Visualizations for genetic assignment analyses using the saddlepoint approximation method

L. F. McMillan* and R. M. Fewster

Department of Statistics, The University of Auckland, Private Bag 92019, Auckland, New Zealand

**email:* lmcm177@aucklanduni.ac.nz

Summary: We propose a method for visualizing genetic assignment data by characterizing the distribution of genetic profiles for each candidate source population. This method enhances the assignment method of Rannala and Mountain (1997) by calculating appropriate graph positions for individuals for which some genetic data are missing. An individual with missing data is positioned in the distributions of genetic profiles for a population according to its estimated quantile based on its available data. The quantiles of the genetic profile distribution for each population are calculated by approximating the cumulative distribution function (CDF) using the saddlepoint method, and then inverting the CDF to get the quantile function. The saddlepoint method also provides a way to visualize assignment results calculated using the leave-one-out procedure. This new method offers an advance upon assignment software such as geneclass2, which provides no visualization method, and is biologically more interpretable than the bar charts provided by the software structure. We show results from simulated data and apply the methods to microsatellite genotype data from ship rats (*Rattus rattus*) captured on the Great Barrier Island archipelago, New Zealand. The visualization method makes it straightforward to detect features of population structure and to judge the discriminative power of the genetic data for assigning individuals to source populations.

Key words: Genetic assignment; Leave-one-out; Lugannani-Rice formula; Multilocus genotype data; Population structure; Saddlepoint approximation.

1. Introduction

Genetic assignment methods compare genetic data from individual animals to genetic profiles of reference samples from multiple candidate source populations, to determine the appropriate source population, if any, for a given animal. The methods can also be used to analyse the reference samples themselves, to assess the amount of genetic overlap or separation between the populations.

Since their inception in the 1990s, genetic assignment techniques have served many purposes in ecology, biology and conservation. They have often been used to detect underlying population structure (e.g. Bergl and Vigilant 2007; Underwood et al., 2007). Berry et al. (2004) used assignment methods to estimate patterns of dispersal, while Taylor et al. (2006) used assignment to detect hybridisation. Genetic assignment has also been used for a range of forensic analyses, from detecting fishing competition fraud (Primmer et al., 2000) to detecting illegal translocations (Frantz et al., 2006) and illegal poaching (Manel et al., 2002). The same methods can help conserve endangered species by monitoring dispersal, or to assist with the management and eradication of invasive pests (Rollins et al., 2009).

Genetic assignment studies typically involve examination of several polymorphic genetic loci, usually microsatellites (Rannala and Mountain, 1997; Fewster, 2017). The basis of the analysis is that separate populations have different allele frequencies at these loci. This is partly due to genetic drift, whereby the randomness of mating changes the allele frequencies in each generation; this has greater effect in small populations, where rare alleles may disappear after only a few generations. Additionally, populations may be subject to founder effects, whereby individuals with some alleles that are rare in the wider community form a new population in which those alleles become common. Founder effects are especially relevant in studies of invasive species.

Populations are therefore characterized in genetic terms by their allele frequencies, which can be estimated by sampling from the populations. We can then compare the alleles of an individual animal to the allele frequencies in each of the populations, and combine those results over multiple loci. From this we obtain measures of fit for the animal with respect to each of the populations.

Ideally, one would then visualize the overall genetic structure, for example using scatterplots of the measures of fit with axes representing populations, where each point represents an individual animal (Paetkau et al., 2004). However, there is a fundamental difficulty with plotting these measures of fit, because some sampled individuals will be missing data at one or more loci. In the dataset we describe below, 2.6% of the locus records are missing. The measures for different animals will therefore be on different scales, preventing them from being plotted together. In this paper we propose methodology to address this difficulty.

In the absence of established methods for plotting measures of fit, other representations of assignment results have been used. The popular software program STRUCTURE (Pritchard et al., 2000) performs assignment, and, under the admixture option, displays genetic structure as bar charts. These charts have one bar per individual, with portions of the bar colored differently to show the proportions of that individual's genome estimated to originate from the different populations. That is, the current set of individuals is assumed to be the result of past mixing between the populations, and each individual is assumed to have components of its genome that were inherited from those populations. However, this method of visualization does not distinguish between an individual with a poor fit to all populations and an individual with a good fit to all populations, because it expresses the individual's profile as a relative composition and does not display the individual's absolute measures of fit. Additionally, although STRUCTURE is a useful tool for exposing cryptic genetic structure, the admixture model may be hard to interpret in biological terms.

Another commonly-used genetic assignment program, GENECLASS2 (Piry et al., 2004) provides measures of fit for individuals with respect to each population, but does not provide any visualization of the results; nor is there a common standard for reporting the results numerically. Kotze et al. (2007) report the results of both the Bayesian and the frequentist analysis options on a few individuals with unknown origins. Rollins et al. (2009) show detailed assignment results for the few individuals that were identified as dispersers (i.e. estimated to have migrated from their original population to a new population) but not for other individuals, so calibration is lacking. Results from GENECLASS2 are typically shown as tables indicating how many individuals were assigned to each population, but the absolute measures of fit used for the assignment are not shown (e.g. Glover et al. 2009).

Other software packages do not perform genetic assignment, but calculate genetic distances and population diversity measures, and some of them provide visualizations of the genetic data for the purpose of interpreting population structure. Visualization methods used include: factor analysis applied to raw multilocus data, implemented in Genetix 4.05 (Belkhir et al., 1996-2004: e.g. Taylor et al., 2006); and principal coordinate analysis (PCoA) of individual pairwise distance data, implemented in GenAlEx 6.05 (Peakall and Smouse, 2012). The `adegenet` package for R (Jombart and Ahmed 2011) performs principal component analysis combined with discriminant analysis on the raw multilocus data. However, none of these packages calculates the fit of an individual into candidate source populations.

Here we propose a method that enables us to visualize genetic population structure on a scatterplot, in which each axis depicts the individual's fit to the corresponding reference population. We accommodate individuals with missing locus data by plotting them at the quantiles of fit they attain in the reduced-locus data for the reference populations, corresponding to the loci they possess. This involves characterizing the posterior distribution of fit for each reference population, and all reduced-locus combinations, which we do by deriving close approximations via the saddlepoint method.

We show that our visual assignment method creates compelling displays of underlying population structure, such as the level of separation or overlap between populations, the genetic spread of individuals in each population, and whether one population is a genetic subset of another. Visualization also shows clearly individual fit to each population, including individuals from the reference samples that do not fit their own source populations, and the best placements for query samples. Visualization is an important initial step in the assignment process, indicating whether or not it is appropriate to assign individuals. If the populations exhibit substantial overlap, this indicates that many individuals have a plausible fit to multiple populations, and thus cannot be conclusively assigned to a single population.

An advantage of the saddlepoint method for characterizing distributions of genetic fit to each population is that the quantiles of the distributions can be calculated accurately and plotted on the same display as the assignment results. This displays the within-population variance, and shows whether one population is a genetic subset of another. These quantiles can also be used as thresholds for assignment by exclusion. A further advantage of the saddlepoint approximation method is that it enables assignment results produced under the leave-one-out procedure to be plotted in the same way as standard results. We will demonstrate the importance of using a leave-one-out approach when sample sizes are small.

2. Genetic Assignment

We now outline the established approach to genetic assignment analysis. The aim is to compare an animal's DNA profile to two or more candidate source populations, which we shall hereafter refer to as *reference populations*. Reference populations are typically defined by their location. Samples from each reference population are required to provide an estimate of the allele frequencies for that population; we shall use the term *reference samples* to denote these samples. The provenance of these samples is not thought to be in question, although this assumption is not critical because any anomalous reference samples will be exposed by our graphical method. Often, there is also a further set of samples whose provenance is unknown. We use the term *query samples* for these additional animals. These samples do not contribute to the estimates of allele frequencies for any reference populations.

The first step in the assignment process is to estimate allele frequencies in each reference population. We then build a genetic profile for each reference population by examining the fit of its own reference animals into their own population: some populations are tight-knit while others are diverse and diffuse. We build a picture of genetic structure among populations by similarly examining the fit of reference samples from other reference populations into each target reference population. These steps require us to define a measure of genetic fit of an individual into a population, which will be the log posterior genotype probability defined below. Finally, if there are any query samples (animals to be assigned), we use the same measure of genetic fit to examine their fit into each of the reference populations.

2.1 *Estimating allele frequencies*

Consider a single reference population R, and a single genetic locus L with k allele types labeled $1, 2, \ldots, k$. Let $\boldsymbol{p} = (p_1, \ldots, p_k)$ be the frequencies of alleles $1, \ldots, k$ in this population, where $0 \leqslant p_i \leqslant 1$ for each i and $\sum_{i=1}^{k} p_i = 1$. We aim to estimate $p_1, \ldots, p_k$. In this development we restrict to a single population R, but the number of available allele types k is determined by pooling observations from all individuals in the reference and query samples.

Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ be the observed allele counts at locus L for diploid reference animals $1, 2, \ldots, n_R$ from population R, where $\sum_{i=1}^{k} X_i = 2n_R$. The likelihood for obtaining the observed allele counts at locus L, given population allele frequencies $\boldsymbol{p}$, is gained from

$$\boldsymbol{X} \mid \boldsymbol{p} \sim \text{Multinomial}(2n_R\,; \boldsymbol{p}). \tag{1}$$

The multinomial model involves an assumption that alleles are independent within genotypes, which holds if the population is in Hardy-Weinberg equilibrium.

