SCIENCE
DEPARTMENT OF STATISTICS

Visualizing Population Genetics

Supervisors: Rachel Fewster, James Russell, Paul Murrell

Louise McMillan

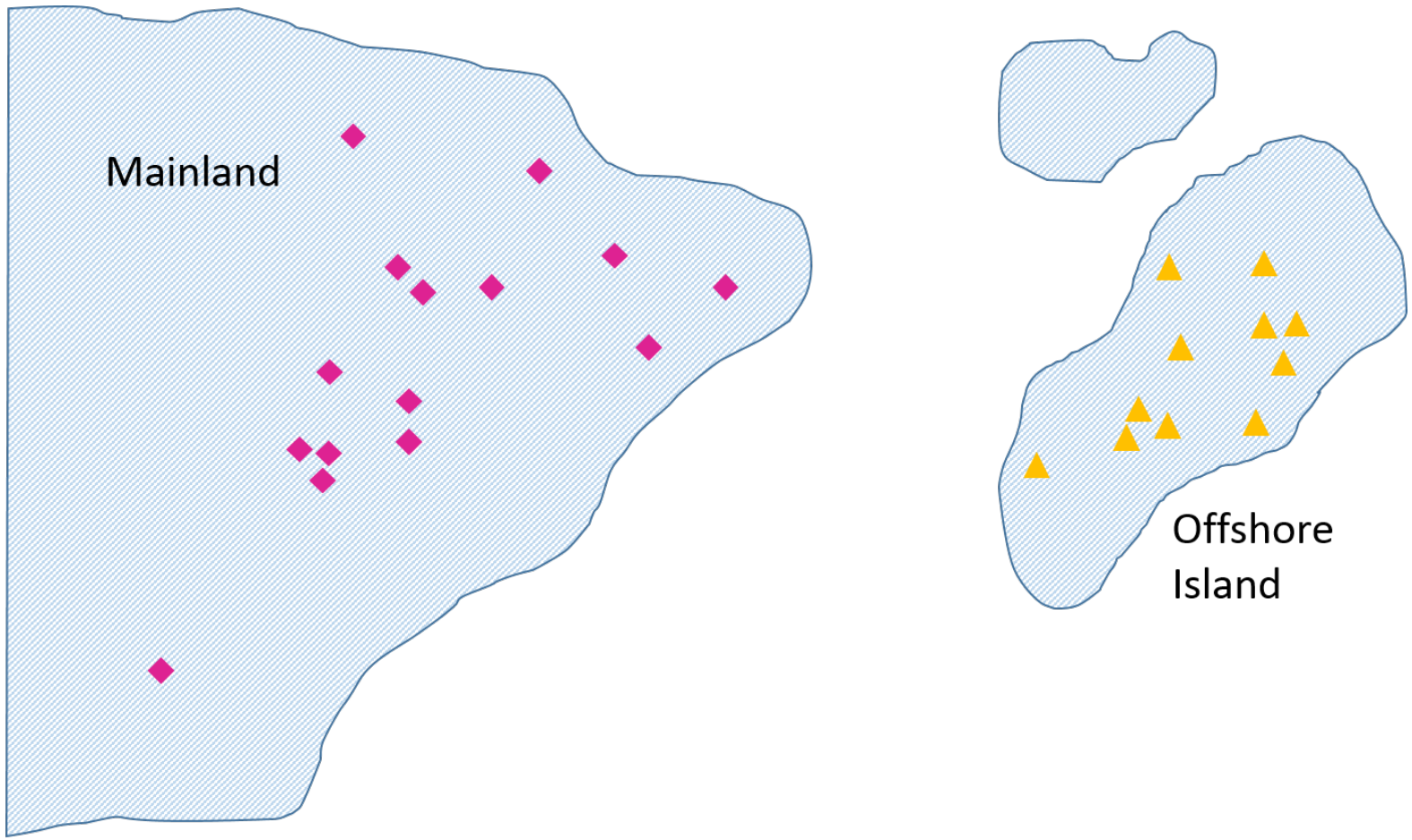
1 December 2015

Outline

