2

Statistical methods for identifying hybrids and groups

ERIC C. ANDERSON

INTRODUCTION

Recently, statistical geneticists have developed a number of model-based methods that use genetic data to infer the population of origin of the gene copies within an individual. In this chapter we focus on three of these methods which are known by the software that implements them: STRUCTURE (Pritchard *et al.* 2000), NEWHYBRIDS (Anderson and Thompson 2002) and BAYESASS+ (Wilson and Rannala 2003). These programs are increasingly used in animal conservation for population assignment, detection of hybridization and estimation of recent migration rates. Unlike more generic statistical approaches (Bowcock *et al.* 1994; Roques *et al.* 2001), the three methods we review here are all based on an underlying probability model that is intended to mimic the inheritance of genes and the sampling of individuals. Such model-based inference has a number of advantages. First, it typically uses more of the information in the data than approaches that are not based explicitly on genetic models, and second, the variables appearing in genetically based statistical models relate directly to genetic phenomena, so they are easily interpreted.

The statistical genetic models underlying STRUCTURE, NEWHYBRIDS and BAYESASS+ are simple and quite similar. The primary goal of this chapter is to describe these models with as few equations as possible. In lieu of mathematical equations we will explore the structure of these models in terms of simple, intuitive diagrams called directed acyclic graphs (DAGs), which show the relationship between variables in a model. This should allow users to better understand what the methods do, how they are similar, and the important ways in which they differ. Though the softwares implementing these techniques are user-friendly, they are certainly not 'plug-and-play' methods. I hope that this chapter will allow users to understand the

Population Genetics for Animal Conservation, eds. G. Bertorelle, M. W. Bruford, H. C. Hauffe, A. Rizzoli and C. Vernesi. Published by Cambridge University Press. © Cambridge University Press 2009.

methods enough to ensure they get reasonable results and they can interpret them appropriately.

After discussing the three different models, we focus on practical issues. Because previous reviews (Pearse and Crandall 2004; Manel *et al.* 2005) have summarized when these various methods (and many related ones, e.g. Rannala and Mountain 1997; Dawson and Belkhir 2001; Corander *et al.* 2004; Piry *et al.* 2004) are useful, and have offered many general guidelines for their use, this final section is devoted to the simple proposition that it is important to assess the results of these programs by comparison to simulated data that look like your own.

CONCEPTUAL MODELS AND GRAPHICAL MODELS

All statistical inference depends in some way on a probability model. This model may be completely specified in terms of the equations describing the statistical distributions involved; though if you simply want to understand the assumptions of the model, it is usually sufficient to understand the verbal description of the model. As a model gets more complex, however, it is helpful to have a visual roadmap as well as a verbal description. A DAG is such a roadmap, providing a diagram of the relationship between components in a model and a comparison of the structure of different models. We will illustrate our first DAG by considering the estimation of allele frequencies in a closed population.

Let us imagine that we are interested in estimating the frequency of alleles at a single locus in a lake population of fish. An obvious course of action would be to draw a sample of *M* fish from the lake, genotype them at the locus, and estimate the allele frequencies from the observed proportion of alleles in the sample. The conceptual model underlying this procedure is one in which each fish carries two gene copies drawn at random from a large pool of alleles whose proportions are the unknown allele frequencies in the lake.

The relationship between the allelic types in the fish we sample and the population allele frequencies is captured in the DAG of Fig. 2.1a. The circles (called nodes) in the graph represent the different variables in the model. The frequencies of the alleles are denoted by $\boldsymbol{\theta}$. The node associated with $\boldsymbol{\theta}$ is unshaded, representing the fact that the values of the population allele frequencies are unknown. The type of the allele is denoted by $Y_{i,1}$ for the first gene copy of the *i*th sampled fish, and $Y_{i,2}$ for the second gene copy. The nodes associated with these variables are shaded black to denote that they are observed – i.e. the fish are genotyped. The allelic type of each gene they can inter-

›ractical issues.
ıel *et al.* 2005)
›lated ones, e.g.
Corander *et al.*
ıeral guidelines
›sition that it is
ırison to simu-

)DELS

ity model. This
; describing the
: to understand
ınderstand the
ılex, however, it
:iption. A DAG
ı between com-
fferent models.
ıation of allele

.e frequency of
vious course of
notype them at
:ved proportion
ıis procedure is
:andom from a
ele frequencies

sample and the
2.1a. The circles
s in the model.
sociated with $\boldsymbol{\theta}$
›pulation allele
l by $Y_{i,1}$ for the
ond gene copy.
: to denote that
pe of each gene

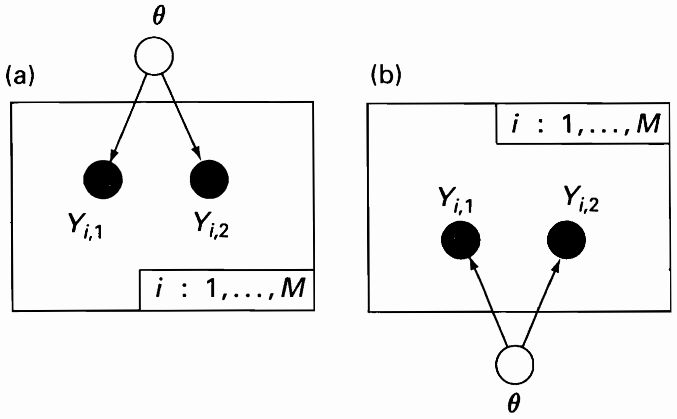


Figure 2.1. DAGs describing the sampling to estimate allele frequency $\boldsymbol{\theta}$: (a) and (b) describe identical models.

copy is independent (under the assumption of Hardy–Weinberg equilibrium) and depends only on the frequency of alleles in the population. Hence, there are distinct arrows drawn from the node at $\boldsymbol{\theta}$ to the nodes for $Y_{i,1}$ and $Y_{i,2}$. The meaning of the arrow can be read as, for example, 'the allelic type, $Y_{i,1}$, depends on $\boldsymbol{\theta}$'. Finally, the two Y nodes are placed inside a box which is known as a *plate* (or, in this case, an M-plate). The legend at the lower right of the plate indicates that the variables within the plate are duplicated M times, over the subscript i. This shorthand expresses that M fish are sampled independently from the lake. The node for $\boldsymbol{\theta}$ is not included on the plate because each of the M fish is assumed to be sampled from the same population with the same allele frequencies. Figure 2.1b represents exactly the same model. This figure is included to emphasize that the spatial position of variables in the graph is unimportant; only the orientation of the arrows, and the connections they make, are relevant.

