Colias Redundancy Analysis

Lily Durkee | Final Project, BZ562

2024-05-05

Load the raw VCF file

I used the `adegenet` package to load raw VCF file in `genlight` format. To convert my VCF file to a `.raw` file, I used the `--recode A` function in `PLINK`.

```
library(adegenet)
```

```
## Loading required package: ade4
```

```
##
##    /// adegenet 2.1.10 is loaded ////////////
##
##    > overview: '?adegenet'
##    > tutorials/doc/questions: 'adegenetWeb()'
##    > bug reports/feature requests: adegenetIssues()
```

```
# Set the path to your PLINK .raw file (without the file extension)
raw_vcf <- "Data/colias.4x.merged_gatk.rm.relate.SNP.filtered_gatkVQSR2.PASS.8miss.recode.vcf.gl_impute

# Read the .raw file into R
colias_vcf <- read.PLINK(raw_vcf)
```

```
##
##  Reading PLINK raw format into a genlight object...
##
##
##  Reading loci information...
##
##  Reading and converting genotypes...
## .
##  Building final object...
##
## ...done.
```

```
# view the number of NAs
sum(is.na(colias_vcf))
```

```
## Warning in is.na(colias_vcf): is.na() applied to non-(list or vector) of type
## 'S4'
```

```
## [1] 0
```

```
# isolate sample names
samples_all <- data.frame(ID=pop(colias_vcf))

# number of SNPs
snps_n <- nLoc(colias_vcf)
```

Subsample the VCF file

Since my VCF file has 1.4 million SNPs, I subsampled down to 100k SNPs for this analysis.

```
# sample 100k
set.seed(123)  # for reproducibility
snps_select <- sample(snps_n, 100000)
colias_vcf.sub <- colias_vcf[,snps_select]
```

Load bioclimatic variables

These variables were extracted from raster files from WorldClim 2 (data from 1970-2000) using the lat/long coordinates of where each individual was collected during the collection season, June-August. I extracted values for each variable for each month, and then averaged over those three values. I used four bioclim variables:

- Maximum temperature (`tmax`)
- Precipitation (`prec`)
- Wind speed (`wind`)
- Solar radiation (`srad`)

Citation for dataset:

Fick, S.E. and R.J. Hijmans, 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37 (12): 4302-4315.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# this dataset contains the extracted bioclim variables
colias_clim1 <- read.csv("colias_bioclim.csv") %>% dplyr::select(-X)
```

Load elevation using package `elevatr`

I then extracted an elevation value for each butterfly using the package `elevatr`. This package uses data from the Amazon Web Services Terrain Tiles and the USGS Elevation Point Query Service.

Citation:

Hollister J, Shah T, Nowosad J, Robitaille A, Beck M, Johnson M (2023). elevatr: Access Elevation Data from Various APIs. doi:10.5281/zenodo.8335450, R package version 0.99.0, https://github.com/jhollist/elevatr/.

```
library(elevatr)
```

```
## elevatr v0.99.0 NOTE: Version 0.99.0 of 'elevatr' uses 'sf' and 'terra'.  Use
## of the 'sp', 'raster', and underlying 'rgdal' packages by 'elevatr' is being
## deprecated; however, get_elev_raster continues to return a RasterLayer.  This
## will be dropped in future versions, so please plan accordingly.
```

```
library(tidyverse)
library(sf)
```

```
## Linking to GEOS 3.10.2, GDAL 3.4.2, PROJ 8.2.1; sf_use_s2() is TRUE
```

```
latlong <- colias_clim1 %>% dplyr:: select(x, y)
crs <- st_crs("+proj=longlat +datum=WGS84")

# convert elevation.df data frame to an sf object
elevation.sf <- st_as_sf(latlong, coords = c("x", "y"), crs = crs)

elevation <- get_elev_point(elevation.sf)
```

```
## Downloading point elevations:

## Note: Elevation units are in meters
```

```
colias_clim.final <- cbind(colias_clim1, elevation)
```

Run the Redundancy Analysis (RDA)

To run the RDA, I used the tutorial made by Brenna Forster at https://popgen.nescent.org/2018-03-27_RDA_GEA.html. This analysis uses the `vegan` package in R.

