R-evolutionary insights

Deep dive into SNP-based population genomics

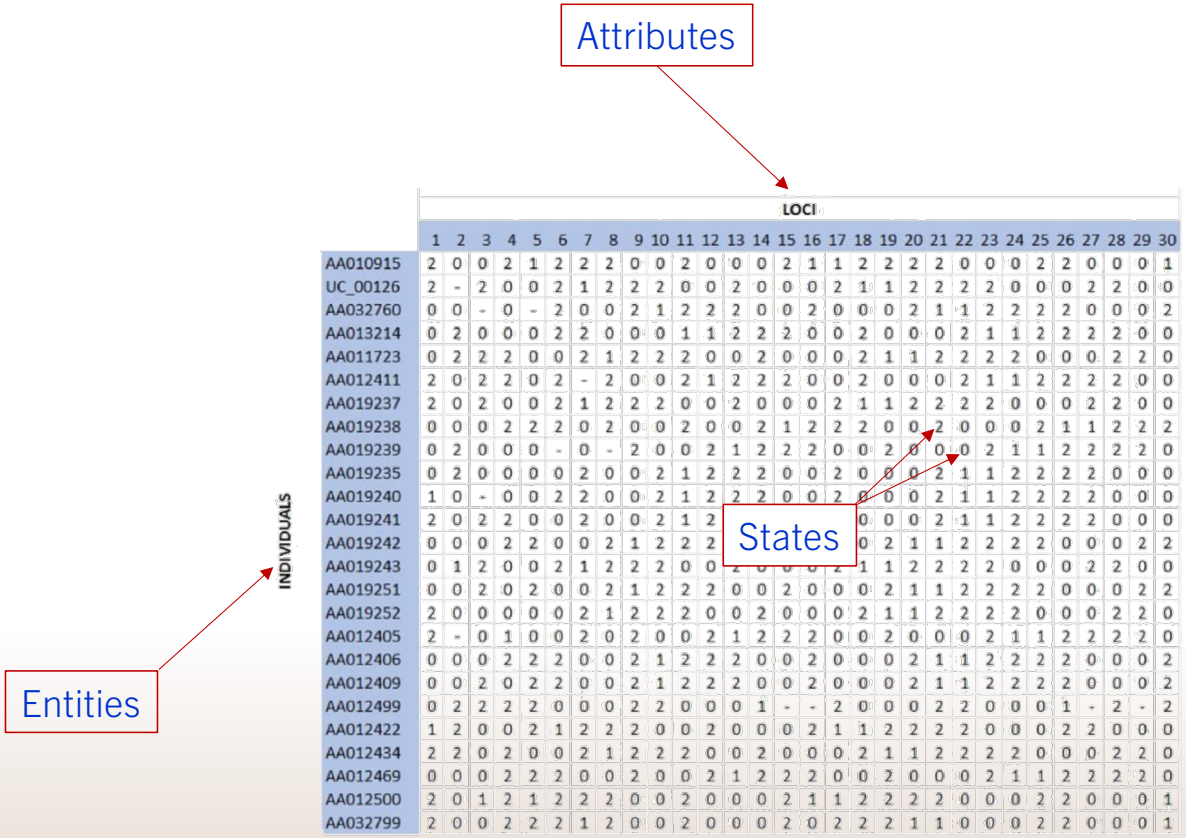
SESSION 12

Genetic Structure

- Luis Mijangos – Introduction [10 mins]
- Arthur Georges – Genetic Distance and Visualization [40 mins]
- Bernd Gruber – Structure and FastStructure [40 mins]
- Bill Sherwin – New developments [20 mins]
- Arthur Georges – Discussion and Close [10 mins]

Distances and their visualization

- Genetic Distance
- Visualization



How can we define a notion of similarity, or conversely, a notion of dissimilarity or genetic distance?

Distances and their visualization

- Genetic Distance
- Visualization

What is a well-behaved distance?

Metric

$d(AB) \geq 0$
$d(AA) = 0$
$d(AB) = d(BA)$
$d(AB) < d(AC) + d(BC)$

An absolute, independent of direction

Euclidean

Shortest distance between two points is a line.

Distances and their visualization

- Genetic Distance
- Visualization

What is a well-behaved distance?

Metric

$d(AB) \geq 0$
$d(AA) = 0$
$d(AB) = d(BA)$
$d(AB) < d(AC) + d(BC)$

If you have to go through a third point to get there, it is on that line.

Euclidean

Distances and their visualization

- Genetic Distance
- Visualization

What is a well-behaved distance?

Metric

$d(AB) \geq 0$
$d(AA) = 0$
$d(AB) = d(BA)$
$d(AB) < d(AC) + d(BC)$

Given three distances between A, B and C, the positions of all three are uniquely defined.

Euclidean

Distances can be uniquely represented in a space defined by Cartesian Coordinates

Private alleles?

Nei's distance?

Rogers D?

Euclidean D?

Distances and their visualization

- Genetic Distance
- Visualization
- Population Structure

What is a well-behaved distance?

Metric

$d(AB) \geq 0$
$d(AA) = 0$
$d(AB) = d(BA)$
$d(AB) < d(AC) + d(BC)$

Euclidean

Distances can be uniquely represented in a space defined by Cartesian Coordinates

Do not overthink

Modest departures have no appreciable impact

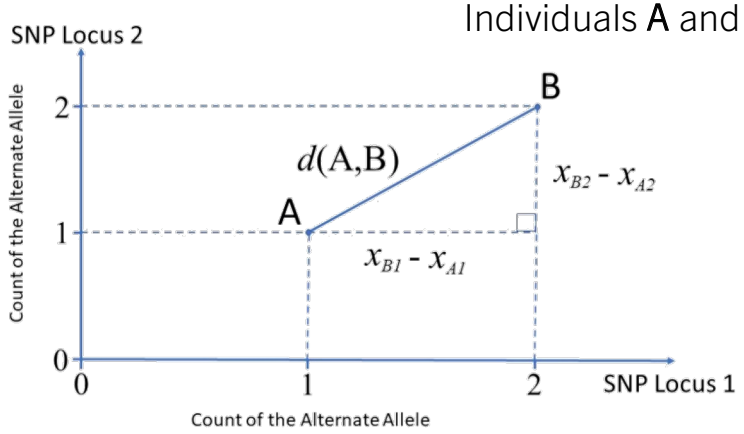
Distances and their visualization

- Genetic Distance
- Visualization

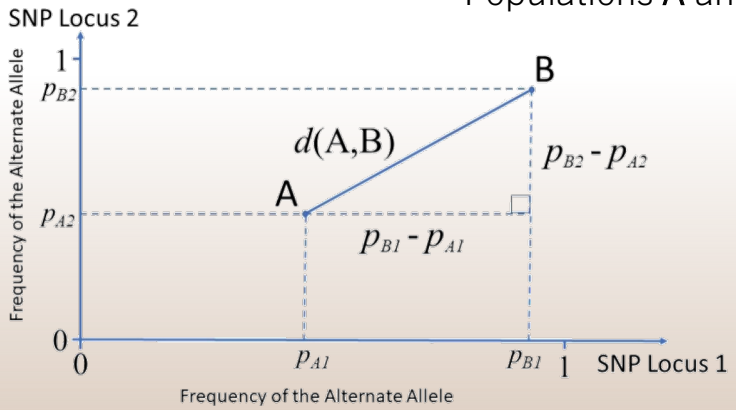
How can we define a simple measure dissimilarity or genetic distance for SNPS?

Standard Euclidean Distance

Individuals **A** and **B**



Populations **A** and **B**



Distances and their visualization

- Genetic Distance
- Visualization

Genetic Distance and dartR

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 112
table(pop(gl)
 signata    farnorth    krefftii       nigra  gunabarra    emmottii
      11          10          10          10          10          10
 macquarii subglobosa    worrelli tanybaraga  victoriae
        10          10           9          12          10
D <- gl.dist.ind(gl)
```

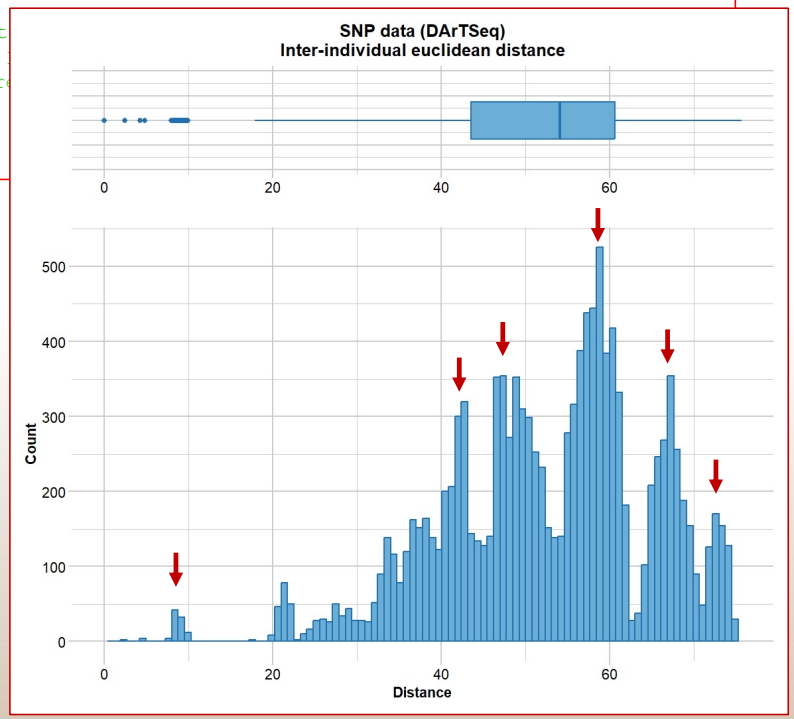
Follow along if you like. Maybe use your own genlight object?

Distances and their visualization

- Genetic Distance
- Visualization

Genetic Distance

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 113
table(pop(gl)
 signata    farnorth    krefft
      11          10
 macquarii subglobosa    worr
      10          10
D <- gl.dist.ind(gl)
```



SNPs :

- Euclidean Distance
- Scaled Euclidean Distance
- Simple Mismatch Distance
- Absolute Mismatch Distance
- Czekanowski (Manhattan) Distance

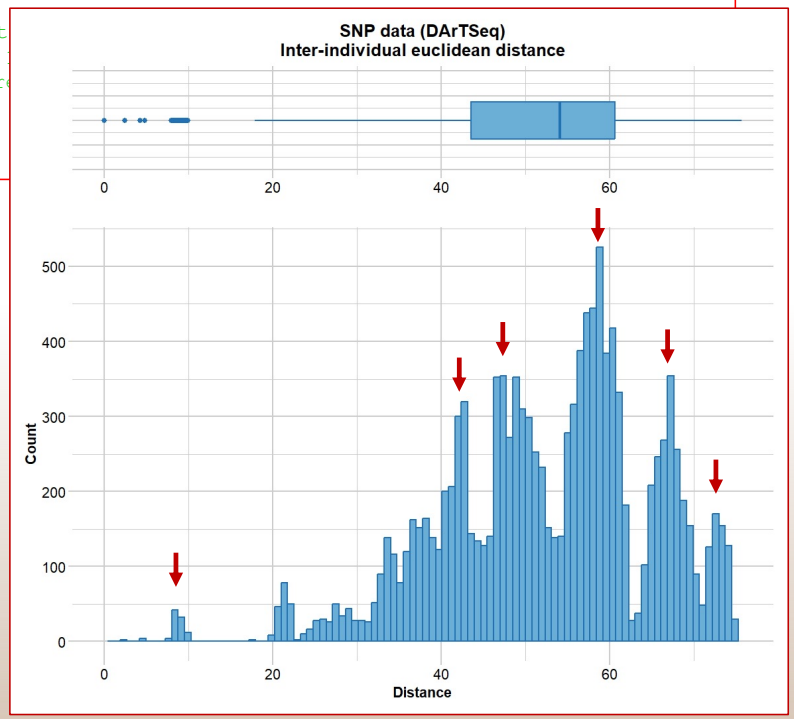
Refer to https://doi.org/10.1101/2023.03.22.533737

Distances and their visualization

- Genetic Distance
- Visualization

Genetic Distance

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 113
table(pop(gl))
 signata    farnorth    krefft
      11          10
 macquarii subglobosa    worr
      10          10
D <- gl.dist.ind(gl)
```



SilicoDArT :

- Euclidean Distance
- Scaled Euclidean Distance
- Simple Matching Distance
- Jaccard Distance
- Bray-Curtis Distance

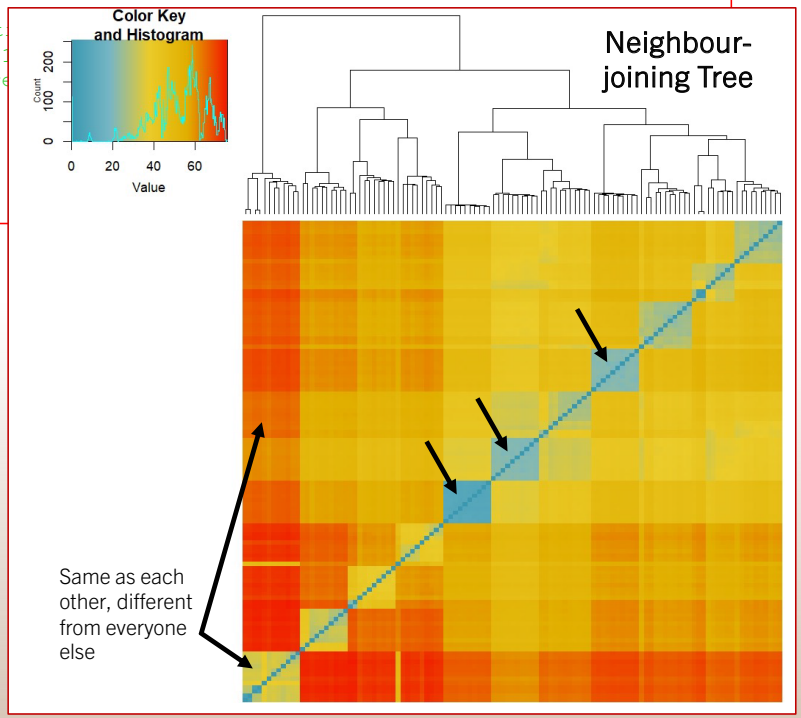
Refer to https://doi.org/10.1101/2023.03.22.533737

Distances and their visualization

- Genetic Distance
- Visualization

Genetic Distance

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 113
table(pop(gl)
 signata    farnorth    krefft
      11          10         1
macquarii subglobosa    worre
      10          10
D <- gl.dist.ind(gl)
gl.plot.heatmap(D)
```



Out of Session Activity

Maybe use your own dataset and look at a heat map with different distance measures?

