**Noname manuscript No.**
(will be inserted by the editor)

# Some applications of genetics in statistical ecology

**R. M. Fewster**

Received: August 2015 / Accepted: date

**Abstract** Genetic data are in widespread use in ecological research, and an understanding of this type of data and its uses and interpretations will soon be an imperative for ecological statisticans. Here we provide an introduction to the subject, intended for statisticians who have no previous knowledge of genetics. Although there are numerous types of genetic data, we restrict attention to multilocus genotype data from microsatellite loci. We look at two application areas in wide use: investigating population structure using genetic assignment and related techniques; and using genotype data in capture-recapture studies for estimating population size and demographic parameters. In each case we outline the conceptual framework and draw attention to both the strengths and weaknesses of existing approaches to analysis and interpretation.

**Keywords** Microsatellite · Population genetics · Genetic assignment · Invasion ecology · Mark-recapture · Misidentification · Genotyping errors

## 1 Introduction

Genetic data contain a wealth of information about numerous processes in ecology and evolution, many of which can only be studied through the genetic lens. However, it is often difficult for statisticians to acquire the background knowledge necessary to contribute to the field. Statistical analysis of genetic data requires at least some understanding of how the data arise, encompassing both the biological mechanisms underlying genetic structures and inheritance, and the laboratory processes by which the data are extracted and reported.

R. M. Fewster
Department of Statistics, University of Auckland,
Private Bag 92019, Auckland, New Zealand
Tel.: +64-9-9233946
E-mail: r.fewster@auckland.ac.nz

The associated biological literature is often impenetrable to those from outside the field, and even familiar concepts such as statistical independence are often expressed in unfamiliar terminology.

The aim of this article is to provide insight into the use of genetic data in ecological research, for statisticians with no previous knowledge of genetics. There is a vast array of genetic structures in nature, as well as numerous ways of extracting the data, and many different approaches to data analysis to serve a plethora of objectives, so it is impossible to capture the breadth of the field in a single paper. Instead we focus on just one type of genetic data — microsatellite genotype data — and two application areas likely to be of particular interest to statistical ecologists: (i) investigating population structure and provenance of individuals using genetic assignment and related techniques; and (ii) using DNA samples for individual identification in capture-recapture studies for estimating population size and demographic parameters. Even within these areas we restrict discussion to a subset of the available methodologies. The intention is to convey sufficient understanding in these areas to generate insight that is transferable to other genetic contexts and data types. For example, much of the statistical foundation that has been developed for microsatellite data is directly transferable to new data emerging from next-generation sequencing (NGS), but NGS data does demand a review of the underlying statistical assumptions, and also presents many opportunities for developing new analysis methods. We focus on microsatellite data here — despite the rapid escalation of cheap NGS technologies and consequent decline in the use of microsatellites — because most of the literature in ecological population genetics to date is based on this data type and the particular errors that arise from the way that it is generated.

## 1.1 Genotype data

A *genetic locus* (plural *loci*) is a position on a chromosome. It describes a genetic location or address. The different genetic choices available at a locus are called *alleles*. For example, we can imagine that humans have a genetic locus for eye colour, at which the available alleles are the genetic sequences for blue eyes, brown eyes, and so on. We shall only treat the case where each individual inherits two alleles at each locus: one from its mother, and one from its father. The set of two specific alleles that an individual possesses is called its *genotype* at this locus. For example, at a locus for eye colour a human might possess one allele for blue eyes and one for brown eyes, in which case we could describe his or her genotype as ‘blue, brown’. It is not generally known which of the two alleles was inherited from the mother, and which from the father.

The selection of alleles available at a locus is a result of the accumulation of genetic mutations over hundreds of thousands of years. For genetic loci that control biological functions, known as *coding* loci, some mutations change the gene’s function and might be eliminated or promoted by natural or sexual selection. However, in much of the genome, mutations are of neutral impact and are said to be *selectively neutral.* This is particularly true in the case of *non-coding DNA:* genetic code that serves no obvious purpose but seemingly surrounds the functional DNA like packaging in a box. Non-coding loci have traditionally been of primary interest in population genetics studies, because the accumulated mutations provide a choice of numerous different alleles that can discriminate between individuals and populations. All loci behave alike in terms of genetic inheritance, but non-coding loci tend to have more allele types, a property known as *polymorphism.*

A *microsatellite locus* is a type of locus that has proved particularly useful in population genetics studies. Microsatellite loci consist of short fragments of DNA that are repeated multiple times. For instance, the sequence $ACACACACAC$ consists of the short fragment $AC$ repeated five times. The term *satellite* is used because repetitive DNA has a higher density than typical DNA and tends to separate into a satellite band in a centrifuge. Microsatellites are also known as short tandem repeats (STRs), simple sequence repeats (SSRs), or variable number tandem repeats (VNTRs). They are typically selected in the belief that they are non-coding loci, although there remains the possibility that some may have biological functions that have not been recognised.

Microsatellites have a relatively high mutation rate, because it is easy for the DNA to 'slip' during replication — effectively losing count of the number of repeats. Consequently, microsatellite loci often exhibit several different alleles that are distinguishable by their different lengths, such as the two alleles $ACACAC$ and $ACACACACACAC$. This type of genetic structure offers two key advantages. Firstly, the relatively large number of available alleles enables good discrimination between individuals, so a suite of about 10 such loci is often sufficient for each individual in a population to have a unique genetic profile. Secondly, the ability to distinguish different alleles by their lengths, instead of having to inspect their precise genetic sequences, means that microsatellite genotypes were for many years relatively inexpensive to obtain. This situation is now in flux with the emergence of cheap next-generation technologies. The primary difference between microsatellite and next-generation protocols is that microsatellite studies target a small number of highly polymorphic loci, whereas next-generation technologies target a massive number of loci but there are typically only two alleles available at each locus.

It is worth having a sketch understanding of how microsatellite genotypes are obtained in the laboratory, because the process involves a small but non-negligible error rate that needs to be taken into account in statistical analysis. The description that follows is not biologically precise but is sufficient for understanding the process of error-generation. Microsatellite loci may initially be identified for a species by genetic sequencing. The genetic sequences on either side of the microsatellite are noted: these regions are called the *binding sites.* It is hoped (but not guaranteed) that the binding sequences are the same for all individuals of the species. For example, the fragment $GCTAAT$-$ACACAC$-$TTATA$ has a left binding sequence of $GCTAAT$ and a right binding sequence of $TTATA$. In reality, the binding sequences are sufficiently long to identify the correct region of DNA uniquely.

The microsatellite is genotyped by bombarding a DNA sample with *primers* consisting of sequences that will bond strongly with the left and right binding sequences. The primers fix on the two binding sites and a reaction follows in which the DNA between them is copied, therefore doubling the number of target microsatellite fragments in the mixture. The process is repeated in several cycles, each of which doubles the occurrence of the microsatellite fragment until it dominates the DNA mixture. This process of *amplifying* the microsatellite fragment is known as *polymerase chain reaction* (PCR). Once it is complete, an electric current is applied to propel the fragments across a gel: a process called *electrophoresis*. Shorter fragments encounter less resistance in the gel and move faster than longer fragments, enabling microsatellite lengths to be deduced. The output of the electrophoresis is plotted on an *electropherogram* or *chromatogram*, which plots allele length on the horizontal axis and intensity on the vertical axis. The allele lengths present in the mixture appear as peaks on the plotted output.

The end product of genotyping is a numeric label for each allele, such as 128 or 130, corresponding to the length of the fragment including the binding sequences. The labels are automatically generated from the electropherogram using computer software, but they should be checked by humans because labelling decisions are not always clear-cut. The absolute number 128 does not have much relevance, but the difference between allele lengths can be relevant. For example, if the microsatellite constitutes repeats $ACAC\dots AC$, the two alleles 128 and 130 are likely to differ by just one repeat of the $AC$ motif.

If the procedure works perfectly, the output trace of an individual's genotype contains either one or two peaks, corresponding to the allele lengths of the individual's alleles. If there is just one peak, say at allele 128, the individual is assumed to be *homozygous* at this locus and its genotype is deduced to be 128,128. If there are two peaks, for instance at 128 and 130, the individual is *heterozygous* with genotype 128,130.

In a typical study, each individual is genotyped at several microsatellite loci: typically from 10 to 20. With $\ell$ loci, the resulting suite of $2\ell$ alleles is the individual's *multilocus genotype*. As long as each locus has a reasonable number of allele types available, these numbers of loci are usually enough to give very high discrimination between individuals. A measure of discriminative power is called the *probability of identity*, PID (Paetkau and Strobeck 1994), and gives the probability that two individuals have the same $\ell$-locus genotype by chance. Ideally, this is extremely small, typically less than $10^{-8}$. For closely related individuals, the equivalent quantity is termed PIDsib (Evett and Weir 1998). Although PIDsib is commonly several orders of magnitude larger than PID, it is still typically very small: perhaps $10^{-3}$ or less. Whether this is sufficient depends upon the population under study and the objectives of the analysis. If PID or PIDsib are not sufficiently small, this can be addressed by adding more loci to the study. However, adding more loci is likely to come at the cost of genotyping fewer individuals within the available budget. It is

**Table 1** Example of multilocus genotype data from New Zealand ship rats (*Rattus rattus*). Each row corresponds to a single rat, whose ID and sampling location are specified. Genotype data from four loci are shown. The locus names are D10, D11, D15, and D16, and each rat possesses two alleles at each locus, denoted by numeric labels. The label 0 denotes missing data for that rat at that locus.

| Rat ID | Location | D10 | | D11 | | D15 | | D16 | |
|---|---|---|---|---|---|---|---|---|---|
| B42 | Broken Islands | 96 | 130 | 276 | 276 | 250 | 262 | 155 | 165 |
| B43 | Broken Islands | 96 | 96 | 276 | 280 | 234 | 262 | 165 | 165 |
| B44 | Broken Islands | 96 | 96 | 276 | 280 | 236 | 262 | 155 | 165 |
| A45 | Aotea | 126 | 128 | 276 | 278 | 234 | 236 | 155 | 165 |
| A46 | Aotea | 96 | 128 | 0 | 0 | 234 | 236 | 155 | 165 |
| A47 | Aotea | 120 | 122 | 276 | 284 | 238 | 238 | 167 | 167 |

highly advisable to run a pilot study to establish how best to balance the number of loci versus the number of individuals that can be genotyped.

## 1.2 Data format

Table 1 shows an example of multilocus microsatellite genotype data from New Zealand ship rats (*Rattus rattus*). The data format shown is typical and is similar to that used by the popular software Genepop (Rousset 2008). Rats were sampled from the Great Barrier Island archipelago, including the large main island (Aotea) and a small island cluster called the Broken Islands about 300m offshore from the main island. Each rat was genotyped at 10 microsatellite loci, of which four are shown in Table 1 (Fewster et al. 2011; Jacob et al. 1995). From the table, it is already evident that it will be challenging to visualize the data in a meaningful way. However, even from this small snippet of data, there is a hint that the Broken Island rats have lower allelic diversity than the Aotea rats. This will be confirmed by the visualisations that we introduce later.

