Dahn-Young Dong Final Project

Dahn-Young Dong

9/27/2020

Population Genetic Structure–Ancestry Proportions of 16 individual wood frogs each sampled from a distinct pond.

Introduction and Goals

Yale Myer's Forests (YMF) in northeastern Connecticut has distributions of populations of wood frogs (*Rana sylvatica*), and Skelly's lab has been recording long-term population dynamics, phenotypic and abiotic environment data that show rapid evolution is occurring (Freidenburg and Skelly, 2004; Skelly, 2004; Ligon and Skelly, 2009). That is, the populations are diverging in their physiological and morphological traits based on the local environments over generations. What requires to substantiate this ecological observation further is to explore the genetics. Previous effort to study microsatellite markers have not been successful in showing the population divergence or clustering at such microgeographical scale.

With the availability of genome-wide SNPs for a few dozens of populations of *Rana sylvatica* from the year of 2018, and the understanding of presumably genetic-based and non-plastic local adaptations for populations that are physically in proximity to each other, we can now start to explore the neutral and adaptive genetic variations over the landscape. Given the pilot data of 16 individuals/genotypes, I would like to know about its standing microgeographic genetic variation.

Two questions I want to find answers for:

1. how are these 16 unique individuals from 16 populations related to each other by their ancestry (shared allelic frequencies across the genome by inheritance)?
2. qualitatively, how much correspondence is there between the individuals genetic ancestry to environmental variables, such as elevation and river system (visual interpolations and comparisons)? I hypothesize that the river and elevation would partition the populations into two, due to isolation by resistance (Richardson, Urban 2012).

My collaborator A. Z. Andis Arietta did the collection, ipyrad processing of the genomic data. For the pilot dataset, in addition, he filtered the SNPs data so that there is no missing data across loci or individual samples.https://github.com/laninsky/GBS_SNP_filter

The pilot dataset is a VCF file with post-processed genomic data. I used this file for the project. The dataset contains 16 wood frog individuals' SNP biallelic sequences, each from one distinct pond population at the focal site, YMF.

Methods

The whole run should be less than 1 minutes (some chunks are quoted out).

I closely follow the scripts from https://github.com/jdalapicolla/LanGen_pipeline_version2, particularly the ones related to manipulating and visualizing genomic data given spatial coordinates— 1.1 filtering SNPS and 1.2 genetic structure. They included 3-4 different parallel analyses and I chose TESS3 as my main analyses, which is also what the author preferred.

I used VCFtools(Danecek et al. 2011), a linux tool, interactively linked through "r2vcftools" package in R. This is for manipulating genomic data in VCF format. I also used "TESS3R" to analyze population structure, produce ancestry proportions and spatial mapping. Lastly, I used package "sp" to create an elevation and raster map of the location, though the detailed wrangling is not shown in this project.

TESS3 uses model-free, individual based, geographically constrained least-squares algorithm for calculating ancestry coefficient, different from Bayesian algorithm that uses MCMC posteriors, which increases computational speed in larger genotype matrices without compromising accuracy (Caye et al. 2016).

```
source("functions_LanGen.R") #some functions provided by the tutorial author, only used a few
library(remotes) # R Package installation from Remote Repositories, including "GitHub"
library(BiocManager) #Access the Bioconductor project package Repository
library(pacman) # package management tool
library(devtools) # Collection of package development tools.
library(r2vcftools) # an interface that brings vcftools analyses into R
library(LEA) # Landscape and ecological association studies
library(tess3r) # inference for spatial population structure
library(raster) #geographic data analysis and modeling
library(sp) #classes and methods for spatial data

#set folders for results
create_dir(c("./adapt_var_mapping/Filtering"))
```

```
## [1] "./adapt_var_mapping/Filtering has already been created. Be careful you can overwrite files"
```

1. Load Files and Verify Data Quality

```
vcf_file <- "vcf/testdata_nomissing.vcf" #biallelic SNP dataset with no missing data- from Andis Ariett
project_name <- "rasy_ymfProj"
coord_file <- "coords/testdata_geoinfo.csv" ## this is using smaller dataset
coords <- read.csv(coord_file)
```

1.1 Set up

```
#Creates an object of class vcfLink, which acts as an interface to a temporary VCF file. Also loads met
snps_unind <-  r2vcftools::vcfLink(vcf_file, overwriteID=T)
VCFsummary(snps_unind)
```

1.2 Load vcf file

```
## 16 individuals and 18077 SNPs.
```

```
#Converts the VCF file to genotype matrix format and returns this matrix
genotypes <- r2vcftools::GenotypeMatrix(snps_unind)
```

1.3 Verify the quality of data

```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee2d158151
##  --012
##  --out /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee5246e7b2
##
## After filtering, kept 16 out of 16 Individuals
## Writing 012 matrix files ... Done.
## After filtering, kept 18077 out of a possible 18077 Sites
## Run Time = 0.00 seconds
```

```
genotypes[1:16, 1:10] ## -1 is missing; otherwise, gives the number of derived alleles in the genotype
```

```
##                         snp1 snp2 snp3 snp4 snp5 snp6 snp7 snp8 snp9 snp10
## YMF2018_009_S25001.trim    1    0    0    1    0    0    1    0    0     0
## YMF2018_011_S27001.trim    0    0    0    0    0    0    1    1    0     0
## YMF2018_012_S29001.trim    0    1    0    0    1    0    0    0    1     0
## YMF2018_013_S31001.trim    0    0    0    0    0    0    0    0    1     0
## YMF2018_017_S33001.trim    0    0    0    1    0    0    1    0    0     0
## YMF2018_018_S35001.trim    0    0    0    0    0    0    0    0    1     0
## YMF2018_019_S37001.trim    0    0    1    0    0    0    0    0    1     0
## YMF2018_022_S39001.trim    0    0    0    0    0    0    1    0    1     0
## YMF2018_025_S26001.trim    0    0    0    0    0    1    1    0    0     0
## YMF2018_028_S28001.trim    0    0    0    0    0    0    2    0    0     0
## YMF2018_030_S30001.trim    0    0    0    1    0    0    0    0    1     0
## YMF2018_032_S32001.trim    0    0    0    0    0    0    0    0    0     0
## YMF2018_033_S34001.trim    0    0    0    1    0    0    0    0    1     0
## YMF2018_036_S36001.trim    0    0    0    0    0    0    1    0    0     1
## YMF2018_037_S38001.trim    1    0    0    0    0    1    1    0    2     0
## YMF2018_039_S40001.trim    0    0    0    0    0    0    0    0    1     0
```

```
#specify which analyses or conversions to perform on the data that passed through all specified filters

#Look at depth, quality, HWE, HE, allele frequencies, and Pi

site.depth <- r2vcftools::Query(snps_unind, type="site-mean-depth")
```

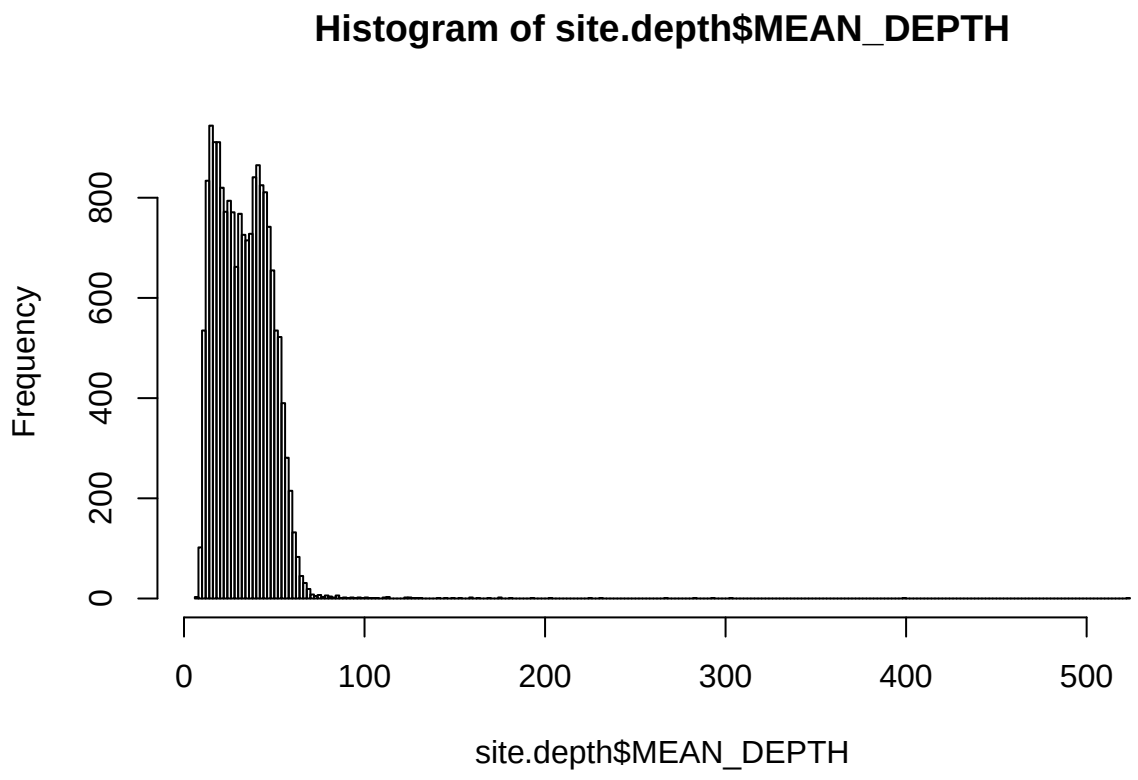
```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
## Parameters as interpreted:
```

```
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee69f679a0
##  --site-mean-depth
##  --stdout
##
## After filtering, kept 16 out of 16 Individuals
## Outputting Depth for Each Site
## After filtering, kept 18077 out of a possible 18077 Sites
## Run Time = 1.00 seconds
```

```
summary(site.depth$MEAN_DEPTH) #Mean = 33 / Median = 33
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.688  20.688  32.562  33.324  44.062 523.812
```

```
hist(site.depth$MEAN_DEPTH, breaks=200)
```




```
#most SNPs have coverage of 33. This is a good number for most SNPs.

#average number of nucleotide differences per site, denoted by pi
PI <- Query(snps_unind, type="site-pi")
```

