This Competitive Revision is submitted in response to Notice NOT-0D-09-058 under the title
*NIH Announces the Availability of Recovery Act Funds for Competitive Revision Applications*.

Under the parent grant, we have developed methods for the analysis of trait and marker data observed on members of pedigree structures, to aid in the identification of genes contributing to increased risk of complex traits, including cardiovascular, neurological and behavioral disorders. These methods include the development of Monte Carlo methods to infer gene descent in extended pedigrees, given data at a dense genome screen of markers, and the use of these gene descent patterns in joint multilocus linkage and segregation analysis of complex traits. The proposed new research direction addresses methods for the use of modern dense genomic marker data available on individuals. The methods we have employed with success to infer identical-by-descent (*ibd*) genome segments among pedigree members will be combined with new methods to infer *ibd* among individuals and families not known to be related. This approach has the potential to combine the linkage detection power of family-based studies with the resolution and sample-size advantages of population-based studies. We propose to develop methods to infer *ibd* of genome segments in the $\sim 10^6$ bp range, within and among members of populations using dense SNP marker data (e.g. 500K SNP chips). We will investigate the accuracy and power of these methods, and their robustness to population subdivision, stratification, and admixture. We will investigate the impact on *ibd* inferences of meiotic map heterogeneity (recombination hot-spots), linkage disequilibrium, and copy number variants. We will extend the methods to make use of within-pedigree *ibd* and resulting haplotype (phase) information, to detect *ibd* genome segments among members of different pedigrees. We will extend methods for genetic analyses of trait data conditional on the inferred joint pattern of *ibd* among observed pedigree members to accommodate also between-pedigree *ibd*. Methods will be evaluated primarily on data sets in which the *ibd* is simulated, but in which real-data chromosomes are used to found the simulated population. Additional testing and evaluation will be done using real data sets including collections of pedigrees segregating cardiovascular or behavioral disorders. These real data sets include several on which are available genome-wide SNP genotype screens or more localized multigene haplotypes. Finally, software will be developed that implements these methods within the MORGAN-3 software package of the parent grant. The software will be documented and released for use by practitioners.

# 1 Introduction to the Application

Modern dense genomic marker data (primarily SNP genotypes) are increasingly available on members of study populations. In family studies, not only are pedigree relationships among extant generations well established by these data, but in fact genome segments shared identical by descent (*ibd*) from recent common ancestors are well determined, and form the basis of our methods for the genetic epidemiology of complex traits. These methods increasingly focus on first using marker data to infer (probabilistically) the inheritance of genome from the pedigree founders to all observed members of the pedigree structure. Specifically, we obtain realizations of the *founder genome labels* (FGL) jointly at all marker locations conditional on all available marker data on the pedigree. These FGL are then used in a variety of methods to detect and map genes for complex traits. They enable lod score computations under multiple trait models, investigation of sensitivity of lod scores to a variety of trait models, analyses of a variety of allele-sharing statistics or variance-component analyses, all from a single set of marker-based FGL realizations. This has facilitated lod-score analyses of models with two linked trait loci (Sung et al. 2007) and has led to new methods of testing for, and resolving, multiple linked trait loci (Di and Thompson [3]; see Appendix).

Even in a small well-sampled pedigree there is often considerable FGL uncertainty. For example, the two diploid genomes of each founder are exchangeable, and if even the members of a founder couple are unobserved the gene descent from the couple to their descendants may be unclear. On the other hand, chromosomal segments of *ibd* among observed descendant relatives are often very well determined [3; Appendix]. Accordingly, we have developed a reduction of the FGL to *distinct genome labels* (DGL), which are arbitrarily labeled, designating only that individuals sharing a DGL at some genome location share that genome *ibd* at that location. In 2009-10 **under the parent grant** we will extend methods for computation of trait likelihoods based on realized FGL but that work directly on the reduced DGL graph. This will greatly enhance our ability to examine multiple trait models, and to use trait controlled by several loci, and/or by loci with several segregating alleles (haplotypes).

With modern dense genomic data, not only is *ibd* within small pedigrees clearly determined, but so also is *ibd* between pedigrees or population members (Purcell et al. 2007; Browning 2008). Where a study consists of a large number of small well-sampled pedigrees, either ascertained through cases and/or from a small population, remote relationships either among “founders” of a given pedigree, or between “founders” of different pedigrees are likely. Detection of resulting between-pedigree *ibd* at the population level could be a powerful signal for genetic linkage. The first theme of this revision application is to develop methods for the detection of *ibd* segments of genome between members of different pedigrees, using recently developed models (Thompson 2008b). The within pedigree information informs founder haplotypes, and enhances the between-pedigree inferences.

Where such between pedigree *ibd* is present, the DGL graphs of the separate pedigrees become connected, though this shared DGL. Trait model analyses using this combined DGL graph remain computationally feasible, and this approach provides a way to combine population and pedigree data. The second (year-2) goal of this revision is to extend within-pedigree trait analyses developed **under the parent grant** to permit use of within-pedigree and within-population *ibd* in a combined analysis of trait linkage signals.

Software under the parent grant is implemented in our MORGAN-3 software package, which is made freely available vie the web at http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml. Under this Competitive Revision, we will develop programs both for trait model analyses conditional on the DGL graph and for better inference of *ibd* between pedigrees and in populations, and these will be integrated into the MORGAN framework. Integrating our software under the parent and revision awards will greatly increase the range of MORGAN programs, allowing both projects to gain from the shared library, input and program structures, and hence increase the range of analyses of the genetic epidemiology of complex traits that may be accomplished using MORGAN.

# 2 Specific Aims

## 2.1 Summary of aims of the parent grant: 12/2007-11/2012

The overall objective is the development of techniques for the analysis of the genetic basis of complex familial traits, with the current focus toward linkage analysis using dense SNP markers. The aims include the development of Markov chain Monte Carlo (MCMC) methods for realization of descent patterns in pedigrees conditional on multilocus marker data, including
1) improved sampling and estimation methods based on joint realization at multiple genome locations;
2) new approaches to joint linkage and segregation analysis of complex genetic traits;
3) development of methods for assessment of statistical significance of linkage findings;
4) inference and use of gene descent patterns in pedigrees accommodating marker model uncertainty;
5) comparison of the new approaches with existing methodology through analysis of sample data sets;
6) development of publicly available software (MORGAN-3) for performing necessary computations.

## 2.2 Specific aim for the parent grant in 2009-2010

In addition to other work progressing under the parent grant, we will further develop methods that will support the revision application. Specifically, we will develop methods for computation of trait likelihoods on pedigree structures using inferred identity-by-descent (*ibd*) by:
(i) improvement of methods for generating multiple realizations of gene descent conditional on dense marker data, jointly over pedigree members and across the chromosome, and the output of realizations in compact format,
(ii) development of efficient algorithms to transform gene descent output to the *DGL graph*, which defines the pattern of *ibd* jointly among observed individuals at each genome location,
(iii) development of programs to compute trait data likelihoods directly on DGL graphs jointly at several genome locations, enabling analysis under more complex trait models, and
(iv) implementation of these methods in the MORGAN-3 software package.

## 2.3 Specific aims of the Competitive Revision: 12/2009-11/2011

1. We will develop and assess methods for detection of *ibd* genome segments among members of populations by:
(i) development of a population *ibd* simulation framework using real-data chromosomes as founder haplotypes
(ii) improvement and extension of preliminary models and methods for inference of *ibd*, including methods for parameter estimation and tuning
(iii) assessment of the effects of population subdivision and admixture, and of resulting allele frequency heterogeneity
(iv) assessment of the effects of recombination hotspots, linkage disequilibrium, copy-number variants, and data error.

2. We will develop and assess methods for detection of *ibd* genome segments among pedigree founders and among members of different pedigrees by
(i) expansion of the simulation framework of 1(i) to create data structures of small pedigrees within populations,
(ii) extension of the inference method of 1(ii) to incorporate pedigree-based inference of haplotypes and missing data into the inference of between-pedigree *ibd*,
(iii) comparison of joint and pairwise inference in the detection and estimation accuracy of *ibd* segments among members of different pedigrees,
(iv) assessment of the advantages of incorporating within-pedigree information in addressing issues of recombination hotspots, linkage disequilibrium, copy-number variants, and data error.

3. We will integrate the methods developed under 1. and 2. into parent-grant methods for the genetic analysis of complex traits by
(i) extension of the work under section 2.2 to construct among-pedigree DGL graphs,
(ii) application of methods on real-data large pedigrees, assessing the extent to which *ibd* inferred among current families can substitute for exact use of ancestral pedigree information,
(iii) application of methods on real-data collections of small pedigrees, assessing both the power to detect between-pedigree *ibd* and the effects of uncertainty in the inferred DGL graph on trait model analyses.

4. We will extend and enhance our user software, the MORGAN-3 package of the parent grant, by
(i) implementation of programs and methods developed under 1. and 2. within MORGAN-3
(ii) documentation, distribution and support of the software,
(iii) development of web-based tutorial materials and examples, as is done for other MORGAN-3 programs

# 3 Background and Significance

## 3.1 Designs for genetic 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, 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). Very large population-based designs are yielding results (Wellcome Trust Case Control Consortium 2007), and some of these results have been replicated (Ehret et al. 2008), but through these studies it is becoming recognized that the sample sizes required for adequate power are of the order of thousands. Population, environmental and genetic heterogeneity are major issues, to which family studies are far more robust (Borecki and Province 2008). Overall, there are several reasons to be cautious about replacing the family-based design with only a 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 do 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-cited associations 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, and hence 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 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 is a time-tested and powerful tool to localize such loci.

## 3.2 Pedigrees in populations

Pedigrees exist within populations, but pedigree-based approaches treat the founder members of such pedigrees as independent or “unrelated”. On the other hand, population-based methods often ignore or ”adjust for” the relationships among members of the study population. One approach to combining within-family inheritance with between-family associations is that of family-based association tests (FBAT), which provide a valid test of association in the presence of linkage that is robust to population stratification (Laird et al. 2000). The first and well known such test is the transmission disequilibrium test or TDT (Spielman et al. 1993). Originally a test using the information on transmission of marker alleles from parents to case offspring, it has been extended to alternate pedigree structures (Spielman and Ewens 1998) and to quantitative traits (Allison 1997). Because the within-family inheritance information is orthogonal to the between-family associations, FBAT and TDT tests can be used to follow up on linkage signals from family-based studies, with the goals of achieving greater mapping resolution and/or gene identification.

