Gene coancestry in pedigrees and populations

Thompson, Elizabeth
University of Washington, Department of Statistics Box 354322
Seattle, WA 98115-4322, USA
E-mail: eathomp@uw.edu

Glazner, Chris
University of Washington, Department of Statistics Box 354322
Seattle, WA 98115-4322, USA
E-mail: cglazner@uw.edu

1. Introduction

Related individuals share common ancestors, and hence may carry DNA that is identical by descent (*ibd*) from these ancestors. With high probability, *ibd* DNA is of the same allelic type, leading to trait similarities among relatives. Classically, data on known relatives are used to map the genes underlying genetically mediated traits, and the prior probabilities of *ibd* are then given by the pedigree structure. However, pedigree data are expensive and difficult to collect, and the limited number of meioses within a set of known pedigrees leads to a lack of resolution in gene mapping. When pedigrees are ascertained for extreme trait values or from small populations, there are likely to be unknown relationships among the founder members of the same or of different pedigrees. Modern dense informative genetic marker data permit inference of *ibd* resulting from these unknown relationships, and this inferred *ibd* may be combined with *ibd* imputed within pedigrees to increase both the power and the resolution of mapping of genes contributing to complex quantitative traits.

In this paper, we consider first the analysis of data within pedigrees, in terms of the *ibd graph*. This graph, defined among observed individuals and across the genome, specifies the segments of genome shared *ibd* among these individuals. Once the *ibd* graph is known, analyses of trait data may be carried out conditionally on the graph, and the pedigree relationships and genetic marker data are no longer relevant. We then show how *ibd* resulting from unknown more remote relationships can be estimated using a population-genetic based *ibd* model. Merging of the *ibd* graphs inferred within and among pedigrees provides a combined *ibd* graph, which may be used for trait-data analyses.

We illustrate these methods with a small simulated-data example. We first examine the effect of genetic marker density on the inference of *ibd* in an extended pedigree. We then remove knowledge of some ancestors to create small subpedigrees, and analyze the *ibd* within and between these subpedigrees. Using the subpedigrees alone, linkage information is lost, but it is almost fully regained by inference of *ibd* among the subpedigrees. Software implementing these methods is available in the MORGAN-3 package (MORGAN V3.0.1 2010).

2. Pedigree-based lod score as a function of coancestry

Given a genetic model, Γ, for genetic marker data $\mathbf{Y}_M$ and trait data $\mathbf{Y}_T$, the classical statistic for mapping DNA contributing to a trait relative to a known map of genetic markers is the lod score:

$$\log_{10}\frac{\Pr(\mathbf{Y}_T,\mathbf{Y}_M;\Gamma)}{\Pr(\mathbf{Y}_T,\mathbf{Y}_M;\Gamma_0)} \;=\; \log_{10}\frac{\Pr(\mathbf{Y}_T,\mathbf{Y}_M;\Gamma)}{\Pr(\mathbf{Y}_T;\Gamma_T)\Pr(\mathbf{Y}_M;\Gamma_M)} \;=\; \log_{10}\frac{\Pr(\mathbf{Y}_T \mid \mathbf{Y}_M;\Gamma)}{\Pr(\mathbf{Y}_T;\Gamma_T)},$$

where $\Gamma_0 = (\Gamma_T, \Gamma_M)$ is Γ without dependence in inheritance of DNA affecting $\mathbf{Y}_T$ and DNA affecting $\mathbf{Y}_M$. On an extended pedigree, the term $\Pr(\mathbf{Y}_T \mid \mathbf{Y}_M;\Gamma)$ can be computationally intractable, but can

be estimated as a sum over latent variables $\mathbf{S}$ which specify the inheritance at all marker locations:

$$\Pr(\mathbf{Y}_T \mid \mathbf{Y}_M;\Gamma) \;=\; \sum_{\mathbf{S}} \Pr(\mathbf{Y}_T \mid \mathbf{S};\Gamma_T)\Pr(\mathbf{S} \mid \mathbf{Y}_M;\Gamma_M), \tag{1}$$

since, given $\mathbf{S}$, $\mathbf{Y}_T$ and $\mathbf{Y}_M$ are independent. One-time realization of a sample of $\mathbf{S}$ then permits the estimation of the lod score for multiple hypothesized trait locations, multiple trait models, and even multiple traits observed on the same pedigree structures (Lange and Sobel 1991). Newer MCMC sampling methods permit effective realization of $\mathbf{S}$ on large pedigree datasets for multiple closely linked markers (Tong and Thompson 2008; Thompson 2011a). These methods are implemented in the MORGAN program *lm_multiple*.

