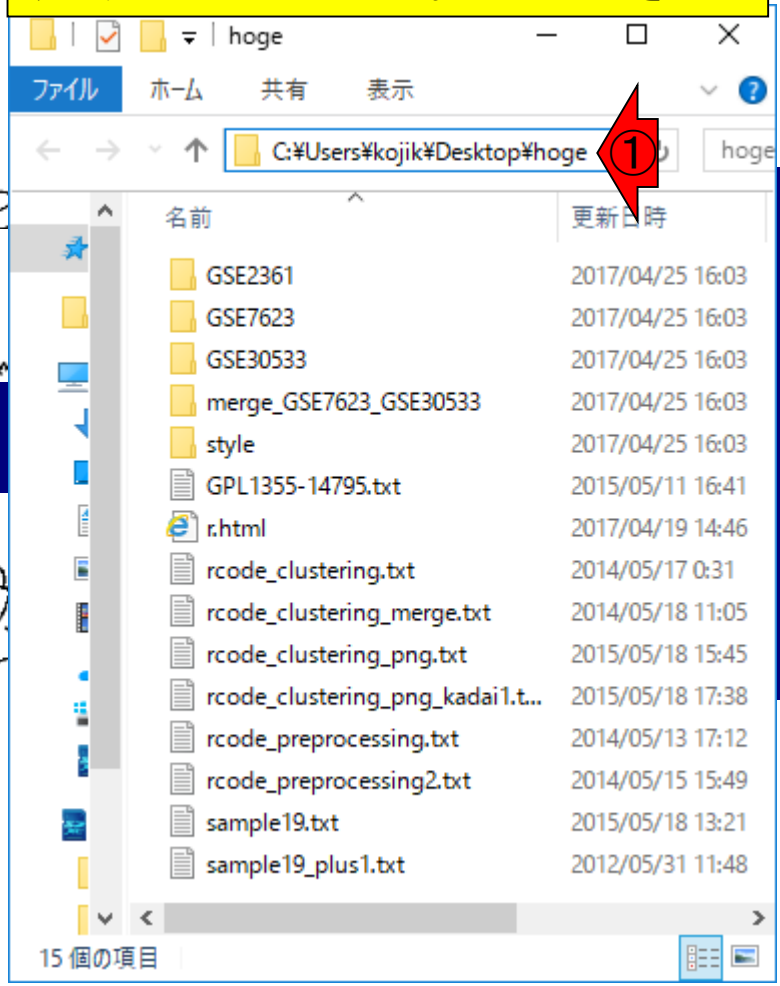
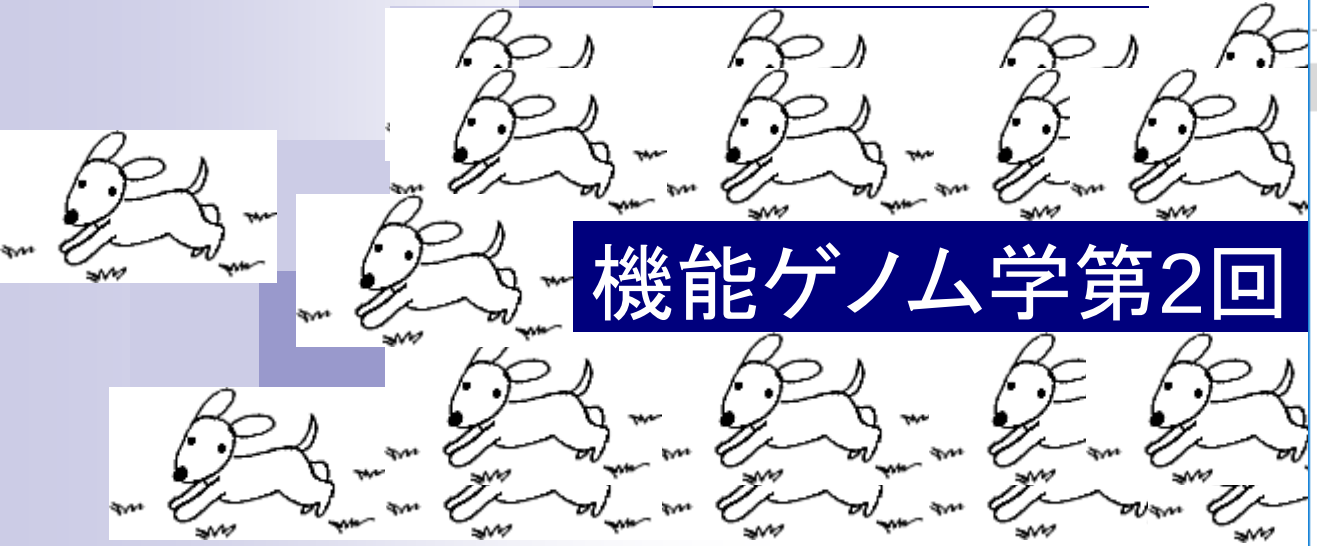
①持ち込みPCのヒトは後ろにある
USBメモリ中のhogeフォルダをデスクトップにコピーしておいてください





機能ゲノム学第2回

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

- 実データ概観
  - GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の§3.2.1)
  - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
  - 階層的 vs. 非階層的、様々な選択肢
  - 距離の定義:ベクトル間、クラスター間
  - 実データで実行:GSE2361 (ヒト)、**課題1**
  - 実データで実行:GSE7623 (ラット)、GSE30533 (ラット)
  - 同一プラットフォームデータ(GSE7623 + GSE30533)をマージして実行、**課題2**
- 実験デザイン(教科書の§3.2.2)
- 2群間比較:発現変動遺伝子(DEG)検出
  - パターンマッチング法(相関係数の利用)
    - コードの中身をおさらい、apply関数の基本的な利用法など

実データ概観

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)

- 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

実データ概観

- 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
  - 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。

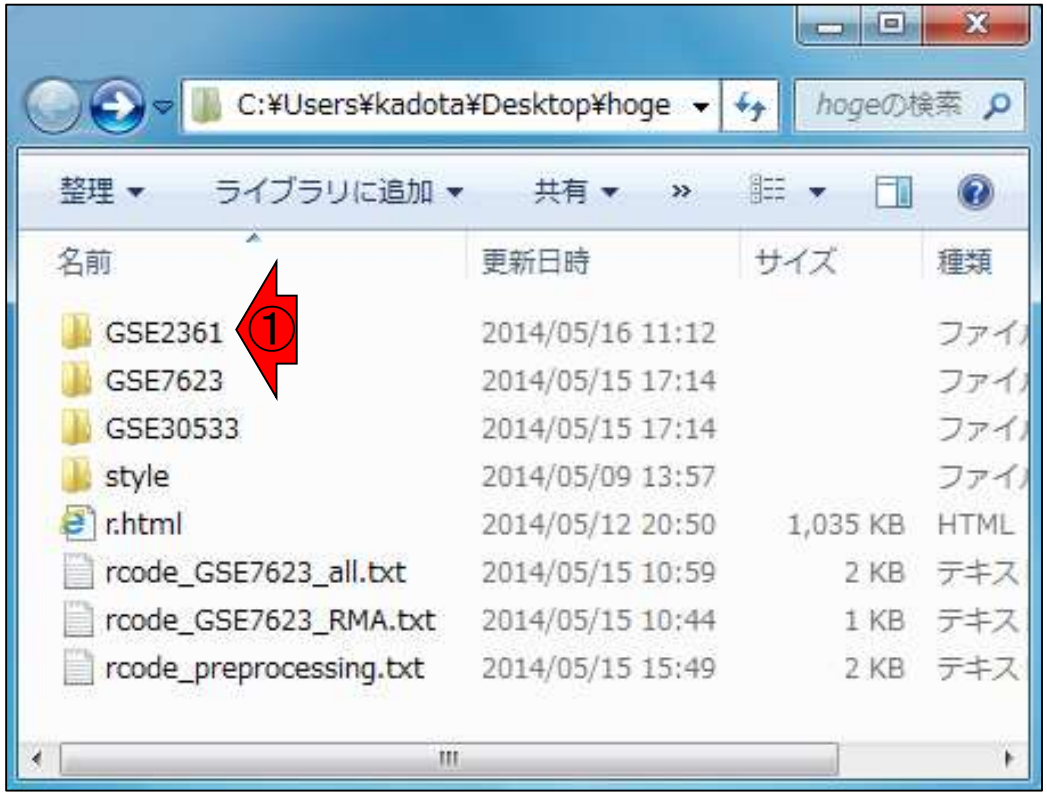
一次データベース

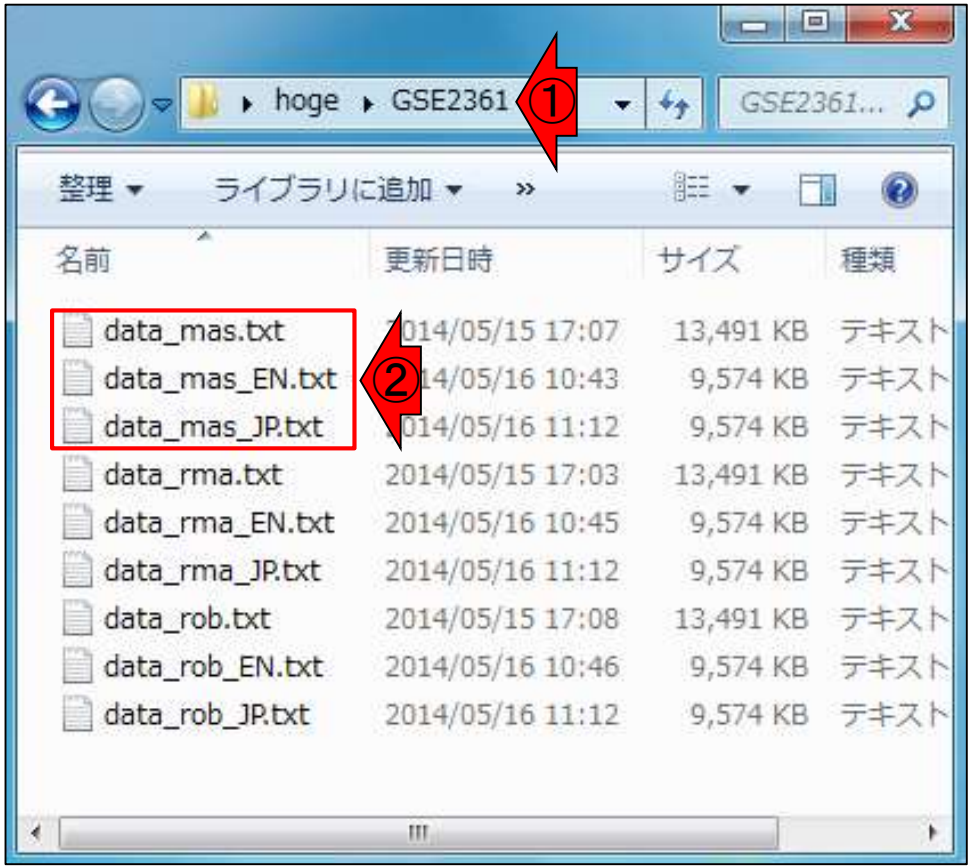
- GEO: Barrett et al., Nucleic Acids Res., 2013
  - GSE7623(ラット24サンプル, 62MB): Nakai et al., BBB, 2008
  - GSE30533(ラット10サンプル, 25MB): Kamei et al., PLoS One, 2013
  - GSE2361(ヒト36サンプル, 130MB): Ge et al., Genomics, 2005
  - GSE10246(マウス182サンプル, 1.1GB): Lattin et al., Immunome Res., 2008
  - GSE1133(ヒトとマウス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

GSE2361(ヒト)

①GSE2361フォルダの中身。②data_mas.txtは前処理法をそのまま適用した結果。*_EN.txtはサンプル名を英語で、*_JP.txtは日本語で書き換えたもの

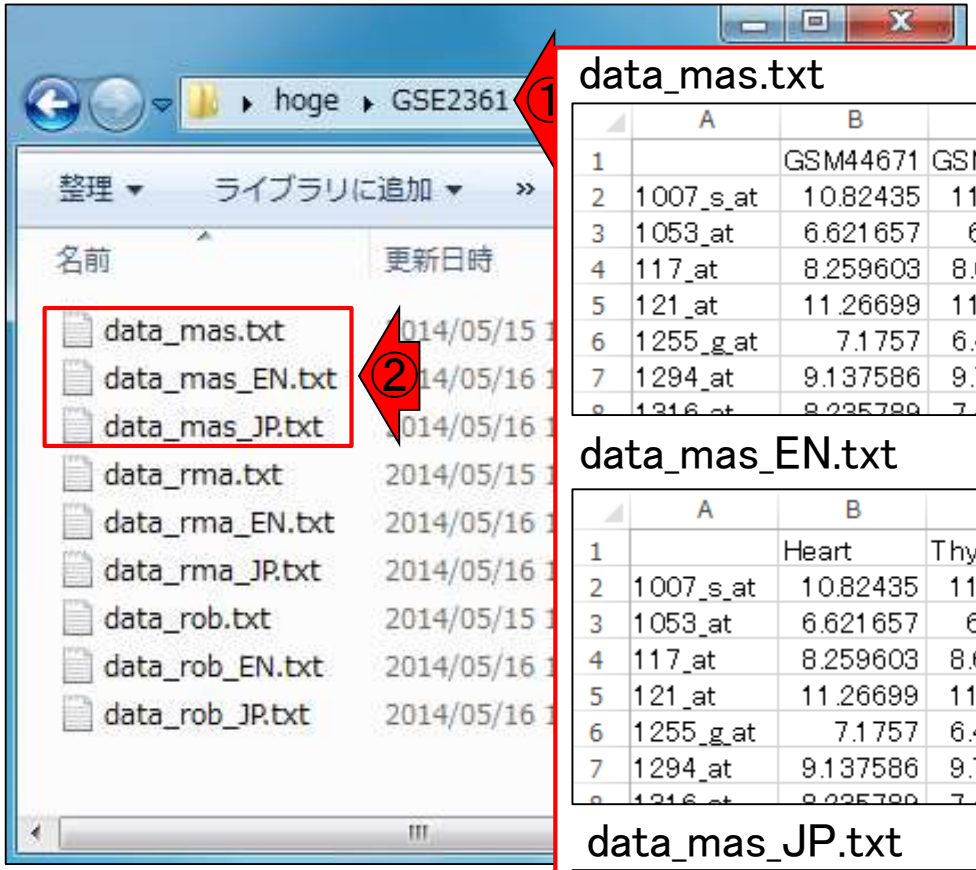
- Affymetrix GeneChip
  - Ge et al., *Genomics*, **86**: 127–141, 2005
    - **GSE2361**、GPL96 (Affymetrix Human Genome U133A Array)、22,283 probesets
    - **ヒト36サンプル**：Heart (心臓)、Thymus (胸腺)、Spleen (脾臓)、Ovary (卵巣)、Kidney (腎臓)、Skeletal Muscle (骨格筋)、Pancreas (膵臓)、Prostate (前立腺)、…





GSE2361(ヒト)

*_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。好きなものをご利用ください。いずれも**対数変換後のデータ**(log-transformed data)



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	11.4
3	1053_at	6.621657	6.47488	7.371465	4.168622	7.350244	7.209918	8.362777	7.176571	7.7
4	117_at	8.259603	8.647364	8.689724	7.875649	5.083001	8.165044	7.96707	8.418082	5.71
5	121_at	11.26699	11.04186	11.39999	11.02855	13.13267	11.39138	12.32899	11.0559	11.2
6	1255_g_at	7.1757	6.477278	6.781766	7.048799	7.30767	6.600267	6.42359	6.694412	7.3
7	1294_at	9.137586	9.718507	9.083742	9.014997	8.813377	8.402562	9.146946	8.421351	10.1

data_mas_EN.txt

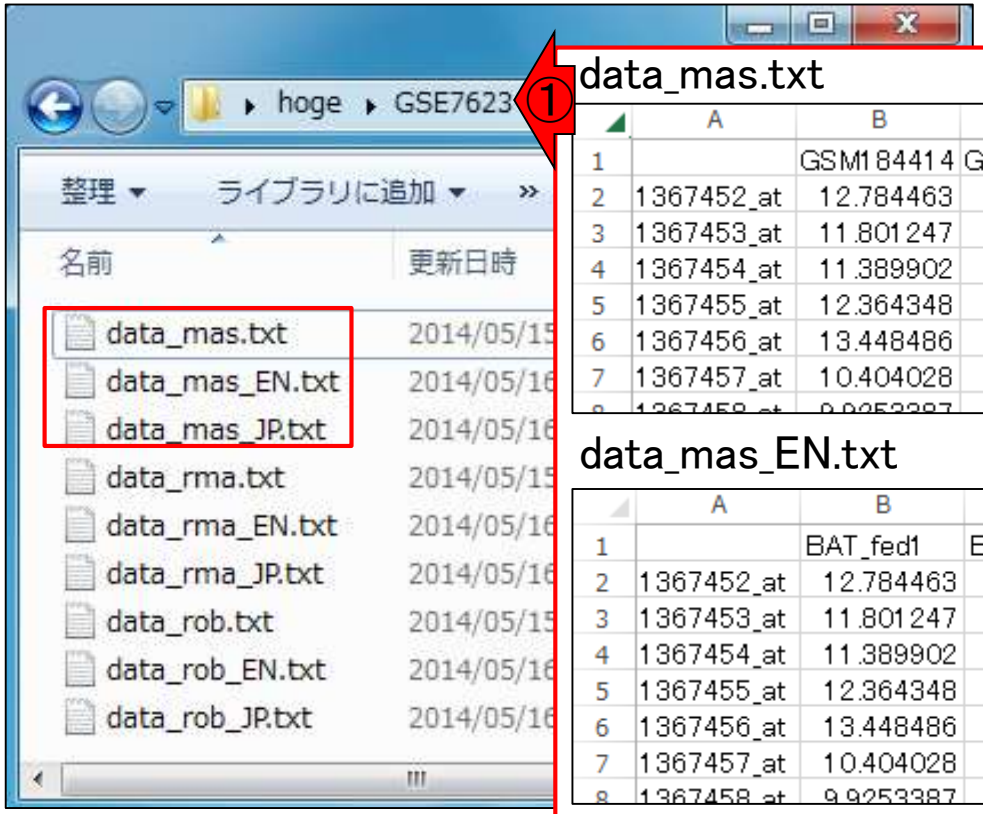
	A	B	C	D	E	F	G	H	I	J
1		Heart	Thymus	Spleen	Ovary	Kidney	Skeletal_Mu	Pancreas	Prostate	Small_I
2	1007_s_at	10.82435	11.21261	9.898072	10.8948	11.75531	10.92032	11.61243	11.22101	11.48
3	1053_at	6.621657	6.47488	7.371465	4.168622	7.350244	7.209918	8.362777	7.176571	7.75
4	117_at	8.259603	8.647364	8.689724	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

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

GSE7623(ラット)

①GSE7623 (Nakai et al., 2008)の対数変換後のデータ



①

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	12.43
3	1367453_at	11.801247	12.152935	11.942227	11.968477	11.845375	11.681727	12.078672	12.04
4	1367454_at	11.389902	11.160757	11.145987	11.212088	11.540652	11.308877	11.49885	11.40
5	1367455_at	12.364348	12.529744	12.432574	12.604011	12.441991	12.249935	12.281827	12.19
6	1367456_at	13.448486	13.543046	13.552794	13.629799	13.36913	13.244278	13.424371	13.32
7	1367457_at	10.404028	10.69632	10.475078	10.45579	10.141921	10.290666	10.146529	10.26

data_mas_EN.txt

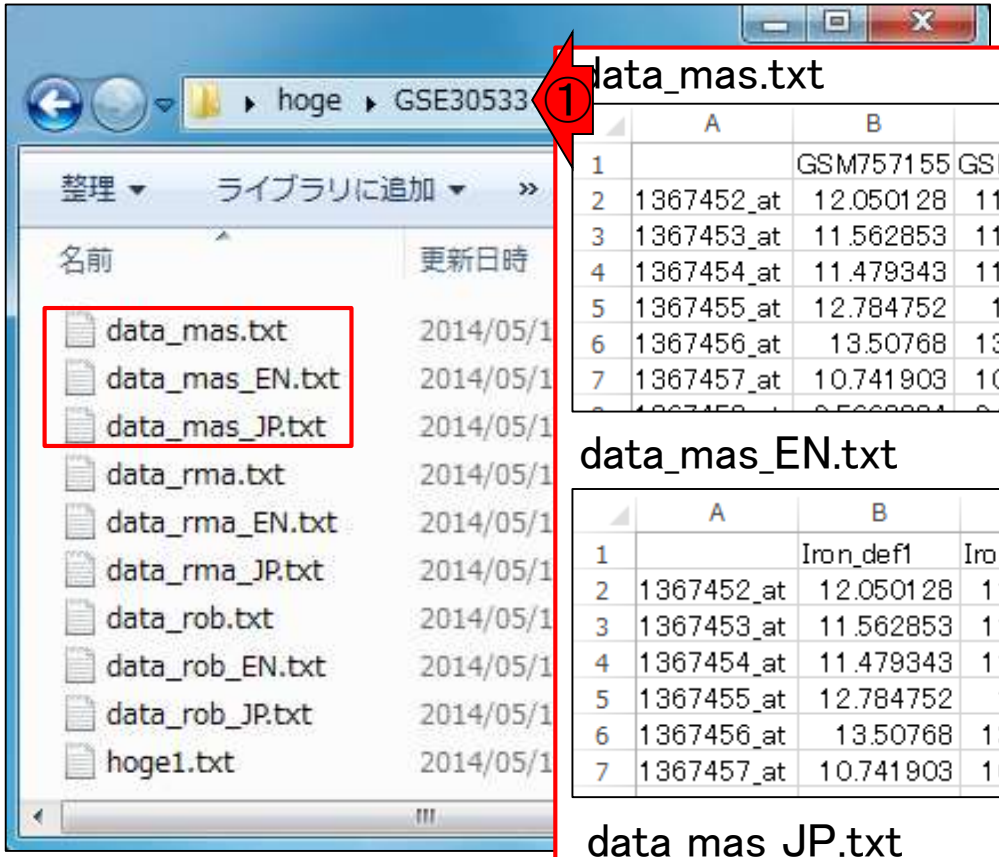
	A	B	C	D	E	F	G	H	I
1		BAT_fed1	BAT_fed2	BAT_fed3	BAT_fed4	BAT_fas1	BAT_fas2	BAT_fas3	BAT_fa
2	1367452_at	12.784463	12.447082	12.805908	12.304718	12.589425	12.607532	11.815378	12.43
3	1367453_at	11.801247	12.152935	11.942227	11.968477	11.845375	11.681727	12.078672	12.04
4	1367454_at	11.389902	11.160757	11.145987	11.212088	11.540652	11.308877	11.49885	11.40
5	1367455_at	12.364348	12.529744	12.432574	12.604011	12.441991	12.249935	12.281827	12.19
6	1367456_at	13.448486	13.543046	13.552794	13.629799	13.36913	13.244278	13.424371	13.32
7	1367457_at	10.404028	10.69632	10.475078	10.45579	10.141921	10.290666	10.146529	10.26

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

GSE30533（ラット）

①GSE30533 (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



①

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

Contents

- 実データ概観
  - GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の§3.2.1)
  - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
  - 階層的 vs. 非階層的、様々な選択肢
  - 距離の定義:ベクトル間、クラスター間
  - 実データで実行:GSE2361 (ヒト)、**課題1**
  - 実データで実行:GSE7623 (ラット)、GSE30533 (ラット)
  - 同一プラットフォームデータ(GSE7623 + GSE30533)をマージして実行、**課題2**
- 実験デザイン(教科書の§3.2.2)
- 2群間比較:発現変動遺伝子(DEG)検出
  - パターンマッチング法(相関係数の利用)
    - コードの中身をおさらい、apply関数の基本的な利用法など

対数変換の有無

①で用いている②hoge1.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 | 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 2014/05/16) 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

① ②

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
```

hoge1.txtは、①§2.2.2で作成

対数変換の有無

- 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) ① 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

p40の網掛け部分(上):

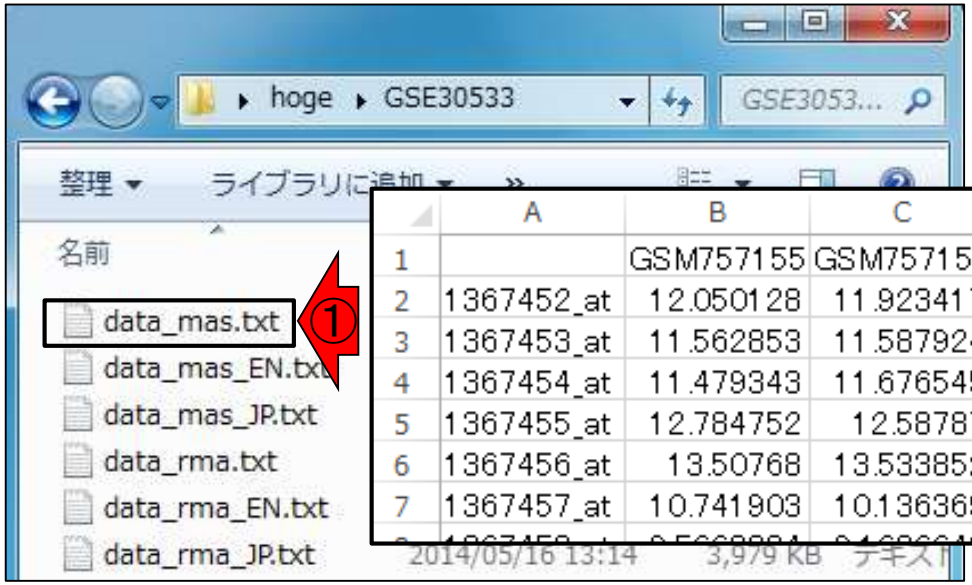
hoge1.txtと同じものができていると思います。

