Daniel Araujo and Henry Wittich

STAT 388/488 Multivariate - Course Paper 2

Paper: doi.org/10.1017/S1751731118001179

Introduction

Phylogenetics is the study of evolutionary relationships among biological entities and it centers around the construction of phylogenies to characterize relatedness of different organisms. The general practice in the field is to study similarities and differences between different organisms to produce evidence of shared common ancestry; the more similarities between two organisms, the more likely it is that they evolved from a recent common ancestor. Today, with next-generation DNA sequencing methods, a wealth of genotypic data has been produced for use in phylogenetic studies. Genetic markers are inherited, therefore they represent signatures of ancestry that can be used to trace evolutionary histories. For this reason, genetic analysis of different organisms' genomes can provide additional evidence of evolutionary relationships and help clarify experimentally-produced phylogenies. This represents the main objective of this paper: to incorporate genetic data into the study of the evolutionary relationships between six different cattle populations from Algeria and Morocco.

In this case, the evolutionary relationship they are investigating is the domestication of different cattle breeds across these population groups. The geographic distribution of these cattle populations has led to different processes of domestication for these breeds, producing different traits in the cattle populations. It is believed that the differential domestication of these cattle breeds has led to artificial selection for different genetic markers in the cattle, which resulted in their different traits. For example, in Morocco, cattle raised in the marginal areas of the country were bred for hardiness and resistance to harsh climates, however these cattle also show reduced milk production and growth rate compared to cattle in the irrigated and water-rich areas of the country, which weren't bred for these hardiness traits. This exemplifies the way that unplanned breeding of domestic animals may cause rare genetic traits to be lost within populations (and, ultimately, in the whole species), or conversely increase the frequency of unwanted, deleterious genetic variants. Thus, the main goal of the study we are analyzing was to detect the genetic population structure among these cattle breeds and provide genetic evidence for the different traits characteristic of each population. The hope is that this genetic study will benefit further study into the genetic markers affecting the traits of these cattle breeds for use in conservation efforts.

Investigation of population structure is extremely important in order to assess genetic diversity both within and between species, thus it is a useful method for constructing phylogenies using genetic data. Principal Component Analysis (PCA) is one method that has been successfully applied to genetic data to detect population structure and partition phylogenies. The PCs produced from PCA or other similarity metrics can then be used to cluster individual organisms at different levels. These are the methods employed by the paper to explore the genetic population structure of these cattle breeds; they hypothesized that if the clusters produced by their PCA/clustering analysis aligned with the known population groups of the cattle, then the differences in traits between the cattle could be attributed to different genetic markers selected for by differential breeding methods. To accomplish this, the authors

combined genotype data from 121 Algerian cattle and 82 Moroccan cattle with genotype data from the WIDDE database. In total, they had genotypes for 732 individuals belonging to 23 different breeds.

Statistical Methods Employed

The genotype data used in the study was single-nucleotide polymorphism (SNP)-level, meaning they were focused on variants between cattle where there was a difference of one nucleotide at one position throughout the genome. They performed quality control on the genotype data, filtering out SNPs that didn't meet 3 criteria: SNPs had to be common (frequency greater than 1% in population), they had to be genotyped in greater than 25% of samples, and they had to pass a test for Hardy-Weinberg equilibrium. To assess population structure, the authors implemented three distinct statistical analyses: i) Principal Components Analysis; ii) maximum likelihood estimation of individual ancestries using ADMIXTURE; and iii) Neighbor-joining (NJ) clustering, the former and latter of which we have tried to replicate.

Genotype data is highly dimensional; the study analyzed the genotype of all 732 individuals for 41,183 different SNPs. Thus, the first step in this paper's analysis was to perform PCA to detect the underlying population structure in their multivariate dataset and reduce its dimensions. In order to represent the genotypes numerically for PCA, the genotypes were stored in dosage format. For diploid organisms such as cows, which inherit 2 different copies of every genetic marker (1 from each parent), dosage format is a way to represent the number of each allele they possess in 1 of 3 numeric codes. A value of 0 indicates the individual is homozygous for the major allele, a value of 1 represents homozygous for the alternate allele, and a value of 0.5 indicates the individual is heterozygous, with a copy of each allele. They performed PCA on the genotype data in R, using the ade4 package to calculate the eigenvectors and eigenvalues for every individual based on their genotype data. They used the eigenvalues to calculate the proportion of variance explained (PVE) by each PC, reporting that the PVE by the first and second PC was 11.6% and 7.2%, respectively. They then used the first two eigenvectors for each individual to produce a scatter plot of PC1 vs. PC2.

