RESEARCH PLAN

a: 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 multilocus marker data. Whereas our focus of the last few years has been effective use of the increasingly available multilocus marker data, we now return to the issue of more complex 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 an understanding of genetic variation among relatives and 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 models and methods, including Markov chain Monte Carlo (MCMC) methods, for realization of gene descent in extended pedigrees conditional on multipoint marker data by
(i) continuing development of block-Gibbs MCMC samplers of gene descent patterns, sampled jointly over individuals and over loci, including diagnostics to ensure accurate sampling of latent variables,
(ii) development of new Metropolis-Hastings MCMC samplers to provide more accurate estimates of gene descent patterns, and
(iii) development of conditional sampling and importance sampling reweighting of marker-based realizations of gene descent patterns, to permit analysis of haplotype identity-by-descent (*ibd*), trait models, linkage likelihoods, genetic map heterogeneity, and meiosis models.

2. To develop further effective Monte Carlo analysis methods for the joint multilocus linkage and segregation analysis of complex traits, by
(i) extending the trait model framework that can be incorporated into Bayesian and likelihood MCMC linkage analyses, including discrete, ordered categorical, censored, and quantitative phenotypes,
(ii) extending the trait model framework that can be incorporated into Bayesian oligogenic linkage analysis, allowing for missing and/or censored covariates, and for genetically correlated traits, and
(iii) combining MCMC estimation of trait model parameters with the sampling of gene descent at marker and trait loci, to develop improved methods for likelihood-based joint segregation and linkage analysis.

3. To develop robust methods for linkage detection using data on extended pedigrees, and to compare linkage detection approaches, by
(i) developing methods for the computation and estimation of the probability distributions of *ibd* measures conditional on multipoint marker data,
(ii) developing scoring statistics and significance measures for robust linkage detection that can be computed on extended pedigrees and make use of the segregation information therein, and
(iii) calibrating and comparing the performance of *ibd*-based, likelihood, and Bayesian approaches to linkage detection for complex traits.

4. To develop Monte Carlo methods using multipoint genetic marker data on extended pedigrees, to analyze genetic maps, meiosis models, and typing and pedigree error, by
(i) analyzing realized recombination events, sampled by MCMC conditional on marker data on pedigrees, to investigate map uncertainty and heterogeneity among individuals,
(ii) using MCMC methods to impute genotypes of specified individuals, and for developing methods to assess probabilities of typing or pedigree errors, and by
(iii) developing likelihood-based tests for genetic interference, likelihood comparisons of alternative meiosis models, and methods for analysis of haplotype *ibd* over multiple tightly linked markers.

5. In addition to appropriate simulation studies, to evaluate methods developed in aims 1-4 on real data, by

(i) using data from multipoint arrays 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 pedigrees of various sizes, with various missing data patterns, to evaluate the utility of such pedigrees for linkage detection, haplotype analysis, and genetic mapping.

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

b: Background and significance

b.1: Complex traits: the classical genetic approach

The risk for developing many common diseases and disease-associated traits is influenced by both genetic and environmental factors. Examples include cardiovascular and neurological diseases, mental illness, cancers, and diseases with an immune-system component such as asthma. Identification of genes which predispose to such traits and diseases (disease genes) provides a key step towards understanding disease etiology. Examples include the early identification of the low density lipoprotein receptor (LDLR) and its role in coronary artery disease (Brown and Goldstein 1986), 4 genes that play a role in Alzheimer's disease (Tanzi and Bertram 2001), as well as more than 30 genes involved in cancers (Knudson 2002; Marsh and Zori 2002), 7 genes involved in Parkinson's disease (Skipper and Farrer 2002), and 1 gene involved in Crohn's disease (Ogura et al. 2001; Hugot et al. 2001). Identification of such genes has often catapulted a field forward and is thus of value, even when the disease genes behave in a Mendelian fashion. For example, identification of the first Parkinson's disease (PD) gene (Polymeropoulos et al. 1997) lead to a radically different understanding of not only the molecular basis of PD, but other neurodegenerative diseases (Skipper and Farrer 2002). Similarly, identification of LDLR was critical for the identification of pathways important in cholesterol transport (Brown and Goldstein 1986).

Identification of disease genes classically involves several stages of gene localization followed by gene identification. As of the beginning of 2003, this approach, which includes positional cloning, is responsible for identification of >1400 genes out of >3337 mapped disease and quantitative trait loci (http://www.ncbi.nlm.nih.gov/LocusLink) - an increase of more than 10-fold in number of genes identified by positional cloning over the past four years. Most of such identified genes were first localized in genome scans to a small genomic region by pedigree-based linkage analysis methods, described more fully in section **b.3**. Once a gene is mapped, the relatively small number of meioses in pedigrees makes it difficult to achieve the fine-scale resolution needed for gene identification (Boehnke 1994). At this stage, identification of linkage disequilibrium, or association, between the trait and nearby markers is often used to limit the size of the region searched for the responsible gene. This proposal is concerned with the steps that precede this fine scale mapping.

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, such as schizophrenia, there have been many attempts to map disease genes, with little evidence of cross-study replication (Baron 2001). For other diseases there are 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). For yet other diseases, such as Hirschsprung disease, evidence for linkage was not apparent until more complex, multilocus models were entertained (Gabriel et al. 2002). Such situations rarely arise for monogenic diseases (Rao et al. 1978).

For several reasons, limitations in the available techniques of analysis contribute to the difficulty, compared to simple monogenic traits, in identifying genes contributing to complex traits. First, there are severe practical problems with performing computations for analyses of complex genetic models, especially 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). For example, a 2-trait locus 2-marker locus analyses on pedigrees of modest size (< 38 individuals) required more than 6 days of computation on a fast workstation (Lin 2000). 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 (Daw et al. 2000b; Cherny et al. 2001). Third, many current methods do not allow efficient use to be made of phenotypic data, covariates, and/or multivariate phenotypes, although there have been some developments along these lines in the past few years in the context of variance-component linkage analysis (Almasy et al. 1997; Williams et al. 1999; Province et al. 2003). Only a limited number of models that relate trait phenotype to genotype have been developed and tested. Current methods that can incorporate covariates cannot handle missing covariate data, yet missing data is extremely common in pedigrees where individuals in older generations may no longer be alive or available for sampling. Finally, available methods that do not require specification of a trait-model are typically limited to restricted, small pedigree structures, which are inefficient for both linkage detection and gene localization (Wijsman and Amos 1997).

b.2: Study designs for mapping of complex traits

There are two fundamentally different study designs available for the detection of genetic linkage, using information on a genomic array of DNA markers. One uses known pedigree relationships among sampled individuals, and the other samples individuals of unknown relationship. The former is based on correlated inheritance patterns within pedigrees, and the latter is based on association between marker genotypes and trait phenotypes among unrelated individuals. The pedigree-based design is that on which most past successes in gene mapping were based (section **b.1**), and does not depend on prior biological knowledge. The population-based design has been used extensively in the investigation of candidate genes (Hirschhorn et al. 2002), and includes as success stories a few notable examples, including ApoE and Alzheimer's disease (Corder et al. 1993), HLA and NIDDM1 (Field 2002), and several loci involve in susceptibility to AIDS (O'Brien et al. 2000). Recently the population-based design has also been suggested as a higher-power alternative approach for linkage detection of complex traits in genome screens (Risch and Merikangas 1996), generating considerable optimism regarding this design for use in linkage detection for complex traits (Sham 2001; Glazier et al. 2002). There are several reasons, however, to be cautious about replacing the family-based design with the population-based design, and it is critical at this stage to remain skeptical of this claim of increased power

Theoretical considerations raise questions about the efficacy of population-based designs for complex-trait gene identification. The study by Risch and Merikangas was based on a limited comparison of models and analysis methods. Examination of other trait models, including multiple disease alleles, multiple mutations, and multiple trait loci, along with comparisons among other analysis approaches, has lead other investigators to question the generality of the initial conclusion favoring a population-based design (Chapman and Wijsman 1998; Morris and Kaplan 2002; Terwilliger and Weiss 1998), or to note that sample sizes grow rapidly with increasing complexity of the trait model (Longmate 2001). Finally, allelic heterogeneity is a major problem for population-based approaches, yet extensive empirical data shows that there can be very large numbers of mutations at identified loci: for example, for early-onset Alzheimer's disease and familial combined hyperlipidemia, over 100 (Tanzi and Bertram 2001) and 600 (Lombardi et al. 2000) mutations are known, respectively. In contrast, allelic heterogeneity has no deleterious effect on family-based analyses.

Empirical data also does not lend support to the claim that population-based designs are more effective than family-based designs for detecting gene linkage. (1) Reproducible associations, such as NIDDM1 and HLA, have so far also been identifiable with family-based designs (Field 2002). (2) Comparison of results of analysis of simulated data that included both population-based and family-based designs indicates that family-based designs, especially based on larger pedigrees, are more likely to be able to identify correctly the location of segregating trait loci (Goldin 2001; Falk 2001). (3) Tests of population-based designs on genes that have already been identified have failed to detect evidence of genes that can be identified through family studies, including Crohn's disease (van Heel et al. 2002) and Alzheimer's disease (AD) (Emahazion et al. 2001; Rogaeva 2002). (4) Reproducibility is difficult to achieve in population-based designs. For example, a comprehensive review of 166 multiply-studied putative associations identified only 6 that were consistently replicated (Hirschhorn et al. 2002). Also, a study of 36 traits that were investigated in multiple studies reported that cumulative odds ratios approached 1, as would be expected under lack of association (Ioannidis et al. 2001).

There are several reasons that may explain discrepancies among theoretical studies, and between some theoretical studies and empirical results. (1) Some studies that have compared population-based and pedigree-based linkage approaches have tended to focus on simplistic pedigree structures such as affected sib pair analyses (Risch and Merikangas 1996). Sib-pair designs are well-known to have low power per sampled individual, relative to use of larger pedigrees (Wijsman and Amos 1997). (2) Theoretical studies have varied in their assumptions and approaches. The sensitivity of conclusions to assumptions suggests sensitivity to parameters about which we have little information. (3) Despite claims for low power in pedigree-designs compared to population designs (Ardlie et al. 2002), little allowance in these discussions has been made for methods of data ascertainment that increase the informativeness of the sample, which again, can markedly improve power (Wijsman and Amos 1997; Goldgar and Easton 1997). As both theory (Zeng 1994; Jansen 2001) and the myriad of linkage studies on crosses in model organisms attest (Mackay 2001; Mauricio 2001; Andersson 2001), pedigree-based linkage mapping for trait loci with common alleles is a powerful tool to detect and map such loci.

b.3: Linkage estimation and testing for complex traits

Within linkage analysis methods based on data from individuals of known relationship, there are again two approaches, sometimes referred to as "parametric" and "non-parametric" or "model-free". We avoid this characterization since all genetic data analysis relies on parametric models for the meiosis process underlying the data, and since the performance of a method must necessarily depend on the underlying true genetic basis of the trait. The fundamental difference between the approaches is in whether a trait model is explicitly included in the analysis. If a trait model is included, then analysis is based on the probabilities under the joint marker and trait model of the combined data $\mathbf{Y} = (\mathbf{Y}_M, Y_T)$ on the marker (M) and trait (T) phenotypes of individuals. That is, the method is based on the likelihood for the joint marker and trait model given the observed data, and we refer to such methods as "likelihood-based". If a trait model is not included, then analysis is based on marker allele sharing probabilities among individuals with selected phenotypes, and we refer to such methods as identity-by-descent (*ibd*) based linkage detection methods.

A trait model may be estimated in a joint segregation and linkage analysis (see **d.2**), but in the linkage assessment this model is normally held fixed. The statistic used for assessment is typically the lod score for the trait locus location λ:

$$\mathrm{lod}(\lambda) = \log_{10}\left(\frac{\mathrm{P}_\psi(\mathbf{Y};\lambda)}{\mathrm{P}_\psi(\mathbf{Y};\lambda=\infty)}\right) = \log_{10}\left(\frac{\mathrm{P}_\psi(Y_T \mid \mathbf{Y}_M,\lambda)}{\mathrm{P}_\psi(Y_T)}\right) \tag{1}$$

where $\lambda = \infty$ denotes that the trait locus is unlinked to the markers on the chromosome under consideration, and ψ summarizes all other parameters of the genetic model. Alternatively, a Bayesian prior distribution may be placed on the parameters of the trait model, and these may be integrated (or sampled) over to estimate the marginal posterior probability for λ. In this case, assessment of the linkage evidence is not straightforward, since the posterior probability of trait locus position depends on the structure of the marker data $\mathbf{Y}_M$. One approach is to compare the posterior probability with that for marker data simulated under the same marker model on the same pedigree structures and missing data pattern as the actual data, but in the absence of linkage (Daw et al. 2003). The likelihood is a fundamental component of any Bayesian approach: the posterior distribution is proportional to the product of prior distribution and likelihood. The term "likelihood-based" does not imply "frequentist", and the adoption of Bayesian prior distribution over trait models does not preclude a frequentist assessment of the significance of a linkage signal.

The earliest pedigree-based human genetic inference methods (Fisher 1934; Haldane 1934) were restricted to small pedigrees, primarily for computational reasons. As improved computational methods became available (Elston and Stewart 1971; Ott 1974; Lathrop et al. 1984) the additional power provided by data on extended pedigrees became well accepted in the mapping of Mendelian traits. Given even an approximate model, extended pedigrees also provide increased power for mapping complex traits (Wijsman and Amos 1997; Chapman et al. 2001), but a more robust approach to linkage detection avoids specification of any trait model through use of *ibd*-based linkage detection methods. Such methods, which are based on concordance of markers in relatives of like phenotype, date back to the sib pair methods of Penrose (1935) and Suarez et al. (1978). In recent years these methods have been developed in terms of *ibd* probabilities computed or imputed conditionally on marker data. Three types of measures have been used. (1) Risch (1990) developed the MLS method for sib pairs, which compares probabilities of marker data $\mathbf{Y}_M$ under estimated and null (no-linkage) values of *ibd* probabilities. (2) The approach developed by Whittemore and Halpern (1994) requires a measure, W_λ, of *ibd* at chromosomal location λ, among individuals of like phenotype. The test statistic is then the expected value of W given the marker data $\mathbf{Y}_M$, which is compared to its expected value in absence of linkage. (3) Thompson (2000a) has considered the probability odds of particular sets of *ibd* configurations, conditional on marker data $\mathbf{Y}_M$, and compared these with the probability odds of the same configurations given only the pedigree structure.

Affected relative pair methods generally lack power. This results from both the desired robustness to trait model assumptions and also because designs based on pairs of relatives lose much of the segregation information inherent in an extended pedigree structure. Newer computational methods have allowed extensions of the approaches of Risch (1990) and Whittemore and Halpern (1994) to small pedigrees (Kruglyak and Lander 1995; Kruglyak et al. 1996). McPeek (1999) has developed and compared alternative *ibd* measures W for use on small pedigrees. Despite these improvements, there remain both computational and statistical difficulties with the use of *ibd*-based linkage detection tests on extended pedigrees. The standard *ibd* measures (Kruglyak et al. 1996; McPeek 1999) may have highly skewed distributions or be computationally intractable on extended pedigrees (Basu et al. 2002). Moreover, as with the Bayesian approach, the correct comparison is not of posterior (given $\mathbf{Y}_M$) and prior (pedigree-based) probabilities of *ibd*, but of posterior distribution given the observed data with that expected in the presence of the same marker data availability but absence of linkage.

The trait-model robustness of the *ibd* approach makes it appropriate for linkage detection for truly complex traits. Of necessity power is lost relative to inference with a trait model that accurately reflects trait model determination. However, with suitable *ibd*-measures W and use of the full distribution of W conditionally on marker data, the information inherent in joint segregation patterns at marker loci can be extracted (see **d.3**).

b.4 Map and genotyping error

Genetic maps are an important component of multipoint linkage analysis. Analysis with such maps is important for pedigree and marker error detection (Boehnke and Cox 1997; Sieberts et al. 2002), increasing the information available in pedigrees for both initial gene detection and for improving the accuracy of gene localization (Lathrop et al. 1984; Wijsman and Amos 1997). A genetic map consists of the order and meiotic intermarker distances between a series of marker loci. Sex-specific maps exist, or, more commonly, 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).

