A: Specific Aims

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. We propose to continue to develop methods of analysis of phenotypic and marker data observed on members of a pedigree. 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 quantitative phenotypes and permit full use of multilocus marker data, including dense SNP markers. We propose methods that will permit analysis of traits that are influenced by environmental factors, and where the data include covariate information and censoring. Traits may be affected by genes at several loci, and a trait locus may affect several genetically correlated quantitative phenotypes. Our methods 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 multilocus marker data by
(i) development of Monte Carlo methods providing independent realizations of latent gene descent patterns jointly across multiple genome locations,
(ii) continuing development of block-Gibbs and Metropolis-Hastings MCMC samplers of gene descent patterns, extending recently developed multiple-meiosis samplers to include joint sampling at tightly linked marker loci,
(iii) developing the use of importance-sampling reweighting methods to permit a single set of realizations to be used to explore a variety of genetic marker maps, and
(iv) developing diagnostics of block-Gibbs and Metropolis-Hastings MCMC samplers to predict required run-times and ensure accurate sampling of latent gene descent.

2. To develop further effective analysis methods to use marker-based realizations of gene descent patterns for joint multilocus linkage and segregation analysis of complex traits, by
(i) extending the trait model framework that can be incorporated into MCMC-based linkage analyses, allowing for epistasis and pleiotropy in the analysis of discrete or quantitative phenotypes,
(ii) developing Monte-Carlo likelihood-ratio estimation methods to explore the sensitivity of lod scores both to trait model parameters and to genetic marker map and model uncertainties, and
(iii) using realized marker-based gene descent patterns in trait-model selection procedures, providing a new graphical-model based approach to joint segregation and linkage analysis.

3. To develop methods for the assessment of statistical significance of linkage findings using the empirical MCMC-generated distribution of gene descent patterns conditional on observed marker data, and correcting for multiple dependent tests, by
(i) developing computationally efficient trait-resimulation and trait-permutation approaches to obtain exact p-values for linkage detection tests or lod scores using data on extended pedigrees,
(ii) developing the approach of analysis of the distribution of latent p-values conditional on observed marker data to obtain an exact measure of the statistical significance of a maximum lod score, for either a given complex trait model or for a lod score integrated over trait model parameters, and
(iii) using importance-sampling reweighting both to obtain distributions of gene descent patterns under linkage alternatives, and hence confidence sets for trait locus locations, and also to explore sensitivity of p-values to trait and marker model assumptions.

4. To develop methods using gene descent patterns sampled conditional on multilocus genetic marker data on extended pedigrees, to analyze genetic maps, marker models, and typing error, by
(i) extending the marker models that can be used in MCMC analyses to allow for not only for typing errors and null alleles at a marker, but also deletions and copy-number variants that may extend over several markers,
(ii) using Monte Carlo likelihood-ratio estimation and importance-sampling reweighting to investigate the effects of genetic marker map uncertainty and heterogeneity on imputed recombination events, *ibd* probabilities and

linkage lod scores, and
(iii) developing exact and Monte-Carlo methods for analysis of haplotype *ibd* over multiple tightly linked markers, incorporating linkage disequilibrium along founder haplotypes into the analysis of family data.

5. In addition to appropriate simulation studies, to evaluate methods developed in aims 1-4 on real data from studies studies of cardiovascular, neurological and behavioral traits. We will explore the use of MCMC-based realizations of gene descent patterns conditional on marker data
(i) to assess and compare methods of joint segregation and linkage analysis of multiple genetically correlated quantitative phenotypes, with covariate information,
(ii) to compare methods of assessing significance of linkage findings, using both data resimulation approaches, and methods that directly use the distribution of latent inheritance patterns conditional on observed data, and
(iii) to evaluate the utility of genetic marker maps at different densities, and at different levels of typing accuracy and missing data, including incorporation of information on haplotypes and/or linkage disequilibrium.

6. To extend and enhance our user software, by
(i) further developing the MORGAN package to contain the full array of our current methods,
(ii) developing our new MORGAN_3 framework which provides for multiple loci affecting a trait, and for trait loci that may affect multiple traits,
(iii) implementing new methodology as proposed in aims 1–4, and
(iv) documenting, distributing and supporting the software, developing web-based tutorial materials and examples, and seeking user response to aid in ongoing software improvement.

B: Background and significance

B.1: Genetic approaches for complex traits

B.1.1: The classical genetic approach. The risk for developing many common diseases and associated traits is influenced by both genetic and environmental factors. Examples include cardiovascular and neurological diseases, mental illness, cancers, and developmental disorders. Identification of genes that predispose to such traits (disease genes) provides a key step towards understanding disease etiology, which, in turn, can be important for prevention and development of treatments. 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 involved in Alzheimer's disease (AD) (Tanzi and Bertram 2005), more than 40 genes involved in cancers (Vogelstein and Kinzler 2004), and 6 genes involved in Parkinson's disease (PD) (Skipper and Farrer 2002; Lorincz 2006). Other studies have also recently identified the CFH gene on chromosome (chr) 1q involved in age-related macular degeneration (AMD) (Klein et al. 2005; Haines et al. 2005; Edwards et al. 2005), and two additional genes implicated in AMD in a small region on chr 10q (Rivera et al. 2005; Dewan et al. 2006; Yang et al. 2006) with resolution regarding which gene is causative not yet available. Identification of such genes has often catapulted a field forward and is thus of value, even when the diseases behave in a Mendelian fashion. For example, identification of the first Parkinson's disease 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; Walsh and Selkoe 2004). Similarly, identification of LDLR was critical for the identification of pathways important in cholesterol transport (Brown and Goldstein 1986). Identification of multiple genes which are involved in a particular disease increasingly often identifies key interacting players in relevant biochemical pathways, as has happened in AD, PD, and cancer (Skipper and Farrer 2002; Walsh and Selkoe 2004; Vogelstein and Kinzler 2004). This, in turn, helps to focus subsequent studies that contribute to, for example, the 48% decrease in death rate from coronary heart disease that occurred between 1983 and 2003 (http://www.nhlbi.nih.gov/about/factbook/chapter4.htm#4_1). The successes of this approach have fueled recent studies in which very large numbers of phenotypes are measured in large data sets, in the hopes of identifying genes contributing to complex traits such as hypertension (Boerwinkle et al. 2002).

Identification of disease genes typically involves several stages of gene localization followed by in-depth investigation of a small number of genes. As of the beginning of 2007, this approach, which includes positional cloning, is responsible for identification of most of >2420 identified genes out of >6134 mapped disease and quantitative trait loci (http://www.ncbi.nlm.nih.gov/Omim/mimstats.html)—an increase of more than 20-fold

in number of genes identified by positional cloning over the past decade. Most such identified genes were first localized to a small genomic region by genome scans using pedigree-based linkage analysis methods, described more fully in section **B.2**. Once a gene is mapped, the relatively small number of meioses in pedigree data makes it difficult to achieve the fine-scale resolution needed for gene identification (Boehnke 1994). At this stage, linkage disequilibrium, or association, between the trait and nearby markers is often used to limit the size of the region to be searched for the responsible gene. This proposal is primarily concerned with the steps that precede fine scale mapping or use of population-based association studies, but also proposes the combination of linkage and association (section **D.4.3**) to address this latter stage.

Despite increasing success with analysis of complex diseases, there remain difficulties. The success rate in mapping such disorders with pedigree-based designs remains lower than for Mendelian traits. While for monogenic diseases failure to replicate evidence for linkage is rare (Rao et al. 1978), such failure to obtain replication is more common for complex traits. For some traits, such as schizophrenia, there have been many attempts to map disease genes, with numerous positive reports of linkage but at best modest evidence of cross-study replication (Owen et al. 2004). For other diseases initial false positive reports of linkage 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). Finally, even for diseases with positive evidence of linkage, it is common that evidence for linkage in individual studies is relatively weak, even when there is consistency across studies for linkage to locations with subsequent gene identification, such as was observed for AMD (Klein et al. 1998; Majewski et al. 2003; Seddon et al. 2003). With weak linkage signals, even small amounts of undetected genotyping error can have large effects. For example, in an analysis of body-mass-index in nuclear families, lod scores more than doubled on two chromosomes, from 1 to 2.8 and from 1.5 to 3, after careful and extensive genotype data cleaning was completed (Chang et al. 2006).

For several reasons, limitations in the available techniques of analysis contribute to the difficulty, compared to simple monogenic traits, in identifying complex trait genes. (1) 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, 2-trait locus multipoint linkage analysis on pedigrees of modest size (< 24 individuals) remains computationally impractical with exact methods even on current fast workstations (Dietter et al. 2004). (2) Use of multipoint analyses for linkage detection is more common for complex than for simple traits, but map error and undetected genotyping errors may be problems for analyses based on many, closely linked, markers (Daw et al. 2000; Fingerlin et al. 2006). This issue has become particularly important with the increasing emphasis on use of dense, diallelic, markers rather than more widely-space multiallelic markers, which forces use of multipoint methods to maintain linkage information (Wilcox et al. 2005). (3) 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 context of variance-component linkage analysis (Almasy and Blangero 1998; Province et al. 2003), with additional recent exploration of methods that incorporate censored or skewed distributions (Diao and Lin 2006; Li et al. 2006; Epstein et al. 2003). Only a limited number of models that explicitly relate trait phenotype to genotype have been developed and tested in a model-based framework, and the effect of undetected genotype errors is hard to assess because of lack of programs that can incorporate a genotyping error model. (4) Available methods that do not require specification of a trait-model are largely limited to restricted, small pedigree structures, which are inefficient for both linkage detection and gene localization (Wijsman and Amos 1997). (5) Finally, for studies involving large numbers of traits, there are computational and practical challenges in identifying the individual or combined traits for which genome scans are most likely to yield useful results, and in carrying out the large number of computations associated with the genome scans and their followup.

B.1.2: 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.1.1**), 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 is beginning to be used for genome wide scans (Klein et al. 2005; Smyth et al. 2006). This approach includes as success stories a few notable examples, including APOE and Alzheimer's disease (Corder et al. 1993), HLA and NIDDM1 (Field 2002), several loci involve in susceptibility to AIDS (O'Brien et al. 2000), and complement factor H and AMD (Edwards et al. 2005; Haines et al. 2005; Klein et al. 2005). 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 (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 broad claims of increased power to map disease.

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). While some traits may be explained by simple allelic variation in a small number of loci, allelic heterogeneity is a major problem for population-based approaches. Extensive empirical data show that there can be very large numbers of mutations at identified loci, thus justifying concerns about allelic heterogeneity. For example, by January 2007, for early-onset Alzheimer's disease > 190 mutations in 3 genes were known (http://www.molgen.ua.ac.be/ADMutations), and for familial hypercholesterolemia, > 770 mutations in one gene (http://www.ucl.ac.uk/fh/muttab.html) were known. 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 evidence for linkage produced with association analyses, such as NIDDM1 and HLA, Alzheimer's disease and APOE, and two widely-touted associations recently identified in genome wide association scans with AMD (Klein et al. 2005; Haines et al. 2005; Edwards et al. 2005; Rivera et al. 2005; Dewan et al. 2006; Yang et al. 2006) have so far also been identifiable with family-based designs (Field 2002; Blacker et al. 2003; Klein et al. 1998; Fisher et al. 2005). (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 large pedigrees, are more likely to be able to identify correctly the location of segregating trait loci (Goldin 2001; Falk 2001). This point is important because the relative costs of ascertainment and phenotyping are typically considerably more than the costs of genotyping samples. (3) Tests of population-based designs on genes that have already been identified have sometimes failed to detect evidence of genes that can be identified through family studies, including Crohn's disease (van Heel et al. 2002), Alzheimer's disease (AD) (Emahazion et al. 2001; Rogaeva 2002), and HDL levels (Cohen et al. 2004). (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 recent attempt at replication, which was based on one of the largest case-control studies published to date with a sample of 11,208 cases and controls, failed to reproduce associations for 13 SNPs and Parkinson's disease (Elbaz et al. 2006).

There are several reasons that may explain discrepancies among theoretical studies, and between 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). However, 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 and gene identification studies on crosses in agricultural and model organisms attest (Mackay 2001; Fridman et al. 2004; Ueda et al. 2005), pedigree-based linkage mapping for trait loci with common alleles remains a time-tested and powerful tool to localize such loci.

B.2: Linkage testing and estimation 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. We include current variance-component (VC) methods (Haseman and Elston 1972; Amos 1994; Almasy and Blangero 1998), which are particularly suitable for continuous traits, in this latter category.

For the past three decades, linkage analysis has frequently been performed with an explicit trait locus model, often capitalizing on the power of relatively large pedigrees. This approach became possible when improved computational methods became available (Elston and Stewart 1971; Ott 1974; Lathrop et al. 1984), and allowed practical implementation of the earliest pedigree-based genetic inference methods (Fisher 1934; Haldane 1934; Haldane and Smith 1947). This framework was developed in the context of Mendelian traits, but has also been used successfully for more complex traits. This approach provides both linkage detection and, with a good trait model, accurate localization. Localization is improved when relatively dense markers in a region of interest can be used to identify the closest flanking recombinations (Boehnke 1994) and/or identify multilocus haplotypes segregating with the disease allele (Levy-Lahad et al. 1995b). This approach also easily handles pedigrees of diverse sizes, which is useful because, in many situations, extended pedigrees provide greater power for the same total sample size than do small pedigrees for mapping trait loci (Wijsman and Amos 1997; Chapman et al. 2001). With the methods commonly used, a trait model may be estimated in a joint segregation and linkage analysis (see **D.2**), but in linkage assessment this model is normally held fixed. The statistic used is typically the the map-specific lod score (Ott 1999). 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 γ (Heath 1997; Logue and Vieland 2004). This approach to genetic analysis can be used for both linkage detection, and for location estimation. In principle either a single-locus or multilocus trait model may be used (Schork et al. 1993). The limiting factors for use of more complex trait models consist of computational constraints, amount of available data, and information about the trait model.

