DNA-based methods for pedigree reconstruction and kinship analysis in natural populations

Michael S. Blouin

Department of Zoology, Oregon State University, Corvallis, OR 97331-2914, USA

The widespread use of microsatellite loci has spurred the recent development of many new statistical methods for inferring kin relationships from molecular data. We now have an unprecedented ability to infer detailed genealogical information about individuals in natural populations, but the best approach for a given problem is not always obvious. Researchers in different fields have also been deriving similar methods independently. Thus, some biologists might not be aware of what is even possible. By adopting these new methods, researchers in ecology and evolution could extract far more pedigree information from natural populations than is currently being exploited.

The ability to infer genealogical relationships among individuals in a population has opened up many areas of research in behaviour, evolution and conservation. Examples include estimating heritabilities in the wild [1–4], minimizing inbreeding in captive populations [5–7], estimating rates of gene flow into a population [8,9], adjusting population allele frequency estimates for the presence of relatives in a sample [10–13], estimating the total number of breeders in a population [14–16] and estimating variance in reproductive success among individuals, which can be used to study selection and estimate effective population sizes [17–20]. Nevertheless, a bewildering array of statistical methods for molecular-based kinship analysis is now available, and choosing the best tool for a particular job can be confusing. Researchers in different fields (e.g. evolution, animal breeding, human genetics and forensics) have been independently deriving similar methods, and so far there has been little effort to bring them together. Many researchers are familiar with parentage analysis (e.g. paternity testing) but not with the other statistical methods for inferring familial relationships in the absence of parentage data. More importantly, they might not be aware of the unique questions that can be asked using some of these other methods. Here, I provide a guide to those other methods, with an emphasis on those that are not yet in widespread use by students of ecology and evolution. My goals are to introduce the logic behind each technique, to highlight interesting applications and to provide practical advice about their use.

Methods of kinship analysis can be divided into two categories: RELATEDNESS (see Glossary) estimation and assignment of pairs or groups of individuals to categories of relationship. Relatedness ($r$) is a continuous measure of overall identity by descent (IBD) between individuals, whereas RELATIONSHIP CATEGORIES are specific pedigree (genealogical) relations, such as full sibs or first cousins (Box 1). Parentage analysis is a unique application in which one searches among candidates for the most likely parents of a target offspring. There are so many variations on basic parentage analysis that it warrants separate treatment and will not be covered here (see [21] for a recent review).

Relatedness estimators

Estimators of $r$ are useful as correlates of genome-wide IBD between individuals (Box 1). For example, one can estimate heritabilities of traits by regressing pairwise estimates of phenotypic similarity against $r$ [3], or one could minimize inbreeding in a captive population by choosing mates based on $r$ [7,22,23] (Box 2). The ability to

Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele-sharing test</td>
<td>a measure of the fraction or total number of alleles shared (identical by state) between two individuals</td>
</tr>
</tbody>
</table>
Box 1. Identity by descent, relationship categories and relatedness

Identity by descent
Alleles are identical by descent if they recently descended from a single ancestral allele. Because all alleles are identical by descent if you look back far enough, recently means within a particular reference population, usually going back just a few generations [65]. Two alleles are identical by state if they have the same allelic state. Alleles that are identical by state might not be identical by descent if they coalesce farther back than the reference pedigree or arose independently via mutation. In practice, we can only score identity by state and must infer probabilities of identity by descent.

Categories of relationship and IBD coefficients
Categories of relationship refer to particular pedigree categories, such as full sibs or half sibs. Avuncular refers to any of the four categories involving aunts or uncles with nieces or nephews. The categories parent–offspring and full sib are collectively referred to as first degree (1st) relatives (50% of alleles shared identical by descent, on average), the categories grandparent–grandoffspring, half sibs, and avuncular as second degree (2nd) (average 25% shared), the categories first cousins and great grandparent–great grandoffspring as third degree (3rd) (12.5% shared), and so on. The probabilities that a dyad of a particular relationship shares 0, 1 or 2 alleles that are identical by descent at any locus are summarized by a three-parameter set of IBD coefficients \(k_0, k_1, k_2\), sometimes called \(k\) coefficients [38]. Most of the common relationship categories have different expected IBD coefficients (Table I).

