BSG - Homework4: Haplotype estimation

Marti Cardoso, Pau Ferrer

December 5th, 2018

1. Myoglobin is an oxygen-binding protein found in muscle tissue. The protein is encoded by the MB gene, which resides on the long arm of chromosome 22. The file MB.rda contains genotype information of unrelated individuals for a set of SNPs in the MB gene. The file contains genotype information in object Y. Load this data into the R environment.

```r
load("MB.rda")
```

2.(1p) How many individuals and how many SNPs are there in the database? What percentage of the data is missing?

```r
(n <- nrow(Y)) #Individuals
```

```
## [1] 139
```

```r
(m <- ncol(Y)) #SNPs
```

```
## [1] 28
```

```r
#Percentage of missing
sum(is.na(Y))/(n*m)*100
```

```
## [1] 39.54265
```

```r
#Percentage of individuals with no missings
sum(apply(Y,1,function(l) sum(is.na(l))==0))/n*100
```

```
## [1] 25.17986
```

```r
#Percentage of SNPs with no missings
sum(apply(Y,2,function(l) sum(is.na(l))==0))/m*100
```

```
## [1] 3.571429
```

There are a total of 139 individuals (number of rows) and 28 SNPs (number of rows). 39.54% of the data is missing data. There is only 1 SNPs (3.57%) that do not have any missing values and 35 individuals that do not have missings (25.18%).

3. (1p) Assuming all SNPs are bi-allelic, how many haplotypes can theoretically be found for this data set?

```r
(nHaploPossible <- 2^m)
```

```
## [1] 268435456
```

Theoretically, with 28 SNPs, we can find 268.435.456 haplotypes.

4. (1p) Estimate haplotype frequencies using the haplo.stats package (set the minimum posterior probability to 0.001). How many haplotypes do you find? List the haplotypes and their estimated probabilities. Which haplotype is the most common?

```
Geno <- cbind(substr(Y[,1],1,1),substr(Y[,1],2,2))
for(i in 2:m) {
  Geno <- cbind(Geno,substr(Y[,i],1,1),substr(Y[,i],2,2))
}
Haplo.Res <- haplo.em(Geno,locus.label=colnames(Y),control=haplo.em.control(min.posterior=0.001))
Haplo.Res
```

```
## ==========================================================================
##                                 Haplotypes
## ==========================================================================
##   rs1056680 rs2076141 rs16995880 rs16995883 rs5999890 rs11705485 rs5750130
## 1         C         A          A          C         C          A         G
## 2         T         A          A          C         C          A         G
## 3         T         A          A          C         C          A         G
## 4         T         A          A          C         C          A         G
## 5         T         A          G          G         C          A         G
## 6         T         C          A          C         C          A         C
##   rs2899254 rs5755790 rs7292 rs7293 rs2283962 rs916230 rs2179870 rs5750132
## 1         T         A      G      C         G        C         C         A
## 2         C         G      G      C         A        A         T         G
## 3         C         G      G      C         A        A         T         G
## 4         T         A      G      C         A        A         T         G
## 5         T         A      G      C         G        C         C         A
## 6         T         A      A      T         G        C         C         A
##   rs5755793 rs5755794 rs5755795 rs5755798 rs5755799 rs5750135 rs1997882
## 1         A         C         C         C         C         G         C
## 2         G         T         C         T         C         A         T
## 3         G         T         C         T         C         G         C
## 4         A         C         C         T         C         G         C
## 5         A         C         C         C         C         G         C
## 6         A         C         A         C         G         G         C
##   rs5750136 rs8136856 rs8140868 rs8140754 rs8141048 rs13056550 hap.freq
## 1         C         C         C         A         C          T  0.02158
## 2         T         C         C         A         C          T  0.05665
## 3         T         C         C         A         C          T  0.01511
## 4         T         C         C         A         C          T  0.00378
## 5         C         C         C         A         C          T  0.17626
## 6         C         C         C         A         C          T  0.72662
## ==========================================================================
##                                   Details
## ==========================================================================
## lnlike =  -197.5975
## lr stat for no LD =  1868.693 , df =  -16 , p-val =  NA
```

```
(Haplo.Res$nreps)
```

```
## indx.subj
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##   1   1   1   3   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
## 127 128 129 130 131 132 133 134 135 136 137 138 139
##   1   1   1   1   1   1   1   1   1   1   1   1   1
```

It is found 6 different haplotypes:

- CAACCAGTAGCGCCAACCCCGCCCCACT (freq 0.02158)
- TAACCAGCGGCAATGGTCTCATTCCACT (freq 0.05665)
- TAACCAGCGGCAATGGTCTCGCTCCACT (freq 0.01511)
- TAACCAGTAGCAATGACCTCGCTCCACT (freq 0.00378)
- TAGGCAGTAGCGCCAACCCCGCCCCACT (freq 0.17626)
- TCACCACTAATGCCAACACGGCCCCACT (freq 0.72662)

So, the most common haplotype has frequency 0.72662 and it is: TCACCACTAATGCCAACACGGCCCCACT

5. (2p) Is the haplotypic constitution of any of the individuals in the database ambiguous or uncertain? If so, for which individuals? What is, in case, the most likely haplotypic constitution of any possibly uncertain individuals?

The individuals with ambiguous haplotypic constitutions will be those that are double-side heterozygous.

```
isAmbiguousOrUncertain <- function(x){
  # If some is NA, the haplotypic is uncertain
  isUncertain <- any(is.na(x))
  # If it is double-side heterozygous, the haplotypic is ambiguous
  isAmbiguous <- sum( sapply(x, function(y) substr(y,1,1)!=substr(y,2,2)) )> 1
  return(isUncertain || isAmbiguous)
}

Y.isAmbiguousOrUncertain <- apply(Y, 1, isAmbiguousOrUncertain)
Y.ambiguousOrUncertain <- Y[Y.isAmbiguousOrUncertain,]
#Number of ambiguous or uncertain and names
nrow(Y.ambiguousOrUncertain)
```

```
## [1] 121
```

```
row.names(Y.ambiguousOrUncertain)
```

```
##   [1] "NA18524" "NA18525" "NA18526" "NA18527" "NA18528" "NA18531" "NA18532"
##   [8] "NA18533" "NA18534" "NA18536" "NA18537" "NA18538" "NA18539" "NA18540"
##  [15] "NA18541" "NA18543" "NA18544" "NA18545" "NA18546" "NA18547" "NA18548"
##  [22] "NA18553" "NA18557" "NA18559" "NA18560" "NA18564" "NA18567" "NA18568"
##  [29] "NA18569" "NA18570" "NA18572" "NA18573" "NA18576" "NA18579" "NA18580"
##  [36] "NA18583" "NA18591" "NA18592" "NA18593" "NA18594" "NA18595" "NA18596"
##  [43] "NA18597" "NA18599" "NA18602" "NA18608" "NA18610" "NA18611" "NA18613"
##  [50] "NA18614" "NA18615" "NA18616" "NA18617" "NA18618" "NA18619" "NA18621"
##  [57] "NA18622" "NA18623" "NA18624" "NA18626" "NA18627" "NA18628" "NA18629"
##  [64] "NA18630" "NA18631" "NA18632" "NA18633" "NA18634" "NA18635" "NA18636"
##  [71] "NA18637" "NA18638" "NA18639" "NA18640" "NA18641" "NA18642" "NA18643"
##  [78] "NA18644" "NA18645" "NA18647" "NA18648" "NA18649" "NA18739" "NA18740"
##  [85] "NA18741" "NA18742" "NA18743" "NA18745" "NA18747" "NA18748" "NA18749"
##  [92] "NA18750" "NA18751" "NA18752" "NA18755" "NA18757" "NA18758" "NA18759"
##  [99] "NA18760" "NA18761" "NA18762" "NA18763" "NA18765" "NA18769" "NA18771"
## [106] "NA18772" "NA18773" "NA18774" "NA18777" "NA18778" "NA18779" "NA18780"
## [113] "NA18783" "NA18784" "NA18785" "NA18787" "NA18790" "NA18792" "NA18794"
## [120] "NA18795" "NA18798"
```

```r
#Set NA to '--'
Y.ambiguousOrUncertain[is.na(Y.ambiguousOrUncertain)] <- "--"
Y.ambiguousOrUncertain.str = apply(Y.ambiguousOrUncertain, 1, function(v) paste(v,collapse = ""))
# Compute the most common haplotypic consitituion.
ambiguousOrUncertain.count = table(Y.ambiguousOrUncertain.str)
ambiguousOrUncertain.count[ambiguousOrUncertain.count == max(ambiguousOrUncertain.count)]
```

```
## TT--AA--CC----TTAAAATT--------AACC----GGGGCC------------
##                                                       50
```

There are a lot of ambiguous or uncertain individuals, a total of 121. The individuals are shown in the previous R print text.

The most likely haplotypic constitution of the uncertain individuals is 'TT -- AA -- CC -- -- TT AA AA TT -- -- -- -- AA CC -- -- GG GG CC -- -- -- -- -- --' (where '--' is NA), this constitution appears 50 times in the dataset.

