

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。

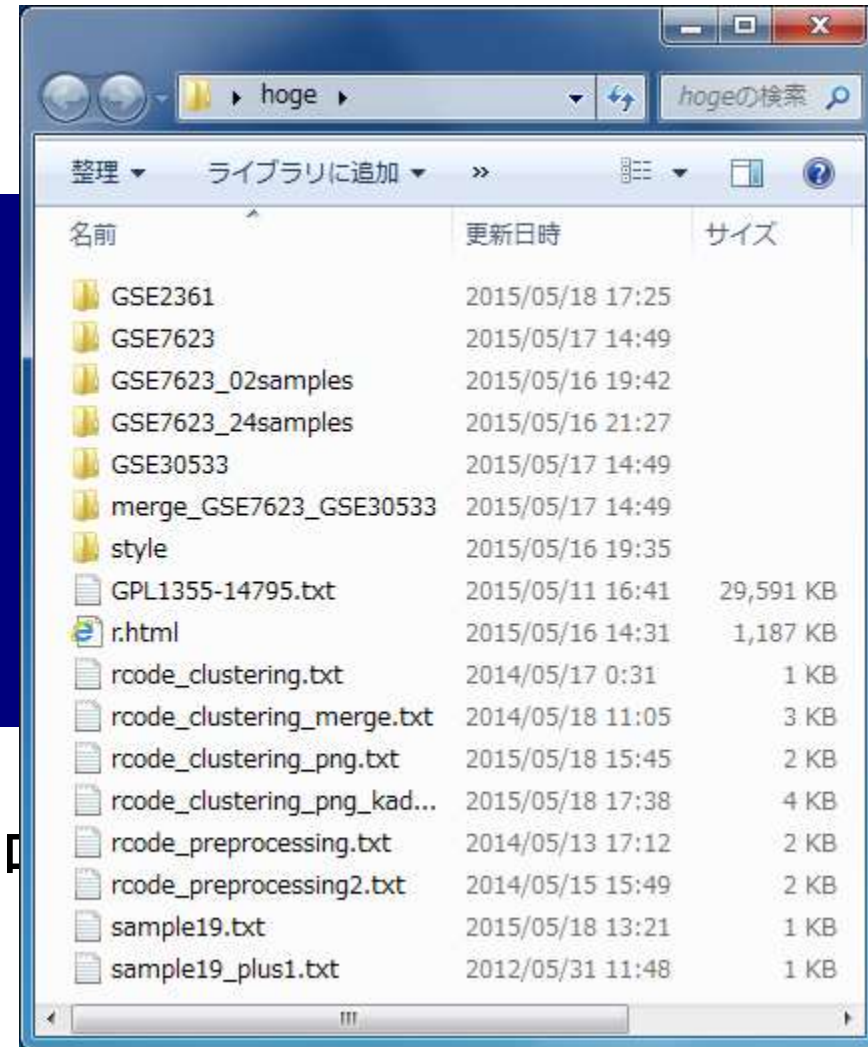
機能ゲノム学 第2回

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ジェクト

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/>



前回(5/12)のhogeフォルダがデスクトップに残っているかもしれないのでご注意ください。

講義予定

細胞中で発現している全転写物(トランスクリプトーム)の解析技術は、マイクロアレイから次世代シーケンサ(RNA-seq)に移行しつつあります。しかしRNA-seqデータ解析の多くは、マイクロアレイの知識を前提としています。本科目では、マイクロアレイデータを主な例として、各種トランスクリプトーム解析手法について解説します。

■ 第1回(2015年5月12日)

- 原理、各種データベース、生データ取得
- 教科書の1.2節、2.2節周辺

■ 第2回(2015年5月19日)

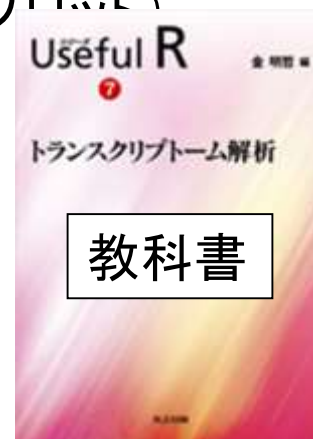
- 遺伝子発現行列作成(データ正規化)
- クラスタリング(データ変換や距離の定義など)、実験デザイン、分布
- 教科書の3.2節周辺

■ 第3回(2015年5月26日)

- 発現変動解析(多重比較問題)、各種プロット(M-A plotや平均-分散プロット)
- 教科書の3.2節と4.2節周辺

■ 第4回(2015年6月9日)

- 機能解析(Gene Ontology解析やパスウェイ解析)、分類など



Contents

- 前処理法の適用(プローブレベルデータ → 発現行列データ)
 - MAS法、RMA法、RMX法 (RobLoxBioC)、IRON法 (教科書の § 2.2.2~2.2.4)
 - データの正規化 (グローバル正規化、quantile正規化)、課題1
 - 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2
 - 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3
- 実験デザイン (教科書の § 3.2.2)

R経由で生データ取得

■ Affymetrix GeneChip

- Ge et al., *Genomics*, 86: 127–141, 2005
 - GSE2361、ヒト36サンプル、GPL96を利用
- Nakai et al., *BBB.*, 72: 139–148, 2008
 - GSE7623、ラット24サンプル、GPL1355を利用
- Kamei et al., *PLoS One*, 8: e65732, 2013
 - GSE30533、ラット10サンプル、GPL1355を利用

■ Illumina BeadChip

- Sharma et al., *Cancer Cell*, 23: 35–47, 2013
 - GSE28680、ヒト24サンプル、GPL10558を利用

■ NGSデータも…

- Neyret-Kahn et al., *Genome Res.*, 23: 1563–1579, 2013
 - GSE42213、ヒト26サンプル、GPL10999とGPL11154を利用
 - GSE42211、ヒト20サンプル、GPL10999とGPL11154を利用 (ChIP-seq)
 - GSE42212、ヒト6サンプル、GPL10999を利用 (RNA-seq)
- Huang et al., *Development*, 139: 2161–2169, 2012
 - GSE36469、シロイヌナズナ8サンプル、GPL13222を利用

R経由で生データ取得

復習: R経由でダウンロードしたzip圧縮ファイルの解凍を行うだけで、目的のCELファイル群のみからなるフォルダを得ることができる。

- 書籍 | トランスクリプトーム解析 | [4.2.3 多群間比較\(特異的発現パターン\)](#) (last modified 2014/04/20)
- イントロ | 発現データ取得 | [公共DBから](#) (last modified 2014/05/11)
- イントロ | 発現データ取得 | [inSilicoDb\(Taminau 2011\)](#) (last modified 2015/05/11) **NEW**
- イントロ | 発現データ取得 | **①** [ArrayExpress\(Kauffmann 2009\)](#) (last modified 2014/05/15) **推奨**
- イントロ | 発現データ取得 | [GEOquery\(Davis 2007\)](#) (last modified 2013/08/20)
- イントロ | アノテーション情報取得 | [公共DB\(GEO\)から](#) (last modified 2013/08/18)

イントロ | 発現データ取得 | [ArrayExpress\(Kauffmann_2009\)](#) **NEW**

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

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

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

② 3. Affymetrixデータ [GSE7623 \(Nakai et al., BBB, 2008\)](#) のCELファイルを取得したい場合:

```
out_f <- "E"
param <- "E"

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

```
param <- "GSE7623" #入手したいIDを指定

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


#前処理(データ取得)
hoge <- getAE(param, type="raw", extract=F) #paramで指定したIDの生データをダウン
```

データ解析の全体像

マイクロアレイ

プローブレベル数値データ(CELファイル)を入力として、発現行列データを出力するのが前処理法(preprocessing method)。

RNA-seq

公共データ取得	GEO, ArrayExpress	GEO, ArrayExpress, NCBI SRA, EBI ENA, DDBJ SRA (DRA)
解析対象生物種	配列情報既知(アレイが提供されているもののみ)	モデル・非モデル問わず
生データ	プローブレベル数値データ	塩基配列(数億リード程度、数百塩基長)
		QC (Quality Control): クオリティチェック、フィルタリング、トリミング アセンブリでトランスクリプトーム配列取得(マッピング時のリファレンスとしても利用) マッピング(bowtie2, TopHat2など)でSAM/BAMファイル取得
発現行列作成	前処理法(MAS5, RMAなど)適用後に遺伝子発現行列を得る	アノテーションファイルを利用してカウントデータ、配列長補正後のRPKM/FPKM、転写物レベルの発現情報など取得
発現変動遺伝子(DEG)同定	基本Rを利用(limma, SAM, Rank productsなど)	基本Rを利用(cuffdiff2, edgeR, DESeq2, TCCなど)
機能解析	GSEA, GSA, CytoscapeなどR/パッケージSeqGSEAなどを利用。	

よく使われているのはMAS5とRMAです

前処理法

(Rで)マイクロアレイデータ解析 (last modified 2015/05/16, since 2005)

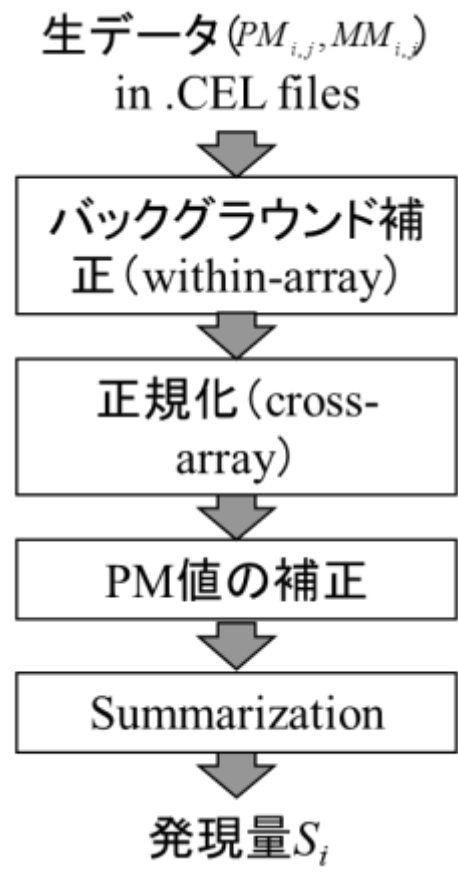
- What's new
- 門田幸一解析
- どについて
- このページ
- お知らせ
- 関連の
- ら連れま
- はじめに
- 過去のお
- インスト
- インスト
- インスト
- インスト
- インスト

- 正規化 | Illumina BeadArray | [BeadDataPackR\(Smith 2010\)](#) (last modified 2013/05/30)
- 正規化 | Illumina BeadArray | [lumi\(Du 2008\)](#) (last modified 2013/05/30)
- 正規化 | Illumina BeadArray | [beadarray\(Dunning 2007\)](#) (last modified 2013/05/30)
- 正規化 | Affymetrix GeneChip | [|について](#) (last modified 2015/05/16) **NEW**
- 正規化 | Affymetrix GeneChip | [frma\(McCall 2010\)](#) (last modified 2013/08/21)

正規化 | Affymetrix GeneChip | について **NEW**

2015年5月に調査した結果をリストアップします。

- [MBEI: Li and Wong, PNAS, 2001](#)
- [VSN: Huber et al., Bioinformatics, 2002](#)
- [MAS5.0: Hubbell et al., Bioinformatics, 2002](#)
- [RMA: Irizarry et al., Biostatistics, 2003](#)
- [GCRMA: Wu et al., J. Am. Stat. Assoc., 2004](#)
- [multi-mgMOS: Liu et al., Bioinformatics, 2005](#)
- [FARMS: Hochreiter et al., Bioinformatics, 2006](#)
- [Extrapolation Averaging \(EA\): Goldstein, DR, Bioinformatics, 2006](#)
- [refRMA: Katz et al., BMC Bioinformatics, 2006](#)
- [DFW: Chen et al., Bioinformatics, 2007](#)
- [libaffy: Eschrich et al., Bioinformatics, 2007](#)
- [RefPlus\(RMA++ and RMA+\): Harbron et al., Bioinformatics, 2007](#)
- [Hook: Binder et al., Algorithms Mol. Biol., 2008](#)
- [GRSN: Pelz et al., BMC Bioinformatics, 2008](#)
- [frMA: McCall et al., Biostatistics, 2010](#)
- [tRMA: Giorgi et al., BMC Bioinformatics, 2010](#)
- [rmx: Kohl and Deigner, BMC Bioinformatics, 2010](#)
- [KDL and KDQ \(SAS code\): Hsieh et al., BMC Bioinformatics, 2011](#)
- [RPA: Lahti et al., Nucleic Acids Res., 2013](#)
- [IRON in libaffy: Welsh et al., BMC Bioinformatics, 2013](#)



よく使われているのはMAS5とRMAです

前処理法

- **MAS5** (Hubbell et al., *Bioinformatics*, 2002)
 - 特徴: アレイごとに独立して前処理を実行 (per-array basis)
 - 正規化: グローバル正規化
- **RMA** (Irizarry et al., *Biostatistics*, 2003)
 - 特徴: 読み込んだ複数サンプル(複数アレイ)の情報を用いて前処理を実行 (multi-array basis)
 - 正規化: **quantile正規化** (プローブレベルデータに対して実行)

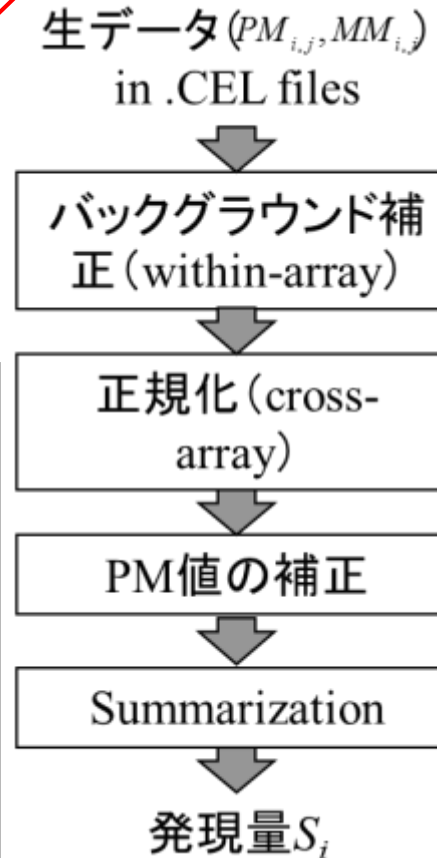


Table 1: Frequency of preprocessing algorithms used during 2003 – 2008

	2003	2004	2005	2006	2007	2008
MAS (2002)	8	34	53	42	47	16
RMA (2003)		8	15	29	20	9
MBEI (2001)	0	3	7	16	8	3
GCRMA (2004)			0	5	8	4
VSN (2002)	0	0	0	4	0	2

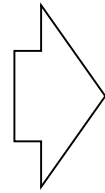
Our investigation was performed for 394 different papers with analyses performed using the Affymetrix HG-U133A array (Gene Expression Omnibus (GEO) ID: GPL96) [32]. These results were

要素技術(グローバル正規化)

- 「各サンプルから測定されたmRNAの全体量は一定」と仮定
 - アレイ上の遺伝子数が少ない場合は非現実的だが、数千~数万種類の遺伝子が搭載されているので妥当

	sample1	sample2
gene1	10.5	12.4
gene2	6.4	7.1
gene3	8.0	8.5
gene4	10.8	11.4
gene5	5.6	6.7
gene6	8.4	8.9
gene7	6.2	7.0
gene8	6.1	6.8
gene9	6.6	6.5
gene10	5.1	5.8
平均値	7.4	8.1

正規化



	sample1	sample2
gene1	14.2	15.3
gene2	8.7	8.8
gene3	10.9	10.5
gene4	14.7	14.1
gene5	7.6	8.3
gene6	11.4	11.0
gene7	8.4	8.6
gene8	8.3	8.4
gene9	9.0	8.0
gene10	6.9	7.2
平均値	10.0	10.0

チップごとに独立して正規化(per-array basis)。他のアレイの影響を受けない。補正後の平均値を10にしたい場合は、sample1の正規化係数 = $10/7.4$ 、sample2の正規化係数 = $10/8.1$ とする。RNA-seqの補正法であるRPM (RPKMの一部)も基本的に同じ考え方。サンプルごとの総カウント数を100万に揃えたいので、正規化係数 = $1,000,000$ /補正前の総リード数としているだけ。尚、総和(sum)と平均(mean)は数学的には等価。

要素技術 (Quantile正規化)

データセット中のサンプル数が変わると結果が変わる (multi-array basis)。他のアレイの影響を受ける。

- 「シグナル強度の順位が同じなら値も同じ」と仮定

正規化前

s1	s2
10.5	12.4
6.4	7.1
8.0	8.5
10.8	11.4
5.6	6.7
8.4	8.9
6.2	7.0
6.1	6.8
6.6	6.5
5.1	5.8

列ごとに
ソート
→

s1	s2
5.1	5.8
5.6	6.5
6.1	6.7
6.2	6.8
6.4	7.0
6.6	7.1
8.0	8.5
8.4	8.9
10.5	11.4
10.8	12.4

行ごとの平
均を算出
→

Ave.
5.5
6.1
6.4
6.5
6.7
6.9
8.3
8.7
11.0
11.6

対応する行の要素の元の位置に
平均値を代入
→

正規化後

s1	s2
11.0	11.6
6.7	6.9
8.3	8.3
11.6	11.0
6.1	6.4
8.7	8.7
6.5	6.7
6.4	6.5
6.9	6.1
5.5	5.5

sample19.txt

要素技術 (Quantile正規化)

データセット中のサンプル数が変わると結果が変わる (multi-array basis)。他のアレイの影響を受ける。

- 「シグナル強度の順位が同じなら値も同じ」と仮定

正規化前

s1	s2	s3
10.5	12.4	9.3
6.4	7.1	13.2
8.0	8.5	10.7
10.8	11.4	7.8
5.6	6.7	5.2
8.4	8.9	9.0
6.2	7.0	6.1
6.1	6.8	7.3
6.6	6.5	7.7
5.1	5.8	3.5

列ごとに
ソート

s1	s2	s3
5.1	5.8	3.5
5.6	6.5	5.2
6.1	6.7	6.1
6.2	6.8	7.3
6.4	7.0	7.7
6.6	7.1	7.8
8.0	8.5	9.0
8.4	8.9	9.3
10.5	11.4	10.7
10.8	12.4	13.2

行ごとの平
均を算出

Ave.
4.8
5.8
6.3
6.8
7.0
7.2
8.5
8.9
10.9
12.1

対応する行の要
素の元の位置に
平均値を代入

正規化後

s1	s2	s3
10.9	12.1	8.9
7.0	7.2	12.1
8.5	8.5	10.9
12.1	10.9	7.2
5.8	6.3	5.8
8.9	8.9	8.5
6.8	7.0	6.3
6.3	6.8	6.8
7.2	5.8	7.0
4.8	4.8	4.8

sample19_plus1.txt

データの正規化

BMC Bioinformatics. 2013 Apr 11;14:124. doi: 10.1186/1471-2105-14-124.

The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis.

Qiu X, Wu H, [RNA](#). 2012 Jun;18(6):1279-88. doi: 10.1261/rna.030916.111. Epub 2012 Apr 24.

Department of
Rochester, New York

Evaluation of normalization methods in mammalian microRNA-Seq data.

Garmire LX, [Subramaniam S](#).

Abstract

BACKGROU

designed to
such as gen
study, we fin
expression a

RESULTS: V
expression a
with fixed sa
extensive sir

CONCLUSIO
not always d
design and v

Department of Bi
92093-0412, US, [Brief Bioinform](#). 2012 Sep 17. [Epub ahead of print]

Abstract

Simple total ta
the next gener
methods on m
used normaliz
Method (TMM)
method. We a
statistical metr
we evaluate th
that Lowess n
applied to the
normalization
microRNA-Seq
the primary fac
normalization

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.

[Dillies MA](#), [Ra](#)
[Estelle J](#), [Gue](#)
on behalf of T

Abstract

During the la
emerged in t
adopted. How
appropriate n
analysis. In t
normalization
varied real ar
data character
practical rec
differential ar

[Algorithms Mol Biol](#). 2012 Apr 5;7(1):5. doi: 10.1186/1748-7188-7-5.

A normalization strategy for comparing tag count data.

[Kadota K](#)¹, [Nishiyama T](#), [Shimizu K](#).

⊕ **Author information**

Abstract

BACKGROUND: High-throughput sequencing, such as ribonucleic acid sequencing (RNA-seq) and chromatin immunoprecipitation sequencing (ChIP-seq) analyses, enables various features of organisms to be compared through tag counts. Recent studies have demonstrated that the normalization step for RNA-seq data is critical for a more accurate subsequent analysis of differential gene expression. Development of a more robust normalization method is desirable for identifying the true difference in tag count data.

RESULTS: We describe a strategy for normalizing tag count data, focusing on RNA-seq. The key concept is to remove data assigned as potential differentially expressed genes (DEGs) before calculating the normalization factor. Several R packages for identifying DEGs are currently available.

前処理法の違いを実感してみよう

- MAS5 (Hubbell et al., *Bioinformatics*, 18: 1585–92, 2002)
 - 特徴: アレイごとに独立して前処理を実行 (per-array basis)
 - 正規化: グローバル正規化
- RMA (Irizarry et al., *Biostatistics*, 4: 249–64, 2003)
 - 特徴: 読み込んだ複数サンプル(複数アレイ)の情報を用いて前処理を実行 (multi-array basis)
 - 正規化: quantile正規化 (プローブレベルデータに対して実行)
- RMX (Kohl et al., *BMC Bioinformatics*, 11: 583, 2010)
 - 教科書中のRobLoxBioCと同じ方法

- [正規化 | Affymetrix GeneChip | について](#) (last modified 2015/05/16) **NEW**
- 正規化 | Affymetrix GeneChip | [frma\(McCall 2010\)](#) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | [rmx\(Kohl 2010\)](#) (last modified 2013/11/19) 推奨
- 正規化 | Affymetrix GeneChip | [GRSN\(Pelz 2008\)](#) (last modified 2013/05/27)
- 正規化 | Affymetrix GeneChip | [Hook\(Binder 2008\)](#) (last modified 2013/05/30)
- 正規化 | Affymetrix GeneChip | [DFW\(Chen 2007\)](#) (last modified 2013/08/20)
- 正規化 | Affymetrix GeneChip | [FARMS\(Hochreiter 2006\)](#) (last modified 2013/08/20)
- 正規化 | Affymetrix GeneChip | [multi-mgMOS\(Liu 2005\)](#) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | [GCRMA\(Wu 2004\)](#) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | [PLIER\(Affymetrix 2004\)](#) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | [VSN\(Huber 2002\)](#) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | [RMA\(Irizarry 2003\)](#) (last modified 2013/08/21)
- 正規化 | Affymetrix GeneChip | [MAS5.0\(Hubbell 2002\)](#) (last modified 2013/11/25)
- 正規化 | Affymetrix GeneChip | [MBEI\(Li 2001\)](#) (last modified 2013/08/21)

正規化 | Affymetrix GeneChip | RMA (Irizarry_2003)

Affymetrix chip (GeneChip™)を用いて得られた*.CELファイルを元に、RMA(Irizarry et al., Biostatistics, 2003)アルゴリズムを用いてSummary scoreを算出。

「ファイル」-「ディレクトリの変更」

1. (CELファイルがあるディレクトリ)

```

out_f <- "hoge1.txt"
#必要なパッケージをロード
library(affy)
#データファイルの読み込み
hoge <- ReadAffy()
#本番
eset <- rma(hoge)
#ファイルに保存
write.exprs(eset, file=out_f)

```

3つのコードの主な違いは、前処理法の違いを表す関数名とパッケージ名部分

正規化 | Affymetrix GeneChip | MAS5.0 (Hubbell_2002)

Affymetrix chip (GeneChip™)を用いて得られた*.CELファイルを元に、MAS5.0 (Hubbell & Bioinformatics, 2002)アルゴリズムを用いてSummary scoreを算出するやり方を示します。低発現領域で

のばらつきが大きいことが指摘をすれば決して悪い方法ではない。レイごとに独立して正規化を行う利点があります。

「ファイル」-「ディレクトリの変更」

1. (CELファイルがあるディレクトリ)

```

out_f <- "hoge1.txt"
#必要なパッケージをロード
library(affy)
#データファイルの読み込み
hoge <- ReadAffy()
#本番
eset <- mas5(hoge)
#対数変換
summary(exprs(eset))
exprs(eset)[exprs(eset) < 1] <- 1
summary(exprs(eset))
exprs(eset) <- log(exprs(eset), 2)
#ファイルに保存
write.exprs(eset, file=out_f)

```

正規化 | Affymetrix GeneChip | rmx (Kohl 2010)

RobLoxBioCというRパッケージ中に実装されているrobsummarization法です。論文中にMAS5の拡張版と書いてありますが、論文に書かれているように、サンプルごとに独立して正規化を前のデータになっているので、robloxbioc関数を用いた代替したものを出力しています。

「ファイル」-「ディレクトリの変更」で適切なディレクトリへ移動

1. (CELファイルがあるディレクトリ上で)手元にあるCELファイルを読み込む

