Molecular Ecology (2003) **12**, 1039–1047



# Prospects for inferring pairwise relationships with single nucleotide polymorphisms

JEFFREY C. GLAUBITZ, O. EUGENE RHODES JR and J. ANDREW DEWOODY
*Purdue University, Department of Forestry & Natural Resources, 195 Marsteller St, West Lafayette, Indiana, USA 47907–2033, USA*

## Abstract

**An extraordinarily large number of single nucleotide polymorphisms (SNPs) are now available in humans as well as in other model organisms. Technological advancements may soon make it feasible to assay hundreds of SNPs in virtually any organism of interest. One potential application of SNPs is the determination of pairwise genetic relationships in populations without known pedigrees. Although microsatellites are currently the marker of choice for this purpose, the number of independently segregating microsatellite markers that can be feasibly assayed is limited. Thus, it can be difficult to distinguish reliably some classes of relationship (e.g. full-sibs from half-sibs) with microsatellite data alone. We assess, via Monte Carlo computer simulation, the potential for using a large panel of independently segregating SNPs to infer genetic relationships, following the analytical approach of Blouin *et al.* (1996). We have explored a 'best case scenario' in which 100 independently segregating SNPs are available. For discrimination among single-generation relationships or for the identification of parent–offspring pairs, it appears that such a panel of moderately polymorphic SNPs (minor allele frequency of 0.20) will provide discrimination power equivalent to only 16–20 independently segregating microsatellites. Although newly available analytical methods that can account for tight genetic linkage between markers will, in theory, allow improved estimation of relationships using thousands of SNPs in highly dense genomic scans, in practice such studies will only be feasible in a handful of model organisms. Given the comparable amount of effort required for the development of both types of markers, it seems that microsatellites will remain the marker of choice for relationship estimation in nonmodel organisms, at least for the foreseeable future.**

*Keywords*: computer simulation, microsatellites, pairwise relationship, pedigree reconstruction, relatedness, single nucleotide polymorphisms

*Received 1 October 2002; revision received 9 January 2003; accepted 9 January 2003*



## Introduction

Scientists have long been interested in the use of genetic markers to unravel the web of genealogical relationships present in a population (Thompson 1975; Pamilo & Crozier 1982). This pursuit has relevance to both the study of wild populations and to the genetic management of captive and/or threatened populations. Knowledge of relationships in wild populations is valuable in studies of kin selection and social behaviour/organization (e.g. Hamilton 1964; Morin *et al.* 1994), mating systems (e.g. Heg & van Treuren 1998; Engh *et al.* 2002), dispersal, isolation by distance, and spatial genetic structure (e.g. Goodisman & Crozier 2002), and for the estimation of quantitative genetic parameters such as heritability (Ritland 2000). In captive populations, knowledge of relationships and shared ancestry allow for the minimization of inbreeding by permitting matings only between the most distantly related individuals (e.g. Jones *et al.* 2002).

In the absence of a known pedigree, the only way to determine relationships is to employ genetic markers. However, in practice, true genetic relationships can be difficult to infer, primarily because of the large sampling variance of pairwise relatedness ($r$) among loci. In addition to variance among sampled loci, inherent variance in $r$ exists within a relationship category; for example, the proportion of genome shared between full-sibs varies (Wang 2002).

Correspondence: J. C. Glaubitz. Fax: 765–496–2422; E-mail: glaubitz@fnr.purdue.edu

Because of the high level of sampling variance, the estimation of $r$ (or the inference of genetic relationships) would be ideally conducted with a large number of independently segregating loci, each with a large number of alleles of roughly even frequencies (Lynch & Ritland 1999).

Although kinship has been evaluated with allozymes (e.g. Avise & Shapiro 1986), microsatellites are currently the marker of choice for pedigree reconstruction (e.g. Taylor *et al.* 1997). However, there are practical limits on the number of microsatellites that can be efficiently surveyed in most studies. Blouin *et al.* (1996) have shown that, on the basis of the expected distributions of pairwise $r$, full-sibs can be reliably distinguished (misclassification error of less than 5%) from unrelated individuals with reasonably sized panels of independently segregating microsatellites (i.e. fewer than 20 loci). However, intermediate degrees of relationship are usually present in natural populations and these will be more difficult to differentiate. For example, to discriminate between full-sibs and half-sibs with greater than 90% reliability using Blouin *et al.*'s method would require about 40 independently segregating microsatellites — a number that would be impractical in most cases.

A potential alternative to using microsatellites to resolve pairwise relationships is to employ single nucleotide polymorphisms (SNPs). SNPs have recently proven to be very informative in human genetics (Wang *et al.* 1998) and are rapidly being incorporated into studies of evolutionary biology (Hacia *et al.* 1999; Primmer *et al.* 2002). The fact that 1.42 million SNPs have been placed on a genetic map of the human genome (The International SNP Map Working Group 2001) illustrates the potential abundance of these markers in the genome of any higher organism. Furthermore, recent advances in microarray technology now permit genotyping of large samples of individuals at hundreds or even thousands of SNP loci (reviewed in Kwok 2001).

However, these potential advantages of SNPs are offset by the amount of sequencing effort required for their development (substantial in nonmodel organisms), and by their lack of informativeness per locus. The vast majority of SNPs have only two alleles, and their allele frequencies are usually skewed (Marth *et al.* 2001), with one allele (the 'major allele') much more frequent than the other (the 'minor allele'). This results in low heterozygosity and hence, on a single locus basis, SNPs are considerably inferior to microsatelllites for relatedness and/or relationship estimation. In theory, the paucity of information per SNP locus can be compensated for by increasing the number of loci assayed. However, conventional analytical methods for relationship estimation typically assume independence among the marker loci employed, and the number of independent loci potentially available is limited by the size of the genome — in terms of the numbers of chromosomes and their length (in map units).

