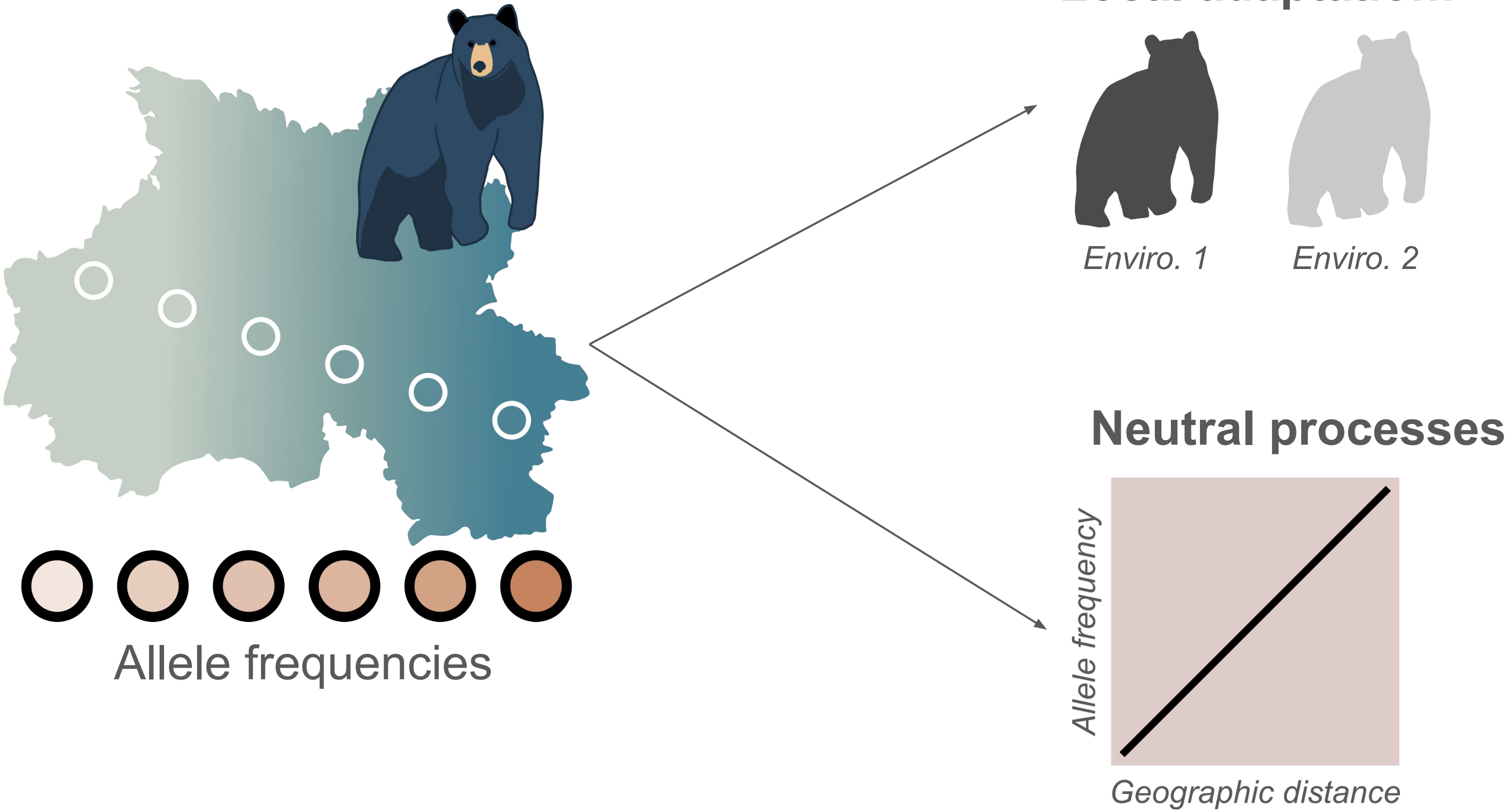
Genotype-environment associations

E. Anne Chambers
March 21, 2024
eachambers@berkeley.edu

Berkeley
UNIVERSITY OF CALIFORNIA

Genotype-environment association (GEA) methods

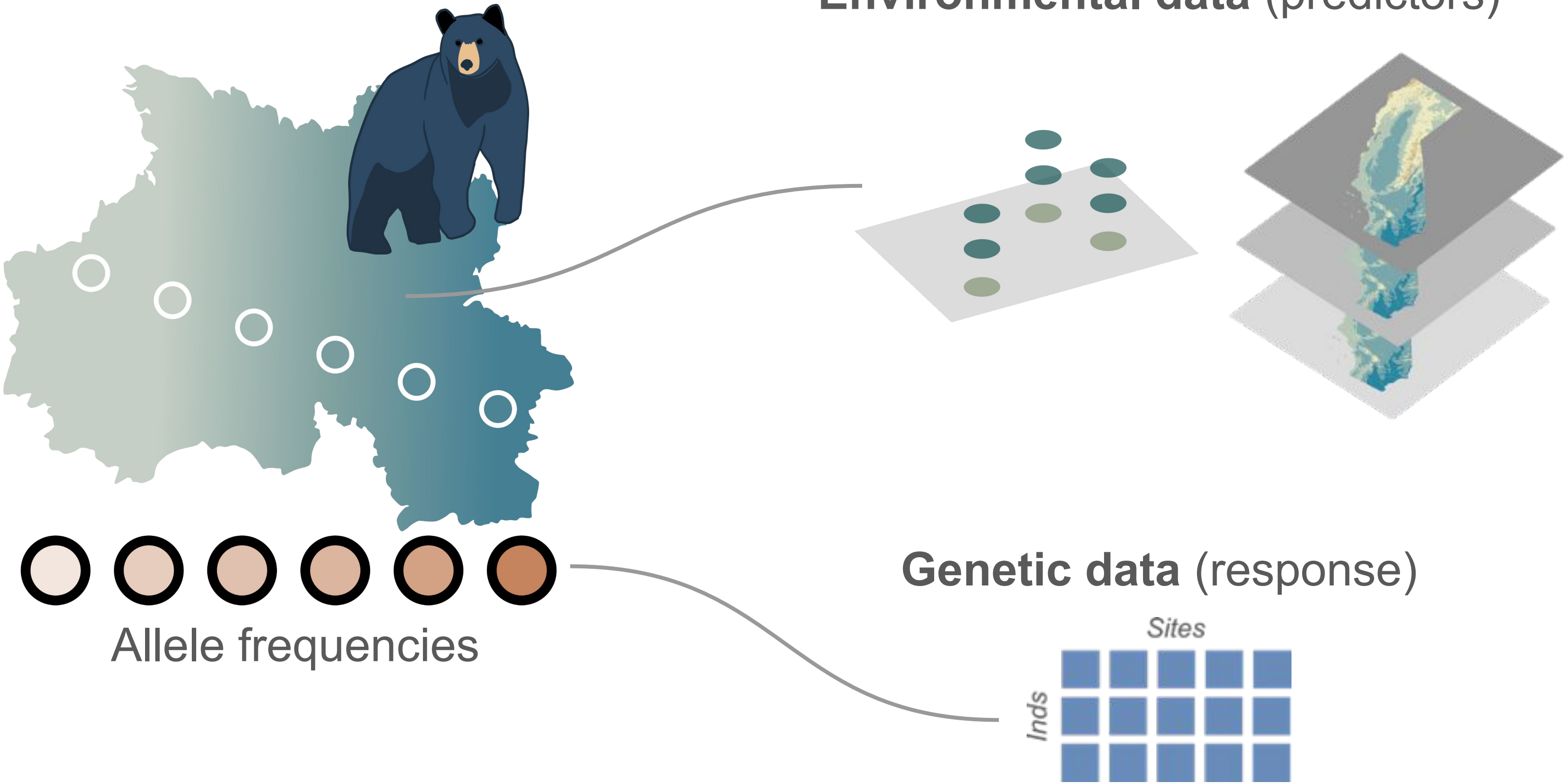


Allele frequencies

Local adaptation!