Relatedness coefficients
The coefficient of consanguinity (also coefficient of kinship or of co-ancestry) between individuals \(I\) and \(J\), \(f_{IJ}\), is the probability that two alleles, one chosen randomly from each individual, are identical by descent. If those two individuals could reproduce, then \(f_{IJ}\) would be the inbreeding coefficient of their offspring. The relatedness between two individuals, \(r\), (also coefficient of relatedness or of relationship) can be interpreted as the expected fraction of alleles that are shared identical by descent [66]. More formally, \(r\) is the genetic similarity between two individuals relative to that between random individuals from some reference population [66]. Thus, \(r\) is the correlation or regression of genetic values of individuals, and so is usually of more interest than \(f_{IJ}\) because of its central place in quantitative genetics and kin selection theory [66,67]. Note that \(r\) need not be symmetrical between two individuals according to the IBD coefficients.

\[ \begin{align*}
&\text{Table I. Identity by descent coefficients } \{k_0, k_1, k_2\} \text{ and relatedness, } r \text{, for some common relationship categories} \\
&\text{Relationship category} \quad k_0 \quad k_1 \quad k_2 \quad r \\
&\text{Monozygotic twins or self} \quad 0 \quad 0 \quad 1 \quad 1 \\
&\text{Parent-offspring} \quad 0 \quad 1 \quad 0 \quad 0.50 \\
&\text{Full sibs} \quad 0.25 \quad 0.50 \quad 0.25 \quad 0.50 \\
&\text{2\textsuperscript{*}} \text{ (e.g. half sibs, avuncular)} \quad 0.50 \quad 0.50 \quad 0 \quad 0.25 \\
&\text{3\textsuperscript{*}} \text{ (e.g. first cousins)} \quad 0.75 \quad 0.25 \quad 0 \quad 0.125 \\
&\text{Unrelated} \quad 1 \quad 0 \quad 0 \quad 0 \\
\end{align*} \]

Fig. I. Why full sibs have identity by descent coefficients \(k_0 = 0.25, k_1 = 0.5, k_2 = 0.25\) and relatedness \(r = 0.5\). The genotypes of the parents (A and B) at a locus are 12 and 34 (where alleles 1–4 are unique by descent (a), so each offspring (C and D) can have one of four genotypes, 13, 14, 23 or 24. Out of the 16 ways to pair two offspring, the dyad can share two alleles that are identical by descent in four ways, one allele in eight ways and 0 alleles in four ways (b). Thus, \(k_0 = 0.25, k_1 = 0.5, k_2 = 0.25\). On average, a pair of siblings shares one out of two alleles identical by descent, which gives \(r = 0.5\).

estimate \(r\) between interacting individuals is very useful in the study of kin selection [24,25]. Relatedness can also be used to assign pairs (dyads) to relationship categories [26,27], but there are better ways to do this.

Several estimators of \(r\) have been proposed, and their relative precision and accuracy depends on allele-frequency distributions and the true relationship [28–30]. Wang’s [30] modification of Li et al.’s [31] similarity index appears to have the most desirable properties, including: (1) low sensitivity to the sampling error that results from estimation of population allele frequencies; and (2) a low sampling variance that decreases asymptotically to the theoretical minimum with increasing numbers of loci and alleles per locus. Lynch and Ritland’s [29] and Queller and Goodnight’s [32] estimators also perform well, although the Lynch–Ritland estimator can have some undesirable properties when loci are highly polymorphic and true \(r\) is high [30]. The original Queller–Goodnight estimator is undefined for heterozygotes at bi-allelic loci. This is not true for its implementation in the RELATEDNESS computer programme (Table I), in which heterozygotes are assigned a value of 1 at bi-allelic loci.

