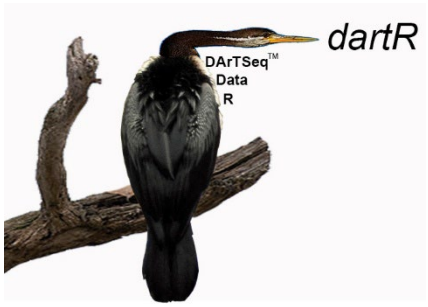
Wildlife Genomics@UC

UNIVERSITY OF CANBERRA

SNP Analysis using dartRverse



Population Assignment using dartRverse

Version 1

Copies of the latest version of this tutorial are available from:

The Institute for Applied Ecology
University of Canberra ACT 2601
Australia

Email: arthur.georges@biomatix.com.au

Citation: Georges, A., Gruber, B., and Mijangos, J.L. (2025). Population Assignment. Version 1. Technical Note 3, Biomatix Pty Ltd, Sutton, Australia..

dartRverse is a collaboration between the University of Canberra, CSIRO and Diversity Arrays Technology, and is supported with funding from the ACT Priority Investment Program, CSIRO and the University of Canberra.

Contents

Introduction

Population assignment refers to the suite of analyses used to decide which population, stock, lineage, or genetic cluster an individual most likely belongs to on the basis of its genotype.

You may already have reference groups, such as animals sampled from different rivers, islands, or management units. Allele-frequency differences across SNPs are then used to assess, probabilistically, which reference population best matches an unknown individual.

Alternatively, reference populations may not be defined in advance. Instead, methods such as Principal Components Analysis (PCA) may identify genetic groupings from the SNP data itself. Individuals can then be assigned to those clusters on a probabilistic basis in subsequent analyses.

Assignment analysis can be particularly useful for identifying migrant individuals.

Population assignment needs to be distinguished from analyses designed to identify a source location in a geographical context. Population assignment asks, “Which genetic group does this individual belong to?” As such, it requires genetic differentiation among groups, whether predefined or identified post hoc.

Location assignment asks, “From which geographic place did this individual most likely come?” It requires finer-scale spatial structure, denser sampling, and usually many more markers than population assignment.

We do not cover the topic of location assignment in this tutorial.

What you will Learn

- Understanding of the fundamentals of population assignment, including the different analytical approaches to the analysis.
- Practical experience in exploring SNP data in the context of population assignment – playing in the Sand Pit.
- Various options for formally undertaking population assignment using the latest available tools.
- A sound appreciation of software options for population assignment and the nuances of analysis.

Who is this Tutorial for?

This tutorial is for users of the R package dartRverse designed for genetic analysis of single nucleotide polymorphisms (SNPs) and associated SilicoDArT data.

Session 1: Basic Population Assignment

Overview

Background

Early population assignment methods in genetics were developed in the 1990s with the advent of highly polymorphic markers like microsatellites. One of the first individual-based assignment methods was introduced by Paetkau *et al.* (1995) who used multilocus microsatellite genotypes to successfully assign Canadian polar bears to their region of origin. Theirs was a likelihood-based approach used to compute the probability of an individual's genotype in each candidate population using observed allele frequencies, assigning the individual to the population where this likelihood was highest. Rannala and Mountain (1997) developed a Bayesian refinement that treated allele frequencies as random variables with prior distributions (e.g. Dirichlet priors) to account for sampling uncertainty. Their method could identify first-generation migrants by flagging individuals whose genotypes were much more probable in a population other than the one where they were sampled (Rannala & Mountain, 1997). These studies laid the groundwork for modern assignment techniques.

Some early assignment approaches used *distance-based* metrics (Degen *et al.* 2017). For instance, one can compute a genetic distance (e.g. Nei's distance or Euclidean distance in allele frequency space) between an individual's genotype and each population and assign to the nearest population. However**,** likelihood-based methods, especially Bayesian approaches, consistently outperformed distance-based methods across various simulation scenarios (Cornuet *et al.* 1999).

Assignment tests gained popularity in population genetics for addressing diverse questions from measuring population connectivity to detecting dispersal and migration events (Manel *et al.*, 2005; Wilson & Rannala, 2003). The development of model-based clustering algorithms like STRUCTURE (Pritchard *et al.*, 2000), while primarily designed to infer population structure, could also probabilistically assign individuals to inferred populations. Note however that STRUCTURE has been shown to provide erroneous results if population sample sizes are unbalanced (Puechmaille, 2016; Wang, 2016).

Manel *et al.* (2005) emphasized choosing appropriate assignment techniques for different biological questions.

Allele Frequency Methods

Allele frequency assignment methods rely directly on the observed allele frequencies in putative source populations to assign individuals. The classic implementation is to calculate the likelihood of an individual's multilocus genotype arising from each candidate population, assuming Hardy–Weinberg equilibrium and linkage equilibrium within populations (Paetkau *et al.*, 1995; Cornuet *et al.*, 1999). The individual is then assigned to the population with the highest likelihood (or sometimes highest posterior probability if assuming equal priors for populations). A small fudge factor is required to handle alleles present in the individual to be assigned but not in the putative source population.

In salmon management, extensive SNP datasets have enabled assignment of individual fish to river of origin with high confidence using frequency-based

methods, assisting the regulation of harvests (Beacham *et al.*, 2018). In conservation, frequency-based assignment has helped identify source populations of confiscated or captive animals by comparing genotypes to reference databases (Ogden & Linacre, 2015). These methods are relatively easy to interpret and implement, but they assume reference populations are well characterized. Accuracy diminishes if sampling of putative source populations is incomplete or if once cannot be confident that the focal individual comes from one of the sampled putative source populations.

Bayesian Assignment Approaches

Bayesian approaches to population assignment extend the likelihood framework by incorporating prior information and treating the allele frequencies of the unknown focal individual as random variables. In the Bayesian assignment test, allele frequencies in each population are assumed to follow a Dirichlet distribution (a conjugate prior to the multinomial sampling of alleles) (Rannala and Mountain, 1997). This approach adds a prior pseudo-count to each allele, overcoming the issue with the frequency-based methods outlined above, and improving assignment accuracy when sample sizes are small. The assignment of an individual is based on the posterior probabilities that it originates from each putative source population, given its genotype. An individual can be assigned to the population with the highest posterior probability, or considered a migrant if none of the posterior probabilities (including for the population where it was found) are high enough. Empirical and simulation studies showed that this Bayesian method tends to outperform the simple frequency method, particularly in challenging scenarios of low differentiation (Cornuet *et al.*, 1999).

