Analysing Genomic Data with **dartRverse**: Accessible Tools for Conservation

ICCB JUNE 15-20 2025

UNIVERSITY OF CANBERRA

DArT diversity arrays technology

MONASH University

Centre for Biodiversity Analysis

dartR
Population assignment

Assignment of an individual of unknown provenance to a source population ……

Applications?

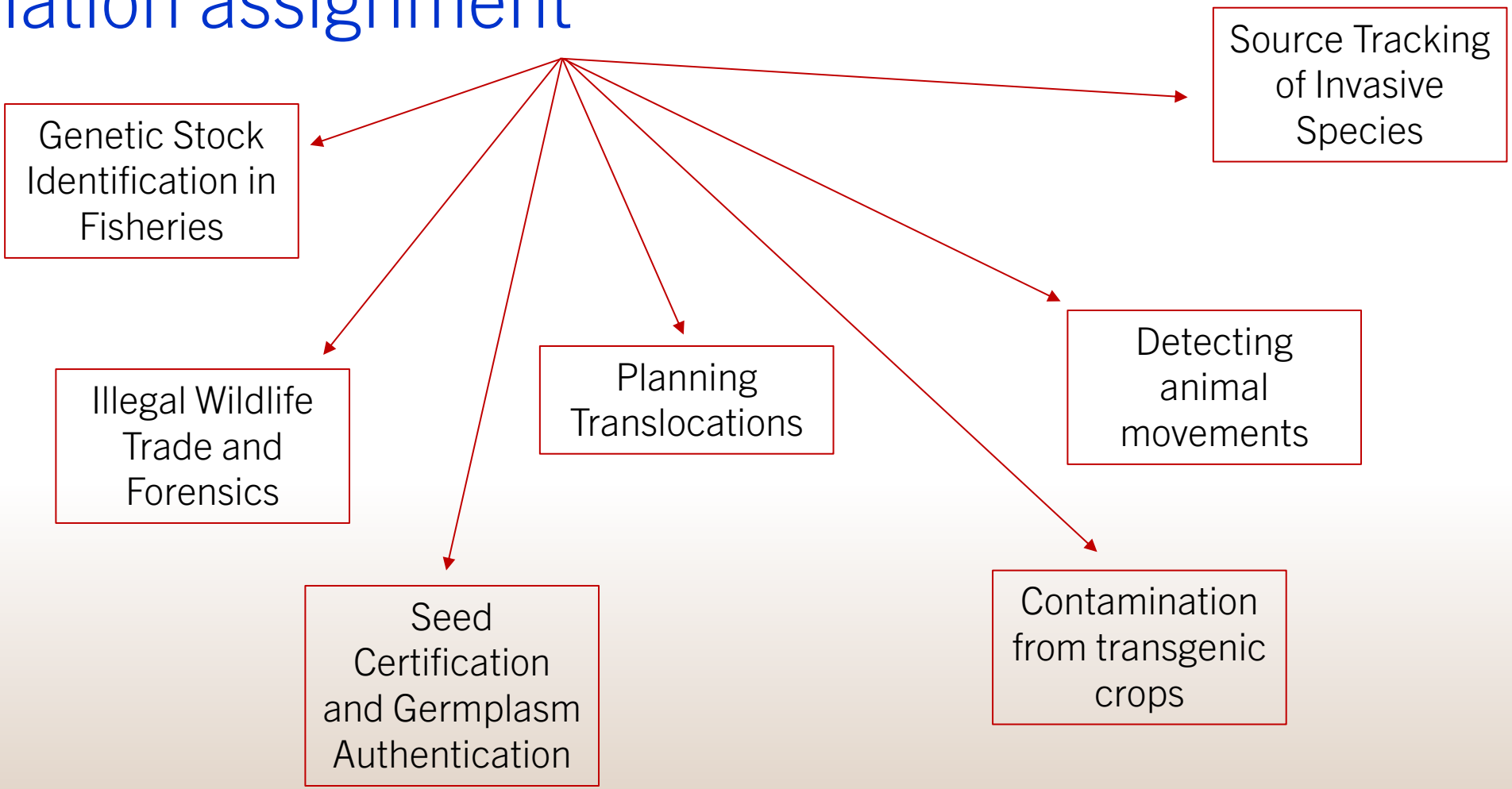
Bernd Gruber

Elise Furlan

Arthur Georges

UNIVERSITY OF CANBERRA

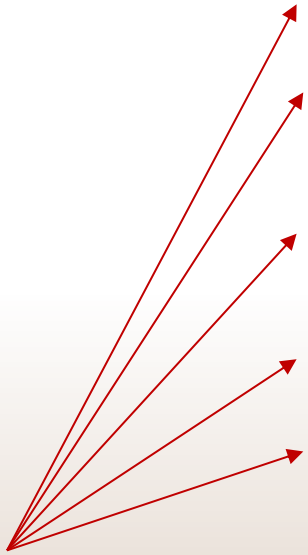
dartR
Population assignment



Anderson et al. (2008). *Canadian Journal of Fisheries and Aquatic Sciences* 65:1475–1486
Beacham et al. (2005). *Canadian Journal of Fisheries and Aquatic Sciences* 62:2425–2434
Wasser, S. K., et al. (2015). *Science* 349:84–87

dartR
3rd Party Software

Software	Approach	Input	Usage
STRUCTURE	Bayesian	SNPs/microsats	Structure, admixture
GENECLASS2	Bayesian & freq-based	SNPs/microsats	Migrant detection, forensics
assignPOP	Machine learning	SNPs	High-accuracy assignment
rubias	Bayesian, SNP-specific	SNPs	Fish stock assignment
ONCOR	Maximum likelihood	SNPs/microsats	Stock ID in fisheries



dartR

e.g. gl2faststructure()