## 1.3 Missing data

Missing data are unavoidable in genetic studies, and statistical methodologies must be capable of handling them. Missing data are denoted by the allele label ‘0’ in Table 1, and indicate that the PCR amplification failed for that individual at that locus. Such failures are common when the DNA sample is of low quality — for example, derived from hair, feathers, or faeces — but also occur for high-quality tissue samples that have degraded due to inadequate or delayed preservation. Missing data can also result from random chance, or from a systematic cause such as null alleles or long allele dropout, which are described below. Sometimes missing records can be restored after repeated efforts to extract and profile the DNA, but typically some records will remain missing.

### 1.4 Allelic dropout

*Allelic dropout* describes the situation where a heterozygous genotype such as 120,130 is misreported as 130,130, because the 120 allele failed to amplify during PCR. The 120 allele is said to have 'dropped out' of the profile. Dropout, where just one of the two alleles fails to amplify, differs from missing data, where both of the alleles fail to amplify, because dropout is not observable from the output trace. Whereas missing data delivers no genotype, dropout delivers the wrong genotype.

Dropout errors have the effect of exaggerating the apparent *homozygosity* in the sample, such that more individuals appear to be homozygous than is really the case. High rates of homozygosity are prevalent in very small populations, or in populations with inbreeding among close kin, and are associated with poor population health because the homozygosity applies not only to non-coding microsatellite loci but also to functional genetic loci. An individual with two copies of the same detrimental allele will suffer from its effects, whereas if the individual has only a single copy, the problematic effects are often overridden by its other allele. A high rate of homozygosity is a real concern for endangered populations, because there is such a vast number of genetically-controlled traits that it is very likely that at least one life-threatening defect, susceptibility, or behaviour will become prevalent in the population. When estimating the homozygosity levels in a population, therefore, researchers must be aware of the possible exaggerating effects of allelic dropout.

### 1.5 Systematic causes of allelic dropout and missing data

Although missing data and allelic dropout are often due to poor sample condition or random chance, there are also some systematic effects that will tend to guarantee that particular alleles or genotypes fail to be reported correctly. Such effects interfere with assumptions that data are missing-at-random, and the consequent impact on the proposed analysis should be considered. For some analyses, systematic missingness of certain alleles might be of little concern, whereas for others it might invalidate the results. Here we describe two mechanisms for systematic dropout, known as null alleles and long allele dropout.

*Null alleles* occur when an individual has a mutation in the binding site used by the primers in the PCR process, so the primers fail to bind and the microsatellite allele flanked by the binding sites is not amplified (Chapuis and Estoup 2007; Pompanon et al. 2005). If an individual possesses such a mutation, its genotype will never be correctly read at this locus. An individual with two copies of the null allele will be reported as missing data (profile 0,0), whereas if it has only one copy — for example, if its real profile is 120,130 but the 120 allele is flanked by a mutation and will not amplify — then the genotype will be misreported as 130,130, corresponding to allelic dropout. Null alleles are alleles like any others, inherited according to the same processes, but they are only 'observable' when the individual is homozygous for the null allele — in which case its profile is always reported as missing at that locus.

*Long allele dropout*, or short allele dominance, describes another artefact of PCR whereby longer alleles might be more likely to suffer dropout than shorter ones, because longer alleles take longer to replicate so there is greater risk that the reaction is not completed within each PCR cycle. In some cases an allele might be sufficiently long that the reaction is never completed, so it always drops out. In other cases, a partially-successful amplification might be readable from the output trace when both of an individual's alleles are long, but otherwise a small peak from a longer allele might be obscured by a much more dominant peak from a shorter allele. Thus, for example, genotype 120,160 might tend to be consistently misread as 120,120, whereas genotype 156,160 might tend to be correctly read because the signal from the two peaks is roughly equal, albeit weak.

## 1.6 Stutter and false allele reads

There are numerous other causes of error in microsatellite genotype data (Pompanon et al. 2005; Taberlet and Luikart 1999). Sometimes a microsatellite allele length is simply misreported, giving a false allele record. This could happen due to sample contamination — for example, from not cleaning equipment properly between dealing with samples from different individuals — or due to laboratory handling or labelling errors that ascribe a locus reading to the wrong sample. A more systematic reason for false allele reads comes from the PCR process itself. During PCR, the microsatellite fragment flanked by the binding site is repeatedly replicated. If the replication is occasionally incomplete, or if slippage occurs during replication in much the same way as mutations occur in reality, the replicated fragment might be a different length from the microsatellite it is aiming to replicate. This fragment is then itself amplified in the next PCR cycle, leading to a mixture of correct and incorrect lengths in the final solution. It is common for this to occur to some small degree, leading to a *stutter* on the output genotyping trace consisting of minor peaks at allele lengths slightly different from that of the target microsatellite fragment. Usually the peak at the correct microsatellite length clearly dominates the minor peaks caused by stutter; but occasionally it is difficult to distinguish between the cases where the true genotype is (say) 128,130, and where it is 130,130 with a minor peak at 128 caused by stutter.

The occurrence of problems such as stutter and false allele reads can in some cases be reduced by a careful selection of which microsatellite loci to genotype, as some are more error-prone than others. This again highlights the importance of setting aside funds for pilot studies.

### 1.7 Importance of error-handling

Errors and missing data are inherent in genotyping data, and statistical analyses must be designed to accommodate them. The importance of errors and error-handling depends upon the context of the analysis. If genotyped samples are used to reconstruct capture histories for a capture-recapture analysis, then the validity of the analysis rests on the correct matching of samples to individual animals. Failing to allow for common errors such as dropout in this context will lower the apparent recapture rate and lead to a systematic overestimation of population size (Wright et al. 2009; Vale et al. 2014). Similarly, if genetic data are used to investigate parentage, errors in individual reads could falsify conclusions: a single instance of allelic dropout could appear to exclude the true parent from having produced a particular offspring. By contrast, studies of population structure and connectivity do not rely to such a degree on the accuracy of individual genotypes, and occasional misreads are of lesser importance. In Section 3 we give examples where possible to calibrate the level of genetic error and missingness that might be expected in modern studies.

## 2 Genetic assignment and population structure

The term *population genetics* describes the study of the structure, connectivity, and evolutionary history of ecological populations based on their contemporary genetic profiles. It is a fascinating field of study, because genetic data offer insights that would be difficult or impossible to gain by other means. However, the information encoded in the genetic record can be hard to extract and interpret, which for statisticians creates considerable opportunities for innovative development. The genetic processes underlying today's populations are immensely complex, including processes that unfold over vast time-scales such as natural selection, mutation, and genetic drift, down to the complicated business of sexual reproduction which involves the scrambling of parental genes at every generation. Alongside these is a vast array of species mating systems, migration and dispersal patterns, and other behavioural considerations that conspire to ensure that inference from genetic data is far from straightforward.

In this section we describe the statistical foundation of some widely-used techniques for exploring population structure and estimating the provenance of individual animals. The term *population structure* is used to signal the existence of identifiable subpopulations within a larger population, contrasting with so-called unstructured situations where the whole population is genetically homogeneous. For example, if the genetic profiles of different islands in an archipelago are distinct from one another, it might be relatively easy given an individual from any of the islands to determine which island it has come from. The archipelago population in this case is said to be highly structured. On the other hand, an unstructured population would present no clear genetic differences between the islands, either because of ongoing mixing of island individuals or because of a common genetic heritage.

It is clear from this that concepts of population structure are linked with the question of assigning the provenance of individuals, and indeed both types of investigation may be tackled using the same statistical foundation, which we outline below.

### 2.1 Principles of genetic assignment

Genetic assignment is the process of estimating the source population of an individual by comparing the individual's genotype with the profiles of candidate source populations. A precise mathematical treatment quickly cascades into volumes of notation and subscripts — concerning alleles within loci, loci within individuals, and individuals within populations — and this level of abstraction readily conceals the common-sense principles of the subject. For this reason we defer the mathematical treatment to Section 2.4, and here focus on introducing the ideas in a simple everyday context. The aim is to anchor the principles in common-sense reasoning which will aid interpretation in more complex and abstract scenarios.

Genetic assignment techniques rely upon the observation that populations that are isolated from one another develop differences in *how common* are the different allele types within them. For example, among humans, alleles for blond hair are very common in Sweden and less common in Italy. However, blond alleles are present in both populations: it is the difference in prevalence that underpins the process of genetic assignment.

The statistical basis of genetic assignment is very simple, and can be illustrated by the same example. Suppose we wish to decide upon a native source country for a blond (fair-haired) person, out of three candidate countries: Sweden, Italy, and England. In Sweden, we take reference samples and estimate that blonds constitute 75% of the population. The blond person of interest is therefore given 75% chance of arising in Sweden: $P(\text{blond} \,|\, \text{Swedish}) = 0.75$. Similarly, we estimate that blonds constitute 10% of the population in Italy, and 40% of the population in England, yielding $P(\text{blond} \,|\, \text{Italian}) = 0.10$ and $P(\text{blond} \,|\, \text{English}) = 0.40$. (Figures are rough estimates based on genetic maps of Europe from `www.eupedia.com`.)

The total of the genetic evidence in this case is the trio of numbers $(0.75, 0.10, 0.40)$, giving the probability of finding the blond genotype in the three candidate countries. Specifically, the genetic evidence constitutes the three conditional probabilities $P(\text{blond} \,|\, \text{Swedish})$, $P(\text{blond} \,|\, \text{Italian})$, and $P(\text{blond} \,|\, \text{English})$. The order of the conditioning is important, and should not be confused with $P(\text{nationality} \,|\, \text{blond})$.

The trio of probabilities $(0.75, 0.10, 0.40)$ has the interpretation that the blond person could have been born in any of the three countries, but that the examined genes are very common in Sweden, common in England, and less common in Italy. As such, it tells us little about the provenance of the individual, except that all three candidate countries are plausible sources. However, the multi-dimensional probability vector is often simplified to a more succinct summary, and in the process some potential for misinterpretation is created. The software GeneClass2 (Piry et al. 2004) uses the following calculation to rescale the numbers into *assignment scores*:

$$\begin{aligned}
&\text{Assignment to Sweden:} && \frac{0.75}{0.75+0.10+0.40}\times 100 = 60\% \\
&\text{Assignment to Italy:} && \frac{0.10}{0.75+0.10+0.40}\times 100 = 8\% \\
&\text{Assignment to England:} && \frac{0.40}{0.75+0.10+0.40}\times 100 = 32\% \qquad (1)
\end{aligned}$$

Assignment scores are therefore a simple rescaling of the three numbers (0.75, 0.10, 0.40) so that they add up to 100. The calculation could also be thought of as an application of Bayes' rule using equal prior probabilities for each population. Under this rescaling, the blond person is sometimes said to assign 60% to Sweden, 8% to Italy, and 32% to England. However, as we discuss below, such terminology is rather misleading.

There are several problems with the transformation from the genetic evidence (0.75, 0.10, 0.40) into assignment scores (60%, 8%, 32%). Most importantly, all information about the size of the genotype probabilities is lost. Large probabilities such as ours, that signal that the blond genotype is common in all three populations, are given the same assignment output as the vector of tiny probabilities (0.00075, 0.00010, 0.00040) that signals the opposite conclusion. A better interpretation of the evidence is needed that acknowledges the differing conclusions from these two results: in the first case the genotype is universally common and all three populations are plausible sources, whereas in the second case it is universally rare and raises doubt over whether any of the populations is the true source.

