Inferring coancestry of genome segments in populations

Thompson, Elizabeth
University of Washington, Department of Statistics
Box 354322,
Seattle WA 98195-4322, USA
E-mail: eathomp@u.washington.edu

1. Introduction

Identity by descent (*ibd*) underlies all similarities among relatives, and hence is the basis of linkage mapping both in pedigrees (Albers et al. 2008) and in populations (Te Meerman et al. 1995). As dense genomic marker data become increasingly available, *ibd* among observed individuals can be accurately inferred. This is leading to new methods of genetic analysis of complex traits whereby *ibd* is first inferred, and then trait data are analyzed conditionally on the inferred *ibd* (Di and Thompson 2008). In pedigrees, the pedigree structure provides a prior distribution on *ibd*, but for remotely related individuals in populations pedigree relationships are unknown. The strength of methods for inference of *ibd* from genomic data lie in the fact that, while remote relatives have low probability of sharing any of their genome *ibd*, the lengths of *ibd* segments, if present, extend on average over millions of base pairs (Donnelly 1983). In this paper we present models for inference of *ibd* among multiple genomes sampled from a population, given either haplotypic (phased) or genotypic (unphased) data on a set of individuals. For convenience we will refer to the underlying DNA as genomes, even when considering only *ibd* at a single genome location, to the allelic types of the DNA as haplotypes, and to the unphased pair of haplotypes as genotypes.

On a less dense genomic scale, Leutenegger et al. (2003) produced the first model to infer *ibd* among chromosomes in populations from genetic data at multiple linked loci. Although she considered only the two chromosomes within each individual, a key feature of her model is that it permits error in the observations. Browning (2008) considered dense data on a genomic scale. She again only considered pairs of chromosomes, and her model did not allow for error, but her key contribution was the incorporation of linkage disequilibrium (LD) among the dense genetic markers upon which inferences are to be based. Although the model of this paper could be extended to include LD (Thompson 2008a), it is not computationally feasible (as yet) to include both LD and multiple genomes. Also, since LD is a reflection of the coancestry we aim to infer (although at a longer time frame), it is questionable as to whether LD should be modeled. Finally, Purcell et al. (2007) considered estimation of *ibd* between two individuals, given genotypic data. However, this model also does not allow for data error or LD, and considers only *ibd* between the individuals and not between the two genomes of each. Since, in most human populations, our parents are at least as closely related to each other than each of us is to other members of a study population, this seems an inherently undesirable constraint.

The Leutenegger model for *ibd* is a two-parameter Markov model for changes in *ibd* along a chromosome; *ibd* is gained at rate g and lost at rate h, giving the first form of the two-state Markov rate matrix Q between non-*ibd* ($Z = 0$) and *ibd* ($Z = 1$):

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

There are two equivalent interpretations of this rate matrix in terms of population-genetic parameters α and β. The marginal probability of *ibd* is β, and the relative rate of gain vs loss of *ibd* is $g/h = \beta/(1-\beta)$. The parameter α measures a rate of change in *ibd* status along the chromosome, and hence the "typical length" of *ibd* segments. Specifically, *ibd* segment lengths are exponential with expected length

$(\alpha(1-\beta))^{-1}$. In reality, *ibd* is not Markov and expected segment lengths are heterogeneous, depending on the number of meioses from the common ancestor, but our objective here is a flexible "prior" model that will allow the data to inform the inference of *ibd*. An alternative interpretation of the model (1) is given by the final expression in the equation. Here, we have a model of breakpoints, occurring randomly and independently at rate α, and the intervals between segments are independently of type 0 (non-*ibd*) with probability $(1-\beta)$, and of type 1 (*ibd*) with probability β. Both these interpretations will be useful in developing the more general model of this paper. Finally, values of α and β are required for data analyses. Leutenegger et al. (2003) developed an EM-algorithm for parameter estimation in this model. However, here we regard them simply as tuning parameters, with β depending on the overall *ibd* level, and α on the typical *ibd* segment length we wish to detect.

