Project 4: Resolving complex traits through inferred coancestry of genome segments

Abstract

The overall objective is the development of methods for the enhanced detection and resolution of genes contributing to complex quantitative genetic traits observed in individuals not known to be related. The approach will be through using dense SNP marker data for the detection and estimation of segments of gene identity by descent (*ibd*) shared among sets of individuals. Locus-specific inferred *ibd* among individuals will be analyzed in conjunction with their phenotypic similarities and differences, in order to detect and resolve causal loci. We will develop and assess hidden Markov models (HMM) and methods for detection of *ibd* genome segments between pairs of members of populations from dense SNP data or sequence variants. We will assess the effects on performance of our methods of linkage disequilibrium, data error and copy-number variants, and the efficacy of prior haplotype imputation, data cleaning, and screening for regions of allelic similarity. We will extend our models and methods to the inference of *ibd* among larger sets of chromosomes using both HMM and coalescent models, and develop Markov chain Monte Carlo methods for sampling of *ibd* genome segments, conditional on dense SNP marker or sequence variant data in candidate gene regions. We will develop and assess methods for analyzing trait data on individuals conditional on the patterns of *ibd* genome segments inferred among them, by assessing location-specific levels and regional chromosomal extent of *ibd* segments among sampled chromosomes in relation to quantitative trait values. We will assess our methods and compare with alternative approaches, by first testing methods in simulated population structures, where latent *ibd* is known, but in which founder haplotypes are provided by real-data population samples. Then, in real data sets available to us, where latent *ibd* is unknown, we will compare results of our methods with those of other approaches developed both within the P01 group and elsewhere. We will develop software implementing our methods, and document, distribute and support this software.

Specific Aims

The overall objective is the development of methods for the enhanced detection and resolution of genes contributing to complex quantitative genetic traits. The approach will be through using dense SNP marker data for the detection and estimation of segments of gene identity by descent (*ibd*) shared among sets of individuals. These individuals are members of a population, but any relationships among them are not close and pedigree relationships are unknown. Locus-specific inferred coancestry among individuals will be analyzed in conjunction with their phenotypic similarities and differences, in order to detect and resolve causal loci.

1. Inference of pairwise *ibd* between individuals in populations

We will develop and assess *hidden Markov models* (HMM) and methods for detection of *ibd* genome segments between pairs of members of populations by: • improvement and extension of preliminary models and methods for inference of *ibd* at marker locations from dense SNP genotype data, including methods for parameter estimation and tuning: • assessment of the effects of population subdivision and admixture, and of resulting allele frequency and local haplotype heterogeneity: • assessment of the effects of linkage disequilibrium (LD) on *ibd* inference, and of the efficacy of prior haplotype imputation: • assessment of the effects of data error and copy-number variants, and the efficacy of prior data cleaning, and of prior screening for regions of allelic similarity.

2. Inference of *ibd* among larger sets of chromosomes

We will develop and assess *Markov chain Monte Carlo* (MCMC) methods for sampling of *ibd* genome segments, conditional on dense SNP marker or sequence variant data in candidate gene regions, on collections of haplotypes in a population by: • extension of the *ibd* models of **(1)** above, to provide models of *ibd* among collections of chromosomes; • development of MCMC methods to sample segments of *ibd* under population-based *ibd* models, conditional on dense SNP or sequence variant data; • development of MCMC-based methods for parameter estimation including modeling and parameter estimation for data error; • development of MCMC methods based on the *ancestral recombination graph* (ARG), and comparison of results under the *ibd* and ARG approaches; • investigation of inclusion of LD into Monte Carlo based methods of *ibd* inference.

3. From inferred *ibd* segments to trait data analysis

We will develop and assess methods for analyzing trait data on individuals conditional on the patterns of *ibd* genome segments inferred among them, by development of: • methods assessing location-specific levels and regional chromosomal extent of *ibd* segments among sampled chromosomes in relation to trait value; • methods of genomic control using genome-wide levels of *ibd* inferred in the same set of sampled chromosomes;
• model-based methods of analysis of quantitative trait data, computing trait data likelihoods given *ibd* inferred or sampled conditionally on marker data; • methods that translate measures of uncertainty in the inferred *ibd* among population members to measures of confidence in resulting trait inferences.

4. Methods testing, assessment and comparison

We will assess our methods and compare with alternative approaches, by: • testing methods on our simulated population structures, where latent *ibd* is known, but in which founder haplotypes are provided by real-data population samples; • testing *ibd* inference methods on real data sets available to us, where latent *ibd* is unknown, comparing results with those of other approaches both within the P01 group and elsewhere;
• testing *ibd* inference methods and consequent trait-data analyses on quantitative traits constructed within real SNP marker haplotypes; • use data available under the Program Project to assess trait-analysis methods on gene expression profiles relative to *ibd* inferred from sequence variants in candidate gene regions, comparing results with those of other approaches developed by the P01 group.

5. Software development

We will extend and enhance our preliminary *IBD_Haplo* software, by • implementation of programs and methods developed under **(1)-(3)** using both R-packages and C; • insure compatibility of input and output formats with other software of the Program Project; • documentation, distribution and support of the software; • development of web-based tutorial materials and examples.

1 Significance

There have been 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. However, while very large population-based designs are yielding results (Wellcome Trust Case Control Consortium 2007), it is becoming well recognized that the individuals of a case-control study are never truly ”unrelated”, and that relationships among individuals, whether known or unknown, are key in the analysis of phenotypic information in order to resolve complex genetic traits (Visscher et al. 2008; Yang et al. 2011).

Gene identity by descent (*ibd*) (Cotterman 1940) provides the fundamental framework for the analysis of phenotypic data on individuals. It is the descent of segments of DNA from common ancestors to extant individuals that provides for all genetically mediated phenotypic similarities, since such *ibd* genome has high probability of being of the same allelic type. The allelic associations maintained by linkage and resulting in linkage disequilibrium (LD) are likewise a result of coancestry. Conversely, modern genetic marker data, such as dense SNP data, permits the detection of segments of DNA shared *ibd* among extant individuals. The power of *ibd* detection lies in the fact that *ibd* segments in remote relatives are rare but not short (Donnelly 1983). Thus, for remote relatives, separated perhaps by 20 to 40 meioses, *ibd* genome segments are easily detected, and are typically much longer that the extent of LD present in the population.

Over the last five years, there has been a dramatic increase in the genomic density of genetic marker data. On a genome-wide scale there are data for $\sim$10M SNP markers (The 1000 Genomes Project Consortium 2010), while in targeted candidate regions and even for whole exomes sequence data are becoming increasingly available (Ng et al. 2009). These data are available on ever-increasing numbers of individuals both in disease studies and in population samples (Novembre et al. 2008). Such data challenge current statistical and computational methodologies, but also present new opportunities. In particular, using dense informative marker data, accurate estimates of segments of genome shared *ibd* among sampled individuals have become computationally feasible and statistically practical.

