2 PROJECT PROGRESS REPORT; 12/2007-5/2011

2.1 Summary of previous specific aims

The overall objective remains the development of techniques for the analysis of the genetic basis of complex familial traits, with the current focus toward methods of analysis that can make effective use of increasingly available genomic data including dense SNP markers and DNA sequence variants. The aims include: 1) the further development of Markov chain Monte Carlo (MCMC) methods for realization of descent patterns in pedigrees jointly across multiple genome locations and conditional on multilocus marker data, 2) new approaches to using realized descent patterns in the joint linkage and segregation analysis of complex genetic traits; 3) development of methods for assessment of statistical significance of linkage findings, using the empirical MCMC-generated distribution of gene descent; 4) development of methods for inference and use of gene descent patterns in pedigrees accommodating marker model and data uncertainty, including typing error, linkage diseqilibrium, and allele frequency uncertainty; 5) comparison of the new approaches with existing methodology through appropriate simulation studies and the analysis of sample real data sets; and 6) development, documentation, and support of our publicly available MORGAN software, together with additional related software, to enable broad use of our methods in the analysis of complex traits on general pedigrees.

Additionally, under an ARRA competitive supplement (9/2009-9/2011), we are extending methods for inference of *ibd* among remotely related individuals whose pedigrees are unknown, and developing methods for combining the *ibd* inferred within pedigrees with *ibd* inferred among pedigrees.

2.2 Importance of the research and of findings

Family-based designs remain important in the study of complex traits for several reasons. (1) Although modern dense genotyping and sequencing technologies can nominate disease risk variants, a standard step used to help prioritize the large number of variants is demonstration of cosegregation, or evidence for linkage, with the trait of interest in a pedigree. (2) For risk variants with low frequency, use of large pedigrees allows efficient sampling of the multiple copies of a variant needed to evaluate effects of such a variant. (3) Variants identified through pedigree designs tend to have larger effects than those identified in population-based samples, leading to easier translation to biological mechanisms. (4) Samples of pedigrees are often the first used to investigate the genetic basis of novel phenotypes, because the relatively small samples permit collection of large numbers of phenotypes and sometimes costly or difficult new phenotype measurements. This contrasts with large population-based samples that rely on inexpensive and widely-available standard measures. (5) Pedigree data are efficient, thus reducing sample size requirements. This is the consequence of Mendelian transmission, which guarantees that genetic information propagates through the pedigree, even when some or even many subjects in older generations are unavailable for sampling.

Many studies have a rich set of existing marker and trait data that has been collected over time and with different scales of completeness. Many such studies also have existing biosamples in various stages of completion/depletion. It is critical to be able to integrate these different sources of data. As new technologies for genotyping become available that add potential additional information, it is common and useful to add such newer genotypes to a sample. However, because of sample availability and/or cost, it is not always possible to add genotypes from new technologies for all subjects. To maximize the use of the data from these samples, which often also includes rich and costly phenotype information, it is useful to be able to integrate the different sources and levels of existing data with a modest amount of newer data to maximize the information gained, without incurring unnecessarily high costs. Development of methods that achieve this integration is a focus of our current research.

As described more fully below, an important technical and methodological advance has been in direct use of patterns of identity by descent (ibd) inferred from marker data, through the use of the *ibd graph*. This

development of our approach provides important gains in efficiency, and also lays the framework for a more flexible approach to analyses of complex traits. It also provides for integrating analyses at different scales both across the genome and across subjects. By specification of descent in terms of recombination breakpoints rather than by marker, it permits the use of modern dense marker data. By specification of shared genome in terms of *ibd* rather than descent from pedigree founders, it can express coancestry on the population scale as well as the pedigree scale. Thus, it allows for combining dense data at the population level with genomically sparser data on pedigrees, increasing the power and resolution of our methods.

2.3 Report of research accomplished

Numbers in square brackets refer to papers resulting from this award, and are listed in section 5.1. A reprint/preprint of each of these papers is included as Appendix material. Three major papers currently in preparation are listed for ease of reference and cited as [P#], while project-generated software resources are cited as [S#] and listed in section 5.2.

1 MCMC methods for sampling gene descent on pedigrees: Aim 1

The fundamental basis of our approach is the sampling of gene descent paths on pedigrees conditional on genetic marker data, using Markov chain Monte Carlo (MCMC) methods. We have continued to improve our MCMC algorithms [13,17], such that we can now efficiently use SNP data at a 0.5cM spacing, even on extended pedigrees with several generations of ancestral individuals unobserved. With increasing availability of dense SNP data and complexity of trait data, our approach has been to obtain samples of descent conditional on marker data once only, and to use these in subsequent analysis of multiple trait models and trait data sets. Generally, marker maps, model parameters, and data are much more firmly established than are the corresponding elements for complex phenotypes.