```

out_f <- "hoge1.txt"
#必要なパッケージをロード
library(RobLoxBioC)
#データファイルの読み込み(*.CELファイル)
hoge <- ReadAffy()
#本番
eset <- robloxbioc(hoge)
#対数変換
summary(exprs(eset))
exprs(eset)[exprs(eset) < 1] <- 1
summary(exprs(eset))
exprs(eset) <- log(exprs(eset), 2)
#ファイルに保存
write.exprs(eset, file=out_f)

```

hoge - GSE7623_24samples フォルダ中には、実行後のファイルがある。実際にやるのはGSE7623_02samplesのみ

hoge - GSE7623_24samples フォルダにディレクトリ変更して前処理法を実行。ウェブページは出力ファイル名が同じことに注意

#出力ファイル名を指定してout_fに格納

#パッケージの読み込み

#*.CELファイルの読み込み

#rmxを実行し、結果をesetに保存

#得られたesetの遺伝子発現行列のシグナル強度
#対数変換 (log2) できるようにシグナル強度が
#上記処理後のシグナル強度分布を再び表示させ
#底を2として対数変換

#結果を指定したファイル名で保存

門田のやり方

メモ帳やワードパッドなどのテキストエディタを開いて、出力ファイル名などを適宜変更した一連のコードをファイル (rcode_preprocessing.txt) として保存しています。プログラムの実行時間は7~8分程度。

rcode_preprocessing.txt

```
#####↓
### MAS5 ###↓
#####↓
out_f <- "data_mas.txt" ←
library(affy)
hoge <- ReadAffy()
eset <- mas5(hoge)
summary(exprs(eset))
exprs(eset)[exprs(eset) < 1] <- 1
summary(exprs(eset))
exprs(eset) <- log(exprs(eset), 2)
write.exprs(eset, file=out_f)
↓
#####↓
### RMA ###↓
#####↓
out_f <- "data_rma.txt" ←
library(affy)
hoge <- ReadAffy()
eset <- rma(hoge)
write.exprs(eset, file=out_f)
↓
#####↓
### RMX (RobLoxBioC) ###↓
#####↓
out_f <- "data_rob.txt" ←
library(RobLoxBioC)
hoge <- ReadAffy()
eset <- robloxbioc(hoge)
summary(exprs(eset))
exprs(eset)[exprs(eset) < 1] <- 1
summary(exprs(eset))
exprs(eset) <- log(exprs(eset), 2)
write.exprs(eset, file=out_f)
←
```

#出力ファイル名を指定してout_fに格納↓
#パッケージの読み込み↓
#*.CELファイルの読み込み↓
#MASを実行し、結果をesetに保存↓
#得られたesetの遺伝子発現行列のシグナル強度分布を表
#対数変換 (log2) できるようにシグナル強度が1未満のも
#上記処理後のシグナル強度分布を再び表示させて確認↓
#底を2として対数変換↓
#結果を指定したファイル名で保存↓

#出力ファイル名を指定してout_fに格納↓
#パッケージの読み込み↓
#*.CELファイルの読み込み↓
#RMAを実行し、結果をesetに保存↓
#結果を指定したファイル名で保存↓

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

門田のやり方

R Console画面上でコピー。作業ディレクトリの変更と.CELファイルが2つあることを確認。

```
#####↓
### MAS5 ###↓
#####↓
out_f <- "data_mas.txt"
library(affy)
hoge <- ReadAffy()
eset <- mas5(hoge)
summary(exprs(eset))
exprs(eset)[exprs(eset) < 1] <- 1
summary(exprs(eset))
exprs(eset) <- log(exprs(eset), 2)
write.exprs(eset, file=out_f)
↓
#####↓
### RMA ###↓
#####↓
out_f <- "data_rma.txt"
library(affy)
hoge <- ReadAffy()
eset <- rma(hoge)
write.exprs(eset, file=out_f)
↓
#####↓
### RMX (RobLoxBioC) ###↓
#####↓
out_f <- "data_rob.txt"
library(RobLoxBioC)
hoge <- ReadAffy()
eset <- robloxbioc(hoge)
summary(exprs(eset))
exprs(eset)[exprs(eset) < 1] <- 1
summary(exprs(eset))
exprs(eset) <- log(exprs(eset), 2)
write.exprs(eset, file=out_f)
↓
```

rcode_preprocessing.txt

```
R Console
#
# R version 3.1.3 (2015-03-09) -- "Smooth Sidewalk"
# Copyright (C) 2015 The R Foundation for Statistical Computing
# Platform: x86_64-w64-mingw32/x64 (64-bit)
#
# R は、自由なソフトウェアであり、「完全に無保証」です。
# 一定の条件に従えば、自由にこれを再配布することができます。
# 配布条件の詳細に関しては、'license()' あるいは 'licence()' と$
#
# R は多くの貢献者による共同プロジェクトです。
# 詳しくは 'contributors()' と入力してください。
# また、R や R のパッケージを出版物で引用する際の形式については
# 'citation()' と入力してください。
#
# 'demo()' と入力すればデモをみることができます。
# 'help()' とすればオンラインヘルプが出ます。
# 'help.start()' で HTML ブラウザによるヘルプがみられます。
# 'q()' と入力すれば R を終了します。
#
# > getwd()
# [1] "C:/Users/kadota/Desktop/hoge/GSE7623_02samples"
# > list.files()
# [1] "GSM184414.CEL" "GSM184415.CEL"
# > |
```


途中経過(MAS5)

R Console画面では見られないが、rat2302cdf_2.15.0.zipというファイルを自動でダウンロードしている。

```
#####↓
### MAS5 ###↓
#####↓
out_f <- "data_mas.txt"
library(affy)
hoge <- ReadAffy()
eset <- mas5(hoge)
summary(exprs(eset))
exprs(eset)[exprs(eset) < 1] <- 1
summary(exprs(eset))
exprs(eset) <- log(exprs(eset), 2)
write.exprs(eset, file=out_f)
↓
#####↓
### RMA ###↓
#####↓
out_f <- "data_rma.txt"
library(affy)
hoge <- ReadAffy()
eset <- rma(hoge)
write.exprs(eset, file=out_f)
↓
#####↓
### RMX (RobLoxBioC) ###↓
#####↓
out_f <- "data_rob.txt"
library(RobLoxBioC)
hoge <- ReadAffy()
eset <- robloxbioc(hoge)
summary(exprs(eset))
exprs(eset)[exprs(eset) < 1] <- 1
summary(exprs(eset))
exprs(eset) <- log(exprs(eset), 2)
write.exprs(eset, file=out_f)
<
```

```
code_preprocessing.txt
R Console
> hoge <- ReadAffy()
> eset <- mas5(hoge)
background correction: mas
PM/MM correction : mas
expression values: mas
background correcting... URL 'http://bioconductor.org/packages$
Content type 'application/zip' length 2385529 bytes (2.3 MB)
開かれた URL
downloaded 2.3 MB

The downloaded binary packages are in
C:\Users\kadota\AppData\Local\Temp\Rtmp8MLM6q\download$
done.
31099 ids to be processed
|
|#####|
> summary(exprs(eset))
GSM184414.CEL      GSM184415.CEL
Min.      :    0.13   Min.      :    0.37
1st Qu.:   35.14   1st Qu.:   38.82
Median :  113.97   Median :  131.40
```

#*.CELファイルの読み込み
#MASを実行し、結果をesetに格納

#得られたesetの遺伝子

途中経過(MAS5)

exprs(eset)がMAS5法実行結果の遺伝子発現行列。
summary関数を実行して列ごと(つまりサンプルごと)の要約統計量を表示している。①デフォルト出力は対数変換前のデータなので、②シグナル強度1未満の数値を1に置換して、③確認。④対数変換(底は2)。

```
#####↓  
### MAS5 ###↓  
#####↓  
out_f <- "data_mas.txt"  
library(affy)  
hoge <- ReadAffy()  
eset <- mas5(hoge)  
summary(exprs(eset))  
exprs(eset)[exprs(eset) < 1] <- 1  
summary(exprs(eset))  
exprs(eset) <- log(exprs(eset), 2)  
write.exprs(eset, file=out_f)  
↓  
#####↓  
### RMA ###↓  
#####↓  
out_f <- "data_rma.txt"  
library(affy)  
hoge <- ReadAffy()  
eset <- rma(hoge)  
write.exprs(eset, file=out_f)  
↓  
#####↓  
### RMX (RobLoxBioC) ###↓  
#####↓  
out_f <- "data_rob.txt"  
library(RobLoxBioC)  
hoge <- ReadAffy()  
eset <- robloxbioc(hoge)  
summary(exprs(eset))  
exprs(eset)[exprs(eset) < 1] <- 1  
summary(exprs(eset))  
exprs(eset) <- log(exprs(eset), 2)  
write.exprs(eset, file=out_f)  
↓
```



```
R Console  
> summary(exprs(eset))  
GSM184414.CEL      GSM184415.CEL  
Min.   : 0.13      Min.   : 0.37  
1st Qu.: 35.14     1st Qu.: 38.82  
Median : 113.97    Median : 131.40  
Mean   : 803.82    Mean   : 812.80  
3rd Qu.: 492.37    3rd Qu.: 513.97  
Max.   :51495.93   Max.   :56539.61  
  
> exprs(eset)[exprs(eset) < 1] <- 1  
> summary(exprs(eset))  
GSM184414.CEL      GSM184415.CEL  
Min.   : 1.00      Min.   : 1.00  
1st Qu.: 35.14     1st Qu.: 38.82  
Median : 113.97    Median : 131.40  
Mean   : 803.82    Mean   : 812.80  
3rd Qu.: 492.37    3rd Qu.: 513.97  
Max.   :51495.93   Max.   :56539.61  
  
> exprs(eset) <- log(exprs(eset), 2)  
> write.exprs(eset, file=out_f)  
>  
> #####  
> ### RMA ###  
> #####
```

#得られたesetの遺伝子\$
#対数変換(log2)でき\$
#上記処理後のシグナル\$
#底を2として対数変換
#結果を指定したファイル

途中経過(RMA)

RMAは非常に早く終わります。それも流行った理由かも。。。やたら長いメッセージが延々と続きますが、特にエラーではなさそうなので、門田は気にしていません。

```
#####↓  
### MAS5 ###↓  
#####↓  
out_f <- "data_mas.txt"  
library(affy)  
hoge <- ReadAffy()  
eset <- mas5(hoge)  
summary(exprs(eset))  
exprs(eset)[exprs(eset) < 1] <- 1  
summary(exprs(eset))  
exprs(eset) <- log(exprs(eset), 2)  
write.exprs(eset, file=out_f)
```

```
#####↓  
### RMA ###↓  
#####↓  
out_f <- "data_rma.txt"  
library(affy)  
hoge <- ReadAffy()  
eset <- rma(hoge)  
write.exprs(eset, file=out_f)  
↓  
#####↓  
### RMX (RobLoxBioC) ###↓  
#####↓  
out_f <- "data_rob.txt"  
library(RobLoxBioC)  
hoge <- ReadAffy()  
eset <- robloxbioc(hoge)  
summary(exprs(eset))  
exprs(eset)[exprs(eset) < 1] <- 1  
summary(exprs(eset))  
exprs(eset) <- log(exprs(eset), 2)  
write.exprs(eset, file=out_f)
```

rcode_preprocessing.txt

R Console

```
> #####  
> ### RMA ###  
> #####  
> out_f <- "data_rma.txt"  
> library(affy)  
> hoge <- ReadAffy()  
> eset <- rma(hoge)  
Background correcting  
Normalizing  
Calculating Expression  
> write.exprs(eset, file=out_f)  
>  
> #####  
> ### RMX (RobLoxBioC) ###  
> #####  
> out_f <- "data_rob.txt"  
> library(RobLoxBioC)  
要求されたパッケージ RobLox をロード中です  
要求されたパッケージ distrMod をロード中です  
要求されたパッケージ distr をロード中です  
要求されたパッケージ startupmsg をロード中です  
:startupmsg> Utilities for start-up messages  
:startupmsg> (version 0.9)
```

#出力ファイル名を指定\$
#パッケージの読み込み
#*.CELファイルの読み\$
#RMAを実行し、結果をe\$

#結果を指定したファイ\$

#出力ファイル名を指定\$
#パッケージの読み込み

途中経過(RMX)

RMXも①デフォルト出力は対数変換前のデータなので、②シグナル強度1未満の数値を1に置換して、③確認。④対数変換(底は2)。という処理をしているが、この場合はいきなり④のlog変換をやってもよい。

```
#####↓  
### MAS5 ###↓  
#####↓  
out_f <- "data_mas.txt"  
library(affy)  
hoge <- ReadAffy()  
eset <- mas5(hoge)  
summary(exprs(eset))  
exprs(eset)[exprs(eset) < 1] <- 1  
summary(exprs(eset))  
exprs(eset) <- log(exprs(eset), 2)  
write.exprs(eset, file=out_f)
```

```
#####↓  
### RMA ###↓  
#####↓  
out_f <- "data_rma.txt"  
library(affy)  
hoge <- ReadAffy()  
eset <- rma(hoge)  
write.exprs(eset, file=out_f)
```

```
#####↓  
### RMX (RobLoxBioC) ###↓  
#####↓  
out_f <- "data_rob.txt"  
library(RobLoxBioC)  
hoge <- ReadAffy()  
eset <- robloxbioc(hoge)  
summary(exprs(eset))  
exprs(eset)[exprs(eset) < 1] <- 1  
summary(exprs(eset))  
exprs(eset) <- log(exprs(eset), 2)  
write.exprs(eset, file=out_f)
```

rcode_preprocessing.txt

R Console

```
> eset <- robloxbioc(hoge)  
Background correcting ... done.  
PM/MM correcting ... done.  
Computing expression values ... done.  
> summary(exprs(eset))  
GSM184414.CEL      GSM184415.CEL  
Min.   : 32.04      Min.   : 32.08  
1st Qu.: 41.27      1st Qu.: 41.82  
Median : 60.17      Median : 63.05  
Mean   : 225.76     Mean   : 219.60  
3rd Qu.: 150.84     3rd Qu.: 150.68  
Max.   :12662.67    Max.   :13103.49  
> exprs(eset)[exprs(eset) < 1] <- 1  
> summary(exprs(eset))  
GSM184414.CEL      GSM184415.CEL  
Min.   : 32.04      Min.   : 32.08  
1st Qu.: 41.27      1st Qu.: 41.82  
Median : 60.17      Median : 63.05  
Mean   : 225.76     Mean   : 219.60  
3rd Qu.: 150.84     3rd Qu.: 150.68  
Max.   :12662.67    Max.   :13103.49  
> exprs(eset) <- log(exprs(eset), 2)  
> write.exprs(eset, file=out_f)
```



#得られたesetの遺伝子\$

#対数変換(log2)でき\$
#上記処理後のシグナル\$

#底を2として対数変換
#結果を指定したファイル

うまく実行できれば、list.files()の結果として、3つのファイルができているはずです。

実行結果

```
#####↓  
### MAS5 ###↓  
#####↓  
out_f <- "data_mas.txt"  
library(affy)  
hoge <- ReadAffy()  
eset <- mas5(hoge)  
summary(exprs(eset))  
exprs(eset)[exprs(eset) < 1] <- 1  
summary(exprs(eset))  
exprs(eset) <- log(exprs(eset), 2)  
write.exprs(eset, file=out_f)  
↓  
#####↓  
### RMA ###↓  
#####↓  
out_f <- "data_rma.txt"  
library(affy)  
hoge <- ReadAffy()  
eset <- rma(hoge)  
write.exprs(eset, file=out_f)  
↓  
#####↓  
### RMX (RobLoxBioC) ###↓  
#####↓  
out_f <- "data_rob.txt"  
library(RobLoxBioC)  
hoge <- ReadAffy()  
eset <- robloxbioc(hoge)  
summary(exprs(eset))  
exprs(eset)[exprs(eset) < 1] <- 1  
summary(exprs(eset))  
exprs(eset) <- log(exprs(eset), 2)  
write.exprs(eset, file=out_f)  
↓
```

rcode_preprocessing.txt

```
#出力ファイル名を指定してout_fに格納↓  
#パッケージの読み込み↓  
#*.CELファイルの読み込み↓  
#MASを実行し、結果をesetに保存↓  
#RMAを実行し、結果をesetに保存↓  
#RobLoxBioCを実行し、結果をesetに保存↓  
#  
# R Console  
#  
# Mean : 225.76 Mean : 219.60  
# 3rd Qu.: 150.84 3rd Qu.: 150.68  
# Max. :12662.67 Max. :13103.49  
#  
> exprs(eset)[exprs(eset) < 1] <- 1  
> summary(exprs(eset))  
# GSM184414.CEL GSM184415.CEL  
# Min. : 32.04 Min. : 32.08  
# 1st Qu.: 41.27 1st Qu.: 41.82  
# Median : 60.17 Median : 63.05  
# Mean : 225.76 Mean : 219.60  
# 3rd Qu.: 150.84 3rd Qu.: 150.68  
# Max. :12662.67 Max. :13103.49  
#  
> exprs(eset) <- log(exprs(eset), 2)  
> write.exprs(eset, file=out_f)  
> list.files()  
# [1] "data_mas.txt" "data_rma.txt" "data_rob.txt"  
# [4] "GSM184414.CEL" "GSM184415.CEL"  
#  
> |  
#  
#
```

#対数変換(log2)でき\$
#上記処理後のシグナル\$

#底を2として対数変換
#結果を指定したファイ\$



24サンプルの実行

ディレクトリ変更を正しくできていれば、同じコードを使いませるので便利です。
 ①実行前、②実行後。③dim関数で遺伝子発現行列の行数(=31,099 probesets)と列数(=24 samples)を表示

```
#####↓
### MAS5 ###↓
#####↓
out_f <- "data_mas.txt"
library(affy)
hoge <- ReadAffy
eset <- mas5(ho
summary(exprs(e
exprs(eset)[exp
summary(exprs(e
exprs(eset) <-
write.exprs(ese
↓
#####↓
### RMA ###↓
#####↓
out_f <- "data_
library(affy)
hoge <- ReadAff
eset <- rma(hog
write.exprs(ese
↓
#####↓
### RMX (RobLo
#####↓
out_f <- "data_rob.txt"
library(RobLoxBioC)
hoge <- ReadAffy()
eset <- robloxbioc(hoge)
summary(exprs(eset))
exprs(eset)[exprs(eset) < 1] <- 1
summary(exprs(eset))
exprs(eset) <- log(exprs(eset), 2)
write.exprs(eset, file=out_f)
```

rcode_preprocessing.txt

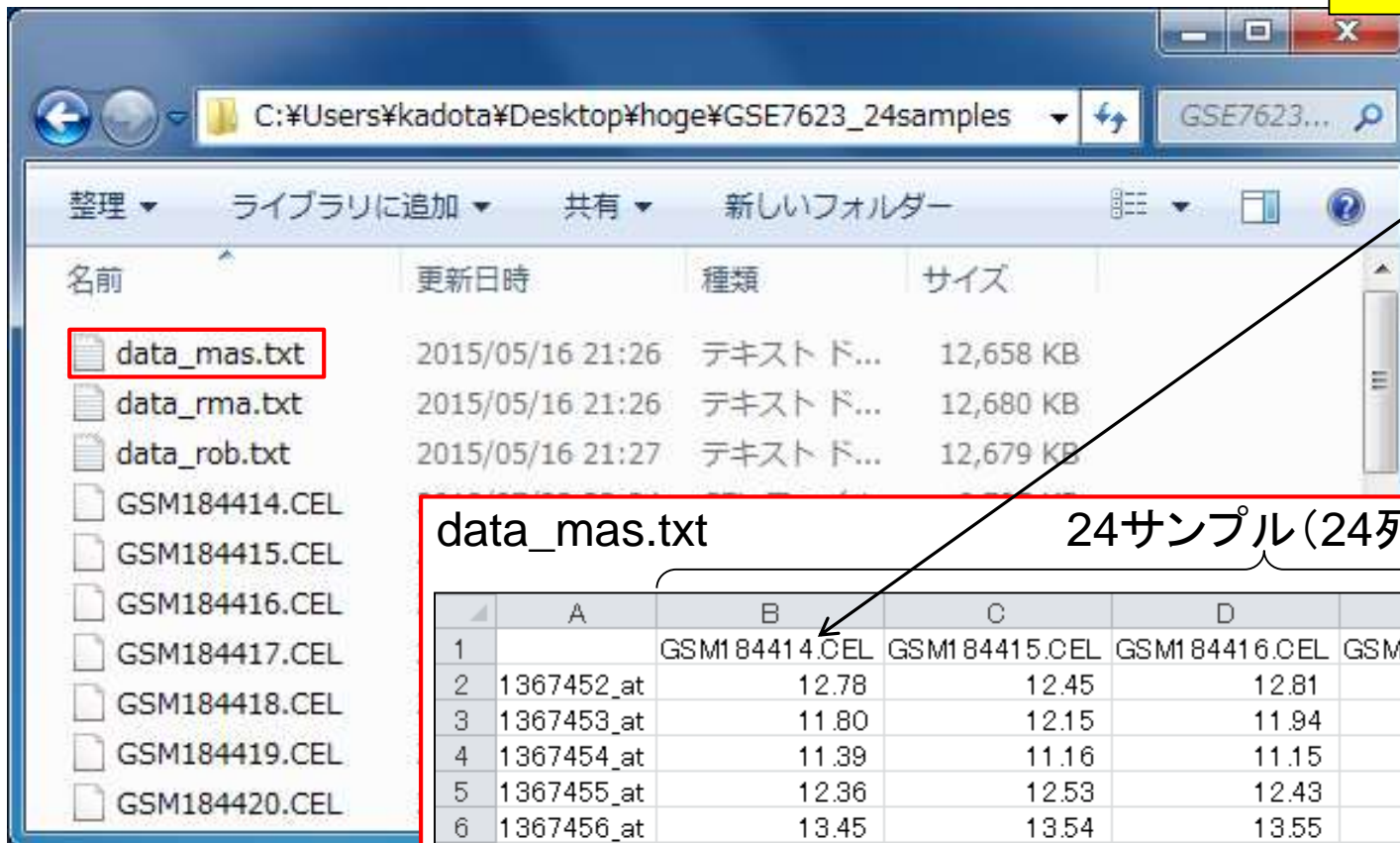
#出力ファイル名を指定してout_fに格納↓
 #パッケージの読み込み↓

```
R Console
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/GSE7623_24samples"
> list.files() ①
[1] "GSM184414.CEL" "GSM184415.CEL" "GSM184416.CEL"
[4] "GSM184417.CEL" "GSM184418.CEL" "GSM184419.CEL"
[7] "GSM184420.CEL" "GSM184421.CEL" "GSM184422.CEL"
[10] "GSM184423.CEL" "GSM184424.CEL" "GSM184425.CEL"
[13] "GSM184426.CEL" "GSM184427.CEL" "GSM184428.CEL"
[16] "GSM184429.CEL" "GSM184430.CEL" "GSM184431.CEL"
[19] "GSM184432.CEL" "GSM184433.CEL" "GSM184434.CEL"
[22] "GSM184435.CEL" "GSM184436.CEL" "GSM184437.CEL"
```