More generally, it is well recognized that that between-pedigree associations can be used to provide greater resolution to pedigree-based linkage results, and that mutations shared by members of different pedigrees are powerful confirmatory evidence of gene identification. Often pedigrees are studied or grouped on the basis of ethnic origin with the goal of increasing trait homogeneity. The sharing of an associated STR allele and ultimately of the identified PS2 mutation for early-onset Alzheimer disease among a subset of pedigrees of Volga German ethnic origin (Levy-Lahad et al. 1995) were instrumental in resolving that trait. The sharing of a diverse collection of BRCA2 mutations among different subsets of Ashkenazi Jewish families segregating early-onset breast cancer has aided in the assessment of cancer risks (Fodor et al. 1998; Neuhausen et al. 1996). In studies of type-2 Diabetes under the GENNID study (Edwards et al. 2008), the subdivision of the pedigrees into three ethnic groups clarifies but does not resolve a very complex pattern. Some linkage signals are shared among groups, or by a subset of pedigrees within a group, but others are not, and subdividing the sample reduces power. Last but not least, the entire Finnish population has offered a basis for resolving genes contributing to traits phenotypes, using information both within and among pedigrees. The relative genetic homogeneity and presence of founder effects, together with extensive demographic, historical, and pedigree data, have led to successes in genetic analyses of complex traits (Peltonen et al. 1995).

In pedigrees, the pedigree structure provides a prior distribution on identity by descent (*ibd*), but for members of different pedigrees any remote ancestral relationships are unknown. Subdividing a set of pedigrees by ethnic origins or phenotypic characteristics can create difficulties. Even if a pedigree can be designated, say, as Hispanic, this designation may not apply to all individuals within it, and is unlikely to apply to the entire genomes of any individuals within it. As dense genomic marker data become increasingly available, *ibd* of localized genome segments can be accurately inferred among observed individuals, not only within pedigrees but also among individuals not *a priori* known to be related. Thus, without any pre-imposed subsets, information from different pedigrees and specific to particular genome regions can be combined. This will enable weak linkage signals from multiple small families to be strengthened through inferred *ibd* at specific genome locations, and will address the problems of allelic and locus heterogeneity that make so difficult the replication and validation of association study findings. Sharing even between two separate pedigrees within a study will provide additional support for linkage, for a shared variant impacting the trait of interest, and perhaps significant additional resolution.

### 3.3 Pedigrees of populations

While many pedigree-based studies are challenged by the power and resolution limitations of small pedigrees (Boehnke 1994), the huge pedigrees of genetic isolates present other challenges: (1) On extended multi-generation pedigrees with several generations of missing genetic data, explicit use of the presumed pedigree in analyses is a huge computational burden particularly for modern multi-marker SNP data (Albers et al. 2008). (2) Even with large pedigrees, the information in terms of the number of informative meioses is often not large. (3) Due to the missing genetic data, it not possible to validate the pedigree; there are potential errors and biases in the recorded structures, and the possibility of unknown relationships and relevant coancestry among founder members of the pedigree. (4) Such pedigrees are very costly to collect, and there are always privacy and confidentiality concerns caused simply by the specific structures recorded.

On the other hand, there is much information potentially in the remote relationships of extended pedigrees in connection with resolving the genetic basis of complex traits. (Thomas et al. 2008). Such distant cousins have a much lesser degree of shared environment than do close relatives: such shared environment decreases both the sensitivity and specificity of findings in family studies. As importantly, these distant cousins have a greater degree of both allelic and locus homogeneity with regard to the genetic basis of their trait values than do completely “unrelated” individuals. Breaking up the pedigree is sometimes necessary, as in the analysis of a 4645-member Dutch pedigree split into 35 smaller ones (Liu et al. 2007), but ancestral connections between parts of the population pedigree can be critical in obtaining linkage signals, as shown in an analysis of asthma data on the Hutterite population made available for Genetic Analysis Workshop 13 (Chapman et al. 2001). Thompson (1981), in a study of a Newfoundland isolate, found that inferences were robust to most details of the ancestral pedigree, but showed also there were certain ancestral links whose validity was essential to the conclusions. Analyses of neurological disorders in a very complex 10-generation pedigree that is a part of the population of Guam (Poorkaj et al. 2001), even with STR markers, are challenging our best computational methods (see [6]: publications list). Currently, these methods could not address SNP data on this pedigree.

With the advent of dense SNP genotyping assays, there have been approaches on how best to use such data for genetic analyses of complex traits in the context of very large pedigrees. Albers et al. (2008) use an approximation to the descent of genome down ancestral lineages, resulting in a Markov approximation across the genome for the *ibd* segments of genome shared by descendant individuals, while Thomas et al. (2008) use identity-by-state, together with the pedigree structure, to identify these segments from dense SNP marker data for use in subsequent trait analyses. Both these approaches use the pedigree structure as a prior model for *ibd* among descendant individuals affected by some trait. Our proposal would be to ignore (even if available) any pre-imposed remote ancestral pedigree structure, evading immediately the issues of pedigree accuracy, some part of confidentiality concerns arising for pedigree data, and the huge computational burden or even infeasibility of using the full details of an extended pedigree with many generations of missing data. Instead we would use the small pedigrees of observed individuals in the extant population. We would use our methods applied to modern dense genomic data possible to identify localized *ibd* genome segments among these current pedigrees as in section 3.2 above.

## 3.4 Inference of identity by descent

Identity by descent (*ibd*) underlies all methods of gene mapping. Shared descent of genes within pedigree structures leads to individuals sharing a phenotype having increased probability of *ibd* for DNA affecting the trait values, and hence increased probability of marker similarity for nearby markers. Precisely the same phenomenon, albeit at scales of less than $10^5$ base pairs (bp) of DNA rather than several million, underlies the allelic associations (linkage disequilibrium: LD) that are the basis of population-based approaches to gene mapping. The key premise of this proposal is that, from modern dense genomic marker data we can infer *ibd*, albeit imperfectly, among the chromosomes of observed individuals in different pedigrees. This will enable us to combine information from relationships within pedigrees with population-based strengths of coancestry among pedigrees, in the detection and mapping of DNA affecting a complex trait of interest. We thus provide a framemwork in which pedigree-based and population-based approaches to mapping and identifying genes are complementary and not competing.

With the advent of dense SNP genotyping arrays is coming the recognition by numerous authors that coancestry of genome segments can be inferred without exact ancestral pedigree information (Purcell et al. 2007; Thomas et al. 2008; Browning 2008; Leibon et al. 2008) and there is a range of approaches. At one extreme, the methods of Thomas et al. (2008) and Albers et al. (2008) use the ancestral pedigree information to inform the inferences. At the other, the approach of Leibon et al. (2008) is based on identity-by-state (*ibs*) and requires a summary of pedigree information in terms of numbers of non-founder ancestors only in assessing significance of a detected shared segment. The *ibs* approach also does not use genetic map information nor even variation in SNP allele frequencies among markers, and therefore cannot gain from these genetic and population parameters, nor from well-established models for genetic data and patterns of *ibd*.

The approaches of Purcell et al. (2007), Browning (2008), and ourselves (Thompson 2008b) may be regarded as intermediate. They do not require a known pedigree, but they do directly model the *ibd* outcomes that might arise from the ancestral pedigree of a population. The power of this approach arises from the fact that *ibd* segments shared by remote relatives are few rather than small (Donnelly 1983). That is, the probability of *ibd* is small, but conditional on *ibd* the expected lengths of genome shared *ibd* extend over many megabases. For example, consider a pair of relatives separated by $m = 20$ meioses, sharing perhaps one diploid common ancestor 10 generations ago. The probability that these individuals share genome *ibd* at any specific locus is $2 \times (1/2)^m \approx 2 \times 10^{-6}$. The probability they share any genome at all *ibd* from this ancestor is approximately $(1 - \exp(-(m-1)L/2^{m-1})) \approx 10^{-3}$ for a human genome of total length $L \approx 30$ Morgans (Donnelly 1983). On the other hand, conditional on the existence of a genome segment shared *ibd*, its length is of order $m^{-1}$ Morgans or approximately 5 cM ($5 \times 10^6$ bp). By comparison, the extent of allelic association (LD) is at most $5 \times 10^5$ bp and is often much less.

There are several advantages to approaches which model underlying *ibd* among the genomes of a population, and the probabilities of haplotypic (phased) or genotypic (unphased) genetic marker data given that latent *ibd*. They allow for the heterogeneity of SNP locations and SNP informativeness. In both respects, SNPs are far from evenly distributed. Of course, there are approximations inherent in such models. *ibd* segments result from recombination events in the ancestry of individuals, and hence are dependent on genetic (meiotic) distance, while dense SNPs are known only by physical base-pair (bp) positions. There are uncertainties in SNP allele frequencies and haplotype frequencies. While allele frequencies can be well estimated from population samples, haplotype frequencies are a function of LD and large samples are required for accurate estimation (Browning 2008).

On a less dense genomic scale, Leutenegger et al. (2003) produced the first model to infer *ibd* among chromosomes in populations from genetic marker data at multiple linked loci. Although she considered only the two chromosomes within each individual, a key feature of her model is that it permits error in the observations. Browning (2008) considered dense data on a genomic scale. She again only considered pairs of chromosomes, and her model did not allow for error, but her key contribution was the incorporation of LD among the dense genetic markers upon which inferences are to be based. Purcell et al. (2007) considered estimation of *ibd* between two individuals, given dense genotypic data. However, this model also does not allow for data error or LD, and considers only *ibd* between the individuals and not between the two genomes of each. Since, in modern human populations, our parents are at least as closely related to each other as is each of us to other members of a study population, this seems an undesirable constraint.

