On identification problems requiring linked autosomal markers

Thore Egeland[a*] Nuala Sheehan[b]

[a]Department of Medical Genetics,

Ulleval University Hospital, 0407 Oslo, Norway

[b]Departments of Health Sciences and Genetics,

University of Leicester, 2nd Floor Adrian Building

University Road, Leicester LE1 7RH, UK

[0]* Corresponding author. Phone: +4722118579 Fax: +4722119899

E-mail address: Thore.Egeland@medisin.uio.no

On identification problems requiring linked autosomal markers

Thore Egeland[a*] Nuala Sheehan[b]

[a]Department of Medical Genetics,

Ulleval University Hospital, 0407 Oslo, Norway

[b]Departments of Health Sciences and Genetics,

University of Leicester, 2nd Floor Adrian Building

University Road, Leicester LE1 7RH, UK

Abstract

This paper considers identification problems based on DNA marker data. The topics we discuss are general, but we will exemplify them in a simple context. There is DNA available from two persons. There is uncertainty about the relationship between the two individuals and a number of hypotheses describing the possible relationship is available. The task is to determine the most likely pedigree. This problem is fairly standard. However, there are some problems that cannot be solved using DNA from independently segregating loci. For example,

[0]* Corresponding author. Phone: +4722118579 Fax: +4722119899

E-mail address: Thore.Egeland@medisin.uio.no

the likelihoods for (i) grandparent-grandchild, (ii) uncle-niece and (iii) half-sibs coincide for such DNA data and so these relations cannot be distinguished on the basis of markers normally used for forensic identification problems: the likelihood ratio comparing any pair of hypotheses will be unity.

Sometimes, but not in the examples we consider, other sources of DNA like mtDNA or sex chromosomes can help to distinguish between such equally likely possibilities. Prior information can likewise be of use. For instance, age information can exclude alternative (i) above and also indicate that alternative (iii) is apriori more likely than alternative (ii).

More generally, the above problems can be solved using linked autosomal markers. To study the problem in detail and understand how linkage works in this regard, we derive an explicit formula for a pair of linked markers. The formula extends to independent pairs of linked markers. While this approach adds to the understanding of the problem, more markers are required to obtain satisfactory results and then the Lander-Green algorithm is needed. Simulation experiments are presented based on a range of scenarios and we conclude that useful results can be obtained using available freeware (MERLIN and R).

The main message of this paper is that linked autosomal markers deserve greater attention in forensic genetics and that the required laboratory and statistical analyses can be performed based on existing technology and freeware.

Keywords: Identification; likelihoods; linked autosomal markers

1 Introduction

This paper deals with relationship estimation based on DNA-data. There is an extensive literature on this general problem for unlinked markers and a recent review is provided in [1]. There are some distinguishing features of the problems we address and the solutions we propose. First, we restrict attention to pairwise problems assuming that DNA is available from two persons and the task is to determine the relationship between these individuals. This is a problem of great practical importance arising in various contexts. For example, consider the situation where a disaster has wiped out a large part of an individual's family. A body is found, and DNA data is available from the two individuals. The problem is to estimate the relationship between the deceased and the survivor. There is no theoretical problem in the extension from pairwise to joint relationship. Second, the possible relationships are listed and the objective is to determine the most likely. The problem is much harder if the alternatives are unspecified. Thirdly, and this is an important distinction between this and previous work, we consider problems that cannot be solved using DNA from any number of independently segregating loci. For example, the likelihoods for (i) grandparent-grandchild, (ii) uncle-niece and (iii) half-sibs coincide for such DNA data and so these relations cannot be distinguished on the basis of markers normally used for forensic identification problems. Thompson [2] provides an early discussion of this problem and Thompson and colleagues have revisited and extended the discussion in subsequent writings including [3] and [4]. In the latter paper the relevance of linked markers is summarised as follows "...the

three relationships have distinct consequences for data at linked loci, since each provides a different probability that the two relatives share one gene identical by descent at both of two loci". A large number of markers might be required to distinguish between alternatives that have equal likelihoods for independently segregating loci. In [5] as many as 399 markers are used. The number of markers is determined by the chosen distance between markers explaining the odd figure 399. The calculations of the latter paper are only approximate for avuncular relations like alternative (ii) above. Our calculations will be exact, based on an explicit formula in a simple case and on the freeware MERLIN [6] in the more general case. The number of markers used in [5] may be too small for some purposes and we provide examples with 3820 markers.

The next section presents the basic methods. Linked autosomal markers will be the main focus, but some alternative or supplementary approaches based on mtDNA, sex chromosomes and prior information will be mentioned. In the results section identification problems are solved that are unsolvable based on standard forensic markers. Our main message is that linked autosomal markers deserve greater attention in forensic applications.

2 Methods

We formulate the problem in a Bayesian context. This is done since this approach handles cases with more than two alternatives conveniently. Furthermore, if there is prior, non-DNA, information that the user would like to include, this can be easily accommodated. How-

ever, our approach by no means implies that a Bayesian analysis is required.