```
library(vegan) # to run the RDA
```

```
## Loading required package: permute

## Loading required package: lattice

## This is vegan 2.6-4
```

```
library(psych) # for visualizing correlations
```

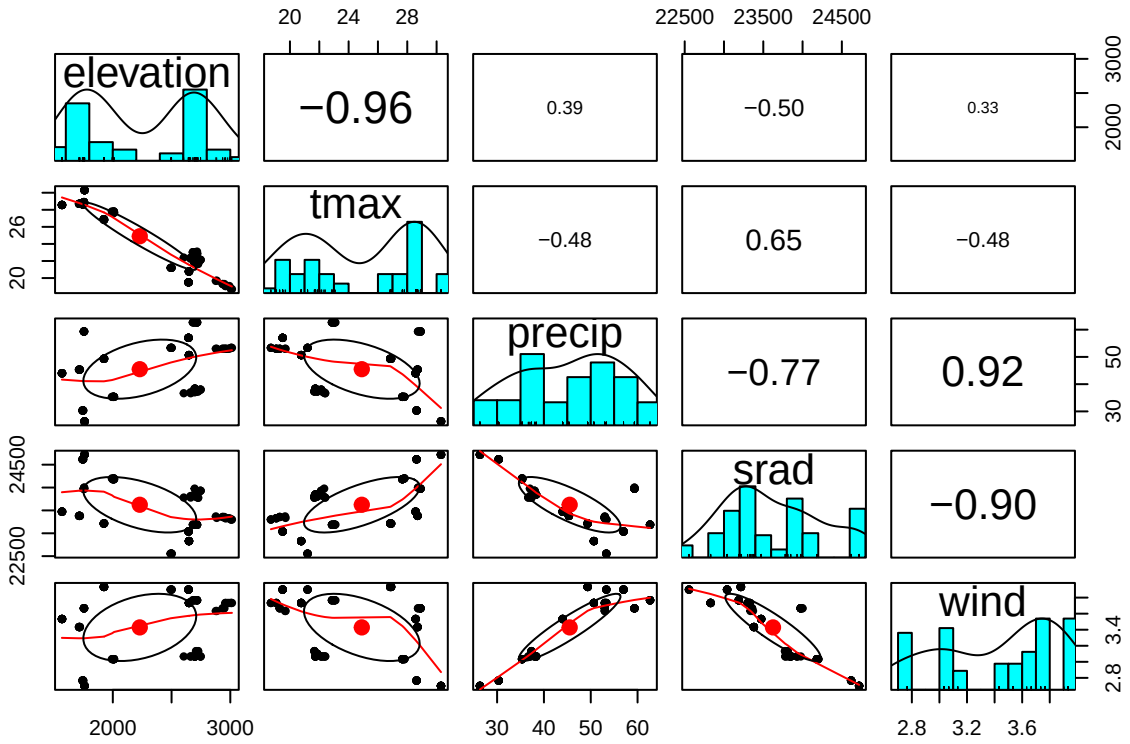
```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