While PCA is effective at dimension-reduction, and can produce a 2D visualization of population clusters by creating PC plots, PCA does not directly assign individuals to clusters, thus additional clustering analysis was performed. Thus, they performed NJ clustering on a genotype distance matrix for around 300 cattle from 10 of the 23 populations to generate a phylogenetic tree. In their methods section, they state that they used the R package ape to calculate the fixation index (F_{st} value) for every pair of populations, which is essentially a distance metric that looks at the average number of alleles shared between individuals within each population compared to the average number of alleles share between individuals between each population. The equation they likely used for calculating this index, though they don't include this in the paper, can be found in the appendix (figure 1). However, it appears that this tree was never created, or at least not included in the paper. F_{st} values between the 6 study populations from Algeria and Morocco are reported in table 3 of the

paper, meanwhile the tree plotted in figure 3 was constructed at the individual level, not the population level. Each branch in the tree represents one individual cow and they are all colored by the population that the cattle belong to. The tree was built off a distance matrix of allele sharing distances (ASD) between each cow, likely using the ape R package. This clustering method is performed by assigning each individual to an external node and iteratively joining the most similar individuals to common internal nodes according to the distance matrix until the entire tree is formed.

Replication

To replicate their analysis, we downloaded their dataset from an online data portal, Data INRAE. The data came in csv format, with ~730 rows - one for each individual - and ~40,000 columns - the first column was the population ID for each cow and the remaining columns contained the genotype data for each SNP used in the study. Based on the number of markers included in the file, we were provided their data post-quality control, thus we didn't have to re-perform any of the filtering discussed above. We did, however, have to make some alterations to the dataset when loading it into R in order to replicate both the PCA and clustering analysis. First, we discovered that there were non-numeric values in the dataframe, so we performed the as.numeric() function on every column to ensure everything was formatted correctly. This introduced NA values for any data points that couldn't be converted into numeric values, which the pca function from the ade4 package wasn't accepting. Since there were very few NA values and our dataset was mostly intact, we just replaced NA values with the average of the genotype values in the column - while this workaround doesn't make biological sense, it does reduce the effect that the NAs will have on the downstream analysis. A better way to replace NAs with actual genotype calls would be to perform genotype imputation, a process in which missing genotypes are inferred using linkage-disequilibrium data. This, however, would require reference panels for each population analyzed in this study, which are not available. The final alteration we made to the dataset was removing any SNPs where all individuals had the same genotype, or in other words where the variance was 0, as these provided no information for the PCA and clustering analysis.

Despite not being sure if the alterations we made to the dataset exactly matched the analysis the authors performed, as there is no information about this in the paper, we were able to successfully replicate their principal component analysis. We used the dudi.pca() method from the ade4 R package to perform the PCA. We chose not to scale the data, because all dosages are reported in the same "units" on the same scale, but centered the data. With our PCA results, we produced a PC1 vs. PC2 plot like they did in the paper, but we additionally decided to produce a PC1 vs. PC3 plot and a PC2 vs. PC3 plot (figure 2). Looking at the plots, our PC1 vs. PC2 plot matches the authors' almost exactly, indicating that we were able to accurately replicate their analysis despite not knowing exactly how the authors prepared the data for PCA. We also calculated and plotted the PVE for the first 10 PCs and notably, our values matched what the authors reported for the first two PCs (figure 3).

