9. RESEARCH PLAN

9a: 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 conditional probability estimation on complex data structures. There is a need for techniques for analysis of complex models that not only combine linkage and segregation analysis, but also allow for complex multivariate phenotypes, including covariate information and censoring, and permit full use of the multilocus marker data now becoming available. Only then can we resolve the components of traits that are influenced by environmental factors as well as by genes at several loci. We propose to develop methods of analysis of phenotypic and marker data observed on members of a pedigree. This will aid in the genomic localization of genes which contribute to complex genetic traits. Our present objectives build on those accomplished in our currently funded research.

Specific objectives:

1. To develop further our Markov chain Monte Carlo (MCMC) methods for multilocus linkage analyses on extended pedigrees, using an array of linked markers, by
(i) estimating conditional gene identity-by-descent (IBD) patterns jointly over individuals and over loci, for linkage detection and gene resolution,
(ii) continuing to develop MCMC samplers of gene IBD patterns and multilocus haplotypes, to permit more efficient analysis of data observed at tightly linked markers, and
(iii) developing MCMC methods of haplotype analysis at the pedigree and population levels, for disequilibrium analyses and fine-scale mapping.

2. To develop further effective Monte Carlo analysis methods for the joint multilocus linkage and segregation analysis of complex traits, by
(i) extending Bayesian MCMC oligogenic linkage analyses to ordered categorical phenotypes,
(ii) allowing for missing and/or censored covariates, and for censored phenotypes,
(iii) extending the oligogenic models that underlie quantitative and categorical traits, and
(iv) extending MCMC oligogenic linkage analyses to multiple correlated phenotypes.

3. To develop methods to analyze and interpret the outputs of our Monte Carlo methods, by
(i) calibrating the MCMC realizations of oligogene locations, to obtain measures of the reliability of linkage inferences,
(ii) assessing the gains of multilocus linkage analyses using oligogenic trait models with multiple QTL, and
(iii) developing methods for the interpretation of gene IBD patterns, estimated jointly over loci and over individuals.

4. To develop Monte Carlo methods to analyze genetic maps, genetic interference, and meiosis models, by
(i) studying map heterogeneity, uncertainty, and error, and their impact on haplotype imputation, and on genotype and pedigree error detection,
(ii) extending current MCMC methods to permit linkage detection and haplotype estimation under certain interference models, investigating the impact on linkage detection for oligogenic traits, and
(iii) developing MCMC methods to realize underlying meiosis processes conditional upon genetic marker data, in order to fit and use more general meiosis models that incorporate interference.

5. In addition to appropriate simulation studies, to evaluate methods developed under #1, #2, #3, and #4, on real data, by
(i) using data from a genomic array of markers to evaluate MCMC approaches to linkage detection and

oligogenic linkage analysis
(ii) evaluating methods of joint segregation and linkage analysis applied to complex correlated phenotypes, including censored and categorical phenotypes, with covariate information, and
(iii) exploring the use of MCMC methods on very large multi-generation pedigrees, to evaluate the utility of such pedigrees for linkage detection, haplotype analysis, and fine-scale mapping.

6. To extend and enhance our user software, by
(i) further developing the user-friendly MORGAN_2 package to contain the full array of our current methods and programs,
(ii) implementing new methodology as proposed in #1, #2, #3 and #4, and
(iii) documenting, distributing and supporting the software, serving users through electronic mail, and seeking user response, to aid in ongoing software improvement.

9b: Background and significance

9b.1: Complex traits: the genetic approach

The risk for developing many common diseases is believed to be influenced by both genetic and environmental factors. Examples include cardiovascular disease, neurodegenerative diseases, mental illness, and cancers. Identification of genes which predispose to such diseases (disease genes) provides a key step towards developing understanding disease etiology. Examples include identification of the low density lipoprotein receptor and its role in coronary artery disease (Brown and Goldstein 1986), the two presenillin genes and their roles in Alzheimer's disease (Sherrington et al. 1995; Levy-Lahad et al. 1995a), as well as genes involved in several cancers (Miki et al. 1994; Papadopoulos et al. 1994), and diabetes (Yamagata et al. 1996).

Identification of disease genes usually involves several stages of gene localization followed by gene identification. Initial mapping by population-level association studies, using identity-by-state (IBS) measues, may be possible when there already exists considerable biological information, such as a known candidate gene. However, both empirical data (Pennisi 1998) and theoretical considerations (Chapman and Wijsman 1998) are starting to raise questions about claims (Risch and Merikangas 1996; Brown and Hartwell 1998) that this approach will succeed for complex traits in the absence of such prior biological information. On the other hand, the positional cloning strategy has been responsible for identifying genes for >100 diseases (http://www.nhgri.nih.gov/DIR/GTB/CLONE), representing >150% increase over the number identified just four years ago. Each of these was first localized to a small genomic region by pedigree-based linkage analysis methods, which are founded on analysis of gene identity-by-descent (IBD). An additional > 3250 genes have been mapped to specific genomic regions with such linkage analysis methods (http://www3.ncbi.nlm.nih.gov/Omim). Once a gene has been mapped, however, it is difficult to use family studies to achieve the resolution needed for gene identification since pedigree data have relatively small numbers of meioses (Boehnke 1994). Here the allelic disequilibria in populations which are the outcome of recombination events on a coalescent ancestry provide the necessary resolution because of the large number of meioses, although these meioses are very incompletely observed. Use of such data for fine-scale localization of disease genes has included both IBS approaches (Xiong and Guo 1997; van der Meulen and te Meerman 1997; Lazzaroni 1998) and IBD approaches (Kaplan et al. 1995; Rannala and Slatkin 1998; Graham and Thompson 1998). Conceptually, linkage analysis and fine-scale mapping by linkage disequilibrium methods are equivalent: both in pedigrees and populations, coancestry at a trait locus implies patterns of IBD in relatives in surrounding genome segments.

Despite some success with analysis of complex diseases, there remain difficulties. The success rate in mapping such disorders has been lower than for Mendelian traits, and for the most part, the examples represent unusual forms of the diseases. For some traits, there have been multiple genome screens that failed to identify clearly new gene locations (e.g. Sawcer et al. 1996; Ebers et al. 1996; Haines et al. 1996), and some examples of false positive reports of linkage that were later exposed by identification of the correct location and gene (e.g. St.George-Hyslop et al. 1987 vs. Sherrington et al. 1995). Such situations rarely arise for monogenic diseases (Rao et al. 1978).

Limitations in the available techniques of analysis contribute to the difficulty, compared to simple monogenic traits, in identifying genes contributing to complex traits. These limitations exist for several reasons. First, there are severe practical problems with performing computations for analyses of complex genetic models on large pedigrees with many markers, yet these may be the conditions necessary to obtain sufficient power and resolution to detect linkage and clone the genes (Wijsman and Amos 1997). Second, use of multipoint analyses for linkage detection is more common for complex than for simple traits, but map error, poor meiosis models, and undetected genotyping errors may all be problems for analyses based on many, closely linked, markers. Third, current methods do not allow efficient use to be made of such information as covariates and multivariate phenotypes. Finally, recent methodological developments that do not require specification of a trait-model ("model-free" methods) are typically limited to restricted, small pedigree structures, which are inefficient for both linkage detection and gene localization (Wijsman and Amos 1997).

9b.2 Linkage analysis and genetic maps

A variety of methods exist for linkage analysis of complex traits. Statistical methods for linkage detection include "model-based" methods that either require a pre-specified trait-locus model (Morton 1955), or fit a trait-locus model jointly with an estimate of gene location in variance-component (Amos 1994; Almasy and Blangero 1998) and quantitative trait locus (QTL) methods (Heath 1997). "Model-free" methods are also available which do not require specification of a trait locus model. These range from a variety of sib-pair methods (Suarez et al. 1978) to methods based on more extended pedigrees (Weeks and Lange 1988; Risch 1990; Kruglyak et al. 1996). Use of these methods has recently become much more widely used due to the ready availability of IBD posterior probabilities out of linkage programs such as GENEHUNTER (Kruglyak et al. 1996). Tests have become formalized through theoretical studies of IBD statistics by, among others, Whittemore and Halpern (1994), Whittemore and Tu (1998), and Feingold and Siegmund (1997). "Model-free" methods for small pedigrees have the advantage of rapid computation, ease of understanding, and lack of need to assume a trait-locus model. However, they are often restricted to pedigrees of limited size and/or structure, and have low power relative to use of model-based methods. The limitations in pedigree structure imposed by most current model-free methods means that significant proportions of some data sets may be unusable, which is of concern given the high cost of collecting such data (see **9b.3**). Model-based methods, on the other hand, can be used on pedigrees of essentially arbitrary size and structure, and are more efficient than are "model-free" methods, thereby requiring smaller sample sizes for linkage detection in ideal cases. However, model misspecification for a complex trait may cause severe reduction in power to detect linkage (Risch and Giuffra 1992), and computational requirements can be extreme— measured in weeks to months of CPU time, even for analyses with only a few marker loci (Schellenberg et al. 1992).

All linkage analysis methods can in principle be used either with one marker or many. Multipoint linkage analysis has long been known to increase power and to provide more accurate gene localization than single-marker analyses (Lathrop et al. 1984). Widely used analysis programs that can be used with a single trait-locus but more than a single marker and that employ exact computational methods include FASTLINK (Cottingham et al. 1993) and VITESSE (O'Connell and Weeks 1995) for model-based lod score analysis, and GENEHUNTER (Kruglyak et al. 1996) for either model-based or model-free analysis. Each approach has limitations. FASTLINK and VITESSE can handle large pedigrees, but VITESSE is restricted to simple-structure pedigrees. GENEHUNTER is restricted to small pedigrees with fewer than about 12 non-founders, but can incorporate very large numbers of markers into the analysis. VITESSE and, even more so, FASTLINK, can usually only handle a few markers in the analysis before either time or computer memory even on a fast workstation become limiting factors. Exact computations for 2-trait-locus linkage analysis implemented in TMLINK (Lathrop and Ott 1990) are only computationally practical in analysis with single markers, although multipoint analyses are theoretically possible. SOLAR (Almasy and Blangero 1998), which implements the variance-component approach and can perform analyses for certain oligogenic trait models, can also perform multipoint analyses on large pedigrees, but the multipoint computations are performed with approximation methods (Fulker et al. 1995). Finally, **Loki** (Heath et al. 1997; Heath 1997) implements a computationally practical Bayesian MCMC approach to oligogenic-trait multipoint linkage analysis in the context of QTL models, thereby substituting estimation for approximation in the multipoint computations.

Genetic maps are an important component of multipoint linkage analysis. A genetic map consists of the order and intermarker distances between a series of marker loci, these genetic distances being a function of recombination frequencies estimated from meiotic data. While physical distance (e.g., in base pairs) is an unarguable measure of distance, meiotic map distances can vary considerably from the distances predicted from physical maps (Nachman and Churchill 1996) so that maps must be estimated from meiotic data. Distance estimates for meiotic maps are based on certain models and assumptions about the recombination process. Simple mapping functions are generally used to estimate distances from meiotic data (e.g., Haldane 1919) even though there is evidence that interference and other high-order processes do occur (Schmitt et al. 1994). Maps may assume equal recombination fractions in males and females even though it is well established that these rates are not the same (Broman et al. 1998; Donis-Keller et al. 1987) and that there is extensive variation with map position of the male-female recombination ratio (Kwiatkowski et al. 1993). Sex-specific maps also exist, but even these assume that recombination fractions are constant across individuals, which also may not be true (Broman et al. 1998). Finally, map distances are derived from real data, and thus are subject to statistical issues associated with sample size. Most maps in use today are based on a very limited set of ~ 200 meiotic events (~ 100 for each sex): the 95% confidence interval for a typical 10 cM distance between markers in current maps (e.g., http://www.marshmed.org/genetics) ranges from 4-19 cM in sex-averaged maps, and 2-22 cM in sex-specific maps (Yuan et al. 1997). For higher-density maps of diallelic markers (SNPs) detected with new automated detection systems (Wang et al. 1998), this sample size is inadequate for accurate map estimation, especially for sex-specific maps.

Interference models become important for dense maps and when there is heterogeneity in the recombination process. As maps become denser, recombination frequencies used to construct maps are increasingly estimated from events between tightly linked loci; although a map function can be used to adjust these, this has almost no impact on these small genetic distances. Using such maps with a computational method which does not allow for genetic interference can severely distort, relative to those observed in actual meioses, joint meiosis patterns at loci at distance 30–50cM (Thompson and Meagher 1998). With the availability of an increasing array of genetic markers, likelihood analyses of alternative interference models becomes practical (McPeek and Speed 1995; Speed 1996). Heterogeneity in recombination can give the appearance of genetic interference, or can mask the effects of interference. Heterogeneity in recombination may lead to false positive indications of linkage (Daw et al. 1998b). It also leads to heterogeneity in the extent of linkage disequilibria, leading to misinterpretation if the heterogeneity is undetected or unaccounted. The impact of interference on linkage analysis and linkage detection is unclear, although Goldstein et al. (1995) suggest there may be substantial loss of linkage information through failure to model interference.

Detection of error becomes increasingly important with use of multipoint linkage analyses based on published map estimates. Many pedigree and typing errors elude detection in locus-by-locus analyses, particularly if connecting members of a pedigree are unobserved. In multipoint analyses, such errors result in apparent low-probability multiple-recombinants in certain meioses (Boehnke and Cox 1997), or low-probability patterns of patterns of gene IBD among relatives (Browning 1998). If undetected, this can have serious consequences for the resulting analyses. Programs such as Mapmaker/Sibs (Kruglyak and Lander 1995) can be used to identify potential errors in small pedigrees. For large pedigrees, however, since computations needed for multipoint error detection are the same as those used for linkage analysis, such haplotype-based error detection is impractical.

