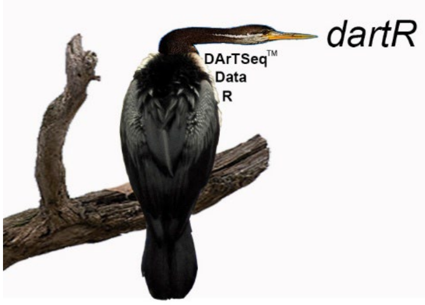
Wildlife Genomics@UC

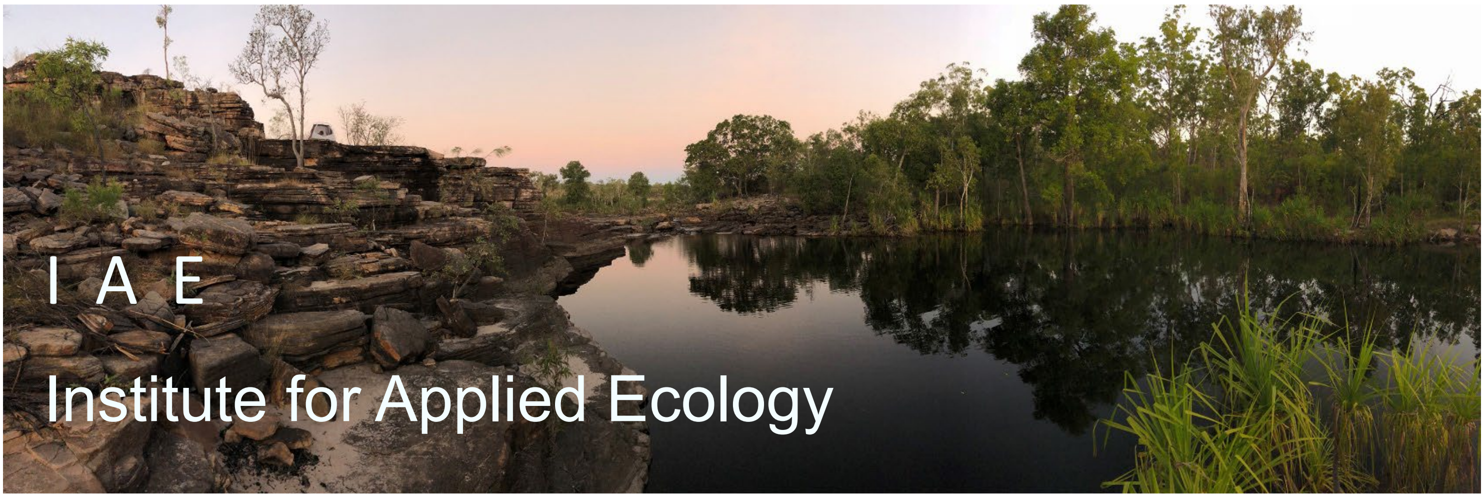
UNIVERSITY OF CANBERRA

SNP Analysis using dartR



Guide to Population Assignment

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., Mijangos, J.L and Furlan, E. (2025). Guide to population assignment. Version 1. Institute for Applied Ecology, University of Canberra.

dartR 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

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 Atlantic 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 (Fogel *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 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 dartR for Assignment

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

The focus of dartR scripts is on exploratory analysis. We present three 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. They can be used individually or to progressively eliminate unlikely source populations in sequence.

The assignment scripts in dartR take as input a 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 dartR 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 will not 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 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 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 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 dartR approaches should be seen as a preliminary examination, setting the expectation for the outcome of more sophisticated analyses.

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.

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 the thinking behind the intuitive approaches applied in dartR.
- 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.