```
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
8	1367458_at	758.1	575.5	568.6	494.5	681.0	610.6	443.0	565

対数変換の有無

①data_mas.txtは、GSE30533 (Kamei et al., 2013)の、②対数変換（log_2変換）後のMAS5データ



	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

```
##########################################################↓
###  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)                       #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↓
↓
```

②

対数変換の有無

②hoge – GSE30533フォルダ中の③hoge1.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) ①modified 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

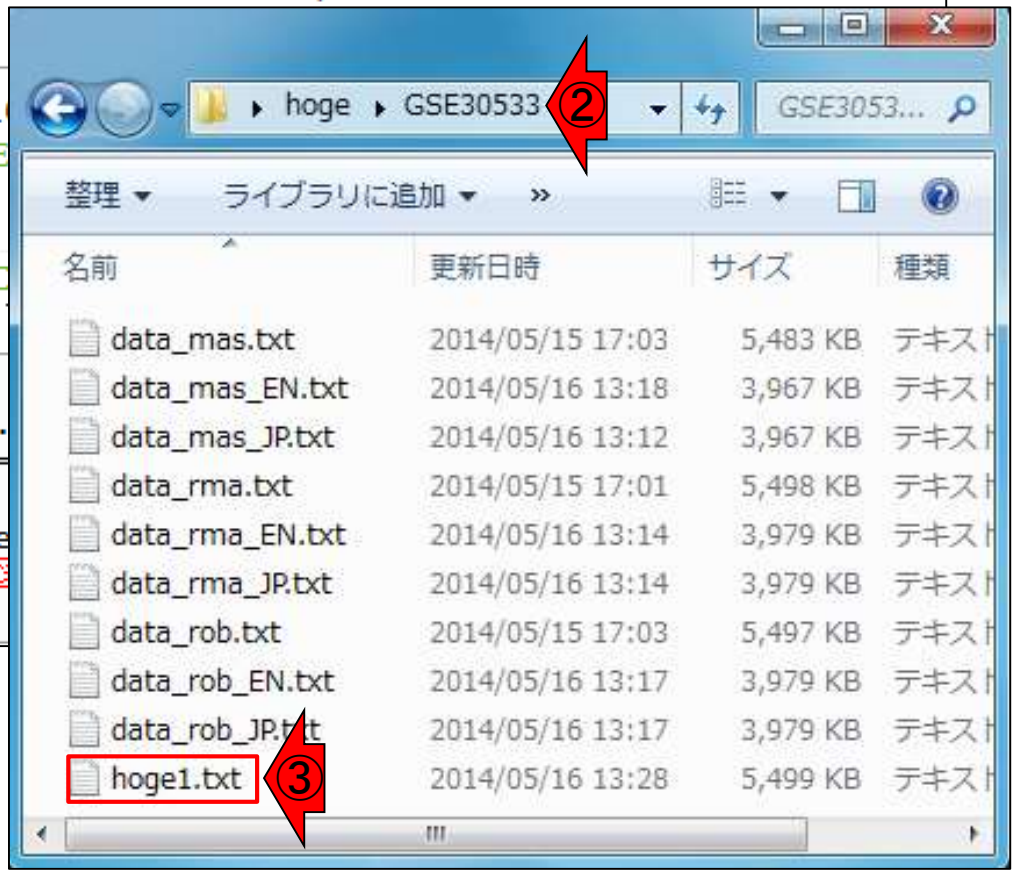
シリーズ Useful R 第7巻トランスクリプトーム解析のp99-1

「ファイル」-「ディレクトリの変更」でデスクトップ上の"E-GE

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
MAS5データファイル(hoge1.txt)を置いてあるディレクトリ

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



教科書p99

参考

Tips : list.files

①list.files()のみだと全ファイルになるが、②や③で示すように任意の文字列を含むファイル名のみにすることもできる

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, heade
colnames(data) <- c(paste("G1_
data.dist <- as.dist(1 - cor(
out <- hclust(data.dist, meth
plot(out)
```

R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/GSE30533"
> list.files()
 [1] "data_mas.txt"    "data_mas_EN.txt" "data_mas_JP.txt"
 [4] "data_rma.txt"    "data_rma_EN.txt" "data_rma_JP.txt"
 [7] "data_rob.txt"    "data_rob_EN.txt" "data_rob_JP.txt"
[10] "hoge1.txt"
> list.files(pattern="hoge")
[1] "hoge1.txt"
> list.files(pattern="EN")
[1] "data_mas_EN.txt" "data_rma_EN.txt" "data_rob_EN.txt"
>
```

①
②
③

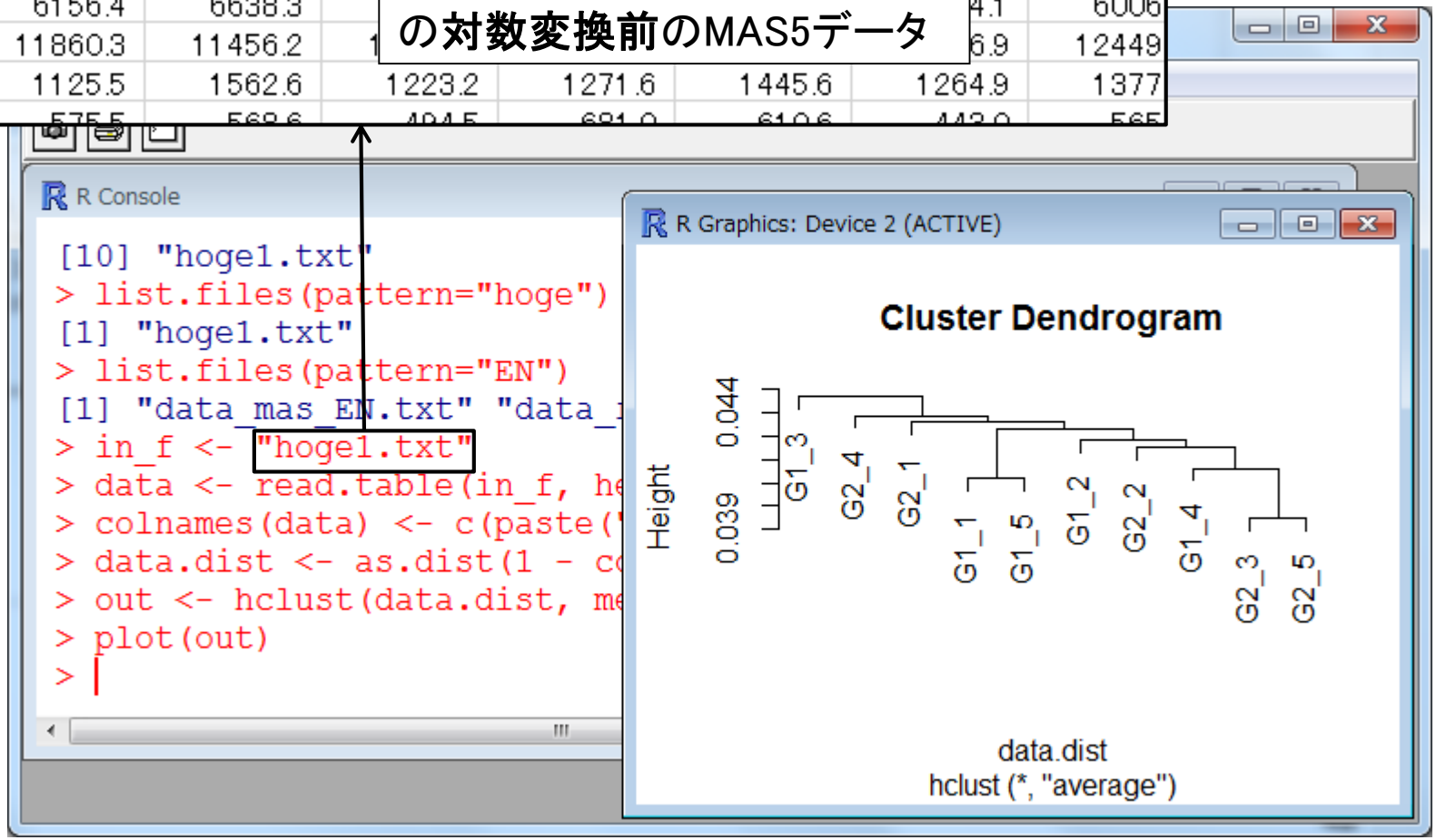
教科書p99

サンプル名のところが変わっている理由を説明します

対数変換の有無

	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

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



参考

①read.table関数で読み込んだ直後のdataオブジェクトの列名。②列名変更後

Tips

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
```

R Console

①

```
> 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"
```

②

```
> 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"
> |
```

参考

Tips

黒下線部分がG1群のサンプル名作成に相当する部分。①その部分のみ実行。②同じ結果だがやり方を微妙に変えている。③Iron_defに文字を変更

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
```

R Console

```
> 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"
① > paste("G1_", 1:5, sep="")
[1] "G1_1" "G1_2" "G1_3" "G1_4" "G1_5"
② > paste("G1", 1:5, sep="_")
[1] "G1_1" "G1_2" "G1_3" "G1_4" "G1_5"
③ > paste("Iron_def", 1:5, sep="")
[1] "Iron_def1" "Iron_def2" "Iron_def3" "Iron_def4" "Iron_def5"
> paste("Iron_def", 1:3, sep="")
[1] "Iron_def1" "Iron_def2" "Iron_def3"
> paste("Iron_def", c(2,4,5), sep="")
[1] "Iron_def2" "Iron_def4" "Iron_def5"
>
```

対数変換の有無

①GSE30533 (Kamei et al., 2013)の対数変換後のMAS5データ(data_mas_EN.txt)でもクラスタリングを行い、②対数変換前(hoge1.txt)の結果と比較する

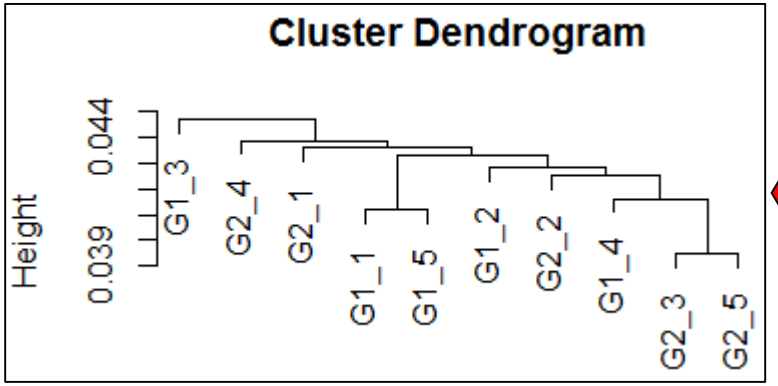
rcode_clustering.txt

```
#####################################↓
###  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)↓
←
```

①

	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	1		0589	11.975747	12.0095	
5	1367455_at	12.784752	12.58787	1		3645	12.780148	12.5522	
6	1367456_at	13.50768	13.533852	1		2369	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



②

対数変換の有無

①対数変換後の結果。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

rcode_clustering.txt

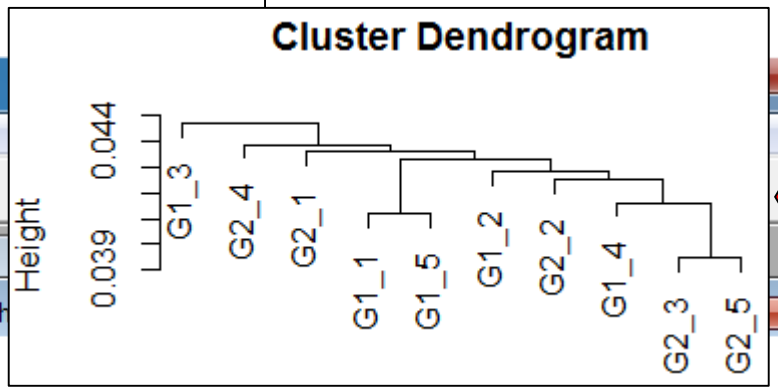
```
#####################################↓
###  MAS5データのクラスタリング  ###↓
#####################################↓
in_f <- "data_mas_EN.txt"↓
data <- read.table(
#colnames(data) <-
data.dist <- as.dis
out <- hclust(data.
plot(out)↓
←
```

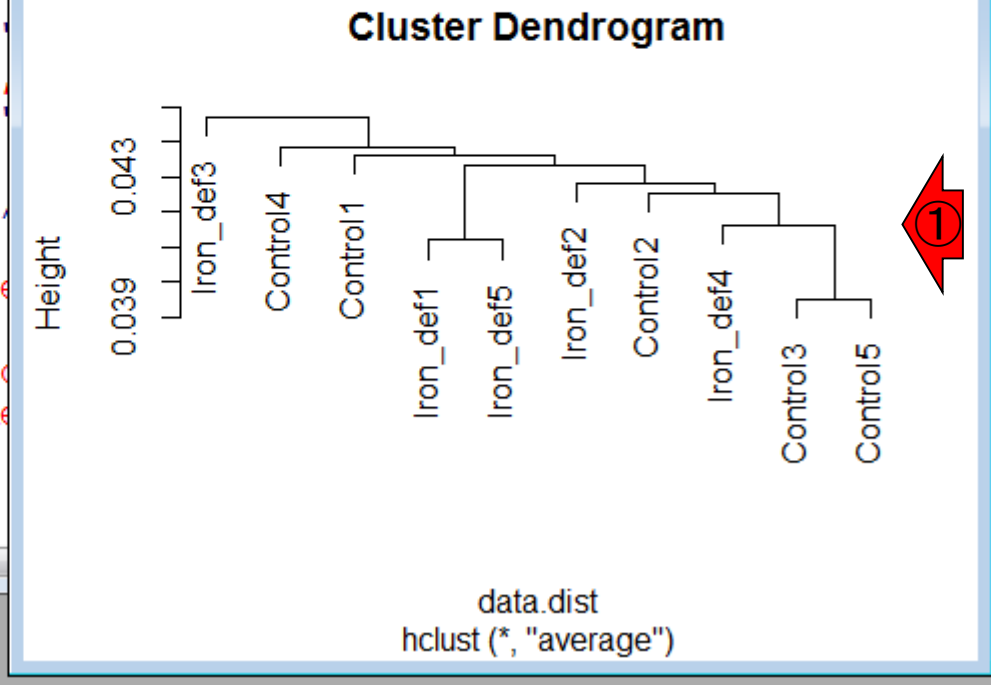
RGui (64-bit)

ファイル　履歴　サイズ変更　ウインドウ

R Console

```
[1] "Iron_def1" "Iron_def2" "
> paste("Iron_def", 1:3, sep=
[1] "Iron_def1" "Iron_def2" "
> paste("Iron_def", c(2,4,5),
[1] "Iron_def2" "Iron_def4" "
> getwd()
[1] "C:/Users/kadota/Desktop/
> in_f <- "data_mas_EN.txt"
> data <- read.table(in_f, he
> #colnames(data) <- c(paste
> data.dist <- as.dist(1 - co
> out <- hclust(data.dist, me
> plot(out)
>
```





Contents

- 実データ概観
  - GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の § 3.2.1)
  - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
  - 階層的 vs. 非階層的、様々な選択肢
  - 距離の定義:ベクトル間、クラスター間
  - 実データで実行:GSE2361 (ヒト)、**課題1**
  - 実データで実行:GSE7623 (ラット)、GSE30533 (ラット)
  - 同一プラットフォームデータ(GSE7623 + GSE30533)をマージして実行、**課題2**
- 実験デザイン(教科書の § 3.2.2)
- 2群間比較:発現変動遺伝子(DEG)検出
  - パターンマッチング法(相関係数の利用)
    - コードの中身をおさらい、apply関数の基本的な利用法など

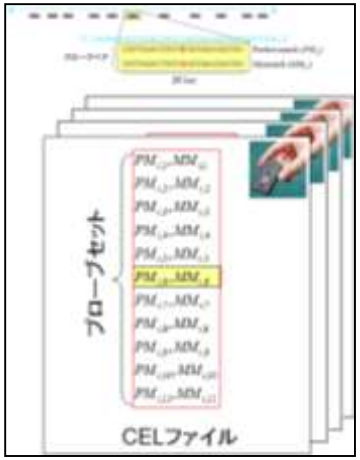
どれを使うかはほぼ趣味の問題

階層的 vs. 非階層的

- 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）

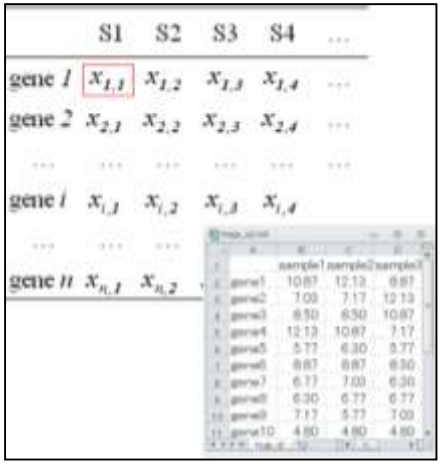
得られる樹形図の可能性は無数

様々な選択肢



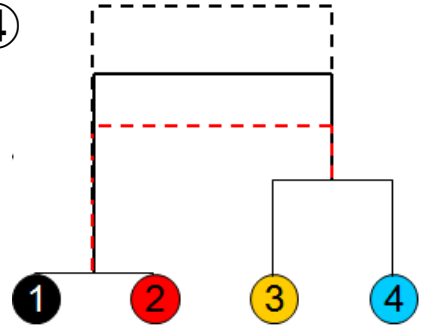
生データファイル

①前処理法



遺伝子発現行列②

クラスタリング③と④



①前処理法	×	②スケーリング	×	③距離	×	④群の併合
・MAS5		**・対数変換**		**・1ー相関係数**		・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
…		…		…		…

様々な選択肢

- 決めておくべき2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

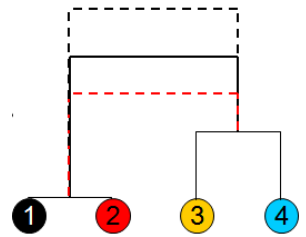
得られた結果の妥当性を何らかの知見に基づいて評価するため、結果の正当性を主張する視点が複数存在しうる。私は、「外れサンプルのチェック」や「発現変動遺伝子の有無や数」の見当をつける目的で行う

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



Danielsson et al., *Brief Bioinform.*, **16**: 941-949, 2015
Tang et al., *BMC Bioinformatics*, **16**: 361, 2015

「1－相関係数」を**距離**として定義することができます

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 → $r\approx -1$

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

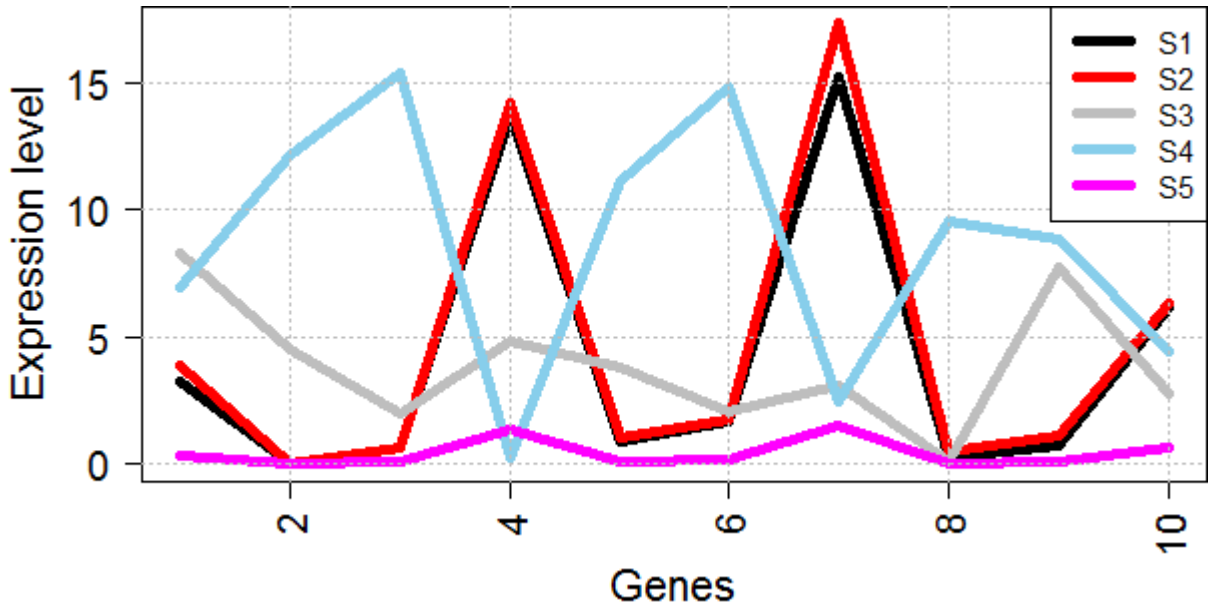
- $r=1 \rightarrow D=1-1=0$
- $r=0 \rightarrow D=1-0=1$
- $r=-1 \rightarrow D=1-(-1)=2$

相関係数 → 距離（計算例）

パターンが似ていれば0に近い値、逆パターンに近ければ最大値の2に近い値になっていることが分かる

- ベクトルxとyの発現パターンの距離$D(x,y)$

	S^1	S^2	S^3	S^4	S^5
g1	3.24	3.89	8.27	6.93	0.32
g2	0.01	0.03	4.55	12.17	0
g3	0.65	0.69	1.98	15.39	0.07
g4	13.73	14.21	4.83	0.28	1.38
g5	0.89	1.05	3.84	11.16	0.09
g6	1.65	1.74	2.11	14.82	0.17
g7	15.21	17.33	3.13	2.49	1.51
g8	0.26	0.52	0.08	9.53	0.03
g9	0.73	1.11	7.76	8.88	0.07
g10	6.18	6.36	2.81	4.47	0.62



$$\text{相関係数 } r_{S^1S^2} = 0.998 \rightarrow \text{距離} D_{S^1S^2} = 1 - 0.998 = 0.002$$

$$\text{相関係数 } r_{S^1S^3} = 0.035 \rightarrow \text{距離} D_{S^1S^3} = 1 - (0.035) = 0.965$$

$$\text{相関係数 } r_{S^1S^4} = -0.851 \rightarrow \text{距離} D_{S^1S^4} = 1 - (-0.851) = 1.851$$

