# Identification of population structure in genotype matrix

This lab is built on the class materials prepared by Prof. John Novembre (thanks a lot!). We only cover a fraction of the entire exercise, but interested readers can visit the original github.

For questions and bug reports, please contact Yongjin Park `ypp@csail.mit.edu`

## 0. Installation of software

**Note for the Windows users:** You could install virtual box or seriously consider using docker since most of computational biology pipelines run on unix-like environment.

First download our favorite software PLINK and unzip.

To estimate admixture structure of population genetics we will use another software Admixture.

For instance in MAC OS, you can download

```
wget https://www.genetics.ucla.edu/software/admixture/binaries/admixture_macosx-1.3.0.tar.gz
```

simply unzip:

```
tar xzvf admixture_macosx-1.3.0.tar.gz
```

Check if the binary file is excitable. You should be able to see output like this.

```
./admixture_macosx-1.3.0/admixture
****                   ADMIXTURE Version 1.3.0                 ****
****                  Copyright 2008-2015                      ****
****           David Alexander, Suyash Shringarpure,           ****
****                John  Novembre, Ken Lange                  ****
****                                                           ****
****                 Please cite our paper!                    ****
****   Information at www.genetics.ucla.edu/software/admixture  ****

Usage: admixture <input file> <K>
See --help or manual for more advanced usage.
```

To make files more organized, let's put all the binary files under `./bin` subdirectory.

## 1. Preparation of data

Here are files you will find in the data directory:

```
   32 -rw-r--r--  1 staff    13K Apr 20 16:53 1KG.samples.gz
   48 -rw-r--r--  1 staff    21K Apr 20 16:32 H938.clst.txt
39920 -rw-r--r--  1 staff    19M Apr 20 16:32 H938_Euro.bed
35824 -rw-r--r--  1 staff    17M Apr 20 16:32 H938_Euro.bim
    8 -rw-r--r--  1 staff   2.9K Apr 20 16:32 H938_Euro.fam
60664 -rw-r--r--  1 staff    30M Apr 20 16:53 chr22.bed
 2744 -rw-r--r--  1 staff   1.3M Apr 20 16:53 chr22.bim
  128 -rw-r--r--  1 staff    61K Apr 20 16:53 chr22.fam
```

Let's identify population structure of `H938` fileset. First we subsample the maker SNPs by LD-pruning.

```
mkdir ./result/
cd ./result/
../bin/plink --bfile ../data/H938_Euro --indep-pairwise 50 10 0.1
../bin/plink --bfile ../data/H938_Euro --extract plink.prune.in --make-bed --out H938_Euro.LDprune
```

## Short questions

- Why is the LD-pruning step useful or even necessary?

# 2. Estimate admixture models

## What is admixture model?

A genetic (a mixture of mixture) model can be best understood in terms of a generative model. First of all, we need to assume SNPs are exchangeable, meaning we treat each individual as "a bag of SNPs."

1. Suppose there are `K` ancestral populations (or `K` colors).
2. For each population `k` and each SNP `j`, we sample minor allele frequency `F[k,j]` between `0` and `1`.
3. For each individual `i`, we sample a propensity of populations `Q[i,k]` such that `sum Q[i,k] = 1`.
4. Within this individual `i`, for each SNP `j`, we sample the origin `k` of the SNP `j` proportional to individual-specific population propensity, which is `Q[i,k]`.
5. Then for this SNP `j` we twice sample the haplotype using the population `k`-specific allele frequency on this location `j`, which is `F[k,j]`.

In sum, admixture estimate two matrices: individual `i`-specific population `k` propensity `Q[i,k]` and population `k`-specific allele frequency of SNP `j` `F[k,j]` given the observed genotype matrix `G[i,j]`.