Next, a model for the genotypic or haplotypic data given latent *ibd* is required. At each locus, allele frequencies q_i of alleles A_i are assumed known: in reality they can be well estimated from genotypic samples. The model of (Leutenegger et al. 2003) may be written

$$
\begin{array}{llllll}
 & P(A_iA_i|Z=0) & = & q_i^2 & P(A_iA_i|Z=1) & = & (1-\epsilon)q_i + \epsilon P(A_iA_i|Z=0) \\
(2) & P(A_iA_j|Z=0) & = & 2q_iq_j \ (i<j) & P(A_iA_j|Z=1) & = & \epsilon P(A_iA_j|Z=0)
\end{array}
$$

That is, in principle, *ibd* $(Z=1)$ implies the same allelic type and non-*ibd* $(Z=0)$ implies independent allelic types. However a small "error" probability ϵ (or order 0.01) allows alleles scored as of different types to be *ibd*. An advantage of this simple model is that it is easily extended to the joint probability of a larger set of allelic types. Note the model under non-*ibd* $(Z=0)$ is of Hardy Weinberg proportions. As a population model, this seems more satisfactory than the model of homozygote deficiency and heterozygote excess implied by the finite-sample approach of Purcell et al. (2007).

Equations (1) and (2) define a Hidden Markov Model (HMM) for latent *ibd* and observed genotypes. The standard forward-backward algorithm (Baum et al. 1970) provides the conditional probability of *ibd* at every location on the chromosome, given the are allelic types on the chromosomes jointly over all loci.

In this paper, the aim is to extend the model of Leutenegger et al. (2003) to multiple genomes, enabling the joint analysis of *ibd* patterns among a set of haplotypes, or within and among a set of individuals for whom genotypic data are available. In section 2, we review the classical specification of *ibd* at a single locus, relating models of labeled (ordered) haplotypes (Nadot and Vayssiex 1973) to those of unlabeled partitions of a set (Ewens 1972). We then review the way in which the very large numbers of *ibd* states among n genomes group into classes that are *genotypically equivalent* (Thompson 1974). In section 3, we generalize the 2-haplotype Markov model (1) to multiple genomes using the marginal (single-locus) model of section 2. The model is developed first for labeled haplotypes, and it is shown that the reduction to genotypically equivalent classes of states remains Markov. The data model of equation (2) is also extended. Section 4 provides as illustrative example based on artificially constructed *ibd* among HapMap chromosomes (International Hapmap Consortium 2005), and section 5 concludes the paper.

2. Single-locus model for ibd among multiple genomes

We start by reviewing the classical single-locus framework for *ibd* among multiple genomes. A canonical labeling and accounting of the partitions of n genomes into *ibd* subsets was given by (Nadot and Vayssiex 1973): the case of the 15 states arising for $n=4$ is given in Table 1. For an unordered set of n genomes, Balding and Nichols (1994) modeled the partition into *ibd* groups using the Ewens sampling formula Ewens (1972). The partition into k sets is specified by $\mathcal{A} = (a_i; i = 1, ..., n)$ where

a_i is the number of sets of size i ($k = \sum_i a_i$, $n = \sum_i i a_i$). Then

$$(3) \qquad \boldsymbol{\pi}_n(a_1, ..., a_n) \quad = \quad \frac{n!\beta^{n-k}(1-\beta)^{k-1}}{(1+\beta)(1+2\beta)....(1+(n-2)\beta)} \prod_{j=1}^{n} (j^{a_j} a_j!)^{-1}$$

where $\beta = \pi_2(0,1)$ is the probability two genomes are *ibd*.