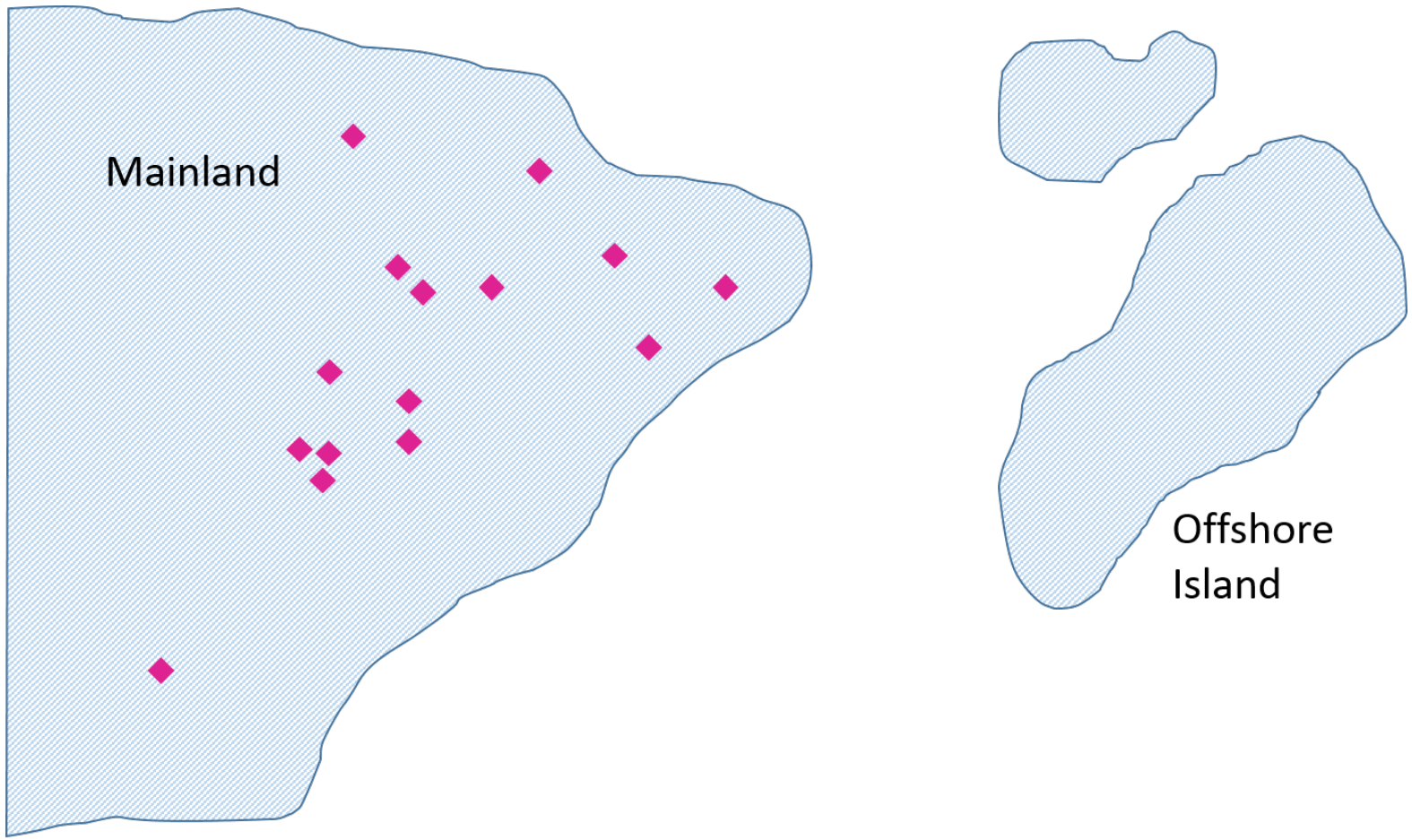
Assignment

- Uses genetic data (microsatellites or Single Nucleotide Polymorphisms (SNPs))
- Compare individuals to populations
- Infer likely source populations