There are competing hypotheses $H_1, ..., H_n$ having prior probabilities $\pi_1, ..., \pi_n$, respectively. One hypothesis corresponds to a specific pedigree. The values $\pi_i = 1/n$ reflect a flat prior whereby all hypotheses are assumed to be equally likely in the absence of data and will be used for our examples. More general priors are discussed in [7] and further exemplified in [8]. Let $L_i \equiv L(data|H_i)$ be the likelihood of the data calculated assuming hypothesis H_i to be true. By Bayes' Theorem, the posterior probability of H_i is

$$P(H_i|data) = \frac{L_i\pi_i}{\sum_{i=1}^{n} L_i\pi_i} = \frac{L_i}{\sum_{i=1}^{n} L_i}, \tag{1}$$

where the last equality applies for a flat prior. This last equality leads to a meaningful frequentist version: the likelihood of one hypothesis is compared to the sum of the others. However, this is not the traditional forensic approach and in particular it does not yield the classical paternity index for the case of two alternatives. Rather, classical pairwise comparisons are made:

$$\frac{P(H_i|data)}{P(H_j|data)} = \frac{L_i}{L_j}\frac{\pi_i}{\pi_j} = \frac{L_i}{L_j} \text{ for any } i \neq j \tag{2}$$

expressing the posterior probability ratio on the left hand side as the product of the likelihood ratio, L_i/L_j, and the prior ratio, π_i/π_j. Again, the right hand side of the equation assumes a flat prior and coincides with the conventional LR (likelihood ratio). There is also a simple relation to Essen-Möller's W [9] since $W = P(H_i|data) = LR/(1+LR)$ is the posterior probability corresponding to two equally likely prior alternatives.

The pedigrees of Figure 1, corresponding to the following hypotheses

$$\begin{aligned} H_1 &: \text{A is the grandparent of B,} \\ H_2 &: \text{A is the niece of B,} \\ H_3 &: \text{A is the half-sib of B.} \end{aligned} \tag{3}$$

will be used to exemplify the methods throughout as they all have equal likelihood for unlinked markers . However, we emphasize that the approach applies generally and is not restricted to this example.

Sometimes, but not in the examples we consider, other sources of DNA like mtDNA [10], X-chromosomes [11, 5] or Y-chromosomes [12] can be helpful. Prior information can likewise be of use. For instance, age information can exclude hypothesis H_1 above by assigning $\pi_1 = 0$ in (1). Prior information can also indicate that H_3 is apriori more likely than alternative H_2. In this paper, we will not assume that such prior information is available.

The remaining part of this section discusses the calculation of the likelihoods required for Equation (2). We will present likelihood calculations for each of the following cases:

1. one marker,
2. two linked markers,
3. independent pairs of linked markers,
4. general case.

2.1 One marker

The likelihoods can be calculated analytically for the pedigrees corresponding to Figure 1 in several ways. In our context the IBD concept will prove convenient to show that the likelihoods L_i corresponding to the hypotheses H_i, $i = 1, 2, 3$, coincide. Alleles that have descended from a single ancestral allele are said to be identical by descent, IBD. The likelihood for a pair of individuals for one marker depends on the pedigree describing their relationship only through the IBD-probabilities. For pedigrees 1, 2 and 3 of Figure 1, individuals A and B share no, one or two alleles with probabilities 0.5, 0.5 and 0 respectively. Since these probabilities are identical, so are the likelihoods. This is noted in [2] along with a more detailed account of IBD probabilities and reference to earlier work. The likelihoods can also be calculated explicitly. Note that for $i = 1, 2, 3$

$$\begin{aligned} L(data|H_i) &= L(data|I=0)P(I=0) + L(data|I=1)P(I=1) \\ &+ L(data|I=2)P(I=2) \end{aligned}$$

where I is the number of IBD alleles. For the pedigrees of Figure 1,

$$L(data|H_i) = L(data|I=0)0.5 + L(data|I=1)0.5.$$

The right hand side of the above equation can be evaluated for specific marker data using Table 1, based on [2]. For instance, if both individuals are homozygous a,a and the allele frequency is p_a then

$$L(data|H_i) = p_a^4 0.5 + p_a^3 0.5.$$

The above equation as well as remaining likelihood calculations of this paper assumes Hardy-Weinberg equilibrium.

2.2 Two linked markers

The distinguishing feature of this paper compared to *forensic* science texts like [13] and [14] is the need to consider linked autosomal markers. At least two linked markers are required to distinguish the pedigrees of Figure 1. The required number of markers depends on how informative they are and we elaborate on this in the discussion section. Some concepts from linkage analysis are needed to explain the methods. There are several classical introductions to linkage analysis like [15] and there are also more recent reviews [16]. We will briefly review the required background when the need arises.

Consider two markers on the same chromosome string. The distance between the markers can be measured by r, the *recombination probability*. Generally $0 \leq r \leq 0.5$ where $r = 0.5$ corresponds to the markers being unlinked. For $r < 0.5$ the markers are linked. Let $k_{11}^i(r)$ denote the probability that two individuals whose relation is described by pedigree i have one allele IBD at two markers separated by a distance of r. For the pedigrees of Figure 1

$$k_{11}^i(r) = \begin{cases} (1-r)/2 & i = 1, \\ R/2 & i = 2, \\ (2(1-r)R + r)/4 & i = 3. \end{cases} \tag{4}$$

where $R = r^2 + (1-r)^2$. These functions are plotted in Figure 2. A derivation of the above equation based on [3] is provided in the appendix. Equation 4 is also reproduced in slightly different form as Table 1 of [4]. The function values coincide for $r = 0.0$ corresponding to complete linkage, i.e., there is effectively only one marker and $r = 0.5$ when there is no linkage and the loci are segregating indepen-

dently. If the distance between markers can be chosen, it would be wise based on power considerations to select a value of r maximizing the difference between the k-functions. For instance, $r = 0.25$ maximises the difference $k^1_{11}(r) - k^2_{11}(r)$ and so this choice is optimal if the purpose is to distinguish between pedigrees 1 and 2 of Figure 1. Other comparisons lead to other optimal choices for r. In the absence of exact information, $r = 0.25$ is a good choice. The curves corresponding to $i = 2$ and 3 are the closest and we can anticipate that the corresponding pedigrees will be the hardest to distinguish.

