Concepts of statistical analysis, visualization, and communication in population genetics

Louise Fastier McMillan

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Statistics,
The University of Auckland, 2019.

Abstract

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 & 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. We call the resulting plots *GenePlots*.

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.

We then advance these techniques further by proposing methods for quantifying population genetic structure, and associated tests of significance. The measures we propose are closely related to GenePlots, and enable visual features obvious from the plots to be expressed more formally. One measure is the *interloper detection probability*: for two random genotypes arising from populations A and B, the probability that the one from A has the better fit to A and thus the genotype from B would be correctly identified as the 'interloper' in A. Another measure is the *correct assignment probability*: this corresponds to the probability that a random genotype arising from A would be correctly assigned to A rather than B.

Using permutation tests, we can test two populations for significant population structure. These permutation tests are sensitive to subtle population structure, and are particularly useful for eliciting asymmetric features of the populations being studied, e.g. where one population has undergone extensive genetic drift but the other population has remained large enough to retain greater genetic diversity. We illustrate the new methods using microsatellite and SNP data, as well as simulation studies. We also compare the results with existing measures of genetic diversity and differentiation.

Dedication

This work is dedicated to Giles Turner, without whose wondrous willingness to move to a distant land this PhD would never have started, and without whose multifaceted support this PhD would never have been completed.

Acknowledgements

I would first like to thank my supervisor, Rachel Fewster, for her guidance, her support and her persistence in the face of my stubbornness. I have always felt like we were on similar wavelengths, looking at the world from similar viewpoints. Even when we were walking through fog in the midst of a particularly convoluted technical discussion, I knew she was only a short distance away and reachable if I could only walk a few steps further. It has been a pleasure to work with her and I am extremely grateful to her for proposing that I begin this PhD.

I would also like to thank my co-supervisors James Russell and Paul Murrell. Without James Russell I would have had a much narrower view of the field of invasive pests, and without Paul Murrell my graphs would have been a great deal less elegant, which would have been a pity in a thesis focused on visualization. I also thank Dr Jesse Goodman for his help refining the saddlepoint implementation and making it more practically useful. I thank Emma Carroll for her very helpful suggestions and for providing the southern right whale data.

I would like to thank the University of Auckland for my Doctoral Scholarship, and the staff of the Department of Statistics for making it such a pleasant place to work. I thank Joei Mudaliar for her help on so many occasions, and for her cheerful and welcoming friendship. I thank Ilze Ziedins for backing me on several projects, and giving me the valuable opportunity to lecture STATS10X during my PhD. I thank Russell Millar for his friendship, support and sound advice, and Brian McArdle for his delightful humour and cynicism.

I would like to thank Rebecca Turner for all her help and friendship, and for our joint effort at making the Maths and Stats Departments a little more sociable. I would like to thank James Brock for giving me an interesting problem to work on, and for making the crazy world seem like a friendlier place. I would also like to thank Daisy Shepherd for many enjoyable chats about PhD life, Niffe Hermansson for his support and many even longer discussions, Liza Bolton for being such a happy bouncy ray of sunshine that I cheer up whenever I see her, and finally Matt Edwards, Blake Seers, David Huijser and Cindy Gunawan for not only helping me out during my PhD but also being eager participants in many many sessions playing board games.

I thank my parents, Alice and Brian McMillan, for their support from afar, and for putting up with me working whenever I come to visit them in the UK, and I thank my brother Euan McMillan for always encouraging me.

I would most of all like to thank Giles Turner, who has not let his initial concerned reaction to my suggestion of doing a PhD get in the way of supporting me before I started and

throughout the whole long process. I have often felt like I was following an alternative PhD path to the norm, never descending to the point of hating my PhD but instead having other struggles to cope with, and Giles has been my invaluable friend every step of the way.

Contents

APPENDICES FOR CHAPTER 2 **43**

APPENDICES FOR CHAPTER 3 **129**

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Population genetics concepts

Population genetics is the study of genetic variation within and among populations of organisms. It can be applied to both plant and animal species, and to genetic data of any ploidy, although we will only deal with diploid data here.

There are two common questions in population genetics. One question is whether there is any evidence of population structure, defined as divisions between groups of organisms, with preferential mixing within groups and reduced dispersal among groups. Population structure can be basic, with just simple divisions, or there could be more complex hierarchical structure; and the divisions may be caused by behavioural patterns or, more commonly, by geographical barriers that discourage or prevent dispersal. The divisions may be detected by calculating various diversity measures (see Section 1.3) and testing them for significance, or by using multivariate methods such as discriminant analysis, or via clustering methods such as those used within the popular STRUCTURE software (see Sections 1.2.2 and 1.2.3).

The other main question is whether individuals can be accurately assigned to the populations they originated from[1]. Assignment can be used to identify the source populations of new individuals based on a reference sample from each of several candidate source populations. It can also be used to identify migrants within that reference sample, who originated from populations different from those they were sampled in, or admixed individuals, who have inherited genetic data from ancestors in more than one of the source populations.

Assignment involves two distinct stages: calculation of how well each individual fits to each of the candidate source populations; and assignment of each individual to the most appropriate source population, following a specific assignment protocol. The most common

[1] Individuals originate from population A if they were born in population A to at least one parent from population A, or were born in population B to two parents from population A.

assignment protocol is *best population assignment*, which assigns individuals to the source population for which they have the best fit. The fit of an individual into a population is assessed by how well-aligned the individual's genotype is with those of individuals in the reference sample from that population. The complication is that the reference sample of individuals presumed to have originated from that population is itself subject to sampling variability.

Detection of population structure and assignment of individuals both attempt to determine the genetic patterns within populations or clusters of individuals. These patterns are often expressed in terms of *allele frequencies*: the frequencies of different allele types within each group, usually assessed at multiple points on the genome, known as *loci*. The data can come from many different types of genetic markers, including but not limited to: allozymes; amplified fragment length polymorphisms (AFLPs); short tandem repeats (STRs), commonly known as microsatellites; and single nucleotide polymorphisms (SNPs).

We use the term *meta-population* throughout to refer to the overall set of populations studied, although it is also common in the literature to see the term "population" for the meta-population and "subpopulation" for a single population.

1.2 Population genetics software

The current ecosystem of population genetics software programs is vast and still growing. Excoffier & Heckel (2006) provided a "survival guide" to the available genetics software packages.

Since many of the programs have non-overlapping functionality, most ecological studies make use of several different programs to analyse their data. For example, Dussex et al. (2016), whose genetic data we reassess in Section 3.5 using our own metholodogy, cite twelve different software programs used in their study: GenAlEx 6[2] (Peakall & Smouse 2006); FSTAT 2.9.3 (Goudet, 1995[3]); GENEPOP 3.1d (Raymond & Rousset 1995); ARLEQUIN 3.1 (Excoffier & Lischer 2010); STRUCTURE 2.2 (Pritchard et al. 2000, Falush et al. 2003), STRUCTURE HARVESTER (Earl & vonHoldt 2012); CLUMPP (Jakobsson & Rosenberg 2007); DISTRUCT (Rosenberg 2004); the `adegenet` package (Jombart & Ahmed 2011) for the R language (R Core Team 2017); BAYESASS (Wilson & Rannala 2003); GENECLASS2 (Piry et al. 2004); and DIYABC v1.0.4 (Cornuet et al. 2010), as well as other statistical and genetic analysis methods.

Here we discuss the various software packages according to the type of functionality they provide. We begin with GENECLASS (Piry et al. 2004), which is closely related to our methodology, and STRUCTURE, which is the most commonly used software for analysing population structure.

[2] They cite Peakall & Smouse (2001) but the citation for GenAlEx 6 should be Peakall & Smouse (2006).

[3] The citation gives an incorrect year, 2001.

1.2.1 GENECLASS

GENECLASS (Cornuet et al. 1999, Piry et al. 2004) is a commonly used program for genetic assignment of individuals. It implements the Bayesian assignment method of Rannala & Mountain (1997), as well as the frequentist method of Paetkau et al. (1995) and the distance method of Cornuet et al. (1999). Rannala & Mountain released their own software called IMMANC, but it has been much less popular than GENECLASS.

Paetkau et al. (1995) propose a frequentist approach for assigning individuals. The method calculates the fit of each individual to each candidate source population as the expected frequency of the individual's genotype in that the source population, using the allele frequencies observed in the reference sample. The individual is then assigned to the population for which its genotype has the highest expected frequency. This is effectively a maximum-likelihood method. Paetkau et al. modify the basic method by adding the alleles of each individual's genotype to the reference samples for all the populations they were not sampled in before calculating the likelihoods. This reduces the bias produced by each individual's alleles being present in the population they were sampled in, and ensures that there are no candidate source populations that do not contain at least one copy of the individual's alleles (the null frequency problem). The method assumes random mating within each population, and linkage equilibrium among all the loci.

Rannala & Mountain (1997) propose an alternative method for assigning individuals, which estimates the posterior allele frequencies of the populations rather than using the observed allele frequencies in the sample, before calculating the fit of individuals to populations. Further details of this method are given in Section 2.2.

The method of Rannala & Mountain (1997) is described as a "Bayesian" method by Cornuet et al. (1999), but it has also been called a "partial Bayesian" method (see e.g. Manel et al., 2002) because, like the Paetkau et al. (1995) method, it uses a likelihood calculation to assess the fit of each individual to a population, given the estimated posterior allele frequencies in the population. Unlike Paetkau et al., however, they do not suggest a maximum likelihood, or best assignment, protocol, but rather recommend comparing the fit of an individual to different populations as a likelihood ratio, and determine an appropriate threshold for assignment using Monte Carlo simulations of genotypes.

Cornuet et al. (1999) describe the methods used in the first version of GENECLASS. This includes the methods of Paetkau et al. (1995) and Rannala & Mountain (1997), but also a new distance-based method. In this approach the fit of an individual to a population is calculated using a measure such as one of Nei's measures for population comparisons(Nei, 1987), and by treating the individual as a sample population of its own, described by its two alleles at each locus.

Cornuet et al. (1999) also introduce a new assignment protocol. They first calculate the fit of the individual to each population, relative to typical genotypes from those populations. These "typical genotypes" are simulated based on either the observed allele frequencies in the reference sample from that population, or based on the posterior allele frequencies estimated according to the method of Rannala & Mountain. For each population, Cornuet et al. then calculate the percentage of simulated genotypes from that population whose fit to their own population is lower than or equal to the fit of the target individual to the population, and treat this is as a "probability that the individual belongs to the population".

Cornuet et al. suggest the possibility of assigning the individual to the population for which they have the highest "probability of belonging", similar to the best assignment protocol of Paetkau et al. (1995), but the probabilities can also be used within a stricter exclusion protocol, under which the individual is only assigned to a population if there is only one population for which it has a fit above a given threshold (see Manel et al., 2002). If there are multiple populations for which the individual has a fit above the threshold, or there are no populations for which the individual has a fit above the threshold, then the individual is not assigned to any population. Such a method allows for the possibility that there is insufficient genetic differentiation among populations to assign that individual conclusively, or the possibility that the individual's true source population was not among those sampled.

Paetkau et al. (2004) update the method of Paetkau et al. (1995) by using leave-one-out to remove an individual from the population in which it was sampled, rather than adding the individual's genotype to all other sampled populations. They deal with null frequency alleles by setting the frequency of those alleles to a small non-zero value.

Paetkau et al. (2004) calculate two scores for each individual and each population. One is the likelihood L_h of the individual's genotype arising in the population within which it was sampled. The other is a test statistic $\Lambda = L_h/L_{\text{max}}$, where L_{max} is the maximum of the individual's likelihoods over all sampled populations. They then rank all the individuals by Λ or by L_h, depending on whether all possible source populations are thought to have been sampled, and treat any individuals with values below a threshold as migrants.

Paetkau et al. (2004) calculate the threshold for identifying migrants by simulating large numbers of genotypes from the populations, either by drawing alleles at random from the allele frequencies in each population and combining them into genotypes, or by simulating breeding of individuals from the population reference samples, picking two parents at random from the sample and combining half of the alleles from each parent. Paetkau et al. then either treat the simulated genotypes from a population as a single large sample to calculate allele frequencies for the population and calculate the likelihoods of the simulated individuals based on those allele frequencies, or group the simulated individuals into subsets the same size as the original population reference samples, and calculate allele frequencies and likelihoods within each subset before collating the likelihoods of all the subsets. The aim of the latter procedure is to reflect the sampling uncertainty inherent in the finite reference sample when establishing population quantiles.

Baudouin et al. (2004) propose an alternative prior for the population allele frequencies than the one used by Rannala & Mountain: their alternative prior was originally proposed by Baudouin & Lebrun (2001). They also introduce the concept of assigning a group of new individuals rather than assigning each separately.

GENECLASS2 (Piry et al. 2004) updates the first version of GENECLASS by incorporating the prior of Baudouin & Lebrun as an option within the Rannala & Mountain assignment method, and by adding the group-assignment option of Baudouin et al. (2004). GENECLASS2 calculates the L_h and Λ criteria defined by Paetkau et al. (2004), and an additional criterion, which is the ratio of the individual's likelihood in its own population to the maximum of its likelihoods over all sampled populations excluding its own.

GENECLASS2 also calculates what they term as the "probability of belonging", for each individual to each population, namely the quantile of the individual's fit among the fit of individuals simulated according to the methods in Paetkau et al. (2004). Alternatively, if the probability of belonging method is not used, the software instead calculates a score for each individual and each population which is the ratio of the individual's likelihood in that population to the sum of the likelihoods of the individual in all populations[4].

1.2.2 STRUCTURE and related packages

STRUCTURE occupies the centre of a sub-ecosystem of its own, surrounded by various programs that either carry out additional processing on the output of STRUCTURE or extend its functionality for different scenarios such as those with high levels of inbreeding (e.g. InStruct: see Gao et al., 2007).

1.2.2.1 Versions of STRUCTURE

Pritchard et al. (2000) propose the first version of the STRUCTURE methodology, with the primary purposes of assigning individuals to populations and detecting cryptic, or previously unidentified, population structure, and determining which apparent population divisions are spurious. To this end, the software uses Bayesian Markov Chain Monte Carlo (MCMC) methods to iteratively reassign individuals to different clusters. The MCMC process alternates between estimating the allele frequencies for a fixed cluster membership combination and estimating the best cluster membership for a fixed set of allele frequencies within those clusters.

The initial version assumes that each of the true populations is in linkage equilibrium and Hardy-Weinberg equilibrium, and attempts to find a cluster membership combination that minimises deviations from those equilibria within the clusters. The software does have an optional correlated allele frequencies model, which allows for the allele frequencies to be correlated across the clusters, because without that the software has a tendency to merge clusters that have similar allele frequencies.

[4] See http://www1.montpellier.inra.fr/CBGP/software/GeneClass/GeneClass2/Help.pdf for a corrected version of the score equation given in Piry et al (2004).

The initial version has two models: (a) the no-admixture model, in which each individual is assumed to have ancestry from only one cluster; and (b) the admixture model, in which each individual is allowed to have portions of its ancestry from multiple clusters. In the admixture model, the software outputs a vector $\boldsymbol{q}$ for each individual that gives the proportion of the individual's ancestry that comes from each cluster.

The initial version of STRUCTURE also allows the user to incorporate a limited form of prior population information by specifying the fixed value of a parameter, ν, that gives the probability of an individual being a recent immigrant to the sampling location it was found in. Hubisz et al. (2009) stress that this feature is limited, and is mostly intended for the identification of a few recent immigrants in the samples.

STRUCTURE has many other detailed parameters and settings, some of which relate to the MCMC method, such as the number of burn-in iterations, the number of final iterations, and the number of replicate runs to use. Running replicate runs for each combination of settings is recommended because the method, like all MCMC methods, is non-deterministic.

Falush et al. (2003) update the STRUCTURE methodology to allow for linkage disequilibrium within admixed clusters, and to extend the software to work with any ploidy. Pritchard et al. (2000) had assumed that the allele frequencies at a single locus in a single cluster were drawn from a symmetric Dirichlet distribution parameterized by λ, with a separate value of λ for each cluster. The admixture proportions for a single individual were also assumed to be drawn from a symmetric Dirichlet distribution with the same parameter α for all clusters.

Falush et al. (2003) describe three forms of linkage disequilibrium (LD): (i) "mixture LD", caused by the fact that individuals who mostly have ancestry from a particular cluster will have an excess of alleles that are common in that cluster; (ii) "admixture LD", which is the correlation between loci that are near to each other on the same chromosome; and (iii) "background LD". The method of Pritchard et al. (2000) accounted for mixture LD, whereas the method of Falush et al. (2003) accounts for admixture LD by allowing the admixture parameter α to vary among clusters, although they still ignore background LD.

The newer model assumes that the loci are divided into "chunks" that are near to each other on the same chromosome and are inherited together, so that the vector of cluster membership along a single chromosome is a Markov model, where the length of each chunk varies randomly. This Markov model gives rise to a Hidden Markov Model for the observed genotype data.

Falush et al. (2003) note that the software will make better inferences if some of the data from each individual is from unlinked or weakly linked genetic regions.

Although the model of Falush et al. (2003) does not account for background LD directly, they found from simulations with two clusters that in cases where background LD was a problem, STRUCTURE tended to infer admixture across all individuals in both clusters, whereas admixture LD tended to be asymmetrical, affecting some clusters much more than others. For cases with more than two clusters and high levels of background LD, they found that more admixture was inferred between closely related clusters, and so they concluded that admixture found between distantly related populations is unlikely to be due to background LD.

Falush et al. (2007) extend STRUCTURE further by providing a method that allows for ambiguous ascertainment of genotypes in diploid data with dominant markers, or polyploid data with dominant or co-dominant markers. The method infers the true genotypes given the incomplete observations as well as the allele frequencies and admixture proportions.

Hubisz et al. (2009) extend STRUCTURE by allowing the inclusion of sampling location priors, to give more weight to clustering outcomes that are correlated with the sampling locations. Previous versions of STRUCTURE treated all possible cluster membership combinations as equally likely and therefore any given specific combination was very unlikely. The new method, in non-admixture mode, allows for non-equal probabilities of membership of different clusters, and in admixture mode allows for non-equal admixture priors for the different clusters.

Hubisz et al. (2009) designed the priors for the parameters of cluster membership so that if the sampling location labels are strongly uncorrelated with the inferred ancestry proportions then the sampling locations will not bias the results. They recommend the new model for cases with limited data and weak or no signal of population structure. They do not incorporate geographical data, only the sampling location labels.

1.2.2.2 Choosing the best value of K

STRUCTURE runs the MCMC process for a fixed value of K, the number of clusters, and does not directly infer the value of K. Instead, Pritchard et al. (2000) recommend running the process for multiple candidate values of K and then selecting the best value of K. Much effort since then has been focused on this problem of determining the true number of clusters.

Pritchard et al. (2000) offer a rough method of estimating the log-likelihood for the observed genotypes given K, and use that to obtain a posterior estimate for K. They emphasise that this is an "*ad hoc* guide to which models are most consistent with the data", and note that the posterior distribution of K seems to be more dependent on modelling assumptions and choice of priors than the posterior distributions of the other parameters.

Falush et al. (2003) comment further on the difficulty of estimating K: "the number of populations supported by the data may depend on how different one would expect the allele frequencies in different populations to be *a priori*, which is often difficult to specify."

Evanno et al. (2005) note that many studies chose K to maximise the log-likelihood $L(K) = \log \mathbb{P}(X \mid K)$, and that method evaluation studies had shown STRUCTURE to be a good tool for assigning individuals to populations, but had not tested the suitability of STRUCTURE for detecting the correct value of K, especially in cases with hierarchical population structure.

Evanno et al. tested STRUCTURE using simulated dispersal scenarios with three types of structure: simple island model, with dispersal between all pairs of populations; hierarchical model, with preferential dispersal within groups of populations and with all groups linked to each other; and a “contact zone” model akin to the “stepping stone” model of Easypop (Balloux, 2001), with dispersal only between adjacent populations and reduced dispersal between the two central populations, marking the contact zone between otherwise isolated groupings. They used Easypop to simulate microsatellite and amplified fragment length polymorphism (AFLP) data at 100 loci, with 10 allele types at each locus with mutation rate $\mu = 10^{-3}$. They used the admixture mode of STRUCTURE with correlated allele frequencies between populations.

Evanno et al. (2005) found that the log-likelihood $L(K)$ did not tend to have a clear mode at the correct value of K. They instead propose a new criterion, now commonly known as the ΔK criterion of Evanno et al. (2005), which is calculated based on the second-order rate of change of $L(K)$ with respect to successive values of K. This new criterion tended to have a clear peak at a value of K correctly matching the higher-level population structure. For the hierarchical model, the ΔK method correctly indicated the number of population groupings, and for the contact zone model, the method correctly indicated that there were two population groups.

1.2.2.3 STRUCTURE output

The basic output of STRUCTURE is the inferred cluster membership of each individual for the no-admixture mode, or the inferred admixture proportions of each individual for the admixture model. Although these proportions are not the probabilities of membership of the clusters for the individual, they are often used as such (see e.g. Manel et al., 2002).

The other main output of STRUCTURE in admixture mode is a bar chart showing the admixture proportions of each individual, coloured by cluster. This may be an implementation of the method used in the DISTRUCT software (Rosenberg, 2004), although that program is also used alongside STRUCTURE in many studies. Figure 1.1 shows an example of this display.

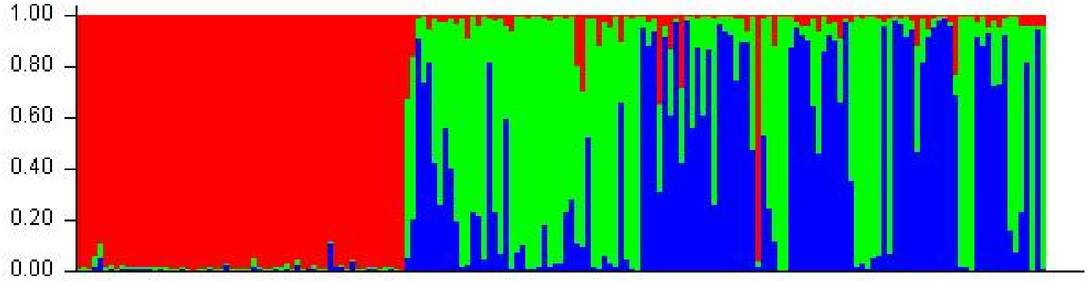


Figure 1.1: Example of a STRUCTURE bar plot of ship rats captured on Kaikoura Island, the Broken Islands and Aotea, New Zealand, between 2005 and 2008. The red bars correspond to cluster 1, the green bars to cluster 2 and the blue bars to cluster 3. See Chapter 2 for a detailed discussion of this case study.

Additional software packages have been developed to make the output of STRUCTURE more useful or more configurable. CLUMPP (Jakobsson & Rosenberg 2007) aims to collate the results of multiple replicate runs from clustering software such as STRUCTURE and deal with the issue of permutation of cluster labels between different replicate runs. CLUMPAK (Kopelman et al. 2015) groups replicate runs into subsets with similar outcomes that may indicate different modes in the space of possible solutions. STRUCTURE PLOT takes the output of STRUCTURE or CLUMPP and provides an interactive interface for editing the resulting bar plots. Earl & vonHoldt (2012) created STRUCTURE HARVESTER to collate STRUCTURE results, visualize likelihoods for K, calculate Evanno's ΔK criterion, and reformat data for programs such as CLUMPP and DISTRUCT. ObStruct (Gayevskiy et al. 2014) processes output from STRUCTURE, InStruct or BAPS (Corander et al., 2003 and 2004) and aims primarily to assess how the inferred cluster membership relates to a factor of interest encoded as the population labels of the individuals.

Many other packages exist to run STRUCTURE in batch mode or parallelize it.

1.2.2.4 Further developments

Kalinowski (2011) assesses the ability of STRUCTURE to identify, from a large number of populations, a smaller number of clusters representing the major divisions within the species, and argues that the clusters identified by STRUCTURE do not fit well with the evolutionary history of the species. Kalinowski demonstrates, via simulations, that the STRUCTURE clusters can be strongly influenced by sample size variation and by the relative amount of differentiation among the populations. Kalinowski also reanalysed human genetic data used in studies by Rosenberg et al. (2005) and Tishkoff et al. (2009) and disputes the conclusions they had drawn from STRUCTURE analyses.

Lawson et al. (2012) note that STRUCTURE and similar methods treat loci individually, without incorporating information about the relative positions of the loci in the genome. They introduce two new software programs, ChromoPainter and fineSTRUCTURE, for biallelic genomic data. fineSTRUCTURE is intended as an improvement on STRUCTURE and another software program, ADMIXTURE (Alexander et al. 2009, Alexander & Lange 2011).

ChromoPainter "paints" stretches of each individual's genome to indicate which of the other individuals in the sample is the nearest neighbour for each stretch, and produces a coancestry matrix showing, for each individual, what proportions of its genome have each other individual as its nearest neighbour.

fineSTRUCTURE uses this coancestry matrix to perform model-based inference akin to the STRUCTURE methods, or non-model-based inference akin to principal components analysis (PCA). The authors recommend using it not to find major splits between populations, but rather to find finer subdivisions. Lawson et al. (2012) comment that fineSTRUCTURE performs better than STRUCTURE for large numbers of clusters ($K > 10$).

Lawson et al. (2018) criticise the way that STRUCTURE is commonly applied to detect population structure, in particular the apparent tendency to assume that the estimated value of K

is correct and that each of the inferred clusters corresponds to a non-admixed ancestral population, ignoring the possibility of admixture within or relationships between ancestral populations. Lawson et al. (2018) simulated three scenarios that all produce similar STRUCTURE output: (i) one population was formed by recent admixture of the other three populations; (ii) one population was formed by admixture between one of the other populations and an unsampled "ghost" population; and (iii) two of the populations became separated and one of them underwent a strong recent bottleneck event. Lawson et al. introduce a new software program called badMIXTURE that assesses whether scenario (i), recent admixture, is the most plausible scenario for the given dataset. They also show that the results of both STRUCTURE and ADMIXTURE (Alexander et al. 2009, Alexander & Lange 2011) are very strongly affected by sample size.

1.2.3 Other assignment and clustering software

Other software for assignment and clustering includes WHICHRUN (Banks & Eichert 2000), for assigning individuals to populations, and ONCOR (Anderson et al. 2008, Kalinowski et al. 2008), commonly used for mixed stock analysis.

Partition (Dawson & Belkhir 2001) uses MCMC methods to partition the sample into groups of individuals, rather than attempting to infer the proportions of populations in a mixture. BayesAss (Wilson & Rannala 2003) uses MCMC methods to estimate rates of recent immigration, individual ancestries and other parameters of the population structure.

BAPS 2.1 (Corander et al., 2003 and 2004) provides a similar clustering method to STRUCTURE except that it directly estimates the number of clusters. Further modules were added in later versions of BAPS, including the optional incorporation of spatial data (Corander et al., 2008).

1.2.4 Multi-purpose software

There are several multi-purpose packages that can be used to calculate various diversity measures, carry out significance tests, check for deviations from Hardy-Weinberg equilibrium and estimate population parameters.

An early popular package was BIOSYS-1 (Swofford & Selander, 1981), which was later replaced in popularity by GENETIX (Belkhir et al. 1996-2004), GENEPOP (Raymond & Rousset 1995, and Rousset 2008) and FSTAT (Goudet 1995).

Later popular packages include GenAlEx 6.5 (Peakall & Smouse, 2006 and 2012) and ARLEQUIN 3.5 (Excoffier & Lischer 2010). ARLEQUIN implemented the analysis of molecular variance (AMOVA) method introduced by Excoffier et al. (1992) (see Section 1.3.3).

`adegenet` (Jombart & Ahmed 2011) is a general-purpose package for the R language, but focuses on multivariate analyses of genetic data, particularly using the Discriminant Analysis of Principal Components (DAPC) method (Jombart et al. 2010) (see Appendix E).

There are more recent software packages for calculating newer diversity measures such as G'_{ST} and Jost's D (see Section 1.3), including GENODIVE (Meirmans & Van Tienderen 2004), SMOGD[5] (Crawford, 2010), and the `SpadeR` (Chao et al. 2015), `diveRsity` (Keenan et al. 2013), `DEMEtics` (Gerlach et al. 2013) and `mmod` (Winter 2012) packages for the R language (R Core Team 2017).

Easypop (Balloux 2001) is a small, fast program for simulating populations with different types of simple or hierarchical structure and different levels of migration and mutation.

[5] The software package is available at https://github.com/ngcrawford/SMOGD but lacks a user guide.

1.3 Classical measures of population differentiation

As well as the methods of genetic assignment and clustering described in Section 1.2, genetic differentiation measures based on the classical Wright-Fisher model of population genetics (Wright 1951) have proliferated in recent years, leading to much disagreement about which measures should be used to assess population structure in different scenarios and using different genetic markers.

Here, we give a brief survey of measures in common use, and the numerous adjustments and disagreements that they have spawned. Our aim is to give a flavour of the context in which we propose our alternative approach to measuring population differentiation in Chapter 3, which is motivated by the visualizations we derive in Chapter 2. We will not explore the measures below further, except to use them for discussion and comparison in Sections 3.9 and 4.2.

1.3.1 F_{ST}

Wright (1951) initially proposed F_{ST}, and the other F-statistics F_{IS} and F_{IT}, defined for biallelic loci. Weir & Cockerham (1984) propose the method most commonly used for estimating F_{ST}, for which they use the notation θ after establishing its equivalence to F_{ST}. They also describe corresponding estimates for f and F, equivalent to F_{IS} and F_{IT}.

Weir & Cockerham split up the variance of the frequency of an allele into the variance among populations, the variance among individuals within populations, and the variance among gametes within individuals, called a, b and c. They then define

$$\theta = \frac{\mathbb{E}(a)}{\mathbb{E}(a) + \mathbb{E}(b) + \mathbb{E}(c)},$$

where $\mathbb{E}(x)$ is the expectation of quantity x, and they define F and f similarly. Estimates for θ, f and F are calculated using the components of the observed variance.

Holsinger & Weir (2009), in their summary of methods for estimating F_{ST}, describe the Weir & Cockerham (1984) method as "method-of-moments" analysis, carried out via ANOVA for allele frequencies. The initial definitions of $\hat{F}$, $\hat{\theta}$ and $\hat{f}$ in Weir & Cockerham (1984) are for a single biallelic locus. Estimates for multiallelic loci are obtained using a weighted average over the alleles, and similarly the estimate for multiple loci is obtained using a weighted average over loci.

The Weir & Cockerham method assumes equal-sized populations, descended from a single ancestral population in Hardy-Weinberg equilibrium and linkage equilibrium. Sample sizes are not required to be equal, but population sizes are required to be equal because the derivation involves expected values of f, θ and F and would give multiple expected values for unequal population sizes. The expectations are taken over all possible replicate populations and all possible samples from those populations, thus accounting for both genetic sampling uncertainty and statistical sampling uncertainty.

Weir & Cockerham (1984) and Holsinger & Weir (2009) distinguish between genetic sampling from generation to generation within populations, and statistical sampling caused by taking samples smaller than the populations. Genetic sampling uncertainty leads from the fact that most single finite generations in the population will not have allele frequencies that exactly match the underlying allele frequencies in the population. The authors distinguish between these two forms of sampling uncertainty because only statistical sampling uncertainty can be reduced, by taking larger samples from the population, whereas genetic sampling causes inherent and unavoidable uncertainty.

Weir & Cockerham comment that already by 1984 the earlier studies that used F_{ST} were inconsistent in their documentation of the calculations used and the assumptions required. They also argue that Nei & Chesser (1983) do not account for replication over replicate populations. However, most studies using or citing F_{ST} use the method of Weir & Cockerham without meeting the assumptions of equal population size and absence of mutation. Many studies, and software such as GenAlEx 6.5 (Peakall & Smouse, 2006 and 2012) calculate F_{ST} for pairs of populations, which contradicts the assumption in Weir & Cockerham of a constant value of F_{ST} across all populations. Weir & Hill (2002) extended Weir & Cockerham's method to obtain estimates for F_{ST} that vary between the populations, but in many studies it is unclear whether Weir & Hill's method is the one used.

Alongside the method-of-moments process of using ANOVA calculations to estimate F-statistics, Holsinger & Weir (2009) describe alternative methods involving maximum likelihood or Bayesian methods, including the methods of Weir & Hill (2002). Maximum likelihood methods for estimating F-statistics require specification of a likelihood model that describes how the allele frequencies vary among populations, and multinomial distributions for genotypes sampled from the populations. Bayesian methods for estimating F-statistics combine the likelihood model with prior distributions, often uniform, for θ and f and the allele frequencies.

Again, these maximum likelihood and Bayesian methods are for single loci but are extended to multiple loci by assuming equal θ and f over all loci and that "genotype counts are sampled independently across loci and populations" (Holsinger & Weir 2009). Holsinger & Weir (2009) suggest that Bayesian and method-of-moments estimates of F_{ST} will only be similar for large numbers of populations (>10) and sufficient average numbers of individuals per population (>20).

Gaggiotti & Foll (2010) complain that Holsinger and Weir only discuss overall estimates of F_{ST} for a set of populations, and recommend Balding (2003) for a more detailed discussion of F_{ST} methods, including a method for calculating F_{ST} separately for each population, which Gaggiotti & Foll prefer to calculating F_{ST} pairwise across populations.

Despite its ubiquity, therefore, there is only limited consensus about how F_{ST} should be defined or estimated and how it should be used, although the majority of studies using F_{ST} appear to use the form proposed by Weir & Cockerham (1984).

FSTAT (Goudet 1995) and Genepop (Raymond & Rousset 1995) both calculate F_{ST} using code developed by Goudet from FORTRAN code in Weir (1990) which follows the method of Weir and Cockerham (1984). Arlequin 3.5 (Excoffier & Lischer 2010) is not clear about the details of F_{ST} estimation but cites both Weir and Cockerham (1984) and Slatkin's (1995) adjusted method for estimating F_{ST} with microsatellite data. GenAlEx 6.5 (Peakall & Smouse, 2006 and 2012), which has become one of the most popular general software packages for population genetics, describes in Tutorial 1 a manual calculation for F_{ST} akin to Nei's (1987) method, but also offers AMOVA analysis including F_{ST} estimates[6].

1.3.2 Nei's G_{ST} and D_{ST}

Nei (1977) defines F_{ST} for a single locus as

$$F_{\mathrm{ST}} = \frac{H_{\mathrm{T}} - H_{\mathrm{S}}}{H_{\mathrm{T}}}$$

where H_{T} is the total heterozygosity of the meta-population and H_{S} is the within-population heterozygosity averaged over all populations. Under the appropriate conceptual Wright-Fisher model, this is equivalent to the previous definitions for a single biallelic locus; however, its generalization to multiple alleles and loci differs from that of Weir & Cockerham (1984). Nei (1973, 1987) therefore names his multiallelic analogue G_{ST}, to distinguish it from other concepts of F_{ST}, although many authors use F_{ST} and G_{ST} interchangeably.

Nei (1973) begins by defining the "gene identity" in population r at a single locus:

$$J_r = \sum_i p_{ri}^2, \tag{1.1}$$

where p_{ri} is the sample frequency of the ith allele in population r. He also defines the gene identity in the total population:

$$J_{\mathrm{T}} = \sum_i p_{\cdot i}^2, \tag{1.2}$$

where $p_{\cdot i}^2$ is the weighted average frequency of allele i over all populations. Nei names these quantities gene identity because they give the probabilities of gene identity of two randomly chosen genes, either from population r or from the meta-population. He also defines corresponding quantities $H_r = 1 - J_r$ and $H_T = 1 - J_T$, which he calls gene diversity because the commonly-used term "heterozygosity" is "not appropriate for a nonrandom mating population".

Nei defines D_{ST} as the difference between the average gene identity within populations, J_{S} and the gene identity in the meta-population, J_{T}:

$$\begin{aligned} D_{\mathrm{ST}} &= J_{\mathrm{S}} - J_{\mathrm{T}} \\ &= \frac{1}{K} \sum_r J_r - J_{\mathrm{T}}, \end{aligned}$$

[6] GenAlEx guides and tutorials refer for details to a separate document titled "Appendix 1", which does not appear to be available on the GenAlEx website http://biology-assets.anu.edu.au/GenAlEx/Welcome.html.

where K is the number of populations. Thus D_{ST} is a multiallelic calculation of $H_{\mathrm{T}} - H_{\mathrm{S}}$ because J_{T} represents $1 - H_{\mathrm{T}}$ and J_{S} represents $1 - H_{\mathrm{S}}$. Nei then defines

$$G_{\mathrm{ST}} = D_{\mathrm{ST}}/H_{\mathrm{T}}.$$

Nei also defines bias-corrected versions of H_{S}, H_{T}, G_{ST} and D_{ST} estimators (Nei 1978, and Nei & Chesser 1983). Nei & Roychoudhury (1974) considers the inter- and intra-locus variability of the heterozygosity.

Holsinger & Weir (2009) suggest that, unlike estimates of F_{ST}, which are designed to account for genetic sampling, Nei's G_{ST} is only concerned with the actual population allele counts observed in a generation, rather than the underlying allele frequencies. Similarly, Balding (2003) describes Nei's estimates for F_{ST}, G_{ST} and D_{ST} as "descriptive statistics, dealing with the observed data rather than inferring from them."

1.3.3 Φ_{ST} and R_{ST}

F_{ST} treats all allele types equally and does not use the information that some allele types might be more similar than others, in the sense that it is easier for some types of mutations to occur than others. The aim of Φ_{ST} (Excoffier et al. 1992) and R_{ST} (Slatkin 1995) is to exploit this information by using a measure of distance between allele types. Φ_{ST} does this for haplotype data. Analysis of molecular variation (AMOVA, Excoffier et al. 1992) derives the hierarchy of variances within and between populations, based on the sums of squared distances between pairs of individuals, and constructs Φ_{ST} as a corresponding analogue of F_{ST}.

Microsatellite markers, or "short tandem repeats" (STRs) were also considered problematic for estimating F_{ST}. Microsatellite allele types are defined by their sequence length and are often considered to follow a stepwise mutation model, in which each mutation adds or subtracts a small number of repeats, and therefore alleles are more likely to mutate into allele types of similar length than those of very different lengths. Slatkin (1995) introduced the measure R_{ST} specifically for microsatellite markers, to accommodate their stepwise mutation patterns. However, it has been suggested (Holsinger & Weir 2009) that estimates of R_{ST} may be unreliable unless a very large number of loci are used, and Balloux & Lugon-Moulin (2002) argue that no actual markers have been found that follow a stepwise mutation pattern.

1.3.4 Standardized forms of F_{ST} and other measures

Estimates of F_{ST} using biallelic markers, for which F_{ST} was originally defined, vary between 0 and 1: they reach 0 when the populations have exactly the same allele frequencies as each other, and 1 when the populations have totally different alleles from each other. Several studies, in particular Charlesworth (1998,) Hedrick (1999), Hedrick (2005) and Meirmans & Hedrick (2011), have demonstrated that estimates of F_{ST} for multiallelic markers do not exhibit the same behaviour.

In particular, using Nei's definitions of F_{ST}, the range of values taken by F_{ST} is heavily dependent on the value of the mean within-population heterozygosity H_{S}. As H_{S} increases, the maximum possible value of F_{ST} decreases so that for $H_{\mathrm{S}} = 0.9$, for example, $F_{\mathrm{ST}} = 0.1$ would indicate maximal differentiation with no alleles shared among the populations.

For this reason, F_{ST} is very commonly standardized as F'_{ST} by dividing F_{ST} by its maximum possible value for the given heterozygosity. However, there are alternative standardization approaches.

Nei (1987) defined a standardized form of G_{ST}. He defined

$$D'_{\mathrm{ST}} = (H_{\mathrm{T}} - H_{\mathrm{S}})\frac{K}{K-1}$$

where K is the number of populations in the meta-population, and from this he obtained

$$H'_{\mathrm{T}} = (H_{\mathrm{T}} - H_{\mathrm{S}})\frac{K}{K-1} + H_{\mathrm{S}}$$

and

$$G'_{\mathrm{ST(Nei)}} = \frac{H'_{\mathrm{T}} - H_{\mathrm{S}}}{H'_{\mathrm{T}}} = \frac{K(H_{\mathrm{T}} - H_{\mathrm{S}})}{KH_{\mathrm{T}} - H_{\mathrm{S}}}.$$

Hedrick (2005) defined a different standardized form of G_{ST} similar to the typical standardization of F_{ST}:

$$G'_{\mathrm{ST(Hedrick)}} = \frac{G_{\mathrm{ST}}}{G_{\mathrm{ST(max)}}} \tag{1.3}$$

where

$$G_{\mathrm{ST(max)}} = \frac{H_{\mathrm{T(max)}} - H_{\mathrm{S}}}{H_{\mathrm{T(max)}}}.$$

$H_{\mathrm{T(max)}}$ is the maximum heterozygosity in the total meta-population given the observed heterozygosity within populations:

$$H_{\mathrm{T(max)}} = \frac{1}{K}(K - 1 + H_{\mathrm{S}}). \tag{1.4}$$

Meirmans (2006) defined a standardized form of Φ_{ST},

$$\Phi'_{\mathrm{ST}} = \Phi_{\mathrm{ST}}/\Phi_{\mathrm{ST(max)}}.$$

This adjusts for the high levels of within-population variance for highly polymorphic markers. The calculations for estimating $\Phi_{\mathrm{ST(max)}}$ depend on the population model, mainly on whether there is only simple structure or a hierarchical structure with groupings of populations within the meta-population. Then AMOVA methods used to calculate Φ_{ST} are extended to obtain $\Phi_{\mathrm{ST(max)}}$ and Φ'_{ST}.

Meirmans & Hedrick (2011) define a further standardization of G_{ST}, based on the fact that the maximum possible value of Nei's G'_{ST} is $1 - H_{\mathrm{S}}$. They define

$$G''_{\mathrm{ST}} = \frac{G'_{\mathrm{ST(Nei)}}}{G'_{\mathrm{ST(Nei,max)}}} = \frac{K(H_{\mathrm{T}} - H_{\mathrm{S}})}{(KH_{\mathrm{T}} - H_{\mathrm{S}})(1 - H_{\mathrm{S}})}.$$

Meirmans & Hedrick (2011), in the appendix, provide a useful summary of the standardized forms of G_{ST}. They consider G''_{ST} and Φ'_{ST} to be estimates of F'_{ST}, i.e. standardized estimates of F_{ST}.

1.3.5 Jost's D and D_{est}

Alongside G'_{ST} and G''_{ST} there is another alternative measure of differentiation, defined by Jost (2008). Jost pointed out that the dependency of G_{ST} and F_{ST} on H_{S} was not a consequence of statistical sampling, but holds true even when the actual population allele frequencies are used in the calculations.

Jost also argued that Nei's differentiation estimate D_{ST} is not independent of H_{S} and thus the total heterozygosity H_{T} cannot simply be additively partitioned into D_{ST} and H_{S}. Moreover, G_{ST} may be unsuitable as a measure of differentiation for highly polymorphic loci, and under some conditions G_{ST} decreases as the amount of differentiation between allele frequencies in the populations increases.

Finally, Jost considered it a misconception to equate heterozygosity with the "intuitive concept of diversity as actually used by geneticists, ecologists, and other scientists," because there are scenarios in which the populations share no alleles but the value of H_{S} is still close to the value of H_{T}, and thus the estimates of measures such as G_{ST} would be low.

Jost's definition of fully differentiated populations is populations that share no alleles, and his measure D is developed from the idea that if populations share no alleles, then the overall diversity of the meta-population should be the sum of the diversities of the individual populations. Jost defines the gene identity as in Nei (1973), given in (1.1), and likewise defines J_{S} as the gene identity averaged over all populations r and J_{T} as the meta-population gene identity in (1.2). From these he obtains Δ_{ST}, the effective number of distinct populations:

$$\Delta_{\text{ST}} = \frac{J_{\text{S}}}{J_{\text{T}}},$$

and then scales Δ_{ST} to obtain a relative measure of differentiation, D, that ranges from 0 to 1.

Jost also defines D_{est}, a "nearly unbiased" estimator of D, which is constructed using Nei and Chesser's (1983) unbiased estimators of H_{T} and H_{S}:

$$D_{\text{est}} = \frac{H_{\text{T\_est}} - H_{\text{S\_est}}}{1 - H_{\text{S\_est}}} \frac{K}{K-1}$$

1.3.6 Comparisons of G_{ST} and D

Following Jost (2008) there has been much discussion in the literature about the various merits and flaws of G_{ST} and D.

Heller & Siegismund (2009) evaluated G_{ST} and D_{est} using a meta-analysis of 34 earlier genetic studies on 43 species, published in *Molecular Ecology*. They calculated G_{ST} and D_{est} for each, though they used a slightly different version of unbiased estimator for H_{T} than the one used in Jost (2008). They confirmed that G_{ST} tends to underestimate the level of differentiation and overall they support Jost's findings about G_{ST} and recommend the use of either G'_{ST} or D_{est}.

Ryman & Leimar (2009) argued in favour of G_{ST}, showing that D is not independent of H_{S} as claimed by Jost (2008), and that D is also more strongly affected by the mutation rate than G_{ST} is. They demonstrated this using simulations with varying mutation rates and simulations with varying H_{S} and fixed mutation rate, and showed that D takes longer to reach its equilibrium value than G_{ST}. Their conclusion was that D "cannot be interpreted exclusively in terms of basic population genetics quantities such as population size and gene flow" because it confounds these with mutation rate.

In response to Heller & Siegismund (2009) and Ryman & Leimar (2009), Jost (2009) argued that G_{ST} is not logically consistent, unlike D. He provided a counterexample to Ryman & Leimar (2009) in which a set of populations with no migration between them have a constant value $D = 1$ but G_{ST} decreases over time as mutations increase the value of H_{S}. Jost dismisses the complaint about D taking a long time to reach equilibrium, commenting that "D, like G_{ST}, cannot be used to estimate any parameters of the finite island model unless the populations are in equilibrium" but that D is nonetheless useful as a measure of differentiation and, over short timescales, can also be used as a genetic distance.

Gerlach et al. (2010) developed the `DEMEtics` package for the R language as part of an evaluation of G_{ST} and D. They simulated 2 populations, starting from full differentiation with no migration, and gradually increased the level of migration. They found that D always took the value 1 when the populations shared no alleles, but the maximum value of G_{ST} was affected by the number of alleles per population. They constructed simulation cases with different values of G_{ST} but the same value of D, and cases with the same value of G_{ST} but different values of D. They also found cases with negative values of D_{est}. Finally, they found that the results of significance tests on the bias-corrected form of G_{ST} were sometimes conclusive even for very low values of bias-corrected G_{ST}, but bias-corrected D_{est} was still more useful as it displayed a wider range of values.

Meirmans & Hedrick (2011) commented that mutation affects all the existing measures of differentiation, and that even measures that attempt to account for mutation such as Φ_{ST} and R_{ST} are still only applicable in cases where the mutation patterns actually match the model used. They also commented that F'_{ST} is generally affected by population size and migration rate, whereas D is affected by the migration rate and the number of populations. They provided an example with fixation of alleles in the populations to demonstrate that F_{ST} and F'_{ST} are better for assessing demographic events, whereas D is better for measuring the actual differentiation in allele frequencies. They considered G''_{ST} to be an estimator for F'_{ST}.

Further contributions to the discussion came from Leng & Zhang (2011) and Alcala et al. (2014).

Whitlock (2011) argued in favour of F_{ST} rather than G'_{ST} or D, particularly when using a coalescent approach to estimate F_{ST} as a description of the relative time to the most recent common ancestor (Slatkin 1995). In scenarios with high migration and low mutation F_{ST} or G_{ST} may be more suitable than G'_{ST} or D, because the former are less affected by heterozygosity when there is low mutation and G'_{ST} and D may overestimate the amount of differentiation. In scenarios with low migration and high mutation the coalescent estimate of F_{ST} may still be more suitable than G'_{ST} or D because those two measures are dominated by the effects of mutation.

Whitlock argued that F_{ST} estimates a quantity of direct biological interest, unlike G'_{ST} and D, particularly since Jost defines D and D_{est} as single-locus measures, without providing a method of averaging over loci. We note that three packages for the R language, `diveRsity` (Keenan et al. 2013), `DEMEtics` (Gerlach et al. 2013) and `mmod` (Winter 2012), each use different methods of calculating the global value of D. G'_{ST}, G''_{ST} and D can also be calculated using software such as RecodeData (Meirmans 2006), SMOGD (Crawford 2010), and GenAlEx 6.5 (Peakall & Smouse, 2006 and 2012).

1.4 Thesis aims

The aim of this thesis is to provide new methodology focused on the interpretability of the output. We propose a new framework that, for the first time, produces visualizations that are closely linked to numeric measures and subsequent inference. Using this framework, we can reveal fine details of population structure not shown by any of the existing methods. The original motivation for the new methodology was to elicit the source of invasive individuals on sanctuary islands, but we found the same methods to be informative about genetic structure in a much wider range of scenarios.

The material in Chapter 2 and Appendices A–G has been published in *Biometrics* as McMillan & Fewster (2017) and supplementary materials. In the future, we intend to publish an abridged form of the material in Chapter 3.

Chapter 2

Visualizations for genetic assignment analyses using the saddlepoint approximation method

2.1 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 showing the degree of similarity to each population, 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 drifted 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 drifted 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.

2.2 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.

2.2.1 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 $0 \leq p_i \leq 1$ 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

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

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

$$f(\boldsymbol{p}; \tau) = \frac{\Gamma(\tau k)}{\Gamma(\tau)^k} \prod_{i=1}^{k} p_i^{\tau - 1}, \text{ for } \boldsymbol{p} \in [0, 1]^k; \sum_{i=1}^{k} p_i = 1. \tag{2.1}$$

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

$$[\,\boldsymbol{p} \,|\, \boldsymbol{X} = \boldsymbol{x}] \sim \text{Dirichlet}(x_1 + \tau,\, x_2 + \tau,\, \ldots,\, x_k + \tau), \tag{2.2}$$

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.

2.2.2 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:

$$\mathbb{P}(\boldsymbol{a} \,|\, \boldsymbol{X} = \boldsymbol{x}) = \int_{\boldsymbol{p}} \mathbb{P}(\boldsymbol{a} \,|\, \boldsymbol{p}\,;\boldsymbol{x}) \pi(\boldsymbol{p} \,|\, \boldsymbol{x}) d\boldsymbol{p}\,, \tag{2.3}$$

where $\pi(\boldsymbol{p} \,|\, \boldsymbol{x})$ is the posterior density of allele frequencies for population R given by (2.2).

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 (2.3) simplifies to:

$$\mathbb{P}(\boldsymbol{a}) = \begin{cases} \dfrac{(x_r + \tau)(x_r + \tau + 1)}{(2n_R + k\tau)(2n_R + k\tau + 1)} & a_r = 2;\, a_j = 0 \text{ for } j \neq r; \\ \dfrac{2(x_r + \tau)(x_s + \tau)}{(2n_R + k\tau)(2n_R + k\tau + 1)} & a_r = a_s = 1; \\ & a_j = 0 \text{ for } j \neq r, s. \end{cases} \tag{2.4}$$

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 (2.4) over all loci $L = 1, 2, \ldots, N_L$. Taking logs gives the log genotype probability (LGP) for individual I in population R:

$$\mathrm{LGP}_I^R = \sum_{L=1}^{N_L} \log\{\, \mathbb{P}(\boldsymbol{a}^L)\, \}, \tag{2.5}$$

where $\log\{\, \mathbb{P}(\boldsymbol{a}^L)\}$ is given by (2.4) based on sampled alleles $\boldsymbol{X}^L$ from population R for each locus $L = 1, 2, \ldots, N_L$. Adding the terms in (2.5) 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 Appendix G.

The overall LGP given by (2.5), 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.

The LGPs of individual I with respect to different populations can be compared to determine which population the individual has better fit to. If just two populations, R and S, are being compared, then the difference between the LGPs is the log of the likelihood ratio for the two populations. A potential assignment method would be to assign individual I to population R rather than population S if the log-likelihood ratio for R and S is above a certain threshold; however, in some cases this would not be a suitable method (see Section 2.3).

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.

2.3 Visualization of assignment results

We now outline our novel approach to assignment visualization. From (2.5) 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 2.1, which we henceforth call a GenePlot, shows individuals with complete data, sampled from two reference populations.

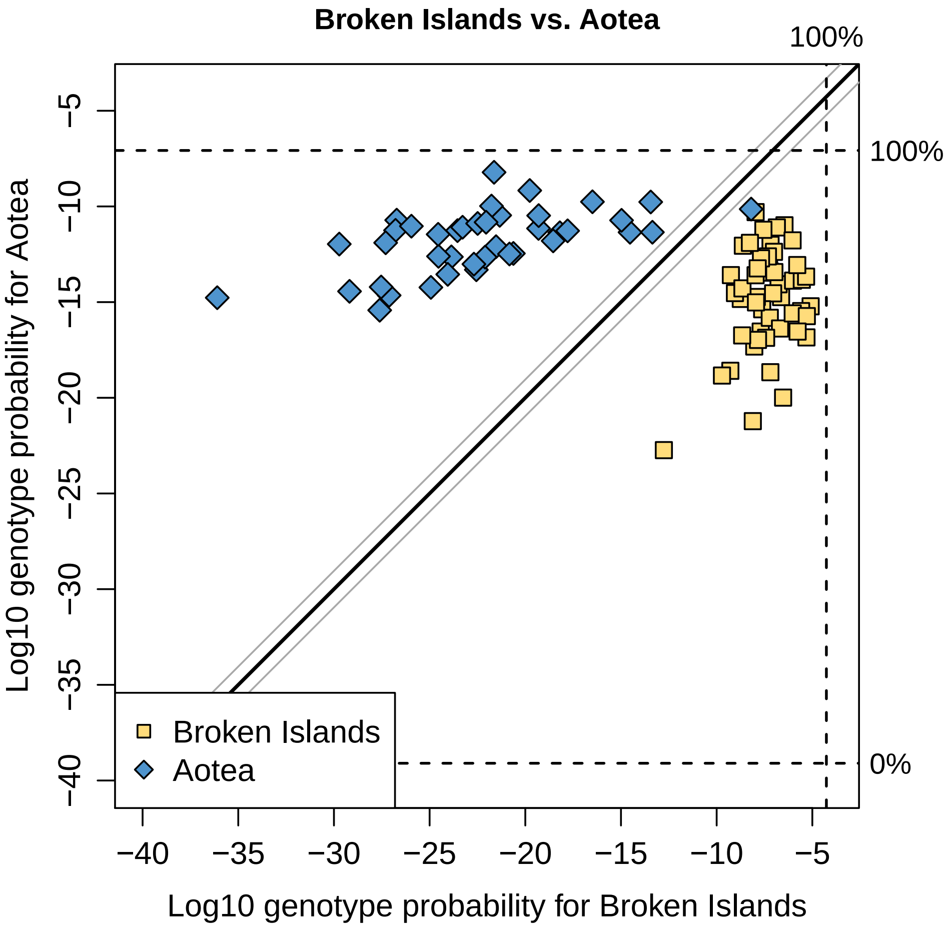


Figure 2.1: GenePlot based on microsatellite data extracted from ship rats (*Rattus rattus*). Each point represents an individual rat. Rats with missing data (i.e. data at fewer than 10 loci) are excluded. The graph shows rats captured on Aotea (diamonds) and the Broken Islands (squares). The horizontal 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.

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 (2.1) with $\tau = 1/k$.

The data in Figure 2.1 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 Appendix A.

The solid diagonal line in Figure 2.1 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 2.1 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 (2.4) for known $x_1, x_2, \ldots, x_k$ (Appendix B).

Figure 2.1 can be used to assess population structure: it suggests that the Broken Islands population may be a drifted 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 2.4. 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 2.1 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 2.1 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 Figure A.1 for a map of sampling locations. The display in Figure 2.1 is consistent with the suggestion that the Broken Islands population was founded from Aotea, retaining only a subset of the alleles in the larger Aotea population, and has then drifted relative to that population.

Figure 2.1 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 2.1, 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 Table E.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 2.1. For example, any individual on the lower thin diagonal line in Figure 2.1 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 (2.5), 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.

2.4 Quantile method for plotting individuals with missing data

Let $\mathcal{L} = \{1, 2, \dots, N_L\}$ 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 (2.5) for $\mathcal{L}_I$. We cannot simply rescale the LGP based on $\mathcal{L}_I$ to fit the scale of $\mathcal{L}$, because the loci that are missing may have very different allele frequency patterns than $\mathcal{L}_I$. We want to include information on those loci based on samples that have data at those loci, and we additionally wish 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:

$$\widetilde{\mathrm{LGP}}_I^R = Q_{\mathcal{L}}\left[F_{\mathcal{L}_I}\left\{\, \mathrm{LGP}(\mathcal{L}_I)\,\right\}\right]. \tag{2.6}$$

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 $\mathcal{L}$.

2.4.1 Characterizing the posterior LGP distribution

The multilocus posterior LGP in (2.5) 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 (2.4), 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.

2.4.2 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 2.1 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.

2.4.3 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, $K(t) = \log\{M(t)\} = \log\big[\,\mathbb{E}\{\exp(tY)\}\,\big]$, 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:

$$\hat{F}(y) = \begin{cases} \Phi(\hat{w}) + \phi(\hat{w})\left(\dfrac{1}{\hat{w}} - \dfrac{1}{\hat{v}}\right) & \text{if } y \neq \mu, \\ \dfrac{1}{2} + \dfrac{K'''(0)}{6\sqrt{2\pi}\,K''(0)^{3/2}} & \text{if } y = \mu, \end{cases} \tag{2.7}$$

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

$$\hat{w} = \text{sign}(\hat{s})\sqrt{2\{\hat{s}y - K(\hat{s})\}}\,, \tag{2.8}$$

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

$$K'(\hat{s}) = y. \tag{2.9}$$

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 (2.7) to (2.9) 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 (2.7) 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 (2.9), to obtain the corresponding value for $\hat{s}$; this incurs a small computational cost. Additionally, it is simple to solve (2.9) 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 (2.8), giving

$$\hat{w} = \text{sign}(\hat{s})\sqrt{2\{\hat{s}K'(\hat{s}) - K(\hat{s})\}}. \tag{2.10}$$

The inaccuracy caused by using (2.10) instead of (2.8) is negligible, particularly when compared with the numerical instability that can arise from (2.8).

Using this method gives $\hat{F}_{\mathcal{L}^*}$ for any $\mathcal{L}^* \subseteq \mathcal{L}$, as required in (2.6): see 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.

