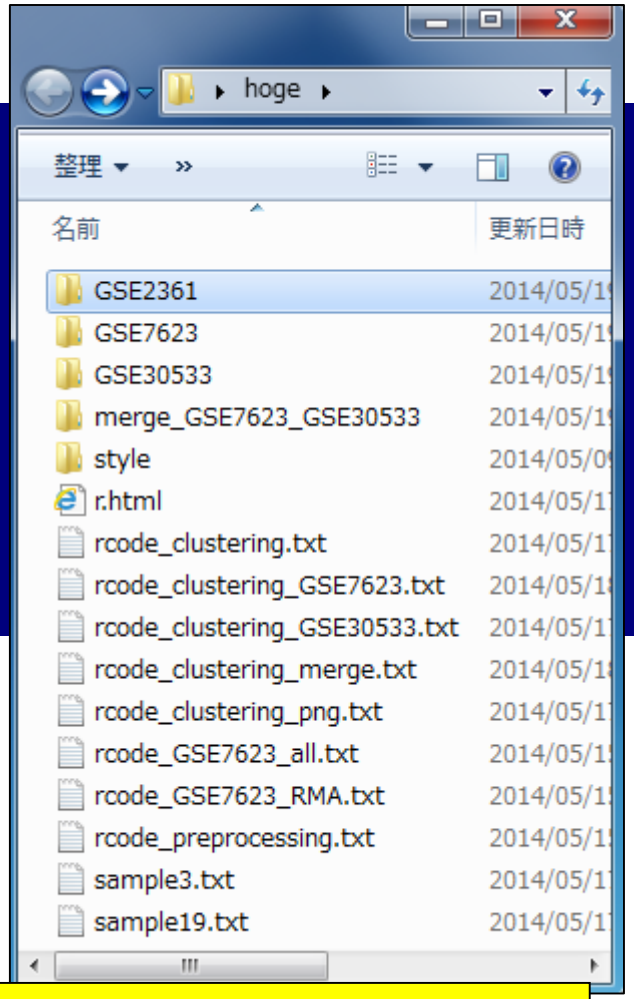
講義室後ろにあるUSBメモリ中のhogeフォルダをデスクトップにコピーしておいてください。



機能ゲノム学 第2回

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ユニット
門田幸二
kadota@iu.a.u-tokyo.ac.jp

前回(5/14)のhogeフォルダがデスクトップに残っているかもしれないのでご注意ください。

講義予定

- 第1回（2014年5月14日）
  - 原理、各種データベース、生データ取得、遺伝子発現行列作成（データ正規化）
  - 教科書の1.2節、2.2節周辺
- 第2回（2014年5月21日）
  - クラスタリング（データ変換や距離の定義など）、実験デザイン、分布
  - 教科書の3.2節周辺
- 第3回（2014年5月28日）
  - 発現変動解析（多重比較問題）、各種プロット（M-A plotや平均-分散プロット）
  - 教科書の3.2節と4.2節周辺
- 第4回（2014年6月4日）
  - 機能解析（Gene Ontology解析やパスウェイ解析）、分類など

授業の目標・概要

細胞中で発現している全転写物（トランスクリプトーム）の解析技術は、マイクロアレイから次世代シーケンサ（RNA-seq）に移行しつつあります。RNA-seqデータ解析の多くは、マイクロアレイの知識を前提としています。また、ニュートリゲノミクス（食品系）分野では、マイクロアレイは現在でも主流派です。**マイクロアレイデータを主な例として**、各種トランスクリプトーム解析手法について解説します。

Contents（第2回）

- プローブレベルデータ取得および前処理法適用（発展形）
  - Rパッケージ経由でダウンロードしたデータをそのまま前処理法にかける
  - サンプル名の順番が変わることもあるので注意
  - 何度も同じ作業をしなくていいように有効活用
  - 3つの論文の遺伝子発現行列をコピペで作成
    - ヒトの36サンプルからなるデータ（Ge et al., 2005）
    - ラットの24サンプルからなるデータ（Nakai et al., 2008）
    - ラットの10サンプルからなるデータ（Kamei et al., 2013）
- サンプル間クラスタリング
  - 階層的クラスタリング、距離の定義、クラスターをまとめる方法など
  - 前処理法（MAS5, RMA, RobLoxBioC）と相関係数（Spearman and Pearson）の違い
  - 同一アレイ由来データセットのマージ（ラット24サンプル＋ラット10サンプル）
    - 3’発現アレイの長所（教科書p7）
    - ラット10サンプル（通常　対　鉄欠乏）クラスタリング結果の印象は、外群（ラット24サンプル）の有無でずいぶん異なる（教科書p106-107）
- 実験デザイン

CELファイル取得 → 前処理法の実行

- サンプルデータ (last modified 2013/11/25)
- 書籍 | について (last modified 2014/04/17) NEW
- 書籍 | トランスクリプトーム解析 | 1.1 はじめに (last modified 2014/05/09) NEW
- 書籍 | トランスクリプトーム解析 | 1.2.1 原理(Affymetrix 3'発現アレイ) (last modified 2014/05/12) NEW
- 書籍 | トランスクリプトーム解析 | 1.2.2 最近の知見 (last modified 2014/05/09) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.1 生データ(プローブレベルデータ)取得 (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.2 データの正規化(基礎) (last modified 2014/04/17) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.3 データの正規化(計算例) (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.4 データの正規化(その他) (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.5 アノテーション情報 (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.1 クラスタリング(データ変換や距離の定義など) (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.2 実験デザイン, データ分布, 統計解析との関係 (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.3 多重比較問題 (last modified 2014/04/19) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.4 各種プロット(M-A plotや平均-分散プロットなど) (last modified 2014/04/19) NEW
- 書籍 | トランスクリプトーム解析 | 4.2.1 2群間比較 (last modified 2014/04/19) NEW
- 書籍 | トランスクリプトーム解析 | 4.2.2 他の実験デザイン(paired, multi-factor, 3群間) (last modified 2014/04/19) NEW
- 書籍 | トランスクリプトーム解析 | 4.2.3 多群間比較(特異的発現パター
- イントロ | 発現データ取得 | 公共DBから (last modified 2014/05/11) NE
- イントロ | 発現データ取得 | inSilicoDb(Taminau_2011) (last modified 2(
- イントロ | 発現データ取得 | ArrayExpress(Kauffmann_2009) (last modif
- イントロ | 発現データ取得 | GEOquery(Davis_2007) (last modified 2013
- イントロ | アノテーション情報取得 | 公共DB(GEO)から (last modified 2(
- イントロ | アノテーション情報取得 | GEOquery(Davis_2007) (last modifi
- イントロ | アノテーション情報取得 | Rのパッケージ*.dbから (last modifi

一気にやる方法もあります

①CELファイル取得
②前処理法の実行

- 正規化 | Affymetrix GeneChip | について (last modified 2013/09/02)
- 正規化 | Affymetrix GeneChip | frma (McCall_2010) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | rma++ (Kohl_2010) (last modified 2013/11/19)推奨
- 正規化 | Affymetrix GeneChip | GRSN (Pelz_2008) (last modified 2013/05/27)
- 正規化 | Affymetrix GeneChip | Hook (Binder_2008) (last modified 2013/05/30)
- 正規化 | Affymetrix GeneChip | DFW (Chen_2007) (last modified 2013/08/20)
- 正規化 | Affymetrix GeneChip | FARMS (Hochreiter_2006) (last modified 2013/08/20)
- 正規化 | Affymetrix GeneChip | multi-mgMOS (Liu_2005) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | GCRMA (Wu_2004) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | PLIER (Affymetrix_2004) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | VSN (Huber_2002) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | RMA (Irizarry_2003) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | MAS5.0 (Hubbell_2002) (last modified 2013/11/25)
- 正規化 | Affymetrix GeneChip | MBEI (Li_2001) (last modified 2013/08/21)

CELファイル取得 → 前処理法の実行

- 書籍 | トランスクリプトーム解析 | 4.2.2 他の実験デザイン(paired, multi-factor, 3群間) (last modified 2014/04/19)
- 書籍 | トランスクリプトーム解析 | 4.2.3 多群間比較(特異的発現パターン) (last modified 2014/04/20) NEW
- イントロ | 発現データ取得 | 公共DBから (last modified 2014/05/11) NEW
- イントロ | 発現データ取得 | inSilicoDb(Taminau_2011) (last modified 2013/08/20)
- イントロ | 発現データ取得 | ArrayExpress(Kauffmann_2009) (last modified 2013/08/29)推奨
- イントロ | 発現データ取得 | GEOquery(Davis_2007) (last modified 2013/08/20)
- イントロ | アノテーション情報取得 | 公共DB(GEO)から (last modified 2013/08/18)
- イントロ | アノテーション情報取
- イントロ | アノテーション情報取
- イントロ | プローブ配列情報取
- イントロ | トランスクリプトーム
- イントロ | トランスクリプトーム
- イントロ | Affymetrix CELファイ

イントロ | 発現データ取得 | ArrayExpress(Kauffmann_2009) NEW

マイクロアレイデータベースArrayExpressに登録されているデータをArrayExpressというRパッケージで取得するやり方を示します。GEO IDでも検索可能であり、CELファイルデータも取得可能、任意のpreprocessing法を適用可能、などの利点からこのパッケージ経由での利用をお勧めします。
「ファイル」-「ディレクトリの変更」でファイルを保存したいディレクトリに移動し以下をコピペ。

1. AffymetrixデータE-MEXP-1422 (Bourgon et al., PNAS, 2010)のCELファイルを取得し、RMA法(Irizarry et al., Biostatistics, 2003)を実行して得られた発現情報を取得したい場合:

以下のArrayExpress関数のオプションをsave=Fからsave=Tに変更すると、CELファイルなどを含む全データのダウンロードも同時に行ってくれます。が、そんなことをいちいちやらなくてもReadAffy関数を用いて読み込んだ状態と同じなので直接RMA(Irizarry et al., Biostatistics, 2003)などの任意の正規化法を適用可能です。

```
out_f <- "data_rma.txt"                 #出力ファイル名を指定してout_fに格納
param <- "E-MEXP-1422"                  #入手したいIDを指定

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

#前処理(データ取得)
hoge <- ArrayExpress(param, save=F)     #paramで指定したIDのデータを取得した結果をhogeに格納

#本番
eset <- rma(hoge)                       #RMAを実行し、結果をesetに保存

#ファイルに保存
write.exprs(eset, file=out_f)           #結果を指定したファイル名で保存
```

これが一気にやるための
テンプレートスクリプトです

R経由で生データ取得(教科書の§2.2.1)

- Affymetrix GeneChip
  - Ge et al., *Genomics*, **86**: 127–141, 2005
    - **GSE2361**、ヒト36サンプル、GPL96を利用
  - Nakai et al., *Biosci Biotechnol Biochem.*, **72**: 139–148, 2008
    - **GSE7623**、ラット24サンプル、GPL1355を利用
  - Kamei et al., *PLoS One*, **8**: e65732, 2013
    - **GSE30533**、ラット10サンプル、GPL1355を利用
- Illumina BeadChip
  - Sharma et al., *Cancer Cell*, **23**: 35–47, 2013
    - **GSE28680**、ヒト24サンプル、GPL10558を利用
- NGSデータも…
  - Neyret-Kahn et al., *Genome Res.*, **23**: 1563–1579, 2013
    - **GSE42213**、ヒト26サンプル、GPL10999とGPL11154を利用
      - **GSE42211**、ヒト20サンプル、GPL10999とGPL11154を利用(ChIP-seq)
      - **GSE42212**、ヒト6サンプル、GPL10999を利用(RNA-seq)
  - Huang et al., *Development*, **139**: 2161–2169, 2012
    - **GSE36469**、シロイヌナズナ8サンプル、GPL13222を利用

GSE7623のプローブレベルデータ取得からRMA前処理法の実行までを一気にやる

イントロ | 発現データ取得 | ArrayExpress(Kauffmann_2009) NEW

マイクロアレイデータベースArrayExpressに登録されているデータをArrayExpressというRパッケージで取得するやり方を示します。
GEO IDでも検索可能であり、CELファイルデータも取得可能、任意のpreprocessing法を適用可能、などの利点からこのパッケージ経由での利用をお勧めします。
「ファイル」-「ディレクトリの変更」でファイルを保存したいディレクトリに移動し以下をコピペ。

1. AffymetrixデータE-MEXP-1422 (Bourgon et al., PNAS, 2010)のCELファイルを取得し、RMA法(Irizarry et al., Biostatistics, 2003)を実行して得られた発現情報を取得したい場合:

以下のArrayExpress関数のオプションをsave=Fからsave=Tに変更すると、CELファイルなどを含む全データのダウンロードも同時に行ってくれます。が、そんなことをいちいちやらなくてもReadAffy関数を用いて読み込んだ状態と同じなので直接RMA (Irizarry et al., Biostatistics, 2003)などの任意の正規化法を適用可能です。

```
out_f <- "data_rma.txt"                 #出力ファイル名を指定してout_fに格納
param <- "E-MEXP-1422"                  #入手したいIDを指定

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

#前処理(データ取得)
hoge <- ArrayExpress(param, sav

#本番
eset <- rma(hoge)

#ファイルに保存
write.exprs(eset, file=out_f)
```

```
##############################################################↓
###  GSE7623のCELファイルダウンロードとRMA前処理実行  ###↓
##############################################################↓
out_f <- "data_rma.txt"                 #出力ファイル名を指定してout_fに格納↓
param <- "GSE7623"                      #入手したいIDを指定↓
↓
#必要なパッケージをロード↓
library(ArrayExpress)                   #パッケージの読み込み↓
library(affy)                           #パッケージの読み込み↓
↓
#前処理(データ取得)↓
hoge <- ArrayExpress(param, save=F)     #paramで指定したIDのデータを取得した結果をhogeに格納↓
↓
#本番↓
eset <- rma(hoge)                       #RMAを実行し、結果をesetに保存↓
↓
#ファイルに保存↓
write.exprs(eset, file=out_f)           #結果を指定したファイル名で保存↓
←
```

必要最小限の変更後

R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> ############################################################
> ###  GSE7623のCELファイルダウンロードとRMA前処理実行  ###
> ############################################################
> out_f <- "data_rma.txt"                    #出力ファイル名を指定してout_fに格納
> param <- "GSE7623"                         #入手したいIDを指定
>
> #必要な
> library
 要求された
 要求された
 要求された

 次のパッケ

 以下のオブ

     clus
   clust
   parLa

 以下のオブ
```

R Console

```
ArrayExpress: Reading pheno data from SDRF
ArrayExpress: Reading data files
Read 49 items

 GSE7623  was successfully loaded into  AffyBatch

 警告メッセージ:
1: In strsplit(g, "\t") :  入力文字列 1 はこのロケールでは不適切です
2: In strsplit(g, "\t") :  入力文字列 1 はこのロケールでは不適切です
>
> #本番
> eset <- rma(ho

Background corre
Normalizing
Calculating Expr
 警告メッセージ:
1: replacing pre
2: replacing pre
>
> #ファイルに保存
> write.exprs(es
>
```

列の順番は異なっているものの、中の数値は同じ

data_rma.txt - Excel

ファイル　ホーム　挿入　ページ レイアウト　数式　データ　校閲　表示　アドイン　門田幸二

L16　9.81808731077457

	A	B	C	D	E	F	G	H	I
1		GSM184434	GSM184414	GSM184418	GSM184427	GSM184437	GSM184420	GSM184421	GSM184
2	1367452_at	9.65660613	10.524571	10.195674	10.3803459	9.45655687	9.74547642	10.0793083	10.3261
3	1367453_at	9.56071583	9.66424171	9.65657585	9.61752258	9.24437968	9.86596543	9.9082944	9.72840
4	1367454_at	9.60223688	9.65369911	9.79645967	9.61747771	9.85195553	9.727593	9.78282946	9.47288
5	1367455_at	11.4411168	10.7645706	10.5480622	10.9425437	11.2800063	10.7885641	10.7542919	11.1411
6	1367456_at	11.5355052	11.7127252	11.3259473	11.6187254	11.6465211	11.3108027	11.4757781	11.5824
7	1367457_at	9.07839762	8.96244355	8.76459798	9.26592163	9.09357188	8.38811185	8.45840226	9.11513
8	1367458_at	8.50569408	8.28183868	7.5182244	7.59895175	7.8007724	7.85110536	7.98179758	8.08825
9	1367459_at	11.8201046	11.8071	11.6800452	11.7443567	11.6770477	11.5597179	11.5977639	11.914
10	1367460_at	11.1040311	11.6317936	11.5323152	12.0131555	10.996116	11.6244301	11.6467453	11.8150

data_rma

準備完了　100%

hogeオブジェクトの中身は実質的に同じ

正規化 | Affymetrix GeneChip | RMA (Irizarry_2003)

Affymetrix chip (GeneChipTM)を用いて得られた*.CELファイルを元に、RMA(Irizarry et al., Biostatistics, 2003)アルゴリズムを用いてSummary scoreを算出。
「ファイル」-「ディレクトリの変更」で適切なディレクトリに移動し以下をコピペ。

1. (CELファイルがあるディレクトリ上で)手元にあるCELファイルの読み込みから行う場合:

```
out_f <- "hoge1.txt"                    #出力ファイル名を指定してout_fに格納

#必要なパッケージをロード
library(affy)

#データファイルの読み込み(*.CELファイル)
hoge <- ReadAffy()

#本番
eset <- rma(hoge)

#ファイルに保存
write.exprs(eset, file=out_f)
```

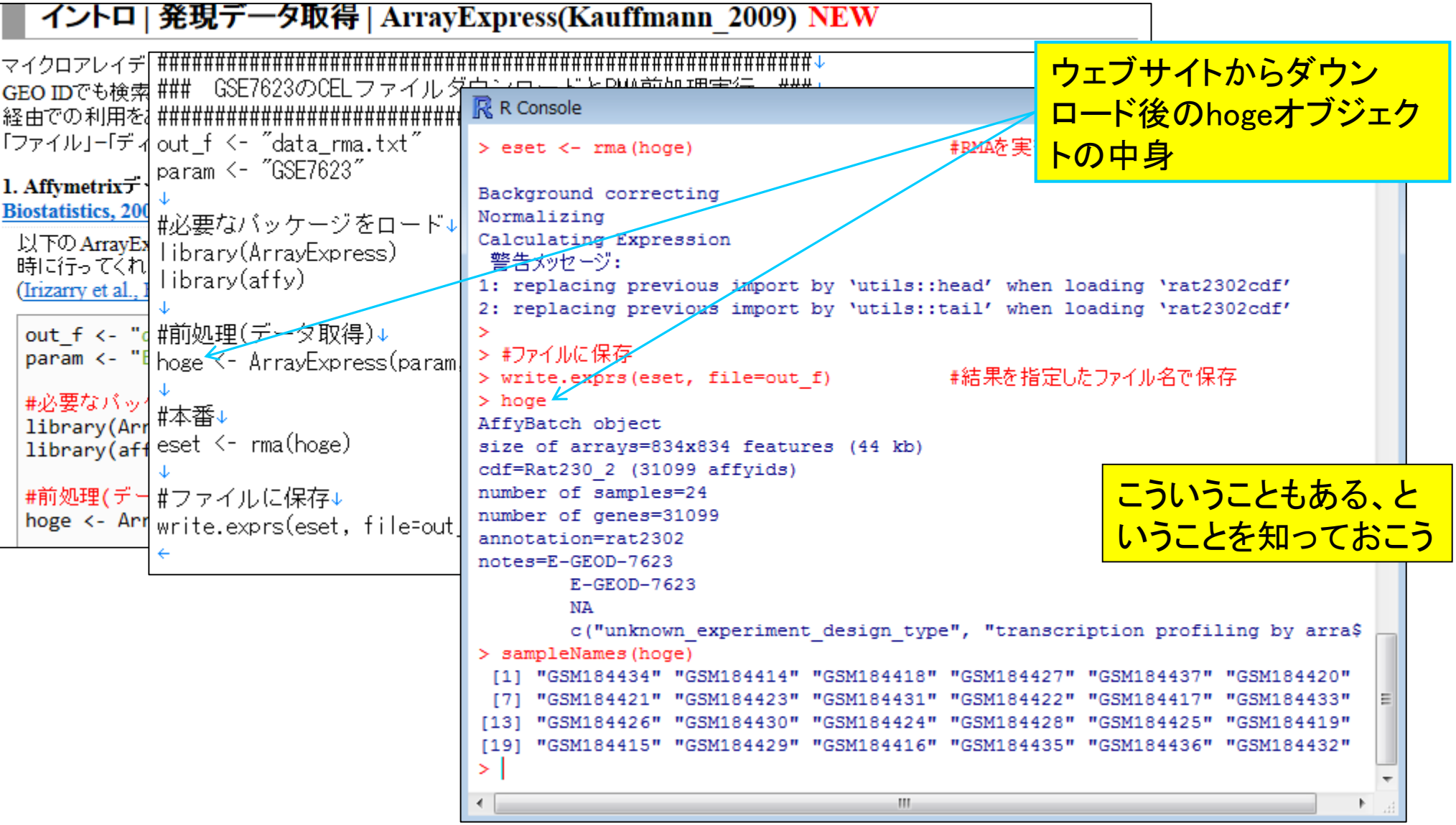
作業ディレクトリ中のCELファイルを読み込んだ後のhogeオブジェクトの中身

R Console

```
> library(affy)                         #パッケージの読み込み
>
> #データファイルの読み込み(*.CELファイル)
> hoge <- ReadAffy()                    #*.CELファイルの読み込み
> hoge

AffyBatch object
size of arrays=834x834 features (24 kb)
cdf=Rat230_2 (31099 affyids)
number of samples=24
number of genes=31099
annotation=rat2302
notes=
 警告メッセージ:
1: replacing previous import by 'utils::head' when loading 'rat2302cdf'
2: replacing previous import by 'utils::tail' when loading 'rat2302cdf'
> sampleNames(hoge)
 [1] "GSM184414.CEL" "GSM184415.CEL" "GSM184416.CEL" "GSM184417.CEL"
 [5] "GSM184418.CEL" "GSM184419.CEL" "GSM184420.CEL" "GSM184421.CEL"
 [9] "GSM184422.CEL" "GSM184423.CEL" "GSM184424.CEL" "GSM184425.CEL"
[13] "GSM184426.CEL" "GSM184427.CEL" "GSM184428.CEL" "GSM184429.CEL"
[17] "GSM184430.CEL" "GSM184431.CEL" "GSM184432.CEL" "GSM184433.CEL"
[21] "GSM184434.CEL" "GSM184435.CEL" "GSM184436.CEL" "GSM184437.CEL"
>
```

hogeオブジェクトの中身は実質的に同じ



hogeオブジェクトを有効利用

```
############################################################
###  GSE7623のCELファイルダウンロードとRMA前処理実行  ###
############################################################
out_f <- "data_rma.txt"                  #出力ファイル名を指定してout_fに格納
param <- "GSE7623"                       #入手したいIDを指定

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

#前処理(データ取得)
hoge <- ArrayExpress(param, save=F)      #paramで指定したIDのデータを取得した結果をhogeに格納

#本番
eset <- rma(hoge)

#ファイルに保存
write.exprs(eset, file=out_f)
```

CELファイルのダウンロードと前処理法実行を切り分けることもできます

```
##########################################
###  GSE7623のCELファイルダウンロード  ###
##########################################
param <- "GSE7623"                       #入手したいIDを指定
library(ArrayExpress)                    #パッケージの読み込み
hoge <- ArrayExpress(param, save=F)      #paramで指定したIDのデータを取得した結果をhogeに格納

##########################
###  RMA前処理法実行  ###
##########################
out_f <- "data_rma.txt"                  #出力ファイル名を指定してout_fに格納
library(affy)                            #パッケージの読み込み
eset <- rma(hoge)                        #RMAを実行し、結果をesetに保存
write.exprs(eset, file=out_f)            #結果を指定したファイル名で保存
```

```
##########################################↓
###  GSE7623のCELファイルダウンロード  ###↓
##########################################↓
param <- "GSE7623"                    #入手したいIDを指定↓
library(ArrayExpress)                 #パッケージの読み込み↓
hoge <- ArrayExpress(param, save=F)   #paramで指定したIDのデータを取得した結果をhogeに格納↓
↓
##########################↓
###  RMA前処理法実行  ###↓
##########################↓
out_f <- "data_rma.txt"               #出力ファイル名を指定してout_fに格納↓
library(affy)                         #パッケージの読み込み↓
eset <- rma(hoge)                     #RMAを実行し、結果をesetに保存↓
write.exprs(eset, file=out_f)         #結果を指定したファイル名で保存↓
↓
##########################↓
###  MAS5前処理法実行  ###↓
##########################↓
out_f <- "data_mas.txt"               #出力ファイル名を指定してout_fに格納↓
library(affy)                         #パッケージの読み込み↓
eset <- mas5(hoge)                    #MASを実行し、結果をesetに保存↓
exprs(eset)[exprs(eset) < 1] <- 1     #対数変換（log2）できるようにシグナル強度が1未満のもの
exprs(eset) <- log(exprs(eset), 2)    #底を2として対数変換↓
write.exprs(eset, file=out_f)         #結果を指定したファイル名で保存↓
↓
#####################################↓
###  RMX (RobLoxBioC)前処理法実行  ###↓
#####################################↓
out_f <- "data_rob.txt"               #出力ファイル名を指定してout_fに格納↓
library(RobLoxBioC)                   #パッケージの読み込み↓
eset <- robloxbioc(hoge)              #rmxを実行し、結果をesetに保存↓
exprs(eset)[exprs(eset) < 1] <- 1     #対数変換（log2）できるようにシグナル強度が1未満のもの
exprs(eset) <- log(exprs(eset), 2)    #底を2として対数変換↓
write.exprs(eset, file=out_f)         #結果を指定したファイル名で保存↓
↓
```

rcode_GSE7623_all.txt

CELファイル情報を含むhogeオブジェクトを有効利用して、複数の前処理法を一気に実行

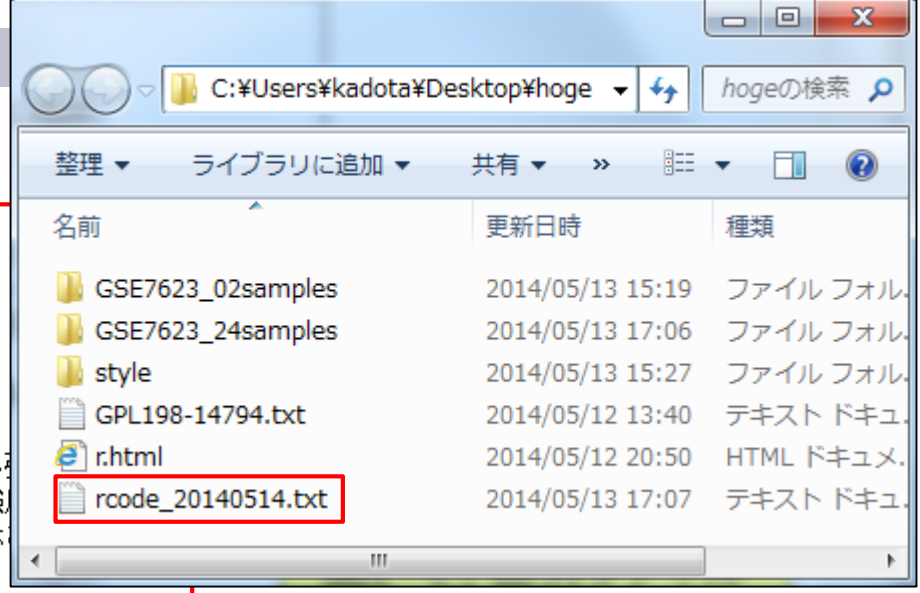
利点：作業ディレクトリがどこであろうが、任意のGSE IDを指定して前処理法実行結果ファイルを得ることができる

門田のやり方

```
###############
###  MAS5  ###
###############
out_f <- "data_mas.txt"              #出力ファイル名を指定してout_fに格納
library(affy)                        #パッケージの読み込み
hoge <- ReadAffy()                   #*.CELファイルの読み込み
eset <- mas5(hoge)                   #MASを実行し、結果をesetに保存
summary(exprs(eset))                 #得られたesetの遺伝子発現行列のシグナル
exprs(eset)[exprs(eset) < 1] <- 1    #対数変換（log2）できるようにシグナル強
summary(exprs(eset))                 #上記処理後のシグナル強度分布を再び表示
exprs(eset) <- log(exprs(eset), 2)   #底を2として対数変換
write.exprs(eset, file=out_f)        #結果を指定したファイル名で保存

###############
###  RMA  ###
###############
out_f <- "data_rma.txt"              #出力ファイル名を指定してout_fに格納
library(affy)                        #パッケージの読み込み
hoge <- ReadAffy()                   #*.CELファイルの読み込み
eset <- rma(hoge)                    #RMAを実行し、結果をesetに保存
write.exprs(eset, file=out_f)        #結果を指定したファイル名で保存

############################
###  RMX (RobLoxBioC)  ###
############################
out_f <- "data_rob.txt"              #出力ファイル名を指定してout_fに格納
library(RobLoxBioC)                  #パッケージの読み込み
hoge <- ReadAffy()                   #*.CELファイルの読み込み
eset <- robloxbioc(hoge)             #rmxを実行し、結果をesetに保存
summary(exprs(eset))                 #得られたesetの遺伝子発現行列のシグナル強度分布を表
exprs(eset)[exprs(eset) < 1] <- 1    #対数変換（log2）できるようにシグナル強度が1未満の
summary(exprs(eset))                 #上記処理後のシグナル強度分布を再び表示させて確認
exprs(eset) <- log(exprs(eset), 2)   #底を2として対数変換
write.exprs(eset, file=out_f)        #結果を指定したファイル名で保存
```



rcode_20140514.txt

2014/05/14のスライド75のコードはただの基本形。これは毎回作業ディレクトリ中のCELファイルの読み込みを行っており非効率

- 正規化 | Affymetrix GeneChip | RMA (Irizarry 2003)
- 正規化 | Affymetrix GeneChip | MAS5.0 (Hubbell 2002)
- 正規化 | Affymetrix GeneChip | rmx (Kohl 2010)

