RESEARCH PLAN

1: Specific Aims

Long-term objective:

The overall objective is to continue to develop methods and software for the analysis of the genetic basis of complex familial traits, using a Monte Carlo approach to problems of likelihood evaluation and maximization on complex data structures. There is a need for techniques for likelihood analysis of complex models that combine linkage and segregation analysis; only then can we address questions about traits, such as coronary artery disease, that are influenced by environmental factors as well as by genes at several loci. Using genetic models we propose to develop methods of likelihood analysis of phenotypic and marker data observed on members of a pedigree. This will aid in the identification of markers which are linked to loci which contribute to the determination of complex genetic traits. Our present objectives build on those accomplished in our currently funded research.

Specific objectives:

1. To develop effective Monte Carlo computational methods for the computation of multilocus linkage likelihoods on extended and/or complex pedigrees:
(a) for Mendelian traits mapped against an array of linked markers,
(b) for traits that are genetically heterogeneous, of incomplete penetrance, or delayed onset.
(c) for genetically determined quantitative risk factors of complex genetic disease.

2. To develop further our Monte Carlo methods for maximizing likelihoods and fitting models for complex genetic traits, including linkage analysis of marker data,
(a) using censored age-of-onset data for diseases of delayed onset,
(b) analyzing data on genetically determined quantitative risk factors, jointly with disease phenotype information,
(c) making appropriate ascertainment corrections for pedigrees ascertained through affected individuals.

3. To develop Monte Carlo methods for multilocus linkage analyses based on the affected members of a pedigree, using an array of linked markers,
(a) extending current methods for homozygosity mapping of affected inbred individuals with complex ancestry,
(b) considering jointly several affected individuals in an extended or complex pedigree,
(c) investigating the potential for application to complex traits, and to affection status defined by extreme values of a quantitative trait.

4. To evaluate the procedures of specific aims #1, #2 and #3 by:
(a) analyzing Alzheimer's disease data, and other qualitative genetic disease data available to us, using data on multiple markers,
(b) analyzing Apolipoprotein B levels, and other quantitative measurements, using available data on multiple markers,
(c) performing simulation studies to assess the performance of methods and comparing results to those obtained by other computational approaches.

5. To extend Monte Carlo likelihood to further areas of multilocus linkage analysis,
(a) including locus ordering problems at a range of mapping scales,
(b) developing probability estimation procedures for models of a continuous genome,

(c) applying such models to genome-matching analyses of marker data on sets of related affected individuals, as in #3.

6. To produce user software,

(a) integrating our current methods and programs into a coherent package,

(b) implementing new methodology as proposed in #1, #2, #3 and #5,

(c) distributing and supporting the software, serving users through electronic mail, and seeking user response, to aid in ongoing software improvement.

2: Background and significance.

2.1: The genetic approach.

The risk for developing many common diseases is believed to be influenced by both genetic and environmental factors. Examples include coronary artery disease, hypertension, Alzheimer's disease, diabetes, several forms of mental illness, and some cancers. While the data necessary to understand different complex traits is disease-specific, the underlying principles used to identify and dissect the genetic component are applicable to a variety of diseases.

The identification of genes which influence the development of diseases provides a powerful tool for the study of disease. Identification of such genes can be done through one of two mechanisms: through study of loci which are suspected to play a role in the disease (candidate genes), or through mapping and identification of previously unknown loci (positional cloning), which then can be used to identify other relevant biochemical pathways. The candidate gene approach can be effective when considerable information about the biochemistry and physiology of the disease is already known, but is generally ineffective otherwise.

The positional cloning approach can be extraordinarily useful when little prior information about the underlying biology of a disease is available. An excellent example is the identification of the low density lipoprotein receptor (LDLR), which is defective in rare individuals with familial hypercholesterolemia (Brown & Goldstein, 1986). Proteins identified through interactions with the LDLR primary defect were studied as candidate genes for more common causes of coronary artery disease, leading to increased understanding of the biochemistry and physiology leading to such disease.

The positional cloning approach involves two initial types of studies. First, a genetic model is developed which at least approximates the mode or modes of inheritance of the disease. The model is needed to choose strategies for identifying the gene(s) of interest, and to provide parameter estimates for gene mapping studies. The model need not be a perfect reflection of the mode of inheritance of the trait, but must reflect the general dominance characteristics of the trait (Clerget-Darpoux et al., 1986; Greenberg & Hodge, 1989) Second, one or more of the relevant genes is localized (mapped) to a chromosomal region. This localization is followed by more precise localization to a small enough region so that laboratory techniques can be used to find the gene of interest. The positional cloning approach has been enormously successful for identifying genes in which defects cause simple genetic diseases. Examples of genes identified this way include those responsible for Huntington's disease (Huntington's disease collaborative group, 1992), cystic fibrosis (Riordan et al., 1989), and achondroplasia (Shiang et al., 1994) and more than 30 others.

2.2 Complex diseases.

Complex genetic disorders are also starting to yield to the genetic approach. Such disorders have more than one interacting genetic component and/or genetically influenced risk factor. This may include the relatively simple case where two or more loci each can individually produce the same phenotype (genetic heterogeneity), or it may describe the situation where several genes interact with each other and with different environments to produce manifestation(s) of the trait. Examples of mapped genes involved in such complex disorders include early-onset breast cancer gene on chromosomes 17 (Hall et al., 1990) and 13 (Wooster et al., 1994), loci contributing to early-onset Alzheimer's disease on chromosome 21 (Goate et al., 1991) and 14 (Schellenberg et al., 1992a), and a recent report of at least four loci on chromosomes 6p, 6q, 11p and 11q, each of which contributes to risk for insulin dependent diabetes (Davies et al., 1994). Recently, the chromosome 17 gene for familial breast cancer (BRCA1) has been identified (Miki et al., 1994), demonstrating that once a gene contributing to a complex disorder is mapped, it is possible to identify it.

Despite some success with complex diseases, there are many difficulties with the genetic analyses. The success rate in mapping such disorders has been much lower than for Mendelian traits, and for the most part, the examples of mapped complex diseases above represent unusual forms of the diseases, with particularly severe symptoms or early onset. Progress has been slow for even these forms of the diseases. With the exception of BRCA1, no other locus contributing to a complex disease has yet been identified through positional cloning, and few genes have been mapped or identified as having effects on more typical forms of complex diseases. There have also been many examples of apparent false-positive mapping results for complex diseases, a situation which rarely arises for Mendelian diseases (Rao et al., 1978). Examples of false or unconfirmable positive reports of linkage exist for Alzheimer's disease (St George-Hyslop et al., 1987), familial combined hyperlipidemia (Wojiechowski et al., 1991), and manic depression (Egeland et al., 1988), each of which was followed up by reports of non-confirmation (Schellenberg et al., 1988, Wijsman et al., 1992), extensions to the original pedigree which destroyed the evidence in favor of linkage (Kelsoe et al., 1989), or later mapping of the gene defect in the original families to a different chromosome (St George-Hyslop et al., 1992).

Limitations in the available techniques of analysis undoubtedly contribute to the failure to identify genes contributing to complex traits. These limitations exist for two reasons. First, there are severe practical problems with performing computations for analyses of complex genetic models. Second, most of the techniques of genetic analysis in human genetics have been designed around discrete single-locus disorders. Problems with implementation, power, and interpretation of results of analyses exist when these techniques are used for analysis of complex disorders.

2.3 Linkage analysis

(i) Pairwise linkage analysis

Various methods of linkage analysis exist which can be used to identify genomic regions containing disease loci. The regions thus identified are generally very large and need to be refined with analyses using more marker loci. The methods which are currently available can be used to map genes for disease loci which have a substantial contribution to risk of disease within families segregating for such loci. However, the current methods may be inadequate for mapping loci contributing to complex traits, where the trait model may be poorly specified, or where the loci may have small contributions to disease risk.

The most common technique of linkage analysis is the method of lod scores (Morton, 1955). This method, set in a sequential testing framework, efficiently provides both positive and negative evidence of linkage for single-locus traits (Morton, 1955), as well providing an estimate of the genetic distance between the disease locus and linked marker loci. Because the method is based on likelihood, it maximizes the use of the information contained in the data on extended pedigrees, subject to assumptions about the genetic model. Use of a technique of analysis which makes efficient use of the data is important for gene mapping projects because it is expensive and difficult to collect samples for mapping studies from pedigree members.

Alternative techniques for performing linkage analysis also exist when complications such as incomplete penetrance or interactions between loci may interfere with use of the lod score method. These methods, based on identity-by-descent or identity-by-state methods, include the affected-pedigree-member method (Weeks & Lange, 1988) for analysis of discrete traits on extended pedigrees, affected sib-pair methods (Suarez & Van Eerdewegh, 1984; Blackwelder & Elston, 1985), and methods for analysis of quantitative traits in sib-pairs (Haseman & Elston, 1972) and in extended pedigrees (Olson & Wijsman, 1993). Although these techniques do not require specification of a genetic model for the disease, they can provide only a test of linkage, and not an estimate of gene location. Computation of a test statistic, with a distribution often known only under the null hypothesis, cannot substitute for a likelihood analysis. Power is always an issue, and without the ability to compute probabilities of data under alternative hypotheses, no rigorous inference is possible (Elston, 1994).

A number of problems exist in mapping gene(s) for complex diseases. Genetic heterogeneity (the presence of more than one locus causing the disease), incomplete penetrance, possible interactions between loci contributing to risk of the disease, and missing data all reduce the ease by which genes for complex disorders can be initially mapped. With the exception of the problems created by interactions between loci, the other scenarios primarily cause reduction in power to detect evidence of linkage (Clerget-Darpoux et al., 1986), and cause biased estimates of the location of the disease locus (Cavalli-Sforza & King, 1986) especially when parameters are misspecified in the analysis. The loss of power can be substantial. For example, the sample size necessary to detect linkage if the penetrance is 50% is about 4 times that needed if the penetrance is 100%, and the mean sample size necessary to detect linkage if 50% of the families are unlinked to the region is about 13 times that needed if the trait is caused by the same locus in all families (Ott, 1991).

(ii) Multipoint linkage analysis

It is difficult to refine the location of disease loci, even for Mendelian traits. Use of multiple markers is critical for pinpointing the location of disease genes, because the limiting factor in localization of a disease locus is often the available trait data on pedigree members, with different families and even different individuals within families being informative for different markers. Multipoint mapping methods much more accurately localize a gene to an interval flanked by markers used in the analysis than do 2-point methods (Lathrop et al., 1984). Although a small number of non-parametric multipoint techniques have been proposed (Goldgar, 1990; Fulker and Cardon, 1994), to date the most successful method has been use of the multipoint lod score.

A multipoint analysis is the only way that all the available information can be used so that the location of the gene can be pinpointed to a small enough region that laboratory methods can be used for gene identification. Multipoint methods are particularly important for gene localization for diseases with incomplete penetrance, delayed onset, or substantial missing marker data because of late onset of the disease. For such diseases, very few phase-known meioses can be identified. It takes 400 phase-known meioses to localize a gene to an interval of less than 1 centimorgan with 90% confidence (Boehnke, 1994). Since localization to such small intervals is desired before attempts to clone the gene are made, it is clear that every effort to maximize the available data in pedigrees segregating for complex diseases will be necessary.

