

比較トランスクリプトーム解析とその周辺： モデル、正規化、発現変動検出など

東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット

門田 幸二(かどた こうじ)

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

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

スライドPDFはウェブから取得可能です

http://www.iu.a.u-tokyo.ac.jp/~kadota/ 門田 幸二のホームページ

門田 幸二のホームページ

名前 門田 幸二(かどた こうじ)



所属 [東京大学 大学院農学生命科学研究科](#)
[アグリバイオインフォマティクス教育研究ユニット](#)

身分

講演など(上記講義以外) (last modified: 2014.02.17)

研究分野

27. 題目:「比較トランスクリプトーム解析とその周辺:モデル、正規化、発現変動検出など」, [よく分かる次世代シーケンサー解析ワークショップ](#), 九州大学(福岡), 2014.03.19

所属学会

26. 題目:「Rでゲノム・トランスクリプトーム解析」, [HPCIチュートリアル・バイオインフォマティクス実習コース](#), 生命情報工学研究センター(東京), 2014.03.07

研究テーマ(1)

トランスクリプトーム解析などによって得られるデータの応用を目指します。これまでの主な研究テーマとして、[「Rで塩基配列解析](#)

25. 題目:「[トランスクリプトーム解析の現況2013\(詳細版\)](#)」, 東京大学大学院農学生命科学研究科第124回アグリバイオインフォマティクスセミナー, 東京大学(東京), 2013.11.01

24. 題目:「[トランスクリプトーム解析の現況:マイクロアレイ vs. RNA-seq](#)」, [生命医薬情報学連合大会「オミックス・計算そして創薬」・オミックス解析における実務者意見交換会](#), タワーホール船堀(東京), 2013.10.30

23. 題目:「[食品機能解析研究とバイオインフォマティクス](#)」, [日本農芸化学会2013年度大会・シンポジウム4SY08](#), 東北大学(宮城), 2013.03.27

22. 題目:「[Rでトランスクリプトーム解析](#)」, [HPCIチュートリアルセミナー](#), 生命情報工学研究センター(東京), 2013.03.07

21. 題目:「[Rでトランスクリプトーム解析](#)」, [HPCIチュートリアルセミナー](#), 生命情報工学研究センター(東京), 2012.03.09

20. 題目:「[Rによるトランスクリプトーム解析～NGS由来塩基配列データを自在に解析する～](#)」, [Rでつなぐ次世代オミックス情報統合解析研究会](#), 理化学研究所横浜研究所(神奈川), 2012.02.22

19. 題目:「[RNA-Seqデータ解析リテラシー](#)」, [Illumina Webinar Series: RNAシーケンスを始めよう・セッション3:データ解析](#), イルミナ株式会社(東京), 2011.11.17



自己紹介

- 1995年3月
 - 高知工業高等専門学校・工業化学科 卒業
- 1997年3月
 - 東京農工大学・工学部・物質生物工学科 卒業
- 1999年3月
 - 東京農工大学・大学院工学研究科・物質生物工学専攻 修士課程修了
- 2002年3月
 - 東京大学・大学院農学生命科学研究科・応用生命工学専攻 博士課程修了
 - 学位論文:「cDNAマイクロアレイを用いた遺伝子発現解析手法の開発」
(指導教官:清水謙多郎教授)
- 2002/4/1~
 - 産総研・生命情報科学研究センター(CBRC) 産総研特別研究員
- 2003/11/1~
 - 放医研・先端遺伝子発現研究センター 研究員
- 2005/2/16~
 - 東京大学・大学院農学生命科学研究科
特任助手→...

参考URL

(Rで)塩基配列解析(主にNGS、RNA-seq、トランスクリプトーム解析)

(last modified 2014/03/16, since 2010)

What's new?

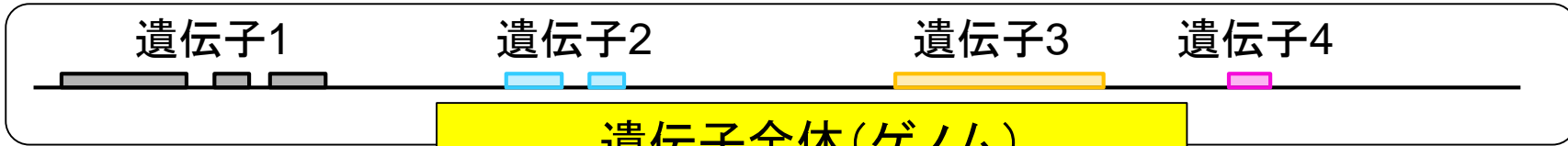
- 門田幸二 著 [シリーズ Useful R 第7巻トランスクリプトーム解析](#)が2014年4月10日に共立出版から出ます。
- [参考資料\(講義、講習会、本など\)](#)の項目を追加しました。(2014/03/16) **NEW**
- 私の所属する[アグリバイオインフォマティクス教育研究プログラム](#)では、平成26年度も(東大生に限らず)バイオインフォ関連講義を行います。受講希望者は平成26年4月7日18:00-18:45に東大農学部二号館二階化学第一講義室にて開催予定の受講ガイダンスに出席してください。例年東大以外の企業の方、研究員、学生が二割程度は受講しております。このウェブページと直接関連する講義は「[ゲノム情報解析基礎](#)」と「[農学生命情報科学特論](#)」ですが、背景理論の説明などは「[機能ゲノム学](#)」でも行います。興味ある科目のみの受講も可能ですので、お気軽にどうぞ。(2014/03/03) **NEW**
- 一連の解析パイプライン(RNA-seqデータ取得 -> マッピング -> カウントデータやRPKMデータ取得 -> サンプル間クラスターリングや発現変動解析およびM-A plot描画まで)のクラスターリング部分をアップデートしました。項目名の一番下のほうです。(2014/02/26) **NEW**
- 2014年3月17-19日に九州大学にて、ワークショップ([よく分かる次世代シーケンサー解析～最先端トランスクリプトーム解析～](#))が開催されます。私は3日目(3/19, 13:00-16:30)を担当します。興味ある方はどうぞ。締切は確か2/21です。(2014/02/17) **NEW**
- 発現変動解析用Rパッケージ [TCC](#) (ver. 1.2.0; [Sun et al., BMC Bioinformatics, 2013](#))がBioconductorよりリリースされました。最新版を利用したい方は、R (ver. 3.0.2)をインストールしたのち、Bioconductor (ver. 2.13)をインストールしてください。(2013/10/17)

- [はじめに](#) (last modified 2014/01/30)
- [参考資料\(講義、講習会、本など\)](#) (last modified 2014/03/16)
- [過去のお知らせ](#) (last modified 2014/03/03) **NEW**
- [Rのインストールと起動](#) (last modified 2013/09/27)
- [サンプルデータ](#) (last modified 2014/03/05) **NEW**
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2013/09/29)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2013/10/10)
- イントロ | 一般 | [任意のキーワードを含む行を抽出\(基礎\)](#) (last modified 2014/02/06)
- イントロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2013/09/29)

自前PCでやる場合はここを参考にして必要なパッケージを予めインストールしておかねばなりません。数時間程度かかります。

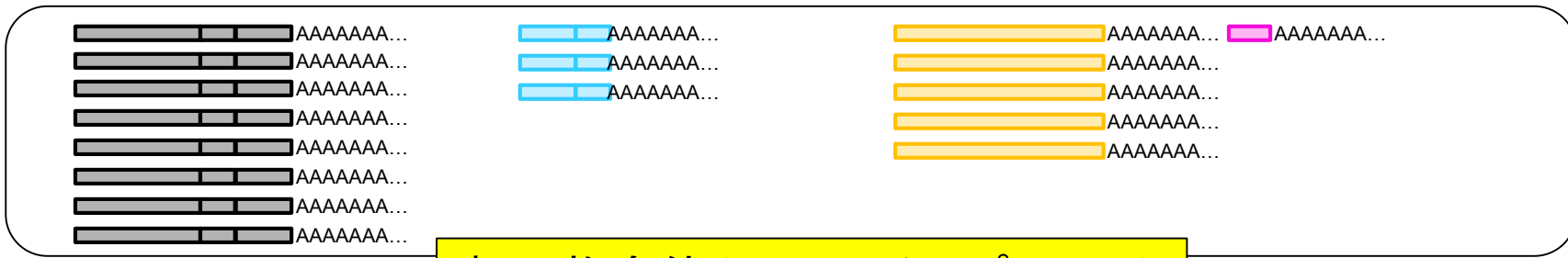
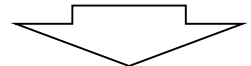
トランスクリプトームとは

- ある状態のあるサンプルのあるゲノムの領域



遺伝子全体(ゲノム)

どの染色体上のどの領域にどの遺伝子があるかは調べる個体が同じなら、目だろうが心臓だろうが不変

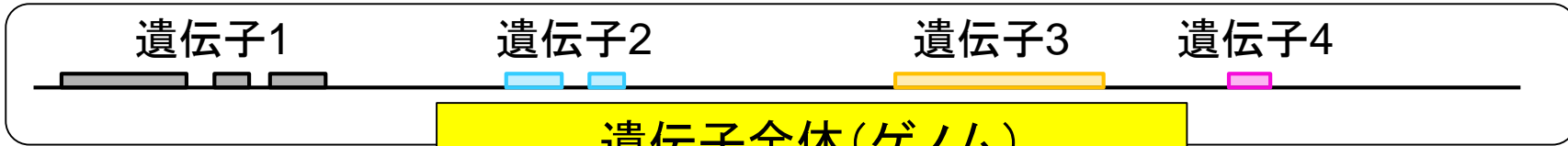


転写物全体(トランスクリプトーム)

- ・遺伝子1は沢山転写されている(発現している)
- ・遺伝子4はごくわずかしか転写されていない
- ・...

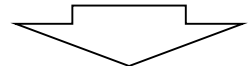
トランスクリプトームとは

- ある状態のあるサンプルのあるゲノムの領域



遺伝子全体(ゲノム)

どの染色体上のどの領域にどの遺伝子があるかは調べる個体が同じなら、目だろうが心臓だろうが不変



転写物全体(トランスクリプトーム)

- ・遺伝子2は光刺激に应答して発現亢進
- ・遺伝子4も光刺激に应答して発現亢進



トランスクリプトーム解析の目的は様々

- トランスクリプトーム配列取得
 - ゲノム配列既知の場合 : CufflinksやTIGERなどを用いて遺伝子構造推定
 - ゲノム配列未知の場合 : Trinityなどのトランスクリプトーム用アセンブラを実行
- 遺伝子またはisoformごとの発現量の正確な推定(サンプル内比較)
 - TIGER (Nariai et al., *Bioinformatics*, 2013)などを利用して発現量情報を得る
 - ある特定のサンプル内での遺伝子間の発現量の大小関係を知りたい
 - 配列長やGC含量などの各種補正がポイント
- サンプル間で発現変動している遺伝子の同定(サンプル間比較)
 - TCCパッケージなどを利用して発現変動遺伝子(DEG)を得る
 - ライブラリサイズ(総リード数)や発現している遺伝子の組成の補正がポイント

3/19は比較トランスクリプトーム解析の話

Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

研究目的別留意点：サンプル内比較

■ 発現量補正の基本形：カウント数 × $\frac{\text{定数}}{\text{配列長} \times \text{総リード数}}$

- RPK (Reads per kilobase)
- RPM (Reads per million)
- RPKM (Reads per kilobase per million)

■ 同一サンプル内での異なる遺伝子間の発現レベル比較の場合

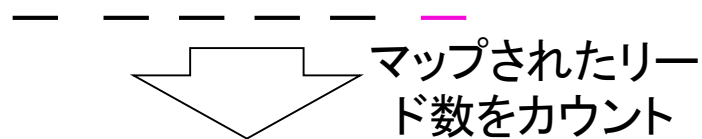
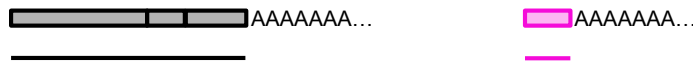
- 配列長由来bias: 長いほど沢山sequenceされる
 - RPKMやFPKMなどの配列長を考慮して正規化されたデータで解析
- GC含量由来bias: カウント数の分布がGC含量依存的である
 - Risso et al., *BMC Bioinformatics*, 12: 480, 2011
 - Benjamini and Speed, *Nucleic Acids Res.*, 40: e72, 2012



総リード数(ライブラリサイズ or sequence depth)補正は不必要
理由: 遺伝子間の発現レベルの大小関係は定数倍しても不変

配列長の補正

- 配列長が長い遺伝子ほど沢山sequenceされる
 - それらの遺伝子上にマップされる生のリード数が増加傾向
 - 配列長が長い遺伝子ほど発現レベルが高い傾向になる

発現レベルが同じで長さの異なる二つのmRNAs



mRNA	リード数
 AAAAAAA...	5
 AAAAAAA...	1

1つのサンプル内での異なる遺伝子間の発現レベルの高低を配列長を考慮せずに比較することはできない

配列長を考慮した発現量推定のイメージ

- gene1: 3 exons (middle length), 14 reads mapped (**low** coverage)
- gene2: 3 exons (middle length), 56 reads mapped (**high** coverage)
- gene3: 2 exons (**short** length), 12 reads mapped (middle coverage)
- gene4: 2 exons (**long** length), 31 reads mapped (middle coverage)

マップされたリード分布



生リードカウント結果

補正度の発現量

- ・長さが同じならリード数の多い方が発現量高い (gene 1 vs. 2)
- ・長いほどマップされるリード数が多くなる効果を補正する必要がある (gene 3 vs. 4)

1つのサンプル内で転写物または遺伝子間の発現レベルの大きさを比較したい場合には配列長を考慮すべきである

配列長の補正

mRNA	リード数	配列長 (in bp)
 AAAAAAA...	5	1500
 AAAAAAA...	1	300

■ 前提条件: 配列長が既知

■ 補正の基本戦略: 配列長で割る

□ 「1 / 配列長」を掛ける場合

→ 「塩基あたりの平均のリード数」を計算しているのと等価

□ 「1000 / 配列長」を掛ける場合

→ 「その遺伝子の配列長が1000bpだったときのリード数(or カウント数)」と等価

Reads Per Kilobase (RPK)

Counts Per Kilobase (CPK)



研究目的別留意点：サンプル間比較

■ 発現量補正の基本形：カウント数 × $\frac{\text{定数}}{\text{配列長} \times \text{総リード数}}$

- RPK (Reads per kilobase)
- RPM (Reads per million)
- RPKM (Reads per kilobase per million)

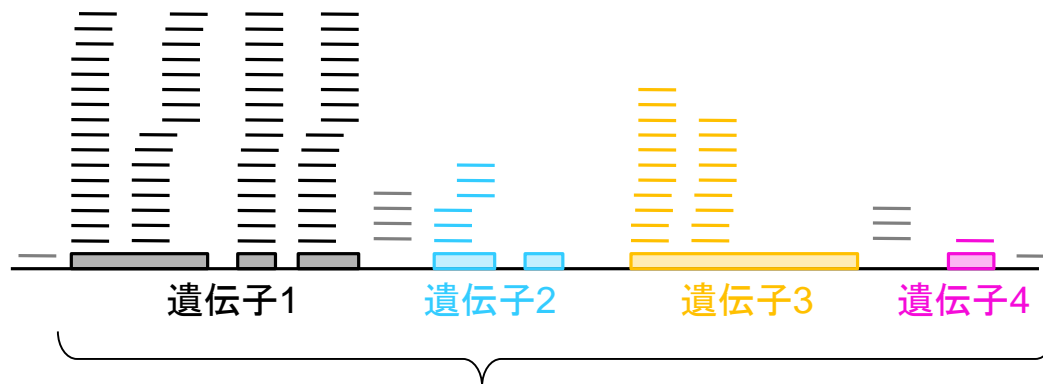
■ 異なるサンプル間での同一遺伝子間の発現レベル比較の場合

- 総リード数の違い：総リード数が x 倍違うと全体的に x 倍変動…
 - RPM正規化で全体を揃えることは基本
- 組成の違い：サンプル特異的高発現遺伝子の存在で比較困難に…
 - TMM正規化法(Robinson and Oshlack, *Genome Biol.*, 11: R25, 2010)
 - TbT正規化法(Kadota et al., *Algorithms Mol. Biol.*, 7: 5, 2012)

配列長やGC bias補正は少なくとも理論上は不必要
理由：同一遺伝子に対して掛かる係数はサンプル間で同じ

RNA-Seqデータの正規化の一部

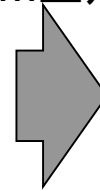
- 発現しているRNA量の総和はサンプル間で一定



	T1	T2
遺伝子1	40	7
遺伝子2	6	15
遺伝子3	20	5
遺伝子4	1	1

総リード数 **67** **28**

RPM正規化



	T1	T2
遺伝子1	597014.9	250000.0
遺伝子2	89552.2	535714.3
遺伝子3	298507.5	178571.4
遺伝子4	14925.4	35714.3

総リード数 1000000 1000000

Reads Per Million mapped reads (RPM)
 正規化後の総リード数が100万 (one million) になるように補正
 例: T1の正規化係数 = $1000000 / 67$

RPKM

- Reads per kilobase (of exon) per million (mapped reads)
- 配列長が1000 bpだったとき、かつ総リード数が100万だったときのカウント数

$$\text{RPKM} = \text{カウント数} \times \frac{1,000}{\text{配列長}} \times \frac{1,000,000}{\text{総リード数}} = \text{カウント数} \times \frac{1,000,000,000}{\text{配列長} \times \text{総リード数}}$$

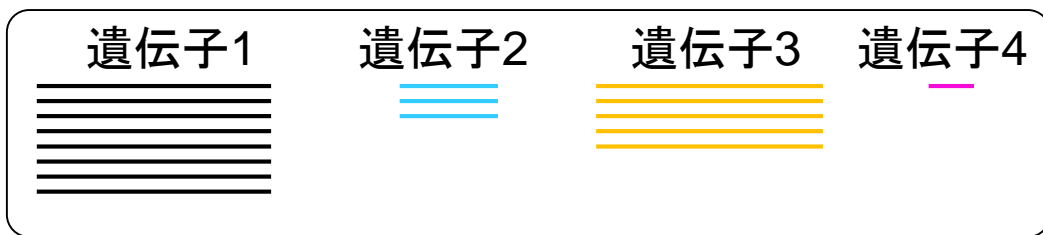
sample_length_count.txt			hoge1.txt		
ID	Length	Count	rownames(data)	Length	Count
NM_203348.1	3543	3	NM_203348.1	3543	0.355
NM_001008737.1	1897	19	NM_001008737.1	1897	4.199
NM_001037228.1	537	7	NM_001037228.1	537	5.465
NM_033183.2	886	0	NM_033183.2	886	0
NM_138368.3	4443	56	NM_138368.3	4443	5.284
NM_152833.2	2844	85	NM_152833.2	2844	12.53
NM_001100111.1	682	0	NM_001100111.1	682	0
NM_001102659.1	1376	0	NM_001102659.1	1376	0
NM_001104548.1	888	3	NM_001104548.1	888	1.416

総リード数 = 2385273

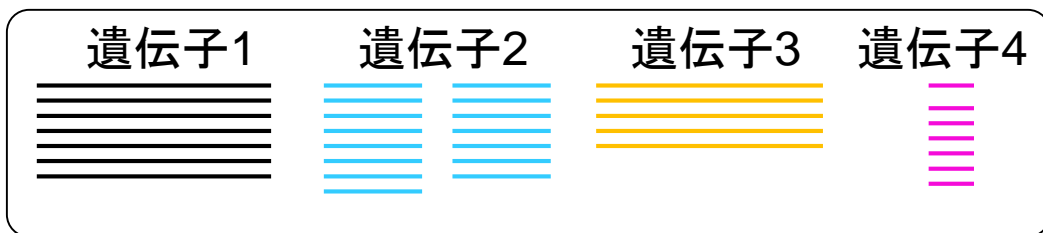
比較トランスクリプトーム解析

- 比較するサンプルまたはグループ間での発現変動遺伝子 (Differentially Expressed Genes; DEGs) 検出が解析の主要部分

光刺激前 (T1) の目のトランスクリプトーム



光刺激後 (T2) の目のトランスクリプトーム

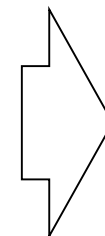


数値化



	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	1	7
...

発現変動解析

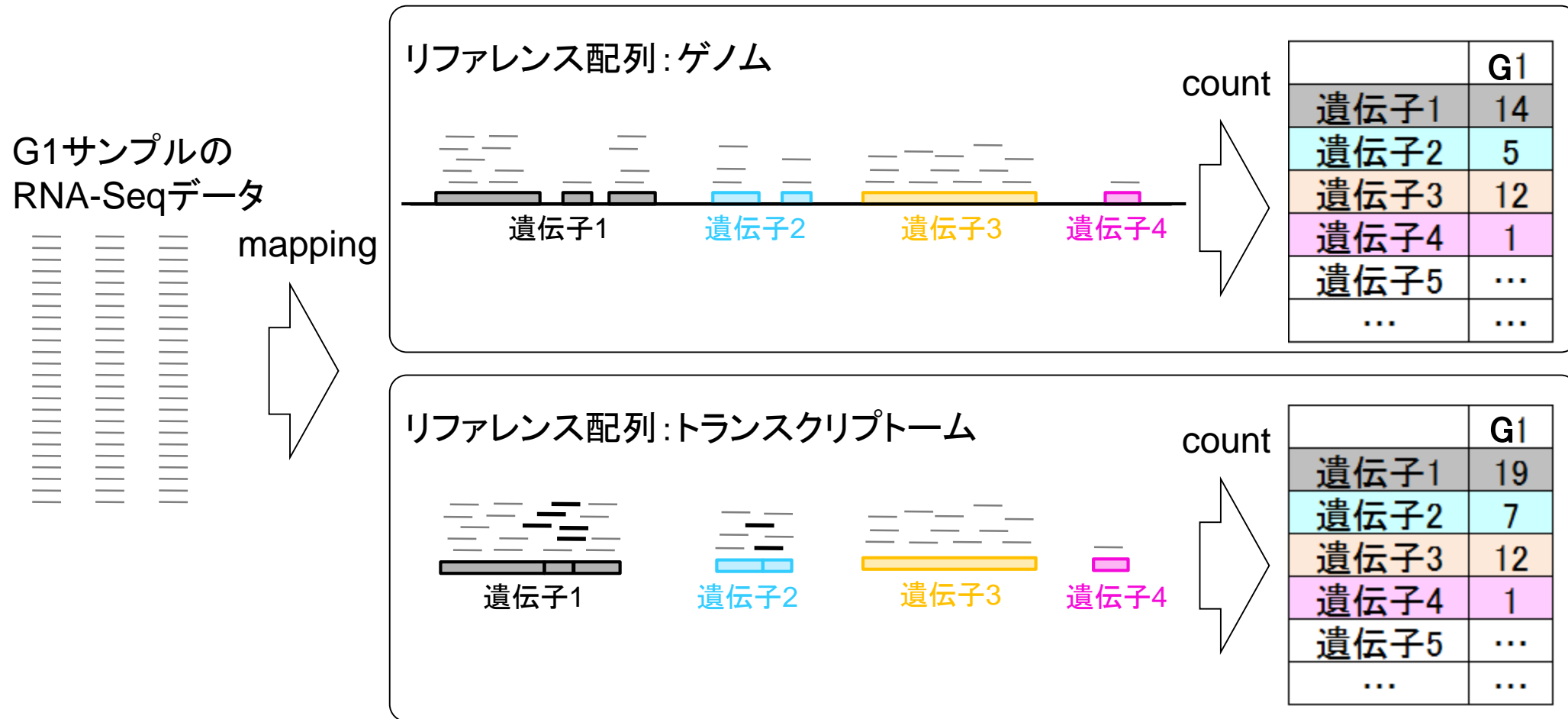


DEGリスト
(遺伝子2, 4)

- ・実験デザインは **biological replicates**
- ・発現変動解析用Rパッケージは **TCC** (*edgeR* や *DESeq* ではない)
 - ・パッケージの入力データは **カウントデータ** (RPM や RPKM ではない)
 - ・データ正規化手段は **DEGES** (RPM や TMM ではない)

比較解析の入力データの基本はカウントデータ!

■ 基本的なマッピングプログラム (bowtieなど) を用いた場合



発現変動解析用パッケージTCCの入力データは、リファレンス配列上にマップされたリード数(カウント数)からなるカウントデータ行列です

Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

今はLinuxコマンド抜きで一通り解析可能

- *SRadb* (Zhu et al., *BMC Bioinformatics*, **14**: 19, 2013)
 - 公共DBからのRNA-seqデータ(FASTQファイル)取得
- *QuasR* (Lerch et al., unpublished)
 - リファレンス配列(ゲノム or トランスクリプトーム)へのマッピング
 - Bowtie (Langmead et al., 2009) or SpliceMap (Au et al., 2010)を選択可能
 - 出力はBAM形式ファイル、QCレポートも
 - 遺伝子アノテーション情報をもとにカウントデータ取得
 - *GenomicFeatures* (Lawrenceら, 2013)で得られるTranscriptDbオブジェクトを利用
 - UCSC known genesやEnsembl genesのカウントデータなど
- *TCC* (Sun et al., *BMC Bioinformatics*, **14**: 219, 2013)
 - 内部的に*edgeR* (Robinson et al., 2010)や*DESeq* (Anders et al., 2010)などを用いて頑健な発現変動解析を実行

アセンブル以外ならWindows(のR)上でどうにかなる時代がやってきました

QuasRパッケージを用いてマッピング

- Basic alignerの1つであるbowtie (Langmead et al., 2009)を利用
 - マッピング時に多くのオプションを指定可能
 - “-v”:許容するミスマッチ数を指定するオプション。“-v 0”は、リードがリファレンスに完全一致するもののみレポート。“-v 2”は、2塩基ミスマッチまで許容してマップされうる場所を探索。
 - “-m”:出力するリード条件を指定するオプション。“-m 1”は、複数個所にマップされるリードを除外して、1か所にのみマップされたリードをレポート。“-m 3”は、合計3か所にマップされるリードまでをレポート。
 - “--best --strata”:最も少ないミスマッチ数でマップされるもののみ出力する、という意思表示。これをつけずに“-v 2 -m 1”などと指定すると、たとえ完全一致(ミスマッチ数0)で1か所にのみマップされるリードがあったとしても、どこか別の場所で1塩基ミスマッチでマップされる個所があれば、マップされうる場所が2か所ということを意味し、そのリードは出力されなくなる。それを防ぐのが主な目的
 - ...

デフォルトである程度よきに計らってくれるが...実際の挙動を完全に把握できる状況で様々なオプションを試したい

マッピング = (大量高速)文字列検索

- マップされる側のリファレンス配列: hoge4.fa
- マップする側のRNA-seqデータ(リードと呼ばれる): “AGG”

```
hoge4.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>contig_1
CGGACAGCTCCTCGGCATCCGGAT
>contig_2
GTCTGCCTCAAGCGCCCCAAGTGGGTTTGGAGGCCTAACATCGCAAGTCG
ACACTCAGTCCGGCCGTCTGGTTGGCAGGGGCAGAGACCCAGCACACCCT
GTC
>contig_3
TGTAGGAGAAGGGCGGTATCAGCGTCCACTTACACGATCCGTTACTAATT
GTATGAGGTCGGGCA
>contig_4
CGTGCTGATTCCACACAGCAGTAAACGCGGACCTCTACCTATGAACATG
```

出力ファイル

	start	end
contig_2	31	33
contig_2	77	79
contig_3	4	6
contig_3	10	12
contig_3	56	58

マッピングプログラムの出力: (どのリードが)リファレンス配列上のどの位置から転写されたものかという位置(座標)情報

(Rで)塩基配列解析(主にNGSやRNA-seq解析)

(last modified 2014/03/05, since 2010)

What's new?

- 参考資料(講義、講習会、本など)
- 私の所属するアグリバイオインフ...
行います。受講希望者は平成26...
出席してください。例年東大以外...
義は「ゲノム情報解析基礎」と「農...
目のみの受講も可能ですので、お
- 一連の解析パイプライン(RNA-se...
現変動解析およびM-A plot描画
- 2014年3月17-19日に九州大学に...
開催されます。私は3日目(3/19、1
- 発現変動解析用RパッケージTCC...
を利用したい方は、R (ver. 3.0.2)を

- はじめに (last modified 2014/01/30)
- 参考資料(講義、講習会、本など)
- 過去のお知らせ (last modified 2014/01/30)
- Rのインストールと起動 (last modified 2014/01/30)
- サンプルデータ (last modified 2014/01/30)
- イントロ | 一般 | ランダムに行を抽出 (last modified 2014/01/30)
- イントロ | 一般 | 任意の文字列を抽出 (last modified 2014/01/30)
- イントロ | 一般 | 任意のキーワード抽出 (last modified 2014/01/30)
- イントロ | 一般 | ランダムな塩基配列抽出 (last modified 2014/01/30)
- イントロ | 一般 | 任意の長さの塩基配列抽出 (last modified 2014/01/30)
- イントロ | 一般 | 任意の位置の塩基配列抽出 (last modified 2014/01/30)
- イントロ | 一般 | 指定した範囲の塩基配列抽出 (last modified 2014/01/30)
- イントロ | 一般 | 翻訳配列(translation) (last modified 2014/01/30)

赤矢印あたりの項目の例題を参考にすればマッピングやカウントデータ取得のイメージがつかめますが、トレーニングではやりません

- アセンブル | について (last modified 2011/07/26)
- アセンブル | ゲノム用 (last modified 2013/06/17)
- アセンブル | トランスクリプトーム(転写物)用 (last modified 2013/08/08)
- アセンブル | ゲノム既知で転写物構造推定用 (last modified 2014/02/04) **NEW**
- マッピング | 備忘録 | について (last modified 2013/10/25)
- マッピング | 備忘録 | basic aligner (last modified 2013/10/17)
- マッピング | 備忘録 | splice-aware aligner (last modified 2014/02/04) **NEW**
- マッピング | 備忘録 | Bisulfite sequencing用 (last modified 2014/02/04) **NEW**
- マッピング | 備忘録 | (ESTレベルの長さの)contig (last modified 2010/12/06)
- マッピング | 基礎 (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | QuasR(Lerch XXX) (last modified 2013/10/25)
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Lerch XXX) (last modified 2013/10/25)
- マッピング | single-end | ゲノム | splice-aware aligner | QuasR(Lerch XXX) (last modified 2013/10/25)
- マップ後 | について (last modified 2013/06/19)
- マップ後 | 出力ファイル形式について (last modified 2013/11/05)
- マップ後 | 出力ファイルの読み込み | BAM形式 (last modified 2013/09/30)
- マップ後 | 出力ファイルの読み込み | Bowtie形式 (last modified 2013/06/18)
- マップ後 | 出力ファイルの読み込み | SOAP形式 (last modified 2013/06/19)
- マップ後 | 出力ファイルの読み込み | htSeqTools (Planet 2012) (last modified 2013/06/19)
- マップ後 | カウント情報取得 | ゲノム | アノテーション有 | QuasR(Lerch XXX) (last modified 2013/11/06)
- マップ後 | カウント情報取得 | ゲノム | アノテーション無 | QuasR(Lerch XXX) (last modified 2014/02/12) **NEW**
- マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから (last modified 2013/10/13)
- マップ後 | 配列長とカウント数の関係 (last modified 2013/10/27)
- 正規化 | について (last modified 2013/06/21)
- 正規化 | 基礎 | RPK or CPK (配列長補正) (last modified 2013/07/03)
- 正規化 | 基礎 | RPM or CPM (総リード数補正) (last modified 2014/02/13) **NEW**

トップページへ

マッピング時に用いるオプションの理解

■ マップされる側のリファレンス配列: ref_genome.fa

```

ref_genome.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
  
```

chr3とchr5の違いは、2番目と7番目の塩基のみ。主に“-m”オプションの違いの把握が可能。

マッピング時に用いるオプションの理解

- マップする側のRNA-seqデータ: sample_RNAseq1.fa

```
ref_genome.fa - メモ帳
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATC
AGCATCTAGTCGCATCAGAAGGGTGTAGTC

sample_RNAseq1.fa - メモ帳
>chr1 11 45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCGCTTGCAGGAATCGTGTC
```

許容するミスマッチ数による違いや、マップされるべき場所が完全に把握できるように、リードのdescription行に記述されている

マッピングオプションと結果の解釈

- “-m 1 --best --strata -v 0”: 0 mismatches with 1 location only mapped reads output

```
ref_genome.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGA chr1 11 45 TGGG
>chr2
AGGGAGGGGGTCCAGTATC chr2 1 35
ACGCAGGTAGGCTGAGGAT chr2 16 50 GAGGAG
CTCGGGTATGGTTAGTCTT chr3 1 35 GGGCTG
TGACGCCCTG chr3 3 37 GGTTC
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

```
sample_RNAseq1.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCGCTTGCAGGAATCGTGTC
```

マップされなかったのは、
計8リード中3リード

マッピングオプションと結果の解釈

- “-m 1 --best --strata -v 0”: 0 mismatches with 1 location only mapped reads output

```
ref_genome.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

```
sample_RNAseq1.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATAACAGG
>chr3_3_37
GGGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCGCTTGCAGGAATCGTGTC
```

完全一致でも複数個所にマップされるために落とされたリード

マッピングオプションと結果の解釈

- “-m 1 --best --strata -v 0”: 0 mismatches with 1 location only mapped reads output

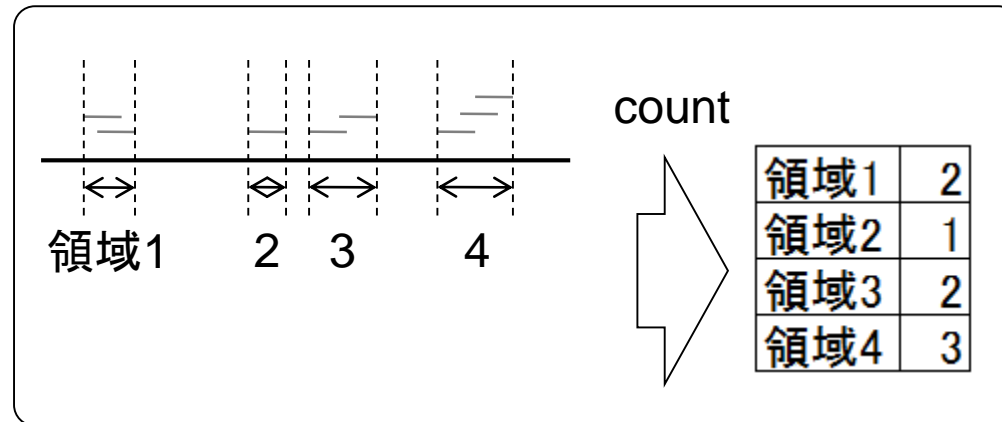
```
ref_genome.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

```
sample_RNAseq1.fa - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGACTATTTCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGACTATTTCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGGGGTCTATTTCCCGCTTGCAGGAATCGTGTC
```

1か所にのみマップされるが mismatches のため落とされたリード

マッピング結果からのカウント情報取得

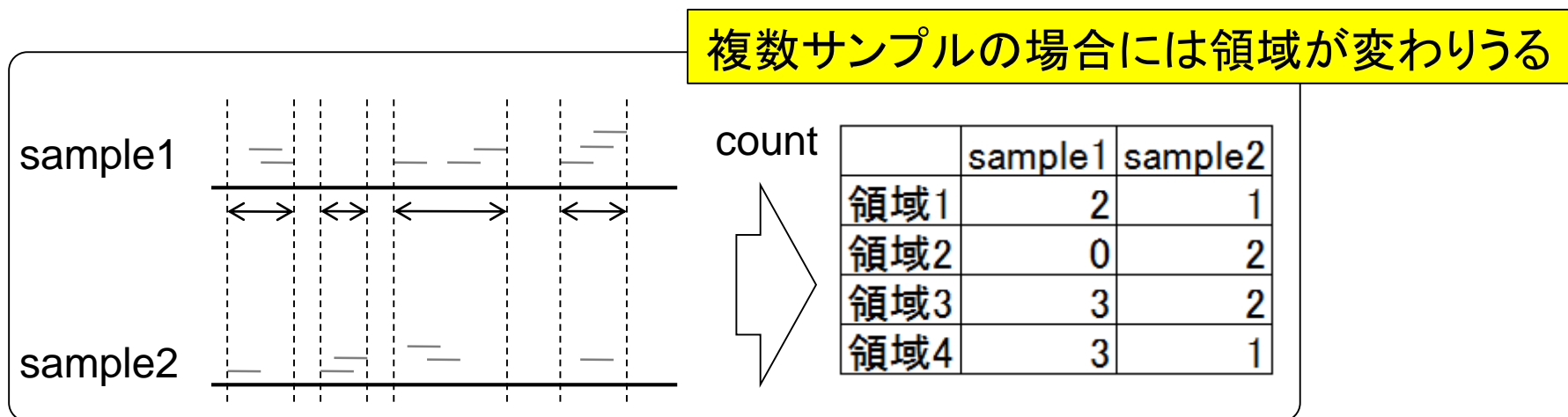
- アノテーション情報を利用する場合
 - UCSC Genes, Ensembl Genesなど様々なテーブル名を指定可能
 - gene, exon, promoter, junctionなど様々なレベルを指定可能
- アノテーション情報がない場合
 - マップされたリードの和集合領域を同定したのち、領域ごとのリード数をカウント
 - BEDtools (Quinlan et al., 2010)中のmergeBedプログラムを実行して和集合領域同定後、intersectBedプログラムを実行してリード数をカウントする作業に相当



基本的なイメージ

マッピング結果からのカウント情報取得

- アノテーション情報を利用する場合
 - UCSC Genes, Ensembl Genesなど様々なテーブル名を指定可能
 - gene, exon, promoter, junctionなど様々なレベルを指定可能
- アノテーション情報がない場合
 - マップされたリードの和集合領域を同定したのち、領域ごとのリード数をカウント
 - BEDtools (Quinlan et al., 2010)中のmergeBedプログラムを実行して和集合領域同定後、intersectBedプログラムを実行してリード数をカウントする作業に相当



(Rで)塩基配列解析(主にNGSやRNA-seq解析)

(last modified 2014/03/05, since 2010)

What's new?

- [参考資料\(講義、講習会、本など\)](#)
- 私の所属するアグリバイオインフ...
行います。受講希望者は平成26...
出席してください。例年東大以外...
義は「[ゲノム情報解析基礎](#)」と「[農](#)...
目のみの受講も可能ですので、お
- 一連の解析パイプライン(RNA-se...
現変動解析およびM-A plot描画
- 2014年3月17-19日に九州大学に...
開催されます。私は3日目(3/19、1
- 発現変動解析用Rパッケージ[TCC](#)...
を利用したい方は、R (ver. 3.0.2)を

- [はじめに](#) (last modified 2014/01/30)
- [参考資料\(講義、講習会、本など\)](#)
- [過去のお知らせ](#) (last modified 2014/03/05)
- [Rのインストールと起動](#) (last modified 2014/03/05)
- [サンプルデータ](#) (last modified 2014/03/05)
- インポート | 一般 | [ランダムに行を抽出](#)
- インポート | 一般 | [任意の文字列を抽出](#)
- インポート | 一般 | [任意のキーワード](#)
- インポート | 一般 | [ランダムな塩基配列](#)
- インポート | 一般 | [任意の長さの可能](#)
- インポート | 一般 | [任意の位置の塩基](#)
- インポート | 一般 | [指定した範囲の配列](#)
- インポート | 一般 | [翻訳配列\(translation\)](#)

- [アセンブル | について](#) (last modified 2011/07/26)
- [アセンブル | ゲノム用](#) (last modified 2013/06/17)
- [アセンブル | トランスクリプトーム\(転写物\)用](#) (last modified 2013/08/08)
- [アセンブル | ゲノム既知で転写物構造推定用](#) (last modified 2014/02/04) **NEW**
- [マッピング | 備忘録 | について](#) (last modified 2013/10/25)
- [マッピング | 備忘録 | basic aligner](#) (last modified 2013/10/17)
- [マッピング | 備忘録 | splice-aware aligner](#) (last modified 2014/02/04) **NEW**
- [マッピング | 備忘録 | Bisulfite sequencing用](#) (last modified 2014/02/04) **NEW**
- [マッピング | 備忘録 | \(ESTレベルの長さの\)contig](#) (last modified 2010/12/06)
- [マッピング | 基礎](#) (last modified 2013/06/19)
- [マッピング | single-end | ゲノム | basic aligner\(基礎\)](#) | [QuasR\(Lerch XXX\)](#) (last modified 2013/10/25)
- [マッピング | single-end | ゲノム | basic aligner\(応用\)](#) | [QuasR\(Lerch XXX\)](#) (last modified 2013/10/25)
- [マッピング | single-end | ゲノム | splice-aware aligner](#) | [QuasR\(Lerch XXX\)](#) (last modified 2013/10/25)
- [マップ後 | について](#) (last modified 2013/06/19)
- [マップ後 | 出力ファイル形式について](#) (last modified 2013/11/05)
- [マップ後 | 出力ファイルの読み込み | BAM形式](#) (last modified 2013/09/30)
- [マップ後 | 出力ファイルの読み込み | Bowtie形式](#) (last modified 2013/06/18)
- [マップ後 | 出力ファイルの読み込み | SOAP形式](#) (last modified 2013/06/19)
- [マップ後 | 出力ファイルの読み込み | htSeqTools \(Planet 2012\)](#) (last modified 2013/06/19)
- [マップ後 | カウント情報取得 | ゲノム | アノテーション有](#) | [QuasR\(Lerch XXX\)](#) (last modified 2013/11/06)
- [マップ後 | カウント情報取得 | ゲノム | アノテーション無](#) | [QuasR\(Lerch XXX\)](#) (last modified 2014/02/12) **NEW**
- [マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから](#) (last modified 2013/10/13)
- [マップ後 | 配列長とカウント数の関係](#) (last modified 2013/10/27)
- [正規化 | について](#) (last modified 2013/06/21)
- [正規化 | 基礎 | RPK or CPK \(配列長補正\)](#) (last modified 2013/07/03)
- [正規化 | 基礎 | RPM or CPM \(総リード数補正\)](#) (last modified 2014/02/13) **NEW**

[トップページへ](#)

マッピング結果からのカウント情報取得

マップ後 | カウント情報取得 | ゲノム | アノテーション無 | QuasR(Lerch_XXX)

*.bed

chr1	11	45
chr2	1	35
chr2	16	50
chr3	1	35
chr3	3	37

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウント取得までの一連の流れを示します。アノテーション情報がない場合を想定しているため、GenomicRangesパッケージで、マップされたリードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。RNA-seqデータのほうのリード数は少ないですが、リファレンス配列の前処理でかなり時間がかかるようです(2時間「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18-20の複数のRNA-seqデータ(sample RNAseq1.fa)をref_genome.faにマッピングする場合(mapping_single_genome1.txt):

複数サンプルのマッピング結果をまとめて和集合領域を定め、カウント情報を得るやり方です。サンプル間比較の