Inference of *ibd* has typically been restricted to pairs of individuals (Purcell et al. 2007; Browning and Browning 2010), or even to pairs of haplotypes (Leutenegger et al. 2003; Browning 2008). In order to gain the full power of an *ibd* approach it is necessary to consider the joint patterns of *ibd* among sampled haplotypes in a given region of the genome. When *ibd* patterns are clear, joint *ibd* can be constructed from pairwise inferences. Where there is uncertainty, methods that can deal directly with multiple chromosomes have an advantage. Previous approaches to joint analysis of allelic associations over multiple chromosomes have modeled latent ancestral similarities (Scheet and Stephens 2006) or taken a coalescent approach (Zöllner and Pritchard 2005). Our proposed approach that models directly the *ibd* pattern among multiple chromosomes or individuals has both computational and statistical advantages.

Analyses of population data have used allelic and haplotypic associations to assess evidence for genes associated with disease. However, these associations are reflections of *ibd* among individuals of the study, and the extent of *ibd*, both among individuals and along the chromosome, provides more direct evidence for involvement of genes in contributing to a trait. Haplotypic associations often require selection of variants to characterize a region, while all variants may be used in the assessment of *ibd*. Allelic association approaches have significant difficulties with rare variants (Dickson et al. 2010), but sharing of rare variants and haplotypes provides powerful evidence of *ibd*, and, within a causal locus, different rare variants may be shared by case subsets, allowing an *ibd* approach to implicate the locus where an association approach can not. The combination of close and remote relatives, through inferred *ibd*, has the potential to provide both linkage detection power and resolution for complex traits.

This project complements and is complemented by the other projects of this Program Project proposal. As for the other three methodological projects, this project addresses the problems of “relatedness” in resolving genes affecting complex quantitative traits, using using dense genetic marker data on individuals not known to share coancestry. In any large sample of individuals, there is necessarily significant structure, due to demographic diversity, heterogeneous ethnicity and cryptic relatedness. Together with data available through the Wellcome Trust Case Control Consortium (2007) and the NIH GENEVA studies (Weir; Project 1), the cohort-based data to be generated by Gibson (Project 3) on sequence variants and quantitative gene expression in 96 genes will provide an invaluable resource for the development of methods that can be used to quantify the genetic architecture of complex traits in diverse populations.

The projects of Weir (Project 1) and Visscher (Project 5) both consider relatedness and structure in a genome-wide average sense, considering the impact of this relatedness on “missing heritability” and in resolving gene-gene and gene-environments interactions and regulatory effects. This project (Project 4) and that of Browning (Project 2) focus on coancestry of chromosome segments; while such segments may exist genome-wide, the measure of any particular instance is specific to the local genome region. Their focus is on using inferred coancestry (*ibd*) as a means of addressing allelic and local haplotypic heterogeneity at underlying quantitative trait loci associated with disease. While Browning (Project 2) has focused on modeling population-level LD to improve *ibd* inference, this project (Project 4) focuses more directly on modeling and inference of *ibd* in a set of individuals; these two approaches are strongly complementary. All four methodological groups will develop software implementing their different approaches. Through the synergistic interactions of these groups, and in collaboration with the Computational Core, we expect the emergence of publicly available software that can address many aspects of resolving the genetic architecture of complex quantitative traits from individual data from diverse structured populations.

2 Innovation

It has long been understood that gene identity by descent (*ibd*) underlies all genetically mediated patterns of phenotypic similarity among related individuals (Cotterman 1940). However, methods of genetic association analysis have focused on allelic correlations, and, apart from pairwise *ibd* used in regression and variance component approaches (Haseman and Elston 1972; Almasy and Blangero 1998), the methods of genetic linkage analysis have typically not focused directly on gene *ibd*. A major innovation of our work, both in pedigrees and in populations, is the focus on *ibd* not only at specific marker locations, but across the genome. Once patterns of joint patterns of *ibd* are known, the marker data and any information on relationships among individuals observed for a phenotype of interest are no longer needed in subsequent trait analyses.

While Thompson may be best known for her work in the analysis of genetic data on pedigree structures (Thompson 2000), she has a long record of working in the area of joint patterns of gene *ibd* among relatives (Thompson 1974), likelihood estimation of relationships from genetic marker data (Thompson 1975a), and inference of the coancestry structure among populations (Thompson 1975b). More recently she has considered allelic associations or linkage disequilibrium (LD) in populations in terms of population coancestry (Thompson and Neel 1997). With students, she has incorporated coancestry into LD mapping (Graham and Thompson 1998), analysed extent of segments of genome shared *ibd* between remote relatives (Donnelly 1983; Chapman and Thompson 2003), investigated the effects of population growth and subdivision on the distribution of such segments (Chapman and Thompson 2002), and developed methods for inference of *ibd* within individuals whose parents are not known to share coancestry (Leutenegger et al. 2003).

The standard paradigm, both in pedigrees and in populations, has been to evaluate test statistics at marker positions, even where marker data are used jointly (for example, as haplotypes) in the computation of these test statistics. As marker data become ever denser, this is no longer practical. A major advantage of working in terms of the *ibd* is that the patterns of *ibd* among sampled individuals remains constant over many markers. Only where there is a change in *ibd* is it necessary to re-evaluate linkage evidence. In the pedigree context, we have developed methods to store these *ibd graphs* by chromosome segment rather than by marker (Koepke and Thompson 2010), together with IBDgraph software to identify *ibd* graphs and their chromosomal extent. The IBDgraph software can equally be applied to patterns of *ibd* inferred from population genetic marker data, greatly increasing efficiency of trait-data analyses.

An additional shift in genetic data analyses in recent years has been towards increasingly complex traits. Use of quantitative trait measures and inclusion of covariate information increase power of linkage detection (Wijsman and Amos 1997), but more complex models raise issues of model estimation and sensitivity. Again, the use of marker data to first inform patterns of gene *ibd* among individuals, and then to analyze trait data conditional on these *ibd* patterns provides a major advantage. A single analysis of marker data, stored compactly as inferred *ibd* patterns across the genome, can be used for multiple test locations, multiple trait models, and for multiple traits observed on the same set of individuals.

Finally, both in pedigrees and in populations, there has been a shift from exact computation of probabilities, likelihoods, and test statistics, towards Monte Carlo estimates of these, either for computational feasibility (Thompson 2000) or to integrate across diverse information (Stephens and Balding 2009). Typically statistics are averaged to provide an expectation or posterior probability. However, there is additional information beyond an average in a set of Monte Carlo realizations, of, for example, the across-chromosome *ibd* graphs. The distribution of *latent p-values* derived from these realizations provides measures not only of the strength of the evidence, but also the uncertainty in that measure (Thompson and Geyer 2007). This approach has been applied in the pedigree context (Thompson 2008b) to assess uncertainty in lod-score evidence for linkage. The same approach can be applied to marker-based realizations of *ibd* in the population paradigm.