Secondly, the rescaling of the genotype probabilities, and consequent loss of information about their magnitude, is applied not only to the single sample of interest (the blond person), but to all samples, including the reference samples of known origin that were used to establish the estimates of 0.75, 0.10, and 0.40 in the first place. This means that we lose calibration of what constitutes a 'common' or 'rare' genotype in each of the populations under consideration, and what level of variability in 'commonness' is exhibited among genotypes genuinely drawn from these populations. In our example we have asserted that a genotype with probability 0.75 is 'common' and one with probability 0.00075 is 'rare', but in reality we do not have any basis for asserting that 0.00075 denotes a rare genotype without knowing more about the range of genotypes available, and their probabilities.

Finally, the addition of the percentage sign to the assignment scores (60%, 8%, 32%) is unfortunate because it suggests that the assignment scores should be interpreted as probabilities or proportions. While it might be argued that these numbers reflect the probabilities $P(\text{nationality} \mid \text{blond})$ by an application of Bayes' rule with equal prior probability on each candidate source population, such a prior preempts the point of the analysis and distorts the evidence. In particular, the Bayes' rule interpretation forfeits relevant information on the absolute size of the conditional probabilities $P(\text{blond} \mid \text{nationality})$, to emphasise instead relative values among different populations, which is misleading in the absence of a calibration of magnitude and variability within populations and degree of overlap between them. The implied conclusion that there is a 60% probability that each blond person is Swedish is a reflection of the imposed priors and set of candidate source populations rather than an objective summary of the genetic evidence. Another misleading consequence of the transformation is the subconscious assumption that these probabilities will tend to apply as long as the sample sizes are large enough. This is not the case, as there is no law of large numbers that can be invoked: there is no sense in which a large sample of blond people must inevitably converge to 60% Swedes.

We note that Piry et al. (2004) do not imply that the assignment scores should be interpreted as probabilities: they simply describe them as 'scores' and do not make any further comment about how they should be interpreted or used. The assumption by some practitioners that they can be treated as source probabilities is a misinterpretation, but one that is perhaps encouraged by the unfortunate use of the percentage sign.

Instead of transforming meaningful conditional probabilities through the use of an arbitary rescaling or prior, we propose that the genetic evidence is best presented on graphical displays that demonstrate both the magnitude and variability of the raw genetic evidence $P(\text{blond} \mid \text{nationality})$. We outline how such graphics may be constructed in the next section.

Some studies take the output from assignment analyses to an even greater extreme and select a single 'best' population source for each individual, this being the population with the highest assignment score. We call this practice *best-population assignment*. Best-population assignment can be justifiable when the genetic evidence is very conclusive, but when applied without proper consideration of the wider genetic and scientific context, it can generate absurd conclusions. For example, on the basis of genes for hair colour, every blond human in the world should be assigned to Sweden. Indeed, we can take the argument to an even more ludicrous extreme: on the basis of genetic sex, every male human in the world should be assigned to the tiny nation of Liechtenstein — because according to census data in Wikipedia (2015), Liechtenstein is the country in the world with the highest proportion of males at birth. The fact that this is probably an artefact of the relatively small sample size available in Liechtenstein to establish the sex ratio there only serves to reinforce the risks of best-population assignment, as the same possibility of sampling flukes in small reference samples applies to real studies.

Although our example takes best-population assignment to an absurd extreme, it contributes two important points that sometimes get lost amidst greater levels of abstraction. Firstly, there is no basis for assuming that an individual must have been born in the population in which its genotype is most common. The individual might fit well into all of the candidate populations, or into none of them, and there will still be a 'best' population in either case. There is no reason to assume that this must be the individual's birth population. Secondly, and similarly to assignment scores, there is no law of large numbers or other rationale for converting best-population assignments into sample compositions. For example, if 80% of a sample have a best-population assignment to population A, there is no reason to suppose that 80% of the sample were born in population A because the results will tend to be right 'on average', any more than we believe that 50% of the worldwide human population was born in Liechtenstein. Instead, the same mistake (assignment to Liechtenstein) is repeated over again for every human male in the sample. The aim of the graphical displays we describe below is to supply the missing information on genetic context that determines whether practices such as best-population assignment can be supported.

### 2.2 Visualising population structure

Rather than converting genetic assignment evidence to scores, we recommend visualising the data as a way of addressing the points raised in the previous section. We begin by looking at raw data on allele frequencies drawn from different populations. We then describe how genetic assignment data such as the trio (0.75, 0.10, 0.40) can be portrayed on a chart to reveal population structure.

Figure 1 shows the sample data of ship rats from the Broken Islands and Aotea, New Zealand, at the first four of ten genotyped loci as featured in Table 1. The barcharts show the frequency of each allele encountered in the data: in other words, the number of times the allele appeared in the sample data divided by $2n$ where $n$ is the number of rats in the sample from the population of interest. Missing data are shown with allele label 0. The sample sizes are $n = 60$ and $n = 56$, which are fairly large for this type of study.

A number of features of Fig. 1 are evident. Firstly, it is clear that the allele frequencies of the two populations are substantially different, despite the fact that the Broken Islands lie only 300 metres offshore from the much larger island Aotea, and ship rats are capable of swimming this distance. This difference in allele frequencies at each locus between the two populations is the basis on which genetic assignment works, so the evident differences here will contribute to a successful analysis.

Secondly, the Broken Islands profile appears to be largely a subset of the Aotea profile, as we would expect if the Broken Islands were colonized by founders from Aotea. Among the four loci shown, there are 13 alleles that were found in the Aotea sample but not in the Broken Islands sample, and only two alleles for which the reverse is true. This is consistent with the possibility that the Broken Islands were colonized by a small group of founders from Aotea, such that Broken Island alleles are drawn from the Aotea gene-pool, but much of the genetic diversity of the larger island is absent from the small island group.

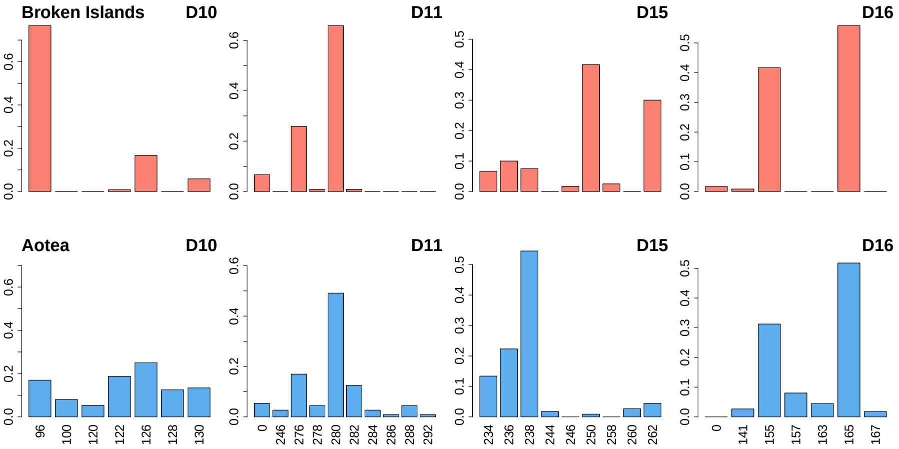

**Fig. 1** Allele frequencies at the four microsatellite loci shown in Table 1 for ship rats sampled on the Broken Islands ($n = 60$ rats: top row) and on Aotea ($n = 56$ rats: bottom row). Alleles are vertically aligned between the two rows, and barcharts for the two populations are plotted on the same vertical scale for each locus.

Thirdly, rare alleles are common — in other words, both populations exhibit a number of alleles that were sampled with very low frequency. This is a common feature of genetic data: there are often many rare alleles in a population sample, meaning that many individuals will possess at least one unusual allele in their multilocus genotype.

Finally, despite the subsetting hypothesis, allele frequencies in the Aotea profile are not good predictors of those in the Broken Islands. Although there are some loci where the Broken Islands profile mirrors the pattern on Aotea, such as D16, this is not true in general: for example, the most common allele sampled on the Broken Islands at locus D15 is very rare in the Aotea sample. This is consistent with so-called *founder effects*. The Broken Islands population was probably founded by a small number of rats sourced from Aotea, and as mentioned it is likely that these rats possessed some alleles that are rare on Aotea, but would henceforth become very common in the newly-founded Broken Islands population by descent from the founders. Subsetting and founder effects have combined to give the Broken Islands a substantially different genetic profile from the nearby Aotea, so we can expect genetic assignment to be a powerful discriminatory tool.

While plotting the raw allele frequency data as in Fig. 1 is instructive, the barcharts are not effective as an overall display of population structure. Fig. 2 shows a more succinct chart for the ship rat data encompassing information from all ten microsatellite loci. We call these charts *GenePlots*. Each individual rat corresponds to one plotted point. Its horizontal coordinate is the estimated log-probability of finding its genotype in the Broken Islands population, and its vertical coordinate is the same for the Aotea population. Thus, each rat has coordinates given by $\log \{P\,(\text{rat's genotype} \mid \text{population}_i)\}$ for populations

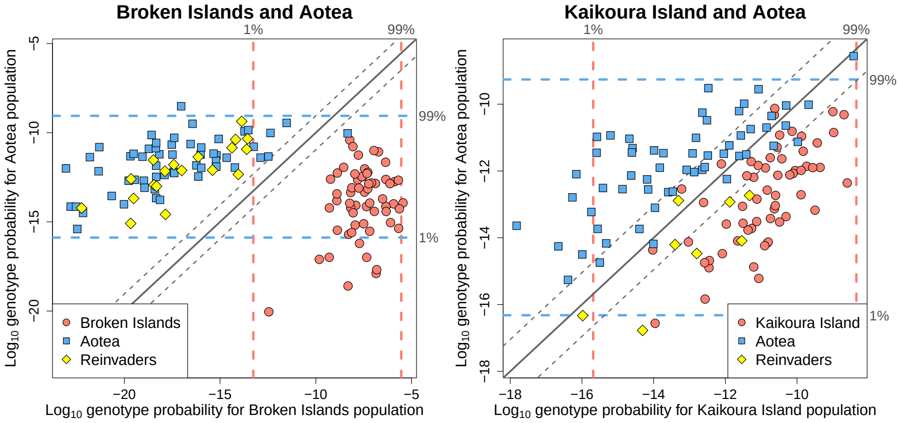

**Fig. 2** GenePlots of ship rats in the Great Barrier Island archipelago. Left: pre-2009 Broken Islands rats (circles); Aotea rats (squares); and rats found on the Broken Islands in 2010 after the eradication attempt (diamonds). Right: pre-2008 Kaikoura Island rats (circles); Aotea rats (squares); and rats found on Kaikoura Island in 2009 after the eradication attempt (diamonds).

$i = 1, 2$, maintaining the correct order of conditioning and therefore retaining information about the size of the genotype probabilities. These estimated genotype probabilities are based on the allele frequencies found in the reference samples, which are rats from each population whose origin is assumed to be known. The reference samples are plotted on the GenePlot along with any samples of unknown provenance, so as to calibrate the magnitude and variability of genotype probabilities that naturally arise in each population. In the first panel of Fig. 2, the reference samples are plotted as circles and squares corresponding respectively to rats sampled before 2009 on the Broken Islands, and rats sampled on Aotea. Because the probability of finding an exact 10-locus genotype in any population will always be extremely small, we plot genotype probabilities on a logarithmic scale. We use base-10 logarithms so that the orders of magnitude covered by the chart can easily be deduced.