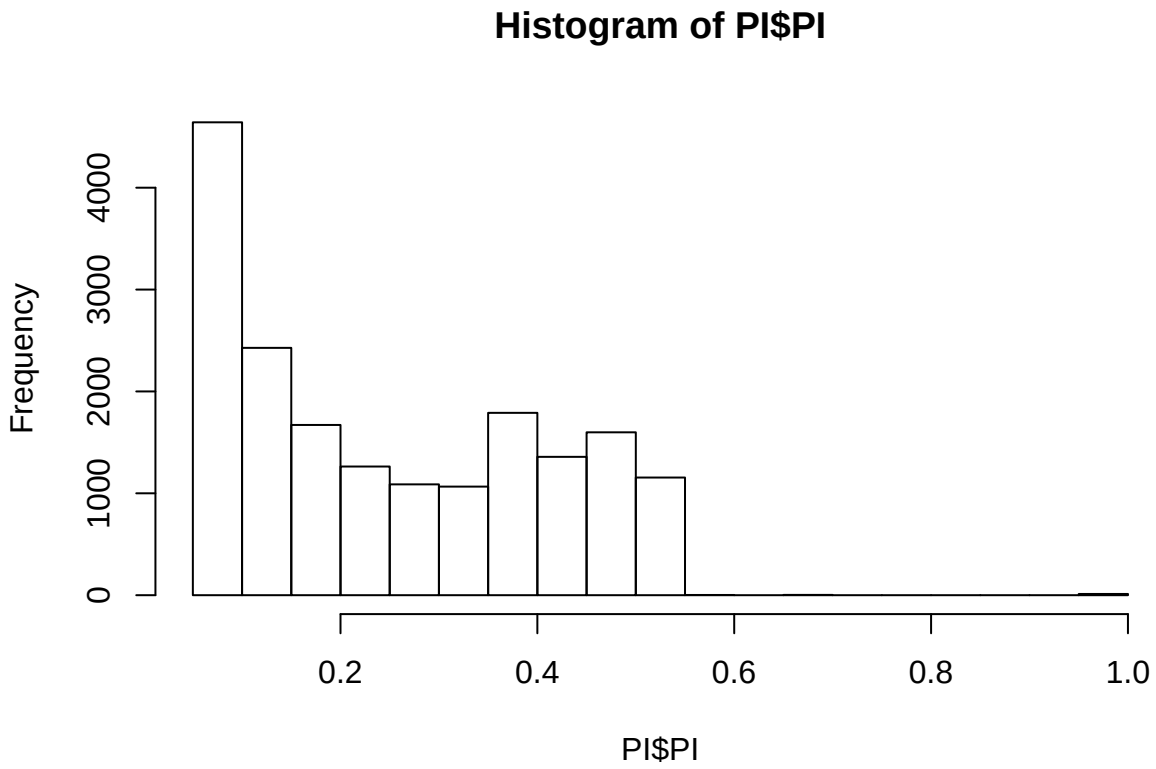
```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
```

```
##
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee69f679a0
##  --site-pi
##  --stdout
##
## After filtering, kept 16 out of 16 Individuals
## Outputting Per-Site Nucleotide Diversity Statistics...
## After filtering, kept 18077 out of a possible 18077 Sites
## Run Time = 0.00 seconds
```

```
mean(PI$PI) ## Mean nucleotide divergency per-site 0.244. At each SNP site, on average, 20% of the geno
```

```
## [1] 0.2447176
```

```
hist(PI$PI)
```




```
#Reports a p-value for each site from a Hardy-Weinberg Equilibrium test
HWE <- Query(snps_unind, type="hardy")
```

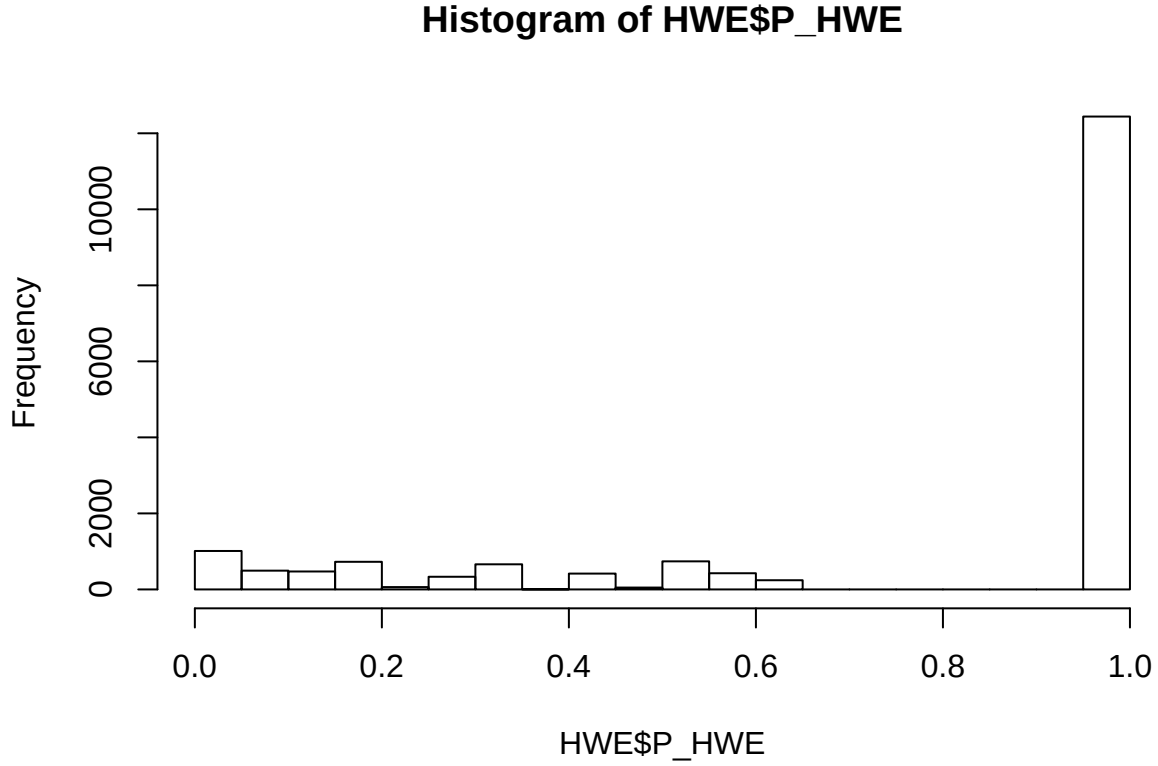
```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
```

```
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee69f679a0
##  --hardy
##  --stdout
##
## After filtering, kept 16 out of 16 Individuals
## Outputting HWE statistics (but only for biallelic loci)
## After filtering, kept 18077 out of a possible 18077 Sites
## Run Time = 0.00 seconds
```

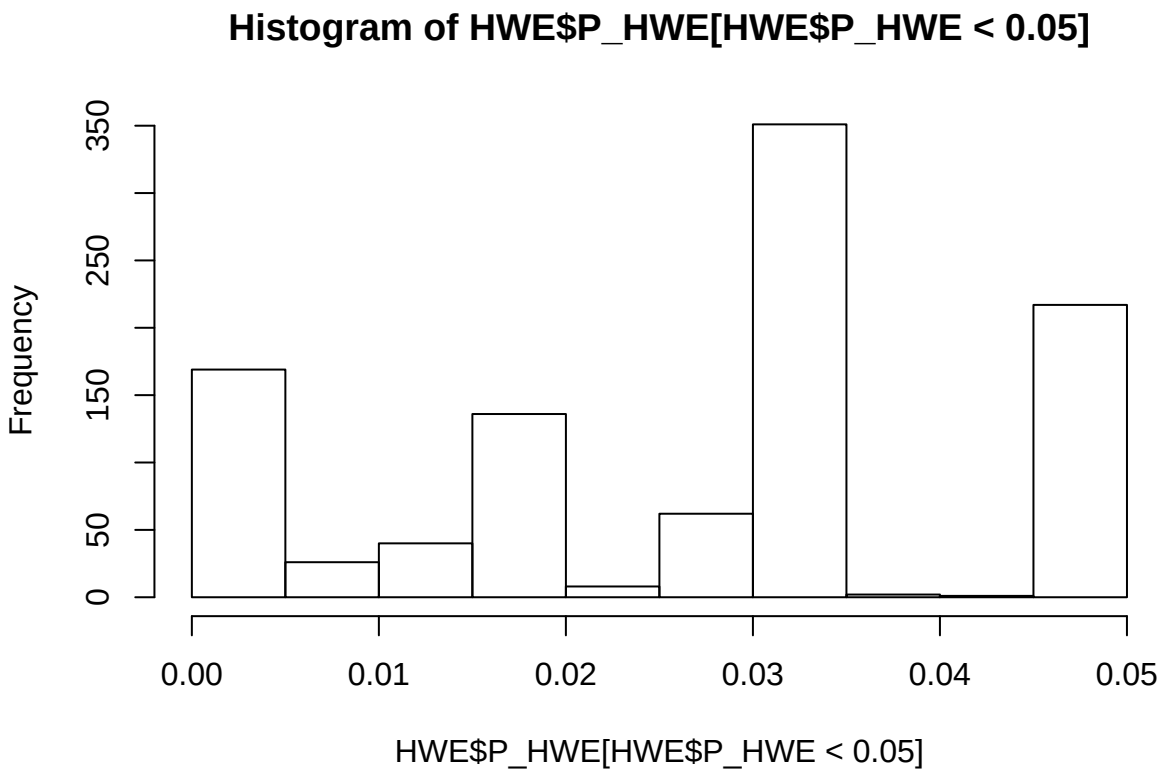
```
summary(HWE$P_HWE) #Mean = 0.7775 / Median = 1
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000214 0.5301446 1.0000000 0.7754875 1.0000000 1.0000000
```

```
hist(HWE$P_HWE) #almost all fail to reject the null hypothesis, meaning most SNPs are at HWE, that is n
```




```
hist(HWE$P_HWE[HWE$P_HWE<0.05]) #around 800 snps our of 18000 are significantly out of HWE, violating s
```




```
#Add coordinates to vcf file.

#add the information.
snps_unind@meta$local_ID <- coords[,4] #locality pond ID
snps_unind@meta$longitude <- coords[,3] #longitude
snps_unind@meta$latitude <- coords[,2] #latitude
snps_unind@meta$ind_ID <- coords[,1] #sample ID
#verify the file
head(snps_unind@meta) #verify
```