The model we propose is based on that of Leutenegger et al. (2003). It is extended to multiple genomes (Thompson 2008b), since even to consider the genotypes of a pair of diploid individuals, we must consider their four genomes. It allows for error, which we regard as of key importance. Without allowance for error, a single discrepant marker allele may destroy the inference of an entire *ibd* segment (Leutenegger et al. 2003). We do not propose to include LD in our analysis models. Although the model could be extended to include LD (Thompson 2008a), LD modeling is complex and requires large samples for accurate estimation (Browning 2008). Also, since LD is a reflection of the coancestry we aim to infer (although at a longer time frame), it is questionable as to whether LD should be modeled.

### 3.5 Inferred *ibd* in the analysis of complex traits

The approach of first using marker data to compute *ibd* probabilities, and then modeling the trait data conditionally on these probabilities is of long standing, used in early sib-pair linkage detection approaches for quantitative traits by Haseman and Elston (1972) and for qualitative traits by Suarez et al. (1978). More recently, it has become the basis of allele-sharing methods (Whittemore and Halpern 1994; Kruglyak et al. 1996; Kong and Cox 1997; McPeek 1999) and is the the foundation of variance-component methods for quantitative traits (Amos 1994). Indeed, the variance-component approach implemented in SOLAR (Almasy and Blangero 1998) is very widely used and provides a benchmark against which new methods for quantitative trait analysis must be judged.

The above methods all use probabilities of *ibd* estimated from marker data. Recently, a different approach to deal with uncertainty in *ibd* has been proposed. The statistics of Thompson and Basu (2003), Thompson and Geyer (2007), and Di and Thompson [3; Appendix] are all direct functions of the *ibd* itself, and significance is assessed by considering the probability distribution of p-values over the probability distribution of *ibd* states conditional on marker data. When *ibd* is fully determined by the marker data, there is no difference between these approaches, but where there is *ibd* uncertainty the ability to separate this marker-based uncertainty from the uncertainty in trait inferences provides additional information in analyses of complex traits (Thompson and Geyer 2007). It also makes easier the development of linkage test statistics that use information from all observed pedigree members (not only affected), can be used for both qualitative and quantitative data on general pedigrees, and make use of higher-order (not only pairwise) *ibd* sharing [3; Appendix].

With the advent of dense SNP data on pedigree members, methods that use marker-based estimates of *ibd* state in the subsequent analyses of complex trait data have clear advantages. Marker data is analysed only once, and the resulting *ibd* estimates can be used in analyses of many traits or for many trait models. Additionally, within well sampled pedigrees *ibd* uncertainty will be small, and, of particular relevance for this proposal, *ibd* can be inferred not only within-pedigrees but also between. The use of test statistics that are functions of *ibd* state rather than of *ibd* probabilities, while not essential, makes straightforward the combination of population-based *ibd* and within-pedigree *ibd* to provide an integrated pedigree and population-based approach to genetic analysis of complex traits.

# 4 Progress Report and Preliminary Studies

In section 4.1, we report progress on the parent grant that is related to the research proposed in the Competitive Revision, while in section 4.2 we describe preliminary studies that provide the theoretical support and practical evidence that the proposed approach is viable and practical. Numbers refer to publications resulting from the parent award and listed in the Progress Report Publications (Section 7), while other references are included in the Bibliography. Items [3] and [9] are included as Appendix materials.

## 4.1 Progress on the parent grant

### 4.1.1 Improvements in computational and sampling methods

We have continued improvement of our MCMC methods for the sampling of gene descent in large pedigrees given dense marker data [10], extending our meiosis sampler to permit joint updates of multiple meioses. The same algorithm permits exact computation and independent realization of inheritance patterns conditional on marker data on small pedigrees. The two approaches have been combined to provide optimal performance on large data sets including both small and large pedigrees. These improvements give accurate lod score estimates an order of magnitude faster than our previous MCMC programs, as indicated by our analyses for Genetic Analysis Workshop 16 using dense SNP markers [5].

### 4.1.2 Use of realized gene descent for linkage detection and localization

The approach of realizing gene descent from marker data leads to a flexible framework for analysis of discrete or quantitative complex genetic traits [8,9]. We have developed trait-model-robust procedures which test for associations between marker inheritance and trait similarities or differences. We have modeled these associations [2], and developed new test statistics for linkage detection [3]. We have used the conditional independence of inheritance at non-adjacent markers to develop new test procedures for localizing multiple linked QTL [3].

### 4.1.3 Assessment of the significance of linkage signals

We have worked on two new approaches to the assessment of significance: both involve resampling methods conditional on the fixed observed marker data. In one approach, trait data are resimulated conditional on marker data, under multiple genetic models; preliminary analyses are promising [4]. The second approach involves multiple realizations of inheritance patterns jointly across loci conditional on marker data. Trait-related test statistics of trait-data likelihoods are then computed conditionally on each realized inheritance pattern, to obtain an empirical distribution of lod-score contributions or latent p-values (Thompson and Geyer 2007). This distribution provides not only an assessment of the strength of the evidence, but also of the uncertainty in that assessment. This approach has been compared with permutation and other resimulation approaches for discrete traits [1], used to assess localization information for quantitative traits [3], and has been further developed to provide valid confidence sets for trait locus localization [8].

### 4.1.4 Software development

Software under the parent grant is implemented in the software package MORGAN, for which source code is made freely available. In 2008, software development was three-fold.
(1) New programs and code: In MORGAN 2.8.3 and 2.9 we have implemented new samplers [10] and exact computation of lod scores on small pedigree components. MORGAN 2.9 also includes the new program lm_twoqtl, which performs lod score computations for two linked QTL (Sung et al. 2007). MORGAN 2.9 has involved extensive clean-up of code at all levels, and testing on multiple operating systems.
(2) The MORGAN tutorial, documentation and example files have been updated and released to the web. The new example files run under MORGAN 2.9.

(3) Software development proceeds under the new MORGAN-3 framework in which trait loci are distinguished from traits, permitting a far greater range of complex trait models. A limited beta-test version of MORGAN-3 was released in 2008.

In March 2009 a fully functional (although still beta-test) version of MORGAN-3 was released, including all the programs of MORGAN 2.9 and two additional programs. Programs for more effective sampling of within-pedigree inheritance patterns conditional on dense SNP marker data on individuals have been implemented. A program to output multiple realizations of inheritance patterns across a chromosome in a compact format has been included in the new MORGAN-3 beta-test release. Under the parent grant, we will continue to develop these methods and implement these programs in MORGAN-3. The website for download of our software is *http//:www.stat.washington.edu/thompson/Genepi/pangaea.shtml*. A complete release history and other software information may also be found at this site.

## 4.2 Preliminary studies

### 4.2.1 A model for *ibd* among multiple genomes in a population

In this section we present models for inference of *ibd* among multiple genomes sampled from a population, given either haplotypic (phased) or genotypic (unphased) data on a set of individuals. For convenience we will refer to the underlying DNA as genomes, and to the phased genotypes as haplotypes, even when considering only *ibd* or genetic marker data at a single genome location. More details are given in [9], which is included as Appendix material.

Leutenegger et al. (2003) first considered the inference of *ibd* between the two genomes of an individual, given genome-wide genotypic data. The *ibd*-model of that paper is two-parameter Markov model for changes between non-*ibd* and *ibd* along a chromosome; *ibd* is gained at rate $g$ and lost at rate $h$ giving Markov rate matrix

$$Q \;=\; \begin{pmatrix} -g & g \\ h & -h \end{pmatrix} \;=\; \begin{pmatrix} -\alpha\beta & \alpha\beta \\ \alpha(1-\beta) & -\alpha(1-\beta) \end{pmatrix} \;=\; \alpha(-I \;+\; \begin{pmatrix} 1 \\ 1 \end{pmatrix}(1-\beta,\beta)). \tag{1}$$

The parameter $\beta = h/(h+g)$ is the marginal *ibd* probability, and $\alpha$ measures expected lengths of *ibd* segments. The relative rate of gain and loss of *ibd* is $g/h = \beta/(1-\beta)$ [7]. The final expression of equation (1) shows that the model may also be interpreted as one of independent segments independently non-*ibd* (probability $(1-\beta)$) and *ibd* (probability $\beta$), with recombination breaks occurring at rate $\alpha$.

The model of Leutenegger et al. (2003) for the genotypic data given latent *ibd* is that, in principle, *ibd* implies the same allelic type and non-*ibd* implies independent allelic types. However a small “error” probability $\epsilon$ (of order 0.01) allows alleles scored as of different types to be *ibd*. An advantage of this simple model is that it is easily extended to the joint probability of a larger set of genotype. At each locus, allele frequencies of alleles are assumed known: in reality they can be well estimated from genotypic samples. Hardy-Weinberg proportions (HWE) are implicit in the assumption of independent allelic types of non-*ibd* alleles. As a population model, we prefer HWE to the model of homozygote deficiency and heterozygote excess implied by the finite sample approach of Purcell et al. (2007). Any deviation from overall HWE due to population stratification and heterogeneity will be in the direction of homozygote excess, not homozygote deficiency.

The *ibd* model (equation (1)) and data model define a Hidden Markov Model (HMM) for latent *ibd* and observed genotypes. The standard forward-backward algorithm (Baum et al. 1970) provides the conditional probability of *ibd* at every location on the chromosome, given the allelic types on the chromosomes jointly over all loci.