In our human blond example, the chart would comprise three axes — one each for Sweden, Italy, and England — and the blond individual's three-dimensional coordinate would be the logarithm of (0.75, 0.10, 0.40), corresponding to $\log_{10}\{P(\text{blond}\,|\,\text{population}_i)\}$ for $i = 1, 2, 3$. As in Fig. 2, these three estimated genotype probabilities are based on reference samples from the three countries. The primary difference between this and the chart shown in Fig. 2 is that the genotype probabilities in Fig. 2 correspond to the full 10-locus genotype, whereas our blond human example has focused on a single genetic trait. The probability of a 10-locus genotype is gained from the product of the 10 single-locus probabilities, and in practice we use a Bayesian posterior predictive genotype probability: details are given in Section 2.4. Otherwise, the principle is the same, so we can verify that the charts resolve the issues raised in Section 2.1:

- The GenePlot genuinely plots (log posterior) genotype probabilities, a quantity of biological relevance, so we need not be concerned about misinterpreting a score as a probability.
- The GenePlot retains information about the size of estimated genotype probabilities. An individual with estimates (0.75, 0.10, 0.40) has a different plotted point from an individual with estimates (0.00075, 0.00010, 0.00040).
- By plotting all reference samples on the same chart, we calibrate the range of genotype probabilities that can be expected for individuals genuinely drawn from the populations of interest. This enables us to calibrate whether a genotype probability such as 0.00075 is genuinely 'rare' for a population or whether it is within the typical range. We can also plot quantiles of the posterior distribution for log genotype probabilities from each population, as shown by the dashed 1% and 99% lines on Fig. 2.

In essence, the GenePlot plots 'belongingness' or fit of an individual to each of $K$ populations, with one axis for each population. The measure of belongingness we use is the posterior log-genotype probability, or LGP for short: the estimated probability of finding the individual's genotype in the population concerned. That is, the LGP is the estimated $\log\{P(\text{genotype} \,|\, \text{population})\}$. If there are more than $K = 2$ populations, a dimension-reduced plot can be used, for example using principal components analysis.

We now go through the features that we identified from the barcharts in Fig. 1 and show how these can be seen on the GenePlot in Fig. 2.

1. *Allele frequencies are substantially different between the Broken Islands and Aotea.* This has created a clear separation on the GenePlot between reference samples from the Broken Islands (circles) and those from Aotea (squares). With the exception of one rat from Aotea that clusters with the Broken Islands reference samples, there is no overlap between the reference samples on the chart. Interestingly, this single rat was sampled on the part of Aotea directly opposite the Broken Islands, and might have been a swimmer sourced from the Broken Islands.
2. *The Broken Islands population is largely a genetic subset of the Aotea population.* This feature is evident on the GenePlot by looking at the quantiles of the posterior LGP distributions. Most of the Broken Islands rats fall between the horizontal dashed lines marking the 1% and 99% quantiles of the posterior LGP distribution for Aotea. This means that most of the Broken Islands rats have an acceptable belongingness to the larger Aotea population. However, very few of the Aotea rats fall between the vertical lines marking the 1% and 99% quantiles of the posterior LGP distribution for the Broken Islands. This means that very few of the Aotea rats have an acceptable belongingness to the Broken Islands. This happens because Aotea is much more allele-rich than the Broken Islands, so a typical Aotea rat possesses alleles that are not found on the Broken Islands. Indeed, it is very unlikely that an Aotea rat by chance possesses only those alleles

found on the Broken Islands, because it is conceived from a much richer allele pool. The impoverished genetic profile of the Broken Islands is what makes these rats distinctive from Aotea rats.

3. *Rare alleles are common.* As mentioned, it is common for an individual to possess one or more alleles that are rare in its own source population. Just one such allele will substantially lower the individual's log genotype probability for its own source population, and two or more such alleles will further exaggerate the effect. The result is that the range of LGPs exhibited in each source population is very large. Broken Islands rats are typically plotted from LGP $-10$ to $-5$, meaning that the 'likeliest' native Broken Island rats are judged 100,000 times more likely than the 'unlikeliest' native Broken Island rats in their own source population. For Aotea, the range is even larger at nearly 10 million. This enormous range emphasizes the danger of simplifying assignment output using assignment scores or best-population assignments, which ignore the inherent variability of belongingness within each source population.

The diagonal lines on Fig. 2 depict the transformation from log genotype probabilities to assignment scores. In our human blond example, this is the transformation from (0.75, 0.10, 0.40) to (60%, 8%, 32%). Points on the left, central, and right diagonal lines would be given an assignment score to the Broken Islands of 10%, 50%, and 90% respectively. A score of 90% to the Broken Islands means that the posterior probability of finding the rat's genotype in the Broken Islands is 9 times greater than the posterior probability of finding it in Aotea. It is worth pointing out that a multiplier of 9 is not very impressive when seen in the context of the within-population ranges of $10^5$ to $10^7$ described above; and from the chart it is clear that the band from 10% to 90% is very narrow. However, due to the substantial genetic differences in this example between the Broken Islands and Aotea populations, there are almost no points in this range. Almost all animals in this example would be given assignment scores of greater than 90% to their source population. It would be reasonable to undertake best-population assignment with this level of genetic distinction between populations.

### 2.3 Ecological interpretation

The ecological context of the Broken Islands study is invasive species management. New Zealand has no native land mammals, so its native ecosystems are extremely vulnerable to impacts of introduced mammals, including ship rats. Considerable efforts are devoted to establishing mammal-free island sanctuaries. Rats eat seeds and fruit, and predate directly on invertebrates, reptiles, and birds' nests. Through forest damage, competition, and direct predation, they have been solely responsible for the global extinction of several endemic bird and reptile species (e.g. Bell et al. 2016).

An eradication of ship rats on the Broken Islands was attempted in 2009 (Fewster et al. 2011). The Broken Islands reference population shown in Fig. 2

was sampled on the islands before the eradication took place. However, as rats are capable swimmers, there is a constant threat of reinvasion from the extant population on Aotea just 300 m away. In 2010, presence of rats was detected on the Broken Islands and 19 rats were trapped. The question of management interest is whether these rats had survived the eradication attempt, or whether they had swum from Aotea, perhaps with subsequent breeding on the Broken Islands. If they were survivors, this would necessitate a revision of eradication protocols, whereas if they were swimmers, this would inform and reinforce the need for managing the ongoing threat. Genetic assignment results from these 19 rats are plotted as diamonds on Fig. 2, from which it is clear that all 19 cluster convincingly with the Aotea population. It is not credible that any of these rats was sourced from the pre-eradication Broken Islands population, because they each possess too many alleles not found in the impoverished Broken Islands profile. The chart shows convincing evidence that the 'reinvader' rats were swimmers.

It is interesting that the genetic separation between Aotea and the Broken Islands prior to 2009 is decisive, despite the rapid reinvasion of the islands after the eradication. This might be due to a behavioural pattern called the *incumbent effect*, whereby the pre-2009 Broken Islands rats might have rejected swimmers from Aotea so that they did not contribute to the breeding population or genetic profile. We speculate that the readiness of incumbents to accept immigrants might be affected by the frequency of immigrants. The Broken Islands are buffered from Aotea by rugged cliffs on the Aotea side, so immigrants might occur at relatively low frequency and this might exaggerate an incumbent effect. Although speculative, these ideas have been reinforced by subsequent events. Since the reinvaders were trapped in 2010, the essential rat-free status of the Broken Islands has been maintained. However, new invaders were detected each year from 2011 to 2014, sometimes taking hold into a small population with genetic evidence of breeding on the islands, but always genetically aligned with the Aotea population rather than the pre-eradication Broken Islands population or the previous year's in-situ breeding. The level of reinvasion is frequent but not overwhelming, enabling the islands to be managed as sanctuary islands with strong reinvasion response procedures.

The right panel of Fig. 2 shows a second island system about 3 km north of the Broken Islands. Kaikoura Island (530 ha) is a larger island than the Broken Islands group (125 ha). Its closest approach to Aotea is over a water gap of only 80 m, although the terrain at this point is rugged; however there is also frequent boat traffic between Aotea and Kaikoura, and rats are known to hitch-hike on small craft. From Fig. 2, we see that the genetic profile of Kaikoura Island rats (circles) is much harder to distinguish from that of Aotea rats (squares) than was the case for the Broken Islands. Nearly all rats from either Aotea or Kaikoura fit between the 1% and 99% posterior LGP quantiles of the other population, meaning that they have an acceptable genetic fit to either population. This can be seen at a glance by noting that most rats are plotted inside the central box marked by the dashed quantile lines on Fig. 2. There is a hint of genetic subsetting for Kaikoura Island, but it is very minor. The diagonal lines show that several rats from both Aotea and Kaikoura would be given an assignment score greater than 50% for the wrong population, and occasionally greater than 90%. This underlines the danger of relying only on assignment scores without viewing the overall genetic variability in the populations to establish context.

A rat eradication was attempted on Kaikoura Island in 2008. The diamonds on the plot show eight rats captured on the island in 2009. The genetic evidence for any of these rats alone is inconclusive, as any of them could have originated either from the Kaikoura population denoting survivors, or from the Aotea population denoting swimmers or hitch-hikers. However, the eight rats do show a greater tendency to group with the Kaikoura population, and it is very unlikely that a group of eight Aotea rats would yield seven or more with higher genotype probabilities in Kaikoura than in Aotea ($p = 0.001$: Fewster et al. 2011). This gives us evidence that the post-eradication sample contains at least some survivors, although we do not wish to pronounce on the provenance of any of the rats individually. The conclusion of incomplete eradication is corroborated by the discovery of a different rat species, kiore (*Rattus exulans*) on Kaikoura Island from 2009 onwards. Kiore are thought to be non-swimmers, so it is likely that they were present on the island before the 2008 eradication attempt, undetected due to the presence of the more dominant ship rats, and that small numbers of both species survived the eradication attempt.

Subsequent events on Kaikoura Island have reinforced the conclusions from the genetic analysis. The ship rat population has persisted on the island since 2009 and is now managed as a controlled, low-density population. The genetic chart suggests that there is little isolation of the island population from Aotea, so it would be a significant challenge to maintain as a rat-free population. In 2013, it was confirmed that rats swim from Aotea to both Kaikoura Island and the Broken Islands using direct evidence from Rhodamine B dye (Bagasra et al. 2016). Bait laced with the dye was distributed on Aotea, and the dye was found during the following month in two males out of 39 ship rats trapped on Kaikoura Island, and in two isolated male ship rats found on the Broken Islands.

We give one final example of the insights that can be gained from genetic assignment data. Figure 3 shows GenePlots from two different species of rats taken from the Bay of Islands region in Northland, New Zealand. These GenePlots differ from Figure 2 because they involve more than two reference populations. The multi-dimensional LGP data is depicted on a two-dimensional chart by plotting the first two principal components. We lose the ability to depict posterior quantiles and assignment scores on these multi-population charts, but we still gain considerable insight into population structure and variation.

