Bioinformatics and Statistical Genetics (taught by Jan Graffelman), HW4, Haplotype estimation

Sten-Oliver Salumaa, Denys Kovalenko

13 December 2017

Task1 (1p)

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.

Task2 (1p)

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

```
## [1] "Individual count:"
```

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

```
## [1] "SNP count:"
```

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

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

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

Task3 (1p)

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

```
## [1] "Theoretical number of haplotypes:"
```

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

Task4 (2p)

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?

```
##
## Attaching package: 'haplo.stats'
```

```
## The following object is masked from 'package:genetics':
##
##     locus
```

```
## [1] "Total haplotypes:"
```

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

```
## [1] "Most popular haplotype:"
```

```
##  [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] "Probability of the most popular haplotype:"
```

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

Task5 (2p)

Is the haplotypic constitution of any of the individuals in the database ambiguous or uncertain? If so, for which individuals? What is, in this case, the most likely haplotypic constitution of any possibly uncertain individuals?

```
## [1] "Individuals with ambiguous haplotypes:"
```

```
##  [1] "NA18524" "NA18525" "NA18527" "NA18528" "NA18532" "NA18534" "NA18536"
##  [8] "NA18537" "NA18538" "NA18540" "NA18541" "NA18545" "NA18547" "NA18548"
## [15] "NA18564" "NA18567" "NA18570" "NA18572" "NA18573" "NA18576" "NA18579"
## [22] "NA18583" "NA18591" "NA18592" "NA18593" "NA18594" "NA18595" "NA18597"
## [29] "NA18599" "NA18608" "NA18611" "NA18621" "NA18622" "NA18623" "NA18629"
## [36] "NA18631" "NA18632" "NA18633" "NA18636" "NA18637" "NA18640" "NA18641"
## [43] "NA18645" "NA18649" "NA18739" "NA18740" "NA18742" "NA18751" "NA18757"
## [50] "NA18758" "NA18761" "NA18762" "NA18765" "NA18774" "NA18777" "NA18779"
## [57] "NA18780" "NA18783" "NA18785" "NA18787" "NA18790" "NA18794"
```

```
## [1] "The most likely haplotypic constitution for these ambiguous 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] "And its probability:"
```

```
## [1] 0.4516129
```

Task 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.

It shouldn't affect results, cause SNP rs5999890 has only CC values, so it doesn't change possible haplotypes.

```
## [1] "Testing the hypothesis..."
```

```
## [1] "Total haplotypes:"
```

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

```
## [1] "Most popular haplotype:"
```

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

```
## [1] "Probability of the most popular haplotype:"
```

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

The test proves that actually we got same results.

```
## [1] "Listing haplotypes and their estimated frequencies:"
```

```
## ================================================================================
##                                 Haplotypes
## ================================================================================
##   snp1 snp2 snp3 snp4 snp5 snp6 snp7 snp8 snp9 snp10 snp11 snp12 snp13
## 1    C    A    A    C    A    G    T    A    G     C     G     C     C
## 2    T    A    A    C    A    G    C    G    G     C     A     A     T
## 3    T    A    A    C    A    G    C    G    G     C     A     A     T
## 4    T    A    A    C    A    G    T    A    G     C     A     A     T
## 5    T    A    G    G    A    G    T    A    G     C     G     C     C
## 6    T    C    A    C    A    C    T    A    A     T     G     C     C
##   snp14 snp15 snp16 snp17 snp18 snp19 snp20 snp21 snp22 snp23 snp24 snp25
## 1     A     A     C     C     C     C     G     C     C     C     C     A
## 2     G     G     T     C     T     C     A     T     T     C     C     A
## 3     G     G     T     C     T     C     G     C     T     C     C     A
## 4     G     A     C     C     T     C     G     C     T     C     C     A
## 5     A     A     C     C     C     C     G     C     C     C     C     A
## 6     A     A     C     A     C     G     G     C     C     C     C     A
##   snp26 snp27 hap.freq
## 1     C     T  0.02158
## 2     C     T  0.05665
## 3     C     T  0.01511
## 4     C     T  0.00378
## 5     C     T  0.17626
## 6     C     T  0.72662
## ================================================================================
##                                   Details
## ================================================================================
## lnlike =  -197.5975
## lr stat for no LD =  1868.693 , df =  -16 , p-val =  NA
```

Task 7 (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?

In this case we have 6-allele system. We can label fist haplotype as A allele, second as B .. 6th as F. In this case haplotypes' estimated frequencies would be corresponding allele frequencies. Then, we would need to build a table 6 by 6 of genotype frequencies. But as we know, homozygotes AiAi will have frequency p(i)^2, and all heterozygotes AiAj will have frequency 2p(i)*p(j).

```
## ================================================================================
##                                 Haplotypes
## ================================================================================
##   snp1 snp2 snp3 snp4 snp5 snp6 snp7 snp8 snp9 snp10 snp11 snp12 snp13
## 1    C    A    A    C    C    A    G    T    A     G     C     G     C
## 2    T    A    A    C    C    A    G    C    G     G     C     A     A
## 3    T    A    A    C    C    A    G    C    G     G     C     A     A
## 4    T    A    A    C    C    A    G    T    A     G     C     A     A
## 5    T    A    G    G    C    A    G    T    A     G     C     G     C
## 6    T    C    A    C    C    A    C    T    A     A     T     G     C
##   snp14 snp15 snp16 snp17 snp18 snp19 snp20 snp21 snp22 snp23 snp24 snp25
## 1     C     A     A     C     C     C     C     G     C     C     C     C
## 2     T     G     G     T     C     T     C     A     T     T     C     C
## 3     T     G     G     T     C     T     C     G     C     T     C     C
## 4     T     G     A     C     C     T     C     G     C     T     C     C
## 5     C     A     A     C     C     C     C     G     C     C     C     C
## 6     C     A     A     C     A     C     G     G     C     C     C     C
##   snp26 snp27 snp28 hap.freq
## 1     A     C     T  0.02158
## 2     A     C     T  0.05665
## 3     A     C     T  0.01511
## 4     A     C     T  0.00378
## 5     A     C     T  0.17626
## 6     A     C     T  0.72662
## ================================================================================
##                                   Details
## ================================================================================
## lnlike =  -197.5975
## lr stat for no LD =  1868.693 , df =  -16 , p-val =  NA
```

As we can see, 6th (F) haplotype has frequency of 0.72, second most frequent haplotype 5 (E) has 0.17 frequency. Obviously (as there is no other haplotype, for which frequency*2 would be more than 0.72), most frequent genotype would be FF with frequency

```
## [1] 0.5279747
```

second most frequent would be F with second most frequent allele - E.

```
## [1] 0.1280731
```

So we don’t see need to calculate all others. Most frequent genotype would be 66 (FF), second one - 65 (FE)

Task8 (1p)

Simulate a set of independent markers using the 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?

```
## [1] "The most likely haplotypic constitution for these ambiguous haplotypes (having allele
s 'A' and 'B':"
```

```
##  [1] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
## [18] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
```

```
## [1] "And its probability:"
```

```
## [1] 0.1832642
```

```
## [1] "Total haplotypes:"
```

```
## [1] 131
```

The results are very different. Previously under the same settings we found 6 probable haplotypes but now we see close to 130. This suggests that there is an association between different loci of myoglobin dataset (linkage disequilibrium) thus making some haplotypes more probable than others. In the simulated dataset we assumed no linkage (totally independent SNPs) and so we got more uniform distribution of haplotypes and their probabilities.