Practical 05 SG: Association analysis and Relatedness research

Quim Aguado, Marcel Cases

08-Jan-2022

Association analysis

Ex 1 What is the sample size? What is the number of cases and the number of controls? Construct the contingency table of genotype by case/control status.

```
df.aa <- read.table("rs394221.dat", header=FALSE)
nrow(df.aa)
```

```
## [1] 1167
```

```
ncol(df.aa)
```

```
## [1] 2
```

```
df.aa.cases.tot <- length(df.aa[,2][df.aa[,2]=="case"])
df.aa.cases.tot
```

```
## [1] 509
```

```
df.aa.control.tot <- length(df.aa[,2][df.aa[,2]=="control"])
df.aa.control.tot
```

```
## [1] 658
```

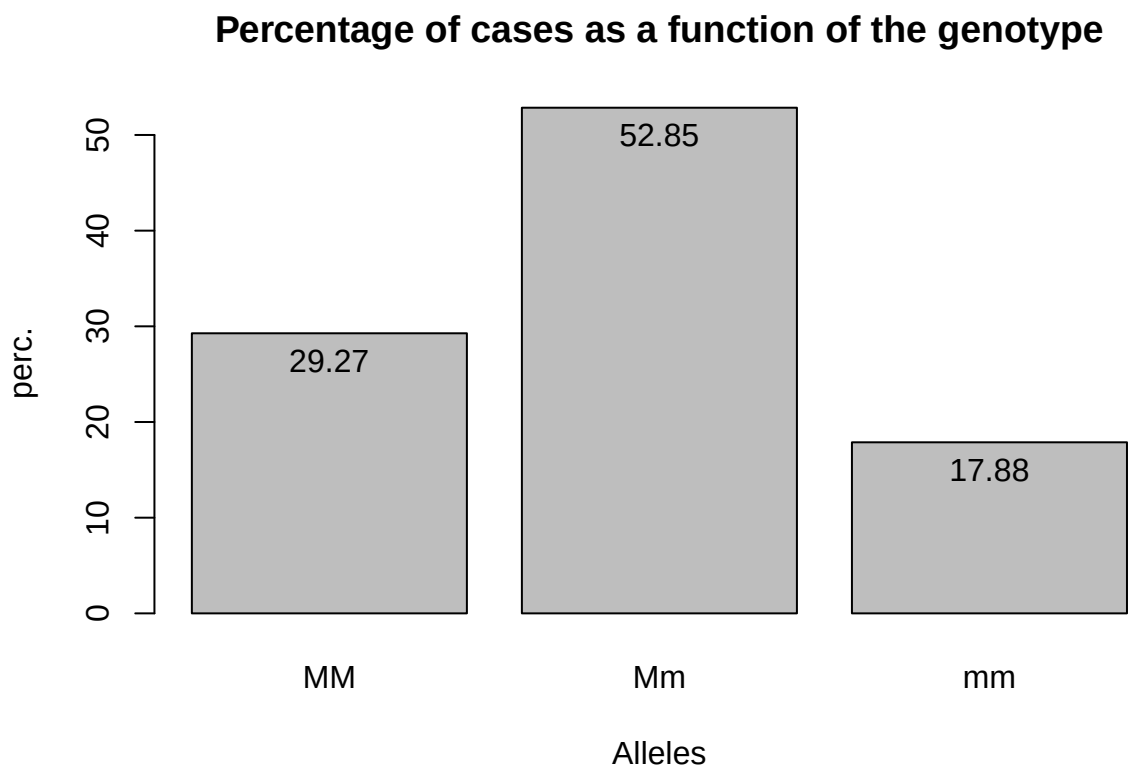
- Sample size: 1167
- Number of cases: 509
- Number of controls: 658

```
df.aa.cases <- c(length(df.aa[,2][df.aa[,2]=="case"&df.aa[,1]=="MM"]),
                 length(df.aa[,2][df.aa[,2]=="case"&df.aa[,1]=="Mm"]),
                 length(df.aa[,2][df.aa[,2]=="case"&df.aa[,1]=="mm"])
                 )
df.aa.controls <- c(length(df.aa[,2][df.aa[,2]=="control"&df.aa[,1]=="MM"]),
                    length(df.aa[,2][df.aa[,2]=="control"&df.aa[,1]=="Mm"]),
                    length(df.aa[,2][df.aa[,2]=="control"&df.aa[,1]=="mm"])
                    )
con.tab.geno <- rbind(df.aa.cases,df.aa.controls)
rownames(con.tab.geno) <- c("Cases","Controls")
colnames(con.tab.geno) <- c("MM","Mm","mm")
con.tab.geno
```

```
##           MM  Mm  mm
## Cases    149 269  91
## Controls 153 325 180
```

