Adjusting for Founder Relatedness in a Linkage Analysis Using Prior Information

N.A. Sheehan\textsuperscript{a}  T. Egeland\textsuperscript{b}

\textsuperscript{a}Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, UK;
\textsuperscript{b}Department of Medical Genetics, Ulleval University Hospital, Oslo, Norway

1 Introduction

The issue of relatedness in a given set of individuals is of central importance in many situations. In animal and plant wildlife and population management settings, knowledge of relatedness of the individuals under study is required to assess patterns of inbreeding and mate choice, or pollen and seed dispersion, and is vital for devising conservation programs for endangered species or breeding schemes for future stock \cite{1–5}. Relationship questions are also fundamental to forensic science applications, including paternity, inheritance, immigration and DNA profile matching investigations. In human genetics, family relationships are required for disease mapping methods that use affected relatives. Moreover, identification of sets of more distantly related individuals from a large population study can be important for homozygosity mapping as they are more likely to share longer haplotype blocks around susceptibility genes with relatively small effects on common complex diseases \cite{6–8}.

The inferential problem of estimating relationships from observed genetic marker data can be one of two main types: either the data are used to distinguish between specific alternative relationships or they are used to estimate the unknown relationship \cite{9}. The former approach is standard for the classic paternity testing scenario in a forensic application. It is also relevant when checking for pedigree errors or lack of relatedness prior to conducting a linkage or association study. The latter is
more difficult due to the probabilistic nature of genetic inheritance whereby unrelated individuals can have genetic information that indicates that they are related, while the observed genetic data may give higher support to an incorrect relationship than the true one.

All relationship estimation problems can be described in terms of reconstructing the pedigree connecting the individuals of interest. In principle, this simply involves consideration of all possible alternatives and finding the one with the highest likelihood [10]. While it can be shown that this is possible for small problems, it is not generally practical unless the set of possible alternatives can be considerably reduced [11, 12]. Sequential reconstruction methods which build a single structure starting with the assumption that all individuals are unrelated and gradually accepting sibships based on the associated increase in log-likelihood, will not necessarily assign individuals to their true relationships [13, 14].

Although it has been known for a long time that the results of a genetic analysis can be extremely sensitive to misspecification of the relationships amongst the individuals under study [15], an increase in false positive findings in a linkage study can also arise when supposed founders are mistakenly assumed to be unrelated [16, 17]. Indeed, as demonstrated on a complex Hutterite pedigree, genetic association studies which assume that all individuals are unrelated can also yield dramatic increases in false positive signals if this assumption is violated [18]. Consanguineous populations are important for homozygosity mapping of rare disease susceptibility loci so the standard linkage statistics based on the assumption of outbreeding populations must be appropriately adjusted. Leutenegger et al. [19] adapted the Maximum Lod Score statistic for affected sib-pair analyses by modifying the expected IBD sharing vector to reflect known parental relatedness as parameterised by the kinship coefficient and both parental inbreeding coefficients. They also noted that the accepted threshold value of 3:36 corresponding to an overall genome-wide significance level of 5% is not appropriate for related parents and that the correct threshold depends on this relationship. Abney et al. [20] present methods for homozygosity and association mapping in large complex pedigrees with known genealogical structure. Hössjer [21] describes a hidden Markov model for the IBD configuration of a family with possible inbreeding among founders. Although the focus of the paper is on nonparametric linkage analysis for affected sib-pair designs, the method is purportedly valid for arbitrary family structures. It is, however, sensitive to the choice of input values for the kinship coefficient and expected length of chromosome segments shared IBD. Ideally these should be estimated, perhaps along the lines suggested by Leutenegger et al. [22] for estimating an individual’s inbreeding coefficient, but genotype data on the founders would be required.

In this paper, we also consider the case where we suspect that the founders of a pedigree are related but do not know the precise nature of this relationship. However, we will assume that there is sufficient prior information, both on global features of the overall population from which these founders derive and on local features pertaining to individuals or specific pairwise relationships, that enables us to restrict the set of possible alternative extended pedigree structures to a manageable size. We are not interested in using the marker data to try to infer the true extended pedigree as we argue that this background relatedness is more of a nuisance factor in these applications. Rather, we propose a general method that allows for such background relatedness to be integrated out of the linkage analysis. Specifically, posterior probabilities for the alternative pedigrees in the sample space are calculated from the marker data that appears to be unlinked to the disease and then used to adjust the lod scores based on the full data. We will begin by discussing the method [described in detail in Sheehan & Egeland, 12] as it applies to this particular context. We will then demonstrate the potential usefulness of the approach by assessing its performance in a realistic simulated scenario.