Linkage analysis is sensitive to map estimates. Poor map estimates can result in both false negative (Halpern and Whittemore 1999) and false positive (Daw et al. 2000b) evidence of linkage, and empirical results illustrate the impact of variability in estimated map distances. For example, an analysis of familial stroke with the latest Icelandic map (Kong et al. 2002) gave a maximum lod score of 4.4, while analysis with the widely used Marshfield map (http://research.marshfieldclinic.org/genetics/) gave a lod score of only 3.4 (Gretarsdottir et al. 2002). This variability in results among maps may be a function of sample size used to construct the maps. Because meiotic map distances can vary considerably from the distances predicted from physical maps (Nachman and Churchill 1996), maps must be estimated from meiotic data and thus the map accuracy obtained is limited by the size of the meiotic mapping panels. The Marshfield map, as well as other commonly used maps, is based on a very limited set of $\sim$ 200 meiotic events: the 95% confidence interval for a typical 10 cM distance between markers in current maps ranges from 4-19 cM in sex-averaged maps (Yuan et al. 1997), and is wider for sex-specific maps. For higher-density maps of diallelic markers (SNPs) that are starting to be generated (Sachidanandam et al. 2001), this sample size is inadequate for accurate map estimation. Two new maps have recently been constructed which are based on independent and larger data sets - the Icelandic map (Kong et al. 2002), and the TSC map (Matise et al. 2002). The effect of large sample size on map accuracy is underscored by the observation that these two maps are more similar to each other (99% identical in marker order) than to the Marshfield map (97% identical in marker order) (Matise et al. 2002), despite the fact that the TSC map includes as a subset the pedigrees used to construct the Marshfield map.

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 RELPAIR (Boehnke and Cox 1997) can be used to identify potential errors in small pedigrees, and **Eclipse** has extended this approach to include an error model and additional relationships (Sieberts et al. 2002). However, for large pedigrees, since computations needed for multipoint error detection are the same as those used for linkage analysis, such haplotype-based error detection is impractical.

b.5: Monte Carlo based methods for analysis of data on pedigrees

Over the last twelve years, MCMC 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 makes feasible and practical the MCMC estimation of probabilities and likelihoods. The Metropolis algorithm is an older MCMC algorithm 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 developing effective MCMC methods for genetic analysis with multiple marker loci, we have relied increasingly on the framework of latent variables provided by the meiosis indicators $\mathbf{S} = \{S_{i,j}\}$. These specify the parental origins of genes at each locus j in each meiosis i:

$$S_{i,j} \;=\; 1 \text{ or } 0 \text{ as the gene is grandpaternal or grandmaternal} \tag{2}$$

The meiosis indicators $\mathbf{S}$ define the descent of genes through a pedigree structure, and determine the founder gene origin of each allele of every pedigree member at every locus considered. Together with an assignment of allelic type to each founder allele that is represented, $\mathbf{S}$ determines the genotypes underlying observed marker or trait phenotypes. While earlier MCMC methods on pedigrees (e.g., Thompson and Guo (1991)) sampled latent genotypes directly, MCMC sampling of $\mathbf{S}$ provides a more flexible and powerful framework. The ability to sample $\mathbf{S}$ from its full joint distribution $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y})$ under genetic model ψ conditional on observed marker and trait data $\mathbf{Y}$, provides provides direct estimates of gene *ibd* jointly over loci. Recombination events are also a function of $\mathbf{S}$: in any meiosis i there is a recombination between loci j and j' if $S_{i,j} \neq S_{i,j'}$. Thus sampling $\mathbf{S}$ conditionally upon marker data provides provides information for analysis of genetic maps and meiosis patterns.

The dependence structure of $\mathbf{S}$ is a key to effective computation and sampling. Meioses i are *a priori* independent, while, in the absence of genetic interference, $S_{i,j}$ are Markov along a chromosome. The vector of $S_{i,j}$ values over the m meioses of a pedigree ($S_{\bullet,j} = \{S_{i,j}, i = 1,\ldots,m\}$) is the inheritance vector at locus j (Lander and Green 1987). Suppose that the data are similarly structured, so that $Y_{\bullet,j}$ are the marker or trait data determined by underlying genes at locus j , $j = 1,\ldots,L$. Then the probability of observed data $\mathbf{Y} = \{Y_{\bullet,j}\}$ under genetic model ψ takes the form

$$\mathrm{P}_\psi(\mathbf{Y}) \;=\; \sum_{\mathbf{S}} \mathrm{P}_\psi(\mathbf{Y} \mid \mathbf{S})\; \mathrm{P}_\psi(\mathbf{S}) \;=\; \sum_{\mathbf{S}} \left(\prod_{j=1}^{L} \mathrm{P}(Y_{\bullet,j}|S_{\bullet,j})\right) \mathrm{P}(S_{\bullet,1}) \left(\prod_{j=2}^{L} \mathrm{P}(S_{\bullet,j}|S_{\bullet,j-1})\right) \tag{3}$$

This Markov structure of the latent variables $\mathbf{S}$ permits computation using the Baum algorithm (Baum et al. 1970), either for exact computation on small pedigrees (Lander and Green 1987; Kruglyak et al. 1996) or for sampling based MCMC methods (Thompson 2000b).

A major difficulty with MCMC methods is to ensure proper mixing of the samplers, and hence efficient Monte Carlo estimation. This is a particular problem in sampling large discrete probability spaces, such as that defined by the conditional probability of $\mathbf{S}$ given data $\mathbf{Y}$, on large pedigrees with a significant proportion of unobserved individuals, using models or data involving multiple linked loci. Together, the constraints of Mendelian segregation, of linkage among loci, and of observations at multi-allelic markers, make this an exceptionally difficult space to sample effectively. Since 1995, several improved samplers have been proposed, all involving simultaneous updating of blocks of latent variables (Jensen et al. 1995; Jensen and Kong 1999; Thomas et al. 2000). Jensen et al. (1995) update genotypes of blocks of individuals jointly at several loci, while Heath (1997) updates all genotypes at a single locus, and Thompson and Heath (1999) update jointly meiosis indicators for all loci in a single meiosis. Our current methods use as a core engine a combination of the locus sampling (L-sampler) of Heath (1997) with the meiosis (M-sampler) sampling of Thompson and Heath (1999). The combination of these two orthogonal approaches (LM-sampler) greatly improves performance (Thompson 2000b). Development of more effective MCMC sampling procedures for specific purposes continues (George and Thompson 2003), while the integration of diverse sampling and estimation proposals by many researchers brings continued improvement.

Recent years have also seen improvements in efficiency of exact computational methods in pedigree analysis, as well as in the speeds and processor power of computers. New computational algorithms are based either on the Baum algorithm (Baum et al. 1970), or the more general stochastic network approach of HUGIN (Andersen et al. 1989). New programs such as GENEHUNTER 2 (Kruglyak and Lander 1998; Markianos et al. 2001), Allegro (Gudbjartsson et al. 2000) and Merlin (Abecasis et al. 2002) are increasing the size of pedigrees on which multilocus computations can be undertaken, although size limitations still create problems in many cases (Goedken et al. 2000). Exact computation is intrinsically limited in the number of loci, size of pedigree, or pedigree complexity. Nonetheless, newer exact methods provide much more flexible and powerful tools for validating Monte Carlo approaches, and elements of the computational algorithms can be incorporated into MCMC samplers, as has been done in block updating MCMC samplers. Even where exact computation of probabilities $\mathrm{P}_{\psi}(\mathbf{Y})$ and $\mathrm{P}_{\psi}(S_{\bullet,j}|\mathbf{Y})$ is feasible, MCMC has advantages. MCMC provides a joint realization from $\mathrm{P}_{\psi}(\mathbf{S}|\mathbf{Y})$. MCMC is also flexible: for example, while still using the basic form (3) for sampling proposals, genetic interference can be incorporated (Thompson 2000a).

MCMC methods can also sample elements of ψ in a Bayesian or pseudo-Bayesian approach. The elements of ψ selected for sampling are usually either trait segregation model parameters, β, or trait locus positions, λ, the marker model Λ_M being assumed known. In a pseudo-Bayesian approach, the objective remains a likelihood-based analysis. A prior distribution is placed on a limited number of parameters solely in order to improve MCMC mixing (Geyer and Thompson 1995; George and Thompson 2003). By contrast, in a fully Bayesian approach prior distributions are placed on parameters, and inferences are based on the posterior probability distributions of parameters of interest. For models with a large number of parameters, the Bayesian approach provides a powerful and effective tool. Methods for the oligogenic linkage analysis of quantitative traits continue to develop (Heath 1997; Uimari and Sillanpaa 2001; Bink et al. 2002), and these methods have been successful in analyses of complex quantitative traits in the presence of heterogeneity (Yuan et al. 2000; Gagnon et al. 2002). Bayesian MCMC methods permit the inference of a variable number of quantitative trait loci (QTL), different ones of which may segregate in different families. The number of QTL is also accorded a prior distribution, and the MCMC output provides an estimate of the posterior distribution of QTL of non-negligible effect (Daw et al. 1999a). Bayesian approaches provide not only posterior probabilities of the descent of genes in pedigrees, the genotypes of individuals, and of patterns of genotype sharing, but also posterior probabilities of linkage, trait gene effects and locations.

MCMC methods, providing either estimates of posterior probabilities or Monte Carlo likelihood functions, are ideally suited to all forms of "incomplete data", including censored traits and latent disease liabilities. Considering latent variables between the level of genotype and phenotype, such as genotypically determined latent liabilities, greatly extends 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 used successfully for simpler traits can be brought to bear on more truly complex phenotypes.

C: Progress report

c.1: Summary of previous specific aims

The specific aims of the previous award period (1999-2003) were

1. To develop further our Markov chain Monte Carlo (MCMC) methods for multilocus linkage analyses on extended pedigrees, using an array of linked markers, including estimation of conditional gene identity-by-descent (*ibd*) patterns jointly over individuals and over loci, for linkage detection and gene resolution, continuing development of MCMC samplers of gene *ibd* patterns and multilocus haplotypes, and development of MCMC methods of haplotype analysis at the pedigree and population levels.

2. To develop further effective Monte Carlo analysis methods for the joint multilocus linkage and segregation analysis of complex traits, by extending Bayesian MCMC oligogenic linkage analyses to ordered categorical

phenotypes, and extending the oligogenic models that underlie quantitative and categorical traits.

3. To develop methods to analyze and interpret the outputs of our Monte Carlo methods, by calibrating the MCMC realizations of oligogene locations in relation to other measure of linkage evidence, by assessing the gains of multilocus linkage analyses using oligogenic trait models with multiple QTL, and developing methods for the interpretation of conditional gene *ibd* patterns.

4. To develop Monte Carlo methods to analyze genetic maps, genetic interference, and meiosis models, by studying map heterogeneity, uncertainty, and error, and their impact on haplotype imputation, and on genotype and pedigree error detection. To extend MCMC methods to permit linkage detection and haplotype estimation under interference models, and to develop methods to realize underlying meiosis processes conditional on genetic marker data, in order to fit more general meiosis models.

5. In addition to appropriate simulation studies, to evaluate methods developed under #1, #2, #3, and #4, on real data, by using data from a genomic array of markers to evaluate MCMC approaches to linkage detection and oligogenic linkage analysis, by evaluating methods of joint segregation and linkage analysis applied to complex phenotypes, including censored and categorical phenotypes with covariate information, and by exploring the use of MCMC methods on very large multi-generation pedigrees, to evaluate their utility for linkage detection, haplotype analysis, and fine-scale mapping.

6. To extend and enhance our user software, by further developing the user-friendly MORGAN_2 package to contain the full array of our current methods and programs, by implementing new methodology as proposed in #1, #2, #3 and #4, and documenting, distributing and supporting the software.

c.2 Progress in meeting the aims

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

1. We have made significant advances in the development of MCMC methods for sampling multilocus marker haplotypes on extended pedigrees with extensive missing data [15], and used these both in *ibd* scoring methods for linkage detection [**2**,**16**,27,33], and for improved estimation of linkage likelihoods [6,17,18]. The samplers are based on a combination of locus (L) and meiosis (M) block Gibbs updates of the latent inheritance array **S**, and we refer to it as the LM-sampler, with programs which use it having the prefix **lm_**. We have also implemented sequential imputation (Kong et al. 1994) as a component of sampling, both for start-up and as a mechanism for Metropolis-Hastings restarts in the course of an MCMC run. The improved samplers provide better imputation of gene *ibd* on extended pedigrees with missing data, as implemented in our program **lm_auto**, and we have begun to develop multilocus *ibd* scoring statistics for linkage detection [24,33]. For comparison purposes, we have continued to study classical linkage likelihood methods and to compare MCMC approaches with other methods [**2**,9,19,20,21,**30**,31] We have also studied haplotype structure at the population level [1,3], developed methods for fine scale mapping from population haplotype data [10,11,34], and implemented likelihood estimation of inbreeding from population genome screen data [38].

2. The implementation of our Bayesian MCMC methods towards complex traits has not proceeded as fast as expected, due to a change in postdoctoral personnel. However, we have developed new methods for MCMC estimation of linkage likelihoods which show promise for complex traits. In addition to improving previous estimators of linkage likelihoods implemented in our older **lm_lods** program [29] we have implemented two new estimators. One is based on the original MCMC estimator of location scores introduced by Lange and Sobel (1991), but uses our LM-sampler and other MCMC enhancements: this is implemented in the program **lm_markers**. Additionally, we have developed a new pseudo-Bayes MCMC estimation method [**28**,36]. In this approach the trait segregation model and genetic marker model are fixed within any given MCMC run, but in addition to sampling over the space of latent inheritance vectors **S**, the MCMC samples also over a discrete set of trait locus locations. A linkage likelihood is then regained from the posterior distribution, by dividing by the one-dimensional prior distribution on trait locus positions. Since the prior plays no role in the final inference, serving only to enhance MCMC performance, the method is pseudo-Bayesian in the sense of Geyer and Thompson (1995). This approach is implemented in the program **lm_bayes** and has been applied to ordered categorical trait data [27,37], using a new penetrance routine. This and several other new penetrance

routines can be used also in our other MCMC linkage likelihood programs, **lm_lods** and **lm_markers**, and so enable us to address slightly more complex traits than previously, but a major extension to complex traits remains to be undertaken (see **d.2**). We have also continued to develop our MCMC program **schnell** for linkage likelihoods for quantitative traits: each of **lm_lods** and **schnell** has incorporated enhancements of the other.

3. Our assessment of methods has proceeded primarily through comparison of MCMC and other methods on real and simulated data. We have compared exact and MCMC methods for model selection and hypothesis testing [13], and compared a variety of exact and MCMC methods for linkage detection and mapping in Genetic Analysis Workshop papers [**2**,6,27], with a focus on the utility of extended pedigrees. Using MCMC linkage likelihood methods we have investigated the utility of multi-marker analyses for improved trait locus localization on extended pedigrees with substantial missing data, validating MCMC results by comparison with exact methods where feasible [**5**,24,**28**,29,36,37], and by comparison to other MCMC and sequential imputation approaches [32]. We have assessed MCMC methods for oligogenic traits both for model selection and for joint segregation and linkage analysis [22,**23**]. Additionally, we have developed a score for linkage detection in oligogenic traits, using the relative frequencies of trait locus realization on a chromosome compared with frequencies on an unlinked chromosome with the same marker characteristics and data availability [**26**]. This score shows promise for linkage detection of QTL in Bayesian analyses using oligogenic trait models.

4. We have undertaken a study of the effects of genetic map misspecification on linkage inferences [8]. We have developed methods for detection of pedigree and typing errors from genome scan marker data [**14**], and investigated their impact on haplotype inference. We have investigated the effect of unknown parental relationships on sib pair linkage likelihood methods [12]. We have not yet used MCMC in these studies. Although the computations for multilocus probabilities use the same latent meiosis array, **S**, and Baum algorithm framework (Baum et al. 1970), we restricted initial studies to cases where exact computation is feasible, including sibship, avuncular, and cousin relationships. In particular, we have explored the additional information on pairwise relationship provided by marker data on an additional close relative [**14**]. In fact, joint information on trios of individuals provides for much more powerful relationship inference and typing error detection, and we have investigated the effects of alternative error models, genetic maps, and marker polymorphism, on our conclusions.

We have initiated the use of MCMC methods both in the assessment of genetic map uncertainty and estimation, in the detection of typing errors in more extended pedigrees, and in the assessment of such errors on the performance of MCMC estimates of linkage likelihoods. We have implemented MCMC methods permitting genetic interference [**16**], and used these to estimate the potential impact of interference on genetic map and linkage likelihood estimation.

5. In addition to the inclusion of real-data examples in analyses [**14**,26,33–37], we have used our methods in the analysis of real data [**2**,**5**,6,**7**,25,27,35]. These cases have provided useful tests of the methods because we have been able to compare results to those from other analyses of the same or similar data, or there are other measures, such as an identified gene, that could be used to validate results.