9b.3: Efficient data use and large pedigrees

Genetic studies are expensive. This is particularly true for studies of complex traits where often many different phenotypic measures are taken, some at considerable cost. While genotyping costs are non-trivial, these costs are rapidly dropping through improved technology: genotyping costs at the NHLBI Marshfield mammalian genotyping service halved in the 2-year period ending in 1997, dropping to \$0.67/genotype (Zhao et al. 1997). The relative costs attributable to sample collection are now currently at least twice the cost of the genotyping in a genome scan (e.g., NIH grant HD33812, at \$950/sampled individual) and are often an order of magnitude or more greater (e.g., NIH grant AG05136 at > \$3000/ sampled individual). Since the cost of such sample collection is not likely to drop significantly through technological improvements, it is imperative that methods of statistical analysis are made available that can make efficient use of all the data.

Considerable information exists concerning data analysis methods and efficient use of data in genetic studies. The results of a large simulated data set analyzed as part of Genetic Analysis Workshop 10 (GAW10) provide a number of convenient contrasts. Four key factors that are relevant here had large effects on ability to detect linkage and to refine estimates of gene location (Wijsman and Amos 1997). First, large pedigrees are more efficient, on a per-sample basis, than are small pedigrees. For example, for the simulations in GAW10, lod scores were 1.7-3 times higher in data sets consisting of large pedigrees than in equivalently-sized data sets of small pedigrees (Wijsman and Amos 1997), depending on details of the analysis. The accuracy of gene localization was also better in the large than in the small pedigrees. Similar results were reported by Xie et al. (1998) for alternative designs used in agricultural settings, with information differences of 6-7 fold in pedigree data sets consisting of single large vs. many nuclear pedigrees. Second, the GAW10 simulations demonstrated that use of covariate adjustments could result in a 2-3 fold increase in mapping information over analyses without such use of covariates. Third, careful attention to choice of an analysis model that summarized the full phenotype also yielded approximately twice the information as did less carefully constructed models. Finally, use of multipoint analysis with up to 25 markers produced gains of up to 40% over use of single markers. Presumably for less polymorphic markers the gains achieved by multipoint analysis would be even higher.

A survey of recent mapping studies suggests that use of large pedigrees remains a popular design, in support of these simulation results. For purposes of discussion here, we define a large pedigree as one for exact multipoint computations with many markers is impossible with the inheritance-vector approach implemented in GENEHUNTER (Kruglyak et al. 1996): a pedigree with no more than about 12 non-founders and 16 transmissions from common ancestors. Roughly 2/3 of the gene mapping studies (21 out of 32 studies) reported in the June - August, 1998 issues of the American Journal of Human Genetics included at least one such large pedigree. A survey of the projects approved for genotyping between 1995 and 1997 by the NHLBI through the Marshfield Mammalian Genotyping service (NO1-HV-48141) indicated that even for complex traits, approximately 2/3 of the studies use large pedigrees. Of the 24 studies supported for genotyping by the NHLBI during this period, 12 involved large pedigrees, 6 involved only small pedigrees, and the information on pedigree size for 6 is not yet readily available. Among the 12 studies using large pedigrees, there are several with pedigrees of 50-100 individuals. Most of these projects involve studies of complex traits: lipid-level abnormalities, asthma, cancer, psychiatric illness, etc. Past successes in mapping and identifying genes for complex support this strategy: most gene identification of complex genetic traits has involved large pedigrees, including Alzheimer's disease (Levy-Lahad et al. 1995b; Levy-Lahad et al. 1995a; Schellenberg et al. 1992; Sherrington et al. 1995), colon cancer (Bodmer et al. 1987; Kinzler et al. 1991; Lindblom et al. 1993; Peltomaki et al. 1993), breast cancer (Miki et al. 1994; Hall et al. 1990; Wooster et al. 1994; Wooster et al. 1995), and diabetes (Froguel et al. 1992; Gidh-Jain et al. 1993; Vaxillaire et al. 1995).

The need for efficient methods of analysis coupled with the information on these influential factors suggests three guiding themes for the development of statistical genetic methodology. First, the methods must be powerful. It is important that they can use all the available data efficiently so that data collection costs can be minimized. Model-based methods are advantageous here, when a good model can be developed. Second, the methods must be flexible. It is critical that they are capable of analyzing the full spectrum of possible data sets including a full range of pedigree structures and types of phenotypes. It is important for the success of genetic studies that data should be collected and phenotypes defined according to scientific questions of interest, rather than according to constraints imposed by the limitations of statistical methodology, and that data need not be discarded because of such limitations. Finally, the methods must lead to the development of tools that provide computationally tractable analyses with interpretable results.

9b.4: Sampling based methods for structured stochastic systems

Over the last ten years, Monte Carlo sampling based methods of analysis have become increasingly used in a wide variety of structured stochastic systems (Gilks et al. 1996). In such systems, exact computation of probabilities is often infeasible, but the conditional independence structure makes local conditional probabilities readily obtainable. Realization of latent variables from the local distributions makes feasible and practical the Monte Carlo estimation of probabilities and likelihoods using Markov chain Monte Carlo (MCMC). The Metropolis algorithm is an MCMC algorithm dating to Metropolis et al. (1953), while the Gibbs sampler (Geman and Geman 1984) is a more recent innovation, which has become widely using in Bayesian analyses (Gelfand and Smith 1990). However, both are special cases of the wider class of Metropolis-Hastings MCMC methods (Hastings 1970), while reversible-jump MCMC (Green 1995) is a generalization of these ideas, permitting realizations from more general state spaces.

In the genetics context, there are two key structures that underlie computational methods: the pedigree structure of relationships among individuals and the linear structure of a chromosome. Correspondingly there are two sets of latent variables, whose conditional independence structure is a key property facilitating likelihood computations. Considering the underlying genotypes $\mathbf{G} = \{G_{ij}\}$ for all members i of a pedigree structure at all loci j, the probability of observed marker and/or trait phenotypes $\mathbf{Y}$ is

$$P(\mathbf{Y}) \quad = \quad \sum_{\mathbf{G}} P(\mathbf{Y} \mid \mathbf{G}) \; P(\mathbf{G}) \tag{1}$$

which underlies computational approaches based on peeling (Elston and Stewart 1971). Alternatively, we may consider the inheritance vectors or meiosis indicators $\mathbf{S} = \{S_{ij}\}$, specifying the parental origins of genes at each locus j in each meiosis i:

$$S_{ij} \quad = \quad 1 \text{ or } 0 \text{ as the gene is grandpaternal or grandmaternal} \tag{2}$$

Then

$$P(\mathbf{Y}) \quad = \quad \sum_{\mathbf{S}} P(\mathbf{Y} \mid \mathbf{S}) \; P(\mathbf{S}) \tag{3}$$

and the Markov dependence of S_{ij} along a chromosome, gives rise to alternative computational approaches (Lander and Green 1987). MCMC sampling-based methods of analysis have used the genotypes (Thompson and Guo 1991), the meiosis indicators (Thompson 1994b), or both (Lange and Matthysse 1989).

A major difficulty with MCMC methods is to ensure proper mixing of the samplers, and hence efficient Monte Carlo estimation. On large pedigrees, with models or data involving multiple linked loci, single-variable MCMC updating methods are not effective. Recently a variety of joint-updating schemes have been developed. For example, Jensen et al. (1995) update genotypes of blocks of individuals jointly at several loci, while Heath (1997) updates all genotypes at a single locus. Thompson and Heath (1998) update jointly meiosis indicators for all loci in a single meiosis. All these MCMC provide, directly or indirectly, realizations of the descent of genes in pedigrees and the genotypes of individuals, and hence Monte Carlo estimates of likelihoods for linkage and segregation analysis, posterior probabilities of gene identity by descent and haplotype sharing, or, in a Bayesian framework, posterior probabilities of linkage and trait gene effects and locations.

However, the latent variables of gene type and gene descent, do not provide the full depth of structure from gametogenesis to complex phenotype. Below the level of meiosis indicators lie the events of chiasma formation. Above the level of genotypes, lie the consequent liabilities that may underlie the observable phenotype relating to a complex genetic disease. Also underlying such phenotypes are liabilities resulting from environmental risk factors. Such factors may be unknown, but shared among relatives, or known but not observed for all individuals. The currently proposed research will improve current methods for Monte Carlo analyses, using meiosis indicators and genotypes, but will also extend the range of models by extending the latent variable framework.

By considering the latent process of chiasma formation and gametogenesis, we can extend models for meiosis to allow for interference and map heterogeneity (for example, between male and female meioses: see **9b.2**). Estimation of map parameters, individual haplotypes, and linkage model likelihoods, under such meiosis models, will provide improved map estimation and allow assessment of the effects of interference and heterogeneity on trait linkage detection and on trait model inferences. The results of Thompson and Meagher (1998) suggest that failure to allow for interference impacts estimates of joint transmission of genes. Bishop and Thompson (1988) and Goldstein et al. (1995) have shown that failure to account for interference adversely affects map estimation and linkage detection efficiencies.

MCMC and other Monte Carlo methods, providing either estimates of posterior probabilities or Monte Carlo likelihood functions, are ideally suited to all forms of "incomplete data", including censored traits (age-of-onset analyses) and latent disease liabilities. By considering latent variables between the level of genotype and phenotype, such as genotypically determined latent liabilities, we can greatly extend the range of complex-trait models and data observation structures it is possible to analyze. Data observations may be censored not only by unavailability or death, but also by delayed onset. Quantitative traits or risk factors may also be age-dependent. Ages of some individuals may be unavailable. While liability may be best modeled as a quantitative trait, phenotypes may be multivariate and/or categorical. By considering unavailable or unobservable variables as model components to be sampled, the same approaches we have used successfully for simpler quantitative and qualitative traits can be brought to bear on more truly complex phenotypes.

9c: Progress report

9c.1: Summary of previous specific aims

The specific aims of the previous award period (1995-99) were

1. To develop effective Monte Carlo computational methods for the computation of multilocus linkage likelihoods on extended and/or complex pedigrees, for Mendelian traits mapped against an array of linked markers, for traits that are genetically heterogeneous, of incomplete penetrance, or delayed onset, and for genetically determined quantitative risk factors.
2. To develop further our Monte Carlo methods for maximizing likelihoods and fitting models for complex genetic traits, including linkage analysis of marker data, using censored age-of-onset data for diseases of delayed onset, and analyzing data on genetically determined quantitative risk factors.

3. To develop Monte Carlo methods for multilocus linkage analyses based on the affected members of a pedigree, using an array of linked markers, extending current methods for homozygosity mapping of affected individuals with complex ancestry, considering jointly several affected individuals in an extended or complex pedigree, and investigating the potential for application to complex traits.
4. To evaluate the procedures of items #1, #2 and #3 by performing simulation studies to assess the performance of methods and comparing results to those obtained by other computational approaches, and by analyzing Alzheimer's disease data, and other qualitative genetic disease data available to us, using data on multiple markers.
5. To extend Monte Carlo likelihood to further areas of multilocus linkage analysis, developing probability estimation procedures for models of a continuous genome, and applying such models to genome-matching analyses of marker data on related individuals.
6. To produce user software, integrating our current methods and programs into a coherent package, implementing new methodology as proposed in #1, #2, #3 and #5, and distributing and supporting the software.

9c.2 Progress in meeting the aims

Numbered references refer to the list of publications in **9c.3**: those included in Appendix materials are in **bold**. References in the software table are given in the literature cited section **9g**.

1. Building on earlier work from 1991-4, we have further developed methods for Markov chain Monte Carlo (MCMC) likelihood analyses of Mendelian traits and multiple linked markers, where data are observed for some members of extended and/or complex pedigrees [14, 18, 19, 26]. Improved MCMC methods involve some form of joint updating. We have developed MCMC Bayesian methods for analysis of data on extended pedigrees; in this approach genotypes of all individuals at a given locus are updated jointly conditional upon the current genotype configuration at other loci. The methods developed have been applied to oligogenic genetically heterogeneous traits [**6**], traits with delayed onset [1, **24**], and to genetically determined quantitative risk factors of complex genetic diseases [8, 23]. We have developed methods for finding feasible genotypic configurations on extended complex pedigrees, with a high proportion or complex pattern of unobserved phenotypes [5, 7]. Resolution of complex traits is aided by data on extended pedigrees, at multiple linked markers, for quantitative risk factors [**20**]. Together, our MCMC methods provide the tools that can analyze these data.
2. In analysis of censored (age-of-onset) data we have adopted two approaches. One models disease onset with a genotype-dependent proportional-hazards risk function [10, **11**], while the other treats age-of-onset as a quantitative trait censored by the current ages of individuals [1, **24**]. The latter approach permits multipoint linkage analysis and an oligogenic trait [**6**], while the former allows incorporation of familial effects and other covariate data [10]. We have addressed questions of carrier detection under both approaches [2, **11**]. We have also extended methods to analyses of a disease trait with three ordered categories which are the expression of an underlying quantitative liability [23]; this approach will be extended to more general ordered categorical phenotypes.
3. The latent variables in a pedigree analysis may be considered either the genotypes of individuals (equation (1)), or the inheritance vectors (equation (3)) which specify the multilocus segregation of each meiosis [26]. We have further developed MCMC methods for the analysis of gene ancestry in an extended complex pedigree, using both samplers of ancestral genotypes [4] and samplers of inheritance vectors [17]. Using improved multilocus sampling of inheritance vectors, these same approaches have been used to compute posterior probabilities of gene identity-by-descent (IBD) jointly in multiple relatives, conditional on data on these or other members of the pedigree [**27**]. Such analyses of gene IBD in affected relatives, jointly over loci and individuals, provide greater power than (for example) pairwise analyses, particularly for heterogeneous or complex traits [**15**]. Analysis of gene IBD in affected relatives can serve not only as a method of linkage detection, but also as a method for detecting the pedigrees

segregating a putative gene for a heterogeneous oligogenic trait [23]. Multilocus sampling of inheritance vectors is an effective MCMC approach with multiple tightly linked loci, but the method can perform poorly on pedigrees with extended ancestral paths. Joint updating of genotypes for all individuals, but at a single locus, is very effective on extended pedigrees, but can perform poorly where loci are very tightly linked. Preliminary work on combining these two MCMC approaches shows promise for problems involving both extended pedigrees and tightly linked loci [9].