The likelihoods for two linked markers corresponding to the pedigrees of Figure 1 depend on the pedigree only through the IBD probabilities given in Equation (4). An explicit formula for this likelihood, $L(\text{data}|\text{ped. i})$ is derived in the appendix and appears as Equation (10).

2.3 Independent pairs of linked markers

While one pair of markers may be relevant for the understanding of the problem, more markers are of course required to obtain useful results. The first obvious extension is to consider independent pairs of linked markers. Let j denote one such pair on chromosome j and assume that one pair of markers is available on each autosome. Then

$$L(\text{data}|\text{ped. i}) = \prod_{j=1}^{22} L(data_j|\text{ped. i}) \tag{5}$$

It may be possible to extend the number of markers if independent pairs of markers can be obtained on the same chromosome. Recall,

however, that the markers in the pair should be separated by some distance to be of use.

2.4 General case

The approaches described so far only use a small fraction of the markers available. It is obviously of interest to use a much larger number of markers. Likelihoods must then be calculated numerically and the Lander-Green algorithm [17] is the basic engine in modern computing packages. This algorithm is based on a hidden Markov model for the unobserved IBD status along the chromosome. There are several freeware implementations and we will be using the program MERLIN [6]. For large complex pedigrees simulation based methods may be required and MCMC has been implemented in the freeware programs SIMWALK2 [18] and Morgan [19].

3 Results

This section consists of two examples. The first illustrates the analytical approach based on Equation (5) and illustrates how the recombination fraction or distance between markers influences the result. The second example uses a much larger number of markers and numerical results are obtained using MERLIN. The data for Examples 1 and 2 are simulated in MERLIN for individuals A and B of Figure 1 using Haldane's map function. For Example 1, 400 simulations were performed whereas Example 2 is more computer intensive and the number of simulations was reduced to 100. The results reported in Tables 2 and 3 below and Figures 3 and 4 are based on these simulations. Mark-

ers are assumed to be in linkage equilibrium and there are four alleles with equal allele frequencies. There is a number of parameter settings that can be varied. This has not been given priority in the coming examples; we have chosen to emphasise more fundamental issues in the examples rather than provide detailed sensitivity analyses. Some of these assumptions are discussed further in Section 4.

3.1 Example 1

For this first example we consider the case motivating this paper, i.e., the hypotheses formulated in Equation (3). In the appendix, analytical results are worked out for one pair of linked markers and the influence of parameters on the resulting likelihoods is discussed. One pair of markers is obviously of little practical use and the immediate extension is to consider pairs of independent markers and the likelihood given in (5). We simulated data for 22 pairs of markers using MERLIN. The calculations are implemented in R; numerical results have been confirmed for selected cases using MERLIN. The distance between the markers in a pair was varied from 0 to 0.5 with steps of 0.05. Figure 3 shows the posterior probabilities when data were simulated assuming H_1, the grandparent - grandchild alternative, to be true. The true alternative comes out as the most likely when it should, but only marginally so. Figure 4 displays the same information as Figure 3 but the LR-s are presented rather than posterior probabilities. The relation between LR-s and posterior probabilities is given in Equation (2). LR-s require a reference pedigree or hypothesis and the uncle-niece alternative has been chosen in Figure 4. From

Figure 3 and 4 we note that alternatives 2 and 3 are the closest alternatives and the hardest to distinguish. This confirms the observations based on the $k-$ functions of Equation (4) and Figure 1.

3.2 Example 2

This example expands on the previous by considering a much larger number of markers. An extra alternative, H_4, corresponding to A and B being sibs, is also added to allow for extra comparisons.

The resulting posterior probabilities or equivalently scaled likelilihoods, are given in Table 2 based on Equation 1. The first column of the table gives information on the markers used. For instance '20 chr; 3820 markers' indicates that 3820 markers evenly spread on 20 chromosomes have been used. The distance between the markers is 1cM, corresponding roughly to $r = 0.01$. The second column shows the 'True R', i.e., the relationship from which data has been simulated. For the alternative '20 unlinked markers', the posteriors for the first three relationships are the same as explained earlier. For instance, when data is simulated from the grandparent-grandchild alternative, this posterior probability is 0.302 while the corresponding figure for the sibs alternative is 0.093. Observe that readers preferring likelihood ratios can obtain these easily: For the above example the likelihood ratio is obtained as 0.302/0.093=3.2 for a flat prior. As more and more linked markers are introduced results improve and for the largest data set the posterior for the grandparent-grandchild relationship is 0.976. Observe that there is a considerable improvement moving from 400 markers (inter marker distance 10cM), corresponding to the amount

of data used in [5] to 3820 markers. From Table 2 it again appears to be hardest to distinguish between 'half-sibs' and 'uncle-nephew' and the posterior probability for the true relationship exceeds 0.5 only when the greatest amount of data is used. This is consistent with the previous example.

