Pop Gen In Conservation & Restoration

William Sherwin,
Carlo Pacioni and
Renee Catullo

Population genetics: theory is hard but it matters when filtering

Renee Catullo

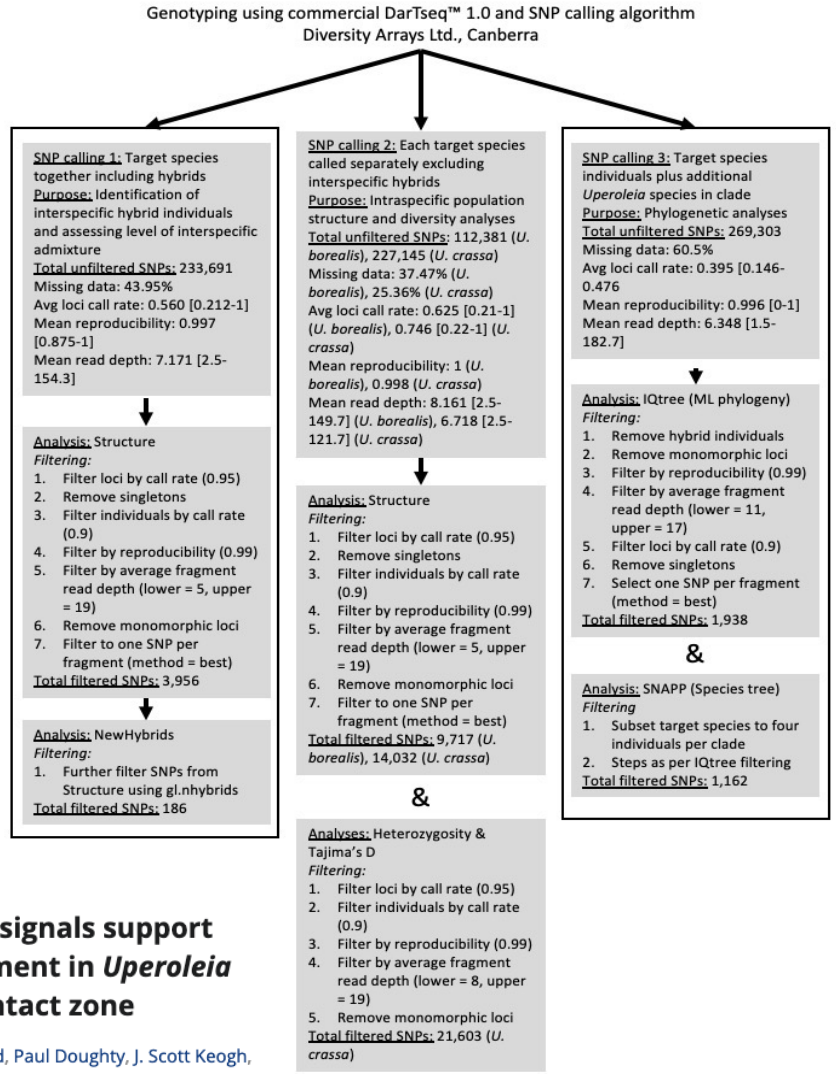
University of Western Australia

Filtering strategies (or the Art of Filtering)

- One set of SNPs and one set of filters cannot answer all your questions about a single species
- It will give you an answer but there is a good chance it is wrong

This week I will cover:

1. What are the basic filters and considerations? (this session)
2. What are some key considerations for filtering for specific conservation questions? (session 2)
3. What do you need to think about when calling SNPs? (session 3)



Population genomics and sexual signals support reproductive character displacement in *Uperoleia* (Anura: Myobatrachidae) in a contact zone

Frederick R. Jaya, Jessie C. Tanner, Michael R. Whitehead, Paul Doughty, J. Scott Keogh, Craig C. Moritz, Renee A. Catullo

Filtering for call rate (method = “loc”)

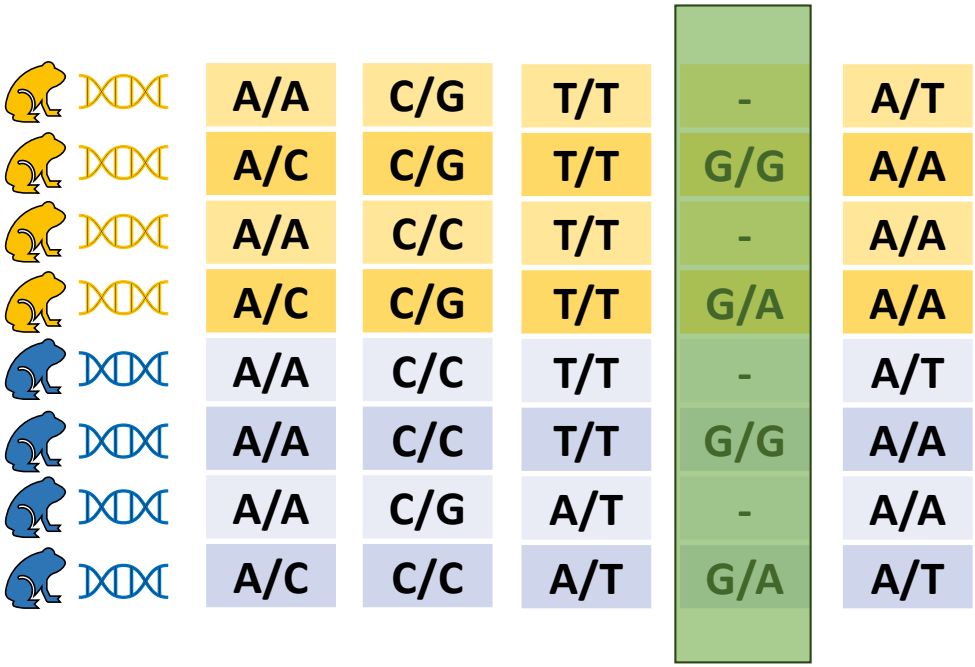
What is it?

- Removing SNPs because they didn’t work across individuals to a set level

- Why is it important?
  - Missing loci can add “noise” and computation time to an analysis

Key Considerations:

- Trade-off between strength of filter and number of loci
- Call rate filters matter for different metrics
  - Strong filters disproportionately remove heterozygous sites
  - Many population structure methods work better without high levels of missing data



A/A	C/G	T/T	-	A/T
A/C	C/G	T/T	G/G	A/A
A/A	C/C	T/T	-	A/A
A/C	C/G	T/T	G/A	A/A
A/A	C/C	T/T	-	A/T
A/A	C/C	T/T	G/G	A/A
A/A	C/G	A/T	-	A/A
A/C	C/C	A/T	G/A	A/T

Keep up …

testset.gl

```
gl <- testset.gl
nInd(gl)
nLoc(gl)
nPop(gl)
popNames(gl)
indNames(gl)
locNames(gl)
table(pop(gl))
as.matrix(gl)[1:7,1:5]

gl <- testset.gl
gl.set.verbosity(3)
gl.report.callrate(gl)
gl.report.callrate(gl,method="ind")
gl.report.reproducibility(gl)

gl.report.callrate(gl,method="ind")
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80)

gl <- testset.gl
gl.smearplot(gl)
gl <- gl.filter.callrate(gl,verbose=0)
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80,
verbose=0)
gl.smearplot(gl)
```

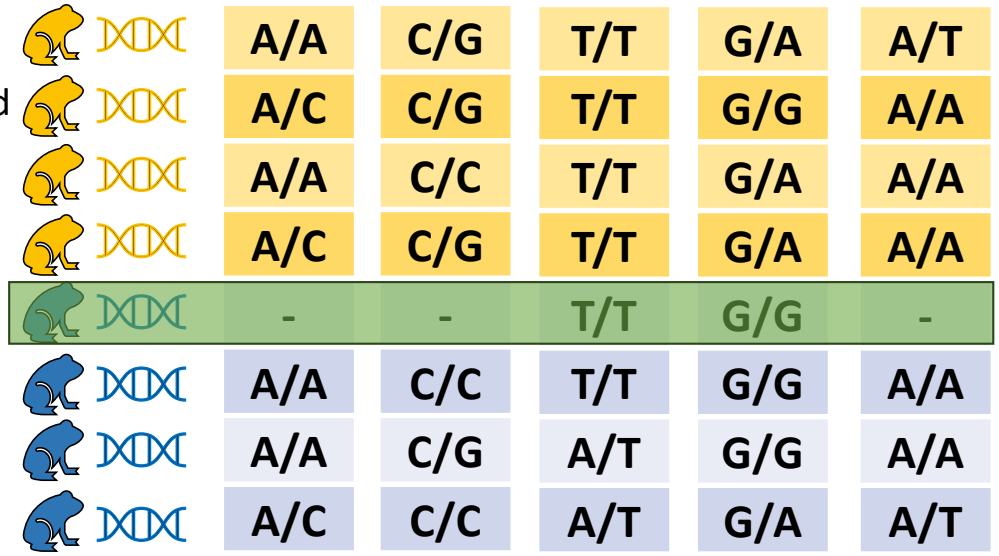
Filtering for call rate (method = “ind”)

What is it?