The *ibd* graph specifies patterns of identity by descent (*ibd*) among individuals and across a chromosome. At a locus, the edges of the *ibd* graph are labelled by the individuals observed for the trait or by their trait values. Edges connect two nodes which correspond to the two haploid genomes descending to the individual. Two different edges impinging on a node indicate genome shared *ibd* at this locus by the corresponding individuals. If the two genomes of an individual are *ibd* at a locus, both ends of his edge connect to a single node. Thus the nodes of the *ibd* graph are intrinsically unlabelled, showing only *ibd* among individuals. Nodes are defined only through the edges that impinge upon them (Thompson 2011b).

At genetic marker locations, the *ibd* graph is a function of $\mathbf{S}$. The probability of trait data $\mathbf{Y}_T$ depends on $\mathbf{S}$ only through the *ibd* graph. Instead of computing the lod score contribution for each realized $\mathbf{S}$, the MORGAN program *gl_auto* samples $\mathbf{S}$ but converts each scored realization to an *ibd* graph. A sample of *ibd* graphs may be stored in compact format; only change-points across a chromosome are stored. The MORGAN program *gl_lods* then computes lod score contributions for each stored *ibd* graph. For modern dense informative marker data, and where complex phenotypes often provide little information on inheritance, the one-time analysis of marker data has clear computational advantages, permitting easy analysis of many trait models and many trait phenotypes. There are also data-security advantages; the *gl_auto* program requires only pedigree information, marker data, and marker model. Once the *ibd* graphs are sampled, the pedigree structure and marker data are no longer relevant. The *gl_lods* program requires only the *ibd* graphs, trait data, and trait model.

Use of the sampled *ibd* graphs for the computation of trait-model lod score contributions has other significant computational advantages. First, computation on the *ibd* graph of observed individuals is often significantly faster than computation on a pedigree using $\mathbf{S}$. Particularly when few individuals are observed, the disjoint components of the *ibd* graph tend to be much smaller than the pedigree graph. More importantly, many realizations of $\mathbf{S}$ may be the same and many distinct values of $\mathbf{S}$ give the same unlabelled *ibd* graph. In a pedigree, recombination breakpoints are relatively few, and realized *ibd* graphs remain constant over many markers. Recognizing when *ibd* graphs are the same is key to efficient computation, since lod score computations need be computed only for each distinct graph. Software to recognize *ibd*-graph equivalence has been implemented in the IBDgraph package (Koepke and Thompson 2010), and can reduce the lod-score component of computation by orders of magnitude (Thompson 2011b).

3. Inferring coancestry among pedigrees

When relationships between individuals are not known, *ibd* can be inferred using a Hidden Markov Model, which we implement in the MORGAN program *ibd_haplo*. The hidden states of the model are the possible *ibd* patterns among two individuals and form a Markov chain as described in Thompson (2008, 2009). The transition matrix is parametrized by the expected degree of relatedness among the individuals and the expected length of *ibd* segments, both of which are derived from attributes of the population containing the individuals. The hidden states emit observed alleles in

accordance with population allele frequencies; *ibd* chromosomes will emit the same allele in the absence of measurement error, while non-*ibd* alleles are modeled as random draws from the population.

Studies using simulated haplotypes showed the model detected nearly all *ibd* segments longer than 1 Mbp (Glazner et al. 2010). Linkage disequilibrium (LD) in the founder population created many short segments of detected *ibd*. Because LD is itself a reflection of coancestry more recent than the time required to break down haplotypes, these segments can be interpreted as a form of *ibd* sharing.

The *ibd* detected in this manner can be used to recover unobserved coancestry among individuals in different pedigrees. A set of families drawn from the same population is likely to have some shared ancestry, but pedigrees reflecting these relationships will typically be far larger and deeper than can be realistically observed. The *ibd_haplo* model infers the *ibd* produced by these unobserved relationships.

To combine a set of MCMC realizations of the *ibd* graphs on a pair of pedigrees, *ibd_haplo* is first run on the genotypes of every possible pairing of individuals between the two pedigrees. This produces, at each locus and for each pair, the marginal probability that the two individuals are in any of 9 possible *ibd* states at that locus. The most probable *ibd* state from each pair is selected, and the pairs are ranked according to the probability of the most probable state.