Neutral processes

Genotype-environment association (GEA) methods

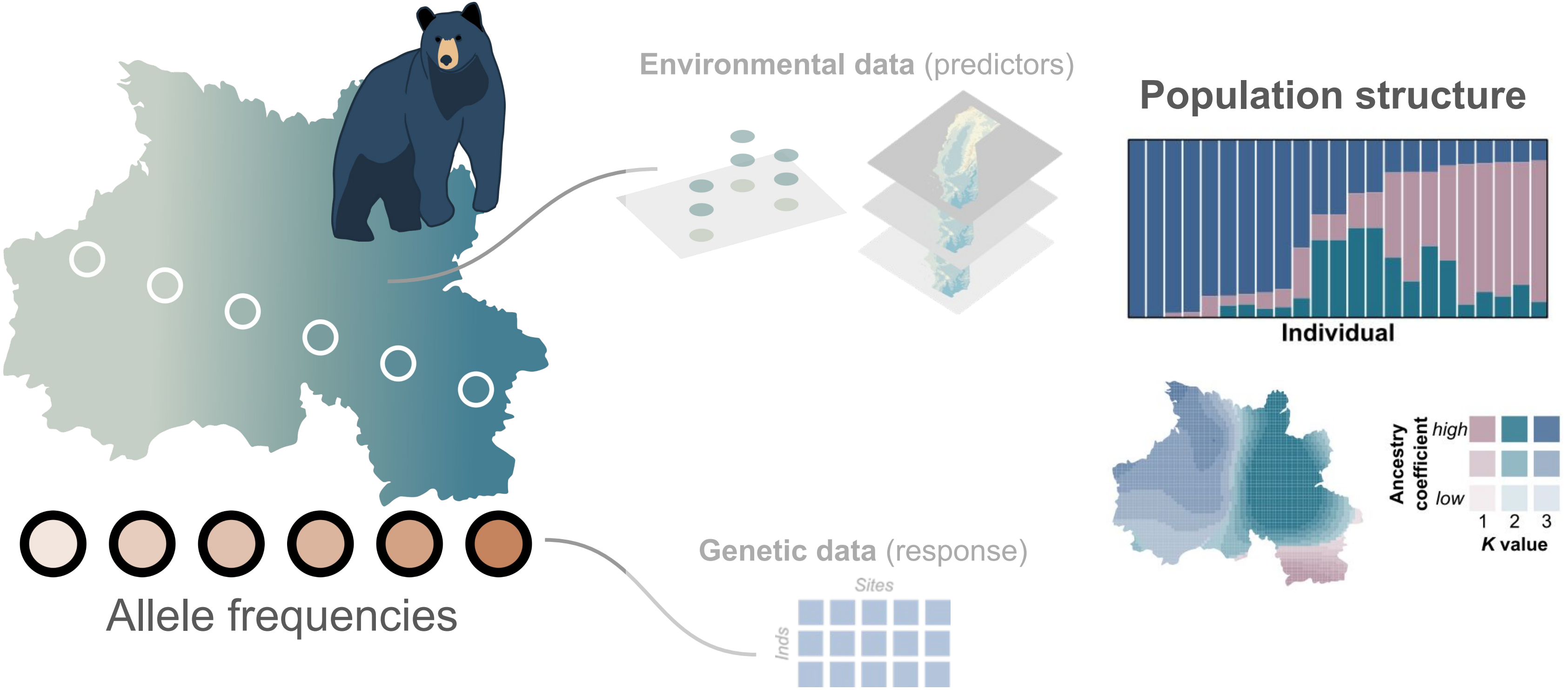


Environmental data (predictors)

Allele frequencies

Genetic data (response)

Genotype-environment association (GEA) methods



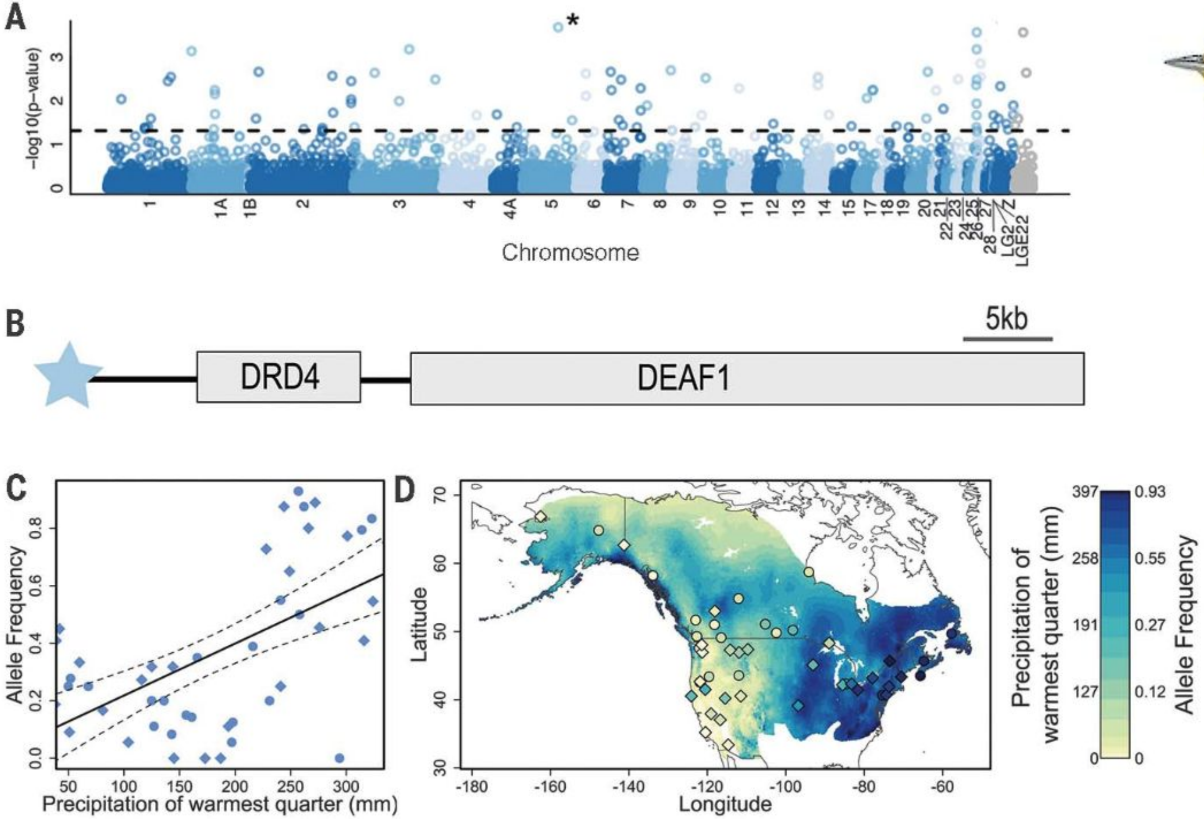
Environmental data (predictors)

Population structure

Allele frequencies

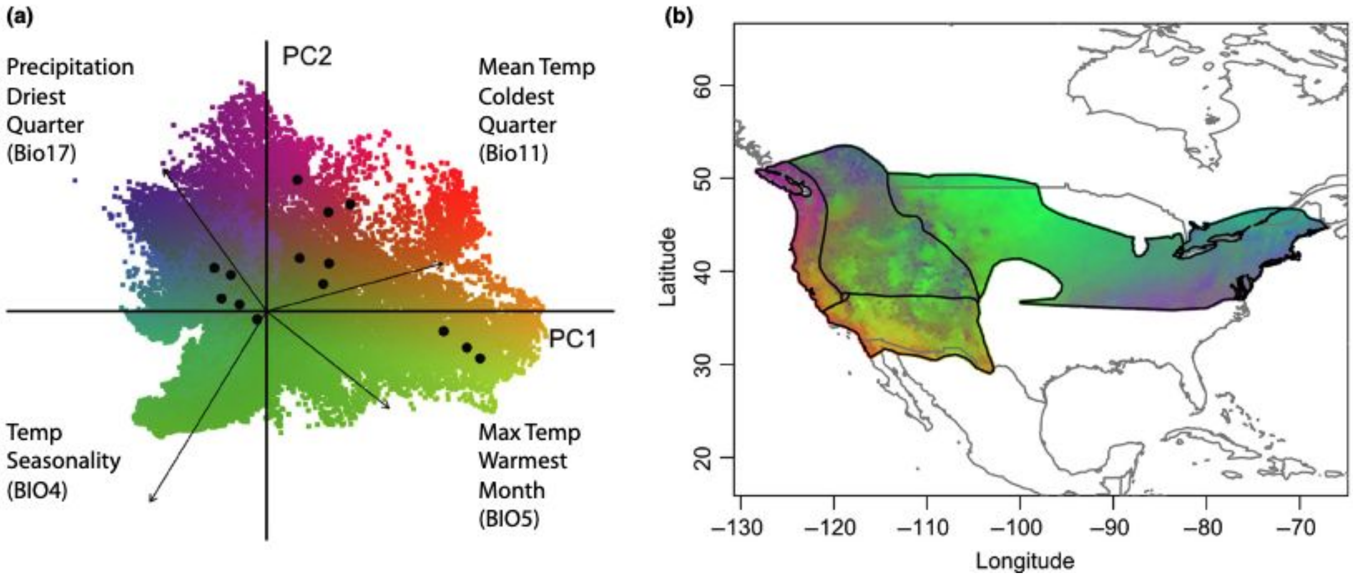
Genetic data (response)

What questions can we answer using GEA?



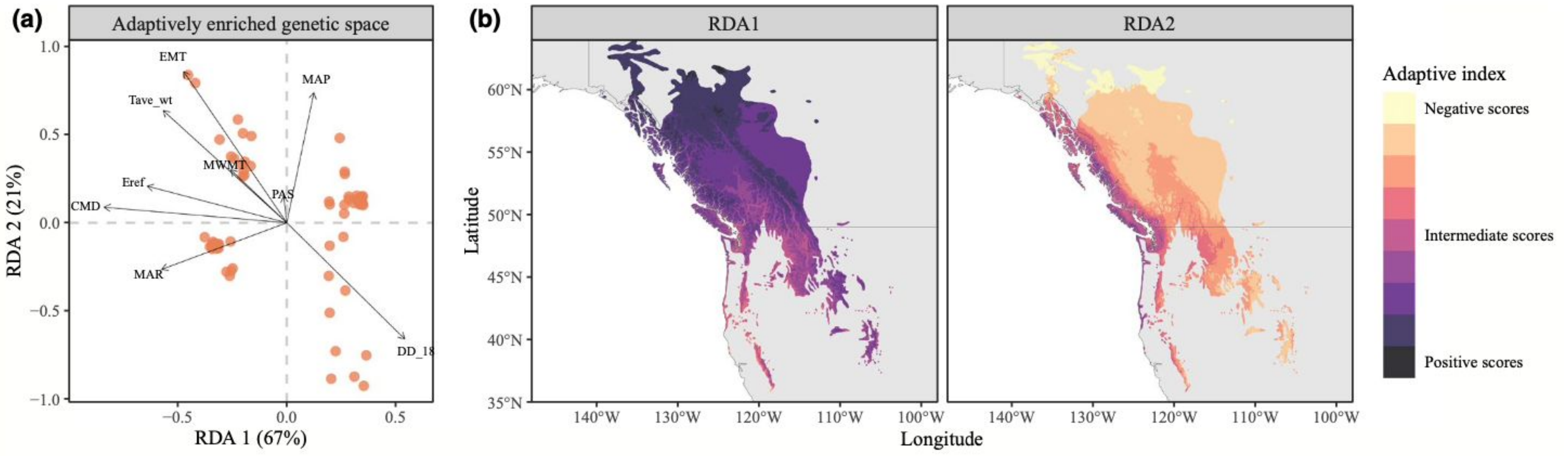
Bay et al. (2018) *Science*

What questions can we answer using GEA?



