On the Use of Dense SNP Marker Data for the Identification of Distant Relative Pairs *

M Sun¹, M A Jobling², D Taliun⁴,⁵, P P Pramstaller⁴, T Egeland³, N A Sheehan¹,² †

¹ Department of Health Sciences, University of Leicester UK
² Department of Genetics, University of Leicester UK
³ IKBM Norwegian University of Life Sciences, Norway
⁴ Center for Biomedicine, European Academy of Bolzano (EURAC), Bolzano, Italy
⁵ Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA

Abstract

There has been recent interest in the exploitation of readily available dense genome scan marker data for the identification of relatives. However, there are conflicting findings on how informative these data are in practical situations and, in particular, sets of thinned markers are often used with no concrete justification for the chosen spacing. We explore the potential usefulness of dense single nucleotide polymorphism (SNP) arrays for this application with a focus on inferring distant relative pairs. We distinguish between relationship estimation, as defined by a pedigree connecting the two individuals of interest, and estimation of general relatedness as would be provided by a kinship coefficient or a coefficient of relatedness. Since our primary interest is in the former case, we adopt a pedigree likelihood approach. We consider the effect of additional SNPs and data on an additional typed relative, together with choice of that relative, on relationship inference. We also consider the effect of linkage disequilibrium. When overall relatedness, rather than the specific relationship, would suffice,

†Corresponding author. E-mail address: Nuala.Sheehan@leicester.ac.uk
we propose an approximate approach that is easy to implement and appears to compete well with a popular moment-based estimator and a recent maximum likelihood approach based on chromosomal sharing. We conclude that denser marker data are more informative for distant relatives. However, linkage disequilibrium cannot be ignored and will be the main limiting factor for applications to real data.

**Keywords:** Relationship, Relatedness, Pedigree, Likelihood, Bayes’ Theorem

1 Introduction

There are many reasons why one would wish to identify sets of relatives from genetic marker data. These include forensic identification questions, whereby disaster victims are identified from their relatives, for example, and legal issues such as paternity, adoption and immigration cases. In conservation genetics, some knowledge of the relationships amongst existing individuals is crucial to the design of effective breeding strategies for endangered species and is informative for gene flow, estimates of heterozygosity and reproductive success in ecological genetics [28]. In human genetics, family data are essential for segregation and linkage analyses and genetic counselling applications. Moreover, it is often important to be able to verify stated pedigree relationships from the observed data and to check for errors. For example, the results of an association analysis on a case-control study could be seriously biased if some cases and controls were related and linkage analyses are sensitive to undeclared relationships between pedigree founders [11, 23, 51, 56]. More recently, family-based designs have been recommended for association studies concerned with parent-of-origin effects and the control of population stratification [40]. Pedigrees have also been recommended as an efficient design, both for identifying rare disease-risk alleles in association analyses and for initial sequencing of selected affected family members [12, 60].

Sometimes a general measure of overall pairwise relatedness, such as a kinship coefficient or coefficient of relationship, will suffice but in many cases, the precise relationship, or pedigree, is of interest. In the latter situation, choosing between a particular pedigree and several specified alternatives is an easier problem to address than estimating the true unknown relationship. Firstly, there is a potentially enormous number of possible alternatives to consider, even in ‘non-deficient’ cases where all the pedigree members have
been observed. Secondly, the observed genetic sharing amongst relatives may not reflect the expected sharing, due to the probabilistic nature of genetic inheritance [25, 26, 59].

Likelihood-based approaches to pedigree reconstruction tend to be heuristic and typically deliver a high likelihood, but not necessarily optimal, solution [3, 14, 45, 53]. Approaches that find an overall guaranteed maximum likelihood pedigree are computationally intensive and require either an exhaustive or complete search. An exhaustive enumeration of all possible pedigrees is implemented in the programme Familias (http://familias.name/) for a related problem of maximising likelihood ratios for forensic identification cases and is restricted in practice to small numbers (about 12) of individuals [18]. A dynamic programming approach can yield a maximum likelihood solution for pedigrees of up to 30 individuals [13] while recent work exploiting efficient integer linear programming optimisation algorithms to perform a complete search can deliver a guaranteed maximum likelihood pedigree for over a thousand individuals [15, 46]. Both the above make simplifying assumptions about the genetic model and, crucially, require complete data, by which we mean that all pedigree members are present in the genotyped sample.

Here, we will primarily focus on the problem of choosing between a particular pairwise relationship specified by a pedigree and a set of clearly specified alternatives as might be available in forensic identification, animal breeding and genetic counselling applications. The consideration of pairwise relationships immediately takes us out of the complete data situation and so it is not practical to try to find the overall most likely pedigree. It is well known in ‘deficient’ forensic identification applications, for example, that the standard sets of unlinked short tandem repeat (STR) markers are not sufficient to resolve certain relationship testing cases [6, 36]. For example, a first-degree relative of the true father could be compatible with paternity since exclusion at a small number of loci could be explained by mutation. Several authors have suggested adding a carefully selected set of single nucleotide polymorphism (SNP) data to the standard STRs for solving relationship testing problems typically involving close relationships e.g. up to first cousins [7, 43]. SNPs have the advantages of having negligible mutation rates, being very numerous and are more reliably typed on degraded DNA which can yield incomplete STR profiles. Moreover, some problems can never be resolved by any number of unlinked markers [17, 19, 55].

It has recently been proposed that large dense sets of linked markers,
such as are routinely collected in genetic association studies, could have the potential to considerably extend the range of identification problems that can be satisfactorily addressed [49]. However, smaller subsets of SNPs tend to be used in practical applications [5, 27, 42] without much discussion as to why and several authors have considered the problem of selecting an optimal subset size for any given application [29, 32, 44]. Hence, it would seem that there is some lack of agreement on how useful these dense markers really are in practice.

In this paper, we will return to the basic idea proposed in [49] and extend it to further explore the exploitation of dense genome-wide autosomal SNP data for identifying distant relatives where by ‘distant’ we will mean second cousin (or second half-cousin) relationships and beyond. We will thus adopt a pedigree likelihood approach which is straightforward to implement using freely available software and will address the problem as one of choosing the best pedigree from a given set of alternatives [47, 49]. Specifically, we now consider the effects of denser SNP arrays, an additional typed individual, and linkage disequilibrium (LD) on relationship estimation. Furthermore, we propose a new pedigree likelihood approach to assessing general relatedness when the particular relationship (or pedigree) is not of primary interest. We compare our method with a popular moment-based estimator of kinship which is also easy to construct and we compare our results with reported findings from a recently proposed maximum likelihood approach based on chromosomal segment sharing [27]. Our investigations involve simulated data based on real allele frequencies and real data from a complex human pedigree.

The structure of the paper is as follows. We begin by describing our likelihood framework, the family of pedigree structures we consider and the data and software used for the simulations and calculations. We then present our results followed by a discussion of our findings and some general conclusions.