- Probability of individual `i` being homozygous recessive at SNP `j` = `(QF)[i,j]^2` or = $\left[\sum_k Q_{ik} F_{kj}\right]^2$

- Probability of individual `i` being heterozygous at SNP `j` = `2(QF)[i,j] * (Q(1-F))[i,j]` or = $2\left[\sum_k Q_{ik}F_{kj}\right]\left[\sum_k Q_{ik}(1-F_{kj})\right]$
- Probability of individual `i` being homozygous dominant at SNP `j` = `(Q(1-F))[i,j]` or = $\left[\sum_k Q_{ik}(1-F_{kj})\right]^2$

Assuming genotype `G[i,j]` were sampled based on multinomial model, for each individual `i` on the SNP `j` we have likelihood (Eq 2 of the admixture paper):

$$\ln P(G_{ij}|Q,F) = G_{ij}\ln\left(\sum_k Q_{ik}F_{kj}\right) + (2-G_{ij})\ln\left(\sum_k Q_{ik}(1-F_{kj})\right)$$

## Try it out

Let's run the `admixture` software to see if we can estimate admixture structures with some arbitrary number of populations.

```
../bin/admixture H938_Euro.LDprune.bed 6
```

You will have the following results:

```
H938_Euro.LDprune.6.P
H938_Euro.LDprune.6.Q
```

### Short questions

- What are these files? See the manual or try `../bin/admixture --help`.
- How do you determine the number of populations in the model (model complexity)? See the manual.
- Repeat the same thing with different random seed option `--seed=X`. Why is it necessary (hint: EM algorithm)?

# 3. Interpret the results

Probably the best way to interpret the result is to visualize them. Here is an example with the `K=6` model. Since the individuals in our data were sampled from known European populations, we can compare the inferred population structures with the actual populations.

```
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
source('../Util.R')

.delim <- function(...) read_delim(..., delim = ' ')

Q <- .delim('H938_Euro.LDprune.6.Q', col_names = FALSE)
P <- .delim('H938_Euro.LDprune.6.P', col_names = FALSE)
```

```
fam.tab <- .delim('H938_Euro.LDprune.fam', col_names = 'iid', col_types = '_c____') %>%
    mutate(xpos = 1:n())

clust.tab <- .delim('../data/H938.clst.txt', col_names = c('iid', 'pop'), col_types = '_cc')

fam.tab <- fam.tab %>% left_join(clust.tab) %>%
    arrange(pop)

colnames(Q) <- 1:ncol(Q)

Q.melt <- Q %>% mutate(xpos = 1:n()) %>%
    gather(key = 'k', value = 'Q', -xpos) %>%
    left_join(fam.tab)

Q.argmax <- Q.melt %>%
    group_by(iid) %>%
    slice(which.max(Q)) %>%
    arrange(pop) %>%
    as.data.frame()

Q.melt.sort <- Q.melt %>%
    mutate(iid = factor(iid, Q.argmax$iid))

fam.tab.sort <- fam.tab %>%
    mutate(iid = factor(iid, Q.argmax$iid))

## plot the hidden population components
p1 <- ggplot(Q.melt.sort, aes(x = iid, y = Q, fill = k, color = k)) +
    theme_bw() +
    geom_bar(position = 'stack', stat = 'identity') +
    scale_x_discrete(position = 'top') +
    xlab('individuals i') +
    scale_fill_discrete(guide = guide_legend(nrow = 1)) +
    theme(axis.text.x = element_text(angle = 80, hjust = 0, vjust = 0, size = 3),
          legend.position = 'top')

## compare with known population
p2 <- ggplot(fam.tab.sort, aes(x = iid, y = pop)) +
    geom_tile() +
    xlab('individuals i') +
    theme(axis.text.x = element_text(angle = 80, hjust = 0, vjust = 0, size = 3))

out <- grid.vcat(list(p1, p2), heights = c(2, 1))

ggsave(filename = 'Fig_pop.pdf', plot = out, width = 8, height = 5)
```

### Short questions

- What is your interpretation of your findings? Do you see the inferred structures agree with known population structures? If not, why?
- We haven't looked at the `P` file (which contains the population-specific allele frequency). Can you prioritize most informative markers? Plot your results and justify your answers.

## Lab questions

1. Follow the steps and make a report.

2. Repeat the same analysis with the 1000 genomes data (chr22) located in `data/chr22` with population labels `data/1KG.samples.gz`.

3. Briefly discuss potential utility of these types of results in GWAS.

## Optional

Read the original structure paper, Pritchard *et al.*.