1.4 Add geographic info by individuals to a vcf file

```
##                         sample_num             sample_name local_ID longitude
## YMF2018_009_S25001.trim          1 YMF2018_009_S25001.trim       BS -72.12360
## YMF2018_011_S27001.trim          2 YMF2018_011_S27001.trim       MI -72.14694
## YMF2018_012_S29001.trim          3 YMF2018_012_S29001.trim       WF -72.16048
## YMF2018_013_S31001.trim          4 YMF2018_013_S31001.trim       PB -72.12569
## YMF2018_017_S33001.trim          5 YMF2018_017_S33001.trim       BO -72.14963
## YMF2018_018_S35001.trim          6 YMF2018_018_S35001.trim       GB -72.14886
##                         latitude                 ind_ID
## YMF2018_009_S25001.trim 41.95493 trimYMF2018_009_S25001
## YMF2018_011_S27001.trim 41.93590 trimYMF2018_011_S27001
## YMF2018_012_S29001.trim 41.91979 trimYMF2018_012_S29001
```

```
## YMF2018_013_S31001.trim 41.96234 trimYMF2018_013_S31001
## YMF2018_017_S33001.trim 41.91794 trimYMF2018_017_S33001
## YMF2018_018_S35001.trim 41.91800 trimYMF2018_018_S35001
```

2. Filter for neutral loci

```
#From dataset with all individuals
snps_fil <- Filter(snps_unind, filterOptions(max.missing = 1, min.meanDP=20, max.meanDP=500, hwe=0.05))
```

2.1 Filter dataset by missingness, min and max coverage, and hardy-weinberg equilibrium (HWE):

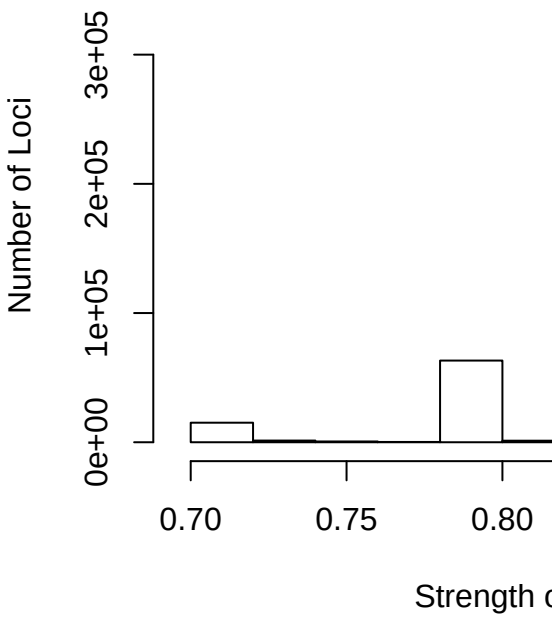
```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee69f679a0
##  --max-alleles 2
##  --max-meanDP 500
##  --hwe 0.05
##  --min-meanDP 20
##  --max-missing 1
##  --out /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee6577824
##  --recode
##
## After filtering, kept 16 out of 16 Individuals
## Outputting VCF file...
## After filtering, kept 12609 out of a possible 18077 Sites
## Run Time = 0.00 seconds
```

```
VCFsummary(snps_fil)  ## 16 individuals and 12609 SNPs left.
```

```
## 16 individuals and 12609 SNPs.
```

```
# review of linkage disequilibrium https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124487/#FD1


#Define R² value:
r2 = 0.7 # filtering out most strongly correlated/linked loci but still maximizing available SNPs left
#remove snps with R²
#ld_between <- Linkage(snps_fil, type="interchrom-geno-r2", linkageOptions(min.r2=r2))
ld_between <- read.csv(paste0("adapt_var_mapping/Filtering/ld_between_", r2, "_hwe_test2.csv")) #load t
hist(ld_between$R.2, main = "Distribution of Linked Loci", xlab = "Strength of Linkage Disequilibrium",
```



2.2 Filter dataset by linkage disequilibrium (LD) between contigs.

```
#write.csv(ld_between, file= paste0("adapt_var_mapping/Filtering/ld_between_", r2, "_hwe_test2.csv"))
```

```
#Select one set of the correlated snps (ID1 or ID2) # I went for ID2
ld2_snps <- ld_between$ID2
nold2_snps <- snps_fil@site_id[!(snps_fil@site_id %in% ld2_snps)]
snps_fil_ldF <- Subset(snps_fil, sites=nold2_snps) # Keep snps that are not in LD (excluding 6000+ SNPs
```

```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee6577824
##  --out /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee5191acae
##  --recode
##  --snps /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee405800a2
##
## After filtering, kept 16 out of 16 Individuals
## Outputting VCF file...
## After filtering, kept 3658 out of a possible 12609 Sites
## Run Time = 1.00 seconds
```

```
neutralLDBetween_SZ <- capture.output(VCFsummary(snps_fil_ldF))
neutralLDBetween_SZ # "16 individuals and 3658 SNPs."
```

```
## [1] "16 individuals and 3658 SNPs."
```

```
site.depth2 <- Query(snps_fil_ldF, "site-mean-depth")
```

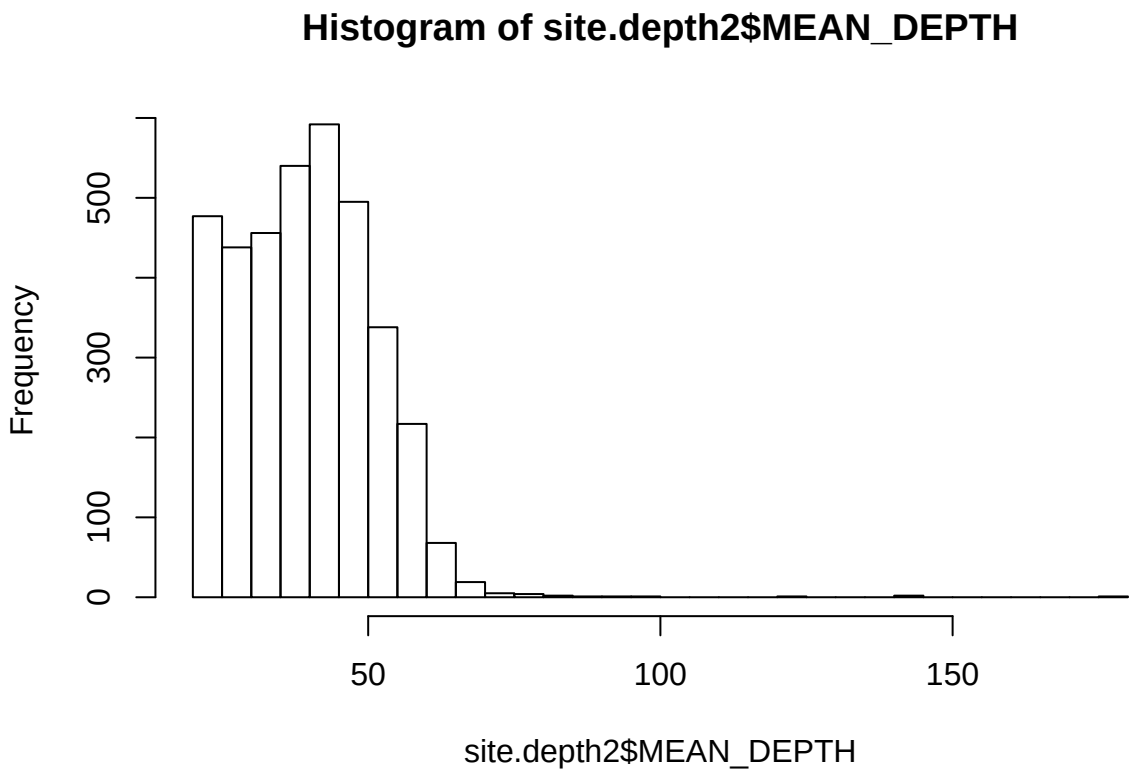
2.3 Verify the quality of the filtered dataset

```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee5191acae
##  --site-mean-depth
##  --stdout
##
## After filtering, kept 16 out of 16 Individuals
## Outputting Depth for Each Site
## After filtering, kept 3658 out of a possible 3658 Sites
## Run Time = 0.00 seconds
```

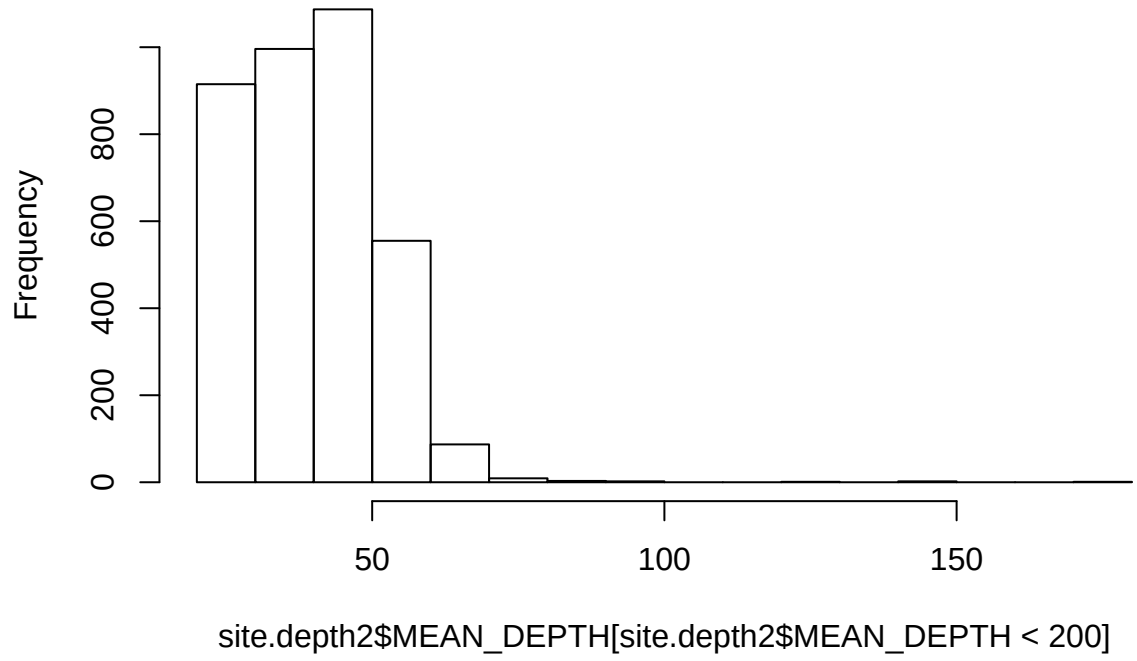
```
HWE2 <- Query(snps_fil_ldF, type="hardy")
```

```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee5191acae
##  --hardy
##  --stdout
##
## After filtering, kept 16 out of 16 Individuals
## Outputting HWE statistics (but only for biallelic loci)
## After filtering, kept 3658 out of a possible 3658 Sites
## Run Time = 0.00 seconds
```