State class	*ibd* states $m_1p_1\ m_2p_2$	partition $\mathcal{A} = (a_1, a_2, a_3, a_4)$	equilibrium state prob	equilibrium class probability
1	11 11	all *ibd*; (0,0,0,1)	$6\eta\beta^3$	$6\eta\beta^3$
2	11 22	two pairs; (0,2,0,0)	$\eta\beta^2(1-\beta)$	$\eta\beta^2(1-\beta)$
3	11 12 and 11 21	three *ibd*; (1,0,1,0)	$2\eta\beta^2(1-\beta)$	$4\eta\beta^2(1-\beta)$
4	11 23	one pair; (2,1,0,0)	$\eta\beta(1-\beta)^2$	$\eta\beta(1-\beta)^2$
5	12 11 and 12 22	three *ibd*; (1,0,1,0)	$2\eta\beta^2(1-\beta)$	$4\eta\beta^2(1-\beta)$
6	12 33	one pair; (2,1,0,0)	$\eta\beta(1-\beta)^2$	$\eta\beta(1-\beta)^2$
7	12 12 and 12 21	two pairs; (0,2,0,0)	$\eta\beta^2((1-\beta)$	$2\eta\beta^2((1-\beta)$
8	12 13, 12 31, 12 23 and 12 32	one pair; (2,1,0,0)	$\eta\beta(1-\beta)^2$	$4\eta\beta(1-\beta)^2$
9	12 34	no *ibd*; (4,0,0,0)	$\eta(1-\beta)^3$	$\eta(1-\beta)^3$

Table 1: Single-locus states of gene identity for 4 genomes

In Table 1, $\eta = ((1+\beta)(1+2\beta))^{-1}$ is the normalizing constant of the distribution (3). Note also that different ordered states have the same unordered partition $\mathcal{A}$. For example, state 4, state 6, and the four states in class 8, all have $\mathcal{A} = (2,1,0,0)$. The total probability given by equation (3) is $6\eta\beta(1-\beta)$; each state has probability $\eta\beta(1-\beta)$. However, the seven states with $k = 2$ correspond both to $\mathcal{A} = (0,2,0,0)$ (state 2, and the two states of class 7) and to $\mathcal{A} = (1,0,1,0)$ (the four states in classes 3 and 5). For the former, $\boldsymbol{\pi}_4(0,2,0,0) = 3\eta\beta^2(1-\beta)$ and each of the three states has probability $\eta\beta^2(1-\beta)$, but for the latter the total probability $\boldsymbol{\pi}_4(1,0,1,0) = 8\eta\beta^2(1-\beta)$, and each has probability $2\eta\beta^2(1-\beta)$. For general n and k, there is no simple accounting of the numbers of *ibd* states of labeled genomes (Nadot and Vayssiex 1973) corresponding to a given partition of unlabeled genomes (Ewens 1972).

The number of *ibd* states among n labeled genomes increases very rapidly with n, being over 4×10^6 for $n = 12$ (Nadot and Vayssiex 1973; Thompson 1974). However many of these *ibd* states are *genotypically equivalent*, in the sense of providing equal probabilities of genotypic data and hence being non-identifiable from genotypic data. This potential reduction was first considered for two individuals ($n = 4$) by Cotterman (1940), but it was Jacquard (1972) who gave us the nine state classes in the now standard form. These nine state classes are as shown in Table 1. Although for two individuals ($n = 4$) we obtain a reduction only from 15 states to 9 state classes, the reduction becomes of increasing importance as n increases. For 6 individuals ($n = 12$), the over 4×10^6 states fall into only 198,091 state classes (Thompson 1974).

The genotype (at a single locus) of an individual is the unordered pair of his alleles. Hence if the n genomes are those of $n/2$ labeled individuals, any states which can be obtained from each other through the interchange of the two genomes within each of a subset of the individuals are genotypically equivalent. Thompson (1974) considered the group of transformations generated by elements T_i:

$$(4) \qquad G \quad = \quad < T_1, ..., T_{n/2} >$$

where T_i operates on the *ibd* states among n labeled genomes by interchanging genomes $2i-1$ and $2i$; the two genomes of individual i, $i = 1, ..., n/2$. Equivalence classes of states under G form the sets of genotypically equivalent *ibd* states. G has $2^{n/2}$ elements, but not all equivalence classes are the same