- Removes *individuals* that did not sequence to the specified level of completeness
- Why is it important?
  - Deletes the key thing you need to do a study
  - Removes individuals that may be misleading as they are sequencing outliers

Key Considerations:

- The filtering step that costs you the very most amount of money
- Generally not a first step – recommend filtering lightly here in the first go and seeing if other filters improve the individuals completeness



Keep up …

```
testset.gl

gl <- testset.gl
nInd(gl)
nLoc(gl)
nPop(gl)
popNames(gl)
indNames(gl)
locNames(gl)
table(pop(gl))
as.matrix(gl)[1:7,1:5]

gl <- testset.gl
gl.set.verbosity(3)
gl.report.callrate(gl)
gl.report.callrate(gl,method="ind")
gl.report.reproducibility(gl)

gl.report.callrate(gl,method="ind")
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80)

gl <- testset.gl
gl.smearplot(gl)
gl <- gl.filter.callrate(gl,verbose=0)
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80,
verbose=0)
gl.smearplot(gl)
```

Filtering for reproducibility (gl.filter.reproducibility)

What is it?

- Diversity Arrays duplicates individuals during library prep, to assess whether the same answer is found for every locus
- A control

- Why is it important?
  - Provides confidence in your base calls which is pretty fundamental

Key Considerations:

- If reproducibility is low, you need your SNPs re-assessed
- If doing sequencing through AGRF or other facilities, make sure to include duplicated individuals (this is an additional cost but do it anyway)

Keep up …

testset.gl

```
gl <- testset.gl
nInd(gl)
nLoc(gl)
nPop(gl)
popNames(gl)
indNames(gl)
locNames(gl)
table(pop(gl))
as.matrix(gl)[1:7,1:5]

gl <- testset.gl
gl.set.verbosity(3)
gl.report.callrate(gl)
gl.report.callrate(gl,method="ind")
gl.report.reproducibility(gl)

gl.report.callrate(gl,method="ind")
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80)

gl <- testset.gl
gl.smearplot(gl)
gl <- gl.filter.callrate(gl,verbose=0)
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80,
verbose=0)
gl.smearplot(gl)
```

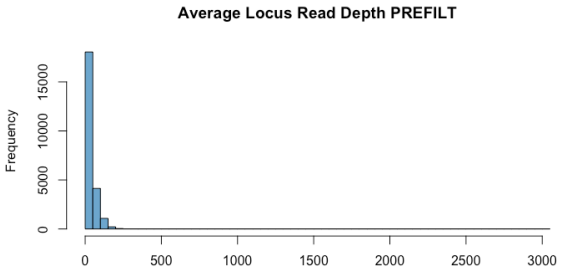
Filtering for read depth (gl.filter.rdepth)

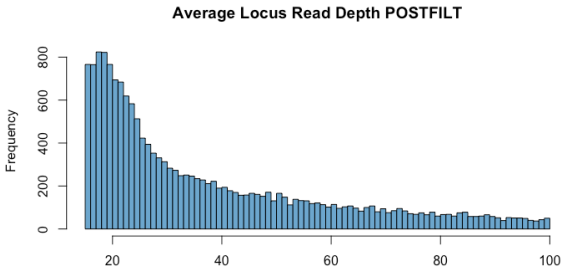
What is it?

- The mean number of sequencing reads for a particular locus, across all individuals
- Why is it important?
  - Number of reads in a stack tells you how confident you can be in your base calls
  - A low read depth means that your heterozygous sites have pretty low coverage

Key Considerations:

- If your read depth is not high enough, there are questions you can't confidently answer (e.g., heterozygosity)
- Very high read depth suggests paralogs (genes with multiple copies) being assembled in to one fragment





Keep up …

testset.gl

gl <- testset.gl
nInd(gl)
nLoc(gl)
nPop(gl)
popNames(gl)
indNames(gl)
locNames(gl)
table(pop(gl))
as.matrix(gl)[1:7,1:5]

gl <- testset.gl
gl.set.verbosity(3)
gl.report.callrate(gl)
gl.report.callrate(gl,method="ind")
gl.report.reproducibility(gl)

gl.report.callrate(gl,method="ind")
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80)

gl <- testset.gl
gl.smearplot(gl)
gl <- gl.filter.callrate(gl,verbose=0)
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80,
verbose=0)
gl.smearplot(gl)

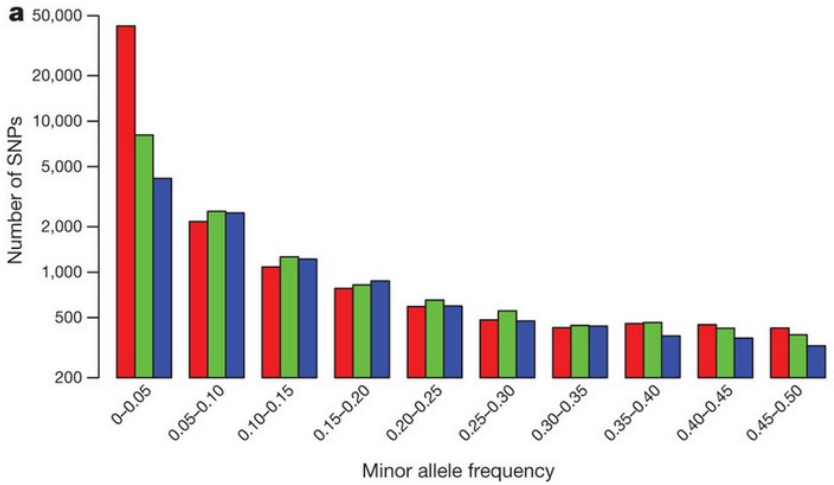
Filtering out minor alleles (gl.filter.maf)

What is it?

- Minor allele frequency is the frequency of the second most common allele in a given population
- Removing SNPs based on their relative proportion
- In the literature, often 0.05 BUT…

Key Considerations:

- 0.05 can be high if you have lot of individuals
  - If you have 300 diploid individuals, then you delete SNPs with less than 30 copies of the minor allele
- Or singleton if you have a small number of individuals
- Should depend on your question
  - Rare alleles can be very important to certain questions
  - Most SNPs are rare alleles, as are most heterozygous sites
  - Rare alleles are important in expansion processes
  - NOT important in structure or phylogenetic analyses



https://www.nature.com/articles/nature11174

Keep up …

```
testset.gl

gl <- testset.gl
nInd(gl)
nLoc(gl)
nPop(gl)
popNames(gl)
indNames(gl)
locNames(gl)
table(pop(gl))
as.matrix(gl)[1:7,1:5]

gl <- testset.gl
gl.set.verbosity(3)
gl.report.callrate(gl)
gl.report.callrate(gl,method="ind")
gl.report.reproducibility(gl)

gl.report.callrate(gl,method="ind")
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80)

gl <- testset.gl
gl.smearplot(gl)
gl <- gl.filter.callrate(gl,verbose=0)
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80,
verbose=0)
gl.smearplot(gl)
```

Filtering secondaries (gl.filter.secondaries)

What is it?

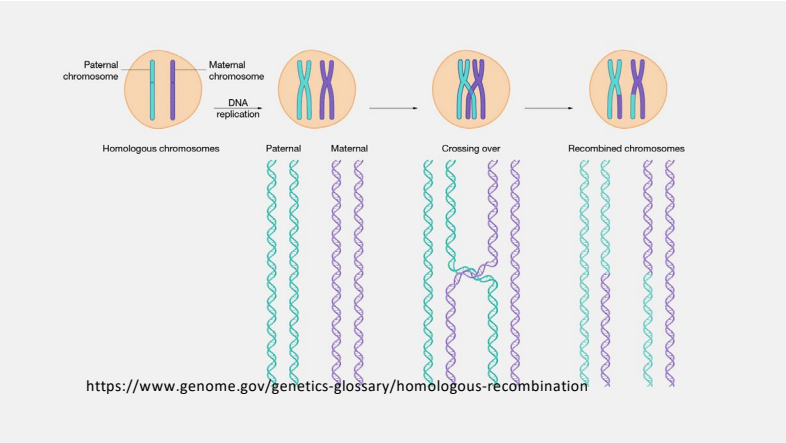
- When there are two SNPs on a single fragment, choosing to keep only one.