Ex 2 Explore the data by plotting the percentage of cases as a function of the genotype, ordering the latter according to the number of M alleles. Which allele increases the risk of the disease?

```
df.aa.cases.perc <- 100*df.aa.cases/df.aa.cases.tot
bp <- barplot(df.aa.cases.perc,names.arg=c("MM","Mm","mm"),xlab="Alleles",ylab="perc.", main="Percentag
text(x=bp,y=df.aa.cases.perc,label=round(df.aa.cases.perc,digits=2),pos=1)
```



The data follows a *co-dominant model*. The allele "Mm" has a higher risk of contracting the disease.

Ex 3 Test for equality of allele frequencies in cases and controls by doing an alleles test. Report the test statistic, its reference distribution, and the p-value of the test. Is there evidence for different allele frequencies?

```
con.tab.allele <- cbind(2*con.tab.geno[,1]+con.tab.geno[,2],2*con.tab.geno[,3]+con.tab.geno[,2])
colnames(con.tab.allele) <- c("M","m")
con.tab.allele
```

```
##            M   m
## Cases    567 451
## Controls 631 685
```

```
con.tab.allele.chisq <- chisq.test(con.tab.allele,correct=FALSE)
con.tab.allele.chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  con.tab.allele
## X-squared = 13.797, df = 1, p-value = 0.0002037
```

```
con.tab.allele.chisq$expected
```

```
##                M       m
```

```
## Cases    522.521 495.479
## Controls 675.479 640.521
```

- Reference distribution: Chi-square
- p-value = 0.0002037

Expected cases for allele M are lower compared to real cases, while expected cases for allele m are higher compared to real cases.

Ex 4 Which are the assumptions made by the alleles test? Perform and report any additional tests you consider adequate to verify the assumptions. Do you think the assumptions of the alleles test are met?

```
fisher.test(con.tab.allele)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  con.tab.allele
## p-value = 0.0002368
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.154074 1.614038
## sample estimates:
## odds ratio
##   1.364592
```

```
or <- (con.tab.allele[1,1]*con.tab.allele[2,2])/(con.tab.allele[1,2]*con.tab.allele[2,1])
or
```

```
## [1] 1.364796
```

```
se.lor <- sqrt(sum(1/con.tab.allele))
se.lor
```

```
## [1] 0.08381887
```

```
ll.logodds <- log(or) - qnorm(0.975)*se.lor
ul.logodds <- log(or) + qnorm(0.975)*se.lor
ll.odds <- exp(ll.logodds)
ul.odds <- exp(ul.logodds)
ll.odds
```

```
## [1] 1.158033
```

```
ul.odds
```

```
## [1] 1.608476
```

Ex 5 Perform the Armitage trend test for association between disease and number of M alleles. Report the test statistic, its reference distribution and the p-value of the test. Do you find evidence for association?

```
con.tab.geno
```

```
##           MM  Mm  mm
## Cases    149 269  91
## Controls 153 325 180
```

```
df.aa.cases
```

```
## [1] 149 269  91
```

```
df.aa.controls
```

```
## [1] 153 325 180
```

```
df.aa.cases.rep <- rep(c(0,1,2),df.aa.cases)
df.aa.controls.rep <- rep(c(0,1,2),df.aa.controls)
x <- c(rep(1,sum(df.aa.cases)),
       rep(0,sum(df.aa.controls)))
y <- c(df.aa.cases.rep,df.aa.controls.rep)
length(x)
```

```
## [1] 1167
```

```
length(y)
```

```
## [1] 1167
```

```
r <- cor(x,y)
r
```

```
## [1] -0.1097624
```

```
n <- sum(con.tab.geno)
A <- n*(r^2)
A
```

```
## [1] 14.05977
```

```
pvalue <- pchisq(A,df=1,lower.tail=FALSE)
pvalue
```

```
## [1] 0.0001770917
```

- Reference distribution: Chi-square
- p-value = 0.0001082287

p-value is very close to zero, which means evidence for association.