4. Methods developed under items 1, 2, and 3, have been tested on simulated data, and on real data where the truth is known. Comparisons with other methods have been made, using data simulated under a variety of models, both for questions of parameter estimation [**12**] and model selection [**13**, 21, 28]. It is found that, for segregation analyses of quantitative traits, MCMC estimation of likelihood ratios between alternative models is an effective model selection procedure [**13**]. Monte Carlo methods for genome-wide linkage analyses of oligogenic traits have also been tested, and have performed well, on simulated data from Genetic Analysis Workshops [**6**, 8], and on Alzheimer's disease age-of-onset data [1, **24**].

5. Our work under this aim has diverged slightly from that originally proposed, and is a main focus for the final currently funded year of our research. Using ideas of gene IBD within the ancestral coalescent of a sample of disease haplotypes, we have developed methods for Monte Carlo likelihood analysis of multiple tightly linked markers for fine-scale mapping [16, **25**]. We have applied our MCMC methods to problems of haplotype estimation with multiple tightly linked markers, but the results raise questions regarding the assumptions and models for marker maps. For denser maps, interference or apparent interference resulting from heterogeneity in recombination frequencies impacts posterior probabilities of gene sharing among relatives, conditional upon marker data. Allelic associations among markers also impact multilocus likelihoods. Lack of data for accurate estimation of marker recombination frequencies and marker allele frequencies reduce the power to detect genes for complex traits and for fine-scale mapping of detected genes [3, 22]. Working with a student supported under a fellowship, we have developed Monte Carlo likelihood methods for analyses on the basis of a continuous genome [17], and used these to analyze genome sharing between relatives (Browning 1998). We are using the same approach to analyze models of position and count interference, and to use such models in haplotyping and fine-scale mapping.

6. Compressed tar files of the source code of all software developed under this award are made available to interested researchers via the web and anonymous ftp. The web address is *http://www.stat.washington.edu/thompson/pangaea.html* and the direct anonymous ftp address is *ftp://ftp.u.washington.edu/pangaea/PANGAEA*. The site contains all our available software. MORGAN (programs for MOnte caRlo Genetic ANalysis) and **Loki** are the parts developed under this grant. The table on the next page summarizes the available software; **Loki** is a stand-alone program while the other programs are parts of the MORGAN package.

Currently, there are two versions of the MORGAN package implementing the programs for Monte Carlo Genetic Analysis. MORGAN 1.1 (updated version) was released July 1997: this version contains a larger array of programs, but is not user-friendly. MORGAN 2.0.1 (released February 1998) provides a more friendly user-interface, but does not yet contain the full array of programs: progress has been impeded by the serious illness of our staff programmer. Additionally available is the package **Loki** for the Monte Carlo Analysis of multiple QTL, developed by postdoctoral research associate Simon Heath (most recent release **Loki 2.1.4**, March 1998). Both MORGAN and **Loki** have been used by other researchers. Outside of our own group and close collaborators, two groups at Genetic Analysis Workshop 11 (1998) used MORGAN as a primary tool in analysis and one other group used **Loki**. Many others researchers have accessed the PANGAEA web page and downloaded the programs. User email queries are increasing.

Package	Calculates/Models	Technique	Main references
PolyEM	polygenic model with additive effects and multiple traits	deterministic EM equations	Thompson and Shaw (1990, 1992) Heath et al. (1997)
Genedrop	full genome marker data and oligogenic quantitative traits	simulation of test data	
MIXED	mixed model (major Mendelian genes + polygenic additive effects)	Monte Carlo EM using Gibbs sampler	Thompson and Guo (1991) Guo and Thompson (1994) Thompson (1994c)
Multipnt	Mendelian quantitative trait model with Mendelian markers	lod score multipoint linkage analysis using MCMC	Lin et al. (1993,1994) Lin (1996)
Loki	Mendelian quantitative oligogenic trait model with Mendelian markers	fully Bayesian MCMC QTL analysis	Heath et al. (1997) Heath (1997)
PBGibbs	Mendelian qualitative trait model	Monte Carlo likelihood using pseudo-Bayes MCMC sampler	Geyer and Thompson (1995)
Autozyg	multiple Mendelian marker loci, using meiosis indicators	Monte Carlo estimation of multilocus gene IBD probabilities	Thompson (1994a, 1994b, 1998) Thompson and Heath (1998)

9c.3(i) Published papers, 1995-8.

1. Daw, E.W., Heath, S. C. and Wijsman, E. M. (1997) Multipoint oligogenic analysis of age-of-onset data with applications to large Alzheimer's disease pedigrees. *American Journal of Human Genetics* **61:** *Suppl.* A273.

2. Daw, E.W., Thompson, E.A., and Wijsman E.M. (1998) Identifying gene carriers for heterogeneous and complex traits. *Genetic Epidemiology* **15:** 537.

3. Daw, E.W., Thompson, E.A., and Wijsman E.M. (1998) Bias in multipoint linkage analysis arising from map misspecification. *American Journal of Human Genetics* **63:** (Suppl.) A17.

4.* Geyer, C. J. and Thompson, E. A. (1995) Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference. *Journal of the American Statistical Association* **90:** 909–920.

5. Heath, S. C. (1996) Generation of valid genotypic configurations for multi-allelic loci and complex pedigrees. *American Journal of Human Genetics* **59S:** A383.

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

7. Heath, S. C. (1998) Generating consistent genotypic configurations for multi-allelic loci and large complex pedigrees. *Human Heredity* **48:** 1–11.

8. Heath, S. C., Snow, G. L., Thompson, E. A., Tseng, C., and Wijsman, E. M. (1997) MCMC Segregation and linkage analysis. *Genetic Epidemiology* **14:** 1011–1016.

9. Heath, S.C., and Thompson, E. A. (1997) MCMC samplers for multilocus analyses on complex pedigrees. *American Journal of Human Genetics* **61:** *Suppl.* A278.

10. Li, H., and Thompson, E. A. (1997) Semiparametric estimation of major gene and random environmental effects. *Biometrics* **53:** 282–293.

11.* Li, H., Thompson, E. A. and Wijsman, E.M. (1998) Semiparametric estimation of major gene effects for age of onset. *Genetic Epidemiology* **15:** 279–298.

12.* Olshen, A. B. and Wijsman, E. M. (1996) Pedigree Analysis Package vs. MIXD: Fitting the mixed model on a large pedigree. *Genetic Epidemiology* **13:** 91–106.

13.* Snow, G.L. and Wijsman, E. M. (1998) Pedigree analysis package (PAP) vs. MORGAN: model selection and hypothesis testing on a large pedigree. *Genetic Epidemiology* **15:** 355–369.

14. Thompson, E. A. (1996) Likelihood and linkage: from Fisher to the future. *Annals of Statistics* **24:** 449–465.

15.* Thompson, E. A. (1997) Conditional gene identity in affected individuals. In: *Genetic Mapping of Disease Genes,* I. H. Pawlowitzki, J. H. Edwards and E. A. Thompson (eds.). Pp. 137–146. Academic Press: London.

16. Thompson, E. A. and Neel, J. V. (1997). Allelic association and allele frequency distribution as a function of social and demographic history. *American Journal of Human Genetics* **60:** 197–204.

17. Thompson, E. A. (1998) Inferring gene ancestry; Estimating gene descent. *International Statistical Review* **66:** 29–40.

18. Wijsman, E. M. and Lin, S. (1995) Monte Carlo methods in genetic epidemiology; opportunities and problems. *American Statistical Association 1995: Proceedings of the Epidemiology Section.* Pp. 38–47.

19. Wijsman E. M. (1997) Monte Carlo methods and model selection in genetic analysis. *Animal Biotechnology;* **8:** 129–144.

20.* Wijsman, E. M. and Amos, C. I. (1997) Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: Summary of GAW10 contributions. *Genetic Epidemiology* **14:** 719–735.

21. Wijsman, E. M. and Snow, G. L. (1996) Selection among segregation analysis models with Monte Carlo Markov chain methods. *American Journal of Human Genetics* **59S:** A193.

9c.3(ii) Technical reports, and papers submitted or in press.

22. Daw, E.W., Thompson, E.A., and Wijsman E.M. (1998) Bias in multipoint linkage analysis arising from map misspecification. Submitted.

23. Daw, E.W., Kumm, J., Snow, G.L., Thompson, E.A., and Wijsman E.M. (1999) MCMC methods for genome screening. *Genetic Epidemiology:* submitted.

24.* Daw, E.W., Heath, S.C., and Wijsman, E. M. (1998) Multipoint oligogenic analysis of age-of-onset data with applications to Alzheimer's disease pedigrees. Submitted.

25.* Graham, J., and Thompson, E. A. (1998) Disequilibrium likelihoods for fine-scale mapping of a rare allele. *American Journal of Human Genetics* **63:** 1517–1530.

26. Thompson, E. A. (1995) Monte Carlo in Genetic Analysis. Technical report No. 294. Department of Statistics, University of Washington.

27.* Thompson, E.A., and Heath, S. C. (1998) Estimation of conditional multilocus gene identity among relatives. *IMS lecture note series,* F. Seillier-Moseiwitch, P. Donnelly and M. Waterman (eds): in press.

28. Wijsman E. M. (1998) Monte Carlo methods and model selection in genetic epidemiology. *Computational Statistics and Data Analysis:* in press.

Five collated sets of copies of the ten items marked * are submitted as Appendix material.
(List of publications not included in 25-page limit.)

9d: Research Design and Methods

9d.1 Multilocus gene mapping on extended pedigrees.

(i) IBD methods for linkage detection and gene resolution

A frequent current approach to linkage detection for complex traits is to score, for pairs of affected individuals, posterior probabilities of gene identity by descent (IBD) at gene locations across the genome, conditional on marker data (see **9b.2**). Patterns of gene IBD are a simple function of the meiosis indicators $\mathbf{S} = \{S_{ij}\}$ of equation (2) in **9b.4**. Our current MCMC methods thus allow accurate estimation of conditional probabilities of patterns of gene IBD, jointly among up to eight haplotypes (e.g. in affecteds) from data $\mathbf{V} = (V^{(0)}, \ldots, V^{(K)})$ at $K+1$ multiple linked marker loci $0, \ldots, K$ observed for members of a pedigree. The pedigree may be extended and/or complex, and the individuals for whom marker data are observed may or may not be those among whom the IBD patterns are of interest (Heath and Thompson 1997; Daw et al. 1998a).

Thus, first, our programs allow accurate estimation of IBD probabilities, in cases where exact multilocus computations are infeasible due to the size of complexity of relationships between pairs of affected individuals. Remote relatives, where available, provide more power for linkage detection. At 12-14 meioses distance, the prior probability of sharing any segment of genome is small (Donnelly 1983); thus any evidence of IBD among affected relatives is evidence of genes contributing to the trait. However, since the prior probability of IBD is small, a single locus cannot provide strong evidence for IBD. Moreover, segments of IBD will be small; data at multiple tightly linked markers must be assessed, but assumptions of perfect infinitely dense information leads to unacceptably conservative procedures (Kong and Cox 1997). Accurate estimates of exact multilocus IBD probabilities are necessary.

In linkage detection and fine-scale mapping there are questions of *specificity* and *scale* (Thompson 1997). Sibs share more of their genome, and may be more homogeneous in the genetic causes of their shared traits. However, the genome segments shared by the members of a given sib pair are large, and there may be considerable heterogeneity among the very large number of sib pairs required for gene mapping. More remote relatives allow finer genetic resolution, and the potentially much larger number of affected in an extended pedigree may insure greater genetic homogeneity among affecteds. However, for remote relatives, the segments of genome shared will be too small to be readily detected, and within-pedigree trait heterogeneity may cause problems. Using estimates provided by our programs, we propose to examine the balance between closer and more remote relatives, depending on the extent of genetic heterogeneity, and the density of the available genetic marker map.

Use of data on many affected individuals within a single pedigree structure not only provides more power due to increased trait homogeneity among affecteds, but also allows use to be made of patterns of IBD among several relatives jointly. Scoring of posterior probabilities of IBD patterns jointly in several affected individuals, using data on these and/or other pedigree members at multiple linked marker loci, provides both greater specificity and greater sensitivity for linkage detection and gene resolution, relative to tests based only on pairs of relatives (Thompson 1997). Nor is IBD scoring limited to affected individuals; posterior probabilities of gene IBD among several affecteds, jointly with non-sharing with specified non-affecteds is possible. Again, using our MCMC methods of estimation of IBD and non-IBD patterns, jointly over loci and among individuals, conditional on marker data on an extended pedigree, we propose to investigate this approach to linkage detection for complex traits. We will investigate the use of IBD jointly among several individuals compared with the use of the data on each pair. We will examine the extent of joint IBD among affecteds to be expected for a heterogeneous trait; statistics that score only extreme patterns of sharing will fail to detect important genes. We will investigate scoring and use of IBD patterns involving simultaneously IBD among like phenotypes, and non-IBD with contrasting phenotypes. We will develop test statistics making use of the dependent IBD scores at several loci jointly.

In addition to providing MCMC estimates of linkage likelihoods (Thompson 1994b) and of gene IBD statistics for linkage detection (Thompson 1997, 1998), MCMC estimates of gene descent patterns **S** can address the question of which pedigrees segregate a putative gene influencing a heterogeneous or oligogenic trait. This provides better resolution of the trait model, and more precise estimates of trait gene locations (Boehnke and Moll 1989; Levy-Lahad et al. 1995a). Samplers realizing genotypes are well suited to the detection of individuals carrying putative trait genes, and have been used for this purpose (Geyer and Thompson 1995; Li et al. 1998), but a meiosis sampler provides direct estimates of gene IBD probabilities. Gene IBD jointly among affecteds, and jointly over loci, will discriminate those pedigrees in which there is excess gene sharing among affecteds, at specified locations in the genome. We will investigate the potential of this approach both on simulated data for a quantitative oligogenic trait, and for pedigrees segregating a heterogeneous trait where the true gene status in different pedigrees is known (Levy-Lahad et al. 1995a).

