Assignment5 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.

Prompt 1

It is possible that a SNP is polymorphic in the world-wide sample, but not polymorphic in a given population sample. Count the number of SNPs in "abbgen1k.csv" for which the ALT allele does NOT have frequency 0 in the sample of Europeans individuals. Also count this number for the sample of African individuals. Due to the out-of-Africa hypothesis, we expect that this number is greater for the African sample than the European sample. Is this what we observe?

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

Examine file format

```
# look across all columns
testdata[1,]
```

```
##   ï..CHROM      POS          ID REF ALT QUAL FILTER
## 1       22 16050115 rs587755077   G   A  100   PASS
##
## 1 AC=32;AF=0.00638978;AN=5008;NS=2504;DP=11468;EAS_AF=0;AMR_AF=0.0014;AFR_AF=0.0234;EUR_AF=0;SAS_AF=(
##   FORMAT  E1  E2  E3  E4  E5  A1  A2  A3  A4  A5
## 1     GT 0|0 0|0 0|0 0|0 0|0 1|0 0|1 1|1 0|0 0|0
```

```
# confirm columns 1:9 are SNPs details
testdata[1, 1:9]
```

```
##   ï..CHROM      POS          ID REF ALT QUAL FILTER
## 1       22 16050115 rs587755077   G   A  100   PASS
##
## 1 AC=32;AF=0.00638978;AN=5008;NS=2504;DP=11468;EAS_AF=0;AMR_AF=0.0014;AFR_AF=0.0234;EUR_AF=0;SAS_AF=(
##   FORMAT
## 1     GT
```

```
# confirm E columns
testdata[1, 10:14]
```

```
##    E1  E2  E3  E4  E5
## 1 0|0 0|0 0|0 0|0 0|0
```

```
# confirm A columns
testdata[1, 15:19]
```

```
##    A1  A2  A3  A4  A5
## 1 1|0 0|1 1|1 0|0 0|0
```

Subset testdata

```
testdataE <- testdata[, 10:14]
testdataE
```

```
##     E1  E2  E3  E4  E5
## 1  0|0 0|0 0|0 0|0 0|0
## 2  0|0 0|0 0|0 0|0 0|0
## 3  0|0 0|0 0|0 0|0 0|0
## 4  0|0 0|0 0|1 0|1 0|0
## 5  0|1 0|0 0|0 0|0 0|0
## 6  0|0 1|0 0|0 0|0 0|0
## 7  0|0 0|0 1|0 0|0 0|0
## 8  1|1 0|1 1|0 0|1 1|0
## 9  0|1 1|0 1|0 0|0 0|0
## 10 0|0 0|0 0|0 1|1 0|0
```

```
testdataA <- testdata[, 15:19]
testdataA
```

```
##     A1  A2  A3  A4  A5
## 1  1|0 0|1 1|1 0|0 0|0
## 2  0|1 0|0 0|0 0|1 0|0
## 3  0|0 0|0 0|0 0|0 1|0
## 4  0|1 1|1 1|0 0|0 0|0
## 5  0|0 1|0 0|0 1|0 0|0
## 6  0|1 0|0 0|0 0|0 0|0
## 7  1|0 0|1 0|0 0|0 0|0
## 8  0|0 0|0 0|0 0|0 0|0
## 9  0|0 0|1 1|0 0|1 1|1
## 10 0|0 0|1 1|0 0|0 1|1
```

Code for assessing if any SNPs present

```
rowSNP <- function(v) {
  cnt = sum(v == "1|0") + sum(v == "0|1") + 2*sum(v == "1|1")
  return(cnt > 0)  # returns TRUE or FALSE
}
```

Code for counting the number of SNPs in dataframe for which the ALT allele does NOT have frequency 0

```
cntSNPs <- function(x) {
  z = apply(x,1,rowSNP)
  # the 1 means apply to every row
  # is now a vector: the ith element is the answer for the ith row of x
  return(sum(z))
}
```

Test functions for sample data

```
cntSNPs(testdataA)
```

```
## [1] 9
```

```
cntSNPs(testdataE)
```

```
## [1] 7
```

The test code worked appropriately. Let's now tackle the assignment questions:

Count the number of SNPs in "abbgen1k.csv" for which the ALT allele does NOT have frequency 0 in the sample of Europeans individuals. Columns 10 to 90 are unrelated individuals from Europe.

```
abbEuro <- abbgen1k[, 10:90]  # subset European individuals
#dim(abbEuro)  # dimensions for European subset
cntSNPs(abbEuro)  # counts number of non-zero frequency SNPs for European individuals
```

```
## [1] 7959
```

Count this number for the sample of African individuals. Columns 91 to 179 are unrelated individuals from Africa.

```
abbAfr <- abbgen1k[, 91:179]  # subset Afridcan individuals
#dim(abbAfr)  # dimensions for African subset
cntSNPs(abbAfr)  # counts number of non-zero frequency SNPs for African individuals
```

```
## [1] 12705
```

Due to the out-of-Africa hypothesis, we expect that this number is greater for the African sample than the European sample. Is this what we observe? In accordance with the out-of-African hypothesis, the number of non-zero frequency SNPs is indeed greater for the African sample than the European sample.