Recall that the original problem was to estimate the allele frequencies, $\boldsymbol{\theta}$, given the observed genotypes of a sample of fish. This problem is also apparent in the DAG because $\boldsymbol{\theta}$ is something we wish to know about, and yet it is unknown (as signified by its unshaded node). Generally, the problem of estimation can be interpreted in a DAG as the process of learning about variables or parameters with unshaded nodes given what is observed in the data (the shaded nodes).

As a final word on Fig. 2.1, we should keep in mind that we would have obtained the same DAG if we were sampling any objects, two at a time, from a large population of objects. In fact, it is often easier to think of the sampling process as that of randomly drawing coloured balls, two at a time, out of a large barrel. In this case, each ball is a gene copy, its colour is its allelic type, and the barrel full of balls is the population of gene copies carried by fish in the lake. We have explored this example in detail because the 'balls-in-barrels' conceptual model, and the DAG that goes with it, are basic building blocks for understanding more complex models. In the next section we use these building blocks to describe a class of models called mixture models.

MIXTURE MODELS

The problem recently called 'population assignment' in the molecular ecology literature is a special case of inference in a finite mixture model. In statistics, a finite mixture model is one in which the collection from which the sample is taken is a mixture of individuals from different populations. Such models were applied to the problem of population assignment and 'genetic stock identification' as early as 1981 in the fisheries management literature (Milner *et al.* 1981). The programs STRUCTURE, NEWHYBRIDS and BAYESASS+ are all elaborations of the basic mixture model. In fact, the version of STRUCTURE 'without admixture' employs the same mixture model as an earlier method used to estimate proportions of Columbia River tributary salmon caught in a mixed stock fishery (Smouse *et al.* 1990).

This salmon-fishery mixture model arises from a scenario such as the following: K separate spawning populations of salmon, each with its own unknown allele frequencies, reproduce in different tributaries of a river. Fish from each population migrate through the same place (for example, the mouth of the river), where they are subject to a fishery. By sampling M fish in the fishery and genotyping them at L loci we hope to estimate the proportion of fish from each of the K tributaries that were at the fishery site when the sample was taken. We might also want to infer the population of origin of each of the sampled fish.

Figure 2.2(a) shows the DAG for the mixture model described above. This DAG is composed of a number of elements that look suspiciously like the DAG of Fig. 2.1. Working our way through the graph, from top to bottom, we first have $\boldsymbol{\pi}$, which denotes the unknown proportions of fish from the K different populations at the fishery site. W_i is a variable that denotes the population of origin of the ith fish. It can be thought of as a ball that is tied to the fin of the fish, with the colour of the ball telling us where the fish comes from. Under this interpretation, each fish is sampled from the fishery as if it were a coloured ball drawn from a barrel in which the different colours of balls are in the unknown proportions $\boldsymbol{\pi}$. Of course, the node associated with W_i is unshaded because we don't know where the fish

(a)

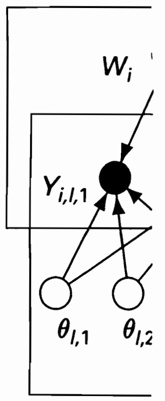


Figure 2.2. D
(b) includes n
specification

come from – th
looks complica
lower of the tw
fish, there are L
given the fish's
at locus ℓ in pop
and $Y_{i,\ell,2}$ are the
the arrows in the
on W_i and on tl
of this depende
drawn from the

The whole s
erating a sampl
a barrel with f
population to sa
you draw two ba
Each barrel rep

As before, tl
can be seen in
et al. 1995; Ran
values of the W
of fish from d
value of $\boldsymbol{\pi}$. Fina

detail because
ıes with it, are
ꞏls. In the next
models called

he molecular
ixture model.
ıllection from
different pop-
lation assign-
the fisheries
S STRUCTURE,
ɔasic mixture
' employs the
ıroportions of
ıery (Smouse

o such as the
with its own
es of a river.
(for example,
' sampling M
estimate the
ıe fishery site
ɔopulation of

:ribed above.
piciously like
from top to
:tions of fish
variable that
ht of as a ball
ing us where
ımpled from
in which the
f course, the
here the fish

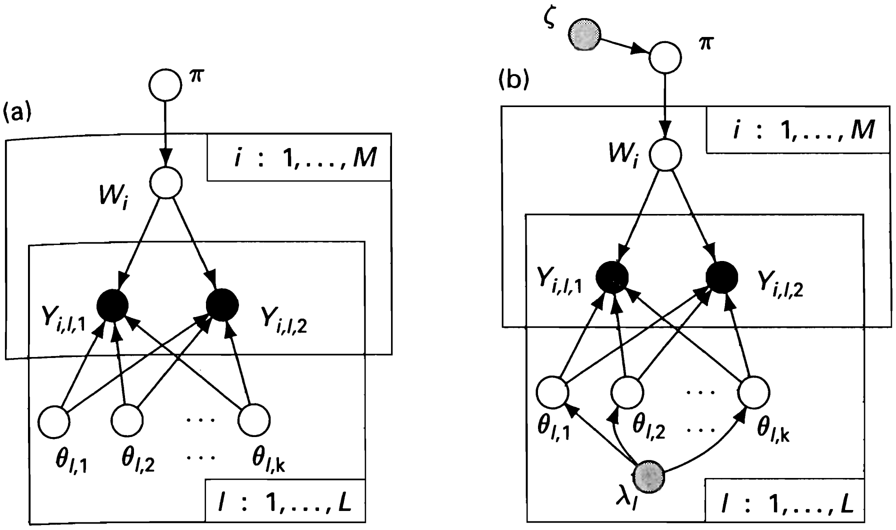


Figure 2.2. DAGs for the mixture model: (a) represents the likelihood model, (b) includes nodes associated with the prior distributions for a Bayesian specification of the problem.

come from – that is what we would like to learn. The remainder of the graph looks complicated, but we can break it down as follows: the L-plate (the lower of the two plates, with the legend 'ℓ : 1, ..., L') signifies that for each fish, there are L loci genotyped, and that their allelic types are independent, given the fish's population of origin, W_i. $\boldsymbol{\theta}_{\ell,k}$ is the frequency of alleles at locus ℓ in population k (where k denotes any one of the K populations). $Y_{i,\ell,1}$ and $Y_{i,\ell,2}$ are the allelic types of the gene copies carried by fish i at locus ℓ. As the arrows in the graph show, the allelic types of these gene copies depend *both* on W_i and on the allele frequencies in the different populations. The nature of this dependence is straightforward: the two allelic types $Y_{i,\ell,1}$ and $Y_{i,\ell,2}$ are drawn from the population that the fish is from – which is denoted by W_i.