MCMC linkage analyses have worked well in analysis of real data. Oligogenic multipoint linkage analysis of Alzheimer's disease (AD) pedigrees demonstrated correct localization of the known PSEN1 gene, even in a large and genetically heterogeneous data set [**5**]. An analysis with **lm_bayes** of a single complex AD pedigree with a known mutation also gave correct lod scores and gene localization; time required for analysis with 2 markers was only 6 min of CPU time, compared to 3.3 hrs. with the standard FASTLINK program, and time required for analysis with 10 markers was 8.75 hrs while for exact methods this problem was infeasible [37]. Analysis of the Framingham (FRM) data set as part of GAW13 [27,32] allowed a comparison of estimated *ibd* sharing in large pedigrees among three independent Monte Carlo implementations with different algorithms: Simwalk2 (Sobel and Lange 1996), SIMPLE (Skrivanek et al. 1999), and our programs. Correlations between *ibd* estimates from our programs and these others exceed .96, supporting the conclusion that this approach gives valid estimates. With an oligogenic model, analysis of an alcoholism data set (COGA) gave evidence for linkage of to chromosome 1 [6], and was corroborated by similar results in the same data obtained with standard analysis methods (Almasy and Borecki 1999). Similarly, linkage analysis in a complex asthma pedigree identified a region on chromosome 1 [**2**], which was also identified with other methods [4]. Multipoint linkage analysis provided evidence of linkage of HDL levels to chromosome 11, which was undetectable in this data set until a multipoint analysis was made feasible with MCMC methods. Finally, multipoint analysis of a dyslexia phenotype (DYS) produced concordant linkage signals with both Bayesian oligogenic linkage and with variance component analyses, and analysis of an autism data set (AUT) with phenotypic measures has obtained concordant evidence for linkage to the same markers with MCMC methods and with Haseman-Elston regression methods.

We have performed MCMC analyses to obtain and investigate aspects of the trait model. For AD, we showed that a censored quantitative trait model can localize known genes [**5**]. We used Bayesian oligogenic segregation analysis to provide evidence of 4-7 genes affecting age of onset in late-onset AD [**7**], and 2-3 contributing to prostate cancer [25,35]. Estimates of total genetic variance obtained through this approach vs. through traditional complex segregation analysis are similar in both the COGA data, and in the DYS data; e.g, for the latter data, estimated genetic variance was .28 and .3, respectively. For the DYS data, we also found that it is possible to summarize the parameters of the model for a QTL mapping to a particular region, and that these parameters are useful for use in a traditional model-based (lod=2.7). Initial genome scans of the DYS and FCHL data indicate that the mean number of genes fit by the oligogenic trait analysis predicts the number of potentially interesting linkage signals: for example, the mean number of QTL contributing to a pseudoword reading measure is about 2, with two strong linkage signals subsequently found in the genome scan. Finally, attempts to use Bayesian oligogenic analysis on several data sets, including AD, FRM, FCHL, DYS, and AUT have identified the prior distribution on the genetic variance of QTL effects as key to getting good results with these methods [**23**], and that there are phenotypes for which current models do not work well, as assayed by problems with convergence in parameter estimates.

6. In addition to testing and maintaining up-to-date versions of the **Loki** package and the pedigree drawing **Pedfiddler** package, we developed a new package **Eclipse** for analysis of pedigree and typing errors [**14**], and continued development of our MORGAN package. Details of these software packages are available at our grant software web page http://www.stat.washingon.edu/thompson/Genepi/pangaea.html. The most recent MORGAN release is version 2.5 (11/2002). This version includes several new MCMC lod score programs [18,**28**] as well as improvements to the methods of earlier programs [17,29].

Since 08/2000, we have been using the CVS system (Concurrent Versions System: www.cvshome.org) to maintain and develop our MORGAN package. This has enabled us to integrate development, local-user, and public-release versions, to incorporate newly developed algorithms by several different researchers in the group, to make major changes increasing data format flexibility, and in general to speed up software development without compromising software reliability and integrity. We have also implemented the documentation system doxygen (www.doxygen.org). Doxygen extracts documentation from the source code into various formats. The html format serves as hyperlinked source code browser, greatly facilitating overview of a package which now contains over 56,500 lines of C-code.

While focusing on Unix/Linux platforms, we experimented with CodeWarrior versions of MORGAN under Mac OS 8.6 and 9.1 and with MS Windows under Cygwin. With the availability of Mac OSX, a separate Mac version is no longer necessary, but we have now ported MORGAN to compile natively under MS Windows with MinGW compilers (www.mingw.org). The MinGW port has removed the dependence on Cygwin's POSIX emulation layer (cygwin1.dll), is easier to use, and should have better performance. A single MORGAN release can run under all major operating systems (Unix/Linux, Windows, and Mac OSX). MORGAN has been used in several short courses and other tutorials settings, leading to testing on multiple platforms and the development of examples and improved documentation.

c.3(i) Papers published since previous renewal submission.

1. Chapman, N.H., and Thompson, E.A. (2000) Linkage disequilibrium mapping: the role of population history, size, and structure. *Advances in Genetics, Vol 42: Genetic Dissection of Complex Traits* D.C.Rao and M.A.Province (eds). Pp. 413–437. Academic Press: San Diego, CA.

2.* Chapman, N.H., Leutenegger, A-L., Badzioch, M. D., Bogdan, M., Conlon, E.M., Daw, E.W., Gagnon, F., Li, N, Maia, J.M., Wijsman, E.M. and Thompson, E. A. (2001) The importance of connections: Joining components of the Hutterite pedigree. In *Genetic Analysis Workshop 12. Genetic Epidemiology* **21**(Suppl 1): S230–S235.

3. Chapman, N.H. and Thompson E.A. (2002) The effect of population history on the lengths of ancestral chromosome segments. *Genetics* **162:** 449–458.

4. Chapman, N.H., and Wijsman, E.M. Introduction: Linkage analysis in the Hutterites. In *Genetic Analysis Workshop 12. Genetic Epidemiology* **21**(Suppl 1): S222–S223.

5.* Daw, E.W., Heath, S.C., and Wijsman, E. M. (1999) Multipoint oligogenic analysis of age-of-onset data with applications to Alzheimer's disease pedigrees. *American Journal of Human Genetics* **64:** 839–852.

6. Daw, E.W., Kumm, J., Snow, G.L., Thompson, E. A., and Wijsman, E. M. (1999) Monte Carlo Markov chain methods for genome screening. In *Genetic Analysis Workshop 11. Genetic Epidemiology* **17:** Supplement 1, S133–138.

7.* Daw, E.W., Payami, H., Nemens, E.J., Nochlin, D., Bird, T.D., Schellenberg, G.D., and Wijsman, E.M. (2000) The Number of Trait Loci in Late-Onset Alzheimer Disease. *American Journal of Human Genetics* 66: 196–204.

8. Daw, E. W., Thompson, E. A. and Wijsman, E. M. (2000) Bias in multipoint linkage analysis arising from map misspecification. *Genetic Epidemiology* **19:** 366–380.

9. Elston, R,C. and Thompson, E.A. (2000) A century of biometrical genetics. *Biometrics* **56:** 659–666.

10. 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.

11. Graham, J., and Thompson, E. A. (2000) A coalescent model for ancestry of a rare allele. *Genetics* **156:** 375–384.

12. Leutenegger, A-L., Genin, E., Thompson, E. A., and Clerget-Darpoux, F. (2002) Impact of parental relationships in maximum lod score affected sib pair method. *Genetic Epidemiology* **23:** 413–425.

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.* Sieberts, S. K., Wijsman, E. M. and Thompson, E. A. (2002) Relationship inference from trios of individuals in the presence of typing error. *American Journal of Human Genetics* **70:** 170–180.

15. Thompson, E.A., and Heath, S. C. (1999) Estimation of conditional multilocus gene identity among relatives. In: *Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology.* F. Seillier-Moiseiwitsch (Ed), IMS Lecture Note–Monograph Series Volume 33, Pp.95–113. Institute of Mathematical Statistics (Hayward, CA).

16.* Thompson, E. A. (2000) MCMC estimation of multi-locus genome sharing and multipoint gene location scores. *International Statistical Review* **68:** 53–73.

17. Thompson, E.A. (2000) Monte Carlo methods on genetic structures. In *Complex Stochastic Systems* O.E. Barndorff-Nielsen, D.R.Cox, and C. Klüppelberg (eds.) (Chapter 4, Pp.176–218) Proceedings of the fourth Séminaire Européen de Statistique (Eindhoven, March 1999). CRC Press: Boca Raton, FL.

18. Thompson, E. A. (2000) *Statistical Inferences from Genetic Data on Pedigrees.* NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 6. Institute of Mathematical Statistics, Beachwood, OH.

19. Thompson, E.A. (2001) Linkage Analysis. In *Handbook of Statistical Genetics.* D.J. Balding, M.Bishop, C. Cannings (eds). Pp. 541–563. Wiley; Chichester, UK.

20. Thompson, E. A. (2002) Detection of human genetic linkage: Foundations. *Theoretical Population Biology* **60:** 203–213.

21. Thompson, E. A. (2002) Human Genetics: Overview. In *Biostatistical Genetics and Genetic Epidemiology* Pp. 386–390. R. C, Elston, Olson J., and Palmer (eds.). Wiley; New York.

22. Wijsman E. M. (2000) Monte Carlo methods and model selection in genetic epidemiology. *Computational Statistics and Data Analysis* **32:** 349–360.

23.* Wijsman, E. M. (2002) Joint segregation and linkage analysis using Markov chain Monte Carlo methods. In *Quantitative trait loci: methods and protocols.* N. Camp and A. Cox, eds. Pp. 139–161. Humana Press, Totowa, NJ.

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

24. Basu, S. and Thompson, E. A. (2002) A study of an allele sharing statistic on extended pedigrees. Technical report No. 413, Department of Statistics, University of Washington.

25. Conlon, E. M., Goode, E. L., Gibbs, M., Stanford, J. L., Badzioch, M., Janer, M., Kolb, S., Hood, L., Ostrander, E. A., Jarvik, G. P., Wijsman, E. M. (2003) Segregation analysis of hereditary prostate cancer pedigrees: evidence for multiple loci affecting age-at-onset. *International Journal of Cancer:* in press.

26.* Daw, E. W., Wijsman, E. M. and Thompson, E. A. (2003) A score for Bayesian genome screening. *Genetic Epidemiology:* in press.

27. George, A.W., Basu, S., Li, N., Rothstein, J.H., Sieberts, S.K., Stewart, W., Wijsman, E.M., and Thompson, E. A. (2003) Approaches to mapping genetically correlated traits. *Genetic Analysis Workshop 13:* in press.

28.* George, A. W. and Thompson, E. A. (2002) Multipoint linkage analyses for disease mapping in extended pedigrees: A Markov chain Monte Carlo approach. *Statistical Science* **18:** in press. (Also, Technical report No. 405, Department of Statistics, University of Washington.)

29. Stewart, W. C. L., George, A. W. and Thompson, E. A. (2002) Using Markov chain Monte Carlo for multipoint linkage analysis: Improved estimates of lod scores. Technical report No. 412, Department of Statistics, University of Washington.

30.* Thompson, E.A. (2003) Linkage Analysis. In *Handbook of Statistical Genetics.* D.J. Balding, M.Bishop, C. Cannings (eds). Wiley; Chichester, UK. Revised edition: in press.

31. Thompson, E. A. (2003) Lod Scores. In *Encyclopedia of the Human Genome* Macmillan: London, UK. In press.

32. Wijsman, E. M. (2003) Summary of group 8: development and extension of linkage methods. *Genetic Analysis Workshop 13 summary paper:* Submitted.

c.3(iii) Relevant published abstracts not yet otherwise represented by published or submitted papers

33. S. Basu, E. M. Wijsman, and E. A. Thompson, (2002) Allele-sharing methods on large pedigrees. *Genetic Epidemiology* **23:** 267.

34. Chapman, N. H., Crumley, J., Fujiwara, T. M., Morgan, K., and Thompson, E. A. (2001) Estimation of small map distances using chromosome data from an isolated population. *American Journal of Human Genetics* **69:** S410 (#1336).

35. Daw, E. W., Bird, T.D., Nemens, E. J., Nochlin, D., Schellenberg, G., and Wijsman, E. M. (2000) ApoE and other loci affect age of Alzheimer's disease onset in families with PS2 mutation. *American Journal of Human Genetics* **67:** S192.

36. George, A. W., Bogdan, M., Wijsman, E. M. and Thompson, E. A. (2001) Markov chain Monte Carlo methods for the calculation of likelihoods in genetic linkage studies. *American Journal of Human Genetics* **69:** S411 (#1337).

37. George, A. W., Wijsman, E. M. and Thompson, E. A. (2002) Detecting disease genes via a new Markov chain Monte Carlo approach for multipoint linkage analysis. *Genetic Epidemiology* **23:** 283.

38. Leutenegger, A-L., Prum, B., Clerget-Darpoux, F., and Thompson, E. A. (2002) Estimation of the inbreeding coefficient from multipoint marker data. *Genetic Epidemiology* **23:** 292.

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

d: Research Design and Methods

d.1 Methods and models for gene descent in pedigrees.

(i) Block-Gibbs MCMC sampling and diagnostics for gene descent at marker loci

The primary basis of our methods is the Monte Carlo realization of meiosis indicators $\mathbf{S} = \{S_{i,j}\}$ of equation (2) in **b.5**, conditional on marker and/or trait data $\mathbf{Y}$ observed for some members of an extended pedigree.

First, we focus on improved methods for sampling inheritance vectors $\mathbf{S}$ at the marker loci of a linkage group conditional on marker data $\mathbf{Y}_M$, in the absence of genetic interference Realizations under more general meiosis and data models then follow (see **(iii)** below). Our methods are based on a combination of two block Gibbs samplers developed previously under this grant. These are the L-sampler (Heath 1997) and M-sampler (Thompson and Heath 1999) combined to provide what we have called the LM-sampler (Thompson 2000b).

First, since $\mathbf{S}$ determines the descent of genes within the pedigree, these realizations may be used to estimate gene identity by descent (*ibd*) among specified relatives. Second, since there is a recombination between locus j and j' in meiosis i if and only if $S_{i,j} \neq S_{i,j'}$, realizations of $\mathbf{S}$ provide immediate estimates of recombination frequencies under the assumed genetic map and conditional on observed marker data $\mathbf{Y}_M$. Third, from equation (3) the complete-data log-likelihood is

$$\log \mathrm{P}_\psi(\mathbf{Y}_M, \mathbf{S}) \;=\; \sum_{j=1}^{L} \log(\mathrm{P}_\psi(Y_{\bullet,j}|S_{\bullet,j})) + \log \mathrm{P}_\psi(S_{\bullet,1}) + \sum_{j=2}^{L} \log_\psi(\mathrm{P}_\psi(S_{\bullet,j}|S_{\bullet,j-1})). \tag{4}$$

Thus computation of expected conditional complete-data log-likelihoods (ECCDL) given $\mathbf{Y}_M$ using $\mathbf{S}$-values realized by MCMC is straightforward, and these can be partitioned by penetrance contribution $\mathrm{P}_\psi(Y_{\bullet,j}|S_{\bullet,j})$ for each locus j and recombination contribution $\mathrm{P}_\psi(S_{\bullet,j}|S_{\bullet,j-1})$ from each marker interval (Thompson 2000b). We routinely compute these general diagnostic gene descent, recombination count, and ECCDL statistics in our MCMC programs. We propose now to formalize these diagnostic criteria. By considering the component terms of the ECCDL, we will identify MCMC runs with unusual recombination or gene descent patterns,

New programs by several groups are increasing the size of pedigrees on which it is possible to compute exact probabilities of each inheritance vector $S_{\bullet,j}$ conditional on the multilocus data $\mathbf{Y}_M$ (see **b.5**), or even, on very small pedigrees, the pairwise probabilities $\mathrm{P}_\psi(S_{\bullet,j}, S_{\bullet,j+1} \mid \mathbf{Y}_M)$ for adjacent marker locations (Thompson 2000b). On small pedigrees, it is thus possible to compute exact values for our marginal single-locus gene descent measures, expected single-interval recombination frequencies, and ECCDL statistics. We propose extensive validation of our MCMC sampling by comparing exact and MCMC estimates whenever exact estimates are available, using our standard diagnostic statistics.

The increasing feasibility of exact computation does not obviate the need for Monte Carlo estimation. Whereas exact methods can compute only the marginal, or at best pairwise, probability for inheritance vectors $S_{\bullet,j}$ conditional on the full data $\mathbf{Y}_M$, Monte Carlo methods provide realizations of the full array $\mathbf{S}$ from the conditional distribution $\mathrm{P}_\psi(\mathbf{S} \mid \mathbf{Y}_M)$. Thus from Monte Carlo realizations, it is possible to score *ibd* across loci, to identify shared multilocus haplotypes among pedigree members, and to investigate recombination events jointly over marker intervals. We propose to expand our statistics to score such multi-locus and multi-interval events. We will then exploit this joint information in the development of *ibd*-based statistics for linkage detection (see **d.3**) and map heterogeneity analyses (see **d.4**). Thus our methods will combine the confidence of exact marginal computation with the greater scope of joint MCMC realizations across multiple loci.