2 The Method and the Context

In many practical situations, such as in a genetic counselling environment, families may be included in a linkage study because they appear to have an unusual number of cases of a rare autosomal disease. For a recessive disease due to a single mutation with prevalence $p$ at a particular locus, unaffected carriers are assumed homozygous with one copy of the disease predisposing allele. Ignoring new mutations as a possible explanation, both parents of an affected individual have to be carriers. For the diseases we have in mind, the mutated allele will be extremely rare and so carriers are far more likely to be related than not [23]. For instance, assuming Hardy-Weinberg equilibrium, the probability that they are carriers when they are first cousins is 7 times the probability that they are unrelated carriers when $p = 0.01$ but is 636 times that probability when $p = 0.0001$. Failure to account for such relatedness may lead to false positive linkage signals as the expected levels of allele sharing will be misleadingly low and the differences between expected and observed sharing hence more marked than they are in reality [9].

It would be extremely unusual to have no additional information whatsoever on what the possible founder relationships could...
be in these genetic counselling situations. A formal structured approach to incorporating such prior information is hence useful and the problem may be sufficiently tractable to formulate in the four-stage framework of Sheehan and Egeland [12]: define a sample space of (extended) pedigrees, assign a prior probability distribution on this space, calculate the likelihood of the genetic data for each pedigree and compute posterior probabilities using Bayes’ Theorem. These posterior probabilities can be used to infer the true relationship [11] but, as we will show in Section 3 below, they can also be used to account for the presence of background relatedness as a weighting on the pedigree-specific likelihoods for linkage calculations.

In order to consider relationships between the founders of a pedigree, extra individuals for whom genetic data are probably not available must be included. In principle, there could be an infinite number of these. Reducing the set of alternative pedigree structures is usually the difficult step in this approach but, in this situation, the information provided by the families concerned should provide considerable restrictions. For each pedigree $g$ in the sample space, we assign a prior probability of the form:

$$Pr(g) = c \prod_{i} M_{i}^{l_{i}(g)} \prod_{j \neq k} R_{jk}^{o_{jk}(g)}$$ \hspace{1cm} (1)

where

$$c = \frac{1}{\sum_{g} \prod_{i} M_{i}^{l_{i}(g)} \prod_{j \neq k} R_{jk}^{o_{jk}(g)}}$$

is the normalisation constant and $n$ is the number of individuals in the pedigree. $M_{i}, ..., M_{s}$ are non-negative parameters that allow pedigrees to be weighted according to $s$ specified global characteristics which might include demographic features such as cultural prejudices and marriage laws. Each integer exponent $l_{i}(g)$ is a particular measure of the relevant characteristic, is internal to pedigree $g$ and provides the degree of the relative weightings of different pedigrees for the $i$-th characteristic. For example, if the $i$-th characteristic is inbreeding, we could define a pedigree to be inbred if it contains marriages between individuals whose parents are related via other members of the pedigree. In this case, $l_{i}(g)$ could be the number of such marriages (with $l_{i}(g) = 0$ for non-inbred structures) and is hence one measure of the extent of inbreeding. The $R_{jk}$ are the local parameters and relate to specific parts of the pedigree or to particular individuals. As given here, they allow prior information on parent-offspring links to be incorporated. The corresponding pedigree-specific exponents are defined as $o_{jk}(g) = 1$ if $j$ is the parent of $k$ in pedigree $g$ and $o_{jk}(g) = 0$ if $j$ is not the parent of $k$. (Clearly, $o_{jk}(g) + o_{kj}(g) = 1$ for $j \neq k$ in any pedigree $g$). Defining $0^{0} = 1$ means that a value of 0 for any global parameter $M_{i}$ will eliminate all pedigrees for which $l_{i}(g) = 1$ (e.g. all inbred pedigrees in the sample above). Likewise, $R_{jk} = 0$ rules out the pedigrees featuring $j$ as a parent of $k$, i.e. those structures $g$ for which $o_{jk}(g) = 1$. Setting all local and global parameters to 1 amounts to assigning a flat prior whereby all generated pedigrees have equal probability a priori and there is no penalty associated with any particular feature. A value between 0 and 1 decreases the prior probability of the relevant characteristic while a value exceeding 1 increases it. Thus, $R_{jk} > 1$ would favour all pedigrees featuring $j$ as a parent of $k$. For a sample space of $k$ pedigrees $g_{1}, ..., g_{k}$ connecting the individuals of interest, and given the genetic data, the likelihood for any pedigree is $L_{i} = P(data | g_{i})$. By Bayes’ Theorem, the posterior probability of the $i$-th pedigree is thus