All relatedness estimators have very large variances owing to stochastic differences in true IBD among loci and to the chance sharing of alleles that are identical by state. Tens of microsatellite loci (e.g. 30–40) or three to four times that many single nucleotide polymorphism (SNP) loci are needed to obtain even moderate confidence around a single pairwise estimate (standard deviations of, e.g. 0.1) [26,27,33]. In the absence of enough loci to accurately estimate \(r\) for individual pairs, one might still be able to estimate the average relatedness within groups with reasonable accuracy [32]. With relatively few loci, one can also ask a different type of question. Here, one assumes that the group includes individuals of two or more relationship categories. The goal is to estimate the fraction of each type of category comprising the group, but without worrying which pairs belong to each category. The distribution of all pairwise \(r\) in a group is modeled as a mixture of several underlying distributions, and the fraction of
Box 2. Case studies

Correlation between allele-sharing and true identity by descent in an inbred pedigree

In many captive breeding situations (e.g. livestock breeding or wildlife conservation), it is essential to control the rate of inbreeding in populations that are descended from a few founders. Thoroughbred horses represent an essentially closed population that is now highly inbred, and for which there exist detailed pedigree records [23]. Cunningham et al. [23] genotyped 211 thoroughbreds at 13 microsatellite loci and regressed the proportion of alleles shared between pairs of individuals, AS, on the coefficient of co-ancestry, \( f_{ij} \) (Box 1), estimated from the pedigree. The equation for the line was \( AS = 0.309 + (1.017)f_{ij} \), where \( f_{ij} \) explained 98% of the variance in AS (the intercept can be interpreted as the background allele sharing in the founders of the population). This almost perfect, one-to-one relationship shows that even a simple allele-sharing statistic estimated from 13 loci captured most of the information about pairwise identity by descent in a complex, inbred pedigree.

The relationship categories comprising a group can be inferred from the distribution of pair-wise \( r \) estimates

Colonies of the social wasp Polistes dominulus are founded by multiple females, and one foundress assumes complete reproductive dominance over the others. The nonreproductive, helping behavior of the other foundresses was assumed to result from kin selection among closely related foundresses. However, using seven microsatellite loci, Queller et al. [34] estimated that the distribution of pairwise \( r \) among nestmate foundresses was composed of 35% unrelated, 9% cousin and 56% full sister dyads (Figure I). This result re-jects kin selection as the sole explanation for non-reproductive helping behavior among subordinate foundresses.

Use of estimated IBD coefficients and likelihood tests of relationship category

For linkage analysis, one begins with a pedigree that is assumed to be correct. Human pedigrees often contain errors owing to, for example, mis-specified paternity or mis-handled samples (e.g. duplicates or switched identities). To error-check pedigrees, all the individuals in the pedigree are genotyped and all the putative (null) pairwise relationships specified by the pedigree are tested by likelihood or allele-sharing methods. McPeek and Sun [37] discuss the interesting example in Figure IIa. Here, not all the individuals could be genotyped. All testable pairwise relationships were consistent with expectations except for the expected first-cousin relationship of individual 18 with individuals 14 and 15. The expected identity by descent (IBD) coefficients for a first cousin pair are (0.75, 0.25, 0.0) (Box 1). The maximum likelihood estimates of the IBD coefficients between 18 and 14 were (0.28, 0.56, 0.16), and between 18 and 15 were (0.27, 0.57, 0.16). These values are between those expected for half and full sibs (Box 1). There is no misfit between 18 and his half sib 19, or between 14, 15 or 18 and their avuncular relatives. One plausible explanation is that individuals 5 and 10 are actually the same person (i.e. 5 is also the father of 18). In that case, the relationship of 18 to 14 or 15 is that of half sib plus first cousin, as illustrated in Figure IIb.

Fig. I. Observed distribution of pairwise \( r \) estimates among Polistes dominulus foundresses (filled squares) and expected distributions for three other plausible relationship categories (open symbols). Values are grouped into intervals of width 0.1. The expected distributions were obtained via simulation. Open squares = unrelated (true \( r = 0 \)), open circles = cousins (true \( r = 3/16 \)), open triangles = full sisters (true \( r = 3/4 \)). True \( r \) values for cousins and sisters are higher than shown in Box 1 because wasps are haplodiploid. Reproduced, with permission, from [34].

