Assignment 4

Krishna Kalyan

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.

```
## [1] "Y"
```

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

```
## [1] "Number of Individuals  139"
```

```
## [1] "Number of SNPs  28"
```

```
## [1] "Percentage of missing data  39.5426515930113"
```

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

```
2^ncol(X)
```

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

4. 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?

```
## [1] "Number of Haplotypes  6"
```

```
## [1] "Hapltypes and their probabilities "
```

```
## ================================================================================
##                                  Haplotypes
## ================================================================================
##   loc-1 loc-2 loc-3 loc-4 loc-5 loc-6 loc-7 loc-8 loc-9 loc-10 loc-11
## 1     C     A     A     C     C     A     G     T     A      G      C
## 2     T     A     A     C     C     A     G     C     G      G      C
## 3     T     A     A     C     C     A     G     C     G      G      C
## 4     T     A     A     C     C     A     G     T     A      G      C
## 5     T     A     G     G     C     A     G     T     A      G      C
## 6     T     C     A     C     C     A     C     T     A      A      T
##   loc-12 loc-13 loc-14 loc-15 loc-16 loc-17 loc-18 loc-19 loc-20 loc-21
## 1      G      C      C      A      A      C      C      C      C      G
## 2      A      A      T      G      G      T      C      T      C      A
## 3      A      A      T      G      G      T      C      T      C      G
## 4      A      A      T      G      A      C      C      T      C      G
## 5      G      C      C      A      A      C      C      C      C      G
## 6      G      C      C      A      A      C      A      C      G      G
##   loc-22 loc-23 loc-24 loc-25 loc-26 loc-27 loc-28 hap.freq
## 1      C      C      C      C      A      C      T  0.02158
## 2      T      T      C      C      A      C      T  0.05665
## 3      C      T      C      C      A      C      T  0.01511
## 4      C      T      C      C      A      C      T  0.00378
## 5      C      C      C      C      A      C      T  0.17626
## 6      C      C      C      C      A      C      T  0.72662
## ================================================================================
##                                    Details
## ================================================================================
## lnlike =  -197.5975
## lr stat for no LD =  1868.693 , df =  -16 , p-val =  NA
```

```
## [1] "Common Haplotypes"
```

```
##  [1] "T" "C" "A" "C" "C" "A" "C" "T" "A" "A" "T" "G" "C" "C" "A" "A" "C"
## [18] "A" "C" "G" "G" "C" "C" "C" "C" "A" "C" "T"
```

```
## [1] "Maximum Probability "
```

```
## [1] 0.7266187
```

5. Is the haplotypic constitution of any of the individuals in the database ambiguous or uncertain? If so, for which individuals?

```
## [1] "Uncertain Individual  NA18547"
```

	Index	Hap1code	Hap2code
22	22	4	6
23	22	3	6
24	22	2	6

Indiviudal number 22 is ambigous in the database.

6. 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.

```
head(X[,'rs5999890'])
```

```
## NA18524 NA18525 NA18526 NA18527 NA18528 NA18529
##    "CC"    "CC"    "CC"    "CC"    "CC"    "CC"
```

We see that SNP rs5999890 is monomorpihic. Generally monomorphic SNPS do not provide any useful information.

```
X_rem <- subset(X, select=-c(rs5999890))
Geno_rem = matrix(apply(X_rem, MARGIN = 2, function(x) return(c(substr(x,1,1),substr(x,
2,2)))),nrow(X_rem), 2*ncol(X_rem))
HapEM_rem = haplo.em(Geno_rem,control=haplo.em.control(min.posterior=.001))
print(HapEM_rem$hap.prob)
```

```
## [1] 0.021582734 0.056654669 0.015107916 0.003776984 0.176258993 0.726618705
```

```
HapEM_rem
```

```
## ================================================================================
##                                  Haplotypes
## ================================================================================
##   loc-1 loc-2 loc-3 loc-4 loc-5 loc-6 loc-7 loc-8 loc-9 loc-10 loc-11
## 1     C     A     A     C     A     G     T     A     G      C      G
## 2     T     A     A     C     A     G     C     G     G      C      A
## 3     T     A     A     C     A     G     C     G     G      C      A
## 4     T     A     A     C     A     G     T     A     G      C      A
## 5     T     A     G     G     A     G     T     A     G      C      G
## 6     T     C     A     C     A     C     T     A     A      T      G
##   loc-12 loc-13 loc-14 loc-15 loc-16 loc-17 loc-18 loc-19 loc-20 loc-21
## 1      C      C      A      A      C      C      C      C      G      C
## 2      A      T      G      G      T      C      T      C      A      T
## 3      A      T      G      G      T      C      T      C      G      C
## 4      A      T      G      A      C      C      T      C      G      C
## 5      C      C      A      A      C      C      C      C      G      C
## 6      C      C      A      A      C      A      C      G      G      C
##   loc-22 loc-23 loc-24 loc-25 loc-26 loc-27 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
```