$$P(g_{i} | data) = \frac{L_{i} Pr(g_{i})}{\sum_{i=1}^{n} L_{i} Pr(g_{i})}$$

and the posterior distribution can be used to infer the true pedigree.

The prior probability function is generally not as cumbersome in practice as is perhaps suggested by (1) since many parameters will be set to unity. Despite interpretability issues with regard to the choice of values for the $M$ and $R$ parameters, precise definition of the global characteristics of interest and its multiplicative form, it is interesting to note that many existing approaches to incorporating essential non-genetic information can be expressed as special cases of this prior and are less transparent (see Sheehan and Egeland [12] for further discussion).

3 An Example Linkage Scenario

Consider a linkage study with a nuclear family comprising a couple 1 and 2 and their seven children, four of whom have been affected with a rare lethal disease. By definition, the founding members, 1 and 2, are assumed to be unrelated as depicted in figure 1(a). In fact, there is some uncertainty about this and there is a suggestion that 1 and 2 might be related via the father of 1 and the mother of 2. All relationships via other parental combinations have been ruled out. In particular, given the rarity of the disease, together with the information that they come from a community where cousin marriages are quite common, the choice of relationships can be narrowed down considerably. Suppose that, in this case, the family members feel that the most likely possibilities are that they are either half-first cousins or full first cousins and thus have one or two grandparents in common. By use of what Sheehan and Egeland [12] refer to as ‘hard’ prior information, the sample space of possible alternative pedigrees for this particular genetic counselling application can be restricted to the four shown in figure 1. Note that 4 unobserved males and 4 unobserved females have to be included with the original 9 typed family members to describe these four possibilities. However, as most of the pairwise relationships are fixed (1 and 2 are the parents of all seven children, 3 is the father of 1 and 4 is the mother of 2), we only need to consider the pairwise relationships between a possible common grandfather 5 and grandmother 6 with each of 3 and 4. Note that this is sufficient to cover all orderings of the four relevant pedigrees by the prior of (1). In the absence of data on these extra individuals, pedigrees figure 1(b) and (c) will be indistin-
guishable on the basis of likelihoods and lod scores unless mtDNA or Y-chromosome data are incorporated. Their adjusted lod scores may differ, however, if prior information indicates that one alternative is more likely than the other.

The local part of the prior model in Equation 1 is hence formulated by the four parameters $R_{5,3}$, $R_{5,4}$, $R_{6,3}$ and $R_{6,4}$. The first cousin alternative figure 1(d) is most likely a priori when these are all greater than unity and the strength of such prior belief is reflected by the magnitude of the parameter values. The half-cousin option with common grandfather figure 1(b) will be favoured when $R_{5,3}$ and $R_{5,4}$ are assigned values greater than 1 while $R_{6,3}$ and $R_{6,4}$ are both given values less than 1. Setting all four parameters to 1 will make all four possibilities equally likely a priori. For simplicity, a flat prior over all global

| Pedigree | Prior probability distribution |
|----------|--------------------------------|
| 1 (a)    | $c$                           |
| 2 (b)    | $cR_{5,3}R_{5,4}M_P$          |
| 3 (c)    | $cR_{6,3}R_{6,4}M_P$          |
| 4 (d)    | $cR_{5,3}R_{5,4}R_{6,3}R_{6,4}$ |

where $c = 1 + R_{5,3}R_{5,4}M_P + R_{6,3}R_{6,4}M_P + R_{5,3}R_{5,4}R_{6,3}R_{6,4}$

parameters could be assumed but, for illustration, we will also include the ‘promiscuity’ parameter $M_P$ of Egeland et al. [11] as it might be the case that multiple marriages are socially discouraged. Here, the degree of promiscuity in a pedigree is defined in terms of $b_P(g)$, the number of pairs of offspring in pedigree $g$ with exactly one common parent (i.e. half-siblings) who is also in $g$. Thus, $M_P = 0$ excludes the two half-first cousin pedigrees while $M_P > 1$ favours them but, as noted in Sheehan and Egeland [12], different definitions of this parameter are possible. The corresponding pedigree-specific index, $b_P(g)$, is simply a count of the number of pairs of half-siblings in pedigree $g$. The prior probabilities for the four pedigrees in figure 1 are given in table 1.

