1. INTRODUCTION

The past few years have seen a dramatic increase in the use of multigenerational pedigrees of natural populations in evolutionary biology studies. In this review I outline the origins and benefits of this trend, summarize the available approaches for recovering pedigrees, discuss the consequences of imperfect information and show why recovered pedigrees are superior to proposed alternative approaches.

A pedigree is one of the simplest concepts in biology and probably one of the best understood biological concepts among non-scientists; after all, we each have a family tree, as do our pets and our farm animals. For more than a century, geneticists have recognized the value of pedigrees for studying the inheritance of polymorphisms, inbreeding depression and quantitative genetic variation. It has taken a great deal longer for wild pedigrees to be used—why?

Pedigree analysis within studies of individuals living in the wild has only been made possible by a series of developments. First, the intensive study of breeding success and other traits for all individuals of a species living in a particular area in the wild over several years, although initiated as early as 1936 (Richdale 1957), only became fashionable in the 1980s as ecologists recognized the value of individual life-history data for understanding population processes, and behavioural ecologists sought to measure the results of behavioural strategies in the currency of reproductive success (Clutton-Brock 1988). In many cases, measuring the reproductive success of individuals amounts to recording parentage, meaning pedigrees can be constructed. The first uses of pedigrees for socially monogamous birds (‘social pedigrees’) to investigate inbreeding (Bulmer 1973) and quantitative genetic variation (Boag & Grant 1978) followed soon after.

A second major contribution to modern wild pedigree analysis was made by the discovery of abundant, highly variable neutral genetic markers. The first breakthrough was multilocus DNA fingerprinting with minisatellites (Jeffreys et al. 1985a,b), which was rapidly applied to wild populations to assign parentage (Burke & Bruford 1987; Witt et al. 1987). The second breakthrough was DNA profiling using microsatellites (Litt & Luty 1989; Tautz et al., 1989), which soon superceded DNA fingerprinting for wild population studies. When combined with appropriate statistical analysis (see §3 below), these techniques enable us to confirm suspected pedigree links, or infer parentage or sibship among groups of individuals, with far greater accuracy than is possible from behavioural data alone. Within virtually every social system observed in the field, a great variety of actual mating systems has been revealed. In socially monogamous birds, extra-pair paternity EPP rates range up to 55% across species and vary between populations within species (reviewed by Griffith et al. 2002). Among cooperative breeders, the dominant male in a meerkat (Suricata suricatta) group fathers, 60–80% of the offspring born in the group (Griffith et al. 2003) while in the superb fairy-wren (Malurus cyaneus) all group males together sire just 24% of offspring in the local nest (Mulder et al. 1994). In haplodiploid social hymenoptera, worker relatedness ranges from the often-predicted 0.75 right down to ~0, depending on the number of queens and their number of mates (Avise 2004). Among polygynous breeders, harbour seals (Phoca vitulina) show remarkably low variance in male mating success (Coltman et al. 1998), while red deer (Cervus elaphus) show higher variance in mating success than behavioural data suggest (Pemberton et al. 1992).

Keywords: pedigree; parentage; microsatellite; heritability; inbreeding; kin selection
So far, two studies of non-social passerines using pedigree coancestry and realistic null models have found little evidence for a behavioural inbreeding avoidance strategy (Keller & Arcese 1998; Hansson et al. 2007).

3. PEDIGREE CONSTRUCTION APPROACHES

(a) Field observation

Pedigrees are formed from the accumulation of parent–offspring or sib–sib links. Field observations are the key starting point for pedigree construction, since they often supply hypotheses for genealogy, and if the hypotheses are correct, they make genetic analysis more powerful. For example, knowing the identity and having genotype information for a mother greatly increases the power of paternity analysis. In many birds and mammals, multiple offspring, likely to be sibs, are reared in the same place, a nest or burrow, where they can be marked and sampled. In many birds, parental care is indicative of parentage, though should not be assumed without some molecular analysis (see above). In most mammals, pregnancy and lactation provide excellent maternal information. In some species, additional information can be obtained through modest intervention; in their study of side-blotched lizards
(Uta stansburiana), Sinervo and co-authors bring gravid females into the laboratory briefly for the egg-laying period, hatch the eggs in captivity, and then sample and release the offspring at the mother’s capture site, giving perfect maternal information with minimal influence on reproductive success (Sinervo & Zamudio 2001). At the very least, field observations of marked individuals are useful in determining which candidate parents were in the study area during the mating and parturition periods.

