

# NGSハンズオン講習会

## RNA-seq、カウントデータ取得以降の統計解析

東京大学・大学院農学生命科学研究科  
アグリバイオインフォマティクス教育研究プログラム

門田幸二(かどた こうじ)

kadota@iu.a.u-tokyo.ac.jp

<http://www.iu.a.u-tokyo.ac.jp/~kadota/>

# Contents (全体)

- 7月22日(水): 84→83名。Bio-Linux 8とRのインストール状況確認。基本自習(門田・寺田先生)
- 7月23日(木): 92→90名。Linux基礎。LinuxコマンドなどUNIXの基礎の理解(門田)
- 7月24日(金): 85→83名。スクリプト言語。シェルスクリプト(アメリエフ株式会社 服部恵美先生)
- 7月27日(月): 93→91名。スクリプト言語。Perl(アメリエフ 服部先生)
- 7月28日(火): 91→90名。スクリプト言語。Python(アメリエフ 服部先生)
- 7月29日(水): 94→88名。データ解析環境R(門田)
- 7月30日(木): 96→91名。データ解析環境R(門田)
- 8月3日(月): 89→84名。NGS解析。基礎(アメリエフ 山口昌雄先生)
- 8月4日(火): 85→80名。NGS解析。ゲノムReseq、変異解析(アメリエフ 山口先生)
- 8月5日(水): 86 →81名。NGS解析。RNA-seq、統計解析(前半:山口先生、後半:門田)
- 8月6日(木): 104 →98名。NGS解析。ChIP-seq(理研 森岡勝樹先生)
- 8月26日(水): 23名。NGS解析。基礎(アメリエフ 山口昌雄先生)
- 8月27日(木): 24名。NGS解析。ゲノムReseq、変異解析(アメリエフ 山口先生)
- 8月28日(金): 26名。NGS解析。RNA-seq、統計解析(前半:アメリエフ 山口先生、後半:門田)

# Contents

## ■ サンプル間クラスタリング

- 実行手順のおさらい
- 計算の一部を解説、結果の解釈

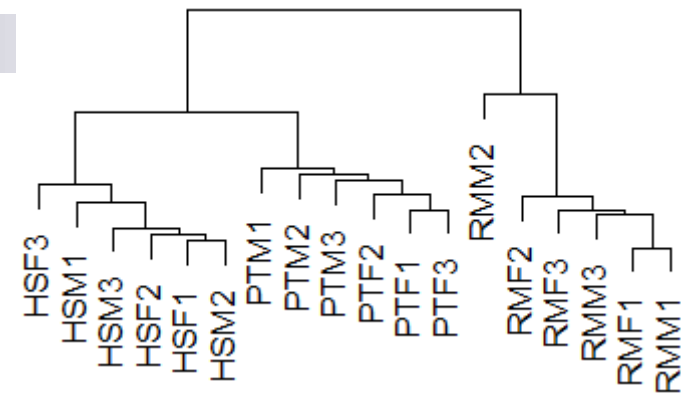
■ 発現変動解析(2群間比較) : 発現変動遺伝子(DEG)が多数存在する場合

■ モデル、分布、統計的手法、2群間比較でDEGがそれほど多くない場合

■ 2群間比較でDEGがほとんどない同一群の場合

■ 倍率変化(2倍以上、1/2以下の発現変動)を用いた場合

■ 発現変動解析: 3群間比較など



# クラスタリング

20,689遺伝子 × 18サンプルのbiological replicatesのみからなるカウントデータ (Blekhman et al., 2010)のサンプル間クラスタリング。データの取得や整形については、2015.07.29の講義資料を参考。

- ・ 解析 | [発現量推定\(トランスクリプトーム配列を利用\)](#) (last modified 2014/07/09)
- ・ 解析 | [クラスタリング | について](#) (last modified 2014/02/05)
- ・ 解析 | [クラスタリング | サンプル間 | hclust](#) (last modified 2015/02/26) **NEW**
- ・ 解析 | [クラスタリング | サンプル間 | TCC\(Sun\\_2013\)](#) (last modified 2015/03/02) **NEW**
- ・ 解析 | [クラスタリング | 遺伝子間 | MBCluster.Seq\(SiRNA\)](#) (last modified 2014/02/05)



## 解析 | クラスタリング | サンプル間 | TCC(Sun\_2013) **NEW**

TCCパッケージを用いてサンプル間クラスタリングを行うやり方を示します。clusterSample関数を利用した頑健なクラスタリング結果を返します。  
「ファイル」→「デスクトップの変更」で解析したいファイルを置いてあるデスクトップに移動し、以下をコピー



### 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

1. 59

[Blekhman et al., Genome Res., 2010](#)の 20,689 genes×18 samplesのカウントデータです。

Neyret-  
ンゲル

in\_f  
out\_f  
param

#必要  
libra

#入力  
data  
dim(d

#本番  
out <

```

in_f <- "sample_blekhman_18.txt"      #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png"                 #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400)              #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC)                          #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルを読み込み
dim(data)                             #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman",#クラスタリング実行結果をoutに格納
                     hclust.method="average", unique.pattern=TRUE)#クラスタリング実行結果をoutに格納

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各種パラメータを指定
par(mar=c(0, 4, 1, 0))                #下、左、上、右の順で余白(行)を指定
plot(out, sub="", xlab="", yaxp=lab=1, 2)#樹形図(デンドログラム)の表示

```

# 入力ファイル



2015/7/29受講者は、デスクトップ上のhogeフォルダ中にsample\_blekhman\_18.txtが存在するはず。未受講者およびファイルが存在しないヒトは、①右クリック、②「対象をファイルに保存」で基本デスクトップ上のhogeに保存。

## 8. サンプルデータ42のリアルデータ(sample\_blekhman\_18.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×18

```
in_f <- "sample_blekhman_18.txt" #入力ファイル
out_f <- "hoge8.png" #出力ファイル
param_fig <- c(700, 400) #パラメータ
```

```
#必要なパッケージをロード
library(TCC) #ロード
```

```
#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1)
dim(data) #次元
```

```
#本番実行
out <- plotHeatmap(data, out_f, param_fig, cex=1, dev.off=TRUE)
```

開く(O)  
新しいタブで開く(N...)  
新しいウィンドウで開く(N)  
対象をファイルに保存(A)... **②** (はピクセル)  
対象を印刷(P)  
切り取り  
コピー(C)  
ショートカットのコピー(T) (fで指定したファイル)  
貼り付け(P)

	HSF1	HSF2	HSF3	HSM1	HSM2	HSM3	PTF1	PTF2	PTF3	PTM1	PTM2	PTM3	RMF1	RMF2	RMF3	RMM1	RMM2	RMM3
ENSG000000000003	329	300	168	121	421	359	574	429	386	409	685	428	511	464	480	424	1348	705
ENSG000000000005	0	0	0	0	1	0	1	4	1	0	1	1	0	1	2	2	0	0
ENSG000000000419	81	61	56	39	78	62	100	66	65	59	58	93	67	72	57	49	82	90
ENSG000000000457	91	62	76	114	73	95	131	229	87	274	239	149	89	69	118	117	114	163
ENSG000000000460	6	17	12	15	7	17	8	8	5	12	7	10	4	4	10	7	3	4
ENSG000000000938	44	65	210	73	43	65	84	104	76	198	31	58	73	28	54	80	34	72
ENSG000000000971	4765	7225	3405	3600	6383	5546	5382	8331	4335	2568	5019	2653	13566	9964	18247	14236	5196	11834
ENSG000000001036	297	251	189	200	234	249	305	301	313	254	151	331	292	106	379	201	88	140
ENSG000000001084	630	737	306	336	984	459	417	328	885	298	569	218	1062	786	1110	873	664	1752
ENSG000000001167	36	30	36	29	33	28	63	80	25	69	74	41	62	34	108	97	35	61
ENSG000000001460	3	1	5	1	4	2	0	1	1	1	1	3	1	1	1	0	1	3
ENSG000000001461	49	37	34	28	62	32	75	69	40	90	69	60	210	92	176	247	81	117
ENSG000000001487	117	93	88	80	131	110	125	98	75	108	130	131	138	95	187	137	158	172

# 入力ファイル

このデータは、3種類の生物種間比較。ヒト(*Homo sapiens*; HS)、チンパンジー(*Pan troglodytes*; **PT**)、アカゲザル(*Rhesus macaque*; **RM**)。生物種ごとにメス3匹、オス3匹。雄雌を考慮しなければbiological replicates (生物学的な反復)は6。

20,689 genes

	ヒト ( <i>Homo sapiens</i> ; HS)						チンパンジー ( <i>Pan troglodytes</i> ; <b>PT</b> )						アカゲザル ( <i>Rhesus macaque</i> ; <b>RM</b> )					
	メス(Female)			オス(Male)			メス			オス			メス			オス		
	HSF1	HSF2	HSF3	HSM1	HSM2	HSM3	PTF1	PTF2	PTF3	PTM1	PTM2	PTM3	RMF1	RMF2	RMF3	RMM1	RMM2	RMM3
ENSG000000000003	329	300	168	121	421	359	574	429	386	409	685	428	511	464	480	424	1348	705
ENSG000000000005	0	0	0	0	1	0	1	4	1	0	1	1	0	1	2	2	0	0
ENSG000000000419	81	61	56	39	78	62	100	66	65	59	58	93	67	72	57	49	82	90
ENSG000000000457	91	62	76	114	73	95	131	229	87	274	239	149	89	69	118	117	114	163
ENSG000000000460	6	17	12	15	7	17	8	8	5	12	7	10	4	4	10	7	3	4
ENSG000000000938	44	65	210	73	43	65	84	104	76	198	31	58	73	28	54	80	34	72
ENSG000000000971	4765	7225	3405	3600	6383	5546	5382	8331	4335	2568	5019	2653	13566	9964	18247	14236	5196	11834
ENSG000000001036	297	251	189	200	234	249	305	301	313	254	151	331	292	106	379	201	88	140
ENSG000000001084	630	737	306	336	984	459	417	328	885	298	569	218	1062	786	1110	873	664	1752
ENSG000000001167	36	30	36	29	33	28	63	80	25	69	74	41	62	34	108	97	35	61
ENSG000000001460	3	1	5	1	4	2	0	1	1	1	1	3	1	1	1	0	1	3
ENSG000000001461	49	37	34	28	62	32	75	69	40	90	69	60	210	92	176	247	81	117
ENSG000000001487	117	93	88	80	131	110	125	99	75	108	130	131	139	95	197	137	159	172

# Rを起動

このデータは、3種類の生物種間比較。ヒト(Homo sapiens; HS)、チンパンジー(Pan troglodytes; PT)、アカゲザル(Rhesus macaque; RM)。生物種ごとにメス3匹、オス3匹。雄雌を考慮しなければbiological replicates (生物学的な反復)は6。

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の

Blekhman et al., Genome Res., 2010の 20,689 genes×18 samplesの

```

in_f <- "sample_blekhman_18.txt" #入力ファイル名
out_f <- "hoge8.png" #出力ファイル名
param_fig <- c(700, 400) #ファイルサイズ

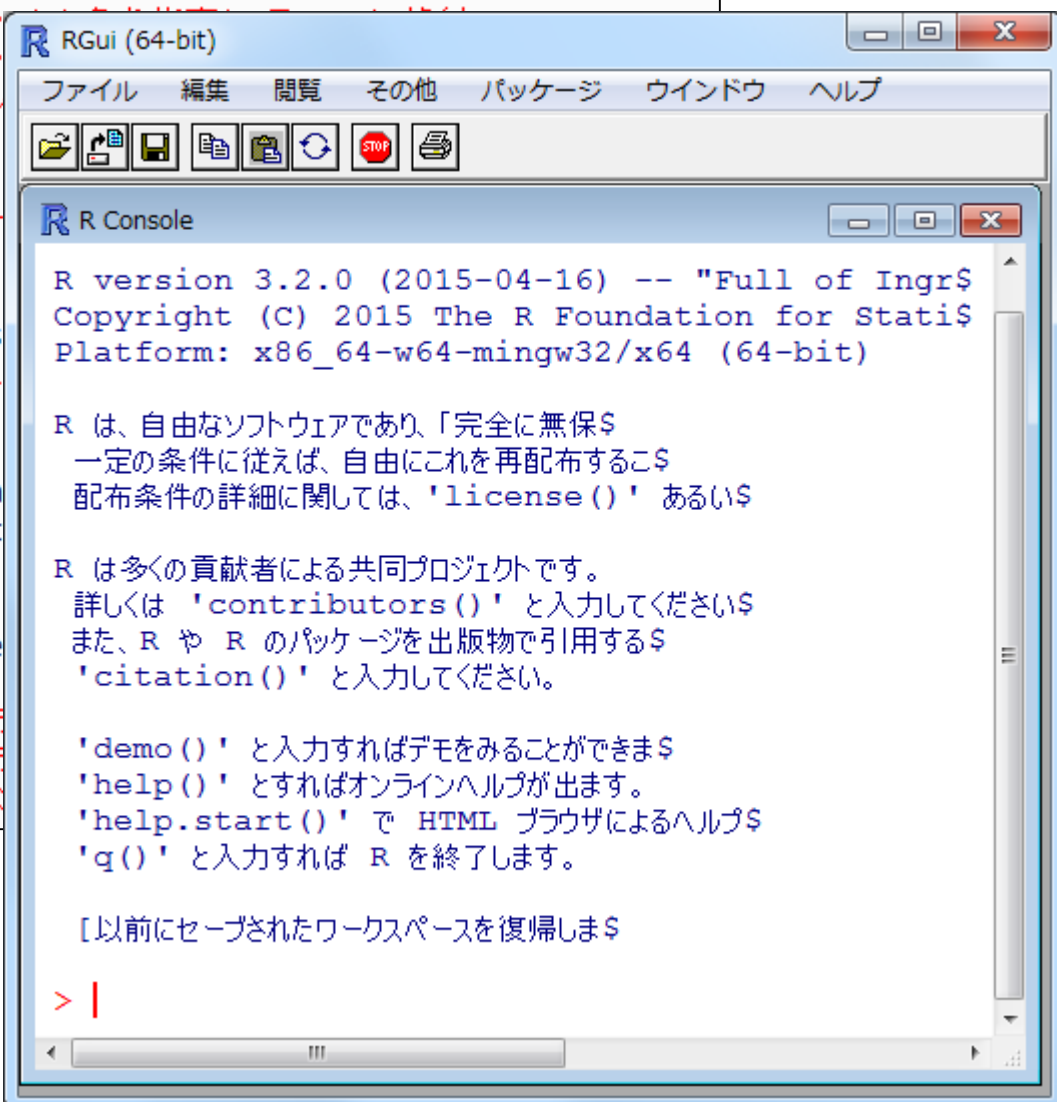
#必要なパッケージをロード
library(TCC) #パッケージ名

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=colnames, as.is=TRUE) #オブジェクト
dim(data)

#本番
out <- clusterSample(data, dist.method="spearman", hclust.method="average", unique.pat=1)

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2],
par(mar=c(0, 4, 1, 0)) #下、左、右、上の余白
plot(out, sub="", xlab="", cex.lab=1.2, #樹形図のx軸ラベル
cex=1.3, main="", ylab="Height") #樹形図のy軸ラベル
dev.off() #おまじな

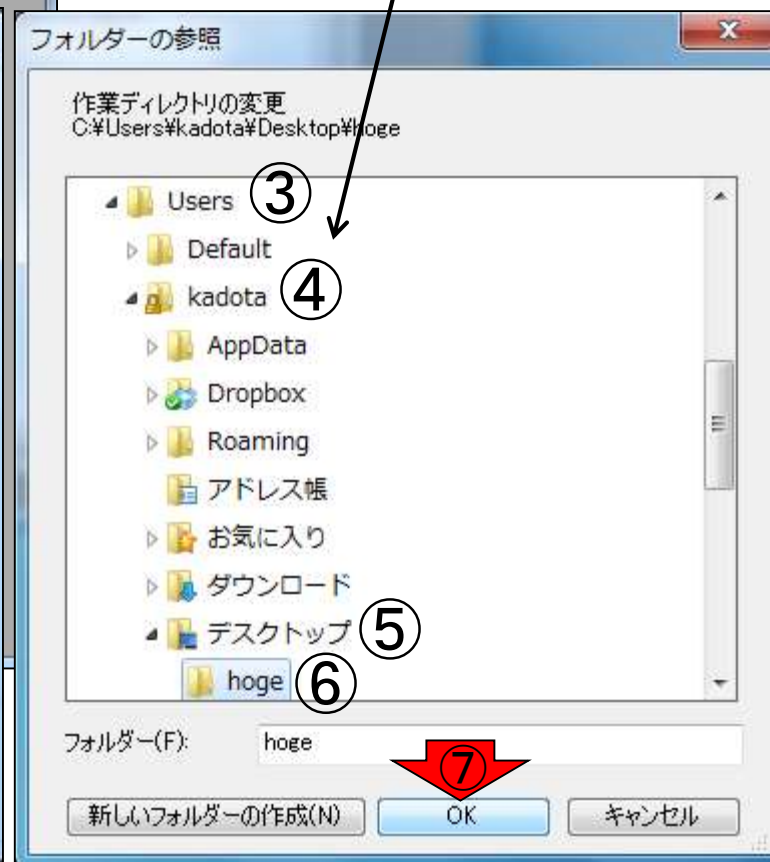
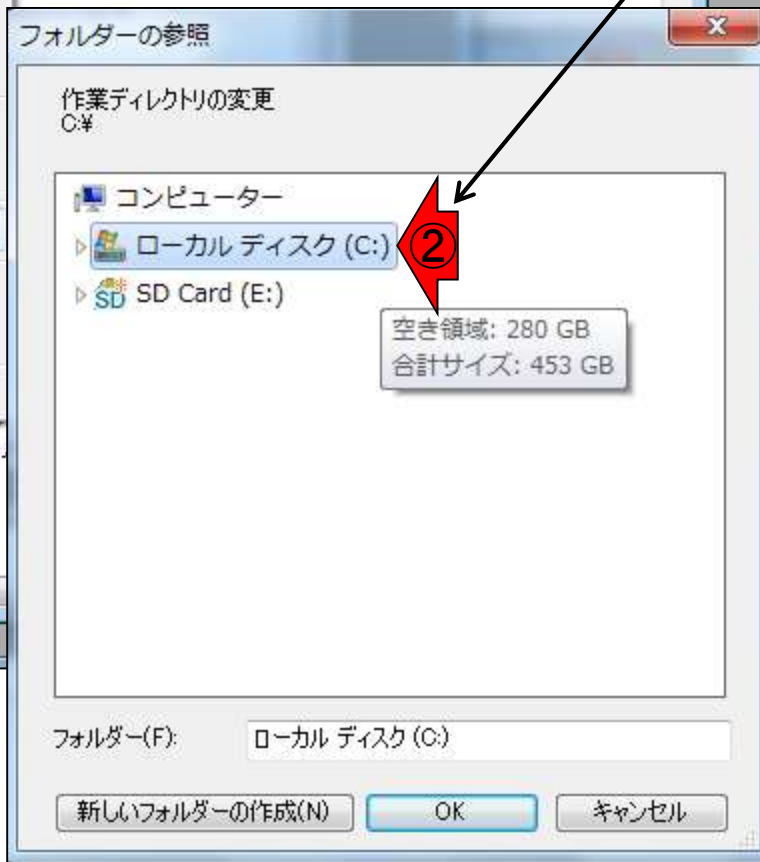
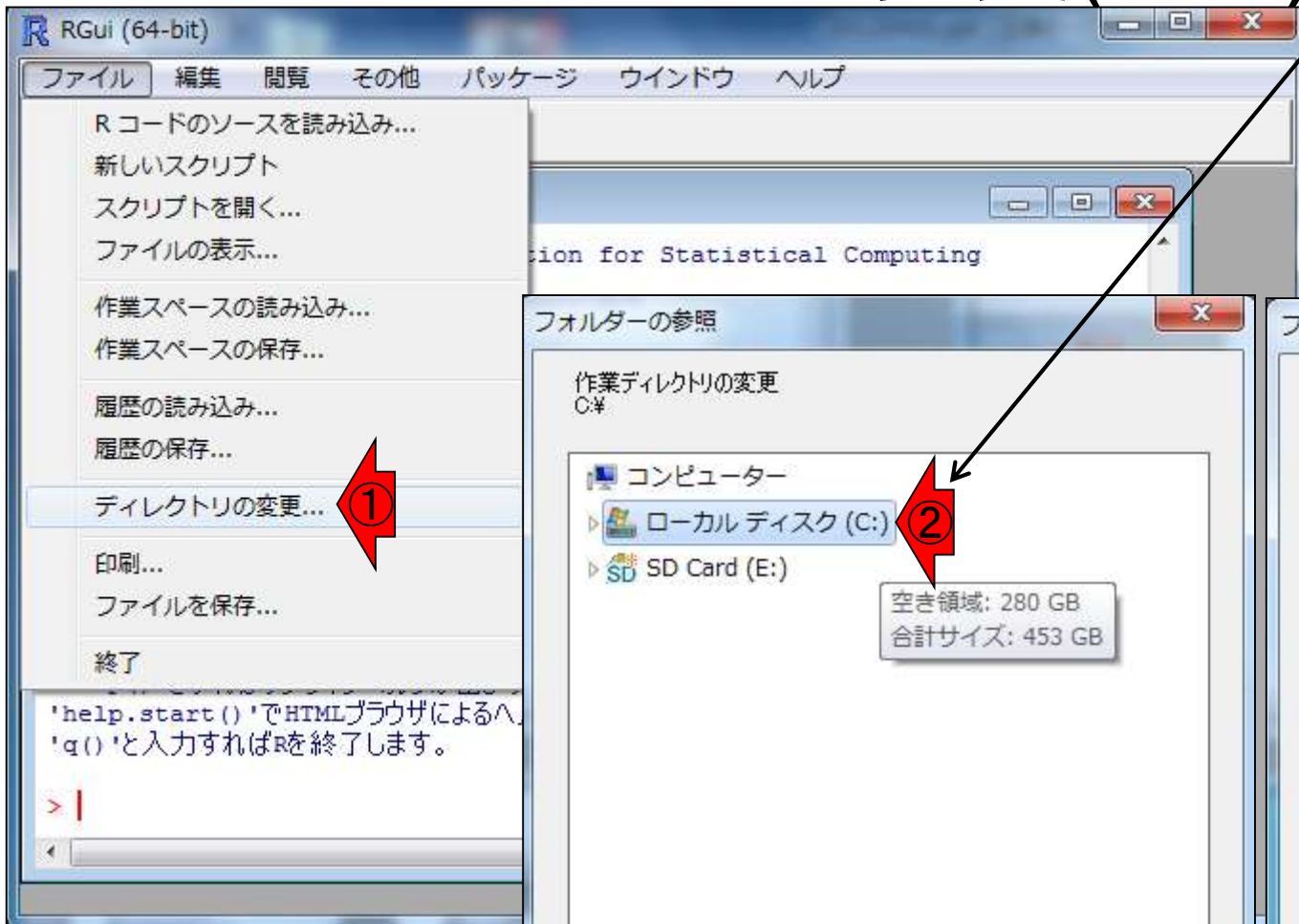
```





# 作業ディレクトリの変更 (Win)

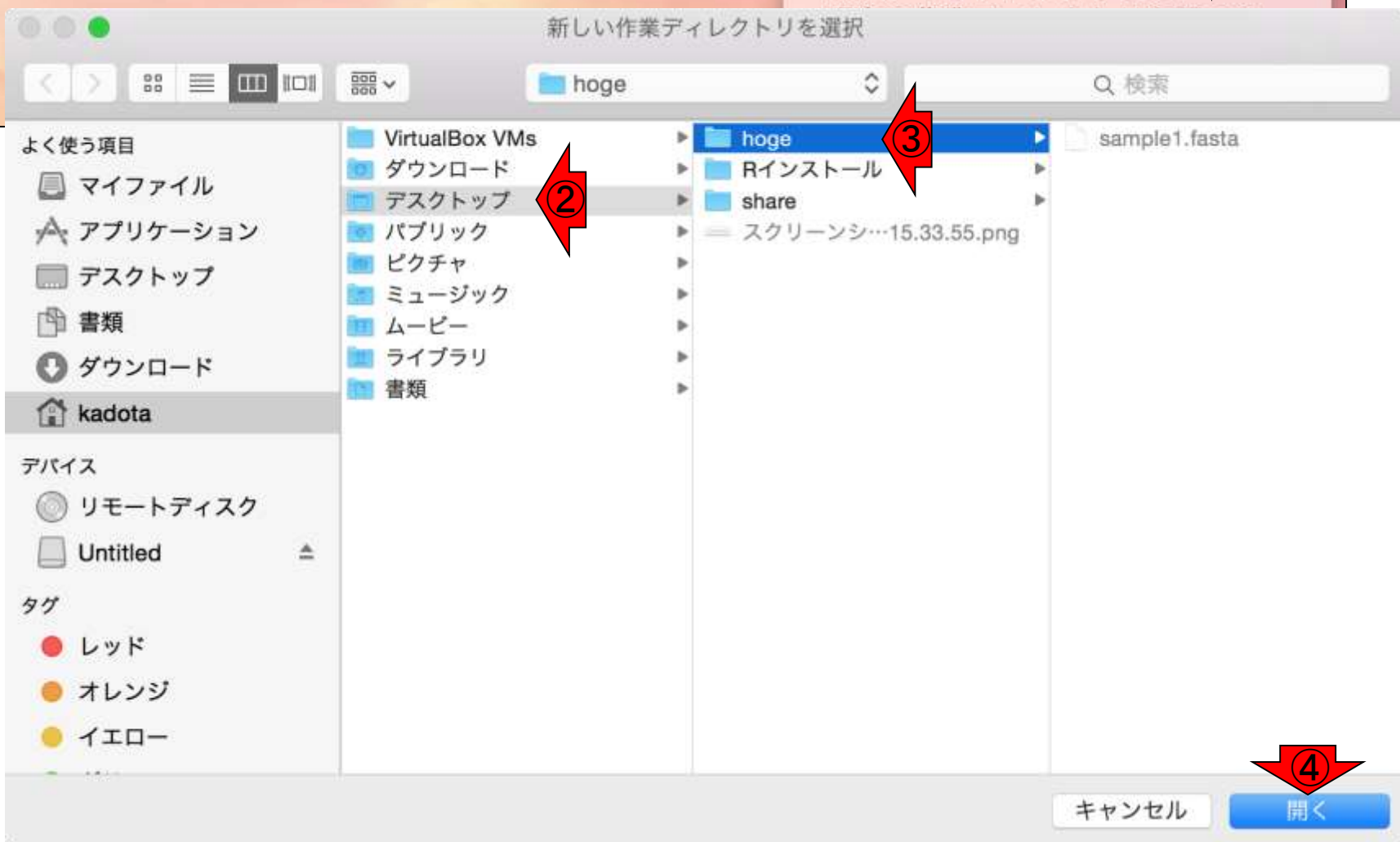
②「Windows(C:)」となっている場合もあるが、気にしない。④はヒトによって異なる





デスクトップのhoge  
を指定して、④開く

# 作業ディレクトリの変更(Mac)



# 確認

①getwd()で作業ディレクトリの確認、②list.files()で入力ファイルの存在確認。ここでは"blekh"というキーワードを含むファイルのみ表示させている。

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

[Blekhman et al., Genome Res., 2010](#)の 20,689 genes×18 samplesのカウントデータです。

```
in_f <- "sample_blekhman_18.txt" #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png" #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC) #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, #オブジェクト
dim(data))

#本番
out <- clusterSample(data, dist.method="spearman",
hclust.method="average", unique.patter

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2],
par(mar=c(0, 4, 1, 0)) #下、左、上、
plot(out, sub="", xlab="", cex.lab=1.2, #樹形図(デン
cex=1.3, main="", ylab="Height") #樹形図(デン
dev.off() #おまじない
```

```
R Console
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$
'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$
'q()' と入力すれば R を終了します。

[以前にセーブされたワークスペースを復帰します]

> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files(pattern="blekh")
[1] "sample_blekhman_18.txt"
> |
```

# コピーで実行

①一連のコマンド群をコピーして、②R Console画面上でペースト。ブラウザがInternet Explorerの場合は、CTRLとALTキーを押しながらコードの枠内で左クリックすると、全選択できます。

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×18 samplesのカウントデータです。

```
in_f <- "sample_blekhman_18.txt" #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png" #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400) #出力時の横幅と縦幅を指定(単位はピクセル)
```

```
#必要なパッケージをロード
library(TCC)
```

```
#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, as.is=TRUE)
dim(data)
```

```
#本番
out <- clusterSample(data, distance="j", method="hclust", hclust.method="a")
```

```
#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2],
par(mar=c(0, 4, 1, 0)))
plot(out, sub="", xlab="", cex=1.3, main="", ylab="Heatmap of sample_blekhman_18.txt")
dev.off()
```

A context menu is open over the code editor. The 'コピー(C)' (Copy) option is highlighted with a red arrow and the number 1. Other options include '貼り付け' (Paste), 'すべて選択(A)' (Select All), '印刷(I)...' (Print...), '印刷プレビュー(N)...' (Print Preview...), 'Bing でマップ' (Map with Bing), 'Bing で翻訳' (Translate with Bing), 'Google で検索' (Search with Google), '電子メール (Windows Live Mail)...' (Email...), 'すべてのアクセラレータ' (All Accelerators), and 'Send to OneNote'.