相関係数 → 距離

②の赤枠全体をコピペで実行。③距離計算結果を格納したdata.distの中身を眺める

①

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。

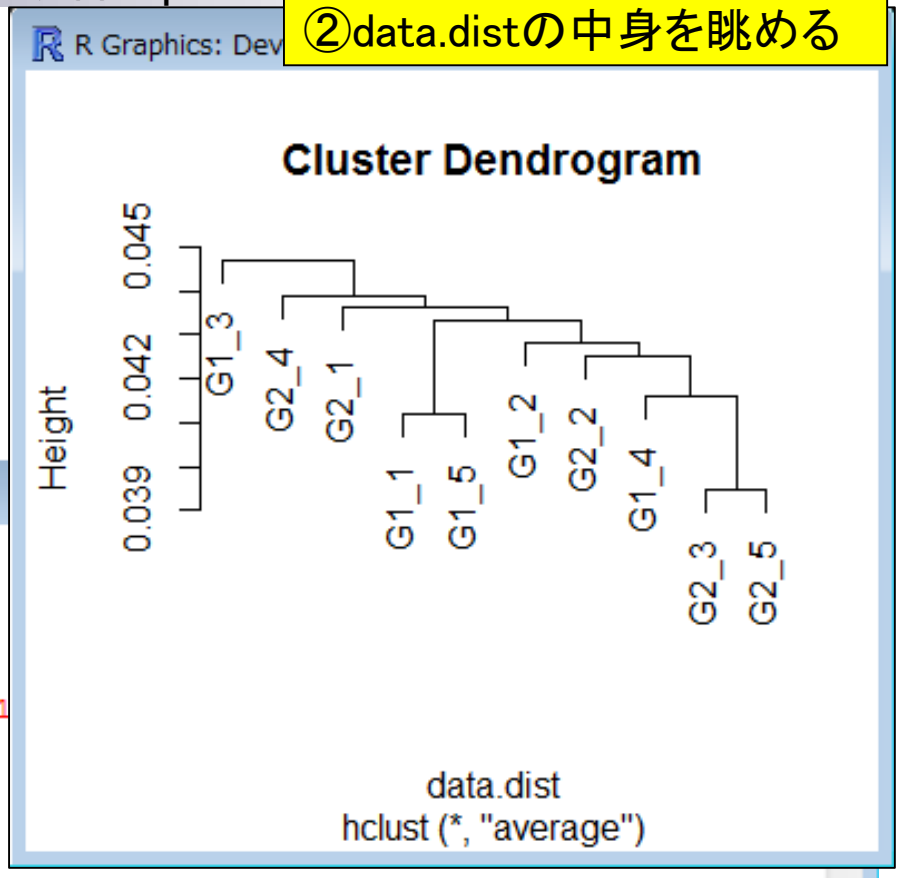
```
in_f <- "hoge1.txt"
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))
data.dist <- as.dist(1 - cor(data, method = "spearman"))
out <- hclust(data.dist, method = "average")
plot(out)                    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

③ ②

教科書p99

相関係数 → 距離

①距離計算結果を格納した
②data.distの中身を眺める

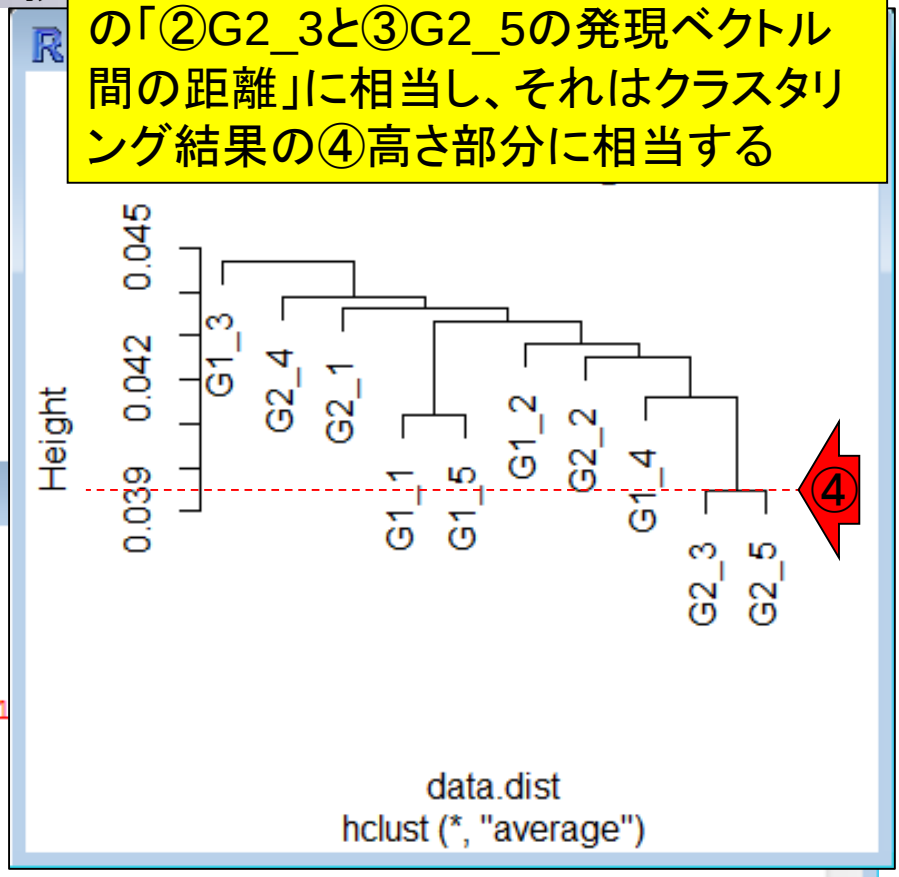


R Console

```
> 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_
> 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
>
```

相関係数 → 距離

①の0.03949559がクラスタリング結果の「②G2_3と③G2_5の発現ベクトル間の距離」に相当し、それはクラスタリング結果の④高さ部分に相当する



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
> 
```

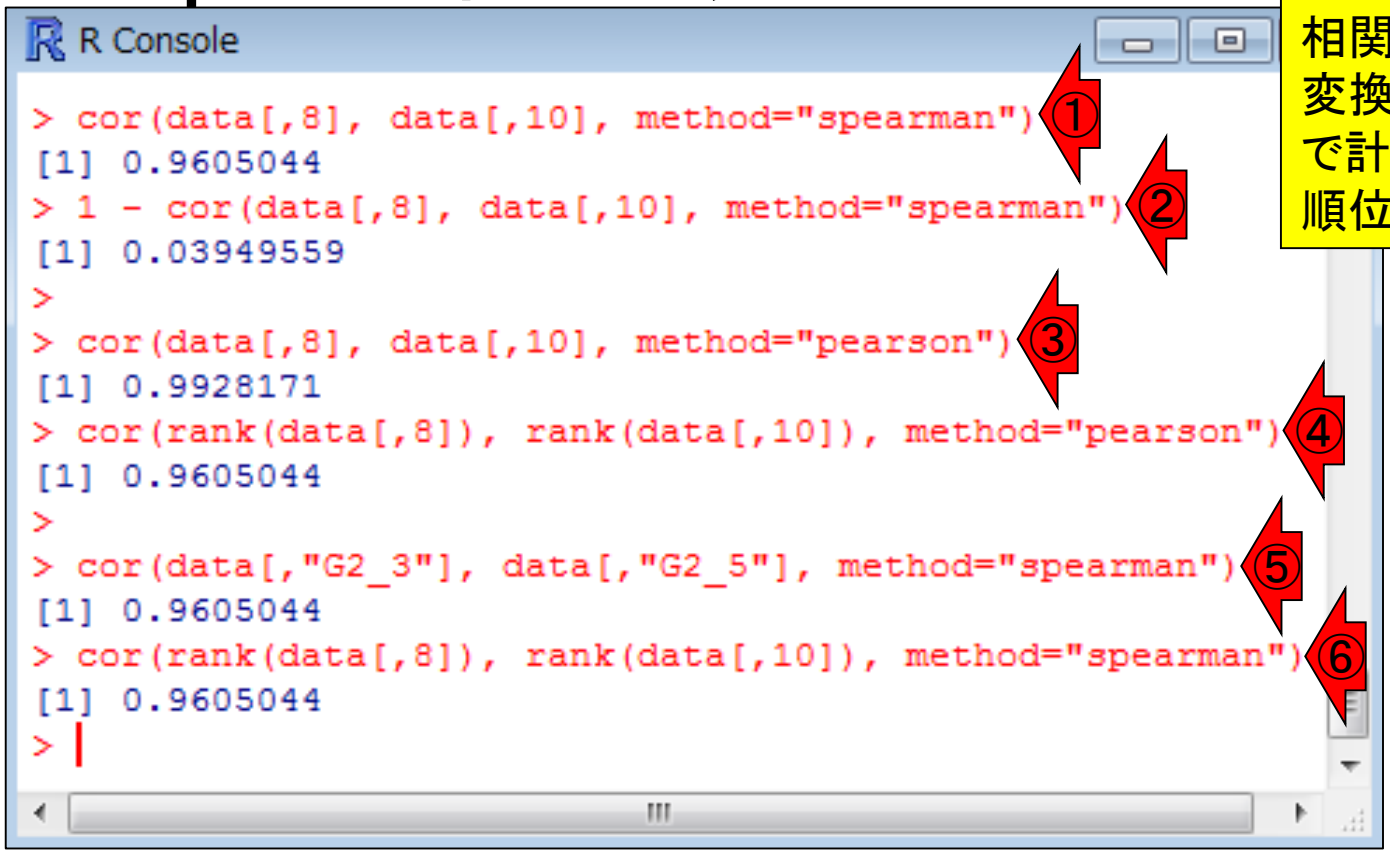
②

③

①

Tips：相関係数

教科書p100　参考



Spearman相関係数とPearson相関係数の関係。①Spearman相関係数、②1 − Spearman相関係数、③Pearson相関係数、④rank関数を用いて順位変換後のPearson相関係数、⑤列名で計算することもできるというTips、⑥順位変換後のSpearman相関係数

教科書p100-106

他の類似性尺度

- 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 $D = \sum_{i=1}^{n} \frac{| x_i - y_i |}{| x_i + y_i |}$
  - …

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。**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**

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

参考

計算例（サンプルxとy間の距離D）

名前が仰々しいだけで計算自体は大したことありません

	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

• 解析 | 基礎 | ベクトル間の距離

参考

①実際に自分で確認したいヒトはこちら

?関数名

(Rで)マイクロアレイデータ解析

(last modified 2015/05/16, since 2005)

What's new

- 門田幸二
  イ解析に
  どについ
  このページ
- お知らせ
  イ関連の
  ら辿れま

- 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 | ID変換 | について(last modified 2014/06/03)
- 前処理 | ID変換 | probe ID --> gene symbol(last modified 2014/06/03)
- 前処理 | ID変換 | probe ID --> Entrez ID(las
- 前処理 | ID変換 | probe ID --> その他(last m
- 前処理 | ID変換 | 同じ遺伝子名を持つもの
- 解析 | 基礎 | 共通遺伝子の抽出(last modifi
- 解析 | 基礎 | ベクトル間の距離(modifie
- 解析 | 基礎 | 遺伝子ごとの各種要約統計量
- 解析 | 基礎 | 最大発現量を示す組織の同定
- 解析 | 基礎 | 似た発現パターンを持つ遺伝
- 解析 | 基礎 | 平均-分散プロット (last modifie
- 解析 | クラスタリング | 階層的 | について (la

①

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

②

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

```
in_f <- "sample19.txt"                    #入力ファイル名を指定してin_fに格納

#データファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#i

#本番
dist(t(data), method="euclidean")         #ユークリッド(Euclidean)距離
dist(t(data), method="manhattan")         #マンハッタン(Manhattan)距離
dist(t(data), method="maximum")           #チェビシェフ(Chebyshev)距離
dist(t(data), method="canberra")          #キャンベラ(Canberra)距離
1 - cor(data, method="pearson")           #1 - Pearson相関係数

dist(t(data), method="binary")            #ハミング(Hamming)距離
dist(t(data), method="minkowski")         #ミンコフスキー(Minkowski)距離
1 - cor(data, method="spearman")          #1 - Spearman相関係数
```

参考

①dist関数で、他にどんな距離を利用可能か調べたい場合は、「?dist」

?関数名

解析 | 基礎 | ベクトル間の距離

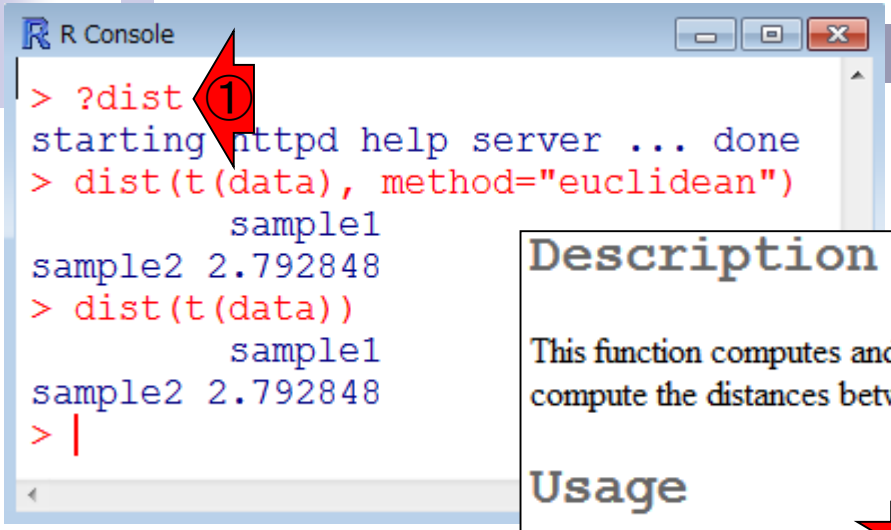
二つのベクトル間の距離を定義する方法は多数存在します。ここでは10 genes ×2 samplesのデータファイル(sample19.txt)を読み込んで二つのサンプル間の距離をいくつ
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレ
コピペ

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

```
in_f <- "sample19.txt"                    #入力ファイル名

#データファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, se

#本番
dist(t(data), method="euclidean")         #ユークリッド(E
dist(t(data), method="manhattan")         #マンハッタン(M
dist(t(data), method="maximum")           #チェビシェフ(C
dist(t(data), method="canberra")          #キャンベラ(Can
1 - cor(data, method="pearson")           #1 - Pearson相

dist(t(data), method="binary")            #ハミング(Hamm
dist(t(data), method="minkowski")         #ミンコフスキー
1 - cor(data, method="spearman")          #1 - Spearman
```

①

```
> dist(t(data), method="euclidean")       #ユークリッド(Euclid$
         sample1
sample2 2.792848
> dist(t(data), method="manhattan")       #マンハッタン(Manhat$
        sample1
sample2     7.6
> dist(t(data), method="maximum")         #チェビシェフ(Chebys$
        sample1
sample2     1.9
> dist(t(data), method="canberra")        #キャンベラ(Canberra$
          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$
        sample1
sample2       0
> dist(t(data), method="minkowski")       #ミンコフスキー(Mink$
         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
>
```



参考

①「?dist」。②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。③“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

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

②

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

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

## 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, ...)
```

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.

③

参考

クラスター間の距離の定義

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

- 最短距離法(単連結法;single-linkage)
- 最長距離法(完全連結法;complete-linkage)
- 群平均法(平均連結法;average-linkage)
- 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…

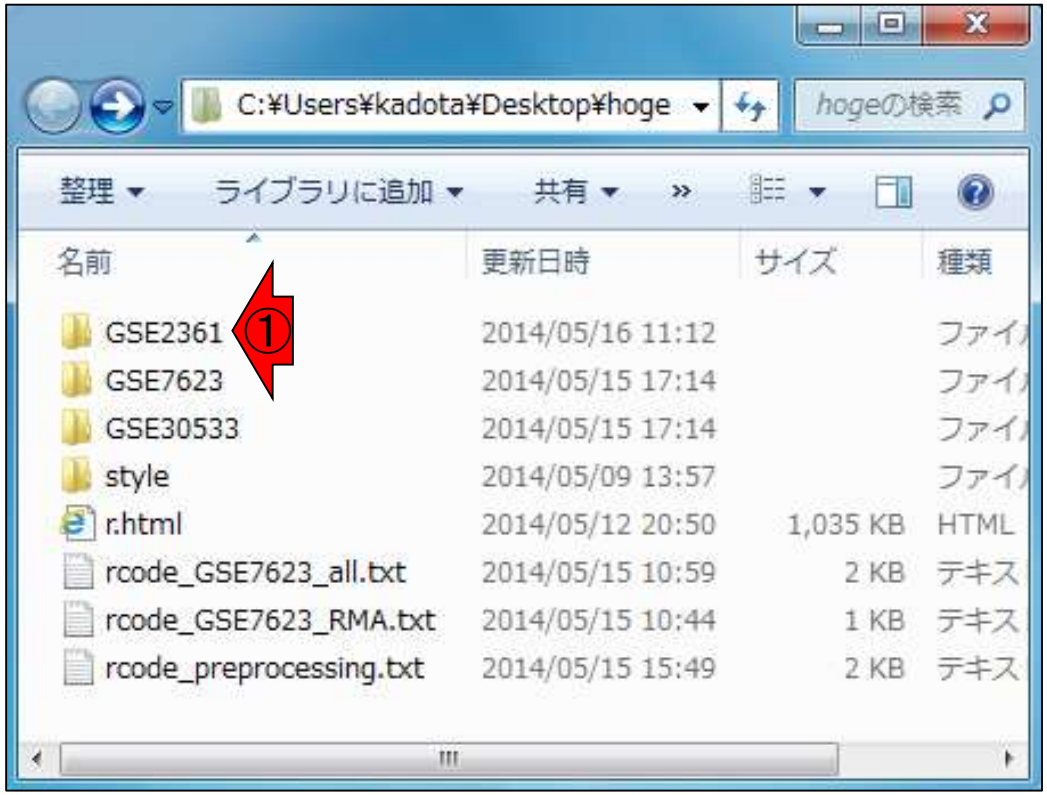
Contents

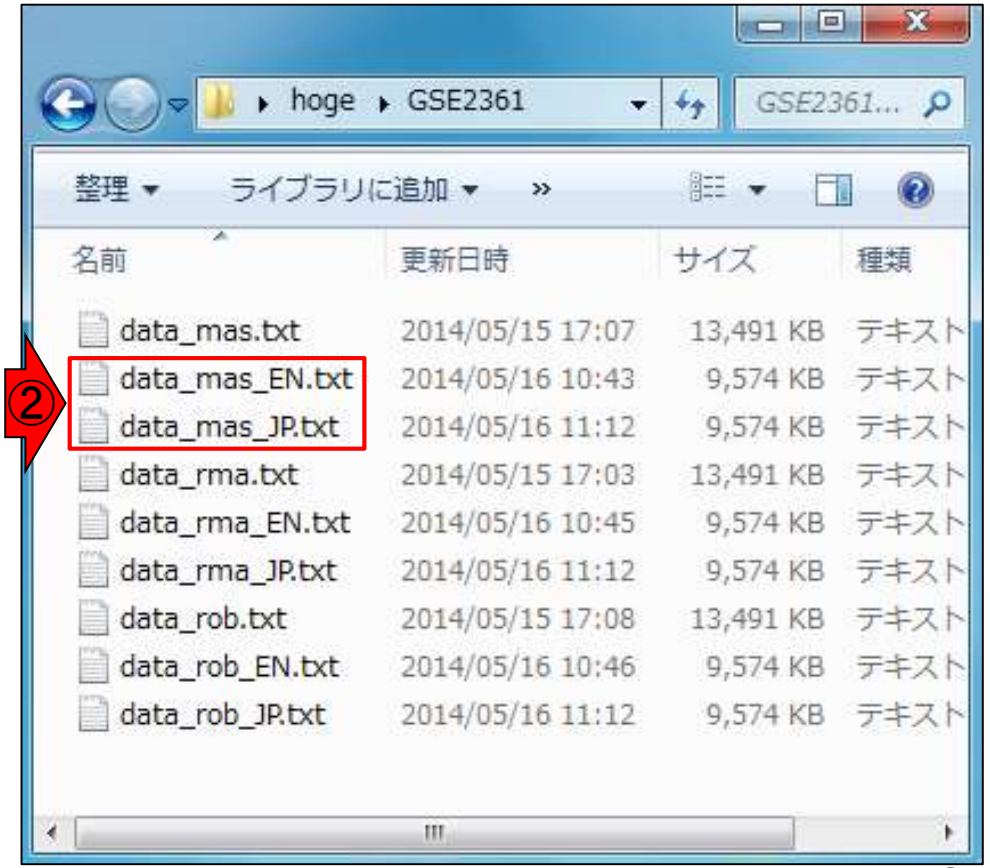
- 実データ概観
  - GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の§3.2.1)
  - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
  - 階層的 vs. 非階層的、様々な選択肢
  - 距離の定義:ベクトル間、クラスター間
  - **実データで実行:GSE2361 (ヒト)、課題1**
  - 実データで実行:GSE7623 (ラット)、GSE30533 (ラット)
  - 同一プラットフォームデータ(GSE7623 + GSE30533)をマージして実行、**課題2**
- 実験デザイン(教科書の§3.2.2)
- 2群間比較:発現変動遺伝子(DEG)検出
  - パターンマッチング法(相関係数の利用)
    - コードの中身をおさらい、apply関数の基本的な利用法など

GSE2361(ヒト)

①hoge – GSE2361フォルダ中の、②MAS5データを用いて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)、…





GSE2361（ヒト）

例題との違いは、②入力ファイル名と③Spearman相関係数の部分。④rcode_clustering.txt

(Rで)マイクロアレイデータ解析

- 解析 | クラスタリング | 階層的 | について (last modified 2009/08/12)
- 解析 | クラスタリング | 階層的 | pvclust (Suzuki 2006) (last modified 2010/08/05)
- 解析 | クラスタリング | 階層的 | hclust (last modified 2014/05/17)
- 解析 | クラスタリング | 階層的 | hclust後の詳細な解析 (last modified 2009/8/7)

①

解析 | クラスタリング | 階層的 | hclust

階層的クラスタリングのやり方を示します。1.
(対数変換やZ-scoreなど)、3.距離(または非
ターをまとめる方法(平均連結法やウォード法
「ファイル」-「ディレクトリの変更」で解析したい

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
に表示するやり方です。

```
in_f <- "sample3.txt"
param <- "average"

#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

#本番
data.dist <- as.dist(1 - cor(data, method="pearson"))#サンプル間の距離を計算した
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
plot(out)                              #樹形図（デンドログラム）の表示
```

②

④ 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)↓
←
```

②

③

リーズナブルだと私は思います

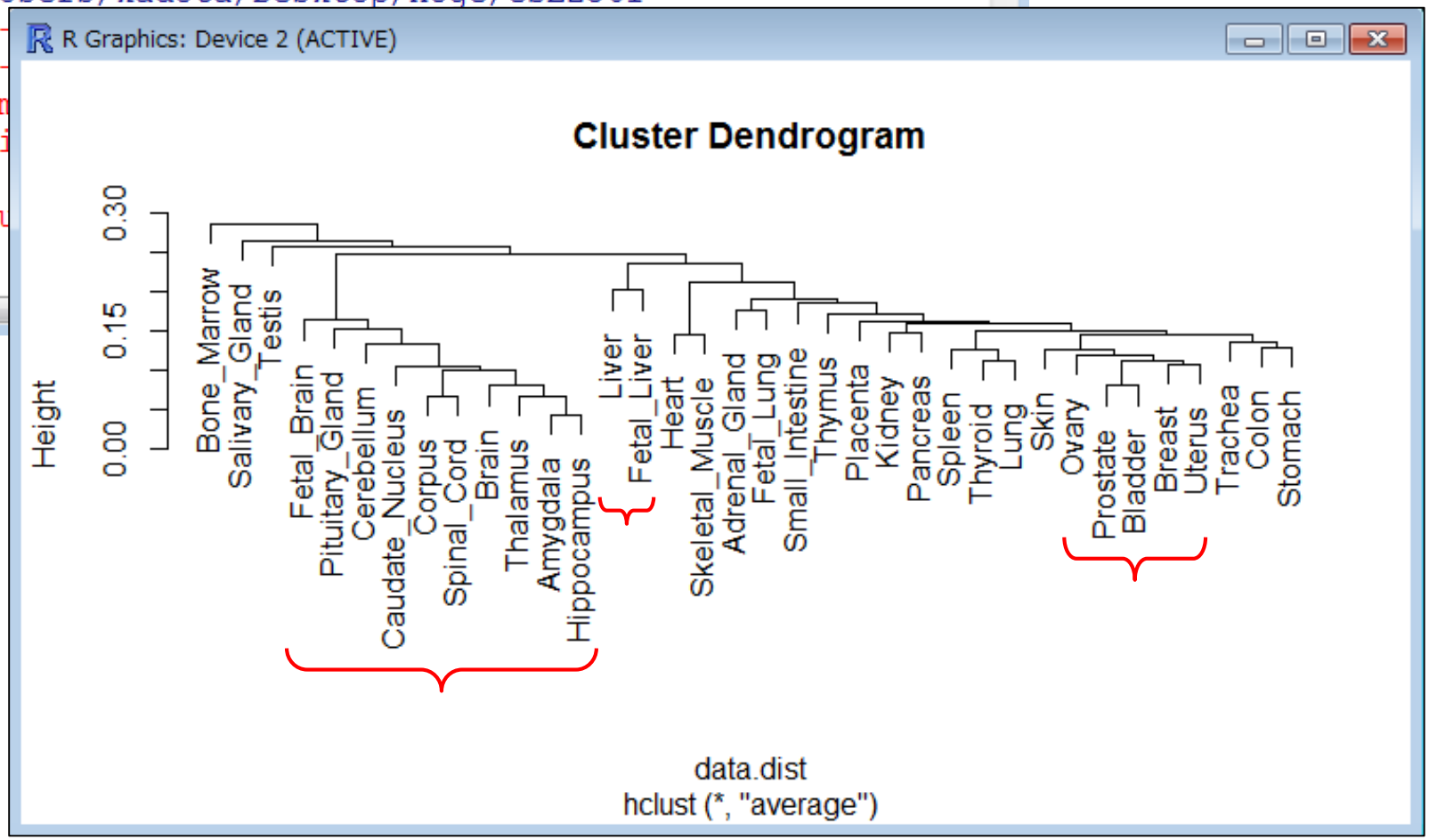
GSE2361 (ヒト)

rcode_clustering.txt

```
#####################################↓
###  MAS5データのクラスタリング  ###↓
#####################################↓
in_f <- "data_mas_EN.txt"↓
data <- read
#colnames(da
data.dist <-
out <- hclus
plot(out)↓
←
```

