RNAseq解析チュートリアル

松波　雅俊（大学院医学研究科先進ゲノム検査医学講座）

2023年12月23日作成

サーバへのログイン

まずsshコマンドを用いて配布したIDでサーバにログインしてください。

"server.address"のアドレスは当日お知らせします。

```
ssh [共有サーバのアカウント名]@server.address
xxxxx@server.address's password:
# [パスワードの入力]
Last login: XXX 2021 from XX
```

Rの起動とデータ読み込み

サーバにログイン出来たら、作業ディレクトリを作成し、Rを起動しましょう。

```
mkdir RNAseq_class #作業ディレクトリの作成
cd RNAseq_class            #作業ディレクトリに移動
R  #Rの起動

R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R は、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、'license()' あるいは 'licence()' と入力してください。

R は多くの貢献者による共同プロジェクトです。
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で引用する際の形式については
'citation()' と入力してください。

'demo()' と入力すればデモをみることができます。
'help()' とすればオンラインヘルプが出ます。
'help.start()' で HTML ブラウザによるヘルプがみられます。
'q()' と入力すれば R を終了します。

>
```

Rが起動している状態では行頭に">"が表示されます。

この状態でコマンドを入力し解析を進めます。

まず、使用するRのライブラリーを読み込みます。

```
library(edgeR)
library(heatmap3)
```

次に解析に使用するデータを読み込みます。

今回はシミュレーションで作った2グループ（それぞれ繰り返し3回）の遺伝子発現データを使います。

6サンプル（2 groups x 3 replicates）, 20,000遺伝子の発現データは　https://raw.githubusercontent.com/wachinakatada/23_seimeijoho/main/Matsunami/SimData.txt に置かれています。

このファイルはタブ区切りで 1 行目にサンプル名、1列目に遺伝子名が記載されています。

よって、sep="\t"でタブ区切りを指定して, header=Tで 1 行目を行名とし, row.names=1で1列目を列名として、これを読み込みます。

```
d <-
read.table("https://raw.githubusercontent.com/wachinakatada/23_seimeijoho/main/Matsunami/SimData.txt",sep="\t",header=T,row.names=1)
```

データは、行列として変数"d"に格納されました。

正しく読み込めたか中身を確認しましょう。

```
head(d)      #行列の先頭数行を表示
#        G1_rep1 G1_rep2 G1_rep3 G2_rep1 G2_rep2 G2_rep3
#gene_1       0       0       0       0       0       0
#gene_2      83     256     190      42     102      95
#gene_3     214     169     208      45      67      31
#gene_4     494     622     575     159     128     109
#gene_5     110     239     256      19      17      35

d$G1_rep1    #G1_rep1の列を表示
d[,1]        #1列目を表示
d[1,]        #1行目を表示
```

他にもいろいろな表示方法があるので、各自調べてみてください。

scatter plot

データが読み込めたので、サンプルごとの遺伝子発現にどれくらいばらつきがあるのかをscatter plotを用いて確認します。

```
#値が0だとエラーが出るので1を足す
G1_1 <- d$G1_rep1 + 1
G1_2 <- d$G1_rep2 + 1
G1_3 <- d$G1_rep3 + 1
G2_1 <- d$G2_rep1 + 1
G2_2 <- d$G2_rep2 + 1
G2_3 <- d$G2_rep3 + 1

#1 pair表示
jpeg('test.scatter.jpeg',width = 1024, height = 768)
plot(G1_1, G1_2, log="xy", col="black", pch=16, cex=0.5, xlim = c(1, 10000), ylim =
c(1, 10000))
dev.off()
```