- Why is it important?
  - Loci that are very close together in the genome are not independently inherited as they are too close together for recombination to split.

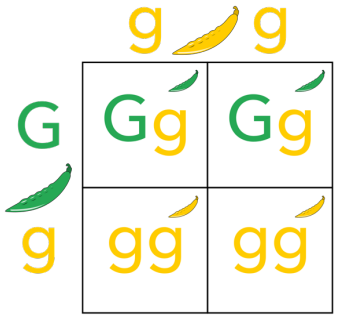
Key Considerations:

- Can mislead structure and phylogenetic analyses
- Can improve representations of heterozygosity

Recombination



Genetic diversity



OFFICIAL

Keep up …

testset.gl

```
gl <- testset.gl
nInd(gl)
nLoc(gl)
nPop(gl)
popNames(gl)
indNames(gl)
locNames(gl)
table(pop(gl))
as.matrix(gl)[1:7,1:5]

gl <- testset.gl
gl.set.verbosity(3)
gl.report.callrate(gl)
gl.report.callrate(gl,method="ind")
gl.report.reproducibility(gl)

gl.report.callrate(gl,method="ind")
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80)

gl <- testset.gl
gl.smearplot(gl)
gl <- gl.filter.callrate(gl,verbose=0)
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80,
verbose=0)
gl.smearplot(gl)
```

What's the right filtering order?

- **There isn't**
- **Be iterative – test different options**
- **Might do the same filter twice (call rate)**

As a starting point I would get rid of loci I don't believe in and then individuals that didn't work properly at all:

- gl.filter.callrate(gl, method = "loc", threshold = 0.7)
  - *Get rid of really poorly sequenced loci*
  - *But don't cut hard*
- gl.filter.callrate(gl, method = "ind", threshold = 0.25)
  - *Very low filter – this is only to get rid of your really bad individuals*
- gl.filter.monomorphs(gl)
  - *Always run this after removing individuals – removes loci that are no longer variable*
- gl.filter.reproducibility(gl)
  - *Get rid of unreliable loci*
- gl.filter.rdepth(gl, lower = X, higher = X)
  - *Get rid of low and super high read depth loci*

Keep up …

```
testset.gl

gl <- testset.gl
nInd(gl)
nLoc(gl)
nPop(gl)
popNames(gl)
indNames(gl)
locNames(gl)
table(pop(gl))
as.matrix(gl)[1:7,1:5]

gl <- testset.gl
gl.set.verbosity(3)
gl.report.callrate(gl)
gl.report.callrate(gl,method="ind")
gl.report.reproducibility(gl)

gl.report.callrate(gl,method="ind")
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80)

gl <- testset.gl
gl.smearplot(gl)
gl <- gl.filter.callrate(gl,verbose=0)
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80,
verbose=0)
gl.smearplot(gl)
```

What's the right filtering order?

- **There isn't**
- **Be iterative – test different options**
- **Might do the same filter twice (call rate)**

Then I would filter more strongly as appropriate for my question. For a population structure analysis I would,

- gl.filter.callrate(gl, method = "loc", threshold = 0.95)
  - *Structure dislikes missing data*
- gl.filter.maf(gl, threshold = 1/(2*nInd(gl))
  - *I usually set up the threshold so it is just removing singletons to improve computation time*
- gl.filter.secondaries(gl)
  - *Always do this as the last loci filter so that you've cut for quality before you cut because there are two SNPs*
- gl.filter.callrate(gl, method = "ind", threshold = .9)
  - *Filter on individuals. You can usually be a bit flexible at this point.*
- gl.filter.monomorphs(gl)
  - *Always run this after removing individuals*

Keep up …

```
testset.gl

gl <- testset.gl
nInd(gl)
nLoc(gl)
nPop(gl)
popNames(gl)
indNames(gl)
locNames(gl)
table(pop(gl))
as.matrix(gl)[1:7,1:5]

gl <- testset.gl
gl.set.verbosity(3)
gl.report.callrate(gl)
gl.report.callrate(gl,method="ind")
gl.report.reproducibility(gl)

gl.report.callrate(gl,method="ind")
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80)

gl <- testset.gl
gl.smearplot(gl)
gl <- gl.filter.callrate(gl,verbose=0)
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80,
verbose=0)
gl.smearplot(gl)
```

What is right for one situation is probably wrong for a different situation

- Be flexible
- Know what each filter is doing to your data
- Think carefully about whether the filter is appropriate to the test you want to run
- Analyse your data many different ways
- Don’t over-interpret your PCoA

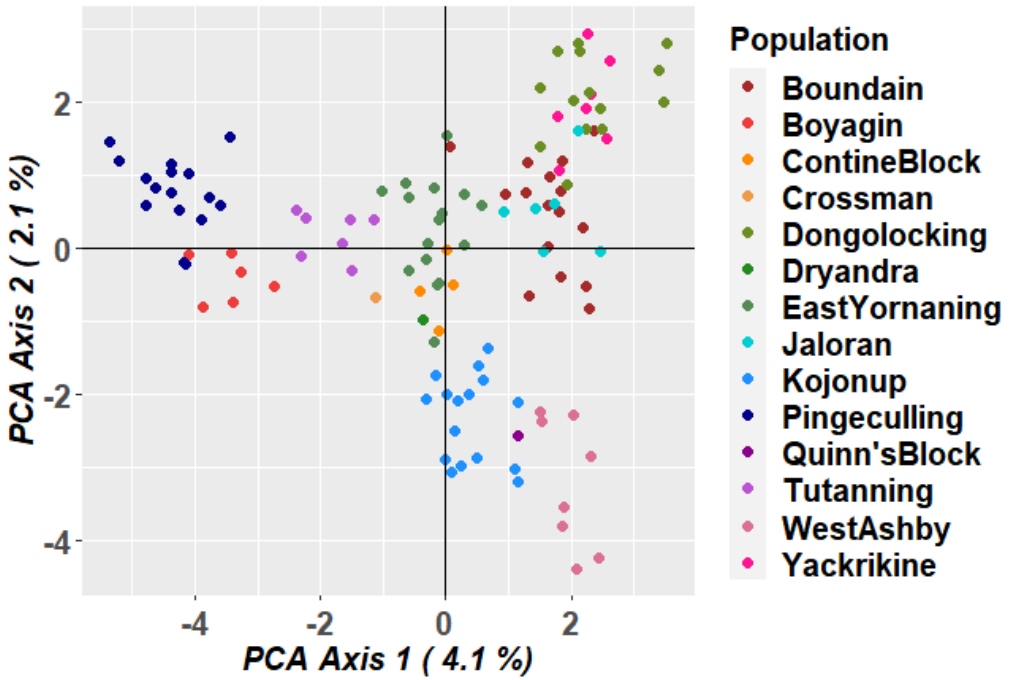
Keep up …

testset.gl

```
gl <- testset.gl
nInd(gl)
nLoc(gl)
nPop(gl)
popNames(gl)
indNames(gl)
locNames(gl)
table(pop(gl))
as.matrix(gl)[1:7,1:5]

gl <- testset.gl
gl.set.verbosity(3)
gl.report.callrate(gl)
gl.report.callrate(gl,method="ind")
gl.report.reproducibility(gl)

gl.report.callrate(gl,method="ind")
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80)

gl <- testset.gl
gl.smearplot(gl)
gl <- gl.filter.callrate(gl,verbose=0)
gl <- gl.filter.callrate(gl,
method="ind", threshold=0.80,
verbose=0)
gl.smearplot(gl)
```

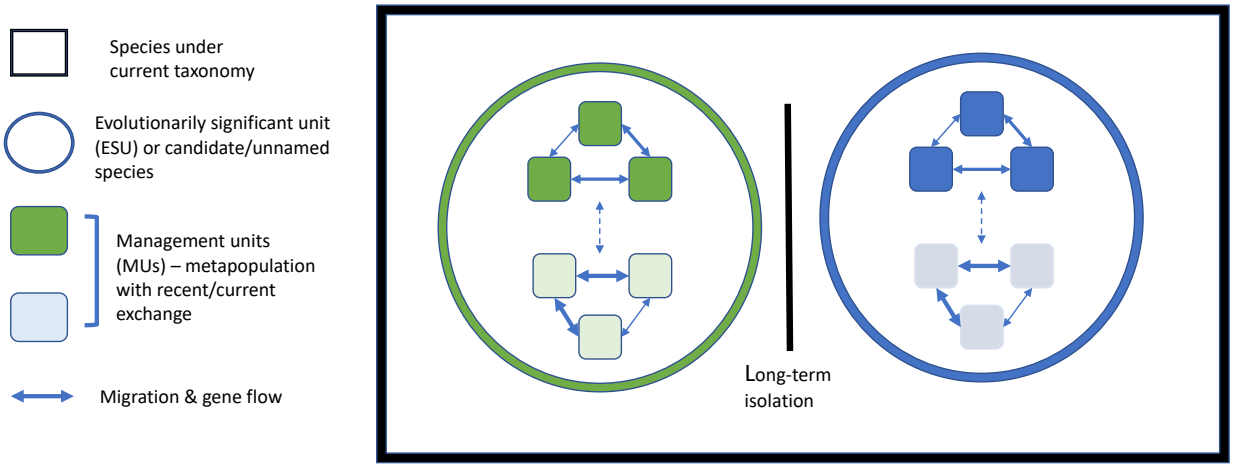
Filtering for your conservation question

