



Detecting Selection

with EAA and Fst outlier tests



Rachael Dudaniec
Macquarie University



Workshop on Population and Speciation Genomics 2020

Practical learning goals:

- Become familiar with running differentiation-based and environmental association analyses (EAA)
- Apply different parameter sets and evaluate how different settings impact outlier detection.
- Plot and interpret results of different tests

Data for the practical come from:

Dudaniec RY, Yong CJ, Lancaster LT, Svensson EI, Hansson B (2018) Signatures of local adaptation along environmental gradients in a range-expanding damselfly (*Ischnura elegans*). *Molecular Ecology*. 27(11): 2576-2593

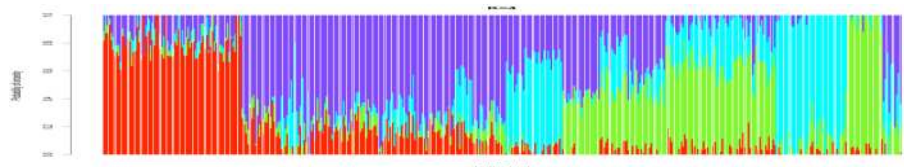
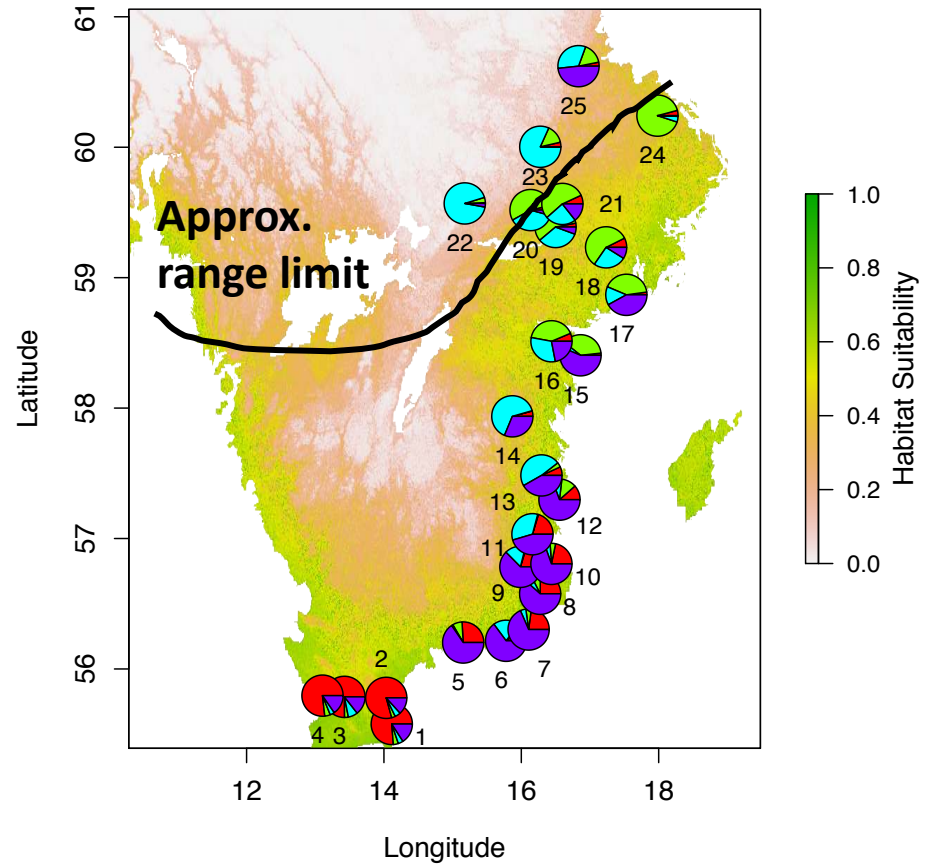


THE GENETIC DATA

Ischnura elegans –
Blue tailed damselfly

- 426 *I. elegans* (2013)
- RADseq/13,612 SNPs
- 500km gradient

4 admixed clusters



THE ENVIRONMENTAL DATA



Percentage tree cover



Mean Annual Temperature

Mean Maximum Temp Warmest Quarter (Summer)