Ex 6 Test for association between genotype and disease status by a logistic regression of disease status on genotype, treating the latter as categorical. Do you find significant evidence for association? Which allele increase the risk for the disease? Give the odds ratios of the genotypes with respect to base line genotype *mm*. Provide 95% confidence intervals for these odds ratios.

```
cas <- rep(c("MM","Mm","mm"),con.tab.geno[1,])
con <- rep(c("MM","Mm","mm"),con.tab.geno[2,])
ncas <- length(cas)
ncon <- length(con)
y <- c(rep(1,ncas),rep(0,ncon))
x <- factor(c(cas,con))
out.lm <- glm(y~x, family = binomial(link = "logit"))
summary(out.lm)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1662  -1.0982  -0.9046   1.2587   1.4773
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  -0.6821     0.1286  -5.303 1.14e-07 ***
## xMm           0.4930     0.1528   3.227 0.001251 **
## xMM           0.6556     0.1726   3.798 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1598.7  on 1166  degrees of freedom
## Residual deviance: 1582.7  on 1164  degrees of freedom
## AIC: 1588.7
##
## Number of Fisher Scoring iterations: 4
```

```
or <- exp(coefficients(out.lm))
```

Allele Mm has a higher risk of contracting the disease. ($\Pr(>|z|) = 0.001251$).

Relatedness research

Ex 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).

Ex 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, using the PLINK instruction:

```
plink --bfile CHD --recodeA --out CHD
```

```
system("./plink.exe --bfile CHD --recodeA --out CHD")
```

```
## [1] 0
```

Ex 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?

```
df <- fread("CHD.raw",data.table = FALSE)
df[1:5,1:6]
```

```
##       FID     IID PAT MAT SEX PHENOTYPE
## 1 NA17970 NA17970   0   0   2        -9
## 2 NA17977 NA17977   0   0   2        -9
## 3 NA17981 NA17981   0   0   2        -9
## 4 NA17993 NA17993   0   0   2        -9
## 5 NA18101 NA18101   0   0   2        -9
```

There is only information on the sex of each individual, nothing else. No family relationships are contained in the dataset.

Ex 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.

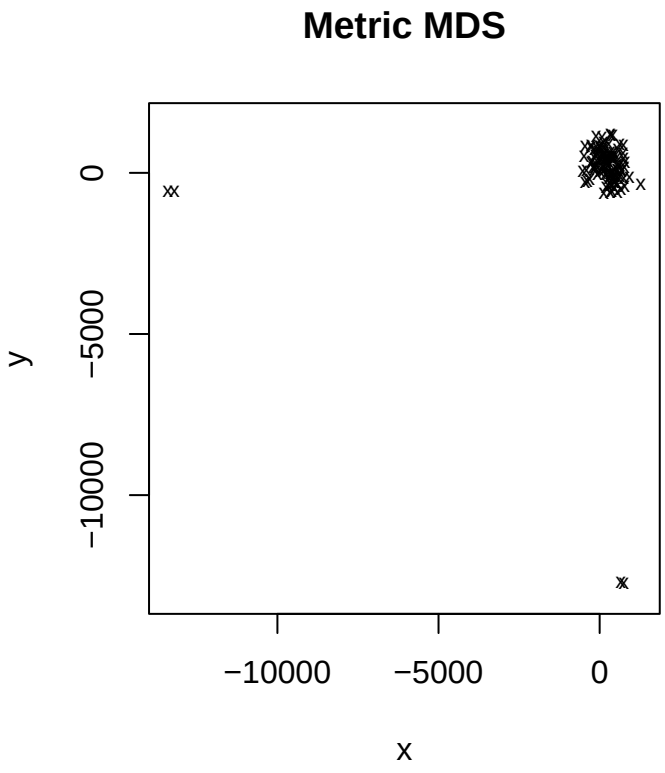
```
gendata <- df[,7:ncol(df)]
sum(is.na(gendata))
```

```
## [1] 0
```

```
Dobs <- as.matrix(dist(gendata, method = "manhattan"))
Dobs[1:5,1:5]
```

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

```
n <- nrow(gendata) # number of individuals
p <- ncol(gendata) # number of variants
mds.fit <- cmdscale(Dobs,eig=TRUE,k=n-1) # k is the max dimension
X <- mds.fit$points[,1:2]
par(pty="s") # square
plot(X[,1],X[,2],xlab="x",ylab="y",main="Metric MDS",type="n",asp=1)
text(X[,1],X[,2],labels="x",cex=.7)
```



We clearly observe that most of the individuals are related to a larger group (upper-right), while there are two smaller groups that are independent from one another. There are three different groups in total.