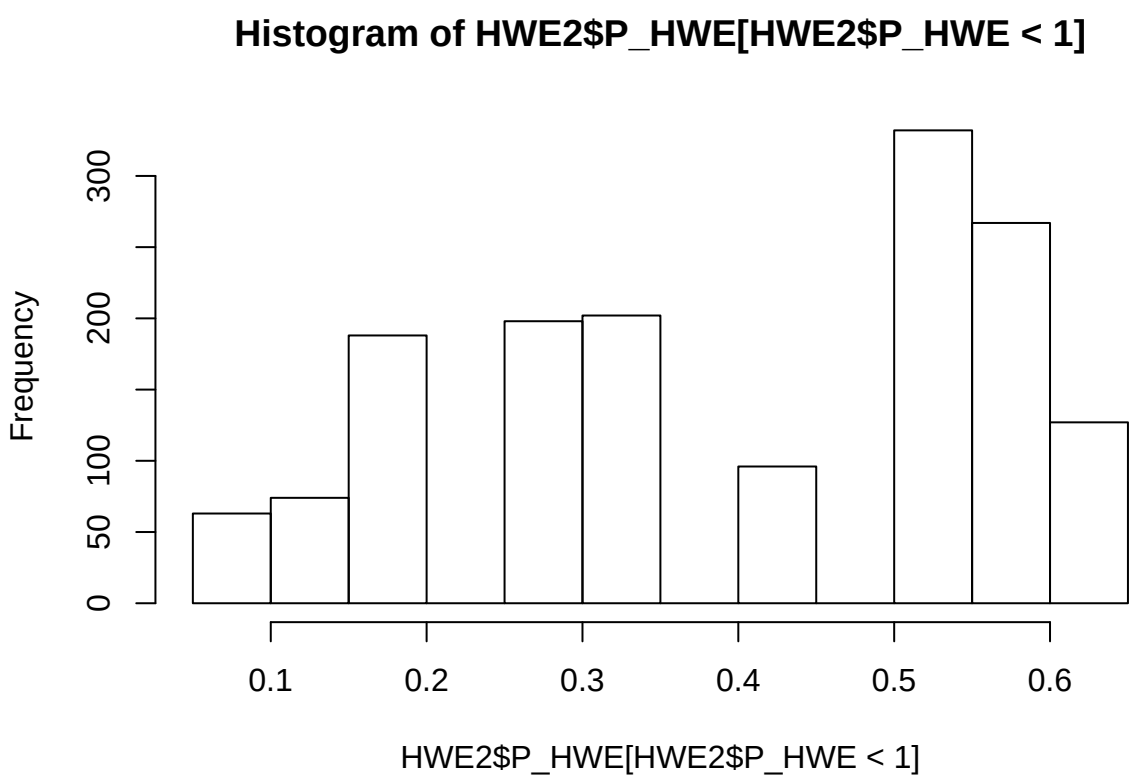
```
hist(site.depth2$MEAN_DEPTH, breaks=50)
```




```
hist(site.depth2$MEAN_DEPTH[site.depth2$MEAN_DEPTH <200])
```




```
hist(HWE2$P_HWE[HWE2$P_HWE< 1]) # all the SNPs with <0.05 p values are excluded already. good sign
```



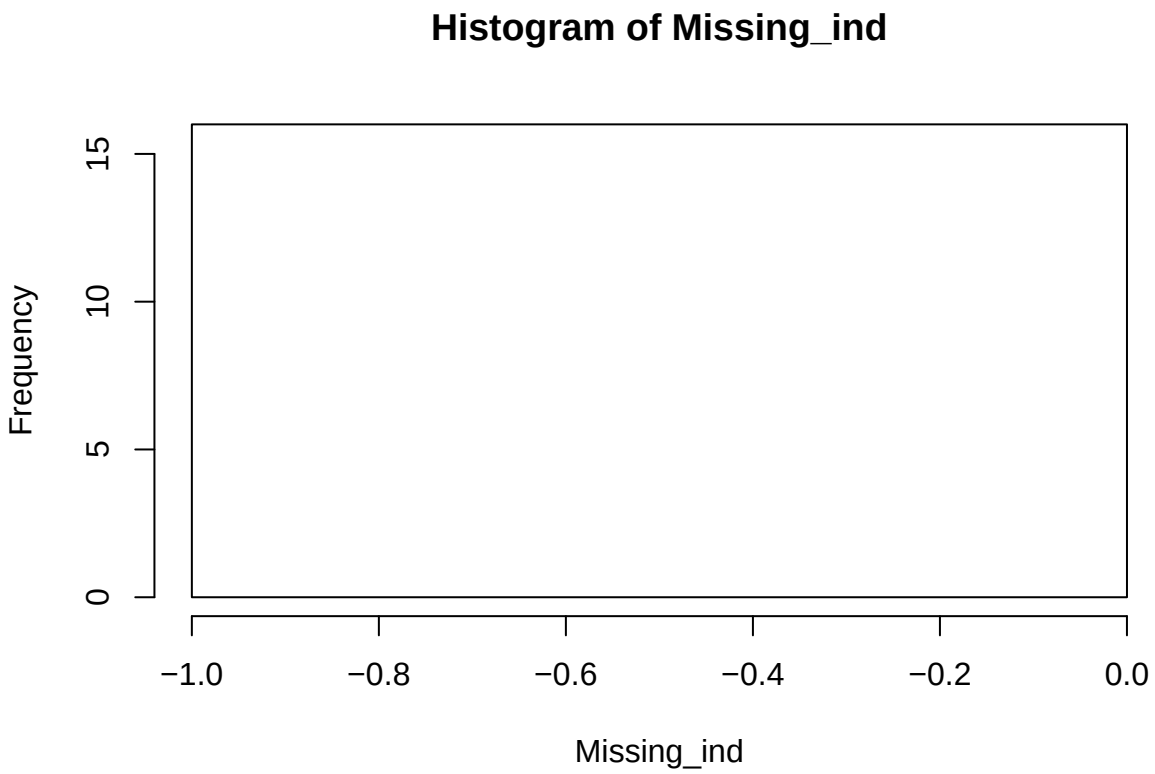

```
#Verify the real missing data per individual:
Missing_ind <- apply(GenotypeMatrix(snps_fil_ldF),1, function(x) sum(x<0)/length(x)*100)
```

```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee314201b2
##  --012
##  --out /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee1d399c43
##
## After filtering, kept 16 out of 16 Individuals
## Writing 012 matrix files ... Done.
## After filtering, kept 3658 out of a possible 3658 Sites
## Run Time = 0.00 seconds
```

```
summary(Missing_ind) ## no more missing individuals
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

```
hist(Missing_ind)
```



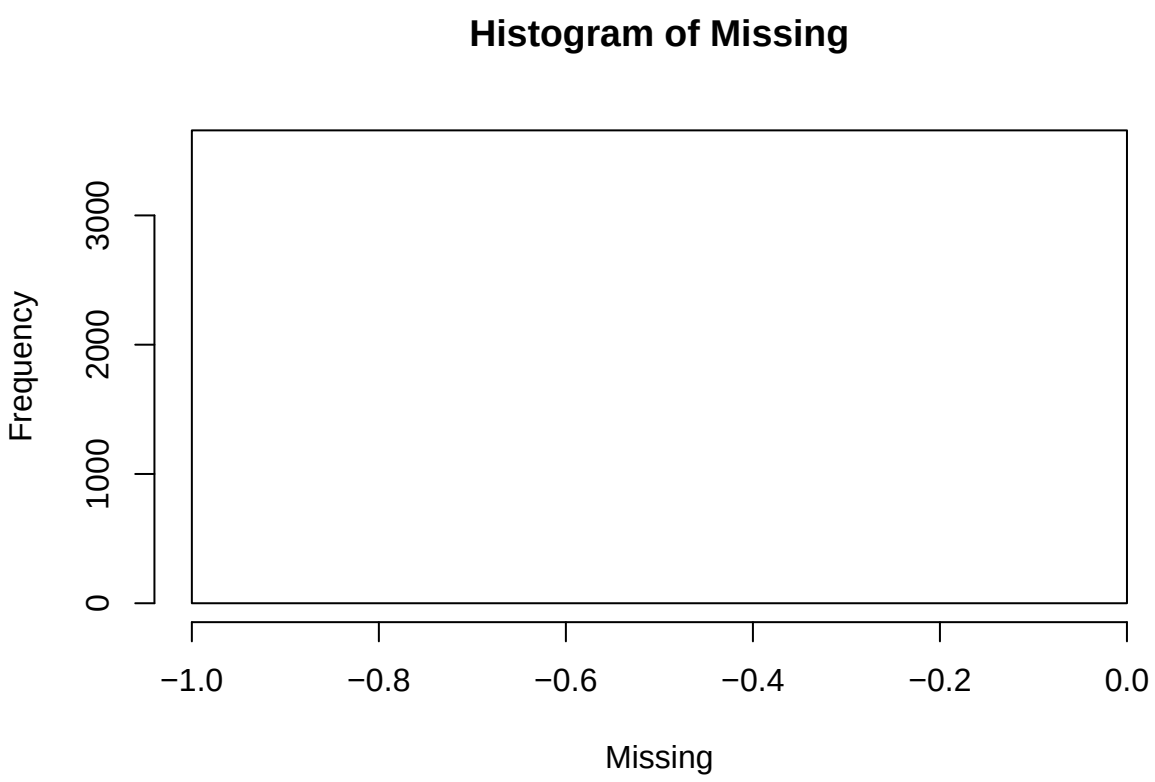

```
#Verify the real missing data per locus:
Missing <- apply(GenotypeMatrix(snps_fil_ldF), 2, function(x) sum(x < 0)/length(x)*100) ## Actual perce
```

```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee4a5a33fd
##  --012
##  --out /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee358d380a
##
## After filtering, kept 16 out of 16 Individuals
## Writing 012 matrix files ... Done.
## After filtering, kept 3658 out of a possible 3658 Sites
## Run Time = 0.00 seconds
```

```
summary(Missing) # no more missing locus- good sign for successful filtering above.
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

```
hist(Missing)
```




```
VCFsummary(snps_unind) #16 individuals and 18077 SNPs.
```

```
## 16 individuals and 18077 SNPs.
```

```
VCFsummary(snps_fil_ldF) ##16 individuals and 3658 SNPs. # would this change the resulting ancestry sig
```

```
## 16 individuals and 3658 SNPs.
```

2.4. Save the vcf with only neutral SNPs.

3. TESS3 inference of Spatial Population Genetic Structure (https://bcm-uga.github.io/TESS3_encho_sen/articles/main-vignette.html)

```
#Load the .VCF file with only neutral SNPs:
snps <- vcfLink(paste0("vcf/", project_name,"_filtered_neutral_partial.vcf"), overwriteID=T)
VCFsummary(snps) ##16 individuals and 3658 SNPs.
```

3.1. Input files/objects for TESS analyses

```
## 16 individuals and 3658 SNPs.
```

```
#Create a Genotype matrix
genotypes <- GenotypeMatrix(snps) # only returns biallelic
```

```
##
## VCFtools - 0.1.17
## (C) Adam Auton and Anthony Marcketta 2009
##
## Parameters as interpreted:
##  --vcf /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee3a62a867
##  --012
##  --out /var/folders/fg/x245ssxd4w76gcsndt4lhgxr0000gp/T//Rtmp6eYPFg/vcfLink_filef3ee5af3c657
##
## After filtering, kept 16 out of 16 Individuals
## Writing 012 matrix files ... Done.
## After filtering, kept 3658 out of a possible 3658 Sites
## Run Time = 0.00 seconds
```

```
genotypes[1:16, 3000:3010] ## -1 is missing # should have no missing because I filtered out everything
```

```
##                         snp3000 snp3001 snp3002 snp3003 snp3004 snp3005 snp3006
## YMF2018_009_S25001.trim       0       1       1       1       1       1       0
## YMF2018_011_S27001.trim       0       0       0       0       0       0       2
## YMF2018_012_S29001.trim       1       0       0       1       1       2       1
## YMF2018_013_S31001.trim       0       0       1       0       0       0       2
## YMF2018_017_S33001.trim       1       0       0       0       0       2       0
## YMF2018_018_S35001.trim       1       1       1       1       2       0       1
## YMF2018_019_S37001.trim       1       1       2       2       1       1       0
## YMF2018_022_S39001.trim       1       1       2       2       1       0       1
## YMF2018_025_S26001.trim       2       2       1       0       1       0       1
## YMF2018_028_S28001.trim       0       1       0       2       0       1       1
## YMF2018_030_S30001.trim       0       2       1       2       1       0       0
## YMF2018_032_S32001.trim       1       1       1       0       0       1       1
## YMF2018_033_S34001.trim       0       0       1       1       0       2       1
## YMF2018_036_S36001.trim       0       0       0       0       0       0       1
## YMF2018_037_S38001.trim       0       1       0       0       1       0       2
## YMF2018_039_S40001.trim       2       2       0       0       2       1       0
##                         snp3007 snp3008 snp3009 snp3010
## YMF2018_009_S25001.trim       1       1       0       2
## YMF2018_011_S27001.trim       1       1       0       0
## YMF2018_012_S29001.trim       1       1       0       1
## YMF2018_013_S31001.trim       1       1       0       1
## YMF2018_017_S33001.trim       1       0       1       1
## YMF2018_018_S35001.trim       0       0       1       1
## YMF2018_019_S37001.trim       0       2       0       1
## YMF2018_022_S39001.trim       1       1       2       1
## YMF2018_025_S26001.trim       0       0       0       0
## YMF2018_028_S28001.trim       1       0       0       0
## YMF2018_030_S30001.trim       1       0       1       0
## YMF2018_032_S32001.trim       2       0       0       0
## YMF2018_033_S34001.trim       2       1       0       0
## YMF2018_036_S36001.trim       2       1       1       1
## YMF2018_037_S38001.trim       0       1       0       0
## YMF2018_039_S40001.trim       2       0       1       2
```

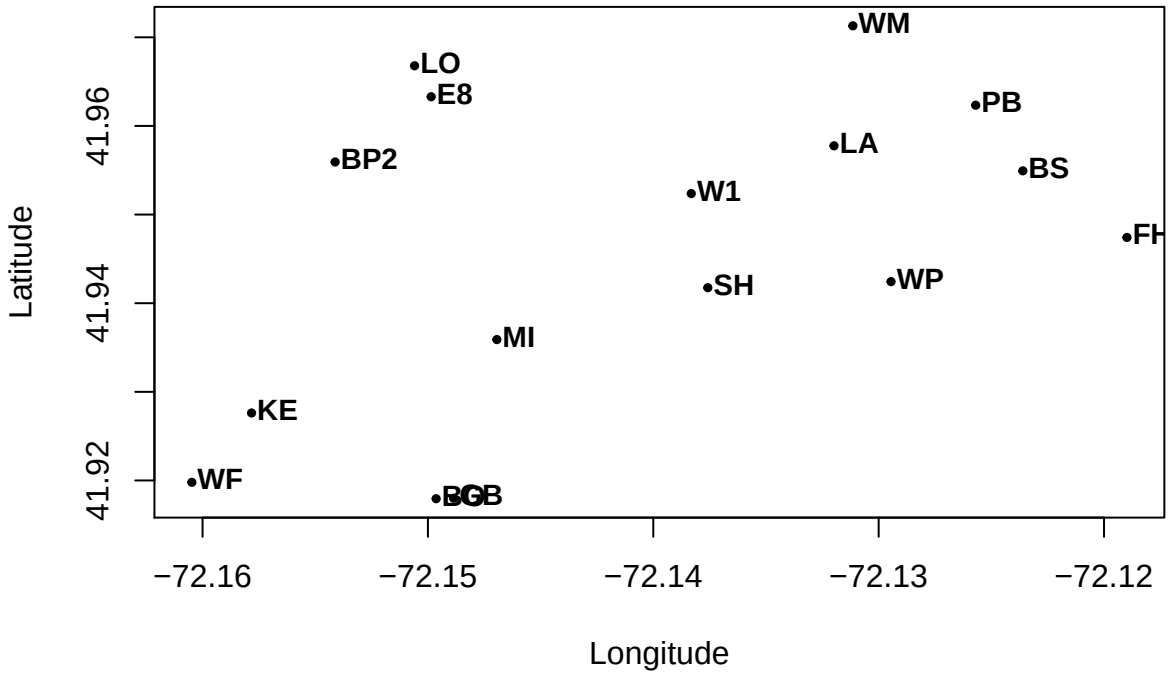
```
#Create a Matrix with long and lat
coordinates <- snps@meta[,4:5] #added pond ID here
class(coordinates)
```

```
## [1] "data.frame"
```

```
coordinates <- data.matrix(coordinates, rownames.force = NA) #force all data type into numeric
class(coordinates)
```

```
## [1] "matrix"
```

```
#verify the coords
plot(coordinates, pch = 19, cex = .5, xlab = "Longitude", ylab = "Latitude")
text(coordinates, labels=snps@meta[,3],data=snps@meta, cex=0.9, font=2, pos = 4, offset = 0.15)
```



3.2 Running the TESS3R function Population genetic structure analysis includes three main steps:

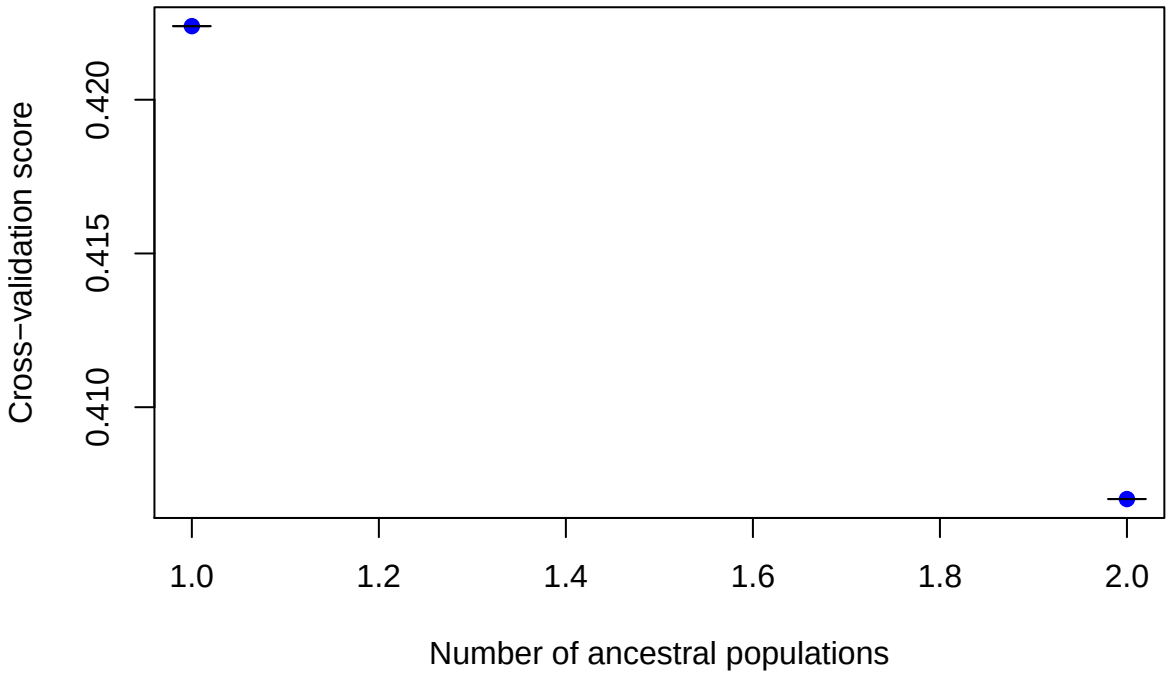
1.running one or more inference algorithms, 2.choosing the number of ancestral populations or genetic clusters, 3.showing bar-plots of ancestry coefficients or displaying them on geographic maps.

```
#Customize parameters for the run
K = c(1:2) # set the number of K to be tested
ploidy = 2 # species ploidy
CPU = 4 #Number of cores for run in parallel
```

```r
#Estimate ancestry coefficients
#set.seed(1)
tess3.ls <- tess3(X = genotypes, coord = coordinates, K = K,
                  method = "projected.ls", ploidy = ploidy, openMP.core.num = CPU)
```

```
## == Computing spectral decomposition of graph laplacian matrix: done
## ==Main loop with 4 threads: done
## == Computing spectral decomposition of graph laplacian matrix: done
## ==Main loop with 4 threads: done
```

```r
#to choose best K, generates a plot for root mean-squared errors computed on a subset of loci used for
plot(tess3.ls, pch = 19, col = "blue",
     xlab = "Number of ancestral populations",
     ylab = "Cross-validation score")
```




```r
#Smaller values of the cross-validation criterion mean better runs. The best choice for the K value is

# My scores keep decreasing as K increases, so the more K, the better. So my decision here is fairly ar
```