(b) Markers for parentage and sibship inference
For inference of family relationships, microsatellites have been the marker of choice for several years (Parker et al. 1998; Jones & Arden 2003). No other marker type combines the following desirable features: single locus information, codominance, high variability due to many alleles at low frequency, potential for high throughput through automation and short DNA fragments amenable to analysis of forensic samples obtained from wild populations. Identifying microsatellite markers for novel species is through de novo discovery or by taking advantage of their cross-species utility (Barbára et al. 2007). In recent years, centrally funded facilities and commercial companies specializing in finding microsatellites have arisen, so that obtaining loci for parentage analysis is now often a matter of time and money. However, it would be wrong not mention some difficulties. In some taxa, microsatellites are hard to find or insufficiently polymorphic for the task at hand. Genotyping is error-prone, mutations occur and an appreciable proportion of loci has segregating null alleles (Dakin & Avise 2004), all of which can cause false parentage exclusion. A technical issue affecting long-term studies is that microsatellite allele sizes change, and not necessarily in a consistent fashion, between detection platforms (J. M. Pemberton 1999, personal observation).

In future, single nucleotide polymorphisms (SNPs) may be commonly used in pedigree reconstruction in natural populations. Although individually much less informative than microsatellites, they exist in large numbers and scoring is potentially less error-prone than with microsatellites. As a result, discriminatory power for both identifying individuals and parents is potentially very high (Anderson & Garza 2006). Panels of SNPs have now been developed for farm animals and humans and studies confirm their usefulness compared with standard microsatellite panels (e.g. Phillips et al. 2007; Rohrer et al. 2007); the development of SNP panels for well-established long-term natural study populations seems probable in the near future.

(c) Parentage assignment
Parentage analysis using genetic markers requires careful statistical analysis. There is a substantial literature, an array of freeware computer programs and a recent review on the subject (Jones & Arden 2003). In brief, the simplest approach is to use exclusion with associated exclusion probabilities, calculated from allele frequencies, to provide statistical support. However, few studies of natural populations use an exclusionary approach, since candidate sampling is almost never complete, marker panels are not always powerful enough to exclude all but one candidate and genotyping error can easily cause false exclusion of a true parent. Instead, most workers adopt a likelihood approach, which makes better use of candidate genotype information as well as using allele frequencies. Specifically, among those candidates not excluded at a locus, an individual that is homozygous for a required allele is twice as likely to be the true parent as an individual that is heterozygous for the required allele. The nine freeware programs comprehensively reviewed and tabulated by Jones & Arden (2003) take different approaches to dealing with the range of complexities encountered in wild populations such as the existence of large numbers of candidates (of one, both or even unknown sexes), some or all of which may not be sampled or even enumerated; mutations, genotyping errors and null alleles; insufficiently informative marker data; relatives among the candidates; and the assessment of statistical confidence.

There have been some advances in parentage analysis outwith and since the Jones & Arden (2003) review of parentage inference methods. The authors omitted mention of the first full probability, Bayesian, approach to parentage analysis in the absence of any parental information (Emery et al. 2001) which is presented as the program PARENTAGE, available at www.mas.ncl.ac.uk/~nijw/. Duchesne et al. (2005) present PASOS, available at www.bio.ulaval.ca/louisbernatchez/, an open-system (i.e. allows for unsampled candidates) stable mate for their previous program, PAPA. A useful feature of PASOS is that it explicitly estimates the number of unsampled candidates. CERVUS (Marshall et al. 1998) has proved one of the most popular programs, but Kalinowski et al. (2007) point out an error in the way its likelihood equations accounted for genotyping error. This has been corrected in CERVUS v. 3.0, which can now also conduct simultaneous analysis of maternity and paternity and is available from a new website www.fieldgenetics.com.