Realizations of gene descent may be specified through founder genome labels (FGL) that identify the founder origin of DNA present in each individual at each genome location of interest. At a given location, the FGL determine the pattern of gene identity by descent (*ibd*) among individuals, in the form of a descent graph. The descent graph has labeled edges that represent individuals. These edges connect nodes that represent the two FGL genomes carried by the individuals. With many markers, the descent graph is not an efficient way either to store *ibd* patterns inferred from marker data or to use in the analysis of trait data. Different descent patterns from founders (different FGL) can give rise to the same *ibd* pattern among individuals observed for trait or marker data. Different MCMC realizations often give rise to the same *ibd* patterns (with the same or different FGL). Along a chromosome, the pattern of *ibd* can change only at recombination breakpoints in meioses ancestral to the individuals, and hence remains constant over many markers. We have therefore developed a version of the descent graph which we call the *ibd* graph [12].

Like the descent graph, the location-specific *ibd* graph has labeled edges that represent individuals. These edges connect (unlabeled) nodes that represent genome carried by the individuals; individuals connecting to a given node share DNA *ibd* at the given location [12,17]. Across the genome however, *ibd*-graphs are specified in terms of their change-points resulting from ancestral recombination breakpoints. Hence the storage of multiple realizations of *ibd* graphs, genome-wide and for multiple pedigrees, becomes practical. While genetic markers may be available at a scale of 1 per 10^4 base pairs (bp), recombination breakpoints are on the scale of 1 per 10^8 bp per meiosis. For example, storage of 1000 realizations of the *ibd* graph across a 200cM chromosome on 31 observed individuals in one of our small test data analyses [19] requires 1.45Mb. The same information, stored by marker for the $\sim$350 markers used in the within-pedigree analyses on this data set would require 67Mb. For the 10,188 dense markers used in the between-pedigree analyses, 2Gb would be required. For data sets with more pedigrees, or more observed individuals, or requiring more MCMC realizations, storage of results by marker would be prohibitive.

The compact storage of *ibd* graphs realized conditional of genetic marker data enables marker analyses on a pedigree data set to be accomplished once only, and hence provides much greater efficiency and flexibility in the subsequent analyses of trait data on the pedigrees [12]. We have implemented this approach of generating

and storing *ibd* graphs in a program, *gl_auto*, within our MORGAN-3 software [S3]. The program samples realizations of gene descent conditional on genetic marker data, using the same improved MCMC algorithms and sampling options as the newer versions of our MORGAN lod score programs [13,17]. Instead of using these directly in lod-score computation, they are output in compact format and stored for subsequent analyses.

The gain in efficiency through once-only MCMC analyses of marker data is only a small part of the advantage of the *ibd*-graph approach. Previous methods have computed lod-score contributions or other linkage test statistics for each descent-graph realization at each marker. However, for analyses of trait data, only the pattern of *ibd* among the individuals observed for the trait is relevant. Under a given model, for given trait data, lod-score contributions need be computed only for each distinct *ibd* graph. We have developed *IBDgraph* software [20], which recognizes equivalence of descent graphs in their compressed *ibd*-graph format, across FGL labelings, across realizations, and across genome locations. In this way, an *ibd* graph may be defined as an equivalence class of descent graphs across these dimensions. Any subsequent analysis, such as computation of a lod-score or test statistic, depends only on the *ibd* graph.

The *IBDgraph* software has been released [S6], and is already providing major yields in linkage analyses of data of extended pedigrees [21,23]. In some cases, computations associated with the trait-data portion of the analysis can be be reduced by two orders of magnitude [12]. For example, in analyses of a 26-member real-data pedigree [23], 1666 realized *ibd* graphs fell into only 24 equivalence classes. Over these 1666 realizations, only 24 lod-score contributions need be computed [12]. Even greater gains are obtainable in studies on larger or more complex pedigrees where more MCMC realizations are required, since the number of distinct *ibd* graphs sampled increases slowly compared to the number of realizations.

The *ibd*-graph approach also has significant practical importance for analyses of public-health-related human data in terms of data confidentiality. For the marker-based MCMC, pedigree data and marker data are required, but no trait data. The resulting *ibd* graphs contain no genetic data of any kind. Consequent trait-data analyses require trait data, but pedigree information is no longer necessary. The three files (pedigree and marker data, *ibd* graphs, and trait data) are linked only though fully de-identified IDs. This approach provides the potential for easier but secure sharing of pertinent information among collaborators in large multi-site studies of complex genetic traits.

Finally, as described in item 4 below, the *ibd* graph is not constrained to defined pedigrees, but can also be used to specify segments of *ibd* inferred at the population level. This flexibility of scope permits the combination of *ibd* within and among pedigrees, which has the potential to increase the power and resolution of linkage analyses [18,19].