```
in_f1 <- "mapping_single_genome1.txt" #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa" #入力ファイル名を指定してin_f2に格納(リファレンスゲノム)
param_mapping <- "-m 1 --best --strata -v 0" #マッピング時のオプションを指定
```

*_range.txt

seqnames	start	end	width	strand	name
chr1	11	45	35	+	1
chr2	1	50	50	+	2
chr3	1	37	37	+	2

	A	B
1	FileName	SampleName
2	sample_RNAseq1.fa	name

カウント数はこちら

マッピング結果からのカウント情報取得

5. サンプルデータ18-20の複数のRNA-seqデータ(sample_RNAseq1.faとsample_RNAseq2.fa)をref_genome.faにマッピングする場合(mapping_single_genome4.txt):

全部のマッピング結果をまとめて和集合領域を定め、カウント情報を得るやり方です。一般的なカウントデータ行列の形式(2列目以降がカウント情報)にし、配列長情報と別々のファイルにして保存するやり方です。

```
in_f1 <- "mapping_single_genome4.txt" #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa" #入力ファイル名を指定してin_f2に格納(リファレンス配列)
out_f1 <- "hoge4_count.txt" #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge4_gene.length.txt" #出力ファイル名を指定してout_f2に格納
param_mapping <- "-m 1 --best --strata -v 1" #マッピング時のオプションを指定
```

FileName	SampleName
sample_RNAseq1.fa	sample1
sample_RNAseq2.fa	sample2

tmp	sample1	sample2
chr1_11_45_35_+	1	0
chr2_1_60_60_+	2	1
chr3_1_37_37_+	2	0
chr4_6_65_60_+	0	1
chr5_1_35_35_+	1	0

リストファイル中で指定したサンプル名がカウントデータ行列の列名となる

これがカウントデータ

よく見かけるカウントデータ取得手段

- 基本的なマッピングプログラム (Bowtieなど) を利用
- 最大2塩基ミスマッチまで許容してリファレンス配列の1か所とのみ一致するリード (uniquely mapped reads or unique mapper) 数をカウント
 - Marioni et al., *Genome Res.*, **18**:1509-1517, 2008
 - Bullard et al., *BMC Bioinformatics*, **11**:94, 2010
 - Risso et al., *BMC Bioinformatics*, **12**:480, 2011
 - ReCount (Frazee et al., *BMC Bioinformatics*, **12**:449, 2011)
 - ...

“-m 1 --best --strata -v 2”: 2塩基ミスマッチまで許容して1か所にのみマップされるリード

TCCなどの発現変動解析用パッケージの入力データは、基本的に曖昧なものを使わずに何の補正もかけていないカウントデータ

Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- **実データ解析例: 結果の解釈やM-A plotの見方など**
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

実データ解析例: SRP017142

■ SRADBを用いたgzip圧縮FASTQ形式ファイルのダウンロード

□ Neyret-Kahn et al., *Genome Res.*, **23**: 1563-1579, 2013

- 複製あり2群間比較用ヒトRNA-seqデータ(3 Ras vs. 3 Proliferative)

FileName	SampleName
SRR616151.fastq.gz	Pro_rep1
SRR616152.fastq.gz	Pro_rep2
SRR616153.fastq.gz	Pro_rep3
SRR616154.fastq.gz	Ras_rep1
SRR616155.fastq.gz	Ras_rep2
SRR616156.fastq.gz	Ras_rep3

G1群

G2群

計6GB程度。QuasRパッケージは圧縮ファイルのままでもマッピング可能

■ QuasR (Bowtie)を用いたヒトゲノムへのマッピング

□ *BSgenome.Hsapiens.UCSC.hg19*パッケージを利用

□ 18種類程度の生物種のゲノム配列がRパッケージとして利用可能

- シロイヌナズナの場合: *BSgenome.Athaliana.TAIR.TAIR9*

- ショウジョウバエの場合: *BSgenome.Dmelanogaster.UCSC.dm3*

実データ解析例: SRP017142

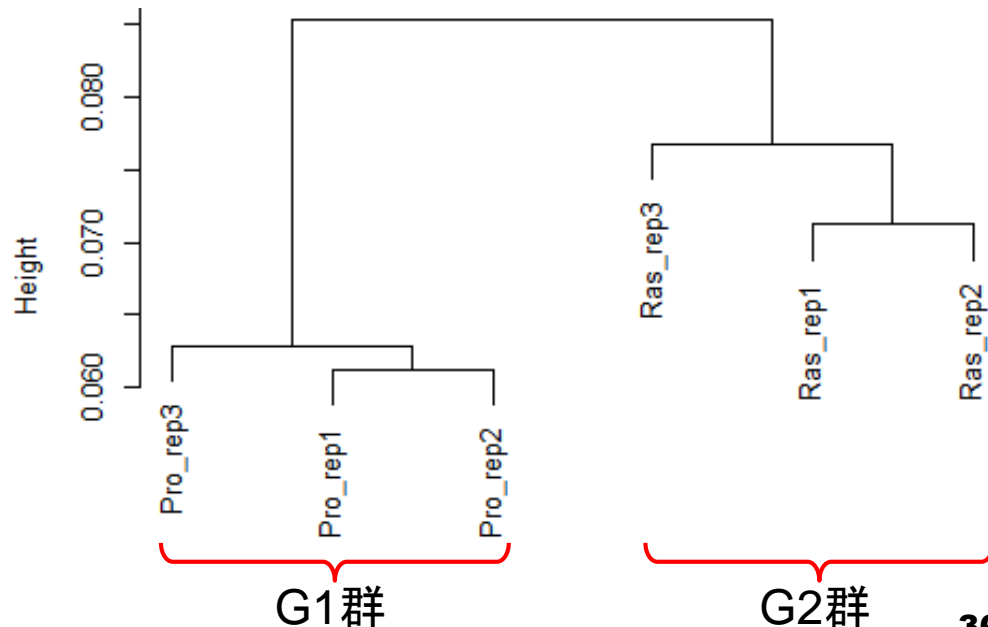
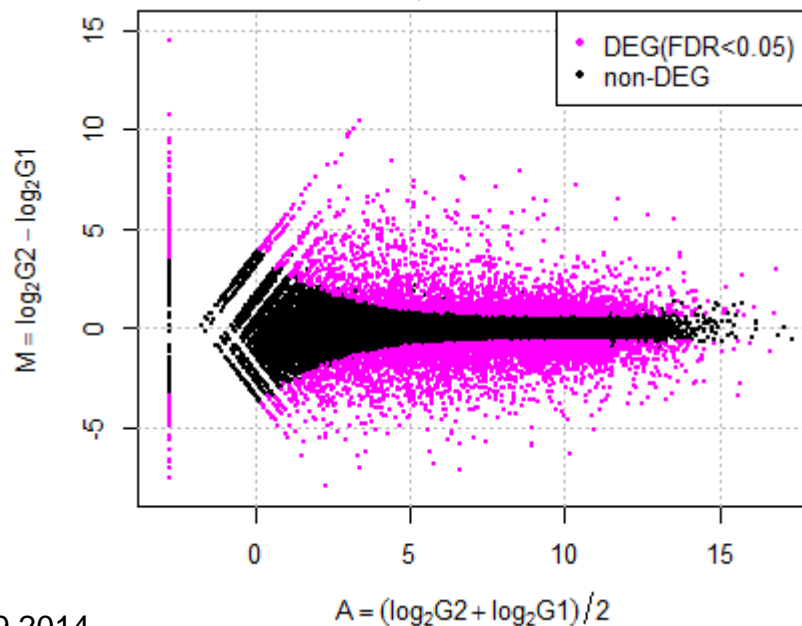
- QuasR (Bowtie)を用いたカウント情報取得
カウントデータ

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

59,857 genes

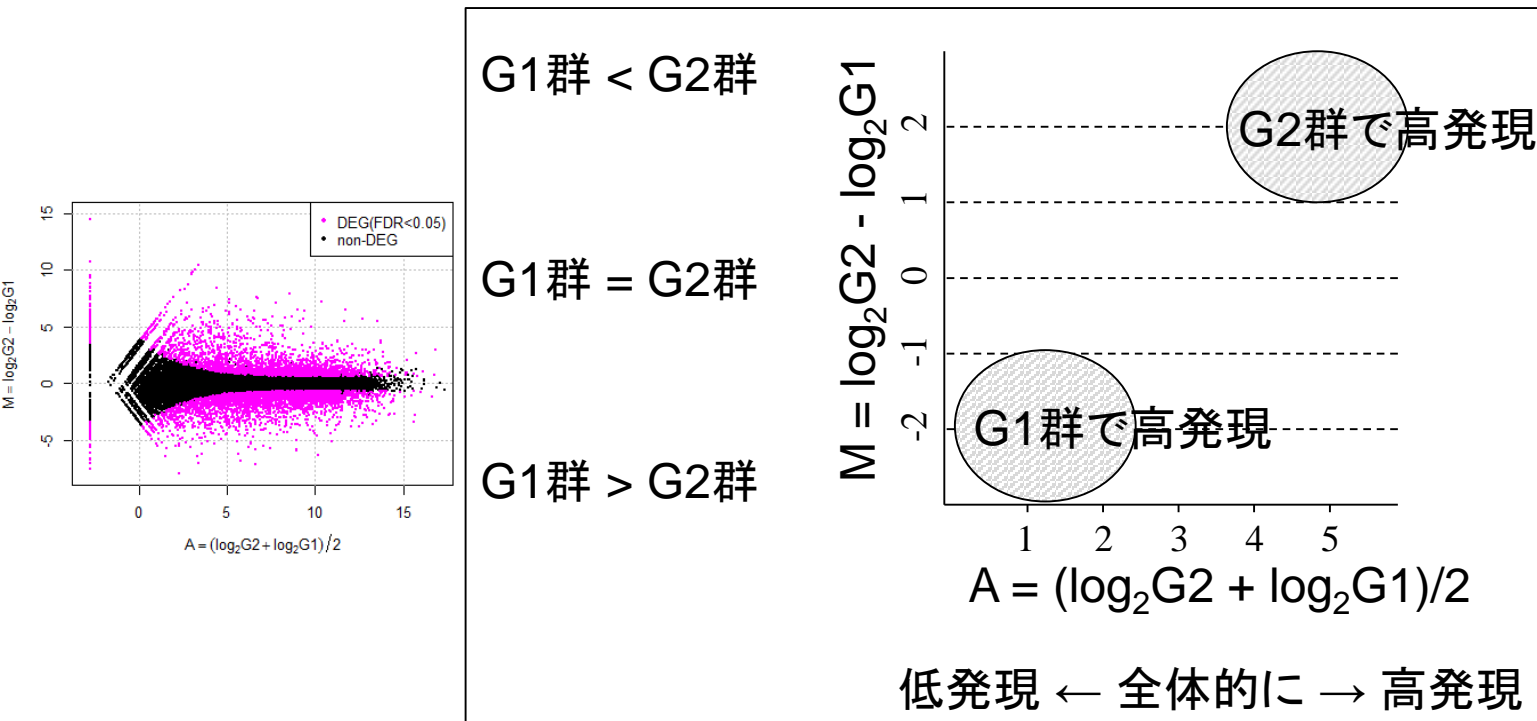
TCCを用いた発現変動遺伝子 (DEG) 同定

サンプル間クラスタリング



M-A plot

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



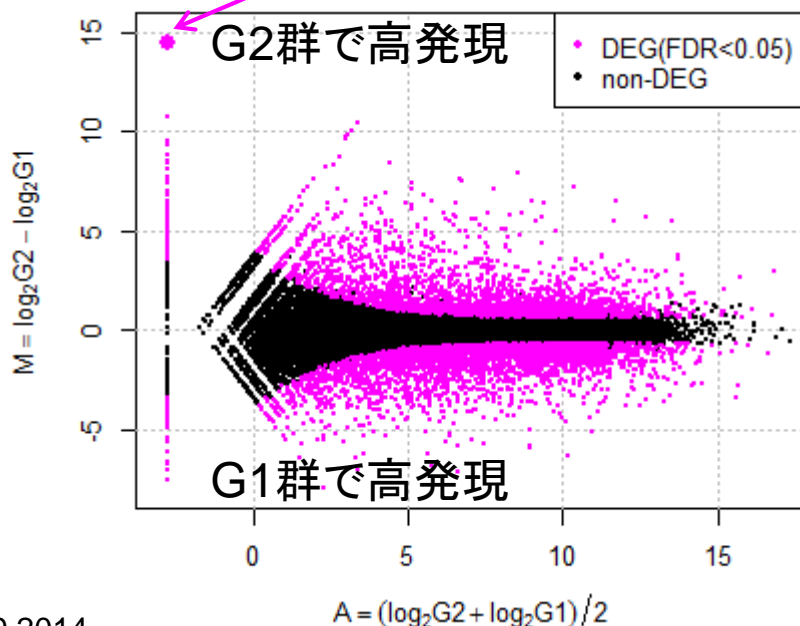
DEGが存在しないデータのM-A plotを眺めることで、縦軸の閾値のみに相当する倍率変化を用いたDEG同定の危険性が分かります

実データ解析例: SRP017142

TCCを用いたDEG同定結果ファイル

p-valueとその順位

rownames(tcc\$count)	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000240386	0.0	0.0	0.0	5608.1	5097.7	8188.0	ENSG00000240386	-2.78	14.48	1.75E-139	1.04E-134	1	1
ENSG00000128564	15.4	22.3	19.1	2856.6	2462.6	2555.3	ENSG00000128564	7.80	7.11	4.03E-107	1.21E-102	2	1
ENSG00000188064	7.7	5.8	10.1	1425.5	1254.0	1486.8	ENSG00000188064	6.71	7.47	2.67E-98	5.33E-94	3	1
ENSG00000101188	6.0	5.0	11.1	1477.4	1407.0	1254.9	ENSG00000101188	6.65	7.55	3.81E-97	5.70E-93	4	1
ENSG00000163431	3716.6	3244.5	4185.2	70.1	78.8	49.0	ENSG00000163431	8.95	-5.82	2.90E-80	3.48E-76	5	1
ENSG00000204291	1215.5	1236.8	1339.3	22.4	27.8	21.5	ENSG00000204291	7.44	-5.72	3.45E-78	3.44E-74	6	1
ENSG00000181634	107.0	158.6	66.5	12842.0	16014.1	19820.5	ENSG00000181634	10.39	7.20	1.04E-76	8.88E-73	7	1
ENSG00000178726	53.1	46.3	62.5	6006.1	5567.6	3166.0	ENSG00000178726	9.01	6.51	2.39E-74	1.79E-70	8	1
ENSG00000117600	576.9	518.9	602.6	7.0	5.6	2.4	ENSG00000117600	5.73	-6.83	1.14E-72	7.59E-69	9	1
ENSG00000158050	7.7	9.1	11.1	552.3	536.6	523.5	ENSG00000158050	6.14	5.85	1.16E-70	6.91E-67	10	1
ENSG00000124126	50.5	44.6	53.4	1819.4	1683.2	1413.9	ENSG00000124126	8.15	5.05	5.35E-69	2.91E-65	11	1



M-A plotのA値とM値

q-value

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

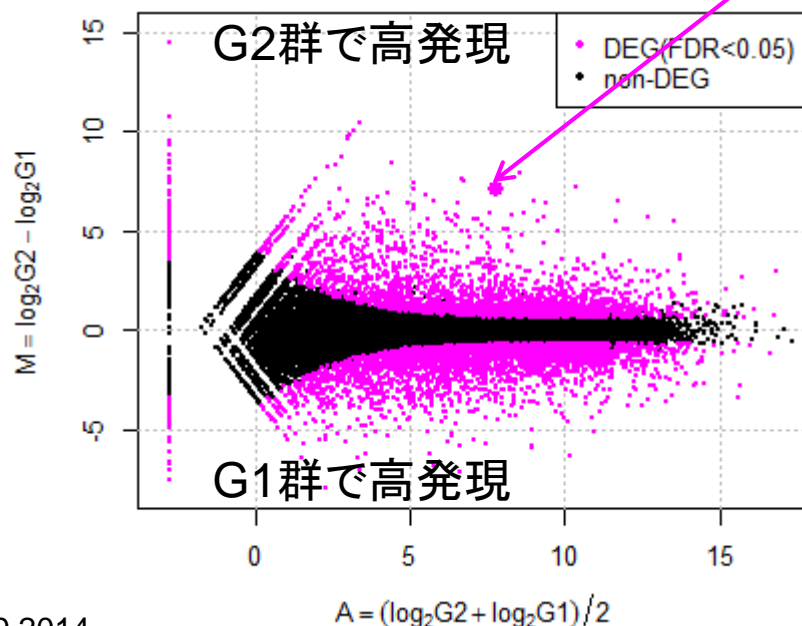
基本的には、これらが解析結果です
1位はRas群(G2群)で高発現のDEG

実データ解析例: SRP017142

TCCを用いたDEG同定結果ファイル

p-valueとその順位

rownames(tcc\$count)	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000240386	0.0	0.0	0.0	5608.1	5097.7	8188.0	ENSG00000240386	-2.78	14.48	1.75E-139	1.04E-134	1	1
ENSG00000128564	15.4	22.3	19.1	2856.6	2462.6	2555.3	ENSG00000128564	7.80	7.11	4.03E-107	1.21E-102	2	1
ENSG00000188064	7.7	5.8	10.1	1425.5	1254.0	1486.8	ENSG00000188064	6.71	7.47	2.67E-98	5.33E-94	3	1
ENSG00000101188	6.0	5.0	11.1	1477.4	1407.0	1254.9	ENSG00000101188	6.65	7.55	3.81E-97	5.70E-93	4	1
ENSG00000163431	3716.6	3244.5	4185.2	70.1	78.8	49.0	ENSG00000163431	8.95	-5.82	2.90E-80	3.48E-76	5	1
ENSG00000204291	1215.5	1236.8	1339.3	22.4	27.8	21.5	ENSG00000204291	7.44	-5.72	3.45E-78	3.44E-74	6	1
ENSG00000181634	107.0	158.6	66.5	12842.0	16014.1	19820.5	ENSG00000181634	10.39	7.20	1.04E-76	8.88E-73	7	1
ENSG00000178726	53.1	46.3	62.5	6006.1	5567.6	3166.0	ENSG00000178726	9.01	6.51	2.39E-74	1.79E-70	8	1
ENSG00000117600	576.9	518.9	602.6	7.0	5.6	2.4	ENSG00000117600	5.73	-6.83	1.14E-72	7.59E-69	9	1
ENSG00000158050	7.7	9.1	11.1	552.3	536.6	523.5	ENSG00000158050	6.14	5.85	1.16E-70	6.91E-67	10	1
ENSG00000124126	50.5	44.6	53.4	1819.4	1683.0	1413.9	ENSG00000124126	8.15	5.05	5.35E-69	2.91E-65	11	1



M-A plotのA値とM値

q-value

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

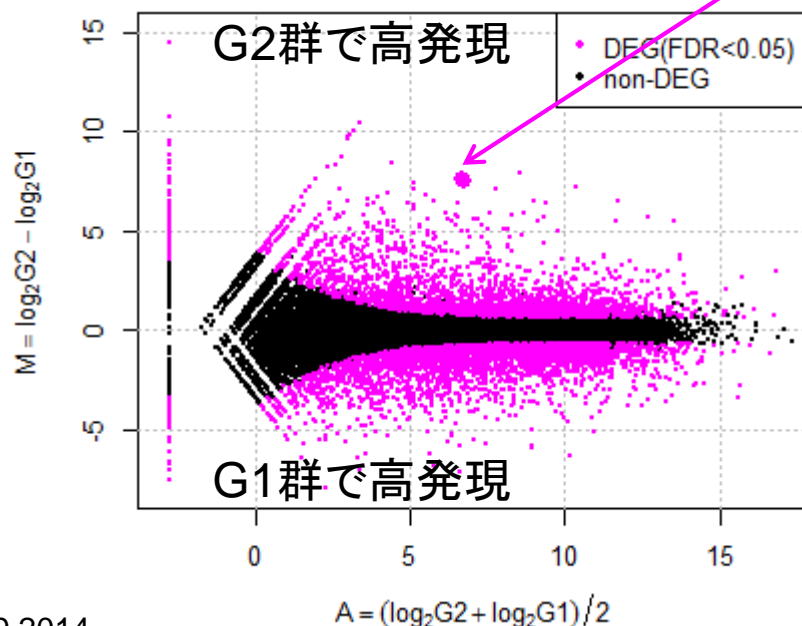
2位もRas群(G2群)で高発現のDEG

実データ解析例: SRP017142

TCCを用いたDEG同定結果ファイル

p-valueとその順位

rownames(tcc\$coun	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000240386	0.0	0.0	0.0	5608.1	5097.7	8188.0	ENSG00000240386	-2.78	14.48	1.75E-139	1.04E-134	1	1
ENSG00000128564	15.4	22.3	19.1	2856.6	2462.6	2555.3	ENSG00000128564	7.80	7.11	4.03E-107	1.21E-102	2	1
ENSG00000188064	7.7	5.8	10.1	1425.5	1254.0	1486.8	ENSG00000188064	6.71	7.47	2.67E-98	5.33E-94	3	1
ENSG00000101188	6.0	5.0	11.1	1477.4	1407.0	1254.9	ENSG00000101188	6.65	7.55	3.81E-97	5.70E-93	4	1
ENSG00000163431	3716.6	3244.5	4185.2	70.1	78.8	49.0	ENSG00000163431	8.95	-5.82	2.90E-80	3.48E-76	5	1
ENSG00000204291	1215.5	1236.8	1339.3	22.4	27.8	21.5	ENSG00000204291	7.44	-5.72	3.45E-78	3.44E-74	6	1
ENSG00000181634	107.0	158.6	66.5	12842.0	16014.1	19820.5	ENSG00000181634	10.39	7.20	1.04E-76	8.88E-73	7	1
ENSG00000178726	53.1	46.3	62.5	6006.1	5567.6	3166.0	ENSG00000178726	9.01	6.51	2.39E-74	1.79E-70	8	1
ENSG00000117600	576.9	518.9	602.6	7.0	5.6	2.4	ENSG00000117600	5.73	-6.83	1.14E-72	7.59E-69	9	1
ENSG00000158050	7.7	9.1	11.1	552.3	536.6	523.5	ENSG00000158050	6.14	5.85	1.16E-70	6.91E-67	10	1
ENSG00000124126	50.5	44.6	53.4	1819.4	1682.2	1413.9	ENSG00000124126	8.15	5.05	5.35E-69	2.91E-65	11	1



M-A plotのA値とM値

q-value

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

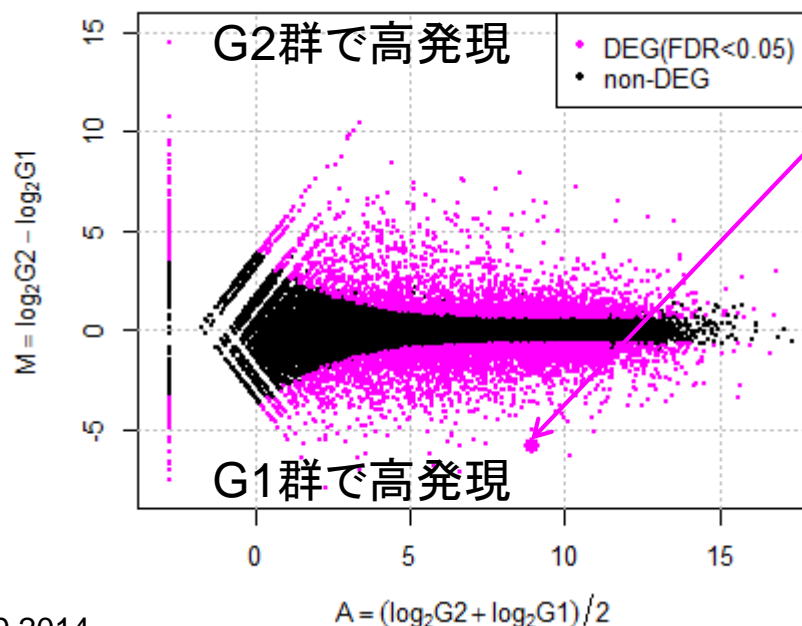
3,4位もRas群(G2群)で高発現のDEG

実データ解析例: SRP017142

TCCを用いたDEG同定結果ファイル

p-valueとその順位

rownames(tcc\$coun	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000240386	0.0	0.0	0.0	5608.1	5097.7	8188.0	ENSG00000240386	-2.78	14.48	1.75E-139	1.04E-134	1	1
ENSG00000128564	15.4	22.3	19.1	2856.6	2462.6	2555.3	ENSG00000128564	7.80	7.11	4.03E-107	1.21E-102	2	1
ENSG00000188064	7.7	5.8	10.1	1425.5	1254.0	1486.8	ENSG00000188064	6.71	7.47	2.67E-98	5.33E-94	3	1
ENSG00000101188	6.0	5.0	11.1	1477.4	1407.0	1254.9	ENSG00000101188	6.65	7.55	3.81E-97	5.70E-93	4	1
ENSG00000163431	3716.6	3244.5	4185.2	70.1	78.8	49.0	ENSG00000163431	8.95	-5.82	2.90E-80	3.48E-76	5	1
ENSG00000204291	1215.5	1236.8	1339.3	22.4	27.8	21.5	ENSG00000204291	7.44	-5.72	3.45E-78	3.44E-74	6	1
ENSG00000181634	107.0	158.6	66.5	12842.0	16014.1	19820.5	ENSG00000181634	10.39	7.20	1.04E-76	8.88E-73	7	1
ENSG00000178726	53.1	46.3	62.5	6006.1	5567.6	3166.0	ENSG00000178726	9.01	6.51	2.39E-74	1.79E-70	8	1
ENSG00000117600	576.9	518.9	602.6	7.0	5.6	2.4	ENSG00000117600	5.73	-6.83	1.14E-72	7.59E-69	9	1
ENSG00000158050	7.7	9.1	11.1	552.3	536.6	523.5	ENSG00000158050	6.14	5.85	1.16E-70	6.91E-67	10	1
ENSG00000124126	50.5	44.6	53.4	1819.4	1683.2	1413.9	ENSG00000124126	8.15	5.05	5.35E-69	2.91E-65	11	1



M-A plotのA値とM値

q-value

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

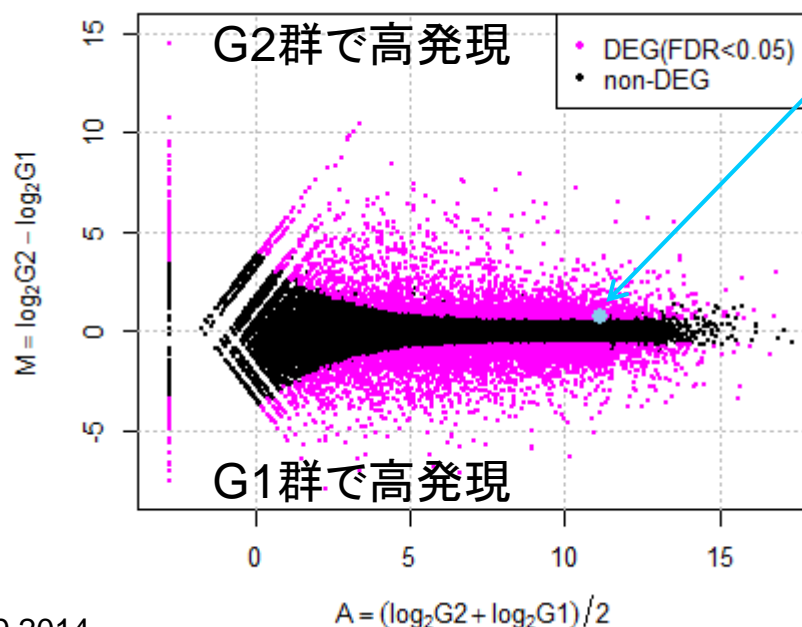
5位はPro群(G1群)で高発現のDEG

実データ解析例: SRP017142

TCCを用いたDEG同定結果ファイル

p-valueとその順位

rownames(tcc\$count)	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000148848	286.8	327.2	262.0	486.4	475.5	419.5	ENSG00000148848	8.52	0.66	0.004726	0.049922	5666	1
ENSG00000186603	16.3	13.2	10.1	23.8	16.7	69.3	ENSG00000186603	4.46	1.47	0.004727	0.049927	5667	1
ENSG00000168556	161.8	142.9	146.1	218.7	257.7	236.6	ENSG00000168556	7.56	0.66	0.004729	0.049936	5668	1
ENSG00000189159	1794.1	1668.1	1774.6	2377.2	2307.9	4183.1	ENSG00000189159	11.15	0.76	0.004731	0.049954	5669	1
ENSG00000177096	621.4	575.0	600.6	322.4	317.0	468.5	ENSG00000177096	8.88	-0.70	0.004739	0.050031	5670	0
ENSG00000103148	1707.7	1452.5	1820.0	2347.8	2142.0	4082.7	ENSG00000103148	11.09	0.78	0.004746	0.050088	5671	0
ENSG00000156011	918.5	1103.8	882.8	605.5	685.9	271.3	ENSG00000156011	9.47	-0.89	0.00475	0.050127	5672	0
ENSG00000089818	472.5	544.5	478.7	685.4	845.3	815.1	ENSG00000089818	9.29	0.65	0.004751	0.050127	5673	0
ENSG00000160007	4551.2	4256.6	4650.7	3080.9	3115.1	1459.3	ENSG00000160007	11.72	-0.81	0.004752	0.05013	5674	0
ENSG00000105778	900.5	1027.8	984.6	1529.2	1904.7	1239.4	ENSG00000105778	10.26	0.68	0.004765	0.050255	5675	0
ENSG00000246451	46.2	91.7	73.6	46.3	33.4	29.9	ENSG00000246451	5.66	-0.95	0.004771	0.050317	5676	0



M-A plotのA値とM値

q-value

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

5,669位の指定したFDR閾値
(0.05)をギリギリ満たす遺伝子

Contents



■ セミナー(13:00-14:00)

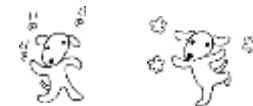
- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- **多重比較問題: FDRって何?**
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

多重比較問題：FDRって何？

- **DEG**かnon-DEGかを判定する閾値を決める問題
 - 有意水準5%というのが $p\text{-value} < 0.05$ に相当
 - False discovery rate (FDR) 5%というのが $q\text{-value} < 0.05$ に相当
- 発現変動ランキング結果は不変なので上位 x 個という決め打ちの場合にはこの問題とは無関係



rownames(tcc\$coun	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000148848	286.8	327.2	262.0	486.4	475.5	419.5	ENSG00000148848	8.52	0.66	0.004726	0.049922	5666	1
ENSG00000186603	16.3	13.2	10.1	23.8	16.7	69.3	ENSG00000186603	4.46	1.47	0.004727	0.049927	5667	1
ENSG00000168556	161.8	142.9	146.1	218.7	257.7	236.6	ENSG00000168556	7.56	0.66	0.004729	0.049936	5668	1
ENSG00000189159	1794.1	1668.1	1774.6	2377.2	2307.9	4183.1	ENSG00000189159	11.15	0.76	0.004731	0.049954	5669	1
ENSG00000177096	621.4	575.0	600.6	322.4	317.0	468.5	ENSG00000177096	8.88	-0.70	0.004739	0.050031	5670	0
ENSG00000103148	1707.7	1452.5	1820.0	2347.8	2142.0	4082.7	ENSG00000103148	11.09	0.78	0.004746	0.050088	5671	0
ENSG00000156011	918.5	1103.8	882.8	605.5	685.9	271.3	ENSG00000156011	9.47	-0.89	0.00475	0.050127	5672	0
ENSG00000089818	472.5	544.5	478.7	685.4	845.3	815.1	ENSG00000089818	9.29	0.65	0.004751	0.050127	5673	0
ENSG00000160007	4551.2	4256.6	4650.7	3080.9	3115.1	1459.3	ENSG00000160007	11.72	-0.81	0.004752	0.05013	5674	0
ENSG00000105778	900.5	1027.8	984.6	1529.2	1904.7	1239.4	ENSG00000105778	10.26	0.68	0.004765	0.050255	5675	0
ENSG00000246451	46.2	91.7	73.6	46.3	33.4	29.9	ENSG00000246451	5.66	-0.95	0.004771	0.050317	5676	0

DEG数に関するよりよい結果を得たい場合には、 $p\text{-value}$ ではなく $q\text{-value}$ 閾値を利用しましょう



多重比較問題：FDRって何？

■ p -value (false positive rate; FPR)

- 本当はDEGではないにもかかわらずDEGと判定してしまう確率
- 全遺伝子に占めるnon-DEGの割合(分母は**遺伝子総数**)
- 例：10,000個のnon-DEGからなる遺伝子を p -value < 0.05で検定すると、
 $10,000 \times 0.05 = 500$ 個程度のnon-DEGを間違っ**て**DEGと判定することに相当
 - 実際のDEG検出結果が900個だった場合：500個は偽物で400個は本物と判断
 - 実際のDEG検出結果が510個だった場合：500個は偽物で10個は本物と判断
 - 実際のDEG検出結果が500個以下の場合：全て偽物と判断

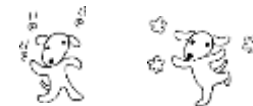
■ q -value (false discovery rate: FDR)

- DEGと判定した中に含まれるnon-DEGの割合
- DEG中に占めるnon-DEGの割合(分母は**DEGと判定された数**)
- non-DEGの期待値を計算できれば、 p 値でも上位 x 個でもDEGと判定する手段はなんでもよい。以下は10,000遺伝子の検定結果でのFDR計算例
 - $p < 0.001$ を満たすDEG数が100個の場合：FDR = $10,000 \times 0.001 / 100 = 0.1$
 - $p < 0.01$ を満たすDEG数が400個の場合：FDR = $10,000 \times 0.01 / 400 = 0.25$
 - $p < 0.05$ を満たすDEG数が926個の場合：FDR = $10,000 \times 0.05 / 926 = 0.54$



多重比較問題：FDRって何？

- **DEG**かnon-DEGかを判定する閾値を決める問題
 - 有意水準5%というのが $p\text{-value} < 0.05$ に相当
 - False discovery rate (FDR) 5%というのが $q\text{-value} < 0.05$ に相当
- 発現変動ランキング結果は不変なので上位 x 個という決め打ちの場合にはこの問題とは無関係



rownames(tcc\$count)	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000148848	286.8	327.2	262.0	486.4	475.5	419.5	ENSG00000148848	8.52	0.66	0.004726	0.049922	5666	1
ENSG00000186603	16.3	13.2	10.1	23.8	16.7	69.3	ENSG00000186603	4.46	1.47	0.004727	0.049927	5667	1
ENSG00000168556	161.8	142.9	146.1	218.7	257.7	236.6	ENSG00000168556	7.56	0.66	0.004729	0.049936	5668	1
ENSG00000189159	1794.1	1668.1	1774.6	2377.2	2307.9	4183.1	ENSG00000189159	11.15	0.76	0.004731	0.049954	5669	1
ENSG00000177096	621.4	575.0	600.6	322.4	317.0	468.5	ENSG00000177096	8.88	-0.70	0.004739	0.050031	5670	0
ENSG00000103148	1707.7	1452.5	1820.0	2347.8	2142.0	4082.7	ENSG00000103148	11.09	0.78	0.004746	0.050088	5671	0
ENSG00000156011	918.5	1103.8	882.8	605.5	685.9	271.3	ENSG00000156011	9.47	-0.89	0.00475	0.050127	5672	0
ENSG00000089818	472.5	544.5	478.7	685.4	845.3	815.1	ENSG00000089818	9.29	0.65	0.004751	0.050127	5673	0
ENSG00000160007	4551.2	4256.6	4650.7	3080.9	3115.1	1459.3	ENSG00000160007	11.72	-0.81	0.004752	0.05013	5674	0
ENSG00000105778	900.5	1027.8	984.6	1529.2	1904.7	1239.4	ENSG00000105778	10.26	0.68	0.004765	0.050255	5675	0
ENSG00000246451	46.2	91.7	73.6	46.3	33.4	29.9	ENSG00000246451	5.66	-0.95	0.004771	0.050317	5676	0

5%の偽物(本当はnon-DEGだがDEGと判定してしまう誤り)を許容すると5,669遺伝子がDEGとみなせます。
 → $5,669 \times 0.05 = 283.45$ 個が理論上偽物だということ



多重比較問題：FDRって何？

- **DEG**かnon-DEGかを判定する閾値を決める問題
 - 有意水準5%というのが $p\text{-value} < 0.05$ に相当
 - False discovery rate (FDR) **1%**というのが $q\text{-value} < 0.01$ に相当
- 発現変動ランキング結果は不変なので上位 x 個という決め打ちの場合にはこの問題とは無関係



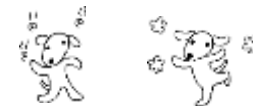
rownames(tcc\$coun	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEC
ENSG00000106211	11551.4	9071.7	13133.9	5924.8	5090.3	7692.0	ENSG00000106211	13.03	-0.85	0.000695	0.009934	4185	1
ENSG00000266714	21.4	30.6	22.2	14.0	9.3	3.6	ENSG00000266714	3.90	-1.46	0.000696	0.00995	4186	1
ENSG00000205002	1.7	3.3	0.0	7.0	14.8	6.0	ENSG00000205002	1.98	2.47	0.000696	0.009956	4187	1
ENSG00000272198	28.2	28.9	30.2	12.6	11.1	13.1	ENSG00000272198	4.24	-1.24	0.000698	0.009973	4188	1
ENSG00000116745	5.1	5.0	8.1	19.6	17.6	15.5	ENSG00000116745	3.37	1.54	0.000698	0.009975	4189	1
ENSG00000123395	2071.5	1531.8	2072.9	3076.7	2969.6	3983.5	ENSG00000123395	11.30	0.82	0.0007	0.010002	4190	1
ENSG00000100867	26.5	19.0	20.2	8.4	5.6	9.6	ENSG00000100867	3.71	-1.48	0.0007	0.010002	4191	1
ENSG00000171861	321.8	271.8	310.4	486.4	472.7	633.4	ENSG00000171861	8.64	0.82	0.000703	0.010036	4192	1
ENSG00000178972	27.4	32.2	30.2	9.8	18.5	6.0	ENSG00000178972	4.21	-1.39	0.000705	0.010066	4193	1
ENSG00000160013	297.9	264.4	302.3	510.2	456.0	512.7	ENSG00000160013	8.56	0.77	0.000706	0.010077	4194	1
ENSG00000091622	671.9	651.0	861.6	426.1	381.9	135.1	ENSG00000091622	8.90	-1.21	0.000707	0.010094	4195	1

1%の偽物(本当はnon-DEGだが**DEG**と判定してしまう誤り)を許容すると4,189遺伝子が**DEG**とみなせます。
 → $4189 \times 0.01 = 41.89$ 個が理論上偽物だということ



多重比較問題：FDRって何？

- **DEG**かnon-DEGかを判定する閾値を決める問題
 - 有意水準**0.1%**というのが **p -value** < **0.001**に相当
 - False discovery rate (FDR) 5%というのが **q -value** < 0.05に相当
- 発現変動ランキング結果は不変なので上位 **x** 個という決め打ちの場合にはこの問題とは無関係



rownames(tcc\$count)	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000139116	0.9	1.7	0.0	4.2	11.1	3.6	ENSG00000139116	1.20	2.91	0.000997	0.013507	4418	1
ENSG00000205362	2.6	0.8	3.0	11.2	6.5	12.0	ENSG00000205362	2.20	2.21	0.000997	0.013507	4419	1
ENSG00000268592	752.4	480.9	883.8	455.5	351.3	208.0	ENSG00000268592	8.93	-1.06	0.000998	0.013507	4420	1
ENSG00000099622	2392.4	2684.3	2510.3	1460.5	1299.4	1740.2	ENSG00000099622	10.93	-0.75	0.000998	0.013507	4421	1
ENSG00000248958	3.4	3.3	1.0	9.8	5.6	19.1	ENSG00000248958	2.45	2.16	0.000998	0.013507	4422	1
ENSG00000227644	5.1	3.3	11.1	0.0	0.9	1.2	ENSG00000227644	1.10	-3.20	0.001001	0.01354	4423	1
ENSG00000176018	483.6	600.6	454.5	217.3	373.5	151.8	ENSG00000176018	8.48	-1.05	0.001002	0.013552	4424	1
ENSG00000155962	1.7	7.4	5.0	16.8	15.8	12.0	ENSG00000155962	3.07	1.65	0.001003	0.013569	4425	1
ENSG00000232549	63.3	62.8	77.6	25.2	38.0	39.4	ENSG00000232549	5.59	-0.99	0.001003	0.013569	4426	1
ENSG00000116701	105.3	86.8	81.6	148.6	144.6	215.1	ENSG00000116701	6.96	0.89	0.001009	0.013638	4427	1
ENSG00000213996	98.4	90.9	102.8	50.5	34.3	64.5	ENSG00000213996	6.12	-0.97	0.001012	0.013674	4428	1

有意水準**0.1%**で59,857遺伝子を検定すると、4,422個が棄却された(p < **0.001**を満たすものは59,857遺伝子中4,422個でした)



多重比較問題：FDRって何？

- **DEG**かnon-DEGかを判定する閾値を決める問題
 - 有意水準**0.1%**というのが **p -value** < **0.001**に相当
 - False discovery rate (FDR) 5%というのが **q -value** < 0.05に相当
- 発現変動ランキング結果は不変なので上位 **x** 個という決め打ちの場合にはこの問題とは無関係



rownames(tcc\$count)	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000139116	0.9	1.7	0.0	4.2	11.1	3.6	ENSG00000139116	1.20	2.91	0.000997	0.013507	4418	1
ENSG00000205362	2.6	0.8	3.0	11.2	6.5	12.0	ENSG00000205362	2.20	2.21	0.000997	0.013507	4419	1
ENSG00000268592	752.4	480.9	883.8	455.5	351.3	208.0	ENSG00000268592	8.93	-1.06	0.000998	0.013507	4420	1
ENSG00000099622	2392.4	2684.3	2510.3	1460.5	1299.4	1740.2	ENSG00000099622	10.93	-0.75	0.000998	0.013507	4421	1
ENSG00000248958	3.4	3.3	1.0	9.8	5.6	19.1	ENSG00000248958	2.45	2.16	0.000998	0.013507	4422	1
ENSG00000227644	5.1	3.3	11.1	0.0	0.9	1.2	ENSG00000227644	1.10	-3.20	0.001001	0.01354	4423	1

p 値の定義から、59,857遺伝子 \times 0.001 = 59.857個分の真のnon-DEGをDEGと判定ミスするのを許容することに相当



$p < 0.001$ を満たす4,422個の中に占める偽物の割合は $59.857 / 4,422 = 0.013536$ と計算することができる



これ(0.013536)がFDR!!



Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

■ トレーニング(14:30-16:30)

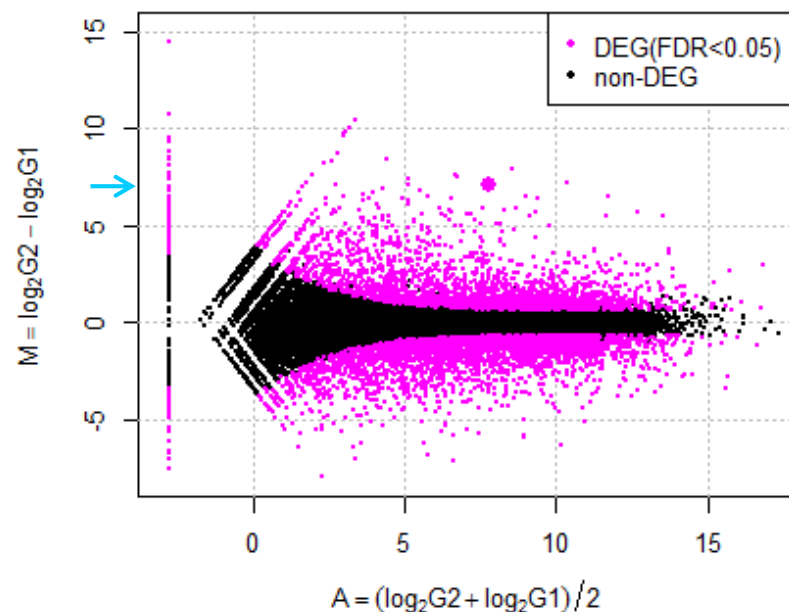
- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

分布やモデルのイントロ

- TCC (Sun et al., 2013)を用いたDEG同定

59,857 genes

	G1群			G2群		
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						



rownames(tcc\$coun	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000240386	0.0	0.0	0.0	5608.1	5097.7	8188.0	ENSG00000240386	-2.78	14.48	1.75E-139	1.04E-134	1	1
ENSG00000128564	15.4	22.3	19.1	2856.6	2462.6	2555.3	ENSG00000128564	7.80	7.11	4.03E-107	1.21E-102	2	1
ENSG00000188064	7.7	5.8	10.1	1425.5	1254.0	1486.8	ENSG00000188064	6.71	7.47	2.67E-98	5.33E-94	3	1
ENSG00000101188	6.0	5.0						5	7.55	3.81E-97	5.70E-93	4	1
ENSG00000163431	3716.6	3244.5						5	-5.82	2.90E-80	3.48E-76	5	1
ENSG00000204291	1215.5	1236.8						4	-5.72	3.45E-78	3.44E-74	6	1
ENSG00000181634	107.0	158.6						9	7.20	1.04E-76	8.88E-73	7	1
ENSG00000178726	53.1	46.3						1	6.51	2.39E-74	1.79E-70	8	1
ENSG00000117600	576.9	518.9						3	-6.83	1.14E-72	7.59E-69	9	1
ENSG00000158050	7.7	9.1											
ENSG00000124126	50.5	44.6											

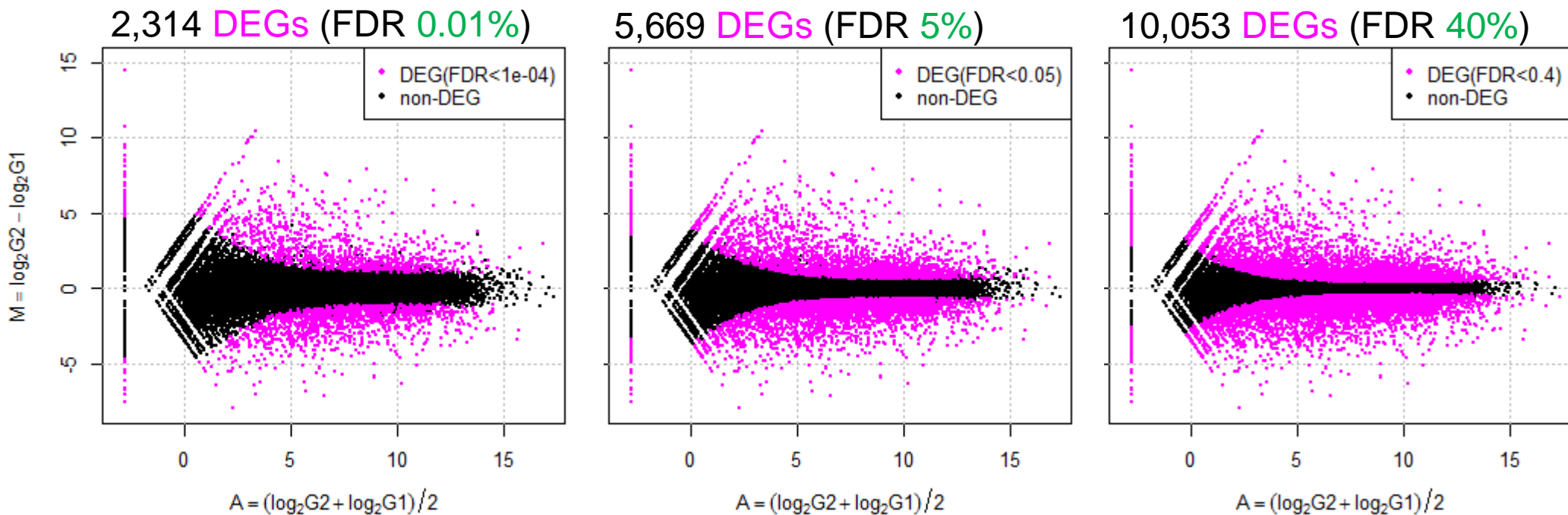
```