```
R Console
> 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に格納
> list.files() ②
[1] "data_mas.txt" "data_rma.txt" "data_rob.txt"
[4] "GSM184414.CEL" "GSM184415.CEL" "GSM184416.CEL"
[7] "GSM184417.CEL" "GSM184418.CEL" "GSM184419.CEL"
[10] "GSM184420.CEL" "GSM184421.CEL" "GSM184422.CEL"
[13] "GSM184423.CEL" "GSM184424.CEL" "GSM184425.CEL"
[16] "GSM184426.CEL" "GSM184427.CEL" "GSM184428.CEL"
[19] "GSM184429.CEL" "GSM184430.CEL" "GSM184431.CEL"
[22] "GSM184432.CEL" "GSM184433.CEL" "GSM184434.CEL"
[25] "GSM184435.CEL" "GSM184436.CEL" "GSM184437.CEL"
> dim(exprs(eset)) ③
[1] 31099 24
> |
```

24サンプルの実行結果

31,099 probesets × 24 samples
の遺伝子発現行列です。列名は
入力ファイル名と同じ。



data_mas.txt

24サンプル(24列)

	A	B	C	D	E	F	G
1		GSM184414.CEL	GSM184415.CEL	GSM184416.CEL	GSM184417.CEL	GSM184418.CEL	...
2	1367452_at	12.78	12.45	12.81	12.30	12.59	
3	1367453_at	11.80	12.15	11.94	11.97	11.85	
4	1367454_at	11.39	11.16	11.15	11.21	11.54	
5	1367455_at	12.36	12.53	12.43	12.60	12.44	
6	1367456_at	13.45	13.54	13.55	13.63	13.37	
7	1367457_at	10.40	10.70	10.48	10.46	10.14	
8	1367458_at	9.93	10.24	9.97	9.96	8.70	
9	1367459_at	13.83	13.71	13.95	13.70	13.77	
10	1367460_at	13.36	13.55	13.48	13.43	13.54	
11	1367461_at	10.84	11.32	10.98	11.05	10.40	
12	1367462_at	13.47	13.39	13.44	13.43	13.37	
13	1367463_at	14.05	14.06	14.10	13.86	13.79	
14	1367464_at	10.84	11.16	11.09	11.34	11.46	
15	...						

31099 probesets
(31099行)

MAS5法の比較

MAS5はアレイごとに独立して前処理
を実行するので(サンプル数の増減に
かかわらず)同一サンプル間で得られ
る数値情報は不変(per-array basis)。

GSE7623_24samples

24サンプル(24列)

	A	B	C	D	E	F	G
1		GSM184414.CEL	GSM184415.CEL	GSM184416.CEL	GSM184417.CEL	GSM184418.CEL	...
2	1367452_at	12.78	12.45	12.81	12.30	12.59	
3	1367453_at	11.80	12.15	11.94	11.97	11.85	
4	1367454_at	11.39	11.16	11.15	11.21		
5	1367455_at	12.36	12.53	12.43	12.60		
6	1367456_at	13.45	13.54	13.55	13.63		
7	1367457_at	10.40	10.70	10.48	10.46		
8	1367458_at	9.93	10.24	9.97	9.96		
9	1367459_at	13.83	13.71	13.95	13.70		
10	1367460_at	13.36	13.55	13.48	13.43		
11	1367461_at	10.84	11.32	10.98	11.05		
12	1367462_at	13.47	13.39	13.44	13.43		
13	1367463_at	14.05	14.06	14.10	13.86		
14	1367464_at	10.84	11.16	11.09	11.34		
15	...						

GSE7623_02samples 2サンプル(2列)

	A	B	C
1		GSM184414.CEL	GSM184415.CEL
2	1367452_at	12.78	12.45
3	1367453_at	11.80	12.15
4	1367454_at	11.39	11.16
5	1367455_at	12.36	12.53
6	1367456_at	13.45	13.54
7	1367457_at	10.40	10.70
8	1367458_at	9.93	10.24
9	1367459_at	13.83	13.71
10	1367460_at	13.36	13.55
11	1367461_at	10.84	11.32
12	1367462_at	13.47	13.39
13	1367463_at	14.05	14.06
14	1367464_at	10.84	11.16
15	...		

RMA法の比較

RMAは同一サンプル間で得られる数値が異なっていることがわかる。この理由はサンプル間正規化時にquantile normalizationを行っているから(multi-array basis)。

GSE7623_24samples

24サンプル(24列)

	A	B	C	D	E	F	G
1		GSM184414.CEL	GSM184415.CEL	GSM184416.CEL	GSM184417.CEL	GSM184418.CEL	...
2	1367452_at	10.52	10.23	10.35	10.11	10.20	
3	1367453_at	9.66	10.05	9.90	9.82	9.66	
4	1367454_at	9.65	9.40	9.44	9.45		
5	1367455_at	10.76	11.10	10.82	11.03		
6	1367456_at	11.71	11.60	11.59	11.49		
7	1367457_at	8.96	8.77	8.74	8.74		
8	1367458_at	8.28	8.55	8.66	8.43		
9	1367459_at	11.81	11.66	11.70	11.52		
10	1367460_at	11.63	11.62	11.48	11.51		
11	1367461_at	9.38	9.47	9.41	9.33		
12	1367462_at	11.94	11.75	11.89	11.66		
13	1367463_at	12.38	12.08	12.34	12.02		
14	1367464_at	9.48	9.61	9.63	9.58		
15	...						

GSE7623_02samples 2サンプル(2列)

	A	B	C
1		GSM184414.CEL	GSM184415.CEL
2	1367452_at	10.58	10.22
3	1367453_at	9.91	10.17
4	1367454_at	9.68	9.54
5	1367455_at	10.69	10.85
6	1367456_at	11.60	11.51
7	1367457_at	8.89	8.93
8	1367458_at	8.17	8.47
9	1367459_at	11.98	11.73
10	1367460_at	11.60	11.76
11	1367461_at	9.02	9.11
12	1367462_at	11.88	11.68
13	1367463_at	12.41	12.20
14	1367464_at	9.43	9.62
15	...		

課題1: RMX法の比較

GSE7623_24samples

24サンプル(24列)

	A	B	C	D	E	F	...
1		GSM184414.CEL	GSM184415.CEL	GSM184416.CEL	GSM184417.CEL	GSM184418.CEL	...
2	1367452_at	10.90	10.59	10.45	10.46	10.53	
3	1367453_at	9.91	10.14	9.65	9.96	9.82	
4	1367454_at	9.52	9.18	8.89	9.25		
5	1367455_at	10.41	10.49	9.91	10.48		
6	1367456_at	11.47	11.42	11.15	11.39		
7	1367457_at	8.66	8.71	8.25	8.56		
8	1367458_at	7.96	8.21	7.74	7.92		
9	1367459_at	11.67	11.48	11.32	11.50		
10	1367460_at	11.33	11.40	11.01	11.28		
11	1367461_at	9.28	9.34	9.00	9.13		
12	1367462_at	11.44	11.29	11.09	11.37		
13	1367463_at	11.89	11.78	11.65	11.61		
14	1367464_at	8.97	9.07	8.78	9.26		
15	...						

GSE7623_02samples 2サンプル(2列)

	A	B	C
1		GSM184414.CEL	GSM184415.CEL
2	1367452_at	10.90	10.59
3	1367453_at	9.91	10.14
4	1367454_at	9.52	9.18
5	1367455_at	10.41	10.49
6	1367456_at	11.47	11.42
7	1367457_at	8.66	8.71
8	1367458_at	7.96	8.21
9	1367459_at	11.67	11.48
10	1367460_at	11.33	11.40
11	1367461_at	9.28	9.34
12	1367462_at	11.44	11.29
13	1367463_at	11.89	11.78
14	1367464_at	8.97	9.07
15	...		

教科書 § 2-2-2～ § 2-2-4について

■ § 2-2-2 データの正規化(基礎)

- 行列データへのアクセスの基本をおさらい。列名変更。
- summary関数やapply関数。箱ひげ図をpng形式で保存。

■ § 2-2-3 データの正規化(計算例)

- MAS5前処理法を例として、警告メッセージへの対応やサブセットでの実行、プローブごとのシグナル強度の抽出、プローブ配列情報取得(GGRNAと同じような機能)。
- 折れ線グラフの作成手順などを折りませながら、数式の解読が苦手なヒト向けに、重みつき平均の一種であるTukey's biweight estimator計算手順の解説を通じて、重みをつけるという概念の具現化や用いるパラメータの意味合いや感覚を述べている。また、一連の作業を繰り返して、より頑健な値を得るというひらめきやその具体的事例としてRMX (RobLoxBioC)の計算例を示している。本書の醍醐味的部分!

■ § 2-2-4 データの正規化(その他)

- RMAの改良版開発に至る背景(quantile正規化時にサンプル数の増減で結果が変わること)、およびプローブ効果、バッチ効果、トレーニングセット、リファレンス分布の例や基本的な考え方を述べている。また、refRMA, frozen RMA, IRON, frmaTools周辺の比較的最近提唱された方法の特徴についても述べている。

原著論文の引用

Rパッケージやプログラムの多くは原著論文が存在する。各項目の最後のほうにRパッケージとその原著論文のPubMedへのリンクを張ってあります。

イントロ | 発現データ取得 | [ArrayExpress\(Kauffmann 2009\)](#)

マイクロアレイで取得するGEO IDなどの利用「ファイル」

1. Affymetrix法(Trizari)

以下の全データ関数を用意の正

6. AffymetrixデータGSE781 ([Lenburg et al., BMC Cancer, 2003](#))のCELファイルを取得したい場合:

GSE781は2種類のアレイ(GPL96 and GPL97)を使っています。ファイルサイズが大きい(全部で1GB程度?)ので注意してください。

```
param <- "GSE781" #入手したいIDを指定
#必要なパッケージをロード
library(ArrayExpress) #パッケージの読み込み
#前処理(データ取得)
hoge <- getAE(param, type="raw", extract=F
```

- [ArrayExpress: Kauffmann et al., Bioinformatics, 2009](#)
- [affy: Gautier et al., Bioinformatics, 2004](#)

Bioinformatics. 2009 Aug 15;25(16):2092-4. doi: 10.1093/bioinformatics/btp354. Epub 2009 Jun 8.

Importing ArrayExpress datasets into R/Bioconductor.

[Kauffmann A¹](#), [Rayner TE](#), [Parkinson H](#), [Kapushesky M](#), [Lukk M](#), [Brazma A](#), [Huber W](#).

Author information

Abstract

SUMMARY: ArrayExpress is one of the largest public repositories of microarray datasets. R/Bioconductor provides a comprehensive suite of microarray analysis and integrative bioinformatics software. However, easy ways for importing datasets from ArrayExpress into R/Bioconductor have been lacking. Here, we present such a tool that is suitable for both interactive and automated use.

AVAILABILITY: The ArrayExpress package is available from the Bioconductor project at <http://www.bioconductor.org>. A users guide and examples are provided with the package.

PMID: 19505942 [PubMed - indexed for MEDLINE] PMCID: PMC2723004 [Free PMC Article](#)

Contents

- 前処理法の適用(プローブレベルデータ → 発現行列データ)
 - MAS法、RMA法、RMX法(RobLoxBioC)、IRON法(教科書の § 2.2.2~2.2.4)
 - データの正規化(グローバル正規化、quantile正規化)、課題1
 - 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2
 - 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3
- 実験デザイン(教科書の § 3.2.2)

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)

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 (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

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 Biotech Bioinform*

- GSE7623、GPL1355 (ラット24サンプル: Brown adipose tissue (白色脂肪組織))
 - BAT 8サンプル: 3
 - WAT 8サンプル: 3
 - 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 (Iron deficiency diet))

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

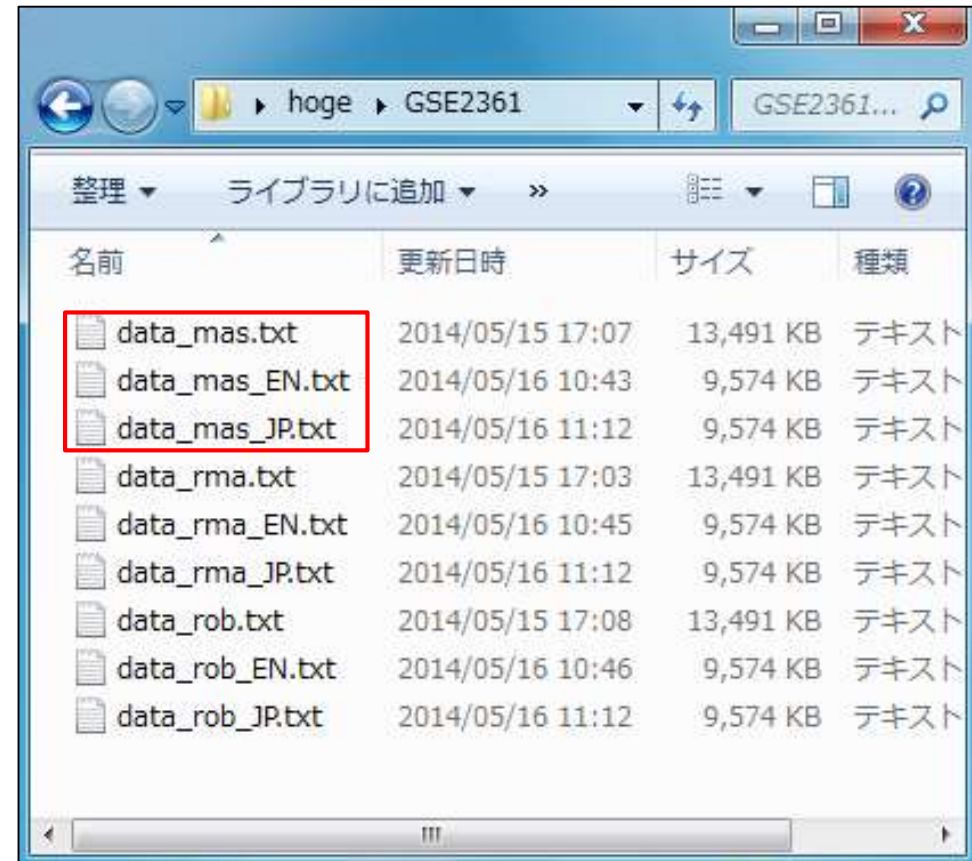
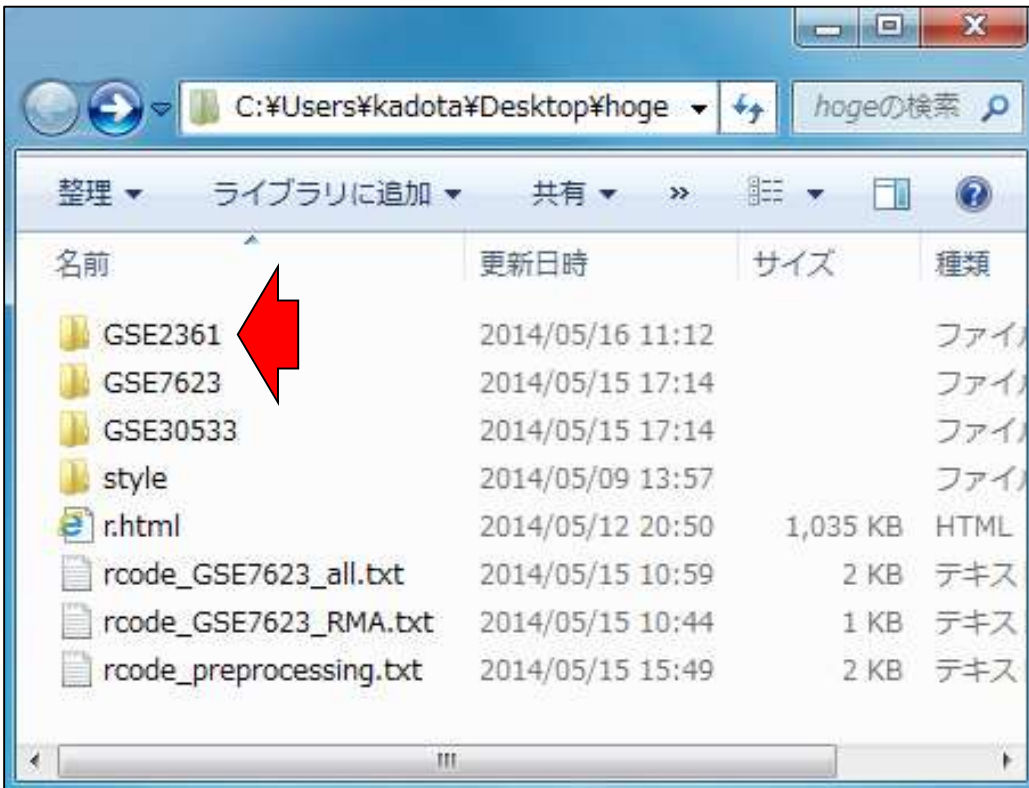
実データ概観

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 (骨格筋)、Pancreas (膵臓)、Prostate (前立腺)、…

GSE2361フォルダの中身。
data_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。



実データ概観

*_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です。

名前	更新日時
data_mas.txt	2014/05/15
data_mas_EN.txt	2014/05/16
data_mas_JP.txt	2014/05/16
data_rma.txt	2014/05/15
data_rma_EN.txt	2014/05/16
data_rma_JP.txt	2014/05/16
data_rob.txt	2014/05/15
data_rob_EN.txt	2014/05/16
data_rob_JP.txt	2014/05/16

data_mas.txt

	A	B	C	D	E	F	G	H	I	J
1		GSM44671	GSM44672	GSM44673	GSM44674	GSM44675	GSM44676	GSM44677	GSM44678	GSM44679
2	1007_s_at	10.82435	11.21261	9.898072	10.8948	11.75531	10.92032	11.61243	11.22101	11.48
3	1053_at	6.621657	6.47488	7.371465	4.168622	7.350244	7.209918	8.362777	7.176571	7.75
4	117_at	8.259603	8.647364	8.689724	7.875649	5.083001	8.165044	7.96707	8.418082	5.719
5	121_at	11.26699	11.04186	11.39999	11.02855	13.13267	11.39138	12.32899	11.0559	11.28
6	1255_g_at	7.1757	6.477278	6.781766	7.048799	7.30767	6.600267	6.42359	6.694412	7.30
7	1294_at	9.137586	9.718507	9.083742	9.014997	8.813377	8.402562	9.146946	8.421351	10.12
8	1316_at	8.225789	7.418895	8.07469	7.280095	8.228176	7.600147	7.422262	7.289894	7.67

data_mas_EN.txt

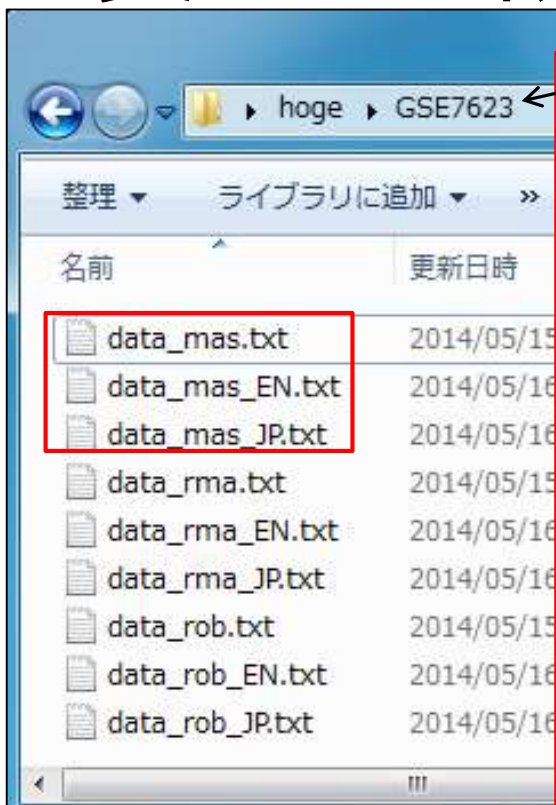
	A	B	C	D	E	F	G	H	I	J
1		Heart	Thymus	Spleen	Ovary	Kidney	Skeletal_Mu	Pancreas	Prostate	Small_I
2	1007_s_at	10.82435	11.21261	9.898072	10.8948	11.75531	10.92032	11.61243	11.22101	11.48
3	1053_at	6.621657	6.47488	7.371465	4.168622	7.350244	7.209918	8.362777	7.176571	7.75
4	117_at	8.259603	8.647364	8.689724	7.875649	5.083001	8.165044	7.96707	8.418082	5.719
5	121_at	11.26699	11.04186	11.39999	11.02855	13.13267	11.39138	12.32899	11.0559	11.28
6	1255_g_at	7.1757	6.477278	6.781766	7.048799	7.30767	6.600267	6.42359	6.694412	7.30
7	1294_at	9.137586	9.718507	9.083742	9.014997	8.813377	8.402562	9.146946	8.421351	10.12
8	1316_at	8.225789	7.418895	8.07469	7.280095	8.228176	7.600147	7.422262	7.289894	7.67

data_mas_JP.txt

	A	B	C	D	E	F	G	H	I	J
1		心臓	胸腺	脾臓	卵巣	腎臓	骨格筋	膵臓	前立腺	小腸
2	1007_s_at	10.82435	11.21261	9.898072	10.8948	11.75531	10.92032	11.61243	11.22101	11.48
3	1053_at	6.621657	6.47488	7.371465	4.168622	7.350244	7.209918	8.362777	7.176571	7.75
4	117_at	8.259603	8.647364	8.689724	7.875649	5.083001	8.165044	7.96707	8.418082	5.719
5	121_at	11.26699	11.04186	11.39999	11.02855	13.13267	11.39138	12.32899	11.0559	11.28
6	1255_g_at	7.1757	6.477278	6.781766	7.048799	7.30767	6.600267	6.42359	6.694412	7.30
7	1294_at	9.137586	9.718507	9.083742	9.014997	8.813377	8.402562	9.146946	8.421351	10.12
8	1316_at	8.225789	7.418895	8.07469	7.280095	8.228176	7.600147	7.422262	7.289894	7.67

実データ概観

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



data_mas.txt

	A	B	C	D	E	F	G	H	I
1		GSM184414	GSM184415	GSM184416	GSM184417	GSM184418	GSM184419	GSM184420	GSM184421
2	1367452_at	12.784463	12.447082	12.805908	12.304718	12.589425	12.607532	11.815378	12.439876
3	1367453_at	11.801247	12.152935	11.942227	11.968477	11.845375	11.681727	12.078672	12.048912
4	1367454_at	11.389902	11.160757	11.145987	11.212088	11.540652	11.308877	11.49885	11.402134
5	1367455_at	12.364348	12.529744	12.432574	12.604011	12.441991	12.249935	12.281827	12.190123
6	1367456_at	13.448486	13.543046	13.552794	13.629799	13.36913	13.244278	13.424371	13.329876
7	1367457_at	10.404028	10.69632	10.475078	10.45579	10.141921	10.290666	10.146529	10.261234
8	1367458_at	9.9253387	10.244544	9.9720008	9.9576072	8.702884	9.3578792	9.2134367	9.499876

data_mas_EN.txt

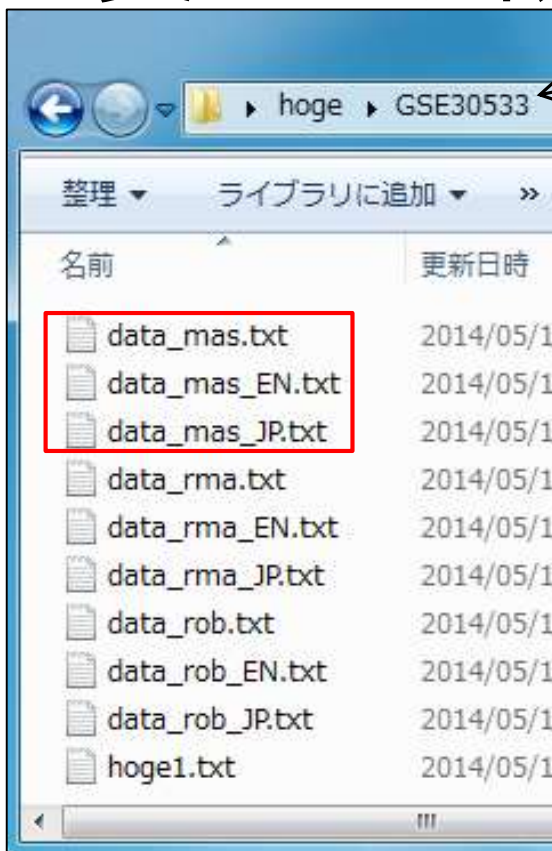
	A	B	C	D	E	F	G	H	I
1		BAT_fed1	BAT_fed2	BAT_fed3	BAT_fed4	BAT_fas1	BAT_fas2	BAT_fas3	BAT_fas4
2	1367452_at	12.784463	12.447082	12.805908	12.304718	12.589425	12.607532	11.815378	12.439876
3	1367453_at	11.801247	12.152935	11.942227	11.968477	11.845375	11.681727	12.078672	12.048912
4	1367454_at	11.389902	11.160757	11.145987	11.212088	11.540652	11.308877	11.49885	11.402134
5	1367455_at	12.364348	12.529744	12.432574	12.604011	12.441991	12.249935	12.281827	12.190123
6	1367456_at	13.448486	13.543046	13.552794	13.629799	13.36913	13.244278	13.424371	13.329876
7	1367457_at	10.404028	10.69632	10.475078	10.45579	10.141921	10.290666	10.146529	10.261234
8	1367458_at	9.9253387	10.244544	9.9720008	9.9576072	8.702884	9.3578792	9.2134367	9.499876

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
2	1367452_at	12.7844634	12.44708219	12.80590758	12.30471769	12.58942538	12.6075319
3	1367453_at	11.80124704	12.15293493	11.94222741	11.96847729	11.84537542	11.6817274
4	1367454_at	11.38990178	11.16075717	11.14598707	11.21208786	11.54065185	11.3088766
5	1367455_at	12.36434768	12.52974368	12.43257392	12.60401124	12.44199125	12.2499348
6	1367456_at	13.44848649	13.54304603	13.55279359	13.62979898	13.36912977	13.2442783
7	1367457_at	10.40402803	10.69631952	10.47507777	10.4557902	10.14192076	10.2906657
8	1367458_at	9.92533749	10.24454359	9.97200015	9.957607169	8.70288404	9.35787919

実データ概観

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



data_mas.txt

	A	B	C	D	E	F	G	H	I
1		GSM757155	GSM757156	GSM757157	GSM757158	GSM757159	GSM757160	GSM757161	GSM757162
2	1367452_at	12.050128	11.923417	11.991814	11.921645	11.728346	12.079674	12.058299	11.979674
3	1367453_at	11.562853	11.587924	11.62175	11.747779	11.776342	11.633797	11.50868	11.47941
4	1367454_at	11.479343	11.676545	11.608875	11.652704	11.86008	11.70589	11.975747	12.00952
5	1367455_at	12.784752	12.58787	12.696588	12.789029	13.002298	12.678645	12.780148	12.55229
6	1367456_at	13.50768	13.533852	13.483843	13.524296	13.452217	13.472369	13.594191	13.603852
7	1367457_at	10.741903	10.136369	10.609696	10.256467	10.312411	10.49749	10.304808	10.42741

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	11.979674
3	1367453_at	11.562853	11.587924	11.62175	11.747779	11.776342	11.633797	11.50868	11.47941
4	1367454_at	11.479343	11.676545	11.608875	11.652704	11.86008	11.70589	11.975747	12.00952
5	1367455_at	12.784752	12.58787	12.696588	12.789029	13.002298	12.678645	12.780148	12.55229
6	1367456_at	13.50768	13.533852	13.483843	13.524296	13.452217	13.472369	13.594191	13.603852
7	1367457_at	10.741903	10.136369	10.609696	10.256467	10.312411	10.49749	10.304808	10.42741

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	11.979674
3	1367453_at	11.562853	11.587924	11.62175	11.747779	11.776342	11.633797	11.50868	11.47941
4	1367454_at	11.479343	11.676545	11.608875	11.652704	11.86008	11.70589	11.975747	12.00952
5	1367455_at	12.784752	12.58787	12.696588	12.789029	13.002298	12.678645	12.780148	12.55229
6	1367456_at	13.50768	13.533852	13.483843	13.524296	13.452217	13.472369	13.594191	13.603852
7	1367457_at	10.741903	10.136369	10.609696	10.256467	10.312411	10.49749	10.304808	10.42741

Contents

- 前処理法の適用(プローブレベルデータ → 発現行列データ)
 - MAS法、RMA法、RMX法(RobLoxBioC)、IRON法(教科書の § 2.2.2~2.2.4)
 - データの正規化(グローバル正規化、quantile正規化)、課題1
 - 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2
 - 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3
- 実験デザイン(教科書の § 3.2.2)

対数変換の有無

§ 3.2.1 クラスタリング(データ変換や距離の定義など)

- 書籍 | トランスクリプトーム解析 | [1.1 はじめに](#) (last modified 2014/05/09) **NEW**
- 書籍 | トランスクリプトーム解析 | [1.2.1 原理\(Affymetrix 3'発現アレイ\)](#) (last modified 2014/05/12) **NEW**
- 書籍 | トランスクリプトーム解析 | [1.2.2 最近の知見](#) (last modified 2014/05/09) **NEW**
- 書籍 | トランスクリプトーム解析 | [2.2.1 生データ\(ブロープレベルデータ\)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**

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
```