Ex 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.

```
gen.matrix <- as.matrix(gendata)
ibs.mean <- function(x,y) {
   y <- mean(2 - abs(x - y),na.rm=TRUE)
```

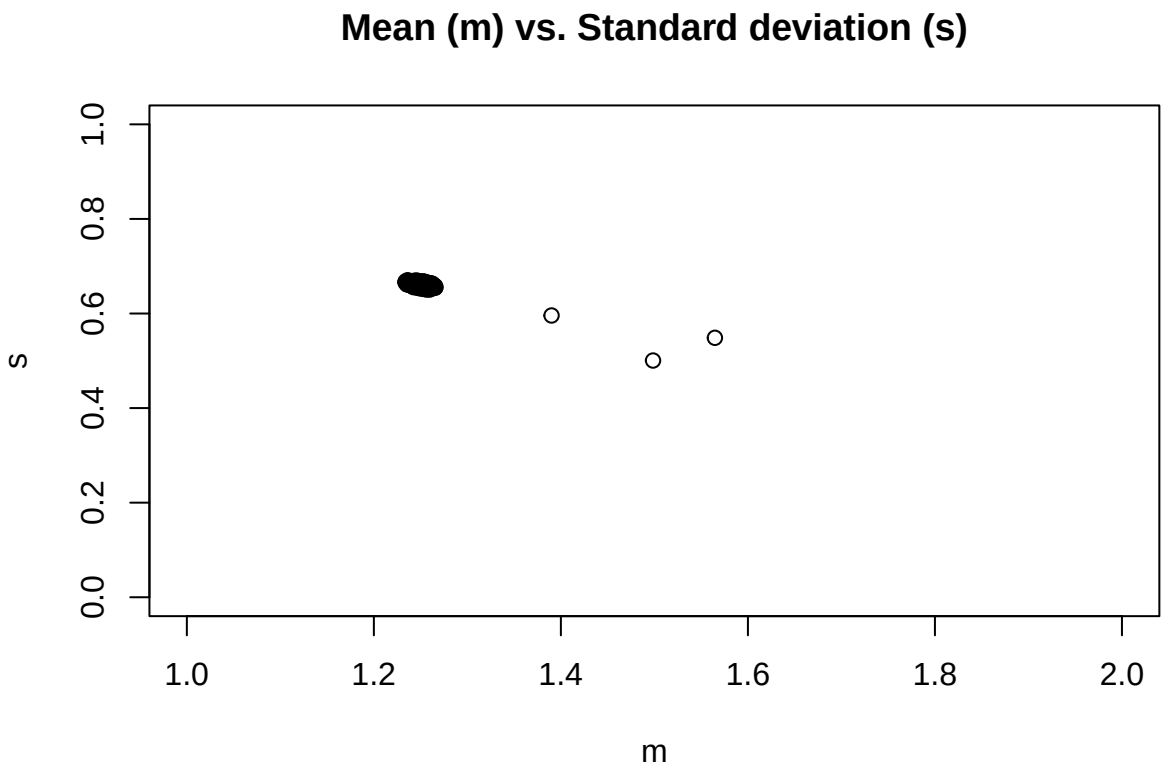
```
   return(y)
}
ibs.sd <- function(x,y) {
   y <- sd(abs(x-y),na.rm=TRUE)
   return(y)
}
#ibs.mean(gen.matrix[1,],gen.matrix[2,])
#ibs.sd(gen.matrix[1,],gen.matrix[2,])

ibs.mean.matrix <- matrix(nrow=n,ncol=n)
ibs.sd.matrix <- matrix(nrow=n,ncol=n)
for(i in 1:n) {
  for(j in 1:n) {
    ibs.mean.matrix[i,j] <- ibs.mean(gen.matrix[i,],gen.matrix[j,])
    ibs.sd.matrix[i,j] <- ibs.sd(gen.matrix[i,],gen.matrix[j,])
  }
}
ibs.mean.matrix[1:5,1:5]
```

```
##          [,1]     [,2]     [,3]     [,4]     [,5]
## [1,] 2.000000 1.249094 1.245294 1.250195 1.253569
## [2,] 1.249094 2.000000 1.246005 1.257085 1.251722
## [3,] 1.245294 1.246005 2.000000 1.252362 1.253818
## [4,] 1.250195 1.257085 1.252362 2.000000 1.253321
## [5,] 1.253569 1.251722 1.253818 1.253321 2.000000
```

```
ibs.sd.matrix[1:5,1:5]
```

```
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 0.0000000 0.6626760 0.6614932 0.6619661 0.6597397
## [2,] 0.6626760 0.0000000 0.6604233 0.6576489 0.6598009
## [3,] 0.6614932 0.6604233 0.0000000 0.6626725 0.6617673
## [4,] 0.6619661 0.6576489 0.6626725 0.0000000 0.6594852
## [5,] 0.6597397 0.6598009 0.6617673 0.6594852 0.0000000
```

```
ibs.m <- ibs.mean.matrix[lower.tri(ibs.mean.matrix)]
ibs.s <- ibs.sd.matrix[lower.tri(ibs.sd.matrix)]
plot(ibs.m,ibs.s,xlim=c(1,2),ylim=c(0,1),xlab="m",ylab="s",main="Mean (m) vs. Standard deviation (s)")
```




```
which(ibs.mean.matrix >= 1.3 & ibs.mean.matrix < 2, arr.ind = TRUE)
```

```
##      row col
## [1,]  18   3
## [2,]  17  12
## [3,]  12  17
## [4,]   3  18
## [5,]  89  62
## [6,]  62  89
```

```
#which(ibs.sd.matrix <= 0.6 & ibs.sd.matrix > 0, arr.ind = TRUE)
```

Pairs of individuals whose coordinates are in the list above have larger allele sharing average compared to the main subgroup, which means a closer family relationship. These three pairs are:

```
print(paste(df[18,1], "and", df[3,1]))
```

```
## [1] "NA17986 and NA17981"
```

```
print(paste(df[17,1], "and", df[12,1]))
```

```
## [1] "NA17980 and NA18150"
```

```
print(paste(df[89,1], "and", df[62,1]))
```

```
## [1] "NA18116 and NA17976"
```

Ex 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.

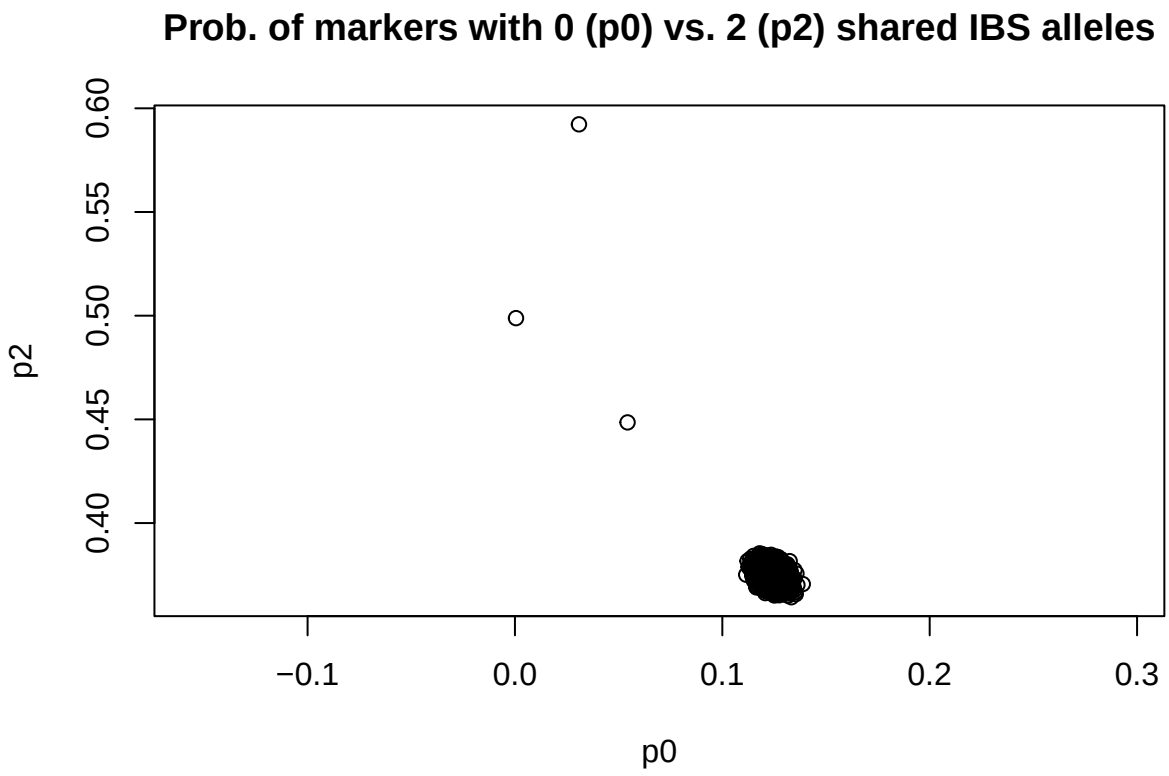
```
gen.matrix[1:5,1:5]
```

```
##      rs1110052_T rs9442373_C rs3813204_A rs2274264_G rs9439458_C
## [1,]           1           1           2           0           2
## [2,]           1           0           1           2           1
## [3,]           1           0           0           0           1
## [4,]           0           2           1           1           1
## [5,]           1           1           2           0           1
```

```
stats <- function(x,y) {
  aux <- 2-abs(x-y) # number of shared alleles
  n0 <- sum(aux==0,na.rm=TRUE)
  n1 <- sum(aux==1,na.rm=TRUE)
  n2 <- sum(aux==2,na.rm=TRUE)
  n <- sum(!is.na(aux))
  p0 <- n0/n
  p1 <- n1/n
  p2 <- n2/n
  y <- c(p0,p1,p2)
  return(y)
}

sum(stats(gen.matrix[1,],gen.matrix[2,]))
```

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

```
Mp0 <- matrix(NA,nrow=n,ncol=n)
Mp1 <- matrix(NA,nrow=n,ncol=n)
Mp2 <- matrix(NA,nrow=n,ncol=n)

for(i in 1:n) {
   for(j in 1:n) {
      statsofapair <- stats(gen.matrix[i,],gen.matrix[j,])
      Mp0[i,j] <- statsofapair[1]
      Mp1[i,j] <- statsofapair[2]
      Mp2[i,j] <- statsofapair[3]
   }
}

p0vec <- Mp0[lower.tri(Mp0)]
p2vec <- Mp2[lower.tri(Mp2)]

plot(jitter(p0vec,amount=.005),p2vec,asp=1,xlab="p0",ylab="p2",main="Prob. of markers with 0 (p0) vs. 2
```




```
which(Mp2 >= 0.4 & Mp2 < 1, arr.ind = TRUE)
```

```
##      row col
## [1,]  18   3
## [2,]  17  12
## [3,]  12  17
## [4,]   3  18
## [5,]  89  62
## [6,]  62  89
```

We have found the same three pairs of individuals that have a stronger relationship compared to the main sub-group. These pairs of users have at least 40% of their variants with two shared alleles:

```
print(paste(df[18,1], "and", df[3,1]))
```

```
## [1] "NA17986 and NA17981"
```

```
print(paste(df[17,1], "and", df[12,1]))
```

```
## [1] "NA17980 and NA18150"
```

```
print(paste(df[89,1], "and", df[62,1]))
```

```
## [1] "NA18116 and NA17976"
```

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

According to mean vs. std and p vs. p2 results, the following pairs of individuals might have some kind of family relationship:

- "NA17986 and NA17981"

- "NA17980 and NA18150"
- "NA18116 and NA17976"