R Console
> (15.4 + 22.3 + 19.1)/3
[1] 18.93333
> (2856.6 + 2462.6 + 2555.3)/3
[1] 2624.833
> (log2(2624.833) + log2(18.93333))/2
[1] 7.800433
> log2(2624.833) - log2(18.93333)
[1] 7.115154
> |
    
```

M-A plotのM値は倍率変化(log比)に相当(2^{7.11}倍G2群で高発現)

DEG同定結果:FDR閾値の違い

- TCC (Sun et al., 2013)を用いたDEG同定

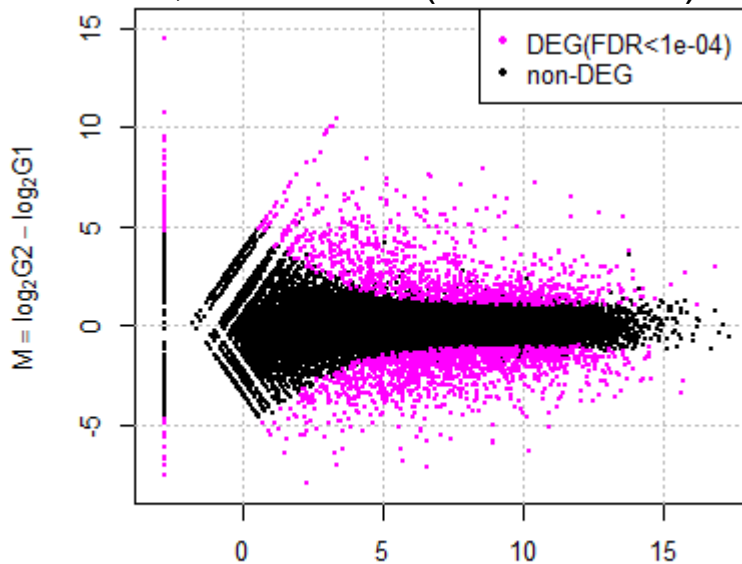


FDR閾値を緩めると得られるDEG数は増える傾向
厳しめ ← FDR閾値 → 緩め

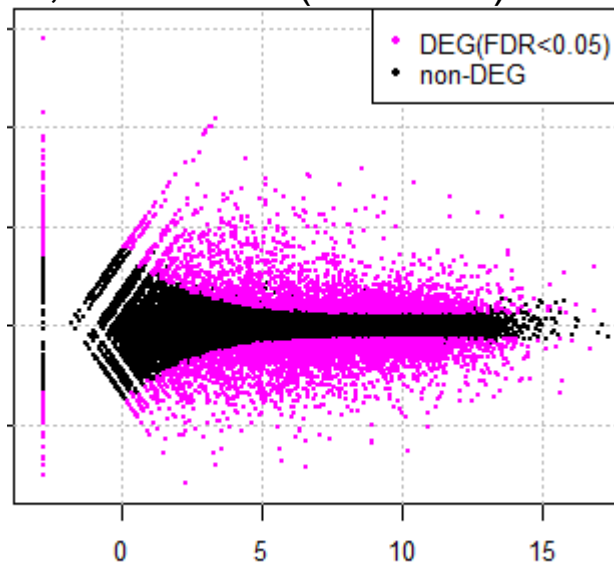
分布やモデル

- TCC (Sun et al., 2013)を用いたDEG同定

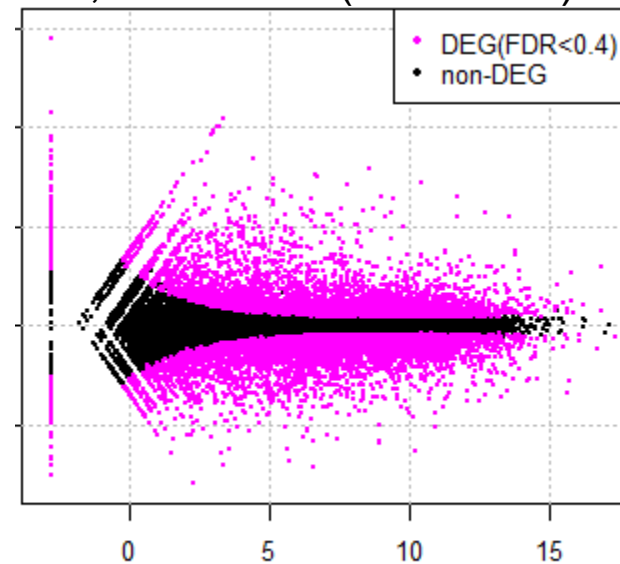
2,314 DEGs (FDR 0.01%)



5,669 DEGs (FDR 5%)



10,053 DEGs (FDR 40%)



$$A = (\log_2 G_2 + \log_2 G_1) / 2$$

$$A = (\log_2 G_2 + \log_2 G_1) / 2$$

$$A = (\log_2 G_2 + \log_2 G_1) / 2$$

黒の分布はnon-DEGの分布に相当



分布やモデル

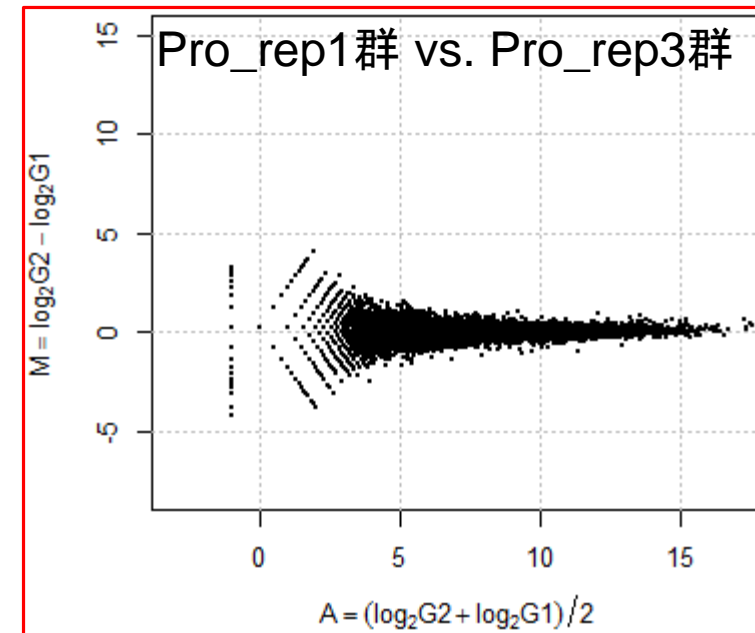
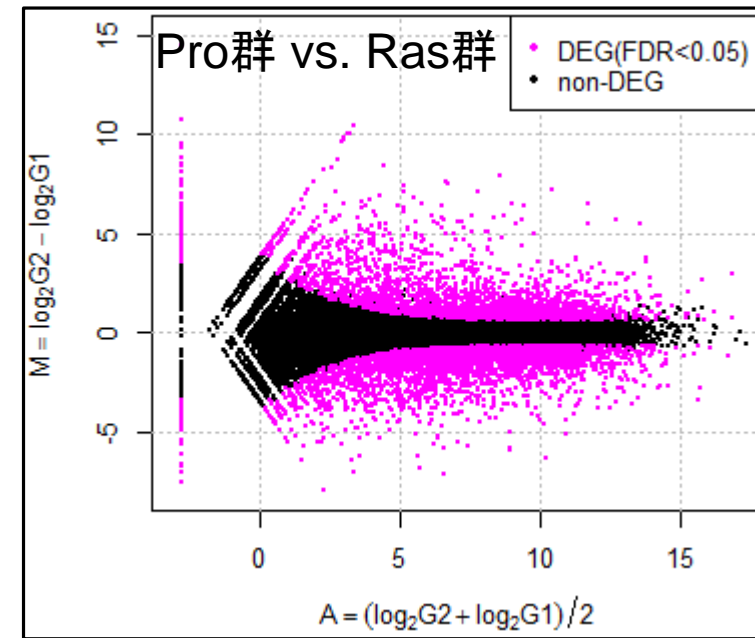
59,857 genes

	同一群 (G1群)			同一群 (G2群)		
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

G1群

G2群

黒の分布はnon-DEGの分布に相当



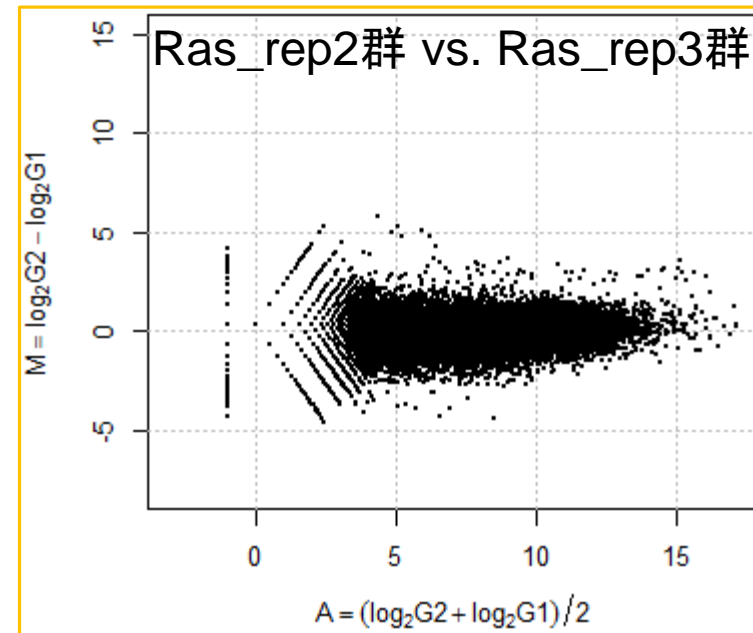
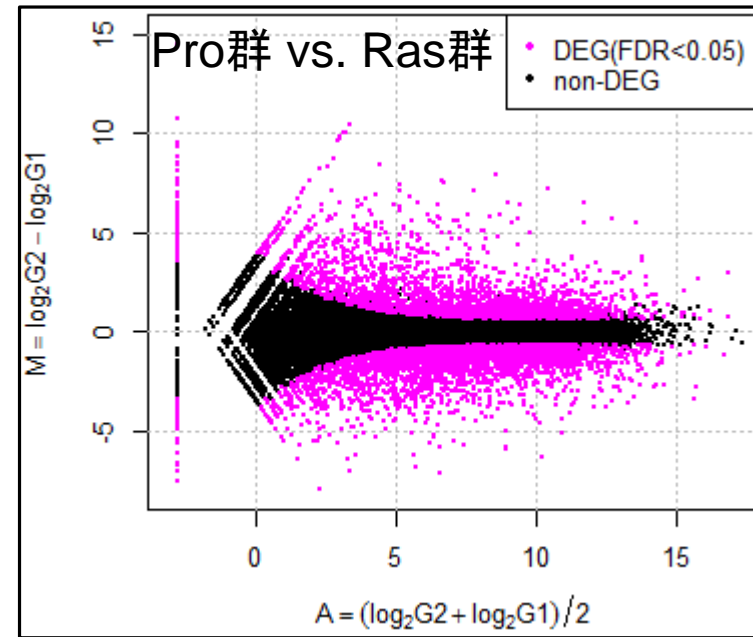
分布やモデル

59,857 genes

	同一群 (G1群)			同一群 (G2群)		
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

G1群 G2群

黒の分布はnon-DEGの分布に相当



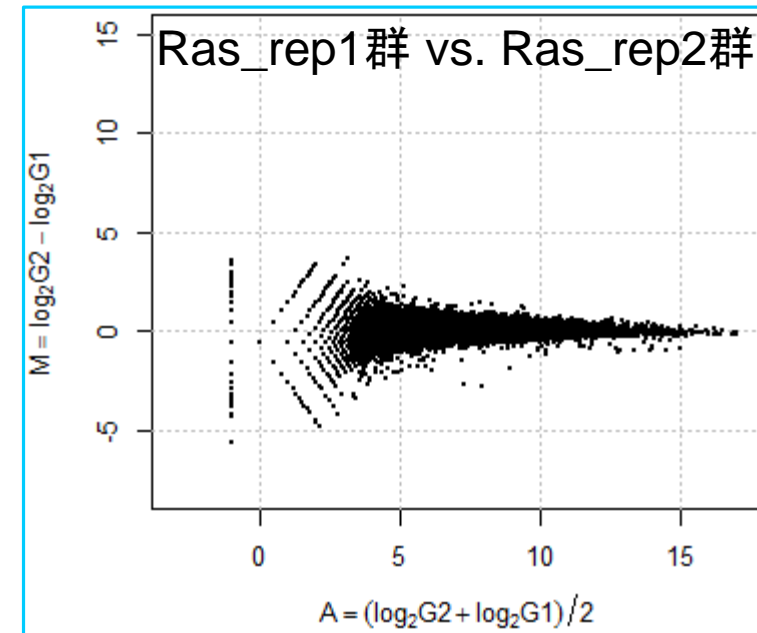
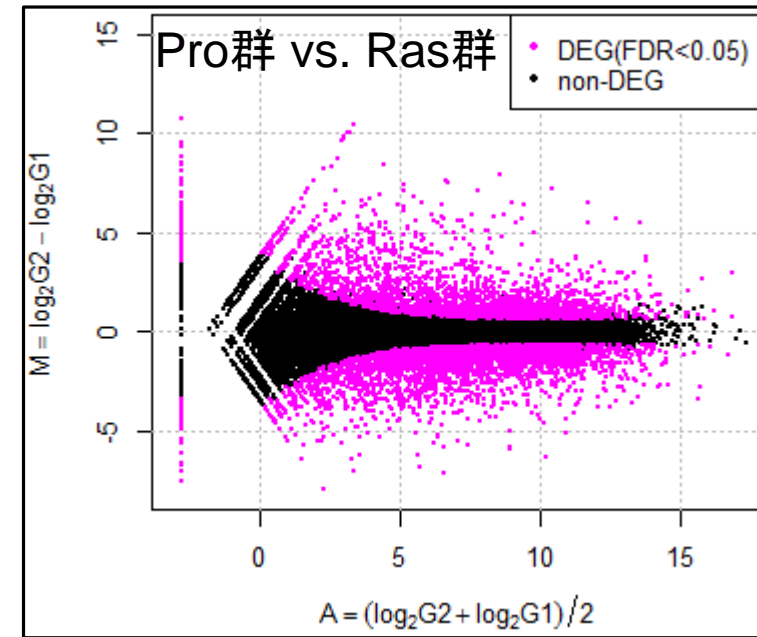
分布やモデル

59,857 genes

	同一群 (G1群)			同一群 (G2群)		
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG000000240386	0	0	0	4001	5500	6851
...						
ENSG000000128564	18	27	19	2038	2657	2138
...						

G1群 G2群

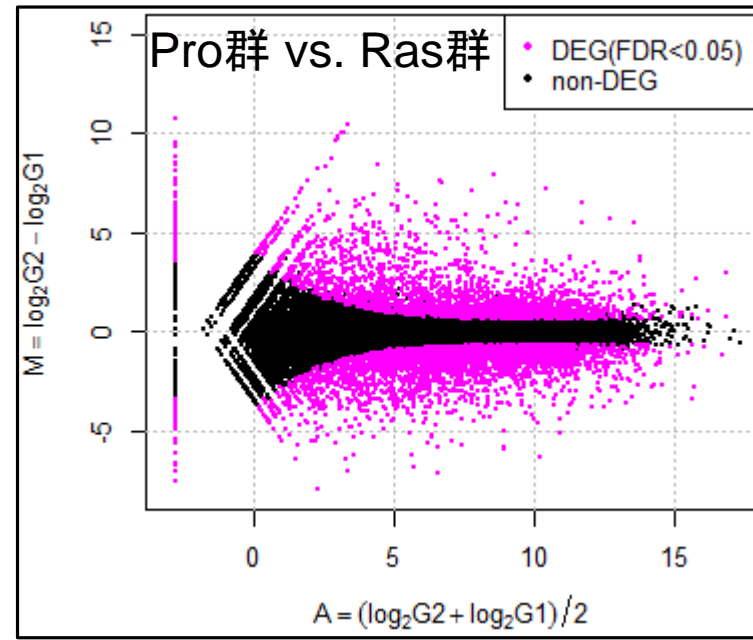
黒の分布はnon-DEGの分布に相当



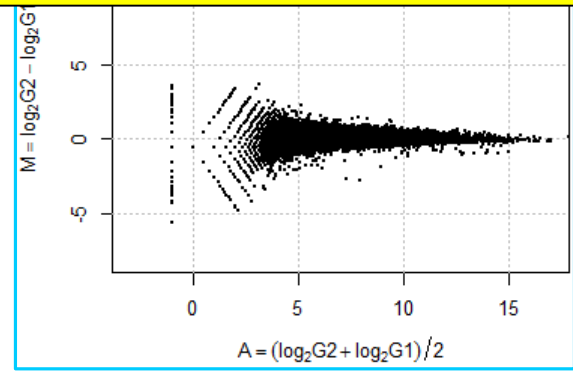
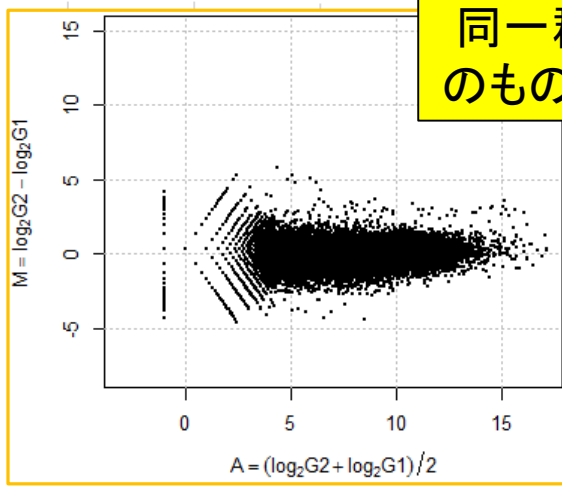
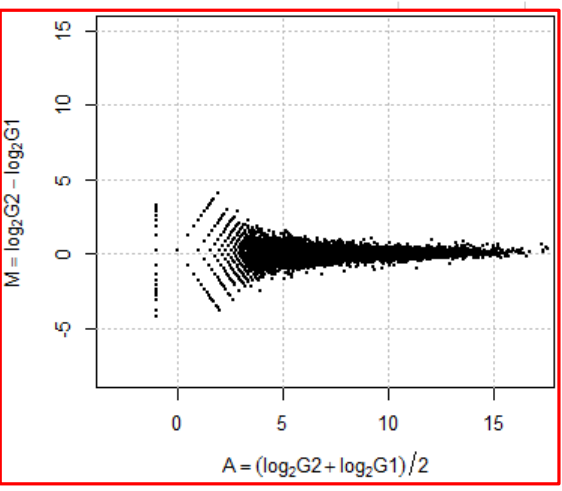
分布やモデル

59,857 genes

	同一群 (G1群)			同一群 (G2群)		
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						



同一群内のばらつきの分布 (non-DEG分布) 以外のものがDEGと判定されるのが統計的手法の結果

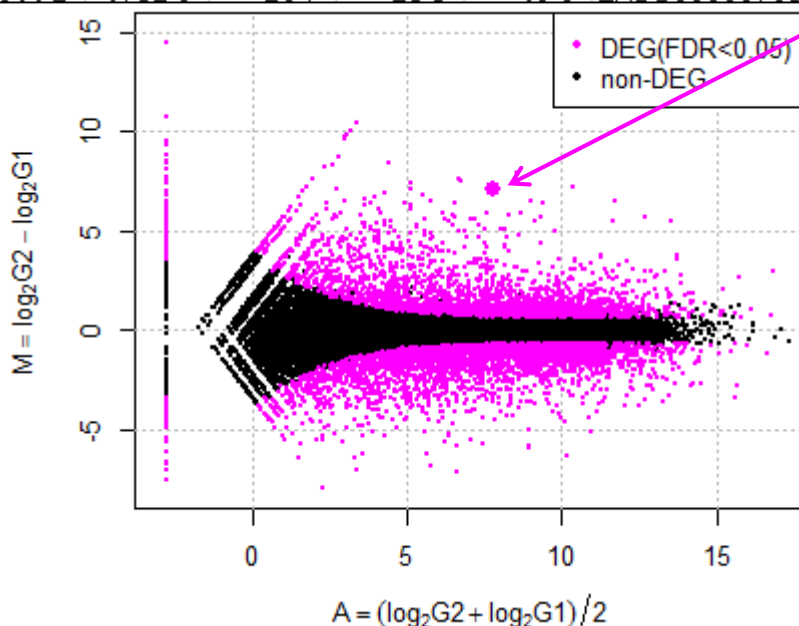




統計的手法とは

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

rownames(tcc\$coun	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000240386	0.0	0.0	0.0	5608.1	5097.7	8188.0	ENSG00000240386	-2.78	14.48	1.75E-139	1.04E-134	1	1
ENSG00000128564	15.4	22.3	19.1	2856.6	2462.6	2555.3	ENSG00000128564	7.80	7.11	4.03E-107	1.21E-102	2	1
ENSG00000188064	7.7	5.8	10.1	1425.5	1254.0	1486.8	ENSG00000188064	6.71	7.47	2.67E-98	5.33E-94	3	1
ENSG00000101188	6.0	5.0	11.1	1477.4	1407.0	1254.9	ENSG00000101188	6.65	7.55	3.81E-97	5.70E-93	4	1



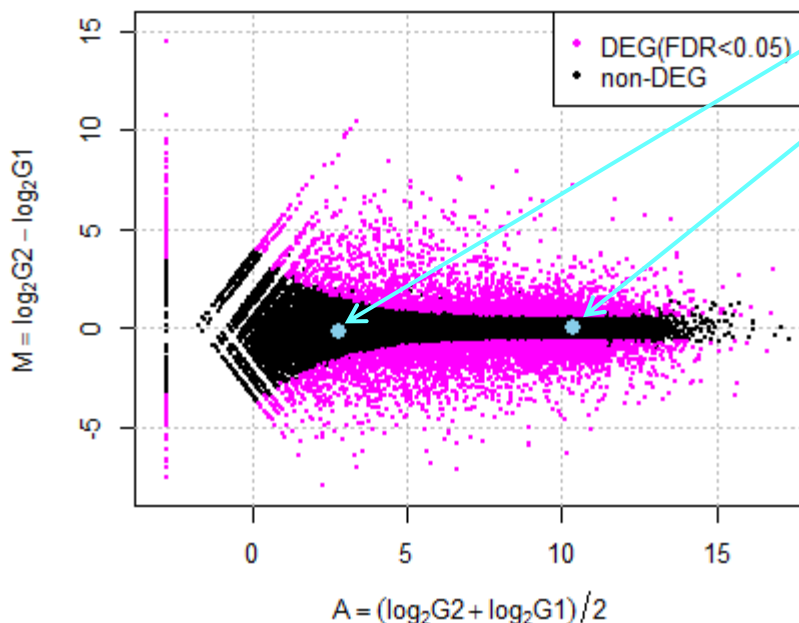
同一群内のばらつきの分布(non-DEG分布)から遠く離れたところに位置するものは0に近いp-value



統計的手法とは

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

rownames(tcc\$coun	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000165660	404.0	390.0	301.3	333.6	386.5	350.2	ENSG00000165660	8.50	-0.03	0.893466	1	24460	0
ENSG00000166359	4.3	7.4	10.1	9.8	3.7	6.0	ENSG00000166359	2.78	-0.16	0.893944	1	24461	0
ENSG00000146676	1141.9	1420.2	1272.8	1156.4	1558.0	1204.7	ENSG00000146676	10.34	0.03	0.89404	1	24462	0
ENSG00000229880	112.1	114.8	94.7	81.3	114.9	133.9	ENSG00000229880	6.76	0.04	0.894049	1	24463	0



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



Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

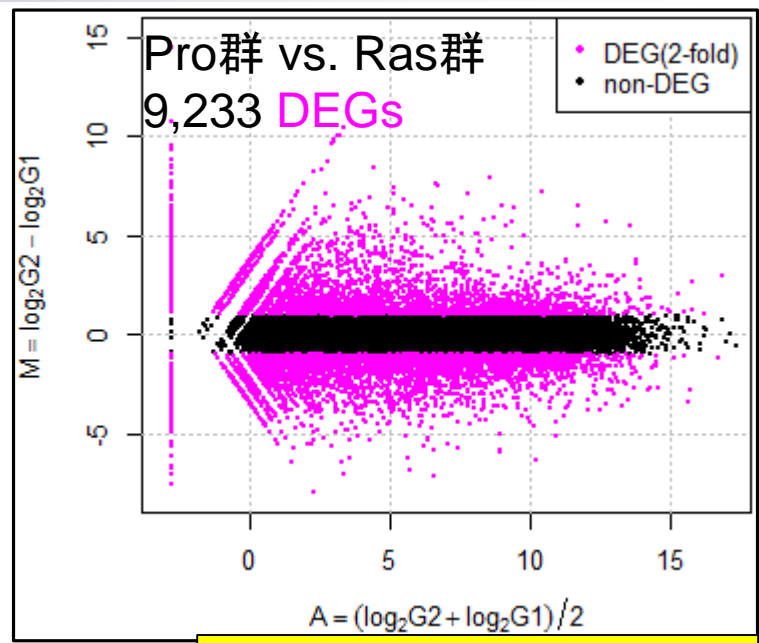
■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

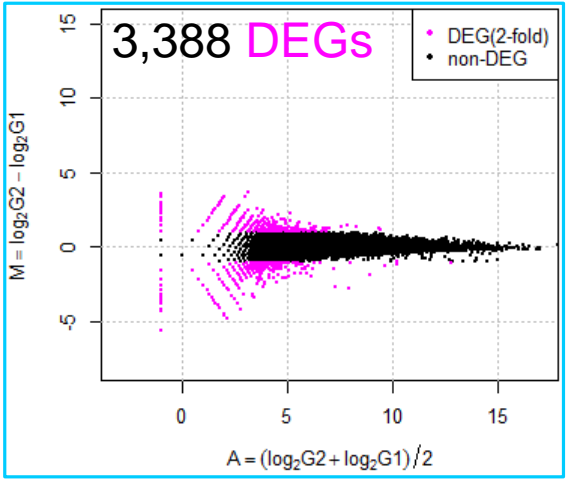
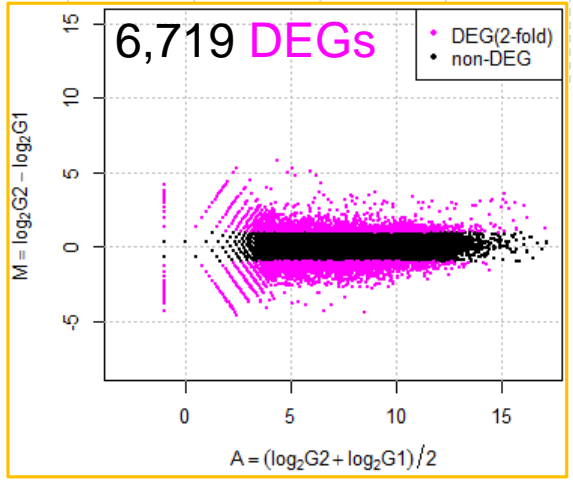
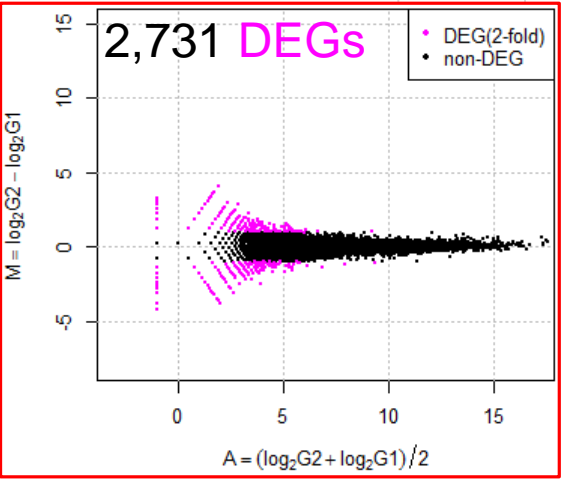
倍率変化の結果

59,857 genes

	同一群 (G1群)			同一群 (G2群)		
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						



同一群内比較でも多数の偽陽性が検出されている

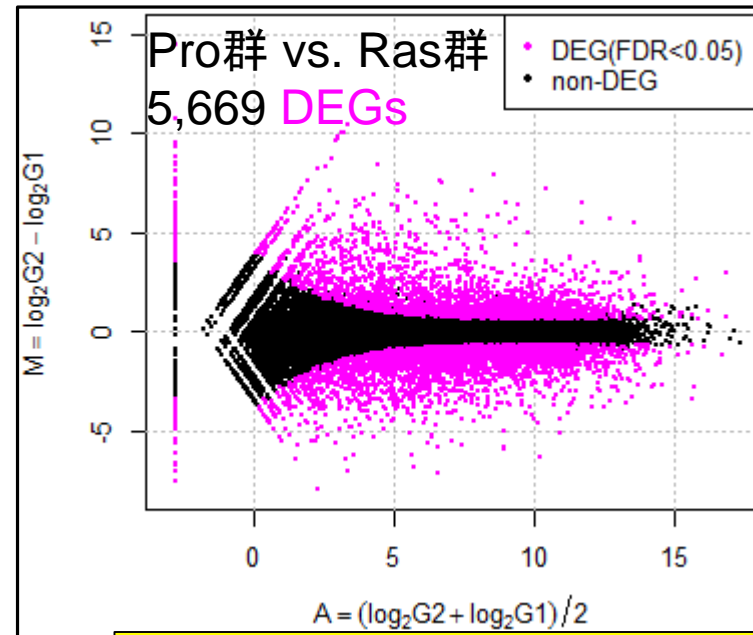


...

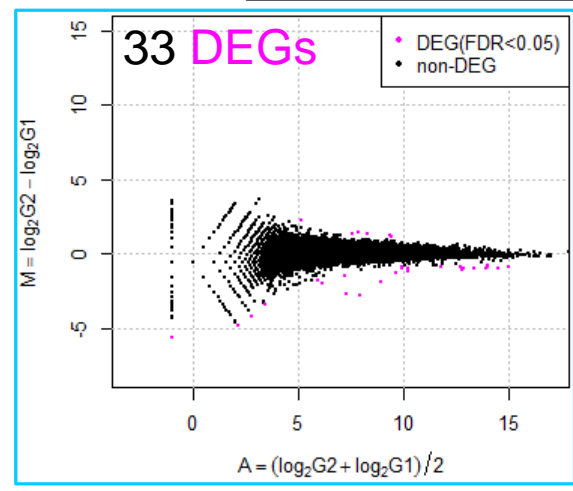
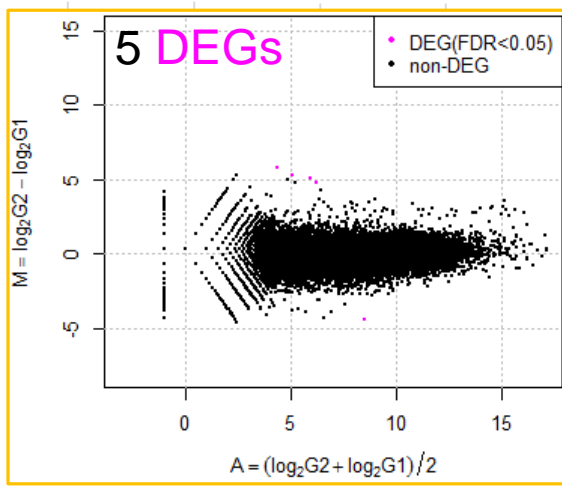
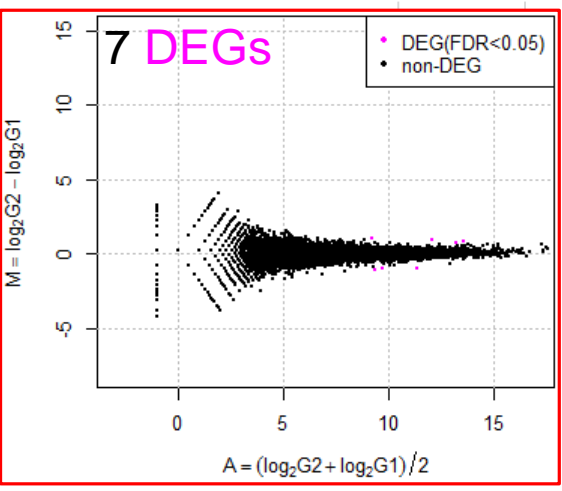
統計的手法 TCCの結果

59,857 genes

	同一群 (G1群)			同一群 (G2群)		
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						



同一群内比較でも多少の偽陽性が検出されるが許容範囲



...

Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- **理想的な実験デザイン**
- データの正規化

■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

理想的な実験デザイン

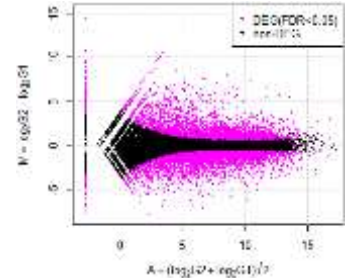
■ 腎臓 vs. 肝臓のようなG1群 vs. G2群の比較の場合

□ 生のリードカウントのデータ(基本的には整数値)



Gene ID	A1	A2	A3	A4	...	B1	B2	B3	B4	...
Gene1										
Gene2										
Gene3										
Gene4										
Gene5										
Gene6										
Gene7										
...										

G1_rep1: ある生物の腎臓
 G1_rep2: 同じ生物種の別個体の腎臓
 G1_rep3: 同じ生物種のさらに別個体の腎臓
 ...
 G2_rep1: ある生物の肝臓
 G2_rep2: 同じ生物種の別個体の肝臓
 ...



Biological replicatesのデータ
 生物学的なばらつき(個体間の違い)をできるだけ正確に捉えて本物のDEGを感度・特異度高く検出

2群間比較: technical replicatesデータ

data_marioni.txt (ヒトのデータ)



kidney(腎臓)



liver(肝臓)

rownames(data)	R1L1Kidney	R1L3Kidney	R1L7Kidney	R2L2Kidney	R2L6Kidney	R1L2Liver	R1L4Liver	R1L6Liver	R1L8Liver	R2L3Liver
ENSG00000000003	178	167	179	172	151	138	178	175	187	169
ENSG00000000005	0	0	0	0	1	0	0	0	0	0
ENSG00000000419	53	78	64	72	71	30	42	41	33	43
ENSG00000000457	22	33	30	27	30	47	60	37	42	62
ENSG00000000460	9	7	18	14	9	19	9	13	14	19
ENSG00000000938	14	18	7	27	15	42	34	39	37	45
ENSG00000000971	28									
ENSG00000001036	154									
ENSG00000001084	77									
ENSG00000001167	41									
ENSG00000001460	24									
ENSG00000001461	23									
ENSG00000001497	55									
ENSG00000001561	139									
ENSG00000001617	136									

18,110 genes

Technical replicatesのデータ
 レーン間の違いなどサンプル内の技術的なばらつきを調べるための同一個体由来データ。このようなデータで2群間比較し、**発現変動遺伝子がどの程度あるか**といった数に関する議論は**ほぼ無意味**。
 理由: Biological variation > Technical variation。得られた結果はその**個体内のみで成立する**ものであり、同じ生物種の別個体においても同様な事象が観測されるわけではない。

Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

データの正規化

59,857 genes

	同一群 (Pro群)			同一群 (Ras群)		
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

G1群

G2群

G1群

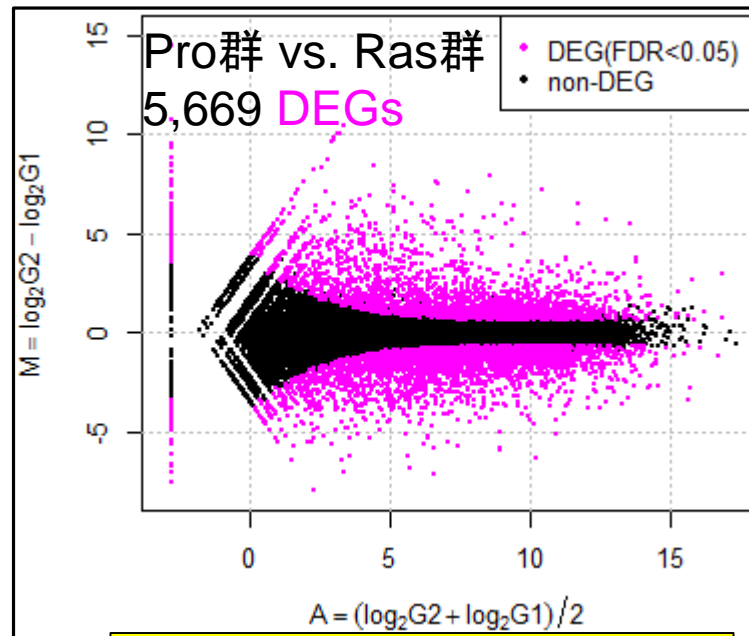
G2群

G1群

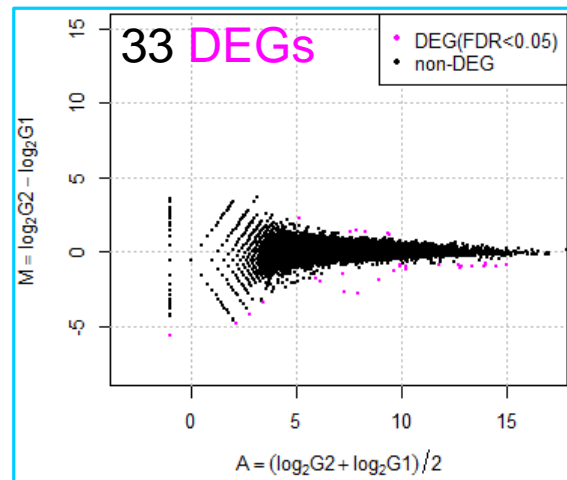
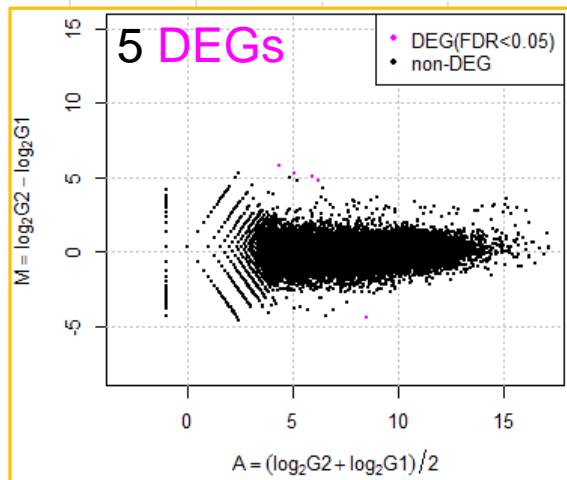
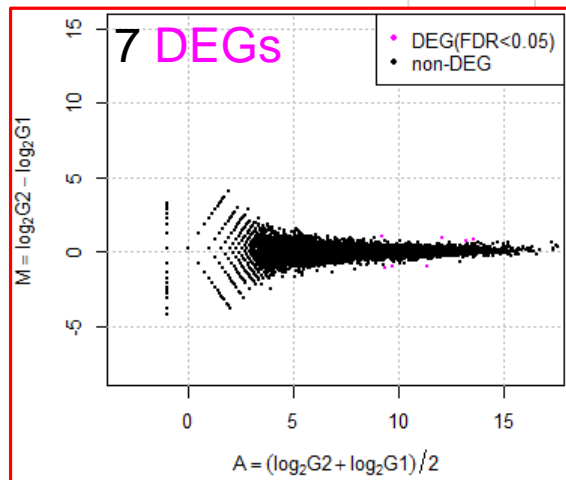
G2群

G1群

G2群



これらの結果は、正規化後のデータで描画したものです



...

データの正規化

59,857 genes

	同一群 (Pro群)			同一群 (Ras群)		
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

G1群

G2群

G1群

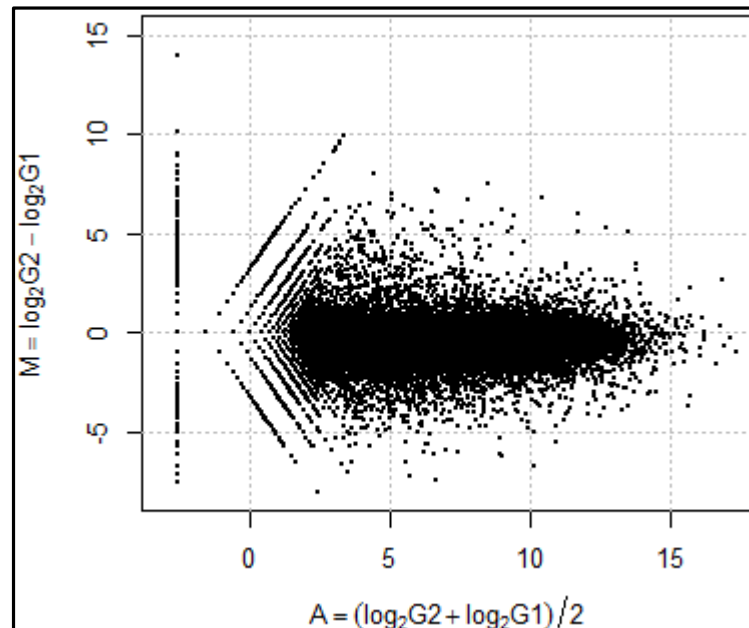
G2群

G1群

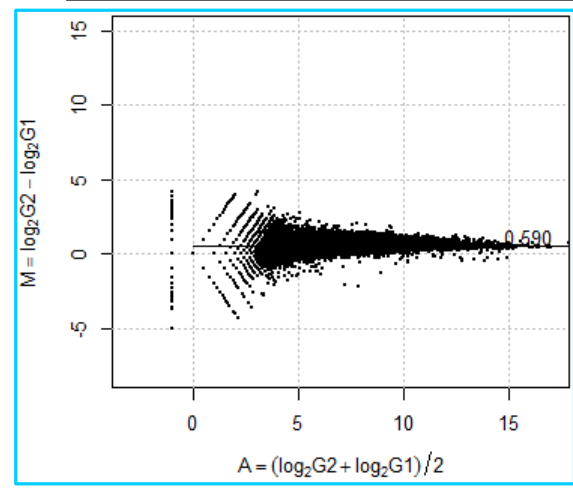
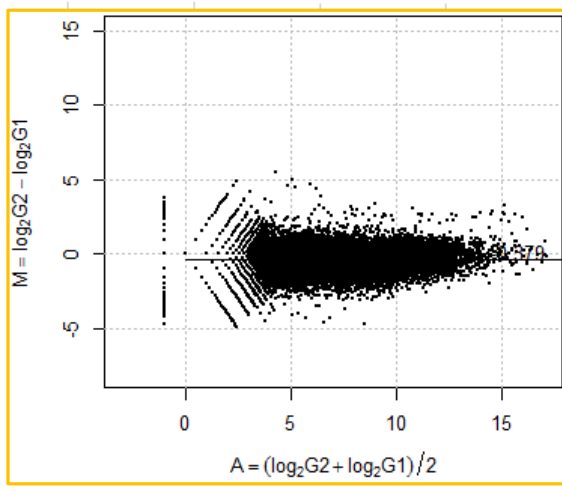
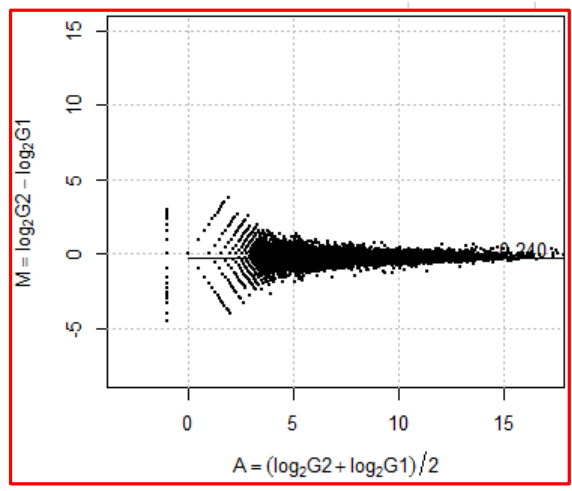
G2群

G1群

G2群



正規化前のデータでプロットすると全体的にM = 0からずれていることがわかる



...

データ正規化の目的

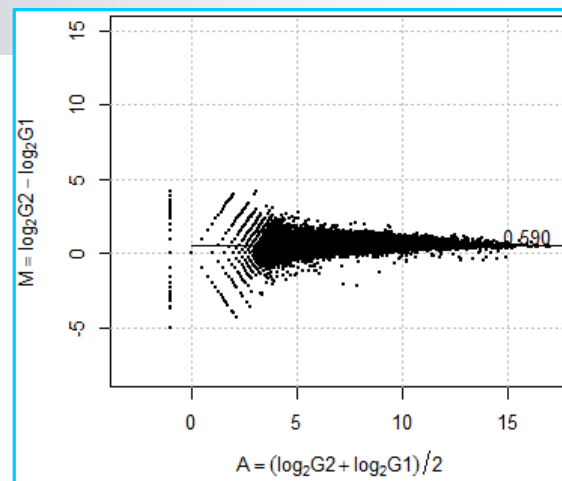
59,857 genes

同一群 (Pro群)

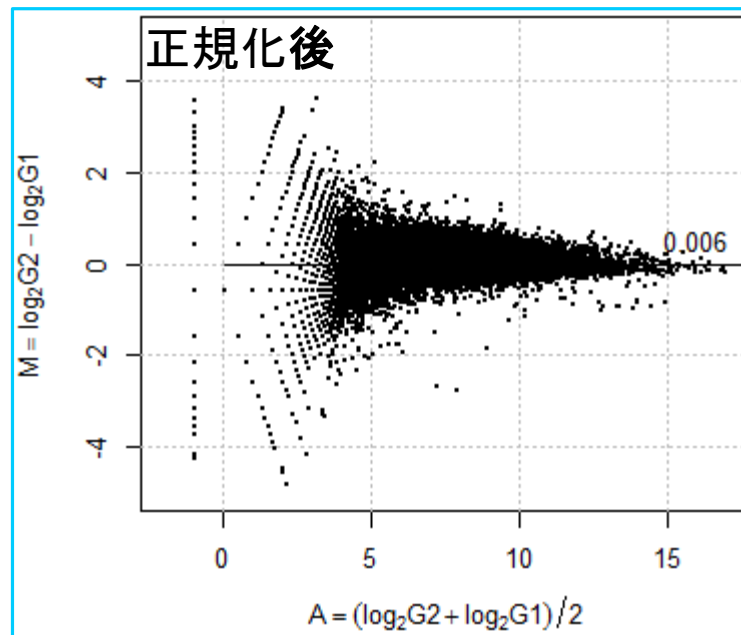
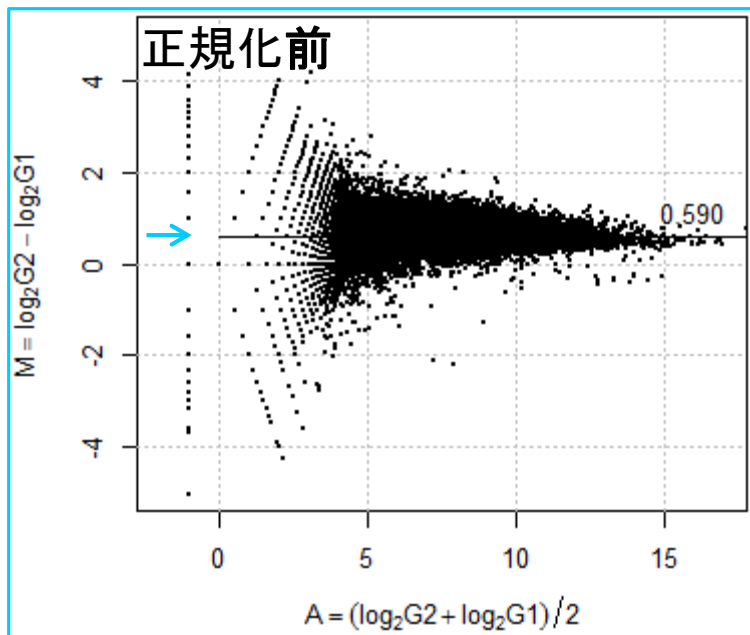
同一群 (Ras群)

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	
...						

G1群 G2群



M-A plotのM値は $\log_2(G2/G1)$ に相当する。
 正規化前: G2群で $2^{0.590}$ (= 1.505)倍高発現
 正規化後: G2群で $2^{0.006}$ (= 1.004)倍高発現

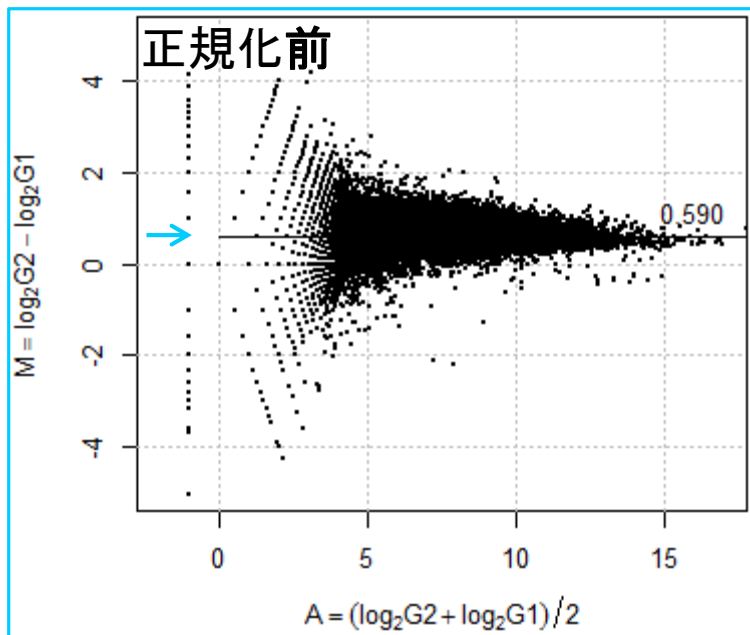


データ正規化の目的

	A	B	C	D	E	F	G
1		Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
59856	ENSG00000272543	0	0	0	1	1	0
59857	ENSG00000272544	0	1	0	0	0	0
59858	ENSG00000272545	0	0	0	0	0	0
59859		22669407	22521535	19989914	13630668	20268177	18126870

G1群 G2群

G2群で $2^{0.590}$ (= 1.505)倍高発現となっているのは、G2群の総リード数(または総カウント数)がG1群に比べて約1.5倍多いから



```
R Console
> in_f <- "srp017142_count_bowtie.txt"
> data <- read.table(in_f, header=TRUE, row.names=1, s$
> colSums(data)
Pro_rep1 Pro_rep2 Pro_rep3 Ras_rep1 Ras_rep2 Ras_rep3
22669407 22521535 19989914 13630668 20268177 18126870
> 20268177/13630668
[1] 1.486954
> |
```

データ正規化の目的

生のカウントデータ: srp017142_count_bowtie.txt

	A	B	C	D	E	F	G	H	I
1		Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3		G2/G1
2	ENSG000000000003	480	513	366	124	271	366		2.19
3	ENSG000000000005	0	0	0	1	0	0		0.00
4	ENSG000000000419	282	354	208	165	301	209		1.82
5	ENSG000000000457	167	198	155	156	248	129		1.59
6	ENSG000000000460	114	112	101	55	81	59		1.47
7	ENSG000000000938	0	0	0	2	2	1		1.00
8	ENSG000000000971	712	867	570	237	394	142		1.66
9	ENSG00000001036	2939	2860	2338	1612	2672	2341		1.66
10	ENSG00000001084	811	937	599	433	759	421		1.75
11	ENSG00000001167	731	843	666	764	1314	920		1.72
12	ENSG00000001460	417	427	411	241	390	166		1.62
13	ENSG00000001461	6629	6144	5384	1430	2312	839		1.62
14	ENSG00000001497	680	752	648	487	689	680		1.41
15	ENSG00000001561	3	5	3	0	4	0		#DIV/0!
16	ENSG00000001617	2770	2647	2690	1366	1917	1388		1.40
17	ENSG00000001626	11	12	3	0	2	0		#DIV/0!

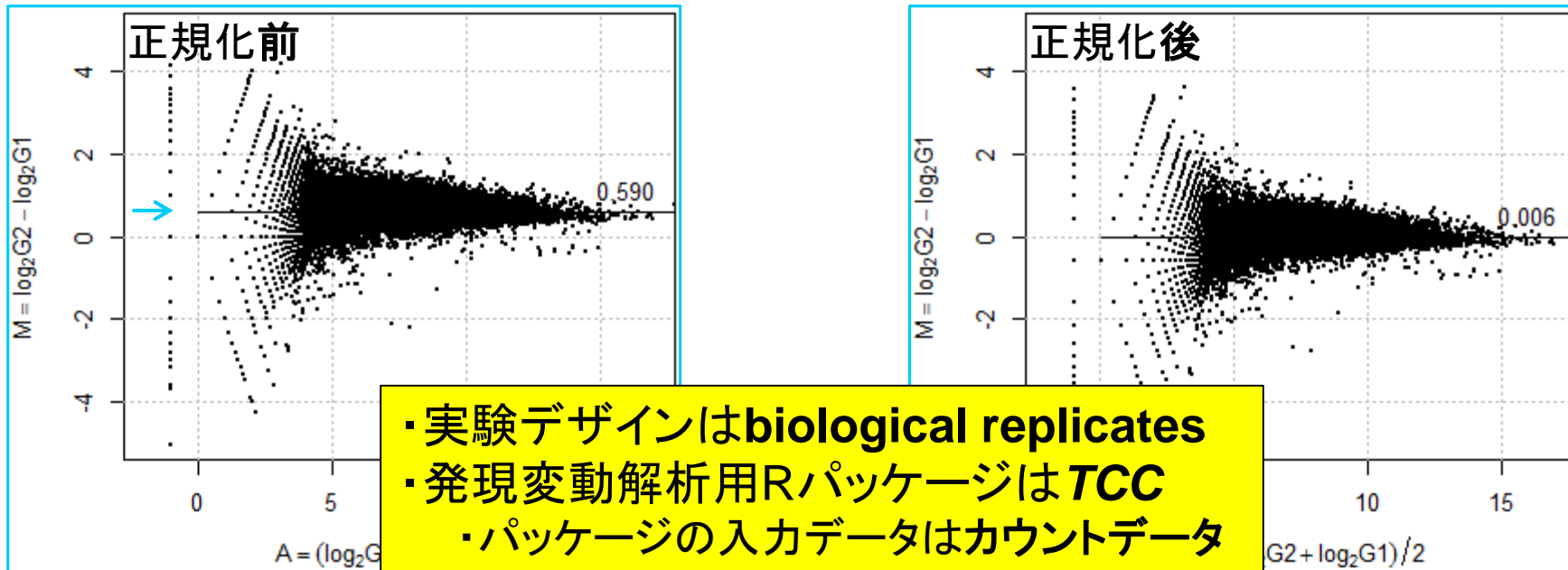
G1群 G2群

全体的にG2群で $2^{0.590}$ (= 1.505)倍高発現とは、こういう意味です



データ正規化の目的

- 比較可能にすること
- DEGはDEGと、non-DEGはnon-DEGと正しく判定されるように揃えること
- よりよい正規化法とは？
 - 発現変動ランキング結果で、真のDEGが上位に、真のnon-DEGが下位にランキングされるような感度・特異度が高い方法 (AUC値が高い方法)
 - non-DEGが全体的に発現変動していないと判定される方法。例えば、正規化後のデータのlog比が0に近い方法。



Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

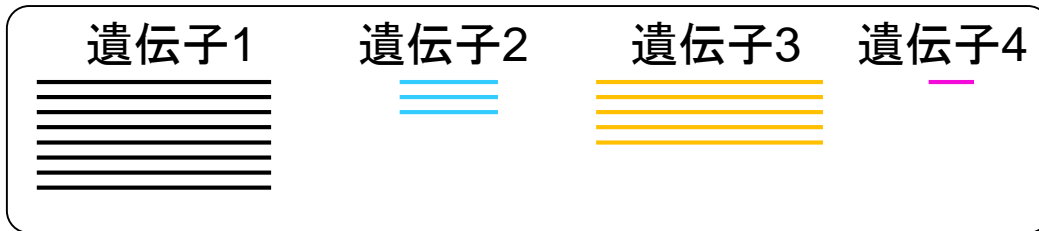
■ トレーニング(14:30-16:30)

- **TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ**
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

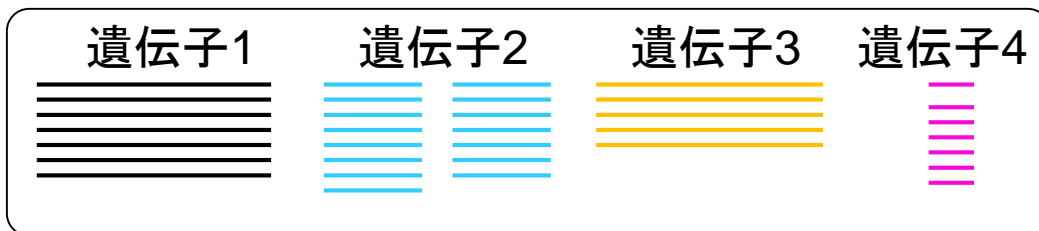
比較トランスクリプトーム解析

- 比較するサンプルまたはグループ間での発現変動遺伝子 (Differentially Expressed Genes; DEGs) 検出が解析の主要部分

光刺激前 (T1) の目のトランスクリプトーム

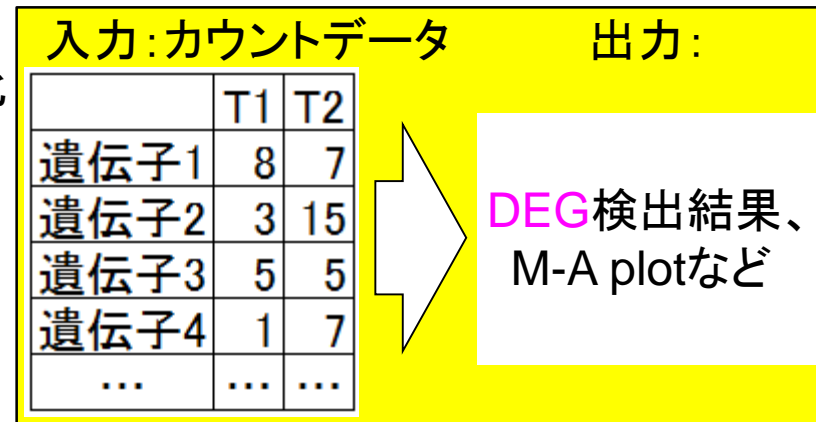


光刺激後 (T2) の目のトランスクリプトーム



TCCの守備範囲

数値化



No TCC, No better result!



Rパッケージ TCC ver. 1.2.0 (次期版)

□ 頑健な正規化法を実装した比較トランスクリプトーム解析パイプライン

■ カウントデータを入力として、発現変動解析結果を出力

■ RNA-seq用

- 2群間比較用 | 対応なし | 複製あり
- 2群間比較用 | 対応なし | 複製なし
- 3, 4群間比較用 | 対応なし | 複製あり
- シミュレーションデータ作成機能: 2, 3, ... 多群
- 2群間比較用 | 対応あり | 複製なし

■ マイクロアレイ用

- 多群間比較用: ROKU法 (Kadota et al., 2006)
- 2群間比較用: WAD法 (Kadota et al., 2008)

■ 両方

- サンプル間クラスタリング

一般的な実験デザインである対応なし、複製ありの2群間比較解析をやってみよう

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help

Home » Bioconductor 2.13 » Software Packages » TCC

TCC

TCC: Differential expression analysis for tag count data with robust normalization strategies

Bioconductor version: Release (2.13)

This package provides a series of functions for performing differential expression analysis from RNA-seq count data using robust normalization strategy (called DEGES). The basic idea of DEGES is that potential differentially expressed genes or transcripts (DEGs) among compared samples should be removed before data normalization to obtain a well-ranked gene list where true DEGs are top-ranked and non-DEGs are bottom ranked. This can be done by performing a multi-step normalization strategy (called DEGES for DEG elimination strategy). A major characteristic of TCC is to provide the robust normalization methods for several kinds of count data (two-group with or without replicates, multi-group/multi-factor, and so on) by virtue of the use of combinations of functions in other sophisticated packages (especially edgeR, DESeq, and BaySeq).

Author: Jianqiang Sun, Tomoaki Nishiyama, Kantaro Shimizu, and Koji Kadota

Maintainer: Jianqiang Sun <jksun@at.b.i.u.-tokyo.ac.jp>, Tomoaki Nishiyama <tomoaki@staff.kanazawa-u.ac.jp>

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("TCC")
```

To cite this package in a publication, start R and enter:

```
citation("TCC")
```

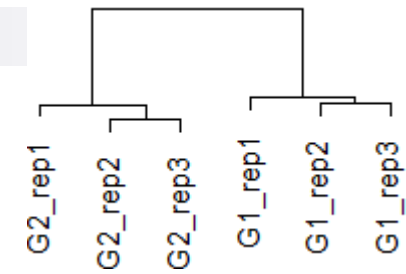
Documentation

PDF [R Script](#) TCC
 PDF Reference Manual
 Text NEWS

Details

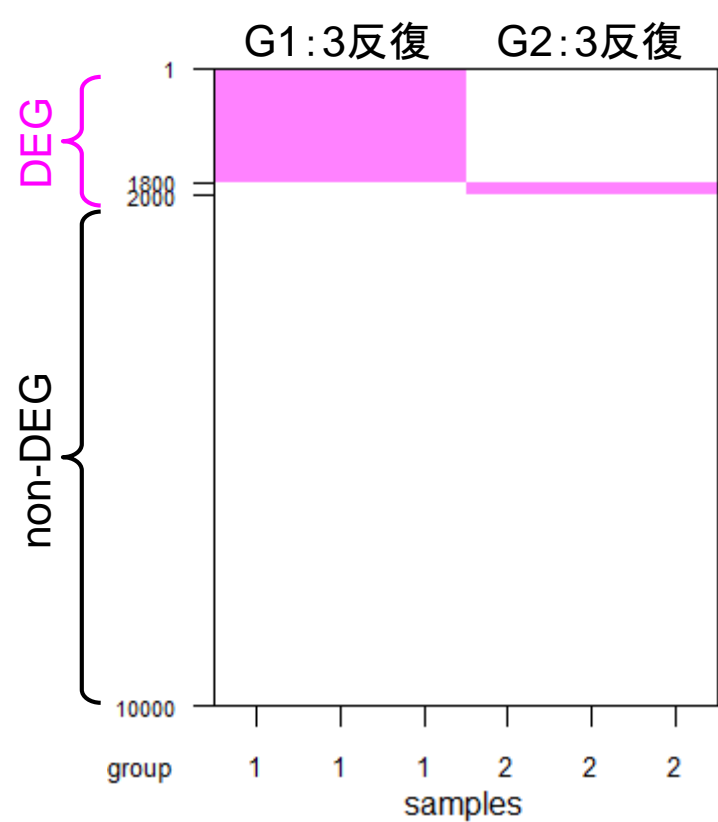
bioViews [Differential Expression](#), [Half Throughput Sequencing](#), [RNAseq](#), [Software](#)
 Version 1.2.0
 In Bioconductor since BioC 2.13 (R-3.0)
 License GPL-3
 Depends R (>= 2.15), methods, [DESeq](#), [edgeR](#), [BaySeq](#), [RGC](#)
 Imports [edgeR](#), [stats](#)
 Suggests [limma](#), [BioGenerics](#)

TCCで複製あり2群間比較



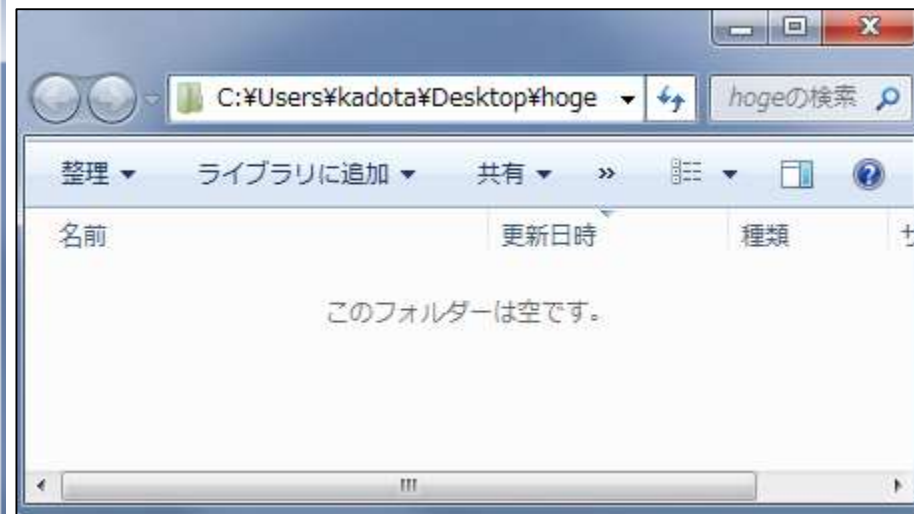
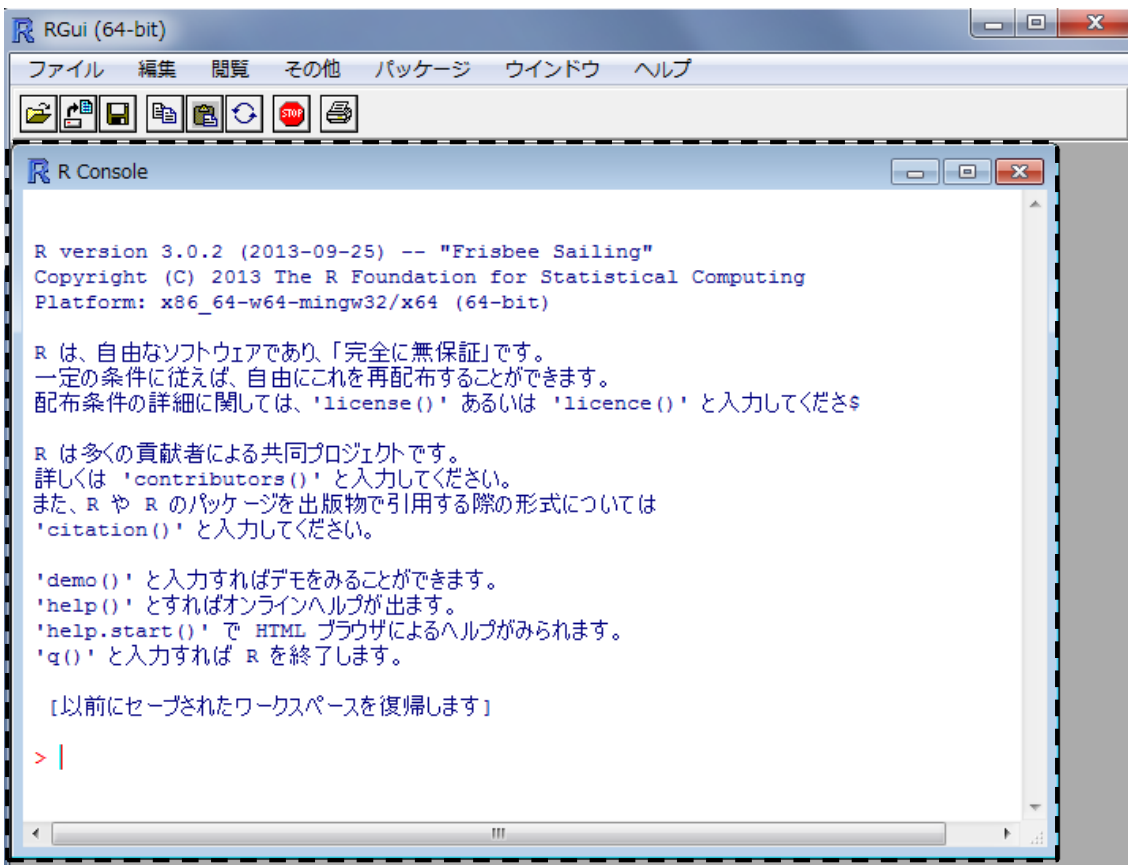
- data_hypodata_3vs3.txt (2群間比較用)
 - G1群:3サンプル、G2群:3サンプル
 - 全部で10,000行×6列。最初の2,000行分が発現変動遺伝子 (DEG)

	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene_1	36	56	144	2	1	0
gene_2	84	152	124	52	37	28
gene_3	592	840	800	151	257	200
gene_4	0	8	4	1	1	3
gene_5	32	32	0	1	1	0
...						
gene_1801	34	86	24	284	180	364
gene_1802	5	1	3	0	160	24
gene_1803	57	56	51	248	192	220
gene_1804	29	25	32	128	204	160
gene_1805	42	29	44	184	156	92
...						
gene_2001	4	8	9	13	12	4
gene_2002	88	139	40	22	44	21
gene_2003	933	667	462	889	396	443
gene_2004	48	37	14	36	57	71
gene_2005	290	338	553	319	210	504
...						
gene_9996	107	67	104	35	65	45
gene_9997	145	220	120	80	95	156
gene_9998						
gene_9999						
gene_10000						



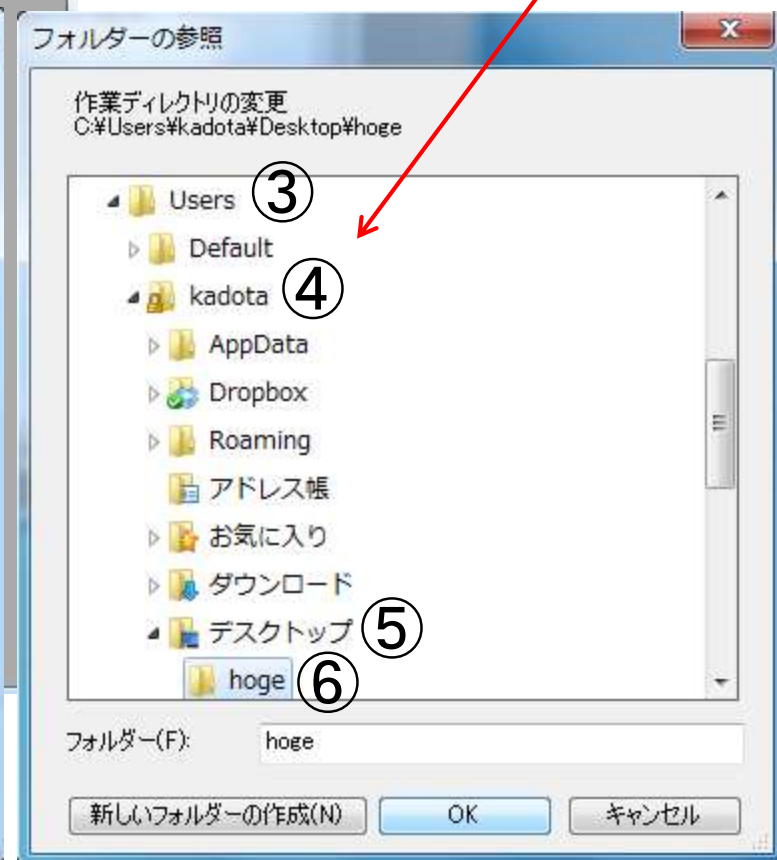
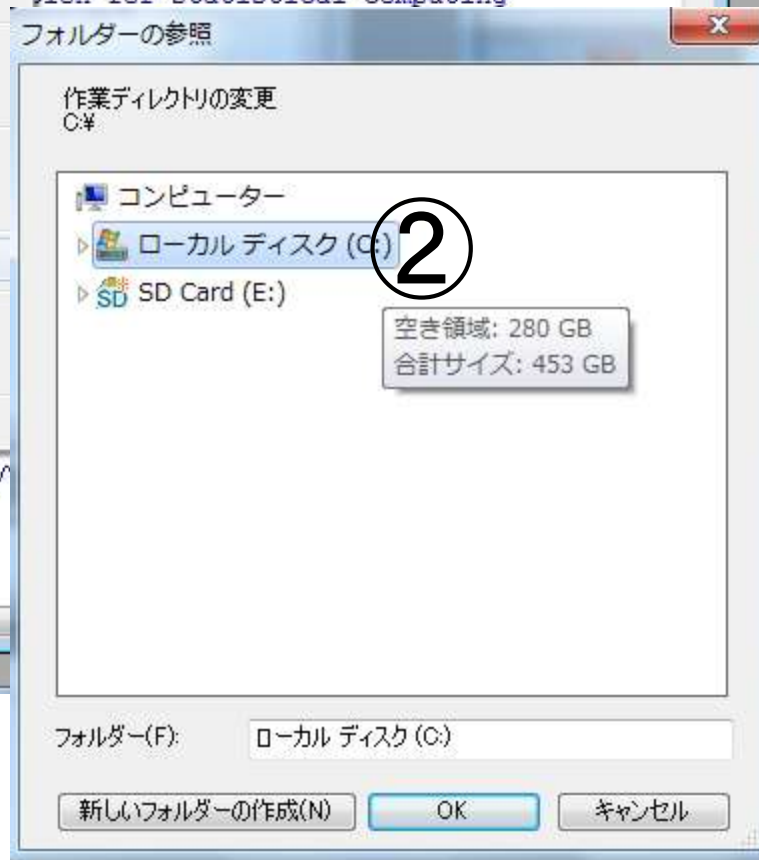
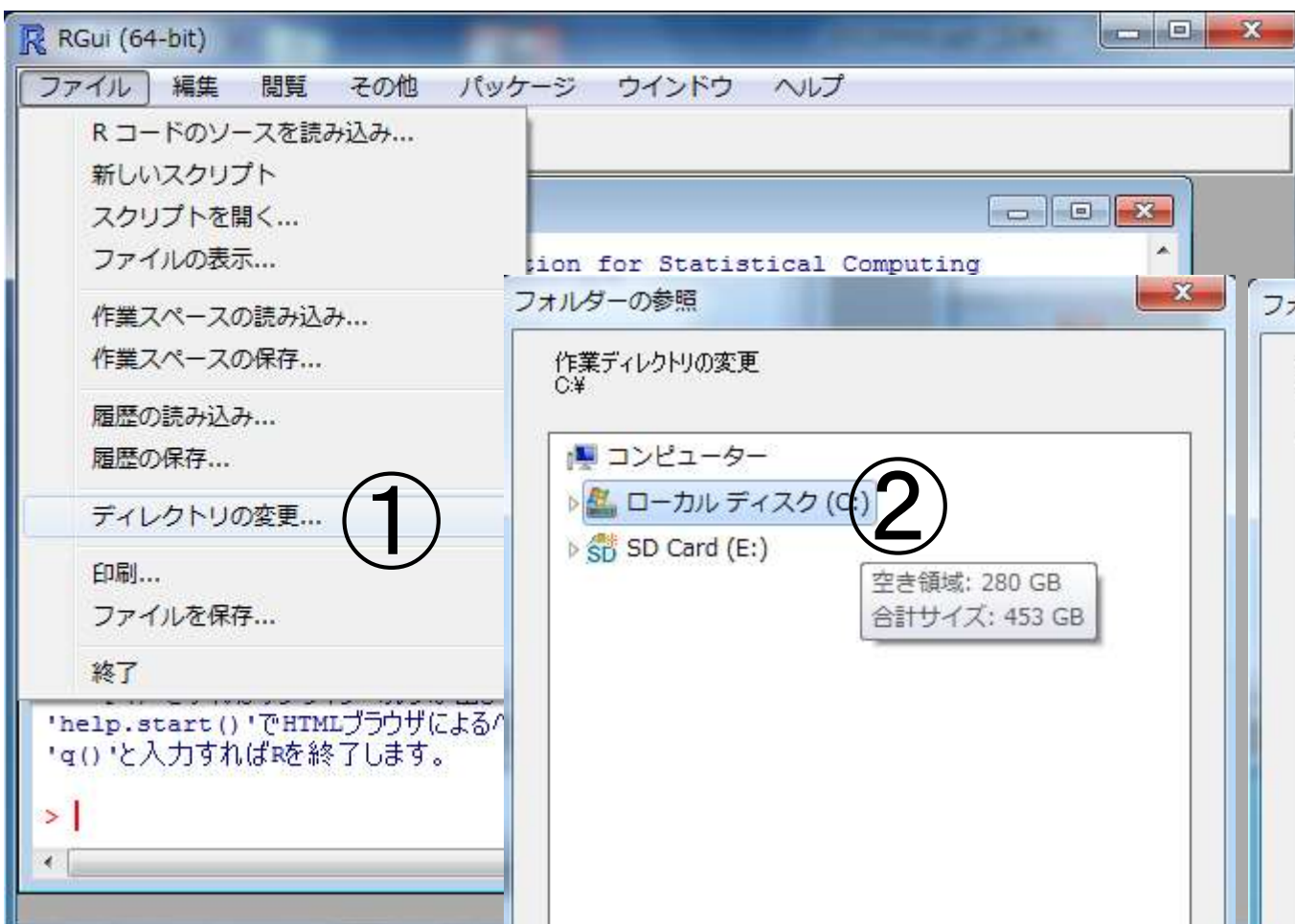
DEG同定結果として、gene_1~gene_2000
が上位にランキングされていれば正解！

Rの起動とhogeフォルダの作成

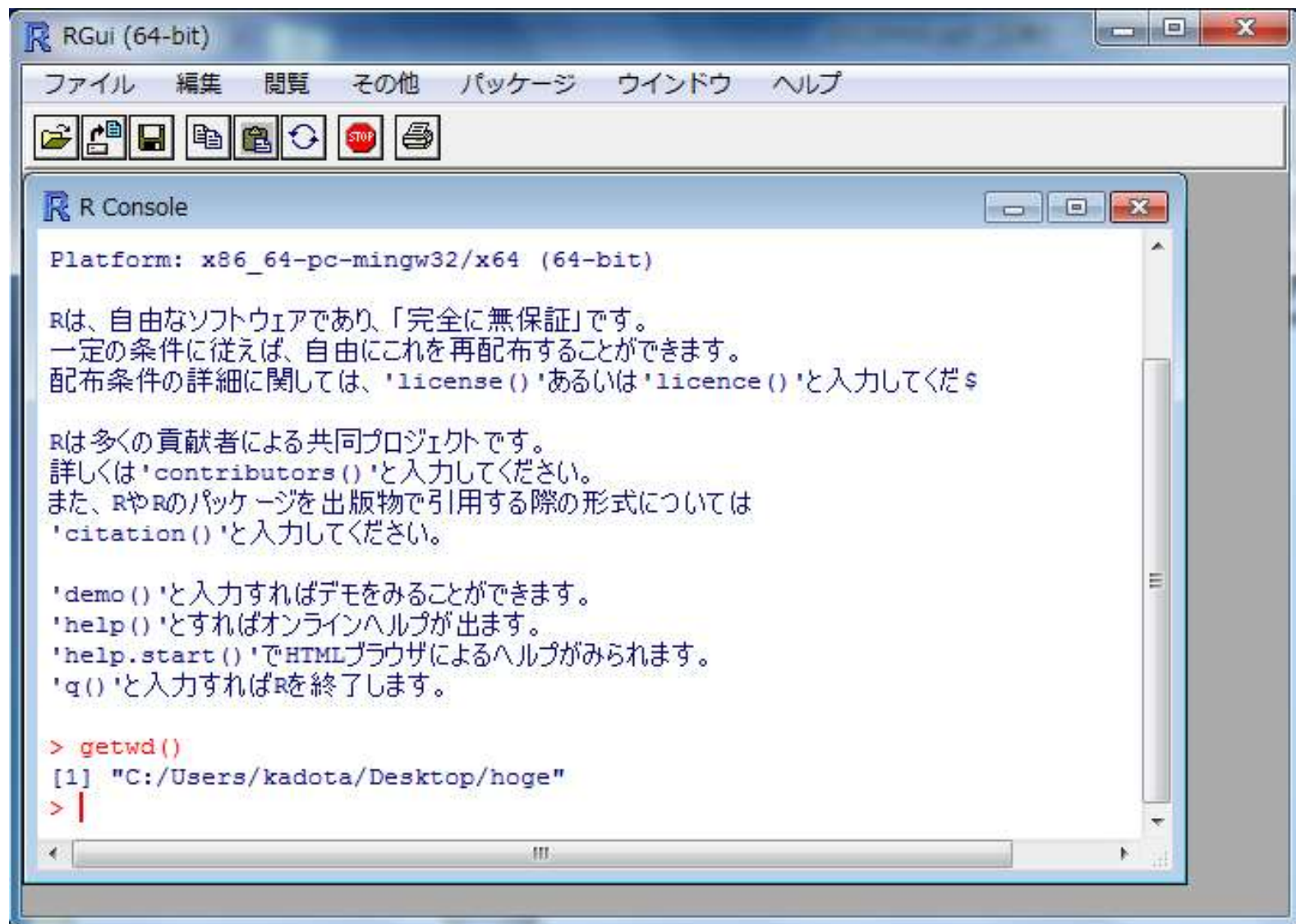


トレーニングでは、デスクトップにあるhogeフォルダ中のファイルを解析するやり方として説明します

作業ディレクトリの変更



getwd()と打ち込んで確認



```
RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ
[Icons: File Explorer, Print, Save, Copy, Paste, Refresh, Stop, Print]

R Console
Platform: x86_64-pc-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()'と入力してくださ

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"
> |
```


(Rで)塩基配列解析(主にNGSやRNA-seq解析)

(last modified 2014/02/26, since 2010)

What's new?

- 一連の解析パイプライン(RNA-seqデータ取得 -> マッピング -> 発現変動解析および M-A plot 描画まで)のクラスタリング部分
- 2014年3月17-19日に九州大学にて、ワークショップ(よく)が催されます。私は3日目(3/19, 13:00-16:30)を担当します
- 項目名の整理を行っています。3C (Hi-C)やBS-seq周辺
- 発現変動解析用Rパッケージ **TCC** (ver. 1.2.0; [Sun et al.](#)) を利用したい方は、R (ver. 3.0.2)をインストールしたのち
- どのブラウザからでもエラーなく見られる([W3C validation](#) (2013/07/30))
- 2013年7月29日まで公開していた以前の「(Rで)塩基配列解析」です(110MB程度)。(2013/07/30)

- [はじめに](#) (last modified 2014/01/30)
- [Rのインストールと起動](#) (last modified 2013/09/27)
- [サンプルデータ](#) (last modified 2014/02/20) **NEW**
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2013/08/29)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2013/08/29)
- イントロ | 一般 | [任意のキーワードを含む行を抽出\(基礎\)](#) (last modified 2013/08/29)
- イントロ | 一般 | [ランダムな塩基配列を生成](#) (last modified 2013/08/29)
- イントロ | 一般 | [任意の長さの可能な全ての塩基配列を生成](#) (last modified 2013/08/29)
- イントロ | 一般 | [任意の位置の塩基を置換](#) (last modified 2013/08/29)
- イントロ | 一般 | [指定した範囲の配列を取得](#) (last modified 2013/08/29)
- イントロ | 一般 | [翻訳配列\(translate\)を取得](#) (last modified 2013/08/29)

赤矢印部分をクリック

解析	一般	GC含量 (GC contents) (last modified 2013/06/24)
解析	一般	Sequence logos (Schneider 1990) (last modified 2012/06/27)
解析	一般	上流配列解析 LDSS (Yamamoto 2007) (last modified 2012/07/17)
解析	一般	上流配列解析 Relative Appearance Ratio (Yamamoto 2011) (last modified 2012/07/17)
解析	基礎	平均分散プロット (Technical replicates) (last modified 2014/02/18) NEW
解析	基礎	平均分散プロット (Biological replicates) (last modified 2014/02/21) NEW
解析	クラスタリング	クラスタリング について (last modified 2014/02/05) NEW
解析	クラスタリング サンプル間	hclust (last modified 2014/02/21) NEW
解析	クラスタリング 遺伝子間	MBCluster Seq (Si 2014) (last modified 2014/02/05) NEW
解析	発現変動	ポアソン分布 シミュレーションデータ (Technical replicates) (last modified 2011/09/16)
解析	発現変動	負の二項分布 シミュレーションデータ (Biological replicates) (last modified 2013/07/01)
解析	発現変動	 について (last modified 2013/08/29)
解析	発現変動	2群間 対応なし について (last modified 2013/08/29)
解析	発現変動	2群間 対応なし 複製あり TCC (Sun 2013) (last modified 2014/02/21) 推奨 NEW
解析	発現変動	2群間 対応なし 複製あり edgeR (Robinson 2010) (last modified 2014/01/30)
解析	発現変動	2群間 対応なし 複製あり SAMseq (Li 2013) (last modified 2014/01/30)
解析	発現変動	2群間 対応なし 複製なし TCC (Sun 2013) (last modified 2014/02/20) 推奨 NEW
解析	発現変動	2群間 対応なし 複製なし DESeq (Anders 2010) (last modified 2014/01/30)
解析	発現変動	2群間 対応なし BitSeq (Glaus 2012) (last modified 2013/01/08)
解析	発現変動	2群間 対応なし DSS (Wu 2012) (last modified 2013/01/18)
解析	発現変動	2群間 対応なし NOISeq (Tarazona 2011) (last modified 2013/01/08)
解析	発現変動	2群間 対応なし NBPSeq (Di 2011) (last modified 2012/03/15)
解析	発現変動	2群間 対応あり について (last modified 2013/08/29)
解析	発現変動	2群間 対応あり 複製なし TCC (Sun 2013) (last modified 2014/02/07) 推奨 NEW
解析	発現変動	2群間 対応あり 複製なし edgeR (Robinson 2010) (last modified 2014/01/07)
解析	発現変動	3群間 対応なし について (last modified 2013/08/29)
解析	発現変動	3群間 対応なし 複製あり TCC (Sun 2013) (last modified 2014/02/04) 推奨 NEW
解析	発現変動	3群間 対応なし 複製あり edgeR (Robinson 2010) (last modified 2013/09/16)
解析	発現変動	時系列データ Bayesian model-based clustering (Nascimento 2012) (last modified 2012/07/01)
解析	選択的スプライシング	 について (last modified 2014/02/04) NEW



解析 | 発現変動 | について (last modified 2013/08/29)

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [TCC \(Sun_2013\)](#) (last modified 2013/08/29)

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [TCC \(Sun_2013\)](#) (last modified 2014/02/04) 推奨 NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [edgeR \(Robinson et al. 2010\)](#) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [SAMseq \(Li 2013\)](#) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [TCC \(Sun_2013\)](#) (last modified 2014/02/04) 推奨 NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [TCC \(Sun_2013\)](#) NEW

TCCを用いたやり方を示します。

内部的に [iDEGES/edgeR \(Sun_2013\)](#) 正規化を実行したのち、[edgeR](#) パッケージ中の exact test で発現変動遺伝子 (Differentially expressed Genes; DEGs) 検出を行っています。TCC 原著論文中の iDEGES/edgeR-edgeR という解析パイプラインに相当します。全て [TCC](#) パッケージ ([Sun et al., BMC Bioinformatics, 2013](#)) 内で完結します。

「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. サンプルデータ13の10,000 genes×6 samplesのカウントデータ([data_hypodata_3vs3.txt](#))の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)の10,000個の遺伝子 (gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現)、gene_10000までがnon-DEGであることが既知です。

```

in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を指定
out_f1 <- "hoge1.txt" #出力ファイル名を指定
out_f2 <- "hoge1.png" #出力ファイル名を指定
param_G1 <- 3 #G1群のサンプル数
param_G2 <- 3 #G2群のサンプル数
param_FDR <- 0.05 #DEG検出率
param_fig <- c(400, 380) #ファイル名とサイズ

```

```

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

```

```

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

```

```

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2)) #G1群を1, G2群を2と指定
tcc <- new("TCC", data, data.cl) #TCCクラスオブジェクトの作成

```

```

#本番(iDEGES/edgeR正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="exact",
                      iteration=3, FDR=0.1, floorPDEG=1e-5)

```

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [TCC \(Sun_2013\)](#)

- 開く(O)
- 新しいタブで開く(W)
- 新しいウィンドウで開く(N)
- 対象をファイルに保存(A)...
- 対象を印刷(P)
- 切り取り

解析したいファイル [data_hypodata_3vs3.txt](#) をデスクトップ上の hoge フォルダに保存しましょう

- Bing で翻訳
- 電子メール (Windows Live Hotmail)
- すべてのアクセラレータ
- 要素の検査(L)
- お気に入りに追加(F)...
- Send to OneNote
- プロパティ(R)

実際のhogeフォルダとR操作画面の関係

ファイル保存前

ファイル保存後

The screenshot shows a Windows Explorer window with the address bar set to `C:\Users\kadota\Desktop\hoge`. The folder is empty, with the message "このフォルダは空です。" (This folder is empty). Below it, the RGui (64-bit) window is open. The R Console shows the following commands and output:

```
[以前にセーブされたワークスペースを復帰します]
> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
character(0)
> |
```

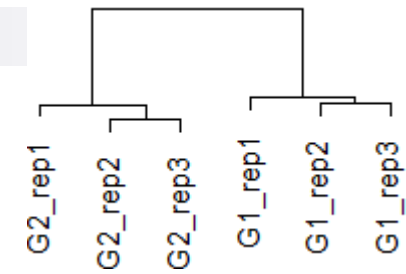
The screenshot shows the same Windows Explorer window, but now it contains a file named `data_hypodata_3vs3.txt` with a modification date of 2014/03/03 13:46. The RGui window below shows the following commands and output:

```
> getwd()
[1] "C:/Users/kadota/Desktop/hoge"
> list.files()
character(0)
> list.files()
[1] "data_hypodata_3vs3.txt"
> |
```

当たり前ですが、解析したいディレクトリ(またはフォルダ)を正しく指定できていなければエラーに遭遇します。また、解析したいファイルが存在しない状態でもエラーが出ます



目的をおさらい



■ data_hypodata_3vs3.txt (2群間比較用)

- G1群:3サンプル、G2群:3サンプル
- 全部で10,000行×6列。最初の2,000行分が発現変動遺伝子 (DEG)

	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene_1	36	56	144	2	1	0
gene_2	84	152	124	52	37	28
gene_3	592	840	800	151	257	200
gene_4	0	8	4	1	1	3
gene_5	32	32	0	1	1	0
...						
gene_1801	34	86	24	284	180	364
gene_1802	5	1	3	0	160	24
gene_1803	57	56	51	248	192	220
gene_1804	29	25	32	128	204	160
gene_1805	42	29	44	184	156	92
...						
gene_2001	4	8	9	13	12	4
gene_2002	88	139	40	22	44	21
gene_2003	933	667	462	889	396	443
gene_2004	48	37	14	36	57	71
gene_2005	290	338	553	319	210	504
...						
gene_9996	107					
gene_9997	145					
gene_9998	42					
gene_9999	5	1	2	3	4	11
gene_10000	2	4	5	2	0	0

DEG

G1で高発現

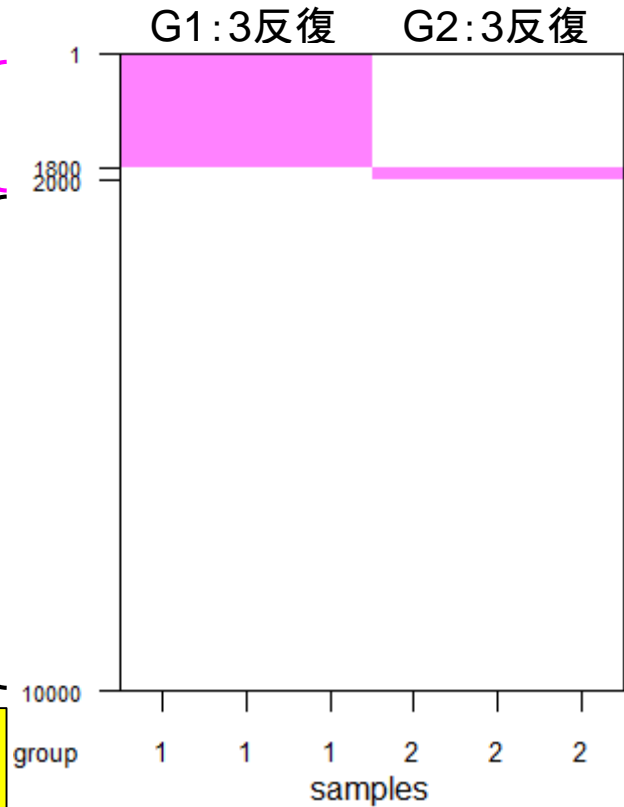
G2で高発現

non-DEG

DEG

non-DEG

TCCパッケージを用いて、複製あり2群間比較を行い、DEG同定結果を得る



基本はコピー

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) **NEW**

2013年7月以降のリニューアルで、コードのコピーがやりやすくなっています。**CTRLとALTキー**を押しながらコードの枠内で左クリックすると、全選択できます。

TCCを用いたやり方を示します。

内部的にiDEGES/edgeR(Sun_2013)正規化を実行したのち、edgeRパッケージ中のexact testで発現変動遺伝子(Differentially expressed Genes; DEGs)検出を行っています。TCC原著論文でのiDEGES/edgeR-edgeRという解析パイプラインに相当します。全てTCCパッケージ(Sun et al.)「ファイル」-「ディレクトリの変更」で解析したいファイルを選択

1. サンプルデータ13の10,000 genes×6 samplesのカウント

Biological replicatesを模倣したシミュレーションデータ(Gene 10000までがDEG (最初の1800個がG1群で高発現、gene_10000までがnon-DEGであることが既知です。)

```
in_f <- "data hypodata 3vs3.txt"
out_f1 <- "hogel1.txt"
out_f2 <- "hogel1.png"
param_G1 <- 3
param_G2 <- 3
param_FDR <- 0.05
param_fig <- c(400, 380)
```

```
#必要なパッケージをロード
library(TCC)
```

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

```
#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2))
tcc <- new("TCC", data, data.cl)
```

```
#本番(iDEGES/edgeR正規化)
```

```
tcc <- calcNormFactors(tcc, norm.method="tcc", iteration=3, FDR=0.05)
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ

R Console

Platform: x86_64-pc-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()'と入力してください。

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。

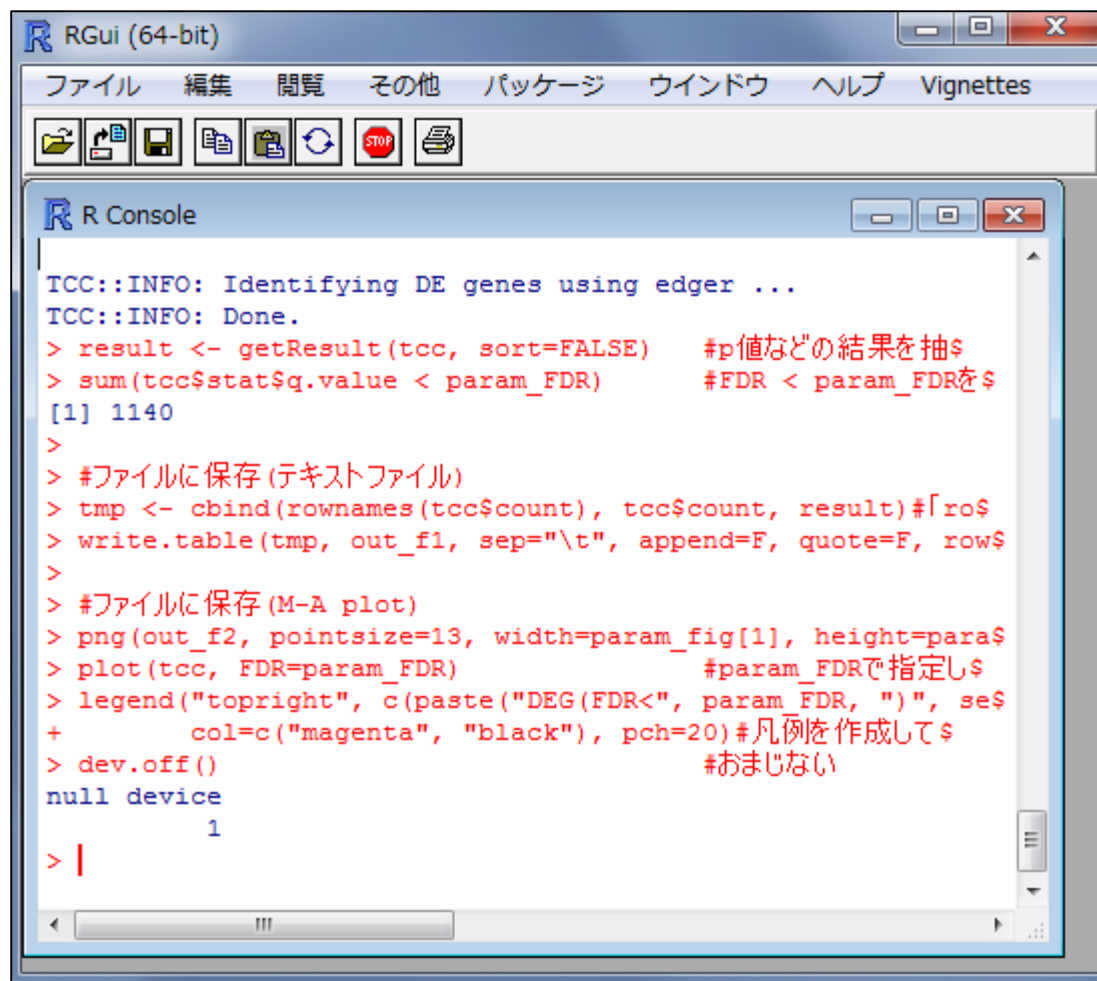
'demo()'と入力すればデモをみることができます。'help()'とすればオンラインヘルプが出ます。'help.start()'でHTMLブラウザによるヘルプが表示されます。'q()'と入力すればRを終了します。

> getwd()
[1] "C:/Users/kadota/Desktop/hogel1.txt"
> |

コピー Ctrl+C
ペースト Ctrl+V
コマンドのみペースト
コピー&ペースト Ctrl+X
ウィンドウの消去 Ctrl+L
全て選択
✓ ハッファに出力 Ctrl+W

- ①一連のコマンド群をコピーして
- ②R Console画面上でペースト

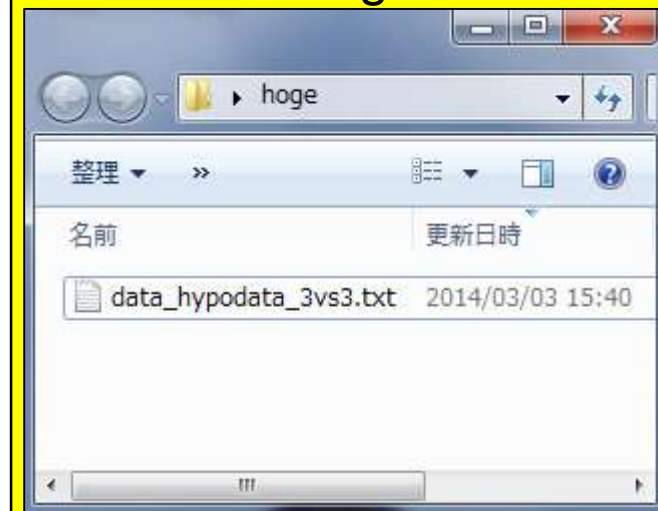
実行結果



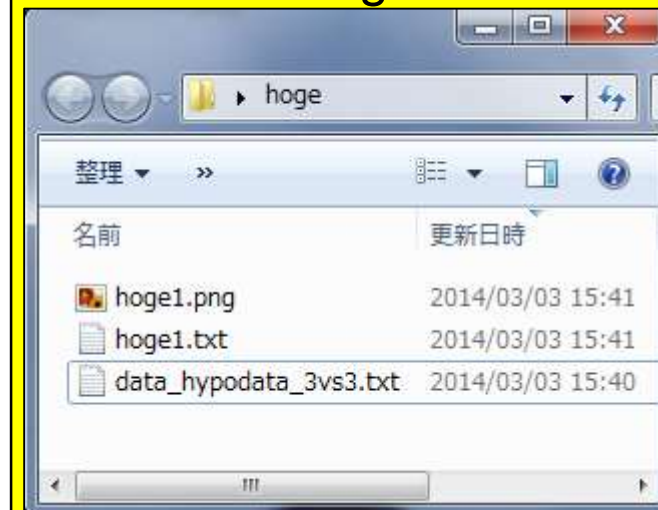
```
RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ Vignettes

R Console
TCC::INFO: Identifying DE genes using edgeR ...
TCC::INFO: Done.
> result <- getResult(tcc, sort=FALSE) #p値などの結果を抽$
> sum(tcc$stat$q.value < param_FDR) #FDR < param_FDRを$
[1] 1140
>
> #ファイルに保存(テキストファイル)
> tmp <- cbind(rownames(tcc$count), tcc$count, result)#「ro$
> write.table(tmp, out_f1, sep="\t", append=F, quote=F, row$
>
> #ファイルに保存(M-A plot)
> png(out_f2, pointsize=13, width=param_fig[1], height=para$
> plot(tcc, FDR=param_FDR) #param_FDRで指定し$
> legend("topright", c(paste("DEG(FDR<", param_FDR, ")"), se$
+ col=c("magenta", "black"), pch=20)#凡例を作成して$
> dev.off() #おまじない
null device
      1
> |
```

実行前のhogeフォルダ



実行後のhogeフォルダ



指定したパラメータ解説

同一群のサンプルがまとまっているという前提です

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun 2013) NEW

TCCを用いたやり方を示します。

内部的にiDEGES/edgeR(Sun 2013)正規化を実行したのち、edgeRパッケージ中の(Differentially expressed Genes; DEGs)検出を行っています。TCC原著論文でのiDパイプラインに相当します。全てTCCパッケージ(Sun et al., BMC Bioinformatics, 2013)「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てください。

1. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata)

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)のcount data (gene_1からgene_10000まで)がDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現)であることが既知です。

```
in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を指定し
out_f1 <- "hoge1.txt" #出力ファイル名を指定し
out_f2 <- "hoge1.png" #出力ファイル名を指定し
param_G1 <- 3 #G1群のサンプル数を指定
param_G2 <- 3 #G2群のサンプル数を指定
param_FDR <- 0.05 #DEG検出時のfalse disc
param_fig <- c(400, 380) #ファイル出力時の横幅と縦幅

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

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

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2)) #G1群を1、G2群を2
tcc <- new("TCC", data, data.cl) #TCCクラスオブジェクトの作成

#本番(iDEGES/edgeR正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="
iteration=3, FDR=0.1, floorPDEG=0.05)
```

入力ファイル: data_hypodata_3vs3.txt

G1群: 3サンプル

G2群: 3サンプル

	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene_1	36	56	144	2	1	0
gene_2	84	152	124	52	37	28
gene_3	592	840	800	151	257	200
gene_4	0	8	4	1	1	3
gene_5	32	32	0	1	1	0
...						
gene_1801	34	86	24	284	180	364
gene_1802	5	1	3	0	160	24
gene_1803	57	56	51	248	192	220
gene_1804	29	25	32	128	204	160
gene_1805	42	29	44	184	156	92
...						
gene_2001	4	8	9	13	12	4
gene_2002	88	139	40	22	44	21
gene_2003	933	667	462	889	396	443
gene_2004	48	37	14	36	57	71
gene_2005	290	338	553	319	210	504
...						
gene_9996	107	67	104	35	65	45
gene_9997	145	220	120	80	95	156
gene_9998	42	73	67	62	44	37
gene_9999	5	1	2	3	4	11
gene_10000	2	4	5	2	0	0

想定外の入力ファイル例

同一群のサンプルがまとまっていないものはダメ!

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

TCCを用いたやり方を示します。

内部的にiDEGES/edgeR(Sun 2013)正規化を実行したのち、edgeRパッケージ中の(Differentially expressed Genes; DEGs)検出を行っています。TCC原著論文でのiDパイプラインに相当します。全てTCCパッケージ(Sun et al., BMC Bioinformatics, 2013)「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てください。

1. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata)

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)のcount data (gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で低発現)、gene_10000までがnon-DEGであることが既知です。

```
in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を指定し
out_f1 <- "hoge1.txt" #出力ファイル名を指定し
out_f2 <- "hoge1.png" #出力ファイル名を指定し
param_G1 <- 3 #G1群のサンプル数を指定
param_G2 <- 3 #G2群のサンプル数を指定
param_FDR <- 0.05 #DEG検出時のfalse disc
param_fig <- c(400, 380) #ファイル出力時の横幅と高さ

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

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

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2)) #G1群を1、G2群を2
tcc <- new("TCC", data, data.cl) #TCCクラスオブジェクトの作成

#本番(iDEGES/edgeR正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="tmm",
iteration=3, FDR=0.1, floorPDEG=0.05)
```

	G1群:3サンプル			G2群:3サンプル		
	G1_rep1	G2_rep1	G1_rep2	G2_rep2	G1_rep3	G2_rep3
gene_1	36	2	56	1	144	0
gene_2	84	52	152	37	124	28
gene_3	592	151	840	257	800	200
gene_4	0	1	8	1	4	3
gene_5	32	1	32	1	0	0
...						
gene_1801	34	284	86	180	4	364
gene_1802	5	0	1	160	3	24
gene_1803	57	28	56	192	51	220
gene_1804	29	128	25	24	32	160
gene_1805	42	184	29	156	44	92
...						
gene_2001	4	13	8	12	9	4
gene_2002	88	22	139	44	40	21
gene_2003	933	889	667	26	462	443
gene_2004	48	8	37	5	14	71
gene_2005	290	319	338	210	553	504
...						
gene_9996	16	35	67	65	10	45
gene_9997	145	80	220	95	120	156
gene_9998	42	62	73	44	67	37
gene_9999	5	3	1	4	2	11
gene_10000	2	2	4	0	5	0

指定したパラメータ解説

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) N

出力ファイル名をhoge1.txtとhoge1.pngとしているので、エラーなく実行できれば指定した通りのファイルが生成されます

TCCを用いたやり方を示します。

内部的にiDEGES/edgeR(Sun_2013)正規化を実行したのち、edgeRパッケージ中のexact testで発現変動遺伝子(Differentially expressed Genes; DEGs)検出を行っています。TCC原著論文中のiDEGES/edgeR-edgeRという解析パイプラインに相当します。全てTCCパッケージ(Sun et al., BMC Bioinformatics, 2013)内で完結します。「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. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1~gene_2000までがDEG(最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

```
in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge1.txt" #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge1.png" #出力ファイル名を指定してout_f2に格納
param_G1 <- 3 #G1群のサンプル数を指定
param_G2 <- 3 #G2群のサンプル数を指定
param_FDR <- 0.05 #DEG検出時のfalse discovery rate (FDR)閾値を
param_fig <- c(400, 380) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

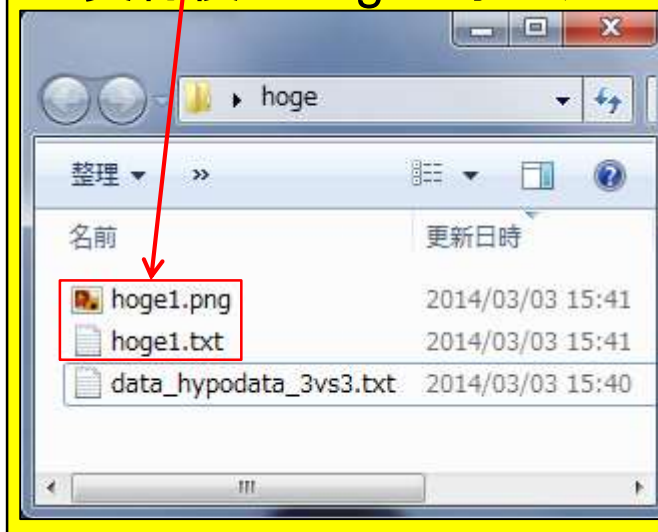
#必要なパッケージをロード
library(TCC) #パッケージの読み込み

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

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2))#G1群を1、G2群を2としたベクトルdata.
tcc <- new("TCC", data, data.cl) #TCCクラスオブジェクトtccを作成

#本番(iDEGES/edgeR正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger",#正規化を実行した
iteration=3, FDR=0.1, floorPDEG=0.05)#正規化を実行した結果をt
```

実行後のhogeフォルダ



指定したパラメータ解説

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

指定したFDR閾値を満たすものが
DEGとしてマゼンタ色になります

TCCを用いたやり方を示します。

内部的にiDEGES/edgeR(Sun 2013)正規化を実行したのち、edgeRパッケージ中のexact testで発現変動遺伝子(Differentially expressed Genes; DEGs)検出を行っています。TCC原著論文中のiDEGES/edgeR-edgeRという解析パイプラインに相当します。全てTCCパッケージ(Sun et al., BMC Bioinformatics, 2013)内で完結します。

「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

1. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_2000までがDEG(最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_10000までがnon-DEGであることが既知です。

```
in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge1.txt" #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge1.png" #出力ファイル名を指定してout_f2に格納
param_G1 <- 3 #G1群のサンプル数を指定
param_G2 <- 3 #G2群のサンプル数を指定
param_FDR <- 0.05 #DEG検出時のfalse discovery rateを指定
param_fig <- c(400, 380) #ファイル出力時の横幅と縦幅を指定