これで、group"G1_1"/group"G1_2"間のscatter plotがファイル"test.scatter.jpeg"に保存されました。

各自、ファイルをサーバから自分のPCに転送して中身を確認してください。

全ての結果をまとめて1つの画像ファイルに保存する場合は、例えば下記のように書きます。

これで、6つのscatter plotがまとめて"DE.scatter.jpeg"に出力されます。

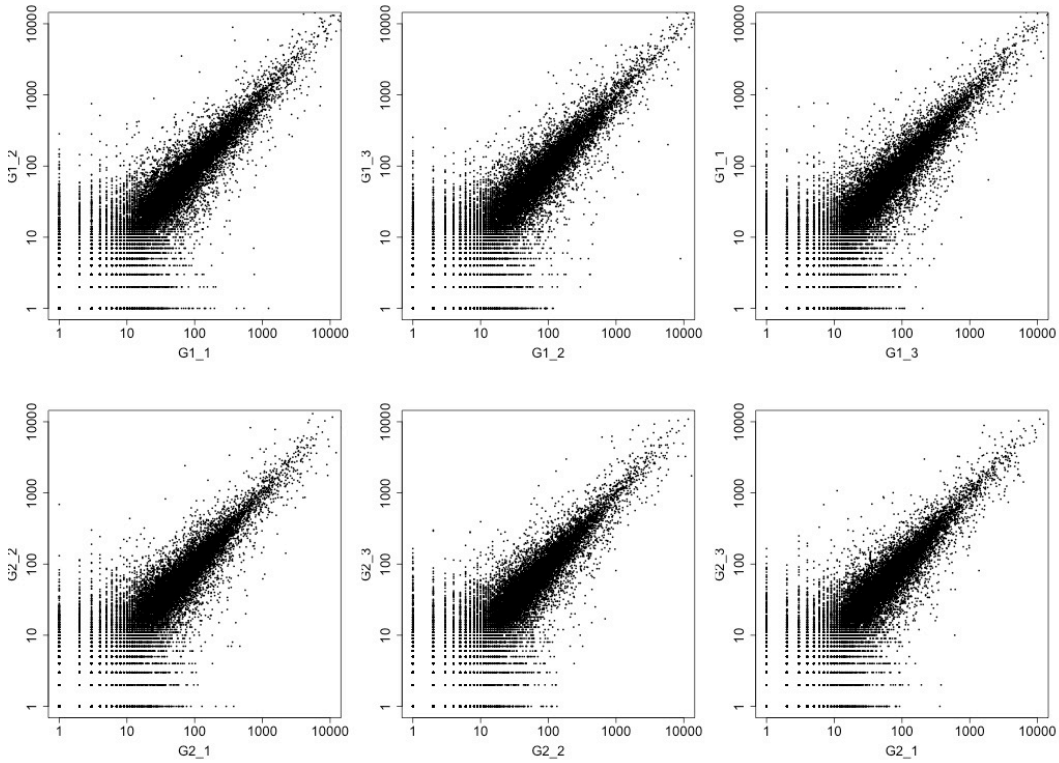
```
jpeg('DE.scatter.jpeg',width = 1024, height = 768)
par(mfrow=c(2,3))
par(ps = 20)
plot(G1_1, G1_2, log="xy", col="black", pch=16, cex=0.5, xlim = c(1, 10000), ylim =
c(1, 10000))
plot(G1_2, G1_3, log="xy", col="black", pch=16, cex=0.5, xlim = c(1, 10000), ylim =
c(1, 10000))
plot(G1_3, G1_1, log="xy", col="black", pch=16, cex=0.5, xlim = c(1, 10000), ylim =
c(1, 10000))
plot(G2_1, G2_2, log="xy", col="black", pch=16, cex=0.5, xlim = c(1, 10000), ylim =
c(1, 10000))
plot(G2_2, G2_3, log="xy", col="black", pch=16, cex=0.5, xlim = c(1, 10000), ylim =
c(1, 10000))
plot(G2_1, G2_3, log="xy", col="black", pch=16, cex=0.5, xlim = c(1, 10000), ylim =
c(1, 10000))
dev.off()
```

出力結果は下記のようになるはずです。

各々の点が遺伝子を表しており、各軸がそれぞれのグループにおける発現量を表しています。

対角線上に位置する点はグループ間でほぼ同じ発現量の遺伝子です。

一方で、対角線から離れるに従って、グループ間で発現量の差が大きくなることを示しています。



発現解析

scatter plotでざっくりとした遺伝子発現の傾向を把握した後、いよいよ遺伝子発現解析を実施します。

今回は、RのedgeRという遺伝子発現解析のライブラリを使用します。

まずは、授業で説明したTMM (Trimmed Mean of M values)という方法でデータを正規化します。

```
#比較グループを設定
grp <- c("G1","G1","G1","G2","G2","G2")
D <- DGEList(d, group=grp)

#中身を確認
ls(D)       #"counts"  "samples"
D$samples

#TMM正規化
D <- calcNormFactors(D)

#中身を確認
ls(D)       #"counts"  "samples"
D$samples   #norm.factorsの値が変わっている
```

