Untitled

Ryo uchimido

10/31/2017

```r
library(tidyr)
```

separate(sumstats,SNP_HGLT,c(“chr”,“pos”),sep=“”)

Introduction

You’ve recently discovered that there are in fact two types of cholesterol– both good (HDL) and bad (LDL). You are worried that you may have a genetic predisposition to having high levels of bad cholesterol and decide to `investigate what genes may be associated with bad cholesterol`. Here, we’ll do a basic exploration of the data before investigating the specific genetic effects in subsequent lectures.

Problem 1

** -Access the LDL summary statistics from the GLGC Consortium’s genome-wide association study. -Visualize the summary statistics both as a Manhattan Plot and as a Q-Q plot. -What `chromosome(s)` appear to have genome-wide significant hits?

GWAS Summary Statistics

1 jointGwasMc_LDL.txt.gz

```r
pathToDataFile <- "/Users/uchimidouryou/Documents/HSPH:MIT:Catalyst/BST227/HW1_1031/jointGwasMc_LDL.txt
sumstats1 <- as.data.frame(data.table::fread(paste0("zcat < ", pathToDataFile), showProgress = FALSE))
head(sumstats1)
```

```
##           SNP_hg18        SNP_hg19       rsid A1 A2   beta     se     N
## 1   chr10:10000135   chr10:9960129 rs4747841  a  g 0.0037 0.0052 89138
## 2   chr10:10000265   chr10:9960259 rs4749917  c  t 0.0033 0.0052 89138
## 3 chr10:100002729 chr10:100012739  rs737656  a  g 0.0099 0.0054 89888
## 4 chr10:100002880 chr10:100012890  rs737657  a  g 0.0084 0.0054 89888
## 5 chr10:100003553 chr10:100013563 rs7086391  c  t 0.0075 0.0067 89888
## 6 chr10:100003805 chr10:100013815  rs878177  c  t 0.0073 0.0055 89888
##   P-value Freq.A1.1000G.EUR
## 1 0.71580            0.4908
## 2 0.77480            0.4908
## 3 0.04000            0.3206
## 4 0.08428            0.3206
## 5 0.26890            0.7810
## 6 0.13760            0.6517
```

```r
newdt1<-separate(sumstats1,SNP_hg18,c("CHR","POS"),sep=":")
```

```
## Warning: Too few values at 3 locations: 2251295, 2311852, 2396132
```

```r
head(newdt1)
```

```
##     CHR       POS        SNP_hg19      rsid A1 A2   beta     se     N
## 1 chr10  10000135   chr10:9960129 rs4747841  a  g 0.0037 0.0052 89138
## 2 chr10  10000265   chr10:9960259 rs4749917  c  t 0.0033 0.0052 89138
## 3 chr10 100002729 chr10:100012739  rs737656  a  g 0.0099 0.0054 89888
## 4 chr10 100002880 chr10:100012890  rs737657  a  g 0.0084 0.0054 89888
## 5 chr10 100003553 chr10:100013563 rs7086391  c  t 0.0075 0.0067 89888
## 6 chr10 100003805 chr10:100013815  rs878177  c  t 0.0073 0.0055 89888
##   P-value Freq.A1.1000G.EUR
## 1 0.71580            0.4908
## 2 0.77480            0.4908
## 3 0.04000            0.3206
## 4 0.08428            0.3206
## 5 0.26890            0.7810
## 6 0.13760            0.6517
```