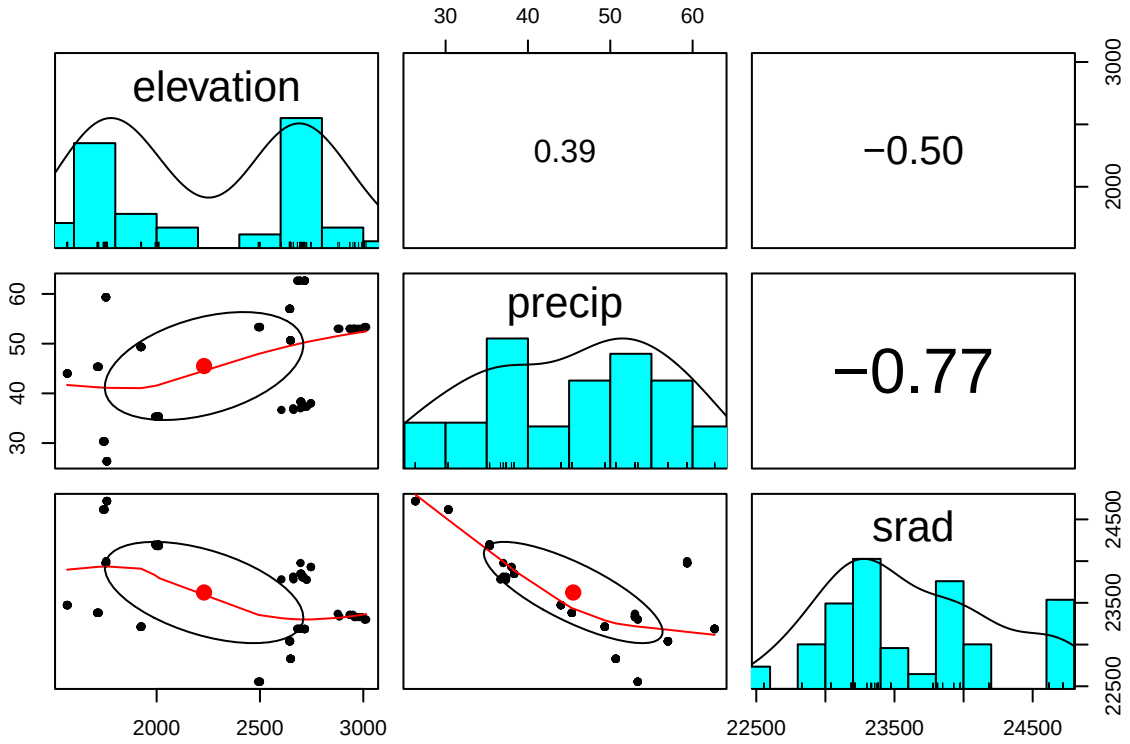
```
## confirms environmental data are in the same order as the genlight file
identical(as.character(samples_all[,1]), colias_clim.final[,1])
```

```
## [1] TRUE
```

```
#colias_clim$ID <- as.character(colias_clim$ID)

# create ENV file that has the environmental variables to include
colias_env.all <- colias_clim.final %>% dplyr::select(c(elevation, tmax, precip, srad, wind))
pairs.panels(colias_env.all, scale=T)
```




```
colias_env <- colias_clim.final %>% dplyr::select(c(elevation, precip, srad))
pairs.panels(colias_env, scale=T)
```



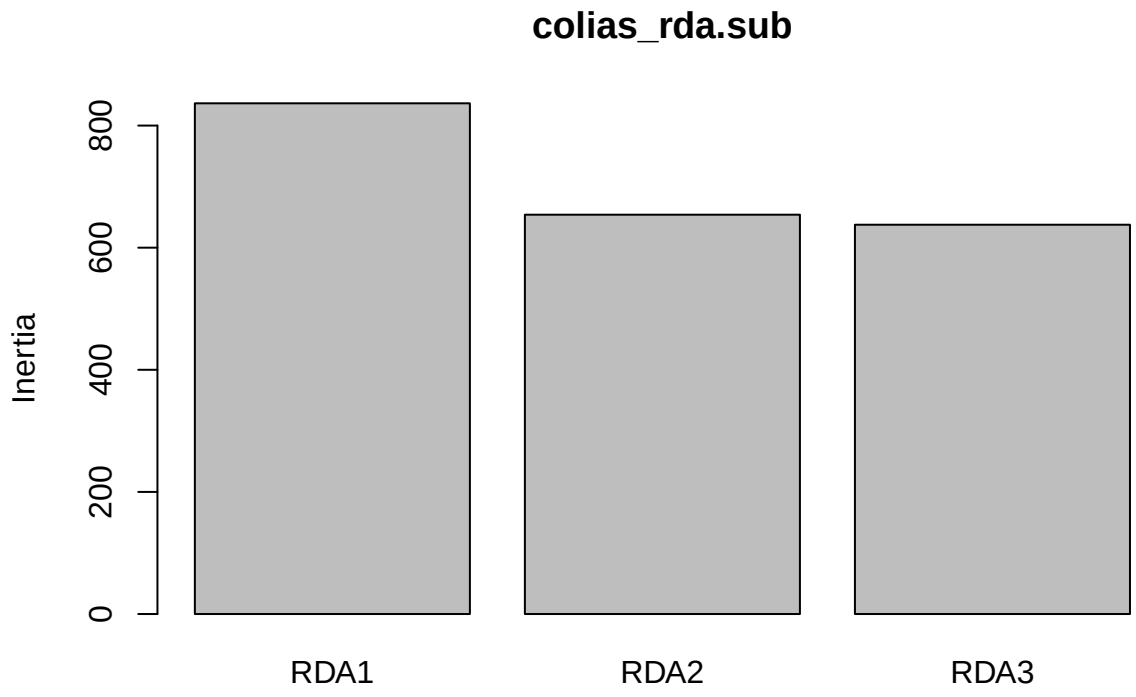

```
#png("colias_bioclim.corr_final.png")

#### run the RDA on a subset of 100k SNPs ####
colias_rda.sub <- rda(colias_vcf.sub ~ ., data=colias_env, scale=T)
colias_rda.sub
```

```
## Call: rda(formula = colias_vcf.sub ~ elevation + precip + srad, data =
## colias_env, scale = T)
##
##                Inertia Proportion Rank
## Total         9.996e+04  1.000e+00
## Constrained   2.128e+03  2.129e-02    3
## Unconstrained 9.783e+04  9.787e-01  152
## Inertia is correlations
##
## Eigenvalues for constrained axes:
##  RDA1  RDA2  RDA3
## 836.4 654.0 637.6
##
## Eigenvalues for unconstrained axes:
##   PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
## 836.7 766.1 733.3 719.8 712.0 711.1 707.7 704.9
## (Showing 8 of 152 unconstrained eigenvalues)
```

```r
# R squared
RsquareAdj(colias_rda.sub)
```

```
## $r.squared
## [1] 0.02129039
##
## $adj.r.squared
## [1] 0.00197375
```

