Question: What are the multivariate relationships between the taxa?

Use a principal component analysis that decomposes data of high dimensionality.

In command line

Step 1. Preparing the data set and software

Downloading the file into home computer, decpmpressing the file, browsing the vcf file.
Downloading another file into computer, but do noy decompress this one.
Activating conda environment. Installing **vcftools** and **plink** in home computer.

Step 2. removing 'Naxos2' using vcftools

```
vcftools --gzvcf ProjTaxa.vcf.gz --remove-indv Naxos2 --mac 1 --recode --stdout | gzip -c > test.vcf.gz
```

The output file is `test.vcf.gz`

Step 3. Doing PCA, using plink

Changing the format of vcf file.

```
plink --vcf test.vcf.gz --recode --out test --const-fid --allow-extra-chr
```

Getting files with the extension name `.map, .nosex, .ped`.

Generating a bed file based on .ped fiel

```
plink --allow-extra-chr --file test -noweb --make-bed --out test
```

Getting 2 file with extension name `.bim, .bed`.

doing PCA analysis

```
plink -allow-extra-chr --threads 20 -bfile test --pca 20 --out test
```

Getting two file with extension name `.eigenval, .eigenvec`. `.eigenval` contains the weight of each PC, and `.eigenvec` contain the vector of each individual on each PC. moving the .eigenval, .eigenvec file into local computer.

In R-studio

Step 4 editing the matrix for ploting

Installing or loading the package

```
 library(car)
library(GGally)
library(ggplot2)
library(gridExtra)
library(Hmisc)
library(lmtest)
library(MVN)
```

import the .eigenval, .eigenvec file

Setting working directory at first.

```
 eigenval <- read.table("test.eigenval")
eigenvec <- read.table("test.eigenvec")
```

Editing the format of eigenvec, adding the name of population for each row , column 2-17 is the useful value of eigenvec.
Combing two parts together, getting the a "eigenvec" that can be used to plot.

```
blend <- cbind(substr(eigenvec[c(seq(1,15)),2],1,2) , eigenvec[,c(seq(2,17))])
```

The matrix used to ploting PCA result is called `blend`.

Step 5 ploting PCA result

```
ggplot(blend, aes(x=blend[,3],y=blend[,4])) +
  geom_point(aes(color=blend[,1], shape=blend[,1]),size=3)+
  labs(x=paste("PC1", round(eigenval$percentage[1],2),"%"),
       y=paste("PC2", round(eigenval$percentage[2],2),"%" ))+
  theme_bw()+theme(legend.title = element_blank())
```

Step 6 plotting the percentage of PCs

change the name of column at the first, called `PC`

```
names(eigenval)[1] <- "PC"
```

creating a column that is the percentage of each PC called `Percentage`

```
eigenval$percentage <- eigenval$PC/sum(eigenval$PC)*100
```

creating a column that is the index of PC, called `name`

```
eigenval$name <- as.numeric(rownames(eigenval))
```

plot the percentage bar chart

```
ggplot(eigenval, aes(x = name)) +
  geom_bar(stat="identity" ,aes(y = percentage)) +
  labs(x="PC",y="Percentage of variation that explan")
```

Finish, the next step is to analyse the plot