```
#########################################################↓
###  作業ディレクトリ中のCELファイルの読み込み  ###↓
#########################################################↓
library(affy)                          #パッケージの読み込み↓
hoge <- ReadAffy()                     #*.CELファイルの読み込み↓
↓
↓
##########################↓
###  RMA前処理法実行  ###↓
##########################↓
out_f <- "data_rma.txt"                #出力ファイル名を指定してout_fに格納↓
library(affy)                          #パッケージの読み込み↓
eset <- rma(hoge)                      #RMAを実行し、結果をesetに保存↓
write.exprs(eset, file=out_f)          #結果を指定したファイル名で保存↓
↓
##########################↓
###  MAS5前処理法実行  ###↓
##########################↓
out_f <- "data_mas.txt"                #出力ファイル名を指定してout_fに格納↓
library(affy)                          #パッケージの読み込み↓
eset <- mas5(hoge)                     #MASを実行し、結果をesetに保存↓
exprs(eset)[exprs(eset) < 1] <- 1      #対数変換（log2）できるようにシグナル強度が1未満のもの
exprs(eset) <- log(exprs(eset), 2)     #底を2として対数変換↓
write.exprs(eset, file=out_f)          #結果を指定したファイル名で保存↓
↓
#####################################↓
###  RMX (RobLoxBioC)前処理法実行  ###↓
#####################################↓
out_f <- "data_rob.txt"                #出力ファイル名を指定してout_fに格納↓
library(RobLoxBioC)                    #パッケージの読み込み↓
eset <- robloxbioc(hoge)               #rmxを実行し、結果をesetに保存↓
exprs(eset)[exprs(eset) < 1] <- 1      #対数変換（log2）できるようにシグナル強度が1未満のもの
exprs(eset) <- log(exprs(eset), 2)     #底を2として対数変換↓
write.exprs(eset, file=out_f)          #結果を指定したファイル名で保存↓
↓
```

rcode_preprocessing.txt

CELファイル情報を含むhogeオブジェクトを有効利用して、複数の前処理法を一気に実行

利点：CELファイルを含むディレクトリの指定さえすれば、プログラムのどこも変更する必要がない

遺伝子発現行列データは作成済み

- Affymetrix GeneChip
  - Ge et al., *Genomics*, **86**: 127–141, 2005
    - **GSE2361**、GPL96 (Affymetrix Human Genome U133A Array)、22,283 probesets
    - **ヒト36サンプル**：Heart (心臓)、Thymus (胸腺)、Spleen (脾臓)、Ovary (卵巣)、Kidney (腎臓)、Skeletal Muscle (骨格筋)、Pancreas (膵臓)、Prostate (前立腺)、…
  - Nakai et al., *Biosci Biotechnol Biochem.*, **72**: 139–148, 2008
    - **GSE7623**、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
    - **ラット24サンプル**：Brown adipose tissue (褐色脂肪組織; BAT)8サンプル、White adipose tissue (白色脂肪組織; WAT)8サンプル、Liver (肝臓; LIV)8サンプル
      - BAT 8サンプル：通常(BAT_fed) 4サンプル　対　24時間絶食(BAT_fas) 4サンプル
      - WAT 8サンプル：通常(WAT_fed) 4サンプル　対　24時間絶食(WAT_fas) 4サンプル
      - LIV 8サンプル：通常(LIV_fed) 4サンプル　対　24時間絶食(LIV_fas) 4サンプル
  - Kamei et al., *PLoS One*, **8**: e65732, 2013
    - **GSE30533**、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
    - **ラット10サンプル**：全てLiver (肝臓)サンプル
    - iron-deficient diet (Iron_def) 5サンプル　対　control diet (Control) 5サンプル

hogeフォルダ中に3つの前処理法の実行結果ファイルがあります。
MAS5 (data_mas.txt)、RMA (data_rma.txt)、RMX (data_rob.txt)

遺伝子発現行列データは作成済み

- Affymetrix GeneChip
  - Ge et al., *Genomics*, **86**: 127–141, 2005
    - **GSE2361**、GPL96 (Affymetrix Human Genome U133A Array) 22,283 probesets
    - **ヒト36サンプル**：Heart
      臓)、Skeletal Muscle
  - Nakai et al., *Biosci Biot*
    - **GSE7623**、GPL1355 (
    - **ラット24サンプル**：Brov
      tissue (白色脂肪組織;
      - BAT 8サンプル：
      - WAT 8サンプル：
      - LIV 8サンプル：通
  - Kamei et al., *PLoS One*,
    - **GSE30533**、GPL1355
    - **ラット10サンプル**：全て
    - iron-deficient diet (Iro

イントロ | 発現データ取得 | 公共DBから NEW

遺伝子発現(主にマイクロアレイ)データベースをリストアップします。

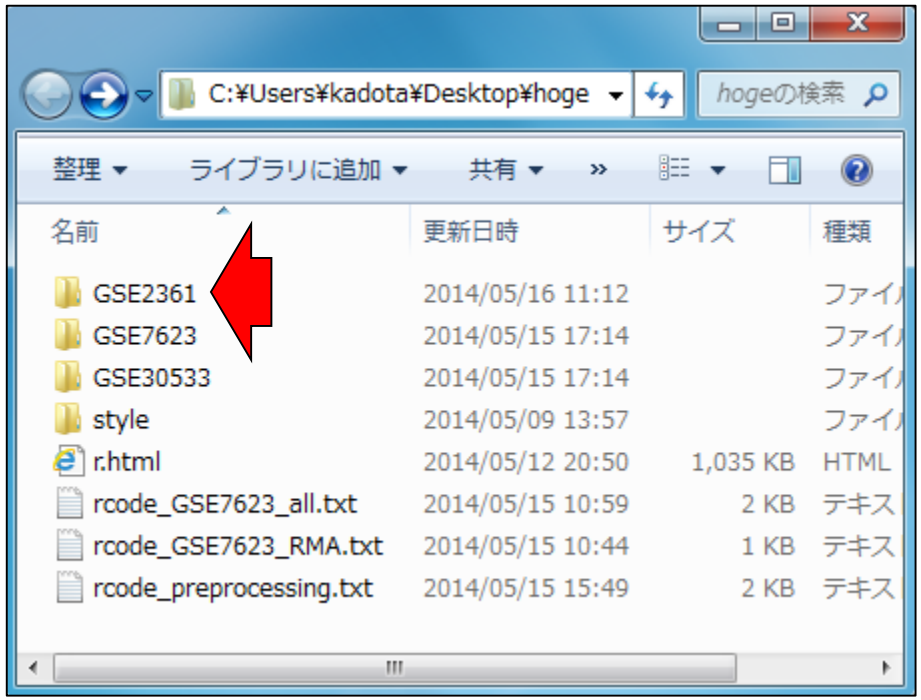
一次データベース

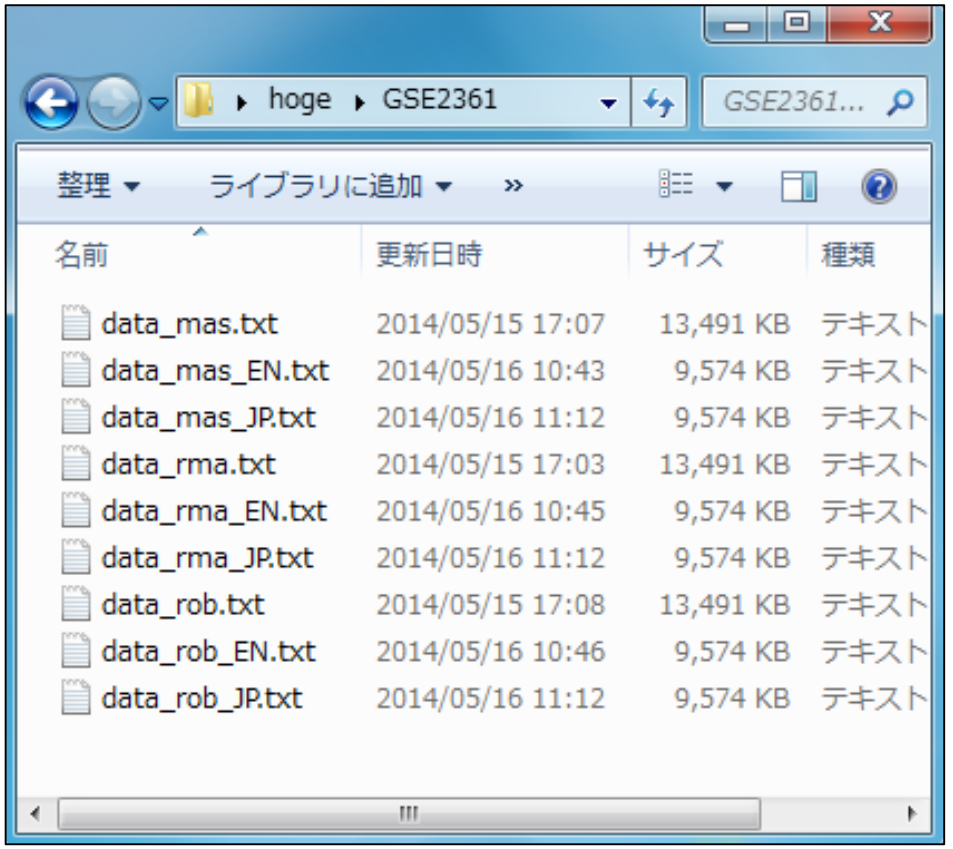
- GEO: Barrett et al., Nucleic Acids Res., 2013
  - GSE7623(ラット24サンプル, 62MB): Nakai et al., BBB, 2008
  - GSE30533(ラット10サンプル, 25MB): Kamei et al., PLoS One, 2013
  - GSE2361(ヒト36サンプル, 130MB): Ge et al., Genomics, 2005
  - GSE10246(マウス182サンプル, 1.1GB): Lattin et al., Immunome Res., 2008
  - GSE1133(ヒトとマウス438サンプル, 1.7GB): Su et al., Proc Natl Acad Sci U S A, 2004
  - GSE5364(ヒト341サンプル, 生データなし): Yu et al., PLoS Genet., 2008
  - GSE15998(マウス106サンプル, 4.0GB): 原著論文はなし?!エクソンアレイ
- ArrayExpress: Rustici et al., Nucleic Acids Res., 2013
  - GSE7623(ラット24サンプル, 62MB): Nakai et al., BBB, 2008
  - GSE30533(ラット10サンプル, 25MB): Kamei et al., PLoS One, 2013
  - GSE2361(ヒト36サンプル, 130MB): Ge et al., Genomics, 2005
  - GSE10246(マウス182サンプル, 1.1GB): Lattin et al., Immunome Res., 2008
  - GSE1133(リンク先なし): Su et al., Proc Natl Acad Sci U S A, 2004
  - GSE5364(ヒト341サンプル, 生データなし): Yu et al., PLoS Genet., 2008
  - GSE15998(マウス106サンプル, 4.0GB): 原著論文はなし?!エクソンアレイ

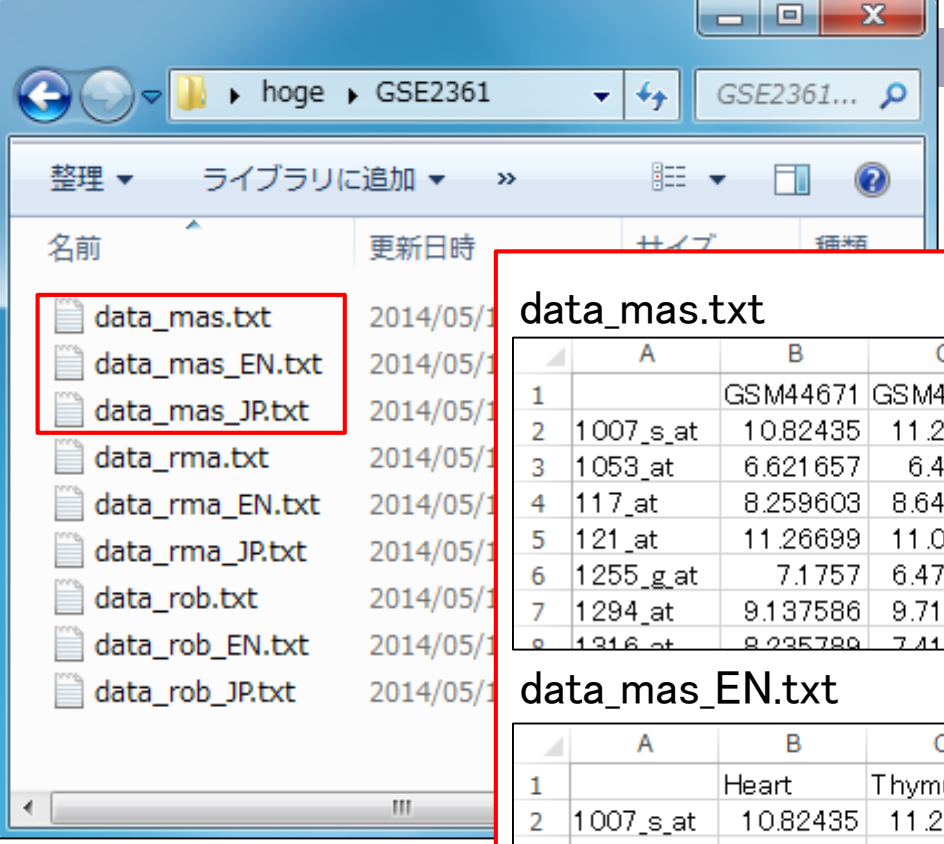
これらのファイルを用いて各種データ解析を行っていきます

遺伝子発現行列データは作成済み

- Affymetrix GeneChip
  - Ge et al., *Genomics*, **86**: 127–141, 2005
    - **GSE2361**、GPL96 (Affymetrix Human Genome U133A Array)、22,283 probesets
    - **ヒト36サンプル**：Heart (心臓)、Thymus (胸腺)、Spleen (脾臓)、Ovary (卵巣)、Kidney (腎臓)、Skeletal Muscle (骨格筋)、Pancreas (膵臓)、Prostate (前立腺)、…







*_EN.txtや*_JP.txtのように入力ファイルの段階で(手作業で)解析結果を見やすくするのが一般的。好きなのをご利用ください。いずれも**対数変換後**のデータです。

data_mas.txt

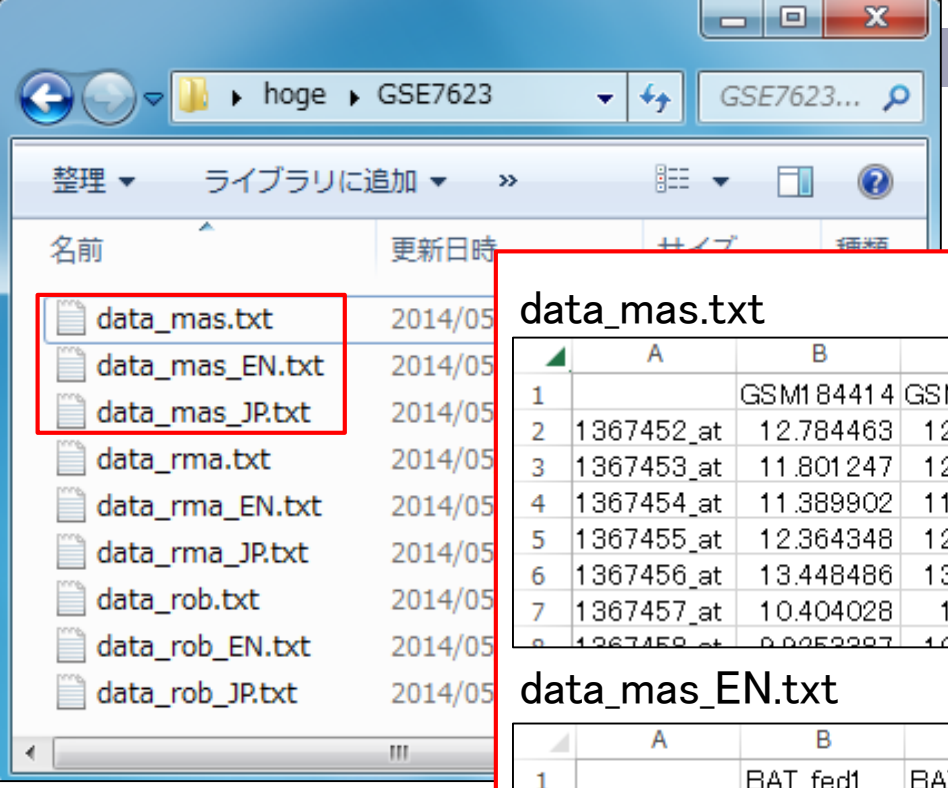
	A	B	C	D	E	F	G	H	I	J
1		GSM44671	GSM44672	GSM44673	GSM44674	GSM44675	GSM44676	GSM44677	GSM44678	GSM4
2	1007_s_at	10.82435	11.21261	9.898072	10.8948	11.75531	10.92032	11.61243	11.22101	11.48
3	1053_at	6.621657	6.47488	7.371465	4.168622	7.350244	7.209918	8.362777	7.176571	7.75
4	117_at	8.259603	8.647364	8.689724	7.875649	5.083001	8.165044	7.96707	8.418082	5.719
5	121_at	11.26699	11.04186	11.39999	11.02855	13.13267	11.39138	12.32899	11.0559	11.28
6	1255_g_at	7.1757	6.477278	6.781766	7.048799	7.30767	6.600267	6.42359	6.694412	7.30
7	1294_at	9.137586	9.718507	9.083742	9.014997	8.813377	8.402562	9.146946	8.421351	10.12
8	1316_at	8.235789	7.418895	8.07462	7.280035	8.238176	7.600147	7.422262	7.399891	7.671

data_mas_EN.txt

	A	B	C	D	E	F	G	H	I	J
1		Heart	Thymus	Spleen	Ovary	Kidney	Skeletal_Mu	Pancreas	Prostate	Small_I
2	1007_s_at	10.82435	11.21261	9.898072	10.8948	11.75531	10.92032	11.61243	11.22101	11.48
3	1053_at	6.621657	6.47488	7.371465	4.168622	7.350244	7.209918	8.362777	7.176571	7.75
4	117_at	8.259603	8.647364	8.689724	7.875649	5.083001	8.165044	7.96707	8.418082	5.719
5	121_at	11.26699	11.04186	11.39999	11.02855	13.13267	11.39138	12.32899	11.0559	11.28
6	1255_g_at	7.1757	6.477278	6.781766	7.048799	7.30767	6.600267	6.42359	6.694412	7.30
7	1294_at	9.137586	9.718507	9.083742	9.014997	8.813377	8.402562	9.146946	8.421351	10.12
8	1316_at	8.235789	7.418895	8.07462	7.280035	8.238176	7.600147	7.422262	7.399891	7.671

data_mas_JP.txt

	A	B	C	D	E	F	G	H	I	J
1		心臓	胸腺	脾臓	卵巣	腎臓	骨格筋	膵臓	前立腺	小腸
2	1007_s_at	10.82435	11.21261	9.898072	10.8948	11.75531	10.92032	11.61243	11.22101	11.48
3	1053_at	6.621657	6.47488	7.371465	4.168622	7.350244	7.209918	8.362777	7.176571	7.75
4	117_at	8.259603	8.647364	8.689724	7.875649	5.083001	8.165044	7.96707	8.418082	5.719
5	121_at	11.26699	11.04186	11.39999	11.02855	13.13267	11.39138	12.32899	11.0559	11.28
6	1255_g_at	7.1757	6.477278	6.781766	7.048799	7.30767	6.600267	6.42359	6.694412	7.30
7	1294_at	9.137586	9.718507	9.083742	9.014997	8.813377	8.402562	9.146946	8.421351	10.12
8	1316_at	8.235789	7.418895	8.07462	7.280035	8.238176	7.600147	7.422262	7.399891	7.671



GSE7623 (Nakai et al., 2008)の対数変換後のデータ

data_mas.txt

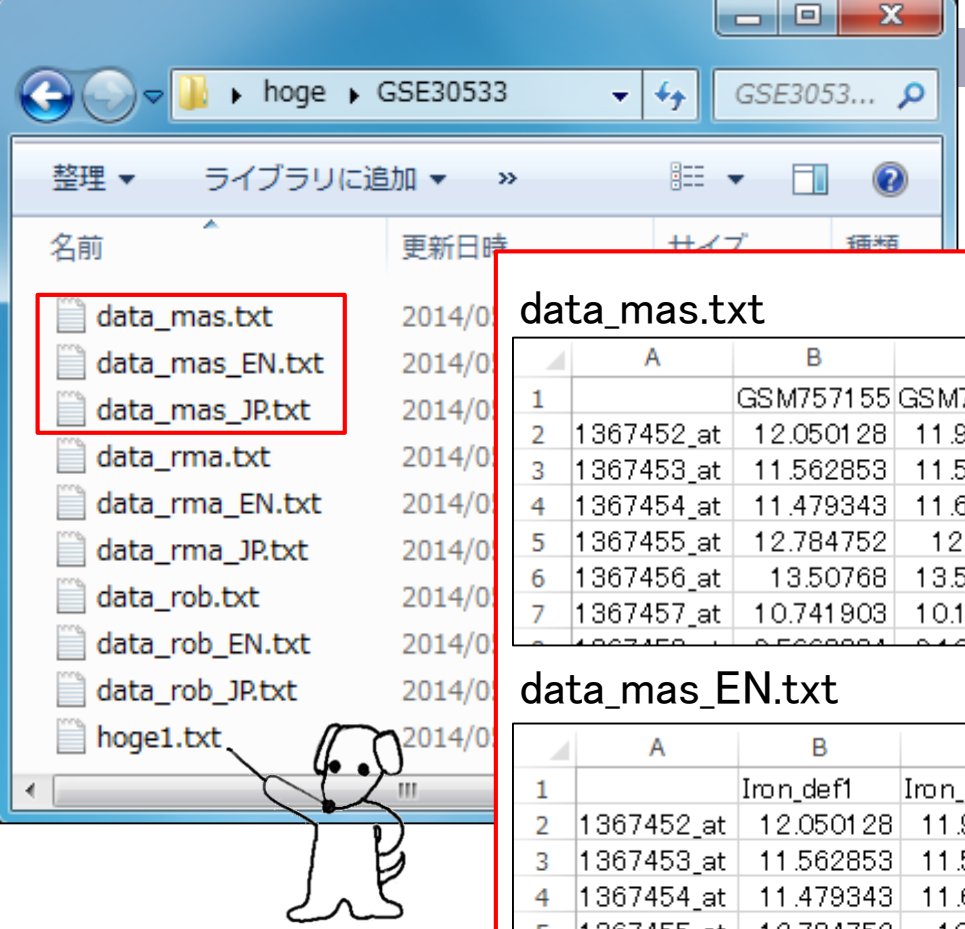
	A	B	C	D	E	F	G	H	I
1		GSM184414	GSM184415	GSM184416	GSM184417	GSM184418	GSM184419	GSM184420	GSM184
2	1367452_at	12.784463	12.447082	12.805908	12.304718	12.589425	12.607532	11.815378	12.439
3	1367453_at	11.801247	12.152935	11.942227	11.968477	11.845375	11.681727	12.078672	12.048
4	1367454_at	11.389902	11.160757	11.145987	11.212088	11.540652	11.308877	11.49885	11.402
5	1367455_at	12.364348	12.529744	12.432574	12.604011	12.441991	12.249935	12.281827	12.190
6	1367456_at	13.448486	13.543046	13.552794	13.629799	13.36913	13.244278	13.424371	13.329
7	1367457_at	10.404028	10.69632	10.475078	10.45579	10.141921	10.290666	10.146529	10.26
8	1367458_at	9.9253387	10.244544	9.9720008	9.9576072	8.702884	9.3578792	9.2134367	9.4998

data_mas_EN.txt

	A	B	C	D	E	F	G	H	I
1		BAT_fed1	BAT_fed2	BAT_fed3	BAT_fed4	BAT_fas1	BAT_fas2	BAT_fas3	BAT_fas
2	1367452_at	12.784463	12.447082	12.805908	12.304718	12.589425	12.607532	11.815378	12.439
3	1367453_at	11.801247	12.152935	11.942227	11.968477	11.845375	11.681727	12.078672	12.048
4	1367454_at	11.389902	11.160757	11.145987	11.212088	11.540652	11.308877	11.49885	11.402
5	1367455_at	12.364348	12.529744	12.432574	12.604011	12.441991	12.249935	12.281827	12.190
6	1367456_at	13.448486	13.543046	13.552794	13.629799	13.36913	13.244278	13.424371	13.329
7	1367457_at	10.404028	10.69632	10.475078	10.45579	10.141921	10.290666	10.146529	10.26
8	1367458_at	9.9253387	10.244544	9.9720008	9.9576072	8.702884	9.3578792	9.2134367	9.4998

data_mas_JP.txt

	A	B	C	D	E	F	G
1		褐色脂肪_満腹1	褐色脂肪_満腹2	褐色脂肪_満腹3	褐色脂肪_満腹4	褐色脂肪_空腹1	褐色脂肪_空腹
2	1367452_at	12.7844634	12.44708219	12.80590758	12.30471769	12.58942538	12.6075319
3	1367453_at	11.80124704	12.15293493	11.94222741	11.96847729	11.84537542	11.6817274
4	1367454_at	11.38990178	11.16075717	11.14598707	11.21208786	11.54065185	11.3088766
5	1367455_at	12.36434768	12.52974368	12.43257392	12.60401124	12.44199125	12.2499348
6	1367456_at	13.44848649	13.54304603	13.55279359	13.62979898	13.36912977	13.2442783
7	1367457_at	10.40402803	10.69631952	10.47507777	10.4557902	10.14192076	10.2906657
8	1367458_at	9.9053387	10.244544	9.9720008	9.9576072	8.702884	9.3578792



GSE30533 (Kamei et al., 2013)の**対数変換後の**データ。教科書中で用いているデータセットです。

data_mas.txt

	A	B	C	D	E	F	G	H	I
1		GSM757155	GSM757156	GSM757157	GSM757158	GSM757159	GSM757160	GSM757161	GSM7571
2	1367452_at	12.050128	11.923417	11.991814	11.921645	11.728346	12.079674	12.058299	11.979
3	1367453_at	11.562853	11.587924	11.62175	11.747779	11.776342	11.633797	11.50868	11.4794
4	1367454_at	11.479343	11.676545	11.608875	11.652704	11.86008	11.70589	11.975747	12.0095
5	1367455_at	12.784752	12.58787	12.696588	12.789029	13.002298	12.678645	12.780148	12.5522
6	1367456_at	13.50768	13.533852	13.483843	13.524296	13.452217	13.472369	13.594191	13.603
7	1367457_at	10.741903	10.136369	10.609696	10.256467	10.312411	10.49749	10.304808	10.4274

data_mas_EN.txt

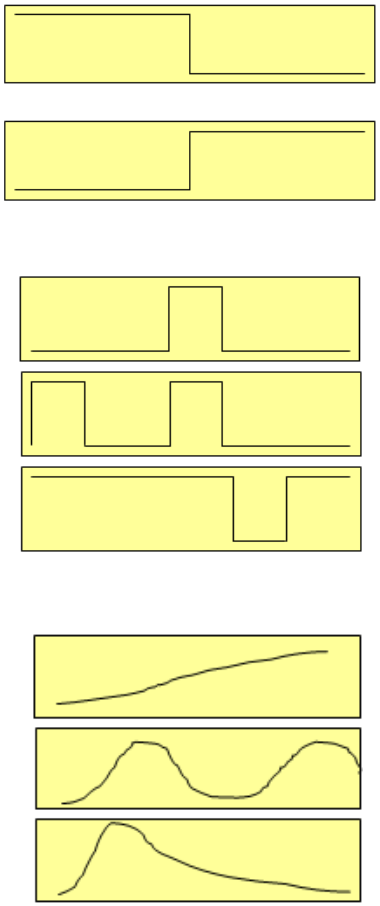
	A	B	C	D	E	F	G	H	I
1		Iron_def1	Iron_def2	Iron_def3	Iron_def4	Iron_def5	Control1	Control2	Control3
2	1367452_at	12.050128	11.923417	11.991814	11.921645	11.728346	12.079674	12.058299	11.979
3	1367453_at	11.562853	11.587924	11.62175	11.747779	11.776342	11.633797	11.50868	11.4794
4	1367454_at	11.479343	11.676545	11.608875	11.652704	11.86008	11.70589	11.975747	12.0095
5	1367455_at	12.784752	12.58787	12.696588	12.789029	13.002298	12.678645	12.780148	12.5522
6	1367456_at	13.50768	13.533852	13.483843	13.524296	13.452217	13.472369	13.594191	13.603
7	1367457_at	10.741903	10.136369	10.609696	10.256467	10.312411	10.49749	10.304808	10.4274

data_mas_JP.txt

	A	B	C	D	E	F	G	H	I
1		鉄欠乏1	鉄欠乏2	鉄欠乏3	鉄欠乏4	鉄欠乏5	通常1	通常2	通常3
2	1367452_at	12.050128	11.923417	11.991814	11.921645	11.728346	12.079674	12.058299	11.979
3	1367453_at	11.562853	11.587924	11.62175	11.747779	11.776342	11.633797	11.50868	11.4794
4	1367454_at	11.479343	11.676545	11.608875	11.652704	11.86008	11.70589	11.975747	12.0095
5	1367455_at	12.784752	12.58787	12.696588	12.789029	13.002298	12.678645	12.780148	12.5522
6	1367456_at	13.50768	13.533852	13.483843	13.524296	13.452217	13.472369	13.594191	13.603
7	1367457_at	10.741903	10.136369	10.609696	10.256467	10.312411	10.49749	10.304808	10.4274

データ解析もいろいろ

発現変動遺伝子同定



遺伝子発現行列

二群間比較用

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

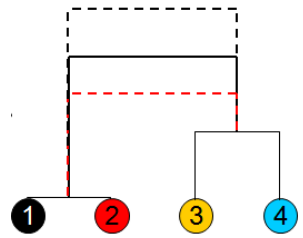
様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

時系列データ

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

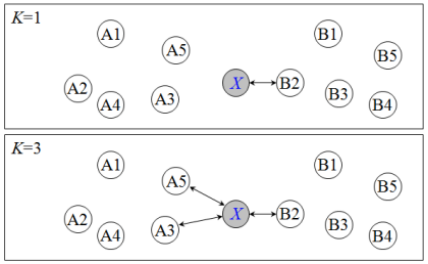
クラスタリング



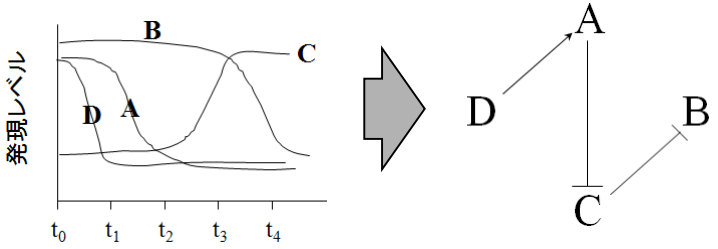
機能解析

- Gene Ontology（GO）
- パスウェイ解析

分類（診断）



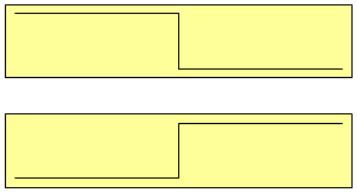
遺伝子ネットワーク推定



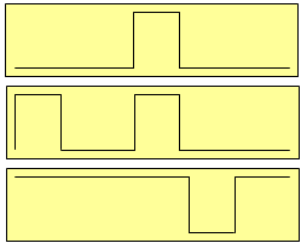
バイオインフォマティクス要素技術

- 相関係数やエントロピーなどの応用例を紹介

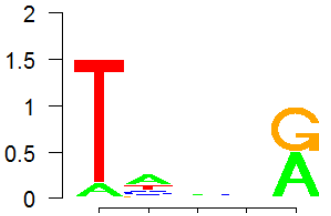
二群間比較



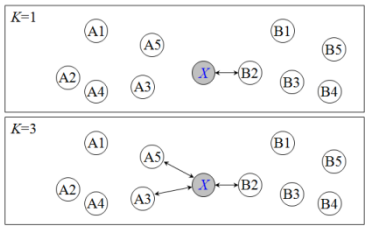
組織特異的遺伝子



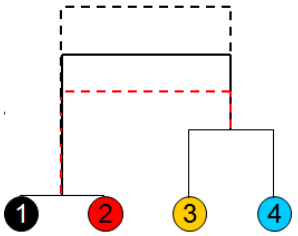
Sequence logo



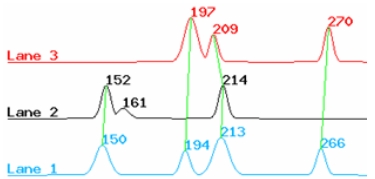
分類（診断）



クラスタリング



同一ピーク同定



基本スキルのみでいろいろなことができます

第3章　データ解析（基礎）

- § 3.2.1 クラスタリング（データ変換や距離の定義など）

- 書籍 | トランスクリプトーム解析 | 1.1 はじめに (last modified 2014/05/09) NEW
- 書籍 | トランスクリプトーム解析 | 1.2.1 原理(Affymetrix 3'発現アレイ) (last modified 2014/05/12) NEW
- 書籍 | トランスクリプトーム解析 | 1.2.2 最近の知見 (last modified 2014/05/
- 書籍 | トランスクリプトーム解析 | 2.2.1 生データ(プローブレベルデータ)取
- 書籍 | トランスクリプトーム解析 | 2.2.2 データの正規化(基礎) (last modifie
- 書籍 | トランスクリプトーム解析 | 2.2.3 データの正規化(計算例) (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.4 データの正規化(その他) (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.5 アノテーション情報 (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.1 クラスタリング(データ変換や距離の定義など) (last modified 2014/05/16) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.2 実験デザイン、データ分布、統計解析との関係 (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.3 多重比較問題 (last modified 2014/04/19) NEW
- 書籍 | トランスク
- 書籍 | トランスク
- 書籍 | トランスク
- 書籍 | トランスク

教科書中で用いているhoge1.txtは、GSE30533 (Kamei et al., 2013)の**対数変換（$\log_2$変換）前の**MAS5データ

書籍 | トランスクリプトーム解析 | 3.2.1 クラスタリング(データ変換や距離の定義など) NEW

シリーズ Useful R 第7巻トランスクリプトーム解析のp99-107のRコードです。
「ファイル」-「ディレクトリの変更」でデスクトップ上の"E-GEOD-30533.raw.1"など任意のディレクトリに移動し以下をコピペ。

p99の網掛け部分:

書籍中では作業ディレクトリがデスクトップ上の"E-GEOD-30533.raw.1"という前提になっていますが、p40で作成したMAS5データファイル(hoge1.txt)を置いてあるディレクトリであればどこでも構いません。

```
in_f <- "hoge1.txt"
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))
data.dist <- as.dist(1 - cor(data, method = "spearman"))
out <- hclust(data.dist, method = "average")
plot(out)                                    # 図3-1作成部分
```

第3章　データ解析（基礎）

- 書籍 | トランスクリプトーム解析 | 1.1 はじめに (last modified 2014/05/09) NEW
- 書籍 | トランスクリプトーム解析 | 1.2.1 原理(Affymetrix 3'発現アレイ) (last modified 2014/
- 書籍 | トランスクリプトーム解析 | 1.2.2 最近の知見 (last modified 2014/05/09) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.1 生データ(プローブレベルデータ)取得 (last modified
- 書籍 | トランスクリプトーム解析 | 2.2.2 データの正規化(基礎) (last modifie
- 書籍 | トランスクリプトーム解析 | 2.2.3 データの正規化(計算例) (last modif
- 書籍 | トランスクリプトーム解析 | 2.2.4 データの正規化(その他) (last modif
- 書籍 | トランスクリプトーム解析 | 2.2.5 アノテーション情報 (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.1 クラスタリング(データ変換や距離の定義など) (last
- 書籍 | トランスクリプトーム解析 | 3.2.2 実験デザイン, データ分布, 統計解析との関係 (last
- 書籍 | トラン

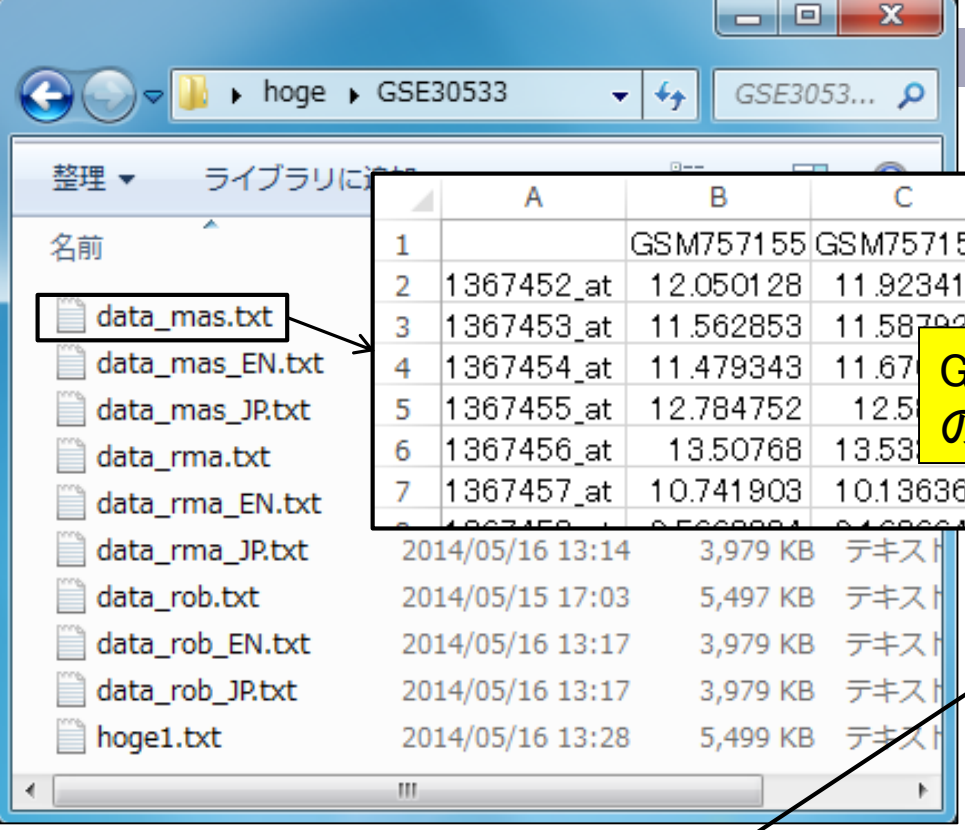
教科書中で用いているhoge1.txtは、GSE30533 (Kamei et al., 2013)の**対数変換（log_2変換）前のMAS5データ**

p40の網掛け部分(上):

hoge1.txtと同じものができていると思います。

```
out_f <- "hoge1.txt"                    #出力ファイル名を指定してout_fに格納
library(affy)                           #パッケージの読み込み
hoge <- ReadAffy()                      #*.CELファイルの読み込み
eset <- mas5(hoge)                      #MAS5を実行し、結果をesetに保存
write.exprs(eset, file=out_f)           #結果をout_fで指定したファイル名で保存
```

	A	B	C	D	E	F	G	H	I
1		GSM757155	GSM757156	GSM757157	GSM757158	GSM757159	GSM757160	GSM757161	GSM7571
2	1367452_at	4240.8	3884.2	4072.8	3879.5	3393.0	4328.6	4264.9	4039
3	1367453_at	3025.3	3078.3	3151.3	3439.0	3507.8	3177.8	2913.8	2855
4	1367454_at	2855.1	3273.3	3123.3	3219.7	3717.4	3340.6	4027.7	4123
5	1367455_at	7056.6	6156.4	6638.3	7077.5	8205.1	6556.2	7034.1	6006
6	1367456_at	11647.1	11860.3	11456.2	11782.0	11207.8	11365.5	12366.9	12449
7	1367457_at	1712.5	1125.5	1562.6	1223.2	1271.6	1445.6	1264.9	1377



	A	B	C	D	E	F	G	H	I
1		GSM757155	GSM757156	GSM757157	GSM757158	GSM757159	GSM757160	GSM757161	GSM7571
2	1367452_at	12.050128	11.923417	11.991814	11.921645	11.728346	12.079674	12.058299	11.979
3	1367453_at	11.562853	11.587924	11.621175	11.747779	11.776342	11.633797	11.50868	11.4794
4	1367454_at	11.479343	11.67				11.70589	11.975747	12.0095
5	1367455_at	12.784752	12.5				12.678645	12.780148	12.5522
6	1367456_at	13.50768	13.53				13.472369	13.594191	13.603
7	1367457_at	10.741903	10.136369	10.609696	10.256467	10.312411	10.49749	10.304808	10.4274

GSE30533 (Kamei et al., 2013)
の対数変換後のMAS5データ

```
##################################################↓
###  CELファイルの読み込みとMAS5前処理法実行  ###↓
##################################################↓
out_f <- "data_mas.txt"              #出力ファイル名を指定してout_fに格納↓
library(affy)                        #パッケージの読み込み↓
hoge <- ReadAffy()                   #*.CELファイルの読み込み↓
eset <- mas5(hoge)                   #MAS5を実行し、結果をesetに保存↓
exprs(eset)[exprs(eset) < 1] <- 1    #対数変換（log2）できるようにシグナル強度が1未満のものを1にしておく
exprs(eset) <- log(exprs(eset), 2)   #底を2として対数変換↓
write.exprs(eset, file=out_f)        #結果を指定したファイル名で保存↓
↓
```

対数変換部分

§ 3.2.1 クラスタリング

- 書籍 | トランスクリプトーム解析 | 2.2.2 データの正規化(基礎) (last modified 2014/04/17) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.3 データの正規化(計算例) (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.4 データの正規化(その他) (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 2.2.5 アノテーション情報 (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.1 クラスタリング(データ変換や距離の定義など) modified 2014/05/16) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.2 実験デザイン、データ分布、統計解析との関係 (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | 3.2.3 多重比較問題 (last modified 2014/04/19) NEW
- 書籍 | トランスク
- 書籍 | トランスク
- 書籍 | トランスク
- 書籍 | トランスク

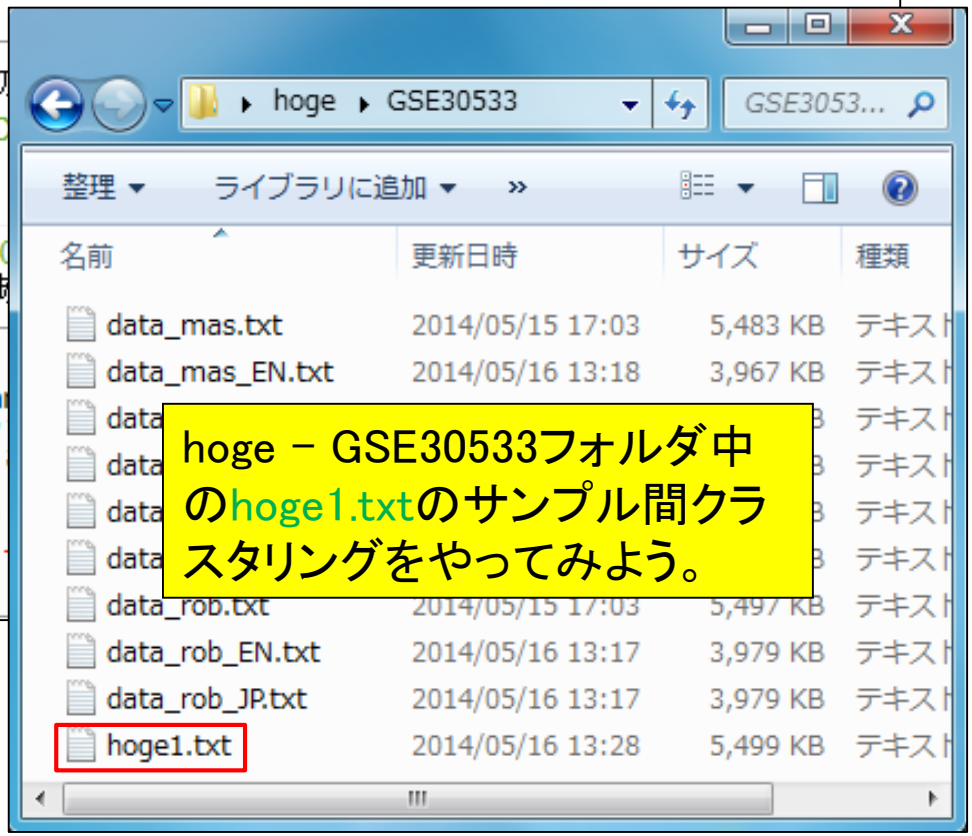
書籍 | トランスクリプトーム解析 | 3.2.1 クラスタリング(データ変換や距離の定義など) NEW

シリーズ Useful R 第7巻トランスクリプトーム解析のp99-107の
「ファイル」−「ディレクトリの変更」でデスクトップ上の"E-GEOD

p99の網掛け部分:

書籍中では作業ディレクトリがデスクトップ上の"E-GEOD-3
MAS5データファイル(hoge1.txt)を置いてあるディレクトリであ

```
in_f <- "hoge1.txt"
data <- read.table(in_f, header=TRUE, row.na
colnames(data) <- c(paste("G1_", 1:5, sep=""
data.dist <- as.dist(1 - cor(data, method =
out <- hclust(data.dist, method = "average")
plot(out)                                  # 図3-
```



hoge − GSE30533フォルダ中のhoge1.txtのサンプル間クラスタリングをやってみよう。

書籍 | トラン
ど) NEW

シリーズ Useful R 第

「ファイル」-「ディレク

p99の網掛け部分:

書籍中では作業デ
MAS5データファイル(hoge1.txt)と置いてあるフォルダにセットしてから実行してください。

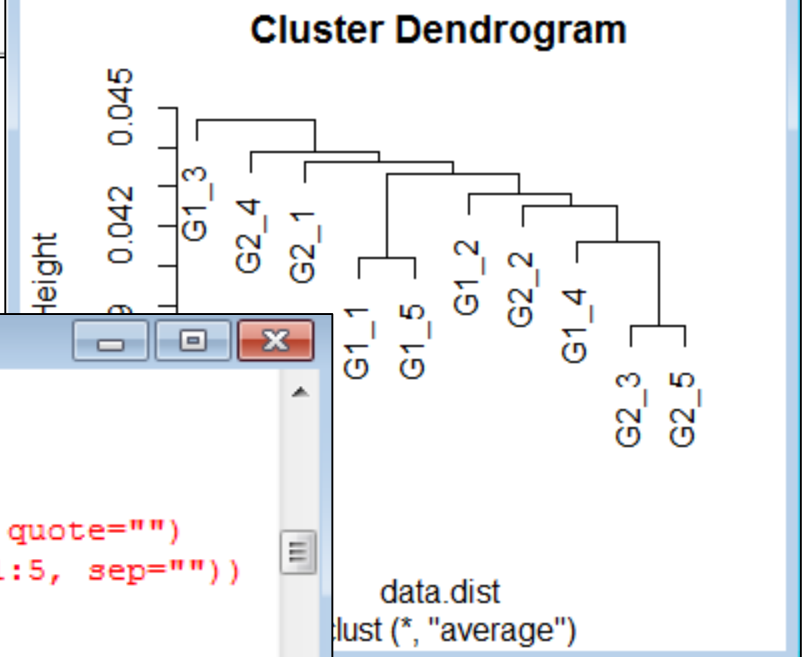
	A	B	C	D	E	F	G	H	I
1		GSM757155	GSM757156	GSM757157	GSM757158	GSM757159	GSM757160	GSM757161	GSM7571
2	1367452_at	4240.8	3884.2	4072.8	3879.5	3393.0	4328.6	4264.9	4039
3	1367453_at	3025.3	3078.3	3151.3	3439.0	3507.8	3177.8	2913.8	2855
4	1367454_at	2855.1	3273.3	3123.3				7.7	4123
5	1367455_at	7056.6	6156.4	6638.3				4.1	6006
6	1367456_at	11647.1	11860.3	11456.2				6.9	12449
7	1367457_at	1712.5	1125.5	1562.6	1223.2	1271.6	1445.6	1264.9	1377
8	1367458_at	758.1	575.5	568.6	494.5	681.0	610.6	443.0	565

GSE30533 (Kamei et al., 2013)
の対数変換前のMAS5データ

```
in_f <- "hoge1.txt"
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))
data.dist <- as.dist(1 - cor(data, method = "spearman"))
out <- hclust(data.dist, method = "average")
plot(out)                                    # 図3-1作成部分
```

なぜサンプル名のところが変
わっているのかを説明します

R Graphics: Device 2 (ACTIVE)



R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/GSE30533"
> in_f <- "hoge1.txt"
> data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
> colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))
> data.dist <- as.dist(1 - cor(data, method = "spearman"))
> out <- hclust(data.dist, method = "average")
> plot(out)                                    # 図3-1作成部分
```

p99の網掛け部分:

書籍中では作業ディレクトリがデスクトップ上の"E-GEOD-30533.raw.1"という前提になっていますが、p40で作成したMAS5データファイル(hoge1.txt)を置いてあるディレクトリであればどこでも構いません。

```
in_f <- "hoge1.txt"
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))
data.dist <- as.dist(1 - cor(data, method = "spearman"))
out <- hclust(data.dist, method = "average")
plot(out)                                    # 図3-1作成部分
```

サンプル名部分に相当するcolnames(data)に任意の文字列ベクトルを代入しています

R Console

```
> data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
> colnames(data)
 [1] "GSM757155_.Fe_short_27.CEL" "GSM757156_.Fe_short_31.CEL"
 [3] "GSM757157_.Fe_short_33.CEL" "GSM757158_.Fe_short_35.CEL"
 [5] "GSM757159_.Fe_short_37.CEL" "GSM757160_control_28.CEL"
 [7] "GSM757161_control_30.CEL"   "GSM757162_control_32.CEL"
 [9] "GSM757163_control_34.CEL"   "GSM757164_control_36.CEL"
> colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))
> colnames(data)
 [1] "G1_1" "G1_2" "G1_3" "G1_4" "G1_5" "G2_1" "G2_2" "G2_3" "G2_4" "G2_5"
> paste("G1_", 1:5, sep="")
[1] "G1_1" "G1_2" "G1_3" "G1_4" "G1_5"
> paste("G1", 1:5, sep="_")
[1] "G1_1" "G1_2" "G1_3" "G1_4" "G1_5"
> paste("Iron_def", 1:5, sep="")
[1] "Iron_def1" "Iron_def2" "Iron_def3" "Iron_def4" "Iron_def5"
>
```

上矢印キーを押すと直前に打ったコマンドが出る。有効に利用し最小限の労力で打つ。

§ 3.2.1 クラスタリング

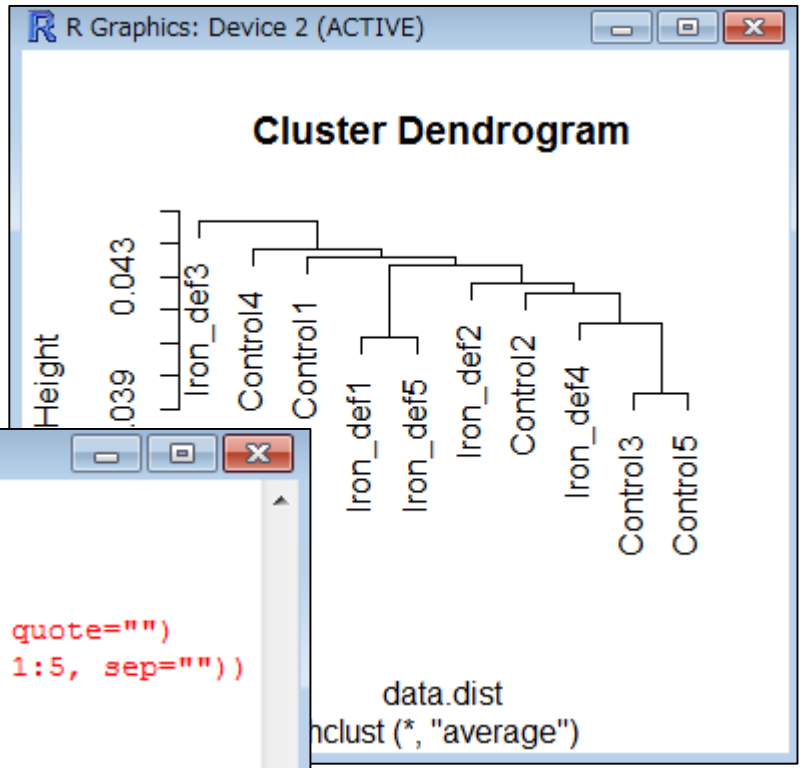
	A	B	C	D	E	F	G	H	I
1		Iron_def1	Iron_def2	Iron_def3	Iron_def4	Iron_def5	Control1	Control2	Control3
2	1367452_at	12.050128	11.923417	11.991814	11.921645	11.728346	12.079674	12.058299	11.979
3	1367453_at	11.562853	11.587924	11.				.50868	11.4794
4	1367454_at	11.479343	11.676545	11.				975747	12.0095
5	1367455_at	12.784752	12.58787	12.				780148	12.5522
6	1367456_at	13.50768	13.533852	13.				594191	13.603
7	1367457_at	10.741903	10.136369	10.609696	10.256467	10.312411	10.49749	10.304808	10.4274

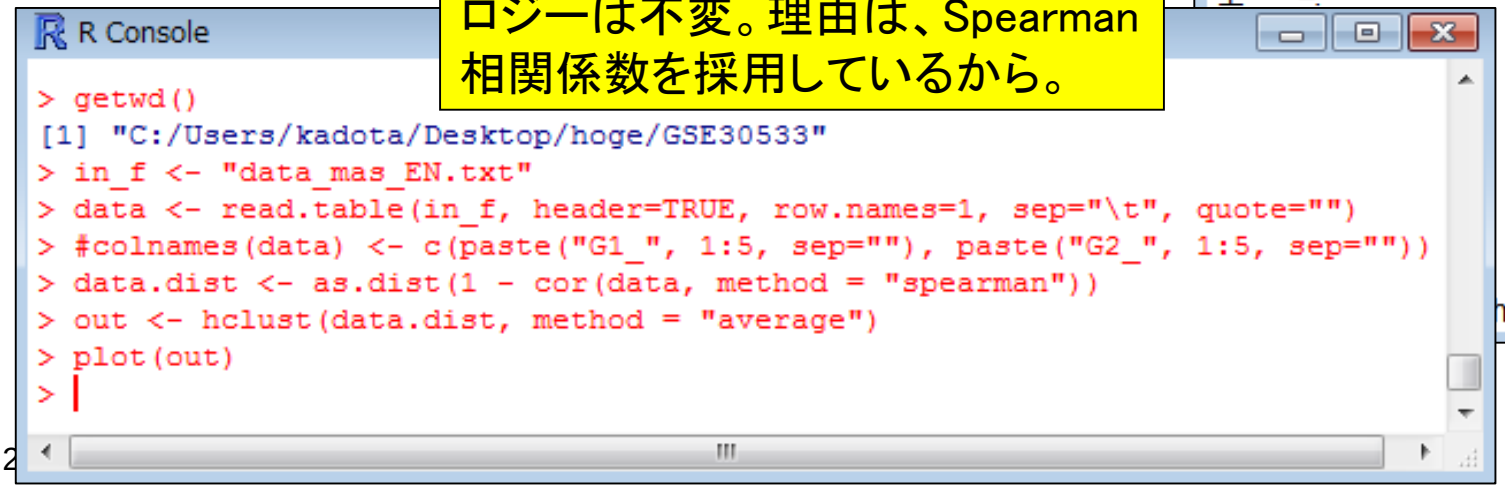
GSE30533 (Kamei et al., 2013) の対数変換後のMAS5データ

rcode_clustering.txt

```
#####################################↓
###  MAS5データのクラスタリング  ###↓
#####################################↓
in_f <- "data_mas_EN.txt"↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")↓
#colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))↓
data.dist <- as.dist(1 - cor(data, method = "spearman"))↓
out <- hclust(data.dist, method = "average")↓
plot(out)↓
↓
```

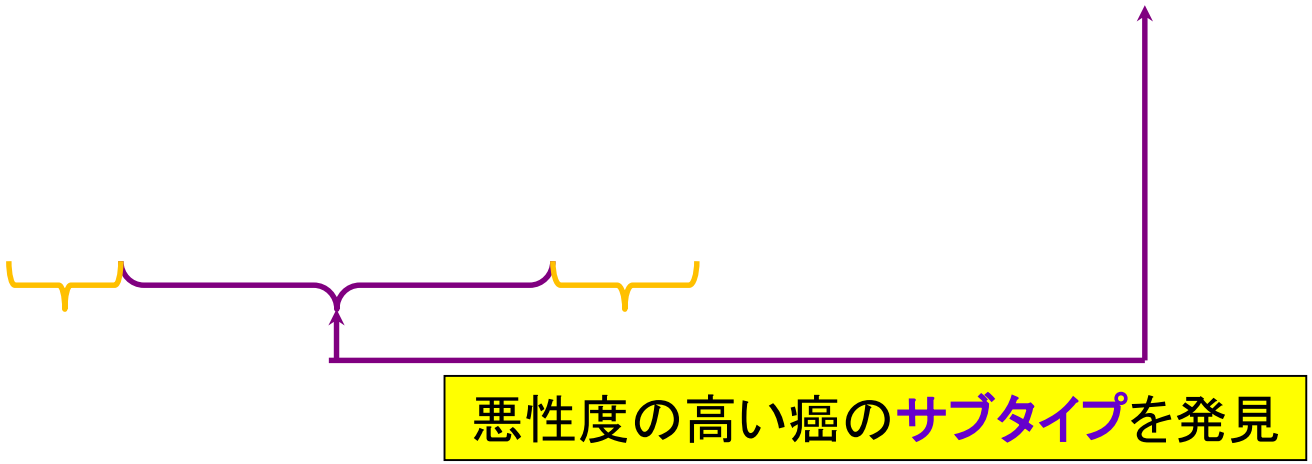
対数変換の有無にかかわらずクラスタリング結果(樹形図)のトポロジーは不変。理由は、Spearman相関係数を採用しているから。





（サンプル間）クラスタリングの実例

- 悪性黒色腫（メラノーマ）31サンプルのデータ

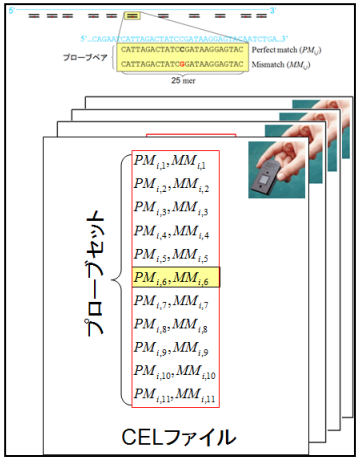


クラスタリング（教師なし学習）

- 階層的クラスタリング
  - 発現パターンの類似した遺伝子を集めて系統樹を作成
- 非階層的（分割最適化）クラスタリング
  - K-meansクラスタリング
    - 「K個のクラスターに分割（Kの数は主観的に決定）する」と予め指定し、各クラスター内の遺伝子（サンプル）間の距離の総和が最小になるようなK個のクラスターを作成
  - 自己組織化マップ（SOM）
  - 主成分分析（PCA）

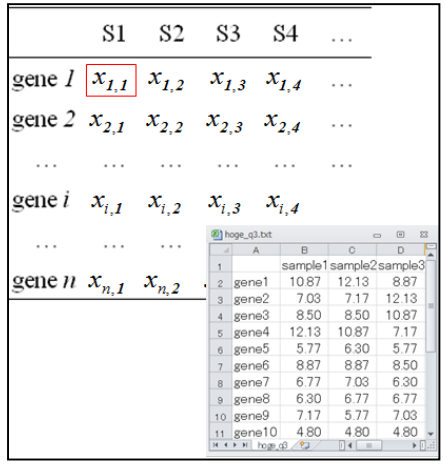
どれを使うかはほぼ趣味の問題です

得られる樹形図の可能性は無数



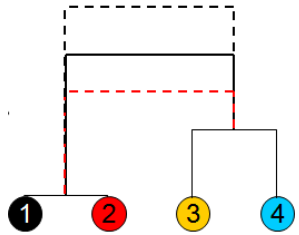
生データファイル

前処理法



遺伝子発現行列

クラスタリング



①前処理法	②スケーリング	③距離	④群の併合
・MAS	・**対数変換**	・**1－相関係数**	・単連結法
・RMA	・相対値(0~1)	・ユークリッド	・完全連結法
・RobLoxBioC	・Z-score化	・マンハッタン	・**平均連結法**
…	…	・キャンベラ	・ウォード法
…	…	…	…

×　×　×

クラスタリング（教師なし学習）

- 決めておくべき二つの基準（事柄）
  - 距離（類似度）の定義
    - ユークリッド距離、マンハッタン距離など
  - クラスターをまとめる（併合する）方法
    - クラスター間の距離を定義する方法、とほぼ同じ
    - 最短距離法、平均連結法、ウォード法など

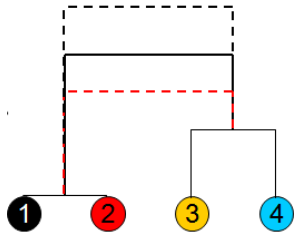
得られた結果の妥当性を何らかの知見に基づいて評価するため、結果の正当性を主張する視点が複数存在しうる。
→見当をつける、程度に利用すべし

入力例

	x^1	x^2	x^3	x^4
Sample 1	38	42	204	204
Sample 2	0	3	182	182
Sample 3	6	6	19	168
Sample 4	141	106	11	11
Sample 5	8	5	79	179
Sample 6	16	20	96	126
Sample 7	2	5	164	164
Sample 8	132	101	222	0
Sample 9	7	9	164	164
Sample 10	2	8	16	116
Sample 11	66	79	195	195
Sample 12	8	9	241	241
Sample 13	6	8	177	177

クラスタリング

出力例



距離（類似度）の定義

- ベクトルxとyの発現パターンの距離$D(x,y)$

$$\text{相関係数 } r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (-1 \le r_{xy} \le 1)$$

i	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
…	…	…
n	x_n	y_n

- xとyの発現パターンが酷似 → $r \approx 1$
- xとyの発現パターンがばらばら → $r \approx 0$
- xとyの発現パターンがほぼ正反対 → $r \approx -1$

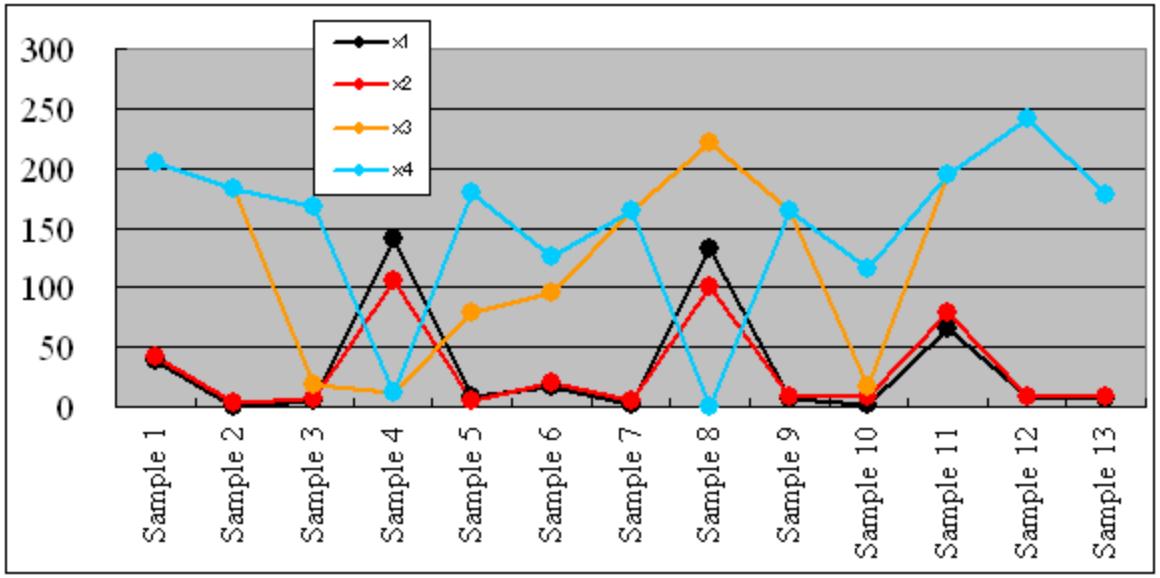
$$\text{距離 } D(x,y) = 1 - r \quad (0 \le D \le 2)$$

- $r = 1 \rightarrow D = 1-1 = 0$
- $r = 0 \rightarrow D = 1-0 = 1$
- $r = -1 \rightarrow D = 1-(-1) = 2$

「1－相関係数」を距離として定義することができます

相関係数 → 距離(計算例)

	x^1	x^2	x^3	x^4
Sample 1	38	42	204	204
Sample 2	0	3	182	182
Sample 3	6	6	19	168
Sample 4	141	106	11	11
Sample 5	8	5	79	179
Sample 6	16	20	96	126
Sample 7	2	5	164	164
Sample 8	132	101	222	0
Sample 9	7	9	164	164
Sample 10	2	8	16	116
Sample 11	66	79	195	195
Sample 12	8	9	241	241
Sample 13	6	8	177	177



相関係数 $r_{x^1x^2} = 0.98 \rightarrow$ 距離 $D_{x^1x^2} = 1 - 0.98 = 0.02$

相関係数 $r_{x^1x^3} = -0.01 \rightarrow$ 距離 $D_{x^1x^3} = 1 - (-0.01) = 1.01$

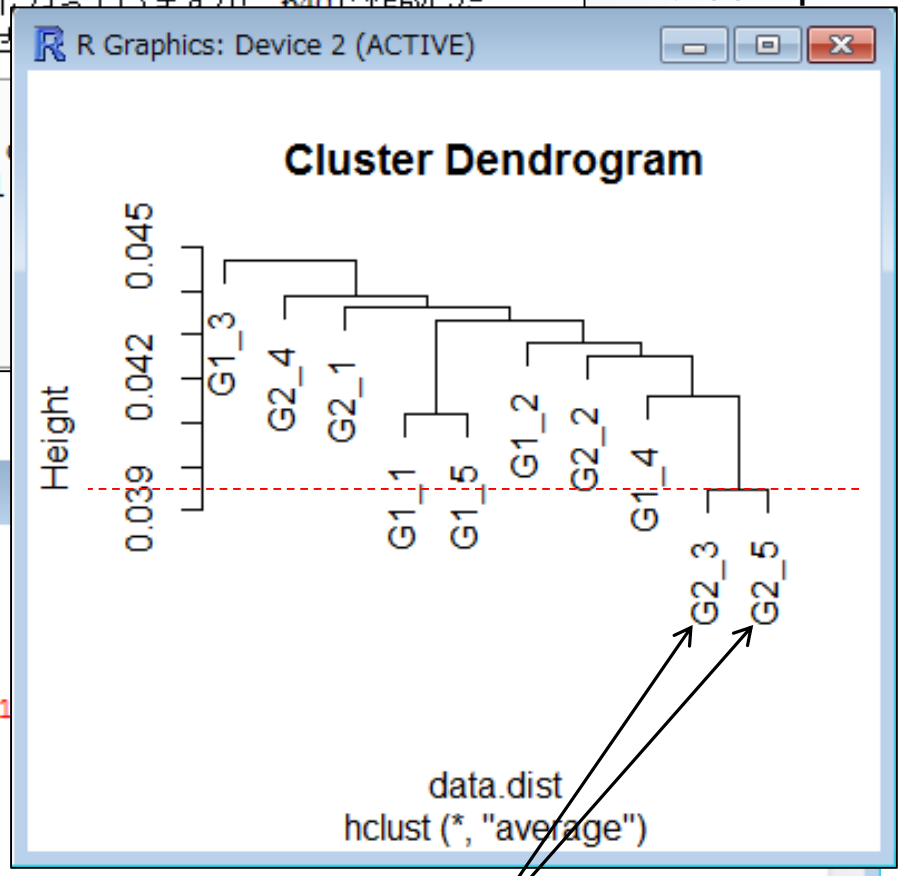
相関係数 $r_{x^1x^4} = -0.78 \rightarrow$ 距離 $D_{x^1x^4} = 1 - (-0.78) = 1.78$

教科書p99

p99の網掛け部分:

書籍中では作業ディレクトリがデスクトップ上の"E-GEOD-30533.raw.1"という前提になっていますが、p40で作成したMAS5データファイル(hoge1.txt)を置いてあるディレクトリであればどこでも構いません

```
in_f <- "hoge1.txt"
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t",
colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1
data.dist <- as.dist(1 - cor(data, method = "spearman"))
out <- hclust(data.dist, method = "average")
plot(out)                                    # 図3-1作成部分
```

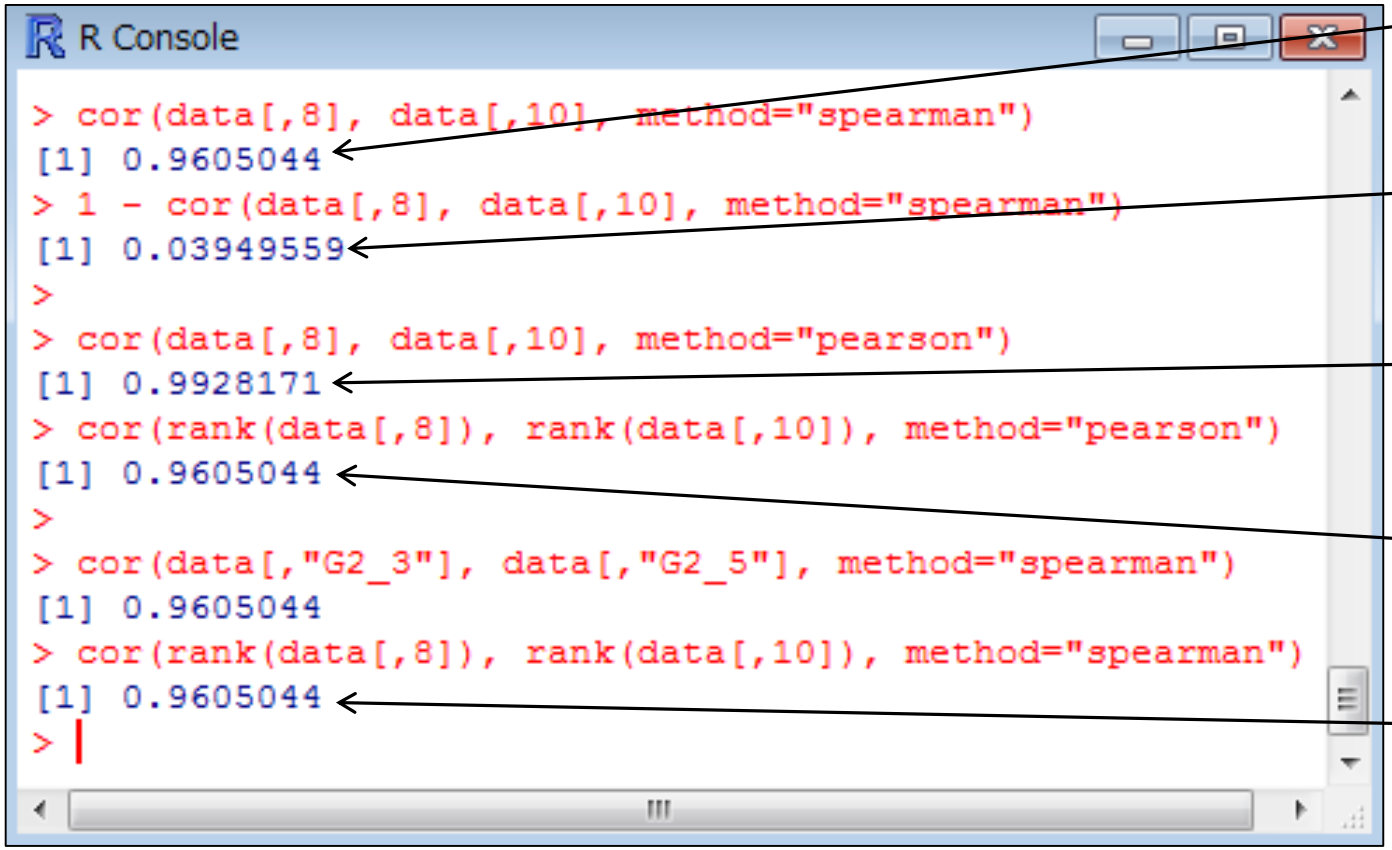


R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/GSE30533"
> in_f <- "hoge1.txt"
> data <- read.table(in_f, header=TRUE, row.names=1, sep="\t",
> colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1
> data.dist <- as.dist(1 - cor(data, method = "spearman"))
> out <- hclust(data.dist, method = "average")
> plot(out)                                    # 図3-1作成部分
> data.dist
           G1_1       G1_2       G1_3       G1_4       G1_5       G2_1       G2_2       G2_3       G2_4
G1_2 0.04533075
G1_3 0.04530167 0.04403145
G1_4 0.04298677 0.04437527 0.04491689
G1_5 0.04119100 0.04173156 0.04377032 0.04466745
G2_1 0.04336350 0.04423320 0.04547613 0.04395383 0.04331610
G2_2 0.04550683 0.04330957 0.04650139 0.04415664 0.04429918 0.04386371
G2_3 0.04243402 0.04060421 0.04257708 0.04034358 0.04158203 0.04243729 0.04110153
G2_4 0.04462469 0.04483793 0.04526199 0.04423116 0.04244358 0.04430399 0.04405626 0.04210250
G2_5 0.04244397 0.04304633 0.04445892 0.04287979 0.04232089 0.04416588 0.04230360 0.03949559 0.04432153
>
```

G2_3とG2_5の発現ベクトル間の距離

相関係数



対数変換**前**の数値でSpearman相関係数

対数変換**前**の1－Spearman相関係数

対数変換**前**のPearson相関係数

対数変換**前**の順位でPearson相関係数

対数変換**前**の順位でSpearman相関係数

Spearman相関係数を用いれば、対数変換の有無に関わらず、距離の値が変わらないようにすることもできる。しかし、ユークリッド距離などそれ以外の多くの場合には対数変換の有無によって値が変わるため、マイクロアレイデータは対数変換後の値で取り扱うのが一般的である。p100-106。

他の距離(類似度)を定義する手段

- ベクトルxとyの発現パターンの距離$D(x,y)$
  - ユークリッド距離 $D = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
  - マンハッタン距離 $D = \sum_{i=1}^{n} | x_i - y_i |$
  - 最大距離 $D = \max(| x_1 - y_1 |,\ldots | x_i - y_i |,\ldots, | x_n - y_n |)$
  - キャンベラ距離 $D = \sum_{i=1}^{n} \frac{| x_i - y_i |}{| x_i + y_i |}$
  - …

i	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
…	…	…
n	x_n	y_n

計算例（サンプルxとy間の距離D）

	A	B	C	D	E	F
1		*x*	*y*		|xi − yi|	|xi − yi|/|xi + yi|
2	gene1	10.5	12.4		1.9	0.0830
3	gene2	6.4	7.1		0.7	0.0519
4	gene3	8	8.5		0.5	0.0303
5	gene4	10.8	11.4		0.6	0.0270
6	gene5	5.6	6.7		1.1	0.0894
7	gene6	8.4	8.9		0.5	0.0289
8	gene7	6.2	7		0.8	0.0606
9	gene8	6.1	6.8		0.7	0.0543
10	gene9	6.6	6.5		0.1	0.0076
11	gene10	5.1	5.8		0.7	0.0642

$$D = \sum_{i=1}^{n} | x_i - y_i |$$ マンハッタン距離 = 1.9+0.7+0.5+0.6+1.1+0.5+0.8+0.7+0.1+0.7 = 7.6

$$D = \max(| x_i - y_i |)$$ 最大距離 = max(1.9, 0.7, 0.5, 0.6, 1.1, 0.5, 0.8, 0.7, 0.1, 0.7) = 1.9

$$D = \sum_{i=1}^{n} \frac{| x_i - y_i |}{| x_i + y_i |}$$ キャンベラ距離 = 0.0830+0.0519+0.0303+…+0.0642 = 0.4972

- 前処理 | フィルタリング | 分散が小さいものを除去 (last modified 2013/11/15)
- 前処理 | 同じ遺伝子名を持つものをまとめる (last modified 2013/7/31)
- 解析 | 基礎 | 共通遺伝子の抽出 (last modified 2013/6/2)
- 解析 | 基礎 | ベクトル間の距離 (last modified 2013/11/22)
- 解析 | 基礎 | 遺伝子ごとの各種要約統計量の算出 (last modified 2013/6/2)

解析 | 基礎 | ベクトル間の距離

二つのベクトル間の距離を定義する方法は多数存在します。ここでは10 genes ×2 samplesのデータファイル(sample19.txt)を読み込んで二つのサンプル間の距離をいくつかの方法で算出します。
「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、以下をコピペ

1. 10 genes ×2 samplesのデータファイル(sample19.txt)の場合:

```
in_f <- "sample19.txt"

#データファイルの読み込み
data <- read.table(in_f

#本番
dist(t(data), method="e
dist(t(data), method="m
dist(t(data), method="m
dist(t(data), method="c
1 - cor(data, method="p

dist(t(data), method="b
dist(t(data), method="m
1 - cor(data, method="s
```

R Console

```
> #本番
> dist(t(data), method="euclidean")        #ユークリッド(Euclidean)距
         sample1
sample2 2.792848
> dist(t(data), method="manhattan")        #マンハッタン(Manhattan)距
         sample1
sample2     7.6
> dist(t(data), method="maximum")          #チェビシェフ(Chebyshev)距
         sample1
sample2     1.9
> dist(t(data), method="canberra")         #キャンベラ(Canberra)距離
          sample1
sample2 0.4972074
> 1 - cor(data, method="pearson")          #1 - Pearson相関係数
           sample1    sample2
sample1 0.00000000 0.02414407
sample2 0.02414407 0.00000000
>
> dist(t(data), method="binary")           #ハミング(Hamming)距離
         sample1
sample2       0
> dist(t(data), method="minkowski")        #ミンコフスキー(Minkowski)距離
         sample1
sample2 2.792848
> 1 - cor(data, method="spearman")         #1 - Spearman相関係数
           sample1   sample2
sample1 0.0000000 0.1333333
sample2 0.1333333 0.0000000
```

sample19.txt

	A	B	C
1	ID	sample1	sample2
2	gene1	10.5	12.4
3	gene2	6.4	7.1
4	gene3	8	8.5
5	gene4	10.8	11.4
6	gene5	5.6	6.7
7	gene6	8.4	8.9
8	gene7	6.2	7
9	gene8	6.1	6.8
10	gene9	6.6	6.5
11	gene10	5.1	5.8

sample19

他にどんな距離を利用可能か調べたい場合は…

?関数名で詳細な使用法を学ぶ

R Console

```
> ?dist
starting httpd hel
>
> dist(t(data))
          sample1
sample2 2.792848
>
```

Description

This function computes and returns the distance matrix computed by using the specified distance measure to compute the distances between the rows of a data matrix.

Usage

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)

as.dist(m, diag = FALSE, upper
## Default S3 method:
as.dist(m, diag = FALSE, upper

## S3 method for class 'dist'
print(x, diag = NULL, upper = NULL,
      digits = getOption("digits"), justify = "none",
      right = TRUE, ...)

## S3 method for class 'dist'
as.matrix(x, ...)
```

ユークリッド距離でよければ、「method="xxx"」のところを記述しなくてもいいようだ

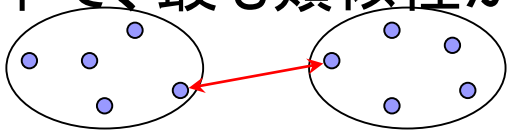
"binary"や"minkowski"というものも指定できるようだが「1-相関係数」を指定することはできないようだ…orz

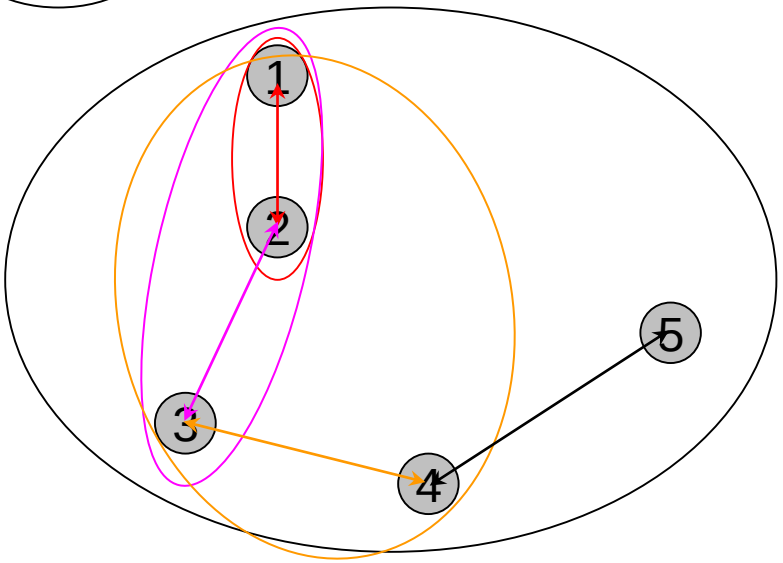
Arguments

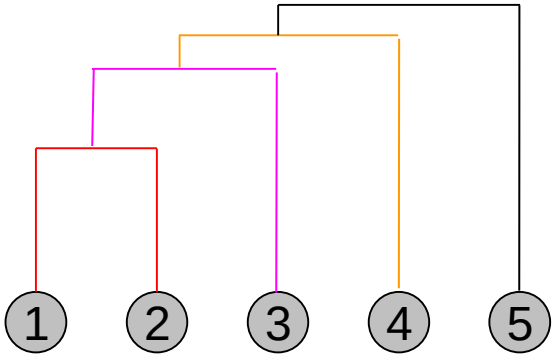
x	a numeric matrix, data frame or "dist" object.
method	the distance measure to be used. This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski". Any unambiguous substring can be given.

クラスターをまとめる方法

- 最短距離法（単連結法；single-linkage）
  - 二つのクラスター間の類似度を、それぞれに含まれる要素対の中で、最も類似性が高い対の間の類似度で定義

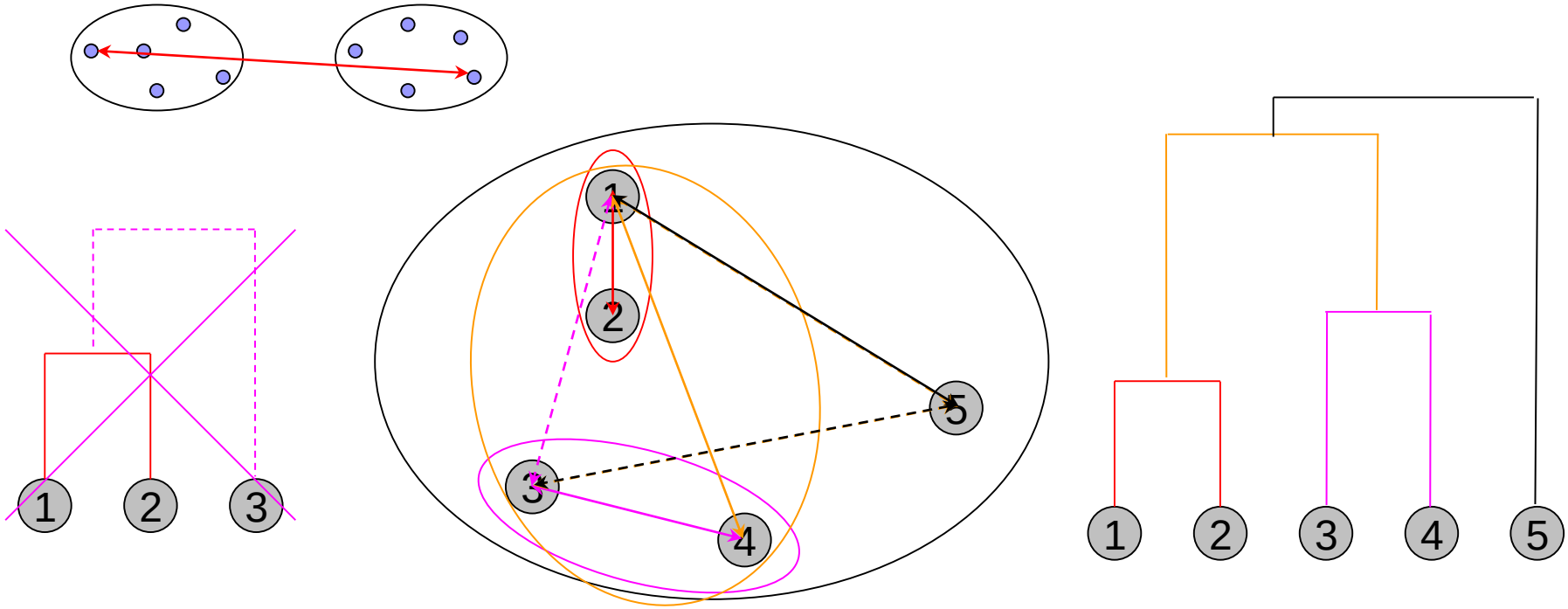






クラスターをまとめる方法

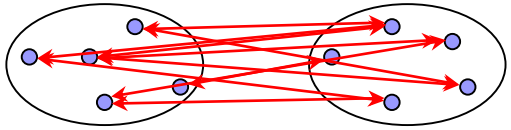
- 最長距離法（完全連結法；complete-linkage ）
  - クラスターをmergeするときの基準として、最遠距離を用いる

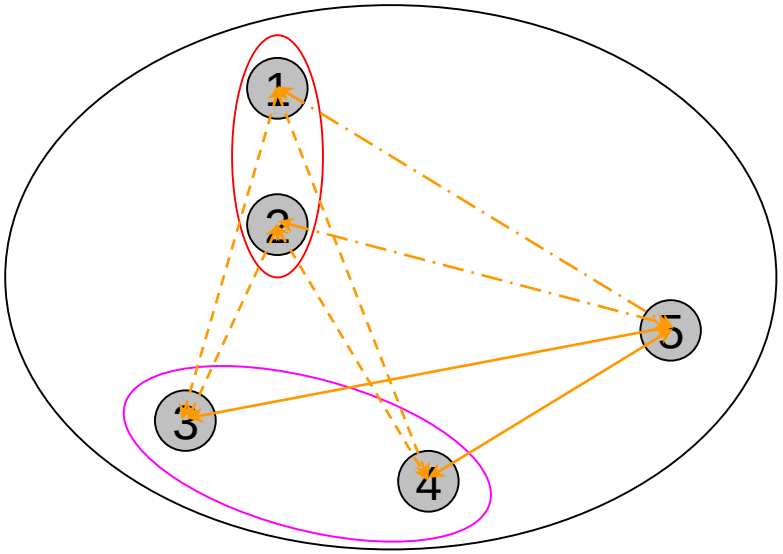


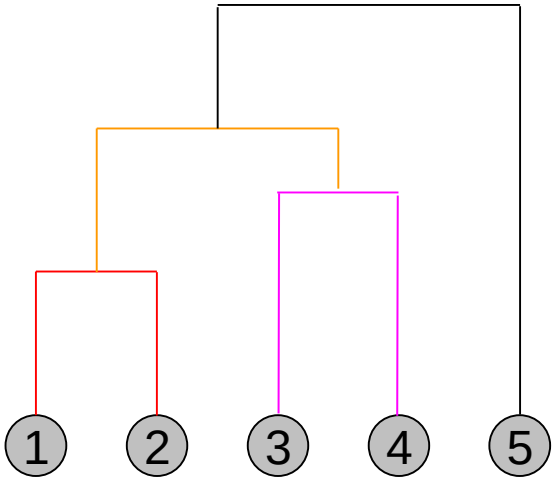
電気泳動波形解析（ピークアラインメント）にも応用可能

クラスターをまとめる方法

- 群平均法（平均連結法；average-linkage ）
  - クラスターをmergeするときの基準として、群間平均距離を用いる







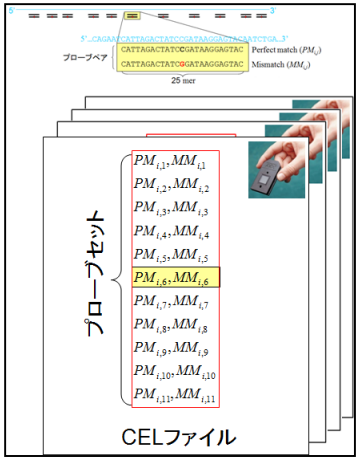
用いる方法によって得られるデンドログラム（樹形図）が異なる

他のクラスターをまとめる手段

- 重心法（Centroid）：重心間距離を利用
- ウォード法：群内平方和の増加量が最小となるクラスターと併合
- メディアン（Median）法：群間距離の中央値を利用
- McQuitty法･･･
- 可変（flexible）法･･･

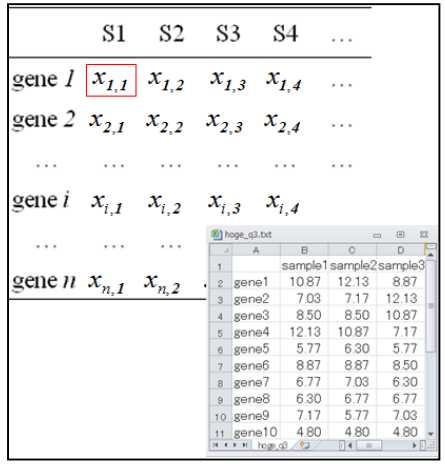
分野にもよるらしいが**群平均法**が最もよく利用されている?!（**ウォード法**も?!）…。
いろいろ試して総合的に判断することが重要

得られる樹形図の可能性は無数



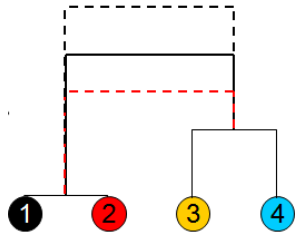
生データファイル

前処理法



遺伝子発現行列

クラスタリング



①前処理法
- **MAS**
- **RMA**
- **RobLoxBioC**
- …

…

×

②スケーリング
- **対数変換**
- 相対値(0~1)
- Z-score化
- …

…

×

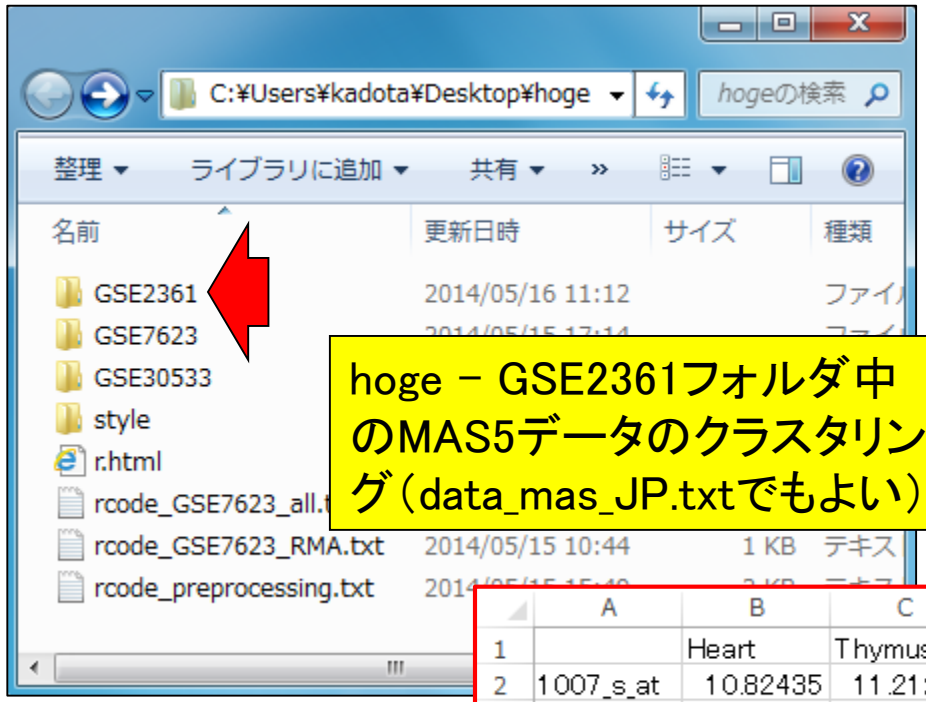
③距離
- **1ー相関係数**
- ユークリッド
- マンハッタン
- キャンベラ
- …

×

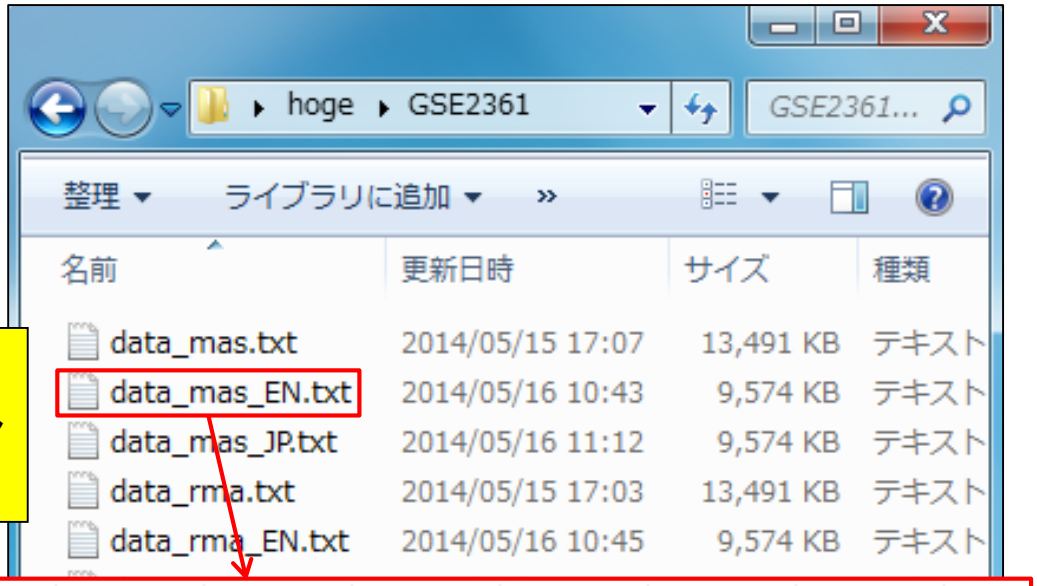
④群の併合
- 単連結法
- 完全連結法
- **平均連結法**
- ウォード法
- …

サンプル間クラスタリングをやってみよう

- Affymetrix GeneChip
  - Ge et al., *Genomics*, **86**: 127–141, 2005
    - **GSE2361**、GPL96 (Affymetrix Human Genome U133A Array)、22,283 probesets
    - **ヒト36サンプル**：Heart (心臓)、Thymus (胸腺)、Spleen (脾臓)、Ovary (卵巣)、Kidney (腎臓)、Skeletal Muscle (骨格筋)、Pancreas (膵臓)、Prostate (前立腺)、…



hoge – GSE2361フォルダ中のMAS5データのクラスタリング（data_mas_JP.txtでもよい）



	A	B	C	D	E	F	G	H	I	J
1		Heart	Thymus	Spleen	Ovary	Kidney	Skeletal_Mu	Pancreas	Prostate	Small_I
2	1007_s_at	10.82435	11.21261	9.898072	10.8948	11.75531	10.92032	11.61243	11.22101	11.48
3	1053_at	6.621657	6.47488	7.371465	4.168622	7.350244	7.209918	8.362777	7.176571	7.75
4	117_at	8.259603	8.647364	8.689724	7.875649	5.083001	8.165044	7.96707	8.418082	5.719
5	121_at	11.26699	11.04186	11.39999	11.02855	13.13267	11.39138	12.32899	11.0559	11.28
6	1255_g_at	7.1757	6.477278	6.781766	7.048799	7.30767	6.600267	6.42359	6.694412	7.30
7	1294_at	9.137586	9.718507	9.083742	9.014997	8.813377	8.402562	9.146946	8.421351	10.12

- 解析 | 基礎 | ベクトル間の距離(last modified 2013/11/22)
- 解析 | 基礎 | 遺伝子ごとの 各種要約統計量の 算出(last modified 2013/6/2)
- 解析 | 基礎 | 最大発現量を示す組織の同定(last modified 2013/6/2)
- 解析 | 基礎 | 似た発現パターンを持つ遺伝子の同定 (last modified 2013/6/2)
- 解析 | 基礎 | 平均-分散プロット (last modified 2013/11/25)
- 解析 | クラスタリング | 階層的 | について (last modified 2009/8/12)
- 解析 | クラスタリング | 階層的 | pvclust (zuki 2006) (last modified 2010/8/5)
- 解析 | クラスタリング | 階層的 | hclust odified 2013/11/19)
- 解析 | クラスタリング | 階層的 | hclust後 詳細な解析 (last modified 2009/8/7)
- 解
- 解

- 解析 | クラスタリング | 階層的 | hclust

解析 | クラスタリング | 階層的 | hclust NEW

階層的クラスタリングのやり方を示します。1.用いた前処理法(MAS5やRMAなど)、2.スケーリング方法(対数変換やZ-scoreなど)、3.距離(または非類似度)を定義する方法(ユークリッド距離など)、4.クラスターをまとめる方法(平均連結法やウォード法など)でどの方法を採用するかで結果が変わってきます。
「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、以下をコピペ

1. サンプルデータ3の sample3.txtの場合:

サンプル間クラスタリング(距離: 1-Pearson相関係数、方法: 平均連結法(average))でR Graphics画面上に表

```
in_f <- "sample3.txt"                    #入力ファイル名を指定してin_fに格納
param <- "average"                       #方法(method)を指定

#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の読み

#本番
data.dist <- as.dist(1 - cor(data, method="pearson"))#サンプル間の距離を計算した結果をdata.distに格
out <- hclust(data.dist, method=param) #階層的クラスタリングを実行した結果をoutに格納
plot(out)                              #樹形図（デンドログラム）の表示
```

左記のテンプレートスクリプトとの違いは、入力ファイル名とSpearman相関係数の部分のみ

```
#########################################↓
###  MAS5データのクラスタリング  ###↓
#########################################↓
in_f <- "data_mas_EN.txt"↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")↓
#colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))↓
data.dist <- as.dist(1 - cor(data, method = "spearman"))↓
out <- hclust(data.dist, method = "average")↓
plot(out)↓
←
```

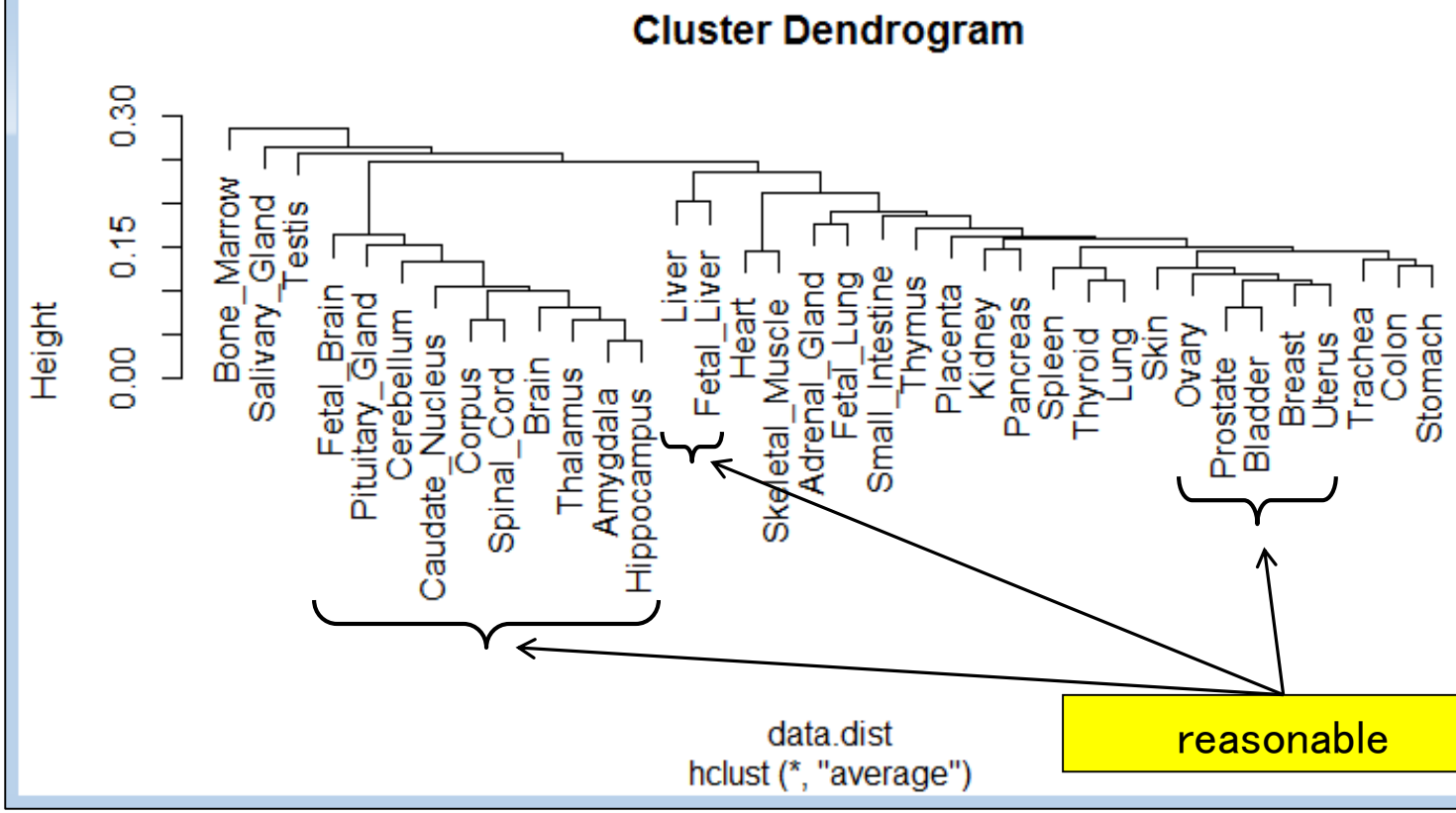
rcode_clustering.txt

```
#####################################↓
###  MAS5データのクラスタリング  ###↓
#####################################↓
in_f <- "data_mas_EN.txt"↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")↓
#colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))↓
data.dist <- as.dist(1 - cor(data, method = "spearman"))↓
out <- hclust(data.dist, method = "average")↓
plot(out)↓
←
```

R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/GSE2361"
> in_f <- "data_mas_EN.txt"
> data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
> #colnames(data) <- c(paste("G1 ", 1:5, sep=""), paste("G2 ", 1:5, sep=""))
```

R Graphics: Device 2 (ACTIVE)

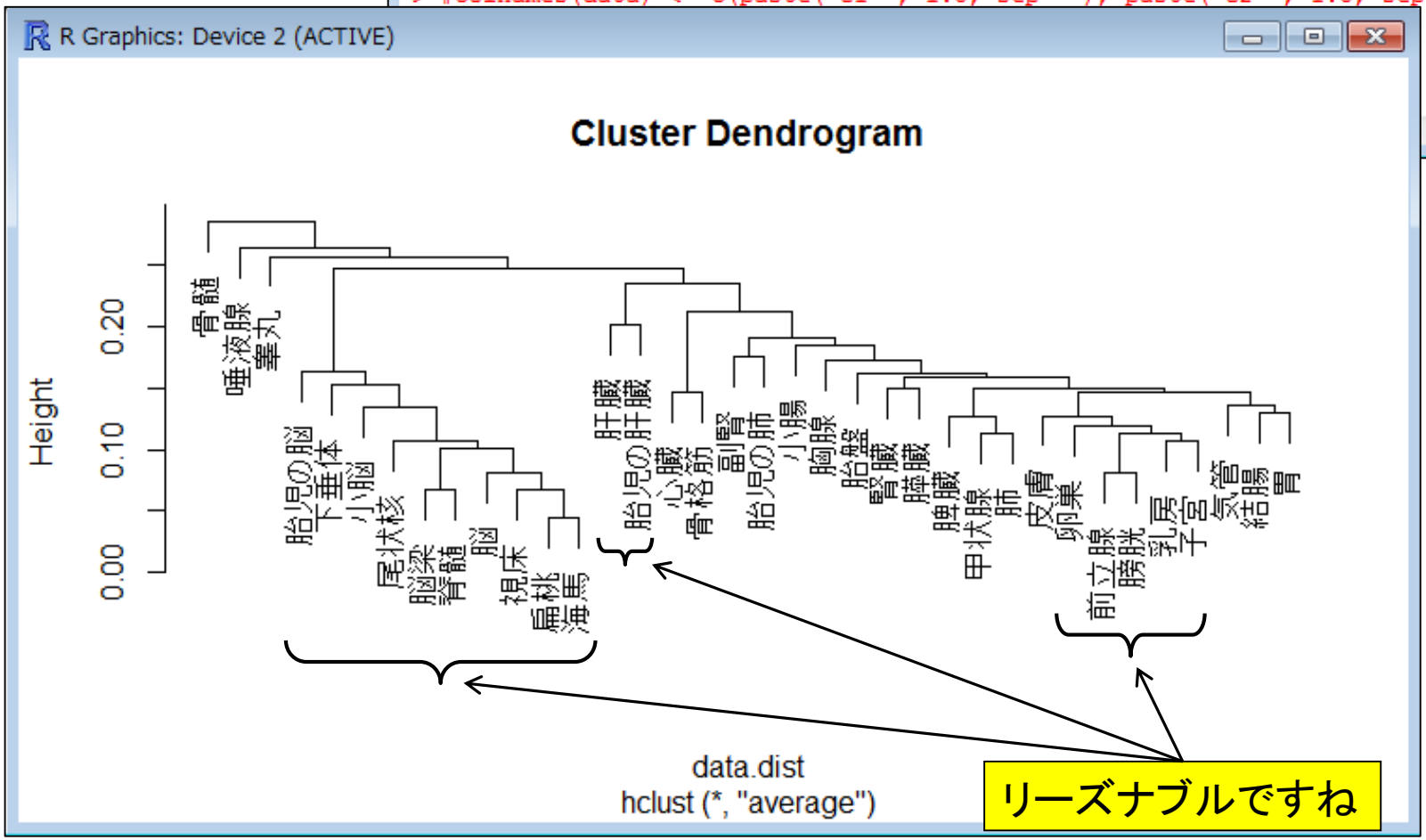


```
#####################################↓
###  MAS5データのクラスタリング  ###↓
#####################################↓
in_f <- "data_mas_JP.txt"↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")↓
#colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))↓
data.dist <- as.dist(1 - cor(data, method = "spearman"))↓
out <- hclust(data.dist, method = "average")↓
plot(out)↓
←
```

R Console

```
> in_f <- "data_mas_JP.txt"
> data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
> #colnames(data) <- c(paste("G1 ", 1:5, sep=""), paste("G2 ", 1:5, sep=""))
```

R Graphics: Device 2 (ACTIVE)



クラスタリング結果をファイルに保存

解析 | クラスタリング | 階層的 | hclust NEW

階層的クラスタリングのやり方を示します。1.用いた前処理法(MAS5やRMAなど)、2.スケーリング方法(対数変換やZ-scoreなど)、3.距離(または非類似度)を定義する方法(ユークリッド距離など)、4.クラスターをまとめる方法(平均連結法やウォード法など)

「フ

1. サン

サンプ

```
in_f
para

#入力
data

#本番
data
out
plot
```

3. サンプルデータ3のsample3.txtの場合:

サンプル間クラスタリング(距離: 1-Spearman相関係数、方法: 平均連結法(average))で図の大きさを指定してpng形式ファイルで保存するやり方です。

```
in_f <- "sample3.txt"                  #入力ファイル名を指定してin_fに格納
out_f <- "hoge3.png"                   #出力ファイル名を指定してout_fに格納
param <- "average"                     #方法(method)を指定
param_fig <- c(500, 400)               #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#入力ファイ
data <- re

#本番
data.dist
out <- hcl

#ファイルに
png(out_f,
plot(out)
dev.off()
```

rcode_clustering_png.txt(変更点は赤矢印部分のみ)

```
#################################################################################↓
###  MAS5データのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###↓
#################################################################################↓
in_f <- "data_mas_JP.txt"              #入力ファイル名を指定してin_fに格納↓
out_f <- "data_mas_JP.png"             #出力ファイル名を指定してout_fに格納↓
param <- "average"                     #方法(method)を指定↓
param_fig <- c(720, 400)               #ファイル出力時の横幅と縦幅を指定(単位はピクセル)↓
↓
#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の読み込み↓
↓
#本番↓
data.dist <- as.dist(1 - cor(data, method="spearman"))#サンプル間の距離を計算した結果をdata.distに格納↓
out <- hclust(data.dist, method=param) #階層的クラスタリングを実行した結果を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↓
plot(out)                              #樹形図（デンドログラム）の表示↓
dev.off()                              #おまじない↓
```

クラスタリング結果をファイルに保存

rcode_clustering_png.txt （変更点は赤矢印部分のみ）

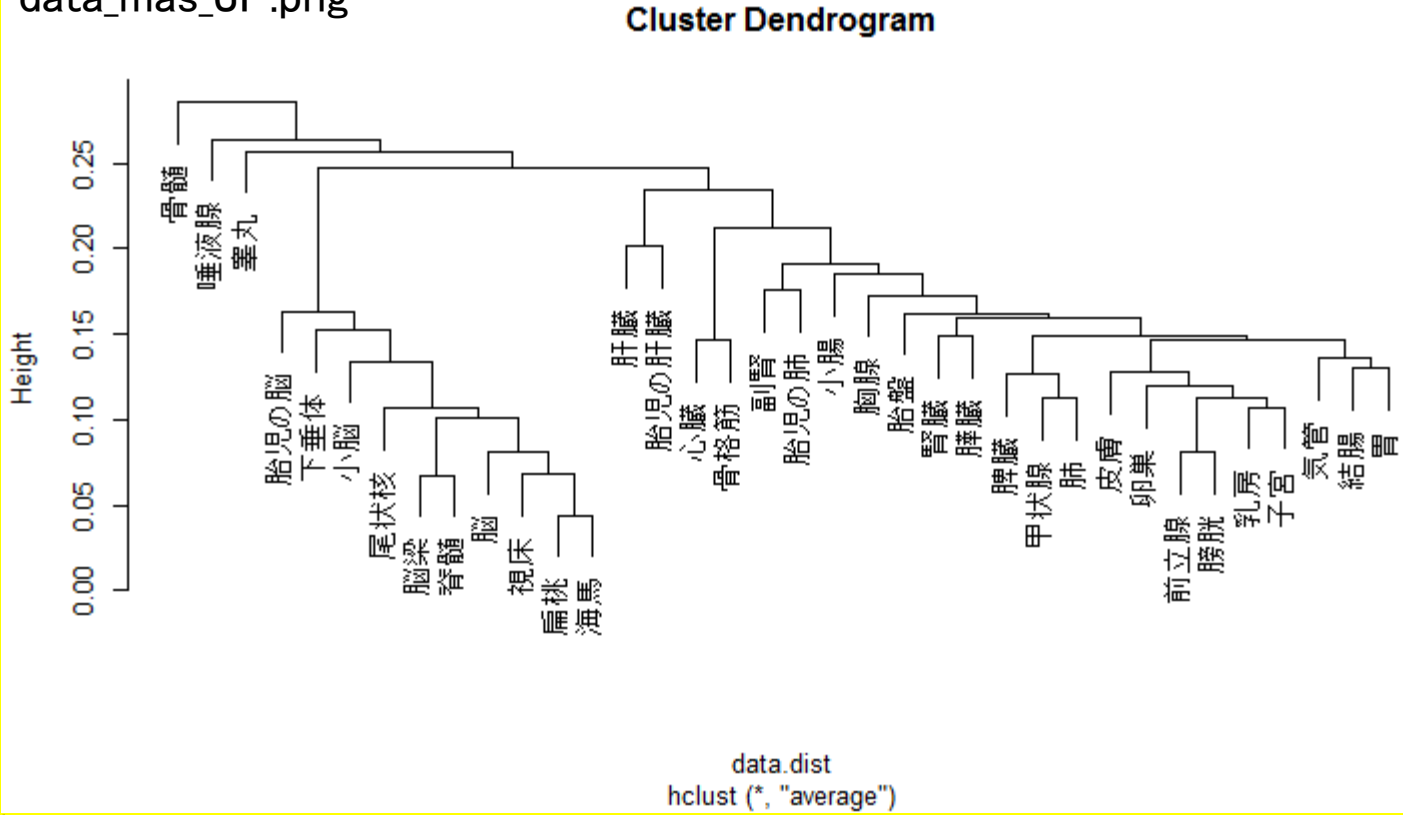
```
##############################################################################
###  MAS5データのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###
##############################################################################
in_f <- "data_mas_JP.txt"              #入力ファイル名を指定してin_fに格納
out_f <- "data_mas_JP.png"             #出力ファイル名を指定してout_fに格納
param <- "average"                     #方法(method)を指定
param_fig <- c(720, 400)

#入力ファイルの読み込み
data <- read.table(in_f, h

#本番
data.dist <- as.dist(1 - c
out <- hclust(data.dist, m

#ファイルに保存
png(out_f, pointsize=13, w
plot(out)
dev.off()
```

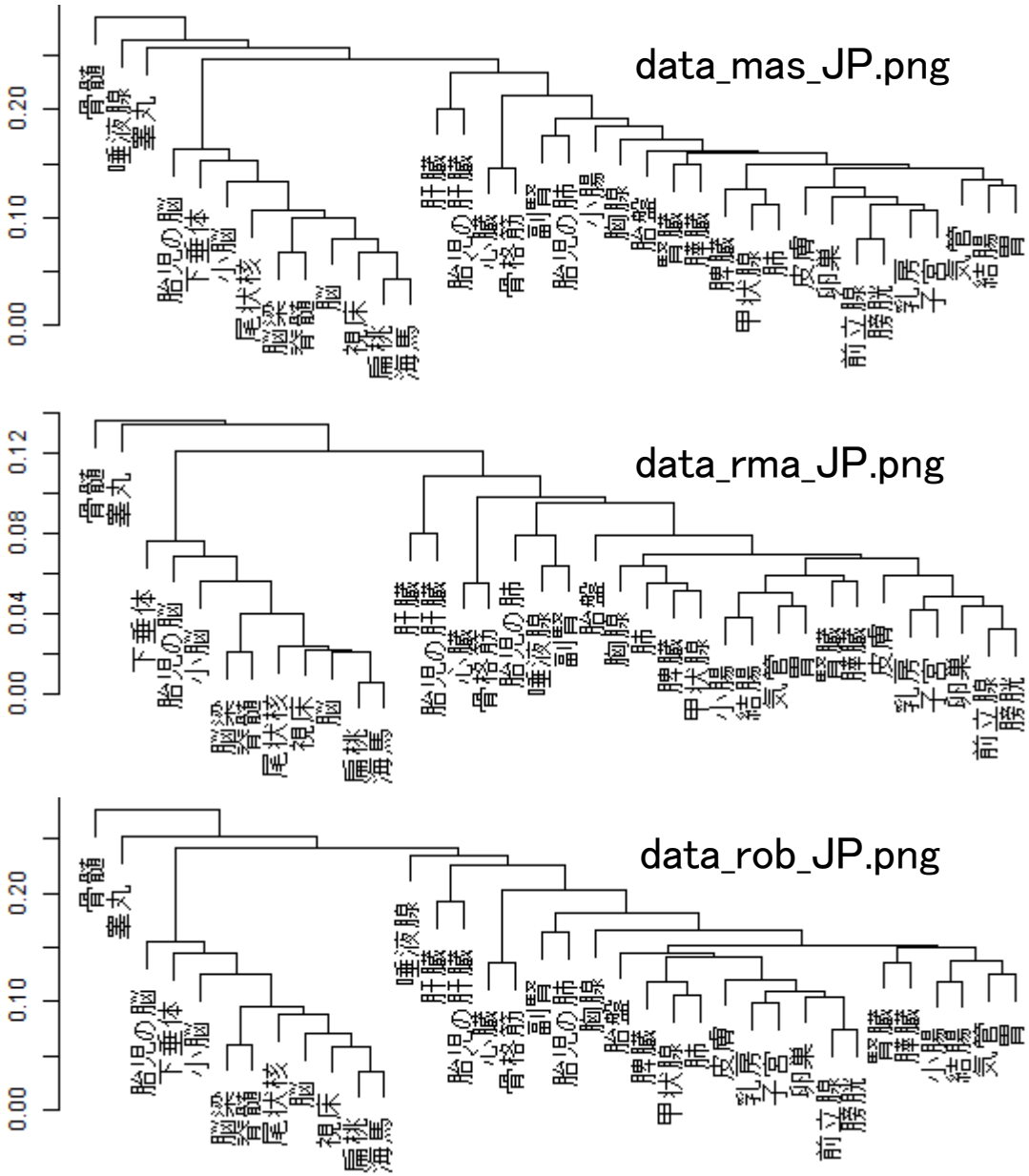
data_mas_JP.png



サンプル間クラスタリング（GSE2361）

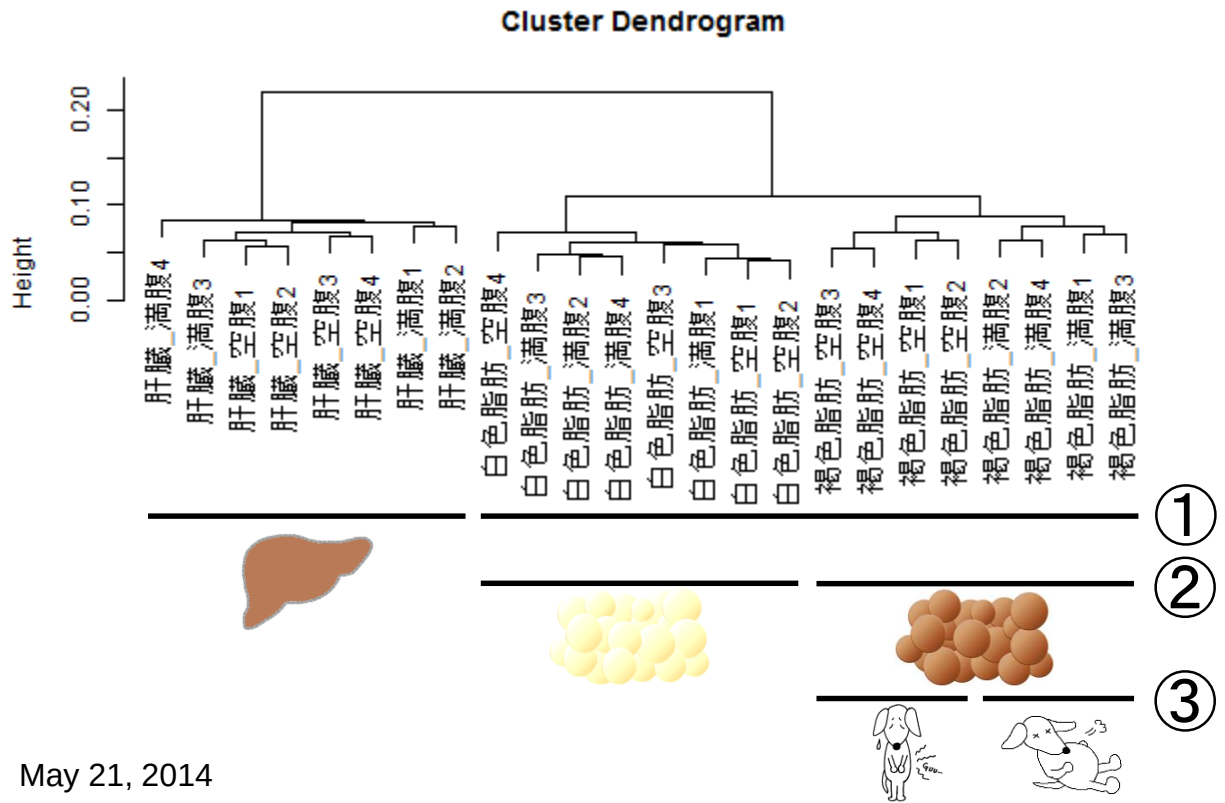
- GSE2361のサンプル間クラスタリングをMAS5, RMA, およびRobLoxBioC前処理法を適用したデータについても行ってみよう。
  - 1 − Spearman相関係数
  - 平均連結法

ヒント：rcode_clustering_png.txtをテンプレートとすればよい。
正解：右図の結果。どの前処理法を適用しても似たような樹形図が得られていることがわかる。



サンプル間クラスタリング（GSE7623）

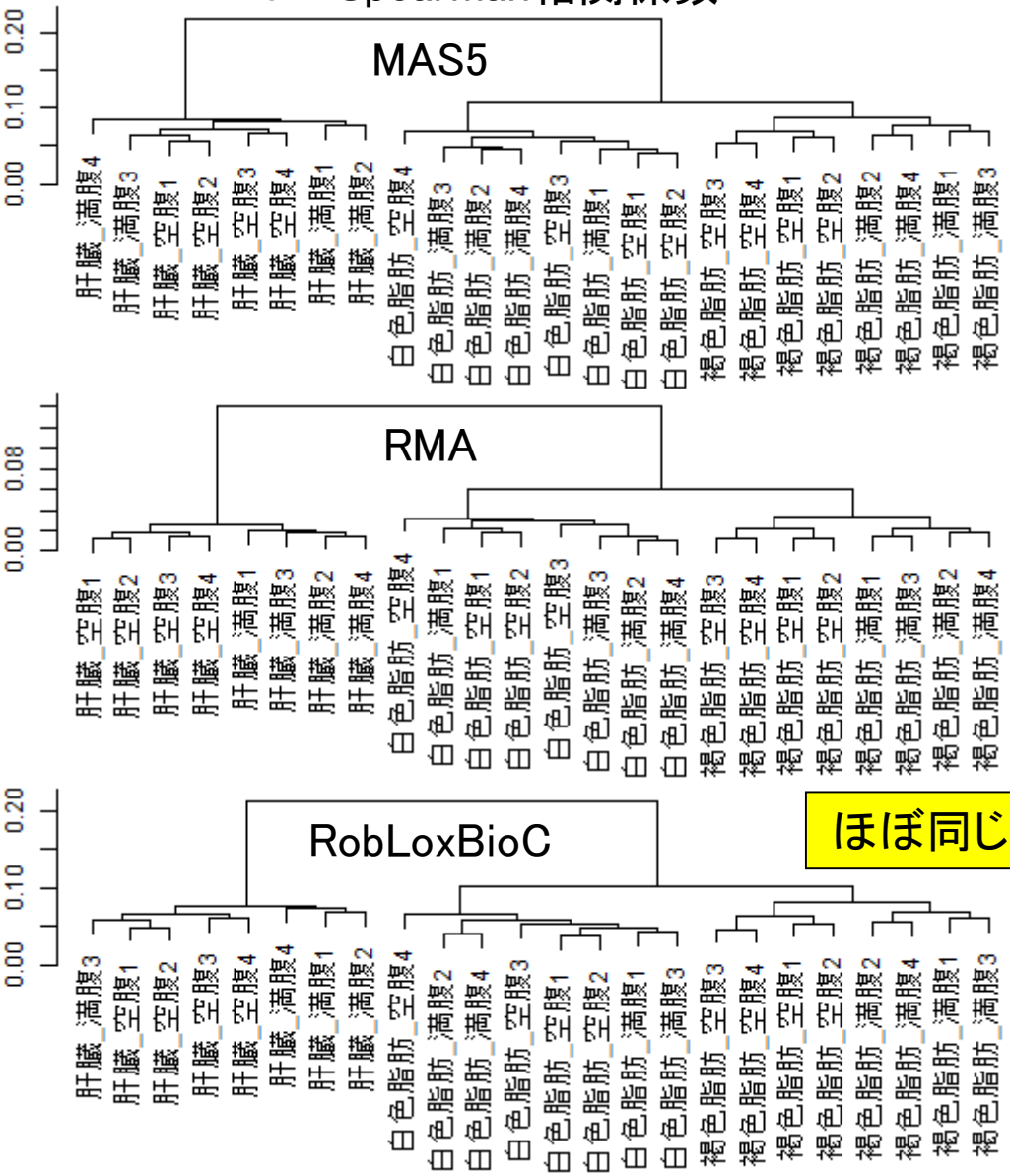
- Nakai et al., *Biosci Biotechnol Biochem.*, **72**: 139-148, 2008
  - **GSE7623**、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
  - **ラット24サンプル**：Brown adipose tissue（褐色脂肪組織; BAT）8サンプル、White adipose tissue（白色脂肪組織; WAT）8サンプル、Liver（肝臓; LIV）8サンプル
    - BAT 8サンプル：通常（BAT_fed）4サンプル　対　24時間絶食（BAT_fas）4サンプル
    - WAT 8サンプル：通常（WAT_fed）4サンプル　対　24時間絶食（WAT_fas）4サンプル
    - LIV 8サンプル：通常（LIV_fed）4サンプル　対　24時間絶食（LIV_fas）4サンプル



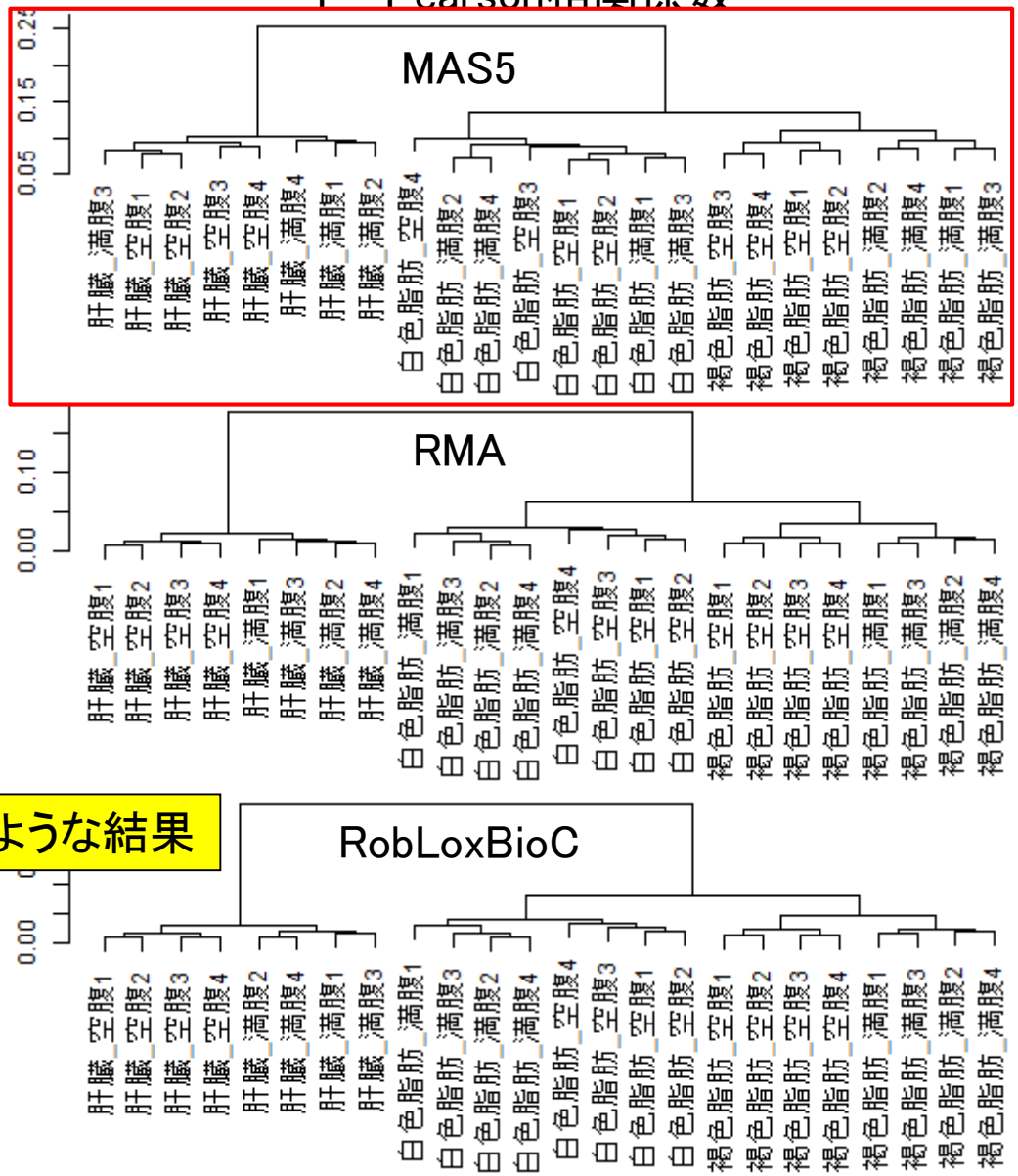
rcode_clustering_png.txtの実行結果。
①肝臓と脂肪間で大きく二つのクラスターに分かれている。
②脂肪の中でも白色脂肪と褐色脂肪に分かれている。
③褐色脂肪は空腹（24時間絶食）と満腹（通常）できれいに分かれている。

サンプル間クラスタリング（GSE7623）

1 − Spearman相関係数

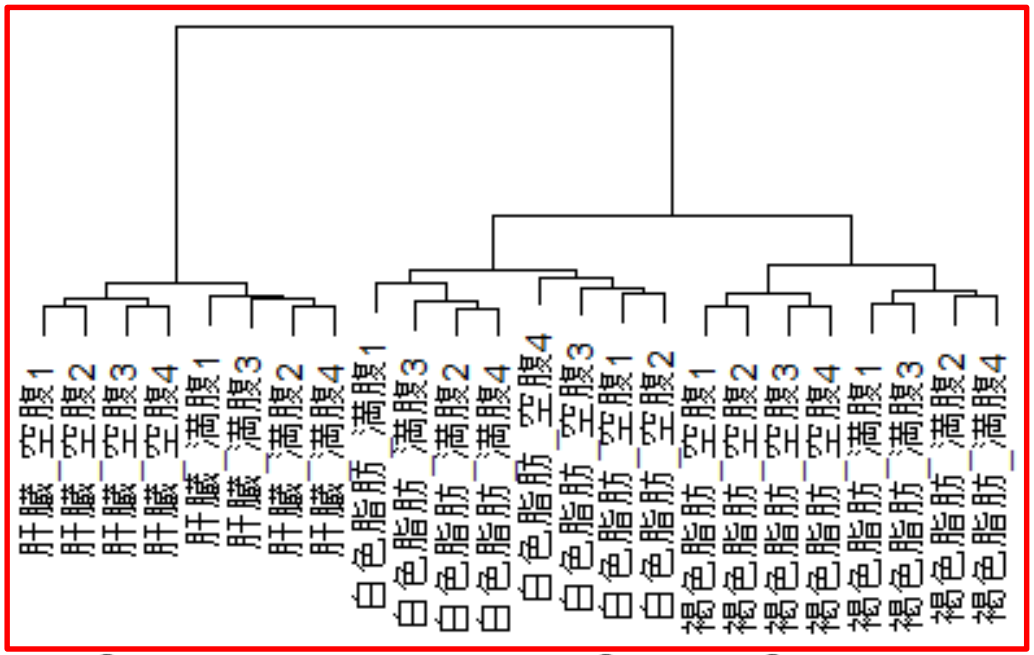


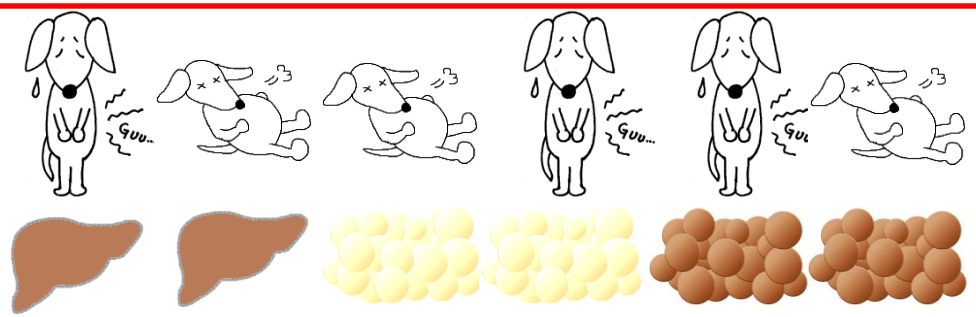
1 − Pearson相関係数



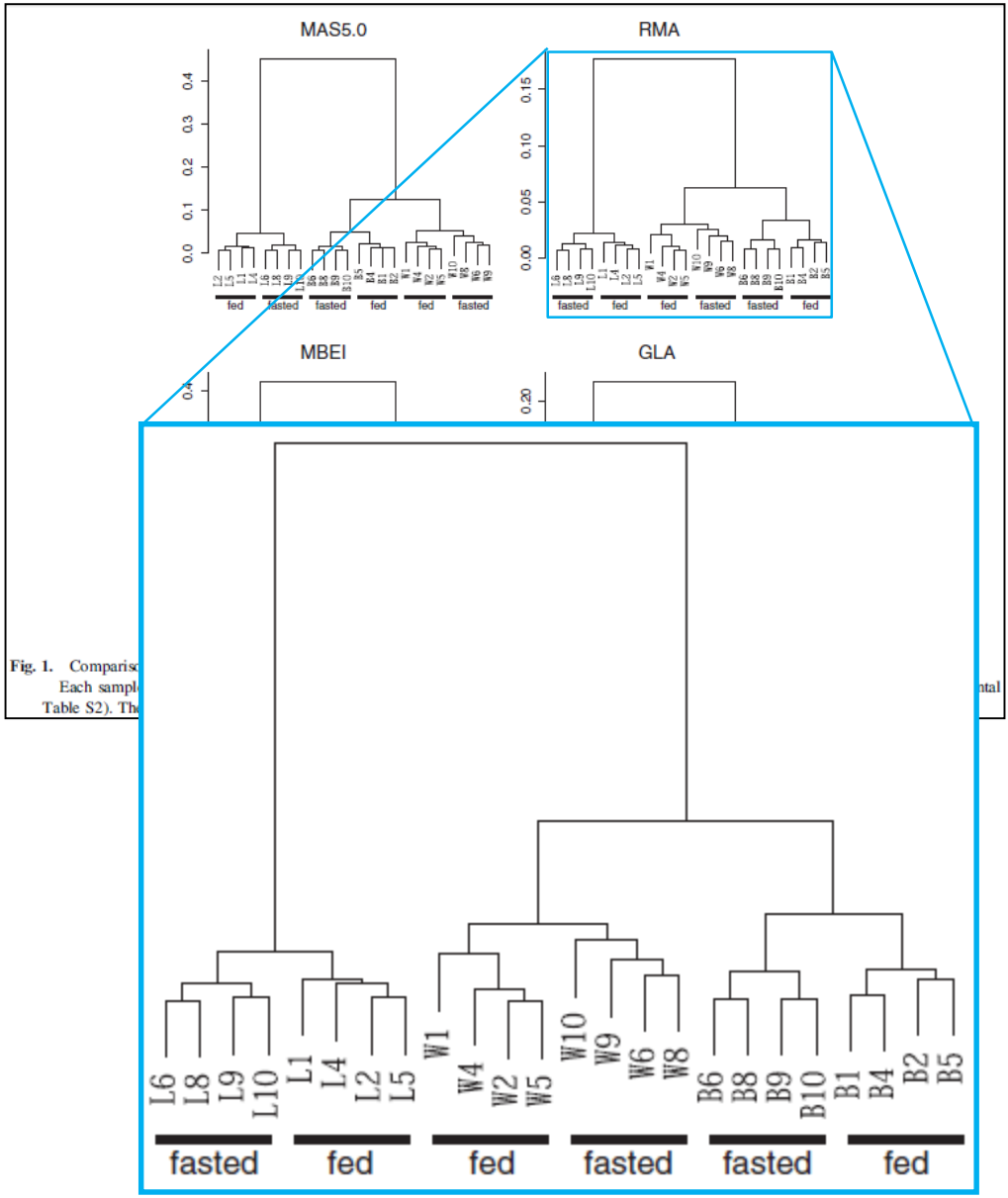
ほぼ同じような結果

原著論文と比較



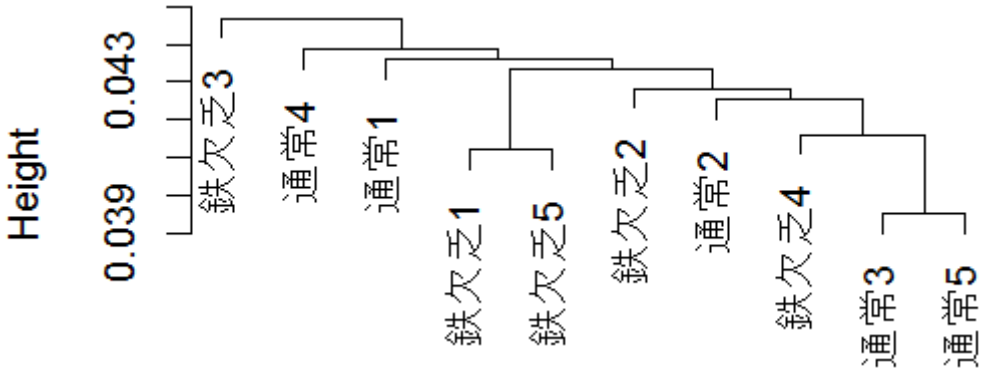


ばっちりFig. 1（の一部）を再現できました

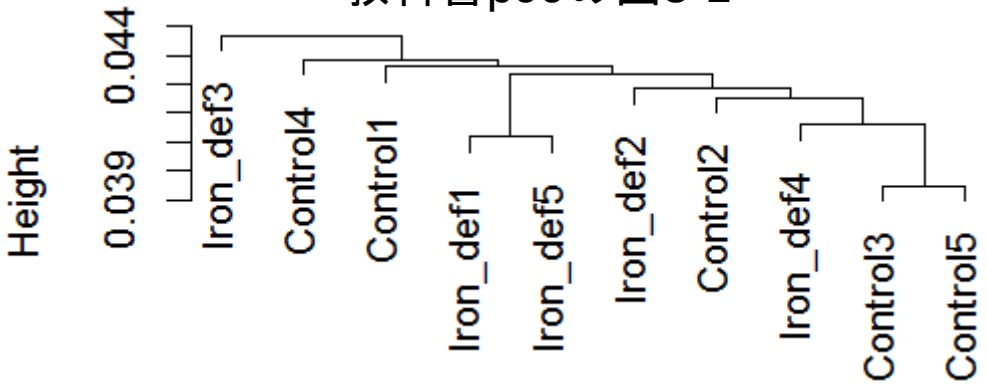


サンプル間クラスタリング（GSE30533）

- Kamei et al., *PLoS One*, **8**: e65732, 2013
  - **GSE30533**、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
  - **ラット10サンプル**：全てLiver (肝臓) サンプル
  - iron-deficient diet (Iron_def) 5サンプル　対　control diet (Control) 5サンプル



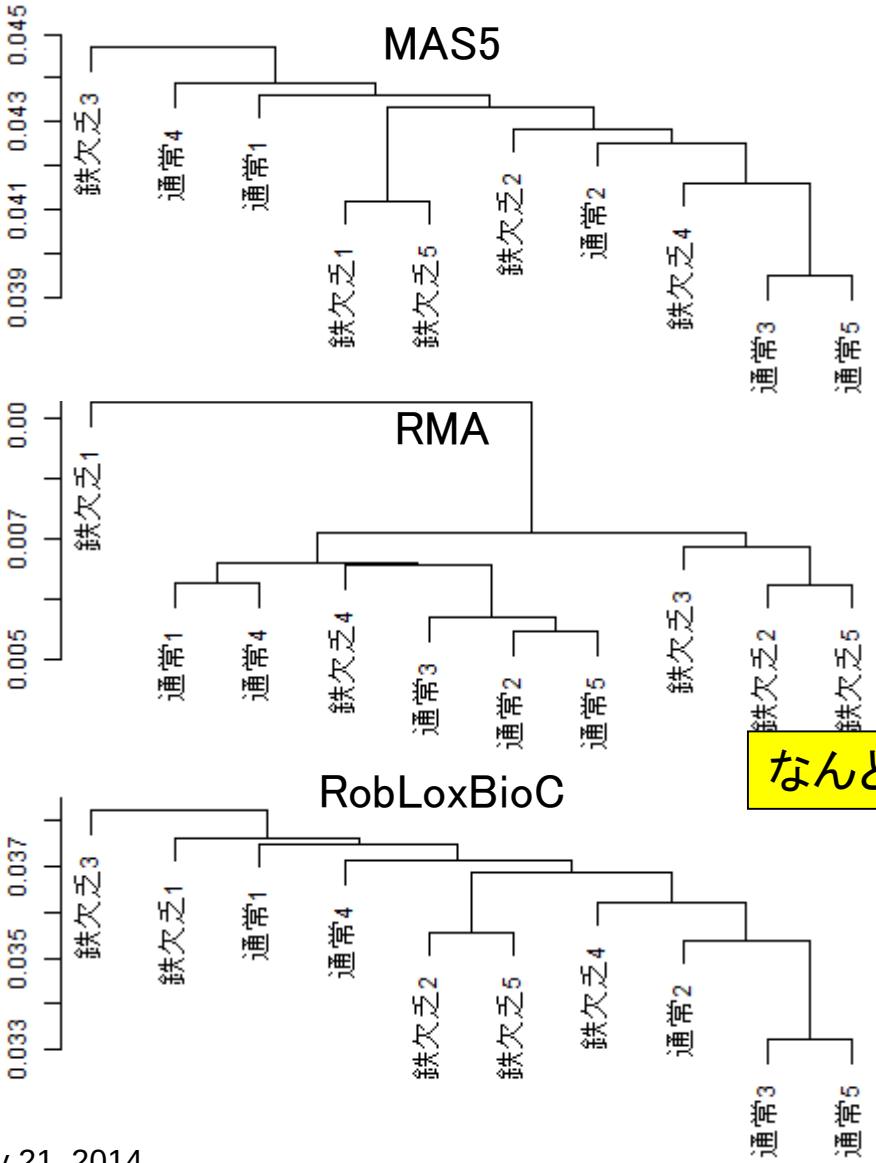
教科書p99の図3-1



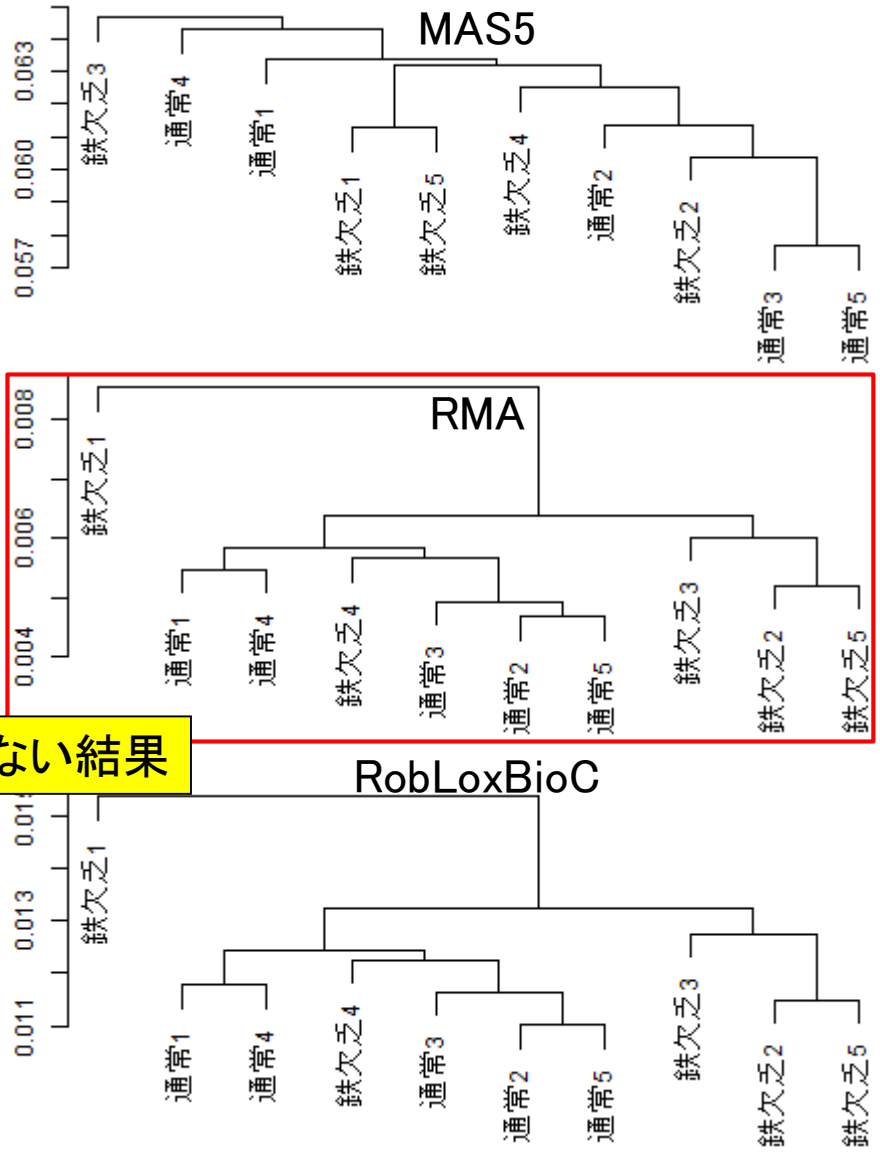
rcode_clustering_png.txtの実行結果。肝臓全体の発現プロファイルが通常状態と鉄欠乏状態という違い程度では明確に区別できない、ということかもしれない…。

サンプル間クラスタリング（GSE7623）

1 − Spearman相関係数

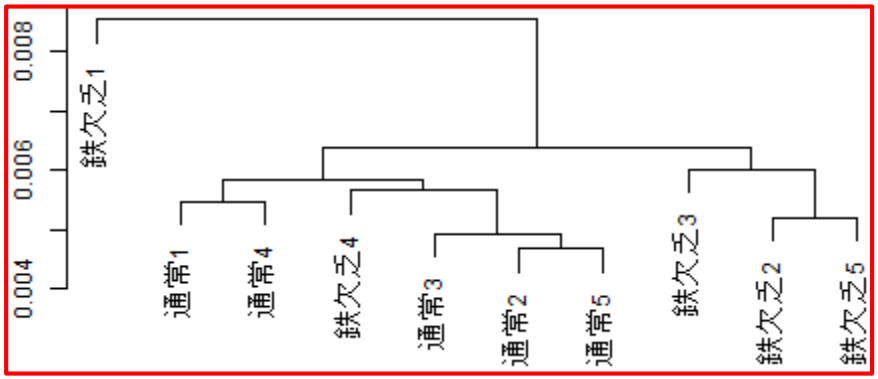


1 − Pearson相関係数

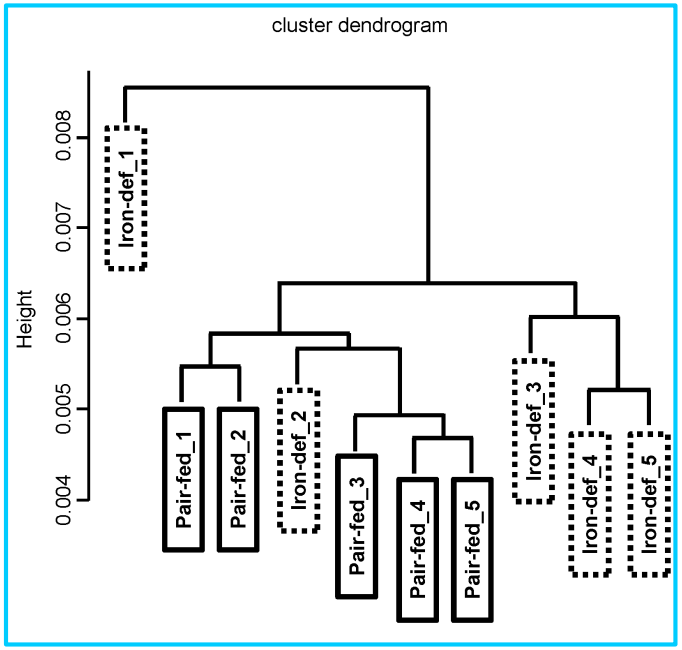


なんともいえない結果

原著論文と比較



原著論文のFigure S1



（サンプルのラベル番号が異なるだけで実質的には）同じ結果

同一アレイデータはマージ可能

- Affymetrix GeneChip
  - Ge et al., *Genomics*, **86**: 127–141, 2005
    - **GSE2361**、GPL96 (Affymetrix Human Genome U133A Array)、22,283 probesets
    - **ヒト36サンプル**：Heart (心臓)、Thymus (胸腺)、Spleen (脾臓)、Ovary (卵巣)、Kidney (腎臓)、Skeletal Muscle (骨格筋)、Pancreas (膵臓)、Prostate (前立腺)、…
  - Nakai et al., *Biosci Biotechnol Biochem.*, **72**: 139–148, 2008
    - **GSE7623**、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
    - **ラット24サンプル**：Brown adipose tissue (褐色脂肪組織; BAT) 8サンプル、White adipose tissue (白色脂肪組織; WAT) 8サンプル、Liver (肝臓; LIV) 8サンプル
      - BAT 8サンプル：通常 (BAT_fed) 4サンプル　対　24時間絶食 (BAT_fas) 4サンプル
      - WAT 8サンプル：通常 (WAT_fed) 4サンプル　対　24時間絶食 (WAT_fas) 4サンプル
      - LIV 8サンプル：通常 (LIV_fed) 4サンプル　対　24時間絶食 (LIV_fas) 4サンプル
  - Kamei et al., *PLoS One*, **8**: e65732, 2013
    - **GSE30533**、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
    - **ラット10サンプル**：全てLiver (肝臓) サンプル
    - iron-deficient diet (Iron_def) 5サンプル　対　control diet (Control) 5サンプル

この2つの論文は同一プラットフォーム (同一アレイ) を利用している。
3' 発現アレイ用いることで、他の多くのデータセットとの比較が可能

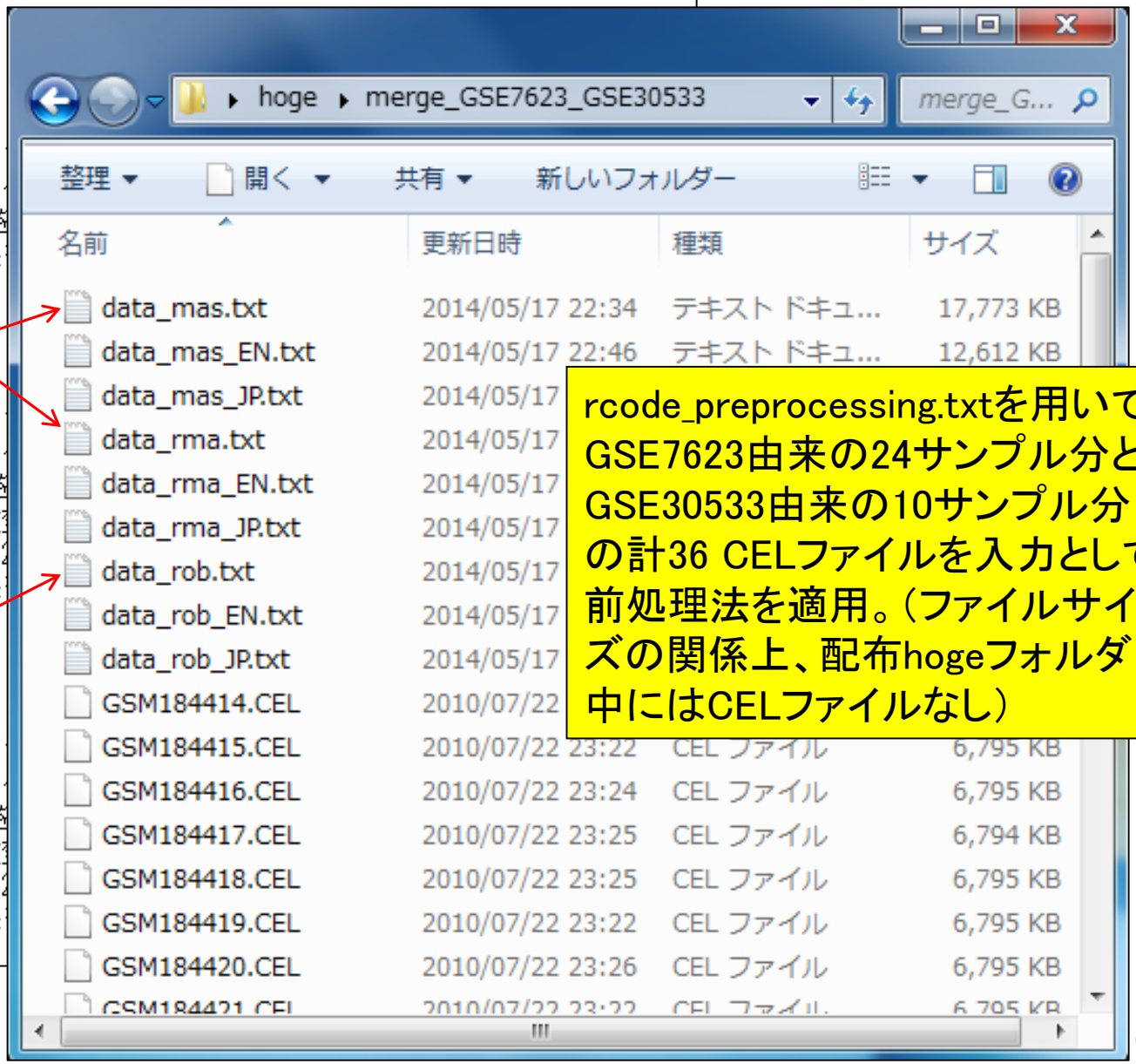
- 正規化 | Affymetrix GeneChip | RMA (Irizarry 2003)
- 正規化 | Affymetrix GeneChip | MAS5.0 (Hubbell 2002)
- 正規化 | Affymetrix GeneChip | rmx (Kohl 2010)

```
##########################################################
###  作業ディレクトリ中のCELファイルの読み込み  ###
##########################################################
library(affy)                          #パッケージの読み込み
hoge <- ReadAffy()                     #*.CELファイルの読み込み

##########################
###  RMA前処理法実行  ###
##########################
out_f <- "data_rma.txt"                #出力
library(affy)                          #パッ
eset <- rma(hoge)                      #RMAを
write.exprs(eset, file=out_f)          #結果

##########################
###  MAS5前処理法実行  ###
##########################
out_f <- "data_mas.txt"                #出力
library(affy)                          #パッ
eset <- mas5(hoge)                     #MASを
exprs(eset)[exprs(eset) < 1] <- 1      #対数
exprs(eset) <- log(exprs(eset), 2)     #底を
write.exprs(eset, file=out_f)          #結果

#######################################
###  RMX (RobLoxBioC)前処理法実行  ###
#######################################
out_f <- "data_rob.txt"                #出力
library(RobLoxBioC)                    #パッ
eset <- robloxbioc(hoge)               #rmxを
exprs(eset)[exprs(eset) < 1] <- 1      #対数
exprs(eset) <- log(exprs(eset), 2)     #底を
write.exprs(eset, file=out_f)          #結果
```



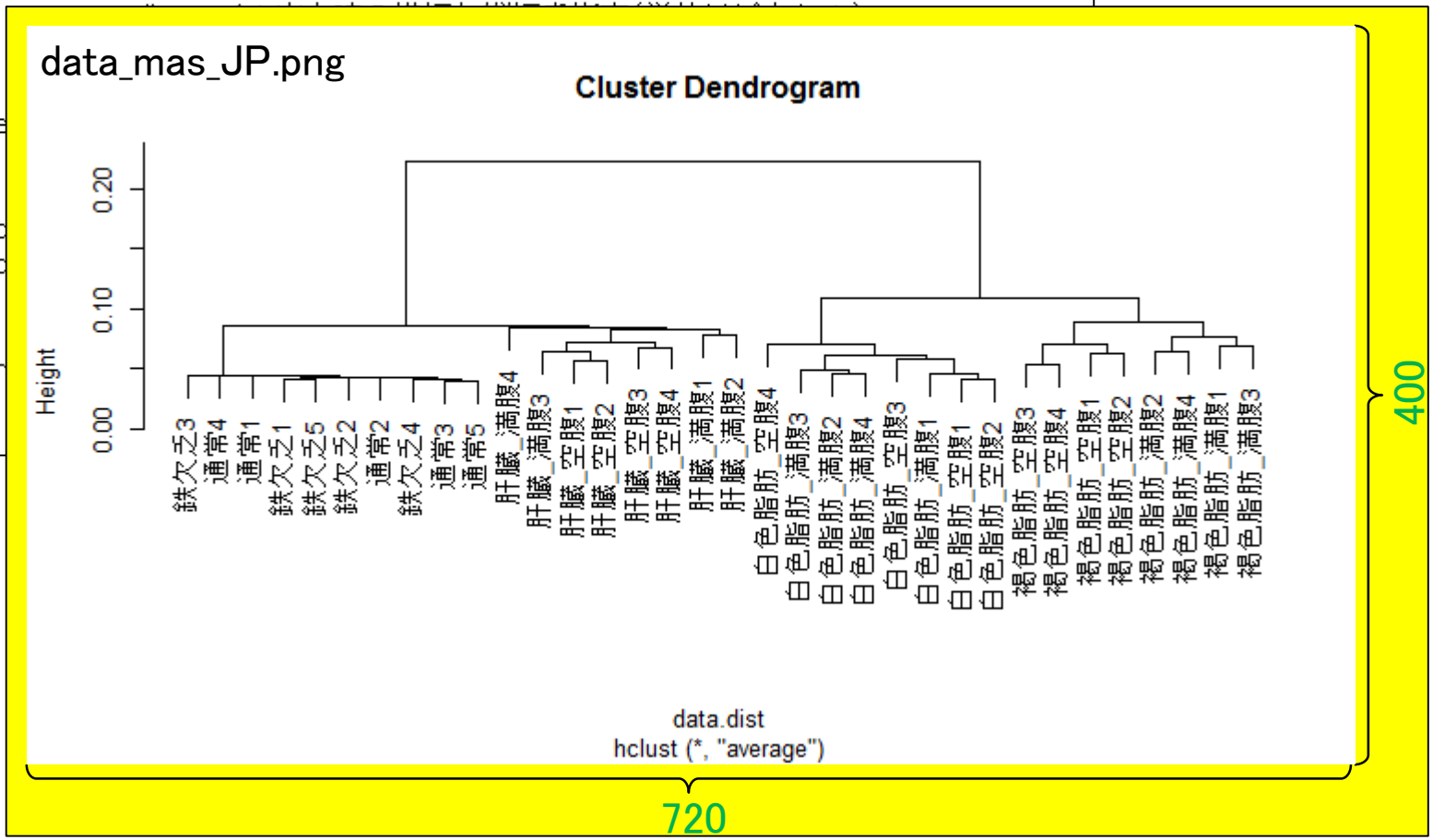
rcode_preprocessing.txtを用いてGSE7623由来の24サンプル分とGSE30533由来の10サンプル分の計36 CELファイルを入力として前処理法を適用。(ファイルサイズの関係上、配布hogeフォルダ中にはCELファイルなし)

クラスタリング結果をファイルに保存

rcode_clustering_png.txt

```
#################################################################################↓
###  MAS5データのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###↓
#################################################################################↓
in_f <- "data_mas_JP.txt"              #入力ファイル名を指定してin_fに格納↓
out_f <- "data_mas_JP.png"             #出力ファイル名を指定してout_fに格納↓
param <- "average"                     #方法(method)を指定↓
param_fig <- c(720, 400)
↓
#入力ファイルの読み込み↓
data <- read.table(in_f, heade
↓
#本番↓
data.dist <- as.dist(1 - cor(d
out <- hclust(data.dist, metho
↓
#ファイルに保存↓
png(out_f, pointsize=13, width
plot(out)
dev.off()
```

data_mas_JP.png

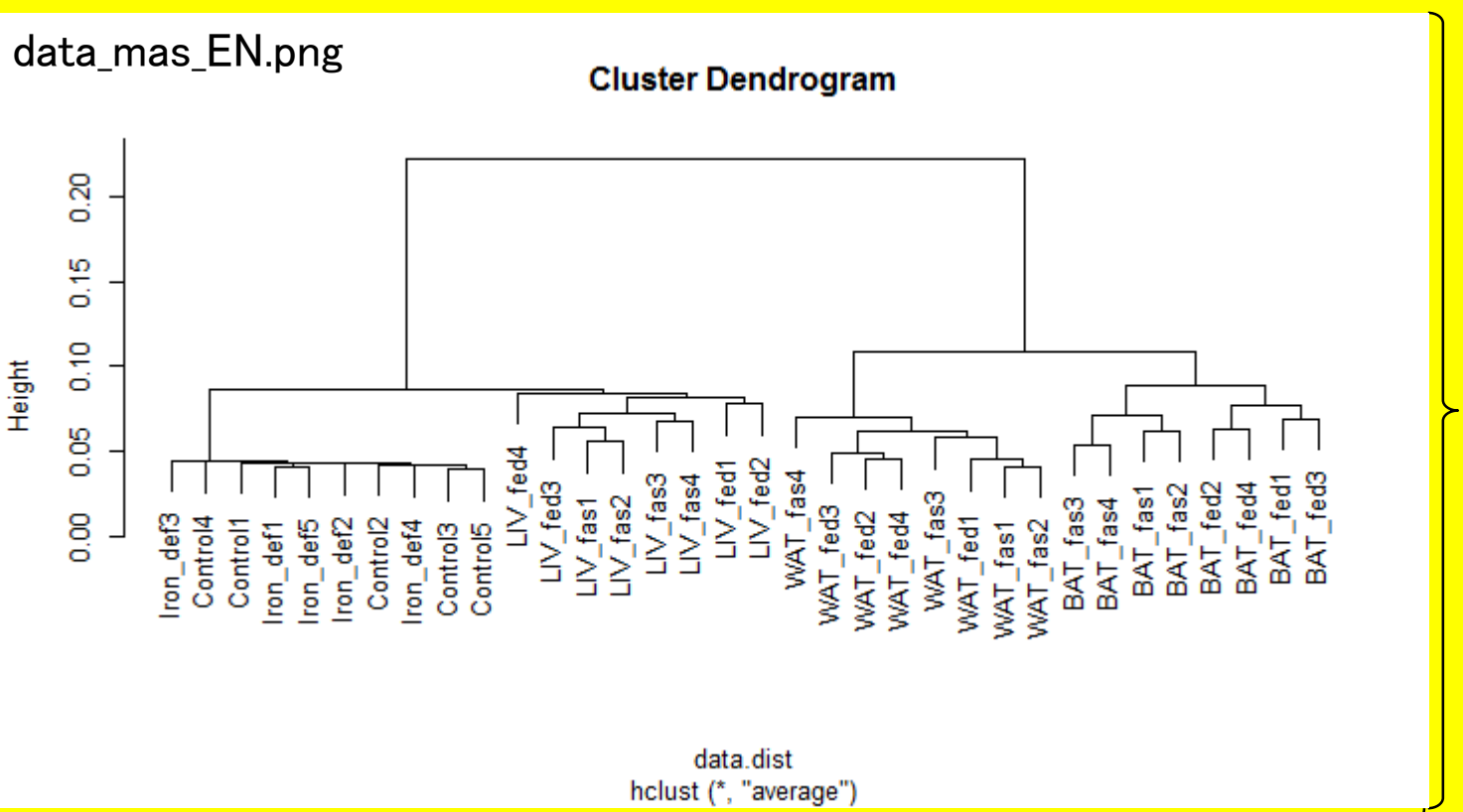


クラスタリング結果をファイルに保存

rcode_clustering_png.txt

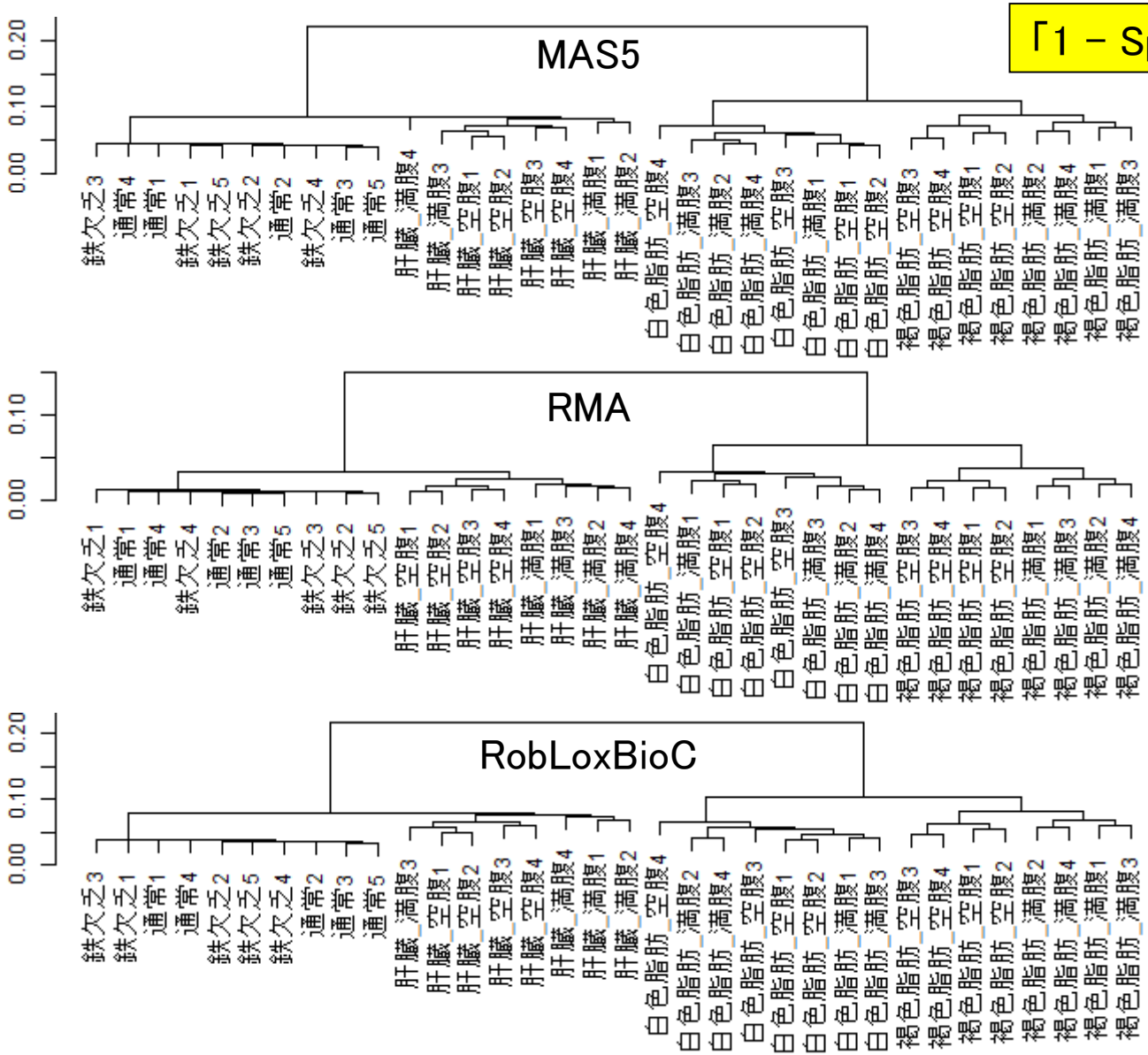
```
#################################################################################################↓
###  MAS5データのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###↓
#################################################################################################↓
in_f <- "data_mas_EN.txt"                #入力ファイル名を指定してin_fに格納↓
out_f <- "data_mas_EN.png"               #出力ファイル名を指定してout_fに格納↓
param <- "average"                       #方法(method)を指定↓
param_fig <- c(720, 400)
↓
#入力ファイルの読み込み↓
data <- read.table(in_f, heade
↓
#本番↓
data.dist <- as.dist(1 - cor(d
out <- hclust(data.dist, metho
↓
#ファイルに保存↓
png(out_f, pointsize=13, width
plot(out)
dev.off()
```

data_mas_EN.png



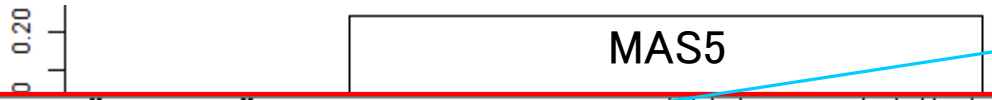
サンプル間クラスタリング（マージ後）

「1 − Spearman相関係数」の結果



rcode_clustering_merge.txt

サンプル間クラスタリング（マージ後）



「1 − Spearman相関係数」の結果

```
param1 <- "average"                    #方法(method)を指定↓
param2 <- "spearman"                   #相関係数の種類を指定↓
param_fig <- c(720, 310)               #ファイル出力時の横幅と縦幅を指定(単位はピクセル)↓
↓
##########################################################################################↓
###  MAS5データのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###↓
##########################################################################################↓
in_f <- "data_mas_JP.txt"              #入力ファイル名を指定してin_fに格納↓
out_f <- "data_mas_JP.png"             #出力ファイル名を指定してout_fに格納↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み
data.dist <- as.dist(1 - cor(data, method=param2))#サンプル間の距離を計算した結果をdata.distに格納↓
out <- hclust(data.dist, method=param1) #階層的クラスタリングを実行した結果をoutに格納↓
png(out_f, pointsize=12, 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↓
plot(out)                              #樹形図（デンドログラム）の表示↓
dev.off()                              #おまじない↓
↓
##########################################################################################↓
###  RMAデータのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###↓
##########################################################################################↓
in_f <- "data_rma_JP.txt"              #入力ファイル名を指定してin_fに格納↓
out_f <- "data_rma_JP.png"             #出力ファイル名を指定してout_fに格納↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み
data.dist <- as.dist(1 - cor(data, method=param2))#サンプル間の距離を計算した結果をdata.distに格納↓
out <- hclust(data.dist, method=param1) #階層的クラスタリングを実行した結果をoutに格納↓
```

rcode_clustering_merge.txt

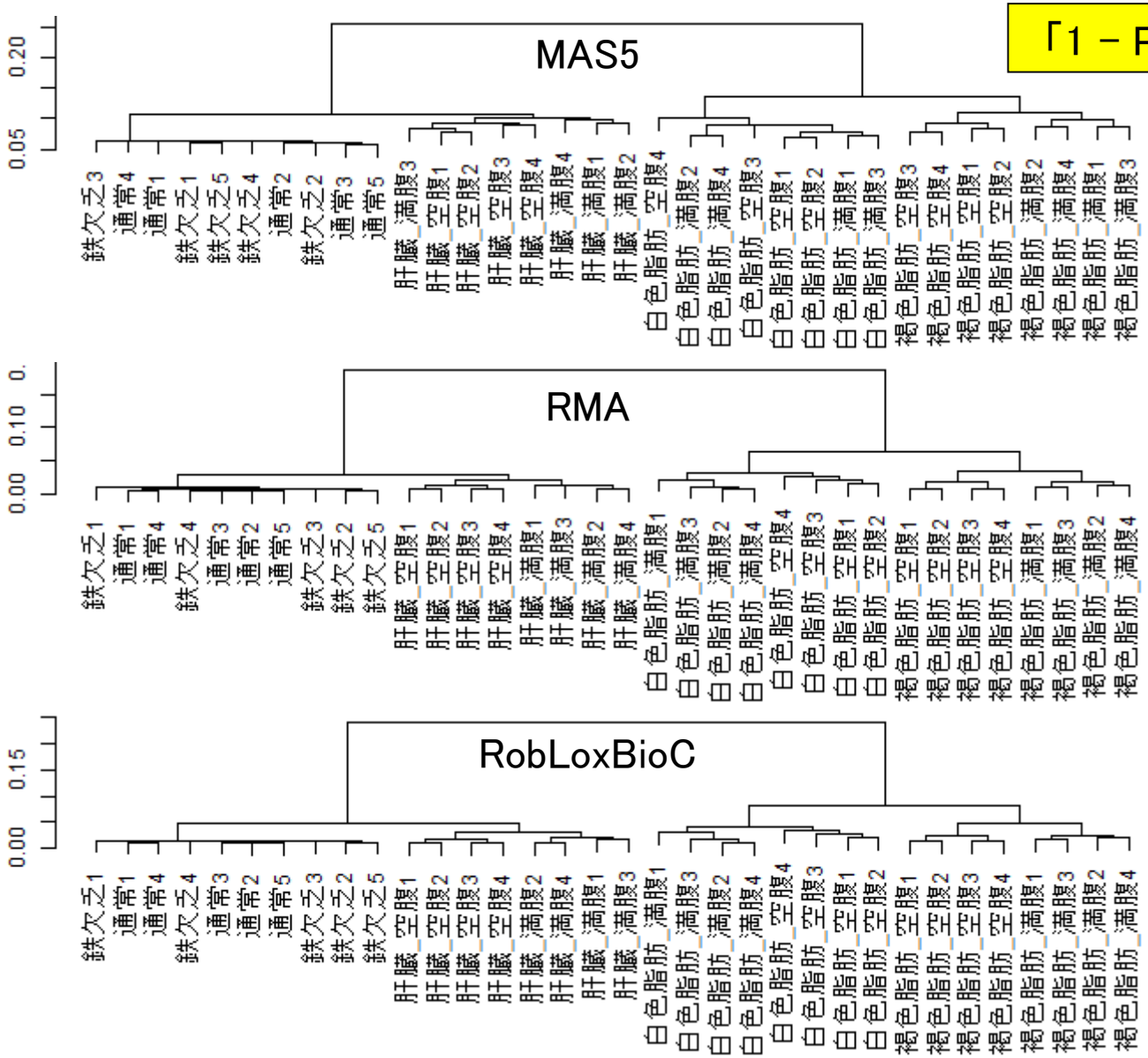
サンプル間クラスタリング（マージ後）

「1 – Pearson相関係数」の結果にしたいとき

```
param1 <- "average"                    #方法(method)を指定↓
param2 <- "pearson"                    #相関係数の種類を指定↓
param_fig <- c(720, 310)               #ファイル出力時の横幅と縦幅を指定(単位はピクセル)↓
↓
##########################################################################################↓
###  MAS5データのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###↓
##########################################################################################↓
in_f <- "data_mas_JP.txt"              #入力ファイル名を指定してin_fに格納↓
out_f <- "data_mas_JP.png"             #出力ファイル名を指定してout_fに格納↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み↓
data.dist <- as.dist(1 - cor(data, method=param2))#サンプル間の距離を計算した結果をdata.distに格納↓
out <- hclust(data.dist, method=param1) #階層的クラスタリングを実行した結果をoutに格納↓
png(out_f, pointsize=12, 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↓
plot(out)                              #樹形図（デンドログラム）の表示↓
dev.off()                              #おまじない↓
↓
##########################################################################################↓
###  RMAデータのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###↓
##########################################################################################↓
in_f <- "data_rma_JP.txt"              #入力ファイル名を指定してin_fに格納↓
out_f <- "data_rma_JP.png"             #出力ファイル名を指定してout_fに格納↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み↓
data.dist <- as.dist(1 - cor(data, method=param2))#サンプル間の距離を計算した結果をdata.distに格納↓
out <- hclust(data.dist, method=param1) #階層的クラスタリングを実行した結果をoutに格納↓
```

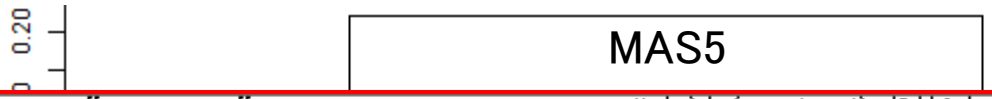
サンプル間クラスタリング（マージ後）

「1 − Pearson相関係数」の結果



rcode_clustering_merge.txt

サンプル間クラスタリング（マージ後）



Japanese → English

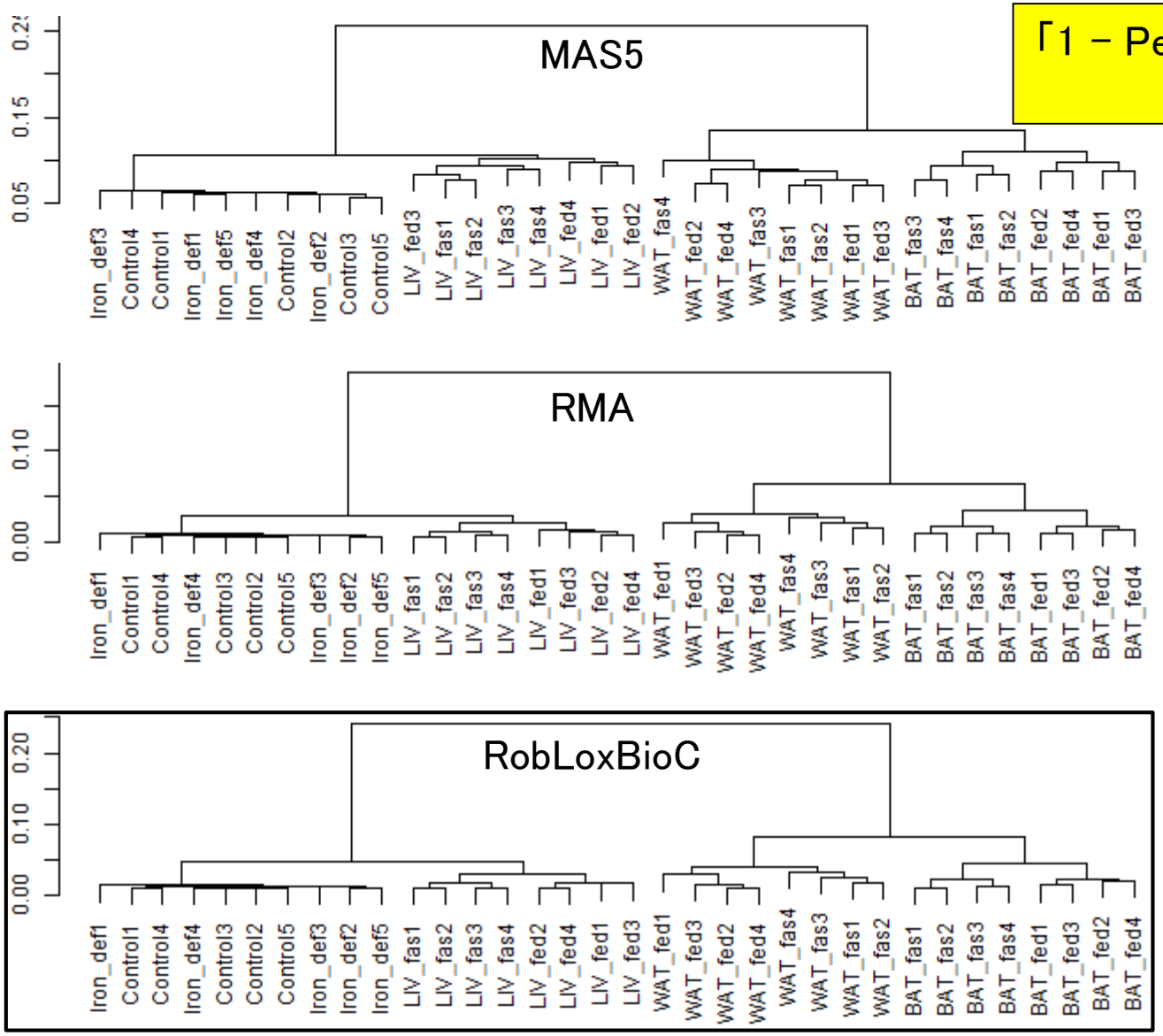
```
param1 <- "average"                    #方法(method)を指定
param2 <- "pearson"                    #相関係数の種類を指定
param_fig <- c(720, 310)               #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#########################################################################################
###  MAS5データのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###
#########################################################################################
in_f <- "data_mas_EN.txt"              #入力ファイル名を指定してin_fに格納
out_f <- "data_mas_JP.png"             #出力ファイル名を指定してout_fに格納
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み
data.dist <- as.dist(1 - cor(data, method=param2))#サンプル間の距離を計算した結果をdata.distに格納
out <- hclust(data.dist, method=param1) #階層的クラスタリングを実行した結果をoutに格納
png(out_f, pointsize=12, 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
plot(out)                              #樹形図（デンドログラム）の表示
dev.off()                              #おまじない

