Genetic Epidemiology 32: 513–519 (2008)

Calculation of IBD Probabilities with Dense SNP or Sequence Data

Jonathan M. Keith,[1]* Allan McRae,[2] David Duffy,[2] Kerrie Mengersen,[1] and Peter M. Visscher[2]

[1]*School of Mathematical Sciences, Queensland University of Technology, Brisbane, Qld. 4001, Australia*
[2]*Queensland Institute of Medical Research, Royal Brisbane Hospital, Herston, Qld. 4006, Australia*

The probabilities that two individuals share 0, 1, or 2 alleles identical by descent (IBD) at a given genotyped marker locus are quantities of fundamental importance for disease gene and quantitative trait mapping and in family-based tests of association. Until recently, genotyped markers were sufficiently sparse that founder haplotypes could be modelled as having been drawn from a population in linkage equilibrium for the purpose of estimating IBD probabilities. However, with the advent of high-throughput single nucleotide polymorphism genotyping assays, this is no longer a reasonable assumption. Indeed, the imminent arrival of individual sequencing will enable high-density single nucleotide polymorphism genotyping on a scale for which current algorithms are not equipped. In this paper, we present a simple new model in which founder haplotypes are modelled as a Markov chain. Another important innovation is that genotyping errors are explicitly incorporated into the model. We compare results obtained using the new model to those obtained using the popular genetic linkage analysis package Merlin, with and without using the cluster model of linkage disequilibrium that is incorporated into that program. We find that the new model results in accuracy approaching that of Merlin with haplotype blocks, but achieves this with orders of magnitude faster run times. Moreover, the new algorithm scales linearly with number of markers, irrespective of density, whereas Merlin scales supralinearly. We also confirm a previous finding that ignoring linkage disequilibrium in founder haplotypes can cause errors in the calculation of IBD probabilities. *Genet. Epidemiol.* 32:513–519, 2008.

Key words: identity by descent; linkage disequilibrium; single nucleotide polymorphism

Contract grant sponsor: NHMRC; Contract grant numbers: 38982, 219178, 389891, 442915; Contract grant sponsor: Australian Research Council; Contract grant number: DP0556631.
*Correspondence to: Jonathan M. Keith, School of Mathematical Sciences, Queensland University of Technology, Brisbane, Qld. 4001, Australia. E-mail: j.keith.@qut.edu.au
Received 24 October 2007; Accepted 8 February 2008
Published online 20 March 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20324

INTRODUCTION

A wide range of strategies for mapping disease genes and quantitative trait genes involve estimating probabilities that a pair of related individuals share 0, 1, or 2 alleles identical by descent (IBD) at a point along a chromosome. For example, the affected sib-pair method [Cudworth and Woodrow, 1975; Motro and Thomson, 1985] involves investigating whether the proportion of affected sib pairs sharing 0, 1, or 2 alleles IBD at a marker locus differs significantly from expectations, assuming a null model in which there is no linkage between the marker and the disease. Standard parametric and non-parametric tests for linkage also involve estimating IBD probabilities [Kong and Cox, 1997; Krugylak et al., 1996; Whittemore and Halpern, 1994]. IBD probabilities are also required for certain family-based tests of linkage disequilibrium (LD), such as those implemented in the QTDT package [Abecasis et al., 2000a,b].

A number of methods for estimating IBD probabilities have been developed. The Elston-Stewart algorithm [Elston and Stewart, 1971; Lange and Elston, 1975], as implemented in the VITESSE software package [O'Connell, 2001; O'Connell and Weeks, 1995], is appropriate for large pedigrees, but is only feasible for a small number of markers. The leading method when the number of marker loci is large is the Lander-Green algorithm [Krugylak et al., 1995, 1996; Lander and Green, 1987]. Importantly for large numbers of loci, the Lander-Green algorithm has time and memory requirements proportional to the number of loci. Markov chain Monte Carlo methods suitable for small or large pedigrees have also been developed [Heath, 1997; Liu et al., 2007; Wijsman et al., 2006]. However, these are not suitable for dense marker data and, in particular, do not take into account marker-marker LD. The state-of-the-art software for evaluating IBD probabilities in small-to-medium size pedigrees is the popular Merlin package [Abecasis et al., 2002], which includes a highly efficient implementation of the Lander-Green algorithm. An assumption implicit in straightforward implementations of the Lander-Green algorithm, including the original implementation in Merlin, is that parental haplotypes are drawn from a population in linkage equilibrium. In other words, the allele occurring at any given locus in the haplotype is assumed to be independent of the alleles occurring at all other loci. However, Merlin has recently been modified to allow for LD between markers in founder haplotype population [Abecasis and Wigginton, 2005]. The modification involves identifying clusters of markers that represent haplotype blocks and estimating the population frequencies of each haplotype in each cluster. This is an important development, since information about parental haplotypes can be very helpful in

resolving uncertainties about the number of IBD alleles at each locus. Moreover, ignoring LD can create bias in IBD estimation when parental genotypes are unobserved [Schaid et al., 2002].