Wind Speed



Mean Annual Precipitation



What we'll cover

Each Part has a separate R script file:

Part 1.0 The Data:

Load and examine genetic data, plot and correlate environmental variables

Part 2.0 Outliers:

Differentiation-based (F_{st}) outlier detection with *pcadapt* and *OutFLANK*

Part 3.0 Multivariate EAA:

Ordination approach using *Redundancy Analysis (RDA)*

Part 4.0 Univariate EAA:

Single locus tests with *Latent Factor Mixed Models (LFMM)*



PART 1.0 The Data

- 1) Visualise and examine the dataset
- 2) Plot and examine environmental variables
- 3) Check for correlations in environmental variables



PART 2.0. Outliers

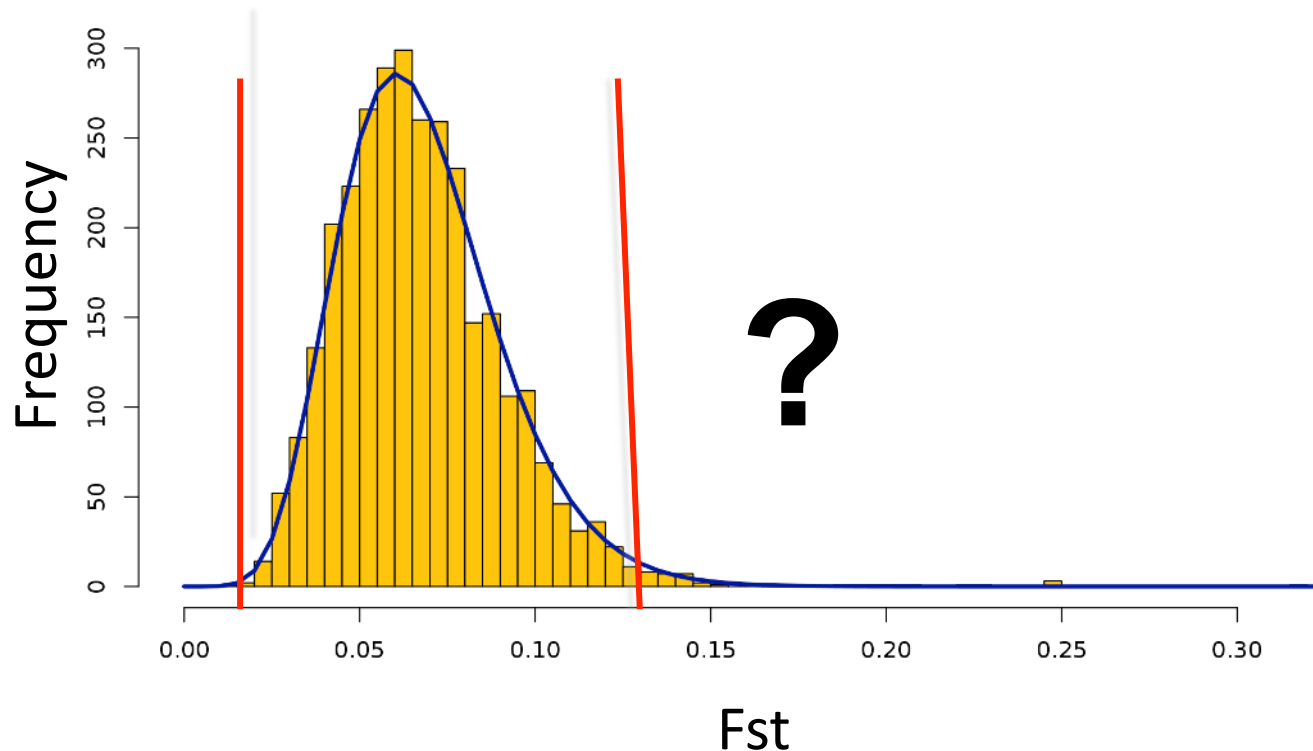
- 1) Run *OutFLANK* (Whitlock and Lotterhos 2015)
- 2) Modify parameters, fill in Table 1 – compare N outliers with different parameters:
 - Left and Right Trim Fraction
 - q- threshold (FDR rate)



