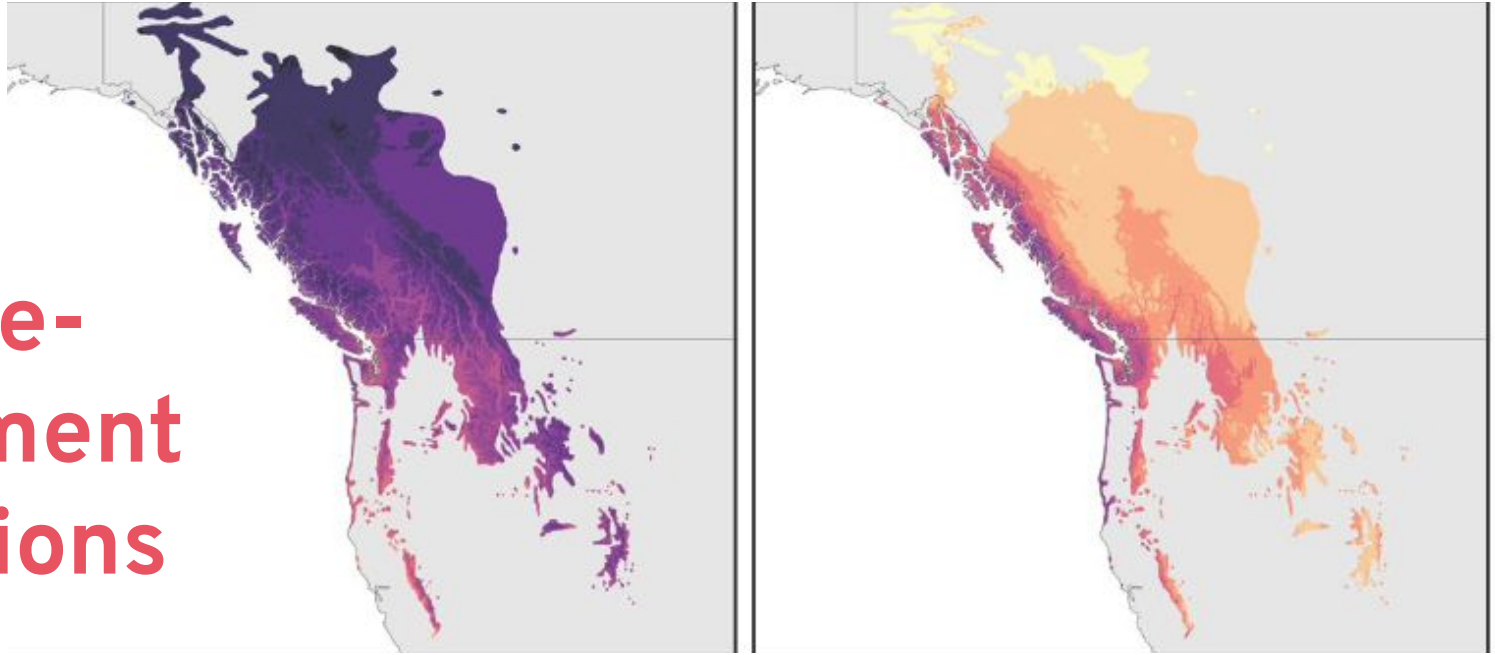
Genotype-environment associations

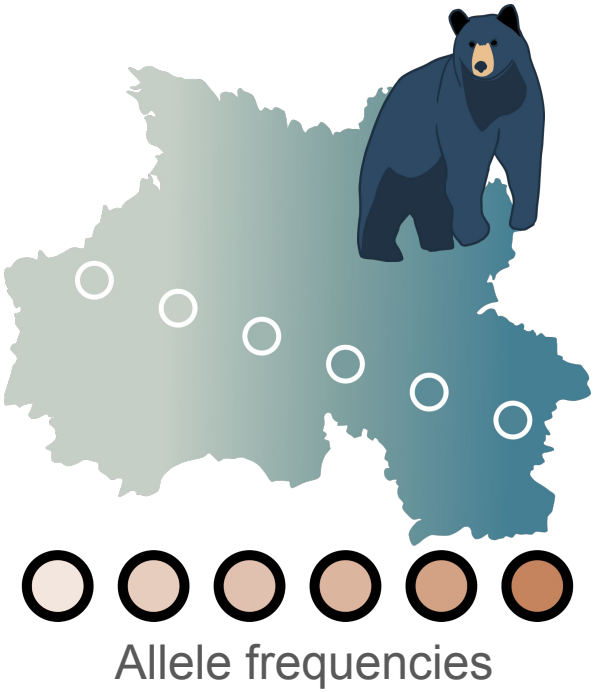


Capblancq & Forester (2021) *Methods Ecol. Evol.*

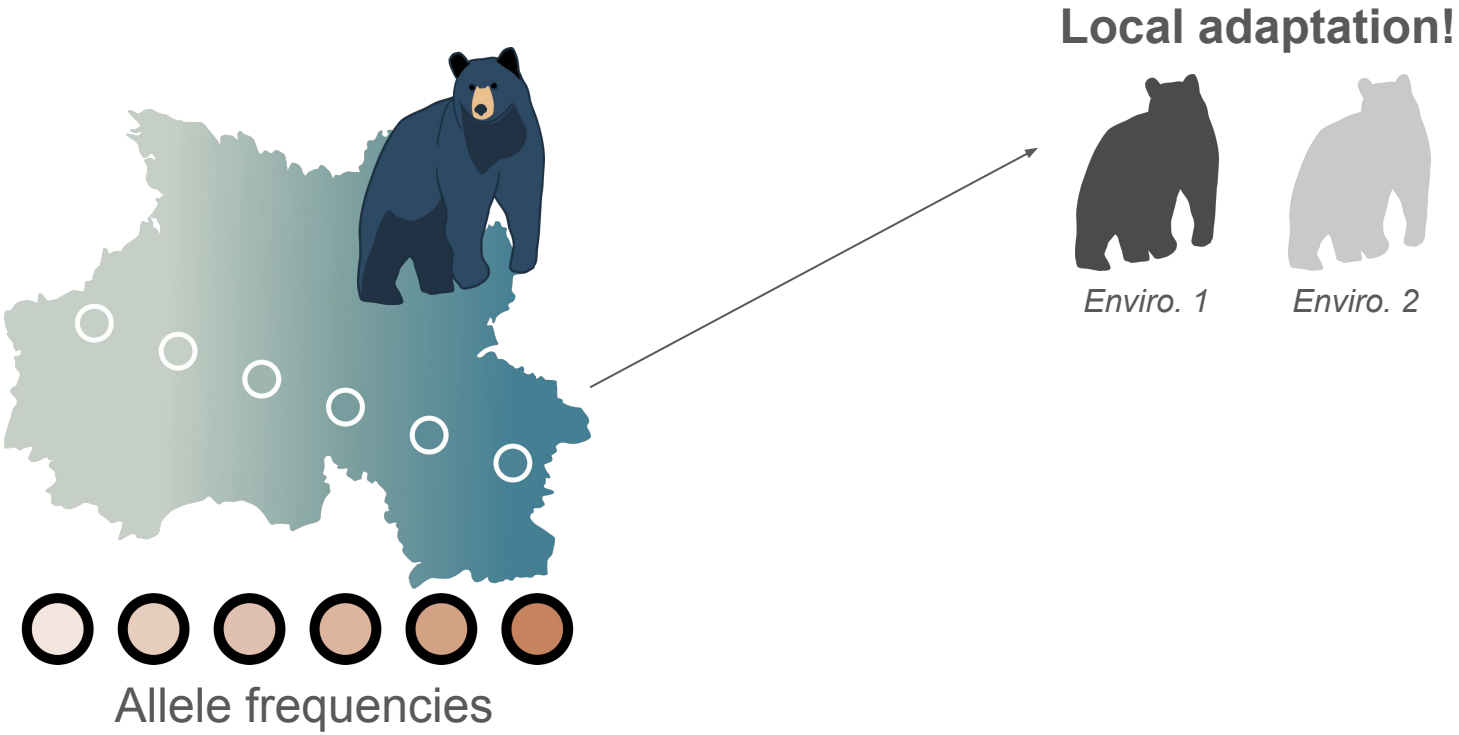
Anne Chambers & Anusha Bishop (2024)

Genotype-environment association (GEA) methods

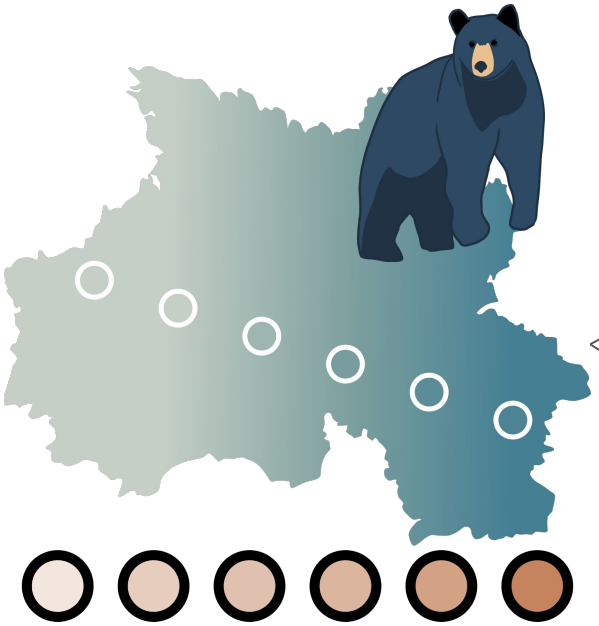
Genotype-environment association (GEA) methods



Genotype-environment association (GEA) methods

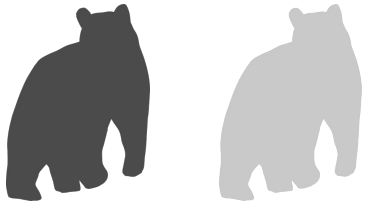


Genotype-environment association (GEA) methods



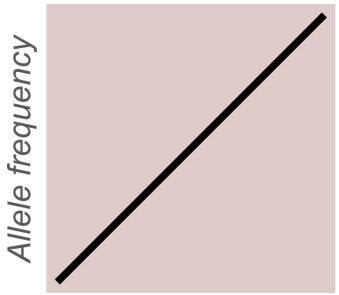
Allele frequencies

Local adaptation!