Given an *ibd*-graph realization, these states can be translated into statements about pairs of founder haplotypes (nodes in the *ibd* graph) being *ibd*. For example, suppose two individuals carry founder labels {1,2} and {7,8}, respectively, in a particular pair of *ibd* graphs. If we infer from *ibd_haplo* that they share one allele *ibd*, then we conclude that one of the four possible pairings 1-7, 1-8, 2-7, or 2-8 must be a pair of labels which are *ibd*.

The (ambiguous) founder label statements implied by each pair's state are successively added, in order of probability, to a collection of statements whose consistency is checked at each step using the MiniSat program. (Eén and Sörensson 2003) If the addition of a set of statements conflicts with the previously included statements, then that set of statements is excluded. In this manner more probable inferences are given priority over less probable ones. When all sets of statements have been tried, the program produces a consistent solution to the set of included statements, which corresponds to the presence or absence of pairings between founder labels in the two *ibd* graphs. The nodes whose labels are paired are then combined in the two graphs, creating a new, possibly connected graph.

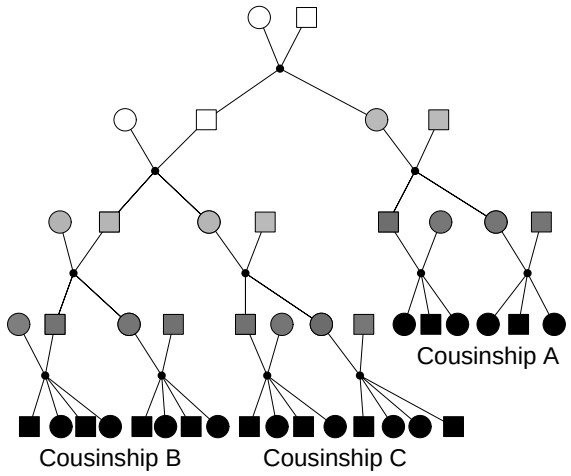

Figure 1: The Ped44 example pedigree. The 22 dark-shaded, last-generation, individuals are observed for trait and marker data. To create the three cousinships, the 4 unshaded ancestors are removed. To create the six sibships, the light-shaded grandparents of the observed individuals are also removed.

4. The Ped44 example; missing pedigree information

As an illustrative example, we describe results for simulated data on a single 44-member pedigree, Ped44 (Figure 1). A locus affecting a quantitative trait was placed at the centre of a 100 Mbp chromosome, and descent of genome over the chromosome was simulated conditional on the trait data, using the MORGAN *markerdrop* program. Three marker data sets were then simulated, conditional on the single descent pattern; 51 SNP markers at 2 Mbp spacing, 13 STR markers at 7.5 Mbp spacing,

and 201 SNPs at 0.5 Mbp spacing. Only the 22 final individuals of the pedigree were assumed observed for marker and trait data (Figure 1).

We first considered the lod scores assuming the whole Ped44 to be known. Lod scores were estimated using the MORGAN *lm_multiple* program, with sampling for 30,000 MCMC scans, and scores realized every 30 scans. While the overall lod scores do not differ greatly among the three marker densities (Figure 2(a)), the 1000 MCMC-generated contributions to the overall score (equation (1)) (shown in grey in Figure 2) show different patterns. With only 51 SNPs, there is very high uncertainty in latent *ibd*, as reflected in highly variable lod score contributions (Figure 2(b)). With the more widely spaced but individually more informative STR markers, uncertainty is reduced, but resolution is poor (Figure 2(c)). With 201 SNPs, we have low uncertainty and high resolution (Figure 2(d)). Since the data are simulated, we in fact know the lod score that would be found were the true *ibd* on this pedigree known. This lod score is shown in Figure 2(e), and the 201-SNP lod score follows it closely. These results show also that the MCMC methods of Tong and Thompson (2008) work well at this 0.5 Mbp scale on this extended pedigree with no observed data on 50% of the individuals.

While reduction using IBDgraph (Koepke and Thompson 2010) was not used for this small example, it was verified that identical results were obtained when lod score contributions were computed on the basis of *ibd* graphs generated with the same MCMC sampling options by the MORGAN *gl_auto* program. Further, running IBDgraph on these *ibd* graphs showed that at the 50 Mbp position, the 1000 realizations for the three marker datasets generate only 265, 70 and 5 distinct *ibd* graphs, with the size of the largest group being 51, 495, and 996, respectively. For the 201 SNP dataset, the number of realizations in the largest group averages 932 over the 30 markers from 42 to 57 Mbp, with many of these *ibd* graphs remaining unchanged across these 30 markers. Clearly, computing lod score contributions only for distinct *ibd* graphs would greatly reduce *gl_lods* computation time.