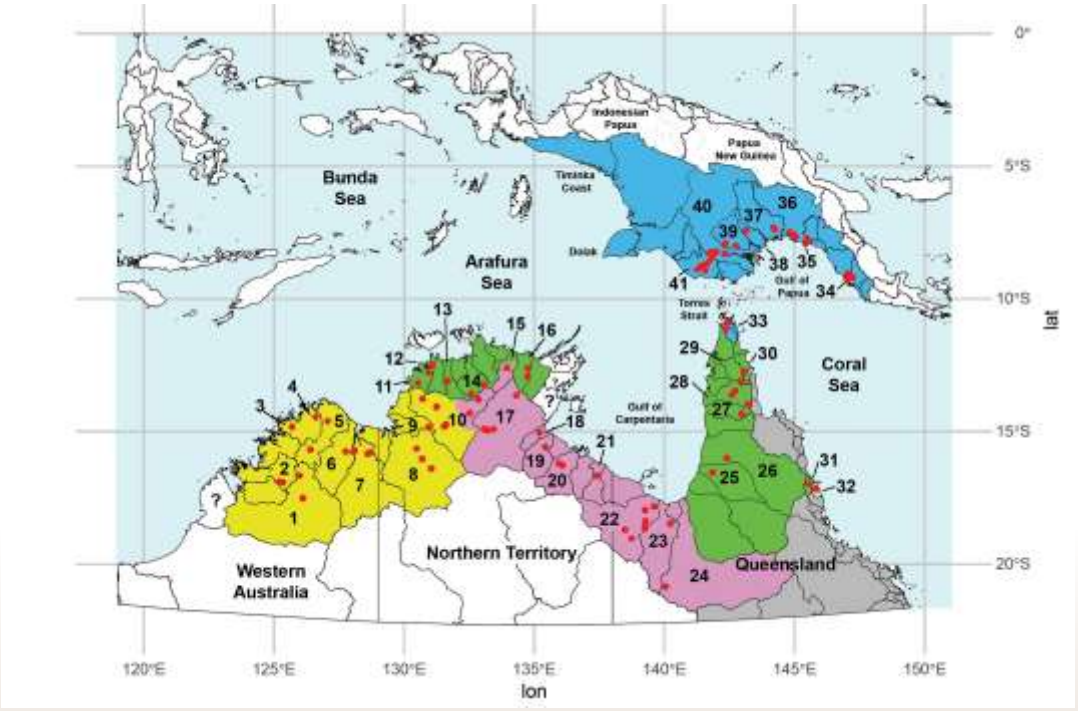
dartR
Population assignment

dartR Assignment Scripts – Exploratory Analysis

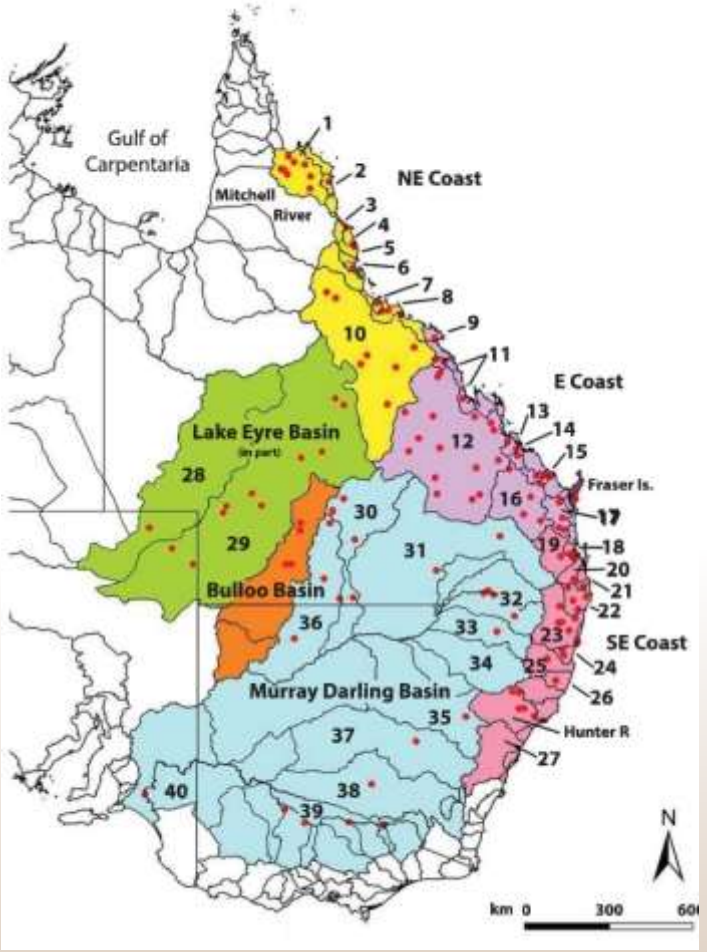
- **Genotype Likelihood:** The likelihood of drawing the unknown from a population with the observed allele frequencies is calculated assuming Hardy-Weinberg equilibrium.
- **Private Alleles:** A focal unknown individual is likely to have fewer private alleles in comparison with its source population 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.

Example dataset

Emydura river turtles



Georges et al. 2025, in review



Georges et al. 2018, Molecular Ecology

Example dataset

Emydura river turtles

- Read the data into dartR

```
setwd(<directory path> # Your working directory
gl.set.verbosity(3)    # Globally set verbosity
gl <- readRDS("assignment.example1.Rdata")
```

- Examine the contents

```
gl  #How many individuals, how many loci?
nLoc(gl)
nInd(gl)
nPop(gl)
```

- Tabulate the populations

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

Example dataset

Workflow

```
gl.set.verbosity(3)
gl <-
readRDS("assignment.example1.Rdata")
gl
nLoc(gl)
nInd(gl)
nPop(gl)
table(pop(gl))
```

- Tabulate the populations

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

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

NOTE: Some population sizes less than the recommended minimum of 10

Example Analysis -- likelihood

Workflow

```
gl.set.verbosity(3)
gl <-
readRDS("assignment.example1.Rdata")
gl
nLoc(gl)
nInd(gl)
nPop(gl)
table(pop(gl))
gen.result<-
gl.assign.on.genotype(gl,
  unknown="AA011731", nmin=10)
```