size. T_i has no effect on a state if the two genomes of i are *ibd*. Also, if a subset of individuals I all have the same two non-*ibd* genomes not shared by any other individuals, then $\prod_{i\in I} T_i$ does not change the state. For example,in Table 1, neither T_1 nor T_2 changes state 6 = (12 33), since individual 2 has two *ibd* genes, and individual 2 two distinct genomes not shared with individual 2. On the other hand, for states in group 8 both T_1 and T_2 are effective, creating an equivalence class of 4 states. In group 7, $T_1(12\ 12) = T_2(12\ 12) = (12\ 21)$, and T_1T_2 has no effect, providing an equivalence class of size 2.

3. Genomic model for ibd among multiple genomes

We must now extend the marginal model of section 2 to a model along a chromosome. We generalize the model (1). As in equation (1), along the genome, for each chromosome, *ibd* is gained at rate g and lost at rate, with $g/h = \beta/(1-\beta)$, where β is the probability a pair of chromosomes are *ibd* at any given location. Specifically, two types of transitions among *ibd* states are modeled:
(1) All pairs of singletons in states with $a_1 \geq 2$ become *ibd* at rate g. All doubletons in states with $a_2 \geq 1$ become non-*ibd* at rate h. We note that

$$\frac{\boldsymbol{\pi}_n(a_1,...,a_n)}{\boldsymbol{\pi}_n(a_1-2,a_2+1,a_3,...,a_n)} = \frac{2}{ga_1(a_1-1)}h(a_2+1). \tag{5}$$

Now there are $a_1(a_1-1)/2$ possible pairs of singletons, all becoming *ibd* at rate g, and in the resulting state a_2+1 doubletons, each becoming non-*ibd* at rate h.
(2) Each singleton in a state with $a_1 \geq 1$ joins with an *ibd* group size $(j-1) \geq 2$ at rate $(j-1)g$. Each member of an *ibd* group size j leaves it at rate h. We note that

$$\frac{\boldsymbol{\pi}_n(a_1,...,a_n)}{\boldsymbol{\pi}_n(a_1-1,a_2,...a_{j-2},a_{j-1}-1,a_j+1,a_{j+1},...a_n)} = \frac{jh}{(j-1)g}\frac{a_j+1}{a_1a_{j-1}}. \tag{6}$$

In this case there are a_1a_{j-1} choices of singleton and *ibd* set size $(j-1)$, and for each the larger group is formed at rate g, while in the resulting state there are a_j+1 *ibd* sets size j, and from each, each of the j elements becomes non-*ibd* at rate h.

Thus, under the transitions both of equation (5) and (6), detailed balance w.r.t. $\boldsymbol{\pi}_n$ is maintained. Hence $\boldsymbol{\pi}_n$ is the unique equilibrium distribution. For the case $n = 4$, the transition rate matrix under this model was given by Thompson (2008b).

The model for labeled states could be used even for genotypic data, but this would be computationally inefficient. As noted by Thompson (2008b) the 15-state Markov model for 4 genomes reduces to a 9-state Markov model for the genotypically equivalent classes of states among an ordered pair each of two unordered genome pairs (i.e. *ibd* among the four genomes of two individuals). Using the notation of the equivalence classes under group G (equation (4)), we note that this is true for any number of genomes under the transition model of equations (5) and (6). Under this model: (1) every transition changes the number of *ibd* groups, and hence the equivalence class, (2) The sojourn time within the state has the same distribution for every member of the class, and (3) The transition probability to a class A from class B is the same for every member $b \in B$. These conditions are sufficient to ensure that the process on the reduces space of equivalence classes of states remains Markov. Only (3) requires comment. It follows from the fact that for every member T of the group of transformations G generating the equivalence classes of states $\mathcal{A}$ and $\mathcal{B}$, and $a \in \mathcal{A}$ and $b \in \mathcal{B}$, the transition from b to a has the same rate as Tb to Ta. Note it is possible that $Ta = a$ and/or $Tb = b$; the equivalence classes $\mathcal{A}$ and $\mathcal{B}$ need not be the same size.

