2016年度第4回
バイオインフォマティクス実習

発現変動遺伝子の抽出

発現変動遺伝子の抽出

- 前回までの実習で次世代シーケンスデータのマッピング→タグのカウントデータを取得しました。
- 遺伝子毎の発現プロファイルの扱いはマイクロアレイもRNA-seqも同じ。
- 二群間の比較を行い発現変動遺伝子を抽出する。

仮説検定

① 帰無仮説：棄却されるべき仮説

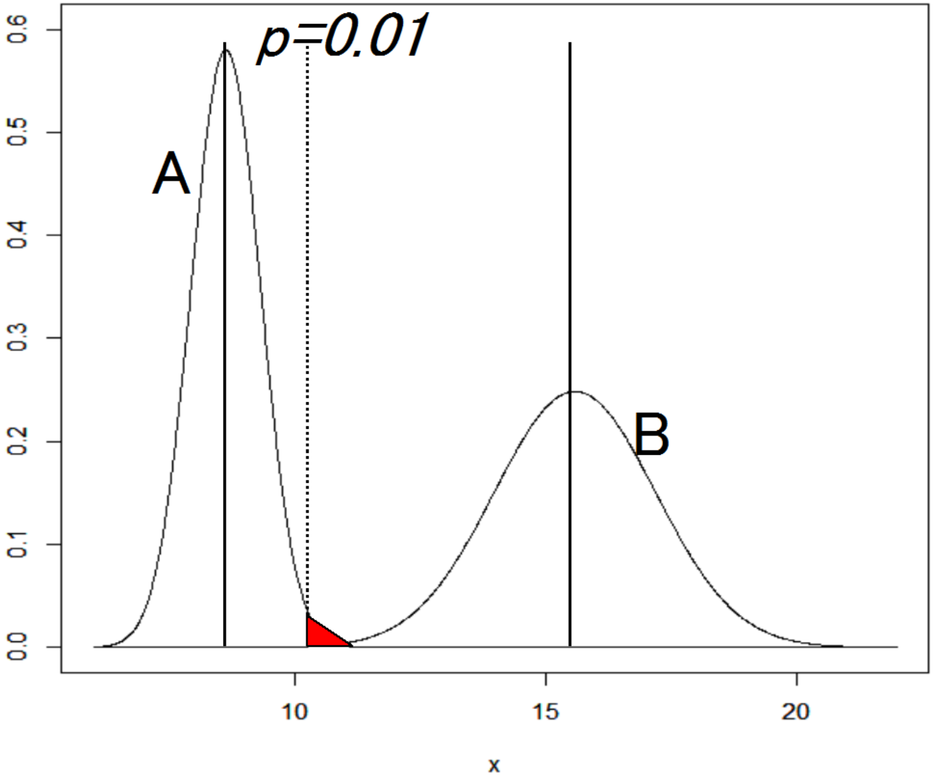
② 帰無仮説の元で観測されたデータが取得される確率を計算する

③ 有意水準から仮説を棄却するか否か判定する

二群間の比較

A群	B群
7.89	16.28
9.60	16.75
9.07	13.21
8.31	17.01
8.30	14.69

帰無仮説：A群、B群は平均の等しい分布から得られた



B群のデータが得られる確率は1%以下
↓
帰無仮説を棄却し、二群間には差があると判定する

検定による過誤と多重検定問題

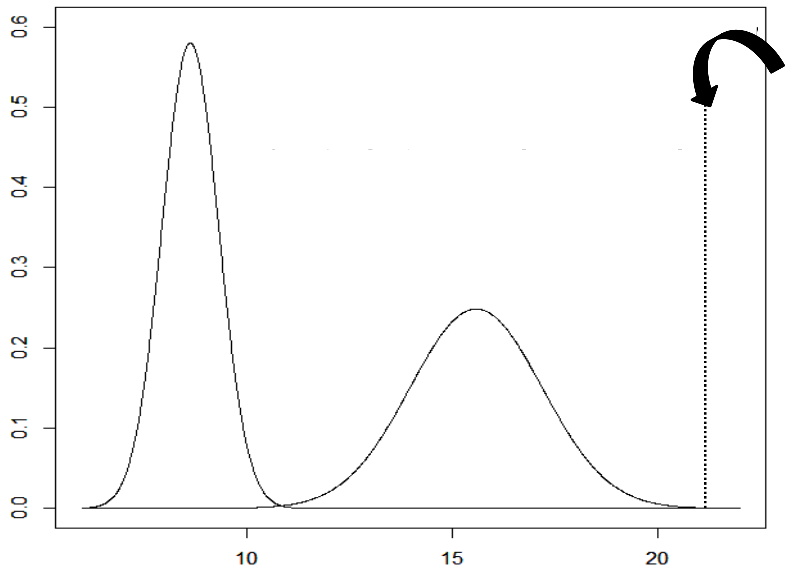
有意水準　$P<0.01$　：実測データが得られる確率は1%未満
遺伝子発現データ20000超→200個以上のFalse Positive

実際/判定	陰性	陽性
陰性	True Negative　(TN)	False Positive　(FP)
陽性	False Negative　(FN)	True Positive　(TP)

Bonferroni補正
有意水準を検定回数で割る

$$\hat{p} = p / N$$

$$\hat{p} = 0.01/20000 = 5\times10^{-7}$$



非現実的な有意水準が
要求される

False Discovery Rate (FDR)