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\(基礎\)](#) (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

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。

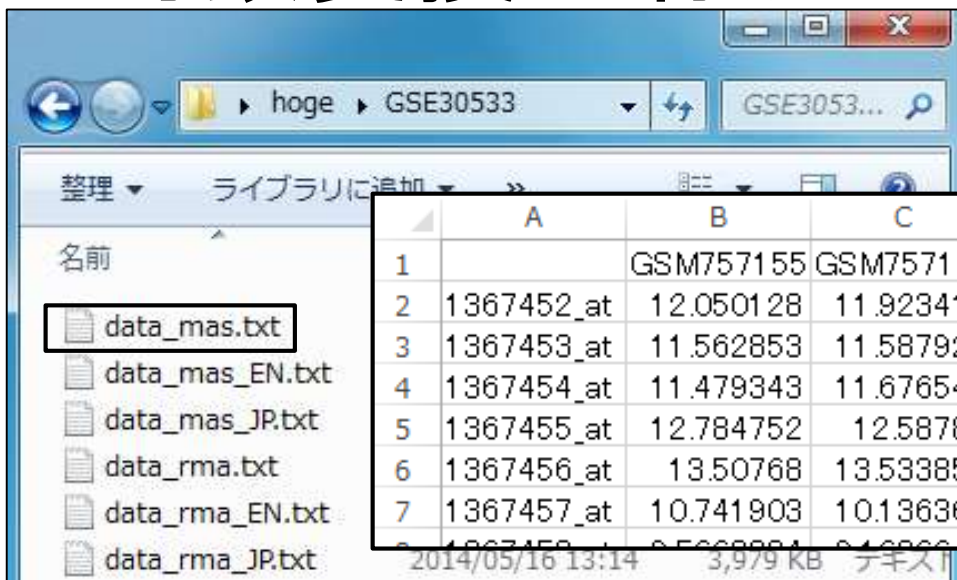
```
out_f <- "hoge1.txt"
library(affy)
hoge <- ReadAffy()
eset <- mas5(hoge)
write.exprs(eset, file=out_f)
```

```
#出力ファイル名を指定してout_fに格納
#パッケージの読み込み
#*.CELファイルの読み込み
#MAS5を実行し、結果をesetに保存
#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	691.0	610.6	443.0	565

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

対数変換の有無



	A	B	C	D	E	F	G	H	I
1		GSM757155	GSM757156	GSM757157	GSM757158	GSM757159	GSM757160	GSM757161	GSM757162
2	1367452_at	12.050128	11.923417	11.991814	11.921645	11.728346	12.079674	12.058299	11.979128
3	1367453_at	11.562853	11.587924	11.62175	11.747779	11.776342	11.633797	11.50868	11.47941
4	1367454_at	11.479343	11.676545	11.608875	11.652704	11.86008	11.70589	11.975747	12.00951
5	1367455_at	12.784752	12.58787	12.696588	12.789029	13.002298	12.678645	12.780148	12.55221
6	1367456_at	13.50768	13.533852	13.483843	13.524296	13.452217	13.472369	13.594191	13.60312
7	1367457_at	10.741903	10.136369	10.609696	10.256467	10.312411	10.49749	10.304808	10.42741

```
#####↓
### 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↓
↓
```

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

対数変換の有無

- 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 アンテーション情報](#) (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | [3.2.1 クラスタリング\(データ変換や距離の定義など\)](#) (last modified 2014/05/16) NEW
- 書籍 | トランスクリプトーム解析 | [3.2.2 実験デザイン, データ分布, 統計解析との関係](#) (last modified 2014/04/18) NEW
- 書籍 | トランスクリプトーム解析 | [3.2.3 多重比較問題](#) (last modified 2014/04/19) NEW

書籍 | トランスクリプトーム解析 | 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p99-100

「ファイル」-「ディレクトリの変更」でデスクトップ上の"E-GEODATA"フォルダに

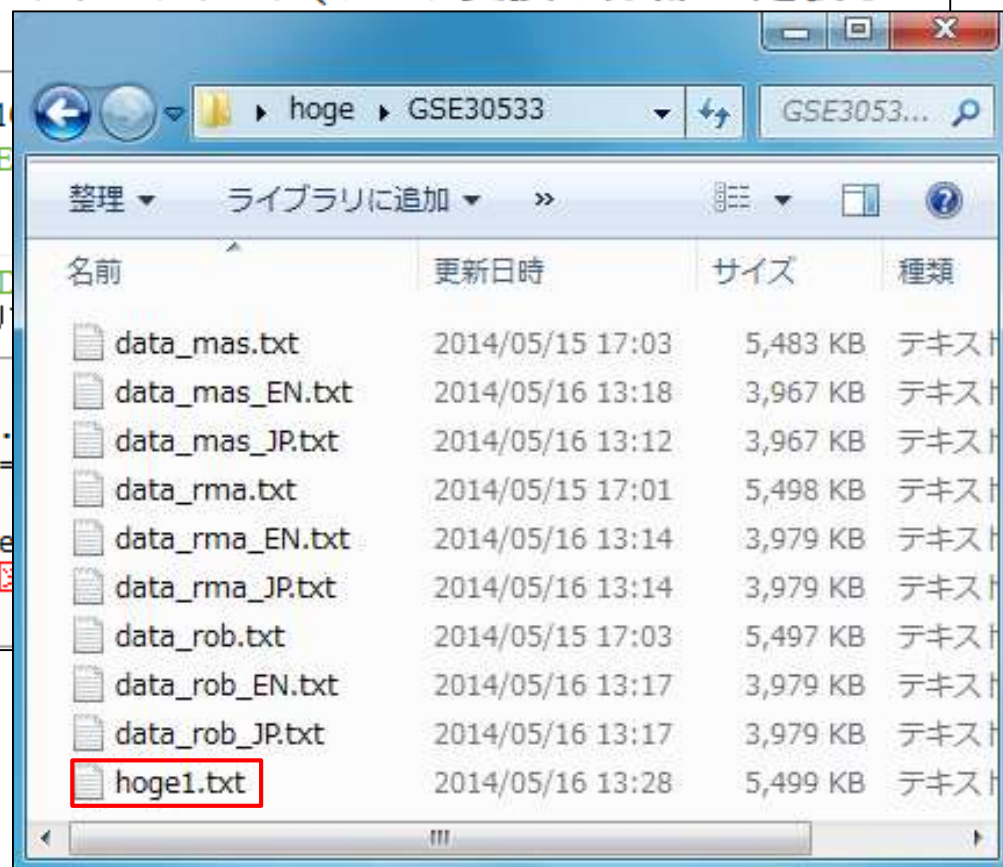
p99の網掛け部分:

書籍中では作業ディレクトリがデスクトップ上の"E-GEODATA"フォルダ内の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.names=1)
colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, sep=""))
data.dist <- as.dist(1 - cor(data, method="spearmanr"))
out <- hclust(data.dist, method = "average")
plot(out) #

```



任意の文字列を含むファイル名のみをリストアップすることもできます。

Tips

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ー。

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)
colnames(data) <- c(paste("G1", 1:10))
data.dist <- as.dist(1 - cor(data))
out <- hclust(data.dist, method="ward.D2")
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"
> |
```

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
8	1367458_at	759.4	575.5	569.6	494.5	691.0	610.6	442.0	565

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

R Console

```

[10] "hogel.txt"
> list.files(pattern="hoge")
[1] "hogel.txt"
> list.files(pattern="EN")
[1] "data_mas_EN.txt" "data_
> in_f <- "hogel.txt"
> data <- read.table(in_f, he
> colnames(data) <- c(paste("
> data.dist <- as.dist(1 - co
> out <- hclust(data.dist, me
> plot(out)
> |
                    
```

R Graphics: Device 2 (ACTIVE)

Cluster Dendrogram

data.dist
hclust(*, "average")

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
> data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
① > 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"
> |
```

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

GSE30533 (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、対数変換前のMAS5データ(hoge1.txt)のクラスタリング結果を比較する。

	A	B	C	D	E	F
1		Iron_def1	Iron_def2	Iron_def3	Iron_def4	Iron_def5
2	1367452_at	12.050128	11.923417	11.991814	11.921645	11.728346
3	1367453_at	11.562853	11.587924	11.62175	11.747779	11.776342
4	1367454_at	11.479343	11.676545	11.6589	11.975747	12.0095
5	1367455_at	12.784752	12.58787	12.8645	12.780148	12.5522
6	1367456_at	13.50768	13.533852	13.2369	13.594191	13.600
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

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



対数変換の有無

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.dist
out <- hclust(data.
plot(out)↓
←
```

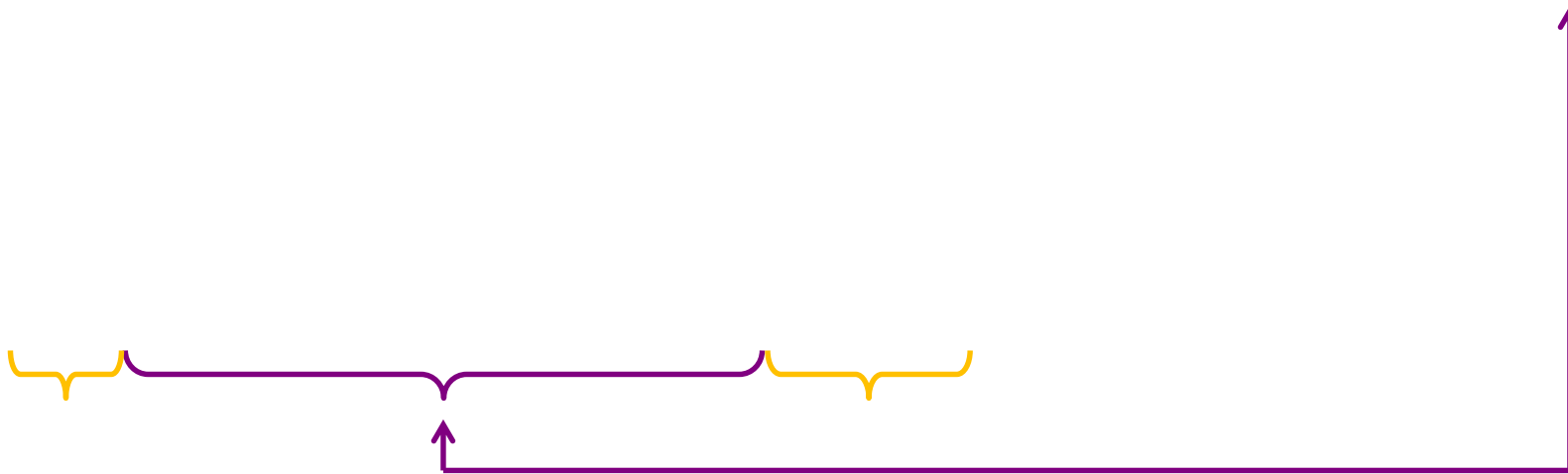
The screenshot shows the R GUI interface. The R Console window contains the following code and output:

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

The R Graphics window displays a Cluster Dendrogram titled "Cluster Dendrogram". The y-axis is labeled "Height" and has tick marks at 0.039 and 0.043. The x-axis labels are Iron_def3, Control4, Control1, Iron_def1, Iron_def5, Iron_def2, Control2, Iron_def4, Control3, and Control5. The dendrogram shows the hierarchical clustering of these samples based on the distance matrix.

他のクラスタリング例

悪性黒色腫(メラノーマ)31サンプルのデータ。悪性度の高い癌のサブタイプを発見。



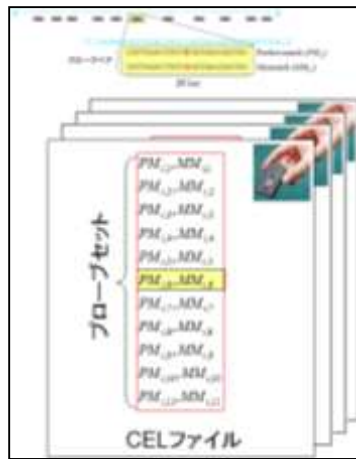
Contents

- 前処理法の適用 (プローブレベルデータ → 発現行列データ)
 - MAS法、RMA法、RMX法 (RobLoxBioC)、IRON法 (教科書の § 2.2.2~2.2.4)
 - データの正規化 (グローバル正規化、quantile正規化)、課題1
 - 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2
 - 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3
- 実験デザイン (教科書の § 3.2.2)

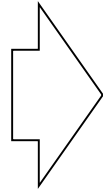
階層的 vs. 非階層的

- 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)

様々な選択肢



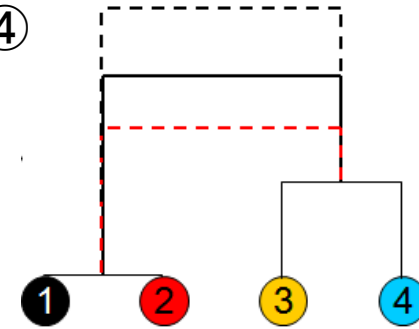
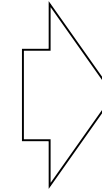
①前処理法



	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	$x_{n,1}$	$x_{n,2}$

遺伝子発現行列②

クラスタリング③と④



①前処理法

- ・MAS
- ・RMA
- ・RobLoxBioC

...

...

×

②スケールング

- ・対数変換
- ・相対値(0~1)
- ・Z-score化

...

...

×

③距離

- ・1-相関係数
- ・ユークリッド
- ・マンハッタン
- ・キャンベラ

...

×

④群の併合

- ・単連結法
- ・完全連結法
- ・平均連結法
- ・ワード法

...

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

□ クラスタをまとめる(併合する)方法

- クラスタ間距離を定義する方法、とほぼ同じ
- 最短距離法、平均連結法、ワード法など

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する。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。

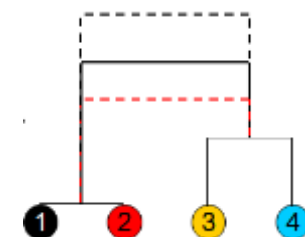
入力例

	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



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 $D(x,y)$

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

- 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 → $r \approx -1$

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

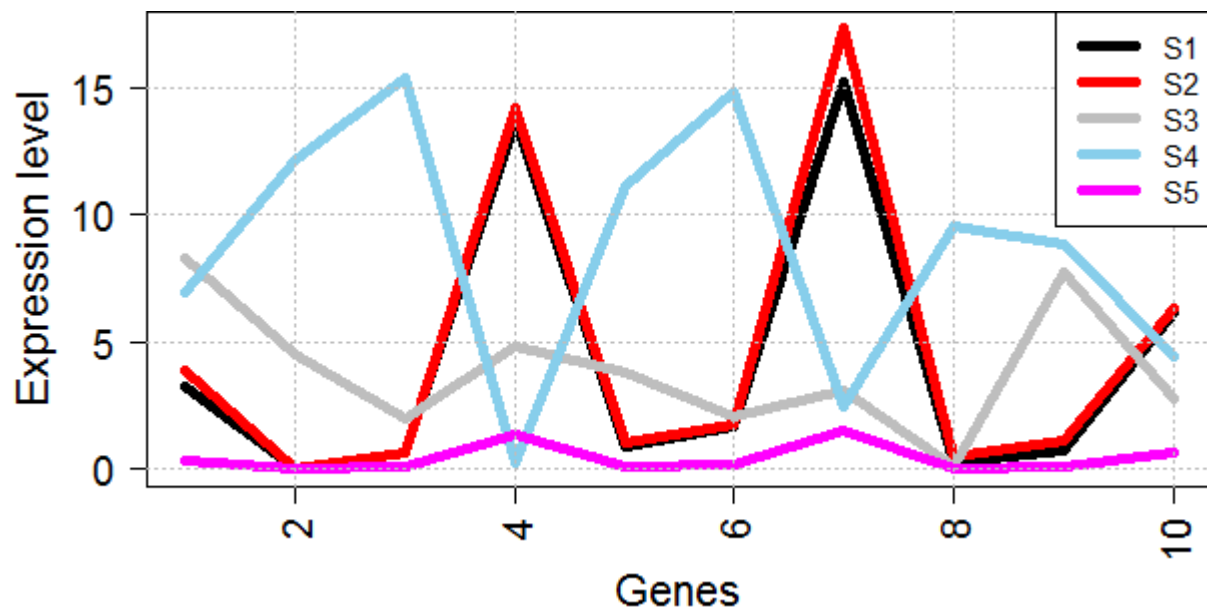
$$\text{距離 } D(x,y) = 1 - r \quad (0 \leq D \leq 2) \quad \begin{cases} r = 1 \rightarrow D = 1 - 1 = 0 \\ r = 0 \rightarrow D = 1 - 0 = 1 \\ r = -1 \rightarrow D = 1 - (-1) = 2 \end{cases}$$

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



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

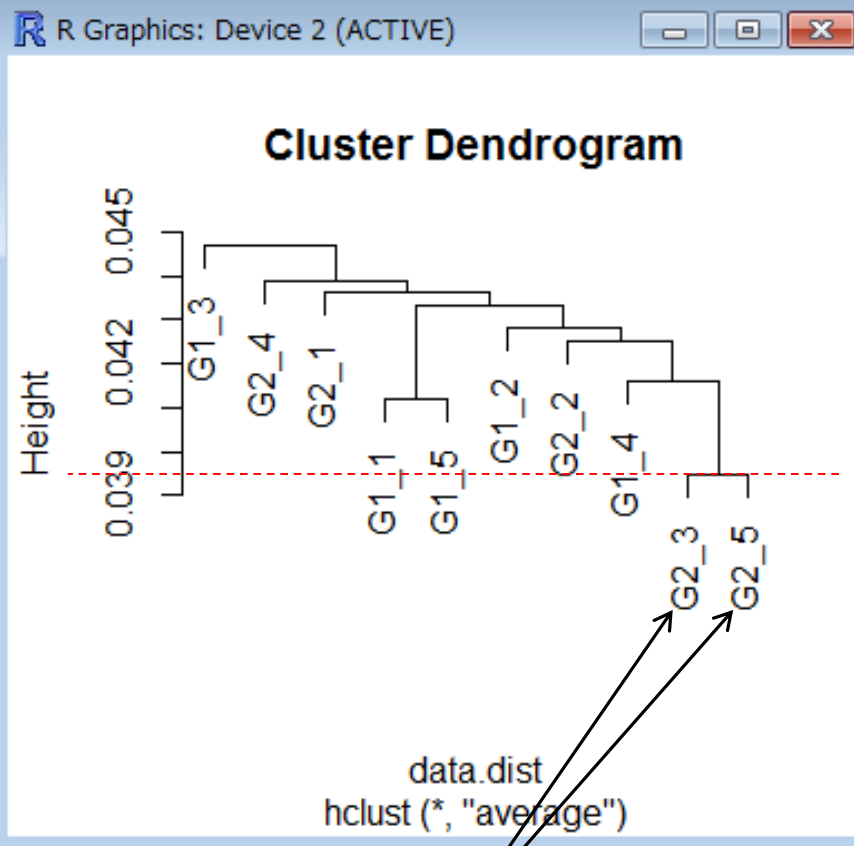
$$\text{相関係数 } r_{S^1S^3} = 0.035 \rightarrow \text{距離 } D_{S^1S^3} = 1 - (0.035) = 0.965$$

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

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
 MAS5データファイル(hoge1.txt)を置いてあるディレクトリであればどこでも構いません

```
in_f <- "hoge1.txt"
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t",
  colnames(data) <- c(paste("G1_", 1:5, sep=""), paste("G2_", 1:5, 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
```



G2_3とG2_5の発
現ベクトル間の距離

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_", 1:5, 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
> data.dist
```

	G1_1	G1_2	G1_3	G1_4	G1_5	G2_1	G2_2	G2_3	G2_4	G2_5
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	

Tips: 相関係数

Spearman相関係数とPearson相関係数の関係。①Spearman相関係数、② $1 - \text{Spearman相関係数}$ 、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

```
R Console  
> cor(data[,8], data[,10], method="spearman")  
[1] 0.9605044  
> 1 - cor(data[,8], data[,10], method="spearman")  
[1] 0.03949559  
>  
> cor(data[,8], data[,10], method="pearson")  
[1] 0.9928171  
> cor(rank(data[,8]), rank(data[,10]), method="pearson")  
[1] 0.9605044  
>  
> cor(data[, "G2_3"], data[, "G2_5"], method="spearman")  
[1] 0.9605044  
> cor(rank(data[,8]), rank(data[,10]), method="spearman")  
[1] 0.9605044  
> |
```

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 $D = \max(|x_1 - y_1|, \dots, |x_i - y_i|, \dots, |x_n - y_n|)$

□ 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である。p100-106。

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

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

計算例 (サンプルxとy間の距離D)

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

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

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

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

サンプルファイルは、右クリックでダウンロード。hogeフォルダ中にあり。

?関数名

(Rで)マイクロアレイデータ解析

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

- 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](#)
- 前処理 | ID変換 | [probe ID -> Entrez ID](#)
- 前処理 | ID変換 | [probe ID -> その他](#)
- 前処理 | ID変換 | [同じ遺伝子名を持つ](#)
- 解析 | 基礎 | [共通遺伝子の抽出](#)
- 解析 | 基礎 | [ベクトル間の距離](#) (last modified 2015/05/16)
- 解析 | 基礎 | [遺伝子ごとの各種要約統計](#)
- 解析 | 基礎 | [最大発現量を示す組織](#)
- 解析 | 基礎 | [似た発現パターンを持つ](#)
- 解析 | 基礎 | [平均-分散プロット](#) (last modified 2015/05/16)
- 解析 | クラスタリング | [階層的](#) | [について](#)

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、以下をコピー

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

```

?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か「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コピー

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, sep="\t")

#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
```

```
R Console

> dist(t(data), method="euclidean") #ユークリッド (Euclidean) 距離
      sample1
sample2 2.792848

> dist(t(data), method="manhattan") #マンハッタン (Manhattan) 距離
      sample1
sample2    7.6

> dist(t(data), method="maximum") #チェビシェフ (Chebyshev) 距離
      sample1
sample2    1.9

> dist(t(data), method="canberra") #キャンベラ (Canberra) 距離
      sample1
sample2 0.4972074

> 1 - cor(data, method="pearson") #1 - Pearson相関係数
      sample1 sample2
sample1 0.0000000 0.02414407
sample2 0.02414407 0.00000000

> dist(t(data), method="binary") #ハミング (Hamming) 距離
      sample1
sample2    0

> dist(t(data), method="minkowski") #ミンコフスキー (Minkowski) 距離
      sample1
sample2 2.792848

> 1 - cor(data, method="spearman") #1 - Spearman相関係数
      sample1 sample2
sample1 0.0000000 0.1333333
sample2 0.1333333 0.0000000

> |
```

```
R Console
> ?dist
starting httpd help server ... done
> dist(t(data), method="euclidean")
      sample1
sample2 2.792848
> dist(t(data))
      sample1
sample2 2.792848
> |
```

参考

①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。②"binary"や"minkowski"というものも指定できるようだが「1-相関係数」を指定することはできないようだ...orz

Description

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

Usage



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

as.dist(m, diag = FALSE, upper = 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.