A plethora of analytical approaches has been developed for inference of relationships from genetic marker data. As we have noted, most of these assume independent segregation of the employed marker loci. A maximum likelihood framework for the determination of pairwise relationships, in which the likelihood of the marker data is calculated for each of many possible relationships, was first developed by Thompson (1975). An alternative approach has been formulated by Goodnight & Queller (1999), in which a primary relationship hypothesis (or several hypotheses in turn) can be tested against a null hypothesis via a likelihood ratio test. Rather than calculating likelihoods for the observed genotypic data, Blouin *et al.* (1996) first estimated pairwise relatedness and then classified each pair of individuals as either full-sib, half-sib, or unrelated according to predetermined cut-off values for $r$ — Calafell *et al.* (1999) used a similar approach. Likelihood-based methods for partitioning a population into full-sib or half-sib families have recently been improved by moving beyond pairwise analysis to joint likelihood analysis of all individuals using Markov Chain Monte Carlo (MCMC) techniques (Thomas & Hill 2000; Smith *et al.* 2001). Although such joint analyses are more powerful than pairwise analyses, they require prior knowledge that the population in question is in fact solely composed of full-sib (or half-sib) families. Such knowledge is seldom available for natural populations.

The availability of dense genome scan data from studies that seek to map disease-causing loci in humans using sib-pairs, combined with the need to exclude pairs that are not actually full-sibs from such analyses, led to the development of methods for the detection of half-sib and unrelated pairs that account for tight genetic linkage between the markers employed (Boehnke & Cox 1997; Goring & Ott 1997). These multipoint methods have recently been extended by McPeek & Sun (2000) and by Epstein *et al.* (2000) to test for a greater variety of possible relationships. Such multipoint methods are able to extract additional information regarding the genealogical relationship of a pair from the pattern and lengths of shared segments along their genomes (Donnelly 1983; Thompson & Meagher 1998). In the absence of crossover interference, the generation of such identity by descent (*ibd*) patterns can be modelled as a Markov process (Feingold 1993). Simulations by McPeek & Sun (2000) and by Epstein *et al.* (2000) suggest that with very dense genomic scans of thousands of highly informative SNPs [i.e. biallelic markers with a minor allele frequency of 0.3 and 1 cM spacing in McPeek and Sun (2000), or with equifrequent alleles and 4 cM spacing in Epstein *et al.* (2000)], their multipoint methods will be able to determine pairwise relationships with high accuracy.

Although the application and continued development of multipoint methods holds much promise, high-density genomic scans with such highly informative SNPs separated by known recombination distances are only possible in a select few model organisms (e.g. humans, mice,

*Arabidopsis*). Such resources and information are not readily available in nonmodel organisms; however, these are the species that are most often of interest in ecological or evolutionary studies. Hence, in this study we explore the potential for relationship estimation with more modest numbers of SNP loci, under the assumption of independent segregation. Our question is, 'Assuming a realistic number of independent loci, can SNPs equal or exceed the power of practical numbers of microsatellite loci in relationship estimation?' We describe the results of Monte Carlo computer simulations designed to evaluate the efficacy of SNPs in discriminating between a variety of relationships likely to occur in natural populations. In our simulation we examine a 'best-case scenario' in which 100 independently segregating SNPs are available in a hypothetical large genome. Pairwise misclassification rates are determined according to the approach of Blouin *et al.* (1996) and we compare our results to those that they obtained for microsatellite loci.

## Materials and methods

We used Monte Carlo computer simulation to evaluate the efficacy of SNPs for the determination of pairwise relationships. The simulated SNPs were assumed to segregate independently, which limits the number available because of the finite size of the genome of any given higher organism. To assess the prospects under a best-case scenario, we made the rather generous assumption of an ideal genome with 20 autosomes, each 200 cM in length (total genome size of 4000 cM). This allowed us to place five independent SNPs on each autosome, with a 50-cM spacing, so that our panel consisted of 100 SNPs in total. For ease of interpretation, we also assumed that the frequency of the minor allele was constant among the SNP loci. Minor allele frequencies of 0.05, 0.1, 0.2 and 0.5 were simulated separately.

The simulation began by generating a pseudopopulation of 10 000 diploid, sexually reproducing and unrelated individuals. Each individual was assigned 100 SNP loci with genotypes determined at random (i.e. under Hardy–Weinberg equilibrium) in accordance with one of the allele frequencies given above. After this, an array of pedigrees was constructed to allow the formation, via Mendelian inheritance, of noninbred pairs of a variety of relationships. The simulated relationship types fell into five categories according to relationship order (Table 1). The abbreviations that we have used and the expected relatedness for each relationship also are given in Table 1. Example pedigrees that illustrate all of the relationship types examined are shown in Fig. 1. We refer to cousins related by full-sibs on one side and half-sibs on the other as 'almost double-cousins' (ADC; Fig. 1).

Each pedigree was formed by drawing at random (without replacement) the appropriate number of unrelated founder individuals from the pseudopopulation at large, and then mating them in the required configuration. A given pedigree was used only once (i.e. to generate a single pair of individuals of the required relationship). To generate additional pedigrees, either for obtaining pairs of different relationship or for replication, the founders were once again drawn at random from the pseudopopulation, without replacement. Ten million replicate pairs were produced for each relationship type.

