3-2005

Assignment methods: matching biological questions with appropriate techniques

Stephanie Manel
Universite Joseph Fourier (Grenoble I), stephanie.manel@ujf-grenoble.fr

Oscar Gaggiotti
Université Joseph Fourier, oscar.gaggiotti@ujf-grenoble.fr

Robin Waples
NOAA, robin.waples@noaa.gov

Review TRENDS in Ecology and Evolution Vol.20 No.3 March 2005

Assignment methods: matching biological questions with appropriate techniques

Stephanie Manel[1], Oscar E. Gaggiotti[1] and Robin S. Waples[1,2]

[1]Laboratoire d'Ecologie Alpine (LECA), Génomique des Populations et Biodiversité, Université Joseph Fourier, Grenoble, France
[2]Northwest Fisheries Science Center, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

Assignment methods, which use genetic information to ascertain population membership of individuals or groups of individuals, have been used in recent years to study a wide range of evolutionary and ecological processes. In applied studies, the first step of articulating the biological question(s) to be addressed should be followed by selection of the method(s) best suited for the analysis. However, this first step often receives less attention than it should, and the recent proliferation of assignment methods has made the selection step challenging. Here, we review assignment methods and discuss how to match the appropriate methods with the underlying biological questions for several common problems in ecology and conservation (assessing population structure; measuring dispersal and hybridization; and forensics and mixture analysis). We also identify several topics for future research that should ensure that this field remains dynamic and productive.

Biologists are becoming increasingly interested in using population genetic approaches to answer questions of ecological, evolutionary, or conservation relevance (e.g. 'What proportion of individuals captured in population A are immigrants from population B?'; 'Is this individual an endangered species or a hybrid?') that involve the contemporary dynamics of natural populations. By contrast, most traditional population genetic models are based on equilibrium assumptions; that is, they attempt to characterize long-term genetic processes that have achieved a stable balance between opposing evolutionary forces. Events occurring on much shorter (ecological) timescales, which are the focus of much current interest, are typically considered 'noise' in equilibrium models.

Contemporary events can be studied using a variety of genetic approaches that collectively can be called assignment methods (AMs) (see Glossary). AMs use genetic information to ascertain population membership of individuals or groups of individuals. One approach, the 'assignment test' (AT; Box 1), has seen widespread use over the past decade [1–7]. More recent advances enable one to address questions such as, 'How many populations exist in my study area?' and 'Is this individual a migrant?'

Glossary

Admixture: a composite gene pool in which at least some individuals (F_1, F_2, … F_x and/or backcross) can trace ancestry to more than one population.
Assignment index: the expected frequency of the multilocus genotype of an individual in the population from which it was sampled. Individuals with a relatively low assignment index have genotypes that are unlikely for that population and, thus, are potential immigrants.
Assignment method (AM): any of several related statistical methods that use genetic information to ascertain population membership of individuals or groups of individuals (Table 1).
Assignment test (AT): a statistical test of the hypothesis that the multilocus genotype of an individual in question arose from a particular population (Box 1).
Bayesian analysis: a method of statistical analysis that begins with prior distributions for the model parameters and updates these based on observed data to arrive at a posterior probability distribution.
Classification: a method for assigning individuals to predefined categories, based on a suite of characters (e.g. multilocus genotype) measured for the individual and for samples from each category (e.g. potential source populations).
Clustering: a method for decomposing a mixture into its component parts (e.g. gene pools or populations) in the absence of information to characterize the units *a priori*.
Discriminant function: a linear combination of variables that maximizes the contrast among different groups of interest. Unknowns (e.g. individuals) are classified into one of the groups (e.g. populations) based on their score on the discriminant function. An AT is a type of discriminant function.
Frequentist methods: statistical methods that test hypotheses about an event based on the expected frequency of that event happening over a large number of trials (frequency distribution). If no such information is available (e.g. from a theoretical frequency distribution), randomization techniques are used to generate an empirical frequency distribution.
Likelihood: the probability of obtaining the observed data under a certain model or hypothesis. The assignment index can be viewed as the likelihood of an individual occurring in the population in which it was sampled. An ML approach finds or approximates the parameter values that maximize the likelihood.
Linkage disequilibrium: the non-random association of alleles at different gene loci. A standard index of linkage disequilibrium, *D*, is defined as the difference between observed and expected frequencies of a two-locus gamete. If the two loci are independent, the expected frequency of a gamete is the product of the frequencies of the two alleles.
Markov Chain Monte Carlo (MCMC): a simulation technique to generate samples from a probability distribution of interest. MCMC can estimate complex multivariate distributions that cannot be generated by standard simulations methods. It has also been used to approximate likelihood surfaces in the context of ML methods.
Mixture: a group of F_0 individuals originating in different populations.
Mixture analysis: estimation of the proportion of individuals that different source populations contribute to a genetic mixture or admixture. This is typically done using classification models (Box 2) when data for potential source populations are available, but clustering models can be used when such data are not available.
Parentage analysis: a classification method for determining the parents of an individual or group of individuals (Box 3).