Bayesian approaches remain a cornerstone of assignment testing, particularly when integrating additional uncertainty or needing probability-based interpretations of assignment (e.g. “assignment probability” or “posterior assignment odds” for each individual).

Machine Learning Approaches

Machine learning using supervised classifiers such as support vector machines (SVM), random forests, and naïve Bayes classifiers has been applied to SNP genotype data for population discrimination. These methods can handle high-dimensional input (thousands of SNP features) and often include built-in regularization or feature selection that is useful for avoiding overfitting. For example, a random forest classifier can rank SNPs by importance, helping to identify a subset of informative markers (Sylvester *et al.*, 2016). The R package assignPOP (Chen *et al.*, 2018) provides a machine-learning framework for population assignment. It allows the user to train and evaluate multiple classifiers (LDA, SVM, decision tree, random forest, etc.) with k-fold cross-validation into the assignment model (Chen *et al.*, 2018). The emphasis on rigorous cross-validation in {assignPOP} addresses a critical point for ML methods, that is, to ensure that the model’s accuracy is assessed on independent data to prevent overfitting, especially when the number of SNP predictors is very large relative to sample size.

A recent innovative approach (KLFDAPC, Kernel Local Fisher Discriminant Analysis of Principal Components) combines kernel methods with neural networks to improve assignment of individuals to geographic origin (Qin *et al.*, 2022). KLFDAPC first uses a kernel-based extension of adegenet’s DAPC (Jombart et al. 2010) to capture nonlinear genetic structure, then trains a neural network to predict the

latitude/longitude of origin for each individual (Qin *et al.*, 2022). This method significantly improved the accuracy of geographic origin prediction in human genomic datasets compared to standard PCA or DAPC, highlighting how machine learning can integrate spatial prediction with genetics. While such complex models are still emerging, they illustrate the potential for machine-learning approaches to capture subtle patterns in SNP data (e.g. signals of isolation by distance or admixture) that might be missed by simpler methods.

Overall, machine learning and multivariate methods offer powerful alternatives and complements to traditional allele frequency approaches. They can be especially useful when dealing with thousands of correlated SNPs, where dimension reduction and classification algorithms can outperform likelihood-based models that struggle with high dimensionality. Unlike classical methods, ML models may not provide clear biological interpretation (e.g., they won't directly give allele frequency-based probabilities), so their use is often guided by practical accuracy considerations rather than theoretical population genetics.

Recommended R Software: assignPOP (Chen *et al.*, 2018); KLFDAPC1.r (Qin *et al.*, 2022)

Using dartRverse for Assignment

In the spirit of dartRverse, we do not attempt to duplicate the innovations of others in the space of population assignment, focusing instead on smoothing the path between the dartR genlight object and other publicly available packages.

The focus of dartRverse scripts is on exploratory analysis. We present four basic approaches.

- **Genotype Likelihood:** The likelihood of drawing the unknown from a population with the observed allele frequencies is calculated assuming Hardy-Weinberg equilibrium.
- **Private Alleles:** A focal unknown individual is likely to have fewer private alleles in comparison with its source population than in comparison with other putative source populations.
- **PCA:** The genotype of a focal unknown individual is likely to lie within the confidence envelope of its source population than within the confidence envelope of other putative source populations.
- **Mahalanobis Distance:** The distances of the focal unknown individual from the centroids of the standardized confidence envelops of its putative source populations are used to calculate a z-scores and associated probabilities of assignment.

These approaches are more of value in eliminating putative source populations from consideration than in making a definitive assignment to a source population based on rigorous statistical assessment. A large p-value is not positive support for source population; it is merely a failure to exclude. This is an important caveat.

These approaches can be used individually or to progressively eliminate unlikely source populations in sequence.

The assignment scripts in dartRverse take as input a dartR genlight object with multiple populations that are taken to be the putative source populations for an unknown focal individual. The genotype of the focal individual is included in a population called `unknown`.

The approaches are sound only if the sample sizes for the putative source populations are relatively large, and a warning is issued if the user specifies a minimum sample size `nmin` of less than 10. Populations with sample sizes less than `nmin` are eliminated from the analyses.

Genotype Likelihood

The script `gl.assign.on.genotype()` calculates the likelihood of drawing the observed genotype of the unknown individual from each putative source population on the assumption that the population is in Hardy-Weinberg equilibrium.

A list of populations, the likelihoods, and AIC value and AIC weights are output to the screen. The population with the highest AIC weight is chosen as the source population. This decision carries the risk that the actual source population may not be among those sampled.

The best `n.best` populations are retained if `n.best` is specified otherwise only those assignments that have AIC weights greater than a user specified threshold are retained.

Private Alleles

The script `gl.assign.pa()` calculates the distribution of counts of private alleles for each individual in a putative source population against the remaining individuals in that population. This information is used to generate an expectation for the private allele count for the unknown focal individual. Comparing the unknown focal individual with the expectation yields a z score and p value (under negative binomial assumptions) that can be used for a decision on assignment.

The best `n.best` populations are retained if `n.best` is specified otherwise only those assignments that have p-values less than a specified `alpha` value are retained. The best putative source population (the one with the largest p-value) may be chosen in support of a decision.

PCA

The PCA approach implemented in dartRverse as `gl.assign.pca()` is used as a first cut to eliminate putative source populations from consideration. This might be done for example to reduce computational load when applying other approaches.

A classical PCA analysis is applied to the data to generate confidence ellipses with a user specified level of alpha (typically small to avoid over-exclusion of putative source populations). This is done in only the top two dimensions, justified because if a focal unknown individual lies outside the confidence ellipse of a putative source population in 2D, then examination of deeper dimensions are unlikely to draw it into the confidence envelope.

Only putative source populations for which the focal unknown individual falls within their confidence ellipses are retained in the dartR genlight object passed

back by the function, along with the focal individual in a population called `unknown`.

Mahalanobis Distance

The script `gl.assign.mahalanobis()` first undertakes a classical PCA and retains only those dimensions that are considered to contain structural information. The "noise" dimensions are discarded. The distinction between the informative and noise dimensions is made using the broken-stick criterion.