2 Using gene descent in the genetic analysis of complex traits: Aims 2 and 3

While the use of MCMC-based realizations of gene descent in the estimation of multilocus lod scores has been well established [13], their use in testing more complex trait hypotheses is more novel. In [1], we developed methods for exact trait-model-free linkage detection tests, and in [2] we used a likelihood approach to test association of trait data with inheritance inferred from marker data. The methods of these two papers have been implemented in our released MORGAN software [S1,S3]. Almost all model-based trait analyses have assumed trait loci with only two (classes of) alleles that differentiate disease risk. The presence of several alleles of varying effect can confound these models. In [9], an important generalization was made in developing joint linkage and segregation analysis methods for oligogenic multiallelic traits.

The variation inherent in sampled descent patterns provides measures of uncertainty in resulting inferences, and hence of significance of linkage findings. We investigated two general approaches. First, we developed a trait-resimulation approach to the estimation of linkage p-values [5], and have released software for this approach [S7]. Second, we used the variation over the sampled descent patterns to provide confidence measures for lod scores [15] and p-values for tests that localize trait genes [4]. In the latter paper [4], we investigated a new approach to localizing trait genes through conditioning on realizations of gene descent at positions bounding a test region of the chromosome. The conditional tests developed can also be used to detect additional linkage signals in the presence of previously detected causal genes, and hence provide

an approach to oligogenic traits where linked loci may contribute to a single phenotype. This may become increasingly important with dense haplotyping or sequencing of regions containing several trait loci.

The use of the *ibd* graph not only greatly improves efficiency of lod-score analyses under single trait-locus models [21,23], but also provides an approach to efficient computation of bivariate lod-score surfaces under models of two linked quantitative trait loci (two-QTL). Even on a large pedigree, the components of the *ibd* graph are relatively small and simple [12]. Thus we have been able to develop methods for exact computation of lod score contributions conditional jointly on the realized *ibd* graphs at two (or even more) linked or unlinked genome locations, and hence determine lod scores under two-QTL models. Our earlier (2006-7) published approaches to this two-QTL problem were either limited to small pedigrees or required extensive Monte Carlo computation to estimate the lod-score contribution for each MCMC-realized descent pattern. On large pedigrees, with extensive missing data, this two-level hierarchy of Monte Carlo was often computationally intensive, and, more importantly, did not always perform well. Instead, we first use our *gl_auto* program to generate and store realizations of *ibd* graphs across the chromosome(s) conditional on the genetic marker data. Given these *ibd* graphs, exact computation of two-QTL lod scores is of the same order of computational complexity as established single QTL models. A paper is in preparation [P3].

3 Using gene descent in genotype imputation and error detection: Aims 2 and 4

A main focus of our work has been the development of our methods to meet the challenges of modern data. Applications range from testing Hardy-Weinberg Equilibrium for data quality control of dense SNP data in population studies [28], to using inferred gene descent for selecting individuals for next-generation sequencing in large pedigrees [27].

As SNP marker panels become denser and next-generation sequence data becomes available, the integration of population-based association methods and pedigree-based linkage methods provides the benefits of both approaches in the analysis of complex traits [7]. We implemented several methods as part of Genetic Analysis Workshop 16 (GAW16), including the option of storing MCMC-generated sampled inheritance vectors and re-using these in subsequent analyses [6,7]. Our results demonstrated the advantages of these stored inheritance vectors, although there remained challenges in using the very dense markers. In addition, we found that our methods could easily handle dense markers to make inferences about *ibd*, and that increased number and density of markers increased the stability of *ibd* estimates, and decreased the number of pairs of subjects with high estimated kinship coefficients, especially among reportedly unrelated individuals. We also showed that existence of *ibd* across reportedly independent pedigrees affects association analyses.

The GAW16 experience combined with availability and importance of increasingly dense genotype data stimulated investigation of alternative approaches to combine pedigree and marker information. Linkage analysis alone does not require extremely dense genotyping. In fact, very dense markers hamper estimation of *ibd* graphs on pedigrees, both by increasing computation time, and by including markers that violate the linkage equilibrium assumption that underlies the methods implemented in MORGAN. However, for analyses that also incorporate marker-based association, allelic information is important. For example, a particular rare variant from sequence data may explain segregation of a phenotype in a pedigree, or a particular combination of alleles might identify a shared haplotype across pedigrees. Our strategy [24,25] involves two stages. First, at a marker density that is appropriate for our existing MCMC methods [12,13], we use the *gl_auto* program to sample inheritance jointly at the positions of these framework markers. Then, we use all the available (denser) marker data to impute alleles between the framework markers, using the jointly sampled inheritance realizations, observed dense genotypes, and a probability-based computation marginally for each marker and subject, with a threshold to determine whether or not to impute an allele.