Corresponding author: Waples, R.S. (robin.waples@noaa.gov).
Available online 6 January 2005

www.sciencedirect.com 0169-5347/$ - see front matter. Published by Elsevier Ltd. doi:10.1016/j.tree.2004.12.004

Box 1. Principles of traditional assignment tests

ATs address classification problems and attempt to 'assign' unknown individuals to their population of origin, based on the multilocus genotype of an individual and the expected probabilities of that genotype occurring in each of the potential sources. In the original formulation [1], expected genotypic probabilities are computed from samples from each potential source population, and genotypes are assigned to the population in which that genotype is most likely to occur. Fundamental assumptions are that all potential source populations are defined in advance, sampled randomly, and are in Hardy–Weinberg and linkage equilibrium.

Each AT has an exclusion-method counterpart, which provides a measure of the confidence associated with individual assignments [12]. For each potential source population, a distribution of genotypes is generated by Monte Carlo simulations based on the sample allele frequencies. By repeating this procedure many times (e.g. 10 000), one obtains the expected distribution of genotypes in that population, which, in turn, is used to generate a distribution of genotypic likelihood values. The likelihood of a particular genotype of interest is then compared to the empirical distribution for each candidate population. If the likelihood of the genotype falls in the tail of the distribution (e.g. less than a critical value such as $\alpha=0.01$ or 0.001), one can exclude that population as the origin of the individual. If all but one population are excluded, the individual is assigned to the non-excluded population. Exclusion methods thus provide a check of the standard AT assumption that the true population of origin has been sampled.

ATs combined with exclusion methods can have an important role in conservation; for example, in ensuring broodstock integrity of captive propagation programs. When a hatchery program for endangered winter-run Chinook salmon in the Sacramento River in California was at risk of contamination from a nearby spring-run population (the population names refer to the season in which adults enter fresh water to begin their spawning migration), the AT software WHICHRUN [51] was developed to enable real-time screening of maturing adults based on microsatellite data from fin clips. Genetic differences between the two populations were sufficient to enable managers to exclude >99.9% of spring-run fish from broodstock collection while excluding only 1% of fish that were winter run [20].

Various AT programs are available from http://www.bio.ulaval.ca/louisbernatchez/links_fr.htm.

Although these approaches hold great promise, the recent proliferation of publications on assignment problems has left researchers with a sometimes bewildering array of potential methods to assess. Furthermore, with the availability of sophisticated new computer programs, it is easy for researchers to apply methods without adequately considering the underlying biological questions.

Here, we consider biological questions that are commonly asked by researchers interested in AMs and attempt to match the most appropriate method(s) with each

Box 2. Genetic mixture analysis

Many fish species migrate extensively and, as a result, harvests often include numerous different populations or stocks. Shaping mixed-stock fisheries to take advantage of abundant populations without imposing excessive risks on less abundant and productive populations is one of the most difficult challenges of fishery management. Here, the question of interest is not the origin of individual fish, but rather the stock composition of a fishery and how it changes in space and time.

Figure I shows contrasting results of genetic mixture analysis using two different classification methods. Both start by calculating the likelihood of each individual originating from each potential source, based on genetic data from the fishery as well as from all potential source populations (Figure Ia). Next, ATs (Figure Ib) classify each individual to the population with the highest likelihood (individual 1 to source B and individual 2 to source C), and estimated mixture proportions are the sum of the individual assignments. However, this approach ignores information regarding the uncertainty of individual assignments. During the late 1970s, maximum likelihood (ML) methods for genetic stock identification (GSI) were developed to estimate directly both the mixture proportions and the posterior source probabilities for each individual. In a procedure analogous to fractional paternity assessments (Box 3), each individual is 'carved up' and allocated to each source in proportion to these posterior probabilities (Figure Ic). Mixture proportions are estimated as the sum of all fractional assignments; this is done iteratively, with the posterior probability of population membership at each iteration being used as the prior for the next iteration. In this way, genetic mixture analysis is performed jointly on all unknown individuals, rather than independently as in ATs. The few direct empirical comparisons of the two methods have demonstrated greater accuracy of GSI [20,52], although performance of ATs is comparable if genetic differences among populations are large and individual assignments can be made with confidence [53].

The first GSI approaches [49,54] used a conditional ML model that assumed that source population allele frequencies were known without error and that all potential sources had been sampled. These assumptions were relaxed in an unconditional ML model [50] that treated source population data as estimates and allowed for the possibility of unsampled sources. The unconditional model, which takes advantage of information about source population allele frequencies from individuals in the mixture, shares attributes of the classification and clustering approaches and presaged some features of later AM models (e.g. [17]). Statistical aspects of GSI have recently been reviewed [55]. A recent ML implementation (SPAM; [56]) is available from http://www.cf.adfg.state.ak.us/geninfo/research/genetics/software/spampage.php, and a Bayesian analysis (BAYES; [18]) from ftp://ftp.afsc.noaa.gov/sida/mixture-analysis/bayes/.