Standardized confidence envelopes at a user-specified level of alpha are generated in the reduced ordination space for each putative source population. These are classical PCA confidence envelopes but the extent of them along each axis is measured in standard deviations. The result is a series of confidence "spheres".

The distance of the focal unknown individual from the centroids of the standardized confidence envelope is calculated for each putative population. Note that this is a z-score that can be used to generate a p-value for assignment. This p-value is compared to the user specified alpha value as the basis of a decision.

As a final refinement, it is noted that each individual genotype in a diploid organism carries more information than is represented by its position in the PCA. To accommodate this, the script will optionally (the default) generate 100 individuals under Hardy-Weinberg equilibrium assumptions for the generation of the confidence ellipses.

A list of populations, a z-score and associated p-value are output to the screen.

Only putative source populations for which the focal unknown individual falls within their confidence envelopes are retained in the dartR genlight object passed back by the function, along with the focal individual in a population called `unknown`.

Caveats

Each of these approaches is highly intuitive and may be used collectively as a basis for a decision on the source population of an individual of unknown provenance. They are not presented as an alternative to the more sophisticated approaches based on the maximum likelihood approaches, the Bayesian approaches or the Machine Learning methods. The dartRverse approaches should be seen as a preliminary examination, setting the expectation for the outcome of more sophisticated analyses.

PCA assignment

There is a caveat on the use of PCA for eliminating populations from the pool of possible source populations. The confidence ellipse calculated for the data in the two-dimensional ordination space is not the same as the confidence envelope defined in the space of informative dimensions projected on to the two most informative dimensions. A subtle distinction to be sure, but it can matter. It is for this reason that we set the default for p at 0.001 which brings the false exclusion rate down to acceptable levels (ca 0.0009). So do not be tempted to set this parameter `plevel` to 0.05. If you do, the false exclusion rate will rise to ca 5%, which is consequential.

A second concern with the PCA approach is that the individuals that are falsely excluded because of the mismatch in confidence envelopes (all informative

dimensions vs just the top two) are not random. There is a systematic bias toward excluding individuals that fall close to the boundary, so the false exclusions systematically affect the borderline cases that matter most.

So the PCA approach is visual and informative, but should not be used as the sole basis for the exclusion of a putative source population.

Sample sizes

All approaches depend critically on the estimate of the allele frequency profiles of the putative source populations. Without adequate sample sizes, ideally 30 individuals per population, there is a risk of mis-assignment. This risk is managed to a practical extent by insisting on sample sizes of 10 individuals or greater.

Source population not sampled

Finally, there is the possibility that the focal unknown individual has been sourced from a population that is not among those sampled as putative sources. A decision that is based on picking the best supported assignment thus carries with it a risk. The PCA approach and the Mahalanobis approach presented here will assist you in managing that risk, because both approaches admit the possibility that the focal unknown was not sourced from any of the putative source populations.

Where have we come?

The above Session was designed to give you a basic overview of approaches to population assignment.

Having completed this Session, you should now:

- Appreciate the different approaches to population assignment.
- Understand some of the thinking behind the intuitive approaches applied in dartRverse.
- Be aware of some of the limitations in applying population assignment tools, particularly the asymmetry between eliminating putative populations from consideration (definitive) and assigning an individual to a particular population (always with uncertainty).

Session 2: Worked Example

Scenario

The sample data are taken from an unpublished study of diversity across ranges (Figure 1) of the freshwater turtles of the genus *Emydura* from Australia and southern New Guinea. There are currently five taxa recognised – the southern Emydura (*Emydura macquarii*), the northern redfaced turtle (*Emydura australis*), the northern yellowfaced turtle (*Emydura tanybaraga*), the diamondhead (*Emydura worrelli*) and the New Guinea painted turtle (*Emydura subglobosa*). An objective of the analysis is to determine if it is possible to reliably assign an individual of unknown

provenance to its source. This is of obvious relevance too monitoring of illicit wildlife trade. In this worked example, we will first explore this dataset to examine the number of populations sampled, the number of individuals per population, the number of loci scored for each individual and other information.

