Project 1

Ezgi Karaesmen

October 5, 2016

Answers

Question 1

Hardy-Weinberg equilibrium is used to estimate genotype and allelic frequencies, under the assumptions of infinite population, random mating, no genetic forces, equal genotype frequencies in both sexes and genotype frequencies are constant after one generation of mating. Testing for HWE can be useful to detect genotyping errors, population stratification and suggestive association with a certain outcome. Testing for HWE can be done using the Chi-Square method, by comparing the observed and HWE-expected genotype frequencies. The resulting test statistic would have a χ_1^2 distribution with degrees of freedom depending on the number of genotypes for the locus under testing. The hypotheses for this case then would be:

H_0 : The locus is in HWE
H_a : The locus deviates from HWE

With an $\alpha = 0.05$, for a $p - value < 0.05$ H_0 is rejected and H_a is accepted. The HWE test was applied to both GIH and MEX populations via `HWE.chisq` function, p-values were extracted and filtered for < 0.05. In the GIH population 3 SNPs that were detected to be deviating from HWE were: rs9617528 rs7288972 rs7293121. In the MEX population only SNP that was deviating from HWE was rs7155494. Determining that different loci are deviating from HWE in these populations indicated that the genetic structure of the population is different. Therefore analyzing these populations jointly is inappropriate and could generate false positive of false negative results.

Question 2

Linkage disequilibrium (LD) is the non-random association of alleles at different loci. Presence of LD can be determined by calculating Linkage Disequilibrium Coefficient (D) and its normalized measures D', r^2 (or Δ^2), which indicate the strength of the LD. D for alleles A and B at 2 respective loci is defined as $D_{AB} = P_{AB} - P_A \cdot P_B$ and for the cases of no LD, $D_{AB} = O$. Therefore, the hypotheses to be tested are defined as:

$H_0 : D_{AB} = 0$
$H_a : D_{AB} \neq 0$

The estimate of D_{AB} is $\hat{D}_{AB} = \hat{P}_{AB} - \hat{P}_A \cdot \hat{P}_B$ and under H_0, $\hat{D}_AB$ has mean 0 and variance of $\frac{1}{2n}(\hat{P}_A, \hat{P}_a, \hat{P}_B, \hat{P}_b)$ the score test statistic is then defined as:

$$T = \frac{\hat{D}_{AB}}{\sqrt{\frac{1}{2n}(\hat{P}_A, \hat{P}_a, \hat{P}_B, \hat{P}_b)}} \quad asymp \quad \sim \quad N(0,1)$$

With $\alpha = 0.05$, for a $P - value < 0.05$, H_0 is rejected and H_a is accepted, indicating that the tested loci pair is in LD.

Approximately 15 SNPs that showed significant LD with other SNPs were common between GIH and MEX populations. Significant LD pairs for 20 SNPs and their LD pairs for both populations are presented in Table 1. It can be seen that many common SNPs are in LD between the populations, but have different LD pairs.

Table 1: LD pairs for 20 SNPs in GIH and MEX populations. Rows indicate each SNP that is tested for LD and SNPs in columns GIH and MEX indicate the detected LD pairs for that SNP in the respective population.

	GIH	MEX
rs2334386	rs8140723, rs7354790, rs12163493	rs7293121
rs12628452	rs12163493	rs6423472
rs7289830	rs9617528, rs6423472, rs7293121	rs9617528, rs7354790, rs2334336, rs9604698, rs2845199, rs8138488
rs9617528	rs6423472, rs2334336, rs7293121, rs9617337, rs8138488, rs9617160	rs7354790, rs2334336, rs11089128, rs9604698, rs2845199, rs9617337, rs8138488, rs9617160
rs10154759	rs11167319, rs9617160	rs7288972, rs9604698, rs9617337, rs9617160
rs8140723	rs7354790, rs2845199, rs9617337, rs9617160	rs9604698, rs9617337, rs9617160
rs7354790	rs2334336, rs11089128, rs715549, rs9617337, rs11167319, rs8138488	rs7288972, rs9604698, rs2845199, rs9617337
rs6423472	rs7288972, rs7293121, rs9617337, rs9617160	rs9617337, rs8138488, rs9617160
rs2334336	rs11089128, rs12163493, rs8138488	rs7288972, rs9604698, rs11167319
rs11089128	rs11167319, rs8138488	rs9617337, rs9617160
rs7288972	rs715549, rs2845199, rs8138488	rs9604698, rs9617337, rs9617160
rs7293121	rs2845199, rs9617337, rs9617160	rs9617160
rs715549	rs9617337, rs9617160	rs2845199
rs9604698		rs2845199, rs9617337, rs9617160
rs2845199		rs9617337, rs8138488, rs9617160
rs9617337	rs9617160	rs8138488, rs9617160
rs12163493		rs11167319
rs11167319	rs8138488	rs8138488
rs8138488		
rs9617160		