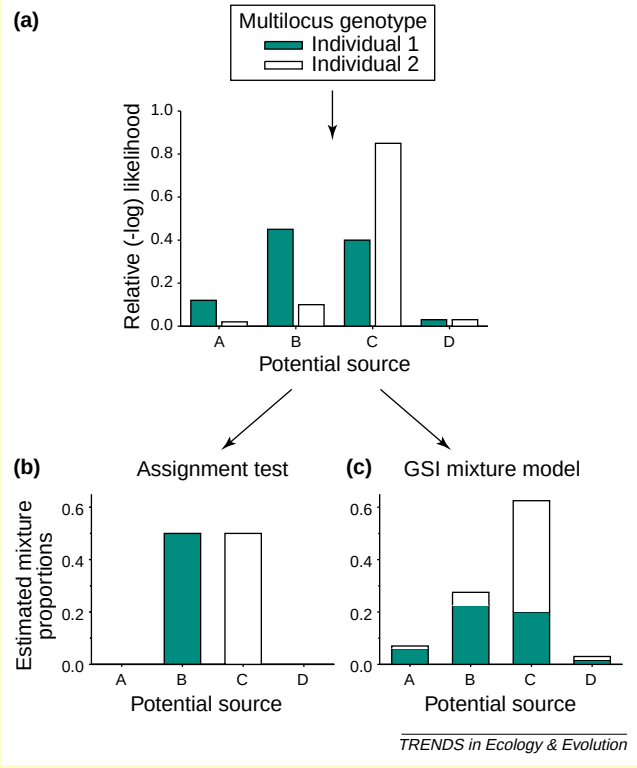

Figure I.

Box 3. Parentage analysis

Parentage analysis, which involves identifying the parents of specific individuals [57,58], is a particular case of ATs and is affected by many of the same statistical issues as ATs are [59]. Ideally, all potential parents are genotyped and all but one pair can be excluded categorically, based on the multilocus genotype of the progeny. When this is not possible, fractional parentage assignments (analogous to the method for analyzing genetic mixtures, Box 2) can be made, based on the relative likelihoods of different potential parents [60]. In some cases, it is possible to reconstruct genotypes of unknown parents [58,61]. Blouin [62] recently reviewed DNA-based methods for the related field of kinship analysis.

The methods described elsewhere in this paper all depend on some degree of genetic differentiation among candidate source populations. In some cases, it is possible to address the same types of question using parentage analysis, even when the groups in question do not differ genetically in any systematic way. This can be accomplished by genetically identifying parents of each individual (i.e. 'assigning' each individual to its two parents) and then grouping the parents (and their respective numbers of offspring) according to the trait of interest, such as size or time of reproduction [63] or hatchery versus wild origin [64]. This type of analysis has also been used to study male selection gradients in plants [65], and recent modifications [66] have been used to estimate pollen dispersal curves.

In the study of mating systems, one might also be interested in different questions, such as whether a family is a mixture of individuals descended from more than two parents, whether mating is random or assortative, or whether males and females have the same number of mates. Parentage analysis has often provided novel and surprising insights into mating structure and behavioural ecology of the studied species [57]. Typically, this involves measuring the contribution of different males to the offspring of a single female. A recent variation is designed to detect subfamilies within a single batch of offspring, based on the presence of linkage disequilibrium caused by a mixture of progeny groups [67].

question. We use the term 'assignment method' broadly to include genetic mixture analysis (Box 2) and parentage analysis (Box 3), because they use the same basic principles and are also active fields of research.

What types of problems can AMs address?

AMs can address two basic types of problem: classification and clustering (Table 1). Which formulation is more appropriate depends on how much prior information one has about the categories of interest (e.g. family, population and species).

In classification problems, individuals are assigned to predefined categories. A standard approach is to compute a discriminant function based on samples from potential sources and then classify unknowns to the group with the highest discriminant score. In the case of AMs, the discriminant function is the expected genotypic frequency distribution under the assumption of Hardy–Weinberg and linkage equilibrium in each source population. AMs that use classification at least in part include ATs, genetic mixture analysis, and parentage analysis.

Clustering problems [8] are more challenging because the categories are not predefined; instead, they must themselves be constructed from the data. In the absence of source population data to guide classification, clustering methods rely on the presence of linkage disequilibrium, which occurs in a mixture of individuals from different populations [9,10] even if all contributing populations are

Table 1. Characterization of the principal assignment methods in relation to the biological question addressed[a]

Approach	Statistical method[b]		Distinctive features of method	Questions[d]	Refs
	Assignment	Estimate of allele freq.[c]			
Classification					
Assignment test (AT)	ML	Freq	First formulation of an AT; for codominant markers	O	[1]
	ML	Freq	AT for dominant markers	O	[48]
	ML – Freq	Bay	AT allowing identification of migrants	O,D	[4]
	Freq	Freq	Exclusion tests	O,D	[12]
Genetic mixture analysis	ML	Freq	Assumes all sources sampled and gene frequencies known without error	M	[49]
	ML	ML	Allows for unsampled sources and estimation of source gene frequencies	M	[50]
	Bay	Bay	Bayesian implementation	M	[18]
	Bay	Bay	For studying colonization processes	M	[16]
Methods for answering specific questions	Bay	Bay	Estimates migration rates	D	[30]
	Bay	Bay	Identifies hybrid individuals	H	[47]
Clustering					
Methods for delineation of populations	Bay	Bay	First method for the identification of populations. Inference about number of populations requires trying different values of this parameter	O,S,D,H	[17]
	Bay	Bay	As above, but allows for linked loci	O,S,D,H	[15]
	Bay	Bay	Number of populations is estimated and the range of plausible values for this parameter has the number of sampled individuals as upper boundary	O,S,D,H	[14]
	Bay	Bay	The range of plausible values for the number of populations has the number of observed subpopulations as upper boundary	O,S,D	[13]