- Filter as appropriate to the theoretical assumptions of what you are calculating
- Don’t over-interpret your output
  - Know what you can and can’t know from any particular analysis



Q1: Delimiting conservation units

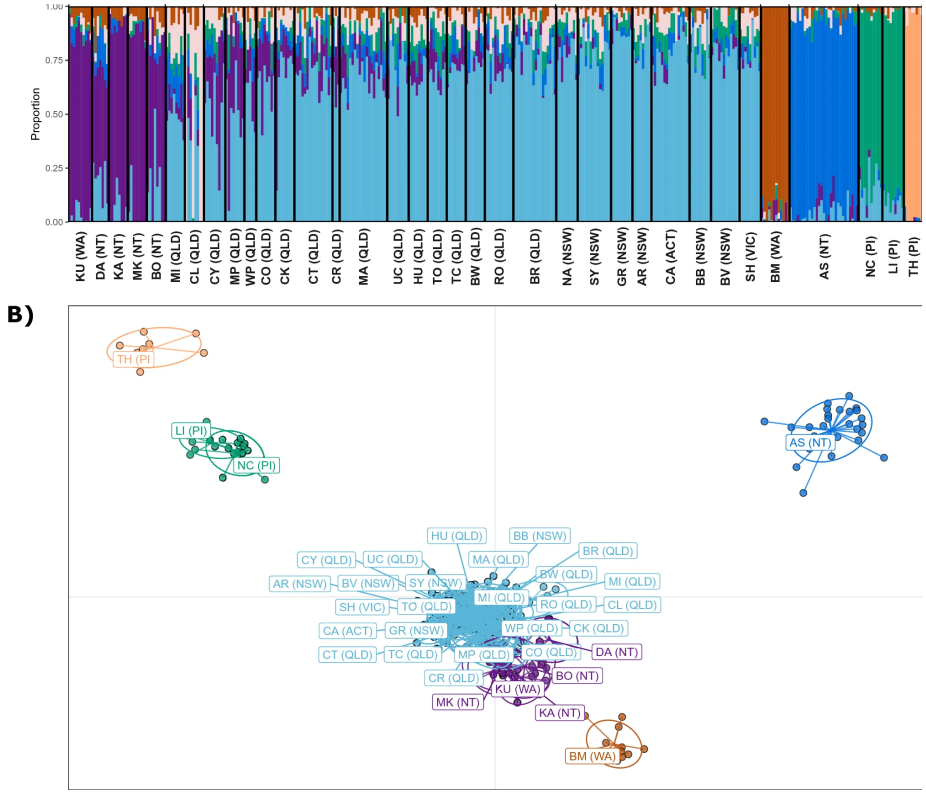
Intraspecific units for conservation and management



Q1. Delimiting conservation units

Is there gene flow?

- Structure or structure-like analyses
- Phylogenetics
- Fixed difference analysis/private alleles
- PCoA



Q1. Delimiting conservation units

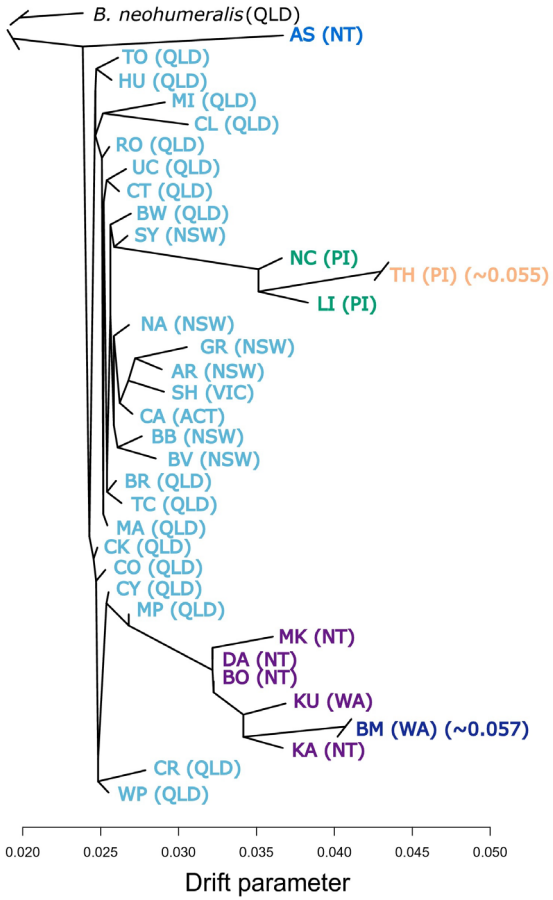
Is there gene flow?

- Structure or structure-like analyses
- Phylogenetics
- Fixed difference analysis/private alleles
- PCoA

Genome-wide patterns of differentiation over space and time in the Queensland fruit fly

Ángel-David Popa-Báez, Renee Catullo, Siu Fai Lee, Heng Lin Yeap, Roslyn G. Mourant, Marianne Frommer, John A. Sved, Emily C. Cameron, Owain R. Edwards, Phillip W. Taylor & John G. Oakeshott

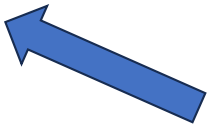
Scientific Reports **10**, Article number: 10788 (2020) | Cite this article



Fixed differences {gl.fd.sim}

- Fixed differences are genetic variants that are not shared between populations AND have no variation between populations
- The detection of fixed differences strongly supports a lack of gene flow between populations
- Whoo hooo! New species/subspecies/ESU!!!

```
> fd <- gl.fdsim(testset.gl[,1:100],poppair=c('EmsubRopeMata','EmmacBurnBara'),
+                sympatric=TRUE,verbose=3)
Starting gl.fdsim
  Processing genlight object with SNP data
    Populations EmsubRopeMata vs EmmacBurnBara [sympatric]
    Sample sizes: 11 5
    No. of loci: 15
  Calculating false positive rate with 1000 replications. Please be patient
    Threshold minor allele frequency for generating a false positive: 0.02
    Estimated mean count of false positives: 0
    Estimated SD of the count of false positives: 0
    Prob that observed count of 3 are false positives: 0
Completed: gl.fdsim
```



It depends!

Hardy-Weinberg Equilibrium

Genetic variation in a population will remain constant from one generation to the next in the absence of disturbing factors