### 4 A Simulation Study

100 sets of microsatellite marker data were generated on pedigree figure 1(b), i.e. 1 and 2 are half-first cousins via a common grandfather, using Allegro [24]. Recall (Section 3) that individuals 1 and 2 and their seven children have marker data. All additional individuals are unobserved. Each data set comprised 10 markers on each of 10 chromosomes, labelled 1, ..., 100 and positioned 10 cM apart which is an acceptable spacing interval to avoid issues arising from possible linkage disequilibrium. All markers were simulated with 5 equifrequent alleles. We considered an autosomal recessive disease allele with frequency 0.0001 and penetrance probabilities of 0.98 for homozygous carriers and 0 for both non-carriers and heterozygous carriers. The disease locus was set at 49 cM on chromosome 1, i.e. 1 cM to the left of marker 5. Three different specifications of the prior function (1) were used: (1) a flat prior where all local and global parameters are set to 1: $M_P = R_{5,3} = R_{5,4} = R_{6,3} = R_{6,4} = 1$; (2) a prior that downweights multiple marriages, and makes the half-sib options half as likely a priori as the alternatives but doesn’t
favour any parent-offspring relationships: $M_P = 0.5, R_{5,3} = R_{5,4} = R_{6,3} = R_{6,4} = 1$, and (3) a prior that favours the common grandfather structure: $M_P = 1, R_{5,3} = R_{5,4} = 10, R_{6,3} = R_{6,4} = 1$. This parametrisation implies that pedigree figure 1(b) is 100 times more likely than figure 1(c) a priori and also a posteriori in this case, since the likelihoods coincide.

4.1 An Adjusted Lod Score Analysis

A multipoint linkage analysis was carried out for each of the four pedigrees in figure 1 using Merlin [25]. The rest of the analysis in this section was done in R (http://www.r-project.org). It was assumed that marker allele frequencies, disease allele frequency, disease penetrance and marker genetic map were known and equal to those used in the simulations. We used Haldane’s map function.

In this section we must weight the likelihood components of the corresponding ratio. We could do this naively by using our prior probabilities for the pedigrees in the sample space or, for that matter, by using any set of probabilities that have been otherwise ascertained as reasonable. We have chosen to update our prior knowledge with the likelihoods based on the unlinked data (i.e. with chromosome 1 excluded) and will use these pedigree posterior probabilities, $P(g_i \mid data)$ where $i = 1, ..., k$ and $k = 4$ here, as the weighting factors. Specifically, the likelihoods for marker $i$ are

$$L_{Ai} = L_{Ai,1} P(g_1 \mid data) + \ldots + L_{Ai,k} P(g_k \mid data),$$
$$L_0 = L_{0,1} P(g_1 \mid data) + \ldots + L_{0,k} P(g_k \mid data)$$

and the ratio

$$lod_i = \log_{10}(L_{Ai}/L_0)$$

is the pedigree-adjusted lod score.

Note that an expected lod score could be calculated for each marker along the lines suggested by Ott [15]:

$$elod_i = \log_{10} \left( \frac{L_{Ai,1}}{L_{0,1}} P(g_1 \mid data) + \ldots + \frac{L_{Ai,k}}{L_{0,k}} P(g_k \mid data) \right).$$

This is similar to Ott’s ELOD but here the expectation is taken over the posterior pedigree distribution and not over the distribution of the number of recombinations. We should stress, however, that the adjusted lod score (2) is a proper likelihood ratio statistic and hence has all the required asymptotic properties for likelihood ratio hypothesis testing. The expected lod score (3) does not have such a rigorous theoretical justification. On these grounds, we will prefer to use adjusted lod scores rather than expected lod scores.