The increasing scope and flexibility of exact computation also enables us to improve our block-Gibbs samplers. The performance of the L-sampler can be improved by peeling and resampling jointly at very tightly linked loci, where this is feasible. Likewise the M-sampler can be improved by choice of meioses for joint updating: Thomas et al. (2000) have proposed some easily implemented joint meiosis updates that improve mixing. A multi-meiosis M-sampler update is computationally equivalent to a Baum-algorithm *ibd* computation on a pedigree with that number of meioses. Thus improved exact computational algorithms can always be used to enhance MCMC performance. We propose to investigate the feasibility, practicality, and performance of MCMC samplers that update several loci and several meioses jointly, and to investigate the choice of such meioses and loci. We will also continue to investigate the optimal balance of L-steps and M-steps in the LM-sampler for improved computational efficiency. This balance depends on the pedigree structures, marker polymorphism, and the extent of missing data which all affect the computational burden of an M-step or L-step in different ways.

In addition to diagnostic statistics computed after an MCMC run, we propose to implement and test dynamic visualization as a means of assessing sampler mixing (or failure thereof). We have experimented with the OpenGL/Glut-based library for dynamic visualization of MCMC realizations (www.opengl.org). In a prototype version of one MORGAN program, we can track the latent variable state changes in real time. The code is portable and has been tested on various platforms including Linux, Tru64 Unix, and Windows. We use the Mesa 3D Graphics Library (www.mesa3d.org) which is a freely available open-source implementation of the OpenGL specification: OpenGL is the industrial standard for high performance graphics with many different vendor implementations. We propose to implement the option of dynamic visualization options into our sampling of latent variables **S** that specify the descent of genes on chromosomes. Since **S** determine the descent of founder genes on a pedigree, we will assign a color to each of a limited number (say, 10) founder haplotypes of interest, all other founder haplotypes remaining uncolored. Then as successive **S** are sampled by MCMC we will display in real time the colorings of the implied haplotypes in current individuals of interest. These may be the haplotypes of affected individuals, of key members in the pedigree lines on descent, or of individuals whose meioses suggest unusual recombination on the basis of recombination event counts. While complete visualization of gene descent in an entire pedigree is impractical, selective choice can show alternative descent modes on the pedigree and help to identify and resolve mixing problems.

(ii) Metropolis-Hastings MCMC methods for gene descent in pedigrees

While providing much more accurate realizations than previous samplers, the LM sampler still encounters mixing problems for multiple linked loci on extended pedigrees with substantial missing data. We propose to enhance the performance of the LM-sampler by adding a variety of Metropolis-Hastings (M-H) steps (Hastings 1970) to the block-Gibbs updates of the LM-sampler. Such hybrid samplers are not new in MCMC (Tierney 1994), and not new to our methods. The Bayesian oligogenic linkage analysis methods of Heath (1997) include M-H proposals for the update of quantitative trait loci (QTL) locations and effect parameters. Our new pseudo-Bayes sampling over trait locus locations also involves a M-H proposal and acceptance step. We now propose to incorporate M-H proposals also into the sampling of **S**.

An initial step in this direction has been taken with our implementation of sequential imputation (SI) realizations of **S** (Kong et al. 1994). We have tested these both to provide starting configurations for an MCMC run, and as an additional M-H component of the overall MCMC sampling procedure (George and Thompson 2003). For small numbers of loci, the M-H SI proposals worked well, and in several examples served to allow the MCMC to sample alternative haplotype patterns over 3 or 4 loci whereas our previous samplers had not (George et al. 2001). However, over large numbers of linked loci SI proposals were almost never accepted by the M-H procedure. We propose to modify SI proposals, proposing update only of a block of inheritance vectors $\{S_{\bullet,j}; j \in J\}$ conditional on inheritance patterns at the markers flanking the block J. We believe this will provide better acceptance probabilities, and will be computationally feasible in many cases in which the joint updating of loci in J via the L-sampler is not. We will also investigate the order of imputation for loci within the block. In the importance-sampling estimation of linkage likelihoods via SI the chosen order of loci can have a strong impact (Irwin et al. 1994), and we have seen the similar effects on acceptance probabilities even for small numbers of loci.

Our implementation of SI to provide starting values for an MCMC run had unexpectedly poor performance. Assessment of the reasons for this leads to a new proposal and potential major improvement in the MCMC sampling of **S**. As expected, these SI starting configurations had much higher values of $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y}_M)$ than those produced by our previous locus-by-locus based initial configurations. However, linkage likelihoods estimated from MCMC runs initiated by SI were often much more poorly estimated than from MCMC runs initiated by locus-by-locus sampling. SI focuses initial realizations on a particular high-probability haplotypic *ibd* configuration, and thereafter MCMC sampling failed adequately to explore the full probability space $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y}_M)$. Using a locus-by-locus initial configuration $\mathbf{S}^{(0)}$, the MCMC must first eliminate large numbers of recombination events implied by $\mathbf{S}^{(0)}$, before $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y}_M)$ can reach acceptable values. This we accomplish by burn-in MCMC using the marker data alone (treating any trait loci as absent or unlinked). For the few examples as yet studied, in about 30% of cases, the configuration reached will local to that provided by SI. In as many as 50% of cases, the MCMC will fail to reach a mode with high value of $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y}_M)$, due to failures to eliminate initial recombination events between tightly linked loci. However, in the remaining 20% of cases, MCMC sampling from an initial locus-by-locus configuration produces a haplotypic *ibd* configuration substantially different to that produced by SI, but with comparable $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y}_M)$. That is, a quite different but comparable local mode of $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y}_M)$ has been identified.

The definition and determination of modes of $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y}_M)$ in the discrete high-dimensional structured space of meiosis indicators $\mathbf{S}$ or latent multi-locus genotypes is a long-standing problem (Lin 1996). If such modes can be identified characterized by proposed state values local to this mode, then a variety of mode-hopping MCMC procedures (Besag and Green 1993) are available to us. Note that the characterizing state need not be the local maximum, it need only have sufficiently high $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y}_M)$ in the complex space of $\mathbf{S}$ values to be accepted by a M-H proposal step. Thus we propose to use locus-by-locus initial configurations, and MCMC burn-in possibly in conjunction with gradual annealing of the genetic map, to produce a collection of maybe 500 potentially quite different configurations $\mathbf{S}$ all having comparable probability $\mathrm{P}_\psi(\mathbf{S}|\mathbf{Y}_M)$. We need not attempt to define modes, or to characterize or group these candidates: each is a candidate that can be proposed at any M-H step. These configurations are "over-dispersed" in the sense of Gelman and Rubin (1992). Generally, finding overdispersed candidates in a discrete highly structured space is a difficult task, since there is no metric to define the dispersion. Here the deficiencies of the locus-by-locus sampling work to our advantage.

(iii) Conditional sampling and importance sampling reweighting of gene descent patterns

The advantage of sampling the meiosis array only at marker loci conditional only on joint marker data $\mathbf{Y}_M$ is that this is relatively easily and efficiently done, and that realizations can be simply combined over independent sampling on disjoint linkage groups (chromosomes) and disjoint pedigrees. The disadvantage is that these realizations are rarely all that are required. We show here how subsequent conditional sampling and importance sampling reweighting can address issues of haplotype sharing, trait model incorporation and assessment, and map, meiosis, and marker model heterogeneity.

Consider first the realization of gene descent at inter-marker locations, λ, conditional on the marker data $\mathbf{Y}_M$. It is inefficient to include non-data chromosomal locations λ in the MCMC sampling. A simple and effective procedure is to realize $S_{\bullet,\lambda}$ given each realized $\mathbf{S}$ using the L-sampler (Skrivanek et al. 2002). We propose to compare this approach with a method that will permit the imputation of the descent of entire chromosome segments. Since the markers in a genome scan are often less than 10cM apart, the probability of more than one crossover event in a single gamete and single marker interval is small, even under a no interference model. In reality, genetic interference suggests that precluding this possibility may be a more accurate model (King et al. 1991). Then, given the inheritance vectors at flanking markers j and j', either $S_{i,j} = S_{i,j'}$ in which case we assume the entire segment in inherited intact in meiosis i, or $S_{i,j} \neq S_{i,j'}$ in which case the single crossover event is uniformly distributed (in terms of genetic distance) between j and j'. Although, on an extended pedigree with many meioses i, there may in all be many recombination events between two adjacent markers, the assumption that only one can occur in any given meiosis leads to practical methods for Monte Carlo estimation of the descent pattern of the segment (j, j') conditional on $S_{\bullet,j}$ and $S_{\bullet,j'}$. Hence, since $\mathbf{S}$ is realized jointly at all markers j, we have realization of the descent patterns of each entire chromosome (see also **d.3** and **d.4**).

Lange and Sobel (1991) developed the first MCMC approach for multipoint estimation of linkage lod scores for a trait locus location, λ, given trait data Y_T and marker data $\mathbf{Y}_M$. The parameters of the marker model and map are denoted by Λ_M, and the remaining trait model parameters by β. In our current notation, Lange and Sobel (1991) developed the likelihood in the form

$$\mathrm{P}(Y_T, \mathbf{Y}_M \mid \lambda, \beta, \Lambda_M) \quad \propto \quad \mathrm{P}(Y_T \mid \mathbf{Y}_M, \lambda, \beta) \quad = \quad \sum_{\mathbf{S}} \mathrm{P}(Y_T|\mathbf{S}, \lambda, \beta)\ \mathrm{P}(\mathbf{S} \mid \mathbf{Y}_M, \Lambda_M) \tag{5}$$

where $\mathbf{S}$ is the meiosis array at marker loci only. This form has many advantages, since the MCMC sampling need be done once only, under the fixed marker model Λ_M. The meiosis array $\mathbf{S}$ is repeatedly realized conditional on $\mathbf{Y}_M$ by MCMC, and the probability $\mathrm{P}(Y_T|\mathbf{S}, \lambda, \beta)$ is then computed for realizations $\mathbf{S}$ for each location λ and trait model β of interest.

Where a trait allele is rare, and the trait close to Mendelian, this approach can perform poorly, since the trait data themselves convey much of the trait locus segregation information on an extended pedigree. In considering the ancestry of a rare recessive allele, on an extended pedigree with few marker data, it is essential to include the trait locus in the MCMC sampling (Thompson 2000b). For an incompletely dominant trait we have also found increased computational efficiency in including the trait locus in the MCMC sampling using our pseudo-Bayesian approach (George and Thompson 2003). However, for a complex trait the trait distribution on a pedigree neither can nor should dominate over marker data in the imputation of gene descent, and thus the method characterized by equation (5), should be computationally efficient. For simple traits, we have implemented this approach, using our current MCMC samplers in the MORGAN program **lm_markers**: for proposed developments for complex traits, see section **d.2**.

Our primary tool is the realization of $\mathbf{S}$ at marker data, given marker data $\mathbf{Y}_M$, in the absence of interference, under either sex-averaged or sex-specific genetic maps. However, realizations may be readily reweighted to accommodate alternative marker models that do not differ too strongly from that under which the sampling is done. The basic importance-sampling formula result is that each realization $\mathbf{S}$ sampled under marker model Λ_M conditional on marker data $\mathbf{Y}_M$ is equivalent to a realization of weight $\mathrm{P}(\mathbf{S}|\mathbf{Y}_M, \Lambda_M^*)$ / $\mathrm{P}(\mathbf{S}|\mathbf{Y}_M, \Lambda_M)$ under the model Λ_M^*. Thus instead of resampling, realizations may be simply reweighted to alternative maps and models. This provides very effective methods for analysis under deviations from the MCMC sampling model Λ_M, although sampling under sex-specific maps may still be more efficient if map differences are extreme. It provides for analysis under interference, provided the interference model Λ_M^* is not so extreme as to give 0 weight to almost all sampled realizations. Since a realization $\mathbf{S}$ together with observed marker data $\mathbf{Y}_M$ determine (partial) haplotypes of founders, incorporation of disequilibrium among loci in the founder haplotypes can also be achieved by reweighting. The extent to which this is computationally feasible, or more effective than MCMC sampling under different marker models is addressed in section **d.4**.

d.2 Likelihood and Bayesian methods for complex traits.

(i) A framework of trait models for MCMC sampling

We propose to develop a general framework of trait models that can be integrated into our MCMC methods. We have previously implemented trait models in the context of specific analyses. For a quantitative trait, we have used a mixed model (Guo and Thompson 1994; Snow and Wijsman 1998), while for multiple correlated traits we have used polygenic models (Daw et al. 1999b; George et al. 2003). The Bayesian approach in **Loki** permits multiple QTL contributing additively to a trait (Heath 1997) and this model has also been extended to censored age-of-onset data (Daw et al. 1999a). In MCMC-based linkage likelihood analyses we have used a single-locus model with incomplete penetrance (George et al. 2003), and have incorporated age-of-onset data using liability classes (George et al. 2002). However, with the exception of the oligogenic model in **Loki**, trait models have been *ad hoc*. We have not integrated a general framework for trait modeling and specification into our MCMC methods. This we propose now to do.

In any MCMC incorporating trait information Y_T, the key component is the probability $\mathrm{P}(Y_T|S_{\bullet,\lambda})$ of the trait data given the inheritance vector at hypothesized trait location λ. This in turn provides, either by exact or Monte Carlo integration over $S_{\bullet,\lambda}$, the probability of Y_T given the inheritance vectors at the flanking marker loci. A complex trait model need not be incorporated into the MCMC sampling itself. Importance sampling and equation (5) allow phenotypic data to be incorporated into likelihood and probability estimates using realizations of the joint meiosis array $\mathbf{S}$ only at marker loci given marker data $\mathbf{Y}_M$ (see **d.1 (iii)**).

The computation of $\mathrm{P}(Y_T|S_{\bullet,\lambda})$ is not straightforward for general trait models, but for single-locus trait models an algorithm due to Heath (Thompson and Heath 1999) has been implemented in MORGAN. Through conditioning in MCMC sampling, the single-locus computational framework may be extended. For example, for QTL contributing additively to a quantitative trait (Heath 1997), or a quantitative trait with polygenic contributions (Guo and Thompson 1994), a trait locus may be updated conditionally on the contributions of other loci or heritable effects. In effect, the computation is reduced to the single-locus case. Initially, we will restrict analyses to trait models falling within the framework of locus-by-locus computation, but we propose a more general trait model specification.

A key requirement for a general trait model framework is thus to distinguish traits from loci. We propose to model four trait types: genotypic, discrete (qualitative), ordered categorical, and quantitative. Each trait will have specified sets of associated loci and covariates. A locus has a known or hypothesized chromosomal location, alleles, allele frequencies, and possible allelic population associations with other loci. The models for loci derive from population models for genotypes pedigree founders, and the meiosis models that determine segregation to non-founders. The covariates are those that affect the phenotype-genotype relationship. These may include sex, age, and other qualitative or quantitative measures. Genotypes at other loci may also be treated as covariates, such as the treatment of ApoE as a genetic covariate in analysis of Alzheimer's disease

(Daw et al. 2000a).

In general, several traits may be associated with a given locus, or a single trait may be affected by several loci. The model for a discrete trait is specified by a penetrance matrix $\mathbf{Q} = (Q_{g,y})$, where $Q_{g,y}$ is the probability of phenotype y given genotype g specified for each possible individual genotype combination g at the contributing loci. There is no distinction between disease phenotypes and marker phenotypes. A genetic marker observed without error is a genotypic trait, and $\mathbf{Q}$ is the identity matrix. With an error model, it becomes a discrete trait. A genotypic trait will often correspond to a single mapped marker locus, but, for example, the haplotypes at tightly linked markers could be modeled as a single genotypic trait. For a discrete trait, without associated covariates, in principle one simply specifies the matrix $\mathbf{Q}$. However, some constraints on $\mathbf{Q}$ are necessary. Order constraints on penetrance values define the disease predisposing alleles or genotypes, and equality constraints reduce the number of parameters. At each locus we will allow for the specification of (incomplete) dominant and recessive traits, with multiplicative or additive action of alleles on penetrance probabilities. We will allow for multiplicative or additive effects across loci, and for some simple forms of epistasis. Specifically, we will implement the epistatic and heterogeneity models developed by Neuman and Rice (1992). This requires methods for computation of the probabilities $\mathrm{P}(Y_T|\{S_{\bullet,\lambda_k}\})$ of phenotypic data Y_T given the inheritance vectors jointly at the hypothesized locations λ_k of the set of interacting trait loci.