We have developed preliminary software, *GIGI*, and have tested the imputation approach on three different data sets consisting of 4-5 generation large pedigrees (52-95 members) with missing data in the oldest two generations: a simulated pedigree, and two real pedigrees. All three data sets have both sparser multiallelic STR and dense diallelic SNP markers, with more missing SNP than STR data in the older generations. In the real data sets, accuracy was measured by masking the dense SNPs in a portion of the genotyped subjects

prior to imputation. For a genotype call probability threshold of 80%, accuracy of imputed alleles is $95 - 99\%$, depending on the dataset. Accuracy is inversely related to allele frequency, increases with marker information in the panel of markers used to sample the framework inheritance, and increases also with the number of subjects with measured genotype data. This research has been presented at national meetings [24,25], and a paper is in preparation [P2].

The sampling of *ibd* on pedigrees, first at framework markers, selected for being informative and reliable, and then at intervening dense markers, also provides a computationally efficient approach to identifying probable Mendelian consistent errors for the dense markers. This is important because of the paucity of reliable tools to identify such errors in large pedigrees, coupled with the potential influence of undetected errors on subsequent analyses. If the observed data at a non-framework SNP are inconsistent with a high proportion of realized *ibd* graphs, an error is indicated. This approach has the advantage of requiring neither allele frequencies nor an error model for the non-framework SNPs. For example, on the same simulated 52-member pedigree used above to measure imputation accuracy, we simulated genotypes for 25,000 SNPs spaced at 0.004 cM, with a genotyping error rate of 0.001. Of the generated errors, 88% were Mendelian consistent (MC). Using a framework panel of markers at 0.5 cM density and a 95% threshold for the probability of error for error detection, we identified 98% of the MC errors that are be detectable with perfect information about inheritance at the marker position, for a positive predictive value of 86%.

The descent patterns on pedigrees inferred from framework markers can also be important in the selection of individuals for sequencing at a candidate locus [27], by indicating which related affected individuals are most likely to share *ibd* the variant that has provided the linkage signal on the pedigree.

4 Inference of *ibd* among pedigrees in populations: Aim 4 and ARRA

With modern genetic marker data, relationships among observed pedigree members may be readily validated, but in an extended multi-generation pedigree, ancestral relationships may be uncertain. Even if correct, the stated pedigree may be biased in that descent from prominent individuals may be known, but other unknown relationships may exist among individuals specified as founders. Among smaller pedigrees of a genetic epidemiological study, there may exist unknown ancestral relationships, particularly in study populations where the degree of relatedness in the sample is substantial due to population structure, admixture, or history, or due to the sample ascertainment. Modern genetic marker data permit the inference of segments of genome shared *ibd* among individuals not known to be related, and hence the combination of the power of data on known pedigrees with the resolution of population data.

Under an ARRA Competitive Supplement (funded 9/30/2009-9/29/2011) to the R37 award, we have been investigating methods for the estimation of gene *ibd* among individuals sampled from a population. We have developed a hidden Markov model (HMM) for segments of *ibd* among the four chromosomes of two individuals [11], and investigated its performance using both genotypic and haplotypic SNP data [16]. More recently, we have conducted an extensive simulation study, investigating parameter sensitivity and model performance, and shown that with realistic data at a density of $\sim$50 SNPs per cM, we can reliably detect *ibd* segments of length 1 cM using genotypic data, with better performance if phased haplotypes are available [26]. Thus, shared inheritance of DNA can be detected even when the individuals are separated by up to 100 meioses.

We have implemented our methods in IBD_Haplo [S4], as a separately released part of our main MORGAN-3 package [S3]. The method is fast; for example, 500 pairs of individuals, for a chromosome of 7,000 SNP markers, can be analyzed in under 90 seconds. However, output is large and cumbersome; for each pair, at each marker, the probability of each of the 15 possible *ibd* states among the four chromosomes. We have therefore written and released a small R-package, IBDhaploRtools [S5], to facilitate analysis of these output files.

We have extended these methods in several ways. First we have improved the latent HMM model for *ibd* among haplotypes sampled from a population, and related this model to the coalescent ancestry of the sample [29]. Second, our HMM analysis for pairs of individuals now permits the input of data that is partly genotypic and partly haplotypic. This is of particular importance in analyzing *ibd* between members of different

pedigrees, where partial within-pedigree information on phase of sampled individuals contributes substantially to the accurate estimation of between-pedigree *ibd* [26].

Finally, we are undertaking an extensive study of the effect of varying levels of linkage disequilibrium (LD) on our *ibd* inferences. Specifically, since LD is itself a reflection of remote coancestry, the resolution of *ibd* segments is limited to larger scales than the extent of LD. In this connection, we have developed a novel simulation strategy, *beaglesim*, to generate realistic dense SNP haplotypes at varying levels of LD. An LD model is first fit to a collection of real haplotypes with high LD, using the publicly available BEAGLE package. Simulated haplotypes are then generated from the model, but in generating each haplotype along the chromosome LD is "broken" with probabilities that determine the scale of resulting LD [19]. The generated haplotypes exhibit quite realistic LD patterns, and are used to populate our simulated population samples and pedigree founders. A paper describing these methods and studies is in preparation [P1].