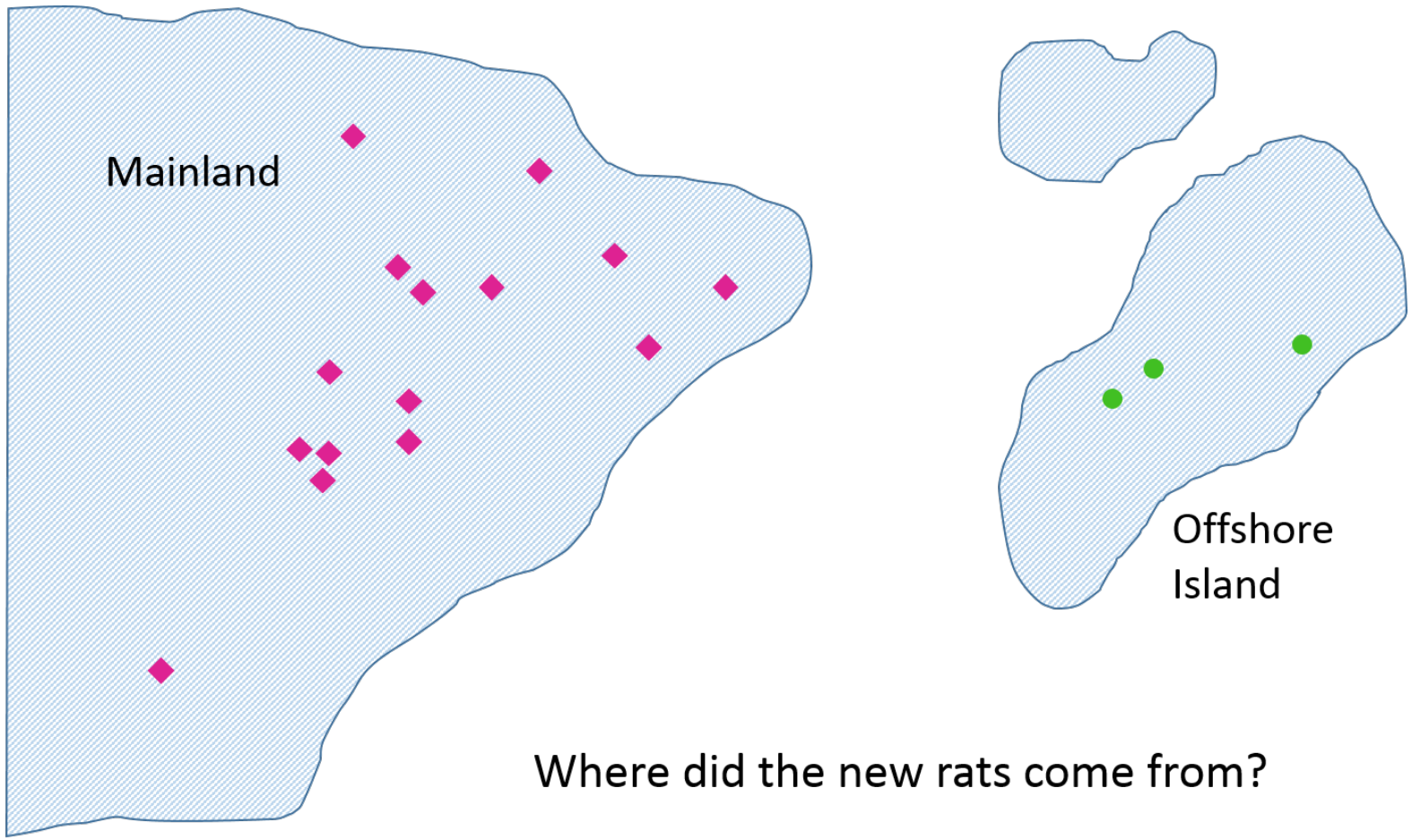
Post-eradication Assignment



Post-eradication Assignment



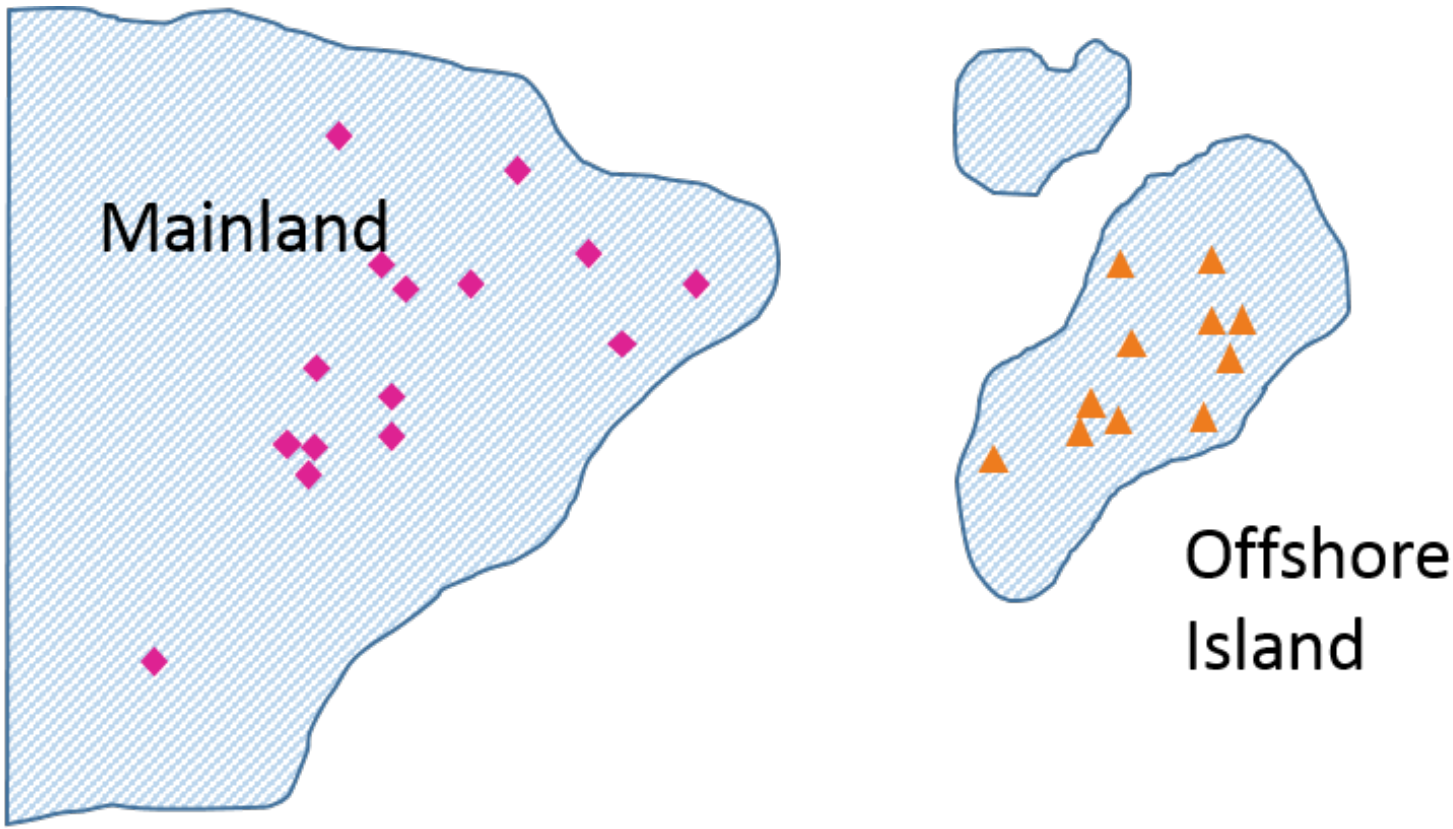
Post-eradication Assignment



Where did the new rats come from?