An interesting recent advance concerns the direct incorporation of field information into parentage analysis. In principle, it makes efficient use of the data, and reduces certain biases, to incorporate information about candidates, for example, spatial proximity, into the same analysis as the genetic marker information. Hadfield et al. (2006) took this approach in the case of the Seychelles warbler (Acrocephalus sechellensis), in which microsatellite variation is low, helpers at the nest of both sexes, which are relatives of the dominant pair, are potential parents and there are extra-territory fertilizations. This Bayesian approach (available as MASTERBAYES at http://www.R-project.org) found several different extra-group paternity assignments compared with previous methods. In the future, using developments of this approach, it will be possible to estimate quantitative genetic parameters or inbreeding depression at the same time as the pedigree (Hadfield et al. 2006).

The development of such a diversity of parentage inference programs is a reflection of the diversity of problems encountered during parentage analysis in natural populations. However, by far the most common problems in parentage analysis are that candidate parents are poorly sampled and the amount of marker information available is marginal for confident resolution of parentage links, even when the true parents are sampled (Marshall et al. 1998; Jones & Arden 2003), suggesting that we should never skimp on sampling effort, the number of loci screened and the accuracy with which the loci are screened.
(d) Sibship reconstruction

Another approach to partial pedigree construction using genotype data is to recover full and half sibships from samples of individuals. Here, methods have developed rapidly over the last few years (Blouin 2003). Butler et al. (2004) tested four algorithms for full sibship reconstruction ranging from an exclusionary approach through methods using MCMC to maximize the likelihood of partitions between sibships, and showed that they varied in accuracy depending on the structure of the data in terms of family size, and that all were sensitive to genotyping error. The approach of reconstructing sibships using MCMC laid out by Thomas & Hill (2002) has been developed further, particularly to deal with genotyping error, by Wang (2004) and is available as the program COLONY from www.zoo.cam.ac.uk/ioz/software.htm. Although some downstream analyses can be carried out with sibship data, they do not of themselves allow, for example, the analysis of inbreeding depression, and the challenge now is to combine sibship analyses with data on parentage analysis to construct more complete pedigrees. One possible approach was demonstrated in a study of bighorn sheep (Ovis canadensis) in which candidate sires were only partially sampled. Colman et al. (2005) genotyped offspring without identifiable sires at 32 loci and used COLONY to infer 38 half sib clusters among 167 offspring. This information substantially increased the number of pedigree links available for quantitative genetic analysis.

(e) Pedigree reconstruction without field data

With enough polymorphic markers, it should be possible to reconstruct a pedigree of a sample of individuals without the need for any field information. Methods using simulated annealing algorithms have already been proposed and tested by Almudevar (2003) and Fernandez & Toro (2006) and this field seems likely to expand greatly as marker information increases for natural populations. However, from the perspective of downstream analyses, ecological information about individuals will nearly always be useful. For example, information on year of birth often explains trait variation and is usefully fitted as an additional effect in quantitative genetic analyses.

4. PEDIGREE QUALITY

Pedigrees for wild populations vary in depth, accuracy, size, completeness and structure, and a fast-growing literature describes the effect of this variation on the results obtained from evolutionary genetic analyses. This is useful both from the perspective of those planning to try and recover pedigrees for wild populations and for those analysing pedigrees for which there is no prospect of retrospective pedigree improvement, because the individuals have died or dispersed without sampling.