Unfortunately, for real data on individuals among whom there is substantial *ibd*, this model may be insufficient. For example, in the case $n = 4$ suppose that at some point all four are *ibd* ($a_4 = 1$), and that the next transition event along the genome arises from an ancestral recombination event shared in the ancestry of two of these chromosomes. The result of this shared ancestral junction is a

transition from $a_4 = 1$ to $a_2 = 2$, not permitted under the model of equations (5) and (6) above. To adjust for this, we use the second form of the Leutenegger model of equation (1). If Q denotes the Q-matrix under the model of equations (5) and (6), then we adjust Q to become:

$$(7) \qquad Q^{\dagger} \;=\; (1-\delta)Q + \delta(-I + \mathbf{1}\boldsymbol{\pi}_n')$$

Under $Q^{\dagger}$ there are two kinds of breakpoints the first, occurring at rate $(1-\delta)$ times the previous rates with transitions forced to occur, and the the second, at rate δ where the new state is chosen from $\boldsymbol{\pi}_n()$ independently of the current state (and may be that current state). In th model (7) all transitions are possible, and choice of the mixing parameter δ can be tuned to the extent shared ancestral junctions are suspected within the ancestral time-frame that is the focus of the analysis. This model clearly retains the same equilibrium probabilities $\boldsymbol{\pi}_n()$ of *ibd* states as the previous Q-matrix, and additionally, since after a "breakpoint" of the second kind the new state is independent of the previous state, the retention of the Markov property in reducing to genotypically equivalent state classes is assured.

Our model for data at a locus given the underlying *ibd* follows that of equation (2). In principle, *ibd* haplotypes must have the same allelic type, and non-*ibd* ones are of independent types. As in equation (2) this is modified to allow for error by mixing this idealized distribution with a proportion ϵ of the distribution where no haplotypes are *ibd*. Since we do not here model LD, only the single-locus haplotype probabilities need be specified. The model for genotypes is simply given by reducing to the unordered pair of haplotypes within each individual. As for the model of section 1, we have a simple HMM. The standard forward-backward algorithm (Baum et al. 1970) provides the conditional probability of *ibd* at every location on the chromosome, given the are allelic types on the chromosomes jointly over all loci.

4. Inference of IBD: an Illustrative Example

As an illustration, we take genetically realistic chromosomes, using phased HapMap SNP data from Chromosome 19 (International Hapmap Consortium 2005) YRI (African) individuals. Markers with minor allele frequency less than 6% are eliminated, leaving markers at any average density of 1 per 10^4 base pairs. A pool of 60 chromosomes of length 10^8 bp (10,000 SNP markers) are then created, retaining the SNP marker physical locations, allele frequencies, and LD patterns of the original HapMap data.

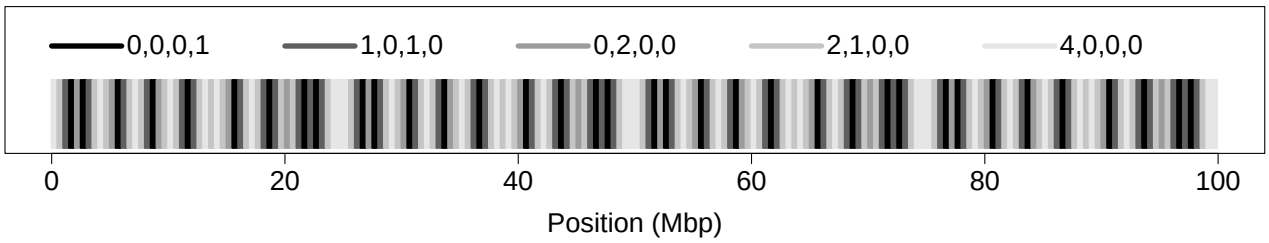

Figure 1: True ibd pattern constructed for the example

Patterns of *ibd* among sets of chromosomes are then artificially created. For illustration here, we present just one set on 4 chromosomes, with a complex pattern of *ibd*. Every 0.5 Mbp (on average, every 50 markers), one chromosome currently *ibd* to other(s) will switch to copying from one in the pool but not currently represented in the quartet, or a singleton in the quartet will switch to copying from a random one of the other three. Additionally some transitions between the states $\mathcal{A} = (0,0,0,1)$ (all *ibd*) and $\mathcal{A} = (0,2,0,0)$ (two pairs *ibd*) are introduced. The resulting pattern among the 15 *ibd* states is shown in Figure 1; all 15 states are represented multiple times in these 200 segments, although it is only possible to show the 5 groups of unlabeled partitions. Finally errors are introduced, at a rate of 1%, switching the SNP allelic type independently in each chromosome and at each marker.

For the analysis of these data, parameter values must be chosen. We use $\epsilon = 0.01$ for the error probability, and $\delta = 0.2$ in equation (7), since we have constructed many transitions reflecting shared ancestral junctions. The choice of β reflects the overall level of pairwise *ibd*. The marginal probabilities of the five unlabeled state groups are given in Table 2, and the results shown in this paper are for $\beta = 0.3$. Finally, the absolute values of h and g are determined as a function of β and δ, in such a way that the mean length of chromosome in an *ibd* state is 0.5 Mbp under the model, reflecting the data of Figure 1.

State description	partition $\mathcal{A} = (a_1, a_2, a_3, a_4)$	number of states	Total probability under $\beta = 0.2$	$\beta = 0.3$
all *ibd*	(0,0,0,1)	1	0.029	0.078
three *ibd*	(1,0,1,0)	4	0.152	0.242
two pairs *ibd*	(0,2,0.0)	3	0.057	0.091
one pair *ibd*	(2,1,0,0)	6	0.457	0.424
no-*ibd*	(4,0,0,0)	1	0.305	0.165

Table 2: Marginal probabilities of unlabeled state classes

The data are analyzed first as haplotypic data. The results are shown in Figure 2. At each position across the 10^8 bp, the proportion of each unit height is the probability, given the marker data, of each of the five unlabeled *ibd* states. At each position, the states are ordered, from bottom to top, in order of decreasing *ibd*, from $\mathcal{A} = (0, 0, 0, 1)$ ($a_4 = 1$; all *ibd*) to $\mathcal{A} = (4, 0, 0, 0)$ ($a_1 = 4$; no *ibd*). Finally, the true *ibd* pattern of Figure 1 is shown as a strip across the center, for a visual assessment of the accuracy of the *ibd* inference.