A robust approach to linkage detection avoids specification of any trait model through use of one of many *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), and are popular for analysis of complex traits. In recent years many variants on these methods have been developed in terms of *ibd* probabilities computed or imputed conditionally on marker data. The methods include likelihood-based VC methods that are useful for analysis of continuous traits on both nuclear and extended pedigrees (Amos 1994; Almasy and Blangero 1998). However, while *ibd*-based methods can be used for linkage detection, they do not provide good trait localization even in the presence of dense marker genotyping (Atwood and Heard-Costa 2003). In addition, affected relative pair methods used for discrete traits generally lack power. Loss of power 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 extended pedigree structures and in

contrasting affected with unaffected individuals. Because more trait information is used, power of VC methods to detect linkage is consequently better than for discrete traits. 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) as well as allowing development of additional *ibd* measures (McPeek 1999). However, there remain both computational and statistical difficulties with the use of *ibd*-based linkage detection tests on extended pedigrees. In particular, some *ibd* measures may be computationally intractable on extended pedigrees (Basu et al. 2002) or have highly skewed distributions that affect testing, either because of the characteristics of the measure (Kruglyak et al. 1996; McPeek 1999) or because of the ascertainment procedures (Forrest and Feingold 2000).

B.3: Assessment of linkage evidence

It is not simple to assess the statistical evidence for linkage that corresponds to a specific value of a linkage statistic obtained for a specific data set. Use of an assumed significance level that is overly conservative relative to the correct value results in loss of power and consequently need for excessive amounts of data, while use of an anti-conservative significance level results in too many false positive results. Approaches that are used to evaluate significance of linkage analysis results fall broadly into one of three categories, which we review below: (1) Reliance on previously-proposed thresholds; (2) resorting to asymptotic, or rarely, exact, distributions; and (3) simulation-based approaches.

Historically, reliance on previously-proposed thresholds and use of asymptotic distributions have been the predominant approaches to providing statistical significance of linkage analysis results. In the context of a known trait model, the widely-used lod score threshold suggested by Morton (1955) appears to be conservative in real analyses (Rao et al. 1978), as would be predicted from the underlying theory used to define the threshold. In contrast, a particular fixed threshold for VC analysis may be either conservative or anti-conservative, depending on the pedigree ascertainment and phenotype distribution (Blangero et al. 2000; Forrest and Feingold 2000). Other investigators use threshold values of statistics that yield specific genomewide p-values derived under asymptotic assumptions for particular data configurations (Lander and Kruglyak 1995). With these approaches, comparison of the observed data to those expected under some version of a null hypothesis is typical, with several types of measures used. For example, the MLS method for sib pairs Risch (1990) compares probabilities of marker data $\mathbf{Y}_M$ under estimated and null (no-linkage) values of *ibd* probabilities. The approach developed by Whittemore and Halpern (1994) requires a measure of *ibd* at chromosomal location γ, among individuals of like phenotype. The test statistic is then the expected value of this measure given the marker data $\mathbf{Y}_M$, which is compared to its expected value in absence of linkage, using an estimate of the variance computed under one of several possibilities. Problems with use of published and asymptotically-derived thresholds are that both the value and accuracy of a threshold needed to yield, say, a specific p-value, are data-specific, and different variance approximations yield different results (Jung et al. 2006).

To avoid use of approximations or threshold values for statistics, simulation-based approaches allow evaluation of significance in the context of a particular data set. Simulation approaches circumvent problems with use of published thresholds or distributional assumptions by retaining key elements of the original data by conditioning on some aspects of the real data. The result is an empirical distribution of data expected under the null hypothesis of no linkage. In human genetics, the most common simulation approach has been to retain the trait data but to simulate many times new marker data under the marker model assumed in the original analysis - the marker map and allele frequencies (Jung et al. 2006). An alternative for nuclear families with a range of possible phenotypes is to permute trait data among siblings, which retains but decouples both the original trait and marker data (Churchill and Doerge 1994). Among other advantages, marker simulation-based approaches can provide an assessment of the significance of a linkage signal obtained with Bayesian approaches (Daw et al. 2003). 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 $\mathbf{Y}_M$. However, the likelihood is a fundamental component of any Bayesian approach: the posterior distribution is proportional to the product of prior distribution and likelihood. Thus 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. Thus simulation-based approaches are of interest. However, they raise

computational challenges: even for analysis of nuclear family data, several weeks of computer time may be required to estimate the significance level of a single strong linkage signal for a chromosome with ~ 20 markers because of the need for large numbers of replicates needed to obtain enough extreme events to estimate the tail of the distribution accurately (Schellenberg et al. 2006). In situations where computations are challenging, such as for large pedigrees or where there are large numbers of traits and/or markers to be analyzed, use of current simulation-based approaches is currently too computationally-intensive to be practical.

B.4: Maps, markers, and genotyping error

Genetic maps are an important component of linkage analysis. Analysis with such maps facilitates pedigree and marker error detection (Boehnke and Cox 1997; Sieberts et al. 2002), increases the information available in pedigrees for initial linkage detection, and improves the accuracy of gene localization (Lathrop et al. 1984; Wijsman and Amos 1997) prior to gene identification. A genetic map consists of the order and meiotic intermarker distances between a series of marker loci. The construction and use of such a map assumes that the marker order is same in all individuals, and that each individual has exactly two alleles for each autosomal marker. Intermarker distances are assumed to be constant - either among all meioses for a sex-averaged map, or among all meioses from individuals of a particular sex for sex-specific maps. Although sex-averaged maps are commonly used in initial genome scans, it is well established that recombination rates differ in males vs. females, with extensive variation of the male-female recombination ratio (Broman et al. 1998; Donis-Keller et al. 1987). Most current linkage analysis programs allow incorporation of sex-specific maps when desired (Dietter et al. 2007).

Recent discoveries have identified submicroscopic polymorphic structural polymorphisms that result in violation of current assumptions used in gene mapping. Such variation has been identified both in marker order with identification of > 77 polymorphic inversions (Giglio et al. 2001; Stefansson et al. 2005; Tuzun et al. 2005; Feuk et al. 2006), and in genome content in the form of several thousand copy number variants (CNV) (Redon et al. 2006; Feuk et al. 2006; Wong et al. 2007). Both inversions and CNV have been implicated in genetic diseases (Yu et al. 2002; Sharp et al. 2006; Feuk et al. 2006; Rovelet-Lecruz et al. 2006), yet have only recently started being systematically studied. Polymorphic inversions, in particular, are currently likely to be underrepresented among known structural variants because balanced inversions are difficult to identify with current data and technologies. In contrast, CNV, to some extent can be identified with current genotyping panels, including dense SNP panels that identify CNV through tracts of apparent homozygosity, and even sparse microsatellite panels for which use of family data identifies deletions through Mendelian inconsistencies (Yu et al. 2002). Results of investigations that used parent-child trios suggest that use of relationship information combined with marker data improves identification of such CNV (Amos et al. 2003; Conrad et al. 2006).

Linkage analysis is sensitive to map and marker assumptions. Assumed intermarker distances, marker order, and relationship between inferred marker genotype and observed marker phenotype may all affect results. Poor map estimates can lead to either inflation or deflation of evidence for linkage(Daw et al. 2000; Fingerlin et al. 2006), and empirical results illustrate the impact of variability in estimated map distances. For example, an analysis of familial stroke using different map estimates on one chromosome gave lod scores ranging from 3.4-4.4 on the same data (Gretarsdottir et al. 2002). Similarly, careful data cleaning to identify and correct Mendelian-consistent genotype errors produced increases in maximum lod scores of 10%-100% in regions also identified in other studies of blood-pressure related traits compared to analyses that did not eliminate such genotypes (Chang et al. 2006).

Genotyping error and map assumptions become increasingly important with use of multipoint linkage analyses based on published map estimates. Many pedigree and genotyping errors elude detection in locus-by-locus analyses, particularly if connecting members of a pedigree are unobserved and/or if markers are individually relatively uninformative, as are SNPs. In multipoint analyses, Mendelian-consistent errors result in apparent low-probability multiple-recombinants in certain meioses (Boehnke and Cox 1997), or low-probability patterns of gene *ibd* among relatives (Browning 1998). If undetected, this can have serious consequences for the resulting analyses (Chang et al. 2006). Programs such as RELPAIR (Boehnke and Cox 1997) and our Eclipse program (Sieberts et al. 2002) can be used to identify potential errors in small pedigrees. 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. Also, even for analysis of small pedigrees, identifying and removing most genotyping errors in the context of a SNP genome scan is a time consuming and difficult task because of the large number of genotypes involved.

C: Progress report

Numbered references refer to the list of publications from this award; those numbered in **bold** are the ones we consider key to this renewal. Other references are given in the **Literature Cited** section.

C.1: Summary of previous specific aims (2003-2007)

1. To develop further Markov chain Monte Carlo (MCMC) methods, for realization of gene descent in extended pedigrees conditional on multilocus marker data, including new block-Gibbs samplers and importance sampling reweighting. Then, to use these sampling methods for improved analyses 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 both Bayesian and likelihood-based joint multilocus linkage and segregation analysis of complex traits, by extending the model framework of MCMC-based analyses to include discrete, ordered categorical, censored, and quantitative phenotypes. To extend methods to permit analyses of oligogenic traits and of several genetically correlated traits.

3. To develop robust methods for linkage detection using data on extended pedigrees, and to compare linkage detection approaches, by computing, conditional on multilocus marker data on extended pedigrees, the probability distributions of *ibd* measures, scoring statistics and significance measures. Also, to compare the performance of *ibd*-based, likelihood, and Bayesian approaches to linkage detection for complex traits.

4. To develop Monte Carlo methods using multilocus genetic marker data on extended pedigrees, to analyze genetic maps, meiosis models, and typing and pedigree error. To develop likelihood-based methods to investigate map uncertainty and map heterogeneity, alternative meiosis models, and models for haplotypes over multiple tightly-linked markers.

5. In addition to appropriate simulation studies, to evaluate methods developed in aims 1-4 on real data, by using data from dense SNP markers to evaluate MCMC approaches to linkage detection and oligogenic linkage analysis of correlated quantitative phenotypes with covariate information. Also, to explore the performance of MCMC methods on pedigrees of various sizes, with various missing data patterns, to evaluate the utility of data on such pedigrees.

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

C.2.1: Improved Monte Carlo methods

We have shown that our basic block-Gibbs LM-sampling method is competitive with other MCMC-based sampling methods on pedigrees for microsatellite markers, and superior for dense SNPs [**28**]. For example, for a set of 67 dense SNP markers on a 52-member pedigree with the first two generations of data missing, our **lm_markers** MORGAN program took 11 minutes to compute the lod score curve, while SIMWALK2 (Sobel et al. 2001) took 11 hours. Even with this difference in time, **lm_markers** gave more accurate results [**28**]. We therefore use our LM-sampler, and the **lm_markers** program in particular, as the standard against which to measure improvements.

We have continued to develop improved MCMC sampling methods to obtain realizations of multilocus inheritance patterns conditional on marker and/or trait data. These methods include Metropolis-Hastings sampling of trait locus positions [7], importance-sampling reweighting [7,**8**,34], and multiple-meiosis block-Gibbs samplers [36]. Implementation of computations using a factorial HMM system (Fishelson and Geiger 2004) have greatly increased the capacity for exact likelihood computation [32] and increased the efficiency of multiple-meiosis samplers [34,36]. Additionally, we have augmented the latent space of meiosis indicators with the haplotypes of key individuals who divide an extended pedigree [36]. This enables parts of likelihood computation to be performed exactly, and combined with MCMC on the remainder of the pedigree. Using a

simulated dataset with tightly linked microsatellite markers, constructed to give problems to the previous LM-sampler, we have shown that multiple-meiosis sampling and augmentation of the latent space by key individual haplotypes each gives significant improvement in MCMC performance and provides more accurate lod score estimates in shorter CPU time [36]. We have also improved other exact computation algorithms such as gene-graph peeling for computation of probabilities of trait data given underlying inheritance patterns [19], and extended sampling methods to more general models, such as two-locus trait models [**15**].

C.2.2: Methods for multilocus linkage and segregation analysis

Using the computational methods of the previous paragraph, we have improved estimation of likelihoods for multilocus linkage analyses [18, **21**, 34]. We have shown how sampling over trait-locus positions in a pseudo-Bayesian approach can be used to obtain MCMC estimates of multipoint lod scores [7] and that this approach works well on extended pedigrees with a high proportion of missing data [**8**], using a general liability-class model for age-of-onset. For simpler binary-trait models, we have shown that augmenting the space of latent variables with the haplotypes of a few key individuals provides much more accurate MCMC-estimated lod scores in less time on similar extended pedigrees [36], particularly when multiallelic markers are tightly linked. For more complex trait models, this approach remains to be implemented and evaluated.

We have also extended our methods to address more complex quantitative traits, including correlated traits [6], quantitative subphenotypes of a complex trait [10], models allowing for two QTL in addition to a polygenic component [**15**] and epistatic interactions [**16**]. We have investigated use of our MCMC sampling methods for joint oligogenic linkage and segregation analysis [29], using more complex trait models [24] including multiallelic trait loci [38]. We have also investigated their use for genome scans and compared them both to other MCMC programs and to exact methods, using either multiallelic markers or dense diallelic markers [**28**].

While we have not yet been able to substantively address quantitative trait models with covariate information, nor issues of missing covariates, our two-QTL models represent a major development for lod-score linkage analyses of complex traits. Our new program [**15**,**16**] is the first linkage program that can perform lod-score analyses using a parametric trait model including two (possibly linked) QTL and a polygenic component (2Q+P model). Additionally, as an MCMC-based program, it has (at least, in principle) no restriction on the number of markers nor on complexity or size of the pedigrees. Our simulation study [**15**] shows that parametric linkage analyses with two QTL provide higher power for linkage detection and better localization than analyses with simpler one-locus models [12] or variance-component (VC) models (Almasy and Blangero 1998). For example, over both different pedigree sizes and different spacings between two QTL contributing to a trait, lod scores for the strong-QTL location computed under the 2Q+P model were 1.2 to 2 times higher than the VC lod scores, which were very similar to lod scores using a single-QTL model [**15**]. Additionally, conditioning on the estimated location of the strong QTL, the 2Q+P model gave stronger evidence and better localization of the weaker QTL than did the VC lod scores.