Distances and their visualization

- Genetic Distance
- Visualization

PCA

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 113
table(pop(gl))
  signata    farnorth    krefftii       nigra   gunabarra    emmottii
       11          10          10          10          10          10
 macquarii subglobosa    worrelli tanybaraga   victoriae
       10          10          10          12          10
D <- gl.dist.ind(gl)
gl.plot.heatmap(D)
pc <- gl.pcoa(gl)
Starting gl.pcoa
  Processing genlight object with SNP data
  Performing a PCA, individuals as entities, loci as attributes, SNP genotype as state
  Ordination yielded 13 informative dimensions from 112 original dimensions
    PCA Axis 1 explains 19.1 % of the total variance
    PCA Axis 1 and 2 combined explain 29.4 % of the total variance
    PCA Axis 1-3 combined explain 38.8 % of the total variance
Completed: gl.pcoa
```

Kaiser-Guttman criterion

How much are you willing to throw away?

Out of Session Activity

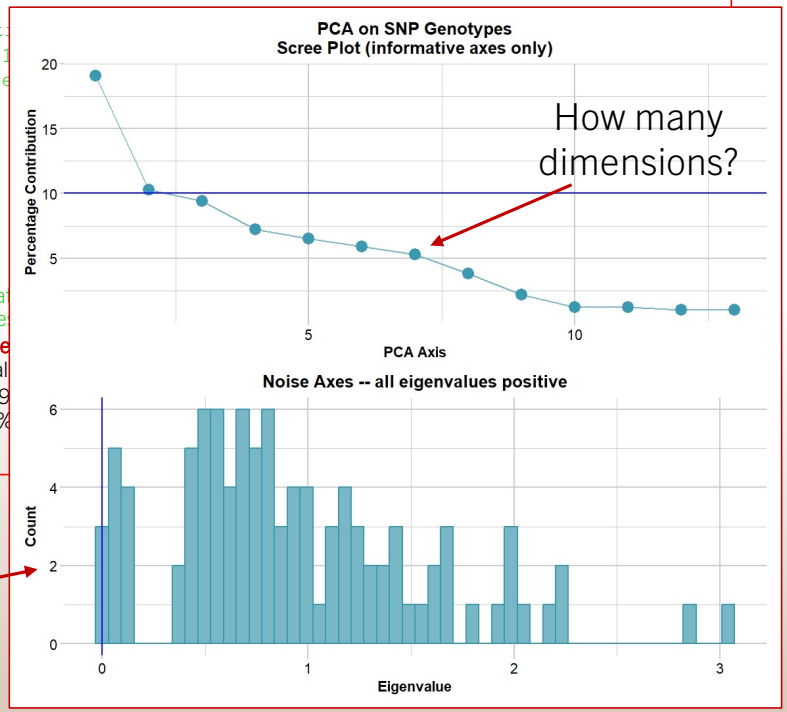
Maybe use your own dataset and look at a PCA with different distance measures?

Distances and their visualization

- Genetic Distance
- Visualization

PCA

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 113
table(pop(gl))
 signata    farnorth    krefft
      11          10         1
macquarii subglobosa    worre
      10          10
D <- gl.dist.ind(gl)
gl.plot.heatmap(D)
pc <- gl.pcoa(gl)
Starting gl.pcoa
  Processing genlight object with SNP dat
  Performing a PCA, individuals as entitie
  Ordination yielded 13 informative dime
    PCA Axis 1 explains 19.1 % of the total
    PCA Axis 1 and 2 combined explain 29
    PCA Axis 1-3 combined explain 38.8 %
Completed: gl.pcoa
```



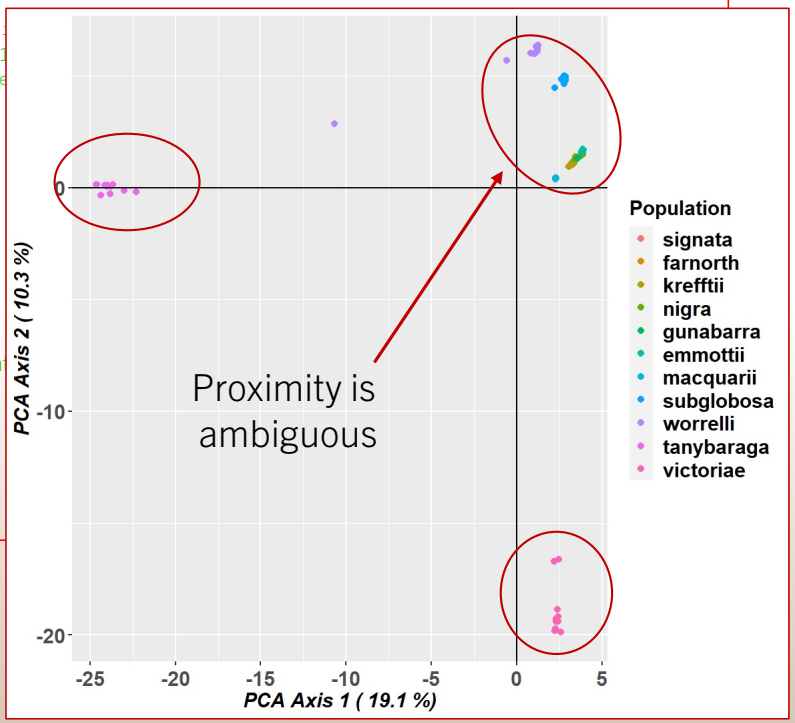
Is my distance well behaved?

Distances and their visualization

- Genetic Distance
- Visualization

PCA Plot

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 113
table(pop(gl))
 signata    farnorth    krefft
      11          10         1
macquarii subglobosa    worre
      10          10
D <- gl.dist.ind(gl)
gl.plot.heatmap(D)
pc <- gl.pcoa(gl)
gl.pcoa.plot(pc,gl)
Starting gl.pcoa.plot
  Processing an ordination file (glPca)
  Processing genlight object with SNP da
  Plotting populations in a space defined
  Preparing plot .... please wait
Completed: gl.pcoa.plot
```

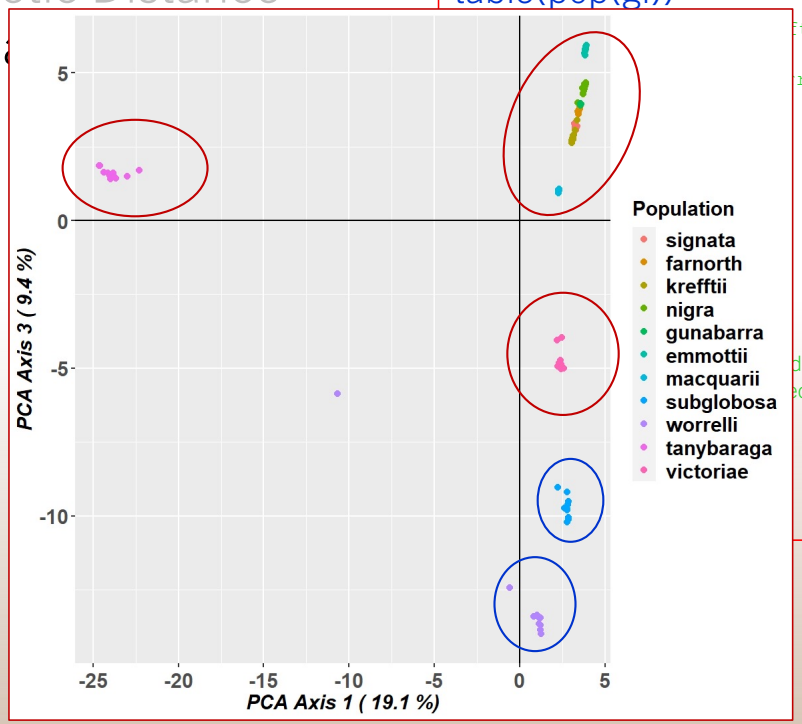


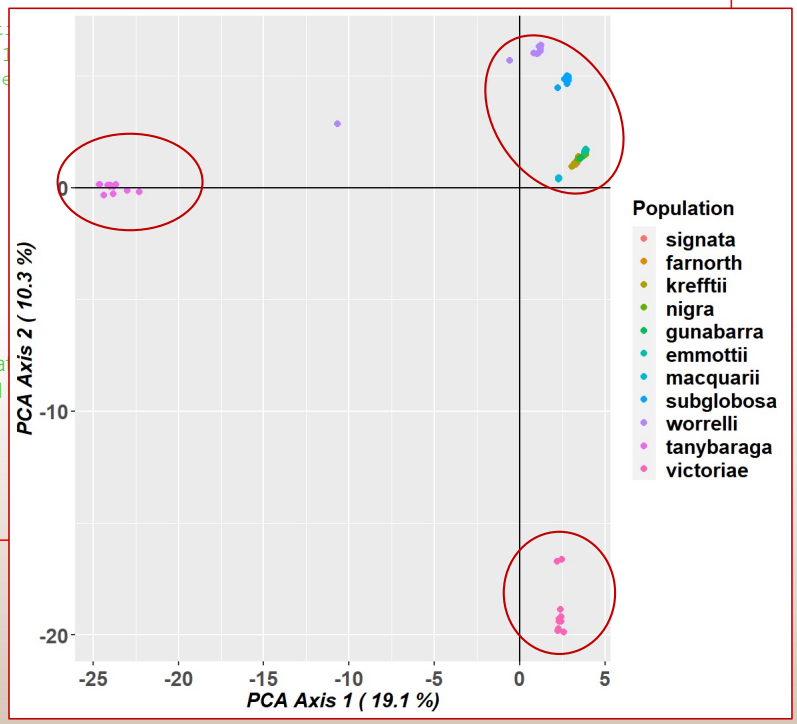
Distances and their visualization

- Genetic Distance
- Visu

PCA Plot

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 113
table(pop(gl))
```



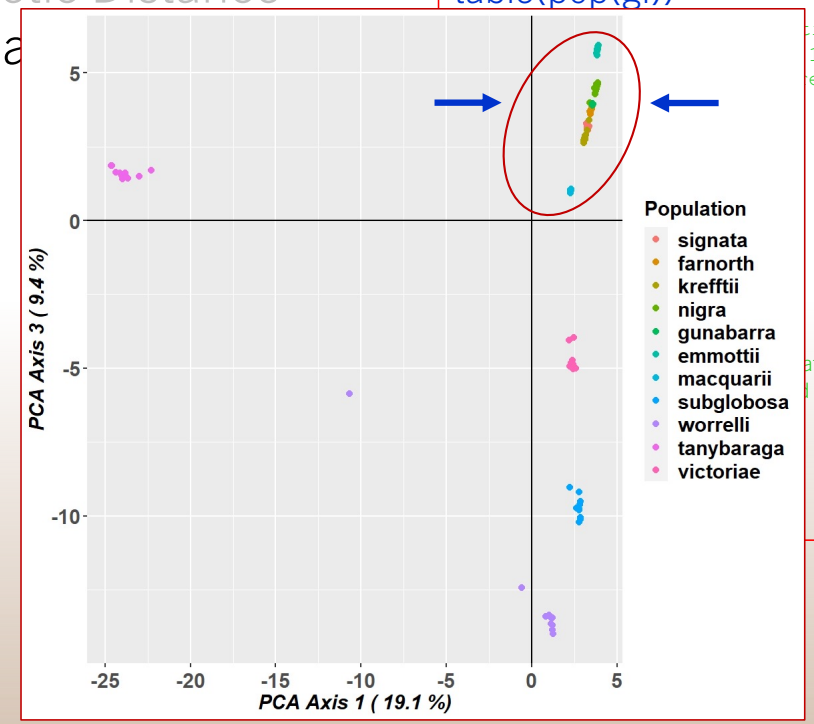


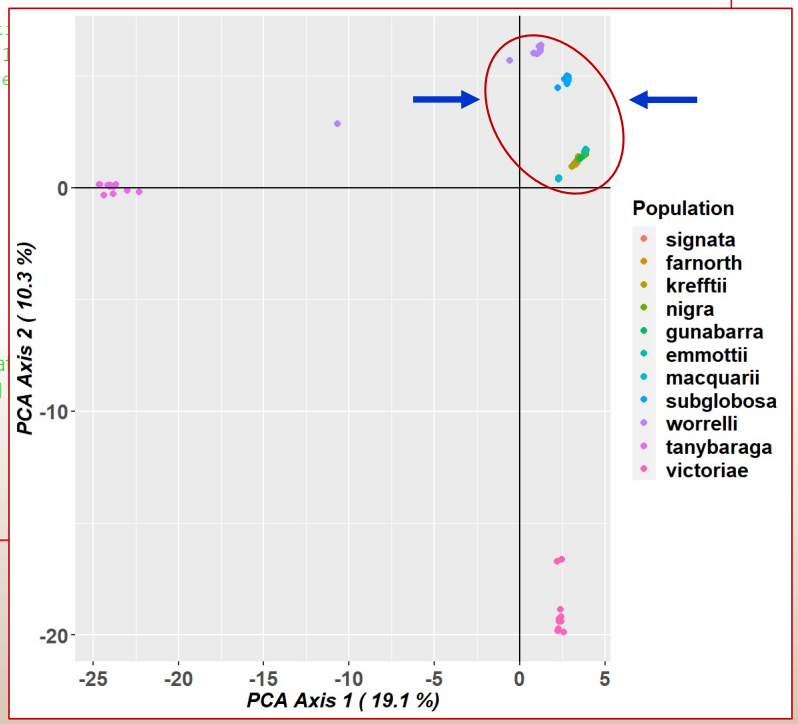
Distances and their visualization

- Genetic Distance
- Visua

PCA Plot

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 113
table(pop(gl))
```





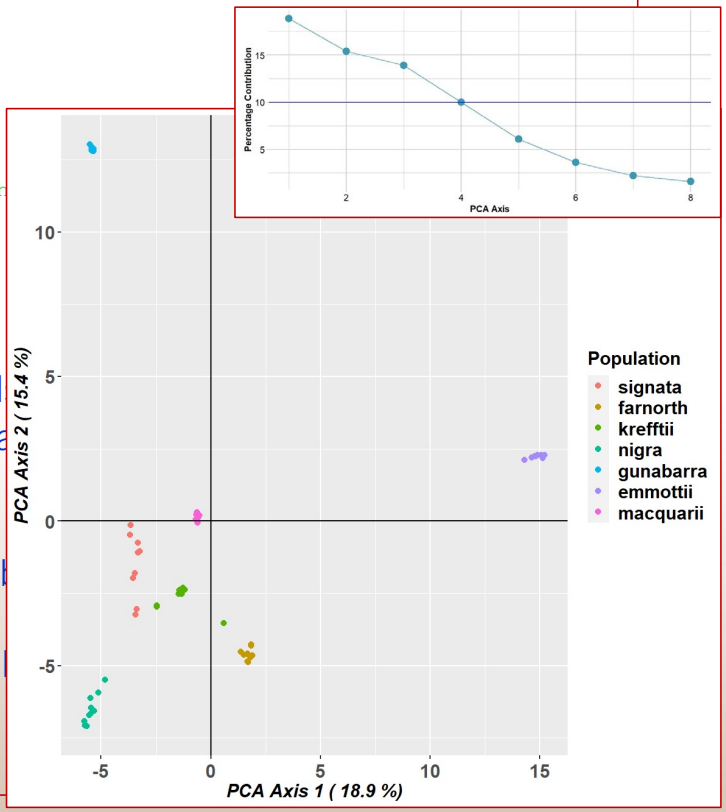
Distances and their visualization

- Genetic Distance
- Visualization

PCA Plot

```
gl <- gl.load("./data/Example_12.1_SNP.Rdata")
nLoc(gl)
[1] 10529
nInd(gl)
[1] 113
table(pop(gl))
 signata    farnorth    krefftii
      11          10          10
macquarii subglobosa    worrelli tan
      10          10          10
D <- gl.dist.ind(gl)
gl.plot.heatmap(D)
pc <- gl.pcoa(gl)
gl. pcoa.plot(pc,gl,pop.label
gl.plot.pcoa(pc,gl,xaxis=1,ya

temp <- gl.drop.pop(gl,
   pop.list=c("victoriae","sul
pc <- gl.pcoa(temp)
gl.pcoa.plot(pc,temp,pop.la
```



Distances and their visualization

Herd of Elephants in the Room

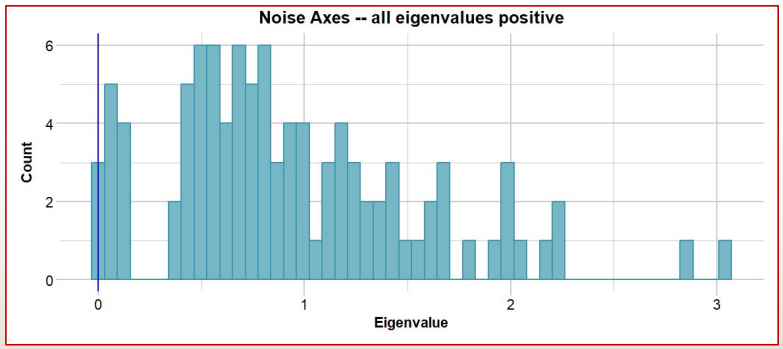
No. 1: Missing Data
Compromises Metric/Euclidean properties

- Genetic Distance
- Visualization

Distortion

↓

Misinterpretation



Keep your eye on the PCA diagnostics

Distances and their visualization

Herd of Elephants in the Room

No. 2: Missing Data
PCA does not like missing values

- Genetic Distance
- Visualization

Either delete all loci with one or more missing values

or

Delete all individuals with one or more missing values

Most algorithms (e.g. gIPCA) fill missing values with the global average

↓

Distortion

↓

Misinterpretation

Distances and their visualization

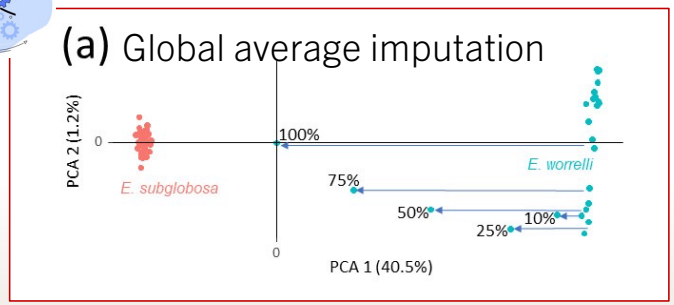
Herd of Elephants in the Room

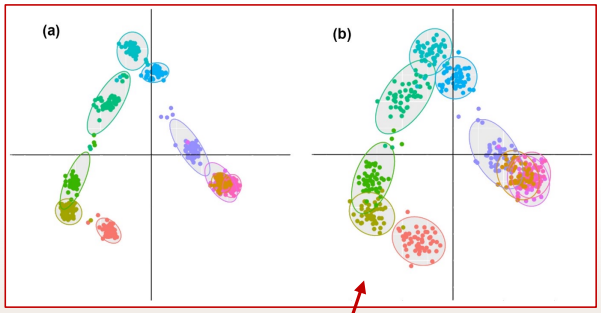
No. 2: Missing Data
PCA does not like missing values

- Genetic Distance
- Visualization

Be a bit more stringent if filtering on Call Rate
dartr::gl.filter.callrate()

More appropriate imputation
dartr::gl.impute()





Particularly critical if using distances for population assignment

Distances and their visualization

Herd of Elephants in the Room

No. 2: Missing Data
Worried about imputation?

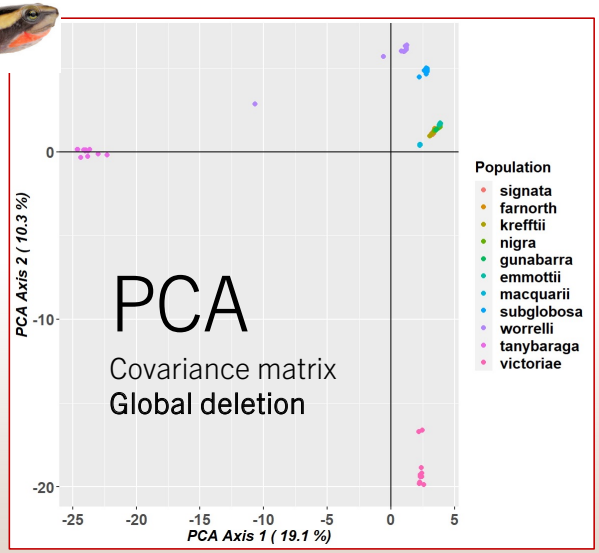
- Genetic Distance
- Visualization

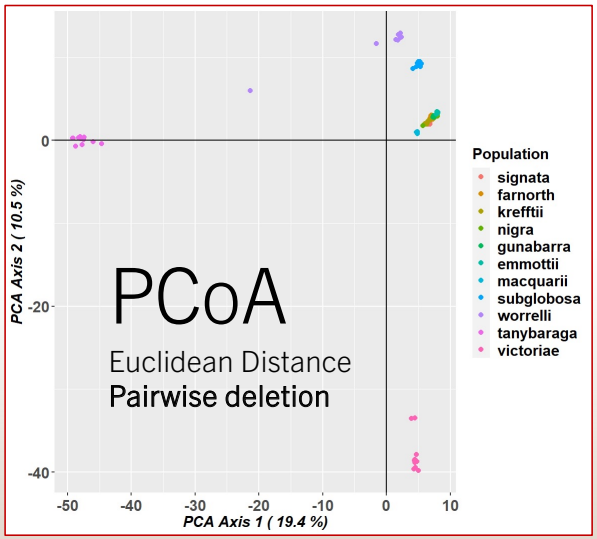
PCA