The current move towards denser linkage maps entails that LD between adjacent markers is also increasing. Moreover, large numbers of individual genome sequences will soon be available. The first individual genome has been sequenced [Levy et al., 2007] and an international collaboration known as the 1,000 genome project, which aims to sequence 1,000 individual genomes, has been launched [Hayden, 2008; Qiu and Hayden, 2008]. Consequently, extremely dense marker data at a very large number of loci will soon be available. Although individual sequencing will produce essentially the same type of data—single nucleotide polymorphism (SNP) genotypes—two new problems will arise. The first is the sheer scale of data that will be produced—millions of SNPs for thousands of individuals. The second is the increased density, and hence the increasing importance of marker-marker LD. Efficient algorithms that account for marker-marker LD will thus become essential.

The algorithm implemented in Merlin involves identifying clusters of markers that represent haplotype blocks and estimating the population frequency of each haplotype in each cluster. However, the method has a number of drawbacks:

(i) it relies on knowing or estimating a large number of population parameters,
(ii) it assumes no recombination within a cluster and no LD between clusters, and
(iii) it scales supralinearly in run time and memory requirements with respect to the number of markers, and thus does not scale up to whole-genome, dense-marker data.

Here, we describe a new model and method for estimating IBD probabilities that addresses these three drawbacks, by incorporating a Markov model for the founder haplotypes. In particular, the computation time and memory requirements scale linearly with the number of markers, and recombination and LD between any pair of adjacent markers are permitted. An additional advantage of the new model is that it allows for and detects likely genotyping errors. We present results using real and simulated data to demonstrate the improved speed and accuracy of the new method compared to Merlin. Our results also confirm that accounting for LD can substantially improve the estimation of IBD probabilities.

Source code (written in C) implementing the new method is available from the contact website: http://www.maths.qut.edu.au/~keithj/.

MATERIALS AND METHODS

THE MODEL

Figure 1 illustrates the various parameters of the model and their conditional dependencies. The given data consists of observed genotypes, which we group together into two vectors, G and G'. These are shown at the bottom of Figure 1, and represent observed genotypes for founders and non-founders, respectively. The components of G are written as G_{ij}, representing the observed genotype

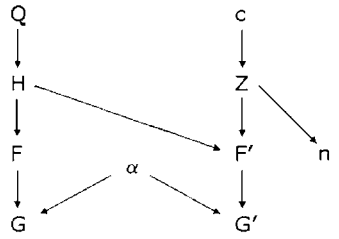

Fig. 1. The parameters of the model and their conditional dependencies. The parameter at the head of each arrow is conditionally dependent on the parameter at the tail. From the bottom up, G and G' are the observed genotypes of founders and non-founders, respectively. These depend on the true founder and non-founder genotypes F and F', and on the parameters of the error model α. Founder genotypes depend on the founder haplotypes H, whereas non-founder genotypes depend on both H and the inheritance vector Z. The number of alleles n shared at a locus by two relatives is a function of the inheritance vector. Founder haplotypes are here modelled as products of a Markov process with parameters Q and inheritance vectors result from independent random recombinations with the probabilities of recombinations between adjacent markers collected as vector c.

for founder i at locus j, and similarly for the non-founders. We assume that all markers are biallelic SNPs with alleles labelled "0" and "1," and we denote genotypes "0," "1," and "2" representing "00," "01," and "11" pairs, respectively. Some of the observed non-founder genotypes, and some or all of the observed founder genotypes, may be missing; these are denoted by "-."

Although Figure 1 does not show it, all computations are conditional on the pedigree structure. Throughout this paper, for the sake of simplicity, the data consist of nuclear families with two offspring. However, the algorithm is readily generalized for larger pedigrees.

In this section, we assume that IBD probabilities are to be calculated for a single pair of individuals, which may be founders, non-founders, or one of each. Let the number of alleles shared IBD between these two individuals at marker j be n_j, which may be 0, 1, or 2 and let $n = (n_1,\ldots,n_J)$, where J is the number of markers. The problem is to estimate the probability of each possible value of n_j, for $j = 1,\ldots,J$, given G and G'. These probabilities are denoted $p(n_j \mid G, G')$.

The values n_j are completely determined at each locus by the inheritance vectors, which we collect into a single vector Z. The inheritance vector has binary components $Z_{ii'j}$ for each parent-child pair (i,i') in the pedigree and for each marker locus j. Each component indicates which copy of the chromosome child i' inherited from parent i at locus j, where "0" indicates the chromosome inherited from the father of i and "1" indicates the chromosome inherited from the mother of i. The arrow from Z to n in Figure 1 indicates that n depends on Z.

By tracing back the sequence of inheritance events implied by a given inheritance vector, one can determine which two founder chromosomes were inherited by each individual in the pedigree at each marker. Thus, if the founder haplotypes and the inheritance vector were known, all founder and non-founder genotypes would be completely determined. Let H be the vector of founder haplotypes with components H_{kj} representing the allele at locus j of founder haplotype k,

where the number of founder haplotypes is twice the number of founders. Each H_{kj} may therefore be either "0" or "1." In Figure 1, the true founder and non-founder genotypes are denoted F and F', respectively. These vectors have components of the form F_{ij} and $F_{ij'}$, representing the true genotypes of each individual i and locus j. Note that in Figure 1 the founder genotypes F depend only on H, not on Z, whereas the non-founder genotypes F' depend on both H and Z.

The observed founder genotypes G_{ij} may differ from the true founder genotypes F_{ij}, and similarly for the observed and true non-founder genotypes. We employ a two-parameter error model relating observed genotypes to true genotypes. The model involves the transition matrix shown in Figure 2, which has parameter vector $\alpha = (\alpha_0, \alpha_1)$. The entry in row i and column j is the probability that a true genotype i will be measured as genotype j, for $i,j = 0,1,2$. Thus, Figure 1 indicates that the observed genotypes G and G' are dependent on the true genotypes F and F' and on the parameter vector α. In this paper, we use $\alpha_0 = \alpha_1 = 0.001$. Note that in Figure 1, the observed founder genotypes G and non-founder genotypes G' are dependent on the same error parameter vector α. This does not imply that the location of errors in G is in any way correlated with the location of errors in G'. Rather, it merely indicates that the same genotyping technique has been used for all individuals, whether founder or non-founder, and hence the probability of any particular type of error (such as misreading 01 as 00) is the same for all individuals.

According to our model, the inheritance vector Z is dependent upon the vector c of recombination probabilities between adjacent markers, as indicated in Figure 1. We assume that the probability of an odd number of crossing-over events occurring between markers j and $j+1$ is independent of whether recombination occurs anywhere else. That is, we assume that there is no crossing-over interference. Let this probability be c_j, and let $c = (c_1, \ldots, c_{J-1})$. Various ways of estimating c are possible, but in this paper recombination frequencies are estimated using genomic positions. Specifically, for the human chromosomes that we consider in this paper—7 and 15—the ratio of map length to physical length was calculated using published measurements [Kong et al., 2004] and then Haldane's mapping function ($c = 0.5[1-\exp[-2d/100]]$) was used to estimate the probability of recombination.

Finally, the remaining part of the model specifies how the founder haplotypes H are distributed, and it is here that we introduce a Markov chain to model LD in the founder haplotype population. The model assumes that the allele at the first (i.e. leftmost) locus is "1" with probability q, but that the probability of a "1" allele at the second locus depends on which allele was selected at the first locus, and similarly the probability of a "1" allele at locus j depends on which allele was selected at locus $j-1$,

$$\begin{bmatrix} 1-\alpha_0 & \alpha_0 & 0 \\ \alpha_1 & 1-2\alpha_1 & \alpha_1 \\ 0 & \alpha_0 & 1-\alpha_0 \end{bmatrix}$$

Fig. 2. Transition matrix for the error model. The entry in row *i* and column *j* is the probability that genotype *i* will be observed as genotype *j*.

for $j > 2$. Let the probability of a "1" allele at locus $j \geq 2$ be Q_{j0} if a "0" allele was selected at locus $j-1$ and Q_{j1} if a "1" allele was selected. Let $Q = (q, Q_{20}, Q_{21}, \ldots, Q_{J0}, Q_{J1})$. The value of Q we estimate using genotype data, as described in the following subsection.

In summary, the data in Figure 1 consist of observed genotypes G and G'. The parameters Q, c, and α are estimated directly from the data. The unknowns are H, Z, F, F', and n and we want to determine $p(n_j \mid G, G')$ for $n_j = 0$, 1, or 2 and $j = 1,\ldots,J$.

We can expand $p(n_j \mid G, G')$ as follows:

$$\begin{aligned} p(n_j|G,G') &= \frac{p(n_j,G,G')}{p(G,G')} \\ &= \frac{1}{p(G,G')}\sum_{H,Z}[p(H|Q)p(Z|c)p(G|F(H),\alpha) \\ &\quad \times p(G'|F'(H,Z),\alpha)p(n_j|Z)], \end{aligned}$$

where F is written as $F(H)$ and F' is written as $F'(H,Z)$ to indicate that H and Z completely determine F and F'. The term $p(n^j \mid Z)$ is 1 or 0, depending on whether n^j is compatible or incompatible with the components of the inheritance vector at locus j. The term $p(G \mid F(H),\alpha)$ can be expanded as follows:

$$p(G|F(H),\alpha) = \prod_{i,j} p(G_{ij}|F_{ij}(H_{ij}),\alpha),$$

where terms on the right-hand side are obtained from the matrix shown in Figure 2. The expression $p(G' \mid F'(H,Z),\alpha)$ can be similarly expanded. The term $p(Z \mid c)$ can be expanded as

$$p(Z|c) = \prod_{(i,i')} p(Z_{ii'1}) \prod_{j=2}^{J} p(Z_{ii'j}|Z_{ii'(j-1)}, c_{j-1}),$$

where the first product is over all parent-child pairs (i,i'), $p(Z_{ii'j})$ is 0.5 and $p(Z_{ii'j}|Z_{ii'(j-1)}, c_{j-1})$ is $(1 - c_{j-1})$ if $Z_{ii'j} = Z_{ii'(j-1)}$ and c_{j-1} otherwise.

The main novelty of our method is that we expand $p(H \mid Q)$ in terms of our Markov model as follows:

$$p(H|Q) = \prod_{k} p(H_{k1}) \prod_{j=2}^{J} p(H_{kj}|H_{k(j-1)}, Q_{j0}, Q_{j1}),$$

where $p(H_{k1})$ is $1-q$ or q according to whether H_{k1} is 0 or 1 and $p(H_{kj}|H_{k(j-1)}, Q_{j0}, Q_{j1})$ is $1 - Q_{j0}$, Q_{j0}, $1 - Q_{j1}$, or Q_{j1}, according to whether $(H_{k(j-1)}, H_{kj})$ is (0,0), (0,1), (1,0), or (1,1), respectively.

The summations over H and Z are efficiently computed via dynamic programming. Note that the Lander-Green algorithm is also essentially just dynamic programming.

ESTIMATING TRANSITION PROBABILITIES

To estimate the transition probabilities Q_{j0} and Q_{j1} for each locus j, we used the genotype data supplied as input. Consequently, the larger the number of individuals genotyped, the more accurate the estimates will be. Let the covariance of the haplotypes at locus $j-1$ and j be $\mathrm{cov}(H_{j-1}, H_j)$ where the haplotypes are encoded as "0" or "1." Similarly, let the covariance of the genotypes at locus $j-1$ and j be $\mathrm{cov}(G_{j-1}, G_j)$ where the haplotypes are encoded

as "0," "1," or "2." A straightforward argument (see the Appendix) shows that $\mathrm{cov}(H_{j-1}, H_j) = 0.5\,\mathrm{cov}(G_{j-1}, G_j)$. The right-hand side can be estimated using the sample covariance of the genotypes, giving an estimate also of $\mathrm{cov}(H_{j-1}, H_j)$. The proportion of the "1" allele at loci j and $j-1$ can also be estimated directly from the data. Let the proportion of the "1" allele at locus j be q_j. Then another straightforward argument (see the Appendix) gives that

$$Q_{j0} = q_j - \frac{\mathrm{cov}(H_{j-1}, H_j)}{1 - q_{j-1}}$$

and

$$Q_{j1} = q_j + \frac{\mathrm{cov}(H_{j-1}, H_j)}{q_{j-1}}.$$

An important refinement of this method is to recognize that

$$\begin{aligned} &\max\{-q_{j-1}q_j, -(1-q_{j-1})(1-q_j)\} \\ &\quad \le \mathrm{cov}(H_{j-1}, H_j) \\ &\quad \le \min\{(1-q_{j-1})q_j, q_{j-1}(1-q_j)\}. \end{aligned}$$

If either the upper or lower of these limits is violated, then we set the covariance equal to that limit.

RESULTS

The new model was applied to three data sets. The first involved real parental haplotypes obtained from the HapMap project and used to simulate siblings with known IBD sharing at each locus. The second was a very large simulated data set involving 1,000 parental haplotypes and more than one million markers. This approaches the scale of SNP data that will arise from individual sequencing projects. The third was based on twin-study data supplied by the Queensland Institute of Medical Research.

DATA SET 1—SIMULATED SIBLING GENOTYPES

The raw data from which sibling genotypes were simulated were obtained from the International HapMap Project [Gibbs et al., 2005] and were downloaded from the project website: http://www.hapmap.org/. The data consisted of phased haplotypes obtained for 60 Utah residents with ancestry from northern and western Europe. We used data from chromosome 7 only, consisting of 144351 SNPs. The haplotypes represent the inherited and non-inherited alleles of a single offspring for each of 30 pairs of parents.

We simulated meioses resulting in two offspring for each of the 30 pairs of parents. The probability of recombination occurring between any pair of adjacent markers was set in accordance with the distance between SNPs, using the approximation 1 Mb ≈ 1.2 cM and Haldane's mapping function. The resulting data set consisted of genotype data for 30 sibling pairs, plus known IBD status for each pair and each SNP.