C.2.3: Linkage detection on extended pedigrees: statistical significance and robustness

Computation of a multipoint lod score, linkage detection test statistic, or Bayesian linkage signal is only a first stage. A statistical interpretation must be given to the result, and for genome scans of complex traits this is not a straightforward procedure. In the case of a Bayesian signal we earlier developed a score enabling a frequentist (significance) assessment [3], but an exact p-value is hard to determine. Another approach to linkage detection is through a likelihood-based analysis of the dependence (under linkage) between trait data and the patterns of inheritance implied by marker data (Kong and Cox 1997). We developed an alternative model which models this dependence directly rather than through an *ibd* measure [37]. Asymptotic theory can be derived, but again an exact p-value is elusive. In the case of segregation analyses, with or without joint linkage analysis, issues of ascertainment also affect the significance of results. For the case of multivariate traits using polygenic models we addressed this ascertainment issue [14].

In this area, our greatest success has been in developing computationally efficient MCMC-based methods for the determination of exact p-values for linkage detection using *ibd*-based test statistics (for example, S-pairs (Whittemore and Halpern 1994)) that are robust to the trait model. Our methods avoid resimulation of marker data, which is computationally desirable if MCMC is required to impute latent inheritance patterns for every marker dataset, and also statistically desirable for robustness to the marker model Γ. First, we developed

permutation tests that permute trait data against the marker data. Our permutations provide correct exact p-values using data on multiple affected relatives in extended pedigrees, and our methods permit multiple permutations to be scored in a single MCMC run. However, on extended human pedigree structures, there may be few valid permutations. Thus we have also we have investigated use of the probability distribution of latent inheritance patterns and *ibd* measures (e.g. S-pairs) conditional on observed marker data to quantify statistical significance, using not a single value but a distribution of p-values [17,22]. More recently, this approach has been developed more formally in the framework of the fuzzy p-values of Geyer and Meeden (2005). Given observed data $\mathbf{Y}$, a fuzzy p-value is a random variable. In the current context, we consider first the p-value that would result were we able to observe latent patterns of genome inheritance. The fuzzy p-value is then the distribution of such p-values, given observed marker data $\mathbf{Y}_M$ [20]. This approach has been shown to provide useful measures simultaneously of the strength of evidence for linkage and of the uncertainty about such evidence [35]. We have compared power and robustness of resimulation-based approaches, fuzzy p-values, and permutation. When sufficient permutations are available, a permutation test is both powerful and robust to marker model misspecification. When insufficient valid permutations are available, the fuzzy p-value approach provides a computationally efficient alternative, much more robust to the marker model Γ than is resimulation of marker data [31]. The fuzzy p-value approach also permits extension to many other contexts. Specifically, we have already applied it to a lod score analysis, and shown how it provides a correct measure of significance of a maximum lod score, correcting naturally for multiple testing [**33**].

C.2.4: Genetic maps, meiosis models, marker models, and typing error

As data at increasingly dense genomic arrays of markers become available for genetic studies, we have placed increasing emphasis on questions relating to such maps [25]. On the COGA data for GAW14 we compared marker types and map assumptions in MCMC-based linkage analyses [12,30]. We also considered related issues such as comparison of STRPs and SNPs in the estimation of population structure from genomic data [11].

Publicly available genetic marker maps are based on relatively few meioses, and the coefficient of variation in estimates of small recombination fractions may be large. Large genetic epidemiological studies using pedigree data may contain at least a comparable number of meioses. To obtain improved genetic maps it is therefore important to be able to use such data, and to combine data over studies in a way that does not compromise privacy. In [**13**], we developed MCMC methods for efficient genetic map estimation, using data on extended pedigrees in which many ancestral individuals may be unobserved. Additionally, we showed how maps may be combined over studies, using MCMC based estimates of the Fisher information to determine an *equivalent number of meioses* on which each (possibly sex-specific) recombination fraction estimate is based.

Finally, we have begun to address issues of linkage disequilibrium among tightly linked markers in the analysis of pedigree data. We have investigated the effects of population subdivision and structure on the lengths of genome segments shared *ibd* in small populations [23]. We have incorporated a simple LD model into exact lod-score and *ibd*-based analyses of sib-pair data [39]. The model assumes that SNP alleles along founder haplotypes follow a Markov chain. Although an oversimplification of true LD patterns, this model permits the use of standard HMM computational algorithms (Baum et al. 1970; Fishelson and Geiger 2004) to obtain likelihoods and *ibd* probabilities conditional on marker data, and also permits easy investigation of the effect of varying LD intensity through varying the transition probabilities of the founder haplotype process. Using data simulated under our model, we found not only that failure to incorporate LD produces biases in *ibd*-based statistics and in lod scores (as is well known), but also investigated the loss of power caused by the presence of LD even when a correct LD model is used in the analysis [39]. For example, for simulated sets of 1000 affected sib pairs analyzed under the simulation LD model, weak LD had little effect on the lod score at the true trait location, but strong LD reduced this maximum lod score by 30% and very strong LD can reduce the lod score by over 60%.

C.2.5: Method comparison, including real-data analyses

We conducted detailed comparisons of our methods for identifying trait models with other exact approaches, and showed that there are advantages to estimating complex trait models with MCMC methods. In analyses of real quantitative traits, we used MCMC-based oligogenic segregation analysis to provide evidence for multiple loci affecting LDL particle size [1] and HDL levels [4] in large pedigrees segregating for hyperlipdemia, and

affecting age-of-onset in hereditary prostate cancer [2]. In each of these studies, validity of the estimated number of QTL is supported by results from the subsequent genome scans. We found that one of the QTL in the oligogenic model typically is also identified by classical complex segregation analysis, and that results from the MCMC-based oligogenic model are more robust to extreme data points than is complex segregation analysis with its single assumed Mendelian locus [10]. Using the same oligogenic approach, we also were able to find evidence that APOE on chromosome 19 is a genetic modifier of age-of-onset in pedigrees segregating the PS2 mutation on chromosome 1 [27]. Finally, we used polygenic models to investigate multiple correlated quantitative traits relating to autism, including an ascertainment adjustment [14], which demonstrated the sensitivity of such multivariate models to the sample ascertainment.

We showed that our MCMC linkage analysis approaches yield accurate results with modest computational requirements compared to other approaches. Linkage analysis of HDL variation [5] illustrates key points identified in several such comparisons: (1) the lod score (2.97) obtained with a trait model for 25 markers and MCMC computation was virtually identical to that obtained with 2 highly polymorphic markers and exact computation (2.98) with similar total computation times (1.5 hrs) for the large pedigrees used, for which exact computation is impractical with > 3 markers; (2) model-based lod scores were 15-28% higher than were VC lod scores, illustrating the advantage of a model-based analysis. We carried out similar comparisons for early-onset [**8**] and late-onset Alzheimer's disease [26] and for subphenotypes of dyslexia [9]. While most of these studies were based on standard microsatellite markers, we also showed that our MCMC linkage analysis methods could be used on data sets genotyped with dense SNPs for the COGA data of GAW14 [12]. Problems with long required run times identified in the GAW14 analysis stimulated improvements and reevaluation with the gene expression and SNP data of GAW15. For GAW15 [32], MCMC analysis with our improved samplers [36] yielded lod scores that differed by no more than 2% from those obtained with exact computation with much improved computational speed.

We normally carry out initial validation and testing of methods on simulated data with follow-up analyses using real data. For example, we have made extensive comparisons with our new multipoint analysis methods that permit a trait model with two QTL and a polygenic component [**16**, 32]. This again showed that a more complex model gives both higher lod scores and better trait resolution than a simpler VC model (see **C.2.2** above). However, real-data analyses not only produce important results in their own right [1,2,4,27], but also invariably raise new methodological challenges and suggest new performance studies to be done. For example, the challenges of the dense SNP maps of GAW14 [12] lead to studies using simulated data both of alternative MCMC-based approaches to multipoint linkage analyses [**28**] and of MCMC samplers [36].

C.2.6: Software development, implementation and distribution

Our methods are implemented, documented, and released in the MORGAN [40] package. Over the last 4 years we have moved from MORGAN_2.5 to MORGAN_2.8. In 2003, MORGAN_2.5 first included the pseudo-Bayesian estimation of lod scores [7,**8**]. In 2004, MORGAN_2.6 includes the first version of our **lm_markers** program, which has become the most flexible and versatile of our MCMC multipoint lod score programs, allowing lod scores under quantitative trait models, as well as liability-class models for discrete traits [18, **21**]. In 2005, we released MORGAN_2.7, which contained new programs for both MCMC estimation of multipoint lod scores and for multiple imputation of latent *ibd* given marker data, to assess significance and uncertainty in linkage findings [17,22,35]. MORGAN_2.8 (released in 2006) augments MORGAN_2.7 with two new MCMC programs, one performing a variety of likelihood-based and *ibd*-based linkage detection tests [31, 37], and the other a program for genetic map estimation using data on extended pedigrees [**13**].

Morgan programs not yet released include improved MCMC samplers and exact lod score computation methods for quantitative and qualitative traits [32,34,36]. Additionally we have a prototype MORGAN_3.0. Whereas currently-released MORGAN programs assume each trait to be controlled by a single trait locus (with possible polygenic component), MORGAN_3.0 separates trait loci from traits, and thus provides a much more general framework allowing for both epistasis and pleiotropy in trait models [**15**, **16**]. Additionally, we have maintained and improved the MORGAN web-based tutorial and examples [41], adding more realistic examples of newly-released programs. The use of the tutorial and examples in short courses both in US and in Europe has increased MORGAN use and given useful feedback on documentation and tutorial.

D: Research Design and Methods

In this section, all references both from this award and other literature are included in the **Literature Cited** component.

For ease of reference, we first introduce some general notation used in the remainder of this proposal. Generally, bold type is used for variables over multiple loci, while plain upper case is used for a single vector of variables over all the individuals of the dataset. Thus $\mathbf{S}$ refers to the complete set of meiosis indicators (Thompson 2000) at all locations or interest, while $S_{\bullet,j}$ refers to the inheritance vector (Lander and Green 1987) at location j. Multilocus marker data is denoted $\mathbf{Y}_M$, while for a single trait the data are denoted Y_T. Unphased genotypes at marker and trait loci are denoted $\mathbf{G} = (\mathbf{G}_M, G_T)$. The array $\mathbf{S} = \{S_{i,j}\}$ over all meioses i and loci j specifies the descent of genome in the pedigree. The complete genotypes $\mathbf{G}$ of all individuals are determined by $\mathbf{S}$ and the allelic types at all loci in the two haploid genomes of each founder. To specify identity by descent (*ibd*) we assign a unique label to each haploid founder genome in the pedigree data-set. We refer to these *founder genome labels* as FGL (Thompson 2005). General penetrance parameters relating trait data Y_T to underlying genotypes G_T are denoted β, trait locus allele frequencies are p, and the genomic location of a trait locus is γ. The general marker model is denoted Γ consisting minimally of marker locations $\boldsymbol{\lambda}$ and marker allele frequencies $\mathbf{q}$. Additionally, we will use $\mathbf{Z}$ for various auxiliary latent variables over loci, in our models for pedigree inheritance and population haplotypes.

D.1: MCMC and other computational algorithms

The focus of the new renewal will be on use of inheritance indicators $\mathbf{S}$ realized at marker loci conditional on marker data $\mathbf{Y}_M$ under marker model Γ, either by MCMC or by other Monte Carlo methods. The founder genome labels (FGL) of all pedigree members at all marker loci are determined by $\mathbf{S}$ and (with rare artificial exceptions) determine $\mathbf{S}$. Thus we may equivalently output either $\mathbf{S}$ or FGL, whichever is more convenient for the subsequent analyses. These $\mathbf{S}$ or FGL may be used in analyses of complex traits as detailed in later subsections. In the absence of genetic interference, the joint probability of marker data over n ordered marker loci $\mathbf{Y}_M = (Y_{\bullet,1}, ..., Y_{\bullet,n})$ and corresponding latent inheritance vectors $\mathbf{S} = (S_{\bullet,1}, ..., S_{\bullet,n})$ may be written

$$\mathrm{P}_\Gamma(\mathbf{Y}_M, \mathbf{S}) \;=\; \left(\prod_{j=1}^{n} \mathrm{P}(Y_{\bullet,j} \mid S_{\bullet,j})\right) \left(\frac{1}{2}\right)^m \left(\prod_{j=2}^{n}\prod_{i=1}^{m} \mathrm{P}(S_{i,j} \mid S_{i,j-1})\right) \tag{1}$$

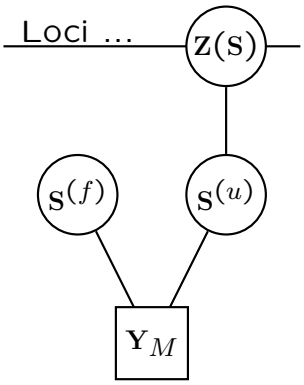
where $S_{i,j},\ i = 1, ..., m$ are the components of $S_{\bullet,j}$ over the m meioses of the pedigree dataset.

Equation (1) is the basis both of exact computational methods and of MCMC samplers. It defines the hidden Markov model (HMM) structure of the data $\mathbf{Y}_M$ over inheritance indicators $\mathbf{S}$ and hence allows either computation of $\mathrm{P}_\Gamma(\mathbf{Y}_M) = \sum_{\mathbf{S}} \mathrm{P}_\Gamma(\mathbf{Y}_M, \mathbf{S})$ or of $\mathrm{P}(S_{\bullet,j} \mid \mathbf{Y}_M) = \mathrm{P}(\mathbf{Y}_M, S_{\bullet,j})/\mathrm{P}(\mathbf{Y}_M)$ (Baum et al. 1970). The factorization of the locus-to-locus transition probabilities over meioses i in equation (1) provides a factored HMM (FHMM) system for which the computations can be achieved in time of order $mn2^m$ (Fishelson and Geiger 2004).

Aim 1(i): As computers improve in speed and memory, the size of pedigree on which exact computation is feasible becomes larger. Then also, at any genome location, computation of the probability distribution of inheritance indicators at any *single* position j conditional on data at multiple marker loci is feasible, as for example is done by Merlin (Abecasis et al. 2002). However, in order to score $\mathbf{S}$ and hence *ibd* jointly over locations j, Monte Carlo is still necessary. If a single forwards computation of $\mathrm{P}_\Gamma(\mathbf{Y}_M)$ is feasible, albeit computationally intensive, then backwards Monte-Carlo realization provides multiple independent realizations of $\mathbf{S}$ from the conditional distribution of $\mathrm{P}_\Gamma(\mathbf{S}|\mathbf{Y}_M)$ (Thompson 2000). We propose to investigate the limits of exact computation and independent realizations of $\mathbf{S}$ both as a tool for analyzing haplotypic regions of *ibd* in pedigrees (see **D4.3**), and also as a standard against which to compare the mixing performance of our MCMC methods when the data include multiple tightly linked SNP markers.