```
pc <- gl.pcoa(gl)
gl.pcoa.plot(pc,gl,pop.labels="legend")
```

PCoA

```
D <- gl.dist.ind(gl)
pco <- gl.pcoa(D)
gl.pcoa.plot(pco,gl,pop.labels="legend")
```





Concordance not usually THIS good – Example dataset

Distances and their visualization

Herd of Elephants in the Room

No. 3: Structural Variants (cryptic and not-so-cryptic)

- Genetic Distance
- Visualization

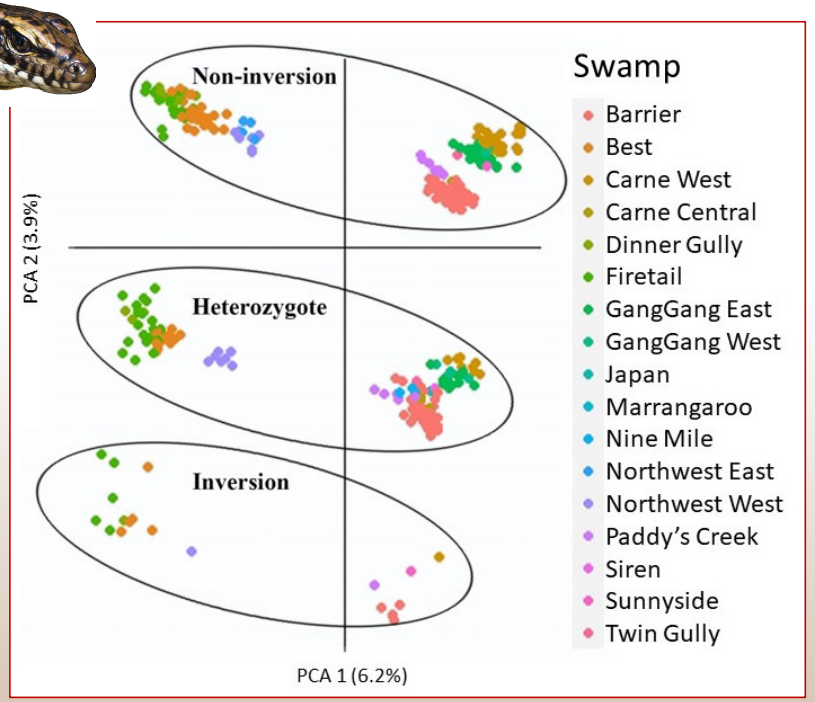
  - Batch Effects
  - Sex-linked Haploblocks
  - Large PolymorphicStructural Variants

↓

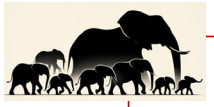
Distortion

↓

Confounded interpretation



Distances and their visualization



Herd of Elephants in the Room

No. 4: Kinship
(closely related individuals)

- Genetic Distance
- Visualization

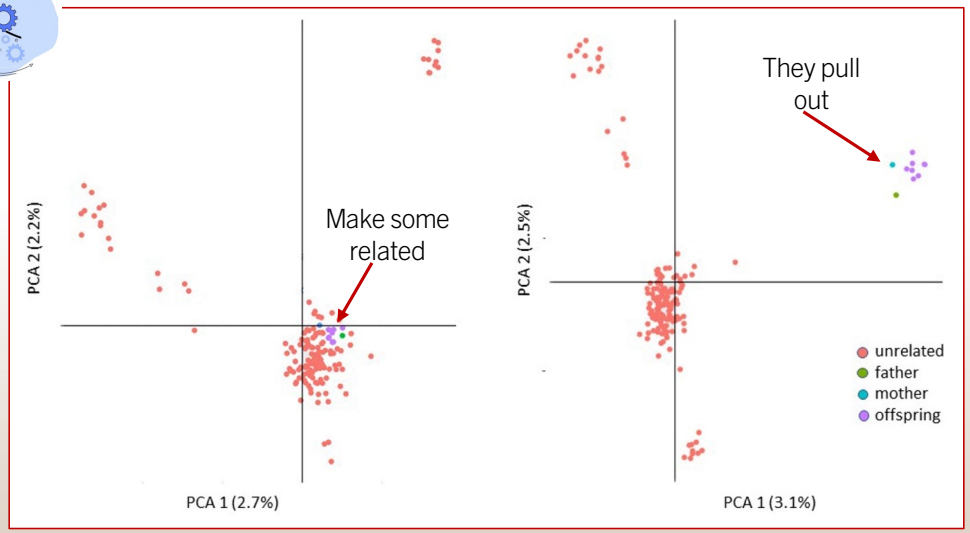
High relatedness among some individuals (kin) against a background of lower relatedness

↓

Distortion

↓

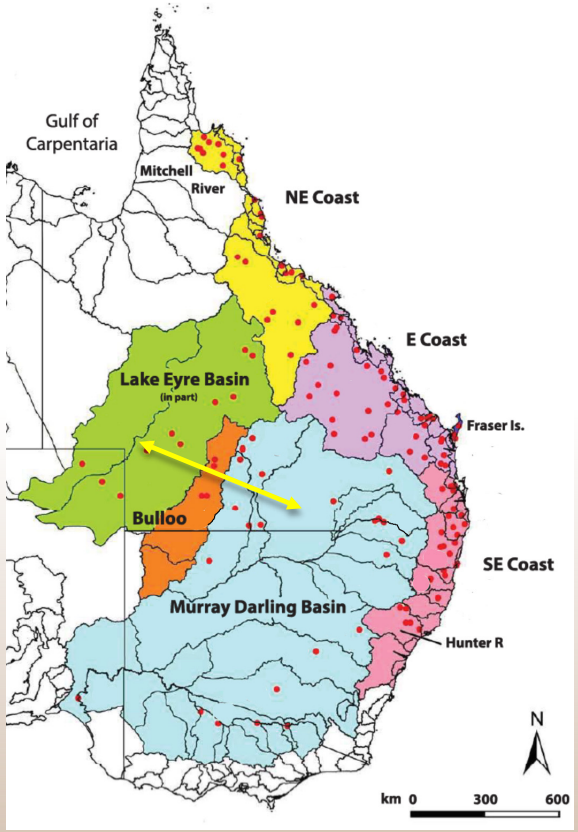
Misinterpretation

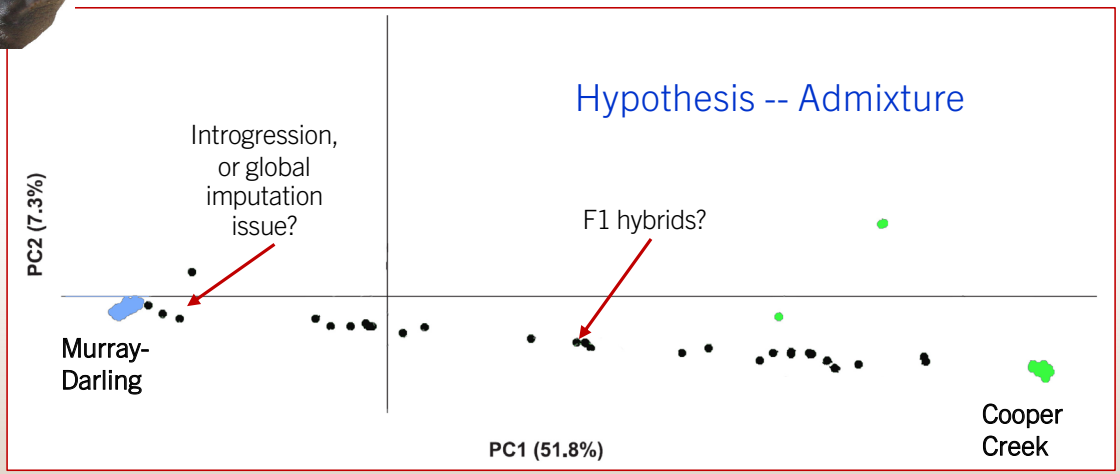


Distances and their visualization

Herd of Elephants in the Room

No. 5: Not a Panacea (Hypothesis generation)

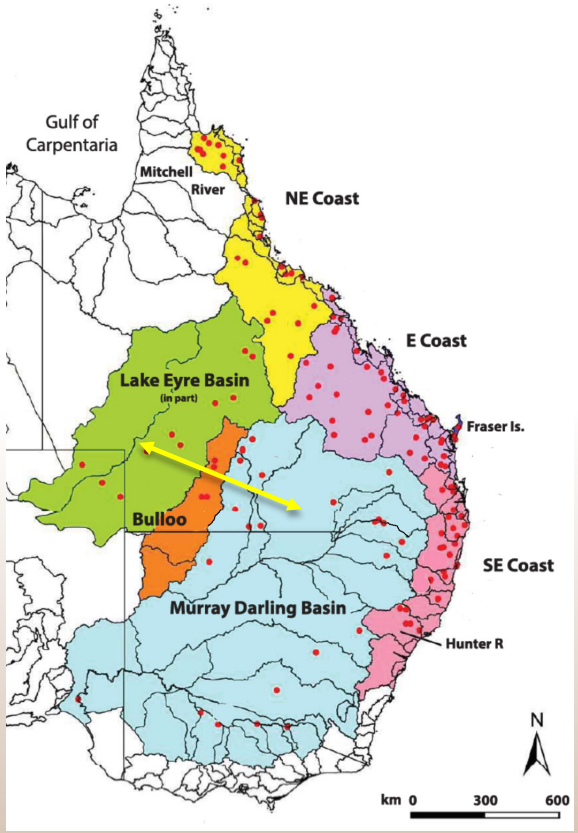


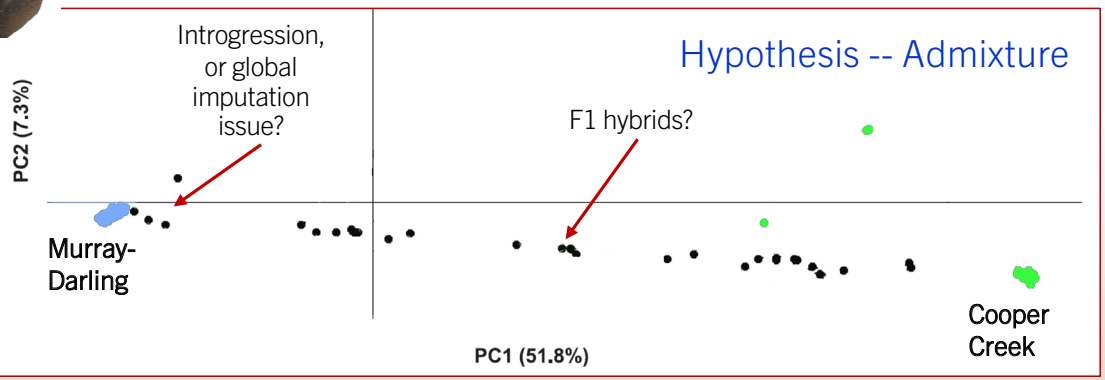


Distances and their visualization

Herd of Elephants in the Room

No. 5: Not a Panacea (Hypothesis generation)

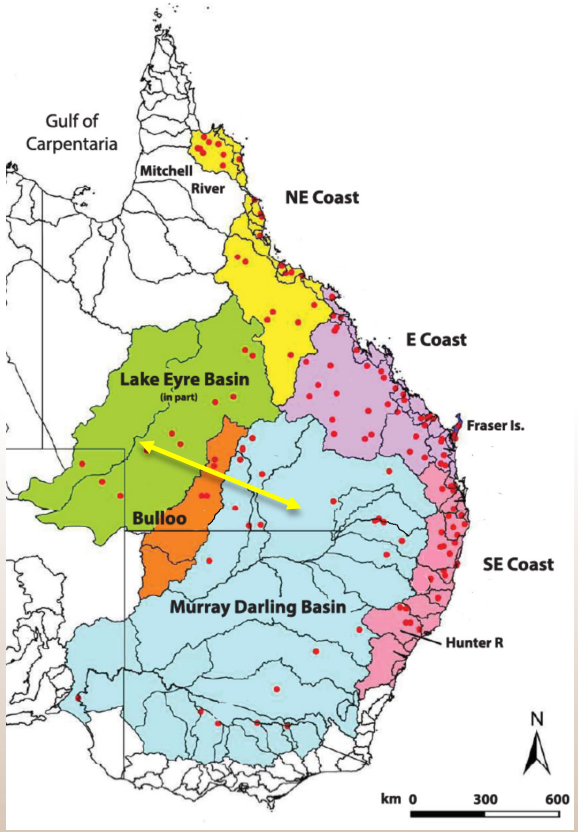


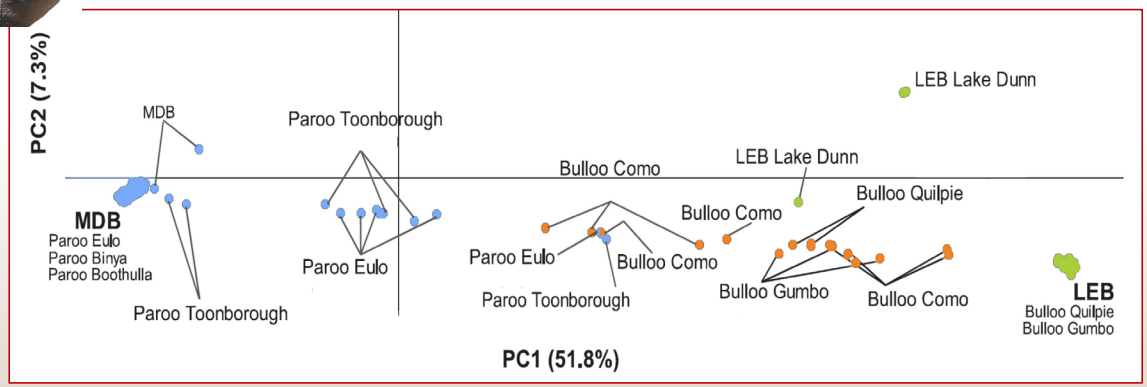


Distances and their visualization

Herd of Elephants in the Room

No. 5: Not a Panacea
(Hypothesis generation)

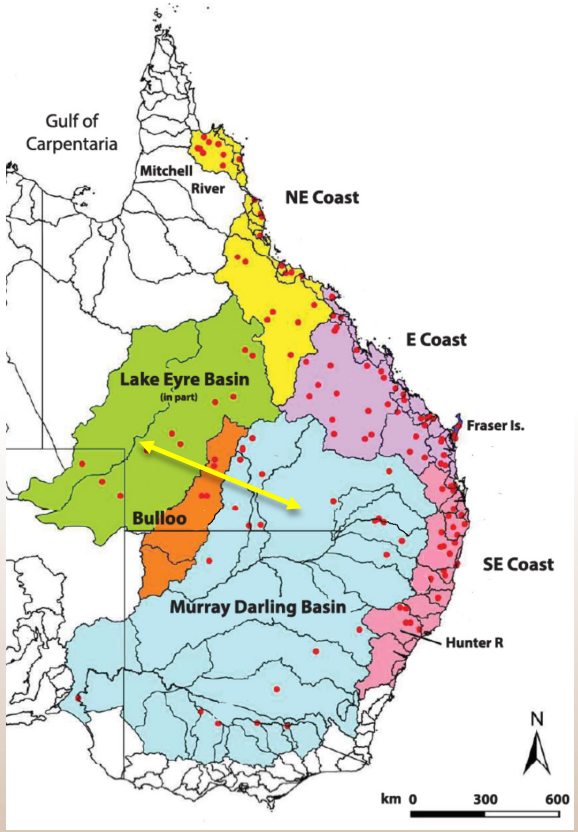




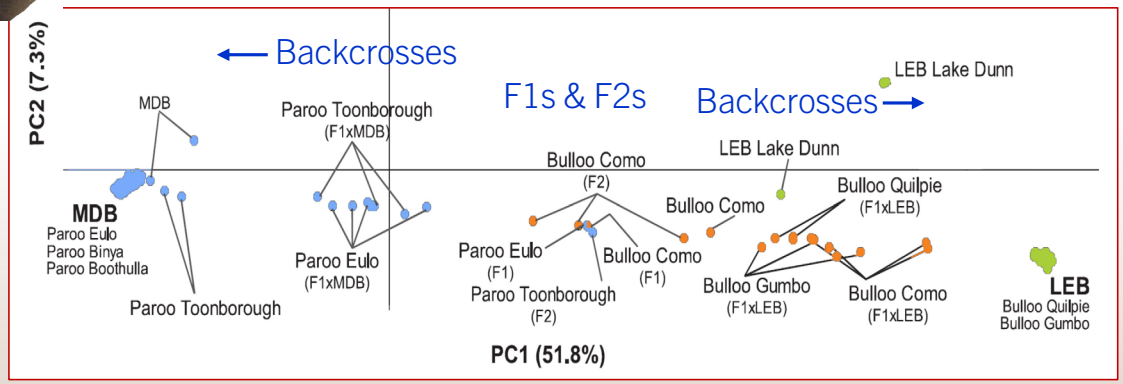
Distances and their visualization

Herd of Elephants in the Room

No. 5: Not a Panacea (for hypothesis generation)



New Hybrids -- gl.newhyb()



PCA hypothesis → New Hybrids definitive analysis

Conclusion
Where have we come?

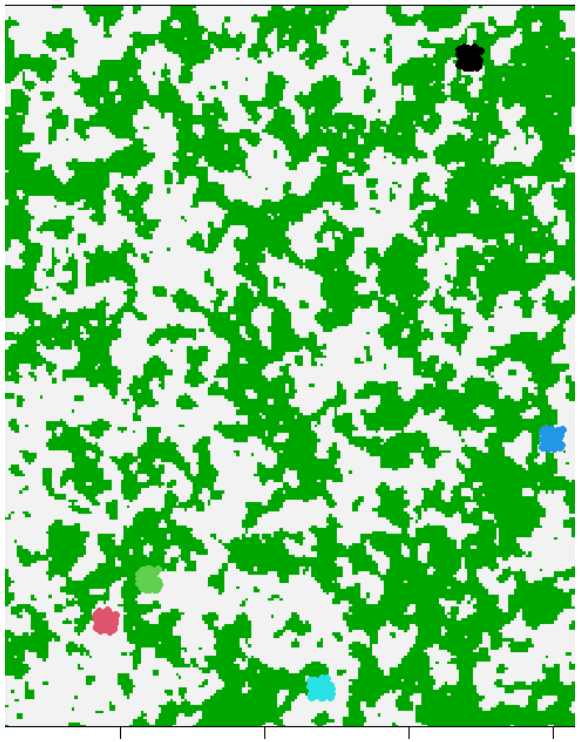
- Genetic Distance
- Visualization

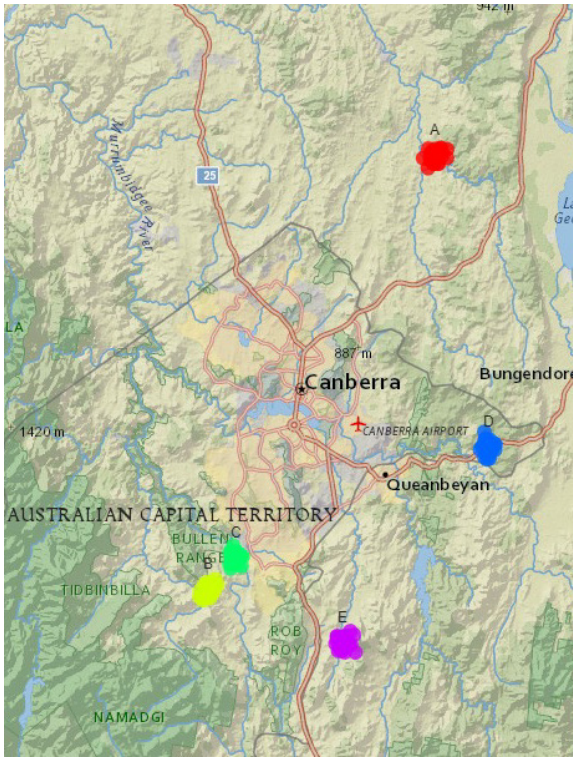
- PCA is a great tool for visualizing structure in a multidimensional dataset
- Take care with missing values
  gl.impute()
- Attend to any possible batch effects, sex-linked haploblocks, structural variants, kinship relationships
  gl.filter.sexlinkage()
  gl.filter.factorloadings()
- PCA is an exploratory tool. Do not infer process from pattern. Follow hypotheses with definitive analysis.

Structure

- What is “structure”
  - barriers to random mating, notably geographic and reproductive choices, lead to allele frequency divergence, causing population structure (= differences in allele or haplotype frequencies among groups within a species)
  - Ancestry is a measure of a individuals genome originating from various defined groups or populations.
  - But keep in mind: models that approximate population structure and ancestry examine a tangled web of past migrations and mixtures of ancestors and populations.
  - So there is not a single answer if you try to put such a complex history into “boxes” and as they vary in scale and also between methods

Simulated possums in Canberra





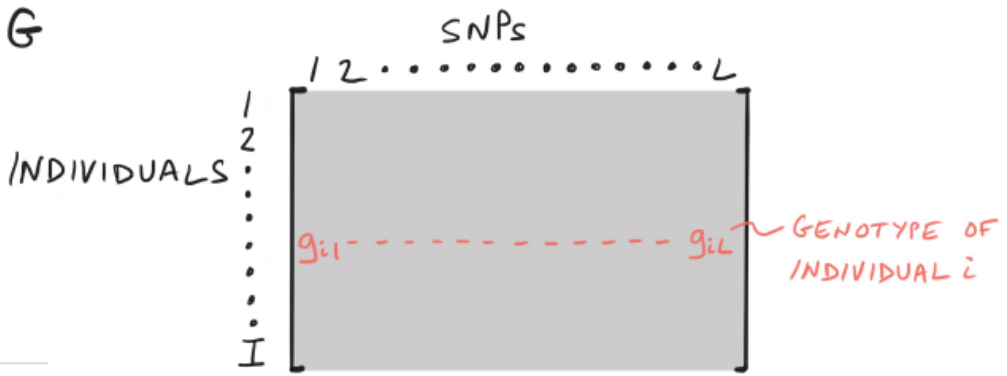
- 30 individuals in 5 (10) populations
- Typical metapopulation (individuals migrate at a low rate following and isolation by distance (actually resistance)

Clustering

An Owner's Guide to the Human Genome:
an introduction to human population genetics, variation and disease

by Jonathan Pritchard, Stanford University

- Genotype Matrix
  - Each row is the genotype of an individual