Prompt 2

Generalize this to compute the allele frequency spectrum. The allele frequency spectrum is the number of SNPs for which there are k haplotypes that have the ALT allele in the sample. k ranges from 1 to the number of haplotypes in the sample (see slide 65 in the Population Genetics lecture slides). Use the built-in R function "barplot" to make two plots: the allele frequency spectrum for the European individuals and the allele frequency spectrum for the African individuals. Comment on the similarity and the differences between the plots.

Code for assessing frequency of ALT allele among haplotypes, across individuals

```
rowSNP <- function(v) {
  cnt = sum(v == "1|0") + sum(v == "0|1") + 2*sum(v == "1|1")  # counts each ALT allele across individu
  return(cnt)  # returns number of ALT alleles
}
```

Code for allele frequency count vector

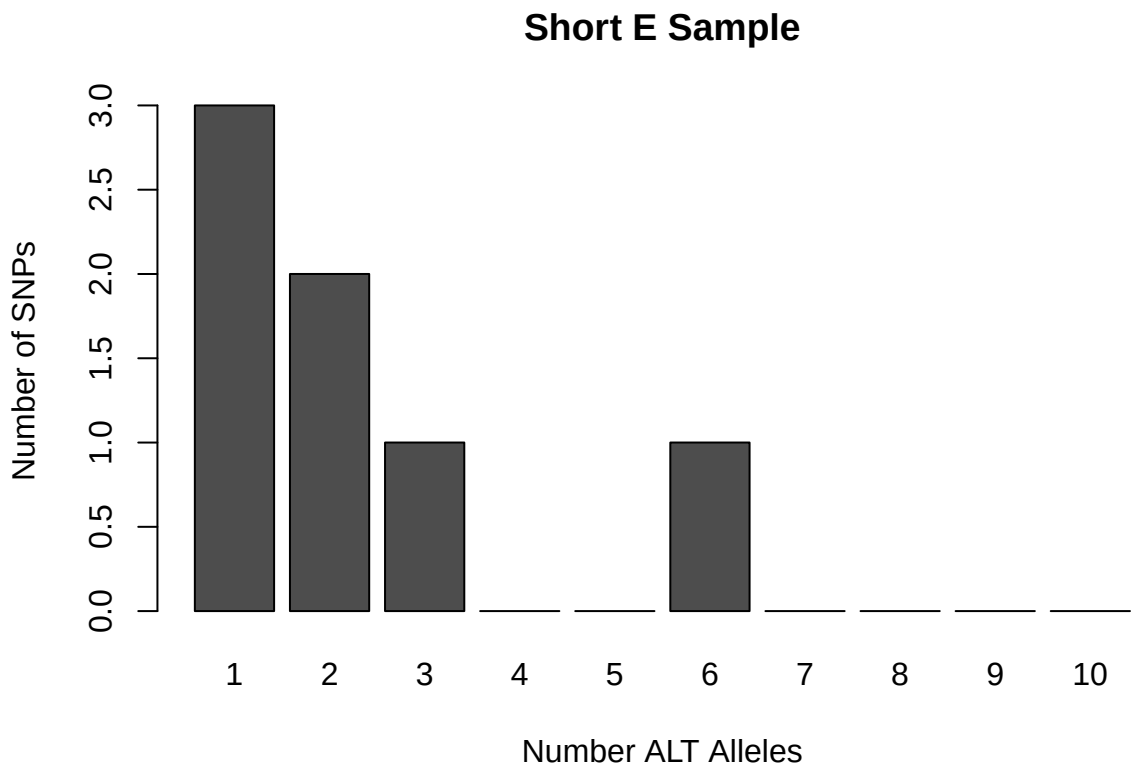
```
altFreq <- function(x) {
  z <- apply(x, 1, rowSNP)  # list of number of ALT allele for EACH SNP
  return(z)
}
```