Pairwise relatedness at a single locus ($r_l$) between individuals $x$ and $y$ was estimated according to Li *et al.* (1993):

$$r_l = (S_{xy} - U)/(1 - U)$$

**Table 1** Abbreviations and expected relatedness for all relationship types examined

| Order | Relationship type | Abbrev. | $r$* |
|---|---|---|---|
| First-order | parent–offspring | PO | 0.5 |
| | full-sibs | FS | 0.5 |
| Second-order | half-sibs | HS | 0.25 |
| | double-cousins | DC | 0.25 |
| | aunt–nephew | AN | 0.25 |
| | grandparent–grandchild | GG | 0.25 |
| Third-order | cousins | C | 0.125 |
| | double half-cousins | DHC | 0.125 |
| | half aunt–nephew | HAN | 0.125 |
| Fourth-order | half-cousins | HC | 0.0625 |
| Others | almost double-cousins | ADC | 0.1875 |
| | unrelated | U | 0 |

*$r$, expected relatedness.

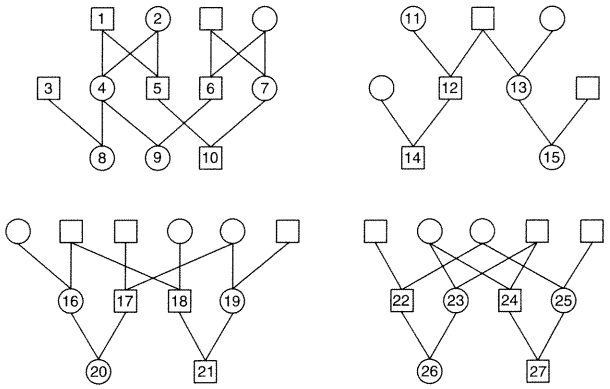


**Fig. 1** Example pedigrees illustrating the various simulated relationship types. PO: e.g. 1 & 4, 11 & 12; FS: 4 & 5, 6 & 7, 23 & 24; HS: 12 & 13, 16 & 18, 17 & 19, 22 & 25; DC: 9 & 10; AN: e.g. 4 & 10, 5 & 8; GG: e.g. 1 & 8, 11 & 14; C: 8 & 10; DHC: 20 & 21; HAN: e.g. 13 & 14, 12 & 15, 17 & 21, 19 & 20; HC: 14 & 15; ADC: 26 & 27; U: e.g. 1 & 2, 1 & 3, 3 & 4. Abbreviations follow those given in Table 1. Note that the aunt–nephew relationship (AN) includes aunt–niece, uncle–nephew, and uncle–niece (likewise for HAN).

where the similarity index $S_{xy} = 1$ when genotype $x = ii$ and genotype $y = ii$, or when $x = ij$ and $y = ij$, $S_{xy} = 0.75$ when $x = ii$ and $y = ij$, or vice versa, $S_{xy} = 0.5$ when $x = ij$ and $y = ik$, and $S_{xy} = 0$ when the two individuals have no alleles in common at the locus. The similarity as a result of chance alone

$$U = 2\sum_{i=1}^{a} p_i^2 - \sum_{i=1}^{a} p_i^3$$

where $p_i$ is the frequency of allele $i$ in the (random mating) population, and $a$ is the number of alleles at the locus. Overall $r$ was computed as an average across all loci.

The above Li *et al.* (1993) estimator is a slightly modified, improved version of Lynch's (1988) estimator. A number of other, more complicated, estimators of relatedness are available (Queller & Goodnight 1989; Ritland 1996; Lynch & Ritland 1999; Wang 2002). However, the Li *et al.* (1993) estimator is relatively simple to compute and has been shown by simulation to perform as well, in terms of accuracy and precision, as any of the others under a variety of realistic conditions (Van de Casteele *et al.* 2001; Wang 2002). Furthermore, three of the alternative estimators (but not the Wang (2002) estimator) would have yielded undefined estimates of $r$ under some of the conditions of the simulation (e.g. allele frequency of 0.5; Wang 2002).

To evaluate the performance of the 100 simulated SNPs in the classification of pairs of individuals by relationship, we quantified the misclassification rate for selected, single-generation relationships [full-sibs (FS), half-sibs (HS), cousins (C) and unrelated (U)] following the approach of Blouin *et al.* (1996). This allowed us to compare the performance of SNPs in our simulation directly with the results that Blouin *et al.* (1996) obtained with microsatellites. As in Blouin *et al.* (1996), the probability that a pair of one relationship type (e.g. FS) would be misclassified as that of another (e.g. HS) was determined using a cut-off set at the midpoint between the expected values of $r$ for the two relationship types under consideration. For example, for FS (expected $r$ of 0.5) and HS (expected $r$ of 0.25) comparisons, the cut-off was set at 0.375. A FS pair was considered misclassified as HS if the $r$ estimate for the pair was less than 0.375 and an HS pair was considered misclassified as FS if $r$ was greater than 0.375. Pairs with $r$-values that fell exactly on the cut-off were considered misclassified with probability 1/2.

In addition to having an expected value of $r$ of 0.5, the parent–offspring relationship type can be further distinguished from others on the basis that, barring mutation, parent–offspring pairs must share at least one allele in common at every locus. Hence, as in Blouin *et al.* (1996), the percentage of pairs of each relationship type that were not excluded from being parent–offspring was determined by simply counting the number of pairs that met this criterion.