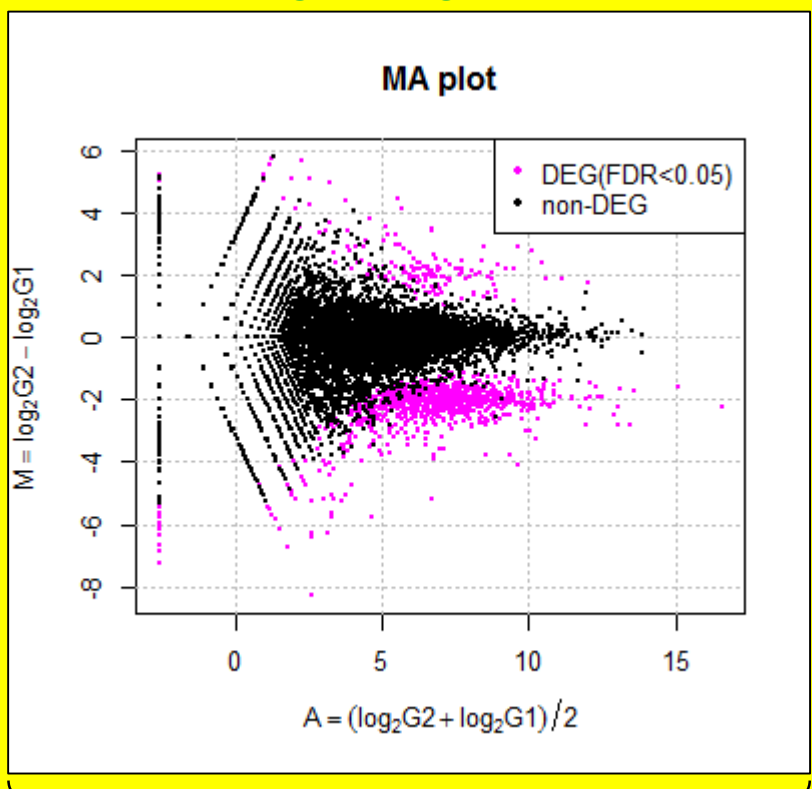
#必要なパッケージをロード
library(TCC) #パッケージの読み込み

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

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2)) #G1群を1、G2群を2としたベクトルを作成
tcc <- new("TCC", data, data.cl) #TCCクラスオブジェクトtccを作成

#本番(iDEGES/edgeR正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger", #正規化を実行
iteration=3, FDR=0.1, floorPDEG=0.05) #正規化を実行
```

M-A plot (hoge1.png)



380

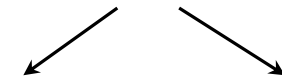
400

実行結果

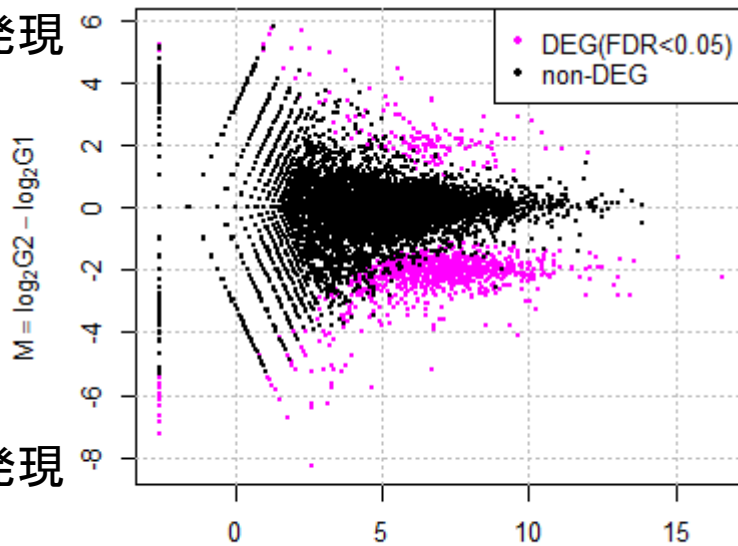
TCCを用いたDEG同定結果ファイル(hoge1.txt)

rowname	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
gene_1	36	56	144	2	1	0	gene_1	3.15	-6.26	1.72E-10	1.14E-07	15	1
gene_2	84	152	124	52	37	28	gene_2	6.10	-1.59	6.58E-04	7.31E-03	900	1
gene_3	592	840	800	151	257	200	gene_3	8.60	-1.84	2.91E-06	6.81E-05	427	1
gene_4	0	8	4	1	1	3	gene_4	1.37	-1.23	4.81E-01	1.00E+00	4523	0
gene_5	32	32	0	1	1	0	gene_5	1.92	-4.97	3.28E-03	3.10E-02	1060	1
gene_6	4	0	24	4	10	0	gene_6	2.72	-0.96	5.54E-01	1.00E+00	5047	0
gene_7	344	240	236	76	67	71	gene_7	7.13	-1.90	1.52E-06	4.15E-05	367	1
gene_8	1264	784	1060	212	183	179	gene_8	8.80	-2.40	5.79E-10	2.26E-07	25	1
gene_9	92	88	84	21	22	33	gene_9	5.56	-1.76	3.68E-04	4.37E-03	841	1

p-valueとその順位



G2群で高発現



G1群で高発現

M-A plotのA値とM値

q-value

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

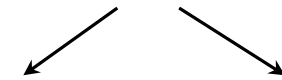
真のDEGはgene_1 ~ gene_2000。右端のestimatedDEG列が1のものが多いため結果としては妥当

実行結果

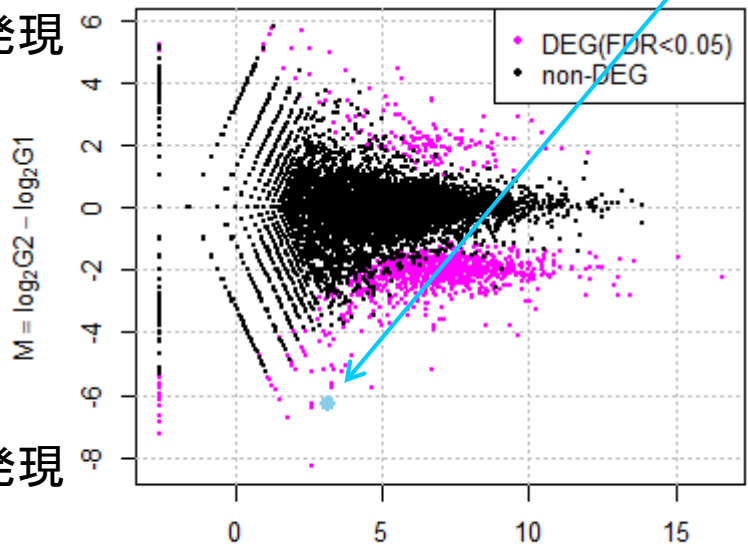
TCCを用いたDEG同定結果ファイル(hoge1.txt)

rowname	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
gene_1	36	56	144	2	1	0	gene_1	3.15	-6.26	1.72E-10	1.14E-07	15	1
gene_2	84	152	124	52	37	28	gene_2	6.10	-1.59	6.58E-04	7.31E-03	900	1
gene_3	592	840	800	151	257	200	gene_3	8.60	-1.84	2.91E-06	6.81E-05	427	1
gene_4	0	8	4	1	1	3	gene_4	1.37	-1.23	4.81E-01	1.00E+00	4523	0
gene_5	32	32	0	1	1	0	gene_5	1.92	-4.97	3.28E-03	3.10E-02	1060	1
gene_6	4	0	24	4	10	0	gene_6	2.72	-0.96	5.54E-01	1.00E+00	5047	0
gene_7	344	240	236	76	67	71	gene_7	7.13	-1.90	1.52E-06	4.15E-05	367	1
gene_8	1264	784	1060	212	183	179	gene_8	8.80	-2.40	5.79E-10	2.26E-07	25	1
gene_9	92	88	84	21	22	33	gene_9	5.56	-1.76	3.68E-04	4.37E-03	841	1

p-valueとその順位



G2群で高発現



G1群で高発現

M-A plotのA値とM値

q-value

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

gene_1はDEGと判定されているが、M-A plot上で眺めても妥当

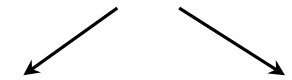
$$A = (\log_2 G2 + \log_2 G1) / 2$$

実行結果

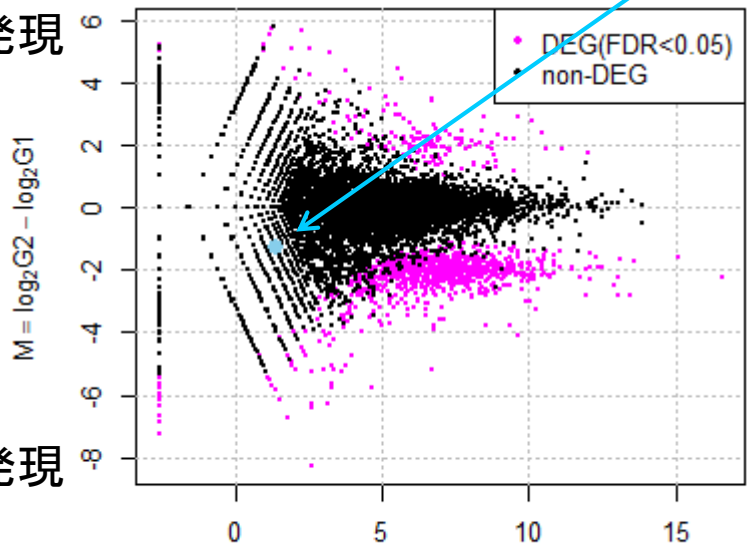
TCCを用いたDEG同定結果ファイル(hoge1.txt)

rowname	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	gene_id	a.value	m.value	p.value	qvalue	rank	estimatedDEG
gene_1	36	56	144	2	1	0	gene_1	3.15	-6.26	1.72E-10	1.14E-07	15	1
gene_2	84	152	124	52	37	28	gene_2	6.10	-1.59	6.58E-04	7.31E-03	900	1
gene_3	592	840	800	151	257	200	gene_3	8.60	-1.84	2.91E-06	6.81E-05	427	1
gene_4	0	8	4	1	1	3	gene_4	1.37	-1.23	4.81E-01	1.00E+00	4523	0
gene_5	32	32	0	1	1	0	gene_5	1.92	-4.97	3.28E-03	3.10E-02	1060	1
gene_6	4	0	24	4	10	0	gene_6	2.72	-0.96	5.54E-01	1.00E+00	5047	0
gene_7	344	240	236	76	67	71	gene_7	7.13	-1.90	1.52E-06	4.15E-05	367	1
gene_8	1264	784	1060	212	183	179	gene_8	8.80	-2.40	5.79E-10	2.26E-07	25	1
gene_9	92	88	84	21	22	33	gene_9	5.56	-1.76	3.68E-04	4.37E-03	841	1

p-valueとその順位



G2群で高発現



G1群で高発現

M-A plotのA値とM値

q-value

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

gene_4はnon-DEGと判定されており、false negativeに相当する。G2群で $2^{-1.23}$ (= 0.425)倍高発現、つまりG1群で2.35倍高発現ではあるが、全体的に低発現でnon-DEG分布の範囲内にすっぽり収まっているので判定結果自体は妥当

色についての説明

(Rで)塩基配列解析(主にNGS、RNA-seq、トランスクリプトーム解析)

(last modified 2014/03/16, since 2010)

What's new?

- 門田幸二 著 [シリーズ Useful R 第7巻トランスクリプトーム解析](#)が2014年4月10日に共立出版から出ます。
- [参考資料\(講義、講習会、本など\)](#)の項目を追加しました。(2014/03/16) **NEW**
- 私の所属する[アグリバイオインフォマティクス教育研究プログラム](#)では、平成26年度も(東大生に限らず)バイオインフォ関連講義を行います。受講希望者は平成26年4月7日18:00-18:45に東大農学部二号館二階化学第一講義室にて開催予定の受講ガイダンスに出席してください。例年東大以外の企業の方、研究員、学生が二割程度は受講しております。このウェブページと直接関連する講義は「[ゲノム情報解析基礎](#)」と「[農学生命情報科学特論](#)」ですが、背景理論の説明などは「[機能ゲノム学](#)」でも行います。興味ある科目のみの受講も可能ですので、お気軽にどうぞ。(2014/03/03) **NEW**
- 一連の解析パイプライン(RNA-seqデータ取得 -> マッピング -> カウントデータやRPKMデータ取得 -> サンプル間クラスターリングや発現変動解析およびM-A plot描画まで)のクラスターリング部分をアップデートしました。項目名の一番下のほうです。(2014/02/26) **NEW**
- 2014年3月17-19日に九州大学にて、ワークショップ([よく分かる次世代シーケンサー解析～最先端トランスクリプトーム解析～](#))が開催されます。私は3日目(3/19, 13:00-16:30)を担当します。興味ある方はどうぞ。締切は確か2/21です。(2014/02/17) **NEW**
- 発現変動解析用Rパッケージ [TCC](#) (ver. 1.2.0; [Sun et al., BMC Bioinformatics, 2013](#))がBioconductorよりリリースされました。最新版を利用したい方は、R (ver. 3.0.2)をインストールしたのち、Bioconductor (ver. 2.13)をインストールしてください。(2013/10/17)

[はじめに](#) (last modified 2014/01/30)

- [参考資料\(講義、講習会、本など\)](#) (last modified 2014/03/16) **NEW**
- [過去のお知らせ](#) (last modified 2014/03/03) **NEW**
- [Rのインストールと起動](#) (last modified 2013/09/27)
- [サンプルデータ](#) (last modified 2014/03/05) **NEW**
- イントロ | 一般 | [ランダムに行を抽出](#) (last modified 2013/10/10)
- イントロ | 一般 | [任意の文字列を行の最初に挿入](#) (last modified 2013/10/10)
- イントロ | 一般 | [任意のキーワードを含む行を抽出\(基礎\)](#) (last modified 2014/02/06)
- Mar 19 2014 | 一般 | [ランダムな塩基配列を生成](#) (last modified 2013/09/29)

コメント

特にやらなくてもいいコマンド
プログラム実行時に目的に応じて変更すべき箇所

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | edgeR (Robinson et al 2010) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq (Li 2013) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013)

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

TCCを用いたやり方を示します。

内部的にiDEGES/edgeR(Sun 2013)正規化を実行したのち、edgeRパッケージ中のexact testで発現変動遺伝子(Differentially Expressed Genes; DEGs)検出を行っています。TCC原著論文でのiDEGES/edgeR-edgeRという解析パイプラインは、まずTCC (Sun 2013)で正規化を行い、edgeR (Robinson et al 2010)で検出を行います。

2. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1～gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001～gene_10000までがnon-DEGであることが既知です。
正規化後のテキストファイルデータを出力し、平均-分散プロットのpngファイルを出力しています。

1. サンプルデータ

Biological replicates
gene_2000までがDEG
gene_10000までがnon-DEG

```
in_f <- "data_hypodata_3vs3.txt"
out_f1 <- "hoge2.txt"
out_f2 <- "hoge2.png"
param_G1 <- 3
param_G2 <- 3
param_FDR <- 0.05
param_fig <- c(380, 420)

#必要なパッケージをロード
library(TCC)

#入力ファイル
```

```
in_f <- "data_hypodata_3vs3.txt"
out_f1 <- "hoge2.txt"
out_f2 <- "hoge2.png"
param_G1 <- 3
param_G2 <- 3
param_FDR <- 0.05
param_fig <- c(380, 420)

#必要なパッケージをロード
library(TCC)

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

#入力ファイル名を指定してin_fに格納
#出力ファイル名を指定してout_f1に格納
#出力ファイル名を指定してout_f2に格納
#G1群のサンプル数を指定
#G2群のサンプル数を指定
#DEG検出時のfalse discovery rate (FDR)閾値を
#ファイル出力時の横幅と縦幅を指定(単位はピクセル)

#パッケージの読み込み

1つの項目内でも様々な例題を提供しています

コード内のコピーは
CTRL + ALT + 左クリック

```
data.cl <- list(param_G1, rep(2, param_G2))#G1群を1、G2群を2としたベクトルdata.cl
tcc <- TCC::TCC(data, data.cl) #TCCクラスオブジェクトtccを作成

#正規化を実行した結果をtccに格納
norm <- edgeR::exact(tcc, norm.method="tmm", test.method="edgeR", #正規化を実行した結果をtccに格納
iteration=3, FDR=0.1, floorPDEG=0.05)#正規化を実行した結果をtccに格納
normalized <- edgeR::normalizedData(tcc) #正規化後のデータを取り出してnormalizedに格納
```



1. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | [TCC \(Sun 2013\)](#)

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1~gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

```
in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を指定してin_f1に格納
out_f1 <- "hoge1.txt" #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge1.png" #出力ファイル名を指定してout_f2に格納
```

rowname	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
gene_1	36	56	144	2	1	0	gene_1	3.148	-6.26	1.7E-10	1.1E-07	15	1
gene_2	84	152	124	52	37	28	gene_2	6.097	-1.59	0.00066	0.00731	900	1
gene_3	592	840	800	151	257	200	gene_3	8.602	-1.84				1
gene_4	0	8	4	1	1	3	gene_4	1.371	-1.23				0
gene_5	32	32	0	1	1	0	gene_5	1.917	-4.97				1
gene_6	4	0	24	4	10	0	gene_6	2.722	-0.96	0.55429		1 5047	0
gene_7	344	240	236	76	67	71	gene_7	7.125	-1.9	1.5E-06	4.2E-05	367	1

1.では正規化前の入力データを出力させましたが...

2. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1~gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

正規化後のテキストファイルデータを出力し、平均-分散プロットのpngファイルを出力しています。

```
in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を指定してin_f1に格納
out_f1 <- "hoge2.txt" #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge2.png" #出力ファイル名を指定してout_f2に格納
```

rowname	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
gene_1	35.4	55.7	141.9	2.0	1.0	0.0	gene_1	3.148	-6.26	1.7E-10	1.1E-07	15	1
gene_2	82.7	151.2	122.2	52.5	37.6	28.3	gene_2	6.097	-1.59	0.00066	0.00731	900	1
gene_3	582.9	835.5	788.2	152.6	261.0	202.1	gene_3	8.602	-1.84				1
gene_4	0.0	8.0	3.9	1.0	1.0	3.0	gene_4	1.371	-1.23				0
gene_5	31.5	31.8	0.0	1.0	1.0	0.0	gene_5	1.917	-4.97				1
gene_6	3.9	0.0	23.6	4.0	10.2	0.0	gene_6	2.722	-0.96	0.55429		1 5047	0
gene_7	338.7	238.7	232.5	76.8	68.0	71.7	gene_7	7.125	-1.9	1.5E-06	4.2E-05	367	1

2.で示すように、正規化後の数値を出力させることもできます

1. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data hypodata 3vs3.txt)の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1～gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001～gene_10000までがnon-DEGであることが既知です。

```
tcc <- new("TCC", data, data.c1) #TCCクラスオブジェクトtccを作成

#本番(iDEGES/edgeR正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edgeR", #正規化を実行した
iteration=3, FDR=0.1, floorPDEG=0.05)#正規化を実行した結果をt

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

#ファイルに保存(テキストファイル)
tmp <- cbind(rownames(tcc$count), tcc$count, result)#「rownames情報」、「カウントデ
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定し
```

2. サンプルデータ13の10,000 genes×6 samplesの正規化後のテキストファイルデータ

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1～gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001～gene_10000までがnon-DEGであることが既知です。

```
#本番(iDEGES/edgeR正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edgeR", #正規化を実行した
iteration=3, FDR=0.1, floorPDEG=0.05)#正規化を実行した結果をt
normalized <- getNormalizedData(tcc) #正規化後のデータを取り出してnormalizedに格納

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

#ファイルに保存(テキストファイル)
tmp <- cbind(rownames(tcc$count), normalized, result)#「rownames情報」、「正規化後の
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定し
```

両者の違いはこの部分

コードの中身が分かると応用範囲も拡大



2. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

対応なし | 複製あり | TCC (Sun 2013)

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1~gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

正規化後のテキストファイルデータを出力し、平均-分散プロットのpngファイルを出力しています。

#本番(iDEGES/edgeR正規化)

```
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger", #正規化を実行した  
                      iteration=3, FDR=0.1, floorPDEG=0.05) #正規化を実行した結果をt  
normalized <- getNormalizedData(tcc) #正規化後のデータを取り出してnormalizedに格納
```

#本番(DEG検出)

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

#ファイルに保存(テキストファイル)

```
tmp <- cbind(rownames(tcc$count), normalized, result) # 「rownames情報」、「正規化後の  
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F) #tmpの中身を指定し
```

hoge2.txt

rowname	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
gene_1	35.4	55.7	141.9	2.0	1.0	0.0	gene_1	3.148	-6.26	1.7E-10	1.1E-07	15	1
gene_2	82.7	151.2	122.2	52.5	37.6	28.3	gene_2	6.097	-1.59	0.00066	0.00731	900	1
gene_3	582.9	835.5	788.2	152.6	261.0	202.1	gene_3	8.602	-1.84	2.9E-06	6.8E-05	427	1
gene_4	0.0	8.0	3.9	1.0	1.0	3.0	gene_4	1.371	-1.23	0.4809		1 4523	0
gene_5	31.5	31.8	0.0	1.0	1.0	0.0	gene_5	1.917	-4.97	0.00328	0.03098	1060	1
gene_6	3.9	0.0	23.6	4.0	10.2	0.0	gene_6	2.722	-0.96	0.55			0
gene_7	338.7	238.7	232.5	76.8	68.0	71.7	gene_7	7.125	-1.9	1.5E			1

コードの中身が分かると応用範囲も拡大



2. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

対応なし | 複製あり | TCC (Sun 2013)

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1~gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

正規化後のテキストファイルデータを出力し、平均-分散プロットのpngファイルを出力しています。

#本番(iDEGES/edgeR正規化)

```
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger", #正規化を実行した
                        iteration=3, FDR=0.1, floorPDEG=0.05) #正規化を実行した結果をt
normalized <- getNormalizedData(tcc) #正規化後のデータを取り出してnormalizedに格納
```

#本番(DEG検出)

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

#ファイルに保存(テキストファイル)

```
tmp <- cbind(rownames(tcc$count), normalized, result) # 「rownames情報」、「正規化後の
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F) #tmpの中身を指定し
```

```
R Console
> head(normalized, n=2)
      G1_rep1  G1_rep2  G1_rep3  G2_rep1  G2_rep2  G2_rep3
gene_1 35.44451  55.69717 141.8786  2.020687  1.015669  0.00000
gene_2 82.70385 151.17802 122.1732 52.537871 37.579743 28.29417
> head(result, n=3)
  gene_id  a.value  m.value      p.value      q.value  rank  estimatedDEG
1  gene_1  3.148364 -6.261971 1.717077e-10 1.144718e-07   15           1
2  gene_2  6.096850 -1.588288 6.581803e-04 7.313114e-03  900           1
3  gene_3  8.601852 -1.841508 2.906231e-06 6.806161e-05  427           1
> head(rownames(tcc$count))
[1] "gene_1" "gene_2" "gene_3" "gene_4" "gene_5" "gene_6"
> head(rownames(tcc$count), n=4)
[1] "gene_1" "gene_2" "gene_3" "gene_4"
> |
```

コードの中身が分かると応用範囲も拡大



解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | edgeR (Robinson et al 2010) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq (Li 2013) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

TCCを用いたやり方を示します。

内部的にiDEGES/edgeR(Sun_2013)正規化を実行したのち、edgeRパッケージ中のexact testで発現変動遺伝子(Differentially expressed Genes; DEGs)検出を行っています。TCC原著論文でのiDEGES/edgeR-edgeRという解析パイプラインは、まずedgeRを用いて正規化されたデータから、edgeRのexact testを用いてDEGsを検出します。

7. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1~gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

1.と基本的に同じで、出力のテキストファイルが正規化前のデータではなく正規化後のデータになっていて、発現変動順にソートしたのになっています。

1. サンプルデータ

```
Biological replicates
gene_2000までがDEG
gene_10000までがnon-DEG

in_f <- "data_hypodata_3vs3.txt"
out_f1 <- "hoge7.txt"
out_f2 <- "hoge7.png"
param_G1 <- 3
param_G2 <- 3
param_FDR <- 0.05
param_fig <- c(400, 380)

#必要なパッケージをロード
library(TCC)

#入力ファイル
```

```
in_f <- "data_hypodata_3vs3.txt"
out_f1 <- "hoge7.txt"
out_f2 <- "hoge7.png"
param_G1 <- 3
param_G2 <- 3
param_FDR <- 0.05
param_fig <- c(400, 380)

#必要なパッケージをロード
library(TCC)

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

#入力ファイル名を指定してin_fに格納
#出力ファイル名を指定してout_f1に格納
#出力ファイル名を指定してout_f2に格納
#G1群のサンプル数を指定
#G2群のサンプル数を指定
#DEG検出時のfalse discovery rate (FDR)閾値を
#ファイル出力時の横幅と縦幅を指定(単位はピクセル)

発現変動順にソートした状態で出力させることもできます

コード内のコピーは
CTRL + ALT + 左クリック

```
#パッケージの読み込み
library(TCC)

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

```
#正規化後のデータを取り出してnormalizedに格納
normalized <- TCC::NormalizedData(tcc)

#正規化後のデータを取り出してnormalizedに格納
```

```
#正規化後のデータを取り出してnormalizedに格納
normalized <- TCC::NormalizedData(tcc)

#正規化後のデータを取り出してnormalizedに格納
```

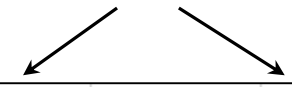


実行結果

TCCを用いたDEG同定結果ファイル(hoge7.txt)

rownames	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
gene_1211	4324.2	2100.6	3550.9	267.74	67.034	256.67	gene_1211	9.661	-4.08	2.88E-13	2.63E-09	1	1
gene_615	248.11	190.96	157.64	25.259	11.172	17.179	gene_615	5.898	-3.48	7.48E-13	2.63E-09	2	1
gene_1399	1421.7	2824.6	2963.7	299.06	142.19	92.967	gene_1399	9.354	-3.75	7.89E-13	2.63E-09	3	1
gene_833	23968	17031	25688	3239.2	3878.8	2535.4	gene_833	13.05	-2.79	1.81E-12	4.54E-09	4	1
gene_1811	20.676	27.849	46.308	484.96	349.39	210.19	gene_1811	6.713	3.461	3.08E-12	6.16E-09	5	1
gene_1760	7825.4	14720	8713.7	1369	1672.8	1334.9	gene_1760	11.93	-2.84	4.09E-12	6.82E-09	6	1
gene_1379	964.88	767.83	748.8	139.43	87.348	141.47	gene_1379	8.316	-2.75	8.48E-12	1.21E-08	7	1

p-valueとその順位

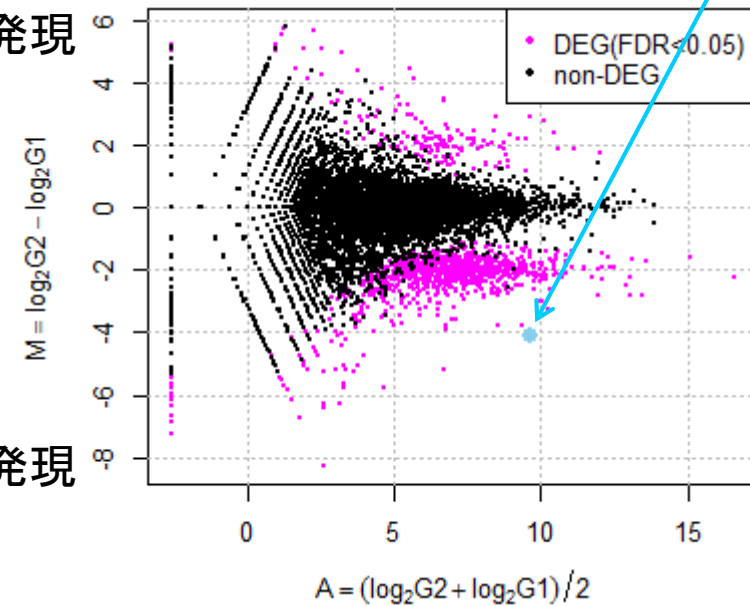


M-A plotのA値とM値

q-value

G2群で高発現

G1群で高発現



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

真のDEGはgene_1~gene_2000。上位7個は、真のDEGで占められており妥当

結果の解釈: FDRの定義をおさらい

■ TCCを用いたDEG同定結果ファイル(hoge7.txt)

rownames	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
gene_360	228.4	43.8	118.2	29.3	26.4	66.7	gene_360	6.187	-1.67	0.0055	0.0483	1138	1
gene_1153	122.1	71.6	130.1	38.4	45.7	50.5	gene_1153	6.121	-1.27	0.0056	0.0488	1139	1
gene_260	7.9	0.0	67.0	0.0	0.0	2.0	gene_260	2.036	-5.21	0.0057	0.0499	1140	1
gene_1424	31.5	23.9	23.6	3.0	8.1	5.1	gene_1424	3.576	-2.29	0.0057	0.0501	1141	0
gene_1798	3.9	35.8	0.0	0.0	0.0	0.0	gene_1798	-2.59	-5.3	0.0057	0.0501	1142	0
gene_894	220.5	99.5	161.6	37.4	102.6	16.2	gene_894	6.514	-1.63	0.0057	0.0502	1143	0
gene_1635	90.6	11.9	59.1	9.1	11.2	17.2	gene_1635	4.697	-2.11	0.0057	0.0502	1144	0

p 値の定義から、10,000遺伝子 \times 0.0057 = 57個分の真のnon-DEGをDEGと判定ミスするのを許容することに相当



$p < 0.0057$ を満たす1,140個の中に占める偽物の割合は $57/1,140 = 0.05$ と計算することができる



これ(0.05)がFDR!!



False positiveに相当する偽物混入率を5%まで許容すると、1,140個がDEGと判定される

結果の解釈: FDRの定義をおさらい

■ TCCを用いたDEG同定結果ファイル(hoge7.txt)

rownames(G1_rep1 G1_rep2 G1_rep3 G2_rep1 G2_rep2 G2_rep3 gene_id a.value m.value p.value q.value rank estimatedDEG
gene_1303 55.1 59.7 86.7 20.2 39.6 30.3 gene_1303 5.489 -1.16 0.0288 0.1981 1453 0
gene_9031 0.0 0.0 0.0 1.0 3.0 9.1 gene_9031 -2.59 3.74 0.0290 0.1993 1454 0
gene_7976 0.0 0.0 1.0 7.1 0.0 20.2 gene_7976 0.789 4.791 0.0291 0.1999 1455 0
gene_96 204.8 194.9 169.5 131.3 67.0 109.1 gene_96 7.124 -0.89 0.0293 0.2010 1456 0
gene_8971 85.7 95.5 79.8 135.4 164.5 166.7 gene_8971 6.862 0.839 0.0294 0.2018 1457 0
gene_6333 8.9 16.9 6.9 17.2 24.4 67.7 gene_6333 4.316 1.742 0.0295 0.2025 1458 0
gene_998 11.8 15.9 3.9 3.0 0.0 1.0 gene_998 1.915 -2.97 0.0297 0.2038 1459 0

p 値の定義から、10,000遺伝子 \times 0.0291 = 291個分の真のnon-DEGをDEGと判定ミスするのを許容することに相当



$p < 0.0291$ を満たす1,455個の中に占める偽物の割合は $291/1,455 = 0.20$ と計算することができる



これ(0.20)がFDR!!



False positiveに相当する偽物混入率を20%まで許容すると、1,455個がDEGと判定される

7. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodexa_3vs3.txt)の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_10000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

1.と基本的に同じコマンドで、発現変動順にソートした結果を

コメント

特にやらなくてもいいコマンド

プログラム実行時に目的に応じて変更すべき箇所

結果ファイルを眺めて調べなくても、R Console画面上に表示されています。ウェブ上で灰色なのは特にやらなくてもいいコマンドだから

```
#本番(DEG検出)
tcc <- estm
result <- ge
sum(tcc$stat$q.value < param_FDR) #FDR < param_FDRを満たす遺伝子数を表示

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

#ファイルに保存(M-A plot)
png(out_f2, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各
plot(tcc, FDR=param_FDR) #param_FDRで指定した閾値を満たすDEGをマゼンタ
legend("topright", c(paste("DEG(FDR<", param_FDR, ")"), sep=""), "non-DEG"),#凡例を
col=c("magenta", "black"), pch=20)#凡例を作成している
dev.off() #おまじない
sum(tcc$stat$q.value < 0.05) #FDR < 0.05を満たす遺伝子数を表示
sum(tcc$stat$q.value < 0.10) #FDR < 0.10を満たす遺伝子数を表示
sum(tcc$stat$q.value < 0.20) #FDR < 0.20を満たす遺伝子数を表示
sum(tcc$stat$q.value < 0.30) #FDR < 0.30を満たす遺伝子数を表示
```

R Console

```
> sum(tcc$stat$q.value < 0.05) #FDR < 0.05を満たす遺伝子数を表示
[1] 1140
> sum(tcc$stat$q.value < 0.10) #FDR < 0.10を満たす遺伝子数を表示
[1] 1270
> sum(tcc$stat$q.value < 0.20) #FDR < 0.20を満たす遺伝子数を表示
[1] 1455
> sum(tcc$stat$q.value < 0.30) #FDR < 0.30を満たす遺伝子数を表示
[1] 1613
```

- [TCC: Sun et al., BMC Bioinformatics, 2011](#)
- [edgeR: Robinson et al., Bioinformatics, 2009](#)
- [TMM正規化法: Robinson and Oshlack, 2010](#)
- [an exact test for negative binomial distr](#)

Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- **TCC発現変動解析: 複製あり2群間比較用実データ**
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

実データ解析例: SRP017142

■ Step1: SRADBを用いたgzip圧縮FASTQ形式ファイルのダウンロード

□ Neyret-Kahn et al., *Genome Res.*, **23**: 1563-1579, 2013

約6GBで1時間程度

- 複製あり2群間比較用ヒトRNA-seqデータ(3 Ras vs. 3 Proliferative)

FileName	SampleName	
SRR616151.fastq.gz	Pro_rep1	} G1群
SRR616152.fastq.gz	Pro_rep2	
SRR616153.fastq.gz	Pro_rep3	
SRR616154.fastq.gz	Ras_rep1	} G2群
SRR616155.fastq.gz	Ras_rep2	
SRR616156.fastq.gz	Ras_rep3	

■ Step2: QuasR (Bowtie)を用いたヒトゲノムへのマッピング

計6サンプルのマッピングに10時間程度

□ *BSgenome.Hsapiens.UCSC.hg19*パッケージを利用

□ 18種類程度の生物種のゲノム配列がRパッケージとして利用可能

- シロイヌナズナの場合: *BSgenome.Athaliana.TAIR.TAIR9*

- ショウジョウバエの場合: *BSgenome.Dmelanogaster.UCSC.dm3*

実データ解析例: SRP017142

■ QuasR (Bowtie)を用いたカウント情報取得

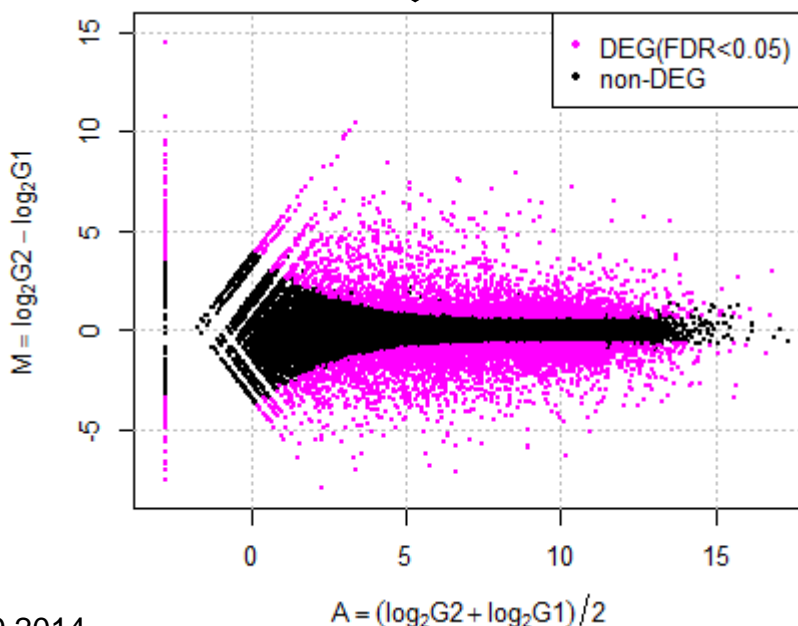
カウントデータ: [srp017142_count_bowtie.txt](#)

このファイルを入力として2群間比較解析を行う

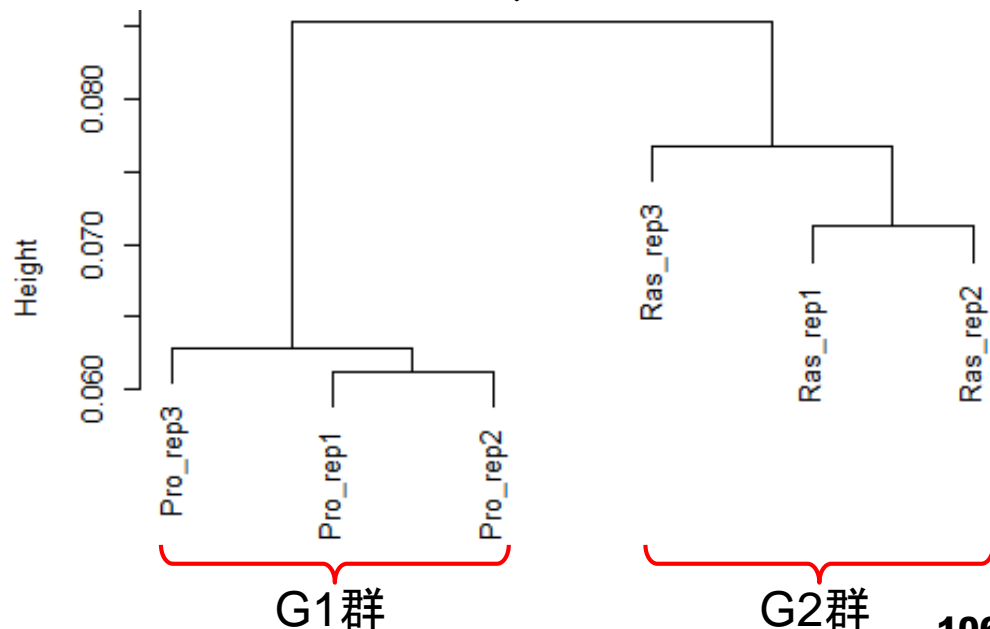
59,857 genes

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG000000000003	480	513	366	124	271	366
ENSG000000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

Step4: TCCを用いた発現変動遺伝子(DEG)同定



Step3: サンプル間クラスタリング



• 解析 | small RNA | [segmentSeq \(Hardcastle 2012\)](#) (last modified 2014/02/04) **NEW**

• 作図 | [Iについて](#) (last modified 2012/09/10)

• 作図 | [M-A plot \(基本編\)](#) (last modified 2012/10/01)

• 作図 | [M-A plot \(ggplot2編\)](#) (last modified 2013/07/30)

• 作図 | [ROC曲線](#) (last modified 2012/10/01)

• 作図 | [SplicingGraphs](#) (last modified 2013/08/07)

• [パイプライン Iについて](#) (last modified 2013/10/17)

• [パイプライン | ゲノム | 発現変動 | 2群間 | 対応なし | 複製あり | \[SRP017142 \\(Neyret-Kahn 2013\\)\]\(#\)](#)

• [パイプライン | ゲノム | small RNA | \[SRP016842 \\(Nie 2013\\)\]\(#\)](#) (last modified 2013/11/12)

• [リンク集](#) (last modified 2012/03/29)

• [パイプライン | ゲノム | 発現変動 | 2群間 | 対応なし | 複製あり | \[SRP017142 \\(Neyret-Kahn 2013\\)\]\(#\)](#)

この記載通りに行えば、公共データの発現変動解析までが一通りできますが、トレーニングでは、カウントデータ取得以降を行います。

[パイプライン | ゲノム | 発現変動 | 2群間 | 対応なし \(Neyret-Kahn 2013\)](#) **NEW**

はじめに

このページは、次世代シーで行うための一連の手続き

[Neyret-Kahn et al., Genome Res., 2013](#)の2群間比較用ヒトRNA-seqデータ (3 proliferative samples vs. 3 Ras samples)が [GSE42213](#)に登録されています。ここでは、[SRADB](#)パッケージを用いたそのFASTQ形式ファイルのダウンロードから、[QuasR](#)パッケージを用いたマッピングおよびカウントデータ取得、そして[TCC](#)パッケージを用いた発現変動遺伝子(DEG)検出までを行う一連の手順を示します。