Code for allele frequency matrix

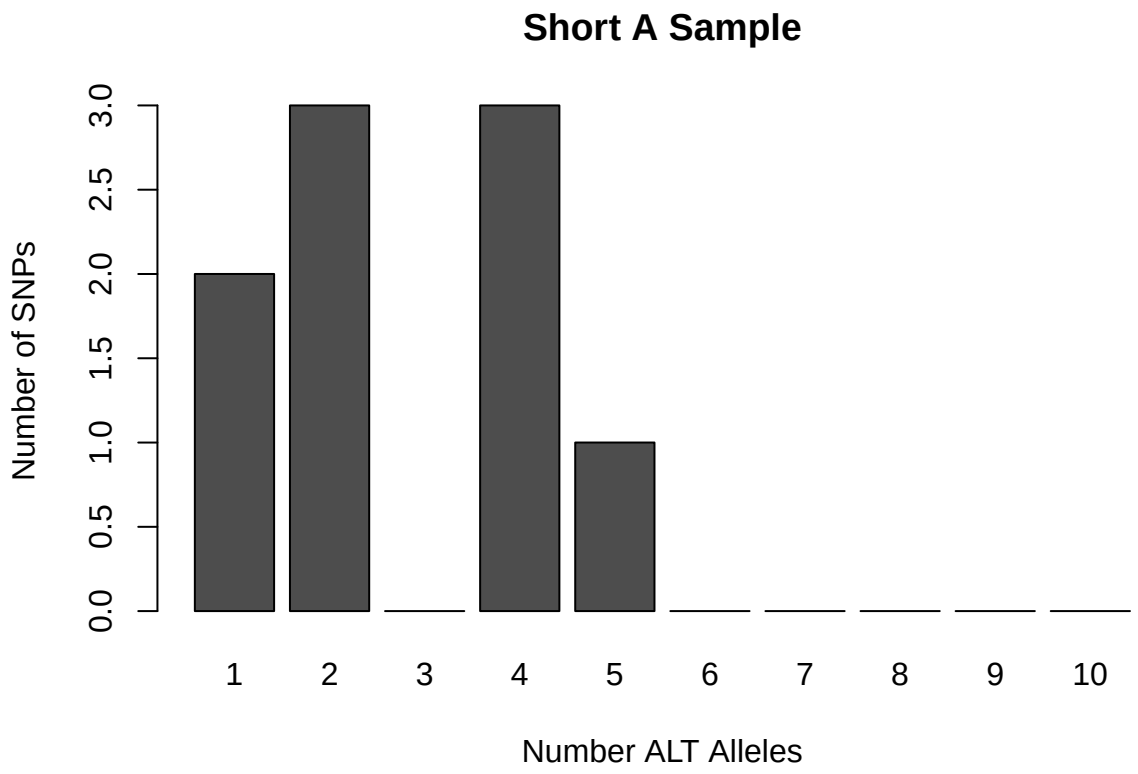
```
spectrum <- function(x) {
  n <- length(x[1,])
  m <- matrix(0, nrow = 1, ncol = 2*n)  # empty matrix for up to maximum number of ALT alleles for a gi
  listAltFreq <- altFreq(x)  # list of number of ALT allele for EACH SNP
  for (i in listAltFreq) {
    if (i > 0) {
      m[1,i] = m[1,i] + 1  # increment appropriate position in matrix for number of ALT allele for each
    }
  }
  colnames(m) <- c(1:(2*n))  # applies number header to matrix
  return(m)  # return allele frequency matrix
}
```

Barplots for Allele Frequency Spectrum

```
spectrumE <- spectrum(testdataE)
#spectrumE
barplot(spectrumE,
        xlab = "Number ALT Alleles",
        ylab = "Number of SNPs",
        main = "Short E Sample",
        )
```




```
spectrumA <- spectrum(testdataA)
#spectrumA
barplot(spectrumA,
        xlab = "Number ALT Alleles",
        ylab = "Number of SNPs",
        main = "Short A Sample",
        )
```



The test code worked appropriately. Let's now tackle the assignment questions:

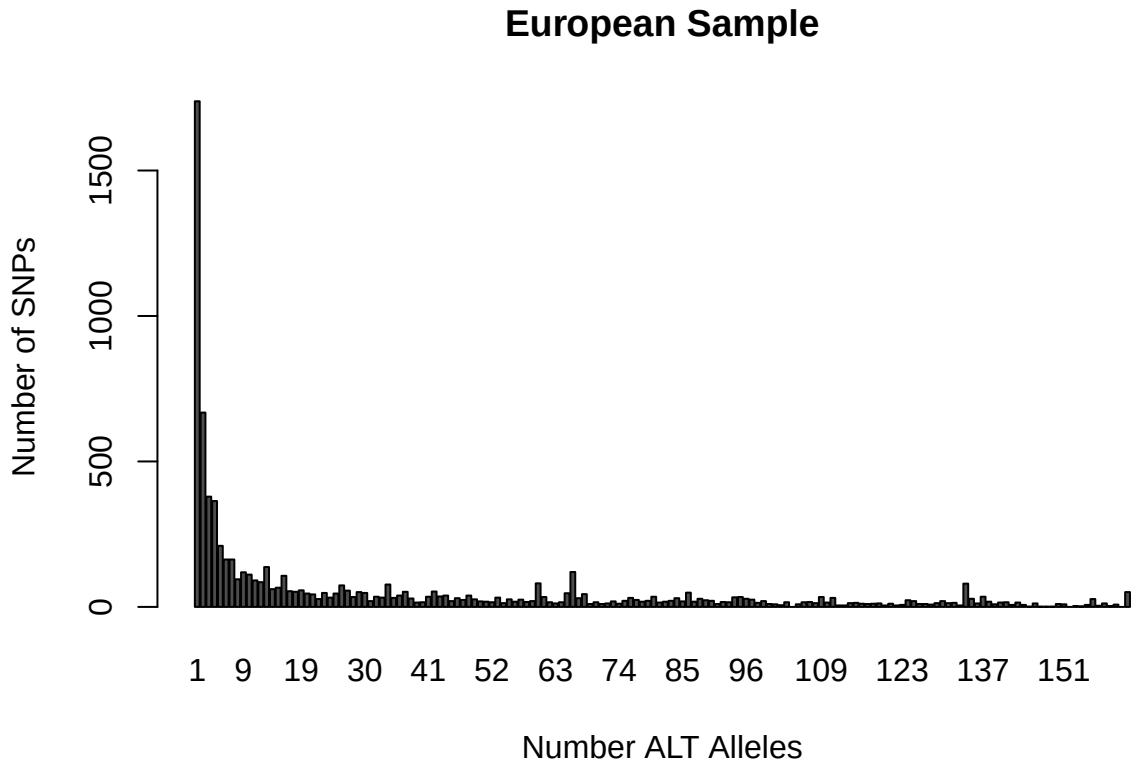
The allele frequency spectrum for the European individuals

```
spectrumEuro <- spectrum(abbEuro)
spectrumEuro
```

```
##         1   2   3   4   5   6   7  8   9  10 11 12  13 14 15  16 17 18 19 20 21
## [1,] 1738 668 379 364 210 163 163 95 119 111 91 85 137 61 66 107 54 52 57 46 43
##      22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
## [1,] 27 48 32 46 74 56 34 51 48 20 35 32 77 31 39 52 29 15 16 35 53 36 39 20 30
##      47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65  66 67 68 69 70
## [1,] 24 39 26 19 18 16 32 13 26 18 25 17 20 81 34 16 12 16 47 120 30 44 10 16
##      71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95
## [1,] 10 12 19 11 21 31 24 18 21 35 15 18 21 30 19 49 18 28 23 21 10 17 16 33 34
##      96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114
## [1,] 28 25 14 20  10   9   6  16   0   9  16  17  13  34  15  31   5   5  13
##      115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132
## [1,]  14  11  10  11  12   5  11   5   7  23  20  10  10   8  13  20  13  14
##      133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150
## [1,]   5  80  28  12  35  18   9  15  16   7  15   7   1  12   1   1   1  10
##      151 152 153 154 155 156 157 158 159 160 161 162
## [1,]   9   0   3   2   7  27   4  12   3   8   0  51
```

```
barplot(spectrumEuro,
        xlab = "Number ALT Alleles",
        ylab = "Number of SNPs",
        main = "European Sample",
```

```
)
```

