6.6609	6.6609	6.6609	8.2079
LOCUS 2	6.6609	6.6609	6.6609	6.6609	8.2077
LOCUS 3	6.6609	6.6609	6.6609	6.6609	8.2077
LOCUS 4	6.6609	6.6609	6.6609	6.6609	8.2077
LOCUS 5	6.6609	6.6609	6.6609	6.6609	8.5109

Table C4

Pedigree Analysis by Parameter Penetrance for Locational Position

PENETRANCE	EXACT	ORIGINAL	A	B	C
PEN 1	6.6609	6.6609	6.6609	6.6609	8.2079
PEN 2	6.6609	6.9703	6.9703	6.9703	8.2077
PEN 3	6.6609	7.2797	7.2797	7.2797	8.2077
PEN 4	6.6609	6.6609	6.6609	6.6609	8.2077
PEN 5	6.6609	7.2797	7.2797	6.9703	8.2077

Table C5

Pedigree Analysis by Parameters Allele Frequency and Penetrance for Locational Position

PARAMETERS	EXACT	ORIGINAL	A	B	C
LOCUS 3, PEN 2	6.6609	6.9703	6.9703	6.9703	9.2178
LOCUS 3, PEN 4	6.6609	6.9703	6.9703	6.9703	8.9149
LOCUS 3, PEN 5	6.6609	7.2797	7.2797	7.2797	9.2178

Table C6

Pedigree Analysis by Parameters Allele Frequency and Penetrance (Severe Case) for Locational Position

PARAMETERS	EXACT	ORIGINAL	A	B	C
LOCUS 4, PEN 2	6.6609	6.9703	6.9703	6.9703	9.1169
LOCUS 4, PEN 4	6.6609	6.9703	6.9703	6.9703	9.1169
LOCUS 4, PEN 5	6.6609	9.2178	7.5891	7.5891	9.2178

Table C7

Pedigree Analysis by Parameter Allele Frequency for Location Score

LOCUS	EXACT	ORIGINAL	A	B	C
LOCUS 1	3.56	3.609	3.291	2.718	3.291
LOCUS 2	3.56	3.597	3.258	2.442	2.992
LOCUS 3	3.56	3.59	3.241	2.353	2.892
LOCUS 4	3.56	3.563	3.185	2.168	2.669
LOCUS 5	3.56	3.537	3.143	2.079	2.553

Table C8

Pedigree Analysis by Parameter Penetrance for Location Score

PENETRANCE	EXACT	ORIGINAL	A	B	C
PEN 1	3.56	3.609	3.291	2.718	3.291
PEN 2	3.56	3.265	2.954	1.745	1.677
PEN 3	3.56	2.783	1.86	0.282	0.259
PEN 4	3.56	3.176	2.863	1.63	1.537
PEN 5	3.56	2.397	1.436	0.161	0.146

Table C9

Pedigree Analysis by Parameters Allele Frequency and Penetrance for Location Score

PARAMETERS	EXACT	ORIGINAL	A	B	C
LOCUS 3, PEN 2	3.56	3.084	2.76	1.843	2.255
LOCUS 3, PEN 4	3.56	2.939	2.634	1.781	2.175
LOCUS 3, PEN 5	3.56	2.414	2.126	1.28	1.537

Table C10

Pedigree Analysis by Parameters Allele Frequency and Penetrance (Severe Case) for Location Score

PARAMETERS	EXACT	ORIGINAL	A	B	C
LOCUS 4, PEN 2	3.56	2.876	2.543	1.623	1.992
LOCUS 4, PEN 4	3.56	2.714	2.412	1.586	1.948
LOCUS 4, PEN 5	3.56	2.112	1.838	1.087	1.303

Table C11

Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency for Location Score

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
LOCUS 1	5	16.469	3.2938	0.1254437
LOCUS 2	5	15.849	3.1698	0.2259602
LOCUS 3	5	15.636	3.1272	0.2672887
LOCUS 4	5	15.145	3.029	0.3655935
LOCUS 5	5	14.872	2.9744	0.4167978
EXACT	5	17.8	3.56	0
ORIGINAL	5	17.896	3.5792	0.0008412
A	5	16.118	3.2236	0.0035038
B	5	11.76	2.352	0.0626105
C	5	14.397	2.8794	0.0832643