- Assign based on genotype likelihood

```
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

.....

On the mark!!! But with a caveat ……

Example Analysis – Private Alleles

Workflow

```
gl.set.verbosity(3)
gl <-
readRDS("assignment.example1.Rdata")
gl
nLoc(gl)
nInd(gl)
nPop(gl)
table(pop(gl))
gen.result<-gl.assign.on.genotype(gl,
    unknown="AA011731", nmin=10)
pa.result <- gl.assign.pa(gl,
  unknown="AA011731", nmin=10,
  alpha=0.05)
```

- Assign based on Private Alleles

```
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
```

Private alleles are alleles possessed by the focal unknown that are absent from a putative source population.

Three putative sources in adjacent drainages ……

Example Analysis -- PCA

Workflow

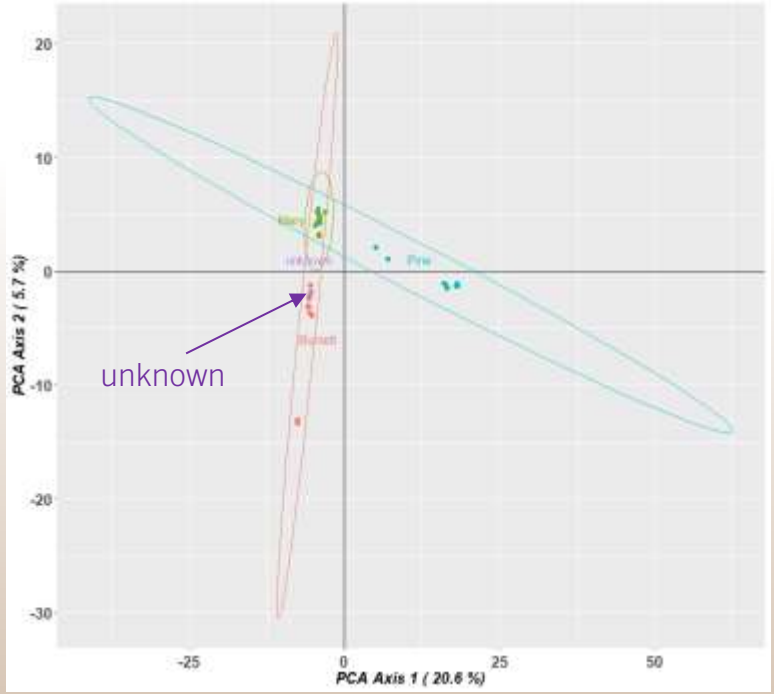
```
gl.set.verbosity(3)
gl <-
readRDS("assignment.example1.Rdata")
gl
nLoc(gl)
nInd(gl)
nPop(gl)
table(pop(gl))
gen.result<-gl.assign.on.genotype(gl,
    unknown="AA011731", nmin=10)
pa.result <- gl.assign.pa(gl,
  unknown="AA011731", nmin=10,
  alpha=0.05)
pca_pa_result <-
  gl.assign.pca(pa.result,
  unknown="AA011731")
```

- Assign based on PCA

Note