4.2 Results

Figure 2 shows the posterior probability distributions (over the 100 simulated datasets) calculated from the unlinked data on chromosomes 2, ..., 10, corresponding to each of the priors described in Section 4 for the four pedigrees in the sample space. It is clear from the plots corresponding to the flat prior, where no particular features have been favoured or downweighted, that there is enough information in the unlinked marker data to indicate that a common grandparent (i.e. pedigree fig. 2(b) or (c)) is generally the most likely relationship between the founding individuals 1 and 2. This relationship is deliberately downweighted by the second prior making the first cousin relationship of pedigree figure 2(d) more likely than the others. In this case, the evidence provided by the marker data largely counteracts the downweighting effect but a more extreme prior would induce more marked differences. Marker data alone will never distinguish between the true pedigree figure 2(b) and the common grandmother option of pedigree figure 2(c): strong prior information, such as is provided by the third prior, is required to make such a distinction.

Mean posterior probabilities for each pedigree and each prior are given in table 2. For the first and third prior, the true relationship (b) between 1 and 2 has the largest posterior probability on average. As can be seen from figure 2, the posterior distributions are actually quite skewed and taking medians rather than means would accentuate this trend. For the second prior where the true relationship has been deliberately downweighted, the first cousin relationship (d) is now the most likely option a posteriori.

Figure 3 shows the distribution (over 100 datasets) of multipoint lod scores at marker 5, the marker closest to the disease locus at 49 cM. The first four columns are the
raw lod scores calculated for each of the four pedigrees in the sample space. The last three columns depict the distributions of adjusted lod scores at this marker position corresponding to each prior with the likelihoods weighted by the posterior probabilities of figure 2, as given in (2). Lod scores around the disease location of 49 cM on chromosome 1 tend to be higher for the half-first cousin relationships (b) and (c) than for the unrelated (a) and first-cousin (d) alternatives, as would be anticipated. However, even when the posterior probabilities were calculated from a flat prior which does not favour any particular structure a priori, the adjusted lod scores all tended to be much higher than those for (a) and slightly higher than those for (d). Average adjusted lod scores in this area reflected this trend (table 3).

As is consistent with findings reporting increased false linkage signals [17, 18], multipoint lod scores generally seemed to be higher at locations away from the disease locus when the founder relationships were misspecified. Some indication of this can be seen in figure 4 showing the relevant distributions over the 100 datasets for a randomly sampled location away from chromosome 1. Although we did not look for a particular location that showed clear false linkage signals, lod scores for the unrelated case (a) are generally higher than those that take account of relatedness, either when a particular relationship is assumed, or when the lod scores are adjusted for several alternatives. There is not a marked difference between the distributions for first and half-first cousin relationships in figure 4 and the adjusted lod scores all have similar distributions and are close to the true relationship. Similar patterns were observed for other sampled locations.

4.3 Sensitivity Issues

We have presented a general method for adjusting lod scores in a linkage analysis to reflect uncertainty about founder relatedness for applications where prior infor-
Table 3. Mean lod scores at marker 5 for the four pedigrees of figure 1 and adjusted lod scores corresponding to each of the three priors of Section 4

| Pedigree | Prior 1 | Prior 2 | Prior 3 |
|----------|---------|---------|---------|
| (a)      | 1.21    | 2.22    | 2.00    |
| (b)      | 2.00    | 2.22    | 2.19    |
| (c)      | 2.19    | 2.17    |         |

Fig. 3. Multipoint lod score distributions over 100 simulated datasets for the four pedigrees of figure 1 at marker 5 together with the adjusted lod score distributions corresponding to each prior.

Fig. 4. Distribution of 100 multipoint lod scores for an arbitrarily chosen location away from chromosome 1 for each of the pedigrees in figure 1 with adjusted lod scores corresponding to each of the three prior parameterisations.

Table 4. Mean posterior probabilities (based on unlinked data) from 100 simulations for each pedigree of figure 1 and each prior in Section 4 when the marker allele frequencies are estimated from the data

| Pedigree | Posterior 1 | Posterior 2 | Posterior 3 |
|----------|-------------|-------------|-------------|
| (a)      | 0.1949      | 0.2505      | 0.0093      |
| (b)      | 0.3023      | 0.2302      | 0.6604      |
| (c)      | 0.3023      | 0.2302      | 0.0066      |
| (d)      | 0.2005      | 0.2890      | 0.3237      |
mation on the nature of such relationships is typically available. Our simulation study illustrates that the approach is potentially useful. However, sensitivity of the approach to choice of marker polymorphism and allele frequencies, disease location and genetic model requires investigation. We present some preliminary discussion on some of these issues here but note that a full sensitivity analysis is beyond the scope of this paper. We note that choice of sample space and local and global prior parameters in the prior specification is also important but these will be particular to the application at hand and cannot be discussed satisfactorily in general terms.