3 Approach

The definition of identity by descent (*ibd*) has been discussed since the early work of Sewall Wright, who considered both correlations between uniting gametes and probabilities of descent of these gametes from a common ancestor, these being equivalent in the case of an idealized infinite population (Wright 1922). However, in real populations the reference population may depend on the purpose of the analysis, and both correlations relative to an extant population and probabilities of coancestry have an important role in the analysis of genetic variation (see Wright (1969), Chapter 12).

In the pedigree context, *ibd* is normally defined relative to the founders of the pedigree, but even here the issue of the base population can arise. For example, it may happen that subsets of a collection of pedigrees derive from clearly different ethnic backgrounds, and there is clear within-subset haplotypic coancestry between pedigrees (Edwards et al. 2008). When pedigrees are placed in their population context, and the possibility of unknown relationships among the pedigrees is entertained, the definition of a “founder” of a pedigree is no longer clear.

In the population context, Powell et al. (2010) have recently argued that in large-scale association studies the primary use of *ibd* is to impute allelic types at unobserved markers, and that then a definition of relatedness relative to the extant population is preferred. However, missing marker imputation is far from the only purpose of *ibd* inference. Coancestry resulting in *ibd* segments of genome in current individuals can aid haplotype imputation (Kong et al. 2008) and detect aspects of population structure not found by other methods (Browning and Browning 2010). Moreover, where there is heterogeneity of variants within a causal locus, an *ibd* approach can detect excess segments of *ibd* sharing associated with a trait, automatically combining the groups of individuals who share the local haplotype of each variant. Marker-based allelic associations are proving a powerful tool in the estimation of relatedness and consequent trait associations (see Yang et al. (2011) and Visscher; Project 5). Inference of the chromosomal extent of *ibd* segments provides an additional dimension to such inferences, not dependent on the precise SNP markers observed and polymorphic in a data set.

While the exact base population for the definition of *ibd* may be unclear, *ibd* segments inherited intact from a common ancestor can often be detected. This is because *ibd* segments in remote relatives are rare but not short (Donnelly 1983). More precisely, for a pair of individuals separated by $k+1$ meioses, the probability of *ibd* at any genome location is 2^{-k} and the probability of sharing any of an autosomal genome length L Morgans is approximately $1 - \exp(-k2^{-k}L)$. However, any *ibd* segment is of length order k^{-1} Morgans. For example, for $k = 20$, the probability of sharing any autosomal genome segment is about 10^{-3}, but if such a segment exists its expected length is 0.05 Morgans, or about 5×10^6 base pairs (5 Mbp). Thus, for remote relatives *ibd* genome segments are easily detected, and are typically much longer that the extent of LD present in the population. In case populations, or populations ascertained for extreme values of a quantitative trait, such *ibd* segments have increased probability of occurring at causal loci (Thompson 1997).

3.1 Models and inference of *ibd* for pairs of individuals

A hidden Markov model (HMM) for the inference of segments of *ibd* between the two chromosomes of an individuals was presented by Leutenegger et al. (2003). The latent state of *ibd* or non-*ibd* switches according to a Markov process such that the pointwise probability of *ibd* is β and the expected length of an *ibd* segment is $1/\alpha(1-\beta)$. The data model is of Hardy-Weinberg genotype frequencies in non-*ibd* segments. In *ibd* segments, the two alleles should be of the same allelic type, and of type A_i with probability q_i, the population frequency. However, in order that a rare mistyping should not preclude an *ibd* segment, we mix a fraction $(1 - \varepsilon)$ of this required homozygous distribution with a small fraction ε of the Hardy-Weinberg frequencies. Inclusion of an “error” model such as this is important in any HMM; the data should serve to reweight probabilities of latent state but should not preclude them. However, inferences are usually robust to the exact form of the “error” model (see Section 4). Given marker data, allele frequencies, and parameter values, a standard HMM computation provides the conditional probability of the *ibd* state at all locations across the genome. Browning (2008) extended this model to allow for the linkage disequilibrium (LD) that arises when markers are dense, but assumed that phase was known, and did not allow for mistyping.

In fact, data come as genotypes of individuals, and in analyzing *ibd* between individuals one must consider at least their four chromosomes. In developing the PLINK software, Purcell et al. (2007) presented an HMM for inference of *ibd* between two individuals. However, their model did not allow for LD, nor for mistyping. Moreover, their *ibd* model assumes that there is no *ibd* between the two chromosomes within an individual, and the data-model presented assumes that observation of a given allele decreases the probability of the subsequent allele being of that same allelic type. Albrechtsen et al. (2009) extended the model to allow for pairwise LD, and Browning and Browning (2010) used their general BEAGLE LD model, but their *ibd* model was limited to 2 states at any location: any-*ibd* or no-*ibd* among the four chromosomes. While the work of Browning (Project 2) is focused on LD, using the powerful BEAGLE model and software, this project focuses instead on models for *ibd* segments resulting from coancestry of chromosomes and individuals.

An *ibd* model founded in population genetics is that of Balding and Nichols (1994) who used the Ewens sampling formula (ESF) as a model for the partition of n homologous genes into k *ibd* groups. The ESF is normally given in the form (Ewens 1972; Tavare and Ewens 1997)

$$\pi_n(a_1, ..., a_n) \quad = \quad (\Gamma(\theta)\theta^k n!)/\Gamma(\theta + n)) \prod_{j=1}^{n} (j^{a_j} a_j!)^{-1} \qquad (1)$$

where a_i is the number of *ibd* subsets of size i, so that $n = \sum_i i a_i$ and $k = \sum_i a_i$. This one-parameter model assumes only exchangeability of the chromosomes. In particular, the probability of *ibd* between the two chromosomes within an individual is the same as between chromosomes in different individuals. The number k of subgroups is a sufficient statistic for θ (Ewens 1972), which is thus (inversely) related to the population level of *ibd*. Specifically, $\beta = \pi_2(0, 1) = 1/(1 + \theta)$ is the kinship coefficient, the marginal probability of *ibd* between a pair of chromosomes. Equation (1) is easily reparametrized in terms of β.

Thompson (2008a) developed a model for *ibd* across the genome that has the ESF (equation (1)) as the marginal *ibd* model among $n = 4$ chromosomes, and has one additional rate parameter α for changes in *ibd* state. Parametrization in terms of the pointwise pairwise probability of *ibd* (β) and a single rate parameter (α) controlling expected lengths of *ibd* segments, permits appropriate choice of these parameter values for specific applications. The model for changes in *ibd* state allows single chromosomes to join or disassociate from larger *ibd* groups, but unfortunately does not allow all transitions in *ibd* state that occur in reality. A very simple but unrealistic model would be that potential changes in *ibd* state occur along the chromosome at rate α, and that at such points a new (possible unchanged) state is sampled according to (1). This model has the advantage that all state changes are possible, but transition probabilities are far from those realized by any population model when $n > 2$. In preliminary analyses (see Section 4), we have used a mixture of these two models with a fraction $(1 - \delta)$ for the model of Thompson (2008a) and δ of the simple unrealistic model.

