Principal Component Analysis

Tutorial

1. **Download** out98copy vcf of the dataset from https://datadryad.org/stash/dataset/doi:10.5061/dryad.s5d698d

2. **Moving data** from local directory to graham network on the /scratch/rmuhtadi/genomics_methods_w23/Project4 directory using the code below

```
Scp username@graham.computecanada.ca:/scratch/rmuhtadi/genomics_methods_w23/Project4
```

3. **Perform linkage pruning**

Linkage pruning using plink was performed on the dataset. To complete this step, it is important to be in the same directory as the dataset, Project4 directory in this case. This step will produce two files with .prune.in and .prune.out suffix after the code below is used.

```
plink --vcf Castanea.vcf --double-id --allow-extra-chr --set-missing-var-ids @:# --indep-pairwise
50 10 0.1 --out castanea
```

4. **Perform PCA using Plink**

The code below will run PCA and produce 5 files, but only two will be used for further analysis. The files ending with .eigenval and .eigenvec will be imported to R for analysis.

```
plink --vcf Castanea.vcf --double-id --allow-extra-chr --set-missing-var-ids @:# --extract
castanea.prune.in --make-bed --pca --out castanea
```

5. **Read the data into R after exporting eigenval and eigenvec to local computer**

```
pca <- read_table2("castanea.eigenvec", col_names = FALSE)
eigenval <- scan("castanea.eigenval")
```

6. **Filter the data and change column names**

```
pca <- pca[,-1]
names(pca)[1] <- "Individuals"
names(pca)[2:ncol(pca)] <- paste0("PC", 1:(ncol(pca)-1))
pca1 <- as_tibble(pca)
```

7. **Convert to percentage variance explained**

Create a dataframe from the eigenval data and plot the data for better visualization.

```
pve <- data.frame(PC = 2:21, pve = eigenval/sum(eigenval)*100)
a <- ggplot(pve, aes(PC, pve)) + geom_bar(stat = "identity")
a + ylab("Percentage variance explained") + theme_light()
```

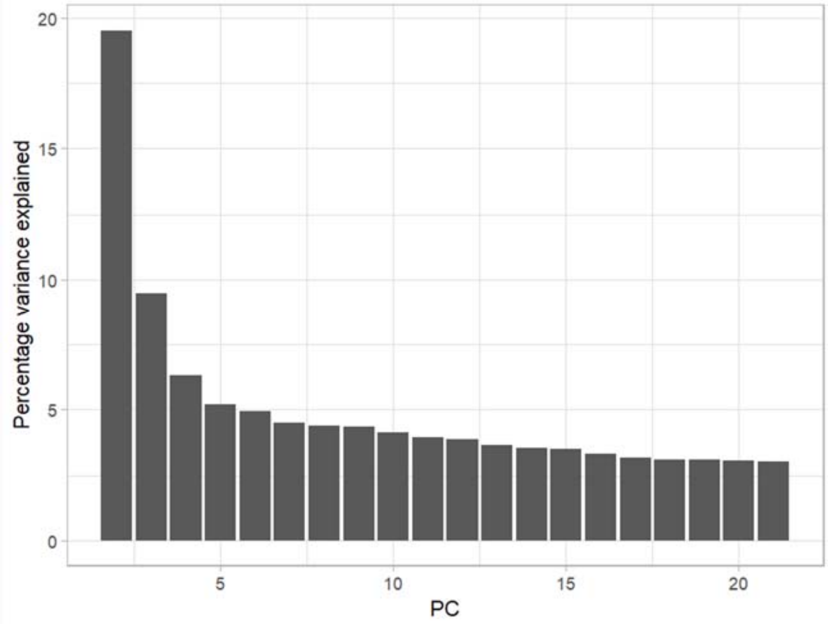

Figure 1: Percentage variance explained by PCA of the Castanea alabamensis dataset.

8. Plot PCA

```
b <- ggplot(pca1, aes(PC1, PC2)) + geom_point(size = 3)
b + xlab(paste0("PC1 (", signif(pve$pve[2], 3), "%)")) + ylab(paste0("PC2 (", signif(pve$pve[3],
3), "%)"))
```

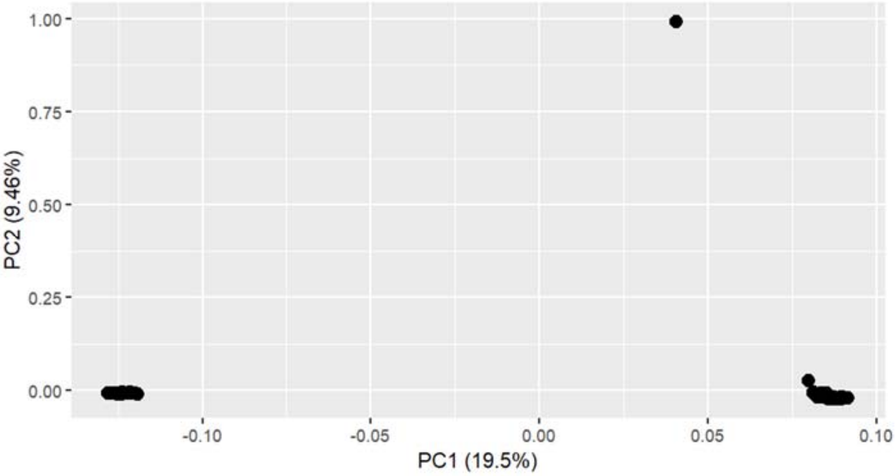

Figure 2: PCA plots of the Castanea alabamensis dataset. The plot shows 3 different clusters of the organism.