We now summarize the extension of this model to multiple genomes. At a single locus, Balding and Nichols (1994) and Weir (2001) have modeled multigene *ibd* probabilities using Ewens' sampling formula

(Ewens 1972). In a sample of size $n$, let $a_i$ be the number of allelic types present in $i$ copies. Then the number of distinct alleles is $k = \sum a_i$, $n = \sum ia_i$, and Ewens' sampling formula may be written in terms of the population coancestry $\beta$ as

$$\boldsymbol{\pi}_n(a_1, ..., a_n) \quad = \quad \frac{n!\beta^{n-k}(1-\beta)^{k-1}}{(1+\beta)(1+2\beta)....(1+(n-2)\beta)} \prod_{j=1}^{n} (j^{a_j} a_j!)^{-1} \tag{2}$$

In inference of *ibd* we will normally wish to consider more than a pair of chromosomes. Minimally, we may wish to consider the four genomes of a pair of individuals, and additionally may have only genotypic data, without phase. In preliminary work, Thompson [7] extended the *ibd* model of equation (1) to the case of four chromosomes in such a way that the marginal model (2) is retained, and in [9] has extended this to multiple (arbitrary $n$) genomes. Specifically, two types of transitions $A$ and $B$ among *ibd* states are modeled: ($A$) All pairs of singletons in states with $a_1 \geq 2$ become *ibd* at rate $g$. All doubletons in states with $a_2 \geq 1$ become non-*ibd* at rate $h$. ($B$) Each singleton in a state with $a_1 \geq 1$ joins with an *ibd* group size $(j-1) \geq 2$ at rate $(j-1)g$. Thompson [9] shows that under the transitions ($A$) and ($B$) detailed balance with respect to $\boldsymbol{\pi}_n$ of equation (2) is maintained and hence $\boldsymbol{\pi}_n$ is the unique equilibrium distribution.

The model for $n$ chromosomes relates directly to haplotypic data. It could be used for genotypic data simply by using the appropriate genotype probabilities for each *ibd* state, but this would be computationally inefficient. It was noted in [7] that the 15-state Markov model for 4 genomes reduces to a 9-state Markov model for the nine genotypically equivalent classes of states (Jacquard 1972) and the reduced model can be used directly for genotypic data on a pair of individuals. Thompson [9] has shown that the same reduction holds true for any number of individuals. That is, the reduction to genotypically equivalent classes of *ibd* states remains a Markov process on the smaller space.

Unfortunately, for real data on individuals among whom there is substantial *ibd*, this model may be insufficient, due to the presence of shared ancestral recombination events [9]. To adjust for this, we have proposed a generalization using the second form of equation (1). If $Q^*$ denotes the Markov rate matrix under the transitions ($A$) and ($B$) then we adjust $Q^*$ to become:

$$Q^\dagger \quad = \quad (1-\delta)Q^* + \delta(-I + \mathbf{1}\boldsymbol{\pi}_n'). \tag{3}$$

Under $Q^\dagger$ there are two kinds of breakpoints. In the first, occurring at rate $(1-\delta)$, transitions follow the previous rate matrix $Q^*$, and in the second, at rate $\delta$, the new state is chosen from $\boldsymbol{\pi}_n$ independently of the current state. In the model (3) all transitions are possible. Additionally, the matrix $Q^\dagger$ has the same equilibrium probabilities as $Q^*$, namely $\boldsymbol{\pi}_n$ of equation (2). Additionally, $Q^\dagger$ has the same characteristic as $Q^*$ of retaining the Markov property in reducing to genotypically equivalent state classes [9].

Our model for data at a locus given the underlying *ibd* follows that of Leutenegger et al. (2003). In principle, *ibd* haplotypes must have the same allelic type, and non-*ibd* ones are of independent types. As before, this is modified to allow for error by mixing this idealized distribution with a proportion $\epsilon$ of the distribution where all haplotypes are of independent allelic types. Since we do not here model LD, only the single-locus allele frequencies need be specified. The model for genotypes is simply given by reducing to the unordered pair of haplotypes within each individual. As before, we have an HMM. The standard forward-backward algorithm (Baum et al. 1970) provides the conditional probability of *ibd* at every location on the chromosome, given the allelic types on the chromosomes jointly over all loci.

In summary, a model for *ibd* among four haplotypes, and hence applicable to the genotypes of a pair of individuals, was presented in [7]. This has been extended to an arbitrary number of haplotypes [9; included as Appendix material]. The model may be applied to either haplotypic or genotypic data, and allows for marker typing error. This framework thus provides for the combination of pedigree and population data in the inference of *ibd* from dense SNP marker data, in situations where phase may be unknown or only partially inferred from the pedigree data.

### 4.2.2 A preliminary application of the model

We have undertaken some preliminary analyses of simulated data using the above model [9]. We constructed data chromosomes, using phased HapMap SNP data from Chromosome 19 YRI (African) individuals (International Hapmap Consortium 2005). Markers with minor allele frequency less than 6% are eliminated, leaving markers at any average density of 1 per $10^4$ base pairs. A pool of 60 artificial chromosomes of $10^8$ bp (10,000 SNP markers) are then created, retaining the SNP marker physical locations, allele frequencies, and LD patterns of the original HapMap data. Patterns of *ibd* among sets of chromosomes are then artificially constructed. Finally errors are introduced, at a rate of 1%, switching the SNP allelic type independently in each chromosome and at each marker.

The analysis presented in [9] is of just one set on 4 chromosomes, with a complex pattern of 200 *ibd* segments, encompassing all 15 states of *ibd* among them, with a change in the *ibd* pattern every 0.5 Mbp (on average, every 50 SNP markers). The analysis was run using several different choices of parameter values, but the results shown [9] were for an error rate 0.01, $\delta = 0.2$, $\beta = 0.3$. The absolute value of $h + g$ was determined as a function of $\beta$ and $\delta$, in such a way that the mean length of chromosome in an *ibd* state is 0.5 Mbp under the model. Clearly, these values are chosen to reflect the artificially constructed data; the chromosomes of actual populations would have values of $\beta$ and $\delta$ an order of magnitude smaller. The empirical SNP allele frequencies of the HapMap chromosomes were used, and no distinction was made between physical and genetic distance. Additionally, the six pairwise summaries of *ibd* among the four haplotypes from the joint analysis were compared with six pairwise analyses.

Using the haplotypic data on the four chromosomes, the *ibd* pattern was generally well inferred. Along most of the chromosome, most probabilities are close to 1 for some state, and to 0 for the remainder. There are some intermediate probabilities, and it is of interest that these seem to occur in regions of high LD in the HapMap chromosomes [9]. With the parameter values chosen, the regions of high *ibd* are well estimated, but generally *ibd* is over-estimated with some segments of no *ibd* among the four chromosomes missed entirely in the reconstruction. The general problem of parameter tuning or estimation remains to be addressed. For this example, there was no clear difference in the accuracy of pairwise *ibd* inferred from joint and from pairwise analyses. However, it appears that uncertainty is better calibrated by the joint analyses. The pairwise analyses had *ibd* probabilities close to 0 or 1, even where incorrect. Under our model, joint analysis of four chromosomes using 10,000 SNP markers, including summarizing the output both jointly and as six pairwise analyses, takes only 2 CPU seconds on a small laptop. Although genome-wide analyses of multiple chromosome sets will take proportionately longer, this approach is computationally practical.

Also shown in [9] is an analysis of the same four haplotypes paired into two genotypes, and reanalyzed as genotypic data on two individuals. As would be expected, with only genotypic data there is generally less certainty indicated by many more intermediate probabilities of *ibd* states. The *ibd* pattern no longer as well inferred, but the inference of more as compared to less *ibd* is mostly correct. Real haplotypes from real populations will have much less complex *ibd* patterns than those of our example, but the importance of phase information is clear. In the context of inferring *ibd* among pedigrees sampled from a population, the pedigree data will provide at least partial phase information.

Our *ibd*-inference programs have also been applied to real SNP genotypes arising from the Framingham Study supplied for Genetic Analysis Workshop 16 [5]. Here our analyses used only 2132 SNP markers on Chr-7, but the analysis required extension of the program to deal with missing SNP genotype data. Large (30 cM) segments of *ibd* were detected between individuals not specified as related, but later confirmed as sibs. This validates our approach and its application to true genotypic data, although the *ibd* segments we aim to detect in our studies are at least an order of magnitude smaller.

### 4.2.3 From inheritance to *ibd* in modeling complex traits

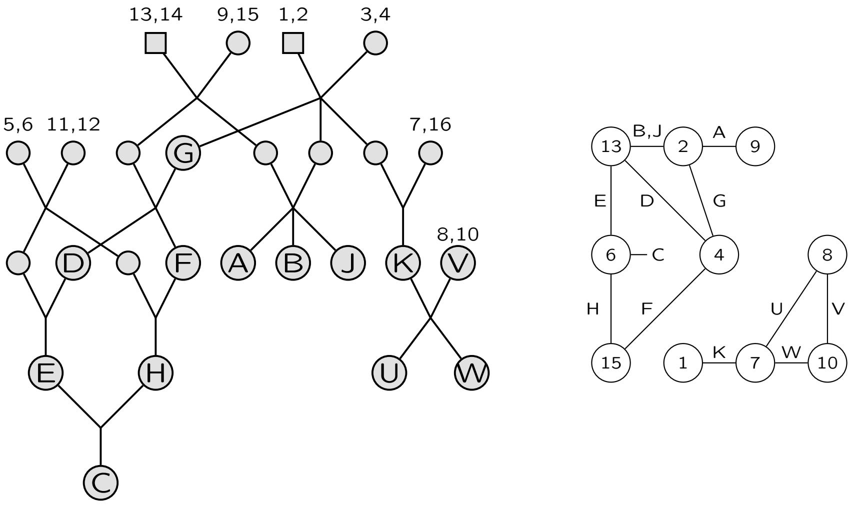
This section describes work which has been initiated **under the parent grant**, and will continue to be developed **under the parent grant** in 2009-10. Hence, in the second year of the Competitive Revision, the methods and programs will be available for the proposed extension from within-pedigree *ibd*-based analyses of complex traits, to make use also of *ibd* inferred between pedigrees where relationship are not *a priori* known to exist.

The foundation of phenotypic similarity among relatives is identity-by-descent (*ibd*) resulting from descent of DNA through repeated transmissions from the same DNA in a common ancestor. Relatives are similar because of *ibd*, in that *ibd* DNA is of the same allelic type $\mathcal{A}$ with high probability, while non-*ibd* segments of DNA are of close to independent allelic types $\mathcal{A}$. Within a pedigree, *ibd* is measured relative to the founders, and it is normally assumed that the alleles carried by the founders are of independent types.