We have studied *ibd* inference for pairs in individuals, under this augmented *ibd*-model, using a data-model that is a direct generalization to four chromosomes of that of Leutenegger et al. (2003). We have developed HMM software, IBD_Haplo, and have undertaken an extensive study to investigate the potential and limitations of this approach to inference of *ibd*. A brief summary of some of the results of this study are given in Section 4. Using a simulated population descent but with real-data haplotypes, with a SNP density of about 50 per Mbp, we find that with haplotypic data we can reliably detect *ibd* segments down to 0.5 Mbp, and with unphased genotypic data down to 1 Mbp. An important recent advance in our software is that it permits the use of partially phased data, improving performance in chromosome regions where phase can be estimated.

We propose more extensive testing and assessment of inference of *ibd* segments under this 4-chromosome model using IBD_Haplo, particularly in admixed and structured populations. We propose comparison of direct use of unphased data, with prior use of haplotype imputation (Browning: Project 2), and assessment of both the ease and the importance of prior phasing as a function of the level of LD. With phased data, we propose comparison of pairwise analysis of chromosomes, with joint analysis of sets of 4, in terms of computational and statistical efficiency.

As commented above, Browning (Project 2) focuses on powerful LD-modeling in the inference of *ibd* segments, whereas this project focuses on more realistic population-genetic based *ibd* modeling. Fitting a good LD model requires large numbers of individuals (Browning 2006), but in large cohort populations the *ibd* may be of simple form. In smaller or more heterogeneous populations, it may be impossible to fit a good LD model, and more complex patterns of interrelatedness, and hence *ibd*, may arise. It is not computationally practical to include both model aspects in large-scale applications (either genome-wide or involving many individuals), but it is important to assess the impact of the simplifying assumptions made. We propose extensive comparison of our full *ibd*-model no-LD approach, with the general LD but 2-state *ibd* of Browning and Browning (2010).

HMM computations are rapid for the 15 states of latent *ibd* among four chromosomes (or 9 between two genotypes); for example, a large chromosome analysis of 45 pairs of individuals requires only a few seconds CPU. However, genome-wide analysis of all pairs from a large case-control cohort becomes prohibitive, and prior screening becomes necessary. One simple screen involves copy-number-variant (CNV) detection (Itsara et al. 2009). We will extend our current IBD_Haplo tests of CNV detection to assess the impact and efficiency of prior screening for homozygous segments, incorporating intensity data to distinguish *ibd* segments from chromosomal aberrations (See Laurie: Project 1). More generally, screening for high levels of allelic similarity across a region may increase efficiency of *ibd* detection, but the resulting biases must be assessed.

3.2 Models for *ibd* among multiple chromosomes

While pairwise analysis is well suited to genome-wide analysis, inference of the joint patterns of *ibd* among a set of chromosomes is a more powerful approach, particularly where there is uncertainty in the inference. However the number of possible *ibd* states at any genome location increases very rapidly with the number of chromosomes (Nadot and Vayssiex 1973), and HMM methods and genome-wide analysis become computationally infeasible. Instead, we propose Markov chain Monte Carlo (MCMC) methods to sample *ibd* patterns among a set of chromosomes, over small genomic regions, for example regions of sequence variation in the exons of a candidate gene (Gibson: Project 3). An additional advantage of the MCMC approach is that incorporating LD by importance sampling reweighting may be feasible (Thompson 2011a).

As before, equation (1) provides the prior distribution on the numbers a_j of *ibd* groups of size j in a set of n chromosomes. Since for a given $(a_1, ..., a_n)$, the a_j groups of size j may be permuted and the j elements of each of the a_j groups of size j may be permuted, the number of unordered labeled partitions with given

$(a_1, ..., a_n)$ is $n!/\prod_j (j!)^{a_j} a_j!$. Hence the probability of each unordered labeled partition z of the n chromosomes is (Ewens 1972)

$$\pi_n(z) \;=\; \pi_n(a_1, ..., a_n)\prod_j (j!)^{a_j} a_j!/n! \;=\; (\Gamma(\theta)\theta^k/\Gamma(\theta+n))\prod_j \Gamma(j)^{a_j} \tag{2}$$

Again, also, we require a model of changing *ibd* state along a chromosome that maintains the marginal distribution (2). Although the model of Thompson (2008a) can be extended to any number of chromosomes, n, and retains its Markov property when applied to reduced genotypic *ibd* states among a set of $n/2$ diploid individuals (Thompson 2009), it becomes increasingly restrictive for larger sets of chromosomes. Chaozhi Zheng, the postdoctoral researcher in Thompson's current group, has recently proposed a better model, based on the *Chinese Restaurant Problem* distribution (Tavare and Ewens 1997). To model transitions from a current *ibd* state, first a *supplementary* chromosome is proposed as a singleton with probability $\theta/(\theta + n)$, and to join each group of size j with probability $j/(\theta + n)$. Then, one of the $n + 1$ chromosomes is selected for deletion, and, if not deleted, the supplementary chromosome is given the identity of the deleted chromosome. Under this model any one chromosome is permitted to move among groups, which provides a much wider class of permitted transitions along the chromosome, and a much better fit to transitions realized in our simulated population (see Section 4).

The data model of our current IBD_Haplo software is also easily generalized, with different *ibd* groups having independent allelic types and alleles within an *ibd* group having high probability of being observed as the same allelic type, but again allowing mistyping with some probability ε. We have initiated the development of MCMC sampling of latent *ibd* state, among sets of 10 and 20 chromosomes given SNP or sequence variants, for example for 100 SNPs over 100 kbp, but much further development is needed. We propose to enhance these MCMC sampling methods, and to develop software. We propose to test the performance of this approach, and to assess the balance of increased information from larger sets of chromosomes against the increased computational complexity. We will compare the accuracy of MCMC estimates based on 10 or 20 chromosomes with the results of exact HMM computations on subsets of size 4.

For our HMM 15-state *ibd* model, we have not estimated HMM parameters, but have instead focused on robustness to parameter values (see Section 4.4). An advantage of an MCMC approach is that parameters of the model can also be sampled, using a Bayesian approach (Stephens and Balding 2009). We propose to sample not only the latent *ibd* state but also the parameters θ of the ESF, α relating to rate changes in *ibd* and thus to recombination, and ε the data error rate. We will develop appropriate priors for these parameters, and sample their posterior distributions given SNP or sequence variants, and compare prior and posterior distributions to assess performance of the MCMC and information in the data about these parameters.

Assessment of accuracy of inferences also requires development. For four chromosomes, with 15 states of *ibd* we can hope to recover the correct state. For 12 chromosomes, with over 4×10^6 *ibd* states, this is impractical.

Suppose $\{\{1,4,5\},\{2\},\{3,6,7\},\{8,11,12\},\{9,10\}\}$ is the true state.
Then $\{\{1,4,5\},\{11,8,12\},\{3,7,6\},\{9,10\},\{2\}\}$ is also the true state,
and $\{\{1,4,5\},\{2,3,6,7\},\{8,11\},\{9,10,12\}\}$ is a 2-step change,
but $\{\{1\},\{4,5\},\{2,3,6,7\},\{8,9,11\},\{10,12\}\}$ requires more steps.