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Allele Frequency	0.31105976	4	0.0777649	4.2931525	0.0150593	3.0069174
Pedigree	5.31451616	4	1.328629	73.34934	4.392E-10	3.0069174
Error	0.28981944	16	0.0181137			
Total	5.91539536	24				

Table C12

Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Penetrance for Location Score

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
PEN 1	5	16.469	3.2938	0.1254437
PEN 2	5	13.201	2.6402	0.7660037
PEN 3	5	8.744	1.7488	2.1833467
PEN 4	5	12.766	2.5532	0.8456157
PEN 5	5	7.7	1.54	2.1676355
EXACT	5	17.8	3.56	0
ORIGINAL	5	15.23	3.046	0.21805
A	5	12.404	2.4808	0.6258547
B	5	6.536	1.3072	1.1627587
C	5	6.91	1.382	1.636039

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Penetrance	10.1683548	4	2.5420887	9.2388045	0.0004572	3.0069174
Pedigree	19.9497264	4	4.9874316	18.126002	8.618E-06	3.0069174
Error	4.4024548	16	0.2751534			
Total	34.520536	24				

Table C13

Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency and Penetrance for Location Score

SUMMARY	*Count*	*Sum*	*Average*	*Variance*		
LOCUS 3, PEN 2	5	13.502	2.7004	0.4557823		
LOCUS 3, PEN 4	5	13.089	2.6178	0.4718697		
LOCUS 3, PEN 5	5	10.917	2.1834	0.7963658		
EXACT	3	10.68	3.56	0		
ORIGINAL	3	8.437	2.8123333	0.1242583		
A	3	7.52	2.5066667	0.1126493		
B	3	4.904	1.6346667	0.0953023		
C	3	5.967	1.989	0.154828		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Parameters	0.77135853	2	0.3856793	15.220366	0.0018758	4.4589683
Pedigree	6.69335373	4	1.6733384	66.03628	3.646E-06	3.8378545
Error	0.20271747	8	0.0253397			
Total	7.66742973	14				

Table C14

Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency and Penetrance (Severe Case) for Location Score

SUMMARY	*Count*	*Sum*	*Average*	*Variance*		
LOCUS 4, PEN 2	5	12.594	2.5188	0.5730627		
LOCUS 4, PEN 4	5	12.22	2.444	0.57539		
LOCUS 4, PEN 5	5	9.9	1.98	0.9474415		
EXACT	3	10.68	3.56	0		
ORIGINAL	3	7.702	2.5673333	0.1620573		
A	3	6.793	2.2643333	0.1406103		
B	3	4.296	1.432	0.089611		
C	3	5.243	1.7476667	0.1487803		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Parameters	0.85199413	2	0.4259971	14.809314	0.0020453	4.4589683
Pedigree	8.15345293	4	2.0383632	70.861428	2.776E-06	3.8378545
Error	0.23012387	8	0.0287655			
Total	9.23557093	14				

Table C15

Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency for Locational Position

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
LOCUS 1	5	34.8515	6.9703	0.4786418
LOCUS 2	5	34.8513	6.97026	0.478518048
LOCUS 3	5	34.8513	6.97026	0.478518048
LOCUS 4	5	34.8513	6.97026	0.478518048
LOCUS 5	5	35.1545	7.0309	0.6845
EXACT	5	33.3045	6.6609	1.42109E-14
ORIGINAL	5	33.3045	6.6609	1.42109E-14
A	5	33.3045	6.6609	1.42109E-14
B	5	33.3045	6.6609	1.42109E-14
C	5	41.3419	8.26838	0.018379992

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Allele Frequency	0.01470399	4	0.003676	1	0.436207616	3.006917382
Pedigree	10.3359678	4	2.58399195	702.9360922	9.40787E-18	3.006917382
Error	0.05881597	16	0.003676			
Total	10.4094878	24				

Table C16

Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Penetrance for Locational Position

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
PEN 1	5	34.8515	6.9703	0.4786418
PEN 2	5	35.7795	7.1559	0.36366258
PEN 3	5	36.7077	7.34154	0.306244128
PEN 4	5	34.8513	6.97026	0.478518048
PEN 5	5	36.3983	7.27966	0.334956448
EXACT	5	33.3045	6.6609	1.42109E-14
ORIGINAL	5	34.8515	6.9703	0.09572836
A	5	34.8515	6.9703	0.09572836
B	5	34.5421	6.90842	0.067009852
C	5	41.0387	8.20774	8E-09