In order to consider the effects of allele frequency mispecification, we reanalysed our 100 sets of simulated microsatellite data by estimating the allele frequencies from the data in Merlin. The same three priors were used and the prior probability distributions over the pedigree sample space, together with the mean posterior probabilities when frequencies were assumed to be known and equal to those used in the simulations, are given in Table 2.

Comparing the original results with those for estimated allele frequencies in Table 4, it can be seen that the two sets of posterior distributions are very similar. The mean adjusted lod scores at the marker closest to the disease locus are identical for pedigree (a) for known and estimated frequencies since the founders are typed, and are very similar for the other three. In this particular example, the effects of estimating allele frequencies are practically negligible. However, this will not generally be the case.

The approach we have presented in this paper is not restricted to microsatellite markers. Diallelic SNP data can also be used. To illustrate this, we also simulated 100 data sets of 100 SNP markers with equifrequent alleles on 10 chromosomes at 10 cM intervals on pedigree (b). These were analysed in Merlin, as described above, when allele frequencies were known (and equal to 0.5), and when they were estimated from the data. Now the posterior probability distributions based on the unlinked data on chromosomes 2, ..., 10 were not informative enough to place more weight on pedigrees (b) and (c). Lod scores calculated when allele frequencies were estimated (not shown) were again practically identical to those calculated for known frequencies. Table 5 shows the mean adjusted lod scores at the marker closest to the disease locus, for both SNP and microsatellite data. Expected lod scores, calculated as described in Equation (3), are also shown in Table 5 for comparison although, for the reasons given in Section 4.1, we would advise that adjusted, rather than expected, lod scores be reported. Not surprisingly, the lod scores based on SNPs are lower than those for microsatellites as 100 microsatellites are more informative than 100 SNPs. However, with huge numbers of SNP data becoming more readily available, this lack of informativeness is not really an issue. The SNP mean adjusted lod scores, however, are also all higher than the mean lod score calculated under the assumption that pedigree (a) was the true relationship i.e. when 1 and 2 were assumed to be unrelated. The same pattern is observed for the expected lod scores.

How does this approach perform when the true relationship is not actually one of those listed in the sample space? For the example presented here, suppose that individuals 1 and 2 could be second cousins. We simulated microsatellite data on 100 datasets exactly as before for the second cousin pedigree of Figure 5 and then analysed them as before using the original sample space of Figure 1. Since the strength of the second cousin relationship is between the unrelated and half first cousin options, it is perhaps not surprising (Table 6) that the posterior probability distribution based on the unlinked data for the flat prior assigns more weight to pedigrees (b) and (c). The mean adjusted lod scores at marker 5 (Table 7) are all higher than the mean lod score for the unrelated case (a) and so the method appears to work reasonably well even when the true relationship is not among those considered. However, we should note that the true relationship was still reasonably close to those of the sample space and further investigations on more distant possibilities are required.

| Data         | (a) | (b) | (c) | (d) | Prior 1 | Prior 2 | Prior 3 |
|--------------|-----|-----|-----|-----|---------|---------|---------|
| Microsatellite | 1.21 | 2.22 | 2.22 | 2.00 | 2.22 (1.99) | 2.19 (1.91) | 2.17 (2.14) |
| SNP          | 1.13 | 1.73 | 1.73 | 1.64 | 1.69 (1.56) | 1.66 (1.51) | 1.69 (1.69) |
5 Discussion

This paper presents an approach to adjusting the lod scores of a linkage analysis when pedigree founders are believed to be related. We exploit our previous method [12] for formal incorporation of prior information in relationship estimation problems. Importantly, we rely on being able to restrict the set of alternative extended pedigrees to a reasonable size and so consider applications where such information is often available. The focus is quite different, however. Here we are not interested in estimating the founder relationship as we believe this to be a nuisance factor for these applications. Instead, we update our prior probabilities with genetic data believed to be unlinked to the disease locus and use these posterior probabilities to weight the relevant likelihoods for the marker lod scores. The prior probabilities could be heavily influential when the data are not as informative as they were for our simulations and will always be required to separate pedigrees that have the same degree of relatedness. The choice of parameters for the prior distribution depends heavily on the particular application and general recommendations are not really sensible. However any reasonable set of probabilities on pedigrees, based possibly on expert knowledge, could in principle be used. We also note that adjusted lod scores should be calculated in preference to expected lod scores. Lod scores can be parametric (as presented here) or non-parametric.