For discrete and ordered categorical traits, a convenient way both to specify penetrance and incorporate covariate information into a trait model is through the use of a quantitative liability or a qualitative liability class (Ott 1999). Penetrance for a dichotomous trait is then a function of the observable liability x and genotype g. We propose to use a logistic form as our penetrance in terms of liability. Let

$$F_Z(z; x, \alpha_g, \gamma_g) = (1 + \exp(-\gamma_g(z - x - \alpha_g)))^{-1}$$

Here Z may be regarded as an unobservable latent liability, having a logistic distribution centered at $x + \alpha_g$ and with scale parameter $\gamma_g > 0$. Then, for example, for a dichotomous trait we may define

$$\mathrm{P}(\text{unaffected } |g, x) = 1 - \mathrm{P}(\text{affected } |g, x) = F_Z(0) = (1 + \exp(\gamma_g(x + \alpha_g)))^{-1}$$

Choosing the threshold $z_1 = 0$ imposes no restriction: it simply defines location on the latent liability axis, providing identifiability of parameters α_g. For an ordered categorical trait with K classes it is necessary to define $K - 2$ additional thresholds $z_i,\ i = 2, ..., k - 1$. Then, with $z_0 = -\infty$ and $z_K = \infty$,

$$\mathrm{P}(\text{phenotype i} \mid g,\ x) = F_Z(z_i; x, \alpha_g, \gamma_g) - F_Z(z_{i-1}; x, \alpha_g, \gamma_g) = \mathrm{P}(z_{i-1} < Z \leq z_i; x, \alpha_g, \gamma_g)$$

To reduce the number of parameters, it may be desirable to have γ_g independent of genotype, but we will make our framework flexible enough to accommodate genotype-dependent values.

Finally, a form is required for the covariate liability x as a function of observed covariates (age, gender, and other quantitative or qualitative observations on individuals). We propose to do this though specification of a linear multiple regression form, with regression coefficients for quantitative covariates and additive effects on x for qualitative covariate classes. The user may also specify the inclusion of specified interaction effects on covariate liability x of either qualitative or quantitative covariates. Age-of-onset may be treated either as a covariate or as a quantitative trait. As a covariate, age is inversely related to liability: individuals affected at earlier ages have higher liability, while those unaffected at older ages have lower liability. However, there may also be pedigree effects; for example, age-of-onset may vary among pedigrees (Bird et al. 1989). The pedigree will be included as a potential qualitative covariate. Of course, the data from a given study will normally only support the *estimation* of a small number of penetrance parameters, and the user will need to exercise caution in the inclusion of effects. However, through the flexibility in our parameter specification, the user will be able to investigate alternative penetrance scenarios. Covariate information and the resulting liabilities are required only for individuals whose phenotypes are observed. The use of a covariate liability to summarize covariate information, also makes it easy to deal with missing covariate observations. Using covariate models we will impute missing covariates (see **(ii)** below).

Quantitative traits also have lists of contributing loci and covariates. Genetic effects may be specified as the additive and dominance effects of QTL, and possible interaction (epistasis) effects. Covariate effects may be specified either as regression parameters for a quantitative covariate or as the additive effect of a qualitative effect class, and hence use the same framework of liability specification as for the ordered categorical traits. Additionally, there may be other genetic components of quantitative traits. Additional heritable variation may be modeled as a polygenic component with additive and residual variances. We will allow also for the inclusion of a sibship effect. In human pedigrees, sibs are the only common bilateral relatives, and a sibship variance component allows for excess similarity among sibs without the modeling and computational complications of general heritable variance components (Thompson and Shaw 1992). Quantitative traits are considered more fully in **(ii)** below.

(ii) Bayesian inference for oligogenic models

For models with a large number of parameters, Bayesian analysis offers many advantages. Integrating over parameters, with respect to prior distributions, avoids the need to maximize over nuisance parameters or to view a high-dimensional likelihood surface. Inferences are based on posterior distributions of a few parameters of interest (Daw et al. 1999a). MCMC procedures are particularly well suited to such procedures, since the integration is accomplished by sampling in the course of the MCMC process. These procedures have proved particularly successful for quantitative traits (Yuan et al. 2000; Gagnon et al. 2002), where the number of contributing QTL, their locations, additive and dominance contributions, allele frequencies, and interaction effects can all be sampled over.

There are often many missing covariate data in genetic studies. This currently necessitates dropping associated trait data, leading to inefficient use of the available data. We propose to implement, in the context of MCMC procedures, data augmentation techniques, based on multiple imputation (Rubin 1987). The parameters estimated for the data augmentation are nuisance parameters in the context of linkage analysis, with the MCMC process integrating over these parameters. We will obtain parameters that allow us to sample values for the missing covariate data, which can then be used with all the trait data in a linkage analysis. Three types of covariate data exist, two of which also have a genetic basis: major gene covariates, other traits that may also be used as covariates (trait-covariates), and other types of covariates that are not believed to have a genetic basis. For all covariate types, it would be important to not only implement an approach to impute the missing data, but also to determine what the optimal frequency of re-imputing data is, since it may be inefficient to resample missing covariate data on every MCMC cycle.

For genetic covariates we propose to build on our genetic model framework in order to obtain samples from the missing covariate data. We already obtain missing values for major gene covariates (**d.2(i)**). We propose to take this idea one step further in imputation of trait-covariate data. For such a phenotypic covariate, we will use sampled genotypes and model parameter estimates from segregation analysis based on the observed data to sample missing phenotype values: these are then used as covariates for the main trait analysis. For a trait analysis that conditions on such trait-covariates, we propose to use a cycle of trait-covariate phenotype imputation followed by linkage analysis for the trait based on complete covariate data. The trait-covariate phenotypic sampling could be done during segregation analysis alone, which would be computationally rapid, or as part of a joint linkage and segregation analysis for the trait-covariate. With this framework it is straightforward to incorporate the various models that we develop to handle different kinds of trait phenotypes (**d.2(i)**).

For non-genetic covariate data X it is also possible to use MCMC to sample missing values. To do this, we still will need to specify distribution(s) for the missing values. For this purpose, however, it may be possible to pre-specify a small number of fairly general distributions with few parameters (e.g., a gamma distribution) since the goal is to capture, possibly only approximately, the missing value distribution. With this, it should be possible to construct a Markov chain in which a sample of missing (X_{mis}) values is sampled on the basis of observed (X_{obs}) data and the parameters of the model, followed by model parameter estimation (or sampling), given the complete phenotypic data.

One additional aspect that we propose to address in the context of Bayesian oligogenic analysis is incorporation of multiallelic QTL models. For common, complex traits (e.g., the triallelic APOE system and AD risk) the current diallelic QTL model may be inadequate. However, the number of alleles at a QTL would not ordinarily be known ahead of time. We propose to use reversible-jump MCMC (Green 1995) to address this shortcoming. Because the number of genotypes becomes large rapidly with the number of alleles, it will be necessary to restrict to additive contributions of alleles to limit the number of parameters.

(iii) Model estimation in combined segregation and linkage analysis

While a Bayesian analysis may always be used to integrate over model parameters, where there are few nuisance parameters it may be desirable to estimate parameters and compare likelihoods of alternative models. The oligogenic model is well suited to Bayesian analysis, but a mixed model with a single major locus and additional heritable variation has a more limited parameter set. For a mixed model, likelihood analysis is appropriate. For convenience, we divide the genetic model parameters as $\psi = (\beta, \lambda, \Lambda_M)$ where β summarizes the parameters of the trait model, λ is a hypothesized trait locus location, and Λ_M summarizes all the parameters of the genetic marker map and loci.

We propose to combine our earlier development of MCMC methods for estimating genetic models with our newer approaches for sampling marker inheritance configurations. While an earlier implementation combined Monte Carlo EM (MCEM) segregation analysis for quantitative traits (Guo and Thompson 1994) with single diallelic marker linkage analysis (Guo and Thompson 1992), our newer LM-sampler (Thompson and Heath 1999) and pseudo-Bayesian samplers (George and Thompson 2003) now make possible a combine segregation and multilocus linkage analysis under fixed marker model Λ_M. First, we will reimplement mixed model MCEM approaches, combining MCEM over trait model parameters β with MCMC sampling of marker and trait locus inheritance patterns $\mathbf{S}$ and trait variance component effects, $\mathbf{Z}$. We will use the LM-sampler to sample $\mathbf{S}$ conditionally on data $\mathbf{Y}$ and current values of $\mathbf{Z}$, and will update $\mathbf{Z}$ jointly over the pedigree, conditional on $\mathbf{Y}$ and $\mathbf{S}$. Current peeling programs for a QTL provide the probability $\mathrm{P}_\beta(Y_{\bullet,T}^{(\mathbf{Z})}|S_{\bullet,T})$ of trait data $Y_{\bullet,T}$ adjusted by $\mathbf{Z}$, given the inheritance vector $S_{\bullet,T}$ at the trait locus. Resampling of $\mathbf{Z}$ can be achieved by using polygenic peeling programs.

We will compare the ability of MCEM and stochastic approximation (SA) to provide joint estimates of λ and β. SA (Younes 1988; Moyeed and Baddeley 1991) is an alternative when MCEM becomes too computationally intensive. This can occur because MCEM requires multiple realizations of latent variables at each step of the EM process, in order to estimate the conditional expectations required for the EM algorithm. A large number of EM steps may be required. In SA, rather than taking multiple realizations in order to estimate the conditional expectations used in the EM algorithm, at each step a single realization of the latent variables is made, and a small step taken in the direction of the gradient of the complete-data log-likelihood:

$$\bigtriangledown \log \mathrm{P}_\psi(\mathbf{Y}, \mathbf{S}, \mathbf{Z}) \;=\; \bigtriangledown \log \mathrm{P}_\beta(\mathbf{Y}|\mathbf{S}, \mathbf{Z}) \;+\; \bigtriangledown \log \mathrm{P}_\beta(\mathbf{Z}) \;+\; \bigtriangledown \log \mathrm{P}_{(\lambda,\Lambda_M)}(\mathbf{S})$$

Given an already existing MCEM program, SA is easily implemented. The complete-data log-likelihood and its gradient take a simple form, and can be computed.

Once a model is estimated, we will derive Monte Carlo estimates of the lod score (equation (1)) under a given trait model β, and explore lod score sensitivity to changes in β. A basic equation in exploring a local likelihood surface is

$$\frac{L(\psi)}{L(\psi_0)} \;=\; \frac{\mathrm{P}_\psi(\mathbf{Y})}{\mathrm{P}_{\psi_0}(\mathbf{Y})} \;=\; \mathrm{E}_{\psi_0}\left(\frac{\mathrm{P}_\psi(\mathbf{Y}, \mathbf{S})}{\mathrm{P}_{\psi_0}(\mathbf{Y}, \mathbf{S})} \;\Big|\; \mathbf{Y}\right) \tag{6}$$

(Thompson and Guo 1991). MCMC is used to sample $\mathbf{S}$ conditionally on data $\mathbf{Y}$ under model ψ_0, and the likelihood ratio $L(\psi)/L(\psi_0)$ is estimated by averaging the ratio of complete-data likelihoods $\mathrm{P}_\psi(\mathbf{Y}, \mathbf{S})/\mathrm{P}_{\psi_0}(\mathbf{Y}, \mathbf{S})$ over the MCMC run. For fixed β and Λ_M, equation (6) reduces to

$$\frac{L(\lambda)}{L(\lambda_0)} \;=\; \mathrm{E}_{(\beta,\lambda_0,\Lambda_M)}\left(\frac{\mathrm{P}_\lambda(S_{\bullet,T}|\mathbf{S}_M)}{\mathrm{P}_{\lambda_0}(S_{\bullet,T}|\mathbf{S}_M)} \;\Big|\; \mathbf{Y}\right) \tag{7}$$

We have previously used equation (7) to estimate location lod score curves for λ (Thompson 2000b). Although providing accurate and reliable estimates of the relative likelihoods for positions λ within a marker interval, equation (7) does not provide good estimates of the relative likelihoods between marker intervals. This is because the probabilities $\mathrm{P}_\lambda(S_{\bullet,T}|\mathbf{S}_M)$ can differ dramatically if λ and λ_0 are in different marker intervals. Thus we will use our pseudo-Bayes method (George and Thompson 2003) to estimate relative likelihood of locations in different marker intervals under each trait model β.

We will exploring likelihood changes due to changes in trait model parameters β. Where a likelihood surface itself and the distribution of latent variables $\mathbf{S}$ are both smoothly varying in the changing parameter, equation (6) provides an effective method of exploring local likelihood surface. For a fixed (λ, Λ_M), equation (6) becomes

$$\frac{L(\beta)}{L(\beta_0)} = \mathrm{E}_{(\beta,\lambda_0,\Lambda_M)}\left(\frac{\mathrm{P}_\beta(Y_T|S_{\bullet,T})}{\mathrm{P}_{\beta_0}(Y_T|S_{\bullet,T})} \mid \mathbf{Y}\right) \tag{8}$$

Once a model β_0 is estimated, the LM-sampler provides realizations of $\mathbf{S}$ conditional on data $\mathbf{Y}$ at any given λ_0, and sensitivity of the linkage likelihood to the trait model β can be explored. Alternatively, our pseudo-Bayes approach realizes $(\mathbf{S}, \lambda)$ jointly from the distribution $\mathrm{P}(\mathbf{S}|\mathbf{Y}, \lambda)\pi(\lambda)$ where $\pi(\lambda)$ is the chosen one-dimensional discrete pseudo-prior on λ. Thus, by reweighting by $\pi(\lambda)$ and marginalizing with respect to λ, the realizations $\mathbf{S}$ obtained in the course of the linkage likelihood estimation can be used directly in equation (8).

d.3 Linkage detection: comparison of methods

(i) Measures of *ibd* on extended pedigrees

Measures W of *ibd* in a pedigree are functions of the inheritance vector $S_{\bullet,\lambda}$ at a given location λ of interest. Whittemore and Halpern (1994) developed tests derived from such *ibd* measures W for trait linkage at a chromosomal location λ. Based on marker data $\mathbf{Y}_M$ the test statistic takes the form

$$T(\lambda, \mathbf{Y}_M) = \mathrm{E}(W(S_{\bullet,\lambda}) \mid \mathbf{Y}_M) = \sum_{\mathbf{S}} \mathrm{E}(W(S_{\bullet,\lambda}) \mid \mathbf{S})\mathrm{P}(\mathbf{S} \mid \mathbf{Y}_M). \tag{9}$$

This approach has been extended to small pedigrees by Kruglyak et al. (1996) (see **b.3**). An approach using such a measure is also ideally suited to analyses on extended pedigrees, using data at multiple linked marker loci. Where the probabilities $\mathrm{P}(S_{\bullet,\lambda}|\mathbf{Y}_M)$ cannot be computed exactly, they may be estimated by MCMC. MCMC realizations from $\mathrm{P}(\mathbf{S}|\mathbf{Y}_M)$ provide a direct estimate of T at marker locations: we consider below the estimation of $T(\lambda, \mathbf{Y}_M)$ at inter-marker locations. In fact, MCMC provides far more: it provides realizations of the meiosis array $\mathbf{S} = \{S_{\bullet,j}\}$ at all marker loci j given marker data $\mathbf{Y}_M$, and hence realizations $W(S_{\bullet,\lambda})$ given $\mathbf{Y}_M$. These in turn provide an estimate of the full probability distribution of $W(S_{\bullet,\lambda})$ given $\mathbf{Y}_M$, and not simply its expectation T. Inclusion of an unlinked "null" locus with no data in the MCMC provides the pedigree-based prior probability distribution for W. We propose to develop linkage detection methods that are based on an *ibd* measure W, but which are based on a comparison of the full probability distributions of W conditionally on marker data $\mathbf{Y}_M$, in the presence and absence of linkage (see **(ii)** below).

Next, we address the choice of *ibd* measure W. Two widely-used *ibd* measures are the S-pairs and S-all measures of Whittemore and Halpern (1994). The S-pairs measure considers the mean over all pairs of affected individuals of the *ibd* between the two members of the pair. In considering only pairs, it ignores much of the joint segregation information in a pedigree structure, and thus defeats the purpose of using such pedigrees. At the other extreme, for A affected individuals in a pedigree with f founders, the S-all *ibd* measure requires summation over all 2^A choices of one gene from each affected individual of a product over all $2f$ founder alleles of a non-trivial occurrence count. This measure cannot even be computed on a pedigree with large numbers of affected individuals. The measures considered by McPeek (1999) (see **b.3**) are mostly intermediate between S-pairs and S-all in their use of information jointly on affected individuals and their computational practicality for use on pedigrees.