At any given locus $j$, inheritance within a pedigree is most easily specified using the *inheritance vector* $S_j$ (Lander and Green 1987). Clearly, within a pedigree, $S_j$ determines *ibd*$(j)$, the pattern of *ibd* among pedigree members at locus $j$. Closer to the phenotypic data are the genotypes $G_j$, the types of the two alleles locus $j$ carried by each individual. Denoting by $\mathcal{A}_j$ the allelic types at locus $j$ of the founder genomes entering the pedigree, we see that $S_j$ and $\mathcal{A}_j$ together determine the genotypes of all individuals at locus $j$.

The latent structure underlying phenotypic data on pedigrees may be thought of in terms of genotypes $G_j$, inheritance vectors $S_j$, or identity by descent *ibd*$(j)$. There is no single correct framework, and indeed the most computationally effective framework has changed with the type of genetic data available. Traditionally (Elston and Stewart 1971), genotypes were considered. For multiple genetic markers to be considered jointly, the approach of inheritance vectors becomes the method of choice, since the dependence structure of the data is then that of an HMM (Lander and Green 1987). This is the basis of most current computational approaches for genome-wide linkage analyses, whether using exact computation on small pedigrees (Kruglyak et al. 1996; Abecasis et al. 2002) or Monte Carlo (Skrivanek et al. 1999; Sobel and Lange 1996; Thompson 2000).

The first direct use of within-pedigree *ibd* to compute probabilities of observed data was for marker locus genotypes observed without error. Sobel and Lange (1996) and Kruglyak et al. (1996) introduced what we will call the *distinct-genome-label* (DGL) graph in order to compute the probability of data at a marker locus, conditional on the inheritance vector. The figure shows an example pedigree (left) with its associated DGL graph (right). The labeled individuals (larger icons) are assumed to be observed. Given the inheritance $S_j$ at a marker locus $j$, and a set of distinct labels (the DGL) assigned to each of the two genomes of each founder, the DGL descending to each individual of the pedigree are known.



The DGL graph construction is most easily explained through this example. The nodes of the DGL graph are the DGL present in any observed individual. An edge, labelled by the individual, connects the pair of DGL present in each observed individual. For example, individual $A$ receives DGL “2” and “9”. The DGL graph shown here is consistent with Mendelian inheritance and the specified labeling of founder genomes and observed individuals. In this example, it is supposed that, for the inheritance vector $S_j$ at locus $j$, individual C received two copies of the founder genome labeled “6”, and the notation for $C$ is a single short line subtended from DGL “6”. All other observed individuals have two distinct DGL. For marker data, there are 2, 1 or 0 assignments of allelic types to the DGL that are consistent with observed marker data at a locus, and rapid and efficient computation of marker-data probabilities given the inheritance vector $S_j$ follows (Sobel and Lange 1996). In our example, we have labeled the founder genomes for easier explanation of the DGL graph. In any computation, the founder origins of the DGL are irrelvant. Only the *ibd* that results from shared DGL affects the probabilities of data on the pedigree.

Typically, other methods have been used to compute the probabilities of *trait* data given latent inheritance vectors $S_j$ at flanking marker loci. (Abecasis et al. 2002; Sobel and Lange 1996). These methods efficiently reduce the space $S_j$ using pedigree symmetries, but the data-dependent reductions implicit in the DGL graph are far greater. Hence, as was noted by Simon Heath, in many cases direct computation on the DGL graph at an hypothesized trait locus is a simpler approach (Thompson and Heath 1999). In fact, in 1997, a routine *gen_pen()* to compute trait-data probabilities using the DGL graph was implemented by Heath in our MORGAN software. Computation proceeds through the DGL graph, accumulating the overall probability of observed quantitative or qualitative phenotypes, just as on any conditional independence graph (Lauritzen 1992). An example is given by Thompson (2005). The fact that disjoint DGL components are independent, and that only DGL present in observed individuals need be considered, means that DGL graphs tend to be small. Hence computation on the DGL graph for an hypothesized trait locus is often much easier than computing on the pedigree structure, where summation over unobserved trait-locus genotypes is required.

As one moves along a chromosome, DGL graphs change due to ancestral recombination events. For example, a recombination in the meiosis to individual $K$ may result in $K$ carrying DGL 4 instead of 1, or a recombination in the meiosis to $J$ may result in $J$ carrying 3 instead of 2. In the first instance, *ibd* among pedigree members is gained and the number of DGL nodes is reduced: Node “1” is no longer present. In the latter case, *ibd* is lost and sibs $B$ and $J$ no longer share both their alleles *ibd*: Node “3” is added. However, changes are typically small, so that parts of the computation of trait-data likelihoods can be re-used, and even trait models dependent on the DGL graphs at several genetic loci are computationally feasible.

For simplicity, we have here presented our approach as though there were a single DGL graph fully determined by the marker data. In practice, there will be uncertainty, although where *ibd* is well determined by marker data, there may be a single DGL graph with a probability close to 1. More generally, our MCMC methods produce a set of realizations of $S_j$ conditional on marker data, and the DGL graph at a locus $j$ is a function of the inheritance vector $S_j$. Thus, we will have a sample of DGL graphs realized conditional on the marker data. In MCMC-based estimates of lod scores (Lange and Sobel 1991) trait-data likelihoods are computed for each realized $S_j$ and averaged. Exactly the same applies here for the realized sample of DGL.

As our methods **under the parent grant**, move toward inference of *ibd* from dense genomic data, and the subsequent analyses of trait data under complex trait models conditionally on this *ibd*, it becomes computationally effective to implement these analyses within the DGL framework. Where several (linked or unlinked) loci are to be considered as potentially underlying a trait, often only the DGL approach is computationally feasible. In 2009-2010, we will generalize algorithms for DGL-graph trait-model computations within the MORGAN software framework. This is not only important for our current pedigree-based analyses, but is an essential preliminary for the combination of within-pedigree and between-pedigree approach to *ibd*-based trait-model analyses proposed for Year-2 of this Competitive Revision.

# 5 Research Design and Methods

In this section, all references both from the parent award and other literature are included in the Bibliography.

## 5.1 Inference of coancestry (*ibd*) in populations

### 5.1.1 Simulation studies

We propose to develop further the model of section 4.2.1 for the inference of *ibd* among multiple genomes, and to assess its performance. Since our approach to performance assessment will rely heavily on simulation studies, we describe first our basic approach to construction of simulated data haplotypes with known patterns of *ibd* among them. The basic approach will then be generalized to accommodate more complex realities of genetic and population heterogeneity, as described in the following sections.

We propose first to generate the recombination breakpoints in a population of $\sim 20,000$ chromosomes of length $2 \times 10^8$ bp segregating over 200 generations. The original founder genome label (1 through 20,000) and the segment breakpoints in descendant generations will be stored in the same compact format used by our new MORGAN-3 gl_auto program. At each generation, the 20,000 chromosomes are paired into 10,000 “individuals”, and two offspring chromosomes are independently generated. Recombination breakpoints for each offspring chromosome are generated randomly, at a rate of 1 per $10^8$ bp (corresponding on average to approximately 0.01 per centiMorgan), and one of the two resulting gametes is randomly selected as the offspring chromosome. Hence, we will have, due to the constancy of offspring number, a collection of chromosomes at up to 200 generation depth of coancestry, in a random-mating population of effective size greater than $10^4$.

Data haplotypes will then be imposed on the chromosomes, using publicly available HapMap data (http://www.hapmap.org/downloads/). Our goal here is the creation of founder chromosomes that have true SNP allele frequencies, SNP densities, and patterns of linkage disequilibrium (LD). However, there are not 20,000 HapMap chromosomes available. Therefore, given for example 120 phased HapMap chromosomes available, we use the procedure of the previous paragraph to first grow them to a population of 20,000, by having each chromosome pair produce 4 rather than 2 offspring chromosomes for the next generation. A few generations of doubling (7 or 8 in this example) provides the necessary population size. Under this scheme, at generation 200 a random pair of chromosomes will normally share several (up to 5) small ($< 5 \times 10^5$) *ibd* segments. The population provides ample overall *ibd*, in a wide range of size segments, to provide an excellent testing ground for our methods.

Different population recombination histories can be simulated, and the same population history can be populated with HapMap founder chromosomes in different independent realizations. The latter may be more useful, since *ibd* relative to the founder chromosomes will be known, and sets of chromosomes with specific *ibd* patterns can be selected. Using different assignments of HapMap chromosomes to the founders, we propose to study the variation in performance of *ibd* inference resulting from the natural levels of similarity and difference among true HapMap chromosomes.

### 5.1.2 Improved models and methods for *ibd* inference

We propose to improve on the model of section 4.2.1 for the inference of *ibd* among multiple genomes, and to assess its performance. We will use sets of chromosomes selected from our population generated in section 5.1.1 to assess the performance of our method. We will address not only the accuracy of inferred *ibd*-segments, but also the calibration of uncertainty. For example, in the simple case of *ibd* or non-*ibd* at some location and between a pair of chromosomes, results that have 90% probability (conditional on the haplotypic data) of being *ibd*, should in fact be *ibd* in 90% of these cases.

Although our focus will be on sets of four chromosomes, representing genetic data on a pair of individuals (Thompson 2008b), we will assess the inference advantages (if any) and computational disadvantages of making inferences jointly among larger sets of chromosomes. We will assess the ability of the modified model of equation (3) of section 4.2.1 to deal with cases where there are shared ancestral recombination breakpoints. We expect to be able to find many such examples in our simulated population. We will investigate whether, in fact, even in this small population, it is necessary to allow for these shared ancestral recombination events, or whether the basic model ($Q^*$ rather than $Q^\dagger$ in equation (3)) can accurately infer the changes in *ibd* pattern.

Leutenegger et al. (2003) estimated parameters of the *ibd* model, but in the current context there is little meaning or relevance of a global estimate of "kinship" or "mean *ibd* segment length". Rather than attempt estimation of parameters in this situation, we will rather regard them as tuning parameters of the method. We will develop ways to assess values for these parameters, in terms both of the accuracy of inferences, and the calibration of uncertainty.