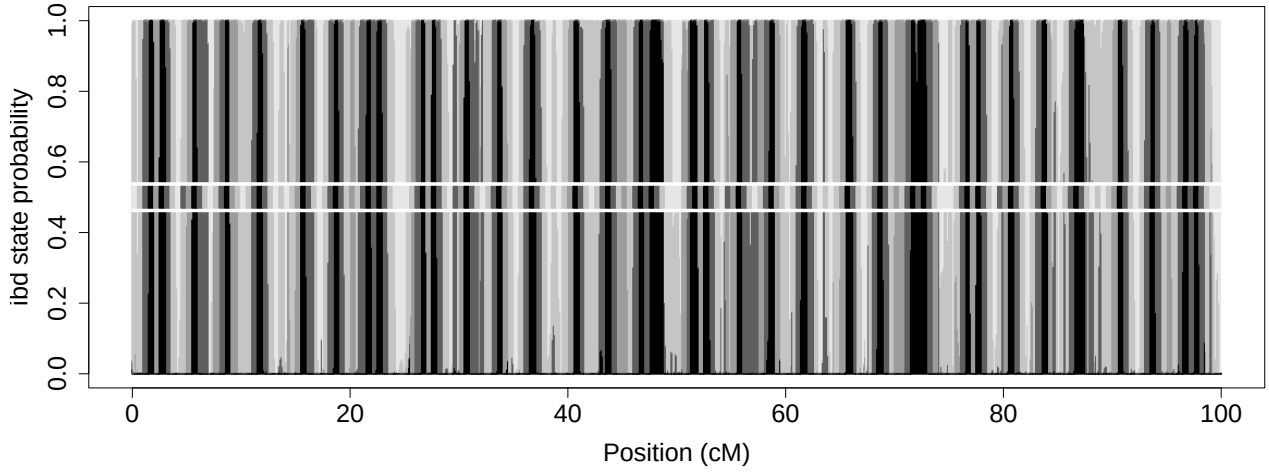

Figure 2:* ibd *probabilities conditional on haplotypic data

We see that most probabilities are close to 1 for some state, and to 0 for the remainder. There are some intermediate probabilities; for example, for $\mathcal{A} = (4, 0, 0, 0)$ (no-*ibd*) and $\mathcal{A} = (2, 1, 0, 0)$ (one pair *ibd*) at around 50 Mbp, and for $\mathcal{A} = (2, 1, 0, 0)$ (one pair *ibd*) and $\mathcal{A} = (1, 0, 1, 0)$ (three *ibd*) at around 90 Mbp. It is of interest that these are both regions of high LD in the original chromosomes. The regions of high *ibd* (the two darkest shades) are well estimated, but generally *ibd* is over-estimated with some sections of no *ibd* ($\mathcal{A} = (4, 0, 0, 0)$; $a_1 = 4$) missed entirely in the reconstruction. This can be remedied by reducing β to 0.2 (Table 2), but the general problem of parameter tuning or estimation remains to be addressed.

The same four haplotypes as above were then paired into two genotypes, and the data reanalyzed as genotypic data on two individuals. The results are shown in Figure 3. Interestingly, in the case, using the same parameter values, more non-*ibd* ($\mathcal{A} = (4, 0, 0, 0)$) is estimated. As would be expected, there is generally less certainty, with many more intermediate probabilities. The *ibd* pattern no longer

well inferred, but at least the inference of more vs less (darker vs lighter) *ibd* is mostly correct. Real haplotypes from real populations will have less complex *ibd* patterns that those of Figure 1, but the importance of accurate phasing or biological haplotypes is significant. In the context of inferring *ibd* between small pedigrees sampled from a population, the pedigree data will provide at least partial phase information.