The barrier to use of multipoint linkage analysis methods is that such methods are notoriously computationally intensive. For example, a multipoint analysis with only 3 marker loci, no more than 4 alleles per marker, and 23 pedigrees segregating for early onset Alzheimer's disease was estimated to require 3 months of computer time on a workstation computer (Schellenberg et al., 1992a). Similar computation times were estimated for breast cancer and bipolar disease pedigrees (Doug Easton, Lynn Goldin; personal communications). If there are many missing data, too many loci, or too many alleles per locus, analysis becomes difficult or infeasible because of both time and computer memory constraints Pedigree complexities introduce an additional computational burden. For example, computations are infeasible for multipoint linkage analyses on Hutterite pedigrees with multiple inbreeding loops (Nancy Cox, personal communication). Even the speed increases introduced into the *LINKAGE* analysis package (*FASTLINK*, Cottingham et al., 1993), have not eliminated these difficulties. Approximation methods such as that of Curtis & Gurling (1993) may provide a partial solution, but in order to assess such approximations, a method for obtaining exact likelihoods is needed at least for representative problems.

While investigators sometimes reduce computation time by resorting to overlapping multipoint analyses, a full analysis which uses all available markers jointly may be necessary for accurate localization of the gene. For example, a recent comparison of results obtained for cardiomyopathy with overlapping 3-point analyses as compared to a full 7-point analysis performed with the methods of sequential imputation (Kong et al., 1992) demonstrated that conclusions about the disease location differed in the two analyses (Kass et al., 1994). However, the method used for these analyses cannot be used when there are loops in the pedigree.

The computational difficulties inherent in use of multipoint methods are therefore a serious barrier in the analysis of complex diseases. Localization of a disease gene to an interval between closely spaced marker loci requires accurate specification of all parameters of the genetic model. This includes marker gene frequencies, genetic heterogeneity, and parameters describing penetrance of the disease. Inaccuracies in these parameter values can create apparent recombinants with markers leading to shifts of the estimated location of the trait locus to intervals outside of the correct interval (Buetow, 1991). Because these parameters are often estimated from small amounts of data, it is usually necessary to construct location score curves, needed for multipoint analysis, for a number of parameter values. This is frequently necessary even in analysis of Mendelian diseases with few complications, and even more necessary for complex diseases with many unknowns in the parameter values of the genetic model.

2.4 Monte Carlo likelihood.

In genetic analysis, the scientific facts of Mendelian segregation, and of linkage and recombination, have long provided models that mimic the fundamental genetic realities. For this reason, likelihood analysis has had a prominence it has not enjoyed in other areas. However, problems of estimation, fitting, and testing of models have been compounded by the difficulty of likelihood computation. We emphasize here that, in discussing linkage likelihoods, we use this term in the broad sense—the computation of probabilities of data under the hypotheses of alternative genetic maps. The data in question may consist of all available information on a pedigree, or set of pedigrees, as we consider in sections 4.1 and 4.2, or it may consist of a subset of such data, such as marker types of affected individuals (section 4.3), or even a specification of shared chromosome segments (section 4.5).

Markov chain Monte Carlo (Hastings, 1970) has recently become of great interest in statistical analysis, with the recognition that it permits realizations from complex joint and conditional distributions (Gelfand & Smith, 1990). In areas ranging from gamma camera images to agricultural field trials (Besag et al., 1994) this has led to a radical reassessment of models; statistical inference on the basis of realistic scientifically based models is now possible. Monte Carlo likelihood (Geyer & Thompson, 1992) is one aspect of this rapid expansion of Monte Carlo (MC) methods; complex dependencies, missing data, conditional inference, and constraints, can all be easily handled. Thus in order to introduce both our current and proposed work it is necessary to review, very briefly, the basis of MC likelihood computation.

Monte Carlo estimation of expectations is not new, but it has received greatly increased attention with the advent of fast workstations that make accurate MC computations possible. The key ideas are due to Hammersley & Handscomb (1964). Suppose we wish to sum $h(\mathbf{X})$ over a very large space of $\mathbf{X}$-values. For example, $h(\cdot)$ could be the joint probability of phenotypes and genotypes on a pedigree to be summed over all genotypic configurations to provide a likelihood. (For simplicity, all integration, whether of discrete or continuous variables will be written as discrete sums.) Then

$$H \;=\; \sum_{\mathbf{X}} h(\mathbf{X}) = \sum_{\mathbf{X}} \frac{h(\mathbf{X})}{g(\mathbf{X})} g(\mathbf{X}) \;=\; \mathrm{E}_g\left(\frac{h(\mathbf{X})}{g(\mathbf{X})}\right) \tag{1}$$

where $g(\cdot)$ is any probability distribution over the space of $\mathbf{X}$-values. If N realizations $\mathbf{X}^{(t)}$, $t = 1, ..., N$, are obtained from $g(\cdot)$, then

$$\hat{H} \;=\; \frac{1}{N} \sum_{t=1}^{N} \frac{h(\mathbf{X}^{(t)})}{g(\mathbf{X}^{(t)})} \tag{2}$$

is an unbiased estimator of H. However, in MC integration over large multidimensional spaces, it is critical that the values of $h(\mathbf{X})/g(\mathbf{X})$ be as uniform as possible, or no useful estimator will result. That is, the form of $g(\cdot)$ should mimic the form of $h(\cdot)$ are closely as possible. This is the basis of importance sampling (Hammersley & Handscomb, 1964); $g(\cdot)$ should assign greater sampling frequency to values of $\mathbf{X}$ that are "important" in evaluating the sum (1)—that is, having larger values of $h(\mathbf{X})$.

In genetic analysis, we wish to compare likelihoods for alternative models, θ, on the basis of phenotypic data, $\mathbf{Y}$, observed on some subset of members of a pedigree structure. Here, $\mathbf{Y}$ denotes all observed data, for traits and markers, for all observed individuals, and θ all the linkage and segregation analysis parameters that comprise the genetic model. The likelihood is most conveniently considered as a sum over latent unobserved variables, such as genotypes; these we denote by $\mathbf{X}$. Then

$$L(\theta) \;=\; \sum_{\mathbf{X}} P_\theta(\mathbf{Y}, \mathbf{X}) \;=\; \sum_{\mathbf{X}} P_\theta(\mathbf{Y} \mid \mathbf{X}) P_\theta(\mathbf{X}) \;=\; \mathrm{E}_\theta(P_\theta(\mathbf{Y} \mid \mathbf{X})) \;=\; \mathrm{E}_{\theta_0}\left(P_\theta(\mathbf{Y} \mid \mathbf{X}) \frac{P_\theta(\mathbf{X})}{P_{\theta_0}(\mathbf{X})}\right) \tag{3}$$

(Ott, 1979), where θ_0 is any chosen specific value of the set of model parameters θ. However these expectation expressions do not provide useful MC estimates of $L(\theta)$ on a large pedigree: most realizations of pedigree genotypes $\mathbf{X}$ from $P_\theta(\mathbf{X})$ or $P_{\theta_0}(\mathbf{X})$ will not even be compatible with data $\mathbf{Y}$. The last expectation in (3), due to K. Lange (Ott, 1979), does introduce two key ideas. The first is importance sampling reweighting—sampling from $P_{\theta_0}(\mathbf{X})$ instead of $P_\theta(\mathbf{X})$, and multiplying by the factor $P_\theta(\mathbf{X})/P_{\theta_0}(\mathbf{X})$ to obtain an unbiased estimator. The second is that these realizations from a single $P_{\theta_0}(\mathbf{X})$ could simultaneously estimate $L(\theta)$ as a function of θ; this is a key idea in general Monte Carlo likelihood theory (Geyer & Thompson, 1992).

In order to obtain useful realizations for MC analysis, it is necessary to simulate values of $\mathbf{X}$ conditional on data $\mathbf{Y}$:

$$P_{\theta_0}(\mathbf{X}|\mathbf{Y}) = P_{\theta_0}(\mathbf{Y}|\mathbf{X})P_{\theta_0}(\mathbf{X})/P_{\theta_0}(\mathbf{Y}) \tag{4}$$

Only then will the sampling distribution mimic the summand $P_\theta(\mathbf{Y},\mathbf{X})$ sufficiently closely for MC estimation to be feasible. Unfortunately, the normalizing factor $P_{\theta_0}(\mathbf{Y})$ in (4) is precisely the uncomputable likelihood $L(\theta_0)$. There are two ways in which sampling from (4) may be closely approximated. One is to sample from a closely related distribution and reweight the samples appropriately. This is the basis of the method of sequential imputation of Kong et al. (1992). However, where the space of $\mathbf{X}$ values is very large, where there is much missing data on the pedigree, or where the pedigree is complex, any feasible sequential imputation method may produce importance sampling weights ranging over many orders of magnitude, indicating that the imputation distribution is not close enough to (4). Moreover, in estimating a likelihood ratio, or any expectation with respect to the distribution (4), such as those required in implementing a Monte Carlo EM (MCEM) algorithm (Guo & Thompson, 1992), the estimator takes the form of a ratio estimator, with a separate factor providing the estimate of the normalizing factor $P_{\theta_0}(\mathbf{Y})$.

The second approach is that of Markov chain Monte Carlo (MCMC); which is designed to produce realizations from a probability distribution, such as (4), known only up to an unknown normalizing constant, such as $P_{\theta_0}(\mathbf{Y})$. The realizations are dependent; while obviously this does not invalidate the MC estimator (2), it greatly complicates the assessment of the properties of the estimator. However, in estimating expectations, the unknown normalizing factor no longer requires separate estimation. Instead, the method makes use of this factor and so avoids the problems of ratio estimators. In particular, likelihood analysis is based on computation of likelihood ratios, and samples from (4) provide direct estimates of these:

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{X},\mathbf{Y})}{P_{\theta_0}(\mathbf{X},\mathbf{Y})} \mid \mathbf{Y}\right). \tag{5}$$

That is, the MCMC realizations provide a simple direct estimate of the likelihood ratio function, $L(\theta)/L(\theta_0)$, as a function of θ (Thompson & Guo, 1991).

Although equation (5) is the basic core of Monte Carlo likelihood theory for latent-variable problems, it is only the beginning of a solution in the case of genetic analysis. Metropolis-Hastings algorithms are a very general class of MCMC algorithms. The Gibbs sampler (Gelfand & Smith, 1990) is one special case, which, as a local conditional updating sampler, is well suited to genetic analysis, because of the local dependence structure in a pedigree. However, the Gibbs sampler, tends to be very inefficient in sampling large complex spaces. On the other hand, other samplers are much harder to implement, due to the many constraints of Mendelian segregation and constraints imposed by marker data. Thus special MCMC algorithms must be developed for genetic analysis, that permit use of the local dependence structure, but can sample the huge space of possible genotypic configurations efficiently. A second class of problems lies in equation (5) itself. Efficiency in estimation requires that θ_0 be close to θ; that is, only a local likelihood function can be estimated. The development of sampling distributions which span the range of models θ of interest, and can be used in an importance sampling reweighting scheme to produce a likelihood ratio function over a wide model space are also a necessity (Geyer, 1991b). Improved MCMC samplers and improved MC likelihood estimators are the two theoretical thrusts of our current and proposed research.