The final analysis of theirs that we replicated was the NJ clustering method to produce a phylogenetic tree of the cattle in the dataset. We first decided to recreate the tree shown in

figure 3 of the paper by calculating the ASD between individuals from 10 of the 23 total populations. We were able to accomplish this using ape, as the authors describe in the paper, by combining the dist.gene() function to produce the ASD distance matric and the nj() function to generate a tree using that matrix. In comparison to their figure, our individual-level tree looks very similar (figure 4). For example, we see the same population clusters forming, including the CHE and GUE mixed cluster, they are just in different orders. We decided to expand this tree by performing NJ clustering of the ASD between all ~700 individuals from every population, and we saw the same population clusters form (figure 5). Finally, we decided to replicate the NJ clustering analysis they describe in their methods section, but don't represent in the figures. While we couldn't find a function in the ape package to calculate Fst values between populations, we did find the hierfstat R package, which has a function, genet.dist() to do this. So, we used this function to generate a population pairwise Fst distance matrix between the 10 subsetted populations from Algeria, Europe, and Morocco, and again used ape's nj() function to generate the tree (figure 6). When comparing all three of these trees, we can find similar patterns in population clusters. For example, we noticed that populations ABO, MON and TAR are closely related to each other, as well as see that the CHE and GUE populations cluster together. In general, the individual-level tree seems to better show the admixture of the CHE and GUE populations, while the population-level tree effectively generalizes the relationships between populations. Additionally, in the individual-level tree for all populations, we can spot clusters also found in our PCA analysis; for instance, the BRM/GIR/NEL cluster, which is also present in the PC1 vs. PC2 scatter plot (bottom right). Ultimately, the clustering analysis reveals similar results to the ones obtained in the PCA.

Conclusion

Through replicating the methods of this paper, we have demonstrated that they effectively applied PCA and NJ clustering to explore the genetic population structure in 23 different cow breeds. When comparing all three of the PC plots we made, it seems like the PC1 vs. PC2 plot is the best at separating the dataset, validating the author's choice to display their results in this way. The other two plots produce a T shape, with one large vertical line and one horizontal line, indicating that the individuals that are clustered together on one axis get separated on the other. Likewise, the PC1 vs. PC2 plot produces a triangle shape with distinct population clusters; individuals that cluster near each other on PC1 also cluster on PC2. These results can be explained by our scree plot; the PVE by PC3 is considerably lower than the PVE by PC1 and 2, at around 1.5%. For this reason, it makes sense that the PC1 vs. PC2 plot would produce the most informative clusters. Nevertheless, the PVE by the first two PCs is low, at around 18.8% total, suggesting that only a small portion of the variation in the dataset aligns with the variations in location and breeding of the different cattle populations. For this reason, it is appropriate the authors conducted additional clustering methods. We demonstrate that the clusters they found by performing NJ-clustering on the ASD between individuals is replicable, both by reproducing their tree in figure 3, as well as producing an additional tree including all 23 populations and one at the population-level using Fst as the distance metric.

Figures

$$F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}}$$

Figure 1 - One equation for calculating the fixation index (F_{st}). πbetween represents the average number of pairwise ASD between individuals from different populations while πwithin represents the average pairwise ASD between individuals from the same population.

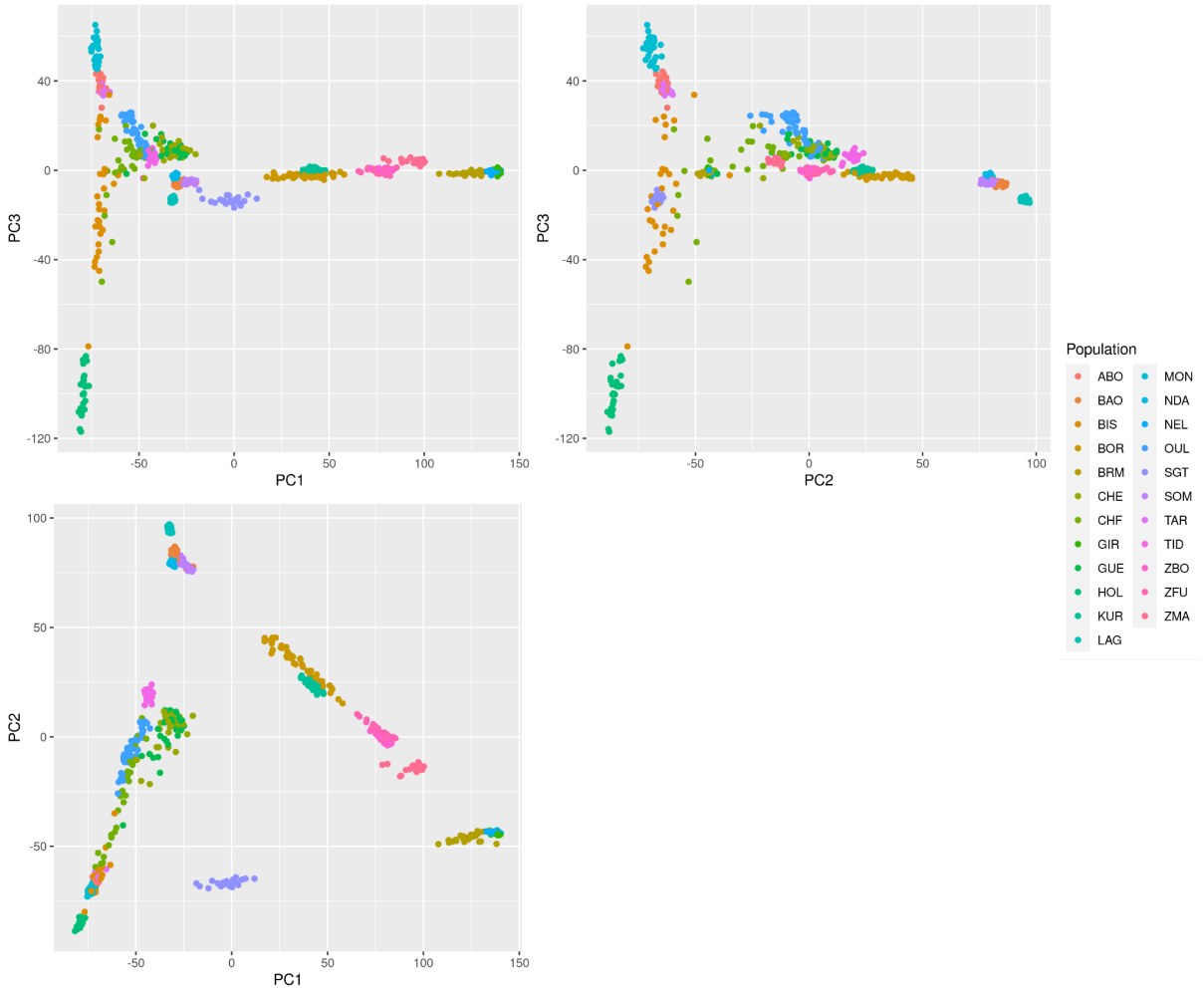

Figure 2 - Scatter plots obtained from the principal component analysis using all 732 individuals in the dataset. Individuals are plotted according to the first, second, or third principal component and colored according to their respective populations.

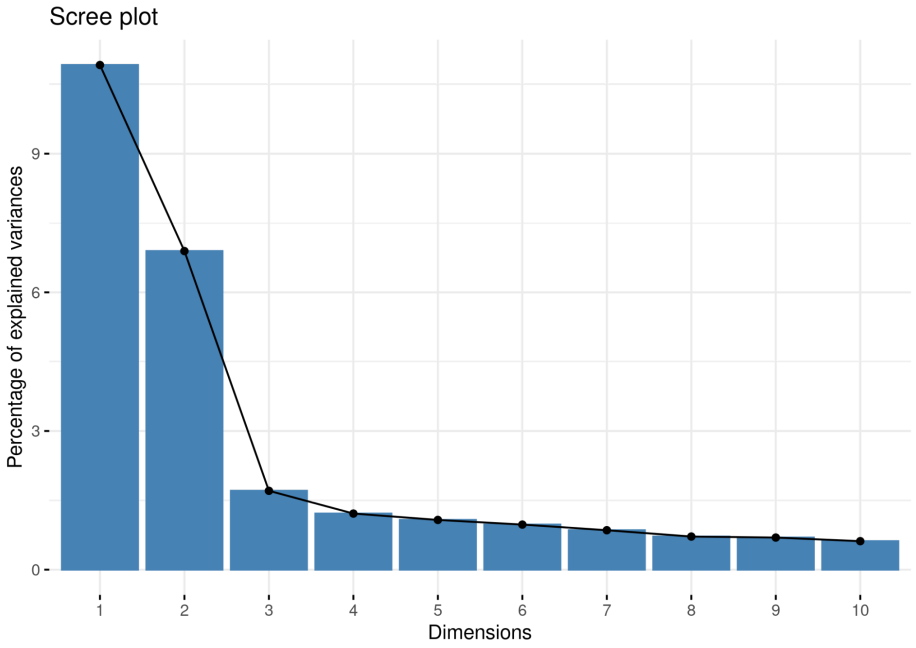

Figure 3 - Scree plot obtained from the principal component analysis using all 732 individuals in the dataset. Only the percentage of variance explained by the first ten principal components are shown.

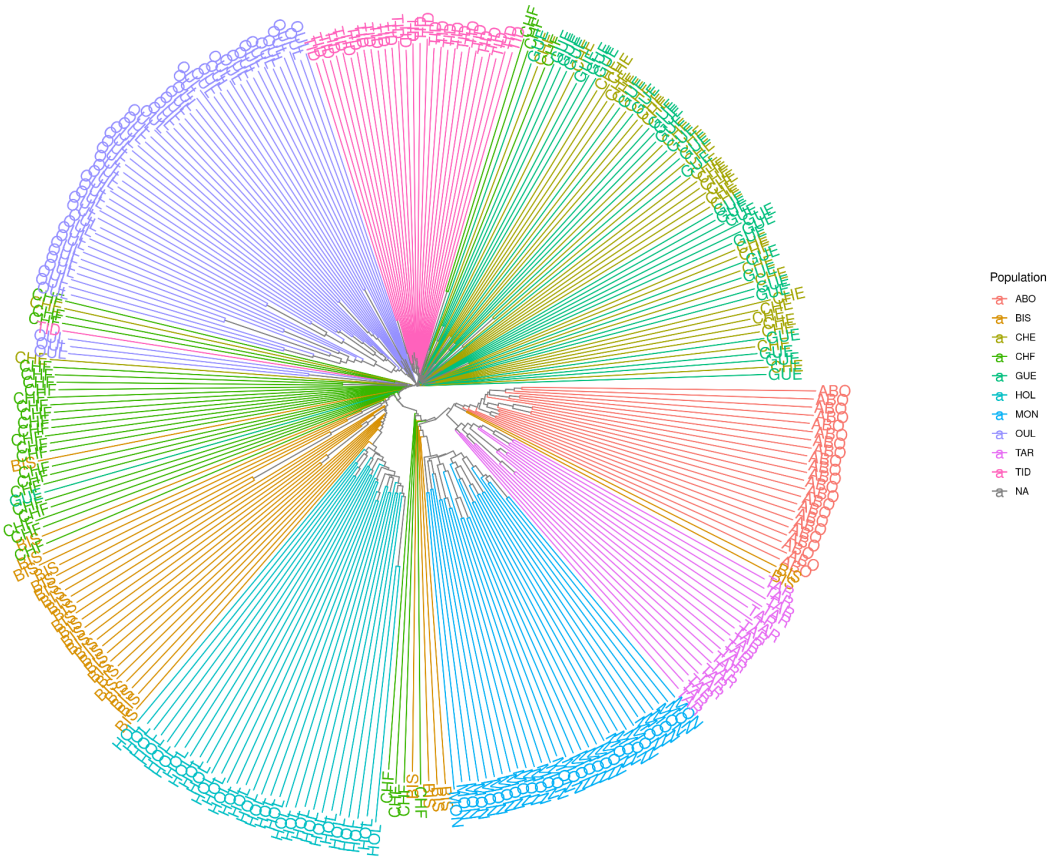

Figure 4 - Neighbor-joining phylogenetic tree built among 303 individuals in the dataset, representing ten populations from Algeria, Europe or Morocco, using the pairwise ASD matrix. Individuals (leaf nodes) and final branches are colored according to their respective populations.