The new algorithm was then run on the full set of 30 sibling pairs. The Merlin package was also run on the same data, with and without accounting for LD using clusters. Run times were ∼4 min for MCIBD, ∼7 min for Merlin without clusters, and ∼4 hr for Merlin with clusters.

To compare the performances of the three algorithms, we first considered only those loci that were known to have 0 alleles shared IBD. For each such locus, we considered the probability of 0 alleles shared IBD as estimated by each algorithm. The cumulative distributions of these probabilities for each algorithm are shown in Figure 3(A). An ideal algorithm would assign probability 1 to all loci with a true IBD status of 0, and thus the cumulative distribution of probabilities would spike at probability 1. Thus, it would appear that our algorithm is performing better than Merlin does without accounting for LD using clusters, but not as well as Merlin does using clusters. Figure 3(B) shows a similar analysis for all loci with a true IBD status of 1. The same conclusion can be drawn. The analysis was also performed for all loci with a true IBD status of 2, but in this case all three algorithms were close to the ideal distribution, that is, all three algorithms assigned a high probability of IBD status 2 to almost all such loci.

DATA SET II—SIMULATED INDIVIDUAL SEQUENCING

In order to investigate whether the algorithms would be able to handle SNP genotype data on the scale that is expected from individual sequencing projects, we simulated a population of 2,000 chromosomes using the computer program ms [Hudson, 2002]. We assumed an effective diploid population size of 10,000, a neutral

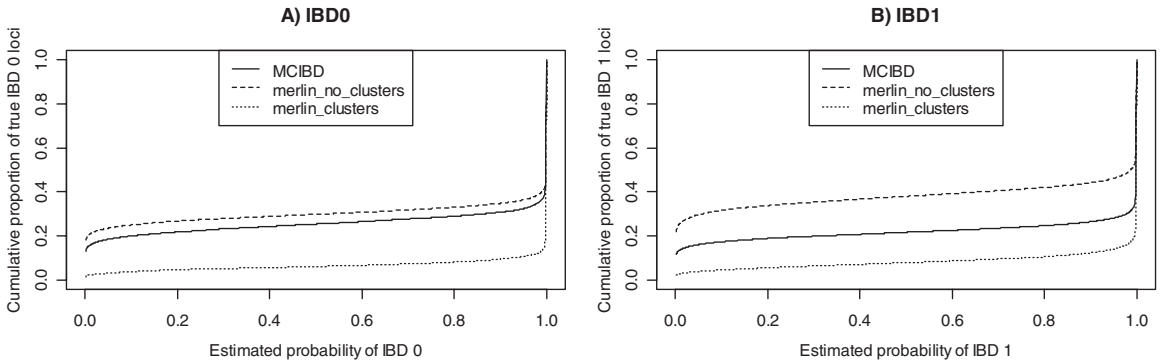

Fig. 3. Cumulative distributions of estimated probabilities of IBD status (A) 0 and (B) 1, for simulated data at loci known to have IBD status 0 (1), using our new algorithm (MCIBD), merlin without clusters, and merlin with clusters. The cumulative proportion is the proportion of loci with true IBD status 0 (1) that has assigned probability of that status less than or equal to a given threshold.

mutation rate of 10^{-8} per base pair per generation, and a recombination rate of 10^{-8} per base pair per generation, and generated a single replicate population. The nominal chromosome length was 300 megabases—approximately the length of the longest human chromosome (chromosome 1). With these parameters, ms produced 1,018,805 SNPs. The 2,000 chromosomes were paired to produce 1,000 diploid individuals. These were then paired to produce 500 pairs of parents. For each pair of parents, genotypes for two children were simulated in the same manner as for Data Set I, except that we used the map to physical distance ratio 1 Mb $\approx$ 1 cM, which is consistent with the parameters supplied to ms.

These data were then analyzed using MCIBD and Merlin with clusters. Parental information was withheld from both algorithms (as for Data Set I), but the true IBD status for each pair of siblings and each marker locus was recorded for the purpose of evaluating the results. Both algorithms were run on a supercomputer with ninety-six 64 bit 1.6 GHz Itanium 2 processor cores and 198 Gigabytes of shared memory. MCIBD completed the task successfully in a little less than $9\frac{1}{2}$ hr, whereas Merlin with clusters had not produced any output after running for 150 hr and was therefore terminated.

The accuracy of MCIBD was comparable to that observed for Data Set I. Almost all loci with 2 alleles IBD were assigned a probability close to 1 of having 2 alleles IBD. However, approximately 25% of loci with 1 allele shared IBD were assigned a probability less than 0.5 of having 1 allele shared IBD, and approximately 29% of loci with no alleles shared IBD were assigned a probability less than 0.5 of having no alleles shared IBD.