For large pedigrees, we propose to adopt a different class of *ibd* measures, taking an approach analogous to a lod score approach developed by J. Terwilliger in Trembath et al. (1997) (see Ott 1999, P.279). Information for linkage detection, whether lod-score-based or *ibd*-based, can be partitioned broadly into information among sibs within sibships, information among sibships within pedigrees, and information among pedigrees. Our measures consider greater than expected *ibd* among affected sibs, and/or lower than expected *ibd* with their unaffected sibs, and can be expressed as the sum over sibships of *ibd* sharing among sibships of certain founder alleles (Basu et al. 2002). We propose to study this class of statistics, which involves several parameters depending on the weighting accorded to the assessed importance of affected vs unaffected, and within-sibship vs among-sibship information. The advantages for our analyses on extended pedigrees are ease of computation and evaluation. The recursive descent computation methods of Thompson (1988) are well suited to exact computation of probability distributions of such measures.

Finally in this subsection, we consider computation of terms such as $\mathrm{E}(W(S_{\bullet,\lambda}) \mid \mathbf{S})$ of equation (9). MCMC methods realize the inheritance vectors $S_{\bullet,j}$ only at (trait or marker) data locations. To consider inter-marker locations we require realizations, expectations, or probabilities of $W(S_{\bullet,\lambda})$ given $\mathbf{S} = \{S_{\bullet,j}\}$. Two alternative Monte Carlo methods for such realizations were given in section **d.1 (iii)**, and we propose to compare their performance in the context of realization of $W(S_{\bullet,\lambda})$. Additionally, we propose to extend the exact recursive gene descent computational algorithms of Thompson (1988) to compute probabilities for $W(S_{\bullet,\lambda})$ conditionally on descent at flanking markers. This is both feasible and practical for large but not too complex pedigrees, and for some choices of the measure $W(S_{\bullet,\lambda})$.

(ii) Tests for linkage detection: estimation of significance

Computation of the statistic alone is insufficient. At a minimum, to assess the significance of any result, we need its distribution in the absence of any linked loci affecting the trait phenotypes of individuals. We propose here three alternative approaches to the assessment of significance: (1) the usual approach using statistics (9) but applied to our new *ibd* measures W, (2) a permutation-based approach, and (3) a new approach, providing direct Monte Carlo estimation of p-values for linkage, both against the pedigree-based null distribution, and also against the true null hypothesis of absence of linkage but presence of marker data. We propose to develop, implement and compare the three approaches.

(1) For an expectation-based test statistic, $T = \mathrm{E}(W(\lambda) \mid \mathbf{Y}_M)$ (equation (9)), the null expectation is $\mathrm{E}^*(T) = \mathrm{E}^*(\mathrm{E}_0(W|\mathbf{Y}^*_M)) = \mathrm{E}_0(W)$, while $\mathrm{var}_0(W) = \mathrm{var}^*(\mathrm{E}(W|\mathbf{Y}^*_M)) + \mathrm{E}^*(\mathrm{var}(W|\mathbf{Y}^*_M)) \geq \mathrm{var}^*(T)$. Here, both superscript $*$ and subscript 0 denote absence of linkage, but 0 denotes the pedigree prior distribution of an *ibd* measure W, while $*$ denotes an expectation over marker data sets $\mathbf{Y}^*_M$ having the same marker loci model, genetic map(s), and data availability as the actual data $\mathbf{Y}_M$, but in the absence of linkage. On an extended pedigree, using MCMC, the pedigree-based null moments of W are very easily estimated by including one unlinked, data-free, ("null") locus in the MCMC run (Thompson 2000b). That is, the null locus estimates both $\mathrm{E}_0(W) = \mathrm{E}^*(T)$ and $\mathrm{var}_0(W)$ which is an upper bound on $\mathrm{var}^*(T)$. This upper bound is often used to provide a conservative test for linkage (Kruglyak et al. 1996).

For given data $\mathbf{Y}_M$, $\mathrm{var}(W(S_{\bullet,j})|\mathbf{Y}_M)$ is also readily computed from the MCMC run, for any given locus j in the marker map, and, with additional sampling, the variance at inter-marker locations is also available (see **d.1 (iii)**). On extended pedigrees, the variance reduction due to presence of marker data can be huge: $\mathrm{var}^*(T) << \mathrm{var}_0(W)$. A test based on standardizing via the pedigree-null variance can be far too conservative. Moreover, the standard more informative *ibd* measures have highly skewed distributions in the presence of marker data but absence of linkage (McPeek 1999; Basu et al. 2002), making normality assumptions untenable. The same is likely true for our new statistics: any statistic which weights highly joint sharing among multiple individuals has the potential to exhibit extreme values.

We propose to investigate the expected distributions of *ibd* measures W, both in the presence and absence of linkage, over data realizations $\mathbf{Y}^*_M$ having the same genetic map, marker model, and marker data availability as actual data $\mathbf{Y}_M$. Using the MORGAN program **genedrop**, we propose to simulate marker data in accordance with the locus characteristics, genetic maps and data availability of the study data. We will investigate the conditional variance $\mathrm{var}(W|\mathbf{Y}_M)$ as a function of the data. We will evaluate the extent to which, on the pedigree structures and marker-data availability we have in actual studies, we can expect $\mathrm{var}_0(W)$ to underestimate $\mathrm{var}^*(T)$ in the absence of linkage. We have also developed a MORGAN program **markerdrop**, that simulates such marker data in the presence of linkage, on the basis of any partially specified pattern of descent at a hypothetical trait locus. Using this program we will investigate the behavior of alternative measures W, in terms of the full distributions of $T = \mathrm{E}(W|\mathbf{Y}_M)$. Thus we will obtain both empirical measures of linkage significance, and be able to estimate the power of alternative test statistics. The goal is to construct standardizations of test-statistics specific to the marker-data availability and pedigree structures of the study, in order to obtain more powerful, but still valid, tests for linkage detection.

(2) Permutation tests provide an approach to testing that is robust to the distributional details of test statistics. In the case of designed crosses, with large sibships, Churchill and Doerge (1994) developed permutation tests for linkage, permuting trait phenotypes but keeping marker data fixed. Such methods are not readily applicable to human pedigrees, since individuals are not exchangeable in their joint relationships with others. However, some individuals are exchangeable within a pedigree structure. These include sibs who have no offspring, and the two members of any founder couple. We propose to develop permutation tests, where the permutations are constrained to result in the same joint relationship among affected individuals. Initially, we

have used the exchangeability of final sibs and of founder couples, but these often provide little power (Basu et al. 2002). For a late-onset disease, sibs without offspring will often be unaffected, while on an extended pedigree founder couples are normally unobserved. We propose to investigate the permutation distribution of the statistic, when the permutation of trait values is not constrained to exchangeable relatives only, and we will investigate the weighting of alternative permutations in terms of the relationships they imply among affected individuals. The MCMC framework of sampling $\mathbf{S}$ conditional on marker data $\mathbf{Y}_M$ is very well suited to such a study. A large number of permutation sets of "affected" individuals are prespecified. The value of the statistic for the actual data, and for all of the trait value permutation sets are then computed in a single MCMC run.

(3) The approach of Thompson (2000a) to assessment of linkage evidence from *ibd* has been to compare the conditional and prior probabilities of specified genome sharing events: see **b.3**. We propose to adapt this approach to use an *ibd* measure W. Given the measure W_λ of *ibd* at location λ, MCMC may be used to estimate the cumulative distribution function of W_λ given marker data $\mathbf{Y}_M$ and this may be compared with the pedigree-based null distribution of W. Supposing larger values of W correspond to evidence for linkage, a p-value for linkage would be

$$p_0(\lambda, \mathbf{Y}_M) \quad = \quad \mathrm{P}(W_\lambda < W_0 \mid \mathbf{Y}_M),$$

where W_0 is a realization from the pedigree-null distribution. This is a p-value: that is, the probability that, under the null distribution, a larger value of the *ibd* measure is obtained than that obtained conditional on the actual data. That W_λ is an imputed measure conditional on $\mathbf{Y}_M$, rather than a statistic which is a deterministic function of $\mathbf{Y}_M$, does not affect the p-value interpretation.

Although our new approach provides a direct estimate of a p-value for linkage, the use of the pedigree-null distribution has the same disadvantages as for statistics $T = \mathrm{E}(W|\mathbf{Y}_M)$. With little more computation, we can obtain p-values relative to the no-linkage null distribution. Specifically,

$$p^*(\lambda, \mathbf{Y}_M) \quad = \quad \mathrm{E}^*(\mathrm{P}(W(\lambda) < W^*(\lambda)) \mid \mathbf{Y}_M, \mathbf{Y}^*_M).$$

To obtain this p-value we first estimate the cumulative distribution function (cdf) of $W(\lambda)$ given the actual observed data $\mathbf{Y}_M$. Next, for each of a number of data sets $\mathbf{Y}^*_M$ having the same map, marker model, and data availability as the actual data $\mathbf{Y}_M$, and obtain the estimated cdf of $W^*(\lambda)$ given $\mathbf{Y}^*_M$. The p-value is then obtained directly from the two cdf's. An expectation (average) is taken over additional realizations of pseudo-data $\mathbf{Y}^*_M$. As an empirical check, we will also estimate the jackknife distribution of p-values for the pseudo-data cdf's against each other, which should lead to a uniform $U(0,1)$ distribution of p-values, and also for each pseudo-data cdf against the pedigree-null cdf of W_0. Note that, on average, the pseudo-data W^*-values should have the same mean as W_0, with reduced variance. The estimation of cdf's of W^* and of W_0 provides an additional assessment of the reduction in variance and power of pedigree-null based tests.

Although gene *ibd* scoring statistics are based on imputed probabilities conditional of all the marker data $\mathbf{Y}_M$ observed on pedigree members over all the loci of the chromosome, the actual *ibd* measures that have been used are single-locus scores, functions of $S_{\bullet,\lambda}$ at location λ. Since MCMC provides realizations of the multilocus patterns of gene descent $\mathbf{S}$ given the marker data $\mathbf{Y}$, it provides the opportunity to investigate the power of statistics that measure *ibd* over several loci jointly or over chromosomal segments (see **d.1 (iii)**). Rather than pointwise sharing, we can impute the lengths of segments shared among different subsets of the individuals. Alternatively, with a complex multivariate phenotype, there may be sharing at some loci jointly with lack of sharing at others among phenotypic subgroups. We propose to develop such statistics, and investigate under what circumstances (if any) they provide additional power for linkage detection.

(iii) Comparison of *ibd*-based, Bayesian and likelihood methods

In addition to the investigation of performance of separate methods, as already outlined, we will be in a position to undertake a thorough comparison of the power of alternative approaches to detect and localize genes contributing to complex traits in the presence of heterogeneity. Using both simulation and real data (see **d.5**), we will compare the Bayesian (**d.2 (ii)**), likelihood (**d.2 (iii)**), and *ibd*-based (**d.3 (ii)**) approaches outlined in previous sections. It is important that comparisons assess the methods, rather than relative performance of alternative software or MCMC samplers. **Loki** and MORGAN share many features, including the LM-sampler for realization of meiosis array $\mathbf{S}$ conditional on marker and/or trait data.

Using simulated data, we will compare Bayesian oligogenic approaches using **Loki** with MCEM or SA approaches to joint segregation and linkage analysis using the mixed model, and assess the gains of using an oligogenic trait model with multiple QTL. We will examine the loss of power of likelihood methods in the presence of genetic heterogeneity, and compare Bayesian and *ibd*-based approaches in their ability to resolve such heterogeneity. We will examine the sensitivity of likelihood and Bayesian methods to severe model mis-specification, in comparison to the robustness of *ibd*-based methods. We will examine the loss of power of *ibd*-based methods, relative to Bayesian methods, when the model is at least approximately correct. By using the same data, same generation models, and same analysis models (where model specification is required), an objective comparison will be possible.

d.4 Linkage heterogeneity, data errors, and meiosis models

(i) Map estimation and analysis of map heterogeneity

The development of methods for the detection and analysis of map heterogeneity, on the basis of any available family and pedigree data, is timely given the increasing data-availability and evidence (see **b.4**). The MCMC approach provides a powerful tool to investigate genetic maps and the meiosis process. The sampling of the multilocus descent of genes, $\mathbf{S}$, conditional on marker data, $\mathbf{Y}_M$, provides conditional probability estimates of recombination events (see **b.5**). In fact, our current programs routinely summarize realized recombination counts by meiosis gender and chromosomal interval (see **d.1**). These in turn provide information for improved genetic map estimates and examination of map heterogeneity among individuals.

Given a set of (marker) loci, $j = 1, \ldots, L$, ordered along a chromosome, let $\rho_{i,j-1} = \mathrm{P}(S_{i,j-1} \neq S_{i,j})$ be the recombination frequency for meiosis i in marker interval $j-1$, between loci $j-1$ and j. Given a meiosis array $\mathbf{S}$ the implied recombination events are immediate: $R_{i,j-1} = |S_{i,j} - S_{i,j-1}|$ is a 0/1 indicator of recombination in marker interval $j-1$ in meiosis i. As a function of the recombination frequencies between adjacent marker loci, the complete-data log-likelihood of equation (4) may be written as

$$\log \mathrm{P}_\psi(\mathbf{Y}_M, \mathbf{S}) \;=\; \text{const} \;+\; \sum_{j=2}^{L} \sum_{i} \left(R_{i,j-1} \log \rho_{i,j-1} \;+\; (1 - R_{i,j-1}) \log(1 - \rho_{i,j-1}) \right).$$

In exact computation, computational efficiency demands that $\rho_{i,j-1}$ be constant over meioses i. In an MCMC analysis, variation in genetic maps among meioses creates no added computational burden, and may be incorporated either within the MCMC or by importance sampling reweighting (see **d.1 (iii)**). The realized recombination counts, from MCMC sampling of $\mathbf{S}$ given $\mathbf{Y}_M$, can be scored by meiosis and by marker interval.

The gradient of the complete-data log-likelihood, with respect to the recombination fraction $\rho_{I,j-1}$ for a particular marker interval $(j-1)$ and collection of meioses $i \in I$ is

$$\sum_{i \in I} R_{i,j-1} \;/\; \rho_{I,j-1} \quad - \quad \sum_{i \in I} (1 - R_{i,j-1}) \;/\; (1 - \rho_{I,j-1})$$

Based on realizations of $\mathbf{S}$ conditional of $\mathbf{Y}_M$ under current map estimates, an MCEM or SA algorithm is easily implemented (see **d.2**). Taking two sets of meioses I, male and female, we propose to implement these procedures to estimate gender-specific maps. We will compare effectiveness of MCEM and SA, and will include marker map analysis as a component of analysis in any study of pedigrees involving a large number of meioses. Using the relative likelihoods, we will compare maps estimated from the data with the published gender-specific maps, and assess whether there is a significant difference in likelihood between the maps estimated from study data and the published maps. We will assess whether differences between published and estimated maps lead to differences in linkage findings for the traits of the study.

Since it is undesirable to allow either the study marker data or the published maps to be the sole determinant of the final map estimate, we propose a Bayesian analysis to combine them. Published maps provide a prior distribution for the location of each marker, centered at the consensus location, with a variance determined by the number of informative meioses on which that published location is based. Using this prior, we will implement sampling both over genetic map locations, and also over the space of values of the meiosis array $\mathbf{S}$ given the marker data. We will investigate posterior probabilities of the relative genetic map locations of markers, including posterior probability distributions for the magnitude of differences between prior and posterior maps, and posterior probability intervals for the male-female map ratio in different marker intervals.

Adopting a Bayesian perspective, also allows for more flexible modeling of genetic map heterogeneity, beyond simple gender-specific differences. While it is unrealistic to analyze arbitrary variation by individual and by map interval on the basis of human pedigree data, there is sufficient information in a genome scan for analysis of individual heterogeneity. We propose to model a small number of discrete rate classes, defined by the factor η_i which multiplies the gender-specific map. For example, with 3 classes: low recombination rate $\eta_0 < 1$; standard recombination rates $\eta_1 = 1$; high recombination rate $\eta_2 > 1$. Each parent individual is assumed to fall into one of the three classes. Sampling over numbers of classes, their η-values, assignments of individuals to classes, and over the latent variables $\mathbf{S}$, one obtains the posterior probabilities for the number of classes, the multiplicative map factor for each class, and the posterior probabilities that each parent individual falls into each class. We propose to model individual heterogeneity in recombination rate, to develop tests for the presence of such within-gender heterogeneity, and to examine the effects of heterogeneity on linkage findings.