正規化後の遺伝子発現がサンプル間でどれぐらい似ているかを多次元解析法の 1 つであるMDS plotを用いて確認します。

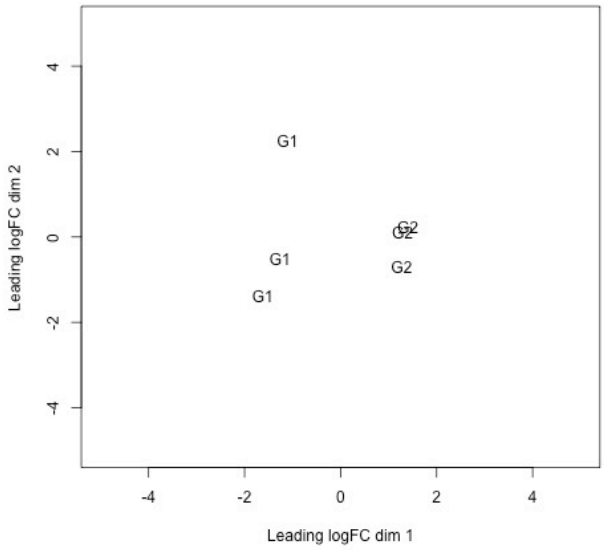
このplotでは、遺伝子発現が類似しているサンプルほと近くにプロットされます。

今回のデータでは、replicatesがあるので、これらの遺伝子発現は類似していることが期待されます。

```
jpeg('DE.MDS.jpeg')
#plotMDS(D, labels=grp)
plotMDS(D, labels=grp, xlim = c(-5, 5), ylim = c(-5, 5))
dev.off()
```

出力された"DE.MDS.jpeg"をサーバから自分のPCに転送して確認しましょう。

確かにreplicatesは、それぞれ近くにプロットされていることがわかります。



それでは、２群間で発現が変動している遺伝子を検出しましょう。

下記のコマンドでは、正規化後のデータの分散を計算した後、正確確率検定で発現変動遺伝子を検出しています。

```
#common dispersionを計算
D <- estimateCommonDisp(D)

#中身を確認
ls(D)   #"AveLogCPM" "common.dispersion" "pseudo.counts" "pseudo.lib.size"が追加されている



#moderated tagwise dispersionを計算
D <- estimateTagwiseDisp(D)

#中身を確認
ls(D)   #さらに"prior.df" "prior.n" "span" "tagwise.dispersion"が追加されている


#exact test (正確確率検定)で発現変動遺伝子を検出
results <- exactTest(D,pair=c("G1","G2"))

#中身を確認
```

```
ls(results) #"comparison" "genes" "table"

#P値でsortしてDEGを検出
tmp <- topTags(results, n=NULL)

#中身を確認
ls(tmp)  #"adjust.method" "comparison" "table" "test"

#P値はtmp$tableに記載されているのでこれをファイルに書き出す
write.table(tmp$table, "DE.results.tsv", sep="\t", quote=F)

#P < 0.05だった遺伝子のみの結果をファイルに書き出す
write.table(tmp[tmp$table$PValue <= 0.05,], "DE.results.p0.05.tsv", sep="\t", quote=F)
```

一般的には、２群間の比較で, false discovery rate (FDR) < 0.05であれば発現変動遺伝子と定義されることが多いです。

解析結果をMA plotで確認しましょう。

FDR < 0.05となる遺伝子は、赤で色分けします。

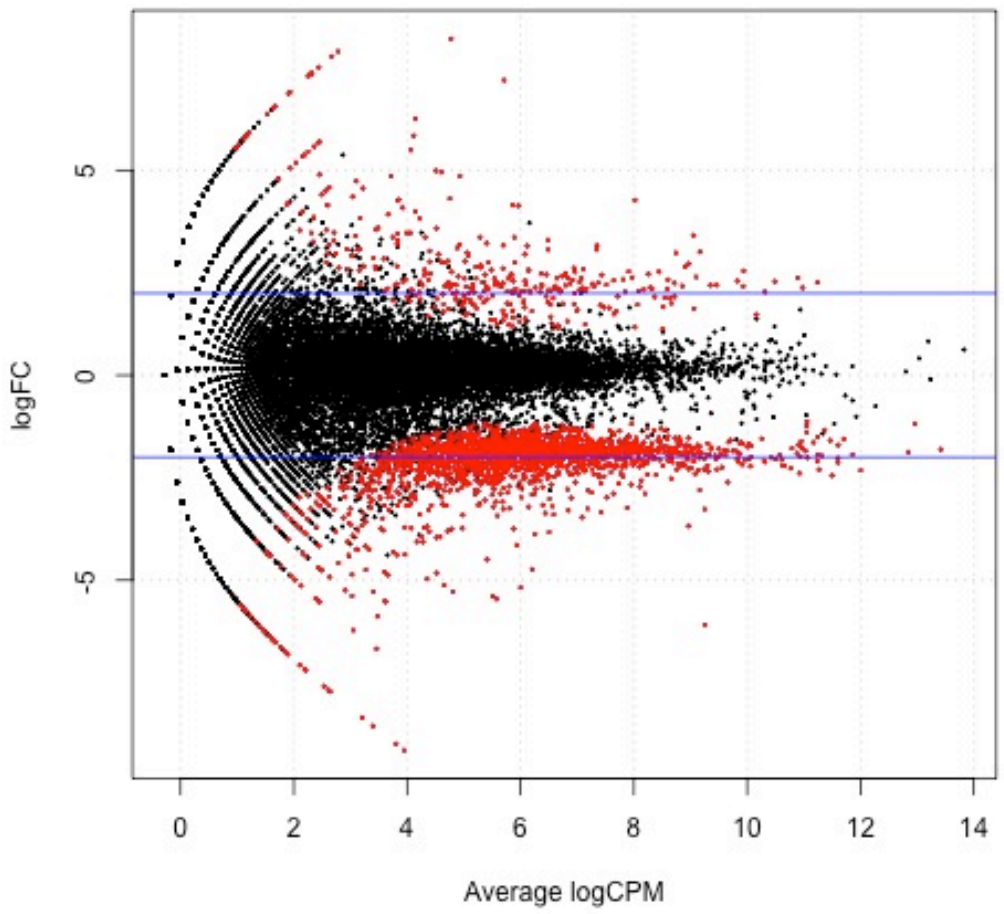
```
#発現変動遺伝子を定義
q.value <- p.adjust(results$table$PValue, method = "BH")
is.DEG <- (q.value < 0.05)
de.tags <- rownames(results$table)[is.DEG]

#MA plot
jpeg('DE.MA.jpeg')
plotSmear(results, de.tags = de.tags)
abline(h=c(-2,2),col="blue")
dev.off()
```

