Gulishana Adilijiang (adiliji2)

STAT 530 Bioinformatics: Homework 4

Problem1:

Perform the non-parametric goodness-of-fit test: **Two-sample Kolmogorov-Smirnov Test**

Null Hypothesis: The two data samples are drawn from the same null distribution.

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution functions of two samples. The empirical distribution is the estimation of one data sample's cumulative distribution function.

In this case, suppose $\boldsymbol{n_A}$ and $\boldsymbol{n_B}$ are the number of *cis*-eQTLs in class A and B respectively. The null hypothesis is that the physical locations of the eQTLs in class A and B are the samples that come from a same null distribution.

Suppose the empirical distribution (estimation of CDF) of the distance between the marker and the gene is $\boldsymbol{F_A(x)}$ and $\boldsymbol{F_B(x)}$ in class A and B respectively.

Then the **Test Statistic** is $\boldsymbol{Dn_A,n_B = sup_x |F_A(x) - F_B(x)|}$

The sup_x means a supremum function.

Reject the null hypothesis at the significance level α, if $Dn_A,n_B > c(\alpha)*\text{sqrt}((n_A+n_B)/n_A,n_B)$, and in general $c(\alpha) = \text{sqrt}(-1/2 *\ln(\alpha/2))$.

Problem2:

According to the marker data, there are **34373** markers.
In each brain region, there are **26493** transcripts.
And the most significant eQTLs in each brain region is shown below:

Brain Region	Most Significant eQTL Pair		p-value
	SNP	Gene (Transcripts)	
CRBL	chr10:122309274:CA_C	t3267563	7.038009e-43
FCTX	chr12:54056747:TGA_T	t3456313	3.905282e-24
HTPP	chr16:77389312	t3700158	3.993966e-22
MEDU	chr16:77389312	t3700158	1.821483e-30
OCTX	chr12:54041192	t3456313	9.334599e-24
PUTM	chr12:54056555	t3456313	1.954332e-14
SNIG	chr16:77390128	t3700158	4.926358e-16
TCTX	chr12:54041192	t3456313	1.68642e-17

THAL	chr16:77397167	t3700158	1.773935e-15
WHMT	chr16:77396849	t3700158	2.543182e-46

The R process of the analysis is shown as following:
(* I used the R in Windows, not in Linux. Therefore, there are a lot of lines of processing informa tion. Here I deleted the processing information and only kept the relative results.)