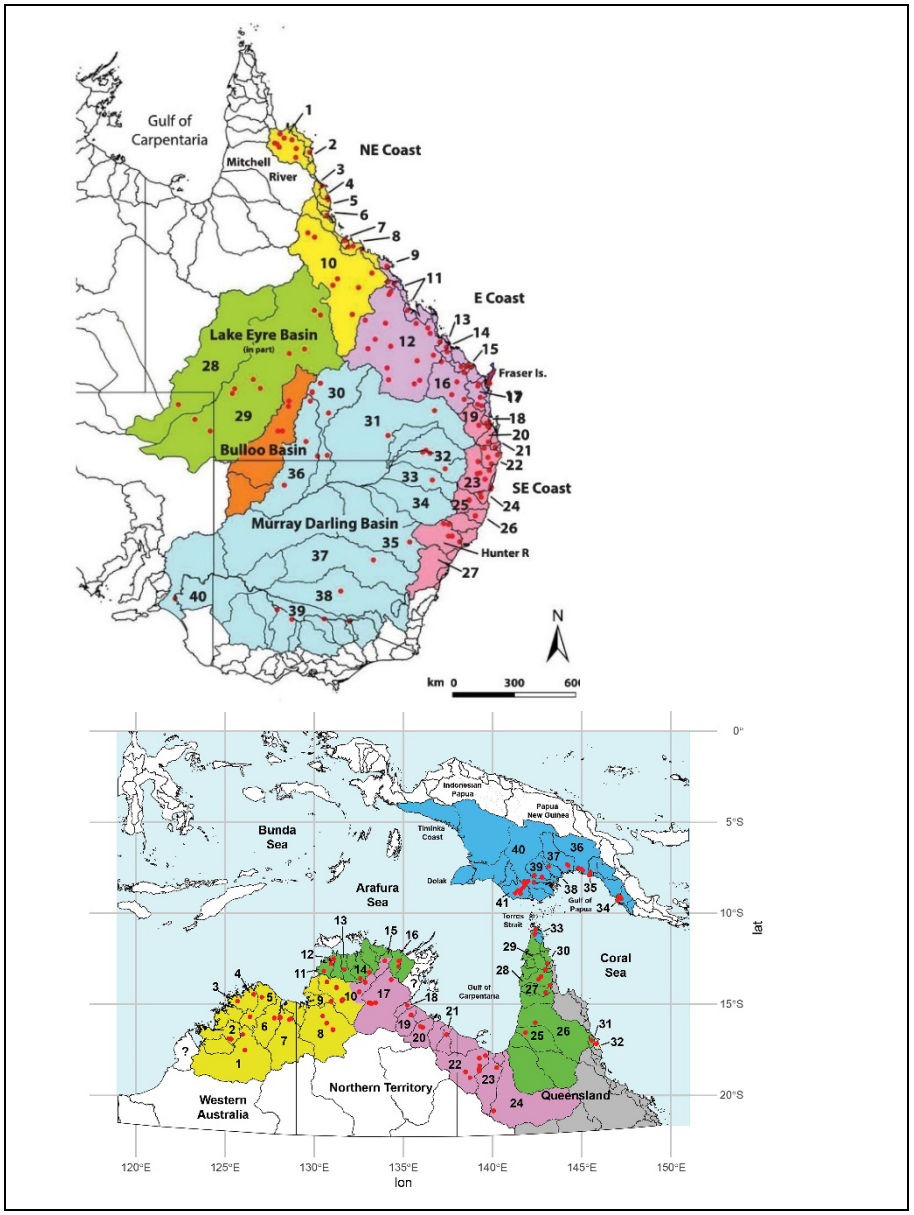


Figure 1. *Maps of Australia showing the comprehensive sampling of Emydura, a species of freshwater turtle, across its range. Each of the points typically represents a sample of at least 10 individuals (Georges et. al., 2018; 2025).*

The Example Data

Reading in the SNP data

The SNP dataset used in this tutorial is `assignment_example1.Rdata` which can be downloaded and read into RStudio using `readRDS()` as below. Set your working directory to the directory with the example data files.

```
setwd(<directory path>)
```

then begin the analysis

```
gl.set.verbosity(3)
```

```
gl <- readRDS("assignment_example1.Rdata")
```

NOTE: This dataset has already been filtered on call rate, reproducibility and read depth. Secondaries have also been filtered (only one SNP per sequence tag retained, at random). Putative admixed individuals have been identified using NewHybrids (Anderson & Thompson, 2002) and removed.

Examining the Contents

Simply typing the name of the dartR genlight object provides a substantial amount of information. We can see that there are 835 individuals scored for 20,688 SNP loci, all but 1.66% having been successfully called. There is a list of individual metrics, such as Genus, Species, Sex etc and a list of locus metrics such as read depth, SnpPosition, CallRate etc.

```
gl
```

```
********************
 *** DARTR OBJECT ***
 ********************
 ** 835 genotypes,  20,688 SNPs , size: 57.3 Mb
   missing data: 289548 (=1.68 %) scored as NA
 ** Genetic data
  @gen: list of 835 SNPbin
  @ploidy: ploidy of each individual  (range: 2-2)
 ** Additional data
  @ind.names:  835 individual labels
  @loc.names:  20688 locus labels
  @loc.all:  20688 allele labels
  @position: integer storing positions of the SNPs [within 69 base sequence]
  @pop: population of each individual (group size range: 3-30)
  @other: a list containing: loc.metrics, ind.metrics, latlon, loc.metrics.flags, verbose, history
   @other$ind.metrics: id, pop, lat, lon, sex, maturity, collector, location, basin, drainage, service,
plate_location
   @other$loc.metrics: AlleleID, CloneID, AlleleSequence, SNP, SnpPosition, CallRate, OneRatioRef,
OneRatioSnp, FreqHomRef, FreqHomSnp, FreqHets, PICRef, PICSnp, AvgPIC, AvgCountRef,
AvgCountSnp, RepAvg, clone, uid, rdepth, monomorphs, maf, OneRatio, PIC, TrimmedSequence
  @other$latlon[g]: coordinates for all individuals are attached
```

We could have used the adegenet accessors to pull this information, for example,

```
nLoc(gl)
  [1] 20688
nInd(gl)
  [1] 835
nPop(gl)
  [1] 81
```

and can in addition, list the individual names and population names

```
indNames(gl)[1:10]
```

```
[1] "AA010915" "AA032703" "UC_00126" "AA032760" "AA013214" "AA011723"
        "AA012411" "AA011893" "AA011896" "AA019237"
```

```
popNames(gl)
```

```
[1] "Brisbane"  "Burdekin"  "Burnett"  "Clarence"  "Cooper_Alvin"
 [6] "Cooper_Cully" "Cooper_Eulbertie" "Dumaresque" "Fitzroy_Alligator" "Fitzroy_Carnavan"
[11] "Fitzroy_Fairburn" "Fraser_Island" "Hunter"  "EmmacJohnWari" "EmmacMaclGeor"
[16] "Mary"  "EmmacMDBBarr" "EmmacMDBBarw" "EmmacMDBBooth" "EmmacMDBBowm"
[21] "EmmacMDBBurr" "EmmacMDBCond" "EmmacMDBCudg" "EmmacMDBDarlBour" "EmmacMDBDarlWeth"
[26] "EmmacMDBDart" "EmmacMDBEulo" "EmmacMDBForb" "EmmacMDBGoul" "GurraGurra"
[31] "EmmacMDBGwyd" "EmmacMDBLach" "EmmacMDBLodd" "EmmacMDBMaci" "EmmacMDBMoon"
[36] "EmmacMDBMurrGunb" "EmmacMDBMurrLock" "EmmacMDBMurrMorg" "EmmacMDBMurrMung"
"EmmacMDBMurrMurr"
```

```
[41] "EmmacMDBMurrTink" "EmmacMDBMurrYarra" "EmmacMDBOven" "EmmacMDBParoBiny"
"EmmacMDBPind"
[46] "EmmacMDBSanf" "EmmacMDBToon" "Normanby"  "Pine"  "EmmacRichCasi"
[51] "EmmacRoss"  "EmmacTweeUki" "EmsubBamuAli" "EmsubBamuAwab" "EmsubMorehead"
[56] "EmsubFlyGuka" "EmsubFlyJikw" "EmsubJardine" "EmsubKerema" "EmsubKikori"
[61] "EmworRoper" "EmtanBlyth" "EmtanFinniss" "EmtanHolrChai" "EmtanMitchell"
[66] "EmtanMitcMitc" "EmtanPascFarm" "EmtanWenlock" "EmvicDaly"  "EmvicDrysdale"
[71] "Fitzroy_WA" "EmvicIsdeBell" "EmvicKingMool" "EmvicOrd"  "EmworClavPung"
[76] "EmworDaly"  "EmworDalySlei" "EmworLeicAlex" "EmworLimmNath" "EmworLiveMann"
[81] "EmworNichGreg"
```