[a]The list is not exhaustive and presents only the latest or most widely used methods in the literature; methods of parentage analysis have been recently reviewed elsewhere [58] and were not included.
[b]Abbreviations indicate method used in each step: Bay, Bayesian; Freq, frequentist; ML, maximum likelihood.
[c]For allele frequency estimation in each population. 'Freq' indicates that the method uses sample allele frequencies.
[d]Abbreviations: D, dispersal; H, hybridization; M, genetic mixture analysis; O, origin of specific individuals; S, population delineation and structure.

in linkage equilibrium. The magnitude of disequilibrium depends on the mixture fractions and the genetic distance between populations [11]. Clustering methods attempt to decompose the mixture by creating groups of individuals within which linkage disequilibrium is minimized; these groups can be considered to be populations or gene pools. Thus, clustering methods can simultaneously delineate clusters of individuals based on their multilocus genotypes and assign individuals to the identified clusters, typically using a Markov chain Monte Carlo (MCMC) approach. Clustering methods are particularly useful when genetic data for potential source populations are not available, population boundaries are uncertain, or when some (but perhaps not all) potential sources have been sampled.

What statistical methods are used with AMs?

AMs have been implemented using frequentist and/or likelihood methods [either maximum likelihood (ML) or Bayesian analysis] (Table 1). Frequentist approaches use statistical hypothesis testing and give a *p*-value derived from a predefined or simulated frequency distribution [12]. Likelihood methods assume that observed data arose from a probabilistic model with unknown parameters; their objective is to use the data to estimate parameters of the model, and to assess the degree of uncertainty associated with these estimates. ML methods produce point estimates of model parameters that maximise the likelihood function, whereas Bayesian methods produce posterior distributions for model parameters; both use numerical integration methods, such as MCMC. Most recent progress in assignment problems has been made within a Bayesian framework [13–18].

Key biological questions

We now consider biological questions commonly asked by users of AMs and identify the method(s) best suited for each question.

What is the origin of a specific individual?

ATs provide the most direct method to determine the population of origin of target individuals. ATs can facilitate detection of illegal harvests and trade routes, help in the management of captive-breeding programs by excluding non-target individuals [19,20], and help to develop control mechanisms for bioinfestation [21]. As in all classification problems, potential source populations must be defined in advance, so bias can result if the true source is not among those sampled. To confirm that the individual truly belongs to the population, Cornuet *et al.* [12] suggest using an exclusion-based AT (Box 1).

The combination of classic AT and exclusion method was used effectively to detect fraud in a fishing competition in Finland [22], where the largest fish presented was a 5.5-kg salmon. Based on data for seven microsatellite loci, the competition location was excluded as a plausible source and, faced with this evidence, the offender confessed to purchasing the 'winning' fish at a local market.

Population structure

Evaluating population structure is of considerable interest to biologists because it is a precursor to answering many other types of question (e.g., estimating migration and identifying conservation units).

Identification of populations Species are commonly subdivided into local breeding populations or less well defined genetic neighbourhoods. In some cases, this structuring is easy to infer from the geographical location of individuals, and standard statistical methods (e.g. a contingency χ^2 or randomization test [23]) provide a direct means of testing the null hypothesis that multiple samples come from a single panmictic population. ATs are an alternative means of testing panmixia: if all samples have come from the same global population, individuals will be no more likely to be 'assigned' to their collection locality than to any other locality. If the proportion of correct assignments is significantly higher than the random expectation, it can be concluded that population structure exists. We are not aware, however, of any studies that have systematically evaluated whether ATs provide increased power to detect population structure compared with standard statistical methods.

For many species, demarcation of geographical populations is problematical and, in this case, clustering methods [13–15,17] provide the best solution. The method of Pritchard *et al.* [STRUCTURE; 17] infers the number of clusters (populations) by comparing the posterior probability for different numbers of putative populations specified by the user. Cegelski *et al.* [24] used STRUCTURE to delineate populations in wolverines *Gulo gulo*, with the objective of identifying the appropriate scale for conservation and management. Based on data for ten microsatellite loci, they showed that Montana wolverines are not panmictic, in spite of their geographical proximity (within 300 km) and dispersal capability. Recent human disturbances might be responsible for this apparent fragmentation, and results have helped focus attention on the importance of identifying continuous areas of undisturbed habitat to protect this species. Falush *et al.* [15] extended the method of Pritchard *et al.* [17] by allowing for physical linkage between loci, illustrated with studies of admixture in African-Americans, recombination in *Helicabacter pylori*, and drift in *Drosophila melanogaster*.