2 Methods

Although we are taking a likelihood approach, we will adopt a Bayesian formulation. In practice, ‘prior’ information on how to weight possible alternative pedigrees may well be available so it should be possible to accommodate this in a formal way [18, 47]. More importantly, a Bayesian approach naturally extends standard likelihood ratios to the consideration of
Figure 1: Examples of simple unlooped extended half-sibling and extended sibling relationships depicting (a) HS-3-3 and (b) S-4-3 pedigrees.

cases with more than two alternative relationships. Thus, for \( n \) competing hypotheses, or pedigrees, \( H_1, ..., H_n \) with associated prior probabilities \( \pi_1, ..., \pi_n \), the posterior probability of \( H_i \) is

\[
P(H_i|\text{data}) = \frac{L_i \pi_i}{\sum_{j=1}^{n} L_j \pi_j}, \tag{1}
\]

where \( L_i \equiv P(\text{data}|H_i) \) is the probability of the data under hypothesis \( H_i \), or the likelihood for \( H_i \). For illustrative purposes, we will assume a ‘flat prior’ reflecting that all pedigrees are equally likely in the absence of any information to the contrary so \( \pi_i = 1/n \) for all \( i \). In this case, the posterior probability (1) of \( H_i \) simplifies to

\[
P(H_i|\text{data}) = \frac{L_i}{\sum_{j=1}^{n} L_j}
\]

and has a meaningful frequentist interpretation as a comparison of the likelihood for the hypothesis under consideration with the sum of the likelihoods for the other possibilities [49]. For pairwise comparisons, the posterior probability ratio equals the conventional likelihood ratio when the prior probabilities for the two alternatives are equal.

We focus initially on the class of simple unlooped pedigree structures considered in [17, 49]. In addition to being useful structures on which to investigate the ability to discriminate between different levels of relationship, we will argue that they can also be considered as general templates for differing degrees of relatedness when the pedigree structure is unknown.
Specifically, two individuals are said to have an extended half-sibling relationship if they share a single common ancestor. We will denote such a relationship as HS-$n_1-n_2$ if there are $n_1$ and $n_2$ meioses separating the first and second individuals from the common ancestor, respectively. Thus, two half-siblings would define a HS-1-1 pedigree and a HS-3-3 relationship, involving a common great-grandparent, is shown in Figure 1(a). Extended sibling, or cousin, relationships where the two individuals of interest share two common ancestors, are defined analogously with an S-1-1 pedigree defining a full sibling relationship. An S-4-3 pedigree depicting second cousins once removed is shown in Figure 1 (b).

All our simulations are based on real allele frequencies and genetic maps. We used allele frequencies for Affymetrix 500K SNP data derived from HapMap and downloaded from http://folk.uio.no/thoree/FEST. After quality control, there were 416,854 SNPs in this set. We also received Affymetrix v6.0 SNP data on 2943 unrelated individuals from the UK National Blood Service cohort from the Wellcome Trust Case Control Consortium (WTCCC) and derived allele frequencies for 711,020 SNPs from this sample. Physical map distances were converted into genetic map distances using the Rutgers Combined Linkage-Physical Map of the Human Genome http://compgen.rutgers.edu/RutgersMap/. Genotype data were simulated using the freely available software packages Merlin [1] and Mendel [31]. Pedigree likelihoods were calculated directly in Merlin using the Lander-Green algorithm [30] with the Haldane map function. The R software (https://www.r-project.org/) was used for data handling and analysis. The program FEST [49], which is an R front-end for Merlin, could have been used for the initial pairwise calculations without LD, but was not used here.

For our real data, we used Illumina 300K SNP data on 1285 individuals from the MICROS study [41]. These individuals are part of a large complex pedigree of about 50,000 individuals connecting three villages in South Tyrol and traced back from 2002 to 1650 via historical records. We consider the 12-generation sub-pedigree comprising the direct ancestors of the 1285 typed individuals. The software Jenti [20] was used to split this 12-generation pedigree into smaller subsets (e.g. 4 and 8-generation pedigrees) enabling the extraction of pairs of individuals with particular relationships. Allele frequencies for 304,383 autosomal SNPs were derived from the genotyped individuals. In the absence of an array-specific linkage map, we used the Illumina 500K linkage map which covers most of the SNPs in the 300K array.
3 Results

For a given pedigree structure connecting two individuals, SNP alleles were assigned independently to founder individuals according to their respective allele frequencies. This random union of gametes ensures Hardy-Weinberg equilibrium among pedigree founder genotypes although we note that this is not essential. Alleles were then segregated through the pedigree from parents to offspring without mutation according to the genetic map provided. Only the simulated marker data for the two individuals of interest were retained for the analyses. Analyses of replicated simulated data for individuals related via the simple unlooped pedigrees of Section 2 confirmed that relative pairs separated by about 6 meioses i.e. up to the level of second or half-second cousins, can be distinguished unambiguously from the a priori equally likely alternative hypothesis of ‘unrelated’ with 500K SNP data. This is consistent with other published findings [16, 26, 27, 29, 49] and the first three lines of Table 1 report posterior probability averages for 6 different HS-\(n-n\) relationships that are very similar to those reported in [49].

<table>
<thead>
<tr>
<th>No. Markers</th>
<th>HS-1-1</th>
<th>HS-2-2</th>
<th>HS-3-3</th>
<th>HS-4-4</th>
<th>HS-5-5</th>
<th>HS-6-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2200</td>
<td>1</td>
<td>0.925</td>
<td>0.605</td>
<td>0.515</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>22000</td>
<td>1</td>
<td>1</td>
<td>0.947</td>
<td>0.685</td>
<td>0.550</td>
<td>0.547</td>
</tr>
<tr>
<td>500K</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.878</td>
<td>0.612</td>
<td>0.551</td>
</tr>
<tr>
<td>1 million</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.872</td>
<td>0.647</td>
<td>0.557</td>
</tr>
</tbody>
</table>

Table 1: Average posterior probabilities of six extended half-sibling relationships for different numbers of markers when the only alternative relationship considered is that the two individuals of interest are unrelated. Averages are taken over 1000 simulated data sets for sets of 2200 and 22,000 markers and over 400 simulated data sets for half a million and a million markers.

Marker arrays of 2200 and 22,000 SNPs with corresponding allele frequencies were derived from the 500K SNP data by selecting 100 and 1000 evenly spaced markers from each autosome, respectively. In this paper, averages were taken over 1000 replicated data sets for arrays of 2200 and 22,000 markers and over 400 replicates for denser SNP arrays. It can be seen that inference improves with increasing numbers of SNPs and, as would be
Figure 2: Frequency histograms showing the distribution of posterior probabilities for the true pedigree versus the single alternative of ‘unrelated’ when the true pedigree is assumed to be HS-3-3, HS-4-4, HS-5-5 and HS-6-6 (columns) for SNP marker sets of size 2200, 22,000 and 500K (rows) respectively. Distributions were obtained over 1000 simulated data sets for the first two marker sets and over 400 replicates for 500K markers.

expected, identification of the true relationship from unrelated becomes more difficult as the relationship becomes more distant. We also confirmed (results not shown) that it is much more challenging to distinguish between the true pedigree and a close alternative structure (rather than just ‘unrelated’) after about 6 separating meioses with 500K SNPs [49, Table 3].