Note: `popNames(gl)` gives a list of population names; `pop(gl)` gives a list of population names against each individual. Samples sizes can thus be obtained using

```
table(pop(gl))
```

Brisbane	Burdekin	Burnett	Clarence	Cooper_Alvin
10	10	11	10	10
Cooper_Cully	Cooper_Eulbertie	Dumaresque	Fitzroy_Alligator	Fitzroy_Carnavan
10	10	10	10	10
Fitzroy_Fairburn	Fraser_Island	Hunter	EmmacJohnWari	EmmacMaclGeor
10	10	10	10	11
Mary	EmmacMDBBarr	EmmacMDBBarw	EmmacMDBBooth	EmmacMDBBowm
10	10	10	9	10
EmmacMDBBurr	EmmacMDBCond	EmmacMDBCudg	EmmacMDBDarlBour	EmmacMDBDarlWeth
10	10	10	10	10
EmmacMDBDart	EmmacMDBEulo	EmmacMDBForb	EmmacMDBGoul	GurraGurra
10	10	10	10	10
EmmacMDBGwyd	EmmacMDBLach	EmmacMDBLodd	EmmacMDBMaci	EmmacMDBMoon
10	10	10	10	10
EmmacMDBMurrGunb	EmmacMDBMurrLock	EmmacMDBMurrMorg	EmmacMDBMurrMung	EmmacMDBMurrMurr
10	10	10	10	10
EmmacMDBMurrTink	EmmacMDBMurrYarra	EmmacMDBOven	EmmacMDBParoBiny	EmmacMDBPind
10	10	10	10	10
EmmacMDBSanf	EmmacMDBToon	Normanby	Pine	EmmacRichCasi
10	11	11	10	10
EmmacRoss	EmmacTweeUki	EmsubBamuAli	EmsubBamuAwab	EmsubMorehead
10	10	10	9	16
EmsubFlyGuka	EmsubFlyJikw	EmsubJardine	EmsubKerema	EmsubKikori
10	30	16	10	4
EmworRoper	EmtanBlyth	EmtanFinniss	EmtanHolrChai	EmtanMitchell
11	10	7	10	9
EmtanMitcMitc	EmtanPascFarm	EmtanWenlock	EmvicDaly	EmvicDrysdale
3	9	10	10	10
Fitzroy_WA	EmvicIsdeBell	EmvicKingMool	EmvicOrd	EmworClavPung
10	12	10	18	10
EmworDaly	EmworDalySlei	EmworLeicAlex	EmworLimmNath	EmworLiveMann
10	7	10	10	9
EmworNichGreg				
12				

Note that some populations have less than 10 individuals. If these are to be used as putative source populations, there will be some additional risk in correct assignment of individuals sourced from these populations.

Analysis

Assign on genotype

Let us take an individual from a river in Queensland, say the Burnett River (n=11), and see how well we can assign this individual to its source population. The individual identity is AA011731.

```
gen.result <- gl.assign.on.genotype(gl,unknown="AA011731",
      nmin=10)
```

```
Starting gl.assign.on.genotype
  Processing genlight object with SNP data
```

```
  Discarding 9 populations with sample size < 10 : EmmacMDBBooth, EmsubBamuAwab,
EmsubKikori, EmtanFinniss, EmtanMitchell, EmtanMitcMitc, EmtanPascFarm,
EmworDalySlei, EmworLiveMann

          population Log Likelihood        AIC        dAIC        AIC.wt assign
3            Burnett      -4926.957   9853.914     0.0000  1.000000e+00    yes
16              Mary      -5341.050  10682.101   828.1863 1.450906e-180     no
1            Brisbane     -19251.444  38502.888 28648.9733  0.000000e+00     no
2            Burdekin     -32844.476  65688.953 55835.0384  0.000000e+00     no
4            Clarence     -31620.048  63240.095 53386.1808  0.000000e+00     no
5        Cooper_Alvin     -42008.293  84016.586 74162.6716  0.000000e+00     no
6        Cooper_Cully     -42849.639  85699.278 75845.3633  0.000000e+00     no
7    Cooper_Eulbertie     -42636.382  85272.764 75418.8497  0.000000e+00     no
8          Dumaresque     -28852.254  57704.509 47850.5946  0.000000e+00     no
9   Fitzroy_Alligator     -12133.240  24266.480 14412.5655  0.000000e+00     no
10   Fitzroy_Carnavan     -13118.904  26237.808 16383.8939  0.000000e+00     no
.....
```

Bang, it is right on the mark – Burnett River. However, note that the result, although convincing, is based on the putative source population with the best AIC weight (hence the zero delta AIC). All other putative populations are measured against this. There is the possibility that there is another population, not sampled or for which the sample size was less than 10, that is the actual source. Caution is required.

Assign on private alleles

Let's try a second approach.

```
pa.result <- gl.assign.pa(gl, unknown="AA011731", nmin=10,
     alpha=0.05)
```

```
Starting gl.assign.pa
  Processing genlight object with SNP data
  Discarding 9 populations with sample size < 10 :
EmmacMDBBooth, EmsubBamuAwab, EmsubKikori, EmtanFinniss, EmtanMitchell,
EmtanMitcMitc, EmtanPascFarm, EmworDalySlei, EmworLiveMann
                 pop count    Z-score  p-value assign
16              Mary    81 -0.1692350 0.567194    yes
3            Burnett    77  0.2743299 0.391916    yes
48              Pine   167  1.1555039 0.123942    yes
21       EmmacMDBCond   785  2.0204271 0.021670     no
46       EmmacMDBToon   668  2.7347470 0.003121     no
15      EmmacMaclGeor  1040  3.4791497 0.000252     no
62          EmvicDaly  1284  3.5437788 0.000197     no
19       EmmacMDBBowm   992  3.6051586 0.000156     no
72      EmworNichGreg  1260  3.8784997 0.000053     no
58         EmworRoper  1273  4.1008215 0.000021     no
24  EmmacMDBDarlWeth   865  4.8762430 0.000001     no
.....
66      EmvicKingMool  1363 24.4944007 0.000000     no
67           EmvicOrd  1333 12.5867638 0.000000     no
68      EmworClavPung  1299 22.5017244 0.000000     no
69          EmworDaly  1307  5.2935238 0.000000     no
70      EmworLeicAlex  1324 15.9637009 0.000000     no
71      EmworLimmNath  1322  5.7857267 0.000000     no
Completed: gl.assign.pa
```

So the count of private alleles held by the focal unknown in comparison to the Mary, Burnett and Pine Rivers is well within expectation. These three populations are putative source populations for specimen AA011731. The remaining 78 populations are no longer under consideration.

