Bioinformatics and Statistical Genetics

Relatedness analysis

Leonardo Ortoleva & Egon Ferri

Contents

1 The file CHD.zip contains genotype information, in the form of PLINK files chd.fam, chd.bed and chd.bim. The files contain genetic information of 109 presumably unrelated individuals of a sample of Chinese in Metropolitan Denver, CO, USA, and corresponds to the CHD sample of the 1,000 Genomes project (www.internationalgenome.org).

2 The chd.bed contains the genetic data in binary form. First convert the .bed file to a text file, chd.raw, with the data in (0, 1, 2) format.

```
## [1] "plink --bfile CHD --recodeA --out CHD"
```

3 Read the genotype data in (0, 1, 2) format into the R environment. Consult the pedigree information. Are there any documented family relationships for this data set?

Table 1: Data

FID	IID	PAT	MAT	SEX	PHENOTYPE	rs1110052_T	rs9442373_C
NA17970	NA17970	0	0	2	-9	1	1
NA17977	NA17977	0	0	2	-9	1	0
NA17981	NA17981	0	0	2	-9	1	0
NA17993	NA17993	0	0	2	-9	0	2
NA18101	NA18101	0	0	2	-9	1	1

Table 2: Summary

FID	IID	PAT	MAT	SEX
Length:109	Length:109	Min. :0	Min. :0	Min. :1.000
Class :character	Class :character	1st Qu.:0	1st Qu.:0	1st Qu.:1.000
Mode :character	Mode :character	Median :0	Median :0	Median :2.000
NA	NA	Mean :0	Mean :0	Mean :1.541
NA	NA	3rd Qu.:0	3rd Qu.:0	3rd Qu.:2.000
NA	NA	Max. :0	Max. :0	Max. :2.000

In the first column there is the $family\ Id$, which permits to recognize the family of each individual (different for each one), plus the column mother and father are vectors of zeros, so we don't have relatedness informations.

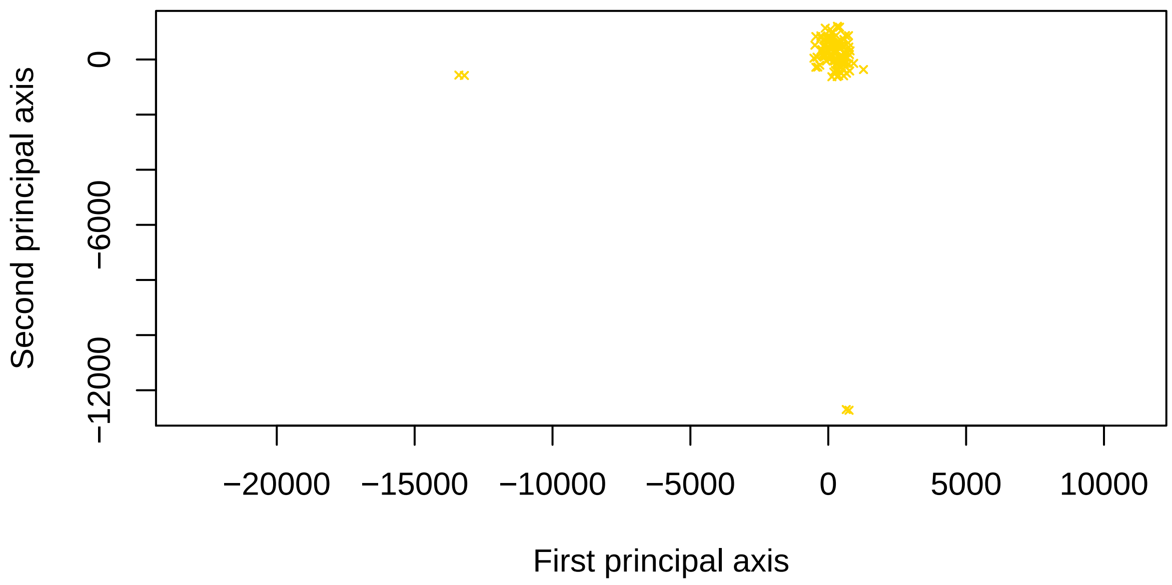
```
## [1] "The number of different families ID is: "
```

```
## [1] 109
```

4 Compute the Manhattan distance between the inviduals on the basis of the genetic data. Use classical metric multidimensional scaling to obtain a map of the indivuals. Are the data homogeneous? Identify possible outliers.