Table 3 is based on the same simulated data, but now *classification rates* comparable to those in [5] are presented, i.e., the fraction of times the indicated relationship has the largest likelihood (or equivalently largest posterior probability when flat priors are used). For instance, simulating from the grandparent-grandchild relation with 3820 markers, the true relationship comes out with the largest likelihood for 395 of the 400 simulations, corresponding to 98.8%.

4 Discussion

The approach using independent pairs of linked markers does not lead to acceptable discrimination between the alternatives. However, for a sufficient number of linked markers, acceptable results are obtained using available freeware for calculations. The main message of this paper is that linked autosomal markers deserve greater attention in forensic genetics.

Consideration of linked autosomal markers comes with a cost. For a fixed number of markers and a specific pedigree, there is more information in unlinked markers as pointed out in [4]. Furthermore, some additional parameters need to be specified for linked markers. In particular, the genetic map describing the location of markers must be specified. The relation beetween distance measured in cM (centi-

Morgan) and recombination fraction must also be specified. A common choice is Haldane's map function [15]. These additional parameters and additional assumptions may complicate matters and according to [2] "...the use of linked markers is best avoided when possible". For court applications it is a great advantage to use methods generally agreed on and using linked markers may be lead to debate. However, there is no alternative for some cases. Moreover, some important cases do not involve court proceedings and controversy may be less of an issue.

The assumption of linkage equilibrium [15] is principally a different problem that may arise when a large number of markers is used for calculations of pedigree likelihoods. When markers are close, this assumption may be violated. It is hard to give definite rules regarding acceptable distance between markers. Linkage disequilibrium varies considerably within an individual genome and there is also considerable difference between populations. The only case where linkage disequilibrium may possibly be a problem for this paper, is when 3820 markers are used. MERLIN produces markers where this assumption holds by construction. The effects of linkage disequilibrium on linkage analysis have been considered [20] and there are also options in MERLIN designated to handle this problem although these are somewhat adhoc. Linked markers and linkage disequilibrium has also been discussed in [21] and [22], the latter with reference to DNA match probabilities for siblings and half-siblings. While the modelling of linkage disequilibrium is still being debated, the effects of any departures from linkage equilibrium on the calculations we have presented are undeniably important and should be central to the sensitivity analyses that

we have deliberately omitted from this particular paper.

We have assumed Hardy-Weinberg equilibrium. This is required for Table 1. It would be possible to include coancestry [23, 14, 24]. Obviously, the majority of case work can be solved satisfactorily with independently segregating loci. However, we maintain that there are important problems that cannot be solved unless linked markers are used. Furthermore, the information on maps and parameters needed for the analyses is becoming increasingly reliable and accurate.

We have restricted attention to pairwise estimation problems. If DNA is available from a person related to both of the individuals, the problem will typically become much easier and there may no longer be a need to consider linked markers [25].

Mutations were not considered for our likelihood calculations and we maintain that it is not probably worthwhile to model mutations for the applications we have considered. The mutation rates for the markers used in linkage and association applications are much smaller than the rates for forensic markers. For the pedigrees we have considered, mutation will be confounded with errors. The large number of markers involved necessarily leads to greater problems related to errors, see [5]. This is a topic that needs further investigation with a view to forensic applications.

Finally, we emphasise that it is important to be aware of the problem of pedigrees with identical likelihoods for independent markers. If, for instance, the result of a case work based on traditional forensic markers is to conclude that two individuals are half sibs, it is important to realise that there is no information in the DNA that allows the uncle-niece or grandparent-grandchild alternatives to be excluded.

5 Appendix

We first provide a derivation of Equation (4) based largely on pages 25 and 26 of [3]. The probability of alleles being IBD for a specific locus is 1/2 for all three relations. For the grandparent - grandchild alternative, the alleles received by the parent must be passed on to the child without recombination. This occurs with probability $1-r$ and so $k_{11}^1(r) = (1-r)/2$. Turning to the half-sib alternative, alleles at the first locus must again be IBD. The second locus is IBD if there is a recombination in the segregation to both offspring (occurring with probability r^2) or to neither (occurring with probability $(1-r)^2$). Consequently, $k_{11}^2(r) = R/2$ where $R = r^2 + (1-r)^2$. It remains to deal with the uncle - niece relationship and some further notation is useful:

$$\begin{aligned} E &= \text{'No recomb. in the maternal chromosome bit received by B'}, \\ I_j &= \text{'The number of IBD alleles for marker } j, j = 1, 2'. \end{aligned}$$

Then

$$\begin{aligned} k_{11}^3(r) &= P(I_1 = 1, I_2 = 1|E)P(E) + P(I_1 = 1, I_2 = 1|E^c)P(E^c) \\ &= P(I_1 = 1, I_2 = 1|E)(1-r) + P(I_1 = 1, I_2 = 1|E^c)r \quad (6) \end{aligned}$$

and

$$P(I_1 = 1, I_2 = 1|E) = R/2. \quad (7)$$

The latter equation holds since in this case the markers passed on to A without recombination from her mother must be IBD to the markers in the uncle. The probability that one marker is IBD is 1/2 and then for the other marker to be IBD there must either be none or two

crossovers. When E^c is true the niece has received one paternal and one maternal allele. The probability that the uncle received the same two alleles is 1/4 and so

$$P(I_1 = 1, I_2 = 1|E^c) = \frac{1}{4}. \tag{8}$$

Inserting Equations (7) and (8) into (6) produces the required result and the argument is completed.