```
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 outsidetheir 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
```

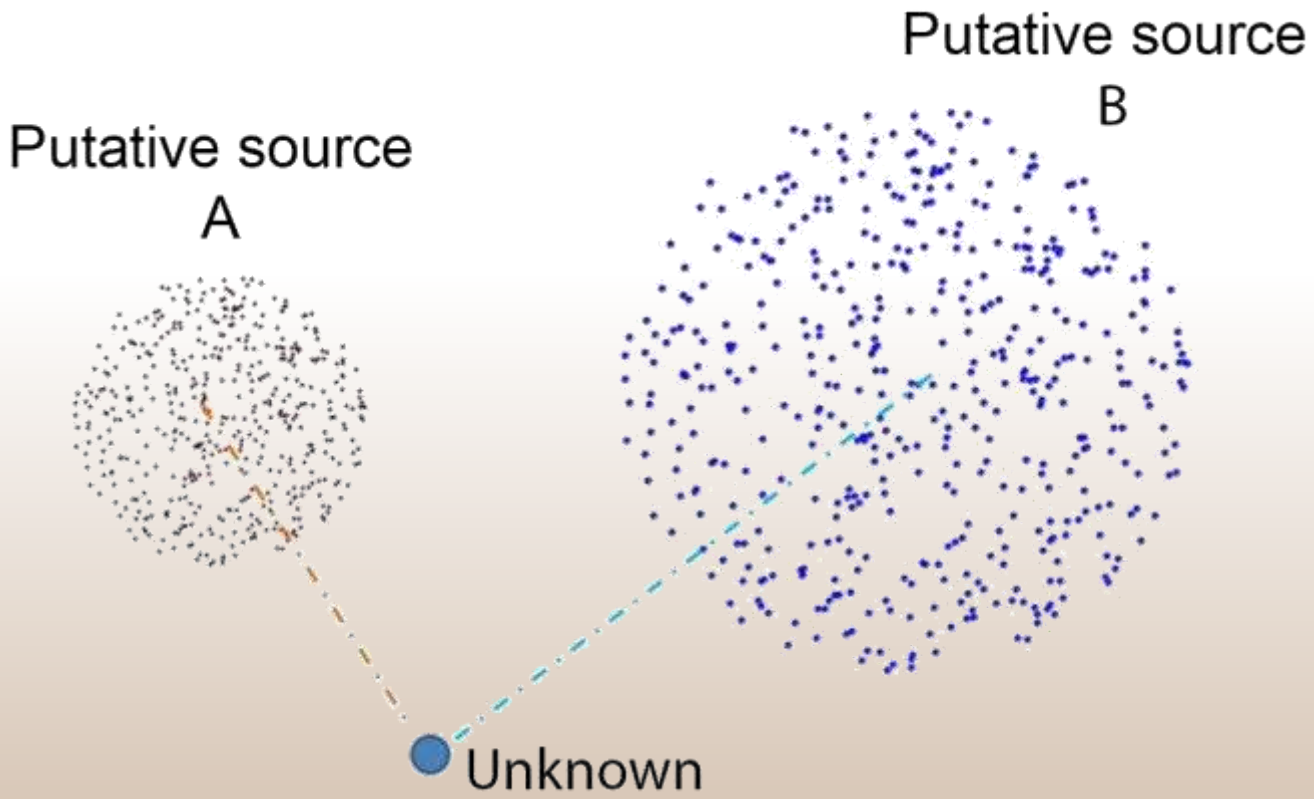


Let's see if we can improve on the Private Alleles approach with PCA?

Again, narrowed down to the Burnett ……

Example Analysis -- Mahalanobis

- Ordinated space in *k* dimensions (selected using PCA and the broken-stick criterion)
- Standardized – units along the principal components expressed in standard deviations. Confidence ellipses become confidence spheres.
- Mahalanobis Distance is a Z-score – is *p* less than alpha or greater than alpha (alpha = 0.01, say)?



Workflow

```
gl.set.verbosity(3)
gl <-
readRDS("assignment.example1.Rdata")
gl
nLoc(gl)
nInd(gl)
nPop(gl)
table(pop(gl))
gen.result<-gl.assign.on.genotype(gl,
    unknown="AA011731", nmin=10)
pa.result <- gl.assign.pa(gl,
  unknown="AA011731", nmin=10,
  alpha=0.05)
pca_pa_result <-
  gl.assign.pca(pa.result,
  unknown="AA011731")
```

Example Analysis -- Mahalanobis

- Assign based on Mahalanobis Distance

Note

```
mahal_result <- gl.assign.mahalanobis(pa.result,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
Best assignment is the population with the largest probability
                of assignment, in this case Burnett
  Returning a dataframe with the Mahalanobis Distances
Completed: gl.assign.mahalanobis
```

p = 0.0550 > 0.001

Workflow