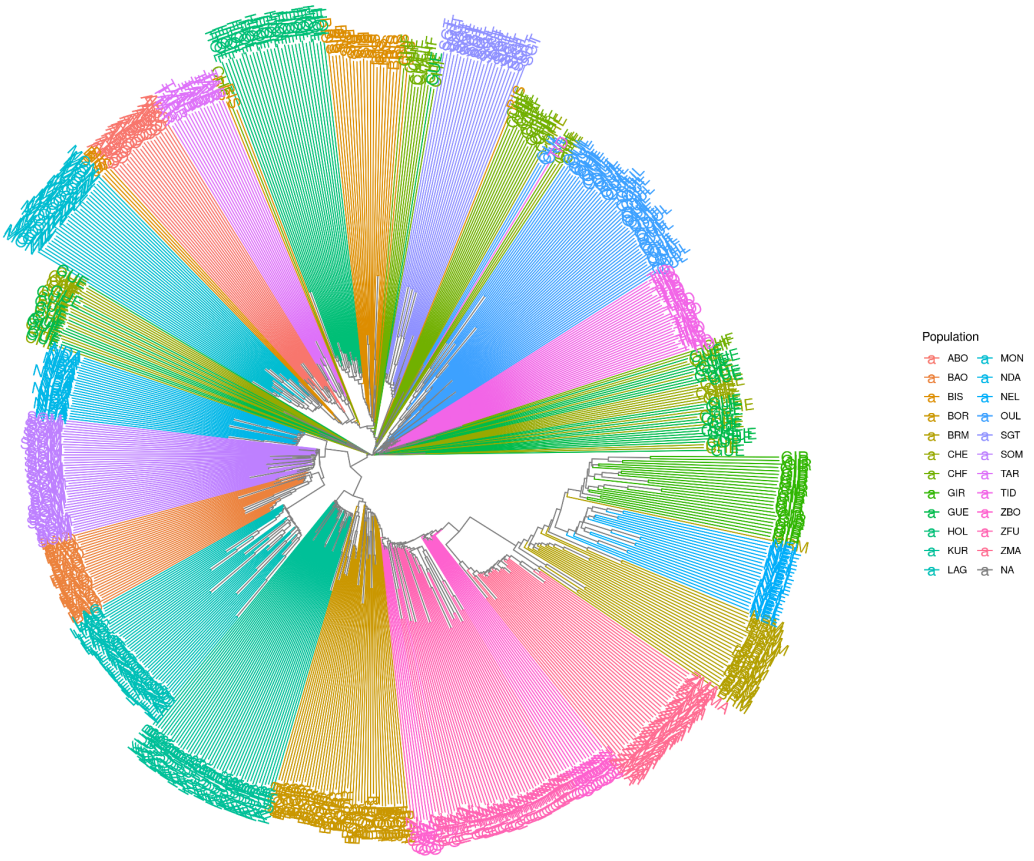

Figure 5 - Neighbor-joining phylogenetic tree built among all 732 individuals in the dataset from the pairwise ASD matrix. Individuals (leaf nodes) and final branches are colored according to their respective populations.