The R Console window is shown with a context menu open over the code. The 'コピー' (Copy) option is highlighted with a red arrow and the number 1, and the 'ペースト' (Paste) option is highlighted with a red arrow and the number 2. The console output shows the execution of 'getwd()' and 'list.files(pattern="blekh")', resulting in the path 'C:/Users/kadota/Desktop/hoge' and the file 'sample\_blekhman\_18.txt'.

# 実行結果

エラーなく実行できると右下のような画面になっているはず。①入力ファイル情報を格納した行列dataの行数が20,689、列数が18となっていることがわかります。

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

[Blekhman et al., Genome Res., 2010](#)の 20,689 genes×18 samplesのカウントデータです。

```

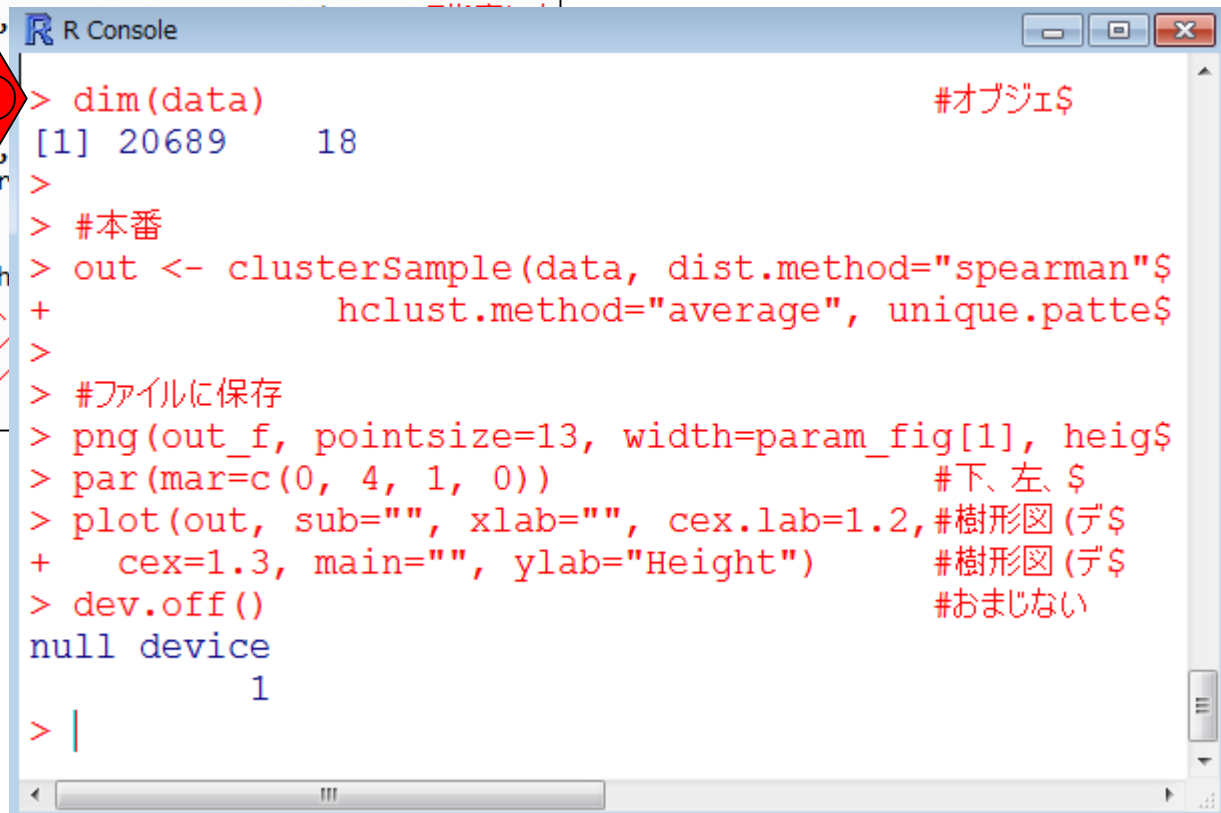
in_f <- "sample_blekhman_18.txt" #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png" #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC) #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, #オブジェクト
dim(data)

#本番
out <- clusterSample(data, dist.method="spearman",
hclust.method="average", unique.patter

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2],
par(mar=c(0, 4, 1, 0)) #下、左、上、右の余白を指定
plot(out, sub="", xlab="", cex.lab=1.2, #樹形図(デフォルト)
cex=1.3, main="", ylab="Height") #樹形図(デフォルト)
dev.off() #おまじない
    
```



```

R Console
> dim(data) #オブジェ$
[1] 20689 18
>
> #本番
> out <- clusterSample(data, dist.method="spearman"$
+ hclust.method="average", unique.patte$
>
> #ファイルに保存
> png(out_f, pointsize=13, width=param_fig[1], heig$
> par(mar=c(0, 4, 1, 0)) #下、左、$
> plot(out, sub="", xlab="", cex.lab=1.2, #樹形図(デ$
+ cex=1.3, main="", ylab="Height") #樹形図(デ$
> dev.off() #おまじない
null device
      1
> |
    
```

①出力ファイル名として指定したhoge8.pngの②情報を表示。

# 出力ファイル

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×18 samplesのカウントデータです。

```

in_f <- "sample_blekhman_18.txt" #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png" #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC) #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, #オブジェクト
dim(data))

#本番
out <- clusterSample(data, dist.method="spearman",
hclust.method="average", unique.patter

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2],
par(mar=c(0, 4, 1, 0)) #下、左、上、右の余白を指定
plot(out, sub="", xlab="", cex.lab=1.2, #樹形図(デフォルト)
+ cex=1.3, main="", ylab="Height") #樹形図(デフォルト)
dev.off() #おまじない
null device
1

```

```

R Console
> #ファイルに保存
> png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2],
> par(mar=c(0, 4, 1, 0)) #下、左、上、右の余白を指定
> plot(out, sub="", xlab="", cex.lab=1.2, #樹形図(デフォルト)
+ cex=1.3, main="", ylab="Height") #樹形図(デフォルト)
> dev.off() #おまじない
null device
1
> file.info("hoge8.png")
      size isdir mode          mtime
hoge8.png 5304 FALSE  666 2015-07-31 16:13:31
              ctime          atime
hoge8.png 2015-07-31 16:13:31 2015-07-31 16:13:31
              exe
hoge8.png no
> |

```

①出力ファイルのサイズを指定しているのので、こんな感じになる

# 出力ファイル

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×18 samplesのカウントデータです。

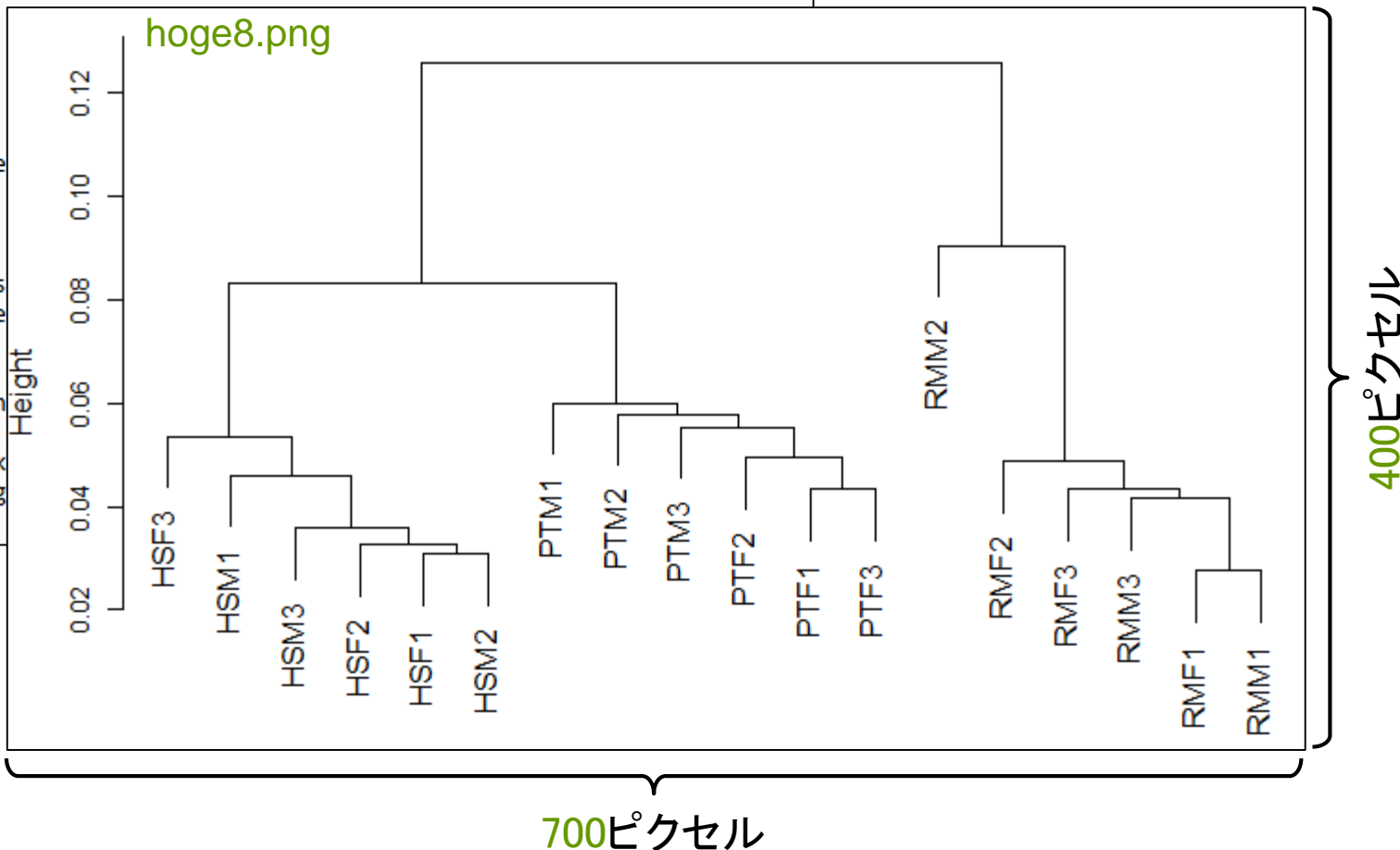
```
in_f <- "sample_blekhman_18.txt" #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png" #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)
```

```
#必要なパッケージをロード
library(TCC)

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, as.is=TRUE)
dim(data)

#本番
out <- clusterSample(data, distance="euclidean", hclust.method="ave")

#ファイルに保存
png(out_f, pointsize=13, width=700, height=400)
par(mar=c(0, 4, 1, 0))
plot(out, sub="", xlab="", cex=1.3, main="", ylab="Height", dev.off())
```





# Contents

## ■ サンプル間クラスタリング

- 実行手順のおさらい
- 計算の一部を解説、結果の解釈

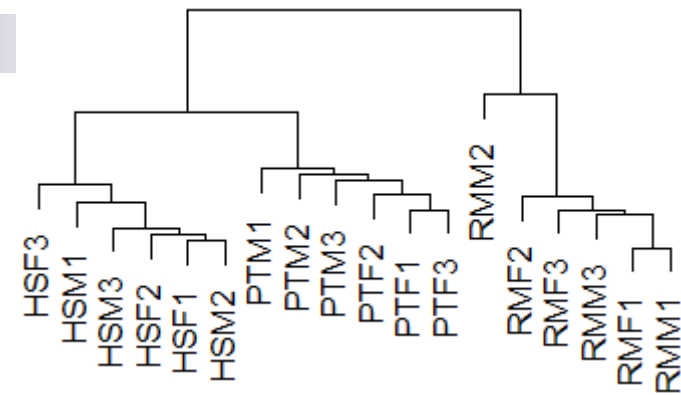
## ■ 発現変動解析(2群間比較) : 発現変動遺伝子(DEG)が多数存在する場合

## ■ モデル、分布、統計的手法、2群間比較でDEGがそれほど多くない場合

## ■ 2群間比較でDEGがほとんどない同一群の場合

## ■ 倍率変化(2倍以上、1/2以下の発現変動)を用いた場合

## ■ 発現変動解析: 3群間比較など



# clusterSample関数

TCCパッケージ中のclusterSample関数のオプションを説明。①ユニークな発現パターンのものでのみフィルタリング。実質的に低発現遺伝子のフィルタリングと同機能。②類似度は「1 - Spearman相関係数」。③平均連結法(average-linkage clustering)を利用してサンプル間クラスタリングしている

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×18 samplesのカウントデータ

```

in_f <- "sample_blekhman_18.txt"      #入力ファイル名を指定
out_f <- "hoge8.png"                 #出力ファイル名を指定
param_fig <- c(700, 400)              #ファイル出力時の横幅と縦幅

#必要なパッケージをロード
library(TCC)                          #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="") #in_fで指定した
dim(data)                              #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman", #クラスタリング実行結果をoutに格納
                     hclust.method="average", unique.pattern=TRUE) #クラスタリング実行結果をoutに格納

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2]) #出力ファイルの各種パラメータを指定
par(mar=c(0, 4, 1, 0)) #下、左、上、右の順で余白(行)を指定
plot(out, sub="", xlab="", cex.lab=1.2, #樹形図(デンドログラム)の表示
      cex=1.3, main="", ylab="Height") #樹形図(デンドログラム)の表示
dev.off()                              #おまじない
    
```



p137-145

# clusterSample関数

- ①「unique.pattern=TRUE」の実体を説明。
- ②rowSums関数を用いて遺伝子ごとの総カウント数を計算し、0でない(0より大きい)遺伝子のみ抽出。
- ③unique関数を用いてユニークな発現パターンのみ抽出。

8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×18 samplesのカウントデー

```

in_f <- "sample_blekhman_18.txt" #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png" #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#必要なパッケージをロード
library(TCC) #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", #オブジェクトdataの行数
dim(data)

#本番
out <- clusterSample(data, dist.method="spearman", #クラスタリ
hclust.method="average", unique.pattern=TRUE) #ク

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2],
par(mar=c(0, 4, 1, 0)) #下、左、上、右の順で余
plot(out, sub="", xlab="", cex.lab=1.2, #樹形図(デンドログラム)
cex=1.3, main="", ylab="Height") #樹形図(デンドログラム)
dev.off() #おまじない

```

```

R Console
> in_f <- "sample_blekhman_18.txt" $
> out_f <- "hoge8.png" $
> param_fig <- c(700, 400) $
>
> #必要なパッケージをロード
> library(TCC) $
>
> #入力ファイルの読み込み
> data <- read.table(in_f, header=$
> dim(data) $
[1] 20689 18
> obj <- (rowSums(data) > 0) $
> hoge <- unique(data[obj,]) $
> dim(hoge)
[1] 16560 18
> |

```



# 出力ファイル

縦軸の高さに相当する数値をどうやって計算しているのか?①最も類似度が高い(=距離が近い)RMF1とRMM1を例として説明。

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., *Genome Res.*, 2010の 20,689 genes×18 samplesのカウントデータです。

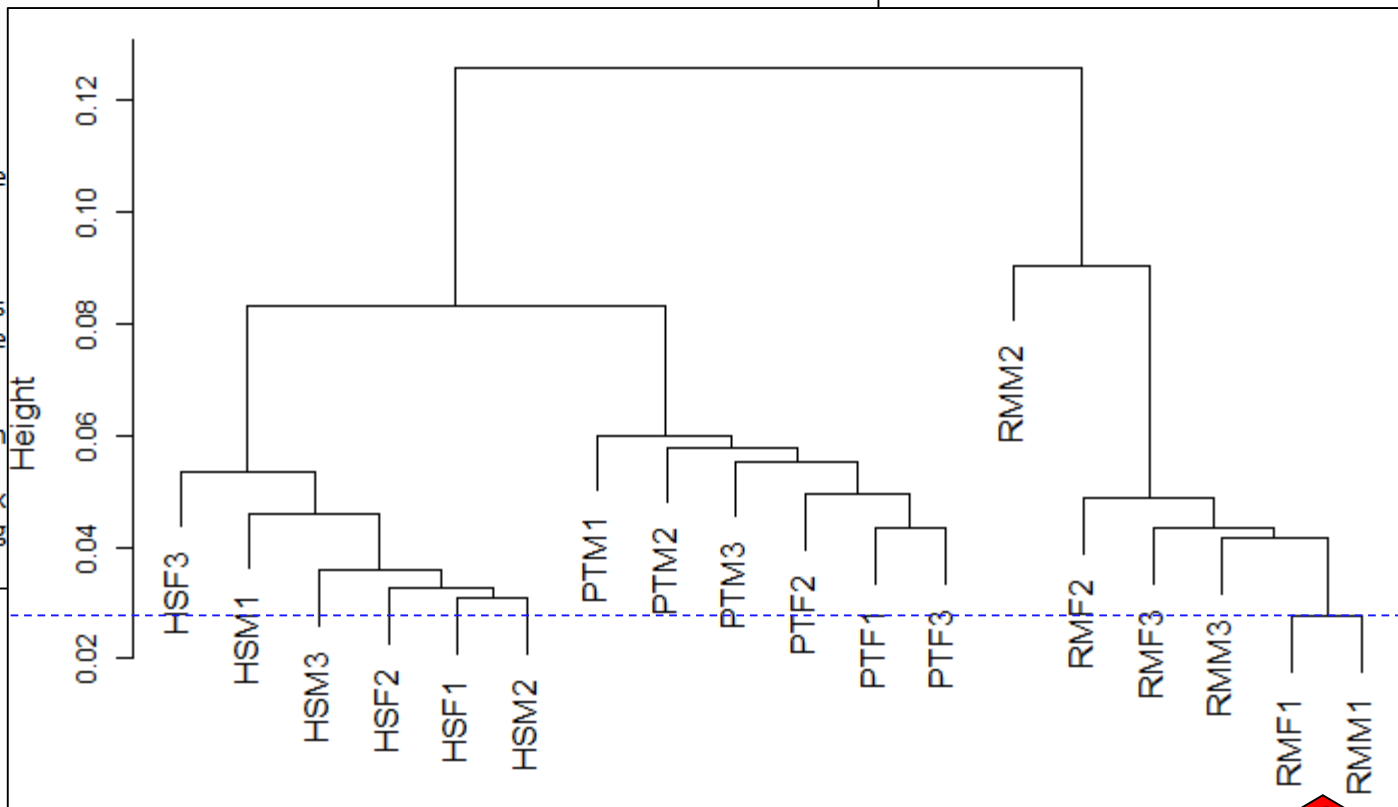
```
in_f <- "sample_blekhman_18.txt" #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png" #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)
```

```
#必要なパッケージをロード
library(TCC)
```

```
#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, as.is=TRUE)
dim(data)
```

```
#本番
out <- clusterSample(data, distance="euclidean", hclust.method="ave")
```

```
#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2])
par(mar=c(0, 4, 1, 0))
plot(out, sub="", xlab="", cex=1.3, main="", ylab="Height", dev.off())
```



# clusterSample関数

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合

Blekhman et al., Genome Res., 2010の 20,689 genes×18 samplesの場合

```
in_f <- "sample_blekhman_18.txt" #入力ファイル名
out_f <- "hoge8.png" #出力ファイル名
param_fig <- c(700, 400) #ファイル出力サイズ

#必要なパッケージをロード
library(TCC) #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="") #in_fで指定したファイルを読み込み
dim(data) #オブジェクトdataの行数と列数を表示

#本番
out <- clusterSample(data, dist.method="spearman", #クラスタリング実行結果をoutに格納
                    hclust.method="average", unique.pattern=TRUE) #クラスタリング実行結果をoutに格納

#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2],
    par(mar=c(0, 4, 1, 0)) #下、右、上、左のマージン
    plot(out, sub="", xlab="", cex.lab=1.2, #樹形図の表示
         cex=1.3, main="", ylab="Height") #樹形図の表示
    dev.off()) #出力終了
```

①「dist.method="spearman"」の実体を説明。② cor関数を用いてRMF1とRMM1ベクトル間の Spearman相関係数を計算した結果は0.9724165。相関係数は1(全く同じ発現パターン)から-1(真逆のパターン)までの値をとる。距離として取り扱いたい場合は、例えば③「1 - 相関係数」とすればよいので、それを距離として定義している。この場合の値の取りうる範囲は[0, 2]。今は最も似ているもの同士(RMF1とRMM1)の距離を調べているので、0に限りなく近い値(=0.02758346)になっている。

```
R Console
> dim(hoge)
[1] 16560 18
> cor(hoge$RMF1, hoge$RMM1, method="spearman")
[1] 0.9724165
> 1 - cor(hoge$RMF1, hoge$RMM1, method="spearman")
[1] 0.02758346
> |
```

①最も類似度が高い(=距離が近い)RMF1とRMM1の縦軸の値(=0.02758346)は、②妥当ですよね。

# 出力ファイル

## 8. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の 20,689 genes×18 samplesのカウントデータです。

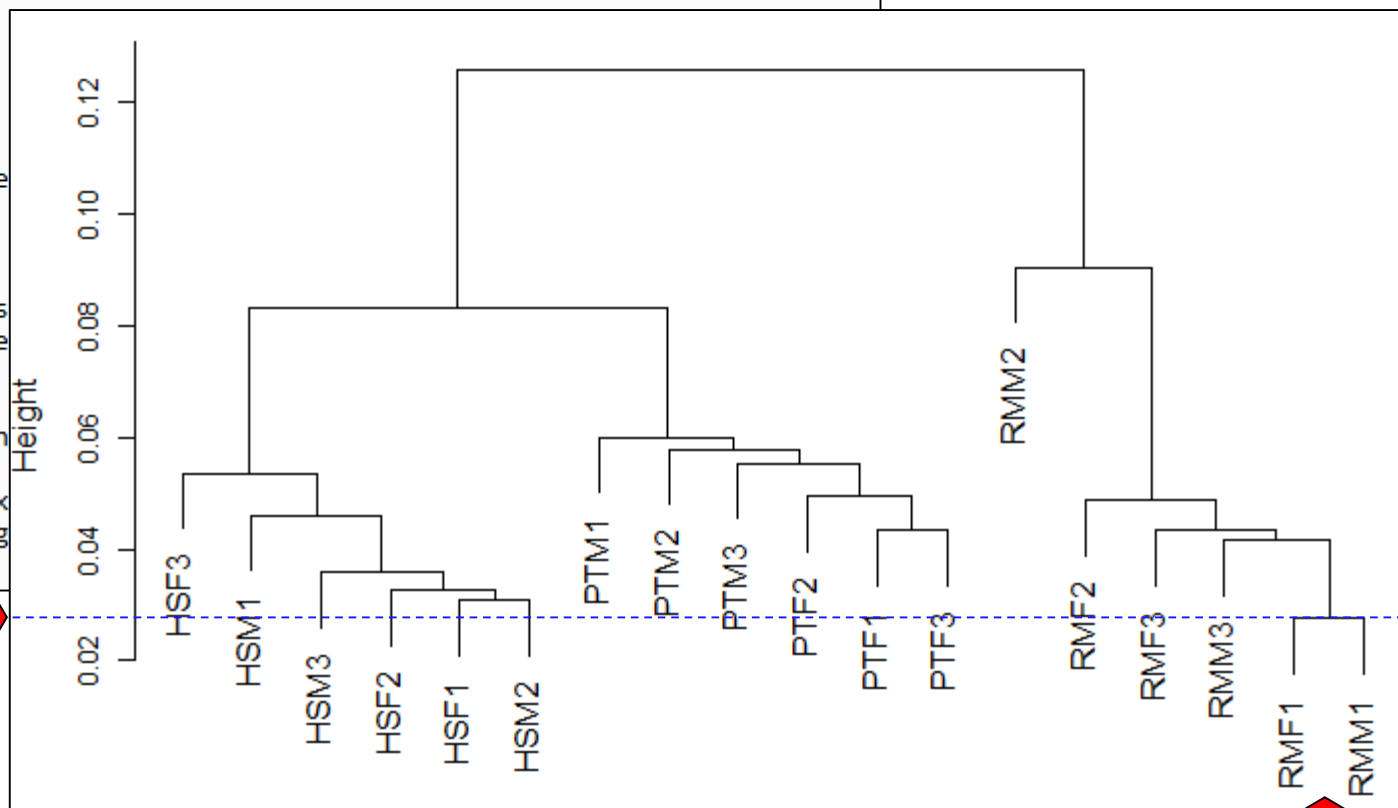
```
in_f <- "sample_blekhman_18.txt" #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png" #出力ファイル名を指定してout_fに格納
param_fig <- c(700, 400) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)
```

```
#必要なパッケージをロード
library(TCC)
```

```
#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, as.is=TRUE)
dim(data)
```

```
#本番
out <- clusterSample(data, distance="euclidean", hclust.method="ave")
```

```
#ファイルに保存
png(out_f, pointsize=13, width=param_fig[1], height=param_fig[2],
    par(mar=c(0, 4, 1, 0)), cex=1.3, main="", ylab="Height", dev.off())
```

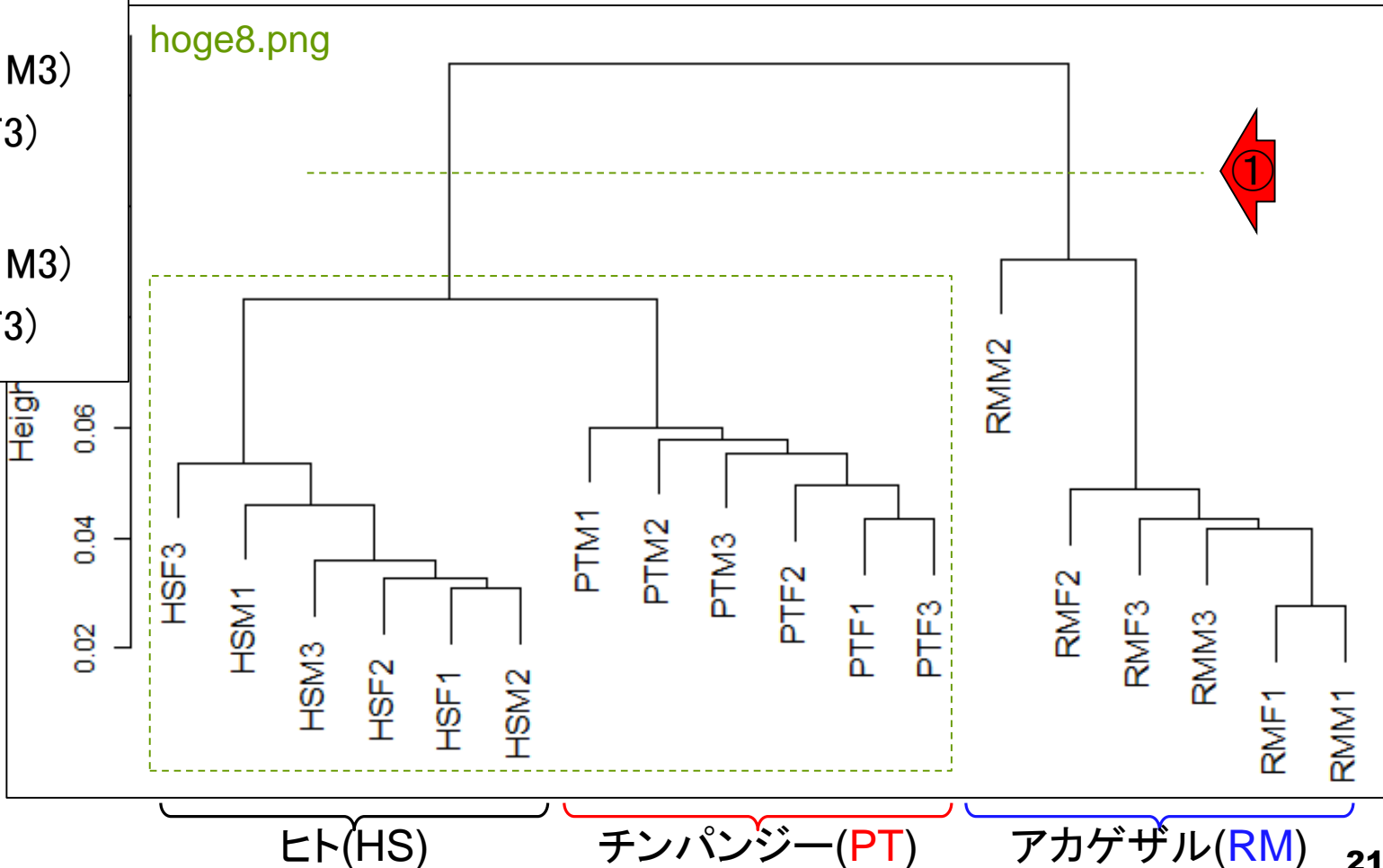




# 結果の解釈

3生物種間全体で眺めると、①ヒト(HS)とチンパンジー(PT)はよく似ている。2群間比較(発現変動遺伝子検出; DEG検出)を行ったときに、「HS vs. RMで得られるDEG数」のほうが「HS vs. PTで得られるDEG数」よりも多そう。