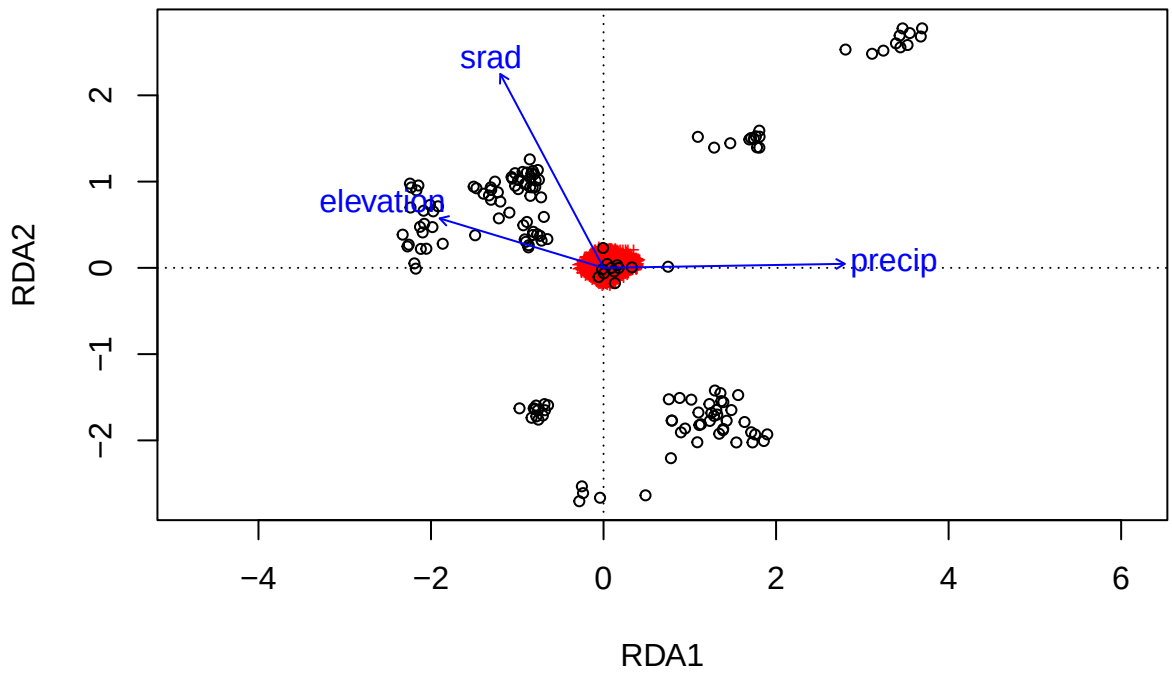
```r
# visualize the axes
screeplot(colias_rda.sub)
```




```r
# variance inflation factor - tells if variables are correlated
# confirm values are < 10
vif.cca(colias_rda.sub)
```

```
## elevation    precip      srad
##  1.331192  2.480259  2.785784
```