Table 3: Submatrix

1	2	3	4	5
0	21144	21251	21113	21018
21144	0	21231	20919	21070
21251	21231	0	21052	21011
21113	20919	21052	0	21025
21018	21070	21011	21025	0



Map of the individuals

After a MDS with n-1 dimensions, the great part of the data is homogenous. There are only two small group of outlier of two elements.

```
## [1] "Outliers on the first principal axis:"
```

```
##         3        18
## -13396.10 -13195.59
```

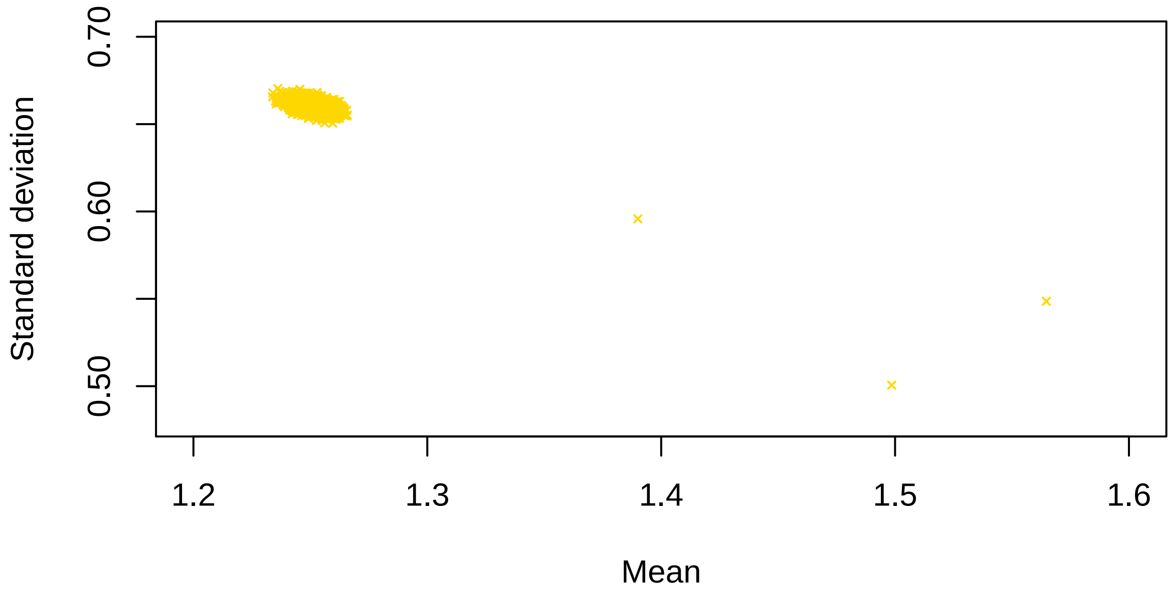
```
## [1] "Outliers on the second principal axis:"
```

```
##        62        89
## -12700.07 -12727.02
```

Table 4: Outliers

FID	IID	PAT	MAT	SEX	PHENOTYPE
NA17981	NA17981	0	0	2	-9
NA17986	NA17986	0	0	1	-9
NA17976	NA17976	0	0	1	-9
NA18116	NA18116	0	0	2	-9

5 Compute the average number of alleles shared between each pair of individuals over all genetic variants. Compute also the corresponding standard deviation. Plot the standard deviation against the mean. Do you think there are any pairs with a close family relationship? How many? Identify the corresponding individuals.