2.4.4 Evaluation of the saddlepoint approximation

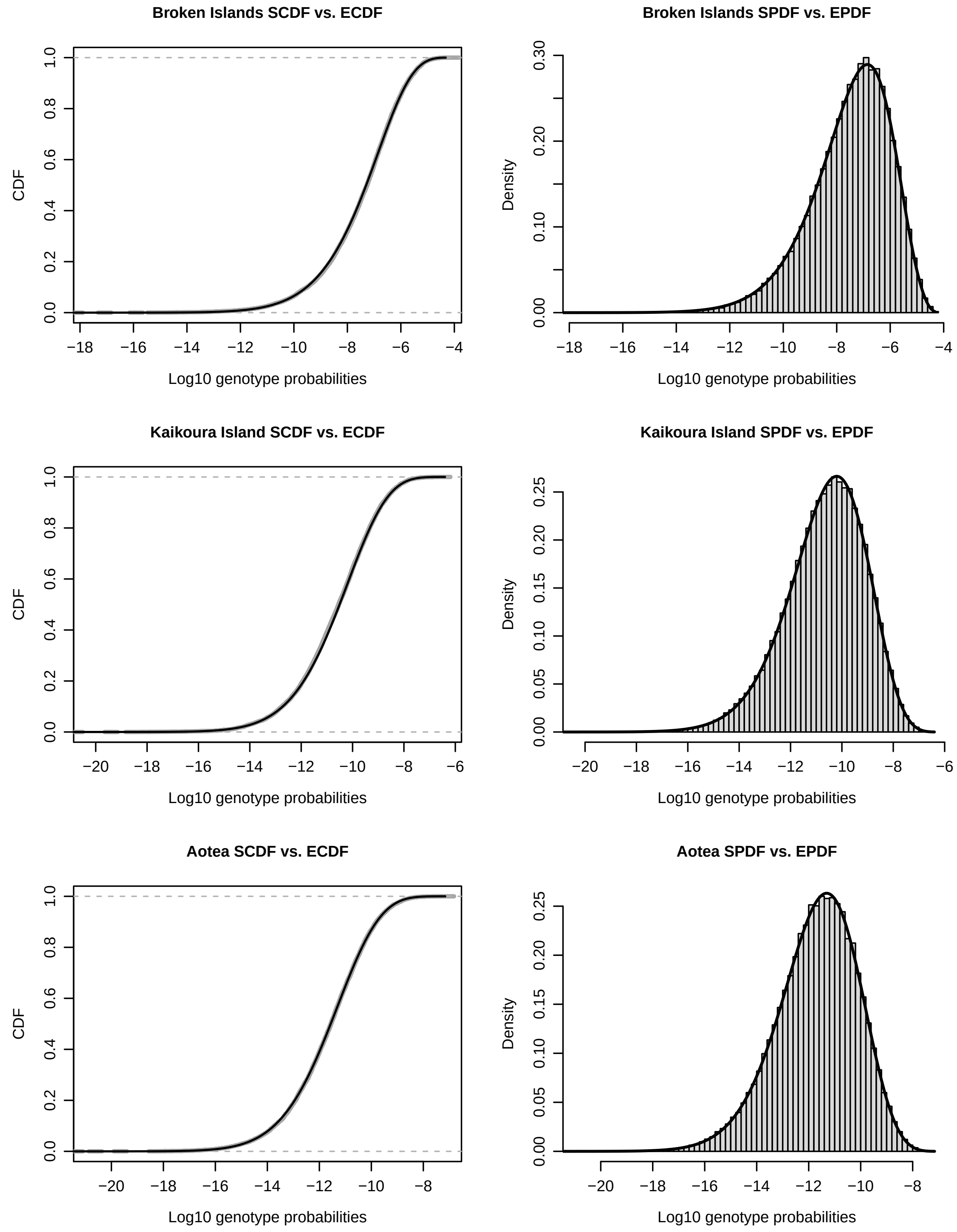
Figure 2.2 shows the application of the saddlepoint approximation to posterior LGP distributions for the ship rat data described above, with the addition of ship rats captured on Kaikoura Island during the same period. Kaikoura (530 ha) lies about 80m from Aotea at closest approach and 3 km north of the Broken Islands (Figure A.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 2.4.2.

The saddlepoint method, without any tuning, adapts to the different shapes of the distributions for the three populations in Figure 2.2 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.

The top row in Figure 2.3 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 2.3 are typical of those we have seen in multiple other data sets including ship rats and Atlantic salmon (*Salmo salar*).

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 2.3, 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 2.3 illustrates a worst-case scenario, since five loci is generally considered to be too


Broken Islands SCDF vs. ECDF

Broken Islands SPDF vs. EPDF

Kaikoura Island SCDF vs. ECDF

Kaikoura Island SPDF vs. EPDF

Aotea SCDF vs. ECDF

Aotea SPDF vs. EPDF

Figure 2.2: 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.

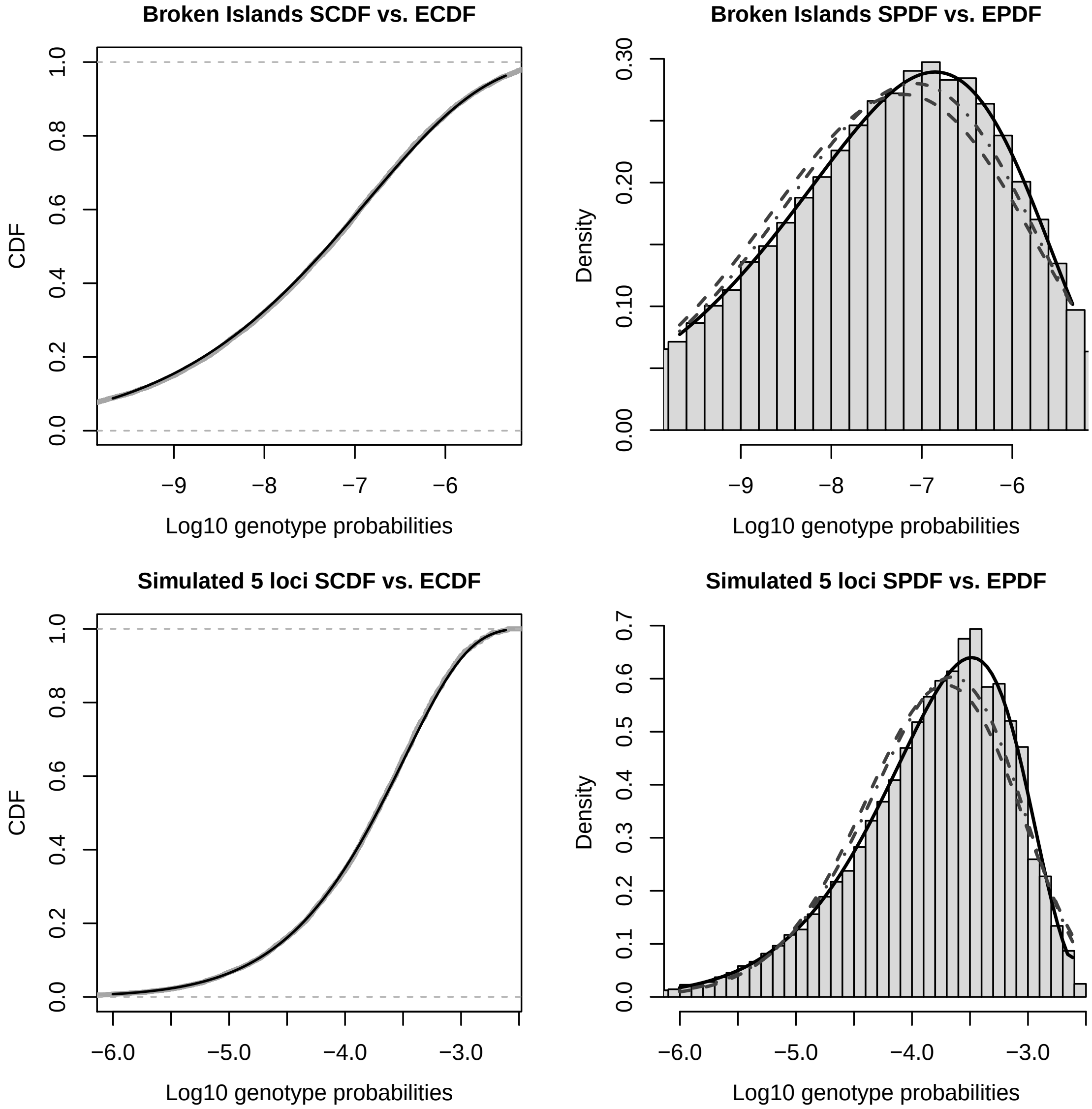


Figure 2.3: CDFs and PDFs of the LGP distributions of two populations, zoomed in to focus on an area around the mean. Details are as for Figure 2.2, with the addition of moment-based gamma and lognormal approximations shown as dashed and dot-dashed lines on the right-hand plots. The top row shows the 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.

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.

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.

2.5 Applications

We now demonstrate how our approach of Sections 2.3 and 2.4 can be used both for visualizing population structure and performing assignment. Figures 2.4-2.6 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 2.4 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 2.4. Calculations using the alternative simulation-based quantile approximations in Section 2.4.2 inflated the LGP estimates for many of these missing-data rats, distorting the overall picture of population structure.

2.5.1 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 2.4. Figure 2.4 shows the 1% quantile lines for the Broken Islands, Kaikoura and Aotea populations. A comparison with Figure 2.1 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 2.4 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 drifted 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 2.4 that the Broken Islands and Aotea populations are easily distinguished genetically, and the Broken Islands population is a

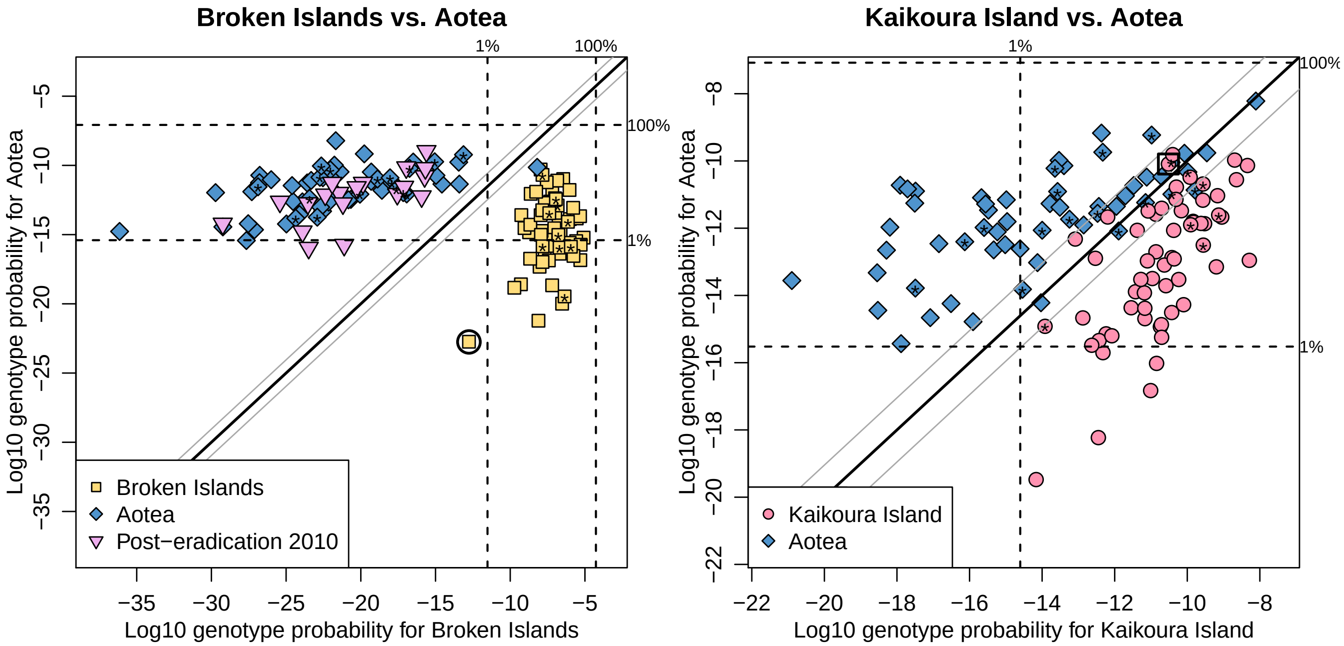


Figure 2.4: GenePlots based on microsatellite data extracted from ship rats. Each point represents an individual rat. Rats marked with asterisks have missing data at one or more loci; rats with data at fewer than 6 loci have been excluded. The left plot shows rats sampled from the Broken Islands and 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 Appendix E.

drifted 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. All of them fit well with the Aotea profile, but not with the Broken Islands profile. 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 2.4 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.

2.5.2 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 2.5. 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 2.5 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.

2.5.3 Comparison with STRUCTURE, GENECLASS2 and DAPC

The bar charts in Figure 2.5 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 2.5 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

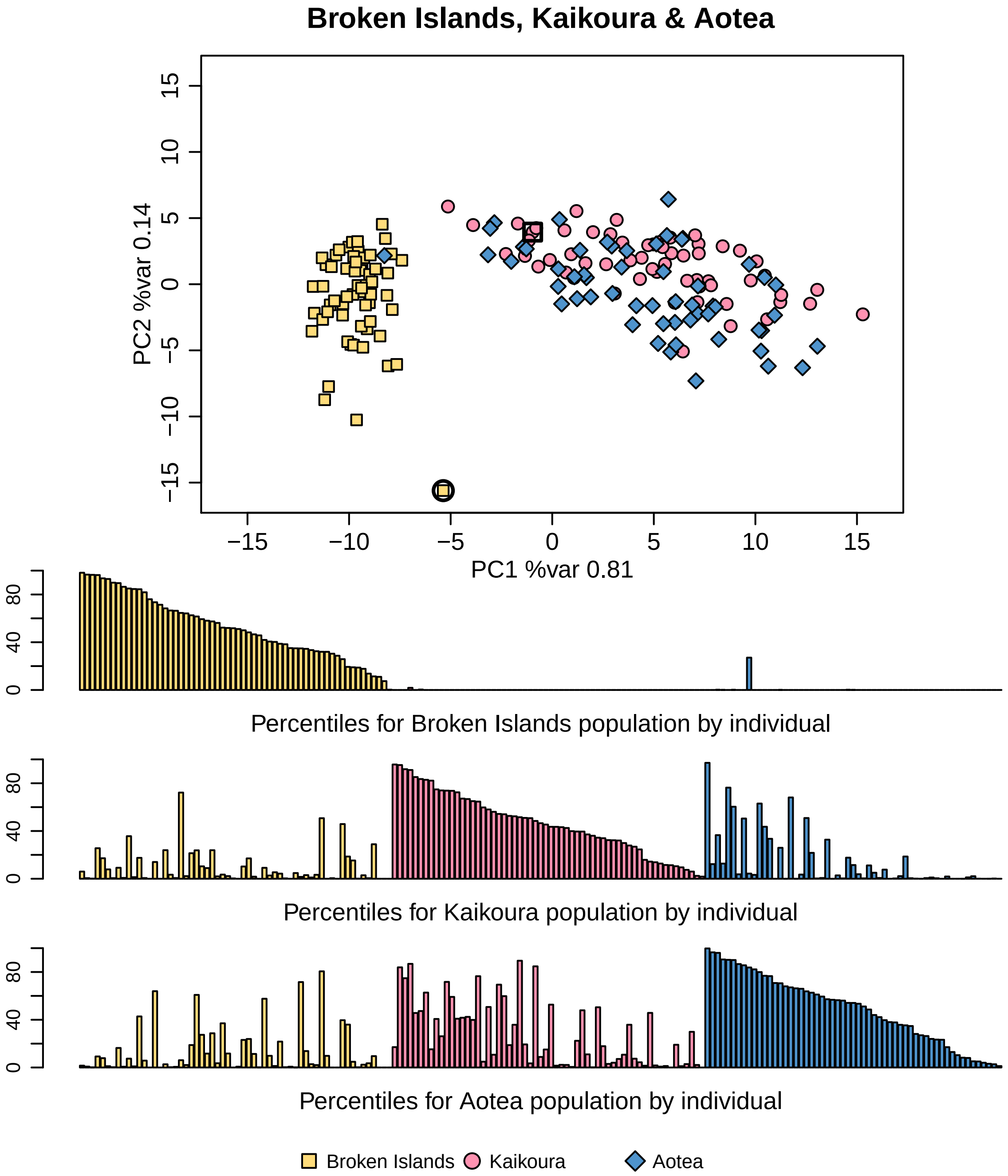


Figure 2.5: Visualizations for multiple reference populations. Individuals are ship rats captured on Kaikoura, the Broken Islands and Aotea between 2005 and 2008. Rats with data at fewer than 6 loci have been excluded. The top plot is a GenePlot for multiple reference populations: the axes show the first 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, and are sorted according to their fit to their own population. 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 Appendix E.

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. 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 2.4 and 2.5 appears, in STRUCTURE, to be well fitted to the Broken Islands (cluster 1), but Figure 2.4 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.

2.5.4 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 2.6 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 distri-

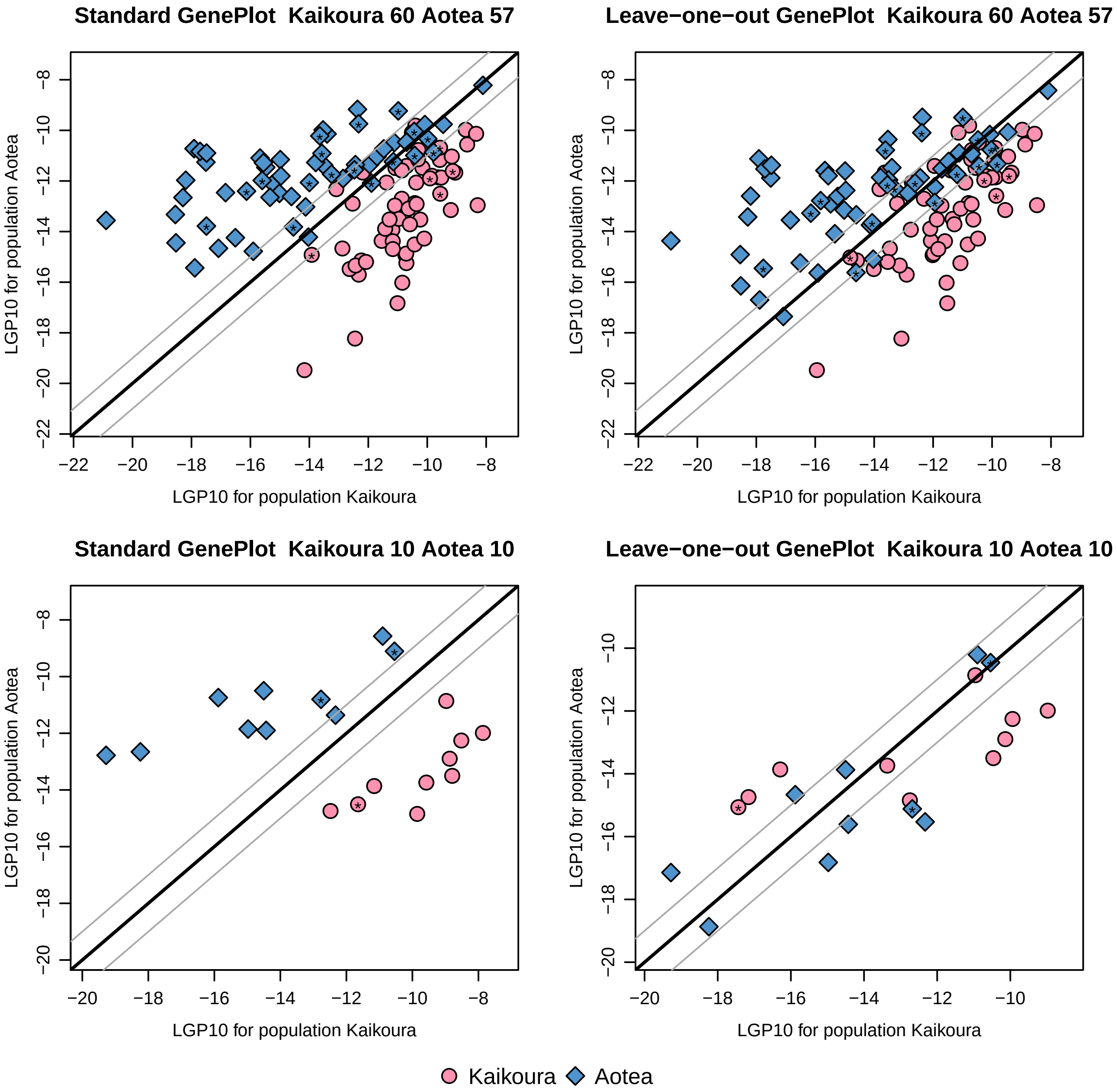


Figure 2.6: Leave-one-out GenePlots using random subsets from samples of rats captured on Kaikoura and Aotea. Rats marked with asterisks have missing data at one or more loci; rats with data at fewer than 6 loci have been excluded. The left column plots are standard GenePlots; the right-column plots 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.

bution 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.

It should be noted that the leave-one-out plots in Chapter 3 use a slightly different form of leave-one-out than the one used here and in McMillan & Fewster (2017). Please see Appendix H.5 for details.

2.5.5 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 Appendix F.

2.6 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 (2.10) 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, drifted 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/`.

Appendices for Chapter 2

Appendix A Map of Great Barrier Island

The data used to demonstrate the visualization method are from ship rats captured on the large island of Aotea/Great Barrier Island, and on Kaikoura Island and the Broken Islands, smaller landmasses off the coast of Aotea. Figure A.1 shows all the locations where samples were captured. The three islands labeled as Motutaiko, Flat and Mahuki are the Broken Islands and contribute 62 samples. Aotea samples used throughout the main text comprise the 58 samples from western Aotea, sampled at locations marked Fitzroy, Red Cliffs, and Mainland. Sample sizes quoted in the text differ slightly from sample numbers marked on the map because some samples had missing data at five or more loci and were excluded from analyses.

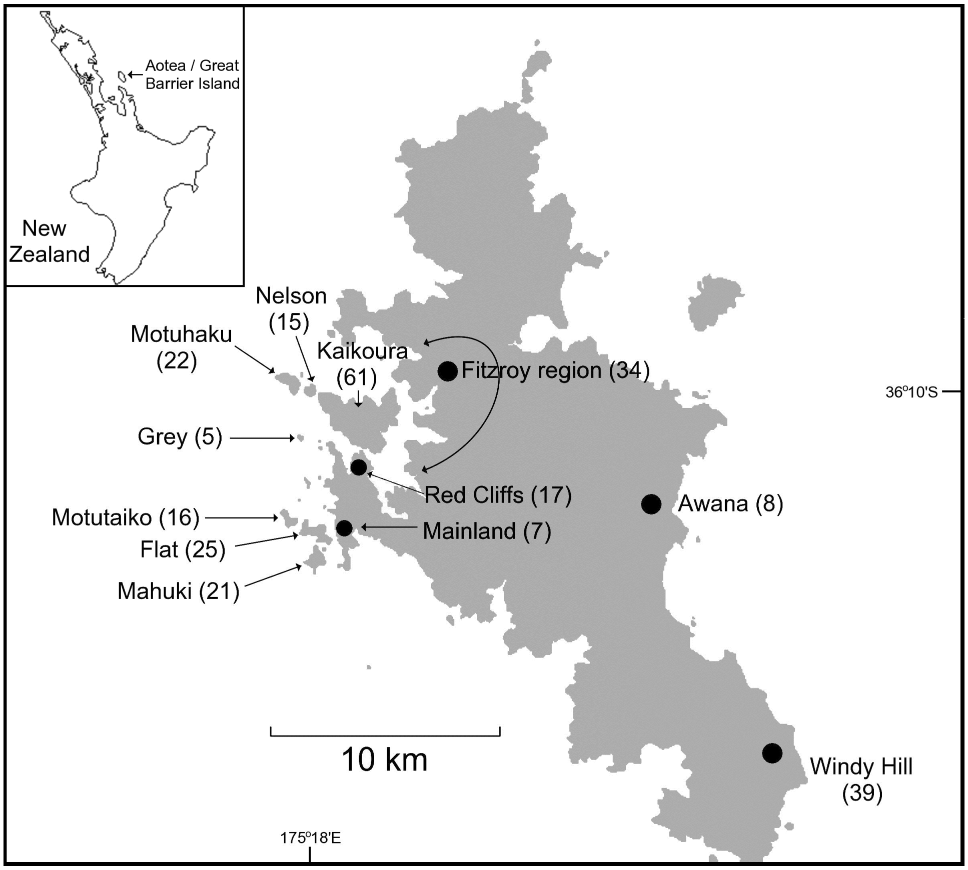


Figure A.1: Rat sampling locations on the Great Barrier Island archipelago. Sample sizes are shown in brackets.

Appendix B Minimum and maximum of the LGP distribution

The distribution by which we characterize a population is the distribution of log-genotype probabilities (LGPs) for all possible multilocus genotypes that may be drawn from the posterior distribution of the allele frequencies for that population.

The posterior distribution of allele frequencies at a single locus is given by

$$\boldsymbol{p}|\boldsymbol{x} \sim \text{Dirichlet}(x_1 + \tau, x_2 + \tau, \ldots, x_k + \tau)$$

where x_i is the observed frequency of allele i in the sample from population R and τ is the parameter for the Dirichlet prior distribution of allele frequencies.

Given the posterior allele frequencies, we can calculate the minimum and maximum of the LGP distribution. We initially calculate the maximum and minimum at each single locus and define $\nu_i = x_i + \tau$ and $N = \sum_{i=1}^{k} \nu_i$. For a single locus, the maximum possible LGP is given by:

$$\begin{cases} \log\left[\dfrac{\nu_1(\nu_1+1)}{N(N+1)}\right] & \text{if } \nu_1 + 1 \geq 2\nu_2\,, \\[2ex] \log\left[\dfrac{2\nu_1\nu_2}{N(N+1)}\right] & \text{otherwise,} \end{cases}$$

where ν_1 is the largest of all the ν_i, and ν_2 is the next largest (possibly equal).

The minimum possible LGP is given by:

$$\begin{cases} \log\left[\dfrac{2\nu_k\nu_{k-1}}{N(N+1)}\right] & \text{if } 2\nu_{k-1} \leq \nu_k + 1\,, \\[2ex] \log\left[\dfrac{\nu_k(\nu_k+1)}{N(N+1)}\right] & \text{otherwise,} \end{cases}$$

where ν_k is the smallest of all the ν_i, and ν_{k-1} is the next smallest (possibly equal).

The maximum and minimum of the multilocus LGP distribution are found by summing over all the loci.

Appendix C Saddlepoint approximation as applied to the LGP distribution

For a single locus L, any genotype $\boldsymbol{a}^L$ arises with probability $\mathbb{P}(\boldsymbol{a}^L)$ given by equation (2.4). Thus the single-locus LGP distribution acquires point mass $\mathbb{P}(\boldsymbol{a}^L)$ at value $\log\{\mathbb{P}(\boldsymbol{a}^L)\}$ for every genotype $\boldsymbol{a}^L$. Recall that $\log\{\mathbb{P}(\boldsymbol{a}^L)\}$ serves as a measure of genetic fit and we are aiming to characterize the probability distribution of this measure.

Let $Y^L = \log\{\mathbb{P}(\boldsymbol{a}^L)\}$ be the random variable representing the single-locus LGP, and let genotypes at locus L be indexed by $g = 1, 2, \ldots, k^L(k^L+1)/2$ where k^L is the number of allele types at locus L. Write $\alpha_g = \mathbb{P}(\boldsymbol{a}_g^L)$ for each g. Then each genotype $\boldsymbol{a}_g^L$ contributes point mass α_g to value $Y^L = \log(\alpha_g)$.

The moment generating function (MGF) for a random variable Z is given by:

$$M(t) = E(e^{tZ})\,.$$

The MGF of the target LGP distribution at a single locus L is thus given by:

$$\begin{aligned} M(t) &= \sum_y P(Y^L = y)\exp(ty) \\ &= \sum_g \alpha_g \exp(t\log\alpha_g) \\ &= \sum_g \alpha_g^{t+1}\,. \end{aligned}$$

Derivatives of the MGF are given by:

$$M^{(r)}(t) = \sum_g (\log\alpha_g)^r \alpha_g^{t+1}\,.$$

The cumulant generating function (CGF) and its derivatives for a single locus can be calculated accordingly. Since the CGF is the log of the MGF, and genotypes are assumed to be independent across loci, the multilocus CGF, K, is the sum of the single locus CGFs, and the derivatives are similarly defined.

The saddlepoint approximation to the multilocus LGP distribution can then be calculated, using a numerical method to find the root of the equation

$$K'(\hat{s}) = y\,,$$

where y is the value of the random variable Y representing the multilocus LGP at which we wish to calculate $F_{\mathcal{L}^*}(y)$.

Technically, the LGP distribution is discrete. However, in most cases it is well-approximated by a continuous distribution, particularly when the number of distinct allele types at each locus is higher than about 4, due to the large number of possible genotype probabilities. The saddlepoint approximation is an approximation to the (undefined) continuous distribution that, in turn, approximates the discrete target LGP distribution.

Appendix D Evaluation of saddlepoint CDF and empirical CDF

We tested the saddlepoint distribution using many simulated populations. For a given simulated population P we calculated the SCDF and generated an ECDF based on 100,000 simulated genotypes, then calculated the values of both the ECDF and the SCDF at 1000 evenly spaced points over the range of the ECDF. We then took the result s_P to be the sum of squared differences between the SCDF and ECDF at those 1000 points. Since different sets of simulated genotypes can produce different ECDFs for the same population, we repeated the procedure 10 times with different sets of simulated genotypes to obtain S_P as the mean of s_P over the 10 replicate ECDFs for population P. Thus S_P is the mean sum of squared differences for a single population with a single set of randomly simulated allele frequencies.

We then simulated many populations. Figure D.2 shows S_P for populations generated with $N_L = 2, \dots, 10$, where the maximum number of allele types k ranges from 2 to 10, with 10 population replicates for each combination of k and N_L constituting 10 different draws of allele frequencies as described in the main text. Within each of those population replicates we generated 10 replicate ECDFs as described above, and summarized the ECDF results as S_P for each population replicate. Figure D.2 illustrates that the SCDF provides a suitable approximation to the LGP distribution for a population.

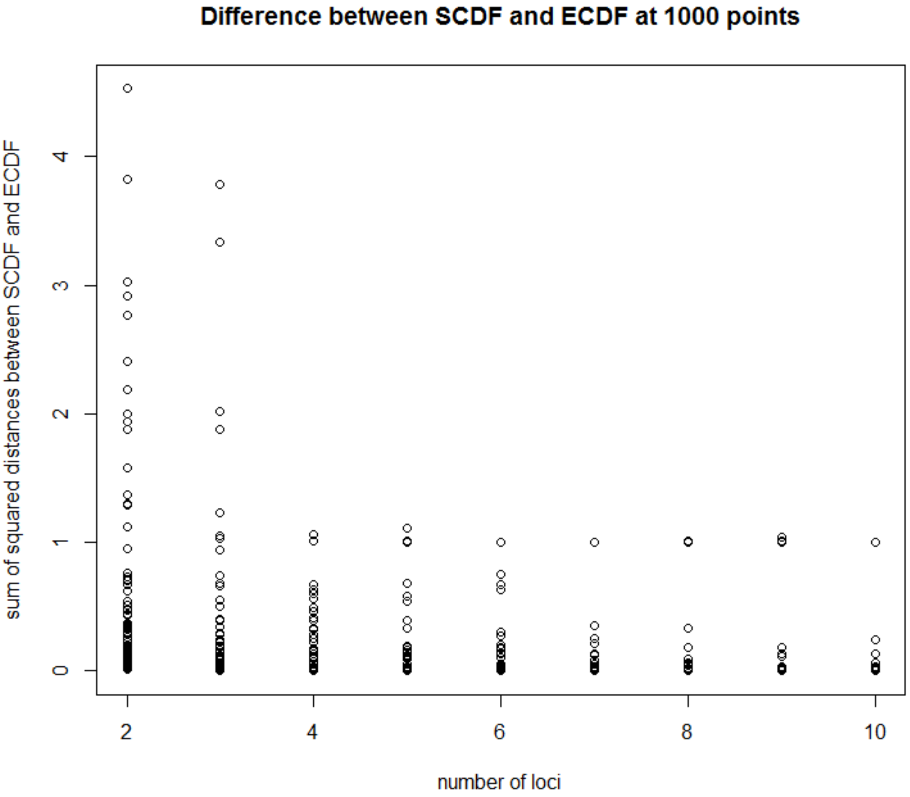


Figure D.2: SCDF vs ECDF at 1000 points for simulated populations. The x-axis shows the number of loci N_L for each simulated population. The y-axis shows the discrepancy measure S_P for each of 90 results at each value of N_L, denoting 10 replicates for each setting of the maximum number of alleles $k = 2, \ldots, 10$.

Appendix E Comparison of GenePlot, STRUCTURE, GENECLASS2 and DAPC

We provide a comparison of our GenePlot method with the popular software packages STRUCTURE and GENECLASS2 by displaying the output of each program applied to the same dataset. We also provide a comparison of GenePlot with the Discriminant Analysis of Principal Components (DAPC) functionality within the R package `adegenet`.

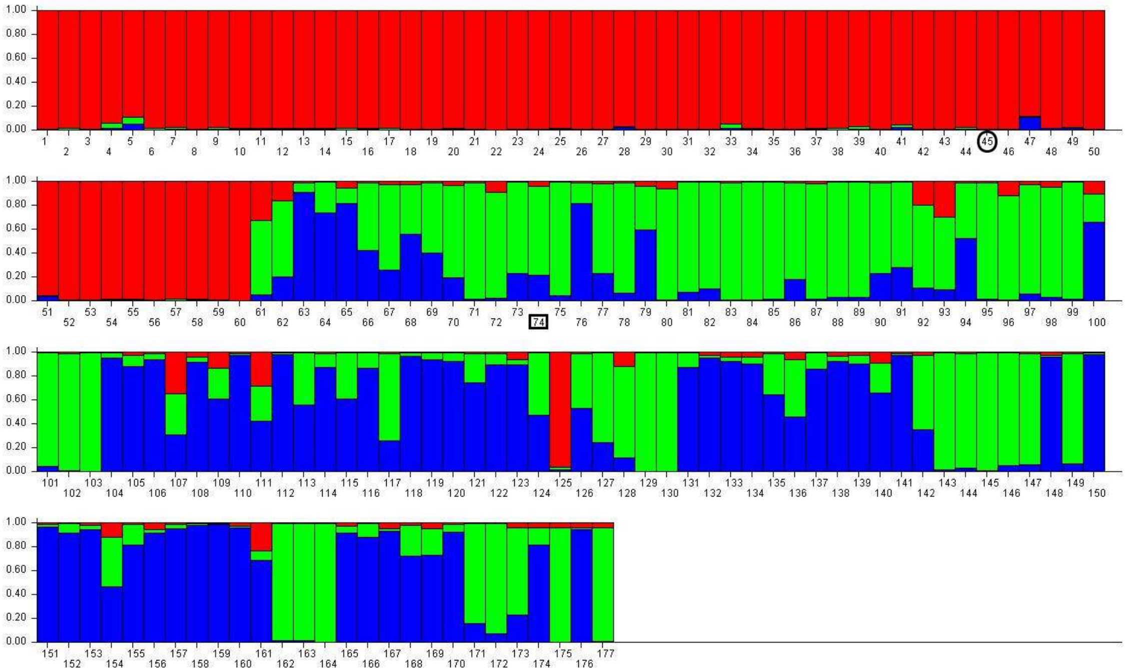


Figure E.3: STRUCTURE bar plot of ship rats captured on Kaikoura Island, the Broken Islands and Aotea between 2005 and 2008. The red bars correspond to cluster 1, the green bars to cluster 2 and the blue bars to cluster 3. Individuals 1 to 60 were captured on the Broken Islands; individuals 61 to 120 were captured on Kaikoura; individuals 121 to 177 were captured on Aotea. Cluster 1 mostly corresponds to the Broken Islands samples.

We ran STRUCTURE (Pritchard et al. 2000, Falush et al. 2003) with the admixture model and without location priors. We used options without correlated allele frequencies, with 10,000 burn-in iterations and 10,000 final iterations. We tested the number of clusters, K, from 1 to 3, running 5 replicates for each value of K. The runs used consecutive random number seeds starting at 1. We found $K = 3$ had the highest likelihood. Figure E.3 shows the results for the run with the highest likelihood using $K = 3$. STRUCTURE uses a clustering algorithm, and without location priors it does not use any location data to assign samples to clusters; instead, it allots each sample a fractional composition of the three estimated clusters. This fractional composition is intended to indicate the estimated proportion of the individual's genome that originated in each of the three clusters, but it is commonly interpreted in studies as the probability that the individual originated from each of the three clusters.

The STRUCTURE results in Figure E.3 indicate that there is one clearly-defined cluster, specifically Cluster 1 in red, and two somewhat overlapping clusters depicted by green and blue. The results do not reveal other information about genetic subsetting between clusters, or the variance of genetic fit within a cluster.

The sample marked with a circle corresponds to the rat circled in Figures 2.4 and 2.5; the sample marked with a square corresponds to the rat marked with a square in Figures 2.4 and 2.5. The circled rat stands out as somewhat anomalous in the GenePlots in Figures 2.4 and 2.5, but this is not apparent from the STRUCTURE output in Figure E.3.

We ran GENECLASS2 (Piry et al. 2004) on the Kaikoura, Broken Islands and Aotea reference samples, using the assign/exclude option for individuals, with assignment threshold 0.05. We used the Rannala and Mountain (1997) Bayesian computation method, without the probability computation from Paetkau et al. (2004). GENECLASS2 uses the leave-one-out method by default if no separate assignment samples are provided (query samples in our terminology), because it assumes that the reference samples are to be assigned. Thus, to supply assignment samples, we entered the reference samples again with renamed populations. This ensured GENECLASS2 displayed results consistent with the GenePlot shown in Figure 2.5, which does not display the leave-one-out method. Table E.1 shows a selection of the results to illustrate the GENECLASS2 output; figures are quoted to the precision returned by GENECLASS2. The full table (not shown) contains 177 rows. The rats marked with circle and square correspond to the rats marked with circle and square in Figures 2.4 and 2.5. When a sample possesses the full complement of 10 loci, the columns marked -log(L) correspond to the negative LGPs as plotted on the GenePlots in Figure 2.4.

The "Assigned sample" column in the GENECLASS2 results shows the population in which each sample was captured, renamed as Brok2, Kai2, and Aotea2 respectively to switch off the leave-one-out option as described above. Some individuals have higher scores for populations they were not found in. For example, the rat marked with a square in Table E.1 was found on Kaikoura. From Table E.1 there appears to be substantially stronger support for Aotea than for Kaikoura. However, Figure 2.4 demonstrates that, seen in context, the rat has a similar fit to both Aotea and Kaikoura, and lies within a region of high overlap between the two populations; its origins can therefore not be determined conclusively. This underlines the importance of viewing the absolute measure of fit shown in the GenePlot. Interpreting only the relative fit to different populations, in the absence of the context displayed on the GenePlot, can lead to misleading conclusions. The same rat is marked with a square on the STRUCTURE plot in Figure E.3. Although there is a strong mapping between Cluster 1 in the STRUCTURE output and rats sampled on the Broken Islands, there is no clear mapping between Clusters 2 and 3 and the sampling locations of Kaikoura and Aotea. The rat marked with a square exhibits a moderate signal of overlapping membership between Clusters 2 and 3.

We ran DAPC (Jombart et al. 2010) from the R package `adegenet` (Jombart 2008, Jombart & Ahmed 2011), a method that performs principal components analysis on raw diploid allele counts, and then performs discriminant analysis on a reduced number of the principal com-

Assigned sample	rank 1	score %	rank 2	score %	rank 3	score %	Brok -log(L)	Kai -log(L)	Aotea -log(L)	Nb. of loci
/Brok2	Brok	100	Kai	0	Aotea	0	8.041	15.224	17.351	10
/Brok2	Brok	99.999	Kai	0.001	Aotea	0	7.265	12.556	12.663	10
/Brok2	Brok	99.999	Kai	0.001	Aotea	0	6.636	11.58	14.787	10
/Brok2	Brok	99.2	Aotea	0.46	Kai	0.34	7.968	10.434	10.301	10
/Brok2 ○	Brok	100	Aotea	0	Kai	0	12.765	25.858	22.818	10
/Kai2 □	Aotea	72.848	Kai	27.151	Brok	0	15.391	10.52	10.091	10
/Kai2	Kai	88.056	Aotea	11.944	Brok	0	22.442	8.968	9.836	9
/Kai2	Kai	72.033	Aotea	27.967	Brok	0	26.611	11.079	11.49	10
/Aotea2	Aotea	99.796	Kai	0.204	Brok	0	23.98	15.336	12.647	10
/Aotea2	Aotea	99.649	Kai	0.351	Brok	0	27.841	17.888	15.435	10
/Aotea2	Aotea	99.963	Kai	0.037	Brok	0	20.609	14.641	11.205	8
/Aotea2	Kai	60.495	Aotea	39.505	Brok	0	27.868	14.034	14.219	10
/Aotea2	Brok	98.869	Aotea	1.13	Kai	0.001	8.199	13.399	10.141	10

Table E.1: Selected GENECLASS2 results for the Broken Islands, Kaikoura and Aotea populations. The rank and score columns show the populations ordered by their corresponding scores for each individual.

ponents. The input data format for DAPC is a table with one row per individual, and columns corresponding to every allele type at every locus. The cell values are 0, 1, or 2, corresponding to the number of alleles of each type that the individual possesses. A principal components analysis is performed on these allele-count variables with values 0, 1, and 2. The user then selects how many of these principal components to use for the subsequent discriminant analysis. We ran DAPC on the combined data from Kaikoura Island, the Broken Islands and Aotea using 50 principal components and two discriminant components. The number of principal components used for the discriminant analysis should not be larger than the smallest reference sample; 50 is the default number of principal components. After the first two discriminant components the remaining discriminant components did not significantly improve the results. Figure E.4 shows the probabilities of membership for each population and each rat, and selected individuals are shown in Table E.2; Figure E.5 shows the scatter graph of the first two discriminant components, with points labelled according to their original population membership. As for the STRUCTURE and GENECLASS2 results, two individuals are marked with a circle and a square, corresponding to the two marked individuals in Figure 2.5; the conclusions in Table E.2 for these two rats are almost identical to those from GENECLASS seen in Table E.1.

Figure E.6, the default DAPC plot when applied to only the two populations of Aotea and the Broken Islands, uses only one discriminant component as the other components do not significantly affect the result. We used only 30 principal components in the PCA stage as there are fewer observations than in the combined dataset with Kaikoura Island.

The GenePlot method has several advantages over DAPC. Figure E.5 shows the separation of the Broken Islands population and the overlap of the Aotea and Kaikoura Island populations, similarly to Figure 2.5, but it is unclear what conclusions should be drawn about individual rats. For example, the circled rat does not stand out relative to other Broken Island rats.

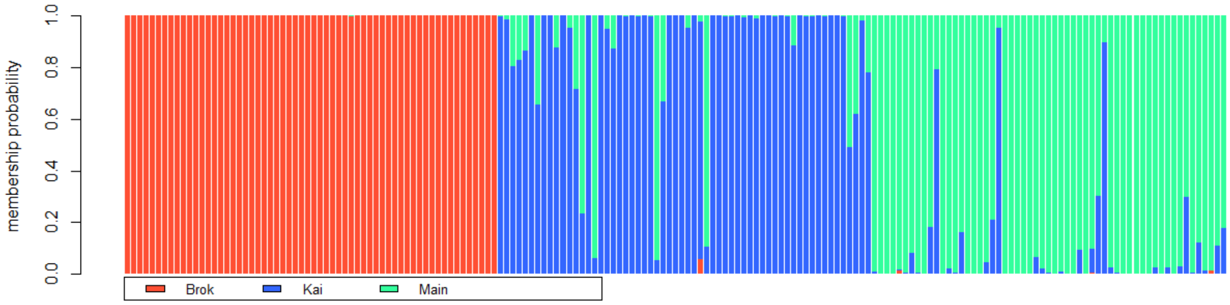


Figure E.4: Population membership graph based on DAPC results from `adegenet` of ship rats captured on Kaikoura Island, the Broken Islands and the main island Aotea between 2005 and 2008. Individuals 1 to 60 were captured on the Broken Islands; individuals 61 to 120 were captured on Kaikoura; individuals 121 to 177 were captured on Aotea.

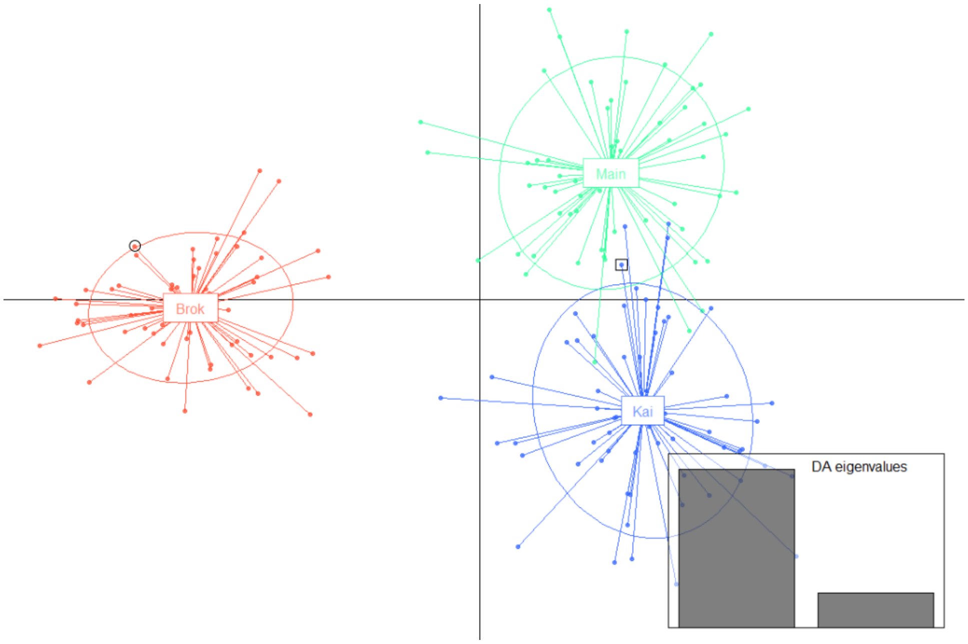


Figure E.5: Cluster plot based on DAPC results from `adegenet` of ship rats captured on Kaikoura Island, the Broken Islands and the main island Aotea between 2005 and 2008. The red cluster, on the left, comprises the rats found on the Broken Islands; the blue cluster comprises the rats found on Kaikoura Island; the green cluster comprises the rats found on Aotea. 50 principal components were used at the PCA stage of the DAPC process.

Figure E.6 gives no indication of the Broken Islands population being a subset of the Aotea population as was indicated by the GenePlot, Figure 2.4. The DAPC plots are also cryptic in interpretation: even in Figure E.6 we cannot see absolute measures of fit, and DAPC does not provide quantile lines akin to those shown in Figure 2.4. The rat sampled on Aotea that was shown in Figure 2.4 to be clustered with the Broken Islands population is not visible in Figures E.5 or E.6, so the procedure has failed to detect what was probably a direct migrant from the Broken Islands to Aotea. DAPC requires the user to choose the number of principal components to use for the discriminant analysis, and this involves not only a loss of information

ID	Brok	Kai	Aotea
Bi39	0.999	0.000	0.001
Bi40	1.000	0.000	0.000
Bi41	1.000	0.000	0.000
Bi49	1.000	0.000	0.000
Bi50	1.000	0.000	0.000
Bi53 ○	1.000	0.000	0.000
Ki015	0.000	1.000	0.000
Ki016	0.001	0.953	0.046
Ki017	0.000	0.717	0.283
Ki018 □	0.000	0.235	0.765
Ki020	0.000	1.000	0.000
Ki021	0.000	0.060	0.940

Table E.2: Selected DAPC results for the Broken Islands, Kaikoura and Aotea populations showing the estimated probability of membership of each cluster.

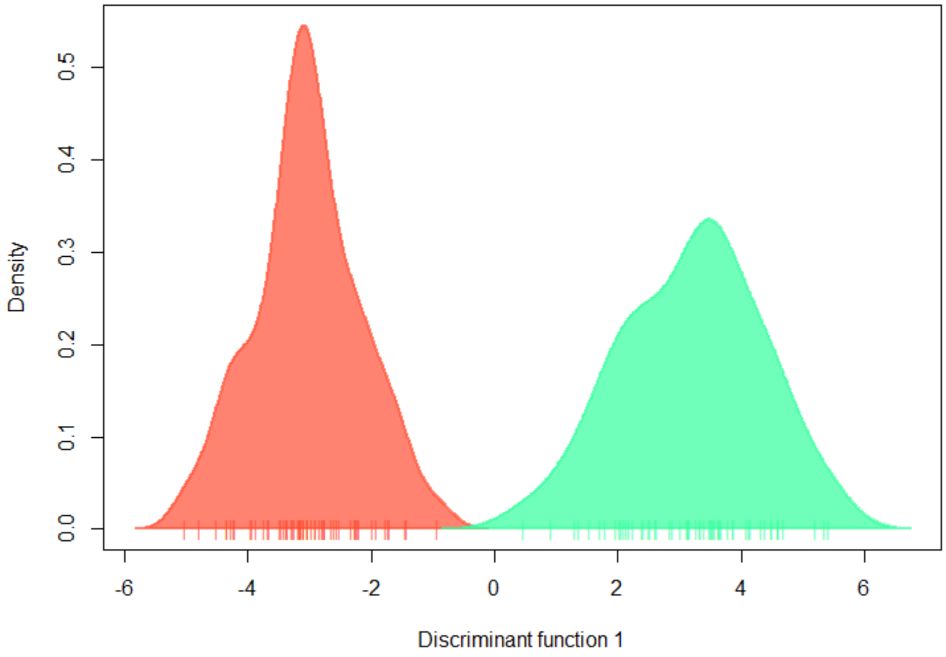


Figure E.6: Cluster plot based on DAPC results from `adegenet` of ship rats captured on the Broken Islands and Aotea between 2005 and 2008. The left, red peak comprises the rats found on the Broken Islands; the right, green peak comprises the rats found on Aotea. 30 principal components were used at the PCA stage of the process.

but also a trade-off between increasing the power to detect cryptic population structure and over-fitting the clusters so that the discriminant functions perform poorly on individuals. By contrast, the GenePlot method has the same settings for every run and uses all the information contained in the genetic data. Finally, for plots of two populations, the GenePlot method also provides additional information about population structure by showing quantile lines and by indicating whether one population is a subset of the other.

E.1 Comparison of GenePlot and DAPC with simulated data

We tested the classification accuracy for GenePlot against that of DAPC from the `adegenet` package using data simulated with Easypop (Balloux 2001). We ran 10 replicates of the hierarchical stepping stone model with 2 groups of 6 populations each, all of size 100, using random mating for 3000 generations, with a migration rate of 0.01 within the groups and 0.001 between groups. We used 30 loci, 50 possible allelic states, 0.0001 mutation rate and free recombination between loci. Figure E.7 shows an example replicate in GenePlot (leave-one-out mode) and Figure E.8 shows the same replicate after running DAPC. Table E.3 shows the misclassification rate for individual assignments. All individuals were assigned to the population for which they had the highest LGP or probability. In the case of missing-data individuals, we used $\widetilde{\mathrm{LGP}}_I^R$ as defined in Section 2.4 for the GenePlot assignment; we could alternatively use $\mathrm{LGP}_I^R(\mathcal{L}_I)$ for such individuals. Misclassified individuals are any individuals assigned to a population that is not their true population of origin. Although we favor assignment using the exclusion method defined in the main text, this is not available in DAPC, so we assigned all individuals to their single 'best' population. Despite DAPC being customized for classification, as opposed to the GenePlot method which is focused on calculating absolute measures of fit, Table E.3 shows that the GenePlot method produces a much lower classification error rate for this scenario, some 3-20 times lower than that of DAPC.

Additional assessment of DAPC and GenePlot results from many scenario simulations (as described in Appendix F) showed that whereas DAPC often shows populations as well-separated clusters, GenePlot gives a clearer view of when the populations are genuinely distinct and when they are poorly differentiated.

Replicate	DAPC errors		GenePlot errors	
	Count	Rate	Count	Rate
1	84	0.070	14	0.012
2	73	0.061	14	0.012
3	78	0.065	23	0.019
4	165	0.138	9	0.008
5	58	0.048	9	0.008
6	131	0.109	17	0.014
7	96	0.080	16	0.013
8	74	0.063	10	0.008
9	69	0.058	12	0.010
10	61	0.052	16	0.013

Table E.3: Misclassification results from DAPC and GenePlot for 10 replicates of data generated in Easypop with the hierarchical stepping stones model, using 2 groups of 6 islands (12 populations in total) of 100 individuals each for 3000 generations, with a migration rate of 0.001 within groups and 0.01 between groups. GenePlot was run in leave-one-out mode. Individuals are assigned to the population for which they have the highest probability or LGP.

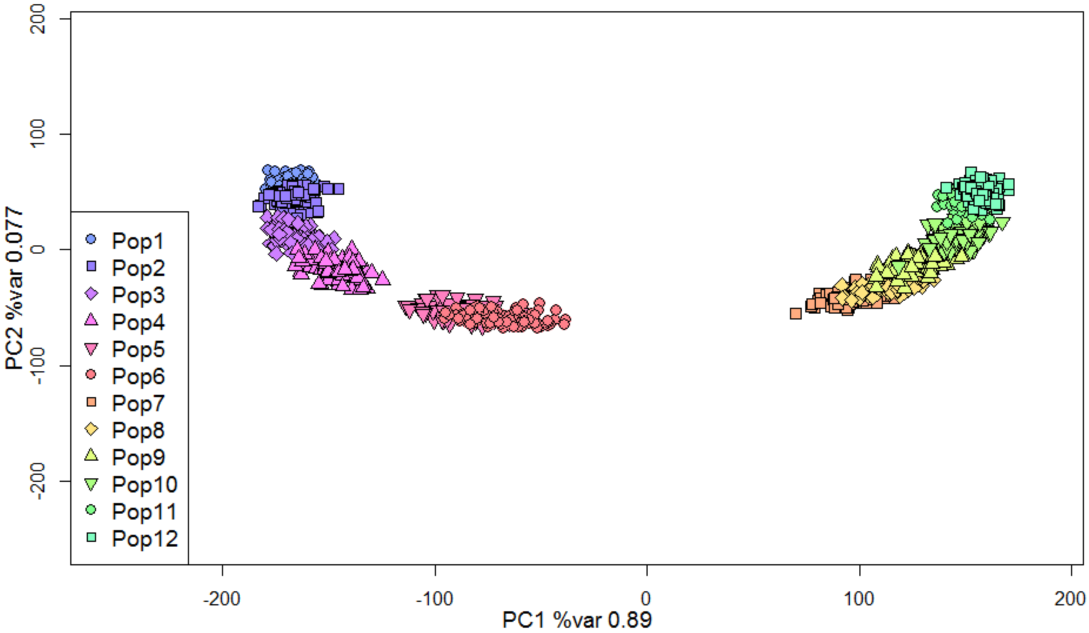


Figure E.7: GenePlot (leave-one-out mode) of data generated in Easypop with the hierarchical stepping stones model, using 2 groups of 6 islands (12 populations in total) of 100 individuals each for 3000 generations, with a migration rate of 0.01 within groups and 0.001 between groups.

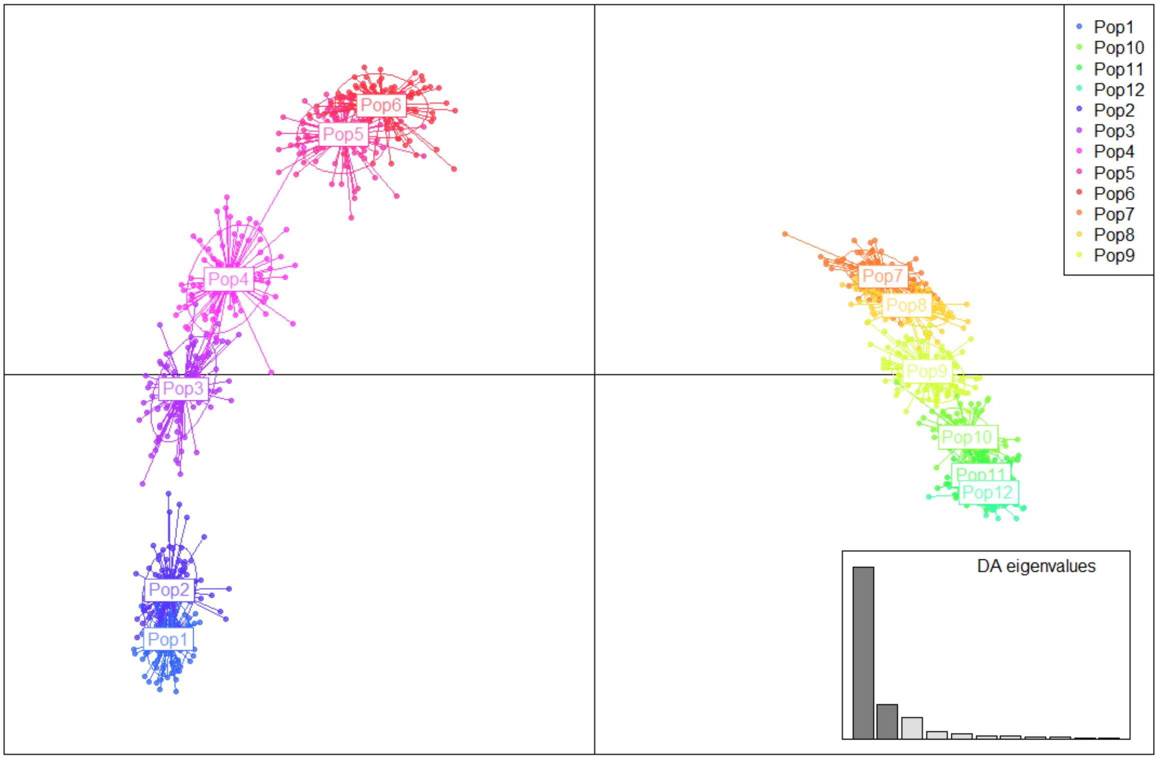


Figure E.8: Cluster plot based on DAPC results from `adegenet` of data generated in Easypop with the hierarchical stepping stones model, using 2 groups of 6 islands (12 populations in total) of 100 individuals each for 3000 generations, with a migration rate of 0.01 within groups and 0.001 between groups.

Appendix F SNPs

Although it was developed for microsatellite data, the GenePlot methodology and code can be applied directly to data from biallelic loci such as single nucleotide polymorphisms (SNPs), without requiring any adaptation of the algorithm.

Running GenePlot on SNPs may incur an increased computational cost due to the typically much higher numbers of loci involved. We tested an analysis of approximately 500 individuals with leave-one-out using 2000, 1000, 500, 250 and 125 SNP loci, and found that the run time to compute the GenePlot results doubled as the number of loci doubled. With 125 loci the run time was about 1 minute, and with 2000 loci the run time was approximately 15 minutes on average. The computer used approximately 30% of the processing capacity of an Intel i7 2.39GHz CPU and up to 2GB RAM. We therefore think that it is practically feasible to run GenePlot on SNP data as well as microsatellite data.

Figure F.9 shows the saddlepoint and genotype simulation approximations to an example distribution of simulated biallelic data from 1000 diploid loci. The distribution is extremely smooth and the distribution is well-approximated by the saddlepoint PDF and CDF. There is a minor discontinuity in the saddlepoint PDF at the mean of the distribution but, as seen from the closeness of the saddlepoint and empirical CDFs, it is of negligible magnitude for the CDF estimates used for visualization and assignment. Note that the SPDF is derived from the SCDF, and is used only for diagnostic plots such as those in Figure F.9; it is not part of the GenePlot procedure.

We tested GenePlot with two sets of simulated biallelic data, to represent SNPs. The first set of simulations uses various scenarios in which populations split and merge, akin to the scenarios used in Falush et al. 2003. The scenarios are shown in Figure F.10; the population sizes are shown in Table F.4. In scenario H, the splitting of the larger population into two populations of size N_L and the merging of one of those populations with one of the smaller populations happens instantaneously between Stage 1 and Stage 2.

Scenario	Ancestral	Stage 1	Stage 2
A	$N_L + N_S$	N_L and N_S	N_L and N_S
B	$N_L + 2N_S$	N_L, N_S and N_S	N_L, N_S and N_S
C	$N_L + 2N_S$	N_L+N_S and N_S	N_L, N_S and N_S
D	$N_L + 2N_S$	N_L and $2N_S$	N_L, N_S and N_S
E	$N_L + 2N_S$	N_L, N_S and N_S	N_L+N_S and N_S
F	$N_L + 2N_S$	N_L, N_S and N_S	N_L and $2N_S$
G	$N_L + 3N_S$	N_L, N_S, N_S and N_S	N_L, N_S and $2N_S$
H	$2N_L + 2N_S$	$2N_L$, N_S and N_S	N_L, N_L+N_S and N_S

Table F.4: Population sizes for simulation scenarios shown in Figure F.10. N_L and N_S are the larger and smaller population size parameters used in the simulations, and range from 100 to 10,000.

When simulating SNP data for these scenarios we tested population size parameters $N_S \in \{100, 200, 500, 1000\}$ and $N_L \in \{100, 200, 500, 1000, 10000\}$, $N_L \geq N_S$ and took samples of size 50 from each population to use in GenePlot. We used 1000 loci and simulated distinct

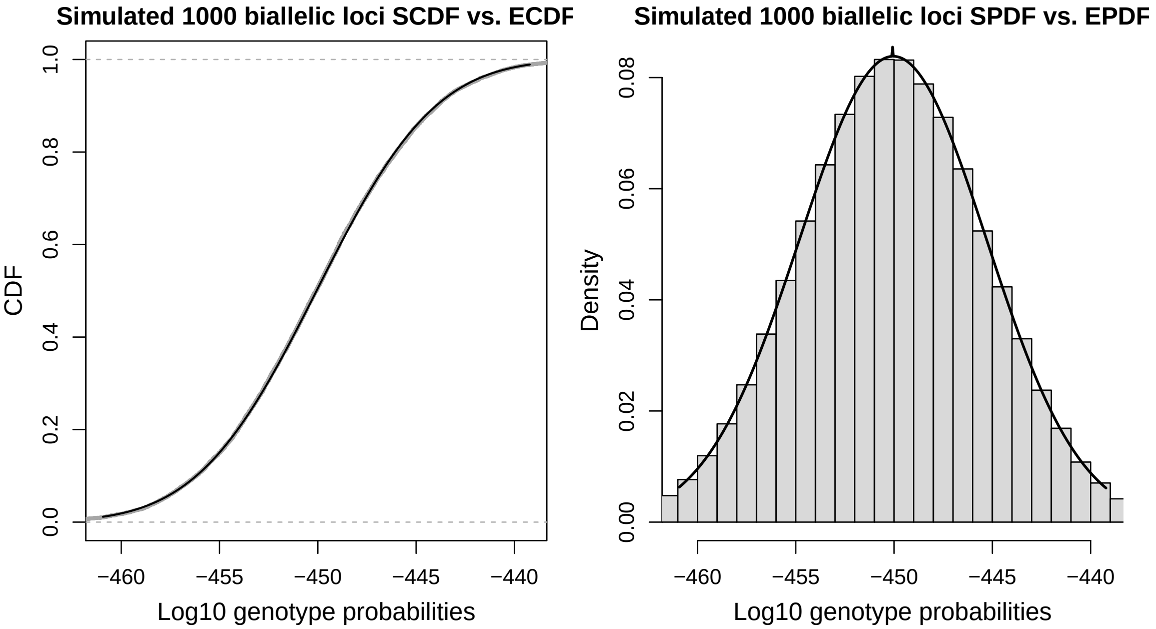


Figure F.9: The left plot shows the CDF of an example multilocus LGP distribution with 1000 biallelic loci. The wide grey line shows the empirical CDF based on 500,000 genotypes simulated from the population distribution (ECDF). The solid black line shows the saddlepoint approximation to the CDF (SCDF). The ECDF is hard to see due to the closeness of the approximations. The histogram in the right plot shows 100,000 log-genotype probabilities for genotypes simulated from the population distribution (EPDF) and the solid line shows the first derivative of the saddlepoint approximation to the CDF, which we denote SPDF. The right plot is truncated to better show the central part of the histogram.

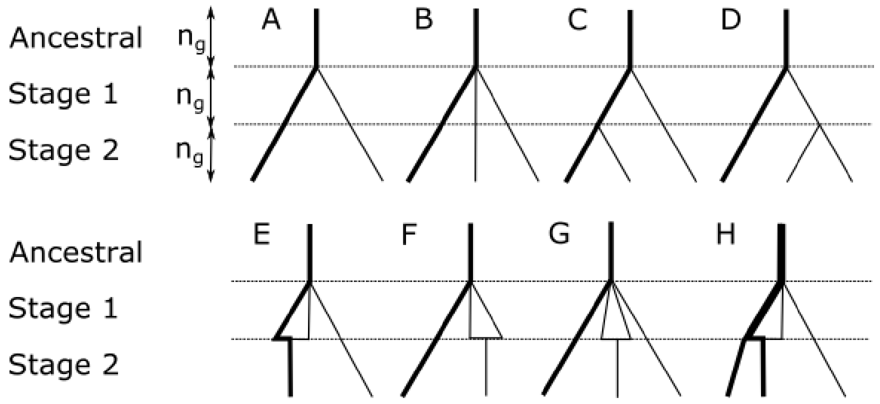


Figure F.10: Scenarios for simulating population data, where at each stage the populations are bred for n_g generations. Thick lines show the larger populations, thin lines show the smaller populations. The ancestral population is the size of all the final populations combined.

generations, each bred by selecting random gametes from random individuals in the previous generation.

We also simulated biallelic data in Easypop, producing 10 replicates of the island migration model with 6 populations each of size 100, using random mating and migration rate 0.05

for 3000 generations. We used 1000 loci, 2 possible allelic states, 0.0001 mutation rate and a recombination rate of 0.1 between adjacent loci to simulate strong linkage.

Figure F.11 shows example GenePlots from these simulations. The GenePlots for the scenario simulations used leave-one-out; the GenePlots for the Easypop simulations used the standard method to reduce computational cost. The GenePlot construction is the same as that for microsatellite data. The first two plots in Figure F.11, showing Scenario H with $n_g = 50$, show how much more distinct the final populations are if the large populations are reduced in size, such that genetic drift and the merge of Pop 1b with Pop2 have a greater impact on the genetic structure. The middle-left plot shows how reducing the number of generations of breeding reduces the level of differentiation, even for the same population sizes, although Pop 3 is still distinct from the other two populations. The results from Scenarios A to G have similar interpretations to those from Scenario H.

The middle-right plot in Figure F.11 shows an example of the Easypop simulations. Although the populations appear to be poorly differentiated in the first two principal components, these do not explain a high proportion of the variance; more separation is exhibited in lower principal components (not shown). The bottom two plots show that the populations are well differentiated pairwise.

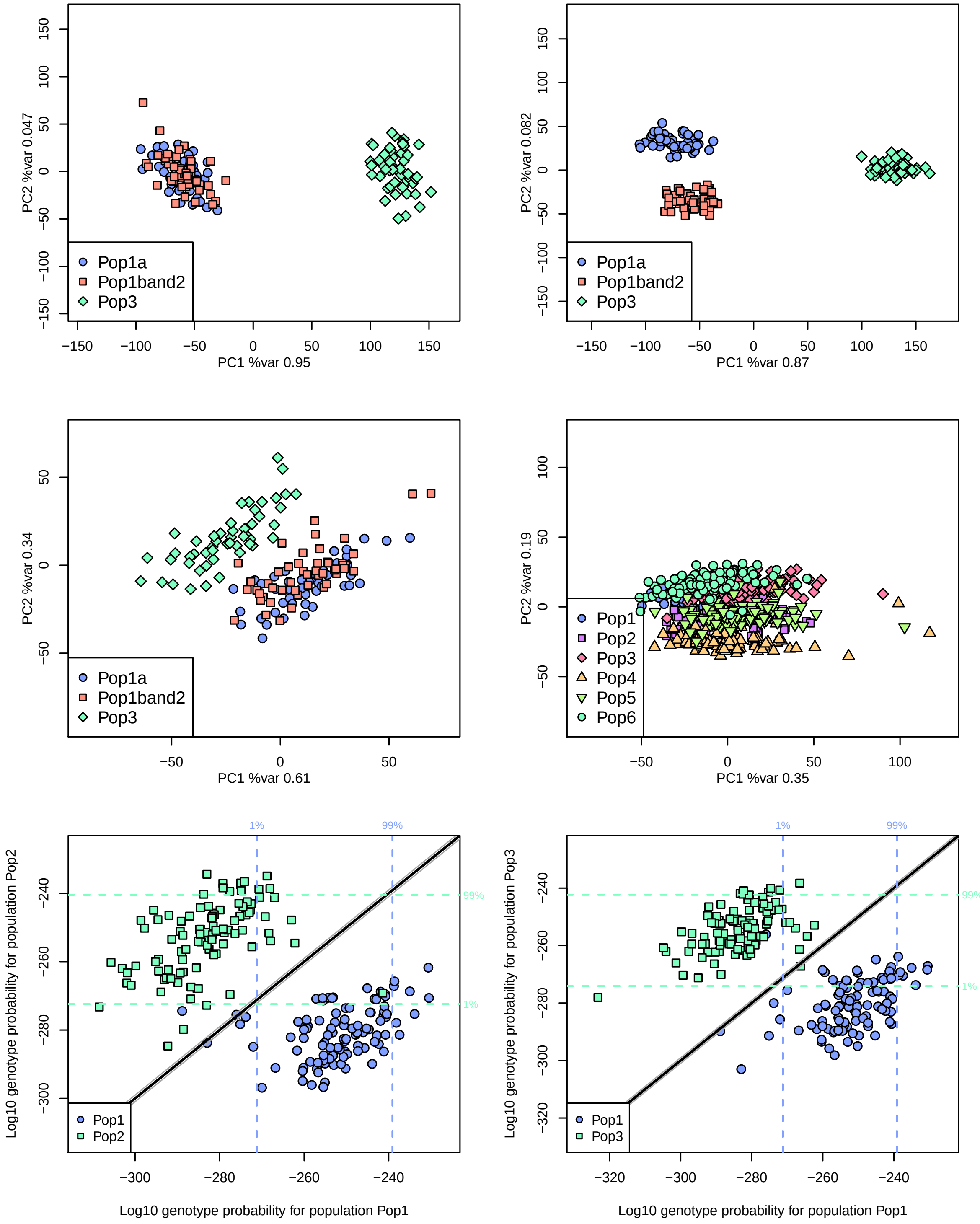


Figure F.11: GenePlots based on simulated biallelic data to represent SNPs. The top two plots show simulated data for two examples of Scenario H with n_g=50; the top-left plot has N_L=10000 and N_S=500; the top-right plot has N_L=1000 and N_S=500. The middle-left plot also shows an example of Scenario H with N_L=1000 and N_S=500 and n_g=10. The middle-right plot shows an Easypop simulation with 6 islands, each of size 100, bred for 1000 generations with a migration rate of 0.05 and recombination rate between adjacent loci of 0.1. The bottom two plots show two pairs of populations from the same Easypop simulation.

Appendix G Linkage disequilibrium

The metholodogy underlying GenePlots involves an assumption that genotypes at different loci are independent within individuals, in other words that there is negligible linkage disequilibrium. To investigate the influence of linkage disequilibrium (LD) on our results, we first estimated LD for the ship rat data from Kaikoura Island, the Broken Islands and Aotea; Table G.5 shows that the estimates are low for all three populations. Here, Δ is the composite disequilibrium measure defined in Schaid 2004 and Zaykin 2004; r^2 is the mean of the squared correlations over all locus pairs, as described below.

The correlation between a given pair of alleles A and B, at a given pair of loci 1 and 2, is calculated with the genotype correction as defined in Zaykin (2004) and Schaid (2004):

$$r_{AB} = \frac{\Delta_{AB}}{\sqrt{\{p_A(1-p_A) + D_A\}\{p_B(1-p_B) + D_B\}}},$$

where p_A and p_B are the estimated frequencies of allele A at locus 1 and allele B at locus 2 respectively. Here, $D_A = P_{AA} - p_A^2$ is the difference between the observed and expected levels of homozygotes of allele A, and similarly for D_B. Define r_{12}^2 for loci 1 and 2 to be the mean of r_{AB}^2 estimates over all allele pairs where allele A is from locus 1 and allele B is from locus 2. The overall mean LD estimate across all locus pairs, r^2, is a weighted mean of r_{12}^2 across locus pairs, weighted by the number of allele types at each of the two loci.

Population	r^2	Δ
Combined	0.010	0.006
Kaikoura Island	0.019	0.010
Broken Islands	0.026	0.014
Aotea	0.023	0.011

Table G.5: Linkage disequilibrium estimates for data from Kaikoura Island, the Broken Islands, Aotea and the combined populations.

We also calculated LD estimates for simulated data sets based on the scenarios in Figure F.10. The data sets for scenarios A to H do not include explicit genetic linkage, but they do include LD from other sources: LD caused by variation in ancestry among individuals, and "background LD" that occurs in all finite-sized populations due to sampling error during random mating of each generation. We ran all eight scenarios for 10 loci, with 5 or 10 initial allele states in the ancestral population, where the population sizes tested were as for the SNP simulations in Appendix F and with $n_g = 50$. We ran 100 replicates of each set of parameters. We tested whether LD affects misclassification rates by selecting a subgroup of runs displaying the highest estimates of r^2 and another subgroup of runs displaying the lowest estimates of r^2 from each set of replicate runs with the same parameters. Within each set of replicate runs, we then randomly paired up runs from the high and low LD subgroups. We used 10 pairs of runs for each scenario and each combination of N_L and N_S. For a given run, we used the highest estimated r^2 among populations in that simulation rather than the overall combined-population r^2, because the GenePlot LGP calculations are conducted separately for each reference population. The misclassification rates were calculated using the exclusion assignment method as described in the main text, using the 1% quantile LGP for each population as the exclusion threshold. Figure G.12 shows the paired differences in LD and differences in misclassification rate.

The concern regarding LD is that loci that are highly differentiating between two or more populations may be correlated with each other, and combining the data from these loci as if they were independent would overstate the level of population differentiation. The opposite is also a risk: that correlation between low-differentiating loci would understate the level of population differentation. However, Figure G.12 indicates that higher LD does not lead to higher error rates, and in fact the parameter sets with the most varied levels of LD are also the ones with the lowest error rates. The pairs with the lowest error rate differences typically had error rates near zero for both of the paired runs.

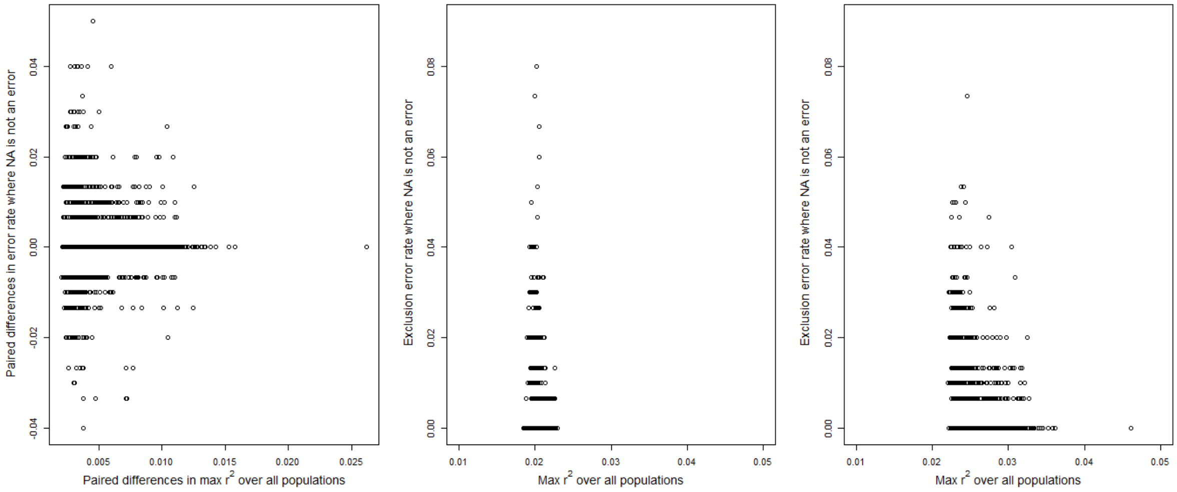


Figure G.12: Linkage disequilibrium versus misclassification rate differences for pairs of runs with high and low r^2. Left plot shows paired differences; middle plot shows the low r^2 runs; right plot shows the high r^2 runs.

We also tested LD with the biallelic/SNP data simulated in Easypop with a recombination rate of 0.1 between adjacent loci, instead of free recombination between all loci, to simulate a high degree of linkage. Table G.6 shows the LD estimates r^2 for the whole data set and the largest population-level r^2. The table also shows the misclassification results from the Easypop SNP simulations, using three different assignment protocols. Using the protocol of assigning an individual to the population for which it has the highest LGP, the misclassification rates are around 6%. An alternative assignment protocol is the exclusion method that is commonly used with GeneClass results (Manel et al. 2002), where the individual is only assigned to a population if it has LGP below a given threshold for all the other populations. This method is preferable to the method of choosing the highest LGP population because it does not assume that the true source population is among those studied, allowing for the possibility of migrants from other unidentified populations. If the individual has LGP above the threshold for more than one population it is not assigned, and is labelled as “NA”. The thresholds chosen for exclusion were the 1% quantile LGPs for each population. The results for misclassification errors under the exclusion method where “NA” is not counted as an error show that about 3% of individuals were misclassified after exclusion, and the results for exclusion where “NA” is counted as an error show that approximately 40% of individuals were not assigned. All other individuals were assigned correctly to their true source population and had very low LGP with respect to the other populations. These results demonstrate that even with a low recombination rate between adjacent loci, denoting high linkage, and a exclusion threshold of 1%, the proportion of incorrectly labelled individuals is very low, and the exclusion method can be used to avoid over-confident assignment where there is poor differentiation between populations. The 1% quantile is a stringent threshold: a low threshold increases the proportion of individuals who are not assigned because their LGP is higher than the threshold for more than one population.

Replicate	Overall	Max pop.	Choose highest LGP		Exclusion		Exclusion	
	r^2	r^2			NA is not an error		NA is an error	
			Count	Rate	Count	Rate	Count	Rate
1	0.00236	0.00848	31	0.052	12	0.020	274	0.457
2	0.00262	0.00892	28	0.047	13	0.022	168	0.280
3	0.00271	0.00876	35	0.058	25	0.042	181	0.302
4	0.00264	0.00900	43	0.072	23	0.038	209	0.348
5	0.00257	0.00927	47	0.078	26	0.043	273	0.455
6	0.00247	0.00926	32	0.053	18	0.030	267	0.445
7	0.00254	0.00911	40	0.067	15	0.025	262	0.437
8	0.00241	0.00886	36	0.060	10	0.017	318	0.530
9	0.00260	0.00905	37	0.062	19	0.032	269	0.448
10	0.00257	0.00924	36	0.060	17	0.028	236	0.393

Table G.6: LD estimates and misclassification results based on the GenePlot method under different assignment protocols, for 10 replicates of data generated in Easypop with the island migration model, using 6 populations of 100 individuals each for 1000 generations, with a migration rate of 0.05 and a recombination rate for adjacent loci of 0.1.

Chapter 3

Directional measures of population differentiation

3.1 Introduction

We propose a new framework of genetic measures based on the distributions of log-genotype probabilities, as introduced in McMillan & Fewster (2017). We derive the methodology starting from the genetic assignment method of Rannala & Mountain (1997), and we consider the differentiation between populations in terms of the fit of individuals to those populations. We consider two populations to be highly differentiated if the individuals from each reference population have a poor fit to the other population. We define fit in terms of the log-probability of an individual's genotype arising in the population, given the population's estimated allele frequencies.

We therefore consider differentiation as two related management questions. Firstly, are the populations distinct enough that, given individuals from two different populations, we would expect to accurately identify which came from each population? Secondly, are the populations distinct enough that a single individual being assigned among the populations is more likely to be assigned to its own population? Within this framework, differentiation between populations may well be directional. If population A has only a restricted set of alleles and population B has a much wider selection of common alleles then individuals arising from B are likely to have a poor fit to A, but individuals from A might have a good fit to B. Such disparity between populations is not captured by existing measures such as F_{ST} (Wright, 1951), G''_{ST} (Meirmans & Hedrick, 2011) or Jost's D (Jost, 2008).

Within the framework of log genotype probability, we can calculate numeric measures to answer these questions, and also run permutation tests for significant population structure.

Many existing studies test for population structure, using STRUCTURE (Pritchard et al. 2000, Hubisz et al. 2009), the AMOVA routine in software Arlequin (Excoffier & Lischer 2010), or by running significance tests on F_{ST} and other diversity measures. Choosing the number of inferred clusters, K, from STRUCTURE, can be thought of as a test for differentiation, based on whether the number of clusters is as large as the number of sampling locations, and it is cast as a model selection exercise with the log-likelihood $L(K)$ or the ΔK method of Evanno et al. (2005) as the selection criterion. This is useful to test for population structure without specifying pre-defined populations.

The fineSTRUCTURE software provides more fine-grained clustering via a coancestry matrix (Lawson et al. 2012) and thus can be more robust for model selection using SNP data. Lawson et al. (2018) argue that the results of STRUCTURE may be strongly affected by sample size, and that fineSTRUCTURE is more suitable than STRUCTURE for large numbers of loci. Janes et al. (2017) argue that STRUCTURE and the ΔK method may be biased towards certain values of K, and it is not always clear how best to match up the inferred clusters to the sampling locations.

Given a predefined set of populations, the DAPC method from the `adegenet` R package (Jombart et al. 2010) can be used to visualize the genetic data, and, like STRUCTURE, can indicate which populations overlap with each other. This method has the advantage of not using any assumptions of Hardy-Weinberg equilibrium or linkage equilibrium, but the plots do not correspond to any biological quantity, so they are often difficult to interpret.

Analysis of molecular variance (AMOVA) in Arlequin, and existing measures such as F_{ST}, G''_{ST} and Jost's D, are used to test whether pre-defined populations are significantly separated. However, as described in Section 1.3, there is still no consensus about which measures to use for assessing population differentiation, because the behaviour of the existing measures varies in different types of scenarios. Moreover, even when applied pairwise to populations, these methods only give indicators of heterozygosity within and among populations, or of the commonality of allele types. The measures we propose describe populations in terms of the fit of individual genotypes, and are thus more easily interpretable and more applicable to management scenarios.

When analysing pre-defined populations, no other existing framework offers visualization combined with numeric measures, to allow users to understand and interpret the results of the significance tests. We develop the first such framework here.

3.2 Distributions of log-genotype probabilities

McMillan & Fewster (2017) introduced *GenePlots*, which extend the methods of Rannala & Mountain (1997) by visualizing the fit of individuals to candidate source populations. We now generalize that method to consider the fit of overall populations to each other, and formalize this into a method of drawing inferences on population differentiation. We motivate this method using a case 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 (28500ha); on Kaikoura Island (530ha), which is 80m off the coast of Aotea at closest approach; and on the Broken Islands, which comprise four islands with total area 125 ha, about 300m from Aotea and about 3km south of Kaikoura Island. A map of the sampling locations can be found in Appendix A.

Figure 3.1 shows the GenePlot of Aotea against the Broken Islands, alongside two additional LGP distribution plots. The GenePlot, in which every point represents an individual rat, shows the multilocus log-genotype probabilities (LGPs) of the sampled rats with respect to the Broken Islands and Aotea. The LGPs were calculated using the leave-one-out method. We use log base 10 rather than the natural log, and the Rannala & Mountain prior (see Chapter 2 or McMillan & Fewster, 2017), for this and all other analyses in this chapter.

The plot shows clear separation between the populations, which fall into two clusters. Apart from one rat sampled on Aotea which clusters with the rats sampled on the Broken Islands, and which is a probable migrant, all the Aotea rats have a poor fit to the Broken Islands, with LGPs to the left of the 1% percentile line in the Broken Islands population. The rats sampled on the Broken Islands, by contrast, commonly have a good fit to Aotea as well as to the Broken Islands.

Figure 3.1 also shows a set of plotted curves, which we call the LGP distribution plots. These are shown along the x- and y-axes of the GenePlots, because they can be thought of as cross-sections through the GenePlot parallel to those axes. These posterior probability density curves show the fit of the overall populations to each other. They are based on the posterior distributions of allele frequencies in the two populations, and thus capture not only the fit of sampled individuals but also the fit of any potential genotypes that might arise in the populations, as well as the uncertainty in estimating allele frequencies from sample data. The LGP plots as well as the GenePlots use leave-one-out to reduce bias in the estimates. Details of the definition and construction of these plots are in Appendix H.

We can see in Figure 3.1 that the overlap of the Broken Islands into the Aotea LGP distribution, as shown on the left-hand side of the plot, is much greater than the overlap of Aotea into the Broken Islands LGP distribution, as shown at the bottom of the plot. The disparity between the two LGP plots is indicative of the fact that the Broken Islands population was most likely founded from the Aotea population, so that the alleles found in the Broken Islands are also found in Aotea, but due to founder effects and genetic drift, many of the alleles on Aotea are not found on the Broken Islands. The LGP plots show that if rats from the Broken Islands population were sampled on Aotea they might well be presumed to be from Aotea, as they are likely to have a reasonable fit to the Aotea population. By contrast, if rats from the Aotea population were sampled on the Broken Islands it would very likely be obvious that they were not from the Broken Islands. We call this situation *drifted subsetting* and we consider the Broken Islands population to be a drifted genetic subset of the Aotea population.

Figure 3.2 shows the GenePlot and LGP distribution plots for rats sampled on Kaikoura Island and Aotea. By contrast with the Broken Islands and Aotea samples, the Kaikoura and Aotea samples are much closer together, and many rats from both populations have a good fit

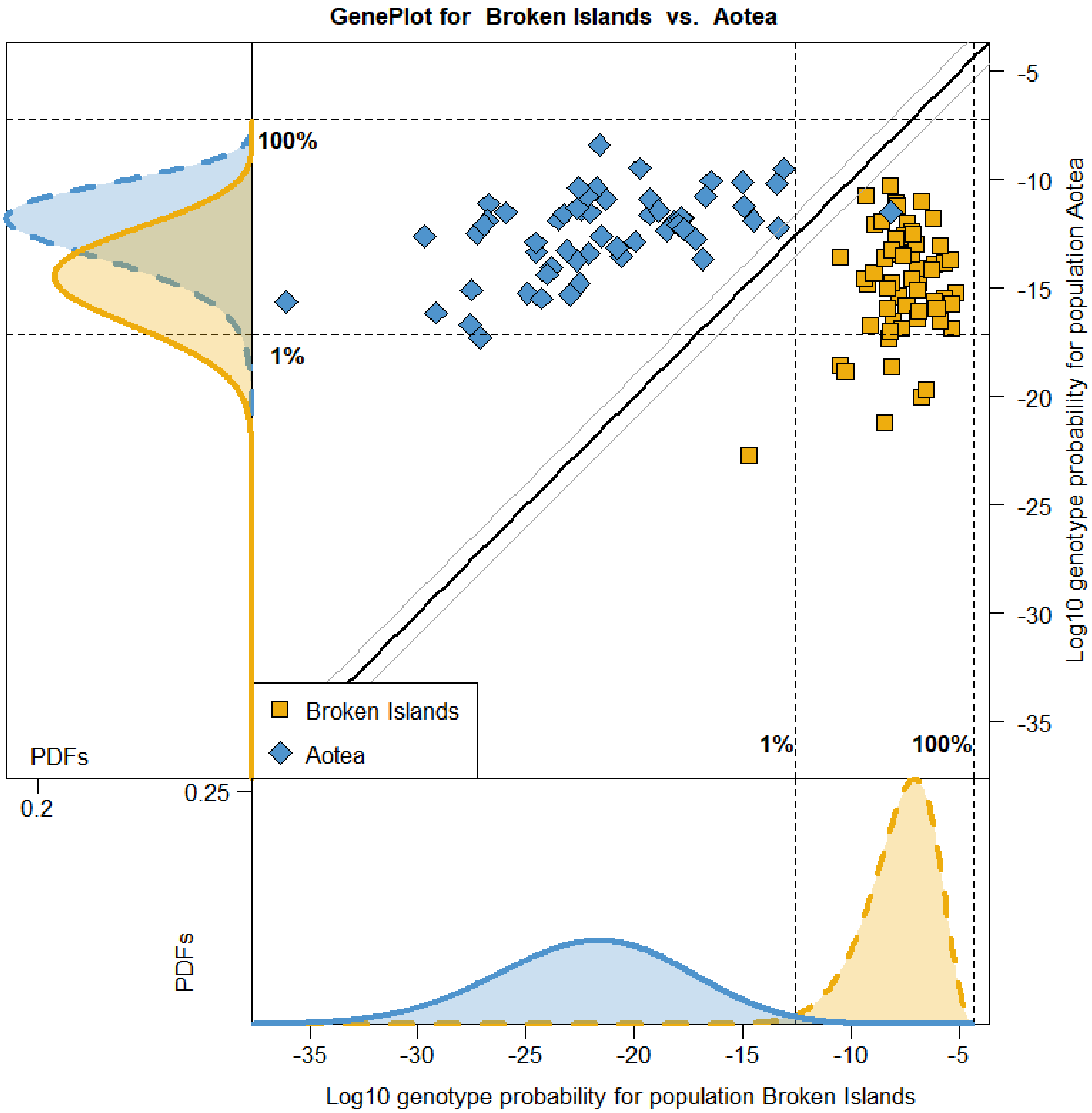


Figure 3.1: GenePlot and LGP distribution plots based on microsatellite data extracted from ship rats (*Rattus rattus*). Each point on the central plot represents an individual rat captured on the Broken Islands (squares, $n = 60$) or Aotea (diamonds, $n = 57$). The main horizontal axes show the posterior log-genotype probability (LGP) of each individual's genotype within the Broken Islands population; the main vertical axes show the same, but with respect to the Aotea population. The thick diagonal line shows equal probability with respect to Aotea and the Broken Islands, and the fine diagonal lines show 10 times higher probability for one population than the other. The vertical dashed lines show the 1% and 100% percentile lines for the Broken Islands; the horizontal lines show the 1% and 100% percentile lines for Aotea. The secondary plot to the left of the main plot shows the LGP distribution of the Aotea population (dashed line) and the cross-population LGP distribution of the Broken Islands population within the Aotea population (solid line). The secondary plot below the main plot shows the LGP distribution of the Broken Islands population (dashed line) and the cross-population LGP distribution of the Aotea population within the Broken Islands population (solid line). The LGPs and the dashed curves were calculated using the leave-one-out method.

to both populations, as indicated by their high LGPs with respect to both populations. Despite the level of overlap between the populations, there is still a certain amount of population structure visible. Most of the individuals from Aotea are on the left-hand side of the thick diagonal line, showing that they have a higher fit to Aotea than to Kaikoura, and similarly most of the Kaikoura rats have a higher fit to Kaikoura than to Aotea.

Again, the accompanying LGP plots show this in terms of the overall populations and not just the observed individuals. The overlap of Kaikoura into Aotea (shown in the left-hand LGP plot) is higher than the overlap of the Broken Islands into Aotea in Figure 3.1, and the overlap of Aotea into Kaikoura is far higher than the overlap of Aotea into the Broken Islands in Figure 3.1.

3.3 Numeric measures of population structure

We can use the LGP distributions of the populations to construct numeric measures of population structure. The first of these we call the *overlap area*. This is the area of the overlap between the two curves in either of the subsidiary plots in Figure 3.1 which show the probability density functions (PDFs) of the LGP distributions for the populations. This is a visual measure, corresponding directly with what can be seen on the plots.

The overlap area in the bottom plot in Figure 3.1 compares the Aotea population with the baseline Broken Islands population. Let $\breve{f}(x)$ be the saddlepoint approximation to the PDF of the leave-one-out LGP distribution for the Broken Islands population, shown as the yellow dashed line in the bottom part of Figure 3.1. Let $\hat{\xi}(x)$ be the saddlepoint approximation to the PDF of the cross-population LGP distribution of Aotea within the Broken Islands, which is the solid blue curve in Figure 3.1, and represents the distribution of how well genotypes arising from Aotea would fit into the Broken Islands population. (See Appendix H for the formal definitions of $\breve{f}$ and $\hat{\xi}$.) Then the overlap area with baseline Broken Islands is:

$$\int \min(\hat{\xi}(x), \breve{f}(x))\, dx.$$

The overlap area with respect to Aotea is similar, but uses the curves defined in the left LGP distribution plot in Figure 3.1 which compares the Broken Islands population with the baseline Aotea population. Let $\breve{g}(x)$ be the saddlepoint approximation to the leave-one-out PDF of the LGP distribution of the Aotea population, shown as the blue dashed line in the left-hand part of Figure 3.1. Let $\hat{\zeta}(x)$ be the saddlepoint approximation to the PDF of the cross-population LGP distribution of the Broken Islands within Aotea, shown as the solid yellow curve in Figure 3.1. Then the overlap area with baseline Aotea is:

$$\int \min(\hat{\zeta}(x), \breve{g}(x))\, dx.$$

We calculate the overlap area using the saddlepoint approximation, as in McMillan and Fewster (2017). Full details are in Appendix H. Overlap area takes values between 0 and 1. A low overlap area signals high separation between the populations.

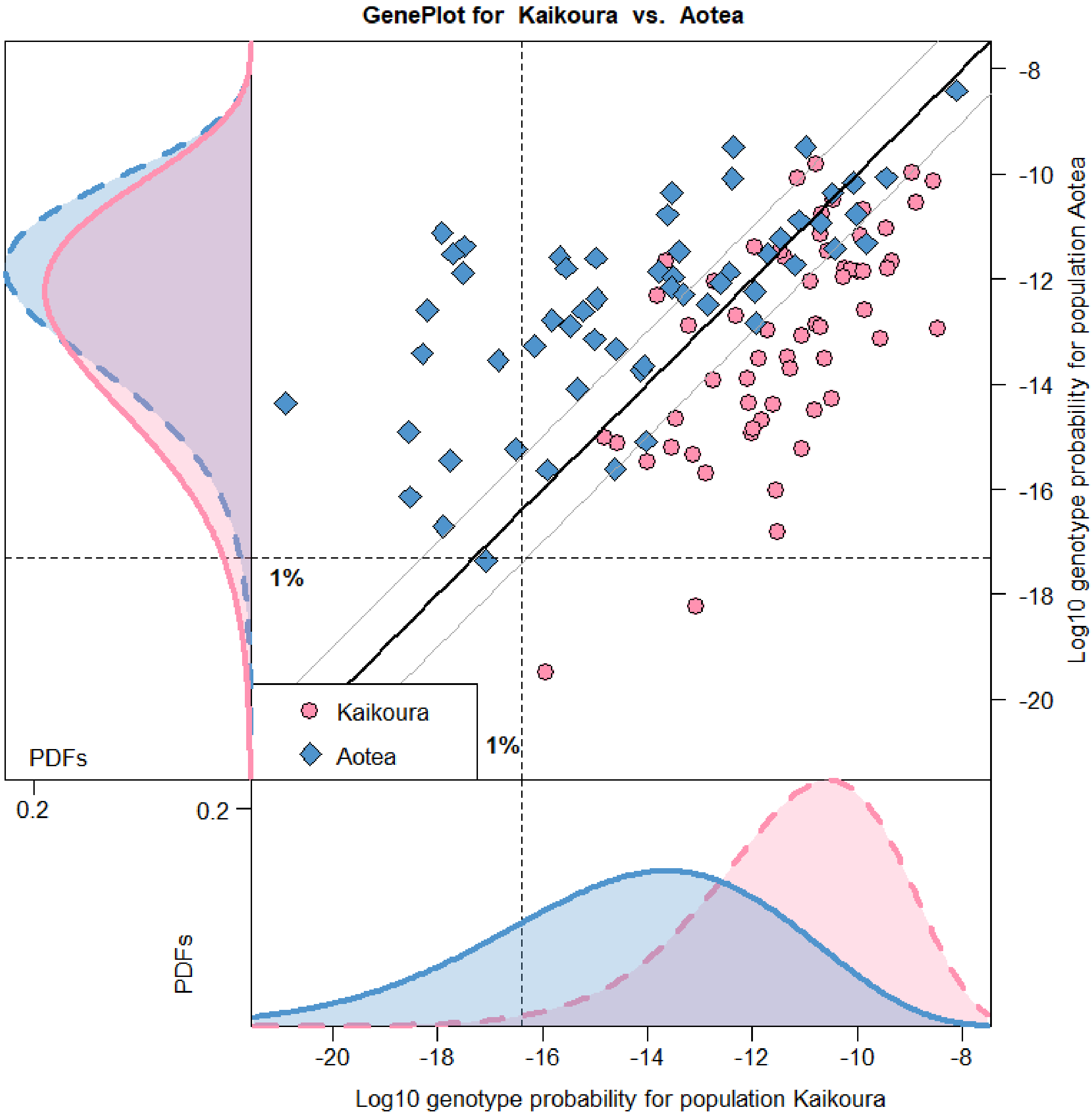


Figure 3.2: GenePlot and LGP distribution plots based on microsatellite data extracted from ship rats (*Rattus rattus*). Details as in Figure 3.1, for samples from Kaikoura Island ($n = 60$) and Aotea ($n = 57$).

The leave-one-out overlap area with respect to the Broken Islands baseline is 0.020 and the leave-one-out overlap area with respect to the Aotea baseline is 0.533. This indicates that individuals from Aotea would not fit well into the Broken Islands population, but some Broken Islands individuals would fit well into the Aotea population.

The overlap area is intuitive in terms of its visualization, but suffers from not having a clear probabilistic interpretation. Moreover, it is a similarity index for which a value of 1 indicates low differentiation between populations, unlike existing measures of differentiation for which a value of 1 indicates high differentiation between populations. Therefore we also provide alternative measures connected to the LGP distributions, which have direct probabilistic interpretations.

The second measure we call the *interloper detection probability*. The interloper detection probability in population A is the probability that for two genotypes randomly arising from populations A and B, the genotype arising from A is the one with the better fit to A. A high interloper detection probability would indicate high separation of population A from population B.

As an analogy, we consider the possibility of a Dutch child and an English child taking an English language test. The interloper detection probability is the probability that the English child would score higher on the English language test than the Dutch child, so that the Dutch child is correctly detected as the "interloper" in the English population.

We can also calculate the interloper detection probability with respect to population B, which in the analogy would correspond to the probability of a randomly selected Dutch child scoring higher on a Dutch language test than a randomly selected English child.

Given the extent of English teaching in Dutch schools and the lack of Dutch teaching in English schools, we would expect the latter probability to be higher than the former. Similarly, in two populations we would expect the interloper detection probability to differ for the two populations.

We calculate these probabilities using the probability density functions (PDFs) and cumulative distribution functions (CDFs) of the LGP distributions for the populations. Let $\hat{\xi}(x)$ be the saddlepoint approximation to the PDF of the cross-population LGP distribution of Aotea within the Broken Islands as before. Let $\breve{F}(x)$ be the saddlepoint approximation to the CDF of the leave-one-out LGP distribution of the Broken Islands population, which is the cumulative integral of the yellow dashed curve in Figure 3.1. Then the interloper detection probability in the Broken Islands is:

$$1 - \int \hat{\xi}(x)\breve{F}(x)\,dx.$$

This corresponds to the probability that a random draw from the yellow dashed PDF curve at the bottom of Figure 3.1 exceeds a random draw from the blue solid curve. We calculate the interloper detection probabilities using the saddlepoint approximation, as in McMillan and Fewster (2017). Full details are in Appendix H.

The interloper detection probability in Aotea is similar. $\hat{\zeta}(x)$ is the saddlepoint approximation to the PDF of the cross-population LGP distribution of the Broken Islands within Aotea, as before. and $\breve{G}(x)$ is the saddlepoint approximation to the CDF of the leave-one-out LGP distribution of the Aotea population, which is the cumulative integral of the blue dashed curve in Figure 3.1. Then the interloper detection probability in Aotea is:

$$1 - \int \hat{\zeta}(x)\breve{G}(x)dx.$$

The Aotea leave-one-out interloper detection probability with respect to the Broken Islands is 0.999 and the Broken Islands leave-one-out interloper detection probability with respect to Aotea is 0.809. Similarly to the overlap areas, these indicate that in most cases we would correctly identify which of two individuals found in the Broken Islands was in its home population, but we would not always correctly identify which of two individuals found in Aotea was in its home population.

The third measure we call the *correct assignment probability*. This is the probability that a genotype randomly arising from population A will have a better fit to population A than to population B, so that under assignment to the best fit population it would be assigned to population A. Once again, high values of the correct assignment probability indicate substantial genetic differentiation between the populations.

The analogy for this would be the probability that an English child would score higher on the English language test than the Dutch language test. The corresponding correct assignment probability for the Dutch population would be the probability that a Dutch child would score higher on the Dutch language test than the English language test. For the interloper detection probability we considered two hypothetical individuals and calculated the probability of correctly discerning which individual had originated in the given population; by contrast for the correct assignment probability we are considering one hypothetical individual and calculating the probability of correctly discerning which population the individual originated in.

We calculate the correct assignment probabilities by generating the distributions corresponding to the difference of the LGPs for the two populations. We use leave-one-out to calculate the individual's probability with respect to its own population, to avoid bias caused by the presence of the individual's alleles within the observed sample. For every possible genotype arising in the Broken Islands, given its posterior allele frequencies, we calculate the genotype's leave-one-out LGP with respect to the Broken Islands and subtract its LGP with respect to Aotea. We then distribute all these LGP differences according to the probabilities of the genotypes arising in the Broken Islands population, using the saddlepoint approximation as usual. Finding the PDF of the difference in LGPs for each genotype ensures we take into account the dependence between the LGPs of a single genotype with respect to the two populations.

We perform a similar process for the Aotea population. Figure 3.3 shows the resulting LGP difference distributions for the Broken Islands and Aotea.

In Figure 3.3(a) the value of 0 on the x-axis, for example, corresponds to genotypes that would have equal LGP with respect to the Broken Islands and Aotea, and the height of the curve gives the associated probability density of such genotypes arising on the Broken Islands. The correct assignment probability in the Broken Islands is the area under the curve to the right of 0 on the x-axis, which is the proportion of all genotypes arising from the Broken Islands that have higher LGP with respect to the Broken Islands than with respect to Aotea. The correct assignment probability for Aotea is the area under the curve in Figure 3.3(b) to the right of 0 on the x-axis.

The Broken Islands and Aotea LGP difference PDF curves both have most of their area to the right of 0 on the x-axis, indicating that rats from either population will almost always have a better fit to the population they originated from. On the GenePlot, this corresponds to individuals being placed on the correct side of the central diagonal line. The leave-one-out correct assignment probability for the Broken Islands is 0.997 and the leave-one-out correct assignment probability for Aotea is also 0.997. These indicate that in almost all cases, individuals from each population would have a better fit to their own population and would be correctly assigned if using the basic assignment protocol of assigning to the population for which the individual has higher LGP. When using a more conservative assignment protocol, some of the individuals may not be assigned to either population if they have similar fit to both populations.

The correct assignment probability could also be calculated using a more conservative assignment protocol. For example, because Figure 3.3 is on a log10 scale, the area to the right of 2 on the x-axis would be the probability of a random individual from population A having a genotype probability with respect to A that is 100 times higher than its genotype probability with respect to population B. In this case, the leave-one-out correct assignment probability for the Broken Islands would be the probability of a random individual from the Broken Islands having a genotype probability for the Broken Islands that is at least 100 times higher than its genotype probability for Aotea. Using this more conservative protocol, the correct assignment probability for the Broken Islands is 0.976 and the correct assignment probability for Aotea is 0.984.

The numeric measures for the Kaikoura Island and Aotea populations show that there is less separation between these populations than between the Broken Islands and Aotea populations. The leave-one-out overlap area with respect to Kaikoura Island is 0.499 and the leave-one-out overlap area with respect to Aotea is 0.883. The Aotea leave-one-out interloper detection probability in Kaikoura Island is 0.828 and the Kaikoura Island leave-one-out interloper detection probability in Aotea is 0.576, barely better than chance.

Figure 3.4 shows the LGP difference distributions for the Kaikoura Island and Aotea populations. These two populations have less area to the right of 0 than the Broken Islands and Aotea populations. This indicates that a non-negligible proportion of rats originating from the posterior genotype distribution for Aotea would actually have a better fit to Kaikoura than to Aotea, and similarly for the rats originating from Kaikoura. The correct assignment

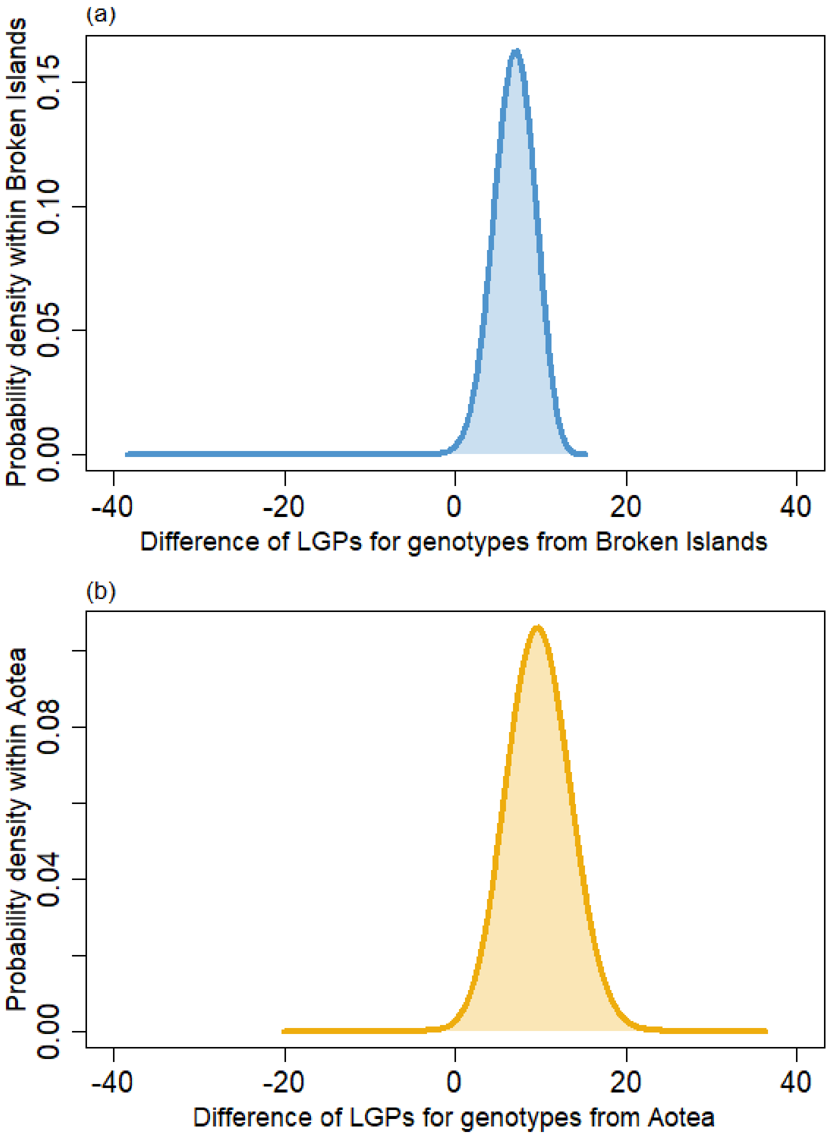


Figure 3.3: LGP difference distribution plots based on microsatellite data extracted from ship rats (*Rattus rattus*) on the Broken Islands and Aotea. (a) Log-genotype probability with respect to the Broken Islands minus log-genotype probability with respect to Aotea for all possible genotypes arising from the Aotea or Broken Islands populations, distributed according to their probability of arising in the Broken Islands population. (b) Log-genotype probability with respect to Aotea minus log-genotype probability with respect to the Broken Islands for all possible genotypes arising from the Aotea or Broken Islands populations, distributed according to their probability of arising in the Aotea population.

probability for Kaikoura Island is 0.860 and the correct assignment probability for Aotea is 0.870.

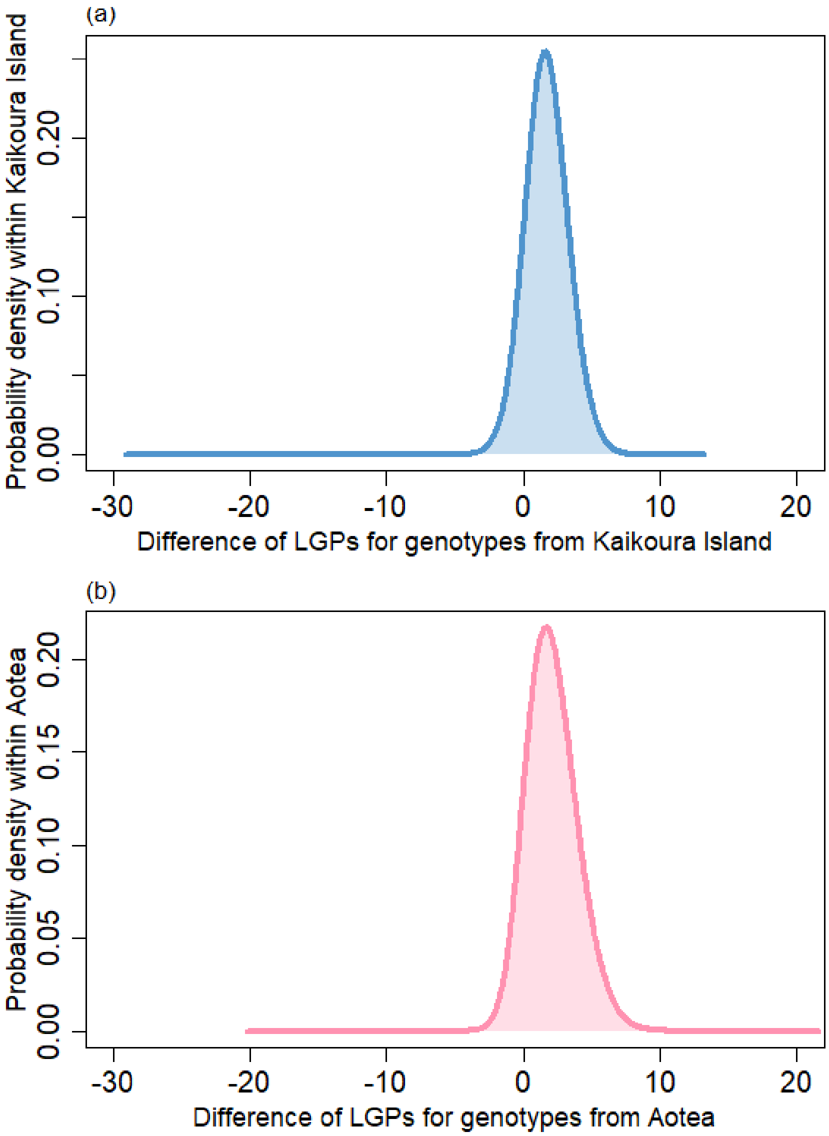


Figure 3.4: LGP difference distribution plots based on microsatellite data extracted from ship rats (*Rattus rattus*) on Kaikoura Island and Aotea. (a) Log-genotype probability with respect to Kaikoura Island minus log-genotype probability with respect to Aotea for all possible genotypes arising from the Aotea or Kaikoura Island populations, distributed according to their probability of arising in the Kaikoura Island population. (b) Log-genotype probability with respect to Aotea minus log-genotype probability with respect to Kaikoura Island for all possible genotypes arising from the Aotea or Kaikoura Island populations, distributed according to their probability of arising in the Aotea population.

3.4 Permutation tests for population structure

Numeric measures of population differentiation allow us to use permutation tests to check for significant population structure, by permuting the population labels in the dataset many times and comparing the values calculated from those randomised datasets with the observed value calculated from the original dataset. Under the null hypothesis of no genetic separation, the population labels of the individuals are immaterial and any separation observed in the numeric measures is due to chance alone.

Similar significance tests exist for measures such as F_{ST} (Excoffier et al 1992; Goudet 1995; Goudet et al. 1996; Peakall and Smouse, 2006 and 2012).

If the observed value of genetic separation is larger than most of the randomised values then we conclude that there is significant evidence of population structure. For example, if we create 1000 randomised datasets and there are only 8 that have larger values of interloper detection probability than the observed value of interloper detection probability, then the p-value of 0.008 based on 1000 randomised values shows strong evidence of structure between those two populations. For interloper detection probability and correct assignment probability, which serve as measures of separation, the p-value is taken from the right tail of the null distribution, whereas for overlap area, which is a measure of similarity, the p-value is taken from the left tail of the null distribution.

When calculating the genetic measures for the randomised datasets, we keep using the original population labels instead of using general names like 'Group 1' and 'Group 2'. For populations of different sizes this is equivalent to keeping the same name for the larger population and the same name for the smaller population, and for equally sized populations this is equivalent to randomising the population names. Leave-one-out LGP distributions are used for all single-population computations in this section (see Appendix H for details).

Figure 3.5 shows the randomisation test results for the Broken Islands and Aotea populations. In each subplot, the histogram indicates the null distribution values obtained from the randomised datasets, and the bold line shows the observed value from the original dataset. We used 100 randomised datasets to reduce computational time. If we obtained results that were less clear-cut, we could use more randomisations to increase the precision of the p-value. However, in Figure 3.5, 100 randomisations is sufficient to show a very clear result. The two populations are significantly separated by every measure. For overlap area, where population separation is indicated by low values, the observed measures are much lower than the randomised values. For the two probabilistic measures, where population separation is indicated by high values, the observed measures are much higher than the randomised values. Permutation tests using F_{ST}, G''_{ST} and D were also significant (results not shown).

Figure 3.6 shows equivalent results for the Kaikoura Island and Aotea populations. Although for these two populations the GenePlot showed the two populations interleaved, with many individuals from each having a good fit to both populations, the randomisation test results show that there is still significant population structure present, although the observed

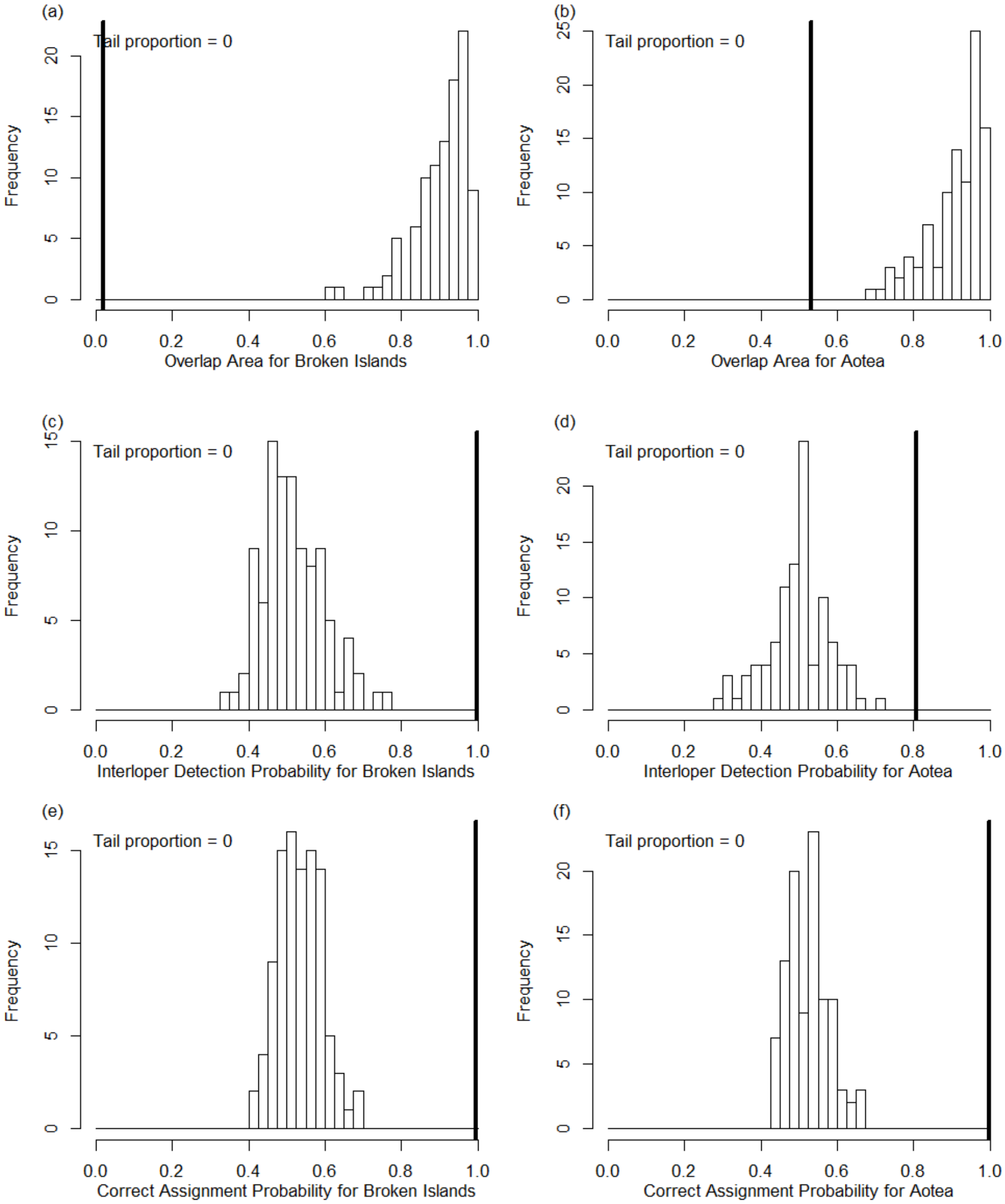


Figure 3.5: Permutation test results based on microsatellite data extracted from ship rats (*Rattus rattus*) on the Broken Islands ($n = 60$) and Aotea ($n = 57$). In each plot, the bold line indicates the observed value and the histogram shows values obtained from 100 datasets in which the population labels were randomly reassigned to all the rats. (a) Overlap area between Broken Islands LGP distribution and the cross-population LGP distribution of Aotea within the Broken Islands. (b) Overlap area between Aotea LGP distribution and the cross-population LGP distribution of the Broken Islands within Aotea. (c) Probability that for a random pair of genotypes arising from the Broken Islands and the Aotea populations, the one arising from the Broken Islands will have a better fit to the Broken Islands population. (d) Probability that of those two genotypes, the one arising from Aotea will have a better fit to the Aotea population. (e) Probability that a random genotype arising from the Broken Islands will have a better fit to the Broken Islands population than to the Aotea population. (f) Probability that a random genotype arising from Aotea will have a better fit to the Aotea population than to the Broken Islands population. The single-population LGP distributions were constructed using leave-one-out.