The whole sampling model can be summarized by thinking about generating a sample from it. The steps in doing so would be: (1) Draw a ball from a barrel with frequencies $\boldsymbol{\pi}$. (2) The colour of the ball tells you which population to sample a fish from. (3) To generate the genotype for that fish you draw two balls, representing the genes in that fish, from each of L barrels. Each barrel represents the alleles in the population at one of the L loci.

As before, the inference problems that can be tackled with this model can be seen in the DAG. The exercise of population assignment (Paetkau *et al.* 1995; Rannala and Mountain 1997) is merely that of inferring the values of the W_i variables. On the other hand, estimating the proportion of fish from different populations is just the process of inferring the value of $\boldsymbol{\pi}$. Finally, if desired, one could also pursue inference of the allele frequencies in the populations. These are all inference problems that are just different uses of the same underlying model. In actual practice, these different inference problems are typically tackled at the same time, but it is still useful to view them as separate inference problems.

Many times, individuals known to be from particular populations may be sampled. Such individuals constitute what are called learning samples or training samples. These would be represented in the DAG simply as individuals for whom the node associated with W_i was shaded. Inference then proceeds much as before – unknown quantities of interest are estimated given the observed data, which in this case includes the W_is of the individuals in the learning samples. Though with multiple loci mixture inference may be possible without learning samples, if there are many populations contributing to the mixture then accurate inference may be impossible without learning samples.

BAYESIAN INFERENCE

STRUCTURE, NEWHYBRIDS and BAYESASS+ all use the Bayesian paradigm for inferring quantities of interest. This means that estimation is conducted by summarizing the posterior distribution of quantities of interest. The posterior distribution of an unknown variable is just its probability distribution conditional on the observed data. Computing the posterior distribution can be difficult, and, indeed, in STRUCTURE, NEWHYBRIDS and BAYESASS+ it is approximated using Markov chain Monte Carlo. However, the fact that the inference is done in a Bayesian manner does not substantially alter the structure of the underlying models. This is illustrated in Fig. 2.2b, which shows the DAG for a Bayesian specification of the mixture model of Fig. 2.2a. It is apparent that the 'heart' of the model is unchanged. In fact, the only modification is the addition of prior distributions parametrized by ζ for $\boldsymbol{\pi}$ and λ_ℓ for the $\boldsymbol{\theta}_\ell$s. The nodes for ζ and λ_ℓ are shaded grey to denote that values of those parameters are assumed rather than observed. Prior distributions are necessary for Bayesian inference. Usually the parameters of the prior distribution are chosen to reflect prior knowledge – or in many cases, ignorance – about the associated variables.

A SURVEY OF METHODS

Having established the language of graphical models, we are now in position to quickly survey the models used in STRUCTURE, NEWHYBRIDS and BAYESASS+.

The STRUCTURE model without admixture

As indicated above, the STRUCTURE model without admixture is identical to the model shown in Fig. 2.2(b), and the details of that model have already been described. It assumes that all individuals descend exclusively from one of K populations, where K can be set by the user. In other words, there is no facility in this model for explicitly dealing with hybrids or admixed individuals. Therefore, the method should be used with collections of organisms that are believed to be non-interbreeding. The data required are the multilocus genotypes of the individuals in a sample. The individuals may belong to 'cryptic' subpopulations. That is, it is not necessary to have prior knowledge of separate groups – the program will automatically infer K subpopulations; however, the inclusion of learning samples can be helpful in resolving groups, especially if K is large, or genetic differentiation between populations is limited. The program computes the posterior probability that each individual belongs to each of the K subpopulations, and, in the process it also estimates the allele frequencies in the K separate subpopulations.

It is worth noting that when STRUCTURE uses the model with no admixture, it assumes that the proportion of individuals from each subpopulation is equal (each subpopulation contributes a proportion $1/K$ to the mixture). This feature will cause STRUCTURE to overestimate the true posterior probability of group membership for individuals from subpopulations that are rare in the mixture. If this is a concern, then it may be preferable to use the program *BAYES* (Pella and Masuda 2001) which was developed for analysing large mixtures of salmon.

The STRUCTURE model with admixture

This model provides a flexible way of accommodating individuals of mixed ancestry. No longer must each individual be purely descended from one of the K subpopulations. Rather, each gene copy within an individual may come from a different one of the K subpopulations. The subpopulations of origin of the two gene copies at locus ℓ in the ith individual are indicated by the unobserved variables $W_{i,\ell,1}$ and $W_{i,\ell,2}$, and the expected proportion of ancestry of the ith individual from each of the K subpopulations is a variable to be inferred, denoted by Q_i. The DAG for this model appears in Fig. 2.3a. Here, α is a parameter that determines whether individuals tend to be mostly admixed (high values of α) or mostly purebred (low values of α). It is a value that can be assumed, or inferred. If it is inferred, its prior distribution is assumed to be uniform on the interval $(0, A)$.

We can use the DAG to follow how we would generate data under the model, given α and the allele frequencies: (1) Conditional on α we would

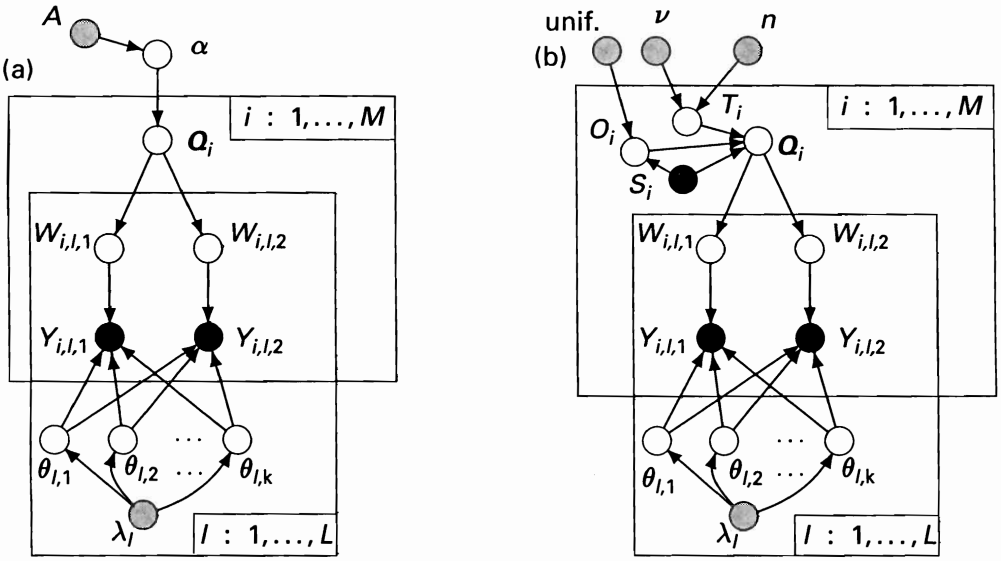


Figure 2.3. (a) The *structure* model with admixture. (b) The *structure* model with admixture *and* prior population information.

simulate a different, random Q_i for each individual i in the sample. Q_i can be thought of as the proportion of balls of K different colours filling a 'Q-barrel' associated with individual i. (2) For each locus, we would draw two balls from individual i's Q-barrel. The colours of the balls drawn tell us which of the K different subpopulations the two gene copies at each locus came from. (3) The allelic type of each gene copy would then be drawn from the allele frequencies in the gene copy's subpopulation of origin.