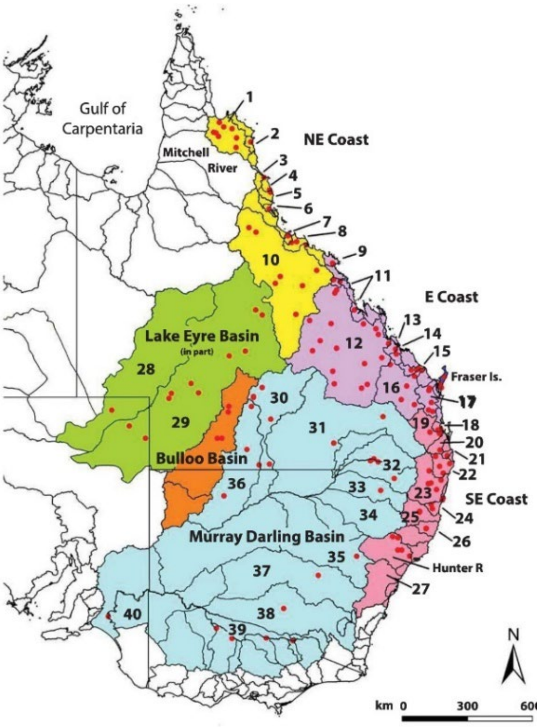
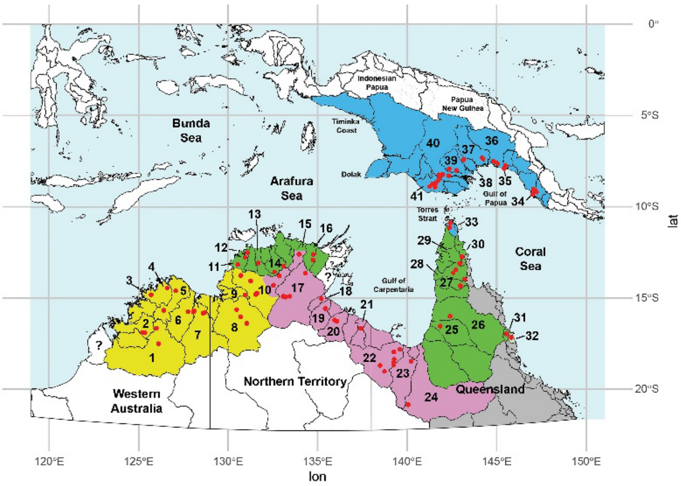


Figure 1. Maps of Australia showing the comprehensive sampling of *Emydura*, a species of freshwater turtle, across its range. Each of the points typically represents a sample of at least 10 individuals (Georges et. al., 2018; 2025).



The Example Data

Reading in the SNP data

The SNP dataset used in this tutorial is `assignment.example1.Rdata which` can be 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 genlight object provides a substantial amount of information. We can see that there are 783 individuals scored for 21,816 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

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.

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.

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

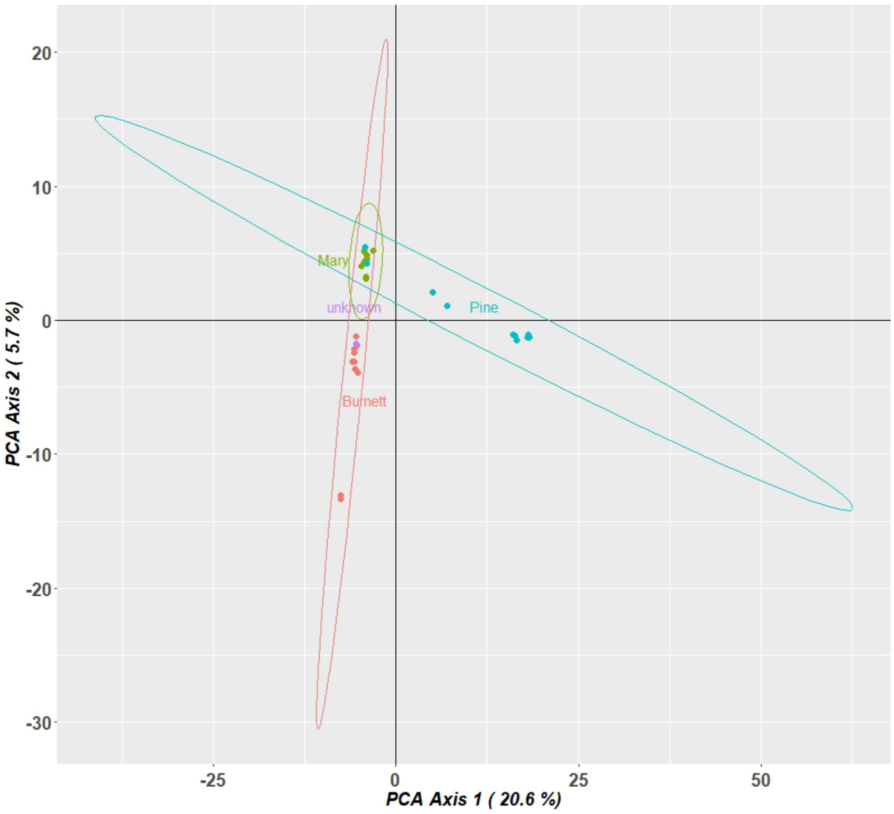


Figure 2. A PCA plot of a focal unknown individual (AA011731) shown in purple in relation to the confidence ellipses for the Burnett (orange), Mary (green) and Pine (blue) Rivers. The unknown falls within the confidence ellipse of the Burnett but outside the confidence ellipses of the Mary and Pine.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