The need to allow for "data error" is clear (section 3.4). Without such allowance, a single discrepant marker may destroy the inference of an entire *ibd* segment (Leutenegger et al. 2003). On the other hand, too large an allowance for error reduces the power to determine *ibd* segments accurately; in fact, *ibd* may be overestimated. We propose to assess whether the simple error model of section 4.2.1 is adequate, and develop methods for the appropriate choice of the error parameter $\epsilon$. We will create data errors in our selected chromosome sets and examine the robustness of inferences to these errors.

### 5.1.3 The realities of population heterogeneity

In our proposed analyses, we will use empirical sample SNP allele frequencies from the descendant individuals in our simulated populations. In a homogeneous population, where we avoid use of SNPs with small ($< 0.06$) minor allele frequencies, preliminary analyses have shown the detection of *ibd* segments robust to the exact allele frequencies used (Thompson 2008a). However, in subdivided or admixed populations, not only is the choice of appropriate allele frequencies a greater concern, but also the subdivision itself affects the probability distributions of lengths of *ibd* segments shared within and between population subdivisions (Chapman and Thompson 2002).

We propose to generate population members as in section 5.1.1, but including both admixture and subdivision. A founding mixture does not require any modification of the *ibd* simulation; we can simply use as the founder chromosomes both African (YRI) and European (CEU) haplotypes. More structured admixture will require more constraints on the way in which haplotypes in founder chromosomes are assigned. In either case, since the *ibd* simulation program tracks the true descent of founder chromosome segments, the (simulated) ethnic origin at every SNP locus in every chromosome of the descendant population is known. We will study the robustness of our *ibd* inferences to admixture, taking descendant chromosomes at different numbers of generations from the initial founding mixture. We will examine the effect of allele frequency differences between the two founding populations, specifically seeking regions where SNP allele frequencies differ widely, and where, although the combined minor allele frequency may be greater than 6%, a SNP allele is rare in one of the founding populations.

We propose to generate descent in subdivided populations, according to the model of Chapman and Thompson (2002), and to investigate the ability of our approach to infer *ibd* segments both within and between subdivisions. Since subdivision leads to heterogeneity of lengths of *ibd* segments, the issue now is one of robustness of inferences to the *ibd*-model parameters used in analysis, as well as to the data-model parameters such as allele frequencies.

### 5.1.4 The realities of genetic heterogeneity

While the issue of allele frequencies is primarily one of population structure, the other genetic parameters relate more closely to the meiotic process underlying the recombination breakpoints that give rise to the *ibd* segments of genome. In our preliminary analyses we have assumed equivalence of genetic (meiotic) and physical distance, modeling rates of change of *ibd* that relate to genetic distance in terms of the base-pair locations of SNPs. There are good maps of the broad relationship between physical and genetic distance (Kong et al. 2002), and where such a relationship can be assumed, the local rates of change of *ibd* pattern can simply be scaled appropriately. However, there are practical issues arising from recombination hot-spots and cold regions. A cold (low-recombination) region leads to low haplotypic diversity, and extended regions of linkage disequilibrium (LD), resulting in falsely inferred *ibd*. Recombination hot-spots break down LD, but may result in there being few SNPs within an *ibd* segment, and hence difficulties in *ibd* detection. We do not propose to include LD in our models, since LD modeling is complex and large samples of haplotypes are required for accurate model estimation (Browning 2008). We will however, investigate the impact of LD and recombination hot-spots on our ability to detect *ibd* segments of genome from both genotypic and haplotypic data.

Another reality of real data is the presence of copy-number variants or CNV (Wong et al. 2007). CNV are of particular interest as potential causal variants in the analysis of complex traits (Itsara et al. 2009), are associated with locally increased genetic variation (Tian et al. 2008), and can result in data errors in the calling of SNP genotypes (MacConaill et al. 2007). We expect that, without additional information, CNV and data errors may adversely affect our methods that use only population data for inference. Specifically, with genotypes deterministically called (without a quantitative measure), note that deletions are indistinguishable from *ibd* between the two genomes of an individual, and that *ibd* sharing of a deletion between individuals would not be detected with our model. Since pedigree data will aid in addressing these limitations, we will consider assessment of the effects of CNV and data error on *ibd* inference in section 5.2.4 below.

## 5.2 *ibd* inference augmented by pedigree data

### 5.2.1 A simulation framework for multiple small pedigrees in populations

We propose to extend the simulation framework of section 5.1.1 to the case of related pedigrees within the larger population framework. The 200-generation random-mating population of section 5.1.1 will contain half-sibs, but very few full-sibs or nuclear families. Thus, rather than extract pedigree segments from the population as simulated, we will take chromosomes from generation 196 (say) and simulate descent in small four-generation pedigrees to obtain the “current” population. We will then form the genotypic data on these individuals exactly as before, and will assume that members of the last three generations of these pedigrees are sampled. Thus we will have small pedigrees within a population for which *ibd* is known over 200 generations of ancestry, but in which the SNP marker data derive from real human chromosomes, with real SNP locations, real SNP allele frequencies, and real LD patterns.

### 5.2.2 Combination of information within and between pedigrees

It is well recognized that within-pedigree segregation of genetic markers can assist in assessing between-pedigree associations. In the context of SNP data, within-pedigree data provide haplotypic information (Li and Li 2007). Additionally, due to the sharing of large chromosome segments among close relatives, data errors can be detected and large savings in genotyping costs can be achieved by imputing genotypes of relatives (Chen and Abecasis 2007).

Thompson (1978), in a very early application of ancestral inference, referred to individuals for whom genetic data were available but whose parents were unavailable as the *senior sampled members* (SSM) of the population. These SSM are the key individuals in inferences of between-pedigree *ibd*. Data on these individuals is enhanced by information from their close relatives, including other SSM such as their sibs. We propose to use our existing within-pedigree inference methods and programs, including the new MORGAN-3 gl_auto program, to infer haplotypic data on the SSM to the extent possible. The population-based *ibd* inference methods of section 5.1.2 can be easily extended to deal with data that is in part haplotypic and in part genotypic. We will use these methods to detect *ibd* between pedigrees in our simulated population, and assess the performance of the method.

As in section 5.1.3 we will assess the impact of admixture and subdivision on the ability of the method to detect *ibd* segments among the SSM of different pedigrees. We will assess the accuracy of *ibd* inferences, and the calibration of the conditional probabilities of *ibd* relative to the performance when no data on close relatives are available. Generally we expect improved performance, especially in the case of subdivision, where each final 4 generation pedigree will exist within a subdivision of the larger population. The case of admixture is less clear. The presence of heterogeneous admixed chromosomes, from original populations with diverse SNP allele frequencies, may lead to greater uncertainty in within-pedigree inferences. We will assess the impact of within-pedigree *ibd* uncertainty, reflected in uncertainty in the haplotypes of the SSM, on the inference of *ibd* among pedigrees.

### 5.2.3 Comparison of joint and pairwise inference of haplotypic *ibd*

To infer *ibd* within and between the diploid genomes of a pair of individuals from genotypic data, at least the four underlying haploid genomes of the pair must be considered jointly. If pedigree data provide clear haplotypic information, pairwise analyses among SSM haplotypes could be performed, but joint analysis of the four haplotypes may still provide more accurate *ibd* inferences or better calibrated *ibd* probabilities (section 4.2.2). Where haplotypes are too uncertain, joint analysis will be necessary. We will investigate the impact of levels of uncertainty in individual haplotypes on the comparison between joint and pairwise analyses of these haplotypes. We will compare analyses using uncertain haplotypes with those using more certain genotypes.

In the context of inferring pedigree relationships, Sieberts et al. (2002) showed that inference of a relationship between two individuals can be much enhanced by including a third individual in the analysis. Similarly here, including additional haplotypes or genotypes in a joint analysis may improve performance, particularly in the case of allelic heterogeneity in admixed populations. Joint analyses should facilitate distinction between general sub-population levels of similarity and the greater segment similarity of true *ibd* segments. We therefore propose to investigate the feasibility and advantages of joint inference, and the loss of accuracy of *ibd* inference due to lack of phase information, specifically with reference to the admixed and subdivided populations.

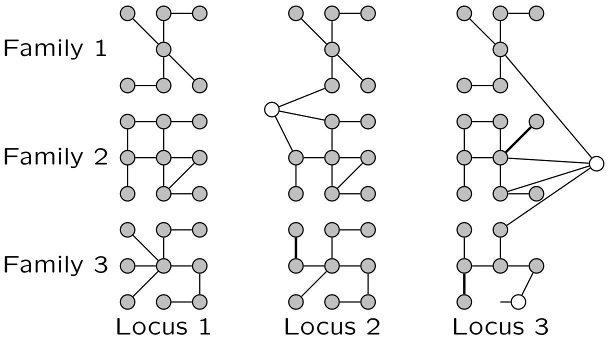
### 5.2.4 Use of pedigree data to address data error and CNV

As suggested in section 5.1.4, typing error and copy number variants (CNV) will likely severely impact inferences of *ibd* based only on population haplotypes or genotypes. While our model can allow for more error, by increasing the error-rate parameter $\epsilon$ (section 4.2.1), too large an allowance for error will reduce power. The ability of within-pedigree data to detect data error and segregating deletions may be essential to dealing with these in the analysis of between-pedigree *ibd*. A general study of this ability of within-pedigree data is beyond the scope of this two-year proposal. However, by artificially creating such errors and deletions, we will investigate the impact on inference of between-pedigree *ibd* of detected and undetected typing errors and of detected and undetected CNV, in the SSM of our simulated pedigrees. In the case of “detected” error, we will delete the data at the aberrant SNPs before analysis. In the case of “detected” deletions, our analysis will recognize the hemizygous region and the resulting locally reduced number of haplotypes under comparison.

## 5.3 Trait modeling and analyses in pedigrees and populations

### 5.3.1 The DGL for multiple pedigrees

Returning now to the DGL-graph Figure of section 4.2.3, note that the founder origins of the DGL are irrelevant in computing the probabilities of observed trait or marker data: only the fact of *ibd* is used.