The simulation study was designed to test the potential usefulness of the method and we used suitably polymorphic loci. There is no reason why SNP marker data cannot be used but the density of SNP marker maps may necessitate consideration of linkage disequilibrium between the markers since ignoring it will lead to severe biases in lod score calculations [26]. Linkage disequilibrium also depends on the population being studied. Our approach will benefit from extensive sensitivity checking to marker polymorphism and allele frequencies, disease location and genetic model as we have barely touched on these issues here. It made sense to ignore the data from markers that we suspected were linked to the disease (i.e. those on chromosome 1 in our simulations) for our posterior probability calculations. Alternatively, given the number of markers one would expect to have in practice, one could simply ignore markers with lod scores exceeding a fairly conservative threshold (e.g. 1) and their immediate neighbours.

As with any statistical analysis, violation of the underlying assumptions could lead to the wrong answers. If the

![Pedigree Diagram](attachment:pedigree.png)

**Fig. 5.** The pedigree representing the situation where 1 and 2 are second cousins via their first cousin parents 3 and 4.

| Pedigree | Posterior 1 | Posterior 2 | Posterior 3 |
|----------|-------------|-------------|-------------|
| (a)      | 0.1912      | 0.2554      | 0.0076      |
| (b)      | 0.3234      | 0.2493      | 0.7009      |
| (c)      | 0.3234      | 0.2493      | 0.0070      |
| (d)      | 0.1621      | 0.2461      | 0.2845      |

**Table 6.** Mean posterior probabilities (based on unlinked data) from 100 simulations for each pedigree of figure 1 and each prior in Section 4 when the data were simulated on the second cousin pedigree of figure 5.

| Pedigree | prior 1 | prior 2 | prior 3 |
|----------|---------|---------|---------|
| (a)      | 2.14    | 2.12    | 2.12    |
| (b)      | 2.18    | 1.97    | 2.14    |
| (c)      | 2.19    | 2.02    | 2.12    |
| (d)      | 2.03    | 2.03    | 2.12    |

**Table 7.** Mean lod scores at marker 5 for the four pedigrees of figure 1 and adjusted lod scores corresponding to each of the three priors of Section 4 for microsatellite data simulated on the second cousin pedigree of figure 5. Expected lod scores are shown in brackets.
true relationship between the founders is known, it should be used and the linkage calculations modified appropriately [19, 20]. Alternatively, parameters describing the relationship could be estimated [21, 22]. However, it is difficult to estimate quantities like the inbreeding coefficient in clinical practice because there is rarely sufficient information. In human genetics, it is customary to take the closest relationship between the founders (i.e. the shortest loop) as the best estimate and ignore all the more distant possibilities, but this can lead to serious underestimation of homozygosity in consanguineous populations [23]. Our approach is to look at a finite set of alternative possible relationships connecting all the individuals of interest, rather than specific relatedness parameters for the founders. In particular, it does not restrict us to considering pairs of founders: relationships between any number of founders can be specified. It does rely on having sufficient prior information on family histories, as we believe would often be the case for the applications we have in mind, to suitably restrict the set of possible alternatives. All the calculations we have presented can be done with existing freely available software.

It would seem plausible that the degree of relatedness or precise relationship amongst the founders in a linkage analysis is not quite as important as the realisation that they are not unrelated. In our simple experiment with the second cousin example, we did not suffer too badly from not having the true relationship in the sample space and our adjusted lod scores were quite insensitive to the priors we used. It would be interesting to see how much of an improvement over the unadjusted analysis can be obtained from a simple modification that allows for ‘general relatedness’, in some sense. Liu et al. [27] consider highly looped inbred pedigree structures which are computationally too complex to handle. Their proposal is to simplify the pedigree but keep the inbreeding values close to the true values thus creating ‘a single hypothetical loop’ for each individual. They demonstrated that this approach yielded a considerable decrease in the number of false positive linkage signals when compared with the standard approach of using only the shortest loops. The authors note that this hypothetical loop will ignore the differences in genome sharing between closely related individuals and those more distantly related, due to the different numbers of recombinations expected. Consequently, the probability of recombination may be underestimated which, in turn will affect the power to detect linkage. Nonetheless, the idea is appealing and could perhaps be adapted to other applications.