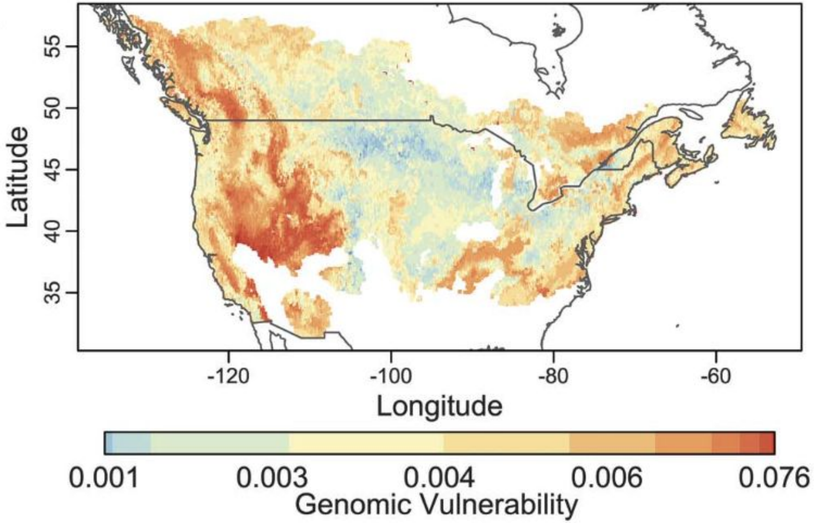
Ruegg et al. (2018) *Ecology Letters*

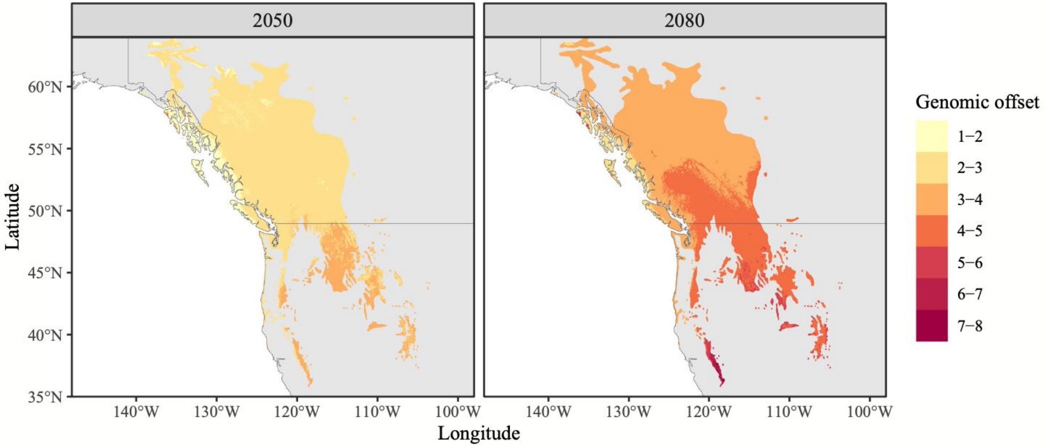
What questions can we answer using GEA?



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

What questions can we answer using GEA?





Bay et al. (2018) *Science*; Capblancq & Forester (2021) *Methods Ecol. Evol.*

Different types of GEA

- BayEnv/BayPASS/BayeScEnv
- Redundancy analysis (RDA)
- Latent factor mixed models (LFMM)
- GLMM
- Gradient 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
GF	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; if lots of data are imputed double-check the relationship between the strength of the association and % missingness (per site)

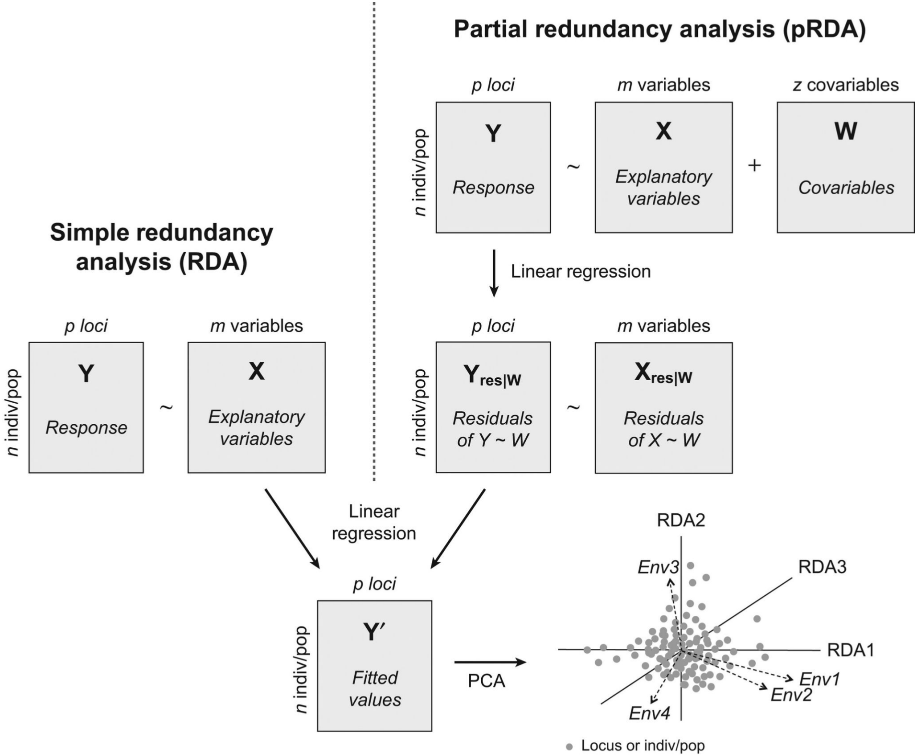
- Prune out sites that are in **linkage disequilibrium**

- Set a reasonable **MAF threshold**

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

Today's methods

Redundancy analysis (RDA)

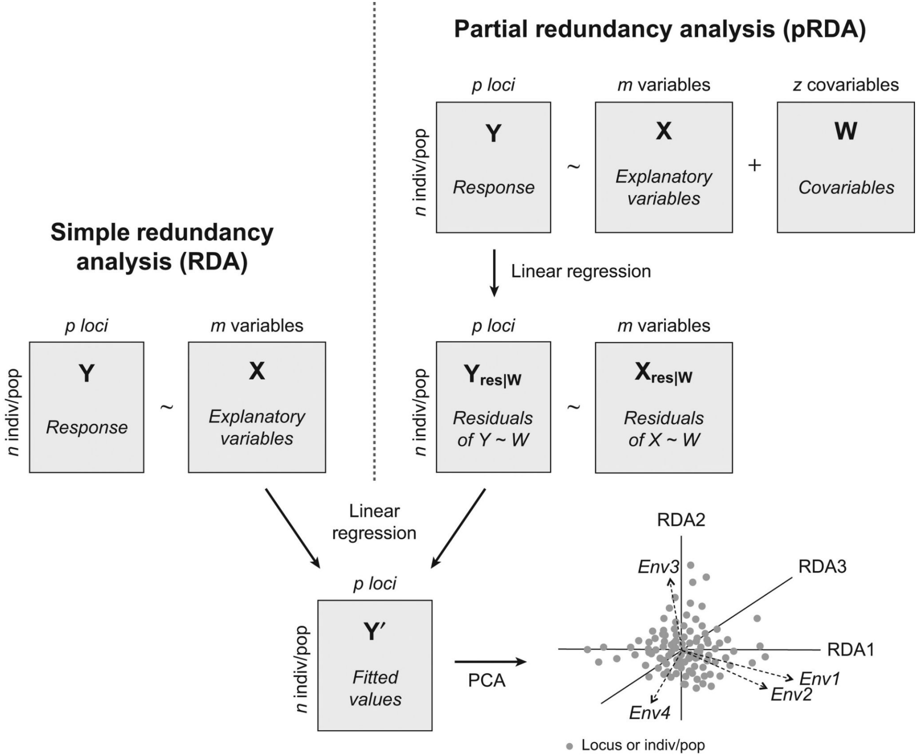


Outlier detection: RDadapt or Z-scores

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

Today's methods

Redundancy analysis (RDA)



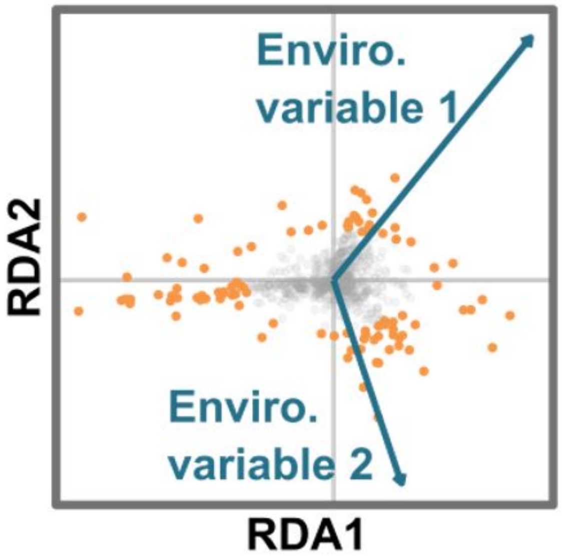
Latent factor mixed models (LFMM)

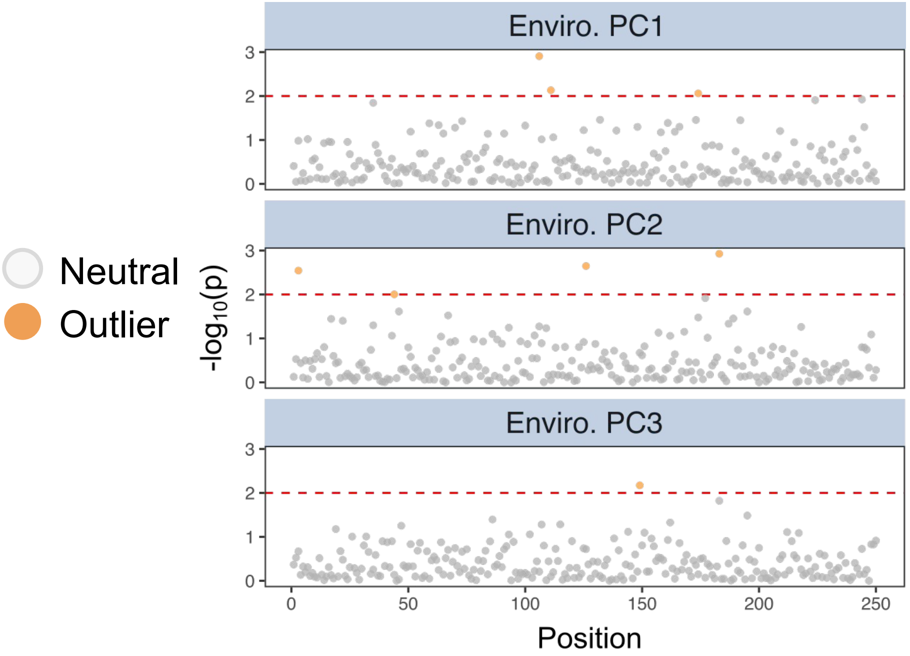
$$Y = XB^{T} + \underset{\substack{\uparrow \\ \text{Latent matrix}}}{W} + E$$

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

Redundancy analysis (RDA)

Latent factor mixed models (LFMM)





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

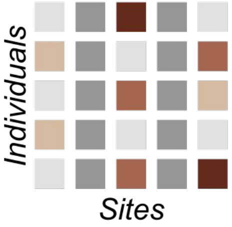
GEA: the logistics

Steps to perform GEA:

1. Gather genetic data
2. Prune out sites that are in linkage disequilibrium
3. Convert to dosage matrix
4. Impute missing values
5. Gather environmental data (or harvest from online given your sampling)
6. Extract environmental data for each sampling locality
7. Decide on model (and covariables)
8. Run the GEA method!

Genetic data processing

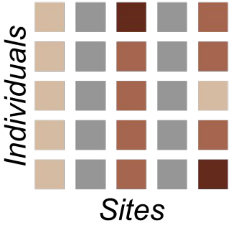
Cleaned genetic data



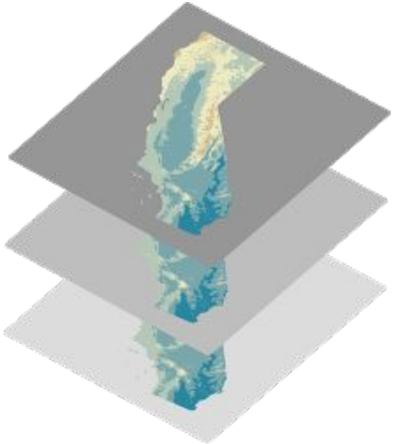
LD prune

Convert to dosage matrix

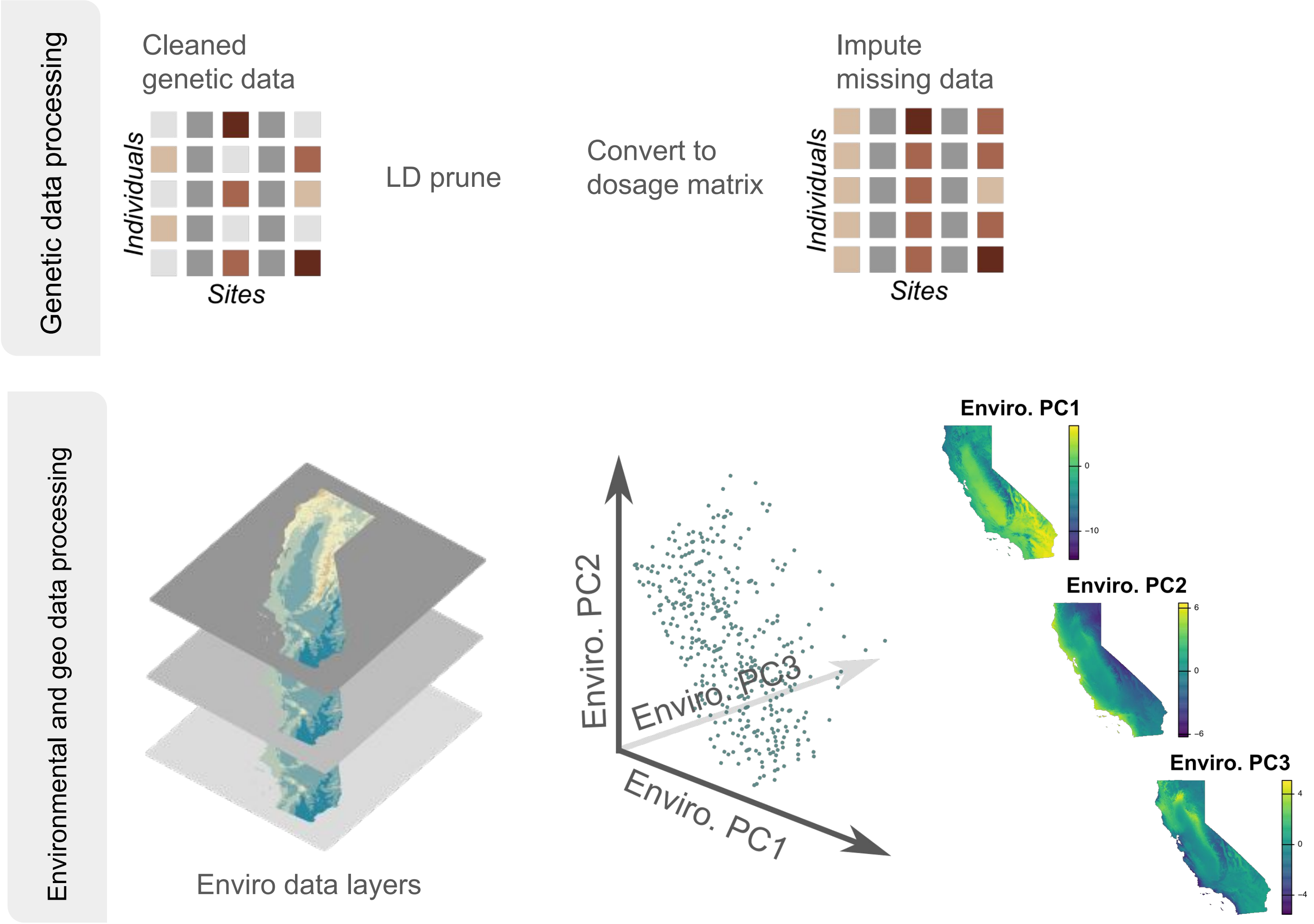
Impute missing data



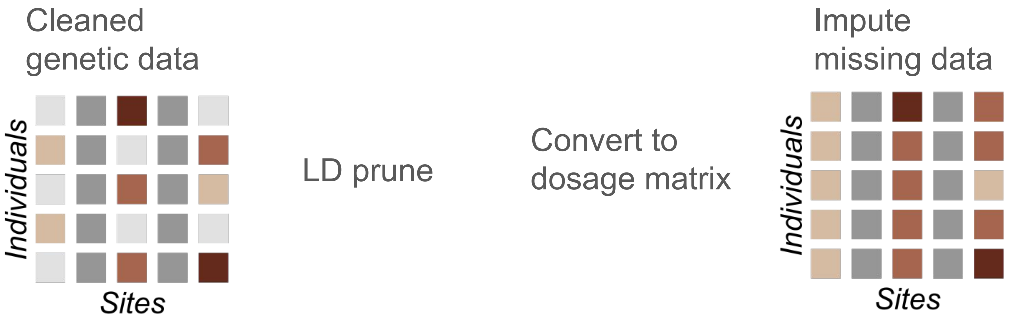
Environmental and geo data processing



Enviro data layers


Genetic data processing
Cleaned genetic data
Individuals
Sites
LD prune
Convert to dosage matrix
Impute missing data
Individuals
Sites
Environmental and geo data processing
Enviro data layers
Enviro. PC2
Enviro. PC3
Enviro. PC1
Enviro. PC1
0
−10
Enviro. PC2
6
0
−6
Enviro. PC3
4
0
−4