R Notebook

```
## Problem 2
setwd("C:/Users/Gulishana/Desktop/HW4_eQTL")

## Part 1: markers report
library(MatrixEQTL);
useModel = modelLINEAR;
SNP_file_name = "hw4_markers.txt";
snps = SlicedData$new();
snps$fileDelimiter = "\t";      # the TAB character
snps$fileOmitCharacters = "NA"; # denote missing values;
snps$fileSkipRows = 1;          # one row of column labels
snps$fileSkipColumns = 1;       # one column of row labels
snps$fileSliceSize = 2000;      # read file in slices of 2,000 rows
snps$LoadFile(SNP_file_name);


## Rows read:  34373  done.



## Part 2: transcripts report of each brain region (10 regions in total)
region <- c("CRBL","FCTX","HIPP","MEDU","OCTX","PUTM","SNIG","TCTX","THAL","W
HMT");
for(r in region){

  expression_file_name = paste("expr_",r,".txt",sep="");
  output_file_name = paste("res_",r,".txt",sep="");
  pvOutputThreshold = 1e-5;
  errorCovariance = numeric();
  gene = SlicedData$new();
  gene$fileDelimiter = " ";       # the SPACE character
  gene$fileOmitCharacters = "NA"; # denote missing values;
  gene$fileSkipRows = 1+291704;   # one row of column labels and all exons
  gene$fileSkipColumns = 1;       # one column of row labels
  gene$fileSliceSize = 2000;      # read file in slices of 2,000 rows
  gene$LoadFile(expression_file_name);
  cvrt = SlicedData$new();        # no covariates
  ## Run the analysis
  me = Matrix_eQTL_engine(
```

```
    snps = snps,
    gene = gene,
    cvrt = cvrt,
    output_file_name = output_file_name,
    pvOutputThreshold = pvOutputThreshold,
    useModel = useModel,
    errorCovariance = errorCovariance,
    verbose = TRUE,
    pvalue.hist = TRUE,
    min.pv.by.genesnp = FALSE,
    noFDRsaveMemory = TRUE);

}
```

```
## Rows read:  26493  done.
## Rows read:  26493  done.
## Rows read:  26493  done.
## Rows read:  26493  done.
## Rows read:  26493  done.
## Rows read:  26493  done.
## Rows read:  26493  done.
## Rows read:  26493  done.
## Rows read:  26493  done.
```

```
## Part 3: find the most significant eQTLs in each brain region (10 regions i
n total)
region <- c("CRBL","FCTX","HIPP","MEDU","OCTX","PUTM","SNIG","TCTX","THAL","W
HMT");
for(r in region){
  res <- read.table(paste("res_",r,".txt",sep=""),header=TRUE);
  print(cbind(r,res[which.min(res$p.value),]));
}
```

```
##         r                   SNP     gene      beta    t.stat      p.value
## 1152 CRBL chr10:122309274:CA_C t3267563 -1.173771 -20.51341 7.038009e-43
##         r                   SNP     gene       beta    t.stat      p.value
## 6437 FCTX chr12:54056747:TGA_T t3456313 -0.3157513 -12.49703 3.905282e-24
##         r             SNP     gene       beta    t.stat      p.value
## 7603 HIPP chr16:77389312 t3700158 -0.6120653 -11.69558 3.993966e-22
##          r             SNP     gene       beta   t.stat      p.value
## 16739 MEDU chr16:77389312 t3700158 -0.7950606 -15.0664 1.821483e-30
##         r             SNP     gene       beta    t.stat      p.value
## 3673 OCTX chr12:54041192 t3456313 -0.3753219 -12.34583 9.334599e-24
##         r             SNP     gene       beta    t.stat      p.value
## 8103 PUTM chr12:54056555 t3456313 -0.3159982 -8.606115 1.954332e-14
##         r             SNP     gene       beta    t.stat      p.value
## 8723 SNIG chr16:77390128 t3700158 -0.5083433 -9.259502 4.926358e-16
##         r             SNP     gene       beta    t.stat     p.value
```

```
## 5650 TCTX chr12:54041192 t3456313 -0.3105819 -9.850797 1.68642e-17
##         r             SNP     gene      beta    t.stat      p.value
## 7439 THAL chr16:77397167 t3700158 -0.404165 -9.033251 1.773935e-15
##         r             SNP     gene      beta    t.stat      p.value
## 6974 WHMT chr16:77396849 t3700158 -1.342786 -22.16242 2.543182e-46
```

Problem3:

While performing the unpooled FDR adjustment, the significant associations in each brain regio n at 1% FDR are shown below. The total number is **1606** significant associations.
After a global FDR adjustment, there are **1574** significant associations at 1% FDR.

R Notebook

```
## Problem 3
## Part 1: perform an unpooled FDR adjustment of p-values at 1% FDR
setwd("C:/Users/Gulishana/Desktop/HW4_eQTL")

n <- 26493*34373;
alpha <- 0.01;
region <- c("CRBL","FCTX","HIPP","MEDU","OCTX","PUTM","SNIG","TCTX","THAL","W
HMT");
for(r in region){
  res <- read.table(paste("res_",r,".txt",sep=""), header=TRUE);
  cat(r,":",sum(p.adjust(res$p.value,method="fdr",n=n)<=alpha),"\n");
}

## CRBL : 300
## FCTX : 141
## HIPP : 129
## MEDU : 156
## OCTX : 290
## PUTM : 26
## SNIG : 53
## TCTX : 143
## THAL : 156
## WHMT : 212

## Part 2: pool p-values first and perform a global FDR adjustment of p-value
s at 1% FDR
region <- c("CRBL","FCTX","HIPP","MEDU","OCTX","PUTM","SNIG","TCTX","THAL","W
HMT");
p <- rep(list(NA),length(region));
for(i in 1:length(region)){
```

```
  res <- read.table(paste("res_",region[i],".txt",sep=""), header=TRUE);
  p[[i]] <- res$p.value;
}
p <- unlist(p);
alpha <- 0.01;
sum(p.adjust(p,method="fdr",n=26493*34373*10)<=alpha);

## [1] 1574
```

Problem4:

There are **684** unique eQTL pairs.