3.1 The Posterior Probability Distribution

It was noted in [49] that there was some variability in the replicates that contributed to the reported average posterior probabilities but this was not investigated any further. Figure 2 shows the posterior probabilities from
our simulations on four extended half-sibling relationships where the only alternative hypothesis is that the two individuals are unrelated. It can be seen that increasing the number of markers to 500K leads to increasingly skewed distributions. In particular, there is a clearer separation between high posterior probabilities where the true relationship is favoured and low posterior probabilities where it is not, reflecting the increase in information from the denser markers. As the relationship becomes more distant, the proportions of high posterior probabilities drop and the actual values of the lower probabilities increase. The decrease in numbers of high posterior probabilities makes sense because the probability of sharing segments of genome identical by descent (IBD) from the common pedigree ancestor, which is implicit in the likelihood calculations, decreases as the number of separating meioses increases. Hence, there is a higher chance that true relatives can have data that make them seem unrelated. However, when distant relatives do share IBD, there is often quite a lot of information as the realised sharing can be considerable [10, 17, 50, 54]. Higher posterior probability values for more distant relationships from simulations where the true pedigree is not favoured reflect the increasing evidence favouring the ‘unrelated’ alternative when the relatives share even less IBD than expected. Similar results (not shown) were obtained for extended sibling (or cousin) relationships.

We examined the realised sharing by simulating data from labelled founder alleles using Mendel [31]. Taking the distant cousin S-6-6 relationship as an example, 400 replicate sets of 500K SNP genotype data were simulated for S-6-6 relative pairs and posterior probabilities calculated for the true relationship when the only alternative is that of unrelated. All founder alleles were tracked through the pedigree so the actual sharing of the two individuals of interest was known for each replicate. In particular, from the realised IBD sharing, we can calculate what we will call the realised kinship coefficient. Specifically, if $X_1$ and $X_2$ are the founder allele labels for the alleles of individual A at a single SNP and $X_3$ and $X_4$ are the corresponding labels for B at the same SNP, we define the realised kinship as

$$\frac{1}{4} \{I_{[X_1=X_3]} + I_{[X_1=X_4]} + I_{[X_2=X_3]} + I_{[X_2=X_4]}\}$$

since there are four possible ways of randomly choosing an allele from each of the two individuals and these may, or may not, be IBD. This is analogous to what has been done elsewhere for moment-based estimation
Figure 3: Posterior probabilities of an S-6-6 relationship versus the only alternative of 'unrelated' from simulated 500K SNP data on 400 relative pairs for those cases where the actual, or realised, IBD sharing was lower than the expected, pedigree-based, sharing. The expected sharing is indicated by the vertical line to the right of the plot.

of kinship using the observed identity-by-state (IBS) sharing [16]. Our overall realised kinship averages the above quantity across all markers. The relative pairs in 220 replicates had no IBD sharing and so should appear to be unrelated. The posterior probability of the ‘unrelated’ alternative was indeed higher than that for the true pedigree in 218 (99%) of these. The expected kinship coefficient for an S-6-6 relationship is 0.0002441406 (i.e. 0.5^{12}) which corresponds to 407 shared IBD alleles for these data. Of the 180 replicates where the two relatives had some IBD sharing, the number of shared alleles ranged from as few as 12 to as many as 4816. When the realised IBD sharing exceeded the expected value, the true pedigree always had a posterior probability of 1. Figure 3 shows the posterior probabilities of the true pedigree when sharing was lower than expected. When the realised
kinship coefficient was greater than 0.0001 (i.e. less than half the expected value and corresponding to about 167 IBD alleles), we could still clearly distinguish the S-6-6 relatives from the unrelated alternative. For very low levels of sharing (about 88 alleles and a realised kinship coefficient of less than 0.00005), the individuals appeared to be unrelated. These findings correspond to the observed polarisation between high and low posterior probabilities for the denser sets of markers in Figure 2.

To investigate whether inference improves with additional SNPs, we also simulated Affymetrix v6.0 SNP data in Merlin for sets of 400 relative pairs with varying degrees of extended half-sibling relationships. Average posterior probabilities of the true pedigree, compared with the alternative of ‘unrelated’ are shown in the last row of Table 1. It would seem that 500K SNPs provide as much information as we can realistically hope for since we do not improve inference, on average, with a denser array. However, as we have already noted, average posterior probabilities can be misleading. The more dense the marker set, the more likely it is to have either a lot of information or else very little on any given relative pair. The more distant the relationship, the harder it is to obtain an informative data set. Generally, more SNPs yielded fewer intermediate values and samples were more clearly categorised. While we would still conclude that, on average, 6 separating meioses would seem like a comfortable upper limit for reliable inference from 500K markers, we note that, as in [27], we also have reasonable power up to 8 separating meioses (e.g. an average posterior probability of 0.878 for a HS-4-4 relationship as shown in Table 1). We would stress, however, that very good inferences for more distant relationships can be made if two individuals actually do share IBD and, in that case, denser marker sets are informative.

### 3.2 An Additional Typed Relative

Clearly, the more individuals in the pedigree that are typed, the more information there will be on the pairwise relationship of interest. In fact, additional marker data on one well-chosen extra relative can greatly improve inference. Mayor & Balding [36] showed that half-sibling pairs can be more easily distinguished from full siblings or unrelated individuals with standard sets of STR markers when maternal genotypes are available. Sieberts et al. [48] considered trios with all possible combinations of the four pairwise relationships of full sibling, half sibling, monozygous twin and unrelated for
STR data using a hidden Markov model for the IBD process to compute likelihoods for the 18 distinct three-way relationships. They concluded that data considered on three individuals jointly can be more informative than those derived from the three separate pairwise comparisons and that fewer markers are required. It has also been shown [57] that a third individual can be used as a control to reduce the chance of identity by state (IBS) being confused with identity by descent (IBD) thus leading to more reliable inferences of pairwise relatedness for unlinked marker data. Such additional information is very easily incorporated in the pedigree likelihood calculation and we are not limited to unlinked markers. Here we consider the addition of a typed relative, without any restriction to a specific set of close relatives, with view to addressing the issue of which relatives one should select for further genotyping should the option be available. Even for our simple unlooped pedigrees, there are potentially many additional relatives to choose from, so it is helpful to know which ones to target.

If we label our two individuals of interest as $A$ and $B$, we will categorise additional pedigree members as ‘within’ and ‘outside’ according to whether or not they lie on a direct line of descent from the common ancestor(s) of either $A$ or $B$. ‘Within’ relatives are thus related to both $A$ and $B$ and are hence expected to share some genome with both, whereas ‘outside’ relatives are either founders, or related via a founder to one of the two, and can hence only share some genome with either $A$ or $B$. Siblings are always ‘within’
whereas half-sibling, cousin or avuncular relatives can be either ‘within’ or ‘outside’ (Figure 4). For direct-line ancestor relatives, we will use ‘within-n’ and ‘outside-n’ to indicate the number of separating meioses between that relative and A or B. So a parent on the direct line of descent is a ‘within-1’ relative and a founder great grandparent is an ‘outside-3’ relative (Figure 4).

In order to make inferences about the relationship between A and B using additional data from a third individual, C, care must be taken with the formulation of the alternative hypothesis. Consider testing the hypothesised HS-3-3 relationship of Figure 5 (a) for A and B with additional data on C, a ‘within-1 relative of B, versus the hypothesis that A and B are unrelated. The alternative pedigree of Figure 5 (b) assumes that the relationship between B and C is known and hence remains unchanged across hypotheses. The alternative shown in Figure 5 (c), although it might seem to be the natural analogue to the standard situation with pairwise data, does not make this assumption and hence represents the alternative that all three individuals are unrelated. For our problem, the correct alternative is (b) since the relationship of interest (i.e. between A and B) is the only one that varies between the two choices. Any differences in likelihood can thus be ascribed to the difference in this relationship. For the alternative (c), all three pairwise relationships involving A, B and C differ between the two options and any of these pairwise differences could induce a change in the likelihood. For instance, pedigree (a) might seem far more likely than pedigree (c) in the presence of overwhelming evidence that C really is the parent of B and may have nothing whatsoever to do with the relationship between A and B.

To distinguish between a given relationship between A and B (e.g. HS-4-4) with an additional typed relative C (e.g. a parent of B) and the simple alternative that A is unrelated to B and C, 500K SNP data were simulated for different structures and different choices of the extra relative. Posterior probabilities were calculated over 400 replicate data sets. Average posterior probabilities of the ‘true’ relationship always increased with the addition of genotype data on a ‘within-n’ relative and the closer this relative was to the common ancestor, the more information was gained. This is because a typed relative on the direct line of descent from a common ancestor essentially makes the relationship closer. For example, in Figure 4, if C, parent of B (within-1) is typed, the S-4-3 relationship between A and B becomes an S-4-2 relationship between A and C which is easier to detect. In dense SNP data, the information on the original individual B is seemingly redundant as the average posterior probabilities for the hypothesised relationship between A
Figure 5: A HS-3-3 pedigree connecting a pair of individuals A and B with an additional typed relative C (a) with two choices for the alternative hypothesis that A and B are unrelated (b) and (c).

and B versus the single alternative of ‘unrelated’ using all three individuals are very close to those obtained for the relationship between A and C, ignoring the data on B. This would actually be the case if the markers were unlinked.

Additional data on relatives at ‘outside’ positions — even for close relatives — had little or no effect on the average posterior probabilities. An ‘outside’ relative of A can only share alleles IBD with B at loci where A and B already share IBD whereas a ‘within’ relative can also share some alleles IBD with B at additional loci. There is already a lot of information in dense SNP data if A and B actually do share parts of their genome IBD. The likelihood of the ‘true’ pedigree is much higher than that of the alternative, in this case, so the posterior probabilities are already very close to one. Using three instead of two related individuals — wherever they are in the pedigree — always increases the likelihood ratio when linked markers are used, but the effect for ‘outside’ relatives is only reflected in the posterior probabilities for less dense sets of SNPs (results not shown).

Table 2 shows the posterior probabilities from simulations for each of HS-2-2, . . . , HS-5-5 and ‘unrelated’ versus all five alternatives. Additional marker data for ‘within-1’ and ‘within-2’ relatives, corresponding to a typed parent or grandparent were considered, as we would not expect to have data on great-grandparents, or more distant relatives, in practice. There is a clear improvement in our ability to distinguish a given relationship from a set of specified close alternatives when data on an additional ‘within’ relative are
available and a typed grandparent is noticeably more informative than a typed parent.

Table 3 illustrates the effect of a third genotyped individual in distinguishing a HS-4-4 relationship from the simple alternative of ‘unrelated’ for a selection of relatives that typically might be available in a practical situation. These include parents (Pa), grandparents (GPa), siblings (sib), half siblings (HS), avuncular (AV) and cousin (C) relatives at both ‘within’ (W) and ‘outside’ (O) positions. Posterior probabilities are averaged over 400 replicated sets of simulated 500K SNP data. As noted above, typing individuals at ‘outside’ positions is not necessarily helpful as they are only informative when there is already a lot of information on the pairwise relationship of interest. Grandparents are more informative than parents and parents are more informative than siblings. ‘Within’ half siblings are as good as siblings and ‘within’ avuncular relatives and cousins are slightly better. In
Table 3: Average posterior probabilities of a HS-4-4 relationship, when the only alternative relationship considered is that the two individuals of interest are unrelated, taking additional marker data for different choices of a close relative into consideration. Averages are taken over 400 simulated data sets.

the absence of any prior information regarding the lines of descent though, it would seem that siblings are the obvious choice to genotype. Alternatively, typing both parents — or all four grandparents — if available would be more informative.

3.3 Effect of Linkage Disequilibrium

So far, we have assumed linkage equilibrium among founder genotypes, both in the simulations and in the likelihood calculations. In practice, this assumption will not hold and will be particularly untrue for dense sets of markers. Merlin [1] has an option for incorporating linkage disequilibrium (LD) into both data simulation and likelihood calculations so, although it is not the only way to model LD, it provides a means to explore the effects of ignoring LD on our results.

Merlin models LD by combining tightly linked markers into clusters and estimating haplotype frequencies within each cluster. It is assumed that markers can be grouped into non-overlapping clusters of consecutive markers such that all markers within a cluster are in LD, markers in different clusters have very low levels of LD and the recombination rate within a cluster is very low [2]. The cluster-specific haplotype frequencies describe the level of LD. To make the likelihood computation tractable, it is assumed that there is no LD between clusters and no recombination within clusters. Merlin can generate clusters from observed genotype data using a specified map distance or a given level of LD, as measured by $r^2$, so that contiguous markers within
Table 4: Posterior probabilities for HS-4-4 and unrelated pairs versus six specified alternatives when data were simulated with and without LD and likelihoods either accounted for or ignored LD. Posterior probabilities were calculated for 400 pairs for each scenario (row) considered. Simulations were based on the 711,020 SNPs typed in the UK National Blood Service cohort.