(a) Pedigree depth

Coancestry and relatedness are greatest between members of the same or adjacent generations (e.g. parent–offspring), and the inbreeding coefficient of an individual is greatest when close relatives mate. This is good news for studies of wild populations, for it means that it is not necessary to have great depth of pedigree, in terms of generations, to capture most of the variance in these parameters. This point was made very clear by Balloux et al. (2004) who investigated the correlation between f calculated over generations 2, 3, ... and f calculated over 50 generations in simulated populations covering four example vertebrate breeding systems and population structures. Within just five generations, 90% of the variance in 50-generation f is captured, regardless of population detail, and some simulated structures reached this figure far sooner (figure 1).

(b) Pedigree accuracy

The accuracy of pedigree links is a major concern for all studies. In general, errors might be expected to result in downward-biased and less precise estimates of heritability (Kruuk 2004) and this has indeed been observed, for example, in Darwin’s finches (Geospiza fortis) using parent–offspring regression (Keller et al. 2001). However, Charmantier & Réale (2005) examined the effect of extra-pair paternities in simulated and real pedigrees of a socially monogamous bird species and showed that, provided the number of families studied is sufficient, animal model heritability estimates are surprisingly robust to EPP rates up to 20%. This finding is also good news for those using molecular parentage analysis with marginal power (see above). Nevertheless, for small sample sizes or highly heritable traits, heritability and other quantitative genetic parameters will be downward biased as the accuracy of pedigree links declines, and systematic patterns in pedigree errors, such as misassignment of paternity to spatially closest males, could cause environmental covariance to be misinterpreted as genetic covariance.

Similarly, estimates of inbreeding depression will be imprecise when pedigree links are inaccurate. Inbreeding coefficients calculated from the social pedigree suggest that Mandarte Island song sparrows experience substantial inbreeding depression in several traits (Keller et al. 1994; Keller 1998). A microsatellite analysis of four cohorts of chicks showed that due to EPPs, 28% of paternal links in the social pedigree are wrong (O’Connor et al. 2006). Using this information, Marr et al. (Amy B. Marr, Louis C. Dallaire and Lukas F. Keller 2007, personal communication) estimated inbreeding depression in the population with increasing proportions of paternity error (28%, the existing social pedigree, to...
100%) and then extrapolated to a predicted inbreeding depression if there was no paternity error. For two of the traits studied, inbreeding depression was predicted to be significantly higher when pedigree errors were zero, suggesting that estimates of inbreeding depression emerging from this study to date are conservative.

(c) Pedigree structure

More subtle issues surround the actual pattern of pedigree links in time and space. Polygonous species yield pedigrees which are good for estimating maternal and shared environment effects, since paternal half sibs have different mothers who may range in habitats of different qualities (Kruuk & Hadfield 2007). Long-lived and/or iteroparas species lend themselves to studies of ontogenetic effects and genetic × environment interactions (Wilson et al. 2005, 2006, 2007). Adding newly collected trait data for recent cohorts to a large pedigree of a short generation time bird (great tit, Parus major) and a smaller pedigree of a more iteroparas long-lived bird (mule swan, Cygnus olor) had contrasting effects (Quinn et al. 2006). In general, quantitative genetic parameters were estimated with greater precision in the great tit pedigree, presumably because sample sizes were greater and first and second degree relatives with measured traits were more likely to occur in adjacent sampling years.

A general difficulty is that owing to the variety of pedigrees and genetic architectures observed, it is hard to determine how powerful a pedigree is for measuring specific parameters and the extent to which errors or gaps in pedigree links will affect results. To address this issue for quantitative genetic studies, Morrissey et al. (2007) suggest a framework in which an empirically acquired pedigree and a user-supplied quantitative genetic architecture for traits can both be manipulated (e.g. wrong pedigree links can be created), and then used in animal models, to discover just how robust results obtained with real trait data are likely to be. A computer package, PEDANTS, is available at http://wildevolution.biology.ed.ac.uk/awilson/pedantics.html for this purpose.