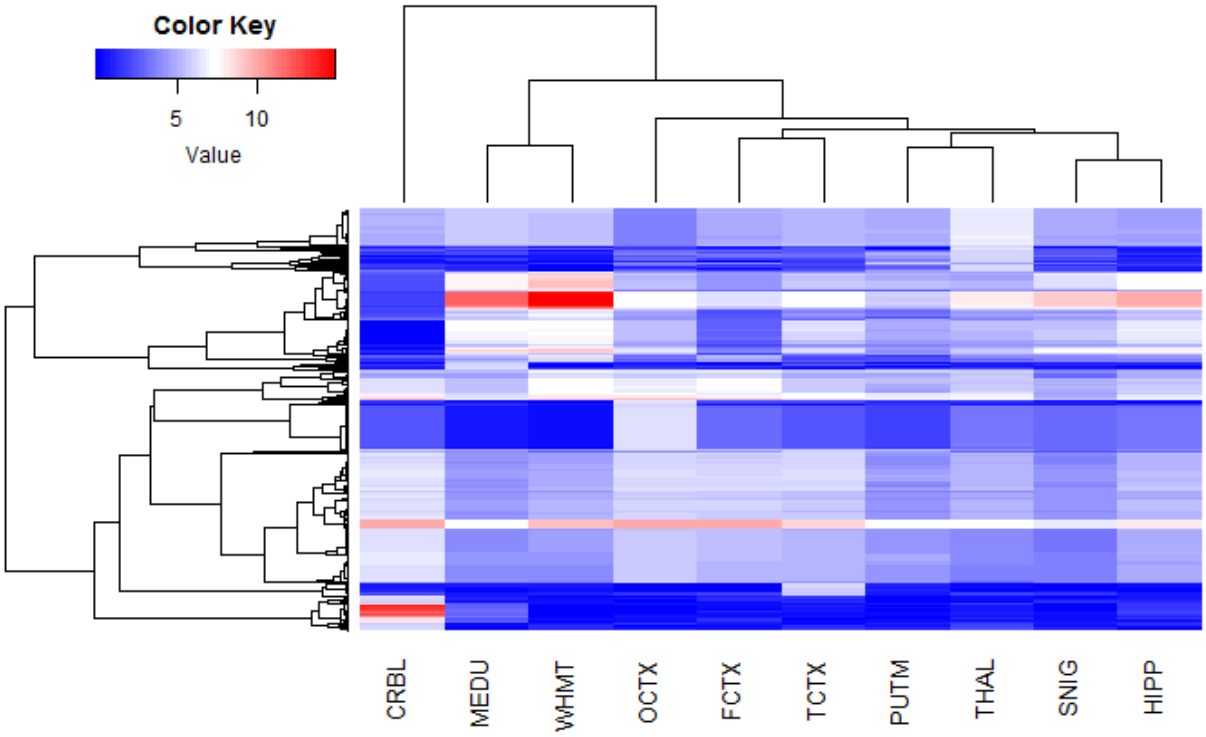
R Notebook

```
## Problem 4
## Identify the significant eQTLs in each brain region
setwd("C:/Users/Gulishana/Desktop/HW4_eQTL")

n <- 26493*34373;
alpha <- 0.01;
pairs <- matrix(NA,nrow=0,ncol=2);
region <- c("CRBL","FCTX","HIPP","MEDU","OCTX","PUTM","SNIG","TCTX","THAL","W
HMT");
for(r in region){
  res <- read.table(paste("res_",r,".txt",sep=""), header=TRUE);
  keep <- p.adjust(res$p.value,method="fdr",n=n)<=alpha;
  pairs <- rbind(pairs,res[keep,1:2]);
}

pairs <- pairs[!duplicated(pairs),];

save(pairs,file = "pairs.Rdata");
nrow(pairs);

## [1] 684
```