The most common technique for estimating allele frequencies is the Bayesian method proposed by Rannala and Mountain (1997) because it enables us to incorporate our uncertainty in estimating $\boldsymbol{p}$ when subsequently measuring the genetic fit of animals into the population. In particular, if $X_i = 0$ for some i, we do not wish to use the maximum-likelihood point estimate $\hat{p}_i = 0$ because this would automatically exclude population R as a source for any animals possessing alleles that were unsampled in population R, regardless of how small the sample is. By using a Bayesian posterior for p_i, we acknowledge the possibility that allele type i is present but unsampled in population R, and the range of posterior values supported for p_i narrows according to the sample size n_R.

The conjugate prior for $\boldsymbol{p}$ is the symmetric form of the Dirichlet distribution, which is a multivariate generalization of the beta distribution, with density

$$f(\boldsymbol{p};\tau) = \frac{\Gamma(\tau k)}{\Gamma(\tau)^k} \prod_{i=1}^{k} p_i^{\tau-1}, \text{ for } \boldsymbol{p} \in [0,1]^k; \sum_{i=1}^{k} p_i = 1. \tag{2}$$

There are two commonly-used options for the choice of prior parameter τ: Rannala and Mountain (1997) proposed $\tau = 1/k$, while Baudouin and Lebrun (2001) proposed $\tau = 1$.

The resulting posterior distribution for $\boldsymbol{p}$ conditional on the sampled alleles is a Dirichlet distribution, given as

$$[\boldsymbol{p} \,|\, \boldsymbol{X} = \boldsymbol{x}] \sim \text{Dirichlet}(x_1 + \tau,\, x_2 + \tau,\, \ldots,\, x_k + \tau), \tag{3}$$

where we take the prior parameter τ to be either 1 or $1/k$, and x_j is the number of alleles of type j observed among all reference samples from population R.

2.2 *Log-genotype probability for a particular individual*

We continue to consider reference population R and locus L. We aim to create a measure of genetic fit of an individual I into population R.

The genotype for individual I at locus L is given by $\boldsymbol{a} = (a_1, a_2, \ldots, a_k)$ where each a_i is 0, 1 or 2, indicating how many alleles of type i are included in its genotype, and $\sum_{i=1}^{k} a_i = 2$. A homozygous genotype will have $a_r = 2$ for some value of r, and all other a_i will be 0, whereas a heterozygous genotype, with two different allele types, will have $a_r = a_s = 1$ for some $r \neq s$, and all other a_i will be 0.

We measure the fit of individual I into population R at locus L by the *posterior genotype probability* of I at locus L:

$$\mathbb{P}(\boldsymbol{a} \,|\, \boldsymbol{X} = \boldsymbol{x}) = \int_{\boldsymbol{p}} \mathbb{P}(\boldsymbol{a} \,|\, \boldsymbol{p} \,;\, \boldsymbol{x}) \pi(\boldsymbol{p} \,|\, \boldsymbol{x}) d\boldsymbol{p}\,, \tag{4}$$

where $\pi(\boldsymbol{p} \,|\, \boldsymbol{x})$ is the posterior density of allele frequencies for population R given by (3).

Now $(\boldsymbol{a} \,|\, \boldsymbol{p}) \sim \text{Multinomial}(2, \boldsymbol{p})$, so the marginal posterior distribution of the single-locus genotype $\boldsymbol{a}$, given the reference allele frequencies $\boldsymbol{x}$, is a Dirichlet-compound-multinomial (DCM) distribution (see Fewster, 2017). The probability in (4) simplifies to:

$$\mathbb{P}(\boldsymbol{a}) = \begin{cases} \dfrac{(x_r + \tau)(x_r + \tau + 1)}{(2n_R + k\tau)(2n_R + k\tau + 1)} & a_r = 2;\ a_j = 0 \text{ for } j \neq r; \\ \dfrac{2(x_r + \tau)(x_s + \tau)}{(2n_R + k\tau)(2n_R + k\tau + 1)} & a_r = a_s = 1;\ a_j = 0 \text{ for } j \neq r, s. \end{cases} \tag{5}$$

Here, we use the notation $\mathbb{P}(\boldsymbol{a})$ as shorthand for $\mathbb{P}(\boldsymbol{a} \,|\, \boldsymbol{X} = \boldsymbol{x})$, where $\boldsymbol{X}$ denotes the reference alleles for population R sampled at locus L. Different loci will have different values of k, $\boldsymbol{X}$, and possibly n_R, due to individuals with missing data at some loci.

Under standard assignment methods, the same calculation is applied whether or not I is one of the reference samples for population R. However, under the leave-one-out method, if I is a reference sample for R, the alleles from I are excluded from the allele counts $(x_1, \ldots, x_k)$ for population R when calculating $\mathbb{P}(\boldsymbol{a})$ for individual I.

The overall posterior probability that the multilocus genotype $(\boldsymbol{a}^1, \boldsymbol{a}^2, \ldots, \boldsymbol{a}^{N_L})$ of individual I could arise from reference population R is calculated as the product of (5) over all loci $L = 1, 2, \ldots, N_L$. Taking logs gives the log genotype probability (LGP) for individual I in population R:

$$\text{LGP}_I^R = \sum_{L=1}^{N_L} \log\{\, \mathbb{P}(\boldsymbol{a}^L) \,\}, \tag{6}$$

where $\log\{\, \mathbb{P}(\boldsymbol{a}^L)\}$ is given by (5) based on sampled alleles $\boldsymbol{X}^L$ from population R for each locus $L = 1, 2, \ldots, N_L$. Adding the terms in (6) assumes independence among the loci.

As with the assumption of independence of alleles within genotypes, any non-independence (termed *linkage disequilibrium*) is unlikely to cause problems unless it is extreme (Fewster, 2017). Assessment of the impact of linkage disequilibrium is provided in Web Appendix G.

The overall LGP given by (6), as defined in Piry et al. (2004), constitutes the measure of fit of animal I into population R: it is the estimated log-probability of the multilocus genotype possessed by individual I arising in population R, in the sample space of all possible multilocus genotypes from population R. This is not the same as the probability that individual I originated in the specific population R out of all populations. A high LGP indicates that the genotype of individual I has high probability of arising in population R; that is, that individual I has a good fit to population R.

We define the *LGP distribution* for population R as being the posterior distribution of LGPs for all possible multilocus genotypes arising from the posterior allele frequencies for population R. For example, if population R comprises loci with only a few common allele types, then any genotype drawn from the posterior allele frequencies for population R will tend to have a high LGP, and the LGP distribution will be tightly focused around a high mean. By contrast, if population R is genetically diverse, with many common allele types at some or all loci, then the LGP distribution will be centered lower, with greater variance.

3. Visualization of assignment results

We now outline our novel approach to assignment visualization. From (6) we see that a set of individuals with complete data at all loci have LGPs that are comparable to each other; the LGPs can thus be plotted without further adjustment. Figure 1, which we henceforth call a GenePlot, shows individuals with complete data, sampled from two reference populations. Each individual I is plotted at coordinates $(\mathrm{LGP}_I^{R_1}, \mathrm{LGP}_I^{R_2})$ for populations R_1 and R_2. Individuals are colored according to which population they were sampled from. We use base-10 logarithms to convey orders of magnitude. We use the Dirichlet prior in (2) with $\tau = 1/k$.

[Figure 1 about here.]

The data in Figure 1 are from a study of invasive ship rats (*Rattus rattus*) in the Great Barrier Island archipelago, New Zealand. Ship rats were captured between 2005 and 2008 on the main island (Aotea, 28500 ha) and the Broken Islands, which comprise four smaller islands with total area 125 ha located about 300m from Aotea at closest approach. A map of the sampling locations can be found in Web Appendix A.

The solid diagonal line in Figure 1 is the line of equal posterior probability for both reference populations; a point lying on that line has the same LGP with respect to both populations. The thin diagonal lines indicate where the genotype probability (the inverse-log of the LGP) for one population is 9 times greater than it is for the other population. The narrow range of these lines highlights the enormous variability in the LGP distributions, which represent genotype probabilities that span many orders of magnitude. Figure 1 also shows the 100% quantiles, and one of the 0% quantiles, of the posterior LGP distributions for the populations: that is, the maximum and minimum possible LGPs for each population. The other 0% quantile is so low it has been excluded from the plot. The maximum and minimum are readily calculated from (5) for known $x_1, x_2, \ldots, x_k$ (Web Appendix B).