Since the precise relationship amongst the founders is not particularly of interest in these applications, sequential approaches to relationship estimation which will quickly provide a good estimate could also be used here. However, even when inbreeding between founders has been accounted for, additional bias may be introduced when combining information from several families in a linkage study. As noted by Leutenegger et al. [19], the sample likelihood is not necessarily the product of the individual family likelihoods when the families all derive from a genetic isolate or a population with a long history of consanguineous marriages. That aside, we have shown that it is not necessary to require founders to be unrelated in linkage applications when that assumption is dubious and when some information on the possible form of their relatedness is available. As more and more marker data become available, estimating relationships will be easier, the effects of the prior on the posterior distribution will be less, and the adjusted lod scores we have presented here will be closer to the ‘true’ lod scores.

Acknowledgment

This work was presented at the symposium ‘Family-based Studies in Genetic Epidemiology’ held at the University of Sheffield, UK, July 31 – August 2, 2006. http://www.fbs.group.shef.ac.uk/index.html.

References

1 Meagher TR, Thompson EA: Analysis of parentage for naturally established seedlings of chamaelirium luteum (liliaceae). Ecology 1987;68:803–812.

2 Bowers J, Bourisiquet J-M, This P, Chu K, Johansson H, Meredith C: Historical genetics: The parentage of Chardonnay, Gamay and other wine grapes of northeastern France. Science 1999;285:1562–1565.

3 Blouin MS: DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evol 2003; 18:503–511.

4 Jones AG, Ardren WR: Methods of parentage analysis in natural populations. Mol Ecol 2003;12:2511–2522.

5 De Woody JA: Molecular approaches to the study of parentage, relatedness, and fitness: Practical applications for wild animals. J Wildlife Management 2005;69:1400–1418.

6 Boehnke M, Cox N: Accurate inference of relationships in sib-pair studies. Am J Hum Genet 1992;61:423–429.
Adjusting for Founder Relatedness in a Linkage Analysis

McPeek MS, Sun L: Statistical test for detection of misspecified relationships by use of genome-screen data. Am J Hum Genet 2000;66:1076–1094.

Stankovich J, Bahlo M, Rubio J, Wilkinson C, Thomson R, Banks A, Ring M, Foote S, Speed T: Identifying nineteenth century links from genotypes. Hum Genet 2005;117:188–199.

Weir BS, Anderson AD, Hepler AB: Genetic relatedness analysis: Modern data and new challenges. Nature Rev Genet 2006;7:771–780.

Cannings C, Thompson EA: Genealogical and Genetic Structure. Cambridge Studies in Mathematical Biology 3, Cambridge University Press, 1981.

Egeland T, Mostad PF, Mevåg B, Stenersen M: Beyond traditional paternity and identification cases. Selecting the most probable pedigree. Forensic Sci Int 2000;110:47–59.

Thompson EA: The estimation of pairwise relationships. Ann Hum Genet 2007;71:501–518.

Thompson EA: Inference of genealogical structure Soc Sci Inf 1976;15:477–526.

Ott J: Analysis of Human Genetic Linkage, ed 3. Baltimore, Johns Hopkins University Press, 1999.

Génin E, Clerget-Darpoux F: Consanguinity and the sib-pair method: An approach using identity by descent between and within individuals. Am J Hum Genet 1996;59:1149–1162.

Miano MG, Jacobson SG, Carothers A, Hanson I, Teague P, Lovell J, Cideciyan AV, Haider N, Stone EM, Sheffield VC, Wright AF: Pitfalls in homozygosity mapping. Am J Hum Genet 2000;67:1348–1351.

Newman DL, Abney M, McPeek MS, Ober C, Cox NJ: The importance of genealogy in determining genetic associations with complex traits. Am J Hum Genet 2001;69:1146–1148.

Leutenegger AL, Génin E, Thompson EA, Clerget-Darpoux F: Impact of parental relationships in maximum lod score affected sib-pair method. Genet Epidemiol 2002;23:413–425.

Abney M, Ober C, McPeek MS: Quantitative-trait homozygosity and association mapping, and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. Am J Hum Genet 2002;70:920–934.

Hössjer O: Modeling the effect of inbreeding among founders in linkage analysis. Theor Popul Biol 2006;70:146–163.