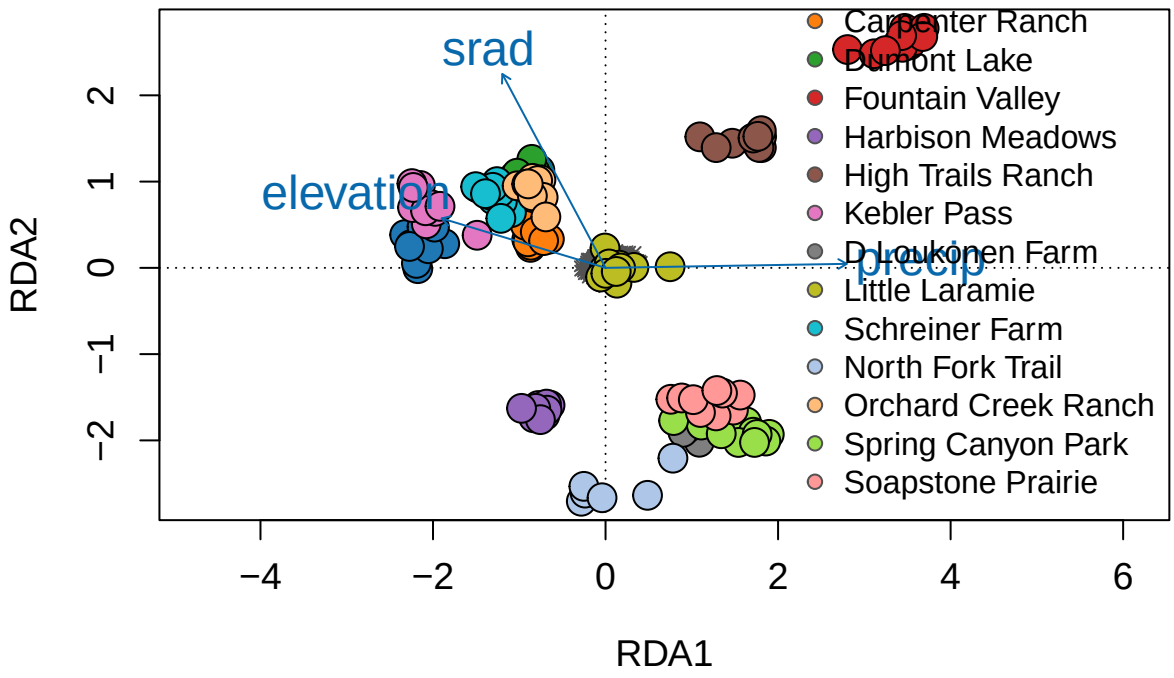
```r
#### RDA plots ####
# simple plot, axes 1 & 2
plot(colias_rda.sub, scaling=3)
```