Figure 1 can be used to assess population structure: it shows clearly that the Broken Islands population is a genetic subset of the Aotea population. Most of the Broken Islands rats have high LGP with respect to the Aotea population (vertical coordinate), and thus could plausibly have originated in the Aotea population, whereas very few of the Aotea rats have high LGP with respect to the Broken Islands population (horizontal coordinate). This property will become even more obvious with the addition of extra quantile lines requiring the more detailed computations we describe in Section 4. The Aotea rats also have more diverse LGP values with respect to both populations than the Broken Island rats, reflecting a genetic profile of many more allele types in Aotea, many of which occur at low frequency. Figure 1 can also be used for assignment using the common method of choosing the population for which the individual has the highest fit. By inspecting the position of each individual relative to the diagonal lines we can assess which individuals have an LGP that is substantially higher for one population than the other. A single Aotea-sampled rat on Figure 1 does have high LGP with respect to the Broken Islands and is grouped with the Broken Islands population; it might therefore have dispersed from the Broken Islands to Aotea. Significantly, this rat was one of only seven rats sampled on the part of Aotea directly opposite the Broken Islands: see Web Figure 1 for a map of sampling locations. The display in Figure 1 is consistent with the suggestion that the Broken Islands population was founded from Aotea but has lost many alleles relative to the much larger adjacent Aotea population.

Figure 1 also shows the shape and the range of the LGP distribution for each population, and whether populations overlap or are genetically distinct.

By contrast with the graphical display in Figure 1, numeric or tabular displays of assignment results can be difficult to interpret. One approach to deal with the enormous span of LGP values and the problem of missing data is to normalize the results. For example, in GENECLASS2 (Piry et al., 2004) the percentage *score* for an individual I in population R_1 is the genotype probability (GP) for R_1 as a percentage of the total GP over all populations: e.g. $100\,\mathrm{GP}_I^{R_1}/(\mathrm{GP}_I^{R_1} + \mathrm{GP}_I^{R_2})$ for two populations. An example of the resulting tabular display is shown in Web Table 1. However, this leads to a loss of information: individuals with a good fit to all populations and individuals with a poor fit to all populations are indistinguishable from each other. Additionally, the tabular display does not provide the context of overall distribution shape and variability shown in Figure 1. For example, any individual on the lower thin diagonal line in Figure 1 satisfies $\mathrm{GP}_I^{R_1} = 9\mathrm{GP}_I^{R_2}$ so is given a seemingly-conclusive score of 90% in favour of population R_1, but it is clear that there are substantial portions of the line for which this conclusion would be inappropriate, especially in the lower left where individuals have poor fits to both populations, making it inappropriate to assign such individuals by choosing the population for which they have the highest LGP.

We conclude that visualization of LGPs is preferable to a numeric display; it conveys both the fit for specific individuals and the overall distribution for a population. However, we need a way to display LGPs for missing-data individuals on the same scale as LGPs for complete-data individuals. LGPs of missing-data individuals will appear artificially low compared with those of complete-data individuals, since they are summed over fewer loci in (6), so they cannot be displayed unadjusted on the same plot. We now derive a method to display missing-data individuals using the quantiles of the LGP distribution. We can also use these quantiles to assess the absolute fit of an individual within a population, rather than the typical practice of only comparing the individual's fit among different populations.

4. Quantile method for plotting individuals with missing data

Let $\mathcal{L} = \{1, 2, \ldots, N_L\}$ be the full set of loci, and let $\mathcal{L}_I \subseteq \mathcal{L}$ be the loci available for individual I. We define $\mathrm{LGP}_I^R = \mathrm{LGP}_I^R(\mathcal{L})$ to be the desired LGP of individual I based on loci $\mathcal{L}$ in population R. For an individual with missing data, i.e. $\mathcal{L}_I \subset \mathcal{L}$, this LGP is unknown so we need to estimate it. We define $\widetilde{\mathrm{LGP}}_I^R$ to be our estimate of LGP_I^R.

We can calculate $\mathrm{LGP}_I^R(\mathcal{L}_I)$ based on the reduced loci $\mathcal{L}_I$, using the analog of (6) for $\mathcal{L}_I$. Our aim is to preserve the 'unusualness' of individual I on the visual chart, so we aim to plot individual I on the full-locus chart at the LGP quantiles that it attains in the reduced-locus distribution based on loci $\mathcal{L}_I$. Therefore we need to characterize the posterior LGP distribution of population R for any set $\mathcal{L}^* \subseteq \mathcal{L}$ so that we can calculate the cumulative distribution function (CDF) $F_{\mathcal{L}^*}$ of this distribution. We also need to calculate the quantile function $Q_{\mathcal{L}} = F_{\mathcal{L}}^{-1}$ of the full-locus distribution. Then our estimated coordinate for I in R based on $\mathcal{L}_I$ is given by:

$$\widetilde{\mathrm{LGP}}_I^R = Q_{\mathcal{L}}\left[F_{\mathcal{L}_I}\{\,\mathrm{LGP}(\mathcal{L}_I)\,\}\right]. \tag{7}$$

For example, if I lies at the 10% quantile of the reduced-locus distribution based on loci $\mathcal{L}_I$, we wish to plot it at the 10% quantile of the full-locus distribution based on $\mathcal{L}$.

4.1 *Characterizing the posterior LGP distribution*

The multilocus posterior LGP in (6) is gained from the sum of N_L discrete random variables corresponding to the N_L single-locus posterior LGPs. Each of these is the logarithm of a Dirichlet compound multinomial (DCM) distribution, as in (5), that can be enumerated in entirety. However, the multilocus LGP is the sum of these N_L random variables, where N_L is typically 10 or more, so it is not feasible to enumerate the probability function as a convolution to gain $F_\mathcal{L}$ and $Q_\mathcal{L}$ directly. The following sections focus on how we may calculate approximations $\hat{F}_\mathcal{L}$ and $\hat{Q}_\mathcal{L}$ to these functions.

4.2 *Simulation approximation method*

It is straightforward to simulate from the multilocus posterior distribution of LGPs for population R by drawing genotypes directly from the constituent single-locus DCM distributions. From this we can obtain the empirical CDF and quantile functions.

The disadvantage of this method is that it becomes particularly inaccurate in the lower tail of the distribution. One feature of the LGP distribution is that many populations have a large number of alleles that occur with very low frequency (Rannala and Mountain, 1997; Fewster, 2017). These low-frequency alleles lead to multilocus LGPs spanning many orders of magnitude in the lower tail. This leads to high variability when estimating the lower quantiles of the distribution. The 0% quantile for the Aotea population shown in Figure 1 is at about -40, indicating that the minimum GP for the population is about 10^{-40} based on all sampled alleles. We will show later that the 1% quantile for the same population is at about -15; thus the lower 1% tail spans a range of about 10^{25}.

Even a very large sample generated from the LGP distribution will often fail to capture the full extent of the distribution, so the simulation method can produce inaccurate estimates of

LGP for missing-data individuals, potentially overestimating their fit to each population. In some datasets, we have found individuals whose calculated LGP is lower than the minimum simulated LGP, despite simulating hundreds of thousands of genotypes. As an alternative to approximation by simulation, we therefore seek an analytic approximation to the CDF $F_{\mathcal{L}^*}$ and the quantile function $Q_{\mathcal{L}}$ of the posterior LGP distributions.

4.3 *Saddlepoint approximation method*

The saddlepoint method was proposed by Daniels (1954), and its initial purpose was to approximate the probability density function (PDF) for the sum of identical distributions. Lugannani and Rice (1980) proposed a related formula for approximating the cumulative distribution function (CDF) of a sum of distributions; their original purpose was to determine tail probabilities, but the formula is accurate over the full range of the distribution.

The saddlepoint approximation formulas apply to any distributions, not just those composed of sums, and have many applications in statistics, including approximation of the bootstrap distribution for a parameter estimate (Davison and Hinkley, 1988) and estimation of marginal tail probabilities for inference about scalar parameters (DiCiccio et al., 1990). Here, we use the Lugannani-Rice saddlepoint CDF approximation, following the exposition of Butler (2007), to derive close approximations to $F_{\mathcal{L}^*}$, the CDF of the posterior LGP distribution based on loci $\mathcal{L}^*$, for any subset of loci $\mathcal{L}^* \subseteq \mathcal{L}$.

The saddlepoint approximation to the CDF of a random variable Y with known mean $\mu = \mathbb{E}Y$ is defined in terms of the cumulant generating function (CGF) of Y, $K(t) = \log\{M(t)\} = \log\left[\,\mathbb{E}\{\exp(tY)\}\,\right]$, where M is the moment generating function. Derivatives of K can be computed in terms of derivatives of M; for example $K''(t) = M''(t)/M(t) - \{M'(t)/M(t)\}^2$. The saddlepoint approximation to the CDF of Y is then:

$$\hat{F}(y) = \begin{cases} \Phi(\hat{w}) + \phi(\hat{w})\left(\dfrac{1}{\hat{w}} - \dfrac{1}{\hat{v}}\right) & \text{if } y \neq \mu, \\[2ex] \dfrac{1}{2} + \dfrac{K'''(0)}{6\sqrt{2\pi}\,K''(0)^{3/2}} & \text{if } y = \mu, \end{cases} \tag{8}$$

where Φ and ϕ are the Gaussian CDF and PDF, and where $\hat{v} = \hat{s}\sqrt{K''(\hat{s})}$,

$$\hat{w} = \text{sign}(\hat{s})\sqrt{2\{\hat{s}y - K(\hat{s})\}}\,, \tag{9}$$

and $\hat{s}$ is the solution to the equation

$$K'(\hat{s}) = y. \tag{10}$$

The saddlepoint approximation is a local approximation to an expansion of the inversion formula that links the PDF of a distribution to its CGF. The point $\hat{s}$ is a saddlepoint because when the integral approximation is converted into the complex plane, $\hat{s}$ remains constant in the imaginary direction, while also acting as a root in the real direction.