Enviro. 1 *Enviro. 2*

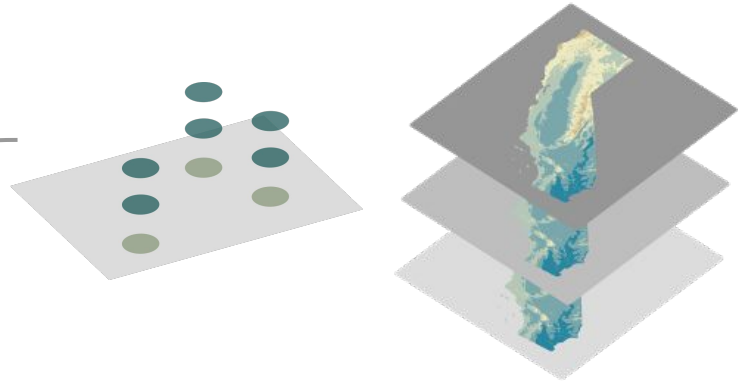
Neutral processes



Genotype-environment association (GEA) methods

Allele frequencies

Environmental data (predictors)

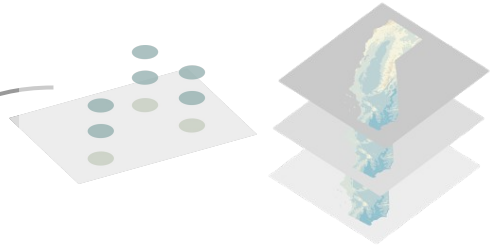


Genetic data (response)

Genotype-environment association (GEA) methods

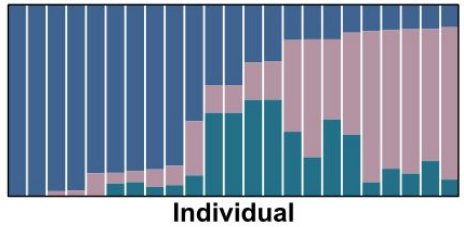
Allele frequencies

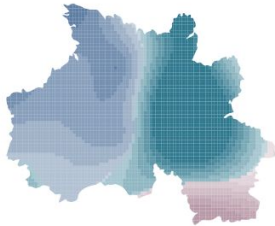
Environmental data (predictors)

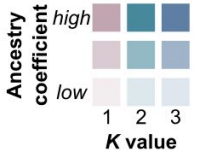


Genetic data (response)

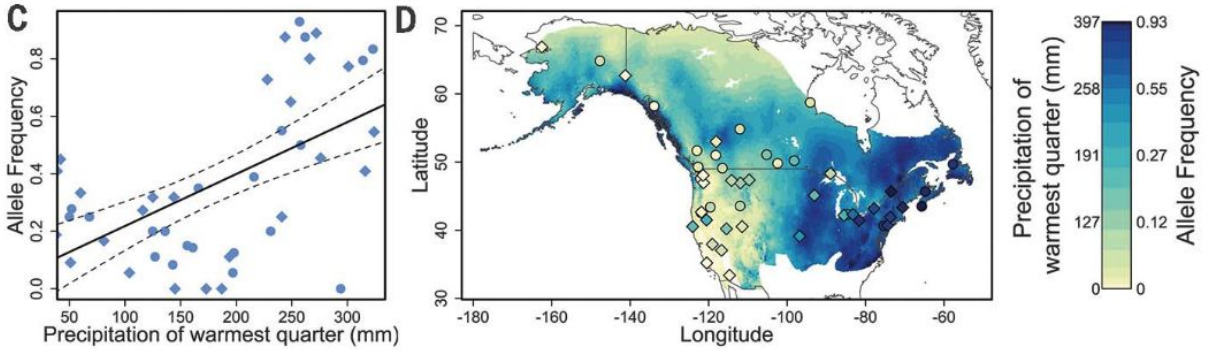
Population structure





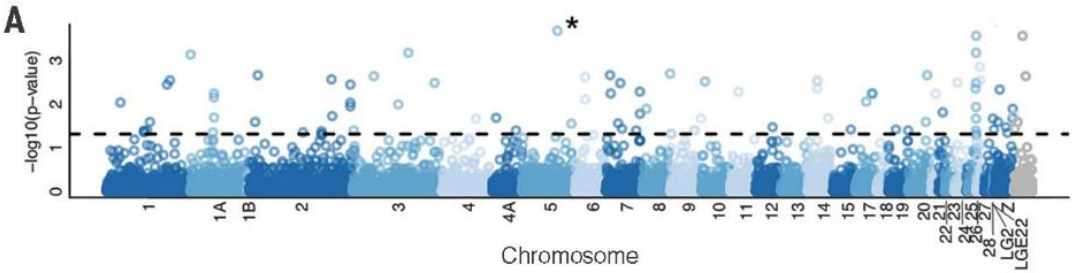


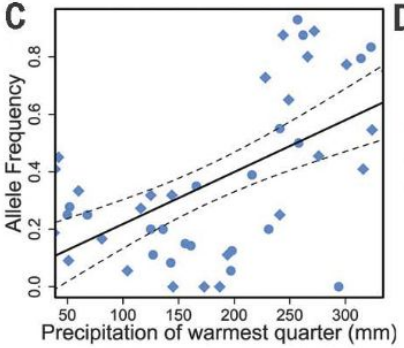
What questions can we answer using GEA?

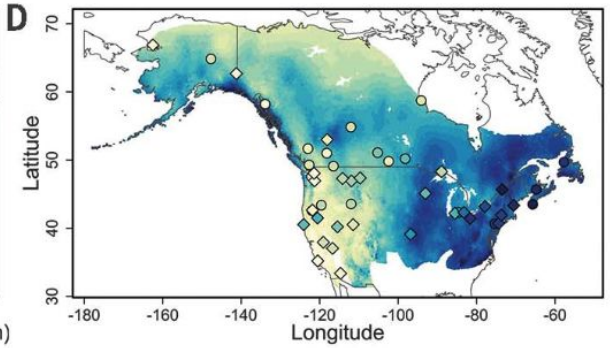


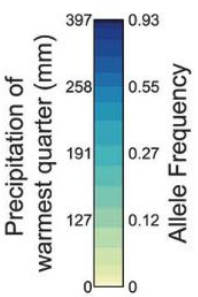
Bay et al. (2018) *Science*

What questions can we answer using GEA?



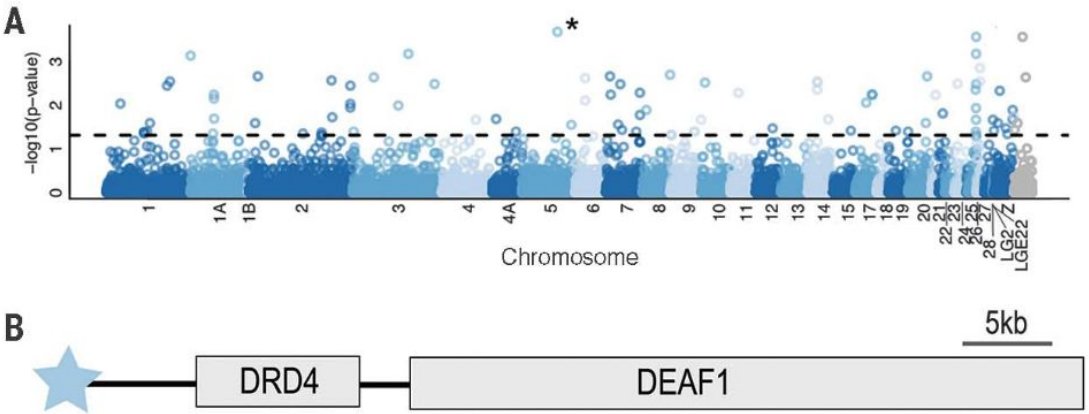


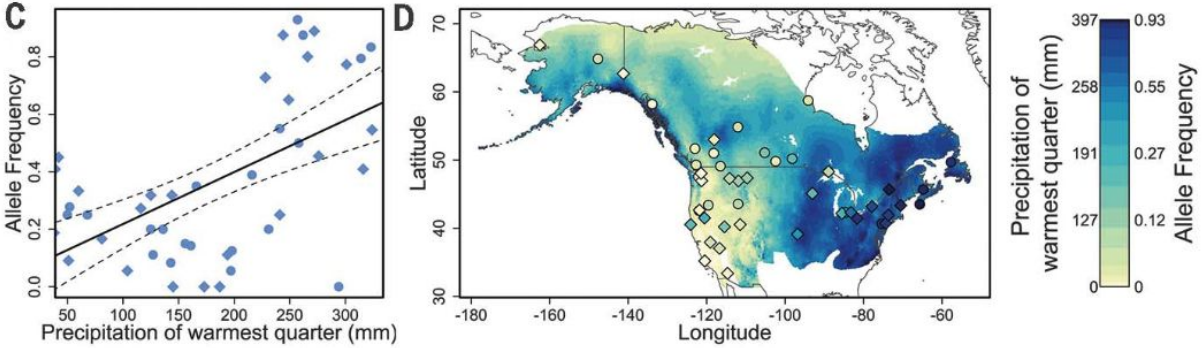




Bay et al. (2018) *Science*

What questions can we answer using GEA?