We see same statistic with same probabilities when the monomorpic SNP is removed.

7. 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?

```
##              H 1          H 2          H 3          H 4         H 5
## H 1 0.0004658144 0.0024455253 0.0006521404 1.630351e-04 0.007608302
## H 2 0.0024455253 0.0032097516 0.0017118683 4.279672e-04 0.019971790
## H 3 0.0006521404 0.0017118683 0.0002282492 1.141246e-04 0.005325813
## H 4 0.0001630351 0.0004279672 0.0001141246 1.426558e-05 0.001331454
## H 5 0.0076083018 0.0199717899 0.0053258130 1.331454e-03 0.031067233
## H 6 0.0313648362 0.0823326849 0.0219553925 5.488849e-03 0.256146162
##             H 6
## H 1 0.031364836
## H 2 0.082332685
## H 3 0.021955392
## H 4 0.005488849
## H 5 0.256146162
## H 6 0.527974743
```

	H 1	H 2	Diplo freq
D 1	6	6	0.5279747
D 2	6	5	0.2561462

- If the combination is same Haplotypes, then the probability is equal to p(allele)^2
- If the combination is different Haplotypes then the probability is equal to 2.p(allele1).p(allele2) The most probable diplotype is (Haplotype 6, Haplotype 6) with a probability of 0.527 followed by (Haplotype 6, Haplotype 5) with a probability of 0.2561.

8. Simulate a set of independent markers using the HWData function of the HardyWeinberg package that mimicks the Myoglobin data in terms of sample size, number of SNPs and minor allele frequencies. Create haplotypes on the basis of the simulated data. Do you find the same number of haplotypes? Can you explain the difference?

```
## ========================================================================
##                                Haplotypes
## ========================================================================
##   loc-1 loc-2 loc-3 loc-4 loc-5 loc-6 loc-7 loc-8 loc-9 loc-10 loc-11
## 1     C     A     A     C     C     A     G     T     A      G      C
## 2     T     A     A     C     C     A     G     C     G      G      C
## 3     T     A     A     C     C     A     G     C     G      G      C
## 4     T     A     A     C     C     A     G     T     A      G      C
## 5     T     A     G     G     C     A     G     T     A      G      C
## 6     T     C     A     C     C     A     C     T     A      A      T
##   loc-12 loc-13 loc-14 loc-15 loc-16 loc-17 loc-18 loc-19 loc-20 loc-21
## 1      G      C      C      A      A      C      C      C      C      G
## 2      A      A      T      G      G      T      C      T      C      A
## 3      A      A      T      G      G      T      C      T      C      G
## 4      A      A      T      G      A      C      C      T      C      G
## 5      G      C      C      A      A      C      C      C      C      G
## 6      G      C      C      A      A      C      A      C      G      G
##   loc-22 loc-23 loc-24 loc-25 loc-26 loc-27 loc-28 hap.freq
## 1      C      C      C      C      A      C      T  0.02158
## 2      T      T      C      C      A      C      T  0.05665
## 3      C      T      C      C      A      C      T  0.01511
## 4      C      T      C      C      A      C      T  0.00378
## 5      C      C      C      C      A      C      T  0.17626
## 6      C      C      C      C      A      C      T  0.72662
## ========================================================================
##                                  Details
## ========================================================================
## lnlike =  -197.5975
## lr stat for no LD =  1868.693 , df =  -16 , p-val =  NA
```

I tried using Make count package. Had issues making it work as I did not have allels in the data. This out is same as the Myoglobin data. We see the same statistical values.