3. Progress Report

3.1. Personnel on the current project

Name	Title	Date of birth	SSN	Dates of service	% appt
Faculty					
Thompson, Elizabeth A.	Professor	5/22/49	529-23-7846	9/1/91-8/31/95	27%
Wijsman, Ellen M.	Res. Assoc. Professor	4/19/54	346-50-7551	9/1/91-8/31/95	25%
Geyer, Charles J.	Asst. Prof. (visiting)	4/29/46	204-42-8556	6/16-8/15/92	25%
				6/16-8/15/93	20%
Graduate Students					
Lin, Shili	Res. Asst.	1/13/63	299-84-8286	9/16/91-6/15/93	48%
	Postdoctoral			6/16-8/15/93	50%
Olshen, Adam	Res. Asst.	09/21/68	549-94-3484	9/1/92-6/15/93	50%
Li, Hongzhe	Res. Asst.	10/11/67	516-19-8611	9/16/93-8/15/95	48%
Snow, Gregory	Res. Asst.	5/2/69	529-21-2758	9/16/94- 8/31/95	50%
Payne, Hannah	Hourly	3/27/69	271-72-8227	6/16-8/15/92	50%
Staff					
Poon, Simon	Syst. Prog. III	11/12/55	539-74-6012	1/1/92-10/1/93	11%
Khalil, Alexandre	Syst. Prog.	11/21/64	645-22-4118	3/16-8/31/94	19%
Godden, Jeff	Syst. Prog. III	10/18/57	531-58-6544	9/1/94-8/31/95	15%

3.2. Summary of previous specific aims

The long-term objective is to develop techniques for the analysis of the genetic basis of complex familial traits—traits, such as coronary artery disease, that are influenced by environmental factors as well as by genes at several loci. Using complex genetic models we proposed to develop Monte Carlo methods of likelihood analysis of phenotypic data observed on members of a pedigree.

The specific objectives were:

1. To develop Monte Carlo methods for the computation of likelihoods on extended and/or complex pedigrees: (a) for multilocus Mendelian models on extended complex pedigrees, (b) for polygenic models with multiple random effects for quantitative traits, (c) for threshold models with multiple thresholds, for rank-ordered categorical traits, and (d) for combining (a) - (c) to compute likelihoods for complex traits under mixed models, including linkage analysis with marker loci.
2. To develop methods for maximizing likelihoods and fitting models for complex genetic traits, using the same sequence of component models as in specific aims #1(a) - #1(d).
3. Evaluating the procedures of specific aims #1 and #2 by: (a) analyzing ancestry of multilocus haplotypes in the large complex genealogies of a genetic isolate, (b) analyzing Apolipoprotein B levels, and other quantitative measurements, both separately and jointly, under multiple-effects polygenic models, (c) analyzing Alzheimer's disease data and low-density lipoprotein measures and undertaking simulation studies to assess the performance of multiple threshold models for rank-ordered categorical traits, (d) segregation and linkage analysis of Apolipoprotein B levels, and other heart disease traits, incorporating genetic effects due to other loci, under complex models for data

on extended pedigrees, and (e) analysis of data simulated under complex models, involving both polygenic and Mendelian effects, and associated with linked markers.
4. To develop a coherent set of computer programs, within a single pedigree analysis computing environment, for items #1, #2 and #3 above.
5. To investigate the statistical power of studies combining segregation and linkage analysis in the analysis of complex genetic traits, and, to the extent that time allows, to investigate efficiencies of alternative sampling designs and pedigree structures.

The above specific aims were planned as a 5-year program. Funding was awarded with an administrative cut to 4 years, and an additional 17% administrative cut in each funded year. In the first 3 years we have substantially addressed goals 1-4, with the exception of threshold models (see below). By the end of the 4-year award, we shall have made additional progress, and commenced work on goal 5.

3.3. The mixed model; [1, 2, 3, 7, 8, 11, 12, 17].

(In this Progress Report references to papers resulting from this research, listed in section 3.10 are given in square brackets. Other references are cited in standard form and are listed in the Literature Cited section, as are all references cited in other sections of this proposal.)

An early focus of current research has been the mixed model for a quantitative genetic trait depending additively on major gene and variance component effects. Using the Gibbs sampler to sample polygenic effects and major genotypes, we developed Monte Carlo EM (MCEM) algorithms to find parameter estimates in the vicinity of the maximum likelihood estimate [1, 3]. These were then refined though Monte Carlo likelihood ratio estimation (MCLR, [11]) to explore the local likelihood surface. Similar methods were used to incorporate marker information in a joint segregation and linkage analysis with a diallelic marker [2].

The simple MCLR method does not provide adequate efficiency for estimating log-likelihood differences between disparate hypotheses; the mixtures method of Geyer (1991b) has been implemented and used to resolve this problem [12]. Other gains in efficiency are achieved by partial exact integration, either over some part of the pedigree (Thompson, 1991), or of some of the latent variables [7]. Such methods do not automatically lead to increased computational efficiency, and require theoretical analysis [7]. An alternative approach is a reparametrization of the model, leading to a rescaling of the latent variables, [8]. This technique can avoid degeneracies that complicate Monte Carlo likelihood comparisons of strictly nested hypotheses, and, combined with other techniques, has provided efficient methods for likelihood analysis of the classical mixed model.

For very large complex pedigrees, in which rare genes (or multilocus haplotypes) are segregating, practical MCMC samplers are hard to construct and harder to validate. One method that has proven successful on a 5000-member Hutterite pedigree in which cystic fibrosis alleles are segregating is a new MCMC method known as simulated tempering [17]. A component of our new proposal focuses on using this method in other problems which are intractable due to involving multiple tightly linked loci rather than due to pedigree size and complexity.

3.4. Linkage analysis and Monte Carlo Lod scores; [4, 5, 6, 12, 13, 19, 20, 21, 22]

Analysis of the mixed model is only a component step in the development of this methodology; for linkage analysis additional new methods are necessary. For multiallelic markers the Gibbs sampler may not be irreducible; therefore, the distribution of the Markov chain realizations is not guaranteed to converge to the correct equilibrium distribution. There are many ways to ensure theoretical irreducibility of the Markov chain; for example, an occasional Metropolis step that permits transition to any genotypic configuration will achieve this. However, such methods are seldom of practical use. Achieving a sampler that will explore the space of feasible genotypic configurations efficiently requires far more than simple irreducibility [20]. For multiple markers, the problems of theoretical irreducibility are unchanged, but practical irreducibility is an increasing challenge.

Lin addressed the problems of MCMC methods for linkage analysis likelihoods in her 1993 Ph.D. thesis [13]. She used the coupled chains method of Geyer (1991a), proposing irreducible coupled chain models that are very close to the true genetic model, and hence couple efficiently [4], together with a form of "heating" of the transition matrices of coupled chains to facilitate more efficient exploration of the space of possible genotypic configurations [6]. These methods were applied in practical analyses of data on extended Alzheimer's disease and coronary artery disease pedigrees [19, 12]. These methods can be applied without the need for prior analysis to determine the data observations causing reducibility of the simple Gibbs sampler. However, where such causes can be identified, improved efficiency is possible. In [5], an algorithm is given to determine these phenotypic events in any specific pedigree, and this is made use of in Lin's more recent work [21, 22].

3.5. Homozygosity mapping; [9], [10], [26].

Although not a specific aim of our original proposal, our approaches to multilocus linkage analysis led us to alternative MCMC methods for multilocus Mendelian traits and markers observed on affected relatives, or single inbred affected individuals [9, 10, 26]. This arose in considering data on Werner's syndrome pedigrees, a rare recessive mapped to a region of chromosome 8 (Goto et al., 1992). Once linkage was confirmed, multilocus linkage analysis was again needed to help pinpoint the gene (Schellenberg et al., 1992b). On a single small but complex pedigree exact analysis was infeasible for more than three marker loci with three alleles each. The computer program was halted after failing to produce even one lod score in 3 days of computing on a VAXstation 3100/76, even with the *FASTLINK* program.

While the methods of Lin et al. [4, 5, 6] could be used to develop efficient MCMC methods in this case, an alternative approach is more effective. Although, in genetic analysis, the genotypes (pairs of multilocus haplotypes) of individuals are a natural choice for the latent variables of interest, this is not the only possibility. Where only one (or a few) individual(s) are observed on each pedigree, it is more effective to use the indicators of grandparental origins of genes at each locus in each segregation as the latent variables. This very substantially reduces the space to be sampled, but the necessary computations for sampling and Monte Carlo likelihood estimation can still be carried out. Using this approach, we have estimated lod scores [9] and posterior probabilities of gene identity by descent [10, 26]. This also is a research aspect to be pursued in future work (section 4.3).

3.6. Complex traits; ascertainment and age of onset [14, 18].

Complexity of a genetic trait can have many features, but one of increasing importance as more complex traits are analyzed is the censoring of data due to delayed onset. Research Assistant Hongzhe Li is addressing this problem, developing Monte Carlo methods for maximization of likelihoods using models for age-of-onset that incorporate a Mendelian gene in addition to other environmental and genetic risk factors [18]. He has implemented methods for estimating the parameters of a survival analysis model for age-of-onset, using a Cox model with a latent Mendelian genetic risk factor. Marker information, and data on quantitative risk factors controlled by the same (unobservable) genes, are also a part of the modeling and estimation process.

In estimation of parameters of genetic models, ascertainment remains a complex issue [14]. We have briefly considered two aspects of ascertainment in our Monte Carlo analyses, but have not made this a major thrust of our work. First, single ascertainment conditional on a specific phenotype, or on the phenotypes of a small specific set of individuals, requires estimation of the probability of that event. If this estimation is done by Monte Carlo, the ascertainment-corrected likelihood becomes a ratio estimate, with consequent statistical problems. However, in many cases it may be possible to compute deterministically the probability of the ascertainment event. In this case, whatever Monte Carlo method is used to give estimates of the unconditioned likelihood is equally applicable to the ascertainment-corrected form. Secondly, in Li's work on age-of-onset, the problem of ascertainment is being addressed in the context of the sampling of a number of small pedigrees through affected probands. For a heterogeneous trait, such as Alzheimer's disease, ascertainment is additionally complicated by the fact that affected probands may not carry the Mendelian disease allele.

3.7. Practical analyses and method comparison

(i) Multipoint linkage analysis; [19, 23].

We have performed two comparative studies of multipoint linkage analyses performed with MCMC methods as compared to exact methods, with computations for exact methods done with the LINKAGE package (Lathrop et al 1984). Both involved a linkage analysis with multi-allelic marker loci, performed on a series of pedigrees segregating for a dominant disease, cyclic neutropenia, available through Dr. Wijsman's participation in NIH DK 18951. In the first comparison, the "heated chain" method [4] was used to sample from the possible configuration of genotypes for a problem with a total of 3 loci, one of which had 3 alleles. The agreement between lod scores obtained with MCMC and LINKAGE was excellent. In a more recent comparison, there were 5 loci, with 3-5 alleles each, and with movement between possible configurations of genotypes done by means of first identifying all "islands" in each pedigree with methods described in [5] and then using "hopping" procedures [22] to move between islands. This procedure also gives good results [23]; see figure 1. Not only were the numerical results similar with MCMC and exact methods, but for this problem, the MCMC analysis was significantly faster: the MCMC methods required about 2.5 hours of CPU time on a SPARC10 computer, which is about 15 times faster than an exact analysis, after adjusting for the different computers used.

(ii) Complex segregation analysis; [12, 24, 25].