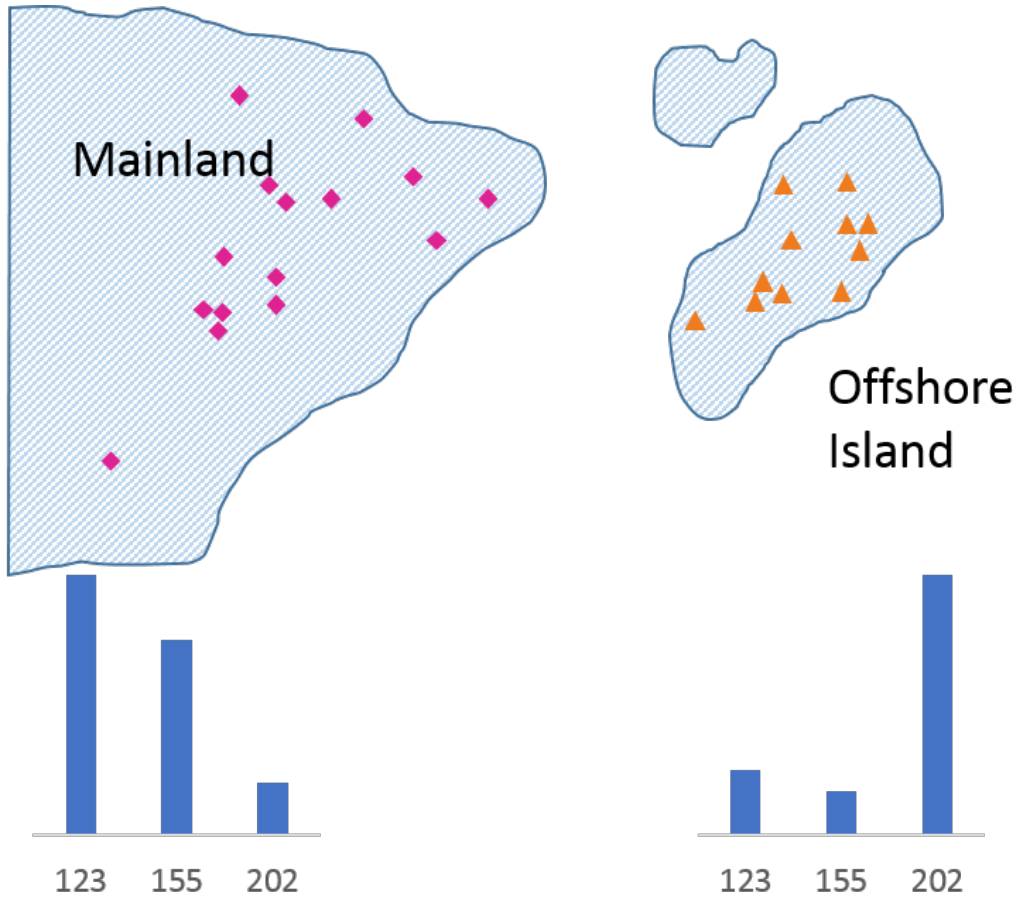
Assignment Process

- Baseline samples from each possible source population



Assignment Process

- Estimate microsatellite allele frequencies of candidate source populations using baseline sample



Assignment Process

- Dirichlet prior for allele frequencies

$$\boldsymbol{p} \sim \text{Dirichlet}(\tau, \tau, \ldots, \tau)$$

- Combine Dirichlet prior and multinomial data to get Dirichlet posterior for baseline allele frequencies

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

Assignment Process

- Compare alleles from new sample with allele frequencies for candidate source population
- Calculate Log Genotype Probability (LGP) using Dirichlet Compound Multinomial (DCM) distribution
- Probability of obtaining individual's genotype from multinomial distribution with estimated allele proportions as probabilities

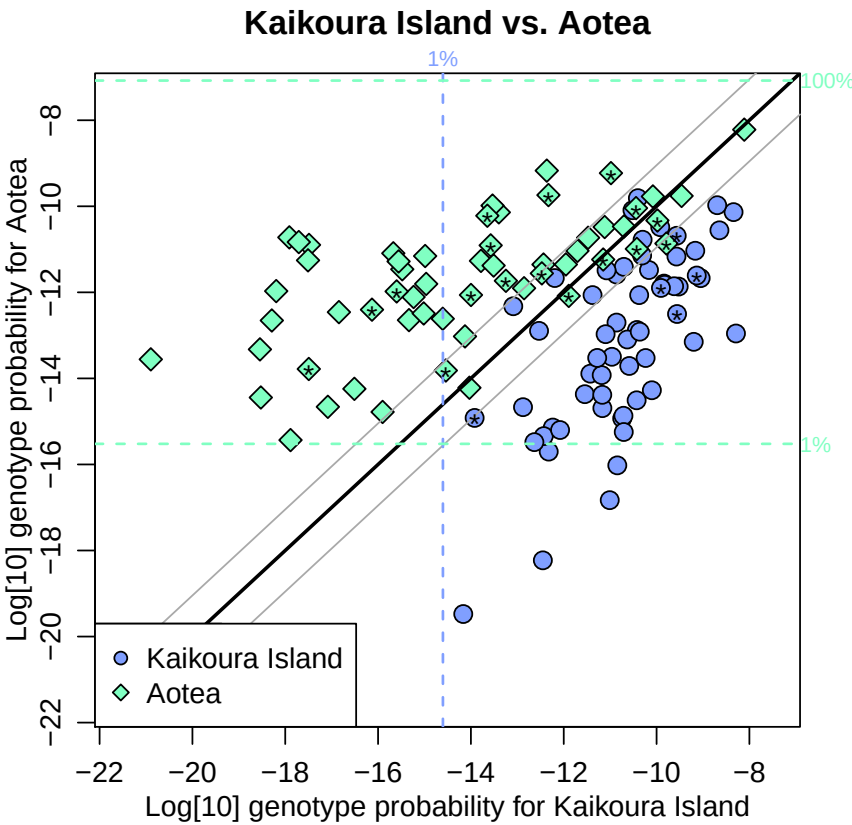
$$\mathbb{P}(\boldsymbol{a}) = \begin{cases} \dfrac{(x_r + \tau)(x_r + \tau + 1)}{(n+1)(n+2)} & a_r = 2,\ a_j = 0 \text{ for } j \neq r \\ \dfrac{2(x_r + \tau)(x_s + \tau)}{(n+1)(n+2)} & a_r = a_s = 1,\ a_j = 0 \text{ for } j \neq r, s \end{cases}$$