Using only the 201-SNP dataset, we next show the result of missing pedigree information. Using first the 3 subpedigrees consisting of cousin-pairs of sibships in Ped44, and then the 6 sibships separately, we computed lod scores, and summed these over the cousinship or sibship families, as would be done if the relationships among the families were unknown. The results for the 1000 realizations for the 3 cousinships are shown in Figure 2(f), and the total lod scores in Figure 2(g). (For the sibships, no MCMC is needed, and exact lod scores are computed.) Clearly, the sibships alone contain little information, but the *ibd* between the two sibships in each cousinship does provide some linkage evidence. With two major exceptions, the sum of the 3 cousinships shows lod score contributions very similar to the overall one (Figure 2(d)), and with slightly less variation among the 1000 realizations. First the lod score in the neighborhood of the trait locus (45-55 Mbp) is significantly reduced. Second, the lod score at 75-85 Mbp is quite high, whereas the overall result and that for the true latent *ibd* (Figure 2(e)) are close to 0 in this region. This result accords with the recognition that over much of the chromosome there is in fact no *ibd* among the 3 cousinships. However, at 45-55 Mbp there is *ibd* that is concordant with trait values, while at 75-85 Mbp there is *ibd* that is discordant with trait similarities among individuals.

Finally, we run the MORGAN *ibd_haplo* program on all pairs of individuals in Cousinships A and B; note these are not the two most closely related cousinships, but, by chance, they have more genome shared *ibd*. The IBDmerge software is then run to produce 1000 *ibd* graphs that combine the *gl_auto* results on the cousinships with the additional *ibd* inferred by *ibd_haplo*. The resulting lod score contributions and overall lod score are shown in Figure 2(h), with the overall value also in Figure 2(g). We see that this procedure has almost fully recaptured the information in the full Ped44. In particular, the high lod score at 45-55 Mbp is regained, and the false signal at 75-85 Mbp is eliminated. Thus our procedures, for combining *ibd* inferred among families not known to be related with the descent patterns within families used in classical linkage analysis, show significant promise both for increasing the possibilities of linkage detection and for eliminating false positive signals.