With the end of the ARRA-funded project, methods for the inference and use of inferred *ibd* in population samples will be pursued outside this R37 award, and therefore we do not focus on these results in this report. However, integration of population *ibd* into the analysis of data on multiple pedigrees has become a major focus of continuing research on the R37 award. We have begun to develop methods for combining within-pedigree inference of genome shared *ibd* among known relatives with between-pedigree inference of genome shared *ibd* due to more remote unknown relationships. We have shown that combination of between- and within-pedigree *ibd* can increase both the power to detect genetic linkage and the degree of resolution of loci contributing to a quantitative trait [18].

The *ibd* graph is a key feature of our approach to merging within- and between-pedigree *ibd*. The within-pedigree *ibd* graph is first realized by MCMC, using again the *gl_auto* program and the methods described above [12,17]. Where two founder genomes within a pedigree are *ibd* due to more remote coancestry, the corresponding genome nodes must be merged. Likewise, where *ibd* is inferred between genome segments in different pedigrees, the genome nodes must again be merged. The result is a combined *ibd* graph over all pedigrees; this *ibd* graph may then be used in lod-score computations or other trait-data analyses. The logical constraints on merging nodes of the *ibd* graph are complex, and *ibd* inferences are probabilistic. Development and improvement of our proposed methods is ongoing; we have implemented our procedures in new *IBDmerge* software.

Preliminary results of our proposed merging procedures are immensely encouraging [18,19]. In a 44-member, 5-generation, pedigree, we used our MCMC methods [13] to estimate lod scores for a simulated quantitative trait using the entire pedigree and using three 3-generation subpedigrees of 12 to 14 members. Only the 22 final-generation individuals, 6 to 8 in each subpedigree, were assumed observed for trait and marker data. The sum of the three sub-pedigree lod scores showed loss of signal at the true trait location, and a weaker but still clear false-positive signal in another region of the chromosome. Merging the *ibd* graphs on the three subpedigrees, and computing lod scores on the resultant combined graphs, essentially recovered the lod score computed on the whole pedigree; this full-pedigree lod score accorded closely with the lod score given the true *ibd* on the pedigree. The signal at the true trait locus was restored and the false signal was eliminated [18]. We are currently working with a much more complex 12-generation 95-member pedigree, with multiple inbreeding loops and interconnected paths of descent. On a 200 cM artificial chromosome we have imposed descent from various founders at various locations, and simulated quantitative trait data for several traits, each associated with descent at some location. Using our *beaglesim* approach, we generated dense SNP data at 10,188 markers across the chromosome. The final three-generation subpedigrees of sizes 11, 8 and 12 individuals are assumed observed for trait and marker data, and $\sim$350 markers were selected for the lod score computations. In this instance, exact computation of lod scores on the subpedigrees is feasible, but even MCMC lod score estimation on the entire pedigree is impractical. However, since these are simulated data, the true descent pattern on the entire pedigree is known, and provides the true-*ibd* lod score. For one particular trait selected for current testing, the subpedigrees provide only very weak signals, with the smallest presenting an almost zero lod score. All 10,188 markers were used in our IBD_Haplo program to estimate segments of *ibd* between individuals from these more remotely related subpedigrees. Merging the within-pedigree and

between-pedigree *ibd* using *IBDmerge* provided appropriate strong positive and negative signals, although this example presents new challenges which we are developing methods to address. Uncertainty in merging of *ibd* can be high, and, on this complex and artificial example, false signals arise in the merged *ibd*, due to the correlated descent from several different founders at different genome locations. These methods and results are reported in [19].

5 Application and testing of methods in the analyses of complex traits: Aim 5

Feedback between application and development of new methods is key for producing effective and useable analysis methods. Analysis of real data with developing methods provides comparison of results obtained with new vs. standard methods, as well as identifying challenges that need to be addressed to effectively address the genetic basis of complex traits, to use new data types, and to interpret results. Such applications also test usability of developing software.

We have used six different real datasets in these contexts. Four of the data sets involve psychiatric disorders: Alzheimer disease (Alz), neurodegenerative disease (Guam), Autism (Aut) and Dyslexia (Dys). The other two studies relate to cardiovascular disease: FCHL and the GAW16 Framingham Heart Study data (FHS). Three applications used new methods to assist in interpretation of results. These were applications to the Dys [3], FHS [6] and Alz [8] studies. Two of these [3, 6] for the first time carried out analysis with a single set of realized inheritance vectors or resulting pairwise ibd estimates in conjunction with multiple, simulated trait data sets to provide relatively fast empirical p-values for complex trait analyses. Three applications contain extensive comparisons between our MCMC-based and standard linkage analysis methods, thus testing and demonstrating their accuracy and computational efficiency. These were applications to FHS [6], Guam [10] and FCHL [14].