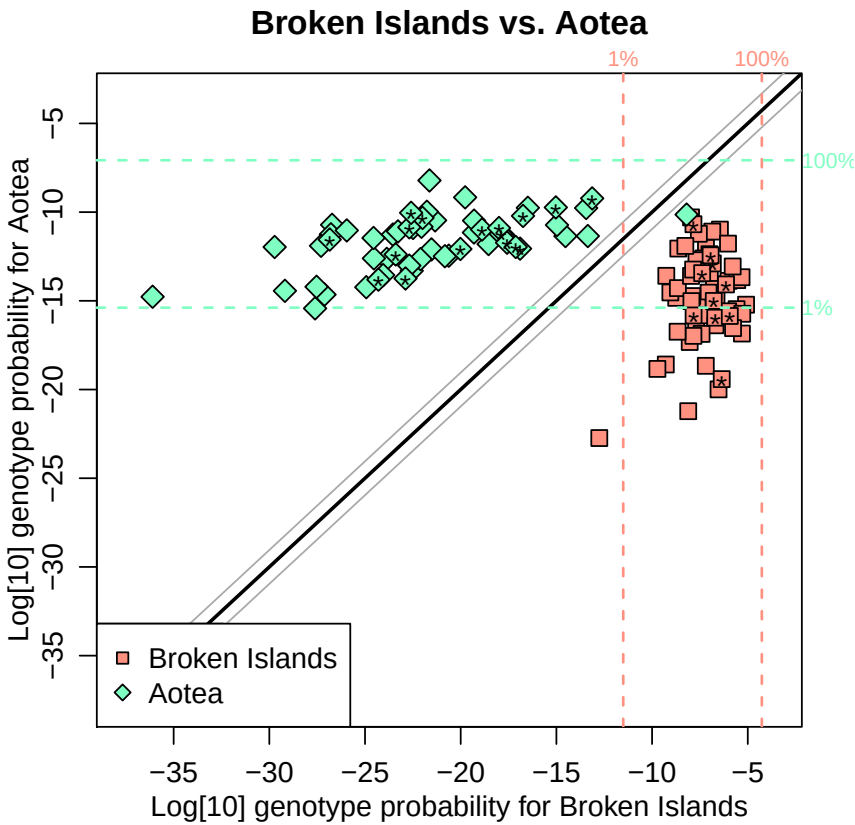
Assignment Process

- Combine over all loci to get overall Log Genotype Probability (LGP)
- Calculate for each candidate source population
- Assign individual to population with highest LGP
- OR only assign if individual has assignment score above threshold for at least one population

Assignment Process

- Starred points indicate individuals with missing data





Alternative Assignment Methods

- GENECLASS2 (Piry et al. 2004) and GenePlot use the above method
- STRUCTURE (Pritchard et al. 2000) uses same Dirichlet posterior for allele frequencies and same LGP calculations
- But STRUCTURE performs clustering, using MCMC to sample from distribution of population membership combinations

GenePlot Method for Missing Data

- Not all DNA samples replicate correctly
- Individual may have missing data at one or more loci
- Calculate LGP for non-missing loci
- But LGP will be on different scale than LGP for individuals with all loci

Characterising Genetic Distribution of Population

- Calculate LGP for non-missing loci
- Find corresponding quantile in full-loci distribution
- Requires characterisation of full-loci and reduced-loci distributions for all candidate source populations
- Distribution is sum of distributions from each locus, difficult to characterise

Characterising Genetic Distribution of Population

- Possible to simulate individuals from each population using allele frequencies
- Distribution of genotype probabilities can be estimated from simulated data
- Slow to calculate, non-repeatable results
- Unreliable in long lower tail of distribution

Characterising Genetic Distribution of Population

- Alternative method is to approximate population genetic distribution analytically
- Saddlepoint approximation method achieves this with high accuracy

Saddlepoint Approximation

- Initial motivation for the saddlepoint approximation:
- Set of random variables $X_1, X_2, \ldots, X_n$ with known distributions
- What is the distribution of their sum, X_{tot}?

Saddlepoint Approximation

- PDF approximation derived by Daniels in 1954
- CDF approximation derived by Lugannani & Rice in 1980
- Can approximate any distribution, not just sums
- Approximations rely on the cumulant generating function (CGF)
- Overall CGF is sum of component CGFs

Multilocus Genetic Distribution

- Distribution of genotype probabilities in a population
- MGF at a single locus is given by:

$$\begin{aligned} M(t) &= \sum P(Y = y)e^{ty} \\ &= \sum p_i e^{t \log p_i} \\ &= \sum p_i^{t+1} \end{aligned}$$

- p_i are the probabilities of the possible genotypes at that locus.