A question of great interest is the potential heritability of recombination levels, which could have a significant impact on linkage estimates. Whereas random variation in recombination rates among individuals may impact imputation of shared haplotypes, recombination events, and hence Bayesian or likelihood-based linkage signals, the effects will be much larger if an increased recombination rate in a chromosomal region of interest segregates in particular families or line of descent in a pedigree. We propose to extend our models of individual variation in genome-wide recombination rate to test for the heritability of such variation. The discrete classes of distinct η-values may themselves be treated as an ordered categorical latent trait. In principle, with sufficient data, it would be possible to use our methods for joint linkage and segregation analysis to detect and map loci that modify an individual's recombination rate.

(ii) Detection and analysis of pedigree and typing error

Both pedigree errors and typing errors can adversely affect linkage analyses (Boehnke and Cox 1997; Douglas et al. 2000), but neither is necessarily apparent as a Mendelian error, particularly on an extended pedigree where there are missing data (Ott 1993). The elimination of data at loci where there are Mendelian errors reduces information for linkage analyses without resolving a pedigree error which impacts linkage likelihoods based on data at the remaining loci (George et al. 2003). While data in error are best eliminated, it is preferable to eliminate an erroneous pedigree member than data on the entire family only at loci where Mendelian errors are detected. Exact computation of relationship likelihoods on the basis of genome scan data in the presence of potential typing error is limited to pairs or trios of individuals (Epstein et al. 2000; Sieberts et al. 2002) (see also **c2**). For larger numbers of individuals, MCMC again provides an alternative approach.

Conditional on observed marker data $\mathbf{Y}_M$ and on the alternative pedigree relationship structures, we sample the latent true marker genotypes $\mathbf{G}$ and the latent meiosis arrays $\mathbf{S}$ that determine gene *ibd*, $\boldsymbol{\phi}$. The conditional independence structure summarized by

$$\boldsymbol{\phi}(\mathbf{S}) \quad \xrightarrow{\text{marker model}} \quad \mathbf{G} \quad \xrightarrow{\text{error model}} \quad \mathbf{Y}_M$$

makes the implementation straightforward. The meiosis array $\mathbf{S}$ is sampled conditional on $\mathbf{G}$ using our standard samplers (see **d.1**), and $\mathbf{G}$ is then sampled conditional on $\mathbf{S}$ and $\mathbf{Y}_M$.

Ott (1993) proposed a single-locus method for detecting typing or pedigree error on extended pedigrees, where Mendelian errors may not be apparent. This involves computation of the genotype probabilities for each individual, conditional on the marker data observed on all other members of the pedigree. Such an approach, extended to multiple linked loci, is particularly well suited to MCMC analysis, since such genotype probabilities can be directly and efficiently estimated from MCMC output. In a preliminary study, such MCMC sampling detected a very probable previously undetected error in one individual of an Alzheimer's disease pedigree, at the marker locus D1S459. This marker is at a distance of about 5cM from the PS2 mutation on chromosome 1. While these particular data had not been used in the original linkage mapping of PS2 (Levy-Lahad et al. 1995b), we found that the the data at this marker, and specifically for this individual, caused severe computational difficulties in our attempts to obtain accurate estimates of linkage lod scores and gene *ibd* using MCMC methods. Eliminating these data provided for much improved MCMC performance.

We propose to develop this MCMC approach to the detection of pedigree and typing errors on extended pedigrees. We will investigate discrepancies between observed data and predictive probabilities provided by data on relatives. We will investigate the power of multilocus data to identify both pedigree errors and typing errors on extended pedigrees with missing data, using data at multiple linked loci. We will examine the effects of typing and pedigree error on the gain in information for linkage provided by an extended pedigree.

(iii) The meiosis model, genetic interference and the impact on haplotype *ibd*

In reality, there is genetic interference, which at the scale of a 5cM to 15cM genome scan has the effect of decreasing, relative to a Haldane map, the probabilities of multiple recombinations in a chromosomal region. Typing errors and genetic map heterogeneity among meioses have the opposite effect, leading to an apparent clustering of recombination events relative to a Haldane map. Thus in order to have maximal efficiency in detecting errors and estimating map heterogeneity it is important to allow for the effects of interference.

Unlike exact computations, MCMC sampling of the meiosis array is not limited to assuming the absence of genetic interference. Either by a M-H step in the MCMC sampler (Thompson 2000a) or by importance sampling reweighting (see **d.1 (iii)**), realizations of the complete meiosis array $\mathbf{S}$ conditional on data $\mathbf{Y}_M$ permit analysis of the conditional probabilities of recombination jointly over marker intervals within each meiosis. Using the LM-sampler and a no-interference model, we propose to develop tests for genetic interference based on deficiencies of same-meiosis recombination events in close marker intervals. In the event such deficiencies are detectable, we will develop a model of positive genetic interference and use our MCMC estimation of likelihood ratios (**d.2 (iii)**) to compute relative likelihoods of alternative meiosis models.

Often in linkage analyses, after detection of a linkage signal, one focuses on a 20 to 30 cM length, perhaps adding information from additional markers in order to attempt a better trait locus localization. In section **d.1 (iii)** we proposed an effective simulation method for gene descent of inter-marker intervals, conditional on marker data, if it could be assumed that no more than one crossover occurs in any meiosis. In effect, complete interference is assumed for the chromosome region (Fisher 1922). In addition to reality, this model has advantages for MCMC computation, even where there are L marker loci within the region. Each meiosis has only a limited number of permitted recombination patterns: in fact, $(L + 1)$. Thus simultaneous update of many more meioses vectors $S_{i,\bullet}$ is computationally feasible. We have not previously pursued this model, owing to the severe constraints it puts on mixing of the MCMC sampler. Normally the sampler will move from one mode in the space of discrete $\mathbf{S}$ values by adding extra recombinations and then removing the previous ones. If these extra recombinations are precluded by the model, the LM-sampler cannot mix. However, our new mode-hopping algorithm of section **d.1** avoids this problem. Using this MCMC sampler, we propose to investigate genome sharing at sets of relatively tightly linked markers (1 to 3 cM spacing), using a complete interference model for the chromosome segment. We will investigate haplotype *ibd* under this model, and compare probabilities of *ibd* conditional on marker data with those obtained with a no-interference model.

d.5 Evaluation of methods on real data

Several real data sets of complex traits are available to us for the purpose of evaluating and extending our methods (Table P.39). Recruitment and phenotyping for two of the studies (SCZ and FCHL) has been completed, two studies (AD and Guam) have large components for which recruitment has been completed, and two (DYS and AUT) are still under active recruitment, although large components have been collected. With the exception of AUT, genome scan data are available at a 10 cM density. AUT should have complete genome scan data available in 1-2 years, and scans of specific chromosomes are already available. All of these six data sets are available to us through Dr. Wijsman's collaborations. Thus more information will become available to us in the next few years, including additional map locations and/or gene identifications. Among these data sets, the existence of complete genome scans on data sets with a variety of pedigree structures, phenotypes, missing data patterns, and marker density provides ample comparative material for evaluation of methods.

The AD data set is an ongoing study of the genetics of Alzheimer's disease and includes 2 known genes in 18 of the 108 families (Poorkaj et al. 1998; Levy-Lahad et al. 1995a). This data set will be useful for testing approaches that combine information on allelic associations among pedigrees with the segregation information within them (**d.1**), joint multilocus linkage analysis and segregation analysis methods (**d.2**), *ibd* scoring in extended pedigrees (**d.3**), calibration of measures to infer the presence of linkage (**d.3**), and evaluation of map heterogeneity on linkage detection (**d.4**). Marker coverage is complete at the 10 cM density or better in 99 pedigrees (1195 individuals). Sizes of pedigrees are reasonably large, up to 53 individuals per pedigree, with

Data Set	SCZ	AD	FCHL	Guam	DYS	AUT
Pedigrees						
Large	+	+	+	+	+	−
Small	+	+	+	−	+	+
Complex	+	(+)	(+)	+	−	−
Phenotype						
Quantitative	−	−	+	−	+	+
Age-of-Onset	+	+	−	+	−	−
Ordered Categorical	−	+	−	(+)	+	+
Affected/Unaffected	+	+	+	+	+	+
Known Covariate	−	−	+	−	+	−
Markers						
Whole 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 will be available to us once data have been generated, or refers to mapping results which have not yet been confirmed in a second sample.
− : Currently absent in data available to us.

extensive missing data: phenotypic data (age at onset or age at censoring) is available on about 80% of the sample, but due to the late onset of the disease, marker data is available for only about 35% of the sample.

The data from the FCHL (familial combined hyperlipidemia) project is ideal for our investigations of phenotypic models that include continuous and multivariate phenotypes with correlated and missing and/or censored covariate data (**d.2**), and for testing methods of typing and pedigree error detection, maps, and meiosis models (**d.4**). This data set is also useful for comparing Bayesian MCMC and standard lod score linkage signals (**d.3**). A full 10 cM genome scan is available for 15 large pedigrees (up to 85 individuals/pedigree, 567 individuals). This data set includes all original marker typing, including typing errors, and is thus good for testing map and genotyping error-checking approaches. This initial data set is augmented by an additional cleaned set of 542 individuals in 19 families which were also genotyped in a full genome scan.

The Guam and schizophrenia (SCZ) data sets play similar roles. Each consists of a huge, complex, multi-generation pedigree. The Guam data consists of a neurodegenerative syndrome with two different phenotypes from the island of Guam, and includes marker and phenotype data from a genome screen from a single highly complex pedigree consisting of 268 individuals, 103 of whom are affected. Because of the late onset and rapid fatality, only 21 affected individuals are sampled, so the missing data problem is large. SCZ also consists of a large, complex pedigree with many inbreeding loops, consisting of 251 individuals. These pedigrees will be useful for methods dealing with estimation of *ibd* in complex pedigrees (**d.1**), and for investigation of models for dealing with categorical, age-of-onset, and/or reduced penetrance (**d.2**).

Finally, the dyslexia (DYS) and autism (AUT) data sets will provide additional testing grounds for our methods. Both data sets include missing covariate data and phenotypic measures with variable measurement error at different points in the phenotypic spectrum (**d.2**). DYS currently contains > 209 families, > 1532 individuals with phenotype data, and over 35 phenotypic measures, including discrete, quantitative, and categorical phenotypes, many of which are strongly correlated (**d.2**). There are a number of large pedigrees. A genome scan is currently available for 500 individuals in 52 pedigrees, along with denser marker coverage in some regions. AUT consists of > 179 nuclear families with at least 2 affected children, as well as some with additional unaffected children, and is increasing at the rate of 50 pedigrees/year. This data set includes several discrete, ordered categorical, and continuous phenotypic measures plus covariate measures, with missing data.

d.6 Software development

Details of our current software are given in the Progress Report **c.2**. Under the current award period, we have extended the publically released version of the package (MORGAN_2.5) to include a range of MCMC programs for linkage detection and linkage mapping on extended pedigrees, using a genomic array of linked markers. In addition to the new range of main programs, several subroutine libraries for MCMC sampling and for marker data management have been added. These have been thoroughly tested and documented. We propose to continue the development of MORGAN, incorporating our earlier Monte Carlo EM methods (Guo and Thompson 1994), but using the new MCMC sampler routines and other more recent improvements to the methodology and software. As older code is revised for MORGAN_2, we will enhance its capabilities, through dynamic memory allocation, for example. All programs released under MORGAN_2 will be thoroughly tested, well documented, and released with test data and corresponding output. For developers, code will be documented using Doxygen (see **c.2**).

As new methods are developed, we will incorporate these into MORGAN. These include the development of improved MCMC sampling methods (see section **d.1**), the new trait model framework and methodology (**d.2**), the development of of *ibd* sharing statistics scored jointly over loci (**d.3**), and programs for the more systematic investigation of genetic maps and genotype and pedigree errors (**d.4**). We will also continue the development of use of tools such dynamic visualization (see **d.1**) to assess performance of our MCMC samplers. Our use of CVS (see **c.2**) will enable us to make fundamental changes in the way traits are specified without disrupting existing programs or their development. We have tested this approach in our successful generalization of individual and locus identifiers, which also required changes to many low-level library routines. As quantitative trait capabilities are integrated into MORGAN programs (**d.2**), this will facilitate comparison with Bayesian analyses under **Loki** (section **d.3**). Indeed, since our pseudo-Bayes likelihood program *lm_bayes* is released under MORGAN_2.5, there is already no hard divide between our likelihood and Bayesian software.

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. Additionally, as MORGAN has been used in short courses and by students, we have developed some tutorial materials. We propose to develop these into a web-based tutorial which will be made publically available with future versions of MORGAN, and to maintain and update this with subsequent releases.

Our current structure of programs and subroutine libraries is developed under the Unix system, and is best suited to a Unix/linux environment. Most of the methods are computationally intensive, and require a fast workstation. We are committed to maintaining portability across a range of Unix/linux platforms, including Tru 64 (Compaq/Alpha), Solaris, and Red Hat linux. Also, we have a Windows version of the software released in Fall 2002 (see **c.2**). As yet we have little experience with the operation of MORGAN programs under Windows, but we propose to continue to make a Windows version available, and we will be willing to work with users who wish to enhance that or other conversions.

d.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 implementing complex trait models (**d.2**) and robust methods for linkage detection (**d.3**) is the top priority for years 1-2, so that substantive comparisons can proceed in years 3-5. Improvements to MCMC sampling and estimation (**d.1**) will continue throughout. Investigations of genetic maps and data errors (**d.4**) will also continue, but will have greater emphasis in years 3-4.

Generally, algorithmic development is the responsibility of Thompson, while Wijsman focuses on comparison and evaluation of methods. However, all members of the research team implement their own algorithmic developments and will test their methods by analyses of real and simulated data, as appropriate (**d.5**). Software for public release (**d.6**) is the responsibility of programmer Myrna Jewett, supervised by Thompson. Each student research assistant has a well-defined project, supervised by Thompson and/or Wijsman. In addition to regular individual meetings with Thompson/Wijsman, all members of the research team meet together weekly to discuss and review goals and progress.

e: Human Subjects Research

The proposed research involves the assessment and comparison of statistical and computational methods, in part through their use on real data on human pedigrees.

1. Risks to the subjects

Human subjects involvement and characteristics

Each real data set derives from an existing study that has Human Subjects approval at the University of Washington, either because the primary study is at the University of Washington, or because the study involves an established research collaboration with the University of Washington. Each study is one in which co-investigator Dr. Ellen Wijsman is involved as consultant or collaborator. The gender, ethnicity, ages, and other characteristics of the subjects involved in each study are such that the study has Human Subjects approval.

Sources of materials

Only computational and statistical analysis of existing data is proposed. There will be no study subjects recruited, nor data collected from subjects, nor biological materials analyzed in connection with this proposal.

Potential risks

The potential risk is to the confidentiality of genetic and disease data on the individual study subjects. Although individual identifiers are anonymized in the computerized data bases for each study, a key to these anonymized identifiers exists for each study. Some studies involve individuals who are members of extended pedigrees with particular pedigree structures. Due to the nature of the methods, these pedigree relationships are an integral part of any analysis. This specificity of pedigree structure increases the potential risk that individuals could be identified.

2. Adequacy of protection against risks

Recruitment and informed consent

There is no recruitment of subjects in connection with this proposal. Each existing study has Human Subjects approval at the University of Washington for its recruitment and informed consent protocols. Each study protocol includes a provision for trying new methodological developments in the computational and statistical analysis of the data.

Protection against risk

The data files which will be used for the proposed research will be fully anonymized, and we will not have access to any key to subject identities. To decrease risk, genders of individuals may be randomly reassigned and data from some individuals excluded in forming data files for the proposed research, provided that this does not compromise the validity of the data set for methodological testing. For the purposes of the proposed research, we will not have access to the full study data on individuals, but only to a subset of the disease traits, genetic marker, and covariate information sufficient to provide suitable data sets for methodological testing.

3. Potential benefit of proposed research to subjects and others

Diseases such as hypertension and psychiatric illness have both environmental and genetic components. However, identification of genes contributing to increased risk of such diseases has been difficult, and computational limitations have constrained the use of the available data. The development of new methods well suited to the analysis of data on extended pedigrees can make maximal use of genetic data, and lead to better methods for resolving the genetic basis of complex traits. Resolution of the genetic and environmental components of a disease trait is an important step in developing methods for the prevention and treatment.