Besides simulating co-dominant, biallelic markers (i.e. SNPs), we also simulated the ideal case of an infinite number of alleles at each locus, or pure identity by descent (*ibd*), in which all of the alleles in all of the founder individuals were unique in state. This was done for three reasons: (i) to illustrate the inherent variance in $r$ particular to each type of relationship (referred to by Wang (2002) as the 'original variance' of $r$), (ii) to verify our simulations, and (iii) to demonstrate the difficulty, even with impossibly ideal genetic markers, of distinguishing certain relationship classes from one another.

## Results

The results from the simulation of pure *ibd* (i.e. an infinite number of alleles per locus) at 100 independent loci are shown in Fig. 2, illustrating that the different pairwise relationships can have different amounts of inherent variance in relatedness. The parent–offspring (PO) and U

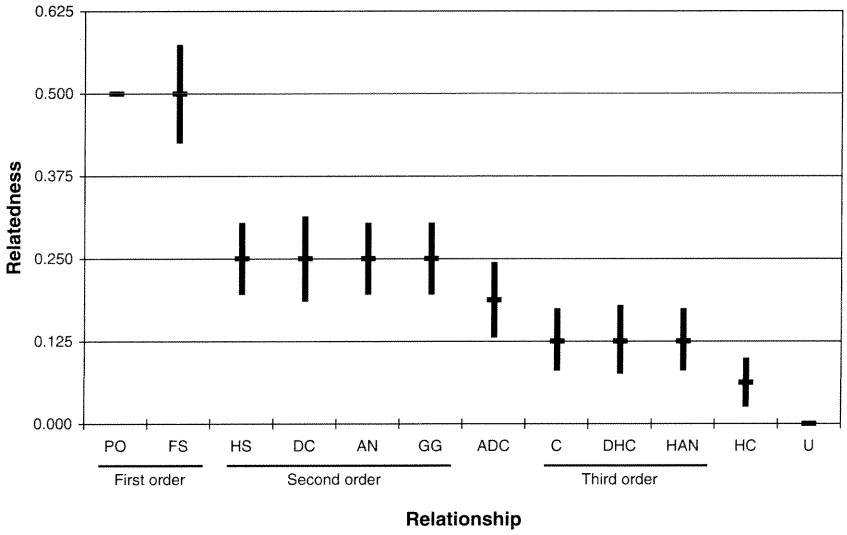


**Fig. 2** The inherent variation in pairwise relatedness at 100 hypothetical, independent loci under pure identity by descent (i.e. with a near-infinite number of alleles at each locus). Shown are the expected values and 95% probability density intervals of relatedness for a variety of noninbred relationships, based upon 10 million simulated pairs of each relationship type. The abbreviations for the various relationships follow those in Table 1.

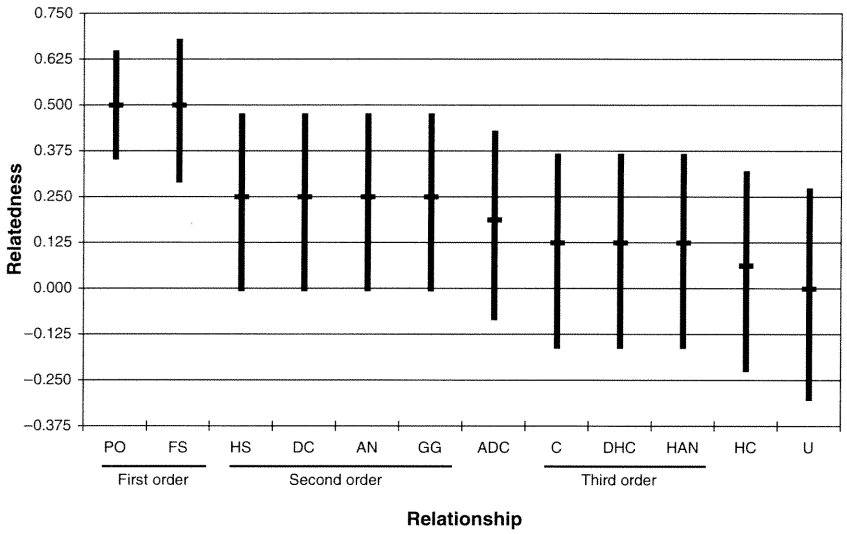


**Fig. 3** Expected values and 95% probability density intervals for pairwise relatedness at 100 independent, biallelic SNP loci each with minor allele frequency of 0.2.

relationships have zero inherent variance, because such pairs always share one or no alleles *ibd*, respectively. The FS relationship has the largest amount of inherent variance (0.00125 with 100 loci), and the double-cousins (DC) relationship has the second largest (0.000937). The inherent variances of HS, aunt–nephew (AN) and grandparent–grandchild (GG) were essentially identical (0.000625), as were those of C and half-aunt–nephew (HAN; 0.000469). The inherent variances obtained from our simulation for FS and HS matched their theoretical expectations ($^1/_8$ per locus and $^1/_{16}$ per locus, respectively; provided by Wang 2002), confirming that our simulations were accurate.