Aim 1(ii): We will continue to improve our MCMC LM-sampler, focusing particularly on enlarging the blocks that can be jointly updated. The currently-released LM-sampler updates inheritance indicators (components of $\mathbf{S}$) either for all meioses at one locus (L-step) or all the loci for a single meiosis (M-step). We have already extended the sampler to update multiple meioses jointly (Tong and Thompson 2007), and we will extend it further to deal with the joint updating of full inheritance vectors $S_{\bullet,j}$ at several SNP markers j. This

will circumvent some of the MCMC mixing limitations involved with very tightly linked marker loci (Sieh et al. 2005). We will also extend the *restricted updates* approach of Tong and Thompson (2007). These restricted updates greatly reduce the space to be considered in any one updating step, while still retaining the size of the blocks of variables updated and the HMM structure that facilitates computations.



The current restricted updates all involve sets of meioses; $\mathbf{S}^{(u)}$ are to be updated while $\mathbf{S}^{(f)}$ are held fixed at this step (see figure). For example, $\mathbf{S}^{(u)}$ might be the inheritance indicators all the meioses (over all loci) from parents to offspring in a large sibship. Dependent on the current $\mathbf{S}$, a process $\mathbf{Z}(\mathbf{S})$ is defined having the same HMM structure as $\mathbf{S}$ along the chromosome, but a much smaller state space than $\mathbf{S}^{(u)}$ at each locus. Other forms of restricted updates involving the full set of meioses in the pedigree but over a contiguous subset of the loci are likely also feasible. Note that in extending to updating of full inheritance vectors $S_{\bullet,j}$ jointly at several marker loci j our marker data become *phenotypic* rather than genotypic. That is, even for observed individuals the multilocus (phased) marker genotype may be uncertain. We address extension of our methods to a variety of types of phenotypic markers in **D.4** below.

Aim 1(iii): The first theme of the research proposed in this renewal is that of using a single set of realizations of $\mathbf{S}$ conditional on $\mathbf{Y}_M$ under marker model $\Gamma = (\boldsymbol{\lambda}, \mathbf{q})$ in multiple analyses (see **D.2** below). The second theme is that importance-sampling reweighting enables the same set of realizations to be used for analogous analyses under marker model $\Gamma^* = (\boldsymbol{\lambda}^*, \mathbf{q})$, which differs from Γ in the marker locations $\boldsymbol{\lambda}^* \neq \boldsymbol{\lambda}$. Each realization of $\mathbf{S}$ generated conditional on $\mathbf{Y}_M$ under map Γ is simply reweighted by the ratio

$$\frac{\mathrm{P}(\mathbf{S} \mid \mathbf{Y}_M; \Gamma^*)}{\mathrm{P}(\mathbf{S} \mid \mathbf{Y}_M; \Gamma)} = \frac{\mathrm{P}(\mathbf{Y}_M \mid \mathbf{S})\mathrm{P}(\mathbf{S}; \Gamma^*)}{\mathrm{P}(\mathbf{Y}_M; \Gamma^*)} \frac{\mathrm{P}(\mathbf{Y}_M; \Gamma)}{\mathrm{P}(\mathbf{Y}_M \mid \mathbf{S})\mathrm{P}(\mathbf{S}; \Gamma)} \propto \frac{\mathrm{P}(\mathbf{S}; \Gamma^*)}{\mathrm{P}(\mathbf{S}; \Gamma)} \tag{2}$$

this equation holding since the probability of marker data $\mathbf{Y}_M$ given $\mathbf{S}$ does not depend on the genetic marker map. Thus a single set of realizations $\mathbf{S}$ under Γ can be used to estimate *ibd* or lod scores under a variety of Γ^*. Equation (2) is valid for models Γ^* provided any $\mathbf{S}$ possible under Γ^* is also possible under the map Γ, and works well provided the maps Γ^* and Γ do not give probabilities to $\mathbf{S}$ that differ by many orders of magnitude. In practice, these constraints are not stringent. Obviously Γ should be a map under which any $\mathbf{S}$ consistent with $\mathbf{Y}_M$ can arise; for example, a sex-averaged map, with strictly positive recombination fractions between markers and no genetic interference. We have used this approach in the past to investigate interference (Thompson 2000) and found accurate and computationally efficient estimation under interference models Γ^*, including models of strong interference that precluded many $\mathbf{S}$. We will explore the limits of importance-sampling genetic-map reweighting of $\mathbf{S}$ in assessing the effect of sex-specific changes in recombination fractions on *ibd* distributions and lod scores.

Aim 1(iv): MCMC methods have advanced to the point where comparison with exact computation on extended pedigrees provides insufficient challenge. Even where exact computation is feasible, it takes orders of magnitude longer than the time required to get an accurate MCMC result (George et al. 2005; Tong and Thompson 2007). Thus we propose to develop alternative methods of assessing MCMC performance. In simulation studies, $\mathbf{S}$ and hence FGL are known. The FGL of observed individuals can be used to mimic a single very highly polymorphic marker, for which an exact lod score can, of course, be easily computed. This provides a comparison with lod scores from dense SNPs in a region, which jointly should provide close to the same level of *ibd* information as such a marker. This approach was used by Wijsman et al. (2006) to compare performance of MCMC programs using dense SNP markers, over multiple simulation datasets. We propose to investigate, validate, and extend this approach more widely. Specifically, we will use the approach of **D.3.2** below to compare the effects of alternative marker types and densities on levels of uncertainty in *ibd*.

The same approach may be used to provide guidelines for how many MCMC iterations will be required to provide accurate estimates of *ibd* probabilities or lod scores. For example, 100 datasets of dense SNPs could be simulated over an ~ 5 cM genomic region with the same data observation pattern and pedigree structures as the actual data, and the average number of scans needed to approximate the FGL-based likelihood or *ibd* probability computed. This will provide a guideline for the number of MCMC iterations required in a genome scan, since this number is determined primarily by the pedigree structure and relationships among observed individuals. Beyond about ~ 5 to ~ 10 cM, the number of MCMC iterations required does not increase, although of course the CPU time required for each MCMC iteration scales linearly with the number of markers. We will investigate the number of datasets and size of region that will be needed to provide reliable guidelines.

D.2: Joint linkage and segregation analyses of complex traits

For any multilocus marker data $\mathbf{Y}_M$ and any trait data Y_T observed on members of the pedigrees of a data set, the key formula that relates $\mathbf{Y}_M$ and Y_T via latent inheritance indicators $\mathbf{S}$ at marker locations is

$$\mathrm{P}_{\beta,\gamma,\Gamma}(Y_T \mid \mathbf{Y}_M) \;=\; \sum_{\mathbf{S}} \mathrm{P}_{\beta,\gamma}(Y_T \mid \mathbf{S})\mathrm{P}_{\Gamma}(\mathbf{S} \mid \mathbf{Y}_M). \tag{3}$$

(Lange and Sobel 1991). The probability of trait data Y_T given $\mathbf{S}$ depends only on the trait location (γ) and model (β), while the probability of $\mathbf{S}$ given $\mathbf{Y}_M$ depends only on the marker model (Γ). Given a collection of N realizations $\mathbf{S}^{(k)},\ k = 1.,,,,N$ from the distribution of $\mathbf{S}$ given marker data $\mathbf{Y}_M$, a Monte Carlo estimate of $\mathrm{P}_{\beta,\gamma,\Gamma}(Y_T \mid \mathbf{Y}_M)$ is obtained by averaging values of $\mathrm{P}_{\beta,\gamma}(Y_T \mid \mathbf{S}^{(k)})$ over the N realizations $\mathbf{S}^{(k)},\ k = 1.,,,,N$. The Monte Carlo estimate of $\mathrm{P}_{\beta,\gamma,\Gamma}(Y_T \mid \mathbf{Y}_M)$ can then be used to estimate the map-specific lod score (Ott 1999):

$$\log_{10}\left(\mathrm{P}_{\beta,\gamma,\Gamma}(Y_T, \mathbf{Y}_M) \;/\; \mathrm{P}_{\beta,U,\Gamma}(Y_T, \mathbf{Y}_M)\right) \;=\; \log_{10}\left(\mathrm{P}_{\beta,\gamma,\Gamma}(Y_T \mid \mathbf{Y}_M) \;/\; \mathrm{P}_{\beta}(Y_T)\right) \tag{4}$$

where U denotes that the trait locus is unlinked to the markers on the chromosome under consideration.

D.2.1: MCMC estimation of lod scores for complex traits (Aim 2(i)). We propose to use realizations of $\mathbf{S}$ conditional on marker data $\mathbf{Y}_M$ in a variety of new approaches to linkage analyses of complex traits. First, for complex quantitative traits, equations (3) and (4) permit efficient MCMC estimation of lod scores, since this approach works well provided trait data themselves do not provide strong evidence regarding inheritance patterns (Thompson 2000). This approach also permits easy investigation of sensitivity of lod scores to the trait model parametrized by β. Given a single set of realizations $\mathbf{S}$, lod scores may be computed not only for alternative locations γ of the (possibly several) trait loci, but also for several different trait models β, allowing study of the sensitivity of the lod score to the trait model. Multiple traits observed on the same pedigree structures may be rapidly analyzed using the same MCMC-based output of $\mathbf{S}$ or FGL.

Further, output $\mathbf{S}$ generated conditional on marker data $\mathbf{Y}_M$, or the derived FGL, can be used as input to other analyses. For example, the usual variance-component approach (Williams et al. 1999) uses *ibd* estimates that are functions of $\mathbf{S}$ generated by MCMC or computed exactly, using (for example) Loki (Heath 1997) or Merlin (Abecasis et al. 2002), respectively. Rather than reduce $\mathbf{S}$ to pairwise *ibd* estimates, we may model quantitative trait data conditional on $\mathbf{S}$ and average the likelihood contributions over the realized $\mathbf{S}$-values. That is averaging is at the level of the likelihood, rather than at the *ibd* level, and thus higher-order information on *ibd* is incorporated. Using this approach, we expect to regain some of the detection power and localization information typically lost in *ibd*-based and variance-component approaches relative to full model-based approaches (Abreu et al. 1999; Sham et al. 2000; Badzioch et al. 2005).

This approach of using a single large set of inheritance indicators $\mathbf{S}$ realized conditional on marker data $\mathbf{Y}_M$ for each chromosome or linked genomic region will be particularly useful in joint analyses of two linked (or unlinked) genomic regions for loci acting jointly on a trait. This applies whether the trait is discrete or quantitative, and whether using a variance-component trait model (Almasy and Blangero 1998), *ibd*-based linkage detection statistics (Biernacka et al. 2005), a Bayesian approach (Biswas et al. 2003), or the two-QTL parametric model of Sung et al. (2007). We propose to develop methods using a single set of MCMC-generated $\mathbf{S}$ from marker data $\mathbf{Y}_M$ and the quantitative trait models of Sung et al. (2007) to search for genomic regions

jointly affecting a trait. With the computational ability to address more complex trait models using this approach, we will also compare the gains of such models in terms of linkage detection and localization as compared to simpler trait models. The approach of Sung et al. (2007) compared favorably with variance component approaches (Almasy and Blangero 1998) in terms of fitting a second weaker linked QTL conditional on the first. We will compare the relative advantages of sequential and joint localization of QTL in terms of the relative strengths of the QTL and their degree of epistatic interaction.

D.2.2 Robustness and sensitivity of lod scores (Aim 2(ii)). When a trait model β_0 and trait loci locations γ_0 have been estimated, it is important to estimate changes in lod score in response to local changes in these parameters. For local exploration of the likelihood surface, MCMC sampling that includes the trait is preferred. The likelihood-ratio formula of Thompson and Guo (1991) becomes

$$\frac{\mathrm{P}_{\beta,\gamma}(Y_T \mid \mathbf{Y}_M)}{\mathrm{P}_{\beta_0,\gamma_0}(Y_T \mid \mathbf{Y}_M)} \;=\; \mathrm{E}_{\beta_0,\gamma_0,\Gamma}\left(\frac{\mathrm{P}_{\beta,\gamma}(Y_T \mid \mathbf{S})}{\mathrm{P}_{\beta_0,\gamma_0}(Y_T \mid \mathbf{S})} \;\mid\; \mathbf{Y}_M, Y_T\right), \tag{5}$$

where now $\mathbf{S}$ is sampled conditional on both Y_T and $\mathbf{Y}_M$. A single set of MCMC realizations permits accurate estimates of changes in the likelihood ratio for small changes to β and γ in the neighborhood of (β_0, γ_0), and hence provides estimates of the curvature of the likelihood surface in the neighborhood of the final estimate. We propose to implement this likelihood-ratio MCMC approach for our newer more complex trait models, enabling more complete exploration of the local likelihood surface around an estimated model. As models increase in complexity and both β and γ increase in dimensionality as additional QTL are added into models, effective methods for this exploration will become increasingly necessary.

Finally, the reweighting approach of equation (2) may be used to explore the impact of alternative genetic maps on a linkage lod score. Consider again the probability formula (3), and suppose that now a lod score under marker map Γ^* is desired:

$$\mathrm{P}_{\beta,\gamma}(Y_T \mid \mathbf{Y}_M;\Gamma^*) \;=\; \sum_{\mathbf{S}} \mathrm{P}_{\beta,\gamma}(Y_T \mid \mathbf{S})\mathrm{P}(\mathbf{S} \mid \mathbf{Y}_M;\Gamma^*) \;=\; \sum_{\mathbf{S}} \mathrm{P}_{\beta,\gamma}(Y_T \mid \mathbf{S})\frac{\mathrm{P}(\mathbf{S} \mid \mathbf{Y}_M;\Gamma^*)}{\mathrm{P}(\mathbf{S} \mid \mathbf{Y}_M;\Gamma)}\mathrm{P}(\mathbf{S} \mid \mathbf{Y}_M;\Gamma)$$

That is, to obtain an MCMC-based estimate of the lod score with marker map Γ^*, we may use realizations of $\mathbf{S}$ generated conditional on $\mathbf{Y}_M$ under map Γ, where marker maps Γ^* and Γ differ in the sex-specific marker locations $\boldsymbol{\lambda}$. Each realization is simply reweighted by the factor of equation (2). We will use this approach to investigate sensitivity of lod scores and linkage findings to trait model and marker map assumptions, both for single-locus models for a quantitative trait and also for more general trait models developed below.