```
> as.matrix(possums.gl)[1:5,1:7]
  X1 X2 X3 X4 X5 X6 X7
1  1  2  1  2  2  2  1
2  1  1  1  1  1  1  0
3  2  0  2  0  1  1  1
4  0  1  2  0  0  1  0
5  1  2  1  2  2  2  1
>
```

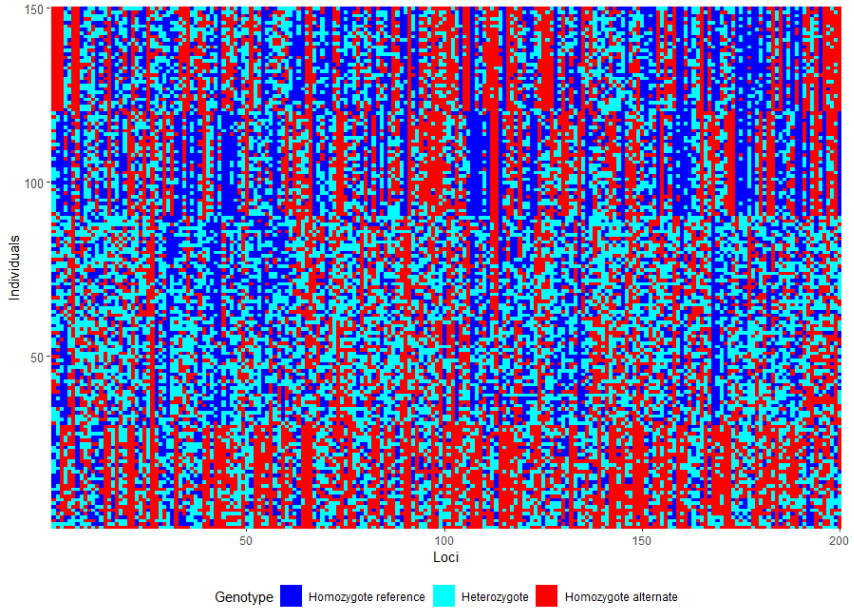
web.stanford.edu/group/pritchardlab/HGbook.html

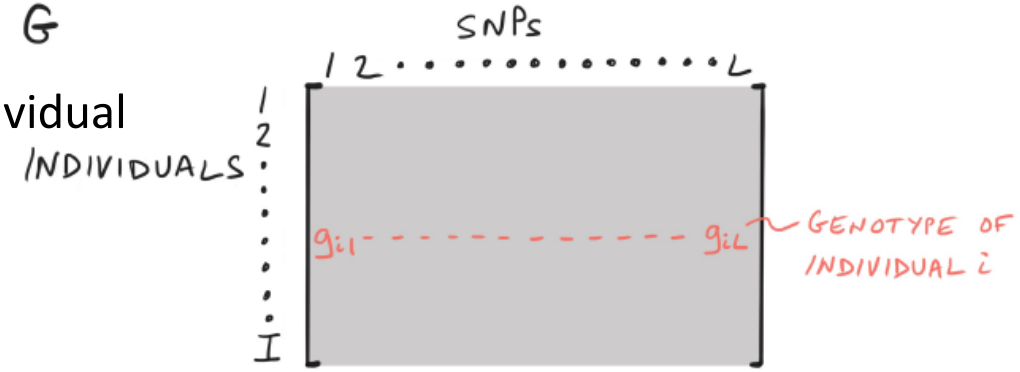
An Owner's Guide to the Human Genome:
an introduction to human population genetics, variation and disease

by Jonathan Pritchard, Stanford University

Clustering

- Genotype Matrix
  - Each row is the genotype of an individual






```
> gl.smearplot(possums)
```

web.stanford.edu/group/pritchardlab/HGbook.html

An Owner's Guide to the Human Genome:
an introduction to human population genetics, variation and disease

by Jonathan Pritchard, Stanford University

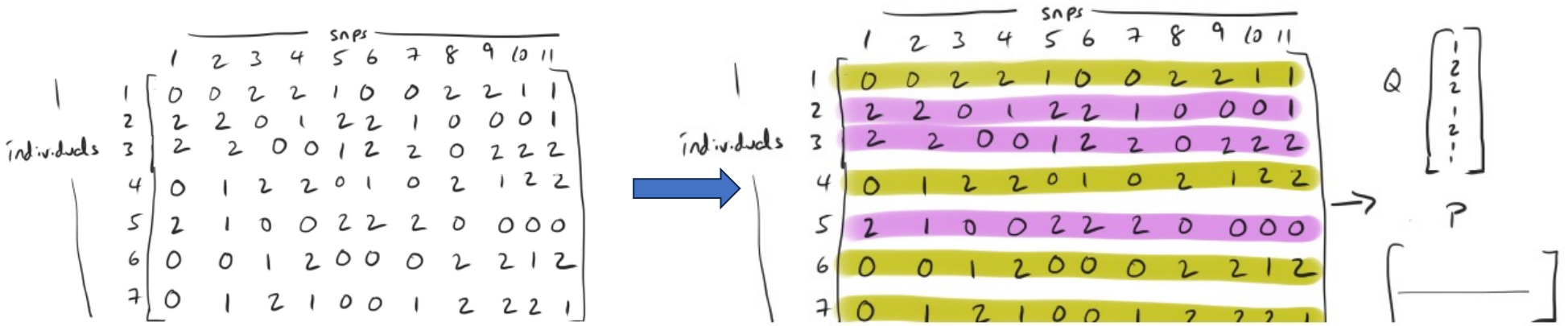
Clustering

- Can you do that by “hand”

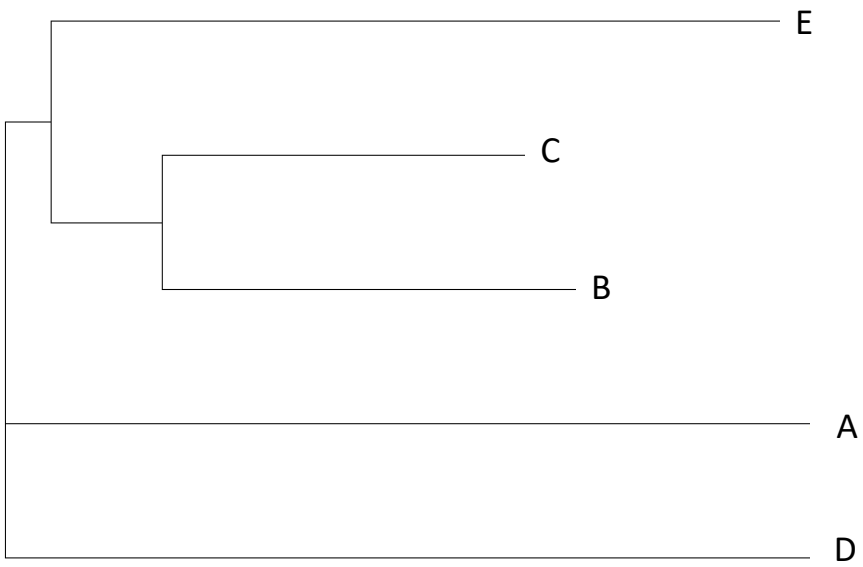
An Owner's Guide to the Human Genome:
an introduction to human population genetics, variation and disease

by Jonathan Pritchard, Stanford University

Clustering



Another way of clustering



- Neighbourjoining tree