(ii) Improved MCMC sampling for gene descent patterns

MCMC sampling methods have improved dramatically in recent years, but sampling efficiency and reliability of MCMC estimates remains a problem for marker data **V** at tightly linked multiple markers, on large complex pedigrees, and for complex trait models (Section **9b.4**). MCMC methods have used both genotypes $\mathbf{G} = \{G_{ij}\}$ (equation (1)) and meiosis indicators $\mathbf{S} = \{S_{ij}\}$ (equation (3)). The early samplers were often single-site samplers; each variable S_{ij} or G_{ij}, for each meiosis or individual i and each locus j, is updated in turn conditionally upon the data and values of the other latent variables.

By combining exact methods of computation with sampling based methods, much more efficient block-updating schemes have been developed. One example of such a block-updating scheme is that used by **Loki** (Heath 1997), where genotypes for all individuals i at each given locus j are updated jointly (the locus or "L-sampler"). This may be computationally infeasible on very complex pedigrees, and mixing of this MCMC sampler is poor when loci are very tightly linked. A second block-updating scheme is our whole-meiosis Gibbs-sampling of inheritance vectors used by the **MORGAN/Autozyg** program (Thompson and Heath 1998). In this meiosis or "M-sampler", indicators S_{ij} at all loci j but at a single meiosis i are updated jointly. This sampler works well even for tightly linked loci, but has problems on very extended pedigrees, with large numbers of observed individuals, or extended ancestral paths of unsampled individuals.

The L-sampler and the M-sampler are in many ways orthogonal. Not only do they have different situations under under which computation becomes intensive, and under which mixing is poor, but also quite different situations under which there is irreducibility. Generally, irreducibility is less of a problem for the meiosis sampler (Thompson and Heath 1998), but either sampler may be reducible (Sobel and Lange 1996). By running both samplers, much increased confidence in the validity and reliability of results is obtained when these agree, and useful avenues for improvement are suggested when they do not. We therefore propose to implement the full range of linkage likelihood and gene-descent probability Monte Carlo estimation procedures under both sampling approaches, in order to investigate and improve the reliability of results.

A further approach to obtaining the benefits of both these samplers is to combine them. The steps of the MCMC process may be either an "L-step" of the type used in **Loki** or an "M-step" of the type used in **MORGAN/Autozyg**, at each stage a cycle of L-steps or of M-steps being chosen with some probability. Initial experiments with this process showed promise on an example (Heath and Thompson 1997). We propose to continue to to improve these samplers, to explore systematically their properties, and to make recommendations as to the correct mixing proportion, dependent upon pedigree structure, proportion and pattern of observed individuals, and marker map density.

The whole-meiosis M-sampler is a Gibbs sampler. In order to update meiosis indicators $S_{i*} = (S_{ij}, j = 0, \ldots, K)$ for given meiosis i but jointly for all loci j, the method uses a forwards-compute, backwards-sampling approach detailed by Thompson and Heath (1998). If $V^{(j)}$ denotes the marker data pertaining to locus j, with the loci ordered $j = 0, \ldots, K$, successive computation of the probabilities $Q_j(s)$, $s = 0, 1$ is required, where $Q_j(s)$ is the probability that $S_{ij} = s$ conditional on $V^{(1)}, \ldots, V^{(j)}$ and on indicators S_{l*} at meioses $l \neq i$. Implementation is computationally efficient, requiring storage only of the two values $Q_j(0)$ and $Q_j(1)$ for each locus j. The M-sampler is readily improved, by joint updating of several meioses. A single chain of several meioses may connect an ancestor and descendant in a pedigree. Such meioses can be treated as a single compound meiosis, by appropriate modification of the probabilities. More generally, in place of $Q_j(s)$, $s = 0, 1$, we will require $Q_j(\mathbf{s})$ where $\mathbf{s}$ is a binary vector for the 2^m values at locus j of the m meioses being updated. Since $2^{10} \approx 1000$, any set of up to about ten meioses could be updated jointly quite efficiently. Conditionally upon the inheritance patterns in other meioses, the computations and sampling for the set to be updated can be performed by an algorithm analogous to the exact computational methods of Lander and Green (1987) and Kruglyak et al. (1996). In this case also, we will explore systematically the properties of alternative joint updating schemes, and develop recommendations as to which sets of meioses should be jointly updated.

(iii) Haplotype estimation, population associations and fine-scale mapping

Once linkage is detected, haplotype analysis both at the pedigree and population levels, is often a key step in refining estimates of gene location (Goddard et al. 1996; Cox et al. 1989; Snell et al. 1989; Theilmann et al. 1989). Algorithms for obtaining point estimates of haplotypes are well developed (Sobel et al. 1996), but MCMC permits realizations from the joint conditional probability distribution of haplotypes on a pedigree. Given marker data and a realization of meiosis indicators $\mathbf{S} = \{S_{ij}\}$ conditional upon those data, the allelic types of all founder genes present in observed individuals are determined, and the multilocus haplotypes on a pedigree are readily imputed. MCMC has been used to estimate haplotypes (Sobel and Lange 1996), but effective mixing of the MCMC sampler is critical, if pedigree haplotypes are to be properly realized. Feasible initial configurations may be found by determining obligatory recombinants, but if the sampler mixes poorly the extent of allelic disequilibria in the pedigree haploptypes may be inaccurately estimated. The improved samplers of **9d.1(ii)** above are a prerequisite for successful MCMC haplotype analyses on extended pedigrees.

As maps become increasingly dense, analyses of genetic data are impacted by allelic associations at the population level, among marker loci and between markers and traits. Using improved MCMC samplers, we will analyze the distributions of haplotypes within pedigrees, given marker data on some pedigree members at multiple tightly linked loci. We will study the information available in marker data on pedigrees to infer allelic associations in the population contributing the founder haplotypes. We will analyze the effects of ignoring allelic associations in founders on the haplotypes imputed in a pedigree and the estimated recombination events. We will study the relationship between map density and the pedigree extent at which allelic associations can be inferred, and between map density and the effects of ignoring founder allelic associations. We will study the patterns of descent of multilocus haplotypes in the very large extended pedigrees of entire populations, and investigate the role of such pedigree structures in maintaining marker disequilibria. We will combine the estimation of multilocus haplotypes and the imputation of recombination break-points within extended pedigrees, with the analysis of coancestry among different pedigrees segregating a trait.

All these approaches, and proposed developments, assume knowledge of marker map and allele frequencies. The effects of uncertainty in these parameters will be considered in section **9d.4**.

9d.2 Resolution of complex genetic traits

(i) Ordered categorical phenotypes from latent liabilities

MCMC approaches to multipoint detection of oligogenic QTL have been developed by several groups (Satagopan et al. 1996; Heath 1997; Uimari and Hoeschele 1997). While using a quantitative trait is usually preferable to using a discrete one (Wijsman and Amos 1997), quantitative measures are not always available. However, more information than binary affectation status often is available. For example, degree of severity of affectation would be an ordered categorical phenotype. Several research groups have developed MCMC approaches to segregation analysis of ordered categorical phenotypes, using a polygenic model for an underlying quantitative liability with categorical phenotypes determined by a series of thresholds (Hoeschele et al. 1995; Sorensen et al. 1995; Wang et al. 1997; Korsgaard et al. 1997; van Tassell et al. 1998). We propose to extend the MCMC joint segregation and linkage analysis of oligogenic QTL (Heath 1997) to the case of ordered categorical phenotypes determined by an underlying quantitative liability.

Both polygenic and oligogenic models allow for both genetic and environmental sources of variance. Whereas a polygenic model provides a normal distribution of genetic effects and hence of phenotype (Falconer 1995), the oligogenic model of (Heath 1997) assumes a finite but unknown number of genes contributing to the trait value. Each diallelic QTL has three genotypic means, specified via additive and dominance effects, and the loci contribute additively to the quantitative trait. This results in a phenotypic distribution which is a sum of normal distributions.

An ordered categorical phenotype, D, is a discrete measure with m categories, on which an order can be placed. One way to treat ordered categorical values, would be to model directly the quantitative values $1, \ldots, m$. However, this imposes distributional assumptions under which MCMC methods may not perform well. A better approach is to model an underlying unobserved quantitative liability, Z, which then determines the categorical phenotype via a series of thresholds, $t_1, \ldots, t_{m-1}$. Since the location (mean) and scale (variance) of the QTL distribution are estimated by the MCMC methods, the placement of up to two threshold values is arbitrary.

In preliminary work to extend the MCMC methods to joint linkage and oligogenic segregation analysis of ordered categorical phenotypes for the case $m = 3$, using the COGA alcoholism data for GAW 11, promising results were obtained (Daw et al. 1998a). In addition to sampling the genetic parameters and QTL genotypes by MCMC (Heath 1997), we treat the liability Z as a latent quantitative oligogenic trait, and sample also values of Z consistent with the observed ordered categorical phenotypes, D.

First, we propose to extend our method for the case $m = 3$, to sample underlying liabilities also for a binary trait $m = 2$. We will test performance on real and on simulated data. We will explore the additional information in a 3-category phenotype, over a binary trait, for joint segregation and linkage analysis, in terms of detecting and localizing QTL. We will evaluate bias and variance in the estimation of genetic parameters. For segregation analysis of a binary trait with a polygenic model for latent liability, bias has been documented (Moreno et al. 1997). We will explore also the gain of genetic marker information in refining estimates of trait-model parameters.

Arbitrary placement of additional thresholds beyond two implies assumptions about the relation between the liability, Z, and the observed phenotype, D, that may not be justified. To develop methods for the multi-category case ($m > 3$), methods are needed for placing these additional threshold values, t_k. The methods, cited above, for polygenic segregation analysis, use Monte Carlo sampling of threshold values. We may be able to adapt these methods, but this may be difficult for the more complex oligogenic model, especially when the segregation analysis is coupled with linkage analysis. This approach may entail estimating too many parameters, both in terms of information available in the data, and in terms of computational efficiency. Thus, we will also compare the results and efficiencies of estimating and of fixing the thresholds. For example, thresholds could be fixed as constant-width intervals on the liability scale ($t_k - t_{k-1} = t_l - t_{l-1}$), or each width could be proportional to the data-proportion in that category ($t_k - t_{k-1} \propto P(D = k)$). Although fixed threshold values are unlikely to work in every case, there may be many cases where they work well.

Once the methods are implemented, we will test them on data sets with ordered categorical phenotypes (see **9d.5**). In addition, once the case of general m is implemented, we will test it as a method to dissect complex traits in cases where the distribution of a quantitative phenotype, Y, seriously violates the assumptions of the QTL model. In such cases, it may be more effective to polychotomize this phenotype, to obtain an ordered categorical phenotype D_Y, and to analyze this phenotype D_Y using an oligogenic latent liability Z.

(ii) Missing covariates and censored phenotypes

Use of covariates is a factor that has been identified as helpful in the genetic dissection of complex traits (Wijsman and Amos 1997). Missing covariate information is a major difficulty for many analysis methods. MCMC methods are well suited to data with missing values, since these may be imputed provided some probability model for them is specified. In the current version of **Loki**, major gene covariates with missing values are permitted. In this case imputation of missing values is straightforward, following the same method as is used to impute missing marker data and estimate QTL effects. For general covariates, the problem is thus to specify a probability distribution over a pedigree. No single method will be appropriate: some covariates are discrete, some continuous, some may be strongly familial, others may be related to other covariate information such as age. Generally, most potential covariates suggest some reasonable way to impute missing values. We will implement a variety of these methods so a wider array of covariates that have missing values can be used in analyses.

The problem of missing covariates is similar to that of censored phenotypes or covariates. Again, given a probability model, MCMC methods can be used to impute the censored variable, conditional on the observed data and current parameter values. We have used this approach to treat age-of-onset as a right-censored quantitative trait (Daw et al. 1998b). We propose to extend this approach to more general censored phenotypes and covariates.

Exposure is a type of covariate that can have additional complexity. While some measures of exposure are continuous or categorical covariates, some phenotypes require exposure for expression. For example, an individual cannot become an alcoholic without exposure to alcohol (COGA; see **9d.5**). In such cases, an exposure variable imposes censoring. Our approach to general censored phenotypes and covariates will encompass this case.

(iii) More flexible oligogenic models for complex traits

Currently (Heath 1997), the oligogenic models used in our QTL analyses are diallelic, and loci act additively on a single quantitative trait. While this has proven sufficiently flexible for successful analyses, we will investigate the gains of relaxing these assumptions slightly, and the robustness of the methods under alternative model generalizations.

Heath (1997) found difficulty in localizing a tri-allelic QTL. To localize multi-allelic QTL more effectively, we will examine methods to relax the assumption of diallelic QTL. We propose a two-stage approach, with the first stage consisting of analysis with the current (diallelic) QTL model. For each putative QTL detected, the second stage would estimate the number of alleles and their effects. The MCMC procedure at this second stage will propose changes in the number of alleles, effect sizes, and small changes of location, but will condition on existence of a QTL in this general location. Other QTL will be updated as before.

We will consider other generalizations of the underlying genetic model similarly. Rather than broadening the model and attempting simultaneous estimation of unreasonable numbers of parameters, we will use the current methods to detect putative QTL, but then allow such QTL to have more general effects. For example, two detected QTL could interact non-additively in their mean effects, or show covariance in their residual variability (individual effects). We will explore the effectiveness of our MCMC methods, both in detecting interacting QTL, and in refining estimation of the genetic parameters.

(iv) Multiple correlated phenotypes

Use of multiple correlated quantitative traits in human data for segregation analysis (Blangero and Konigsberg 1991) and in linkage studies (Blangero 1995; Almasy et al. 1997; Todorov et al. 1997), can improve trait model estimation and localization of trait loci. Our own methods for multiple traits are limited to polygenic models for normally distributed phenotypes (Heath et al. 1997), although others have used MCMC methods in segregation analyses, under polygenic models, of a variety of correlated continuous and/or categorical phenotypes (Hoeschele et al. 1995; Wang et al. 1997; van Tassell et al. 1998; Korsgaard et al. 1997).

We propose to extend our MCMC combined segregation and linkage analysis using oligogenic models to handle correlated phenotypes for general pedigrees. Initially we will consider two quantitative phenotypes. The estimated QTL may have effects on either trait, or on both. Such pleiotropic loci induce genetic and residual covariance between the traits. We will investigate the feasibility and effectiveness of this approach. If successful, we will extend the procedures to additional traits, and to categorical traits with phenotypes determined by an underlying liability. By using multiple correlated binary phenotypes, one can analyze any categorical phenotypes. For example, the ALS-PD syndrome of Guam (see **9d.5**) has four phenotypes: individuals may exhibit only ALS symptoms, or only PD symptoms, or neither, or both (Plato et al. 1969). There is no clear order on the three affected phenotypic classes. Using two correlated binary variables would enable us to incorporate the available phenotypic information.

9d.3 Methods of analysis of Monte Carlo output

(i) Calibration of Bayesian linkage results

A missing component in the Bayesian joint linkage and segregation analysis approach is a framework for interpreting the results. Currently we obtain an estimate of the proportion of iterations in which a QTL, possibly with a variance contribution in a particular range, is placed in a particular map interval by the MCMC procedures. We do not yet have methods for turning this information into a posterior probability of linkage or other measure of reliability (Morton 1998) that can be used to help an investigator determine a level of interest in the results. We plan to address this issue in the next project period.

Figure: Results of linkage analysis for a simulated quantitative trait on a map of 4 markers using a single-trait-locus analysis model with VITESSE, and an oligogenic model with **Loki**. A and B: lod scores from VITESSE and $\log_{10} R$ scores from **Loki**, respectively, for a single-QTL model (dashed lines) contributing 83% of the total variance, and for a 2-QTL-model (solid lines) obtained by adding a second, unlinked, QTL to the first simulation, where the second QTL contributes 10% of the genetic variance. The linked QTL is at 25 cM in both cases. C: Results for VITESSE (lod score, solid line) and **Loki** (**Loki** score, dashed line) for a 2-locus simulation model with QTL at 30 and 70 cM, contributing 57% and 14% of the total variance, respectively. The single-locus model for analysis with VITESSE was obtained with MORGAN. Lod scores and **Loki** scores do not measure the same ratios, and should not be expected to give the same numbers, although these results suggest that a **Loki** score of 2 may be approximately equivalent to a lod score of 4.

We plan to evaluate functions of the ratio of observed to expected QTL placements for use in determining the strength of evidence in favor of linkage. We have preliminary results that suggest that this approach is useful (see figure). Let $R = Q_0/Q_E$, where Q_0 is the observed number of QTL placements in some interval and Q_E is the expected number under some null hypothesis. If R is "large" we have evidence in support of linkage. It is also possible to consider restricting these procedures to QTL of a particular range of magnitudes, both in the observed and expected placement fractions. R may be estimated in several ways. For example:
1) $R_1 = Q_0/[wQ_T/L]$ where w is the interval size of interest, L is the genome length, and Q_T is the total number of QTL placements in all iterations in the entire genome. This is a simple measure of the observed to expected placements under the null hypothesis of uniform placement probability (i.e., equivalent to a complete absence of marker data).
2) $R_2 = \sum_{j=1}^{N} a_j r_j$, where j indexes the number of QTL in the model, N is the maximum number of QTL in the model over all iterations, a_j is a weight for iterations with j QTL, and r_j is an estimate of observed to expected placements in all iterations with j QTL. This measure weights differently iterations with different numbers of QTL. The weight a_j could also contain additional weighting for different numbers of QTL. Without this, a single placement in an iteration with 1 QTL would contribute twice as much to the estimate of the ratio as would a single placement in an iteration with 2 QTL. The weight a_j could also be a function of QTL variance or other measure of effect size. If we assume that a_j is proportional to the number of QTL and the observed fraction of iterations with j QTL, then $R_2 = [L/(Iw)] \sum_{j=1}^{n} Q_{0j}$, where I is the total number of iterations. The figure shows results for R_2 in an analysis of simulated data, along with results from an exact likelihood-based linkage analysis. When a second unlinked QTL contributing 10% to the variance is added, the maximum lod score from VITESSE drops by more than 1, while there is only minor effect on the R_2 score from **Loki**. In the case of two linked QTL (figure C), **Loki** successfully localizes both, whereas single-trait-locus analysis cannot.

The null expectations in the approach above are those based on the assumption of no marker data; the distribution of QTL placements is uniform across the genome. An alternative approach would be to generate unlinked marker data as a comparison, and to calculate the denominator in the ratio from the null hypothesis of no linkage. For example, during the MCMC analyses a second virtual chromosome could be included which exactly paralleled in marker characteristics a chromosome on which QTL placements seemed significantly in excess of null expectation. However, no actual marker data would be included for this chromosome. During the MCMC analyses it is possible to generate periodically marker data for this virtual chromosome, in the absence of a linked QTL. For some iterations these simulated data could be used to count QTL placements in the critical region. By repeating this marker simulation/data analysis procedure many times, an estimate of the null expected mean and distribution of the QTL placements can be obtained. Because an additional chromosome would have to be included in the analysis, this latter approach would be more computationally intensive than would that based simply on the observed placed vs. unplaced QTL. However, it may be possible to define some initial threshold ratio based on the simpler analysis which could then be used to indicate need for reanalysis with this modestly more computationally demanding approach.

We will obtain a measure of strength of evidence in favor of linkage by extending the approach for estimating the score distribution under absence of linkage to the case where there is linkage. We will simulate marker data conditional on realized QTL genotypes for QTL placed into an interval of interest, followed by the usual MCMC analyses carried out as if the simulated marker genotypes were the observed data. Many such iterations of simulated marker configurations should supply the score distribution in the presence of a linked QTL. Together with the prior probability of linkage specified by the prior distribution on number of QTL and the map length, it is possible to obtain an estimate of the posterior probability of linkage given a particular score. We will pursue this simulation-based approach to obtaining posterior probabilities of linkage. As this approach is likely to be computationally intensive, we will also use this approach to calibrate the score obtained from **Loki** with lod scores. If a relatively constant relationship is obtained, this would make it unnecessary to always carry out these more intensive computations.

(ii) The gains of oligogenic linkage analysis

The Bayesian MCMC combined segregation and linkage analysis approach that is implemented in **Loki** seems promising for localizing genes contributing to complex traits. Since this approach uses both a more complex trait model and more markers in the multipoint analysis than is possible with most other analysis methods, it would be useful to determine how much of the gain obtained with this method derives from the oligogenic model, and how much from the full multipoint analysis. Multipoint analysis depends on both a higher level of marker phenotyping quality and map accuracy than does analysis with single markers, so it is important to determine whether the information gain of multipoint analysis is sufficient to make worthwhile the necessary multipoint quality assurance. We therefore will determine, for a complex trait, how much of the apparent gain in linkage signal-detecting ability obtained by performing a combined oligogenic trait segregation and multipoint linkage analysis is from use of a better trait model, and how much is from use of large numbers of markers in multipoint analysis.

We will evaluate the contribution of multiple markers to linkage detection with an oligogenic trait model using both simulated and real data. This will allow us to determine the impact of using successively more markers under both idealized recombination as well as in the context of reality, where assumptions about recombination process, genotype quality, and map accuracy may not be correct. For both simulated and real data we will evaluate the score used to detect linkage as a function of aspects of the map used in the analysis. Using approximately uniform marker spacing, we will evaluate the score as a function of the number of markers used. We will also fix the number of markers and evaluate the score as a function of marker spacing. In case of real data, we will focus primarily on genomic regions from which a trait-gene has already been identified (e.g., the Alzheimer's data set **9d.5**). However, we will also use regions to which there is confirmed evidence of linkage. In the case of simulated data, we will use our program GENEDROP to simulate marker genotypes and traits with ≥ 2 underlying loci, one of which is linked to the markers under analysis. For the simulated data we will also repeat each simulation condition many times so that an estimate of the distribution of the effect can be obtained. In all cases (both real and simulated data) analysis will be performed with **Loki**. We predict that if the quality of the maps and assumptions about the recombination process are important, the value of adding markers in a multipoint analysis will be attenuated in the real data examples as compared to the simulated data examples.

We will evaluate the gain of using an oligogenic trait model vs. using multiple markers as follows. We will simulate data on large pedigrees under a two-locus model, with one of the loci linked to the region for which we also generate multipoint marker data. We will then use **Loki** to fit the oligogenic model in the context of the multipoint map data. Given the fitted oligogenic model, we will reanalyze evidence for linkage to a single, closely linked marker using a 2 trait-locus, 1 marker analysis. Also, we will use MORGAN/MIXED (see table of **9c.2**) to fit a single-locus trait model to the oligogenic trait data, and follow this with linkage analyses with single and with multiple markers. We will use TMLINK for 2-trait-locus analyses, and FASTLINK, VITESSE, or MORGAN/Multipnt (**9c.2**) for the 1-trait-locus linkage analyses.

While MCMC oligogenic combined linkage and segregation analysis allows simultaneous initial localization of multiple contributing loci, identification of which genes are segregating in which families facilitates fine-scale mapping. We have developed a method to use the trait locus genotypes simulated in each step of the MCMC analysis to identify potential "gene carriers" for each locus (Daw et al. 1998c). While this method shows good sensitivity, the specificity was poor. To improve this method, we plan to examine restrictions on the trait model parameter values, longer run times, and alternate weighting schemes on the genotype realized in each MCMC iteration.

(iii) Scoring IBD patterns given multilocus marker data

Our MCMC methods sample inheritance vectors, on extended pedigrees, conditionally on marker data at multiple linked marker loci. Thus the primary output of our program MORGAN/Autozyg (see table of **9c.2**) is the set of prior (pedigree-based) and posterior (given marker-data) probabilities of patterns of gene IBD at specified loci in specified haplotypes of individuals. The individuals among whom patterns of IBD are of interest may or may not be those on whom there are marker data (Thompson and Heath 1998). While computation of prior and posterior probabilities of, for example, all 4140 patterns of gene IBD at any given chromosomal location among up to 8 haplotypes is not hard, interpretation of the output is not straightforward. While tests have been developed for pairwise analysis (see **9d.1**), p-values alone are an inadequate summary of the evidence in more complex situations (Morton 1998).

Consider testing which of two probability distributions, f_0 and f_1, have given rise to data. Suppose the prior probability that f_1 gives rise to the data is π, and the posterior probability is Q. Then $Q = f_1\pi/(f_1\pi + f_0(1-\pi))$, or

$$\log(f_1/f_0) \;=\; \log(Q/(1-Q)) \;-\; \log(\pi/(1-\pi)). \tag{4}$$

Thus the "lod score", or strength of evidence, provided by the data is the difference of the logits of posterior and prior probabilities. Given data at informative, tightly-linked, marker loci, Q is often close to 1 or to 0, the IBD states of interest being those for which Q is close to 1. A probability ratio, such as Q/π, is then influenced primarily by the value of π; the logit-value better reflects the impact of the data. We have proposed the logit-difference measure (4) in analyzing posterior (given marker data) and prior (pedigree-based) probabilities of patterns of gene IBD, and preliminary results were encouraging (Daw et al. 1998a). We will develop these logit measures, and undertake simulation studies to investigate the probability distributions of these statistics in the absence and in the presence of trait genes,

An IBD pattern at a given locus is a partition of the scored haplotypes, and thus IBD patterns exhibit a partial ordering. When there are many possible IBD patterns, it is more effective to group them, computing for example the posterior and prior probabilities of IBD states exhibiting at least a certain amount of IBD. For example, in scoring 6 haplotypes at a particular locus, the pattern (1,2,1,3,2,1). This pattern expresses IBD of haplotypes 1, 3, and 6, and of haplotypes 2 and 5, but non-IBD between the sets $\{1,3,5\}$, $\{2,5\}$, and $\{4\}$. Patterns exhibiting greater IBD are (1,1,1,2,1,1), (1,2,1,1,2,1), (1,2,1,2,2,1) and (1,1,1,1,1,1). We will investigate, theoretically and by simulation, the probability distributions of logit-measures based on groups of IBD states exhibiting at least a specific pattern of IBD. We will use our results to develop recommendations for statistics based on groupings of IBD states that extract linkage information, and have high sensitivity and specificity in detecting pedigrees segregating a trait allele at a putative or candidate locus (see **9d.1**).

Currently our MCMC IBD scoring methods output estimated posterior probabilities or logits (4) of patterns of IBD marginally on a locus-by-locus basis, although of course the estimation process uses jointly the marker data at all linked loci. We will investigate scoring patterns of IBD jointly over loci, in combination with joint patterns over several haplotypes of pedigree members. Since different loci exhibit different IBD patterns, there will be many more (joint) patterns of IBD to be considered. Thus our results on effective groupings and orderings of IBD patterns are a prerequisite for this component.

9d.4 Maps, meioses and interference

(i) Error detection, map uncertainty and map heterogeneity

MCMC realization of meiosis indicators **S**, or of multilocus genotypes **G** on a pedigree under a genetic map, from genetic marker data **V**, leads directly to methods for detecting genotype error, and for genetic map estimates. Using the methods of **9d.1**, we will develop the use of multilocus haplotype imputation for detecting probable genotyping errors that are only detectable with multilocus methods. Methods for detection of such errors in general pedigrees will be important as precursors to other multipoint analysis methods. We will also extend our previous MCMC EM approach to parameter estimation (Guo and Thompson 1994) to the estimation of parameters of genetic maps and meiotic processes. Even the best current meiotic maps (NIH-CEPH Collaborative Mapping Group 1992) are based on very limited numbers of meioses particularly when sex-specific maps are desired. We will use the methods from section **9d.1** coupled with improved interference models (see **9d.4.(ii)** below) to develop methods for accurate map estimation on general pedigrees, using many markers. Such methods are likely to be useful for refining maps in critical regions, where valid likelihoods for linkage detection or association analyses are desired. In this case it is important to assess the accuracy and uncertainty in meiotic maps and to be able to combine data sets from several sources, including large pedigrees used for disease-gene mapping.