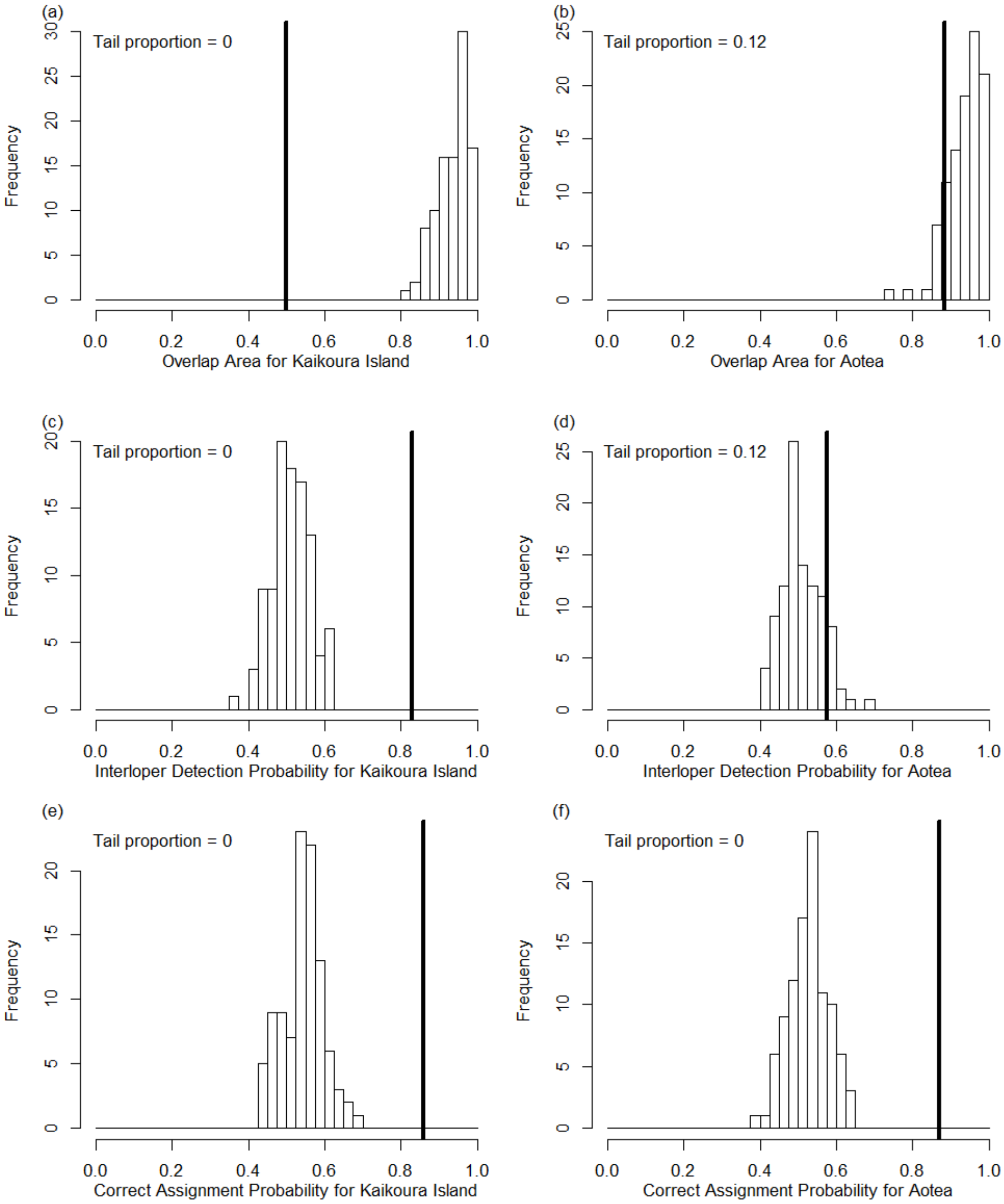


Figure 3.6: Permutation test results based on microsatellite data extracted from ship rats (*Rattus rattus*) on Kaikoura Island ($n = 60$) and Aotea ($n = 57$). Details as for Figure 3.5.

measures show less separation than for the Broken Islands versus Aotea. The overlap area and interloper detection probability are significant for Kaikoura but non-significant for Aotea, showing the directionality of the structure between these two populations.

3.5 Application to data from New Zealand fur seals (*Arctocephalus forsteri*)

We apply our method using microsatellite data at 11 loci from New Zealand fur seals (*Arctocephalus forsteri*), as described by Dussex et al. (2016). The authors of that study defined four regions: NZ North, NZ South, Subantarctic islands and Australia. It should be noted that NZ North and NZ South do not correspond to the North and South Islands of New Zealand, but rather are groupings of sampling locations off the coast of South Island, as well as one single sampling location on North Island. The majority of the seals sampled were linked to the two New Zealand regions, and Dussex et al. (2016) observed weak separation between NZ North and NZ South.

Figure 3.7 shows the GenePlot for the four regions identified by Dussex et al. (2016) and Figure 3.8 shows the multi-bar GenePlot for the same four regions. Both GenePlots agree with the findings of Dussex et al. (2016) that the Australian region is the most distinct: the seals from the Australian region have a poor fit to all other regions and the seals from the other regions have a poor fit to the Australian region. The GenePlots also confirm the findings of Dussex et al. (2016) that there is strong genetic similarity between both NZ regions and some genetic similarity between the NZ regions and the Subantarctic region.

We henceforth investigate the two largest regions, NZ North and NZ South, in more detail.

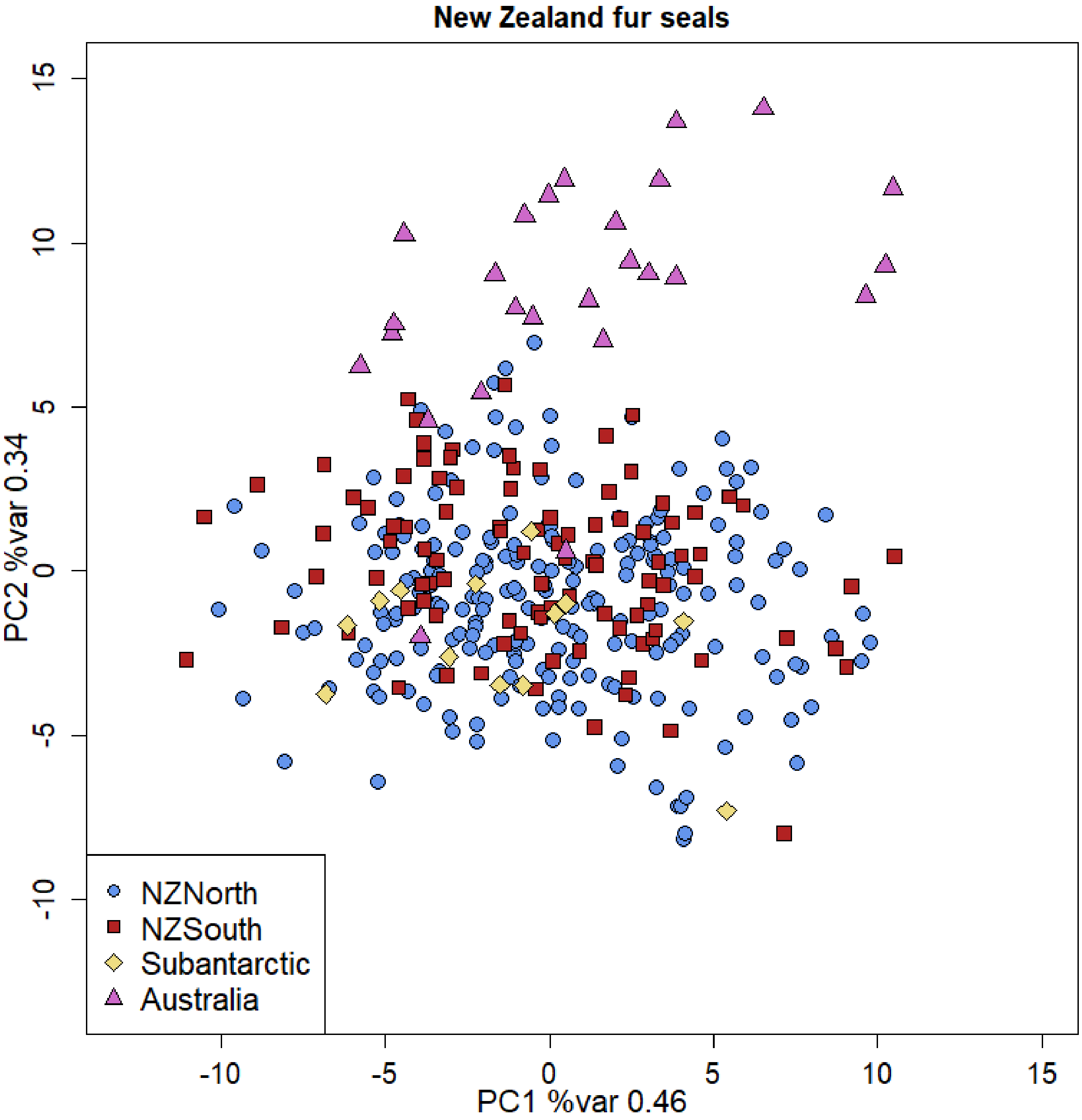


Figure 3.7: GenePlot based on microsatellite data extracted from New Zealand fur seals (*Arctocephalus forsteri*) by Dussex et al. (2016). Each point represents an individual seal from the NZ North ($n = 246$), NZ South ($n = 98$), Subantarctic ($n = 13$) and Australian ($n = 26$) regions identified by Dussex et al. (2016). The axes show the first and second principal components of the log-genotype probabilities with respect to the four regions. The LGPs were calculated using the leave-one-out method.

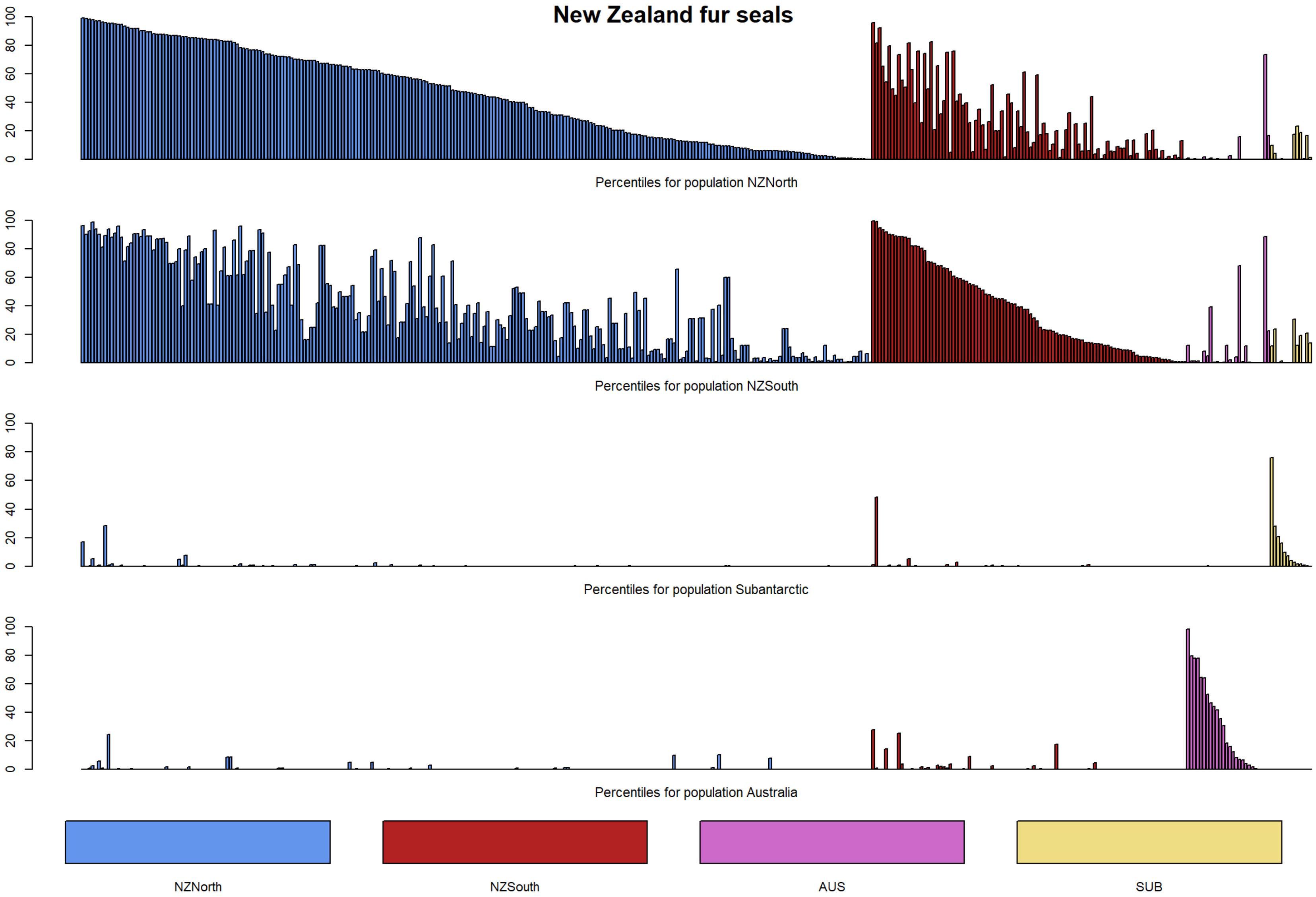


Figure 3.8: GenePlot based on microsatellite data extracted from New Zealand fur seals (*Arctocephalus forsteri*) by Dussex et al. (2016). The multi-bar GenePlot shows one bar chart per population and one bar for each seal from the NZ North ($n = 246$), NZ South ($n = 98$), Subantarctic ($n = 13$) and Australian ($n = 26$) regions identified by Dussex et al. (2016). 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 four bar charts. The LGPs were calculated using leave-one-out.

Figure 3.9 shows the GenePlot and corresponding LGP distribution plots for the NZ North and South regions. This also confirms the findings of Dussex et al. that there is only weak signal of population structure between the two regions. However, there is more overlap between the populations with baseline NZ South (the left hand LGP distribution plot) than with baseline NZ North (the bottom LGP distribution plot), and the permutation test results shown in Figure 3.10 indicate the same directional population structure.

The overlap area and interloper detection probability are non-significant for NZ South, indicating that fur seals from NZ North colonies arriving in NZ South would be indistinguishable from seals originating in NZ South based on this evidence. However, the overlap area and interloper detection probability are significant for NZ North, indicating that seals from NZ South would be less well fitted to NZ North than seals originating in NZ North.

Dussex et al. (2016) used the software DIYABC (Cornuet et al. 2010) to simulate recolonization scenarios. They tested three main scenarios: (i) the Australian population and later the Subantarctic population split from the NZ populations, then the NZ South population was extirpated by commercial sealing and recently recolonized from the NZ North and Subantarctic populations; (ii) the Australian population and later the Subantarctic populations split from the NZ populations, then both the NZ North and NZ South populations were extirpated by commercial sealing and recolonized from refugia on the west coast of New Zealand; (iii) the Australian population split from the Subantarctic and NZ populations, then the Subantarctic and NZ North and NZ South populations all split and experienced no further bottlenecks.

The authors also considered subscenarios of scenario (ii): (iia) NZ North and NZ South were recolonized from refugia on the west coast of the NZ North region; (iib) NZ North and NZ South were recolonized from refugia on the west coast of the NZ-South region; (iic) NZ North and NZ South were recolonized from admixed refugia in the NZ North and NZ South regions.

Dussex et al. (2016) obtained an ambivalent result: the posterior probabilities of all the scenarios (i), (iic) and (iii) were virtually zero, the posterior probability of scenario (iia) was 63% and the posterior probability of scenario (iib) was 37%.

Our result supports scenario (iib) more strongly than scenario (iia). The later-colonized population is more likely to be susceptible to founder effects, so the source population for colonization should possess alleles not found in the later-established population. So we expect the later-established population to fit well into the source population, but we would not expect the source population to fit as well into the later-established population. Here, NZ North fits well into NZ South, but NZ South does not fit so well into NZ North.

The overlap areas for baselines NZ North and South respectively are 0.744 and 0.984. The interloper detection probabilities are 0.678 and 0.496. The correct assignment probabilities are 0.791 and 0.655.

The observed correct assignment probabilities are lower than those found for the ship rats but both are significant, confirming the existence of weak population structure. The results from the overlap area and interloper detection tests yield the same conclusions; both tests are geared towards measuring how well individuals from the comparison population fit into the baseline population, as if we were comparing individuals from 'home' and 'away' sports teams. The correct assignment probability is higher than the interloper detection probability and is significant in both directions: this probability measures the preferential fit of a genotype in its own 'home' population compared with the 'away' population.

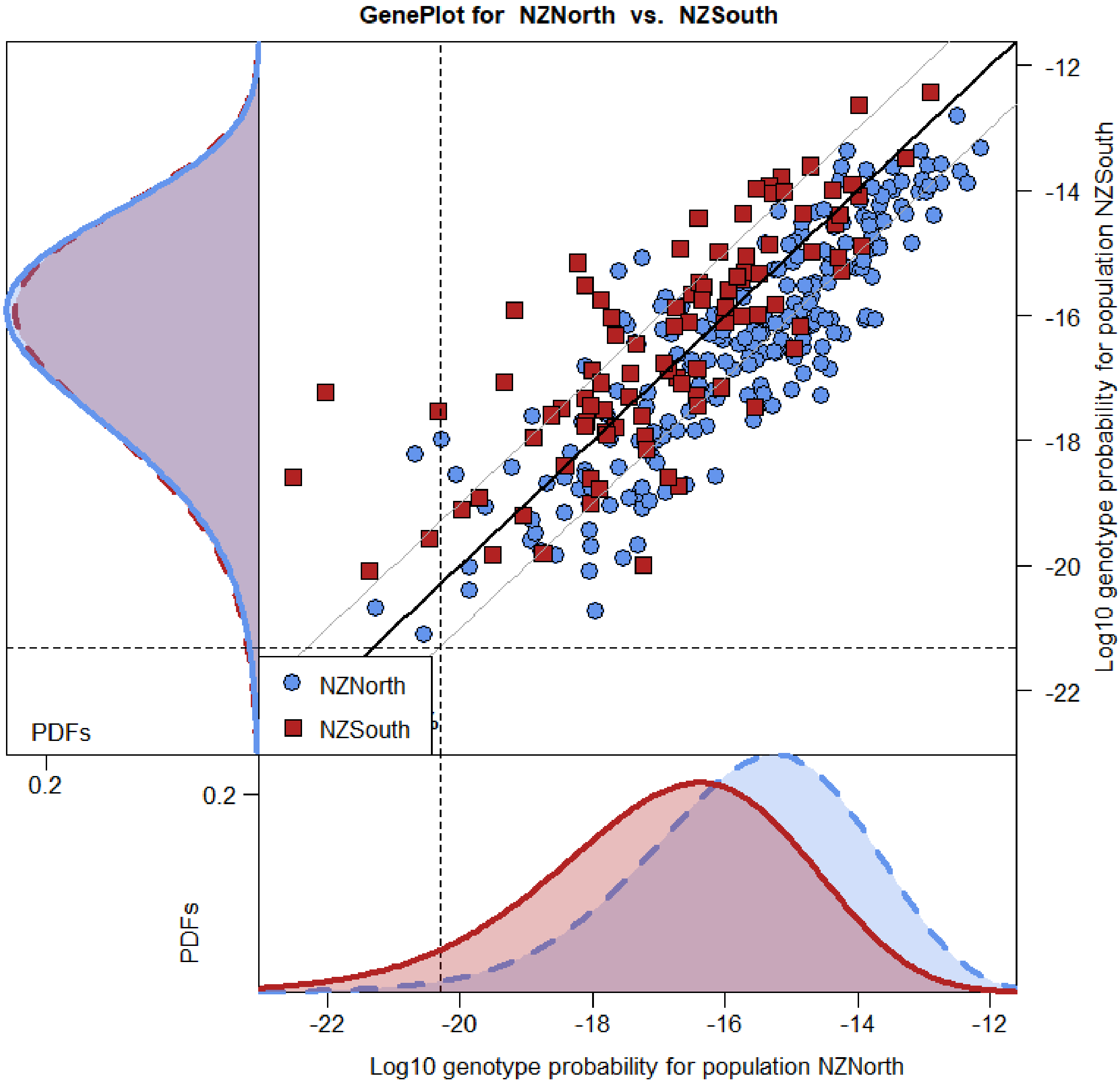


Figure 3.9: GenePlot and LGP distribution plots based on microsatellite data extracted from New Zealand fur seals (*Arctocephalus forsteri*) by Dussex et al. (2016). Each point on the central plot represents an individual seal from the NZ North ($n = 246$) or NZ South ($n = 98$) populations identified by Dussex et al. (2016). The main horizontal axes show the posterior log-genotype probability (LGP) of obtaining each individual's genotype from the NZ North population; the main vertical axes show the same, but with respect to the NZ South population. The thick diagonal line shows equal probability with respect to NZ North and NZ South, and the fine diagonal lines show 10 times higher probability for one population than the other. The vertical dashed line shows the 1% percentile line for NZ North; the horizontal line shows the 1% percentile line for NZ South. The secondary plot to the left of the main plot shows the LGP distribution of the NZ South population (dashed line) and the cross-population LGP distribution of the NZ North population within the NZ South population (solid line). The secondary plot below the main plot shows the LGP distribution of the NZ North population (dashed line) and the cross-population LGP distribution of the NZ South population within the NZ North population (solid line). The LGPs and the dashed curves were calculated using the leave-one-out method.

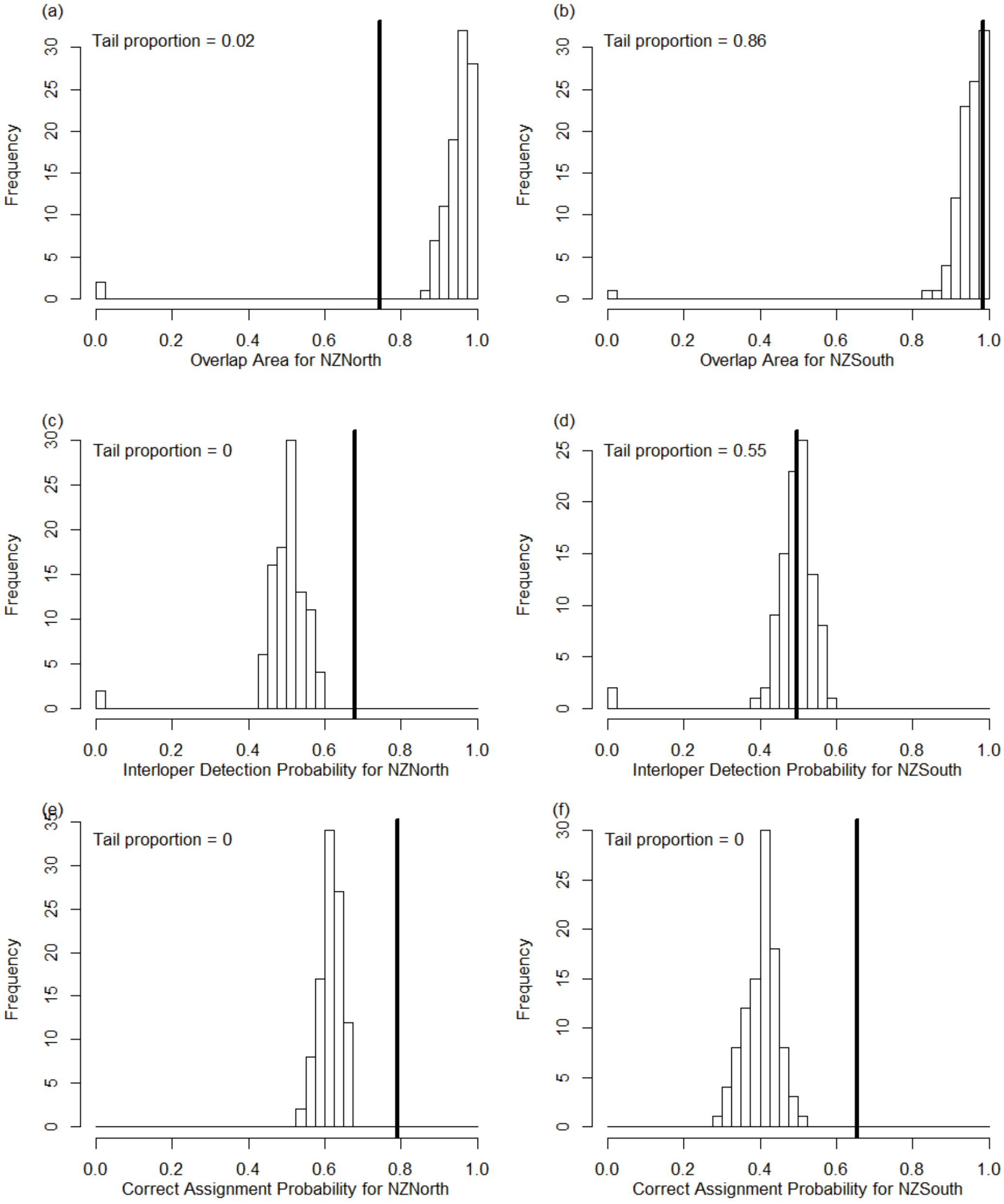


Figure 3.10: Permutation test results based on microsatellite data extracted from NZ fur seals (*Arctocephalus forsteri*), in the NZ North ($n = 246$) and NZ South ($n = 98$) regions defined in Dussex et al. (2016). In each plot, the bold line indicates the observed value and the histogram shows values obtained from 100 datasets in which the population labels were randomly reassigned to all the seals. (a) Overlap area between NZ North LGP distribution and NZ South LGP distribution with respect to NZ North. (b) Overlap area between NZ South LGP distribution and NZ North LGP distribution with respect to NZ South. (c) Probability that for a random pair of genotypes arising from NZ North and the NZ South populations, the one arising from NZ North will have a better fit to the NZ North population. (d) Probability that of those two genotypes, the one arising from NZ South will have a better fit to the NZ South population. (e) Probability that a random genotype arising from NZ North will have a better fit to the NZ North population than to the NZ South population. (f) Probability that a random genotype arising from NZ South will have a better fit to the NZ South population than the NZ North population. The single-population LGP distributions were constructed using leave-one-out.

3.6 Application to data from southern right whales (*Eubalaena australis*)

Microsatellite data at 17 loci were obtained from southern right whales (*Eubalaena australis*). Figure 3.11 shows the GenePlot of the six southern right whale populations identified in Carroll et al. (2018). Five are nursing grounds in Argentina, South Africa, southwestern and southeastern Australia, and New Zealand, and one is the migratory corridor near Australia. Like the previous example of New Zealand fur seals, we deliberately selected this example to trial our methods in a case where population structure is subtle and might or might not exist.

Figure 3.11 shows that the Argentinian and South African populations overlap and the Australian, New Zealand and Australian migratory corridor populations overlap, but there is still visible structure between these two regions.

Figure 3.12 shows the GenePlot and LGP distribution plots for the southern right whale nursing ground populations, pooled as the South Atlantic population (Argentina and South Africa) and the Indo-Pacific population (southwestern Australia, southeastern Australia and New Zealand). These groupings follow Carroll et al. (2018). This plot shows more clearly that there is some population structure between these two regions, with the South Atlantic whales mostly found below the diagonal line and the Indo-Pacific whales mostly found above the diagonal line.

The overlap areas for baselines South Atlantic and Indo-Pacific are 0.727 and 0.704 respectively. The interloper detection probabilities are 0.690 and 0.705. The correct assignment probabilities are 0.861 and 0.836. This indicates that there is some population structure present, and that most of the time we would be able to distinguish between individuals from the two populations or assign an individual to the correct population. Figure 3.13 shows the permutation test results for these populations, which are all significant, confirming the evidence of population structure.

Figure 3.14 shows the GenePlot and LGP distribution plots for the Argentinian and South African populations. As hinted in Figure 3.11, they are strongly overlapping, though there is less overlap with baseline Argentina (bottom plot) than with baseline South Africa (left-hand plot). The overlap areas for Argentina and South Africa are 0.754 and 0.992 respectively. The interloper detection probabilities are 0.668 and 0.499. The correct assignment probabilities are 0.689 and 0.655. This indicates that there is weak directional population structure between these two populations. Figure 3.15 shows the permutation test results for these populations. Argentina is significantly separated from South Africa, but South Africa is not significantly separated from Argentina, and many whales from the Argentina population would fit well into the South African population.

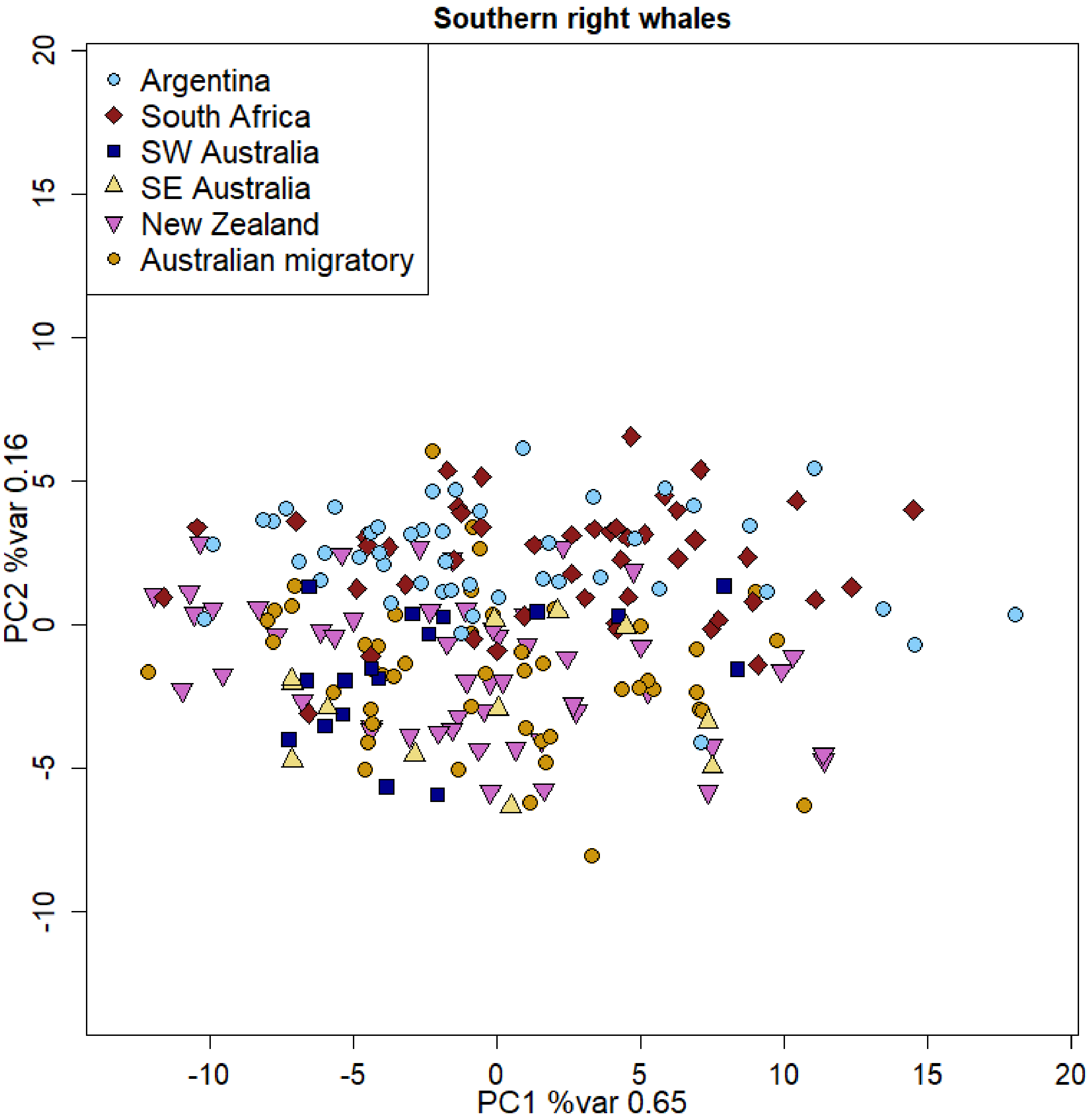


Figure 3.11: GenePlot based on microsatellite data extracted from southern right whales (*Eubalaena australia*) by Carroll et al. (2018). Each point represents an individual whale from the Argentina (n = 46), South Africa (n = 47), southwestern Australia (n = 17), southeastern Australia (n = 12) or New Zealand (n =51) nursing grounds or the Australian migratory corridor (n =49) identified by Carroll et al. (2018). The axes show the first and second principal components of the log-genotype probabilities with respect to the six populations. The LGPs were calculated using the leave-one-out method.

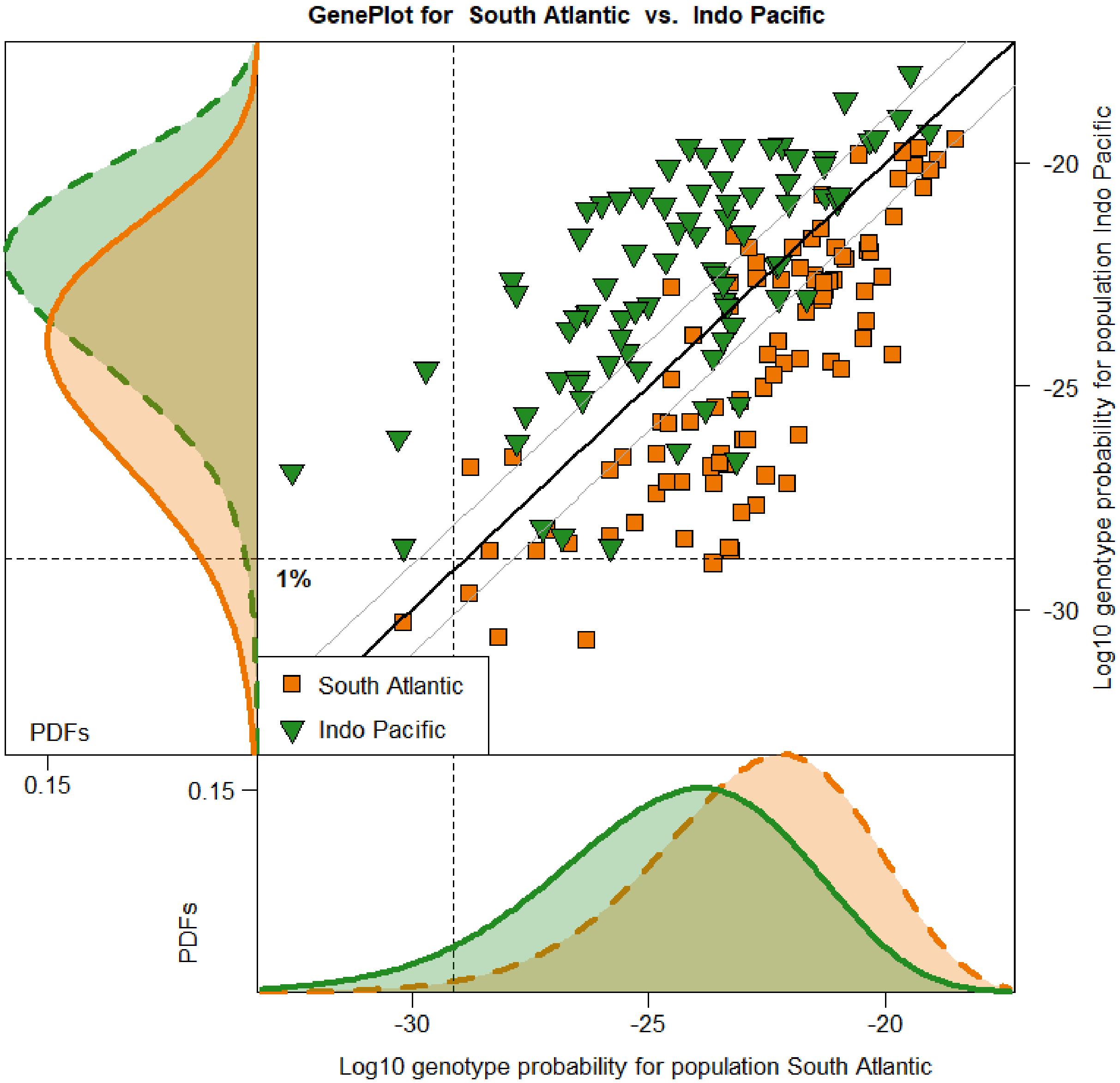


Figure 3.12: GenePlot based on microsatellite data extracted from southern right whales (*Eubalaena australia*) by Carroll et al. (2018). Each point on the central plot represents an individual whale from the South Atlantic (n = 93) or Indo-Pacific (n =80) regions identified by Carroll et al. (2018). The main horizontal axes show the posterior log-genotype probability (LGP) of obtaining each individual's genotype from the South Atlantic population; the main vertical axes show the same, but with respect to the Indo-Pacific population. The thick diagonal line shows equal probability with respect to both populations, and the fine diagonal lines show 10 times higher probability for one population than the other. The vertical dashed line shows the 1% percentile line for South Atlantic; the horizontal line shows the 1% percentile line for Indo-Pacific. The secondary plot to the left of the main plot shows the LGP distribution of the Indo-Pacific population (dashed line) and the cross-population LGP distribution of the South Atlantic population within the Indo-Pacific population (solid line). The secondary plot below the main plot shows the LGP distribution of the South Atlantic population (dashed line) and the cross-population LGP distribution of the Indo-Pacific population within the South Atlantic population (solid line). The LGPs and the dashed curves were calculated using the leave-one-out method.

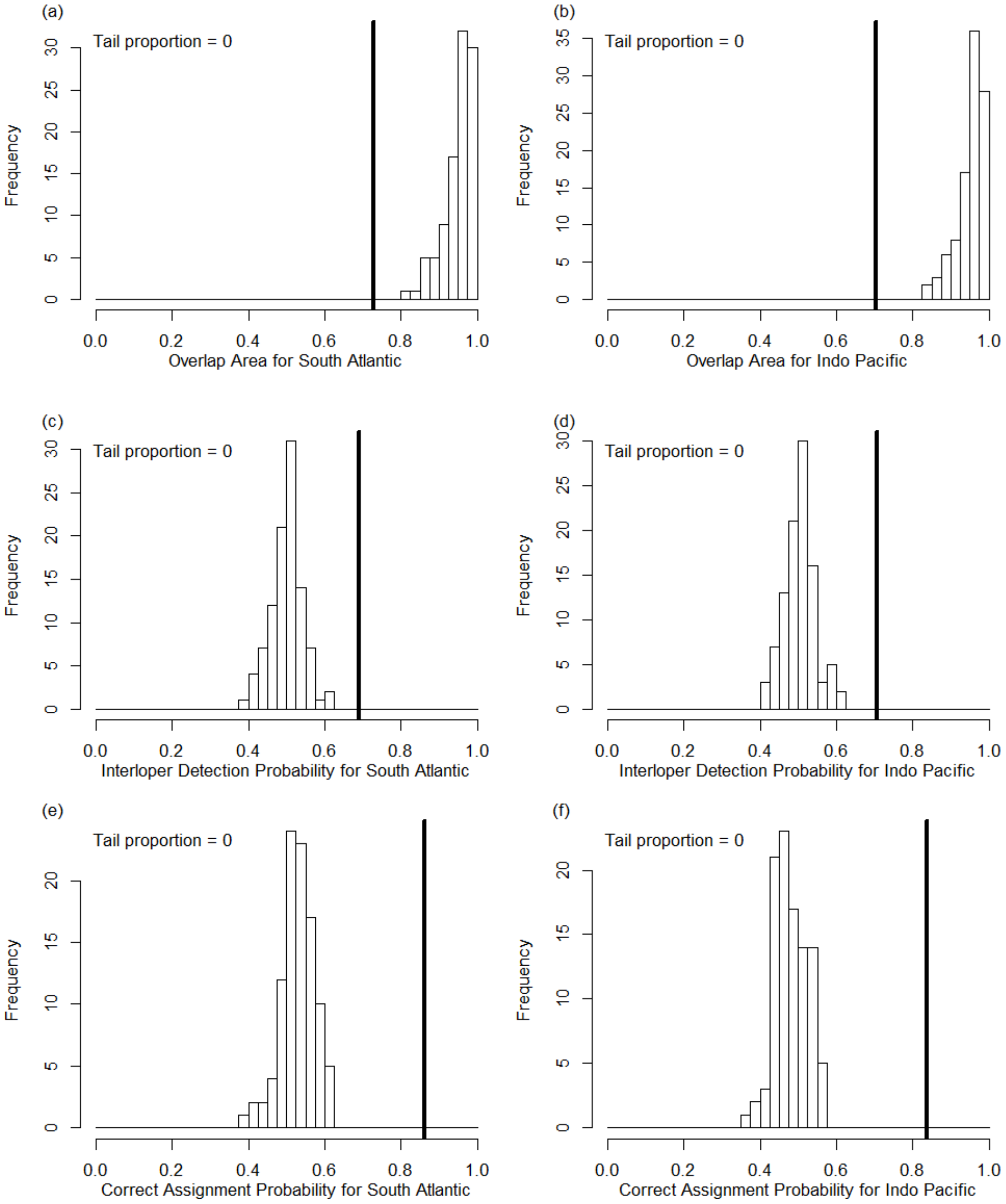


Figure 3.13: Permutation test results based on microsatellite data extracted from southern right whales (*Eubalaena australis*), in the South Atlantic (n =93) and Indo-Pacific (n =80) regions described in Carroll et al. (2018). In each plot, the bold line indicates the observed value and the histogram shows values obtained from 100 datasets in which the population labels were randomly reassigned to all the whales. (a) Overlap area between South Atlantic LGP distribution and LGP distribution of Indo-Pacific with respect to South Atlantic. (b) Overlap area between Indo-Pacific LGP distribution and LGP distribution of South Atlantic with respect to Indo-Pacific. (c) Probability that for a random pair of genotypes arising from South Atlantic and the Indo-Pacific populations, the one arising from South Atlantic will have a better fit to the South Atlantic population. (d) Probability that of those two genotypes, the one arising from Indo-Pacific will have a better fit to the Indo-Pacific population. (e) Probability that a random genotype arising from South Atlantic will have a better fit to the South Atlantic population than to the Indo-Pacific population. (f) Probability that a random genotype arising from Indo-Pacific will have a better fit to the Indo-Pacific population than to the South Atlantic population. The single-population LGP distributions were constructed using leave-one-out.

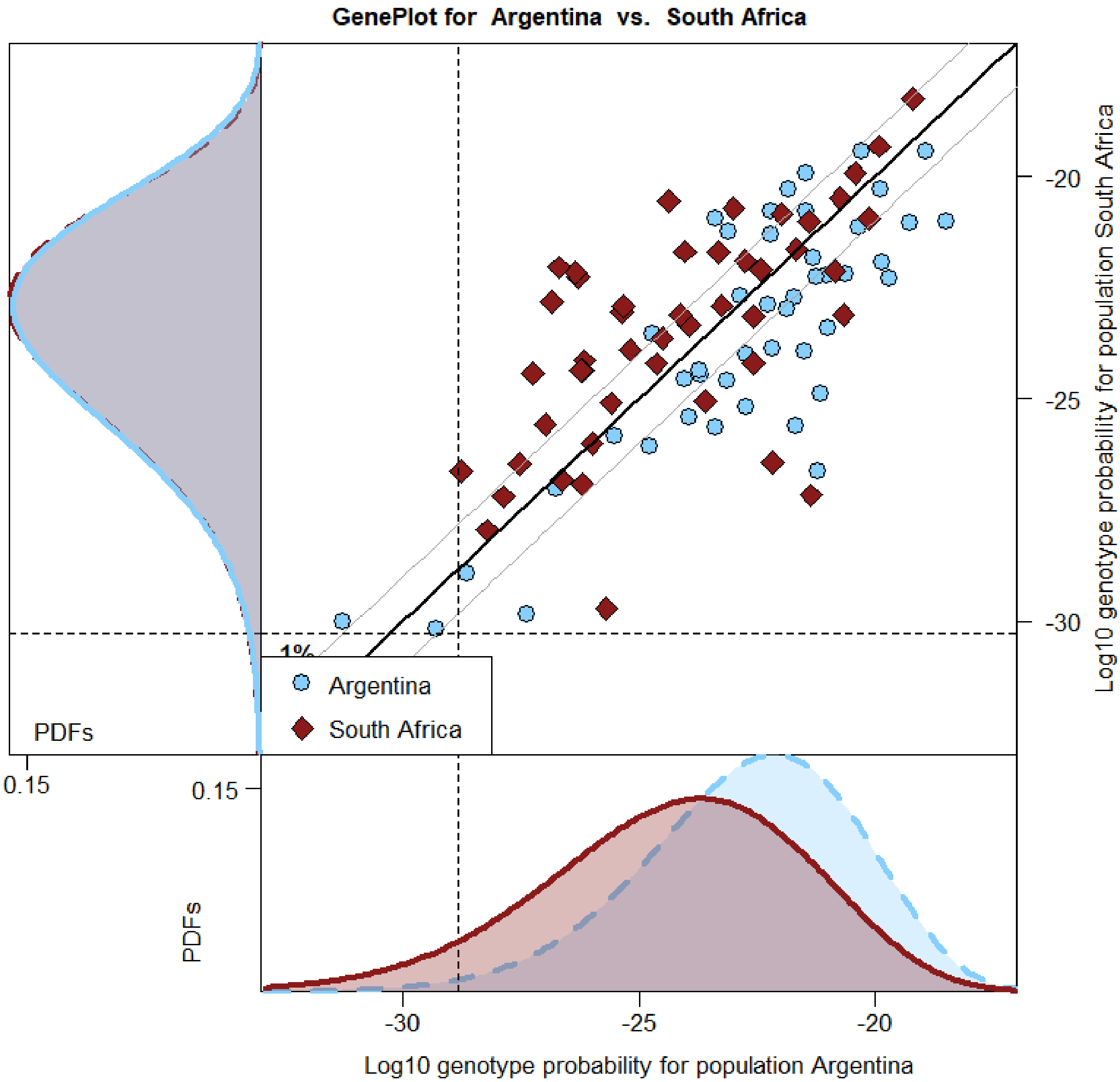


Figure 3.14: GenePlot and LGP distribution plots based on microsatellite data extracted from southern right whales (*Eubalaena australis*) by Carroll et al. (2018). Each point on the central plot represents an individual whale from the Península Valdés, Argentina ($n = 46$) or South African ($n = 47$) nursing ground populations identified by Carroll et al. (2018). Details as in Figure 3.12.

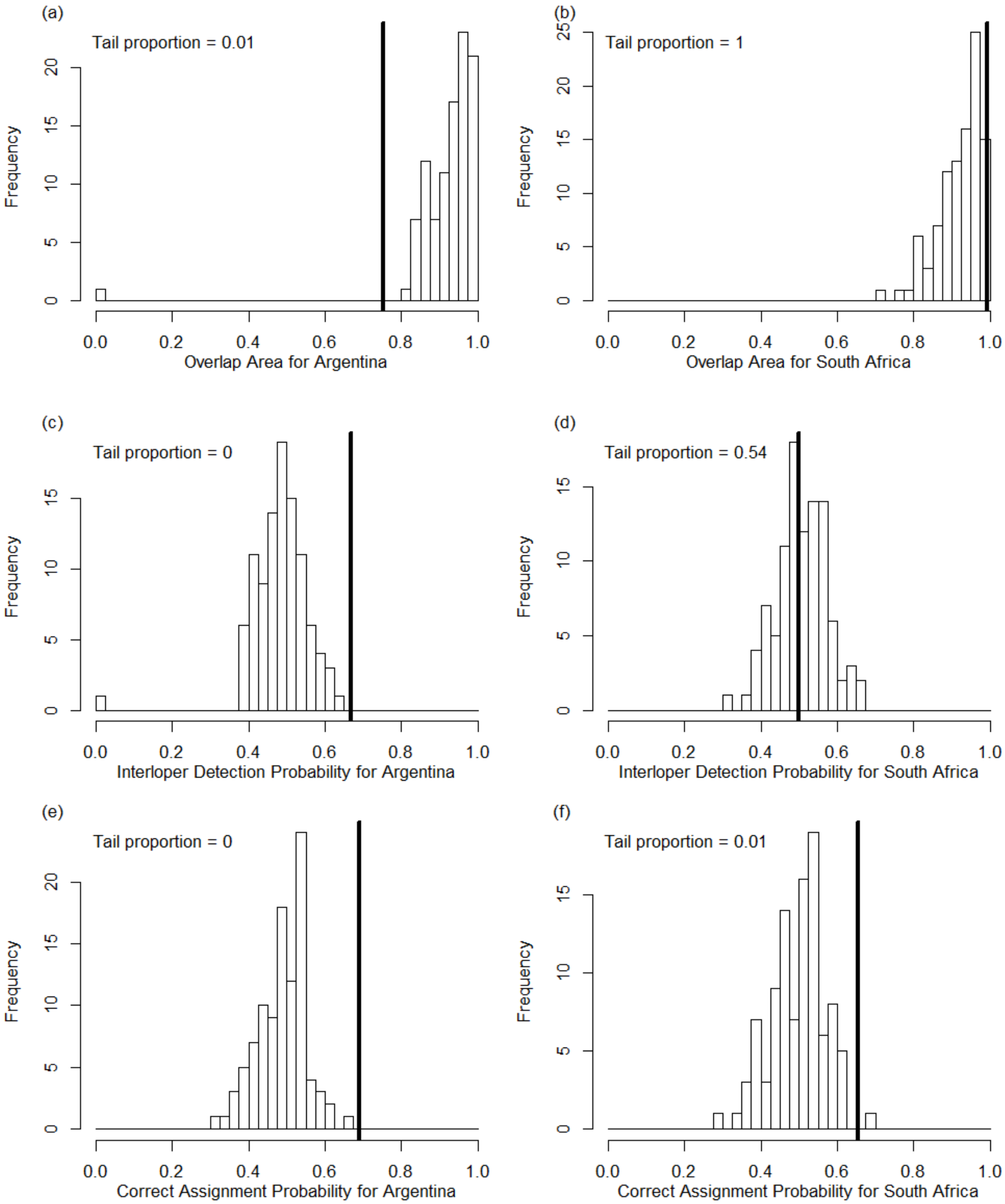


Figure 3.15: Permutation test results based on microsatellite data extracted from southern right whales (*Eubalaena australis*), in the Argentinian ($n = 46$) and South African ($n = 47$) nursing grounds described in Carroll et al. (2018). Details as in Figure 3.13.

For a further example, Figure 3.16 shows the GenePlot and LGP distribution plots for southern right whale data at 13 microsatellite loci from mainland New Zealand published in Carroll et al. (2011), Carroll et al. (2012) and Carroll et al. (2014), and from the Auckland Islands described in Carroll et al. (2013).

Before 19th century whaling almost extirpated them, there was a large breeding population of southern right whales around the southern coast of mainland New Zealand. The subantarctic Auckland Islands population was a refugium during the whaling era, and is now growing healthily (Carroll et al. 2013). Over the last few years a few southern right whales have been sighted around the mainland New Zealand coast, very occasionally with calves. Researchers are interested in whether these occasional sightings are sourced from the Auckland Islands population, potentially signalling a recolonization of mainland New Zealand from the Auckland Islands population.

These two populations are fully overlapping, as anticipated. The overlap areas for the Auckland Islands and Mainland NZ respectively are 0.941 and 0.849 respectively. The interloper detection probabilities are 0.539 and 0.399. The correct assignment probabilities are 0.775 and 0.125.

The smaller magnitude of the correct assignment probability for Mainland NZ is most likely due to the large disparity between the sample sizes of the two populations; as the larger Auckland Islands sample encompasses a much higher level of genetic diversity, we would expect most individuals from either population to have a better fit to the Auckland Islands population than the mainland New Zealand population, even in the absence of any genetic differentiation. This is confirmed by the location of the null distributions in Figure 3.17, which shows the permutation test results for the Auckland Islands and Mainland NZ populations. The null distribution for the correct assignment probability is located substantially lower for Mainland NZ than it is for the Auckland Islands, and the observed correct assignment probability of 0.125 lies within the null distribution.

All six measures are non-significant for this data, implying that there is no evidence of population separation by any measure. This is consistent with the hypothesis that Mainland NZ whales are migrants from the Auckland Islands population or that the Mainland NZ population was recently founded from the Auckland Islands population.

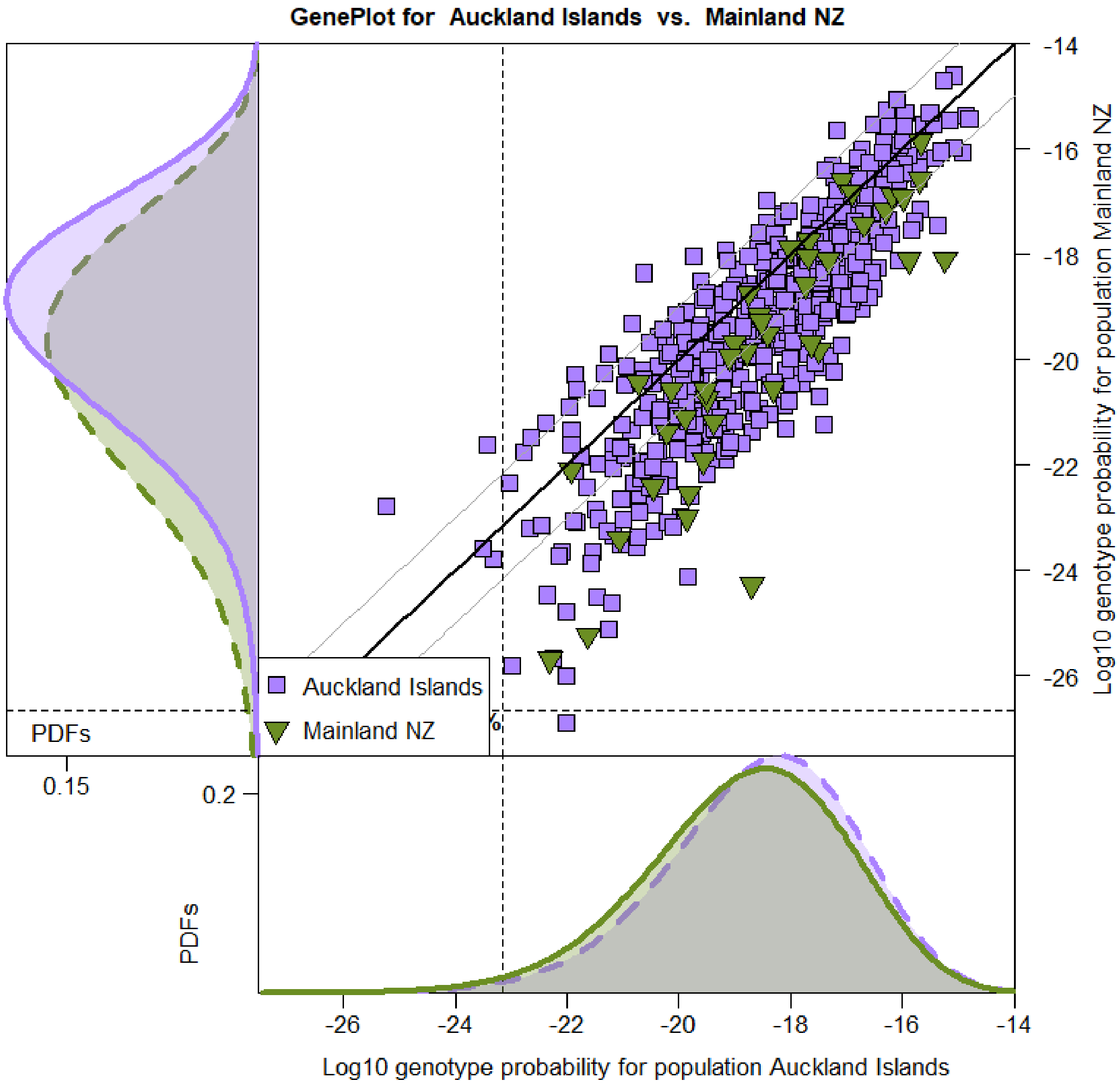


Figure 3.16: GenePlot and LGP distribution plots based on microsatellite data extracted from southern right whales (*Eubalaena australis*) (Carroll et al. 2011, 2012, 2013, 2014). Each point on the central plot represents an individual whale from the Auckland Islands ($n = 516$) or Mainland New Zealand ($n = 41$) populations. Details as in Figure 3.12.