Multilocus Genetic Distribution

- Derivatives of the MGF are given by:

$$M^{(r)}(t) = \sum (\log p_i)^r p_i^{t+1}$$

- Overall CGF and derivatives are sums of those at each locus

Multilocus Genetic Distribution

- Saddlepoint CDF formula:

$$\hat{F}(x) = \left\{ \begin{array}{ll} \Phi(\hat{w}) + \phi(\hat{w})(1/\hat{w} - 1/\hat{u}) & \text{if } x \neq \mu \\ \frac{1}{2} + \frac{K'''(0)}{6\sqrt{2\pi}K''(0)^{3/2}} & \text{if } x = \mu \end{array} \right\}$$

- where

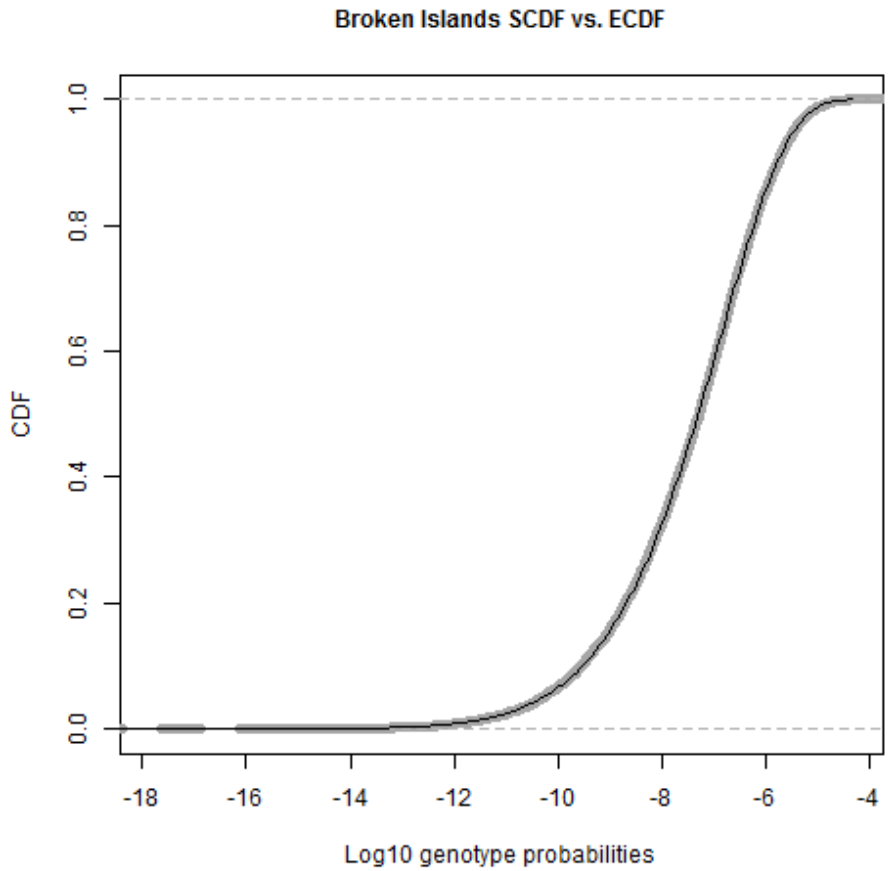
$$K'(\hat{s}) = x$$

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

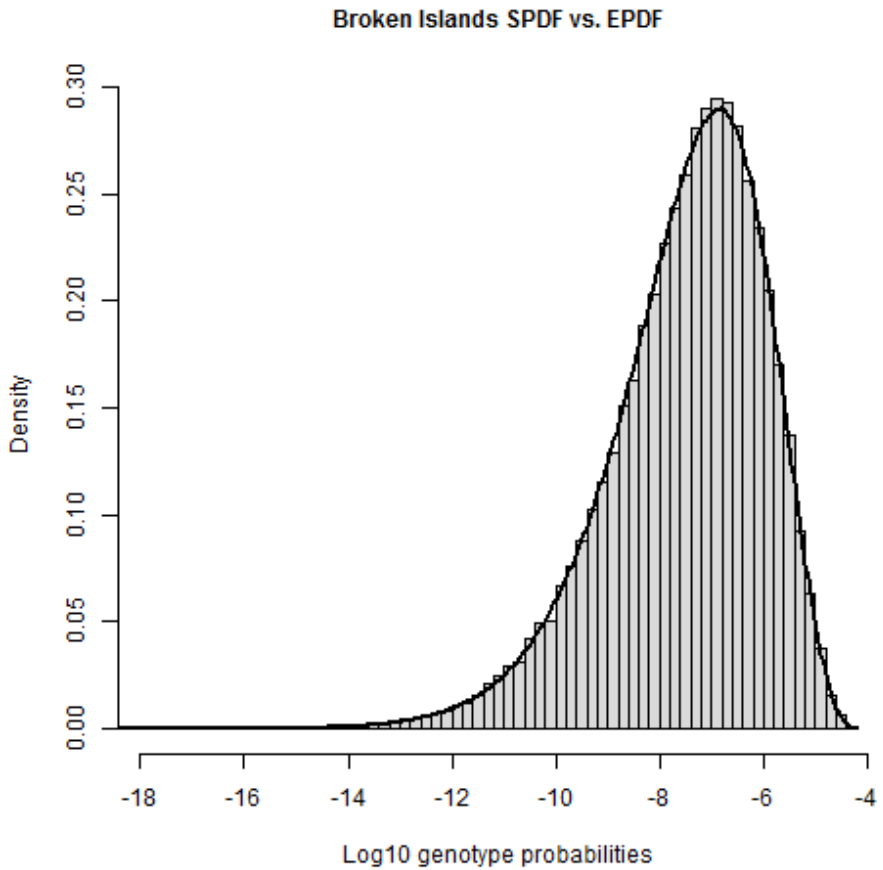
$$\hat{u} = \hat{s}\sqrt{K''(\hat{s})}$$