This is a flexible and general model. It applies generically to many different scenarios: estimating the hybrid index (i.e. Q_i) of individuals in hybrid zones, detecting recent gene flow between populations, and elucidating population structure (cryptic or otherwise). It also provides a facility for estimating the number of subpopulations in a structured population, without prior knowledge about population boundaries.

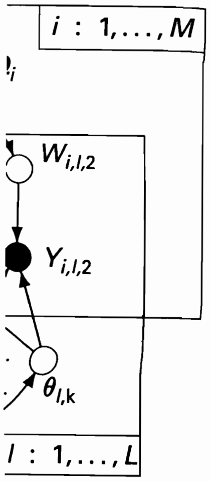
The data required are the multilocus genotypes of sampled individuals. Learning samples are not required, so it is possible to identify cryptic genetic population structure in a sample of individuals from a single location. However, the capacity to detect cryptic structure declines as the degree of admixture of the individuals in the sample increases (Falush *et al.* 2003). In other words, if most individuals in the sample are highly admixed members of a hybrid swarm, it will be more difficult to correctly infer the nature of the population structure than if some of the individuals in the sample retain the genotypes of pure subspecies, and others are admixed.

The STRUCTU
population i
A variant avail
information' ir
known, separa
als in each sar
have recent im
knowledge that
been sampled.
living in a part
is flexible. For
the basis of dis
subspecies.

Figure 2.4a
subpopulations
tion between th
between all sub
probability 1 − ν
subpopulation
individual has i
must be assum

If individua
ancestor in the
ancestor arrive
pling. If $T_i = 0$
then one of ind
four grandpare
unknown. We
individual was

The DAG i
information is
parts 'upstrean
and prior popu
interpreted, pri
ν and n. The arr
ν and n determ
at time T_i; 2)
that migrant d
somewhere *oth*
determined (Fi



ure model with

he sample. Q_i
olours filling a
ve would draw
s drawn tell us
s at each locus
be drawn from
igin.
cally to many
individuals in
ns, and eluci-
vides a facility
ed population,

pled individu-
dentify cryptic
from a single
eclines as the
eases (Falush
ple are highly
ult to correctly
ne individuals
nd others are

The STRUCTURE model with admixture and prior population information

A variant available with STRUCTURE is the model with 'prior population information' in which genotyped individuals have been sampled from K known, separate subpopulations. This model is used to identify individuals in each sample that are migrants from other subpopulations or that have recent immigrant ancestry. In this case, it is necessary to have prior knowledge that there are distinct subpopulations, and that K of them have been sampled. A subpopulation is typically comprised of individuals living in a particular locality; however, the definition of 'subpopulation' is flexible. For example, one might be able to define K subpopulations on the basis of distinct morphological traits possessed by different species or subspecies.

Figure 2.4a is a schematic of the population model in the case of $K=3$ subpopulations of cats. The subpopulations are distinct, but there is migration between them. Immigration is assumed to be symmetrical and equal between all subpopulations. The model specifies that each individual has a probability $1-v$ of being descended purely from ancestors belonging to the subpopulation from which it was sampled. With probability v, however, an individual has immigrant ancestry. If v is unknown (as it usually is) then it must be assumed.

If individual i has immigrant ancestry, then it is assumed that only one ancestor in the last n generations was a migrant, and that this migrant ancestor arrived from subpopulation O_i in the T_i^{th} generation before sampling. If $T_i=0$ then the sampled individual i is itself the migrant; if $T_i=1$ then one of individual i's two parents was a migrant; if $T_i=2$ then one of i's four grandparents was a migrant, and so forth (Fig. 2.4c). O_i and T_i are unknown. We will let S_i denote the subpopulation from which the ith individual was sampled; S_i is an observed variable.

The DAG in Fig. 2.3b shows that this model with prior population information is identical to the original STRUCTURE model, except for the parts 'upstream' from the Q_i node. In effect, the model with admixture and prior population information just establishes a new, and more easily interpreted, prior probability distribution for Q_i that ultimately depends on v and n. The arrows in the DAG appear as they do because 1) the parameters v and n determine the probability that an individual has a migrant ancestor at time T_i; 2) if individual i has a migrant ancestor, then the origin of that migrant depends on S_i because the migrant must have come from somewhere *other* than S_i; and finally 3) given T_i, S_i and O_i, the value Q_i is determined (Fig. 2.4c).