Using multilocus marker data on several real data sets of extended pedigrees **9d.5**, we will use the MCMC methods of **9d.1** and **9d.2** to investigate the extent and pattern of recombination frequency heterogeneity, and heterogeneity and accuracy in map estimates. We will develop methods for the estimation of more accurate meiotic maps by allowing for such heterogeneity. Using these real data, and also simulated data, we will study the impact of heterogeneity on linkage detection and mapping for complex traits, and on patterns of gene IBD among relatives. Using the MCMC samplers of **9d.1** and **9d.2** we will examine the effect of map error on multilocus haplotyping in extended pedigrees, and the robustness of estimation of conditional probabilities of haplotypes to map and allele frequency uncertainties, and to recombination heterogeneity.

(ii) Interference models and likelihoods based on recombination patterns

The question of the impact of interference on linkage analysis and linkage detection can only be resolved through likelihood computation and data analysis under both interference and non-interference models. We propose to use MCMC and Monte Carlo likelihood methods to greatly extend the feasibility of likelihood computation under such models. If more complex meiosis models are to be considered, a meiosis-based inheritance-vector M-sampler (see **9d.1**) has advantages over a locus-based genotypic L-sampler. Since each meiosis i is either from a male or a female, computation using sex-specific maps is straightforward. Moreover, since all the inheritance indicators for all the linked loci in an entire meiosis are resampled jointly, incorporation of an interference model is also feasible. Before detailing our two proposed approaches, it is necessary first to review very briefly some models for meiosis.

To ensure a biologically feasible interference model, a model of chiasma formation in the chromatid tetrad at meiosis is desirable. We shall make the common assumption of absence of chromatid interference: that is, each chiasma involves two randomly chosen non-sister chromatids, independently of those involved in other chiasmata. In this case, Mather (1938) provides the formula for the recombination probability $\theta(w)$ at genetic distance w:

$$\theta(x) = (1/2)(1 - P(N(w) = 0)). \tag{5}$$

Under the no-interference model (Haldane 1919), chiasmata, and hence crossovers, arise as a Poisson process; the count on a chromosome arm has a Poisson distribution, and conditionally on the count their positions are independently and uniformly distributed (all distributions being in terms of genetic, not physical, distance). Thus, in the absence of chromatid interference, there are, broadly, two classes of interference model: count interference and position interference.

In a count-location model, the count of chiasmata on a chromosome arm is no longer Poisson, but conditional upon the count, they are independently and uniformly distributed. Such models have been considered by (Liberman and Karlin 1984). A simple biologically-based model of count interference is that introduced by Sturt (1976); this requires every chromosome to have at least one chiasma in the underlying tetrad. Conditionally upon this, there is no other interference. One simple consequence of this is that every chromosome must have genetic length at least 0.5 Morgans. The Sturt model has been found to fit existing data well (Weeks et al. 1993).

Position interference models allow for more general meiotic processes; we will consider only those where chiasmata arise as a stationary renewal process (Speed 1996; Lange 1997). A flexible and simple renewal-process model is the chi-square model (Zhao et al. 1995). The Sturt model is also a renewal-process model, with modified distribution to the first chiasma on a chromosome, as also are models giving rise to many of the standard mapping functions, including the map function of Kosambi (1944).

Our first approach pertains to the case where exact computation of probabilities of all patterns of recombination and non-recombination in a set of K marker intervals is practical. Two methods of likelihood evaluation under interference have been proposed: both rely on efficient computation of these probabilities. Weeks et al. (1993) provides an approach for models of count interference, while Lin and Speed (1996) provides a method for the renewal process chi-square models of position interference. For a fixed marker map, it is feasible to precompute and store these probabilities for up to about 12 markers ($2^{11} = 2048$). However, as for any multilocus problem, exact likelihood computation on an extended or complex pedigree remains computationally infeasible. Our MCMC "M-sampler" (section **9d.1**) can, however, be extended. In that sampler, given marker data $\mathbf{V}$ at loci $j = 0, \ldots, K$, meiosis indicators at meiosis i, $S_{i*} = (S_{i,0}, \ldots, S_{i,K})$ are realized from

$$P(S_{i*} \mid S_{l*}, l \neq i, \mathbf{V}) \quad \propto \quad P(\mathbf{V} \mid \mathbf{S}) \; P^{(H)}(\mathbf{S}) \tag{6}$$

where $\mathbf{S}$ is the total set of meiosis indicators for all loci at all meioses of the pedigree, and the superscript (H) denotes the Haldane (no-interference) model. We now continue to use equation (6) as our proposal distribution, and add a Metropolis acceptance step (Metropolis et al. 1953), to provide the correct conditional distribution of S_{i*} under interference (denoted $P^{(I)}$). The required Hastings-ratio α (Hastings 1970) for the acceptance probability $\min(1, \alpha)$ for current $\mathbf{S}$ and proposed $\mathbf{S}^\dagger$ is

$$\alpha \quad = \quad \frac{P^{(I)}(\mathbf{S}^\dagger, \mathbf{V})}{P^{(I)}(\mathbf{S}, \mathbf{V})} \; \frac{P^{(H)}(S_{i*} \mid S_{l*}, l \neq i, \mathbf{V})}{P^{(H)}(S_{i*}^\dagger \mid S_{l*}, l \neq i, \mathbf{V})} \quad = \quad \frac{P^{(I)}(S_{i*}^\dagger)}{P^{(I)}(S_{i*})} \frac{P^{(H)}(S_{i*})}{P^{(H)}(S_{i*}^\dagger)}. \tag{7}$$

This considerable reduction, and consequent ease of computation of the acceptance probability $\min(1, \alpha)$ relies on the independence of segregation patterns at different meioses i (when not conditioned on data $\mathbf{V}$), and of the fact that the probability of data $\mathbf{V}$ given meiosis pattern $\mathbf{S}$ does not depend on the interference process giving rise to $\mathbf{S}$.

We propose to implement this combined sampler of whole-meiosis Gibbs-sampler Haldane-model proposals and interference-model Metropolis acceptances, for a range of interference models for which computation is feasible, including the Sturt (1976) model, and the chi-square renewal model. We will examine the properties of the sampler, and its use in genetic analyses, as described below also for our other proposed MCMC approach.

(iii) Monte Carlo likelihood using chiasmata realizations

The multi-interval generalization of Mather's formula (5) is the key to likelihood computations under interference. For K intervals, there are 2^K patterns of occurrence or non-occurrence of chiasmata; explicit results for the absence of chiasmata in each of the 2^K subsets of the set of K marker intervals are required. Conversion of these probabilities to the probabilities of recombination patterns then requires, in general, a $2^K \times 2^K$ matrix (Lange 1997). The approach of **9d.4(ii)** above may therefore fail for several reasons. The number of intervals K may be too large: many genome scans now available have 20–30 loci on a chromosome, and there is the potential for many more. The exact avoidance probabilities for the chiasma process may not be available. For multiple markers, and some data patterns, or certain interference models, the Metropolis acceptance probabilities (7) may be small. Further, the method supposes a fixed map. If map positions of marker loci are to be estimated, or of trait loci are to be explicitly modeled, the method may become impractical.

The latent variables considered above are the meiosis indicators $\mathbf{S} = \{S_{ij}\}$ in a gamete determining the inheritance vectors in the segregations of a pedigree, and hence patterns of gene identity-by-descent (IBD) among related individuals. The simple conditional independence structure of the components of $\mathbf{S}$ in the absence of interference, was key to the development of efficient MCMC methods (Thompson 1994b; Thompson and Heath 1998). Although interference destroys this structure, it may be partially restored by additional latent variables underlying $\mathbf{S}$. Browning has recently extended the Monte Carlo likelihood approach of her earlier paper (Browning 1998) to chi-square interference models by the addition of such auxiliary variables. Depending on the desired interference model, different auxiliary variables may be required, but for ease of presentation we consider the case where these are the locations of the chiasmata in the chromatid tetrad.

Let $\mathbf{C} = \{C_i\}$ specify the locations of chiasmata in the chromatid tetrad for the meioses i of the pedigree. Under both count-location and renewal-process models, chiasmata are very readily realized from their prior (no-data) distributions, and the probability of a realization can be very quickly evaluated. Moreover, conditional upon $\mathbf{S}$, $\mathbf{C}$ is independent of the data $\mathbf{V}$, while C_i depends only on S_{i*}, which are dependent only through data $\mathbf{V}$. In addition to this hierarchical structure, there is substantial flexibility in the relationship between chiasmata and recombination, since each chiasma involves a resultant gamete with probability 1/2. Thus MCMC sampling of $(\mathbf{S}, \mathbf{C})$ conditionally upon data is feasible. For example, one could use a Gibbs sampler for each S_{i*} conditionally upon $\mathbf{C}$ and data $\mathbf{V}$, and a Metropolis sampler for C_i involving the proposed movement, increase, or removal of chiasmata, with acceptance probabilities dependent on the interference model and current values of S_{i*}.

For Monte Carlo likelihood estimates we may use the equation of Thompson and Guo (1991) which, for the current variables, is

$$\frac{L(\psi)}{L(\psi_0)} = \frac{P_\psi(\mathbf{V})}{P_{\psi_0}(\mathbf{V})} = \mathrm{E}_{\psi_0}\left(\frac{P_\psi(\mathbf{V}, \mathbf{S}, \mathbf{C})}{P_{\psi_0}(\mathbf{V}, \mathbf{S}, \mathbf{C})} \mid \mathbf{V}\right) \tag{8}$$

Thus likelihood ratios between models ψ and ψ_0 from the actual data $\mathbf{V}$ may be estimated as the average of the complete-data likelihood ratios from $(\mathbf{V}, \mathbf{S}, \mathbf{C})$ for the values of $(\mathbf{S}, \mathbf{C})$ realized by MCMC, conditionally upon $\mathbf{V}$ under the model ψ_0. Here the model parameters ψ and ψ_0 could specify alternative interference models, alternative genetic maps, or alternative models for marker or trait data. While these meiosis, map, and marker model parameters impinge directly on the probability distributions of $\mathbf{C}$, $\mathbf{S}$, and $\mathbf{V}$ respectively, the actual data $\mathbf{V}$ provides information for inferences about any or all of these model components.

We will study the use of the proposed samplers of **9d.4(ii)** above and of this section for efficient estimation of meiosis patterns in pedigrees, multilocus marker haplotypes, and linkage likelihoods. We will analyze patterns of meiosis, and the consequences for patterns of genome sharing among relatives, under interference models. Using Monte Carlo estimates of likelihood ratios (equation (8)), we will study the estimation of interference patterns from data on multiple linked loci across a chromosome, and the effect of failure to account for interference on map estimation and on linkage analysis, including for oligogenic traits where trait loci may be linked.

9d.5 Evaluation of methods on real data

Several real data sets are available to us for the purpose of evaluating and extending our methods. Eight of these are summarized in the table below. All these data sets are from studies of complex traits, and include various subsets of the original data. Five have been collected and have already been made available to use (COGA, AD, BEL), or have been promised to us (Hutt, PCA—see letters from Drs. Morgan and Ostrander respectively). Additional data from the COGA study have has also been promised to us (see letter from Dr. Reich). The three remaining data sets are in the process of being collected, and are available through Dr. Wijsman's current collaborations. These data sets are all actively growing; what is currently available is likely to be only a subset of what will be available in the next few years, including additional map locations and/or gene identifications. Among these eight data sets, the existence of several data sets with complete genome scans provides ample comparative material for evaluating issues of map quality and meiosis models (**9d.4**) and their effects on linkage detection.

Data Set	COGA	AD	BEL	Hutt	PCA	Guam	Dys	Aut
Pedigrees								
Large	+	+	+	+	+	+	+	−
Small	+	+	−	−	+	−	+	+
Complex	(+)	(+)	(+)	+	−	+	−	−
Phenotype								
Quantitative	+	−	(+)	+	−	−	+	+
Age-of-Onset	+	+	−	−	+	+	−	−
Ordered Categorical	+	+	−	−	(+)	(+)	+	+
Affected/Unaffected	+	−	−	+	+	+	+	+
Known Covariate	+	−	(+)	−	−	−	+	−
Markers								
Whole Genome	+	−	+	−	+	+	−	+
Partial Genome	−	+	−	−	−	−	+	−
Some dense regions	+	+	−	+	+	−	+	+
Gene status								
Known genes	−	+	−	−	−	−	−	−
Mapped genes	(+)	(+)	−	−	+	−	(+)	(+)

+ : Present in data set

(+) : Present in data set, but is a minor component, or will only be available to us once gene has been mapped, or refers to mapping results which have not yet been confirmed in a second sample.

− : Currently absent in data available to us.

COGA is a study of the genetics of alcoholism. We currently have available (with permission to use) data that were submitted to GAW11, consisting of data from 1214 individuals in 105 pedigrees, with a total of 992 sampled individuals. An extended data set which approximately doubles the available data will also be made available to us (see letter from Dr. Reich). 296 marker genotypes from a ∼13cM density scan are included. There are a large number of phenotypic measures, including several discrete and ordered categorical diagnoses of alcoholism, correlated quantitative traits such as event-related-potentials (ERPs), and covariate data such as age-of-onset and smoking behavior. There are no definitive localizations for relevant genes, but there are a number of implicated regions, most notably chromosome 1 (Reich et al. 1998). This data set will be used to evaluate multilocus IBD patterns and haplotypes (**9d.1**), categorical and multiple correlated phenotypes (**9d.2**), missing and censored covariates (**9d.2**), calibration of MCMC realizations (**9d.3**), and evaluation of genetic maps and meiosis models (**9d.4**).