```
gl.set.verbosity(3)
gl <-
readRDS("assignment.example1.Rdata")
gl
nLoc(gl)
nInd(gl)
nPop(gl)
table(pop(gl))
gen.result<-gl.assign.on.genotype(gl,
    unknown="AA011731", nmin=10)
pa.result <- gl.assign.pa(gl,
  unknown="AA011731", nmin=10,
  alpha=0.05)
pca_pa_result <-
  gl.assign.pca(pa.result,
  unknown="AA011731")
mahal_result <-
  gl.assign.mahalanobis(pa.result,
  unknown="AA011731")
```

For computational reasons, let's restrict the application of Mahalanobis Distance to the outcomes of the Private alleles approach?

Again, narrowed down to the Burnett ……

Exercise – Wildlife Forensics

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".

Before you begin the analysis, restrict the populations under consideration to *Emydura subglobosa*.

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

Workflow

```
gl.set.verbosity(3)
gl <-
readRDS("assignment.example1.Rdata")
gl
nLoc(gl)
nInd(gl)
nPop(gl)
table(pop(gl))
gen.result<-gl.assign.on.genotype(gl,
    unknown="AA011731", nmin=10)
pa.result <- gl.assign.pa(gl,
  unknown="AA011731", nmin=10,
  alpha=0.05)
pca_pa_result <-
  gl.assign.pca(pa.result,
  unknown="AA011731")
mahal_result <-
  gl.assign.mahalanobis(pa.result,
  unknown="AA011731")
```

Exercise

```
assignment_example1.Rdata
Unknown = "AA046092"

popNames(gl)

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

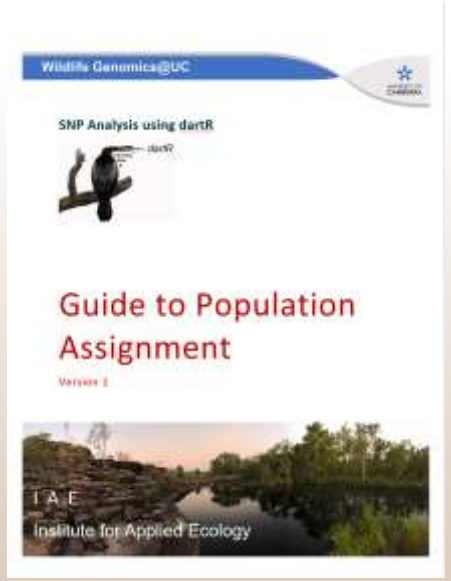
Workflow

```
gl.set.verbosity(3)
gl <-
readRDS("assignment.example1.Rdata")
gl
nLoc(gl)
nInd(gl)
nPop(gl)
table(pop(gl))
gen.result<-gl.assign.on.genotype(gl,
    unknown="AA011731", nmin=10)
pa.result <- gl.assign.pa(gl,
  unknown="AA011731", nmin=10,
  alpha=0.05)
pca_pa_result <-
  gl.assign.pca(pa.result,
  unknown="AA011731")
mahal_result <-
  gl.assign.mahalanobis(pa.result,
  unknown="AA011731")
```

Where have we come?

This 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 four techniques – allele frequency, private alleles, PCA and Mahalanobis Distance.
- Be able to sensibly integrate the results of three approaches in coming to a decision



http://georges.biomatix.org/dartR

Discussion and Questions

Workflow

```
gl.set.verbosity(3)
gl <-
readRDS("assignment.example1.Rdata")
gl
nLoc(gl)
nInd(gl)
nPop(gl)
table(pop(gl))
gen.result<-gl.assign.on.genotype(gl,
    unknown="AA011731", nmin=10)
pa.result <- gl.assign.pa(gl,
  unknown="AA011731", nmin=10,
  alpha=0.05)
pca_pa_result <-
  gl.assign.pca(pa.result,
  unknown="AA011731")
mahal_result <-
  gl.assign.mahalanobis(pa.result,
  unknown="AA011731")
```

Exercise

```
assignment_example1.Rdata
Unknown = "AA046092"

popNames(gl)

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