We next derive the likelihood for the hypotheses of Equation (3) for two linked markers. Then

$$\begin{aligned} & L(\text{data}|\text{ped. i}) = P(\text{data}|\text{ped. i}) \qquad (9) \\ = \;& k^i_{00}(r)P(\text{data}|I_1 = 0, I_2 = 0) \\ + \;& k^i_{10}(r)P(\text{data}|I_1 = 1, I_2 = 0) \\ + \;& k^i_{01}(r)P(\text{data}|I_1 = 0, I_2 = 1) \\ + \;& k^i_{11}(r)P(\text{data}|I_1 = 1, I_2 = 1) \end{aligned}$$

where $k^i_{uv}(r) = P(I_1 = u, I_2 = v)$. The expression for $k^i_{11}(r)$ is given in Equation (4). Equation (9) can be simplified for our application since $k^i_{00}(r) = k^i_{11}(r)$ as shown below:

$$\begin{aligned} k^i_{1,1}(r) = P(I_1 = 1, I_2 = 1) &= P(I_2 = 1|I_1 = 1)P(I_1 = 1) \\ &= (1 - P(I_2 = 0|I_1 = 1))P(I_1 = 0) \end{aligned}$$

since $P(I_1 = 0) = P(I_1 = 1)$ for the pedigrees we consider. The symmetry between markers 1 and 2 implies that

$$\begin{aligned} k^i_{1,1}(r) &= (1 - P(I_1 = 1|I_2 = 0)P(I_1 = 0) \\ &= P(I_1 = 0|I_2 = 0)P(I_1 = 0) = k^i_{0,0}(r) \end{aligned}$$

Using the above equation, the symmetry identity $k^i_{01}(r) = k^i_{10}(r)$ and the fact that the $k-$functions add to unity for fixed r, Equation (9) simplifies to

$$L(\text{data}|\text{ped. i}) = (p_{00} + p_{11} - p_{10} - p_{01})k^i_{1,1}(r) + \frac{1}{2}(p_{10} + p_{01}) \quad (10)$$

where

$$\begin{aligned} p^i_{uv}(r) &= L(\text{data}|I_1 = u, I_2 = v) \quad (11) \\ &= P(\text{data marker 1}|I_1 = u)P(\text{data marker 2}|I_2 = v) \end{aligned}$$

and the right hand side is provided in Table 1.

To illustrate how equation (10) is used, assume individual A is homozygous (1,1) for both markers while B is also homozygous at both markers, but for another allele. It is then impossible that A and B share alleles IBD. Equation (10) simplifies to $L(\text{data}|\text{ped. i}) = p_{00}k^i_{1,1}(r)$ and the LR comparing hypothesis H_1 to H_2 therefore becomes

$$LR = \frac{p_{00}k^1_{1,1}(r)}{p_{00}k^2_{1,1}(r)} = \frac{1-r}{r^2 + (1-r)^2} \quad (12)$$

where $k_{11}(r)$ is given in (4). Observe that this LR is unity for $r = 0$ and $r = 0.5$ as it should. For other values of r the LR exceeds unity and a maximum value of 1.21 occurs for $r = 0.29$ (details omitted). This indicates a modest contribution for data of this type to distinguish between the hypotheses.

References

[1] BS Weir, AD Anderson, and AB Hepler. Genetic relatedness analysis: modern data and new challenges. *Nature Review Genetics*,

7:771–780, 2006.

[2] E A Thompson. The estimation of pairwise relationships. *Annals of Human Genetics*, 39:173–188, 1975.

[3] E A Thompson. *Pedigree Analysis in Human Genetics*. The Johns Hopkins University Press, Baltimore, 1986.

[4] E A Thompson and T R Meagher. Genetic linkage in the estimation of pairwise relationships. *Theoretical and Applied Genetics*, 97:857–864, 1998.

[5] MP Epstein, WL Duren, and M Boehnke. Improved inference of relationship for pairs of individuals. *American Journal of Human Genetics*, 67:1219–1231, 2002.

[6] G.R. Abecasis, S.S. Cherny, W.O. Cookson, and L.R. Cardon. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30:97–101, 2002.

[7] N A Sheehan and T Egeland. Structured incorporation of prior information in relationship identification problems. *Annals of Human Genetics*, 71:501–518, 2007.

[8] N A Sheehan and T Egeland. Adjusting for founder relatedness in a linkage analysis using prior information. *Human Heredity*, 65:221–231, 2008.

[9] E Essen-Möller. Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen. *Mitteilungen der Anthropologische Gesellschaft (Wien)*, 68:9–53, 1938.

[10] W. Parson and H. J. Bandelt. Extended guidelines for mtDNA typing of population data in forensic science. *Forensic Sci Int: Genetics*, 1:13–19, 2007.

[11] M Krawczac. Kinship testing with X-chromosomal markers: Mathematical and statistical issues. *Forensic Sci Int: Genetics*, 1(2):111–114, 2007.

[12] S. Willuweit and L. Roewer. Y chromosome haplotype reference database (YHRD): Update. *Forensic Sci Int: Genetics*, 1(83-87), 2007.

[13] I W Evett and B S Weir. *Interpreting DNA Evidence*. Sinauer, Sunderland MA, 1998.

[14] D.J. Balding. *Weight-of-Evidence for Forensic DNA Profiles*. Wiley, 2005.

[15] J Ott. *Analysis of Human Linkage*. The Johns Hopkins University Press, Baltimore 3rd. ed., 1999.

[16] M Dawn Teare and JH Barrett. Genetic linkage studies. *The Lancet*, 366:1036–1044, 2005.

[17] E S Lander and P Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 84:2263–2267, 1987.

[18] E Sobel and K Lange. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, 58:1323–1337, 1996.

[19] E M Wijsman, J H Rothstein, and E A Thompson. Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov Chain Monte Carlo provides practical approaches for genome scans on general pedigrees. *American Journal of Human Genetics*, 79:846–858, 2006.

[20] G. R. Abecasis and J. E. Wigginton. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet*, 77(5):754–67, 2005.

[21] C. Buckleton, J. Triggs and S. Walsh, editors. *Forensic DNA Evidence Interpretation*. CRC Press, Florida, USA, 2005.

[22] J. Buckleton and C. Triggs. The effect of linkage on the calculation of dna match probabilities for siblings and half siblings. *Forensic Science International*, 160:193–199, 2006.

[23] K.L. Ayres. Relatedness testing in subdivided population. *Forensic Science International*, 114:107–115, 2000.

[24] L R Mayor and D J Balding. Discrimination of half-siblings when maternal genotypes are known. *Forensic Science International*, 159:141–147, 2006.

[25] K.P. Donnelly. The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology*, 23:34–63, 1983.

Figure 1 Three pedigrees are shown. Data is available from individuals A and B and the task is to determine the most likely pedigree.

Figure 2 The probability that two individuals are IBD at each of two loci is shown for the pedigrees of Figure 1.

Figure 3 Posterior probabilities as functions of the recombination fraction, r, for the three hypotheses of Equation (3) based on 400 sets of simulated data.

Figure 4 Likelihood ratios as functions of the recombination fraction, r, for the three hypotheses of Equation (3) based on 400 sets of simulated data.

Table 1: Probabilities for ordered autosomal genotyped genotypes, X, as a function of the number of alleles shared IBD, indicated by I. For instance, when the individuals are (a, a) and (a, b), it is possible that $I = 0$ or $I = 1$ and the probabilities are shown as functions of the allele frequencies.

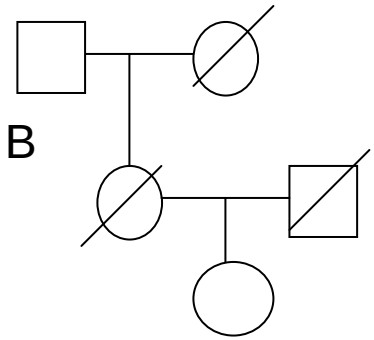
	$P(X\|I)$ for		
Genotype X	$I = 0$	$I = 1$	$I = 2$
(aa, aa)	p_a^4	p_a^3	p_a^2
(aa, ab)	$2p_a^3 p_b$	$p_a^2 p_b$	0
(aa, bb)	$p_a^2 p_b^2$	0	0
(aa, bc)	$2p_a^2 p_b p_c$	0	0
(ab, ab)	$4p_a^2 p_b^2$	$p_a p_b (p_a + p_b)$	$2p_a p_b$
(ab, ac)	$4p_a^2 p_b p_c$	$p_a p_b p_c$	0
(ab, cd)	$4p_a p_b p_c p_d$	0	0

Table 2: Posteriors probabilities are shown. The first column shows the markers used and the second the relation from which data have simulated. The grandparent-grandchild relation is abbreviated grandpar. Observe that it is hard to distinguish between half-sibs and uncle-niece relationships and that only the case with 3820 markers produces useful results.

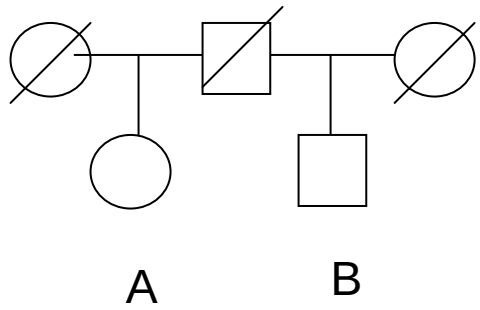
Markers	True R	grandpar	half-sibs	uncle-niece	sibs
20 unlinked markers	grandpar	0.302	0.302	0.302	0.093
1 chr; 20 markers	grandpar	0.236	0.224	0.122	0.419
5 chr; 100 markers	grandpar	0.455	0.288	0.249	0.008
20 chr; 400 markers	grandpar	0.692	0.197	0.112	0.000
20 chr; 3820 markers	grandpar	0.976	0.023	0.001	0.000
20 unlinked markers	half-sibs	0.304	0.304	0.304	0.089
1 chr; 20 markers	half-sibs	0.232	0.229	0.127	0.412
5 chr; 100 markers	half-sibs	0.287	0.349	0.354	0.009
20 chr; 400 markers	half-sibs	0.204	0.406	0.389	0.000
20 chr; 3820 markers	half-sibs	0.031	0.646	0.323	0.000
20 unlinked markers	uncle-niece	0.296	0.296	0.296	0.112
1 chr; 20 markers	uncle-niece	0.234	0.236	0.121	0.408
5 chr; 100 markers	uncle-niece	0.245	0.353	0.386	0.016
20 chr; 400 markers	uncle-niece	0.120	0.400	0.480	0.000
20 chr; 3820 markers	uncle-niece	0.000	0.321	0.678	0.000
20 unlinked markers	sibs	0.095	0.095	0.095	0.715
1 chr; 20 markers	sibs	0.070	0.066	0.526	0.338
5 chr; 100 markers	sibs	0.013	0.015	0.014	0.958
20 chr; 400 markers	sibs	0.000	0.000	0.000	1.000
20 chr; 3820 markers	sibs	0.000	0.000	0.000	1.000

Table 3: Classification rates are shown. The first column shows the markers used and the second the relation from which data have simulated. For instance for '20 chr; 3820 markers', i.e., 3820 markers distributed with 1cM distance on 20 chromosomes, the probability of correctly classifying a grandparent-grandchild (abbreviated grandpar) relation is 0.988.