The AD data set is an ongoing study of the genetics of Alzheimer's disease, which Dr. Wijsman has been studying for over a decade. This data set will be used in a number of situations. Because two disease genes in 13 of the 84 families in this data set are already known (Poorkaj et al. 1998; Levy-Lahad et al. 1995a), this data set will be particularly useful for testing multilocus linkage analysis methods (**9d.1**), calibration of measures to infer the presence of linkage (**9d.3**), and evaluation of meoisis models and map error on linkage detection (**9d.4**). Marker coverage is variable, with some subsets of the pedigrees more extensively typed than others, but genome screening is ongoing and missing data will be collected when possible. Finally, in regions with known genes, marker spacing approaches the centimorgan level, which will be particularly useful in testing haplotype/disequilibrium analysis measures (**9d.1**).

Together with the COGA data set, the marker data from the apoB Elevating Locus (BEL) project is ideal for our investigations of error detection, maps and the multilocus data resulting from the recombination process (**9d.4**). A full genome scan (10cM density, 388 markers) has been carried out by the Marshfield genotyping service on 380 sampled individuals from the total 567 individuals in 15 large pedigrees (up to 85 individuals/pedigree). Unlike the other data sets, this one includes all original marker typing, including any typing errors, and should not include the systematic biases introduced by error correction.

The Hutterite and Guam data sets that are available to us play complementary roles. Each consists of a huge, complex, multi-generation pedigree. The Guam data consists of a neurodegenerative syndrome with two distinctly different phenotypes from the island of Guam, and will include marker and phenotype data from a genome screen (currently in progress) from a single pedigree consisting of >110 individuals selected from a much larger pedigree of more than 5000 individuals. Within this smaller pedigree are currently 11 inbreeding loops. Both the pedigree size and complexity are expected to increase as the data set increases. This pedigree will be useful for methods dealing with complex pedigrees and categorical phenotypes (**9d.2**). The Hutterite genealogical data base consists of >38,000 individuals in one huge pedigree 10-12 generations deep (see letter from Dr. Morgan). In collaboration with Dr. Morgan, we will use this population to study patterns of disequilibria and genome sharing in such a population, and to develop methods of combining pedigree and population studies to fine-map in a founder population and hence narrow the candidate chromosomal regions for genetic risk factors for cardiovascular disease in this population (**9d.1**). Data are available on sib pairs; the pairs are, of course, related (Boright et al. 1998). How much additional power do these relationships among sib pairs contribute? Generally, what is the power in the total Hutterite genealogy for mapping a sparsely distributed trait?

Finally, the prostate cancer (PCA), dyslexia (Dys) and autism (Aut) data sets will provide additional testing grounds for our methods. All three data sets include missing covariate data (**9d.2**). The PCA data set shares similarities with the AD data, except that marker coverage is essentially uniform across subsets of the data. Both fine-scale and coarse mapping data are available on 512 sampled individuals in 141 families. The dyslexia data set currently contains 102 families, more than 500 individuals with phenotype data, and over 30 phenotypic measures, including discrete, quantitative, and categorical phenotypes, many of which are strongly correlated (**9d.2**). Pedigrees are being extended beyond nuclear families. The autism data consist currently of $>$ 179 nuclear families with $\sim$ 250 projected by the end of 2000. This data set includes several discrete and ordered categorical phenotypic measures plus covariate measures, with missing data.

9d.6 Software development

Details of our current software are given in the Progress Report **9c.2**. Under the current award period, we have laid the foundations for a user-friendly version of our MORGAN_2 software. The core subroutine libraries have been thoroughly tested, and input and output formats completely rewritten, with extensive input checking. Thus although only two of our main program directories are released under MORGAN_2.2, the framework data-structures and formats have been established for complex segregation analysis models, jointly with multilocus linkage analyses, using a genomic array of markers. We propose to continue the development of MORGAN_2. The next release will incorporate more of the programs now available under MORGAN_1.1, including the M-sampler of MORGAN/Autozyg and the genotypic Gibbs samplers of MORGAN/MIXED and MORGAN/Multipnt. As code is revised for MORGAN_2, we will enhance capabilities, through dynamic memory allocation, for example. All programs released under MORGAN_2 are, and will be, thoroughly tested, well documented, and released with test data and corresponding output.

As new methods are developed, these have been implemented either in MORGAN_1, or in the stand-alone program **Loki** for Bayesian QTL analyses (Heath 1997). We will incorporate some of the capabilities of **Loki** into MORGAN, both to extend MORGAN and to allow comparison of likelihood and Bayesian MCMC analyses as detailed in section **9d.3**. We will continue to implement and test new MCMC methods as described in **9d.1, 9d.2, 9d.4** within the MORGAN framework. We will continue to make new versions of MORGAN_1 available to interested researchers via anonymous ftp and the web. Programs that prove useful for our own data analyses, or to others, will be promoted to MORGAN_2 status. As the MORGAN_2 structure becomes finalized, implementation of new methods may also be done using MORGAN_2 data structures and subroutine libraries, but we will retain the distinction between our development software (MORGAN_1) and user software (MORGAN_2).

As the new version of the software develops it will continue to be thoroughly documented, with both online help files and PostScript documentation files for hard copy. We will continue to provide response to user queries, and to compile online information on commonly asked questions (with answers where possible!), and on updates to the software.

Our current structure of programs and subroutine libraries is developed under the Unix system, and is best suited to that system. 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. We will continue to maintain easy portability to VMS systems. 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.

9d.7 Timetable and responsibilities

The major segments of our proposed research will proceed in parallel, rather than sequentially; within each segment, development is, of course, sequential. Work on linkage detection for complex traits using IBD methods, and related work on multilocus haplotyping and disequilibrium patterns in extended pedigrees, (section **9d.1**) will be done by Thompson, with a postdoctoral research associate (currently Dr. Kumm). Research Associate Dr. Warwick Daw will undertake the work on complex phenotypes (**9d.2**), under the general supervision of senior investigators Thompson and Wijsman. The studies of maps and meiosis models (**9d.4**) will be directed by Thompson, with a student research assistant (currently Browning).

The development of methods for assessment of MCMC output, and comparison and evaluation of methods (**9d.3**) will be primarily directed by Wijsman, working with research assistant Snow. However, all members of the research team will test their methods by analyses of real and simulated data, as appropriate (**9d.5**). User software is developed by Scientific Programmer Myrna Jewett, under supervision of Thompson (**9d.6**).

In addition to regular meetings with either Thompson or Wijsman, all research assistants, associates, and investigators, meet weekly to discuss ongoing progress. This ensures all participants know and understand developments on related projects and how these may advance their own research, thus maintaining coherent overall progress of the research.

In years 1 and 2, we will focus on developments of our current work. We will improve MCMC methods for sampling multilocus IBD patterns on extended pedigrees and using these methods for linkage detection for complex traits observed on members of extended pedigrees (**9d.1**). We will implement MCMC methods for ordered categorical traits and censored phenotypes, and develop the use of genome scan data for oligogenic linkage and segregation analysis of these traits (**9d.2**). We will undertake comparative analyses of Bayesian approaches and more traditional multilocus lod-score methods, both using MCMC on multilocus data on members of extended pedigrees (**9d.3**). We will continue our investigations of map heterogeneity, and of simple interference models, and their impact on MCMC multilocus linkage analysis methods (**9d.4**). We will also apply these methods to real data (**9d.5**) as needed. By the end of year 2 also, we expect all successful software currently in MORGAN_1.1 to have been implemented in the more user-friendly MORGAN_2 version (**9d.6**).

In years 3-5, we will continue analyses of IBD, focusing on the patterns and lengths of genome IBD, and consequent disequilibria within and between pedigrees (**9d.1**). We will extend censored data approaches to missing and censored covariates, and begin to address multivariate phenotypes (**9d.2**). We will continue evaluation and comparative analyses of different methods (**9d.3**), while testing alternative methods on a variety of complex trait data (**9d.5**). We will extend our MCMC approach to meiosis and map models, allowing for the analysis of data under more complex models (**9d.4**). We will continue the development of our programs, implementing new methods, and developing the user-version of our software (**9d.6**).

9e: 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.

9f: Vertebrate Animals

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

9g: Literature Cited

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. American Journal of Human Genetics 62:1198–1211

Almasy L, Dyer TD, Blangero J (1997) Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. Genetic Epidemiology 14:953–958

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. American Journal of Human Genetics 54:535–543

Bishop DT, Thompson EA (1988) Linkage information and bias in the presence of interference. Genetic Epidemiology 5:107–119

Blangero J (1995) Genetic analysis of a common oligogenic trait with quantitative correlates: summary of GAW9 results. Genetic Epidemiology 12:689–706

Blangero J, Konigsberg LW (1991) Multivariate segregation analysis using the mixed model. Genetic Epidemiology 8:299–316

Bodmer WF, Bailey CJ, Bodmer J, Bussey HJ, Ellis A, Gorman P, Lucibello FC, Murday VA, Rider SH, Scambler P, et al (1987) Localization of the gene for familial adenomatous polyposis on chromosome 5. Nature 328:614–616

Boehnke M (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. American Journal of Human Genetics 55:379–390

Boehnke M, Cox NJ (1997) Accurate inference of relationships in sib-pair linkage studies. American Journal of Human Genetics 61:423–429

Boehnke M, Moll PP (1989) Identifying pedigrees segregating at a major locus for a quantitative trait: an efficient strategy for linkage analysis. American Journal of Human Genetics 44:216–224

Boright AP, Connelly PW, Brunt JH, Morgan K, Hegele RA (1998) Association and linkage of LDLR gene variation with variation in plasma low density lipoprotein cholesterol. Journal of Human Genetics 43:153–159

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination. American Journal of Human Genetics 63:861–869

Brown MS, Goldstein JL (1986) A Receptor-Mediated Pathway for Cholesterol Homeostasis. Science 232:34–47

Brown PO, Hartwell L (1998) Genomics and human disease–variations on variation. Nature Genetics 18:91–93

Browning SG (1998) Relationship information contained in gamete identity by descent data. Journal of Computational Biology 5:323–334

Chapman NH, Wijsman EM (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. American Journal of Human Genetics In Press

Cottingham RW, Idury RM, Schaffer AA (1993) Faster sequential genetic linkage computations. American Journal of Human Genetics 53:252–263

Cox TK, Kerem B, Rommens J, Iannuzzi MC, Drumm M, Collins FS, Dean M, et al. (1989) Mapping of the cystic fibrosis gene using putative ancestral recombinants. American Journal of Human Genetics 45 Suppl:A136

Daw EW, Kumm J, Snow GL, Thompson EA, Wijsman EM (1998a) MCMC methods for genome screening. Genetic Epidemiology Submitted

Daw EW, Thompson EA, Wijsman EM (1998b) Bias in multipoint linkage analysis arising from map misspecification. American Journal of Human Genetics 63:A85

— (1998c) Identifying gene carriers for heterogeneous and complex traits. Genetic Epidemiology 15:537

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, et al (1987) A genetic linkage map of the human genome. Cell 51:319–337

Donnelly KP (1983) The probability that related individuals share some section of genome identical by descent. Theoretical Population Biology 23:34–63

Ebers GC, Kukay K, Bulman DE, Sadovnick AD, Rice G, Anderson C, Armstrong H, Cousin K, Bell RB, et al (1996) A full genome search in multiple sclerosis. Nature Genetics 13:472–476

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

Falconer DS (1995) Introduction to Quantitative Genetics. Longman, Scientific & Technical, Harlow, Essex, England, 4th edn.

Feingold E, Siegmund DO (1997) Strategies for mapping heterogeneous recessive traits by allele-sharing methods. American Journal of Human Genetics 60:965–978

Froguel P, Vaxillaire M, Sun F, Velho G, Zouali H, Butel MO, Lesage S, Vionnet N, Clement K, et al (1992) Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. Nature 356:162–164

Fulker DW, Cherny SS, Cardon LR (1995) Multipoint interval mapping of quantitative trait loci, using sib pairs. American Journal of Human Genetics 56:1224–1233

Gelfand AE, Smith AFM (1990) Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 46:193–227

Geman S, Geman D (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6:721–741

Geyer CJ, Thompson EA (1995) Annealing Markov chain Monte Carlo with applications to ancestral inference. Journal of the American Statistical Association 90:909–920

Gidh-Jain M, Takeda J, Xu LZ, Lange AJ, Vionnet N, Stoffel M, Froguel P, Velho G, Sun F, Cohen D, et al (1993) Glucokinase mutations associated with non-insulin-dependent (type 2) diabetes mellitus have decreased enzymatic activity: implications for structure/function relationships. Proceedings of the National Academy of Sciences USA 90:1932–1936

Gilks WR, Richardson S, Spiegelhalter DJ, eds. (1996) Markov Chain Monte Carlo in Practice. Chapman and Hall, New York

Goddard KA, Yu CE, Oshima J, Miki T, Nakura J, Piussan C, Martin GM, et al. (1996) Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. American Journal of Human Genetics 58:1286–1302

Goldstein DR, Zhao H, Speed TP (1995) Relative efficiencies of chi 2 models of recombination for exclusion mapping and gene ordering. Genomics 27:265–273

Graham J, Thompson EA (1998) Disequilibrium likelihoods for fine-scale mapping of a rare allele. American Journal of Human Genetics 63:1517–1530

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

Guo SW, Thompson EA (1994) Monte Carlo estimation of mixed models for large complex pedigrees. Biometrics 50:417–432

Haines JL, Minassian TM, Bazyk A, Gusella JF, Kim DJ, Terwedow H, Vance PMA, Rimmler JB, Haynes CS, et al (1996) A complete genomic screen for multiple sclerosis underscores a role for the major histocompatability complex. Nature Genetics 13:469–471

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. Journal of Genetics 8:229–309

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250:1684–1689

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

Heath S, Thompson EA (1997) MCMC samplers for multilocus analyses on complex pedigrees. American Journal of Human Genetics 61:A278