Determination of all states with 1 or 2 steps or a given state is feasible, and so this level of accuracy can be assessed, but it is as yet unclear whether this level of accuracy can be achieved with real data in real population cohorts.

An alternative approach to the coancestry of a sample of chromosomes at any locus is through the coalescent, which, considered across the genome, becomes the ancestral recombination graph or ARG (Hudson 1991). The *ibd* of the current chromosomes can be specified relative to any past time point; lineages that have coalesced by that time point are *ibd*, and the *ibd* partition of the chromosomes may be specified as before. A problem inherent in the *ibd* approach is the choice of reference population or ancestral time-point t, when combined with the model assumption that DNA segments that are not *ibd* relative to this timepoint are of independent allelic type. The coalescent framework relative to time t provides a natural *ibd* definition, with earlier time-points corresponding to higher population *ibd* values. However, the allelic types at that ancestral time-point t are not independent, due to their earlier coalescent ancestry.

This problem becomes acute in an MCMC approach, where parameter θ is sampled, rather than being fixed from external information on the degree of relatedness in the population. In this case the time-point t and hence the value of θ is essentially not identifiable. An interesting solution has been proposed by Dr. Chaozhi Zheng. The coalescent back to some time-point t is used to define the *ibd* partition, but the allelic types at that reference time point are not assumed independent. Instead, integration over the earlier coalescent ancestry provides a joint distribution on the allelic types in distinct lineages at time t; in essence we now have an informative prior on θ based on this earlier coalescent ancestry. We plan to implement and test this proposal, to see whether MCMC sampling of parameters will then provide appropriate posterior distribution of the parameters, and assess any improvement of estimates of *ibd* segments in the sample.

Rather than model the *ibd* process, one could also adopt a fully coalescent approach, and consider the latent ARG of the sample. Just as the *ibd* process along a chromosome is well approximated by the Markov process described above, the ARG has also a Markov approximation (McVean and Cardin 2005), which defines the sequence of coalescents of the sample along a chromosome. As in the case of the *ibd* model, the Markov approximation provides an accurate, computationally feasible, prior distribution, although MCMC methods sampling the ARG are computationally much more complex that those for the *ibd* partitions (Kuhner et al. 2000).

However there are significant differences between the probability distributions underlying the ARG and ESF frameworks. First, consider a point in the genome, and the coalescent of the sample at that point. Given the number k of lineages relative to time t, the distribution of the partition of the chromosomes into k groups is as for ESF (equation (1)), since this is a function only of the exchangeability of the chromosomes. However, the distribution of k differs, with that of the ESF having greater variance. Since ESF provides only a prior, to be used in inference given the sample, this greater variance may not be a significant drawback.

There is a greater difficulty in assessing discrepancies in *ibd* partitions under the two models. Consider, for example the two *ibd* partitions $\{\{1,2,6,7\},\{3,4,9\},\{5,8,10\}\}$ and $\{\{1,2,6,7\},\{3,4,5,8,9,10\}\}$. Under a coalescent framework these two states differ only in a single coalescent event joining the two groups $\{3,4,9\}$ and $\{5,8,10\}$. Under the *ibd* model, these states differ by three transitions, since each chromosome or lineage changes *ibd* group as a singleton event.

We propose to develop MCMC sampling methods for ARG framework as well as for the ESF framework, and to compare the two frameworks in their ability to detect segments of coancestry among the sampled chromosomes of case-control studies in candidate gene regions. We also propose to investigate incorporating LD into our models via importance sampling reweighting of MCMC realizations (Thompson 2011a).

3.3 The *ibd* graph and trait data analyses

A major challenge in genome-wide association studies is allelic heterogeneity at a causal locus, particularly where the relevant allelic variants are rare (Dickson et al. 2010). Methods of grouping variants and haplotypes have been proposed (Li and Leal 2008; Madsen and Browning 2009; Price et al. 2010), but these poorly-tagged and rarely-seen variants remain a major challenge. An advantage of an approach based on *ibd* is that it automatically integrates across all allelic variation; *ibd* provides a framework for haplotypic risks, not geared to specific alleles. This approach also combines information across extended haplotypes (say, of order 10^5 bp); again, *ibd* segments are rare but not short (Donnelly 1983). Even where only small numbers of cases share any particular haplotype, their likely *ibd* can be inferred, and excess *ibd* among cases, relative to controls and relative to other genomic regions, can provide strong evidence of a locus contributing to a trait.

We propose to use location-specific *ibd* as a signal for linkage. We have developed use of the *ibd*-graph (Thompson 2011b) to specify *ibd* among individuals observed for a trait. At any genome location this graph has labeled edges, each edge denoting a specific observed individual's qualitative or quantitative trait value, and unlabeled nodes representing *ibd* sharing among the edges (individual phenotypes) that impinge on the node (see Figure 1). In a population sample, there will typically be little *ibd* and the graph will consist of many unconnected edges (not shown in Figure 1), but in a case sample in the neighborhood of causal loci, there will be connections due to shared coancestry. As implemented in our IBDgraph software, the *ibd*-graph is defined not only at each genome location, but collectively over the chromosome, with base-pair or marker indexing (Koepke and Thompson 2010). Any feature of the graph (for example, *ibd* of a set of individuals) has a validity range of markers over which it extends. The IBDgraph software permits efficient querying and identification of *ibd* graph features and their validity ranges.

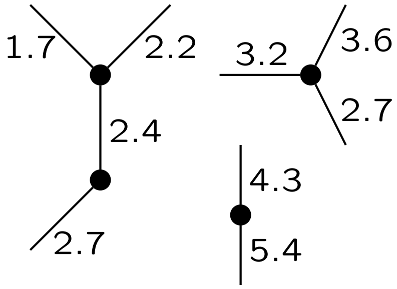

Figure 1: Example *ibd* graph at one genome location. Edges are labeled with the quantitative trait values of the individual represented by that edge. Unlabeled nodes show genome sharing among individuals. Only individuals sharing genome are shown. The *ibd* graph is a sufficient statistic for trait data analysis conditional on inferred *ibd*

Consider first the assessment of *ibd* for a binary trait, for example in a case-control study. While association studies may try to match cases and controls for their population background, and for their level of interrelatedness, this often proves impossible. Heterogeneity of populations, and ascertainment of case samples, may often lead to higher levels of relatedness among cases than among controls. It is therefore important to use genomic control (Devlin et al. 2001). That is, the levels of *ibd* inferred in specific genome regions should be compared, not between cases and controls, but among cases relative to the values for the same individuals for other genomic regions. For this purpose it will be important to have fast and accurate estimates of genomewide levels of *ibd* such as are provided by the methods of Weir (Project 1) and Visscher (Project 5). Note also that we do not propose that *ibd*-based methods should replace those of allelic association, but rather that the strength of *ibd*-based methods to integrate across allelic heterogeneity should provide clear evidence of genomic regions of interest. More detailed investigation of the precise allelic variants shared *ibd* can then follow.