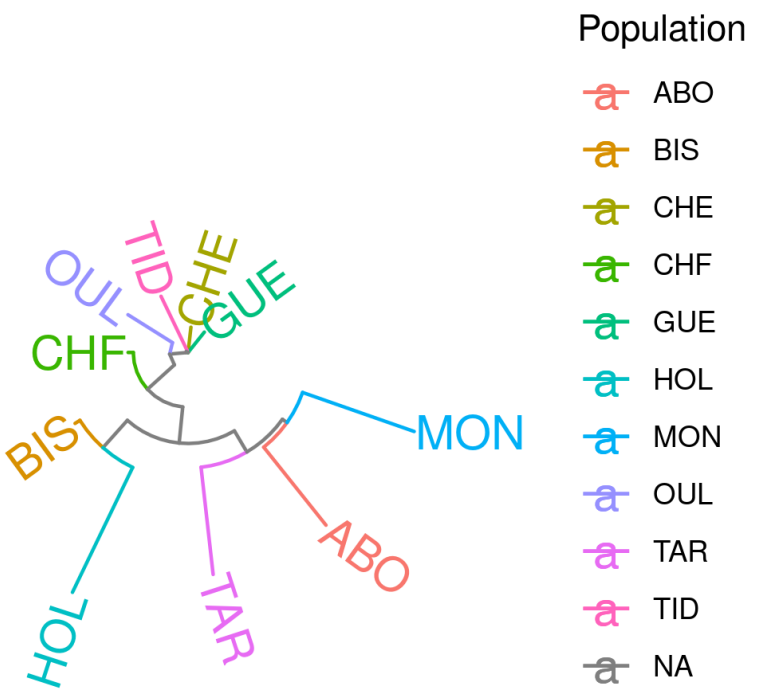

Figure 6 - Neighbor-joining phylogenetic tree built among all populations in the dataset using the Fixation index (Fst) distance matrix. Leaf nodes and final branches are colored according to their respective populations.