A next step might be to examine these three populations further with a PCA assignment.

```
pca_pa_result <- gl.assign.pca(pa.result,unknown="AA011731")
```

```
Starting gl.assign.pca
Calculating a PCA to represent the unknown in the context
                    of putative sources
Eliminating populations for which the unknown is outside
                    their confidence envelope
```

```
Putative source populations: Burnett
Populations eliminated from consideration: Mary, Pine
Returning a genlight object with remaining putative source
                    populations plus the unknown
Completed: gl.assign.pca
```

We see that this analysis restricts the putative source populations further to yield only the Burnett River, which is good because that is the population from which we initially drew the unknown.

Assign on PCA

The result of the PCA assignment shows graphically the unknown AA011731 as lying outside the 2D confidence ellipses for the Mary and Pine Rivers, and within the 2D confidence ellipse of the Burnett River. This is a nice graphical summary of the operation of this script (Figure 2).

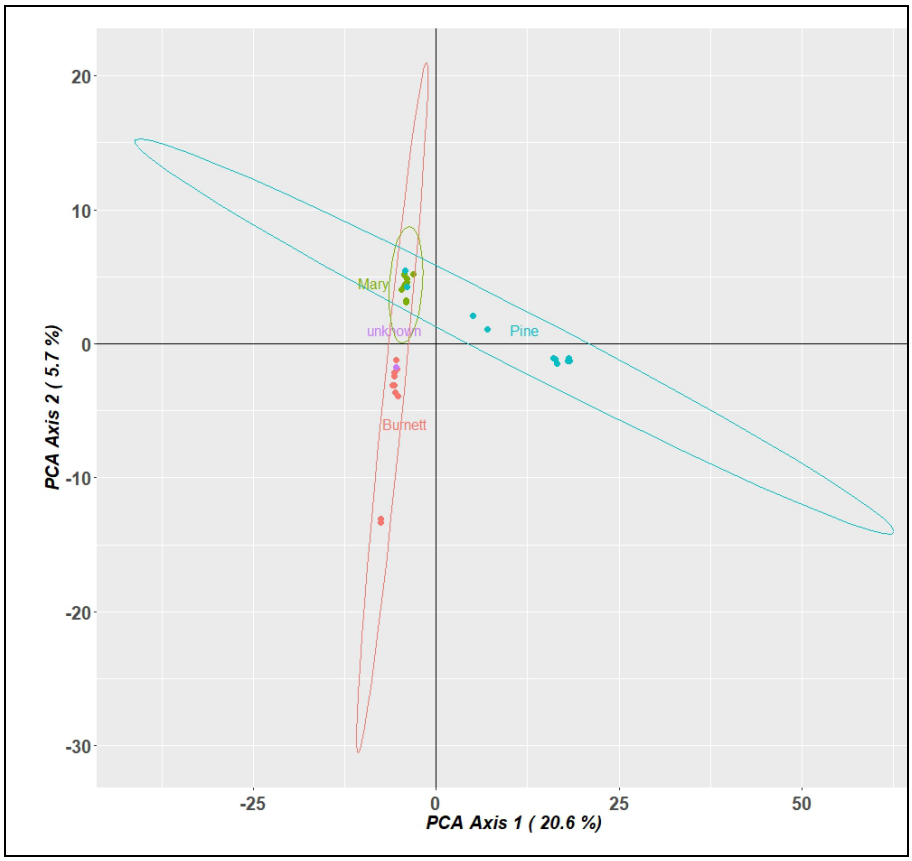


***Figure 2**. A PCA plot of a focal unknown individual (AA011731) shown in purple in relation to the confidence ellipses for the Burnett (orange), Mary (green) and Pine (blue) Rivers. The unknown falls within the confidence ellipse of the Burnett but outside the confidence ellipses of the Mary and Pine.*

Again, it supports the assignment of the unknown to the Burnett River.

Assign on Mahalanobis Distance

We might now like to try the Mahalanobis Distance approach, to see if it provides results that are consistent with the private alleles and PCA approaches.

For computational reasons, lets restrict the candidate putative sources to the 10 best populations identified by the `gl.assign.pa()` script.

```
gl_test <-
     gl.keep.pop(gl,pop.list=c("Mary","Burnett","Pine","EmmacM
     DBCond","EmmacMDBToon","EmmacMaclGeor","EmvicDaly","Emmac
     MDBBowm","EmworNichGreg","EmworRoper"),mono.rm=TRUE)

mahal_result <-
     gl.assign.mahalanobis(gl_test,unknown="AA011731")
```

```
Starting gl.assign.mahalanobis
Number of dimensions with substantial eigenvalues:6.Hardwired limit 20
    Selecting the smallest of the two
    Dimension of confidence envelope set at 6

Assignment of unknown individual: AA011731
Alpha level of significance: 0.001
             pop       MahalD         pval assign
1        Burnett     17.99514 5.504569e-02    yes
2           Mary     39.04125 2.496975e-05     no
3           Pine     74.44904 6.089720e-12     no
4  EmmacMaclGeor  33767.11382 0.000000e+00     no
5   EmmacMDBBowm 734159.09523 0.000000e+00     no
6   EmmacMDBCond  14032.04891 0.000000e+00     no
7   EmmacMDBToon   2665.08984 0.000000e+00     no
8      EmvicDaly  95593.31755 0.000000e+00     no
9  EmworNichGreg 295384.44007 0.000000e+00     no
10    EmworRoper 220440.15550 0.000000e+00     no
  Best assignment is the population with the larges probability
                of assignment, in this case Burnett
  Returning a dataframe with the Mahalanobis Distances
Completed: gl.assign.mahalanobis
```

This analysis again restricts the putative source populations to yield only the Burnett River, which is good because that is the population from which we initially drew the unknown.