Equations (8) to (10) specify the saddlepoint approximation for the CDF of a continuous distribution. The posterior LGP technically follows a discrete distribution; however, even for small to moderate numbers of alleles and loci the number of possible genotypes is extremely large, so the support of the distribution becomes sufficiently dense that we can apply the continuous form of the saddlepoint CDF approximation (8) without applying the correction factors required for discrete distributions.

No tuning is needed to apply the saddlepoint approximation. However, it is necessary to invert K' at each required value of y in (10), to obtain the corresponding value for $\hat{s}$; this incurs a small computational cost. Additionally, it is simple to solve (10) away from $y = \mu$, but near $y = \mu$ the function K' becomes extremely steep and the numerical solution $\hat{s}$ can become inaccurate. When compounded in the calculation of $\hat{w}$, this inaccuracy can lead to numerical instability in the saddlepoint CDF approximation, producing values of $\hat{F}$ outside of the valid range $[0, 1]$. We found that numerical inaccuracy around the mean μ could be reduced by replacing y with $K'(\hat{s})$ in (9), giving

$$\hat{w} = \text{sign}(\hat{s})\sqrt{2\{\hat{s}K'(\hat{s}) - K(\hat{s})\}}. \tag{11}$$

The inaccuracy caused by using (11) instead of (9) is negligible, particularly when compared with the numerical instability that can arise from (9).

Using this method gives $\hat{F}_{\mathcal{L}^*}$ for any $\mathcal{L}^* \subseteq \mathcal{L}$, as required in (7): see Web Appendix C for details. The quantile function approximation $\hat{Q}_{\mathcal{L}}$ is found by inverting $\hat{F}_{\mathcal{L}}$ using standard numerical root-finding methods.

4.4 *Evaluation of the saddlepoint approximation*

Figure 2 shows the application of the saddlepoint approximation to posterior LGP distributions for the ship rat data described above, with the addition of ship rats captured on Kaikoura Island during the same period. Kaikoura (530 ha) lies about 80m from Aotea at closest approach and 3 km north of the Broken Islands (Web Figure 1). The saddlepoint approximation to the CDF is compared with the empirical cumulative distribution function (ECDF) derived by generating samples from the posterior LGP distribution as in Section 4.2.

The saddlepoint method, without any tuning, adapts to the different shapes of the distributions for the three populations in Figure 2 and provides an accurate approximation over the full range of each distribution. The closeness of fit is more visible on the PDF plot than the CDF plot, since the cumulative nature of the CDF makes small inaccuracies less visible. Rather than using the direct saddlepoint PDF approximation we show the derivative of the CDF saddlepoint approximation, because only the CDF is used in our genetic charts.

[Figure 2 about here.]

The top row in Figure 3 shows the same saddlepoint approximation for the Broken Islands LGP distribution, zoomed in to focus on the mean of the distribution. Also shown are gamma and lognormal approximations to the distribution, based on matching the mean and variance of the posterior LGP distribution to the gamma and lognormal mean and variance respectively. The saddlepoint method produces a noticeably better approximation than the gamma and lognormal distributions. Although it would be possible to find gamma and lognormal distributions a different way by fitting the simulated distribution, determining the appropriate parameters for those distributions would rely on the simulated samples so the method is not adequate as an analytic replacement for approximation by simulation. The results for the Broken Islands in Figure 3 are typical of those we have seen in multiple other data sets including ship rats and Atlantic salmon (*Salmo salar*).

[Figure 3 about here.]

We also tested the saddlepoint method extensively using simulated allele frequencies. We tested different numbers of loci and different numbers of allele types per locus, running ten replicates for each combination of parameters. For a given number of loci N_L and a maximum number of allele types per locus, k, we then reduced the number of alleles at some loci, keeping $T_L \sim \text{Binomial}(k, 1 - \frac{1}{k})$ allele types at locus L, truncated at a minimum of $T_L = 2$. We randomly generated allele frequencies by generating individual frequencies from a Beta(0.5, 0.7) distribution and normalizing the frequencies at each locus.

As can be seen in the bottom row of Figure 3, as the number of loci drops and the average number of alleles per locus also drops, the distribution of log-genotype probabilities becomes more ragged and more difficult to approximate. However, the example shown in Figure 3 illustrates a worst-case scenario, since five loci is generally considered to be too few for accurate population analysis. We note that even in this case, the saddlepoint method provides a reasonably good approximation. As the number of loci increases, the distribution rapidly becomes smoother, and the saddlepoint approximation becomes ever more accurate.

Web Appendix D describes the results of a numerical comparison between the saddlepoint approximation (SCDF) and the empirical CDF based on simulations (ECDF), which shows that the SCDF fits extremely well to the ECDF within the central range of the ECDF.

5. Applications

We now demonstrate how our approach of Sections 3 and 4 can be used both for visualizing population structure and performing assignment. Figures 4-6 show the same ship rat data described above. The purpose of the analysis is to understand the dispersal patterns of the invasive rats between islands in the archipelago, so as to inform eradication planning on the smaller islands, and determine the origins of rats detected after eradication attempts.

Rats with missing data are shown in Figure 4 with asterisks. These rats constitute a large minority of the samples, so it is important to include them on the charts, demanding use of the quantile-approximation methods of Section 4. Calculations using the alternative simulation-based quantile approximations in Section 4.2 inflated the LGP estimates for many of these missing-data rats, distorting the overall picture of population structure.

5.1 *Visualization for two reference populations*

A major advantage of the saddlepoint method is that it enables the accurate calculation of all quantile lines for each population, using the function $\hat{Q}_{\mathcal{L}}$ described in Section 4. Figure 4 shows the 1% quantile lines for the Broken Islands, Kaikoura and Aotea populations. A comparison with Figure 1 indicates that there are at least 30 orders of magnitude between the 0% and 1% quantiles for the Broken Islands, and about 25 orders of magnitude between the same quantiles for Aotea. Given the huge range covered by the lowest 1% of these two LGP distributions, indicating a high level of skewness, it is clear that these quantiles provide a useful tool in describing the shape of the genetic profile for each population.

The 1% quantile lines are powerful indicators of inter-population structure. The left plot in Figure 4 shows that most Broken Islands rats lie between the 1% and 100% lines for both the Broken Islands and Aotea, and thus have a reasonable fit to both populations. By contrast, all but one of the Aotea rats lie below the 1% line for the Broken Islands population; it is very unlikely that a random Aotea rat would by chance possess only alleles found in the Broken Islands population because the Aotea population contains many more allele types. This is a clear example of subsetting, where individuals from one population show good fits to both reference populations but individuals from the other population only have reasonable fit into their own sampling population. We can see from Figure 4 that the

Broken Islands and Aotea populations are easily distinguished genetically, and the Broken Islands population is a subset of the Aotea population. The likely explanation is that the Broken Islands population was founded from the Aotea population, so the alleles of Broken Island rats are also common in Aotea; but due to founder effects, isolation and genetic drift the Broken Islands population has lost many of the alleles that are found in Aotea.

The left plot also shows rats sampled from the Broken Islands in 2010, the year after an eradication attempt. We can assign these rats using the exclusion principle, whereby an individual is only assigned to a population if its fit to every other population is below a pre-defined threshold. The quantile lines calculated using the saddlepoint method provide such a threshold, highlighting another advantage of our methodology. Using the 1% quantile lines as thresholds, it is clear that all of the post-eradication rats can be excluded from the Broken Islands population. The graphical display gives compelling evidence that these rats were not survivors of the eradication attempt, but were most likely swimmers or hitch-hikers from Aotea. Similar events have occurred every year since 2010: a small number of rats have arrived annually from Aotea, but have been eliminated before establishing a population.

The right plot in Figure 4 shows that samples from Kaikoura and Aotea up to 2008 were not well separated in genetic terms. Several individuals sampled on either population could plausibly have originated in the other population, and thus it would not be appropriate to attempt to assign individuals to one of these two populations. We infer that the Kaikoura and Aotea populations either separated only a short time earlier, or had ongoing dispersal. The dispersal hypothesis has since been corroborated by Bagasra et al. (2016), who deposited bait laced with Rhodamine B dye on Aotea. The dye was later detected in rats captured on Kaikoura, thus giving direct evidence of rat mobility between the islands. An eradication attempt on Kaikoura in 2008 failed to eliminate rats, and the population is now managed as a controlled, low-density population rather than as a rat-free sanctuary.

We conclude that the levels of pre-eradication genetic separation from Aotea, as seen in the Broken Islands and Kaikoura GenePlots respectively, were predictive of the eventual management outcomes in each case. The link between genetic separation and true separation is not entirely clear-cut, as not all dispersal leads to breeding for behavioral reasons. However, the pre-eradication genetic separation on the charts does seem to be indicative of the level of reinvasion experienced after eradication, so it is helpful in devising management strategies.

[Figure 4 about here.]