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?!(ワード法も?!)...。いろいろ試して総合的に判断することが重要

クラスター間の距離の定義

- 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

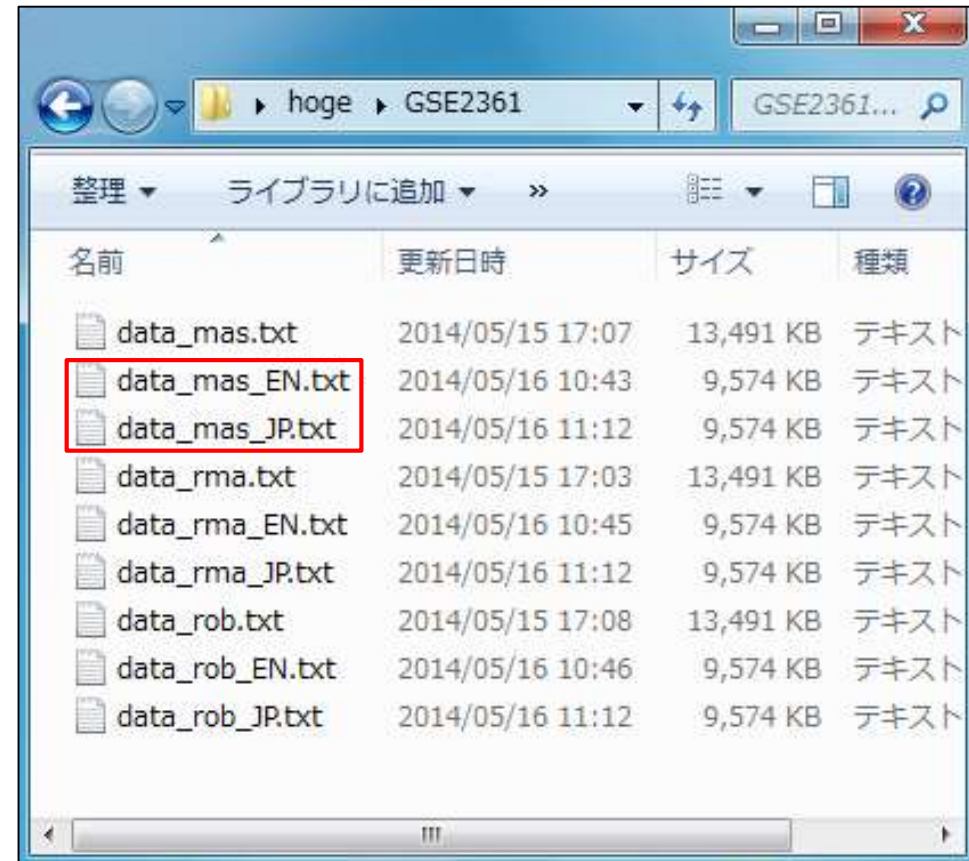
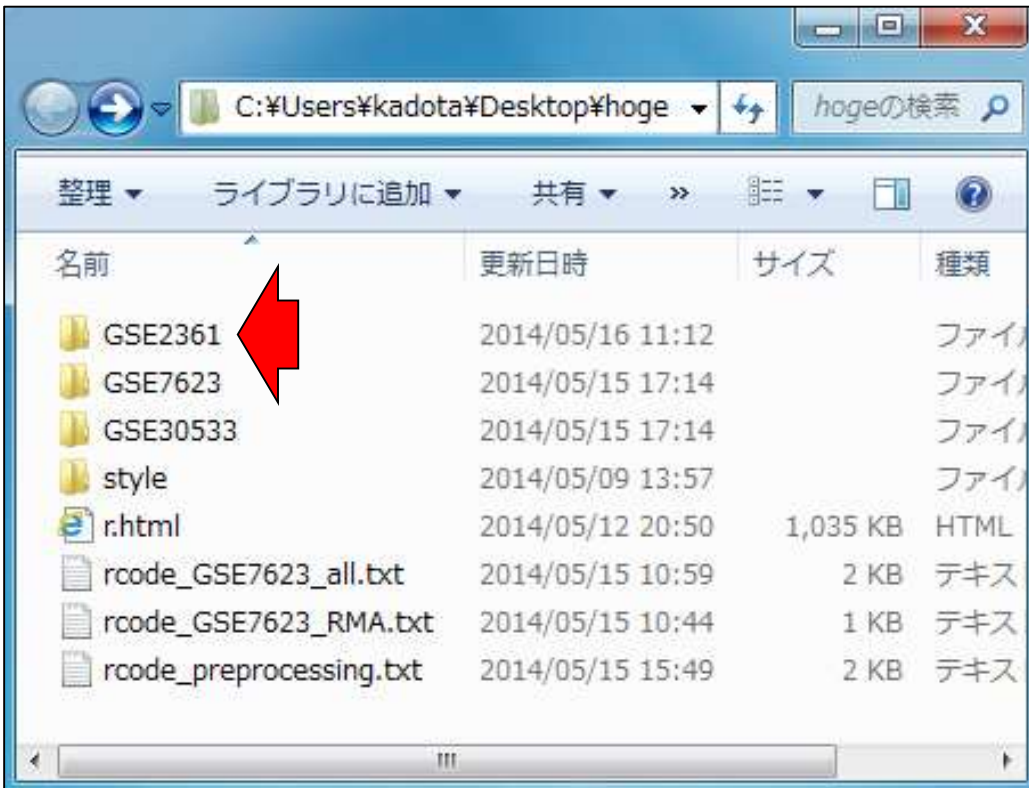
- 前処理法の適用(プローブレベルデータ → 発現行列データ)
 - MAS法、RMA法、RMX法(RobLoxBioC)、IRON法(教科書の § 2.2.2~2.2.4)
 - データの正規化(グローバル正規化、quantile正規化)、**課題1**
 - 実データ概観: GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の § 3.2.1)
 - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
 - 階層的 vs. 非階層的、様々な選択肢
 - 距離の定義: ベクトル間、クラスター間
 - 実データで実行: GSE2361 (ヒト)、**課題2**
 - 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3**
- 実験デザイン(教科書の § 3.2.2)

hoge - GSE2361フォルダ中のMAS5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

GSE2361 (ヒト)

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)

ウェブのテンプレートとの違いは、入力ファイル名とSpearman相関係数の部分のみ

(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 | [hclust](#) (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後で詳細な解析](#) (last modified 2009/8/7)

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.スケール方法(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(平均連結法やワード法など)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、以下をコピー

1. サンプルデータ3の sample3.txt の場合:

サンプル間クラスタリング(距離: 1-Pearson相関係数、方法: 平均連結法(average))でR Graphics画面上に表示するやり方です。

```

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

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

#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="average")
plot(out)

```

```

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

```

rcode_clustering.txt

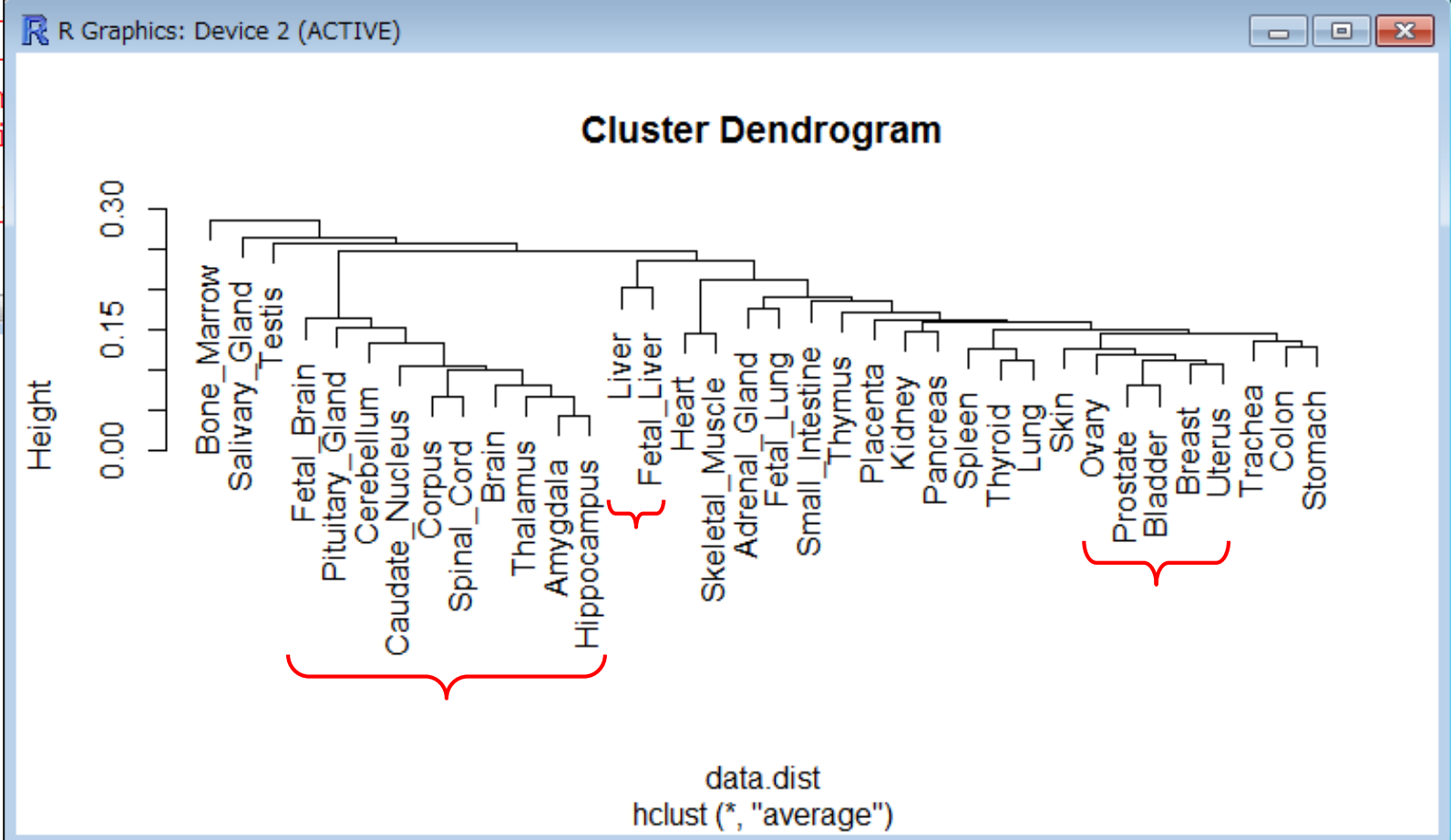
GSE2361 (ヒト)

rcode_clustering.txt

```
#####↓
### MAS5データのクラスタリング ###↓
#####↓
in_f <- "data.mas.EN.txt"↓
```

```
data <- readR
#colnames(da
data.dist <-
out <- hclus
plot(out)↓
<
```

```
R R Console
> getwd()
[1] "C:/Users/kadota/Desktop/hoge/GSE2361"
```



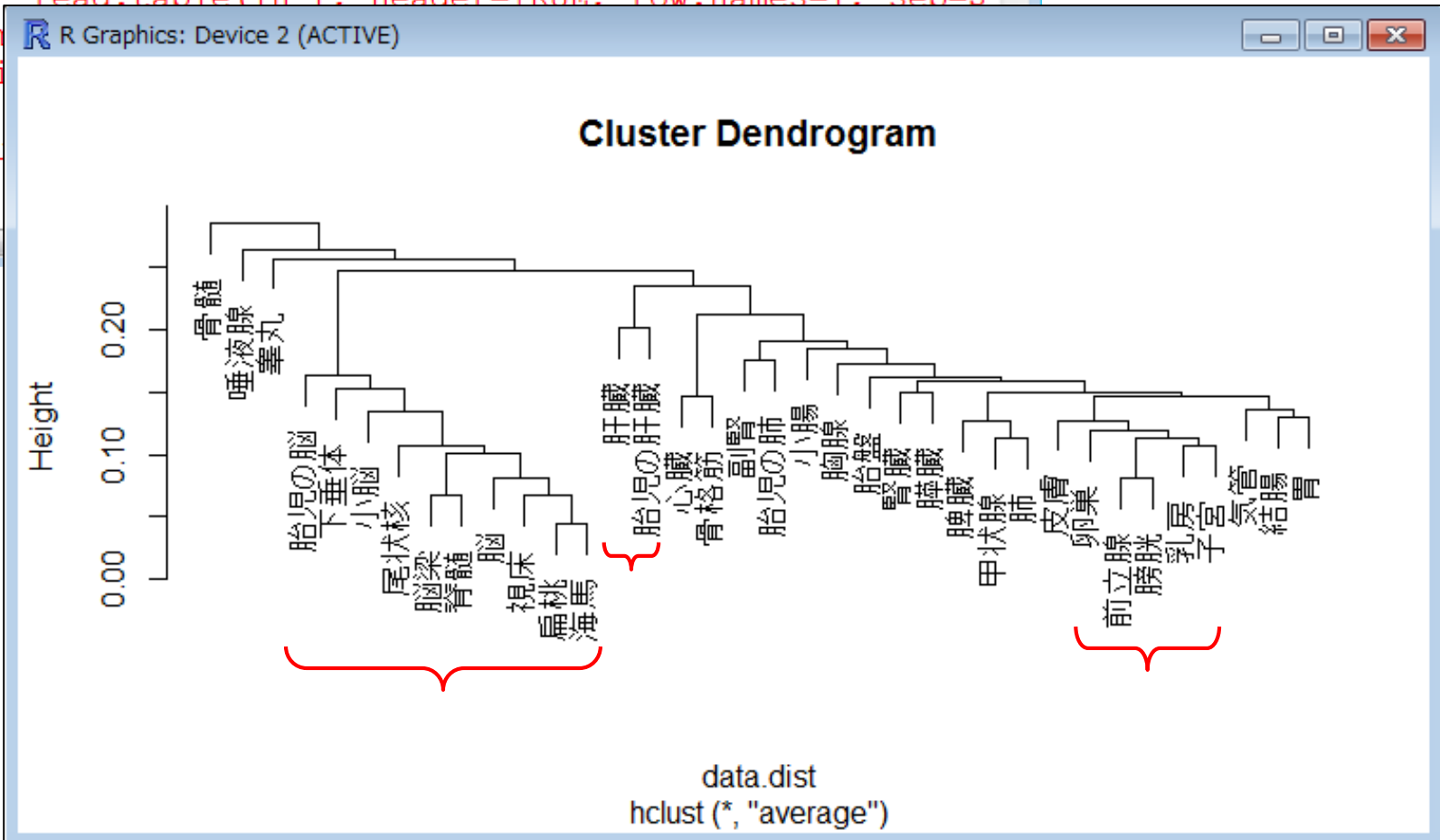
GSE2361 (ヒト)

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
> |
```



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

Tips (ファイル保存)

解析 | クラスタリング | 階層的 | hclust

3. サンプルデータ30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()
```

rancode_clustering_png.txt

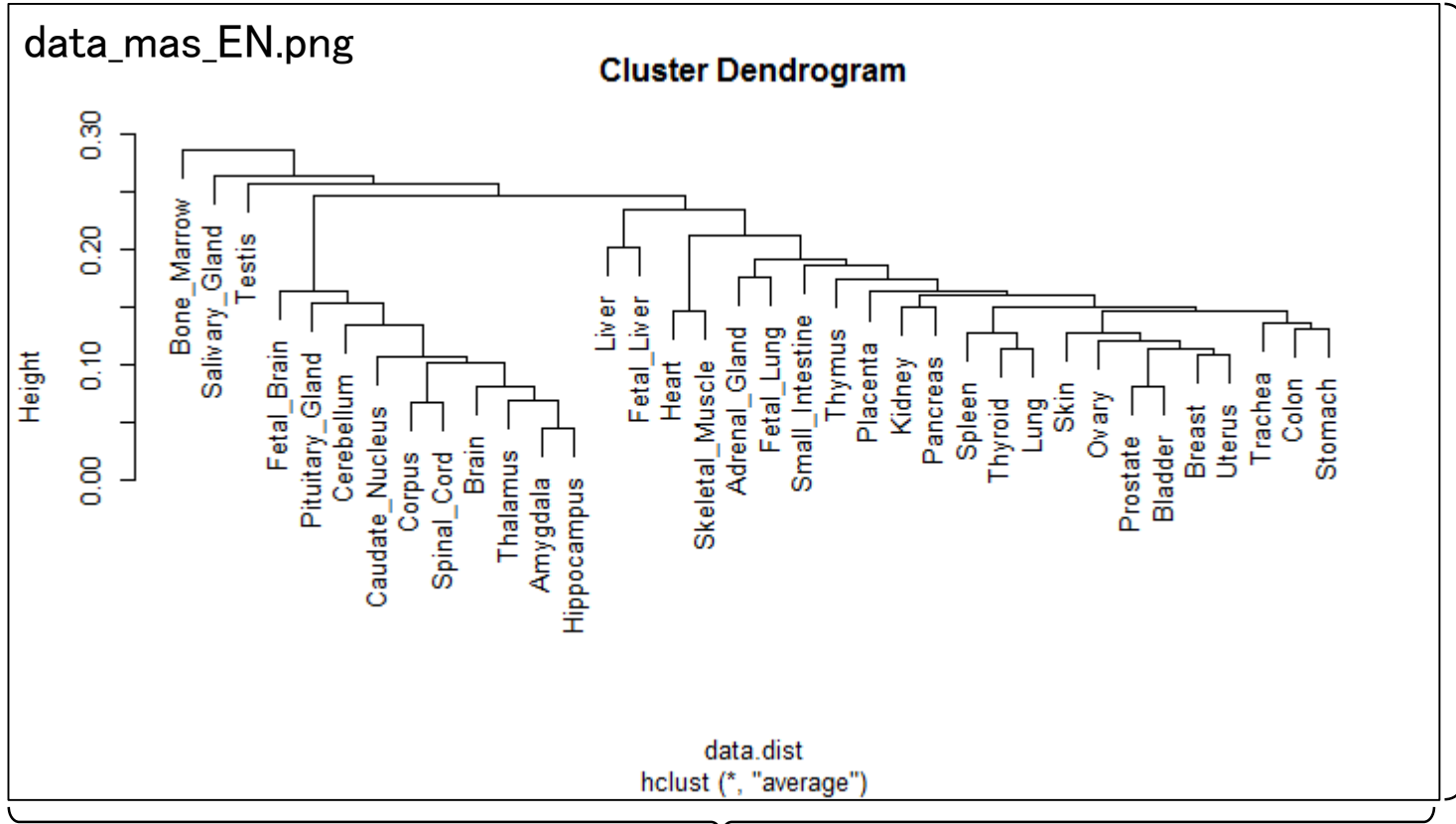
```
##### ↓
### MAS5データのクラスタリング, 距離: 1-Spearman相関係数, 方法: 平均連結法 ### ↓
##### ↓
→ in_f <- "data_mas_EN.txt" #入力ファイル名を指定してin_fに格納 ↓
→ out_f <- "data_mas_EN.png" #出力ファイル名を指定してout_fに格納 ↓
param <- "average" #方法(method)を指定 ↓
→ param_fig <- c(720, 400) #ファイル出力時の横幅と縦幅を指定(単位はピクセル) ↓
↓
#入力ファイルの読み込み ↓
data <- read.table(in_f, 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 ↓
```

常に同じ大ききさで出力されるので便利です

Tips (ファイル保存)

rancode_clustering_png.txt

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



720

400

課題2

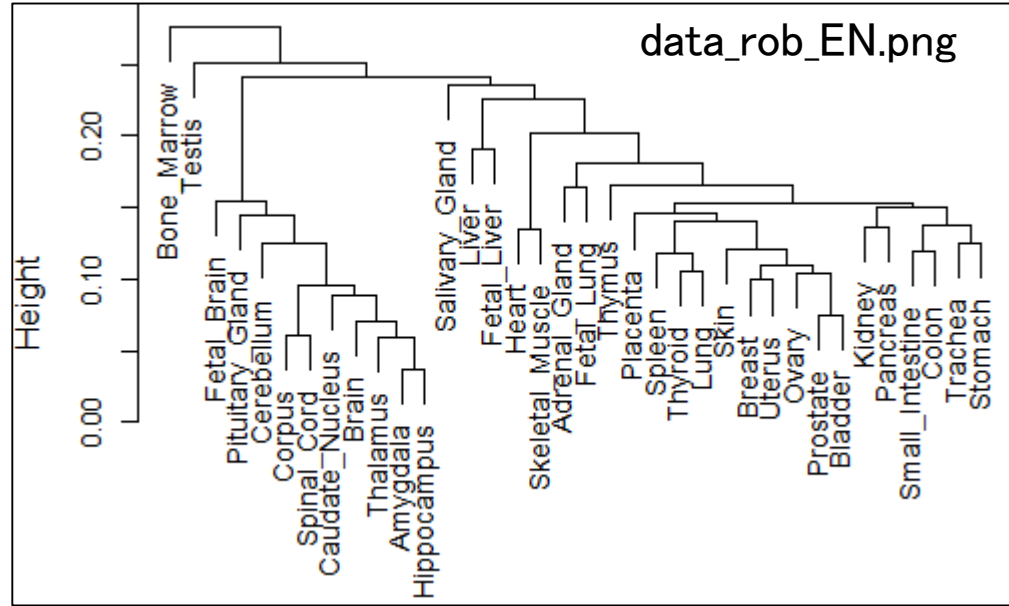
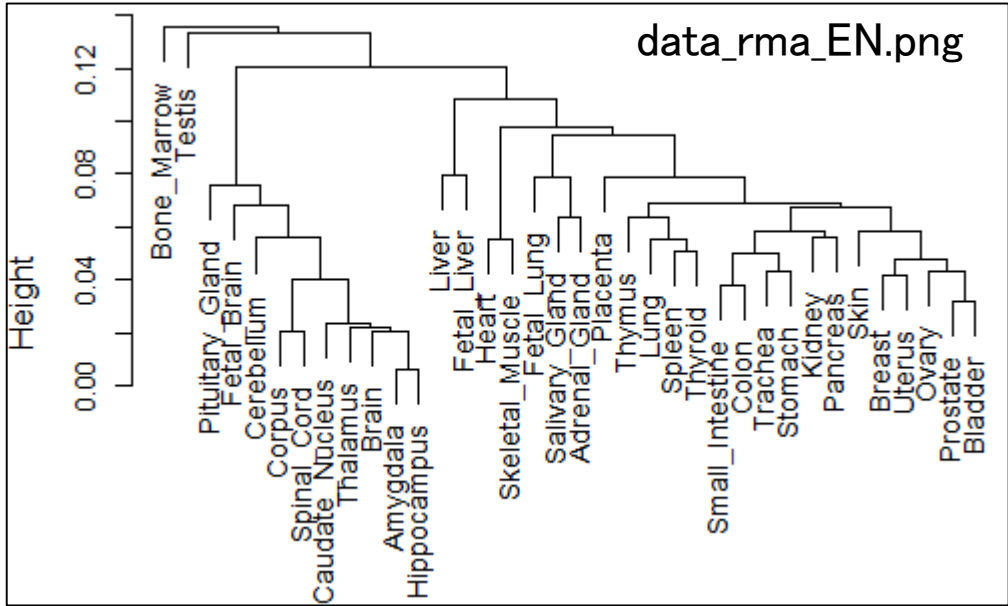
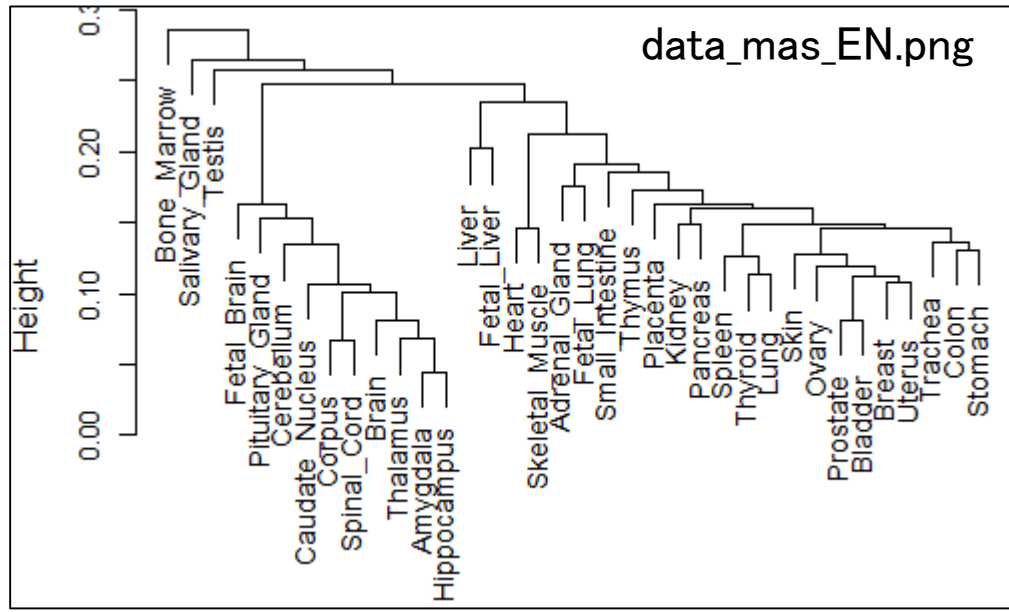
GSE2361のサンプル間クラスタリングをRMA, およびRMX前処理法を適用したデータについても行い、結果を考察せよ。距離の定義はデフォルトのままでよい。