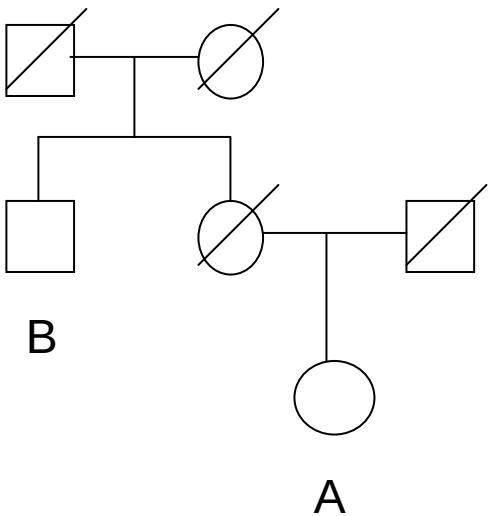
Markers	True R	grandpar	half-sibs	uncle-niece	sibs
20 unlinked markers	grandpar	0.895	0.000	0.000	0.105
1 chr; 20 markers	grandpar	0.320	0.110	0.093	0.478
5 chr; 100 markers	grandpar	0.643	0.198	0.158	0.003
20 chr; 400 markers	grandpar	0.838	0.148	0.015	0.000
20 chr; 3820 markers	grandpar	0.988	0.013	0.000	0.000
20 unlinked markers	half-sibs	0.883	0.000	0.000	0.118
1 chr; 20 markers	half-sibs	0.248	0.178	0.113	0.463
5 chr; 100 markers	half-sibs	0.315	0.238	0.443	0.005
20 chr; 400 markers	half-sibs	0.200	0.408	0.393	0.000
20 chr; 3820 markers	hal-sibs	0.013	0.760	0.228	0.000
20 unlinked markers	uncle-niece	0.858	0.000	0.000	0.143
1 chr; 20 markers	uncle-niece	0.190	0.260	0.095	0.455
5 chr; 100 markers	uncle-niece	0.248	0.188	0.555	0.010
20 chr; 400 markers	uncle-niece	0.083	0.325	0.593	0.000
20 chr; 3820 markers	uncle-niece	0.000	0.248	0.753	0.000
20 unlinked markers	sibs	0.133	0.000	0.000	0.868
1 chr; 20 markers	sibs	0.048	0.018	0.565	0.370
5 chr; 100 markers	sibs	0.008	0.008	0.010	0.975
20 chr; 400 markers	sibs	0.000	0.000	0.000	1.000
20 chr; 3820 markers	sibs	0.000	0.000	0.000	1.000