出力結果を確認しましょう。

MA plotは、横軸に発現の強度、縦軸に２群間での発現の違いが表現されています。

２群間で発現変化がない遺伝子はlogFC = 0のところにプロットされます。

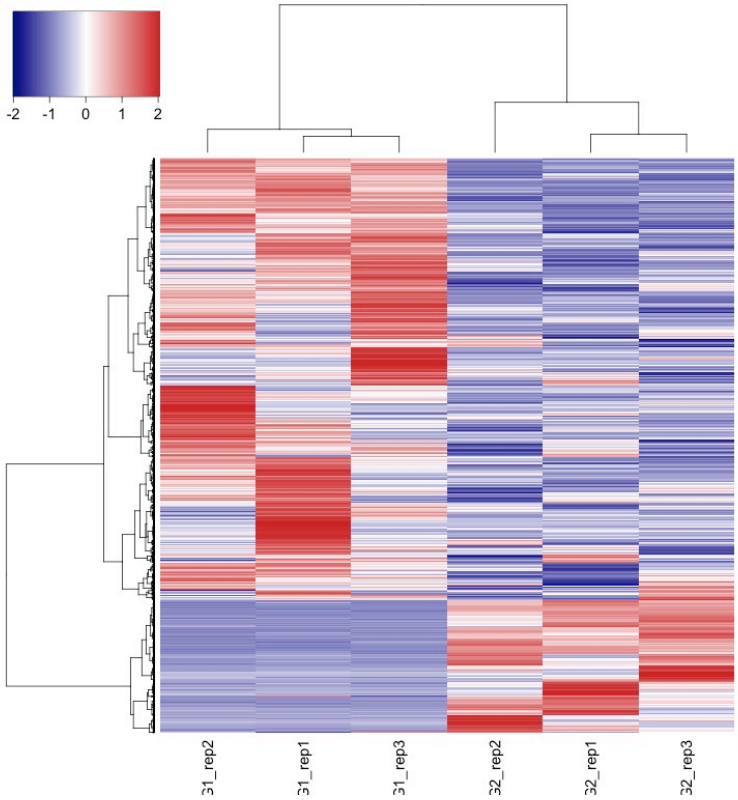
赤で表されているFDR < 0.05となる発現変動遺伝子は、logFC = 0のところから離れているのがわかります。



最後に、P < 0.05だった遺伝子を抽出して、heat mapを書きましょう。

```
#count dataからP < 0.05だった遺伝子のみ抽出
count <- transform(D$count, PValue=results$table$PValue)
count2 <- count[count$PValue <= 0.05,]
count2$PValue <- NULL

#cpmで正規化してplot
m <- as.matrix((cpm(count2)))
jpeg('DE.heatmap.jpeg',width = 1024, height = 768)
heatmap3(m, method="ward.D2", labRow="")
dev.off()
```

出力結果



今回のチュートリアルは以上になります。

さらに詳しく勉強したい場合は、下記のedgeRのuser guideを参考にしてください。

https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf