Haplotype estimation

Barbara Lugar, Giovanni Laganà

1. Apolipoprotein E (APOE) is a protein involved in Alzheimer's disease. The corresponding gene APOE has been mapped to chromosome 19. The file APOE.dat contains genotype information of unrelated individuals for a set of SNPs in this gene. Load this data into the R environment. APOE.zip contains the corresponding .bim, .fam and .bed files. You can use the .bim file to obtain information about the alleles of each polymorphism.

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

```
## [1] "Number of individuals: 107"
```

```
## [1] "Number of SNPs: 162"
```

```
## [1] "Percentage of missing data: 0%"
```

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

```
## [1] "Theoretical possible haplotypes: 5.84600654932361e+48"
```

This number is 2^m where m is the number of SNPs.

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 estimated probabilities in decreasing order. Which haplotype number is the most common?

```
## [1] "Number of observed haplotypes: 31"
```

```
##  [1] 0.3994864027 0.1308411215 0.0744773885 0.0684337821 0.0501816505
##  [6] 0.0467289720 0.0358634245 0.0351614435 0.0225689654 0.0204956857
## [11] 0.0186915888 0.0161150279 0.0086857776 0.0073507292 0.0046728972
## [16] 0.0046728972 0.0046728972 0.0046728972 0.0046728972 0.0046728972
## [21] 0.0046728972 0.0046728972 0.0046728972 0.0046728972 0.0040258310
## [26] 0.0033999885 0.0033021482 0.0028688002 0.0021370639 0.0016597156
## [31] 0.0007955206
```

```
## [1] "The most common haplotype is the one with ID: 27, with 0.39948640273045 frequency"
```

5. Is the haplotypic constitution of any of the individuals in the database ambiguous or uncertain? For how many? What is the most likely haplotypic constitution of individual NA20763? (identify the constitution by the corresponding haplotype numbers).

```
## [1] "There are 19 ambigous subjects:"
```

```
##  [1] "NA20504" "NA20518" "NA20522" "NA20524" "NA20529" "NA20531" "NA20536"
##  [8] "NA20544" "NA20586" "NA20756" "NA20763" "NA20764" "NA20766" "NA20792"
## [15] "NA20796" "NA20798" "NA20804" "NA20815" "NA20818"
```

```
## [1] "The constitution of the individual NA20763 is defined by the following information: "
```

```
##    id hap1code hap2code       post
## 69 59       24       21 0.01536613
## 70 59       18       28 0.96316787
## 71 59       25       20 0.02146601
```

```
## [1] "The most likely pair of possible haplotypes with the posterior probability 0.963167866211377"
```

```
## [1] "is: 18 and 28."
```

6. Suppose we would delete polymorphism rs374311741 from the database prior to haplotype estimation. Would this affect the results obtained? Justify your answer.

```
##   Count Proportion
## C   214          1
```

We notice that this genotype is monomorphic, therefore this cannot change the number of haplotypes. Since there is only one possible allele, it would be always a constant in the haplotypes.

7. Remove all genetic variants that have a minor allele frequency below 0.10 from the database, and re-run haplo.em. How does this affect the number of haplotypes?

```
## [1] "The number of haplotypes now is 8"
```

```
##           8           1           6           4           2           7
## 0.620635582 0.130841121 0.113009278 0.074766355 0.031850535 0.018691589
##           3           5
## 0.005532643 0.004672897
```

8. We could consider the newly created haplotypes in our last run of haplo.em as the alleles of a new superlocus. 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, and what is its probability?

The previous haplotypes probabilities are the new superlocus alleles probabilities, from which we can build the matrix of genotypes:

```
##         p1      p2      p3      p4      p5      p6      p7      p8
## p1 0.01712 0.00417 0.00072 0.00978 0.00061 0.01479 0.00245 0.08120
## p2 0.00417 0.00101 0.00018 0.00238 0.00015 0.00360 0.00060 0.01977
## p3 0.00072 0.00018 0.00003 0.00041 0.00003 0.00063 0.00010 0.00343
## p4 0.00978 0.00238 0.00041 0.00559 0.00035 0.00845 0.00140 0.04640
## p5 0.00061 0.00015 0.00003 0.00035 0.00002 0.00053 0.00009 0.00290
## p6 0.01479 0.00360 0.00063 0.00845 0.00053 0.01277 0.00211 0.07014
## p7 0.00245 0.00060 0.00010 0.00140 0.00009 0.00211 0.00035 0.01160
## p8 0.08120 0.01977 0.00343 0.04640 0.00290 0.07014 0.01160 0.38519
```

The genotype is 8,8. The probability of the most likely genotype is the one in position [8,8]:

```
## [1] 0.38519
```

The genotype is 1,8. The probability of the second most likely genotype is the 2*[1,8]:

```
## [1] 0.1624
```