A comparison of MCEM methods [3] with approximation methods obtained with PAP (Hasstedt & Cartwright, 1981) for estimating parameters of the mixed model indicated that MCEM methods gave parameter estimates which are closer to the "truth" than does PAP (Olshen and

Wijsman, [24]). Data were simulated for a quantitative trait under the mixed model. Three different sets of generating parameter values were generated for each of 100 data sets, corresponding to three different heritabilities for the polygenic component, chosen to represent traits with low, moderate, and high heritabilities. All simulations and analyses were performed on the same large 232-member pedigree structure used in [12]. By creating simulated data, MCEM methods could be compared with the approximation methods used by PAP and also could be compared to the known simulation parameters. The variance in parameters estimates was always less when estimation was done with the MCEM *mixd* program as compared with PAP. In both cases, most parameter estimates seemed to be unbiased. The exception was the overestimation of the heritability by PAP, leading to underestimation of the environmental variance. For most of the parameters, the standard deviation in the estimates obtained for MCEM methods was about half that obtained with PAP. For the additive genetic variance component, however, the ratio of the standard deviations was inversely proportional to the heritability: for a heritability of 0.2, the estimates obtained with PAP had 6 times the standard deviation of estimates obtained with MCEM.

The starting points used in MCEM analysis had no strong effects on the resulting parameter estimates, with one exception. Further exploration of estimation of the additive genetic variance indicated that under MCEM this parameter estimate stayed close to its starting value. While the estimate moves in the correct direction, it does so very slowly, resulting in final estimates (after practical computing time) which are correlated with the starting values. The same is not true for the environmental variance component or any of the major gene parameters: here the final values were uncorrelated with starting values. Other parameters of the analysis investigated, including the length of the burn-in period and the number of EM-steps used in the analysis, had no apparent effect on the resulting parameter estimates.

Segregation analysis with MCEM methods was also performed on a dataset of 3 large pedigrees segregating for unusual lipid phenotypes. These pedigrees are available through Dr. Wijsman's participation in NIH HL 30086. The results of these segregation analyses were similar to analyses obtained previously with PAP.

3.8. Software implementation and development; [15], [16].

In accordance with specific aim #4, we have produced an integrated body of software in ANSII standard C, to carry out the analyses described above. The software has been documented [15], and is available to interested researchers by anonymous ftp. The software was used as a part of a short course on Monte Carlo methods in Genetic Analysis given in May 1994 [16]. Following this several participants have copied the software, and some have installed it. However, the focus has not been on making the software user friendly, and we have not had resources for supporting the software, or answering questions.

In writing programs for the component parts of these analyses, the aim has been not only to build a single integrated package, using the same data file formats, and basic pedigree input and computation subroutines, but also to ensure that the software is portable, at least to other Unix systems. We routinely move programs between DECstations, DEC Alphas, and SUN Sparc2 and Sparc10 workstations, since those machine are in regular use by students in Statistics and Biostatistics. The software for the course [16] was transferred without difficulty to Hewlett Packard Unix workstations. Also since one investigator (Wijsman) works on a VMS VaxStation, we have solved problems of portability to that machine: however, this is less routine. In developing the software further, issues of portability beyond the Unix operating system will have to be addressed.

3.9. Modification of specific aims.

We will have essentially completed aims #1 - #4 of the original 5-year proposal by the end of the 4 years of this award, and will have begun to address the issues of design and sampling that were specific aim #5. Some assessment and comparison of methods remain to be done, but we now have extensive results on the mixed model, and comparisons of linkage analysis methods are in progress. We have begun to address other complexities such as the use of age-of-onset data. In year 4, during the PI's extended visit to McGill University, the extant methods will be applied to data on heart disease indicators in the Hutterite population, in collaboration with Dr. Ken Morgan who has been a consultant on the grant. The performance of linkage methods, and of the joint segregation and linkage programs will be assessed also by investigator Dr. Wijsman and her RA, using Seattle heart disease data available to her. The very rapid evolution of Monte Carlo methodology (both in general, and in genetic analysis) has made it almost impossible for testing of methods to keep pace with new improvements.

As already stated we decided not to pursue MCMC methods for ordered categorical traits; these are being actively investigated in the area of animal breeding genetics (Hoeschele et al., 1987; Sorensen et al., 1994). Instead we have focused on the censoring problem of age-of-onset, which has some similar features, and is more immediately relevant in the analysis of complex traits such as Alzheimer's disease and NIDDM. Additionally our recent work has placed more emphasis on linkage analysis than was in the original proposal; only by using marker data in conjunction with trait data, quantitative risk indicators, and all other available relevant information, will the genetic basis of complex traits be resolved.

3.10. Papers and reports resulting from this research since last competitive review.

[Note: * denotes materials included as appendix.]

(i) Papers published or in press.

1. Guo, S-W. and Thompson, E. A. (1991). Monte Carlo EM for the estimation of multiple random effects models on large pedigrees. *I.M.A. Journal of Mathematics Applied in Medicine & Biology* **8:** 171–189.

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

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

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

5.* Lin, S., Thompson, E. A., and Wijsman, E. M. (1994a) Finding non-communicating sets for Markov chain Monte Carlo estimations on pedigrees. *American Journal of Human Genetics* **54:** 695–704.

6. Lin, S., Thompson, E. A., and Wijsman, E. M. (1994b). A faster mixing algorithm for Hastings-Metropolis updates on complex pedigrees. *Annals of Human Genetics;* in press.

7. Thompson, E. A. (1994a). Monte Carlo likelihood in genetic analysis. In *Probability, Statistics, Optimization: a tribute to Peter Whittle*, F.P. Kelly (ed.). Chapter 21: Pp. 281–293. Wiley; New York.

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

9. Thompson, E. A. (1994c). Monte Carlo likelihood in linkage analysis. *Statistical Science* **9:** in press.

10.* Thompson, E. A. (1994d). Monte Carlo estimation of multilocus autozygosity probabilities. *Proceedings of the 1994 Interface conference* SAS: Cary, NC. In press.

11.* Thompson, E. A. and Guo, S-W. (1991). Evaluation of likelihood ratios for complex genetic models. *I.M.A. Journal of Mathematics Applied in Medicine & Biology* **8:** 149–169.

12.* Thompson, E. A., Lin, S., Olshen, A. B. and Wijsman, E.M. (1993). Monte Carlo analysis on a large pedigree. *Genetic Analysis Workshop 8. Genetic Epidemilogy;* **10:** 677–682.

(ii) Theses, Technical reports

13. Lin, S. (1993) *Monte Carlo methods for the estimation of probabilities on complex structures.* Ph.D. Thesis, University of Washington.

14. Thompson, E. A. (1993) *Sampling and ascertainment in genetic epidemiology; a tutorial review.* Technical Report No. 243, Dept. Statistics, University of Washington.

15.* Thompson, E. A. (1994e) *Monte Carlo Programs for Pedigree Analysis: 1990-93.* Technical Report No. 267, Dept. Statistics, University of Washington.

16. Thompson, E. A. (1994f) *Monte Carlo methods for Genetic Analysis.* A course of 12 lectures given at Wageningen Agricultural University, May 1994.

(iii) Abstracts and submitted papers

17. Geyer, C. J. and Thompson, E. A. (1994). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Revised version: submitted.*

18. Li, H. and Thompson, E.A. (1994). Modeling age-of-onset; Cox model with latent major gene effects. *American Journal of Human Genetics* **55** A156.

19. Lin, S., Thompson, E.A. and Wijsman, E.M. (1992) Multiple alleles and Markov chain Monte Carlo methods in genetic analysis. *American Journal of Human Genetics*, **51:** A14.

20. Lin, S. and Thompson E.A. (1993) Discussion contribution to RSS papers on Markov chain Monte Carlo. *J. Roy. Statist. Soc. (B)*, **55:** 82–83.

21. Lin, S. (1994a) Monte Carlo multipoint linkage analysis. *Submitted.*

22. Lin, S. (1994b). A scheme for constructing an irreducible Markov chain for pedigree data. *Biometrics:* under revision.

23. Lin, S. and Wijsman, E. M. (1994). Monte Carlo multipoint linkage analysis. *American Journal of Human Genetics* **55** A40.

24. Olshen, A. B. and Wijsman, E. M. (1994). PAP vs. Monte Carlo: fitting the mixed model on a large pedigree. *Submitted.*

25. Olshen, A. and Wijsman, E.M. (1994). PAP vs. Monte Carlo: fitting the mixed model on a large pedigree. *American Journal of Human Genetics* **55** A160.

26. Thompson, E.A. and Wijsman, E.M. (1994). Multilocus homozygosity mapping and autozygosity patterns. *American Journal of Human Genetics* **55** A167.

4: Research Design and Methods

4.1: Monte Carlo estimation of linkage likelihoods

Even Mendelian trait loci can be hard to find, if there is heterogeneity, incomplete penetrance, delayed onset, or other standard trait complexities. Once linkage has been established, often using single genetic markers, a location score curve, mapping the gene location against a map of multiple markers in the region is required. Monte Carlo likelihood provides a method for estimating location score curves, but the MC methods proposed so far will not resolve all difficulties. The method of Kong et al. (1992) can be used only if the pedigree is not complex, and may perform poorly if there are too many missing data. The MCMC methods of Lange & Sobel (1991) and of Lin (1994), remain to be extensively tested, but may not be practical with multiple tightly linked polymorphic markers, since the space of multilocus genotype configurations on an extended pedigree is both huge and tightly constrained. Moreover, much more than a single location score curve is needed. Multilocus linkage likelihoods can be sensitive to errors in marker locations or marker allele frequencies; the ability to consider varying these is essential. Additionally, there is sensitivity to genetic heterogeneity or to a second locus contributing to a quantitative trait, creating apparent recombinants with markers. The ability to consider various complex trait models should also be available.

One general MCMC method that offers significant promise for sampling very large and constrained spaces is simulated tempering (Marinari & Parisi, 1992; Geyer & Thompson, 1994). The idea is that although the distribution of interest may be almost impossible to sample from – a very "cold" distribution – it is one of a class of distributions indexed by a parameter ("temperature") – and the hottest member of the class is one that is easily sampled from, obtaining independent realizations. For each distribution of the class, an MCMC method such as Gibbs sampling is employed, while the entire process involves sampling over the members of the class using a Metropolis sampling scheme. Theory shows that each time the process enters the "cold" distribution of interest, it does so with a realization from the appropriate equilibrium distribution, while the independent realizations in the "hot" distribution provide regeneration points of the process, that permit rigorous assessment of precision of the Monte Carlo experiment (Geyer & Thompson, 1994). The method differs from the Metropolis-coupled chains of Geyer (1991a), in that only a single process is realized, and that regeneration provides for independent blocks of realizations. It differs from a Bayesian sampling scheme (Stephens and Smith, 1993), in that the "prior weights" accorded to members of the family do not reflect any prior probability, but are chosen in such a way as to maximize efficiency of the Monte Carlo experiment. The method has been used by Geyer & Thompson (1994), to address ancestry of rare recessive alleles in a complex 5000-member pedigree.

The simulated tempering method is well suited to multilocus linkage analysis, where a trait has been linked to a chromosomal region, but its position within the region is uncertain. With multiple closely linked markers, exact computational methods are often not feasible, and current MCMC methods also have great difficulty sampling the complete space. Our specific proposal for a simulated tempering family of distributions is as follows. The class of distributions would be formed by stretching the marker map. The extreme hot distribution will be of independently segregating loci; genotypes will be sampled conditional on the marker types at that locus by imputation methods (Ploughman & Boehnke, 1989; Ott, 1989; Kong et al. 1992). Successively colder distributions will be formed by shrinking the total map; the coldest distribution will be the true marker map. With the trait locus at different hypothesized positions, the genetic map of marker and trait loci will be stretched ("heated") and squeezed ("cooled"), to provide the indexed family of distributions. By treating the map as a "concertina", squeezing it down to the true map of tightly linked markers, we can obtain efficient MCMC estimates of linkage likelihood curves for the position of a trait locus within this true map.

Use of the multilocus haplotype pairs of individuals as the latent variables sampled in a Monte Carlo analysis automatically produces joint posterior probabilities of these haplotypes. Such posterior probabilities, conditional upon all the data observed on a pedigree are of interest for further analysis. They identify the likely recombination events, and hence are of key importance in identifying individuals and markers that would contribute maximal additional information. They could significantly reduce the burden of additional marker typing and retyping. They also serve to identify uncertainties, and hence possible typing or pedigree errors (Ott, 1993; Kass et al., 1994), or possible genetic heterogeneities. In linkage analysis with single markers, heterogeneity will cause the trait locus to have apparently increased recombination with the linked marker locus. When the trait locus is confined by the intervals of a genetic map of multiple markers, such heterogeneity may cause the locus to "disappear" from an interval due to recombinations in unlinked families with adjacent tightly linked markers. By considering the location score curves and *a posteriori* likely recombination events, as the genetic map is stretched from the true map to the extreme of independent marker loci, we can pinpoint the occurrence of unlinked pedigrees, and of phenocopies within a pedigree. Thence, we will build better trait models that allow for heterogeneity or additional trait loci.

We propose to implement simulated tempering as an MCMC method for estimation of location score curves for data on extended pedigrees. We propose to test the method first for simple genetic traits, including extended pedigrees segregating for a dominant trait, cyclic neutropenia, and complex pedigrees including inbred individuals affected with the recessive disorder, Werner's Syndrome, both of which are available to us through Dr. Wijsman's collaboration on applied projects on these diseases (See "Resources and environment" for details). We propose then to test the method on a collection of extended Alzheimer's disease pedigrees, using multiple markers on chromosome 14, to which the disease maps in some families. We propose also to include unlinked families in our analyses, to investigate the power of our method to accommodate and account for heterogeneity.

We propose also to extend the methods to the case of quantitative traits, using the trait models implemented in our current software. We will apply the method to quantitative lipid phenotypes available to us through the LPPG program (see "Resources and Environment"). We propose then to move to other complex traits, as detailed in the next section.

4.2: Linkage analysis for complex traits.

In addition to the difficulties of heterogeneity and incomplete penetrance that complicate analysis even for Mendelian traits, many human diseases with a genetic component exhibit other complexities. In order to resolve these traits, use must be made of all available information, such as data on linked markers, variation in expression or age-of-inset, or known quantitative risk factors.

Once initial evidence of linkage has been found, joint linkage and segregation analysis (Guo & Thompson, 1992) can help to refine the trait model. Monte Carlo methods for combined segregation and linkage analysis have been implemented, but remain to be more extensively tested. As yet only a single quantitative trait and one diallelic marker are considered. The possibility of a second trait locus, possibly also with a linked marker, will be implemented. Also, the multiple alleles techniques of sections 3.4 and 3.7 will be integrated into the method for joint segregation and linkage analysis. Generally, more complexities of the trait model will be integrated into the *linkreg* program for joint linkage and segregation analysis, including delayed age-of-onset.

Age of onset is another aspect of complexity. A number of methods for addressing this question have been proposed (Elston & George, 1989; Abel & Bonney, 1990). One approach much used in the recent Biostatistics literature is that of the Cox proportional hazards model, with patterns of dependence introduced by latent random effects or frailties (Clayton & Cuzick, 1985; Self & Prentice, 1985). These models have also been used in genetic epidemiology (Mack et al., 1990), but frailty models cannot mimic the patterns of joint dependence implied by Mendelian segregation in an extended pedigree. We have preferred to model censored age-of-onset data using a semiparametric Cox model with a latent Mendelian genotype (Li & Thompson, 1994). We model the rate of onset for an unaffected individual, of age t, covariate information Z and trait genotype G_i as

$$\lambda(t \mid Z, G_i) \; = \; \lambda_0(t) \exp(\beta Z \; + \; \mu_i), \tag{6}$$

where $\lambda_0(t)$ is the baseline hazard which need not be specified. MCMC provides a method for sampling the latent genotypes, $\mathbf{G}$, conditional on data on affectation status, age-of-onset (if affected), and other observable covariates. Hence the risks, μ_i, attributable to the putative Mendelian gene can be estimated, together with the usual risk parameters, β, for observed covariates. The method can be applied to a large complex pedigree, or to a collection of smaller pedigrees; we have tested it on data on some extended early-onset Alzheimer's disease pedigrees, and also on simulated data. We propose to extend this class of models. In particular, we shall implement and test the case where data on a putatively linked marker is also available, to extend joint segregation and linkage analysis to these models. There are numerous marker loci available for such analyses on Alzheimer's disease pedigrees, also available through Dr. Wijsman's collaborations.

For many complex diseases, there are known quantitative risk factors also under genetic control, such as Apolipoprotein B levels for coronary artery disease (Chait et al., 1980) or insulin resistance for NIDDM (Saad et al., 1989; Prochazka et al., 1993). A reasonable model is that the disease genotype itself influences these quantitative traits. Such quantitative indicators can often only be measured on unaffected individuals, but under a model in which the same gene influences both disease and trait, MCMC permits sampling of these latent genotypes, and hence MC estimation of likelihood ratios and of model parameters. Thence, quantitative trait data can be incorporated into a linkage analysis, and, conversely, the frequency distributions of sampled latent individual genotypes provide risk estimates for unaffected individuals.

In analyzing a collection of pedigrees, where estimation of genetic parameters is desired, ascertainment is an issue. Examples include early-onset Alzheimer's disease, coronary artery disease, or NIDDM. Although MCMC provides no general solution, it does permit sampling of latent genotypes, conditional on any subset of genotypes, and hence the construction of conditional likelihood functions. In any EM algorithm a key function is the conditional expected complete-data log-likelihood:

$$\mathrm{E}_{\theta_0}(\ \log\ P_\theta(\mathbf{Y},\mathbf{X})\mid\mathbf{Y}) \tag{7}$$

where $\mathbf{Y}$ are the data (phenotype, age-of-onset, marker and covariate) and $\mathbf{X}$ the latent variables (genotypes). The expectation is considered as a function of θ, the parameters of the genetic model, and expectations are computed at θ_0, the current parameter estimates. Even for the complex likelihood functions provided by censored age-of-onset data, it is possible to sample, and to estimate the expectation (7), conditioning on proband genotypes. Ideally, conditioning on proband phenotype is desired; this is possible, but leads to less satisfactory ratio estimators of the conditional likelihood function. Where trait locus genotypes of a proband are correctly imputed, as will generally be the case for early-onset Alzheimer's disease for example, conditioning on the genotype will not invalidate inferences, and will lose little information (Ewens & Shute, 1986). A preliminary version of this method has proven successful in the analysis of simulated data on a collection of small pedigrees, each ascertained for an early-onset case. We propose to develop further MC procedures for ascertainment corrections, and to investigate their properties.

4.3: Gene identity and affected pedigree member methods.

In any linkage analysis for a disease, the majority of information is derived from affected individuals. For a trait with incomplete penetrance, delayed onset, and uncertain model, linkage analysis based only on data on affected individuals is one way forward (Weeks & Lange, 1988). However, likelihood computation on the basis of only affected individuals in an extended pedigree is often infeasible: there are too many "missing data". The space of all possible multilocus genotype configurations on all relevant individuals is huge; MCMC methods which use genotypes as latent variables are doomed to failure. Multilocus linkage likelihoods are, however, required, once initial evidence of linkage is found. By considering all affected individuals jointly, much power is gained over a pairwise analysis of individuals. By considering several marker loci jointly, again much power will be gained over a pairwise analysis of loci.

One solution is to reduce the space of latent variables. In order to implement any MCMC method it is necessary that

$$P_\theta(\mathbf{Y},\mathbf{X}) \;=\; P_\theta(\mathbf{Y}\mid\mathbf{X})\ P_\theta(\mathbf{X}) \tag{8}$$

can be computed rapidly, where $\mathbf{Y}$ are the data on which the likelihood is to be based, $\mathbf{X}$ the latent variables, and θ the model used in the MCMC simulation. While this is true if $\mathbf{X}$ is the complete genotypic configuration on a pedigree, it is also true if $\mathbf{X}$ is the set of grandparental origin indicators for each segregation and each locus, provided data are considered for only a few individuals in a pedigree (Thompson, 1994a).

The extreme case of the above considerations arises in homozygosity mapping, where data on only one affected inbred individual in each pedigree is used to map a rare recessive trait (Smith, 1953; Lander & Botstein, 1987). As a preliminary exploration we have implemented MCMC algorithms and lod score estimation, using segregation indicators as the latent variables. We have also investigated the method using data on Werner's syndrome, a simple rare recessive of premature aging linked to chromosome 8 markers (Goto et al., 1992). Multiple chromosome 8 markers are available for a collection of inbred affected individuals (Schellenberg et al., 1992b), and multipoint analysis is desired. Although most of the affected individuals are offspring of first-cousin marriages, or other single loop relationships, there are some complex pedigrees for which exact likelihood analysis is impossible with more than three loci. Although containing only 15 individuals, one of these pedigrees (figure 2) defeats standard linkage analysis programs, due to its three interlocking loops (section 3.5; Alejandro Schäffer, personal communication). It also defeats the new exact computation algorithms for multilocus IBD (Guo, 1994) and homozygosity mapping (Kruglyak & Lander, 1994) due to the 18 relevant segregations per locus (Sun Wei Guo, Leonid Kruglyak; personal communications). MCMC methods can provide accurate location score curves in reasonable computing time (Thompson, 1994a,b).

We propose to extend this method to allow analysis of small sets of affected relatives in extended, possibly complex, pedigrees, and so address joint affected-pedigree-member analysis for complex traits, using multiple markers. Although the space of segregation indicators is much easier to sample than the space of individual genotypes, when markers are very closely linked exploration of even this space may be hard. The method of simulated tempering (section 4.1), the temperature parameter indexing the map length of the region, can be applied equally to this MCMC process also, and will help to resolve any such difficulties. We will test the method using those Werner's syndrome pedigrees in which there are several affected individuals, and then apply it also to other pedigrees of more genetically complex traits (see 4.2), using only the data on affected individuals. We will compare the computational and statistical efficiency of using only affected pedigree members with using phenotypic and marker information on all available individuals. Where exact computation is infeasible, Monte Carlo comparisons will be made, using the methods of this section for the affected relative data. For the full data, we will use the Monte Carlo methods of section 4.1, Lin (1994), Lange & Sobel (1991), or Kong et al. (1992) choosing whichever method is best suited to the pedigree structure, marker density and extent of missing information in a particular problem.

As for other MC likelihood methods, the approach also provides posterior probabilities of the latent variables. Here these latent variables are the segregation indicators which determine both recombination events and the multilocus patterns of gene identity by descent (GIBD) among the affected individuals (Thompson, 1994b). These patterns are of interest in interpreting anomalies in the data. For example, in the case of homozygosity mapping, each inbred affected individual has high posterior probability of being IBD at the disease locus, but the probability is small that all such individuals are IBD in a set of over 30 pedigrees. Due to the multiple ways in which an inbred individual can be IBD, or not (NIBD) in even a single-loop pedigree, even prior probabilities of GIBD are highly non-Markov along a chromosome, even in the absence of interference (Thompson, 1994b). Small patches of NIBD intersperse IBD regions, which tend to be clumped (Donnelly, 1983). In considering the occasional marker heterozygosities that occur in otherwise homozygous regions of inbred affected individuals, this clumping has significant impact on the posterior probabilities of recombination events in the pedigree.