The left plot shows Norway rats (*Rattus norvegicus*) sampled in 2005 on five islands in the group: Urupukapuka (URU), Motuarohia (MAH), Waewaetorea (WAE), Okahu (OKA), and Poroporo (POR) (Miller et al. 2009). Four of these islands are in a chain with each pair separated by roughly 200–800 m. Motuarohia is a few kilometres away, separated from the others by two

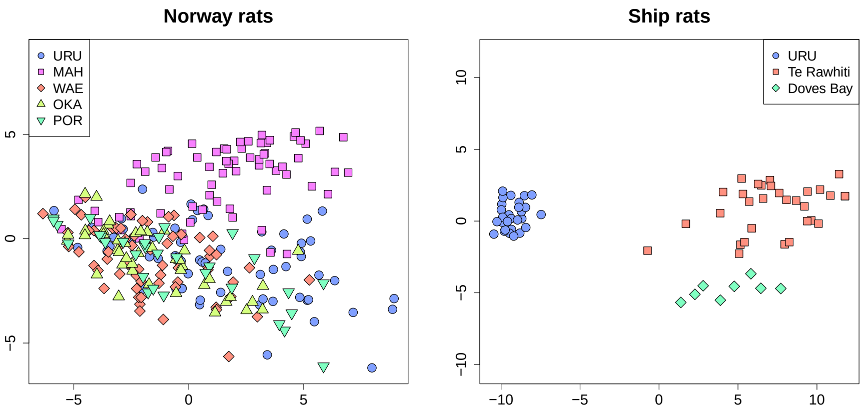

**Fig. 3** GenePlots of Norway rats and ship rats in the Bay of Islands region, New Zealand. Left: Norway rats from five islands in the bay. Islands are of moderate size, up to 208 ha. All five islands are used as reference populations. The first two principal components through the 5-dimensional plot are shown, accounting for a total of 91% variance explained comprising 60% and 31% on principal axes 1 and 2 respectively. Right: ship rats from Urupukapuka island and two mainland sites: Te Rawhiti and Doves Bay. All three populations are used as reference populations. The first two principal components through the 3-dimensional plot are shown and account for a total of 98% variance explained, comprising 92% and 6% on principal axes 1 and 2 respectively.

additional islands on which the rat populations were controlled or eradicated. Norway rats are thought to be more eager swimmers than ship rats (Russell et al. 2005) and the island terrain is mostly gentle with beaches at the entry and exit points on adjacent islands. Fig. 3 shows what we mean by an unstructured genetic profile among the islands. With the exception of the more distant Motuarohia (MAH), which separates from the others only in the direction of the second principal axis (vertical direction), the islands are genetically homogeneous. The first principal axis (horizontal direction) almost entirely describes within-population rather than between-population variation in LGP, and accounts for 60% of the total variance. In a situation like this it would be highly misleading to attempt genetic assignment based on best-population scores.

The right panel of Fig. 3 shows ship rats (*Rattus rattus*) from the same location. Ship rats were mostly absent from the islands, which were overrun by Norway rats, but a small population of ship rats was found on Urupukapuka, which boasts the only passenger ferry among the islands. Two adjacent mainland populations, Te Rawhiti and Doves Bay Marina, were also sampled. Te Rawhiti is about 1 km from Urupukapuka at closest approach, whereas Doves Bay is about 20 km from Urupukapuka by sea, and about 80 km from Te Rawhiti by land (Miller et al. 2009). The GenePlot in Fig. 3 shows a striking separation between Urupukapuka rats and those from the other two sites. Despite the very long land distance between Te Rawhiti and Doves Bay, there is no significant habitat break and the genetic separation between these two populations only occurs in the second principal axis (vertical direction) which explains only 6% of the total variance. The first principal axis (horizontal direction) explains 92% of the total variance and strongly distinguishes Urupukapuka rats from their mainland counterparts. Further inspection confirms that this is a similar process of genetic subsetting to that shown in the Broken Islands in Fig. 2, with the Urupukapuka population being genetically impoverished compared with the mainland populations.

Genetic assignment techniques have been widely applied to invasive species management in New Zealand, including studies of rats (Russell et al. 2010), stoats (*Mustela erminea*) (Veale et al. 2013), and possums (*Trichosurus vulpecula*) (Adams et al. 2014). Different genetic data types can also be used. Robins et al. (2016) use mitochondrial DNA to analyse the source of the disastrous 1964 ship rat invasion of Big South Cape Island, which resulted in the extinction of the last populations of three native bird and bat species (Bell et al. 2016).

### 2.4 Mathematical details

We now give the mathematical details underlying genetic assignment techniques, including the GenePlots described above. We use the Bayesian formulation of Rannala and Mountain (1997), which underpins many similar methods.

Consider a single locus, $L$, at which there are $k$ available allele types, labelled $i = 1, 2, \ldots, k$. The parameters that need to be estimated are the frequencies of alleles $1, 2, \ldots, k$ in each reference population. For a single reference population $R$, let $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$ be the frequencies of the $k$ alleles, where $\sum_{i=1}^{k} p_i = 1$ and $0 \leq p_i \leq 1$ for $i = 1, \ldots, k$. Our aim is to estimate $p_1, \ldots, p_k$ using sample data from the reference population $R$, and then to use these estimates to assess the multilocus genotype probability of any queried individual $I$ with respect to population $R$. This genotype probability is log-transformed to give the LGP of individual $I$ in population $R$.

A Bayesian approach to estimating $(p_1, \ldots, p_k)$ is useful, because it allows for an unsampled allele to have non-zero posterior weight. The amount of posterior weight can be adjusted according to the size of the sample that failed to find the allele. This means that an individual with allele $i$ is not excluded from population $R$ even if allele $i$ was not sampled among the reference samples from population $R$. This is important, because as we have seen there are typically many rare alleles present in a population, and they will not all be exposed in the reference sample.

We use a Dirichlet prior, $(p_1, p_2, \ldots, p_k) \sim \text{Dirichlet}(\tau, \tau, \ldots, \tau)$, where $\tau$ is usually chosen to be either 1 or $1/k$. The prior density is

$$f(p_1, p_2, \ldots p_k) \propto \prod_{i=1}^{k} p_i^{\tau - 1} ,$$

for $0 < p_i < 1$ $(i = 1, \dots, k)$ and $\sum_{i=1}^{k} p_i = 1$.

The data are the numbers of alleles of each type observed among the reference individuals sampled from population $R$, denoted by $\boldsymbol{x} = (x_1, x_2, \dots, x_k)$. Here, $\sum_{i=1}^{k} x_i = 2n$, where $n$ is the number of reference individuals for which genotype data was successfully obtained at locus $L$.

The likelihood is the multinomial density, $\boldsymbol{x} \,|\, \boldsymbol{p} \sim \text{Multinomial}(2n\,; \boldsymbol{p})$. The multinomial model requires that the $2n$ alleles found in the $n$ reference individuals correspond to $2n$ independent draws from the reference population allele frequencies. This implies that there should not be correlation between an individual's two alleles, a requirement that is satisfied if the population is in *Hardy-Weinberg equilibrium* but is violated if the population is substantially inbred. Sources of genetic dropout such as null alleles also interfere with the multinomial assumption. However, empirical investigations suggest that GenePlot charts are quite robust to violations of the multinomial model, because inference derives mainly from the allele frequencies themselves rather than from the particulars of how alleles are combined into genotypes; so issues of inbreeding and dropout tend to be disregarded in assignment analyses.

Because the Dirichlet distribution is the conjugate prior of the multinomial, the posterior allele frequency distribution is also Dirichlet:

$$(p_1, p_2, \dots p_k \,|\, \boldsymbol{x}) \sim \text{Dirichlet}(x_1 + \tau, x_2 + \tau, \dots, x_k + \tau)\,.$$

If allele $i$ is unsampled in reference population $R$ $(x_i = 0)$, there is nonetheless still posterior support for values $p_i > 0$. Larger reference samples drive this support closer to zero, but it never vanishes altogether.

Now consider a query individual, $I$, whose LGP we wish to assess in reference population $R$. The genotype of individual $I$ at locus $L$ consists of two alleles, and can be written as $\boldsymbol{a} = (a_1, a_2, \dots, a_k)$, where each $a_i$ is 0, 1, or 2, and $\sum_{i=1}^{k} a_i = 2$. As before, we assume that the individual's two alleles are independent, so $\boldsymbol{a} \,|\, \boldsymbol{p} \sim \text{Multinomial}(2\,; \boldsymbol{p})$. The marginal distribution of $\boldsymbol{a}$ is the Dirichlet compound multinomial distribution, obtained by integrating the multinomial density over the Dirichlet posterior of $\boldsymbol{p}$, and it simplifies to a simple closed form as follows:

$$P(\boldsymbol{a}) = \begin{cases} \dfrac{(x_r + \tau)(x_r + \tau + 1)}{(2n + k\tau)(2n + k\tau + 1)} & \text{if } a_r = 2 \text{ and } a_j = 0 \text{ for } j \neq r\ : \\ & \quad I \text{ is a homozygote with allele type } r; \\ \dfrac{2(x_r + \tau)(x_s + \tau)}{(2n + k\tau)(2n + k\tau + 1)} & \text{if } a_r = a_s = 1 \text{ and } a_j = 0 \text{ for } j \notin \{r, s\}\ : \\ & \quad I \text{ is a heterozygote with alleles } r \text{ and } s. \end{cases}$$

These expressions demonstrate that the Bayesian procedure has generated a posterior probability for genotype $(r, s)$ that is greater than 0 even if alleles $r$ and $s$ were unsampled in the reference population $(x_r = x_s = 0)$, but that the posterior probability allotted to such a genotype decreases as the size $n$ of the reference sample increases. The posterior log genotype probability (LGP) of individual $I$ at locus $L$ in population $R$ is finally given by $\log_{10}\{P(\boldsymbol{a})\}$.

The typical choices for the prior parameter $\tau$ are $\tau = 1$ (Baudouin and Lebrun 2001) or $\tau = 1/k$ (Rannala and Mountain 1997). The posterior $\boldsymbol{p} \,|\, \boldsymbol{x} \sim \text{Dirichlet}(\boldsymbol{x} + \tau)$ has marginals $p_r \,|\, \boldsymbol{x} \sim \text{Beta}\Big(x_r + \tau,\ \sum_{i=1}^{k}(x_i + \tau) - (x_r + \tau)\Big)$, so the marginal posterior means are

$$E(p_r \,|\, \boldsymbol{x}) = \frac{x_r + \tau}{\sum_{i=1}^{k}(x_i + \tau)} = \frac{x_r + \tau}{2n + k\tau} \,.$$

This shows that the choice $\tau = 1/k$ gives less posterior weight to alleles with low sample frequencies. The choice $\tau = 1$ borrows probability from common alleles to allot to rare ones, so it is more tolerant of rare alleles than the choice $\tau = 1/k$. We use the choice $\tau = 1$ throughout this paper. GenePlots produced with $\tau = 1/k$ are similar but tend to be a little more diffuse because individuals with rare alleles are allotted lower posterior LGPs and therefore drag out the lower tail of the LGP distribution.

The calculation above gives the LGP $\log_{10}\{P(\boldsymbol{a})\}$ for individual $I$ in population $R$ at a single locus $L$, which we could write as $\log_{10}\{P(\boldsymbol{a}_L)\}$. The overall multilocus log-genotype probability for individual $I$ in population $R$ is gained by summing over loci $L = 1, \ldots, \ell$: $\text{LGP}_R^I = \log_{10}\{P(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_\ell)\} = \sum_{L=1}^{\ell} \log_{10}\{P(\boldsymbol{a}_L)\}$. This corresponds to an assumption that an individual's alleles are independent across different loci, which in genetic terminology is to say that the loci are in *linkage equilibrium.* In theory, loci are statistically independent if there is no physical link between them, for example if they are located on different chromosomes or are sufficiently far apart on a single chromosome not to be inherited as a single unit, which is very likely to be the case in practice. However, in small populations, correlation between alleles at different loci can arise as a sampling artefact, where 'sampling' denotes the genetic sampling process of creating offspring from a small number of parents. It is good practice to check for linkage disequilibrium before undertaking an assignment analysis, but violations are unlikely to pose serious problems unless they are extreme, because again inference is based primarily on allele frequencies and not on their assembly into multilocus genotypes.

If there are no missing data, the GenePlot is constructed by plotting the point for individual $I$ at coordinate $(\text{LGP}_1^I, \ldots, \text{LGP}_K^I)$ for populations $R = 1, \ldots, K$. The difficulty comes when individual $I$ has missing data at some loci, because then its log genotype probabilities are on a different scale from those of other individuals. For example, if $I$ has data available for only 8 out of 10 loci, its LGP coordinates are gained by adding the results for only 8 loci, whereas those for full-data individuals are gained by adding the results for 10 loci. This non-comparability is the reason why most studies do not attempt a graphical analysis as shown on the GenePlot. Missing locus data ostensibly imply that the LGPs of different individuals within the sample data are referenced on a multitude of different scales, and these missing data are sufficiently common that it is impracticable to discard all individuals with missing records from the analysis. However, the benefits of plotting the output would seem to outweigh the disadvantages of dealing with missing data, so we proceed by developing a graphical display that can handle these missing records.

We deal with missing data on GenePlots by plotting individuals at the LGP quantiles in the full-data distribution that they obtain from their available loci in the corresponding reduced-locus distribution. Thus, if individual $I$ has data available for only 8 loci, we construct the LGP of $I$ in population $R$ initially for these 8 loci only, and find the quantile of $\text{LGP}_R^I$ in the posterior distribution of the 8-locus $\text{LGP}_R$. If individual $I$ is at the 20th percentile of this 8-locus distribution, then its coordinate for population $R$ in the GenePlot is the 20th percentile of the full 10-locus posterior distribution for $\text{LGP}_R$. This ensures that all individuals are plotted at points that preserve their observed 'rarity' within each population on the basis of the data they possess. We construct the posterior $\text{LGP}_R$ distribution by finding the distribution of LGPs in population $R$ over multilocus genotypes drawn from the posterior Dirichlet compound multinomial distributions for each locus $L$ in population $R$. Quantiles of the posterior $\text{LGP}_R$ distribution may be found either by simulation (Russell et al. 2010; Veale et al. 2013) or using a saddlepoint approximation (McMillan and Fewster in review). A user interface for generating GenePlots using R is available online (McMillan and Fewster in review).

## 2.5 Related genetic assignment methods

There is a large literature on genetic assignment and on eliciting population structure from genetic data, and a range of methods each with their own strengths and caveats. While we cannot attempt a complete survey here, many of the methods in common use share a foundation with the material in the previous sections, so a detailed look at one method as we have provided above enables a quick grasp of many more. Here we briefly mention some additional methods and software in wide use.

— *GeneClass2.* In addition to the tables of percentage assignment scores described in Section 2.1, the GeneClass2 software (Piry et al. 2004) also offers several other options. LGP results and the number of non-missing loci for each individual are returned in the same tabular format, so GenePlots can be plotted directly from the tables as long as they are restricted only to those individuals with no missing data. If the GenePlot is to include individuals with missing data, custom code is needed for quantile calculation (McMillan and Fewster in review).

— *Monte Carlo resampling.* A feature of GeneClass2 that deserves special mention is a suite of Monte Carlo algorithms available under the label 'Probability Computation'. These algorithms create virtual individuals 'bred' from the reference samples to generate a population quantile for the LGP of each real individual $I$ in each population $R$. For example, the algorithm of Paetkau et al. (2004) repeatedly generates new samples of $n$ individuals bred virtually from the $n$ real individuals in the reference sample from population $R$. Each of the $n$ virtual individuals is given an LGP result using the other $n - 1$ individuals in its own batch as a reference population. New batches are simulated until a large sample of LGPs is obtained, pooled across batches. The LGP of a query individual $I$ in the observed data, with reference to the $n$ real reference samples from population $R$, is then compared against this large sample of simulated LGPs. Its ranking among the simulated samples produces a $p$-value against the null hypothesis that individual $I$ could have arisen from population $R$. The output generates similar conclusions to those from the GenePlot posterior quantiles, but the method of estimating the quantiles is different.

— *Mixed stock analysis.* Many migratory species, particularly fish, whales, and other marine species, have high fidelity to breeding sites such as natal rivers or specific coastal areas, but also undergo migrations to areas such as pelagic feeding grounds where the different stocks mix. Mixed stock analyses are used to determine the proportional composition of a population that is thought to contain a mixture of individuals from different breeding stocks. The output of a mixed stock analysis is an estimate of stock composition: for instance, estimating that the mixed stock comprises 20% sourced from population A, 50% from population B, and 30% from population C. As such, the analysis aims to be less specific than individual assignment procedures. However, since it is based on the same principles, it is subject to much the same caveats and considerations. In particular, genetic assignment does not benefit from a law of averages, so there is no reason to expect that a mixed stock analysis will be more successful than an individual assignment exercise if there is only weak genetic discrimination between the source populations.

Mixed stock analysis is conducted by software such as ONCOR (Anderson et al. 2008; Kalinowski et al. 2008), popular in fisheries management. For three reference populations A, B, and C, ONCOR aims to estimate $\boldsymbol{\theta} = (\theta_A, \theta_B, \theta_C)$, where $\theta_R$ is the proportion of the queried stock that is sourced from population $R$ for $R \in \{A, B, C\}$ with $0 \leq \theta_R \leq 1$ and $\sum_R \theta_R = 1$. To estimate $\boldsymbol{\theta}$, ONCOR first calculates the posterior log genotype probabilities (LGPs) for every individual $I$ in the queried stock, following the Rannala and Mountain (1997) method as described above. Taking antilogarithms produces the posterior genotype probabilities themselves, $(\mathrm{GP}_A^I, \mathrm{GP}_B^I, \mathrm{GP}_C^I)$. The probability of finding $I$'s genotype in the mixed stock is then $\theta_A \mathrm{GP}_A^I + \theta_B \mathrm{GP}_B^I + \theta_C \mathrm{GP}_C^I$. This probability is multiplied across all individuals $I$ in the query sample to gain a likelihood for $\boldsymbol{\theta}$, which is maximized to estimate the stock composition.

— *Cryptic population structure.* The software Structure (Pritchard et al. 2000) is an immensely popular package for eliciting cryptic population structure from a genetic sample. By cryptic structure, we mean genetic subsets that are not necessarily associated with their sampling location. For example, humans in a large city might tend to marry within their own ethnic groups, leading to genetic structure in the population that is not connected with location. The Structure software is effectively a genetic clustering algorithm. The number of clusters sought, $K$, is pre-specified by the user. Typically the software is run by trialling several different values for $K$, but alternatives for selecting $K$ are available (e.g. Evanno et al. 2005), including software that uses the Dirichlet process prior (Pella and Masuda 2006; Huelsenbeck and Andolfatto 2007).

Structure operates on much the same principles as the Rannala and Mountain (1997) assignment methods described above, except for extra steps needed to assign cluster memberships for all individuals. Cluster membership is treated as a latent variable and sampled by a Markov chain Monte Carlo algorithm. In essence, Structure elicits its own $K$ reference populations based on clustering the LGP results, as opposed to the GenePlot method which assumes reference populations are known *a priori.* It is often run with an *admixture model* which allows each individual to have mixed cluster membership, with compositions estimated separately for each individual. This model is perhaps rather obscure as far as biological interpretability goes, but it is popular because it enables the uncertainty in cluster membership to be displayed for each individual. The final output is a barchart with clusters discriminated by colour. A bar for each individual displays its estimated composition by cluster: for example, a single individual may be attributed 20% to population 1 (coloured red) and 80% to population 2 (coloured green), leading to a bar split 20-80 between the two colours. A typical output will show a mixture of some individuals estimated to have 'pure' membership of a single cluster, and other individuals estimated to be composed of different clusters to greater or lessser extent.

## 3 Estimating population size with genotype data

In this section we look at a second major area in which genetic data can be useful in statistical ecology: estimating population size. The most straightforward application of genetic data in estimating population size is to treat individual DNA profiles as unique 'marks' for a capture-recapture study. In principle, capture histories can be reconstructed for all sampled animals by treating every unique genotype profile as a unique animal, and applying standard capture-recapture models (Otis et al. 1978). In practice, this is complicated by genotyping errors and missing data which can create differences between DNA profiles obtained from the same animal. We look briefly at how these problems have been approached by statisticians to date.

Other methods for estimating population size rely more directly on properties of genetic inheritance and genetic drift. Here, it is important to distinguish between methods that aim to estimate census population size — the number of animals in a population, $N$ — and those that aim to estimate *genetic effective population size*, $N_e$. Effective population size is a genetic measure that is related to the rate of genetic change in a population, and it does not necessarily relate to census population size in a predictable or temporally-stable way. We comment briefly below on close-kin mark-recapture, an emerging method of estimating census population size, and on the differing aim of estimating genetic effective population size.

### 3.1 Genetic capture-recapture

We use a case study to illustrate issues of genetic sample-matching that arise when reconstructing capture histories from genotype data. Carroll et al. (2011; 2013) describe a capture-recapture study of southern right whales (*Eubalaena australis*) conducted by boat in the New Zealand subantarctic over the four austral winters of 1995-1998. DNA samples from cetaceans are obtained by deploying biopsy darts from a veterinary rifle. The biopsy darts glance off the thick outer blubber of cetaceans, scooping a tiny skin sample on impact, and drop into the sea where they are retrieved by researchers. Darts may be attached to fishing lines and retrieved by reeling in the line after deployment, or they can be fished out of the water using nets. DNA samples obtained from live animal tissue, such as these, are generally of higher quality than those obtained from dropped samples such as hair, feathers, or faeces; however, even with high-quality tissue samples, considerable attention must be paid to genotyping errors.

The following statistics are taken from Vale et al. (2014). Each sample is genotyped at 13 microsatellite loci. There are 132 genetic samples, corresponding to results for $132 \times 13 = 1716$ loci. Of these, 139 or 8% of locus records are missing.

Reconstructing capture histories involves comparisons between all pairs of samples to determine which samples correspond to captures of the same individual. Only one pair of samples in the right whale dataset exhibits a full match on all 13 loci. If problems of genetic errors and missingness were ignored, this would mean only one recapture would be reported for the entire study, and population size would be greatly overestimated. Applying the classical model $M_t$ under this strategy gives an implausible estimate of $\widehat{N} = 6148$ whales (Vale et al. 2014). The true number $N$ is believed to be a few hundred animals.