D.2.3: Graphical models for mapping complex traits (Aim 2(iii)). A graphical model framework provides an alternative approach to linkage detection and estimation jointly with model-selection for a complex trait. This approach has been applied in the context of association studies and population data (Thomas 2005; Verzilli et al. 2006) to model dependence between markers and trait. We propose now to develop this approach in the analysis of pedigree data, where it is the inheritance pattern at marker loci linked to a trait locus that will show this dependence. In a graphical model, outcome variables are represented by nodes, and links show dependencies. Absence of a link implies conditional independence. To facilitate modeling and computation, attention is typically restricted to the decomposable class of graphical models (Wermuth 1976; Whittaker 1984; Verzilli et al. 2006). The goal is to sample from the space of graphs using MCMC to find a model that fits the data. A Bayesian approach is typically used (Verzilli et al. 2006) but a likelihood approach with respect to the graph structure can also be used using AIC (Akaike 1974) or BIC (Schwarz 1978) criteria.

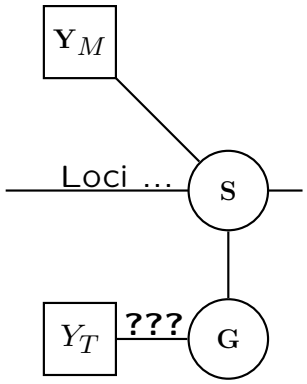
There are two main issues in fitting the graphical dependence structure which will determine genetic markers most directly linked to a trait phenotype. The first is the criterion for model fit, for which we propose to use a penalized likelihood approach (Viallefont et al. 2001). The second is the procedure for searching the space of alternative graphical models. Classically, stepwise backwards selection procedures have been used (Wermuth 1976), in which edges of the graph are removed one-by-one and the models with and without each edge compared using, for example, a deviance criterion (Whittaker 1984). Edwards and Havranek (1987) developed ways of reducing the number of models than must be considered, and Kreiner (1987) discusses exact tests of conditional

independence that determine the presence or absence of links in the graph. More recently, the Lasso method of Tibshirani (1996) has been used in graphical model selection, either to eliminate some dependency links by using Lasso to shrink some regression coefficients to zero (Huang et al. 2006), or by focusing on determination of the non-zero links in the neighborhood dependency structure (Meinshausen and Bühlmann 2006). We propose to investigate the application of modern methods of model selection and model-space search in determining the marker locations whose inheritance patterns show direct dependency to trait phenotypes in pedigree data.

Our approach again separates the genome inheritance indicators $\mathbf{S}$ sampled conditionally on marker data $\mathbf{Y}_M$ from the use of $\mathbf{S}$ to analyze trait data Y_T. Thus, we consider first the unrealistic case where $\mathbf{S}$ is observable, and hence the FGL that are present in each individual at each marker locus are known. Conditional on $\mathbf{S}$, we may employ established graphical-model procedures for model selection and parameter estimation (Madigan et al. 1995; Lauritzen 1996). Each marker locus is represented by a node, with variables that are the FGL of pedigree members. The models of primary interest concern which, and how many, marker loci have direct dependency links to trait phenotypes. Clearly, more links increase a model likelihood or Bayes factor; a penalty for model complexity is required (Viallefont et al. 2001).

Trait-locus allele frequencies, p, and the array $\mathbf{S}$ together determine genotypes $\mathbf{G}$ over all the individuals at a set of potential trait loci. Genotypes $\mathbf{G}$ are related to the trait data Y_T through a penetrance model that may involve many parameters, β:

$$\mathrm{P}(Y_T \mid \mathbf{S}) \;=\; \sum_{\mathbf{G}} \mathrm{P}_\beta(Y_T \mid \mathbf{G})\mathrm{P}_p(\mathbf{G} \mid \mathbf{S}) \qquad (6)$$



Instead of attempting to maximize this expression over β, we instead currently use with each $\mathbf{G}$ the $\hat{\beta}(\mathbf{G})$ which maximizes $\mathrm{P}_\beta(Y_T \mid \mathbf{G})$ obtaining an easily computed score function for the model. In some prototype small-pedigree examples, that this approach appears to be robust to the choice of p, can detect markers linked to trait loci. The goal, then, is to identify causal loci through the conditional independence, given FGL at the hypothesized causal loci, of trait phenotypes and FGL at other linked markers.

We will compare our scoring function of equation (6) with other *ibd*-based scoring statistics for linkage detection (Whittemore and Halpern 1994; McPeek 1999) and with permutation approaches (Basu et al. 2007). Our proposed scoring function contrasts genotypes $\mathbf{G}$ among affected and unaffected individuals. Hence it uses more information than an *ibd* measure that considers only *ibd* among affected individuals. We will examine the increased power, but possible decreased robustness, that this scoring function therefore provides.

We propose to investigate much more fully this graphical model approach to linkage detection and gene localization for complex traits. We will use simulated data on larger (although not large) pedigrees of 18 to 20 individuals, and chromosome-wide arrays of dense SNP markers. We will consider a wide array of phenotypes, including quantitative phenotypes with covariate information. Even on this size of pedigree, it will not be possible to form the summation over $\mathbf{G}$ in equation (6) exactly. We will use a Monte Carlo based approach, sampling $\mathbf{G}$ conditionally on $\mathbf{S}$ to provide a Monte Carlo estimate of the score function.

We will also investigate alternative approaches to accommodating more complex trait models with many model parameters β within the graphical-model framework. For more complex models, simply using for each $\mathbf{G}$ the value $\hat{\beta}(\mathbf{G})$ that maximizes the penetrance term in equation (6) will clearly be inadequate. We will retain a likelihood-based approach to the graphical-model dependence structure, testing which potential trait loci show direct dependence with trait phenotypes. However, it may be advantageous to take a Bayesian approach to the trait model, sampling over trait model parameters. Given a prior distribution, the trait data Y_T and genotypes $\mathbf{G}$ determined by $\mathbf{S}$ and an allele frequency p, it will be possible to sample realizations of β, and integrate (average) a likelihood-based scoring function such as (6) over the realized β as well as over $\mathbf{G}$.

Once the system works for a given $\mathbf{S}$, it will be straightforward (although possibly computationally intensive) to extend the ideas to make use a set of realizations of $\mathbf{S}$ sampled conditionally on marker data. Each realization will provide its appropriate weight to the graphical-model selection procedure. We may average over the sampled

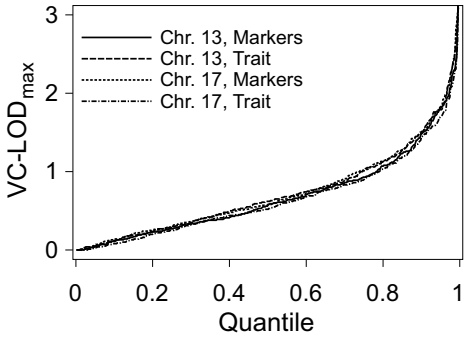
S, analogously to the likelihood evaluation of equation (3), or we may consider the distribution of model scores over the sampled **S** analogously to the latent-p approach of Thompson and Geyer (2007) (see **D.3.2** below). We will compare the properties and performance of these two approaches to dealing with uncertainty in **S**. While this scheme is ambitious, and much remains to be validated and explored, it offers a new way forward for joint segregation and linkage analyses of complex traits using pedigree data.

D.3: Assessment of significance and confidence sets

In the analysis of pedigree data, assessment of significance of *ibd*-based linkage detection test statistics (Whittemore and Halpern 1994), lod scores (Ott 1999), and Bayesian linkage signals (Daw et al. 2003) is a continuing issue. There is little point computing a lod score under a complex trait model, or lod score increase for modeling an additional trait locus, if the significance of the result cannot be assessed. The classical measure of statistical significance is a p-value, which is a function of data random variables **Y**, but the distribution of any function of **Y** is hard to obtain, even under the null hypothesis of no linkage.

D.3.1: Resimulation and permutation approaches (Aim 3(i)). One approach has been to condition on trait data Y_T and resimulate marker data $\mathbf{Y}_M$ independently of Y_T using an assumed marker map and marker allele frequencies, Γ, according to the same pattern of data availability as in the actual data (for example, Davis et al. (1996)). There are two major difficulties in using this approach on extended pedigrees, where ancestral pedigree members are unobserved. One is computational; each such resimulated dataset requires analysis. If MCMC is required to analyze each new marker dataset, this approach becomes computationally impractical (Basu et al. 2007). The other difficulty is statistical. P-values estimated from resimulation of marker data can be very sensitive to the marker model, especially to the marker allele frequencies. This lack of robustness to the marker model can lead to either false-positive or false-negative findings whenever founders of the pedigrees are unobserved (Basu et al. 2007). We propose to develop several alternative approaches that condition on the marker data. By keeping the marker data $\mathbf{Y}_M$ fixed, only one MCMC analysis to generate realizations of **S** given $\mathbf{Y}_M$ is required, and we expect also to gain robustness with respect to the marker model.

Our first approach involves resimulation of trait data. At first, this may seem to pose even greater challenges, since trait models are often problematic, and pedigrees often ascertained through trait data. However, trait models generated (for example) by using Loki (Heath 1997) and used to resimulate quantitative trait data on the pedigrees show similar distributions of trait phenotypes to the actual data. Preliminary exploration of this approach on a real data set for which there was evidence for linkage (Igo et al. 2006) gave a similar distribution of maximum VC lod scores under the null hypothesis for marker and trait resimulation, both for the chromosome with (chromosome 13) and without (chromosome 17) evidence for linkage in the original data set. This is illustrated by VC linkage analysis of 500 replicates of each configuration, sorted by lod score, for which the four distributions are essentially identical (see figure). An advantage of this approach is that we can analyze not only multiple resimulated trait datasets, using the same realizations **S**, but can also use alternative models, and investigate robustness to trait-model changes (equations (3) and (4)).



To validate this approach more fully we will develop pedigree-based measures of similarity in trait-phenotype distributions, and conduct simulation studies to insure valid p-values under the null hypothesis of absence of linkage.

A second approach is permutation, permuting trait data against the marker data. This approach is ideal for experimental crosses (Churchill and Doerge 1994), but can also be applied to pedigree data, permuting, for example, data on sibs who have no offspring. Conditioning on $\mathbf{Y}_M$, we may permute the trait data Y_T together with age, gender, and other covariate information on the siblings. Permutation tests are robust to the marker model, and, where there are sufficient valid permutations in the pedigree dataset, can be almost as powerful as resimulation under the correct marker model (Basu et al. 2007). At this stage, we do not propose to further develop permutation tests, but we will compare our other approaches with permutation-based approaches.

D.3.2: Latent test statistic and latent p-values (Aim 3(ii)). An alternative new approach to assessing significance conditions on both trait data Y_T and marker data $\mathbf{Y}_M$. The defining characteristic of a p-value is that, over the null distribution of data $\mathbf{Y}$, it has a uniform $U(0,1)$ distribution. Geyer and Meeden (2005) define a fuzzy p-value as the probability distribution of a random function of $\mathbf{Y}$, that, over the null distribution of $\mathbf{Y}$, has a $U(0,1)$ distribution. A randomized test of exact type-1 error α is one which rejects the null hypothesis with the probability that the fuzzy-p random variable is less than α. Thompson and Geyer (2007) have applied the same idea to latent variable problems and specifically to the detection of genetic linkage. Here the latent variables are indicators of gene descent $\mathbf{S}$ at a set of locations in the genome. Were $\mathbf{S}$ observable, the ideal test statistic for testing a genetic hypothesis would be a function of $\mathbf{S}$ and the trait data Y_T. Since the tests are conditional on Y_T, we suppress Y_T in the following notation, and denote this statistic as $V(\mathbf{S})$. The p-value $\pi(\mathbf{S})$ is the probability that, for $\mathbf{S}_0$ generated under the null hypothesis, $V(\mathbf{S}_0)$ exceeds $V(\mathbf{S})$. This is readily obtainable, since unconditional simulation of $\mathbf{S}_0$ is trivial. The conditional probability distribution of $\pi(\mathbf{S})$ given marker data $\mathbf{Y}_M$ is a fuzzy p-value, which we here call the *latent p-value*. Thompson and Geyer (2007) have shown: (1) the latent p-value distribution can be generated by a single MCMC run, providing realizations of $\mathbf{S}$ given $\mathbf{Y}_M$; (2) discreteness in $\mathbf{S}$ can be accounted for, to provide an exact p-value and exact type-1 error α for the resulting randomized test; (3) multiple testing can be corrected for, since application of the procedure to an omnibus statistic (for example, a maximum score over the genome) is no more complicated or computationally intensive than a pointwise test (for example, a score at a particular genome location).

In the context of trait-model-free linkage detection tests (Whittemore and Halpern 1994; Kruglyak et al. 1996), we have recently compared the latent-p approach with other exact tests based either on resimulation of marker data under the null hypothesis of absence of linkage or on permutation of marker data against the trait data (Basu et al. 2007). While the latent-p approach has lower power than a resimulation approach, it is computationally orders of magnitude more efficient. More importantly, it is much more robust to marker-model misspecification. Although the latent-p approach does assume a marker map and allele frequencies in order to impute underlying inheritance indicators $\mathbf{S}$, incorrect specification of the marker model has far less impact than a resimulation approach which resimulates marker datasets under this incorrect specification. The permutation test is robust to marker-model misspecification, and also computationally efficient, but on extended pedigrees of general structure valid permutations may be few and hard to determine. We propose to undertake more extensive investigations of the power and marker-model robustness of the latent-p approach, in order to gain understanding of the source of the lower power of this conditional test and hence to improve performance.