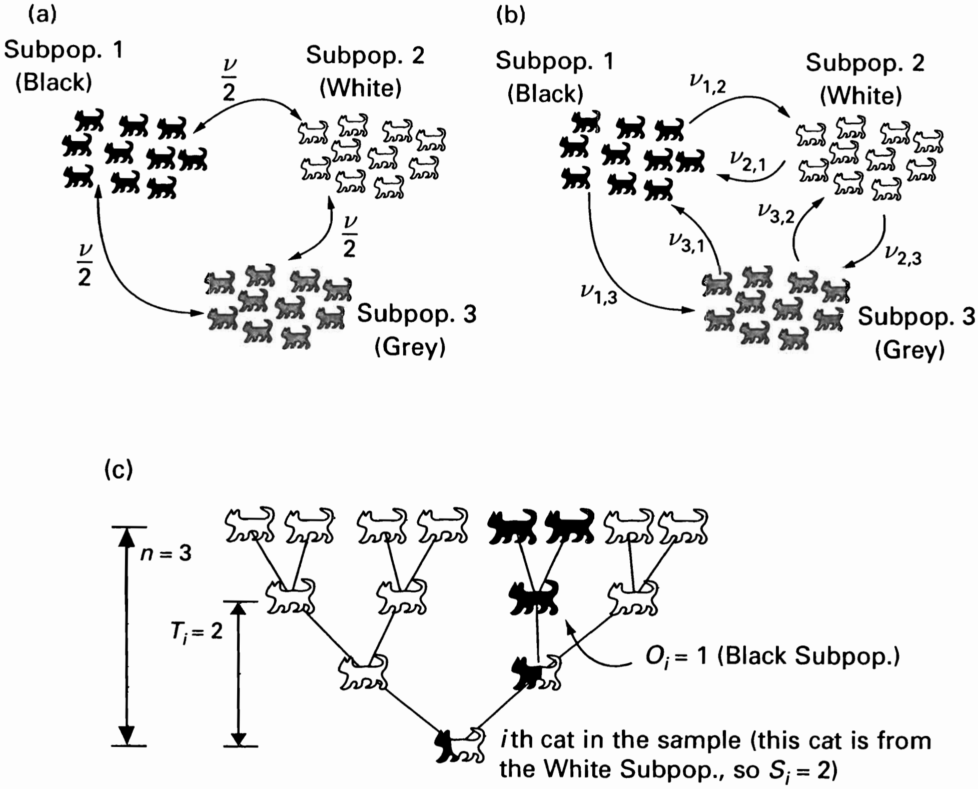


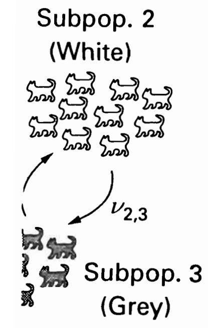
Figure 2.4. (a) A schematic of the *structure* model with prior population information assuming three subpopulations of cats. v is the fraction of individuals in any subpopulation having a single immigrant ancestor in the last n generations from any of the other subpopulations. The other subpopulations are assumed to contribute migrants at the same rate so the probability that an individual has an ancestor from a specific subpopulation is $v/(K-1)$, which in this case is $v/2$ because there are $K=3$ subpopulations. (b) The migration model in *BayesAss+*. $v_{j,k}$ is the fraction of individuals in subpopulation k having immigrant ancestors from subpopulation j in the last n generations. (c) Notation relating to migrants and their descendants. S_i is the location where cat i was sampled. T_i is the number of generations back in time that i had a single migrant ancestor. O_i is the origin of that migrant. n is the total number of generations in the past during which it is assumed an individual might have a migrant ancestor. The cat shown at the bottom of the pedigree was sampled from the White Subpopulation ($S_i=2$) and it has a single migrant ancestor from the Black Subpopulation ($O_i=1$) two generations ago ($T_i=2$). Correspondingly, it is expected to have $\frac{1}{4}$ of its ancestry from the Black Subpopulation, $\frac{3}{4}$ from the White Subpopulation, and no ancestry from the Grey Subpopulation, i.e. $Q_i = \left(\frac{1}{4}, \frac{3}{4}, 0\right)$.

This specialized model is tailored to provide more power than the generic STRUCTURE model for detecting individuals with recent immigrant ancestry. The data required are the multilocus genotypes of the sampled individuals *and* knowledge of the subpopulation each individual was sampled

from. Being n
Specifically, it
rate, and that r
lations. This i
migrants. It is
there were no g
migrant is $1-v$
discover that th
a non-migrant
about the pow
you would hav

Looking at t
genetic data, Y_i
after all, there a
on $Y_{i,\ell,1}$ or $Y_{i,}$
formulation of
another – for e
the frequency
between variab
why it is possib
are orange, the
in the barrel. I
unobserved va
knowing the al
gene copy came
cies. Informati
which, in turn,
that describe w
during the infe
flows througho
served variable
posterior distri
distribution, i.e
individual is a r

There are t
admixture and
account for the
predictable *pat*
subpopulations
second limitat



Subpop.)

; cat is from
= 2)

ılation
on of individuals
ıst *n* generations
are assumed to
dividual has an
se is $\nu/2$ because
Ass+. $\nu_{j,k}$ is the
estors from
migrants and
s the number of
the origin of that
which it is
ıown at the
on ($S_i = 2$) and it
= 1) two
of its ancestry
and no ancestry

ower than the
ent immigrant
f the sampled
ıl was sampled

from. Being more specialized, this model also makes more assumptions. Specifically, it is assumed that migration occurs infrequently at a known rate, and that migration occurs at the same rate from and into all subpopulations. This is a model for detecting migrants; not for detecting non-migrants. It is important to note that given the way the model is set up, if there were no genetic data, the posterior probability that a individual is *not* a migrant is $1 - \nu$. Therefore, if you choose ν to be 0.01 and run STRUCTURE to discover that the posterior probability that each individual in your sample is a non-migrant is 0.99, you *must not* infer that this is telling you anything about the power of your genetic data to distinguish the subpopulations – you would have obtained the same result even if you had no genetic data.

Looking at the DAG of Fig. 2.3b, one might not immediately see how the genetic data, $Y_{i,\ell,1}$ and $Y_{i,\ell,2}$, will influence the posterior distribution of Q_i – after all, there are no arrows from $Y_{i,\ell,1}$ or $Y_{i,\ell,2}$ to Q_i, so how can Q_i depend on $Y_{i,\ell,1}$ or $Y_{i,\ell,2}$? The answer is that, even though it is natural in the formulation of a probability model to speak of one variable depending on another – for example, the colour of a ball drawn from a barrel depends on the frequency of different-coloured balls in the barrel – the influence between variables runs in both directions along the arrow. This is, in fact, why it is possible to do inference: if most of the balls you draw from a barrel are orange, then you may infer that there is a high frequency of orange balls in the barrel. In other words, the observed data influence your belief about unobserved variables. In the case of the STRUCTURE model of Fig. 2.3b, knowing the allelic type $Y_{i,\ell,1}$ gives you some information about where that gene copy came from ($W_{i,\ell,1}$) if you have some idea about the allele frequencies. Information about $W_{i,\ell,1}$, in turn, influences your belief about Q_i which, in turn, influences your belief about T_i and O_i which are variables that describe whether an individual is a migrant or not. In other words, during the inference process information obtained from observed variables flows throughout the graph to influence one's belief about *all* the unobserved variables and parameters. A corollary is that with no data, the posterior distribution of variables or parameters will merely be their prior distribution, i.e. with no genetic data, the posterior probability that an individual is a migrant is merely its prior probability, ν.