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↓
```

課題2

このような結果が得られるはずですが、この図は余白指定や文字の大きさなどをいろいろ変えています。



発展課題

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。param1 (やparam2)などをいろいろいじって結果を眺めてみてください。

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(平均連結法やワード法など)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、以下をコピー

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の結果をpng形式ファイルで保存するやり方

```
in_f <- "sample3.txt"
param <- "average"
#入力ファイル名を指定してin_fに格納
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
#本番
data.dist <- dist(t(data), method=param)
out <- hclust(data.dist, method="average")
plot(out)
```

<

<

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。

```
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で指定したファイルを読み込み

#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
```

<

>

Contents

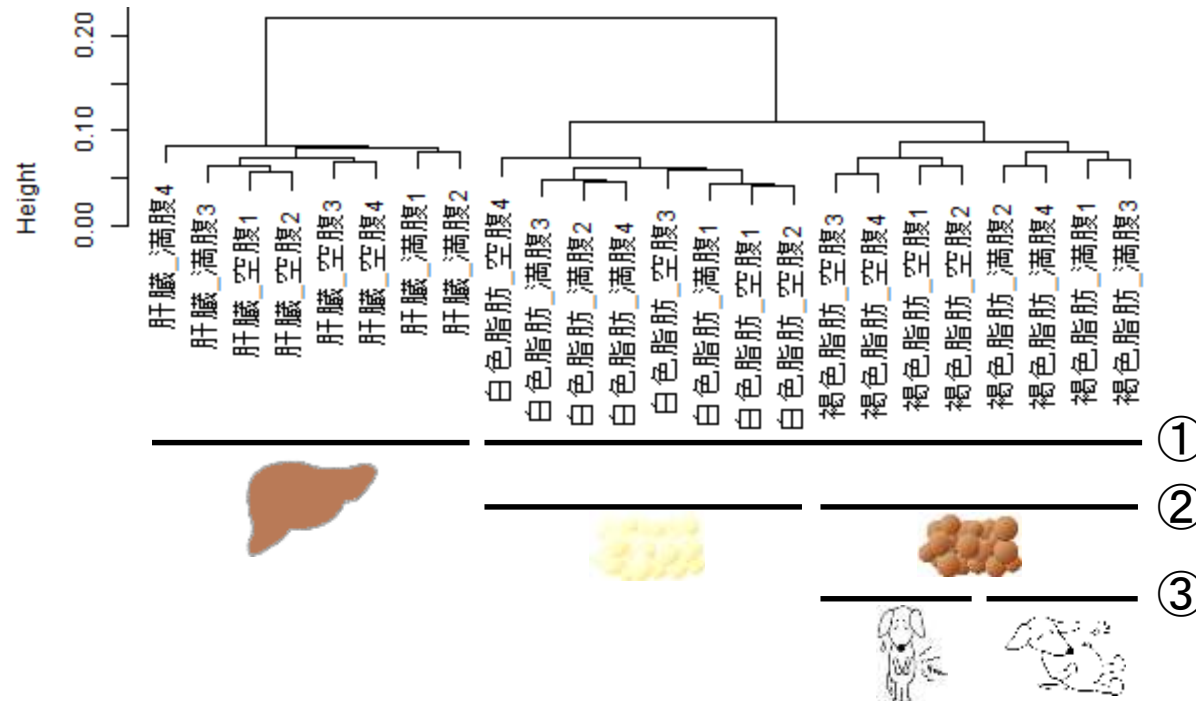
- 前処理法の適用(プローブレベルデータ → 発現行列データ)
 - MAS法、RMA法、RMX法(RobLoxBioC)、IRON法(教科書の § 2.2.2~2.2.4)
 - データの正規化(グローバル正規化、quantile正規化)、課題1
 - 実データ概観: GSE2361 (ヒト)、GSE7623 (ラット)、GSE30533 (ラット)
- クラスタリング(教科書の § 3.2.1)
 - 対数変換の有無(Spearman相関係数を使う場合は気にしなくてよい)
 - 階層的 vs. 非階層的、様々な選択肢
 - 距離の定義: ベクトル間、クラスター間
 - 実データで実行: GSE2361 (ヒト)、課題2
 - 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3
- 実験デザイン(教科書の § 3.2.2)

GSE7623 (ラット)

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

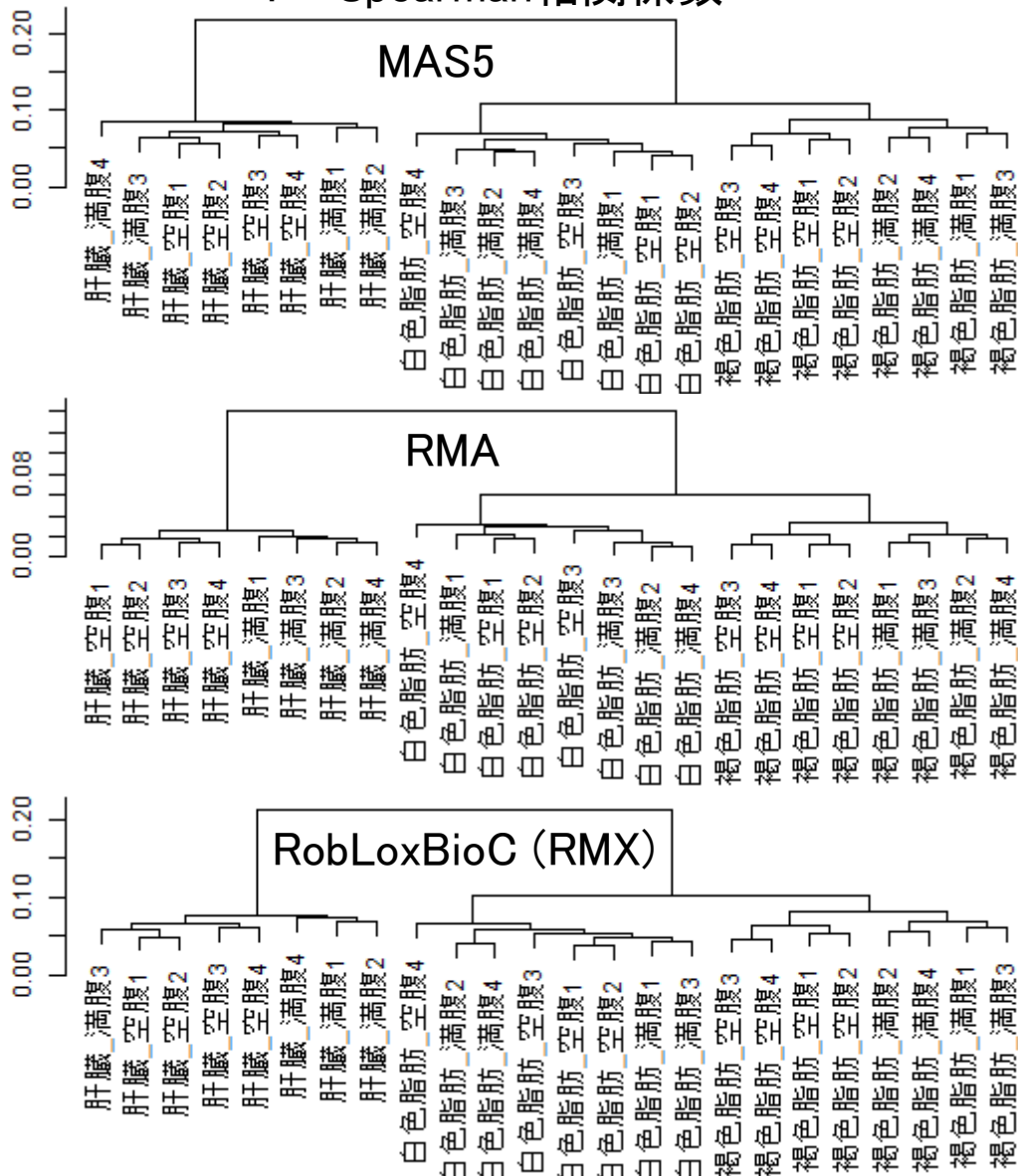
①肝臓と脂肪間で大きく2つのクラスターに分かれている。②脂肪の中でも白色脂肪と褐色脂肪に分かれている。③褐色脂肪は空腹(24時間絶食)と満腹(通常)できれいに分かれている。

- 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

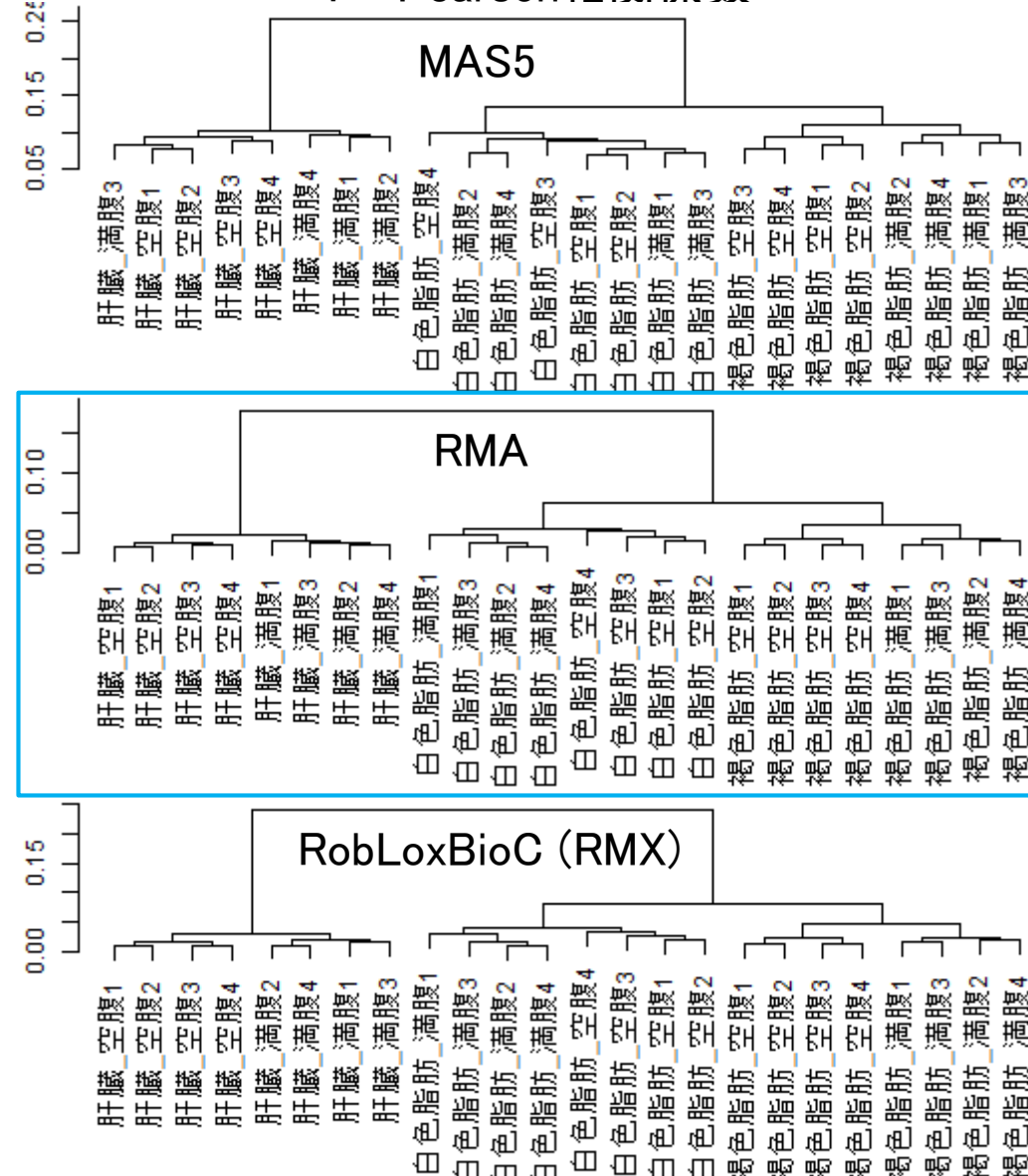


GSE7623 (ラット)

1 - Spearman相関係数



1 - Pearson相関係数



GSE7623 (ラット)

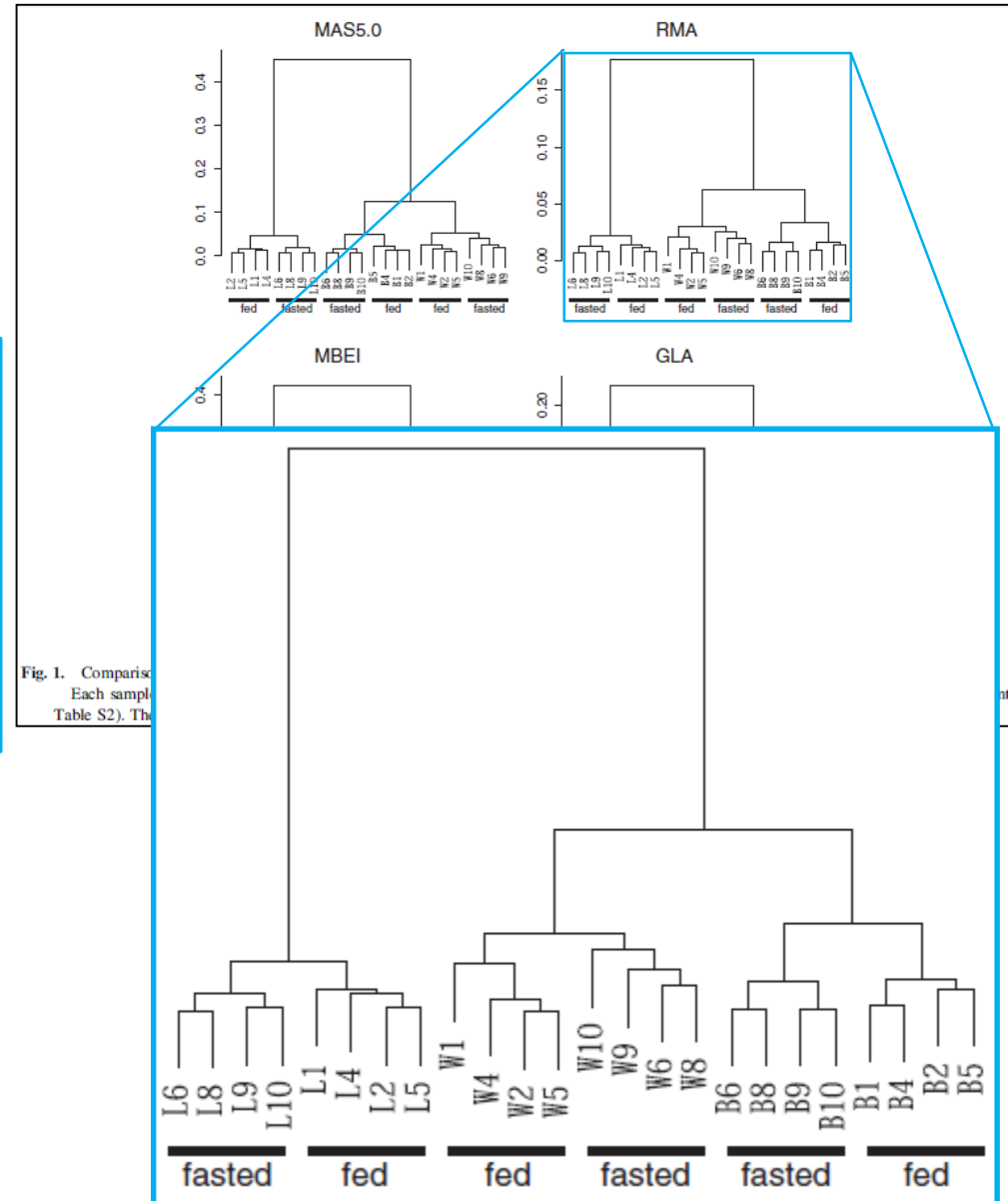
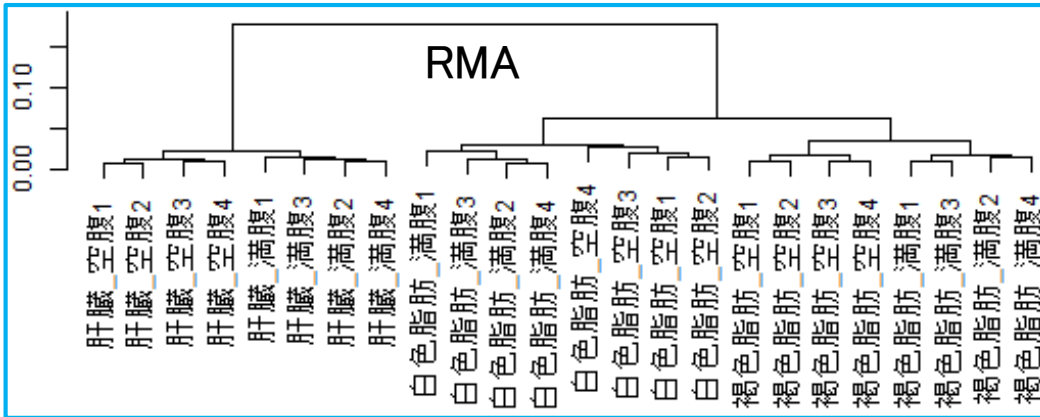


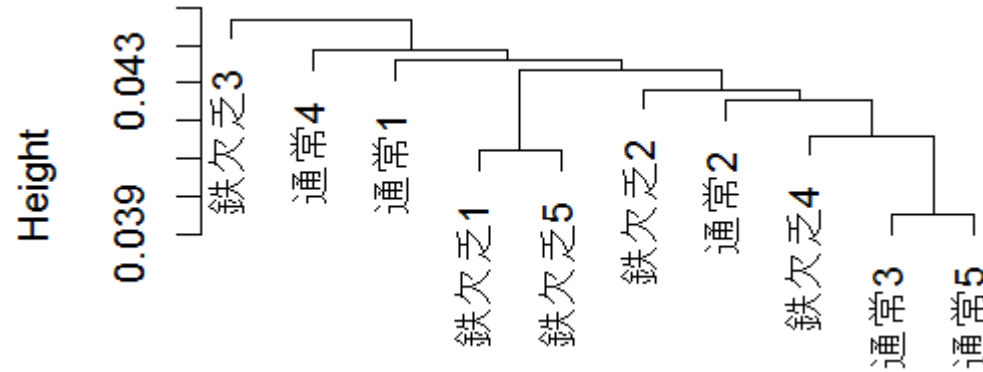
Fig. 1. Comparison of clustering methods. Each sample is labeled as in Table S2. The

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…。

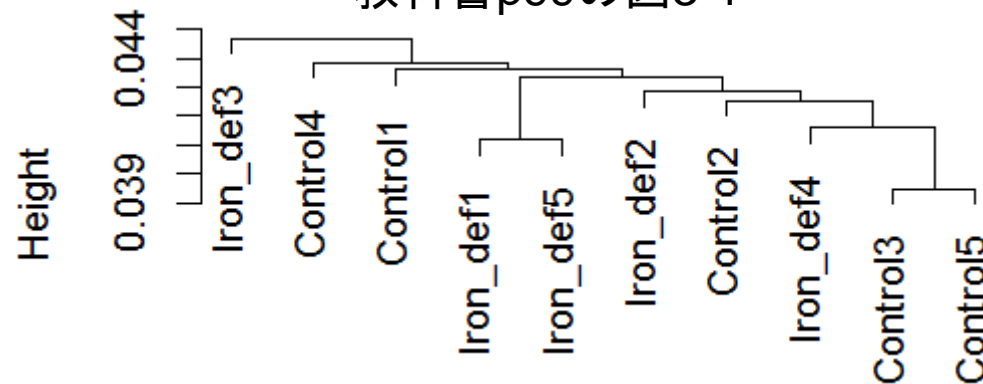
GSE30533 (ラット)

■ Kamei et al., PLoS One, 8: e65732, 2013

- GSE30533、GPL1355 (Affymetrix Rat Genome 230 2.0 Array)、31,099 probesets
- ラット10サンプル: 全てLiver (肝臓) サンプル
- iron-deficient diet (Iron_def) 5サンプル 対 control diet (Control) 5サンプル

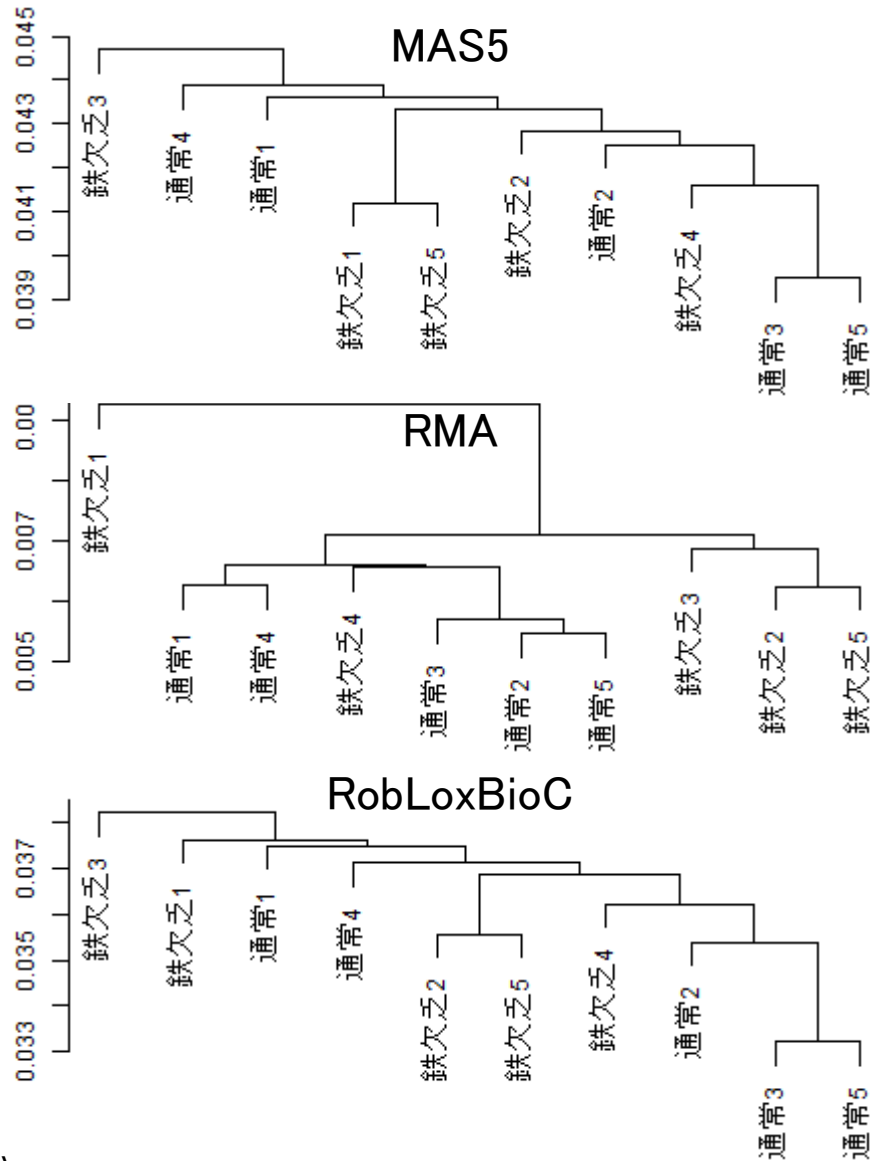


教科書p99の図3-1

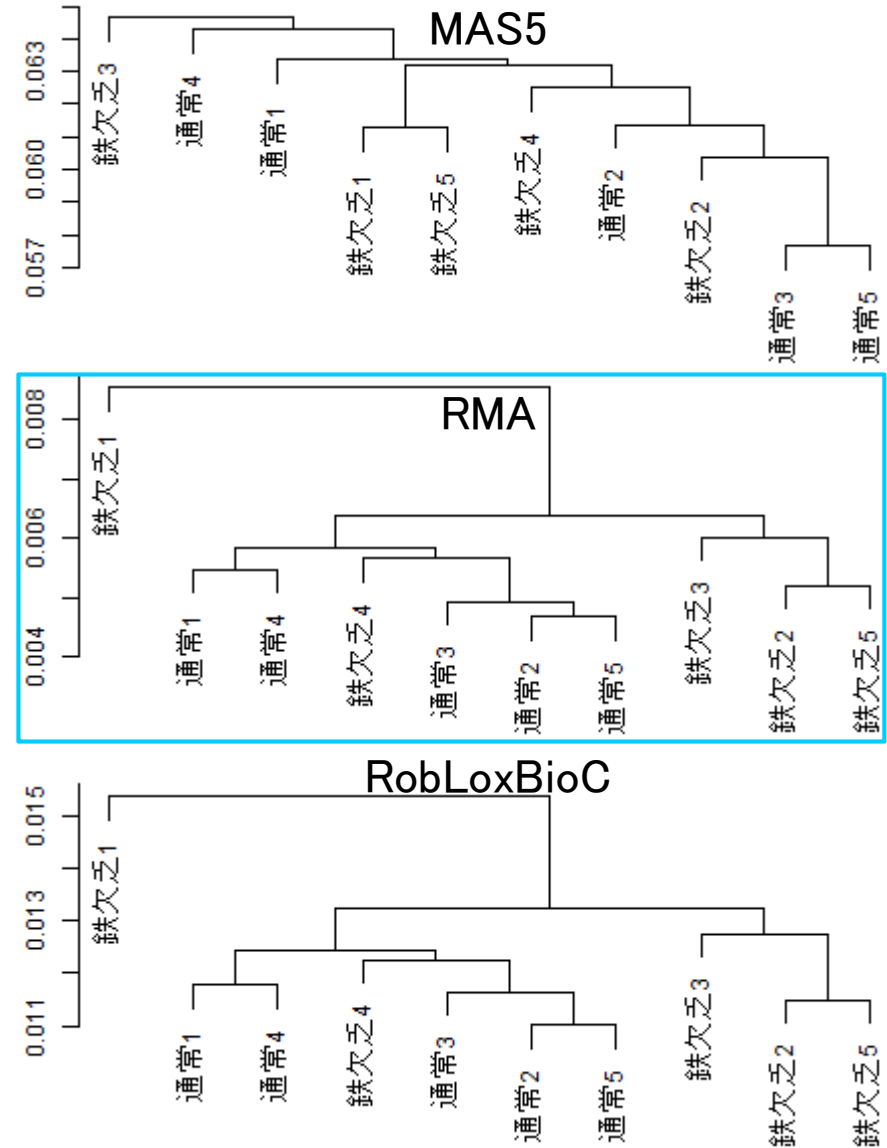


GSE30533 (ラット)

1 - Spearman相関係数

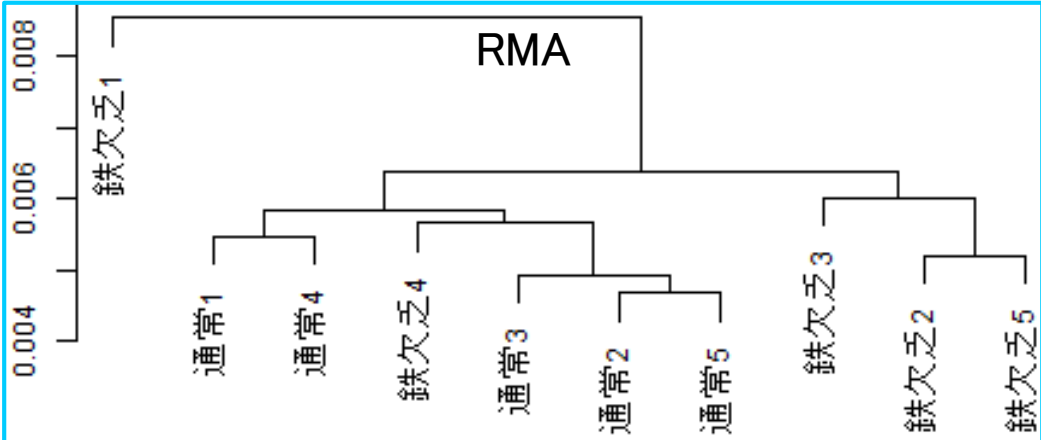


1 - Pearson相関係数

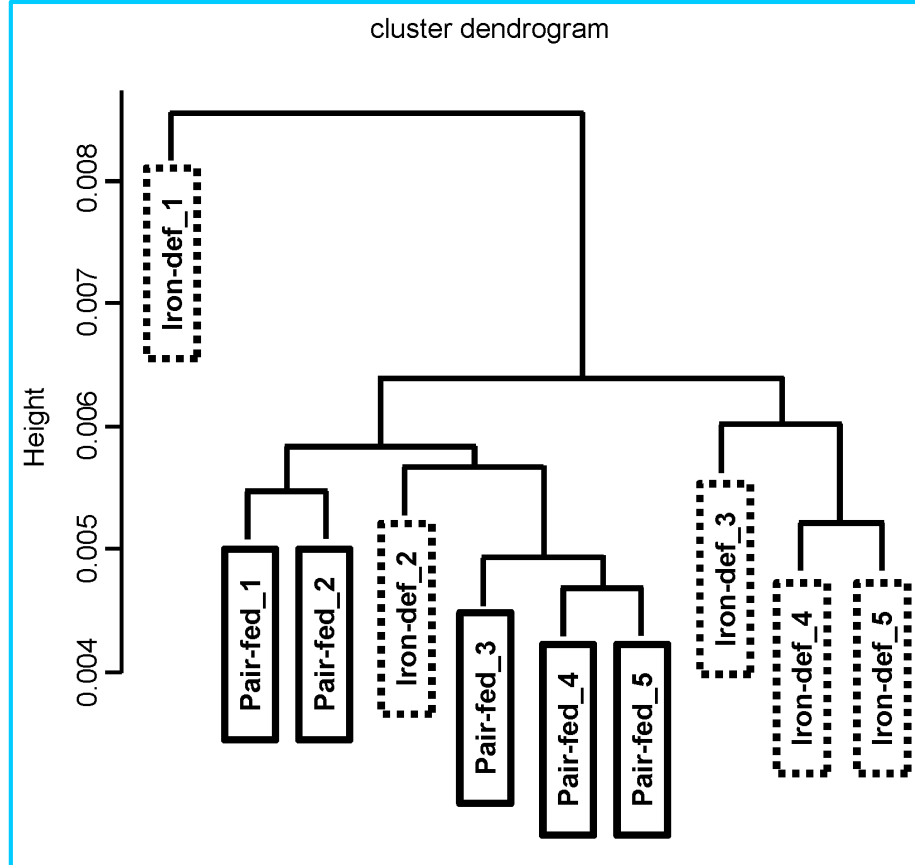


(サンプルのラベル番号が異なるだけで実質的には)同じ結果

GSE30533 (ラット)



原著論文のFigure S1



Contents

- 前処理法の適用(プローブレベルデータ → 発現行列データ)
 - MAS法、RMA法、RMX法(RobLoxBioC)、IRON法(教科書の § 2.2.2~2.2.4)
 - データの正規化(グローバル正規化、quantile正規化)、課題1
 - 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2
 - 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3
- 実験デザイン(教科書の § 3.2.2)

同一アレイデータはマージ可能

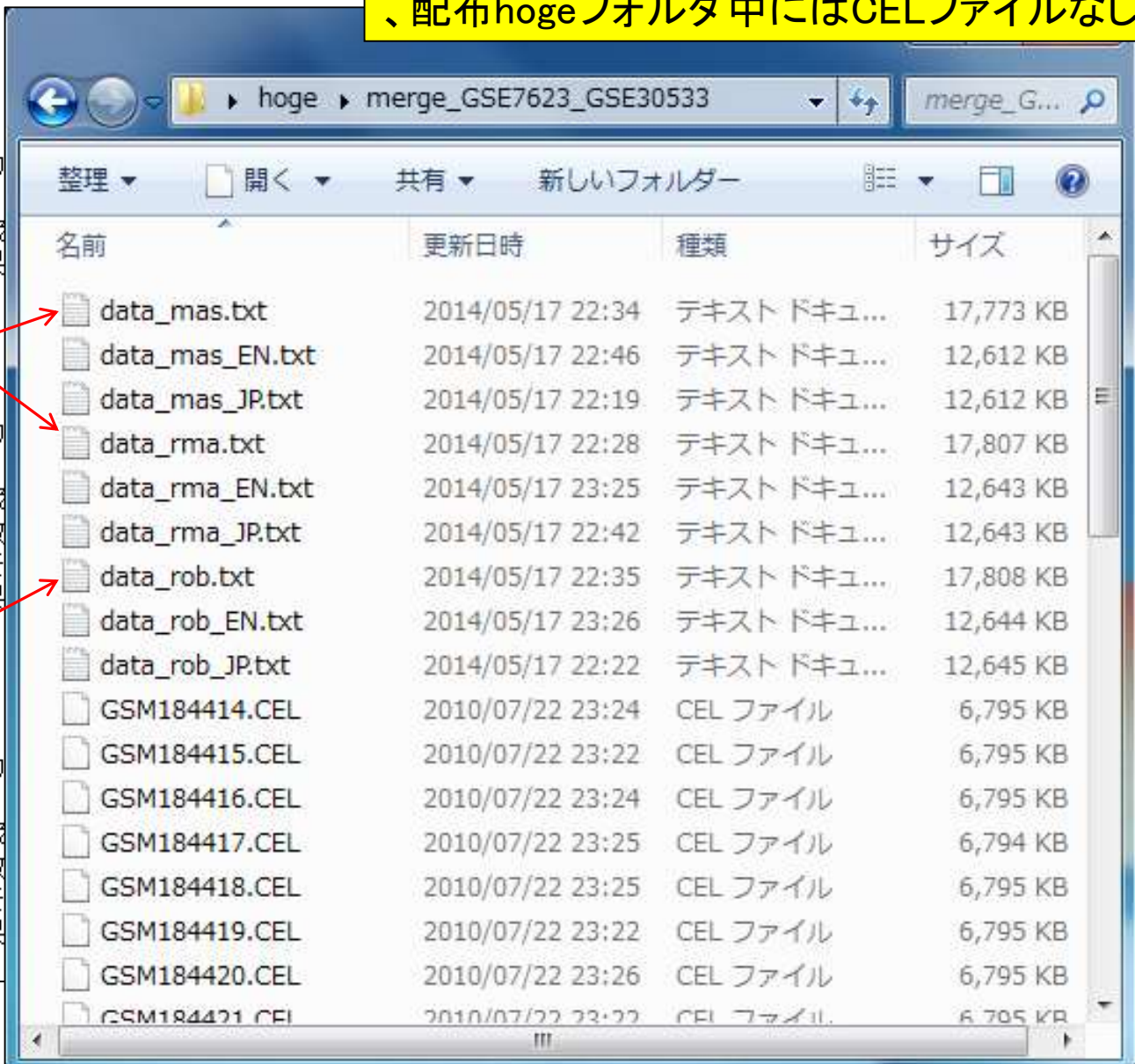
この2つの論文は同一プラットフォーム(同一アレイ)を利用。3' 発現アレイを用いることで、他の多くのデータセットとの比較が可能。

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

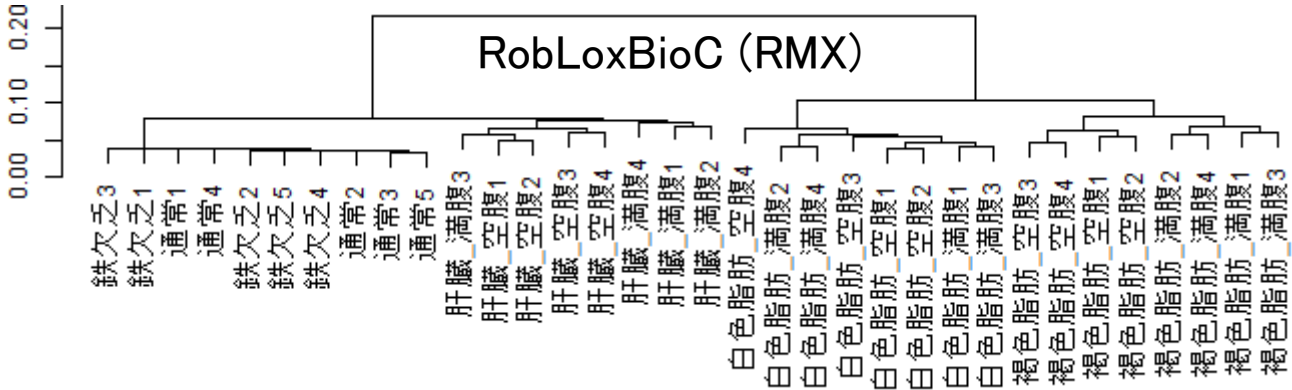
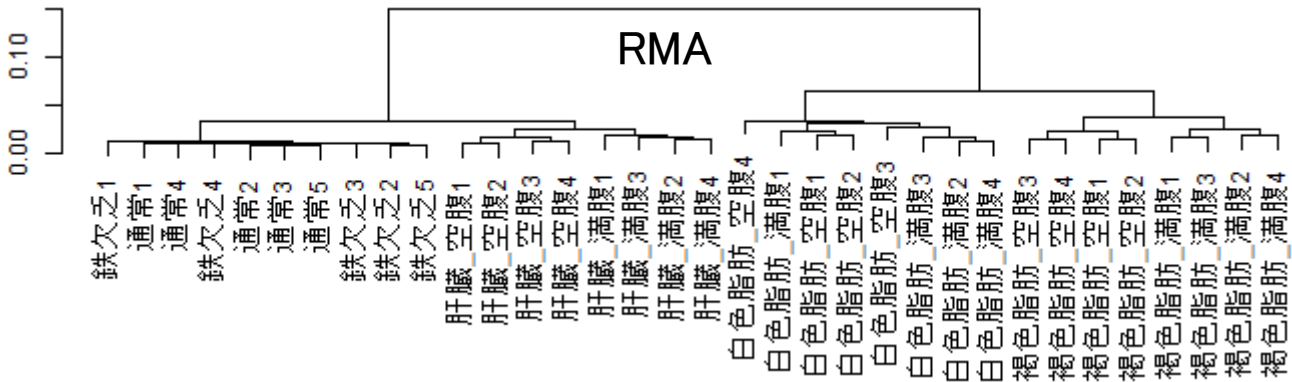
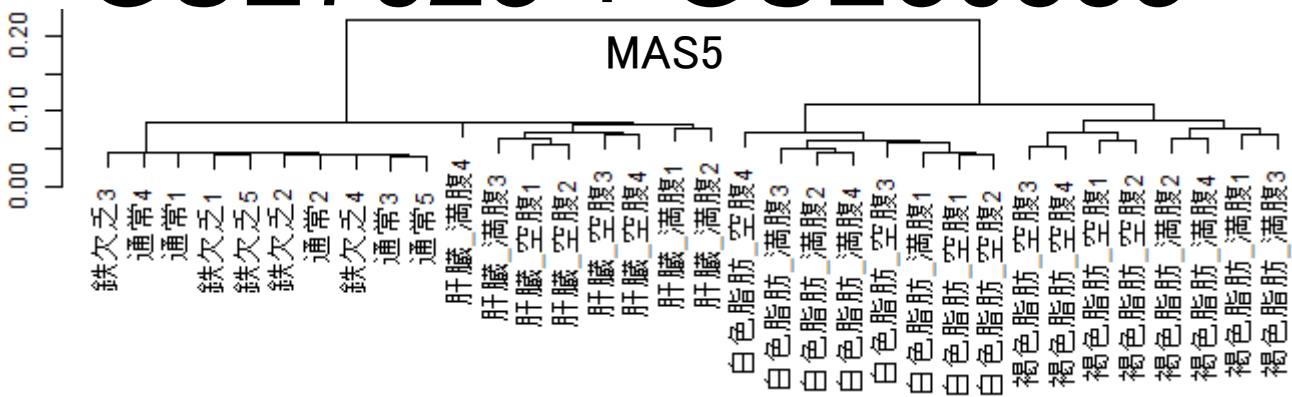
rcode_preprocessing2.txtを用いてGSE7623とGSE30533の計34 CELファイルを入力として前処理法を適用。(ファイルサイズの関係上、配布hogeフォルダ中にはCELファイルなし)