6. (1p) Suppose we would delete SNP rs5999890 from the database prior to haplotype estimation. Would this affect the results obtained? Justify your answer. Delete this SNP from the database and estimate again the haplotype frequencies. List the haplotypes and their estimated frequencies.

```r
unique(Y[,'rs5999890'])
```

```
## [1] "CC"
```

```r
aux <- Y[,-5]
Geno <- cbind(substr(aux[,1],1,1),substr(aux[,1],2,2))
for(i in 2:(m-1)) {
  Geno <- cbind(Geno,substr(aux[,i],1,1),substr(aux[,i],2,2))
}
Haplo.Res2 <- haplo.em(Geno,locus.label=colnames(aux),control=haplo.em.control(min.posterior=0.001))
Haplo.Res2
```

```
## ===========================================================================
##                                  Haplotypes
## ===========================================================================
##   rs1056680 rs2076141 rs16995880 rs16995883 rs11705485 rs5750130 rs2899254
## 1         C         A          A          C          A         G         T
## 2         T         A          A          C          A         G         C
## 3         T         A          A          C          A         G         C
## 4         T         A          A          C          A         G         T
## 5         T         A          G          G          A         G         T
## 6         T         C          A          C          A         C         T
##   rs5755790 rs7292 rs7293 rs2283962 rs916230 rs2179870 rs5750132 rs5755793
## 1         A      G      C         G        C         C         A         A
## 2         G      G      C         A        A         T         G         G
## 3         G      G      C         A        A         T         G         G
## 4         A      G      C         A        A         T         G         A
## 5         A      G      C         G        C         C         A         A
## 6         A      A      T         G        C         C         A         A
##   rs5755794 rs5755795 rs5755798 rs5755799 rs5750135 rs1997882 rs5750136
## 1         C         C         C         C         G         C         C
## 2         T         C         T         C         A         T         T
## 3         T         C         T         C         G         C         T
## 4         C         C         T         C         G         C         T
## 5         C         C         C         C         G         C         C
## 6         C         A         C         G         G         C         C
##   rs8136856 rs8140868 rs8140754 rs8141048 rs13056550 hap.freq
## 1         C         C         A         C          T  0.02158
## 2         C         C         A         C          T  0.05665
## 3         C         C         A         C          T  0.01511
## 4         C         C         A         C          T  0.00378
## 5         C         C         A         C          T  0.17626
## 6         C         C         A         C          T  0.72662
## ===========================================================================
##                                    Details
## ===========================================================================
## lnlike =  -197.5975
## lr stat for no LD =  1868.693 , df =  -16 , p-val =  NA
```

The SNP rs5999890 is monomorphic and homozygous (CC), so the haplotype estimation will not change. Looking at the results, the estimation is the same and the frequencies have also not changed.

7. (1p) Individual NA18525 has missing values for several variants. Does this mean that that haplo.em cannot estimate the constitution of this individual? Investigate the haplotypic constitution of this individual, and what implication the estimation has for his/her missing values

No, the EM algorithm completes the missing data iteratively by using the estimated population parameters. The algorithm can handle missing values.

```
Haplo.Res$nreps[2]
```

```
## 2
## 1
```

```
Haplo.Res$hap1code[2]
```

```
## [1] 6
```

```
Haplo.Res$hap2code[2]
```

```
## [1] 5
```

```
individual <- Y[2,]
Haplo.Res$haplotype[6,]
```

```
##   rs1056680 rs2076141 rs16995880 rs16995883 rs5999890 rs11705485 rs5750130
## 6         T         C          A          C         C          A         C
##   rs2899254 rs5755790 rs7292 rs7293 rs2283962 rs916230 rs2179870 rs5750132
## 6         T         A      A      T         G        C         C         A
##   rs5755793 rs5755794 rs5755795 rs5755798 rs5755799 rs5750135 rs1997882
## 6         A         C         A         C         G         G         C
##   rs5750136 rs8136856 rs8140868 rs8140754 rs8141048 rs13056550
## 6         C         C         C         A         C          T
```

```
pos <- which(is.na(individual))
estimated.allels.missings <- Haplo.Res$haplotype[6,pos]
estimated.allels.missings
```

```
##   rs2076141 rs16995883 rs11705485 rs5750130 rs2283962 rs916230 rs2179870
## 6         C          C          A         C         G        C         C
##   rs5750132 rs5755795 rs5755798 rs5750136 rs8136856 rs8140868 rs8140754
## 6         A         A         C         C         C         C         A
##   rs8141048 rs13056550
## 6         C          T
```

This individual haplotypic constitution is estimated as the haplotype pair (6,5). We will obtain an imputation of the missing values. Indeed as haplotype pair estimated is (6,5) the missing variants will be imputed as monomorphic and genotype corresponding to the genotypes obtained in the position.