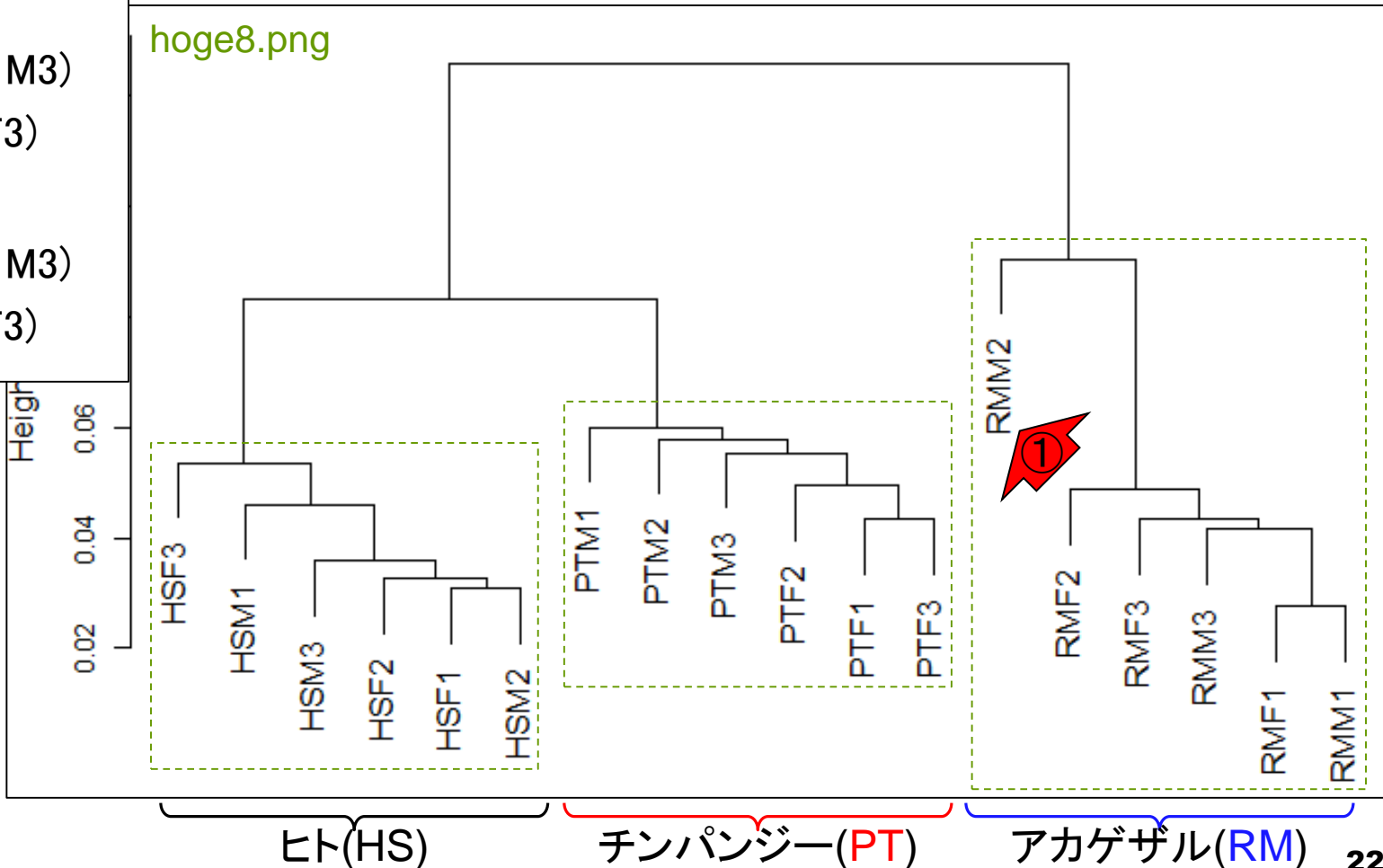
- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)



同一生物種でクラスターを形成している。  
①RMM2は「外れサンプル」っぽい。

# 結果の解釈

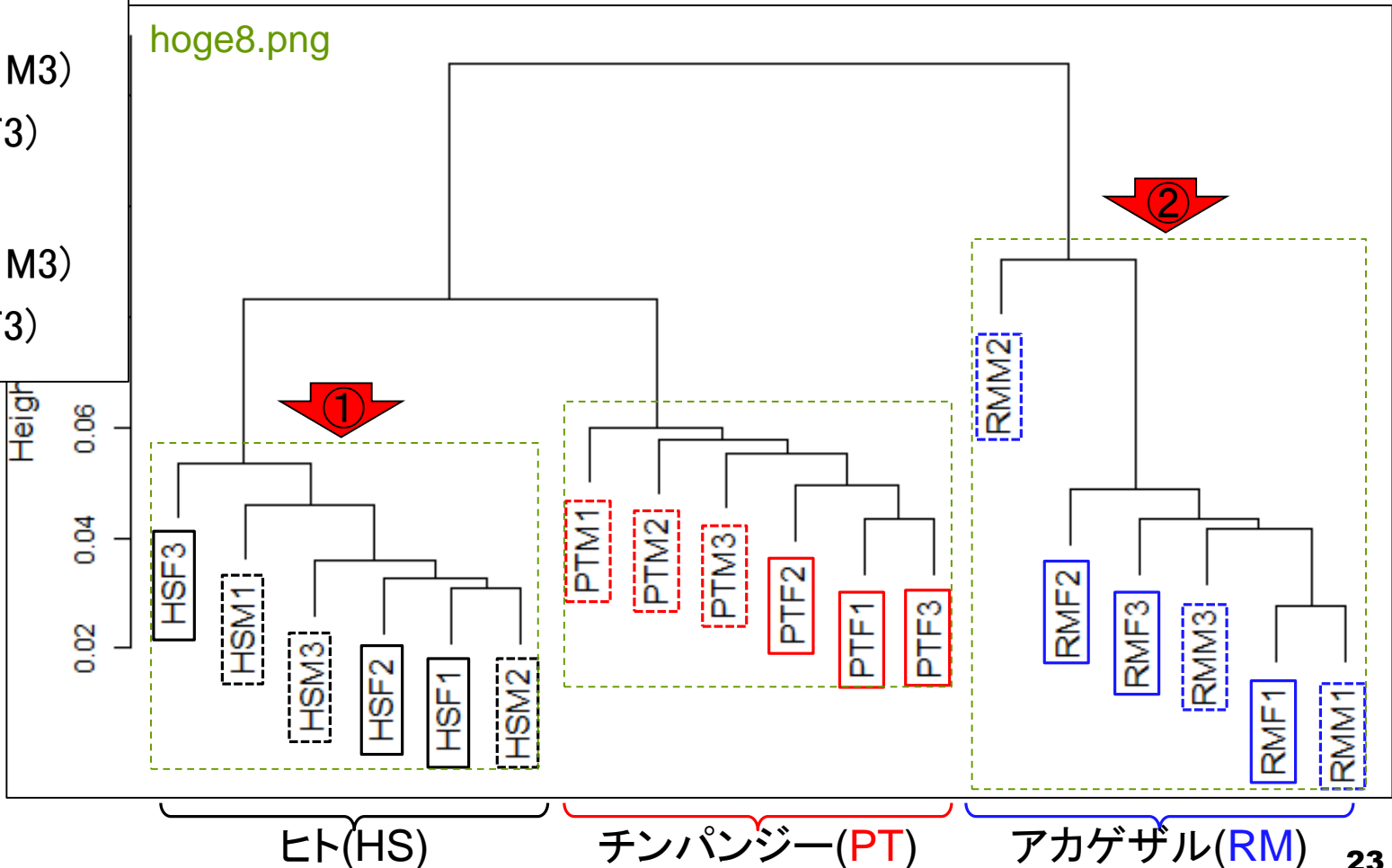
- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)



# 結果の解釈

①ヒト(HS)と②アカゲザル(RM)は、メスとオスのサンプルが入り混じっている。これらの生物種内で、「メス群 vs. オス群」の2群間比較を行ってもDEGはほとんど検出されないだろう。

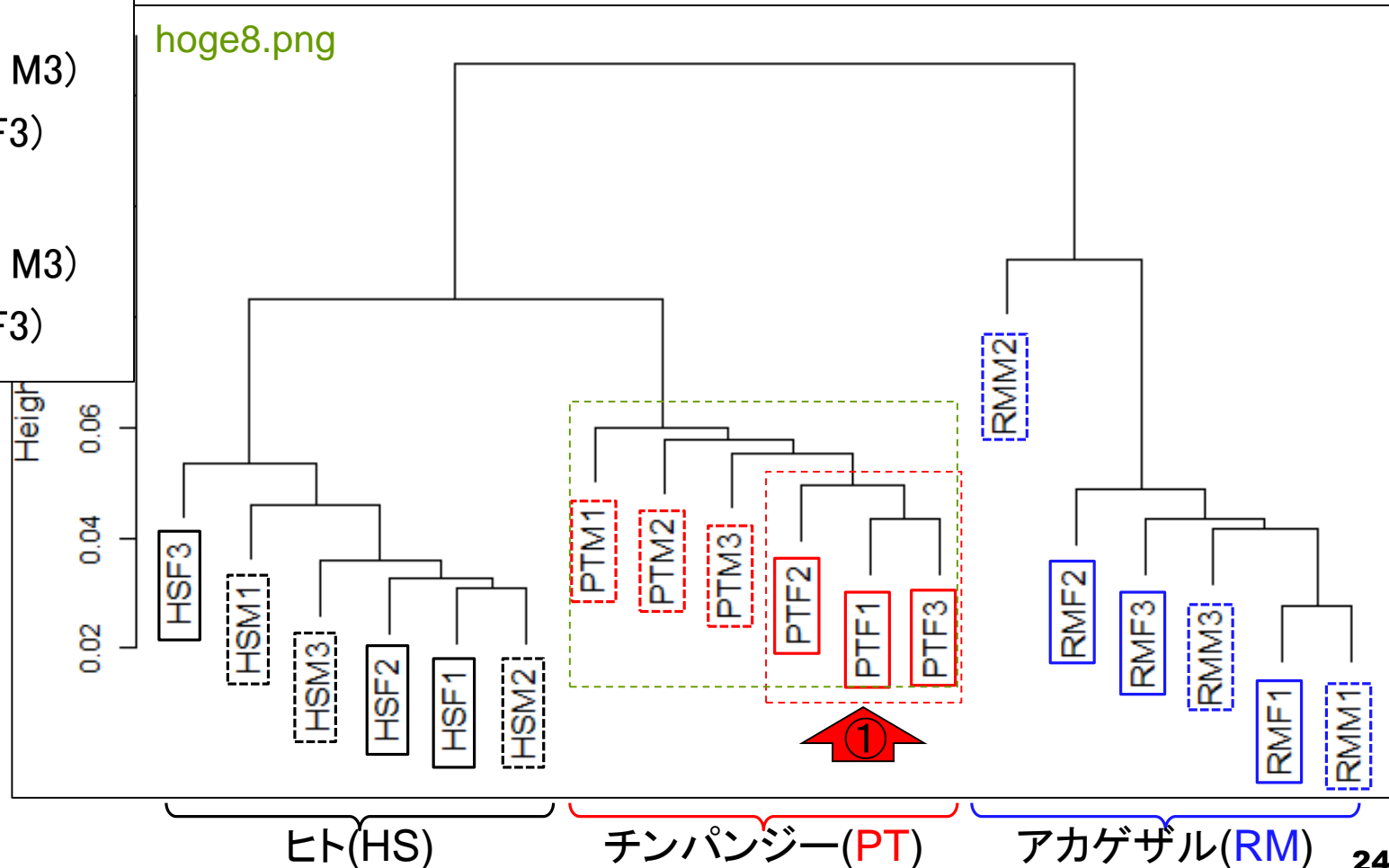
- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)



チンパンジー(PT)に限っていえば、①メス3匹がクラスターを形成しているので、「メス群 vs. オス群」の2群間比較結果として、多少なりともDEGが検出されるだろう。

# 結果の解釈

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)



# Contents

## ■ サンプル間クラスタリング

- 実行手順のおさらい
- 計算の一部を解説、結果の解釈

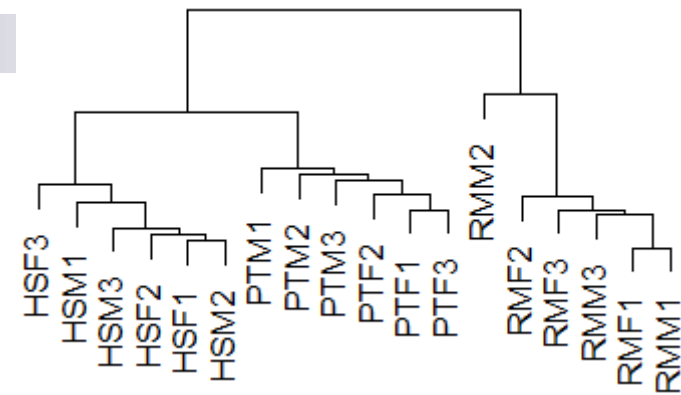
## ■ 発現変動解析(2群間比較) : 発現変動遺伝子(DEG)が多数存在する場合

## ■ モデル、分布、統計的手法、2群間比較でDEGがそれほど多くない場合

## ■ 2群間比較でDEGがほとんどない同一群の場合

## ■ 倍率変化(2倍以上、1/2以下の発現変動)を用いた場合

## ■ 発現変動解析: 3群間比較など







# サブセット抽出

1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル(G2群:RMF1とRMM1)の場合:

1, 4, 13, 16 列目のデータのみ抽出しています。

```
in_f <- "sample_blekhan_18.txt" #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge1.txt" #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge1.png" #出力ファイル名を指定してout_f2に格納
param_subset <- c(1, 4, 13, 16) #取り扱いたいサブセット情報を指定
param_G1 <- 2 #G1群のサンプル数を指定
param_G2 <- 2 #G2群のサンプル数を指定
param_FDR <- 0.05 #DEG検出時のfalse discovery rate (FDR)を指定
param_fig <- c(430, 350) #ファイル出力時の横幅と縦幅を指定(単位はpixel)
param_mar <- c(4, 4, 0, 0) #下、左、上、右の順番で余白を指定(単位はpixel)
```

```
#必要なパッケージをロード
library(TCC) #パッケージの読み込み
```

```
#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep=" ")
```

```
#前処理(サブセットの抽出とTCCクラスオブジェクトの作成)
```

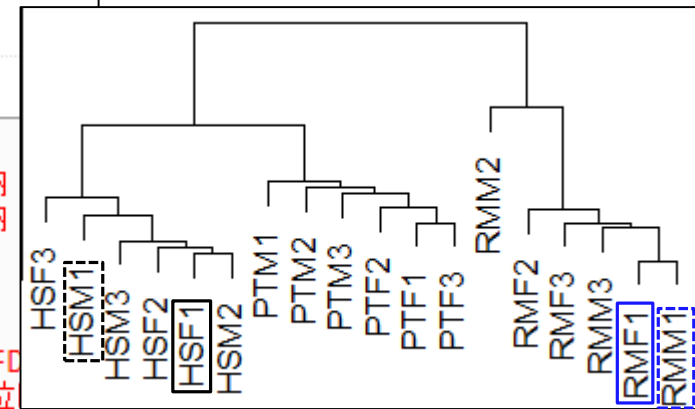
```
data <- data[,param_subset] #param_subsetで指定した列のみを抽出
```

```
data.cl <- c(rep(1, param_G1), rep(2, param_G2)) #G1群を1, G2群を2で指定
```

```
tcc <- new("TCC", data, data.cl) #TCCクラスオブジェクトの作成
```

```
dim(data) #行数と列数を表示
```

```
head(data) #最初の6行分を表示
```



```
R Console
> #前処理(サブセットの抽出とTCCクラスオブジェクトの作成)
> data <- data[,param_subset] # $
> data.cl <- c(rep(1, param_G1), rep(2, param_G2)) # $
> tcc <- new("TCC", data, data.cl) # $
> dim(data) # $
[1] 20689 4
> head(data) # $
      HSF1 HSM1 RMF1 RMM1
ENSG000000000003 329 121 511 424
ENSG000000000005 0 0 0 2
ENSG000000000419 81 39 67 49
ENSG000000000457 91 114 89 117
ENSG000000000460 6 15 4 7
ENSG000000000938 44 73 73 80
>
```

# サブセット抽出

①ここで取得したいサブセットの列番号やグループ情報を指定。②発現変動解析に用いるサブセットは20,689 genes × 4 samplesのデータ。③正しくヒト vs. アカゲザルになっていることが分かる。

1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル(G2群:RMF1とRMM1)

1, 4, 13, 16 列目のデータのみ抽出しています。

```
in_f <- "sample_blekman_18.txt"
out_f1 <- "hoge1.txt"
out_f2 <- "hoge1.png"
param_subset <- c(1, 4, 13, 16)
param_G1 <- 2
param_G2 <- 2
param_FDR <- 0.05
param_fig <- c(430, 350)
param_mar <- c(4, 4, 0, 0)
```



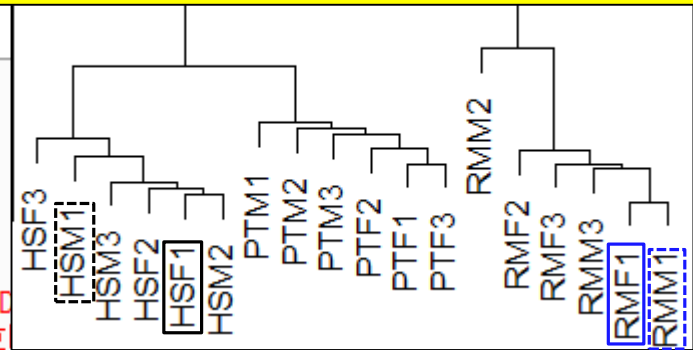
#入力ファイル名を指定してin\_fに格納  
 #出力ファイル名を指定してout\_f1に格納  
 #出力ファイル名を指定してout\_f2に格納  
 #取り扱いたいサブセット情報を指定  
 #G1群のサンプル数を指定  
 #G2群のサンプル数を指定  
 #DEG検出時のfalse discovery rate (FDR)を指定  
 #ファイル出力時の横幅と縦幅を指定(単位はpx)  
 #下、左、上、右の順番で余白を指定(単位はpx)

```
#必要なパッケージをロード
library(TCC)

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep=" ")

#前処理(サブセットの抽出とTCCクラスオブジェクトの作成)
data <- data[,param_subset] #param_subsetで指定した列のみを抽出
data.cl <- c(rep(1, param_G1), rep(2, param_G2)) #G1群を1, G2群を2で指定
tcc <- new("TCC", data, data.cl) #TCCクラスオブジェクトの作成
dim(data) #行数と列数を表示
head(data) #最初の6行分を表示
```

#パッケージの読み込み  
 #param\_subsetで指定した列のみを抽出  
 #G1群を1, G2群を2で指定  
 #TCCクラスオブジェクトの作成  
 #行数と列数を表示  
 #最初の6行分を表示



```
R Console
> #前処理(サブセットの抽出とTCCクラスオブジェクトの作成)
> data <- data[,param_subset] # $
> data.cl <- c(rep(1, param_G1), rep(2, param_G2)) # $
> tcc <- new("TCC", data, data.cl) # $
> dim(data) # $
[1] 20689 4 # $
> head(data) # $
HSF1 HSM1 RMF1 RMM1
ENSG000000000003 329 121 511 424
ENSG000000000005 0 0 0 2
ENSG000000000419 81 39 67 49
ENSG000000000457 91 114 89 117
ENSG000000000460 6 15 4 7
ENSG000000000938 44 73 73 80
>
```



# サブセット抽出

入力ファイル(sample\_blekhman\_18.txt)を眺めるなどして、①該当サンプルの列の位置を把握していることが前提。

1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル(G2群:RMF1とRMM1)の場合:

1, 4, 13, 16 列目のデータのみ抽出しています。

```
in_f <- "sample_blekhman_18.txt"
out_f1 <- "hoge1.txt"
out_f2 <- "hoge1.png"
param_subset <- c(1, 4, 13, 16)
param_G1 <- 2
param_G2 <- 2
param_FDR <- 0.05
param_fig <- c(430, 350)
param_mar <- c(4, 4, 0, 0)
```



#入力ファイル名を指定してin\_fに格納  
 #出力ファイル名を指定してout\_f1に格納  
 #出力ファイル名を指定してout\_f2に格納  
 #取り扱いたいサブセット情報を指定  
 #G1群のサンプル数を指定  
 #G2群のサンプル数を指定  
 #DEG検出時のfalse discovery rate (FDR)を指定  
 #ファイル出力時の横幅と縦幅を指定(単位は行)  
 #下、左、上、右の順で余白を指定(単位は行)

#必要なパッケージをロード  
 library(TCC)

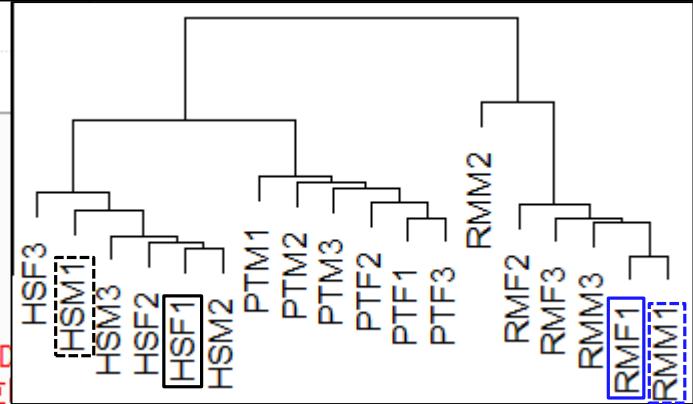
#パッケージの読み込み

#入力ファイルの読み込み

data <- read.table(in\_f, header=TRUE, row.names=1, sep="\t", quote="")#in\_fで

#前処理

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
data	HSF1	HSF2	HSF3	HSM1	HSM2	HSM3	PTF1	PTF2	PTF3	PTM1	PTM2	PTM3	RMF1	RMF2	RMF3	RMM1	RMM2	RMM3	
data	ENSG000000000003	329	300	168	121	421	359	574	429	386	409	685	428	511	464	480	424	1348	705
tcc <	ENSG000000000005	0	0	0	0	1	0	1	4	1	0	1	1	0	1	2	2	0	0
dim(d	ENSG000000000419	81	61	56	39	78	62	100	66	65	59	58	93	67	72	57	49	82	90
head(c	ENSG000000000457	91	62	76	114	73	95	131	229	87	274	239	149	89	69	118	117	114	163
<	ENSG000000000460	6	17	12	15	7	17	8	8	5	12	7	10	4	4	10	7	3	4
	ENSG000000000938	44	65	210	73	43	65	84	104	76	198	31	58	73	28	54	80	34	72
	ENSG000000000971	4765	7225	3405	3600	6383	5546	5382	8331	4335	2568	5019	2653	13566	9064	18247	14236	5196	11834



# FDR

①  $q < 0.05$ を満たす遺伝子数は2,488個。False discovery rate (FDR) = 0.05は、この閾値を満たす2,488個を発現変動遺伝子(Differentially Expressed Genes; DEGs)とみなすと、 $2,488 * 0.05 = 124.4$ 個は偽物であることを意味する。有意水準(false positive rate; FPR)5%と同じような位置づけであり、FDR5%というのは、「許容する偽物(non-DEG)混入割合」に相当する。詳細は2015.05.26の講義資料を参照のこと。

1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル  
1, 4, 13, 16 列目のデータのみ抽出しています。

```

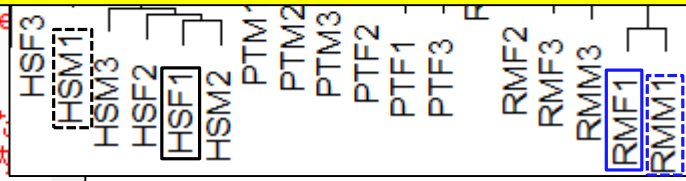
#本番(正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", iteration=3, FDR=0.1)
normalized <- getNormalizedData(tcc) #正規化後のデータを取り出してnormalized

#本番(DEG検出)
tcc <- estimateDE(tcc, test.method="edgeR", FDR=param_FDR) #DEG検出を実行した
result <- getResult(tcc, sort=FALSE) #p値などの結果をした結果をresultに格納
sum(tcc$stat$q.value < param_FDR) #FDR < param_FDRを満たす遺伝子数を表示

#ファイルに保存(テキストファイル)
tmp <- cbind(rownames(tcc$count), normalized, result) #正規化後のデータとDEG検出結果を結合
tmp <- tmp[order(tmp$rank),] #発現変動順にソート
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=FALSE)

#ファイルに保存(M-A plot)
png(out_f2, pointsize=13, width=param_fig[1], height=param_fig[2])
par(mar=param_mar) #余白を指定
plot(tcc, FDR=param_FDR, xlim=c(-2, 17), ylim=c(-10, 10)) #param_FDRで指定
cex=0.9, cex.lab=1.2, #param_FDRで指定
cex.axis=1.2, main="", #param_FDRで指定
xlab="A = (log2(G2) + log2(G1))/2", #param_FDRで指定

```



```

R Console
+ iteration=3, FDR=$
TCC::INFO: Calculating normalization factors
TCC::INFO: (iDEGES pipeline : tmm - [ edgeR ]
TCC::INFO: Done.
> normalized <- getNormalizedData(tcc) # $
>
> #本番 (DEG検出)
> tcc <- estimateDE(tcc, test.method="edgeR", FDR=param_FDR) # $
TCC::INFO: Identifying DE genes using edgeR
TCC::INFO: Done.
> result <- getResult(tcc, sort=FALSE) # $
> sum(tcc$stat$q.value < param_FDR) # $
[1] 2488
>
> |

```





# FDR

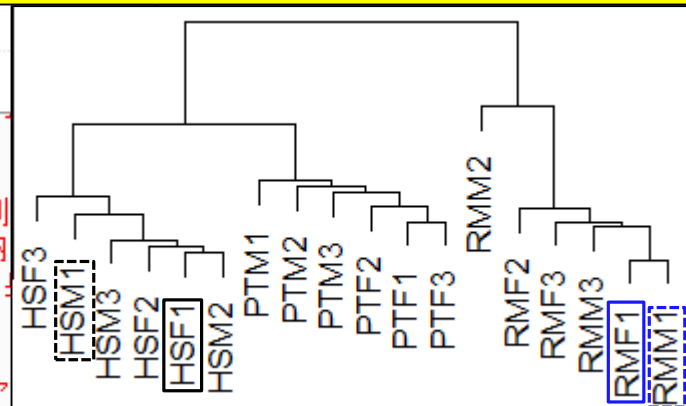
$q < 0.30$ を満たす遺伝子数は4,786個。  
 FDR = 0.30なので、 $4,786 * 0.30 = 1,435.8$ 個は偽物で残りの70%は本物だと判断する。

1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル(G2群:RMF1とRMM1)の場合:

1, 4, 13, 16 列目のデータのみ抽出しています。

```
#ファイルに保存(テキストファイル)
tmp <- cbind(rownames(tcc$count), normalized, result)#正規化後のデータの右側
tmp <- tmp[order(tmp$rank),] #発現変動順にソートした結果をtmpに格納
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F)#tmpの中

#ファイルに保存(M-A plot)
png(out_f2, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイル
par(mar=param_mar) #余白を指定
plot(tcc, FDR=param_FDR, xlim=c(-2, 17), ylim=c(-10, 10), #param_FDRで指定
      cex=0.8, cex.lab=1.2, #param_FDRで指定
      cex.axis=1.2, main="", #param_FDRで指定
      xlab="A = (log2(G2) + log2(G1))/2", #param_FDRで指定
      ylab="M = log2(G2) - log2(G1)" #param_FDRで指定)
legend("topright", c(paste("DEG(FDR<", param_FDR, ")"), #凡例
                      col=c("magenta", "black"), pch=20, cex=1.2)
dev.off() #おまじない
sum(tcc$stat$q.value < 0.05) #FDR < 0.05を満たす遺伝子数
sum(tcc$stat$q.value < 0.10) #FDR < 0.10を満たす遺伝子数
sum(tcc$stat$q.value < 0.20) #FDR < 0.20を満たす遺伝子数
sum(tcc$stat$q.value < 0.30) #FDR < 0.30を満たす遺伝子数
```



```
R Console
+ ylab="M = log2(G2) - log2(G1)" # $
> legend("topright", c(paste("DEG(FDR<", p$
+ col=c("magenta", "black"), pch=20$
> dev.off() # $
null device
1
> sum(tcc$stat$q.value < 0.05) # $
[1] 2488
> sum(tcc$stat$q.value < 0.10) # $
[1] 3122
> sum(tcc$stat$q.value < 0.20) # $
[1] 4049
> sum(tcc$stat$q.value < 0.30) # $
[1] 4786
> |
```



FDR閾値が比較的緩めのところを眺め、20,689 genes中3,300個程度が本物のDEGと判断する。

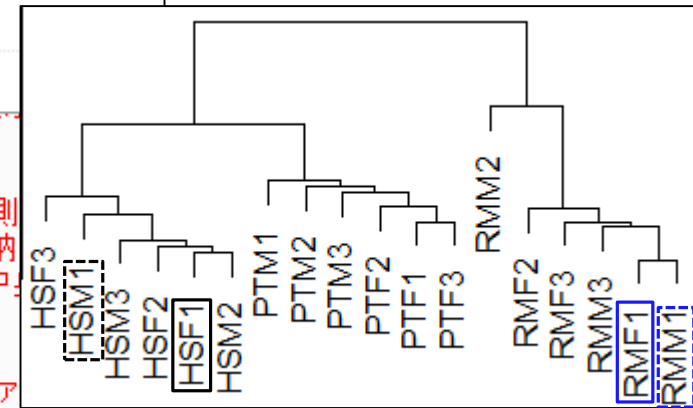
# DEG数の見積もり

1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル(G2群:RMF1とRMM1)の場合:

1, 4, 13, 16 列目のデータのみ抽出しています。

```
#ファイルに保存(テキストファイル)
tmp <- cbind(rownames(tcc$count), normalized, result)#正規化後のデータの右側
tmp <- tmp[order(tmp$rank),]#発現変動順にソートした結果をtmpに格納
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F)#tmpの中

#ファイルに保存(M-A plot)
png(out_f2, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイル
par(mar=param_mar)#余白を指定
plot(tcc, FDR=param_FDR, xlim=c(-2, 17), ylim=c(-10, 10),#param_FDRで指定した閾値を
      cex=0.8, cex.lab=1.2,#param_FDRで指定した閾値を満たすDEGをマゼンタ
      cex.axis=1.2, main="",#param_FDRで指定した閾値を満たすDEGをマゼンタ
      xlab="A = (log2(G2) + log2(G1))/2",#param_FDRで指定した閾値を満たすDEGをマゼンタ
      ylab="M = log2(G2) - log2(G1)"#param_FDRで指定した閾値を満たすDEGをマゼンタ
      legend("topright", c(paste("DEG(FDR<", param_FDR, ")"), "non-DEG"),#凡例を作成
      col=c("magenta", "black"), pch=20, cex=1.2)#凡例を作成
dev.off()#おまじない
sum(tcc$stat$q.value < 0.05)#FDR < 0.05を満たすDEG数
sum(tcc$stat$q.value < 0.10)#FDR < 0.10を満たすDEG数
sum(tcc$stat$q.value < 0.20)#FDR < 0.20を満たすDEG数
sum(tcc$stat$q.value < 0.30)#FDR < 0.30を満たすDEG数
```



```
R Console
> 2488*(1 - 0.05)
[1] 2363.6
> 3122*(1 - 0.10)
[1] 2809.8
> 4049*(1 - 0.20)
[1] 3239.2
> 4786*(1 - 0.30)
[1] 3350.2
> |
```



# 樹形図と一致

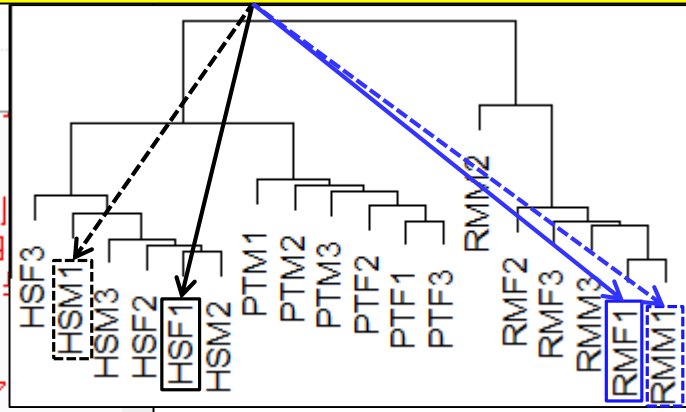
今比較しているのはHS vs. RM。クラスタリング結果からも、これらの発現プロファイルの類似度が低い(距離が遠い)ので妥当

1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル(G2群:RMF1とRMM1)の場合:

1, 4, 13, 16 列目のデータのみ抽出しています。

```
#ファイルに保存(テキストファイル)
tmp <- cbind(rownames(tcc$count), normalized, result)#正規化後のデータの右側
tmp <- tmp[order(tmp$rank),]#発現変動順にソートした結果をtmpに格納
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F)#tmpの中

#ファイルに保存(M-A plot)
png(out_f2, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイル
par(mar=param_mar)#余白を指定
plot(tcc, FDR=param_FDR, xlim=c(-2, 17), ylim=c(-10, 10),#param_FDRで指定した閾値
      cex=0.8, cex.lab=1.2,#param_FDRで指定した閾値を満たすDEGをマゼンタ
      cex.axis=1.2, main="",#param_FDRで指定した閾値を満たすDEGをマゼンタ
      xlab="A = (log2(G2) + log2(G1))/2",#param_FDRで指定した閾値を満たすDEGをマゼンタ
      ylab="M = log2(G2) - log2(G1)"#param_FDRで指定した閾値を満たすDEGをマゼンタ
      legend("topright", c(paste("DEG(FDR<", param_FDR, ")"), "non-DEG"),#凡例を
      col=c("magenta", "black"), pch=20, cex=1.2)#凡例を作成
dev.off()#おまじない
sum(tcc$stat$q.value < 0.05)#FDR < 0.05を満たすDEGの数
sum(tcc$stat$q.value < 0.10)#FDR < 0.10を満たすDEGの数
sum(tcc$stat$q.value < 0.20)#FDR < 0.20を満たすDEGの数
sum(tcc$stat$q.value < 0.30)#FDR < 0.30を満たすDEGの数
```



```
R Console
> 2488*(1 - 0.05)
[1] 2363.6
> 3122*(1 - 0.10)
[1] 2809.8
> 4049*(1 - 0.20)
[1] 3239.2
> 4786*(1 - 0.30)
[1] 3350.2
> |
```

# M-A plot

これがM-A plot。発現変動遺伝子(DEG)と判定されたものが多数存在することがわかる。param\_FDRで指定した閾値(0.05)を満たす遺伝子群がマゼンタ色で表示されている。

1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル(G2群:RMF1とF

1, 4, 13, 16 列目のデータのみ抽出しています。

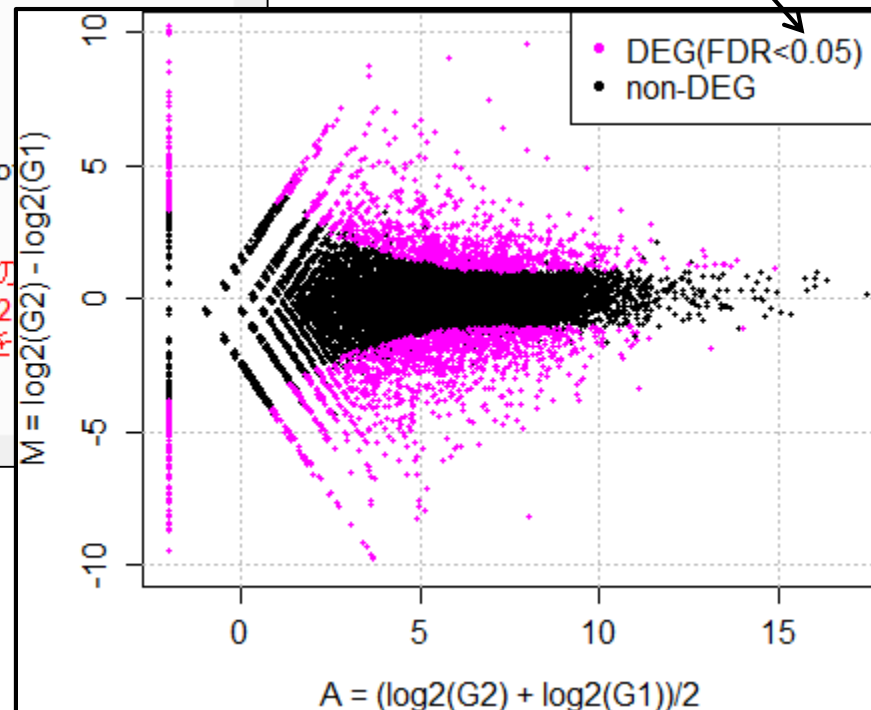
```
in_f <- "sample_blekman_18.txt"
out_f1 <- "hoge1.txt"
out_f2 <- "hoge1.png"
param_subset <- c(1, 4, 13, 16)
param_G1 <- 2
param_G2 <- 2
param_FDR <- 0.05
param_fig <- c(430, 350)
param_mar <- c(4, 4, 0, 0)
```

#入力ファイル名を指定してin\_fに格納  
#出力ファイル名を指定してout\_f1に格納  
#出力ファイル名を指定してout\_f2に格納  
#取り扱いたいサブセット情報を指定  
#G1群のサンプル数を指定  
#G2群のサンプル数を指定  
#DEG検出時のfalse discovery rate (FDR)  
#ファイル出力時の横幅と縦幅を指定(単位はピクセル)  
#下、左、上、右の順で余白を指定(単位は行)

```
#必要なパッケージをロード
library(TCC)
#パッケージの読み込み
```

```
#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quo
```

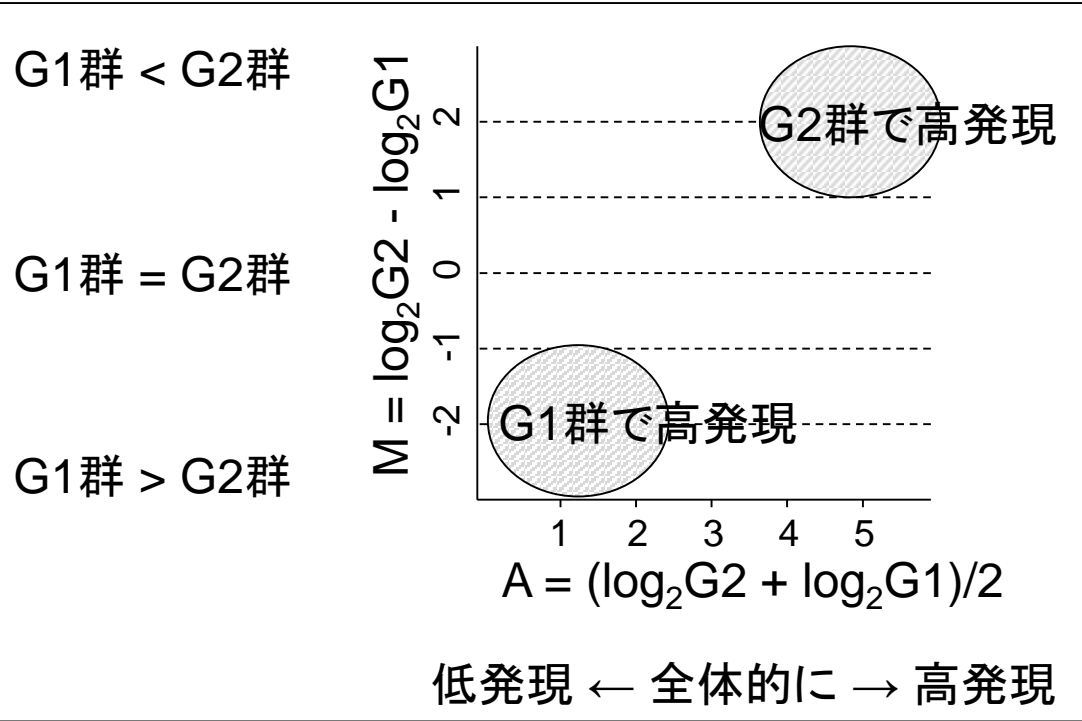
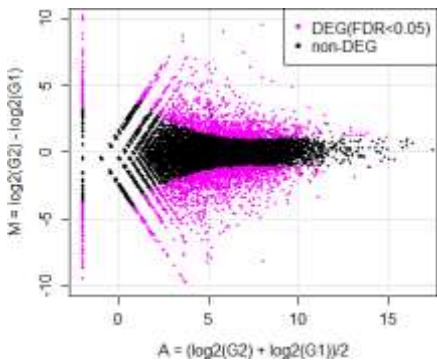
```
#前処理(サブセットの抽出とTCCクラスオブジェクトの作成)
data <- data[,param_subset]
data.cl <- c(rep(1, param_G1), rep(2, param_G2))
tcc <- new("TCC", data, data.cl)
dim(data)
head(data)
#param_subsetで指定した列の抽出
#G1群を1, G2群を2で指定
#TCCクラスオブジェクトtccを作成
#行数と列数を表示
#最初の6行分を表示
```



DEGが存在しないデータのM-A plotを眺めることで、縦軸の閾値のみに相当する倍率変化を用いたDEG同定の危険性が分かります

# M-A plot

- 2群間比較用
- 横軸が全体的な発現レベル、縦軸がlog比からなるプロット
- 名前の由来は、おそらく対数の世界での縦軸が引き算 (Minus)、横軸が平均 (Average)



# DEG検出結果

1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル(G2群:RMF1とRMM1)の場合:

1, 4, 13, 16 列目のデータのみ抽出しています。

```

in_f <- "sample blekhan_18.txt"
out_f1 <- "hogel.txt"
out_f2 <- "hogel.png"
param_subset <- c(1, 4, 13, 16)
param_G1 <- 2
param_G2 <- 2
param_FDR <- 0.05
param_fig <- c(430, 350)
param_mar <- c(4, 4, 0, 0)

#必要なパッケージをロード
library(TCC)

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで
    
```

#入力ファイル名を指定してin\_fに格納  
 #出力ファイル名を指定してout\_f1に格納  
 #出力ファイル名を指定してout\_f2に格納  
 #取り扱いたいサブセット情報を指定  
 #G1群のサンプル数を指定  
 #G2群のサンプル数を指定  
 #DEG検出時のfalse discovery rate (FDR)  
 #ファイル出力時の横幅と縦幅を指定(単位はピクセル)  
 #下、左、上、右の順で余白を指定(単位は行)

#パッケージの読み込み

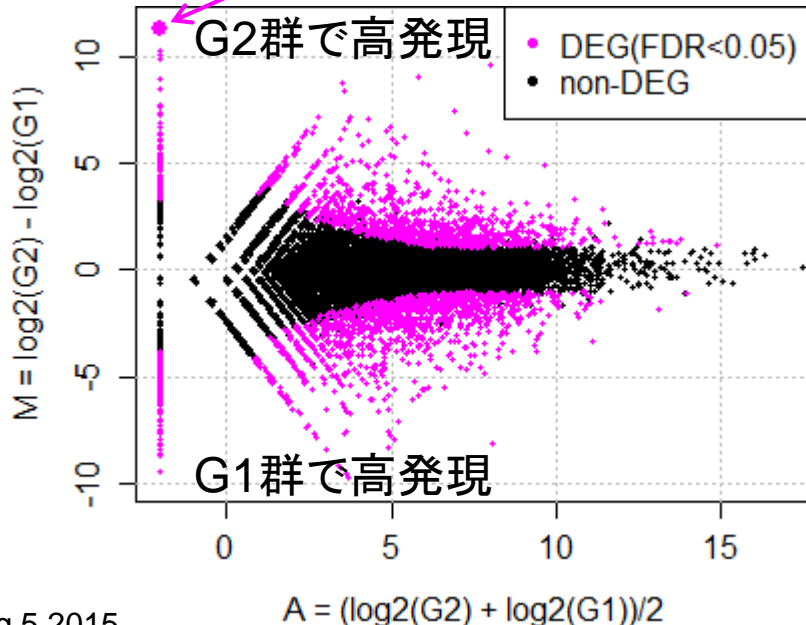
rownames(tcc\$count)	HSF1	HSM1	RMF1	RMM1	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000208570	0.0	0.0	1346.8	1477.0	ENSG00000208570	-2.04	11.29	4.41E-53	9.13E-49	1	1
ENSG00000220191	2.3	2.5	1394.7	1171.1	ENSG00000220191	5.79	9.06	4.54E-47	4.70E-43	2	1
ENSG00000106366	4421.9	4411.0	23.1	8.3	ENSG00000106366	8.04	-8.14	2.46E-45	1.70E-41	3	1
ENSG00000209449	0.0	0.0	644.5	713.1	ENSG00000209449	-2.04	10.23	3.29E-44	1.70E-40	4	1
ENSG00000218007	0.0	0.0	616.1	606.7	ENSG00000218007	-2.04	10.08	1.74E-43	7.21E-40	5	1
ENSG00000070985	0.0	0.0	528.2	650.8	ENSG00000070985	-2.04	10.03	4.68E-42	1.61E-38	6	1
ENSG00000209007	0.0	0.0	615.2	479.6	ENSG00000209007	-2.04	9.92	1.24E-40	3.67E-37	7	1
ENSG00000182327	367.5	363.9	0.9	0.0	ENSG00000182327	3.67	-9.69	1.52E-38	3.93E-35	8	1
ENSG00000156222	367.5	301.5	0.9	0.0	ENSG00000156222	3.61	-9.56	1.09E-36	2.51E-33	9	1
ENSG00000165272	404.8	429.0	2.6	0.9	ENSG00000165272	4.02	-7.59	5.45E-36	1.12E-32	10	1

# DEG検出結果

G1(HS)群    G2(RM)群

p-valueとその順位

rownames(tcc\$count)	HSF1	HSM1	RMF1	RMM1	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000208570	0.0	0.0	1346.8	1477.0	ENSG00000208570	-2.04	11.29	4.41E-53	9.13E-49	1	1
ENSG00000220191	2.3	2.5	1394.7	1171.1	ENSG00000220191	5.79	9.06	4.54E-47	4.70E-43	2	1
ENSG00000106366	4421.9	4411.0	23.1	8.3	ENSG00000106366	8.04	-8.14	2.46E-45	1.70E-41	3	1
ENSG00000209449	0.0	0.0	644.5	713.1	ENSG00000209449	-2.04	10.23	3.29E-44	1.70E-40	4	1
ENSG00000218007	0.0	0.0	616.1	606.7	ENSG00000218007	-2.04	10.08	1.74E-43	7.21E-40	5	1
ENSG00000070985	0.0	0.0	528.2	650.8	ENSG00000070985	-2.04	10.03	4.68E-42	1.61E-38	6	1
ENSG00000209007	0.0	0.0	615.2	479.6	ENSG00000209007	-2.04	9.92	1.24E-40	3.67E-37	7	1
ENSG00000182327	367.5	363.9	0.9	0.0	ENSG00000182327	3.67	-9.69	1.52E-38	3.93E-35	8	1
ENSG00000156222	367.5	301.5	0.9	0.0	ENSG00000156222	3.61	-9.56	1.09E-36	2.51E-33	9	1
ENSG00000165272	404.9	429.0	2.6	0.9	ENSG00000165272	4.02	-7.59	5.45E-36	1.12E-32	10	1



M-A plotのA値とM値

q-value

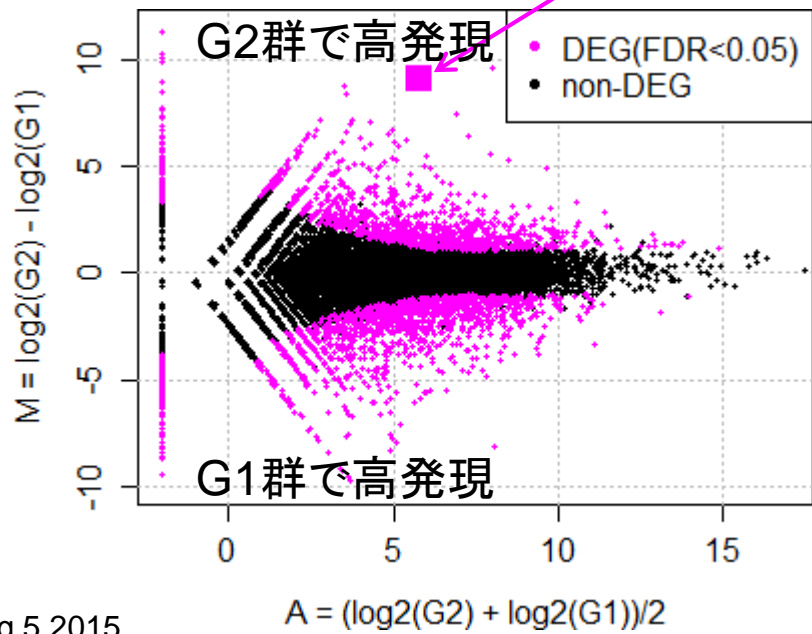
FDR閾値判定結果。q-value < 0.05  
を満たすDEGが1、non-DEGが0。

# DEG検出結果

G1(HS)群    G2(RM)群

p-valueとその順位

rownames(tcc\$count)	HSF1	HSM1	RMF1	RMM1	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000208570	0.0	0.0	1346.8	1477.0	ENSG00000208570	-2.04	11.29	4.41E-53	9.13E-49	1	1
ENSG00000220191	2.3	2.5	1394.7	1171.1	ENSG00000220191	5.79	9.06	4.54E-47	4.70E-43	2	1
ENSG00000106366	4421.9	4411.0	23.1	8.3	ENSG00000106366	8.04	-8.14	2.46E-45	1.70E-41	3	1
ENSG00000209449	0.0	0.0	644.5	713.1	ENSG00000209449	-2.04	10.23	3.29E-44	1.70E-40	4	1
ENSG00000218007	0.0	0.0	616.1	606.7	ENSG00000218007	-2.04	10.08	1.74E-43	7.21E-40	5	1
ENSG00000070985	0.0	0.0	528.2	650.8	ENSG00000070985	-2.04	10.03	4.68E-42	1.61E-38	6	1
ENSG00000209007	0.0	0.0	615.2	479.6	ENSG00000209007	-2.04	9.92	1.24E-40	3.67E-37	7	1
ENSG00000182327	367.5	363.9	0.9	0.0	ENSG00000182327	3.67	-9.69	1.52E-38	3.93E-35	8	1
ENSG00000156222	367.5	301.5	0.9	0.0	ENSG00000156222	3.61	-9.56	1.09E-36	2.51E-33	9	1
ENSG00000165272	404.9	429.0	2.6	0.9	ENSG00000165272	4.02	-7.59	5.45E-36	1.12E-32	10	1



M-A plotのA値とM値

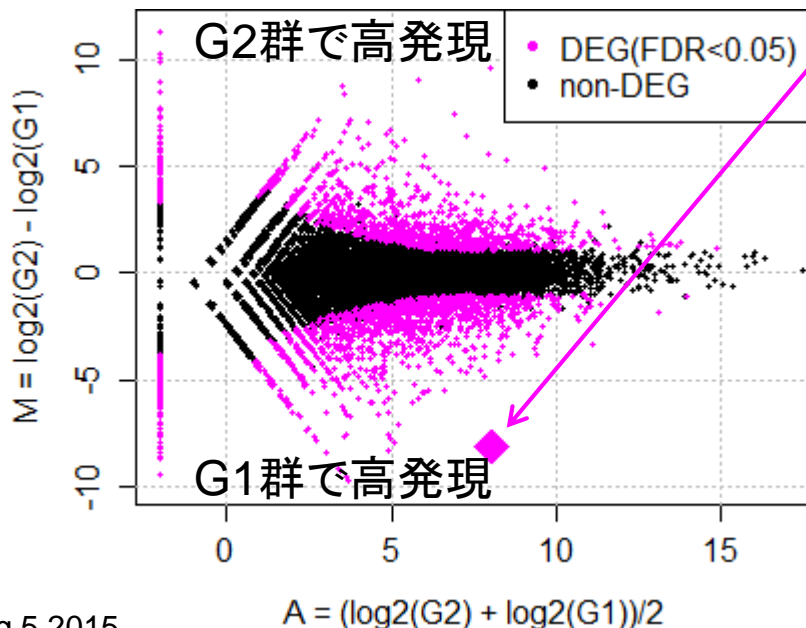
q-value

FDR閾値判定結果。q-value < 0.05を満たすDEGが1、non-DEGが0。



# DEG検出結果

rownames(tcc\$count)	G1(HS)群		G2(RM)群		gene_id	a.value	m.value	p-valueとその順位				estimatedDEG
	HSF1	HSM1	RMF1	RMM1				p.value	q.value	rank		
ENSG00000208570	0.0	0.0	1346.8	1477.0	ENSG00000208570	-2.04	11.29	4.41E-53	9.13E-49	1	1	
ENSG00000220191	2.3	2.5	1394.7	1171.1	ENSG00000220191	5.79	9.06	4.54E-47	4.70E-43	2	1	
ENSG00000106366	4421.9	4411.0	23.1	8.3	ENSG00000106366	8.04	-8.14	2.46E-45	1.70E-41	3	1	
ENSG00000209449	0.0	0.0	644.5	713.1	ENSG00000209449	-2.04	10.23	3.29E-44	1.70E-40	4	1	
ENSG00000218007	0.0	0.0	616.1	606.7	ENSG00000218007	-2.04	10.08	1.74E-43	7.21E-40	5	1	
ENSG00000070985	0.0	0.0	528.2	650.8	ENSG00000070985	-2.04	10.03	4.68E-42	1.61E-38	6	1	
ENSG00000209007	0.0	0.0	615.2	479.6	ENSG00000209007	-2.04	9.92	1.24E-40	3.67E-37	7	1	
ENSG00000182327	367.5	363.9	0.9	0.0	ENSG00000182327	3.67	-9.69	1.52E-38	3.93E-35	8	1	
ENSG00000156222	367.5	301.5	0.9	0.0	ENSG00000156222	3.61	-9.56	1.09E-36	2.51E-33	9	1	
ENSG00000165272	404.9	429.0	2.6	0.9	ENSG00000165272	4.02	-7.59	5.45E-36	1.12E-32	10	1	



M-A plotのA値とM値

q-value

FDR閾値判定結果。q-value < 0.05  
を満たすDEGが1、non-DEGが0。

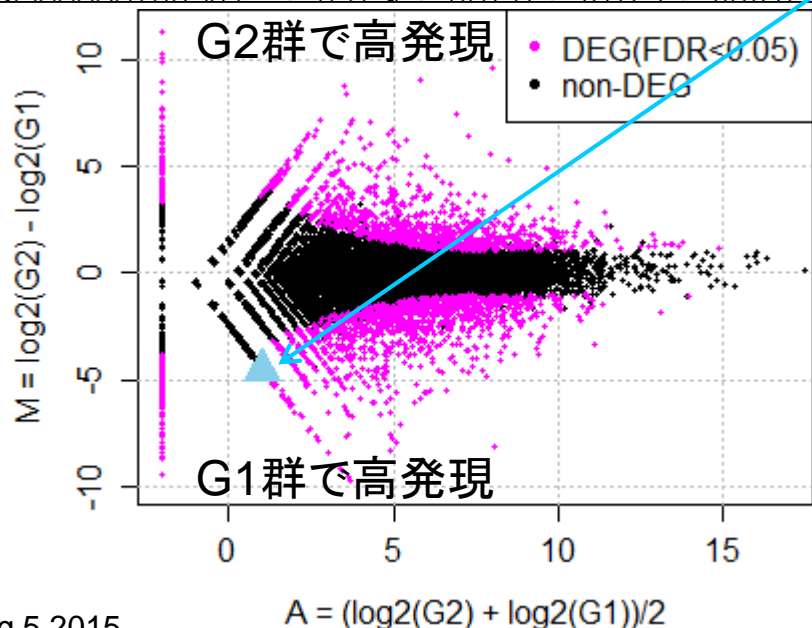




指定したFDR閾値(0.05)をギリギリ満たす2,488位の遺伝子

# DEG検出結果

rownames(tcc\$count)	G1(HS)群		G2(RM)群		gene_id	a.value	m.value	p-valueとその順位				estimatedDEG
	HSF1	HSM1	RMF1	RMM1				p.value	q.value	rank		
ENSG00000180672	9.0	8.9	0.9	1.7	ENSG00000180672	1.76	-2.82	0.00596	0.04967	2484	1	
ENSG00000159899	161.7	89.1	47.9	60.7	ENSG00000159899	6.37	-1.21	0.00597	0.0497	2485	1	
ENSG00000110442	108.5	103.1	214.0	219.4	ENSG00000110442	7.24	1.03	0.00599	0.04987	2486	1	
ENSG00000105327	5.7	24.2	1.8	2.5	ENSG00000105327	2.50	-2.80	0.006	0.04989	2487	1	
ENSG00000139445	17.0	2.5	0.0	0.8	ENSG00000139445	1.01	-4.55	0.006	0.04989	2488	1	
ENSG00000105321	61.1	128.5	14.2	47.4	ENSG00000105321	5.76	-1.62	0.00602	0.05008	2489	0	
ENSG00000118017	1.1	2.5	13.3	10.0	ENSG00000118017	2.21	2.66	0.00603	0.05009	2490	0	
ENSG00000110917	768.8	591.6	1440.9	1334.8	ENSG00000110917	9.92	1.03	0.00603	0.05011	2491	0	
ENSG00000119630	19.2	12.7	34.6	55.7	ENSG00000119630	4.75	1.50	0.00604	0.05011	2492	0	
ENSG00000144567	421.9	402.0	810.5	888.6	ENSG00000144567	9.21	1.01	0.00605	0.05019	2493	0	



M-A plotのA値とM値

q-value

FDR閾値判定結果。q-value < 0.05を満たすDEGが1、non-DEGが0。

# 様々なM-A plot

- ・解析 | 発現変動 | 1について (last modified 2014/07/10)
- ・解析 | 発現変動 | 2群間 | 対応なし | 1について (last modified 2015/06/02)
- ・解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun\_2013) (last modified 2015/07/07)
- ・解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun\_2013) **①**
- ・解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq(Li\_2013) (last modified 2015/07/07)
- ・解析 | 発現変動 | 2群間 | 対応なし | 複製あり | edgeR(Robinson\_2010) (last modified 2015/07/07)
- ・解析 | 発現変動 | 2群間 | 対応なし | 複製あり | WAD(Kadota\_2008) (last modified 2015/07/07)
- ・解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun\_2013) NEW
- ・解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun\_2013) NEW
- ・解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun\_2013) NEW