4. Importance of the knowledge to be gained

While the study of simulated data plays a useful role in the development and testing of new methods, simulated data can never show the complexities and heterogeneity of reality. Testing on real data is of fundamental importance in the development of methods that will serve a useful purpose in the analysis of real data and lead to the resolution of genetic and environmental components of human disease traits.

f: Vertebrate Animals

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

g: Literature Cited

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics 30:97–101

Almasy L, Borecki IB (1999) Exploring genetic analysis of complex traits through the paradigm of alcohol dependence: Summary of GAW11 contributions. Genetic Epidemiology 17 (Suppl 1):S1–S24

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

Andersen SK, Olesen KG, Jensen FV, Jensen F (1989) HUGIN - A shell for building Bayesian belief universes for expert systems. In NS Sridharan, ed., *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, 1080–1085. Morgan Kaufmann, San Mateo, CA

Andersson L (2001) Genetic dissection of phenotypic diversity in farm animals. Nature Reviews Genetics 2:130–138

Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3:299–309

Baron M (2001) Genetics of schizophrenia and the new millennium: Progress and pitfalls. American Journal of Human Genetics 68:299–312

Basu S, Wijsman EM, Thompson EA (2002) Allele-sharing methods on large pedigrees. Genetic Epidemiology 23:267

Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions on Markov chains. Annals of Mathematical Statistics 41:164–171

Besag J, Green P (1993) Spatial statistics and Bayesian computation (with discussion). Journal of the Royal Statistical Society B 55:25–37

Bink MCAM, Uimari P, Sillanpaa MJ, Janss LLG, Jansen RC (2002) Multiple QTL mapping in related plant populations via a pedigree-ana lysis approach. Theoretical and Applied Genetics 104:751–762

Bird TD, Schellenberg GD, Wijsman EM, Martin GM (1989) Evidence for etiologic heterogeneity in Alzheimer's disease. Neurobiology of Aging 10:432–4; discussion 446–8

Boehnke M (1994) Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Hum Genet 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

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

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

Chapman NH, Leutenegger A, Badzioch MD, Bogdan M, Conlon EM, Daw EW, Gagnon F, Li N, Maia JM, Wijsman EM, Thompson EA (2001) The importance of connections: Joining components of the Hutterite pedigree. Genetic Epidemiology 21 (Suppl 1):S230–S235

Chapman NH, Wijsman EM (1998) Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. American Journal of Human Genetics 63:1872–1885

Cherny SS, Abecasis GR, Cookson WOC, Sham PC, Cardon LR (2001) The effect of genotype and pedigree error on linkage analysis: Analysis of three asthma genome scans. Genetic Epidemiology 21, (Suppl 1):S117–S122

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Gene dose of Apolipoprotein-E type-4 allele and the risk of Alzheimer's Disease in late onset families. Science 261:921–923

Daw E, Heath SC, Wijsman EM (1999a) Multipoint oligogenic analysis of age–of–onset data with applications to large Alzheimer's disease pedigrees. American Journal of Human Genetics 64:839–851

Daw EW, Kumm J, Snow GL, Thompson EA, Wijsman EM (1999b) Monte Carlo Markov chain methods for genome screening. Genetic Epidemiology 17 (Suppl 1):S133–S138

Daw EW, Payami H, Nemens EJ, Nochlin D, Bird TD, Schellenberg GD, Wijsman EM (2000a) The number of trait loci in late-onset Alzheimer disease. American Journal of Human Genetics 66:196–204

Daw EW, Thompson EA, Wijsman EM (2000b) Bias in multipoint linkage analysis arising from map misspecification. Genetic Epidemiology 19:366–380

— (2003) A score for Bayesian genome screening. Genetic Epidemiology 24:in press

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden BW, Smith DR, Lander ES, Botstein D, Akots G, Rediker KS, Gravius T, Brown VA, Rising MB, Parker C, Powers JA, Watt DE, Kauffman ER, Bricker A, Phipps P, Muller-Kahle H, Fulton TR, Ng S, Schumm JW, Braman JC, Knowlton RG, Barker DF, Crooks SM, Lincoln SE, Daly MJ, Abrahamson J (1987) A genetic linkage map of the human genome. Cell 51:319–337

Douglas J, Boehnke M, Lange K (2000) A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. American Journal of Human Genetics 66:1287–1297

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

Emahazion T, Feuk L, Jobs M, Sawyer SL, Fredman D, St Clair D, Prince J, Brookes A (2001) SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. Trends in Genetics 17:407–413

Epstein M, Duren W, Boehnke M (2000) Improved inference of relationship for pairs of individuals. American Journal of Human Genetics 67:1219–1231

Falk CT (2001) Introduction: Haplotype analysis of simulated genetic analysis workshop 12 data. Genetic Epidemiology 21 (Suppl 1):S552–S553

Field LL (2002) Genetic linkage and association studies of Type I diabetes: challenges and rewards. Diabetologia 45:21–35

Fisher RA (1922) The systematic location of genes by means of crossover observations. American Naturalist 56:406–411

— (1934) The amount of information supplied by records of families as a function of the linkage in the population sampled. Annals of Eugenics 6:66–70

Gabriel SB, Salomon R, Pelet A, Angrist M, Amiel J, Fornage M, Attie-Bitach T, Olson JM, Hofstra R, Buys C, Steffann J, Munnich A, Lyonnet S, Chakravarti A (2002) Segregation at three loci explains familial and population risk in Hirschsprung disease. Nature Genetics 31:89–93

Gagnon F, Jarvik GP, Motulsky AG, Deeb SS, Brunzell JD, Wijsman EM (2002) Disentangling compound phenotypes and pleiotropic effects in multifactorial diseases using covariate adjustments: FCHL as an example. Genetic Epidemiology 23:281–282

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

Gelman A, Rubin DB (1992) Inference from Iterative Simulation Using Multiple Sequences. Statistical Science 7:457–472

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

George AW, Basu S, Li N, Rothstein JH, Sieberts S, Stewart W, Wijsman EM, Thompson EA (2003) Approaches to mapping genetically correlated traits. Genetic analysis workshop 13, Submitted

George AW, Bogdan M, Wijsman EM, Thompson EA (2001) Markov chain Monte Carlo methods for the calculation of likelihoods in genetic linkage studies. American Journal of Human Genetics 69 (Suppl):411

George AW, Thompson EA (2003) Multipoint linkage analyses for disease mapping in extended pedigrees: A Markov chain Monte Carlo approach. Statistical Science 18:in press

George AW, Wijsman EM, Thompson EA (2002) Detecting disease genes via a new Markov chain Monte Carlo approach for multipoint linkage analysis. Genetic Epidemiology 23:283

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

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

Glazier AM, Nadeau JH, Aitman T (2002) Finding genes that underlie complex traits. Science 298:2345–2349

Goedken R, Ludington E, Crowe RR, Fyer AJ, Hodge SE, Knowles JA, Vieland VJ, Weissman MM (2000) Drawbacks of GENEHUNTER for larger pedigrees: application to panic disorder. American Journal of Medical Genetics Neuropsychiatric Genetics 96:781–783

Goldgar DE, Easton DF (1997) Optimal strategies for mapping complex diseases in the presence of multiple loci. American Journal of Human Genetics 60:1222–1232

Goldin LR (2001) Introduction: Linkage analysis of quantitative traits. Genetic Epidemiology 21 (Suppl 1):S459–S460

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

Gretarsdottir S, Sveinbjornsdottir S, Jonsson HH, Jakobsson F, Einarsdottir E, Agnarsson U, Shkolny D, Einarsson G, Gudjonsdottir HM, Valdimarsson EM, Einarsson OB, Thorgeirsson G, Hadzic R, Jonsdottir S, Reynisdottir ST, Bjarnadottir SM, Gudmundsdottir T, Gudlaugsdottir GJ, Gill R, Lindpaintner K, Sainz J, Hannesson HH, Sigurdsson GT, Frigge ML, Kong A, Gudnason V, Stefansson K, Gulcher JR (2002) Localization of a susceptibility gene for common forms of stroke to 5q12. American Journal of Human Genetics 70:593–603

Gudbjartsson D, Jonasson K, Frigge M, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nature Genetics 25:12–13

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

— (1994) Monte Carlo estimation of mixed models for large complex pedigrees. Biometrics 50:417–432

Haldane JBS (1934) Methods for the detection of autosomal linkage in man. Annals of Eugenics 6:26–65

Halpern J, Whittemore AS (1999) Multipoint linkage analysis. A cautionary note. Human Heredity 49:194–196

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

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

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genetics in Medicine 4:45–61

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603

Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nature Genetics 29:306–309

Irwin M, Cox N, Kong A (1994) Sequential imputation for multilocus linkage analysis. Proceedings of the National Academy of Sciences USA 91:11684–11688

Jansen RC (2001) Quantitative trait loci in inbred lines. In DJ Balding, M Bishop, C Cannings, eds., *Handbook of Statistical Genetics*, 567–597. Wiley, London

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

Jensen CS, Kong A (1999) Block Gibbs sampling for linkage analysis in large pedigrees with many loops. American Journal of Human Genetics 65:885–901

King TR, Dove WF, Guénet J, Hermann BG, Shedlovsky A (1991) Meiotic mapping of murine chromosome 17: The string of loci around *l(17)-2-Pas.* Mammalian Genome 1:37–46

Knudson AG (2002) Cancer genetics. American Journal of Human Genetics 111:96–102

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nature Genetics 31:241–247

Kong A, Liu J, Wong WH (1994) Sequential imputations and Bayesian missing data problems. Journal of the American Statistical Association 89:278–288

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

— (1998) Faster multipoint linkage analysis using Fourier transforms. Journal of Computational Biology 5:1–7

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

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

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

Lin S (1996) Multipoint linkage analysis via Metropolis jumping kernels. Biometrics 52:1417–1427

— (2000) Monte Carlo methods for linkage analysis of two-locus disease models. Annals of Human Genetics 64:519–532

Lombardi MP, Redeker EJW, Defesche JC, Kamerling SWA, Trip MD, Mannens MMAM, Havekes LM, Kastelein JJP (2000) Molecular genetic testing for familial hypercholesterolemia: spectrum of LDL receptor gene mutations in the Netherlands. Clinical Genetics 57:116–124

Longmate JA (2001) Complexity and power in case-control association studies. American Journal of Human Genetics 68:1229–1237

Mackay TFC (2001) Quantitative trait loci in Drosophila. Nature Reviews Genetics 2:11–20

Markianos K, Daly M, Kruglyak L (2001) Efficient multipoint linkage analysis through reduction of the inheritance space. American Journal of Human Genetics 68:963–977

Marsh DJ, Zori RT (2002) Genetic insights into familial cancers - update and recent discoveries. Cancer Letters 181:125–164

Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Chui B, Cohen P, de Toma C, Ehm M, Glanowski S, He C, Heil J, McMullen I, Stein L, Wagner M, Winick J, Winn-Deen ES, Cann HM, Lai E, Holden AL (2002) A high-resolution human SNP linkage map. American Journal of Human Genetics 71 (Suppl):206

Mauricio R (2001) Genetic dissection of phenotypic diversity in farm animals. Nature Reviews Genetics 2:370–381

McPeek MS (1999) Optimal allele-sharing statistics for genetic mapping using affected relatives. Genetic Epidemiology 16:225–249

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

Morris RW, Kaplan NL (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genetic Epidemiology 23:221–233

Moyeed RA, Baddeley AJ (1991) Stochastic approximation for the MLE for a spatial point pattern. Scandinavian Journal of Statistics 18:39–50

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

Neuman RJ, Rice JP (1992) Two-locus models of disease. Genetic Epidemiology 9:347–365

O'Brien SJ, Nelson GW, Winkler CA, Smith MW (2000) Polygenic and multifactorial disease gene association in man: Lessons from AIDS. Annual Review of Genetics 34:563–591

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411:603–606

Ott J (1974) Estimation of the recombination frequency in human pedigrees: Efficient computation of the likelihood for human linkage studies. American Journal of Human Genetics 26:588–597

— (1993) Detecting marker inconsistencies in human gene-mapping. Human Heredity 43:25–30

— (1999) Analysis of Human Genetic Linkage. 3 rd. ed., The Johns Hopkins University Press, Baltimore, MD

Penrose LS (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Annals of Eugenics 6:133–138

Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, Pike B, Root H, Rubenstein J, Boyer R, Stenroos ES, Chandrasekharappa S, Athanassiadou A, Papapetropoulos T, Johnson WG, Lazzarini AM, Duvoisin RC, Di Iorio G, Golbe LI, Nussbaum RL (1997) Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science 276:2045–2047

Poorkaj P, Bird TD, Wijsman E, Nemens E, Garruto RM, Anderson L, Adreadis A, Wiederholt WC, Raskind M, Schellenberg GD (1998) Tau is a candidate gene for chromosome 17 frontotemporal dementia. Annals of Neurology 43:815–825

Province MA, Rice TK, Borecki IB, Gu C, Kraja A, Rao DC (2003) Multivariate and multilocus variance components method, based on structural relationships to assess quantitative trait linkage via SEGPATH. Genetic Epidemiology 24:128–138.

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

Risch N (1990) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. American Journal of Human Genetics 46:242–253

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

Rogaeva E (2002) The solved and unsolved mysteries of the genetics of early-onset Alzheimer's disease. Neuromolecular Medicine 2:1–10

Rubin D (1987) Multiple imputation for nonresponse in surveys. Wiley &Sons, New York

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning ZM, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann NS, Zody MC, Linton L, Lander ES, Altshuler D (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Sham P (2001) Shifting paradigms in gene-mapping methodology for complex traits. 2001 2:195–202

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

Sieberts SK, Wijsman EM, Thompson EA (2002) Relationship inference from trios of individuals in the presence of typing error. American Journal of Human Genetics 70:170–180

Skipper L, Farrer M (2002) Parkinson's genetics: Molecular insights for the new millennium. Neurotoxicology 23:503–514

Skrivanek Z, Irwin M, Lin S, Wright F (1999) SIMPLE: A linkage program that incorporates interference. American Journal of Human Genetics 65 (Suppl):A446

Skrivanek Z, Lin S, Irwin ME (2002) Linkage analysis with sequential imputation. Technical report no. 689, Department of Statistics, The Ohio State University

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

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

St. George-Hyslop PH, Tanzi RE, Polinsky RJ, Haines JL, Nee L, Watkins PC, Myers RH, Feldman RG, Pollen D, Drachman D, Growden J, Bruni A, Foncin JF, Salmon D, Frommelt P, Amaducci L, Sorbi S, Piacentini S, Stewart GD, Hobbs WJ, Conneally PM, Gusella JF (1987) The genetic defect causing familial Alzheimer's disease maps on chromosome 21. Science 235:885–890

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

Tanzi R, Bertram L (2001) New frontiers in Alzheimer's disease genetics. Neuron 32:181–184

Terwilliger JD, Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? Current Opinion in Biotechnology 9:578–594

Thomas A, Gutin A, Abkevich V (2000) Multilocus linkage analysis by blocked Gibbs sampling. Statistics and Computing 10:259–269

Thompson EA (1988) Two-locus and three-locus gene identity by descent in pedigrees. IMA Journal of Mathematics Applied in Medicine and Biology 5:261–280

— (2000a) MCMC estimation of multi-locus genome sharing and multipoint gene location scores. International Statistical Review 68:53–73

— (2000b) Statistical Inferences from Genetic Data on Pedigrees, vol. 6 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Beachwood, OH

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

Thompson EA, Heath SC (1999) Estimation of conditional multilocus gene identity among relatives. In F Seillier-Moiseiwitsch, ed., *Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology*, IMS Lecture Note–Monograph Series Volume 33, 95–113. Institute of Mathematical Statistics, Hayward, CA

Thompson EA, Shaw RG (1992) Estimating polygenic models for multivariate data on large pedigrees. Genetics 131:971–978

Tierney L (1994) Markov chains for exploring posterior distributions. Annals of Statistics 22:1701–1728

Trembath RC, Clough RL, Rosbotham JL, Jones AB, Camp RDR, Frodsham A, Browne J, Barber R, Terwilliger J, Lathrop GM, Barker JNWN (1997) Identification of a major susceptibility locus on chromosome 6p and evidence for further disease loci revealed by a two stage genome-wide search in psoriasis. Human Molecular Genetics 6:813–820

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

van Heel DA, McGovern DPB, Cardon LR, Dechairo BM, Lench NJ, Carey AH, Jewell DP (2002) Fine mapping of the IBD1 locus did not identify Crohn disease-associated NOD2 variants: Implications for complex disease genetics. American Journal of Human Genetics 111:253–259

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

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

Williams JT, Van Eerdewegh P, Almasy L, Blangero J (1999) Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. American Journal of Human Genetics 65:1134–1147

Younes L (1988) Estimation and annealing for Gibbsian fields. Annals of the Institute Henri Poincare for Probability and Statistics 24:269–294

Yuan B, Neuman R, Duan SH, Weber JL, Kwok PY, Saccone NL, Wu JS, Liu KY, Schonfeld G (2000) Linkage of a gene for familial hypobetalipoproteinemia to chromosome 3p21.1-22. American Journal of Human Genetics 66:1699–1704

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

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468

h: Consortium Arrangements

NONE

i: Consultants

NONE