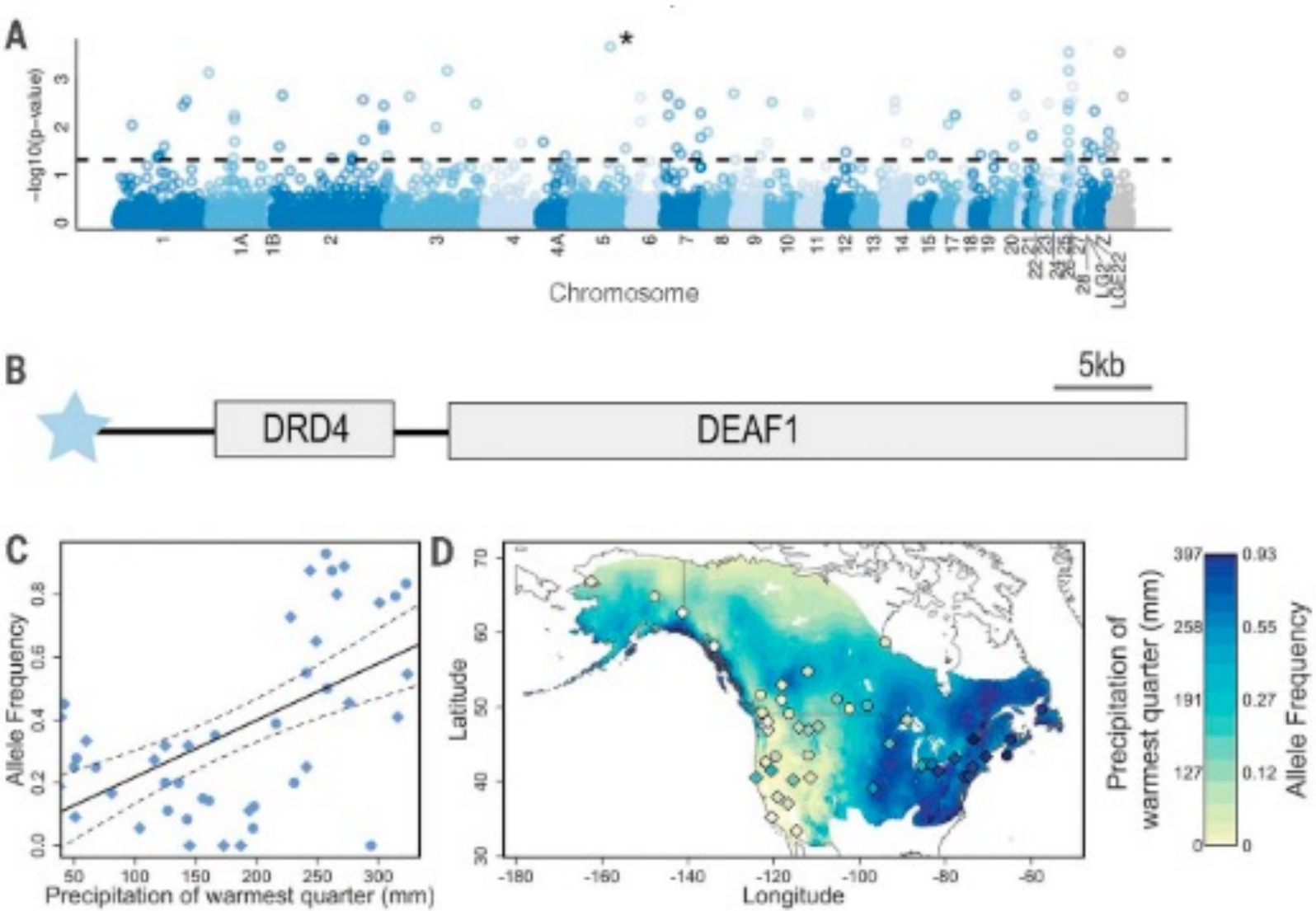