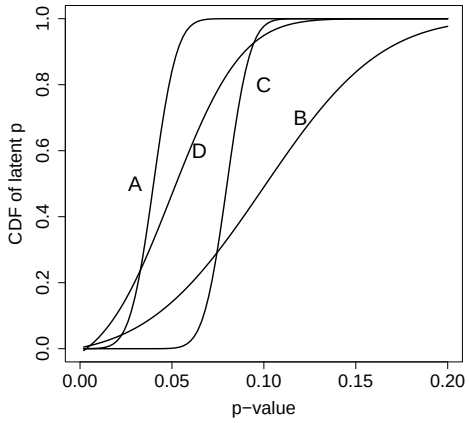
A major advantage of the latent-p approach is that it is not limited to trait-model-free linkage detection tests, but can be applied to any measure of linkage evidence, for example a lod score for a discrete or continuously varying trait. Were inheritance indicators $\mathbf{S}$ observable at all marker loci, the lod score at location γ would be a function of $\mathbf{S}$ and the fixed trait data Y_T. Specifically the latent lod-score test statistic at hypothesized trait location γ is $V_\gamma(\mathbf{S}) = -\log_{10} \mathrm{P}_\gamma(Y_T \mid \mathbf{S})/\mathrm{P}(Y_T)$ (see equations (3) and (4)). The latent p-value $\pi(\mathbf{S})$ is easily obtainable, just as before, by simulating inheritance indicators $\mathbf{S}_0$ unconditionally at locations on the chromosome of the markers. A single set of MCMC-based realizations of $\mathbf{S}$ given marker data $\mathbf{Y}_M$ then provides an estimate of the latent-p random variable. As with the trait-model-free linkage detection tests, correction for multiple testing is straightforward. In particular, the procedure may be applied to a maximum lod score, or to lod scores averaged over windows of contiguous genome locations, or to the maximum of such averages. We propose to explore this approach of assigning valid significance levels to lod scores, and to examine the sensitivity of the latent p-value to changes in the trait model. We will compare the latent p-value measures of significance with the classically accepted limit of a lod score of 3 (Morton 1955), and see whether and how this value should depend either on trait model complexity or genome-wide testing.

A disadvantage of the classical lod score approach for linkage analysis of complex traits is that a trait model β must be specified (equation (3)). Again, the latent p-value approach provides a way forward, since it can be applied to a lod score integrated over trait models β. The sampling of inheritance indicators $\mathbf{S}$ both under the null hypothesis and also conditionally on marker data $\mathbf{Y}_M$ is unaffected. The lod-score test statistic now depends explicitly on β,

$$V\gamma(\mathbf{S};\beta) \;=\; -\log_{10} \mathrm{P}_{\beta,\gamma}(Y_T \mid \mathbf{S})/\mathrm{P}_\beta(Y_T).$$

Trait models β will be sampled conditionally on Y_T and each realized $\mathbf{S}$, to provide a latent test statistic V integrated over trait models as suggested also in our proposed graphical models procedures (see **D.2.3**). Thus, we propose to combine a Bayesian approach to trait models with a significance-testing approach to linkage. We propose to investigate this approach to using latent lod scores integrated over trait models as test statistics for linkage, assigning significance via the latent-p approach.

On extended pedigrees with substantial missing marker data there is often considerable uncertainty about $\mathbf{S}$. The effect of uncertainty in $\mathbf{S}$ on the distribution of test statistics has been considered by Kruglyak et al. (1996), using measures of the expected entropy of the conditional distributions of $t(S_{\bullet,j})$ given $\mathbf{Y}_M$. More recently Nicolae and Kong (2004) have considered several measures of the information in marker data $\mathbf{Y}_M$ relative to what would be available if $\mathbf{S}$ were observed. An advantage of the latent-p approach is that it not only separates the strength of the evidence from uncertainty about that evidence, but that it puts this uncertainty directly onto the p-value scale. Consider, for example, the 4 possible latent p-values for an omnibus test shown in the figure. For case A, the entire probability mass of the latent-p distribution is below 0.05: no additional evidence is necessary. In case B, there is much uncertainty about the strength of the evidence, but the low probability that the latent-p is less than 0.05 indicates that collecting additional data on these pedigree structures is likely futile. In case C, the entire mass is around 0.07; such a finding might warrant a follow-up study, but additional pedigrees would be required. These pedigree structures have yielded all the evidence they can, and there is minimal remaining uncertainty. In case D, however, there is not only substantial remaining uncertainty, but a reasonable probability of a significant finding were this uncertainty to be reduced by typing additional markers or additional individuals on these same pedigree structures. By putting the uncertainty directly on the evidence (p-value) scale, the latent-p approach can guide the collection of additional data, where this is feasible. We propose to explore further this balance of uncertainty and evidence provided by the latent p-value, and the use of the latent p-value to determine potential or current influential marker loci or individual observations that contribute to reduction in uncertainty about $\mathbf{S}$ and/or strength of a linkage signal.



D.3.3: Confidence sets for trait locus locations (Aim 3(iii)). A test of a null hypothesis requires only an estimate of the distribution of a test statistic under that null hypothesis. However, if the null hypothesis of absence of linkage is rejected, then an estimate of the acceptable positions of trait loci is desirable. A genome location is in a $(1-\alpha)$-level confidence set if that location is *not rejected* by a test of type-1 error α. From Fisher (1934) to the present, genetic linkage analyses have focused on testing the null hypothesis of absence of linkage, using, since Smith (1953) and Morton (1955) the lod score statistic (equation (4))

$$\mathrm{lod}(\gamma) \quad = \quad \log_{10}(\mathrm{P}_\gamma(Y_T, \mathbf{Y}_M)/\mathrm{P}_U(Y_T, \mathbf{Y}_M)) \quad = \quad \log_{10}(\mathrm{P}_\gamma(Y_T \mid \mathbf{Y}_M)/\mathrm{P}(Y_T)).$$

Here, Y_T denotes the trait data, $\mathbf{Y}_M$ the marker data, and γ a hypothesized trait locus location, and for convenience parameters of the trait model (β) and marker model (Γ) are omitted. The subscript "U" denotes the null hypothesis: "Unlinked". Testing a given location γ_0 is in principle no harder, but first we require a test statistic, and a natural choice is the log-likelihood-ratio statistic: $-\log(\mathrm{P}_{\gamma_0}(Y_T|\mathbf{Y}_M)/\sup_{\gamma^*}\mathrm{P}_{\gamma^*}(Y_T|\mathbf{Y}_M))$ where the set of γ^* includes both other positions on the chromosome and also the unlinked position. Each γ_0 is in the level-α confidence set if the p-value for testing that γ_0 against others is *larger than* α.

Now given $\mathbf{S}$, Y_T and $\mathbf{Y}_M$ are independent, and thus the corresponding latent test-statistic becomes $-\log(\mathrm{P}_{\gamma_0}(Y_T|\mathbf{S})/\sup_{\gamma^*}\mathrm{P}_{\gamma^*}(Y_T|\mathbf{S}))$. The latent-p approach may be applied to the construction of confidence intervals using this latent test-statistic. A latent p-value for each γ_0 may be obtained, and the probability level for γ_0 in a $(1-\alpha)$-level latent confidence set is the probability that the corresponding latent-p is greater than α: that is, that γ_0 is *not* rejected. We have implemented this in a single protype example, constructing tests for each γ_0 separately, but clearly this is not computationally effective. The issue is that now trait data Y_T and marker-chromosome inheritance indicators $\mathbf{S}$ are dependent under the hypothesis γ_0. Simulation of $\mathbf{S}$ given

Y_T remains little harder than before, requiring first realizations of the inheritance pattern at the hypothesized trait locus. This is readily accomplished using existing MORGAN code. The issue is in the MCMC-based realizations of $\mathbf{S}$ conditional on both Y_T and $\mathbf{Y}_M$ since this is dependent on each particular γ_0.

We propose an importance-sampling approach, in which a single set of MCMC-based realizations of $\mathbf{S}$ are generated conditional on $\mathbf{Y}_M$, and then reweighted to provide their appropriate probabilities conditional also on trait data Y_T. These weights will depend on the γ_0 under consideration, but the MCMC need only be done once. We propose to explore the effectiveness of this importance-sampling approach, and its feasibility under a variety of trait models. It is known that in lod-score estimation, failure to condition on trait data Y_T in MCMC sampling can perform poorly where the trait data provide strong inheritance information or the marker data $\mathbf{Y}_M$ little (Thompson 2000). For a less strong trait the estimation approach of Lange and Sobel (1991) requiring MCMC conditional only on marker data $\mathbf{Y}_M$ works well (equation (3); **D.2.1**). Thus investigation of the effect of alternative types of trait, extent of trait and marker data, and different trait models will be an important part of this investigation.

The confidence interval approach provides different information from testing absence of linkage, U. A position γ_0 may be greatly preferred to the unlinked position, resulting in a low p-value for testing U, but other positions may be much more preferred resulting a low probability level for γ_0 in the confidence set. Conversely, a position γ_0 may be not preferred to the unlinked position in a test of type-1 error α leading to an intermediate p-value, but neither may other positions (including unlinked) be much preferred to γ_0, allowing γ_0 an intermediate level in the level $(1-\alpha)$ confidence set. We have seen both these cases arise for different locations γ in a single prototype test dataset. We propose to explore the use of confidence sets for genome locations, using the latent p-value approach, and to compare these with sets based on levels of the lod score. The latter are more akin to sets based on a pointwise quantile of the latent-p distribution (Thompson 2006), since the test statistic for testing U is the (latent) lod score. Confidence sets based on testing each γ-value may be tighter, rejecting positions close to markers even in a region with strong evidence for linkage.

D.4: Marker models, maps, haplotypes and linkage disequilibrium

Over the current grant cycle, we have paid increasing attention to the effects of new marker-types and to the effects of increased map densities on lod scores (Wijsman 2005) and on MCMC estimation of lod scores (Sieh et al. 2005; Wijsman et al. 2006). We have also developed new MCMC methods for the estimation of genetic marker maps using data on extended pedigrees with extensive missing data in earlier generations (Stewart and Thompson 2006). In published papers, we have not incorporated any error model into the observation of marker data, but Stewart (UW PhD Thesis) has extended the computational framework of our pedigree peeling programs (Thompson 2000) to allow penetrance probabilities for multiallelic loci. Thereby, he permits a marker error model for microsatellite markers to be used in the L-sampler of our MCMC MORGAN programs. We propose now to extend these efforts in several related directions.

D.4.1: Genotyping error and generalized marker types (Aim 4(i)). On average, typing error has little impact on analyses as typing error rates are low. However, where a typing error masks or implies a latent recombination event, the impact can be large (Chang et al. 2006). An error model for diallelic markers (SNPs) is readily implemented in our current LM-sampler framework, since we already have a framework of diallelic trait loci with, in principle, arbitrary penetrance functions. Stewart's setup for microsatellite markers permits extensions to, for example, marker haplotypes over several very tightly linked SNPs. In this case, even without error, phase uncertainty must be accommodated in the marker model: that is, the observed SNP genotypes do not determine the underlying segregating haplotypes. Additionally, SNP or microsatellite markers with null alleles can be detected in family studies (Yu et al. 2002; Amos et al. 2003). In some cases, this may be an "error"; the allele is present but for some reason not detected in typing. In other cases, there may be copy-number variants (CNV) resulting from small deletions (Redon et al. 2006), resulting in absence of any allele over several tightly linked markers. Moreover, detection of such variants and inclusion of them in analysis methods is of particular importance due to their likely impact on many complex disease phenotypes (Sharp et al. 2006). We propose to generalize the marker models that can be accommodated in our linkage analysis methods, allowing several tightly-linked SNPs to be considered as a single marker even when phase information is ambiguous or some part of the multi-SNP observation missing for some individuals. Likewise, we will allow

for CNV that result in "null" segregating alleles at a marker, extending this to the case where a segregating deletion may extend over several markers by jointly sampling inheritance indicators $\mathbf{S}$ over a block of markers. Extending the MCMC methods of Stewart and Thompson (2006) will allow for the detection of such marker variants from pedigree marker data. The same computational algorithms implemented within MORGAN will allow such markers to be used in our MCMC-based linkage analysis methods.

D.4.2: Genetic map heterogeneity and uncertainty (Aim 4(ii)). As markers become denser and more numerous relative to the numbers of meioses that can be scored for genetic map estimation, uncertainty in genetic marker maps becomes an increasing issue. Recombination hot-spots and the effects of small chromosomal rearrangements or deletions also contribute to this uncertainty, while at some scales genetic interference may be significant. In addition to being used directly for joint segregation and linkage analysis, for linkage detection, and haplotype estimation, the realizations of inheritance indicators $\mathbf{S}$ may be reweighted to accommodate alternative genetic maps. For example, indicators generated under a standard sex-averaged map may be reweighted for a sex-specific map, or for a modified genetic map, as described in section **D.1** (equation(2). We will investigate the extent to which a single set of realizations can be used both to explore the evidence for alternate genetic maps and the sensitivity of joint segregation and linkage analyses to the genetic marker map (see also **D.2.1**).

As with marker typing error, on average genetic map error has little effect, although it may reduce power (Daw et al. 2000). However, as with typing error, map errors that affect imputation of specific latent recombination events may have large impact. For example, failure to accommodate widely differing sex-specific maps in the region of a trait locus may lead to aberrant linkage signals (Dietter et al. 2007). MCMC-based inheritance indicators $\mathbf{S}$ conditional on marker data $\mathbf{Y}_M$ under a marker map Γ may be used to explore the effects of alternative genetic maps, analogously to the likelihood ratio formula for trait model robustness of lod scores (equation (5) of **D.2.2**). Using again the approach of Thompson and Guo (1991), a likelihood-ratio comparison of genetic maps Γ^* and Γ based on marker data $\mathbf{Y}_M$ is given by

$$\frac{L(\Gamma^*)}{L(\Gamma)} = \frac{\mathrm{P}(\mathbf{Y}_M;\Gamma^*)}{\mathrm{P}(\mathbf{Y}_M;\Gamma)} = \mathrm{E}_\Gamma\left(\frac{\mathrm{P}(\mathbf{Y}_M,\mathbf{S};\Gamma^*)}{\mathrm{P}(\mathbf{Y}_M,\mathbf{S};\Gamma)} \mid \mathbf{Y}_M\right) = \mathrm{E}_\Gamma\left(\frac{\mathrm{P}(\mathbf{S};\Gamma^*)}{\mathrm{P}(\mathbf{S};\Gamma)} \mid \mathbf{Y}_M\right) \tag{7}$$

the last equation holding since the $\mathrm{P}(\mathbf{Y}_M \mid \mathbf{S})$ does not depend on the marker map or meiosis model. The final probability is very easily computed for any given $\mathbf{S}$, and likelihoods for alternative maps may be explored by realizing $\mathbf{S}$ given $\mathbf{Y}_M$ under map Γ and estimating the likelihood ratio $L(\Gamma^*)/L(\Gamma)$ from these realizations.