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Penetrance	0.5896174	4	0.14740435	5.308892107	0.006464877	3.006917382
Pedigree	7.40384309	4	1.85096077	66.66391507	9.0099E-10	3.006917382
Error	0.44424892	16	0.02776556			
Total	8.43770941	24				

Table C17

Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency and Penetrance for Locational Position

SUMMARY	*Count*	*Sum*	*Average*	*Variance*		
LOCUS 3, PEN 2	5	36.7896	7.35792	1.098934572		
LOCUS 3, PEN 4	5	36.4867	7.29734	0.835605428		
LOCUS 3, PEN 5	5	37.7178	7.54356	0.947758638		
EXACT	3	19.9827	6.6609	0		
ORIGINAL	3	21.2203	7.07343333	0.031909453		
A	3	21.2203	7.07343333	0.031909453		
B	3	21.2203	7.07343333	0.031909453		
C	3	27.3505	9.11683333	0.030582803		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Parameters	0.16459406	2	0.08229703	7.479145442	0.014743539	4.458968306
Pedigree	11.4411663	4	2.86029157	259.9430017	1.68889E-08	3.837854479
Error	0.08802827	8	0.01100353			
Total	11.6937886	14				

Table C18

Two-Factor ANOVA Summary for Missing Data in a Pedigree and Variation in Allele Frequency and Penetrance (Severe Case) for Locational Position

SUMMARY	*Count*	*Sum*	*Average*	*Variance*		
LOCUS 4, PEN 2	5	36.6887	7.33774	1.007139788		
LOCUS 4, PEN 4	5	36.6887	7.33774	1.007139788		
LOCUS 4, PEN 5	5	40.2747	8.05494	1.270462023		
EXACT	3	19.9827	6.6609	0		
ORIGINAL	3	23.1584	7.71946667	1.683752083		
A	3	21.5297	7.17656667	0.127637813		
B	3	21.5297	7.17656667	0.127637813		
C	3	27.4516	9.15053333	0.003393603		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Parameters	1.71458613	2	0.85729307	3.160153905	0.097398111	4.458968306
Pedigree	10.9687099	4	2.74217748	10.10821526	0.003230117	3.837854479
Error	2.17025649	8	0.27128206			
Total	14.8535525	14				

Table C19

SimWalk2 Raw Convergence Results

	Original Pedigree		Pedigree A		Pedigree B		Pedigree C	
Case	**CP(B)**	**CR**	**CP(B)**	**CR**	**CP(B)**	**CR**	**CP(B)**	**CR**
1	1.3374	0.2523	1.5713	0.3636	1.832	0.4541	2.4218	0.5871
2	1.333	0.2498	1.5596	0.3588	1.9475	0.4865	2.4376	0.5898
3	1.3484	0.2584	1.3535	0.2612	1.7583	0.4313	2.4901	0.5984
4	1.3204	0.2427	1.4376	0.3044	1.6661	0.3998	2.2352	0.5748
5	1.4107	0.2911	1.5368	0.3493	1.7053	0.4136	2.3779	0.5795
6	1.4099	0.2907	1.4514	0.311	1.8298	0.4535	2.5108	0.6017
7	1.5313	0.347	1.462	0.316	1.8613	0.4627	2.5374	0.6059
8	1.4769	0.3229	1.619	0.3823	1.899	0.4734	2.5283	0.6045
9	1.8882	0.4704	1.6888	0.4079	2.0007	0.5002	2.5366	0.6058
10	1.5429	0.3519	1.5676	0.3621	2.0104	0.5026	2.2749	0.5604
11	1.4627	0.3163	1.6457	0.3924	1.952	0.4877	2.4014	0.5836
12	1.6699	0.4012	1.7677	0.4343	1.9976	0.4994	2.5162	0.6026
13	1.6658	0.3997	1.7231	0.4197	1.8245	0.4519	2.6008	0.6155
14	1.6657	0.3997	1.5402	0.3507	1.9031	0.4745	2.5793	0.6123
15	1.7827	0.4391	1.8264	0.4525	1.9827	0.4956	2.5891	0.6138
AVG	1.52306	0.335547	1.58338	0.364413	1.87802	0.465787	2.46916	0.595713