Pedigree 1: grandparent-grandchild



Pedigree 2: half sibs



Pedigree 3: uncle-niece

Figure 1

Figure 2. Revised Jan 22, 08

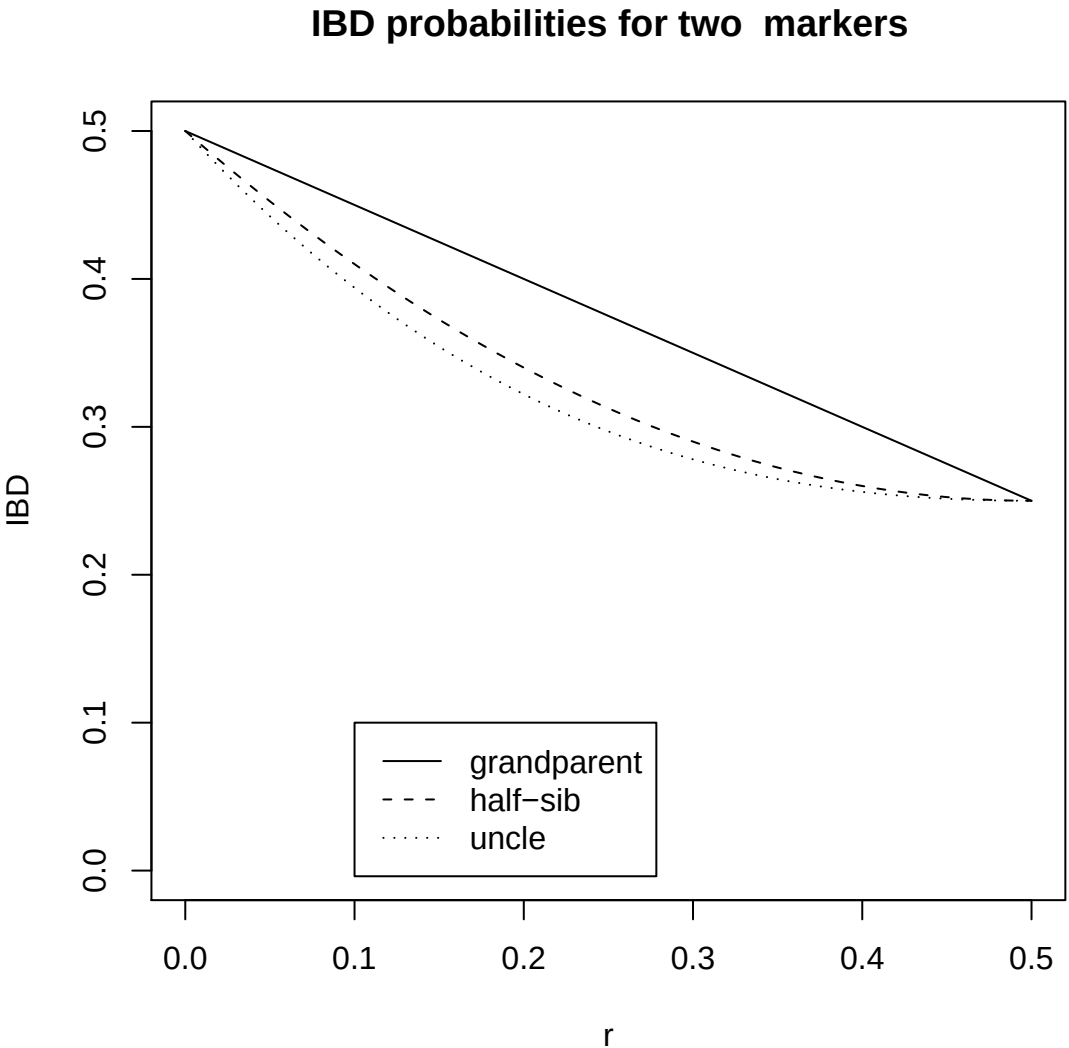


Figure 3. Revised Jan 22, 08

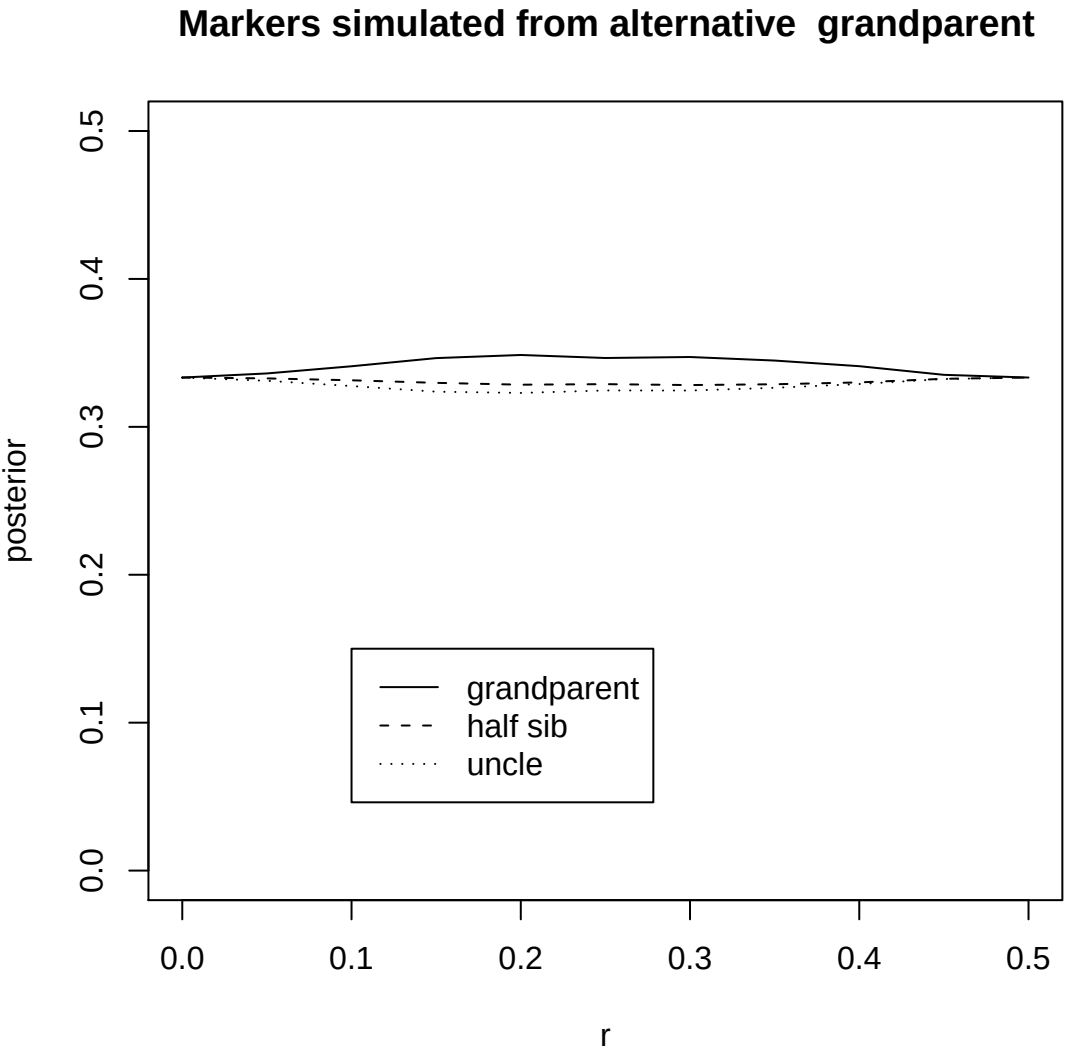


Figure 4. Revised Jan 22, 08