```
##### ↓
### 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
↓
```



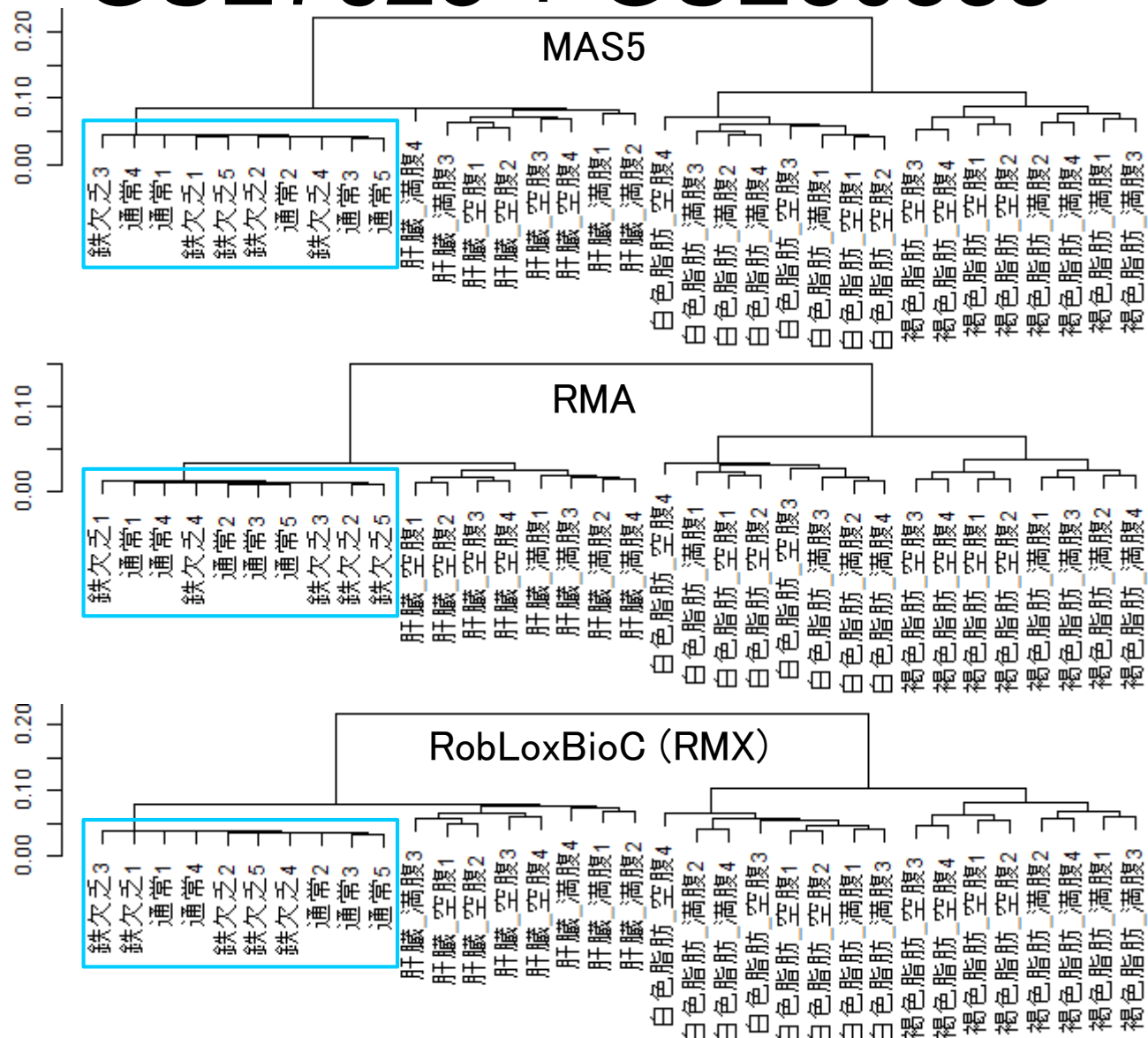
GSE7623 + GSE30533

「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。ラット10サンプル(通常 対 鉄欠乏)クラスティング結果の印象は、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)

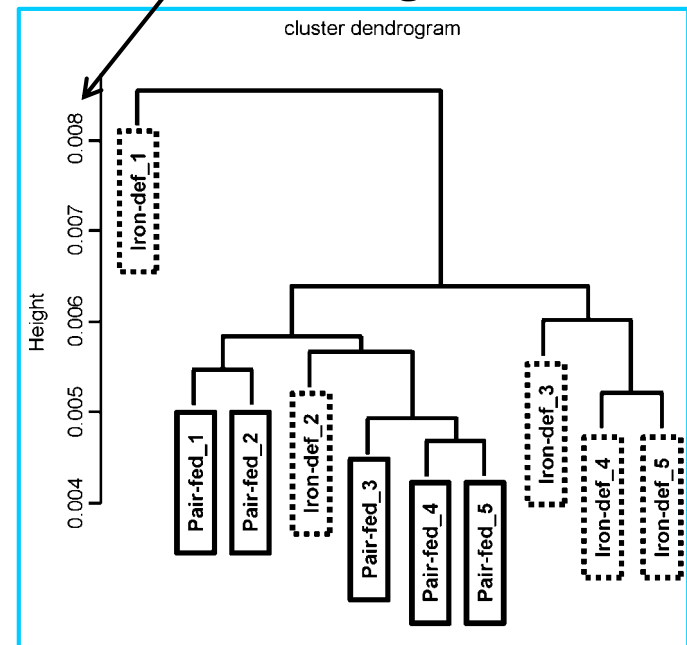


GSE7623 + GSE30533

マージすることによって(事実上初めて)距離が非常に近い(つまりサンプル間の類似度が極めて高い)ので、「Iron-def_1が外れサンプルっぽく見える」といった議論をしていたことに気づく。




原著論文のFigure S1



「1 - Pearson相関係数」にした
 い場合はspearmanをpearson
 とすればよい。

GSE7623 + GSE30533

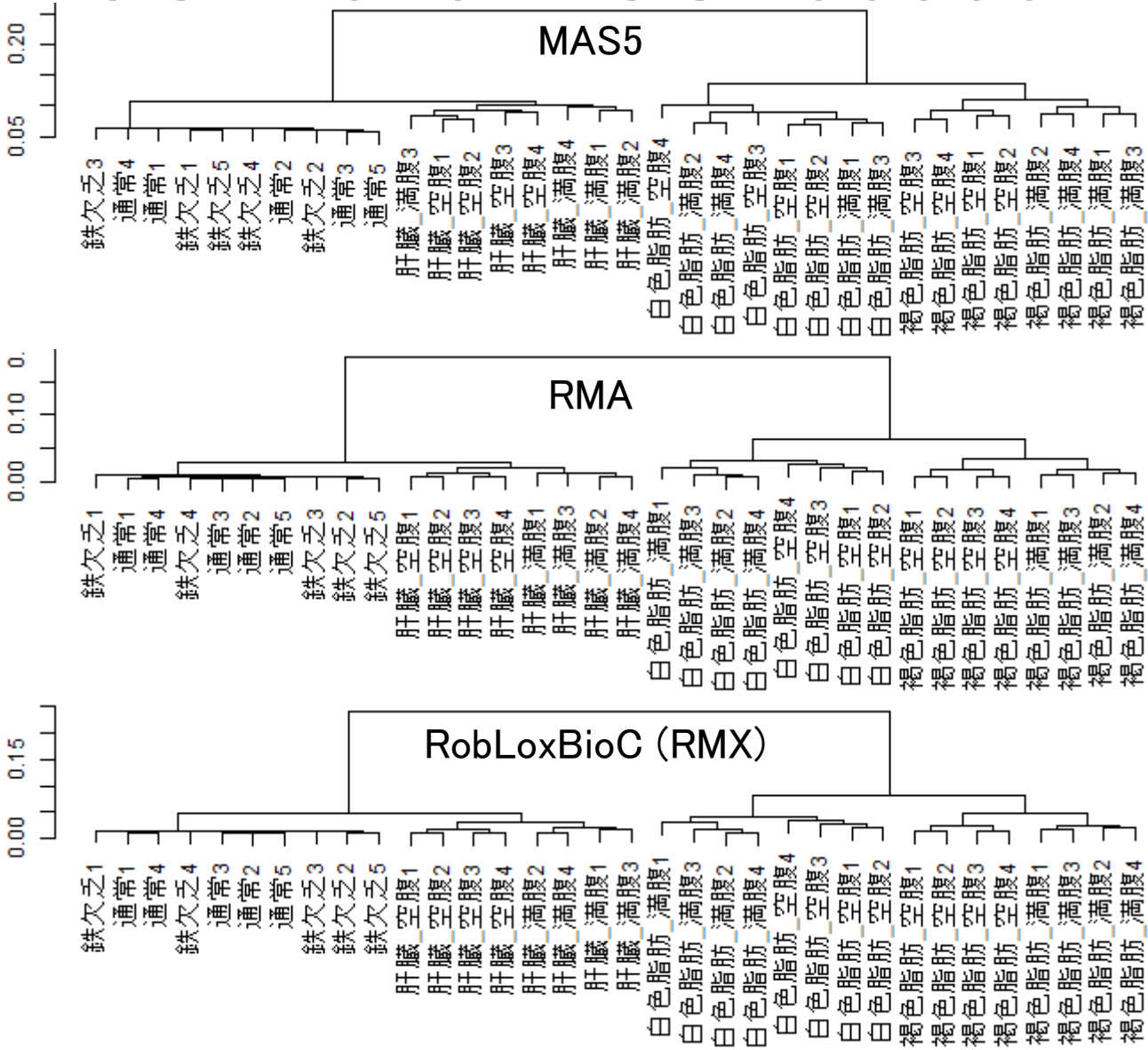