## 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun\_2013) NEW

[Blekhman et al., Genome Res., 2010](#)の公共カウントデータ解析に特化させて、[TCC](#)を用いた様々な例題を示します。入力は全て [サンプルデータ42](#)の 20,689 genes×18 samplesのリアルカウントデータ ([sample blekhman 18.txt](#))です。ヒトHomo sapiens; HS)のメス3サンプル(HSF1-3)とオス3サンプル(HSM1-3), チンパンジー(Pan troglodytes; PT)のメス3サンプル(PTF1-3)とオス3サンプル(PTM1-3), アカゲザル(Rhesus macaque; RM)のメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)の並びになっています。つまり、以下のような感じです。FはFemale(メス)、MはMale(オス)を表します。

ヒト(1-6列目): HSF1, HSF2, HSF3, HSM1, HSM2, and HSM3

チンパンジー(7-12列目): PTF1, PTF2, PTF3, PTM1, PTM2, and PTM3

アカゲザル(13-18列目): RMF1, RMF2, RMF3, RMM1, RMM2, and RMM3

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

### 1. ヒト2サンプル(G1群:HSF1とHSM1) vs. アカゲザル2サンプル(G2群:RMF1とRMM1)の場合:

1, 4, 13, 16 列目のデータのみ抽出しています。

```
in_f <- "sample_blekhman_18.txt"
out_f1 <- "hoge1.txt"
out_f2 <- "hoge1.png"
param_subset <- c(1, 4, 13, 16)
param_G1 <- 2
param_G2 <- 2
param_FDR <- 0.05
param_fig <- c(430, 350)
param_mar <- c(4, 4, 0, 0)
```

```
#入力ファイル名を指定してin_f1に格納
#出力ファイル名を指定してout_f1に格納
#出力ファイル名を指定してout_f2に格納
#取り扱いたいサブセット情報を指定
#G1群のサンプル数を指定
#G2群のサンプル数を指定
#DEG検出時のfalse discovery rate (FDR)閾値を指定
#ファイル出力時の横幅と縦幅を指定(単位はピクセル)
#下、左、上、右の順で余白を指定(単位は行)
```

```
#必要なパッケージをロード
```

```
#パッケージの読み込み
```

# Contents

## ■ サンプル間クラスタリング

- 実行手順のおさらい
- 計算の一部を解説、結果の解釈

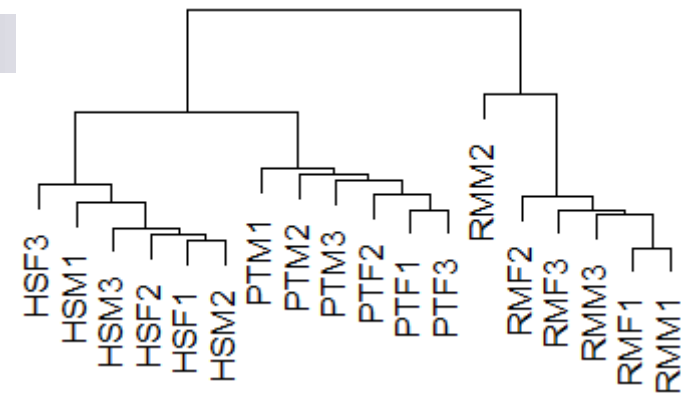
## ■ 発現変動解析(2群間比較) : 発現変動遺伝子(DEG)が多数存在する場合

## ■ モデル、分布、統計的手法、2群間比較でDEGがそれほど多くない場合

## ■ 2群間比較でDEGがほとんどない同一群の場合

## ■ 倍率変化(2倍以上、1/2以下の発現変動)を用いた場合

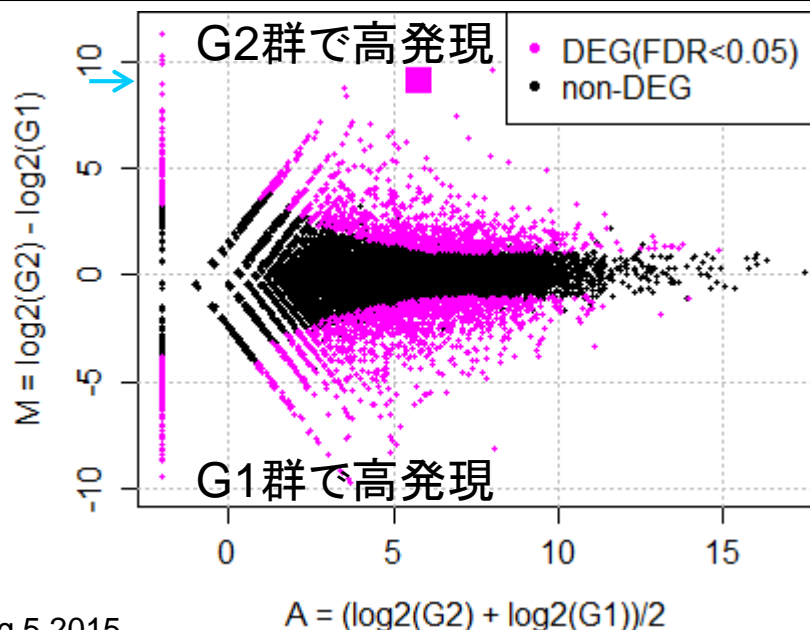
## ■ 発現変動解析: 3群間比較など



# Tips: logの世界

G1(HS)群    G2(RM)群

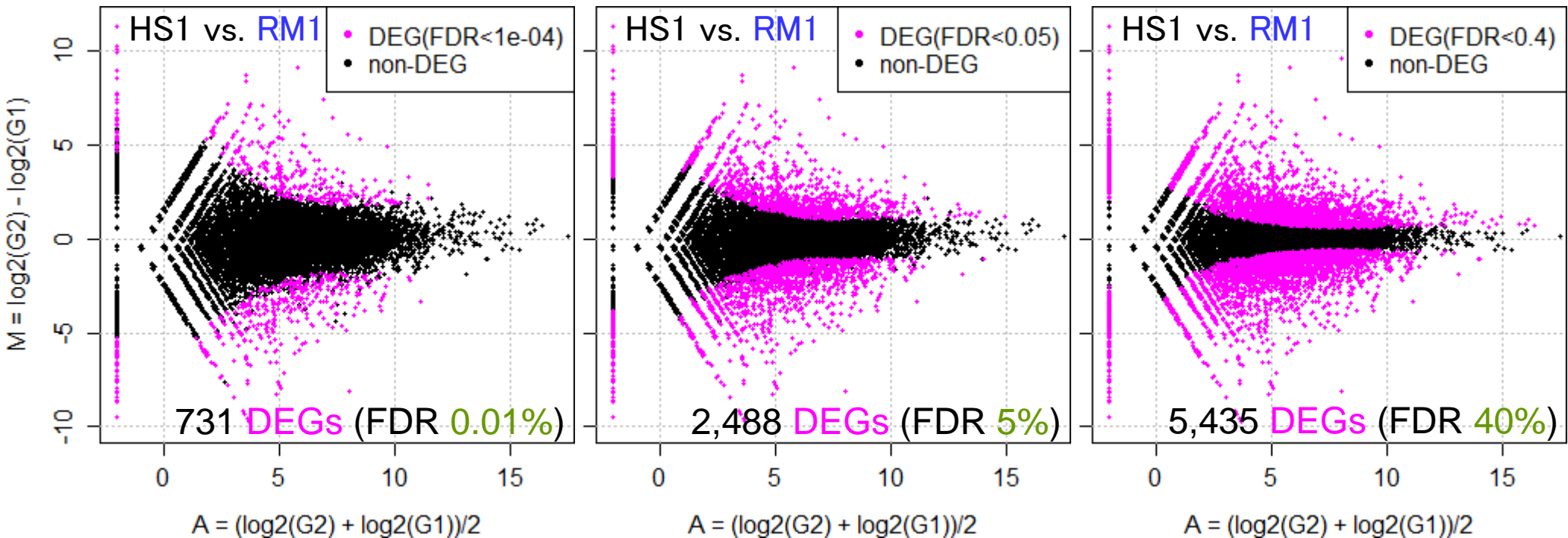
rownames(tcc\$count)	HSF1	HSM1	RMF1	RMM1	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000208570	0.0	0.0	1346.8	1477.0	ENSG00000208570	-2.04	11.29	4.41E-53	9.13E-49	1	1
ENSG00000220191	2.3	2.5	1394.7	1171.1	ENSG00000220191	5.79	9.06	4.54E-47	4.70E-43	2	1
ENSG00000106366	4421.9	4411.0	23.1	8.3	ENSG00000106366	8.04	-8.14	2.46E-45	1.70E-41	3	1
ENSG00000209449	0.0	0.0	644.5	713.1	ENSG00000209449	-2.04	10.23	3.29E-44	1.70E-40	4	1
ENSG00000218007	0.0	0.0	616.1	606.7	ENSG00000218007	-2.04	10.08	1.74E-43	7.21E-40	5	1
ENSG00000070985	0.0	0.0	528.2	650.8	ENSG00000070985	-2.04	10.03	4.68E-42	1.61E-38	6	1
ENSG00000209007	0.0	0.0	615.2	479.6	ENSG00000209007	-2.04					
ENSG00000182327	367.5	363.9	0.9	0.0	ENSG00000182327	3.67					
ENSG00000156222	367.5	301.5	0.9	0.0	ENSG00000156222	3.61					
ENSG00000165272	404.9	429.0	2.6	0.9	ENSG00000165272	4.02					



```
R Console
> (2.3 + 2.5) / 2
[1] 2.4
> (1394.7 + 1171.1) / 2
[1] 1282.9
> (log2(1282.9) + log2(2.4)) / 2
[1] 5.794114
> log2(1282.9) - log2(2.4)
[1] 9.062159
> 1282.9 / 2.4
[1] 534.5417
> log2(1282.9 / 2.4)
[1] 9.062159
> 2^9.062159
[1] 534.5418
> |
```

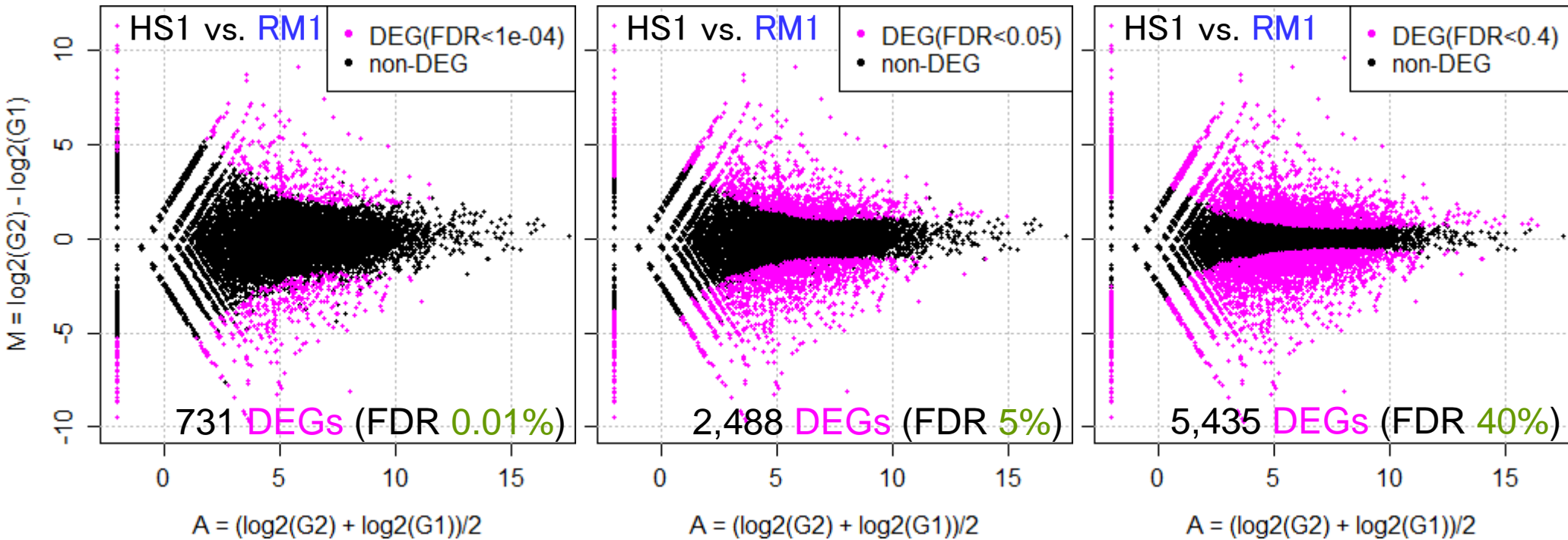
# 分布やモデル

(当たり前だが)FDR閾値を緩めると得られるDEG数は増える傾向にあることがわかる。例題6のコピペで作成。



厳しい ← FDR閾値 → 緩い  
 少ない ← DEG数 → 多い

# 分布やモデル



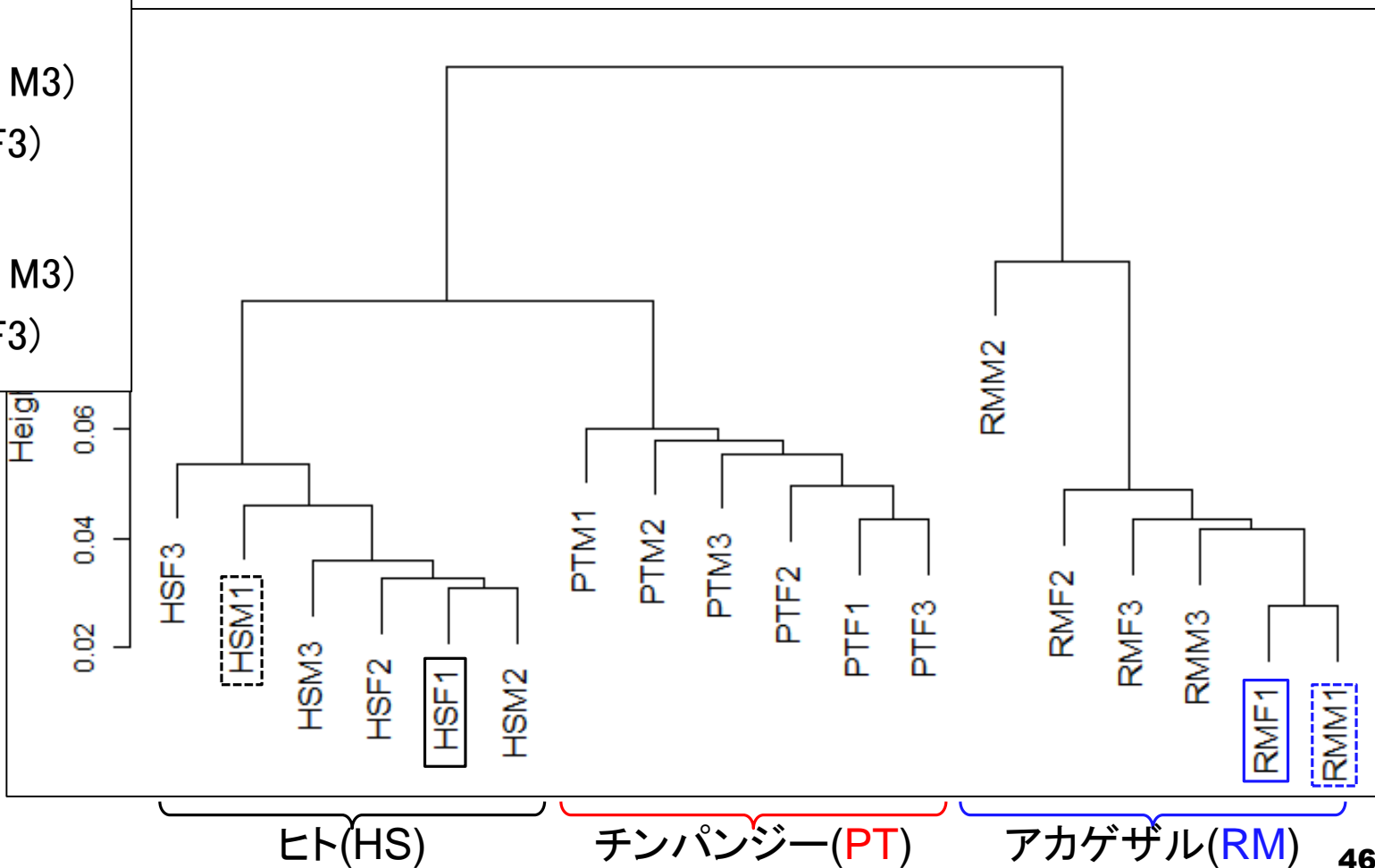
厳しい ← FDR閾値 → 緩い  
 少ない ← DEG数 → 多い



「HS vs. RM」の発現変動解析結果として、20,689 genes中3,300個程度が本物のDEGと判断した。

# おさらい

- ヒト(HS)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- チンパンジー(PT)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)
- アカゲザル(RM)
  - オス3匹(M1, M2, M3)
  - メス3匹(F1, F2, F3)





# HS vs. PT

「HS vs. PT」のDEG同定を行う。ヒト(HS)とチンパンジー(PT)で明瞭にサブクラスターに分かれていることから、DEGは存在すると予想される。しかし、「HS vs. RM」(3,300個程度が本物のDEGと判断した)のときほどDEGは多くないだろうと予想できる。

## ■ ヒト(HS)

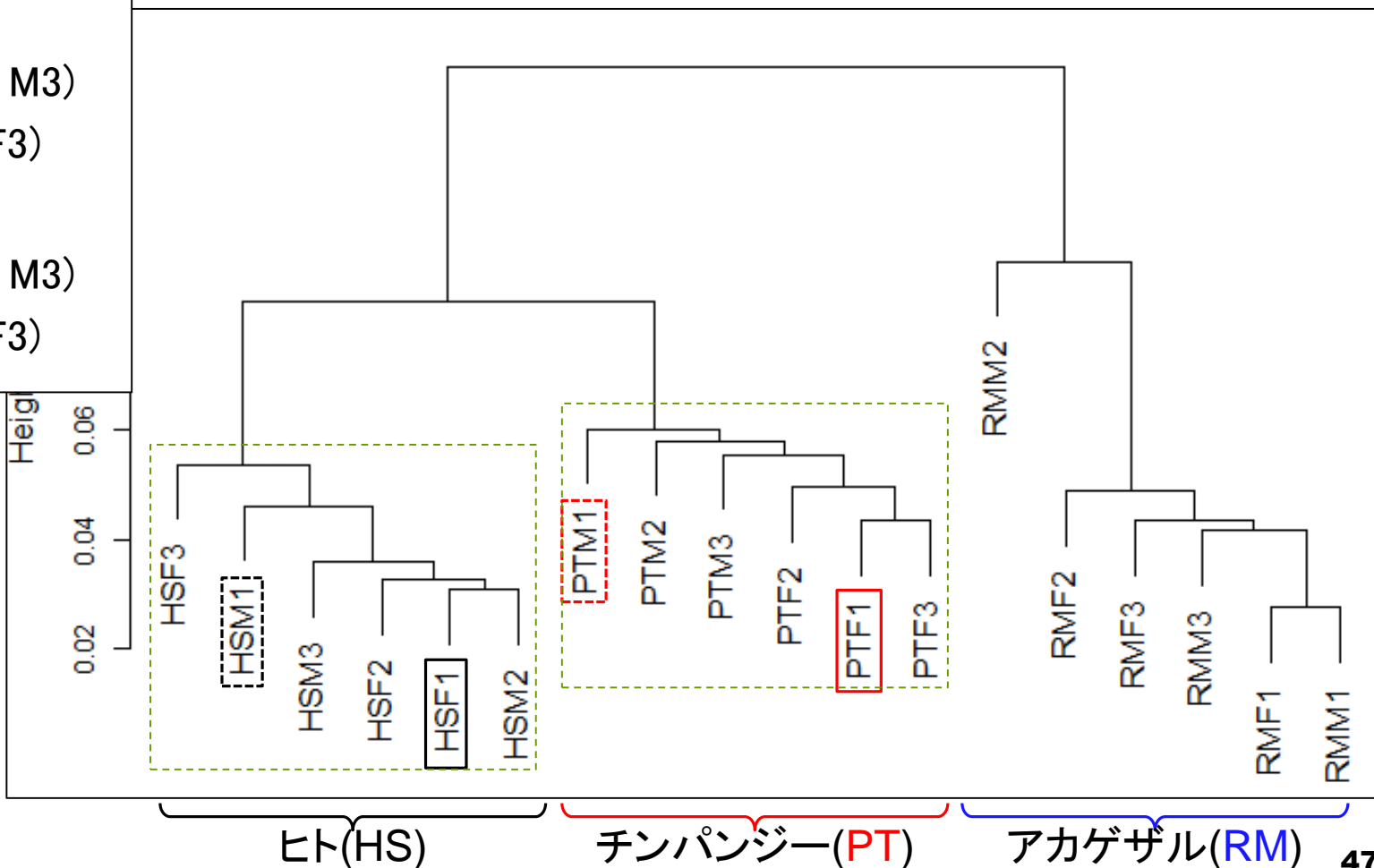
- オス3匹(M1, M2, M3)
- メス3匹(F1, F2, F3)

## ■ チンパンジー(PT)

- オス3匹(M1, M2, M3)
- メス3匹(F1, F2, F3)

## ■ アカゲザル(RM)

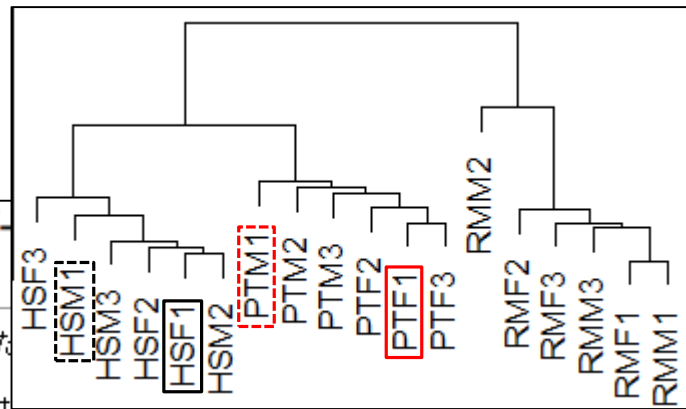
- オス3匹(M1, M2, M3)
- メス3匹(F1, F2, F3)



# TCC実行

- 解析 | 発現変動 | について (last modified 2014/07/10)
- 解析 | 発現変動 | 2群間 | 対応なし | について (last modified 2015/06/02)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun 2013) (last modified 2015/07/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun 2013)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq(Li 2013) (last modified 2015/07/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | LedoP(Rabinovitch 2010) (last modified 2015/07/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun 2013) **NEW**

## 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ (Sun\_2013) NEW



### ② 7. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の20,689 genes×18 samplesのカウントデータです。ヒトのメス3サンプル(HSF1-3)とオス3サンプル(HSM1-3), チンパンジーのメス3サンプル(PTF1-3)とオス3サンプル(PTM1-3), アカゲザルのメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)の並びになっています。ここでは、1, 4, 7, 10 列目のデータのみ抽出して、ヒト2サンプル(G1群:HSF1とHSM1) vs. チンパンジー2サンプル(G2群:PTF1とPTM1)の2群間比較を行います。

```
1, 4, 13, 16 列目のデータ
in_f <- "sample_b
out_f1 <- "hoge1.
out_f2 <- "hoge1.
param_subset <- c
param_G1 <- 2
param_G2 <- 2
param_FDR <- 0.05
param_fig <- c(43
```

```
in_f <- "sample_blekhman_18.txt"
out_f1 <- "hoge7.txt"
out_f2 <- "hoge7.png"
param_subset <- c(1, 4, 7, 10)
param_G1 <- 2
param_G2 <- 2
param_FDR <- 0.05
param_fig <- c(400, 310)
param_mar <- c(4, 4, 0, 0)
```

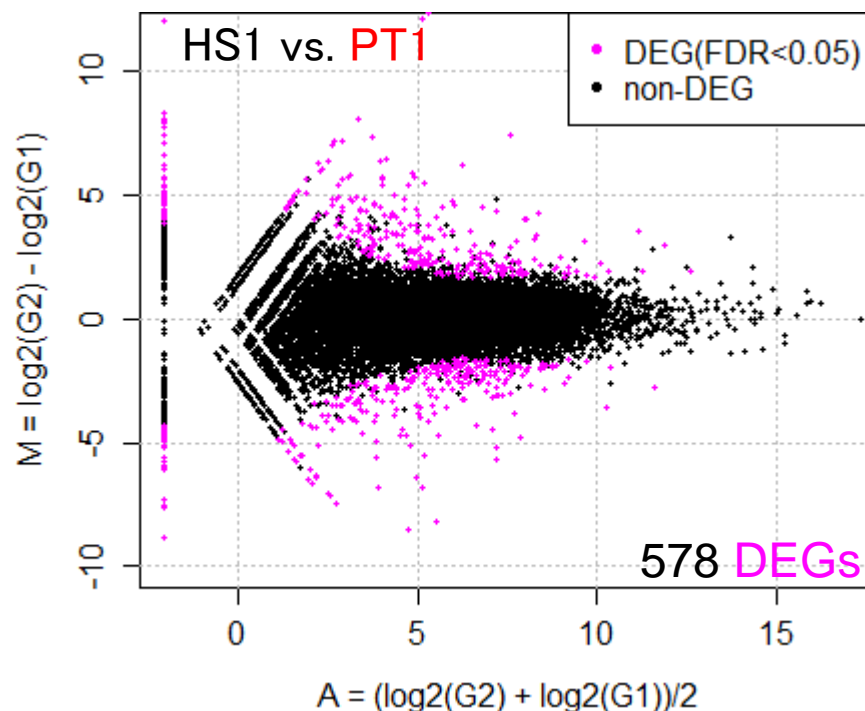
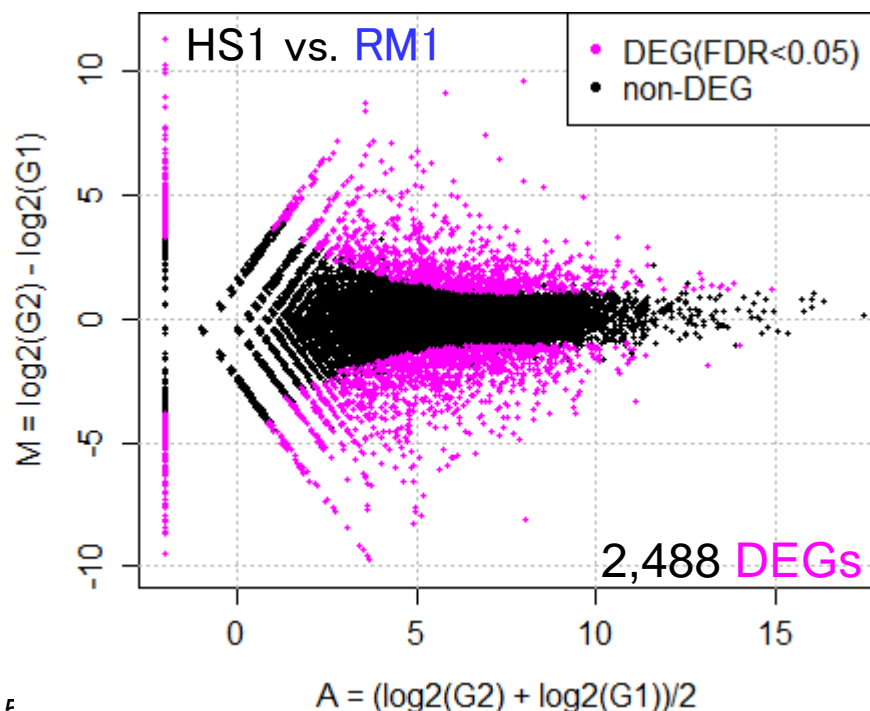
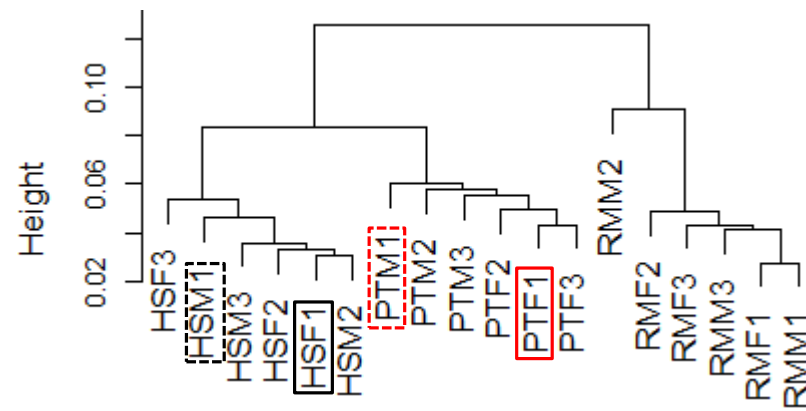
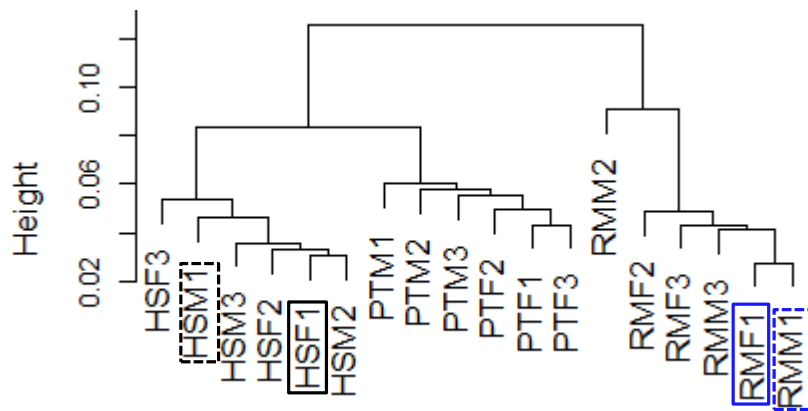
#入力ファイル名を指定してin\_fに格納  
 #出力ファイル名を指定してout\_f1に格納  
 #出力ファイル名を指定してout\_f2に格納  
 #取り扱いたいサブセット情報を指定  
 #G1群のサンプル数を指定  
 #G2群のサンプル数を指定  
 #DEG検出時のfalse discovery rate (FDR)  
 #ファイル出力時の横幅と縦幅を指定(単位はポイント)  
 #下、左、上、右の順で余白を指定(単位はポイント)

```
#必要なパッケージをロード
library(TCC)
```