5.2 *Visualization for multiple reference populations*

Our method provides an intuitive visualization of population structure for two populations. It can also be used to assess multiple populations. Two options for visualization are shown in Figure 5. The top panel shows a GenePlot similar to the two-population case, but generated by performing principal component analysis (PCA) on the multi-dimensional LGP results, after applying the saddlepoint method to accommodate rats with missing data. The first two principal components are the axes of the GenePlot. The lower panel of Figure 5 shows an alternative multidimensional visualization represented as a set of bar charts. The multiple bar chart can be more effective when assessing the specific results for each population or the fit of a single individual with respect to the various populations, whereas the PCA plot is more useful for understanding overall genetic structure such as within-population variability, populations that overlap or are separate on the chart, and identifying anomalous individuals.

[Figure 5 about here.]

5.3 *Comparison with* STRUCTURE, GENECLASS2 *and* DAPC

The bar charts in Figure 5 differ from the bar charts in STRUCTURE (Pritchard et al., 2000) which display the proportions of each individual's genotype estimated to have originated from each population. By contrast, the bar charts in Figure 5 show the percentile of each individual's LGP fit in each population, based on the saddlepoint method, and therefore display an absolute rather than a relative measure of fit into each population. The advantage of this approach is that it distinguishes between individuals who have low genotype probabilities with respect to all candidate source populations, and individuals who have high probabilities with respect to all populations. Web Appendix E shows how results for the same data are reported by STRUCTURE, GENECLASS2, and the DAPC procedure from R package `adegenet`. The rat marked with a circle in Figures 4 and 5 appears, in STRUCTURE, to be well fitted to the Broken Islands (cluster 1), but Figure 4 shows that the rat has a rather poor fit to the Broken Islands. Under the exclusion assignment method with a threshold of 1%, this rat would be excluded from both the Broken Islands and Aotea populations.

5.4 *Leave-one-out*

Our saddlepoint method can be used to calculate LGPs on a leave-one-out basis. In this approach, each individual's alleles are excluded from the population it was sampled from before calculating the posterior distribution of allele frequencies and hence the LGP fit of that individual into its own population. The leave-one-out approach can be necessary when sample sizes are small because each individual's alleles exert a large influence on the estimated allele frequencies for the population. As a result, an individual's fit into its own population is inflated, and the reference populations appear more distinct than they should.

[Figure 6 about here.]

Figure 6 shows the difference between the standard method and the leave-one-out method for the full samples of roughly 60 individuals from each of the Kaikoura and Aotea populations, and for subsets of 10 individuals from each of those samples. We repeated the process for different random subsets, with similar results in each case. The leave-one-out method makes little difference when applied to the full samples, but for small subsets, the standard method shows far more separation between the populations than does the leave-one-out method.

After observing the leave-one-out plot, we conclude there is too much population overlap to accurately assign new individuals to either population, based on those samples.

The saddlepoint method for approximating quantiles allows the leave-one-out results to be plotted on a GenePlot. The axes of the plot correspond to the population LGP distributions based on the full samples from each population. The process for computing the plotted value of an individual from a reference population is similar to the process for plotting a missing-data individual. First we exclude the individual from the population it was sampled from, and re-estimate the allele frequencies for that population. This gives us the leave-one-out distribution. We then calculate the individual's LGP using that leave-one-out distribution, and use the saddlepoint method to find the corresponding CDF value within the leave-one-out distribution. Then we return to the full population distribution, with allele frequencies based on all the samples including this individual, and calculate the corresponding quantile value in that distribution. The individual is then plotted at that quantile value with respect to its own population. This is similar to the method used for missing-data individuals, except that here we are dealing with the leave-one-out distribution instead of the missing-data distribution to calculate the raw LGP for a given individual. We recommend that leave-one-out procedures should be used when sample sizes are below 20 individuals.

5.5 *Application to data from single nucleotide polymorphisms (SNPs)*

Although our examples use microsatellite data, the same development applies to data from SNPs, which are typically biallelic and may be genotyped at large numbers of loci. Examples of GenePlots for simulated biallelic data at 1000 loci are provided in Web Appendix F.

6. Conclusions

Genetic assignment data have many uses in ecology, conservation biology, and forensics, but they are complex to describe and interpret. Effective visualizations are pivotal to conveying the information contained in the data. We have developed a new visualization based on displaying the absolute genetic fit of each individual to each reference population. To achieve this we derived a quantile-based display algorithm to handle individuals with missing data. For quantile estimation we invoked a saddlepoint approximation to the posterior distribution of genetic fit within each reference population. This enabled us to estimate accurate quantiles throughout the support of the distribution, and in particular to capture the enormous span of the distribution's lower tail, which can not be accurately modeled by simulation or by standard distributions such as the gamma and lognormal.

We have shown by simulation that the saddlepoint approximation performs extremely well within the usual realm of genetic assignment analyses. Performance depends only upon the number of loci in the study and the number of different allele types per locus; besides these, the algorithms and conclusions are applicable to data from any diploid species. Saddlepoint approximations have been proved to attain a high level of accuracy. In our case, the saddlepoint approximation is applied in an ideal scenario involving a sum of fully-enumerable distributions, each with a known closed form for the cumulant generating function and its derivatives. Once the adjustment in equation (11) is made, the saddlepoint approximation is easy to code, computationally stable, and incurs a similar computational cost to the less accurate simulation method.

We have shown that the visualization of absolute genetic fit offers a powerful tool for seeing population structure at a glance, including population overlap and separation, genetic subsetting, and within-population variability. Existing visualizations and tabular reports focus on the relative fit of each individual to a selection of reference populations. These relative measures lose valuable information about the absolute measures of fit, which might reveal that an individual does not fit into any of the proposed reference populations, or fits well into all of them. Use of absolute genetic fit avoids errors in individual assignment conclusions that can arise from other commonly-used methods. Our visualization method can nonetheless be used to assign individuals to the population for which they have the highest LGP, or for the more conservative assignment process of excluding individuals from all populations for which they have LGP below a given quantile threshold. Notwithstanding this, genetic data are complex and it is standard practice in applied studies to use a variety of methods and software to gain a holistic understanding of genetic structure. We believe that GenePlots will prove a valuable additional tool in these studies.

An online interface for creating GenePlots is at `catchit.stat.auckland.ac.nz/shiny/geneplot/`.

7. Supplementary Materials

Web Appendices A to G, Tables, and Figures referenced in Sections 2 to 5, and the data and code used, are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

We thank the Editorial Team, and Dr Jesse Goodman for his help refining the saddlepoint implementation. LFM was supported by a University of Auckland Doctoral Scholarship. This work was funded by the Royal Society of New Zealand through Marsden grant 03-UOA-117.

References

Bagasra, A., Nathan, H. W., Mitchell, M. S., Russell, J. C. (2016). Tracking invasive rat movements with a systemic biomarker. *New Zealand Journal of Ecology* **40,** 267–272.

Baudouin, L., and Lebrun, P. (2001). An operational Bayesian approach for the identification of sexually reproduced cross-fertilized populations using molecular markers. *ISHS Acta Horticulturae* **546,** 81–93.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., and Bonhomme F., (1996-2004). GENETIX 4.05, Windows TM software for population genetics. Laboratoire Génome, Populations, Interactions, Montpellier.

Bergl, R. A., and Vigilant, L. (2007). Genetic analysis reveals population structure and recent migration within the highly fragmented range of the Cross River gorilla (*Gorilla gorilla diehli*). *Molecular Ecology* **16,** 501–516.

Berry, O., Tocher, M. D., and Sarre, S. D. (2004). Can assignment tests measure dispersal? *Molecular Ecology* **13,** 551–561.

Butler, R. W. (2007). Saddlepoint approximations with applications. Cambridge, UK: Cambridge University Press. ISBN 9780521872508.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics* **25,** 631–650.

Davison, A. C., and Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika*, **75(3),** 417–431.

DiCiccio, T. J., Field, C. A., and Fraser, D. A. S. (1990). Approximations of Marginal Tail Probabilities and Inference for Scalar Parameters. *Biometrika*, **77(1),** 77–95.

Fewster, R. M. (2017). Some applications of genetics in statistical ecology. *Advances in Statistical Analysis*, in press.

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

Frantz., A. C., Pourtois, J. T., Heuertz, M., Schley, L., Flamand, M. C., Krier, A., Bertouille, S., Chaumont, F., and Burke, T. (2006). Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. *Molecular Ecology* **15,** 3191–3203.

Glover, K. A., Hansen, M. M. and Skaala, O. (2009). Identifying the source of farmed escaped Atlantic salmon (*Salmo salar*): Bayesian clustering increases accuracy of assignment. *Aquaculture* **290,** 37–46.

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27,** 3070–3071.

Kotze, A., Ehlers, K., Cilliers, D. C., and Grobler, J. P. (2007). The power of resolution of microsatellite markers and assignment tests to determine the geographic origin of cheetah (*Acinonyx jubatus*) in Southern Africa. *Mammalian Biology* **73,** 457–462.

Lugannani, R., and Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability* **12,** 475–490.

Manel, S., Berthier, P., and Luikart, G. (2002). Detecting wildlife poaching: identifying the origin of individuals with Bayesian assignment tests and multilocus genotypes. *Conservation Biology* **16,** 650–659.