For quantitative traits, phenotypic similarities among relatives are typically modeled via a variance component model (Almasy and Blangero 1998), using pairwise location-specific *ibd* probabilities, together with other heritable (polygenic) random effects and non-heritable (covariate) fixed effects. There are two drawbacks to this model. First, higher-order *ibd* among individuals can provide much more power. The sharing of genome *ibd* among multiple cases is a strong indicator that cannot be captured by a pairwise variance-component model. Second, the model is pointwise over the genome; the chromosomal extent of *ibd* sharing among a collection of individuals also provides valuable information about the localization and resolution of causal loci. Shared genome in different regions of a gene among different case subsets may be an indicator of allelic heterogeneity in a causal gene. We propose to develop models and methods for the analysis of quantitative trait data, including expression data (Gibson: Project 3), that use the framework of *ibd* inferred jointly among individuals from genetic marker (SNP or sequence) data observed in population samples.

In a model-based approach, once *ibd* is inferred from marker data, the marker data are no longer relevant, and the trait contribution of any genome location can be assessed (Thompson 2011b). We propose to model trait data Y conditional on *ibd*, and to use realizations of *ibd* conditional on marker data M, in a Monte Carlo likelihood approach:

$$\mathrm{P}(Y \mid M) \quad = \quad \sum \mathrm{P}(Y \mid \textit{ibd}) P(\textit{ibd} \mid M)$$

We will use the methods developed in 3.1 and 3.2 above to provide realizations of the *ibd*-graph among observed individuals, and then obtain the likelihood contributions from each realized *ibd*-graph. Monte Carlo realization of *ibd* has not only computational advantages, but also provides measures of uncertainty. Where *ibd* is more uncertain, there will be more variability among *ibd* graphs. Each realized *ibd* graph gives rise to a measure of evidence for particular genome regions to be implicated – e.g. a p-value. The distribution of p-values over realized latent *ibd* graphs is a latent (or fuzzy) p-value (Thompson and Geyer 2007), and can be used to assess the degree of certainty that the marker data provide in associating a phenotype with some chromosomal region. In developing tests of *ibd*-based measures of evidence for linkage, we propose to develop the latent-p-value approach to assess the statistical significance of findings implicating specific chromosome regions.

3.4 Model testing and assessment

Initial tests of methods will be on on simulated population ancestry structures, both from our population (Section 4) and under the coalescent (Section 3.2). Our simulation program permits realizations of populations of varying size and depth of generations. We have preferred this approach of simulating full population descent or sample coalescent ancestry to earlier approaches (Thompson 2009; Browning and Browning 2010) where *ibd* is artificially constructed by transferring segments of genome between individuals. Much of the uncertainty in *ibd* inference is around change-points in the the latent *ibd* state (preliminary results; not shown). Artificial transfer of segments for testing purposes likely creates artificially clear change-points in the resulting *ibd* inferences. Real-data haplotypes will be imposed on the resulting *ibd* structures; for this purpose the data available through the Program Project will provide a much greater variety of levels of LD and admixture structure to assess our methods than do our preliminary data of Section 4.

Pairwise (4-chromosome) analyses of *ibd* (Section 3.1) are quick and efficient and can be carried out on a genome-wide scale. Here we will use GENEVA and HapMap SNP data available to the Program Project. The data-cleaning, CNV detection, and other procedures developed under the GENEVA project (Laurie et al. 2010) will be used to pre-process data for maximal efficiency. We will use these data first on our simulated populations in which the *ibd* imposed by the simulation is known, and later in real-data analyses in which we do not impose any *ibd* beyond that unknown and naturally occurring.

Joint analysis of multiple chromosomes (Section 3.2) is computationally intensive, particularly if LD is to be incorporated, and thus better suited to detailed analysis of candidate gene regions. First, we will use dense SNP data from specific gene regions for preliminary testing, but we will move rapidly to the data on sequence variants developed by Gibson (Project 3).

For our testing of trait loci detection through inferred *ibd* (Section 3.3) we will again initially work with data in which the answer is known, either from prior studies (Wellcome Trust Case Control Consortium 2007) or by generating trait data from SNPs or haplotypes within our HapMap or GENEVA maker data. Our longer-term goal will be to apply these methods to the 96 quantitative expression traits generated by Gibson (Project 3) relating these traits to sequence variation in the relevant genes.

3.5 Software development

Our preliminary IBD_Haplo software has been released (Section 4), but requires much further development. We propose to improve this software, based in the HMM model of Section 3.1, and to provide documentation and tutorial examples. When all pairs of a set of individuals are analyzed using many 1000s of SNP markers, both input and output of IBD_Haplo are large. We propose to write a R package containing functions to create input files in required format, and to sort and summarize output files in various ways. We will also develop and distribute software implementing the MCMC methods of Section 3.2. While the MCMC process is much more complex, the ESF (equation (1)), the transition matrix construction, and the data and error models, are shared by the two approaches. We hope to combine these approaches in a single software package. It is unclear how central our IBDGraph software (also released 2010), will be to the analysis of the sparse *ibd*-graphs of population data Section 3.3, but to the extent it is useful, we propose to include that functionality in our new package also.

Of equal importance will be compatibility with the software of other projects, especially the well-developed BEAGLE software (Browning; Project 2). We propose to work with other project researchers, and through the Computational Core, to ensure that data files can be easily transferred between softwares, and that analysis outputs have shared formats providing for easy comparison, For the large array-oriented input and output files, we will adopt the NetCDF format used by Weir (Project 1) and other researchers. Overall, we will work to insure a coherent package that complements other capabilities.

3.6 Timeline

The three methodological components of this project will develop broadly in parallel, with student research associate Chris Glazner taking the lead in development of HMM-based models and methods (Section 3.1), postdoctoral researcher Dr. Chaozhi Zheng leading the development of Bayesian MCMC-based methods (Section 3.2) and the PI Thompson further developing *ibd*-based approaches to linkage detection and trait-data analysis (Section 3.3), as well as mentoring the research associates. Testing first on simulated or constructed data will follow (Years 2-3), and then application to real data sets (years 3-5) available to and through the Program Project. Software will be developed by these researchers, and integrated into a package with the assistance of research scientist Dr. Steven Lewis; final development of the documented software package will be in years 3-5 of the program, although specific programs should be released sooner.

4 Preliminary Studies: *ibd* estimation in sets of 4 chromosomes.

4.1 Background

Under an ARRA Competitive Supplement (funded 9/30/2009-9/29/2011) to our ongoing R37 GM046255 grant, we have been investigating methods for the estimation of gene *ibd* among individuals sampled from a population. The goal of that research is to combine 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. Additionally, within-pedigree information on phase of sampled individuals contributes substantially to the accurate estimation of between-pedigree *ibd* (Glazner et al. 2010).

Under R37 GM046255 we will continue to develop ibd-based methods for the detection and resolution of loci contributing to complex oligogenic quantitative traits, using data on known pedigrees (Thompson 2011b), and will incorporate developments made under the ARRA-funded supplement. However, the population-based methods we have initiated show great promise also for analysis of data in case-control studies and other studies of individuals not known to be related, particularly in populations where the degree of relatedness in the sample is substantial due to population structure, admixture, or history, or due to the sample ascertainment. We propose therefore to further develop and test these methods in the context of population data, as described in Section 3 above.