DATA SET III—AUSTRALIAN TWINS

The three algorithms were executed on a data set consisting of SNP genotypes for 169 Australian families, each including at least one pair of monozygotic or dizygotic twins. The genotypes were obtained using the Affymetrix Xba 50k SNP array. Only SNPs on chromosome 15 were used, and a subset of 3,030 SNPs was selected. We used the approximation 1 Mb $\approx$ 1.6 cM for chromosome 15. We did not expect much LD with this low SNP density, and thus we did not expect the results to differ much among the three algorithms. However, we found that accounting for LD did indeed make a significant difference for this data set, illustrating the importance of modelling LD.

The entire data set was analyzed using our method, Merlin without clusters and Merlin with clusters. However, we present data for only one pair of dizygotic twins. The genotypes of the parents of these twins were removed from the data set. The estimated probabilities of IBD status 0 or 1 are plotted for each of the three methods in Figure 4. The estimated probability of IBD status 2 at each locus is just one minus the sum of the other two probabilities.

Note firstly that Merlin, with or without clusters, finds a high probability of IBD status 1 at a few loci around marker 2,720, whereas MCIBD finds a high probability of IBD status 2. Note also that all three algorithms assign a high probability of IBD status 2 to the surrounding loci from about marker 2,400 to the end. This entire region consists of loci at which the siblings have two alleles identical by state, except for a single marker at locus 2,720 where there is only one allele identical by state. The most likely explanation is that all loci in this region have two alleles IBD and that a genotyping error has occurred at locus 2,720. Our algorithm allows for genotyping error, and hence can detect this. Merlin, however, cannot allow a locus with only one allele identical by state to be assigned IBD status 2. In fairness, it should be noted that the Merlin package also provides a separate method for detecting probable genotyping errors, which has not been used here.

Secondly, note that the main region of uncertainty about IBD status appears to be between markers 400 and 1,100. Here, Merlin with clusters assigns a probability close to 1 of IBD status 0, whereas Merlin without clusters assigns lower probabilities of IBD status 0 throughout this region, and indeed there are two subregions (between markers 600 and 800 and between markers 900 and 1,100) where IBD status 1 is assigned a higher probability than IBD status 0. The new algorithm produces intermediate results—it also assigns lower probabilities to IBD status 0 in the region between markers 400 and 800, but there is only one subregion (between markers 600 and 800) where IBD status 1 is assigned a higher probability.

These data can also be used to compare the precision of the three algorithms. The precision at a locus can be quantified in terms of information using the formula:

$$\mathrm{INF} = 1 - 8 * \mathrm{PEV},$$

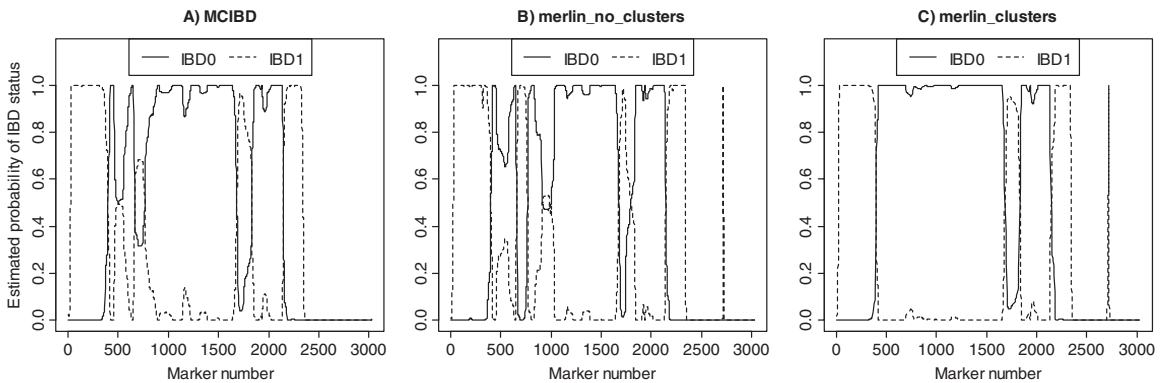

Fig. 4. Estimated probabilities that 0 or 1 alleles are shared IBD at 3,030 loci across chromosome 15 in a pair of dizygotic twins, calculated using (A) MCIBD, (B) Merlin without accounting for LD, and (C) Merlin with haplotype clusters.

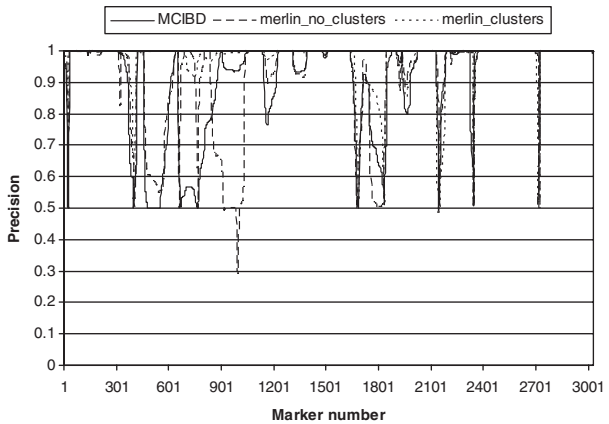

Fig. 5. Precision of the three algorithms for the twin study data.

where

$$\mathrm{PEV} = 0.25P_1(1 - P_1) + P_2(1 - P_2) - P_1P_2$$

and P_1 and P_2 are the estimated probabilities of sharing 1 and 2 alleles IBD, respectively, at a given locus [Visscher et al., 2006]. Figure 5 shows this measure of precision for each locus and each algorithm. The three algorithms do not differ dramatically in their precision for this data, but there are two regions where the differences are noteworthy. The first region is between markers 600 and 800. Here, Merlin without clusters obtains a high precision, but is probably inaccurate, given that it finds the most probable IBD state here to be 1, whereas Merlin with clusters (which our simulated results above suggest is generally the most accurate of the three algorithms) finds it to be 0. In this region, MCIBD also predicts IBD status 1 as being most probable, but recognizes the low precision of this prediction. The second noteworthy region is between markers 900 and 1,100. Here, Merlin without clusters has a substantially lower precision than the other two methods, and indeed it is also likely to be inaccurate in assigning a higher probability to IBD status 1. Both regions illustrate the point that accounting for LD can improve precision and accuracy, even for markers that do not have high density.

DISCUSSION

The new model and algorithm presented here have several features that make it an attractive approach to estimating IBD probabilities in comparison to the algorithm implemented in Merlin. The first is that it accounts for LD in the founder haplotype population using a much simpler model than that employed by Merlin, and as a result it is several orders of magnitude faster (compare the ~4 min run time of our algorithm to the ~4 hr run time of Merlin on the simulated data). Indeed, for the test data presented here, the new algorithm runs in less time than Merlin does even without identifying haplotype blocks.

The gain in efficiency will be of great importance as the SNP density of genotyping arrays increases. It will soon be necessary to estimate IBD probabilities for millions or tens of millions of SNPs across hundreds or thousands of individuals. This is not currently feasible for Merlin using clusters to account for LD, given that the run time of Merlin using clusters scales supralinearly. Indeed, in our analysis of 500 sibling pairs with just over one million simulated SNPs (Data Set II) Merlin had not produced any output after running for 150 hr. However, our new algorithm scales linearly and was able to complete the calculations in $9\frac{1}{2}$ hr.

A second advantage is that it identifies genotyping errors automatically, without requiring a separate step of analysis. This feature will become even more important with the advent of individual whole-genome sequencing data, since errors are expected to be more common under this scenario than for current genotyping platforms, at least initially.

Our results for both real and simulated data emphasize the need to account for LD in the calculation of IBD probabilities. The results shown in Figure 3 indicate that accounting for LD results in the correct IBD state being assigned high probability at many more loci than if LD is not taken into account. Figures 4(A) and 4(C), when compared to Figure 4(B), illustrate that accounting for LD can substantially alter the most probable IBD state in some regions. These results confirm earlier findings [Schaid et al., 2002].

It must be admitted, however, that Figure 3 indicates modelling LD using clusters produces superior results to modelling it using a Markov model. Presumably the reason for this is that the block model captures long-range LD that the Markov model cannot. Moreover, we found that the accuracy of MCIBD for the large-scale Data Set II was comparable to that obtained for Data Set I. This suggests that the discrepancy between MCIBD and Merlin with clusters seen in Figure 3 is not merely a result of the small sample size used in Data Set I. Thus, there is genuine room for improvement to our algorithm. It may be possible to modify the model for LD used here to include a small number of well chosen long-range dependencies, and thus to improve the accuracy of our approach without sacrificing too much efficiency. It may also be possible to identify common features of those sites where our algorithm assigns a high probability to an IBD state other than the true state, and modify the algorithm accordingly. This matter warrants further investigation.

In future work, we intend to enhance the model presented here by including some longer range associations between markers. One possibility to be explored is the use of a second- or higher-order Markov chain to model marker-marker LD. In the related context of Hidden Markov Model haplotype inference, Sun et al. [2007] found that increasing the order of Markov process used in modelling adjacent SNP alleles in "ancestral" haplotypes from first order to third lead to significant improvement in accuracy of imputation. The efficiency of that window size may merely reflect the density of genotyping in the examples studied. Another is to use multiple, interleaved Markov chains with transition probabilities calculated for non-adjacent markers. We also intend to extend the method presented here for larger pedigrees, possibly incorporating the methods here into a Markov chain Monte Carlo approach. One further enhancement that will be useful for analyzing data from individual whole-genome sequences would be to account for fluctuations in sequencing error rates that depend upon how repetitive the surrounding sequence is, and also on the nucleotides that are being sequenced. One could

allow the α parameter to depend upon external factors such as sequence position or nucleotide (and nearby nucleotides).