Fig. II. Test of putative relationships in a human pedigree. (a) Putative pedigree of individuals in the case study discussed in [37]. Shaded individuals were not available for genotyping. All others were scored for at least 250 microsatellite loci. The null relationship of cousins was strongly rejected for individuals 18 and 15, and for 18 and 14, owing to excess allele sharing. (b) Pedigree showing a plausible explanation for the excess allele sharing between 18 and his putative cousins (pedigree condensed to show only the relevant individuals). Here individual 5 is hypothesized to be the true father of all three children. Reproduced, with permission, from [37].
Table 1. Software for implementing methods discussed in the text

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
<th>Comments/Limitations</th>
<th>Web site</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relatedness</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RELATEDNESS 5.0</td>
<td>Pairwise or group average r via Queller–Goodnight method; Symmetrical or asymmetrical estimates for pairs; Standard errors via re-sampling over groups or over loci</td>
<td>Macintosh only; User friendly; Assumes unlinked loci</td>
<td><a href="http://www.gsoftnet.us/GSoft.html">http://www.gsoftnet.us/GSoft.html</a></td>
<td>[32]</td>
</tr>
<tr>
<td>KINSHIP</td>
<td>Expected distribution of $r$ from simulated dyads</td>
<td>Macintosh only; User friendly; Assumes unlinked loci</td>
<td><a href="http://www.gsoftnet.us/GSoft.html">http://www.gsoftnet.us/GSoft.html</a></td>
<td>[41]</td>
</tr>
<tr>
<td>DELRIOUS</td>
<td>Pairwise $r$ and $\Delta$ via Lynch–Ritland method; Standard errors via re-sampling over loci</td>
<td>Requires Mathematica software; Assumes unlinked loci</td>
<td><a href="http://www.zoo.utoronto.ca/stone/delrious/delrious.htm">http://www.zoo.utoronto.ca/stone/delrious/delrious.htm</a></td>
<td>[33]</td>
</tr>
<tr>
<td>MER</td>
<td>$r$, $\Phi$ and $\Delta$ via Wang method</td>
<td>Only does one pair at a time; Assumes unlinked loci</td>
<td><a href="http://www.zoo.cam.ac.uk/ioz/software.htm">http://www.zoo.cam.ac.uk/ioz/software.htm</a></td>
<td>[30]</td>
</tr>
<tr>
<td><strong>Likelihood of belonging to relationship category</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KINSHIP</td>
<td>Estimates likelihood that a dyad belongs to a specified category of relationship; Likelihood ratio test for specified alternate hypotheses; Significance test via simulation</td>
<td>Unlinked loci only; No genotyping error; Flexible method for specifying any possible relationship</td>
<td><a href="http://www.gsoftnet.us/GSoft.html">http://www.gsoftnet.us/GSoft.html</a></td>
<td>[41]</td>
</tr>
<tr>
<td>RELPAIR 2.0</td>
<td>Estimates likelihood of specified relationship for each dyad, accounting for linkage among loci; Allele sharing statistics accounting for linkage</td>
<td>Incorporates X-linked loci; Accounts for genotyping error; Accepts putative pedigrees with input file; Eight possible relationships specifiable</td>
<td><a href="http://www.sph.umich.edu/statgen/boehnke/relpair.html">http://www.sph.umich.edu/statgen/boehnke/relpair.html</a></td>
<td>[36]</td>
</tr>
<tr>
<td>PREST</td>
<td>Estimates likelihood of specified relationship for each dyad, accounting for linkage among loci; Likelihood ratio test of specified null relationship versus a specified alternative or versus the most likely of ten possible alternative relationships; Significance estimated via simulation; Calculates maximum likelihood value of the IBD coefficients for each dyad; Tests of relationship via allele sharing statistics, accounting for linkage</td>
<td>Accounts for genotyping error only with parent-offspring and monozygotic twin pairs; Accepts putative pedigrees with input file; Eleven possible relationships specifiable</td>
<td><a href="http://galton.uchicago.edu/~mcpeek/software/prest/">http://galton.uchicago.edu/~mcpeek/software/prest/</a></td>
<td>[37]</td>
</tr>
<tr>
<td>ECLIPSE (PANGAEA package)</td>
<td>Estimates likelihood of specified relationship for trios, accounting for linkage; Given a known pedigree, can also be used to identify mis-scored loci rather than to estimate likelihood of relationship</td>
<td>Accepts putative pedigrees with input file; Accounts for genotyping error</td>
<td><a href="http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml">http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml</a></td>
<td>[46]</td>
</tr>
<tr>
<td><strong>Partitioning cohorts into sibships</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Almudevar &amp; Field methods</td>
<td>Sibship partitions via enumeration of genetically compatible groups</td>
<td></td>
<td><a href="http://ace.acadiau.ca/~alimudev/pedigree.htm">http://ace.acadiau.ca/~alimudev/pedigree.htm</a></td>
<td>[58,68]</td>
</tr>
<tr>
<td>Thomas &amp; Hill methods</td>
<td>Sibship partitions via likelihood evaluated with MCMC methods</td>
<td>Full sib or nested half sib (e.g. polygamous males, monogamous females) partitions</td>
<td><a href="http://maths.abdn.ac.uk/~i.j.waters/">http://maths.abdn.ac.uk/~i.j.waters/</a></td>
<td>[2,13]</td>
</tr>
<tr>
<td>Smith et al. methods</td>
<td>Sibship partitions via likelihood evaluated with MCMC methods</td>
<td>Two methods available; Advantages of each method depend on details of the data set; Full sib partitions only</td>
<td>Software under development; Contact S.C. Thomas for further information [<a href="mailto:s.thomas@sr0.bio.ed.ac.uk">s.thomas@sr0.bio.ed.ac.uk</a>]</td>
<td>[12]</td>
</tr>
<tr>
<td>PARENTAGE 1.0</td>
<td>Full or half sib partitions via likelihood evaluated with MCMC methods; Infers parental genotypes or contributions per parent to a cohort; Estimates mutation rate</td>
<td>Conveniently incorporates any prior information on family sizes or numbers, and on parental identities or numbers; Accounts for genotyping error; Very flexible</td>
<td><a href="http://maths.abdn.ac.uk/~i.jw/">http://maths.abdn.ac.uk/~i.jw/</a></td>
<td>[57]</td>
</tr>
</tbody>
</table>