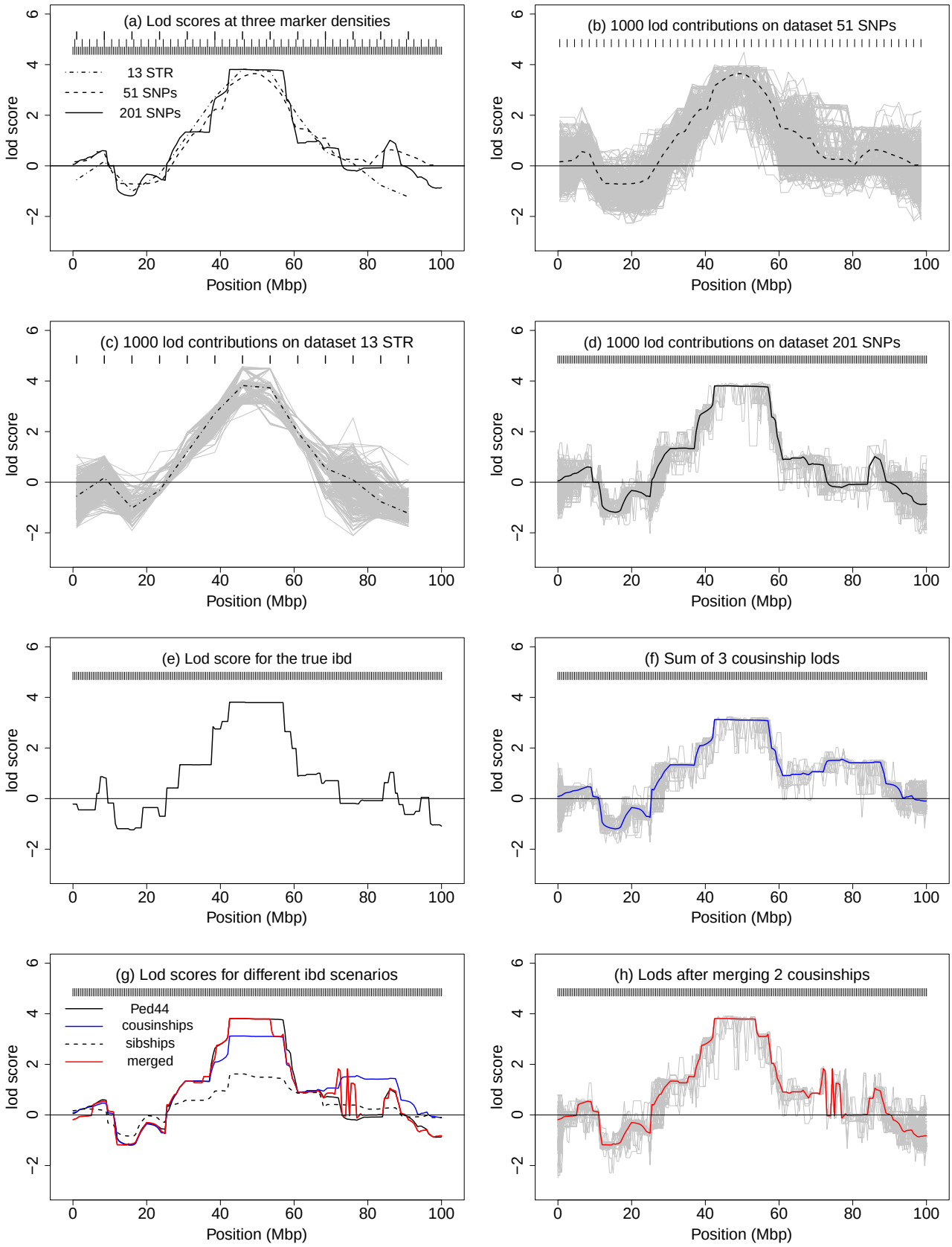

Figure 2: Uncertainty in pedigree-based lod scores: (a) Lod scores at three marker densities, (b,c,d) Full Ped44 lod contributions at three marker densities, (e) True Ped44 lod score, (f) Lod contributions on 3 cousinships, (g) Lod scores with the four inferred ibd ***scenarios. (h) Lod contributions after inferring*** ibd ***between 2 cousinships***

5. Discussion

Lod scores for genetic linkage analysis may be computed on the basis of the *ibd* graph, and, for this purpose, it is irrelevant whether this *ibd* is inferred using known pedigree relationships or from a population model, or from a combination of the two. Our example shows how merging *ibd* inferred among small pedigrees with the *ibd* inferred within these pedigrees can recover the linkage signal that would be obtained were the relationships among pedigrees known.

In our small Ped44 example, we used the same genetic markers for *ibd* inference both between and within pedigrees, and lod scores were computed at all marker locations. The density of markers for *ibd* inference is unrelated to the often lesser density at which lod score computation is desired. Lod scores may be computed at any location at which *ibd* is realized conditional on chromosome-wide marker data and merged among pedigrees. For real examples, with remote unknown relationships among pedigrees, marker densities for between- and within-pedigree *ibd* realization should differ. Within pedigrees, markers at an average spacing of 0.5 Mbp work well. For remote relationships among pedigrees, dense SNP markers (for example, 50 per Mbp) are required for reliable detection of *ibd* segments as small as 1 Mbp. The uncertainty of the lod score based on merged *ibd* at 73 to 77 Mbp (Figure 2(h)) results from discrepancies among single markers at the 0.5 Mbp spacing.

In practice, SNP data are often available at the 50 per Mbp scale. For pedigree-based analyses, markers at an average 0.5 Mbp spacing and exhibiting highest counts of heterozygous individuals in the pedigrees can be subselected. At this scale, potential problems due to LD are avoided. MORGAN programs have been modified so that output information, including *ibd* graphs, is given in terms of the marker indexing in the input file, not in terms of only the selected markers. This makes practical the merging of dense-marker *ibd_haplo* results with those of the pedigree-based *gl_auto* program.

Acknowledgment:
This research was supported in part by NIH grants R37 GM46255 and T32 GM81062.

REFERENCES (RÉFERENCES)

Eén N, Sörensson N (2003) An Extensible SAT-solver. In E Giunchiglia, A Tacchella, eds., *SAT*, vol. 2919 of *Lecture Notes in Computer Science*, 502–518. Springer

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

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

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

MORGAN V3.0.1 (2010) A package for Monte Carlo Genetic Analysis. Available at:, http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml

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

— (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, London, UK

— (2011b) The structure of genetic linkage data: from LIPED to 1M SNPs. Human Heredity 71:88–98

Tong L, Thompson EA (2008) Multilocus lod scores in large pedigrees: Combination of exact and approximate calculations. Human Heredity 65:142–153