The figure shows, schematically, the DGL graphs for three pedigrees from a study at three locations across a chromosome. As before, within families, recombinations change the DGL graph along a chromosome. Two such changes are shown for Family-3, and one for Family-2. Additionally now there may be *ibd* among the founders of a given family, or between members of different families. This results in previously distinct DGL merging into a single DGL. For example, at Locus-2, the lower left node of Family-1 has merged with the top left node of Family-2 creating a single DGL in this region of the chromosome. At Locus-3, the two lower right nodes of Family-3 have merged, so that the individual who previously had copies of each of these DGL now has two copies of a single DGL.

A key point is that such links will be few and sparse (Donnelly 1983). An exception might be for ascertained families in regions where the DNA does affect the trait. In this case several families might share a DGL, or even several disjoint groups of families share different DGL. An example is shown at the right of the Figure where a node from each of the three families has merged into a single DGL. Even in this case, a single DGL links the families, and computation over the graph remains straightforward, since such nodes form a cutset and the allelic type assigned to this DGL can be conditioned on in proceeding through the graph.

We propose to use the same methods developed for single pedigrees under the parent grant (section 4.2.3) to model trait data in collections of pedigrees among which *ibd* may have been inferred from dense SNP data. As in the single-pedigree case there may be DGL-graph uncertainty, and computations are done for each of a sample of DGL graphs realized conditionally on marker data.

### 5.3.2 Real-data analyses on collections of pedigrees

In order to truly validate and assess the usefulness of methods, it is essential they be applied to real data. While population data suitable for methods testing are generally publicly available
(e.g. http://www.hapmap.org/downloads/), family data are not. Under the parent award several real data sets of complex traits on pedigree structures are available to us for the purpose of evaluating and extending our methods (Table: see next page). The first five data sets are available to us through Dr. Wijsman’s collaborations, and all analyses of these pedigree data will be carried out in collaboration with Dr. Wijsman, who is a senior coinvestigator on the parent grant. The last data set (FRM) is a subset of the Framingham data made available for Genetic Analysis Workshop 16, and available to us 2008-2011 under a dbGaP (www.ncbi.nlm.nih.gov/dbgap) approval.

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. Three of them (AUT, AD, FCHL) have a subset of families also with a full SNP genome scan. AD has the Illumina Linkage IVb panel. AUT families have the Affymetrix 10K v2 SNP panel, and a few AUT trios within these families have the Illumina 1M SNP panel, with 92 more trios in progress. FCHL has the $\sim$60K CVD-gene-specific panel on most individuals who have STR typing. 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).

| Data Set | AD | FCHL | Guam | DYS | AUT | FRM |
|---|---|---|---|---|---|---|
| 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.

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. Specifically, the AD, AUT, FCHL and FRM data sets all have pedigrees which, considered independently, provide linkage signals. All have dense SNP data available either genome-wide or in regions, and will provide ideal testing grounds for inference of *ibd* among pedigrees, and the development of methods to use this inferred *ibd* to increase support for linkage and to resolve possible genetic heterogeneity in traits among subsets of the pedigrees. Additionally, the Guam pedigree exhibits extreme complexities, multiple generations of unsampled sampled individuals, and potential uncertainties about the ancestral pedigree, which together challenge our current methods. Although here we have only STR data, this will nonetheless provide a test case of using inferred *ibd* among current population members, and hence the DGL graph at STR locations, to provide more effective methods of trait analysis.

The extent to which evaluation and testing of methods on these data sets can be accomplished within the two-year time-frame of the Competitive Revision is unclear. To the extent this is not achieved within the time-frame of the Revision, this testing will be subsumed into ongoing research under the parent R37 award which continues 2011-12.

### 5.4 Software development and distribution.

New software developed under the Competitive Revision will be integrated into MORGAN-3 the software product of the parent grant.

With the assistance of our new MORGAN programmer Dr.Steven Lewis, there has been rapid progress in the development of MORGAN in 2008-9. There have been significant improvements, most notably in finalizing the transition from MORGAN-2 to MORGAN-3 which enhances our abilities to analyze complex traits. All MORGAN programs from MORGAN 2.8 and 2.9 are now also released in the beta test version of MORGAN-3. (See Progress Report, section 4.1.4.)

The fundamental change with MORGAN-3 is that trait loci are no longer identified with a trait phenotype. 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 phenotype, we enable a far more flexible framework for complex models. Already released in MORGAN-3 is a new program (gl_auto) which outputs multiple realizations of FGL inferred from marker data (section 4.1.4). These FGL determine the DGL graph within a pedigree, and we propose to use these in developing programs for computing trait likelihoods conditional on these DGL (section 4.2.3).

A prototype program for inference of *ibd* genome segments from population genotypes or haplotypes has already been implemented within the MORGAN-3 structure (section 4.2.2), although it is not yet released. We propose that this program be further developed to meet the goals of section 5.1. Additional programs, or a more general version of the same program, will be developed to combine the within-pedigree DGL with *ibd* inferred among pedigrees (section 5.3.1). The program to compute trait likelihoods on the DGL graph will be extended to accommodate these more general multi-pedigree DGL.

As with all MORGAN programs, we will develop “gold-standard” test examples and documentation for these new programs. When fully developed, they will be included in the MORGAN tutorial, together with (simulated) data files to provide examples of their use.

### 5.5 Timeline: Job creation and accelerated research

This proposal both expands the tempo of scientific research under the parent grant and also creates two new positions for the two years of the request, and provides additional salary for current part-time staff.

With increasing availability of dense genomic data, and preliminary work already accomplished, this is a project ready to roll. The theoretical modeling and preliminary examples show it is feasible, and the proposed methods have the potential to combine the advantages of population and of pedigree data in the resolution of complex traits. However, personnel are necessary to expand the research program, and increase the tempo of scientific research.

Specifically, a postdoctoral researcher is needed to take the lead in implementing the proposed studies, a graduate student research associate is needed to assist, and increased support is needed for our part-time software research scientist, in order for programs to be integrated into our existing MORGAN software package. The personnel supported on this award will be fully integrated members of the current parent-grant research group, and will participate in weekly research-group meetings, as well as benefit from other group meetings and discussions.

The proposed research can be accomplished in two years:
**Months 1-6:** Simulation and analysis of *ibd* from population haplotypes and genotypes. Improvement of the analysis models, and data-based methods for tuning parameters of the analysis model. Improvement of current computer programs for population-based *ibd* inference and their integration into MORGAN-3.
**Months 7-12:** Studies of robustness and performance of the methods, and the impact of population subdivision and admixture, and of genetic map variation, linkage disequilibrium, CNV, and data error.

**Months 13-18:** Methods to use within-pedigree data to inform the inference of between-pedigree *ibd*. Comparisons of analyses based on haplotypes or on genotypes, and more generally of joint and pairwise inference. Studies of improvements gained by incorporating within-pedigree information in addressing issues of genetic map variation, linkage disequilibrium, CNV, and data error.

**Months 19-24:** The combination of within-pedigree and between-pedigree *ibd* in the inference of the DGL graph across a chromosome. Trait model computations using the DGL graph inferred both within and between pedigrees. Testing of methods for inference of *ibd* and for subsequent trait models analyses using real pedigree data. Integration of computer programs for these methods into MORGAN-3.

To the extent work under Months 19-24 cannot be accomplished within the time-frame of the Competing Revision, it will be integrated into work in 2011-2012 under the parent grant.

## Bibliography and References Cited

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

AGP (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nature Genetics 39:319–328. PMID: 17322880

Albers CA, Stankovic J, Thomson R, Bahlo M, Kappen HJ (2008) Multipoint approximations of identity-by-descent probabilities for accurate linkage analysis of distantly related individuals. American Journal of Human Genetics 82:607–622

Allison DB (1997) Transmission disequilibrium tests for quantitative traits. American Journal of Human Genetics 60:676–690

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

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

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

Balding DJ, Nichols RA (1994) DNA profile match probability calculations: How to allow for population stratification, relatedness, database selection, and single bands. Forensic Science Int 64:125–140

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

Blacker D, Bertram L, Saunders AJ, Moscarillo TJ, Albert MS, Wiener H, Perry RT, Collins JS, Harrell LE, Go RCP, et al (2003) Results of a high-resolution genome screen of 437 Alzheimer's Disease families. Human Molecular Genetics 12:23–32

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

Borecki IB, Province MA (2008) Genetic and genomic discovery using family studies. Circulation 118:1057–1063

Browning SR (2008) Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics 178:2123–2132

Chapman NH, Leutenegger A, Badzioch MD, Bogdan M, Conlon EM, Daw EW, Gagnon F, Li N, Maia JM, Wijsman EM, Thompson EA (2001) The importance of connections: Joining components of the Hutterite pedigree. Genetic Epidemiology 21 (Suppl 1):S230–S235. PMID: 11793674

Chapman NH, Thompson EA (2002) The effect of population history on the lengths of ancestral chromosome segments. Genetics 162:449–458. PMID: 12242253

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

Chen W, Abecasis GR (2007) Family-based association tests for genomewide association scans. American Journal of Human Genetics 81:913–926

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872

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

Dewan A, Liu MG, Hartman S, Zhang SSM, Liu DTL, Zhao C, Tam POS, Chan WM, Lam DSC, Snyder M, Barnstable C, Pang CP, Hoh J (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. Science 314:989–992

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

Edwards AO, Ritter R, Abel KJ, Manning A, Panhuysen C, Farrer LA (2005) Complement factor H polymorphism and age-related macular degeneration. Science 308:421–424

Edwards KL, Hutter CM, Wan JY, Kim H, Monks SA (2008) Genome-wide linkage scan for the metabolic syndrome: The GENNID Study. Obesity 16:1596–1601

Ehret GB, Morrison AC, OConnor AA, Grove ML, Baird L, Schwander K, Weder A, Cooper RS, Rao DC, Hunt SC, Boerwinkle E, Chakravarti A (2008) Replication of the Wellcome Trust genome-wide association study of essential hypertension. European Journal of Genetics 16:1507–1511

