# Genetic relationship lab assignment

In this lab assignment, the following topics will be covered:

1. IBD: Coefficient of coancestry/fraternity.

2. IBS: Additive/dominance genetic relationship matrix.

3. Genetic distance: Sub-population/genome-wide Fst.

4. Population structure: Principal component analysis (PCA).

You will be using two new datasets which are different from what we used in the lab session. The data files should be provided together with this lab assignment, but in case that it is not, you may download them here. Below are some descriptions of the files in each dataset.

1. **Pine**. This dataset is a subset of individuals from Resende et al 2012. The individuals are derived from crosses between two out of four parents. Between any two individuals, they could be full-siblings, half-siblings or unrelated.
   - pine_geno.csv: genomic marker data (coded as 0/1/2/NA) for 58 individuals and 4,853 markers.
   - pine_ped.csv: pedigree data for 58 individuals and 4 parents.
2. **Oat**. This dataset is a subset of lines from Tinker et al 2014. There are three classifications for the lines: Spring Canada (CA), Spring US, and Winter US. As you may already know, spring oat does not require vernalization but winter oat does.
   - oat_geno.csv: genomic marker data (coded as 0/1/2/NA) for 535 lines and 3,487 markers.
   - oat_info.csv: grouping information for 535 lines.

R scripts are provided for most parts but there are parts where you have to fill in the information - these are usually indicated by XXX, YYY or ZZZ. The analyses are largely sequential and so there may be sections that require previous outputs. Most of the analyses here have been covered in the lab session. If you are unsure how to use any of the functions or packages, please refer to the lab practice. If you wish, you may use different packages that you prefer, or write your own R scripts for custom calculation - in these cases, please include the R scripts in your answers as well. If you are having trouble with installing the package(s) in your computer, you could either do it manually without the package or borrow someone's else computer. Please let me know (cyang@sruc.ac.uk) if you are stuck or having trouble in this lab assignment.

**There are 10 mandatory questions in total that are worth 1-point each. There are also 2 bonus questions that are worth 1-point each, and these are optional. For example, if you score 9 points in the mandatory questions and 2 points in the bonus questions, your final score will be capped at 10 points (100%). Please read the questions carefully as several questions have two or more parts. Please write your answers in a separate document (Word/Markdown/etc) and provide as much justifications to your answers as needed.**.

## Initial preparation

Below are the packages that you may need for this assignment. Load whichever package you need and feel free to use other packages if you want.

```
library(kinship2)
library(AGHmatrix)
library(sommer)
library(hierfstat)
library(ggplot2)
```

Set your working directory to the folder where your data is. For example:

```
setwd("C:/XXX/YYY/ZZZ")
```

Load the data for this lab assignment.

```
pine_geno <- read.csv("pine_geno.csv", row.names=1)
pine_ped <- read.csv("pine_ped.csv")
oat_geno <- read.csv("oat_geno.csv", row.names=1)
oat_info <- read.csv("oat_info.csv")
```

Convert the marker genotype data from data frame to matrix.

```
pine_geno <- as.matrix(pine_geno)
oat_geno <- as.matrix(oat_geno)
```

## IBD: Coefficient of coancestry

Calculate the coefficients of coancestry in the **Pine** dataset using the pedigree information in `pine_ped`. An option to do this is using the `kinship()` function from the *kinship2* package.

```
# check the pedigree.
View(pine_ped)

# calculate the coefficients of coancestry.
pine_coancestry <- kinship(id=XXX, dadid=YYY, momid=ZZZ)
```

Check the results/coefficients.

```
View(pine_coancestry)
```

**Question 1**. What are the coefficients of coancestry for the following three pairs of individuals? [1-point]
AB01 and AB05
AB01 and AD01
AB01 and CD01

**Bonus question 1**. Looking at the pedigree, what are the relationships for the three pairs of individuals? For example, are they full-siblings, half-siblings, or unrelated? [1-point]

## IBD: Coefficient of fraternity

Calculate the coefficients of fraternity in the **Pine** dataset using the pedigree information in `pine_ped`. An option to do this is using the `Amatrix()` function from the *AGHmatrix* package.

```
# calculate the coefficients of fraternity.
pine_fraternity <- Amatrix(data=XXX, dominance=YYY)
```

Check the results/coefficients.

```
View(pine_fraternity)
```

**Question 2**. What are the coefficients of fraternity for the following three pairs of individuals? [1-point]
AB01 and AB01 (to itself)
AB01 and AB05
AB01 and AD01

## IBS: Additive genetic relationship matrix (A-GRM)

Create the A-GRM for the **Pine** dataset using the genomic marker data in `pine_geno`. Before you begin, check the coding for the marker data.

```
# note: unlike our lab practice, there are missing data here.
View(pine_geno)
```

An option to do this is using the `A.mat()` function from the *sommer* package.

```
pine_AGRM <- A.mat(X=XXX-1)
```

Check the results.

```
View(pine_AGRM)
```

**Question 3**. What are the values of A-GRM for the following three pairs of individuals? Round your answers to 3-decimal places. [1-point]
AB01 and AB05
AB01 and AD01
AB01 and CD01

Recall from the lecture and lab that A-GRM is the IBS-version of twice the coefficient of coancestry. Let's compare the values from these two approaches.

```
# the matrix for coefficients of coancestry contains the parents - remove them.
pine_coancestry2 <- pine_coancestry[-c(1:4), -c(1:4)]

# calculate the correlation between the coefficients of coancestry and A-GRM.
cor(pine_coancestry2[lower.tri(x=pine_coancestry2, diag=TRUE)],
    pine_AGRM[lower.tri(x=pine_AGRM, diag=TRUE)])
```

**Question 4**. What is the correlation between the coefficients of coancestry and A-GRM? Round your answer to 3-decimal places. Why is the correlation not 1? [1-point]

## IBS: Dominance genetic relationship matrix (D-GRM)

Create the D-GRM for the **Pine** dataset using the genomic marker data in `pine_geno`. An option to do this is using the `D.mat()` function from the *sommer* package.

```
pine_DGRM <- D.mat(X=XXX-1, nishio=FALSE)
```

Check the results.

```
View(pine_DGRM)
```

**Question 5**. What are the values of D-GRM for the following three pairs of individuals? Round your answers to 3-decimal places. [1-point]
AB01 and AB05
AB01 and AD01
AB01 and CD01

## Genetic distance: Sub-population Fst

Calculate the Fst values for each sub-population in the **Oat** dataset. First, check how the lines are classified into different sub-populations.

```
table(oat_info$Group)
```

Next, check the coding for the marker data.

```
View(oat_geno)
```

An option to calculate the Fst values for each sub-population is using the `fs.dosage()` function from the *hierfstat* package. In the output of this function, the first row is Fis (inbreeding coefficient for each sub-population) and the second row is Fst (fixation index for each sub-population).

```
fs.dosage(dos=XXX, pop=YYY)
```

**Question 6**. What are the Fis values for each sub-population? Are these sub-populations highly inbred or outbred? [1-point]

**Question 7**. What are the Fst values for each sub-population? Which sub-population is the most distant from the overall population? [1-point]

## Genetic distance: Genome-wide Fst

For this section, let's subset the **Oat** dataset to only two sub-populations: Spring_US and Winter_US.

```r
oat_geno_US <- oat_geno[oat_info$Group%in%c("Spring_US", "Winter_US"), ]
oat_info_US <- oat_info[oat_info$Group%in%c("Spring_US", "Winter_US"), ]
```

Recall from the lecture and lab that we can also calculate Fst values for each genomic marker between two sub-populations. To do so, you will need the `varcomp.glob()` function from the *hierfstat* package.

```r
# compute the variance components in each marker - this may take ~5 minutes.
oat_gwfst <- varcomp.glob(levels=oat_info_US[, "Group", drop=FALSE], loci=data.frame(oat_geno_US))

# Fst is taken as the first column of the variance components divided by the sum.
oat_gwfst <- oat_gwfst$loc[,1]/rowSums(oat_gwfst$loc)
```

Check the markers with the highest Fst values.

```r
tail(sort(oat_gwfst))
```

**Question 8**. Which marker has the highest Fst value and what is its Fst value? Round your answer to 3-decimal places. [1-point]

[ *Optional* ] Normally, we would display the results in a Manhattan plot. Unfortunately, we do not have the marker position information here. Instead, you can use the scripts below to compare the linkage disequilibrium (LD) pattern for markers with different Fst values. Because we did not calculate LD in the lab practice, this will be a bonus question and so there is no penalty from not doing this. Replace XXX with one of the markers with high Fst values that you have identified above. The horizontal axis shows the markers sorted from lowest to highest Fst values, while the vertical axis shows the LD between your chosen marker and other markers.

```r
plot(1:ncol(oat_geno_US),
     cor(oat_geno_US[, "XXX"], oat_geno_US[, order(oat_gwfst)], use="pairwise.complete.obs")^2,
     xlab="Markers sorted from lowest to highest Fst",
     ylab="r2")
```

**Bonus Question 2**. Based on the plot, is there any pattern that you can conclude from the LD involving high versus low Fst markers? What might the pattern imply about the markers that differentiate between Spring and Winter Oat? [1-point]

## Population structure: Principal component analysis (PCA)

Let's do a PCA on the **Oat** dataset. Recall that principal components can be calculated using the `eigen()` function on the A-GRM.

```r
# calculate the A-GRM for oat.
oat_AGRM <- A.mat(X=XXX-1)

# calculate the principal components.
oat_pc <- eigen(x=YYY)
```

Plot the cumulative percent variance explained (PVE) for the principal components.

```r
# prepare the data for plotting.
oat_pve <- data.frame(PC=1:nrow(oat_geno), PVE=cumsum(100*oat_pc$values/sum(oat_pc$values)))

# create the plot.
ggplot() +
  geom_point(data=oat_pve, aes(x=PC, y=PVE)) +
  theme_bw() +
  coord_cartesian(ylim=c(0,100))
```

The plot might be a little hard to identify the answers for the following question, so let's look at `oat_pve` instead.

```r
View(oat_pve)
```

**Question 9**. How many principal components are needed to achieve at least 50% PVE? What is the PVE for PC1? Round your answer to 3-decimal places. [1-point]

Let's see what can be inferred from the first two principal components.

```r
# prepare the data for plotting.
oat_pca <- data.frame(PC1=oat_pc$vectors[,1],
                      PC2=oat_pc$vectors[,2],
                      Group=oat_info$Group)

# plot the PCs.
ggplot() +
  geom_point(data=oat_pca, aes(x=PC1, y=PC2, color=Group)) +
  theme_bw() +
  coord_fixed()
```

**Question 10**. What grouping pattern can you conclude from the plot of first two principal components? In other words, what do PC1 and PC2 separate? [1-point]

---

This is the end of the lab assignment. Please email me (cyang@sruc.ac.uk) if you are having any trouble with this.