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics* **28,** 2537–2539.

Paetkau, D., Slade, R., Burden, M., and Estoup, A. (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Molecular Ecology* **13,** 55–65.

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

Primmer, C. R., Koskinen, M. T., and Piironen, J. (2000). The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. *Proceedings of the Royal Society of London Series B* **267,** 1699–1704.

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

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

Rollins, L. A., Woolnough, A. P., Wilton, A. N., Sinclair, R., and Sherwin, W. B. (2009). Invasive species can't cover their tracks: using microsatellites to assist management of starling (*Sturnus vulgaris*) populations in Western Australia. *Molecular Ecology* **18,** 1560–1573.

Taylor, E. B., Boughman, J. W., Sniatynski, M., Schluter, D., and Gow, J. L. (2006). Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (*Gasterosteus aculeatus*) species pair. *Molecular Ecology* **15,** 343–355.

Underwood, J. N., Smith, L. D., Van Oppen, M. J. H., and Gilmour, J. P. (2007). Multiple scales of genetic connectivity in a brooding coral on isolated reefs following catastrophic bleaching. *Molecular Ecology* **16,** 771–784.

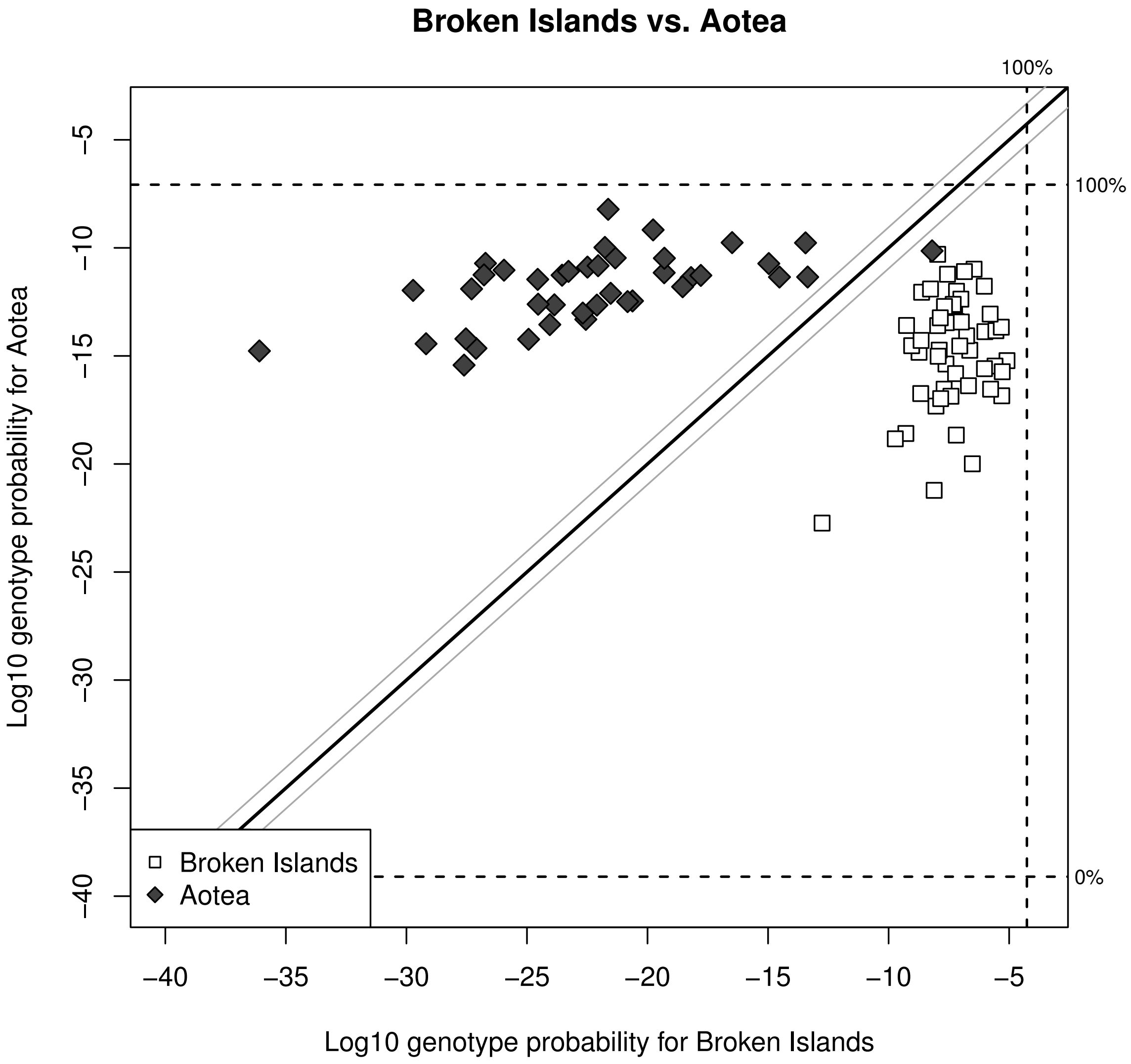

Figure 1. GenePlot based on microsatellite data extracted from ship rats (*Rattus rattus*). Each point represents an individual rat. Rats with missing data (i.e. data at fewer than 10 loci) are excluded. The graph shows rats captured on Aotea (diamonds) and the Broken Islands (squares). The horizontal axis shows the posterior log-probability of obtaining each individual's genotype from the Broken Islands population; the vertical axis shows the same, but with respect to the Aotea population. The thick diagonal line shows equal probability with respect to Aotea and the Broken Islands. The vertical dashed line shows the 100% percentile line, i.e. the maximum log-genotype probability, for the Broken Islands population; the horizontal lines show the 0% and 100% percentile lines for the Aotea population. This figure appears in color in the online version of this article.

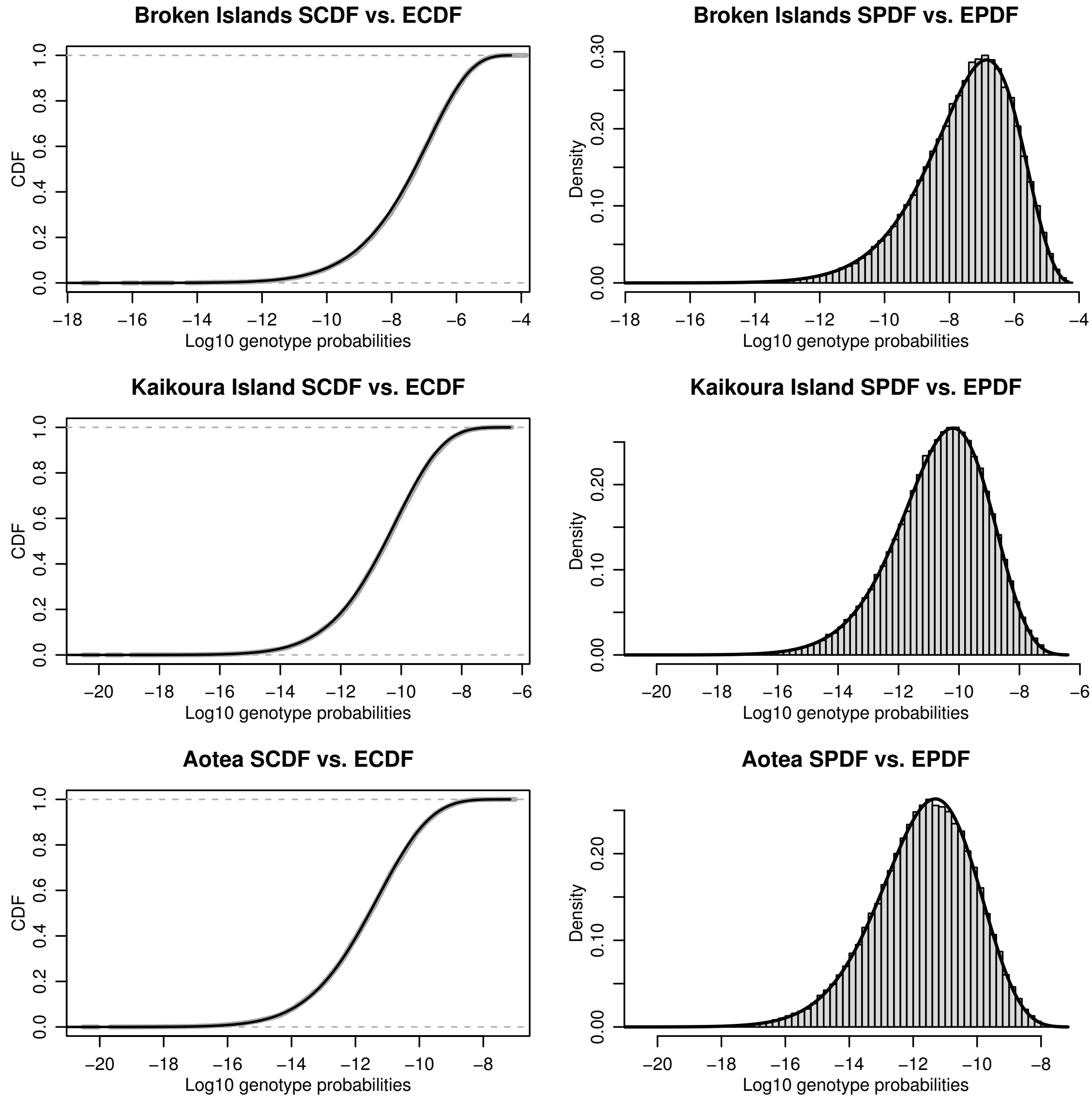

Figure 2. Genetic distributions for three populations. The plots in the left column show the CDFs of the multilocus LGP distributions. Wide, grey dashed lines show the empirical CDF based on 100,000 genotypes simulated from the population distribution (ECDF). Solid black lines show saddlepoint approximations to the CDFs (SCDF). The ECDFs are hard to see due to the closeness of the approximations. The plots in the right column show the corresponding PDFs. The histograms show 100,000 log-genotype probabilities for genotypes simulated from the population distribution (EPDF). The solid line shows the first derivative of the saddlepoint approximation to the CDF, which we denote SPDF. The top row shows the distribution for ship rats (*Rattus rattus*) captured on the Broken Islands; the middle row shows Kaikoura captures; the bottom row shows Aotea captures.

