Practical 6

Statistical Genetics: Relatedness analysis

Anna Putina Marine Mazeau

2023-12-16

```
library(genetics)
library(data.table)
library(SNPRelate)
```

Question 1 : Load the YRI06.raw file into the R environment. How many individuals and how many SNPs are there in the database? What percentage of the data is missing?

```
data <- fread("YRI6.raw")
gen.data <- data[,-c(1:6)]
NA.counter <- sum(is.na(gen.data))
NA.percentage <- NA.counter*100/prod(dim(gen.data))
```

There are 84 individuals and 56574 SNPs in this database.

The percentage of missing data in this database: 0 %.

Question 2 : Compute, for each pair of individuals (and report the first 5), the mean m of the number of alleles shared and the standard deviation s of the number of alleles shared.

```
compute_shared_alleles <- function(individual1, individual2) {
  shared_alleles <- mean(2 - abs(individual1 - individual2))
  return(shared_alleles)
}
shared_alleles_matrix <- matrix(NA, nrow = nrow(gen.data), ncol = nrow(gen.data))

sum(2 - abs(gen.data[1, ] - gen.data[1, ]))

# Compute the number of shared alleles for each pair of individuals
for (i in 1:nrow(gen.data)) {
  for (j in 1:nrow(gen.data)) {
    shared_alleles_matrix[i, j] <- compute_shared_alleles(gen.data[i, ], gen.data[j, ])
  }
}

shared_alleles_matrix <- shared_alleles_matrix
```

It was too long to compute, so we switched to Python at this point.

```r
write.csv(gen.data, file = "gen_data.csv", row.names = TRUE)
write.csv(data, file = "data.csv", row.names = TRUE)
```

Question 6: Use the package SNPRelate to estimate the IBD probabilities, and plot the probabilities of sharing 0 and 1 IBD alleles (k0 and k1) for all pairs of individuals. Use the pedigree information of the `YRI06.raw` file to label the data points in the scatterplot (same as before, one colour for parent-offspring relationship and another colour for unrelated individuals).

```
# Create a gds file
snpgdsCreateGeno("test2.gds", genmat = as.matrix(gen.data),
                 sample.id = data$IID, snp.id = colnames(gen.data),
                 snpfirstdim=FALSE)
# Open the GDS file
genofile <- snpgdsOpen("test2.gds")
# LD pruning
set.seed(10)
snpset <- snpgdsLDpruning(genofile, ld.threshold=0.2)
```