Add a biplot

Bay et al. (2018) *Science*

Different types of GEA

- BayEnv/BayPASS/BayeScEnv
- Redundancy analysis (RDA)
- Latent factor mixed models (LFMM)
- GLMM
- Gradient or random forest
- SAM/SamBada
- Weighted Z-analysis (WZA)

Different types of GEA

Method	Spatially explicit?	Accounts for neutral structure?	Individual- or population-based sampling?	Other tags
BayEnv/BayPASS	No	Yes	Population	Slow, Bayesian, linear
RDA	Optional	Optional	Both	Fast, ordination, linear
LFMM	No	Optional	Both	Fast, linear
GLMM	No	Optional	Both	Slow, linear
Gradient or random forest	Yes	No	Both	Nonlinear, map, machine learning
SAM/SamBada	No	No	Individual	Logistic

Table: Anusha P. Bishop

GEA: the logistics

Some considerations:

- May want to minimize **missing data** so as not to bias results

GEA: the logistics

Some considerations:

- May want to minimize **missing data** so as not to bias results
- Prune out sites that are in **linkage disequilibrium**

GEA: the logistics

Some considerations:

- May want to minimize **missing data** so as not to bias results
- Prune out sites that are in **linkage disequilibrium**
- Set a reasonable **MAF threshold**

GEA: the logistics

Some considerations:

- May want to minimize **missing data** so as not to bias results; if lots of data are imputed double-check the relationship between the strength of the association and % missingness (per site)

- Prune out sites that are in **linkage disequilibrium**

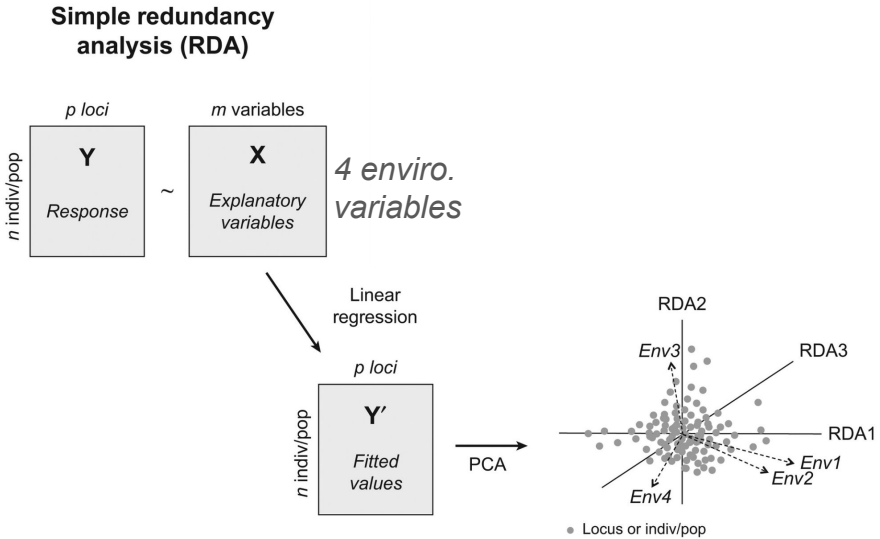
- Set a reasonable **MAF threshold**

- **Environmental data**: use realistic layers that you think are affecting your study species!

Today’s methods

Today's methods

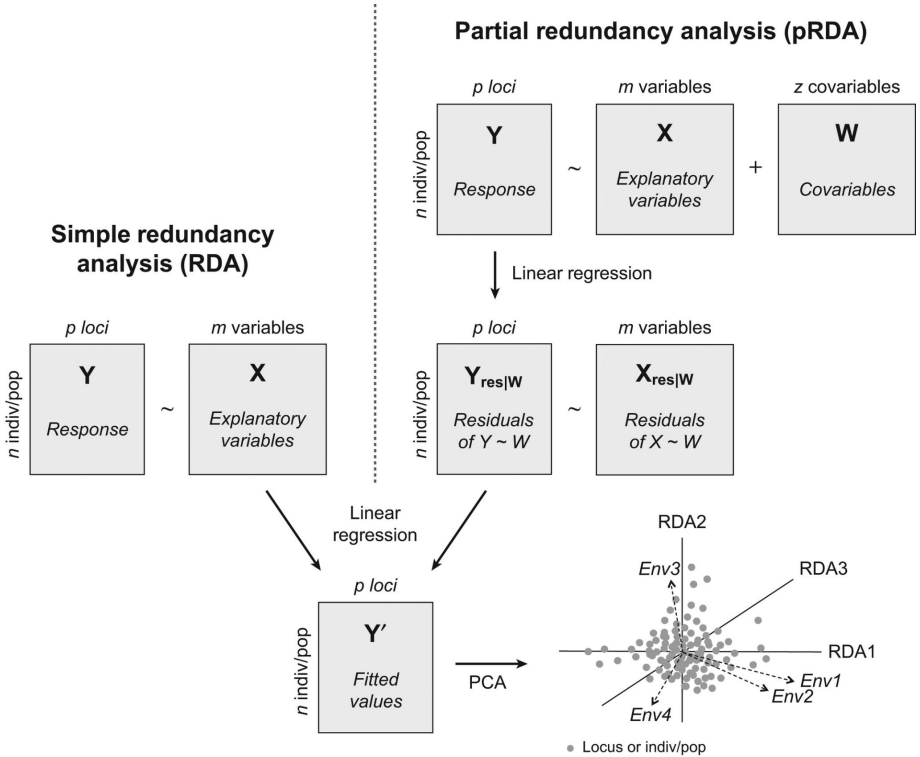
Redundancy analysis (RDA)



Capblancq & Forester (2021) *Methods Ecol. Evol.*

Today's methods

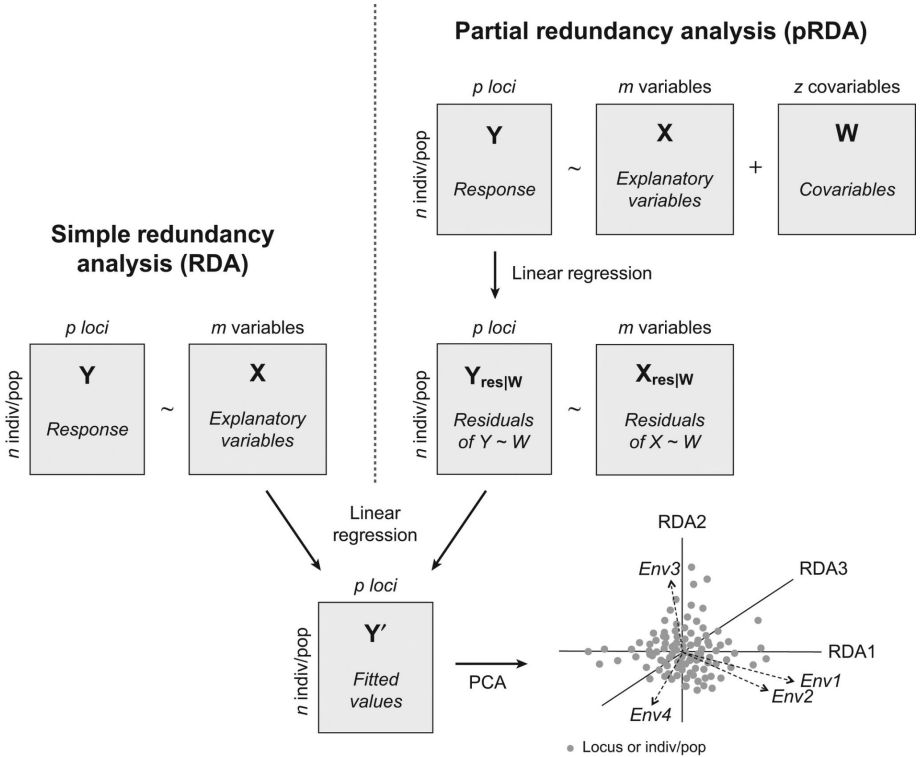
Redundancy analysis (RDA)