It is also clear from Fig. 2 that the second-order relatives (HS, DC, AN and GG) cannot be distinguished from each other via estimation of $r$ with independently segregating loci, as all of these have the same expected relatedness ($^1/_4$) and similar or identical variances. The same goes for the third-order relatives [C, double half-cousins (DHC) and HAN], which have a common expected relatedness of $^1/_8$. Furthermore, even under pure *ibd*, distinguishing between relationship orders would not always be possible: third-order relatives (C, DHC, or HAN) are not fully resolved from fourth-order relatives (such as HC), as their 95% probability density intervals (PDIs) overlap considerably (Fig. 2). Likewise, the 'almost double-cousins' relation (ADC) would not be clearly resolved from either the second- or the third-order relatives. However, with 100 'perfect' loci (i.e. with infinite alleles), the PO relationship would be distinguishable from FS, as the probability of the occurrence of an FS pair with exactly one allele *ibd* at every locus would be infinitesimal ($0.5^{100}$).

As for biallelic markers such as SNPs, even under the optimistic scenario of having one hundred independent loci available (all with a known minor allele frequency of 0.2), our simulation results indicate little promise for discrimination among higher order relationships on the basis of estimates of $r$ (Fig. 3). The lack of resolution among relationship categories is apparent from the overlap between the 95% PDIs of the second-order relatives (HS, DC, AN, GG) with those of the third- or fourth-order relatives (e.g. C, DHC, HAN, or HC). As the 95% PDIs of the second-order relatives also overlap considerably with that of unrelated pairs (U), these classes would also be poorly resolved. Furthermore, even full-sibs and second-order relatives would not be clearly distinguished.

This inability to distinguish among many of the relationships using SNPs is further illustrated by the pairwise misclassification rates (Fig. 4). These are shown for selected comparisons within a single generation, and for the four minor allele frequencies simulated. For instance, with a minor allele frequency of 0.2 there would be a greater than 10% misclassification of FS as HS or vice versa. The acceptable misclassification rate for discrimination among relationship types is likely to vary, depending on the research question. For the purposes of this study, however, we have selected a misclassification rate of $\leq 10\%$ as acceptable. Hence, of the relationship type combinations shown in Fig. 4, only FS–C or FS–U and their reciprocals would be well resolved (misclassification rate of less than 10%) if the minor allele frequency was 0.2. As the frequency of the minor allele decreases across loci, the resolving power deteriorates further. For example, with a minor allele frequency ($q$) of 0.1, cousins are no longer resolved from full-sibs, and when $q$ is 0.05, only full-sibs can be distinguished from unrelated pairs (but not vice versa) with greater than 90% power (Fig. 4).

For our purposes, a 'perfect' panel of biallelic SNPs would consist of one hundred independent loci, each with equifrequent alleles ($q = 0.5$). Even in such an impossibly ideal case the second-order relatives (e.g. HS) would still not be resolved from the third-order relatives (e.g. C), or

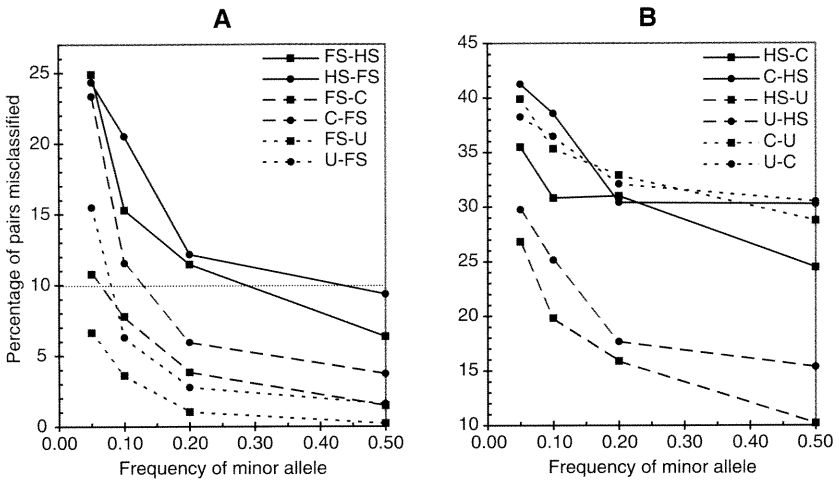


**Fig. 4** Reciprocal misclassification rates for pairwise comparisons among selected relationship types using 100 independent SNPs. For each comparison, simulations were conducted at four different minor allele frequencies. The abbreviations for the various relationships follow those in Table 1. For example, 'FS–HS' refers to 'full-sibs misclassified as half-sibs' while 'HS–FS' refers to 'half-sibs misclassified as full-sibs'. (A) Comparisons involving full-sibs; (B) all other comparisons.

even from unrelated pairs (Fig. 4). However, full-sibs would be distinguished from all other relationship types more than 90% of the time.

The picture is somewhat less bleak in regard to detecting parent–offspring (PO) pairs (parentage analysis). The PO relationship can be excluded if there are no alleles in common between the pair in question at one or more loci (barring mutation and mistyping). Here our 'perfect' panel of 100 biallelic SNPs with $q = 0.5$ would perform quite well, as FS pairs would be misclassified as PO less than 5% of the time, and all higher order relatives would be misclassified as PO less than 1% of the time (Table 2). However, with the somewhat more realistic, but 'best-case' scenario of 100 independent SNPs all with $q = 0.2$, FS pairs would be misclassified as PO more than 27% of the time, and HS pairs would be misclassified as PO more than 7% of the time. With lower minor allele frequencies (e.g. $q = 0.1$ or 0.05) the performance of 100 independent SNPs in parentage analysis remains dismal (Table 2).