R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/GSE2361"
> in_f <-
> data <-
> #colnam
> data.di
> out <-
> plot(ou
>
```



GSE2361（ヒト）

日本語環境ではない場合?!
文字化けすることもあるようですのでご注意ください

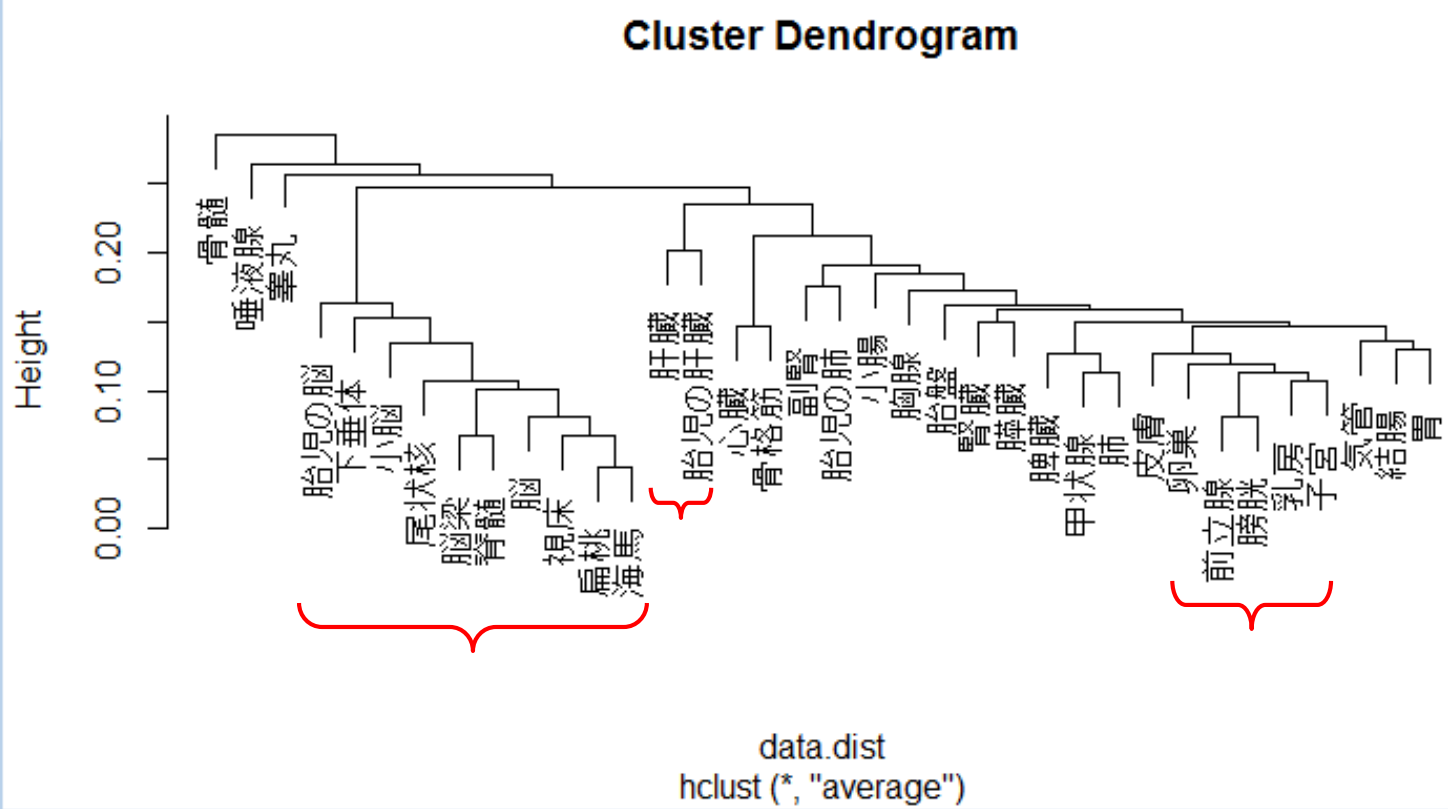
rcode_clustering.txt

```
#####################################
###  MAS5データのクラスタリング  ###
#####################################
in_f <- "data_mas_JP.txt"
data <- read
#colnames(da
data.dist <-
out <- hclus
plot(out)
```

R Console

```
> in_f <- "data_mas_JP.txt"
> data <- read.table(in_f, header=TRUE, row.names=1, sep=$
> #colnam
> data.di
> out <-
> plot(ou
>
```

R Graphics: Device 2 (ACTIVE)



Tips（ファイル保存）

①例題3。PNG形式ファイルとして縦横の大きさを指定して保存することもできる。テンプレートとの違いは赤矢印部分。②rcode_clustering_png.txt

解析 | クラスタリング | 階層的 | hclust

①

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

#入力ファイル
data <- read

#本番
data.dist <-
out <- hclus

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

rcode_clustering_png.txt ②

```
##############################################################################
###  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)             #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
```

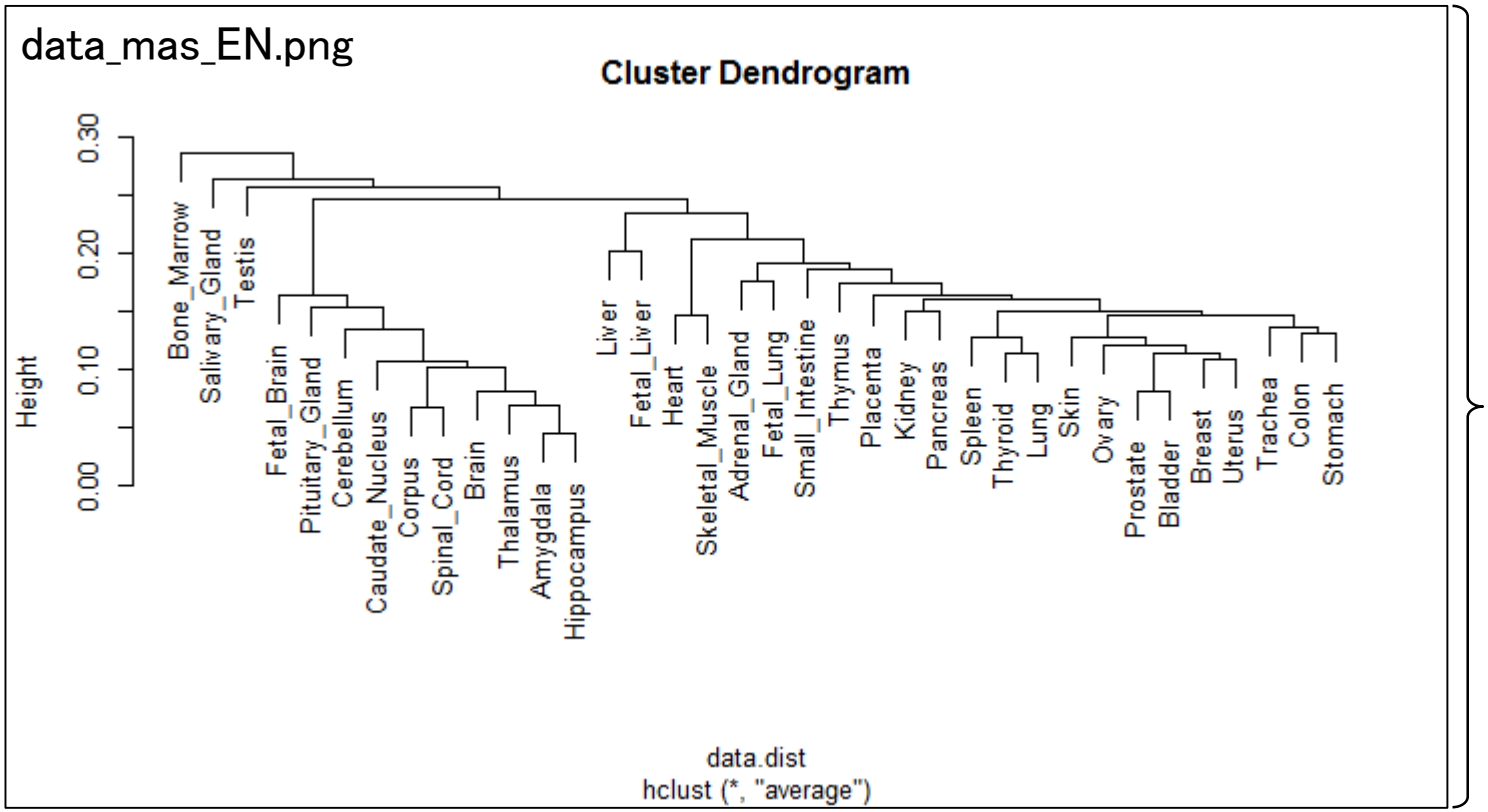
常に同じ大きさで出力されるので便利

Tips（ファイル保存）

rcode_clustering_png.txt ②

```
#################################################################################↓
###  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, hea
↓
#本番↓
data.dist <- as.dist(1 - cor
out <- hclust(data.dist, met
↓
#ファイルに保存↓
png(out_f, pointsize=13, wid
plot(out)
dev.off()
```

data_mas_EN.png



課題1

GSE2361のサンプル間クラスタリングをRMA, およびRMX前処理法を適用したデータについても行い、結果を考察せよ。距離の定義はデフォルトのままでよい。①の部分をdata_rma_...やdata_rob_...に変更すればよいが…

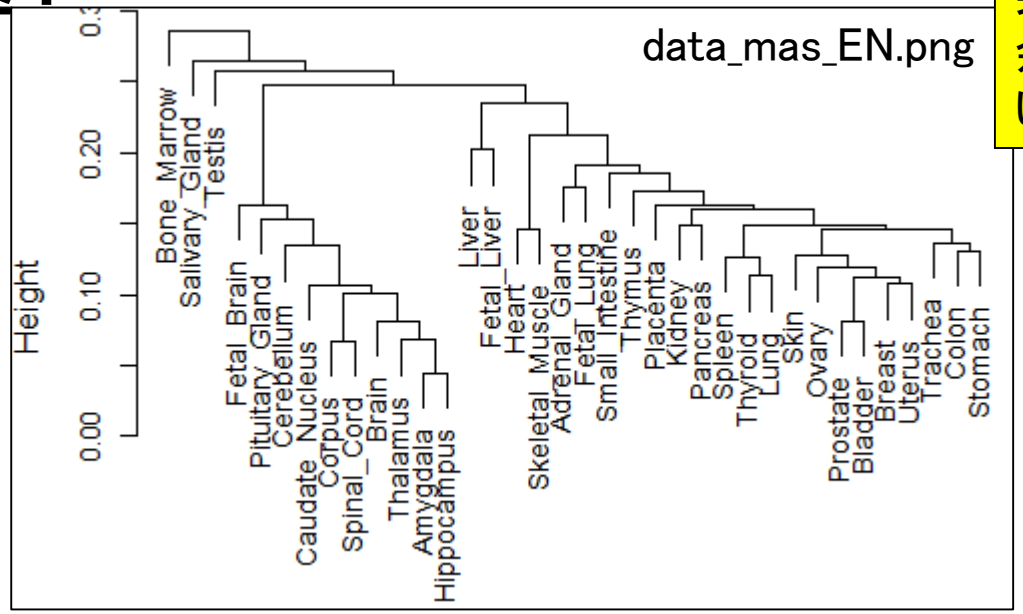
rcode_clustering_png.txt

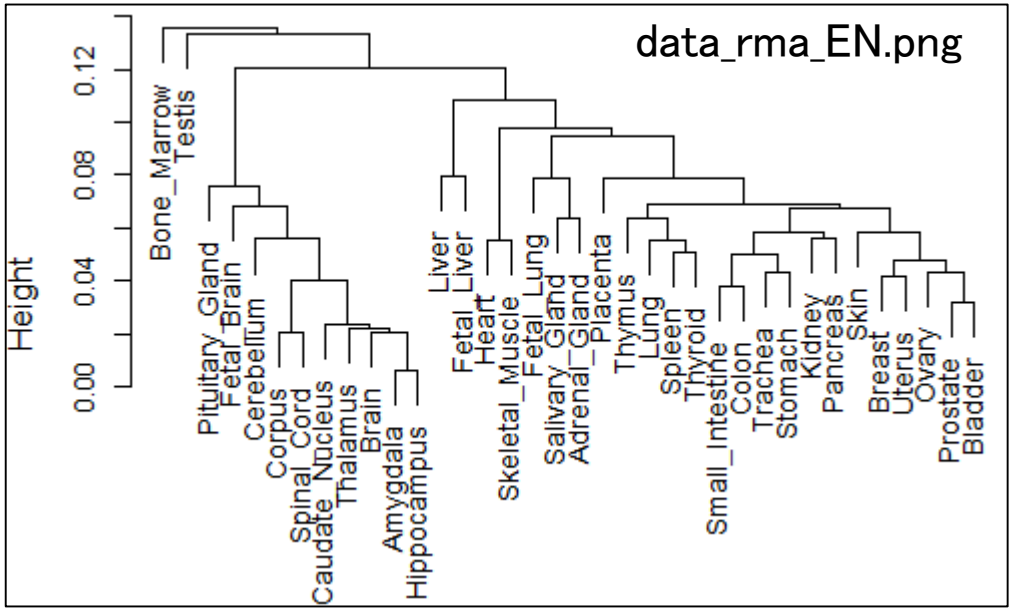
```
#################################################################################
###  MAS5データのクラスタリング，距離: 1-Spearman相関係数 ，方法: 平均連結法  ###
#################################################################################
in_f <- "data_mas_EN.txt"                  #入力ファイル名を指定してin_fに格納
out_f <- "data_mas_EN.png"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

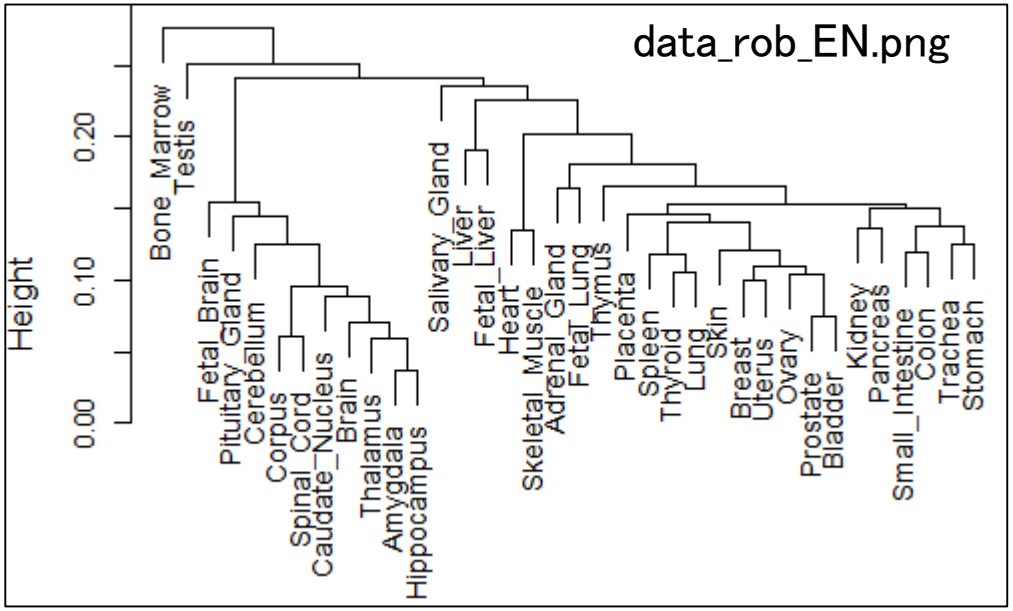
①

課題1

①rcode_clustering_png_kadai2.txt をコピペで実行すれば、このような結果が得られる。実際の論文用の図は、余白指定や文字の大きさなどをいろいろ変えて、見栄えをよくします







発展課題（optional）

①例題5はユークリッド距離を用いる場合のテンプレートです。②param1（やparam2）などをいろいろいじって結果を眺め、可能な範囲で考察してください

解析 | クラスタリング | 階層的 | hclust

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

1. サンプルデー

サンプル間クラス
に表示するやり方

```
in_f <- "samp
param <- "ave

#入力ファイル(
data <- read.

#本番
data.dist <-
out <- hclust
plot(out)
```

5. 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:ユークリッド距離(euclidean)、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。

②

```
in_f <- "sample3.txt"                    #入力ファイル名を指定してin_fに格納
out_f <- "hoge5.png"                     #出力ファイル名を指定してout_fに格納
param1 <- "euclidean"                    #距離(dist)を指定
param2 <- "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

#本番
data.dist <- dist(t(data), method=param1)#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=param2)#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
plot(out)                                #樹形図（デンドログラム）の表示
dev.off()                                #おまじない
```

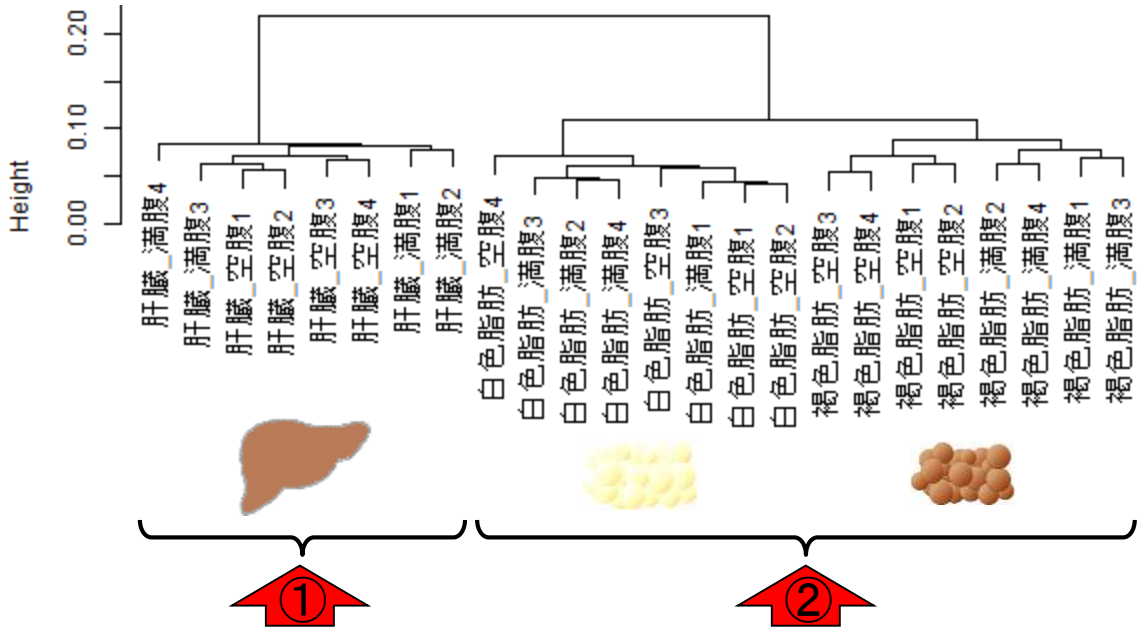
Contents

- 実データ概観
  - GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の§3.2.1)
  - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
  - 階層的 vs. 非階層的、様々な選択肢
  - 距離の定義:ベクトル間、クラスター間
  - 実データで実行:GSE2361 (ヒト)、**課題1**
  - 実データで実行:GSE7623 (ラット)、GSE30533 (ラット)
  - 同一プラットフォームデータ(GSE7623 + GSE30533)をマージして実行、**課題2**
- 実験デザイン(教科書の§3.2.2)
- 2群間比較:発現変動遺伝子(DEG)検出
  - パターンマッチング法(相関係数の利用)
    - コードの中身をおさらい、apply関数の基本的な利用法など

GSE7623 (ラット)

MAS5データを(1 − Spearman)相関係数でクラスタリングした結果。①肝臓と②脂肪間で大きく2つのクラスターに分かれていることがわかる

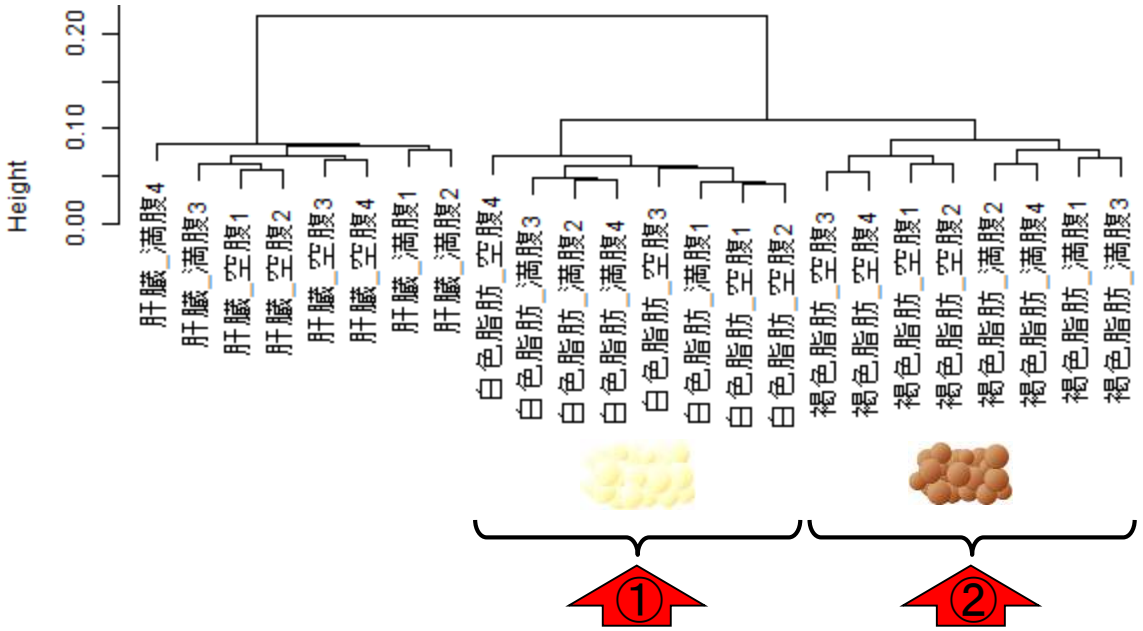
- Nakai et al., *BBB*, 72: 139–148, 2008
  - GSE7623、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
  - ラット24サンプル：Brown adipose tissue (褐色脂肪組織; BAT) 8サンプル、White adipose tissue (白色脂肪組織; WAT) 8サンプル、Liver (肝臓; LIV) 8サンプル
    - BAT 8サンプル：通常(BAT_fed) 4サンプル　対　24時間絶食(BAT_fas) 4サンプル
    - WAT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



脂肪の中でも、①白色脂肪と②褐色脂肪に分かれている

GSE7623 (ラット)

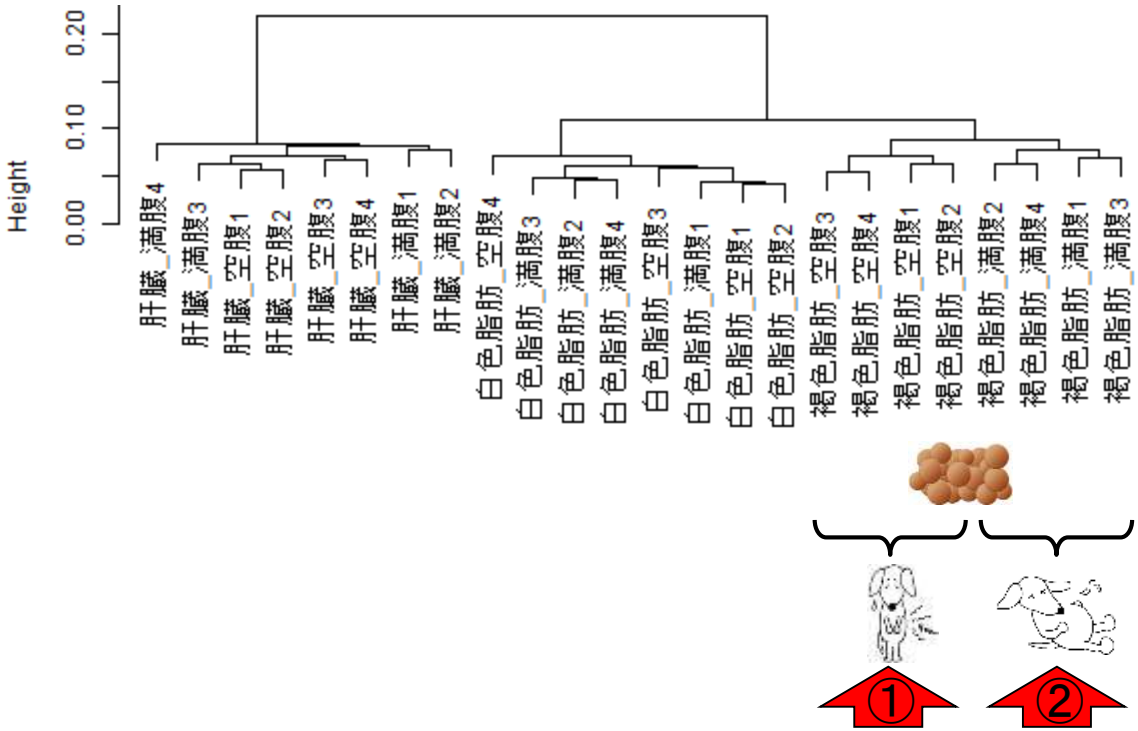
- Nakai et al., *BBB*, 72: 139–148, 2008
  - GSE7623、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
  - ラット24サンプル：Brown adipose tissue (褐色脂肪組織; BAT) 8サンプル、White adipose tissue (白色脂肪組織; WAT) 8サンプル、Liver (肝臓; LIV) 8サンプル
    - BAT 8サンプル：通常(BAT_fed) 4サンプル　対　24時間絶食(BAT_fas) 4サンプル
    - WAT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