Two applications demonstrate the efficiencies in use of classes of equivalent *ibd* graphs [20] for analysis of large pedigrees [21,23], with up to 2-orders of magnitude gain in speed. Three applications identified challenges with use of the increasingly dense marker data available, including use of both dense SNPs [6,14,22] and modern sequence data [22], with such challenges leading to the imputation and error-detection approaches discussed above [24,25]. Finally, two applications demonstrate the effectiveness of methods that allow identification of inherited genomic segments that then inform selection of subjects for sequencing [21,27] or to assist in identification of a causal locus [22].

6 Software development: Aim 6

We have continued to develop our MORGAN software, integrating new programs for linkage detection [1, 2], new MCMC sampling methods [13], and many other computational improvements into the final (V2.9) releases of MORGAN-2 [S1]. The MORGAN Tutorial and Examples have also been fully updated to MORGAN 2.9 [S2]; the tutorial is available online and also in several download formats. Our major effort has been in the development, documentation, and distribution of MORGAN-3 [S3]. MORGAN-3 is now fully functional; V3.0.2 is recently released. While, for older programs, the user may detect little difference between MORGAN-2 and MORGAN-3, the underlying framework is significantly altered, since trait phenotypes are separated from trait loci, permitting several loci to contribute to a trait, as well as other needed complexities. MORGAN-3 also has much greater capacity for handling large numbers of genetic markers, with new options for lod-score computation. Newer methods for assessment of linkage findings [4, 15] are included only in MORGAN-3, as also is the *gl_auto* program for the generation of marker-based *ibd* graphs [12,20,21] and related programs for use of these *ibd* graphs in subsequent trait-data analysis. The tutorial for MORGAN-3, including new examples files, is in preparation for release.

As described above we have also released software tools SimSuite [S7], IBD_Haplo [S4], a related R-package, IBDHaploRtools [S5], and our *IBDgraph* software [20,S6] for determining equivalence of *ibd* graphs. Our IBD_Haplo software is also a component of MORGAN-3, facilitating the combination of population-based and pedigree-based *ibd* in these trait-data analyses [18, 19]. The IBDHaploRtools package and *IBDgraph* software are currently separately released, but we are working towards better interface between these tools and MORGAN-3.

3 Research Plan

The *ibd* graph approach has become central to our analysis methods, and will be the focus of continuing research. Our marker-based MCMC sampling methods have become quite efficient, and the compact storage of *ibd* graphs and the *IBDgraph* reduction provides for effective trait analyses. However, the potential remains for more direct sampling of *ibd* graphs and the recombination break-points that cause changes in *ibd* along a chromosome. We will explore this potential, and evaluate alternate methods of sampling *ibd* within and among pedigrees. We will continue to develop and evaluate methods for merging between-pedigree *ibd* inferred from dense markers with within-pedigree *ibd* sampled using sparser framework markers, and explore the potential for using pedigree-based haplotyping to inform between-pedigree inferences. Our current IBDmerge analyses are preliminary, and much remains to be accomplished.

Our main approach to using realized *ibd* in the analysis of trait data has been through the efficient computation of lod scores [17,12] although we have also developed other tests for linkage detection [1,2] and gene localization and resolution [4] and extended oligogenic linkage analyses [9]. An advantage of the *ibd* approach in the analysis of complex traits is that it integrates over genetic heterogeneity. It permits the detection of multiple causal loci affecting complex traits, and gains power from combining the effects of alternative segregating causal variants within loci. In addition to continuing lod-score approaches, we will develop more trait-mode robust test-statistics for linkage detection and resolution using the correlation of shared segments of *ibd* with phenotypic similarity. We will evaluate these methods, and compare their performance with that of direct model-based statistics.

We will continue to develop methods for combining chromosomally sparse data on many pedigree members with chromosomally dense data that is sparse on the pedigree, with particular reference to genotype and haplotype imputation, selection of individuals for sequencing, and for error detection. A key assumption in these analyses, as in the initial MCMC realization of *ibd* graphs, is that marker allele and local haplotype frequencies are known. We will take two general approaches to address this issue. First, for the framework markers, we will use reweighting of MCMC realizations [17] to determine the sensitivity of *ibd* graphs to allele frequencies within chromosomal regions of interest. This may be of particular importance where a rare variant local haplotype is segregating in a pedigree. Second, we will develop statistically sound methods to combine within-study information on frequencies of allelic and haplotypic variants with relevant publicly available data sources such as HapMap or 1000-Genomes.

In addition testing our methods via simulation studies using publicly available data, a key focus will be the continued testing on data from real pedigree studies, available to us through Dr. Wijsman's collaborations. Specifically, the four studies involving Alzheimer's disease (Alz) [8], FCHL [14], Dyslexia (Dys) [3] and Autism (Aut) [23] all have ongoing data collection, and all include larger pedigrees requiring MCMC analysis. Three, (Alz, FCHL and Aut), have dense SNP genotypes and will have next-generation exome sequence data on a subset of the individuals. Two of the studies (FCHL and Dys) have a wide range of well-defined quantitative phenotypes. The challenges of real marker and trait data include biological and phenotypic complexity, missing data, data error, and the need to integrate older data with that becoming newly available. These challenges lead to significant improvements in our methods. Conversely, our methods contribute to data analyses in these studies, providing ongoing synergistic advances.