Elbaz A, Nelson LM, Payami H, Ioannidis JPA, Fiske BK, Annesi G, Belin AC, Factor SA, Ferrarese C, Hadjigeorgiou GM, et al (2006) Lack of replication of thirteen single-nucleotide polymorphisms implicated in Parkinson's disease: a large-scale international study. Lancet Neurology 5:917–923

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

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

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theoretical Population Biology 3:87–112

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

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

Fisher SA, Abecasis GR, Yashar BM, Zareparsi S, Swaroop A, Iyengar SK, Klein BEK, Klein R, Lee KE, Majewski J, et al (2005) Meta-analysis of genome scans of age-related macular degeneration. Human Molecular Genetics 14:2257–2264

Fodor FH, Weston A, Bleiweiss et al I (1998) Frequency and carrier risk associated with common BRCA1 and BRCA2 mutations in Ashkenazi Jewish breast cancer patients. American Journal of Human Genetics 63:45–52

Fridman E, Carrari F, Liu YS, Fernie AR, Zamir D (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. Science 305:1786–1789

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

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

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

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308:419–421

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 2:3–19

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

International Hapmap Consortium (2005) A haplotype map of the human genome. Nature 237:1299–1319

Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI, Mefford H, Ying P, Nickerson DA, Eichler EE (2009) Population analysis of large copy number variants and hotspots of human genetic disease. American Journal of Human Genetics 84:148–161

Jacquard A (1972) Genetic information given by a relative. Biometrics 28:1101–1114

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

Klein ML, Schultz DW, Edwards A, Matise TC, Rust K, Berselli CB, Trzupek K, Weleber RG, Ott J, Wirtz MK, Acott TS (1998) Age-related macular degeneration - Clinical features in a large family and linkage to chromosome 1q. Archives of Ophthalmology 116:1082–1088

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389

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

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

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

Laird NM, Horvath S, Xu X (2000) Implementing a unified approach to family based tests of association. Genetic Epidemiology 19 (Suppl 1):S36–S42

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

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

Lauritzen SJ (1992) Propagation of probabilities, means and variances in mixed graphical association models. Journal of the American Statistical Association 87:1098–1108

Leibon G, Rockmore DN, Pollak MR (2008) A SNP streak model for the identification of genetic regions identical by dscent. Statistical Applications in Genetics and Molecular Biology 7:Article 16

Leutenegger A, Prum B, Genin E, Verny C, Clerget-Darpoux F, Thompson EA (2003) Estimation of the inbreeding coefficient through use of genomic data. American Journal of Human Genetics 73:516–523. PMID: 12900793

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

Li X, Li J (2007) Comparison of haplotyping methods using families and unrelated individuals on simulated rheumatoid arthritis data. BMC Proceedings 1 (Suppl 1):S55

Liu F, Arias-Vsquez A, Sleegers K, Aulchenko YS, Kayser M, Sanchez-Juan P, Feng BJ, Bertoli-Avella AM, van Swieten J, Axenovich TI, Heutink P, van Broeckhoven C, Oostra BA, van Duijn CM (2007) A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population. American Journal of Human Genetics 81:17–31

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

MacConaill LE, Aldred MA, Lu X, LaFramboise T (2007) Toward accurate high-throughput SNP genotyping in the presence of inherited copy number variation. BMC Genomics 8:211

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

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

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

Neuhausen S, Gilewski T, Norton L, Tran T, McGuire P, Swensen J, Hampel H, Borgen P, Brown K, Skolnick MH, Shattuck-Eidens D, Jhanwar S, Goldgar DE, Offit K (1996) Recurrent BRCA2 6174delT mutations in Ashkenazi Jewish women affected by breast cancer. Nature Genetics 13:126–128

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

Peltonen L, Pekkarinen P, Aaltonen J (1995) Messages from an isolate: lessons from the Finnish gene pool. Biological Chemistry Hoppe Seyler 376:697–704

Poorkaj P, Tsuang D, Wijsman EM, Steinbart E, Garruto RM, Craig UK, Chapman NH, Anderson L, Bird TD, Plato CC, Perl DP, Weiderholt W, Galasko D, Schellenberg GD (2001) TAU as a susceptibility gene for Amyotropic Lateral Sclerosis-Parkinsonism Dementia complex of Guam. Archives of Neurology 58:1871–1878

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool-set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559–575

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

Rivera A, Fisher SA, Fritsche LG, Keilhauer CN, Lichtner P, Meitinger T, Weber BHF (2005) Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. Human Molecular Genetics 14:3227–3236

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

Sieberts SK, Wijsman EM, Thompson EA (2002) Relationship inference from trios of individuals in the presence of typing error. American Journal of Human Genetics 70:170–180, PMID: 11727198

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

Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, et al (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nature Genetics 38:617–619

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

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. American Journal of Human Genetics 62:450–458

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. American Journal of Human Genetics 52:506–516

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, et al (2005) A common inversion under selection in Europeans. Nature Genetics 37:129–137

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

Sung YJ, Thompson EA, Wijsman E (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. PMID: 17123301

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

Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA (2008) Shared genomic segment analysis. Mapping predisposition genes in extended pedigrees using SNP genotype assays. Annals of Human Genetics 72:279–287

Thompson EA (1978) Ancestral inference II: The founders of Tristan da Cunha. Annals of Human Genetics 42:239–253

— (1981) Pedigree analysis of Hodgkin's disease in a Newfoundland genealogy. Annals of Human Genetics 45:279–292. PMID: 7305282

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

— (2005) MCMC in the analysis of genetic data on pedigrees. In F Liang, JS Wang, W Kendall, eds., *Markov Chain Monte Carlo: Innovations and Applications*, 183–216. World Scientific Co Pte Ltd, Singapore

— (2008a) Analysis of data on related individuals through inference of identity by descent. Technical report # 539, Department of Statistics, University of Washington

— (2008b) The IBD process along four chromosomes. Theoretical Population Biology 73:369–373. PMID: 18282591

Thompson EA, Basu S (2003) Genome sharing in large pedigrees: multiple imputation of ibd for linkage detection. Human Heredity 56:119–125

Thompson EA, Geyer CJ (2007) Fuzzy p-values in latent variable problems. Biometrika 90:49–60

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

Tian D, Wang Q, Zhang P, Araki1 H, Yang S, M K, Nagylaki T, Hudson R, Bergelson J, Chen JQ (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455:105–109

Ueda T, Sato T, Hidema J, Hirouchi T, Yamamoto K, Kumagai T, Yano M (2005) qUVR-10, a major quantitative trait locus for ultraviolet-B resistance in rice, encodes cyclobutane pyrimidine dimer photolyase. Genetics 171:1941–1950

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

Weir BS (2001) Forensics. In DJ Balding, M Bishop, C Cannings, eds., *Handbook of Statistical Genetics*, 721–739. Wiley, New York

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678

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

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

Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL (2007) A comprehensive analysis of common copy-number variations in the human genome. American Journal of Human Genetics 80:91–104

Yang ZL, Camp NJ, Sun H, Tong ZZ, Gibbs D, Cameron DJ, Chen HY, Zhao Y, Pearson E, Li X, Chien J, DeWan A, Harmon J, Bernstein PS, Shridhar V, Zabriskie NA, Hoh J, Howes K, Zhang K (2006) A variant of the HTRA1 gene increases susceptibility to age-related macular degeneration. Science 314:992–993

Yu CE, Dawson G, Munson J, D'Souza I, Osterling J, Estes A, Leutenegger AL, Flodman P, Smith M, Raskind WH, Spence MA, McMahon W, Wijsman EM, Schellenberg GD (2002) Presence of large deletions in kindreds with autism. American Journal of Human Genetics 71:100–115

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

## 7 Progress Report Publication List; 2008-9

1. Basu, S, Di, Y., and Thompson, E. A. (2008) Exact trait-model-free tests for linkage detection in pedigrees. *Annals of Human Genetics:* **72:** 676–682. PMID: 18507652

2. Basu, S., Stephens, M., Pankow, J. S. and Thompson E. A. (2009) A likelihood-based trait-model-free approach for linkage detection of a binary trait. *Biometrics:* in press.

3.* Di, Y., and Thompson, E. A. (2009) Conditional tests for localizing trait genes. *Human Heredity*: in press.

4. Igo, R.P. Jr., Wijsman, E.M. (2008) Empirical significance values for linkage analysis: trait simulation using posterior model distributions from MCMC oligogenic segregation analysis. *Genetic Epidemiology* **32:** 119–131. PMID: 17849492.

5. Marchani, E.E., Di, Y., Choi, Y., Cheung, C., Su, M., Boehm, F., Thompson, E.A., and Wijsman, E.M. (2009) Contrasting IBD estimators, association studies, and linkage analyses using the Framingham data. In ”Genetic Analysis Workshop 16.” *BMC Proceedings:* in press.

6. Sieh, W., Choi, Y., Chapman, N.H., Craig, U.-K., Steinbart, E.J., Rothstein, J.H., Oyanagi, K., Garruto, R.M., Bird, T.D., Galasko, D.R., Schellenberg, G.D., and Wijsman, E.M. Identification of novel susceptibility loci for Guam neurodegenerative disease: Challenges of genome scans in genetic isolates. Submitted.

7. Thompson, E. A. (2008) The IBD process along four chromosomes. *Theoretical Population Biology* **73:** 369–373. PMID: 18282591.

8. Thompson, E. A. (2008) Uncertainty in inheritance: assessing linkage evidence. JSM Proceedings, Salt Lake City 2007. Pp. 3751–3758.

9.* Thompson, E. A. (2009) Inferring coancestry of genome segments in populations. *Invited Proceedings of the 57th Session of the International Statistical Institute.* Durban, South Africa. To appear.

10. Tong L. and Thompson, E. A. (2008; Jan.) Multilocus lod scores in large pedigrees: Combination of exact and approximate calculations. *Human Heredity* **65:** 142–153.

* Items [3] and [9] are included as appendix material.