Results

4. Visualizing ancestry and with geography

```
# retrieve tess3 Q matrix for K = 2 clusters
#q.matrix <- qmatrix(tess3.ls, K = 2)
```

```
# plot
#pdf(file = "./Results_TESS/TESS_Ancestry_K?.pdf", width = 4, height = 4)

#my.colors <- c("red","yellow","green","purple","turquoise", "blue")
#my.palette <- CreatePalette(my.colors, 6)
#barplot(q.matrix, border = T, space = 0, main = "Ancestry matrix", xlab = "Individuals", ylab = "Ances
#axis(1, at = 0.5:nrow(q.matrix), labels = snps@meta[,3][bp$order], las = 3, cex.axis = 1)

#dev.off()
```

4.1 Ancestry Matrix in barplot of individuals

```
#prepping objects to plot
#DEM_base <- readRDS("./maps/rasters/DEM_base.rds")
#ext = extent(-72.16100, -72.11800, 41.91700, 41.97200)
#river <- raster::shapefile("./maps/vectors/rivers.shp")
#river <- spTransform(river, DEM_base@crs)
#river <- crop(river, ext)
```

```
#display interpolated values of ancestry coefficients on geographic maps

#pdf("./Results_TESS/TESS_MAP_K?.pdf")

#par(mfrow = c(1, 2))
#par(mar = c(4, 3, 2, 5))

#plot(x = q.matrix, coord = coordinates, method = "map.max", interpol = FieldsKrigModel(10),
    # main = "Ancestry coefficients",
    # xlab = "Longitude", ylab = "Latitude",cex.lab=1,
    # resolution = c(700,700), cex = 1,
    # col.palette = my.palette)
#text(coordinates, labels=snps@meta[,3],data=snps@meta, cex=0.9, font=2, pos = 4, offset = 0.3)

#with geo and elevation (I personally stitched and wrangled the DEM file using geospatial tools during
#plot(DEM_base, col = gray.colors(20, start = 0, end = 1), ext = ext, main = "Elevation and Major river
#points(snps@meta$longitude, snps@meta$latitude, pch = 20, col = "red")
#text(coordinates, labels=snps@meta[,3],data=snps@meta, cex=0.9, font=2, pos = 4, offset = 0.3, col = "
#plot(river, add = TRUE, ext = ext)
#dev.off()
```

4.2 Ancestry coefficients as spatial interpolations

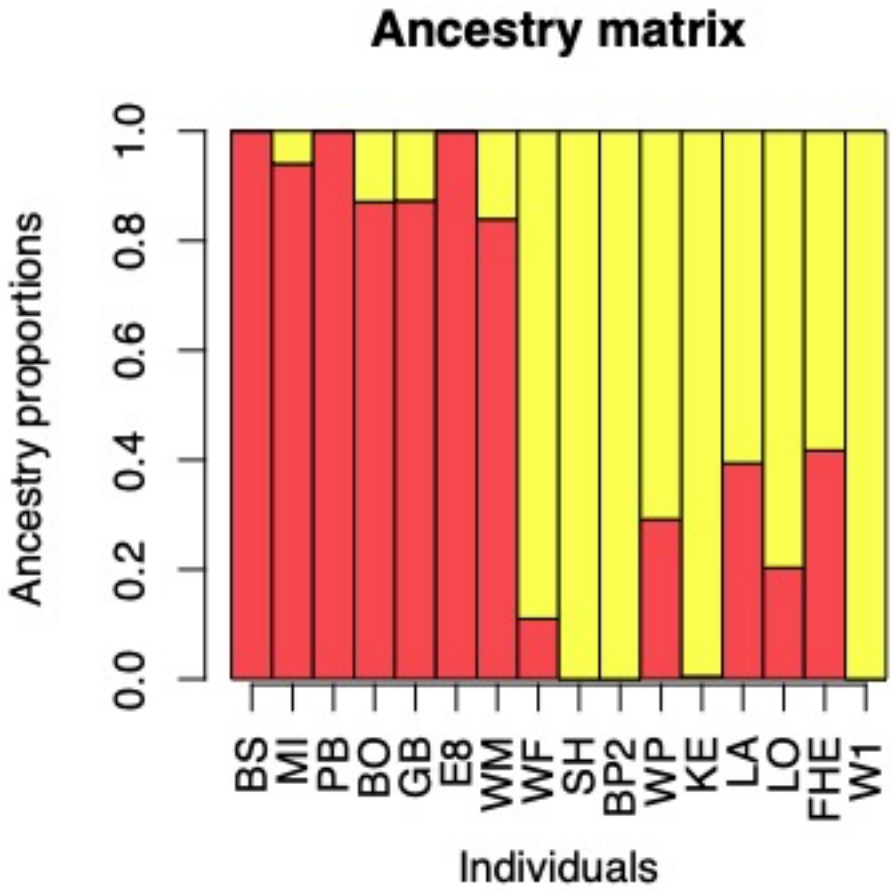
4.3 Comparing results of different K values (might not be useful) Due to the stochastic nature of tess3 function that generated the matrices for these graphs (likely due to small individual sample size and number of loci), the root mean square errors is huge. And each iteration produces different results. See 5.5. So the results might not be interpretable. Based on Caye et al. 2016 Figure 1, the function becomes more reliable when sample size is larger then 200 and loci larger than 50000. Where my dataset is a magnitude smaller than the recommended.

The assignment of colors among different K values are random.

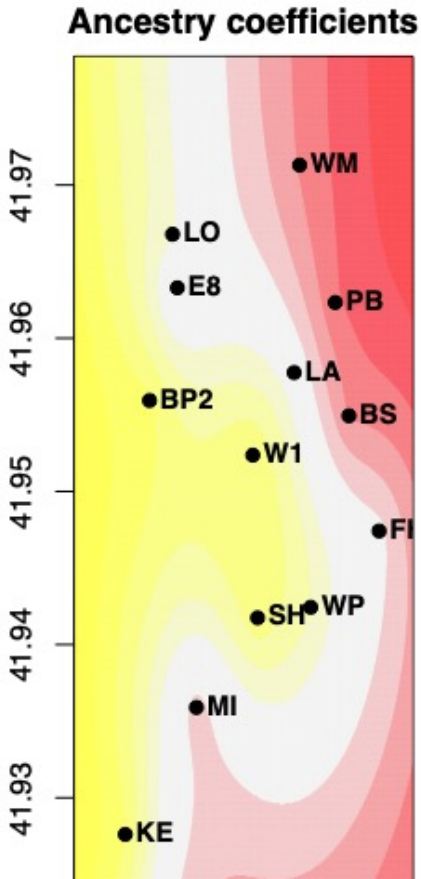
```
# root mean squared error
tess3.ls[[2]]$rmse
```

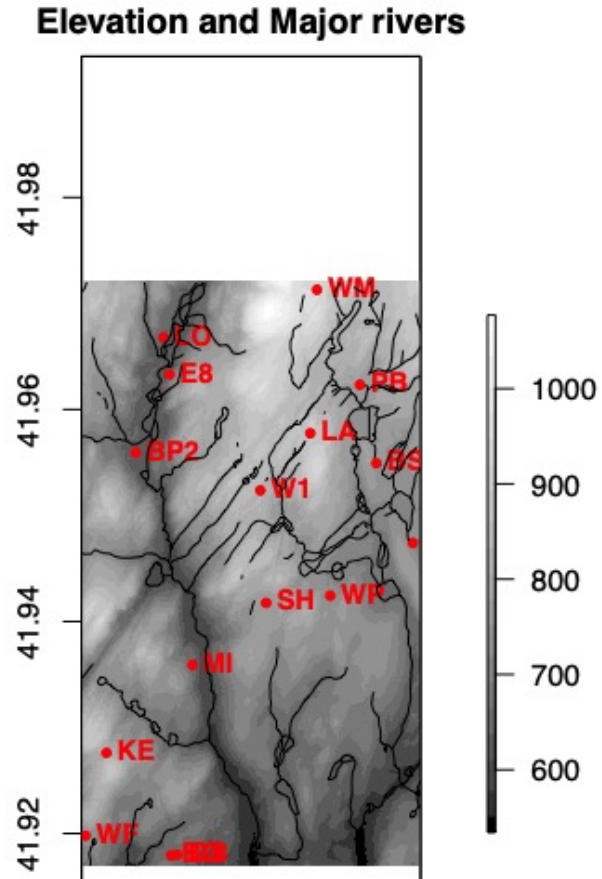
```
## [1] 0.4070118
```

The following results are here to demonstrate my work, though they are not reliable.