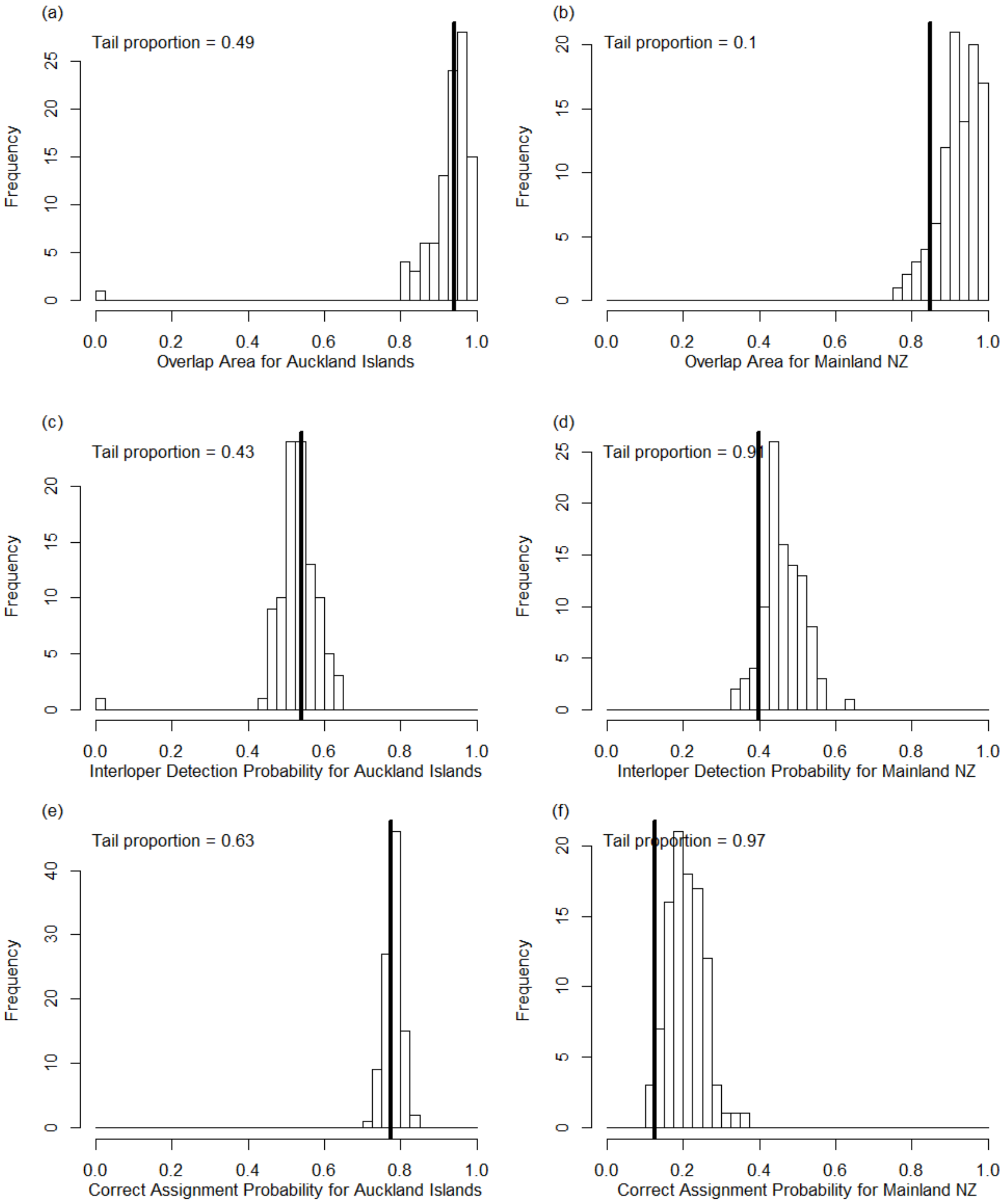


Figure 3.17: Permutation test results based on microsatellite data extracted from southern right whales (*Eubalaena australis*), in the Auckland Islands ($n = 516$) and Mainland NZ ($n = 41$) populations (Carroll et al. 2011, 2012, 2013, 2014). Details as in Figure 3.13.

3.7 Application to data from single nucleotide polymorphisms (SNPs)

GenePlots and the probabilistic measures we propose can also be applied to large numbers of single nucleotide polymorphism (SNP) loci as for microsatellites, though with higher computational cost. We demonstrate this using SNPs from human genomes (The 1000 Genomes Consortium, 2015), with files downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`.

We selected biallelic loci that were at least 1Mbps apart and with minor allele frequency larger than 0.05. Then we removed any loci which were correlated with any other loci on the same chromosome with $r^2 \geq 0.1$ and any loci which were correlated with 2 or more other loci with $r^2 \geq 0.05$, where the correlations were calculated separately within each of the 5 super-populations (Americas, Europe, Africa, South Asia and East Asia). We then had 2317 remaining SNP loci for analysis. We used the VCFtools software (Danecek et al, 2011) to extract the SNP loci and calculate r^2.

We have used leave-one-out computations throughout, and note that it is even more important to use leave-one-out for large numbers of SNP loci than for small numbers of microsatellite loci, because as the number of loci increases, the effect of each observed genotype on the sample increases, and even slight differences in the allele frequencies between the two populations are enough to create the spurious appearance of separation between the populations when not using leave-one-out.

Figures 3.18 and 3.19 show the GenePlot of the five super-populations defined by The 1000 Genomes Consortium (2015). Two of the populations, African Caribbean in Barbados (ACB) and People with African Ancestry in Southwest USA (ASW), may be labelled as belonging to either the African or American super-populations. Figure 3.18 shows them labelled in the American super-population, and Figure 3.19 shows them labelled in the African super-population.

We selected three pairs of populations to demonstrate our methodology with highly overlapping and highly differentiated populations. Figure 3.20 shows the GenePlot of the African populations excluding the ACB and ASW populations. The remaining populations are Yoruba in Ibadan, Nigeria (YRI); Esan in Nigeria (ESN); Gambian in Western Division, Mandinka (GWD); Mende in Sierra Leone (MSL); and Luhya in Webuye, Kenya (LWK). We picked two of those populations which show very low levels of differentiation.

Figure 3.21 shows the GenePlot and LGP distribution plots for the Yoruba (YRI) and Esan (ESN) populations from Nigeria, both within the African superpopulation. This is an example of a pair of strongly overlapping populations, with almost entirely overlapping LGP curves. The overlap areas with baselines Yoruba and Esan are 0.998 and 0.928 respectively. The interloper detection probabilities are 0.498 and 0.551, which is close to the value of 0.5 that we would get if we simply picked one of the two hypothetical individuals purely at random each time. The correct assignment probabilities are 0.688 and 0.576.

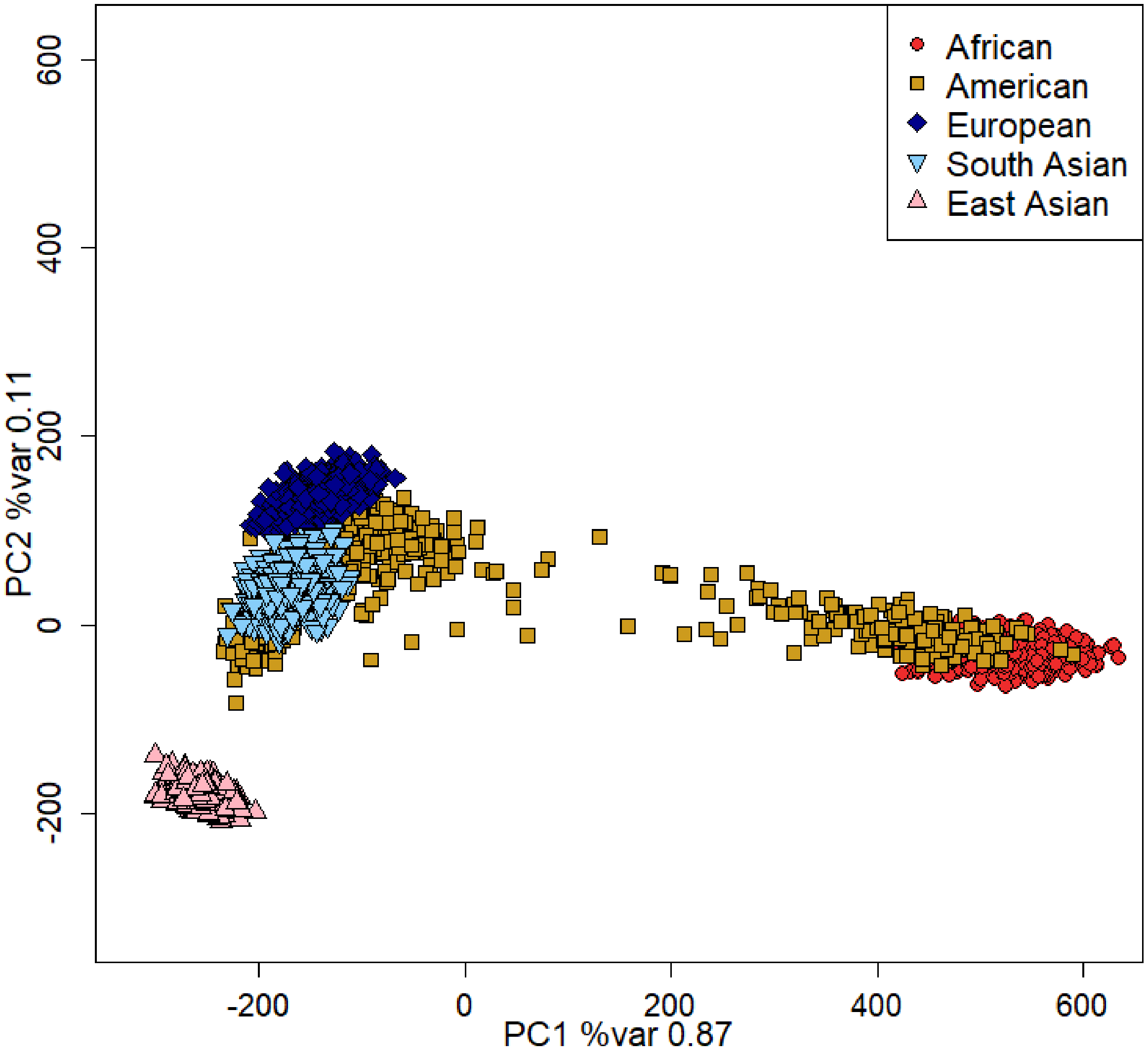


Figure 3.18: GenePlot based on SNP data from human genomes (The 1000 Genomes Consortium, 2015). Each point represents an individual from the African (n =504), American (n =504), European (n =503), South Asian (n =489) or East Asian (n =504) super-populations, with individuals from the Barbados (ACB) and African Ancestry US (ASW) populations labelled in the American super-population. The axes show the first and second principal components of the log-genotype probabilities with respect to the five populations. The LGPs were calculated using the leave-one-out method.

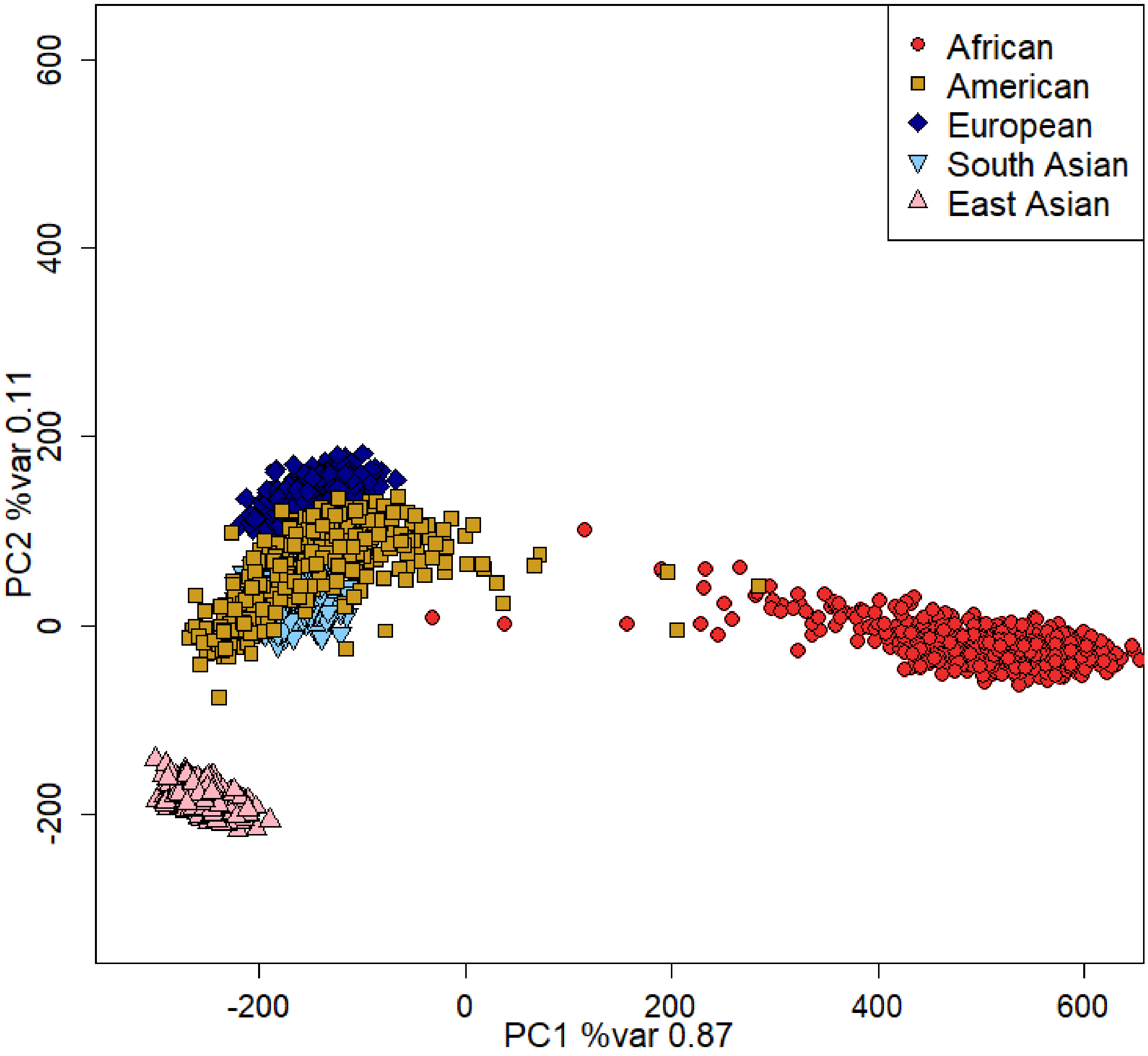


Figure 3.19: GenePlot based on SNP data from human genomes (The 1000 Genomes Consortium, 2015). Each point represents an individual from the African (n =661), American (n =347), European (n =503), South Asian (n =489) or East Asian (n =504) super-populations, with individuals from the Barbados (ACB) and African Ancestry US (ASW) populations labelled in the African super-population. The axes show the first and second principal components of the log-genotype probabilities with respect to the five populations. The LGPs were calculated using the leave-one-out method.

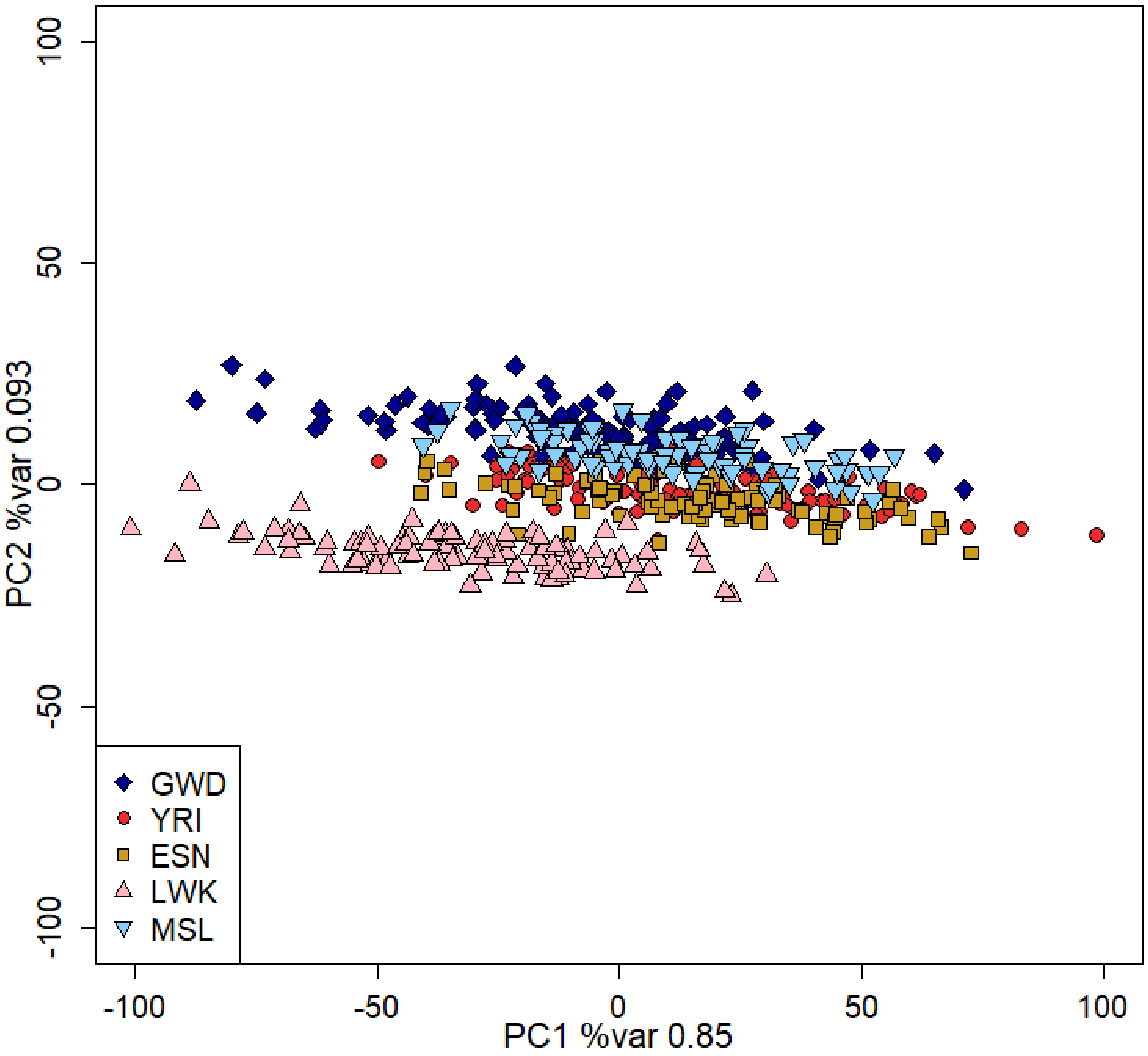


Figure 3.20: GenePlot based on SNP data from human genomes (The 1000 Genomes Consortium, 2015). Each point represents an individual from the Yoruba in Ibadan, Nigeria (YRI, $n = 108$); Esan in Nigeria (ESN, $n = 99$); Gambian in Western Division, Mandinka (GWD, $n = 113$); Mende in Sierra Leone (MSL, $n = 85$); or Luhya in Webuye, Kenya (LWK, $n = 99$) populations in the African superpopulation, excluding individuals from the Barbados (ACB) and African Ancestry US (ASW) populations. The axes show the first and second principal components of the log-genotype probabilities with respect to the five populations. The LGPs were calculated using the leave-one-out method.

Using a more conservative assignment protocol that assigns an individual only if their genotype probability is 100 times higher in one population than the other, the correct assignment probabilities for Yoruba and Esan respectively are 0.439 and 0.325. The fact that these are so much lower than the best-fit assignment probabilities shows that many hypothetical individuals from either population would not be assigned under the more conservative protocol, due to having a similar fit to both populations.

Figure 3.22 shows the GenePlot of the East Asian populations: Japanese in Tokyo, Japan (JPT); Kinh in Ho Chi Minh City, Vietnam (KHV); Chinese Dai in Xishuangbanna, China (CDX); Han Chinese in Beijing, China (CHB); and Southern Han Chinese (CHS). In general we can see that these populations are more differentiated than the African populations in Figure 3.20. We picked two pairs of these East Asian populations which show very high amounts of differentiation.

Figure 3.23 shows the GenePlot and LGP distribution plots for the Japanese (JPT) and Kinh Vietnamese (KHV) populations, both within the East Asian superpopulation, and Figure 3.24 shows the plots for the Japanese and Chinese Dai in Xishuangbanna (CDX) populations. These are examples of pairs of strongly separated populations.

The overlap areas for the Japanese and Kinh Vietnamese populations are 0.302 and 0.327 respectively. The interloper detection probabilities are 0.928 and 0.917. The correct assignment probabilities are both 1.000. Even using the more conservative assignment protocol that assigns an individual only if their genotype probability is 100 times higher in one population than the other, the correct assignment probabilities for Japanese and Kinh Vietnamese respectively are still both 1.000.

The overlap areas for the Japanese and Chinese Dai in Xishuangbanna populations are 0.232 and 0.224 respectively. The interloper detection probabilities are 0.954 and 0.957. The correct assignment probabilities are both 1.000. Even using the more conservative assignment protocol that assigns an individual only if their genotype probability is 100 times higher in one population than the other, the correct assignment probabilities for Japanese and Chinese Dai respectively are still both 1.000.

The overlap areas for these pairs of strongly separated populations are much higher than the overlap area of Aotea within the Broken Islands in Figure 3.1. The overlap area here indicates that the populations are genetically quite similar, but by taking 2317 SNPs we gain the power to discriminate convincingly between them, as demonstrated by the interloper detection probabilities and correct assignment probabilities, which are all close to 1.

Figure 3.25 shows permutation test results for the Yoruba (YRI) and Esan (ESN) populations from Nigeria, the most strongly overlapping out of all pairs of populations. We used 500 permutations for this dataset, because some of the results for 100 permutations were marginally significant. The overlap area probabilities are non-significant for both populations. The interloper detection probability is non-significant for YRI and weakly significant for ESN. Both of the correct assignment probabilities are significant.

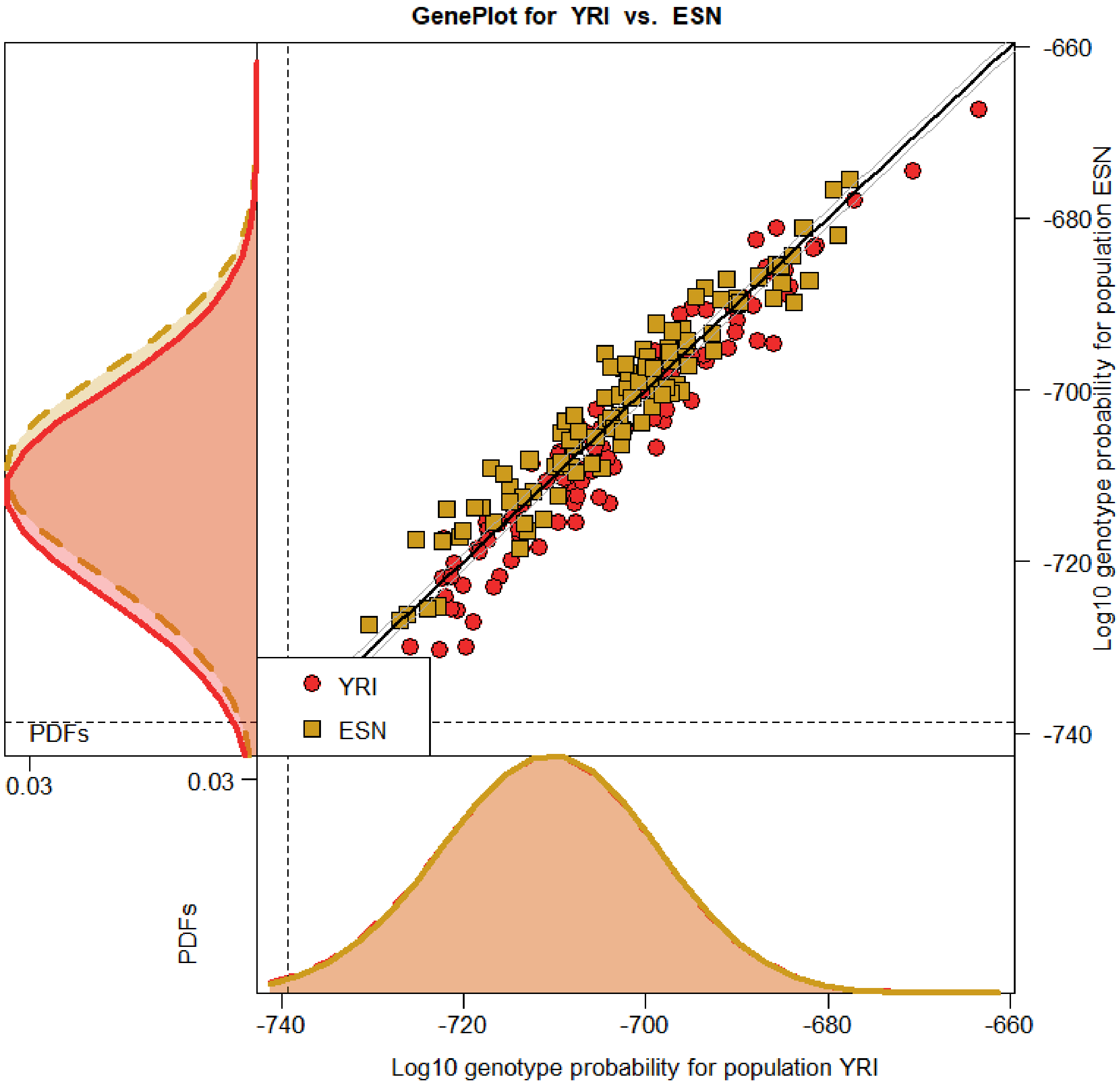


Figure 3.21: GenePlot and LGP distribution plots based on SNP data from the 1000 Genomes Consortium (2015). Each point on the central plot represents an individual from the Yoruba (YRI, $n = 108$) or Esan (ESN, $n = 99$) populations from Nigeria. The main horizontal axes show the posterior log-genotype probability (LGP) of obtaining each individual's genotype from the Yoruba population; the main vertical axes show the same, but with respect to the Esan population. The thick diagonal line shows equal probability with respect to both populations, and the fine diagonal lines show 10 times higher probability for one population than the other. The vertical dashed line shows the 1% percentile line for Yoruba; the horizontal line shows the 1% percentile line for Esan. The secondary plot to the left of the main plot shows the LGP distribution of the Esan population (dashed line) and the cross-population LGP distribution of the Yoruba population within the Esan population (solid line). The secondary plot below the main plot shows the LGP distribution of the Yoruba population (dashed line) and the cross-population LGP distribution of the Esan population within the Yoruba population (solid line). The LGPs and the dashed curves were calculated using the leave-one-out method.

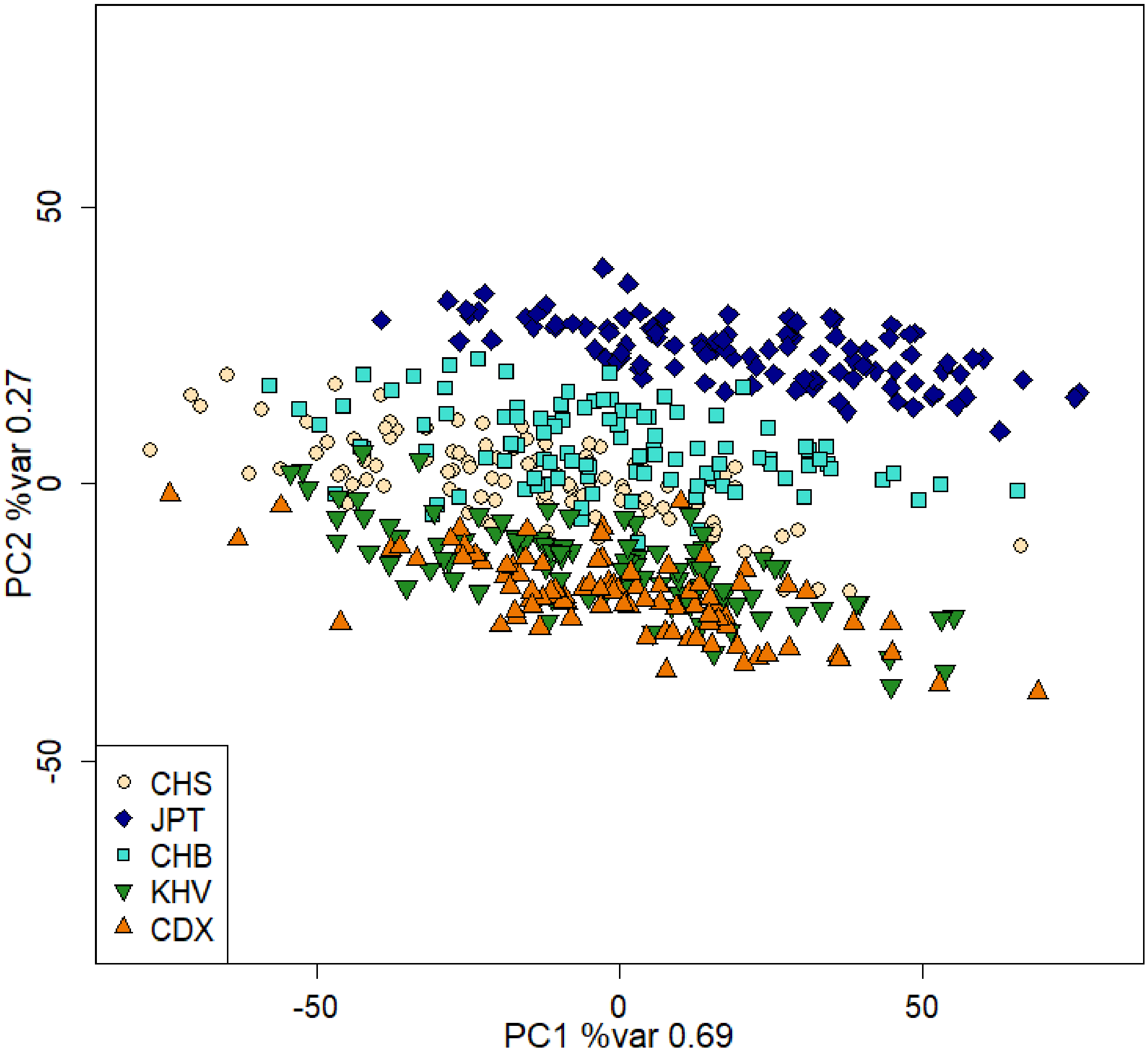


Figure 3.22: GenePlot based on SNP data from human genomes (The 1000 Genomes Consortium, 2015). Each point represents an individual from Japanese in Tokyo, Japan (JPT, $n = 104$); Kinh in Ho Chi Minh City, Vietnam (KHV, $n = 99$); Chinese Dai in Xishuangbanna, China (CDX, $n = 93$); Han Chinese in Beijing, China (CHB, $n = 103$); and Southern Han Chinese (CHS, $n = 105$) populations in the East Asian super-population. The axes show the first and second principal components of the log-genotype probabilities with respect to the five populations. The LGPs were calculated using the leave-one-out method.

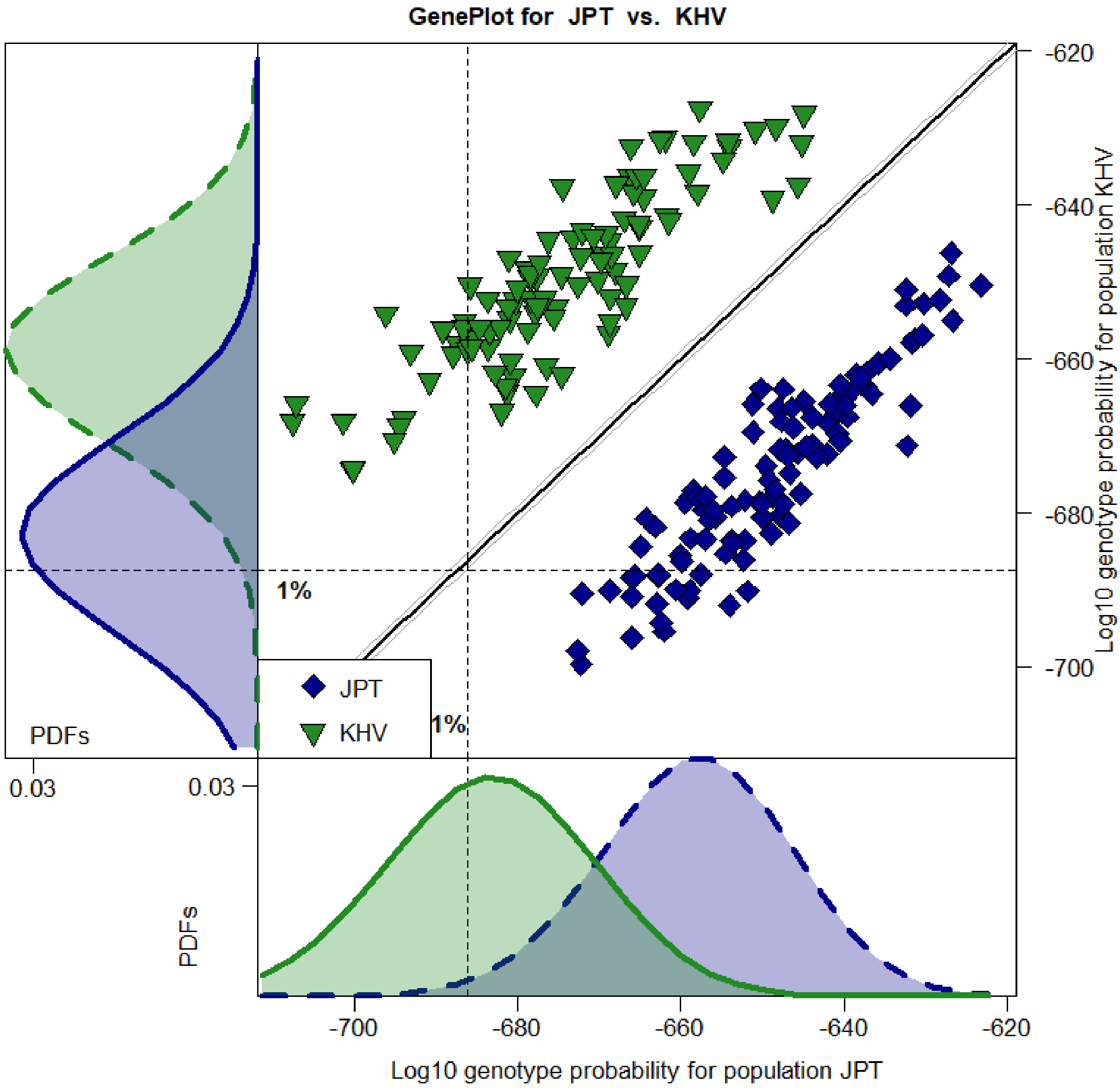


Figure 3.23: GenePlot and LGP distribution plots based on SNP data from the 1000 Genomes Consortium (2015). Each point on the central plot represents an individual from the Japanese (JPT, $n = 104$) or Kinh Vietnamese (KHV, $n = 99$) populations. Details as in Figure 3.21.

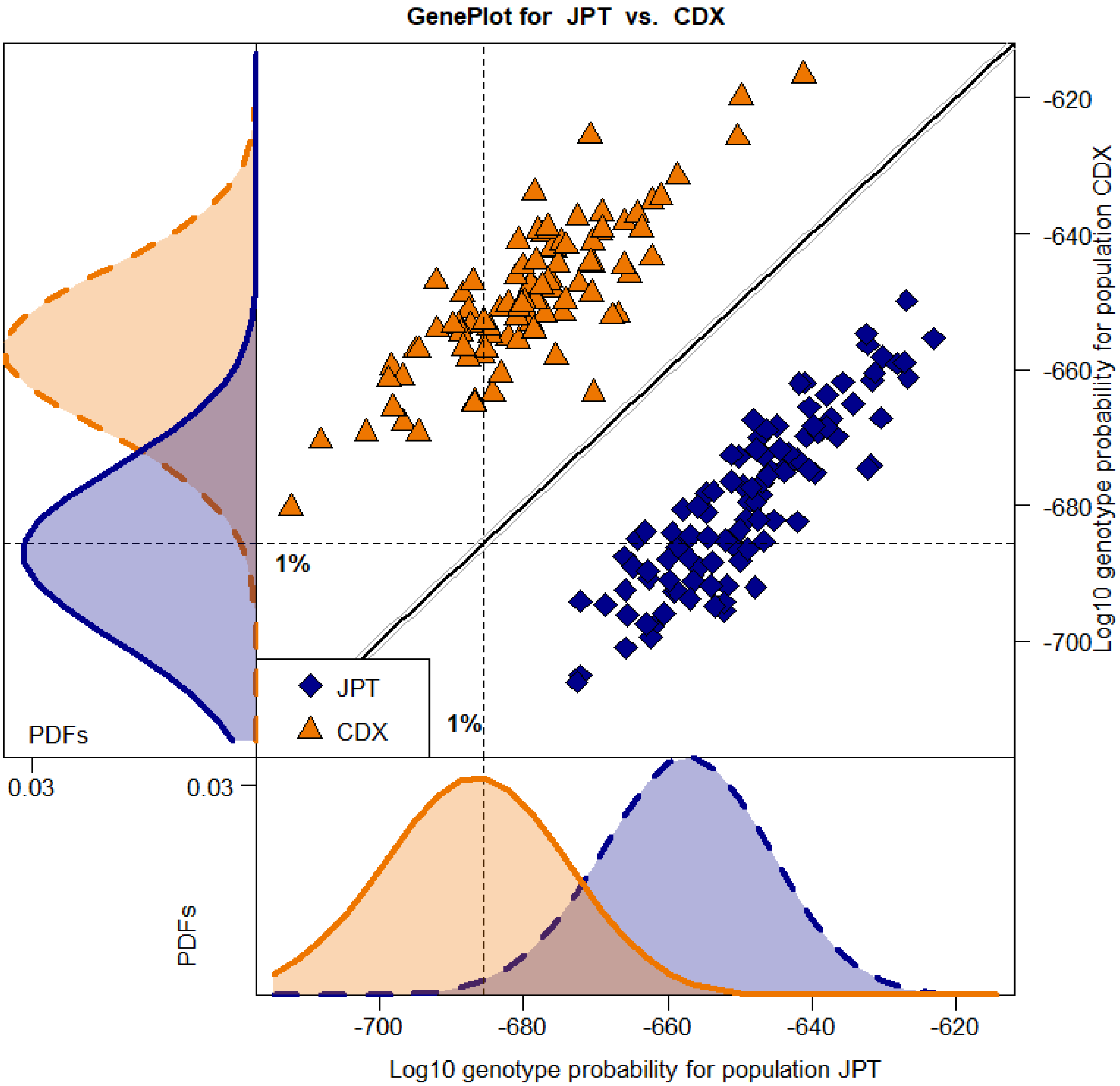


Figure 3.24: GenePlot and LGP distribution plots based on SNP data from the 1000 Genomes Consortium (2015). Each point on the central plot represents an individual from the Japanese (JPT, $n = 104$) or Chinese Dai in Xishuangbanna (CDX, $n = 93$) populations. Details as in Figure 3.21.

Figure 3.26 shows permutation test results for the Japanese (JPT) and Kinh Vietnamese (KHV) populations, which are strongly separated with all measures showing significant results. Figure 3.27 shows similar results for the Japanese and Xishuangbanna Chinese Dai (CDX) populations. These results were based on 100 permutations.

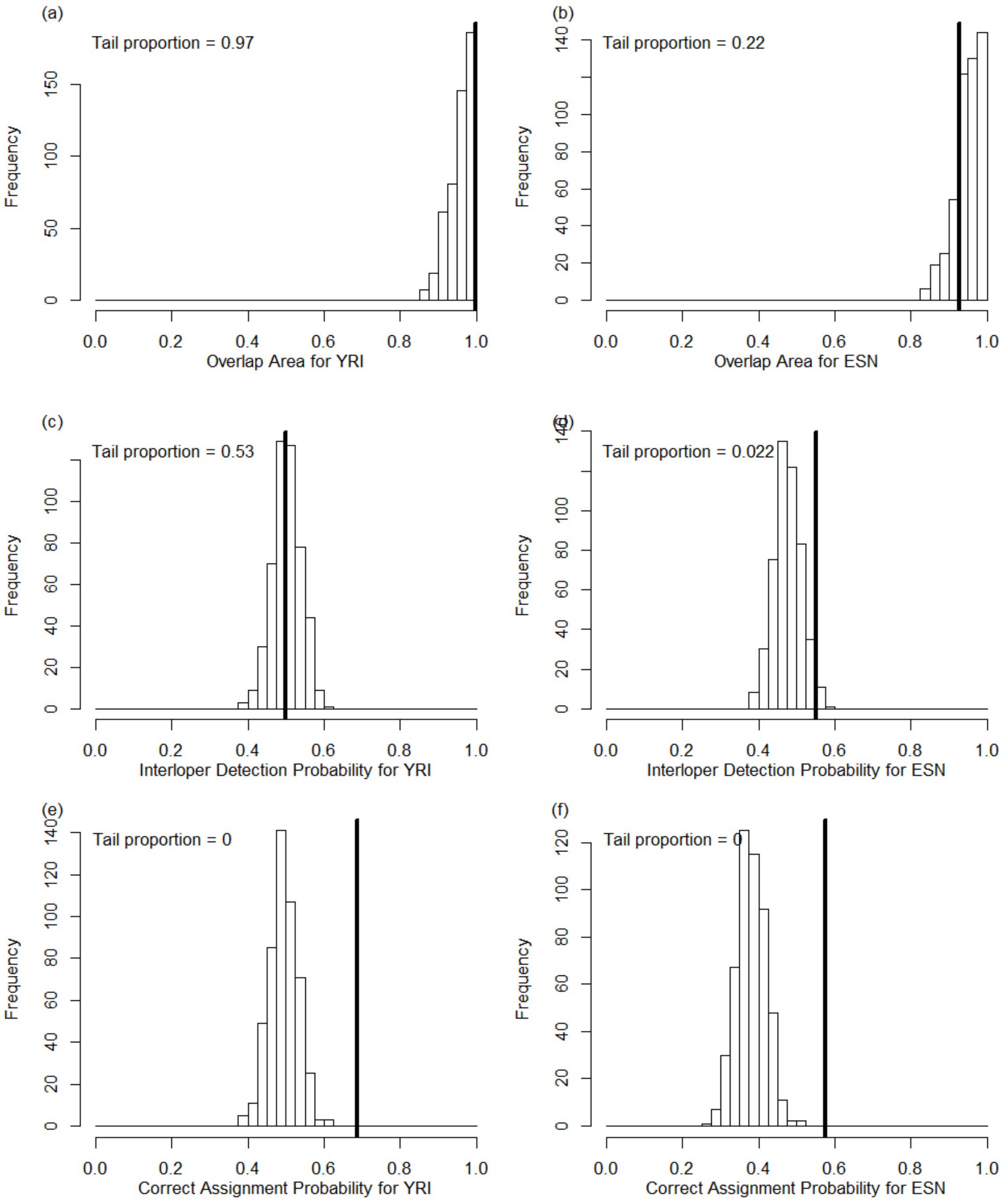


Figure 3.25: Permutation test results based on SNP data from the 1000 Genomes Project, for the Yoruba (YRI, $n = 108$) and Esan (ESN, $n = 99$) populations, both from Nigeria. The tests involved 500 permutations. In each plot, the bold line indicates the observed value and the histogram shows values obtained from 100 datasets in which the population labels were randomly reassigned to all the individuals. (a) Overlap area between Yoruba LGP distribution and LGP distribution of Esan with respect to Yoruba. (b) Overlap area between Esan LGP distribution and LGP distribution of Yoruba with respect to Esan. (c) Probability that for a random pair of genotypes arising from Yoruba and the Esan populations, the one arising from Yoruba will have a better fit to the Yoruba population. (d) Probability that of those two genotypes, the one arising from Esan will have a better fit to the Esan population. (e) Probability that a random genotype arising from Yoruba will have a better fit to the Yoruba population than to the Esan population. (f) Probability that a random genotype arising from Esan will have a better fit to the Esan population than to the Yoruba population. The single-population LGP distributions were constructed using leave-one-out.

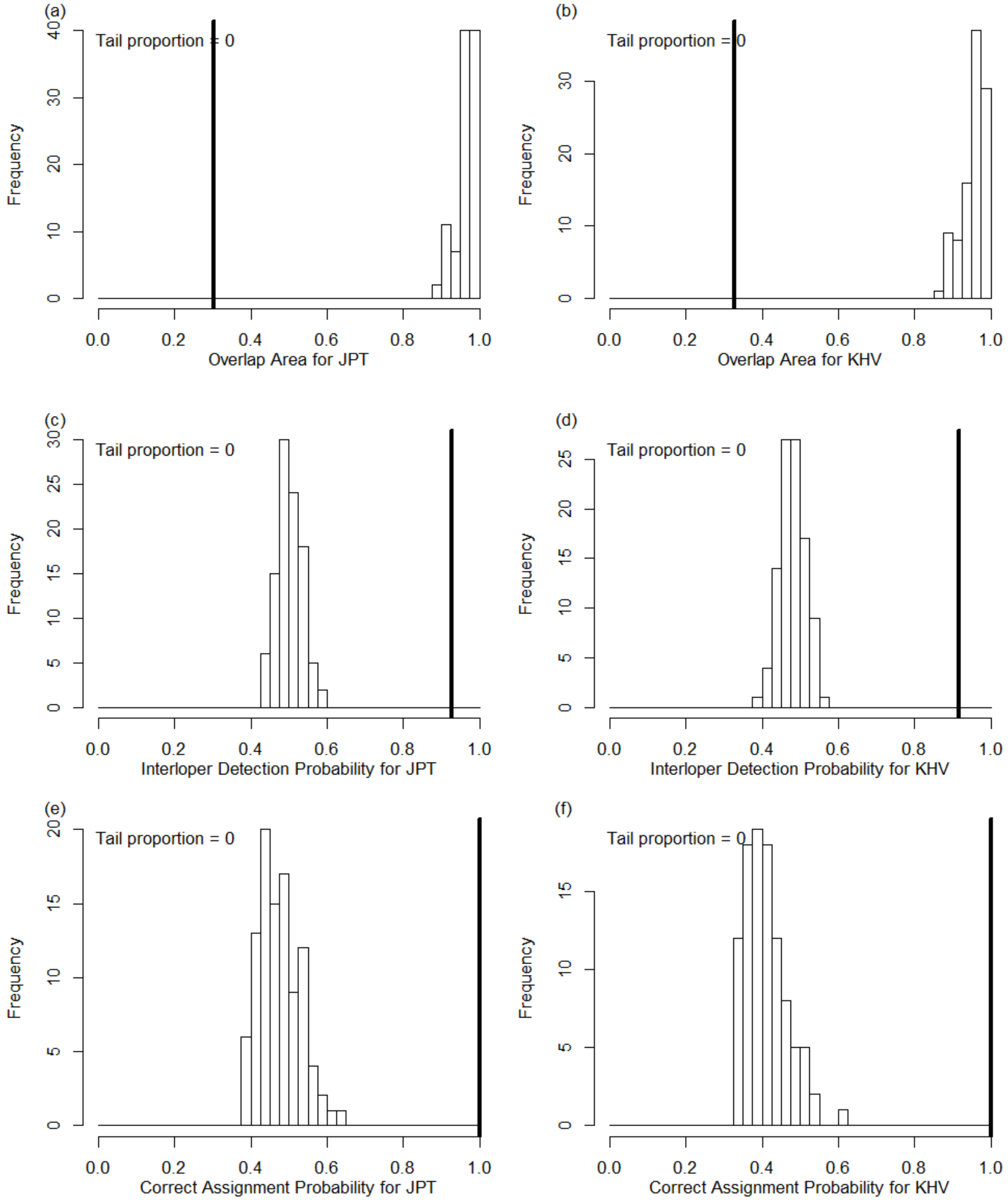


Figure 3.26: Permutation test results based on SNP data from the 1000 Genomes Project, for the Japanese (JPT, $n = 104$) and Kinh Vietnamese (KHV, $n = 99$) populations. Details as in Figure 3.25.

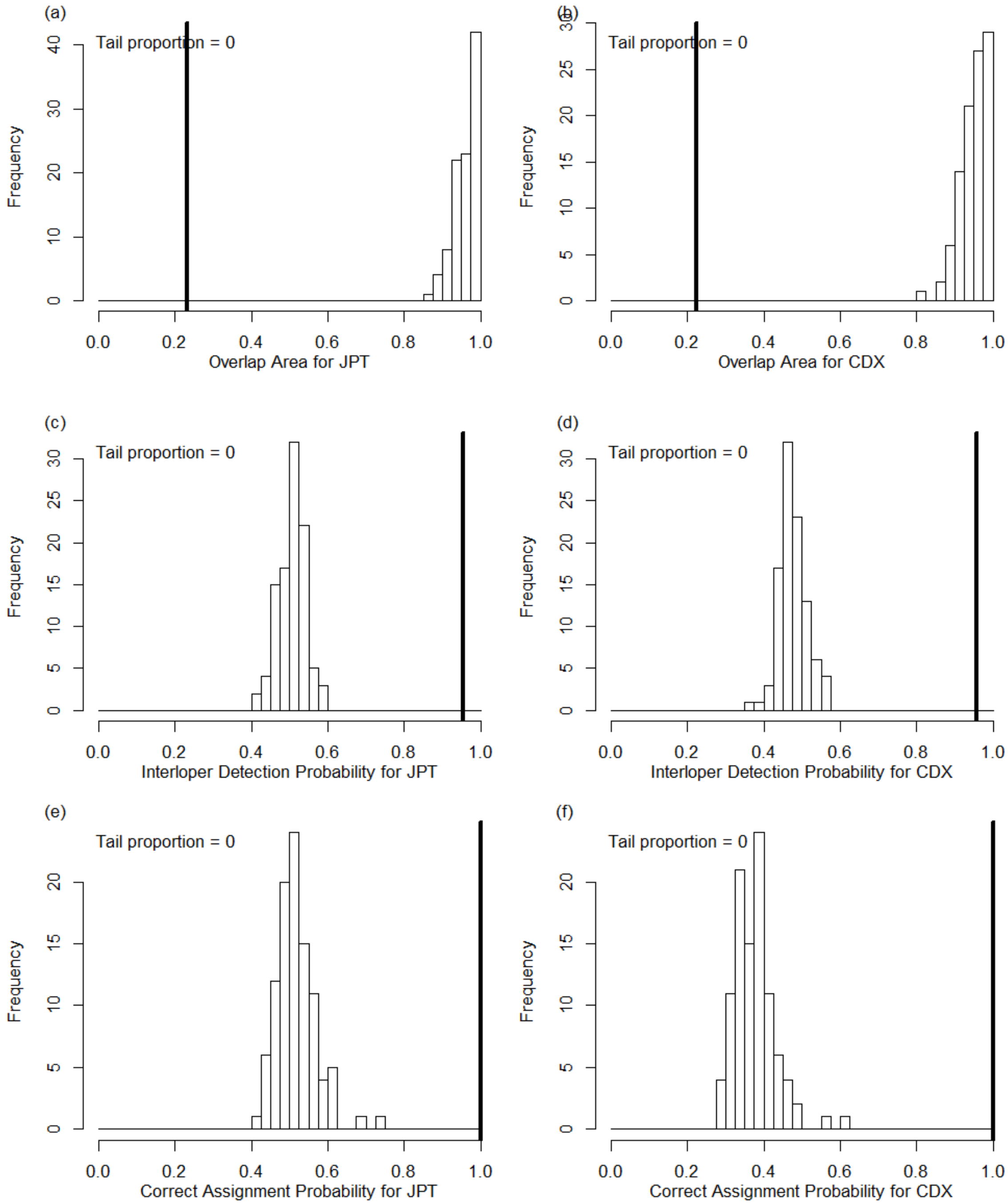


Figure 3.27: Permutation test results based on SNP data from the 1000 Genomes Project, for the Japanese (JPT, $n = 104$) and Chinese Dai in Xishuangbanna (CDX, $n = 93$) populations. Details as in Figure 3.25.

3.8 Simulation study using data from ship rats (*Rattus rattus*) from Aotea

We ran a simulation study using rats from the Aotea ship rat (*Rattus rattus*) sample from Section 3.2, to test the power of the new measures to distinguish between a source population and a splinter population dominated by internal recruitment, compared with the case where substantial ongoing migration occurs from the source into the splinter population.

3.8.1 Methods

We bred a new generation of 40 simulated rats from the Aotea ship rat (*Rattus rattus*) sample ($n = 57$) by selecting 40 pairs of parent rats at random with replacement from the Aotea sample. For each parent we selected one of their alleles at each locus at random, and for any locus at which that parent rat had missing data we selected an allele from that locus at random from the rats in the Aotea dataset that had data at that locus. We then combined the alleles from both parents to form a new individual. The resulting population we labelled as "Founded Population".

For the initial generation we then compared all 40 of those rats with 40 sampled from the Aotea dataset, which were labelled respectively as "Founded Sample" and "Original Sample".

For further generations 1 to 3, we bred n_B simulated rats from the entire previous generation of the Founded Population, and bred n_I simulated rats directly from the original Aotea dataset, and combined the n_B "internally bred" rats and the n_I "immigrant" rats into a new generation of the Founded Population. Then we sampled 40 of that new generation and compared it with a new sample of 40 from the Aotea dataset.

We replicated the process 3 times to assess the amount of sampling variability. Unless otherwise marked, all the results shown are from replicate 1.

3.8.2 Results

Figures 3.28 and 3.29 show combined GenePlots and LGP distribution plots for one replicate of a Founded Population with 20 internally bred individuals and 20 immigrants in each generation. Figure 3.28 shows samples from generation 1 and Figure 3.29 shows samples from generation 3 of the Founded Population. There is a large amount of overlap in the LGP distribution plots for Figures 3.28 and 3.29.

Figures 3.30 and 3.31 show the permutation test results from the same samples as in Figures 3.28 and 3.29, which are all non-significant. We obtained similar results from the other 2 replicates. This implies that if half of the population in each generation are migrants then the Founded Population will maintain sufficient genetic similarity with the original Aotea population for the two samples to be indistinguishable, even after three generations of internal breeding and immigration.

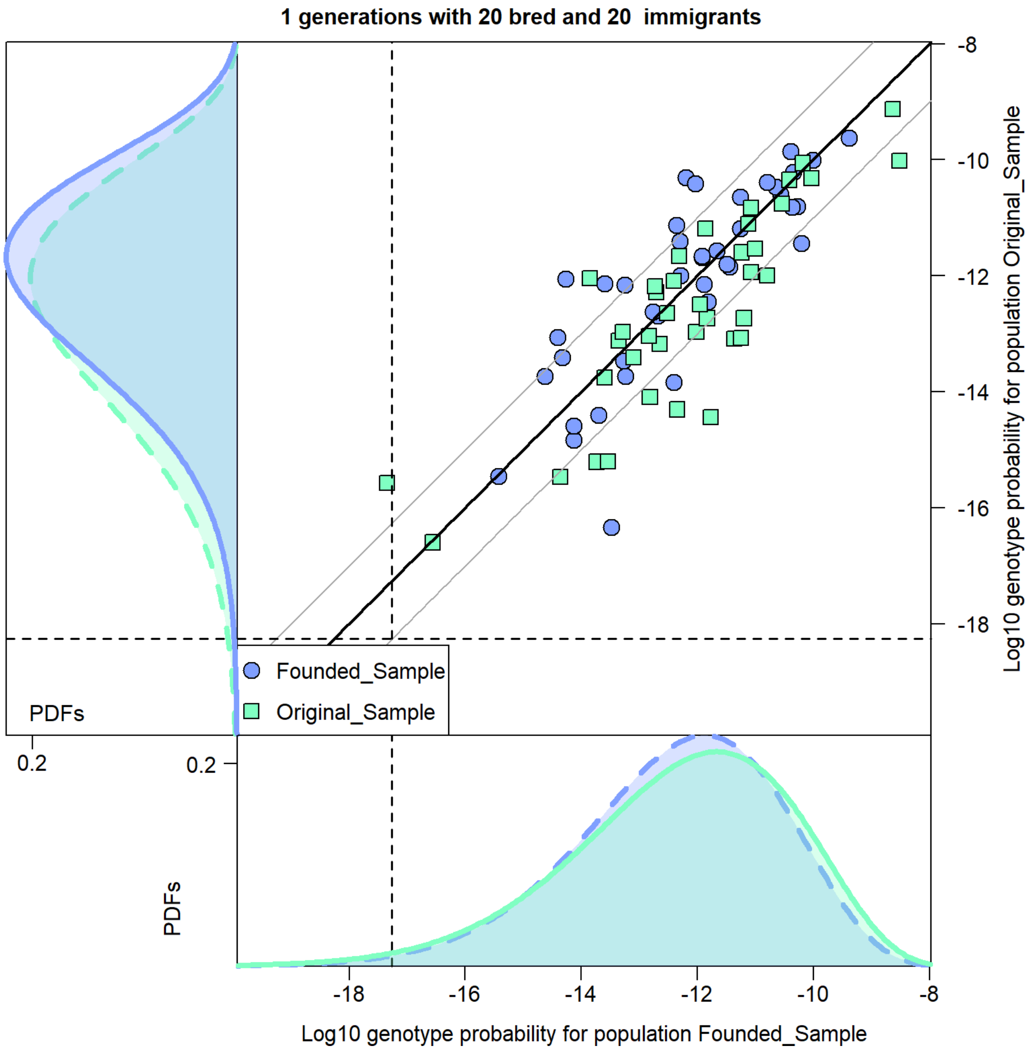


Figure 3.28: GenePlot and LGP distribution plots based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 1 generation, where the Founded Population was made up of 20 internally bred rats and 20 immigrant rats. Details as in Figure 3.1.

We obtained similar results for a Founded population made up of 30 internally bred individuals and 10 immigrants in each generation; after 3 generations, the permutation tests all returned non-significant results for all 3 replicate runs. All the permutation tests were based on 100 randomisations.

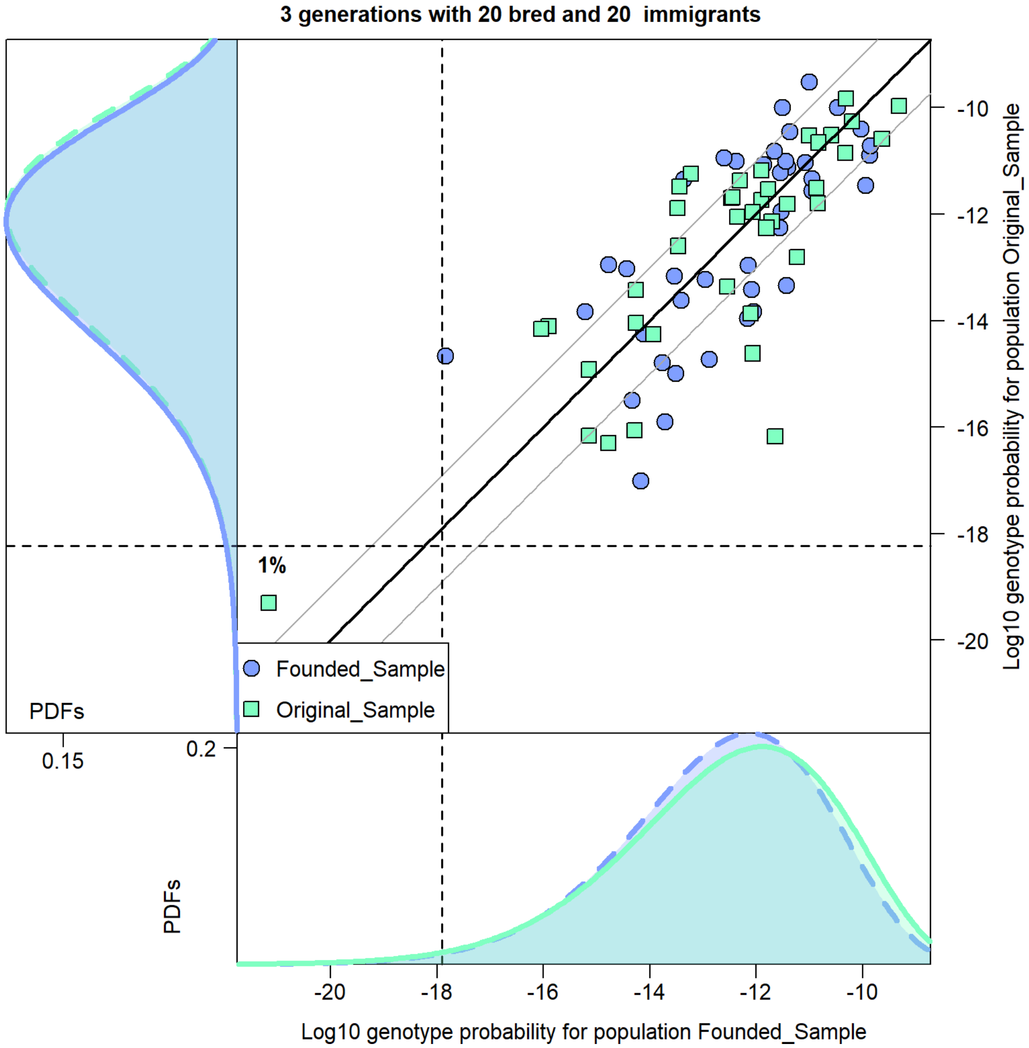


Figure 3.29: GenePlot and LGP distribution plots based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 3 generations, where the Founded Population was made up of 20 internally bred rats and 20 immigrant rats. Details as in Figure 3.1.