```

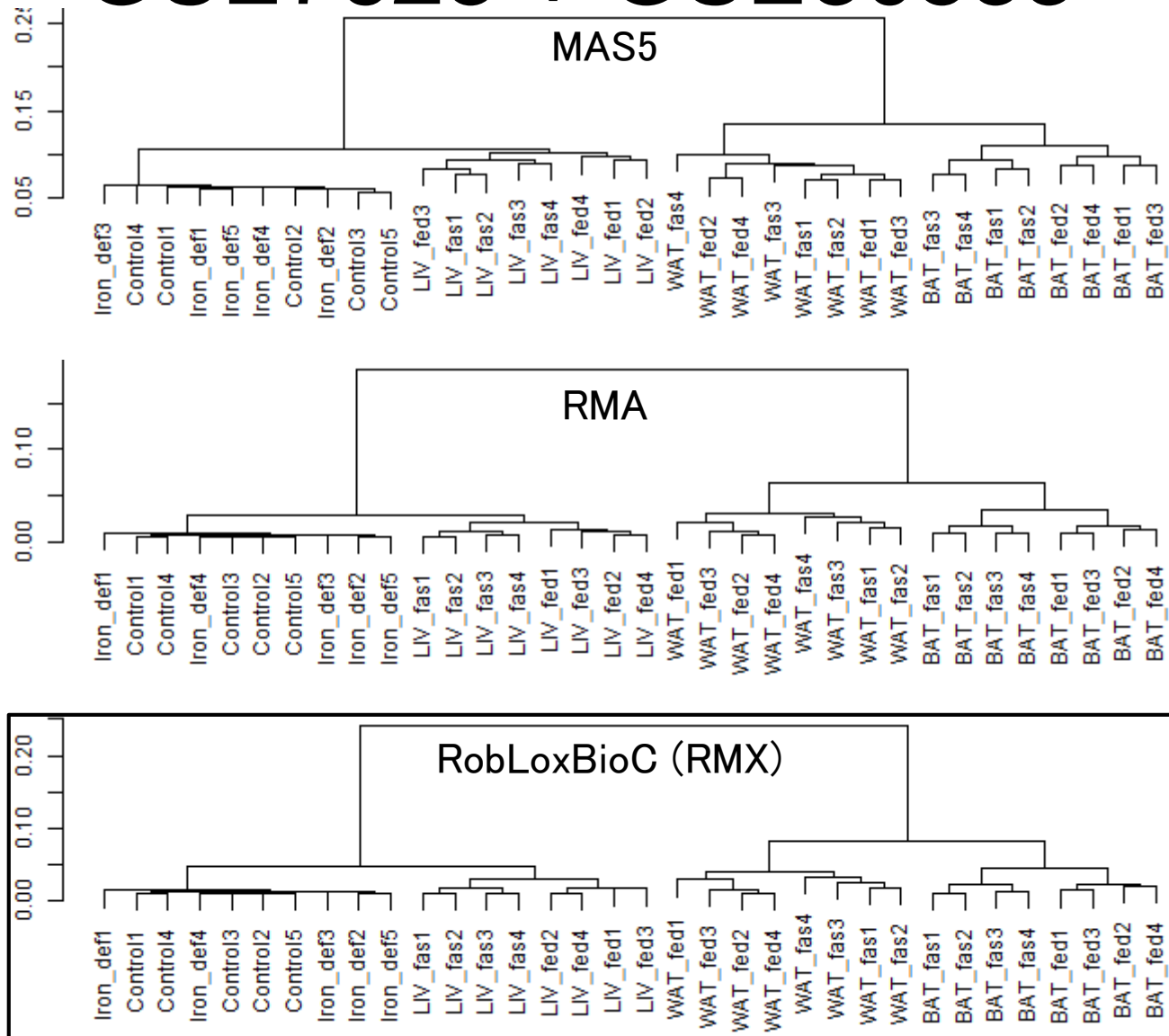
param1 <- "average" #方法(method)を指定↓
param2 <- "spearman" #相関係数の種類を指定↓
param_fig <- c(720, 310) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)↓
↓
#####↓
### MAS5データのクラスタリング, 距離: 1-Spearman相関係数, 方法: 平均連結法 ###↓
#####↓
in_f <- "data_mas_JP.txt" #入力ファイル名を指定してin_fに格納↓
out_f <- "data_mas_JP.png" #出力ファイル名を指定してout_fに格納↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み
data.dist <- as.dist(1 - cor(data, method=param2))#サンプル間の距離を計算した結果をdata.distに格納↓
out <- hclust(data.dist, method=param1) #階層的クラスタリングを実行した結果をoutに格納↓
png(out_f, pointsize=12, width=param_fig[1], height=param_fig[2])#出力ファイルの各種パラメータを指定↓
plot(out) #樹形図 (デンドログラム) の表示↓
dev.off() #おまじない↓
↓
#####↓
### RMAデータのクラスタリング, 距離: 1-Spearman相関係数, 方法: 平均連結法 ###↓
#####↓
in_f <- "data_rma_JP.txt" #入力ファイル名を指定してin_fに格納↓
out_f <- "data_rma_JP.png" #出力ファイル名を指定してout_fに格納↓
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#in_fで指定したファイルの読み込み
data.dist <- as.dist(1 - cor(data, method=param2))#サンプル間の距離を計算した結果をdata.distに格納↓
out <- hclust(data.dist, method=param1) #階層的クラスタリングを実行した結果をoutに格納↓
  
```

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。



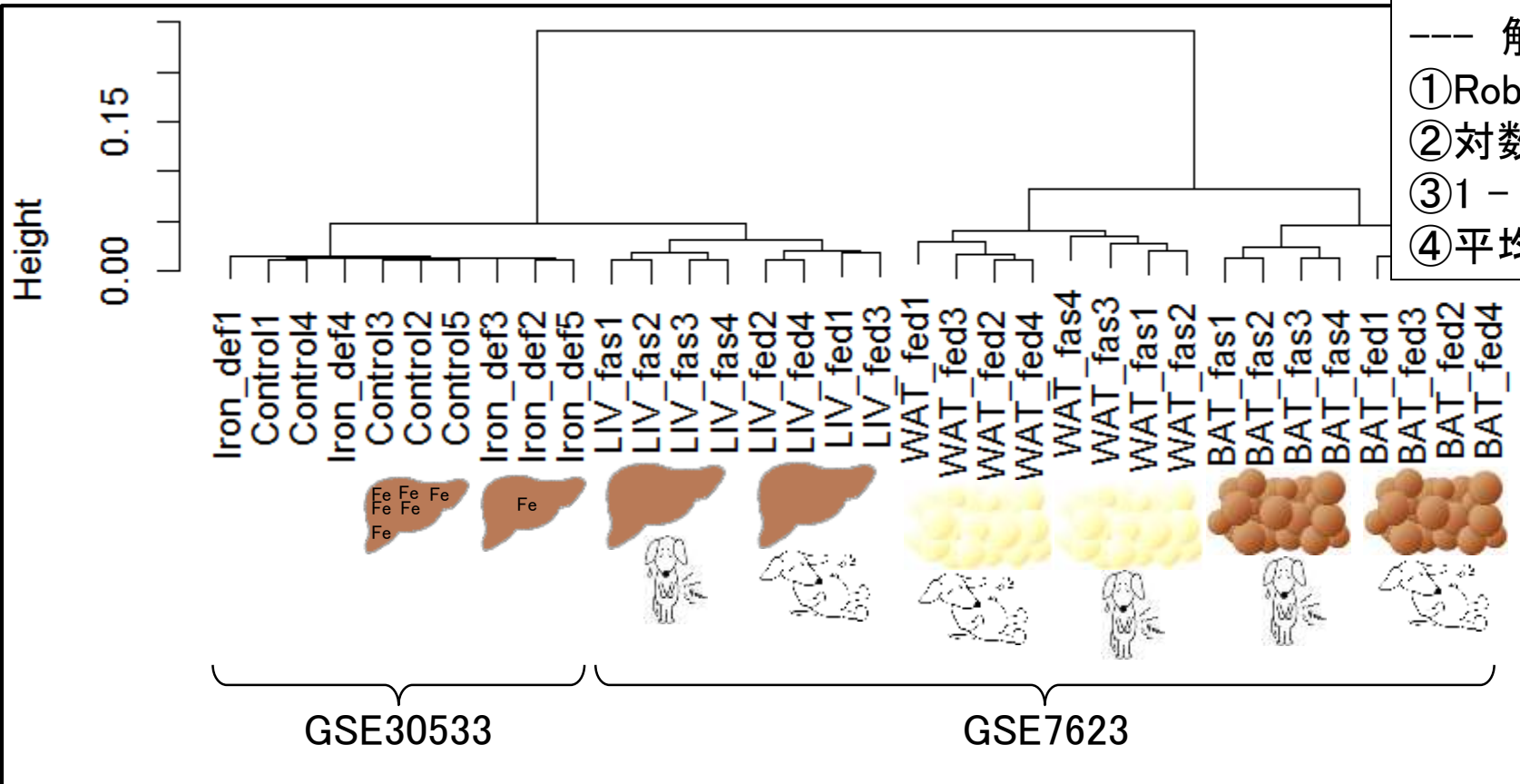
GSE7623 + GSE30533



課題3 (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 ---
- ① RobLoxBioC前処理法
 - ② 対数変換後のデータ
 - ③ 1 - Pearson相関係数
 - ④ 平均連結法



課題3

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ください。

- 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

- 前処理法の適用(プローブレベルデータ → 発現行列データ)
 - MAS法、RMA法、RMX法(RobLoxBioC)、IRON法(教科書の § 2.2.2~2.2.4)
 - データの正規化(グローバル正規化、quantile正規化)、課題1
 - 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2
 - 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3
- 実験デザイン(教科書の § 3.2.2)

実験デザイン (§ 3.2.2)

■ Affymetrix GeneChip

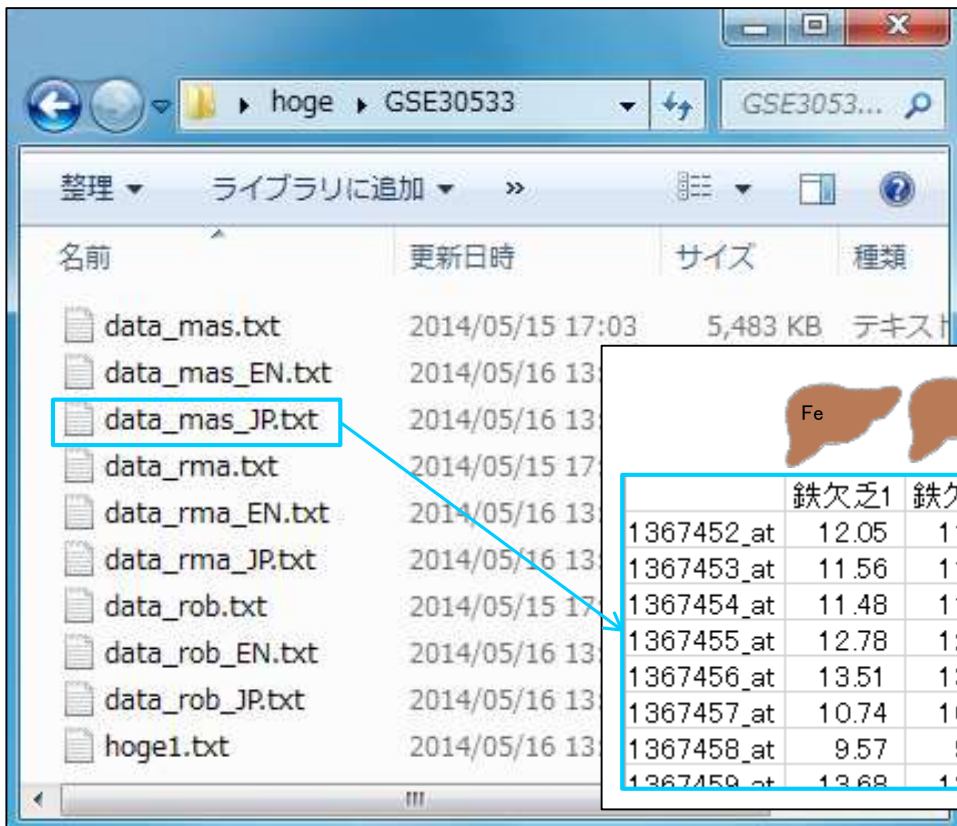
- Ge et al., *Genomics*, **86**: 127–141, 2005
 - GSE2361、GPL96 (Affymetrix Human Genome U133A Array)、22,283 probesets
 - ヒト36サンプル: Heart (心臓)、Thymus (胸腺)、Spleen (脾臓)、Ovary (卵巣)、Kidney (腎臓)、Skeletal Muscle (骨格筋)、Pancreas (膵臓)、Prostate (前立腺)、…
- Nakai et al., *Biosci Biotechnol Biochem.*, **72**: 139–148,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

実験デザイン (§ 3.2.2)

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)、81,999 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
1367459_at	13.68	13.56	13.63	13.57	13.77	13.69	13.61	13.55	13.59	13.69

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

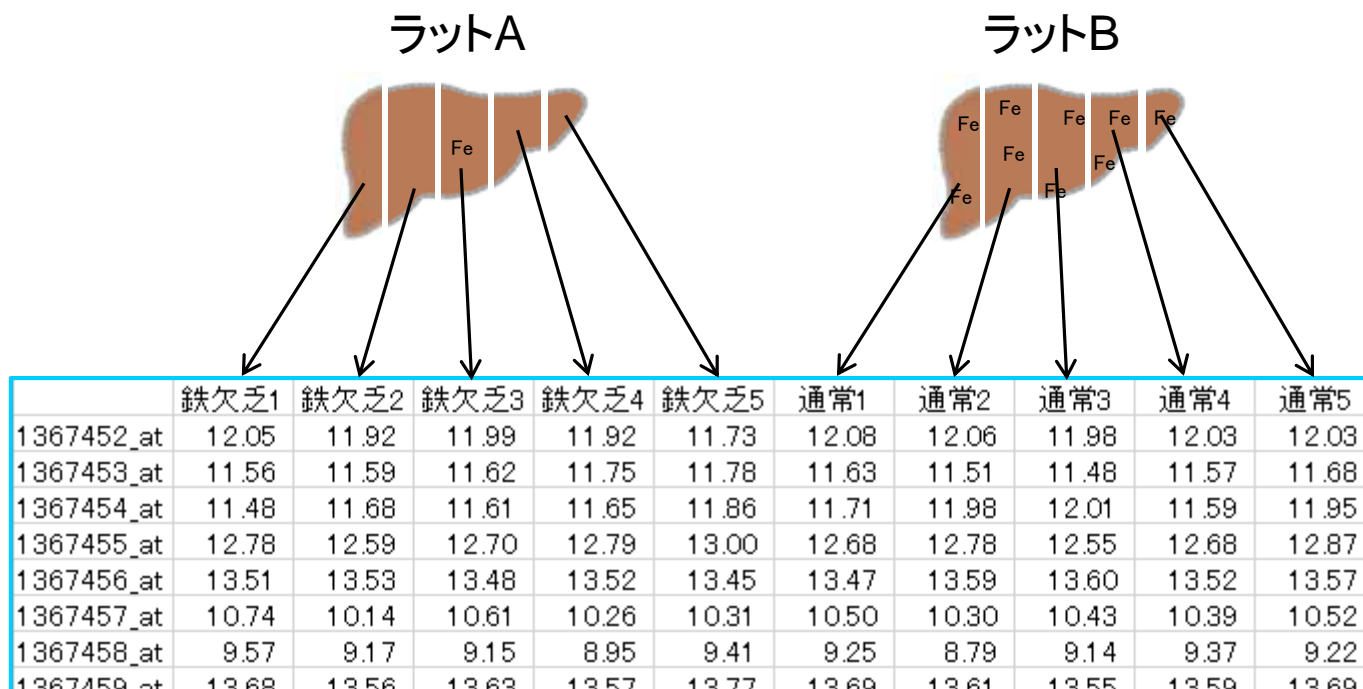
実験デザイン (§ 3.2.2)

Kamei et al., PLoS One 8: e65732 2013

- GSE30533、
- ラット10サン
- iron-deficient diet (iron_def) 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)

31,099 probesets

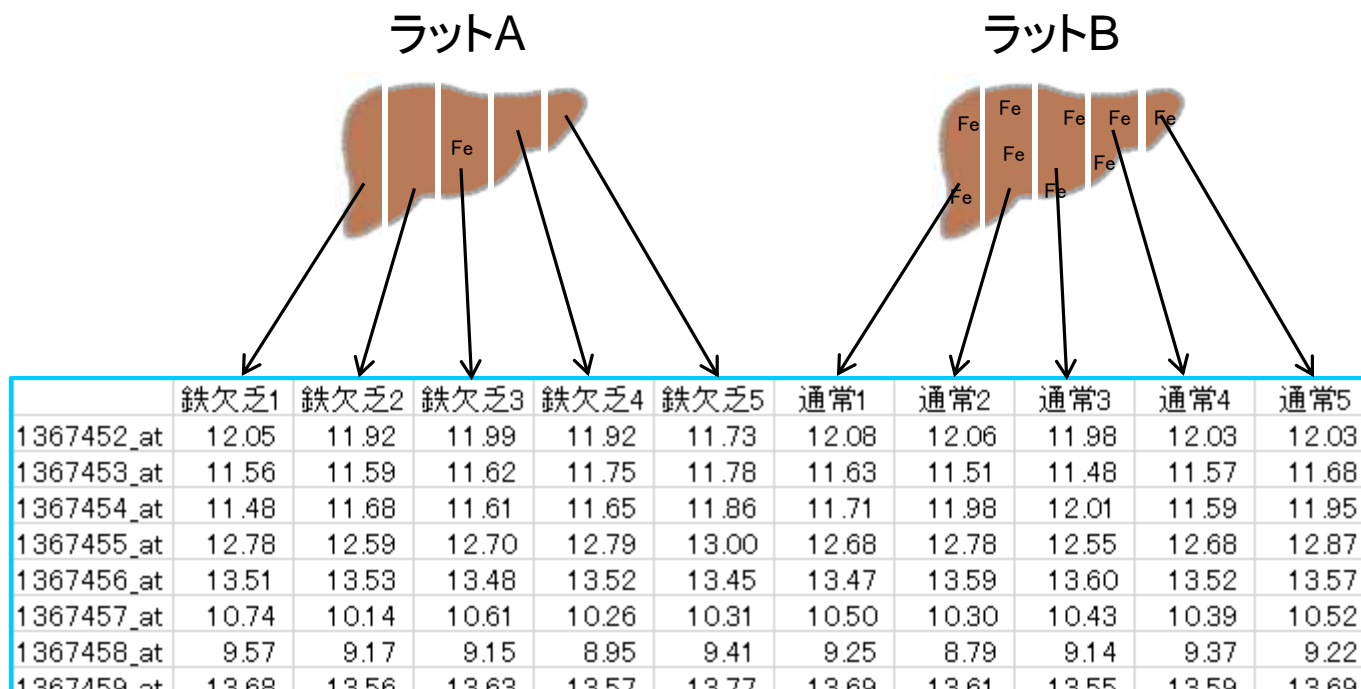


2群間での発現変動遺伝子(DEG) 検出結果は多くなる傾向。多ければいいというものではない!

実験デザイン (§ 3.2.2)

Technical replicatesだと...

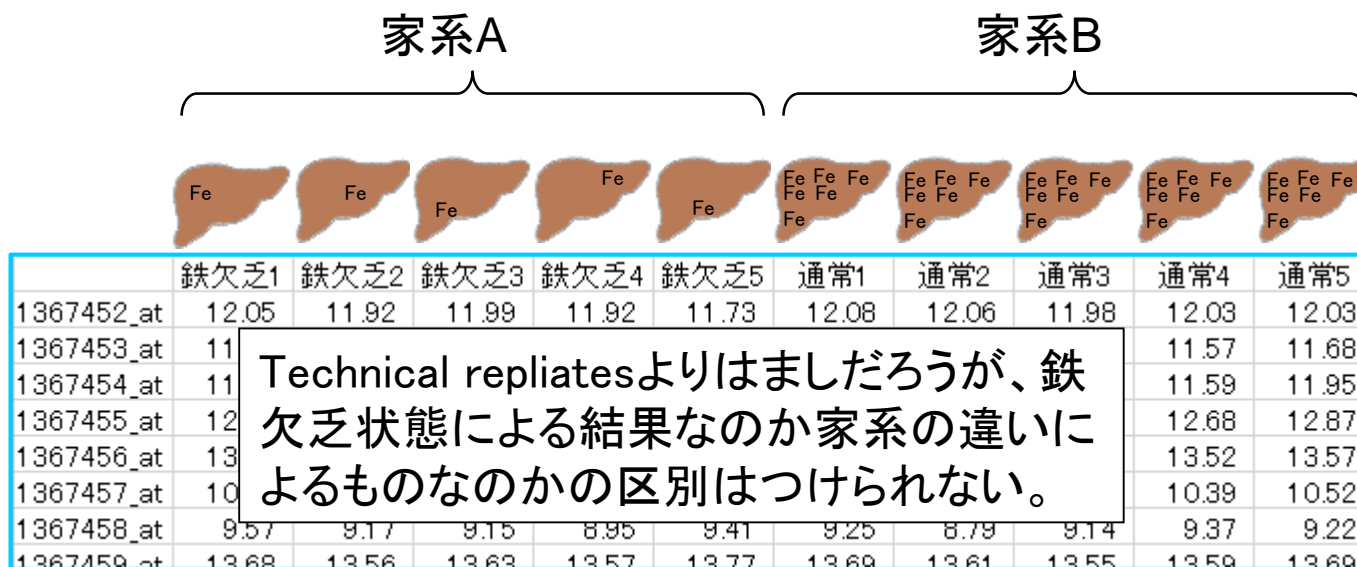
1. 自分は「鉄欠乏 対 通常」の違いを見ているつもりでも、個体間の他の違い(身長、体重など)由来要因との区別がつかない
高身長 対 低身長、低体重 対 高体重、他の病気の有無、家系の違いなど
2. 得られる結果から導き出される結論は、そのラット間のみで成立する事象であり、ラットという生物種全体に適用可能なわけではない



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…



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