Another source of occasional marker heterozygosity is mutation. For microsatellite CA-repeat markers mutation rates may be as high as 0.8%, although on average they appear to be about 0.1% (Weber & Wong, 1993). The chance of a mutation at some marker locus, in one of the 10 to 15 relevant segregations on even a small pedigree, is therefore substantial, and excluding mutation from the model can lead to aberrant conclusions. This is particularly so when only data on affected individuals are analyzed; the marker genotypes that would enable us to detect mutations directly are not observed. Incorporating mutation into deterministic likelihood computations is normally infeasible, due to the huge number of possible marker genotypes which must then be considered. With MC estimation, incorporation is at least feasible. However, to be practical, it will be important that results are not dependent on the details of a complex mutation model for these markers, nor on the precise value of the mutation rate. We propose to investigate the practicality and efficiency of a model including mutation at microsatellite marker loci.

4.4: Gene ordering and genome comparison.

It is not only in the linkage analysis of trait loci that Monte Carlo methods can aid in likelihood analysis. MCMC methods are ideally suited to sampling highly structured discrete hypothesis spaces, as well as spaces of latent variables such as genotype configurations or segregation indicators. The Metropolis algorithm (Metropolis et al., 1953) or improved Hastings-Metropolis algorithms (Hastings, 1970) sample the elements of such spaces in proportion to their posterior probabilities. One class of problems arising at all mapping scales from 100 DNA bases to 10^7 bases is that of ordering. Combinatorial problems arise in inferring order whether considering linkage data (O'Connell et al., 1987), radiation hybrids (Cox et al., 1990), clone contig fingerprint data (Branscomb et al. 1990) or the STS content mapping of YAC contigs (Green & Olson, 1990). Methods proposed have restricted the orders considered using a variety of criteria. The only restriction Monte Carlo sampling imposes on the orders considered is a weighting dependent on the posterior probabilities of the orders themselves.

Branch-and-bound algorithms and simulated annealing have been used to explore the space of possible gene orders in linkage mapping (Thompson, 1987) and in physical mapping using radiation hybrids (Boehnke et al., 1991). Building on these methods, we propose to implement MCMC algorithms that will explore not only likelihoods conditional on a given ordering, but also alternative gene orders. As a first step, a Metropolis algorithm, proposing local changes, can be employed; a "simulated annealing without cooling". However, as for the problem of multilocus linkage analysis, it is likely that better MCMC methods will be needed. Lander and Green (1987) used segregation indicators in their deterministic computation of likelihoods for multilocus marker maps, but this computation was infeasible when there were missing data. By using Monte Carlo sampling of the latent segregation indicators, estimated likelihood ratios can compare alternative maps.

The algorithm of Lander and Green (1987) for maximizing multipoint linkage likelihoods is an EM algorithm (Dempster et al., 1977). The essence of such algorithms is that, as in previous sections, we have data $\mathbf{Y}$ and latent variables $\mathbf{X}$. The E-step of the algorithm requires computation of

$$\mathrm{E}_{\theta_0}(\ \log\ P_\theta(\mathbf{Y},\mathbf{X})\mid\mathbf{Y}) \tag{9}$$

at current parameter estimates θ_0, and the M-step then maximizes the function with respect to θ to provide new estimates of the parameters. Where this expectation is difficult or impossible to compute, but samples from the conditional distribution $P_{\theta_0}(\mathbf{X}\mid\mathbf{Y})$ are available, a Monte Carlo EM algorithm can be implemented, as we have done previously (Guo & Thompson, 1992). As in Monte Carlo likelihood, the choice of latent variables can have a strong impact on the feasibility of an EM or MCEM algorithm. Proposed co-investigiator Dr. Green, in his computational research under R01 HG 00774, has recently developed a two-phase EM algorithm, where two different levels of latent variables are used to increase computational efficiency. With this method, provided marker data are complete, 100 or more loci can be considered together. For trait loci, data are inherently incomplete, but Monte Carlo methods have the facility to accommodate this through imputation from the appropriate distribution. We propose combination of Green's two-phase EM algorithms, with MCMC sampling from the appropriate conditional distributions, to produce improved algorithms for the maximization of multilocus linkage likelihoods.

4.5: Genome comparison.

Closely related to multilocus linkage analysis, to locus ordering problems, and to affected pedigree member methods are questions of genome matching (Nelson et al., 1993). With the increasing density of genome information on individuals, comparison of the genomes of related affected individuals aids in the location of genes influencing complex traits. It has been suggested that such an approach obviates the need for linkage analysis, providing a powerful model-free method of locating trait genes. Feingold et al. (1993) have provided test statistics, and have considered the power of a likelihood ratio test. The potential utility of the method is unquestioned, but it is no more "model-free" than any other test procedure. Power is always an issue and the properties of the test procedure will always depend on the underlying genetic map and genetic basis of the trait. Computation of likelihoods, the probabilities of the data under alternative hypotheses, are still required.

Computation of probabilities of genome matching are most easily placed in the context of models of a continuous genome, rather than the discrete markers at which observations are made. Such models were first introduced by Fisher (1949), and were used by Donnelly (1983) to analyze sections of genome shared identical by descent (IBD) by a pair of relatives, and sections of parental genome present in a set of offspring. Sections of shared genome tend to cluster, interspersed by short lengths of "mismatch". Blossey (1993) used the Poisson clumping heuristic (Aldous, 1989) to investigate this clumping; for a model of discrete loci the same phenomenon was investigated by Thompson (1994b). Thomas et al. (1994) used a Poisson approximation for the total number of distinct segments shared IBD. However, in the analysis of densely packed markers, such an approximation is inadequate; the probability of non-IBD regions in a segment otherwise IBD can have a very significant impact on interpretation of data (Thompson, 1994b).

Several recent papers have considered exact computations of relevant probabilities. The methods of Feingold (1993) provide a test based on affected relative pairs of varying types. Guo (1994) can consider more than two individuals jointly, but only for small simple pedigrees, and for the state of all the considered individuals sharing a genome segment. However, it is important to consider subsets of the individuals, and not only the probability that all affected share a genome segment. If there is very high posterior probability that, for example, four out of five affected members of a pedigree share some section of genome, the region should not be ruled out. Phenocopies, heterogeneity, mutation, or typing error, could explain why the fifth individual shows a discrepancy. For maximal power one should sample relatives, to lessen chances of genetic heterogeneity for the trait, but one should sample remote relatives to lessen the parts of genome shared, other than those that contribute to the shared affected status (Thomas et al., 1994). However, inclusion of more remote relatives decreases the chance that all share the same genetic cause of their affected status, particularly for less rare and more complex traits.

For the case of several relatives jointly, in an extended or complex pedigree, probabilities of the various possible patterns of genome sharing cannot be computed exactly. As in the case of multilocus linkage analysis, the required probabilities are a sum over a large number of latent variables, the recombination points ("junctions": Fisher, 1949) in a continuous genome model. Here again MCMC methods of likelihood estimation are applicable, and provide a method to resample the junction points conditional on the data. Since these points are highly constrained by the data, it is likely that some method to enhance efficiency of the Monte Carlo sampling will be required, as in the case of multilocus linkage analysis. Simulated tempering, with the heating parameter again indexing the total map length, is again a possibility for implementation.

A successful MCMC method will produce a posterior probability curve of genome sharing, for each subset of the affected individuals, conditional upon marker data, at all points along the genome, and hence a direct visual indication of likely positions of genes influencing the trait.

4.6: Software development

Under the current research project, we have focused on producing an integrated body of software, but not on making it user-friendly. The programs have been documented, and are available by anonymous ftp. The documentation is also available by anonymous ftp as a PostScript file. However, the documentation focuses on only a few of the more fundamental programs and subroutine libraries, and even some of these are undergoing continuing development. The programs have been installed by several researchers at their own institutions, particularly following their use in a one-week course on Monte Carlo

Methods for Quantitative Genetic Analysis, but we have not encouraged wide use of the programs, and have not had the resources to offer any support or answers to questions.

Although all current programs use a common core of subroutine libraries for pedigree manipulation, genotype sampling, and shared component computations each developer has added (or substituted) routines specific to their project. Although a common data file format has been imposed for quantitative trait linkage and segregation analysis, the options required to accommodate the varied analyses under development have caused a proliferation of control options in the data files. Many problem-specific subroutines reside in the main program directories, and, although dependencies among programs and routines is clearly specified through the directory Makefiles, the system is not user-friendly.

Under the renewal we have the specific aim of producing software and distributing it to interested users. In the first instance, such users will be computationally experienced researchers who wish to compare methods with their own, or to perform specific data analyses for which other software has been found to be inadequate. The first aim on the software development front is to overhaul thoroughly the structure of our current Monte Carlo software collection. The second aim is to set up a collection of command files and online help files, so that users will no longer have to modify Makefiles for their purposes, and input and output files can be specified more flexibly. Thirdly, we will add a number of features, making the available options more compatible with other software, for easier comparison. For example, we will add the option of varying segregation parameters, τ, in the genotype sampling routines, for testing Mendelian segregation of putative genes underlying complex phenotypes (Elston et al., 1986). We shall also facilitate simultaneous analysis of a set of pedigrees: our MCEM programs still assume an input in the form of a single connected pedigree, although other programs have been modified as the need has arisen. Fourthly, as new methods are implemented, particularly for linkage analysis, including multilocus linkage analysis, these will be integrated also into the new software, using the same subroutine libraries. The new structure of the basic pedigree manipulation routines will be developed with this aim in view.

As the new structure of the software develops it will be thoroughly documented, with both online help files and PostScript documentation files for hard copy. Once the command files, help files, and other improvements of the user interface, are in place, we shall begin distribution to interested users, focusing first on users who have already expressed an interest in testing the software, from whom we can expect helpful feedback. From that point, we plan an electronic newsletter for users, containing updates, commonly asked questions (with answers where possible!), and other information: users comments and contributions to the newsletter will be encouraged. An e-mailing list for this newsletter will be maintained. Initially, the mailing list will be short and the newsletter brief; it may grow in response to demand.

Our current structure of programs and subroutine libraries is developed under the Unix system, and is best suited to that system. We use Unix Makefiles to specify dependencies among programs, subroutines, libraries, and header files, and to provide default names of data input files. Most of the methods are computationally intensive, and require a fast workstation. Initially, at least, the distributed version of the software will be only that on the Unix machine supported under the grant. We will continue to develop the structure primarily for Unix, but are also committed to maintaining portability. One of the co-investigators uses VMS, and must continue to do so (See "Resources and environment"), and there are many other users of VMS. We have streamlined procedures for transferring software to a VMS system, and will continue to ensure that this can be done easily. Initially we do not plan on producing versions outside of the Unix and VMS environments; however we will be willing to work with interested users who wish to make that conversion. Once command files replace the main directory Makefiles, portability outside the Unix environment will be easier.

4.7. Timetable

The separate thrusts of this methodological project will proceed in parallel. Wijsman's primary interests are in sections 4.1 and 4.2, Green's in 4.4, and the programmer will undertake 4.6. Thompson will be closely involved in all the component subprojects, although her current major interests are 4.3 and 4.5. The students also will provide integration of the separate thrusts, and all participants will meet together regularly to review progress.

