# Assignment6 by Elizabeth Nguyen

On Blackboard is a file called: "abbgen1k.csv". This file is a subset of the 1,000 Genomes Project for chromosome 22. The format is the same as we discussed in class: rows are SNPs, columns 1 to 9 (R starts with 1) are details about the SNPs, columns 10 to 90 are unrelated individuals from Europe, and columns 91 to 179 are unrelated individuals from Africa.

Overarching Prompts/Questions: Compute the 95% confidence intervals for the average pairwise diversity within the European samples. Also compute this within the African samples, and between the European and African samples. Due to the out-of-Africa hypothesis, we expect that the average pairwise diversity within Europe will be less than the other two comparisons. Do these three confidence intervals overlap?

```
# read in Assigment and test data
testdata <- read.csv("shorttesteg.csv",stringsAsFactors=F)
abbgen1k <- read.csv("abbgen1k.csv",stringsAsFactors=F)
```

To compute these confidence intervals, first write an R function to compute the average pairwise diversity for n randomly chosen pairs of haplotypes. These pairs of haplotypes can both be chosen from one sample, or they can be chosen from two different samples. The inputs to this function are a dataframe with the format in "abbgen1k.csv", a vector of column numbers for individuals in the first sample, a vector of column numbers for individuals in the second sample (the two samples may be the same or different, if the samples are the same just enter the same vector twice), and a number n. The function will: (1) randomly pick an individual from the first sample and randomly pick an individual from the second sample, (2) for each of these individuals, randomly pick the left or right haplotype, (3) for these two haplotypes count the mean number of SNPs where they differ, (4) repeat this procedure n times (getting different pairs of haplotypes almost every time) and take the average of the mean number of differences for the n randomly chosen pairs to get the average pairwise diversity.

---

Function to return only data for one chromosome, for a given individual

```
library(stringr)  # read in stringr library to implement regular expressions
```

```
## Warning: package 'stringr' was built under R version 3.6.3
```

```
oneChrom <- function(dataframe, individual, side){
  # from 'dataframe', returns only data for one chromosome ('side'), for one 'individual'
  geno <- dataframe[, individual]  # vector of all SNP data for 'individual'
  temp <- str_split(geno,"\\|",simplify=T)  # matrix of all SNP data splitting by chrosomsome
  hap <- temp[, side]  # vector of all SNP for one chromosome ('side')
  return(hap)
}
```

Test 'oneChrom' function on test data.

```
print(testdata[,11])  # SNP data for individual 11, from 'testdata'
```

```
##  [1] "0|0" "0|0" "0|0" "0|0" "0|0" "1|0" "0|0" "0|1" "1|0" "0|0"
```

```
print(oneChrom(testdata, 11, 1))  # returns only first chromosome data for individual 11
```

```
##  [1] "0" "0" "0" "0" "0" "1" "0" "0" "1" "0"
```

Function to count number of SNPs where 2 individuals differ

```
pairDiv <- function(dataframe, s1, s2){
  # from 'dataframe', returns number of SNPs where a random chromosome for a random individual from 's1
  i1 <- sample(s1, 1)  # selects one individual from sample1 's1'
  i2 <- sample(s2[s2 != i1], 1)  # selects one individual from sample2 's2'; exclude individual 'i1' in
  i1side <- sample((1:2), 1)  # randomly selects left or right haplotype, using 1 or 2
  i2side <- sample((1:2), 1)
  hap1 <- oneChrom(dataframe, i1, i1side)  # vector of all SNP for one chromosome
  hap2 <- oneChrom(dataframe, i2, i2side)
  m = mean(hap1 != hap2)  # fraction of SNPs for which the hap1 and hap2 differ
#  print(hap1) # prints hap1 & 2 for testing purposes
#  print(hap2)
  return(m)
}
```

Vectors of column numbers for individuals in test data

```
sampE <- c(10:14)
sampA <- c(15:19)
```

Test 'pairDiv' function on test data.

```
test1 <- pairDiv(testdata, sampE, sampE)
test1
```

```
## [1] 0.3
```

```
test2 <- pairDiv(testdata, sampE, sampA)
test2
```

```
## [1] 0.3
```