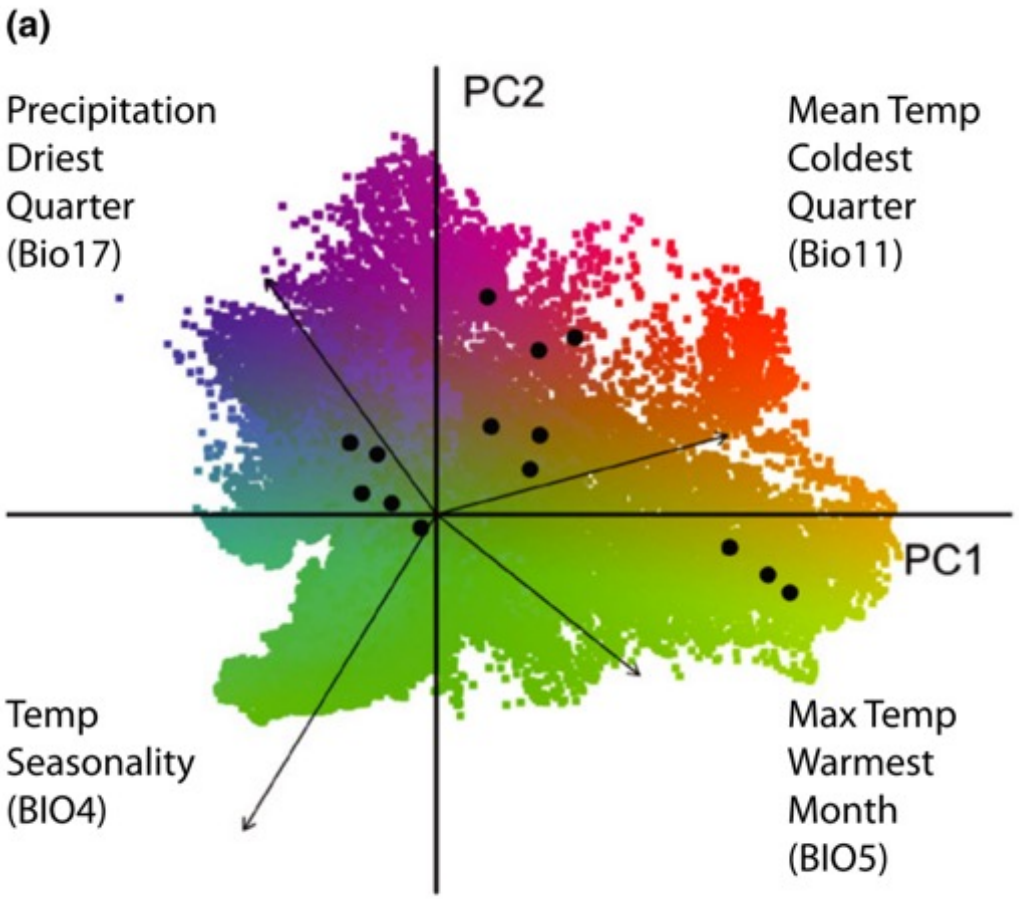
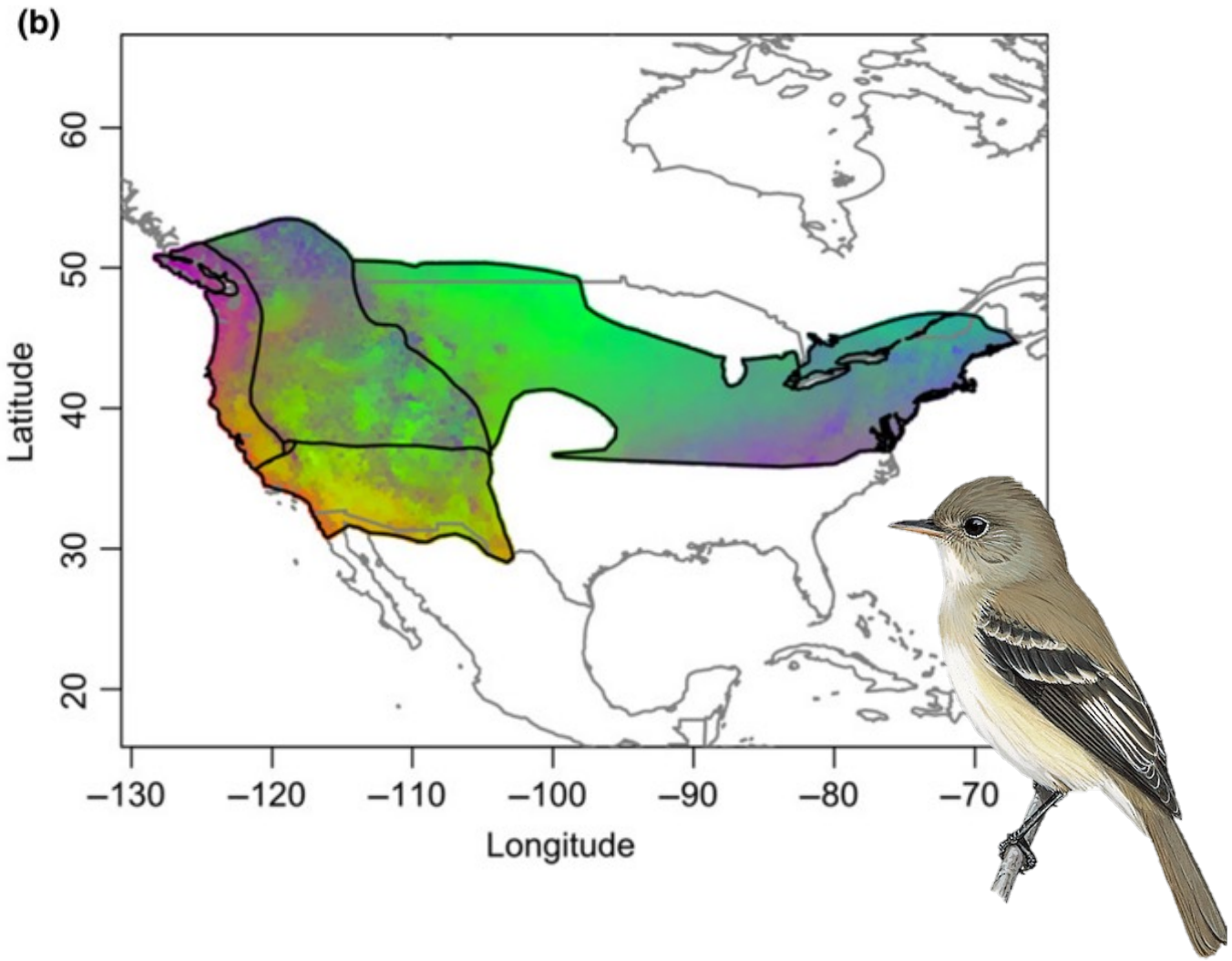
# Part II: Genotype association studies.