Capblancq & Forester (2021) *Methods Ecol. Evol.*

Today's methods

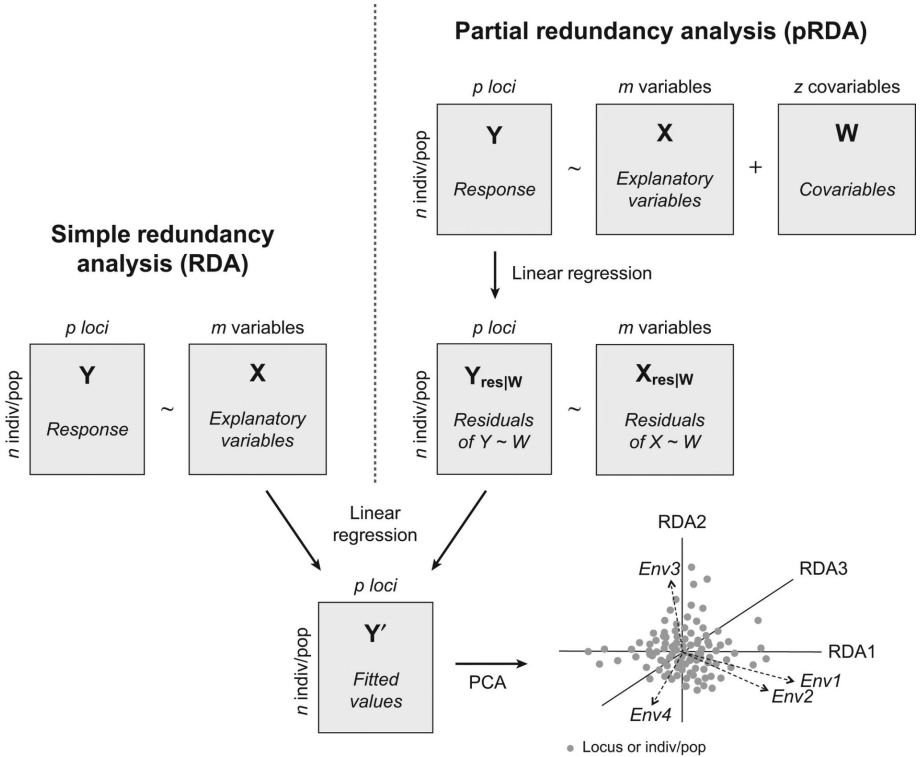
Redundancy analysis (RDA)



Capblancq & Forester (2021) *Methods Ecol. Evol.*

Today's methods

Redundancy analysis (RDA)



Latent factor mixed models (LFMM)

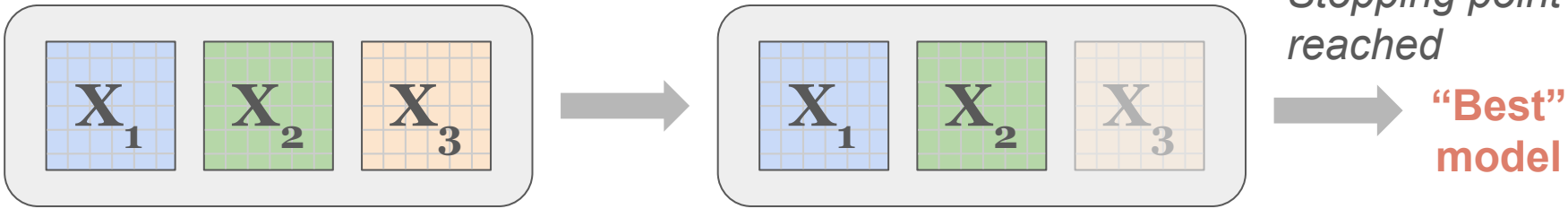
$$Y = XB^T + W + E$$

Latent matrix

Capblancq & Forester (2021) *Methods Ecol. Evol.*; Frichot et al. (2013) *Mol. Biol. Evol.*

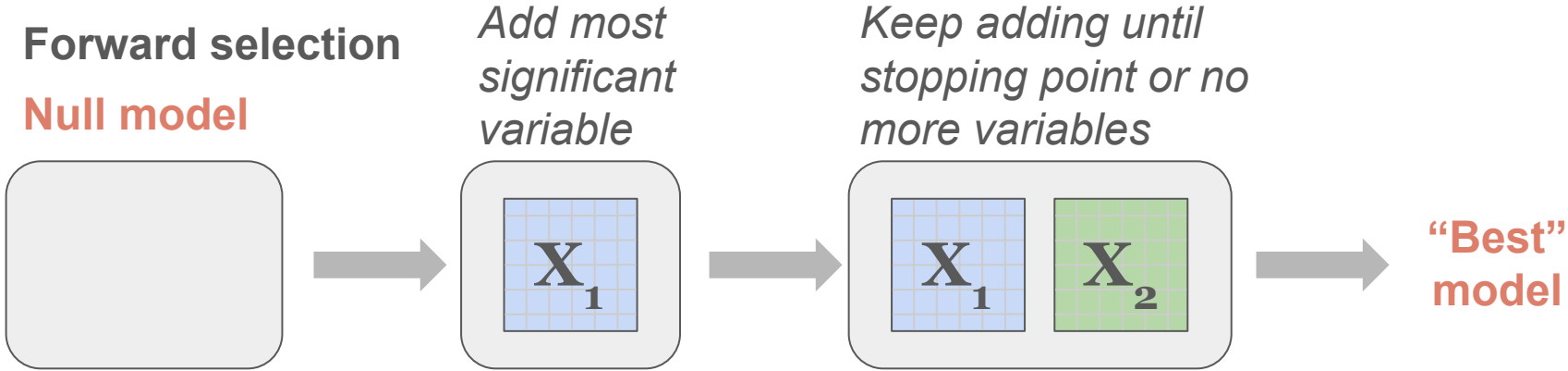
Backward elimination

Full model



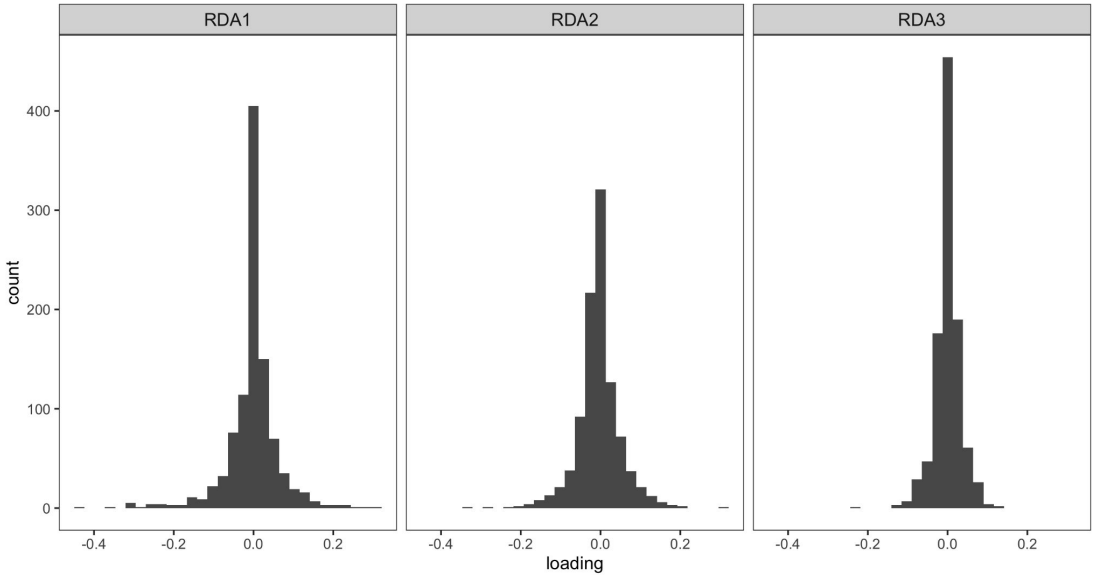
Forward selection

Null model



Variable selection is specified using the **`model = "best"`** and **`model = "full"`** syntax in algatr

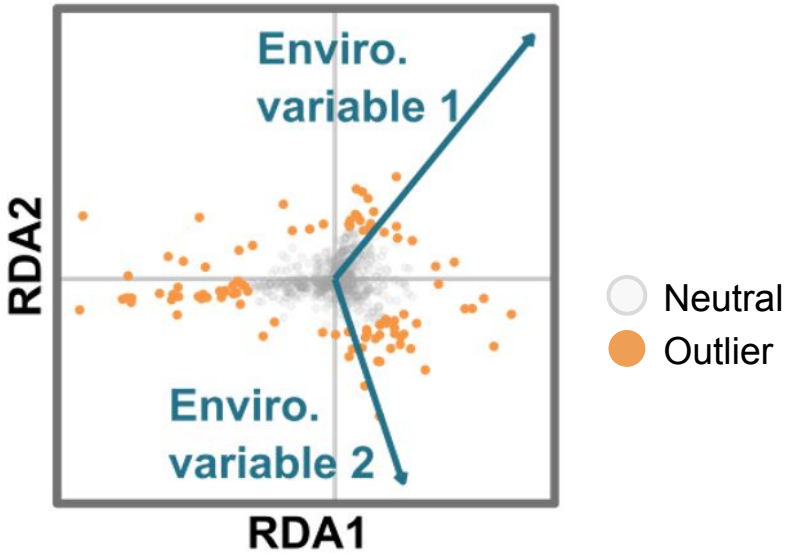
Today's methods: RDA outlier detection



Set a number of standard deviations to identify outliers from extreme loadings: **Z-scores method**