As for the importance-sampling reweighting approach of equation (2), for this approach Γ and Γ^* can differ only in the genetic marker map $\boldsymbol{\lambda}$. Further, the maps Γ^* and Γ should not give probabilities to $\mathbf{S}$ that differ by many orders of magnitude, but the scope for exploring map uncertainty within these limits is broad. We propose to use this approach to investigate evidence for large differences in sex-specific maps and maps containing known inversions at certain points in the genome, perhaps those at which a study has indicated a linkage finding. We propose to investigate evidence for differences in our genetic maps, relative to a published map. Specifically, we will consider small map intervals where our realized $\mathbf{S}$ conditional on marker data $\mathbf{Y}_M$ suggest recombination has occurred.

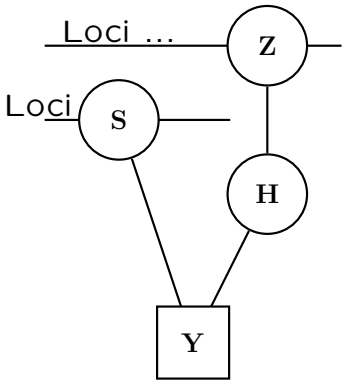
Uncertainty in marker allele frequencies $\mathbf{q}$ may often be as great a problem as uncertainty in the genetic map $\boldsymbol{\lambda}$. However, effects of changes in marker allele frequencies $\mathbf{q}$ cannot be explored directly via equation (7). One possibility is to realize not only $\mathbf{S}$ but also founder haplotypes $\mathbf{H}$, given $\mathbf{Y}_M$ and each realized $\mathbf{S}$. We have avoided this in the past since it greatly increases the size of the latent space and hence decreases accuracy of MCMC estimates based on realizations from this space. We have preferred always to integrate exactly over founder genotypes. However, as we move to denser marker maps, sampling founder haplotypes $\mathbf{H}$ is a possible approach to dealing with deletions that extend over too many loci for feasibility of the approach of **D.1**: that is, joint resampling of the block of inheritance vectors at these loci. Second, it permits both the importance-sampling reweighting of equation (2) and the likelihood-ratio formula of equation (7) to accommodate marker models Γ and Γ^* that differ both in map locations $\boldsymbol{\lambda}$ and marker allele frequencies $\mathbf{q}$. Third, it will enable us to extend our LD analyses of **D.4.3** below to pedigrees that are beyond the bounds of exact computation. Our current MCMC algorithms and MORGAN software will permit this additional sampling of founder haplotypes

H with little modification, and we propose to implement it, and explore its advantages and disadvantages.

D.4.3: The combination of linkage and association methods (Aim 4(iii)). As genetic marker maps become denser, failure to accommodate both linkage disequilibrium (LD) and an appropriate LD model in founder haplotypes can have a large impact on imputation of haplotypes shared *ibd* within a pedigree (Thomas 2007). In turn, this impacts the accurate imputation of founder haplotypes. Linkage analyses provide estimates of regions of *ibd* sharing within pedigrees, but the resolution is limited (Boehnke 1994). SNP haplotypes shared across pedigrees help with a finer-scale localization of a gene, and association methods rely on this among-pedigree sharing. Accurate realization of founder haplotypes in a pedigree under an LD model will serve to identify these shared regions across pedigrees, and lead to improved methods for combining linkage and association mapping. Currently we have only investigated the effects of incorporating LD in parental haplotypes on *ibd* imputation and lod scores from sib-pair data, using an unrealistically simple LD model. We propose now to extend this model and the range of pedigree structures on which we can examine the effect of incorporating LD into the analysis.

Thomas (2007) has recently proposed a framework for combining association and linkage mapping from pedigree data. However, whereas Thomas (2007) has used very high-dimensional models sampled via MCMC, we propose a more structured LD model with many fewer parameters. Our model will permit exact computation on small (3-generation) pedigrees, adopting a factored HMM (FHMM) computation approach. Scheet and Stephens (2006) model the chromosomes of a population as a mixture over a small number of clusters, the different clusters having different allele frequencies at SNP markers. At each marker j, a haplotype has a latent cluster membership Z_j; $\mathbf{Z} = \{Z_j\}$. At marker j the allele of a haplotype is sampled according to the allele frequencies at j for cluster Z_j. In each haplotype, cluster membership changes along the chromosome according to a first-order Markov process, and thus the haplotypes of the population exhibit LD generated by this HMM model.

We propose to use the model of Scheet and Stephens (2006) for founder haplotypes **H**, combining this with the within-pedigree inheritance indicators **S** (see figure). Given marker and/or trait data **Y** this will permit exact computation of *ibd* sharing and lod score computation on small pedigrees in which the founders are typically unobserved. Due to the independence among founder haplotypes in the latent cluster-membership process and independence among latent meioses in the recombination process along a chromosome, both parts of the model have an FHMM process (Fishelson and Geiger 2004). This will permit exact computation using a forwards-backwards algorithm (Baum et al. 1970), provided neither the number of meioses nor number of founder haplotypes is too large. Parameters of the



latent processes may be estimated via an EM algorithm (Dempster et al. 1977). We propose to explore this model in terms of the patterns and strengths of LD it gives rise to, and the fit to publicly available population haplotype data such as of the International Hapmap Consortium (2005).

On a small pedigree, the probability of each *single* inheritance vector $S_{\bullet,j}$ and of the resulting *ibd* may be computed exactly However, in order to analyze patterns of *ibd jointly* over sets of dense SNP markers Monte Carlo is necessary. A single forwards pass of exact computation and stored partial sums along the chromosome enable *independent* realizations of **S** conditional on marker data $\mathbf{Y}_M$ (see **D.1**). We propose to develop methods of estimating the model parameters, using data on small pedigrees, both when founder individuals are themselves observed and when they are not. We propose to use the estimated LD model in developing methods that incorporate population LD into the estimation of *ibd* sharing, computation of lod scores, and imputation of shared haplotypic regions, using multilocus marker data on small pedigrees. We will investigate the effect of incorporating LD on the results of these computations, comparing to the case where no LD is modeled. We will assess the power to detect linkage gained by scoring of *ibd* jointly over dense SNP markers in LD.

Where the pedigree size exceeds the bounds for practical exact computation, MCMC provides an alternative.

For example, updating both founder cluster-memberships $\mathbf{Z}$ and meioses $\mathbf{S}$ over a small block of SNP loci in a *single* multilocus MCMC step, as proposed in section **D.1**, is feasible. For larger blocks, we would need to augment the latent space and sample both $\mathbf{Z}$ and $\mathbf{S}$ as described in section **D.4.2** above, with cluster-membership $\mathbf{Z}$ here replacing the direct sampling of founder haplotypes $\mathbf{H}$. In either case, dependent realizations of the latent variables given the observed marker and/or trait data are obtained. Although an exact EM algorithm is no longer feasible, Monte Carlo EM (Guo and Thompson 1994; Stewart and Thompson 2006) is a practical alternative. For given parameter values and given marker data $\mathbf{Y}_M$, $\mathbf{S}$ and latent class values $\mathbf{Z}$ can be realized, and new parameter values estimated from these realizations. We will investigate the feasibility of this approach for incorporating LD into analyses on 3 to 4 generation pedigrees, specifically to analyze the effect of incorporating LD when unobserved founders are more than one generation removed from the observed marker data.

D.5: Evaluation of methods on real data

In addition to simulation studies of performance, we will continue to test and apply our methods in the context of real data. Several real data sets of complex traits are available to us for the purpose of evaluating and extending our methods (Table). Recruitment and phenotyping for two of the studies (Guam and DYS) has been completed, and the remaining three studies (AD, AUT, and FCHL) have large components for which recruitment has been completed, although some additional data collection, including additional phenotypes and recruitment of a small number of additional subjects, is still in progress. All data sets have complete microsatellite (STR) genome scan data available at an $\sim$ 10 cM density. Two of them (AUT, AD) have a subset of families also with a full SNP genome scan, one (AUT) with the Affymetrix 10K v2 SNP panel, and one (AD) with the Illumina Linkage IVb panel. Two of the data sets (AD, Guam) have families that are segregating the known inversion polymorphism on chromosome 17q (Stefansson et al. 2005), and one of the data sets (AUT) has CNV (Yu et al. 2002; AGP 2007). All of these five 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 sizes and structures, phenotypes, missing data patterns, marker density, and type of marker provides ample comparative material for evaluation of methods.

Data Set	AD	FCHL	Guam	DYS	AUT
Pedigrees					
Large	+	+	+	+	−
Small	+	+	−	+	+
Complex	(+)	(+)	+	−	−
Phenotype					
Quantitative	−	+	−	+	+
Age-of-Onset	+	−	+	−	−
Ordered Categorical	+	−	(+)	+	+
Affected/Unaffected	+	+	+	+	+
Known Covariate	−	+	−	+	+
Markers					
Whole STR Genome	+	+	+	+	+
Whole SNP Genome	+	−	−	−	+
Some dense regions	+	+	−	+	+
Phased haplotypes	+	−	−	−	−
Structural variants					
Inversions	+	−	+	−	−
CNV	−	−	−	−	+
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.

The AD data set is an ongoing study of the genetics of Alzheimer's disease. This data set will be useful for testing approaches for joint multilocus linkage analysis and segregation analysis methods (**D.2**), evaluation of measures to infer the statistical significance of linkage (**D.3**), evaluation of ability to detect effects of inversions (**D.2.2**) and map heterogeneity (**D.4.2**) on linkage detection, and methods for incorporating and evaluating haplotype inferences (**D4.3**). The data set includes 3 known genes (Goate et al. 1991; Levy-Lahad et al. 1995a; Sherrington et al. 1995) in 25 of the available families. A known inversion polymorphism on chromosome 17 (Stefansson et al. 2005) is segregating in a subset of 9 of these pedigrees that have been genotyped for an inversion-specific marker. Molecularly-phased multiple-SNP haplotypes in two regions of chromosome 19 are available for 86 pedigrees (Yu et al. 2004; Sieh et al. 2007) for comparison to methods that model linkage disequilibrium or estimate haplotypes (**D.1, D.4.3**). The 491 available pedigrees, encompassing a total of 4006 individuals, range widely in size: the smallest are sib-pair families, the largest is 161 individuals, 6 pedigrees have 44-73 individuals each, and another 12 pedigrees have 20-35 individuals each. There is 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 in the larger pedigrees, and 50% in the smaller pedigrees.

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 as well as for multilocus trait models (**D.2**), for evaluation of different approaches for estimating statistical significance of results (**D.3**), for testing methods of detection of map heterogeneity(**D.4.2**), for haplotype inference methods (**D.1**), and for investigating approaches for incorporation of linkage disequilibrium among markers (**D.4.3**). A full 10 cM genome scan is available for 15 large pedigrees (up to 88 individuals/pedigree, 567 individuals), with all original marker typing, as well as cleaned data, available. This initial data set is augmented by an additional cleaned data set of 542 individuals in 19 families of up to 136 individuals, which were also genotyped in a full microsatellite genome scan. The data set includes not only initial microsatellite genome scan data, but also dense SNP and STR genotypes in regions with evidence of linkage and/or in candidate genes.

The Guam data set consists of a huge, complex, multi-generation pedigree. The disease phenotype is a neurodegenerative syndrome with two different phenotypes from a population on the island of Guam, and includes marker and phenotype data from a genome screen primarily 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. This pedigree will be useful for methods dealing with estimation of *ibd* in complex pedigrees (**D.1**), for investigation of models for dealing with categorical, age-of-onset, and/or reduced penetrance phenotypes (**D.2.1**), and for investigation of methods to explore the effects of inversions on linkage analyses (**D.2.2**).

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 multivariate phenotypic measures with variable measurement error at different points in the phenotypic spectrum (**D.2**). DYS currently contains 144 families with full genome scan data in 1131 individuals with phenotype and genotype data. Over 35 phenotypic measures are available, including discrete, quantitative, and categorical phenotypes, many of which are strongly correlated (**D.2**). There are also a number of moderately large 3-generation pedigrees ($\sim 18-25$ individuals), which are large enough to provide some computational challenges, but still small enough that some exact computation is possible for comparison. Some regions of the genome in which evidence of linkage has been obtained also have dense STR marker coverage. AUT consists of 267 nuclear families with full genome scan data and at least 2 affected children (Schellenberg et al. 2006), and is projected to consist of 350 fully genotyped and phenotyped families by mid-2007. The pedigrees are small enough that a variety of exact computation is possible for comparison with results obtained with MCMC approaches, including comparison of other simulation-based approaches for obtaining empirical significance levels with standard marker resimulation approaches (**D.3**). Approximately 2/3 of the families also have a dense SNP genome scan (**D.4**), as well as a second, largely non-overlapping, microsatellite genome scan, thus providing in total a denser microsatellite scan than is typically available. This data set also contains CNV, detectable with SNPs and in some cases STRs with confirmation

in some regions based on cloning and sequencing of the regions (Yu et al. 2002; AGP 2007). This AUT data set includes several discrete, ordered categorical, and continuous phenotypic measures (**D.2**).

D.6: Software development

We propose to continue to develop and implement computational methods within the framework of our MORGAN software. In addition to improving and extending the implementation of methods within the current MORGAN_2 framework (MORGAN 2.8.1, released April 2006), we propose to develop our new MORGAN_3 and gradually transfer key programs to this new environment. The fundamental change with MORGAN_3 is that trait loci are no longer tied to single trait models. Thus a trait phenotype may be affected by genotypes at several loci, and the genotypes at a putative trait locus may simultaneously and differentially affect several traits. By breaking the direct link between trait locus and trait, we enable a far more flexible framework for complex models. We will be able also to accommodate multiple traits, for example performing the trait-resimulation approaches of section **D.3** not only within a single MCMC run but also within a single analysis run.