We will continue to develop our MORGAN-3 software, implementing and releasing new methods, as well as continuing the improvement in the underlying structure of the package, allowing for greater flexibility of approaches, and the easier integration of data at different scales, both chromosomally and over pedigrees. We will integrate our *IBDgraph* library [S6] into MORGAN-3, to provide an integrated analysis stream from sampling the *ibd* graphs to the analysis of trait data. We will develop, document and release other related software, currently in early stages of development and testing. This includes *GIGI* for genotype imputation and error detection [P2], *HyperLod* for oligogenic trait-model likelihood computations [P3], and *IBDmerge* for merging *ibd*-graphs among pedigrees [18,19].

5 Progress Report Publication List

5.1 Publications resulting from this award; 2008–2011.

(i) Papers in refereed journals

1. Basu S, Di Y, and Thompson EA. (2008) Exact trait-model-free tests for linkage detection in pedigrees. *Annals of Human Genetics* **72:** 676–682. PMCID: PMC2574967.
2. Basu S, Stephens M, Pankow JS, and Thompson EA. (2010) A likelihood-based trait-model-free approach to linkage detection of binary trait. *Biometrics* **66:** 205–213. PMCID: PMC3118475.
3. Brkanac Z, Chapman NH, Igo RP Jr, Matsushita MM, Nielsen K, Berninger VW, Wijsman EM, and Raskind WH. (2008) Genome scan of a nonword repetition phenotype in families with dyslexia: evidence for multiple loci. *Behavior Genetics* **38:** 462–475. PMCID: PMC2853749.
4. Di Y, and Thompson EA. (2009) Conditional tests for localizing trait genes. *Human Heredity* **68:** 139–150. PMCID: PMC3022037.
5. Igo RP Jr, and Wijsman EM. (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.
6. Marchani EE, Di Y, Choi Y, Cheung C, Su M, Boehm F, Thompson EA, and Wijsman EM. (2009) Contrasting IBD estimators, association studies, and linkage analyses using the Framingham data. In "Genetic Analysis Workshop 16." *BMC Proceedings* **3**(Suppl 7):S102. PMCID: PMC2795873.
7. Marchani EE, Callegaro A, Daw EW, and Wijsman EM. (2009) Combining information from linkage and association methods. *Genetic Epidemiology:* **33**(Suppl 1): S81-S87. PMCID: PMC2910520.
8. Marchani EE, Bird TD, Steinbart EJ, Rosenthal E, Yu CE, Schellenberg GD, and Wijsman EM. (2010) Evidence for three loci modifying age-at-onset of Alzheimer's disease in early-onset PSEN2 families. *American Journal Medical Genetics: B Neuropsychiatric Genetics* **153B:** 1031–1041. PMCID: PMC3022037.
9. Rosenthal EA, and Wijsman EM. (2010) Joint linkage and segregation analysis under multiallelic trait inheritance: Simplifying interpretations for complex traits. *Genetic Epidemiology* **34:** 344-353, PMCID: PMC2914272.
10. Sieh W, Choi Y, Chapman NH, Craig UK, Steinbart EJ, Rothstein JH, Oyanagi K, Garruto RM, Bird TD, Galasko DR, Schellenberg GD, and Wijsman, EM. (2009) Identification of novel susceptibility loci for Guam neurodegenerative disease: Challenges of genome scans in genetic isolates. *Human Molecular Genetics* **18:** 3725–3738. PMCID: PMC2742398.
11. Thompson EA. (2008) The IBD process along four chromosomes. *Theoretical Population Biology* **73:** 369–373. PMCID: PMC2518088.
12. Thompson EA. (2011) The structure of genetic linkage data: from LIPED to 1M SNPs. *Human Heredity,* **71:** 88–98. (PMC journal – in progress).
13. Tong L, and Thompson EA. (2008) Multilocus lod scores in large pedigrees: Combination of exact and approximate calculations. *Human Heredity* **65:** 142–153. PMCID: PMC2701716.
14. Wijsman EM, Rothstein JH, Igo RP Jr, Brunzell JD, Motulsky AG, and Jarvik GP. (2010) Linkage and association analyses identify a candidate region for apoB level on chromosome 4q32.3 in FCHL families. *Human Genetics* **127:** 705–719. PMCID: PMC2877194.

(ii) Published conference proceedings and book chapters

15. Thompson EA. (2008) Uncertainty in inheritance: assessing linkage evidence. Proceedings of the Joint Statistical Meetings, Salt Lake City. Pp. 3751–3758.