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
  Warning: Listed population unknown not present in the dataset --
ignored
  Rendering the data matrix dense by imputation
  Undertaking a PCA
Starting gl.colors
Selected color type 2
Completed: gl.colors
  Dimensions retained: 4
  Number of dimensions with substantial eigenvalues (Broken-Stick
Criterion): 4 . Hardwired limit 10
    Selecting the smallest of the two
    Dimension of confidence envelope set at 4
Assignment of unknown individual: AA011731
Alpha level of significance: 0.001
             pop       MahalD          pval assign
1           Mary 1.032222e+01  4.126933e-01    yes
2        Burnett 2.609197e+01  3.618440e-03    yes
3           Pine 5.515667e+01  2.952235e-08     no
4   EmmacMDBCond 4.926272e+02  1.660274e-99     no
5   EmmacMDBToon 1.275080e+03 1.057785e-266     no
6  EmmacMaclGeor 1.406157e+04  0.000000e+00     no
7   EmmacMDBBowm 1.716823e+04  0.000000e+00     no
8      EmvicDaly 2.656608e+06  0.000000e+00     no
9  EmworNichGreg 9.621106e+05  0.000000e+00     no
10    EmworRoper 2.291194e+05  0.000000e+00     no
  Best assignment is the population with the largest probability
                of assignment, in this case Mary
  Returning a genlight object with the putative source populations and
the unknown
Completed: gl.assign.mahalanobis
```

This analysis again restricts the putative source populations to yield the Mary and the Burnett River, which are adjacent drainages on the east coast of Queensland. The populations in these rivers are very similar genetically.

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`.

For this exercise, you might want to restrict the data to only those for the target species Emydura subglobosa. To do this, use

```
gl2 <- gl.keep.pop(gl,pop.list=c("EmsubBamuAli",
"EmsubMorehead", "EmsubFlyGuka",
"EmsubFlyJikw","EmsubJardine", "EmsubKerema"),
mono.rm=TRUE)
```

`EmsubJardine` is from the tip of Cape York, Australia. The other localities are from southern New Guinea.

The seized specimen has SpecimenID "AA046092"

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

Links to Third-party Software

assignPOP

assignPOP is an R package for population assignment using a machine-learning framework. It employs supervised machine-learning to evaluate the discriminatory power of your data collected from source populations to assign unknown individuals. *assignPOP* is able to analyze large genetic datasets, non-genetic datasets, or undertake analyses that draw upon a combination of genetic and non-genetic data.

To install the package, use

```
install.packages("assignPOP")

library(assignPOP)
```

Refer to the manual https://alexkychen.github.io/assignPOP/.

assignPOP can read the genepop file format using `read.Genepop()`, so the easiest avenue to providing your genlight object to *assignPOP* is via `gl2genepop()`. Let's work with the data used in the above exercise.

```
gl2 <- gl.keep.pop(gl,pop.list=c("EmsubBamuAli",
     "EmsubMorehead", "EmsubFlyGuka",
     "EmsubFlyJikw","EmsubJardine", "EmsubKerema"),
     mono.rm=TRUE)

data.gen <- gl2genepop(gl,outfile="genepop.txt")
```

then you can run assignPOP functions in accordance with the instructions.

```
gen <- read.Genepop("data.gen", pop.names=c(popNames(gl2)))

cross.val <- assign.MC(gen,dir="D:/workspace/assignpop/")
```

A warning: assignPop can be a bit idiosyncratic. You may have to fiddle a bit with syntax.

KLFDAPC

KLFDAPC stands for Kernel Local Fisher Discriminant Analysis of Principal Components. It is a supervised non-linear approach for inferring individual geographic genetic structure that is believed to have advantages over *PCA* or *DAPC*. The developers tested the power of *KLFDAPC* to infer population structure and to predict individual geographic origin using neural networks. Their simulation results showed that *KLFDAPC* has higher discriminatory power than *PCA* and *DAPC*.

To install the package, use

```
requireNamespace("SNPRelate")

if (!requireNamespace("BiocManager", quietly=TRUE))

  install.packages("BiocManager",repos = "http://cran.us.r-
     project.org")

if (!requireNamespace("SNPRelate", quietly=TRUE))

  BiocManager::install("SNPRelate")

if (!requireNamespace("DA", quietly=TRUE))

  devtools::install_github("xinghuq/DA")

if (!requireNamespace("vegan", quietly=TRUE))

  install.packages("vegan")

if (!requireNamespace("PCAviz", quietly=TRUE))

 devtools::install_github("NovembreLab/PCAviz",build_vignettes
     = FALSE)

devtools::install_github("xinghuq/KLFDAPC")


 library(KLFDAPC)
 library(SNPRelate)
 library(vegan)
 library(PCAviz)
```

KLFDAPC expects a GDS file format as input. So first convert your genlight object to GDS then read it in to KLFDAPC.