The common SNPs that in LD with common LD pairs for both populations are shown in Table 2. 9 SNPs were common for both populations and had at least one matching pair. SNPs rs9617528, rs9617337 and rs9617160 occur on Table 2 multiple times, indicating certain regions on chromosome 22 are in strong LD and likely to be conserved in both populations. Assuming that the genotype sampling accurately represents these populations, finding common haplotypes between GIH and MEX can be interesting. GIH and MEX populations are expected to have none or very low gene flow due to geographic isolation from each other. Therefore, finding common LD structures might suggest that these haplotypes were evolutionarily conserved since the divergence from last common ancestor. However, it is difficult to comment on the LD structure only by the p-values. Visualizing the D' or r^2 values would help us better understand the region.

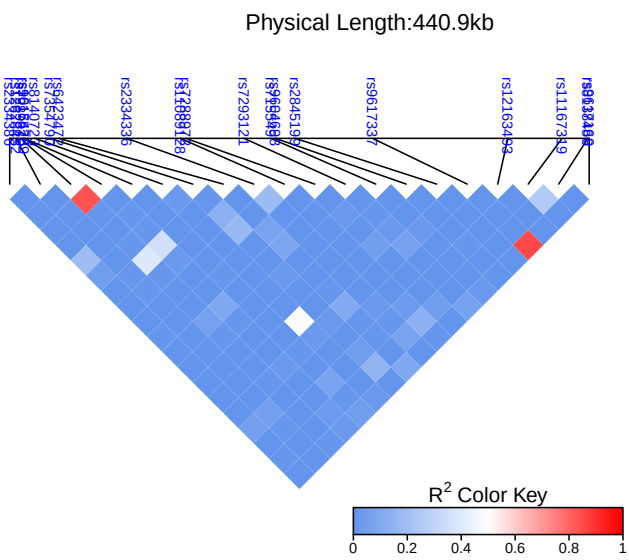
Table 2: SNPs with common LD pairs

	Common SNPs with LD	Common LD pairs in both populations
1	rs7289830	rs9617528
2	rs9617528	rs2334336, rs9617337, rs8138488, rs9617160
3	rs10154759	rs9617160
4	rs8140723	rs9617337, rs9617160
5	rs7354790	rs9617337
6	rs6423472	rs9617337, rs9617160
7	rs7293121	rs9617160
8	rs9617337	rs9617160
9	rs11167319	rs8138488

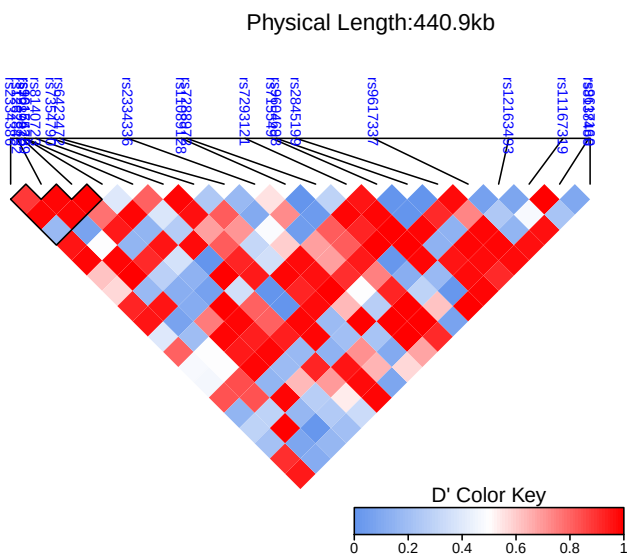
Question 3

LD heatmaps of the GIH and MEX populations are produced for both r^2 and D'. Both of these plots show inconsistent, and similar LD patterns for both populations. LD Heatmap with r^2 does not indicate any significant LD pattern, but suggests significance for the rs9617337-rs9617160 and rs2334386-rs7289830 pairs. These pairs also have significant p-values in both of the populations as shown in Table 1 and 2. However, no signs of LD is seen for other loci in close proximity to these SNPs. Therefore, the signal is questionable and none of the pairs were highlighted for r^2. On the other hand D' LD heatmap does show strong LD pairs, but these pairs do not seem to be in close proximity and no distinct patterns can be observed. Nevertheless, rs2334386, rs12628452, rs7289830 and rs9617528 pairs indicate some plausible degree of LD, and are also significant LD pairs for GIH as well as MEX as shown in Table 1 and 2. Hence the area covering these three SNPs were highlighted. MEX population also show another LD region for rs2845199, rs9617337 and rs12163493 pairs. However, `LDheatmap.highlight` function's `i` and `j` index arguments does not accept numeric values for multiple highlights, hence the region could not be highlighted. These rs2845199, rs9617337 and rs12163493 pairs also show significance for LD in the MEX population. Although LD patterns are very difficult to interpret for this example, results seem to be consistent with those in question 2.

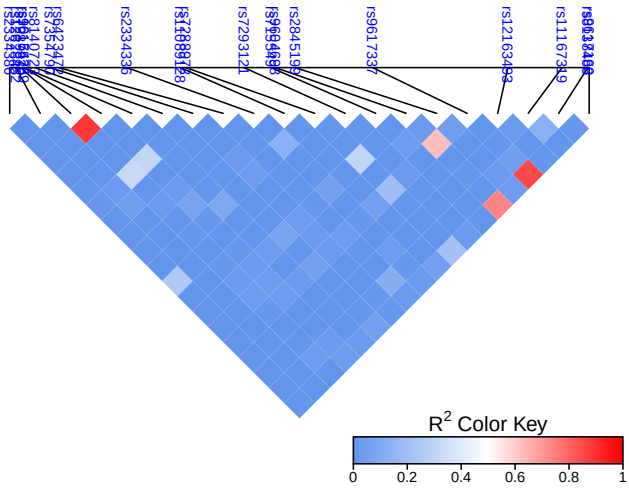
GIH LD Heatmap − r^2



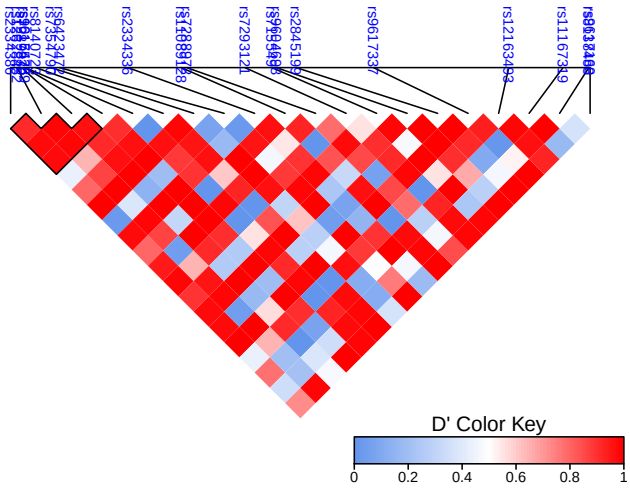
GIH LD Heatmap − D'



MEX LD Heatmap − r^2



MEX LD Heatmap − D'



Question 4

The list of SNP pairs that have the strongest LD value, their physical and genomic distances are shown in Table 3. For the computation of genomic distance $1M \sim 1Mb$ was assumed and was shown in cM.

Table 3: Maximum LD values and their relevant SNP pairs

Population	LD coefficient	SNP Pair	Max LD Value	Physical Lenght (BP)	Genomic Lenght (cM)
GIH	r^2	rs9617337 - rs9617160	0.857	164834	16.483
GIH	D'	rs7289830 - rs9617528	1	2258	0.226
MEX	r^2	rs7289830 - rs9617528	0.889	2258	0.226
MEX	D'	rs9604698 - rs9617160	0.999	238666	23.867

Question 5

For r^2 LD heatmaps of GIH and MEX populations, LD structures are quite similar, as both of the heatmaps suggest weak LD for majority of the SNP pairs, but strong LD for rs7289830 - rs9617528 and rs9617337 - rs9617160. Although LD patterns are much more difficult to interpret for D' heatmaps, no major difference has been observed between two populations. Furthermore, D' coefficient suggests strong LD for rs2334386, rs12628452, rs7289830 and rs9617528 pairs for both populations. Only plausible difference between the populations is that MEX D' LD heatmap suggests strong LD between rs2845199, rs9617337 and rs12163493 pairs and GIH D' LD heatmap does not.

Appendix - R codes

Cleaning and Reshaping the data

The given population data require editing and cleaning to be compatible with the functions will be used in packages `genetics` and `LDheatmap`.

```
wd = "~/Box Sync/Stat Anal of Genetic Data/Project1/"
setwd(wd)

library(genetics)
library(LDheatmap)

#### Load datasets & transform the data to the required format #####

# Gujarati Indian SNP data
gih <- read.table("HapMap_chrom22_20SNP_GIH.txt")

# look at data structure
dim(gih) # n sample = 101
gih[1:3, 1:10]

unique(gih$V8) # whole column 8 is replaced with TRUE istead if "T"
gih$V8 <- "T"  # fix the problem with TRUE on column 8
gih[1:3, 1:10]

# Mexican SNP data
mex <- read.table("HapMap_chrom22_20SNP_MEX.txt")

# look at data structure
dim(mex)  # n sample = 86
mex[1:3, 1:10]

# Create an empty matrix to store the two alleles of a SNP into one column
gih.geno <- data.frame(matrix(nrow=nrow(gih), ncol=ncol(gih)/2))
mex.geno <- data.frame(matrix(nrow=nrow(mex), ncol=ncol(mex)/2))

# Replace N with empty string
combine.snp <- function(snp1, snp2){
  snp1 <- as.character(snp1)
  snp2 <- as.character(snp2)
  snp1[snp1=="N"] <- ""
  snp2[snp2=="N"] <- ""
  snp.geno <- genotype(snp1,snp2)
  return(snp.geno)
}

k <- 1
for(i in 1:ncol(gih.geno)){
  gih.geno[,i] <- combine.snp(gih[,k], gih[,k+1])
  k <- k+2
}
```

```
k <- 1
for(i in 1:ncol(mex.geno)){
  mex.geno[,i] <- combine.snp(mex[,k], mex[,k+1])
  k <- k+2
}

## name the columns the proper rsIDs of the 20 SNPs
snp.info <- read.xls("HapMap_chrom22_20SNP_info.xls")
head(snp.info,3)

colnames(gih.geno) <- as.character(snp.info$SNP.name)
colnames(mex.geno) <- as.character(snp.info$SNP.name)

gih.geno[1:2, 1:4]
mex.geno[1:2, 1:4]
```

Question 1

```
# ?HWE.chisq

## GIH ##
gih.HWE.pval = NULL
for (i in 1:ncol(gih.geno)) gih.HWE.pval[i]=HWE.chisq(gih.geno[,i], B= 100000)$p.value

# SNPs deviating from HWE for GIH
cat("SNP(s) that deviate(s) from HWE in the GIH population:",
    colnames(gih.geno)[gih.HWE.pval < 0.05], sep="\n")

## MEX ##
mex.HWE.pval = NULL
for (i in 1:ncol(mex.geno)) mex.HWE.pval[i]=HWE.chisq(mex.geno[,i], B=100000)$p.value

# SNPs deviating from HWE for MEX
cat("SNP(s) that deviate(s) from HWE in the GIH population:",
    colnames(mex.geno)[mex.HWE.pval < 0.05], sep="\n")
```

Question 2

```
?LD

## GIH ##
gih.LD = LD(gih.geno)
# names(gih.LD)
# head(gih.LD$`R^2`, 3)
# head(gih.LD$`P-value`, 3)

gih.LD.SNPs <- apply(gih.LD$`P-value`, 1,  function(x) names(which(x < 0.05)) )

# LD pairs for GIH
```

```
gih.LD.pairs = NULL
for(i in 1:length(gih.LD.SNPs)) gih.LD.pairs[i] = toString(gih.LD.SNPs[[i]])

## MEX ##
mex.LD = LD(mex.geno)
mex.LD.SNPs <- apply(mex.LD$`P-value`, 1,  function(x) names(which(x < 0.05)))

# LD pairs for MEX
mex.LD.pairs = NULL
for(i in 1:length(mex.LD.SNPs)) mex.LD.pairs[i] = toString(mex.LD.SNPs[[i]])

# Table 1
LD.df = data.frame(GIH = gih.LD.pairs, MEX = mex.LD.pairs, row.names = colnames(gih.geno) )
LD.df.tbl = xtable(LD.df, caption = "LD pairs for 20 SNPs in GIH and MEX populations")
print.xtable(LD.df.tbl, size ="small")


# any common SNPs that show LD?
common.LD = list()
for(i in 1:length(gih.LD.SNPs)){
  common.LD[[names(gih.LD.SNPs)[i]]] = gih.LD.SNPs[[i]][gih.LD.SNPs[[i]] %in% mex.LD.SNPs[[i]]]
}
common.LD <- common.LD[lapply(common.LD, length) > 0]
common.LD.pairs = NULL
for(i in 1:length(common.LD)) common.LD.pairs[i] = toString(common.LD[[i]])
common.LD.df <- data.frame(Common_SNPS_that_show_LD= names(common.LD) ,
                           common.LD.pairs =common.LD.pairs , row.names=NULL)

# Table 2
xtable(common.LD.df, caption="SNPs with common LD pairs")
```

Question 3

```
rgb.palette <- colorRampPalette(rev(c("cornflowerblue", "white", "red")), space = "rgb")

## GIH
# r2
gih.r <- LDheatmap(gih.geno, snp.info$Position,
          color=rgb.palette(100), SNP.name = as.character(snp.info$SNP.name),
          title = expression(paste("GIH LD Heatmap - " ,r^2)), flip=T  )
# D'
gih.d <- LDheatmap(gih.geno, snp.info$Position,
          color=rgb.palette(100), SNP.name = as.character(snp.info$SNP.name),
          LDmeasure = "D'",
          title = expression(paste("GIH LD Heatmap - D'")) , flip=T)
LDheatmap.highlight(gih.d, i=1, j=4, col="red", fill="grey")

## MEX
# r2
mex.r <- LDheatmap(mex.geno, snp.info$Position,
          color=rgb.palette(100), SNP.name = as.character(snp.info$SNP.name),
```

```
          title = expression(paste("MEX LD Heatmap - " ,r^2)) , flip=T )

# D'
mex.d <- LDheatmap(mex.geno, snp.info$Position,
          color=rgb.palette(100), SNP.name = as.character(snp.info$SNP.name),
          LDmeasure = "D'",
          title = expression(paste("MEX LD Heatmap - D'")) , flip=T)
mex.d <- LDheatmap.highlight(mex.d, i=1, j=4, col="red", fill="grey")
```

Question 4

```
LDmats <- list(gih.LD.r$LDmatrix,
               gih.LD.d$LDmatrix,
               mex.LD.r$LDmatrix,
               mex.LD.d$LDmatrix   )

max.LD <- function(m){
mmax = max(m, na.rm = T)
ind <- which(m == mmax, arr.ind = TRUE)
mmax = round(mmax,3)
pair <- paste0(dimnames(m)[[1]][ind[1]], " - ", dimnames(m)[[2]][ind[2]])
LDdist <- abs(snp.info$Position[ind[1]] - snp.info$Position[ind[2]])
LDdist.cM <- round((LDdist/1e+6)*100, 3)
c(pair, mmax, LDdist, LDdist.cM)
}

LDmax.tbl <- lapply(LDmats, max.LD)
LDmax.tbl <- data.frame(do.call(rbind, LDmax.tbl))
colnames(LDmax.tbl) <- c("SNP Pair", "Max LD Value",
                         "Physical Lenght (BP)", "Genomic Lenght (cM)")
LDmax.tbl$Population <- c("GIH", "GIH", "MEX", "MEX")
LDmax.tbl$`LD coefficient` <- c("$r^2$", "$D'$", "$r^2$", "$D'$")
kable(LDmax.tbl[,c(5:6, 1:4)],
      caption = "Maximum LD values and their relevant SNP pairs", align = 'c')
```