There are two important limitations of the STRUCTURE model with admixture and prior population information. The first is that it does not account for the fact that descendants of migrants will inherit genes in predictable *patterns* (not just in predictable *proportions*) from the different subpopulations (more details appear in the following section). The second limitation is the requirement that the migration rate ν must be known, or assumed. It would be preferable to allow the estimation of v from the data.

The NewHybrids model

NewHybrids is designed to identify individuals that are recent hybrids between two species or populations. It can distinguish between genealogical classes like F_1, F_2, and backcrosses in a way that structure cannot because the NewHybrids model takes account of the predictable *patterns* of gene inheritance in hybrids, while the structure model does not. The simplest example occurs in comparing F_1 hybrids (the offspring of parents from different populations or species) with F_2 hybrids (the offspring of two parents who are themselves F_1 hybrids). F_1 hybrids will have *exactly* one gene copy from one population and one gene copy from the other population *at every locus*. An F_2 individual will also have, on average, half of its gene copies from one population and half from the other; however, only in half of its loci, on average, will there be exactly one gene copy from each population. The model in structure is not able to detect differences between F_1s and F_2s because it models admixture strictly in terms of Q_i, which is the proportion of gene copies an individual will have, *on average*, from different subpopulations.

The DAG for the NewHybrids model (Fig. 2.5a), shows that it is a mixture model. In this case, however, the different components of the mixture are different genealogical classes, rather than simply different

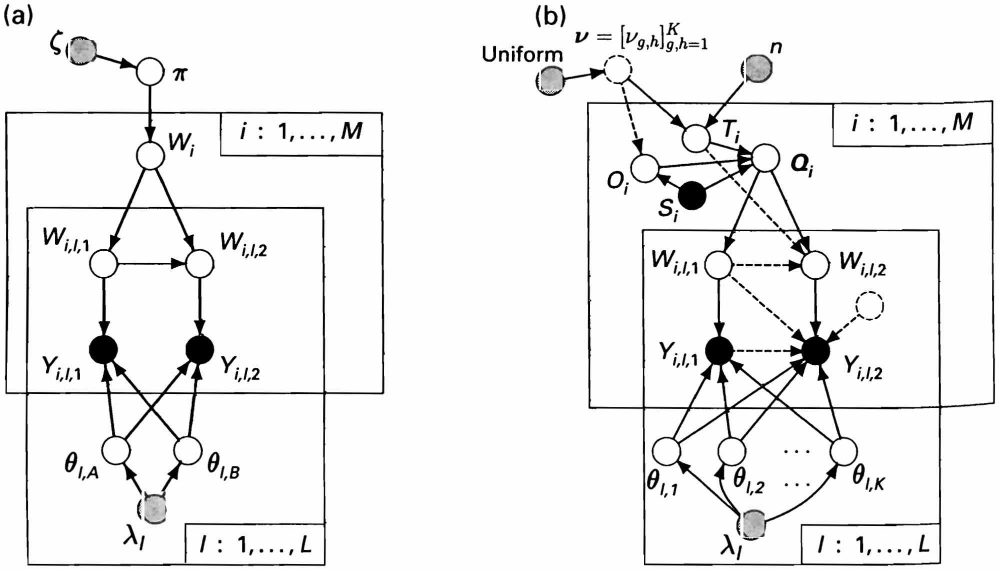
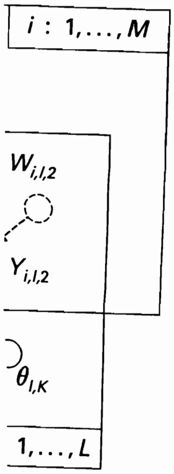


Figure 2.5. (a) The NewHybrids model. (b) The model used in BayesAss+. The additions to the model that make it different from structure with admixture and prior population information are depicted with dashed lines.

e estimation of

recent hybrids
veen genealog-
JCTURE cannot
able *patterns* of
does not. The
ring of parents
ffspring of two
ve *exactly* one
other popula-
half of its gene
only in half of
each popula-
s between F_1s
which is the
rom different

's that it is a
nents of the
ply different



sAss+. The
lmixture and

populations. π denotes the proportions of individuals of different genealogical classes present where the sample is drawn, and Z_i is an unobserved variable that denotes the genealogical class of individual i. There are only two different species or populations (A and B) that an individual's genes may come from.

The model can be described by imagining simulating data from it, conditional on π and the allele frequencies. For the ith individual: (1) A coloured ball is drawn from a barrel with balls in the proportions of π. The colour of the ball gives the genealogical class (Z_i) of the individual. (2) Given the genealogical class, the population of origin of the individual's first gene copy ($W_{i,\ell,1}$) is drawn from a barrel much like the *Q*-barrel described before. (3) The origin of the second gene copy ($W_{i,\ell,2}$) is drawn from a distribution that depends not only on the genealogical class, but also on the origin of the first gene copy. For example, if the genealogical class is F_1, and the first gene copy came from population *A*, then the second gene copy must come from population *B*. (4) The allelic type of each gene copy is drawn from the allele frequencies in their respective populations of origin.

Visible in the DAG are the inference problems that can be tackled with NEWHYBRIDS. The value of π can be estimated, and the genealogical class of each individual in the sample can be inferred. Also, the allele frequencies in populations *A* and *B* may be estimated.

The number of genealogical classes used in NEWHYBRIDS can be determined by the user. The default is six: two pure species categories, F_1, F_2, *A*-backcross, and *B*-backcross categories. A considerable amount of genetic data is required to distinguish genealogical classes, even with as few as six classes (Vähä and Primmer 2006). It is even more difficult to resolve other genealogical classes like second- or third-generation backcrosses. Hence, NEWHYBRIDS is particularly appropriate for the study of hybrid zones in which hybridization has started to occur only recently, or in which the degree of introgression and backcrossing is limited due to selection against hybrids. It is worth pointing out that if only the two pure categories (Pure *A* and Pure *B*) are used, the NEWHYBRIDS model reduces to the standard mixture model of Fig. 2.2b with $K = 2$.

The data required are the multilocus genotypes of the sampled individuals. Learning samples are not necessary, but they may be included. It is not necessary to have prior information about subpopulations or species.

The BAYESASS+ model

The model in BAYESASS+ is a natural extension of the STRUCTURE model with admixture and prior population information. Figure 2.4b gives a schematic of the migration model. Importantly, migration rates are not constrained to be the same between all pairs of populations. Further, with BayesAss+ it is not necessary to assume a value of the migration rate. Rather, BayesAss+ endeavours to estimate the (possibly nonsymmetrical) rates of migration between all subpopulations. The other two advances over structure are the correct modelling of patterns of gene inheritance and the inclusion of an inbreeding parameter $\boldsymbol{F}=(F_1, \ldots, F_K)$ that tries to account for possible departures from Hardy–Weinberg equilibrium within each subpopulation.