Left and Right Trim Fraction

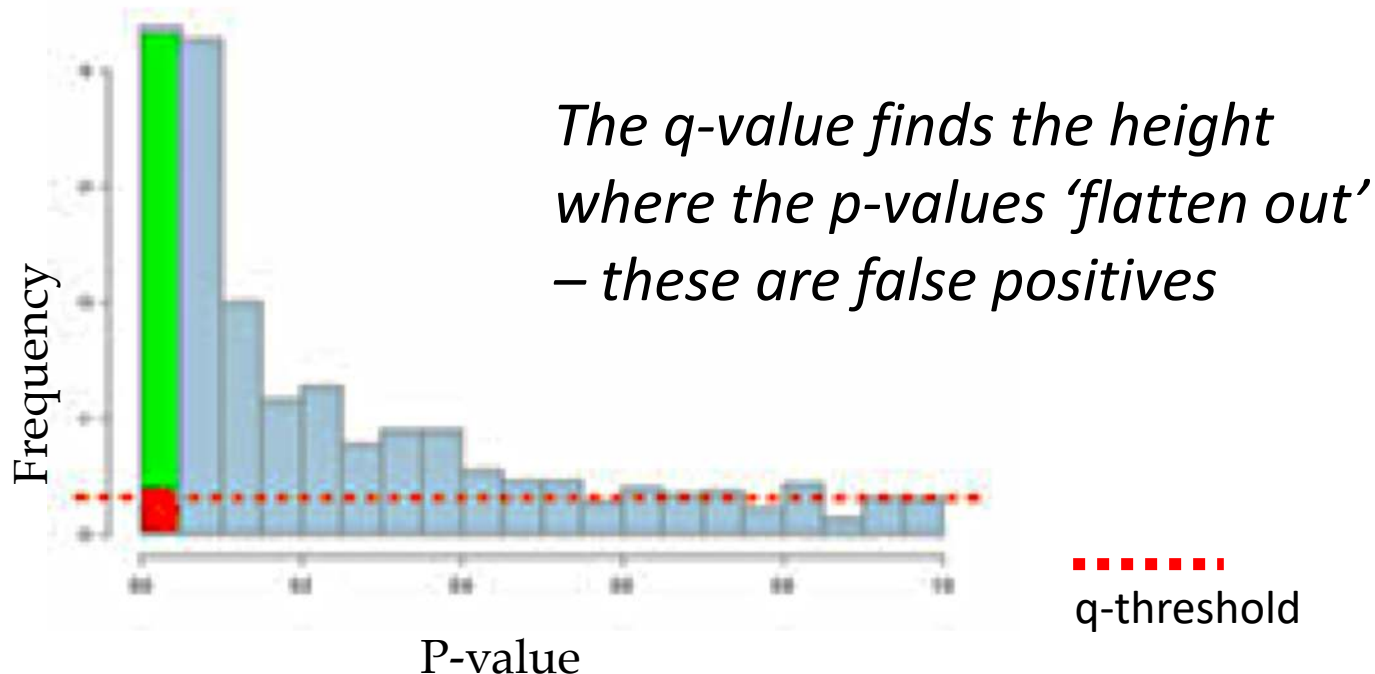
Tells OutFLANK what fraction of extreme Fst values to remove from the left and right tails before estimating the **null Fst distribution**

If you trim too few off the tails, it will result in a much wider chi-square distribution (smaller df) and fewer outliers called.



False discovery rate: 'q threshold'

- q sets the % threshold for determining an 'outlier' = a measure of the false discovery rate
- The q -value is calculated based on the right tail (+) p -values for each locus
- If a locus is below the threshold then it may be included as a candidate