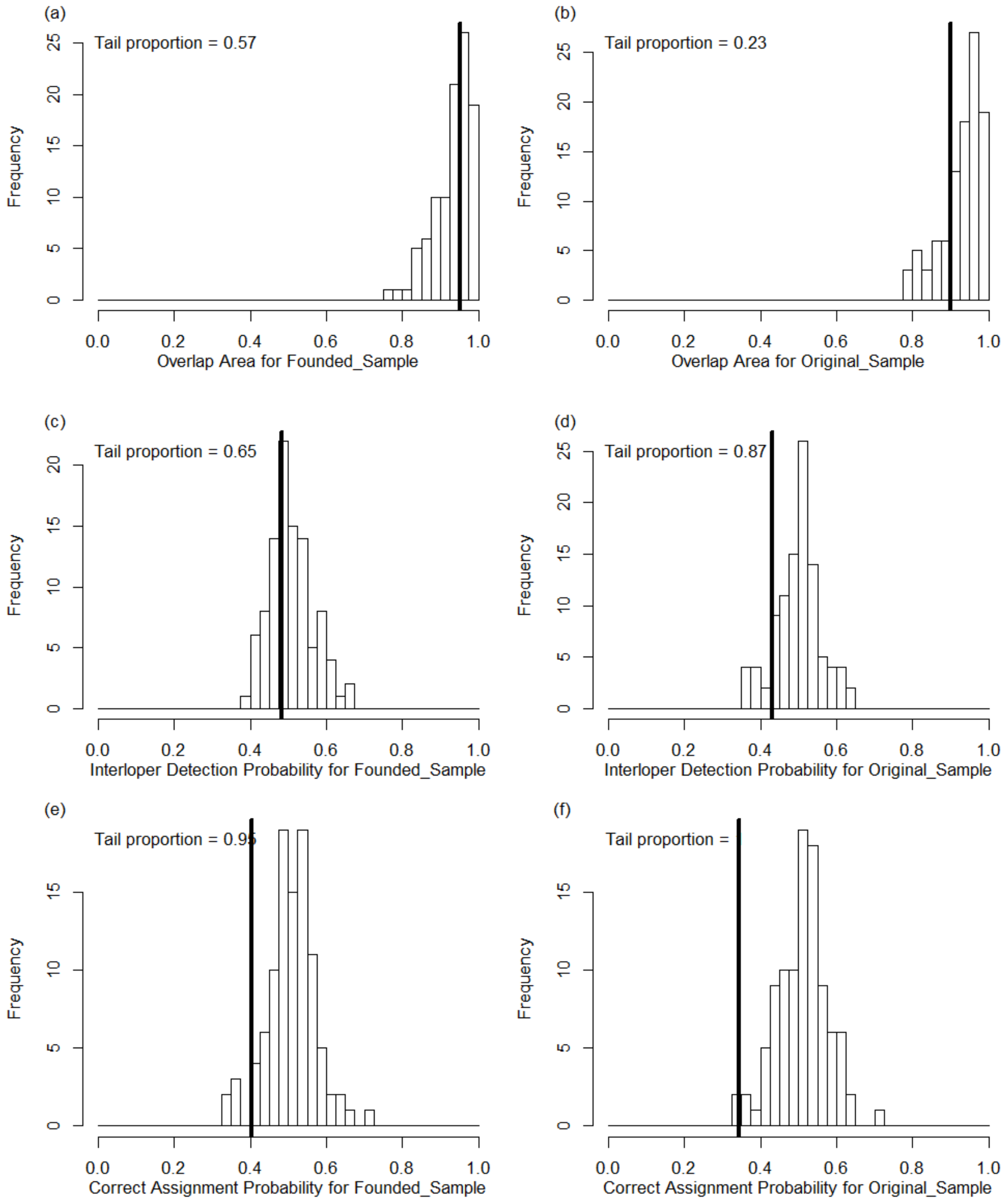


Figure 3.30: Permutation test results based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 1 generation, where the Founded Population was made up of 20 internally bred rats and 20 immigrant rats. Details as in Figure 3.5.

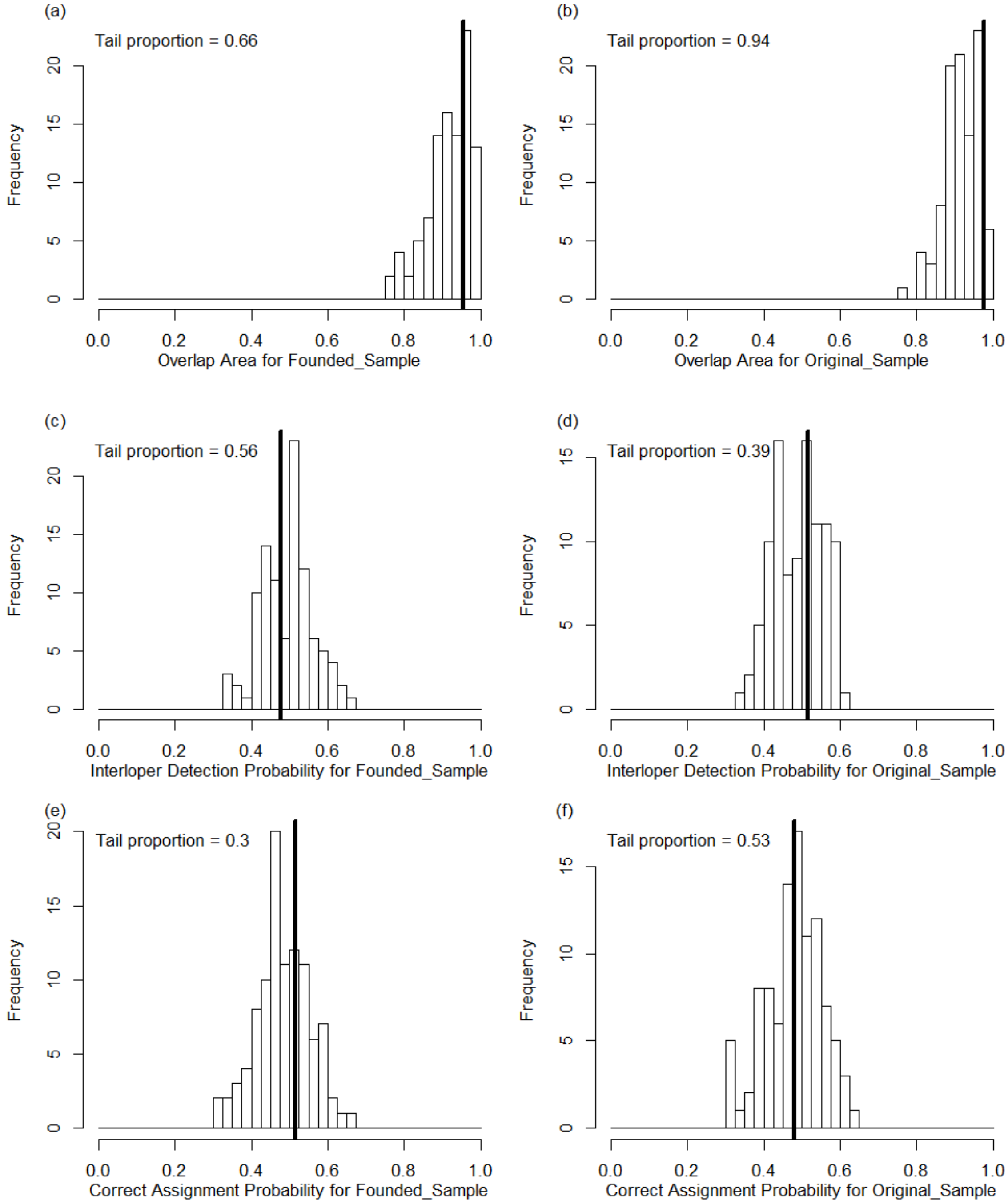


Figure 3.31: Permutation test results based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 3 generations, where the Founded Population was made up of 20 internally bred rats and 20 immigrant rats. Details as in Figure 3.5.

Figures 3.32 and 3.33 show combined GenePlots and LGP distribution plots for one replicate of a Founded Population with 35 internally bred individuals and 5 immigrants in each generation. Figure 3.32 shows samples from generation 1 and Figure 3.33 shows samples from generation 3 of the Founded Population.

After 1 generation, there is much less overlap between the single-population LGP distribution of the Founded Population and the cross-population LGP distribution of the Original sample into the Founded Population (the bottom plot in Figures 3.32 and 3.33) than was seen in Figures 3.28 and 3.29. The sample from the Original (Aotea) population does not fit as well into the Founded Population when there are fewer migrants arriving in the Founded Population.

By contrast, there is still a large amount of overlap between the single-population LGP distribution of the Original sample and the cross-population LGP distribution of the Founded Population into the Original sample (the bottom plot in Figures 3.32 and 3.33). This suggests that the Founded Population is undergoing a small amount of genetic drift but the Founded Population simulated rats would still fit well into the Original (Aotea) population.

Figures 3.34 and 3.35 show the permutation test results from the same samples as in Figures 3.32 and 3.33.

For generation 1, the results for the Founded Population are weakly significant (Figure 3.34). One of the other two replicate runs also had weakly significant results, and the other had significant results for generation 1 of the Founded Population. For generation 3, the results for the Founded Population are significant (Figure 3.35) and the other two replicate runs had weakly significant results.

For generation 1 (Figure 3.34), the results for the Original (Aotea) population are non-significant. For the other two replicate runs, the overlap area and interloper detection probability were non-significant and the correct assignment probability was weakly significant. For generation 3 (Figure 3.35), the results for the Original (Aotea) population are similar to those for generation 1: non-significant for overlap area and interloper detection probability, and weakly significant or non-significant for correct assignment probability.

This implies that if 5 out of the population of 40 (12.5%) in each generation are migrants then within two or three generations the overlap area and interloper detection probability measures will be able to distinguish between the Founded Population and the original Aotea population.

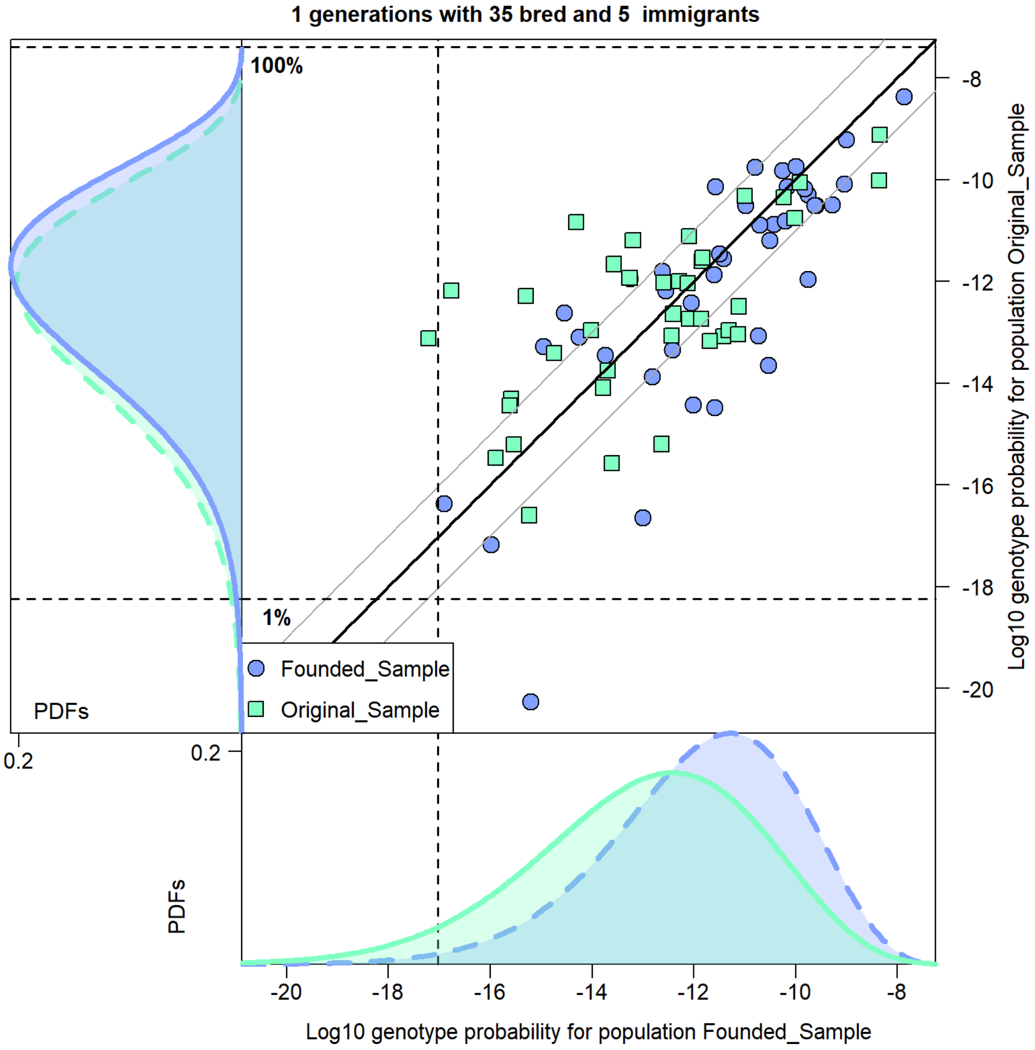


Figure 3.32: GenePlot and LGP distribution plots based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 1 generation, where the Founded Population was made up of 35 internally bred rats and 5 immigrant rats. Details as in Figure 3.1.

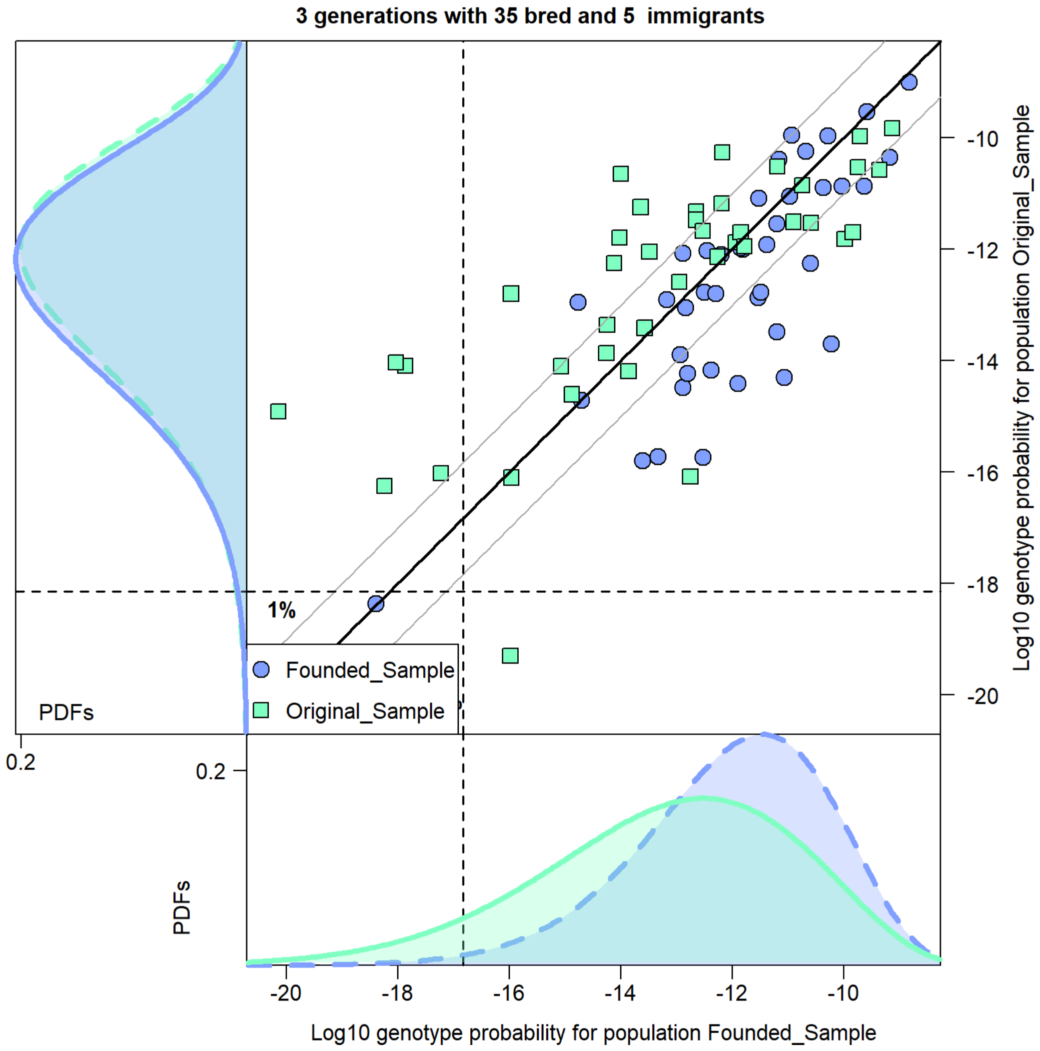


Figure 3.33: GenePlot and LGP distribution plots based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 3 generations, where the Founded Population was made up of 35 internally bred rats and 5 immigrant rats. Details as in Figure 3.1.

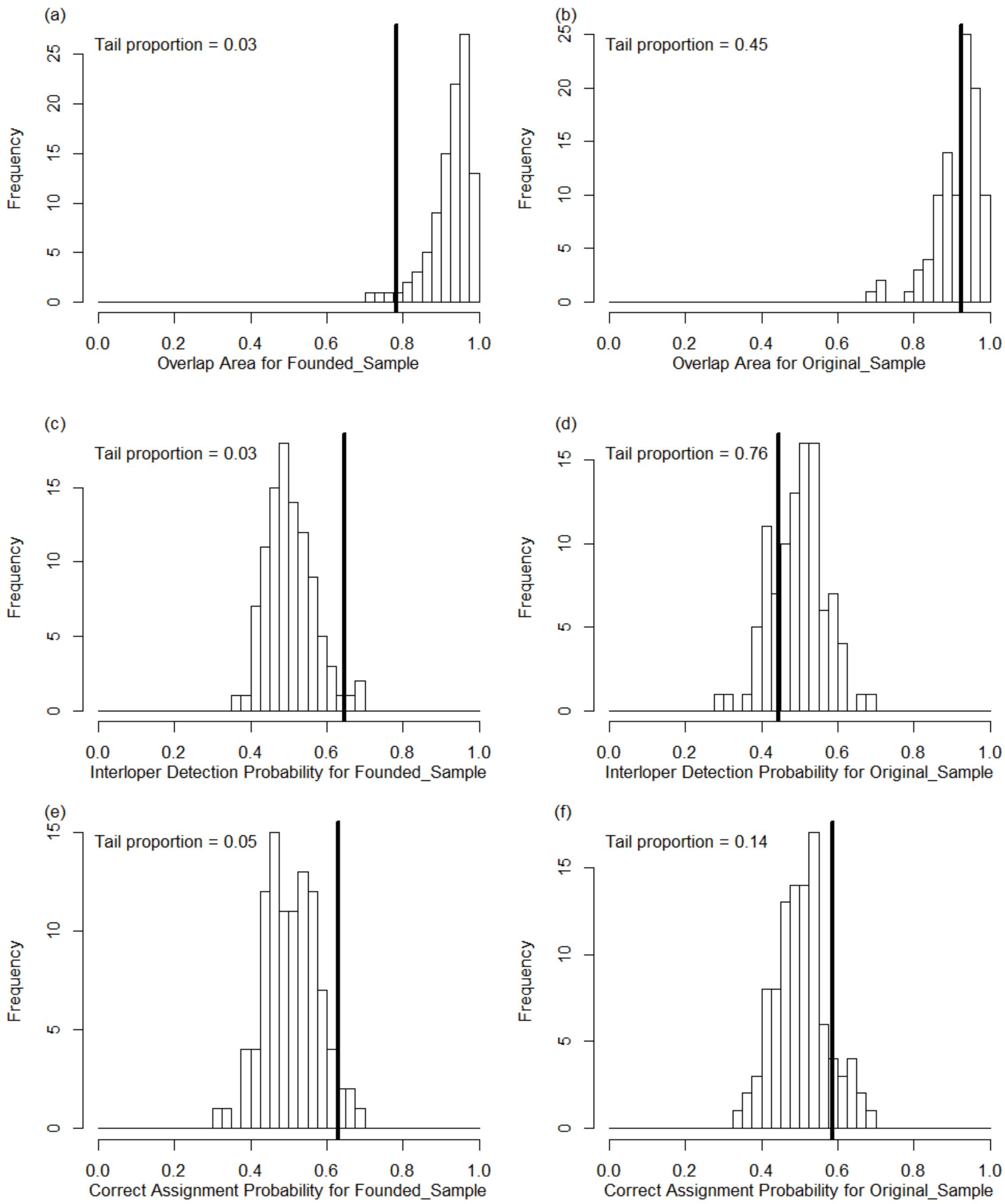


Figure 3.34: Permutation test results based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 1 generation, where the Founded Population was made up of 35 internally bred rats and 5 immigrant rats. Details as in Figure 3.5.

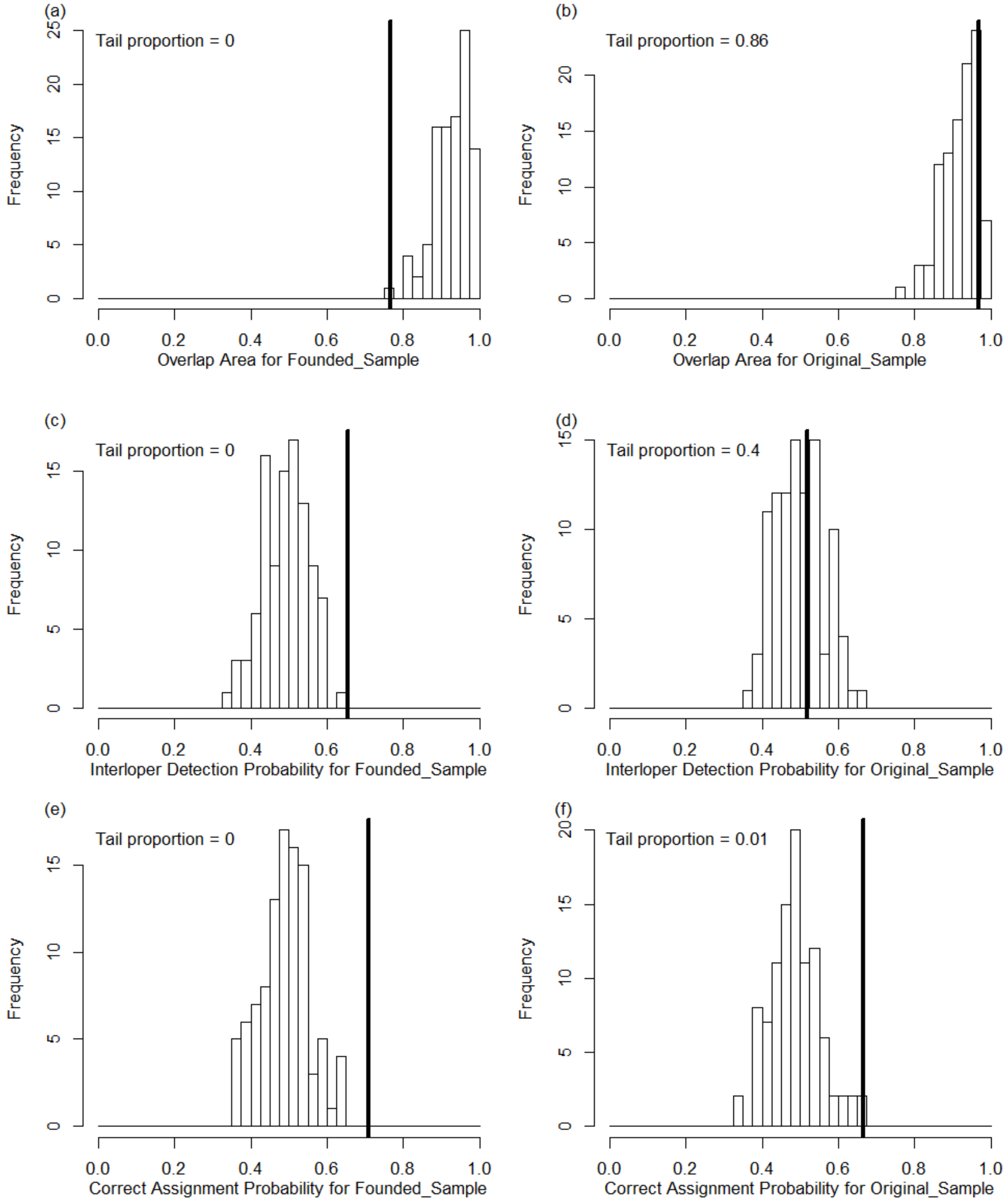


Figure 3.35: Permutation test results based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 3 generations, where the Founded Population was made up of 35 internally bred rats and 5 immigrant rats. Details as in Figure 3.5.

Figures 3.36, 3.37, 3.38 and 3.39 show combined GenePlots and LGP distribution plots for one replicate of a Founded Population with 39 internally bred individuals and 1 immigrant in each generation. Figures 3.36 and 3.37 show samples from generation 1 for replicates 1 and 3. Figure 3.38 shows samples from generation 2 and Figure 3.39 shows samples from generation 3 of the Founded Population.

Figure 3.36 shows more overlap in the bottom LGP distribution plot than was seen in Figures 3.32 and 3.33, but Figure 3.37 shows more overlap than Figure 3.36. Figures 3.38 and 3.39 show similar amounts of overlap in the bottom LGP distribution plot to Figures 3.32 and 3.33.

As for Figures 3.32 and 3.33, the left LGP distribution plots in Figures 3.36 to 3.39, which show the fit of typical Founded Population rats into the Original (Aotea) sample, still show large amounts of overlap. This demonstrates the directionality of the scenario, where the Founded population retains a good fit to the Original (Aotea) population whilst undergoing genetic drift that makes the Original population an increasingly poor fit to the Founded Population.

Figures 3.40 to 3.43 show the permutation test results from the same samples as in Figures 3.36 to 3.39.

For generation 1, the results for the Founded Population in replicate 1 are non-significant (Figure 3.40) and the results for the Founded Population in replicate 3 are non-significant for overlap area and weakly significant for interloper detection probability (Figure 3.41). The remaining replicate had non-significant results for overlap area and interloper detection probability and weakly significant results for correct assignment probability.

For generation 2, the results for the Founded Population are significant for overlap area and interloper detection probability (Figure 3.42) and this was the case for all three replicates. The results for the Founded Population for correct assignment probability were significant for one replicate, weakly significant for another replicate and non-significant for the third replicate.

For generation 3, the results for the Founded Population were all weakly significant or significant for all three replicates (Figure 3.43).

For the Original (Aotea) population, the overlap area and interloper detection probability results were non-significant for all generations and all three replicates (Figures 3.40 to 3.43). The correct assignment probability results for the Original (Aotea) population were non-significant for generation 1 for all three replicates, and significant or weakly significant for generations 2 and 3 for all three replicates.

This implies that if only one of each generation of 40 (2.5%) is an immigrant then within two generations the overlap area and interloper detection probability measures will be able to distinguish between the Founded Population and the original Aotea population.

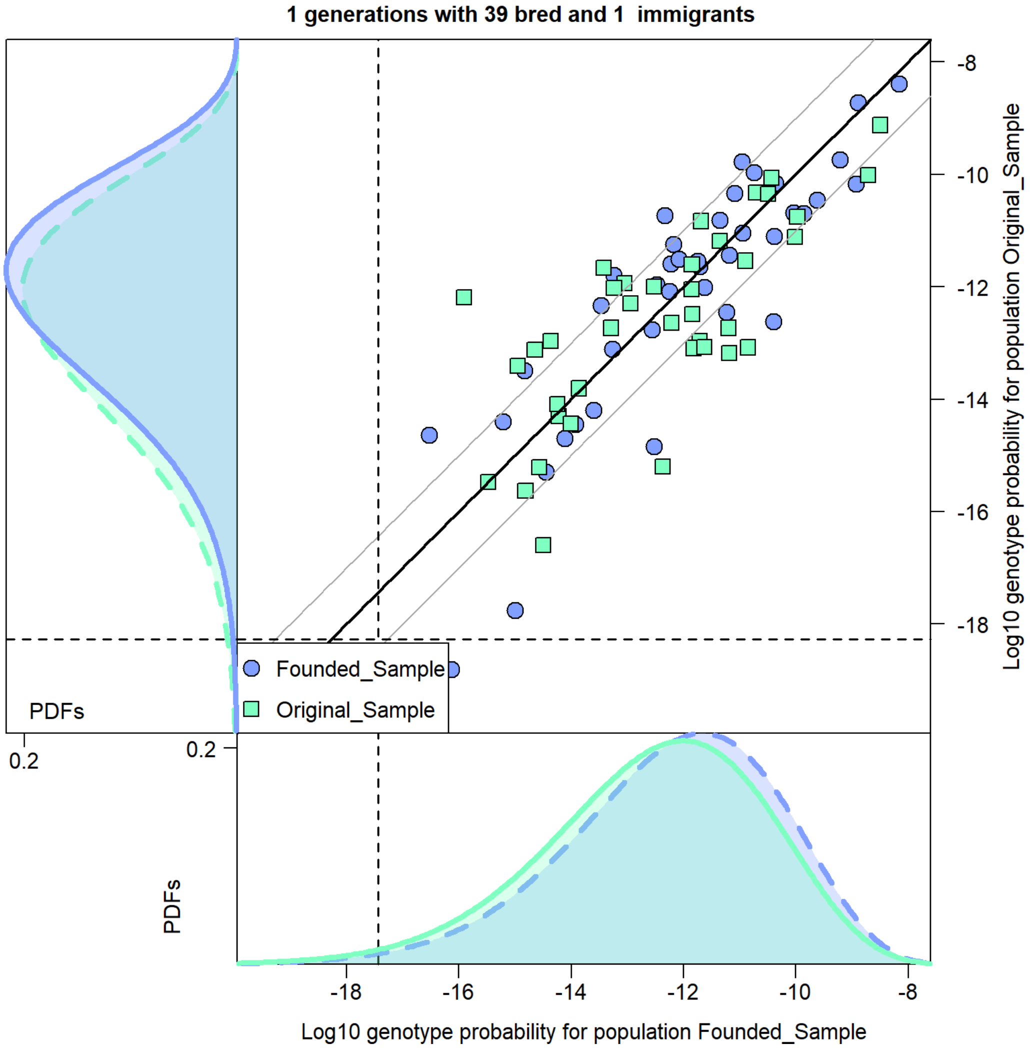


Figure 3.36: GenePlot and LGP distribution plots based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 1 generation, where the Founded Population was made up of 39 internally bred rats and 1 immigrant rat. Results from replicate 1. Details as in Figure 3.1.

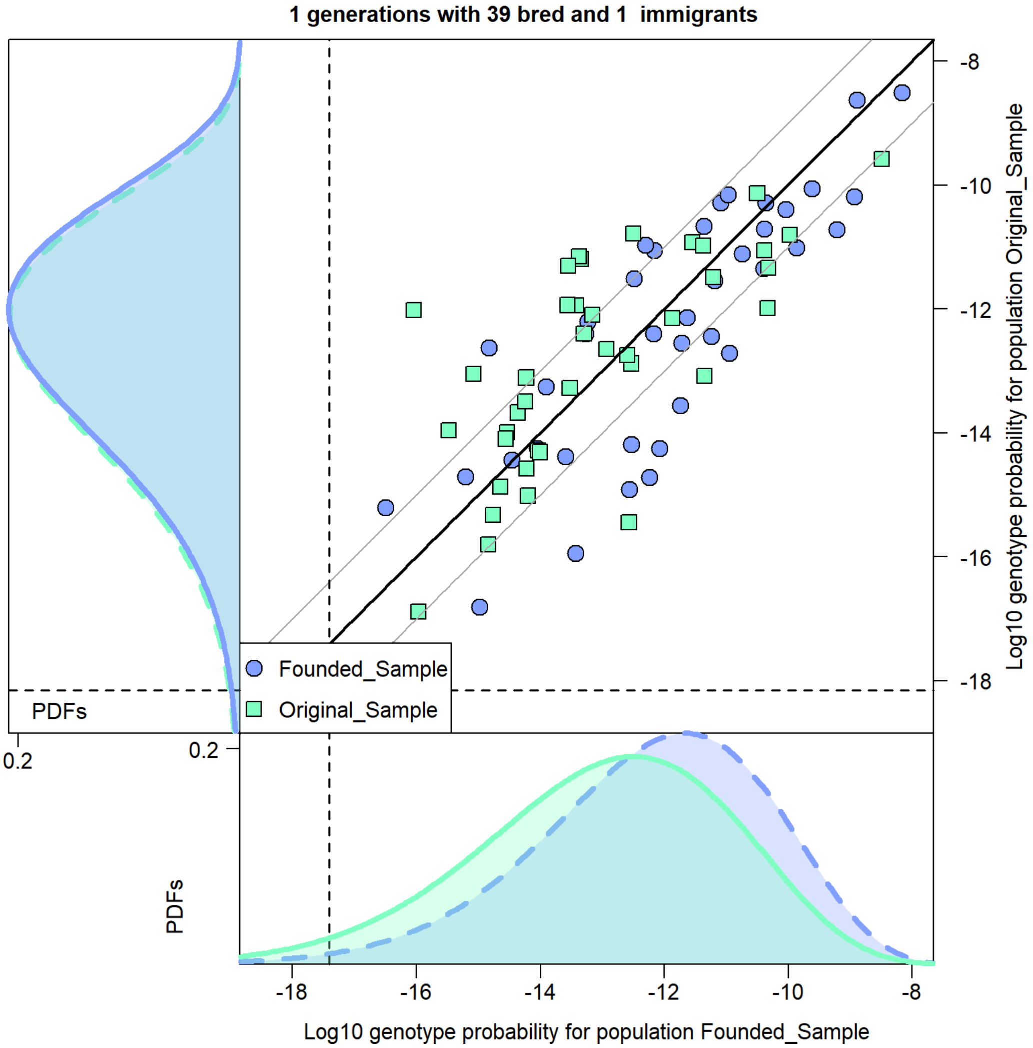


Figure 3.37: GenePlot and LGP distribution plots based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 1 generation, where the Founded Population was made up of 39 internally bred rats and 1 immigrant rat. Results from replicate 3. Details as in Figure 3.1.

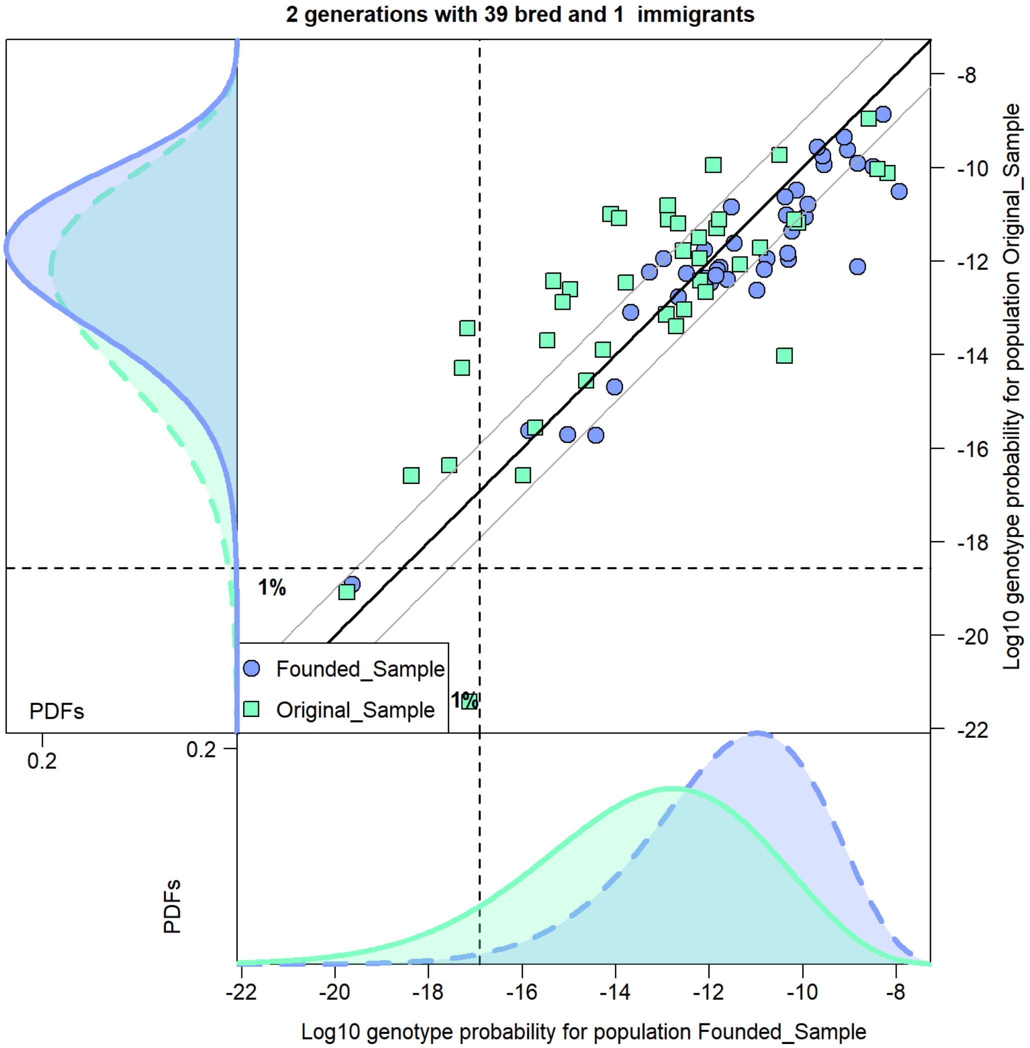


Figure 3.38: GenePlot and LGP distribution plots based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 2 generations, where the Founded Population was made up of 39 internally bred rats and 1 immigrant rat. Details as in Figure 3.1.

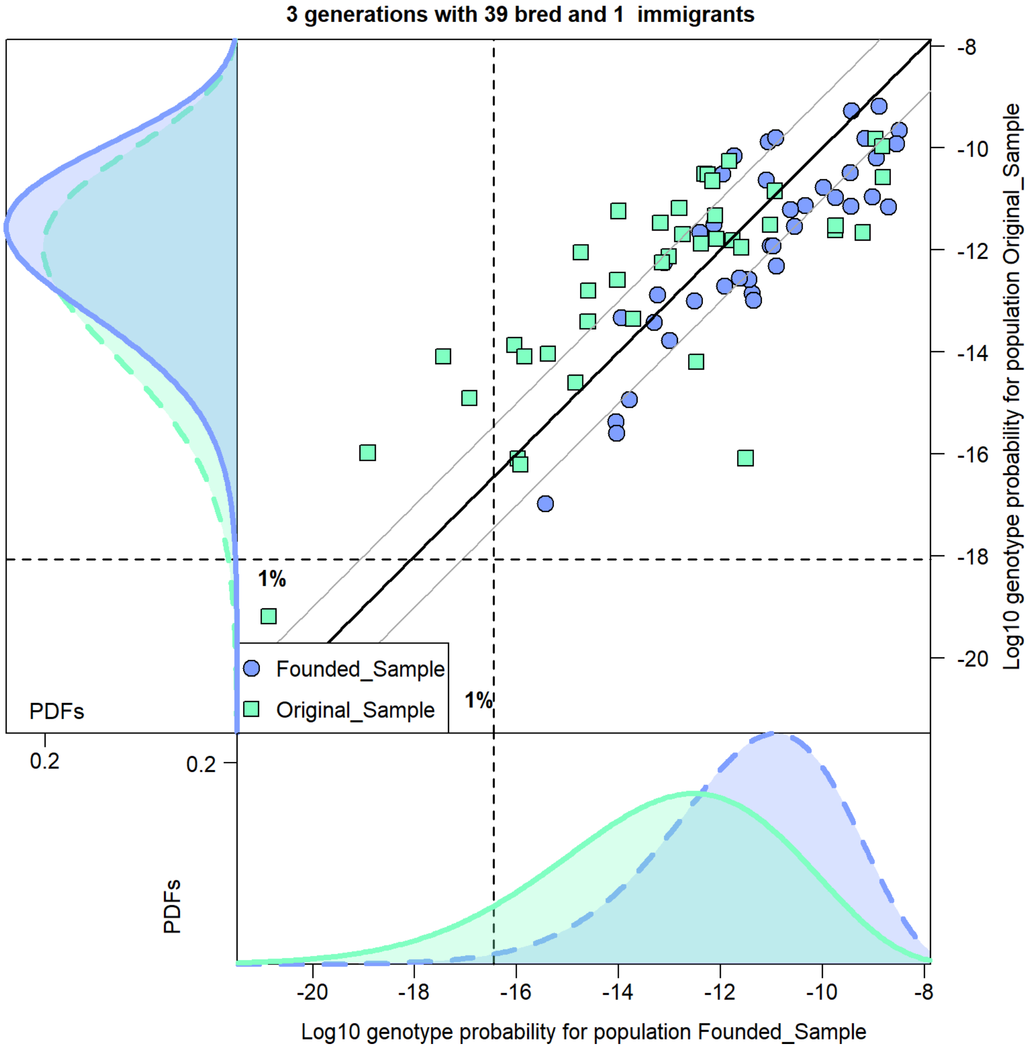


Figure 3.39: GenePlot and LGP distribution plots based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 3 generations, where the Founded Population was made up of 39 internally bred rats and 1 immigrant rat. Details as in Figure 3.1.

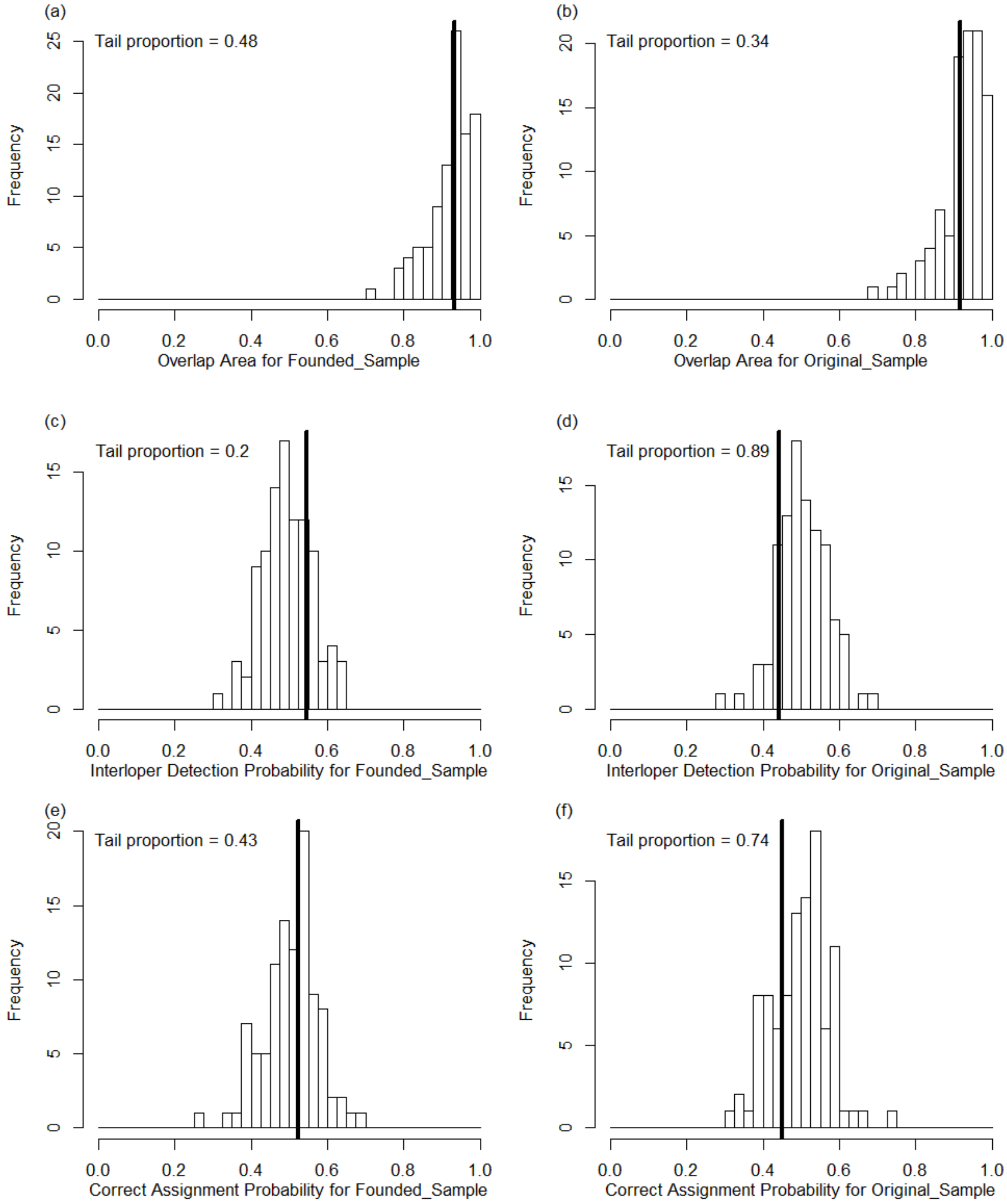


Figure 3.40: Permutation test results based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 1 generation, where the Founded Population was made up of 39 internally bred rats and 1 immigrant rat. Results from replicate 1. Details as in Figure 3.5.

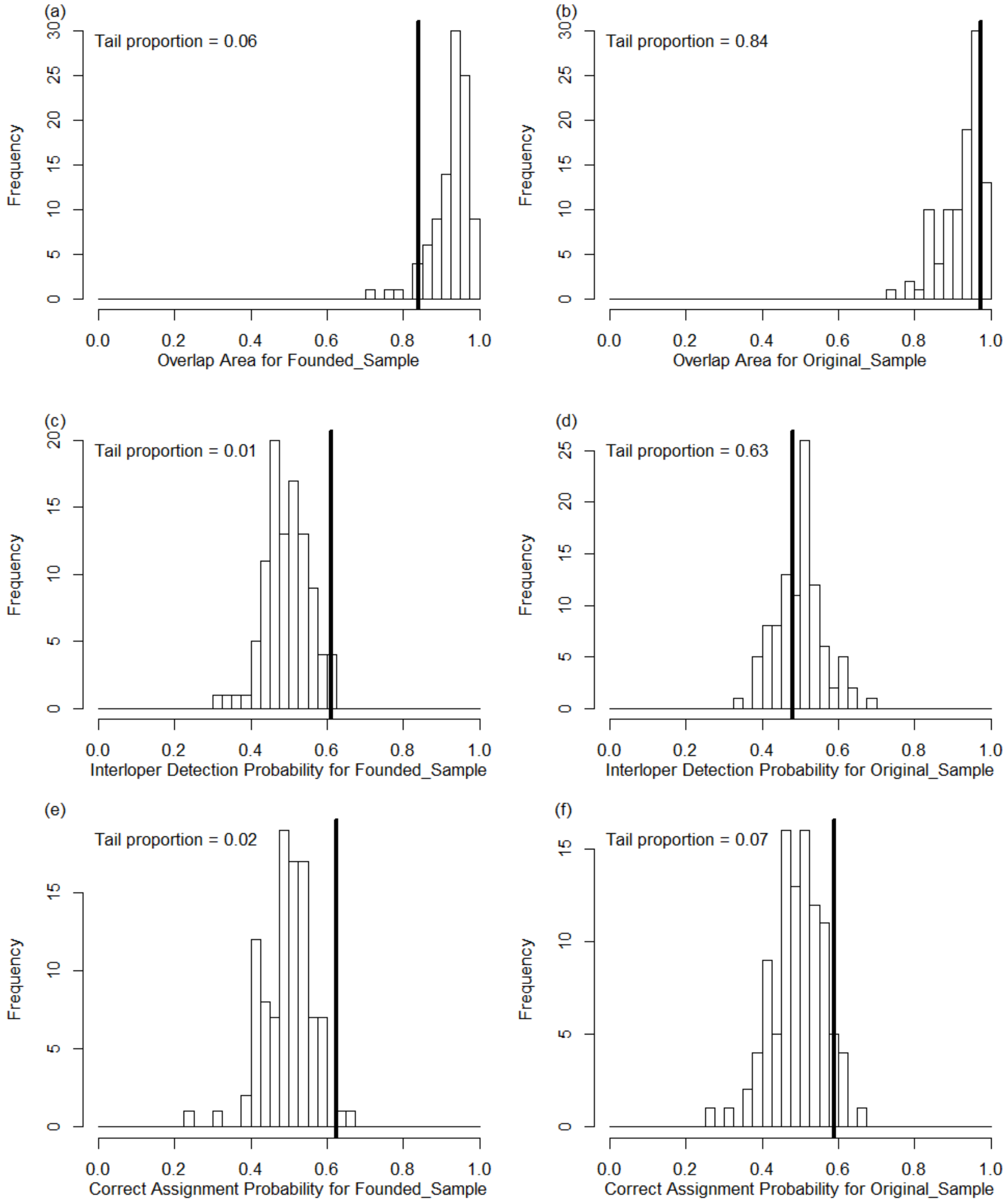


Figure 3.41: Permutation test results based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 1 generation, where the Founded Population was made up of 39 internally bred rats and 1 immigrant rat. Results from replicate 3. Details as in Figure 3.5.

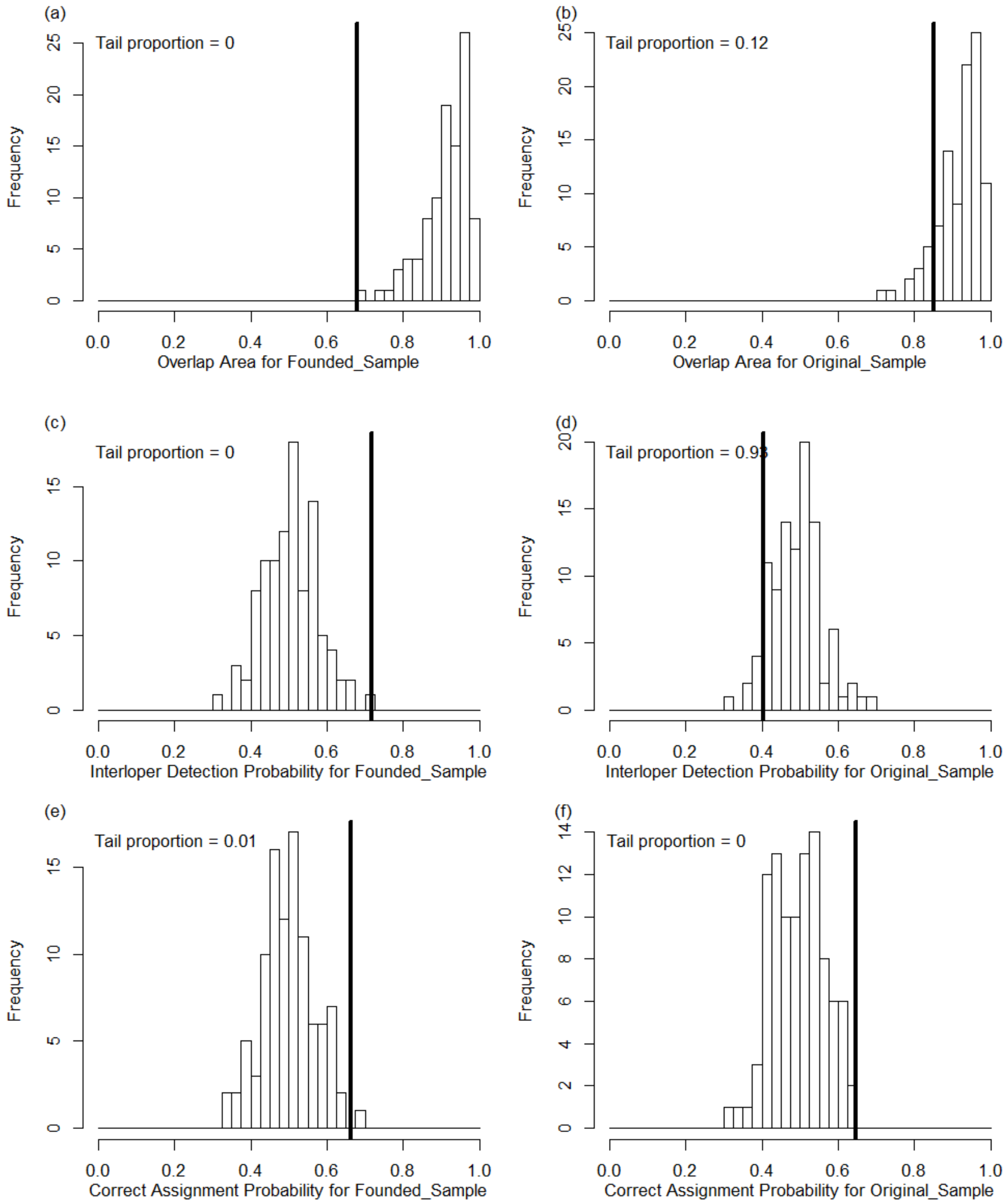


Figure 3.42: Permutation test results based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 2 generations, where the Founded Population was made up of 39 internally bred rats and 1 immigrant rat. Details as in Figure 3.5.

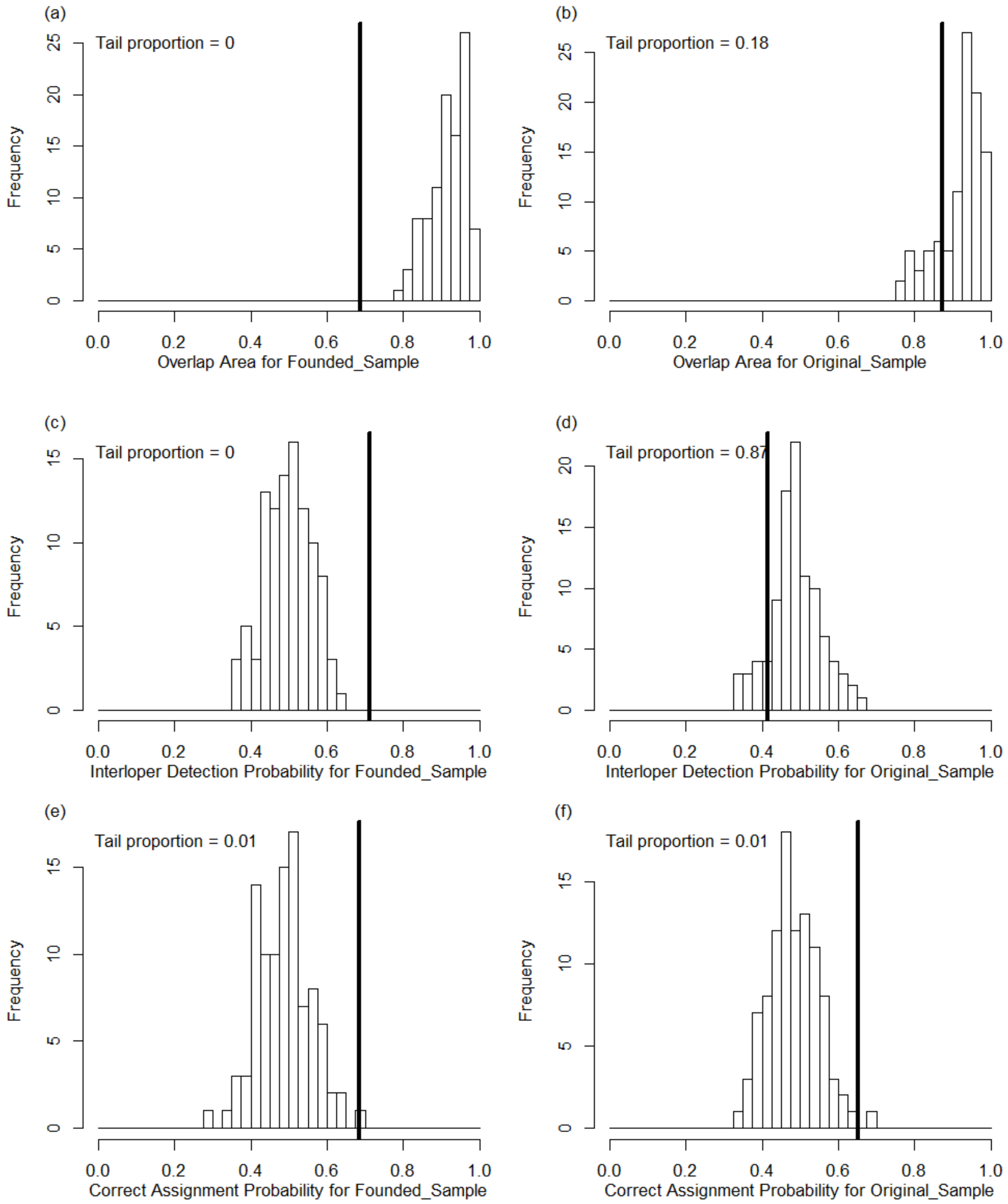


Figure 3.43: Permutation test results based on microsatellite data from 40 ship rats (*Rattus rattus*) sampled on Aotea and 40 rats sampled from the Founded Population after 3 generations, where the Founded Population was made up of 39 internally bred rats and 1 immigrant rat. Details as in Figure 3.5.

3.9 Discussion

The key benefit of our new methodology is that the new measures are closely linked to the GenePlot visualization, greatly improving the interpretability of the results. The GenePlot shows the assignment scores for the observed samples, allowing one to identify any individuals who do not fit well with the rest of their sample. The LGP distribution plots provide an overview of the populations after accounting for sampling uncertainty, and the numeric measures summarize these findings, which can be tested for significance.