#パッケージの読み込み

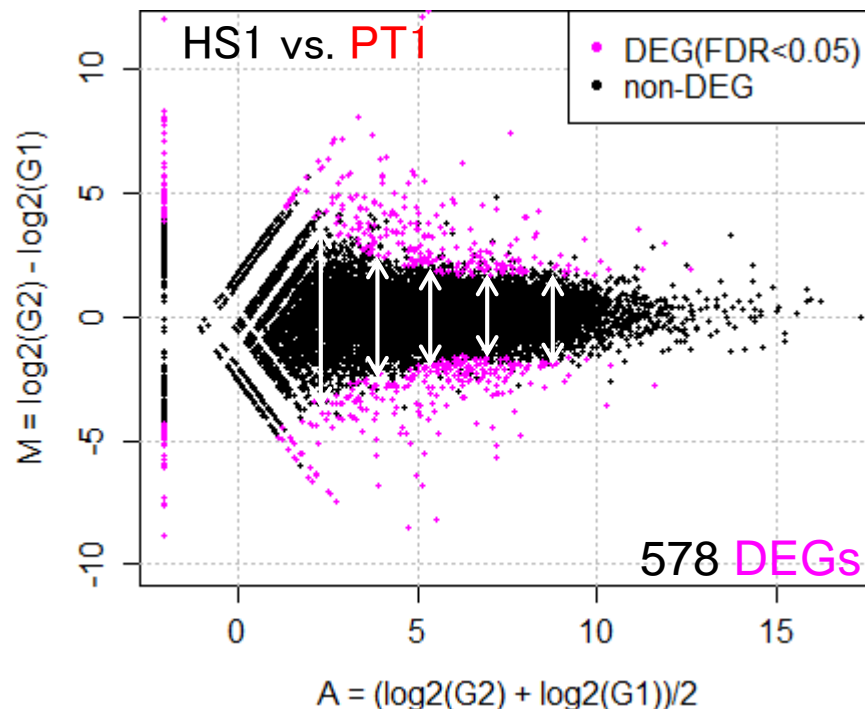
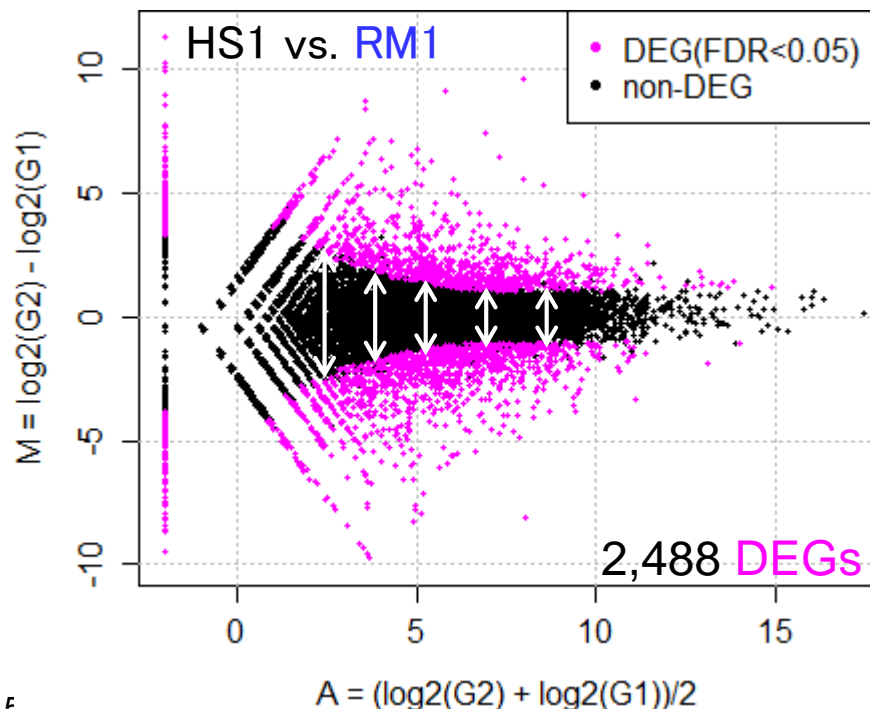
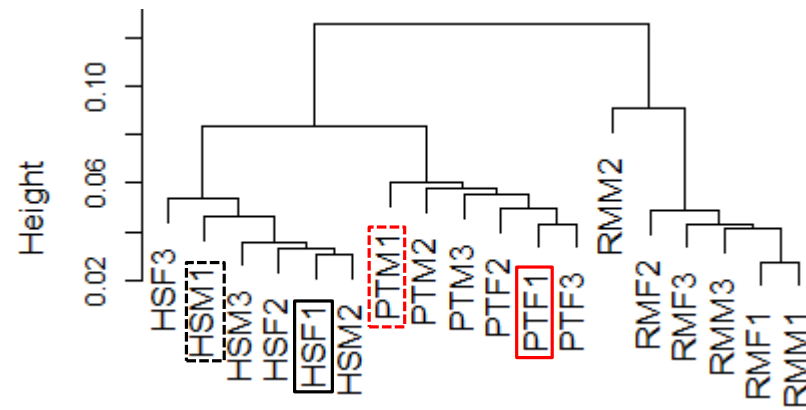
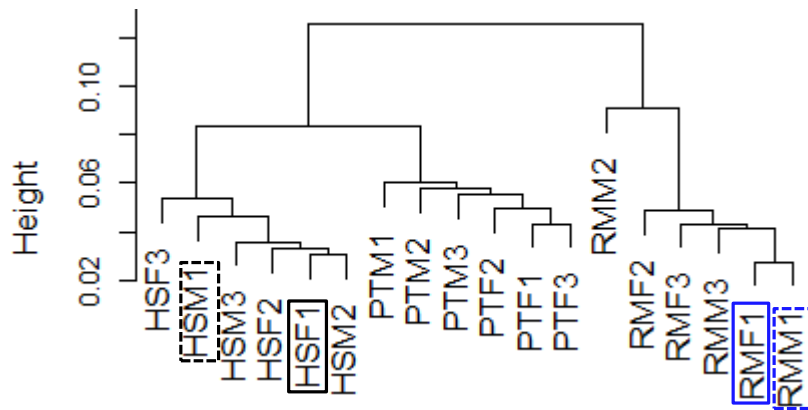
「HS vs. PT」は、「HS vs. RM」に比べて全体的に似ているのでDEG数は少なくなる。

# 結果の比較



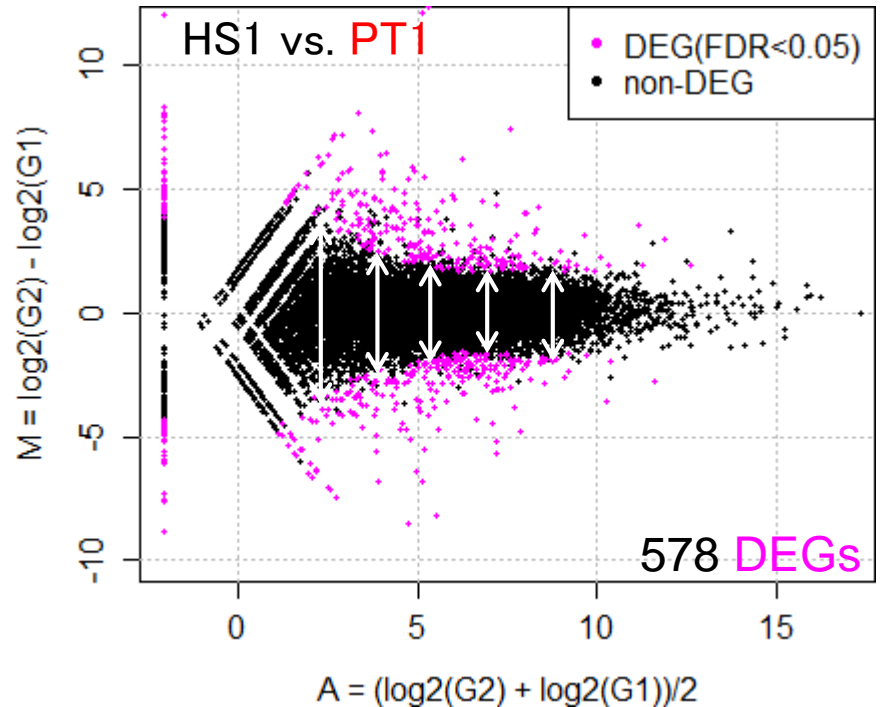
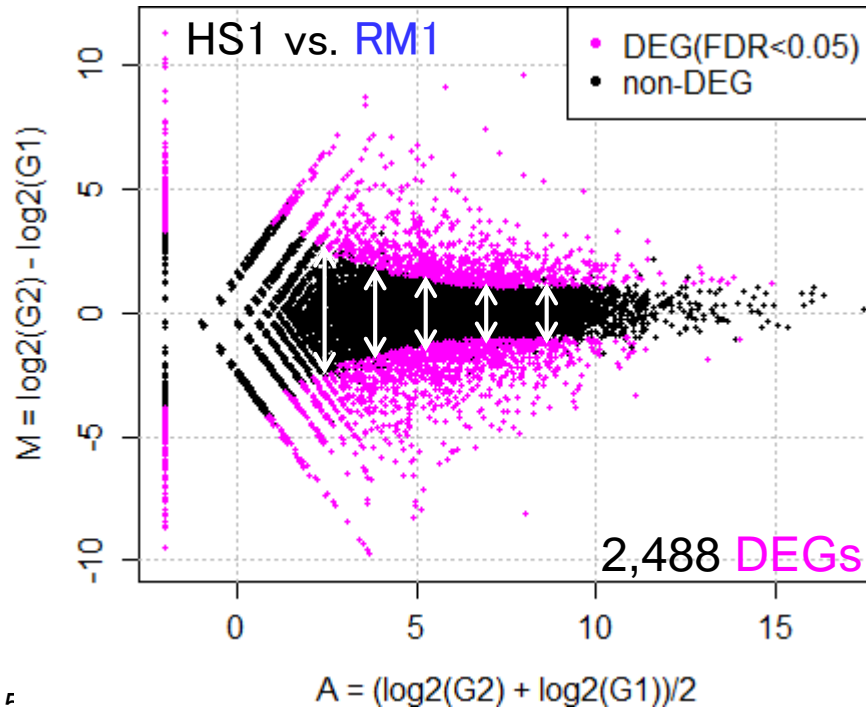
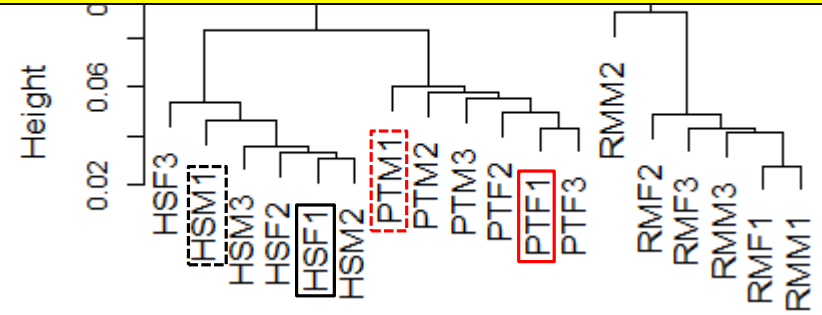
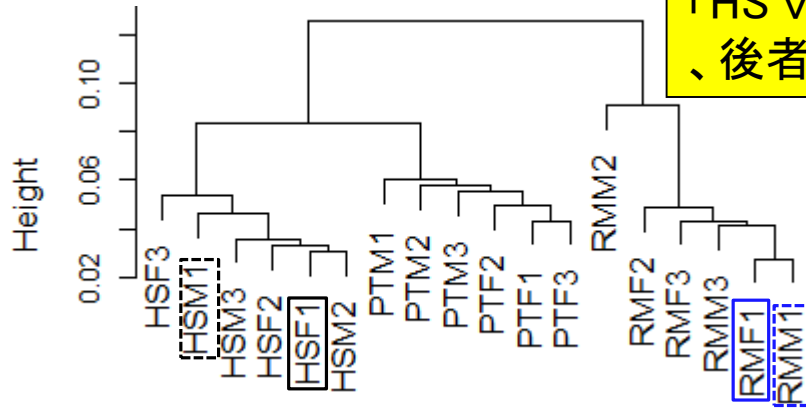
# 素朴な疑問

何故、白矢印で示すように「HS vs. PT」の non-DEG の分布(黒の点の分布)は、「HS vs. RM」に比べて広がっているのか?



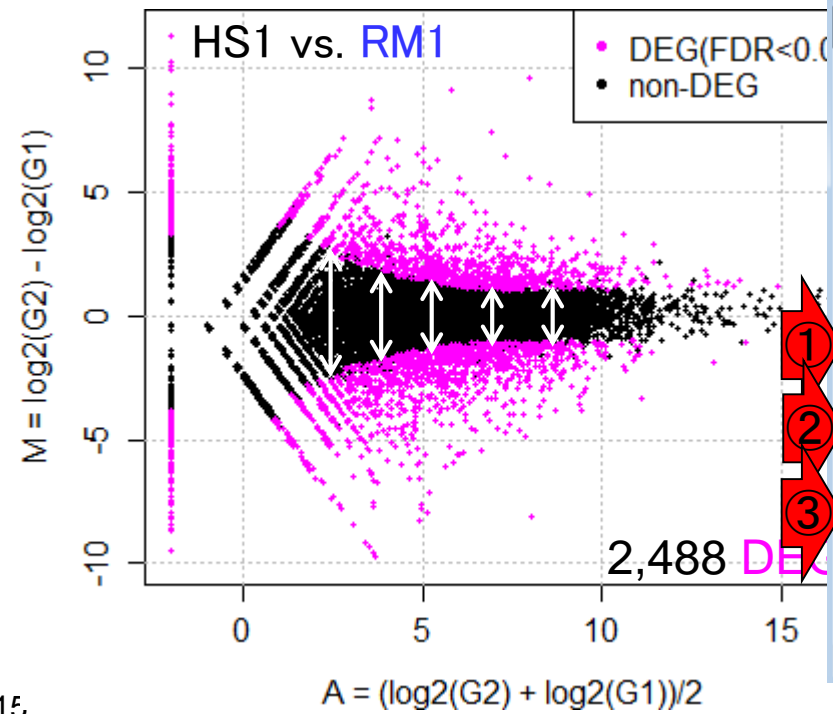
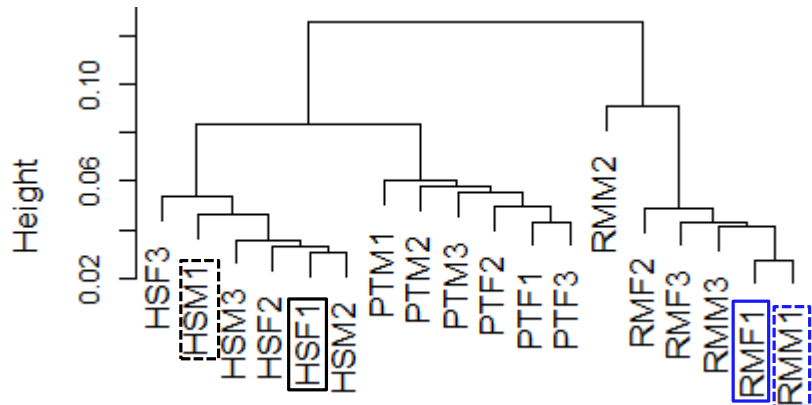
# 統計的手法とは

疑問に対する解答は、統計的手法の手順を再考すればよい。同一群内のばらつきの分布 (non-DEG分布) 以外のものが **DEG** と判定されるのが統計的手法の結果。つまり、「HS vs. **RM**」と「HS vs. **PT**」では、non-DEG分布が異なり、後者のほうが同一群内のばらつきが大きいということ。



# サンプル間類似度

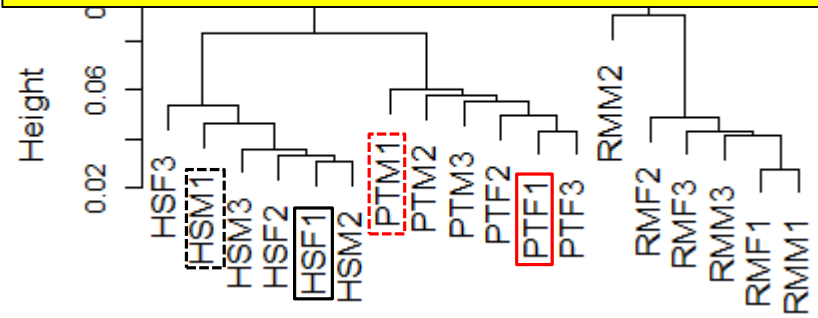
同一群内のばらつきは、サンプル間の類似度で大まかに把握可能。  
 ①HS群内(HSF1 vs. HSM1)のSpearman相関係数は0.950。  
 ②RM群内(RMF1 vs. RMM1)は0.972。  
 ③「HS vs. RM」の群間比較結果は、例えばHSM1 vs. RMM1の相関係数(0.880)が0.950と0.972よりも低いことからDEGの存在を予測可能。



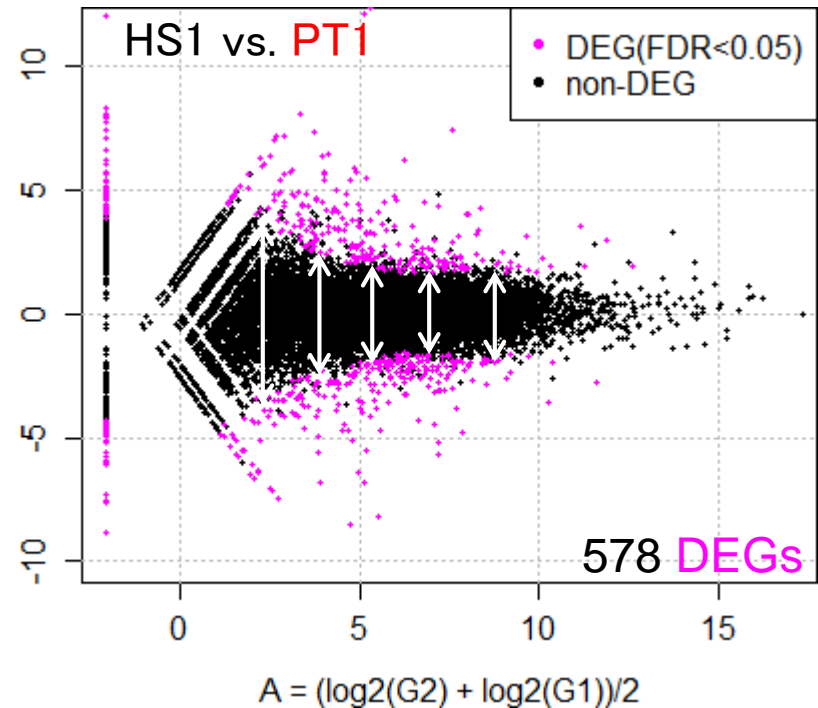
```
R Console
> in_f <- "sample_blekhman_18.txt" #入$
> data <- read.table(in_f, header=TRUE, row.n$
> dim(data)
[1] 20689 18
> data <- unique(data)
> dim(data)
[1] 16561 18
> cor(data$HSM1, data$HSF1, method="spearman")
[1] 0.9502333
> cor(data$RMM1, data$RMF1, method="spearman")
[1] 0.9724166
> cor(data$HSM1, data$RMM1, method="spearman")
[1] 0.8799668
> |
```

# サンプル間類似度

①HS群内(HSF1 vs. HSM1)のSpearman相関係数は0.95。②PT群内(PTF1 vs. PTM1)は0.950。③「HS vs. PT」の群間比較結果は、例えばHSM1 vs. PTM1の相関係数(0.902)が0.950と0.949よりも低いことからDEGの存在を予測可能。



```
R Console
> in_f <- "sample_blekhman_18.txt" #入$
> data <- read.table(in_f, header=TRUE, row.n$
> dim(data)
[1] 20689 18
> data <- unique(data)
> dim(data)
[1] 16561 18
> cor(data$HSM1, data$HSF1, method="spearman")
[1] 0.9502333 ①
> cor(data$PTM1, data$PTF1, method="spearman")
[1] 0.9489023 ②
> cor(data$HSM1, data$PTM1, method="spearman")
[1] 0.9019057 ③
>
```

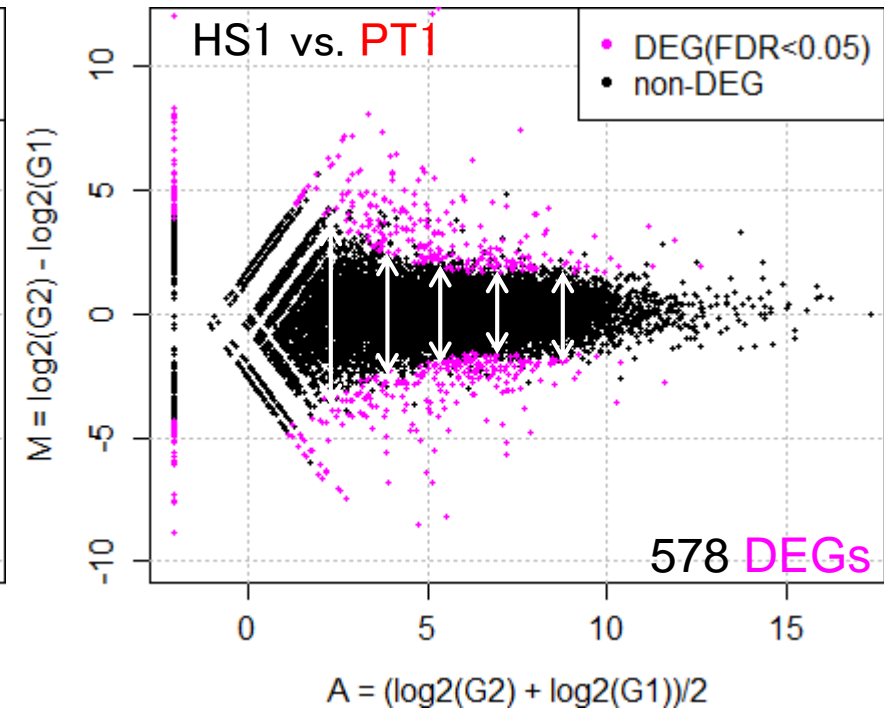
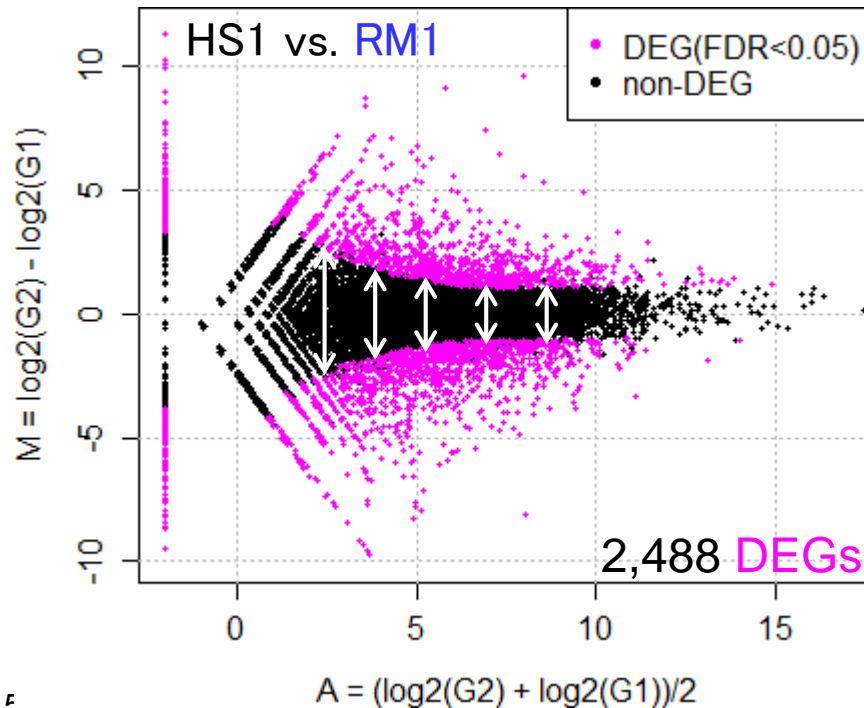
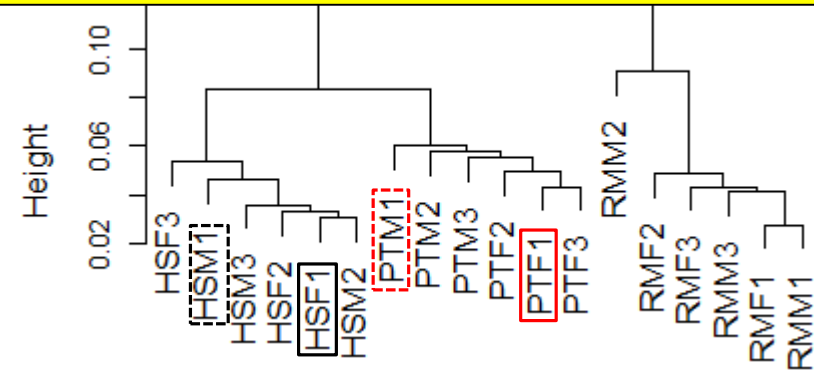
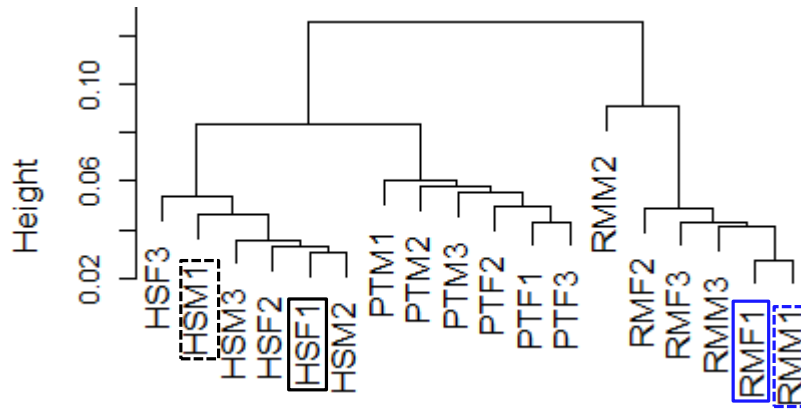




参考

RM群内(RMF1 vs. RMM1)のSpearman相関係数は0.972。一方、PT群内(PTF1 vs. PTM1)は0.949。大まかにいって、この差がnon-DEG分布の違いに寄与しているという理解でよい。

# DEG検出結果の比較



# Contents

## ■ サンプル間クラスタリング

- 実行手順のおさらい
- 計算の一部を解説、結果の解釈

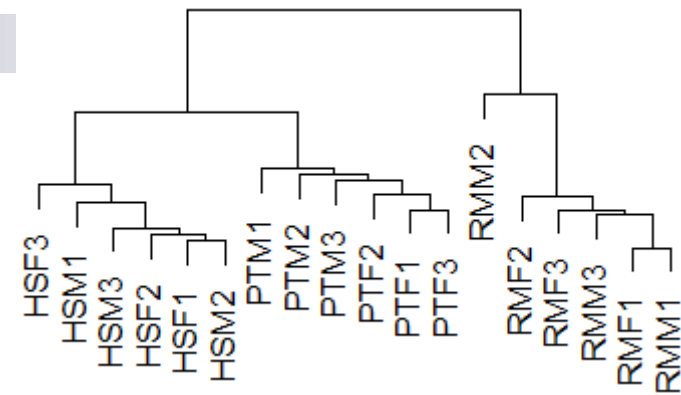
## ■ 発現変動解析(2群間比較) : 発現変動遺伝子(DEG)が多数存在する場合

## ■ モデル、分布、統計的手法、2群間比較でDEGがそれほど多くない場合

## ■ 2群間比較でDEGがほとんどない同一群の場合

## ■ 倍率変化(2倍以上、1/2以下の発現変動)を用いた場合

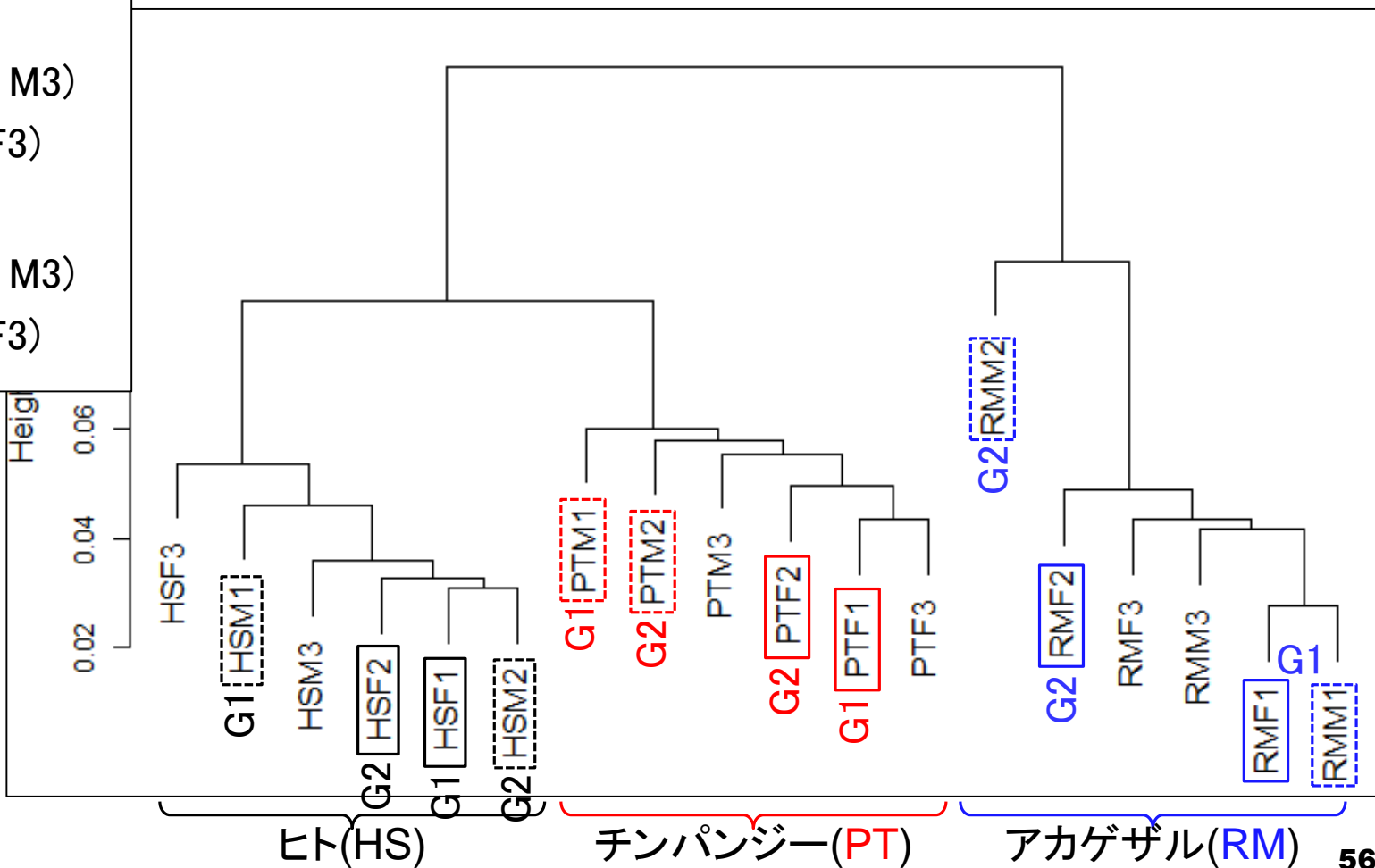
## ■ 発現変動解析: 3群間比較など



同一群内のばらつきの分布 (non-DEG分布) を調べるべく、「G1群(M1とF1) vs. G2群(M2とF2)」の2群間比較を行ってみる。予想はDEGはあったとしてもごくわずか。

# 2群間比較

- ヒト(HS)
  - オス3匹 (M1, M2, M3)
  - メス3匹 (F1, F2, F3)
- チンパンジー(PT)
  - オス3匹 (M1, M2, M3)
  - メス3匹 (F1, F2, F3)
- アカゲザル(RM)
  - オス3匹 (M1, M2, M3)
  - メス3匹 (F1, F2, F3)



「G1群(HSM1とHSF1) vs. G2群(HSM2とHSF2)」の2群間比較結果。7 DEGs。

# TCC実行

- 解析 | 発現変動 | について (last modified 2014/07/10)
- 解析 | 発現変動 | 2群間 | 対応なし | について (last modified 2015/06/02)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun 2013) (last modified 2015/07/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun 2013)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq(Li 2013) (last modified 2015/07/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Ledesma(Rabinovitch 2010) (last modified 2015/07/07)



## 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun\_2013) NEW

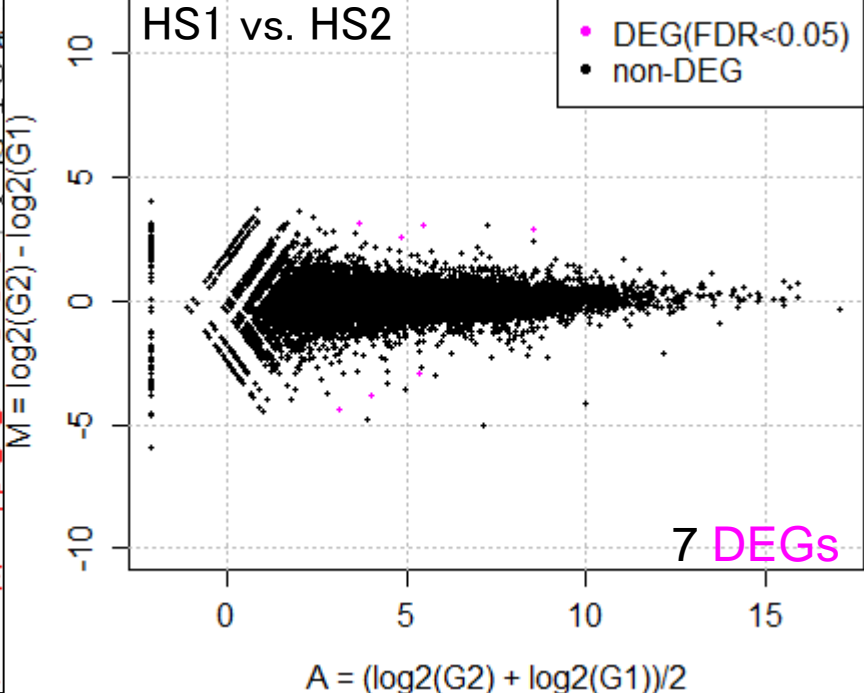
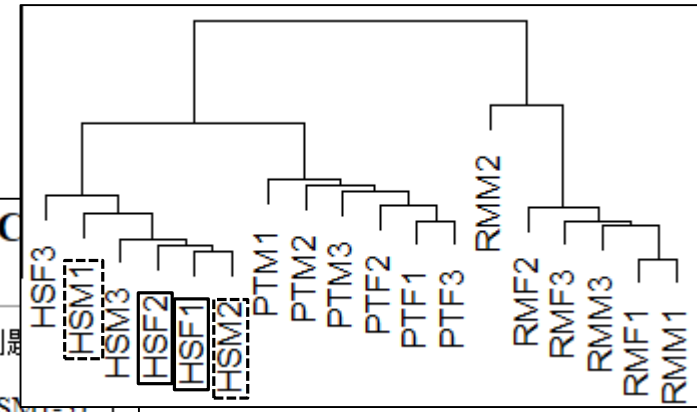
Blekhman et al., Genome Res., 2010の公共カウントデータ解析に特化させて、TCCを用いた様々な例題があります。入力は全てサンプルデータ42の20,689 genes×18 samplesのリアルカウントデータ (sample blekhman 18.txt)です。ヒトHomo sapiens; HS)のメス3サンプル(HSF1-3)とオス3サンプル(HSM1-3)、チンパンジー(Pan troglodytes; PT)のメス3サンプル(PTF1-3)とオス3サンプル(PTM1-3)、アカゲザル(Rhesus macaque)のメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)のリアルカウントデータが提供されています。ここでは、1, 4, 2, 5 列目のデータのみ抽出して、ヒト22サンプル(G2群:HSF2とHSM2)の2群間比較を行います。



Blekhman et al., Genome Res., 2010の20,689 genes×18 sampleのリアルカウントデータ (sample blekhman 18.txt)の22サンプル(G2群:HSF2とHSM2)の2群間比較を行います。

```
1, 4, 13, 14
in_f <- "sample_blekhman_18.txt"
out_f1 <- "hoge8.txt"
out_f2 <- "hoge8.png"
param_subset <- c(1, 4, 2, 5)
param_G1 <- 2
param_G2 <- 2
param_FDR <- 0.05
param_fig <- c(430, 350)
param_mar <- c(4, 4, 0, 0)
```

```
in_f <- "sample_blekhman_18.txt" #入力ファイル
out_f1 <- "hoge8.txt" #出力ファイル
out_f2 <- "hoge8.png" #出力ファイル
param_subset <- c(1, 4, 2, 5) #取り扱うサンプルID
param_G1 <- 2 #G1群のサンプル数
param_G2 <- 2 #G2群のサンプル数
param_FDR <- 0.05 #DEG検出率
param_fig <- c(430, 350) #ファイギュアサイズ
param_mar <- c(4, 4, 0, 0) #下、左の余白
library(TCC) #パッケージをロード
```

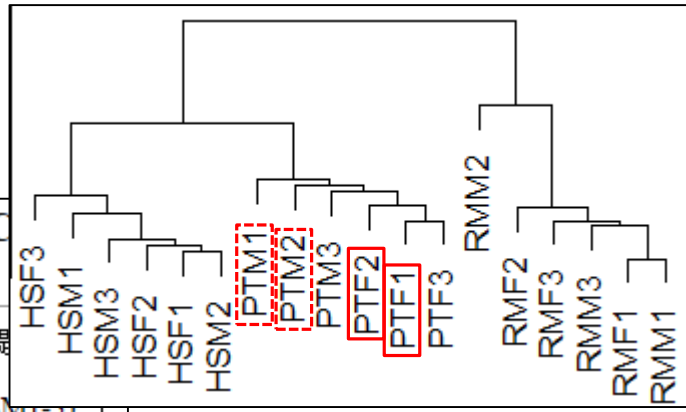


# TCC実行

- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun 2013) (last modified 2015/07/10)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun 2013) (last modified 2015/06/02)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq(Li 2013) (last modified 2015/06/02)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Lederer(Bokros 2010) (last modified 2015/06/02)

## 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun\_2013) NEW

Blekhman et al., Genome Res., 2010の公共カウントデータ解析に特化させて、TCCを用いた様々な例題があります。入力は全てサンプルデータ42の20,689 genes×18 samplesのリアルカウントデータ (sample blekhman 18.txt)です。ヒトHomo sapiens; HS)のメス3サンプル(HSF1-3)とオス3サンプル(HSM1-3), チンパンジー(Pan troglodytes; PT)のメス3サンプル(PTF1-3)とオス3サンプル(PTM1-3), アカゲザル(Rhesus macaque; RM)のメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)のリアルカウントデータが提供されています。ここでは、7, 10, 8, 11列目のデータのみ抽出して、チンパンジー2サンプル(PTM1 vs. チンパンジー2サンプル(G2群:PTF2とPTM2)の2群間比較結果をTCCで実行します。



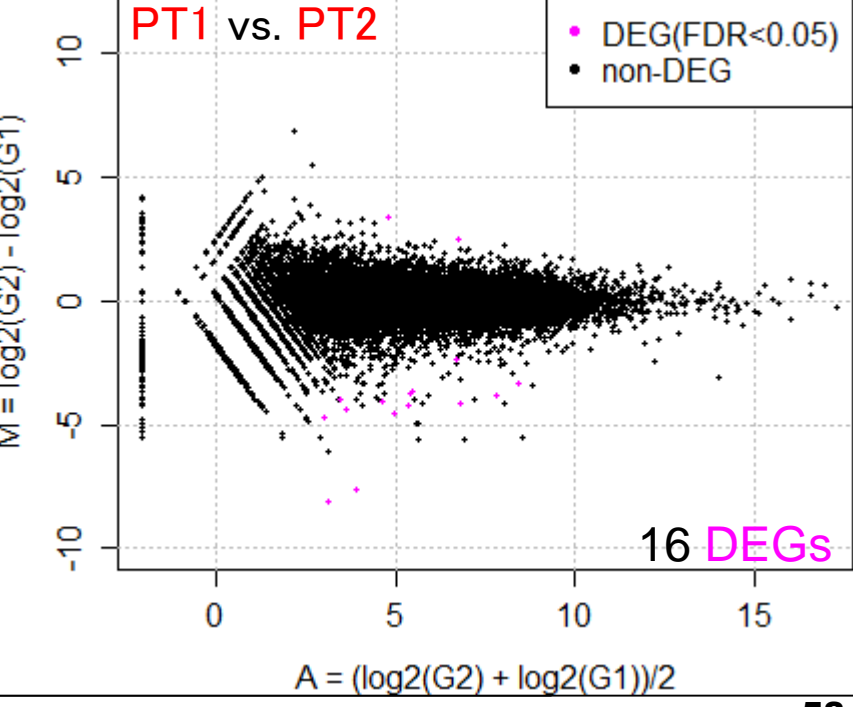
### 9. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の20,689 genes×18 samplesのリアルカウントデータ (sample blekhman 18.txt)です。ここでは、7, 10, 8, 11列目のデータのみ抽出して、チンパンジー2サンプル(PTM1 vs. チンパンジー2サンプル(G2群:PTF2とPTM2)の2群間比較結果をTCCで実行します。

```

in_f <- "sample_blekhman_18.txt" #入力ファイル名
out_f1 <- "hoge9.txt" #出力ファイル名
out_f2 <- "hoge9.png" #出力ファイル名
param_subset <- c(7, 10, 8, 11) #取り扱う列番号
param_G1 <- 2 #G1群のサンプル数
param_G2 <- 2 #G2群のサンプル数
param_FDR <- 0.05 #DEG検出率
param_fig <- c(430, 350) #ファイギュアサイズ
param_mar <- c(4, 4, 0, 0) #下、左、上、右のマージン

#必要なパッケージをロード
library(TCC)
    
```





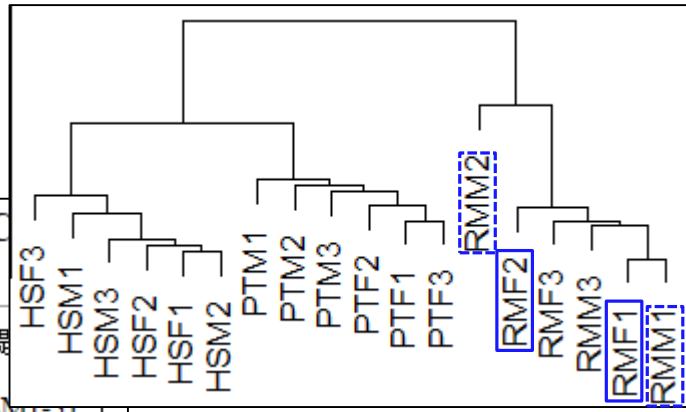
「G1群(RMM1とRMF1) vs. G2群(RMM2とRMF2)」の2群間比較結果。24 DEGs。

# TCC実行

- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun 2013) (last modified 2015/07/10)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun 2013) (last modified 2015/06/02)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq(Li 2013) (last modified 2015/06/02)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Lederer(Babinec 2010) (last modified 2015/06/02)

## 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun\_2013) NEW

Blekhman et al., Genome Res., 2010の公共カウントデータ解析に特化させて、TCCを用いた様々な例題があります。入力は全てサンプルデータ42の20,689 genes×18 samplesのリアルカウントデータ (sample blekhman 18.txt)です。ヒトHomo sapiens; HS)のメス3サンプル(HSF1-3)とオス3サンプル(HSM1-3)、チンパンジー(Pan troglodytes; PT)のメス3サンプル(PTF1-3)とオス3サンプル(PTM1-3)、アカゲザル(Rhesus macaque)のメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)のデータがあります。



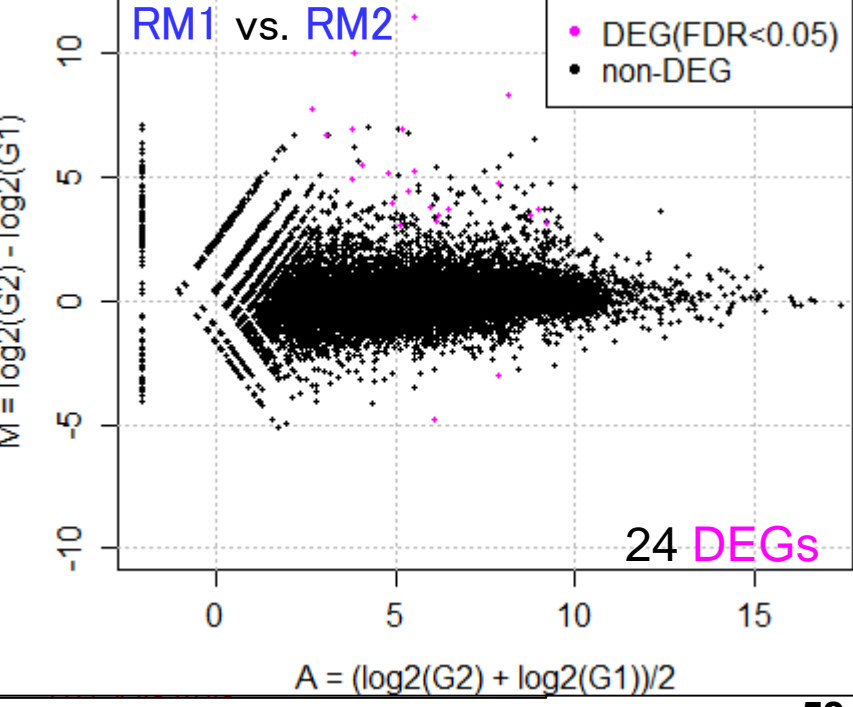
### 10. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の20,689 genes×18 sample サンプル(HSF1-3)とオス3サンプル(HSM1-3), チンパンジーのサンプル(PTM1-3), アカゲザルのメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)のデータがあります。ここでは、13, 16, 14, 17列目のデータのみ抽出して、G1群(RMM1) vs. アカゲザル2サンプル(G2群:RMF2とRMM2)の2群間比較結果。

```

in_f <- "sample_blekhman_18.txt" #入力ファイル
out_f1 <- "hoge10.txt" #出力ファイル
out_f2 <- "hoge10.png" #出力ファイル
param_subset <- c(13, 16, 14, 17) #取り扱う列番号
param_G1 <- 2 #G1群のサンプル数
param_G2 <- 2 #G2群のサンプル数
param_FDR <- 0.05 #DEG検出率
param_fig <- c(430, 350) #ファイギュアサイズ
param_mar <- c(4, 4, 0, 0) #下、左マージン

#必要なパッケージをロード
library(TCC)
    
```





# TCC実行(おまけ)

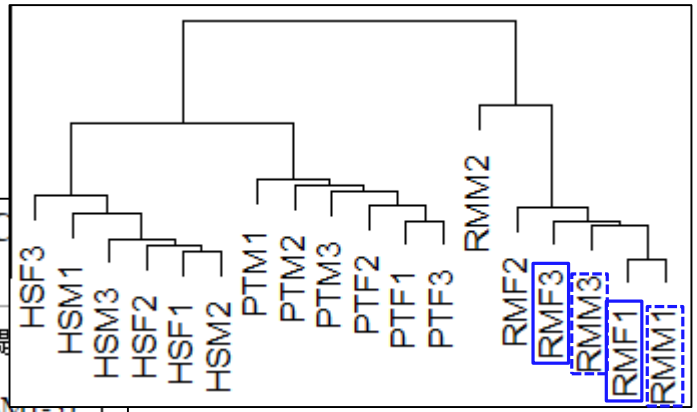
- 解析 | 発現変動 | について (last modified 2014/07/10)
- 解析 | 発現変動 | 2群間 | 対応なし | について (last modified 2015/06/02)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC(Sun 2013) (last modified 2015/07/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun 2013)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq(Li 2013) (last modified 2015/07/07)
- 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Lederer(Babinec 2010) (last modified 2015/07/07)

## 解析 | 発現変動 | 2群間 | 対応なし | 複製あり | Blekhmanデータ | TCC(Sun\_2013) NEW

Blekhman et al., Genome Res., 2010の公共カウントデータ解析に特化させて、TCCを用いた様々な例題があります。入力は全てサンプルデータ42の20,689 genes×18 samplesのリアルカウントデータ (sample blekhman 18.txt)です。ヒトHomo sapiens; HS)のメス3サンプル(HSF1-3)とオス3サンプル(HSM1-3)、チンパンジー(Pan troglodytes; PT)のメス3サンプル(PTF1-3)とオス3サンプル(PTM1-3)、アカゲザル(Rhesus macaque)のメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)のデータがあります。

### 11. サンプルデータ42のリアルデータ(sample blekhman 18.txt)の場合:

Blekhman et al., Genome Res., 2010の20,689 genes×18 sampleのリアルカウントデータ (sample blekhman 18.txt)です。ヒトHomo sapiens; HS)のメス3サンプル(HSF1-3)とオス3サンプル(HSM1-3)、チンパンジー(Pan troglodytes; PT)のメス3サンプル(PTF1-3)とオス3サンプル(PTM1-3)、アカゲザル(Rhesus macaque)のメス3サンプル(RMF1-3)とオス3サンプル(RMM1-3)のデータがあります。ここでは、13, 16, 15, 18列目のデータのみ抽出して、G1群(RMM1とRMF1) vs. アカゲザル2サンプル(G2群:RMF3とRMM3)の2群間比較結果をプロットしています。

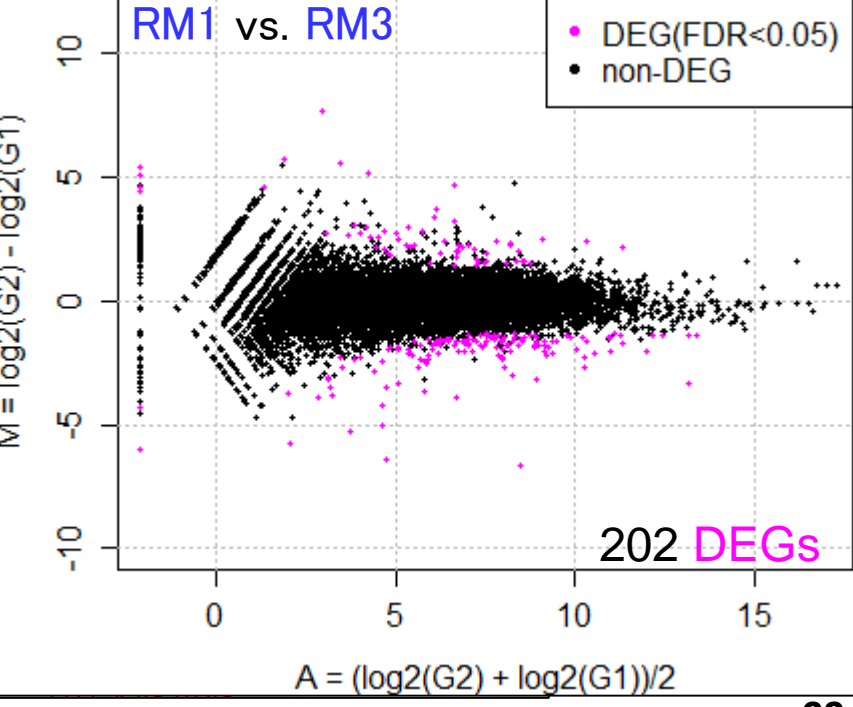


- 1. ヒト2サンプル (1, 4, 13, 16)
- 2. チンパンジー2サンプル (5, 6, 7, 8, 9, 10, 11, 12)
- 3. アカゲザル2サンプル (13, 14, 15, 16, 17, 18)

```

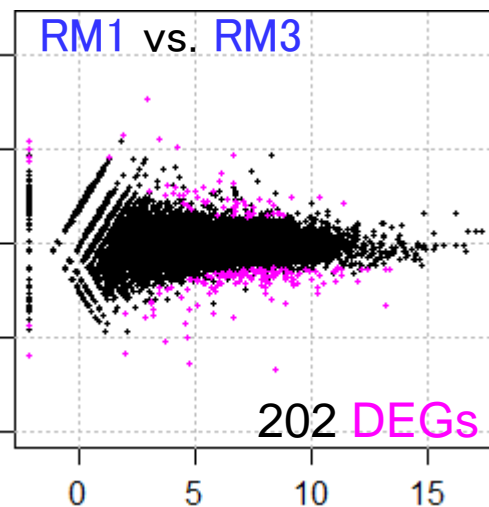
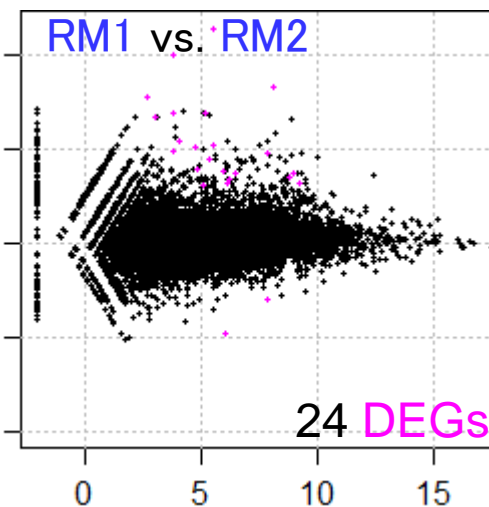
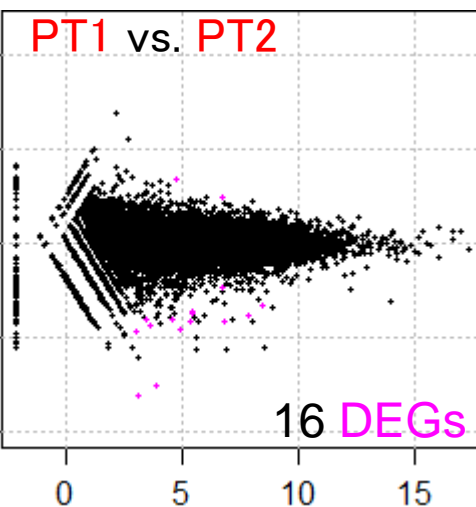
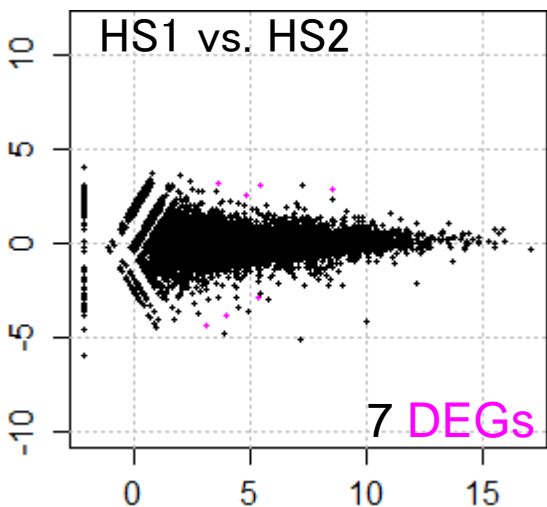
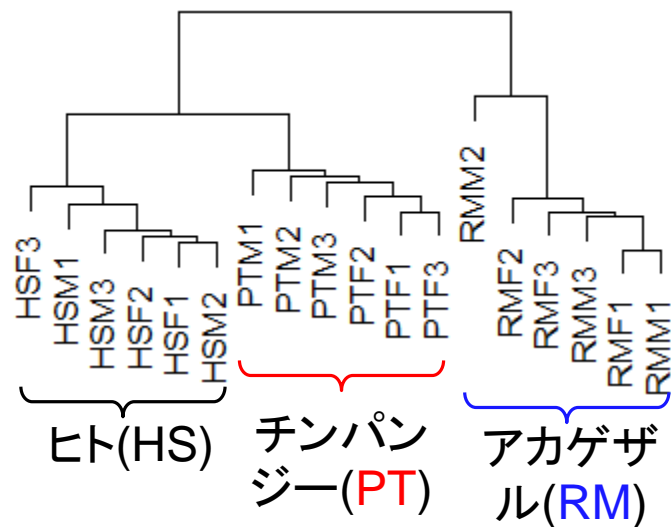
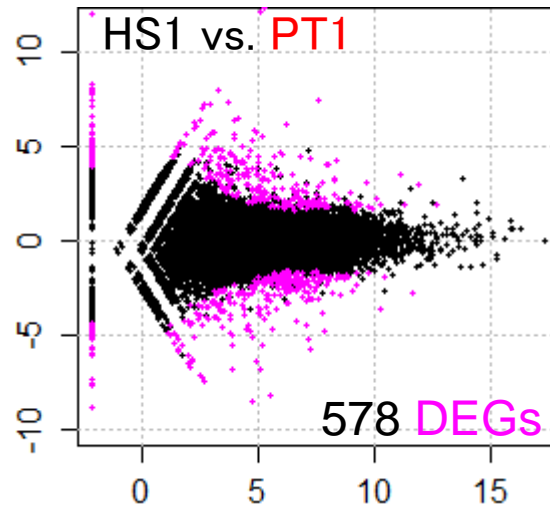
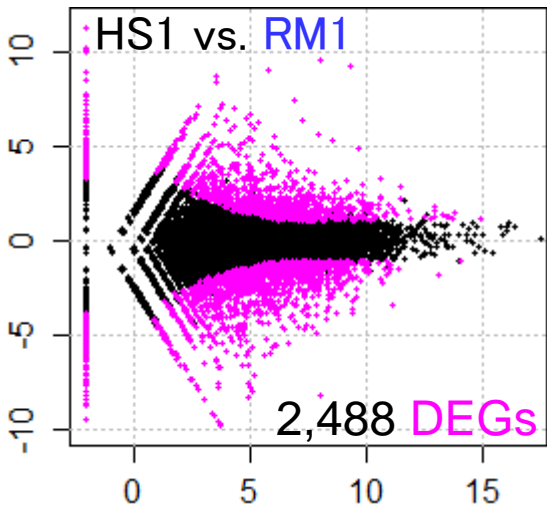
in_f <- "sample_blekhman_18.txt" #入力ファイル
out_f1 <- "hoge11.txt" #出力ファイル
out_f2 <- "hoge11.png" #出力ファイル
param_subset <- c(13, 16, 15, 18) #取り扱うサンプルID
param_G1 <- 2 #G1群のサンプル数
param_G2 <- 2 #G2群のサンプル数
param_FDR <- 0.05 #DEG検出率
param_fig <- c(430, 350) #ファイギュアサイズ
param_mar <- c(4, 4, 0, 0) #下、左、上、右のマージン

#必要なパッケージをロード
library(TCC)
    
```