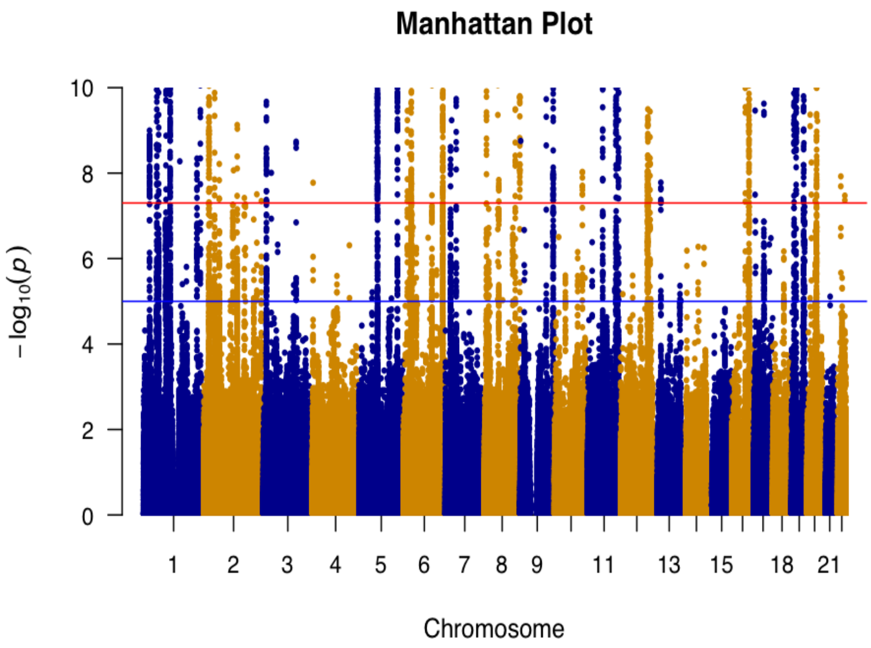
```r
newdt1$chr2 <- as.numeric(gsub("chr","",newdt1$CHR))
```

```
## Warning: NAs introduced by coercion
```

```r
newdt1$pos2 <- as.numeric(gsub("POS","",newdt1$POS))
newdt1$snp2 <- as.character(newdt1$rsid)
newdt1$P <- newdt1$`P-value`
head(newdt1)
```

```
##     CHR       POS        SNP_hg19      rsid A1 A2   beta     se     N
## 1 chr10  10000135   chr10:9960129 rs4747841  a  g 0.0037 0.0052 89138
## 2 chr10  10000265   chr10:9960259 rs4749917  c  t 0.0033 0.0052 89138
## 3 chr10 100002729 chr10:100012739  rs737656  a  g 0.0099 0.0054 89888
## 4 chr10 100002880 chr10:100012890  rs737657  a  g 0.0084 0.0054 89888
## 5 chr10 100003553 chr10:100013563 rs7086391  c  t 0.0075 0.0067 89888
## 6 chr10 100003805 chr10:100013815  rs878177  c  t 0.0073 0.0055 89888
##   P-value Freq.A1.1000G.EUR chr2      pos2      snp2       P
## 1 0.71580            0.4908   10  10000135 rs4747841 0.71580
## 2 0.77480            0.4908   10  10000265 rs4749917 0.77480
## 3 0.04000            0.3206   10 100002729  rs737656 0.04000
## 4 0.08428            0.3206   10 100002880  rs737657 0.08428
## 5 0.26890            0.7810   10 100003553 rs7086391 0.26890
## 6 0.13760            0.6517   10 100003805  rs878177 0.13760
```

```r
newdt1 <- newdt1[newdt1$P > 10^(-100),]
newdt1 <- newdt1[complete.cases(newdt1),]
qqman::manhattan(newdt1, chr="chr2",bp = "pos2",p="P-value", snp="snp2", main = "Manhattan Plot", ylim =
```



The Chromosome number of 1,2,3,4,5,6,7,8,9,10,11,12,13,16,17,19,20,22 appear to have genome-wide significant hits.