In contrast to the above approaches [15,17], the methods of Corander *et al.* [13] and Dawson and Belkhir [14] directly estimate the number of populations but differ in the range of possible values for this parameter. For Corander *et al.* [13], the maximum number of populations allowed is the number of locations sampled, whereas for Dawson and Belkhir [14], it is the total number of individuals. The method of Corander *et al.* [13] cannot, therefore, detect substructuring if it occurs within each location. No study has yet compared the performance of all these methods on either a simulated or empirical data set. All methods based on cluster analysis involve considerable uncertainty unless the true populations are strongly divergent. For example, the number of estimated populations can be affected by model assumptions and cryptic relatedness [15], and results for a specific individual or group can also vary depending on which other individuals are included as unknowns.

When geographical locations of individuals are known but population limits are unclear, methods other than

AMs that enable one to identify genetic boundaries [25] or to find the best partition of the overall sample based on AMOVA and geographical coordinates [26] are more suitable. For species in which individuals are continuously distributed, spatial autocorrelation [27] or regression methods are more appropriate because they do not assume a spatial structuring of populations.

Population differentiation Once populations have been defined, a common question is, 'How different are they?' Traditionally, this is addressed using measures of genetic similarity, such as F_{ST} and genetic distance. Recently, ATs have been used to provide a measure of population differentiation by calculating the proportion of individuals assigned to the population in which they were sampled. In general, however, ATs merely confirm what we suspected from other measures of genetic differentiation: the percentage of correct assignment increases with larger F_{ST} or genetic distance [2,28,29]. For example, Waits *et al.* [28] used F_{ST} and ATs to analyze genetic structure in four populations of Swedish brown bears *Ursus arctos*. All pairwise F_{ST} comparisons indicated differentiation between subpopulations, as did the AT results.

A major shortcoming of ATs for evaluating the magnitude of population differentiation is that the probability of correct assignment depends on not only the degree of population differentiation, but also the sample sizes of individuals and loci, and their level of polymorphism [12,19]. As a result, it is difficult to establish a general scale for evaluating results. Standard measures of population differentiation (e.g. genetic distances and F_{ST}), which are less affected by these factors, thus remain a better way of quantifying levels and patterns of genetic differentiation. Within a given data set, however, the proportion of individuals correctly assigned to each population can provide useful insights regarding the relative patterns of population genetic structure.

Dispersal

What is the rate of dispersal or gene flow? Many researchers want to study dispersal at current (ecological) timescales, and AMs have the potential to provide this real-time information. The original AT [1] has been refined to develop a statistical framework for identifying individuals with immigrant ancestry up to two generations in the past [4] (Box 1). The clustering method of Pritchard *et al.* [17] can also be used as a classification method to identify immigrants. A study of dispersal in the grand skink *Oligosoma grande* in New Zealand [7] demonstrated the high accuracy of the methods of Rannala and Mountain [4] and Pritchard *et al.* [17] in identifying dispersers with known natal origin.

The methods of Rannala and Mountain [4] and Pritchard *et al.* [17] do not explicitly estimate migration rates; however, a rough point estimate can be obtained by dividing the number of individuals identified as migrants by the sample size. A recently developed method [30] is specifically designed to estimate immigration rates; moreover, it relaxes the assumption of Hardy–Weinberg equilibrium made by all previous methods, by estimating a separate inbreeding coefficient for each population. This method provides measures of uncertainty about the migration rate estimates and, therefore, is preferable to the point estimate approach.

An important limitation of AMs is that their power to detect migrants is greatest when populations are genetically divergent, but, in that case, gene flow is generally rare and even large samples might not detect a migration event. A recent empirical evaluation [6] suggests that low levels of population differentiation can be counteracted to some extent by larger samples and additional gene loci, resulting (in some cases at least) in reasonable power to detect migrants.

What are the patterns of dispersal and gene flow? Once identified, migrants can be grouped according to traits (sex, age, dispersal distance [31], etc.) to provide information about evolutionary and conservation relevance. In a study of the shrew *Crocidura russula*, Favre *et al.* [32] calculated a sex-specific assignment index and found that it was significantly lower in females than in males, suggesting female-biased dispersal, which is an unusual pattern in mammals. A recent analysis [33] shows that the assignment index performs best when dispersal is very low ($<10\%$). In other cases, information from maternally inherited mitochondrial DNA and Y-chromosome markers can provide insights into sex-biased dispersal [34].

Genetic mixture and admixture analysis

Sometimes, the interest is not in individual assignments *per se* but rather in the overall composition of a mixture (of individuals from different populations) or an admixture (which results from interbreeding among populations). The linkage disequilibrium generated by the mixing or admixing of individuals from different populations can be used to address two related questions.