Bay et al. 2018 *Science*

# Genotype-environment association (GEA)



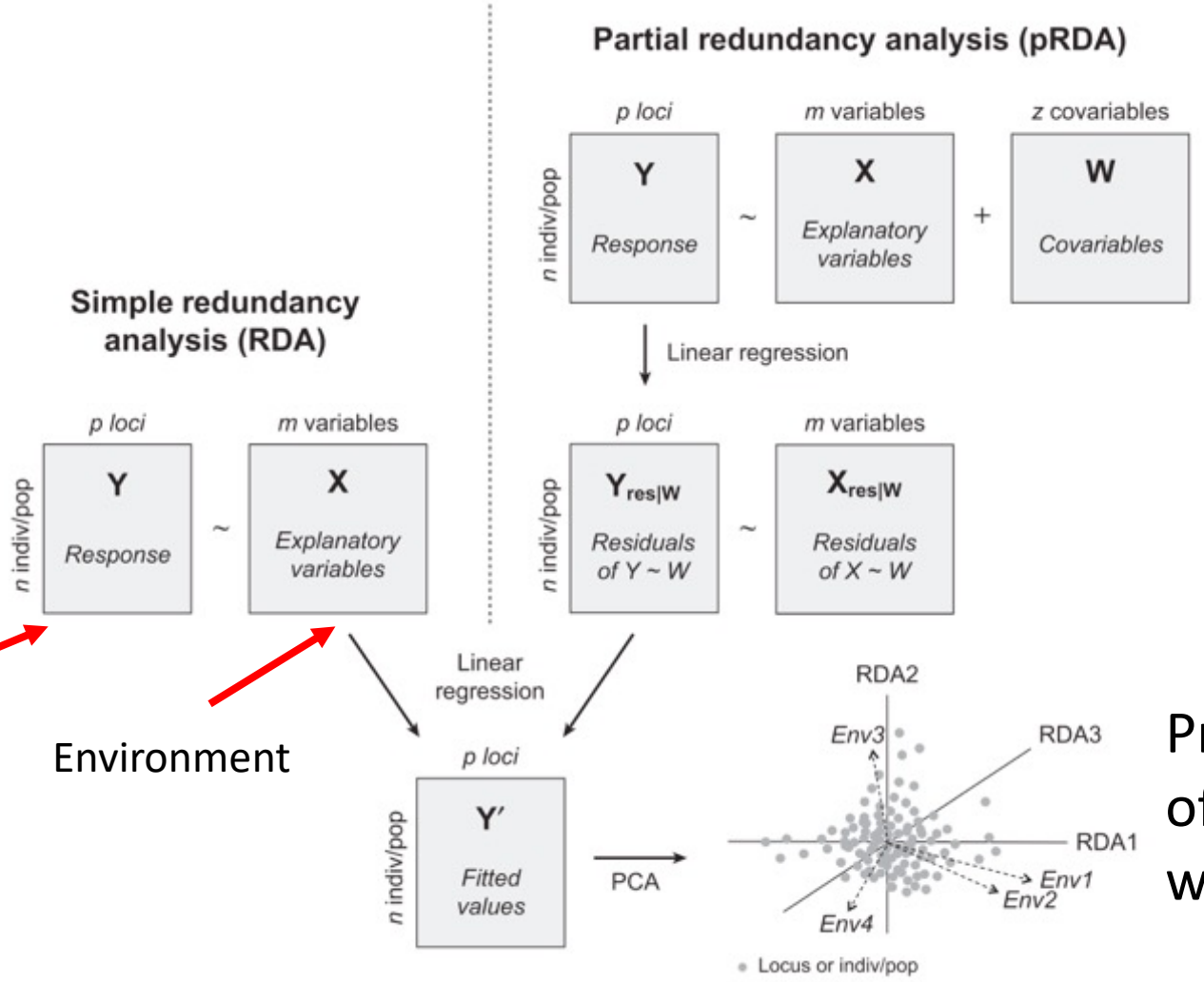




Ruegg et al. 2018 *Ecology Letters*

# Methods for estimating GEA

- BayENV/BayPass– Coop et al. 2010 *Genetics/* Gautier 2015 *Genetics*
- Redundancy analysis (RDA)– Capblancq & Forester *Methods Ecol Evol*
- Latent Factor Mixed Models (LFMM)– Frichot et al. 2013 *Mol. Biol. Evolution*
- Weighted-Z analysis (WZA)– Booker et al. 2023 *Mol. Ecol. Resources*

# Redundancy analysis (RDA)




Genotype

Environment

Produces ordination axes (RDAs) of covarying SNPs that are correlated with a multivariate environment.

Capblancq & Forester 2021 *Methods in Ecology & Evolution*

# Considerations for running RDA

- Does not accept missing data. You can either remove missing data prior to analysis or use an imputation method.
- Assumes that the relationship between genotype and environment is linear. Other models (e.g. RandomForest) can handle non-linear relationships.
- Model performance will be impacted if you have a lot of environmental variables that are tightly correlated. Correlated variables can be pruned from your dataset or you could do a PCA.

# Return to the Savannah Sparrow dataset

95 samples from across North America.

I converted from vcf to plink .raw format beforehand to import into R as a GenLight object.