*Although the published explanation [41] of the algorithm used by KINSHIP contains an error (equations in their Table 2 are actually called once, twice or four times as needed, and then the values are summed), the software outputs the correct likelihoods.*
to incorrectly specified recombination rates [36,37], so maps need not be very accurate. McPeek and Sun [37] show how these techniques could be extended to inbred individuals.

One can use a likelihood approach to ask questions about relationship category in three different ways. (1) Given no hypotheses or other information about a dyad, what is their most likely relationship? One approach is to calculate the probabilities of the data under any hypothesized relationship categories having the same IBD coefficients cannot be distinguished by this method.

Extension to linked loci

If a pair of loci is linked, then the IBD status at one locus (i.e. whether the dyad shares 0, 1 or 2 alleles identical by descent at that locus) is not independent of the IBD status of the adjacent locus. Therefore, one cannot calculate the likelihood of the data across loci by simply multiplying probabilities across loci. Each locus can be in one of three discrete states (0, 1 or 2 alleles identical by descent). Therefore, as you move from locus to adjacent locus in a string of linked loci, the likelihood of IBD states for each locus depend on the recombination rates between the adjacent loci [37,39]. This method can be used to calculate the likelihood of the data under any hypothesized relationship, using any number of linked loci.

Table I. Probabilities that a dyad has a pair of genotypes given that they share m alleles identical by descent

<table>
<thead>
<tr>
<th>Genotype pair</th>
<th>Probabilities*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m = 0</td>
</tr>
<tr>
<td>a0, a0, a0</td>
<td>p0^2</td>
</tr>
<tr>
<td>a1, a1, a1</td>
<td>2 p1 p0</td>
</tr>
<tr>
<td>a2, a2, a2</td>
<td>p1^2</td>
</tr>
<tr>
<td>a0, a1, a1</td>
<td>4 p1 p0</td>
</tr>
<tr>
<td>a0, a2, a2</td>
<td>4 p1 p0</td>
</tr>
</tbody>
</table>

* p0, p1, p2 and p3 are the population allele frequencies of alleles a0, a1, a2 and a3, respectively.