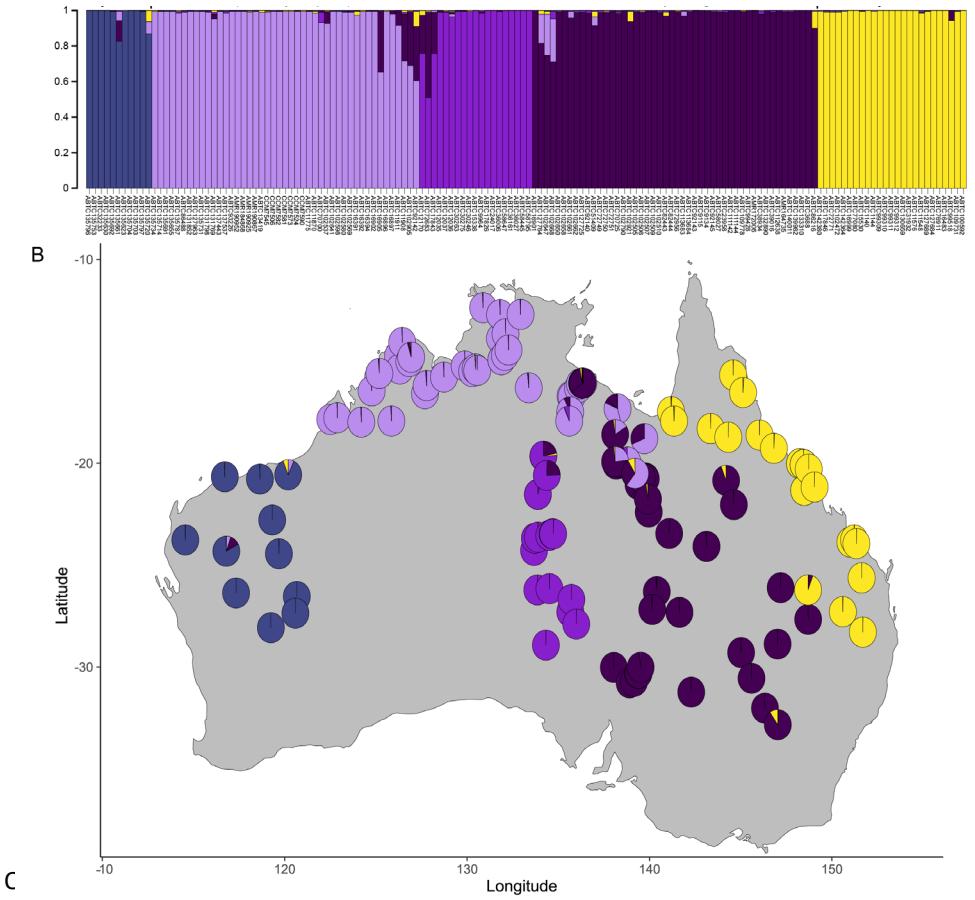
<u>Disturbing Factors</u>

- Mutations
- Natural selection
- Nonrandom mating (i.e. inbreeding or sexual selection)
- Genetic drift due to small population size
- Gene flow

Highly structured species in HWE? Then probably different!

- Is your species:
  - Widespread?
  - Common?

- Species delimitation in *Litoria rubella*
  - Seriously abundant and widespread
  - Lack of gene flow between lineages in eastern Australia



Not in HWE? Possibly not different

- Of particular concern for misuse in conservation genetics
- Mis-identification of *inbred and genetically drifted* populations as discrete conservation units
- Caused by incorrectly thinking analyses show the *causal factor* behind a lack of gene flow

<u>Disturbing Factors</u>

- Mutations
- Natural selection
- Nonrandom mating (i.e. inbreeding or sexual selection)
- Genetic drift due to small population size
- Gene flow

Q1: Identifying conservation units

Q1a: Is there gene flow?

- Structure or structure-like analyses
- Phylogenetics
- Fixed difference analysis/private alleles
- PCoA

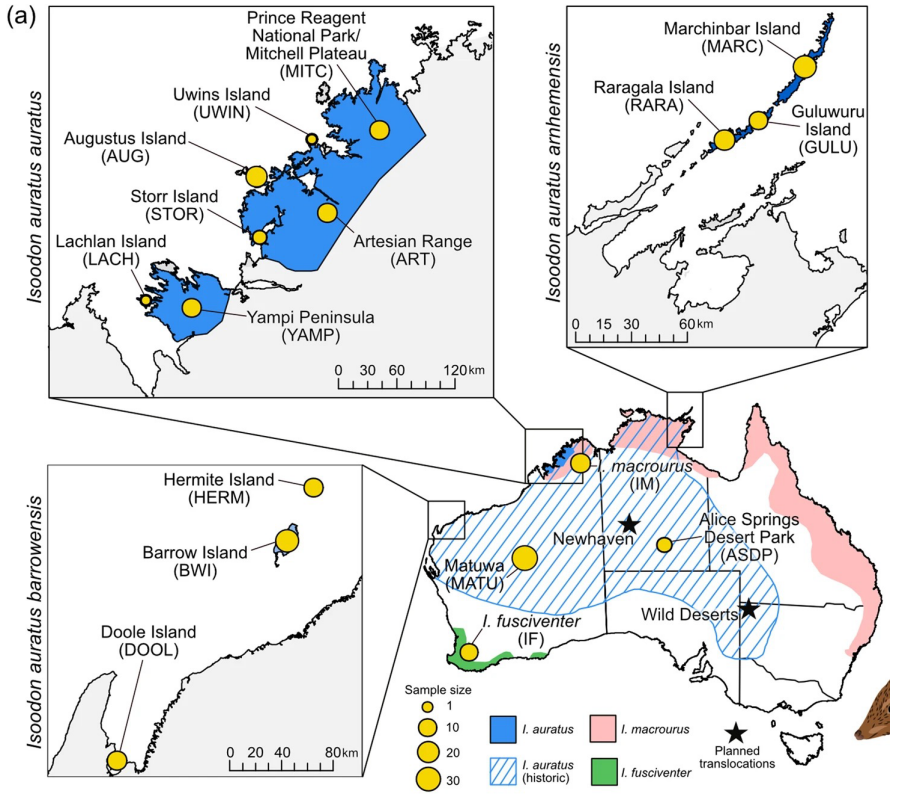
Q1b: Why isn't there gene flow?

- Treemix
- Relatedness
- Biological history/species

Population genomic diversity and structure in the golden bandicoot: a history of isolation, extirpation, and conservation

Kate Rick, Margaret Byrne, Skye Cameron, Steve J. B. Cooper, Judy Dunlop, Brydie Hill, Cheryl Lohr, Nicola J. Mitchell, Craig Moritz, Kenny J. Travouillon, Brenton von Takach & Kym Ottewell

Heredity **131**, 374–386 (2023) | Cite this article



OFFICIAL https://www.nature.com/articles/s41437-023-00653-2

Q1: Identifying conservation units

Q1a: Is there gene flow?

- Structure or structure-like analyses
- Phylogenetics
- Fixed difference analysis/private alleles
- PCoA

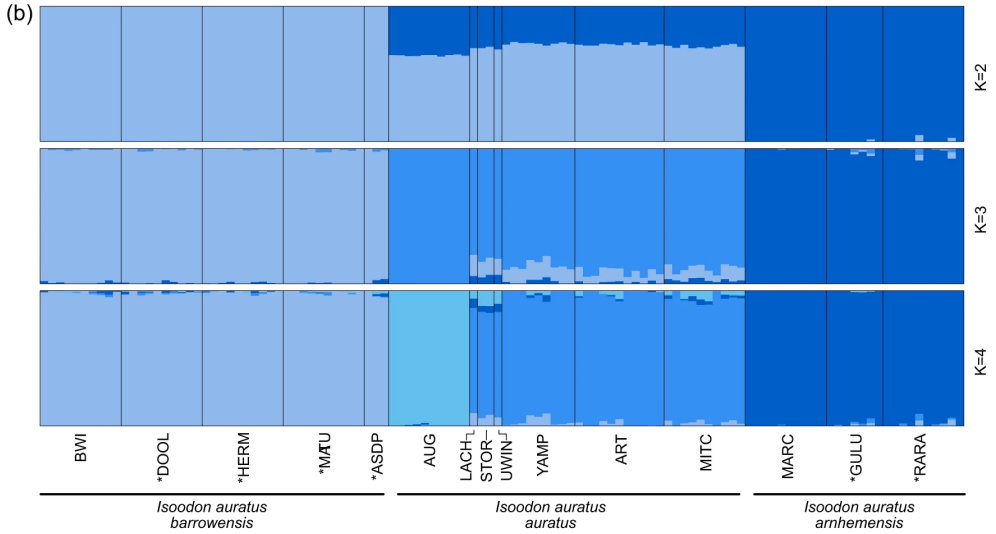
Q1b: Why isn't there gene flow?

- Treemix
- Relatedness
- Biological history/species

Population genomic diversity and structure in the golden bandicoot: a history of isolation, extirpation, and conservation

Kate Rick, Margaret Byrne, Skye Cameron, Steve J. B. Cooper, Judy Dunlop, Brydie Hill, Cheryl Lohr, Nicola J. Mitchell, Craig Moritz, Kenny J. Travouillon, Brenton von Takach & Kym Ottewell

Heredity **131**, 374–386 (2023) | Cite this article

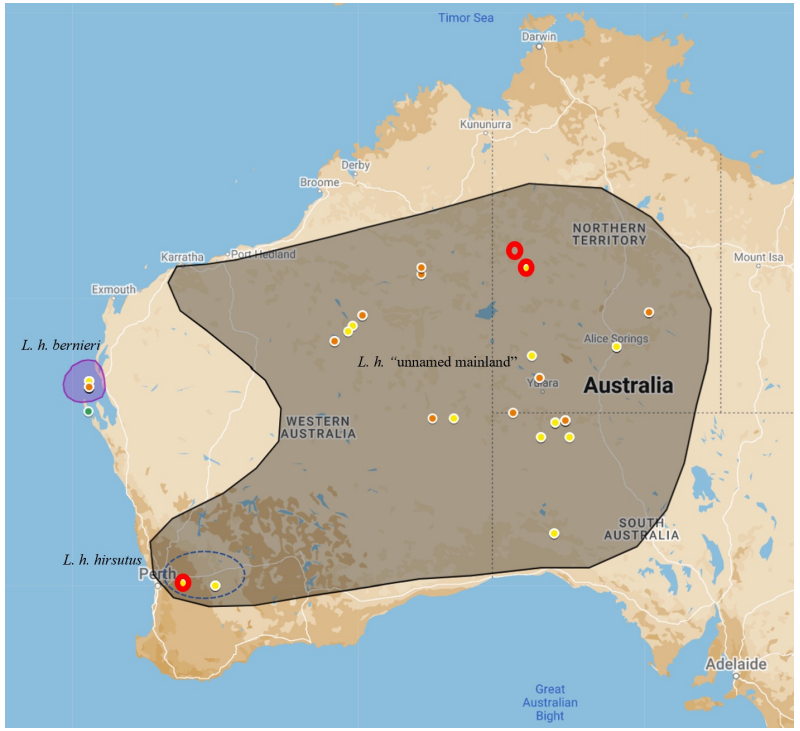


"Given that the Kimberly mainland (Artesian Range and Mitchell Plateau) as well as Barrow Island were key to retaining >90% allelic diversity in the species, we recommend that mixing these populations should be considered for future reintroductions."

OFFICIAL https://www.nature.com/articles/s41437-023-00653-2

Sampling versus Isolation By Distance

- In threatened species, it's not always possible to know whether things did or didn't admix
- Many species were morphologically variable
  - Bergmann's Rule
  - Allen's rule
  - Island rule
- As yourself:
  - Could they have mixed pre-decline and during the Pleistocene?
  - If you don't mix them, are could you be condemning them to inbreeding, drift, and extinction?



Filtering for population structure analyses

1. Filter out bad loci and poorly sequenced individuals

- gl.filter.callrate(gl, method = “loc”, threshold = 0.7)
  - *Get rid of really poorly sequenced loci*
  - *But don’t cut hard*
- gl.filter.callrate(gl, method = “ind”, threshold = 0.25)
  - *Very low filter – this is only to get rid of your really bad individuals*
- gl.filter.monomorphs(gl)
  - *Always run this after removing individuals – removes loci that are no longer variable*
- gl.filter.reproducibility(gl)
  - *Get rid of unreliable loci*
- gl.filter.rdepth(gl, lower = X, higher = X)
  - *Get rid of low and super high read depth loci*

Filtering for population structure analyses

1. Filter out bad loci and poorly sequenced individuals
2. Assess whether this analysis is sensitive to loci on the same fragment (whether recombination matters)
   - For structure and phylogenetics, yes

- gl.filter.callrate(gl, method = “loc”, threshold = 0.95)
  - *Structure dislikes missing data*
- gl.filter.maf(gl, threshold = 1/(2*nInd(gl))
  - *I usually set up the threshold so it is just removing singletons to improve computation time*
- gl.filter.rdepth(gl, lower = X, higher = X)
  - *Stronger read depth filter*
- gl.filter.secondaries(gl)
  - *Always do this as the last loci filter so that you’ve cut for quality before you cut because there are two SNPs*
- gl.filter.callrate(gl, method = “ind”, threshold = .9)
  - *Filter on individuals. You can usually be a bit flexible at this point.*
- gl.filter.monomorphs(gl)
  - *Always run this after removing individuals*

Filtering out singletons {gl.filter.maf}

- Common in the literature to say “we used a minor allele filter of 0.05”
- This means that the allele must be 5% of total alleles to be kept
- Fine to remove from phylogenetics and population structure analyses

Singleton is a single allele in one individual in a dataset

Q2: What is the genetic diversity/population history of my conservation unit?

- Heterozygosity – presence of two alleles at a gene locus, and a way to represent diversity
- Tajima's D - compares an observed nucleotide diversity against the expected diversity under the assumption that all polymorphisms are selectively neutral and constant population size
- Treemix – inference of a population-level tree that shows admixture and genetic drift
- Relatedness/Inbreeding - the likelihood that two individuals have the same gene as a result of common inheritance

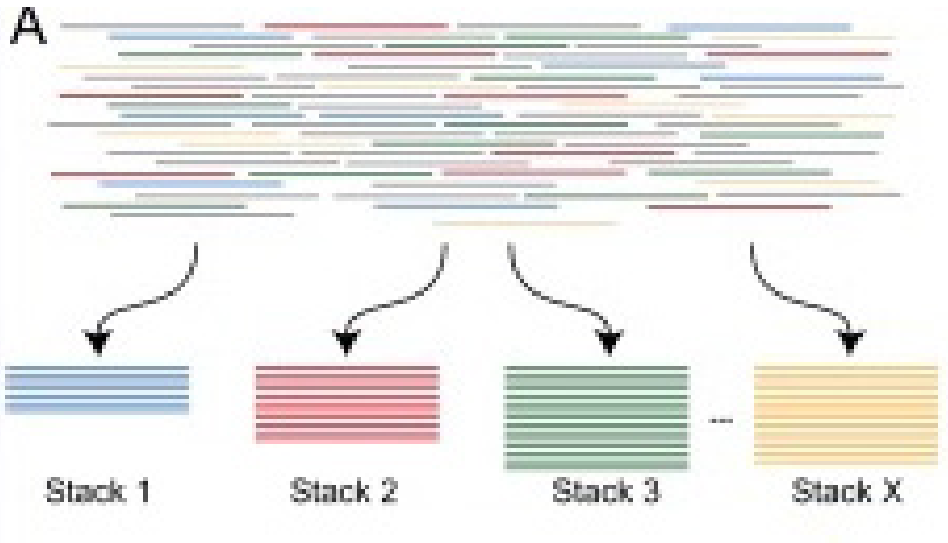
Removing rare alleles can substantially affect all these measures

AND

Not all populations are equally affected

The issue – read error

- If we need to keep our rare alleles, we need to be confident in them.
- This requires filtering hard on read depth
- Preferentially during the assembly stage to keep a certain read depth for every “stack”



***Stacks*: Building and Genotyping Loci *De Novo* From Short-Read Sequences**

Julian M Catchen, Angel Amores, Paul Hohenlohe, William Cresko, John H Postlethwait ✉ Author Notes

G3 Genes|Genomes|Genetics, Volume 1, Issue 3, 1 August 2011, Pages 171–182, https://doi.org/10.1534/g3.111.000240

OFFI

Filtering for diversity analyses

1. Filter out bad loci and poorly sequenced individuals
2. Assess whether this analysis is sensitive to loci on the same fragment (whether recombination matters)
   - For structure and phylogenetics, yes

- gl.filter.callrate(gl, method = "loc", threshold = 0.95)
  - *Structure dislikes missing data*
- ~~gl.filter.maf(gl, threshold = 1/(2*nInd(gl))~~
  - *Remove this filter to keep singletons*
- gl.filter.rdepth(gl, lower = X, higher = X)
  - *STRONG read depth filter – higher than for structure analyses*
- ~~gl.filter.secondaries(gl)~~
  - *I wouldn't do this because we're talking about diversity and recombination isn't expected*
- gl.filter.callrate(gl, method = "ind", threshold = .9)
  - *Filter on individuals. You can usually be a bit flexible at this point.*
- gl.filter.monomorphs(gl)
  - *Always run this after removing individuals*

Not all datasets can answer every question

….even if you can run the analysis and get an output

- You need sufficient population level sampling for demographic analyses
- You need sufficient read depth to keep rare alleles

Molecular Ecology

THE QUESTION(S) KNOW WHAT THE MANAGER AND YOU WANT:

THE POSSIBLE ANSWERS

DO DATASET & APP measure what you want:

DON’T DO THIS:

- See if your field data in the app gives any result at all (with no idea if it is correct)

WHAT THE MANAGER (AND YOU) WANT – SINGLE SPECIES
Population viability analysis

- The species lives in 2+ areas managed by different owners/agencies – can they be managed independently?

  m 'dispersal', MU 'management unit'

- The species has a huge range, are there obvious divisions?

  ESU

- Does small isolate have low genetic variation? Willoughby 2015

  GD/He/Shannon Information

- Does threat (eg pollinator extinct, new predator on dispersers) result in greater 'identity by descent'?

  IBD/Fis

- Are there alleles for adaptation to changed conditions?