PART 2.0. Outliers

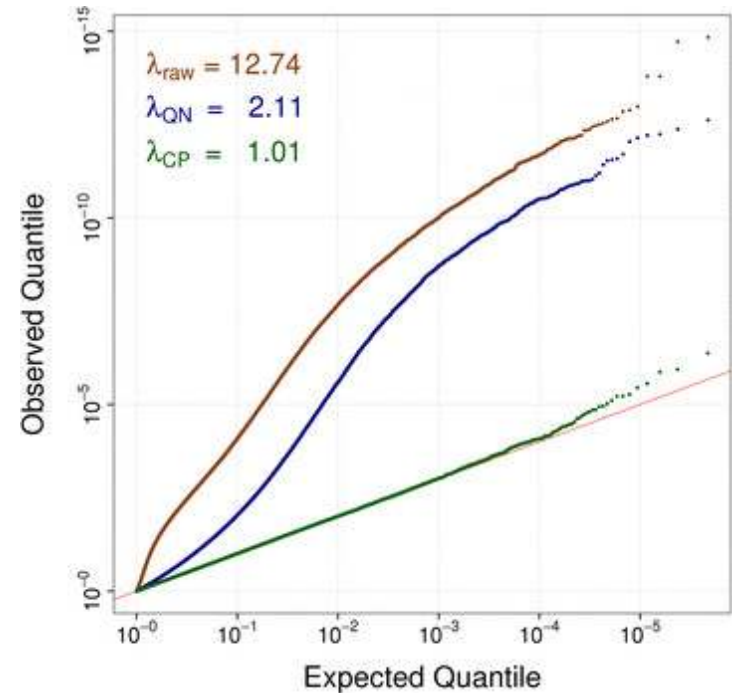
- 1) Run *OutFLANK* (Whitlock and Lotterhos 2015)
- 2) Modify parameters, fill in Table 1 – compare N outliers with different parameters:
 - Left and Right Trim Fraction
 - q- threshold (FDR rate)
- 3) Run *pcadapt*
 - Modify GIF – compare to original GIF
 - Define FDR (q-value cut off)
 - Examine N outliers



Genomic Inflation Factor (GIF)

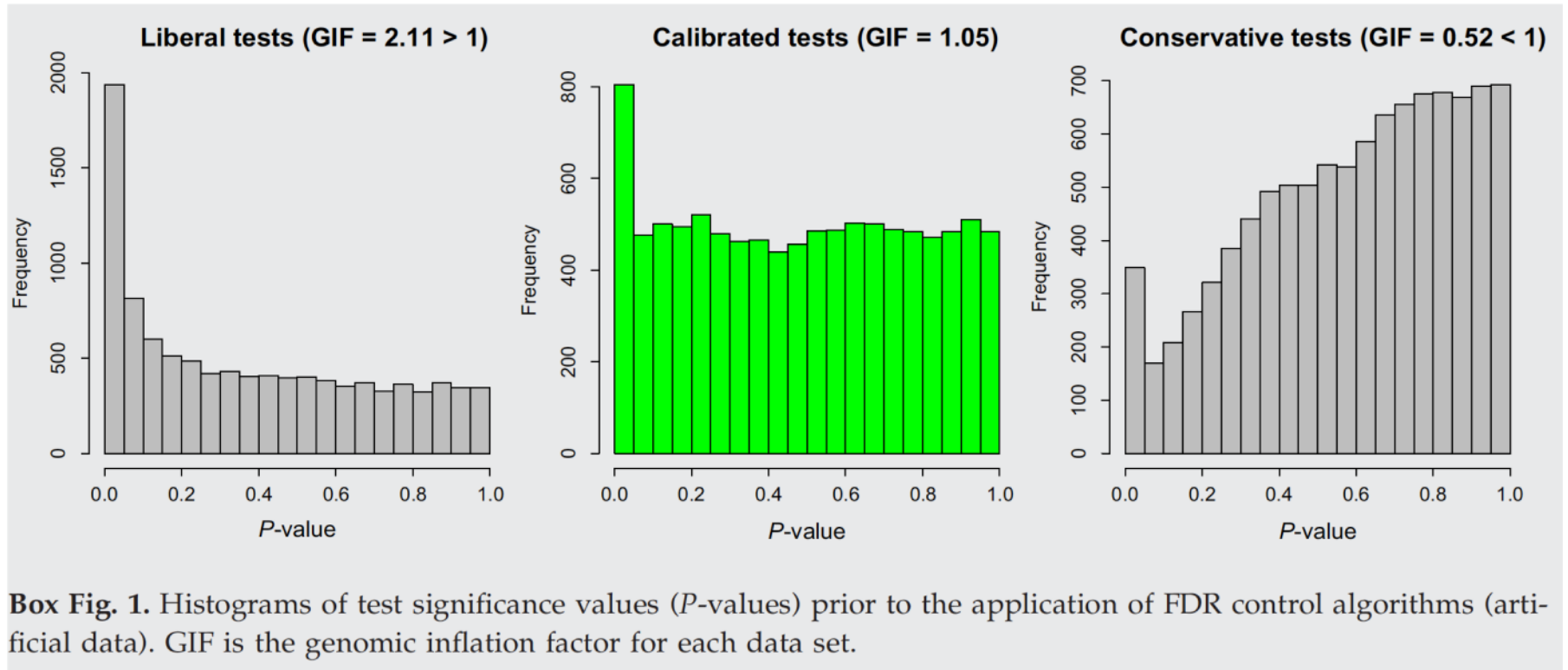
Used to recalibrate test scores (p-values) to control FDR