The new methodology is also distinct from existing measures such as F_{ST}, G'_{ST} and D_{est}, in that it is focused on practical applications of assignment.

F_{ST}, still the most widely used of the existing measures (see Section 1.3), can be interpreted in multiple ways. It has been used to detect loci under selection, and it is commonly interpreted as a fixation index. It is also used as an indicator of gene flow, and can be used to estimate migration rate. Standardization, in the form of alternative estimators such as G'_{ST} and Φ_{ST}, can be used to account for the effects of mutation when analysing F_{ST}. However, this plethora of uses makes it more difficult to interpret F_{ST} in any particular scenario, because it is affected by so many different factors.

D and D_{est} are separated from questions of fixation in order to make clearer statements about the commonality of alleles between populations. But the major disadvantages of D are the lack of a natural global formulation (see Whitlock 2011, and Section 1.3), and its lack of association with population size (Meirmans & Hedrick 2011).

Our measures, overlap area, interloper detection probability and correct assignment probability, are affected by population size and therefore do reflect local drift, unlike D. And although our measures do not, as Whitlock (2011) requested, give "similar results for all neutral loci", rather, we incorporate information from any neutral loci to assess the populations. We could subselect loci that are particularly highly differentiating between populations, but there is no need to do so within our framework, and indeed the choice of loci might produce biased results when analysing additional samples in future.

Unlike existing measures that are mostly defined in terms of heterozygosity, our measures are defined in terms of genotypes. Rather than attempting to make direct inferences about migration rates, we aim to determine whether the evidence of population structure is clear enough to inform assignment or make decisions about the practical management of populations.

Our measures are, as yet, only applicable to pairs of populations rather than a group of multiple populations, unlike measures such as F_{ST}, G'_{ST} and D. For pairs of populations, however, those existing measures are all symmetric, giving only a single value for the two populations, whereas our measures are asymmetric and directional, aiming to capture the disparity that so often exists between populations. Our simulations in Section 3.8 show that our methodology provides a powerful tool for inferring the direction of differentiation between the populations.

A final comment on the three proposed measures is that the permutation results for the overlap area and interloper detection probability are very often aligned with each other, so although users should assess both measures, they can for the most part choose which to use. The overlap area has a more intuitive visual interpretation, whereas the interloper detection probability is designed to be conceptually easier to explain.

The correct assignment probability tends to have higher values than the interloper detection probability for any given pair of populations, although not always. This tendency reflects the fact that assigning of one individual to their own population rather than another population is likely to be easier than distinguishing between a native and a migrant within a given population. Returning to our analogy from Section 3.3, any Dutch child that is good at English is likely to be even better at Dutch!

We conclude that our measures provide a useful complement and a distinct addition to the set of existing diversity measures. F_{ST} will probably still be used for most studies, but there are scenarios where it, and other symmetric measures, are not sufficient to capture the complexity of the population structure. Our methodology offers greater interpretability and a more nuanced understanding of population structure than any other existing measures.

Appendices for Chapter 3

Appendix H Details of overlap area and probability measure calculations

This section describes the calculation of overlap area, interloper detection probability and correct assignment probability. The initial step is to calculate the posterior allele frequencies for the candidate source populations R and S based on the observed genetic samples from those populations, using the method of Rannala & Mountain (1997). From these we can obtain the LGP distributions for each population and the cross-population LGP distributions. As described in Section 2.2.2, we define the LGP distribution for population R as being the posterior distribution of log-genotype probabilities (LGPs) for all possible multilocus genotypes arising from the posterior allele frequencies for population R. In this context, the LGP is interpreted as a measure of genetic fit.

As described in Appendix C, for a single population R and a single locus L, any genotype $\boldsymbol{a}^L$ arises with probability $\mathbb{P}_R(\boldsymbol{a}^L)$ given by equation (2.4). Thus the single-locus LGP distribution acquires point mass $\mathbb{P}_R(\boldsymbol{a}^L)$ at value $\log\{\mathbb{P}_R(\boldsymbol{a}^L)\}$ for every genotype $\boldsymbol{a}^L$. Recall that $\log\{\mathbb{P}_R(\boldsymbol{a}^L)\}$ serves as a measure of genetic fit and we are aiming to characterize the probability distribution of this quantity.

Let $Y_L = \log\{\mathbb{P}_R(\boldsymbol{a}^L)\}$ be the random variable representing the single-locus LGP for population R, and let genotypes at locus L be indexed by $g = 1, 2, \ldots, k(k+1)/2$ where k is the number of allele types at locus L. Write $\alpha_g = \mathbb{P}_R(\boldsymbol{a}_g^L)$ for each g. Then each genotype $\boldsymbol{a}_g^L$ contributes point mass α_g to value $Y_L = \log(\alpha_g)$.

The random variable $Y = \sum_L Y_L$ has as its distribution the multilocus LGP distribution for population R. We can obtain the probability density function (PDF) $f(y)$ and cumulative distribution function (CDF) $F(y)$ via the saddlepoint approximation method.

H.1 Saddlepoint CDF and PDF approximations

The saddlepoint approximation to a distribution is defined in terms of the cumulant generating function (CGF), which is the log of the moment generating function (MGF). Let M_L be the moment generating function for the single-locus distribution of Y_L and let K_L be the cumulant generating function for Y_L.

Since the CGF is the log of the MGF, and genotypes are assumed to be independent across loci, the multilocus CGF, K, is the sum of the single locus CGFs, and the derivatives are similarly defined (see Appendix C).

As given in Section 2.4.3, the saddlepoint approximation to the CDF of Y_L is then:

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

where Φ and ϕ are the Gaussian CDF and PDF, μ is the expected value of Y_L and $\hat{v} = \hat{s}\sqrt{K''(\hat{s})}$,

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

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

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

The saddlepoint approximation to the PDF is the original form of the saddlepoint approximation developed by Daniels (1954), and is given by:

$$\hat{f}(y) = \frac{1}{\sqrt{2\pi K''(\hat{s})}} \exp\{K(\hat{s}) - \hat{s}y\}.$$

H.2 Single and cross-population LGP distributions

If population R is the Broken Islands population, the PDF and CDF of the multilocus LGP distribution for population R are $f(y)$ and $F(y)$ corresponding to the bottom plot in Figure 3.1. The PDF, $f(y)$, is the curve shown by the dashed line, and $F(y)$ is the cumulative integral of that curve.

Now we define the cross-population LGP distribution for population S within the baseline population R. The posterior allele frequencies for R and S are based on the set of all allele types that were sampled in at least one of R or S, so the set of possible genotypes is the same for either population. For a single locus L, the genotypes are written as $\boldsymbol{a}_g^L$.

At locus L, the cross-population LGP distribution within baseline population R has the same values $\log(\alpha_g) = \log\{\mathbb{P}_R(\boldsymbol{a}^L)\}$ as the single-population LGP distribution for R. However, in the cross-population distribution, the point mass at $\log(\alpha_g)$ is $\beta_g = \mathbb{P}_S(\boldsymbol{a}_g^L)$, the probability of finding genotype $\boldsymbol{a}_g^L$ in the comparison population S. Let this be the distribution of variable Z_L. Thus Z_L describes, at locus L, the distribution of genetic fit to population R of individuals arising from population S.

This definition leads naturally to the multilocus cross-population LGP distribution for S within baseline population R, described by random variable $Z = \sum_L Z_L$.

Z_L has the single-locus moment generating function

$$\begin{aligned} \tilde{M}_L(t) &= \sum_z P(Z_L = z)\exp(tz) \\ &= \sum_g \beta_g \exp(t \log \alpha_g) \\ &= \sum_g \beta_g \alpha_g^t\,. \end{aligned}$$

Derivatives of the MGF are given by:

$$\tilde{M}_L^{(r)}(t) = \sum_g \beta_g (\log \alpha_g)^r \alpha_g^t \, .$$

From the moment generating function we can calculate the cumulant generating function (CGF) as its logarithm. The CGF for Z is then the sum of the CGFs of Z_L for all loci L. We then obtain the PDF and CDF of the cross-population LGP distribution within baseline R, using the saddlepoint method. We often refer to this as the LGP distribution of S into R, because it describes the genetic fit of individuals arising from S within the baseline population R.

If we take populations R and S to be the Broken Islands and Aotea populations respectively, the saddlepoint approximation to the PDF of the cross-population LGP distribution of S (Aotea) within R (the Broken Islands) is the function $\hat{\xi}(z)$ given by the solid blue line in the bottom LGP plot in Figure 3.1 and shown in Figure H.1.

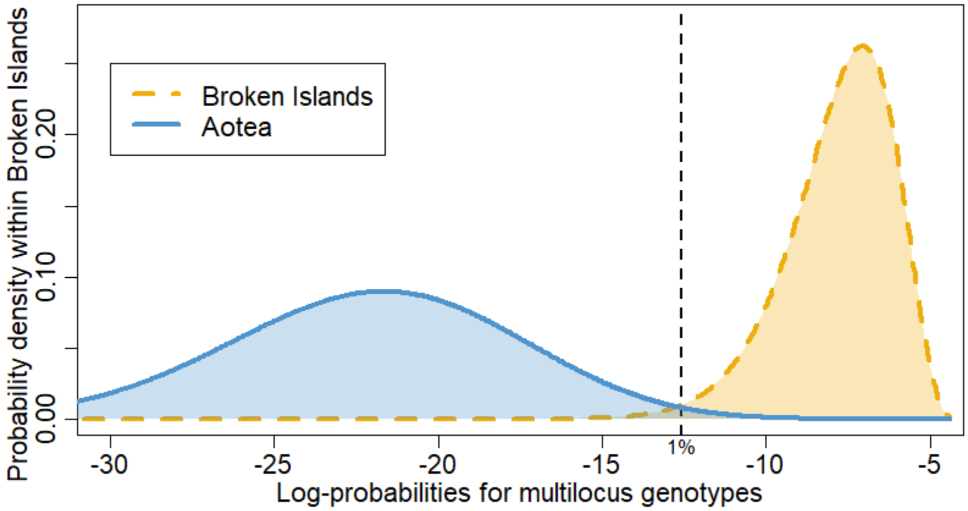


Figure H.1: LGP distribution plot based on microsatellite data extracted from ship rats (*Rattus rattus*) captured on the Broken Islands ($n = 60$) or on Aotea ($n = 57$). The dashed line shows the LGP distribution of the Broken Islands population and the solid line shows the cross-population LGP distribution of the Aotea population within the Broken Islands population. The LGPs and the dashed curves were calculated using the leave-one-out method.

H.3 Overlap area and interloper detection probability

Once we have the saddlepoint approximations $\hat{f}(y)$ and $\hat{F}(y)$ to the PDF and CDF of the LGP distribution for population R, and the saddlepoint approximation $\hat{\xi}(z)$ to the PDF of the cross-population LGP distribution of S within R, then we can calculate the overlap area and interloper detection probabilities, defined as in Section 3.3. We start by normalizing $\hat{f}(y)$ and $\hat{\xi}(z)$ to ensure that the distributions of Y and Z are proper probability distributions.

The overlap area for population S within baseline population R is

$$\mathcal{O}_{S\to R} = \int \min(\hat{\xi}(y), \hat{f}(y)) \, dy,$$

and the interloper detection probability is

$$\mathcal{I}_{S\to R} = 1 - \int \hat{\xi}(y)\hat{F}(y)\, dy.$$

We can integrate over y because the distributions for both Y and Z share the same support values $y = \sum_L \log\{\mathbb{P}_R(\boldsymbol{a}^L)\}$.

H.4 Correct assignment probability

For correct assignment probability in baseline population R, as described in Section 3.3, we consider individuals arising from population R, and calculate the probability that their LGP in R exceeds that in S: in other words, the probability that an individual arising from R would be correctly assigned to R rather than S, using a best-fit assignment criterion.

We calculate the correct assignment probability by constructing the distribution of LGP differences in populations R and S, for genotypes arising in R.

For a single locus L, the LGP difference for genotype $\boldsymbol{a}_g^L$ is the difference between the LGP with respect to R and the LGP with respect to S:

$$\log\alpha_g - \log\beta_g = \log\{\mathbb{P}_R(\boldsymbol{a}^L)\} - \log\{\mathbb{P}_S(\boldsymbol{a}^L)\}.$$

Let W_L be the random variable describing the LGP difference distribution at locus L, for genotypes arising from R. This distribution has point mass α_g at value $w_L = \log\alpha_g - \log\beta_g$.

As in Section H.2, we can obtain the moment generating function for the distribution of W_L:

$$\begin{aligned} \acute{M}_L(t) &= \sum_w P(W_L = w)\exp(tw) \\ &= \sum_g \alpha_g \exp\{t(\log\alpha_g - \log\beta_g)\} \\ &= \sum_g \alpha_g \left(\frac{\alpha_g^t}{\beta_g^t}\right). \end{aligned}$$

Again, we can use this to obtain the single-locus cumulant generating function for W_L and the multilocus cumulant generating function for W, and thus the saddlepoint approximations to the CDF and PDF of W, which we call $\hat{\Upsilon}(w)$ and $\hat{\upsilon}(w)$. Figure 3.3 (a) shows $\hat{\upsilon}(w)$ when the populations R and S correspond to the Broken Islands and Aotea populations.

For assignment under the protocol where each individual is assigned to the population for which they have greater LGP, the correct assignment probability is calculated as the area to the right of 0 on the horizontal axis of Figure 3.3:

$$\mathcal{C}_{R:S} = 1 - \hat{\Upsilon}(0).$$

For more conservative assignment protocols, where individual I is only assigned to R if $\text{LGP}_I^R > \text{LGP}_I^S + \lambda$, the correct assignment probability for R versus S would be

$$\mathcal{C}_{R:S}(\lambda) = 1 - \hat{\Upsilon}(\lambda).$$

H.5 Leave-one-out form of single-population LGP distributions

In practice we generally favour the leave-one-out mode to produce GenePlots and to thenceforth calculate the overlap area, interloper detection probability and correct assignment probability measures.

When producing GenePlots in the leave-one-out mode, the LGP of any individual sampled from one of the reference populations is calculated with respect to their population of origin after temporarily removing that individual's genotype from the reference sample before calculating the posterior allele frequencies in that population. The purpose of this process is to remove the artefactual effect that each individual in the reference sample has on the resulting GenePlot. Without leave-one-out, an individual's fit into its own population is inflated, and the reference populations therefore appear more distinct than they should, particularly when the sample sizes are small or the number of loci is very high. With leave-one-out, each individual's fit is assessed with respect to the rest of the reference sample which is considered to be independent of the individual.

For our measures based on LGP distributions to accurately reflect the leave-one-out GenePlots, we also need to use a similar leave-one-out process when calculating the LGP distribution for each population. This is less straightforward than the leave-one-out process for GenePlots, because the LGP distributions are based on all the possible multilocus genotypes that could arise in the population, whereas GenePlots only plot results for sampled individuals. We need to conceptualize an adjusted measure of genetic fit, the leave-one-out LGP. For any multilocus genotype, observed or not, this corresponds to the LGP of that genotype with respect to the adjusted posterior allele frequencies in the reference population, where the adjusted posterior is obtained by first removing the alleles corresponding to the target genotype from the reference sample. Hypothetical multilocus genotypes are treated in the same way as the genotypes of sampled individuals, in that their alleles are removed from the posterior allele frequencies before calculating the genotype-specific LGP.

We calculate the leave-one-out LGP distribution for a single population R by calculating adjusted genotype probabilities with respect to R for all possible genotypes that could arise from population R.

For a single locus L with k distinct allele types, the genotype for individual I is $\boldsymbol{a} = (a_1, a_2, \ldots, a_k)$ where each a_i is 0, 1 or 2, indicating how many alleles of type i are included in its genotype, and $\sum_{i=1}^{k} a_i = 2$, as in Section 2.2.1.

The unadjusted posterior genotype probability for genotype $\boldsymbol{a}$ with respect to population R, from (2.4), is:

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

where n_R is the number of observed individuals in the reference sample from population R, x_i are the observed counts of alleles in the reference sample, and τ is defined by the choice of prior.

For leave-one-out LGP, we aim to remove the alleles corresponding to the genotype by adjusting the values x_i and n_R as needed. However, some alleles might not have been observed in the sample from this population. The posterior allele frequencies of these alleles are only based on the prior without the addition of any observed data. Thus it might happen that the reference sample does not contain sufficiently many of the target alleles to remove the genotype in question, so a fix is required for these cases.

For homozygous genotypes with $a_r = 2$ we therefore calculate the leave-one-out genotype probability with respect to population R differently depending on how many alleles of type a_r are present in the reference sample:

$$\mathbb{P}(\boldsymbol{a}^L)_{\text{LOO}} = \begin{cases} \dfrac{(x_r+\tau-2)(x_r+\tau-1)}{(2n_R+k\tau-2)(2n_R+k\tau-1)} & x_r \geq 2; \\ \dfrac{(x_r+\tau-1)(x_r+\tau)}{(2n_R+k\tau-1)(2n_R+k\tau)} & x_r = 1; \\ \dfrac{(x_r+\tau)(x_r+\tau+1)}{(2n_R+k\tau)(2n_R+k\tau+1)} & x_r = 0. \end{cases} \tag{H.1}$$

The first expression, in the top row of (H.1), is the standard one, in which x_r is replaced by $x_r - 2$ and n_R is replaced by $n_R - 1$. The second and third expressions progressively account for a deficit of allele r in the reference sample by replacing x_r wth $x_r - 1$ and n_R with $n_R - \frac{1}{2}$ if $x_r = 1$, and making no replacements if $x_r = 0$.

For heterozygous genotypes with $a_r = a_s = 1$, we similarly calculate the leave-one-out genotype probability depending on how many alleles of types a_r and a_s are present in the reference sample:

$$\mathbb{P}(\boldsymbol{a}^L)_{\text{LOO}} = \begin{cases} \dfrac{2(x_r+\tau-1)(x_s+\tau-1)}{(2n_R+k\tau-2)(2n_R+k\tau-1)} & x_r, x_s \geq 1; \\ \dfrac{2(x_r+\tau)(x_s+\tau-1)}{(2n_R+k\tau-1)(2n_R+k\tau)} & x_r = 0; x_s \geq 1; \\ \dfrac{2(x_r+\tau-1)(x_s+\tau)}{(2n_R+k\tau-1)(2n_R+k\tau)} & x_r \geq 1; x_s = 0; \\ \dfrac{2(x_r+\tau)(x_s+\tau)}{(2n_R+k\tau)(2n_R+k\tau+1)} & x_r = x_s = 0. \end{cases} \tag{H.2}$$

The leave-one-out LGP distribution for population R has the same probability point mass at genotype g as the standard LGP distribution for population R, namely $\alpha_g = \mathbb{P}_R(\boldsymbol{a}_g^L)$ for each g at locus L. The values for the leave-one-out LGP distribution are different, however, from the values for the standard LGP distribution, and are given by the logs of the adjusted genotype probabilities, $\log \breve{\alpha}_g = \log\{\mathbb{P}(\boldsymbol{a}^L)_{\text{LOO}}\}$. Thus, $\log \breve{\alpha}_g$ corresponds to a measure of genetic fit of genotype g in population R, when it is assessed relative to an independent reference sample.

The moment generating function for the leave-one-out LGP distribution is also slightly different from that given in Appendix C. The MGF in Appendix C is simplified by the fact that the values of the standard LGP distribution are equal to the logs of the point masses. Let $\breve{Z}_L$ be a random variable describing the leave-one-out LGP in population R at locus L, for a genotype arising from population R. For the leave-one-out distribution, the single-locus MGF is given by:

$$\begin{aligned}\breve{M}_L(t) &= \sum_z P(Z_L = z)\exp(tz) \\ &= \sum_g \alpha_g \exp(t \log \breve{\alpha}_g) \\ &= \sum_g \alpha_g \, \breve{\alpha}_g^t \, .\end{aligned}$$

Derivatives of the MGF are given by:

$$\breve{M}_L^{(r)}(t) = \sum_g \alpha_g (\log \breve{\alpha}_g)^r \, \breve{\alpha}_g^t \, .$$

Let $\breve{Z} = \sum_L \breve{Z}_L$ be the random variable distributed according to the multilocus leave-one-out LGP distribution for population R, and $\breve{f}(z)$ be the saddlepoint approximation to the PDF of that distribution.

Although McMillan & Fewster (2017) converted the leave-one-out results back to the full-population scale for plotting on the GenePlot, we do not use this procedure now and instead treat the leave-one-out GenePlot as a representation of the leave-one-out LGP distribution described here. Thus, individuals from reference population R are plotted on the leave-one-out GenePlot at their raw leave-one-out LGP value in population R, based on equations (H.1) and (H.2). These plotted leave-one-out LGP values for individuals sampled from R can therefore be viewed as a sample from the distribution of $\breve{Z}$, represented by the approximate PDF $\breve{f}(z)$.

We no longer convert the individual leave-one-out LGPs back into the full-population scale because that full-population scale is subject to self-referential bias, where the LGPs for individuals within their own population are biased upwards. This is described in Section H.7, and illustrated in Figures H.2(d), H.3(d) and H.4(c) and (d). As described in Section H.6, we therefore use the leave-one-out form of the LGP distribution for the baseline population, which corresponds to the dashed curves in the LGP plots alongside the central GenePlot, because the leave-one-out form of the baseline population LGP distribution is comparable to the cross-population LGP distribution. The central GenePlot needs to remain consistent with the LGP distribution plots, and therefore we also use the raw individual leave-one-out LGPs rather than converting back to the biased full-population scale.

H.6 Leave-one-out form of numeric measures

The leave-one-out forms of the overlap area, interloper detection probability and correct assignment probability are calculated by using the leave-one-out form for all calculations involving only the baseline population R, and using the standard non-leave-one-out form for all calculations involving a comparison population S as well as baseline population R. This means that all genotypes, whether from population R or S, are assessed with respect to an independent reference sample. If there is genuinely no genetic difference between populations R and S, then for reference samples of approximately equal size the leave-one-out LGP distribution of R will be approximately equivalent to the standard cross-population LGP distribution of S within R.

The leave-one-out form of the overlap area for population S within baseline R is

$$\breve{\mathcal{O}}_{S\to R} = \int \min(\hat{\xi}(y), \breve{f}(y))\, dy,$$

and the leave-one-out form of the interloper detection probability is

$$\breve{\mathcal{I}}_{S\to R} = 1 - \int \hat{\xi}(y)\breve{F}(y)\, dy,$$

where $\breve{f}(y)$ and $\breve{F}(y)$ are saddlepoint approximations to the PDF and CDF of the leave-one-out LGP distribution for population R, and $\hat{\xi}(y)$ is the saddlepoint approximation to $\xi(y)$, unchanged from previous sections.

In order to calculate the leave-one-out form of the correct assignment probability we first need to construct the leave-one-out form of the distribution for W_L, which has point mass α_g at value $\breve{w}_L = \log\breve{\alpha}_g - \log\beta_g$. In other words, the point masses of the distribution are unchanged, and the components of the values that relate to the comparison population S are unchanged, but the components of the values that relate to the baseline population R are the adjusted leave-one-out forms. Then the correct assignment probability with threshold λ is

$$\breve{\mathcal{C}}_{R:S}(\lambda) = 1 - \breve{\Upsilon}(\lambda),$$

where $\breve{\Upsilon}(w)$ is the saddlepoint approximation to the CDF of the leave-one-out distribution of $\breve{W}$.

H.7 Simulation results for leave-one-out LGP distribution

We conducted simulations to test that the leave-one-out LGP distribution for a single population R correctly reproduces the LGP distribution that is obtained when genotypes are assessed relative to an independent reference sample. We considered populations R and S with identical allele frequencies at 2000 biallelic loci. Allele frequencies at each locus were drawn from a $\text{Beta}(1, 10)$ distribution, conditional on the minor allele frequency being at least 0.05. We simulated reference samples of 50 individuals from each population, and drew posterior samples of size 10000 from the single-population and cross-population LGP distributions, variously under leave-one-out and non-leave-one-out protocols.

Figure H.2 shows boxplots and empirical density plots for the cross-population LGP distribution of S within R and the leave-one-out and standard LGP distributions for R. The

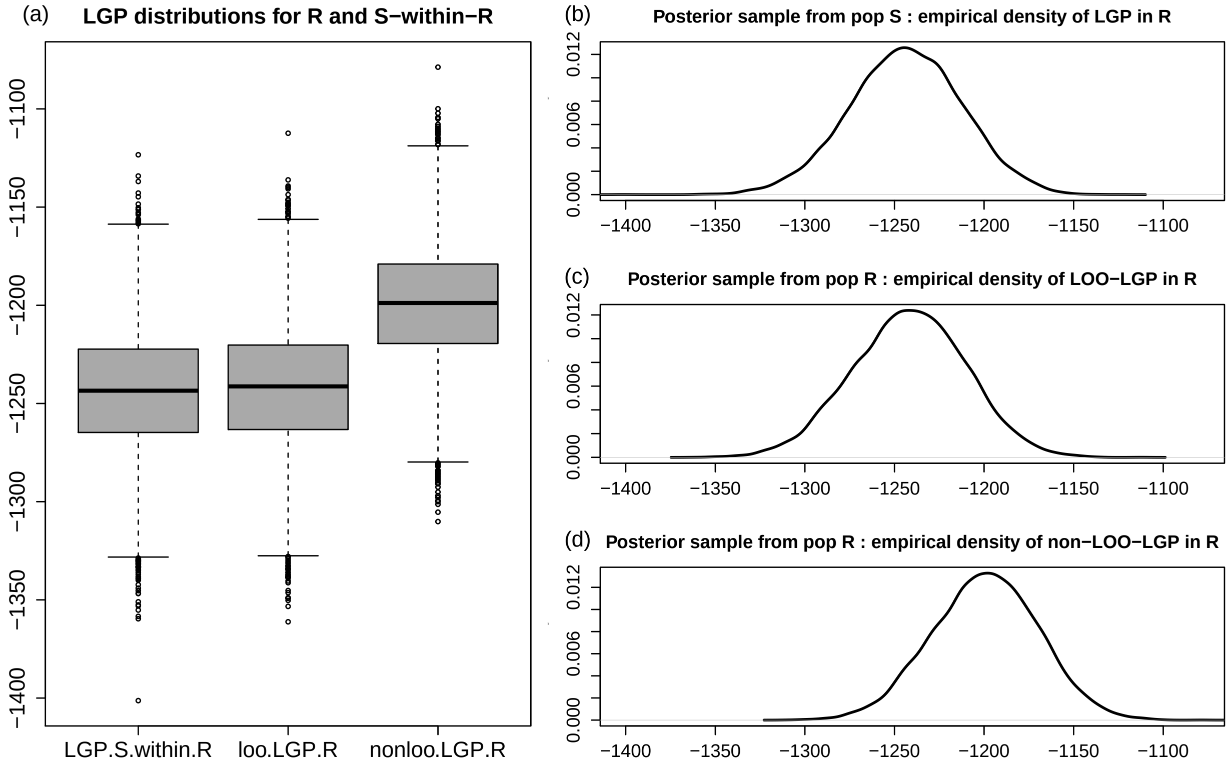


Figure H.2: Cross-population LGP distribution of S within R with the leave-one-out and standard LGP distributions for R. (a) Boxplot of 10000 posterior samples from each of the three distributions. (b) Empirical density plot for samples from the cross-population LGP distribution of S within R. (c) Empirical density plot for samples from the leave-one-out LGP distribution for R. (d) Empirical density plot for samples from the standard LGP distribution for R.

overlap area measure is defined as the overlap area between the cross-population and single-population distributions, and since the allele frequencies are identical in the two populations, we expect the overlap area to be close to one. We see from Figure H.2 that the leave-one-out form of the single-population distribution is the appropriate formulation for constructing the overlap area measure. The leave-one-out form closely matches the cross-population distribution, whereas the the non-leave-one-out form is biased upwards, overestimating the LGPs of individuals within their own population.

Figure H.3 shows boxplots and empirical density plots for the cross-population LGP distribution of R within S and the leave-one-out and standard LGP distributions for R. The correct assignment probability measure is constructed by taking individuals from R and comparing their fit into R with their fit into S, and this can be thought of as a comparison between the LGP distribution of R with the cross-population LGP distribution of R into S, although in practice we also account for correlation between the LGPs of a single individual with respect to two populations R and S by pairing the two results and evaluating the difference between the LGPs for that individual. Since the allele frequencies are identical in the two populations, we expect the cross-population and single-population distributions to be almost identical

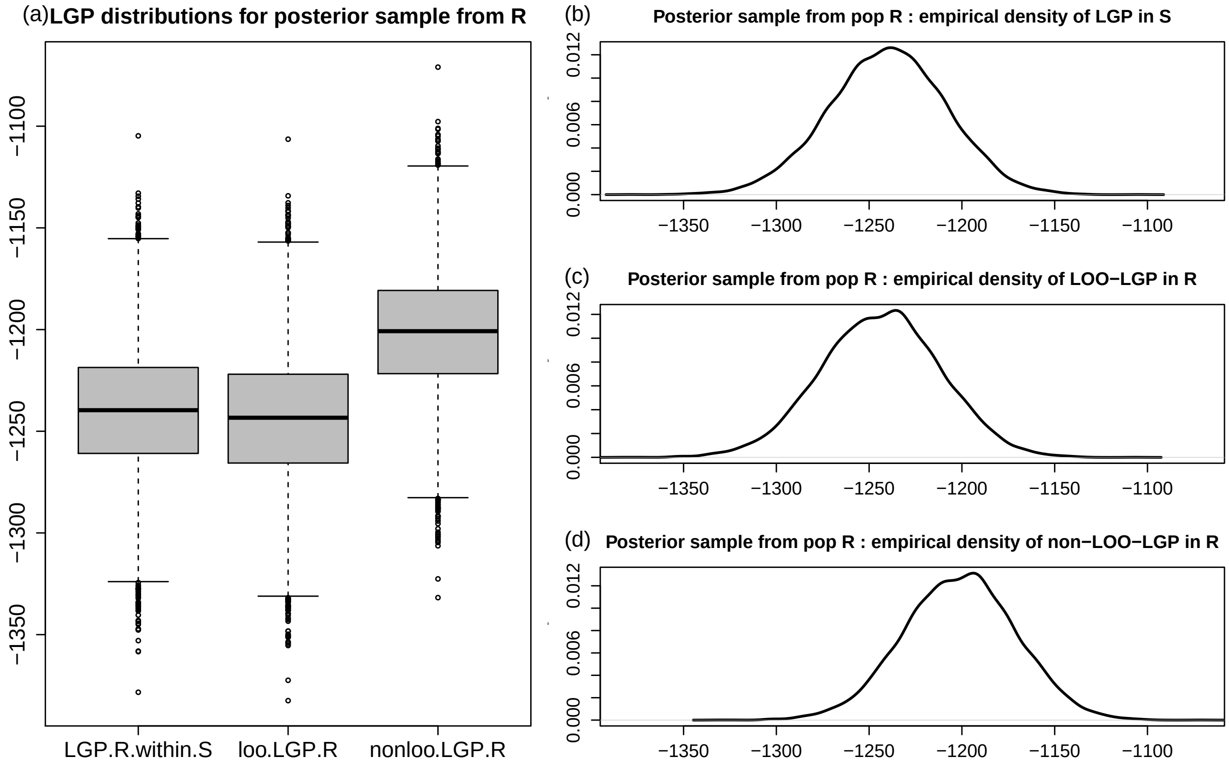


Figure H.3: Cross-population LGP distribution of R within S with the leave-one-out and standard LGP distributions for R. (a) Boxplot of 10000 posterior samples from each of the three distributions. (b) Empirical density plot for samples from the cross-population LGP distribution of R within S. (c) Empirical density plot for samples from the leave-one-out LGP distribution for R. (d) Empirical density plot for samples from the standard LGP distribution for R.

and differences between the LGPs to be close to 0. We see from Figure H.3 that the leave-one-out form of the single-population is again the appropriate form for constructing the correct assignment probability, as it closely matches the cross-population LGP distribution.

Figure H.4 shows QQ plots of the posterior samples. Panels (a) and (b) confirm an excellent correspondence between the leave-one-out single-population distribution for population R and both of the cross-population LGP distributions of S within R and R within S. This indicates that the leave-one-out LGP distribution for a single population R correctly mimics the LGP distribution that is obtained when genotypes are compared against an independent reference sample from a population with identical allele frequencies.

Panels (c) and (d) in Figure H.4 confirm that the non-leave-one-out LGP distribution for population R differs considerably from the cross-population LGP distributions. This indicates that the non-leave-one-out protocol is not suitable for inference on population differentiation, because the two populations are drawn from identical allele frequencies and therefore should have the same distribution. Thus it is important to use the leave-one-out protocol for the single-population LGP distributions to remove this spurious separation.

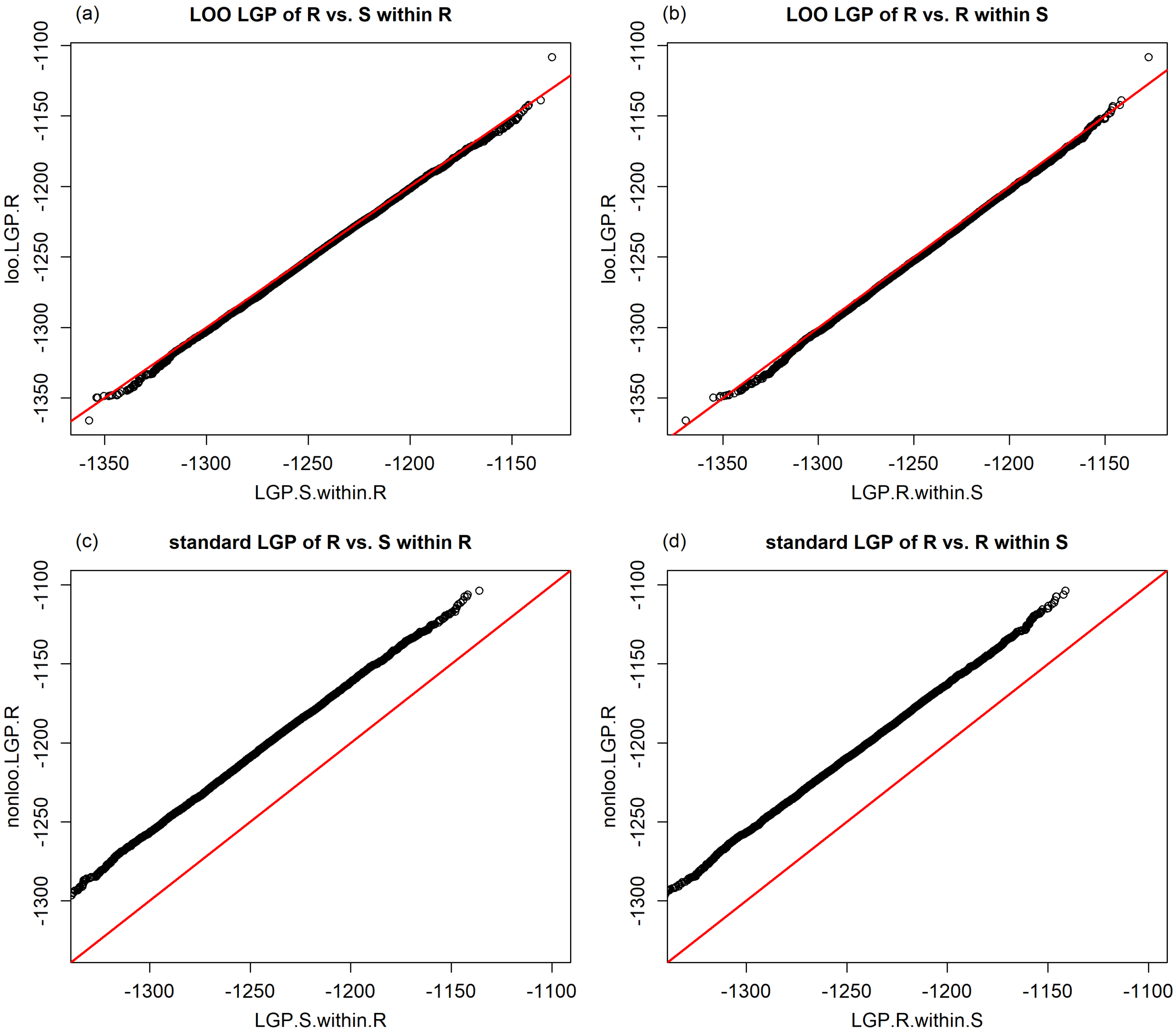


Figure H.4: QQ-plots for the cross-population LGP distributions of R within S and S within R against the leave-one-out and standard LGP distributions for R, using 10000 posterior samples from each distribution. (a) Cross-population distribution of S within R against leave-one-out distribution for R. (b) Cross-population distribution of R within S against leave-one-out distribution for R. (c) Cross-population distribution of S within R against standard distribution for R. (d) Cross-population distribution of R within S against standard distribution for R.

Chapter 4

Simulation study

In Chapter 3 we introduced three new directional measures of population differentiation: overlap area, interloper detection probability, and correct assignment probability. Here, we calculate these measures in various scenarios to see how they vary with allele frequencies, population size and other factors, and also compare them to existing measures such as G'_{ST} and D_{est}. We use purely simulated data, unlike Section 3.8, which used simulated rats bred from a real dataset.

First we test the three measures using simulated genotypes generated directly from allele frequencies, allowing us to control the allele frequencies precisely. We then test the measures using data simulated in Easypop (Balloux 2001).

4.1 Benchmark cases

We assess the behaviour of the measures at both extremes of differentiation, when the populations have identical allele frequencies or have fully non-overlapping alleles. For populations with identical allele frequencies, the overlap areas should be approximately 1 and the interloper detection probabilities and correct assignment probabilities should be approximately 0.5 if samples sizes are equal, corresponding to an even chance of choosing between the two populations. For populations with fully non-overlapping alleles, the overlap areas should be approximately 0 and all the probabilities should be approximately 1.

We also test the behaviour of the measures for populations with similar but non-identical allele frequencies, to assess how the measures change as the allele frequencies become more differentiated.

At each locus we generated a set of genotypes directly from the allele frequencies of each population, then we combined these over all loci, and calculated the measures based on the resulting samples. In most of the scenarios, the finite sample sizes created variability from the stated allele frequencies. In the populations with fully non-overlapping alleles, the alleles with zero frequency never occurred in the generated samples, so the samples always had non-overlapping observed alleles.

Table 4.1 shows the parameter values tested, and Table 4.2 shows the allele frequencies used to generate the populations in each scenario. All the populations generated were of size 10,000.

Parameter	Values tested
Number of loci	10, 50, 100, 500
Sample size	100, 1000

Table 4.1: Parameter values used for testing the directional measures overlap area, interloper detection probability and correct assignment probability in benchmark cases.

Scenario	Pop1 allele frequencies	Pop2 allele frequencies
A.1	(0.5, 0.5)	(0.5, 0.5)
A.2	(0.8, 0.2)	(0.8, 0.2)
A.3	(0.4, 0.4, 0.05, 0.05, 0.05, 0.05)	(0.4, 0.4, 0.05, 0.05, 0.05, 0.05)
A.4	8 alleles with equal frequency	8 alleles with equal frequency
B.1	(0.4, 0.4, 0.2, 0, 0, 0)	(0, 0, 0, 0.2, 0.4, 0.4)
B.2	(0.25, 0.25, 0.25, 0.25, 0, 0, 0, 0)	(0, 0, 0, 0, 0.25, 0.25, 0.25, 0.25)
B.3	(0.6, 0.4, 0, 0)	(0, 0, 0.4, 0.6)
C.1	(0.6, 0.4)	(0.55, 0.45)
C.2	(0.8, 0.2)	(0.75, 0.25)
C.3	(0.8, 0.2)	(0.7, 0.3)

Table 4.2: Allele frequencies used to generate simulated populations for testing the directional measures overlap area, interloper detection probability and correct assignment probability in benchmark cases.

Figure 4.1 shows results from scenarios A.1 to A.4 in Table 4.2, for 10 replicates of 500 loci and samples of size 100. These are all scenarios where the populations have identical underlying allele frequencies, although the actual allele frequencies in each replicate sample differ due to sampling variation. As a result, the overlap area results are not all 1, although they are within about 0.1 of 1, which is as close as we'd expect to see given the level of sampling variability for samples of size 100.

The overlap areas for scenario A.1 are closer to 1 than the overlap areas for the other scenarios. Scenario A.1 has only two alleles at each locus, and they both have the same frequency, and are thus equally common alleles. The other scenarios A.2 to A.4 have some common and some rare alleles at each locus, and this more uneven distribution of allele frequencies increases the effect of the sampling variation on the observed allele frequencies, and thus it is easier to distinguish between the populations in scenarios A.2 to A.4 than in A.1.

The interloper detection probabilities and correct assignment probabilities shown in Figure 4.1 are approximately centred on 0.5, as expected. The results are between about 0.4 and

0.6, which is also as expected, since for samples of size 100 we would expect the observed allele frequencies to differ quite a lot from the true underlying allele frequencies, due to sampling variability. For example, it would not be surprising to see observed allele frequencies 0.45 and 0.55 at a single locus, instead of 0.5 and 0.5, and that sampling variability is then compounded over 500 loci.

Figure 4.2 shows similar results to Figure 4.1, but the sampling uncertainty has been reduced by taking samples of size 1000 instead of 100. The overlap areas are closer to 1 and less variable than in Figure 4.1, and the interloper detection probabilities and correct assignment probabilities are also closer to 0.5 and less variable than in Figure 4.1.

All the results for scenarios A.1 to A.4 were similar for 500 loci or for 10, 50 or 100 loci.

For scenarios B.1 to B.3, with samples of size 100 or 1000, the overlap areas were exactly 0 and the interloper detection probabilities and correct assignment probabilities were exactly 1. These scenarios correspond to populations with fully non-overlapping alleles.

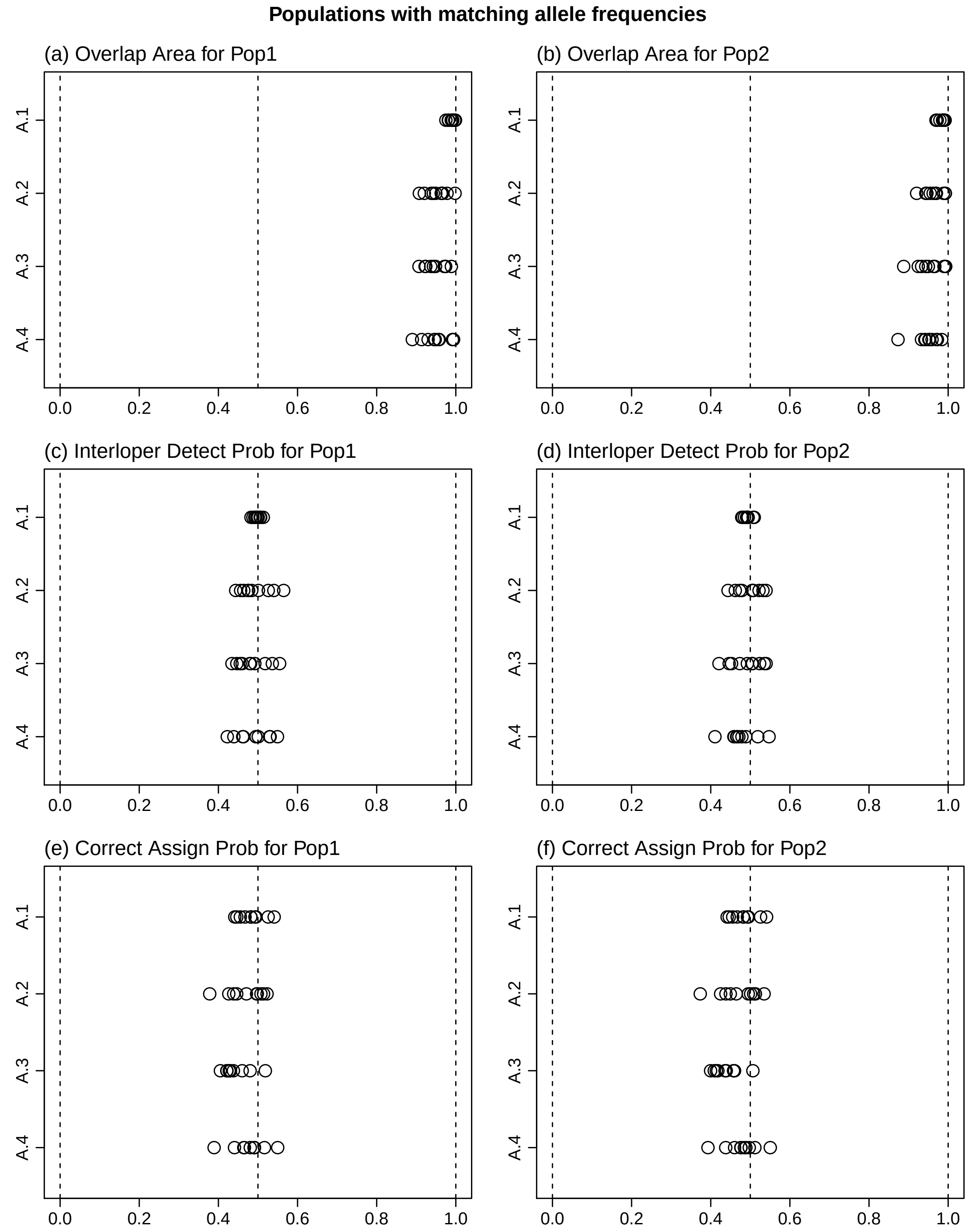


Figure 4.1: Directional measure results from benchmark scenarios A.1 to A.4, involving two populations with identical allele frequencies. Allele frequencies used to generate the populations are shown in Table 4.2. 10 replicates were used with 500 loci and samples of size 100.

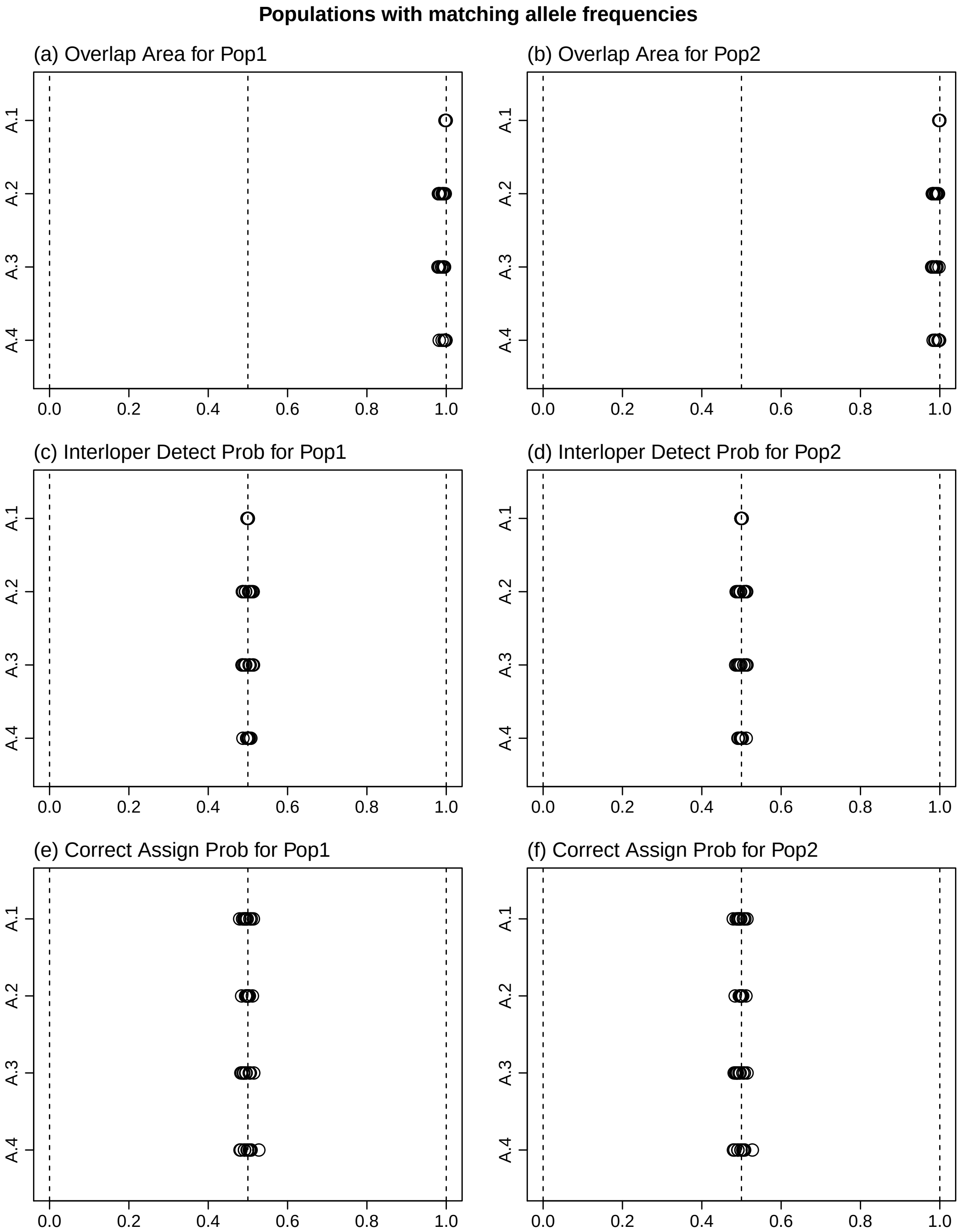


Figure 4.2: Directional measure results from benchmark scenarios A.1 to A.4, involving two populations with identical allele frequencies. Allele frequencies used to generate the populations are shown in Table 4.2. 10 replicates were used with 500 loci and samples of size 1000.

Figure 4.3 shows results from scenarios C.1 to C.3, for 10 replicates of 500 loci and samples of size 100. These are all scenarios where the populations have similar allele frequencies to each other, although again the observed frequencies are subject to sampling variation.

For scenario C.1 the overlap areas for Pop1 are around 0.3 to 0.5, and the overlap areas for Pop2 are around 0.7, and there is some variability in the results. The correct assignment probabilities are all around 0.8 to 0.9, indicating that even a difference of 0.05 between the allele frequencies in the two populations, over 500 loci, is enough to make the populations genetically distinct to the extent that it is fairly easy to assign individuals from either population correctly.

For scenario C.2 the overlap areas are around 0.1, and the correct assignment probabilities are around 0.9, even though the difference between the allele frequencies is 0.05, the same as for scenario C.1. This result is as we would expect: variance is proportional to $p(1-p)$ for a binomial distribution with frequency p, and the allele frequencies are nearer to 0.5 in scenario C.1 than in scenario C.2, and therefore we would expect to see greater tolerance of small deviations of allele frequency from scenario C.1 than from scenario C.2. The same effect presumably explains why overlap area within baseline Pop2 in scenario C.1 is higher than overlap area within baseline Pop1, because Pop2 has allele frequencies closer to 0.5, which creates greater tolerance to small deviations than in Pop1.

For scenario C.3 the allele frequencies differ more than in scenarios C.1 and C.2, and as seen in Figure 4.3 the resulting overlap areas are around 0, and the correct assignment probabilities are around 1, indicating that it has become even easier than in scenario C.2 to distinguish between the populations.

For all three scenarios C.1 to C.3 the interloper detection probabilities are extremely asymmetric. For scenario C.1 the probability for Pop1 is around 0.8 to 0.9, and the probability for Pop2 is around 0.3 to 0.4; for scenarios C.2 and C.3 the probability for Pop1 is close to 1 and the probability for Pop2 is close to 0. This indicates that when two populations have similar allele frequencies and the same major allele, then a choice between two individuals as to which has the best fit to either population will tend to favour those from the population with the higher major allele frequency, and this behaviour becomes stronger as the major allele frequency increases. The interloper detection probabilities within Pop2 are close to zero because individuals from Pop1 would fit better into Pop2 than the individuals from Pop2 themselves. In fact, the fit of Pop1 individuals into Pop2 is so much better that the cross-population LGP distribution of Pop1 into Pop2 sits entirely above the LGP distribution of Pop2 itself, so that the overlap areas within baseline Pop2 are also close to zero. Fortunately this counterintuitive outcome is unlikely to occur in practice, as it is unlikely that one population would have the higher major allele frequency across all loci.

We obtained similar results for samples of size 1000, although with less variability. However, the results for these scenarios are far more affected by the number of loci than the results of scenarios A.1 to A.4 and B.1 to B.3. Figure 4.4 shows the results of scenarios C.1 to C.3 for 10 replicates of 10 loci and samples of size 100. The results are less extreme and show far more variability than those in Figure 4.3. The overlap areas are much greater, the interloper detection results are somewhat less asymmetric, and the correct assignment probabilities are much lower than in Figure 4.3.

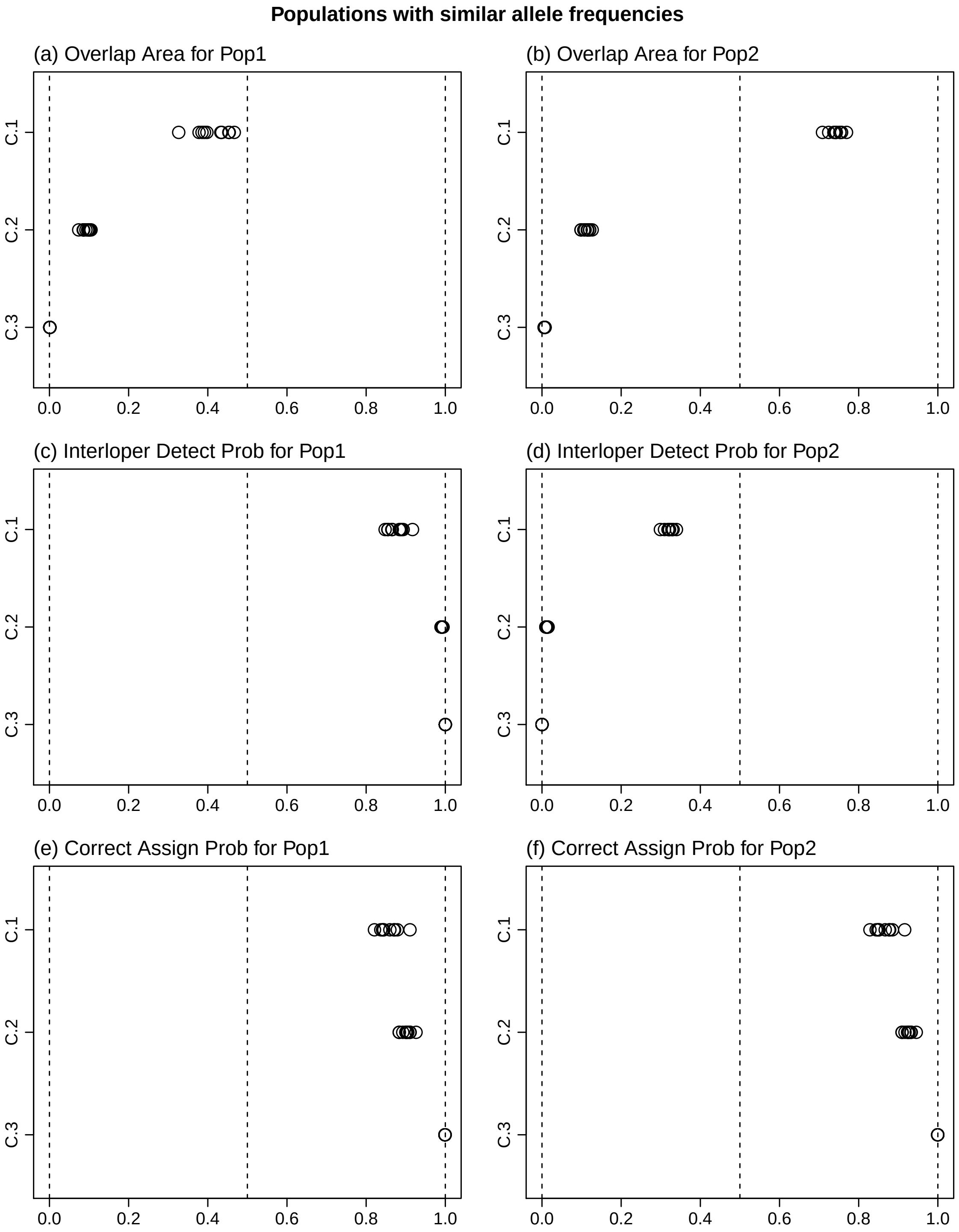


Figure 4.3: Directional measure results from benchmark scenarios C.1 to C.3, involving two populations with similar allele frequencies. Allele frequencies used to generate the populations are shown in Table 4.2. 10 replicates were used with 500 loci and samples of size 100.

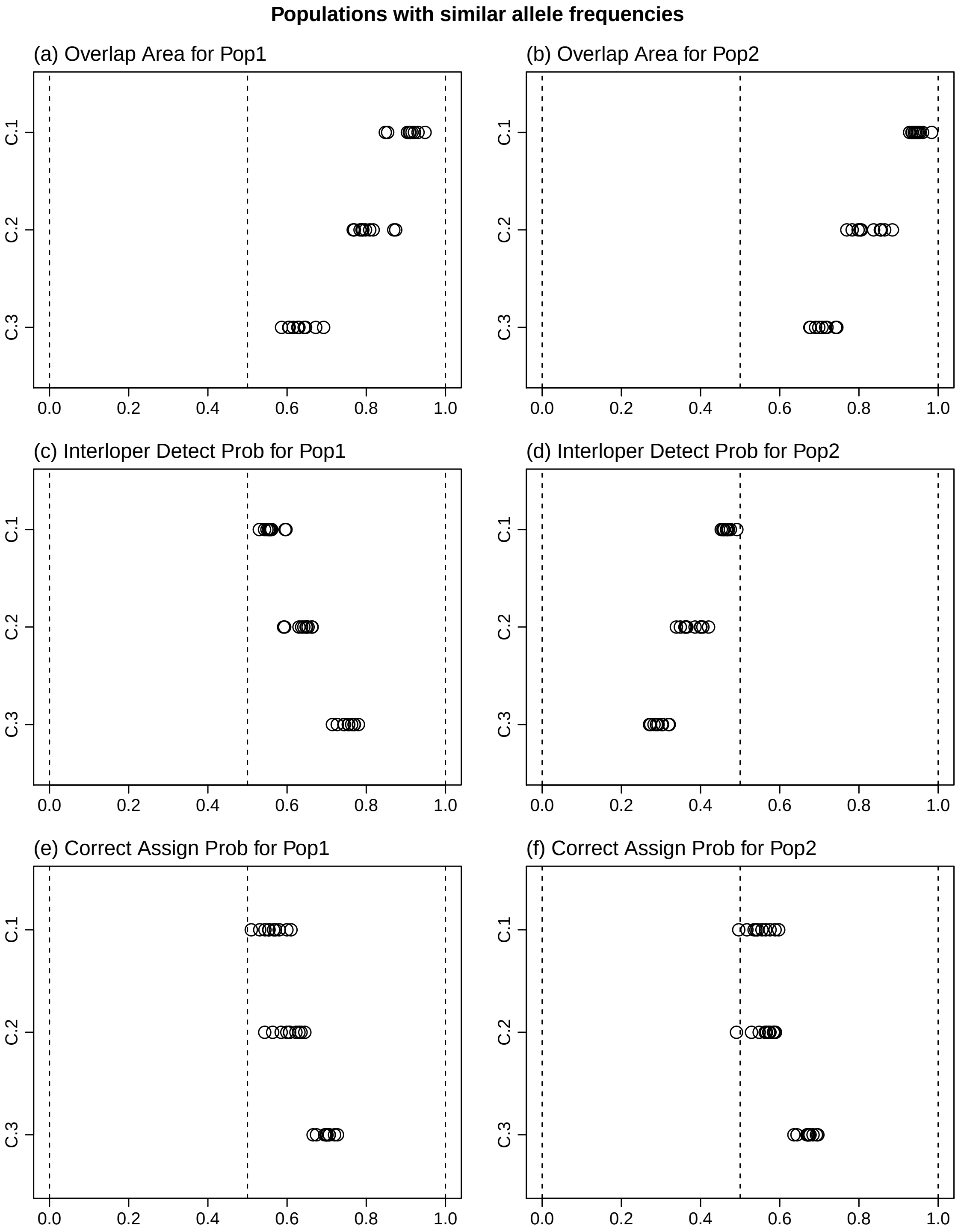


Figure 4.4: Directional measure results from benchmark scenarios C.1 to C.3, involving two populations with similar allele frequencies. Allele frequencies used to generate the populations are shown in Table 4.2. 10 replicates were used with 10 loci and samples of size 100.