Comparing the DAG for the BayesAss+ model (Fig. 2.5b) to that of structure with admixture and prior population information (Fig. 2.3b) shows that the two are similar, differing only in a few variables, and a few extra arrows. Proceeding from top to bottom in the DAG, we first see that ν has been replaced with a matrix $\boldsymbol{\nu}$ of individual migration rates between the populations (Fig. 2.4b). There is a new arrow connecting $\boldsymbol{\nu}$ to O_i because, since immigration rates are no longer symmetrical and equal, the origin of immigrants depends both on their destination S_i *and* on the migration matrix $\boldsymbol{\nu}$. The two new arrows, from T_i and $W_{i,\ell,1}$ to $W_{i,\ell,2}$ are there as a consequence of the fact that BayesAss+ models the inheritance of genes from migrants in the same way that NewHybrids does genes in F_1s and backcrosses. Finally, the arrows from $W_{i,\ell,1}$, $Y_{i,\ell,1}$, and $\boldsymbol{F}$ to $Y_{i,\ell,2}$ describe the interdependence of those variables induced by the possibility of inbreeding (departures from Hardy–Weinberg equilibrium). In words, the type of the second gene copy at a locus is no longer independent of the type of the first gene copy even if they both originate from the same population.

The primary goal of inference using this model is the estimation of the migration matrix. The data requirements for BayesAss+ are the same as they are for structure with admixture and prior population information – it requires multilocus genotypes sampled from K distinct subpopulations. The model provides a more faithful representation of the data than does structure and it is appropriate for estimating recent migration between populations that are well differentiated genetically. However, it is apparent that if the populations are not greatly differentiated, then it may be difficult to estimate the migration rates between them. This could lead to misleading results if attempting to estimate migration rates between demes of a recently fragmented population. The various demes will be similar genetically due to recent common ancestry, and this might lead to inflated estimates of migration rates, even if no migration is presently occurring due to the recent fragmentation. Similarly, users should be suspicious of nonmigration rates close to $\frac{2}{3}$ as this is the minimum allowed by the

program and ma
ated to the level r

PRACTICAL

Quite reasonably,
practical issues i
from 'How large s
to issues like 'Wh
priors for the all
programs?' While
and Primmer 2c
provided some ge
by many differen
between the popu
admixture, etc. It
correspond well t
may have differer
any particular sim
results to the resu
own data set und

An excellent
structure in cod (
(Nielsen *et al.* 20c
genotypes they ob
mixing of pure m
between two pop
cally addresses, s
gram called Hybr
using allele freq
results from the
results from the r
ical mixing scena
ulations may be c

Simulating m
is not a difficult
lation programs.
and Dunham 20
frequencies, and
noaa.gov) simula

t constrained to
BayesAss+ it is
ıer, BayesAss+
ᵌs of migration
structure are
he inclusion of
ınt for possible
ubpopulation.
2.5b) to that of
:ion (Fig. 2.3b)
ıriables, and a
.G, we first see
ıigration rates
:onnecting v to
ical and equal,
ı S_i *and* on the
$_{,1}$ to $W_{i,\ell,2}$ are
he inheritance
ɔs does genes
and F to $Y_{i,\ell,2}$
the possibility
ım). In words,
ıdependent of
rom the same

imation of the
e the same as
information –
bpopulations.
ata than does
ıtion between
it is apparent
ay be difficult
to misleading
demes of a
;imilar genet-
.d to inflated
ıtly occurring
suspicious of
owed by the

program and may indicate that populations are not genetically differentiated to the level required to get reliable results.

PRACTICAL ISSUES

Quite reasonably, an entirely separate chapter could be written dealing with practical issues involved in running the programs described here; issues from 'How large should my samples be?' and 'How many loci should I use?' to issues like 'Why do I get different results in NewHybrids using different priors for the allele frequencies?' and 'Can I trust the results from these programs?' While some recent simulation studies (Evanno *et al.* 2005; Vähä and Primmer 2006) have addressed these sorts of questions, and have provided some general answers, the behaviour of these methods is affected by many different features of the data, including the genetic differentiation between the populations, the number of alleles at each locus, the degree of admixture, etc. It is unlikely that any simulations that have been done will correspond well to all such features in your own data set. Furthermore, you may have different questions in mind than the ones that were addressed in any particular simulation study. In such cases, it is valuable to compare your results to the results obtained by analysing data simulated to look like your own data set under different hypotheses of interest.

An excellent example of this type of effort appears in an analysis of structure in cod (*Gadus morhua*) populations in the seas around Denmark (Nielsen *et al.* 2003). The authors were interested in whether the patterns of genotypes they observed in a contact zone were concordant with mechanical mixing of pure members of two populations, or with a zone of admixture between two populations. This is not a question that structure automatically addresses, so the two different scenarios were simulated with a program called HybridLab (see Nielsen *et al.* 2003 for details of the program) using allele frequencies from the two different pure populations. The results from the simulated admixture scenario were more similar to the results from the real data than were the results from the simulated mechanical mixing scenario, providing evidence that admixture between the populations may be occurring.

Simulating multilocus genotype data from specific allele frequencies is not a difficult task, but is not a standard feature in many genetic simulation programs. In addition to HybridLab, the program spip (Anderson and Dunham 2005) simulates multilocus genotypes from specified allele frequencies, and the program simdata_nh (available from eric.anderson@noaa.gov) simulates genotypes of individuals of different genealogical classes under the NewHybrids model. These programs can be used to test the inferences from the three programs structure, NewHybrids and BayesAss+.

The methods reviewed in this chapter are complex enough that it is difficult (even for the authors of the programs) to make specific predictions about how these methods will behave when confronted with specific data sets. For this reason, the most important practical advice I can give is that it is incumbent upon the careful user of these programs to simulate data that are similar to their own and then analyse them with the program they are using. In order to gain insight about the results of these programs, there really is no substitute for comparing your results to the results achieved using simulated data that look like your own, *but in which you know the truth* (i.e. you know which individuals are F_1s, and F_2s, or which ones are migrants).

ACKNOWLEDGEMENTS

I thank Devon Pearse, Ingrid Spies and Greg Wilson for their insightful comments on a draft of this chapter, and Robin Waples and Kristen Ruegg for helpful discussions on the topic.

REFERENCES

Anderson, E. C. and Dunham, K. K. (2005). spip 1.0: a program for simulating pedigrees and genetic data in age-structured populations. *Molecular Ecology Notes*, 5, 459–461.