Code

```
# loading all necessary packages
library(ade4)
library(ape)
library(data.table)
library(factoextra)
library(ggpubr)
library(ggtree)
library(hierfstat)
library(tidyverse)

# reading genotype data
file <- fread('~/stat488/cattle_genotypes_mod.txt')

# removes first column (population IDs)
file_mod <- file[,-1] %>% as_tibble()

# There are some non-numeric values hidden in the data frame,
# so this should get rid of them.
file_mod <- apply(file_mod, 2, as.numeric) %>% as.data.frame()

# replace NAs by the average genotype because the PCA function used by the
authors
# doesn't handle NAs
file_mod <- replace_na(file_mod, as.list(colMeans(file_mod,na.rm=T)))

# removing columns with variance == 0 (that is, columns in which the genotype is
the same for everyone)
file_mod[is.nan(file_mod)] <- 0
file_mod <- file_mod[,which(apply(file_mod, 2, var)!=0)]

# runs PCA (using the function from the ade4 package, which the authors used)
res.pca <- dudi.pca(file_mod, # input data frame
                    scannf = FALSE, # do not output scree plot
                    center = TRUE, # center it (they don't mention in the
methods, but why not?)
                    scale = FALSE, # genotypes are already in the same scale, so
no need to scale it in here
                    nf = 10) # number of components kept in the results

# scree plot
fviz_eig(res.pca)
ggsave('~/stat488/scree_plot.png',width=7,height=5)

# extract different PC columns and convert to data frame
pcaData <- as.data.frame(res.pca$li[, c(1,2)])
pcaData1.3 <- as.data.frame(res.pca$li[, c(1,3)])
pcaData2.3 <- as.data.frame(res.pca$li[, c(2,3)])

# merge with population IDs
```

```
pcaData <- cbind(pcaData, file$V1)
colnames(pcaData) <- c('PC1', 'PC2', 'Population')
pcaData1.3 <- cbind(pcaData1.3, file$V1)
colnames(pcaData1.3) <- c('PC1', 'PC3', 'Population')
pcaData2.3 <- cbind(pcaData2.3, file$V1)
colnames(pcaData2.3) <- c('PC2', 'PC3', 'Population')

# plot it
a <- ggplot(pcaData, aes(PC1, PC2, color = Population)) + geom_point()
b <- ggplot(pcaData1.3, aes(PC1, PC3, color = Population)) + geom_point()
c <- ggplot(pcaData2.3, aes(PC2, PC3, color = Population)) + geom_point()
ggarrange(b, c, a, common.legend = TRUE, legend='right')
ggsave('~/stat488/PC_plots.png',width=12,height=10)

##Clustering
#Append population IDs onto file_mod
file_mod.pop <- cbind(file$V1, as.data.frame(file_mod))
colnames(file_mod.pop)[1] <- 'V1'

#Calculate pairwise allele sharing distance for all populations
# Use dist.gene() on SNP-level data to produce distance matrix
# From the ape documentation:
# This function is meant to be very general and accepts different kinds of data
# (alleles, haplotypes, SNP, DNA sequences, ...). The rows of the data matrix
represent
# the individuals, and the columns the loci.
ASD_matrix <- dist.gene(x = file_mod.pop,
                        method = 'pairwise', # distance (d) between two
individuals is the number of loci for which they differ
                        pairwise.deletion = TRUE, # Delete columns with
missing values when calculating pairwise distances #shouldn't be a problem but
why not
                        variance = FALSE) # Don't return variances of
distances

#Use nj() on distance matrix to build tree
# From the documentation:
# This function performs the neighbor-joining tree estimation of Saitou and Nei
(1987).
tree <- nj(ASD_matrix) # One argument: distance matrix

# updating tree leaf-nodes labels to be population IDs
tree_tbl <- as_tibble(tree)
tree_tbl$label[1:nrow(file_mod.pop)] <- file_mod.pop$V1
tree_t <- as.phylo(tree_tbl)

write.tree(tree_t, '~/stat488/STAT488_cattle_ALL_analysis_phylo_tree.txt')

ggtree(tree_t, layout='circular', aes(color=label)) + geom_tiplab(size=5,
aes(angle=angle)) +
  labs(color='Population')
```

```
ggsave('~/stat488/all_populations_ASD_tree.png',width=15,height=15)

#Calculate pairwise allele sharing distance, filtered down to ten populations
#Should replicate figure 3 in paper
pops <- c('ABO','BIS','CHE','CHF','GUE','HOL','MON','OUL','TAR','TID')
file_filtered <- file_mod.pop %>% filter(V1 %in% pops)
#unique(file_filtered$V1)

ASD_matrix <- dist.gene(x = file_filtered,
                        method = 'pairwise', # distance (d) between two
individuals is the number of loci for which they differ
                        pairwise.deletion = TRUE, # Delete columns with
missing values when calculating pairwise distances #shouldn't be a problem but
why not
                        variance = FALSE) # Don't return variances of
distances

#Use nj() on distance matrix to build tree
# From the documentation:
# This function performs the neighbor-joining tree estimation of Saitou and Nei
(1987).
tree <- nj(ASD_matrix) # One argument: distance matrix

# updating tree leaf-nodes labels to be population IDs
tree_tbl <- as_tibble(tree)
tree_tbl$label[1:nrow(file_filtered)] <- file_filtered$V1
tree_t <- as.phylo(tree_tbl)

write.tree(tree_t,
'~/stat488/STAT488_cattle_filtered_populations_analysis_phylo_tree.txt')

ggtree(tree_t, layout='circular', aes(color=label)) + geom_tiplab(size=5,
aes(angle=angle)) +
  labs(color='Population')
ggsave('~/stat488/filtered_populations_ASD_tree.png',width=15,height=15)

#Calculate pairwise Fst between populations
#Use genet.dist() from hierfstat package to produce population pairwise Fst
distance matrix
# Fst can be calulcated with the following equation:
# ((avg pairwise ASD between populations)-(avg pairwise ASD within populations))
/ (avg pairwise ASD between populations)
Fst_matrix <- genet.dist(dat = file_filtered, #first column is population, rest
of columns are genotypes in dosage format
                        diploid = TRUE, #cattle are diploid organisms
                        method = 'Fst') #Calculate Fst distance metric

#Use nj() on distance matrix to build tree
# From the documentation:
# This function performs the neighbor-joining tree estimation of Saitou and Nei
(1987).
```

```
tree <- nj(Fst_matrix) # One argument: distance matrix
write.tree(tree,
'~/stat488/STAT488_cattle_filtered_populations_fst_analysis_phylo_tree.txt')

ggtree(tree, layout='circular', aes(color=label)) + geom_tiplab(size=5,
aes(angle=angle)) +
  labs(color='Population')
ggsave('~/stat488/filtered_populations_fst_tree.png',width=4,height=4)
```