Bioinformatics and Statistical Genetic

Haplotype estimation

Leonardo Ortoleva & Egon Ferri

Contents


```
knitr::opts_chunk$set(echo = F)
```

1 Load data into the R environment.

Table 1: Data

id	rs112748686	rs374311741	rs569874826	rs538789834	rs144831227	rs148595630
NA20502	C/C	C/C	C/C	G/G	C/C	G/G
NA20503	C/C	C/C	C/C	G/G	C/C	G/G
NA20504	C/C	C/C	C/C	G/G	C/C	G/G
NA20505	C/C	C/C	C/C	G/G	C/C	G/G
NA20506	C/C	C/C	C/C	G/G	C/C	G/G
NA20507	C/C	C/C	C/C	G/G	C/C	G/G
NA20508	C/C	C/C	C/C	G/G	C/C	G/G
NA20509	C/C	C/C	C/C	G/G	C/C	A/G
NA20510	C/C	C/C	C/C	G/G	C/C	G/G
NA20511	C/C	C/C	C/C	G/G	C/C	G/G

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

```
## [1] "The number of individuals is 107"
```

```
## [1] "The number of variants is 162"
```

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

```
## [1] "The number of haplotypes that can theoretically be found for this data set is: 5.846e+48"
```

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] "The number of haplotypes observed in this data set is 31"
```

```
##           27            8            5           28           18
## 0.3994979012 0.1308411215 0.0744779724 0.0684218323 0.0501816036
##           11           20           30            9           14
## 0.0467289720 0.0358616326 0.0351613306 0.0225591573 0.0204956086
##           26           24           16           23            1
## 0.0186915888 0.0161153410 0.0086858440 0.0073505753 0.0046728972
##            2            6            7           12           13
## 0.0046728972 0.0046728972 0.0046728972 0.0046728972 0.0046728972
##           17           19           22           31           29
## 0.0046728972 0.0046728972 0.0046728972 0.0046728972 0.0040258668
##           21            3           15           25            4
## 0.0034017609 0.0033021999 0.0028688774 0.0021374332 0.0016590801
##           10
## 0.0008053287
```

```
## [1] "The the most common haplotype observed in this data set is the number:"
##        27
## 0.3994979
```

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] "The haplotypic constitution of these 19 individuals is ambiguous or uncertain:"
## indx.subj
##   3  17  21  22  26  28  33  40  44  52  59  60  62  78  80  82  88  99
##   2   2   2   2   2   2   2   2   2   2   3   2   2   2   2   2   2   2
## 100
##   2
```

Infact, the count of haplotype pairs that map to each subjects marker genotypes is greater than 1.

The individual NA20763 (number 59) has these three possibile compositions:

```
## [1] "Diplotype: 24, 21 with probability : 0.0153776935194038"
## [2] "Diplotype: 28, 18 with probability : 0.963150926639683"
## [3] "Diplotype: 20, 25 with probability : 0.0214713798409127"
```

The most likely haplotypic constitution of individual NA20763 is (28, 18) by far.

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

After removing the rs374311741 polymorphism, we redo the experiment and these are the results:

```
## [1] 2
## [1] "The number of haplotypes observed in this data set is 31"
##           27            8            5           28           18
## 0.3995034702 0.1308411215 0.0744783047 0.0684159336 0.0501815487
##           11           20           30            9           14
## 0.0467289720 0.0358613135 0.0351612872 0.0225531477 0.0204955782
##           26           24           16           23            1
## 0.0186915888 0.0161162168 0.0086858581 0.0073505303 0.0046728972
##            2            6            7           12           13
## 0.0046728972 0.0046728972 0.0046728972 0.0046728972 0.0046728972
##           17           19           22           31           29
## 0.0046728972 0.0046728972 0.0046728972 0.0046728972 0.0040258798
##           21            3           15           25            4
## 0.0034021208 0.0033022178 0.0028689077 0.0021369625 0.0016587299
##           10
## 0.0008113382
```

Since we are removing a monomorphic variant, we don't experiment any change in our result. Infact, a monomorphic variant doesn't increase the number of haplotypes configurations,31. And obviously doesn't modify the probability distribution neither.

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 observed in this data set is 8"
```

Removing all the variants with very low minor allele frequency, we remove a lot of different combinations, so the number of different haplotypes drops sharply to 8.

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 probability of the alleles of our superlocus are:

```
##           8           1           6           4           2           7
## 0.620635549 0.130841121 0.113009311 0.074766355 0.031850502 0.018691589
##           3           5
## 0.005532675 0.004672897
```

Under the assumption of Hardy-Weinberg equilibrium, to find the probabilities of genotypes in our new locus we just need the second moment of our vector of probabilities:

Table 2: Genotype Frequencies. Diagonal correspond to homozigotes probabilities, ohter cells represents probability of genotypes formed by the column and row alleles

	p1	p2	p3	p4	p5	p6	p7	p8
p1	0.0171194	0.0041674	0.0007239	0.0097825	0.0006114	0.0147863	0.0024456	0.0812047
p2	0.0041674	0.0010145	0.0001762	0.0023813	0.0001488	0.0035994	0.0005953	0.0197676
p3	0.0007239	0.0001762	0.0000306	0.0004137	0.0000259	0.0006252	0.0001034	0.0034338
p4	0.0097825	0.0023813	0.0004137	0.0055900	0.0003494	0.0084493	0.0013975	0.0464027
p5	0.0006114	0.0001488	0.0000259	0.0003494	0.0000218	0.0005281	0.0000873	0.0029002
p6	0.0147863	0.0035994	0.0006252	0.0084493	0.0005281	0.0127711	0.0021123	0.0701376
p7	0.0024456	0.0005953	0.0001034	0.0013975	0.0000873	0.0021123	0.0003494	0.0116007
p8	0.0812047	0.0197676	0.0034338	0.0464027	0.0029002	0.0701376	0.0116007	0.3851885

The two most likely genotypes probabilities are:

```
## [1] 0.3851885
```

```
## [1] 0.1624093
```

of, respectevly, (8,8), and (8,1) (the sum of the form (8,1) and (1,8)).