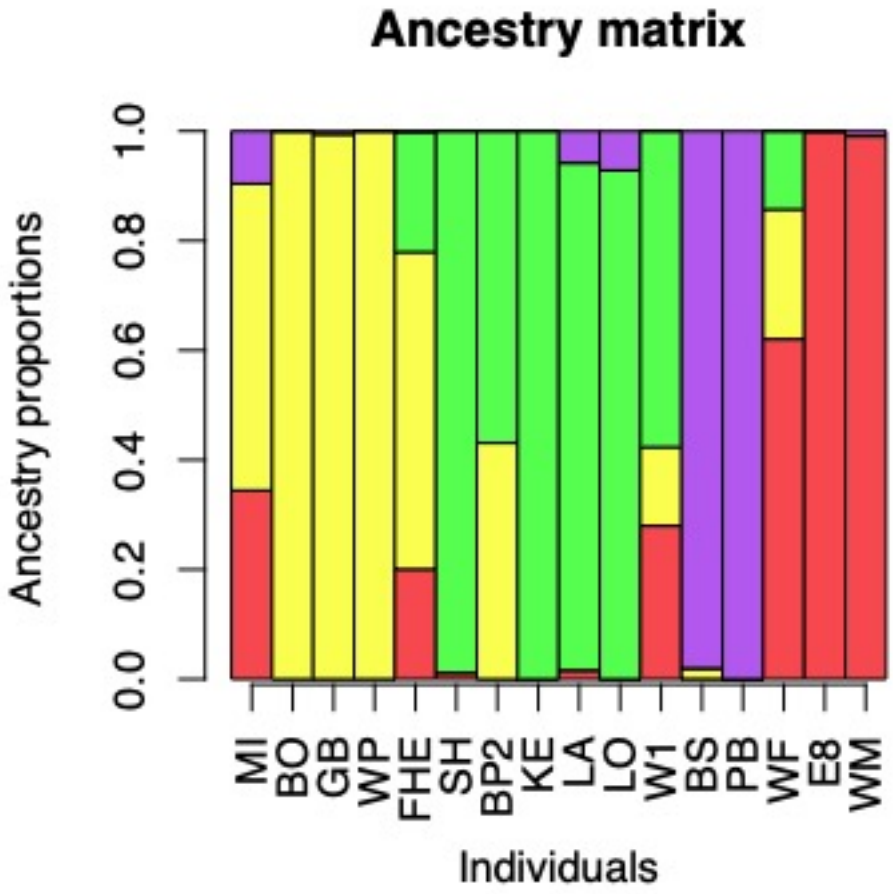


4.3.1 Results when K = 2

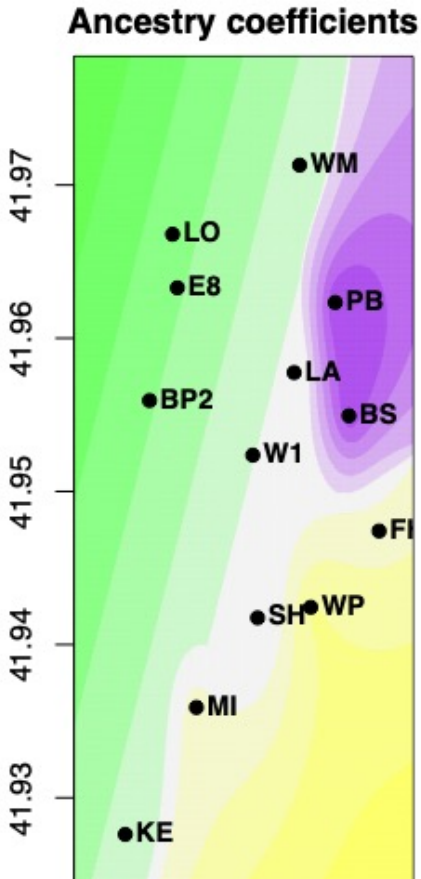


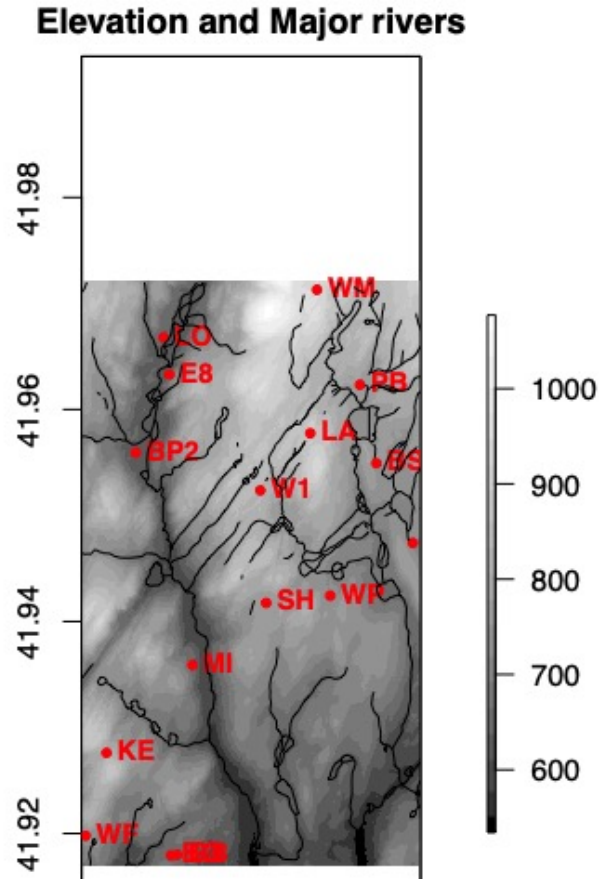


This graph shows that two clear ancestry assignments align with the valley-plataeu- ravine system. With admixture around W1, WP, and SH populations.

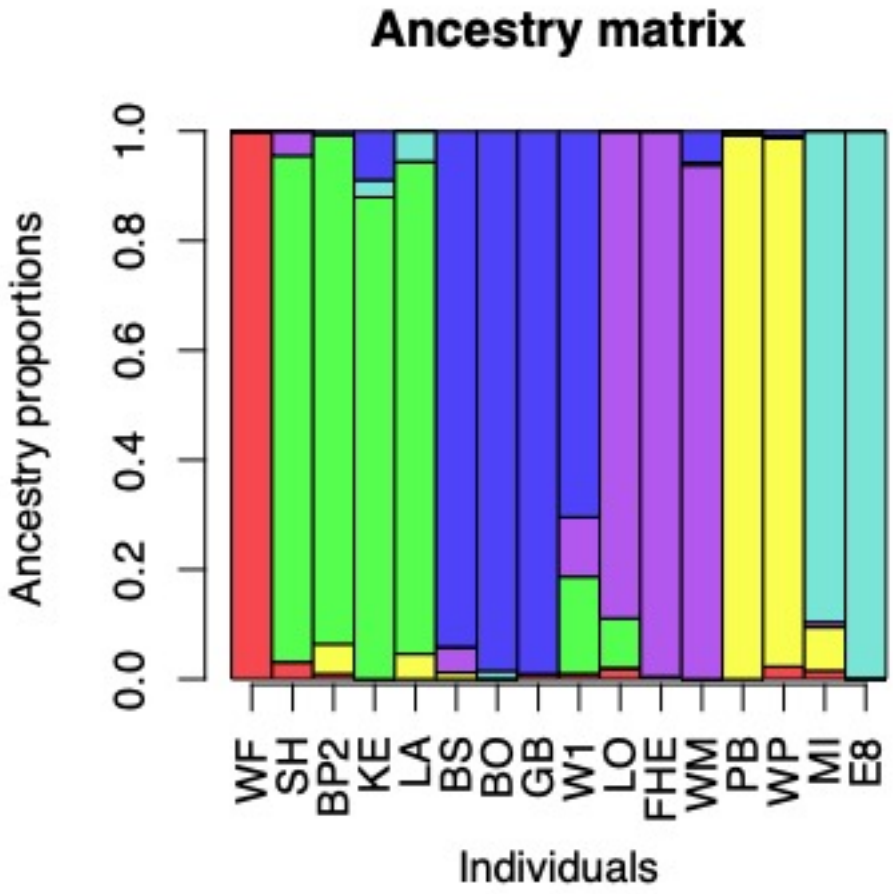


4.3.2 Results when K = 4

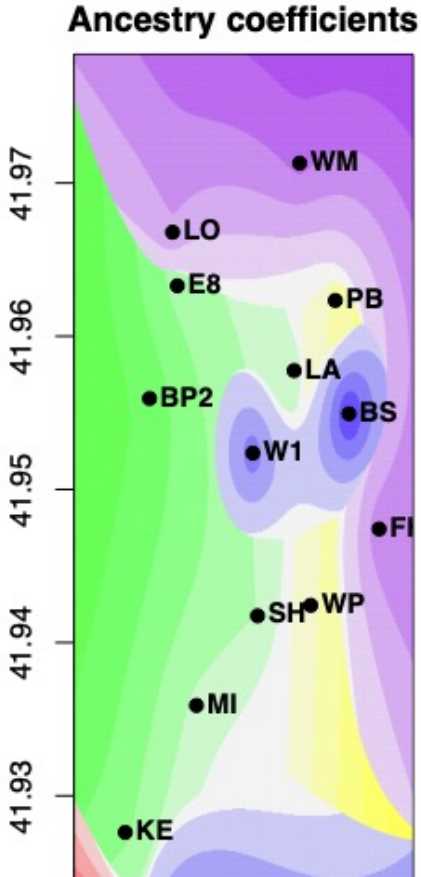


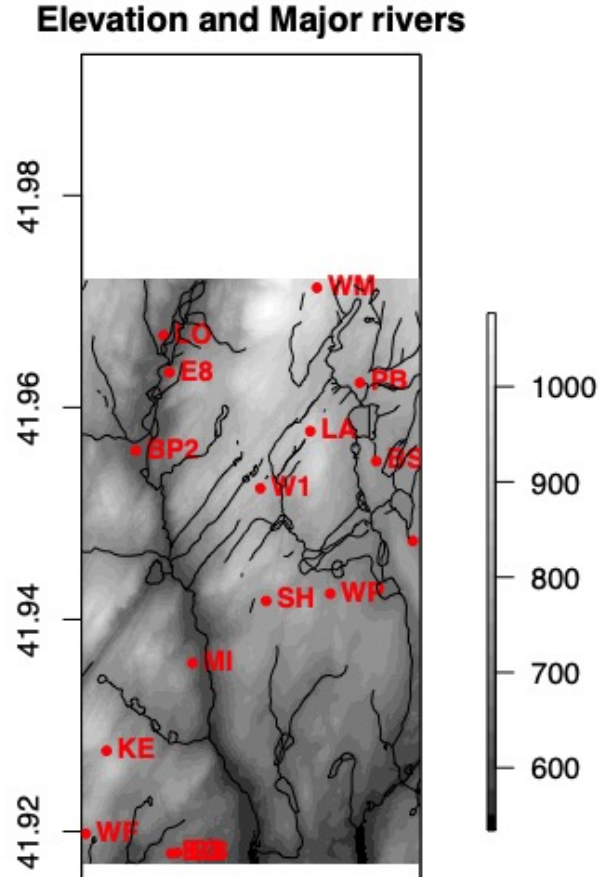


When presumed ancestry number is four, we see higher degree of admixture in all individuals, while ancestry structuring are still generally separated by elevation and river. Though additional partitioning appeared at northeast corner.