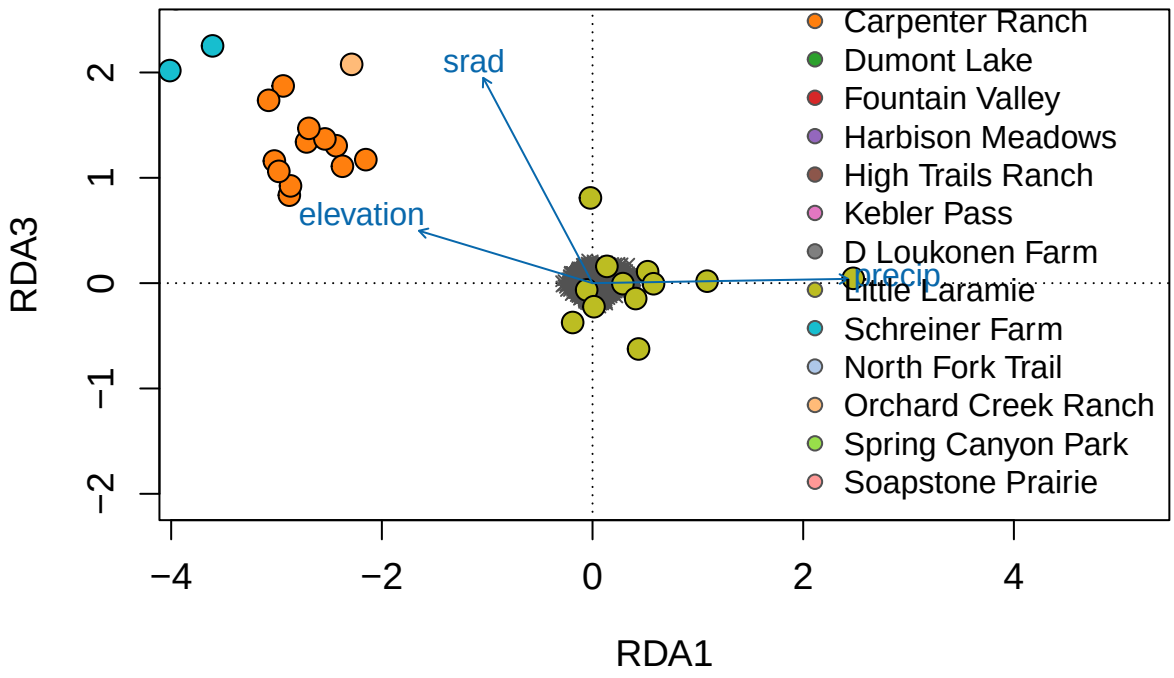

```
# more informative plots - color coded by site
colias_env$site <- colias_clim1$site
unique(colias_clim1$site)
```

```
##  [1] "Babbit Gulch"        "Carpenter Ranch"     "Dumont Lake"
##  [4] "Fountain Valley"     "Harbison Meadows"    "High Trails Ranch"
##  [7] "Kebler Pass"         "D Loukonen Farm"     "Little Laramie"
## [10] "Schreiner Farm"      "North Fork Trail"    "Orchard Creek Ranch"
## [13] "Spring Canyon Park"  "Soapstone Prairie"
```

```
levels(colias_env$site) <- as.factor(c("Babbit Gulch","Carpenter Ranch","Dumont Lake","Fountain Valley"
eco <- colias_env$site
bg <- c("#1f77b4", "#ff7f0e", "#2ca02c", "#d62728", "#9467bd",
                    "#8c564b", "#e377c2", "#7f7f7f", "#bcbd22", "#17becf",
                    "#aec7e8", "#ffbb78", "#98df4a", "#ff9896")

site_colors <- bg[match(eco, levels(eco))]

## plot for axes 1 & 2
plot(colias_rda.sub, type="n", scaling=3, cex.axis=1.2, cex.lab=1.2)
points(colias_rda.sub, display="species", pch=4, cex=.7, col="gray32", scaling=3)  # the SNPs
points(colias_rda.sub, display="sites", pch=21, cex=2, scaling=3, bg=site_colors)  # the butterflies
text(colias_rda.sub, scaling=3, display="bp", col="#0868ac", cex=1.5)  # the predictors
legend("bottomright", legend=levels(eco), bty="n", col="gray32", pch=21, cex=1, pt.bg=bg)
```