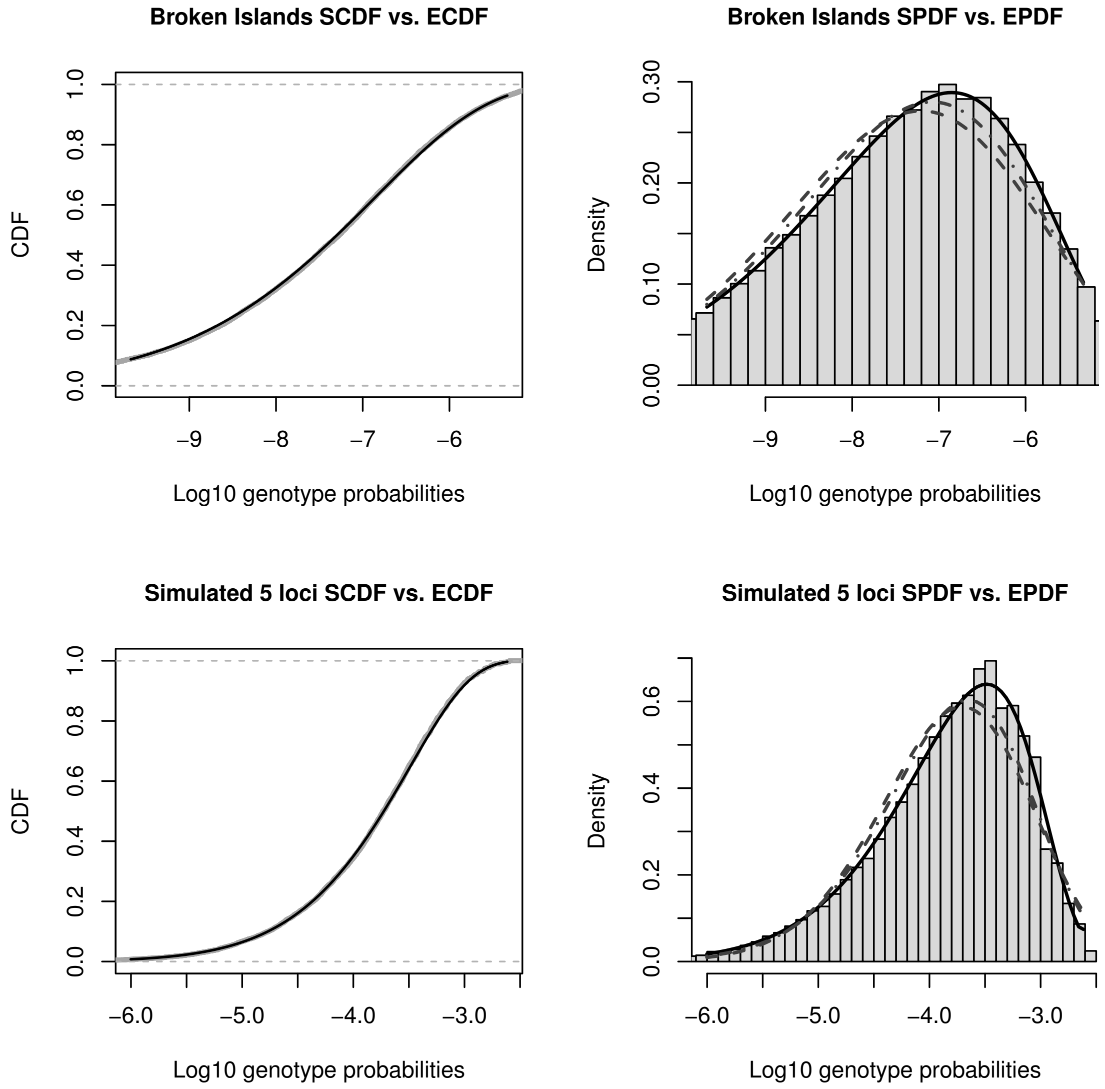

Figure 3. CDFs and PDFs of the LGP distributions of two populations, zoomed in to focus on an area around the mean. Details are as for Figure 2, with the addition of moment-based gamma and lognormal approximations shown as dashed and dot-dashed lines on the right-hand plots. The top row shows the distribution for ship rats captured on the Broken Islands. The bottom row shows the distribution of a simulated population, defined by randomly generated allele frequencies. The simulated data included 5 loci with between 2 and 5 allele types per locus.

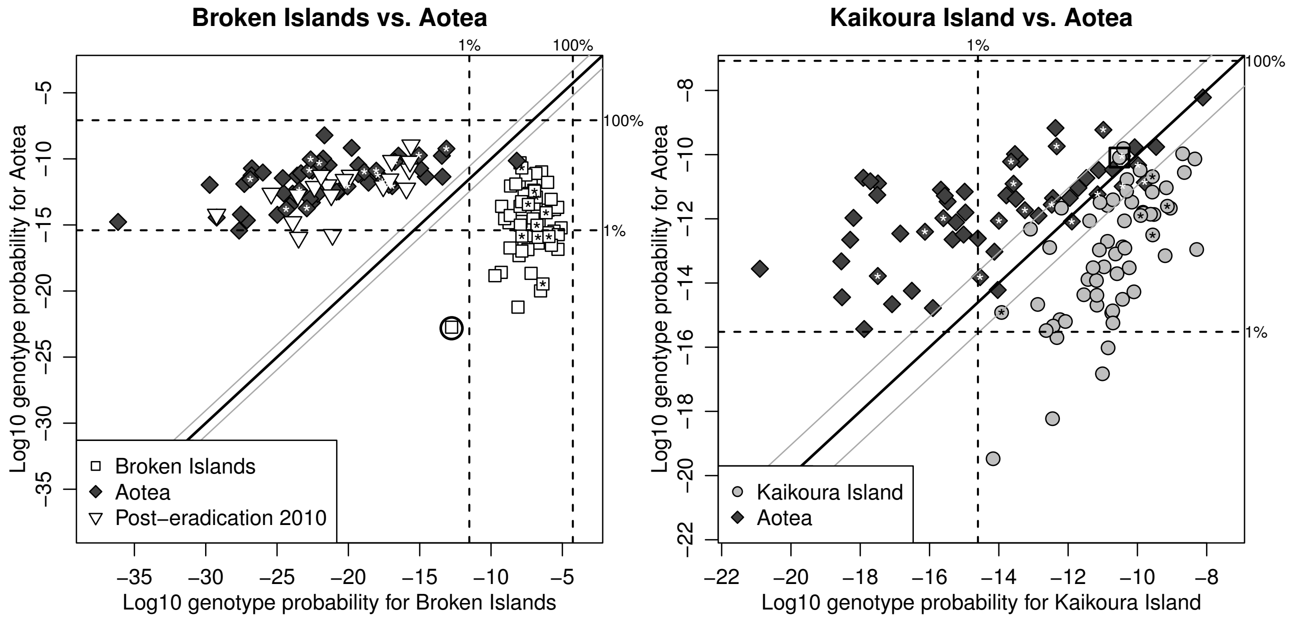

Figure 4. GenePlots based on microsatellite data extracted from ship rats. Each point represents an individual rat. Rats marked with asterisks have missing data at one or more loci; rats with data at fewer than 6 loci have been excluded. The left plot shows rats sampled from the Broken Islands and Aotea between 2005 and 2008, and rats sampled in 2010 on the Broken Islands following an eradication attempt on the Broken Islands. The right plot shows rats captured on Aotea and Kaikoura between 2005 and 2008. The Aotea samples are the same for both plots. The thick diagonal line shows equal posterior genotype probability for both populations; the thin diagonal lines indicate that the probability is 9 times larger for one population than for the other. In both plots, the horizontal lines show the 1% and 100% quantile lines for the Aotea population. In the left plot, the vertical dashed lines show the 1% and 100% quantile lines for the Broken Islands population; in the right plot, the vertical dashed line shows the 1% quantile line for the Kaikoura population. The rat marked with a circle in the left plot and the rat marked with a square in the right plot are marked in the STRUCTURE, GENECLASS2 and adegenet DAPC results shown in Web Appendix E. This figure appears in color in the online version of this article.