褐色脂肪の中でも、①空腹（24時間絶食）と②満腹（通常）できれいに分かれている

GSE7623（ラット）

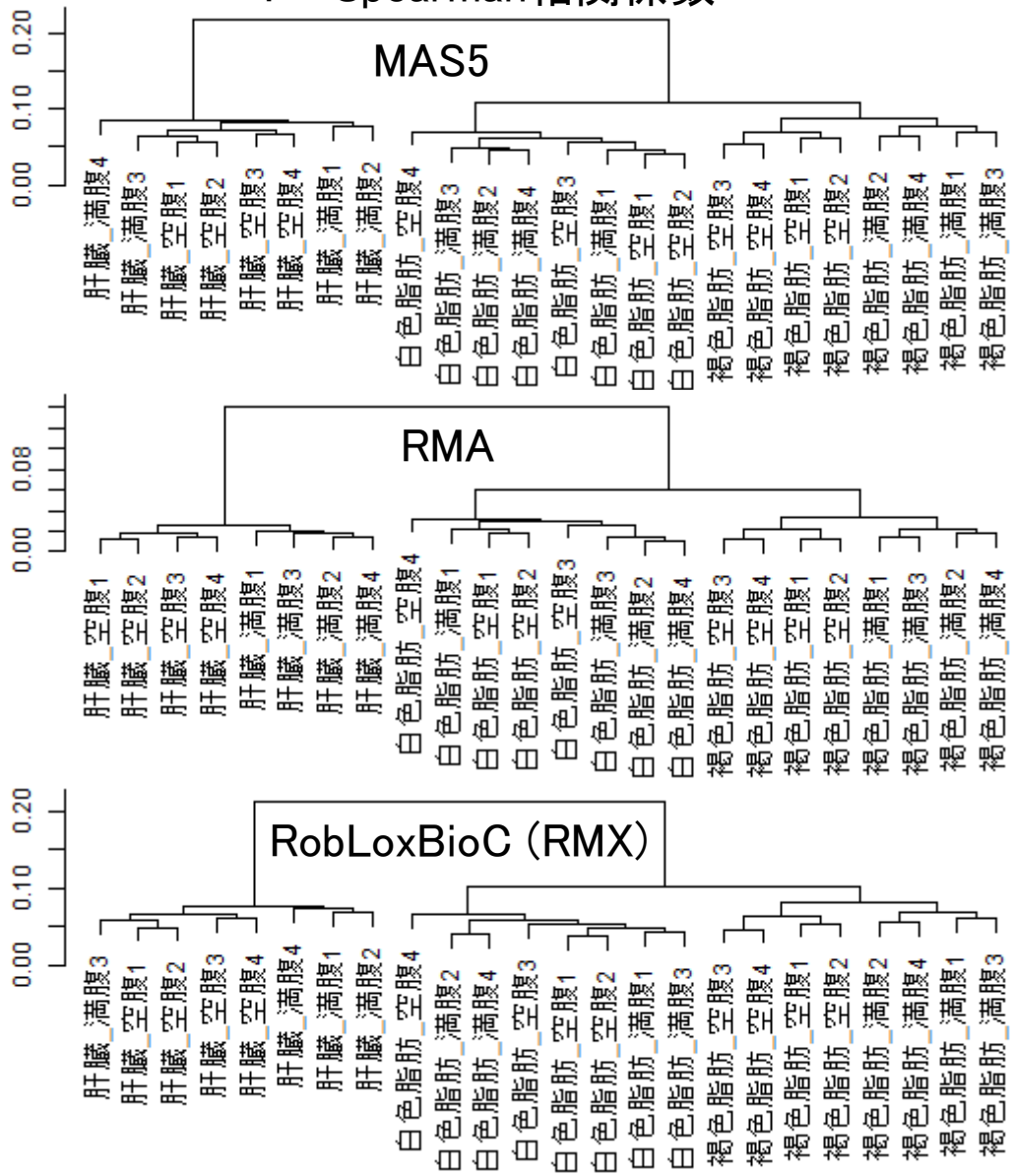
- Nakai et al., *BBB*, 72: 139–148, 2008
  - GSE7623、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
  - ラット24サンプル：Brown adipose tissue（褐色脂肪組織; BAT）8サンプル、White adipose tissue（白色脂肪組織; WAT）8サンプル、Liver（肝臓; LIV）8サンプル
    - BAT 8サンプル：通常（BAT_fed）4サンプル　対　24時間絶食（BAT_fas）4サンプル
    - WAT 8サンプル：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



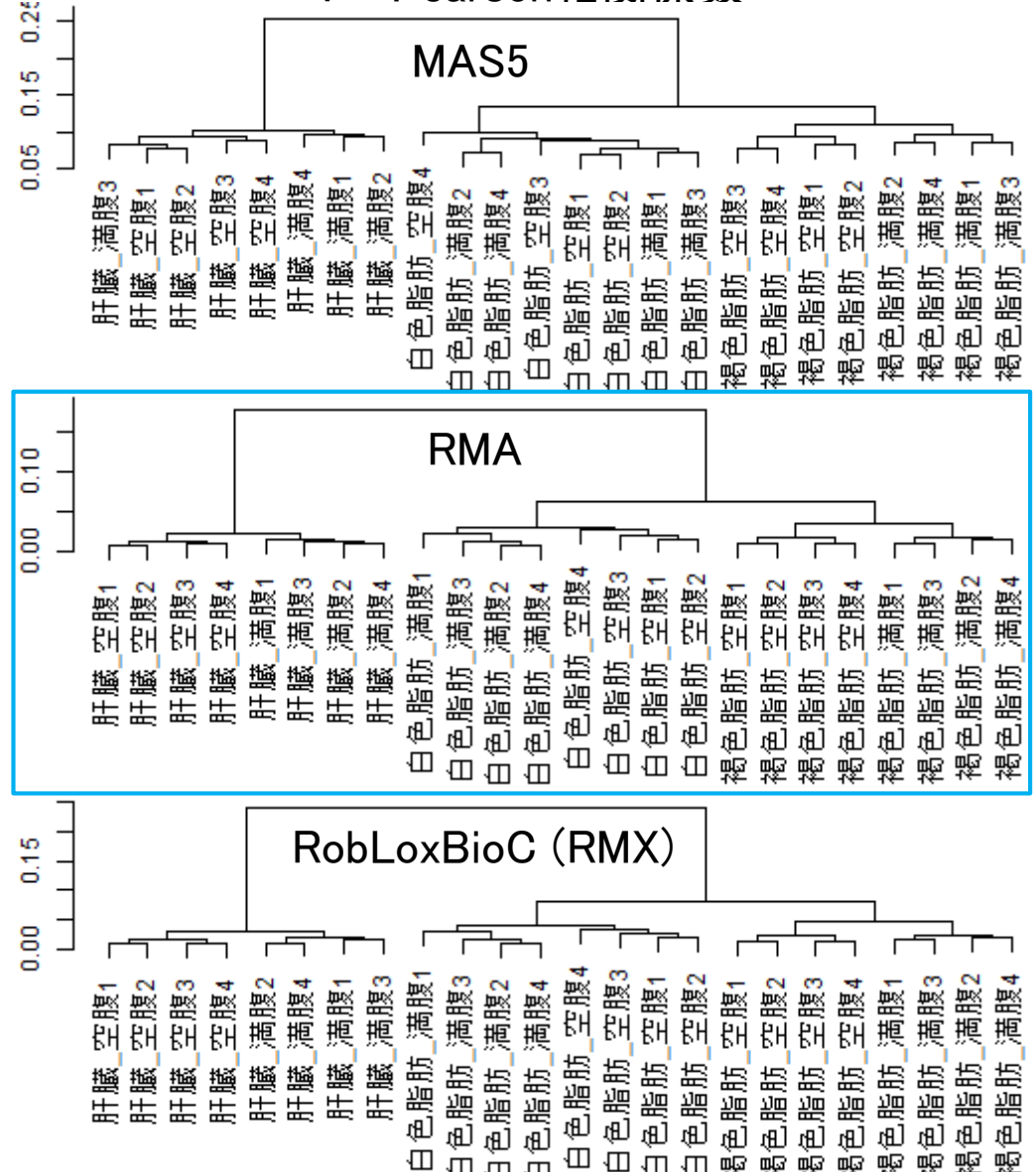
どれもほぼ同じ結果

GSE7623（ラット）

1 − Spearman相関係数

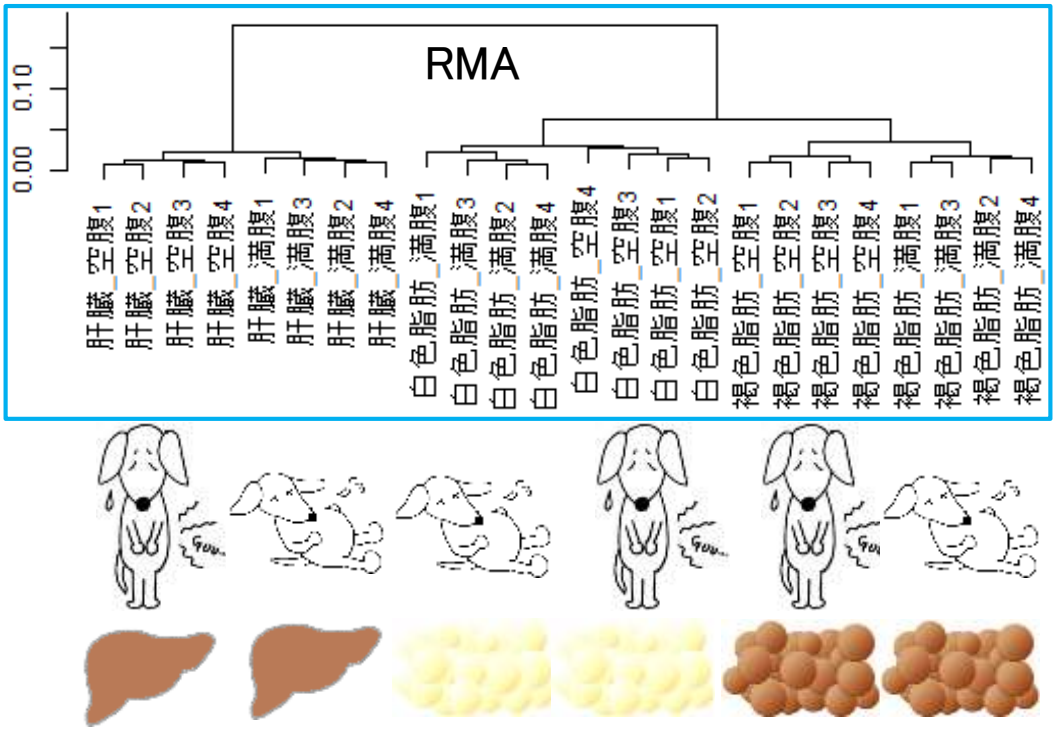


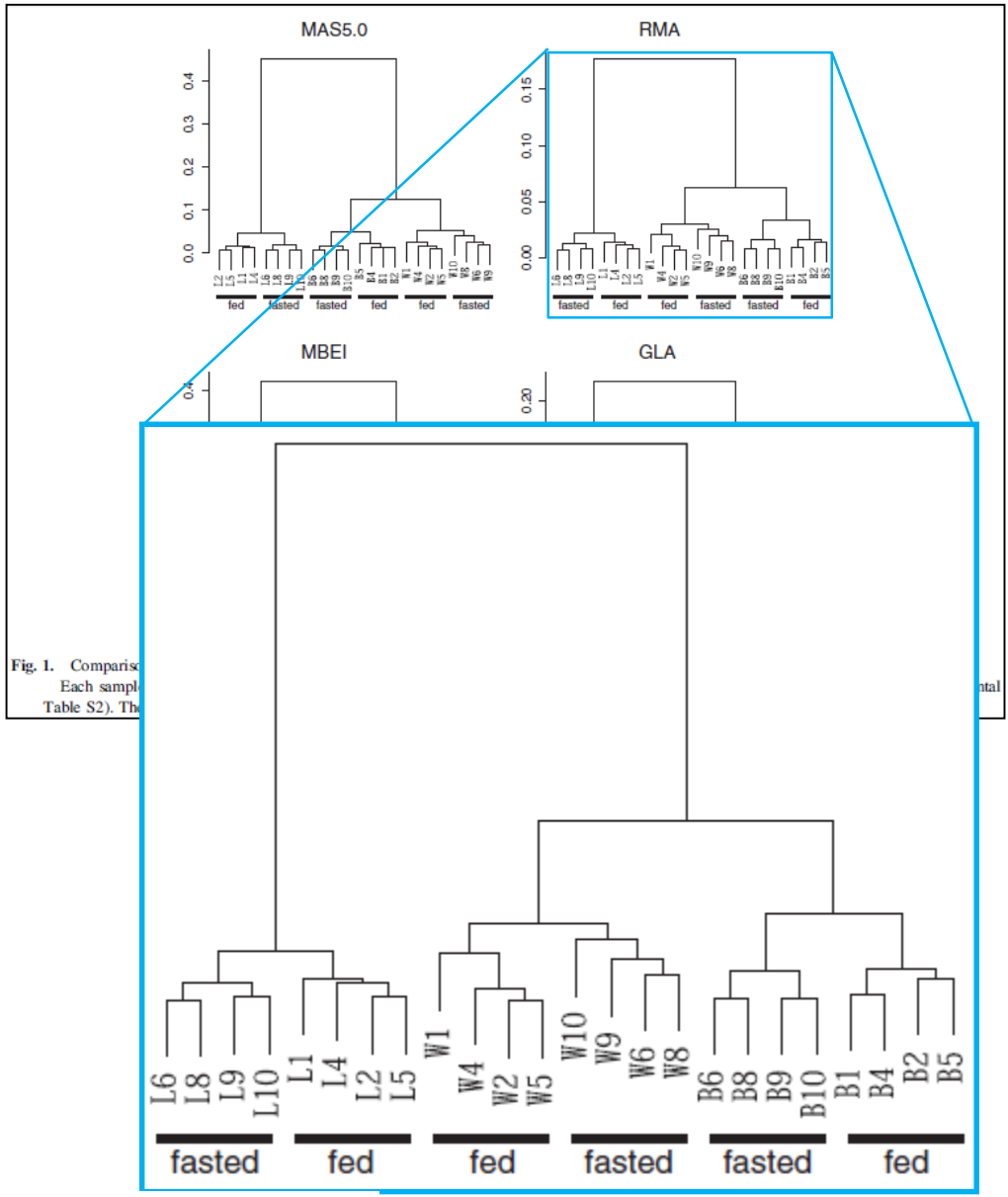
1 − Pearson相関係数



GSE7623（ラット）

①原著論文のFig. 1（の一部）を再現できてます



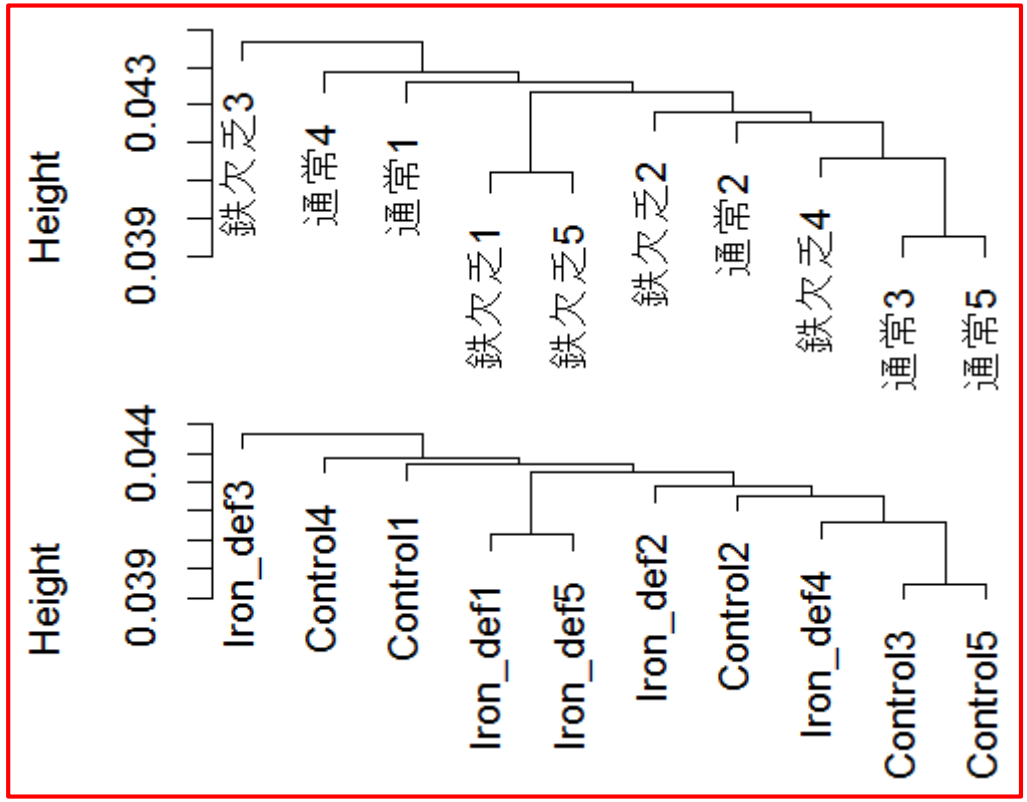


①Nakai et al., *BBB*, **72**: 139–148, 2008

GSE30533（ラット）

①教科書p99の図3-1と基本同じ。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…

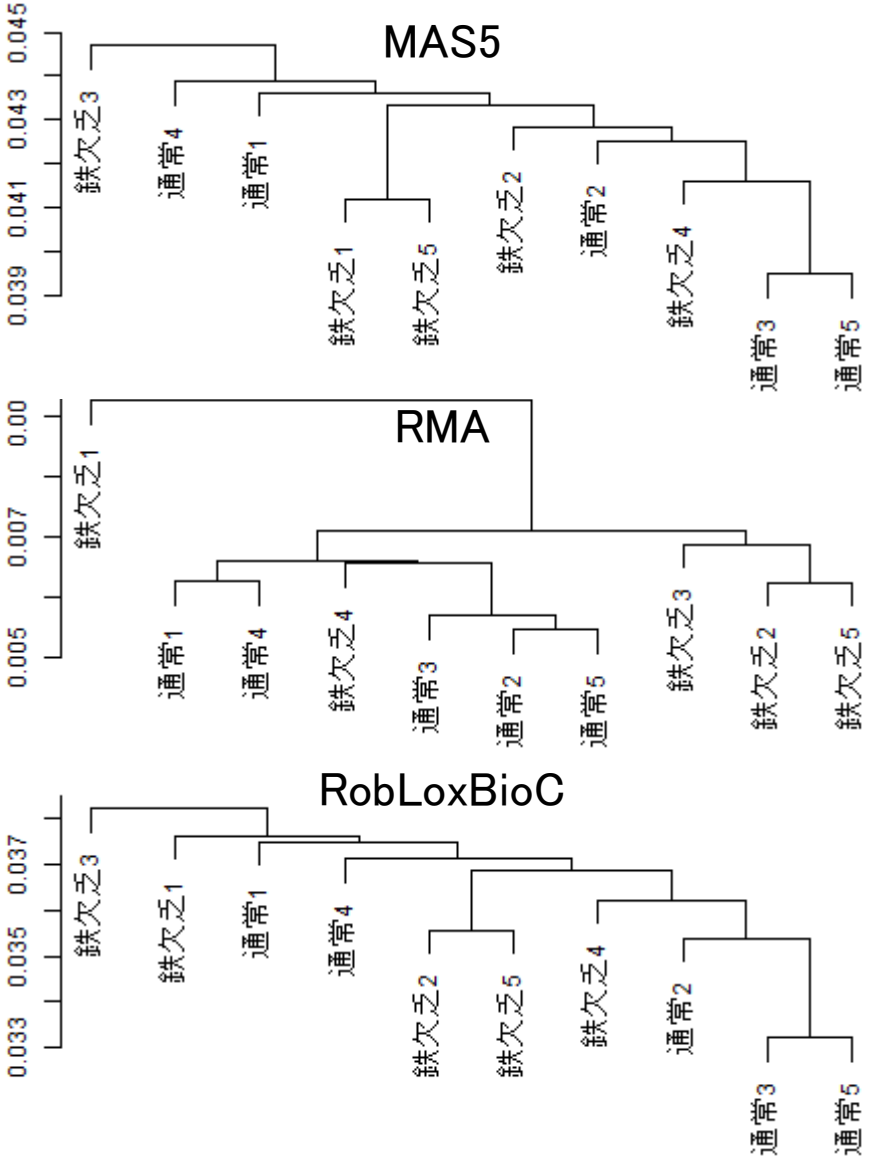
- Kamei et al., PLoS One, 8: e65732, 2013
  - GSE30533、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
  - ラット10サンプル：全てLiver（肝臓）サンプル
  - iron-deficient diet（Iron_def）5サンプル　対　control diet（Control）5サンプル



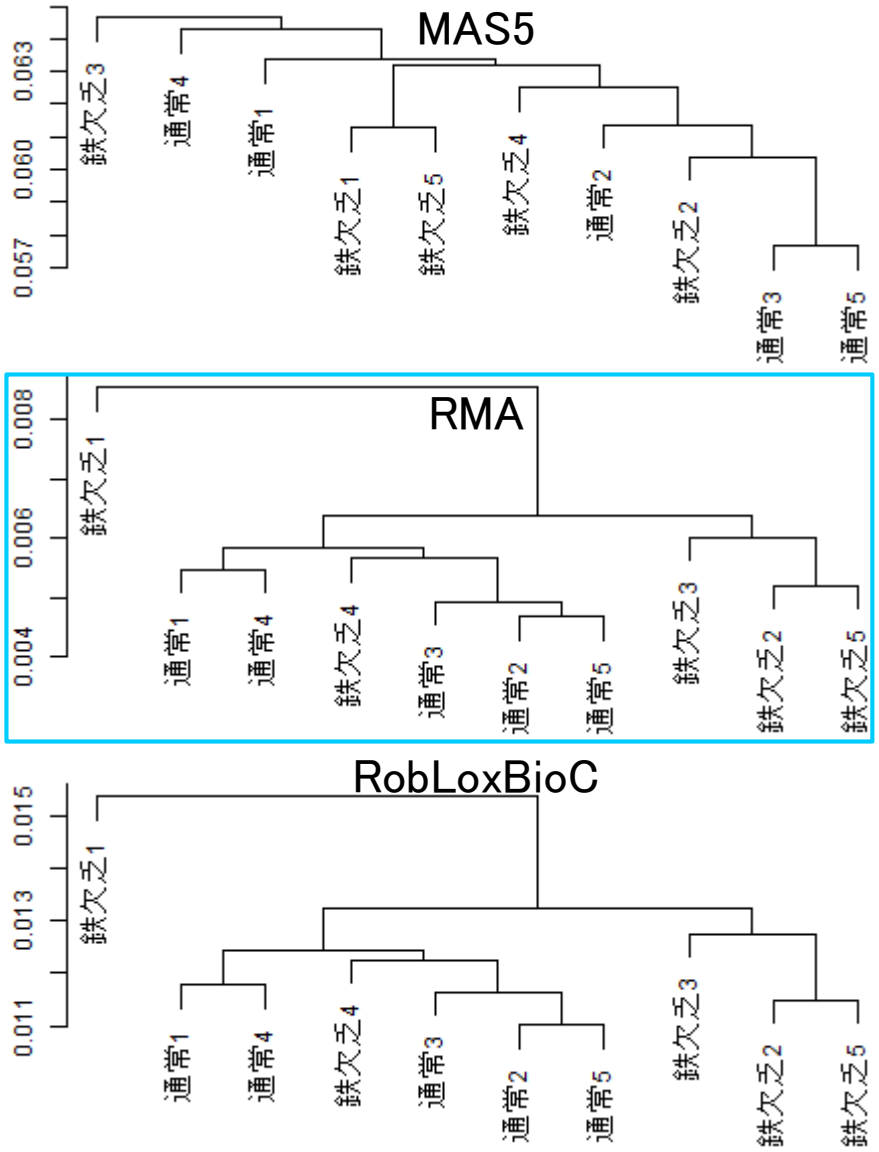
なんともいえない結果

GSE30533（ラット）

1 − Spearman相関係数

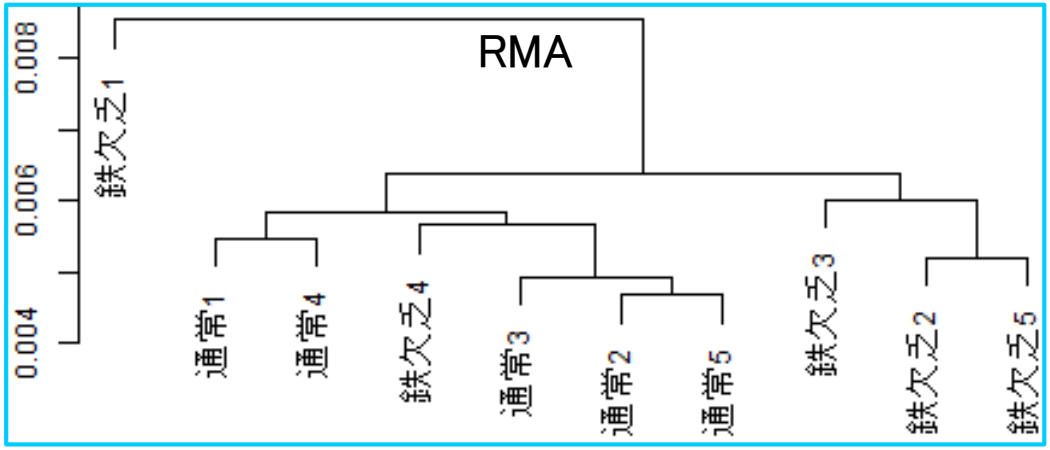


1 − Pearson相関係数

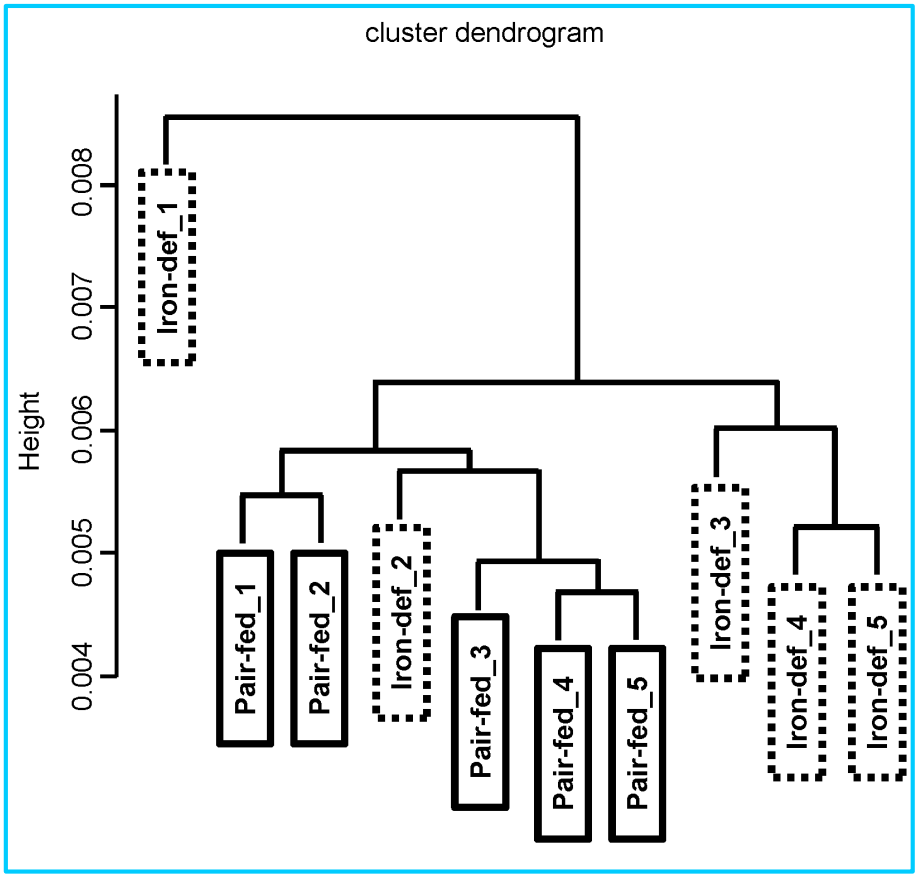


GSE30533（ラット）

（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



①原著論文のFigure S1



①Kamei et al., *PLoS One*, **8**: e65732, 2013

Contents

- 実データ概観
  - GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の§3.2.1)
  - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
  - 階層的 vs. 非階層的、様々な選択肢
  - 距離の定義:ベクトル間、クラスター間
  - 実データで実行:GSE2361 (ヒト)、**課題1**
  - 実データで実行:GSE7623 (ラット)、GSE30533 (ラット)
  - 同一プラットフォームデータ(GSE7623 + GSE30533)をマージして実行、**課題2**
- 実験デザイン(教科書の§3.2.2)
- 2群間比較:発現変動遺伝子(DEG)検出
  - パターンマッチング法(相関係数の利用)
    - コードの中身をおさらい、apply関数の基本的な利用法など

同一アレイデータはマージ可能

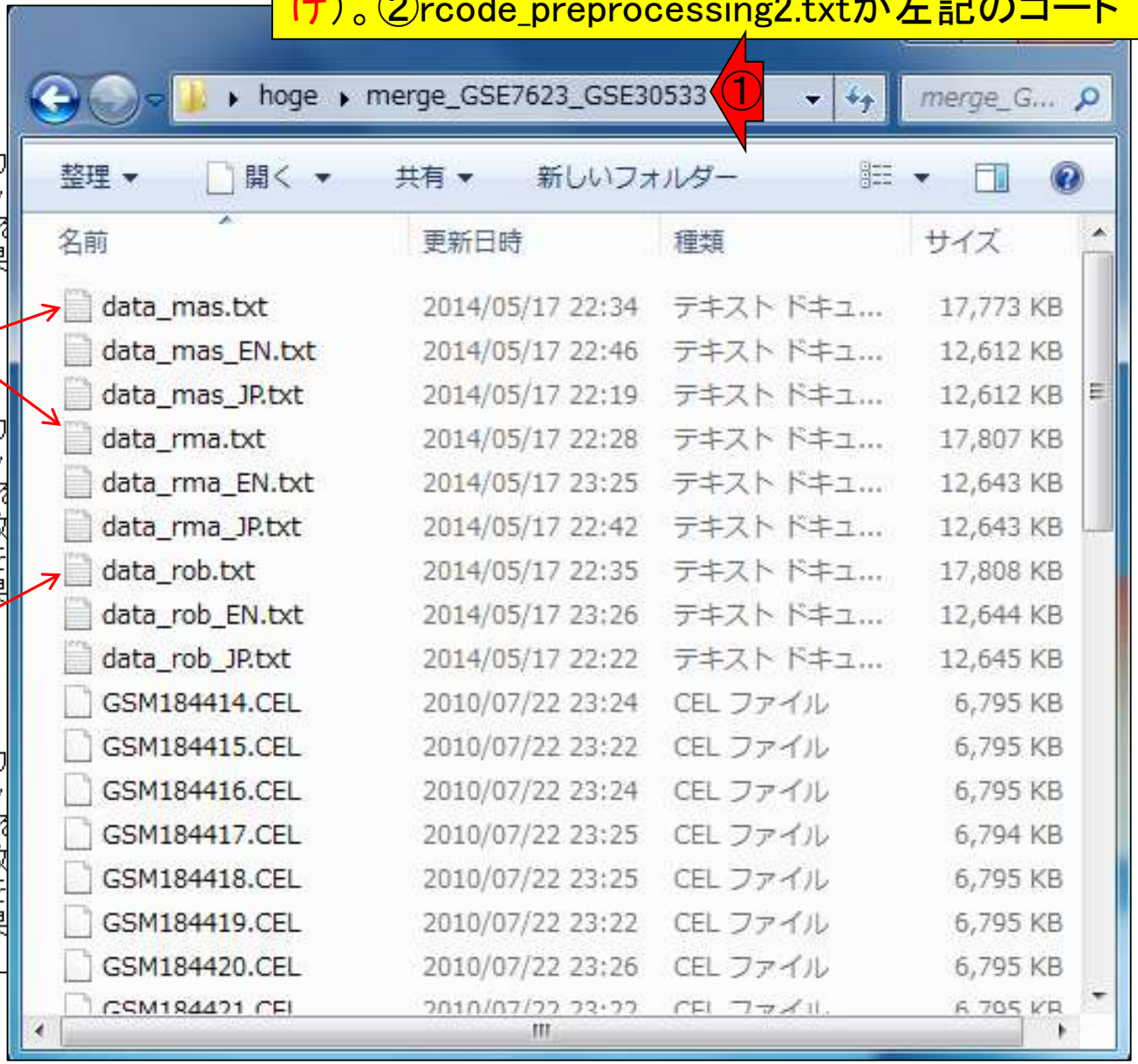
①この2つの論文は同一プラットフォーム（同一アレイ）を利用。3’発現アレイを用いることで、他の多くのデータセットとの比較が可能

- 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

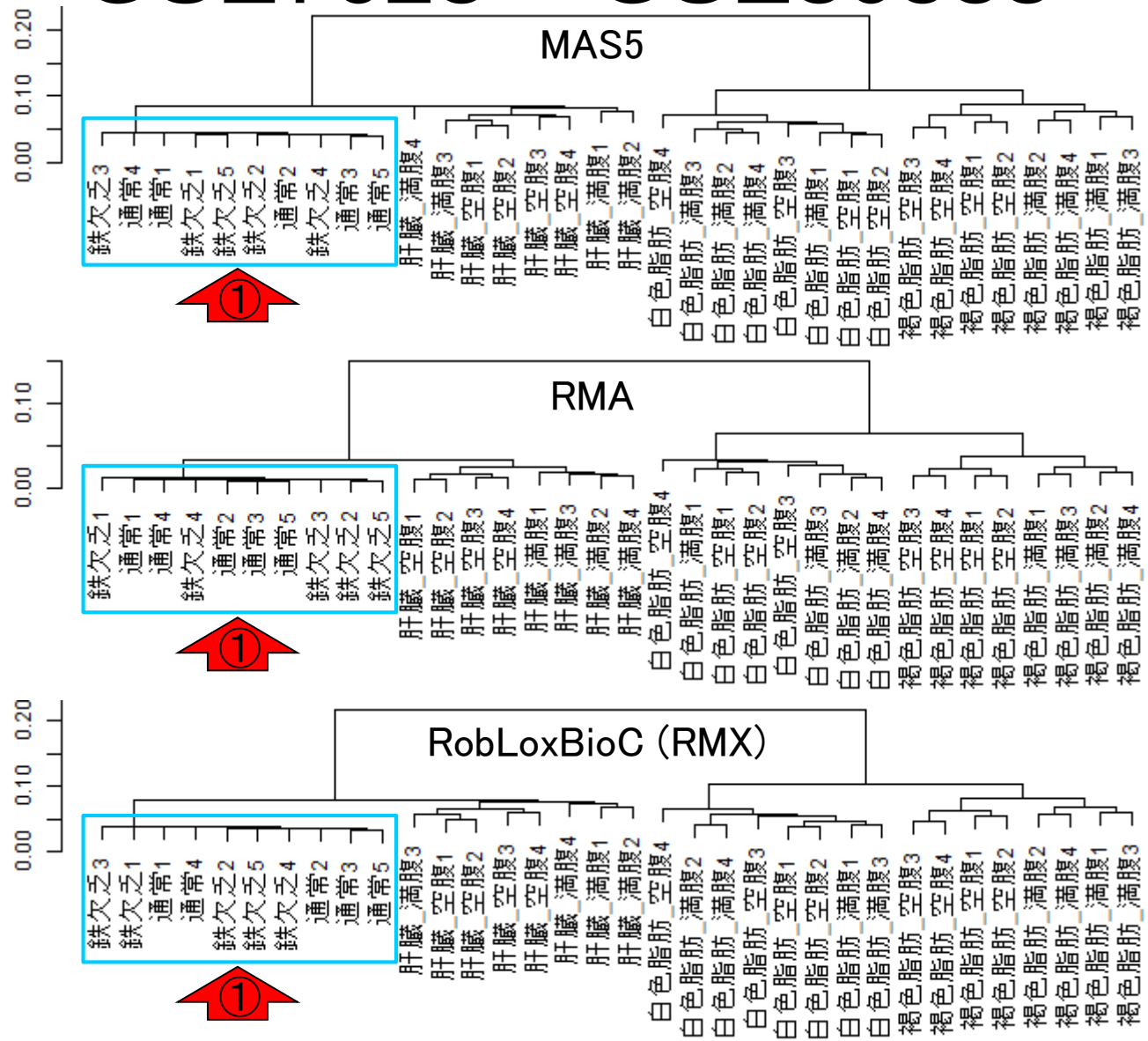
①

①merge_GSE7623_GSE30533上にGSE7623とGSE30533の計34 CELファイルを入力として前処理法を適用した結果のファイルがあります（見るだけ）。②rcode_preprocessing2.txtが左記のコード

```
##########################################################
###  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
```



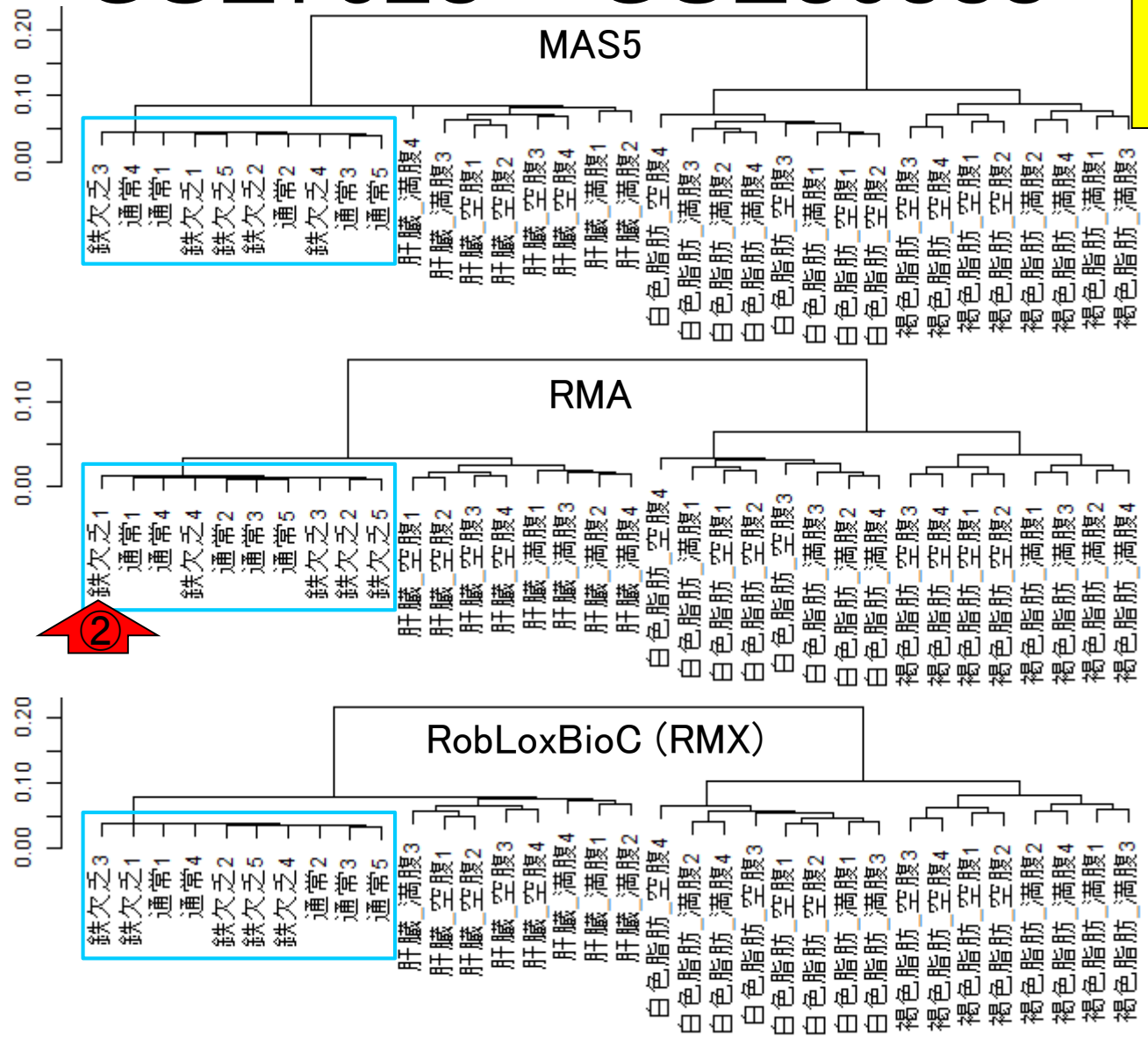
GSE7623 + GSE30533



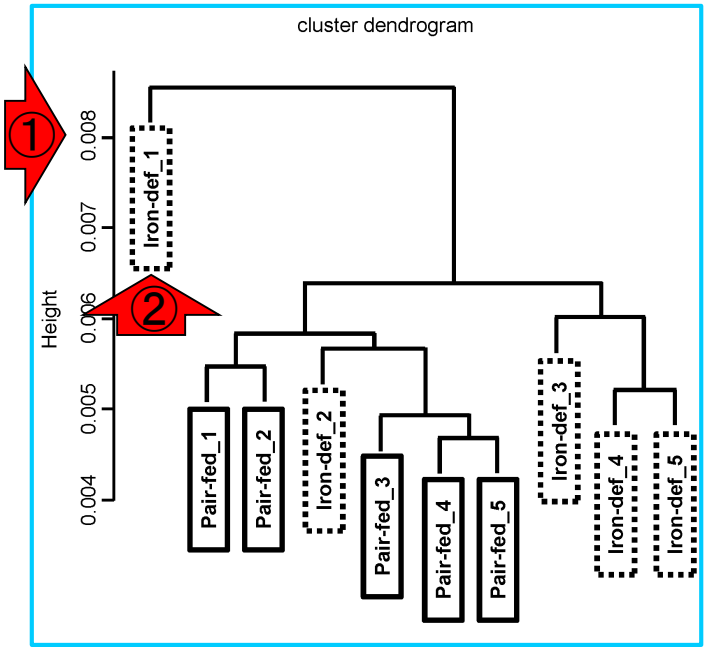
「1 − Spearman相関係数」の結果。どの前処理法でも似たような結果となっているのが分かる。①ラット10サンプル（通常　対　鉄欠乏）のクラスタリング結果の印象は、外群（ラット24サンプル）の有無でずいぶん異なる（教科書p106-107）

GSE7623 + GSE30533

マージすることによって初めて、①距離が非常に近い（サンプル間の類似度が極めて高い）がために、「②Iron-def_1が外れサンプルっぽく見える」といった議論をしてたことに気づく

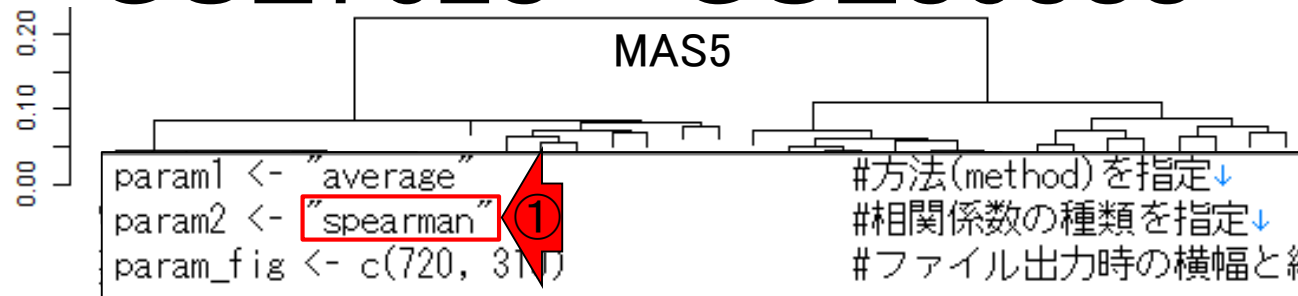


原著論文のFigure S1



「1 – Pearson相関係数」にしたい場合は①spearmanをpearsonに変更

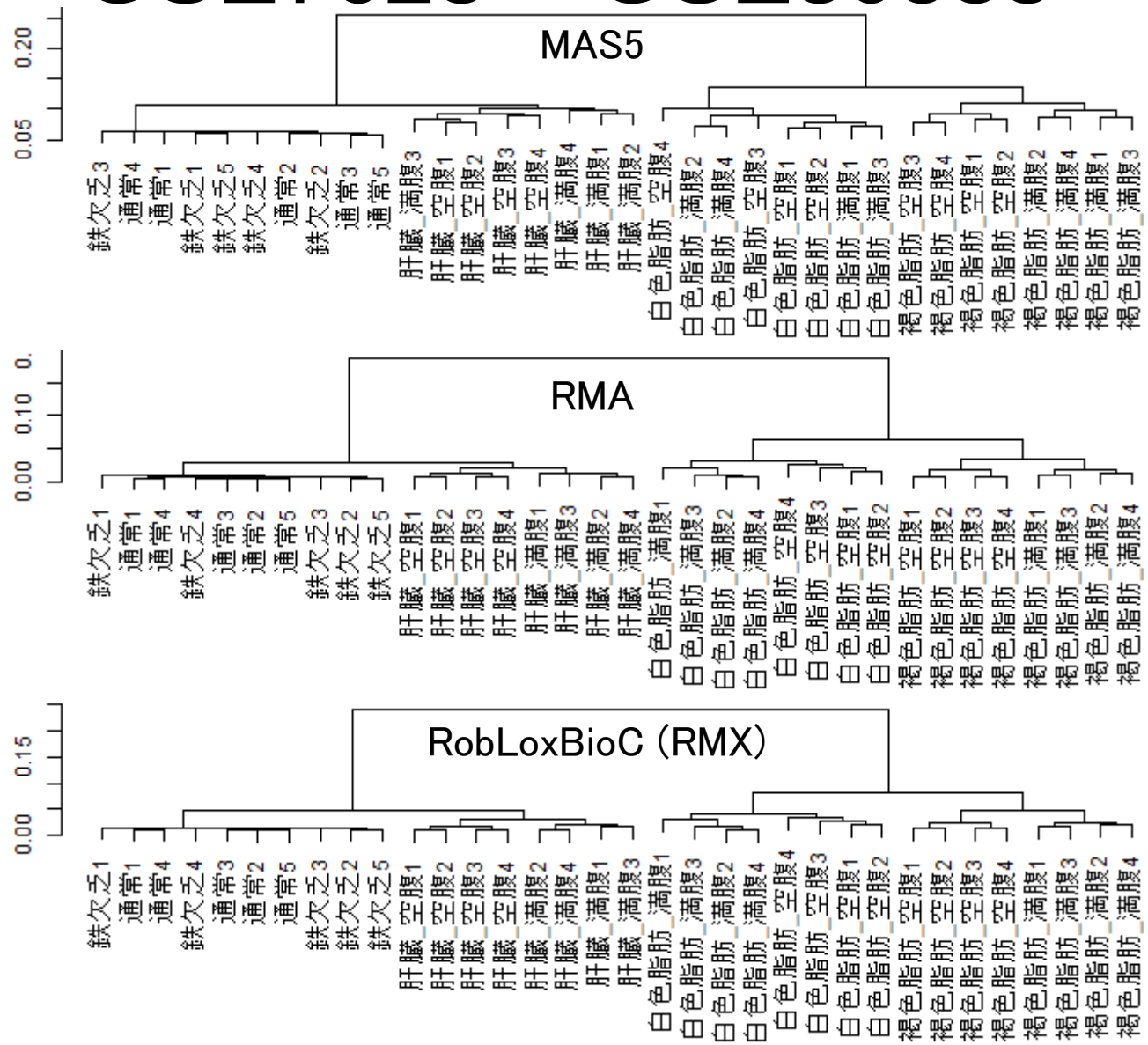
GSE7623 + GSE30533 参考



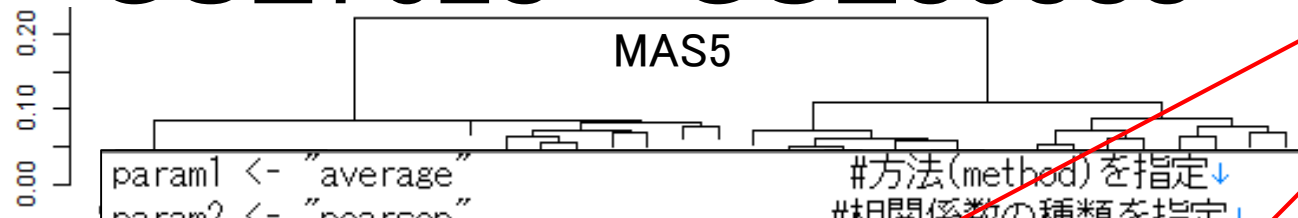

```
param1 <- "average"                    #方法(method)を指定
param2 <- "spearman"  ①                #相関係数の種類を指定
param_fig <- c(720, 31[illegible])      #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

##########################################################################################
###  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
```

GSE7623 + GSE30533

「1 − Pearson相関係数」の結果。どの前処理法でも似たような結果となっているのが分かる



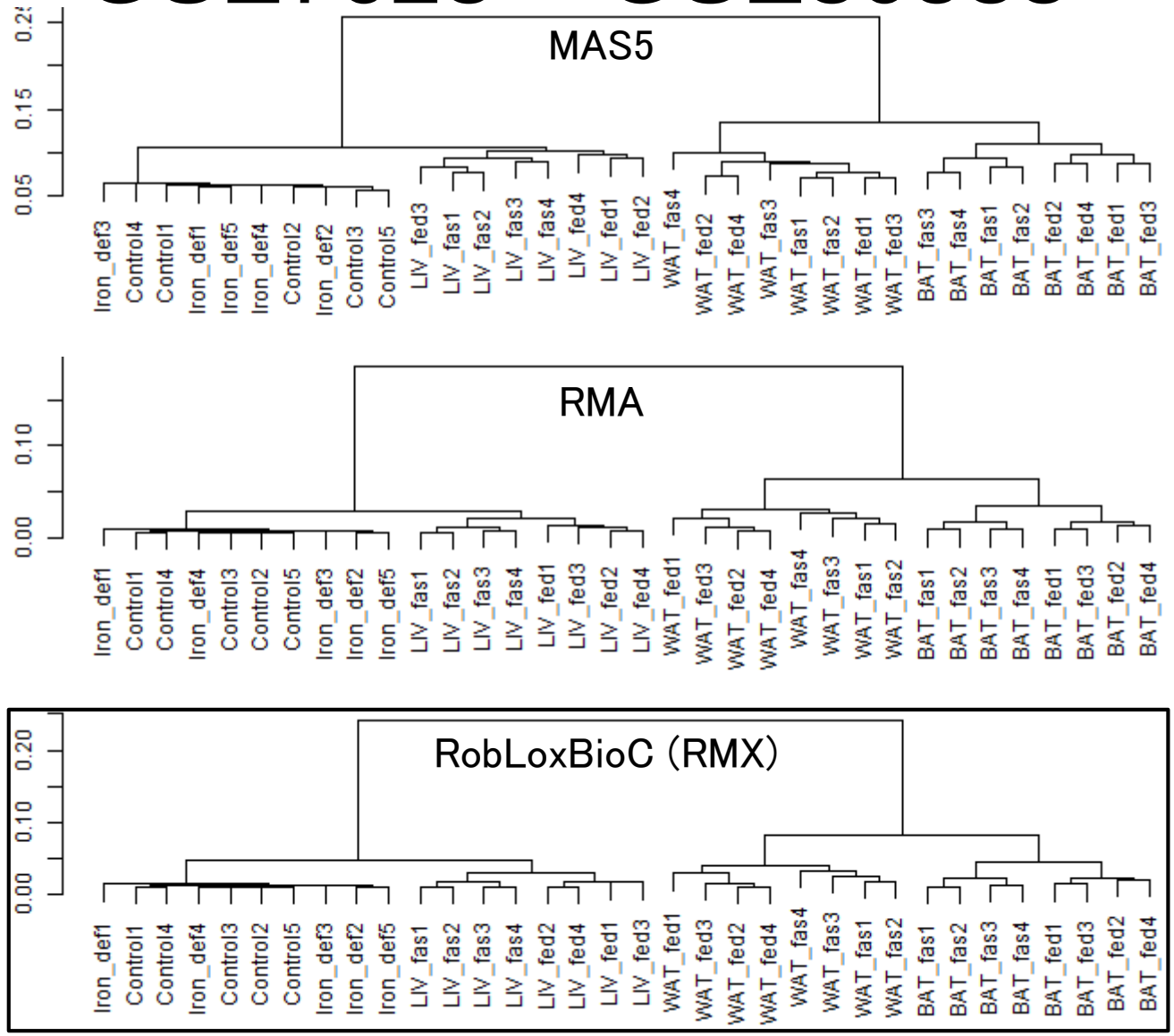
GSE7623 + GSE30533




```
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
```

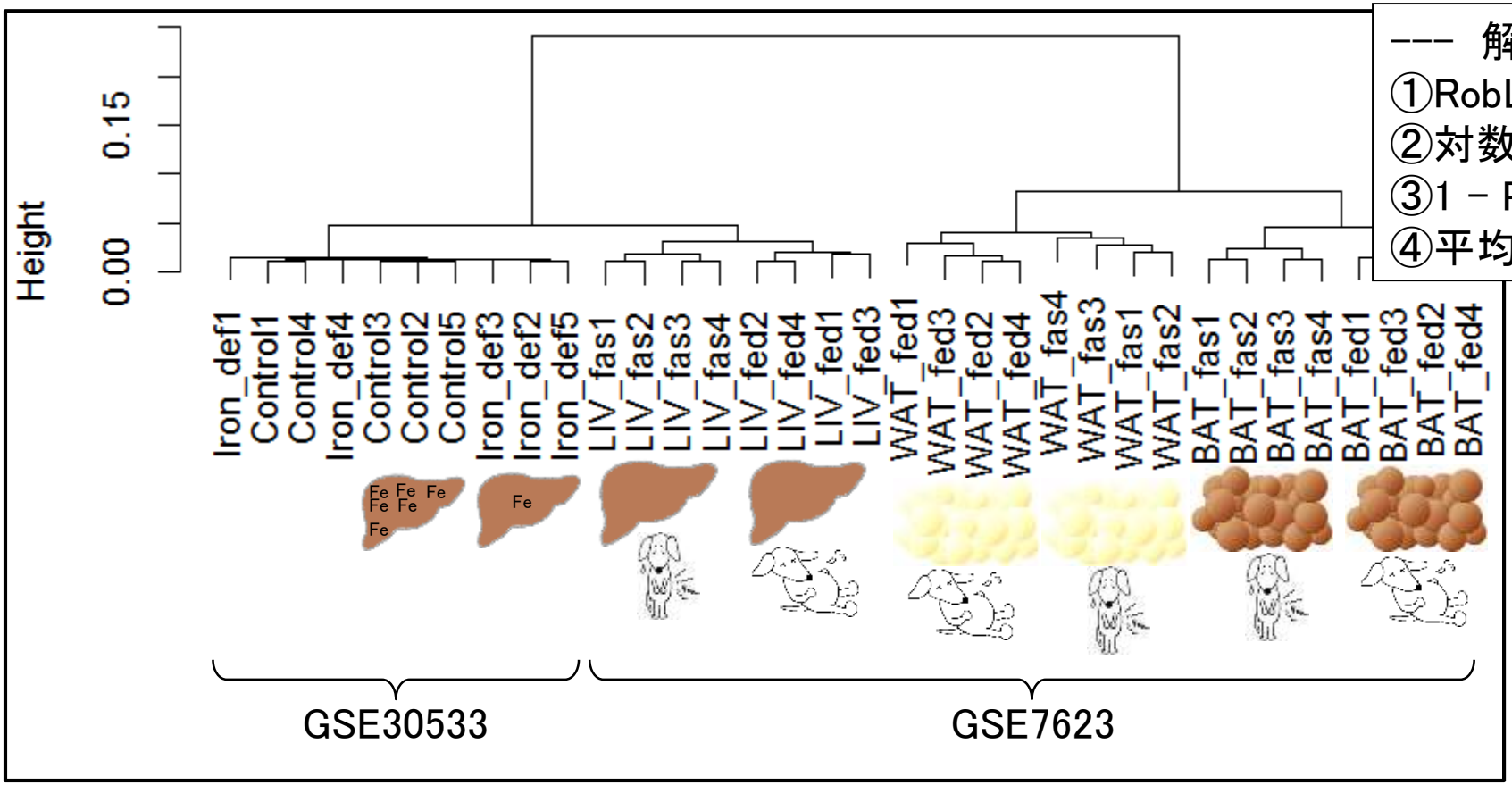
「1 – Pearson相関係数」の結果。English version

GSE7623 + GSE30533



課題2(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簡単に考察せよ



--- 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

課題2

主な論点はここで挙げたような事柄になると思いますが、基本的に自由に考察してよい

- GSE7623とGSE30533は独立した別々の論文
- GSE30533の由来サンプルは?
- 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?
- GSE30533のみでクラスタリングを行った結果のトポロジーは前処理法や距離の定義次第で変わりやすいが…。
- 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)で明瞭に分かれている(MAS5のWATサンプルを除く)。**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?

Contents

- 実データ概観
  - GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の§3.2.1)
  - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
  - 階層的 vs. 非階層的、様々な選択肢
  - 距離の定義:ベクトル間、クラスター間
  - 実データで実行:GSE2361 (ヒト)、**課題1**
  - 実データで実行:GSE7623 (ラット)、GSE30533 (ラット)
  - 同一プラットフォームデータ(GSE7623 + GSE30533)をマージして実行、**課題2**
- 実験デザイン(教科書の§3.2.2)
- 2群間比較:発現変動遺伝子(DEG)検出
  - パターンマッチング法(相関係数の利用)
    - コードの中身をおさらい、apply関数の基本的な利用法など

教科書p107-108

「個体間のばらつき」を考慮した実験デザインを組むべし、という話

実験デザイン（§3.2.2）

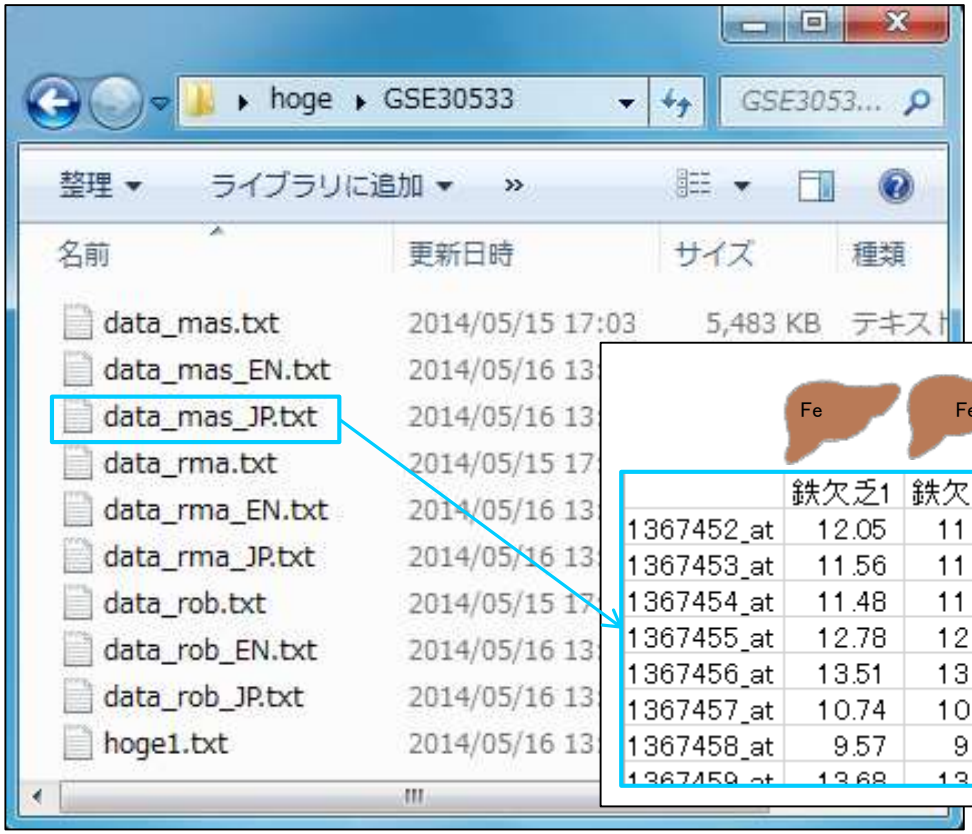
- Affymetrix GeneChip
  - Ge et al., *Genomics*, **86**: 127–141, 2005
    - **GSE2361**、GPL96 (Affymetrix Human Genome U133A Array)、22,283 probesets
    - **ヒト36サンプル**：Heart (心臓)、Thymus (胸腺)、Spleen (脾臓)、Ovary (卵巣)、Kidney (腎臓)、Skeletal Muscle (骨格筋)、Pancreas (膵臓)、Prostate (前立腺)、…
  - Nakai et al., *Biosci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

教科書p107-108

実験デザイン（§3.2.2）

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)

- 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



	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

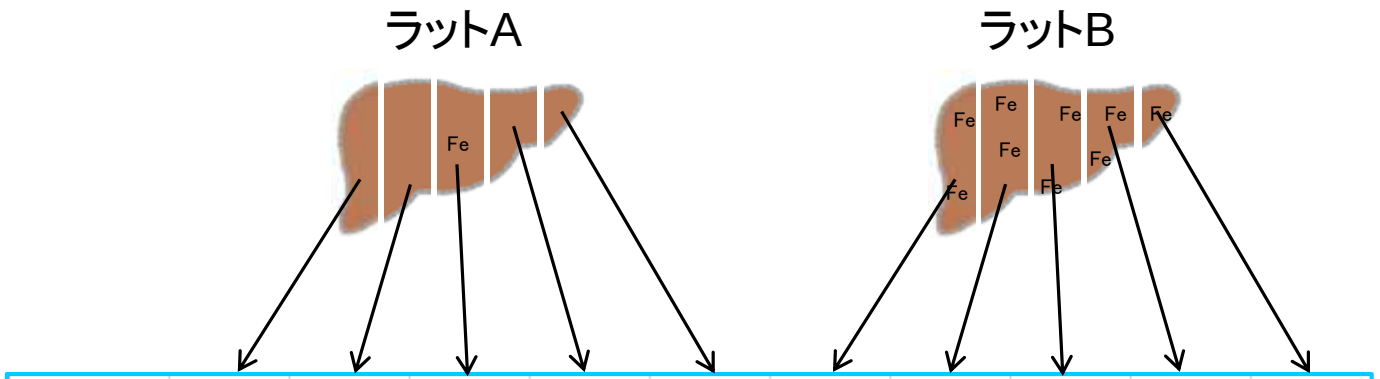
教科書p107-108

実験デザイン（§3.2.2）

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!

- 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）



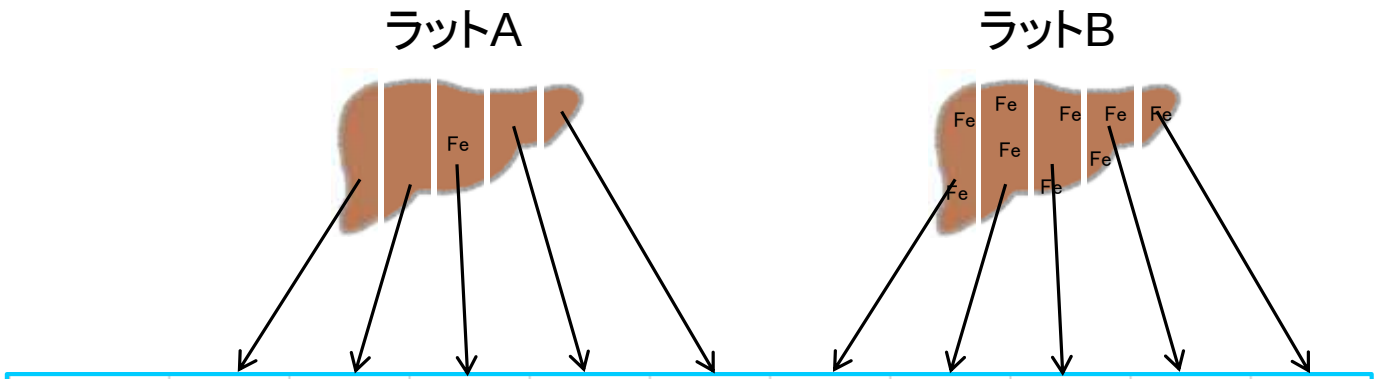
	鉄欠乏1	鉄欠乏2	鉄欠乏3	鉄欠乏4	鉄欠乏5	通常1	通常2	通常3	通常4	通常5
1367452_at	12.05	11.92	11.99	11.92	11.73	12.08	12.06	11.98	12.03	12.03
1367453_at	11.56	11.59	11.62	11.75	11.78	11.63	11.51	11.48	11.57	11.68
1367454_at	11.48	11.68	11.61	11.65	11.86	11.71	11.98	12.01	11.59	11.95
1367455_at	12.78	12.59	12.70	12.79	13.00	12.68	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

実験デザイン（§3.2.2）

2群間での発現変動遺伝子(DEG)検出結果は多くなる傾向。**多ければいいというものではない!**

- 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_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

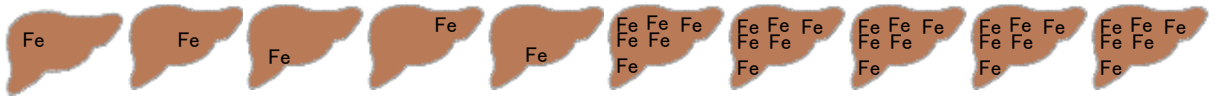
教科書p107-108

実験デザイン（§3.2.2）

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

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

家系A　　家系B

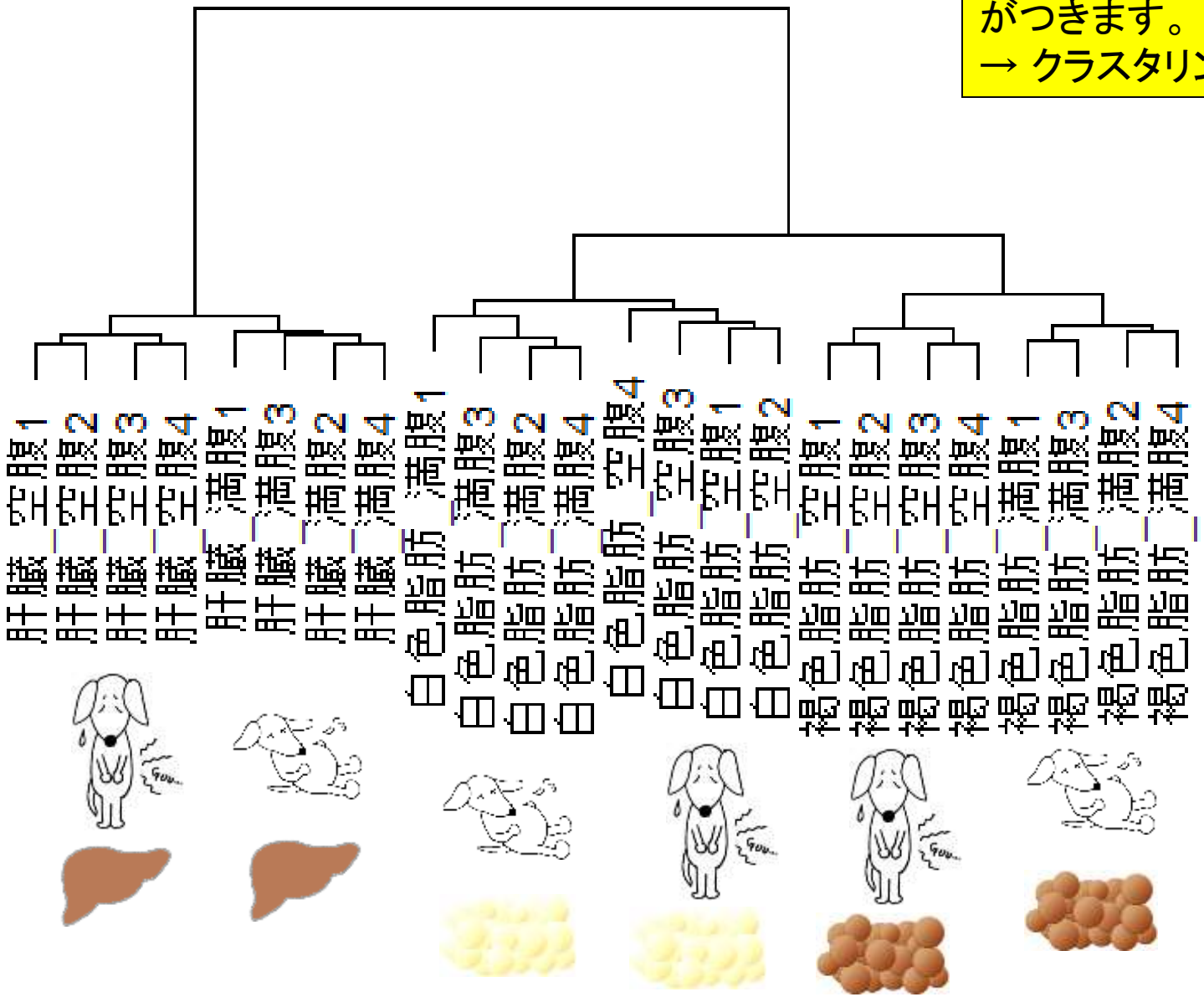


	鉄欠乏1	鉄欠乏2	鉄欠乏3	鉄欠乏4	鉄欠乏5	通常1	通常2	通常3	通常4	通常5
1367452_at	12.05	11.92	11.99	11.92	11.73	12.08	12.06	11.98	12.03	12.03
1367453_at	11								11.57	11.68
1367454_at	11								11.59	11.95
1367455_at	12								12.68	12.87
1367456_at	13								13.52	13.57
1367457_at	10								10.39	10.52
1367458_at	9.57	9.17	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

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。

クラスタリングと発現変動解析

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



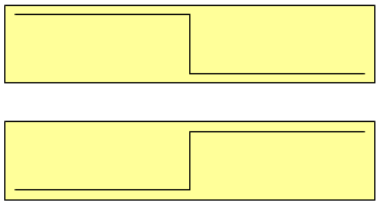
Contents

- 実データ概観
  - GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の§3.2.1)
  - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
  - 階層的 vs. 非階層的、様々な選択肢
  - 距離の定義:ベクトル間、クラスター間
  - 実データで実行:GSE2361 (ヒト)、**課題1**
  - 実データで実行:GSE7623 (ラット)、GSE30533 (ラット)
  - 同一プラットフォームデータ(GSE7623 + GSE30533)をマージして実行、**課題2**
- 実験デザイン(教科書の§3.2.2)
- 2群間比較:発現変動遺伝子(DEG)検出
  - パターンマッチング法(相関係数の利用)
    - コードの中身をおさらい、apply関数の基本的な利用法など

対数変換後のデータを用いて2群間比較

データ解析もいろいろ

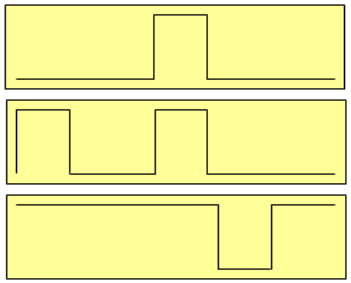
発現変動遺伝子同定



遺伝子発現行列

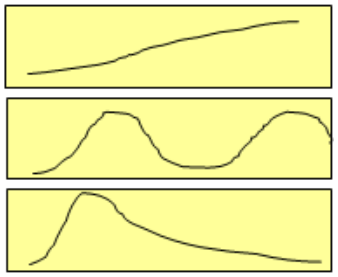
二群間比較用

	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）

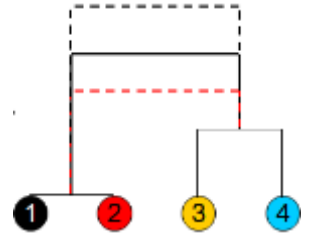
	S1	S2	S3	S4	…
gene 1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene 2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene n	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



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	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

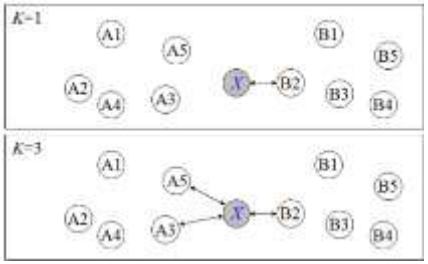
クラスタリング



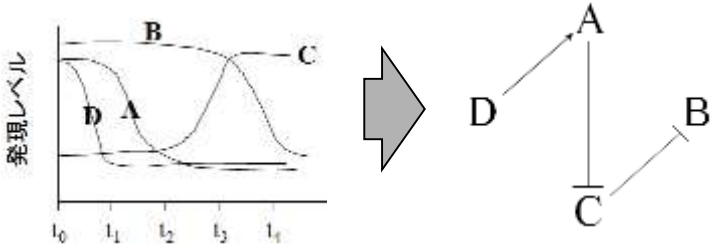
機能解析

- Gene Ontology（GO）
- パスウェイ解析

分類（診断）



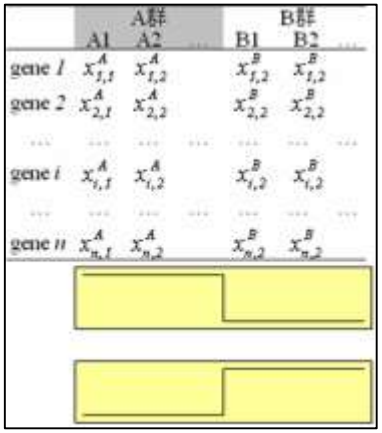
遺伝子ネットワーク推定

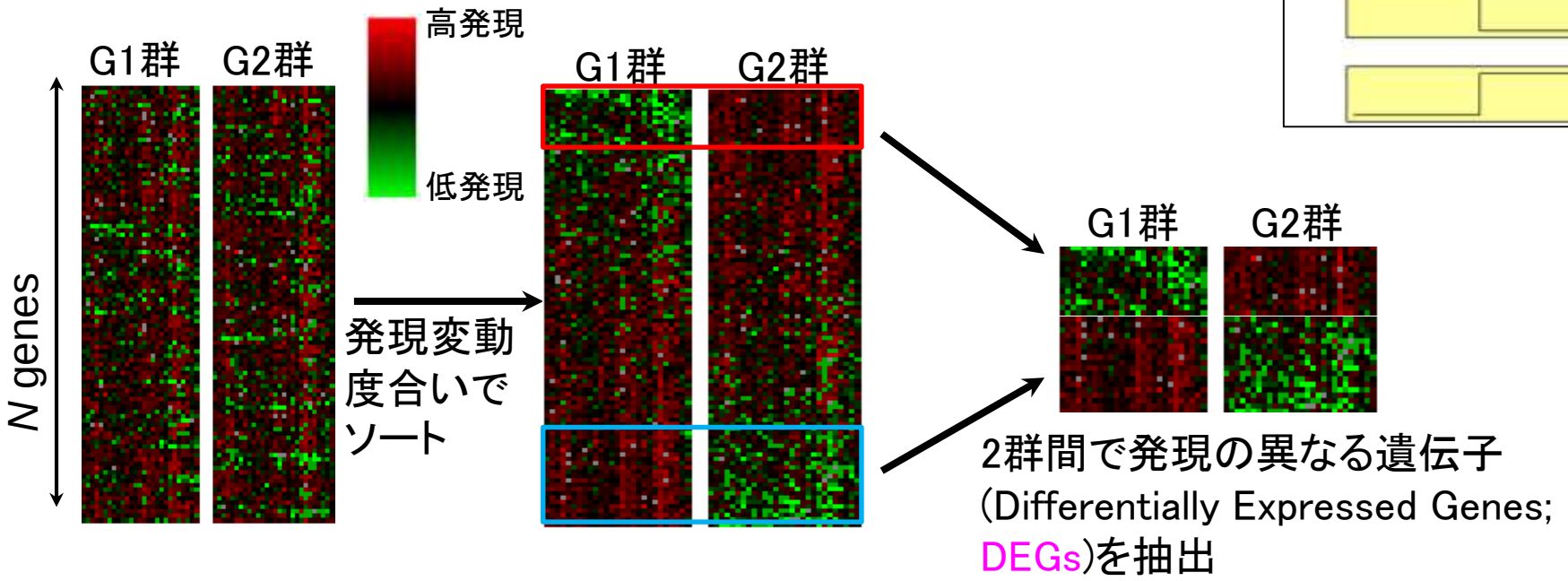


2群間比較

比較したいグループ(群)間で発現変動している遺伝子または転写物を同定することはデータ解析の基本

- 農産物の栽培条件の違い(通常 vs. 低温、通常 vs. 乾燥)
- 味の違い(おいしい vs. まずい)
- サンプルの状態の違い(癌 vs. 正常)



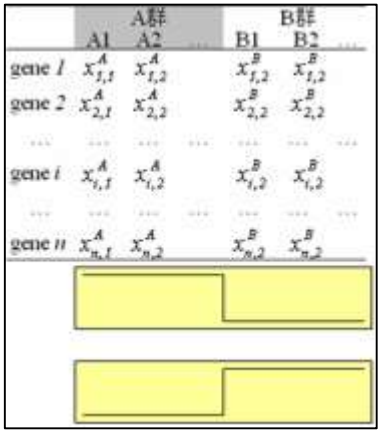


2群間比較

①理想的なパターンyと②遺伝子ベクトル間の
③相関係数rを遺伝子ごとに計算。rの絶対値が
大きいほど発現変動の度合いが大きいと解釈

- パターンマッチング法
  - 理想的なパターンyとの類似度が高い順に遺伝子をランキング

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



①

y	1	1	1	1	1	1	0	0	0	0	0

③

rownan	G1_1	G1_2	G1_3	G1_4	G1_5	G1_6	G2_1	G2_2	G2_3	G2_4	G2_5	r
gene1	6.44	6.30	6.51	6.36	6.49	6.39	3.58	4.39	4.25	3.70	4.09	0.98
gene2	5.81	6.93	6.73	5.55	6.39	6.61	2.81	5.46	1.00	3.46	4.17	0.81
gene3	3.91	4.81	5.04	3.17	4.75	5.36	5.58	5.52	5.70	5.64	5.61	−0.71

②

パターンマッチング法

①パターンマッチング法の、②例題1。③テンプレートパターン情報を含むファイルを読み込んでパターンマッチングを行ってみよう。まだコピペしなくてよい

(Rで)マイクロアレイデータ解析

(last modified 2015/05/16, since 2005)

What's new?

- 門田幸二
イ解析に関
どについて
このページ
- お知らせは
イ関連の講
ら辿れます

- はじめに(1

- 解析 | 発現変動 | 2群間 | 対応なし | Student's t-test(last modified 2014/05/25)
- 解析 | 発現変動 | 2群間 | 対応なし | Welch t-test(last modified 2014/05/23)
- 解析 | 発現変動 | 2群間 | 対応なし | Mann-Whitney U-test(last modified 2013/10/15)
- 解析 | 発現変動 | 2群間 | 対応なし | パターンマッチング法(last modified 2014/05/23)
- 解析 | 発現変動 | 2群間 | 対応あり | について(last modified 2009/11/11)
- 解析 | 発現変動 | 2群間 | 対応あり | SAM (Tusher, 2001) (last modified 2014/06/02)
- 解析 | 発現変動 | 2群間 | 対
- 解析 | 発現変動 | 2群間 | 対
- 解析 | 発現変動 | 2群間 | 対
- 解析 | 発現変動 | 3群間 | 対
- 解析 | 発現変動 | 3群間 | 対
- 解析 | 発現変動 | 3群間 | 対
- 解析 | 発現変動 | 3群間 | 対
- 解析 | 発現変動 | 3群間 | 対
- 解析 | 発現変動 | 多群間 |
- 解析 | 発現変動 | 多群間 |
- 解析 | 発現変動 | 多群間 |
- 解析 | 発現変動 | 多群間 |
- 解析 | 発現変動 | 多群間 |

① ② ③

解析 | 発現変動 | 2群間 | 対応なし | パターンマッチング法

パターンマッチング法を用いて、2群間での発現変動遺伝子の同定を行うやり方を紹介します。
「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. サンプルデータ16のsample16_log.txt(対数変換後のデータ)の場合:

クラスラベル情報ファイル(sample16_cl.txt)を利用するやり方です。

```
in_f1 <- "sample16_log.txt"          #入力ファイル名を指定してin_f1に格納(発現データ)
in_f2 <- "sample16_cl.txt"           #入力ファイル名を指定してin_f2に格納(テンプレート
out_f <- "hoge1.txt"                 #出力ファイル名を指定してout_fに格納
param <- "pearson"                   #相関係数の種類を指定("pearson"または"spearman")

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定し
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]                  #テンプレートパターンベクトルdata.clを作成

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子についてテンプレートパタ

#ファイルに保存
tmp <- cbind(rownames(data), data, r)  #入力データの右側に相関係数rのベクトルを結合した結
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定した
```

入出力の関係

①テンプレートパターン情報ファイル。2列目がテンプレートパターン。②(log変換後の)発現データファイル。赤枠で示すように、列の並びは同じ

解析 | 発現変動 | 2群間 | 対応なし | パターンマッチング法

パターンマッチング法を用いて、2群間での発現変動遺伝子の同定を行うやり方を紹介します。
「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. サンプルデータ16のsample16_log.txt(対数変換後のデータ)の場合:

クラスラベル情報ファイル(sample16_cl.txt)を利用するやり方です。

```
in_f1 <- "sample16_log.txt"          #入力ファイル名を指定してin_f1に格納(発現データ)
in_f2 <- "sample16_cl.txt"           #入力ファイル名を指定してin_f2に格納(テンプレート
out_f <- "hoge1.txt"                 #出力ファイル名を指定してout_fに格納
param <- "pearson"                   #相関係数の種類を指定("pearson"または"spearman")

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定し
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]                  #テンプレートパターンベクトルdata.clを作成

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子についてテンプレートパタ

#ファイルに保存
tmp <- cbind(rownames(data), data, r)  #入力データの右側に相関係数rのベクトルを結合した結
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定した
```

G1_1	1
G1_2	1
G1_3	1
G1_4	1
G1_5	1
G1_6	1
G2_1	0
G2_2	0
G2_3	0
G2_4	0
G2_5	0

①

rownar	G1_1	G1_2	G1_3	G1_4	G1_5	G1_6	G2_1	G2_2	G2_3	G2_4	G2_5
gene1	6.44	6.30	6.51	6.36	6.49	6.39	3.58	4.39	4.25	3.70	4.09
gene2	5.81	6.93	6.73	5.55	6.39	6.61	2.81	5.46	1.00	3.46	4.17
gene3	3.91	4.81	5.04	3.17	4.75	5.36	5.58	5.52	5.70	5.64	5.61

②

入出力の関係

①出力ファイルは、②入力ファイルの右側に相関係数rの計算結果が追加されたもの

解析 | 発現変動 | 2群間 | 対応なし | パターンマッチング法

パターンマッチング法を用いて、2群間での発現変動遺伝子の同定を行うやり方を紹介します。
「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. サンプルデータ16のsample16_log.txt(対数変換後のデータ)の場合:

クラスラベル情報ファイル(sample16_cl.txt)を利用するやり方です。

②

①

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定してin_f1に格納(発現データ)
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定してin_f2に格納(テンプレート
out_f <- "hoge1.txt"                     #出力ファイル名を指定してout_fに格納
param <- "pearson"                       #相関係数の種類を指定("pearson"または"spearman")

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定し
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]                      #テンプレートパターンベクトルdata.clを作成

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子についてテンプレートパタ

#ファイルに保存
tmp <- cbind(rownames(data), data, r)  #入力データの右側に相関係数rのベクトルを結合した結
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定した
```

G1_1	1
G1_2	1
G1_3	1
G1_4	1
G1_5	1
G1_6	1
G2_1	0
G2_2	0
G2_3	0
G2_4	0
G2_5	0

①

rownan	G1_1	G1_2	G1_3	G1_4	G1_5	G1_6	G2_1	G2_2	G2_3	G2_4	G2_5	r
gene1	6.44	6.30	6.51	6.36	6.49	6.39	3.58	4.39	4.25	3.70	4.09	0.984
gene2	5.81	6.93	6.73	5.55	6.39	6.61	2.81	5.46	1.00	3.46	4.17	0.811
gene3	3.91	4.81	5.04	3.17	4.75	5.36	5.58	5.52	5.70	5.64	5.61	−0.708

コードの解説

黒枠内をコピペ。ここは、①入力ファイル情報読み込み部分。②dataオブジェクトを確認

1. サンプルデータ16のsample16_log.txt(対数変換後のデータ)の場合:

クラスラベル情報ファイル(sample16_cl.txt)を利用するやり方です。

```
in_f1 <- "sample16_log.txt"          #入力ファイル名を指定してin_f1に格納(発現データ)
in_f2 <- "sample16_cl.txt"           #入力ファイル名を指定してin_f2に格納(テンプレート情報)
out_f <- "hoge1.txt"                 #出力ファイル名を指定してout_fに格納
param <- "pearson"                   #相関係数の種類を指定("pearson"または"spearman")

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=
hoge <- read.table(in_f2, sep="\t
data.cl <- hoge[,2]

#本番
r <- apply(data, 1, cor, y=data.c

#ファイルに保存
tmp <- cbind(rownames(data), data
write.table(tmp, out_f, sep="\t",
```

①

R Console

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files(pattern="16")
[1] "sample16_cl.txt"  "sample16_log.txt"
> in_f1 <- "sample16_log.txt"          #入力ファイル名を指$
> in_f2 <- "sample16_cl.txt"           #入力ファイル名を指$
> out_f <- "hoge1.txt"                 #出力ファイル名を指$
> param <- "pearson"                   #相関係数の種類を指$
>
> #入力ファイルの読み込みとラベル情報の作成
> data <- read.table(in_f1, header=TRUE, row.names=1, sep="\$
> data
      G1_1 G1_2 G1_3 G1_4 G1_5 G1_6 G2_1 G2_2 G2_3 G2_4 G2_5
gene1 6.44 6.30 6.51 6.36 6.49 6.39 3.58 4.39 4.25 3.70 4.09
gene2 5.81 6.93 6.73 5.55 6.39 6.61 2.81 5.46 1.00 3.46 4.17
gene3 3.91 4.81 5.04 3.17 4.75 5.36 5.58 5.52 5.70 5.64 5.61
>
```

①

②

コードの解説

①テンプレートパターン情報の②読み込み時にheader=TRUEやrow.names=1の記述がない点に注目!

1. サンプルデータ16のsample16_log.txt(対数変換後のデータ)の場合:

クラスラベル情報ファイル(sample16_cl.txt)を利用するやり方です。

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定してin_f1に格納(発現データ)
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定してin_f2に格納(テンプレート情報)
out_f <- "hoge1.txt"                     #出力ファイル名を指定してout_fに格納
param <- "pearson"                       #相関係数の種類を指定("pearson"または"spearman")

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定したファイルの読み込み
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]                      #テンプレートパターンベクトルdata.clを作成

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子についてテンプレートパターンdata.clとの相関

#ファイルに保存
tmp <- cbind(rownames(data), data
write.table(tmp, out_f, sep="\t",
```

G1_1	1
G1_2	1
G1_3	1
G1_4	1
G1_5	1
G1_6	1
G2_1	0
G2_2	0
G2_3	0
G2_4	0
G2_5	0

②

①

R Console

```
> data
      G1_1 G1_2 G1_3 G1_4 G1_5 G1_6 G2_1 G2_2 G2_3 G2_4 G2_5
gene1 6.44 6.30 6.51 6.36 6.49 6.39 3.58 4.39 4.25 3.70 4.09
gene2 5.81 6.93 6.73 5.55 6.39 6.61 2.81 5.46 1.00 3.46 4.17
gene3 3.91 4.81 5.04 3.17 4.75 5.36 5.58 5.52 5.70 5.64 5.61
> hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定$
> data.cl <- hoge[,2]                          #テンプレートパター$
> data.cl
 [1] 1 1 1 1 1 1 0 0 0 0 0
>
```

コードの解説

①テンプレートパターン情報を、②の部分で読み込んで得られた、③hogeの中身は入力ファイルと同じだが…

1. サンプルデータ16のsample16_log.txt(対数変換後のデータ)の場合:

クラスラベル情報ファイル(sample16_cl.txt)を利用するやり方です。

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定してin_f1に格納(発現データ)
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定してin_f2に格納(テンプレート情報)
out_f <- "hoge1.txt"                     #出力ファイル名を指定してout_f1に格納
param <- "pearson"                       #相関係数の種類を指定("

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.nam=1, sep="\t"
hoge <- read.table(in_f2, sep="\t", quote="")  f2で指定したフ
data.cl <- hoge[,2]                      #テンプレートパターンベク

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝

#ファイルに保存
tmp <- cbind(rownames(data), data, r)  #入力データの右側に相関係
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names
```

②

③

R Console

```
> hoge
      V1 V2
1   G1_1  1
2   G1_2  1
3   G1_3  1
4   G1_4  1
5   G1_5  1
6   G1_6  1
7   G2_1  0
8   G2_2  0
9   G2_3  0
10  G2_4  0
11  G2_5  0
> dim(hoge)
[1] 11  2
> hoge[3, ]
    V1 V2
3 G1_3  1
> hoge[, 2]
 [1] 1 1 1 1 1 1 0 0 0 0 0
>
```

G1_1	1
G1_2	1
G1_3	1
G1_4	1
G1_5	1
G1_6	1
G2_1	0
G2_2	0
G2_3	0
G2_4	0
G2_5	0

①

コードの解説

①欲しいのはhoge行列の2列目部分のみなので、②そこの部分のみ抽出

1. サンプルデータ16のsample16_log.txt(対数変換後のデータ)の場合:

クラスラベル情報ファイル(sample16_cl.txt)を利用するやり方です。

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定してin_f1に格納(発現データ)
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定してin_f2に格納(テンプレート情報)
out_f <- "hoge1.txt"                     #出力ファイル名を指定してout_f1に格納
param <- "pearson"                       #相関係数の種類を指定("p

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t",
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したフ
data.cl <- hoge[,2]                      #テンプレートパターンベク

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子

#ファイルに保存
tmp <- cbind(rownames(data), data, r)  #入力データの右側に相関係
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names
```

②

G1_1	1
G1_2	1
G1_3	1
G1_4	1
G1_5	1
G1_6	1
G2_1	0
G2_2	0
G2_3	0
G2_4	0
G2_5	0

①

R Console

```
> hoge
      V1 V2
1   G1_1  1
2   G1_2  1
3   G1_3  1
4   G1_4  1
5   G1_5  1
6   G1_6  1
7   G2_1  0
8   G2_2  0
9   G2_3  0
10 G2_4  0
11 G2_5  0
> dim(hoge)
[1] 11  2
> hoge[3, ]
    V1 V2
3 G1_3  1
> hoge[, 2]
 [1] 1 1 1 1 1 1 0 0 0 0 0
>
```

①

②

Tips

もちろん、読み込み時に①row.names=1をつけてもよい。その場合は黒下線部分を②2から1に変更

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定してin_f1に格納(発現データ)
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定してin_f2に格納(テンプレート情報)
out_f <- "hoge1.txt"                     #出力ファイル名を指定してout_fに格納
param <- "pearson"                       #相関係数の種類を指定("pearson"または"spearman")

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定したファイルの読み込み
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]

#本番
r <- apply(data, 1, cor,

#ファイルに保存
tmp <- cbind(rownames(dat
write.table(tmp, out_f, s
```

R Console

```
> hoge <- read.table(in_f2, row.names=1, sep="\t", quote="")
> hoge
      V2
G1_1   1
G1_2   1
G1_3   1
G1_4   1
G1_5   1
G1_6   1
G2_1   0
G2_2   0
G2_3   0
G2_4   0
G2_5   0
> data.cl <- hoge[, 1]
> data.cl
 [1] 1 1 1 1 1 1 0 0 0 0 0
>
```

①

②

G1_1	1
G1_2	1
G1_3	1
G1_4	1
G1_5	1
G1_6	1
G2_1	0
G2_2	0
G2_3	0
G2_4	0
G2_5	0

コードの解説：apply

apply関数は行ごとや列ごとに同じ関数を繰り返して実行させたい場合に便利。①dataオブジェクトの、②各行に対して、③cor関数を適用せよ。その際、④テンプレートyはdata.clとし、⑤相関係数の種類はparamで指定したものとする

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定し
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定し
out_f <- "hoge1.txt"                     #出力ファイル名を指定し
param <- "pearson"                       #相関係数の種類を指定("pearson" または "spearman")

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定したファイルの読み込み
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]                      #テンプレートパターンベクトルdata.clを作成

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子についてテンプレートパターンdata.clとの相関

#ファイルに保存
tmp <- cbind(rownames(data), data, r)  #入力データの右側に相関係数rのベクトルを結合した結果をtmpに格納。
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名で保存
```

R Console

```
> #本番          ①  ②  ③        ④                ⑤
> r <- apply(data, 1, cor, y=data.cl, method=param)#$
> r
     gene1      gene2      gene3
 0.9839815  0.8107810 -0.7068671
>
```

Tips：cor, as.numeric

①apply関数を使わずに、遺伝子（つまり行）ごとにcor関数を用いて相関係数を計算する基本手順。②as.numeric関数は、data.clオブジェクトとデータの型を揃える目的で利用している。③as.numericなしの場合と比較すればよい

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定して
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定して
out_f <- "hoge1.txt"                     #出力ファイル名を指定して
param <- "pearson"                       #相関係数の種類を指定("pearson" または "spearman")

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定したファイルの読み込み
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]                      #テンプレートパターンベクトルdata.clを作成

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子についてテンプレートパターンdata.clとの相関

#ファイルに保存
tmp <- cbind(r
write.table(tm
```

R Console

```
> r
     gene1       gene2       gene3
 0.9839815   0.8107810  -0.7068671
> data[1, ]
      G1_1 G1_2 G1_3 G1_4 G1_5 G1_6 G2_1 G2_2 G2_3 G2_4 G2_5
gene1 6.44  6.3 6.51 6.36 6.49 6.39 3.58 4.39 4.25  3.7 4.09
> data.cl
 [1] 1 1 1 1 1 1 0 0 0 0 0
> cor(data[1, ], data.cl, method=param)
 以下にエラー cor(data[1, ], data.cl, method = param) :  互換性のない次元です
> as.numeric(data[1, ])
 [1] 6.44 6.30 6.51 6.36 6.49 6.39 3.58 4.39 4.25 3.70 4.09
> cor(as.numeric(data[1, ]), data.cl, method=param)
[1] 0.9839815
>
```

①
②
③

Tips：cor, as.numeric

①ではエラーが出ているが、②as.numericをつけたことで、無事gene1とテンプレートパターン間の相関係数rを計算できていることがわかる

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定してin_f1に格納(発現データ)
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定してin_f2に格納(テンプレート情報)
out_f <- "hoge1.txt"                     #出力ファイル名を指定してout_fに格納
param <- "pearson"                       #相関係数の種類を指定("pearson"または"spearman")

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定したファイルの読み込み
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]                      #テンプレートパターンベクトルdata.clを作成

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子についてテンプレートパターンdata.clとの相関

#ファイルに保存
tmp <- cbind(r
write.table(tm
```

R Console

```
> r
     gene1       gene2       gene3
 0.9839815   0.8107810  -0.7068671
> data[1, ]
      G1_1 G1_2 G1_3 G1_4 G1_5 G1_6 G2_1 G2_2 G2_3 G2_4 G2_5
gene1 6.44  6.3 6.51 6.36 6.49 6.39 3.58 4.39 4.25  3.7 4.09
> data.cl
 [1] 1 1 1 1 1 1 0 0 0 0 0
① > cor(data[1, ], data.cl, method=param)
 以下にエラー cor(data[1, ], data.cl, method = param) :  互換性のない次元です
> as.numeric(data[1, ])
 [1] 6.44 6.30 6.51 6.36 6.49 6.39 3.58 4.39 4.25 3.70 4.09
② > cor(as.numeric(data[1, ]), data.cl, method=param)
[1] 0.9839815
>
```

Tips：データの型

①「互換性のない次元です」と言われているのは、「相関係数を計算するときの入力データの見栄えが異なる」と文句を言われていると解釈すればよい。この「見栄え」は、Rで「データの型」と呼ばれるものに相当します。②と③が同じ見栄え（型）なのでcor関数の要求を満たしています

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定
out_f <- "hoge1.txt"                     #出力ファイル名を指定
param <- "pearson"                       #相関係数の種類を指定

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote="")#in_f1で指定したファイルの読み込み
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]                      #テンプレートパターンベクトルdata.clを作成

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子についてテンプレートパターンdata.clとの相関

#ファイルに保存
tmp <- cbind(r
write.table(tm
```

R Console

```
> r
     gene1       gene2       gene3
 0.9839815   0.8107810  -0.7068671
> data[1, ]
      G1_1 G1_2 G1_3 G1_4 G1_5 G1_6 G2_1 G2_2 G2_3 G2_4 G2_5
gene1 6.44  6.3 6.51 6.36 6.49 6.39 3.58 4.39 4.25  3.7 4.09
> data.cl
 [1] 1 1 1 1 1 1 0 0 0 0 0 ③
> cor(data[1, ], data.cl, method=param)
 以下にエラー cor(data[1, ], data.cl, method = param) :  互換性のない次元です ①
> as.numeric(data[1, ])
 [1] 6.44 6.30 6.51 6.36 6.49 6.39 3.58 4.39 4.25 3.70 4.09 ②
> cor(as.numeric(data[1, ]), data.cl, method=param)
[1] 0.9839815
>
```

Tips：as.vector, mode

慣れてくるとas.numericを使えばいいとすぐにわかるが、はじめのうちはas.character, as.integer, ①as.vectorなど既知のas…関数候補の中からそれっぽいものを試して型（見栄え）が揃うものを探すのが一般的かも…。②modeというデータの型を表示する関数もあるが…私は試行錯誤派です

```
in_f1 <- "sample16_log.txt"              #入力ファイル名を指定してin_f
in_f2 <- "sample16_cl.txt"               #入力ファイル名を指定してin_f
out_f <- "hoge1.txt"                     #出力ファイル名を指定してout_
param <- "pearson"                       #相関係数の種類を指定("pears

#入力ファイルの読み込みとラベル情報の作成
data <- read.table(in_f1, header=TRUE, row.names=1, sep="\t", quote=
hoge <- read.table(in_f2, sep="\t", quote="")#in_f2で指定したファイルの読み込み
data.cl <- hoge[,2]                      #テンプレートパターンベクトルdata.clを作成

#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子についてテンプレートパターンdata.clとの相関

#ファイルに保存
tmp <- cbind(r
write.table(tm
```

R Console

```
> data.cl
 [1] 1 1 1 1 1 1 0 0 0 0 0
> cor(data[1, ], data.cl, method=param)
 以下にエラー cor(data[1, ], data.cl, method = param) :  互換性のない次元です
> as.numeric(data[1, ])
 [1] 6.44 6.30 6.51 6.36 6.49 6.39 3.58 4.39 4.25 3.70 4.09
> cor(as.numeric(data[1, ]), data.cl, method=param)
[1] 0.9839815
① > as.vector(data[1, ])
      G1_1 G1_2 G1_3 G1_4 G1_5 G1_6 G2_1 G2_2 G2_3 G2_4 G2_5
gene1 6.44  6.3 6.51 6.36 6.49 6.39 3.58 4.39 4.25  3.7 4.09
② > mode(data[1, ])
[1] "list"
>
```

コードの解説：cbind

出力部分。cbind関数を用いて出力させたい順番で列(column)方向で結合(bind)したものがtmpオブジェクト。尚、行(row)方向で結合させたい場合は、rbind関数を用いる

```
#本番
r <- apply(data, 1, cor, y=data.cl, method=param)#各（行）遺伝子に

#ファイルに保存
tmp <- cbind(rownames(data), data, r)  #入力データの右側に相関係数rのベクトルを結合した結果をtmpに格納。
write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名で保存
```

R Console

```
> #ファイルに保存
> tmp <- cbind(rownames(data), data, r)  #入力データの右側に相関係数rのベクトルを結合$
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定し$
> rownames(data)
[1] "gene1" "gene2" "gene3"
> data
      G1_1 G1_2 G1_3 G1_4 G1_5 G1_6 G2_1 G2_2 G2_3 G2_4 G2_5
gene1 6.44 6.30 6.51 6.36 6.49 6.39 3.58 4.39 4.25 3.70 4.09
gene2 5.81 6.93 6.73 5.55 6.39 6.61 2.81 5.46 1.00 3.46 4.17
gene3 3.91 4.81 5.04 3.17 4.75 5.36 5.58 5.52 5.70 5.64 5.61
> r
     gene1      gene2      gene3
 0.9839815  0.8107810 -0.7068671
> tmp
      rownames(data) G1_1 G1_2 G1_3 G1_4 G1_5 G1_6 G2_1 G2_2 G2_3 G2_4 G2_5          r
gene1          gene1 6.44 6.30 6.51 6.36 6.49 6.39 3.58 4.39 4.25 3.70 4.09  0.9839815
gene2          gene2 5.81 6.93 6.73 5.55 6.39 6.61 2.81 5.46 1.00 3.46 4.17  0.8107810
gene3          gene3 3.91 4.81 5.04 3.17 4.75 5.36 5.58 5.52 5.70 5.64 5.61 -0.7068671
>
```

①

①