Box 3. Calculating the likelihood of belonging to a relationship category

Given knowledge of population allele frequencies, we can calculate the probabilities, P(G|R), that the observed genotypes are for the two competing categories having the same expected IBD coefficients and so do an equal number of linked loci [42]. However, certain categories have the same expected IBD coefficients and so cannot be distinguished, regardless of how many linked loci are scored (e.g. 2° relatives; Box 1). Nevertheless, these relationship categories do differ in the pattern of meiotic recombination.

(2) If you expect a priori that the dyad will fall into one of a few competing categories, a more useful approach is to calculate the probability of the data under each competing category, and then choose the category giving the highest likelihood [39,40]. In practice, it is most common to test whether the dyad belongs to a pre-specified category (the null hypothesis) versus another pre-specified category (the alternative hypothesis) using a likelihood ratio.

For example, the null hypothesis for a pair of young birds in a nest might be that they are full sibs, whereas a reasonable alternative hypothesis might be half sibs. The problem with the likelihood ratio approach is that one must specify an alternative hypothesis, which might not be obvious. If there are many plausible relationships, one solution is to maximize power to reject the null hypothesis by choosing as your alternative hypothesis the one that gives the highest probability of the data [37]. Simulation is required for testing whether the differences between likelihoods are statistically significant [41].

(3) Approaches (1) and (2) above ask which of several competing relationships is most likely for a given set of individuals. A fundamentally different question is to ask which of several individuals are most likely to have a given relationship (as in parentage analysis or when partitioning a cohort into sibships). In the first situation, the true relationship always has the highest expected likelihood, whereas, in the second situation, the true individuals might not [42].

Power to discriminate among relationship categories

Power to discriminate among relationship categories using likelihood ratio tests depends on the number and polymorphism of the loci and, most importantly, on how different the IBD coefficients are for the competing categories (Box 3). Data from humans show what is possible given dense microsatellite maps. A whole-genome scan used for linkage mapping (300–400 evenly spaced microsatellite loci) yields misclassification rates of close to zero for true monozygotic twins, parent–offspring, full sib and 2° pairs, and of only a few percent for unrelated versus 3° pairs [36,37]. Three to four times that many SNP loci are required for equally high power [37]. Of course, only researchers working on model organisms can currently achieve such discrimination rates. Researchers using fewer loci should obtain estimates of the discrimination power that is possible with their loci via simulation (e.g. KINSHIP software; Table 1). As a rule of thumb, one should be able to discriminate full sib from unrelated dyads with high power (0.9) using 15–20 unlinked microsatellite loci, and parent–offspring pairs from unrelated individuals with ten loci. Around 50 loci might be required for similar power to discriminate 2° pairs from full sibs or unrelatives.

Unlinked loci usually provide a more powerful test than do an equal number of linked loci [42]. However, certain categories have the same expected IBD coefficients and so cannot be distinguished, regardless of how many unlinked loci are scored (e.g. 2° relatives; Box 1). Nevertheless, these relationship categories do differ in the pattern of meiotic recombination.
events separating gametes from the two individuals, and so also differ in the expected length of intact chromosomal regions that are shared identical by descent. Consequently, they can be distinguished using large numbers of linked loci (Box 3). However, the power of these tests is low (e.g. 28–38% misclassification among 2° relatives using whole-genome scans; [36]) and, in this case, an accurate linkage map is important. Browning [43] and Zhao and Liang [44] show how one could, in principle, use data on the lengths of identical-by-descent and non-identical-by-descent regions in gametes sampled from each member of a dyad to improve the discrimination between relationship categories. Including X-linked loci can greatly increase power to distinguish among certain 2° categories because these categories can have very distinct single and multi-locus X chromosome IBD probabilities (e.g. paternal half sisters must share all alleles on one X chromosome; an aunt–niece pair would not). Similarly, Y-linked haplotypes can provide very powerful tests of hypotheses about paternal lineage [45]. Evaluating the joint likelihood for trios of individuals is another good way to distinguish among 2° categories [46]. Indeed, jointly evaluating trios should always be more powerful than three pairwise tests. But this method can be computationally intensive and so, unlike the pairwise method, might be impractical for the routine evaluation of all possible relationships in large data sets [36,47].