The first program to be released within the new MORGAN_3 framework will be our MCMC-based lod score analysis for a trait model incorporating two major genes and a polygenic component (Sung et al. 2007) as the full power of that approach can only be efficiently realized within this new framework. Our lodscore programs will be the next for transfer to the more general trait environment, especially our most flexible and heavily-used **lm_markers** program (Thompson 2005). This program already allows for discrete and continuously varying traits, and for liability class models. The latter are however currently in crude input format. We propose both to widen the classes of trait models which can be used with our MCMC-based lod score programs, and to improve the user-input interface for such models. Although the **lm_markers** program has recently been improved to allow for exact computation on small pedigree components using the factored hidden Markov model (HMM) methods of Fishelson and Geiger (2004), on large pedigree components it still uses the LM-sampler of Thompson (2000). We propose to implement the improved multiple-meiosis sampler of Tong and Thompson (2007), which uses an MCMC-based factored HMM method, into **lm_markers** and, as time permits, into our other MCMC analysis programs.

Our web-based tutorial materials have seen major improvement over the last grant cycle, but much remains to be done, particularly in the area of more realistic examples. To date, the tutorial and examples development have been mainly as needed for a variety of short-courses, and, for efficiency, examples have often been based on the same input files as are the MORGAN test gold-standards for each program. These short-courses have provided very useful feedback, improving the user interface for running examples and the documentation of output. However, gold-standards and tutorial examples serve very different purposes. Gold-standards are designed to test validity of the program and alternative parameter combinations, while the tutorial aims to help users learn the software and provide useful default input options. We now propose a concerted effort to bring tutorial and examples up-to-date, and to maintain them, making examples that mimic the types of data that users have, and providing examples of many of our newer analysis programs. Additionally, we propose to facilitate the use of MORGAN on datasets that may have been previously analyzed by FASTLINK (O'Connell and Weeks 1995) or Merlin (Abecasis et al. 2002) by providing scripts that produce default MORGAN parameter statements and input data files from MORGAN or FASTLINK input files.

Publications resulting from the current grant

References **8**, **13**, **15**, **16**, **21**, **28**, and **33**, numbered in **bold** are the ones we consider key to this renewal. Public web references are provided where available. For references **21** and **33**, PDF copies are attached.

(i) Papers published since previous renewal submission.

1. Badzioch, M.D., Igo, R.P. Jr., Gagnon, F., Brunzell, J.D., Krauss, R.M., Motulsky, A.G., Wijsman, E.M., Jarvik, G.P. (2004) LDL particle size loci in familial combined hyperlipidemia: Evidence for multiple loci from a genome scan. *Arteriosclerosis, Thrombosis and Vascular Biology* **24:** 1942–1950. http://atvb.ahajournals.org/cgi/content/full/24/10/1942
2. 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., and Wijsman, E.M. (2003) Oligogenic segregation analysis of hereditary prostate cancer pedigrees: evidence for multiple loci affecting age-at-onset. *International Journal of Cancer* **105:** 630–635. http://www3.interscience.wiley.com/cgi-bin/fulltext/104085612/PDFSTART
3. Daw, E.W., Wijsman, E.M., and Thompson, E.A. (2003) A score for Bayesian genome screening. *Genetic Epidemiology* **24** 181–190. http://www3.interscience.wiley.com/cgi-bin/fulltext/104083933/PDFSTART
4. Gagnon, F., Jarvik, G.P., Badzioch, M.D., Motulsky, A.G., Brunzell, J.D., and Wijsman, E.M. (2005) Genome scan for quantitative trait loci influencing HDL levels: evidence for multilocus inheritance in familial combined hyperlipidemia. *Human Genetics* **117:** 494–505. http://www.springerlink.com/content/x64565677x107273/
5. Gagnon, F., Jarvik, G..P, Motulsky, A.G., Deeb, S.S., Brunzell, J.D., and Wijsman, E.M. (2003) Evidence of linkage of HDL level variation to APOC3 in two samples with different ascertainment. *Human Genetics* **113:** 522–533. http://www.springerlink.com/content/5ega5b690e2d7f95/
6. George, A.W., Basu, S., Li, N., Rothstein, J.H., Sieberts, S.K., Stewart, W., Wijsman, E.M., Thompson, E.A. (2003) Approaches to mapping genetically correlated complex traits. *Biomed Central Genetics* **4:** (Suppl) 71. http://www.biomedcentral.com/1471-2156/4/S1/S71
7. George, A.W. and Thompson, E.A. (2003) Multipoint linkage analyses for disease mapping in extended pedigrees: A Markov chain Monte Carlo approach. *Statistical Science* **18:** 515–531. http://projecteuclid.org/Dienst/UI/1.0/Summarize/euclid.ss/1081443233
8. George, A. W., Wijsman, E. M. and Thompson, E. A. (2005) MCMC Multilocus Lod Scores: Application of a New Approach. *Human Heredity* **59:** 98–108. http://content.karger.com/produktedb/produkte.asp?typ=fulltext&file=HHE2005059002098
9. Igo, R.P.Jr., Chapman, N.H., Berninger, V.W., Matsushita, M., Brkanac, Z., Rothstein, J.H., Holzman, T., Nielsen, K., Raskind, W.H., and Wijsman, E.M. (2006) Genomewide scan for real-word reading subphenotypes of dyslexia: Novel chromosome 13 locus and genetic complexity. *American Journal of Medical Genetics, Neuropsychiatric Genetics* **141B:** 15–27. http://www3.interscience.wiley.com/cgi-bin/fulltext/112170425/PDFSTART
10. Igo, R.P.Jr., Chapman, N.H., and Wijsman, E.M. (2006) Segregation analysis of a complex quantitative trait: approaches for identifying influential data points. *Human Heredity* **61:** 80–86. http://content.karger.com/produktedb/produkte.asp?typ=fulltext&file=HHE2006061002080
11. Leutenegger, A.-L., Prum, B., Genin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., and Thompson, E.A. (2003) Estimation of the inbreeding coefficient through use of genomic data. *American Journal of Human Genetics* **73:** 516–523. http://www.journals.uchicago.edu/AJHG/journal/issues/v73n3/40058/40058.html
12. Sieh, W., Basu, S., Fu, A.Q., Rothstein, J.H., Scheet, P.A., Stewart, W., Sung, Y.J., Thompson, E.A., and Wijsman, E.M. (2005) Comparison of marker types and map assumptions using MCMC-based linkage analysis of COGA data. *Biomed Central Genetics* **6** (Suppl 1): S11. http://www.biomedcentral.com/1471-2156/6/S1/S11

13. Stewart, W.C.L., and Thompson, E.A. (2006) Improving estimates of genetic maps: A maximum likelihood approach. *Biometrics* **62:** 728–734. http://www.blackwell-synergy.com/doi/abs/10.1111/j.1541-0420.2006.00532.x?cookieSet=1&journalCode=biom

14. Sung, Y.J., Dawson, G., Munson, J., Estes, A., Schellenberg, G.D., and Wijsman, E.M. (2005) Genetic investigation of quantitative traits related to autism: use of multivariate polygenic models with ascertainment adjustment. *American Journal of Human Genetics* **76:** 68–81. http://www.journals.uchicago.edu/AJHG/journal/issues/v76n1/41726/41726.html

15. Sung, Y.J., Thompson, E.A., and Wijsman, E.M. (2007) MCMC-based linkage analysis for complex traits on general pedigrees: multipoint analysis with a two-locus model and a polygenic component. *Genetic Epidemiology* **31:** 103–114. http://www3.interscience.wiley.com/cgi-bin/fulltext/113479694/PDFSTART

16. Sung, Y.J., and Wijsman, E.M. (2007) Accounting for epistasis in linkage analysis of general pedigrees. *Human Heredity*: **63:** 144–152. http://content.karger.com/produktedb/produkte.asp?typ=fulltext&file=HHE2007063002144

17. Thompson, E.A. (2003) Linkage detection for complex traits. In *Invited Proceedings of the 54th Session of the International Statistical Institute*, Berlin, Germany.

18. Thompson, E.A. (2003) Chapter 30: Linkage Analysis. (Revised and expanded) In *Handbook of Statistical Genetics 2 nd ed.* D.J.Balding, M. Bishop and C.Cannings (eds). Pp. 893-918. Wiley: Chichester, UK.

19. Thompson, E.A. (2003) Information from data on pedigree structures. In: *Science of Modeling: Proceedings of AIC 2003.* T. Higuchi, Y. Iba, and M. Ishiguro (eds) Pp. 307–316. Research Memorandum of the Institute of Statistical Mathematics, Tokyo, Japan.

20. Thompson, E.A. (2005) Fuzzy and randomized confidence intervals and p-valuesi (Discussion). *Statistical Science* **20:** 382–383. http://projecteuclid.org/Dienst/UI/1.0/Summarize/euclid.ss/1137076657

21. Thompson, E.A. (2005) MCMC in the analysis of genetic data on pedigrees. In *Markov Chain Monte Carlo: Innovations and Applications.* F Liang, J-S Wang, and W Kendall (eds). Pp. 183–216. Lecture Note Series of the IMS, National University of Singapore. World Scientific Co Pte Ltd, Singapore. (PDF file Attached.)

22. Thompson, E A. and Basu, S. (2003) Genome sharing in large pedigrees: multiple imputation of *ibd* for linkage detection. Human Heredity **56:** 119–125. http://content.karger.com/ProdukteDB/produkte.asp?Aktion=ShowPDF&ProduktNr=224250&Ausgabe=229632&ArtikelNr=73739.pdf

23. Thompson, E. A. and N. H. Chapman (2004) Haplotype blocks in small populations. In Computational methods for SNPs and haplotype inference: DIMACS/RECOMB Satellite Workshop, Piscataway, NJ, Nov.2002. S. Istrail, M. Waterman and A. Clark (eds.). Springer-Verlag Lecture Notes in Computer Science, Vol. 2983, Pp. 74–83.

24. Wijsman, E.M. (2003) Summary of group 8: Development and extension of linkage methods. *Genetic Epidemiology* **25** (Suppl 1): S64-S71. http://www3.interscience.wiley.com/cgi-bin/fulltext/106565537/PDFSTART

25. Wijsman, E.M. (2005) Gene mapping and the transition from STRPs to SNPs. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics.* LB Jorde, PFR Little, MJ Dunn, S Subramaniam, eds. Wiley & Sons, Inc.

26. Wijsman, E.M., Daw, E.W., Yu, C-E., Payami, H., Steinbart, E.J., Nochlin, D., Conlon, E.M., Bird, T.D, and Schellenberg, G.D. (2004) Evidence for a novel late-onset Alzheimer's disease locus on Chromosome 19p13.2. *American Journal of Human Genetics* **75:** 398–401. http://www.journals.uchicago.edu/AJHG/journal/issues/v75n3/41239/41239.html

27. Wijsman, E.M., Daw, E.W., Yu, X., Steinbart, E.J., Nochlin, D., Bird, T.D., and Schellenberg, G.D. (2004) APOE and other loci affect age-at-onset in Alzheimer's disease families with PS2 mutation. *American Journal of Medical Genetics (Neuropsychiatric Genetics)* **132B:** 14–20. http://www3.interscience.wiley.com/cgi-bin/fulltext/109630662/PDFSTART

28. Wijsman, E.M., Rothstein, J.H., and Thompson, E.A. (2006) Multipoint linkage analysis with many multiallelic or dense diallelic markers: MCMC provides practical approaches for genome scans on general pedigrees.

American Journal of Human Genetics: **79:** 846–858.
http://www.journals.uchicago.edu/AJHG/journal/issues/v79n5/43843/43843.html

29. Wijsman, E.M., and Yu, D. (2004) Joint oligogenic segregation and linkage analysis using Bayesian Markov chain Monte Carlo methods. Molecular Biotechnology **28:** 205–226.
http://journals.humanapress.com/index.php?option=com_opbookdetails&task=articledetails
&category=humanajournals&article_code=MB:28:3:205

30. Wilcox, M.A., Pugh, E.W, Zhang, H., Zhong, X., Levinson, D.F., Kennedy, G.C., and Wijsman, E.M. (2005) Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated datasets for Genetic Analysis Workshop 14: Presentation Groups 1, 2 and 3. *Genetic Epidemiology* **29:** S7–S28.
http://www3.interscience.wiley.com/cgi-bin/fulltext/112204115/PDFSTART

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

31. Basu, S., Di, Y., and Thompson, E.A. (2007) Tests for linkage detection in pedigrees. *Genetic Epidemiology*: submitted.

32. Sung, Y., Di, Y., Fu, A.Q., Rothstein, J.H., Sieh, W., Tong, L., Thompson, E.A., and Wijsman, E.M. (2007) Comparison of multipoint linkage analyses for quantitative traits: parametric lod scores, variance component lod scores, and Bayes factors in the CEPH data. *Biomed Central Genetics*: Submitted.

33. Thompson, E.A. (2006) Uncertainty in inheritance: Assessing evidence for linkage. *Proceedings of the Third University of Washington Biostatistics Symposium: Nov. 2005.* Technical Report No. 498, Department of Statistics, University of Washington. (PDF file Attached.)

34. Thompson, E.A. (2007) Linkage Analysis. In *Handbook of Statistical Genetics 3 rd ed.* D.J.Balding, M. Bishop and C.Cannings (eds). Wiley: Chichester, UK. (Significantly revised and updated from 2 nd ed.) Submitted.

35. Thompson, E.A., and Geyer, C.J. (2007) Fuzzy p-values in latent variable problems. *Biometrika:* in press.

36. Tong, L., and Thompson, E.A. (2007) Multilocus lod scores in large pedigrees: Combination of exact and approximate calculations. *Human Heredity*: submitted.

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

37. Basu S. A likelihood-based trait-model-free approach for linkage detection WNAR meeting, Fairbanks, June 2005.

38. Rosenthal, E.A., and Wijsman, E.M. MCMC analysis of complex traits caused by multiallelic loci. American Journal of Human Genetics 73s:499, 2003.

39. Thompson, E.A., and Fu, A.Q. Linkage diseqilibrium in family-based genetic mapping. WNAR Meeting, Flagstaff, June 2006

(iv) Software resources

40. MORGAN: A package for Monte Carlo Genetic Analysis. Version 2.8.1 released 2006. Available at http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml

41. MORGAN Tutorial and examples. Available online and for download in multiple formats at http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml#tut