5. ALTERNATIVES TO PEDIGREES AND INSIGHTS ARISING FROM THEM

The existence of extensive microsatellite genotype data for free-living populations, often combined with information about traits, including behaviour, for the individuals involved, has led to alternative non-pedigree-based approaches to parameter estimation in studies of quantitative genetics, inbreeding and cooperation. These approaches use genotype data as a proxy for the coancestry and inbreeding coefficients outlined above and have great attraction since they avoid the laborious process of parentage analysis and the time required for generations to pass.

(a) Coancestry and relatedness

Conceptually, the sharing of marker alleles between two individuals, after taking account of population allele frequencies, yields an estimate of coancestry. Many different marker-based estimators of pairwise coancestry have been derived over recent years including method-of-moments estimators, maximum-likelihood estimators, two-gene estimators, four-gene estimators and different approaches to allele frequency correction (Queller & Goodnight 1989; Ritland 1996a; Van de Casteele et al. 2001; Thomas 2005; Oliehoek et al. 2006).

Pairwise coancestry estimators based on marker data have been widely used in the behavioural ecology literature in studies of cooperation. Less commonly, they have been used to investigate inbreeding avoidance behaviour (e.g. Reusch et al. 2001). Furthermore, it is in principal possible to use them (with phenotypic data) to infer quantitative genetic parameters without the need to resolve pedigrees (Ritland 1996b, 2000). Heritability estimates using this method were predicted from the start to be highly dependent on the variance in relatedness, and indeed it turns out that heritabilities calculated for outbred vertebrates are erratically different from animal model estimates applied to pedigree data for the same sample (Thomas et al. 2002; Wilson et al. 2003; Colman 2005; Garant & Kruuk 2005; Freniu 2008). Furthermore, in the context of quantitative genetics and inbreeding depression, the lack of information on precise ancestry is a great disadvantage, for it prevents study of additional and often important sources of variance such as maternal effects.

Closer inspection of pairwise marker-based coancestry estimators has shown that at least for outbred vertebrates, they are rather imprecise. Mean and variance in coancestry in real study populations is far lower than has typically been assumed for testing the average performance of coancestry estimators (e.g. compare Van de Casteele et al. (2001) with Castellé et al. (2006); figure 2), and the low precision with which just a few loci can capture this variance merely adds to the difficulties. Pairwise, marker-based coancestry estimators should therefore be used with care in evolutionary studies: they are at their best when applied in scenarios with high variance in pedigree relatedness (e.g. within some haplodiploid hymenopteran colonies or selected samples of individuals likely to show high variance in coancestry). In all other scenarios, including tests of cooperative behaviour, it is questionable how powerful tests using pairwise relatedness really are.

(b) Inbreeding coefficients

Inbred individuals should be more homozygous than outbred individuals after correcting for population allele frequencies, and again a variety of marker-based estimators of individual inbreeding coefficients have been proposed (Hill et al. 1995; Ritland 1996a; Coulson et al. 1998; Colman et al. 1999; Amos et al. 2001). Despite a probable publication bias, there is a certain consistency to findings of a positive correlation between standardized heterozygosity and fitness in natural populations (Colman & Slate 2003).

The idea that heterozygosity or inbreeding coefficient estimated from a few marker loci has precision for measuring inbreeding depression in normally outbred diploids has recently been eroded from several directions. The observed correlation between pedigree inbreeding coefficients and marker-based estimates of inbreeding is often low despite good data (Markert et al. 2004; Slate et al. 2004; Overall et al. 2005; Rodriguez-Ramilo et al. 2007) and this is largely because the mean and variance of inbreeding coefficients are both low in those natural populations so far studied (Slater et al. 2004).

drawn pairs have not altered the view that the overwhelming majority of randomly restriction adds additional classes of relatedness but does analysis to two-generation deep pedigrees; relaxing this For simplicity, these figures were derived by restricting in text except great reed warbler (Acrocephalus arundinaceus) Z 0.25; black bars, r 0.5. The five different scenarios for relatedness structure suggested by the authors are shown. Note that parent–offspring and full sib categories used by the authors have been collapsed into a single category here. (b) The same information for five wild pedigrees analysed by Csalarsityestimators. White bars, r 0; stippled bars, r 0.25; black bars, r 0.5. All species previously identified in text except great reed warbler (Acrocephalus arundinaceus). For simplicity, these figures were derived by restricting analysis to two-generation deep pedigrees; relaxing this restriction adds additional classes of relatedness but does not alter the view that the overwhelming majority of randomly drawn pairs have r 0.