The usual way of dealing with errors is to conduct a thorough manual examination of near-matches, often repeatedly genotyping samples over which there is doubt, and eventually deciding upon a rule for calling matches. This can be a time-consuming and expensive process. In this study, we find that out of $132 \times 131/2 = 8646$ sample pairs, the number of pairs with exact matches at 0–4 loci is 8621; at 5–6 loci is 5; at 7–8 loci is 0; and at 9–13 loci is 20. Missing data are not counted as matches for these statistics. The results exhibit a clear break between samples matching at 6 or fewer loci, and samples matching at 9 or more loci.

Using the least variable 9 loci in the data, the estimated probability that two individuals have the same genotype by chance is $\text{PID} = 6.0 \times 10^{-11}$ (Paetkau and Strobeck 1994), or for closely related individuals, $\text{PIDsib} = 1.5 \times 10^{-4}$ (Evett and Weir 1998). Other selections of 9 loci have even better discrimination, so the probability of 9-locus matches occurring by chance is always lower than about 1 in 10 000. By contrast, for the 6 most variable loci, $\text{PID} = 1.8 \times 10^{-9}$ and $\text{PIDsib} = 1.3 \times 10^{-3}$. For other selections of 6 loci, these probabilities are higher, so the probability of 6-locus matches occurring by chance is greater than 1 in 1000 for close relatives. These figures indicate that, if the population size really is a few hundred animals including numerous close relatives, it is reasonable to assume that a few 6-locus matches but no 9-locus matches will be obtained between samples belonging to different individuals. The break between 6 and 9 matching loci provides a convenient boundary for this study, so we can reconstruct capture histories by assuming that two samples belong to the same individual if they have at least 9 matching loci.

Assuming the 9-loci match rule is correct, we can deduce the level and types of error in the sample. The match rule implies there are 60 non-matching loci among samples assumed to come from the same individual, out of a total of 260 same-locus comparisons. Of the 60 non-matches, 50 are due to missing data. The other 10 mismatched loci can only have arisen from errors: 6 could be due to allelic dropout; 3 have a single allele substitution; and in the remaining case the non-matching loci have no alleles in common. Thus, every type of non-match appears in the data. The error rate by locus among non-missing data is 10/200 (5%).

Once the data are presumed to be corrected, capture-recapture modelling proceeds as usual. Applying model $M_t$ to the data obtained from the 9-loci match rule yields an estimate of $\widehat{N} = 306$ whales with 95% confidence interval (212, 443) (Vale et al. 2014).

For sampling protocols involving low-quality DNA, for example using hair, feathers, or faeces, a much higher error rate may be expected. In principle, identity can be established to near-certainty regardless of the error rate by taking sufficiently many loci. However, this is not always a practical possibility using microsatellite loci, both because of the expense of genotyping, and because microsatellite loci and primers are costly to develop and there might only be a restricted number commercially available for a particular species. These problems may be solved in the future by next-generation sequencing technologies, which allow examination of thousands of loci. Distinguishing individual identity and kinship are among the most straightforward and powerful applications promised by next-generation technologies.

## 3.2 Modelling misidentification

In view of the intrinsic difficulties in matching DNA samples to the same animal, various authors have proposed ways of allowing for genotyping errors in capture-recapture data at the modelling stage. The aims are twofold. Firstly, it is time-consuming and expensive in laboratory work to verify problematic samples: it is sometimes said that 95% of laboratory effort is expended on 5% of samples. Secondly, incorporating genotyping errors at the modelling stage enables quantification of the error rate and the uncertainty that errors contribute to the final results, which are ignored if the data set is patched up to a final version that is treated as fixed and correct for the modelling exercise.

The misidentification model that has perhaps received the most attention is model $M_{t,\alpha}$ (Lukacs and Burnham 2005; Yoshizaki et al. 2011; Link et al. 2010; McClintock et al. 2014; Schofield and Bonner 2015), which is similar to the classical model $M_t$ (Otis et al. 1978) but with the addition of a simple misidentification mechanism. Each sample is considered to be correctly genotyped with probability $\alpha$. With probability $1 - \alpha$, it is incorrectly genotyped and a unique erroneous DNA profile is generated. It is assumed that the same genotyping error never occurs twice, so every error leads to a capture history with a single capture in it. Information to estimate the parameter $\alpha$ comes from the consequent surplus of capture histories with only one capture.

Vale et al. (2014) highlighted two significant problems with model $M_{t,\alpha}$. Firstly, the model itself is too simplistic to capture the genotyping error process adequately. The assumptions that genotypes are either correct or incorrect, and that samples either match or do not match, is not a good description of the error process, and does not make use of the information that a pair of samples matching at 12 out of 13 loci almost certainly belong to the same animal, whereas a pair matching at 4 out of 13 loci almost certainly belong to different animals. Secondly, the model is data-hungry, with very large sample sizes needed for precise maximum likelihood estimates. Gleaning information on the misidentification rate from the surplus of capture histories with only one capture is an ingenious idea, but it is too subtle for the sample sizes often encountered in real studies. A large number of single-entry capture histories could be attributed either to low capture probabilities with a low error-rate, or to high capture probabilities with a high error-rate. Consequently, unless the sample sizes are very large, the $\alpha$ parameter, and consequently the population size, are estimated with low precision (Vale et al. 2014).

For the southern right whale study, Vale et al. (2014) found that model $M_{t,\alpha}$ gave poor results. When applied to the uncorrected data, as it is intended to be, there is only one recapture in the data set so nearly all capture histories contain only one entry. The model drives $\widehat{\alpha}$ as low as possible, yielding a boundary estimate with $\widehat{N} = \max_t\{n_t\} = 51$ and $\widehat{\alpha} = 0.09$. When instead it is applied to the corrected data using the 9-locus match rule, it returns the opposite boundary estimate $\widehat{\alpha} = 1.00$ and gives identical results to model $M_t$, namely $\widehat{N} = 306$, rendering the misidentification mechanism redundant. Vale et al. (2014) note that there does not appear to be a satisfactory application of model $M_{t,\alpha}$ on real data in the literature to date.

A different approach to modelling misidentification is taken by Wright et al. (2009), and further developed in Barker et al. (2014). Instead of estimating error rate indirectly through a surplus of single-entry capture histories, they require all samples to be genotyped at least twice, therefore gaining a direct estimate of error rate at each locus by discrepancies between repeat attempts. Modelling proceeds by treating true genotypes and capture histories as latent variables to be sampled through an MCMC algorithm. The complete data likelihood demands parameters for the probabilities of all genotypes at each locus, so the approach is parameter-intensive. However, it is very appropriate in situations where large numbers of low-quality DNA samples are available: for example, in studies that collect feathers, hair, or faeces. In these cases, potentially large sample sizes mitigate parametrization problems, and for such low-quality DNA samples it is standard protocol to conduct repeat genotyping (Taberlet and Luikart 1999). Because the model uses direct information on genotyping errors, these parameters are well-informed. Wright et al. (2009) report the posterior medians of locus-specific dropout probabilities to be between 0.12 and 0.35 in their study of faeces from European badgers (*Meles meles*) in Gloucestershire, UK. These high estimates, compared with the empirically-estimated error rate of 0.05 in the right whale data set, highlight the differences in genotype quality when using samples from faeces as opposed to high-quality tissue samples.

Despite its advantages, the approach of Wright et al. (2009) is not ideal for situations such as the right whale study, for which repeat-genotyping is not a cost-effective use of resources in view of the high quality of DNA samples and low error rate. Furthermore, cetacean surveys have notoriously low power to detect population change (Carroll et al., 2015), and it is unlikely that data will be capable of supporting a heavy parametrization for genotyping error while still adequately addressing questions of interest. The state of the art for such data is still to correct errors by a manual process prior to modelling, as described for the right whale study. This leaves the field open for further statistical development. It remains to be seen whether developments in next-generation sequencing might largely solve the problem of genetic misidentification. Researchers with substantial microsatellite catalogues from long-running ongoing studies might then face the dilemma of continuing with existing genetic protocols, or recreating their entire catalogues with new technologies.

### 3.3 Close-kin mark-recapture

A promising new direction for estimating population size from genetic data is *close-kin mark-recapture* (Bravington et al. 2016). The simplest formulation relies upon the observation that every individual has two parents. It operates on similar principles to capture-recapture, except that an individual is ‘marked’ by its own presence in the sample, and ‘recaptured’ if one or more of its parents is also present in the sample: an event that is intuitively more likely in a small population than a large one, for a sample of a given size. The close-kin recapture rate therefore contains information about adult population size. Estimation can be conducted on a single sample — in other words from a single capture occasion — but is complicated by the possibility that parents might have died before the sample is taken. This forces the inclusion of a wider demographic model which enables estimation of additional demographic parameters such as parental mortality. With modern genotyping methods, it appears possible to extend the approach to more distant kin such as half-siblings.

Close-kin mark-recapture methods offer innovative new ideas for estimating census population size from genetic data, especially in large-population settings such as commercial fisheries where other data sources can be unreliable. The first large-scale application is a recent study to estimate abundance of southern bluefin tuna (*Thunnus maccoyii*) (Bravington et al. 2014).

### 3.4 Genetic effective population size, $N_e$

In this final section we briefly introduce the concept of genetic effective population size, $N_e$. The primary purpose is to describe what $N_e$ represents, and to distinguish it from the usual meaning of population size, namely the number of individuals in a population. We distinguish this usual quantity by calling it the *census population size*, $N$.

The process of genetic inheritance through the generations can be thought of as a sampling process. Alleles available in the parent generation are sampled to create a new set of alleles for the offspring generation: a process called *genetic sampling* (Weir 1996). As such, familiar effects of sample size come into play. The sample proportion of a particular allele $A$ can change much more from one generation to the next in a small population than it can in a large population. This change in allele frequency from one generation to the next is called *genetic drift*.

Because the rate of change of genetic quantities from one generation to the next depends upon the population size, we should be able to use information about the rate of change of genetic quantities to provide information about population size. Suitable genetic quantities whose rate of change depends upon population size include allele frequencies, inbreeding coefficients, and homozygosity levels. However, genetic models that link the change in these quantities to the population size $N$ are highly idealized, and do not necessarily describe the reproduction of real animals. In particular, in real populations some individuals are more successful breeders than others — massively so in some species — which can be thought of as reducing the pool of alleles available for genetic sampling to a smaller pool belonging to only the successful breeders. The *effective* population size that governs rates of genetic change is therefore typically smaller than the census population size.

The formal definition of effective population size $N_e$ is rather subtle. The effective population size is the size of an idealized population whose genetic parameters change at the same rate as those in the population of interest. The *ideal* population meets the three conditions of equal sex ratio, random mating, and constant census population size over generations, and generations do not overlap. The idea is that the real population, with a census size of $N$ individuals per generation, can then be studied in genetic terms as if it were an ideal population with size $N_e$ individuals.

Generally, because of the uneven breeding success of individuals, the effective population size $N_e$ is smaller than the census population size $N$. How much smaller depends upon the species and mating system. For some species, $N_e$ might be comparable with $N$, whereas for other species it could be millions of times smaller — a ratio observed in some fish species, for example. For any given species, the ratio of $N_e$ to $N$ is not constant or predictable over time, and there are several possible definitions of $N_e$ depending upon which genetic parameters are inspected and the timescale of interest. A comparison of $N_e/N$ ratios across different species is given by Frankham (1995).