ACKNOWLEDGMENTS

We thank Nick Martin and Grant Montgomery for access to data.

REFERENCES

Abecasis GR, Cardon LR, Cookson WOC. 2000a. A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279–292.

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

Abecasis GR, Cookson WOC, Cardon LR. 2000b. Pedigree tests of transmission disequilbrium. Eur J Hum Genet 8:545–551.

Abecasis GR, Wigginton JE. 2005. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77:754–767.

Cudworth AG, Woodrow JC. 1975. Evidence for HL-A-linked genes in ''juvenile'' diabetes mellitus. Br Med J 3:133–135.

Elston RC, Stewart J. 1971. A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542.

Gibbs RA, Belmont JW, Boudreau A, Leal SM, Hardenbol P, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, et al. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Hayden EC. 2008. International genome project launched. Nature 451:378–379.

Heath SC. 1997. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61: 748–760.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Kong A, Cox NJ. 1997. Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188.

Kong X, Murphy K, Raj T, He C, White P, Matise T. 2004. A combined linkage-physical map of the human genome. Am J Hum Genet 75:1143–1148.

Krugylak L, Daly MJ, Lander ES. 1995. Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56:519–527.

Krugylak L, Daly MJ, Reeve-Daly MP, Lander ES. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2263–2367.

Lange K, Elston RC. 1975. Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. Hum Hered 25:95–105.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G and others. 2007. The diploid genome sequence of an individual human. PLoS Biol 5:e254.

Liu J, Liu Y, Liu X, Deng H-W. 2007. Bayesian mapping of quantitative trait loci for multiple complex traits with the use of variance components. Am J Hum Genet 81:304–320.

Motro U, Thomson G. 1985. The affected sib method. I. Statistical features of the affected sib-pair method. Genetics 110:525–538.

O'Connell JR. 2001. Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. Hum Hered 51:226–240.

O'Connell JR, Weeks DE. 1995. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nature Genet 11:402–408.

Qiu J, Hayden EC. 2008. Genomics sizes up. Nature 451:234.

Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN. 2002. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. Am J Hum Genet 71:992–995.

Sun S, Greenwood CMT, Neal RM. 2007. Haplotype inference using a Bayesian Hidden Markov Model. Genet Epidemiol 31:937–948.

Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet 2:316–325.

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

Wijsman EM, Rothstein JH, Thompson EA. 2006. Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. Am J Hum Genet 79:846–858.

APPENDIX

Proof that $\mathrm{cov}(H_{j-1}, H_j) = 0.5\,\mathrm{cov}(G_{j-1}, G_j)$: Recall that genotypes take the values 0, 1, or 2 and haplotypes take the values 0 or 1 and note that $G_j = H_j^{(1)} + H_j^{(2)}$, where $H_j^{(1)}$ and $H_j^{(2)}$ are the haplotypes for locus j on the two homologous chromosomes. If haplotypes are independent, then

$$\begin{aligned}\mathrm{cov}(G_{j-1}, G_j) &= \mathrm{cov}(H_{j-1}^{(1)} + H_{j-1}^{(2)}, H_j^{(1)} + H_j^{(2)}) \\ &= \mathrm{cov}(H_{j-1}^{(1)}, H_j^{(1)}) + \mathrm{cov}(H_{j-1}^{(1)}, H_j^{(2)}) \\ &\quad + \mathrm{cov}(H_{j-1}^{(2)}, H_j^{(1)}) + \mathrm{cov}(H_{j-1}^{(2)}, H_j^{(2)}) \\ &= \mathrm{cov}(H_{j-1}, H_j) + 0 + 0 + \mathrm{cov}(H_{j-1}, H_j)\end{aligned}$$

and the result follows immediately.

Proof that

$$Q_{j0} = q_j - \frac{\mathrm{cov}(H_{j-1}, H_j)}{1 - q_{j-1}}$$

and

$$Q_{j1} = q_j + \frac{\mathrm{cov}(H_{j-1}, H_j)}{q_{j-1}}$$

This result is obtained as the solution to two simultaneous equations. The first equation results from the fact that the probability of a ''1'' allele occurring at locus j is a function of q_{j-1}, Q_{j0} and Q_{j1} as follows:

$$q_j = (1 - q_{j-1})Q_{j0} + q_{j-1}Q_{j1}.$$

The second equation results from the fact that $\mathrm{cov}(H_{j-1}, H_j)$ is given by

$$\mathrm{cov}(H_{j-1}, H_j) = x_{11} - q_{j-1}q_j = q_{j-1}Q_{j1} - q_{j-1}q_{j.}$$

Solving these equations for Q_{j0} and Q_{j1} simultaneously we obtain the expressions above.