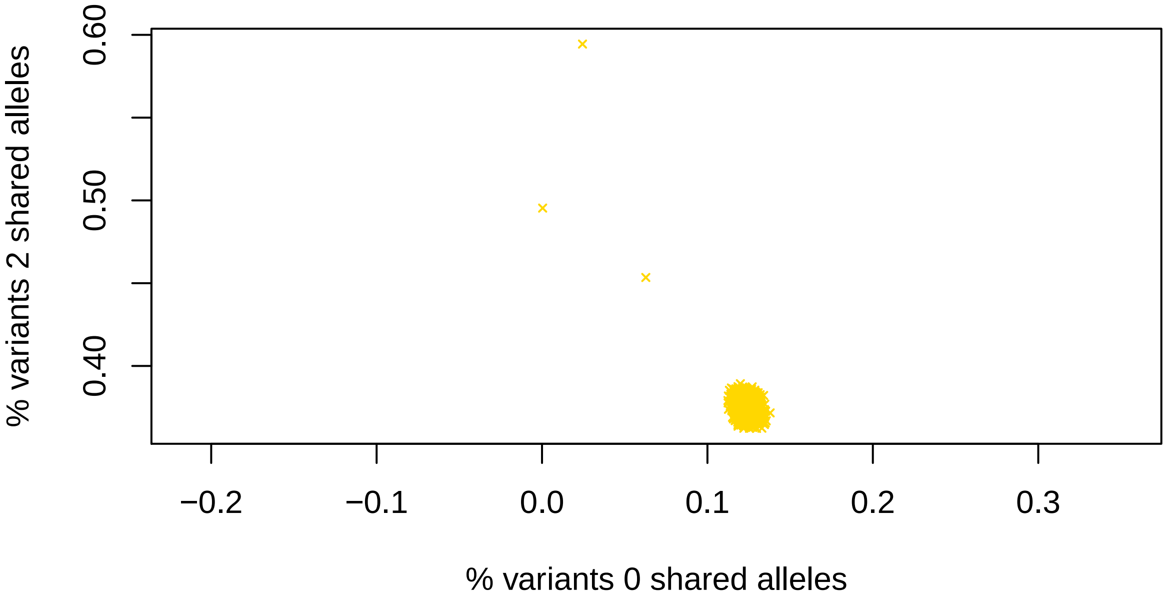
This plot reveals characteristic *clusters* that correspond to the different family relationships. Looking at it, it's possible to see that almost all the individual have not family relationship.

But, if we analyze more in details the plot, there are 3 points with means between 1.4 and 1.6. This means that we have a close family relationship in 3 pairs of individuals. This relationships are probably respectively 2ND, PO and FS (from left to right), because we expepect that full siblings are very similiar, more than parents-sons, that are more similiar than 2nd grade relatives, that are more similiar than unrelated couples.

Table 5: Relevant relationship

	IID1	IID2	Mean	Standar Deviation
230	NA17981	NA17986	1.5647	0.5486
1138	NA18150	NA17980	1.3900	0.5958
4785	NA17976	NA18116	1.4985	0.5006

6 Make a plot of the percentage of variants sharing no alleles versus the percentage of variants sharing two alleles for all pairs of individuals. Do you think there are any pairs with a close family relationship? How many? Identify the corresponding individuals.



Here, again, we have a plot that reveals *clusters* that correspond to the different family relationship. The number of clusters and outliers is consistent with the previous case. We have three outliers which are probably FS, PO and 2ND (from top to bottom), with a p_0 almost 0, but an higher p_2 (percentage of marker with 2 IBS alleles) than the big cluster of Unrelated. We added a little bit of jitter in order to avoid overlaps in the big cluster.

Table 6: Relevant relationship

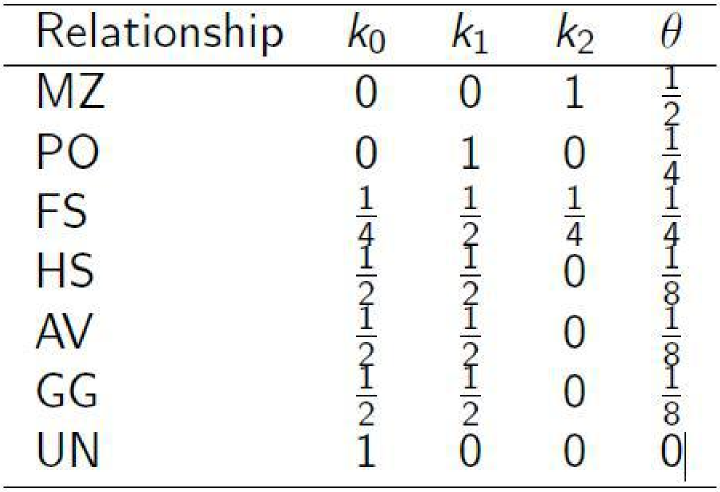
	IID1	IID2	% No shared alleles	% 2 shared alleles
230	NA17981	NA17986	0.0276	0.5923
1138	NA18150	NA17980	0.0585	0.4485
4785	NA17976	NA18116	0.0003	0.4988

7 Can you identify any obvious family relationships between any pairs? Argue your answer.

Yes, actually, our previous results are confirmed. The obvious relationships are the same we obtained and explained before.