Transform RDA loadings into p-values and adjust based on FDR: **p-value method** using the rdadapt function

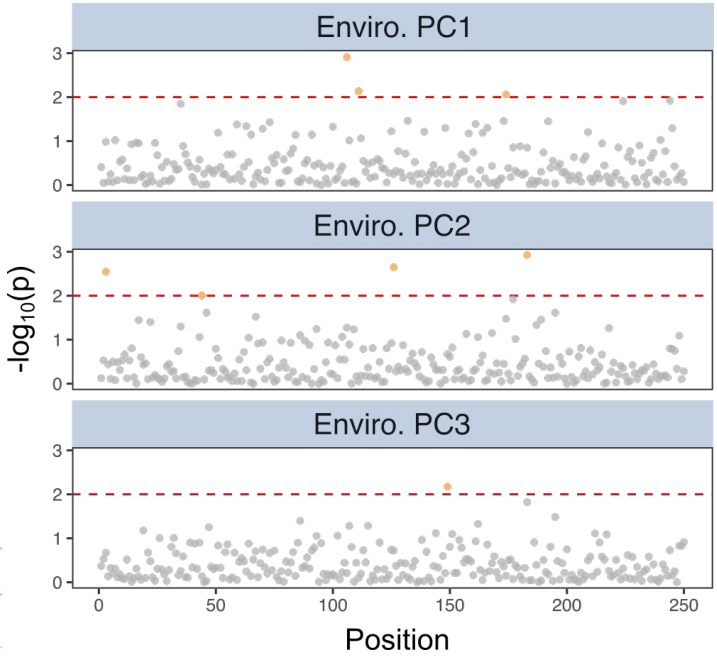
Redundancy analysis (RDA)



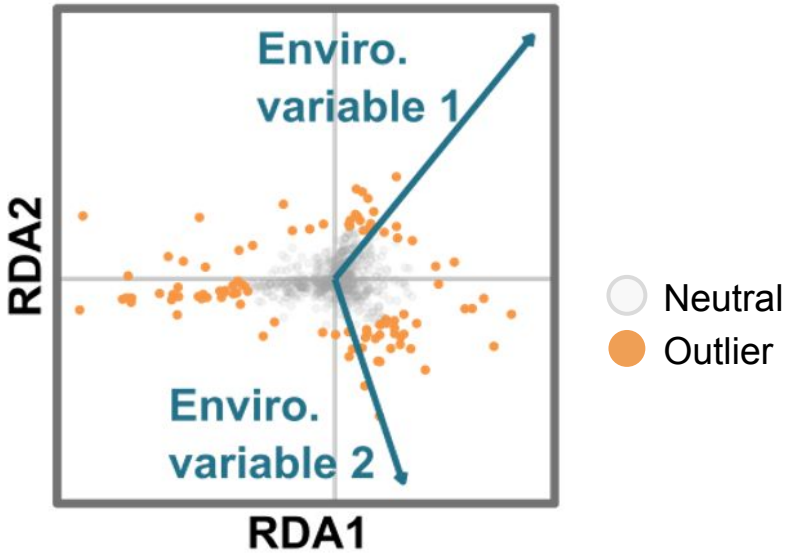
Neutral
Outlier

r	p-value	SNP	Variable
-0.22	0.02	chrom1_SNP5476	Enviro. PC1
0.22	0.02	chrom10_SNP66	Enviro. PC1
-0.24	0.01	chrom13_SNP337	Enviro. PC1
-0.19	0.04	chrom2_SNP779	Enviro. PC1
-0.19	0.04	chrom1_SNP3165	Enviro. PC2

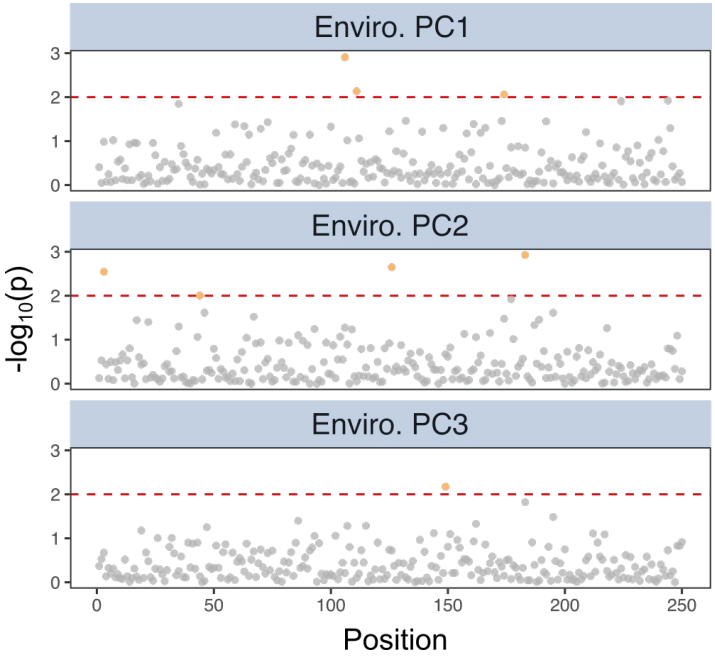
Latent factor mixed models (LFMM)



Redundancy analysis (RDA)



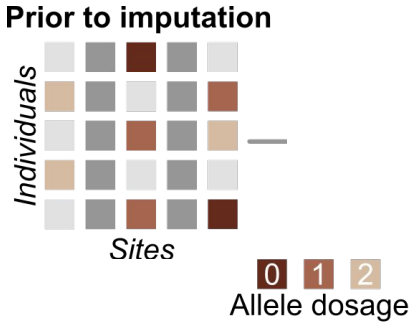
Latent factor mixed models (LFMM)

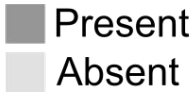


These methods can't accept missing data! Two choices with different tradeoffs…

Imputation for GEA methods

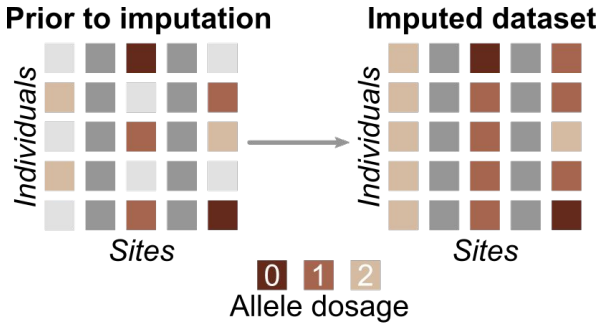
Median-based





Imputation for GEA methods

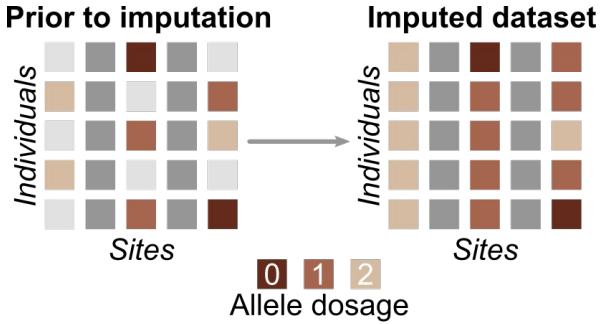
Median-based

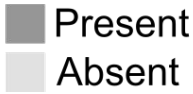


Present
Absent

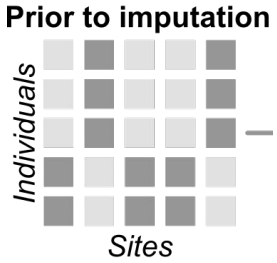
Imputation for GEA methods

Median-based



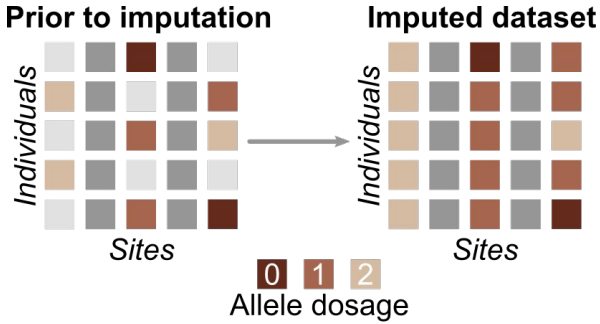


Structure-based

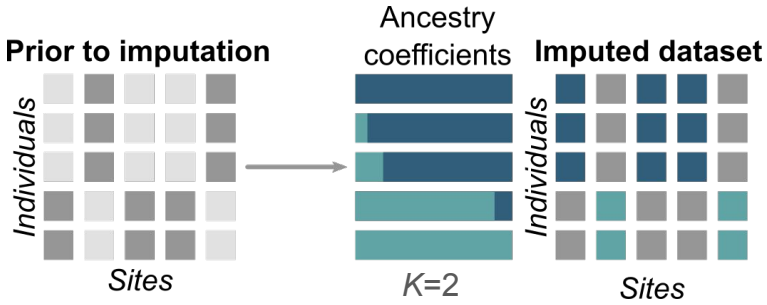


Imputation for GEA methods

Median-based



Structure-based



Present
Absent

EXERCISE 1: RDA

https://thewanglab.github.io/algatr/articles/RDA_vignette.html

Process input data

Run structure-based imputation using **`str_impute()`** and see how this changes your RDA results!

```
# Simple imputation of missing values
gen <- simple_impute(liz_dosage,
                     FUN = median)
```

Dosage matrix (with NAs)

`FUN` = imputation method

```
# Standardize environmental variables
env <- scale(env,
             center = TRUE,
             scale = TRUE)
env <- data.frame(env)
```

Environmental variables, extracted for sampling localities

`center` = transform such that mean=0

`scale` = transform such that SD=1

Turn environmental variables into a data frame

Run simple RDA

```
mod_full <- rda_run(gen,
                    env,
                    model = "full")
```

Dosage matrix (no missing values!)