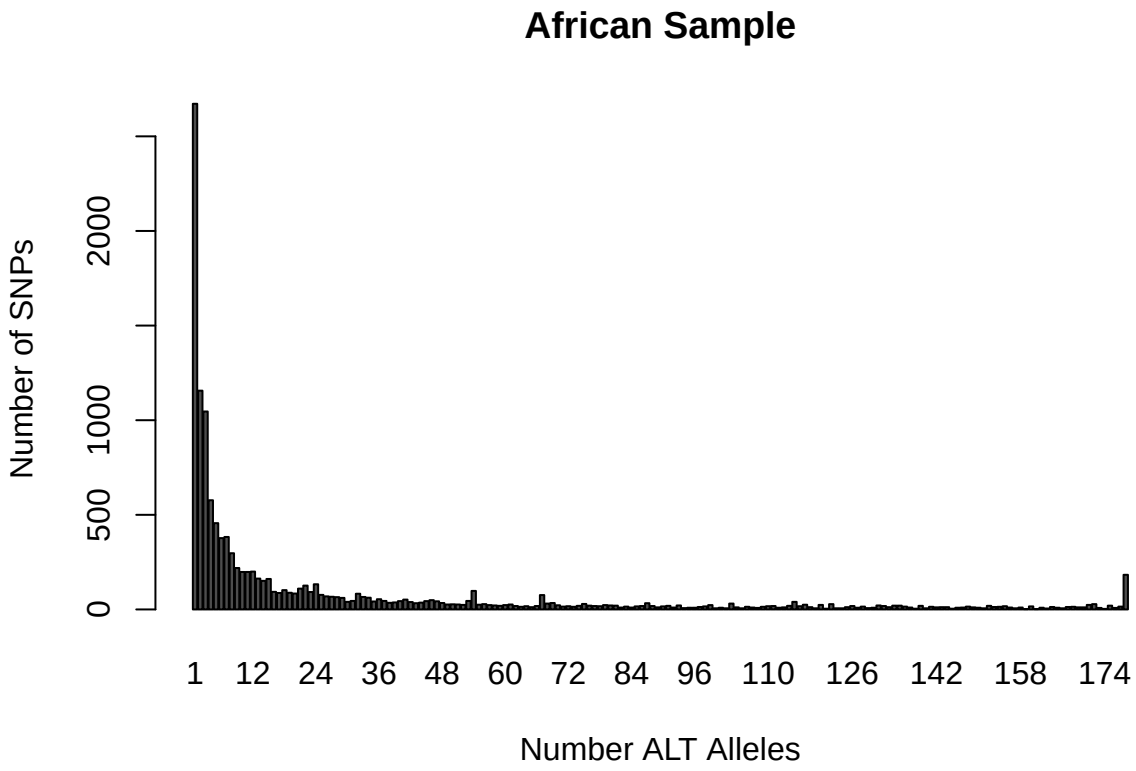


The allele frequency spectrum for the African individuals.

```
spectrumAfr <- spectrum(abbAfr)
spectrumAfr
```

```
##         1    2    3   4   5   6   7   8   9  10  11  12  13  14  15 16 17  18
## [1,] 2672 1156 1046 577 456 377 383 297 219 198 198 200 163 151 161 93 87 102
##      19 20  21  22 23  24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
## [1,] 88 84 110 126 92 133 77 69 67 65 61 40 45 83 66 62 42 54 45 35 37 44 52 39
##      43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67
## [1,] 33 36 44 49 43 34 26 27 26 24 45 98 25 28 23 21 19 23 26 18 14 17 12 18 76
##      68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92
## [1,] 31 34 22 15 17 14 19 29 20 18 17 23 21 20 10 15 10 16 18 33 18 11 16 19 10
##      93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112
## [1,] 21  7  9  9 13 16 23   6   9   6  31  11   6  14  10   8  14  17  18   9
##      113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130
## [1,]  11  19  40  17  25  12   6  24   4  28   6   6  12  18   9  15   7   9
##      131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148
## [1,]  21  18  12  20  20  15  10   4  19   6  14  11  12  12   4   9  10  15
##      149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166
## [1,]  11   9   6  19  13  14  17  10   6  10   2  16   2   9   5  13   9   6
##      167 168 169 170 171 172 173 174 175 176 177 178
## [1,]  13  14  11  11  24  28   8   3  20   8  15 183
```

```
barplot(spectrumAfr,
        xlab = "Number ALT Alleles",
        ylab = "Number of SNPs",
        main = "African Sample",
        )
```



Comment on the similarity and the differences between the plots. Even though the number of non-zero frequency SNPs is greater for the African sample than the European sample, their allele frequency spectrums look pretty similar. Both European and African samples have many SNPs with only one ALT allele across all individuals sampled. And for both groups, most SNPs have only a few ALT alleles across individuals sampled. However, both groups have a handful of SNPs with a fairly high number of ALT alleles across individuals sampled. The European sample includes 81 individuals, while the AFrican sample includes 89 individuals. Both samples have a handful of SNPs that have two ALT alleles across all individuals. The number of possible ALT allels for the African sample is greater than that of the European sample because the African sample includes more people.