Combining approaches

These three approaches can be daisy-chained because the output of each script is compatible with the input to the next. So for example, we could have run

```
pa.result <- gl.assign.pa(gl, unknown="AA011731", nmin=10,
      alpha=0.05, verbose=3)

mahal_result <-
      gl.assign.mahalanobis(pa.result,unknown="AA011731")
```

Because of the caveats associated with assignment on the basis of PCA, it is probably best not to include that approach in a sequential analysis.

Exercise

The authorities have recently raided a premises in Brisbane and found a number of reptiles held without permit. One of these is the painted turtle *Emydura subglobosa*. This species is widespread and common in southern New Guinea, but restricted in Australia to the Jardine River at the tip of Cape York. The Australian population is considered critically endangered under the EPBC Act.

The question is, was the animal sourced from Cape York or imported from New Guinea?

The specimen was genotyped and run in a service with the other available specimens from localities shown in Figure 1. The datafile is `assignment_example1.Rdata`. The SpecimenID is "AA046092"

Can you confidently decide if the animal was sourced from Cape York or New Guinea using the tools we have provided you via dartRverse?

Links to Third-party Software

assignPOP

Software R package assignPOP is for population assignment, that is, determining which reference population an individual of unknown origin most likely came from. Package assignPOP uses machine-learning classifiers trained on SNPs.

- Reads reference (training) and unknown (test) genotypes separately.
- Supports a range of classifiers — linear/quadratic discriminant analysis, naive Bayes, support vector machines, random forest, and neural networks;
- Has validation options such as K-fold cross-validation [`assign.kfold()`] and Monte Carlo resampling [`assign.MC()`] to estimate assignment accuracy on the reference panel before assigning unknowns;
- Can subsample loci (e.g., use the top *n* most informative SNPs) to test how accuracy scales with marker number — useful for panel design;
- Provides assignment probabilities for each candidate population, accuracy metrics, and publication-ready plots.

Compared to the dartRverse methods we played with in the sandpit, assignPOP provides greater flexibility in choice of classifier and better suited to formal accuracy benchmarking; the dartR functions (genotype likelihood, private alleles, PCA ellipse, Mahalanobis) are more interpretable statistically but narrower in scope.

Step 1 — Load the data

```
gl <- gl.load("your_data.rdata", verbose = 0)
```

Step 2 — Split unknowns from the reference set

```
unknown_id <- "AA011731"
gl.unknown <- gl.keep.ind(gl,ind.list=unknown_id,verbose = 0)
gl.ref <- gl.drop.ind(gl, ind.list = unknown_id, verbose = 0)
```

Step 3 — Run the assignment

```
out <- gl.run.assignpop(x = gl.ref, x.unknown = gl.unknown,
  nmin = 10, dir = getwd(), verbose   = 3)
```

`gl.run.assignpop()` handles everything internally: reference populations with fewer than `nmin` individuals are dropped automatically, both genlight objects are reduced to their common loci, allele frequencies are encoded, and `assignPOP::assign.X()` is called.

Results are in `out$results` (assignment data frame), also written to `AssignmentResult.txt` in the working directory. The reference and unknown data objects are available as `out$train` and `out$unknowns` if cross-validation is needed subsequently using

```
assignPOP::assign.kfold(out$train, k.fold = 5, dir = getwd())
```

Refer to the assignPOP documentation for further information.

Results

Assignment result for AA011731

Rank	Pop code	Population name	Probability
1	3	Burnett	0.197
2	25	EmmacMDBDart	0.189
3	27	EmmacMDBForb	0.077
4	11	Fitzroy_Fairburn	0.048

AA011731 is assigned to Burnett with the highest posterior probability (~19.7%). However, EmmacMDBDart is close behind, so we have a ranked result, but not strong discrimination. Full probabilities across all 72 reference populations are in `AssignmentResult.txt`.

KLFDAPC

Software R package KLFDAPC (Kernel Local Fisher Discriminant Analysis of Principal Components) is also for population assignment, that is, determining which reference population an individual of unknown origin most likely came from.

KLFDAPC uses a combination of supervised machine learning and spatial genetic structure analysis. KLFDAPC is designed for datasets with many individuals per population and few populations.

Step 1 — Load the data

```
gl <- gl.load("your_data.rdata", verbose = 0)
```

Step 2 — Split unknowns from the reference set

```
gl.unknown <- gl.keep.ind(gl,ind.list = "AA011731",verbose=0)
gl.ref  <- gl.drop.ind(gl, ind.list = "AA011731", verbose = 0)
```

Step 3 — Run the assignment

```
out <- gl.run.klfdapc(x = gl.ref, x.unknown = gl.unknown,
       n.pc = 10, r = 3, verbose = 3)
out$predictions$posteriors.class1   # Bayesian assignment
```

KLFDAPC is a powerful method for population assignment, but it has strict data requirements that must be met for it to work reliably.

Each reference population must contain at least max(knn, r + 1) individuals. With the default settings (knn = 6, r = 3) this means a minimum of 6 individuals per population. If you increase r to capture more discriminant dimensions (which improves resolution when populations are numerous), the minimum rises accordingly — for example r = 9 requires at least 10 individuals per population.

KLFDAPC works best when the number of putative source populations is small to moderate (roughly 2–20). When there are many populations, two things go wrong simultaneously:

- The discriminant space (r dimensions) becomes too compressed to separate all populations cleanly, so self-assignment accuracy on training individuals drops sharply.

- The Bayesian posterior calculation involves a term that shrinks exponentially with the number of populations and can underflow to zero (producing NaN) when there are more than about 30–40 populations. When this happens, out$predictions$posteriors1 will be NaN and the Bayesian assignment will fail silently.

If your reference panel contains many populations, consider a two-stage approach:

Stage 1 — broad screening. Use `gl.run.assignpop()` or `gl.assign.on genotype()` across the full set of putative source populations. This method scales to any number of source populations and will return a ranked list of posterior probabilities. Identify the top 5–10 candidate source populations for the unknown individual.

Stage 2 — refined assignment. Subset the set of putative source populations to the candidate source populations determined in Stage 1, ensuring each has sufficient individuals, then run `gl.run.klfdapc()` on this reduced panel. With fewer populations the discriminant space is richer, the Bayesian posteriors are numerically stable, and assignment accuracy improves substantially.