Expresses the **deviation** of the distribution of the observed test statistic from the distribution of the expected test statistic, **i.e. inflation of scores**



GIF can vary depending on sample size, relatedness, LD, population substructure, and N causal variants.

Calibrating P-values with GIF



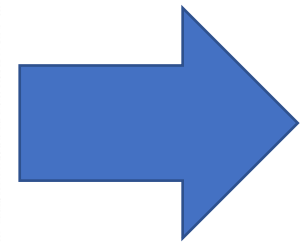
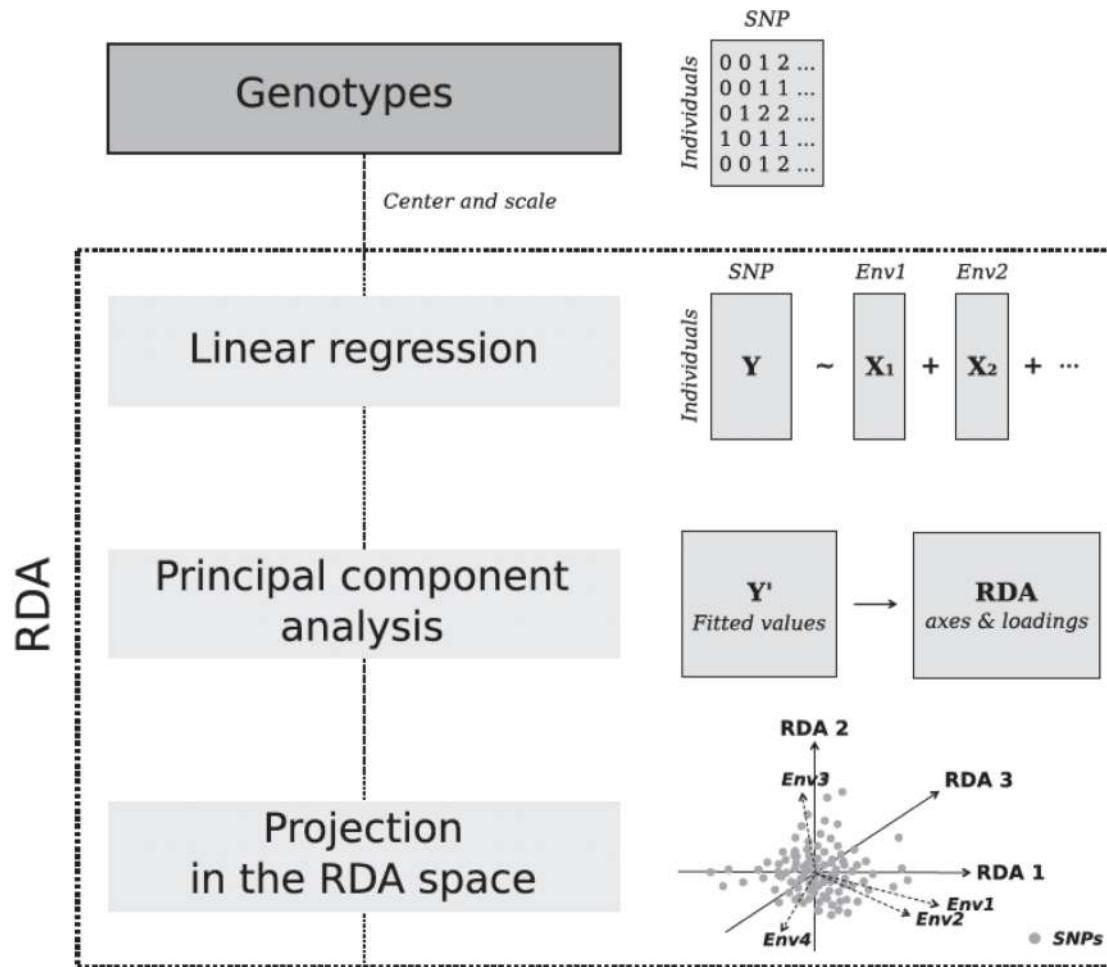
GIF $\sim$ 1= well calibrated