**WorldClim variables**

Bio2: Mean diurnal temperature range
Bio5: Maximum temp. of warmest month
Bio7: Temp. annual range
Bio12: Annual precipitation
Bio13: Precipitation of the wettest month
Bio15: Precipitation seasonality

Excellent tutorial by Brenna Forester can be found here:

https://popgen.nescent.org/2018-03-27_RDA_GEA.html

Call: rda(formula = gen ~ bio2 + bio5 + bio7 + bio12 + bio13, data = pred, scale = T)

| | Inertia | Proportion | Rank |
|---|---|---|---|
| Total | 9.672e+04 | 1.000e+00 | |
| Constrained | 7.028e+03 | 7.266e-02 | 5 |
| Unconstrained | 8.970e+04 | 9.273e-01 | 89 |

Inertia is correlations

Eigenvalues for constrained axes:

| RDA1 | RDA2 | RDA3 | RDA4 | RDA5 |
|---|---|---|---|---|
| 2213.0 | 1337.8 | 1214.7 | 1188.4 | 1073.7 |

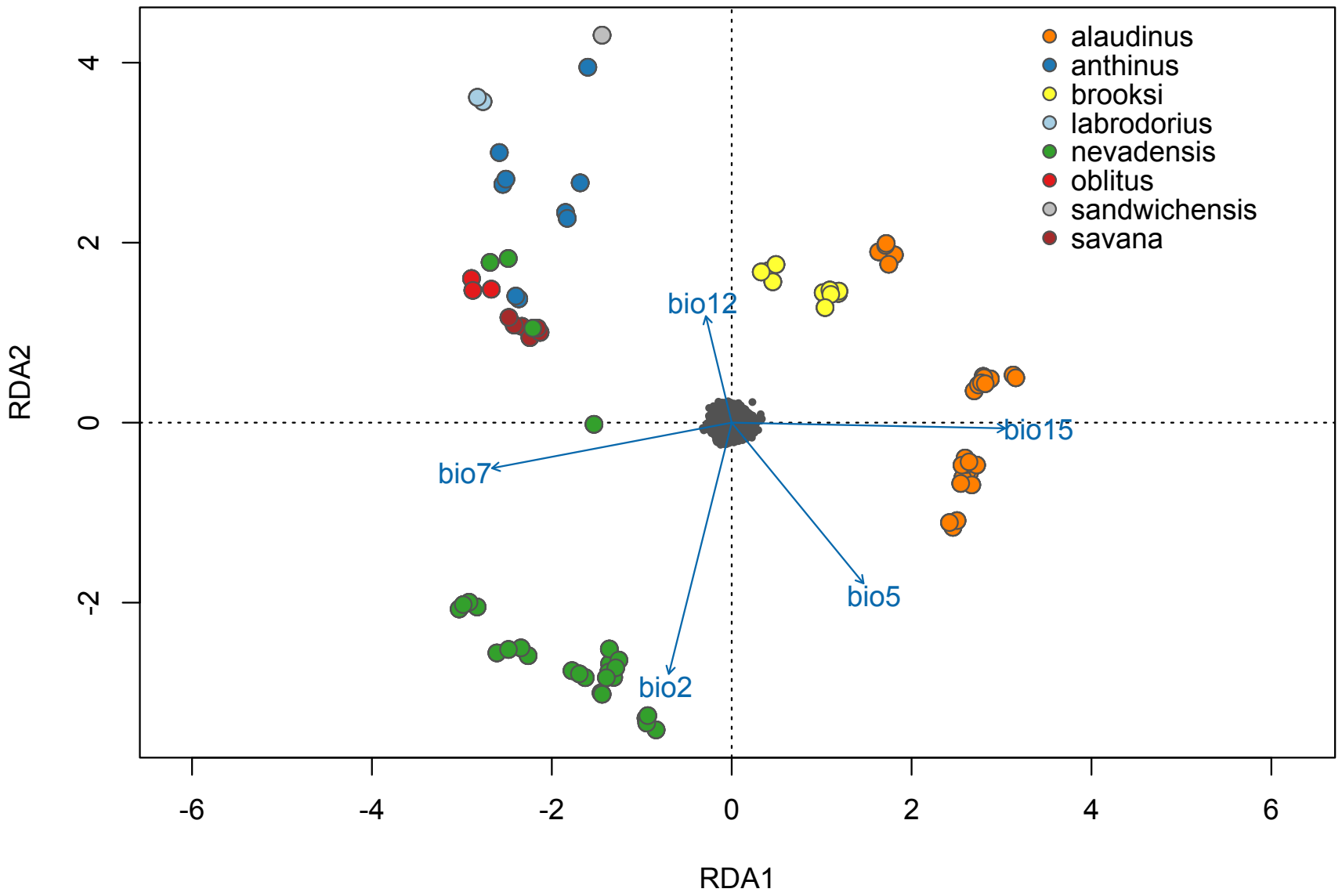
Eigenvalues for unconstrained axes:

| PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 |
|---|---|---|---|---|---|---|---|
| 1387.2 | 1308.4 | 1286.9 | 1203.1 | 1172.8 | 1162.5 | 1137.4 | 1105.5 |

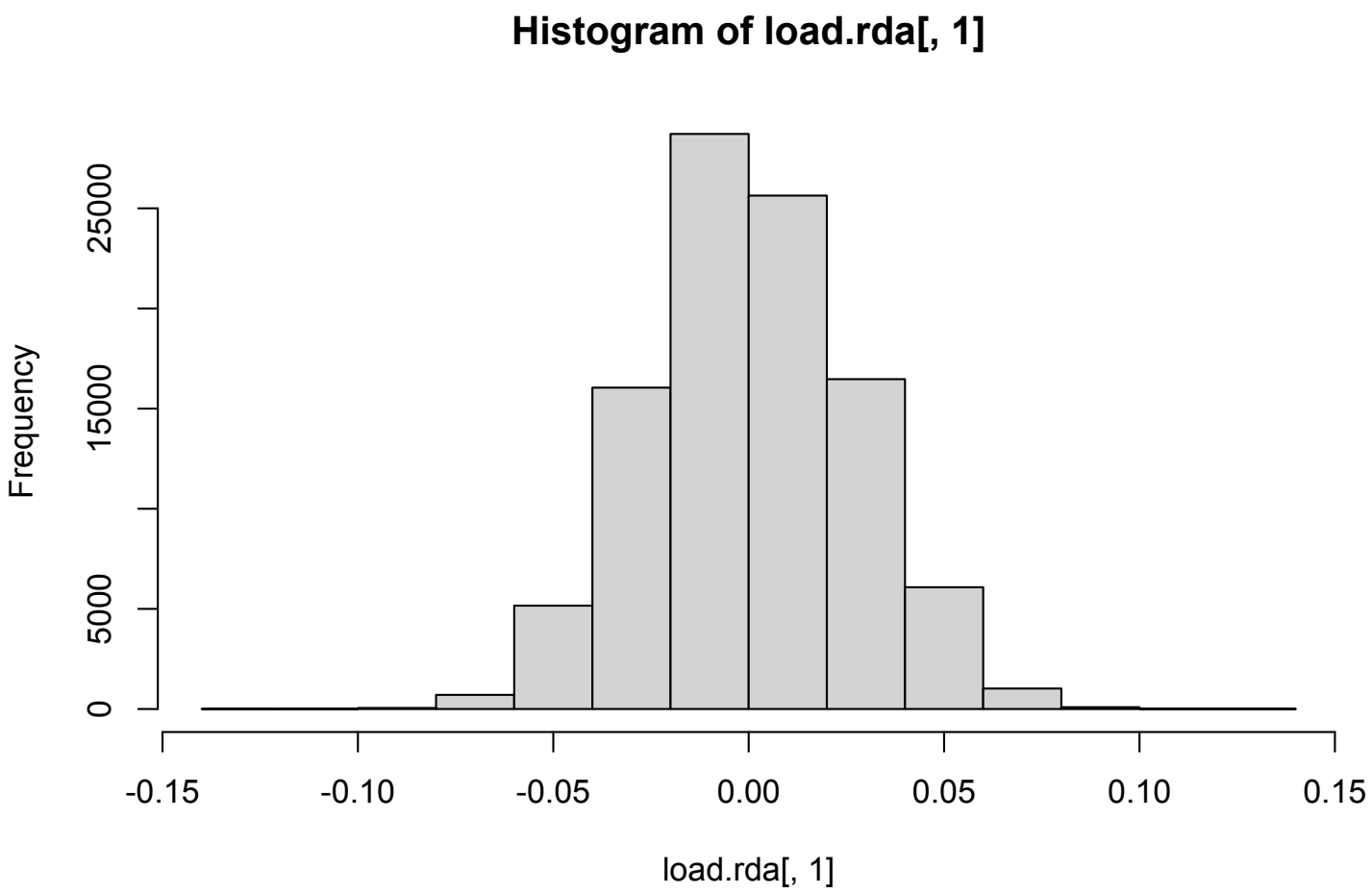
(Showing 8 of 89 unconstrained eigenvalues)

Importance of components:

| | RDA1 | RDA2 | RDA3 | RDA4 | RDA5 |
|---|---|---|---|---|---|
| Eigenvalue | 2213.0266 | 1337.8365 | 1214.6569 | 1188.4078 | 1073.6749 |
| Proportion Explained | 0.3149 | 0.1904 | 0.1728 | 0.1691 | 0.1528 |
| Cumulative Proportion | 0.3149 | 0.5053 | 0.6781 | 0.8472 | 1.0000 |


alaudinus
anthinus
brooksi
labrodorius
nevadensis
oblitus
sandwichensis
savana
bio12
bio15
bio7
bio5
bio2
RDA1
RDA2
-6
-4
-2
0
2
4
6
4
2
0
-2

# Outliers


Histogram of load.rda[, 1]
Frequency
0
5000
15000
25000
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
load.rda[, 1]

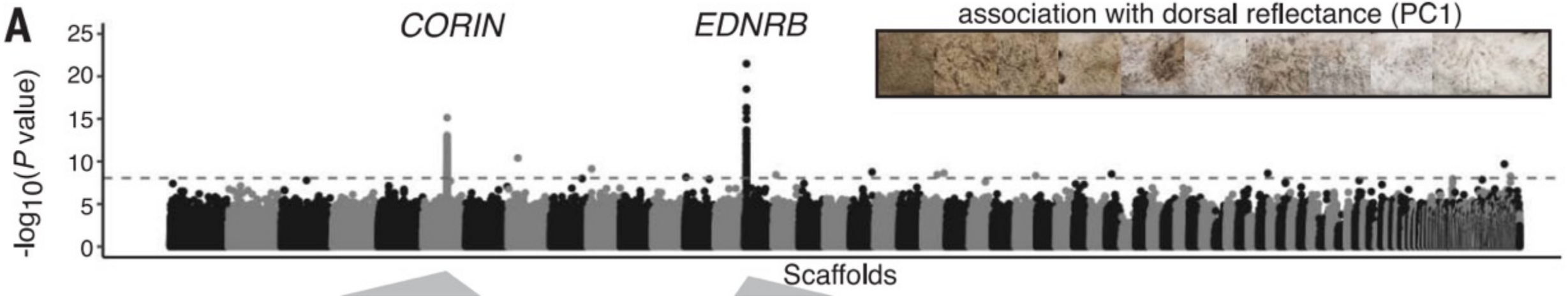
# Final considerations

- Consider using an FDR correction for determining outliers as multiple testing in large datasets can lead to false positives.

- Often best to combine inferences across multiple methods (e.g. RDA & LFMM & Fst outliers).

- Methods are generally robust to some population structure, but the amount of geographic structure in your dataset should be accounted for beforehand.

# Genome wide association study (GWAS)


Ferreira et al. 2023 *Science*

# Practical GWAS considerations

- GWAS analyses can require a substantial number of individual samples.
  - Simple, large effect loci can be detected in smaller datasets <100 individuals (e.g. hare coat color).
  - Complex, polygenic traits can require 1000s of individuals (e.g. body size).

- Population structure can increase false positive inferences. Power to detect causal variants is maximized if you work within a single panmictic population exhibiting substantial variation in the trait of interest. Also beware relatedness.

- Excellent Gemma tutorial here: https://github.com/rcc-uchicago/genetic-data-analysis-2