<table>
<thead>
<tr>
<th>True</th>
<th>HS-1-1</th>
<th>HS-2-2</th>
<th>HS-3-3</th>
<th>HS-4-4</th>
<th>HS-5-5</th>
<th>Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS-4-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No LD at all</td>
<td>0</td>
<td>0</td>
<td>0.173</td>
<td>0.467</td>
<td>0.263</td>
<td>0.097</td>
</tr>
<tr>
<td>LD: unaccounted</td>
<td>0</td>
<td>0.001</td>
<td>0.608</td>
<td>0.339</td>
<td>0.051</td>
<td>0.001</td>
</tr>
<tr>
<td>LD: accounted</td>
<td>0</td>
<td>0</td>
<td>0.164</td>
<td>0.479</td>
<td>0.279</td>
<td>0.077</td>
</tr>
<tr>
<td>LD: 20K SNPs</td>
<td>0</td>
<td>0.001</td>
<td>0.215</td>
<td>0.367</td>
<td>0.246</td>
<td>0.171</td>
</tr>
<tr>
<td>LD: 30K SNPs</td>
<td>0</td>
<td>0.0003</td>
<td>0.218</td>
<td>0.406</td>
<td>0.240</td>
<td>0.136</td>
</tr>
<tr>
<td>LD: 50K SNPs</td>
<td>0</td>
<td>0.0014</td>
<td>0.207</td>
<td>0.435</td>
<td>0.242</td>
<td>0.154</td>
</tr>
<tr>
<td>Unrelated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No LD at all</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.085</td>
<td>0.323</td>
<td>0.590</td>
</tr>
<tr>
<td>LD: unaccounted</td>
<td>0</td>
<td>0</td>
<td>0.256</td>
<td>0.549</td>
<td>0.181</td>
<td>0.014</td>
</tr>
<tr>
<td>LD: accounted</td>
<td>0</td>
<td>0</td>
<td>0.001</td>
<td>0.085</td>
<td>0.324</td>
<td>0.589</td>
</tr>
<tr>
<td>LD: 20K SNPs</td>
<td>0</td>
<td>0</td>
<td>0.021</td>
<td>0.197</td>
<td>0.352</td>
<td>0.429</td>
</tr>
<tr>
<td>LD: 30K SNPs</td>
<td>0</td>
<td>0</td>
<td>0.013</td>
<td>0.168</td>
<td>0.355</td>
<td>0.466</td>
</tr>
<tr>
<td>LD: 50K SNPs</td>
<td>0</td>
<td>0</td>
<td>0.006</td>
<td>0.140</td>
<td>0.352</td>
<td>0.502</td>
</tr>
</tbody>
</table>

Table 4: Posterior probabilities for HS-4-4 and unrelated pairs versus six specified alternatives when data were simulated with and without LD and likelihoods either accounted for or ignored LD. Posterior probabilities were calculated for 400 pairs for each scenario (row) considered. Simulations were based on the 711,020 SNPs typed in the UK National Blood Service cohort.

the given distance or with a pairwise $r^2$ greater than the specified value are placed into one cluster. The cluster haplotype frequencies are then estimated from the data. Alternatively clusters and haplotype frequencies can be generated elsewhere and read into Merlin and incorporated in the likelihood calculations.

17
We used Merlin to model LD from the real Affymetrix v6.0 SNP data on 1285 individuals from the UK National Blood Service cohort and simulated data with and without LD. When LD was simulated but ignored in the likelihood calculations, relative pairs always appear to be more closely related than they actually are. In particular, completely unrelated individuals can appear to be related. Table 4 shows the results for data simulated on pairs of HS-4-4 and pairs of unrelated individuals with and without LD where SNPs were combined into clusters if they had a pairwise genetic distance of 0.001cM or less. Posterior probabilities for the ‘true’ relationship versus several alternatives were calculated as given in Equation (1). The first three rows for each case (HS-4-4 relationship and ‘unrelated’) correspond to the following situations; no LD simulated and no LD accounted for in the likelihood calculations; LD simulated but not accounted for in the likelihood; LD simulated and properly accounted for in the likelihood. The posterior probabilities recorded in rows 1 and 7 yield the inferences we would expect in the absence of LD and correspond closely to values reported elsewhere [49, Table 3]. They confirm that it is difficult to distinguish a HS-4-4 from a HS-5-5 relationship and a HS-5-5 relationship from ‘unrelated’, for example. The very same inferences can be drawn when LD is simulated but is properly accounted for in the likelihood calculations (rows 3 and 9). However, when LD is really present but ignored, HS-4-4 pairs tend to look like HS-3-3 relatives while unrelated individuals can appear to be HS-4-4 relatives.

In this situation, we knew the true data-generating LD model and could account for it perfectly. Not surprisingly, results are sensitive to misspecification of the LD model. An appropriate model for LD will typically not be available for real data and is probably why a thinned set of SNPs is preferred in practice. In order to ‘prevent potential bias due to linkage disequilibrium’, a sparse subset with a distance of at least 300kb between neighbouring markers is recommended in [5] thereby removing about 84% of the available SNPs. The number 10,000 is suggested elsewhere [27] although this number appears to derive from the computational limitations of the software used in the cited reference [42]. No more than 20,000 markers is suggested as an upper limit for testing relationships closer than second cousins [29] and there is considerable focus in the literature on finding the minimal subset of SNPs that will suffice [29, 42, 43].

We considered different subsets of SNPs where we thinned our N markers to a subset of size \( \frac{N}{n} \) by selecting every nth. Thus, for our Affymetrix v6.0
data, a 20,000 SNP subset is obtained by taking roughly every 35th SNP. Posterior probabilities based on thinned subsets of 20,000, 30,000 and 50,000 SNPs are shown in Table 4 for the HS-4-4 (rows 4 to 6) and Unrelated (rows 10 to 12) cases when LD was simulated but not accounted for in the likelihood calculations. It can be seen that the thinned subsets all give highest posterior probability to the true relationship in contrast to when all SNPs are used. In this case, inference improves with increasing density of the subsets with 50,000 SNPs almost doing as well as the whole set when LD is either absent or appropriately modelled.

In the above example, LD was modelled between SNPs within a genetic distance of 0.001cM. This resulted in 145,100 clusters involving 56% (401,027) of the available SNPs. Although there were some large LD blocks (for example, there were 17 that contained at least 50 markers), the number of SNPs per block was generally very low with a median value of 2 SNPs. Increasing the genetic distance would include more SNPs in LD blocks and would increase the average block size but this option is computationally intensive due to the possibility of having some very large blocks. Merlin does not allow very large blocks with the \( r^2 \) blocking criterion as there is a system limit of 21 SNPs per block imposed. Using \( r^2 > 0.2 \) as the clustering criterion yielded fewer blocks (58,950) covering 97% of the available SNPs (690,056) with a median value of 12 markers per block. Results were very similar to those reported in Table 4. Hence, reducing the SNP density to 50,000 was sufficient to remove most of the LD in these data for these LD patterns.

These simulations show that ignoring LD can have severe effects on relationship estimation. While thinning dense SNPs can clearly alleviate bias due to LD, our results are not sufficient to inform on the ‘ideal’ subset. Firstly, this number will depend on the relationship under consideration [29]. In particular, more markers are required for most distant relatives so there is a trade-off between thinning to reduce LD effects and retaining a sufficiently informative set of markers. Secondly, it does not seem sensible to specify an ideal marker set size without consideration of the underlying data-generating model.