```
gl2gds(gl2,outfile="test.gds",outpath=getwd())

gen <- SNPRelate::snpgdsOpen("test.gds")
```

then follow the instruction manual.

Where have we come?

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

- Apply each of the three techniques – allele frequency, private alleles, PCA and Mahalanobis Distance.
- Be able to sensibly integrate the results of three approaches in coming to a decision.
- Examine more formal approaches drawn from the literature, like *assignPOP* and *KLFDAPC*.

Session 3: SNP Panels for Routine Assignment

Overview

Single-nucleotide polymorphism (SNP) panels are targeted sets of SNP markers used for genotyping many loci across the genome —typically ranging from tens to hundreds of markers. These panels enable fast, cost-effective analysis of many samples and yield unambiguous, reproducible data, making them well-suited to high-throughput applications and standardized monitoring across laboratories (von Thaden et al., 2020). SNP panels can be run on array or PCR-based platforms, which is especially valuable for low-quality DNA (e.g. scat, hair, faeces) where whole-genome sequencing (WGS) may fail (Armstrong et al., 2025).

Marker Selection

The utility of a SNP panel depends critically on how markers are selected. Different conservation questions require different types of markers and poor SNP selection can lead to reduced power to address questions and may produce inaccurate results.

- **For population assignment,** highly differentiated loci (e.g. F_{ST} outliers) or loci contributing most to discrimination between populations using DAPC (Discriminant Analysis of Principal Components) are most informative (Bertola et al., 2022, Magliolo et al., 2021)
- **For individual identification,** loci with high minor allele frequencies or high Polymorphic Information Content (PIC) are preferred to maximise genotype uniqueness (Wehrenberg et al., 2024).
- **For parentage or relatedness**, high PIC and low linkage disequilibrium are ideal (Spitzer et al., 2016).
- **To detect hybridisation**, panels should include diagnostic SNPs that are fixed or nearly fixed between species or subspecies (Stronen et al., 2022).

To ensure panel robustness, SNP discovery should use a wide sampling of individuals spanning the species' full distribution, thereby minimizing ascertainment bias and improving the panel’s utility across populations (Quinto-Cortés et al., 2018).

SNP panels are most appropriate when many individuals must be genotyped on a fixed set of informative markers, such as long-term monitoring or enforcement. On the other hand, SNP panels are not suitable for discovering new genetic variants or when no prior SNP data exists. Additionally, SNP panels are economically inefficient for analysing small sample sizes since most platforms require batch processing (e.g. 96-384 samples per run) making urgent one-off samples more difficult to accommodate.

Applications

SNP panels are increasingly being used to address conservation and management questions across taxa. In wildlife trade enforcement, SNP panels designed with highly informative, population-diagnostic SNPs have successfully assigned confiscated animals and products to their geographic origin, aiding prosecutions (see examples in cheetahs and jaguars; Magliolo et al., 2021, Zenato-Lazzari et al., 2025). Similar approach to population assignment have helped clarify population boundaries: for example, custom SNP assays helped assign individual lions to one of four major clades, improving phylogeographic resolution(Bertola et al., 2022). In agriculture, SNP panels can be applied to cattle to reliably assign individuals to their breed of origin (Jasielczuk et al., 2024), and in forensics, SNP panels help resolve the origin of human remains from degraded samples (Terrado-Ortuño & May, 2025).

SNP panels are also effective for detecting hybridisation. A 192-SNP panel designed with species-diagnostic loci correctly distinguished three distinct wolf populations from dogs and jackals and identified first-generation hybrids (Stronen et al., 2022). Such panels are particularly valuable in areas of conservation concern where hybridisation threatens the genetic integrity of endangered species.

For non-invasive genetic monitoring, SNP panels provide reliable genotyping from degraded samples. In European bison, a 96-SNP panel enabled genotyping from faecal, hair, urine and saliva samples for individual and parental assignment, sex determination and breeding line identification (Wehrenberg et al., 2024). Similarly, small SNP panels (~100 markers) have successfully been applied to faecal samples in brown bears to estimated population size and relatedness across a landscape (Spitzer et al., 2016) and identified population assignment and evolutionary lineages of endangered fish for management (Starks et al., 2016).

Trait-linked SNPs can also be incorporated into panels to extend utility beyond population genetics. In agriculture, SNPs associated with agronomic traits enable prediction of genetic merit and assist selective breeding (Ohm et al., 2024). In wildlife or human applications, phenotype-informative markers can enhance individual identification, especially in forensic contexts involving trace or degraded samples (Armstrong et al., 2025, Terrado-Ortuño & May, 2025).

By aligning SNP selection with the intended use—be it assignment, detection of hybrids, demographic monitoring, or trait prediction—researchers can build targeted panels with high resolution and minimal redundancy. When well designed, SNP panels can serve as efficient tools for both conservation research and management, especially for routine applications like population monitoring, enforcement, or management decisions in fisheries.

Using dartR for SNP Panel Selection

Multiple Options

We use nine complementary metrics to identify candidate loci for inclusion in SNP panels. Each metric targets a different aspect of genetic informativeness, allowing flexibility depending on the research question.

- **DAPC (**Discriminant Analysis of Principal Components): Identifies loci that contribute most to discrimination between populations using DAPC.
- **Pahigh** (Private Alleles – High frequency): Selects loci containing private alleles with high frequencies, which can be informative for distinguishing populations.
- **Monopop** (Monomorphic Within Populations): Selects loci that are monomorphic within populations, useful for detecting between-population structure.

- **PIC** (Polymorphic Information Content): Targets loci with high PIC values, maximising individual-level discrimination power for applications such as parentage analysis, individual ID, relatedness, or genetic mark-recapture.
- **PICdart (**Polymorphic Information Content based on presence/absence): Analogous to PIC but based on allele presence/absence rather than frequencies.
- **Hafall** (High Allele Frequency – Across Populations): Selects loci with the highest allele frequencies across all populations, favouring broadly polymorphic markers
- **Hafpop** (High Allele Frequency – Within Populations): Selects loci with the highest allele frequencies within individual populations, enhancing within-population resolution
- **Random**: Selects loci at random, providing an unbiased representation of diversity across the genome
- **Stratified**: Stratified random sampling of loci based on allele frequencies to ensure even representation across the allele frequency spectrum

The choice of metric(s) used to design the SNP panel should align with the specific research objectives (see notes above on marker selection). The script `gl.select.panel()` allows users to either derive all candidate loci from a single metric or combine loci selected across multiple metrics to address multiple questions simultaneously. The number of loci included in the final panel can be tailored accordingly.

Evaluate Panel Performance

Once the final panel has been selected, its performance can be assessed using the `gl.check.panel()` function. This script evaluates the panel across a suite of key genetic metrics, including:

- FST – genetic differentiation
- FIS – inbreeding coefficient
- NALL – number of alleles
- HE – expected heterozygosity
- HO – observed heterozygosity
- NE – effective population size

To assess how well the panel captures overall genetic patterns, results from the panel are compared to those obtained using the full SNP dataset, and R^2 values are calculated to quantify concordance. Panel performance can be evaluated across all available metrics or a selected subset, depending on the specific research questions the panel was designed to address.

Worked example

Scenario

The sample data are taken from an unpublished study of on one of the Australia's smallest and most endangered fish: the red-fin blue eye *Scaturiginichthys vermeilipinnis*. Predation and competition with introduced eastern mosquitofish *Gambusia holbrooki* have likely led to the species' extirpation from all but a single spring (Kerezsy & Fensham, 2013). Human-mediated reintroductions have been implemented for redfin blue eye into several nearby invader-free springs (Furlan et al., 2020). Ongoing genetic assessment of these populations remains essential to monitor levels of genetic diversity yet, due to the small size of the species, current tissue sampling methods are destructive. Various non-invasive DNA applications are currently being explored to determine their use in population monitoring, including swabbing and environmental DNA (eDNA) sampling.

An objective of the analysis is to select a panel of 100 SNPs that would be suitable for the ongoing genetic monitoring of the species. In this worked example, we will first explore an existing SNP dataset to examine the breadth and depth of data available (i.e., number of populations sampled, the number of individuals per population, as well as data quality).

Example Data

Read in the SNP data

The SNP dataset used in this tutorial is assignment_example2.Rdata which can be 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_example2.Rdata ")
```

Examine the Contents

Simply typing the name of the genlight object provides a substantial amount of information. We can see that there are 382 individuals scored for 9,849 SNP loci.

```
gl

********************
 *** DARTR OBJECT ***
 ********************
 ** 383 genotypes,  9,849 SNPs , size: 15.7 Mb

    missing data: 47272 (=1.25 %) scored as NA
 ** Genetic data
   @gen: list of 383 SNPbin
   @ploidy: ploidy of each individual  (range: 2-2)
 ** Additional data
   @ind.names:  383 individual labels
   @loc.names:  9849 locus labels
   @loc.all:  9849 allele labels
   @position: integer storing positions of the SNPs [within 69 base
sequence]
   @pop: population of each individual (group size range: 10-64)
   @other:    a    list    containing:    loc.metrics,    ind.metrics,
loc.metrics.flags, verbose, history
    @other$ind.metrics: id, pop
    @other$loc.metrics:     AlleleID,     CloneID,     AlleleSequence,
TrimmedSequence,              Chrom_Scaturiginichthys_vermeilipinnis2,
ChromPosTag_Scaturiginichthys_vermeilipinnis2,
ChromPosSnp_Scaturiginichthys_vermeilipinnis2,
AlnCnt_Scaturiginichthys_vermeilipinnis2,
AlnEvalue_Scaturiginichthys_vermeilipinnis2,
Strand_Scaturiginichthys_vermeilipinnis2,  SNP,  SnpPosition,  CallRate,
OneRatioRef,  OneRatioSnp,  FreqHomRef,  FreqHomSnp,  FreqHets,  PICRef,
PICSnp, AvgPIC, AvgCountRef, AvgCountSnp, RepAvg, clone, uid, rdepth,
monomorphs, maf, OneRatio, PIC
   @other$latlon[g]: no coordinates attached
```

Pull out the populations and the number of individuals in each population.

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

```
BHA_A2    E504       E508      E509    E518   NW30   NW70   NW72
  19       19         20        20      20     20     20     10
NW80    PJTub1.2.3  PJTub4.5  PJTub6   SE60   SW10   SW20   SW60
  10       64         51        33      20     19     18     20
```

Filter to Reduce Loci

We now filter the data to reduce the number of loci to those that are likely to be the most reliable. We will use default settings.

```
gl <- gl.filter.secondaries(gl)
gl <- gl.filter.callrate(gl)
gl <- gl.filter.reproducibility(gl)
gl <- gl.filter.rdepth(gl)
```

which brings us down to 5,134 loci.

We might like to filter further on minor allele frequency

```
gl <- gl.filter.maf(gl, threshold = 5, by.pop = FALSE)
```

which brings us down to 4,990 loci.

Selecting informative SNPs

We can now select a subsample of SNPs using several different methods. Depending on the aim of your panel you might want to select SNPs that are informative for population structure (Fst) or inbreeding (Ho). It turns out that the DAPC method is a good method to select SNPs that are informative for population structure. The code below uses this method to select 50 SNPs.

```
panel <- gl.select.panel(gl, method="dapc", nl = 50)
```

This may take a while to run.

```
Starting gl.select.panel
  Processing genlight object with SNP data
Starting gl.keep.loc
  Processing genlight object with SNP data
  List of loci to keep has been specified
  Deleting all but the specified loci
Completed: gl.keep.loc
Completed: gl.select.panel
```

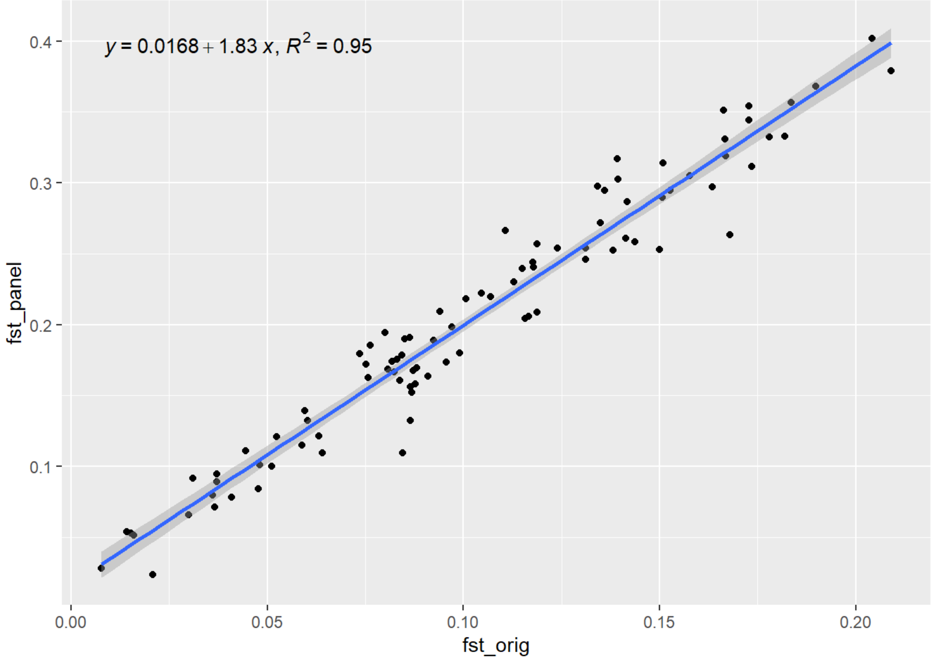
This provides us with a panel with approximately 50 of the most informative loci in the context of assignment.

```
nLoc(panel)
```

```
[1] 82
```

We can now assess the 82 SNPs against the original 4,990 SNPs using F_{ST}.

```
outdapc <- gl.check.panel(panel, rfbe20_6, parameter = "Fst")
```



The reduced SNP data set correlates well with the result we would have got with the full set of filtered loci.

Where have we come?

The above Session was designed to give you an overview of the scripts in dartR for selecting a SNP panel from an existing genomic dataset. Having completed this Session, you should now able to:

- Filter putative SNPs for inclusion in a SNP panel based on suitability of flanking regions for primer design and removing linkage loci
- Modify the selection of suitable loci for inclusion in a panel according to key metrics and number of loci
- Identify concordance between the selected SNP panel and the complete SNP dataset across key measures of genetic diversity i.e., F_{ST}, F_{IS}, N_{ALL}, H_E, H_O, and N_E

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 do you conclude in each case?

Exercise 3: Generate a reduced SNP panel

Here you are asked to design a SNP panel to assign individuals back to their population of origin.

The objective is to see if we can maintain effective assignment of unknown individuals to putative sources with a reduced set of only 100 of the 20,000 SNPs.

Take the same individual used in the earlier example, AA011731, and see how well the SNP panel can assign this individual to its source population, the Burnett River.

You are also asked to check to see how well the selected panel can assign individuals to their population of origin in general.

The datafile is `assignment_example1.Rdata`.

What do you conclude? When might you think such a panel would be of use?

Where have we come?

The above Tutorial was designed to give you an appreciation of the logic behind population assignment and an introduction to some of the tools available in or through dartR to assign unknown individuals to putative source populations. On completion of this tutorial, you should

- Appreciate the strengths and weaknesses of the different approaches to population assignment.
- Be able to apply some of the simpler approaches to population assignment, via assignment by genotype, elimination of putative populations from consideration based on private alleles and PCA, and via considerations using Mahalanobis distance.
- Be aware of some of the third party software for population assignment and how to migrate your genlight data to a format suitable for applying these more sophisticated approaches.
- Know how to reduce your data set to a smaller set of loci while maximally retaining discriminatory power; this can be useful in establishing a small panel of SNPs for routine screening in assignment.
- 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., Thompson, E.A. 2002. A model-based method for identifying species hybrids using multilocus genetic data. Genetics 160: 1217–1229.

Armstrong, E. E., Li, C., Campana, M. G., Ferrari, T., Kelley, J. L., Petrov, D. A., . . . Mooney, J. A. (2025). A Pipeline and Recommendations for Population and Individual Diagnostic SNP Selection in Non-Model Species. Molecular Ecology Resources, 25, e14048. https://doi.org/10.1111/1755-0998.14048.

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.

Bertola, L. D., Vermaat, M., Lesilau, F., Chege, M., Tumenta, P. N., Sogbohossou, E. A., . . . Vrieling, K. (2022). Whole genome sequencing and the application of a SNP panel reveal primary evolutionary lineages and genomic variation in the lion (Panthera leo). BMC Genomics, 23, 321. 10.1186/s12864-022-08510-y

Chen, K.-Y., Marschall, E. A., Sovic, M. G., Fries, A. C., Gibbs, H. L., & 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.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153: 1989-2000.

Degen, B., Blanc-Jolivet, C., Stierand, K., Gillet, E. (2017). A nearest neighbour approach by genetic distance to the assignment of individual trees to geographic origin. Forensic Science International: Genetics 77: 132-141.

Furlan, E. M., Gruber, B., Attard, C. R. M., Wager, R. N. E., Kerezsy, A., Faulks, L. K., . . . Unmack, P. J. (2020). Assessing the benefits and risks of translocations in depauperate species: A theoretical framework with an empirical validation. Journal of Applied Ecology, 57, 831-841. https://doi.org/10.1111/1365-2664.13581

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. and Dissanayake, D.S.B. 2025. Lineages as species or lineages within species – using diagnosability to better inform species delimitation (Chelidae: *Emydura*). Molecular Phylogenetics and Evolution, in review (http://ssrn.com/abstract=5226604).

Jasielczuk, I., Gurgul, A., Szmatoła, T., Radko, A., Majewska, A., Sosin, E., . . . Ząbek, T. (2024). The use of SNP markers for cattle breed identification. Journal of Applied Genetics, 65, 575-589. 10.1007/s13353-024-00857-0

Kerezsy, A. & Fensham, R. (2013). Conservation of the endangered red-finned blue-eye, Scaturiginichthys vermeilipinnis, and control of alien eastern gambusia, Gambusia holbrooki, in a spring wetland complex. Marine and Freshwater Research, 64, 851–863.

Magliolo, M., Prost, S., Orozco-Terwengel, P., Burger, P., Kropff, A. S., Kotze, A., . . . Dalton, D. L. (2021). Unlocking the potential of a validated single nucleotide polymorphism array for genomic monitoring of trade in cheetahs (*Acinonyx jubatus*). Molecular Biology Reports, 48, 171-181. 10.1007/s11033-020-06030-0.

Manel, S., Gaggiotti, O. E., & Waples, R. S. (2005). Assignment methods: Matching biological questions with appropriate techniques. Trends in Ecology & Evolution, 20, 136–142.

Ogden, R. and Linacre, A. (2015). Wildlife forensic science: A review of genetic geographic origin assignment. Forensic Science International: Genetics 18: 152 – 159.

Ohm, H., Åstrand, J., Ceplitis, A., Bengtsson, D., Hammenhag, C., Chawade, A. & Grimberg, Å. (2024). Novel SNP markers for flowering and seed quality traits in faba bean (Vicia faba L.): characterization and GWAS of a diversity panel. Frontiers in Plant Science, 15, 1348014.

Paetkau, D.; Calvert, W.; Stirling, I.; Strobeck, C. Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. 1995, 4, 347–354.

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., 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.

Quinto-Cortés, C. D., Woerner, A. E., Watkins, J. C. & Hammer, M. F. (2018). Modeling SNP array ascertainment with Approximate Bayesian Computation for demographic inference. Scientific Reports, 8, 10209. 10.1038/s41598-018-28539-y

Rannala, B., and J. L. Mountain, 1997. Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA 94: 9197–9201.

Spitzer, R., Norman, A. J., Schneider, M. & Spong, G. (2016). Estimating population size using single-nucleotide polymorphism-based pedigree data. Ecology and Evolution, 6, 3174-3184. https://doi.org/10.1002/ece3.2076

Starks, H. A., Clemento, A. J. & Garza, J. C. (2016). Discovery and characterization of single nucleotide polymorphisms in coho salmon, Oncorhynchus kisutch. Molecular Ecology Resources, 16, 277-287. https://doi.org/10.1111/1755-0998.12430

Stronen, A. V., Mattucci, F., Fabbri, E., Galaverni, M., Cocchiararo, B., Nowak, C., . . . Caniglia, R. (2022). A reduced SNP panel to trace gene flow across southern European wolf populations and detect hybridization with other Canis taxa. Scientific Reports, 12, 4195. 10.1038/s41598-022-08132-0

Terrado-Ortuño, N. & May, P. (2025). Forensic DNA phenotyping: a review on SNP panels, genotyping techniques, and prediction models. Forensic Sciences Research, 10, owae013. 10.1093/fsr/owae013

Von Thaden, A., Nowak, C., Tiesmeyer, A., Reiners, T. E., Alves, P. C., Lyons, L. A., . . . Galián, J. (2020). Applying genomic data in wildlife monitoring: Development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism panels. Molecular ecology resources, 20, 662-680.

Wang, J. (2016). The computer program Structure for assigning individuals to populations: Easy to use but easier to misuse. Molecular Ecology Resources, https://doi.org/10.1111/1755-0998.12650.

Wehrenberg, G., Tokarska, M., Cocchiararo, B. & Nowak, C. (2024). A reduced SNP panel optimised for non-invasive genetic assessment of a genetically impoverished conservation icon, the European bison. Scientific Reports, 14, 1875. 10.1038/s41598-024-51495-9

Zenato-Lazzari, G., Figueiró, H. V., Sartor, C. C., Donadio, E., Di Martino, S., Draheim, H. M. & Eizirik, E. (2025). Development of a SNP Panel for Geographic Assignment and Population Monitoring of Jaguars (Panthera onca). Ecology and Evolution, 15, e71465. https://doi.org/10.1002/ece3.71465

Ende