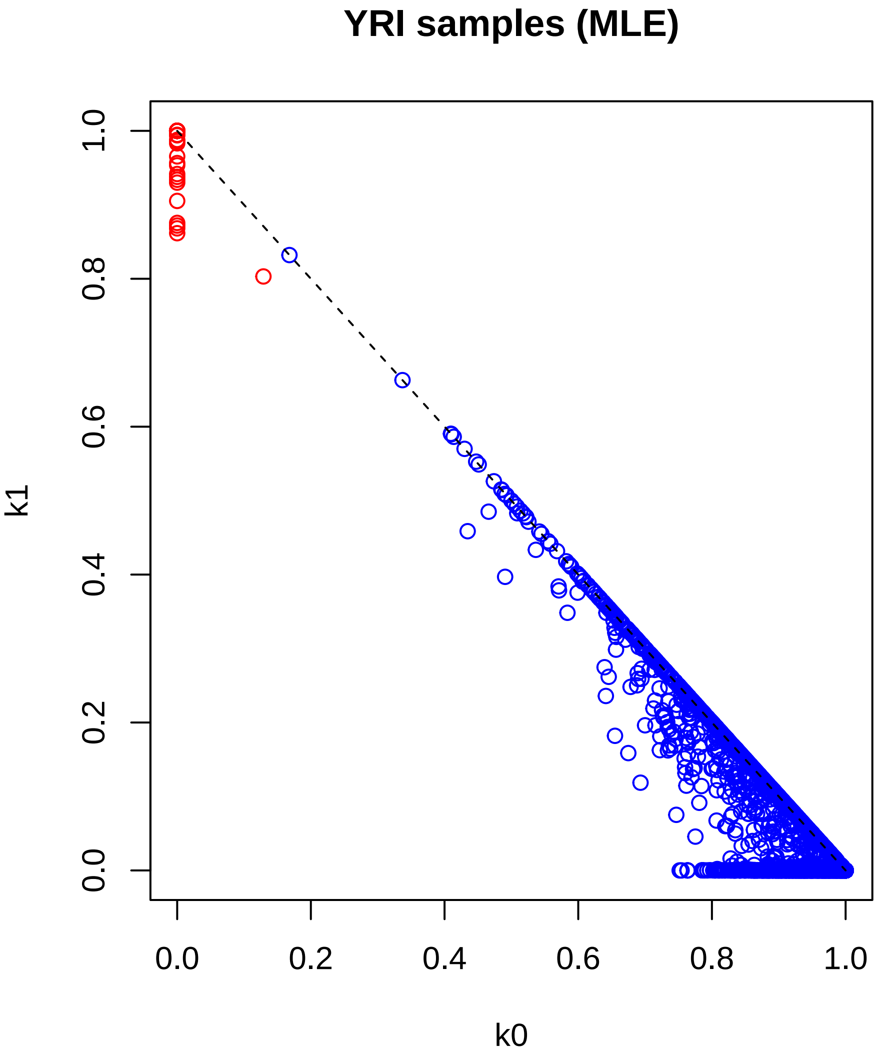
```
## SNP pruning based on LD:
## Excluding 0 SNP on non-autosomes
## Excluding 0 SNP (monomorphic: TRUE, MAF: NaN, missing rate: NaN)
##     # of samples: 84
##     # of SNPs: 56,574
##     using 1 thread
##     sliding window: 500,000 basepairs, Inf SNPs
##     |LD| threshold: 0.2
##     method: composite
## Chromosome 1: 0.12%, 70/56,574
## 70 markers are selected in total.
```

```
snpset.id <- unlist(unname(snpset))
# Estimate IBD coefficients
ibd <- snpgdsIBDMLE(genofile, maf=0.05, missing.rate=0.05,snp.id=snpset.id, num.thread=2)
```

```
## Identity-By-Descent analysis (MLE) on genotypes:
## Excluding 56,504 SNPs (non-autosomes or non-selection)
## Excluding 0 SNP (monomorphic: TRUE, MAF: 0.05, missing rate: 0.05)
##     # of samples: 84
##     # of SNPs: 70
##     using 2 threads
## MLE IBD:    the sum of all selected genotypes (0,1,2) = 5330
## MLE IBD: Tue Dec 19 16:29:55 2023    0%
```

```
# Make a data.frame
ibd.coeff <- snpgdsIBDSelection(ibd)

# Parent-offspring relationship info
ped_info <- data[, c(2, 3, 4)]

col_id1 <- ibd.coeff$ID1
col_id2 <- ibd.coeff$ID2

parent_offspring <- function(id1, id2) {
  if (id2 %in% ped_info[ped_info$IID == id1, c("PAT", "MAT")]) {
    relationship <- TRUE
  } else if (id1 %in% ped_info[ped_info$IID == id2, c("PAT", "MAT")]) {
    relationship <- TRUE
```

```r
  } else {
    relationship <- FALSE
  }
  return(relationship)
}

result_relationship <- mapply(parent_offspring, col_id1, col_id2)

# Create a vector of colors based on family relationship
colors <- ifelse(result_relationship, "red","blue")

# Plot
plot(ibd.coeff$k0, ibd.coeff$k1, xlim=c(0,1), ylim=c(0,1),
     xlab="k0", ylab="k1", main="YRI samples (MLE)", col=colors)
lines(c(0,1), c(1,0), lty=2)
```



Question 7: Do you think the family relationships between all individuals were correctly specified?

The family relationships between all individuals appear to have been correctly specified (*except for 1 misidentification with k0 > 0*).

This is evident in the plot where k0 = 0 and k1 > 0.5, indicating a high probability of sharing 1 IBD allele. These patterns align with expectations for parent-offspring relationships, providing strong support for the accuracy of the specified family relationships. The visual representation in the plot reinforces the consistency between the estimated IBD coefficients and the expected relationships, further validating the correctness of the familial information in the dataset.