Environmental variables,
extracted for sampling localities

`model` = whether to run with (**“best”**)
without variable selection (**“full”**)

Run a simple RDA with variable selection by specifying **`model = “best”`**

Run a partial RDA

```
mod_pRDA <- rda_run(gen,
                    env,
                    model = "full",
                    correctGEO = TRUE,
                    coords = liz_coords,
                    correctPC = FALSE)
```

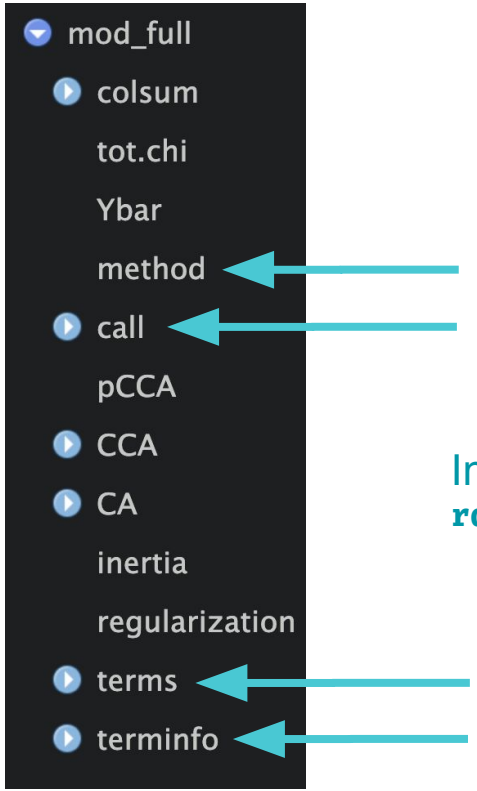
correctGEO = whether to run a partial RDA with geographic coordinates as a covariable

coords = sampling coordinates (required if **correctGEO = TRUE**)

correctPC = whether to run a partial RDA correcting for population structure

How would you run a partial RDA with population structure (i.e., PCs) as a covariable?

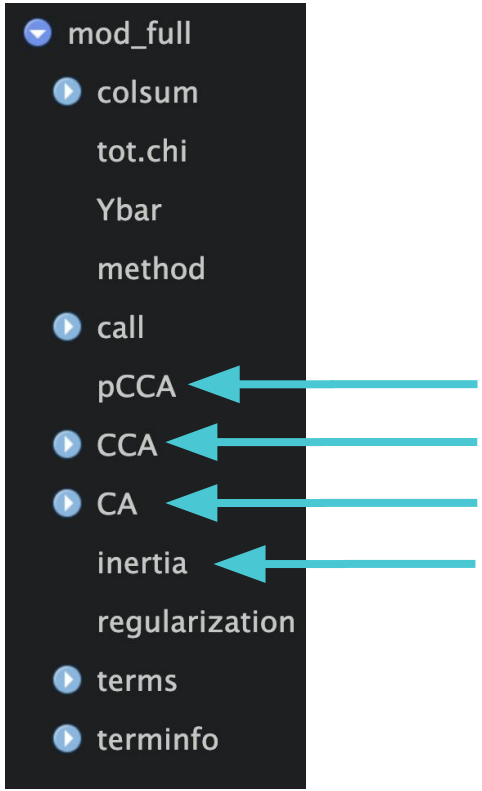
Interpreting RDA results



Information about the method and formula used in the model, e.g.,:

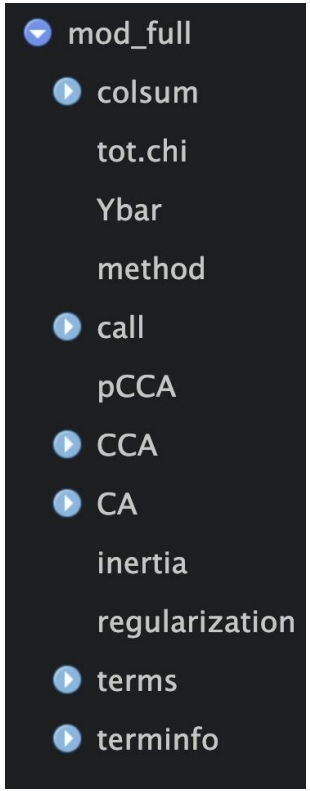
```
rda(formula = gen ~ CA_rPCA1 + CA_rPCA2 + CA_rPCA3, data = moddf)
```

Interpreting RDA results



Information about the ordination performed within RDA

Interpreting RDA results




```
RsquareAdj(mod_full)
```

Identifying candidate SNPs from RDA

```
rda_sig_z <- rda_getoutliers(mod_full,
                             naxes = "all",
                             outlier_method = "z",
                             z = 3,
                             plot = FALSE)
```

RDA model

naxes = number of RDA axes to use

z = number of SDs to use to identify SNPs with the Z-score method

outlier_method = whether to use Z-scores method or *p*-values method (RDadapt) for RDA outlier detection

See what happens when you set **plot = TRUE**.

Identifying candidate SNPs from RDA

```
rda_sig_p <- rda_getoutliers(mod_full,
                             naxes = "all",
                             outlier_method = "p",
                             p_adj = "fdr",
                             sig = 0.01)
```

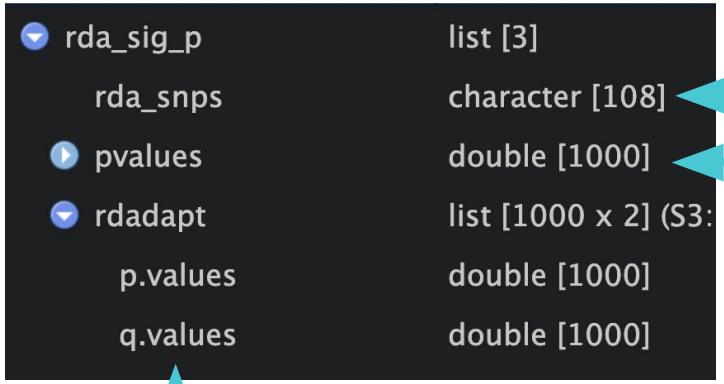
p-value method for outlier detection

`p_adj` = method to use for *p*-value correction

`sig` = significance threshold for *p*-values

How many significant outliers were detected using each of these methods?

Identifying candidate SNPs from RDA



`rda_snps` = significant SNPs

`pvalues` = *p*-values for all SNPs inputted

`rdadapt` = p- and q-values

How does the resulting object differ from the one you produced using the Z-scores method?

Interpreting and plotting RDA results

```
rda_plot(mod_full,
         rda_snps = rda_sig_p$rda_snps,
         pvalues = rda_sig_p$pvalues,
         biplot_axes = c(1, 2),
         rdaplot = FALSE,
         manhattan = TRUE)
```

`rda_snps` = list of outlier SNPs

`pvalues` = *p*-values corresponding to all SNPs considered in RDA model

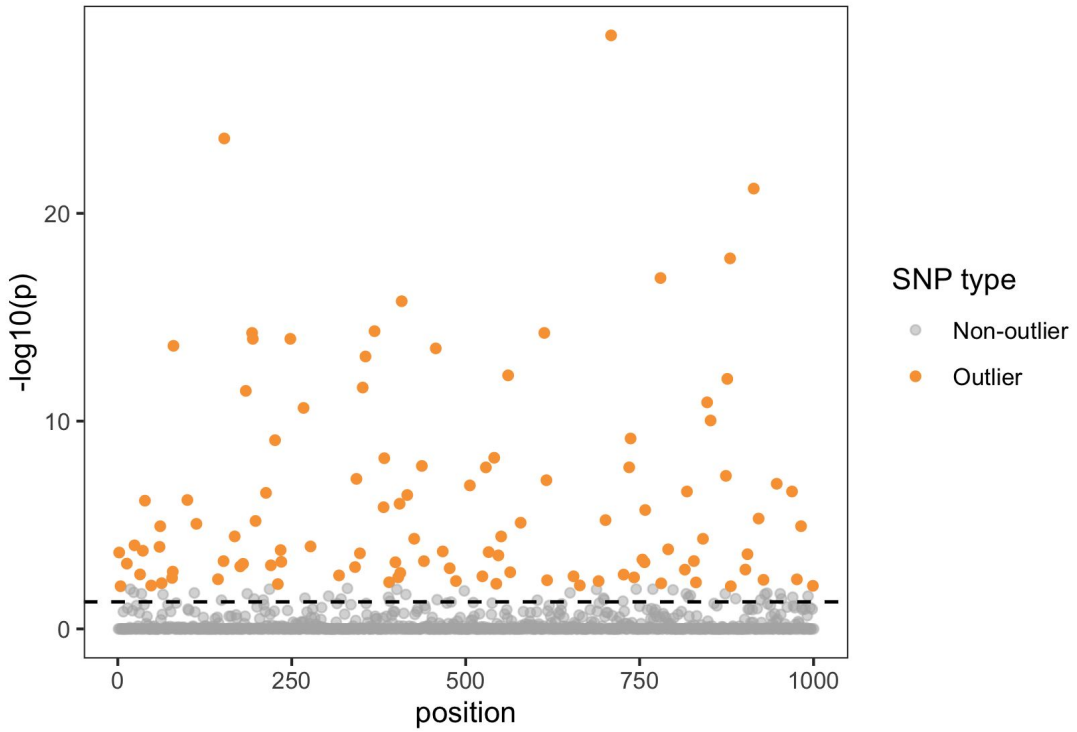
`rdaplot` = whether to make an RDA biplot of results