What proportion of individuals in a mixture come from each source population? This question is a central one for fisheries management (Box 2), but it also arises in the study of natural processes, such as colonization [35]. In the northwestern USA, genetic stock identification (Box 2) has been used to help manage populations of endangered Chinook salmon *Oncorhynchus tshawytscha* from the upper Columbia and Snake Rivers that are harvested in mixed-stock fisheries in the lower river [36]. Weekly samples can be processed rapidly to provide real-time estimates of stock composition of the fishery, which targets relatively abundant, early-returning fish from the nearby Willamette River. The fishery can be closed when genetic results indicate that endangered upriver populations begin to appear in the harvest.

What proportion of genes in an admixture come from each source? Analysis of admixture is common, particularly for human populations. For recent admixture (within a few generations), residual linkage disequilibrium revealed in multilocus genotypes can provide a means of identifying admixed individuals. For example, Beaumont *et al.* [37] used admixture analysis [17] to clarify the status of wildcat *Felis sylvestris* populations in Scotland. They found strong evidence for a unique group that is different from domestic cats and that probably represents remnants of the native gene pool.

If loci are physically linked, insights might be possible for older admixture events, provided the recombination

rate is known [15]. However, if admixture is old enough that recombination has obscured parental gene associations, single-locus models are more appropriate for obtaining information about overall admixture proportions [38]. More recently, several coalescent methods have been proposed to take advantage of genealogical information [39–41].

Hybridization

Identification of hybrid individuals is often a necessary first step in the implementation of management strategies, such as breeding or translocation programs for threatened species, [42,43], and the standard AT has been applied to this problem [44]. A hypothetical hybrid taxon is created by combining randomly sampled alleles from the two parental taxa, and each individual is then assigned to one of three potential sources (two parental taxa and the hybrid taxon). This somewhat *ad hoc* process enables one to identify only first-generation hybrids. A modification [4] performs significance tests for each individual and each degree of relationship (up to the F_2 generation), but does not provide an overall significance for multiple individuals.

When parental taxa are not defined or characterized *a priori*, clustering methods must be used. The program STRUCTURE [15,17] assigns gene copies probabilistically to potential sources; individuals with genes assigned with non-trivial probabilities to two sources are potential hybrids. This method has been used to study hybrids in the wild for several species [45,46]; for example, Randi and Lucchini [45] used it to confirm the introgression of domestic dog genes into the Italian wolf. Although ordination and tree-based methods could not detect introgression, STRUCTURE clearly identified as a hybrid one wolf that was also a suspected hybrid based on morphology (unusually dark fur).

Anderson and Thompson [47] developed a method designed specifically to detect hybrids. Unlike methods that estimate the proportion of an individual genome that originated from each taxon [15,17], this approach distinguishes various hybrid classes (F_1, F_2 and various backcrosses). To date, performance of the two types of method has not been compared. Both are appropriate for identifying purebred individuals; however, if distinguishing various hybrid classes is important, the method of Anderson and Thompson [47] would be more useful.

Conclusions and future directions

AMs have already demonstrated an impressive capacity to provide insights into contemporary ecological and evolutionary processes. Nevertheless, they share with other contemporary methods, such as mark–recapture, the limitation that (for example) migration rates observed during a short study might not accurately reflect long-term patterns of gene flow. In general, therefore, AMs should be viewed as an important complement to, rather than a replacement of, equilibrium models.

In spite of some recent efforts to evaluate the power and sensitivity of AMs (e.g. [6,12]), comparative analyses of performance of the various methods are lacking in most cases. Consequently, it is currently often not possible to say with certainty which of various competing methods performs best and under which conditions. A variety of other problems also merit attention, including:

What is a population? Without consistent criteria for how different gene pools must be to be considered 'populations,' it will be difficult to achieve clarity about the number of population units and relationships among them.

How well do clustering methods perform when genetic differentiation is modest ($F_{ST} \leq 0.05$)?

Under what realistic conditions can AMs detect contemporary migration, given that power is highest when migration events are rare and, therefore, unlikely to be observed?

How do unsampled sources affect analyses based on ATs?

How reliably can AMs detect hybrids and various backcrosses beyond the F_1 generation?

What are the effects of selection, linked loci and genotyping errors on results obtained using AMs?

This list of potential research topics indicates that the next decade should produce new developments as exciting as those of the recent past.

Acknowledgements

We thank Peter Smouse, Jerry Pella and Eric Anderson for stimulating discussions about this topic, and Fred Allendorf, Michael Hansen, Steven Kalinowski, Gordon Luikart, Craig Moritz, Per Palsbøll, David Tallmon and three anonymous reviewers for useful comments. This work was conducted while O.E.G. and S.M. were supported by grant ACI 2004-42-PGDA (from Fond national de la science) and R.S.W. was a visiting scientist at LECA.

