Session IV: Landscape Genomics Pipeline Development & Analysis Overview

The Landscape Genomics Analysis Pipeline for the CCGP

E. Anne Chambers
Anusha Bishop
Ian J. Wang

Berkeley
UNIVERSITY OF CALIFORNIA

CCGP Landscape Genomics Working Group

Monthly meetings over the past year to discuss best approaches, environmental data, specific methods and goals


Victoria Sork


Jason Sexton


Rachael Bay


Erin Toffelmier


Erik Enbody


Ryan Harrigan

Goals

- Characterize spatial genetic variation from across the state
- Map “hotspots” of genetic diversity
- Identify regions, habitats, or landscape features that facilitate or impede population connectivity
- Identify genes or regions of the genome involved in climate adaptation
- Assess vulnerability to climate change and other anthropogenic impacts on natural systems

Challenges

Suitable methods:

- Must work with individual-based sampling (or be adaptable)
- Be computationally tractable for ~150 species x 150 individuals with WGS
- Can be applied to all CCGP species in a consistent way
- Produce output that is comparable across a very diverse set of species
- Must be developed if nothing meeting these criteria currently e

Mission

Our goal was to identify, adapt, and develop a set of methods that could be applied to all CCGP species and that would produce informative output that would allow for downstream comparisons.

Structure How is **genetic variation structured** spatially?

Structure How is **genetic variation structured** spatially?

IBD/IBE What are the effects of **geography and environment** on genetic differentiation?

Structure How is **genetic variation structured** spatially?

IBD/IBE What are the effects of **geography and environment** on genetic differentiation?

GEA What regions of the genome show evidence of **climate associations**?

Structure	How is **genetic variation structured** spatially?
IBD/IBE	What are the effects of **geography and environment** on genetic differentiation?
GEA	What regions of the genome show evidence of **climate associations**?
Diversity	How is intraspecific **genetic diversity** distributed spatially?

A Landscape Genomics Analysis Toolkit in R

A Landscape Genomics Analysis Toolkit in R

algatr

GDM_vignette

Generalized dissimilarity modeling (GDM)

```
library(algatr)
library(gdm)
library(here)
library(tidyverse)
library(raster)
library(rgdal)
library(readr)
devtools::load_all()
```

Generalized dissimilarity modeling is a matrix regression method in which explanatory variables (in our case, genetic data, in the form of a distance matrix) is regressed against a response matrix (environmental variables for sites from which samples were obtained). A GDM calculates the compositional dissimilarity between pairs of sites, and importantly takes into account the fact that genetic data varies nonlinearly across an environmental gradient.

For additional information on GDMs, please see Ferrier et al. 2007 for a description of its basic use in estimating patterns of beta diversity, Freedman et al. 2010 for a classic example of its use, and Fitzpatrick & Keller 2015 for a perspective piece on its applications. Finally, our code primarily uses the gdm package which has excellent documentation (see here).

There is one main function to perform a GDM analysis: `gdm_do_everything()`. This function runs the GDM (using the `gdm()` function within the gdm package), and allows a user to run a GDM with all variables, or with model selection to choose the best-supported variables. This function produces information on the final model, and coefficients for predictor variables.

We can also use the `gdm_plot_isplines()` function to plot I-splines for each environmental variable and geographic distance, and `gdm_map()` to produce a PCA map with GDM results plotted.

There are a few assumptions built within this function that the user must be aware of: (1) the coords and gendist files MUST have the same ordering of individuals; there isn't a check for this, and (2) this function assumes individual-based sampling and that each individual is a sampling site.

Read in and process data files

Running a GDM requires three data files for input: a genetic distance matrix (the `gendist` argument), coordinates for samples (the `coords` argument), and environmental layers on which to run the GDM (the `envlayers` argument).

```
# Load genetic dist matrix and coordinates for 53 inds, and three environmental layers for test a
load_example()
#>
#> ---------------- example dataset ----------------
#>
#> Objects loaded:
#> *liz_vcf* vcfR object (1000 loci x 53 samples)
```

algatr

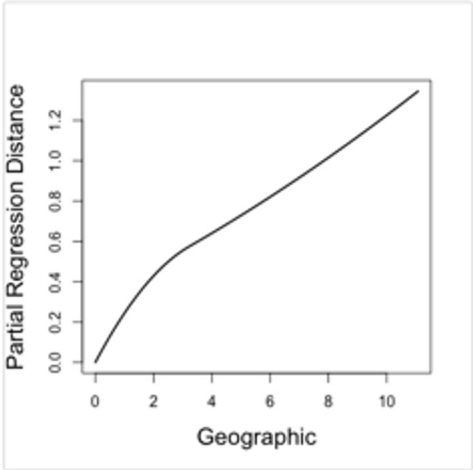
Run GDM

Given that GDM is a regression, the full model (i.e., including all predictor variables) will include all environmental layers in addition to geographic distance, which is also considered a predictor. Thus, in this example, the maximum number of variables you can end up with that are significant is four (three enviro PCs + geographic distance).

GDM with all variables

Let's first run a full GDM model (i.e., including all four variables), specified using the `model` argument. If you have extracted environmental values for each sampling coordinate, this must be specified using the `env` argument, and if genetic distances are not bounded by 0-1, they must be scaled using the `scale` argument. Keep in mind that the `nPerm` argument is only for model selection (see below) and so will not be used in this case.

```
gdm_full <- gdm_do_everything(gendist, liz_coords, CA_env, model = "full", alpha = 0.05, scale =
```



GDM_vignette

Generalized dissimilarity modeling (GDM)

```
library(algatr)
library(gdm)
library(here)
library(tidyverse)
library(raster)
library(rgdal)
library(readr)
devtools::load_all()
```

Generalized dissimilarity modeling is a matrix regression method in which explanatory variables (in our case, genetic data, in the form of a distance matrix) is regressed against a response matrix (environmental variables for sites from which samples were obtained). A GDM calculates the compositional dissimilarity between pairs of sites, and importantly takes into account the fact that genetic data varies nonlinearly across an environmental gradient.