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. American Journal of Human Genetics 61:748–760

Heath SC, Snow GL, Thompson EA, Tseng C, Wijsman EM (1997) MCMC segregation and linkage analysis. Genetic Epidemiology 14:1011–1016

Hoeschele I, Tier B, Graser HU (1995) Multiple-trait genetic evaluation for one polychotomous trait and several continuous traits with missing data and unequal models. Journal of the Animal Science 73:1609–1627

Jensen CS, Kjaerulff U, Kong A (1995) Blocking Gibbs sampling in very large probabilistic expert systems. International Journal of Human Computer Studies 42:647–666

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. American Journal of Human Genetics 56:18–32

Kinzler KW, Nilbert MC, Vogelstein B, Bryan TM, Levy DB, Smith KJ, Preisinger AC, Hamilton SR, Hedge P, Markham A, et al (1991) Identification of a gene located at chromosome 5q21 that is mutated in colorectal cancers. Science 251:1366–1370

Kong A, Cox NJ (1997) Allele-sharing models: scores and accurate linkage tests. American Journal of Human Genetics 61:1179–1188

Korsgaard IR, Sorensen DA, Gianola D (1997) Multivariate Mixed Model analysis of censored normal, ordered categorical and quantitative traits using Gibbs sampling and data augmentation - with applications to animal breeding. Ph.D. Thesis (chaper 3), Department of Theoretical Statistics, University of Aarhus

Kosambi DD (1944) The estimation of map distances from recombination values. Annals of Eugenics 12:172–175

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. American Journal of Human Genetics 58:1347–1363

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. American Journal of Human Genetics 57:439–454

Kwiatkowski DJ, Dib C, Slaugenhaupt SA, Povey S, Gusella JF, Haines JL (1993) An index marker map of chromosome 9 provides strong evidence for positive interference. American Journal of Human Genetics 53:1279–1288

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

Lange K (1997) Mathematical and Statistical Methods for Genetic Analysis. Statistics for Biology and Health. Springer Verlag, New York

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

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

Lathrop GM, Ott J (1990) Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. American Journal of Human Genetics 47S:A188

Lazzaroni L (1998) Linkage disequilibrium and gene mapping: an empirical least-squares approach. American Journal of Human Genetics 62:159–170

Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, Yu CE, Jondro PD, Schmidt SD, Wang K, et al (1995a) Candidate gene for the chromosome 1 familial Alzheimer's disease locus. Science 269:973–977

Levy-Lahad E, Wijsman EM, Nemens E, Anderson L, Goddard KA, Weber JL, Bird TD, Schellenberg GD (1995b) Familial Alzheimer's disease locus on chromosome 1. Science 269:970–973

Li H, Thompson EA, Wijsman EM (1998) Semiparametric estimation of major gene effects for age of onset. Genetic Epidemiology 15:279–298

Liberman U, Karlin S (1984) Theoretical models of genetic map functions. Theoretical Population Biology 25:331–346

Lin S (1996) Multipoint linkage analysis via Metropolis jumping kernels. Biometrics 52:299–309

Lin S, Speed TP (1996) Incorporating crossover interference into pedigree analysis using the χ^2 model. Human Heredity 46:315–322

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

— (1994) Finding noncommunicating sets for Markov chain Monte Carlo estimates on pedigrees. American Journal of Human Genetics 54:695–704

Lindblom A, Tannergard P, Werelius B, Nordenskjold M (1993) Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. Nature Genetics 5:279–282

Mather K (1938) Crossing-over. Biological Reviews of the Cambridge Philosophical Society 13:252–292

McPeek MS, Speed TP (1995) Modeling interference in genetic recombination. Genetics 139:1031–1044

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, 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 PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, et al (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266:66–71

Moreno C, Sorensen D, Garcia Cortes LA, Varona L, Altarriba J (1997) On biased inferences about variance components in the binary threshold model. Genetics Selection Evolution 29:145–160

Morton NE (1955) Sequential tests for the detection of linkage. American Journal of Human Genetics 7:277–318

— (1998) Significance levels in complex inheritance. American Journal of Human Genetics 62:690–697

Nachman MW, Churchill GA (1996) Heterogeneity in rates of recombination across the mouse genome. Genetics 142:537–548

NIH-CEPH Collaborative Mapping Group (1992) A comprehensive genetic linkage map of the human genome. Science 258:67–86

O'Connell JR, Weeks DE (1995) The algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nature Genetics 11:402–408

Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM, et al (1994) Mutation of a mutL homolog in hereditary colon cancer. Science 263:1625–1629

Peltomaki P, Aaltonen LA, Sistonen P, Pylkkanen L, Mecklin JP, Jarvinen H, Green JS, Jass JR, Weber JL, et al (1993) Genetic mapping of a locus predisposing to human colorectal cancer. Science 260:751

Pennisi E (1998) A closer look at SNPs suggests difficulties. Science 281:1787–1789

Plato CC, Cruz MT, Kurland LT (1969) Amyotrophic lateral sclerosis-Parkinsonism dementia complex of Guam: further genetic investigations. American Journal of Human Genetics 21:133–141

Poorkaj P, Sharma V, Anderson L, Nemens E, Alonso ME, Orr H, White J, Heston L, Bird TD, Schellenberg GD (1998) Missense mutations in the chromosome 14 familial Alzheimer's disease presenilin 1 gene. Human Mutation 11:216–221

Rannala B, Slatkin M (1998) Likelihood analysis of disequilibrium mapping, and related problems. American Journal of Human Genetics 62:459–473

Rao DC, B. BJK, , Morton NE, Yee S, Lew R (1978) Variability of human linkage data. American Journal of Human Genetics 30:516–529

Reich T, Edenberg HJ, Goate A, Williams JT, Rice JP, van Eerdewegh P, Foroud T, Hesselbrock V, Schuckit MA, et al. (1998) Genome-wide search for genes affecting the risk for alcohol dependence. American Journal of Medical Genetics 81:207–215

Risch N (1990) Linkage strategies for genetically complex traits.. The power of affected relative pairs. American Journal of Human Genetics 46:229–241

Risch N, Giuffra L (1992) Model misspecification and multipoint linkage analysis. Human Heredity 42:77–92

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

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

Sawcer S, Jones HB, Feakes R, Gray J, Smaldon N, Chataway J, Robertson N, Clayton D, Goodfellow PN, Compston A (1996) A genome screen in multiple sclerosis reveals susceptibility loci on chromosome 6p21 and 17q22. Nature Genetics 13:464–468

Schellenberg GD, Bird TD, Wijsman EM, Orr HT, Anderson L, Nemens E, White JA, Bonnycastle L, Weber JL, et al (1992) Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. Science 258:668–671

Schmitt K, Lazzeroni LC, Foote S, Vollrath D, Fisher EM, Goradia TM, Lange K, Page DC, Arnheim N (1994) Multipoint linkage map of the human pseudoautosomal region, based on single-sperm typing: do double crossovers occur during male meiosis? American Journal of Human Genetics 55:423–430

Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, Lin C, Li G, et al (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature 375:754–760

Snell RG, Lazarou LP, Youngman S, Quarrell OW, Wasmuth JJ, Shaw DJ, Harper PS (1989) Linkage disequilibrium in Huntington's disease: an improved localisation for the gene. Journal of Medical Genetics 26:673–675

Sobel E, Lange K (1996) Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. American Journal of Human Genetics 58:1323–1337

Sobel E, Lange K, O'Connell JR, Weeks DE (1996) Haplotyping algorithms. In T Speed, MS Waterman, eds., *Genetic Mapping and DNA Sequencing*, vol. 81 of *IMA Volumes in Mathematics and its Applications*, 89–110. Springer-Verlag, New York

Sorensen DA, Andersen S, Gianola D, Korsgaard IR (1995) Bayesian inference in threshold models using Gibbs sampling. Genetics Selection Evolution 27:229–249

Speed T (1996) What is a genetic map function? In T Speed, MS Waterman, eds., *Genetic Mapping and DNA Sequencing*, vol. 81 of *IMA Volumes in Mathematics and its Applications*, 65–88. Springer-Verlag, New York

St.George-Hyslop PH, Tanzi RE, Polinsky RJ, Neve RL, Pollen D, Drachman D, Growdon J, Cupples LA, Nee L, et al (1987) Absence of duplication of chromosome 21 genes in familial and sporadic Alzheimer's disease. Science 238:664–666

Sturt E (1976) A mapping function for human chromosomes. Annals of Human Genetics 40:147–163

Suarez BK, Rice J, Reich T (1978) The generalized sib pair distribution: its use in the detection of linkage. Annals of Human Genetics 42:87–94

Theilmann J, Kanani S, Shiang R, Robbins C, Quarrell O, Huggins M, Hedrick A, et al. (1989) Non-random association between alleles detected at D4S95 and D4S98 and the Huntington's disease gene. Journal of Medical Genetics 26:676–681

Thompson EA (1994a) Monte Carlo estimation of multilocus autozygosity probabilities. In J Sall, A Lehman, eds., *Proceedings of the 1994 Interface conference*, 498–506. Fairfax Station, VA

— (1994b) Monte Carlo likelihood in genetic mapping. Statistical Science 9:355–366

— (1994c) Monte Carlo likelihood in the genetic mapping of complex traits. Philosophical Transactions of the Royal Society of London B 344:345–351

— (1997) Conditional gene identity in affected individuals. In IH Pawlowitzki, JH Edwards, EA Thompson, eds., *Genetic Mapping of Disease Genes*, 137–146. Academic Press, London

— (1998) Inferring gene ancestry: estimating gene descent. International Statistical Review 66:29–40

Thompson EA, Guo SW (1991) Evaluation of likelihood ratios for complex genetic models. IMA Journal of Mathematics Applied in Medicine Biology 8:149–169

Thompson EA, Heath SC (1998) Estimation of conditional multilocus gene identity among relatives. IMS Lecture Note Series In Press

Thompson EA, Meagher TR (1998) Genetic linkage in the estimation of pairwise relationship. Theoretical and Applied Genetics 97:857–864

Thompson EA, Shaw RG (1990) Pedigree analysis for quantitative traits: Variance components without matrix inversion. Biometrics 46:399–414

— (1992) Estimating polygenic models for multivariate data on large pedigrees. Genetics 131:971–978

Todorov AA, Siegmund KD, Gu C, Health AC, Rao DC (1997) Multivariate linkage analysis with structural relations. American Journal of Human Genetics 61:A46

Uimari P, Hoeschele I (1997) Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics 146:735–743

van der Meulen M, te Meerman GJ (1997) Association and haplotype sharing due to identity by descent, with an application to genetic mapping. In I Pawlowitzki, J Edwards, E Thompson, eds., *Genetic Mapping of Disease Genes*, 115–136. Academic Press, London

van Tassell CP, van Vleck LD, Gregory KE (1998) Bayesian analysis of twinning and ovulation rates using a multiple-trait threshold model and Gibbs sampling. Journal of the Animal Science 76:2048–2061

Vaxillaire M, Boccio V, Philippi A, Vigouroux C, Terwilliger J, Passa P, Beckmann JS, Velho G, Lathrop GM, Froguel P (1995) A Gene for Maturity Onset Diabetes of the Young) Maps to Chromosome 12q. Nature Genetics 9:418–423

Wang CS, Quaas RL, Pollak EJ (1997) Bayesian Analysis of calving ease scores and birth weights. Genetics Selection Evolution 29:117–143

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Weeks DE, Lange K (1988) The affected-pedigree-member method of linkage analysis. American Journal of Human Genetics 42:315–326

Weeks DE, Lathrop GM, Ott J (1993) Multipoint mapping under genetic interference. Human Heredity 43:86–97

Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118–127

Whittemore AS, Tu IP (1998) Simple, robust linkage tests for affected sibs. American Journal of Human Genetics 62:1228–1242

Wijsman EM, Amos CI (1997) Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genetic Epidemiology 14:719–735

Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G (1995) Identification of the breast cancer susceptibility gene BRCA2. Nature 378:789–792

Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D, et al (1994) Localization of a breast cancer susceptibility gene-BRCA2-to chromosome 13q12-13. Science 265:2088–2090

Xie C, Gessler DD, Xu S (1998) Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics 149:1139–1146

Xiong M, Guo SW (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. American Journal of Human Genetics 60:1513–1531

Yamagata K, Furuta H, Oda N, Kaisaki PJ, Menzel S, Cox NJ, Fajans SS, Signorini S, Stoffel M, Bell GI (1996) Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). Nature 384:458–460

Yuan B, Vaske D, Weber JL, Beck J, Sheffield VC (1997) Improved set of short-tandem-repeat polymorphisms for screening the human genome. American Journal of Human Genetics 60:459–460

Zhao C, Heil J, Dickinson B, Hamm M, Ott L, Vaske D, Christensen C, Weber JL (1997) Computer system for large scale STRP genotyping. American Journal of Human Genetics 61S:A302

Zhao H, Speed TP, McPeek MS (1995) Statistical analysis of crossover interference using the chi-square model. Genetics 139:1045–1056

9h: Consortium/Contractual Arrangements

NONE

9i: Consultants

Professor Simon C. Heath (Rockefeller University) worked as a postdoctoral research associate on the current grant from 1995-1997. He made important contributions to MCMC computational methodology which will be further developed under the current grant. Dr. Heath will assist in this endeavor, and is named as Consultant. A letter from Dr. Heath is included.

Additionally letters are included from Professor Ken Morgan (McGill University), Dr. Elaine Ostrander (Fred Hutchinson Cancer Research Center) and Professor Ted Reich (Washington University). Professor Morgan will collaborate with us in applying our methods to data on the complex genealogy of the Canadian Hutterite population, including haplotype and disequilibrium analyses, assessing the gain of this extended complex pedigree in his endeavors to resolve complex phenotypes. Dr. Ostrander's data on prostate cancer contains markers for both fine-scale and coarse mapping. Professor Ted Reich is enthusiastic about the potential for our methods in application to the COGA database, which includes genome scan data, and information on quantitative and categorical complex phenotypes.