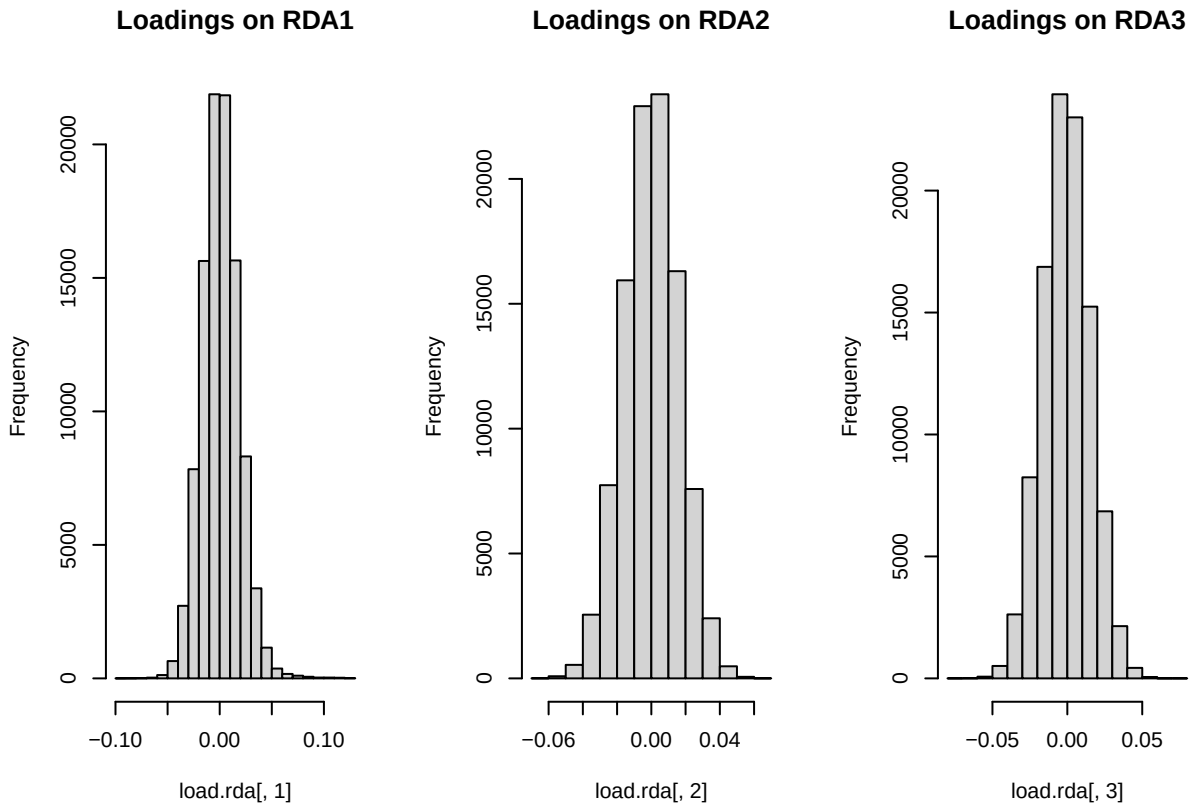

```
## plot for axes 1 & 3
# png("Colias_RDA_attempt1-axis1&3.png", width = 12, height = 12, units = "in", res = 350)

plot(colias_rda.sub, type="n", scaling=3, cex.axis=1.2, cex.lab=1.2, choices=c(1,3))
points(colias_rda.sub, display="species", pch=4, cex=.7, col="gray32", scaling=3)  # the SNPs
points(colias_rda.sub, display="sites", pch=21, cex=1.5, scaling=2, bg=site_colors)  # the butterflies
text(colias_rda.sub, scaling=3, display="bp", col="#0868ac", cex=1)  # the predictors
legend("bottomright", legend=levels(eco), bty="n", col="gray32", pch=21, cex=1, pt.bg=bg)
```