In cases where matings occur only between individuals in the same generation, where it is possible to distinguish among the generations by age class, and where age estimates are available for all individuals of interest, then FS or HS pairs within the same age class can be excluded *a priori* from being PO (as could pairs spanning three or more generations). Here the relationship types that one would hope to be able to exclude with marker information would be AN, HAN and U. With the best-case scenario of 100 independent SNPs all with $q = 0.2$, these would be misclassified as PO less than 8%, 2% and 0.6% of the time, respectively (Table 2). The confounding effect of the presence of aunts and uncles on parentage analysis has been investigated by Olsen *et al.* (2001).

The Li *et al.* (1993) estimator of $r$ proved to be completely free of bias under the conditions of this simulation. The mean estimate of $r$ for each relationship type simulated fell almost exactly upon its theoretical value for all allele frequencies (e.g. see Fig. 3 for $q = 0.2$). This is despite the fact that the Li *et al.* (1993) estimator is much simpler to compute than all of the alternative estimators.

**Table 2** Percentage of pairs not excluded as parent–offspring for all relationships

| | Relationship† | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| q* | FS | HS | DC | AN | GG | ADC | C | DHC | HAN | HC | U |
| 0.50 | 4.18 | 0.16 | 0.07 | 0.16 | 0.16 | 0.02 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 |
| 0.20 | 27.5 | 7.42 | 5.33 | 7.41 | 7.46 | 3.25 | 1.96 | 1.81 | 1.97 | 1.01 | 0.51 |
| 0.10 | 66.6 | 44.3 | 40.0 | 44.3 | 44.2 | 34.3 | 29.4 | 28.6 | 29.4 | 23.9 | 19.3 |
| 0.05 | 89.3 | 79.7 | 77.5 | 79.7 | 79.3 | 74.2 | 71.1 | 70.6 | 71.1 | 67.2 | 62.4 |

*q, frequency of the minor allele.
†Abbreviations for each relationship type defined in Table 1.

## Discussion

Coupled with recent advances in microarray technology, the sheer abundance of SNPs promises to revolutionize many aspects of evolutionary genetics (e.g. the study of nucleotide diversity, recombination rates, linkage disequilibrium, phylogeographic history, etc.). However, our results suggest that SNPs have limited potential for the delineation of immediate genealogical relationships, particularly when independence among the sampled loci is assumed. We simulated a 'best-case scenario' in which it was assumed that 100 independent SNPs each with a minor allele frequency of 0.2 are available in the organism of interest. Under this scenario, about 18% of unrelated individuals would be misclassified as half-sibs. Comparing with Fig. 4 in Blouin *et al.* (1996), this corresponds to the misclassification rate that would be achieved with about 20 microsatellites, each with an expected heterozygosity ($H_E$) of 0.75. For parentage analysis by exclusion (both parents unknown), the expected misclassification rate of full-sibs as parent–offspring for the 100 SNPs is

27.5%, which corresponds to estimates obtained with about 16 microsatellites, each with an $H_E$ of 0.75 (see Fig. 4 in Blouin *et al.* 1996). Hence 16–20 'typical' microsatellites (DeWoody & Avise 2000) would be expected to provide information equivalent to that given by 100 independent SNPs each with a minor allele frequency of 0.2.

However, these comparisons assume that the microsatellites themselves are independent (i.e. not within 50 cM of one another), implying that sufficient numbers of microsatellites have been placed on a linkage map of the organism in question to allow for selection of 'well-spaced' markers. In the absence of a linkage map, the number of microsatellite (or SNP) loci scored must be increased to compensate for the loss of information as a result of nonindependence (i.e. linkage) between some of the markers–a consideration that must be kept in mind when contemplating the use of microsatellites and/or SNPs for relationship estimation or parentage analysis. Testing for linkage disequilibrium in the population of interest (e.g. Blouin *et al.* 1996) usually will not suffice to rule out linkage; for example, in a human population of northern European descent, linkage disequilibrium has been shown to typically extend 60 kb in either direction from common 'core' alleles (Reich *et al.* 2001), corresponding to a genomic segment of less than one-thousandth the length of a typical human chromosome. Hence it is highly feasible that closely linked markers that do not segregate independently will not show any linkage disequilibrium at the population level.

Our assumption of equal minor allele frequencies among all of the utilized SNPs, though unrealistic, was made to simplify interpretation of the simulation results. The fact that over 50% of the more than 1200 human SNPs surveyed by Marth *et al.* (2001) had a minor allele frequency of 0.2 or higher in any given human population (African, Caucasian, or Chinese) suggests that our best case scenario of $q = 0.2$ will often be easily attained or exceeded in practice, especially when many thousands of SNPs are available to choose from in the organism of interest. Having a *minimum* minor allele frequency of 0.2 would lead to improved resolution of relationship types, particularly if loci are weighted according to their information content (Wang *et al.* 2002). On the other hand, the genome size of 4000 cM that would be required to permit the presence of 100 completely independent SNPs is larger than that of most higher organisms: the human, rat, mouse and *Drosophila pseudoobscura* genomes are roughly 3700, 2400, 1600 and 450 cM in length, respectively, and mapped plant genomes range from roughly 500 to 2700 cM (O'Brien 1993). In most organisms the maximum number of completely independent SNPs will be far less than one hundred. Hence, our simulation of 100 loci each with $q = 0.2$ remains an optimistic scenario.

If the prospect that we have painted for the use of independently segregating SNPs for relationship determination in ideal outbreeding populations appears bleak, the picture only becomes darker once additional realism is taken into account. Deviation from Hardy–Weinberg proportions, selection at closely linked loci, genotyping errors, mutation and recent inbreeding (because of small population size and/or the mating system) would all cause further degradation in our ability to distinguish among relationship types. Recent inbreeding will result in elevated $r$, so that, for example, a pair that are half-sibs in their primary relationship, but that share additional prior common ancestors, will be more likely to be mistaken as full-sibs (e.g. Calafell *et al.* 1999). Seen from another point of view, the presence of inbreeding, in combination with matings between generations, leads to a whole continuum of new relationship categories, most of which are yet unnamed. Hence the practice of relationship determination is best restricted to large (or moderately large), strictly outbreeding populations that have not been subjected to a recent bottleneck.

One category of first-order relationships, PO, is of special interest with regard to paternity and/or maternity analyses where one parent is known with certainty. In such analyses, panels of fewer than six microsatellite loci can provide average exclusion probabilities greater than 0.99 (e.g. Jones & Avise 1997; DeWoody *et al.* 2000). If we use the approach of Jamieson & Taylor (1997; equation 4) to determine the number of loci required to achieve a given exclusion probability when one parent is known, we find that at least 32 independently segregating SNPs (each with a minor allele frequency of 0.2 or greater) are required to achieve an exclusion probability of greater than 0.99 (Table 3). Indeed, a real-life example has already been provided in beef cattle by Heaton *et al.* (2002), where an exclusion probability of 99.4% for paternity analysis was estimated from the allele frequencies of 32 SNPs in the American Angus breed.

An alternative to using independently segregating SNPs would be to use independently segregating haplotypes, with each haplotype defined by a cluster of tightly linked SNPs. For example, to improve further the potential power

**Table 3** Number of independently segregating biallelic loci required to achieve specified exclusion probabilities in paternity analysis

| | Desired exclusion probability | | | |
|---|---|---|---|---|
| q* | 0.95 | 0.99 | 0.995 | 0.999 |
| 0.50 | 14.4 | 22.2 | 25.5 | 33.3 |
| 0.25 | 18.1 | 27.9 | 32.1 | 41.8 |
| 0.20 | 20.8 | 31.9 | 36.7 | 47.9 |
| 0.15 | 25.4 | 39.0 | 44.9 | 58.6 |
| 0.10 | 35.1 | 53.9 | 62.0 | 80.8 |
| 0.05 | 64.7 | 99.5 | 114 | 149 |
| 0.01 | 304 | 468 | 538 | 701 |

*q, frequency of the minor allele.

of paternity analyses in cattle, Heaton *et al.* (2002) obtained the DNA sequences surrounding their 32 SNPs, and uncovered an additional 183 polymorphic sites. Such an approach would be very likely to increase the number of alleles (haplotypes) available, resulting in enhanced resolving power for relationship determination. However, haplotype inference in outbred diploid populations (e.g. Excoffier & Slatkin 1995) may prove problematic, unless levels of fine-scale linkage disequilibrium prove to be high (Xu *et al.* 2002). The finding that linkage disequilibrium in a human population typically extends 60 kb in either direction (Reich *et al.* 2001) bodes well for haplotype inference based upon clusters of very tightly linked SNPs. However, for the estimation of $r$ and for the determination of a pairwise relationship, this is a double-edged sword, as higher levels of linkage disequilibrium decrease the number of common haplotypes present in a segment. When linkage disequilibrium among the tightly linked SNPs is at its maximum, the number of haplotypes reduces to two, and we are back to where we started!

We have focused on the use of independently segregating markers (SNPs or haplotype clusters), deliberately excluding consideration of loci that are linked at distances less than 50 cM. The use of more closely linked markers that do not segregate independently will not adversely affect the accuracy of relatedness estimates *per se*, but instead will affect estimates of the variance of $r$. The variance of an $r$ estimate will be underestimated if independent segregation is assumed when some of the loci are in fact linked. In this case, progressive addition of loci — beyond the maximum number that can segregate independently — will yield diminishing returns in terms of improvement in the power to resolve relationships.

To conclude, it appears that realistic numbers of independently segregating SNPs can be useful in determining parent–offspring pairs, but generally no more so than about 16 or fewer microsatellites when both parents are unknown. To take full advantage of the desirable characteristics of SNPs (i.e. their vast abundance in the genomes of most higher organisms and their potential automation) in relationship estimation, it will be necessary to utilize analytical methods that account for tight genetic linkage (McPeek and Sun 2000; Epstein *et al.* 2000). However, because thousands of highly informative SNPs (with minor allele frequencies of at least 0.3) separated by known recombination frequencies are needed to exploit the full power of these multipoint methods, it seems that such studies will only be feasible, in the near future at least, in a handful of model organisms.

## Acknowledgements

We thank T. Belfiore, Y. DeWoody, A. Fiumera, P. Waser and members of the DeWoody laboratory for helpful comments on the manuscript. This is ARP publication #16866 from Purdue University.

## References

Avise JC, Shapiro DY (1986) Evaluating kinship of newly settled juveniles within social groups of the coral reef fish *Anthias squamipinnis*. *Evolution*, **40**, 1051–1059.

Blouin MS, Parsons M, LaCaille V, Lotz S (1996) Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology*, **5**, 393–401.

Boehnke M, Cox NJ (1997) Accurate inference of relationships in sib-pair linkage studies. *American Journal of Human Genetics*, **61**, 423–429.

Calafell F, Shuster A, Speed WC *et al.* (1999) Genealogy reconstruction from short tandem repeat genotypes in an Amazonian population. *American Journal of Physical Anthropology*, **108**, 137–146.

DeWoody JA, Avise JC (2000) Microsatellite variation in animal populations, with special emphasis on marine, freshwater, and anadromous fishes. *Journal of Fish Biology*, **56**, 461–473.

DeWoody JA, Walker D, Avise JC (2000) Genetic parentage in large half-sib clutches: theoretical estimates and empirical appraisals. *Genetics*, **154**, 1907–1912.

Donnelly KP (1983) The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology*, **23**, 34–63.

Engh AL, Funk SM, van Horn RC *et al.* (2002) Reproductive skew among males in a female-dominated society. *Behavioral Ecology*, **13**, 193–200.

Epstein MP, Duren WL, Boehnke M (2000) Improved inference of relationship for pairs of individuals. *American Journal of Human Genetics*, **67**, 1219–1231.

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, **12**, 921–927.

Feingold E (1993) Markov processes for modeling and analyzing a new genetic mapping method. *Journal of Applied Probability*, **30**, 766–779.

Goodisman MAD, Crozier RH (2002) Population and colony genetic structure of the primitive termite *Mastotermes darwiniensis*. *Evolution*, **56**, 70–83.

Goodnight KF, Queller DC (1999) Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology*, **8**, 1231–1234.

Goring HHH, Ott J (1997) Relationship estimation in affected sib pair analysis of late-onset diseases. *European Journal of Human Genetics*, **5**, 69–77.

Hacia JG, Fan J-B, Ryder O *et al.* (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genetics*, **22**, 164–167.

Hamilton WD (1964) The genetic evolution of social behaviour, I and II. *Journal of Theoretical Biology*, **7**, 1–52.

Heaton MP, Harhay GP, Bennett GL *et al.* (2002) Selection and use of SNP markers for animal identification and paternity analysis in US beef cattle. *Mammalian Genome*, **13**, 272–281.

Heg D, van Treuren R (1998) Female–female cooperation in polygynous oystercatchers. *Nature*, **391**, 687–691.

Jamieson A, Taylor StCS (1997) Comparisons of three probability formulae for parentage exclusion. *Animal Genetics*, **28**, 397–400.

Jones AG, Avise JC (1997) Microsatellite analysis of maternity and the mating system in the Gulf pipefish *Syngnathus scovelli*, a species with male pregnancy and sex-role reversal. *Molecular Ecology*, **6**, 203–213.

Jones KL, Glenn TC, Lacy RC *et al.* (2002) Refining the Whooping Crane studbook by incorporating microsatellite DNA and leg-banding analyses. *Conservation Biology*, **16**, 789–799.

Kwok P-Y (2001) Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics*, **2**, 235–258.

Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity*, **43**, 45–52.

Lynch M (1988) Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution*, **5**, 584–599.

Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. *Genetics*, **152**, 1753–1766.

McPeek MS, Sun L (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. *American Journal of Human Genetics*, **66**, 1076–1094.

Marth G, Yeh R, Minton M *et al.* (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genetics*, **27**, 371–372.

Morin PA, Moore JJ, Chakraborty R *et al.* (1994) Kin selection, social-structure, gene flow, and the evolution of chimpanzees. *Science*, **265**, 1193–1201.

O'Brien SJ (1993) *Genetic Maps*, 6th edn. National Institutes of Health, Bethesda, MD.

Olsen JB, Busack C, Britt J, Bentzen P (2001) The aunt and uncle effect: an empirical evaluation of the confounding influence of full sibs of parents on pedigree reconstruction. *Journal of Heredity*, **92**, 243–247.

Pamilo P, Crozier RH (1982) Measuring genetic relatedness in natural populations — methodology. *Theoretical Population Biology*, **21**, 171–193.

Primmer CR, Borge T, Lindell J, Saetre GP (2002) Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Molecular Ecology*, **11**, 603–612.

Queller DC, Goodnight KF (1989) Estimating relatedness using molecular markers. *Evolution*, **43**, 258–275.

Reich DE, Cargill M, Bolk S *et al.* (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, **67**, 175–185.

Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology*, **9**, 1195–1204.

Smith BR, Herbinger CM, Merry HR (2001) Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, **158**, 1329–1338.

Taylor AC, Horsup A, Johnson CN, Sunnucks P, Sherwin B (1997) Relatedness structure detected by microsatellite analysis and attempted pedigree reconstruction in an endangered marsupial, the northern hairy-nosed wombat *Lasiorhinus krefftii*. *Molecular Ecology*, **6**, 9–19.

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

Thomas SC, Hill WG (2000) Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, **155**, 1961–1972.

Thompson EA (1975) The estimation of pairwise relationships. *Annals of Human Genetics*, **39**, 173–188.

Thompson EA, Meagher TR (1998) Genetic linkage in the estimation of pairwise relationship. *Theoretical and Applied Genetics*, **97**, 857–864.

Van de Casteele T, Galbusera P, Matthysen E (2001) A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology*, **10**, 1539–1549.

Wang DG, Fan JB, Siao C-J *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.

Wang J (2002) An estimator of pairwise relatedness using molecular markers. *Genetics*, **160**, 1203–1215.

Xu CF, Lewis K, Cantone KL *et al.* (2002) Effectiveness of computational methods in haplotype prediction. *Human Genetics*, **110**, 148–156.

---

J. Glaubitz is a postdoctoral researcher, while O. E. (Gene) Rhodes and Andrew DeWoody are faculty members at Purdue University. The authors share a common interest in the application of molecular markers and computer simulation to conservation and population genetics.

---