This two-stage workflow plays to the strengths of available methods: `gl.run.assignpop()` or `gl.assign.on genotype()` for broad coverage across many populations, and `gl.run.klfdapc()` for fine-grained discrimination among a shortlist of closely related candidates.

Additional Exercises

Exercise 1: Efficacy within basins

River systems are different from terrestrial systems in that there are, for most aquatic organisms at least, distinct barriers to movement in the form of drainage divides. Accumulation of genetic differences between drainages tends to make population assignment more definitive.

Here you are asked to evaluate the effectiveness of our methods for assignment to population within the Murray-Darling Basin in comparison with effectiveness of assignment to discrete populations on the seaboard. The Murray-Darling Basin is Australia's largest river and is classified into many sub-basins that have been sampled. These sub-basins are of course interconnected.

Here are some individuals to use in your evaluation.

Basin	Sub-basin	popName	Specimen
MDB	Condamine	EmmacMDBCond	AA032809
MDB	Lachlan	EmmacMDBForb	AA010936
MDB	Murray	EmmacMDBMurrYarra	KBF_M1.08
MDB	Lower Murray	GurraGurra	AA032715
Clarence		Clarence	UC_00157
Burnett		Burnett	AA011741
Burdekin		Burdekin	AA019241

The datafile is `assignment_example1.Rdata`.

NOTE: You will need to set `nmin=9` because we are taking one animal out in the evaluation and most populations have only 10 individuals.

	What do you conclude?

Exercise 2: Individual outside sample set

Let us consider what happens when we try to assign an individual that has been collected from a population that is not in our reference set.

The first individual is *Emydura subglobosa*, AA036611, from the Kikori River in Papua New Guinea. Only four animals have been caught there in several years of study.
The second animal is *Emydura tanybaraga*, G121, from the Pascoe River on Cape York.
Both of these populations were eliminated from consideration because they had less than 10 animals sampled.
The datafile is `assignment_example1.Rdata`**.**

How do the three techniques -- private alleles, PCA and Mahalanobis Distance -- perform? What about the third-party software assignPOP and KLFDAPC? What do you conclude in each case?

Where have we come?

The above Session was designed to give you some practical experience in applying the scripts in dartRverse for population assignment. Having completed this Tutorial, you should now able to:

- Apply each of the four techniques – allele frequency, private alleles, PCA and Mahalanobis Distance to get a feel for your data and the assignment of unknowns.
- Sensibly integrate the results of three approaches in coming to a decision.
- Apply established packages available for population assignment, such as assignPOP.
- Be aware of the limitations of the approaches in terms of making a definitive assignment, given that there are risks associated with selecting the assignment on the basis of best statistical support.

Further reading

Anderson, E.C. and Thompson, E.A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* 160:1217–1229. https://doi.org/10.1093/genetics/160.3.1217**.**

Beacham, T.D., Wallace, C., MacConnachie, C., Jonsen, K., McIntosh, B., Candy, J.R. and Withler, R.E. (2018). Population and individual identification of Chinook salmon in British Columbia through parentage-based tagging and genetic stock identification with single nucleotide polymorphisms. *Canadian Journal of Fisheries and Aquatic Sciences* 75:1096–1105. https://doi.org/10.1139/cjfas-2017-0168**.**

Chen, K.-Y., Marschall, E.A., Sovic, M.G., Fries, A.C., Gibbs, H.L. and Ludsin, S.A. (2018). assignPOP: an R package for population assignment using genetic, non-genetic, or integrated data in a machine-learning framework. *Methods in Ecology and Evolution* 9:439–446. https://doi.org/10.1111/2041-210X.12897.

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* 153:1989–2000. https://doi.org/10.1093/genetics/153.4.1989.

Degen, B., Blanc-Jolivet, C., Stierand, K. and Gillet, E. (2017). A nearest neighbour approach by genetic distance to the assignment of individual trees to geographic origin. *Forensic Science International: Genetics* 27:132–141. https://doi.org/10.1016/j.fsigen.2016.12.011.

Georges, A., Gruber, B., Pauly, G.B., Adams. M., White, D., Young, M.J., Kilian, A., Zhang, X., Shaffer, H.B. and Unmack, P.J. (2018). Genome-wide SNP markers breathe new life into phylogeography and species delimitation for the problematic short-necked turtles (Chelidae: *Emydura*) of eastern Australia. *Molecular Ecology* 27:5195-5213

Georges, A., Unmack, P.J., Kilian, A., Zhang, X., Amepou, Y. and Dissanayake, D.S.B. (2025). Diagnosability to inform species delimitation for the genus *Emydura* (Testudines: Chelidae) from northern Australia. *bioRxiv* 2025.07.10.664252. https://doi.org/10.1101/2025.07.10.664252.

Manel, S., Gaggiotti, O.E. and Waples, R.S. (2005). Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution* 20:136–142. https://doi.org/10.1016/j.tree.2004.12.004.

Ogden, R. and Linacre, A. (2015). Wildlife forensic science: a review of genetic geographic origin assignment. *Forensic Science International: Genetics* 18:152–159. https://doi.org/10.1016/j.fsigen.2015.02.008.

Paetkau, D., Calvert, W., Stirling, I. and Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* 4:347–354. https://doi.org/10.1111/j.1365-294X.1995.tb00227.x.

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

Puechmaille, S.J. (2016). The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Molecular Ecology Resources* 16:608–627.

Qin, X., Chiang, C.W.K. and Gaggiotti, O.E. (2022). KLFDAPC: a supervised machine learning approach for spatial genetic structure analysis. *Briefings in Bioinformatics* 23:bbac202. https://doi.org/10.1093/bib/bbac202.

Rannala, B. and Mountain, J.L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America* 94:9197–9201. https://doi.org/10.1073/pnas.94.17.9197.

Sylvester, E.V.A., Bentzen, P., Bradbury, I.R., Clément, M., Pearce, J., Horne, J. and Beiko, R.G. (2018). Applications of random forest feature selection for fine-

scale genetic population assignment. *Evolutionary Applications* 11:153–165. https://doi.org/10.1111/eva.12524**.**

Wang, J. (2017). The computer program STRUCTURE for assigning individuals to populations: easy to use but easier to misuse. *Molecular Ecology Resources* 17:981–990. https://doi.org/10.1111/1755-0998.12650**.**

Wilson, G.A. and Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* 163:1177–1191. https://doi.org/10.1093/genetics/163.3.1177**.**

Ende