Problem5:

R Notebook

```
setwd("C:/Users/Gulishana/Desktop/HW4_eQTL")

load("pairs.Rdata")

library(data.table);
markers <- fread("hw4_markers.txt");

##
Read 34373 rows and 135 (of 135) columns from 0.014 GB file in 00:00:03

setkey(markers,"id");
zscore <- matrix(NA,nrow=684,ncol=0);
region <- c("CRBL","FCTX","HIPP","MEDU","OCTX","PUTM","SNIG","TCTX","THAL","W
HMT");
for(r in region){
  expr <- fread(paste("expr_",r,".txt",sep=""));
  setkey(expr,"ExprID");
  zscore <- cbind(zscore,apply(pairs,1,function(x){
    g <- as.numeric(expr[x[2],-1,with=F]);
    s <- as.numeric(markers[x[1],-1,with=F]);
    p <- summary(lm(g~s))$coef[2,4];
    return(-qnorm(p/2));
  }));
}

colnames(zscore) <- region;
save(zscore,file = "zscore.Rdata");


load("zscore.Rdata")
hc.eqtl <- hclust(as.dist((1-cor(t(zscore)))/2),method="complete");

# 50 colors between blue and red
library(gplots);

## Warning: package 'gplots' was built under R version 3.3.3

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

heatmap.2(zscore,Rowv=as.dendrogram(hc.eqtl),
          labRow=FALSE,
          col=bluered(50),key=TRUE,density.info="none",trace="none",cexCol=
1);
```



The heatmap shows that there is one set of similarly regulated eQTL group in each of the CRBL, MEDU, and WHMT brain regions. The similarly regulated gene group in CRBL is highly specific and only present in this brain region, which has no such regulation similarity in any other brain regions. In the Figure1a, there is also such an only active group in CRBL. However, the similarly regulated gene group in region MEDU and WHMT is similar between these two regions, and no such regulation pattern shown in other regions, or not as significance as in MEDU and WHMT.

Problem6:

R Notebook

```
setwd("C:/Users/Gulishana/Desktop/HW4_eQTL")
load("pairs.Rdata")
load("zscore.Rdata")

hc.eqtl <- hclust(as.dist((1-cor(t(zscore)))/2),method="complete");
clust <- cutree(hc.eqtl,k=3);
table(clust);

## clust
##   1   2   3
## 423 160 101
```

```
for(i in 1:3){
  trans <- sapply(as.character(pairs[clust==i,"gene"]),function(x){
    substr(x,2,nchar(x));
    });
  write.table(unique(trans),file=paste("cluster_",i,".txt",sep=""),
              row.names=FALSE,col.names=FALSE,quote=FALSE);
}

library(data.table);
cluster_1 <- fread("cluster_1.txt"); nrow(cluster_1);

## [1] 13

cluster_2 <- fread("cluster_2.txt"); nrow(cluster_2);

## [1] 10

cluster_3 <- fread("cluster_3.txt"); nrow(cluster_3);

## [1] 40
```

Therefore,

there are **13** unique transcripts in **Cluster 1**;

there are **10** unique transcripts in **Cluster 2**;

there are **40** unique transcripts in **Cluster 3**.

DAVID analysis:

For the 13 genes in cluster 1, the most enriched terms of functions are **ion transport** and **heparin-binding**. And the clustering of the functions is mostly regarding **DNA-binding** and **transcription regulation** that take place in **nucleus**. However, the enrichment score is not high enough due to the low number of genes submitted.

For the 10 genes in cluster 2, the most enriched terms of functions are **transmembrane (helix)** and **integral component of membrane**. And the clustering of the functions is also mostly regarding **transmembrane (helix)** and **integral component of membrane**. Though the number of genes submitted is very low, it still gets a relatively high enough enrichment score. This highly indicates that these 10 genes are encoding transmembrane proteins.

For the 40 genes in cluster 3, the most enriched terms of functions are **myelin sheath**, **axon**, and **phosphoprotein**. And the clustering of the functions is mostly regarding **GTPase activity** and **nucleotide-binding**. The enrichment score is relatively high since there are relatively more genes submitted.