Anderson, E. C. and Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, **160**, 1217–1229.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J. *et al.* (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, **368**, 455–457.

Corander, J., Waldmann, P., Marttinen, P., and Sillanpaa, M. J. (2004). BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.

Dawson, K. J. and Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research*, **78**, 59–77.

Evanno, G., Regnaut S. and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Falush, D., Stephens, M. and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Manel, S., Gaggiotti, O. E. and Waples, R. S. (2005). Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology and Evolution*, **20**, 136–142.

Milner, G. B., Tee
Identification
Northwest an

Nielsen, E. E., Ha
Evidence of a
the Danish Be
12, 1497–1508

Paetkau, D., Calv
of population

Pearse, D. E. and
data for conse

Pella, J. and Mas
from genetic

Piry, S., Alapetite,
assignment a
536–539.

Pritchard, J. K., S
structure usir

Rannala, B. and M
genotypes. *Pr*

Roques, S., Sevig
gressive hybr
Atlantic: a ra

Smouse, P. E., W.
for use with i
and Aquatic S

Vähä, J.-P. and Pr
detecting hyb
different nun

Wilson, G. A. and
using multilo

used to test the
wHYBRIDS and

ough that *it is*
cific predictions
ith specific data
can give is that
o simulate data
e program they
programs, there
esults achieved
u know the truth
which ones are

heir insightful
Kristen Ruegg

simulating
cular Ecology

d for identifying
17–1229.
resolution
Nature, **368**,

04). BAPS 2:
ucture.

entification
netical Research,

er of clusters of
Molecular

ılation structure
requencies.

ethods: match-
Ecology and

Milner, G. B., Teel, D. J., Utter, F. M. and Burley, C. L. (1981). *Columbia River Stock Identification Study: Validation of Method*. Technical report, NOAA. Seattle: Northwest and Alaska Fisheries Center.

Nielsen, E. E., Hansen, M. M., Ruzzante D. E., Meldrup, D., and Gronkjaer, P. (2003). Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. *Molecular Ecology*, **12**, 1497–1508.

Paetkau, D., Calvert, W., Stirling, I. and Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.

Pearse, D. E. and Crandall, K. A. (2004). Beyond F-ST: analysis of population genetic data for conservation. *Conservation Genetics*, **5**, 585–602.

Pella, J. and Masuda, M. (2001). Bayesian methods for analysis of stock mixtures from genetic characters. *Fisheries Bulletin (Seattle)*, **99**, 151–167.

Piry, S., Alapetite, A., Cornuet J.-M. *et al.* (2004). GENECLASS2: A software for genetic assignment and first-generation migrant detection. *Journal of Heredity* **95**, 536–539.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rannala, B. and Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences USA*, **94**, 9197–9201.

Roques, S., Sevigny, J. M. and Bernatchez, L. (2001). Evidence for broadscale introgressive hybridization between two redfish (genus *Sebastes*) in the North-west Atlantic: a rare marine example. *Molecular Ecology*, **10**, 149–165.

Smouse, P. E., Waples, R. S. and Tworek, J. A. (1990). A genetic mixture analysis for use with incomplete source population data. *Canadian Journal of Fisheries and Aquatic Science*, **47**, 620–634.

Vähä, J.-P. and Primmer, C. R. (2006). Efficiency of model-based methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, **15**, 63–72.

Wilson, G. A. and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, **163**, 1177–1191.

Contributors

ERIC C. ANDERSON
Fisheries Ecology Division,
Southwest Fisheries Science Center,
110 Shaffer Road,
Santa Cruz,
CA 95060, USA

JON BEADELL
Center for Conservation and
Evolutionary Genetics,
National Zoological Park and National
Museum of Natural History,
Smithsonian Institution,
P.O. Box 37012 MRC 5503,
Washington,
DC 20013, USA
and
Department of Ecology and Evolutionary
Biology,
Yale University,
ESC 158b,
21 Sachem Street,
New Haven,
CT 06520, USA

MARK BEAUMONT
School of Biological Sciences,
Philip Lyle Research Building,
P.O. Box 68,
University of Reading,
Whiteknights,
Reading RG6 6BX,
United Kingdom

PETER BEERLI
Department of Scientific Computing
and Biological Sciences Department,
150-T Dirac Science Library,
Florida State University,
Tallahassee,
FL 32306, USA

LUCIANO B. BEHEREGARAY
Department of Biological Sciences,
Macquarie University,
Sydney,
NSW 2109, Australia

LOUIS BERNATCHEZ
Department of Biology,
Pavillon Charles-Eugène-Marchand
1030, Avenue de la Médecine,
Room 1145,
Université Laval,
Québec,
Québec G1V 0A6, Canada

GIORGIO BERTORELLE
Department of Biology and Evolution,
University of Ferrara,
44100 Ferrara, Italy

LUIGI BOITANI
Department of Animal and Human
Biology,
University of Rome 'La Sapienza',
viale Università 32,
00185 Roma, Italy

AURÉLIE BONIN
Laboratoire d'Ecologie Alpine,
CNRS-UMR 5553,
Université Joseph Fourier,
BP 53,
38041 Grenoble cedex 09,
France
and
Department of Biology,
Indiana University,
1001 E. Third Street,
Bloomington,
IN 47405, USA

g researchers
basic theory
e series
style targeted
rs.

ford and

and

wistle and

drew
(paperback)
vid
K (paperback)
Mace, Kent
5 (paperback)
521 78340 2

Amanda
521 01110 8

Rabinowitz

d Thomas

uncanand

36 5

85706 6

well, Lesley
back)
ndré
0726

Population Genetics for Animal Conservation

Edited by

GIORGIO BERTORELLE
Department of Biology and Evolution, University of Ferrara, Italy

MICHAEL W. BRUFORD
School of Biosciences, Cardiff University, United Kingdom

HEIDI C. HAUFFE
Edmund Mach Foundation, Trento, Italy

ANNAPAOLA RIZZOLI
Edmund Mach Foundation, Trento, Italy

CRISTIANO VERNESI
Edmund Mach Foundation, Trento, Italy

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521866309

First published 2009

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Population genetics for animal conservation / edited by Giorgio Bertorelle ... [et al.].
p. cm.
ISBN 978-0-521-86630-9
1. Conservation biology. 2. Animal genetics. I. Bertorelle, Giorgio. II. Title.
QH75.P64 2009
639.9–dc22
2009005815

ISBN 978-0-521-86630-9 hardback
ISBN 978-0-521-68537-5 paperback

Conservatio
practitioner
exist, and tc
and probler