  ID environment-specific alleles

Genetic Info & Manager's actions

- m 'dispersal', MU 'management unit' Neutral variant step? (Sentinella Safner 11 Taylor 97 Sherwin 2018)
  'm' across jurisdiction boundary eg coordinate if same MU
- ESU Neutral Phylogenetics
  ESUs high conservation value
- GD eg $H_e=1-\Sigma p^2$ also called ${}^2H$; ${}^1H=-\Sigma p*\log(p)$
  manage N-census; M:F; Family size µσ
- IBD eg $F_{is} = (H_e-H_o)/H_e$
  manage interaction between families (↑pollinators, ↓new predators)
- Adaptive alleles (environment-specific)
  areas with suitable alleles high conservation value

Assumed knowledge

Definitions sent to you before the course

One Alpha (within location) Genetic Diversity measure might need more explanation

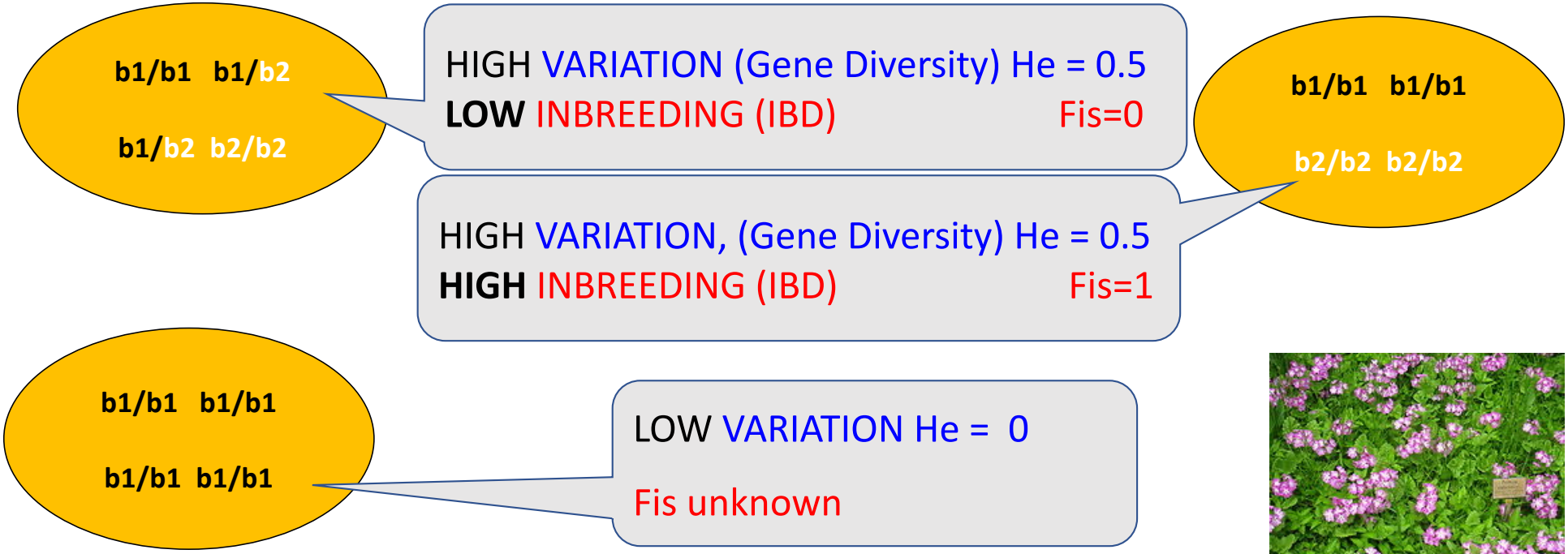
- Shannon Information $^1H=-\Sigma p*\log(p)$
- Commonest Freq-based Diversity measure for species, also used for genes now
- Video: https://ars.els-cdn.com/content/image/1-s2.0-S0169534717302550-mmc2.mp4
- Paper: Sherwin et al 2017,2021. TREE doi.org/10.1016/j.tree.2017.09.012 doi.org/10.1016/j.tree.2021.07.005
- Teaching app: WEB: https://evolutionaryecology.shinyapps.io/learningEE GITHUB: https://github.com/z5168021/InformationTheory

DO DATASET & APP measure what you want?

Documentation & algebra +/or try simulated or constructed data

EG Pitfall #1 : Gene diversity ≠ Fis

Inbreeding doesn't change the population's genetic diversity, but does change the way it is arranged into individuals Washitani 96



DO DATASET & APP measure what you want?

Documentation & algebra +/or try simulated or constructed data

EG Pitfall#2: F_{st} is a poor measure of differentiation $F_{st}=(H_t-H_s)/H_t$

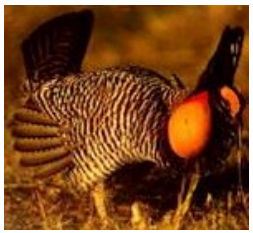
- compare with Shannon Differentiation = Mutual Info / Max MI

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 Fst still! Sherwin 20**

Gray Wolf adaptive genes 90% false discovery? Schweizer 2016

DO DATASET & APP measure what you want?



Documentation & algebra +/or try simulated or constructed data

EG Pitfall#3: "effective population size" N_e

- NOT a population size (~0.1* N-census to 2*N-census)
- Effect of demography on rate of change of H_e *or* F_{is} *or* F_{st} etc
  (different N_e for each of these)
- Calculate from demography (N-census; M:F ratio; Family size μσ)
- Back-calculate from rate of change of H_e or F_{is} or F_{st}

(REFS: Sherwin 1990 Engen, Frankham Holleley, NcMS appendix)

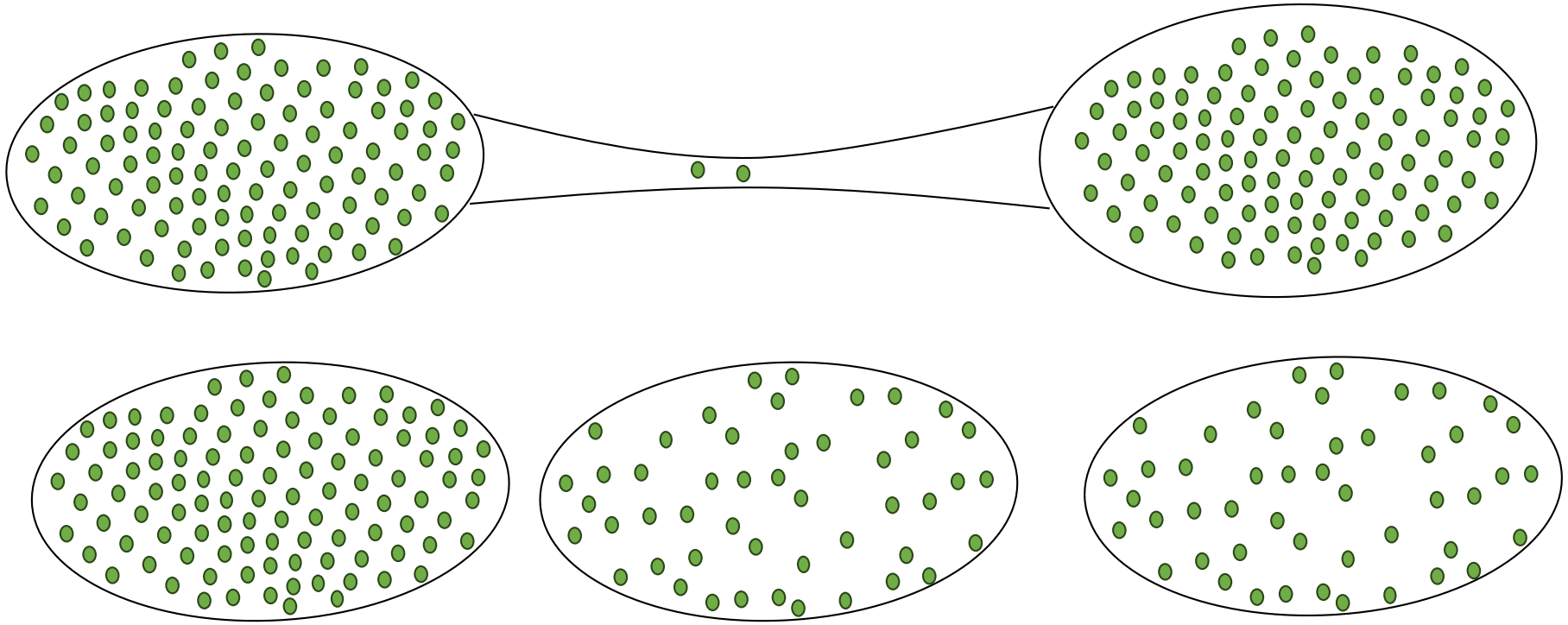
- Managing H_e via N_e : N-census; M:F ratio; Family size μσ,
  also manage via Dispersal [Refs: Westemeier 98 Madsen 00 Newman 97

Task

- You have been asked to evaluate the 'genetic health' of a population at a restoration site of the species A.
- Details are provided in the Task section in the .qmd file in the PopGenR project on the cloud along with how to access the data.
- Compute genetic diversity metrics (e.g. He, Ho, Fis) with the data generated by the 100 samples collected and provide a recommendation on whether the restoration project was successful and whether the studied population is likely to require continued active management in the short term.