#########################################################################################
###  RMAデータのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###
#########################################################################################
in_f <- "data_rma_EN.txt"              #入力ファイル名を指定してin_fに格納
out_f <- "data_rma_JP.png"             #出力ファイル名を指定してout_fに格納
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み
data.dist <- as.dist(1 - cor(data, method=param2))#サンプル間の距離を計算した結果をdata.distに格納
out <- hclust(data.dist, method=param1) #階層的クラスタリングを実行した結果をoutに格納
```

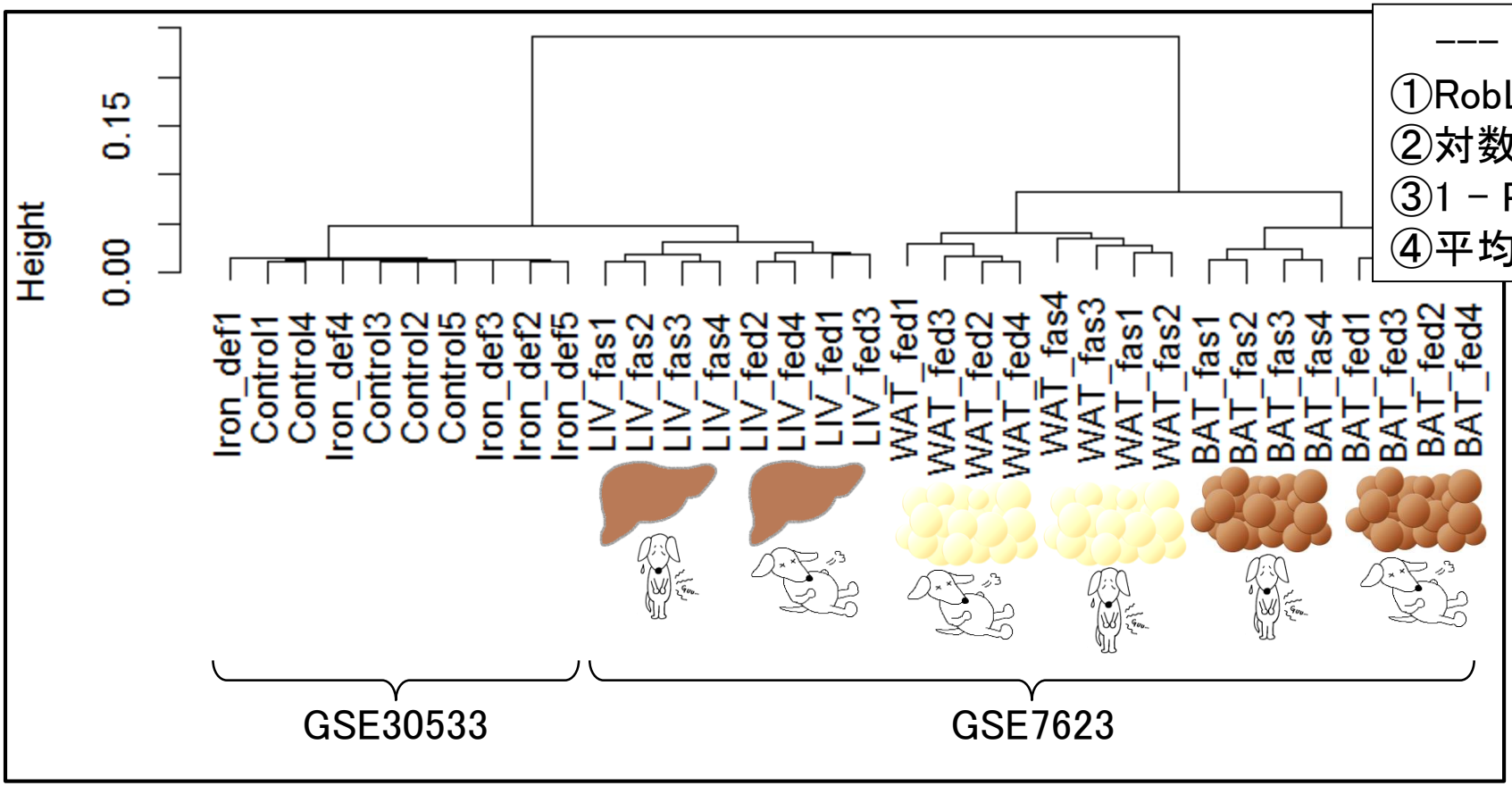
サンプル間クラスタリング（マージ後）

「1 − Pearson相関係数」の結果
English version



課題（クラスタリング結果の解釈）

- ラット（24サンプル＋10サンプル）のクラスタリング結果について考察せよ。

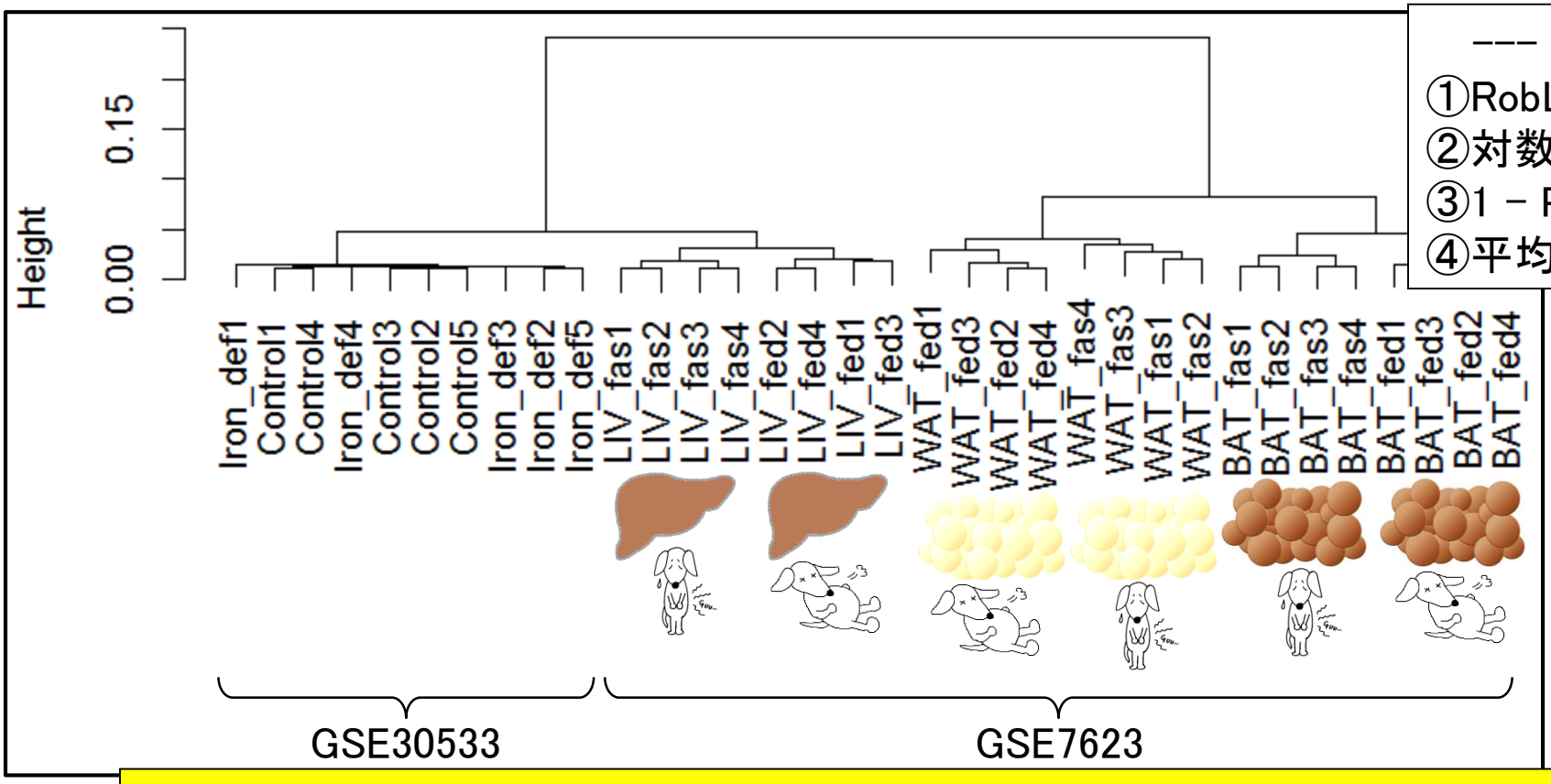


--- 解析手順 ---
①RobLoxBioC前処理法
②対数変換後のデータ
③1 − Pearson相関係数
④平均連結法

Q1：なぜGSE7623とGSE30533のデータはマージ可能か?
Q2：GSE30533の10サンプルからなるクラスターは、GSE7623の3種類の組織（LIV, WAT, BAT）のどの発現パターンに近いか?

課題（クラスタリング結果の解釈）

- ラット（24サンプル＋10サンプル）のクラスタリング結果について考察せよ。



--- 解析手順 ---
①RobLoxBioC前処理法
②対数変換後のデータ
③1 − Pearson相関係数
④平均連結法

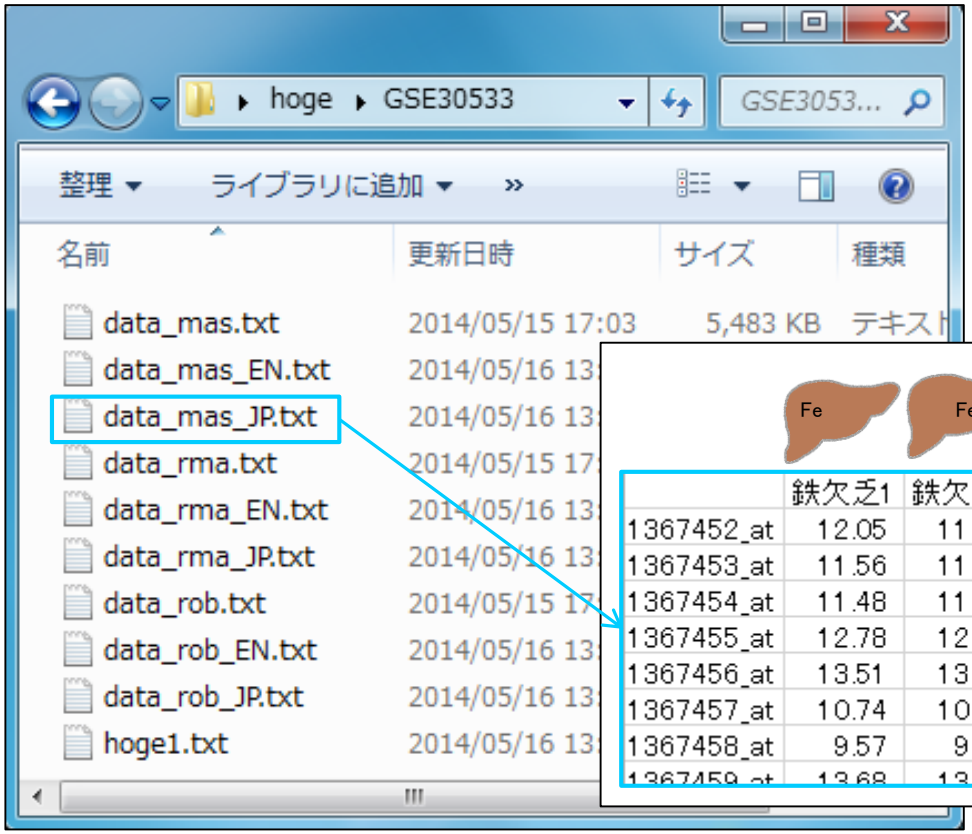
Q3：GSE30533のみのクラスタリング結果は「鉄欠乏（Iron_def）状態と通常（Control）状態」が入り混じっている。その一方で、「満腹（fed）状態と空腹（fas）状態」の違いは3種類の組織（LIV, WAT, BAT）いずれにおいても明瞭に分かれている。**鉄欠乏（Iron_def）状態と空腹（fas）状態の発現プロファイル変化への影響度はどちらか大きいと思われるか？**

実験デザイン（§3.2.2）

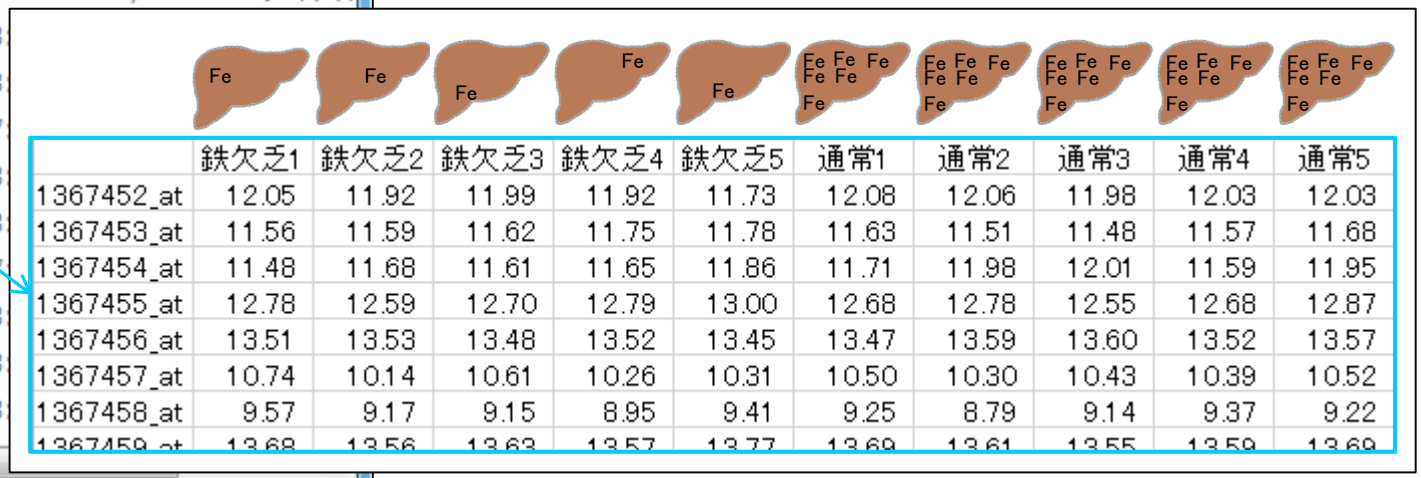
- Affymetrix GeneChip
  - Ge et al., *Genomics*, **86**: 127–141, 2005
    - **GSE2361**、GPL96 (Affymetrix Human Genome U133A Array)、22,283 probesets
    - **ヒト36サンプル**：Heart (心臓)、Thymus (胸腺)、Spleen (脾臓)、Ovary (卵巣)、Kidney (腎臓)、Skeletal Muscle (骨格筋)、Pancreas (膵臓)、Prostate (前立腺)、…
  - Nakai et al., *Biosci Biotechnol Biochem.*, **72**: 139–148, 2008　8匹のラットを使用
    - **GSE7623**、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
    - **ラット24サンプル**：Brown adipose tissue (褐色脂肪組織; BAT)8サンプル、White adipose tissue (白色脂肪組織; WAT)8サンプル、Liver (肝臓; LIV)8サンプル
      - BAT 8サンプル：通常（BAT_fed） 4サンプル　対　24時間絶食（BAT_fas） 4サンプル
      - WAT 8サンプル：通常（WAT_fed） 4サンプル　対　24時間絶食（WAT_fas） 4サンプル
      - LIV 8サンプル：通常（LIV_fed） 4サンプル　対　24時間絶食（LIV_fas） 4サンプル
  - Kamei et al., *PLoS One*, **8**: e65732, 2013　10匹のラットを使用
    - **GSE30533**、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
    - **ラット10サンプル**：全てLiver (肝臓)サンプル
    - iron-deficient diet (Iron_def) 5サンプル　対　control diet (Control) 5サンプル

実験デザイン（§3.2.2）

- Kamei et al., *PLoS One*, **8**: e65732, 2013
  - **GSE30533**、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
  - **ラット10サンプル**：全てLiver (肝臓) サンプル
  - iron-deficient diet (Iron_def) 5サンプル　対　control diet (Control) 5サンプル



2群間比較が主な目的であり、各群につき5反復(five replicates)とっている。生物学的なばらつき(biological variation)を考慮すべく、反復データは別々の個体からとっている(biological replicates)

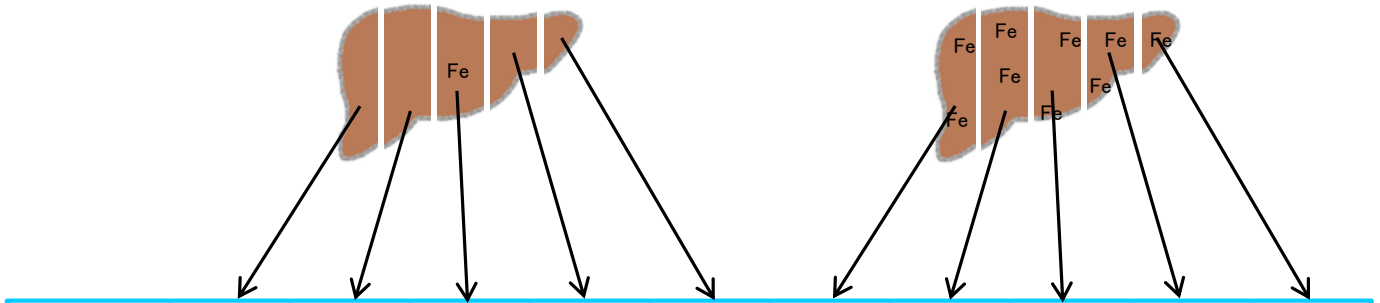


	鉄欠乏1	鉄欠乏2	鉄欠乏3	鉄欠乏4	鉄欠乏5	通常1	通常2	通常3	通常4	通常5
1367452_at	12.05	11.92	11.99	11.92	11.73	12.08	12.06	11.98	12.03	12.03
1367453_at	11.56	11.59	11.62	11.75	11.78	11.63	11.51	11.48	11.57	11.68
1367454_at	11.48	11.68	11.61	11.65	11.86	11.71	11.98	12.01	11.59	11.95
1367455_at	12.78	12.59	12.70	12.79	13.00	12.68	12.78	12.55	12.68	12.87
1367456_at	13.51	13.53	13.48	13.52	13.45	13.47	13.59	13.60	13.52	13.57
1367457_at	10.74	10.14	10.61	10.26	10.31	10.50	10.30	10.43	10.39	10.52
1367458_at	9.57	9.17	9.15	8.95	9.41	9.25	8.79	9.14	9.37	9.22

実験デザイン（§3.2.2）

- Kamei et al., *PLoS One* 8: e65732, 2013
  - GSE30533、 31,099 probesets
  - ラット10サン
  - iron-deficient diet（Iron_def）5サンプル 対 control diet（Control）5サンプル

対比的な用語は技術的なばらつき（technical variation）であり、同一個体由来サンプルを分割して得られた反復データ（technical replicates）



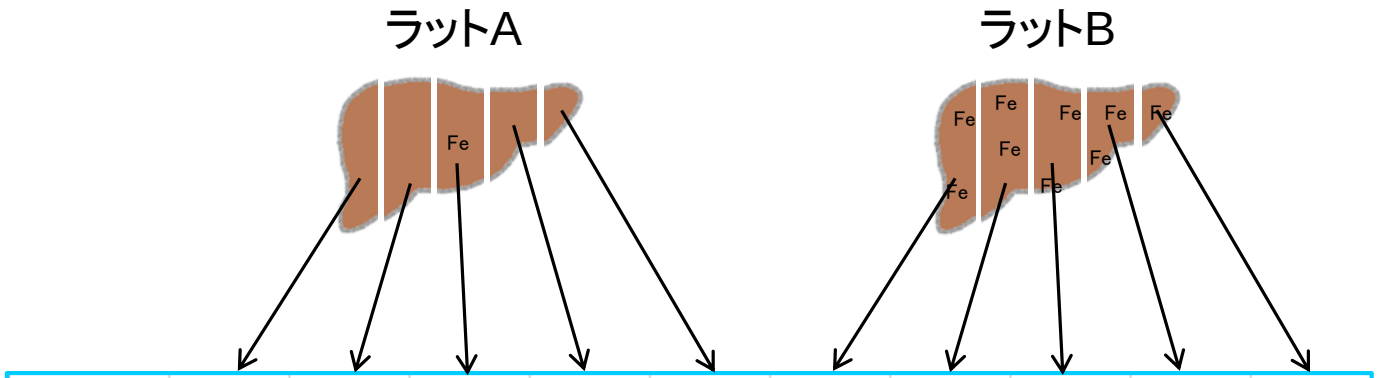
	鉄欠乏1	鉄欠乏2	鉄欠乏3	鉄欠乏4	鉄欠乏5	通常1	通常2	通常3	通常4	通常5
1367452_at	12.05	11.92	11.99	11.92	11.73	12.08	12.06	11.98	12.03	12.03
1367453_at	11.56	11.59	11.62	11.75	11.78	11.63	11.51	11.48	11.57	11.68
1367454_at	11.48	11.68	11.61	11.65	11.86	11.71	11.98	12.01	11.59	11.95
1367455_at	12.78	12.59	12.70	12.79	13.00	12.68	12.78	12.55	12.68	12.87
1367456_at	13.51	13.53	13.48	13.52	13.45	13.47	13.59	13.60	13.52	13.57
1367457_at	10.74	10.14	10.61	10.26	10.31	10.50	10.30	10.43	10.39	10.52
1367458_at	9.57	9.17	9.15	8.95	9.41	9.25	8.79	9.14	9.37	9.22
1367459_at	13.68	13.56	13.63	13.57	13.77	13.69	13.61	13.55	13.59	13.69

このやり方で得られる結論は限定的!できるだけ多様な別個体サンプルを沢山用いるべし!

実験デザイン（§3.2.2）

- Technical replicatesだと…
  1. 自分は「鉄欠乏　対　通常」の違いを見ているつもりでも、個体間の他の違い（身長、体重など）由来要因との区別がつかない

     高身長　対　低身長、低体重　対　高体重、他の病気の有無、家系の違いなど
  2. 得られる結果から導き出される結論は、そのラット間のみで成立する事象であり、ラットという生物種全体に適用可能なわけではない

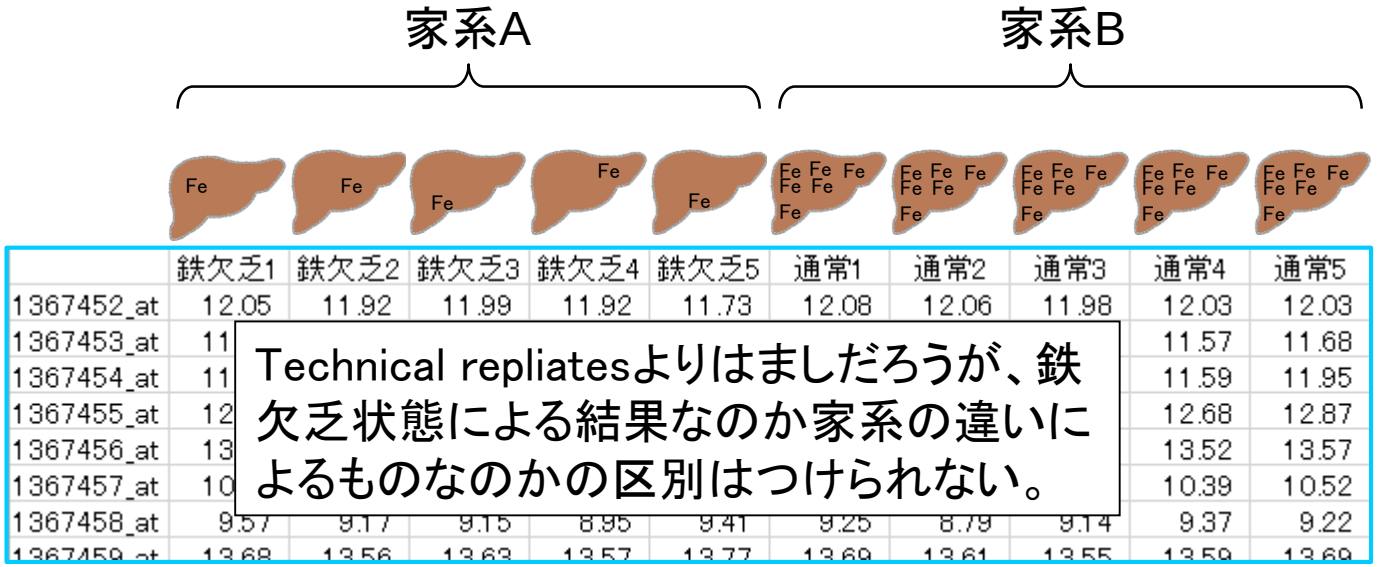


	鉄欠乏1	鉄欠乏2	鉄欠乏3	鉄欠乏4	鉄欠乏5	通常1	通常2	通常3	通常4	通常5
1367452_at	12.05	11.92	11.99	11.92	11.73	12.08	12.06	11.98	12.03	12.03
1367453_at	11.56	11.59	11.62	11.75	11.78	11.63	11.51	11.48	11.57	11.68
1367454										11.95
1367455										12.87
1367456_at										13.57
1367457_at	10.74	10.14	10.61	10.26	10.31	10.50	10.30	10.43	10.39	10.52
1367458_at	9.57	9.17	9.15	8.95	9.41	9.25	8.79	9.14	9.37	9.22
1367459_at	13.68	13.56	13.63	13.57	13.77	13.69	13.61	13.55	13.59	13.69

2群間での発現変動遺伝子検出結果は多くなる傾向

実験デザイン（§3.2.2）

- Biological replicatesでも多様性が不十分な場合はイマイチ…



Technical repliatesよりはましだろうが、鉄欠乏状態による結果なのか家系の違いによるものなのかの区別はつけられない。

普遍的な結果を得たいのなら、できるだけ多様な別個体サンプルを沢山用いるべし!
Expression Atlasも**3 biological replicates**以上を基本としているようだ。

まとめ

- プローブレベルデータ取得および前処理法適用（発展形）
  - Rパッケージ経由でダウンロードしたデータをそのまま前処理法にかける
  - サンプル名の順番が変わることもあるので注意
  - 何度も同じ作業をしなくていいように有効活用
  - 3つの論文の遺伝子発現行列をコピペで作成
    - ヒトの36サンプルからなるデータ（Ge et al., 2005）
    - ラットの24サンプルからなるデータ（Nakai et al., 2008）
    - ラットの10サンプルからなるデータ（Kamei et al., 2013）

6/4の組織特異的遺伝子検出で利用予定

5/28の発現変動遺伝子検出で利用予定

- サンプル間クラスタリング
  - 階層的クラスタリング、距離の定義、クラスターをまとめる方法など
  - 前処理法（MAS5, RMA, RobLoxBioC）と相関係数（Spearman and Pearson）の違い
  - 同一アレイ由来データセットのマージ（ラット24サンプル＋ラット10サンプル）
    - 3’発現アレイの長所（教科書p7）
    - ラット10サンプル（通常　対　鉄欠乏）クラスタリング結果の印象は、外群（ラット24サンプル）の有無でずいぶん異なる（教科書p106-107）
- 実験デザインは重要

クラスタリング結果とDEG数の関係

クラスタリング結果を眺めれば、発現変動遺伝子（DEG）数に関するおおよその見当がつきます
→ クラスタリングって重要