# dartR support for Population Assignment

Assigning individuals of unknown provenance to populations of known provenance is a challenging exercise, and several approaches have been suggested. Perhaps the simplest is to calculate the probabilities of yielding the observed genotype of the unknown individual given the observed allele frequencies in each the target population. Using this approach, the individual is assigned notionally to those populations for which this probability or likelihood is greatest; populations for which the probability or likelihood is lower than some level of significance are eliminated from further consideration. This approach was first applied in a study of microsatellite markers in bear populations (Paetkau, Slade, Burden, & Estoup, 2004) and subsequently applied using classical and Bayesian approaches to estimating probabilities (Blanchong, Scribner, & Winterstein, 2002; Gotz & Thaller, 1998).

For various reasons, including not wishing to duplicate software options that are already available, we have taken an alternate approach. We first eliminate from consideration those target populations where a SNP allele is present in the unknown individual but not in the target – if there are alleles in the unknown individual not present in a particular population, then it is unlikely that the focal individual was drawn from that population.

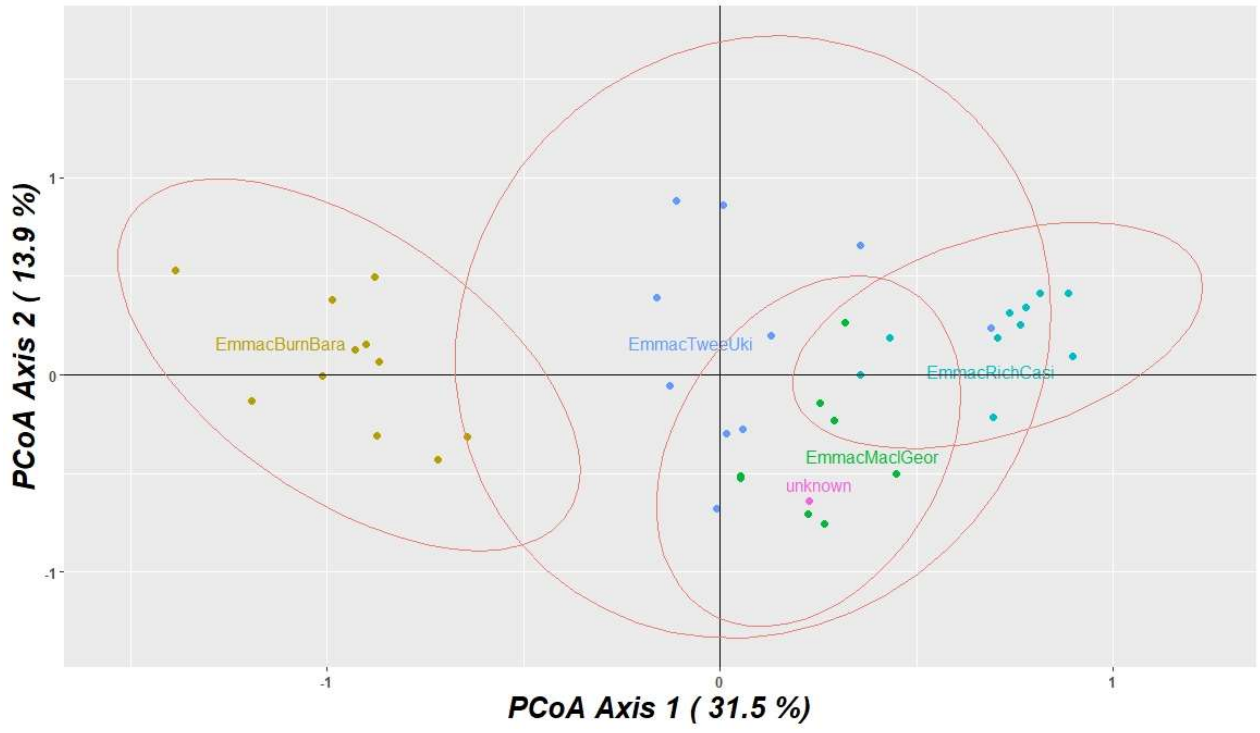
```
x <- gl.assign(testset.gl, id="UC_00146", nmin=10, alpha=0.05, t=1)

   0 EmmacMaclGeor
   1 EmmacBurnBara EmmacRichCasi EmmacTweeUki
   2 EmmacBrisWive EmmacMDBBowm EmmacMDBCudg EmmacMDBForb EmmacMDBMaci EmmacMDBMurrMung
   3 EmmacFitzAllig EmmacMDBCond EmmacMDBSanf EmmacRussEube
   4 EmmacBurdMist EmmacJohnWari EmmacRoss
   6 EmmacCoopEulb
   7 EmmacCoopCully
   16 EmmacCoopAvin
```

Note that the focal unknown individual shares all of its alleles only with EmmacMacGeor, which is the population from which it was drawn for this example.

In many cases, examining private alleles in this way will narrow down the possible source populations considerably, and depending on the spatial resolution required for the assignment (say, Australia or New Guinea), may provide a satisfactory answer.

Second, we examine the position of the unknown individual relative to the target populations in a reduced ordinated locus space using PCoA. This graphic representation is provided by gl.assign().



Addition of confidence ellipses then allows a decision to eliminate some populations from consideration as the source of the unknown individual (shown in pink).

The converse is not true. This approach does not allow assignment of the unknown to populations that contain the unknown within their confidence ellipse. The overall confidence envelope is multidimensional, and separation of the unknown from a target population may occur in deeper dimensions. Hence, as with the private alleles approach, this graphical approach serves to narrow down the candidates for the source of the unknown, and may in that sense, provide a satisfactory answer.

Third, we deal with non-independence (linkage) among the SNP loci by ordinating the space defined by those loci. The resultant axes, linear combinations of the information contained in each locus, are orthogonal and so can be regarded as independent. Subsequent standardization can achieve independent and identically distributed variates, which simplifies analysis of probabilities and likelihoods.

```
      Population   Index      CE Assign
2 EmmacMaclGeor -1.1294 -2.6193    yes
4  EmmacTweeUki -1.2056 -2.6193    yes
3 EmmacRichCasi -3.9405 -2.6193     no
1 EmmacBurnBara -7.8446 -2.6193     no
```

Index is a weighted log-likelihood score, the bigger it is, the more likely does the unknown belong. The CE is the log-likelihood of an individual residing on the confidence envelope. If Index > CE, then membership of the unknown to the respective population is unlikely.

A critical issue is whether the number of individuals in the target populations are sufficient to characterize them in the critical considerations made here. Is the sample size sufficient to support the identification of a private allele in the unknown? Is it sufficient to confidently construct confidence ellipses in the PCoA plot? Is it sufficient to provide a robust estimate of the distribution of individuals along each axis of the ordination in order to adequately estimate the likelihood of the unknown on that axis? We have set a default of nmin=10, but this is a matter of judgement that needs to be considered when planning a study.

## References


Blanchong, J. A., Scribner, K. T., & Winterstein, S. R. (2002). Assignment of individuals to populations: Bayesian methods and multi-Locus genotypes. *Journal of Wildlife Management*, *66*, 321–329.

Gotz, K. U., & Thaller, G. (1998). Assignment of individuals to populations using microsatellites. *Journal of Animal Breeding and Genetics*, *115*, 53–61.

Paetkau, D., Slade, R., Burden, M., & 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.