QQ Plot

```
qqman::qq(newdt1$P)
```

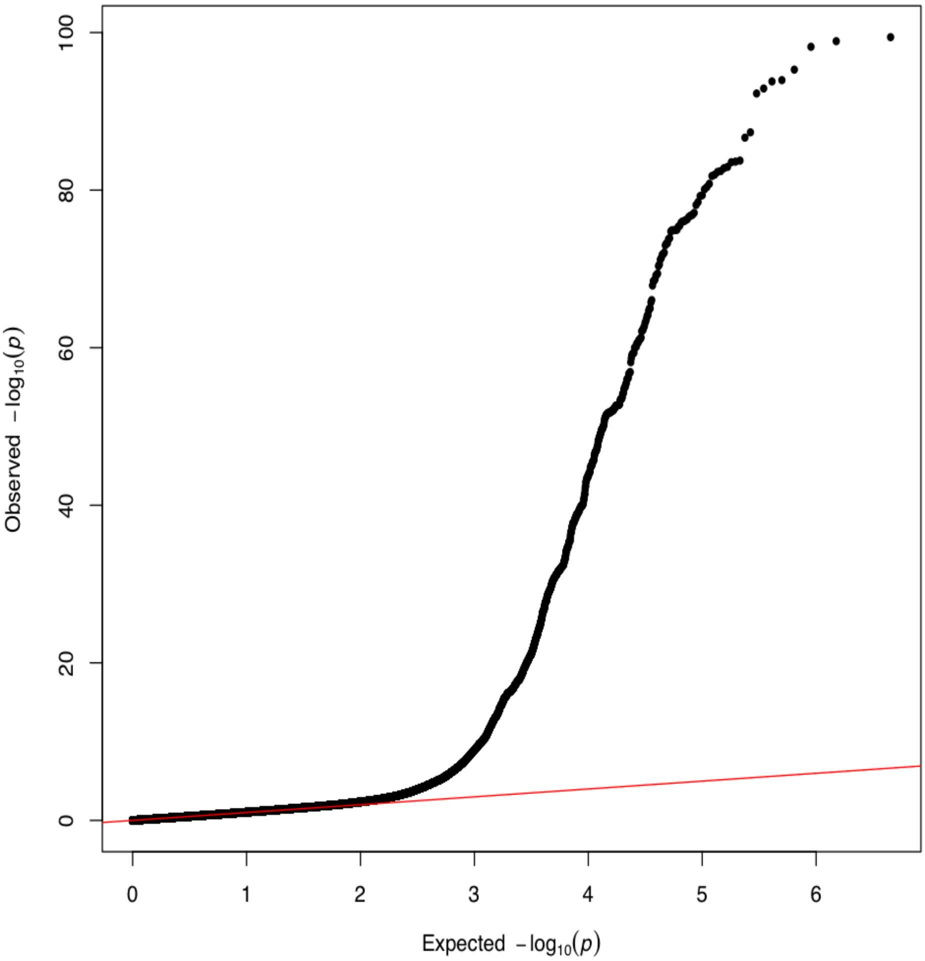


Figure 1: qqplot of summary statistics

Problem 2

As part of the GLGC Consortium, the group analyzed data for a different SNP array, the Metabochip. Visualize the summary statistics both as a Manhattan Plot and as a Q-Q plot. What chromosome(s) appear to have genome-wide significant hits?

Hint: use the Metabochip summary statistics can be found on the same page as the GWAS summary statistics

the Metabochip

1 Mc_LDL.txt.gz

```r
pathToDataFile <- "/Users/uchimidouryou/Documents/HSPH:MIT:Catalyst/BST227/HW1_1031/Mc_LDL.txt.gz"
sumstats2 <- as.data.frame(data.table::fread(paste0("zcat < ", pathToDataFile), showProgress = FALSE))
```

```
## Warning in data.table::fread(paste0("zcat < ", pathToDataFile),
## showProgress = FALSE): C function strtod() returned ERANGE for one or
## more fields. The first was string input '2.07e-651'. It was read using
## (double)strtold() as numeric value 0.0000000000000000E+00 (displayed here
## using %.16E); loss of accuracy likely occurred. This message is designed
## to tell you exactly what has been done by fread's C code, so you can search
## yourself online for many references about double precision accuracy and
## these specific C functions. You may wish to use colClasses to read the
## column as character instead and then coerce that column using the Rmpfr
## package for greater accuracy.
```

```r
head(sumstats2)
```

```
##         SNP_hg18       SNP_hg19       rsid A1 A2   beta     se        N
## 1  chr11:8209625  chr11:8253049   rs110420  c  t 0.0034 0.0051 83030.00
## 2 chr12:69875488 chr12:71589221 rs12227602  t  a 0.0099 0.0124 83102.82
## 3 chr15:92998082 chr15:95197078 rs12442791  g  a 0.0146 0.0112 83118.56
## 4 chr1:167364087 chr1:169097463  rs2000321  g  a 0.0038 0.0085 71499.02
## 5  chr6:29632380  chr6:29524401  rs2745412  t  c 0.0052 0.0091 74156.00
## 6 chr16:52193910 chr16:53636409  rs4784321  c  t 0.0042 0.0064 83097.03
##   P-value Freq.A1.1000G.EUR
## 1  0.5275           0.50260
## 2  0.4332           0.95515
## 3  0.2261           0.95383
## 4  0.6303           0.87990
## 5  0.7921           0.09235
## 6  0.7083           0.77700
```

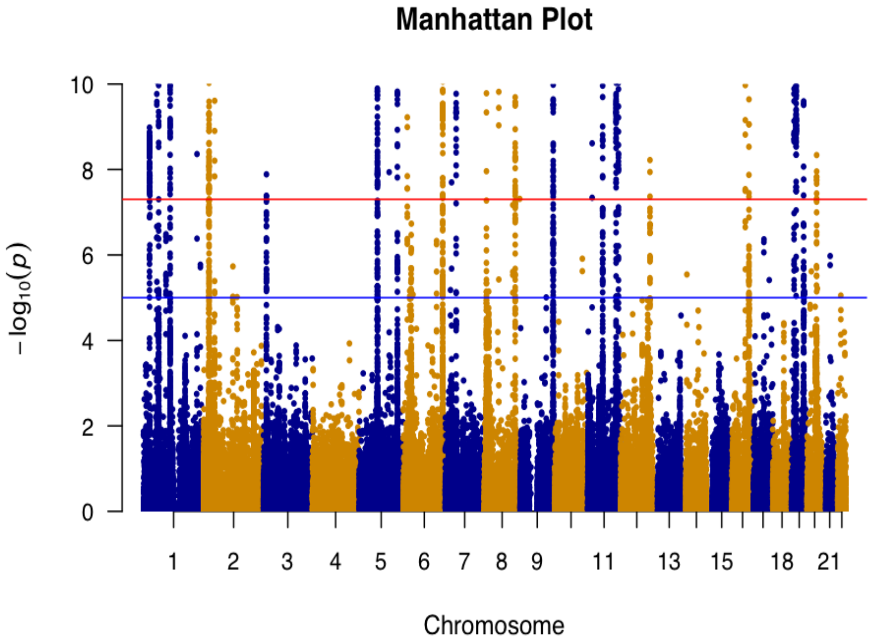
```r
newdt2<-separate(sumstats2,SNP_hg18,c("CHR","POS"),sep=":")
head(newdt2)
```

```
##     CHR       POS       SNP_hg19       rsid A1 A2   beta     se        N
## 1 chr11   8209625  chr11:8253049   rs110420  c  t 0.0034 0.0051 83030.00
## 2 chr12  69875488 chr12:71589221 rs12227602  t  a 0.0099 0.0124 83102.82
## 3 chr15  92998082 chr15:95197078 rs12442791  g  a 0.0146 0.0112 83118.56
## 4  chr1 167364087 chr1:169097463  rs2000321  g  a 0.0038 0.0085 71499.02
## 5  chr6  29632380  chr6:29524401  rs2745412  t  c 0.0052 0.0091 74156.00
## 6 chr16  52193910 chr16:53636409  rs4784321  c  t 0.0042 0.0064 83097.03
##   P-value Freq.A1.1000G.EUR
## 1  0.5275           0.50260
## 2  0.4332           0.95515
## 3  0.2261           0.95383
## 4  0.6303           0.87990
## 5  0.7921           0.09235
## 6  0.7083           0.77700
```

```
newdt2$chr2 <- as.numeric(gsub("chr","",newdt2$CHR))
newdt2$pos2 <- as.numeric(gsub("POS","",newdt2$POS))
newdt2$P <- newdt2$`P-value`
```

```
newdt2 <- newdt2[newdt2$P > 10^(-100),]
newdt2 <- newdt2[complete.cases(newdt2),]
qqman::manhattan(newdt2, chr="chr2",bp = "pos2",p="P-value", main = "Manhattan Plot", ylim = c(0, 10), 
    cex.axis = 0.9, col = c("blue4", "orange3") )
```

```
## Warning in qqman::manhattan(newdt2, chr = "chr2", bp = "pos2", p = "P-
## value", : No SNP column found. OK unless you're trying to highlight.
```



##

The Chromosome number of 1,2,3,5,6,7,8,9,11,12,16,19,20 appear to have genome-wide significant hits.

QQ Plot

```
qqman::qq(newdt2$P)
```

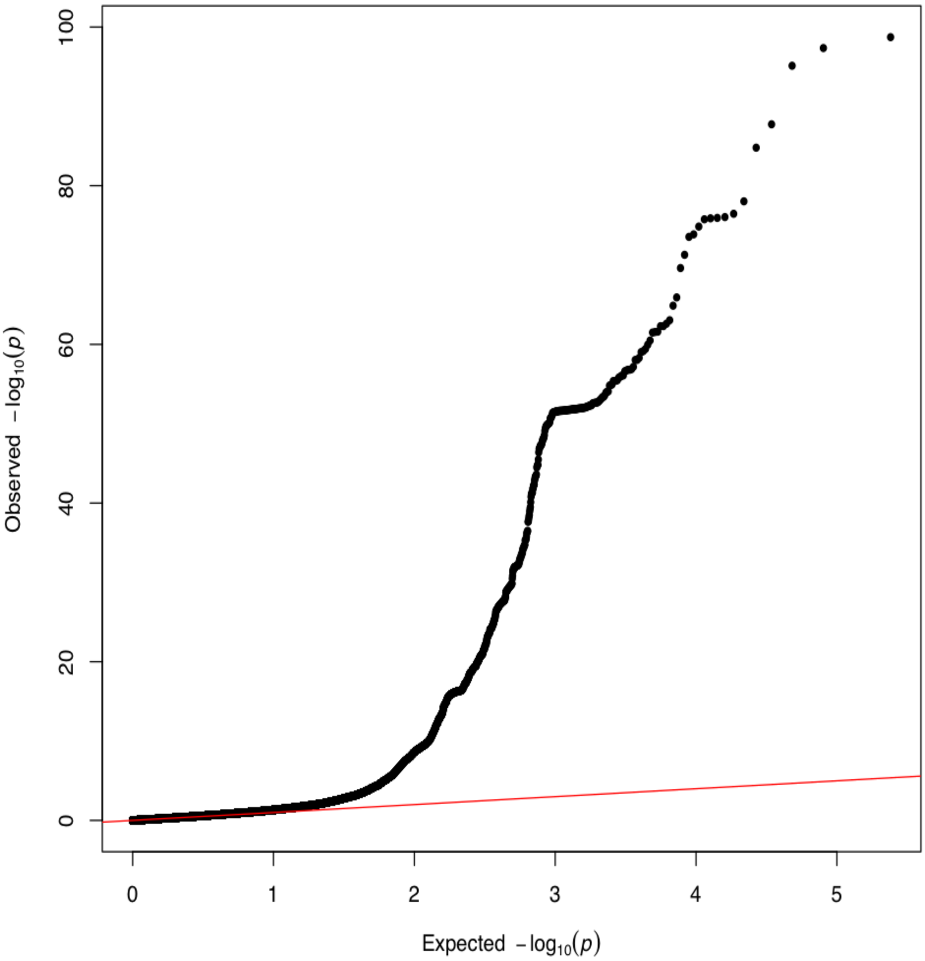


Figure 2: qqplot of summary statistics

Problem 3

Compute the measure of systematic inflation (λ_{GC}) associatd with the summary statistics in Problem 1 and Problem 2. For which SNP array are the summary statistics more inflated?

Hint: $\lambda_{GC} = \text{median}(\chi^2)$ / 0.4549364 *where the last number comes from* qchisq(0.5,1)

A measure of systemic inflation is genomic inflation factor, also known as lambda.

```
# Calculating lambda of the summary statistics in problem 1
chisq1 <- qchisq(1-newdt1$P,1)
lambda1 = median(chisq1)/qchisq(0.5,1)
lambda1
```

```
## [1] 1.015011
```

```
# Calculating lambda of the summary statistics in problem 2
chisq2 <- qchisq(1-newdt2$P,1)
lambda2 = median(chisq2)/qchisq(0.5,1)
lambda2
```

```
## [1] 1.222021
```

From the result above, the lambda in problem 2 is greater than those in problem 1. SNP array in problem 2 appears mor e inflated.