We here provide a brief summary of our results relating to *ibd* estimation from population data.

4.2 The test data, analysis models, and software

In our studies of performance of methods, we have used a simulated population of 7,000 diploid individuals (3,500 male; 3,500 female) generating the descent of a chromosome, typically 140×10^6 bp (140 Mbp) over 200 generations. This provides a broad range of *ibd* patterns among samples of chromosomes from the final generation, with *ibd* segment lengths ranging from a few bp to 3 Mbp. The pointwise probability of *ibd* between a pair of chromosomes is about 1.5%, and the pointwise probability of no *ibd* in a set of 4 chromosomes is about 90%. Along each set of 4 chromosomes, there are typically about 40 changes in *ibd* state. Although at any point in the genome only 90 to 100 founder genomes survive in the 200th generation, over the 140 Mbp chromosome 2,345 founder genomes are represented and as many as 1,400 founder genomes can be represented in a sample of only 10 individuals (20 chromosomes).

Using data from the Framingham study (FHS) (Cupples et al. 2009) we gathered a pool of 1917 real-data SNP haplotypes. Given the pedigree relationships specified none of these share coancestry. Discarding SNPs with significant missing data or minor allele frequencies less than 5%, we retain about 7,000 SNPs over 140 Mbp (50 SNPs per Mbp, on average). In a sample of 10 individuals (20 chromosomes), a unique haplotype from the pool is assigned randomly to each represented founder, to construct the SNP data on the 20-chromosome set. This process is repeated as needed.

We have implemented software (IBD_Haplo) to estimate latent *ibd* among sets of four chromosomes using a hidden Markov model (HMM). The model for the latent *ibd*, either for the 15 states of *ibd* among 4 phased chromosomes or for the 9 states among two (unphased) genotypes, is that developed by Thompson (2008a), augmented to allow, with small probability, transitions not permitted under the simple model (Thompson 2009). There are three parameters of the *ibd* process, the pointwise pairwise probability of *ibd*, β, the rate of change of *ibd* pattern along the chromosome, α, and the probability δ with which which "non-standard" transitions are selected. The data model assumes SNP allele frequencies equal to those in our chromosome pool, and allows for typing error (probability ε). The data analysis model does not incorporate LD.

For given values of the parameters β, α, δ and ε, our IBD_Haplo software analyzes sets of four chromosomes, or pairs of genotypes. The first version of IBD_Haplo was released to the web in 2009, and an improved version was released in Fall 2010 (www.stat.washington.edu/thompson/Genepi/pangaea.shtml).

4.3 Example results; LD is a reflection of *ibd*

Using the IBD_Haplo software, we estimated *ibd* among sets of 4 chromosomes in pairs of individuals. Typically, we run all 45 pairs in sets of 10 individuals. At each marker, a calling threshold of 90% was used to call an *ibd* state (including the state of no-*ibd*). When using the FHS pool of chromosomes, we estimated the state of no-*ibd* in only 78% of calls, although our simulated descent gave this state over 90% of the chromosome. Although this excess *ibd* is a false-positive in the framework of our simulated population, it is real in that it is true *ibd* resulting from cryptic relatedness or LD in the FHS chromosome pool. To assess this we fit an LD model using BEAGLE (Browning 2006), and generated a new chromosome pool with the same markers and marker locations using the fitted BEAGLE model. Very similar results were obtained, indicating the issue is likely one of the high degree of LD in this population. We then produced chromosome pools with reduced LD by random elimination of some of the split-nodes of the BEAGLE model. The following results are for one such case, in which LD extending over more than 5 markers (~ 0.1 Mbp) is reduced. With this level of LD, 89% of calls were for the no-*ibd* state. The "false-positive" rate decreases with decreasing LD, but for real-data situations is hard to estimate; our real-data pool of "unrelated" haplotypes undoubtedly contains many *ibd* segments.

The data were analyzed both as phased haplotypes and as a pair of genotypes. First considering *ibd* detection, we considered the proportion of markers in each *ibd* segment that called any state of *ibd* (Figure 2 (left)). Using haplotypic data, in segment lengths of at least 0.5 Mbp, 85% of markers gave the *ibd* call, and in segments at least 1 Mbp almost 100% of markers did so. Genotypic data are less powerful: recall our analysis model does not incorporate any LD information. The corresponding proportions of markers calling *ibd* in 0.5 and 1 Mbp segments were 60% and 85%. Although some small segments (less than 0.1 Mbp) were detected, many were missed. Interestingly, using haplotypic data, for segments over 0.5 Mbp, the results were little changed when considering calling the correct *ibd* state (Figure 2 (right)), although a few segments over 2 Mbp were called for the wrong state by up to 25% of markers in the segment. Small segments (less than 0.2 Mpb) are often called for the wrong *ibd* state, even where *ibd* is detected. Genotypic data were more prone to calling incorrect *ibd* states, due to the phase uncertainties. The proportions of markers calling the correct *ibd* state in segments lengths 0.5 and 1 Mbp are reduced to 40% and 75%.

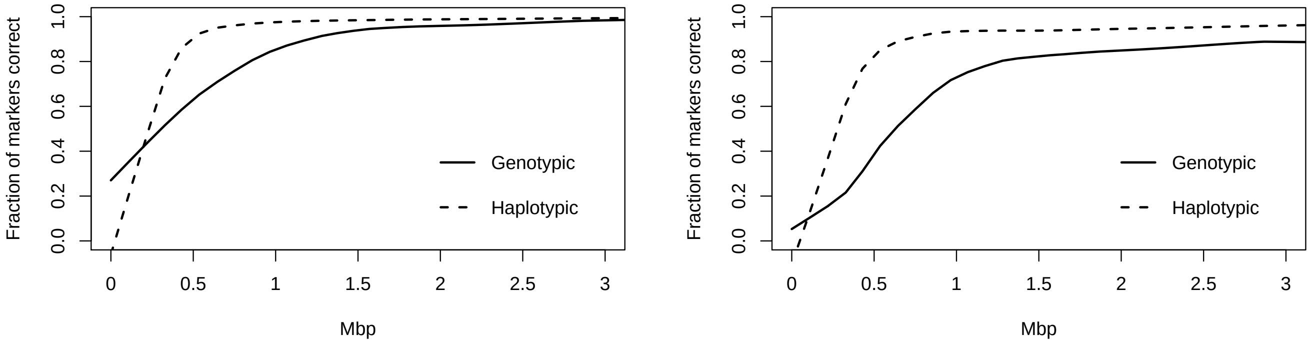

Figure 2: Curved fitted to ~900 *ibd* segments among all 45 pairs in a sample of 10 individuals, analysed both as haplotypes (4 chromosomes) and as genotypes (2 individuals). The left figure shows, by length of segment, the proportion of markers calling any *ibd* state, and the right figure shows the proportion calling the correct *ibd* state. A 90% probability was required to call an *ibd* state.