原著論文([Neyret-Kahn et al., Genome Res., 2013](#))では72-baseと書いてますが、取得ファイルは54-baseしかありません。また、ヒトサンプルなのになぜかマウスゲノム("mm9")にマップしたと書いているのも意味不明です。ちなみに54 bpと比較的長いリードであり、原著論文でもsplice-aware alignerの一つである[TopHat \(Trapnell et al., Bioinformatics, 2009\)](#)を用いてマッピングを行ったと記述していますが、ここでは、(計算時間短縮のため)basic alignerの一つである[Bowtie](#)をQuasRの内部で用いています。

多数のファイルが作成されるので、ここでは「デスクトップ」上に「SRP017142」というフォルダを作成しておき、そこで作業を行うことにします。

Step1. RNA-seqデータのgzip圧縮済みのFASTQファイルをダウンロード:

論文中の記述から[GSE42213](#)を頼りに、RNA-seqデータが[GSE42212](#)として収められていることを見出し、その情報から[SRP017142](#)にたどり着いています。したがって、ここで指定するのは"SRP017142"となります。

計6ファイル、合計6Gb程度の容量のファイルがダウンロードされます。東大の有線LANで一時間弱程度かかります。早く終わらせたい場合は、最後のgetFASTQfile関数のオプションを'ftp'から'fasp'に変更すると時間短縮可能です。

[イントロ | NGS | 配列取得 | FASTQ or SRALite | SRADB \(Zhu 2013\)](#)の記述内容と基本的に同じです。

```
param <- "SRP017142" #取得したいSRA IDを指定
#必要なパッケージをロード
library(SRADB) #パッケージの読み込み
```

Step2. ヒトゲノムへのマッピングおよびカウントデータ取得:

マップしたいFASTQファイルリストおよびそのサンプル名を記述した `srp017142_samplename.txt` を作業ディレクトリに保存したうえで、下記を実行します。

`BSgenome` パッケージで利用可能な `BSgenome.Hsapiens.UCSC.hg19` へマッピングしています。名前から推測できるように "UCSC" の "hg19" にマッピングしているのと同じです。

`basic aligner` の一つである `Bowtie` を内部的に用いており、ここではマッピング時のオプションをデフォルトにしています。原著論文中で用いられた `TopHat` と同じ `splice-aware aligner` ののカテゴリに含まれる `SpliceMap (Au et al., Nucleic Acids Res., 2010)` を利用したい場合は、`qAlign` 関数実行のところで `splicedAlignment` オプションを `Bowtie` に対応する "F" から `SpliceMap` に対応する "T" に変更してください。

`hg19` にマッピングした結果なので、`TranscriptDb` オブジェクト取得時のゲノム情報もそれを基本として `Ensembl Genes ("ensGene")` を指定しているの、`Ensembl Gene ID` に対するカウントデータ取得になっています。

マシンパワーにもよりますが、ノートPCでも10時間程度で終わると思います。

マップ

```

splicedAlignment=F) #マッピングを行ったqAlign関数を実行した結果をout1に格納
time_e <- proc.time() #計算時間を計測するため
qQCReport(out, pdfFilename=out_f1) #QCレポート結果をファイルに保存
    
```

無事マッピングが終了すると、指定した5つのファイルが生成されているはずですよ。

1. QCレポートファイル (`srp017142_QC_bowtie.pdf`): Quality Controlレポートです。よく利用される `FastQC` のようなものです。
2. カウントデータファイル (`srp017142_count_bowtie.txt`): グループ(サンプル)間での発現変動遺伝子同定に用います。
3. 遺伝子配列長情報ファイル (`srp017142_genelength.txt`): 配列長とカウント数の関係を調べたいときなどに用います。これはおまけです。
4. RPKM補正後のファイル (`srp017142_RPKM_bowtie.txt`): 同一サンプル内での発現レベルの大小関係を知りたいときなどに用います。
5. 転写物塩基配列ファイル (`srp017142_transcript_seq.fa`): (遺伝子ではなく)転写物の塩基配列の `multi-FASTA` ファイルです。参考まで。
6. その他の各種情報ファイル (`srp017142_other_info1.txt`): 論文作成時に必要な、マッピング時に用いたオプション情報、マップされたリード数、Rおよび用いたパッケージのバージョン情報などを含みます。

カウントデータファイルをhogeフォルダにダウンロード

Step3. サンプル間クラスタリング:

カウントデータ (`srp017142_count_bowtie.txt`) を用いてサンプル間の全体的な類似度を眺めることを目的として、サンプル間クラスタリングを行います。

類似度は「1-Spearman相関係数」、方法は平均連結法で行っています。TCC論文(Sun et al., 2013)のFig.3でも同じ枠組みでクラスタリングを行った結果を示していますので、英語論文執筆時の参考にどうぞ。PearsonではなくSpearmanで行っているのは、ダイナミックレンジが広いので、順序尺度程度にしておいたほうが良いだろうという思想が一番大きいです。log2変換してダイナミックレンジを圧縮してPearsonにするのも一般的には「アリ」だとは思いますが、マップされたリード数が100万以上あるにも関わらずRPKMデータを用いると、RPKM補正後の値が1未

Neyret-Kahn_2013 samples ウンロー 発現変動 原著論文 せん。ま 54 bpと Bioinform alignerの 多数の で作業を Step1. F 論文中 情報か 計6フ ます。 可能で イント para #必要 libr #必要 libr #マ

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | edgeR (Robinson et al 2010) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq (Li 2013) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | TCC (Sun_2013) NEW

TCCを用いたやり方を示します。

内部的にiDEGES/edgeR(Sun_2013)正規化を実行したのち、edgeRパッケージ中のexact testで発現変動遺伝子(Differentially Expressed Genes; DEGs)検出を行っています。TCC原著論文中的iDEGES/edgeR-edgeRという解析パイプラインが、このTCCの解析に使用されています。

7. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1~gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

1.と基本的に同じで、出力のテキストファイルが正規化前のデータではなく正規化後のデータになっていて、発現変動順にソートしたものになっています。

```

1. サンプルデータ
Biological replicates
gene_2000までがDEG
gene_10000までがnon-DEG

in_f <- "data_hypodata_3vs3.txt"
out_f1 <- "hoge7.txt"
out_f2 <- "hoge7.png"
param_G1 <- 3
param_G2 <- 3
param_FDR <- 0.05
param_fig <- c(400, 380)

#必要なパッケージをロード
library(TCC)

#入力ファイル

```

```

in_f <- "data_hypodata_3vs3.txt"
out_f1 <- "hoge7.txt"
out_f2 <- "hoge7.png"
param_G1 <- 3
param_G2 <- 3
param_FDR <- 0.05
param_fig <- c(400, 380)

#必要なパッケージをロード
library(TCC)

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

#オブジェクトの作成
tcc <- TCC$new(data, param_G1, rep(2, param_G2))#G1群とG2群のサンプル数を指定
#TCCクラスオブジェクトの作成

```

#入力ファイル名を指定してin_fに格納
 #出力ファイル名を指定してout_f1に格納
 #出力ファイル名を指定してout_f2に格納
 #G1群のサンプル数を指定
 #G2群のサンプル数を指定
 #DEG検出時のfalse discovery rate (FDR)閾値を指定
 #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

hogeフォルダにダウンロードしたsrp017142_count_bowtie.txtのDEG同定を行って発現変動順にソートした状態で出力させたい場合は、このスクリプトをテンプレートとして利用します。

コード内のコピーは
 CTRL + ALT + 左クリック



```

#正規化後のデータを取り出してnormalizedに格納
tcc$normalize(tcc, norm.method="tmm", test.method="edgeR", #正規化を実行して
              iteration=3, FDR=0.1, floorPDEG=0.05)#正規化を実行した結果をtcc$normalizedに格納
normalized <- tcc$normalized

```



```
無題 - メモ帳
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge7.txt" #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge7.png" #出力ファイル名を指定してout_f2に格納
param_G1 <- 3 #G1群のサンプル数を指定
param_G2 <- 3 #G2群のサンプル数を指定
param_FDR <- 0.05 #DEG検出時のfalse discovery rate (FDR)閾値を指定
param_fig <- c(400, 380) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

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

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

#前処理(TCCクラスオブジェクトの作成)
data.c1 <- c(rep(1, param_G1), rep(2, param_G2))#G1群を1, G2群を2としたベクトルdata.c1を作成
tcc <- new("TCC", data, data.c1) #TCCクラスオブジェクトtccを作成

#本番(iDEGES/edgeR正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edgeR",#正規化を実行した結果をtccに格納
                      iteration=3, FDR=0.1, floorPDEG=0.05)#正規化を実行した結果をtccに格納
normalized <- getNormalizedData(tcc) #正規化後のデータを取り出してnormalizedに格納

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

#ファイルに保存(テキストファイル)
tmp <- cbind(rownames(tcc$count), normalized, result)#正規化後のデータの右側にDEG検出結果を結合したものをtmpに格納
tmp <- tmp[order(tmp$rank),]#発現変動順にソートした結果をtmpに格納
write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F)#tmpの中身を指定したファイル名で保存

#ファイルに保存(M-A plot)
png(out_f2, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各種パラメータを指定
plot(tcc, FDR=param_FDR) #param_FDRで指定した閾値を満たすDEGをマゼンタ色にしてM-A plotを描画
legend("topright", c(paste("DEG(FDR<", param_FDR, ")"), "non-DEG"),#凡例を作成している
      col=c("magenta", "black"), pch=20)#凡例を作成している
dev.off() #おまじない
sum(tcc$stat$q.value < 0.05) #FDR < 0.05を満たす遺伝子数を表示
sum(tcc$stat$q.value < 0.10) #FDR < 0.10を満たす遺伝子数を表示
sum(tcc$stat$q.value < 0.20) #FDR < 0.20を満たす遺伝子数を表示
sum(tcc$stat$q.value < 0.30) #FDR < 0.30を満たす遺伝子数を表示
```

該当箇所を変更し、R Console画面上でコピー

```
in_f <- "srp017142_count_bowtie.txt" #入力ファイル名を指定して in_fに格納
out_f1 <- "hoge7.txt" #出力ファイル名を指定して out_f1に格納
out_f2 <- "hoge7.png" #出力ファイル名を指定して out_f2に格納
param_G1 <- 3 #G1群のサンプル数を指定
param_G2 <- 3 #G2群のサンプル数を指定
param_FDR <- 0.05 #DEG検出時のfalse discovery rate (FDR)閾値を指定
param_fig <- c(400, 380) #ファイル出力時の横幅と縦幅を指定(単位はピクセル)

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

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

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2)) #G1群を1, G2群を2
tcc <- new("TCC", data, data.cl) #TCCクラスオブジェクト tccを作成

#本番(iDEGES/edgeR正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edgeR",
iteration=3, FDR=0.1, floorPDEG=0.05)
normalized <- getNormalizedData(tcc) #正規化後のデータを取り出す

#本番(DEG検出)
tcc <- estimateDE(tcc, test.method="edgeR", FDR=param_FDR) #edgeRによるDEG検出
result <- getResult(tcc, sort=FALSE) #p値などの結果を抽出して結果表を作成
sum(tcc$stat$q.value < param_FDR) #FDR < param_FDRを満たす遺伝子数を表示
```

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

#ファイルに保存(M-A plot)
png(out_f2, pointsize=13, width=param_fig[1], height=param_fig[2]) #出力するPNGファイルのサイズを指定
plot(tcc, FDR=param_FDR) #param_FDRで指定した閾値を満たすDEGをM-A plotで表示
legend("topright", c(paste("DEG(FDR<", param_FDR, ")"), sep=""), "non-DEG",
col=c("magenta", "black"), pch=20) #凡例を作成している
dev.off() #おまじない

sum(tcc$stat$q.value < 0.05) #FDR < 0.05を満たす遺伝子数を表示
sum(tcc$stat$q.value < 0.10) #FDR < 0.10を満たす遺伝子数を表示
sum(tcc$stat$q.value < 0.20) #FDR < 0.20を満たす遺伝子数を表示
sum(tcc$stat$q.value < 0.30) #FDR < 0.30を満たす遺伝子数を表示
```

```
R Console
> #ファイルに保存(テキストファイル)
> tmp <- cbind(rownames(tcc$count), normalized, result) #正規化後のデータと結果表を結合
> tmp <- tmp[order(tmp$rank),] #発現変動順にソートした結果をtmpに格納
> write.table(tmp, out_f1, sep="\t", append=F, quote=F, row.names=F) #tmpをテキストファイルに保存
> #ファイルに保存(M-A plot)
> png(out_f2, pointsize=13, width=param_fig[1], height=param_fig[2]) #出力するPNGファイルのサイズを指定
> plot(tcc, FDR=param_FDR) #param_FDRで指定した閾値を満たすDEGをM-A plotで表示
> legend("topright", c(paste("DEG(FDR<", param_FDR, ")"), sep=""), "non-DEG",
+ col=c("magenta", "black"), pch=20) #凡例を作成している
> dev.off() #おまじない
null device
1
> sum(tcc$stat$q.value < 0.05) #FDR < 0.05を満たす遺伝子数を表示
[1] 5669
> sum(tcc$stat$q.value < 0.10) #FDR < 0.10を満たす遺伝子数を表示
[1] 6680
> sum(tcc$stat$q.value < 0.20) #FDR < 0.20を満たす遺伝子数を表示
[1] 8110
> sum(tcc$stat$q.value < 0.30) #FDR < 0.30を満たす遺伝子数を表示
[1] 9151
> |
```

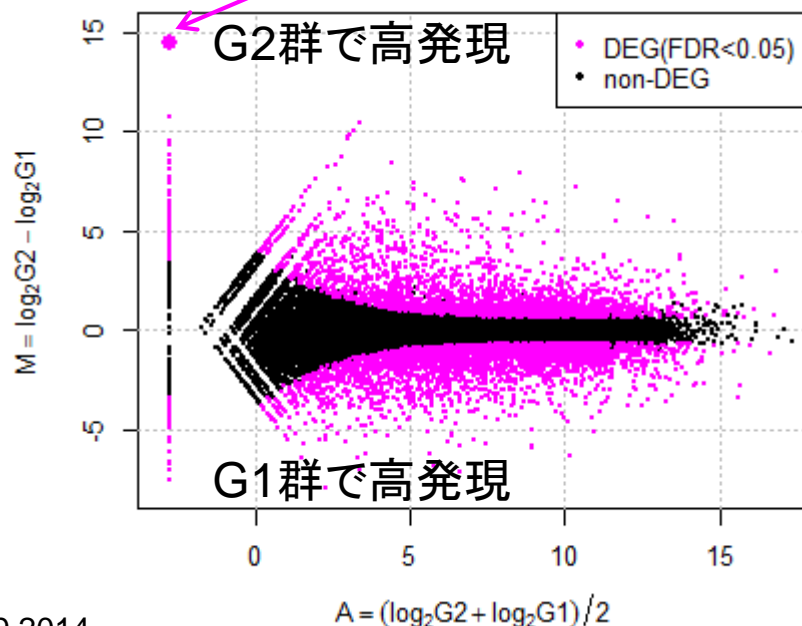
該当箇所を変更し、R Console画面上でコピー

実データ解析結果: SRP017142

TCCを用いたDEG同定結果ファイル

p-valueとその順位

rownames(tcc\$count)	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
ENSG00000240386	0.0	0.0	0.0	5608.1	5097.7	8188.0	ENSG00000240386	-2.78	14.48	1.75E-139	1.04E-134	1	1
ENSG00000128564	15.4	22.3	19.1	2856.6	2462.6	2555.3	ENSG00000128564	7.80	7.11	4.03E-107	1.21E-102	2	1
ENSG00000188064	7.7	5.8	10.1	1425.5	1254.0	1486.8	ENSG00000188064	6.71	7.47	2.67E-98	5.33E-94	3	1
ENSG00000101188	6.0	5.0	11.1	1477.4	1407.0	1254.9	ENSG00000101188	6.65	7.55	3.81E-97	5.70E-93	4	1
ENSG00000163431	3716.6	3244.5	4185.2	70.1	78.8	49.0	ENSG00000163431	8.95	-5.82	2.90E-80	3.48E-76	5	1
ENSG00000204291	1215.5	1236.8	1339.3	22.4	27.8	21.5	ENSG00000204291	7.44	-5.72	3.45E-78	3.44E-74	6	1
ENSG00000181634	107.0	158.6	66.5	12842.0	16014.1	19820.5	ENSG00000181634	10.39	7.20	1.04E-76	8.88E-73	7	1
ENSG00000178726	53.1	46.3	62.5	6006.1	5567.6	3166.0	ENSG00000178726	9.01	6.51	2.39E-74	1.79E-70	8	1
ENSG00000117600	576.9	518.9	602.6	7.0	5.6	2.4	ENSG00000117600	5.73	-6.83	1.14E-72	7.59E-69	9	1
ENSG00000158050	7.7	9.1	11.1	552.3	536.6	523.5	ENSG00000158050	6.14	5.85	1.16E-70	6.91E-67	10	1
ENSG00000124126	50.5	44.6	53.4	1819.4	1683.2	1413.9	ENSG00000124126	8.15	5.05	5.35E-69	2.91E-65	11	1



M-A plotのA値とM値

q-value

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

セミナーで示したものと同じです
このような手順で作成しています

Contents



■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- **TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ**
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

解析 | 発現変動 | について (last modified 2013/08/29)

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | 複製なし | TCC (Sun_2013) (last modified 2013/09/12)

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | edgeR (Robinson_2010) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製あり | SAMseq (Li_2013) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製なし | TCC (Sun_2013) (last modified 2014/02/04) 推奨 NEW

解析 | 発現変動 | 2群間 | 対応なし | 複製なし | DESeq (Anders_2010) (last modified 2014/01/30) NEW

解析 | 発現変動 | 2群間 | 対応なし | BitSeq (Glauz_2012) (last modified 2013/01/08)

解析 | 発現変動 | 2群間 | 対応なし | 複製なし | TCC (Sun_2013)

解析 | 発現変動 | 2群間 | 対応なし | 複製なし | TCC (Sun_2013) NEW

TCCを用いたやり方を示します。

内部的にDEGES/DESeq(Sun_2013)正規化を実行したのち、DESeqパッケージ中のnegative binomial testで発現変動遺伝子(Differentially expressed Genes: DEGs)検出を行っています。TCC原著論文中のDESeq/DESeqと

4. サンプルデータ14の10,000 genes×2 samplesのカウントデータ(data_hypodata_1vs1.txt)の場合:

シミュレーションデータ(G1群1サンプル vs. G2群1サンプル)です。gene_1~gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。1.と基本的に同じで、出力のテキストファイルが正規化前のデータではなく正規化後のデータになっていて、発現変動順にソートしたのになっています。

1. サンプル

シミュレーションデータがG1群で高

```

in_f <- "data_hypodata_1vs1.txt" #入力ファイル名を指定してin_fに格納
out_f1 <- "hoge4.txt" #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge4.png" #出力ファイル名を指定してout_f2に格納
param_G1 <- 1 #G1群のサンプル数を指定
param_G2 <- 1 #G2群のサンプル数を指定
param_FDR <- 0.05 #DEG検出時

```

```

#必要なパッケージをロード
library(TCC) #パッケージ

```

```

#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

```

```

#前処理(正規化前のデータオブジェクトの作成)
data.cl <- list(param_G1, rep(2, param_G2))#G1群を1、G2群を2としたベクトルdata.clを
#TCCクラスオブジェクトtccを作成

```

```

tcc <- TCC::TCC(data.cl, norm.method="deseq", test.method="deseq",#正規化を実行した
norm.method="deseq", test.method="deseq",#正規化を実行した結果をtccに格納
normData(tcc) #正規化後のデータを取り出してnormalizedに格納

```

解析したいファイルdata_hypodata_1vs1.txtをデスクトップ上のhogeフォルダに保存して実行してみましょう。

コード内のコピーは CTRL + ALT + 左クリック



data_hypodata_1vs1.txtの解説

DEG

G1で高発現

G2で高発現

non-DEG

data_hypodata_3vs3.txt

	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene_1	36	56	144	2	1	0
gene_2	84	152	124	52	37	28
gene_3	592	840	800	151	257	200
gene_4	0	8	4	1	1	3
gene_5	32	32	0	1	1	0
...						
gene_1801	34	86	24	284	180	364
gene_1802	5	1	3	0	160	24
gene_1803	57	56	51	248	192	220
gene_1804	29	25	32	128	204	160
gene_1805	42	29	44	184	156	92
...						
gene_2001	4	8	9	13	12	4
gene_2002	88	139	40	22	44	21
gene_2003	933	667	462	889	396	443
gene_2004	48	37	14	36	57	71
gene_2005	290	338	553	319	210	504
...						
gene_9996	107	67	104	35	65	45
gene_9997	145	220	120	80	95	156
gene_9998	42	73	67	62	44	37
gene_9999	5	1	2	3	4	11
gene_10000	2	4	5	2	0	0

data_hypodata_1vs1.txt

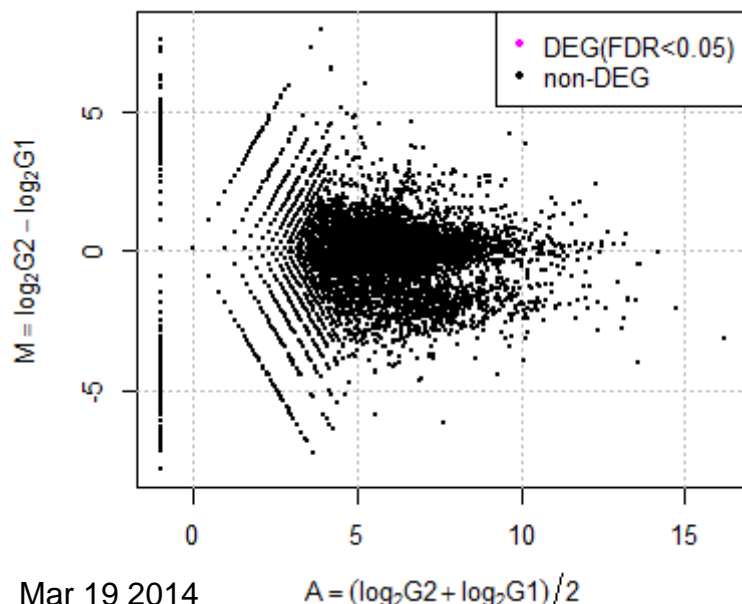
	G1_rep1	G2_rep1
gene_1	36	2
gene_2	84	52
gene_3	592	151
gene_4	0	1
gene_5	32	1
...		
gene_1801	34	284
gene_1802	5	0
gene_1803	57	248
gene_1804	29	128
gene_1805	42	184
...		
gene_2001	4	13
gene_2002	88	22
...		
gene_9996	107	35
gene_9997	145	80
gene_9998	42	62
gene_9999	5	3
gene_10000	2	2

複製ありシミュレーションデータの一部です

実行結果

TCCを用いたDEG同定結果ファイル(hoge4.txt)

rownames(tG1_rep1	G2_rep1	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG	
gene_695	238.9	0.0	gene_695	-0.999	-7.844	0.0019	1	1	0
gene_938	0.0	179.9	gene_938	-0.999	7.545	0.00319	1	2	0
gene_5834	1.0	232.9	gene_5834	3.905	7.918	0.00427	1	3	0
gene_231	1695.2	23.9	gene_231	7.654	-6.147	0.00438	1	4	0
gene_2115	0.0	150.8	gene_2115	-0.999	7.290	0.00473	1	5	0
gene_823	150.3	0.0	gene_823	-0.999	-7.175	0.00524	1	6	0
gene_3852	0.0	138.3	gene_3852	-0.999	7.166	0.00574	1	7	0
gene_1315	134.8	0.0	gene_1315	-0.999	-7.019	0.00666	1	8	0
gene_1214	131.0	0.0	gene_1214	-0.999	-6.977	0.0071	1	9	0
gene_2555	179.2	7691.0	gene_2555	10.197	5.424	0.00821	1	10	0



0.05~0.30のFDR閾値を満たすDEGは0個

```

R Console
> sum(tcc$stat$q.value < 0.05)      #FDR < 0.05を満たす遺伝子数を表示
[1] 0
> sum(tcc$stat$q.value < 0.10)     #FDR < 0.10を満たす遺伝子数を表示
[1] 0
> sum(tcc$stat$q.value < 0.20)     #FDR < 0.20を満たす遺伝子数を表示
[1] 0
> sum(tcc$stat$q.value < 0.30)     #FDR < 0.30を満たす遺伝子数を表示
[1] 0
> |
  
```

実行結果

TCCを用いたDEG同定結果ファイル(hoge4.txt)

rownames(tG1_rep1	G2_rep1	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG	
gene_695	238.9	0.0	gene_695	-0.999	-7.844	0.0019	1	1	0
gene_938	0.0	179.9	gene_938	-0.999	7.545	0.00319	1	2	0
gene_5834	1.0	232.9	gene_5834	3.905	7.918	0.00427	1	3	0
gene_231	1695.2	23.9	gene_231	7.654	-6.147	0.00438	1	4	0
gene_2115	0.0	150.8	gene_2115	-0.999	7.290	0.00473	1	5	0
gene_823	150.3	0.0	gene_823	-0.999	-7.175	0.00524	1	6	0
gene_3852	0.0	138.3	gene_3852	-0.999	7.166	0.00574	1	7	0
gene_1315	134.8	0.0	gene_1315	-0.999	-7.019	0.00666	1	8	0
gene_1214	131.0	0.0	gene_1214	-0.999	-6.977	0.0071	1	9	0
gene_2555	179.2	7691.0	gene_2555	10.197	5.424	0.00821	1	10	0

p 値の定義から、10,000遺伝子 \times 0.005 = 50個分の真のnon-DEGをDEGと判定ミスするのを許容することに相当



$p < 0.005$ を満たす5個の中に占める偽物の割合は $50/5 > 1.0$ と計算することができる



これ(1.0)がFDR!!



統計的手法のおさらい

- 同一群内の遺伝子のばらつきの程度を把握し、帰無仮説に従う分布の全体像を把握しておく(モデル構築)
 - non-DEGのばらつきの程度を把握しておくことと同義
- 実際に比較したい2群の遺伝子のばらつきの程度がnon-DEG分布のどのあたりに位置するかを評価
- 複製なしデータの場合は、モデル構築自体が困難
 - 対策1: 適当なモデルを使う(ポアソンモデルとか...)
 - 対策2: 他のデータから得られたモデルまたはパラメータを拝借
 - 対策3: 複製なしデータ自体を同一群のデータとみなしてモデル構築
 - ...

同一群

	G1_rep1	G2_rep1
gene_1	36	2
gene_2	84	52
gene_3	592	151
gene_4	0	1
gene_5	32	1
...		

複製なしデータの解析結果はどうあがいても信頼性が低いから、むやみに低いp値のものが出力されないように厳しめにせざるをえない



Contents



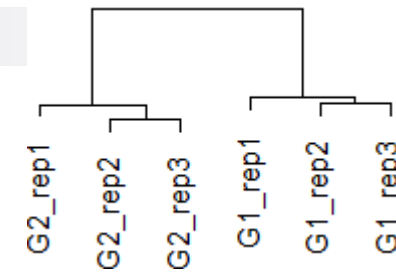
■ セミナー(13:00-14:00)

- 研究目的別留意点: サンプル内とサンプル間の違い
- マッピング → カウント情報取得
- 実データ解析例: 結果の解釈やM-A plotの見方など
- 多重比較問題: FDRって何?
- 分布やモデル
- なぜ x 倍発現変動という議論がだめなんですか?
- 理想的な実験デザイン
- データの正規化

■ トレーニング(14:30-16:30)

- TCC発現変動解析: 複製あり2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり2群間比較用実データ
- TCC発現変動解析: 複製なし2群間比較用シミュレーションデータ
- TCC発現変動解析: 複製あり3群間比較用シミュレーションデータ

TCCで複製あり3群間比較



data_hypodata_3vs3vs3.txt (3群間比較用)

- G1群:3サンプル、G2群:3サンプル、G3群:3サンプル
- 全部で10,000行×9列。最初の3,000行分が発現変動遺伝子(DEG)

DEG

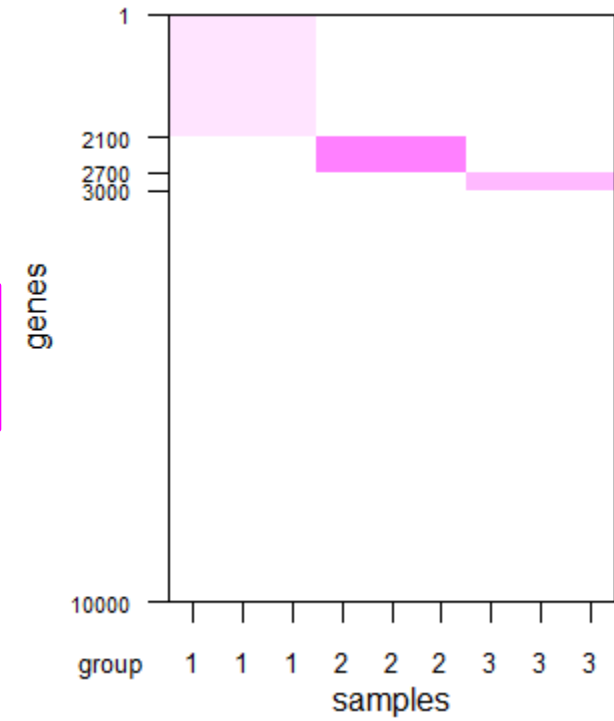
G1で3倍
高発現

G2で10
倍高発現

G3で6倍
高発現

non-DEG

	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	G3_rep1	G3_rep2	G3_rep3
gene_1	245	109	84	52	39	21	68	58	19
gene_2	16	8	5	6	5	4	2	6	3
gene_3	49	52	40	4	15	15	21	36	18
...									
gene_2101	565	757	513	8904	7445	5321	681	399	563
gene_2102	0	7	4	22	0	0	0	9	13
gene_2103	0	1	0	10	5	6	0	0	0
...									
gene_2701	49	72	67	71	94	92	348	370	599
gene_2702	112	101	76	149	105	144	487	526	740
gene_2703	25	49	31	35	14	25	78	82	209
...									
gene_3001	0	0	4	0	0	5	1	1	0
gene_3002	48	54	55	16	55	26	46	23	24
gene_3003	61	80	34	36	60	71	53	52	46
...									
gene_9998	2	0	0	2	0	2	9	0	0
gene_9999	5	3	8	1	2	0	2	3	0
gene_10000	16	16	38	56	50	45	22	8	99



DEG同定結果として、gene_1~gene_3000
が上位にランキングされていれば正解！

- 解析 | 発現変動 | 2群間 | 対応なし | NBPSeq (Di 2011) (last modified 2012/03/15)
- 解析 | 発現変動 | 2群間 | 対応あり | 複製なし | TCC (Sun 2013) (last modified 2014/02/07) 推奨
- 解析 | 発現変動 | 2群間 | 対応あり | 複製なし | edgeR (Robinson 2010) (last modified 2014/01/0)
- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | TCC (Sun 2013) (last modified 2013/08/29)
- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | TCC (Sun 2013) (last modified 2014/02/04) 推奨
- 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | edgeR (Robinson 2010) (last modified 2013/09/1)
- 解析 | 発現変動 | 時系列データ | Bayesian model-based clustering (Nascimento 2012) (last modified 2012/03/15)

• 解析 | 発現変動 | 3群間 | 対応なし | 複製あり | [TCC \(Sun 2013\)](#)

解析 | 発現変動 | 3群間 | 対応なし | 複製あり | TCC (Sun_2013)

TCCを用いたやり方を示します。

内部的に [iDEGES/edgeR \(Sun 2013\)](#) 正規化を実行したのち、[edgeR](#) パッケージ中の `exact test` で発現変動遺伝子 (Differentially expressed Genes: DEGs) 検出を行っています。TCC 原著論文中の `iDEGES/edgeR.edgeR` という解析パイプラインに相当します。全て

5. サンプルデータ15の10,000 genes×9 samplesのカウントデータ([data_hypodata_3vs3vs3.txt](#))の場合:

シミュレーションデータ(G1群3サンプル vs. G2群3サンプル vs. G3群3サンプル)です。gene_1~gene_3000までがDEG (gene_1~gene_2100がG1群で3倍高発現、gene_2101~gene_2700がG2群で10倍高発現、gene_2701~gene_3000がG3群で6倍高発現) gene_3001~gene_10000までがnon-DEGであることが既知です。

1.と基本的に同じで、出力のテキストファイルが正規化前のデータではなく正規化後のデータになっていて、発現変動順にソートしたのになっています。

```
in_f <- "data_hypodata_3vs3vs3.txt"
out_f <- "hoge5.txt"
param_G1 <- 3
param_G2 <- 3
param_G3 <- 3
param_FDR <- 0.05
```

```
in_f <- "data_hypodata_3vs3vs3.txt"
out_f <- "hoge5.txt"
param_G1 <- 3
param_G2 <- 3
param_G3 <- 3
param_FDR <- 0.05
```

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

```
#必要なパッケージをロード
library(TCC)
```

#パッケージの読み込み

```
#入力ファイルの読み込み
```

```
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.cl) #TCCクラスオブジェクトtccを作成
```

```
#必要なパッケージをロード
```

コード内のコピーは
CTRL + ALT + 左クリック



実行結果

このスペースは基本的に2群間比較用。今回の結果は3群間比較用なので意図的にNAと表示させている

TCCを用いたDEG同定結果ファイル(hoge5.txt)

rownames(t)	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3	G3_rep1	G3_rep2	G3_rep3	gene_id	a.value	m.value	p.value	q.value	rank	estimatedDEG
gene_2547	33.3	43.4	38.8	678.1	860.2	502.5	44.6	54.6	47.6	gene_2547	NA	NA	1.00E-36	1.00E-32	1	1
gene_2592	112.6	175.7	150.4	1815.1	2162.0	2016.1	166.2	175.8	160.1	gene_2592	NA	NA	8.15E-36	3.14E-32	2	1
gene_2626	127.3	133.3	126.1	1642.0	1477.3	1205.6	116.6	109.3	135.8	gene_2626	NA	NA	9.43E-36	3.14E-32	3	1
gene_2328	104.8	85.9	84.4	919.9	1134.4	983.7	95.3	81.4	85.1	gene_2328	NA	NA	2.69E-35	6.73E-32	4	1
gene_2164	63.7	61.2	87.3	1348.0	867.2	1294.8	118.6	120.2	113.5	gene_2164	NA	NA	3.97E-35	7.94E-32	5	1
gene_2329	105.8	87.9	105.8	995.7	1141.5	1107.4	96.3	78.5	106.4	gene_2329	NA	NA	6.86E-35	1.14E-31	6	1
gene_2398	112.6	90.8	65.0	1140.1	1428.9	1457.9	125.7	99.3	70.9	gene_2398	NA	NA	1.13E-34	1.61E-31	7	1
gene_2263	174.3	153.0	142.6	1872.5	1723.3	1960.4	159.1	195.7	171.2	gene_2263	NA	NA	1.46E-34	1.83E-31	8	1
gene_2168	60.7	97.7	85.4	801.0	926.7	968.6	64.9	65.6	78.0	gene_2168	NA	NA	3.33E-34	3.70E-31	9	1
gene_2161	110.7	152.0	158.2	1309.1	1659.8	1713.2	101.4	128.1	140.8	gene_2161	NA	NA	6.48E-34	5.91E-31	10	1

G2群での高発現DEGが上位を占めるのはシミュレーション条件的に妥当

```
R Console
> sum(tcc$stat$q.value < 0.05) #FDR < 0.05を満たす遺伝子数を表示
[1] 2003
> sum(tcc$stat$q.value < 0.10) #FDR < 0.10を満たす遺伝子数を表示
[1] 2292
> sum(tcc$stat$q.value < 0.20) #FDR < 0.20を満たす遺伝子数を表示
[1] 2679
> sum(tcc$stat$q.value < 0.30) #FDR < 0.30を満たす遺伝子数を表示
[1] 3058
> |
```

```
R Console
> 2003*0.95
[1] 1902.85
> 2292*0.9
[1] 2062.8
> 2679*0.8
[1] 2143.2
> 3058*0.7
[1] 2140.6
> |
```

検出されたDEG数から理論上の偽物を差し引くと、概ね2,000 DEGsという結果。真は3,000個だが、ノイズに埋もれるものもあるので概ね妥当



東京大学大学院農学生命科学研究科

アグリバイオインフォマティクス教育研究ユニット

Agricultural Bioinformatics Research Unit

[受講生の方へ](#) [研究者の方へ](#)

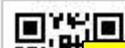
- + ホーム
- + 本ユニットについて
- + メンバー
- + 教育プログラム
- + 研究フォーラム
- + イベント
- + お問い合わせ
- + リンク
- + モバイルサイト

[ホーム](#) > [教育プログラム](#) > [各講義のページ](#)

各講義のページ

(科目名をクリックすると各講義のページに移動します)

先端トピックス セミナー・ 討論形式 研究指導	農学生命情報科学特別演習			
	農学生命情報科学特論 I	農学生命情報科学特論 II	農学生命情報科学特論 III	農学生命情報科学特論 IV
方法論 講義・実習を 一体化	生物配列統計学	システム生物学概論	知識情報処理論	
	オーム情報解析	機能ゲノム学	分子モデリングと分子シミュレーション	
基礎 講義・実習を 一体化	ゲノム情報解析基礎		構造バイオインフォマティクス基礎	
	生物配列解析基礎		バイオスタティクス基礎論	



東大生以外の方も受講可能です(平成26年度もやります)

謝辞

共同研究者

清水 謙多郎 先生(東京大学・大学院農学生命科学研究科)

西山 智明 先生(金沢大学・学際科学実験センター)

孫 建強 氏(東京大学・大学院農学生命科学研究科・大学院生)

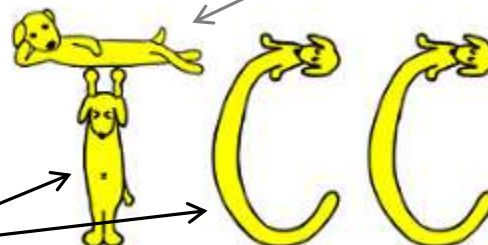
グラント

- 基盤研究(C)(H24-26年度):「シーケンスに基づく比較トランスクリプトーム解析のためのガイドライン構築」(代表)
- 新学術領域研究(研究領域提案型)(H22年度-):「非モデル生物におけるゲノム解析法の確立」(分担;研究代表者:西山智明)




挿絵やTCCのロゴなど

(妻の)門田 雅世さま作



(有能な秘書の)三浦 文さま作



これ以降のスライドは参考資料。なぜRPKM
とNBモデルの相性が悪いのかも載せる。

よく見かけるカウントデータ取得手段

- basic alignerの1つであるBowtieを利用
- 最大2塩基ミスマッチまで許容してリファレンス配列の1か所とのみ一致するリード (uniquely mapped reads or unique mapper) 数をカウント
 - Marioni et al., *Genome Res.*, **18**:1509-1517, 2008
 - Bullard et al., *BMC Bioinformatics*, **11**:94, 2010
 - Risso et al., *BMC Bioinformatics*, **12**:480, 2011
 - ReCount (Frazee et al., *BMC Bioinformatics*, **12**:449, 2011)
 - ...

SpliceMap (Au et al., 2010)などのsplice-aware alignerだと相当時間がかかるという現実的な問題もあるのだろう。講義や講習会では到底無理。
→ ユーザの記憶に残らない → 実際に使われない...

定量化：遺伝子レベル ⇔ isoformレベル

- 全体的な流れとしては遺伝子レベル → isoformレベル
 - 例：新規splice variantの発見 (Twine et al., *PLoS One*, **6**: e16266, 2011)
- 遺伝子セット解析 (Gene Ontology解析やパスウェイ解析など) のための基本情報は遺伝子レベルの解像度
- 複数エクソン → 遺伝子レベルの要約統計量
 - exon union method (Mortazavi et al., *Nat. Methods*, **5**: 621-628, 2008)
 - 全てのisoforms間で用いられているexonの情報 (**union**: 和集合) を利用
 - exon intersection method (Bullard et al., *BMC Bioinformatics*, **11**: 94, 2010)
 - 複数isoforms間で共通して用いられているexonの情報のみ (**intersection**: 積集合) を利用

count情報を得る際に、どのexonの情報を用いるか?

遺伝子のカウント数の定義

- 算出された生リードカウント結果
 - exon union method(和集合)の場合: 20 reads
 - Exon intersection method(積集合)の場合: 11 reads

20 reads

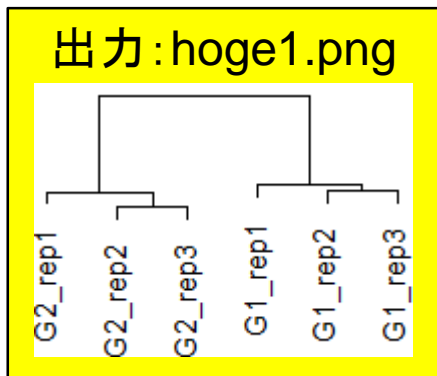
11 reads

11 reads

様々な思想があり、当然その後の解析結果に影響を及ぼします

サンプル間クラスタリング

- 解析 | 一般 | 上流配列解析 | [LDSS\(Yamamoto 2007\)](#)(last modified 2012/07/17)
- 解析 | 一般 | 上流配列解析 | [Relative Appearance Ratio\(Yamamoto 2011\)](#)(last modified 2012/07/17)
- 解析 | 基礎 | [平均分散プロット\(Technical replicates\)](#)(last modified 2013/12/27)
- 解析 | 基礎 | [平均分散プロット\(Biological replicates\)](#)(last modified 2013/12/27)
- 解析 | [クラスタリング | について](#) (last modified 2014/02/05) **NEW**
- 解析 | クラスタリング | サンプル間 | [hclust](#) (last modified 2014/02/06) **NEW**
- 解析 | クラスタリング | 遺伝子間 | [MBClustSeq\(Si 2014\)](#)(last modified 2014/02/05) **NEW**
- 解析 | 発現変動 | ポアソン分布 | [シミュレーションデータ\(Technical replicates\)](#)(last modified 2011/09/16)
- 解析 | 発現変動 | [シミュレーションデータ\(Biological replicates\)](#)(last modified 2011/09/16)



解析 | クラスタリング | サンプル間 | hclust **NEW**

RNA-seqカウントデータのクラスタリング結果は、特にゼロカウント(0カウント; zero count)を多く含む場合に(もちろん距離の定義の仕方によっても変わってきますが)低発現データのフィルタリングの閾値次第で結果が変わる傾向にあります。ここでは、上記閾値問題に悩まされることなく頑健なサンプル間クラスタリングを行うやり方を示します。内部的に行っていることは、以下の通りです:

1. 全サンプルで0カウントとなる行(遺伝子)をフィルタリングした後、unique関数を用いて同一発現パターンのもを1つのパターンとしてまとめる、
2. 「1 - Spearman順位相関係数」でサンプル間距離を定義、
3. Average-linkage clusteringの実行、です。

順位相関係数を用いてサンプルベクトル間の類似度として定義するので、サンプル間正規化の問題に悩まされません。また、低発現遺伝子にありがちな同一発現パターンの遺伝子をまとめることで、(変動しやすい)同順位となる大量の遺伝子が集約されるため、結果的に「総カウント数がx個以下のものをフィルタリング...」という閾値問題をクリアしたことになります。近いうち(2014年中)にTCCパッケージ中でwrapper functionを提供する予定です。

「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13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1~gene_2000までがDEG(最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