Figure 2. Comparison of hypothetical and observed distributions of relatedness in outbred vertebrate populations. (a) Percentage of pairs of individuals which, if drawn at random from a population, would fall into different relatedness categories as used by Van de Casteele et al. (2001) in a study of the average performance of marker-based relatedness estimators. White bars, r 0; stippled bars, r 0.25; black bars, r 0.5. The five different scenarios for relatedness structure suggested by the authors are shown. Note that parent–offspring and full sib categories used by the authors have been collapsed into a single category here. (b) The same information for five wild pedigrees analysed by Csalarsityestimators. White bars, r 0; stippled bars, r 0.25; black bars, r 0.5 (as above). All species previously identified in text except great reed warbler (Acrocephalus arundinaceus). For simplicity, these figures were derived by restricting analysis to two-generation deep pedigrees; relaxing this restriction adds additional classes of relatedness but does not alter the view that the overwhelming majority of randomly drawn pairs have r 0.

conclusions were reached by a simulation study (Balloux et al. 2004). As for coancestry, so for inbreeding coefficients: marker-based estimators of inbreeding are at practice with the screened microsatellites, an idea also known as the ‘local effects hypothesis’ (Hansson & Westerberg 2002).

In conclusion, it is theoretically possible to estimate levels of coancestry and inbreeding from marker data. In practice, this approach is imprecise in several wild populations of outbred organisms studied so far. Greater precision is obtained by using marker data to determine parentage and sibships, yielding a pedigree from which coefficients of coancestry can be calculated and with which fixed effects and a range of variance components can be appropriately assessed.

6. THE FUTURE

Wild pedigrees form a crucial part of a rich seam of data from individual-based projects, analyses of which are likely to stretch for years into the future. There is a correct pedigree for every individual in a population and it is well worth striving to ascertain that pedigree since the information allows better downstream analysis in every way. The main way to resolve pedigrees well is to sample individuals as completely as possible and to use a sufficiently informative panel of markers. Where retrospective social pedigrees cannot be corrected through molecular genetics, or populations are just too large for detailed molecular genetic analysis of all individuals to be practical, the tools are now available to allow estimation of bias and imprecision using pedigree error rates obtained from testing a sample of individuals.

The majority of wild pedigrees analysed to date are for small birds or large mammals, reflecting the relative ease with which individuals can be studied in these groups. In the future, it is to be hoped that the ingenuity of fieldworkers and the power of molecular genetics can greatly expand the taxonomic range of studies to enable exploration of patterns of quantitative genetic variation and inbreeding depression across a wider range of life histories and breeding systems than currently available.

Only some rather general applications of wild pedigrees have been outlined above, but there are more potential topics for investigation in the future. Given enough markers, wild pedigrees can be used to construct linkage maps for study pedigrees, which can then be used to map polymorphisms and quantitative trait loci (addressed elsewhere in this volume (Slate 2008)) and analyse the rate of decay of linkage disequilibrium with genetic distance. No authors have yet investigated dominance variance in a pedigree for a natural population and it is not yet clear whether any wild pedigree structures lend themselves to such analysis. Much further investigation into the effects of imperfect pedigree information can be expected, including the assumptions that founders and immigrants are outbred and unrelated.

Thanks to two referees, both the editors and Tristan Marshall for their comments on the MS.
REFERENCES


Boag, P. T. & Grant, P. R. 1978 Heritability of external morphology in Darwin finches. Nature 274, 793–794. (doi:10.1038/274793a0)