In year 1, we shall focus primarily on multilocus mapping, and implement an initial version of the MCMC simulated tempering scheme. We shall also continue current work on multilocus gene identity by descent, and in year 2 extend this to affected relative set methods on extended pedigrees; again we anticipate that simulated tempering will play a role. In year 1 also we will continue current work on MCMC methods for age-of-onset models, including quantitative risk factors that may be influenced by the same trait genes. In years 2 and 3, we shall address problems of ascertainment in the context of these complex traits, and compare alternative sampling methods for obtaining an ascertainment-corrected Monte Carlo likelihood; this aspect may become the primary focus of the postdoctoral student. By year 3, models of a continuous genome may have become a practical data reality, and we aim to extend the discrete locus affected relative set methods to Monte Carlo sampling of junction points under continuous genome models.

On the software front, year 1 will see the overhaul of the current structure, year 2 the development of a user interface in the form of command files and help files, and year 3 will see the start of distribution. Also in year 3, we shall begin to integrate programs developed in the first two years into the software structure, and so expand the set of programs that are distributed.

In the final 2 years of the grant, we shall focus more towards the comparison and testing of methods, although we expect also further methodological developments in response both to our own project needs, in analyzing data on heart disease and other complex traits, and also to the needs of other users of our methodology and software.

5. Human subjects

The proposed research involves only the study of existing data. The information in the data bases to be used in this research has been recorded by the investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. Details of the available data are provided in the "Resources and environment" section.

6. Vertebrate animals

No research involving vertebrate animals is proposed. Some existing data on animal pedigrees may be used in developing methodology.

7. Consultants and Collaborators

In addition to the PI and Dr Wijsman, who have been collaborators in this research for four years, Dr. Phil Green and Dr. Charles Geyer also become key personnel. Biosketches and Other Support information for Drs. Green and Geyer are included in this proposal, with those of Thompson and Wijsman, and their titles, positions etc. are given with those of other key personnel on P.2.

Dr. Phil Green joins Thompson and Wijsman as an investigator on this grant. Although he does not wish for salary support, he will make a 3% time commitment, and will supervise, jointly with Thompson, a Statistics Research Assistant in his area of genome research, in relation to mapping genes for complex traits (see budget justification). A letter from Dr. Green is appended.

Dr. Charles Geyer has worked on our current grant, as a summer postdoctoral visitor. He has made important contributions to Monte Carlo likelihood theory. He is named as a Consultant on this competitive renewal; he will assist in implementing his general method of simulated tempering for MCMC in the specific application of multilocus linkage analysis. A letter from Dr. Geyer is appended.

8. Consortium and/or Contractual Arrangements

Not applicable.

9. Literature cited.

Abel, L., & Bonney, G. E. (1990). A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial disease. *Genet. Epidemiol.* **7:** 391–407.

Aldous, D. J. (1989) *Probability approximations via the Poisson clumping heuristic.* Springer; Berlin.

Besag, J. E., Green, P., Higdon, D. & Mengerson, K. (1994) Bayesian Computation and Stochastic Systems. *Statistical Science:* in press.

Blackwelder, W. C., & Elston, R. C. (1985). A comparison of sib-pair linkage tests for disease susceptibility. *Genet. Epidemiol.* **2:** 85–97.

Blossey, H. (1993). *The Poisson clumping heuristic and the survival of genome in small pedigrees* Ph.D. Thesis, Dept. Statistics, University of Washington.

Boehnke, M. (1994). Limits of resolution of genetic linkage studies: Implications for positional cloning of human disease genes. *Amer. J. Hum. Genet.* **55:** 379–390.

Boehnke, M., Lange, K., & Cox, D. R. (1991). Statistical methods for multipoint radiation hybrid mapping. *Amer. J. Hum. Genet.* **49:** 1174-1188.

Branscomb, E., Slezak, T., Pae, R., Galas, D., Carrano, A. V., & Watterman, M. (1990). Optimizing restriction fragment fingerprinting method for ordering large genomic libraries. *Genomics* **8:** 351-366.

Brown, M.S. & J.L. Goldstein. (1986). A receptor-mediated pathway for cholesterol homeostasis. *Science* **232:** 34–47.

Buetow, K. H. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *Amer. J. Hum. Genet.* **49:** 984–994.

Cavalli-Sforza, L. L., & King, M-C. (1986). Detecting linkage for genetically heterogeneous diseases and detecting heterogeneity with linkage data. *Amer. J. Hum. Genet.* **38:** 599–616.

Chait, A., Albers, J. J., & Brunzell, J. D. (1980). Very low density lipoprotein overproduction in genetic forms of hypertriglyceridaemia. *European J. of Clin. Invest.* **10:** 17–22.

Clayton, D. & Cuzick, J. (1985) Multivariate generalizations of the proportional hazards model. *J. Roy. Statist. Soc. (A)* **148:** 82–117.

Clerget-Darpoux, F., Bonaiti-Pellie, C., & Hochez, J. (1986). Effects of Misspecifying Genetic Parameters in Lod Score Analysis. *Biometrics* **42:** 393–399.

Cottingham, R.W., Idury, R.M., & Schäffer, A.A. (1993). Faster sequential genetic linkage computations. *Amer. J. Hum. Genet.* **53:** 252–263.

Cox, D. R., Burmeister, M., Price, E.R., Kim, S., & Myers, R. M. (1990). Radiation hybrid mapping: a somatic cel genetic method for constructing high resolution maps of mammalian chromosomes. *Science* **250:** 245–250.

Curtis, D., & Gurling, H. (1993). A procedure for combining two-point lod scores into a summary multipoint map. *Human Heredity* **43:** 173–185.

Davies, J.L., Kawagushi, Y., Bennett, S.T., Copeman, J. B., Cordell, H.J., Pritchard, L.E., Reed, P.W., Gough, S.C.L., Jenkins, S.C., Palmer, S.M., Balfour, K.M., Rowe, B.R., Farrall, M., Barnett, A. H., Bain, S.C., & Todd, J.A. (1994). A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371:** 130–136.

Donnelly, K. P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theor. Pop. Biol.* **23:**, 34–64.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39:** 1–38.

Egeland, J.A., Gerhard, D.S., Pauls, D.L., Sussex, J.N., Kidd, K.K., Allen, C.R., Hostetter, A.M., & Housman, D. E. (1987). Bipolar affective disorders linked to DNA markers on chromosome 11. *Nature* **325:** 783–787.

Elston, R. C. (1994). P-values, power, and pitfalls in the linkage analysis of psychiatric disorders. In *Genetic Approaches to Mental Disorders.* Gerson E.S. & Cloninger C.R. (eds) American Psychiatric Press, Inc. Washington DC. Pp. 3–21.

Elston, R. C., Bailey-Wilson, J. E., Bonney, G. E., Keats, B. J., & Wilson, A. S. (1986). S. A. G. E.: a papckage of computer programs to perform statistical analysis for genetic epidemiology. *7 th. Int. Congr. Hum. Genet.; Berlin.*

Elston, R. C., & George, V. T. (1989). Age of onset, age at examination, and other covariates in the analysis of family data. *Genet. Epidemiol.* **6:** 217–220.

Ewens W. J. & Shute, N. C. E. (1986). A resolution of the ascertainment sampling problem. I. Theory. *Theor. Pop. Biol.* **30:** 388–412.

Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. *J. Appl. Prob.* **30:** 766-779.

Feingold, E., Brown, P. O., & Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Amer. J. Hum. Genet.* **53:** 234–251.

Fisher, R. A. (1949). *The Theory of Inbreeding.* Oliver and Boyd: Edinburgh.

Fulker, D. W., & Cardon, L. R. (1994). A sib-pair approach to interval mapping of quantitative trait loci. *Amer. J. Hum. Genet.* **54:** 1092–1103.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Ass.* **46:** 193–227.

Geyer, C.J. (1991a). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.* Pp 156–163. Interface Foundation of North America, Fairfax Station, VA.

Geyer, C.J. (1991b). Reweighting Monte Carlo Mixtures. Technical Report No. 568, School of Statistics, University of Minnesota.

Geyer, C.J. & Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with Discussion). *J. Roy. Statist. Soc. (B),* **54:** 657–699.

Geyer, C. J. & Thompson, E. A. (1994). Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *J. A. S. A.: submitted.*

Goate, A. M., Chartier-Harlin, C. M., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra, L., Haynes, A., Irving, N., James, L., Mant, R., Newton, P., Rooke, K., Roques, P., Talbot, C., Pericak-Vance, M., Roses, A., Williamson, R., Rosser, M., Owen, M., & Hardy, J. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349:** 704–706.

Goldgar, G. E. (1990). Multipoint analysis of human quantitative genetic variation. *Amer.*

J. Hum. Genet. **47:** 957–967.

Goto, M., Rubenstein, M., Weber, J., Woods, K., & Drayna, D. (1992). Genetic linkage of Werner's syndrome to five markers on chromosome 8. *Nature* **355:** 735–738.

Green, E. D., & Olson, M. V. (1990). Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: A model for human genome mapping. *Science* **250:** 94–98.

Greenberg, D.A. & S.E. Hodge. (1989). Linkage analysis under "random" and "genetic" reduced penetrance. *Genet Epidemiol.* **6:** 259–264.

Guo, S. W. (1994) Marker-based genome comparison as an exploratory tool for mapping complex traits. Genetic analysis workshop 9: *submitted.*

Guo S-W & Thompson E.A. (1992). A Monte Carlo method for combined segregation and linkage analysis. *Amer. J. Hum. Genet.* **51:** 1111–1126.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B., & King, M-C. (1990). Linkage of early-Onset Familial Breast Cancer to Chromosome 17q21. *Science* **250:** 1684–1689.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo Methods.* Methuen & Co., London.

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2:** 3–19.

Hasstedt, S. J., & Cartwright, P. (1981). PAP – Pedigree Analysis Package. Technical report No. 13, Department of Medical Biophysics and Computing, University of Utah.

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

Hoeschele, I., Gianola, D. & Foulley, J.L. (1987). Estimation of variance components with quasi-continuous data using Bayesian methods. *J. Anim. Breed. Genet.* **104:** 334.

Huntington's Disease Collaborative Research Group. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72:** 971–983.

Kass, S., MacRae, C., Graber, H. L., Sparks, E. A., McNamara, D., Boudoulas, H., Basson, C. T., Baker, P. B. III., Cody, R. J., Fishman, M. C., Cox, N., Kong, A., Wooley, C. F., Seidman, J. G., & Seidman, C. E. (1994). A gene defect that causes conduction system disease and dilated cardiomyopathy maps to chromosome 1p1-1q1. *Nature Genetics* **7:** 546–551.

Kelsoe, J.R., Ginns, E.I., Egeland, J.A., Gerhard, D.S., Goldstein, A.M., Bale, S.J., Pauls, D.L., Long, R.T., Kidd, K.K., Conte, G., Housman, D.E., & Paul, S.M. (1989). Re-evaluation of the linkage relationship between chromosome 11p loci and the gene for bipolar affective disorder in the Old Order Amish. *Nature* **342:** 238–243.

Kong, K., Irwin, M., Cox, N. & Frigge, M. (1992). Multiloci problems and the method of sequential imputation. Technical Report No. 351, Department of Statistics, University of Chicago.

Kruglyak, L., & Lander, E. S. (1994). HOMOZ: a software package for homozygosity mapping of recessive traits in children of inbred marriages. In *Gene Mapping and Sequencing.* Cold Spring Habor Laboratory, New York.