Allowing for genotyping error and mutations

Genotyping errors include scoring errors, false homozygotes owing to null alleles or large allele drop out (weak amplification), and mishandled samples. Mutations are essentially scoring errors in their effects on analyses. Typical genotyping error rates for large-scale microsatellite screens are in the range of 0.25% to 2% of genotypes incorrectly specified [47]. At these rates, genotyping errors have little effect on the likelihood of relationship, except in the case of parent–offspring pairs or monozygotic twins, when testing certain trios, or when partitioning cohorts into sibships. In these cases, a single mismatch can make the likelihood under the proposed relationship go to zero [36,46].

The standard way to incorporate genotyping errors is to assume that each diploid genotype is determined correctly with probability 1 – ε, or is chosen at random from the population with probability ε (in which case, the pair is unrelated at that locus) [48,49]. This procedure ensures a non-zero likelihood for all possible relationships. More realistic error models are used for some specialized applications [46,50]. For example, one can model mutation independently for each allele in an individual, and make large-step mutations less likely to occur than are small-step [51]. However, the standard method is computationally efficient and works well in practice [36]. Although it is crucial that some non-zero error rate be incorporated for estimation of relationship in dyads when parent–offspring pairs or monozygotic twins are possible, the actual rate specified does not seem to matter much (ε from hundredths of a percent to a few percent [36,46]). Thus, for most applications, it is not crucial that researchers estimate their actual error rate. A standard rate of 1% should be appropriate for most studies.

Finally, the estimation problem can be turned around and likelihood methods used to test each locus for genotyping errors, assuming the pedigree relationships are known [46,52]. This approach is used for error checking data sets before linkage analysis.

Testing relationship category via allele-sharing statistics: a weaker approach

Another way to test whether a dyad belongs to a particular category is to test whether the number of alleles shared is larger or smaller than expected under the null relationship. One can use as a statistic the number of alleles shared or one of several estimators of the proportion of alleles shared identical by descent [26,37,53,54]. First, generate the expected distributions analytically, via simulation, or via a normal approximation. Then choose cutoff values to control type I and II error rates as appropriate for the question at hand. The drawback of allele-sharing tests is that they have lower power than do likelihood ratio tests when an appropriate alternative hypothesis can be specified [37,53]. One advantage is that ALLELE-SHARING TESTS do not require specifying an alternative hypothesis and so the test can be two-tailed. For example, if putative full sibs share more alleles than expected by chance, then perhaps they are inbred or monozygotic twins; too few shared alleles suggests a more distant relationship. Allele-sharing tests are also insensitive to genotyping errors and are computationally very fast. For example, because likelihood estimation and testing can be extremely slow for large datasets (e.g. when testing all putative pairwise relationships in complex pedigrees using many linked loci), McPeek and Sun [37] recommend an initial screen using allele-sharing tests with a large type I error. Rejected dyads are then re-tested using likelihood ratios. Sun et al. [55] show how plots of expected versus observed allele sharing among all individuals in a putative pedigree can be used to identify inbred or otherwise mis-specified individuals in complex pedigrees.

Partitioning a cohort into sibships

In some situations, the sample of individuals is from a single cohort consisting only of full sibships or of full and half sibships (e.g. tadpoles in a pond or families mixed together in a fish hatchery). The goal is to use molecular marker data to group the individuals into their true sibships. One simple approach is to estimate pairwise genetic distances (e.g. r or simple allele sharing) between all individuals and then graphically cluster them [56]. This method works surprisingly well, even for individuals scored at modest numbers of loci [26]. But the decision of where to draw the family boundaries is simply made by eye and then one requires an ad hoc test of the accuracy of the result (e.g. by verifying that sibships are consistent with mendelian inheritance [11,56], or via pairwise likelihood ratio tests).