Identifying candidate SNPs

This part of RDA looks for candidate loci correlated with the environmental variables, thus identifying SNPs potentially involved in local adaptation.

```
load.rda <- scores(colias_rda.sub, choices=c(1:3), display="species")

# relatively normal distributions
layout(matrix(1:3, nrow = 1))
hist(load.rda[,1], main="Loadings on RDA1")
hist(load.rda[,2], main="Loadings on RDA2")
hist(load.rda[,3], main="Loadings on RDA3")
```




```
# we are interested in SNPs at the tails of the distributions
# this function will be able to detect outlier SNPs
outliers <- function(x,z){
  lims <- mean(x) + c(-1, 1) * z * sd(x)     # find loadings +/-z sd from mean loading
  x[x < lims[1] | x > lims[2]]               # locus names in these tails
}

# candidate loci
cand1 <- outliers(load.rda[,1],3) # 651
cand2 <- outliers(load.rda[,2],3) # 214
cand3 <- outliers(load.rda[,3],3) # 214

# number of candidate loci identified
ncand <- length(cand1) + length(cand2) + length(cand3)
ncand
```

```
## [1] 1098
```

```
#1,079 with 100k SNPs

# now, organize the results in one data frame with axis, SNP name, & correlation with each predictor
cand1 <- cbind.data.frame(rep(1,times=length(cand1)), names(cand1), unname(cand1))
cand2 <- cbind.data.frame(rep(2,times=length(cand2)), names(cand2), unname(cand2))
cand3 <- cbind.data.frame(rep(3,times=length(cand3)), names(cand3), unname(cand3))
```

```
colnames(cand1) <- colnames(cand2) <- colnames(cand3) <- c("axis","snp","loading")

cand <- rbind(cand1, cand2, cand3)
cand$snp <- as.character(cand$snp)

head(cand)
```

```
##   axis snp    loading
## 1    1 ._T 0.11404715
## 2    1 ._G 0.07336634
## 3    1 ._G 0.05701485
## 4    1 ._T 0.06218630
## 5    1 ._C 0.06746767
## 6    1 ._G 0.09409960
```

```
# I will now put them in a single data frame with the 3 env predictors
foo <- matrix(NA, nrow=(ncand), ncol=3)  # 3 columns for 3 predictors
colnames(foo) <- c("elevation", "precip", "srad")

for (i in 1:length(cand$snp)) {
  nam <- cand[i,2]
  snp.gen <- as.matrix(colias_vcf.sub)[,nam]
  foo[i,] <- apply(dplyr::select(colias_env, c(elevation, precip, srad)), 2,
                   function(x) cor(x,snp.gen))
}

cand <- cbind.data.frame(cand,foo)
head(cand)
```

```
##   axis snp    loading    elevation      precip        srad
## 1    1 ._T 0.11404715  0.008390232  0.06897030 -0.05431402
## 2    1 ._G 0.07336634 -0.052353506  0.02240204 -0.05306279
## 3    1 ._G 0.05701485 -0.052353506  0.02240204 -0.05306279
## 4    1 ._T 0.06218630  0.008390232  0.06897030 -0.05431402
## 5    1 ._C 0.06746767 -0.081883070 -0.10685143  0.05609544
## 6    1 ._G 0.09409960 -0.052353506  0.02240204 -0.05306279
```

```
# 1075 total duplicates
length(cand$snp[duplicated(cand$snp)])
```

```
## [1] 1094
```

```
# axis 1
foo <- cbind(cand$axis, duplicated(cand$snp))
table(foo[foo[,1]==1,2])
```

```
##
##   0   1
##   4 652
```

```
# axis 2
table(foo[foo[,1]==2,2])
```

```
##
##   1
## 219
```

```
# axis 3
table(foo[foo[,1]==3,2])
```

```
##
##   1
## 223
```

```
cand_no.dup <- cand[!duplicated(cand$snp),] # remove duplicate detections
# only 4 left

# see which environmental vars they are most correlated with

for (i in 1:length(cand_no.dup$snp)) {
  bar <- cand_no.dup[i,]
  cand_no.dup[i,7] <- names(which.max(abs(bar[3:6]))) # gives the variable
  cand_no.dup[i,8] <- max(abs(bar[3:6]))              # gives the correlation
}

colnames(cand_no.dup)[7] <- "predictor"
colnames(cand_no.dup)[8] <- "correlation"

table(cand_no.dup$predictor)
```

```
##
## loading  precip
##       3       1
```

Conclusions

The RDA plots suggest population structure that is driven by environmental factors. When identifying candidate loci using only 100k SNPs, I was only able to find four that were correlated with environmental factors. I expect to find more loci when using the full VCF file (1.4 million SNPs). RDA is a tool that investigates the genetic basis of local adaptation, and I anticipate these findings will be highly useful to my study, which focuses on local adaptation of a butterfly species (*Colias philodice eriphyle*) to elevation in the Colorado Rocky Mountains.