References

1 Paetkau, D. *et al*. (1995) Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.* 4, 347–354
2 Waser, P.M. and Strobeck, C. (1998) Genetic signatures of interpopulation dispersal. *Trends Ecol. Evol.* 13, 43–44
3 Davies, N. *et al*. (1999) Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics. *Trends Ecol. Evol.* 14, 17–21
4 Rannala, B. and Mountain, J.L. (1997) Detecting immigration by using multilocus genotypes. *Proc. Natl. Acad. Sci. U. S. A.* 94, 9197–9201
5 Hansen, M.M. *et al*. (2001) Assigning individual fish to populations using microsatellite DNA markers. *Fish Fish.* 2, 93–112
6 Paetkau, D. *et al*. (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Mol. Ecol.* 13, 55–65
7 Berry, O. *et al*. (2004) Can assignment tests measure dispersal? *Mol. Ecol.* 13, 551–561
8 Seber, G.A.F. (1984) *Multivariate Observations*, John Wiley & Sons
9 Nei, M. and Li, W-H. (1973) Linkage disequilibrium in subdivided populations. *Genetics* 75, 213–219
10 Makela, M.E. and Richardson, R.H. (1976) The detection of sympatric sibling species using genetic correlation analysis. I. Two loci, two gamodemes. *Genetics* 86, 665–678
11 Waples, R.S. and Smouse, P.E. (1990) Gametic disequilibrium analysis as a means of identifying mixtures of salmon populations. *Am. Fish. Soc. Symp.* 7, 439–458
12 Cornuet, J-M. *et al*. (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* 153, 1989–2000
13 Corander, J. *et al*. (2003) Bayesian analysis of genetic differentiation between populations. *Genetics* 163, 367–374
14 Dawson, K.J. and Belkhir, K. (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res.* 78, 59–77