Function to repeat 'pairDiv' (or any) function n-times to get 'average pairwise diversity'

```
avgPairDiv <- function(dataframe, s1, s2, n){
  # for a given 'dataframe', runs 'pairDiv' function 'n' times
  temp <- 1:n  # vector of pairwise diversity intially set equal to vector 1,2,3,.....n
  for (i in 1:n){
    temp[i] <- pairDiv(dataframe, s1, s2)
  }
#  print(temp)  # for testing purposes
  m <- mean(temp)  # average of all pairwise diverity scores to get 'average pairwise diversity'
  return(m)
}
```

Test 'avgPairDiv' function on test data.

```
test3 <- avgPairDiv(testdata, sampE, sampE, 10)
test3
```

```
## [1] 0.26
```

Function to repeat 'avgPairDiv' (or any) function rep-times to get ci% confidence interval.

```
repFun <- function(dataframe, ci, rep, FUN,...){
  # for a given 'dataframe', implement FUN 'rep' number of times and obtain confidence interval of 'ci'
  temp = 1:rep  # vector of average pairwise diversity intially set equal to vector 1,2,3,.....rep
  for (i in 1:rep){
    temp[i] <- FUN(dataframe,...)
```

```
  }
  temp = sort(temp) # sort temp vector
  spot = round(rep*(1-ci)/2) # the coordinate we need for the confidence interval
#  print(temp)  # for testing purposes
#  print(spot)  # for testing purposes
  return(list(apd=mean(temp), low=temp[spot], high=temp[rep-spot]))  # returns list with (1)average pai
}
```

Test 'repFun' on test data.

```
testAPD <- repFun(testdata, 0.95, 100, avgPairDiv, s1=sampE, s2=sampE, n=100)
testAPD
```

```
## $apd
## [1] 0.22906
##
## $low
## [1] 0.208
##
## $high
## [1] 0.244
```

Now that the function tests all seem to work, let's address the assignment questions: On Blackboard is a file called: "abbgen1k.csv". This file is a subset of the 1,000 Genomes Project for chromosome 22. The format is the same as we discussed in class: rows are SNPs, columns 1 to 9 (R starts with 1) are details about the SNPs, columns 10 to 90 are unrelated individuals from Europe, and columns 91 to 179 are unrelated individuals from Africa.

Time assessment - time in:

```
t1 = proc.time()
```

Compute the 95% confidence intervals for the average pairwise diversity within the European samples.

```
# Vectors of column numbers for European individuals
sampEuro <- c(10:90)

EuroAPD <- repFun(abbgen1k, 0.95, 100, avgPairDiv, s1=sampEuro, s2=sampEuro, n=100)
EuroAPD
```

```
## $apd
## [1] 0.09339444
##
## $low
## [1] 0.09043929
##
## $high
## [1] 0.0957906
```

Also compute this within the African samples,

```
sampAfr <- c(91:179)

AfrAPD <- repFun(abbgen1k, 0.95, 100, avgPairDiv, s1=sampAfr, s2=sampAfr, n=100)
AfrAPD
```

```
## $apd
## [1] 0.1133542
##
## $low
## [1] 0.1109879
##
## $high
## [1] 0.1152105
```

and between the European and African samples.

```
EuroAfrAPD <- repFun(abbgen1k, 0.95, 100, avgPairDiv, s1=sampEuro, s2=sampAfr, n=100)
EuroAfrAPD
```

```
## $apd
## [1] 0.1177228
##
## $low
## [1] 0.1154845
##
## $high
## [1] 0.1198096
```

Due to the out-of-Africa hypothesis, we expect that the average pairwise diversity within Europe will be less than the other two comparisons. Do these three confidence intervals overlap? These three confidence intervals

do not overlap. Europeans have a smaller nucleotdide diversity compared to Africans, while compairing nucleotdie diveristiy between Europeans and Africans has the largest nucleotide diversity.

Time assessment - time out:

```
t2 = proc.time()
t2-t1
```

```
##    user  system elapsed
##  413.70    0.06  413.83
```