For additional information on GDMs, please see Ferrier et al. 2007 for a description of its basic use in estimating patterns of beta diversity, Freedman et al. 2010 for a classic example of its use, and Fitzpatrick & Keller 2015 for a perspective piece on its applications. Finally, our code primarily uses the gdm package which has excellent documentation (see here).

There is one main functi... (using the gdm() functi... model selection to choo... and coefficients for pre...

We can also use the gd... geographic distance, a...

There are a few assump... files MUST have the sa... individual-based sampl...

Read in and process data files

Running a GDM requires three data files for input: a genetic distance matrix (the gendist argument), coordinates for samples (the coords argument), and environmental layers on which to run the GDM (the envlayers argument).

```
# Load genetic dist matrix and coordinates for 53 inds, and three environmental layers for test
load_example()
#>
#> ---------------- example dataset ----------------
#>
#> Objects loaded:
#> *liz_vcf* vcfR object (1000 loci x 53 samples)
```

Run GDM

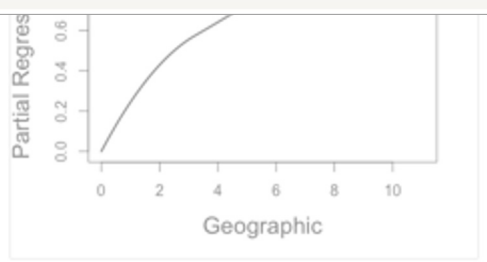
Given that GDM is a regression, the full model (i.e., including all predictor variables) will include all environmental layers in addition to geographic distance, which is also considered a predictor. Thus, in this example, the maximum number of variables you can end up with that are significant is four (three enviro PCs + geographic distance).

GDM with all variables

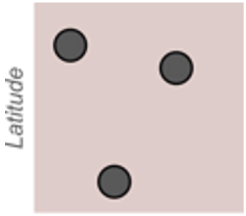
Let's first run a full GDM model (i.e., including all four variables), specified using the model argument. If you have extracted environmental values for each sampling coordinate, this must be specified using the env argument, and if genetic distances are not bounded by 0-1, they must be scaled using the scale argument. Keep in mind that the nPerm argument is only for model selection (see below) and so will not be used in this case.

```
gdm_full <- gdm_do_everything(gendist, liz_coords, CA_env, model = "full", alpha = 0.05, scale =
```

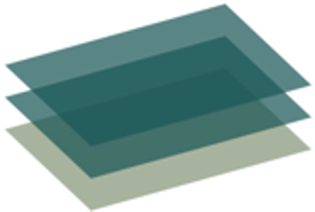
```
gdm_full <- gdm_do_everything(gendist, liz_coords, CA_env, model = "full", ...)
```