```
> gl.tree.nj(possums)
```

Comparison of methods

- Average assignment error

Article | Open access | Published: 04 May 2022

Fast and accurate population admixture inference from genotype data from a few microsatellites to millions of SNPs

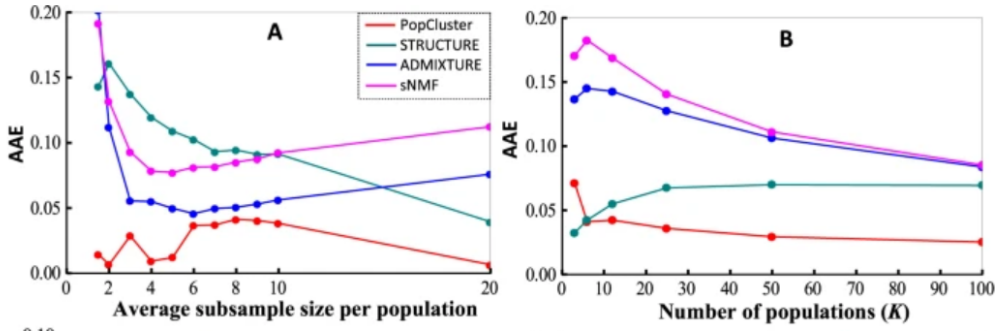
Jinliang Wang

Heredity **129**, 79–92 (2022) | Cite this article

Simulation 1, small samples

STRUCTURE performs poorly when a sample contains a small number of individuals drawn from each population (Fig. 1A). When each population is represented by just a few individuals (say, 10 or fewer), STRUCTURE is less accurate in inferring individual admixture proportions, making more errors than other methods. However, with an increasing sample size for each population, it becomes more accurate than ADMIXTURE and sNMF. Across the entire range of sample sizes considered in the simulation, ADMIXTURE is more accurate than sNMF, and PopCluster is the most accurate among the four methods.

Fig. 1: Simulation results.



Comparison of methods

- Effect of FST

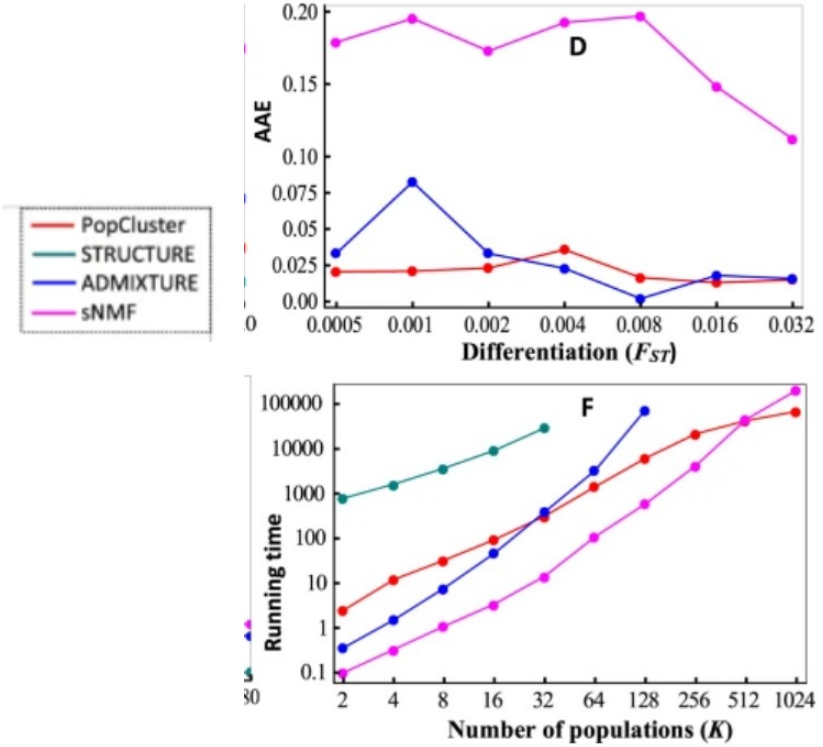
- Runtime

Article | Open access | Published: 04 May 2022

Fast and accurate population admixture inference from genotype data from a few microsatellites to millions of SNPs

Jinliang Wang

Heredity **129**, 79–92 (2022) | Cite this article



Comparison of methods

- Effect of FST

- Runtime

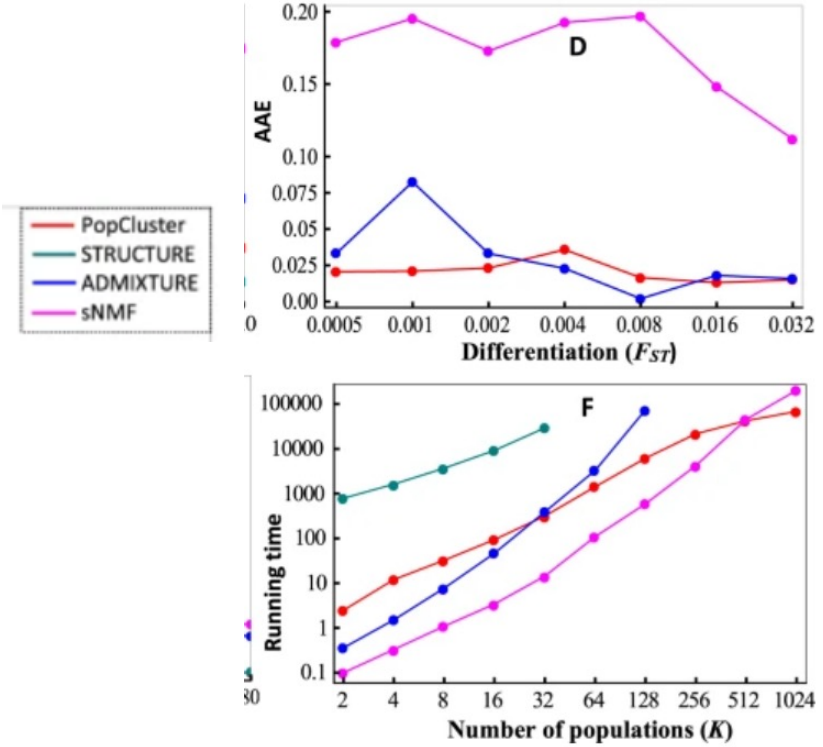
- Anyone uses PopCluster?

Article | Open access | Published: 04 May 2022

Fast and accurate population admixture inference from genotype data from a few microsatellites to millions of SNPs

Jinliang Wang

Heredity **129**, 79–92 (2022) | Cite this article



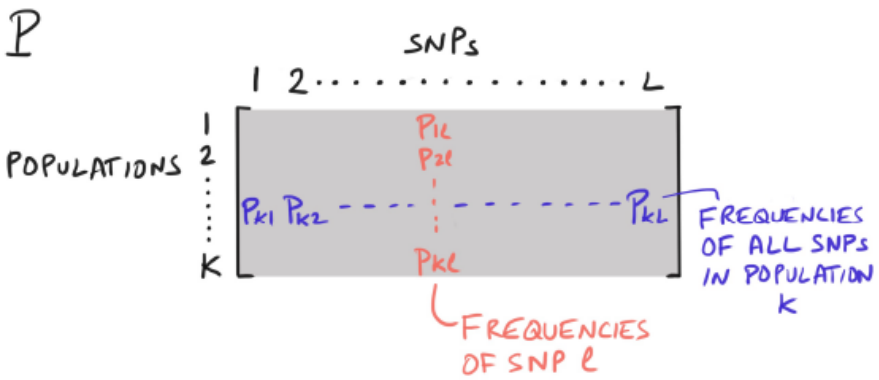
An Owner's Guide to the Human Genome:
an introduction to human population genetics, variation and disease

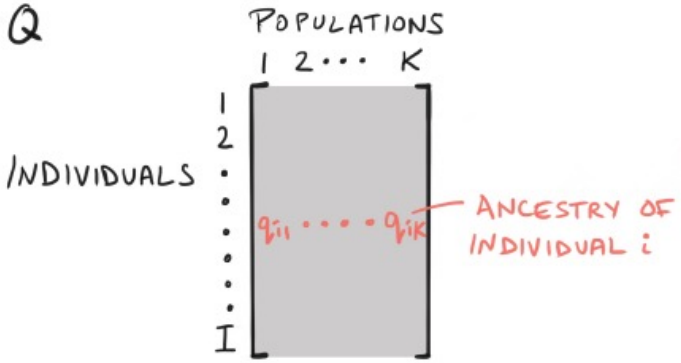
by Jonathan Pritchard, Stanford University

Back to basics - Clustering – P and Q matrix

we assume

- that populations are discrete groups of random-mating individuals.
- that, within populations, SNPs are in Hardy Weinberg equilibrium and pairs of SNPs that are far apart in the genome are in linkage equilibrium.





An Owner's Guide to the Human Genome:
an introduction to human population genetics, variation and disease

by Jonathan Pritchard, Stanford University

No admixture version

- Assume one loci,

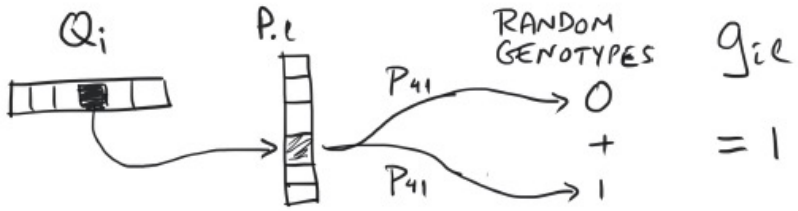


Figure 3.10: **Genotype sampling in the no-admixture model.** *Q_i tells us that individual i is from population k. Therefore the relevant allele frequency at SNP l is $p_{k,l}$. Then the genotype $g_{i,l}$ is the sum of two random allele draws; for each draw the probability of the '1' allele is $p_{k,l}$.*

probability of individual i's entire genotype given their ancestry Qi, by multiplying these probabilities across all SNPs

$$\Pr(G_i|Q_i) = \prod_{l=1}^{L} \Pr(g_{i,l}|Q_i)$$

An Owner's Guide to the Human Genome:
an introduction to human population genetics, variation and disease

by Jonathan Pritchard, Stanford University

Admixture version

- Assume one loci,

Figure 3.11: Genotype sampling in the ad-

Genotype sampling in the admixture model. For each allele we first pick the population-of-origin at random according to the ancestry probabilities in Qi. These then tell us where to look in P.,l for the relevant frequencies for each allele. Genotype gi,l is then the sum of two random allele draws, each according to its

probability of individual i's entire genotype given their ancestry Qi, by multiplying these probabilities across all SNPs

$$\Pr(G_i|Q_i) = \prod_{l=1}^{L} \Pr(g_{i,l}|Q_i)$$

An Owner's Guide to the Human Genome:
an introduction to human population genetics, variation and disease

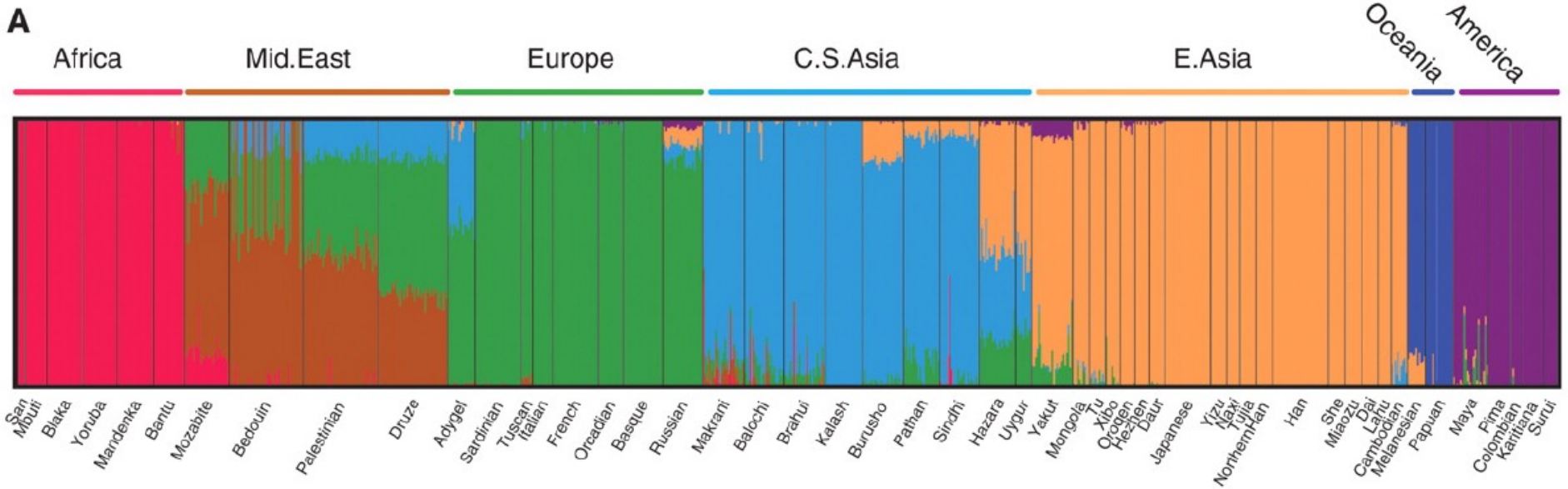
by Jonathan Pritchard, Stanford University

Maximise Likelihood using

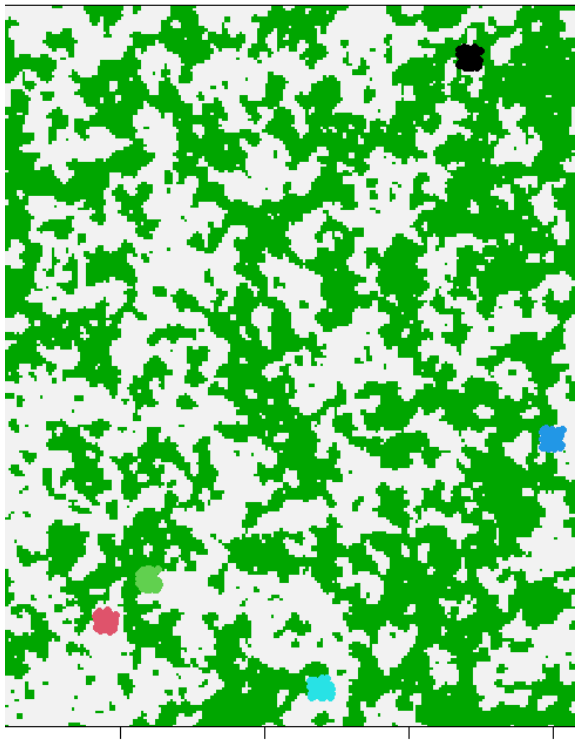
- *Initialization:* Set the ancestry for each individual, Q_i, at random (typically from a uniform distribution).
- *repeat*
  {
  - *Update P given Q:* For each SNP, update the allele frequency estimates according to the allele counts in each population.
  - *Update Q given P:* For each individual, update their ancestry estimate Q_i given the allele frequencies.

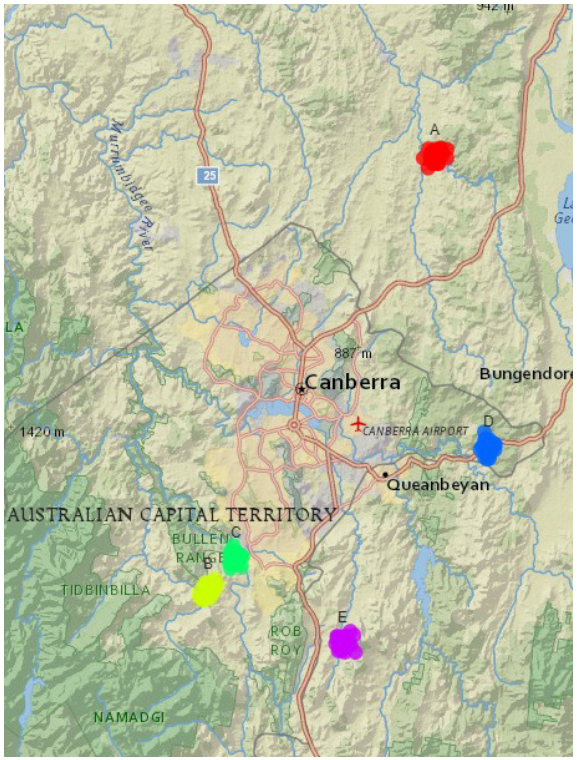
  } until converged.

Structure plots



Simulated possums in Canberra

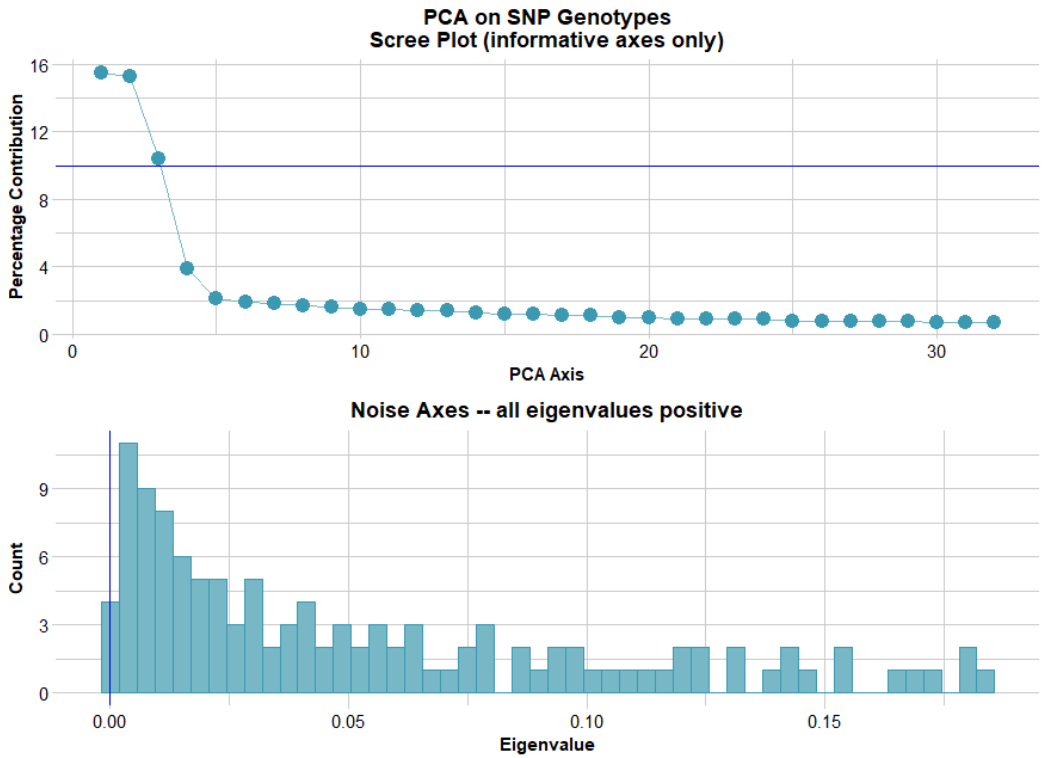




- 30 individuals in 5 (10) populations
- Typical metapopulation (individuals migrate at a low rate following and isolation by distance (actually resistance)

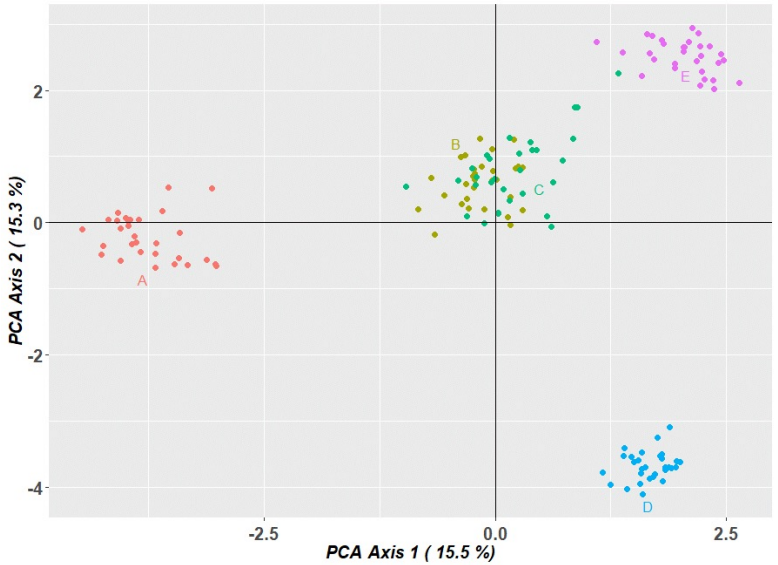
Possums PCA

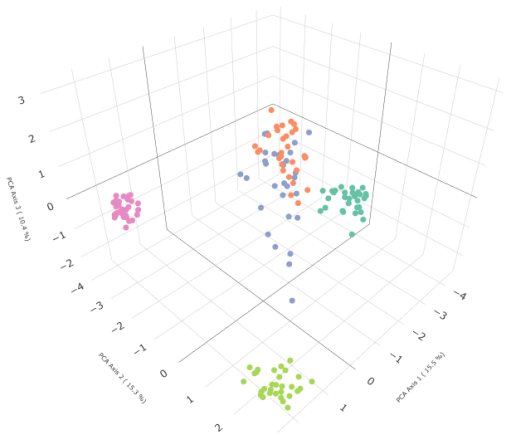
```
> pcoa <- gl.pcoa(possums)
```



Possums PCA

- gl.pcoa.plot(pcoa, possums)
- gl.pcoa.plot(possums, zaxis = 3)





Run structure, no admixture

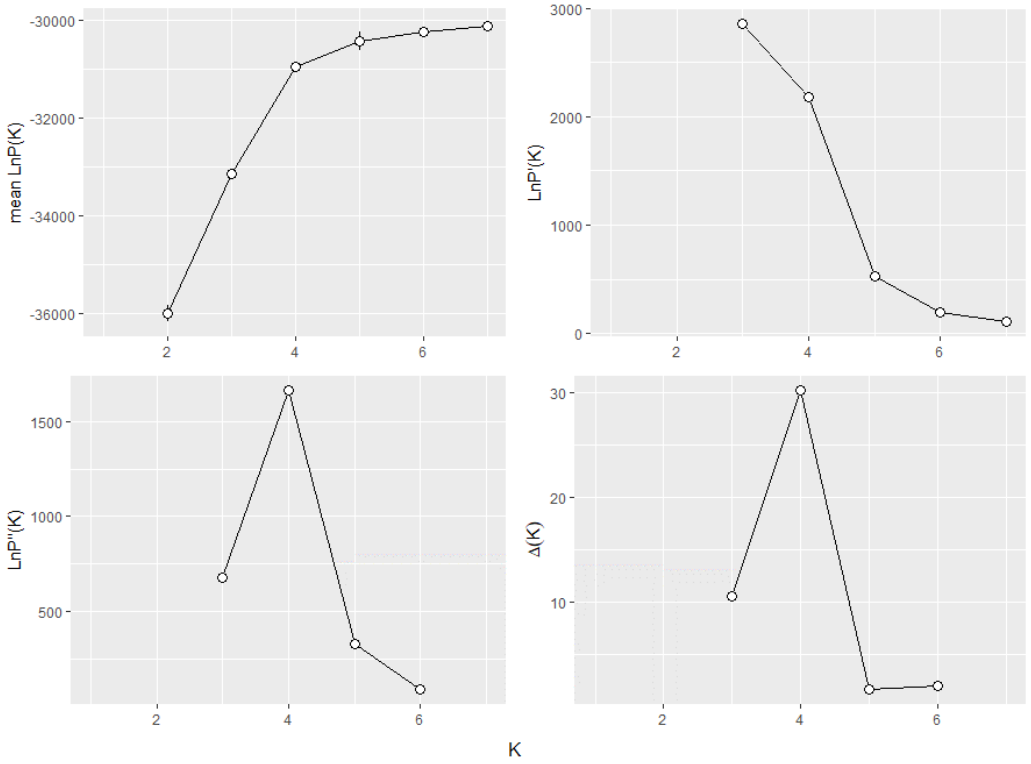
```
srnoad <- gl.run.structure(possums, k.range = 2:7, num.k.rep = 10,
                           exec = '/binaries/linux/structure',plot.out = FALSE,
                           burnin=50000, numreps=100000, noadmix=TRUE)
```

k.range	vector of values to for maxpop in multiple runs. If set to NULL, a single STRUCTURE run is conducted with maxpops groups. If specified, do not also specify maxpops.
num.k.rep	number of replicates for each value in k.range.
exec	name of executable for STRUCTURE. Defaults to "structure".
burnin	number of iterations for MCMC burnin.
numreps	number of MCMC replicates.
noadmix	logical. No admixture?

Evanno Plot

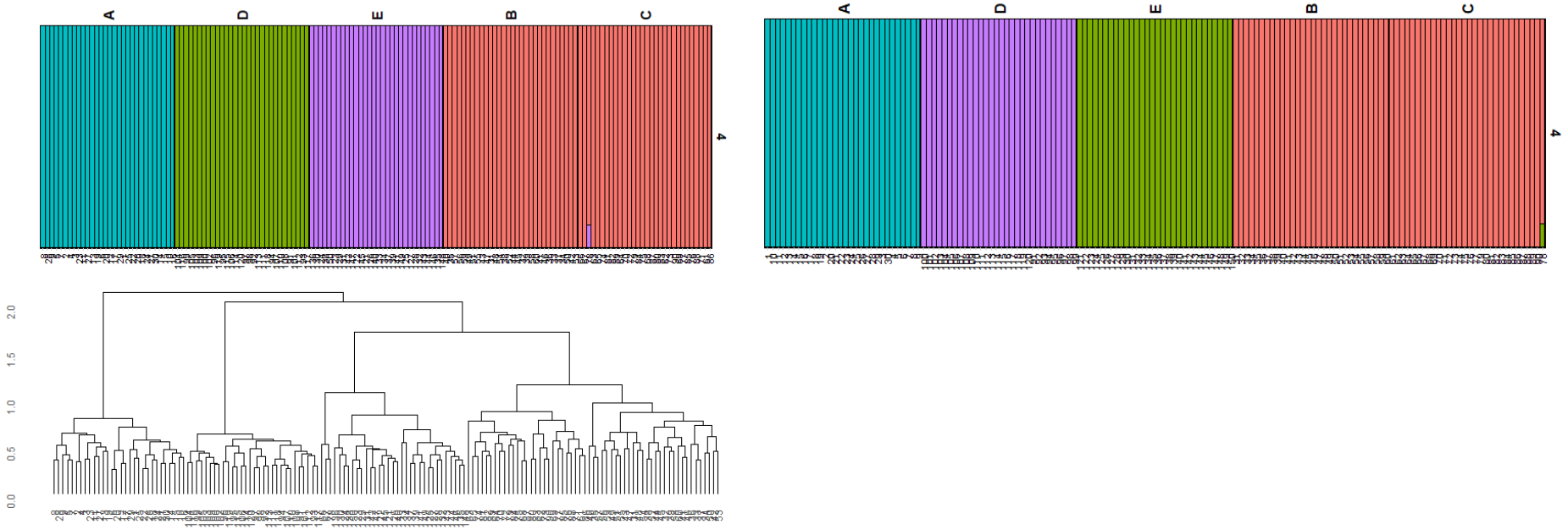
- Plot 1:
  best K = 4: L does not show a substantial increase

- Plot 4: best K=4: maximum of second derivative of L



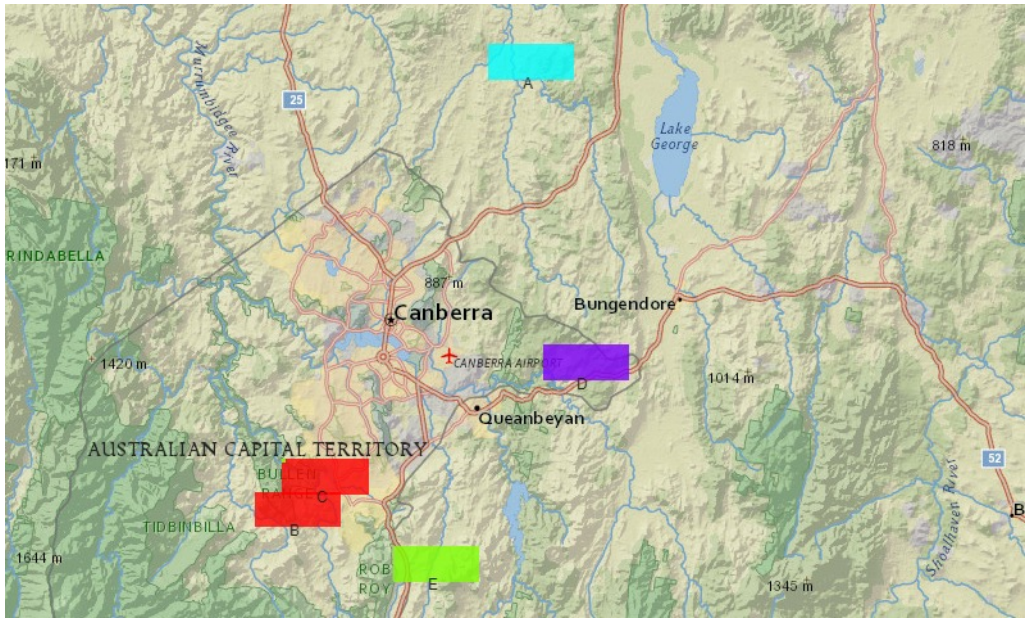
Output: Q-matrix (for K=4)

- `qmat.noad <- gl.plot.structure(sr.noad, K=4,x=possum)s, den = TRUE)`

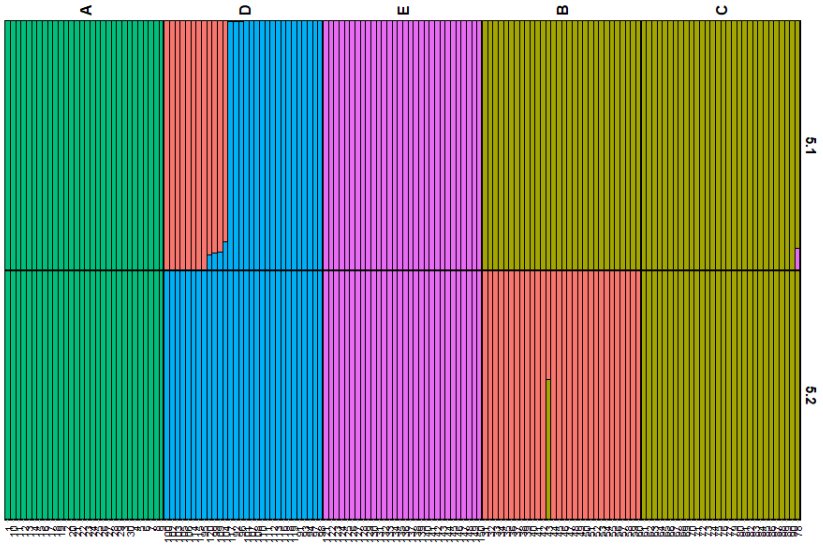


Spatial structure plot

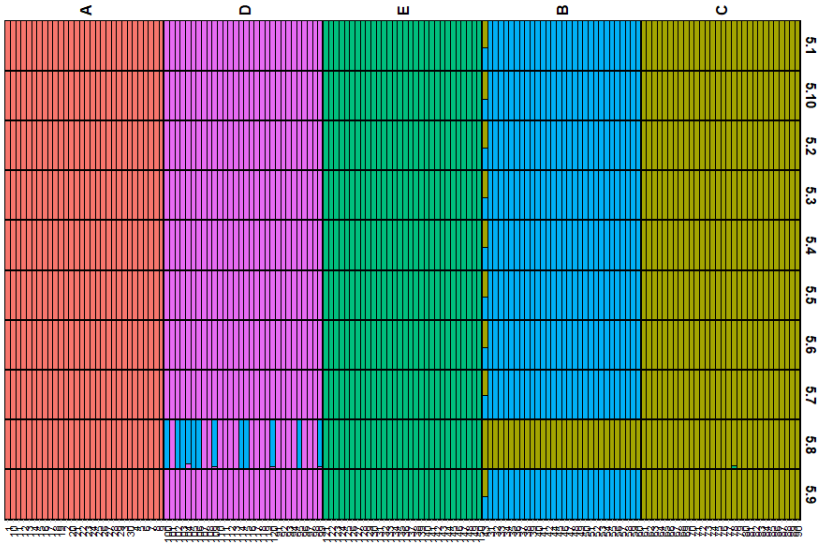
- `gl.map.structure(qmat = qmat.noad, x = possums,K=4)`



A bit more checking: K=5



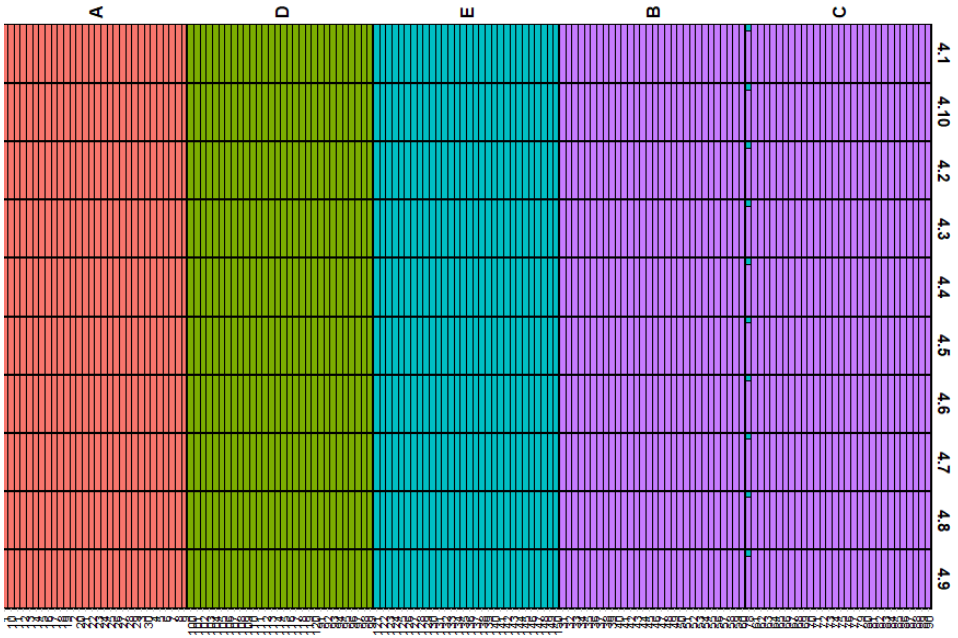