`rdaplot` = whether to make an RDA biplot of results

`manhattan` = whether to make a Manhattan plot of results

Interpreting and plotting RDA results

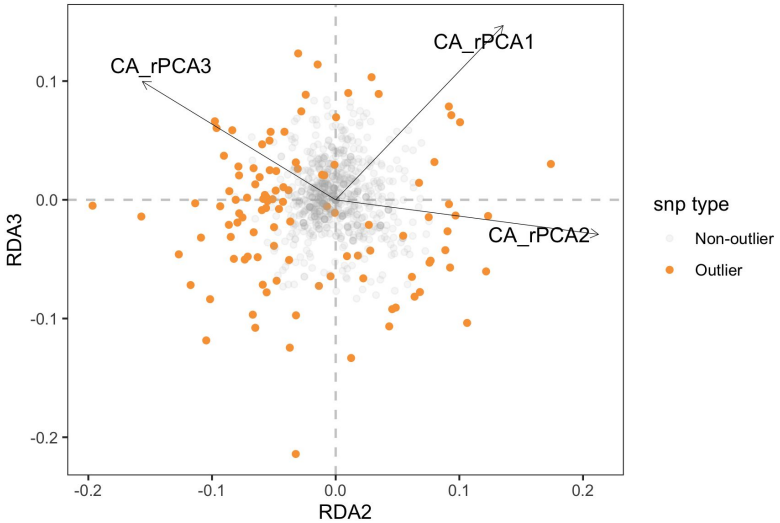
```
rda_plot(mod_full,
         rda_snps = rda_sig_p$rda_snps,
         pvalues = rda_sig_p$pvalues,
         biplot_axes = c(1, 2),
         rdaplot = FALSE,
         manhattan = TRUE)
```



Interpreting and plotting RDA results

```
rda_plot(mod_full,
         rda_snps = rda_sig_p$rda_snps,
         rdaplot = TRUE,
         manhattan = FALSE,
         biplot_axes = c(1, 2))
```

`biplot_axes` = which pairs of RDA axes to plot in RDA biplot



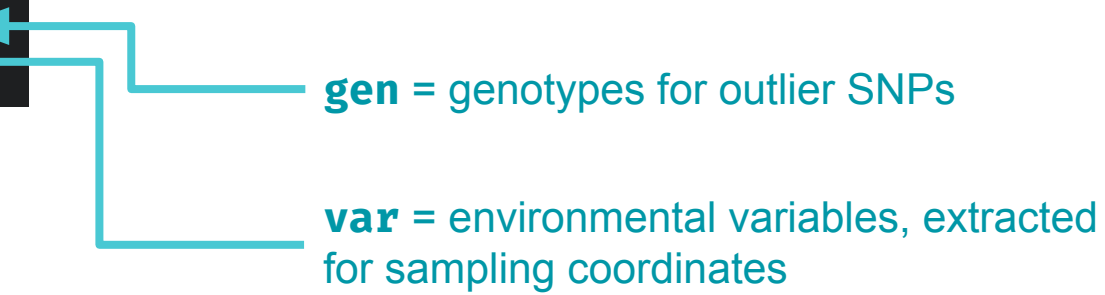
Interpreting and plotting RDA results

```
# Extract genotypes for outlier SNPs
rda_snps <- rda_sig_p$rda_snps
rda_gen <- gen[, rda_snps]
```

Interpreting and plotting RDA results

```
# Extract genotypes for outlier SNPs
rda_snps <- rda_sig_p$rda_snps
rda_gen <- gen[, rda_snps]

# Run correlation test
cor_df <- rda_cor(gen = rda_gen,
                  var = env)
```



Interpreting and plotting RDA results

```
# Extract genotypes for outlier SNPs
rda_snps <- rda_sig_p$rda_snps
rda_gen <- gen[, rda_snps]

# Run correlation test
cor_df <- rda_cor(gen = rda_gen,
                  var = env)

# Make a table from these results
rda_table(cor_df,
          nrow = 5)
```

`cor_df` = data frame with results from correlation test

`gen` = genotypes for outlier SNPs

Interpreting and plotting RDA results

```
# Extract genotypes for outlier SNPs
rda_snps <- rda_sig_p$rda_snps
rda_gen <- gen[, rda_snps]

# Run correlation test
cor_df <- rda_cor(gen = rda_gen,
                  var = env)

# Make a table from these results
rda_table(cor_df,
          nrow = 5)
```

r	p	snp	var
0.27	0.02	Locus_125	CA_rPCA3
0.39	0.00	Locus_166	CA_rPCA3
0.29	0.01	Locus_249	CA_rPCA2
-0.24	0.03	Locus_263	CA_rPCA1
0.27	0.02	Locus_263	CA_rPCA3

Do the top 5 most significantly associated SNPs differ depending on whether you use the Z-scores or *p*-values outlier method?

https://thewanglab.github.io/algatr/articles/RDA_vignette.html

Exercise 1: RDA

1. Load the example dataset
2. Process genetic data:
   a. Impute missing values in dosage matrix using structure-based imputation using `str_impute()`
3. Process environmental data:
   a. Scale extracted environmental values and make into data frame
4. Run simple RDA using `rda_run()`
5. Run partial RDA, correcting for geodist using four PCs using `rda_run()`
6. Get outliers using `rda_getoutliers()`
7. Interpret RDA results using `rda_plot()` and `rda_table()`

EXERCISE 2: LFMM

https://thewanglab.github.io/algatr/articles/LFMM_vignette.html

Run LFMM *K* selection

```
select_K(gen,
         K_selection = "tracy_widom",
         criticalpoint = 2.0234)
```

Dosage matrix with no missing values

`K_selection` = type of *K* selection to be performed

`criticalpoint` = significance levels (2.0234 corresponds to 0.01)

Run LFMM *K* selection

```
select_K(gen,
         K_selection = "tracy_widom",
         criticalpoint = 2.0234)

select_K(gen,
         K_selection = "quick_elbow",
         low = 0.08,
         max.pc = 0.90)
```

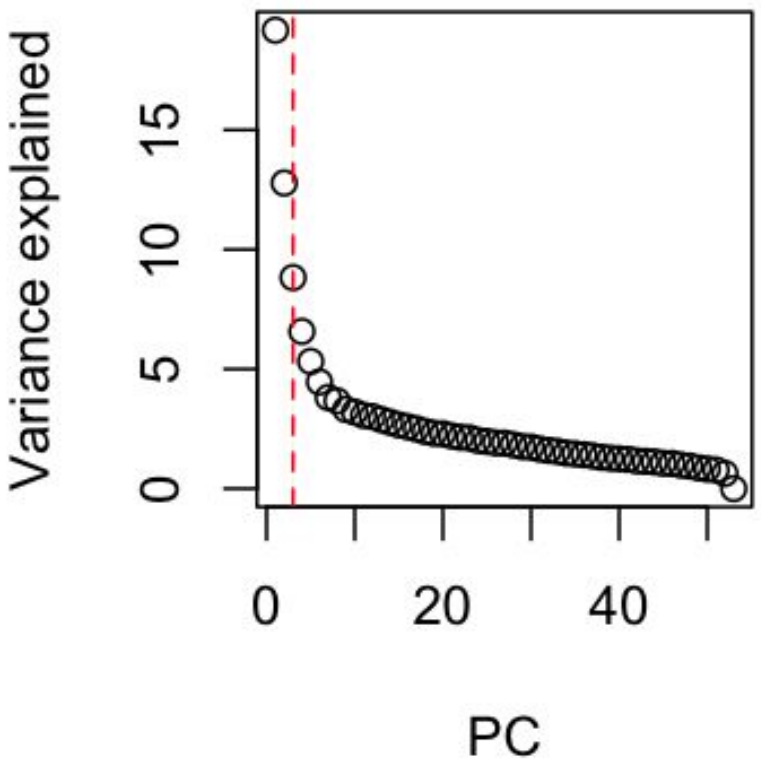
`low` = threshold that defines whether a principal component explains ‘much’ of the variance (between 0 and 1)

`max.pc` = maximum percentage of variance to capture before the elbow

Run LFMM *K* selection

```
select_K(gen,
         K_selection = "tracy_widom",
         criticalpoint = 2.0234)

select_K(gen,
         K_selection = "quick_elbow",
         low = 0.08,
         max.pc = 0.90)
```



See what happens when you use the **“`find_clusters`”** *K* selection method.