# 結果の比較

同一群(下段)の分布は、異なる群(上段)の non-DEG分布とよく一致する。同一群内のばらつきの分布 (non-DEG分布) 以外のものが **DEG**と判定されるのが統計的手法の結果

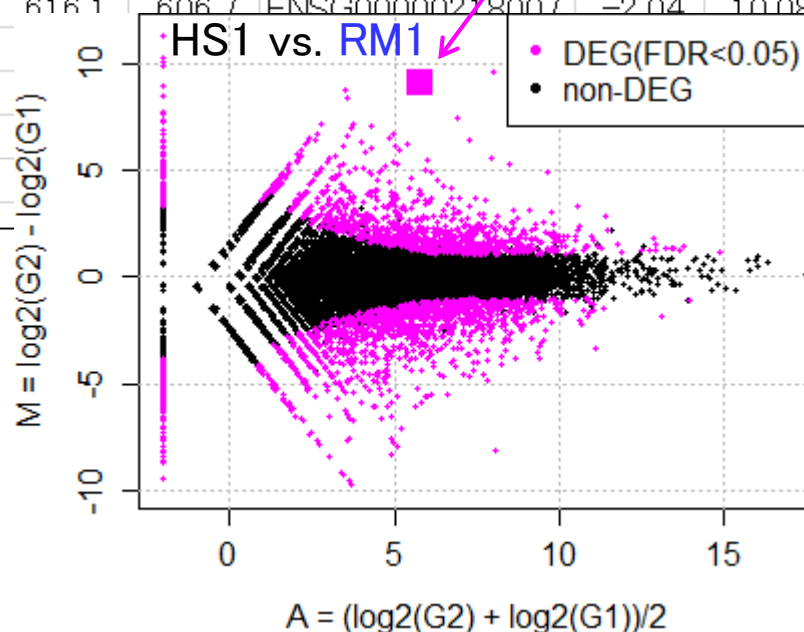


同一群内のばらつきの分布 (non-DEG分布) から遠く離れたところに位置するものは、0に近いp-value

# 統計的手法とは

- 同一群内の遺伝子のばらつきの程度を把握し、帰無仮説に従う分布の全体像を把握しておく (モデル構築)
  - non-DEGのばらつきの程度を把握しておくことと同義
- 実際に比較したい2群の遺伝子のばらつきの程度がnon-DEG分布のどのあたりに位置するかを評価 (検定)

rownames(tcc\$count)	HSF1	HSM1	RMF1	RMM1	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000208570	0.0	0.0	1346.8	1477.0	ENSG00000208570	-2.04	11.29	4.41E-53	9.13E-49	1	1
ENSG00000220191	2.3	2.5	1394.7	1171.1	ENSG00000220191	5.79	9.06	4.54E-47	4.70E-43	2	1
ENSG00000106366	4421.9	4411.0	23.1	8.3	ENSG00000106366	8.04	-8.14	2.46E-45	1.70E-41	3	1
ENSG00000209449	0.0	0.0	644.5	713.1	ENSG00000209449	-2.04	10.23	3.29E-44	1.70E-40	4	1
ENSG00000218007	0.0	0.0	616.1	606.7	ENSG00000218007	-2.04	10.08	1.74E-43	7.21E-40	5	1
ENSG00000070985	0.0	0.0						4.68E-42	1.61E-38	6	1
ENSG00000209007	0.0	0.0						1.24E-40	3.67E-37	7	1
ENSG00000182327	367.5	363.9						1.52E-38	3.93E-35	8	1
ENSG00000156222	367.5	301.5						1.09E-36	2.51E-33	9	1
ENSG00000165272	404.9	429.0						5.45E-28	1.12E-22	10	1

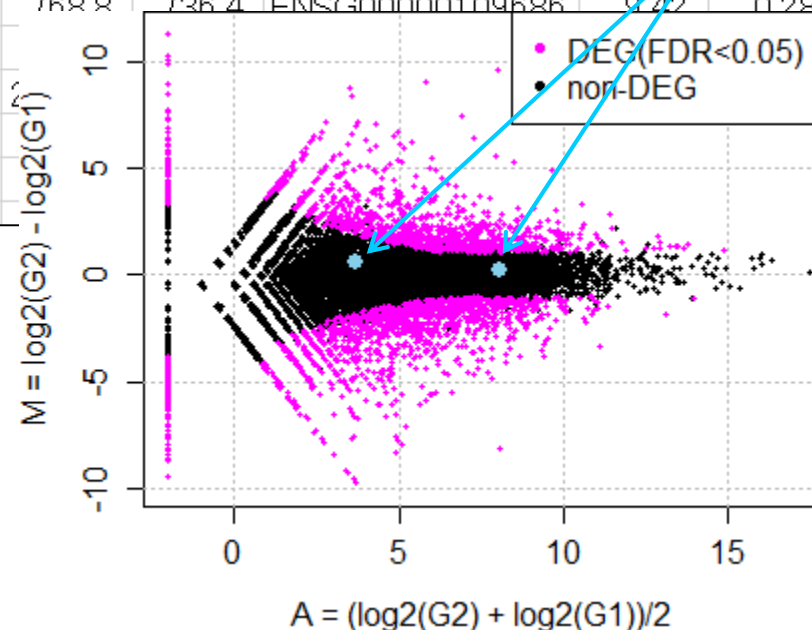


同一群内のばらつきの分布 (non-DEG分布) のど真ん中に位置するものは、1に近い  $p$ -value

# 統計的手法とは

- 同一群内の遺伝子のばらつきの程度を把握し、帰無仮説に従う分布の全体像を把握しておく (モデル構築)
  - non-DEGのばらつきの程度を把握しておくことと同義
- 実際に比較したい2群の遺伝子のばらつきの程度がnon-DEG分布のどのあたりに位置するかを評価 (検定)

rownames(tcc\$count)	HSF1	HSM1	RMF1	RMM1	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000047578	69.0	131.0	98.5	148.8	ENSG00000047578	6.80	0.31	0.4906	1	9727	0
ENSG00000115325	3.4	17.8	16.9	14.1	ENSG00000115325	3.68	0.55	0.49087	1	9728	0
ENSG00000122257	256.7	234.1	253.9	334.1	ENSG00000122257	8.07	0.26	0.49092	1	9729	0
ENSG00000090861	603.8	339.7	542.4	601.8	ENSG00000090861	9.02	0.28	0.491	1	9730	0
ENSG00000109686	451.1	792.6	768.8	736.4	ENSG00000109686	9.42	0.28	0.49109	1	9731	0
ENSG00000032389	53.1	36.9						0.49115	1	9732	0
ENSG00000125844	2299.7	3137.5						0.49127	1	9733	0
ENSG00000180190	52.0	28.0						0.4913	1	9734	0
ENSG00000100351	2.3	12.7						0.49134	1	9735	0
ENSG00000169554	72.5	122.6						0.49139	1	9736	0



# Contents

## ■ サンプル間クラスタリング

- 実行手順のおさらい
- 計算の一部を解説、結果の解釈

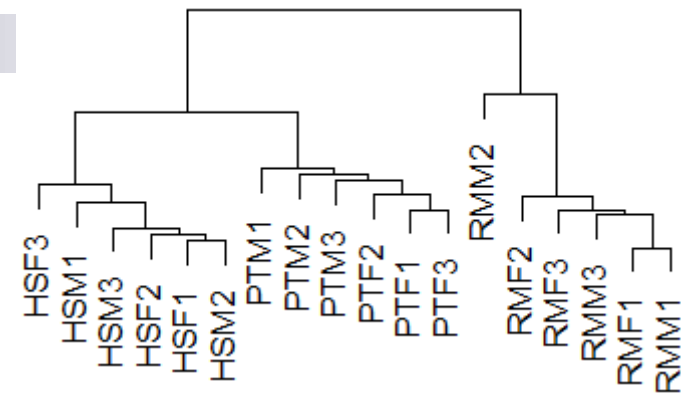
## ■ 発現変動解析(2群間比較) : 発現変動遺伝子(DEG)が多数存在する場合

## ■ モデル、分布、統計的手法、2群間比較でDEGがそれほど多くない場合

## ■ 2群間比較でDEGがほとんどない同一群の場合

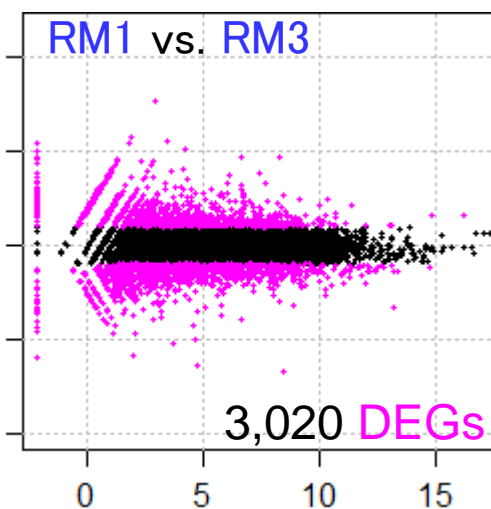
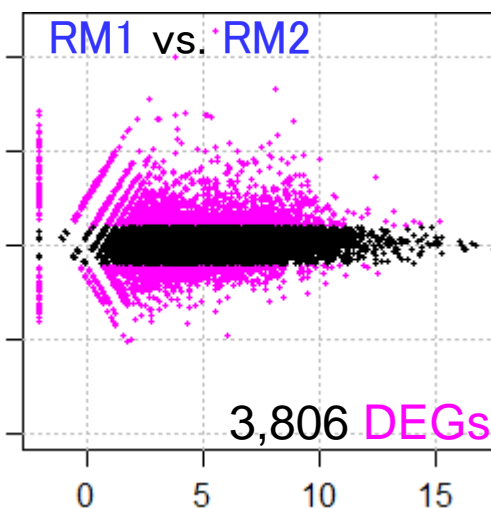
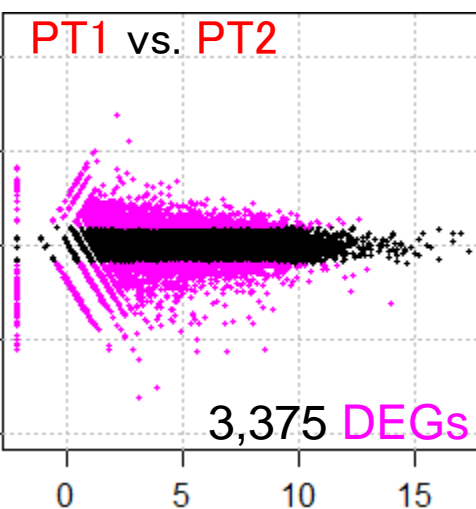
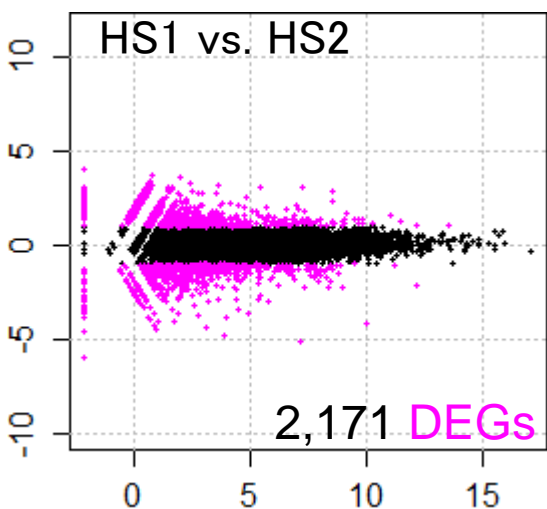
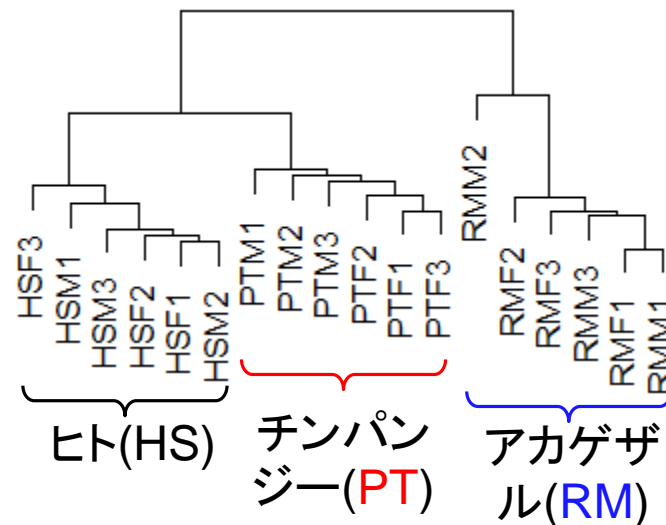
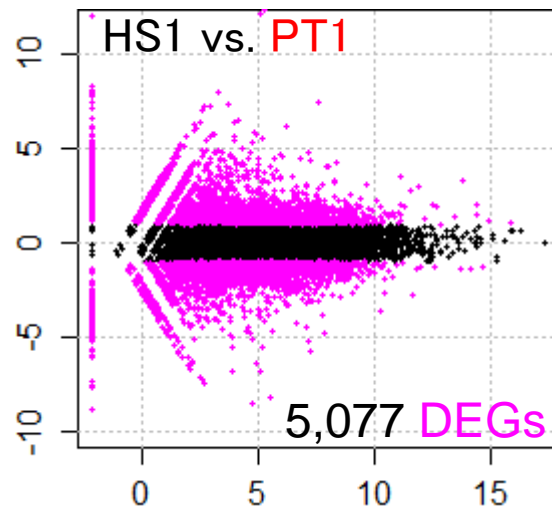
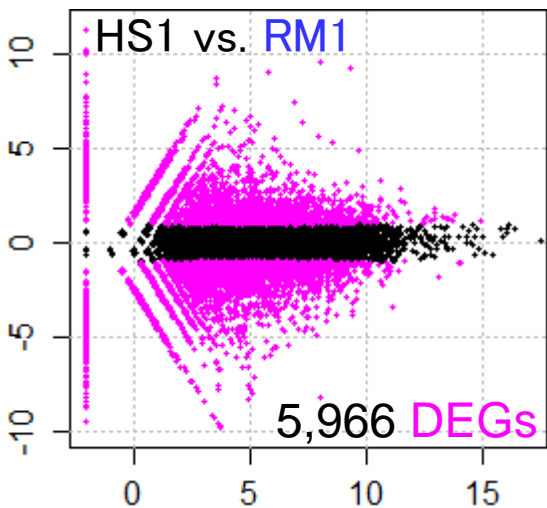
## ■ 倍率変化(2倍以上、1/2以下の発現変動)を用いた場合

## ■ 発現変動解析 : 3群間比較など



# 結果の比較(2倍変化)

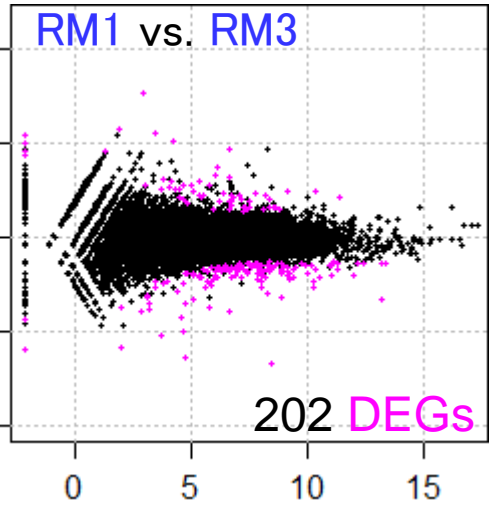
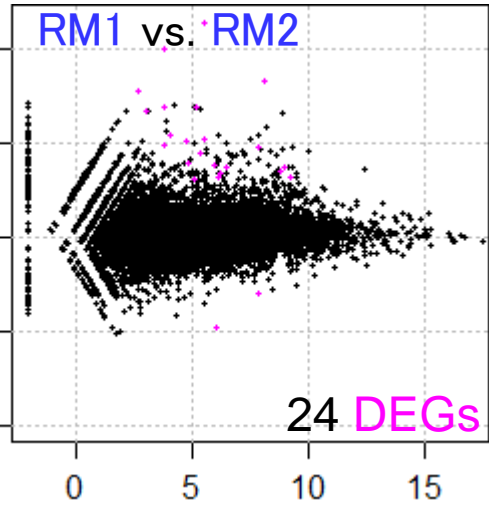
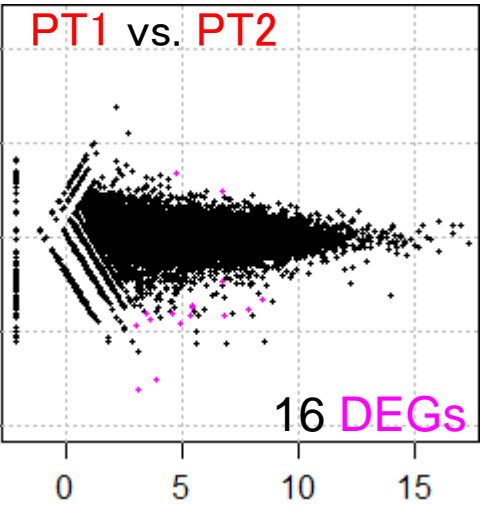
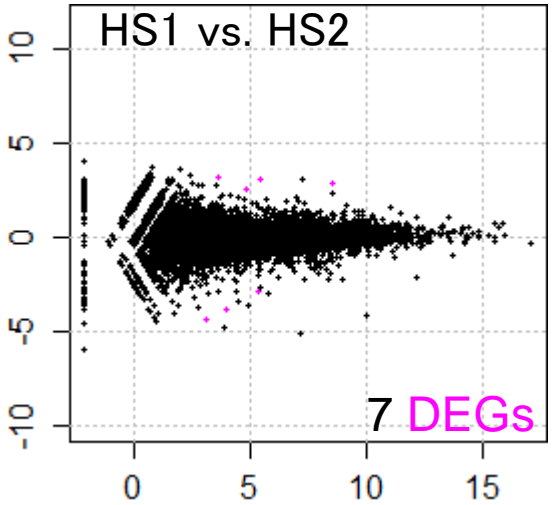
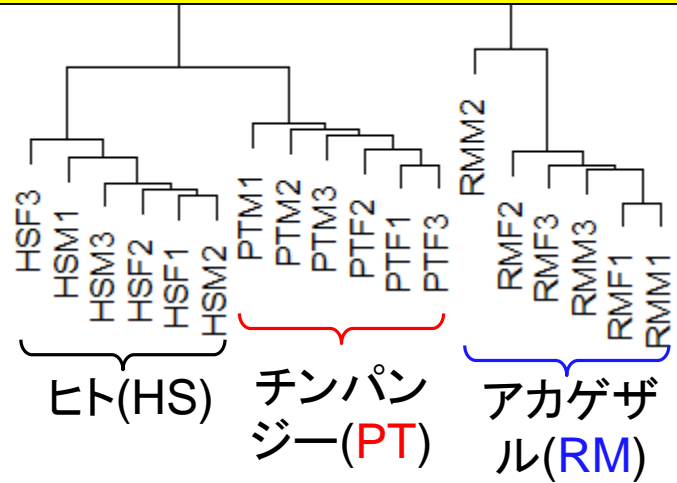
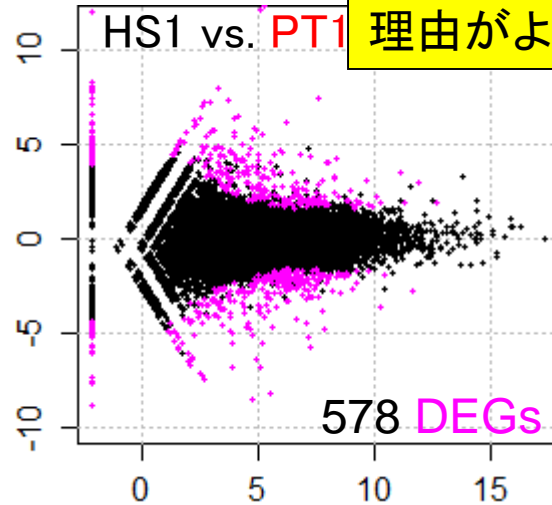
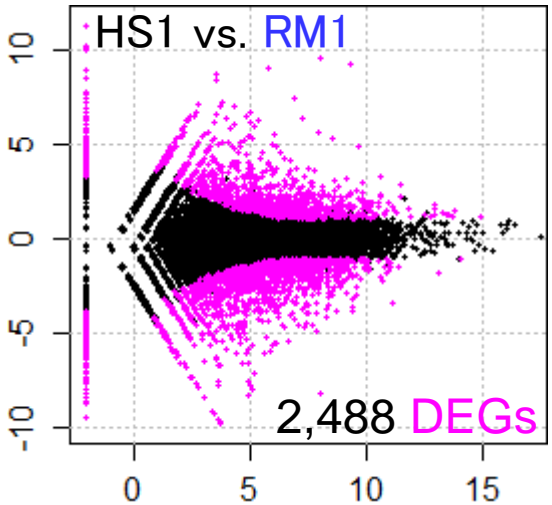
倍率変化(fold-change; FC)でのDEG検出結果。下段の同一群内比較でも多数の偽陽性が検出されている。例題13をベースに作成





統計的手法(TCC)も多少偽陽性が存在するが、倍率変化(FC)ほど凶悪ではないことがわかる。また高発現側のDEGは、FCと比較的よく一致していることがわかる。先人がFCのみで比較的信頼性の高い結果を得てきた理由がよくわかる(高発現側を信頼するという経験則)。

# 結果の比較(FDR)



# 3群間比較

①発現パターンごとの分類もしたい場合に便利。  
②post-hoc test的なことをやりたいときの項目。  
③複製なしデータの場合。③をクリック。

- [解析 | 発現変動 | 3群間 | 対応なし | について](#) (last modified 2015/02/10)
- [解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 基礎 | \[DESeq2\\(Love 2014\\)\]\(#\)](#) (last modified 2015/02/04)
- [解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 基礎 | \[TCC\\(Sun 2013\\)\]\(#\)](#) (last modified 2015/03/04) 推奨
- [解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 基礎 | \[EBSeq\\(Leng 2013\\)\]\(#\)](#) ① (last modified 2015/02/10)
- [解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 基礎 | \[SAMseq\\(Li 2013\\)\]\(#\)](#) (last modified 2015/02/10)
- [解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 基礎 | \[edgeR\\(Robinson 2010\\)\]\(#\)](#) (last modified 2015/02/03)
- [解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 基礎 | \[DESeq\\(Anders 2010\\)\]\(#\)](#) (last modified 2014/03/13)
- [解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 | \[TCC\\(Sun 2013\\)\]\(#\)](#) ② (last modified 2015/01/29) 推奨
- [解析 | 発現変動 | 3群間 | 対応なし | 複製なし | \[TCC\\(Sun 2013\\)\]\(#\)](#) ③ (last modified 2015/07/07) 推奨 **NEW**
- [解析 | 発現変動 | 5群間 | 対応なし | 複製あり | \[TCC\\(Sun 2013\\)\]\(#\)](#) (last modified 2014/08/22) 推奨
- [解析 | 発現変動 | 時系列 | について](#) (last modified 2014/12/19)
- [解析 | 発現変動 | 時系列 | \[Bayesian model-based clustering\\(Nascimento 2012\\)\]\(#\)](#) (last modified 2012/09/10)
- [解析 | 発現変動 | 時系列 | \[maSigPro\\(Nueda 2014\\)\]\(#\)](#) (last modified 2014/07/18)

# 3群間(複製なし)

①の箇所の記述は若干先走ってます。2015年4月リリースのBioconductor ver. 3.1で提供しているTCCパッケージ(ver. 1.8.0)では、まだ内部的に利用するパッケージがデフォルトではDESeqになっています。複製なしの場合はDESeq2を内部的に用いるほうが精度が高いことが分かったので、Bioconductor ver. 3.2(2015年10月リリース予定)では、デフォルトがDESeq2になります。これもバージョンアップの意義の1つであり、デフォルトオプションが変わりうる一例。

- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 基礎 | [DESeq\(Anders 2010\)](#) (last modified 2014/12/19)
- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 | [TCC\(Sun 2013\)](#) (last modified 2014/12/19)
- 解析 | 発現変動 | 3群間 | 対応なし | 複製なし | [TCC\(Sun 2013\)](#) (last modified 2014/12/19)
- 解析 | 発現変動 | 5群間 | 対応なし | 複製あり | [TCC\(Sun 2013\)](#) (last modified 2014/12/19)
- 解析 | 発現変動 | 時系列 | [TCC\(Sun 2013\)](#) (last modified 2014/12/19)
- 解析 | 発現変動 | 時系列 | [maSigPro\(Nueda 2014\)](#) (last modified 2014/07/18)
- 解析 | 発現変動 | 時系列 | [Bayesian model-based clustering\(Nascimento 2014\)](#) (last modified 2014/07/18)

## 解析 | 発現変動 | 3群間 | 対応なし

TCCを用いたやり方を示します。内部的にDESeq2パッケージ中の関数を利用して、複製なしデータに対応済みのDESeq2の通常の手順を複数回繰り返す(DEGES-based normalization; Kadota et al., 2012)ことでより正確なデータ正規化が実現された発現変動解析結果を得ることができます。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピー。

### 1. サンプルデータ43の10,000 genes×3 samplesのカウントデータ(data\_hypodata\_1vs1vs1.txt)の場合:

シミュレーションデータ(G1群1サンプル vs. G2群1サンプル vs. G3群1サンプル)です。gene\_1~gene\_3000までがDEG (gene\_1~gene\_2100がG1群で3倍高発現、gene\_2101~gene\_2700がG2群で10倍高発現、gene\_2701~gene\_3000がG3群で6倍高発現) gene\_3001~gene\_10000までがnon-DEGであることが既知です。2015年7月7日現在(TCC ver. 1.8.0)、デフォルトでは正規化のところでDESeqが動いているので次期リリースでDESeq2になるようにします。約1分かかります。

```
in_f <- "data_hypodata_1vs1vs1.txt"
out_f <- "hoge1.txt"
param_G1 <- 1
param_G2 <- 1
param_G3 <- 1
param_FDR <- 0.05
```

#入力ファイル名を指定してin\_fに格納  
#出力ファイル名を指定してout\_fに格納  
#G1群のサンプル数を指定  
#G2群のサンプル数を指定  
#G3群のサンプル数を指定  
#DEG検出時のfalse discovery rate (FDR)閾値を指定

#必要なパッケージをロード

# シミュレーションデータ

他にも多くの解析用パッケージが存在し、このウェブページ上で紹介しきれていないものが多く存在します。また、バージョンアップなどに追いついていない項目も多くあります。そのため、正しい手順で解析できているのかが不安な局面があるでしょう。そういうときはTCCパッケージ中のシミュレーションデータ作成関数を利用して、「これがDEG検出結果の上位に来ていないやり方はオカシイはず」というようなデータを自分で作成して検証するのです。

- [解析 | クラスタリング | 遺伝子間 | MBCluster.Seq \(Si 2014\) \(last modified 2014/07/10\)](#)
- [解析 | シミュレーションカウントデータ | について \(last modified 2015/01/25\)](#)
- [解析 | シミュレーションカウントデータ | Technical rep.\(ポアソン分布\) \(last modified 2015/01/23\)](#)
- [解析 | シミュレーションカウントデータ | Biological rep. | 基礎編 \(last modified 2015/01/23\)](#)
- [解析 | シミュレーションカウントデータ | Biological rep. | 2群間 | 基礎編 | TCC\(Sun 2013\) \(last modified 2015/01/28\)](#)
- [解析 | シミュレーションカウントデータ | Biological rep. | 2群間 | 応用編 | TCC\(Sun 2013\) \(last modified 2015/01/28\)](#)
- [解析 | シミュレーションカウントデータ | Biological rep. | 3群間 | 基礎編 | TCC\(Sun 2013\) \(last modified 2015/01/28\)](#)
- [解析 | 発現変動 | について \(last modified 2014/07/10\)](#)
- [解析 | 発現変動 | 2群間 | 対応なし | について \(last modified 2015/02/02\)](#)
- [解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC\(Sun 2013\) \(last modified 2015/02/26\)推奨 \*\*NEW\*\*](#)