Test against Generated Samples

- Generate 100,000 samples from multilocus distribution
- Calculate empirical CDF
- Compare with saddlepoint approximation

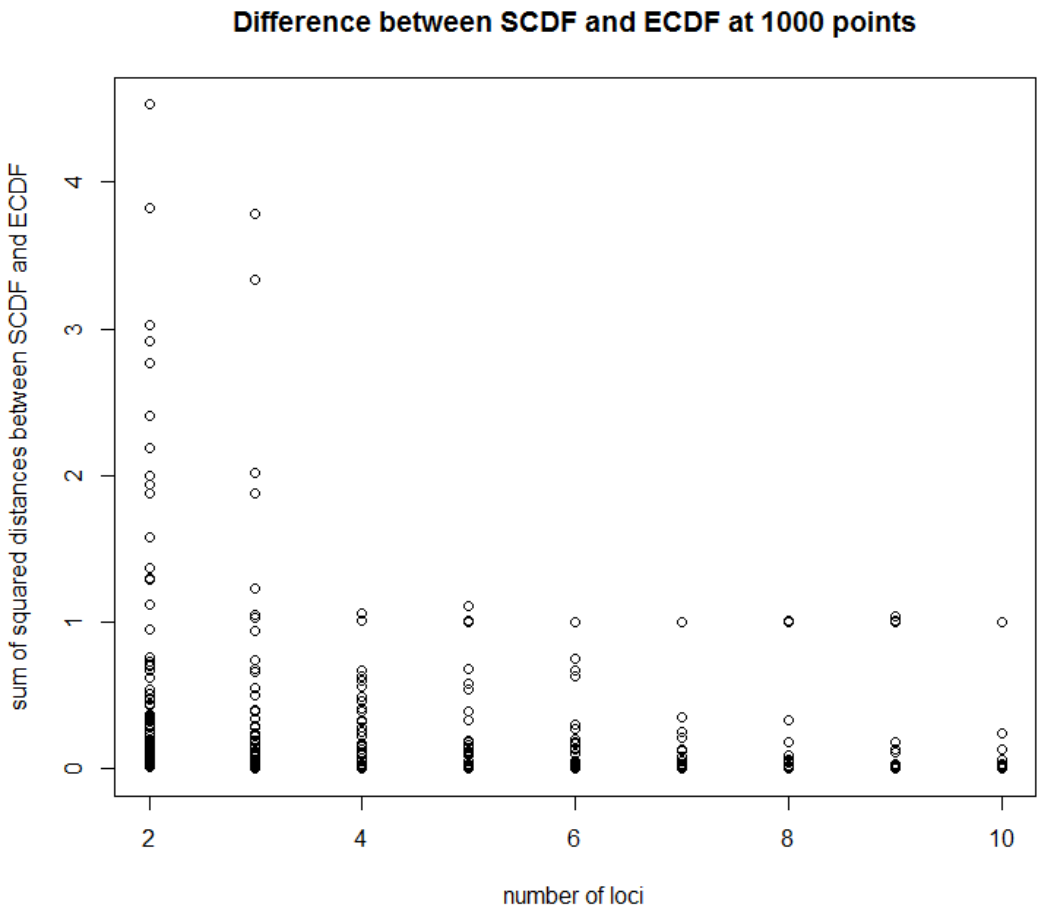


Test against Generated Samples



Test for Simulated Populations

- Calculate sum of squared differences over 1000 points for different numbers of loci