Short tandem repeat (STR) marker sets will typically be less affected by LD and supplementing a thinned set of SNPs with such markers, should they be available, may be a reasonable approach to take when strong LD is suspected and cannot be modelled.
3.4 A Pedigree Likelihood Approach to Inferring Relatedness

In many applications, the precise relationship as defined by a pedigree may not be required and an estimate of overall relatedness may suffice. Here we propose that a set of extended sibling (or half-sibling) pedigree structures could serve as a rough template against which to test a given relative pair. A set of structures representing all reasonable numbers of separating meioses, such as S-1-1, S-1-2, S-2-2, S-2-3 . . . , S-8-8 and unrelated, for example, can be used to estimate the degree of relatedness. A coarser template, such as a sequence of HS-n-n or S-n-n pedigrees, would suffice if the aim is simply to distinguish relatives from non-relatives.

3.4.1 Comparison with a Moment-based Estimator

Moment-based estimators of relatedness consider the genotypes of two individuals and measure their similarity, locus by locus. In contrast with the likelihood approach, marker positions and inter-dependencies are ignored so resulting estimates are not as sensitive to the effects of LD. Estimators differ in how they weight information across multiple loci and are generally easy to implement. Bias tends to be low, but this is largely due to the fact that they can yield negative estimates of coefficients of kinship and relationship. In a likelihood setting, these quantities are always thought of as probabilities but for moment-based estimators, it makes more sense to think of them as correlation coefficients with negative values indicating lower than expected levels of sharing [58]. Indeed, the coefficient of relatedness (twice the kinship coefficient) has a natural interpretation as the correlation, or regression, of genetic values of individuals in quantitative genetics. If the precise relationship between the two individuals is not of interest, such estimators work well up to cousin (S-2-2) relationships. They tend not to perform so well for distant relatives and have high sampling variance [57].

Here we consider just one of these estimators, proposed by Lynch & Ritland [35], which has been shown to work well in other settings [58]. For our simulated data on S-6-6 relative pairs in Section 3.1, we noted that our pedigree likelihood approach could successfully distinguish an S-6-6 relative pair from ‘unrelated’ when the realised sharing was at least half that expected (Figure 3). As can be seen in Figure 6, this is not the case for the method-of-moments (MoM) estimator, which is just as likely to provide negative
estimates of kinship as positive estimates when the realised sharing is similar to the expected sharing. Indeed, positive estimates are possible even where there is no IBD sharing. Even though seemingly unbiased, the variability of the estimates increases with denser SNPs and we found that performance was degraded even for close relationships when dense SNP sets were used. As reported by Milligan [37], the performance of moment-based estimators also depends on the actual relationship under consideration.

In many practical situations, relationship-specific alternatives may not be available and being able to distinguish relatives from non-relatives might suffice. For illustration, we return to our simulated 500K SNP data on 400 pairs of S-6-6 relatives and test them against the alternatives S-1-1, S-2-2, . . . , S-8-8 and ‘unrelated’. The aim is simply to test whether any of the specified alternatives is more likely than the ‘unrelated’ option and no significance is attached to the particular relationship that is seemingly selected. The relationship yielding the highest posterior probability for each pair is plotted against the realised kinship coefficient in Figure 7. As noted previously,
Figure 7: Relationship estimate based on posterior probabilities (or scaled likelihoods) for S-1-1, S-2-2, . . . , S-8-8 and unrelated alternatives versus realised kinship coefficient for simulated 500K SNP data on 400 S-6-6 relative pairs. The value \( n \) on the y-axis represents S-\( n-n \) with 9 standing for 'unrelated'. The vertical line to the left of the plot represents the value 0.0001.

Pairs with no IBD sharing will be assigned to the ‘unrelated’ category with very high probability. Those with realised sharing greater than 0.0001 (less than half the expected sharing for S-6-6) will always be assigned to one of the S-\( n-n \) options and hence can be inferred to be related. The assigned relationship will completely depend on the actual or realised sharing and is not necessarily indicative of the true relationship. Thus, for example, the true S-6-6 pedigree rarely has the highest posterior probability and individuals are more likely to be inferred as S-5-5 relatives. This is because distant relatives are quite likely not to share any part of their genomes IBD and so the realised sharing for those that actually do share is greater than the expected value. Our simulated data here were generated under the assumption of linkage equilibrium. Relationships would seem closer if LD were present and not properly accounted for.
3.4.2 Comparison with a Chromosomal Segment Sharing Approach

A different approach to inferring relationships uses variation in the number, length and position of the shared IBD segments [25]. Great-grandparent—great-grandchild, great-uncle—great-nephew, half-uncle—nephew and first cousin relatives all have the same (pedigree-based) expected sharing, for instance, but the variation in the realised sharing is quite distinct enabling them to be distinguished in principle. A likelihood approach based on the number and lengths of estimated IBD segments is implemented in the software ERSA [27, 34]. A likelihood ratio test is performed where the null hypothesis represents the expected population background sharing and the alternative hypothesis proposes that the two individuals share recent ancestry which is parameterised by the number of common ancestors (1 or 2) and the number of separating meioses. Applications to real data report 80% correct assignment to within one meiosis for sixth- and seventh-degree relatives [27].

Maximising the likelihood over ‘all possible relationships’, as is carried out in ERSA, is very similar to our proposed use of simple unlooped pedigrees as a template for estimating the degree of relatedness. We simulated 400 sets of 500K SNP genotype data for S-3-4 (corresponding to sixth-degree) and S-4-4 (corresponding to seventh-degree) relative pairs and compared results. ERSA was correct to within one separating meiosis in 83% of the sixth-degree relatives and in 75% of the seventh-degree pairs and the results are hence consistent with the published reports of 80% on average. The pedigree likelihood approach, using the set S-1-1, S-1-2, S-2-2, S-2-3 . . . , S-8-8 and ‘unrelated’ to cover all reasonable numbers of separating meioses, had a 94.5% success rate in detecting an S-3-4 relationship (or one that differed by one meiosis) and an 89% success rate for the S-4-4 relationship (to within one meiosis).

The above data were simulated without LD. The pedigree likelihood approach is sensitive to LD but can exploit information in denser sets of SNPs for distant relationships when LD is not an issue, so our approach will possibly have been favoured in this comparison. Taking a thinned subset of 22,000 SNPs from each replicate, as might be sensible if LD were suspected and could not be adjusted for, we obtained an 85.7% success rate for sixth-degree relatives and 81.5% success for seventh-degree relatives with the thinned set which still compare very favourably with those from ERSA. We also simulated 500K SNP data with LD, using a distance threshold of
and artificially created haplotype frequencies. ERSA performed very similarly on these data with 87.5 % and 74 % success rates for sixth- and seventh-degree relatives, respectively, over 400 replicates. The pedigree likelihood approach, on the other hand, yielded corresponding success rates of 64.23% and 4% when LD was not accounted for, but these rates improved dramatically to 93% and 78.25 % on thinning.

3.5 Applications to the MICROS Data

In order to assess how our methods work in real data, we sought genotyped relative pairs in the 8-generation MICROS pedigree. The most distant recent relationship for which we could find sufficient numbers of pairs was that of second cousin i.e. S-3-3. We considered 304,383 autosomal SNP genotypes on 101 such pairs. Allele frequencies, required for both MoM and pedigree likelihood approaches, were estimated from the 1285 observed individuals in the absence of any external information. Expected kinship coefficients were calculated for each pair based on the larger pedigree and mostly exceeded the nominal value of 0.015625 for a second cousin relationship, reflecting the complexity of this pedigree. MoM estimates using the Lynch & Ritland estimator (Section 3.4) tended to be lower than expected, on average, and were quite variable (Figure 8). Some of this variability is due to variability in the realised sharing which, of course, is unobserved but negative bias for inbreeding coefficient estimates using other single point estimators was also noted in a recent comparison [22]. Even though we would not regard second cousins as being particularly distant, it would seem that we are already in a situation where moment-based methods perform poorly.

We now considered our pedigree likelihood approach and tested each of these relative pairs against a range of symmetric extended cousin (S-n-n) relationships plus the alternative of ‘unrelated’. As can be seen from the first row of Table 5, the closest alternative of S-1-1 was the most likely. While we might expect that many of these pairs could appear to be more closely related than S-3-3 due to the possibility of their being related via more distant ancestors in the 8-generation pedigree, we would not expect them all to look like full siblings. This extreme result is clearly due to the overall background relatedness or LD in the population. It concurs with results reported elsewhere [29] although we did not see a tendency to choose the closest alternative in simulated dense SNP data as those authors found
(Section 3.3). We suspect this is due to the fact that we used the ‘true’ allele frequencies in the simulation studies and have estimated them from the data here [57, 58].

Figure 8: Estimated kinship coefficients, using the Lynch & Ritland moment-based estimator, for 101 relative pairs from the MICROS study whose most recent relationship is second cousin plotted against the expected kinship calculated from the larger 8-generation pedigree.

We modelled LD in the MICROS data using Merlin as discussed in Section 3.3 and adjusted for this model in the likelihood calculations. None of the models that we tried seemed to be really appropriate for these data. Results for two choices of $r^2$ are shown in rows 2 and 3 of Table 5. With $r^2 > 0.6$ as the blocking criterion, our relative pairs still tended to look like full siblings. Setting LD to $r^2 > 0.2$ and widening the grid to 10cM resulted in a more reasonable looking distribution although closer relationships were still favoured. Most of the SNPs (283,658) were included in 38,671 LD blocks with this latter option and the median block size was 6 SNPs. LD models based on confidence bounds for $D'$ are also recommended [21] and implemented in popular software such as Haploview [4]. For comparison, we used a fast version of this algorithm [52] on the MICROS data using the standard thresholds. There were 69,280 LD blocks covering 245,341 SNPs.
Table 5: 101 relative pairs from the 8-generation MICROS pedigree whose most recent relationship is second cousin tested against a suite of extended cousin relationships and ‘unrelated’. The counts represent the numbers of pairs having highest posterior probability for the given alternative. Inference is performed using all available 304,384 autosomal SNPs ignoring LD (row 1), accounting for LD using $r^2 > 0.6$ or $>0.2$ as the LD blocking criterion (rows 2 and 3), accounting for LD using standard thresholds on $D'$ as the blocking criterion (row 4), thinned subsets of SNPs (rows 5 to 7) and 1067 STR markers (row 9).

Using this method and the median block size was 3 SNPs. As can be seen in Table 5, the results were even more biased towards the closer alternatives than when $r^2 > 0.6$ was used as the blocking criterion.

We then thinned the SNPs and considered subsets of 50K, 30K, 20K and 10K on which to base our calculations. As can be seen in rows 4 to 7 of Table 5, the bias towards the closest alternatives decreases with decreasing size of marker set. At 10,000 SNPs we begin to see a few cases where a third alternative is more likely than the closest one.
cousin (S-4-4) relationship is favoured indicating that we are now beginning to lose some discriminating power. It would seem that thinning to somewhere between 10 and 20K SNPs is about right for reducing the effects of LD in these data but this is very much arrived at by trial and error and should not be taken as a general recommendation.

Since the MICROS pedigree is so complex, it was not possible to find a sufficiently large set of relative pairs with exactly the same degree of relationship. However, we were able find 148 relative pairs with expected kinship (based on the full pedigree) within 1% of that for an S-4-4 relationship and a further 116 pairs with expected kinship within 1% of an S-4-3 relationship. We tested the pedigree likelihood approach to estimating degree of relatedness, as described in Section 3.4.2, on these relative pairs using 20,000 SNPs. We were accurate to within one meiosis in 77% of the S-4-4 cases and accurate to within one meiosis in 94% of the S-4-3 cases. Implementing ERSA on these data proved to be far less straightforward. The method uses either Germline [24] or Beagle [9] to infer the IBD segments which it requires as input. Neither program could estimate IBD segments for the MICROS data. However, our pedigree likelihood results are very much in line with the reported value of 80% for sixth- and seventh-degree relatives combined using ERSA on real data [27]. Because we have not been successful at modelling the LD in these data, using 20K SNPs is probably as far as we can go.

Individuals in the MICROS study were also typed at 1067 STR markers. We calculated posterior probabilities for our set of S-n-n relationships using the available STR data on our 101 relative pairs. Results (row 8 of Table 5) were comparable with those for the thinned subset of 10,000 SNPs. This supports findings from simulation studies, using allele frequencies estimated from these data, implying that one STR marker is roughly as informative as 10 ‘independent’ SNPs in a dense marker array. Supplementing these markers with a thinned subset of SNPs will improve inference but will not recover the potential information that can be derived from dense SNPs and adjusting for LD with an appropriate model.

Finally, we searched the MICROS pedigree for distant relative pairs who also had additional typed relatives that could be included. In most cases, the pairwise information was sufficient to distinguish the relationship in question from ‘unrelated’. However, we did find one pair of S-6-4 relatives who appeared to be unrelated until we incorporated the genotype of a sibling into the likelihood calculations. LD was not accounted for in these investigations.
but we did use a thinned set of 50K SNPs. As already noted, this is unlikely to have completely resolved the problem and so our inference about the actual relationship is subject to bias for the reasons discussed above.

4 Discussion

We have considered the exploitation of dense SNP arrays for inferring distant relationships and relatedness. We have taken a pedigree likelihood approach and have viewed the problem as one of finding the true pedigree connecting the two individuals of interest from a set of plausible alternatives for which we would typically have prior information. We would warn that average posterior probabilities (and standard errors) are misleading with dense SNP arrays because the posterior distributions are so skewed. The probability that a distant relative pair actually shares part of the genome IBD is quite small and so finding such relatives in practice could be difficult. However, when they do share, they can be detected and denser SNP arrays provide more distinguishing information. Data on an additional relative can also help to distinguish a pairwise relationship. In our approach, such information is very easily accommodated in the pedigree likelihood calculation enabling consideration of data on all three individuals jointly. For dense SNP arrays, parents, grandparents or siblings provide the most information. Data on any relative might suffice for less dense marker sets. Our results will depend on the set of alternative hypotheses considered: if the true pedigree is not listed among the alternatives, inference on the relationship will be incorrect but we may still get a very good idea of relatedness.

Since chromosomal segments, and not individual alleles, are the actual units of inheritance, methods that infer shared IBD segments are obviously attractive for inferring relatedness. They all work on the underlying idea that shared haplotypes of low frequency must be IBD. In the absence of a pedigree where the notion of IBD is clearly defined, they need to have a reference population, or baseline, relative to which IBD is traced [9, 54, 58]. IBD segment detection approaches are either rule-based using a threshold for segment length or employ a probabilistic hidden Markov model (HMM) for underlying IBD status [33]. (See [10] for a recent overview.) For genotype data, a simple approach is to search for shared chromosomal segments that are consistent with IBD and to infer IBD if the segment length exceeds a given threshold [38]. For phased data, haplotypes can be directly compared
and IBD is inferred if shared segments are either sufficiently long or if the estimated frequency is sufficiently low [24, 8]. Most IBD detection methods are restricted to consideration of pairs of individuals [10] and all approaches seem to draw similar conclusions when their various modelling assumptions are met.

We saw that the presence of linkage disequilibrium affected our estimates, typically biasing inference in favour of closer relationships. LD cannot be ignored. Moment-based estimators are not so badly affected and may work reasonably well for close relationships but are unpredictable for distant relatives. A recent comparison of methods for estimating inbreeding coefficients found that they were negatively biased whereas rule-based approaches were reliant on having the correct threshold value. Approaches using HMMs were favoured in that study but it was noted that they were also sensitive to LD. Moreover, HMM methods that modelled LD did not do as well as HMM methods using thinned data [22]. This agrees with our findings for the pedigree likelihood approach as it also requires a good LD model. However, while thinning to a subset of SNPs, as is commonly done in practice, may alleviate the bias due to LD, we would warn against a general recommendation for a marker set size as this will depend on the relationship being considered and the levels of LD in the data. There is clearly a trade-off between alleviating bias due to LD and losing information for more distant relatives when thinning dense sets of SNPs. There are many approaches to modelling LD from genotype data but they are all computationally intensive. Moreover, as we saw with the MICROS example, building a realistic model might prove to be difficult for a given data set and standard settings may not be appropriate.

The underlying assumptions of the ERSA approach based on the number and lengths of estimated IBD segments have been criticised in [26] who propose a simulation-based approach instead. Both methods require accurate detection of IBD segments however, and this may not be straightforward as we encountered with the real data from the MICROS study. Maximising the likelihood over ‘all possible relationships’, as is carried out in ERSA, is very similar to our proposed use of simple unlooped pedigrees as a template for general relatedness. The latter approach also incorporates information on position, number and lengths of shared segments implicitly in the pedigree likelihood via the linkage map, but inference is relative to the hypothesised pedigree. ERSA is more robust to the effects of LD, as background sharing is accounted for, and its performances on simulated data with and without
LD were comparable. On simulated data with no LD, or by using a thinned set of SNPs, the pedigree likelihood method performed just as well if not slightly better. We could not run ERSA on our real data due to the problem with inferring the IBD segments. Our own approach is sensitive to LD and we were not able to model LD satisfactorily in these data. However, we were able to use a thinned set of SNPs and obtained results from these that were comparable with the published findings using denser SNPs in [27].

Our findings based on simulated data assume no genotyping errors, no background relatedness, and that allele frequencies are known. Likelihood calculations on real data would have to incorporate an error rate [48]. Since STRs have high mutation rates, a mutation model should also be incorporated but this can reasonably be ignored for SNP data. All approaches are sensitive to allele frequencies, either directly or at the phasing stage if phased data are required, and estimates of allele frequencies from the data could be poor [57]. In particular, using estimated frequencies rather than externally derived frequencies tends to make individuals look more closely related than they actually are. Population stratification will also affect inference via this sensitivity to allele frequencies [39, 58].

There is obviously a limit on what can be deduced, on average, from the amount of genome shared by distant relatives due to skewness and overlapping of distributions of the realised sharing [25, 26]. While the pedigree likelihood approach that we propose here is not necessarily better than existing approaches, it competes well and has some obvious advantages. In particular, it can be implemented with standard, freely available software, it exploits the available information in dense marker data, it can easily accommodate data on additional individuals provided there is some information on their position in the pedigree and it can, in principle, adjust for LD given an appropriate model. In order to reliably infer relationships beyond second or third cousins, however, dense SNP data and a good model for linkage disequilibrium will be required. Also, rare alleles should be more informative for IBD than common ones and these should have more effect in denser data. Inferring the degree of relatedness is simpler, however, and the method competes very favourably with two popular approaches. Likelihood calculations are computationally intensive for dense SNPs so thinning the data can also have a practical advantage. However, this should not be done without some consideration of the underlying LD model.
5 Acknowledgements

The authors acknowledge research support from Medical Research Council Project Grant G1002312 (NAS) and Wellcome Trust Senior Research Fellowship 087576 (MAJ). The MICROS study was supported by the Ministry for Health of the Autonomous Province of Bolzano and the South Tyrolean Sparkasse Foundation. We are also grateful to Martin Gögele and Cristian Pattaro for discussions and help with the pedigree data and two anonymous referees for their very helpful comments and suggestions. The research in this paper used the ALICE High Performance Computing Facility at the University of Leicester.

References


[17] K.P. Donnelly. The probability that related individuals share some
section of the genome identical by descent. *Theoretical Population

paternity and identification cases. Selecting the most probable pedigree.


[21] S B Gabriel, S F Schaffner, H Nguyen, J M Moore, J Roy, B Blumenstiel,
J Higgins, M DeFelice, A Lochner, M Faggart, S N Liu-Cordero,
C Rotimi, A Adyemo, R Cooper, R Ward, E S Lander, M J Daly, and
D Altshuler. The structure of haplotype blocks in the human genome.

[22] S Gazal, M Sahbatou, H Perdry, S Letort, E Génin, and A L
Leutenegger. Indreeding coefficient estimation with dense SNP data:
comparison of strategies and application to HapMap III. *Human

[23] C Glazner and E A Thompson. Improving pedigree-based linkage
analysis by estimating co-ancestry among families. *Statistical
Applications in Genetics and Molecular Biology*, 11, 2012. Issue 2,
Article 11.

[24] A Gusev, J K Lowe, M Stoffel, M J Daly, D Altschuler, J L Breslow,
J M Friedman, and I Pe’er. Whole population, genome-wide mapping