16. Thompson EA. (2009) Inferring coancestry of genome segments in populations. *Invited Proceedings of the 57th Session of the International Statistical Institute*, Durban, South Africa.
17. Thompson EA. (2011) Chapter 13: MCMC in the analysis of genetic data on related individuals. In *Handbook of Markov Chain Monte Carlo* S. Brooks, A. Gelman, G. Jones, and X. Meng (Eds). Chapman & Hall/CRC Press. Pp. 345–367.
18. Thompson EA, and Glazner CG. (2011) Gene coancestry in pedigrees and populations. *Contributed Proceedings of the 58th Session of the International Statistical Institute*, Dublin, Ireland.

(iii) Submitted Papers and Software Technical Report.

19. Glazner CG, and Thompson EA. (2011) Improving pedigree-based linkage analysis by estimating coancestry among families. *Statistical Applications in Genetics and Molecular Biology:* submitted.
20. Koepke HA, and Thompson EA. (2010) Efficient testing operations on dynamic graph structures using strong hash functions. Technical report No. 567, Department of Statistics, University of Washington.
21. Marchani E, and Wijsman EM. (2011) Estimation and visualization of identity-by-descent within pedigrees simplifies interpretation of complex trait analysis. *Human Heredity;* submitted.
22. Rosenthal EA, Ronald J, Rothstein J, Rajagopalan R, Ranchalis J, Wolfbauer G, Albers JJ, Brunzell JD, Motulsky AG, Reider MJ, Nickerson DA, Wijsman EM, and Jarvik GP. (2011) Linkage and association of phospholipid transfer protein activity to LASS4. *Journal of Lipid Research;* submitted.

(iv) Meeting Abstracts not yet represented by submitted or published papers

23. Chapman N, Estes A, Munson J, Bernier R, Webb SJ, Rothstein J, Schellenberg G, Dawson G, and Wijsman E. (2010) Genome-wide linkage analysis of flexibility/insistence-on-sameness in multiplex families with Autism spectrum disorders. American Society of Human Genetics Meeting (abstract).
24. Cheung CYK, Thompson EA, and Wijsman, EM. (2010) In Silico Genotype imputation on large pedigrees. International Genetic Epidemiology Society Meeting (abstract). Winner of Williams award for best pre-doctoral presentation.
25. Cheung CYK, Thompson EA, and Wijsman EM. (2010) In Silico Genotype imputation on large pedigrees. American Society of Human Genetics Meeting (abstract).
26. Glazner C, Brown MD, Cai Z, and Thompson EA. (2010) Inferring coancestry in structured populations. West North American Region of the IBS Annual Meeting (abstract), Seattle, WA.
27. Marchani E, and Wijsman EM. (2010) Selective sequencing for efficient fine-mapping of disease loci. American Society of Human Genetics Meeting (abstract).
28. Thompson EA. (2008) Testing Hardy Weinberg Equilibrium. Plenary Hardy-Weinberg Centenary Session: *American Society of Human Genetics,* Philadelphia PA.
29. Thompson EA, and Zheng C. (2011) Modeling IBD processes along chromosomes in populations. West North American Region of the IBS Annual Meeting (abstract), St. Luis Obispo, CA.

(v) Papers in preparation

These papers are listed here for ease of citation/description. They will be submitted in 2011.

P1. Brown MD, Glazner CG, Zheng C, and Thompson EA. Inference of coancestry in populations; assessment of models and methods.
P2. Cheung CYK, Thompson EA, and Wijsman, EM. Genotype imputation on large pedigrees.
P3. Su M, and Thompson EA. Lodscores for oligogenic traits through use of multilocus identity by descent.

5.2 Project-generated resources

Compressed tar files of the source code of all software developed under this award are made available to interested researchers via the web. 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.

The following releases have been made since the start of the current award period:

S1. MORGAN 2.9: The first full release of MORGAN 2.9 (the final version of MORGAN-2) was in August 2008, and the final release was in November 2009.

S2. MORGAN 2.9 Tutorial and Examples: Tutorial and Examples for MORGAN 2.9 were updated December 2009, and a revised full update was released September 2010.

S3. MORGAN 3.0: Development of MORGAN-3 has proceeded from beta releases in March 2008 and March 2009, to the first full release in November 2009, with update versions 3.0.1 and 3.0.2 in September 2010 and April 2011, respectively.

S4. IBD_Haplo 2.0; A package for *ibd* inference from population data. First release was in December 2009, with version 2.0 released August 2010. See [26].

S5. IBDhaploRtools 1.1: An R-package to analyze the output of IBD_Haplo; includes example data files and tutorial. First release was in March 2011, with version 1.1 released May 2011.

S6. IBDgraph 2.0: A package to determine equivalent *ibd* graphs over realizations and across the chromosome. First release was in March 2010, with version 2.0 released August 2010. See [20].

S7. SimSuite Ver 1.0: A collection of scripts, with documentation and examples, for marker or trait resimulation. Released March 2009. See [5].