Each of the possible ways to group a set of individuals into sibships is called a partition (Figure 1). In principle, it is possible to evaluate the likelihood of the data under every possible partition, and then choose the most likely
One solution is to use MARKOV CHAIN MONTE CARLO (MCMC) methods to sample from the distribution of likelihoods to identify the most likely partitions. Thomas [MCMC] illustrate a general Bayesian approach for inferring the parents of a cohort (hence doing partitions) that easily incorporates previous information about family size distributions, numbers of parents, or the genotypes of any known parents.

Almudevar and Field [58] proposed an interesting approach to partitioning a set of individuals into full sibships that requires no information about population allele frequencies. They first use an algorithm to find all possible sibling groups that are consistent with mendelian inheritance and are maximally large (i.e. the individuals in such a group could all have been produced by a single pair of parents, and no other individual in the sample could be a member of that sibship). Each possible sibship is then assigned a score that is a function of how probable that sibship was, given the putative parents. For example, a full sibship comprising 20 AA individuals and 20 BB individuals is compatible with mendelian inheritance, but is highly improbable. These scores are then used to find the most likely partition. The algorithm worked well on a test data set comprising known salmon sibships scored at only four microsatellite loci [58]. How this method performs relative to the above likelihood approaches has not yet been investigated.

How best to partition a single-generation sample of individuals into sibships is an active area of research [12,13,57,59], and substantial methodological improvements should be forthcoming. Regardless, these initial studies show surprisingly accurate partitions of individuals that were scored at very few microsatellite loci. Given how quickly and inexpensively one can now score individuals for tens or hundreds of loci, very accurate partitions of even large samples of outbred families should be possible.

Prospects

Researchers studying wild populations now routinely use parentage analysis and, to a lesser extent, estimators of relatedness. But they have been slow to adopt the other methods reviewed here and, consequently, many interesting applications of kinship analysis have been neglected. For example, in captive breeding programs, it should be routine to evaluate the relationships among founders of unknown pedigree [5]. Yet there are few published examples (e.g. [7,6]). Similarly, there have been few attempts to use reconstructed sibships to estimate the effective number of breeders that contributed to a cohort [60], even though many researchers must already have the data to do so. Part of the problem is that researchers in diverse fields such as evolution, animal breeding, forensics and gene mapping have been independently deriving similar methods (e.g. [61–63]). For example, most of the likelihood methods for assigning dyads to relationship category were originally developed for verifying pedigrees in human linkage mapping (Box 2). Another problem might be that wildlife biologists usually work with small numbers of loci and so might be put off by the large amount of computing time required.
of data needed to apply methods other than parent–offspring matching (e.g. to estimate accurately pairwise r or to assign dyads to relationship categories with high power). However, surprisingly few loci are required for some methods, such as tests using trios, estimating the proportion of each type of relationship category that occurs in a sample (Box 2), or for accurate partition of cohorts. Furthermore, microsatellite locus development has become routine, and the only real barrier to using dozens or hundreds of loci with wild species should be the ability to estimate recombination rates.

By using multiple analysis methods and conditioning likelihoods with non-DNA information (e.g. ages of individuals, physical location or behavioural interactions), it might now be possible to largely reconstruct the pedigrees of modestly sized populations. Although it might be awhile before we achieve the ‘Holy Grail’ of reconstructing entire population pedigrees from DNA data alone [42,64], statistical methods are improving and genotyping is becoming faster and cheaper. We should soon be able to extract far more pedigree information from wild populations than was ever thought possible.

Acknowledgements
I thank Steve Arnold, Mark Beaumont, Michael Boekhne, Sharon Browning, Charles Criscione, Eric Hoffman, Michael Lynch, David Lytle, David Queller and Solly Sieberts for helpful discussions or comments on the article.

References
26 Blouin, M.S. et al. (1996) Use of microsatellite loci to classify individuals by relatedness. Mol. Ecol. 5, 393–401
31 Li, C.C. et al. (1993) Similarity of DNA fingerprints due to chance and relatedness. Hum. Hered. 43, 45–52