GIF > 1= liberal (too many small p-values)

GIF < 1= conservative (too many large p-values)

PART 3.0. Multivariate EAA

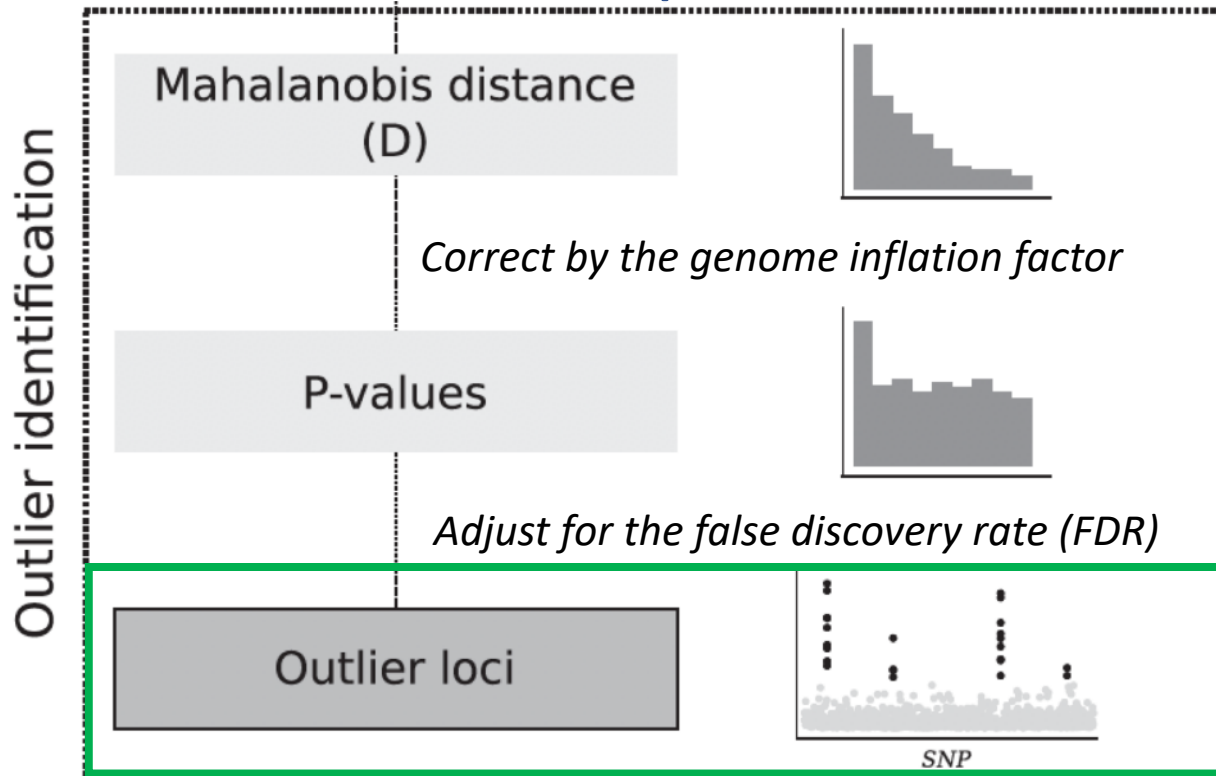
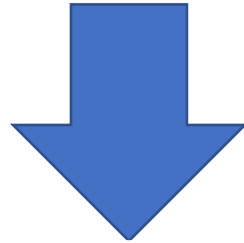
RDA – Redundancy Analysis
(R package **vegan**)



