Practical 03 SG: Haplotype estimation

Lovro Katalinić and Ivan Almer

Hand-in: 05/12/2020

Resolve the following exercise in groups of two students. Perform the computations and make the graphics that are asked for in the practical below. Take care to give each graph a title, and clearly label x and y axes, and to answer all questions asked. You can write your solution in a word or Latex document and generate a pdf file with your solution. Alternatively, you may generate a solution pdf file with Markdown. You can use R packages **genetics**, **haplo.stats**, **LDheatmap** and others for the computations. Take care to number your answer exactly as in this exercise. Upload your solution in **pdf format** to the web page of the course at raco.fib.upc.edu no later than the hand-in date.

```
## Loading required package: combinat
##
## Attaching package: 'combinat'
## The following object is masked from 'package:utils':
##
##     combn
## Loading required package: gdata
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
##
## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
##
## Attaching package: 'gdata'
## The following object is masked from 'package:stats':
##
##     nobs
## The following object is masked from 'package:utils':
##
##     object.size
## The following object is masked from 'package:base':
##
##     startsWith
## Loading required package: gtools
## Loading required package: MASS
## Loading required package: mvtnorm
##
## NOTE: THIS PACKAGE IS NOW OBSOLETE.
##
```

```
##   The R-Genetics project has developed an set of enhanced genetics
##   packages to replace 'genetics'. Please visit the project homepage
##   at http://rgenetics.org for informtion.
##
##
## Attaching package: 'genetics'
## The following objects are masked from 'package:base':
##
##     %in%, as.factor, order
## Loading required package: arsenal
##
## Attaching package: 'haplo.stats'
## The following object is masked from 'package:genetics':
##
##     locus
##
## Attaching package: 'data.table'
## The following objects are masked from 'package:gdata':
##
##     first, last
```

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.

```
apoe <- fread('APOE.dat', data.table=FALSE)
rownames(apoe) <- apoe[,1]
apoe <- apoe[,-c(1)]
#head(apoe)
```

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

```
n <- nrow(apoe)
p <- ncol(apoe)
cat(paste('There are', n, 'individuals and', p, 'SNPs in the database.\n'))
```

```
## There are 107 individuals and 162 SNPs in the database.
```

```
missing_percentage <- 100*sum(as.integer(is.na(apoe)))/(n*p)
cat(paste(missing_percentage, '% of data is missing.'))
```

```
## 0 % of data is missing.
```

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

```
possible_haplotypes <- 2^p
cat(paste('For bi-allelic SNPS, theoretically 2^', p, ' = ', format(possible_haplotypes, digits=3), ' ha
```

```
## For bi-allelic SNPS, theoretically 2^162 = 5.85e+48 haplotypes can be found for this data set.
```

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

```
get_prepared_data <- function(data) {
  prepared_data <- c()
  for(i in 1:ncol(data)) {
    prepared_data <- cbind(prepared_data,
                           substr(data[,i],1,1),
                           substr(data[,i],3,3))
  }
  prepared_data
}
```

```
estimate_haplotypes <- function(data) {
  data_prepared <- get_prepared_data(data)
  haplo_estimation <- haplo.em(data_prepared,
                               locus.label=colnames(data),
                               control=haplo.em.control(min.posterior=1e-3))

  haplos <- haplo_estimation$haplotype
  haplo_num <- dim(haplos)[1]
  cat(paste('Algorithm found', haplo_num, 'haplotypes.\n\n'))

  haplo_probs <- haplo_estimation$hap.prob
  haplo_probs_ordered <- order(haplo_probs, decreasing=TRUE)
  cat('Estimated probabilities in decreasing order:\n')
  cat(haplo_probs[haplo_probs_ordered])
  cat(paste('\n\nMost common is haplotype numbered ', haplo_probs_ordered[1], '.', sep=''))

  return(haplo_estimation)
}

apoe_haplos <- estimate_haplotypes(apoe)
```

```
## Algorithm found 31 haplotypes.
##
## Estimated probabilities in decreasing order:
## 0.3995055 0.1308411 0.07447912 0.06841314 0.05018155 0.04672897 0.03585747 0.03516129 0.02255473 0.0
##
## Most common is haplotype numbered 27.
```

5. (2p) 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).

```
constitutions <- apoe_haplos$nreps
ambiguous <- sum(constitutions > 1)
cat(paste('There are ', ambiguous, ' ambigous haplotypic constituions of an individual.', sep=''))
```

```
## There are 19 ambigous haplotypic constituions of an individual.
```

```
individual_index <- which(rownames(apoe) == 'NA20763')
most_likely_hc <- apoe_haplos$hap1code[individual_index]
cat(paste('\nMost likely haplotypic constitution of individual NA20763 is ', most_likely_hc, '.', sep='
```

```
##
```

```
## Most likely haplotypic constitution of individual NA20763 is 8.
```

6. (1p) Suppose we would delete polymorphism rs374311741 from the database prior to haplotype estimation. Would this affect the results obtained? Justify your answer.

```
apoe_without_polymorphism <- subset(apoe, select=-c(rs374311741))
apoe_without_polymorphism_haplo <- estimate_haplotypes(apoe_without_polymorphism)
```

```
## Algorithm found 31 haplotypes.
##
## Estimated probabilities in decreasing order:
## 0.3994985 0.1308411 0.07447842 0.06842063 0.05018152 0.04672897 0.03586333 0.03516137 0.02255615 0.0
##
## Most common is haplotype numbered 27.
```

Deleting one column from the database resulted in minor changes in haplotype probabilities, but the haplotype count stays the same. Intuitively one of 162 columns cannot have a significant influence on the whole result.

7. (1p) 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?

```
maf <- function(x){
  x <- genotype(x,sep="/")
  out <- summary(x)
  af1 <- min(out$allele.freq[,2],na.rm=TRUE)
  af1[af1==1] <- 0
  af1
}

mafs <- apply(apoe, 2, maf)
apoe_filtered <- apoe[, mafs > 0.10]
cat(paste('By filtering genetic variants that have MAF below 0.1, we reduced their number from ', p, ' t
```

```
## By filtering genetic variants that have MAF below 0.1, we reduced their number from 162 to 21.
```

```
apoe_filtered_haplo <- estimate_haplotypes(apoe_filtered)
```

```
## Algorithm found 8 haplotypes.
##
## Estimated probabilities in decreasing order:
## 0.6206356 0.1308411 0.1130093 0.07476636 0.03185051 0.01869159 0.005532668 0.004672897
##
## Most common is haplotype numbered 8.
```

The results changed dramatically when variants with minor alelle frequency below 0.10 were removed. With filtered dataset like this one, function `haplo.em` found 8 haplotypes.

8. (2p) 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?

```
haplotypes <- apoe_filtered_haplo$haplotype
haplotypes
```

```
##   rs892593 rs892594 rs2722659 rs2722660 rs2571147 rs2571148 rs34762924
## 1        C        A         C         C         G         T          C
## 2        C        G         C         C         G         T          C
## 3        C        G         C         C         G         T          C
```

```
## 4        C        G         C         C         G         T         C
## 5        G        G         T         T         A         C         A
## 6        G        G         T         T         A         C         A
## 7        G        G         T         T         A         C         A
## 8        G        G         T         T         A         C         A
##   rs2571149 rs2722661 rs2571150 rs147663893 rs8102685 rs2571151 rs35570438
## 1         T         A         G           C         T         G          A
## 2         T         A         G           C         T         G          T
## 3         T         A         G           T         T         G          T
## 4         T         A         G           T         T         G          T
## 5         C         G         T           C         C         T          T
## 6         C         G         T           C         C         T          T
## 7         C         G         T           T         C         G          T
## 8         C         G         T           T         C         T          T
##   rs2571152 rs2571153 rs2722662 rs35391606 rs2437014 rs2437013 rs2722664
## 1         T         A         C          A         C         T         A
## 2         T         A         C          A         C         T         A
## 3         T         A         C          A         C         T         A
## 4         T         C         C          A         C         T         A
## 5         G         C         T          A         T         A         G
## 6         G         C         T          C         T         A         G
## 7         T         A         C          A         C         T         G
## 8         G         C         T          C         T         A         G
```

```
superlocus <- c()
for (i in 1:dim(haplotypes)[2]) {
  superlocus <- paste(superlocus, haplotypes[,i], sep='')
}
superlocus
```

```
## [1] "CACCGTCTAGCTGATACACTA" "CGCCGTCTAGCTGTTACACTA" "CGCCGTCTAGTTGTTACACTA"
## [4] "CGCCGTCTAGTTGTTCCACTA" "GGTTACACGTCCTTGCTATAG" "GGTTACACGTCCTTGCTCTAG"
## [7] "GGTTACACGTTCGTTACACTG" "GGTTACACGTTCTTGCTCTAG"
```

```
summary(genotype(superlocus, sep=''))
```

```
##
## Number of samples typed: 8 (100%)
##
## Allele Frequency: (10 alleles)
##                       Count Proportion
## G                         4       0.25
## C                         4       0.25
## GTTACACGTTCTTGCTCTAG      1       0.06
## GTTACACGTTCGTTACACTG      1       0.06
## GTTACACGTCCTTGCTCTAG      1       0.06
## GTTACACGTCCTTGCTATAG      1       0.06
## GCCGTCTAGTTGTTCCACTA      1       0.06
## GCCGTCTAGTTGTTACACTA      1       0.06
## GCCGTCTAGCTGTTACACTA      1       0.06
## ACCGTCTAGCTGATACACTA      1       0.06
##
##
## Genotype Frequency:
##                         Count Proportion
```

```
## C/ACCGTCTAGCTGATACACTA     1       0.12
## G/GTTACACGTTCTTGCTCTAG     1       0.12
## G/GTTACACGTTCGTTACACTG     1       0.12
## G/GTTACACGTCCTTGCTCTAG     1       0.12
## G/GTTACACGTCCTTGCTATAG     1       0.12
## C/GCCGTCTAGTTGTTCCACTA     1       0.12
## C/GCCGTCTAGTTGTTACACTA     1       0.12
## C/GCCGTCTAGCTGTTACACTA     1       0.12
##
## Heterozygosity (Hu)  = 0.9
## Poly. Inf. Content   = 0.8272705
```