Run LFMM

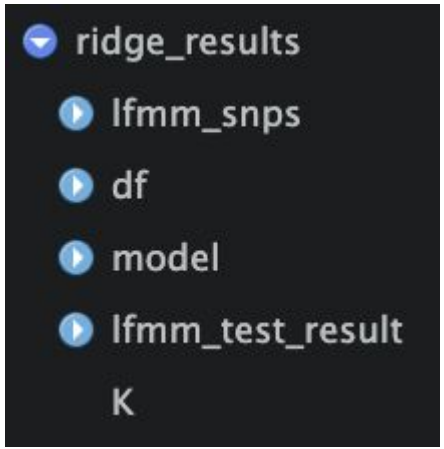
```
ridge_results <- lfmm_run(gen,
                          env,
                          K = 6,
                          lfmm_method = "ridge")
```

Dosage matrix and extracted environmental variables

K = number of latent factors

lfmm_method = method for estimating parameters of LFMM model (based on penalties for minimizing a least-squares problem)

Interpret and plot LFMM results




```
lfmm_table(ridge_results$df,
           order = TRUE,
           rows = 5)
```

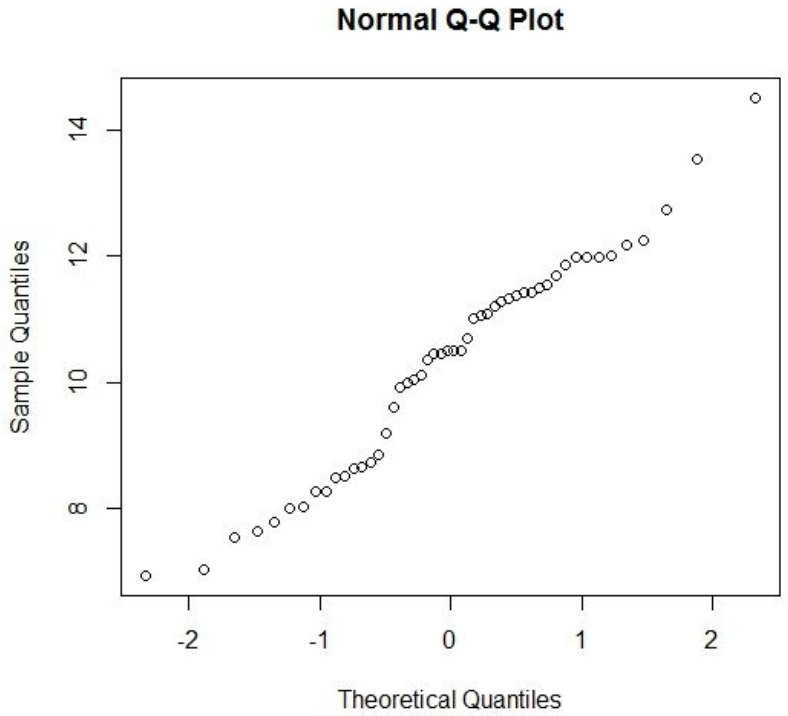
snp	variable	B[1]	z-score	p-value	calibrated z-score	calibrated p-value	adjusted p-value
Locus_2338	CA_rPCA3	-0.40	-5.25	0	25.50	0	0.00
Locus_2947	CA_rPCA2	-0.34	-7.49	0	38.03	0	0.00
Locus_1524	CA_rPCA2	-0.31	-4.84	0	15.89	0	0.00
Locus_249	CA_rPCA2	0.31	5.92	0	23.78	0	0.00
Locus_2338	CA_rPCA2	-0.30	-4.00	0	10.84	0	0.03

[1] LFMM effect size

Does the **`ridge`** or **`lasso`** method produce more LFMM outliers?

Interpret and plot LFMM results

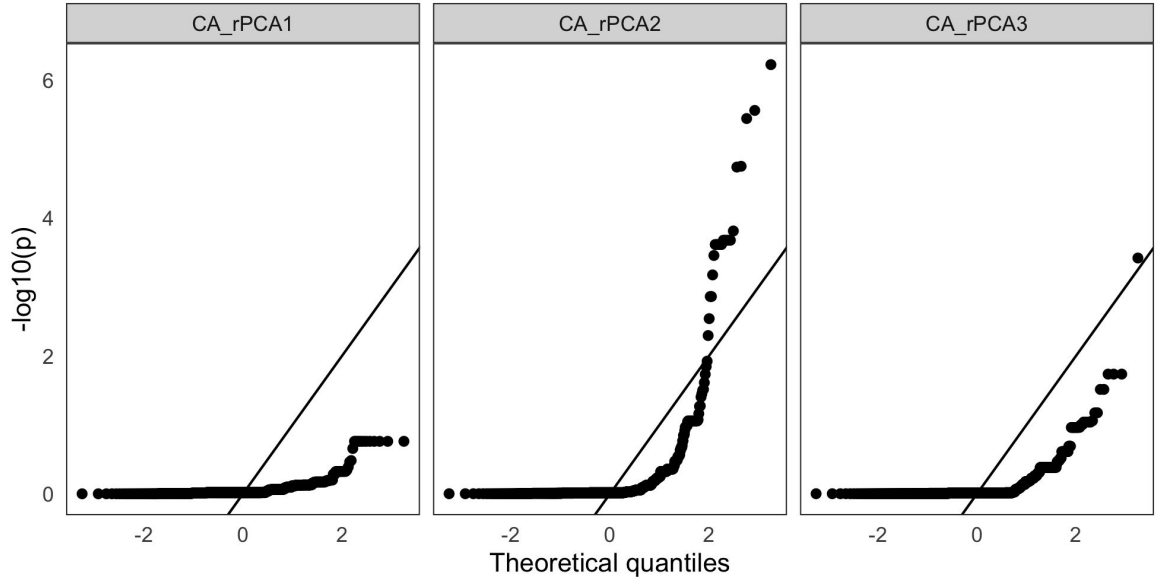
```
lfmm_qqplot(ridge_results$df)
```



https://library.virginia.edu/data/articles/understanding-q-q-plots

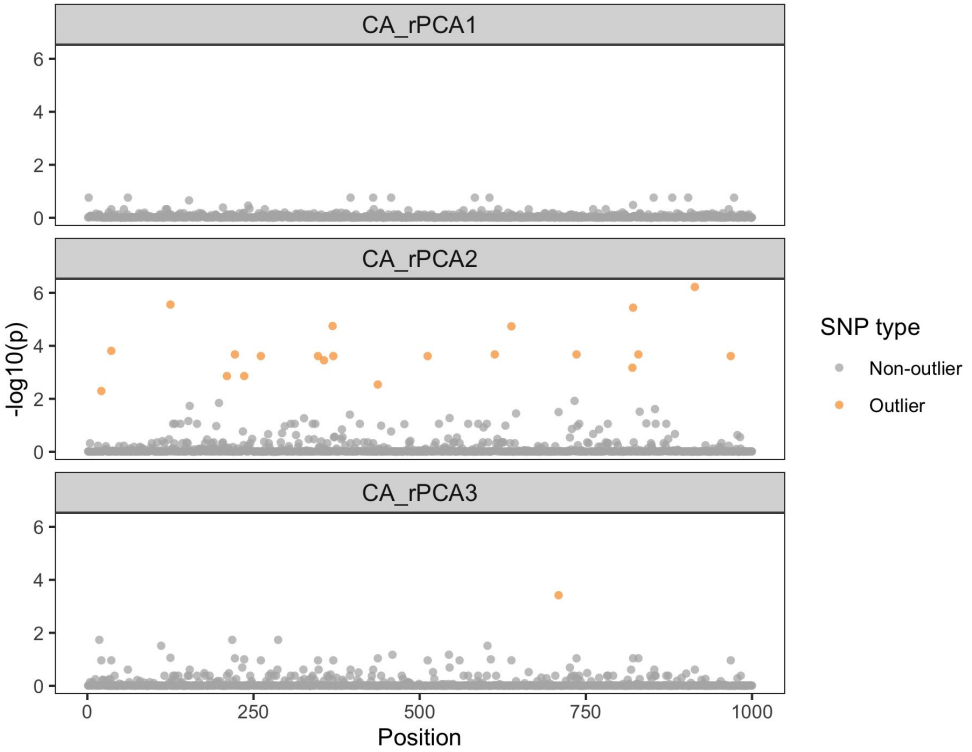
Interpret and plot LFMM results

```
lfmm_qqplot(ridge_results$df)
```



Interpret and plot LFMM results

```
lfmm_manhattanplot(ridge_results$df,
                   sig = 0.01)
```



https://thewanglab.github.io/algatr/articles/LFMM_vignette.html

Exercise 2: LFMM

1. Use same input data from your RDA analysis
2. Perform two types of *K* selection to determine how many latent factors you want to use with `select_K()`
3. Run LFMM using `lfmm_run()`
4. Get summary statistics with `lfmm_table()`
5. Make a Manhattan plot of the results using `lfmm_manhattanplot()`