Lander, E.S., & Botstein, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236:** 1567–1570.

Lander, E.S., & Green, P. (1987). Construction of multilocus linkage maps in humans. *Proceedings of the National Academy of Sciences (USA)* **84:** 2363–2367.

Lange, K., & Sobel, E. (1991) A random walk method for computing genetic location scores. *Amer. J. Hum. Genet.* **49:** 1320-1334.

Lathrop, G.M., Lalouel, J-M., Julier, C., & Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences (USA)* **81:** 3443–3446.

Li, H., & Thompson, E.A. (1994) Modeling age-of-onset; Cox model with latent major gene effects. *Amer. J. Hum. Genet.* **55:** A156.

Lin, S. (1994) Monte Carlo multipoint linkage analysis. *Submitted.*

Mack, W., Langholz, B., & Thomas, D. C. (1990) Survival models for familial aggregation of cancer. *Environmental Health Perspectives* **87:** 27–35.

Marinari, E. & Parisi G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19:** 451–458.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21:** 1087–1092.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennet, L. M., Ding, W., Bell, R., Rosenthal, J., Hussey, C., Tran, T., McClure, M., Frye, C., Hattier, T., Phelps, R., Haugen-Strano, A., Katcher, H., Yakumo, K., Gholami, Z., Shaffer, D., Stone, S., Bayer, S., Wray, C., Bogden, R., Dayananth, Priya., Ward, J., Tonin, P., Narod, S., Brostow, P. K., Norris, F. H., Helvering, L., Morrison, P., Rosteck, P., Lai, M., Barrett, J. C., Lewis, C., Neuhausen, S., Cannon-Albright, L., Goldgar, D., Wiseman, R., Kamb, A., & Skolnick, M. H. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266:** 66–71.

Morton, N.E. (1955). Sequential tests for the detection of linkage. *Amer. J. Hum. Genet.* **7:** 277–318.

Nelson, S. F., McCusker, J.H., Sander, M. A., Kee, Y. Modrish, P. & Brown, P. O. (1993). Genomic mismatch scanning; A new approach to genetic linkage mapping. *Nature Genet.* **4:** 11-18.

O'Connell, P., Lathrop, G. M., Law, M., Leppert, M., Nakamura, Y., Hoff, M., Kumlin, E., Thomas, W., Elsner, T., Ballard, L., Goodman, P., Azen, E., Sadler, J., Cai, G., Lalouel, J-M., & White, R. (1987). A primary genetic linkage map for human chromosome 12. *Genomics* **1:** 93–102.

Olson, J. M. & Wijsman, E. M. (1993). Linkage between quantitative trait and all marker loci: methods using all relative pairs. *Genet. Epidemiol.* **2:** 85–97.

Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *Amer. J. Hum. Genet.* **31:** 161–175.

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

Ott, J. (1991). *Analysis of Human Genetic Linkage.* (2 nd. ed.) The Johns Hopkins University Press; Baltimore, MD.

Ott, J. (1993). Detecting marker inconsistencies in human gene mapping. *Hum Hered.,* **43:** 25–30.

Ploughman, L.M. & Boehnke M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *Amer. J. Hum. Genet.* **44:** 543–551.

Prochazka, M., Lillioja, S., Tait, J. F., Knowler, W. C., Mott, D. M., Spraul, M., Bennet, P. H., & Bogardus, C. (1993). Linkage of chromosomal markers on 4q with a putative gene determining maximal insulin action in Pima Indians. *Diabetes,* **42:** 514–519.

Rao, D.C., Keats, B.J.B., Morton, N.E., Yee, S., & Lew, R. (1978). Variability of human linkage data. *Amer. J. Hum. Genet.* **30:** 516–529.

Riordan, J. R., Rommens, J.M., Kerem, B-S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J-L., Drumm, M.L., Iannuzzi, M.C., Collins, F.S., & Tsui, L-C. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245:** 1066–1073.

Saad, M. F., Knowler, W. C., Pettit, D. J., Nelson, R. G., Mott, D. M., & Bennet, P. H. (1989). Sequential changes in serum insulin concentration during development of non-insulin dependent diabetes. *Lancet,* **1:** 1356-1359.

Schellenberg, G.D., Bird, T.D., Wijsman, E.M., Moore, D.K., Boehnke, M., Bryant, E.M., Lampe, T.H., Nochlin, D., Sumi, S.M., Deeb, S.S., Beyreuther, K., & Martin, G.M. (1988). Absence of linkage of chromosome 21q21 markers to familial Alzheimer's disease. *Science* **241:** 1507–1510.

Schellenberg, G.D., Bird, T.D., Wijsman, E.M., Orr, H.T., Anderson, L., Nemens, E., White, J.A., Bonneycastle, L., Weber, J.L., Alonso, M.E., Potter, H., Heston, L.L., & Martin, G.M. (1992a). Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. *Science* **258:** 668–671.

Schellenberg, G.D., Martin, G. M., Wijsman, E.M., Nakura, J., Miki, T., & Ogihara, T. (1992b). Homozygosity mapping and Werner's syndrome. *The Lancet* **339** 1002.

Self, S. G., & Prentice, R. L. (1985). Incorporating random effects into multivariate relative risk regression models. In *Modern Statistical Methods in Chronic Disease Epidemiology,* S.H. Moolgavkar & R.L Prentice (eds.) John Wiley & Sons, New York.

Shiang, R., Thompson, L.M., Zhu, Y.Z., Church, D.M., Fielder, T.J., Bocian, M., Winokur, S.T., & Wasmuth, J.J. (1994). Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. **Cell 78:** 335–342.

Smith, C. A. B. (1953). Detection of linkage in human genetics. *J. Roy. Statist. Soc. B* **15:** 153–192.

Sorensen, D. A., Andersen, S., Gianola, D. & Korsgaard, I. (1994). Bayesian inference in threshold models using Gibbs sampling. *Genet. Sel. Evol.:* to appear.

St. George-Hyslop, P.H., Tanzi, R.E., Polinsky, R.J., Haines, J.L., Nee, L., Watkins, P.C., Myers, R. H., Feldman, R. G., Pollen, D., Drachman, D., Growden, J., Bruni, A., Foncin, J.-F., Salmon, D., Frommelt, P., Amaducci, L., Sorbi, S., Piacentini, S., Stewart, G. D., Hobbs, W. J., Conneally, P. M., & Gusella, J. F. (1987). The genetic defect causing familial Alzheimer's disease maps on chromosome 21. *Science* **235:** 885–890.

St. George-Hyslop, P., Haines, J., Rogaev, E., Mortilla, M., Vaula, G., Pericak-Vance, M., Foncin, J.-F., Montesi, M., Bruni, A., Sorbi, S., Rainero, I., Pinessi, L., Pollen, D., Polinsky, R., Nee, L., Kennedy, J., Macciardi, F., Rogaeva, E., Liang, Y., Alexandrova, N., Lukiw, W., Schlumpf, K., Tanzi, R., Twuda, T., Farrer, L., Cantu, J-M., Duara, R., Amaducci, L., Bergamini, L., Gusella, J., Roses, A., Crapper McLachlan, D. (1992). Genetic evidence for a novel familial Alzheimer's disease locus on chromosome 14. *Nature Genetics* **2:** 330–334.

Stephens, D.A. & Smith, A. F. M. (1993). Bayesian inference in multipoint gene mapping. *Ann. Hum. Genet.* **57,** 65–82.

Suarez, B.K. & Van Eerdewegh, P. (1984). A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. *Amer. J. Med. Genet.* **18:** 135–146.

Thomas, A. W., Skolnick, M. H., & Lewis, C. M. (1994). Genomic mismatch scanning in pedigrees. *I.M.A. J. Math. Appl. Med. & Biol.,* **11:** 1–16.

Thompson E. A. (1987). Crossover counts and likelihood in multipoint linkage analysis. *IMA J. of Math. Appl. in Med. & Biol.* **4:** 93–108.

Thompson E. A. (1991) Probabilities on complex pedigrees; the Gibbs sampler approach. *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface,* Pp 321-328. Interface Foundation of North America, Fairfax Station, VA.

Thompson, E. A. & Guo, S-W. (1991). Evaluation of likelihood ratios for complex genetic models. *I.M.A. J. Math. Appl. Med. & Biol.,* **8:** 149–169.

Thompson, E. A. (1994a). Monte Carlo likelihood in linkage analysis. *Statistical Science;* in press.

Thompson, E. A. (1994b). Monte Carlo estimation of multilocus autozygosity probabilities. *Proceedings of the 1994 Interface conference* SAS: Cary, NC. In press.

Weber, J. L., & Wong, C. (1993). Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2:** 1123–1128.

Weeks, D.E., & Lange, K. (1988). The affected-pedigree-member-methods of linkage analysis. *Amer. J. Hum. Genet.* **42:** 315–326.

Wijsman, E.M., Motulsky, A.G., Guo, S.W., Yang, M., Austin, M.A., Brunzell, J.B., & Deeb, S. (1992). Evidence against linkage of familial combined hyperlipidemia to the AI-CIII-AIV complex. *Circulation* **86:** I-420.

Wojciechowski, A.P., Farrall, M., Cullen, P., Wilson, M.E., Bayliss, J.D., Farren, B., Griffin, B.A., Caslake, M.J., Packard, C.J., Shepherd, J., Thakker, R., & Scott, J. (1991). Familial combined hyperlipidaemia linked to the apolipoprotein AI-CIII-AIV gene cluster on chromosome 11q23-q24. *Nature* **349:** 161–164.

Wooster, R., Neuhausen, S.L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., Averill, D., Fields, P., Marshall, G., Narod, S., Lenoir, G. M., Lynch, H., Feunteun, J., Devilee, P., Cornelisse, C. J., Menko, F. H., Daly, P. A., Ormiston, W., McManus, R., Pye, C., Lewis, C. M., Cannon-Albright, L. A., Peto, J., Ponder, B. A. J., Skolnick, M. H., Easton, D. F., Goldgar, D. E., & Stratton, M. R. (1994). Localization of breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265:** 2088–2090.

APPENDIX: List of submitted materials.
A. Seven published or in press papers

1. Thompson, E. A. and Guo, S-W. (1991). Evaluation of likelihood ratios for complex genetic models. *I.M.A. J. Math. Appl. Med. & Biol.,* **8:** 149–169.
2. Guo S-W and Thompson E.A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics*, **50:** 417-432.
3. Guo S-W and Thompson E.A. (1992). A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics*, **51:** 1111–1126.
4. Thompson, E. A., Lin, S., Olshen, A. B. and Wijsman, E.M. (1993). Monte Carlo analysis on a large pedigree. *Genetic Analysis Workshop 8. Genetic Epidemilogy;* **10:** 677–682.
5. Lin, S., Thompson, E. A., and Wijsman, E. M. (1994) Finding non-communicating sets for Markov chain Monte Carlo estimations on pedigrees. *Amer. J. Hum. Genet.,* **54:** 695–704.
6. Thompson, E. A. (1994). Monte Carlo likelihood in the genetic analysis of complex traits. Proceedings of the Royal Society (Lond) Series B **344:** 345–351.
7. Thompson, E. A. (1994). Monte Carlo estimation of multilocus autozygosity probabilities. *Proceedings of the 1994 Interface conference* SAS: Cary, NC.

B. Program documentation

Thompson, E. A. (1994) *Monte Carlo Programs for Pedigree Analysis: 1990-93.* Technical Report No. 267, Dept. Statistics, University of Washington.