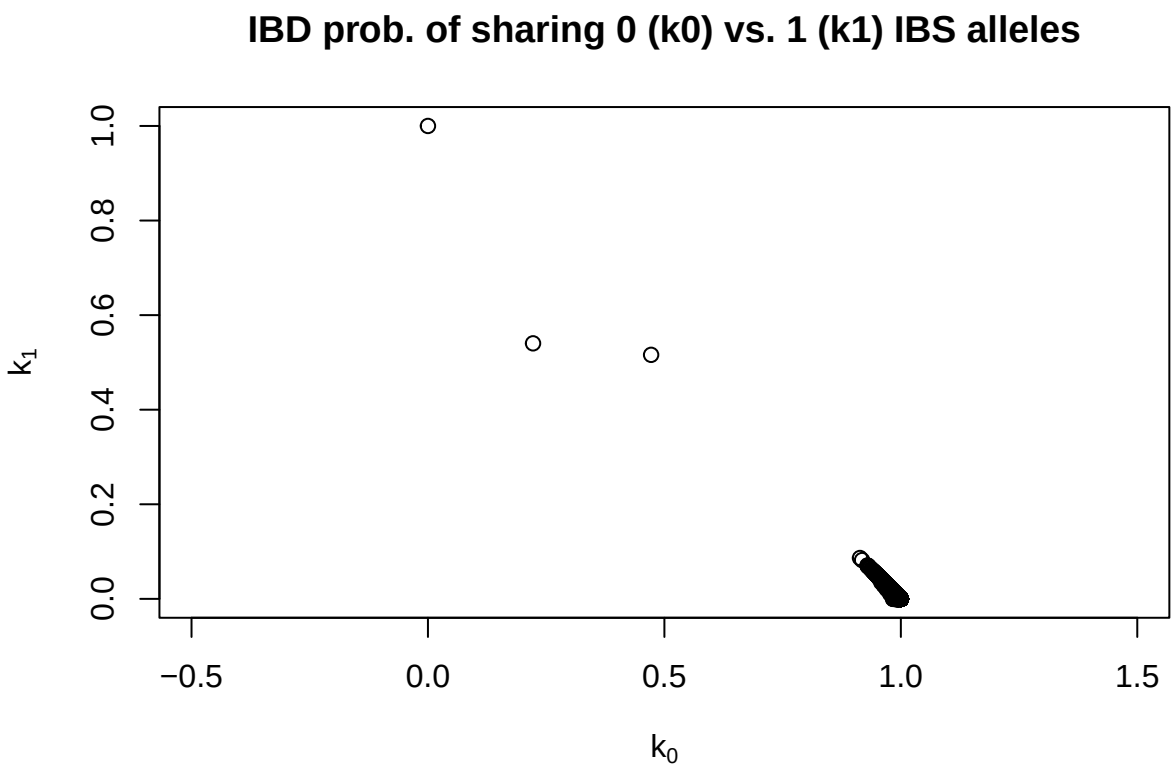
They have a larger allele sharing average, and have at least 40% of their variants with two shared alleles.

Ex 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.

```
system("./plink.exe --bfile CHD --genome --genome-full --out CHD")
```

```
## [1] 0
```

```
CHD.genome <- read.table("CHD.genome",header=TRUE)
plot(CHD.genome$Z0,CHD.genome$Z1,asp=1,xlab=expression(k[0]),ylab=expression(k[1]),main="IBD prob. of sh
```



Ex 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.

```
which(CHD.genome$Z1 >= 0.4, arr.ind = TRUE)
```

```
## [1]  230 1138 4785
```

```
which(CHD.genome$Z0 <= 0.6, arr.ind = TRUE)
```

```
## [1]  230 1138 4785
```

According to k0, k1 results, he following pairs of individuals might have some kind of family relationship (with their Cotterman coefficients):

```
print(paste(CHD.genome[230,1], "and", CHD.genome[230,3], ", with k0=", CHD.genome[230,7], " and k1=", CH
```

```
## [1] "NA17981 and NA17986 , with k0= 0.2222  and k1= 0.5403"
```

```
print(paste(CHD.genome[1138,1], "and", CHD.genome[1138,3], ", with k0=", CHD.genome[1138,7], " and k1="
```

```
## [1] "NA18150 and NA17980 , with k0= 0.4719  and k1= 0.516"
```

```
print(paste(CHD.genome[4785,1], "and", CHD.genome[4785,3], ", with k0=", CHD.genome[4785,7], " and k1="
```

```
## [1] "NA17976 and NA18116 , with k0= 0  and k1= 1"
```

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

```
X[18,1]
```

```
##        18
## -13195.59
```

```
X[3,2]
```

```
##         3
## -565.9969
```

```
X[17,1]
```

```
##      17
## 442.563
```

```
X[12,2]
```

```
##        12
## -337.7617
```

```
X[89,1]
```

```
##       89
## 754.5546
```

```
X[62,2]
```

```
##        62
## -12700.07
```

- "NA17981 and NA17986" have MDS values -13195.59 and -565.9969
- "NA18150 and NA17980" have MDS values 442.563 and 337.7617
- "NA17976 and NA18116" have MDS values 754.5546 and 12700.07

These pairs are the ones that are located the further from the main sub-group in the MDS plot, which means a higher probability of family relationship. This conclusion is the same that we can obtain from (m, s), (p_0, p_2) or (k_0, k_1) plots previously calculated.

Ex 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.

The three representations show the relationships in a similar way. Given that (p_0, p_2) plot and (k_0, k_1) plot leave out one of the three proportions, we would use the (m, s) representation, in which high MAF variants are more informative for discriminating relationship categories.