```
qmat.noad <- gl.plot.structure(sr.noad,
K=5,x=possums,clumpak = TRUE)
```




```
qmat.noad <- gl.plot.structure(sr.noad,
K=5,x=possums,clumpak = FALSE)
```

A bit more checking

- Checking the repeats



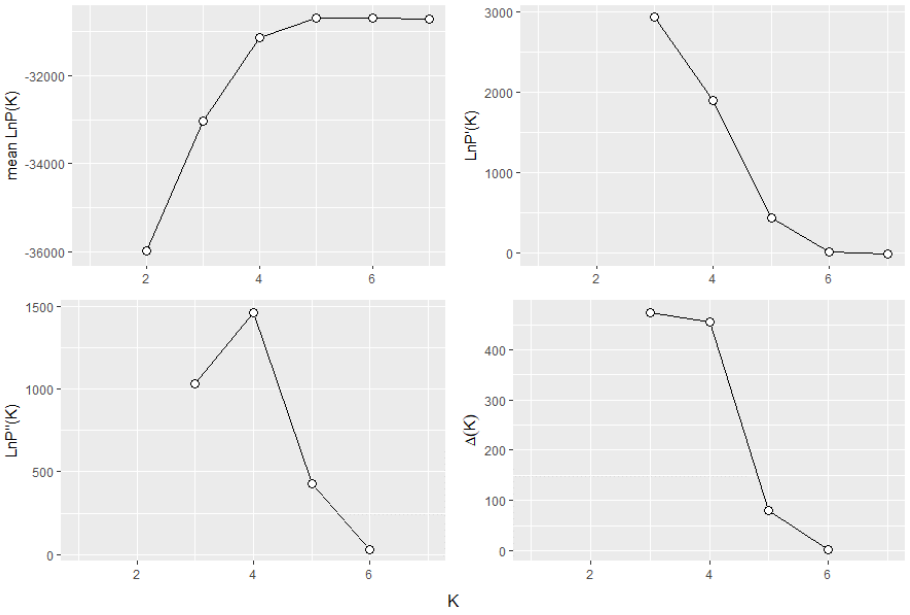

```
qmat.noad <- gl.plot.structure(sr.noad,
K=4,x=possums,clumpak = FALSE)
```

Run structure, with admixture (noadmix=FALSE)

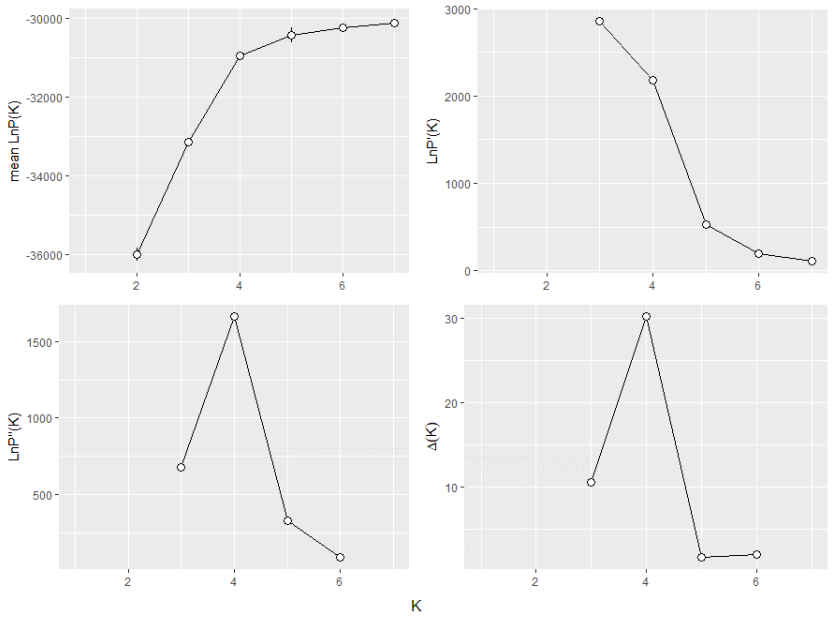
```
srad <- gl.run.structure(possums, k.range = 2:7, num.k.rep = 10,
                         exec = './binaries/linux/structure',plot.out = FALSE,
                         burnin=50000, numreps=100000, noadmix=FALSE)
```

Evanno Plot

- admixture

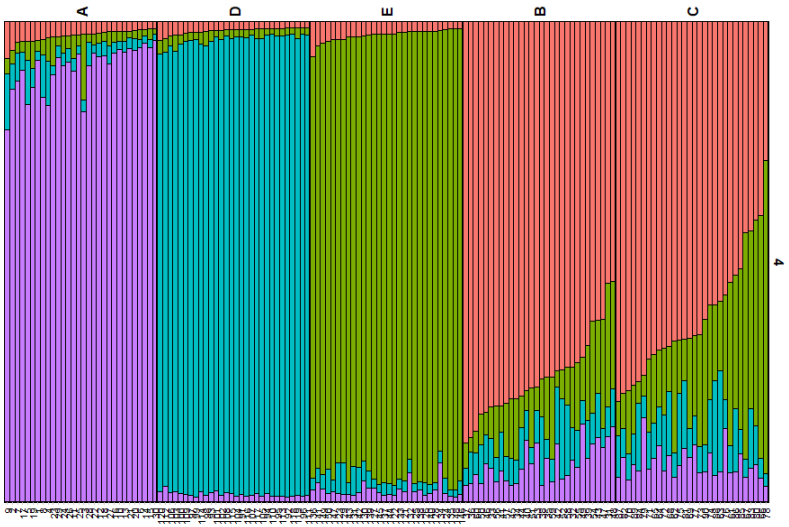


- No admixture

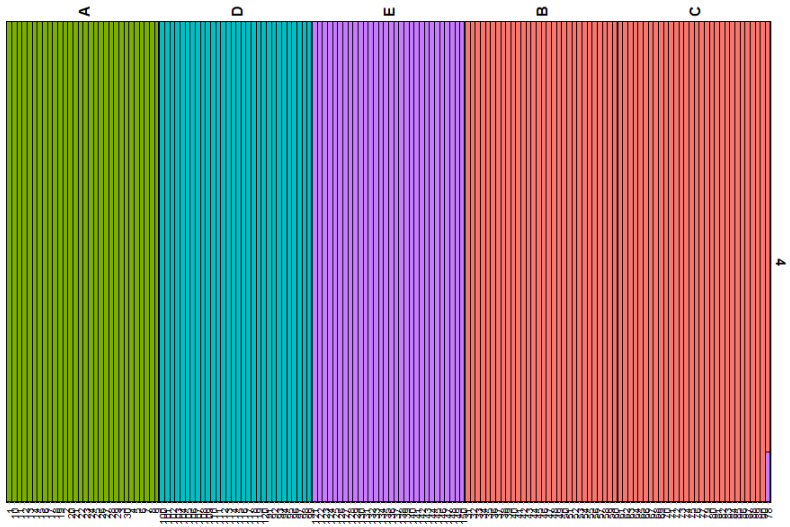


Qmat K=4: admixture vs no admixture

- Admixture



no admixture



Interpretation of clusters

• First: What do the population clusters represent? One way to think about this is in terms of an individual's ancestors. Remember that each of our genomes is made up of many segments of DNA, inherited from many different ancestors. Going backward in time, you can think about these ancestors as occupying a distribution across geographic space or populations.

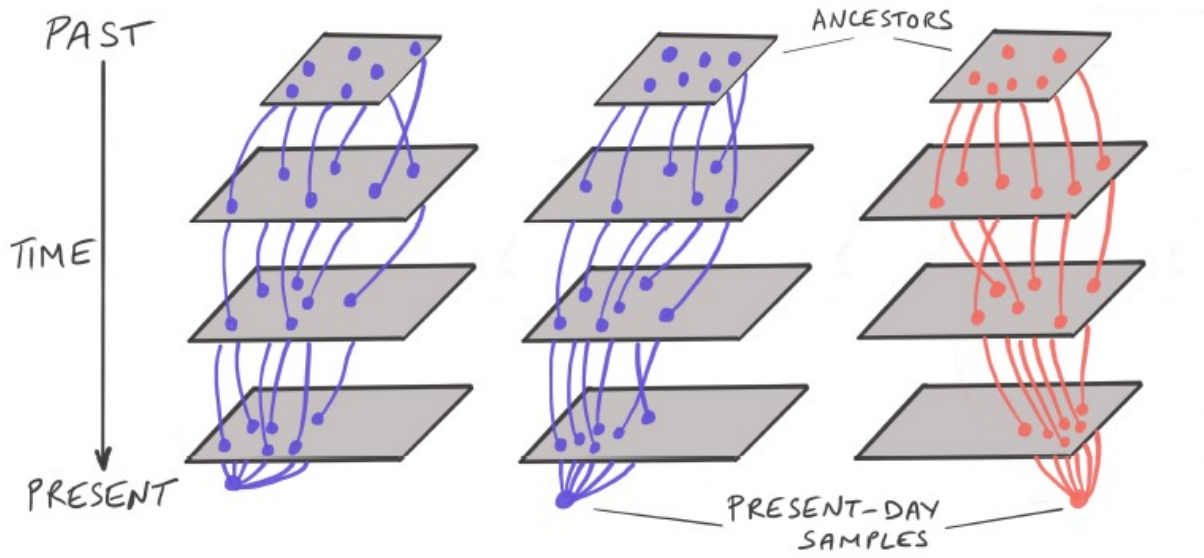


Figure 3.21: **You can think of genetically similar individuals as having similar geographic distributions of ancestors.**
The cartoon shows the geographic distributions of ancestors for three individuals. Each slice shows the locations of ancestors at a single point in time.
Individuals who cluster together have similar geographic distributions of ancestors.

FastStructure

- In general:
- FastStructure uses a model-based clustering approach, where it assumes that the observed genetic data arise from a mixture of underlying populations. Each individual in the dataset can have ancestry from one or more of these populations. The goal is to infer:

  - The number of populations (K) that best explains the observed data.
  - The proportion of each individual's genome that comes from each inferred population, known as ancestry coefficients. (Q-matrix)

Faststructure

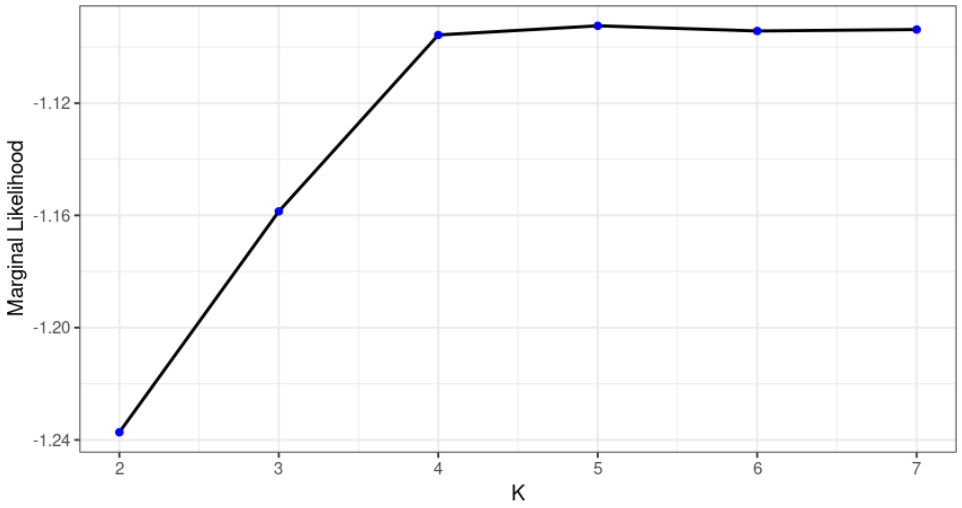
- There is no no-admixture model, model selection different, but much faster. Still same idea of maximising Likelihood for using P and Q matrices.
- Runs only on Linux and Mac (no windows)