15 Falush, D. *et al*. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587
16 Gaggiotti, O.E. *et al*. (2004) Combining demographic, environmental and genetic data to test hypothesis about colonization events in metapopulations. *Mol. Ecol.* 13, 811–825
17 Pritchard, J.K. *et al*. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959
18 Pella, J. and Masuda, M. (2001) Bayesian methods for analysis of stock mixtures from genetic characters. *Fish. Bull.* 99, 151–167
19 Olsen, J.B. *et al*. (2000) Microsatellites reveal population identity of individual pink salmon to allow supportive breeding of a population at risk of extinction. *Trans. Am. Fish. Soc.* 129, 232–242
20 Hedgecock, D. *et al*. (2001) Applications of population genetics to conservation of chinook salmon diversity in the Central Valley. In *Fish Bulletin 179: Contributions to the Biology of Central Valley Salmonids* (Brown, R.L., ed.), pp. 45–69, California Department of Fish and Game
21 Bonizzoni, M. *et al*. (2001) Microsatellite analysis of medfly bioinfestations in California. *Mol. Ecol.* 10, 2515–2524
22 Primmer, C.R. *et al*. (2000) The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. *Proc. R. Soc. Lond. Ser. B* 267, 1699–1704
23 Raymond, M. and Rousset, F. (1995) An exact test for population differentiation. *Evolution* 49, 1280–1283
24 Cegelski, C.C. *et al*. (2003) Assessing population structure and gene flow in Montana wolverines (*Gulo gulo*) using assignment-based approaches. *Mol. Ecol.* 12, 2907–2918
25 Manni, F. *et al*. (2004) Geographic patterns of genetic, morphologic and linguistic variation: how barriers can be detected by using Monmonier's algorithm. *Hum. Biol.* 76, 173–190
26 Dupanloup, I. *et al*. (2002) A simulated annealing approach to define the genetic structure of populations. *Mol. Ecol.* 11, 2571–2581
27 Epperson, B.K. (2003) *Geographical Genetics*, Princeton University Press
28 Waits, L. *et al*. (2000) Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (*Ursus arctos*). *Mol. Ecol.* 4, 421–431
29 McLoughlin, P.D. *et al*. (2004) Genetic diversity and relatedness of boreal caribou populations in western Canada. *Biol. Conserv.* 118, 593–598
30 Wilson, G.A. and Rannala, B. (2003) Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* 163, 1177–1191
31 Castric, V. and Bernatchez, L. (2004) Individual assignment test reveals differential restriction to dispersal between two salmonids despite no increase of genetic differences with distance. *Mol. Ecol.* 13, 1299–1312
32 Favre, L. *et al*. (1997) Female-biased dispersal in the monogamous mammal *Crocidura russula*: evidence from field data and microsatellite patterns. *Proc. R. Soc. Lond. Ser. B* 264, 127–132
33 Goudet, J. *et al*. (2002) Tests for sex-biased dispersal using bi-parentally inherited genetic markers. *Mol. Ecol.* 11, 1103–1114
34 Prugnolle, F. and Meeus, T. (2002) Inferring sex-biased dispersal from population genetic tools: a review. *Heredity* 88, 161–165
35 Gaggiotti, O.E. *et al*. (2002) Patterns of colonization in a grey seal metapopulation. *Nature* 416, 424–427
36 Shaklee, J.B. *et al*. (1999) Managing fisheries using genetic data: case studies from four species of Pacific Salmon. *Fish. Res.* 43, 45–78
37 Beaumont, M. *et al*. (2001) Genetic diversity and introgression in the Scottish wildcat. *Mol. Ecol.* 10, 319–336
38 Thompson, E.A. (1973) The Icelandic admixture problem. *Ann. Hum. Genet.* 37, 69–80
39 Wang, J. (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* 164, 747–765
40 Chikhi, L. *et al*. (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* 158, 1347–1362
41 Bertorelle, G. and Excoffier, L. (1998) Inferring admixture proportions from molecular data. *Mol. Biol. Evol.* 15, 1298–1311
42 Allendorf, F.W. *et al*. (2004) Intercrosses and the U.S. Endangered Species Act: should hybridized populations be included as westslope cutthroat trout? *Conserv. Biol.* 18, 1203–1213
43 Campton, D.E. (1987) Natural hybridization and introgression in fishes: methods of detection and genetic interpretations. In *Population Genetics & Fishery Management* (Ryman, N. and Utter, F., eds), pp. 161–192, University of Washington Press
44 Congiu, L. *et al*. (2001) Identification of interspecific hybrids by amplified fragment length polymorphism: the case of sturgeon. *Mol. Ecol.* 10, 2355–2359
45 Randi, E. and Lucchini, V. (2002) Detecting rare introgression of domestic dog genes into wild wolf (*Canis lupus*) populations by Bayesian admixture analysis of microsatellite variation. *Conserv. Genet.* 3, 31–45
46 Flammand, J. *et al*. (2003) Genetic identification of wild Asian water buffalo in Nepal. *Anim. Conserv.* 6, 265–270
47 Anderson, E.C. and Thompson, E.A. (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* 160, 1217–1229
48 Duchesne, P. and Bernatchez, L. (2002) AFLPOP: a computer program for simulated and real population allocation, based on AFLP data. *Mol. Ecol. Notes* 2, 380–383
49 Pella, J.J. and Milner, G.B. (1987) Use of genetic marks in stock composition analysis. In *Population Genetics & Fishery Management* (Ryman, N. and Utter, F., eds), pp. 247–276, University of Washington Press
50 Smouse, P.E. *et al*. (1990) A genetic mixture analysis for use with incomplete source population data. *Can. J. Fish. Aquat. Sci.* 47, 620–634
51 Banks, M.A. and Eichert, W. (2000) WHICHRUN (version 3.2): a computer program for population assignment of individuals based on multilocus genotype data. *J. Hered.* 91, 87–89
52 Masuda, M. and Pella, J. (2004) Identification of source populations of mixture individuals from their genotypes. *NPAFC Technical Report No. 5*, 103–105 (http://www.npafc.org)
53 Potvin, C. and Bernatchez, L. (2001) Lacustrine spatial distribution of landlocked Atlantic salmon populations assessed across generations by multilocus individual assignment and mixed–stock analyses. *Mol. Ecol.* 10, 2375–2388
54 Grant, W.S. *et al*. (1980) Use of biochemical genetic variants for identification of sockeye salmon (*Oncorhynchus nerka*) stocks in Cook Inlet, Alaska. *Can. J. Fish. Aquat. Sci* 37, 1236–1247
55 Pella, J. and Masuda, M. (2005) Classical discriminant analysis, classification of individuals, and source population composition of mixtures. In *Stock Identification Methods: Applications in Fishery Science* (Cadrin, S. *et al.*, eds), pp. 517–522, Academic Press
56 Debevec, E.M. *et al*. (2000) SPAM (Version 3.2): Statistics Program for Analysing Mixtures. *J. Hered.* 91, 509–510
57 Avise, J.C. *et al*. (2002) Genetic mating systems and reproductive natural histories of fishes: lessons for ecology and evolution. *Annu. Rev. Genet.* 36, 19–45
58 Jones, A.G. and Ardren, W.R. (2003) Methods of parentage analysis in natural populations. *Mol. Ecol.* 12, 2511–2523
59 Burczyk, J. and Chybicky, I.J. (2004) Cautions on direct gene flow estimation in plant populations. *Evolution* 5, 956–963
60 Nielsen, R. *et al*. (2001) Statistical approaches to paternity analysis in natural populations and applications to the Northern Atlantic humpback whale. *Genetics* 157, 1673–1682
61 Emery, A.M. *et al*. (2001) Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Mol. Ecol.* 10, 1265–1278
62 Blouin, M.S. (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol. Evol.* 18, 503–511
63 Seamons, T.R. *et al*. (2004) The effects of adult length and arrival date on individual reproductive success in wild steelhead trout (*Oncorhynchus mykiss*). *Can. J. Fish. Aquat. Sci.* 61, 193–204
64 Flemming, I.A. *et al*. (2000) Lifetime success and interactions of farm salmon invading a native population. *Proc. R. Soc. Lond. Ser. B* 267, 1517–1523
65 Morgan, M.T. and Connor, J.K. (2001) Using genetic markers to directly estimate male selection gradients. *Evolution* 55, 272–281
66 Smouse, P.E. and Sork, V.L. (2004) Measuring pollen flow in forest trees: a comparison of alternative approaches. *For. Ecol. Manage.* 197, 21–38
67 Vines, T.H. and Barton, N.H. (2003) A new approach to detecting mixed families. *Mol. Ecol.* 12, 1999–2002