Although $N_e$ is a parameter of fundamental importance in evolutionary genetics, determining the potential of a population to retain advantageous alleles rather than lose them to genetic drift, it is less clear how useful it is for contemporary conservation or management. While genetic parameters such as inbreeding coefficients are themselves relevant to conservation management, the transformation of these into $N_e$ does not appear to be especially helpful. It might be useful in the management of threatened species as a way of communicating the severity of genetic impoverishment to laypeople, although this benefit is counterbalanced by the difficulty of obtaining a precise estimate of contemporary $N_e$ for small populations, and the difficulty of quantifying what number would represent a 'healthy' $N_e$ for the population in question. Our main aim here is to ensure that the use of genetic data to estimate census population size $N$ is not confused with the estimation of the genetic population size parameter $N_e$. Further discussion and references for estimating $N_e$ can be found in Russell and Fewster (2009) and Luikart et al. (2010).

## 4 Concluding remarks

Our aim in this paper has been to give an introduction to two key applications of genetic data that are likely to be encountered by statistical ecologists: genetic assignment and population structure; and population size estimation using genetic data. There are numerous other applications of genetic data that we have not mentioned. Relatedness studies, including parentage assignment and pedigree reconstruction, have enormous applications, from designing breeding programmes for critically endangered species to avoid inbreeding and maximize the genetic health of a population (so-called *genetic rescue*), to invasive species management with the aim of determining whether a sample of reinvaders comprises independent colonists or a newly-established breeding population. An emerging area that is likely to be the focus of much future statistical work concerns the merging of genetic data with data from other sources for combined inference.

We have barely touched on the enormous field of classical population genetics, including the foundational Wright-Fisher model and associated concepts of inbreeding and coancestry coefficients (Weir 1996). The coancestry coefficient may be loosely referred to as $F_{ST}$ and is often used as a measure of population structure or connectivity: see Fewster et al. (2011) for how this measure can be used for a connectivity analysis of the ship rat data featured in this paper. Despite the many omissions, it is hoped that the concepts covered here will provide a worthwhile introduction to genetic principles and problems, and enable an easier route into further study.

**Acknowledgements** This work was funded by the Royal Society of New Zealand through Marsden grant 03-UOA-117. Thanks to Emma Carroll who checked the manuscript for biological soundness, and to Mark Bravington for comments on close-kin mark-recapture and other genetic methods. Many thanks to the associate editor and referees whose insightful comments greatly improved the paper.

## References

Adams, A.L., van Heezik, Y., Dickinson, K.J.M., Robertson, B.C.: Identifying eradication units in an invasive mammalian pest species. Biol. Invasions 16, 1481–1496 (2014)

Anderson, E.C., Waples, R.S., Kalinowski, S.T.: An improved method for predicting the accuracy of genetic stock identification. Can. J. Fish. Aquat. Sci. 65, 1475–1486 (2008)

Bagasra, A., Nathan, H.W., Mitchell, M.S., Russell, J.C.: Tracking invasive rat movements with a systemic biomarker. NZ J. Ecol. 40, 267–272 (2016)

Barker, R.J., Schofield, M.R., Wright, J.A., Frantz, A.C., Stevens, C.: Closed-population capture-recapture modeling of samples drawn one at a time. Biometrics 70, 775–782 (2014)

Baudouin L., Lebrun P.: An operational Bayesian approach for the identification of sexually reproduced cross-fertilized populations using molecular markers. Acta Hortic. 546, 81–94 (2001)

Bell, B.D., Bell, E.A., Merton, D. The legacy of Big South Cape: rat irruption to rat eradication. NZ J. Ecol. 40, 205–211 (2016)

Bravington, M.V., Skaug, H.J., Anderson, E.: Close-kin mark-recapture. Stat. Sci. in press (2016)

Bravington, M.V., Grewe, P.G., Davies, C.R.: Fishery-independent estimate of spawning biomass of Southern Bluefin Tuna through identification of close-kin using genetic markers. FRDC Report 2007/034, CSIRO, Australia (2014)

Carroll, E.L., Brooks, L., Baker, C.S., Burns, D., Garrigue, C., Hauser, N., Jackson, J.A., Poole, M.M., Fewster, R.M.: Assessing the design and power of capture-recapture studies to estimate demographic parameters for the endangered Oceania humpback whale population. Endanger. Species Res. 28, 147–162 (2015)

Carroll, E.L., Childerhouse, S.J., Fewster, R.M., Patenaude, N.J., Steel, D., Dunshea, G., Boren, L., Baker, C.S.: Accounting for female reproductive cycles in a superpopulation capture-recapture framework: application to southern right whales (*Eubalaena australis*). Ecol. Appl. 23, 1677–1690 (2013)

Carroll, E.L., Patenaude, N.J., Childerhouse, S.J., Kraus, S.D., Fewster, R.M., Baker, C.S.: Abundance of the New Zealand subantarctic southern right whale population estimated from photo-identification and genotype mark-recapture. Mar. Biol. 158, 2565–2575 (2011)

Chapuis M.-P., Estoup A.: Microsatellite null alleles and estimation of population differentiation. Mol. Biol. Evol. 24, 621–631 (2007)

Evanno, G., Regnaut, S., Goudet, J.: Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14, 2611–2620 (2005)

Evett, I., Weir, B.: Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists. Sinauer, Sunderland (1998)

Fewster, R.M., Miller, S.D., Ritchie, J.: DNA profiling — a management tool for rat eradication. In: Veitch, C.R., Clout, M.N., Towns, D.R. (eds.) Island invasives: Eradication and Management, pp. 430–435. IUCN, Gland, Switzerland (2011)

Frankham, R.: Effective population size/adult population size ratios in wildlife: a review. Genet. Res. 66, 95–107 (1995)

Huelsenbeck, J.P., Andolfatto, P.: Inference of population structure under a Dirichlet process model. Genetics 175, 1787–1802 (2007)

Jacob, H.J., Brown, D.M., Bunker, R.K., Daly, M.J., Dzau, V.J., Goodman, A., Koike, G., Kren, V., Kurtz, T., Lernmark, Å., Levan, G., Mao, Y.-P., Pettersson, A., Pravenec, M., Simon, J.S., Szpirer, C., Szpirer, J., Trolliet, M.R., Winer, E.S., Lander, E.S.: A genetic linkage map of the laboratory rat, *Rattus norvegicus*. Nat. Genet. 9, 63–69 (1995)

Luikart, G., Ryman, N., Tallmon, D.A., Schwartz, M.K., Allendorf, F.W.: Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. Conserv. Genet. 11, 355–373 (2010)

Kalinowski, S.T., Manlove, K.R., Taper, M.L.: ONCOR: a computer program for genetic stock identification, v.2. www.montana.edu/kalinowski/Software/ONCOR.htm. Dept. Ecology, Montana State University, Bozeman, USA (2008)

Link, W.A., Yoshizaki, J., Bailey, L.L., Pollock, K.H.: Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. Biometrics 66, 178–185 (2010)

Lukacs, P.M., Burnham, K.P.: Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error. J. Wildl. Manag. 69, 396–403 (2005)

McClintock, B.T., Bailey, L.L., Dreher, B.P., Link, W.A.: Probit models for capture-recapture data subject to imperfect detection, individual heterogeneity and misidentication. Ann. Appl. Stat. 8, 2461–2484 (2014)

McMillan, L.F., Fewster, R.M.: Visualizing genetic distributions for assignment tests using the saddlepoint approximation method. In review.

Miller, S.D., MacInnes, H.E., Fewster, R.M.: Detecting invisible migrants: an application of genetic methods to estimate migration rates. In: Thomson, D.L., Cooch, E.G., Conroy, M.J. (eds.) Modeling Demographic Processes in Marked Populations, pp. 417–437. Environmental and Ecological Statistics Series, Vol 3, Springer, Berlin (2009)

Otis, D.L., Burnham, K.P., White, G.C., Anderson, D.R.: Statistical inference from capture data on closed animal populations. Wildlife Monogr. 62, 3–135 (1978)

Paetkau, D., Slade, R., Burden, M., Estoup, A.: 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 (2004)

Paetkau, D., Strobeck, C.: Microsatellite analysis of genetic variation in black bear populations. Mol. Ecol. 3, 489–495 (1994)

Pella, J., Masuda, M.: The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. Can. J. Fish. Aquat. Sci. 63, 576–596 (2006)

Piry, S., Alapetite, A., Cornuet, J.-M., Paetkau, D., Baudouin, L., Estoup, A.: GeneClass2: A software for genetic assignment and first-generation migrant detection. J. Hered. 95, 536–539 (2004)

Pompanon, F., Bonin, A., Bellemain, E., Taberlet, P.: Genotyping errors: causes, consequences and solutions. Nat. Rev. Genet. 6, 847–859 (2005)

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

Rannala, B., Mountain, J.L.: Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. U.S.A. 94, 9197–9201 (1997)

Robins, J.H., Miller, S.D., Russell, J.C., Harper, G.A., Fewster, R.M. Where did the rats of Big South Cape Island come from? NZ J. Ecol. 40, 229–234 (2016)

Rousset, F.: Genepop'007: A Complete Reimplementation of the Genepop Software for Windows and Linux. Mol. Ecol. Resour. 8, 103–106 (2008)

Russell, J.C., Miller S.D., Harper G.A., MacInnes H.E., Wylie M.J., Fewster R.M. Survivors or reinvaders? Using genetic assignment to identify invasive pests following eradication. Biol. Invasions 12, 1747–1757 (2010)

Russell, J.C., Fewster, R.M.: Evaluation of the linkage disequilibrium method for estimating effective population size. In: Thomson, D.L., Cooch, E.G., Conroy, M.J. (eds.) Modeling Demographic Processes in Marked Populations, pp. 291–320. Environmental and Ecological Statistics Series, Vol 3, Springer, Berlin (2009)

Russell, J.C., Towns, D.R., Anderson, S.H., Clout, M.N.: Intercepting the first rat ashore. Nature 437, 1107 (2005)

Schofield, M.R., Bonner, S.J. Connecting the latent multinomial. Biometrics 71, 1070–1080 (2015)

Taberlet, P., Luikart, G.: Non-invasive genetic sampling and individual identification. Biol. J. Linn. Soc. 68, 41–55 (1999)

Vale, R.T.R., Fewster, R.M., Carroll, E.L., Patenaude, N.J.: Maximum likelihood estimation for model $M_{t,\alpha}$ for capture-recapture data with misidentification. Biometrics 70, 962–971 (2014)

Veale, A.J., Edge, K-A., McMurtrie, P., Fewster, R.M., Clout, M.N., Gleeson, D.M.: Using genetic techniques to quantify reinvasion, survival and in-situ breeding rates during control operations. Mol. Ecol. 22, 5071–5083 (2013)

Weir, B.S.: Genetic Data Analysis II. Sinauer, Sunderland (1996)

Wikipedia: List of countries by sex ratio. Wikipedia (2015) en.wikipedia.org/wiki/List_of_countries_by_sex_ratio

Wright, J.A., Barker, R.J., Schofield, M.R., Frantz, A.C., Byrom, A.E., Gleeson, D.M.: Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. Biometrics 65, 833–840 (2009)

Yoshizaki, J., Brownie, C., Pollock, K.H., Link, W.A.: Modeling misidentification errors that result from use of genetic tags in capture-recapture studies. Environ. Ecol. Stat. 18, 27–55 (2011)