```
gl.run.faststructure(possums, k.range = 2:10, num.k.rep = 3,
                     exec = "./binaries/fastStructure",
                     exec.plink="./binaries", prior = "simple“,
                     cv=30)
```

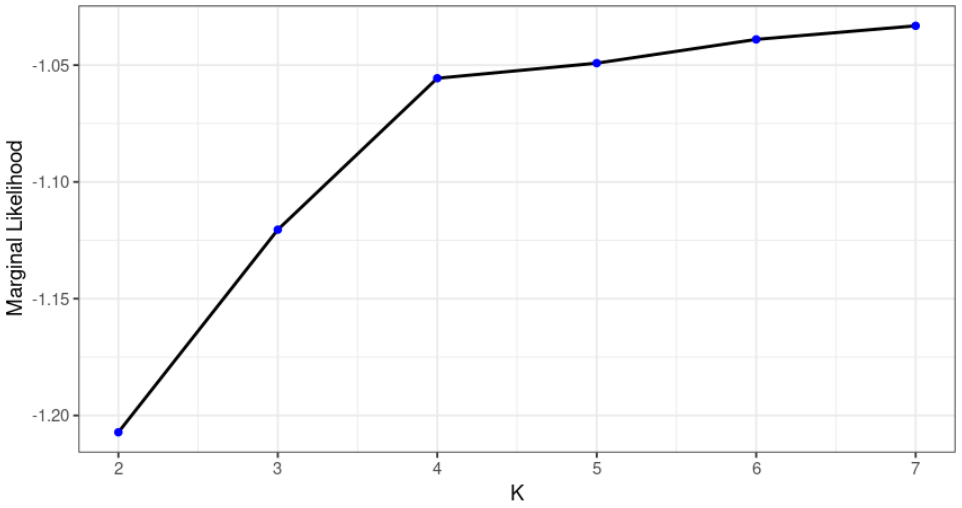
```
Prior Simple or logistic
CV  cross validation
```

Faststructure possums:(simple vs logistic)
K=4

simple

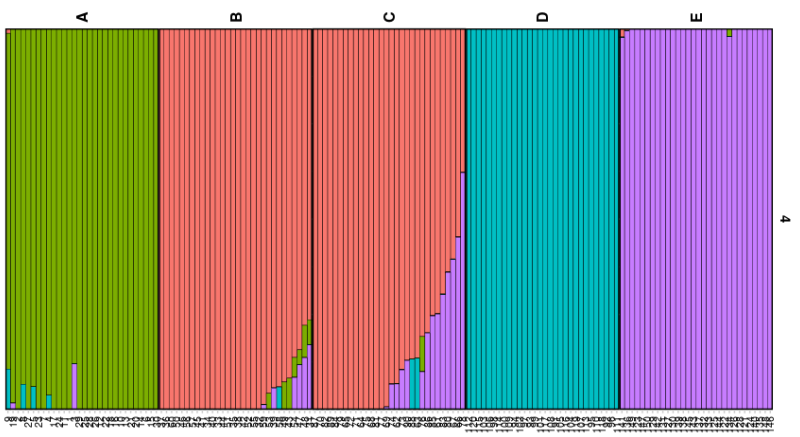


logistic

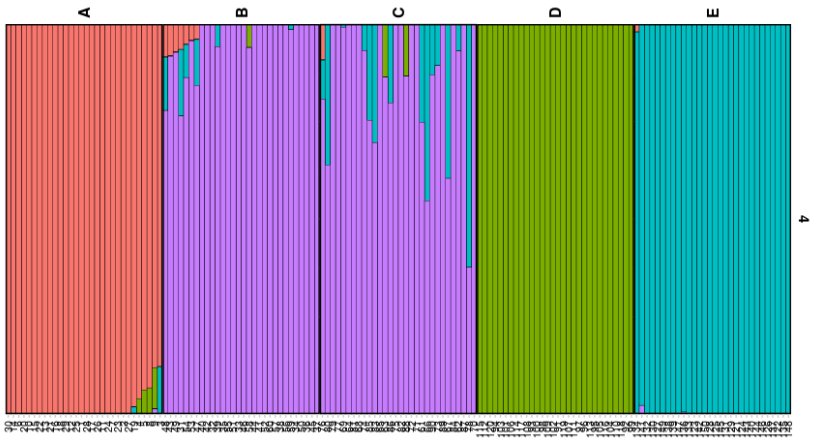


Faststructure possums:(simple vs logistic)
K=4

simple

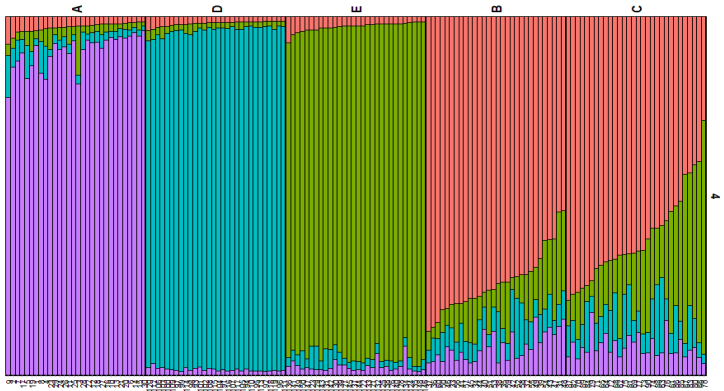


logistic

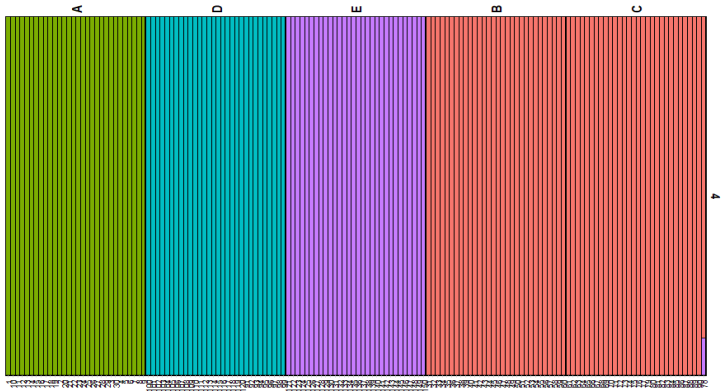


Structure vs Faststructure
K=4

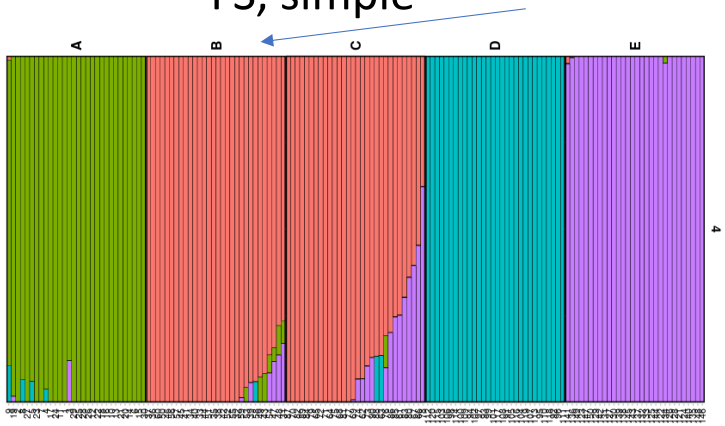
admixture



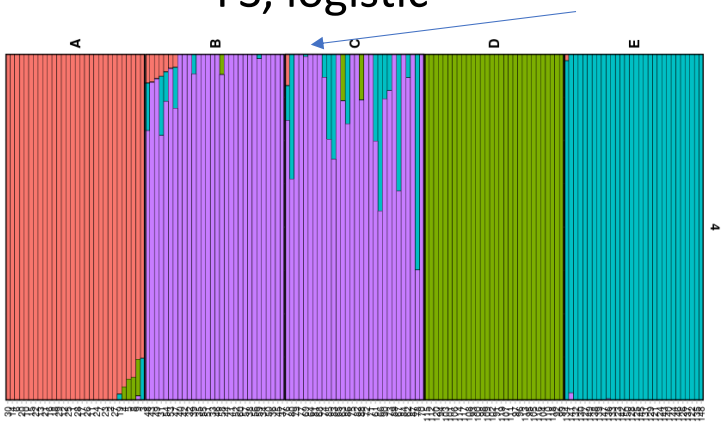
No admixture



FS, simple



FS, logistic



New innovations by Bill Sherwin

Beta (between location) diversity – more than Fst (q=2)

Mutual Information (“ *I* ”) Sherwin et al 2017,2021

- a measure of whether the alleles in an individual tell us where it came from – easier if the locations are highly differentiated
- { Also presented as Shannon Differentiation $\frac{I}{Max\ I}$ }

Adjusted AFD ^{A}A Sherwin 2022

- Not in the Hill-family (q=0,1,2) that contains most other differentiation measure
- Based on Bray-Curtis

DIFFERENTIATION $F_{ST} = (H_T - H_S)/H_T$

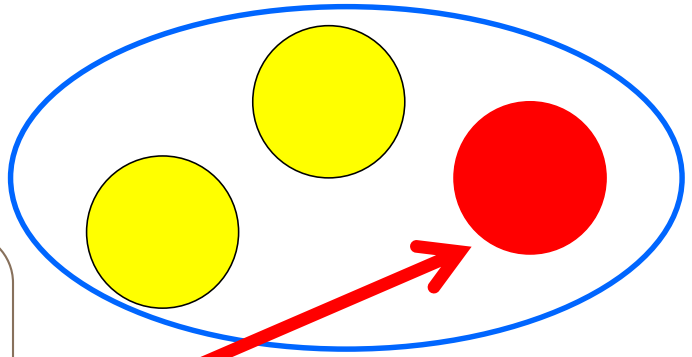
Compare with Shannon Differentiation = Mutual Info / Max MI

b1 b1 b1 b1 b1 b1 b1 b1	VS	b1 b1 b1 b1 b1 b1 b1 b1	All alleles shared F_{st}=0 Shannon differentiation=0
b3 b3 b3 b3 b2 b2 b2 b2	VS	b1 b1 b1 b1 b1 b1 b1 b1	No Shared alleles F_{st}=0.6 Shannon differentiation=1
b2 b2 b2 b2 b2 b2 b2 b2	VS	b1 b1 b1 b1 b1 b1 b1 b1	No Shared alleles F_{st}= 1 Shannon differentiation=1

NB Many apps for adaptive genes use F_{st} still! Sherwin 2022

Gray Wolf adaptive genes 90% false discovery? Schweizer 2016

DIFFERENTIATION



This population is invasive – which of the others is the source?

This population is endangered – will one of the others rescue it naturally?

Do some parts of the genome have alleles that are adapted to conditions here?

MUTUAL INFORMATION (I) – BASED ON SHANNON INFO ^1H

	b1	b7
Pop. 1	50	0
Pop. 2	0	50

- Alleles tell us the individual's population membership.
- "High Mutual Information" (I)
- Sign of LOW dispersal

	b1	b7
Pop. 1	25	25
Pop. 2	25	25

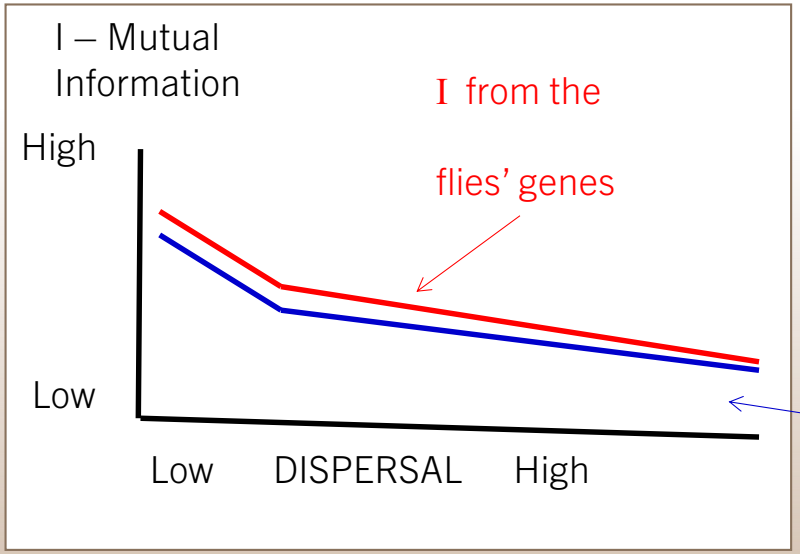
- Alleles DON'T tell us the individual's population membership.
- "Low Mutual Information" (I)
- Sign of HIGH dispersal

Roddy Dewar

MUTUAL INFORMATION I IS GOOD AT ESTIMATING DISPERSAL

Clare Holleley

Flies with known population size (N) & dispersal (m)



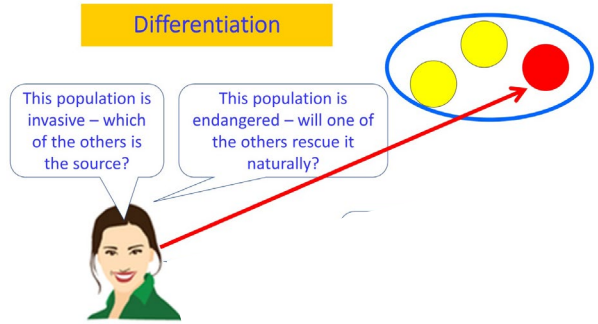
I from equation
I ~ 1/(N,m)

N - pop size
m - dispersal

MUTUAL INFORMATION I OUTPERFORMS OTHER DISPERSAL ESTIMATES

Low bias & random error
OK for

- Invasives
  (Large Pop, high dispersal)
- Endangered
  (Small Pop, low dispersal)



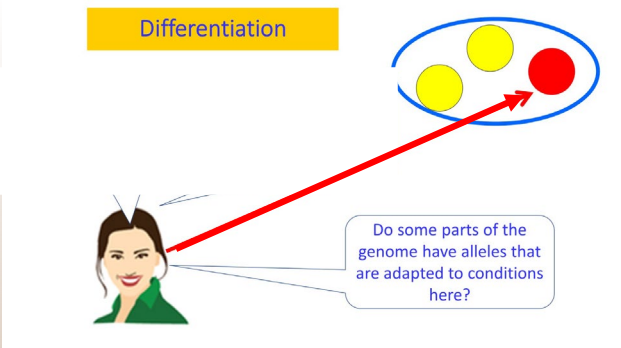
HELP
DETECT GENOMIC LOCATIONS
WHERE ADAPTATION IS HAPPENING

Detecting Loci under Selection (Simulation)

- Single differentiation measures ${}^{A}A$, F_{ST}, D_{EST},I, etc
- 'Candidate Loci' in top 1% of neutral values for genetic differentiation
- << 45% loci true positive !! out of all

BUT:

- Paired measures always much better than single
- Best 3 pairs: $\boldsymbol{{}^{A}A + F_{ST}}$; $\boldsymbol{F_{ST} + D_{EST}}$; $\boldsymbol{D_{EST} + I}$
- 62% to 67% true positive



Sherwin 2022. Ecology and Evolution 12:e9176. dx.doi.org/10.1002/ece3.9176

THANKS

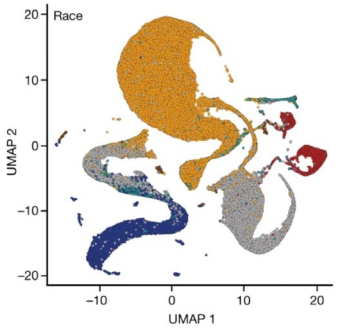
Many Physicists, Mathematicians, Statisticians, Biologists

$$$: Aust. Research Council, EU, Academia Sinica &c

Refs, Apps:

- Sherwin et al 2017,2021, TREE doi.org/10.1016/j.tree.2017.09.012 doi.org/10.1016/j.tree.2021.07.005
- **Sherwin 2022. Ecology and Evolution 12:e9176. dx.doi.org/10.1002/ece3.9176**
- chao.shinyapps.io/SpadeR
- georges.biomatix.org/dartR
- biology-assets.anu.edu.au/GenAlEx

W.Sherwin@unsw.edu.au



https://doi-org/10.1038/s41586-023-06957-x

Discussion

Beta (between location) diversity – more than Fst (q=2)

Mutual Information (“ ***I*** ”) Sherwin et al 2017,2021

- a measure of whether the alleles in an individual tell us where it came from – easier if the locations are highly differentiated
- { Also presented as Shannon Differentiation $\frac{I}{Max\ I}$ }

Adjusted AFD $^{A}\boldsymbol{A}$ Sherwin 2022

- Not in the Hill-family (q=0,1,2) that contains most other differentiation measure
- Based on Bray-Curtis

F_{st} is a poor measure of differentiation $F_{st}= (H_t-H_s)/H_t$

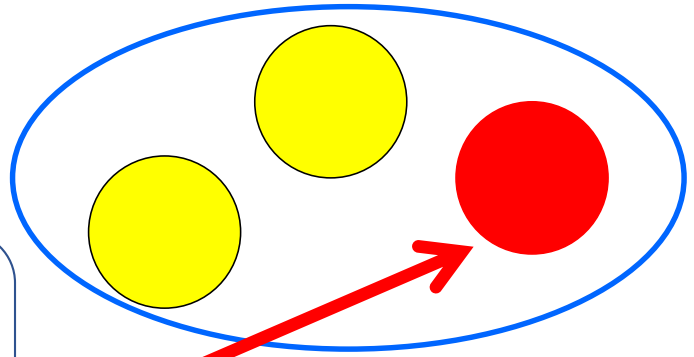
Compare with Shannon Differentiation = Mutual Info / Max MI

b1 b1 b1 b1 b1 b1 b1 b1	VS	b1 b1 b1 b1 b1 b1 b1 b1	All alleles shared F_{st} =0 Shannon differentiation=0
b3 b3 b3 b3 b2 b2 b2 b2	VS	b1 b1 b1 b1 b1 b1 b1 b1	No Shared alleles F_{st} =0.6 Shannon differentiation=1
b2 b2 b2 b2 b2 b2 b2 b2	VS	b1 b1 b1 b1 b1 b1 b1 b1	No Shared alleles F_{st} = 1 Shannon differentiation=1

NB Many apps for adaptive genes use F_{st} still! Sherwin 2022

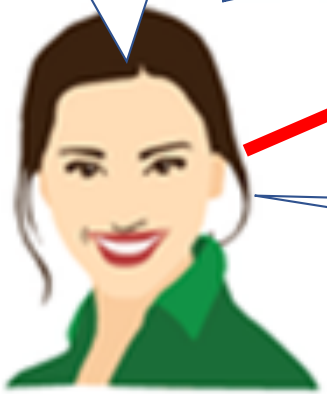
Gray Wolf adaptive genes 90% false discovery? Schweizer 2016

Differentiation



This population is invasive – which of the others is the source?

This population is endangered – will one of the others rescue it naturally?



Do some parts of the genome have alleles that are adapted to conditions here?

Mutual Information (I) – Based on Shannon Info ^{1}H

	b1	b7
Pop. 1	50	0
Pop. 2	0	50

- Alleles tell us the individual's population membership.
- "High Mutual Information" (I)
- Sign of LOW dispersal

	b1	b7
Pop. 1	25	25
Pop. 2	25	25

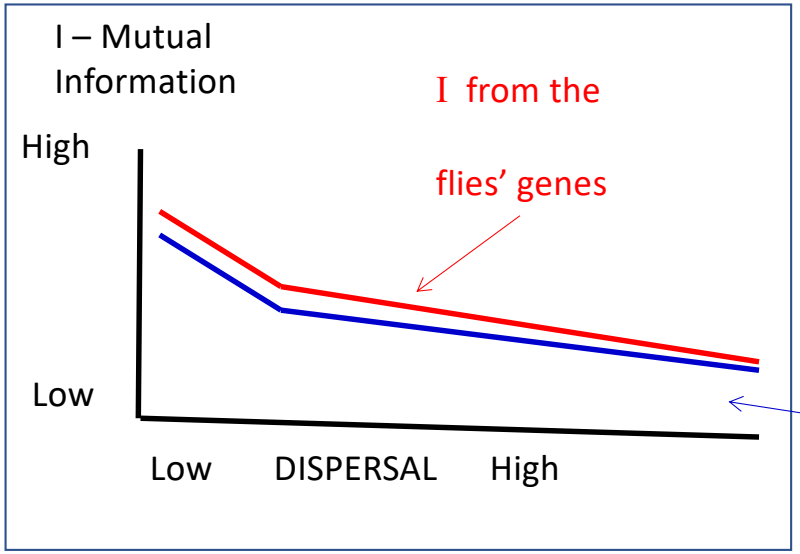
- Alleles DON'T tell us the individual's population membership.
- "Low Mutual Information" (I)
- Sign of HIGH dispersal

Roddy Dewar

Mutual Information **I** is good at estimating dispersal

Clare Holleley

Flies with known population size (N) & dispersal (m)



I from equation
I ~ 1/(N,m)

N - pop size
m - dispersal

Mutual Information $\mathbf{I}$ outperforms other dispersal estimates

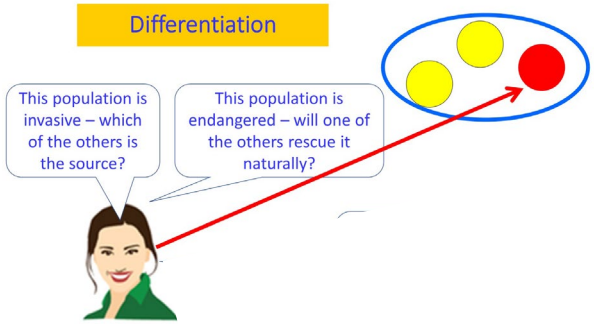
Low bias & random error

OK for

- Invasives

  (Large Pop, high dispersal)
- Endangered

  (Small Pop, low dispersal)



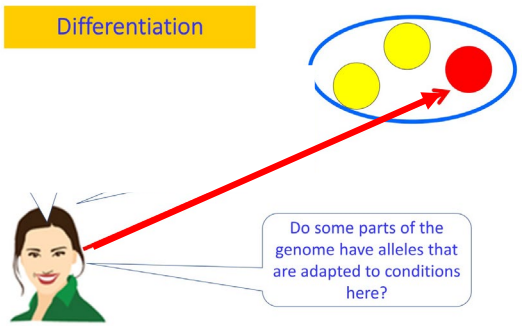
Mutual Information $\mathbf{I}$ & ${}^{A}A$ help detect genomic locations where adaptation is happening

Detecting Loci under Selection (Simulation)

- Single differentiation measures ${}^{A}A$, F_{ST}, D_{EST}, I, etc
- 'Candidate Loci' in top 1% of neutral values for genetic differentiation
- << 45% loci true positive !! out of all

BUT:

- Paired measures always much better than single
- Best 3 pairs: $\boldsymbol{{}^{A}A + F_{ST}}$; $\boldsymbol{F_{ST} + D_{EST}}$; $\boldsymbol{D_{EST} + I}$
- 62% to 67% true positive



Sherwin 2022. Ecology and Evolution 12:e9176. dx.doi.org/10.1002/ece3.9176

THANKS

Many Physicists, Mathematicians, Statisticians, Biologists

$$$: Aust. Research Council, EU, Academia Sinica &c

Refs, Apps:

- Sherwin et al 2017,2021. TREE doi.org/10.1016/j.tree.2017.09.012 doi.org/10.1016/j.tree.2021.07.005
- Sherwin 2022. Ecology and Evolution 12:e9176. dx.doi.org/10.1002/ece3.9176

- chao.shinyapps.io/SpadeR
- georges.biomatix.org/dartR
- biology-assets.anu.edu.au/GenAlEx

W.Sherwin@unsw.edu.au