8. (2p) We could consider the newly created haplotypes as the alleles of a new locus. Which is, under the assumption of Hardy-Weinberg equilibrium, the most likely genotype at this new locus? What is the probability of this genotype? Which genotype is the second most likely?

```
#Compute table of frequencies
alleles.count <- table(unlist(t(Haplo.Res$haplotype)))
alleles.prov <- alleles.count/sum(alleles.count)

alleles.name <- names(alleles.prov)

genotypes <- c(paste(alleles.name[1], alleles.name[1], sep=""),
               paste(alleles.name[1], alleles.name[2], sep=""),
               paste(alleles.name[1], alleles.name[3], sep=""),
               paste(alleles.name[1], alleles.name[4], sep=""),
               paste(alleles.name[2], alleles.name[2], sep=""),
               paste(alleles.name[2], alleles.name[3], sep=""),
               paste(alleles.name[2], alleles.name[4], sep=""),
               paste(alleles.name[3], alleles.name[3], sep=""),
               paste(alleles.name[3], alleles.name[4], sep=""),
               paste(alleles.name[4], alleles.name[4], sep=""))

result <- c(  alleles.prov[1]*alleles.prov[1],
            2*alleles.prov[1]*alleles.prov[2],
            2*alleles.prov[1]*alleles.prov[3],
            2*alleles.prov[1]*alleles.prov[4],
              alleles.prov[2]*alleles.prov[2],
            2*alleles.prov[2]*alleles.prov[3],
            2*alleles.prov[2]*alleles.prov[4],
              alleles.prov[3]*alleles.prov[3],
            2*alleles.prov[3]*alleles.prov[4],
              alleles.prov[4]*alleles.prov[4])
```

```
names(result) <- genotypes
sort(result,decreasing=TRUE)
```

```
##         AC         CC         CG         CT         AG         AT 
## 0.20833333 0.17361111 0.13888889 0.13888889 0.08333333 0.08333333 
##         AA         GT         GG         TT 
## 0.06250000 0.05555556 0.02777778 0.02777778 
```

The most likely genotype at this new locus is the AC genotype, with a probability of 0.20833, and the second most likely genotype is CC with a probability of 0.17361.

9. (1p) Simulate a set of independent markers using the multinomial distribution (R function rmultinom) that mimicks the Myoglobin data in terms of sample size, number of SNPs and minor allele frequencies, assuming HardyWeinberg equilibrium (that is, simulate the markers with multinomial probabilities p2; 2pq and q2, where p is the observed minor allele frequency) Create haplotypes on the basis of the simulated data. Do you find the same number of haplotypes? Can you explain the difference?

```
artificial.snp <- function(aux.y){
  g1 <- genotype(aux.y, sep = "")
  aux <- summary(g1)
  #if monomorphic
  if (aux$allele.freq[1,2] == 1.0){
    return(rep(paste(aux$allele.names[1], aux$allele.names[1], sep=""), n))
  }
  p <- max(aux$allele.freq[c(1,2),2])
  q <- min(aux$allele.freq[c(1,2),2])
  genotypes <- c(paste(aux$allele.names[1], aux$allele.names[1], sep=""),
                 paste(aux$allele.names[1], aux$allele.names[2], sep=""),
                 paste(aux$allele.names[2], aux$allele.names[2], sep=""))

  res <- rmultinom(n,1, prob =  c(p^2, 2*p*q, q^2))
  snp <- apply(res, 2, function(x) genotypes[which(x == 1)])
  return(snp)
}

artificial.snps <- apply(Y, 2, artificial.snp)
Geno <- cbind(substr(artificial.snps[,1],1,1),substr(artificial.snps[,1],2,2))
for(i in 2:m) {
  Geno <- cbind(Geno,substr(artificial.snps[,i],1,1),substr(artificial.snps[,i],2,2))
}
Haplo.Res.art <- haplo.em(Geno,control=haplo.em.control(min.posterior=0.001))
dim(Haplo.Res.art$haplotype)
```

```
## [1] 158  28
```

```
dim(Haplo.Res$haplotype)
```

```
## [1]   6 28
```

We have obtained more different haplotypes (154). This result is different maybe because the SNPs were not independent, or maybe the missing values made the variants not follow the HWE law.