Genetic data processing



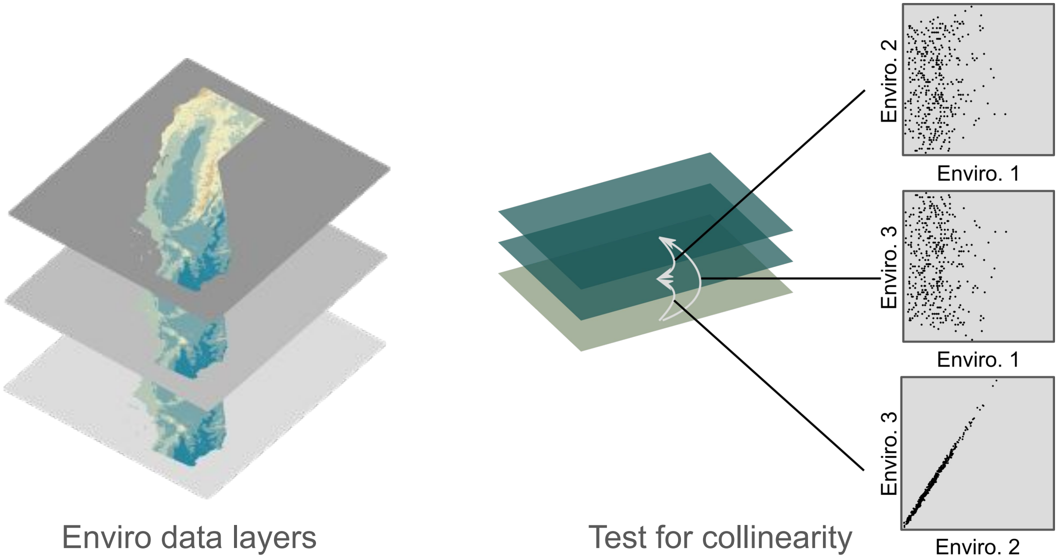
Cleaned genetic data

LD prune

Convert to dosage matrix

Impute missing data

Environmental and geo data processing

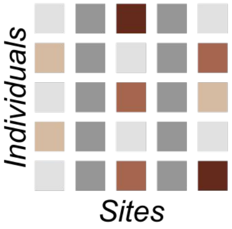


Enviro data layers

Test for collinearity

Genetic data processing

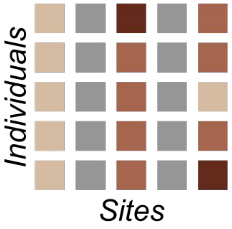
Cleaned genetic data



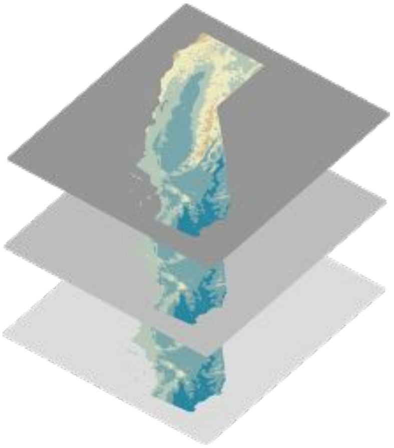
LD prune

Convert to dosage matrix

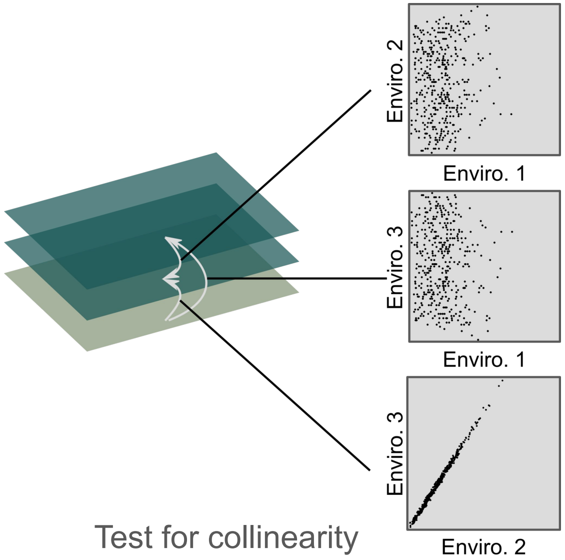
Impute missing data



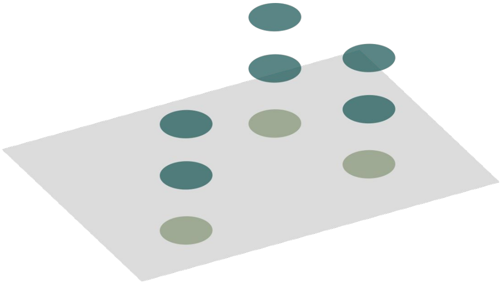
Environmental and geo data processing



Enviro data layers



Test for collinearity



Extract enviro vars at coordinates

Genetic diversity	wingen	
Population structure	TESS	
IBD and IBE	MMRR	GDM
Adaptive genetic variation	LFMM	RDA

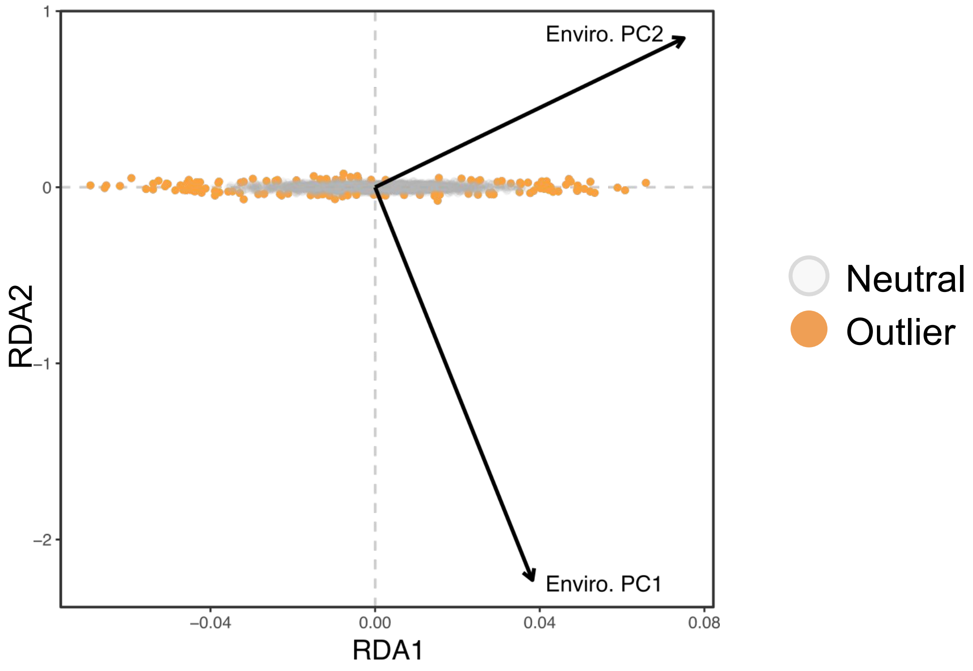
algatr

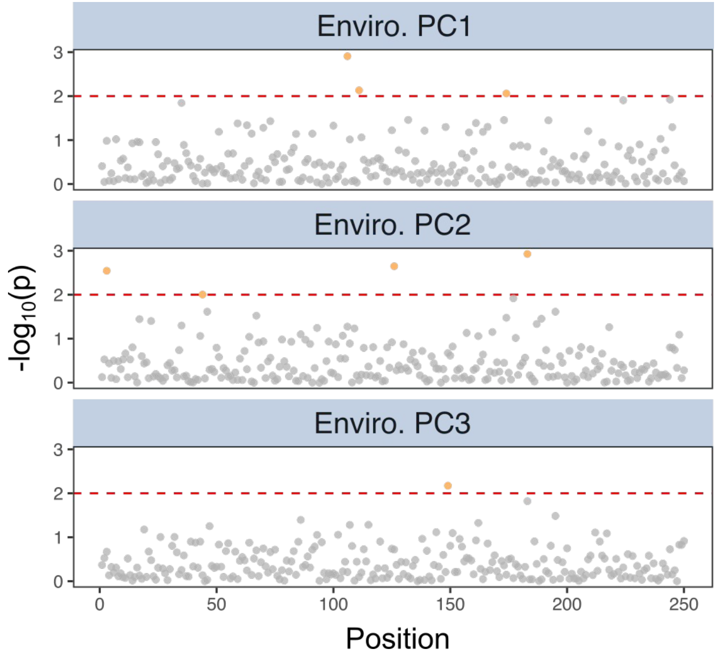
https://github.com/TheWangLab/algatr

Chambers et al. (2024) *Mol. Ecol. Res.*

Redundancy analysis (RDA)

Latent factor mixed models (LFMM)