4.3.3 Results when K = 6





When ancestry assigned at 6, the admixture signature lessened, and more individuals fall into one single ancestry.The ancestry signal due to isolation by resistance from valley and river system is lost. The ancestry needs further variables to explain, which the current model doesn't capture.

Discussion

5.1 Answers to the two questions set out in the introduction

Because of the stochastic results in the ancestry coefficient due to small sample sizes, the results might not be interpretable. Here is just to demonstrate my work.

1. how are these 16 unique individuals from 16 populations related to each other by their ancestry (shared allelic frequencies across the genome by inheritance)?

The best K would be 16, that means all 16 individuals have their own ancestry lineage and none of them cluster. But the best K cross-validation value decreases almost linearly when K increases, so there is no reasonable optimal K assignment from 1 <= K <= 16.

For the interest of the final project's hypothesis with the elevation and river systems which separate the area visually by two halves, I assigned the value to K=2. The spatial ancestry interpolation largely conforms to the elevation and river system as the barrier for gene flow (section 4.3.1).

2. qualitatively, how much correspondence is there between the individuals genetic ancestry to environmental variables, such as elevation and river system (visual interpolations and comparisons)? I hypothesize that the river and elevation would partition the populations into two, due to isolation by resistance.

The correspondence is strong between the ancestries and the geographic barriers of river and elevation difference (section 4.3.1) when K ancestry assignment is 2. This suggests that these environmental variables can explain overall ancestry. When K increases, as seen in 4.3.2 and 4.3.3, we see the ancestry starts to partition in a way that is not associated with the broad pattern of elevation and geography. Other patterns such as local adaptation might be at play but adaptive loci have been filtered out in my case.

5.2 Loading file and running scripts using the Tutorial takes extra steps

To run tess3r and r2vcftools, it requires some development tools and dependencies that configure my computer. The time to create the right environment to run the tutorial scripts is time-consuming. Besides, the tutorial is written for different audiences who look for different methods to apply, so it is like an encyclopedia. I can not simply replicate and run with my data. It takes time to look through the script, pick the chunks that I need and adapt to my circumstance. This also takes a while and sometimes might render some analyses not available.

Lastly, I originally wanted to use the whole dataset (300+ individuals) for the analyses. But each chunk take too long to run, and the developmental stage that requires trial and error would be impossible if I go with the full dataset. So, essential the final project is a complete version of "minimal viability analysis." Next step would be to run the code with the full dataset.

5.3 Very data quality necessitates understanding

Upstream SNPs data filtering is essential for ancestry assignment. We want to build ancestry based on the neutral variation of populations due to gene flow and drift. My collaborator Andis filtered out only missingness and I still needed to the rest of the filtering. Genotype matrix is good for visually examining the heterozygosity of the data across different regions, and whether there is remaining missing data. Site depths determine the trustworthiness of the loci data. Necleotide diversity per site is a good measurement for amount of useful variation for all loci. Hardy Weinberg Equilibrium is good to assess amount of non-neutral variations there are in the SNPs.

5.4 Correct SNPs filtering is based on quality check

Meanwhile, we want to filter out the regions that are under selection and physical linkage. So quality check and applying appropriate filtering threshold is critical. It took me a while to carefully look through the methods of each quality check, understand them, and filter my SNPs accordingly. I don't have a reference number to draw the line of which threshold is good enough, so I'm not sure if I over-filtered or under-filtered. I started out with 18k SNPs, but left with about 3500 SNPs. And the limited number of SNPs also rendered the TESS analyses not accurate (section 5.5).

5.5 Ancestry proportion assignment in my research is stochastic, and is K dependent

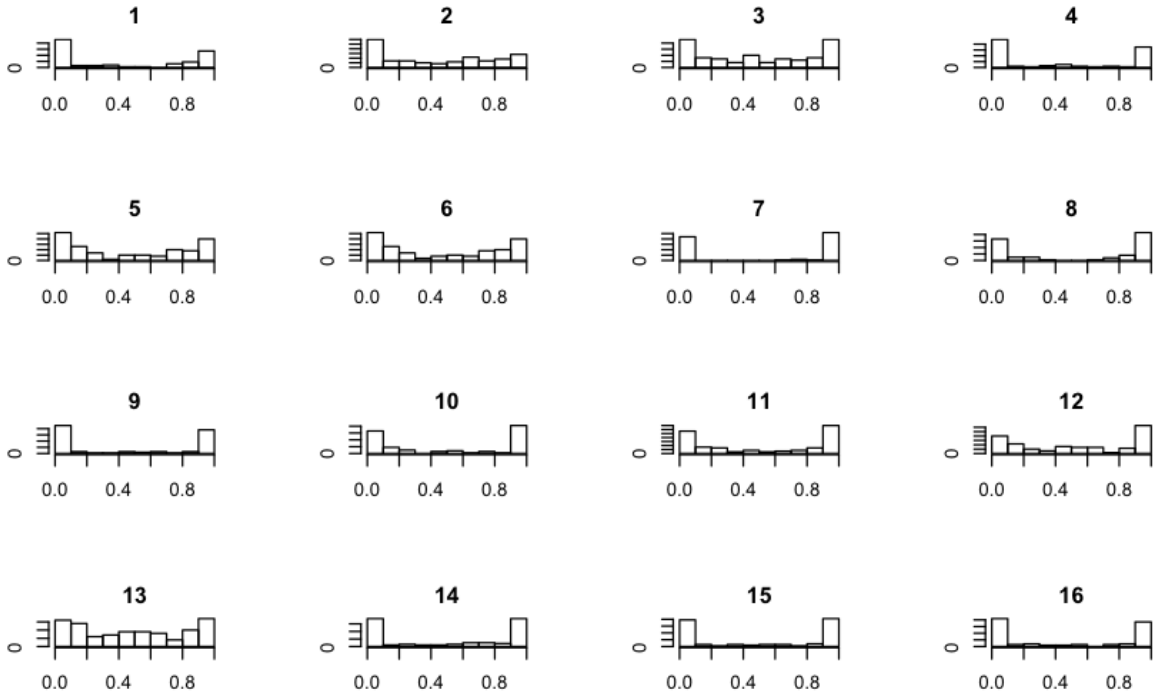


Figure 1: nothing

For the same K assignment, such as K=2, every ancestry coefficient run returns a difference result, which lead to unreliable ancestry proportion calculation and assignment. As the result above shows, each parallel run under the same parameters, return different ancestry probability, and most of them are bimodal, meaning that the ancestry is extremely vague. For example, for individual 14, about 50% of the time within the 100 iterations it belongs to ancestor cluster 1 and 50% of the time cluster 2. This pattern holds strongly in almost all other individuals across the 100 iterations. This suggests the tess ancestry proportion calculation based on 16 individual and 3658 SNPS is not reliable. When individual sample size is larger than 200 and SNPs more than 25000 should we expect stable results from this analyses (Caye et al. 2016). My full dataset has 350 individuals and potentially more SNPs if I allow more flexibility in shallow site depth, Hardy Weinberg Disequilibrium and Linkage Disequilibrium.

5.6 Ancestry mapping is exploratory without testing underlying mechanisms

Spatial Ancestry mapping tells me to which spatial extents the individuals with similar genetic composition dominate the landscape based on the coalescence. The spatial components here only included the coordinates,

and the extent of ancestry is limited by nearest neighbor graphs. This excludes the possibility that far-away individuals can be closely related due to ancestry. Also, I didn't include the environmental heterogeneity into the analyses so I can not test the correlation between elevation + river system and ancestry. Overall, I can not know which factors contributed to the separation of ancestry besides saying that there is a correspondence between elevation/river system and ancestry. Further analyses on landscape resistance by including raster type environmental datasets and associating them with the genotype matrix can test actual relations between genetics and environemnt.

References

Background:
Freidenburg, L. K., and Skelly, D. K. (2004). Microgeographical variation in thermal preference by an amphibian. Ecol. Lett. 7, 369–373. doi:10.1111/j.1461-0248.2004.00587.x.

Ligon, N. F., and Skelly, D. K. (2009). Cryptic divergence: countergradient variation in the wood frog. Evol. Ecol. Res. 11, 1099–1109. http://www.evolutionary-ecology.com/issues/v11/n07/jjar2510.pdf

Skelly, D. K. (2004). Microgeographic countergradient variation in the wood frog, Rana sylvatica. Evolution 58, 160–165. doi:10.1111/j.0014-3820.2004.tb01582.x.

Richardson, J. L. 2012. Divergent landscape effects on population connectivity in two co-occurring amphibian species: divergent landscape effects on two amphibians. Molecular Ecology 21:4437–4451.

VCFTOOLS:
The Variant Call Format and VCFtools, Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group, Bioinformatics, 2011

TESS3R:
Kevin Caye, Timo Deist, Helena Martins, Olivier Michel, Olivier Francois (2016) TESS3: fast inference of spatial population structure and genome scans for selection. Molecular Ecology Resources 16 (2), 540-548. doi: 10.1111/1755-0998.12471.

Kevin Caye, Flora Jay, Olivier Michel, Olivier Francois. Fast Inference of Individual Admixture Coefficients Using Geographic Data. bioRxiv, 2016, doi: http://dx.doi.org/10.1101/080291