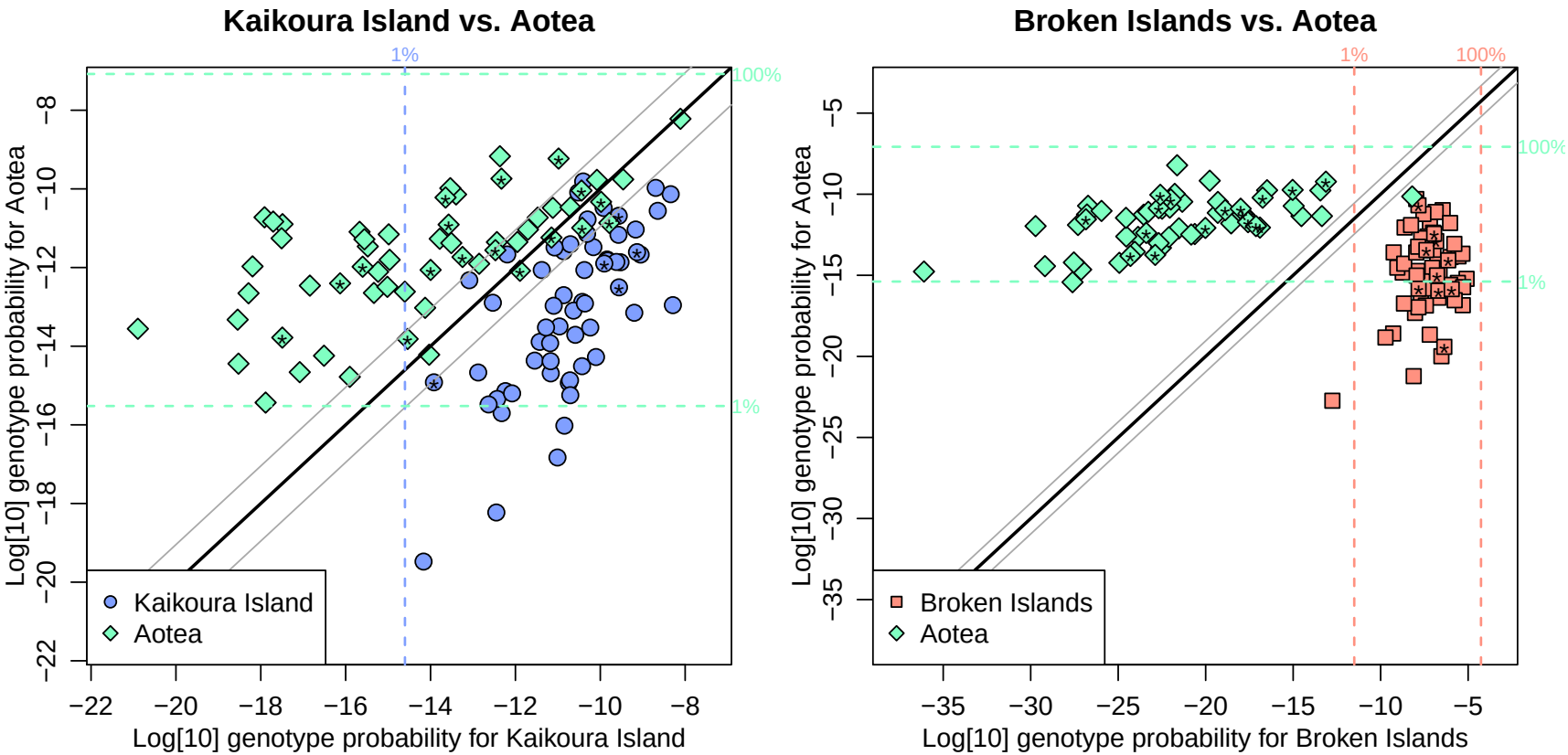


Characterising Genetic Distribution of Population

- Calculate LGP for non-missing loci
- Find corresponding quantile in full-loci distribution
- Requires characterisation of full-loci and reduced-loci distributions for all candidate source populations
- Distribution is sum of distributions from each locus, difficult to characterise

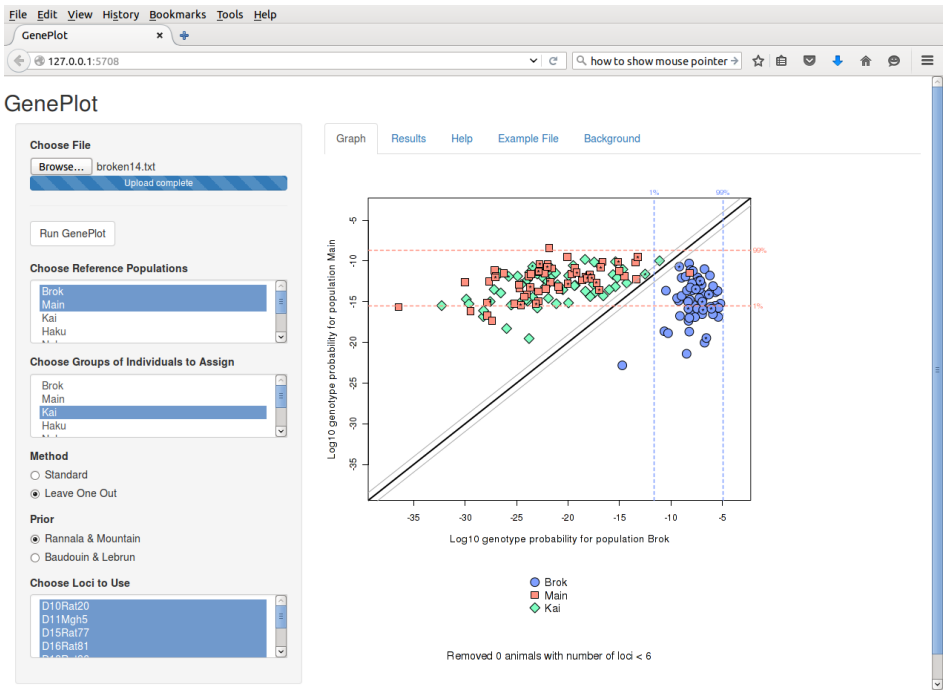
GenePlots

- LGP results for each individual, with respect to all source populations, can be plotted on a graph



Further Work

- Simulate scenarios to test for bias in assignment results
- Investigate relatedness within populations and effects on assignment accuracy
- Online version of GenePlot:



https://lmcm177.shinyapps.io/geneplot-on-the-web

References

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

S. Piry et al (2004) *GENECLASS2: A Software for Genetic Assignment and First-Generation Migrant Detection*.
Journal of Heredity, Vol. 95 (Issue 6).

J. K. Pritchard, M. Stephens and P. Donnelly (2000) *Inference of Population Structure Using Multilocus Genotype Data*.
Genetics, Vol. 155.

H. E. Daniels (1954) *Saddlepoint Approximations in Statistics*.
The Annals of Mathematical Statistics, Vol. 25.

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

C. Goutis and G. Casella (1999)
Explaining the Saddlepoint Approximation. The American Statistician, Vol. 53.

Acknowledgements

Many thanks to my supervisors, particularly
Rachel Fewster