Loss of genetic diversity → $1-\left(1-\frac{1}{2Ne}\right)^{t}$ ← Generations

(Falconer & Mackay 1996)

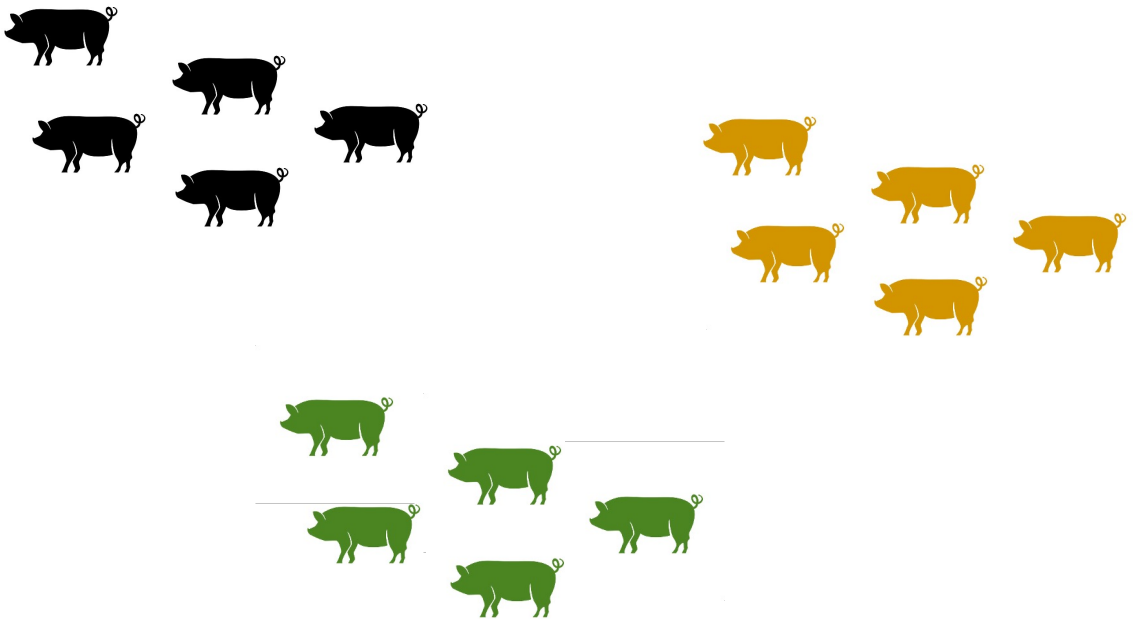


High Fis

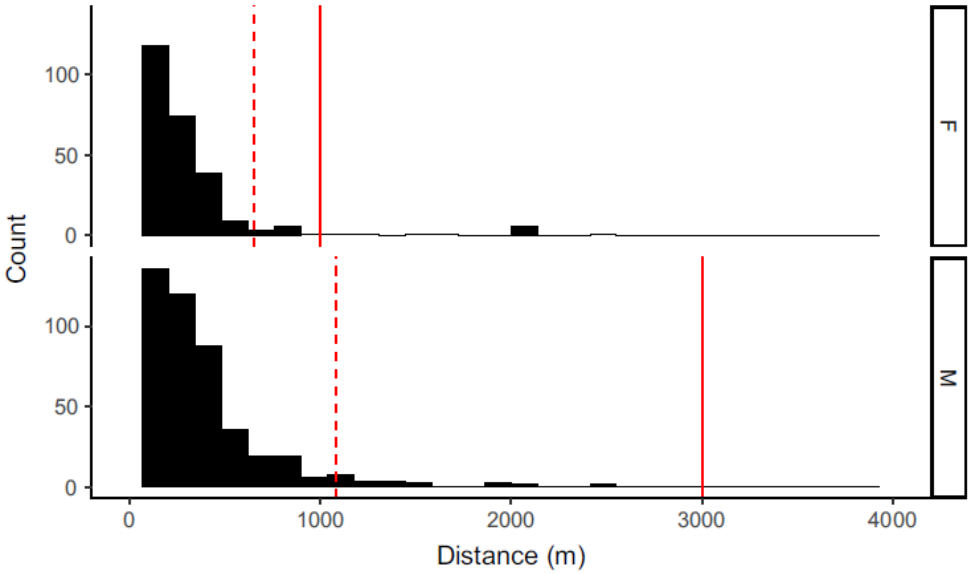
Recall that

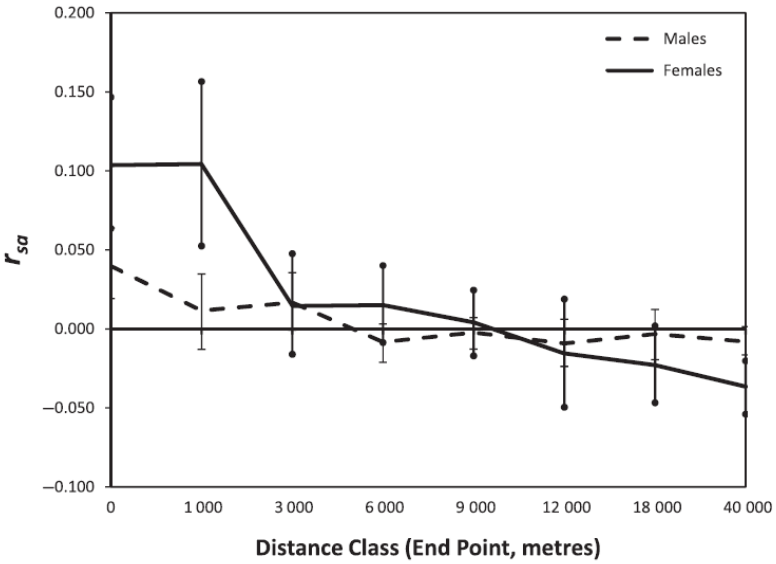
$$Fis = \frac{He - Ho}{He}$$

- Selective sampling of related individuals



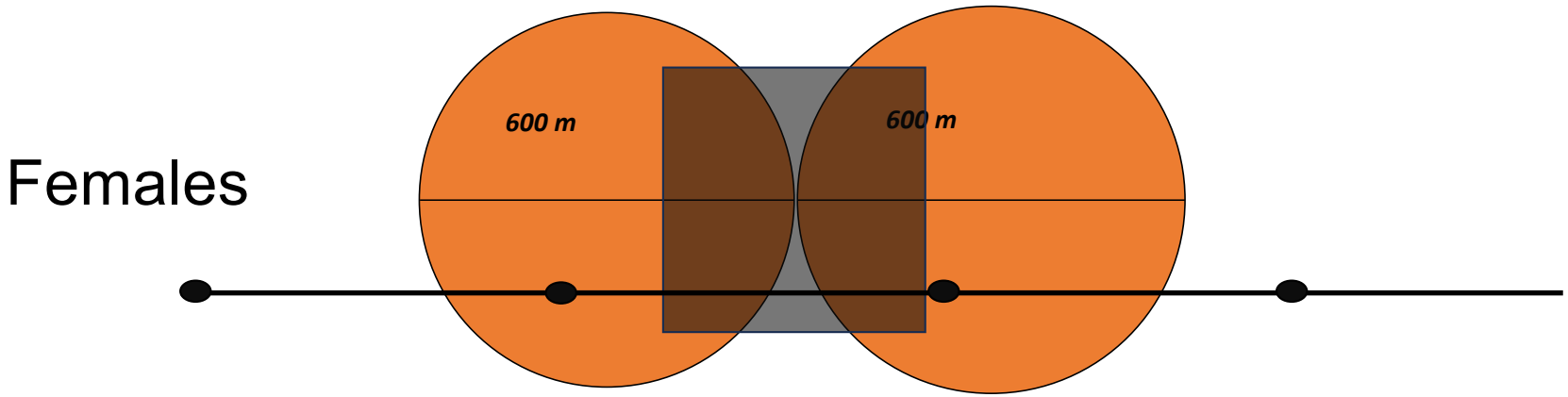
Dispersal - Woylies (*Bettongia penicillata*)





High Fis?

Woylies (*Bettongia penicillata*)



Back to the PopGenR project on the cloud, let’s have a look at an example

High Fis