Recover loci loadings on K number of axes

PART 3.0. Multivariate EAA

RDA – Redundancy Analysis
(R package **vegan**)



= a multi-dimensional measure of how many SDs each locus is from the mean distribution

PART 3.0. Multivariate EAA

RDA – Redundancy Analysis
(R package `vegan`)

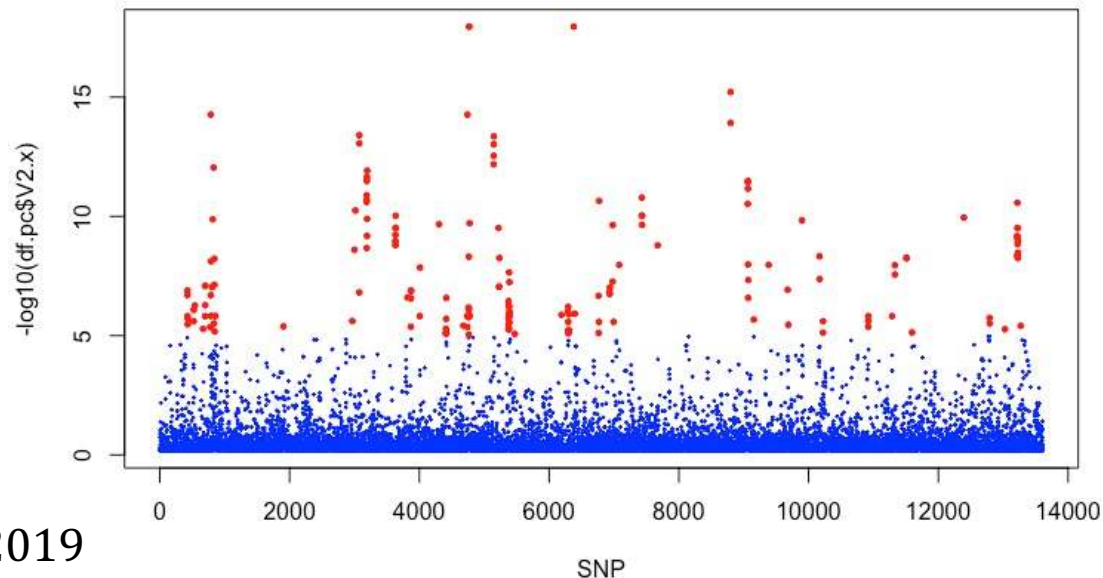
- 1) Identify outliers using *Mahalanobis distance* approach (Capblancq et al. 2018)
- 2) Compare default and modified GIFs
- 3) Identify outliers using a *P-value (SD) cut-off* approach
- 4) Correlate SD outliers with environmental predictors and plot



PART 4.0. Univariate EAA

Latent factor mixed modelling LFMM2

- Linear mixed model that uses 'K' genetic groups as latent factors (representing random effects)
- The environmental variables are the fixed effects
- Each SNP x ENV association is tested separately



See Caye et al. 2019



PART 4.0. Univariate EAA

Latent factor mixed modelling LFMM

1. Run LFMM
2. Examine GIF for each environmental predictor
3. Outlier detection – 1 variable at a time
4. Modify GIF and apply FDR cut-offs
5. *Optional:* Run a PCA on environmental predictors and re-run LFMM - identify PC outliers



What we'll cover

Each Part has a separate R script file:

Part 1.0 The Data:

Feel free to work with others. Parts 2-4 are independent so can be done in any order. Feel free to pick and choose, or do it all 😊

Ordination approach using Redundancy analysis (*RDA*)

Part 4.0 Univariate EAA:

Single locus tests with Latent Factor Mixed Models (*LFMM*)