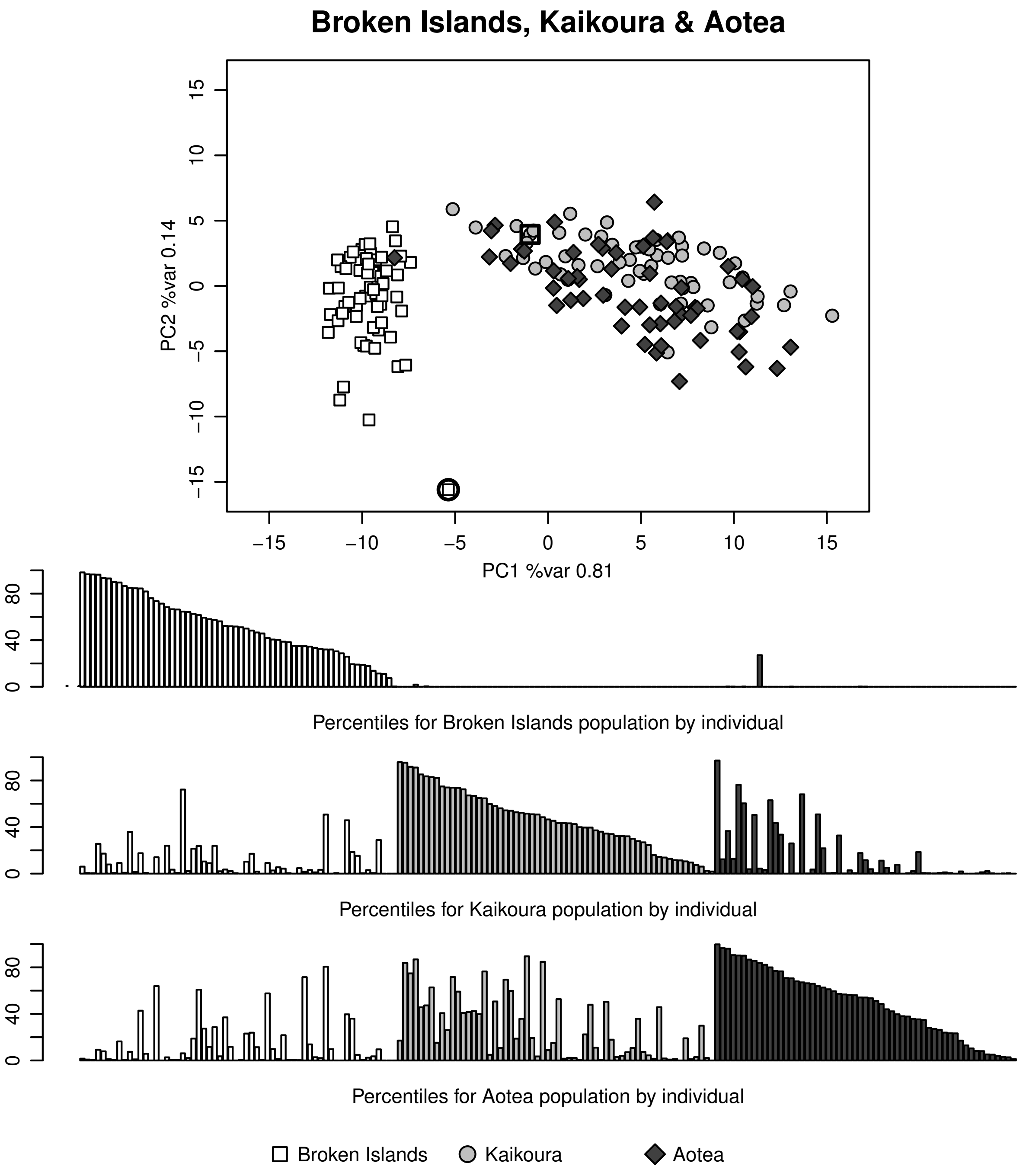

Figure 5. Visualizations for multiple reference populations. Individuals are ship rats captured on Kaikoura, the Broken Islands and Aotea between 2005 and 2008. Rats with data at fewer than 6 loci have been excluded. The top plot is a GenePlot for multiple reference populations: the axes show the first two principal components of the log-genotype probabilities with respect to the three populations. The bottom plot is a multi-bar GenePlot: one bar chart per population. Each bar chart shows all individuals, and a single bar represents the percentile corresponding to that individual's log-genotype probability with respect to the given population. The bars are grouped according to the population from which the rats were sampled. Individuals are vertically aligned among the three bar charts. The rat marked with a circle and the rat marked with a square are marked in the STRUCTURE, GENECLASS2 and `adegenet` DAPC results shown in Web Appendix E. This figure appears in color in the online version of this article.

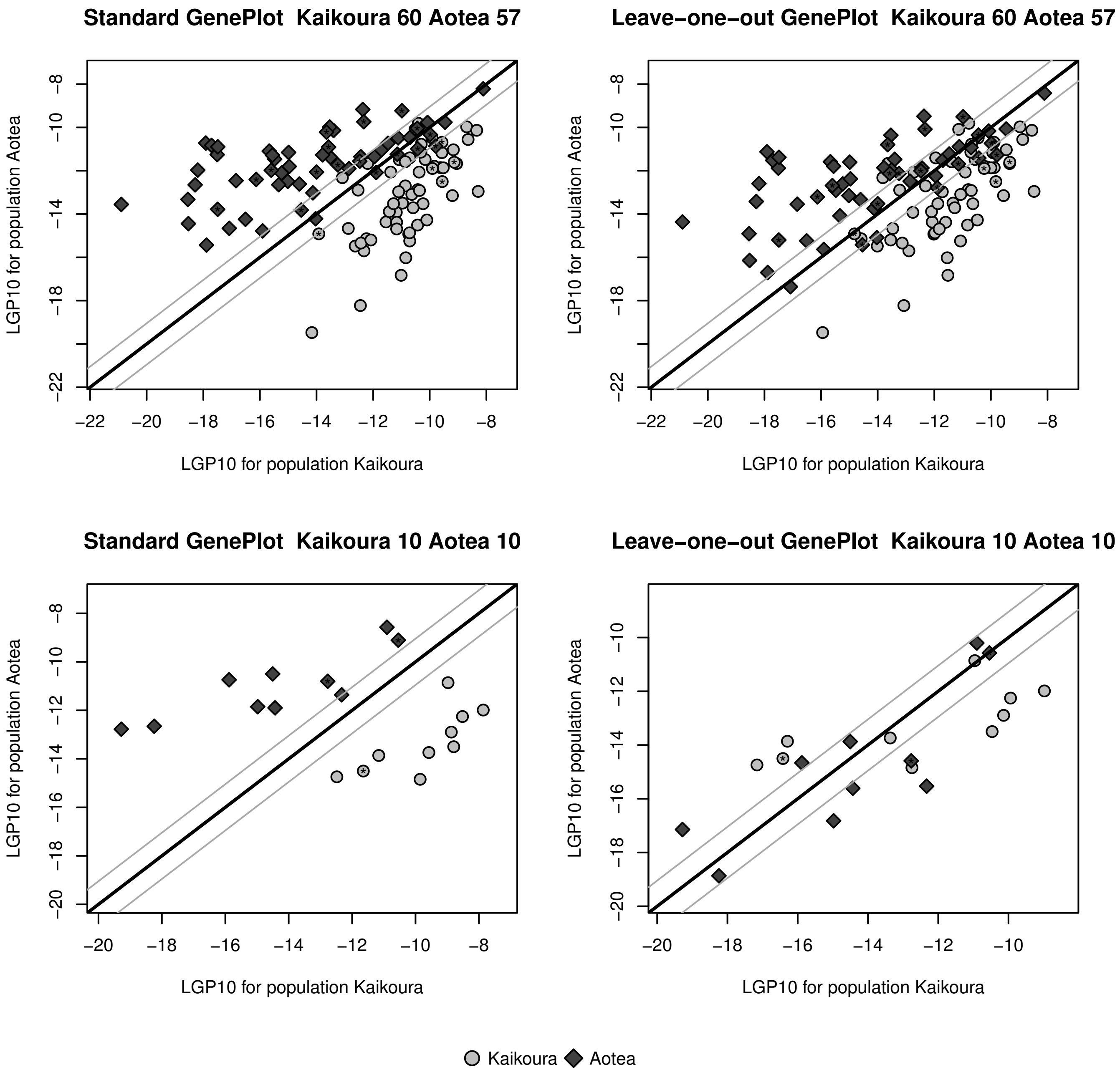

Figure 6. Leave-one-out GenePlots using random subsets from samples of rats captured on Kaikoura and Aotea. Rats marked with asterisks have missing data at one or more loci; rats with data at fewer than 6 loci have been excluded. The left column plots are standard GenePlots; the right-column plots show the same data analysed using the leave-one-out method. The top plots show the full samples of sizes 60 and 57 respectively, and the bottom plots show GenePlots constructed using subsets of 10 individuals from each sample. Using the leave-one-out method has a minor effect for the full samples, but it has a pronounced effect when using small sample sizes. This figure appears in color in the online version of this article.