BH法：1995年にBenjaminiとHochbergによって提唱された。FDR＝FP/(FP+TP)を指標にする手法。ある程度のFPを許容する。

H ： 帰無仮説
m : 帰無仮説の数
p ： 有意水準

H_1	H_2	H_3	…	H_m
p_1	p_2	p_3		p_m
q_1	q_2	q_3		q_m

$$q_{(i)} = \frac{p_{(i)} \times m}{i} \qquad (i = 1, 2, \cdots, m)$$

$$\begin{cases} q_{(m)} = q'_{(m)} \\ q_{(i)} = \min\{q'_{(i)}, q_{(i+1)}\} \end{cases}$$

p値を低いものから並べる
q値を計算
$q_i > q^*$となる帰無仮説を全て棄却する

統計解析ソフトR　Significance Analysis for Microarray (SAM) パッケージ

R console

```
> Sys.setenv(http_proxy = “http://proxy.yokohama-cu.ac.jp:8080”)
> source(“http://bioconductor.org/biocLite.R”)
> biocLite(“samr”)
> library(samr)
```

デモデータ

- GEOデータベース No. GSE40493
- Bcl6KOマウス4サンプル + WT4サンプル
- プラットフォーム Affymetrix Mouse Gene 1.0 ST Array
- 正規化済み 発現量のテキストデータ
  GSE40493_Normalized_ID.txt

SAMのコマンド

R console

```
> data <- read.table(“Z:¥デスクトップ¥GSE40493_normalized_ID.txt”, sep=“¥t”, header=T) ↵
> data.tmp <- list(x = as.matrix(data[,3:10]), y = c(rep(1,4), rep(2,4)), ↵
+ geneid = data$ID, genenames = data$GeneName, ↵
+ logged2 = FALSE) ↵
> out <- samr(data.tmp, resp.type = “Two class unpaired”, nperms = 20)
```

1. データの読み込み
2. データセットの設定（データ、ラベル、遺伝子ID、遺伝子名、log変換の有無）
3. 実行

統計量の計算

R console

```
> p.value <- samr.pvalues.from.perms(out$tt, out$ttstar)
> q.value <- p.adjust(p.value, method = “BH”)
> ranking <- rank(p.value)
> stat_fc <- log2(out$foldchange)
> rank_fc <- rank(-abs(stat_fc))
```

1. p値
2. q値
3. p値によるランキング
4. fold change
5. fold changeによるランキング

ファイルに出力

R console

```
> tmp <- cbind(data$GeneName, data[,3:10], p.value, q.value, ranking, stat_fc, rank_fc)
> write.table(tmp, “result.txt”, quote=F, sep=“¥t”, row.names=F)
```

1. 変数tmpに結果を一時格納
2. ファイルに出力

結果

Bcl6KO WT

q-value

data$GeneName	GSM995228_AD01M008	GSM995227_AD01M007	GSM995226_AD01M006	GSM995225_AD01M005	GSM995224_AD01M004	GSM995223_AD01M003	GSM995222_AD01M002	GSM995221_AD01M001	p.value	q.value	ranking	stat_fc	rank_fc
Gm10568	4.773084	4.742637	4.743848	4.693223	4.767021	4.764441	4.71977	4.456037	0.446271	0.544563	17648	-0.01881	18162
Lypla1	8.1329	8.700074	8.942065	8.058603	8.507404	8.740488	9.771157	9.96973	0.063812	0.249864	5497	0.128629	1225
Tcea1	9.109184	9.335877	9.152534	8.970762	9.289149	9.207187	9.402179	9.339357	0.075108	0.255035	6340	0.026175	16736
Atp6v1h	7.523716	7.914531	7.951814	7.417282	7.67192	7.792107	8.781741	8.900342	0.090457	0.265667	7332	0.105565	2132
Oprk1	5.393743	5.358419	5.242766	5.529793	5.269703	5.154718	4.916978	4.9738	0.016875	0.246514	1319	-0.08343	4219
Rb1cc1	7.186024	7.389256	8.021381	7.023139	6.8452	7.084923	7.727727	7.872987	0.94384	0.95749	21228	-0.00434	20771
Fam150a	5.239859	5.273537	5.158567	5.171551	5.282442	5.143815	5.257763	5.10868	0.825733	0.865269	20551	-0.00352	20904

q-valueの計算

Fam158a	6.831996	6.794337	6.832658	6.979381	7.06143	6.843407	6.670263	6.863003	0.9993708	0.999556	21531	-1.41E-05	21530
Slc35b2	7.124304	7.028796	6.989175	7.221187	7.148212	6.965634	7.059167	7.190294	0.9995565	0.999677	21532	-7.84E-06	21533
Hesx1	5.384937	5.175449	5.295784	5.474797	5.386017	5.529073	5.11388	5.302199	0.9995844	0.999677	21533	1.37E-05	21531
Mkl1	7.849002	7.928032	8.309136	7.923435	7.733856	7.777829	8.269562	8.228277	0.9999303	0.999965	21534	-3.57E-06	21534
Magea4	6.024846	5.876155	5.695813	5.783122	6.033623	6.116748	5.626901	5.602719	0.9999652	0.999965	21535	3.46E-06	21535

$$(0.9999652 \times 21535)/21535 = 0.9999652$$

$$(0.9999303 \times 21535)/21534 = 0.9999767 > 0.9999652$$

$$(0.9995844 \times 21535)/21533 = 0.9996772 < 0.9999652$$

$$(0.9995565 \times 21535)/21532 = 0.9996958 > 0.9996772$$

$$(0.9993708 \times 21535)/21531 = 0.9995565 < 0.9996958$$

FDRの問題点

- 全サンプルでp値が閾値に達する割合が一様分布していると仮定している。
- 実際には有意差あり集団由来+有意差なし集団由来の混合分布である。
- 有意差ありをなしと厳しく判定してしまう。
- Storeyの補正法