algatr



Coordinates

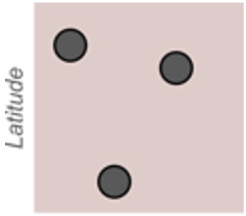


Enviro. layers

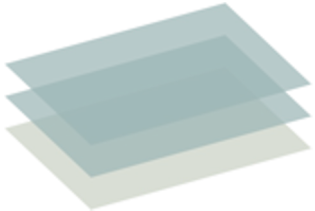
WGS data

algatr

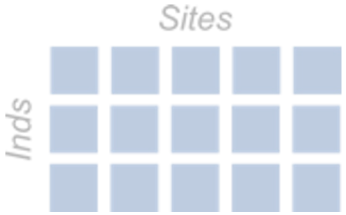
algatr



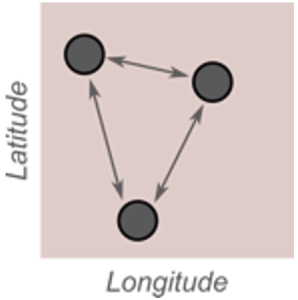
Coordinates

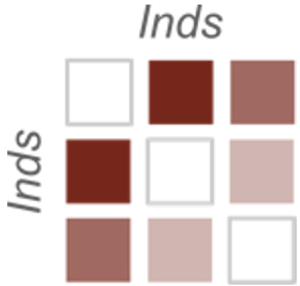


Enviro. layers

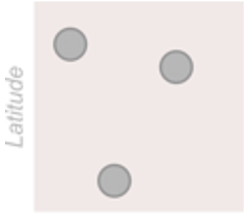


WGS data



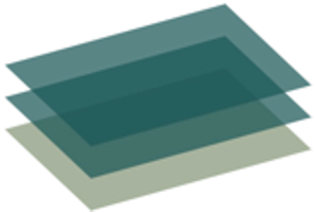


Calculate geographic distances

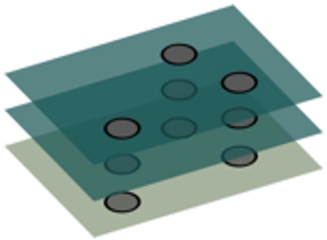


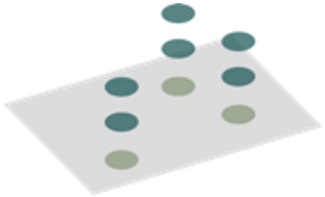
Coordinates

algatr

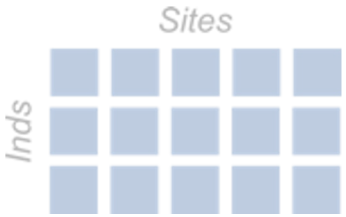


Enviro. layers

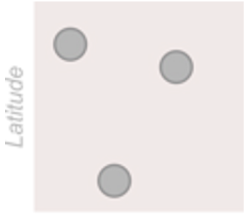




Extract variables at coordinates

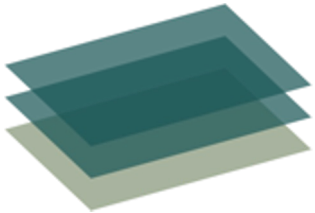


WGS data

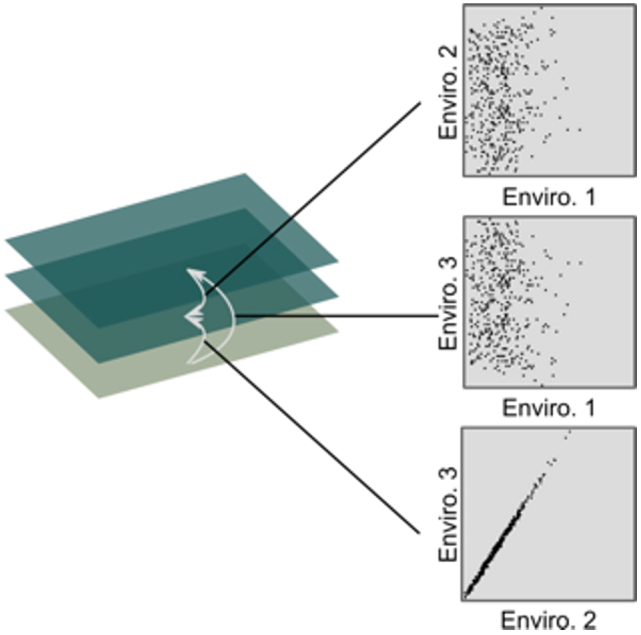


Coordinates

algatr

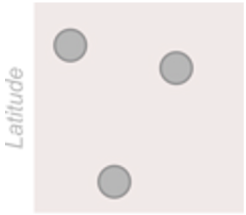


Enviro. layers

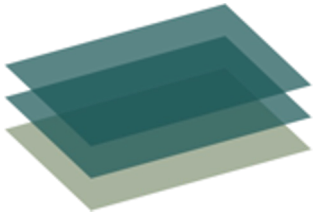


Detect collinearity among layers

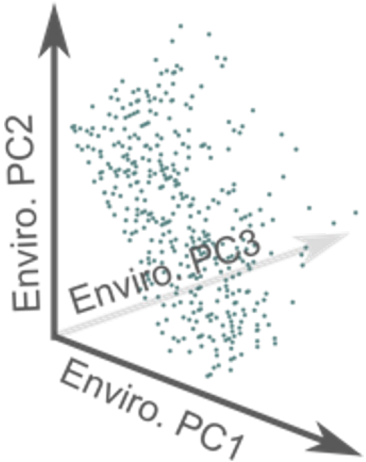
WGS data

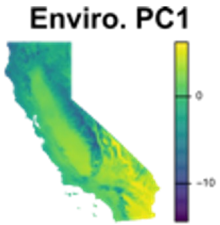


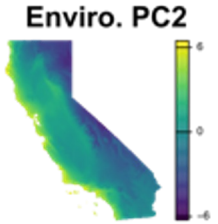
Coordinates

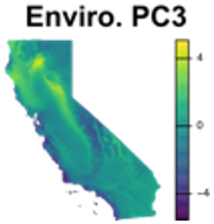


Enviro. layers

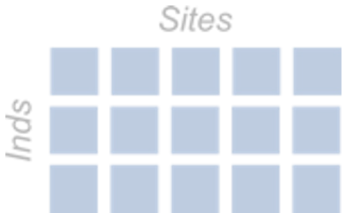






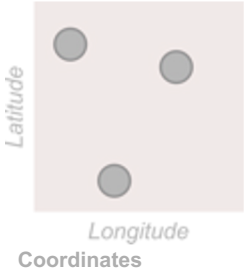


Perform raster PCA

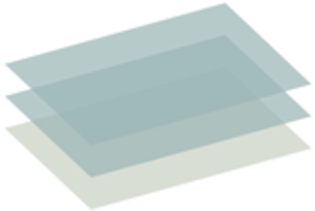


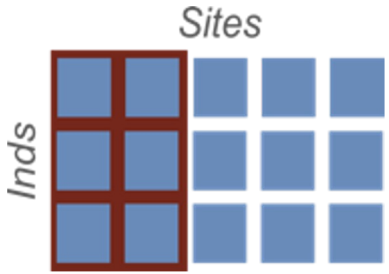
WGS data

algatr



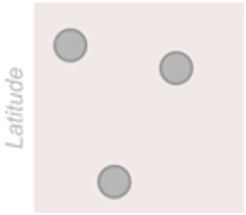
algatr



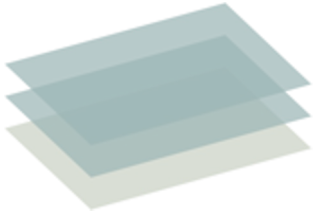


Remove sites in LD

WGS data

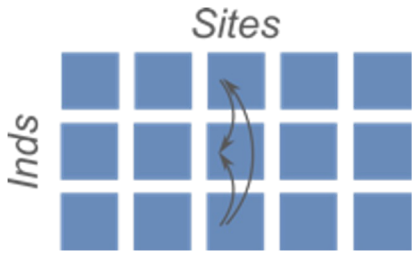


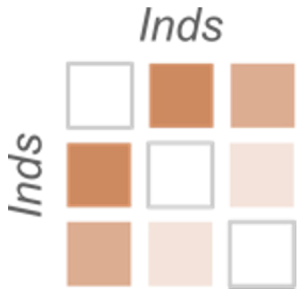
Coordinates



Enviro. layers

WGS data





Calculate genetic distances

algatr

algatr

Structure	How is **genetic variation structured** spatially?
IBD/IBE	What are the effects of **geography and environment** on genetic differentiation?
GEA	What regions of the genome show evidence of **climate associations**?
Diversity	How is intraspecific **genetic diversity** distributed spatially?

algatr

- **Structure** — TESS[1]
- IBD/IBE
- GEA
- Diversity

```
tess_do_everything(gen, coords, grid, Kvals, K_selection, ...)
```

[1] Caye et al. (2016)

algatr

Structure

IBD/IBE

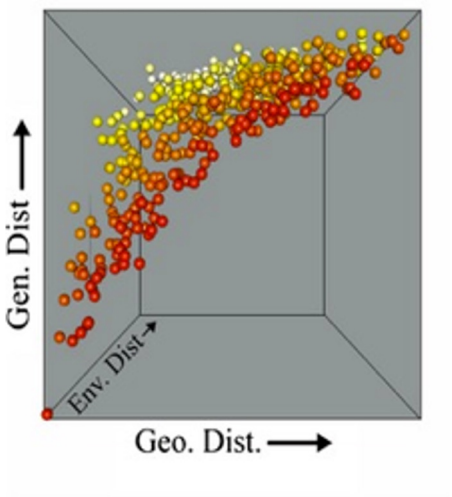
GEA

Diversity

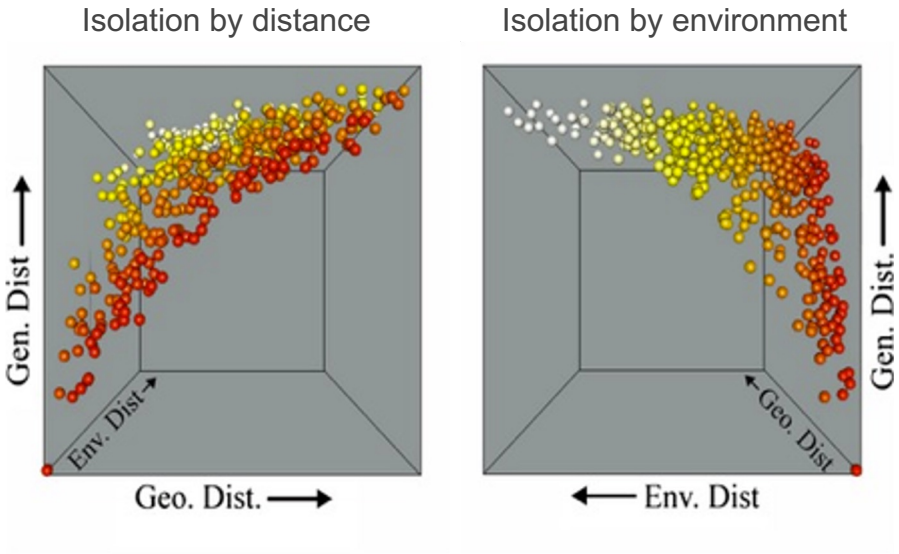
algatr

- Structure
- **IBD/IBE**
- GEA
- Diversity

Isolation by distance



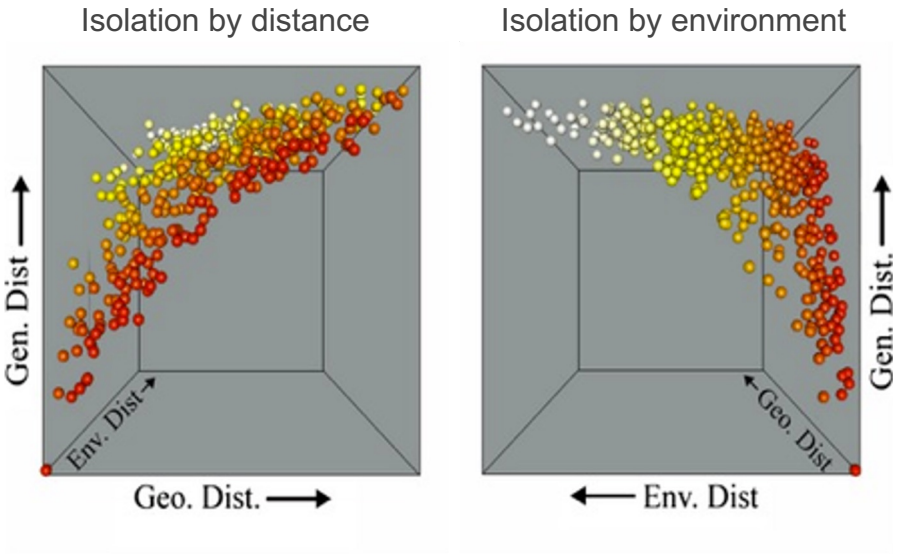
Wang & Bradburd (2014)

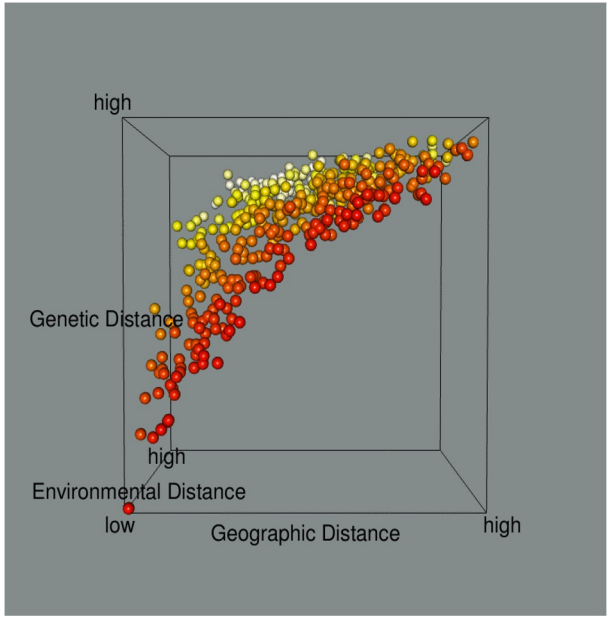


Wang & Bradburd (2014)

algatr

- Structure
- **IBD/IBE**
- GEA
- Diversity





Wang & Bradburd (2014)

algatr

- Structure
- **IBD/IBE**
- GEA
- Diversity

MMRR[2]

```
mmrr_do_everything(gendist, coords, envlayers, model = "best", ...)
```

Multiple matrix regression with randomization

[2]*Wang (2013)*

algatr

- Structure
- **IBD/IBE** — GDM[3]
- GEA
- Diversity

```
gdm_do_everything(gendist, coords, envlayers, model = "best", ...)
```

Generalized dissimilarity modeling

[3]*Ferrier et al. 2007; Freedman et al. 2010; Fitzpatrick & Keller 2015*

algatr

- Structure
- IBD/IBE
- **GEA**
- Diversity

RDA[4]

```
rda_do_everything(gen, coords, env, model = "best", ...)
```

Redundancy analysis

[4]*Capblancq & Forester (2021)*

algatr

- Structure
- IBD/IBE
- **GEA**
- Diversity

LFMM[5]

```
lfmm_do_everything(gen, env, coords, lfmm_method = “ridge”, Kvals, K_selection, ...)
```

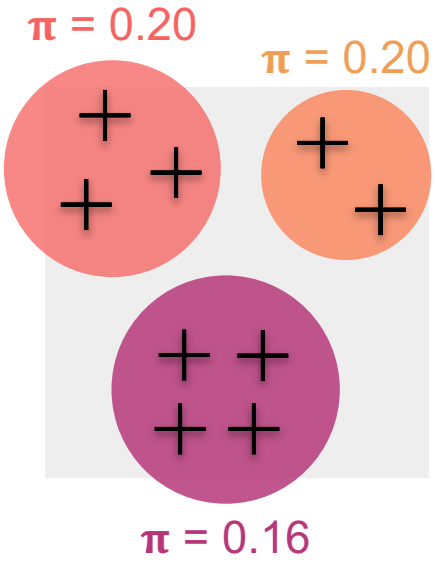
Latent factor mixed models

[5]*Caye et al. 2019*

wingen

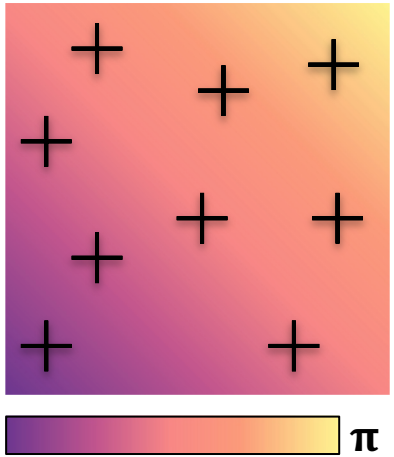
Continuous mapping of genetic diversity using moving windows

wingen

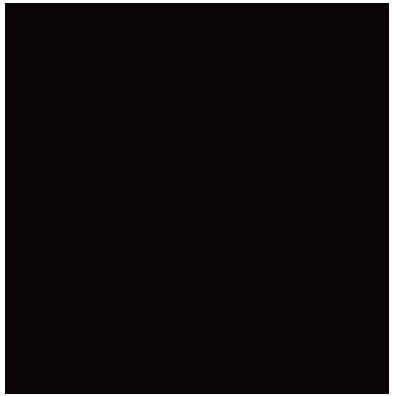


Traditional: calculating genetic diversity by population

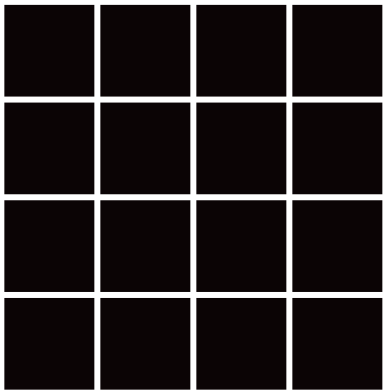
wingen



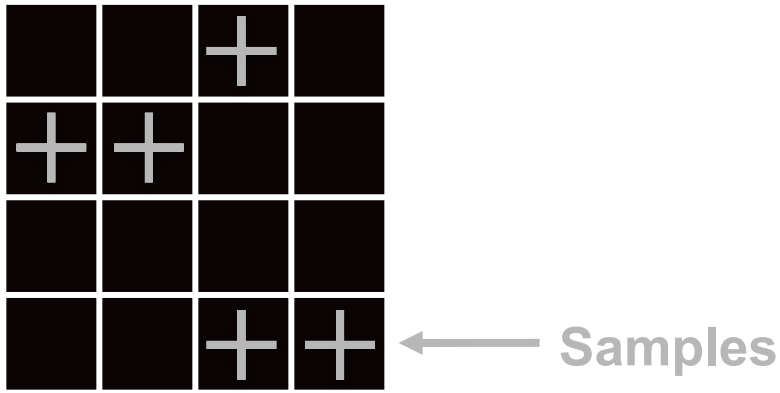
New: Taking advantage of individual based sampling to create continuous maps

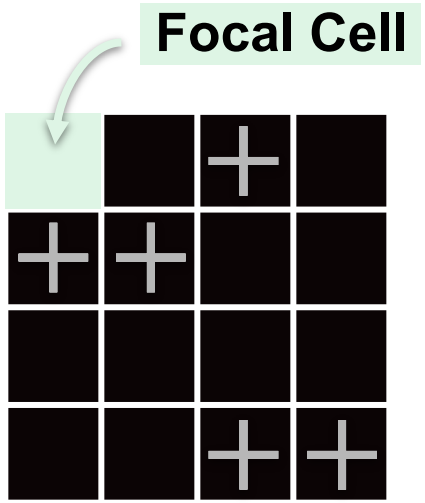


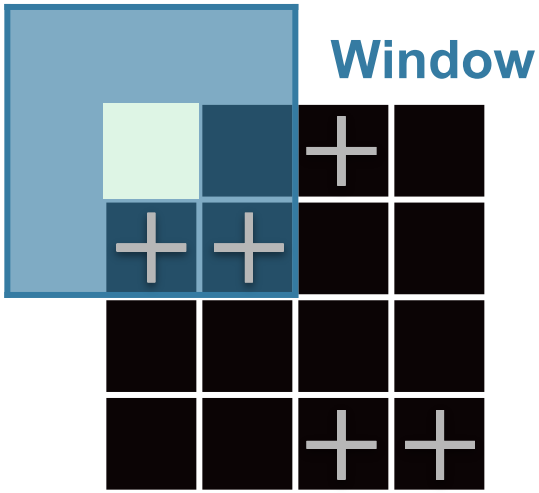
Example Landscape



Example
Landscape


Samples

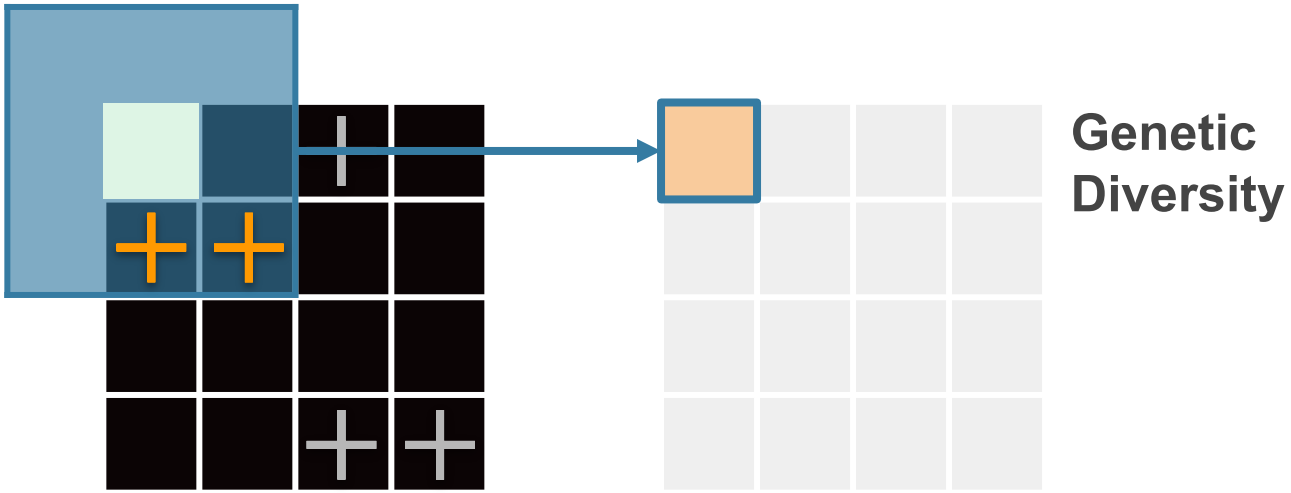

Focal Cell

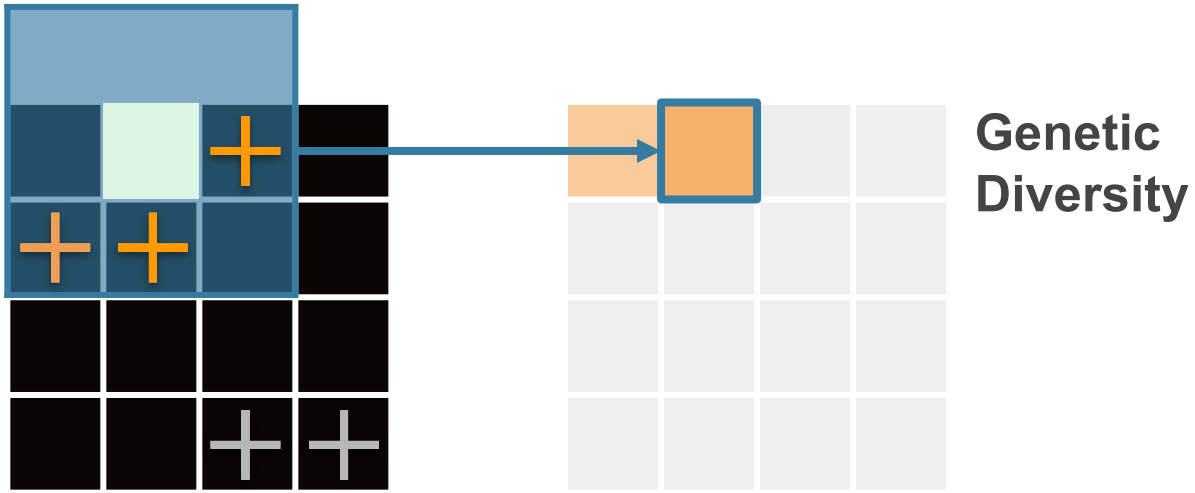

Window

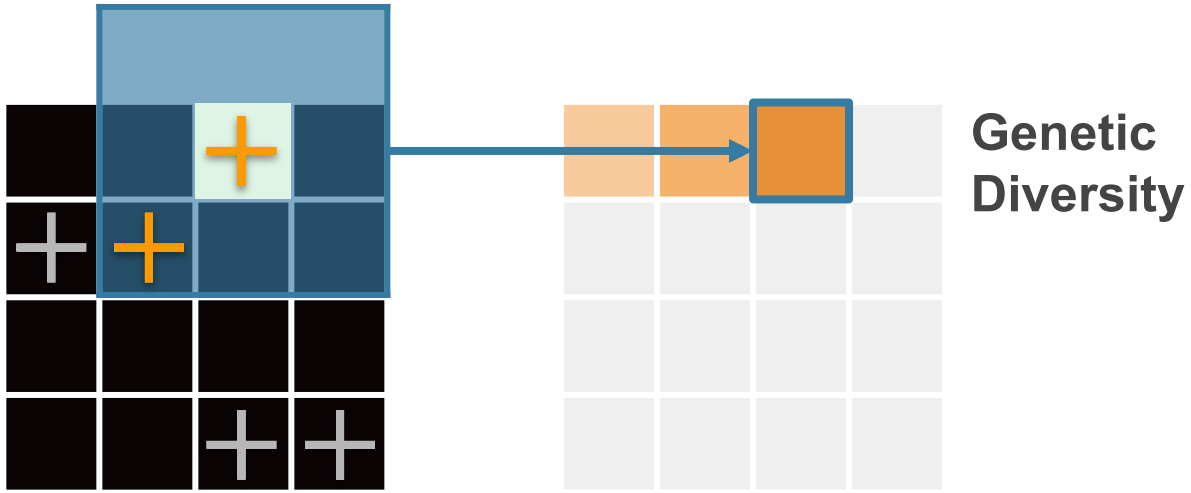
algatr

- Structure
- IBD/IBE
- GEA
- Diversity

wingen




Genetic
Diversity


Genetic
Diversity

algatr

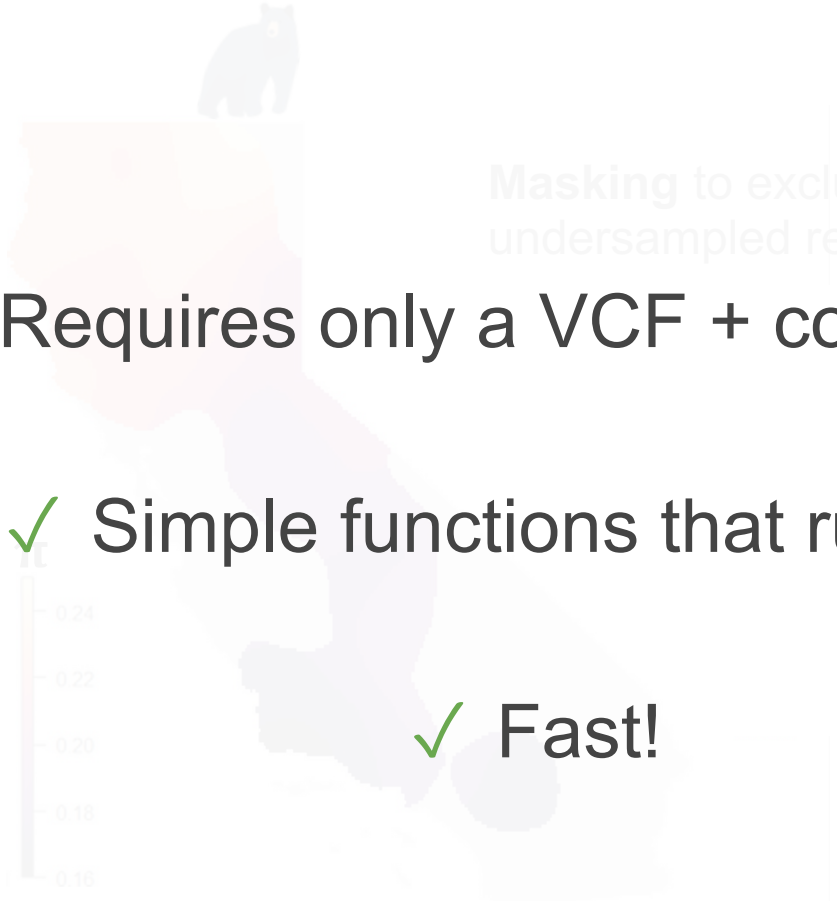
Structure

IBD/IBE

GEA

Diversity

wingen



Masking to exclude undersampled regions

✓ Requires only a VCF + coordinates

✓ Simple functions that run in R

✓ Fast!

algatr is easy to use and fast!

~~Fun~~ Alternative Fact: Alligators can run up to 35 mph.

```
#' @param buff A buffer area around sample points for cropping the data layers, expres
#' @param folder A directory in which to place downloaded worldclim data; if NULL then
#' @details
#' If res = 0.5 then the individual WorldClim tiles that cover the sample coordinates
#' The buffer area maintains a large extent for the final cropped data layers around t
#' @return A SpatRaster of WorldClim layers.
#' @export
#'
#' @examples
get_worldclim <- function(coords, res = 0.5, buff = 0.01, folder = NULL){
  # Raster of worldclim tiles
  r <- raster::raster(vals = 1:60, nrows = 5, ncols = 12, ext = raster::extent(c(-180,

  # Make SpatialPolygons object with convex hull of coords
  ch_pts <- chull(coords)
  ch_poly <- sp::Polygon(coords[ch_pts,])
  ch_polys <- sp::Polygons(list(ch_
  ch_spolys <- sp::SpatialPolygons(
```

26% downloaded

URL: ... u/climate/worldclim/2_1/tiles/tile/tile_14_wc2.1_30s_bio.tif

16:73 get_worldclim(coords, res, buff, folder) R Script

Console Terminal Background Jobs

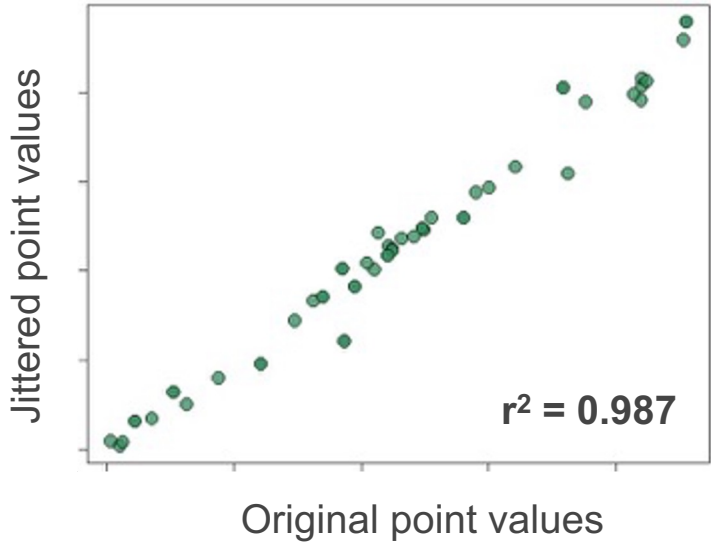
R 4.1.1 · C:/Users/Ian/Dropbox/Projects/algatr/algatr/

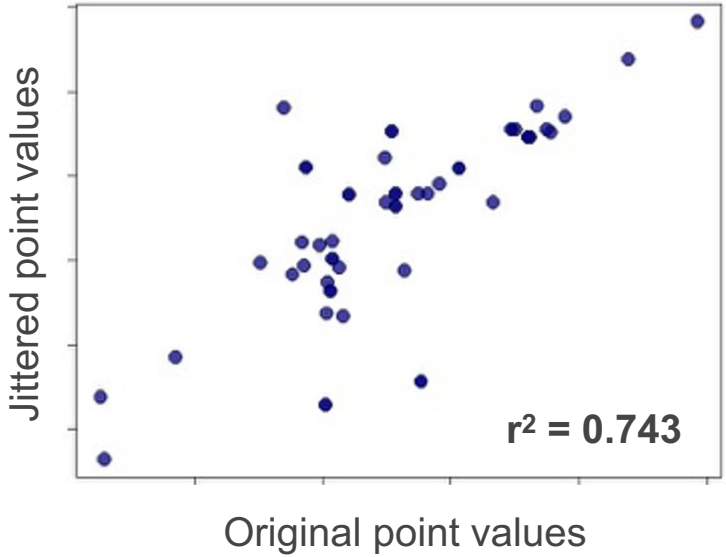
```
> wclim <- get_worldclim(liz_coords)
Downloading WorldClim tile 1...
Fun Fact: Alligators have a bite force of up to 2,980 psi.
```


```
Downloading WorldClim tile 1...
Fun Fact: Alligators have a bite force of up to 2,980 psi.
```

Coordinates are important data!

What happens if we randomly move coordinates by 0-10 km?





Comparative Analyses

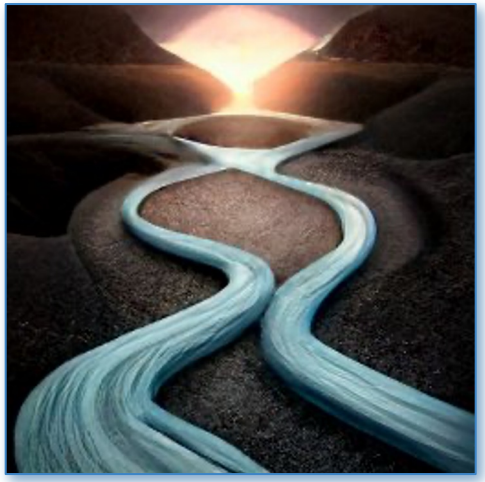
Population Structure, Gene Flow, and Genetic Diversity

- How do signatures of IBD and IBE vary across different taxonomic scales? And which environmental variables drive IBE in different taxa?
- Where do genetic breaks or discontinuities occur, and are these consistent across taxa?
- What are patterns of inbreeding, and are they related to landscape change or habitat fragmentation?
- Are there parts of the state that harbor higher levels of genetic diversity for broad sets of taxa? And what are the drivers of genetic diversity?
- Where do we project to lose genetic diversity under future climate scenarios?

Comparative Analyses

Selection, Adaptation, and Climate

- Are there sets of genes that show evidence of climate adaptation that is consistent across species and/or populations?
- Are there regions of higher genomic vulnerability?
- Are regions of species ranges with more environmental extremes (e.g. hotter, drier) experiencing stronger selection?
- What is the relationship between gene flow / connectivity and the genetic architecture of adaptation?
- What life history traits correlate with genetic diversity, population structure, and signals of adaptation?



Thank you!

Landscape Genomics Working Group:
Victoria Sork
Erin Toffelmier
Jay Sexton
Rachael Bay
Erik Enbody
Ryan Harrigan

With Additional Advice From:
Brad Shaffer
Scott Hodges
Peggy Fiedler

Feel free to reach out to us!
E. Anne Chambers (eachambers@berkeley.edu)
Anusha Bishop (anusha.bishop@berkeley.edu)
Ian Wang (ianwang@berkeley.edu)

PIs Who Shared Preliminary Data:
Beth Shapiro and the black bear team
Greg Grether and the damselfly team
Rachael Bay and the yellow warbler team

Questions for Discussion / Brainstorming

1. What additional questions or analyses would you like to see addressed with the entire CCGP dataset?
2. Is there any other functionality you would like to see in algatr?
3. Are there any considerations pertaining to your species (or related taxa) that we should consider in the landscape genomic analyses (important environmental variables, life history traits, etc.)?