4.4 Testing model robustness and goodness of fit

Although Leutenegger et al. (2003) developed an EM algorithm for estimation of the HMM parameters (our α and β), we have not attempted this for our more complex model, but have instead undertaken analyses of robustness to parameter values, including allele frequencies. As is well known, for *ibd* inference based on few markers (and hence for small segments) there can be high sensitivity to allele frequencies (Ott 1992). However, we find this sensitivity is much reduced in larger segments. When the prior probability of *ibd*, β, is increased, obviously more *ibd* is called, but overall the segments of *ibd* detected were very robust to changes in β, and also to values of the rate parameter α for values giving a 10-fold range of prior expected lengths of *ibd* segments of 0.2 to 2 Mbp. With regard to the error parameter ε, it is important to allow for error, so that a single discrepant marker does not preclude an *ibd* region, but, as in other relationship inference (Sieberts et al. 2002), the precise error model and level assumed has little impact.

Most recently, we have studied the goodness of fit of actual and inferred transitions in *ibd* state to our model assumptions. While the model of Thompson (2008a) captures most realized transitions well, and provides a good prior model for the analysis of data on sets of 4 chromosomes, some transitions are not permitted; it is necessary to allow for these transitions by $\delta > 0$. On the other hand, the model component with proportion δ has no population genetic basis, and allows far too many transitions that do not occur in reality, for example from no-*ibd* to all-4-*ibd* between two close markers. Following these results we are considering both how small δ can be chosen, and also developing alternate models, as detailed in Section 3.2.

4.5 Software development

Our IBD_Haplo software, version 2.0 released Fall 2010, continues to improve, with input from users elsewhere as well as from within the group. Recognizing the advantage of phased data, as described above, our most recent version allows for partially phased data, so that phase information can be used in regions of the chromosome where phase is clear.

The IBD_Haplo program, because under development, has idiosyncratic input and it produces large output files. To assist colleagues we have written scripts to generate the input files to analyze, for example, all pairs among a set of individuals. To process output, we use R. Functions summarize inferred *ibd* call patterns in various ways, specify *ibd* segments, provide distributions of lengths, and so on. These functions are being put into an R-package, which will be released with the next version of IBD_Haplo.

Bibliography and References Cited

Albrechtsen A, Korneliussen TS, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genetic Epidemiology 33:266–274

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

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

Browning SR (2006) Multilocus association mapping using variable-length Markov chains. American Journal of Human Genetics 78:903–913

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

Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. American Journal of Human Genetics 86:526–539

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

— (2003) A model for the length of tracts of identity by descent in finite random matingpopulations. Theoretical Population Biology 64:141–150

Cotterman CW (1940) A Calculus for Statistico-Genetics. Ph.d. thesis, Ohio State University. Published in "Genetics and Social Structure", P.A. Ballonoff ed., Acaddemic Press, New York, 1974.

Cupples LA, Heard-Costa N, Lee M, Atwood LD (2009) Genetics Analysis Workshop 16 Problem 2: the Framingham Heart Study data. BMC Genetics 3(Suppl 7):S3

Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theoretical Population Biology 60:155–166

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLOS Biology 8:e1000294

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

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

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

Glazner C, Brown MD, Cai Z, Thompson EA (2010) Inferring coancestry in structured populations. Abstract, West North American Region of the IBS Annual Meeting

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

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

Hudson R (1991) Gene genealogies and the coalescent process. In R Dawkins, M Ridley, eds., *Oxford Surveys in Evolutionary Biology*, vol. 7, 1–44. Oxford University Press: Oxford

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

Koepke HA, Thompson EA (2010) Efficient testing operations on dynamic graph structures using strong hash functions. Technical report no. 567, Department of Statistics, University of Washington

Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T, Sulem P, Mouy M, F.Jonsson, Thorsteinsdottir U, Gudbjartsson DF, Stefansson H, Stefansson K (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. Nature Genetics 40:1068–1075

Kuhner MK, Yamato J, Felsenstein J (2000) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 156:1393–1401

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut L, Bhangale T, Boehm F, Caporaso N, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs K, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice K, Zheng X, Weir B (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic Epidemiology 34:591–603

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

Li BS, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. American Journal of Human Genetics 83:311–321

Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLOS Genetics 5:e1000384

McVean G, Cardin N (2005) Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society of London Series B 360:1387–1393

Nadot R, Vayssiex G (1973) Algorithme du calcul des coefficients d'identite. Biometrics 29:347–359

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. Nature Reviews Genetics 10:681–690

Ott J (1992) Strategies for characterizing highly polymorphic markers in human genemapping. American Journal of Human Genetics 51:283–290

Powell JE, Visscher PM, Goddard ME (2010) Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11:800–805

Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. American Journal of Human Genetics 86:832–838

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool-set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559–575

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 78:629–644

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

Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10:681–690

Tavare S, Ewens WJ (1997) The multivariate Ewens distribution. In *Discrete Multivariate Distributions*, 232–246. Wiley, New York, NY

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

Thompson EA (1974) Gene identities and multiple relationships. Biometrics 30:667–680

— (1975a) The estimation of pairwise relationship. Annals of Human Genetics 39:173–188

— (1975b) Human Evolutionary Trees. Cambridge University Press, Cambridge, UK

— (1997) Chapter 10: Conditional gene identity in affected individuals. In I Pawlowitzki, JH Edwards, EA Thompson, eds., *Genetic Mapping of Disease Genes*, 137–146. Academic Press, London, UK

— (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

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

— (2008b) Uncertainty in inheritance: assessing linkage evidence. In *Proceedings of the 2007 Joint Statistical Meetings*, 3751–3758. Salt Lake City

— (2009) Inferring coancestry of genome segments in populations. In *Invited Proceedings of the 57th Session of the International Statistical Institute*, IPM13: Paper 0325.pdf. Durban, South Africa

— (2011a) Chapter 13: MCMC in the analysis of genetic data on related individuals. In S Brooks, A Gelman, G Jones, XL Meng, eds., *Handbook of Markov Chain Monte Carlo*, in press. Chapman & Hall/CRC, London, UK

— (2011b) The structure of genetic linkage data: from LIPED to 1M SNPs. Human Heredity in press; accepted 4/14/2010

Thompson EA, Geyer CJ (2007) Fuzzy p-values in latent variable problems. Biometrika 90:45–60

Thompson EA, Neel JV (1997) Allelic association and allele frequency distribution as a function of social and demographic history. American Journal of Human Genetics 60:197–204

Visscher PM, Andrew T, Nyholt DR (2008) Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. European Journal of Human Genetics 16:387–390

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

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

Wright S (1922) Coefficients of inbreeding and relationship. American Naturalist 56:330–338

— (1969) Evolution and the Genetics of Populations. Volume 2: The Theory of Gene Frequencies. University of Chicago Press, Chicago, USA

Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. American Journal of Human Genetics doi:10.1016/j.ajhg.2010.11.011

Zöllner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169:1071–1092