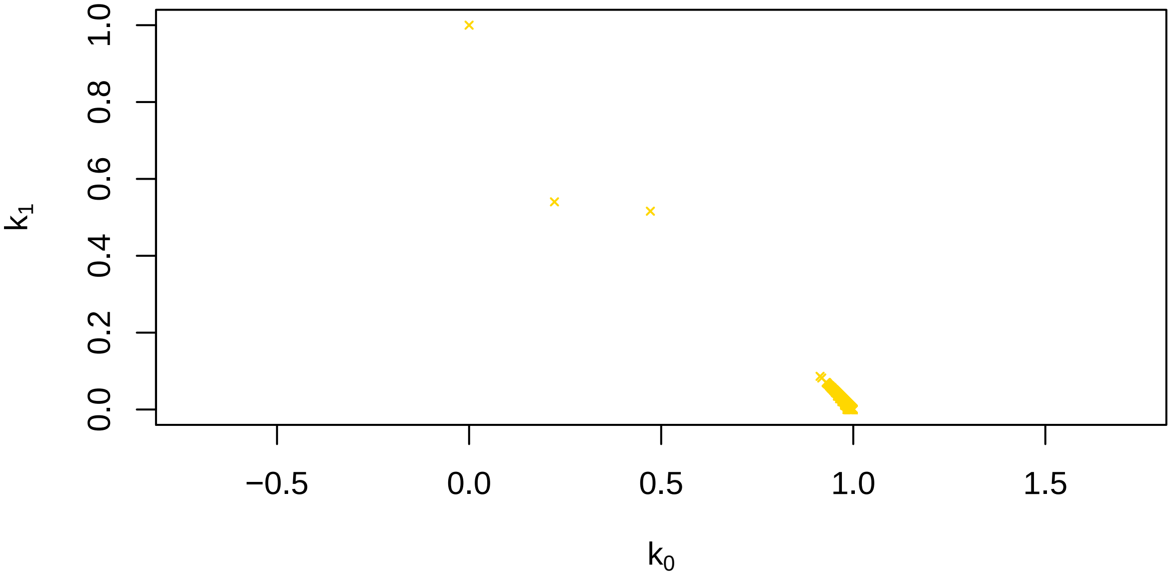
8 Estimate the *Cotterman coefficients* for all pairs using PLINK. Read the coeffients into the R environment and plot the probability of sharing no IBD alleles against the probability of sharing one IBD allele. Add the theoretical values of the Cotterman coefficients for standard relationships to your plot.

```
## [1] "Running the plink command "

## [1] "plink --bfile CHD --genome --genome-full --out CHD"

## [1] " we obatin a CHD.genome file, which contains all the informations"

## [1] " we need about IBD and Cotterman coefficients."
```



Relationship	k_0	k_1	k_2	θ
MZ	0	0	1	$\frac{1}{2}$
PO	0	1	0	$\frac{1}{4}$
FS	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
HS	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{8}$
AV	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{8}$
GG	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{8}$
UN	1	0	0	0

Figure 1: “Cotterman coefficents”



9 Make a table of pairs for which you suspect that they have a close family relationship, and list their Cotterman coefficients. State your final conclusions about what relationship these pairs probably have.

Table 7: Relevant relationship

	IID1	IID2	Z0	Z1	Z2
230	NA17981	NA17986	0.2222	0.5403	0.2376
1138	NA18150	NA17980	0.4719	0.5160	0.0121
4785	NA17976	NA18116	0.0000	1.0000	0.0000

The great part of points are UNRELATED, infact they have a k_0 almost to 1. The point with $k_0 = 0.22$ and $k_1 = 0.54$ approximately is an FS. Then we have a point at $k_0 = 0.47$ and $k_1 = 0.51$, that indicates a 2ND degree relationship (HS, AV or GG). Finally, the top left point is a PO with $k_1 = 1$.

These results confirm all what we said in the previous answers.

10 Is there any relationship between the MDS map you made and the relationships between the individuals? Report your findings.

All the four outliers that we found in the MDS map are part of the three relationship couple.

In particular, $NA17981$ and $NA17986$ both are outlier on the same axis, and from our findings, they are Full Siblings, while $NA17976$ and $NA18116$ are Parent-Offspring. This make sense according to the analysis we did. Of the three pairs found, two are separated from the big cluster in the MDS, but very close to their relatives.

11 Which of the three graphics (m, s), (p_0, p_2) or (k_0, k_1) do you like best for identifying relationships? Argue your answer.

First of all, all the three graphics are coerent each other and give us the same results at the end. We think that the first one, (m, s), is the more intuitive because it's possible to see clearly the clusters and to undestand, thanks to the mean value, their relationship. But, as the Cotterman coefficients are theoretical values, thanks to the third plot we undestood without any doubt which kind of relationship our individuals have.