We also tested one final benchmark scenario, in which Pop1 has frequencies (0.3, 0.3, 0.3, 0.1), three common alleles and one less common, and Pop2 has frequencies (0.8, 0.05, 0.05, 0.1), one very common allele and three rare alleles. Figure 4.5 shows the combined GenePlot and LGP plot for one replicate from the simulations with 10 loci and samples of size 100. The other replicates produced similar results.

Figure 4.5 shows the extremely directional nature of this scenario. Almost all of the simulated individuals are above the 1% percentile line for Pop1, so they would have a reasonable fit to Pop1, but only individuals from Pop2 are above the 1% line for Pop2. The overlap area for Pop1 (0.471) is much higher than the overlap area for Pop2 (0.014), and similarly the interloper detection probability for Pop1 (0.827) is lower than the interloper detection probability for Pop2 (1.000). The correct assignment probabilities are 0.995 and 0.994 respectively, which is reflected in the GenePlot which shows the samples separated on either side of the central diagonal, apart from a single Pop2 individual which is on the Pop1 side.

The results for 50, 100 and 500 loci were more extreme, with the samples becoming more separated as the number of loci increased, so that for 500 loci the overlap areas are both approximately 0 for all replicates, and the interloper detection and correct assignment probabilities are all approximately 1 for all replicates.

Overall, this scenario illustrates that populations with a wider range of common alleles will tend to have higher overlap areas and lower interloper detection probabilities than populations with fewer common alleles. Thus these measures are indicators of diversity, not just of differentiation, whereas the correct assignment probabilities are more directly targeted at differentiation.

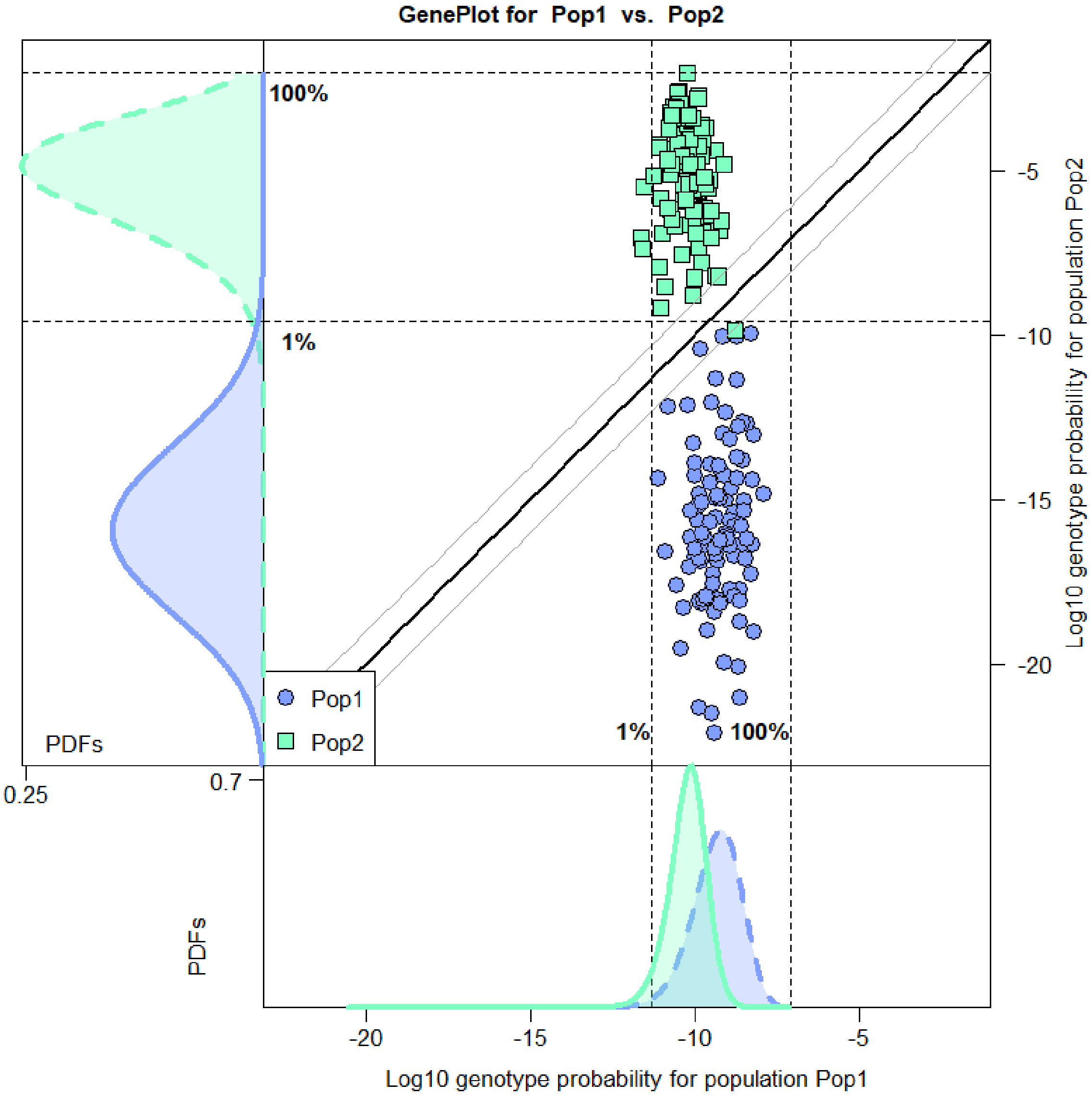


Figure 4.5: GenePlot and LGP distribution plots for benchmark scenario with allele frequencies (0.3, 0.3, 0.3, 0.1) in Pop1 and (0.8, 0.05, 0.05, 0.1) in Pop2, for 10 loci and samples of size 100. One of 10 replicates is shown. Details as in Figure 3.1.

4.2 Easypop simulations

We simulated a set of scenarios using Easypop (Balloux 2001) to assess the behaviour of the three measures in scenarios that are more realistic than the benchmark cases in Section 4.1.

Most simulations used to test population genetics methods use scenarios with equal-sized populations, which is unrealistic. Since our measures can reveal additional information in scenarios with disparate populations, we ran simulations using three populations Pop1, Pop2 and Pop3, with sizes 1000, 500 and 100 respectively.

Table 4.3 shows the parameter values we used in Easypop version 2.0.1, and Table 4.4 shows the migration rate and number of generations used in each scenario. Details of these settings are in the Easypop User Guide:
https://www.unil.ch/files/live/sites/dee/files/shared/softs/EASYPOP_201_userguide.pdf.

Parameter set	Details
Mating scheme	One sex, random mating
Migration scheme	Island model
Loci	20, with free recombination and 20 allelic states
Mutation scheme	KAM model*, with mutation rate $\mu = 0.0001$
Variability of initial population	Maximal

*The KAM model is a form of the infinite alleles model with a finite number of allelic states.

Table 4.3: Settings used for testing the directional measures overlap area, interloper detection probability and correct assignment probability in Easypop scenarios.

Scenario	Migration rate m	Number of generations
A	0.05	500
B	0.01	5000
C	0.01	2000
D	0.01	1000
E	0.005	5000
F	0.005	2000
G	0.005	1000
H	0.001	10000

Table 4.4: Migration rates and number of generations used for testing the directional measures overlap area, interloper detection probability and correct assignment probability in Easypop scenarios.

Scenarios A to H are expected to show approximately increasing levels of differentiation, given the migration rates and number of generations for those scenarios. The initial populations have randomly allocated alleles, and therefore, for a given migration rate, connectivity between the populations will increase over time as the populations gain alleles from immigrant individuals. Scenarios B, C and D have the same migration rate, but scenario B has a longer run time than scenario C, measured as number of generations, so we would expect to see higher connectivity in scenario B than scenario C, and scenario C has a longer run time than scenario D, so we would expect to see higher connectivity in scenario C than scenario D. A similar pattern holds for scenarios E, F and G.

After generating the data using Easypop, we took random samples of size 50 from each population, and calculated the values of our three measures based on those samples.

We also calculated the values of existing measures D_{est} (Jost 2008) and G'_{ST} for each scenario. We calculated D_{est} and G'_{ST} using the bias-corrected forms of H_{S} and H_{T}, as proposed by Nei (Nei & Chesser, 1983), and using Hedrick's (2005) form of G'_{ST} from (1.3):

$$G'_{\text{ST}} = \frac{G_{\text{ST}}}{G_{\text{ST(max)}}}\,, \tag{4.1}$$

where $G_{\text{ST(max)}}$ is calculated using $H_{\text{T(max)}}$ from (1.4):

$$H_{\text{T(max)}} = \frac{1}{K}(K - 1 + H_{\text{S}})\,, \tag{4.2}$$

and K is the number of populations. We tested our single-locus D_{est} and G'_{ST} calculations against the `diveRsity`, `DEMEtics`, and `mmod` R packages, and calculated the global measures by taking the arithmetic mean of each measure over all loci.

Figure 4.6 shows our three measures for Pop1 and Pop2, with sizes 1000 and 500, and samples of size 50. As expected, the overlap areas show a decreasing trend from scenario A to scenario H, although there is a large amount of variability. The overlap values range from 0 to 1; the interloper detection probability values range from about 0.4 to 1; and the correct assignment probability values also range from 0.4 to 1 but tend to be higher than the interloper detection probability values.

We obtained similar results for Pop2 and Pop3 as for Pop1 and Pop2.

Figure 4.7 shows our three measures for Pop1 and Pop3, with sizes 1000 and 100, and samples of size 50. The most notable differences between Figure 4.7 and Figure 4.6 are that in Figure 4.7 the overlap areas for Pop1 only go as low as about 0.2, and the interloper detection probabilities for Pop1 are more variable and tend to be somewhat lower than in Figure 4.6. The measures for Pop2 are similar in Figure 4.6 and Figure 4.7. These results demonstrate that our measures are sensitive to changes in population size.

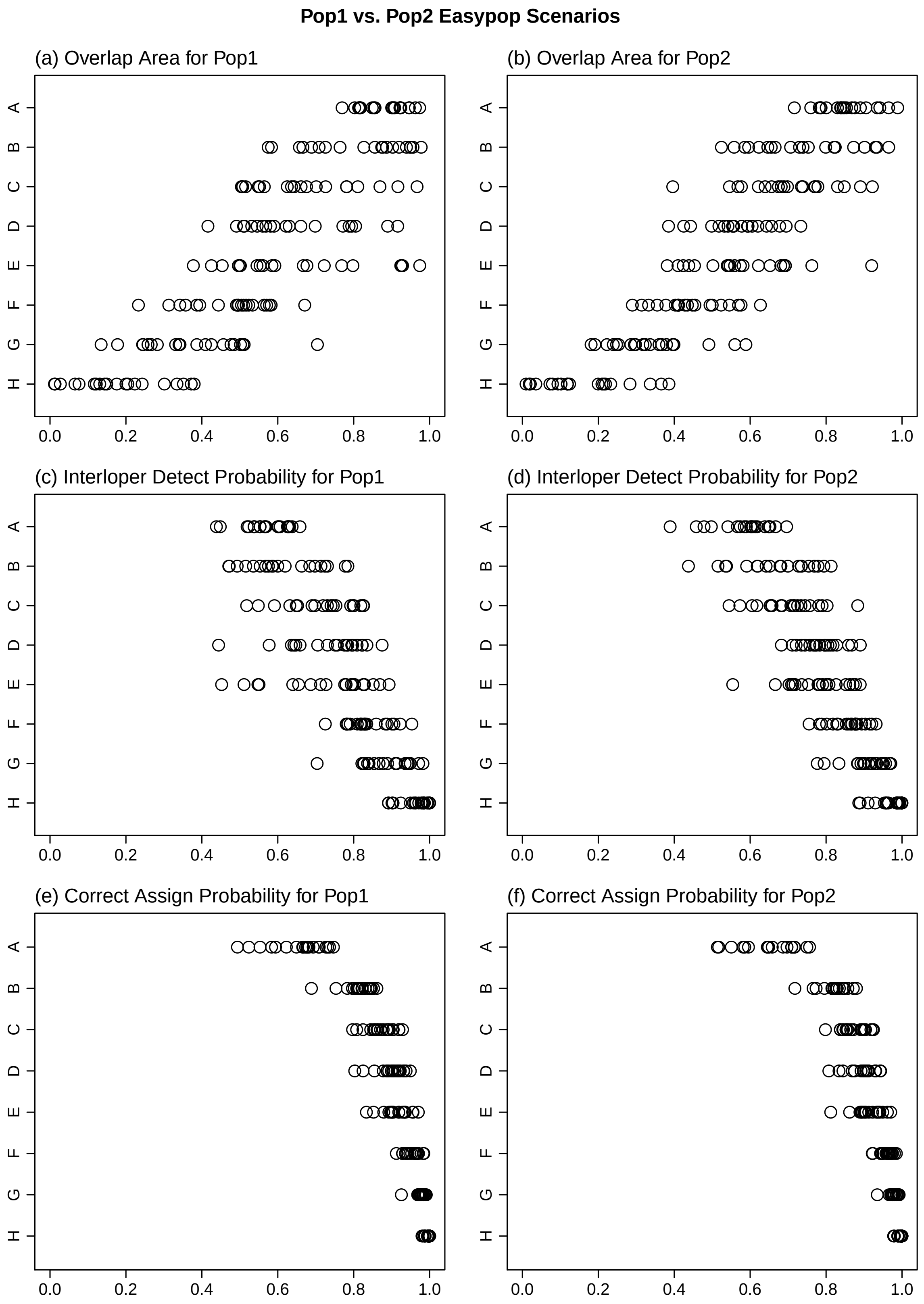


Figure 4.6: Directional measure results for populations Pop1 and Pop2 from Easypop scenarios. Scenario details are shown in Table 4.4. 20 replicates were used for each scenario.

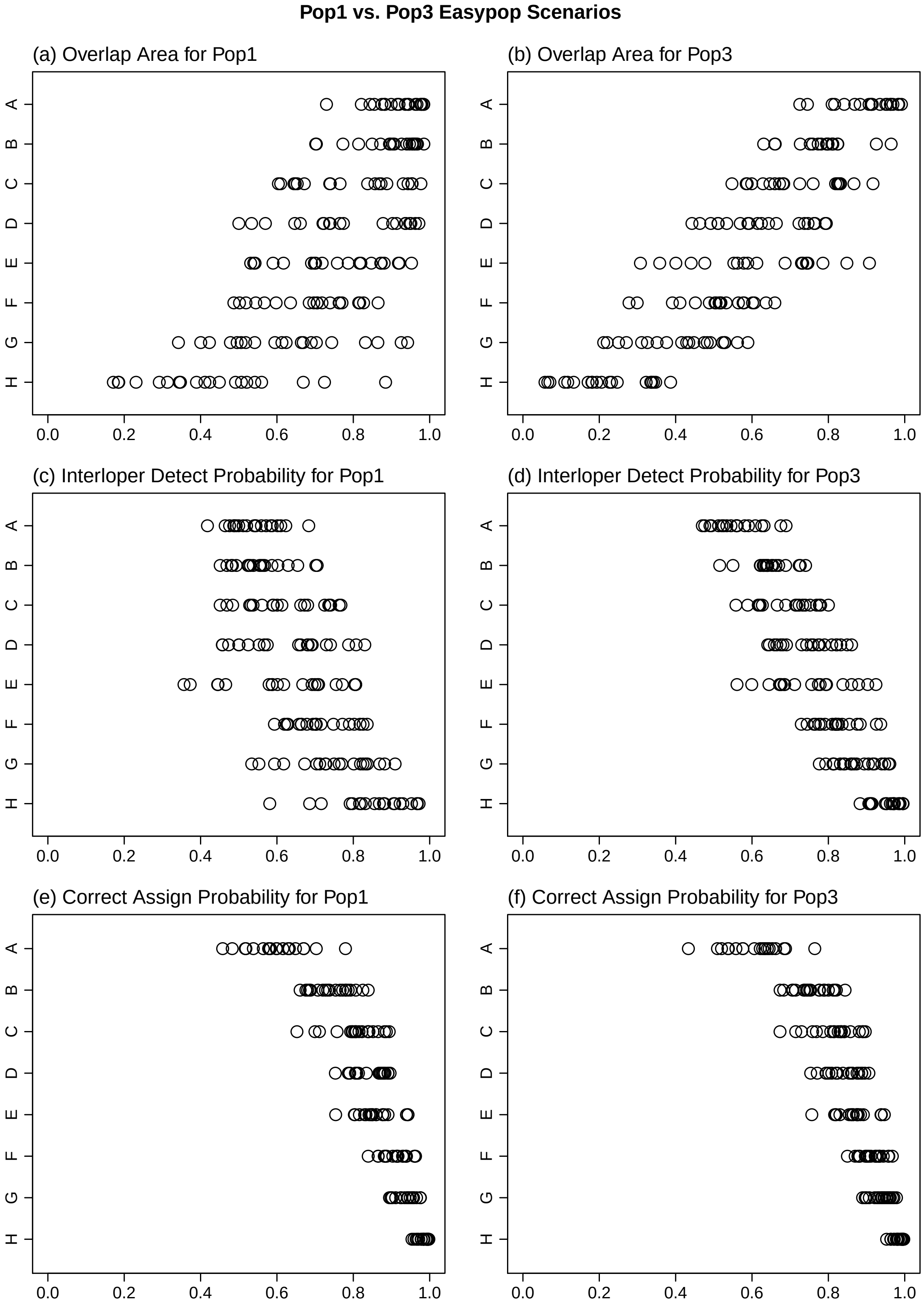


Figure 4.7: Directional measure results for populations Pop1 and Pop3 from Easypop scenarios. Scenario details are shown in Table 4.4. 20 replicates were used for each scenario.

Figure 4.8 shows our three measures for Pop1 and Pop2, plotted against D_{est}. The D_{est} values are below 0.2 in all cases, but the overlap area values range between 1 and 0. This implies that the overlap area measure reaches its minimum value of 0 for populations that may share some alleles, but also that the overlap area measure is more sensitive than D_{est} for populations that are not highly differentiated. The overlap area measure shows approximate negative correlation with D_{est}.

The interloper detection and correct assignment probabilities range from around 0.5 to 1, and are non-linearly correlated with D_{est}. The correct assignment probabilities tend to be higher than the interloper detection probabilities, as we have observed with other datasets. Both probabilistic measures reach their maximum values for fairly low levels of D_{est}. These results suggest that the new measures may be less sensitive for assessing strongly separated populations than classical measures, but more sensitive for populations that show low differentiation according to classical measures. However, in practical situations the effective range of D_{est} values would not be known, so a result of 0.1 or 0.2 would be hard to interpret for a single case.

We obtained similar results for Pop2 and Pop3 plotted against D_{est} as for Pop1 and Pop2 plotted against D_{est}.

Figure 4.9 shows our three measures for Pop1 and Pop3, plotted against D_{est}. These results show weaker correlation with D_{est} than the results in Figure 4.8.

The overlap areas for Pop3 within baseline Pop1 in (a) of Figure 4.9 tend to be higher than the overlap areas for Pop2 with baseline Pop1 in (a) of Figure 4.8. The corresponding interloper detection probabilities in (b) tend to be lower in (c) of Figure 4.9 than in (c) of Figure 4.8. The D_{est} values between Pop1 and Pop2 are similar to the D_{est} values between Pop1 and Pop3.

These results confirm that D_{est} is not affected much by population size, as argued by Meirmans & Hedrick (2011). We would expect Pop3, which is size 100, to have drifted more than Pop2, which is size 500, and most measures of genetic differentiation would reflect this and would show different patterns in Figure 4.8 compared with Figure 4.9. By contrast, the patterns of D_{est} are the same in Figure 4.8 as in Figure 4.9. For most ecological applications, it is useful to be able to detect differences in population size, and thus our measures, which are sensitive to population size, should be more useful in a practical context.

Figure 4.9 also shows more asymmetry between the populations than Figure 4.8. In Figure 4.9, the overlap areas for baseline Pop3 in (b) tend to be lower than the overlap areas for baseline Pop1 in (a), although the D_{est} values are exactly the same, since it is a symmetric measure.

Figure 4.10 shows our three measures for Pop1 and Pop2, plotted against G'_{ST}. The G'_{ST} values are below 0.35 in all cases, covering a wider range than the D_{est} values, and our three measures are more tightly associated with G'_{ST} than D_{est}, although the relationships between the probabilistic measures and G'_{ST} are still nonlinear.

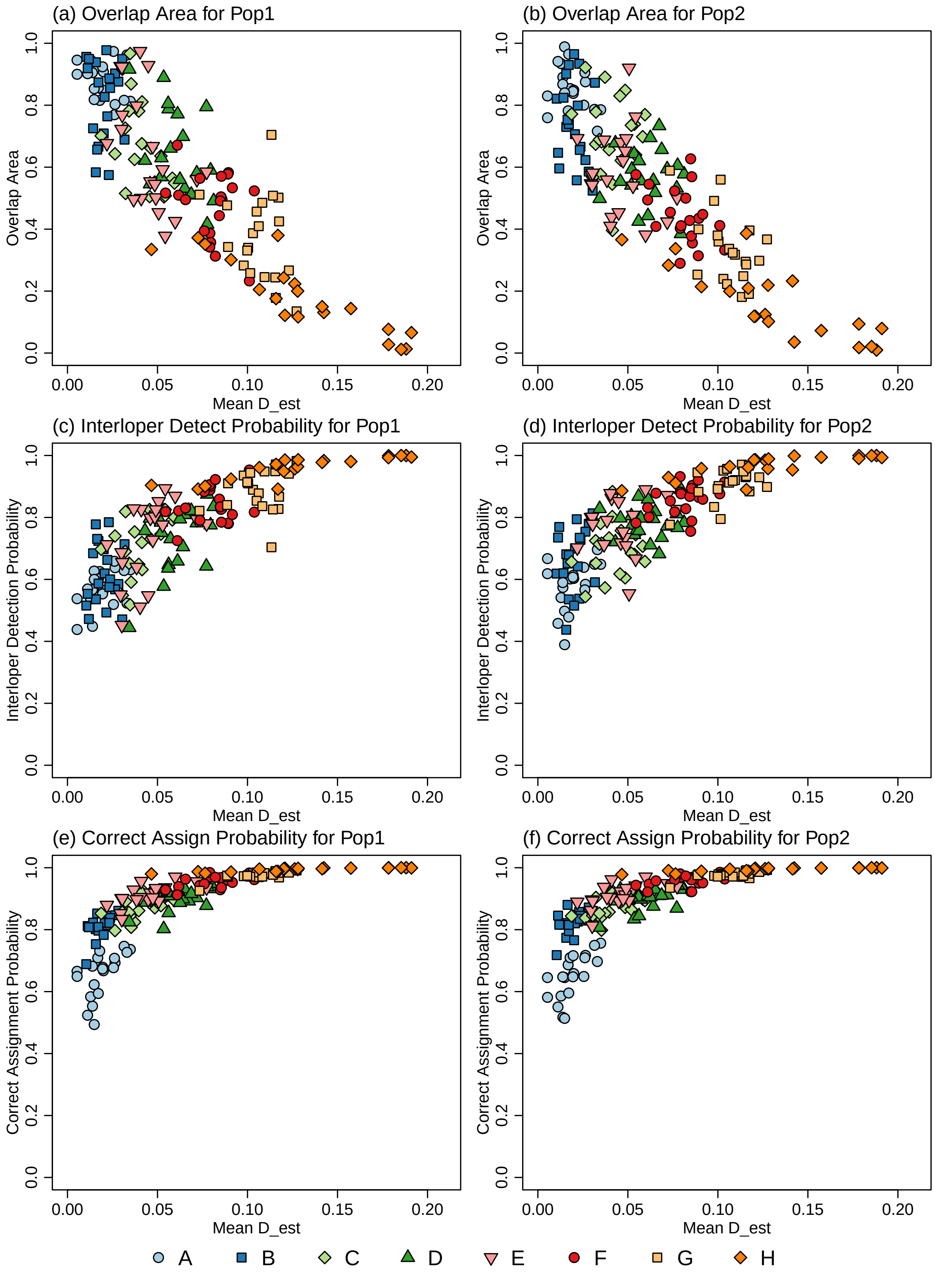


Figure 4.8: Directional measure results for populations Pop1 and Pop2 from Easypop scenarios, plotted against D_{est}. Scenario details are shown in Table 4.4. 20 replicates were used for each scenario.

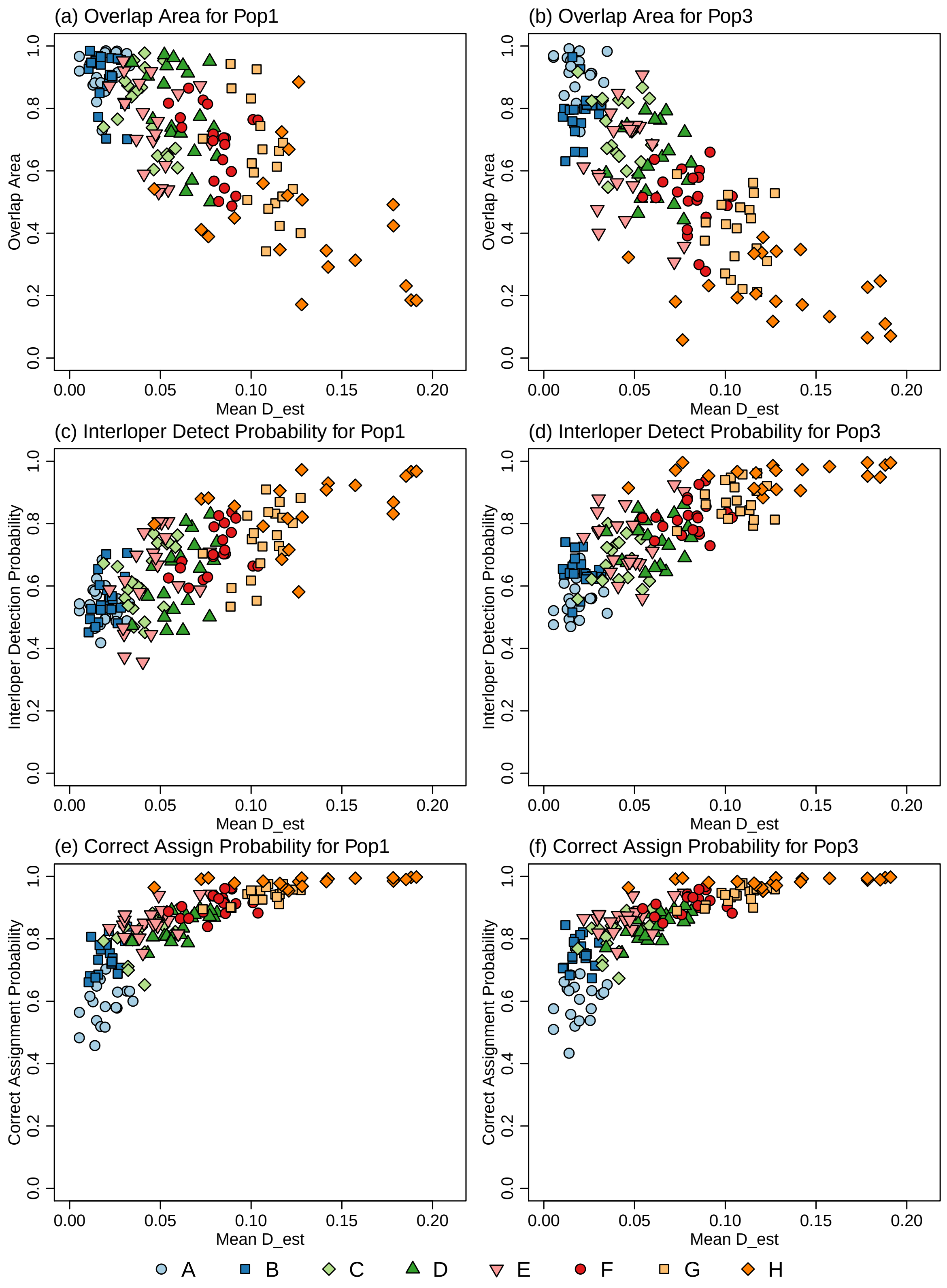


Figure 4.9: Directional measure results for populations Pop1 and Pop3 from Easypop scenarios, plotted against D_{est}. Scenario details are shown in Table 4.4. 20 replicates were used for each scenario.

The comparison between (Pop1, Pop2) and (Pop1, Pop3) for G'_{ST} was similar to the comparison for D_{est} and we drew similar conclusions. Again, we obtained similar results for Pop2 and Pop3 plotted against G'_{ST} as for Pop1 and Pop2 plotted against G'_{ST}.

We reran the measure calculations for samples of size 100 from the same simulated populations, and found that the results were very similar to those for samples of size 50, except that the values of D_{est} were less variable, in general, and the lowest values of D_{est} were slightly higher, with a minimum of about 0.01.

Finally, we reran three of the Easypop scenarios with three populations of size 100 and compared them to the previous scenarios with populations of sizes 1000, 500 and 100. These scenarios are shown in Table 4.5. The new scenarios J, L and N correspond to scenarios C, D and G from Table 4.4.

Scenario	Migration rate m	Number of generations	Population sizes
I	0.01	2000	(100, 100, 100)
J	0.01	2000	(1000, 500, 100)
K	0.01	1000	(100, 100, 100)
L	0.01	1000	(1000, 500, 100)
M	0.005	1000	(100, 100, 100)
N	0.005	1000	(1000, 500, 100)

Table 4.5: Migration rates and number of generations used for testing the directional measures overlap area, interloper detection probability and correct assignment probability in Easypop scenarios with mixed sets of population sizes.

Figure 4.11 shows the results for Pop1 and Pop3, plotted against D_{est}. The results for scenarios M and N in Figure 4.11 differ more than the results for scenarios I and J, or K and L, as we would expect, because the populations in scenarios M and N are more isolated than the populations in scenarios I and J, or K and L. For scenario N the overlap areas and interloper detection probabilities for Pop1 are very different to the overlap areas and interloper detection probabilities for Pop3. The same pattern is seen, to a lesser extent, in scenario L, and, to an even lesser extent, for scenario J.

The D_{est} values for scenarios M and N differ less than the overlap areas or the interloper detection probabilities. This again confirms that our measures are more sensitive to changes in population size than D_{est}. This is appropriate because population size affects genetic diversity and the rate of genetic drift, and our measures aim to detect these features of the population structure.

Figure 4.12 shows the results from plotting Pop1 and Pop3 against G'_{ST}. The G'_{ST} values for scenario N are very different to the G'_{ST} values for scenario M, and a similar, but weaker, pattern is seen for scenarios L and K. G'_{ST} therefore shows more sensitivity to population size than D_{est}, but the overlap area and interloper detection probability measures pick up the disparity between the sizes of populations Pop1 and Pop3 that cannot be shown using G'_{ST} or D_{est}.

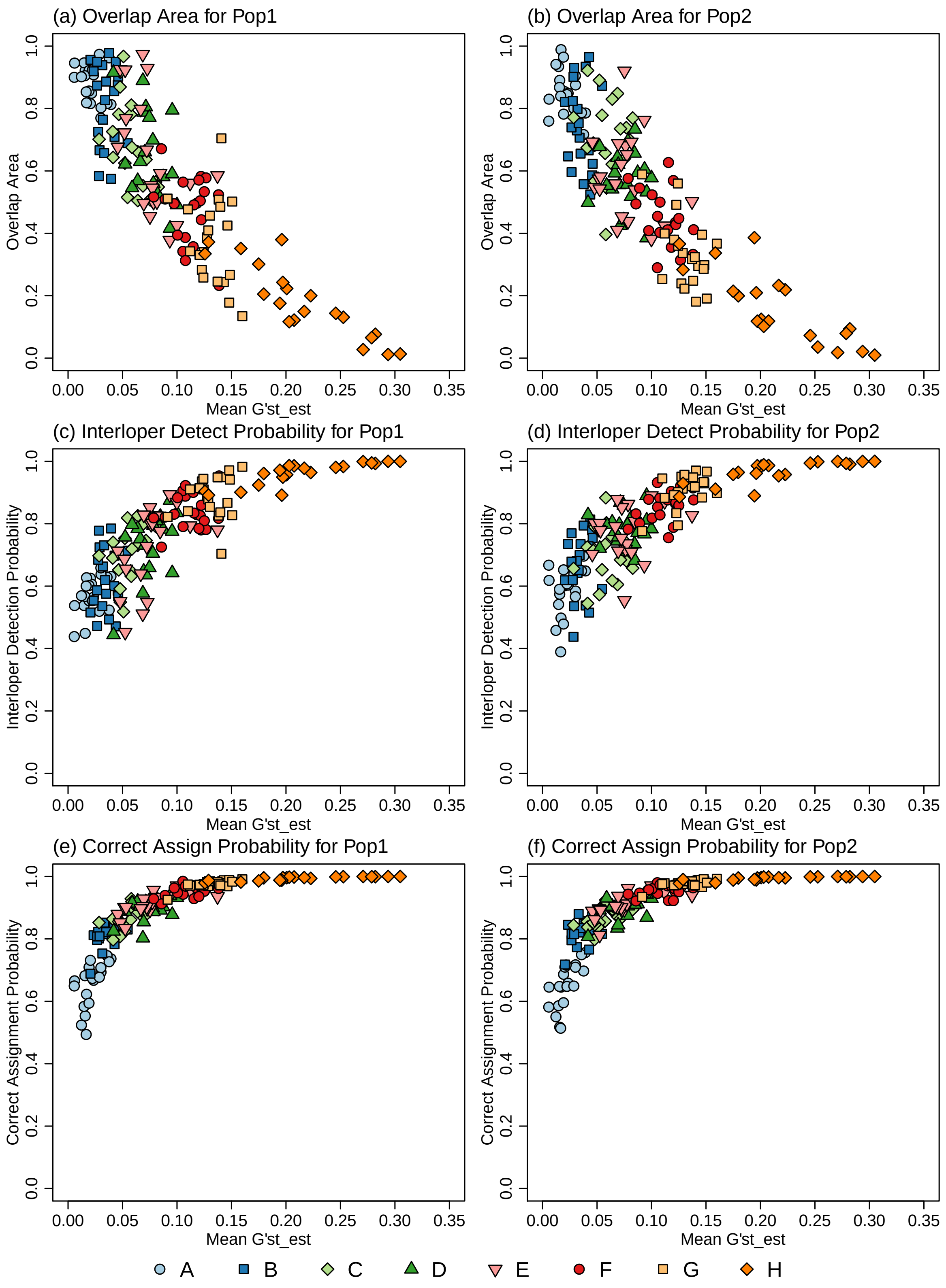


Figure 4.10: Directional measure results for populations Pop1 and Pop2 from Easypop scenarios, plotted against G'_{ST}. Scenario details are shown in Table 4.4. 20 replicates were used for each scenario.

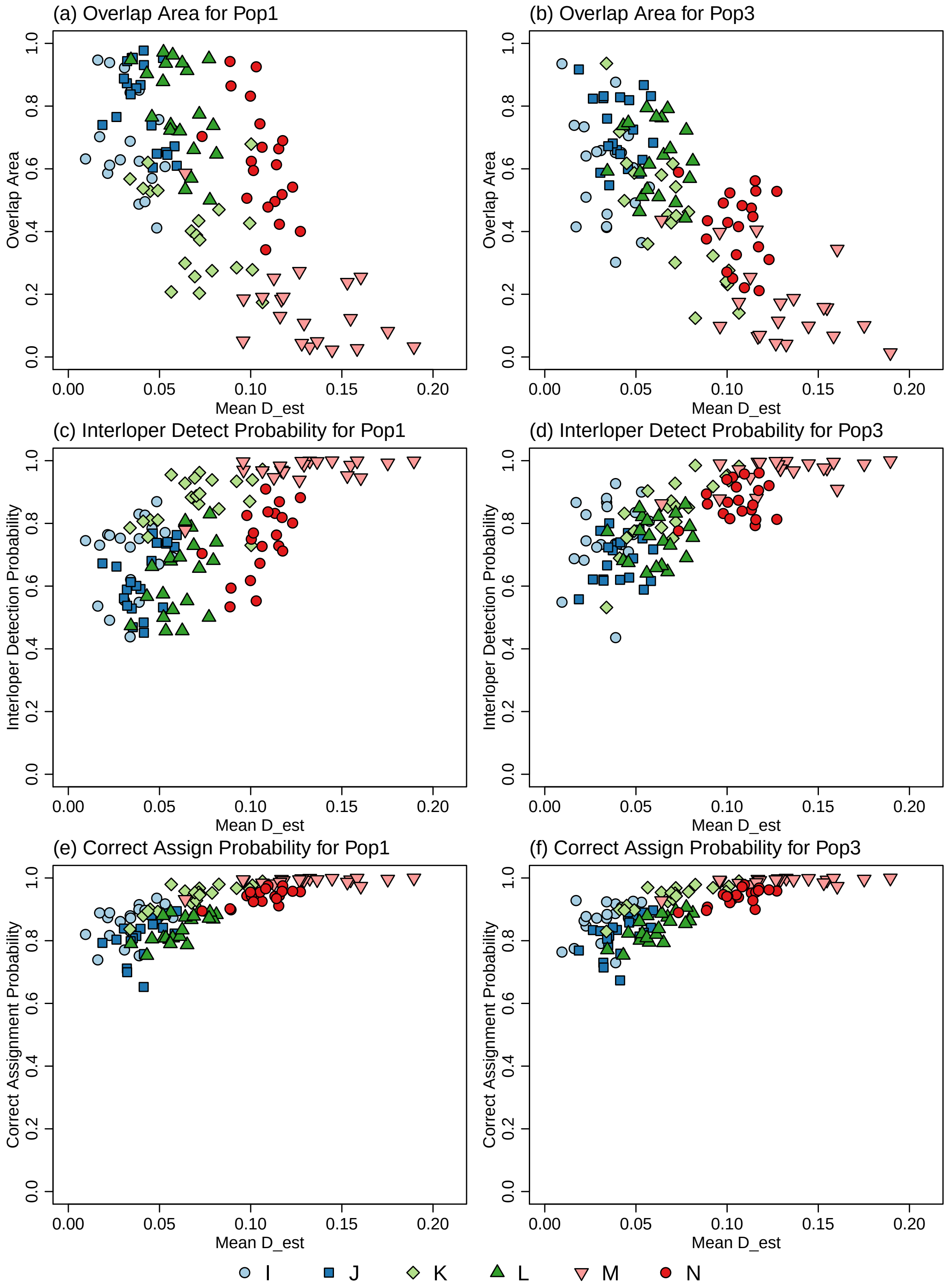


Figure 4.11: Directional measure results for populations Pop1 and Pop3 from Easypop scenarios with mixed sets of population sizes, plotted against D_{est}. Scenario details are shown in Table 4.5. 20 replicates were used for each scenario.

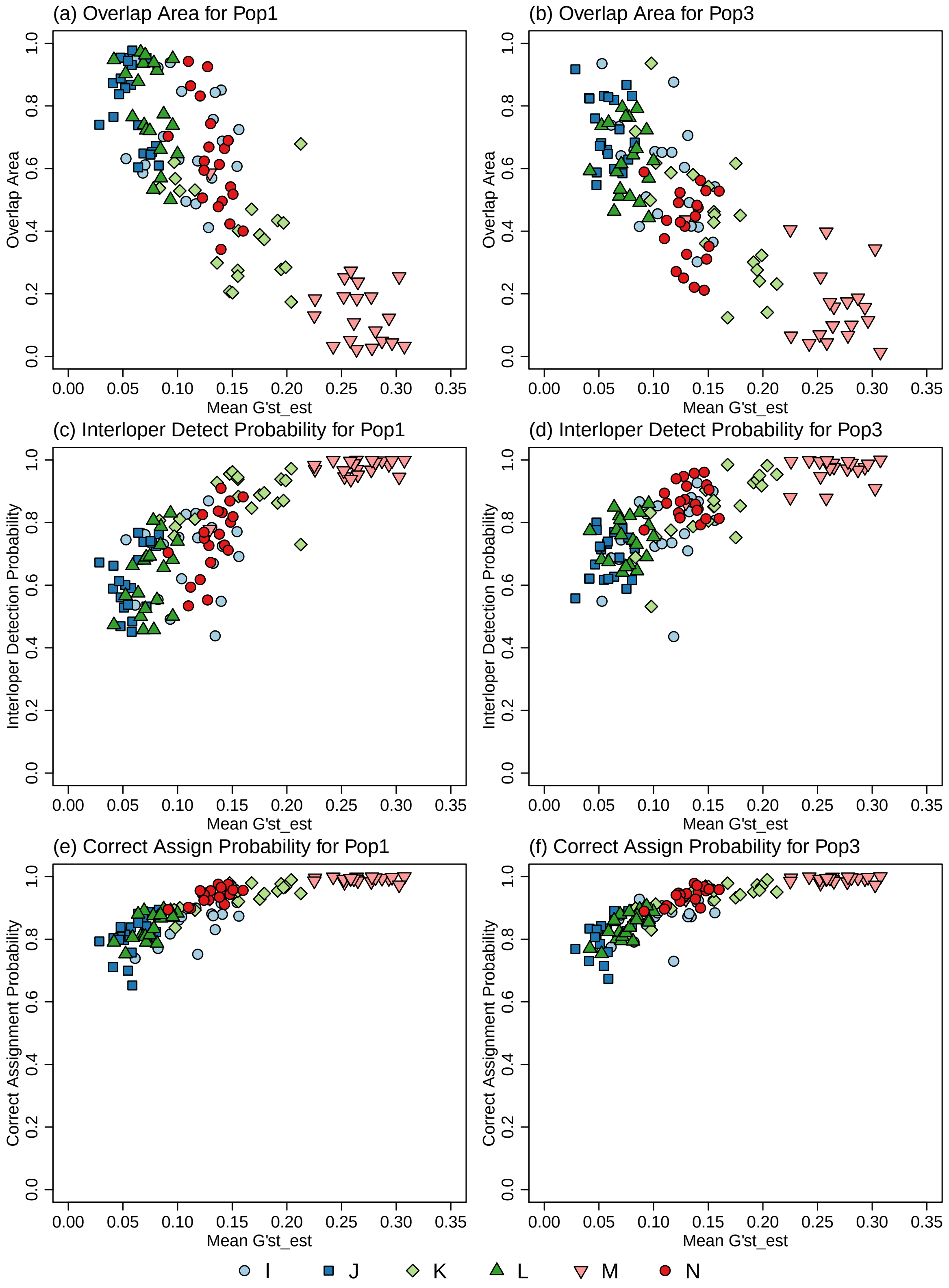


Figure 4.12: Directional measure results for populations Pop1 and Pop3 from Easypop scenarios with mixed sets of population sizes, plotted against G'_{ST}. Scenario details are shown in Table 4.5. 20 replicates were used for each scenario.

Chapter 5

Conclusions

It is often said that "a picture is worth a thousand words". That is not always the case — but it is worth using one hundred words to explain a picture. The human brain processes images in a distinctly different way than language, and thus, if we can combine a meaningful image with appropriate explanatory words, we can provide information in a form that the reader can understand in more ways than one. Typically, a plot is also more useful than a table of numbers, for those numbers require more complex processing to elicit patterns that may be seen at a glance from a suitable plot.

GeneClass2 is still a popular tool for assignment (e.g. Larson et al. 2014, Johnston et al. 2014, Benestan et al. 2015), but it does not provide any visualization of the results. We have shown that our new method for visualizing log-genotype probabilities in GenePlots is valuable for correct interpretation of assignment results. We have demonstrated GenePlots based on microsatellite and SNP data, and shown how they can illuminate new features of the data that were not clear using existing methods.

We have also shown that LGP distributions are very unusual distributions, representing up to hundreds of orders of magnitude depending on the number of loci, as well as being extremely left-skewed. Our saddlepoint method creates an accurate approximation to this distribution, even in the long left tail. This creates a fast, analytic computation of quantiles of the LGP distribution, enabling us to plot all individuals in the same chart even if they have missing data.

We have shown how this analytic approximation to the LGP distribution may be further extended to compare the saddlepoint-PDFs of single-population and cross-population LGP distributions, extending the methodology from an assignment technique into a way to visualize and quantify population structure. The leave-one-out form of the saddlepoint approximation accurately mimics the distribution obtained when all individuals are compared against an independent reference sample, removing the bias in single-population LGP distributions that exaggerates the fit of individuals in the reference sample.

Our methodology combines powerful visualization with three numeric measures to summarize the extent to which we can differentiate between the populations based on the available data. The overlap area measure corresponds to an intuitive visual aspect of the GenePlots, and the interloper detection probability and correct assignment probability answer practical questions about the data. All three measures reveal any directional structure that exists in the data and can be used to test for evidence of such structure. Moreover, we have shown that these three new measures are more sensitive for low-differentiated populations than classical measures such as F_{ST}, G'_{ST} and D. Much effort has gone into eliciting the effects of various population characteristics on F_{ST}, G'_{ST}, D, and similar measures, because such characteristics affect the range of values and subsequent interpretation of a particular value obtained from those measures. Because our measures are based on probabilities that can be described explicitly, our analyses do not rely upon such empirical investigations for a correct interpretation. The correct assignment probability and interloper detection probability retain their meaning regardless of external population characteristics, and directly address questions that are relevant to population management.

We have produced an online interface for producing GenePlots, which can be found at

`catchit.stat.auckland.ac.nz/shiny/geneplot/`

In future, we would like to extend this to incorporate the LGP distribution plots and numeric measures, and we intend to produce an R package to give users full control over the GenePlot settings. Given the plethora of population genetics software tools available, it is critical for us to provide new tools that smoothly complement existing ones, and to communicate clearly to users how to interpret the results.

We intend to adapt the GenePlot methods to be able to combine data of different ploidies, for example to be able to combine mitochondrial DNA haplotypes with microsatellite genotypes within a single analysis. We would also like to find a form of the directional measures that can be applied to multiple populations, and a method for visualizing the pairwise measures from multiple populations on a single diagram. We would also seek to incorporate information about the positions of the loci on the genome when analysing the data.

References

The 1000 Genomes Consortium (2015). A global reference for human genetic variation. *Nature* **526,** 68–74. doi:10.1038/nature15393

Alcala, N., Goudet, J., and Vuilleumier, S. (2014). On the transition of genetic differentiation from isolation to panmixia: what we can learn from GST and D. *Theoretical Population Biology* **93**, 75–84.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655–1664.

Alexander, D. H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12(1)**, 246.

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

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.

Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical population biology* **63(3)**, 221–230.

Balloux, F. (2001). EASYPOP (version 1.7): a computer program for population genetics simulations. *Journal of Heredity* **92**, 301–302.

Balloux, F., and Lugon-Moulin, N. (2002). The estimation of population differentiation with microsatellite markers. *Molecular Ecology* **11(2)**, 155–165.

Banks, M. A., and Eichert, W. (2000). WHICHRUN (Version 3.2): A computer program for population assignment of individuals based on multilocus genotype data. *Journal of Heredity* **91**, 87–89.

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.

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., and Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology* **24**, 3299–3315. doi:10.1111/mec.13245

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.

Carroll, E., Patenaude, N., Alexander, A., Steel, D., Harcourt, R., Childerhouse, S., Bannister, J., Constantine, R., and Baker, C. S. (2011). Population structure and individual movement of southern right whales around New Zealand and Australia. *Marine Ecology Progress Series* **432**, 257–268.

Carroll, E. L., Childerhouse, S. J., Christie, M., Lavery, S., Patenaude, N., Alexander, A., Alexander, A., Constantine, R., Steel, D., Boren, L., and Baker, C. S. (2012). Paternity assignment and demographic closure in the New Zealand southern right whale. *Molecular Ecology* **21(16)**, 3960–3973. doi:10.1111/j.1365-294X.2012.05676.x

Carroll, E. L., Childerhouse, S. J., Fewster, R. M., Patenaude, N. J., Steel, D., Dunshea, G., Boren, L., and Baker, C. S. (2013). Accounting for female reproductive cycles in a superpopulation capture–recapture framework. *Ecological Applications* **23(7)**, 1677–1690.

Carroll, E. L., Rayment, W. J., Alexander, A. M., Baker, C. S., Patenaude, N. J., Steel, D., Constantine, R., Cole, R., Boren, L. J., and Childerhouse, S. (2014), Reestablishment of former wintering grounds by New Zealand southern right whales. *Mar Mam Sci* **30**, 206–220. doi:10.1111/mms.12031

Carroll, E. L., Alderman, R., Bannister, J. L., Bérubé, M., Best, P. B., Boren, L., Baker, C. S., Constantine, R., Findlay, K., Harcourt, R., Lemaire, L., Palsbøll, P. J., Patenaude, N. J., Rowntree, V. J., Seger, J., Steel, D., Valenzuela, L. O., Watson, M., and Gaggiotti, O. E. (2018). Incorporating non-equilibrium dynamics into demographic history inferences of a migratory marine species. *Heredity*. doi:10.1038/s41437-018-0077-y

Chao, A., Ma, K. H., Hsieh, T. C., and Chiu, C. H. (2015). Online Program SpadeR (Species-richness Predictition And Diversity Estimation in R). http://chao.stat.nthu.edu.tw/wordpress/software_download/

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution* **15(5)**, 538–543.

Corander, J., Waldmann, P., and Sillanpää, M. J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.

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

Corander, J., Sirén, J., and Arjas, E. (2008). Bayesian spatial modeling of genetic population structure. *Computational Statistics* **23**, 111–129.

Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A., and Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153(4)**, 1989–2000.

Cornuet, J.-M., Ravigné, V., and Estoup, A. (2010). Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics* **11(1)**, 401.

Crawford, N. G. (2010). SMOGD: software for the measurement of genetic diversity. *Molecular Ecology Resources* **10(3)**, 556–557.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R., Lunter, G., Marth, G., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The Variant Call Format and VCFtools. *Bioinformatics* **27(15)**, 2156–2158. doi:10.1093/bioinformatics/btr330

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.

Dawson, K. J., and Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetical Research* **78**, 59–77. doi:10.1017/S001667230100502x

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.

Dussex, N., Robertson, B. C., Salis, A. T., Kalinin, A., Best, H., and Gemmell, N. J. (2016). Low spatial genetic differentiation associated with rapid recolonization in the New Zealand fur seal *Arctocephalus forsteri*. *Journal of Heredity* **107(7)**, 581–592.

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4(2)**, 359–361.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14,** 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances using DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Excoffier, L., and Heckel, G. (2006). Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics* **7(10)**, 745–758.

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10(3),** 564–567.

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

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* **7(4)**, 574–578.

Fewster, R. M. (2017). Some applications of genetics in statistical ecology. *AStA Advances in Statistical Analysis 101(4)*, 349–379.

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.

Gaggiotti, O. E., and Foll, M. (2010). Quantifying population structure using the F-model. *Molecular Ecology Resources* **10(5)**, 821–830.

Gao, H., Williamson, S., and Bustamante, C. D. (2007). A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multi-locus genotype data. *Genetics* **176(3)**, 1635–51.

Gayevskiy, V., Klaere, S., Knight, S., and Goddard, M. R. (2014). ObStruct: a method to objectively analyse factors driving population structure using Bayesian ancestry profiles. *PLoS One* **9(1)**, e85196.

Gerlach, G., Jueterbock, A., Kraemer, P., Deppermann, J., and Harmand, P. (2010). Calculations of population differentiation based on G(ST) and D: forget G(ST) but not all of statistics! *Molecular Ecology* **19**, 3845–3852.

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

Goudet, J. (1995). FSTAT (version 1.2): A computer program to calculate F-statistics. *Journal of Heredity* **86(6)** 485–486.

Goudet, J., Raymond, M., de Meeüs, T., and Rousset, F. (1996). Testing differentiation in diploid populations. *Genetics* **144(4)**, 1933–1940.

Hedrick, P. W. (1999). Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution* **53(2)**, 313–318.

Hedrick, P. W. (2005). A standardized genetic differentiation measure. *Evolution*, **59(8)**, 1633–1638.

Heller, R., and Siegismund, H. R. (2009). Relationship between three measures of genetic differentiation GST, DEST and G'ST: how wrong have we been? *Molecular Ecology* **18(10)**, 2080–2083.

Holsinger, K. E., and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics* **10(9)**, 639–650.

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**, 1322–1332.

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23(14)**, 1801–1806.

Janes, J. K., Miller, J. M., Dupuis, J. R., Malenfant, R. M., Gorrell, J. C., Cullingham, C. I., and Andrew, R. L. (2017). The K = 2 conundrum. *Molecular Ecology* **26**, 3594–3602.

Johnston, S. E., Orell, P., Pritchard, V. L., Kent, M. P., Lien, S., Niemelä, E., Erkinaro, J., and Primmer, C. R. (2014). Genome-wide SNP analysis reveals a genetic basis for sea-age variation in a wild population of Atlantic salmon (*Salmo salar*). *Molecular Ecology* **23**, 3452–3468. doi:10.1111/mec.12832

Jombart, T. (2008). `adegenet`: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405.

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

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11(1)**, 94. doi:10.1186/1471-2156-11-94

Jost, L. (2008). GST and its relatives do not measure differentiation. *Molecular Ecology* **17(18)**, 4015–4026.

Jost, L. (2009). D vs. GST: response to Heller and Siegismund (2009) and Ryman and Leimar (2009). *Molecular Ecology* **18(10)**, 2088–2091.

Kalinowski, S. T. (2011). The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity* **106**, 625–632.

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

Keenan, K., McGinnity, P., Cross, T. F., Crozier, W. W., and Prodöhl, P. A., (2013). diveRsity: An R package for the estimation of population genetics parameters and their associated errors. *Methods in Ecology and Evolution* **4(8)**, 782–788. doi:10.1111/2041-210X.12067

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* **15**, 1179–1191. doi:10.1111/1755-0998.12387

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.

Larson, W. A., Seeb, L. W., Everett, M. V., Waples, R. K., Templin, W. D., and Seeb, J. E. (2014). Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications* **7(3)**, 355–369. doi:10.1111/eva.12128

Lawson, D. J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS genetics* **8(1),** e1002453.

Lawson, D. J., van Dorp, L., and Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications* **9(1)**, 3258.

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.

Meirmans, P. G. (2006). Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* **60(11)**, 2399–2402.

Meirmans, P. G., and Hedrick, P. W. (2011). Assessing population structure: FST and related measures. *Molecular Ecology Resources* **11(1)**, 5–18.

Meirmans, P. G., and Van Tienderen, P. H. (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* **4**, 792–794. doi:10.1111/j.1471-8286.2004.00770.x

McMillan, L., and Fewster, R. (2017). Visualizations for genetic assignment analyses using the saddlepoint approximation method. *Biometrics* **73**, 1029–1041. doi:10.1111/biom.12667

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences* **70(12)**, 3321–3323.

Nei, M. (1977). F-statistics and analysis of gene diversity in subdivided populations. *Annals of human genetics* **41(2)**, 225–233.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89(3)**, 583–590.

Nei, M. (1987). Molecular evolutionary genetics. New York, USA: Columbia University Press. ISBN 0231063210, 9780231063210.

Nei, M., and Chesser, R. K. (1983). Estimation of fixation indices and gene diversities. *Annals of human genetics* **47(3)**, 253–259.

Nei, M., and Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics* **76(2)**, 379–390.

Paetkau, D., Calvert, W., Stirling, I., and Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* **4(3)**, 347–354.

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.

Peakall, R., and Smouse, P. E. (2001). GenAlEx V5: Genetic Analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288–295.

Peakall, R., and Smouse, P.E. (2006). GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288–295.

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.

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.

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

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

Raymond, M., and Rousset, F. (1995). GENEPOP (version 1.2): Population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86(3)**, 248–249.

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.

Rosenberg, N. A. (2004). DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes* **4(1)**, 137–138.

Rosenberg, N. A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J. K., Feldman, M. W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *PLOS Genetics*, **1**, e70.

Rousset, F. (2008). GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8(1)**, 103–106.

Ryman, N., and Leimar, O. (2009). GST is still a useful measure of genetic differentiation — a comment on Jost's D. *Molecular Ecology* **18(10)**, 2084–2087.

Schaid, D. (2004). Linkage disequilibrium testing when linkage phase is unknown. *Genetics* **166**, 505–512.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139(1)**, 457–462.

Swofford, D. L., and Selander, R. B. (1981). BIOSYS-1: a FORTRAN program for the comprehensive analysis of electrophorectic data in population genetics and systematics. *Journal of Heredity* **72**, 281–283.

Taylor, E. B., Boughman, J. W., Groenenboom, M., 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.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., et al. (2009). The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044.

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.

Weir, B. S. (1990). Genetic data analysis: methods for discrete population genetic data. Sinauer Associates. ISBN 0878938729.

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38(6)**, 1358–1370.

Weir, B. S., and Hill, W. G. (2002). Estimating F-statistics. *Annual Review of Genetics* **36(1)**, 721–750.

Whitlock, M. C. (2011). G'ST and D do not replace FST. *Molecular Ecology* **20(6)**, 1083–1091.

Wilson, G. A., and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163(3)**, 1177–1191.

Winter, D. J. (2012). MMOD: an R library for the calculation of population differentiation statistics. *Molecular Ecology Resources* **12(6)**, 1158–1160.

Wright, S. (1951). The Genetical Structure of Populations. *Annals of Eugenics* **15,** 322–354.

Zaykin, D. (2004). Bounds and normalization of the composite linkage disequilibrium coefficient. *Genetic Epidemiology* **27**, 252–257.