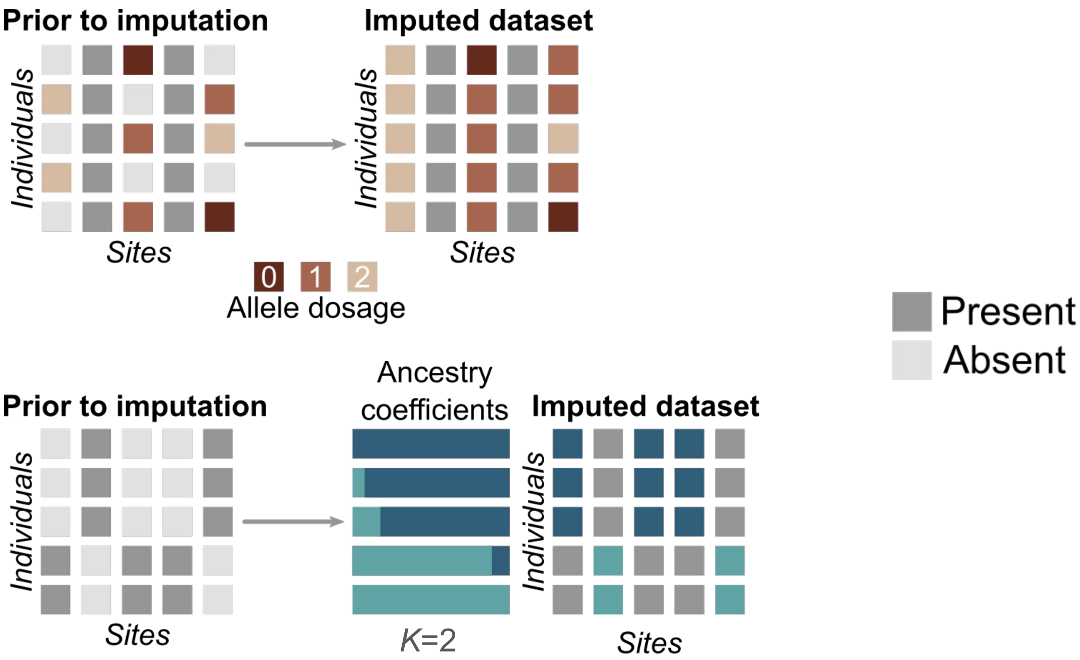


These methods can’t accept missing data! Two choices with different tradeoffs…

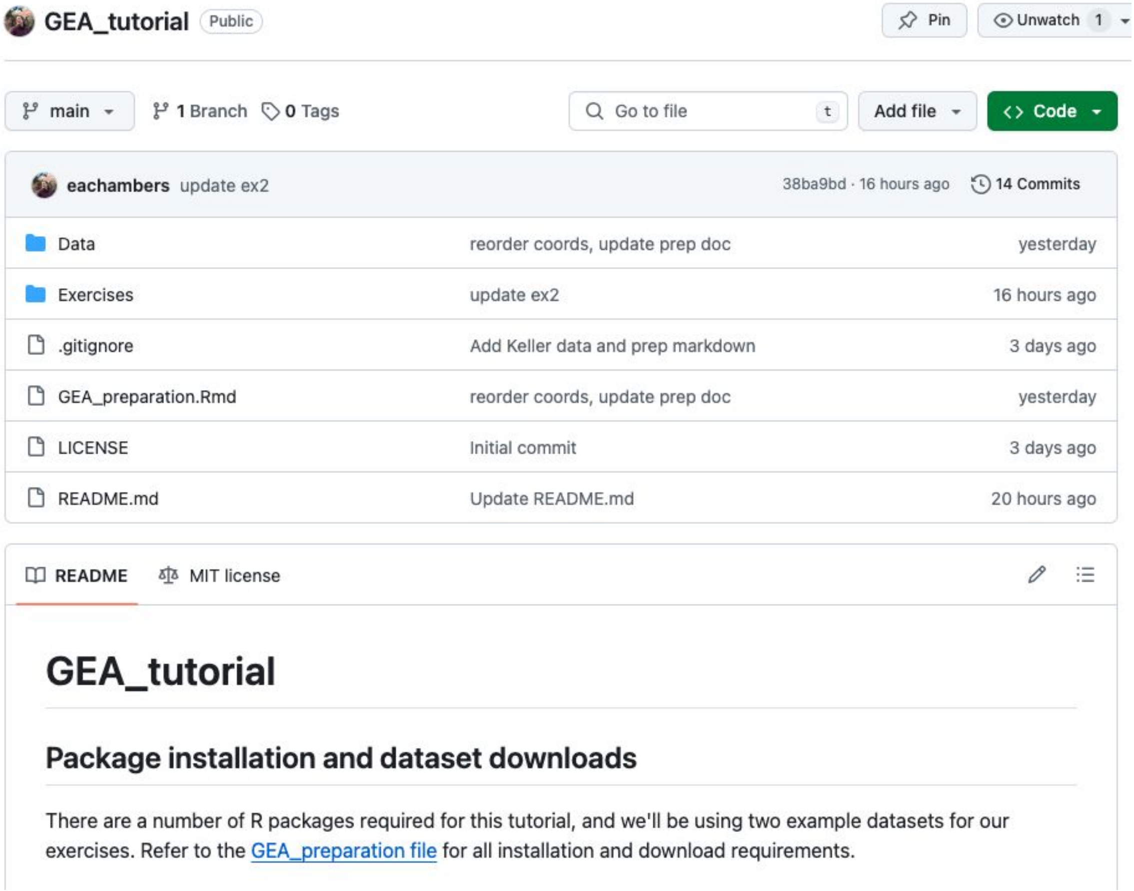
Imputation for GEA methods

Median-based

Structure-based



EXERCISE 1



GEA_tutorial Public

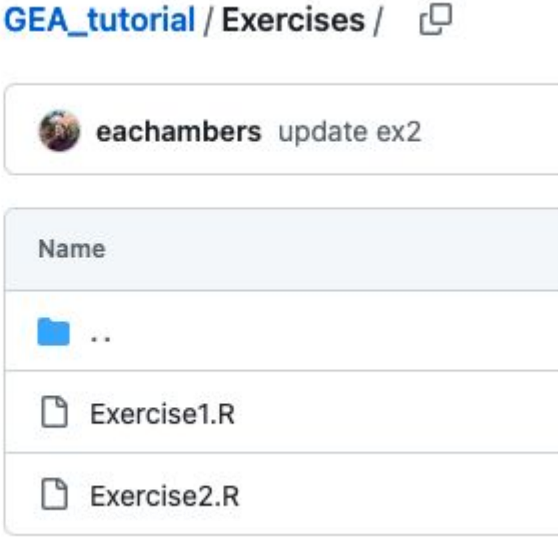
eachambers update ex2 38ba9bd · 16 hours ago 14 Commits

Data	reorder coords, update prep doc	yesterday
Exercises	update ex2	16 hours ago
.gitignore	Add Keller data and prep markdown	3 days ago
GEA_preparation.Rmd	reorder coords, update prep doc	yesterday
LICENSE	Initial commit	3 days ago
README.md	Update README.md	20 hours ago

GEA_tutorial

Package installation and dataset downloads

There are a number of R packages required for this tutorial, and we'll be using two example datasets for our exercises. Refer to the GEA_preparation file for all installation and download requirements.



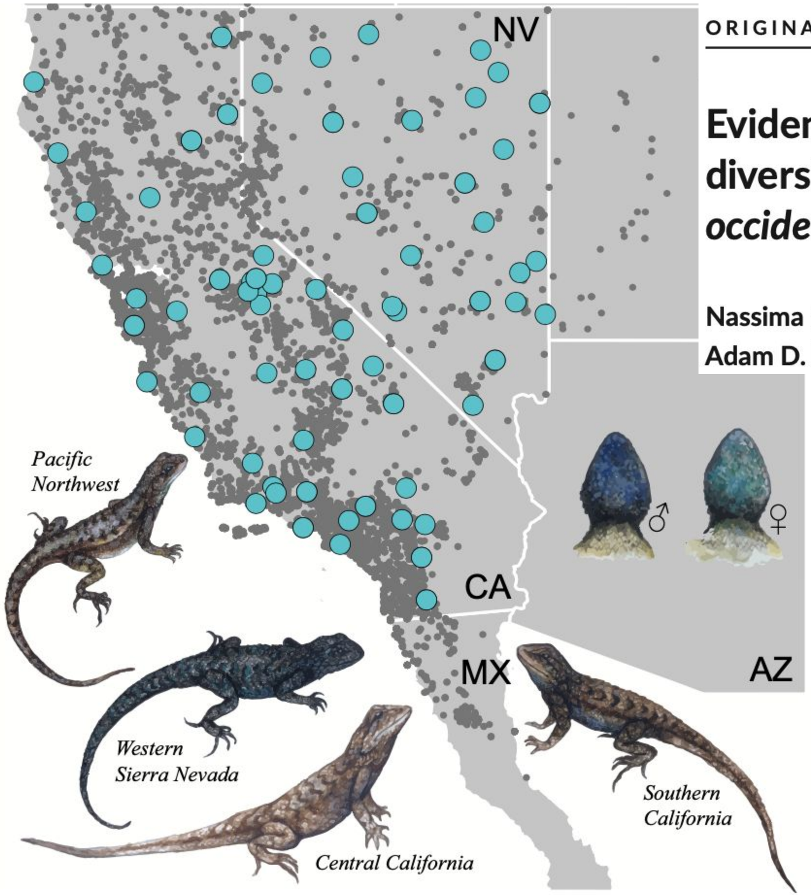
GEA_tutorial / Exercises /

eachambers update ex2

Name
..
Exercise1.R
Exercise2.R

https://github.com/eachambers/GEA_tutorial

Example dataset: Bouzid et al. (2022)



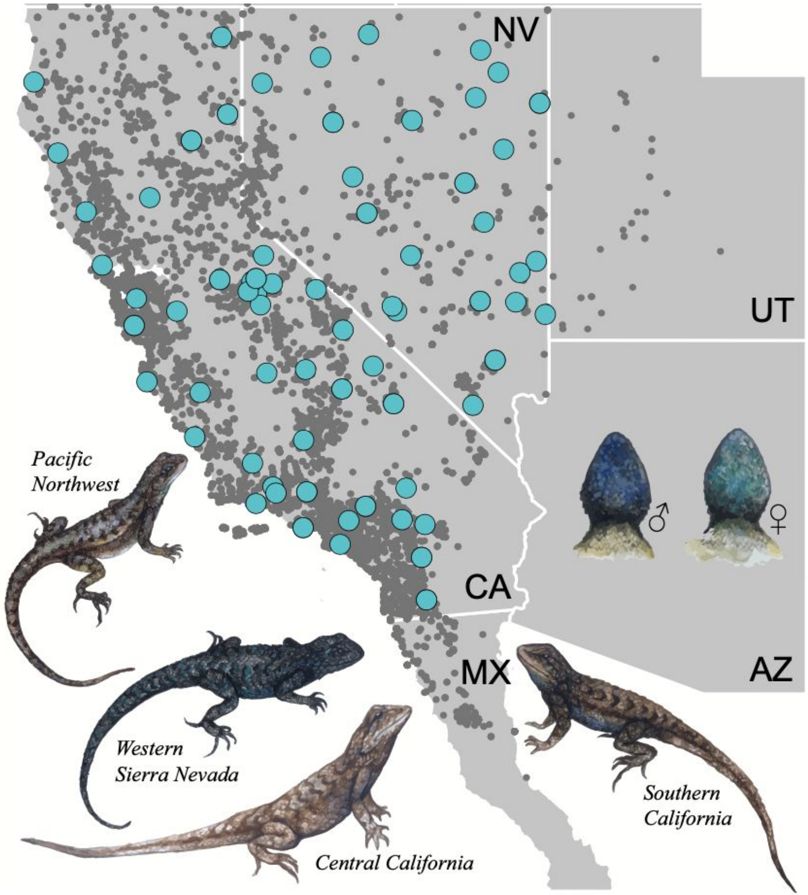
ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Evidence for ephemeral ring species formation during the diversification history of western fence lizards (*Sceloporus occidentalis*)

Nassima M. Bouzid[1] | James W. Archie[2] | Roger A. Anderson[3] | Jared A. Grummer[4] | Adam D. Leaché[1]

Example dataset: Bouzid et al. (2022)



Genetic data:
53 individuals (53 localities)
Individual-based sampling
1,000 SNPs (ddRAD data)

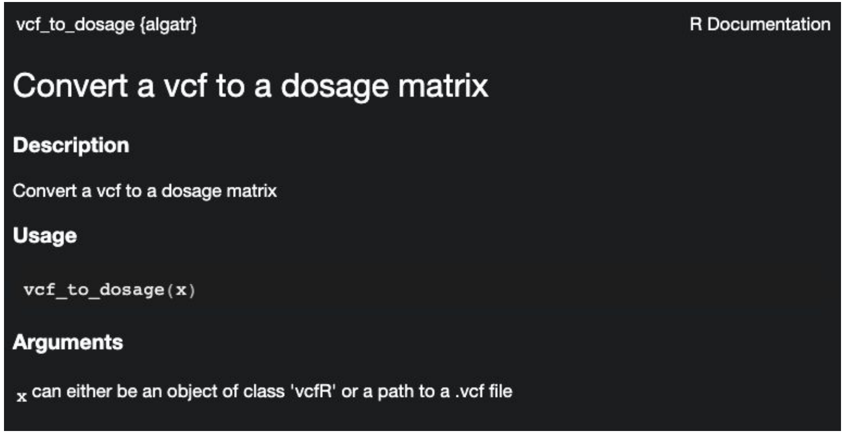
Environmental data:
We're going to gather some for this dataset!

Exercise 1

```
?vcf_to_dosage
```

Exercise 1

```
?vcf_to_dosage
```



vcf_to_dosage {algatr}　R Documentation

Convert a vcf to a dosage matrix

Description

Convert a vcf to a dosage matrix

Usage

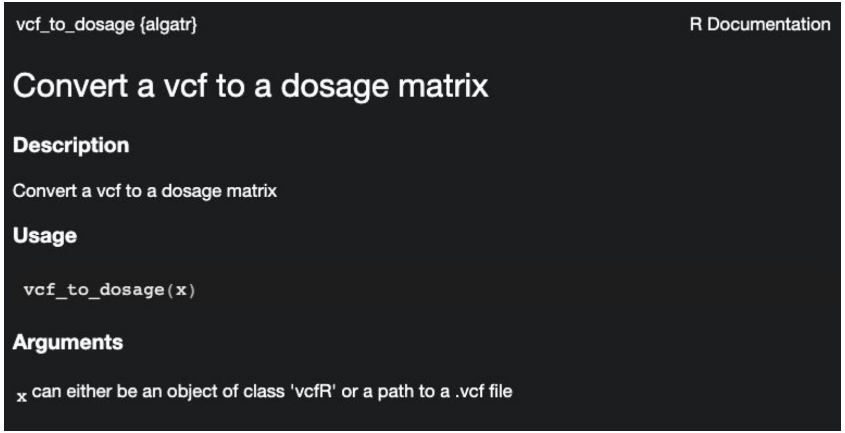
```
vcf_to_dosage(x)
```

Arguments

`x` can either be an object of class 'vcfR' or a path to a .vcf file

Exercise 1

`?vcf_to_dosage`



vcf_to_dosage {algatr} R Documentation

Convert a vcf to a dosage matrix

Description

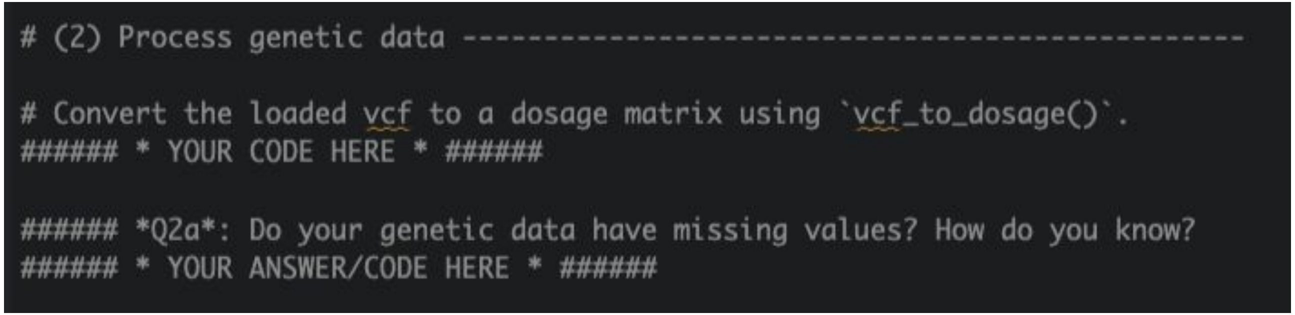
Convert a vcf to a dosage matrix

Usage

```
vcf_to_dosage(x)
```

Arguments

`x` can either be an object of class 'vcfR' or a path to a .vcf file

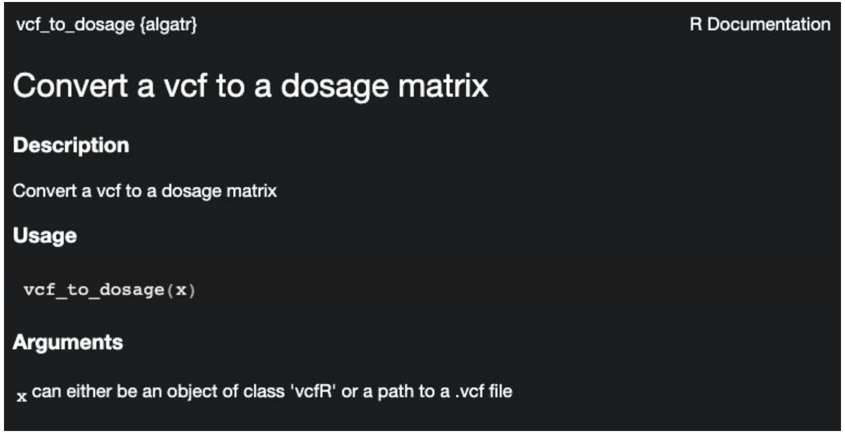


```
# (2) Process genetic data ------------------------------------------------

# Convert the loaded vcf to a dosage matrix using `vcf_to_dosage()`.
###### * YOUR CODE HERE * ######

###### *Q2a*: Do your genetic data have missing values? How do you know?
###### * YOUR ANSWER/CODE HERE * ######
```

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

Exercise 1

?vcf_to_dosage



vcf_to_dosage {algatr} R Documentation

Convert a vcf to a dosage matrix

Description

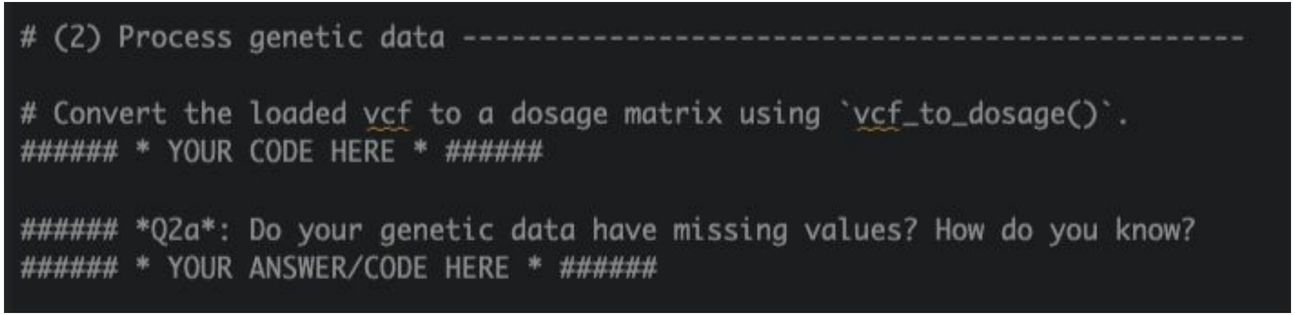
Convert a vcf to a dosage matrix

Usage

```
vcf_to_dosage(x)
```

Arguments

`x` can either be an object of class 'vcfR' or a path to a .vcf file



```
# (2) Process genetic data -----------------------------------------------

# Convert the loaded vcf to a dosage matrix using `vcf_to_dosage()`.
###### * YOUR CODE HERE * ######

###### *Q2a*: Do your genetic data have missing values? How do you know?
###### * YOUR ANSWER/CODE HERE * ######
```

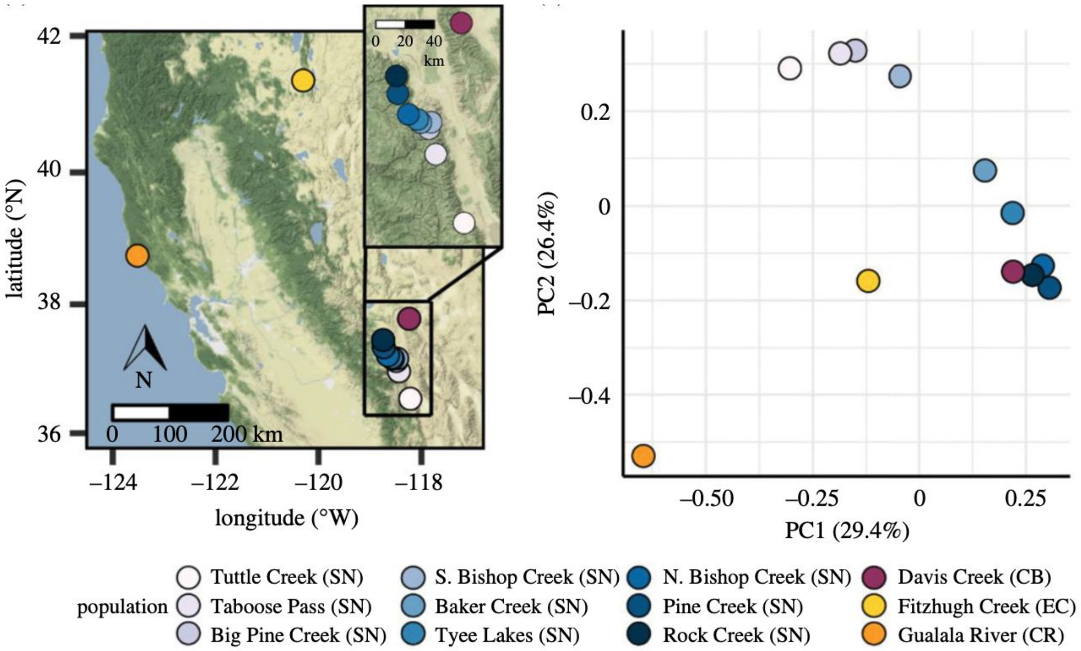
https://thewanglab.github.io/algatr/articles/RDA_vignette.html

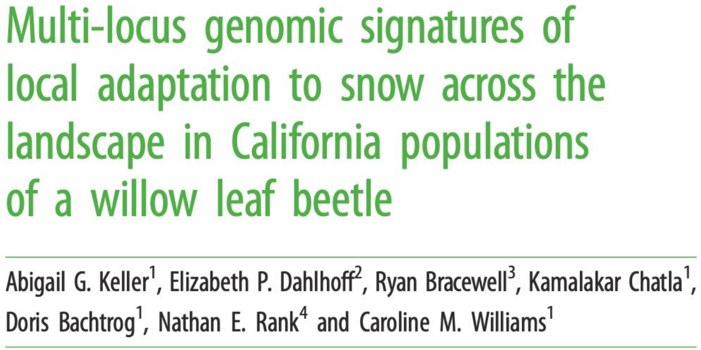
Exercise 1

1. Load the example dataset
2. Process genetic data:
   a. Convert vcf to dosage using `vcf_to_dosage()`
   b. Impute missing values using structure-based imputation using `str_impute()`
3. Process environmental data:
   a. Extract environmental values using coordinates using `raster::extract()`
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. Build a Manhattan plot and an RDA biplot of results using `rda_plot()`

EXERCISE 2

Example dataset: Keller et al. (2023)

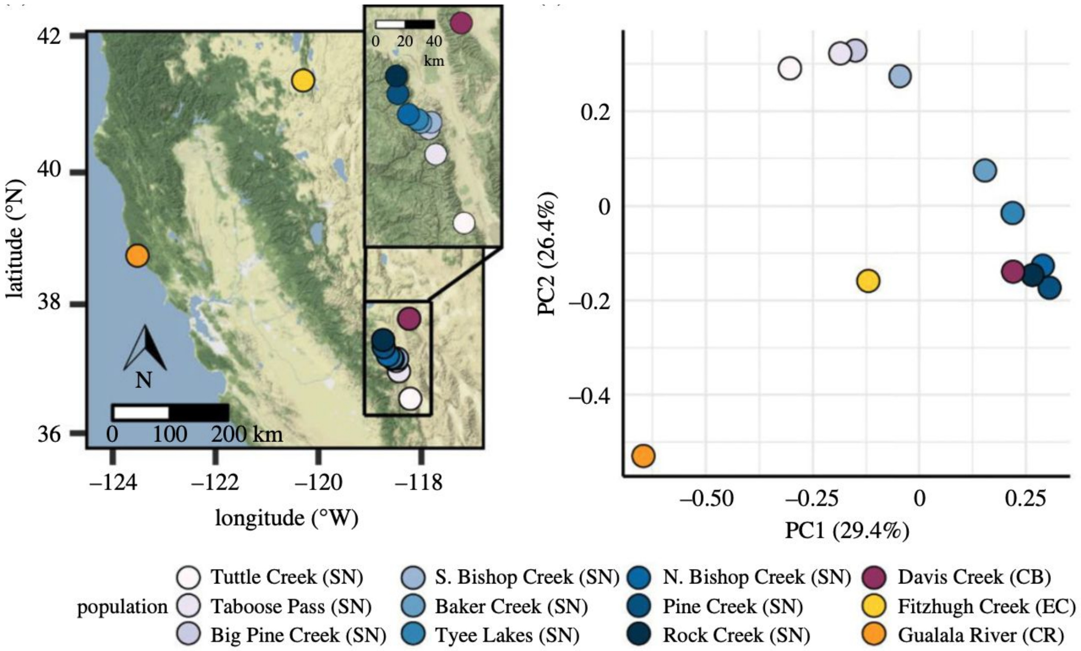




Multi-locus genomic signatures of local adaptation to snow across the landscape in California populations of a willow leaf beetle

Abigail G. Keller[1], Elizabeth P. Dahlhoff[2], Ryan Bracewell[3], Kamalakar Chatla[1], Doris Bachtrog[1], Nathan E. Rank[4] and Caroline M. Williams[1]

Example dataset: Keller et al. (2023)



Genetic data:

175 individuals (12 populations)

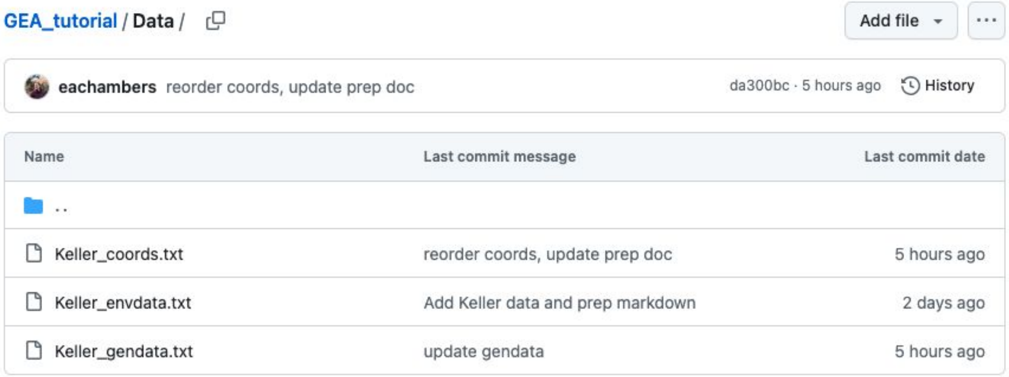
Site-based sampling

22,323 SNPs (WGS data)

Environmental data:

Point data for each population

Example dataset: Keller et al. (2023)



GEA_tutorial / Data /

eachambers reorder coords, update prep doc da300bc · 5 hours ago

Name	Last commit message	Last commit date
..		
Keller_coords.txt	reorder coords, update prep doc	5 hours ago
Keller_envdata.txt	Add Keller data and prep markdown	2 days ago
Keller_gendata.txt	update gendata	5 hours ago

https://github.com/eachambers/GEA_tutorial/tree/main/Data

Genetic data:

175 individuals (12 populations)

Site-based sampling

22,323 SNPs (WGS data)

Environmental data:

Point data for each population

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

Exercise 2

1. Import and process data using the tidyverse
2. Impute missing genetic data using the median using `simple_impute()`
3. Perform two types of *K* selection to determine how many latent factors you want to use with `select_K()`
4. Run LFMM using `lfmm_run()`
5. Get summary statistics with `lfmm_table()`
6. Make a Manhattan plot of the results using `lfmm_manhattanplot()`