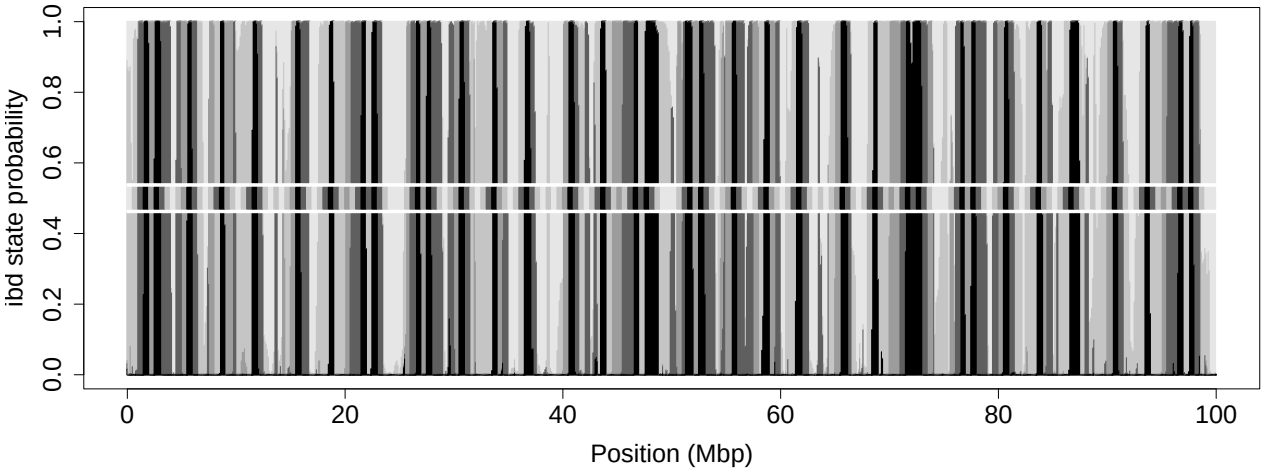

Figure 3:* ibd *probabilities conditional on genotypic data

One may ask whether joint analysis of multiple haplotypes performs better than multiple pairwise analyses. Of course, with genotypic data one has no choice but to analyze jointly the four haplotypes underlying a pair of genotypes. Moreover, analyzing four haplotypes jointly is computationally slightly faster than 6 pairwise analyses, and if multiple pairwise analyses are performed, how are these to be melded together? However, the question of increased power to detect *ibd* segments and improved accuracy of estimation of *ibd* patterns remains open. As an example, we re-analyzed the four haplotypes above in six pairwise analyses, using the same values of h and g (equation (1)) as for the joint analysis (equations (5) and (6)), and compared the results with the six pairwise summaries of the joint analysis (Figure 2). For reasons of space, the results are not shown. Generally, there seemed little change in the accuracy of inferred *ibd* between the two haplotypes of each pair. However, it seemed that uncertainty was better calibrated by the joint analysis. The pairwise analyses had output *ibd* probabilities close to 0 or 1, even where incorrect. In regions where the data apparently suggested an *ibd* pattern other than than constructed in these data, the joint analyses showed intermediate *ibd* probabilities. Joint analysis also better detects regions of high LD, resulting in very short spikes ($\sim 3 - 5 \times 10^4$bp) of inferred *ibd*. These LD segments may reflect true more remote coancestry.

5. Conclusion

Modern dense SNP data permits detection of *ibd* between the chromosomes of observed individuals not only in well-sampled pedigrees but also among population members not known to be related. However, for practical purposes, flexible models with few parameters that permit the inference of, at least, the *ibd* pattern among the four genomes of a pair of individuals are required, and such models must be applicable to both (unphased) genotypic and (phased) haplotypic data.

In this paper, we have shown how the simple Markov process for *ibd* along a chromosome developed by Leutenegger et al. (2003) can be extended to *ibd* among multiple genomes, using the same two rate parameters of gain (g) and loss (h) of *ibd*. The equilibrium distribution of this Markov model of *ibd* among labeled genomes corresponds to the model of Ewens (1972) for the partitions of unlabeled genomes. An additional parameter δ (equation (7)) makes the model more flexible, and can be used in small populations where shared ancestral recombination break-points are likely. In addition

to these three parameters of the *ibd* process, the data-model parameters consist only of marker allele frequencies, for which empirical sample frequencies are used, and an "error rate" ϵ.

Our illustrative example shows that provided *ibd* segments are longer than the range of LD then we can usually detect them, even without any explicit LD model. Phased haplotypic chromosomes provide more accurate *ibd* information. In practice, if these methods are used to infer *ibd* among small pedigrees sampled from a population, the family data provide a least partial phase information.

Many aspects of this approach remain to be investigated, and methods for the estimation or automated tuning of parameters are needed. However, the results are promising for the development of methods of analysis of complex genetic traits that use *ibd* inferred from dense genomic data. We have shown that this *ibd* can be inferred not only within small well-sampled pedigrees, but also between members of different pedigrees not *a priori* known to be related. In this way, family-based and population-based studies can be combined to provide their complementary strengths to the genetic mapping of complex traits.

Acknowledgment: This research was supported in part by NIH grant GM46255.

REFERENCES (RÉFERENCES)

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

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

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

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

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.

Di Y, Thompson EA (2008) Conditional tests for localizing trait genes. Human Heredity in press

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

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

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

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

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

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

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

Te Meerman GJ, Van der Meulen MA, Sandkuijl LA (1995) Perspectives of identity by descent (IBD) mapping in founder populations. Clin Exp Allergy 25 (Suppl 2.):97–102

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

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

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