```
in_f <- "data_hypodata_3vs3.txt" #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.png" #出力ファイル名(クラスタリング結果ファイル)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="")#
dim(data) #オブジェクトdataの行数と列数を表示
```

```
#前処理(フィルタリング)
obj <- as.logical(rowSums(data) > 0) #条件を満たすかどうかを判定した結果をobjに格納
```

ゼロカウントを含む低発現データのフィルタリングは重要です

サンプル間クラスタリング

■ data_hypodata_3vs3.txt (2群間比較用)

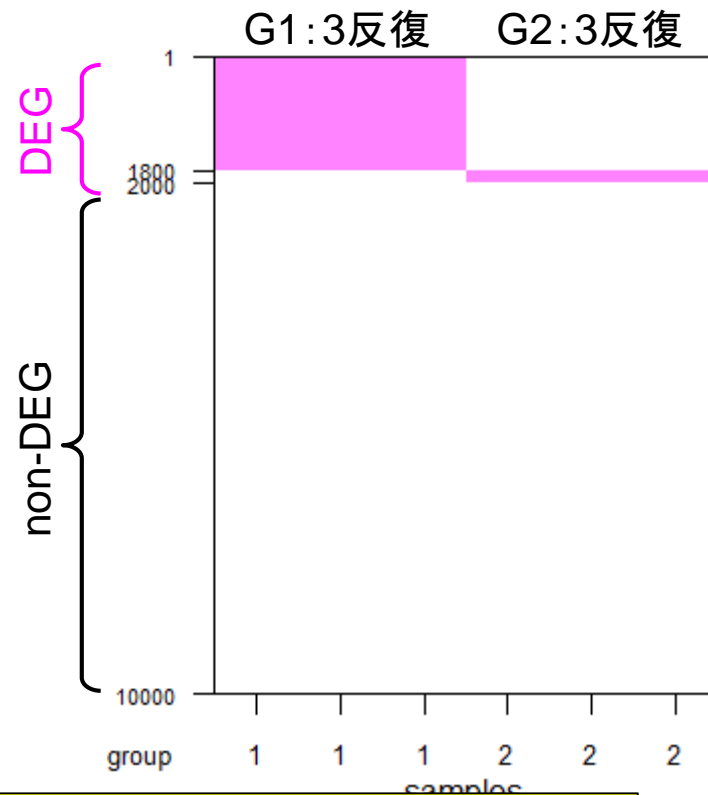
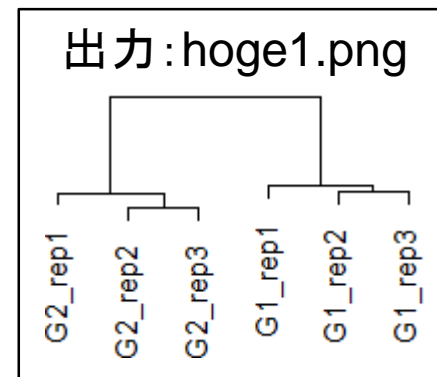
- G1群:3サンプル、G2群:3サンプル
- 全部で10,000行×6列。最初の2,000行分が発現変動遺伝子 (DEG)

	G1_rep1	G1_rep2	G1_rep3	G2_rep1	G2_rep2	G2_rep3
gene_1	36	56	144	2	1	0
gene_2	84	152	124	52	37	28
gene_3	592	840	800	151	257	200
gene_4	0	8	4	1	1	3
gene_5	32	32	0	1	1	0
...						
gene_1801	34	86	24	284	180	364
gene_1802	5	1	3	0	160	24
gene_1803	57	56	51	248	192	220
gene_1804	29	25	32	128	204	160
gene_1805	42	29	44	184	156	92
...						
gene_2001	4	8	9	13	12	4
gene_2002	88	139	40	22	44	21
gene_2003	933	667	462	889	396	443
gene_2004	48	37	14	36	57	71
gene_2005	290	338	553	319	210	504
...						
gene_9996	107	67	104	35	65	45
gene_9997	145					
gene_9998	42					
gene_9999	5					
gene_10000	2					

DEG

G1で高発現
G2で高発現

non-DEG



DEGが多く存在するほど群間で明瞭なクラスターに分かれる傾向
→クラスタリング結果からDEGの有無をある程度把握可能です

サンプル間クラスタリング

2. サンプルデータ13の10,000 genes×6 samplesのカウントデータ(data_hypodata_3vs3.txt)の場合:

Biological replicatesを模倣したシミュレーションデータ(G1群3サンプル vs. G2群3サンプル)です。gene_1~gene_2000までがDEG (最初の1800個がG1群で高発現、残りの200個がG2群で高発現) gene_2001~gene_10000までがnon-DEGであることが既知です。

non-DEGデータのみでクラスタリングを行っています。

```

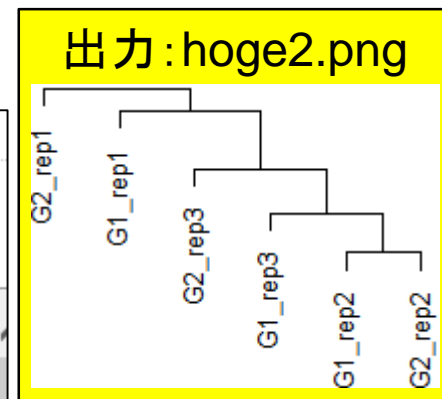
in_f <- "data_hypodata_3vs3.txt"      #入力ファイル名を指定してin_fに格納
out_f <- "hoge2.png"                 #出力ファイル名(クラスタリング結果ファイル)を指定
param_fig <- c(500, 400)             #ファイル出力時の横幅と縦幅を指定(単位はピクセル)
param_nonDEG <- 2001:10000           #non-DEGの位置を指定

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

#前処理(サブセットの抽出)
data <- data[param_nonDEG,]           #指定した行のみ抽出した結果をdata1に格納

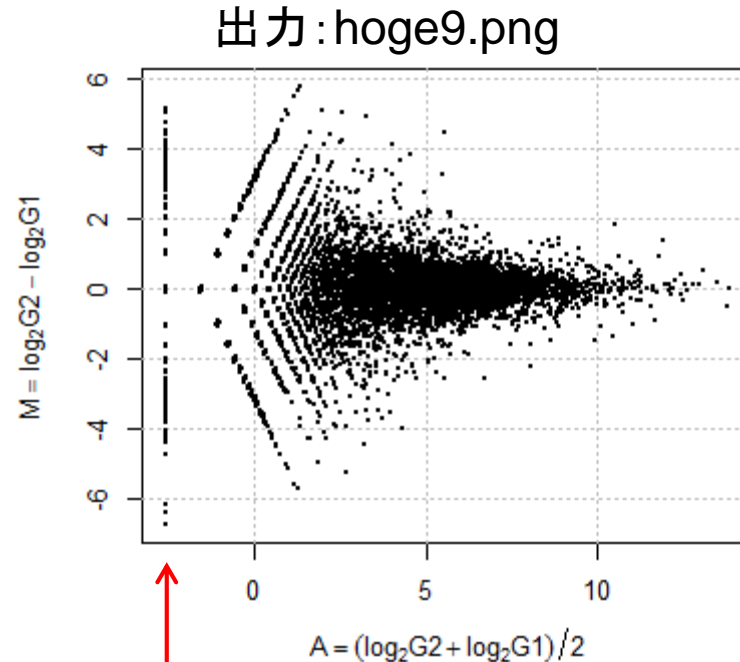
#前処理(フィルタリング)
obj <- as.logical(rowSums(data) > 0)   #条件を満たすかどうかを判定した結果をobjに格納
data <- unique(data[obj,])            #objがTRUEとなる行のみ抽出し、ユニークパターンのみにした結果をdata2に格納
dim(data)                             #オブジェクトdataの行数と列数を表示

#本番
data.dist <- as.dist(1 - cor(data, method="spearman")) #サンプル間の距離を計算し、結果をdata.distに格納
out <- hclust(data.dist, method="average") #階層的クラスタリングを実行し、結果をoutに格納
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
    
```



DEGが存在しないデータの典型的なクラスタリング結果です

M-A plot (TCCパッケージの0カウント対策)



- ①各群について、ゼロでない平均発現量の最小値を取得
- ②0だったところをその値で置換
- ③M値を再計算
- ④M-A plotの左側に、再計算して得られたM値をプロット

性能評価(統計的手法 vs. 倍率変化)

data_marioni.txt (technical replicates)

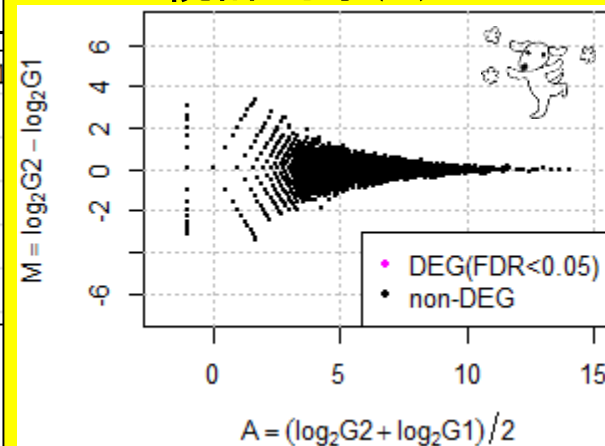
腎臓(Kidney)群

rownames(data)	R1L1Kidney	R1L3Kidney	R1L7Kidney	R2L2Kidney	R2L6Kidney	R2L7Kidney
ENSG00000000003	178	167	179	172	151	167
ENSG00000000005	0	0	0	0	1	0
ENSG000000000419	53	78	64	72	71	67
ENSG000000000457	22	33	30	27	30	28
ENSG000000000460	9	7	18	14	9	12

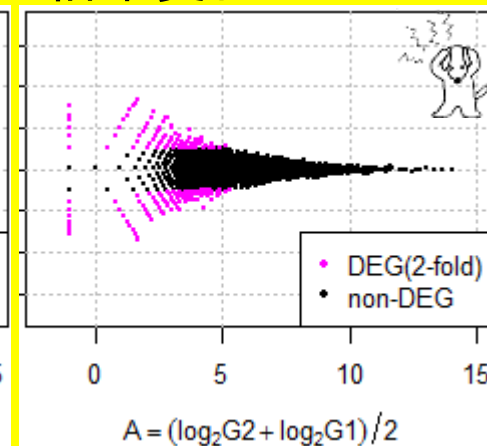
G1群

G2群

統計的手法



倍率変化



data_arab.txt (biological replicates)

mock群

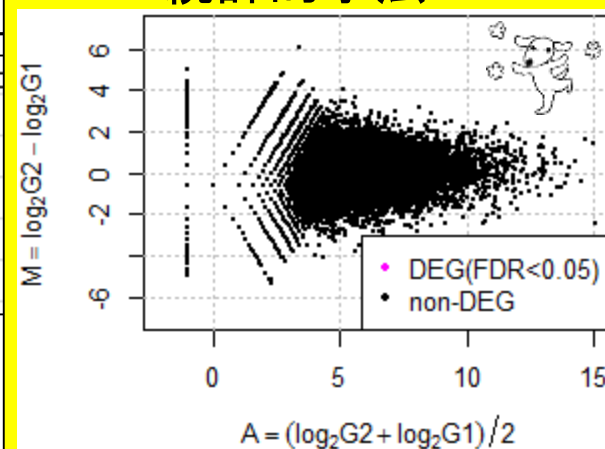
hrcc群

identifier	mock1	mock2	mock3	hrcc1	hrcc2	hrcc3
AT1G01010	35	77	40	46	64	50
AT1G01020	43	45	32	43	39	41
AT1G01030	16	24	26	27	35	31
AT1G01040	72	43	64	66	25	46
AT1G01050	49	78	90	67	45	61

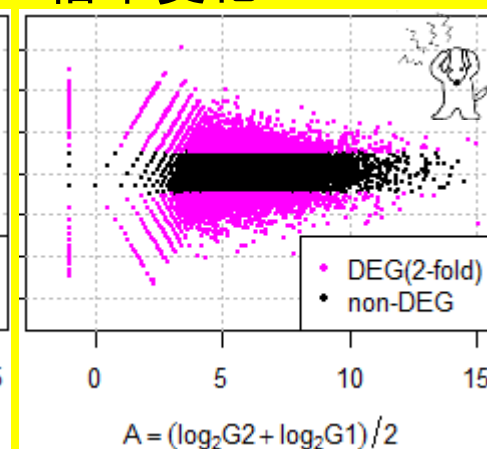
G1群

G2群

統計的手法



倍率変化



統計的手法のほうがnon-DEGをDEGと判定するミス(false positives)が圧倒的に少ない

ばらつき度 (technical vs. biological)

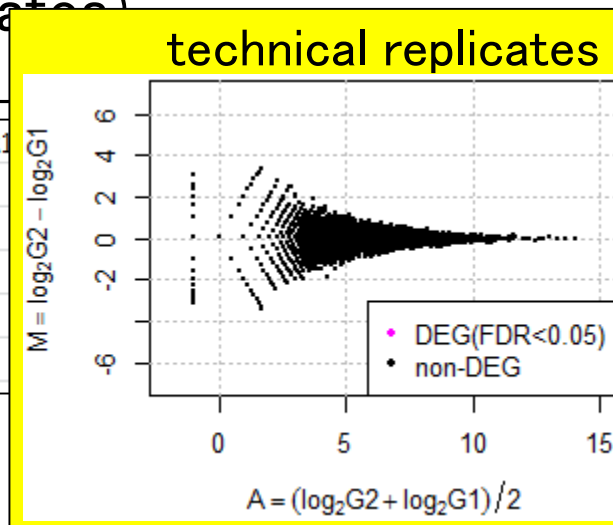
data_marioni.txt (technical replicates)

腎臓(Kidney)群

rownames(data)	R1L1Kidney	R1L3Kidney	R1L7Kidney	R2L2Kidney	R2L6Kidney	R2L7Kidney
ENSG000000000003	178	167	179	172	151	161
ENSG000000000005	0	0	0	0	1	0
ENSG0000000000419	53	78	64	72	71	68
ENSG0000000000457	22	33	30	27	30	28
ENSG0000000000460	9	7	18	14	9	11

G1群

G2群



data_arab.txt (biological replicates)

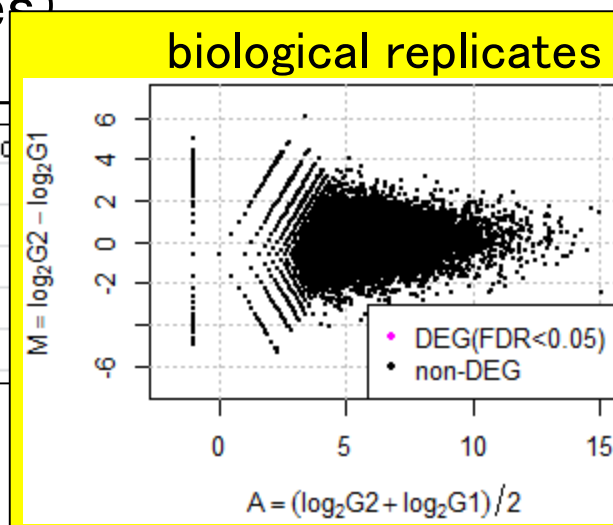
mock群

hrcc群

identifier	mock1	mock2	mock3	hrcc1	hrcc2	hrcc3
AT1G01010	35	77	40	46	64	50
AT1G01020	43	45	32	43	39	42
AT1G01030	16	24	26	27	35	28
AT1G01040	72	43	64	66	25	51
AT1G01050	49	78	90	67	45	71

G1群

G2群



Biological replicatesデータのほうが同一群のばらつきが大きい

他の公共カウントデータでも確認できます

- [はじめに](#) (last modified 2014/01/30) **NEW**
- [Rのインストールと起動](#) (last modified 2013/09/27)
- [サンプルデータ](#) (last modified 2014/02/09) **NEW**
- [イン](#)
- [イン](#)
- [イン](#)
- [イン](#)


使い慣れているので、私はReCountのデータをよく利用しています。自分でもいろいろと試してみましよう。

サンプルデータ **NEW**

1. [Marioni et al., Genome Res., 2008](#)の Supplementary table 2のデータ。
8. [NBPSeg](#)パッケージ([Di et al., SAGMB, 10:art24, 2011](#))中の *Arabidopsis*の Biological replic vs. G2群3サンプル; [Cumbie et al., PLoS One, 2011](#))です。
26,221 genes×6 samplesの「複製あり」タグカウントデータ([data_arab.txt](#))
オリジナルは"AT4G32850"というIDのものが重複して存在していたため、19520行目のテキストファイルにしています。
9. [ReCount](#)データベース([Frazee et al., BMC Bioinformatics, 2011](#))
マッピング済みの遺伝子発現行列形式のデータセットを多数提供しています。
10. Ye

ReCount

A multi-experiment resource of analysis-ready RNA-seq gene count datasets



**JOHNS HOPKINS
BLOOMBERG
SCHOOL OF PUBLIC HEALTH**

Study	PMID	Species	Number of biological replicates	Number of uniquely aligned reads	ExpressionSet	Count table	Phenotype table	Notes
bodymap	22496456	human	19	2,197,622,796	link	link	link	Illumina Human BodyMap 2.0 -- tissue comparison
cheung	20856902	human	41	834,584,950	link	link	link	HapMap - CEU
core	19056941	human	2	8,670,342	link	link	link	lung fibroblasts
gilad	20009012	human	6	41,356,738	link	link	link	liver; males

ReCo
differ
for ea
form
use a
one e

発現変動解析用Rパッケージ

- *DEGSeq* (Wang *et al.*, *Bioinformatics*, **26**: 136-138, 2010)
- *edgeR* (Robinson *et al.*, *Bioinformatics*, **26**: 139-140, 2010)
- *GPseq* (Srivastava and Chen, *Nucleic Acids Res.*, **38**: 1153-1162, 2010)
- *baySeq* (Hardcastle and Kelly, *BMC Bioinformatics*, **11**: 175, 2010)
- *DESeq* (Anders and Huber, *Genome Biol.*, **11**: R106, 2010)
- *NBPSeq* (Di *et al.*, *SAGMB*, **10**: article24, 2011)
- *TPSM* (Auer and Doerge, *SAGMB*, **10**: article26, 2011)
- *BBSeq* (Zhou *et al.*, *Bioinformatics*, **27**: 2672-2678, 2011)
- *NOISeq* (Tarazona *et al.*, *Genome Res.*, **21**: 2213-2221, 2011)
- *PoissonSeq* (Li *et al.*, *Biostatistics*, **13**: 523-538, 2012)
- *SAMseq* (Li and Tibshirani, *Stat Methods Med Res.*, **26**: 101-110, 2012)
- *BitSeq* (Glaus *et al.*, *Bioinformatics*, **28**: 1721-1728, 2012)
- *DEXSeq* (Anders *et al.*, *Genome Res.*, **22**: 2008-2017, 2012)
- *ShrinkBayes* (Van DE Wiel *et al.*, *Biostatistics*, **14**: 111-120, 2013)
- *sSeq* (Yu *et al.*, *Bioinformatics*, **29**: 1275-1282, 2013)
- *TCC* (Sun *et al.*, *BMC Bioinformatics*, **14**: 219, 2013)
- ...

解析 | 発現変動 | 2群間 | 対応なし | について

実験デザインが以下のような場合にこのカテゴリーに属す方法を適用

- Aさんの正常サンプル
- Bさんの正常サンプル
- Cさんの正常サンプル
- Dさんの腫瘍サンプル
- Eさんの腫瘍サンプル
- Fさんの腫瘍サンプル
- Gさんの腫瘍サンプル

2013年8月に調査した結果をリストアップします。

- [DEGSeq: Wang et al., Bioinformatics, 2010](#)
- [edgeR: Robinson et al., Bioinformatics, 2010](#)
- [GPseq: Srivastava et al., Nucleic Acids Res., 2010](#)
- [baySeq: Hardcastle and Kelly, BMC Bioinformatics, 2010](#)
- [DESeq: Anders and Huber, Genome Biol., 2010](#)
- [DESeq2: Anders and Huber, Genome Biol., 2010](#)
- [NBPSeq: Di et al., SAGMB, 2011](#)
- [BBSeq: Zhou et al., Bioinformatics, 2011](#)
- [NOISeq: Tarazona et al., Genome Res., 2011](#)
- [PoissonSeq: Li et al., Biostatistics, 2012](#)
- [SAMseq: Li and Tibshirani, Stat Methods Med Res., 2012](#)
- [BitSeq: Glaus et al., Bioinformatics, 2012](#)
- [easyRNASeq: Delhomme et al., Bioinformatics, 2012](#)
- [ShrinkBayes: Van De Wiel et al., Biostatistics, 2013](#)
- [DSGseq: Wang et al., Gene, 2013](#)
- [sSeq: Yu et al., Bioinformatics, 2013](#)
- [TCC: Sun et al., BMC Bioinformatics, 2013](#)
- [tweeDEseq: Esnaola et al., BMC Bioinformatics, 2013](#)
- [NPEBseq: Bi et al., BMC Bioinformatics, 2013](#)

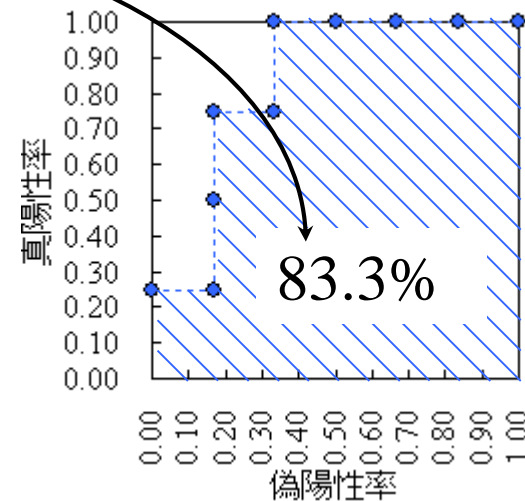
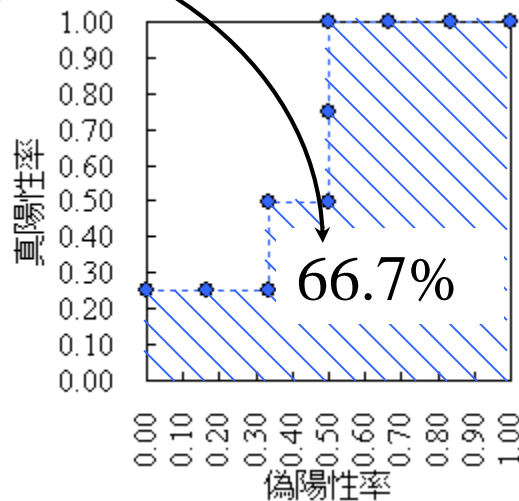
黒字のものたち (+ α) の比較結果は...

- *DEGSeq* (Wang *et al.*, *Bioinformatics*, 26: 136-138, 2010)
- *edgeR* (Robinson *et al.*, *Bioinformatics*, 26: 139-140, 2010)
- *GPseq* (Srivastava and Chen, *Nucleic Acids Res.*, 38: e170, 2010)
- *baySeq* (Hardcastle and Kelly, *BMC Bioinformatics*, 11: 422, 2010)
- *DESeq* (Anders and Huber, *Genome Biol.*, 11: R106, 2010)
- *NBPSeq* (Di *et al.*, *SAGMB*, 10: article24, 2011)
- *TPSM* (Auer and Doerge, *SAGMB*, 10: article26, 2011)
- *BBSeq* (Zhou *et al.*, *Bioinformatics*, 27: 2672-2678, 2011)
- *NOISeq* (Tarazona *et al.*, *Genome Res.*, 21: 2213-2223, 2011)
- *PoissonSeq* (Li *et al.*, *Biostatistics*, 13: 523-538, 2012)
- *SAMseq* (Li and Tibshirani, *Stat Methods Med Res.*, 2011 Nov 28)
- *BitSeq* (Glaus *et al.*, *Bioinformatics*, 28: 1721-1728, 2012)
- *DEXSeq* (Anders *et al.*, *Genome Res.*, 22: 2008-2017, 2012)
- *ShrinkBayes* (*ShrinkSeq*; Van DE Wiel *et al.*, *Biostatistics*, 14: 113-128, 2013)
- *sSeq* (Yu *et al.*, *Bioinformatics*, 29: 1275-1282, 2013)
- *TCC* (Sun *et al.*, *BMC Bioinformatics*, 14: 219, 2013)

よりよい方法とは？

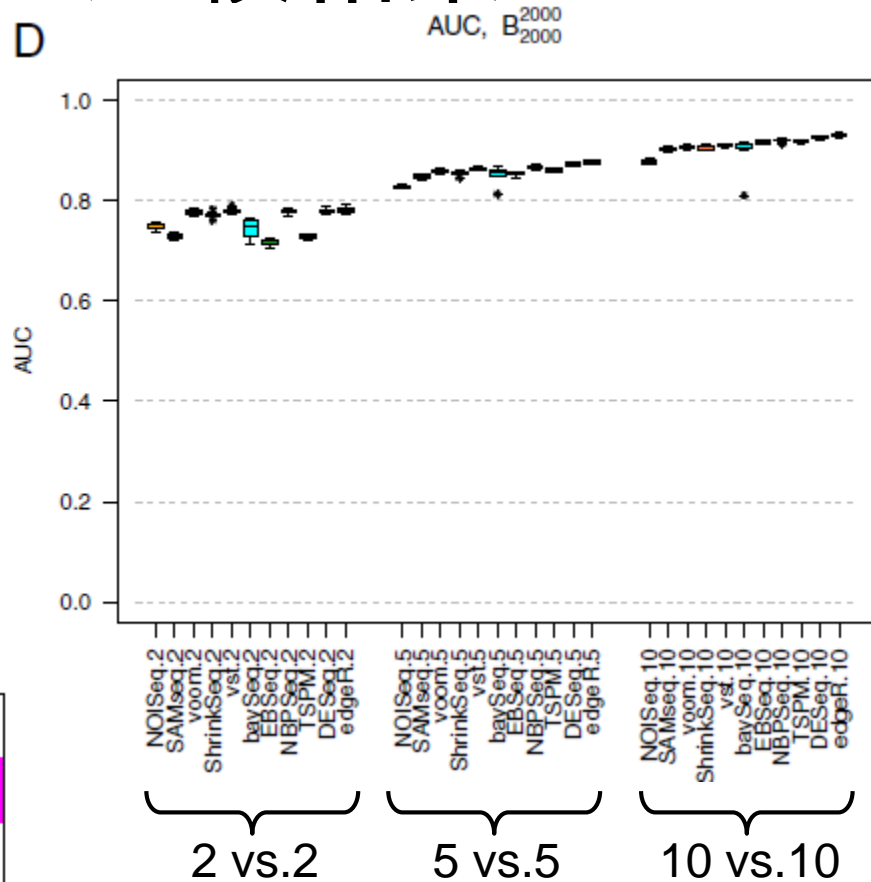
- その方法を用いて発現変動の度合いでランキングしたときに、**真の発現変動遺伝子 (DEG)** がより上位にランキングされる (感度・特異度高い)

ランキング法			
rank	方法1		方法2
1	gene8	真	gene8
2	gene5	偽	gene5
3	gene4	偽	gene3
4	gene3	真	gene2
5	gene7	偽	gene7
6	gene1	真	gene1
7	gene2	真	gene4
8	gene9	偽	gene9
9	gene10	偽	gene10
10	gene6	偽	gene6

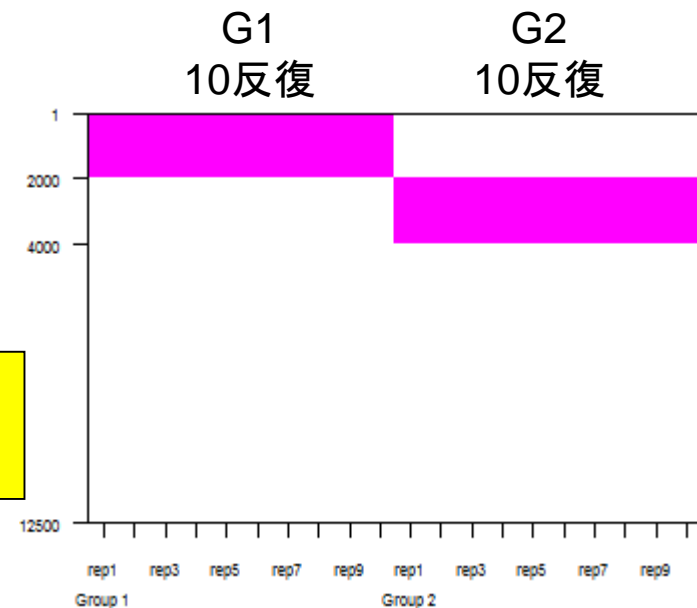
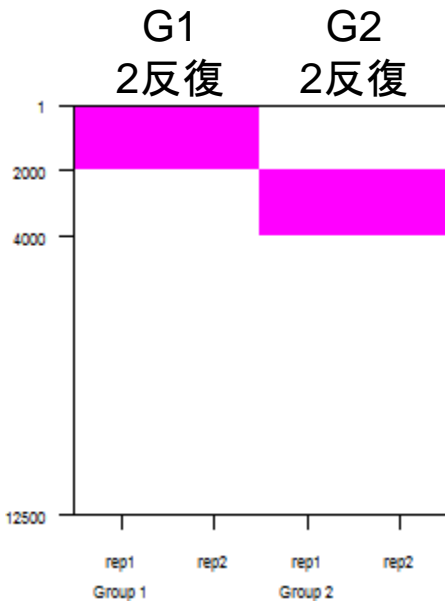


Area Under the ROC Curve (ROC曲線の下部面積: AUC)
 バイオインフォマティクス分野でよく用いられる評価基準です

AUC値の比較結果

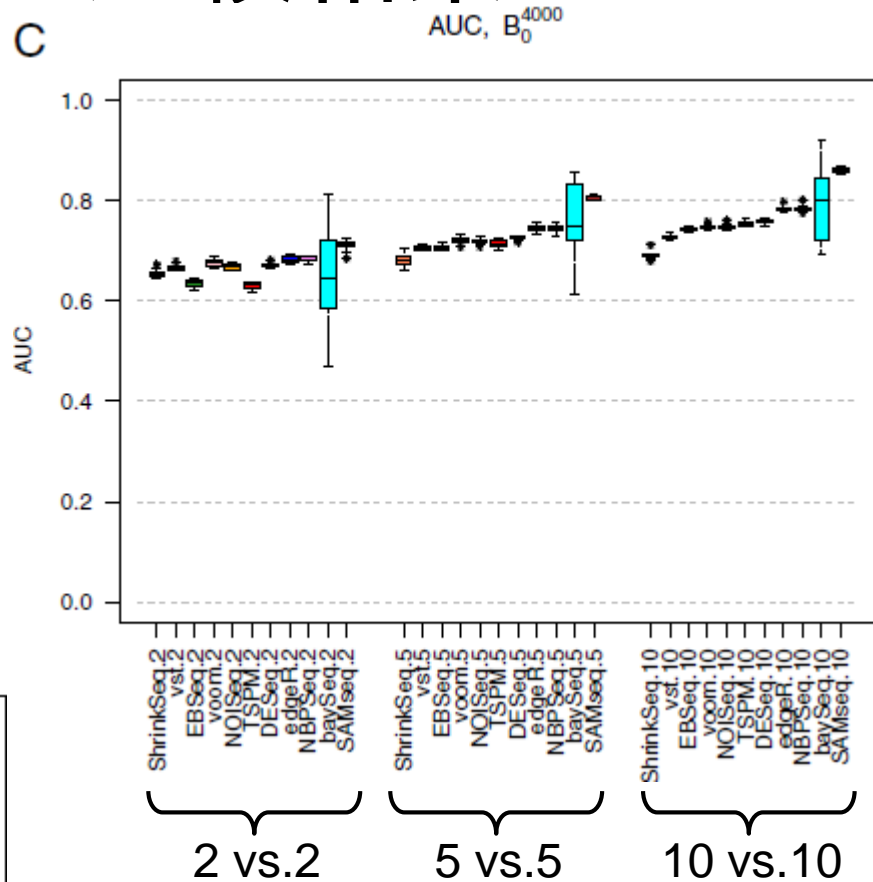


シミュレーション条件: G1 vs. G2
 全遺伝子数: 12500
 発現変動遺伝子(DEG)数: 4000
 G1で高発現: 2000
 G2で高発現: 2000
unbiased DE situation

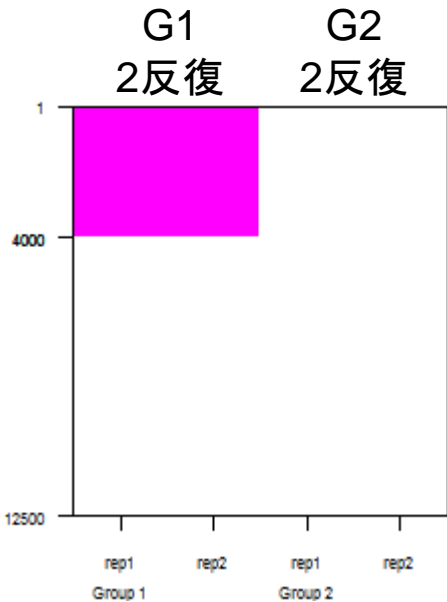


反復実験数を増やすほど精度は上がる
 (これが言いたいわけではない...)

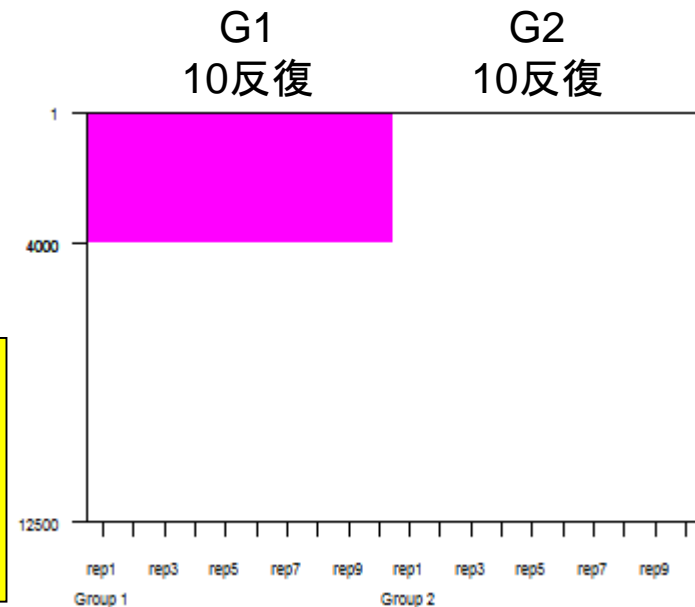
AUC値の比較結果



シミュレーション条件: G1 vs. G2
 全遺伝子数: 12500
 発現変動遺伝子(DEG)数: 4000
 G1で高発現: 4000
 G2で高発現: 0
biased DE situation

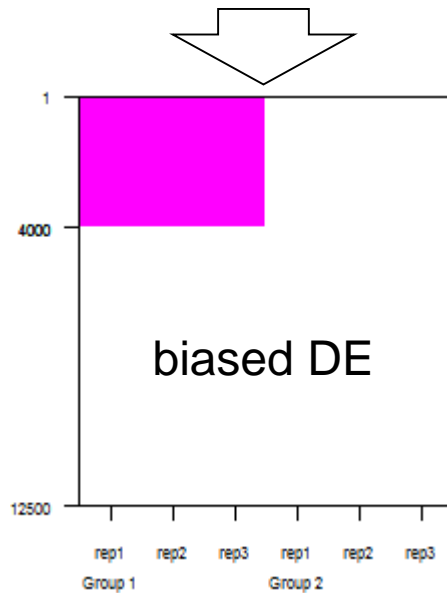
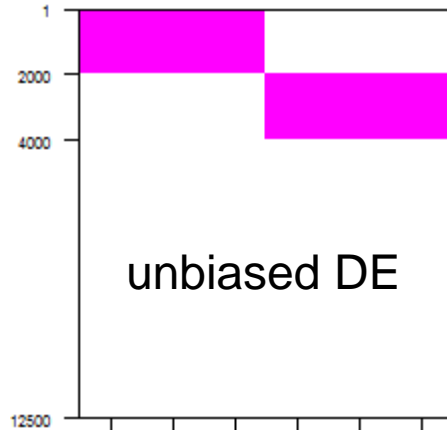


グループ(群)間でDEG数の組成に偏りがあると精度が大幅に低下する
 理由: データ正規化法がDEG数の組成に偏りが無いことを想定しているため



AUC値の比較結果

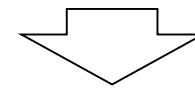
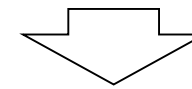
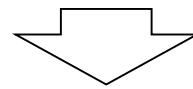
3反復 vs. 3反復



edgeR
90.84%

SAMseq
87.19%

TCC
90.83%



edgeR
82.95%

SAMseq
84.40%

TCC
89.92%

偏りのないデータの場合はedgeRがよい
偏りのあるデータの場合はSAMseqがよい
→ 偏りの有無に関係なくTCCがよい

DEGESって何デゲス？

～ アラフォー達の略称に関する議論 ～

門田:「DESで行くデス」

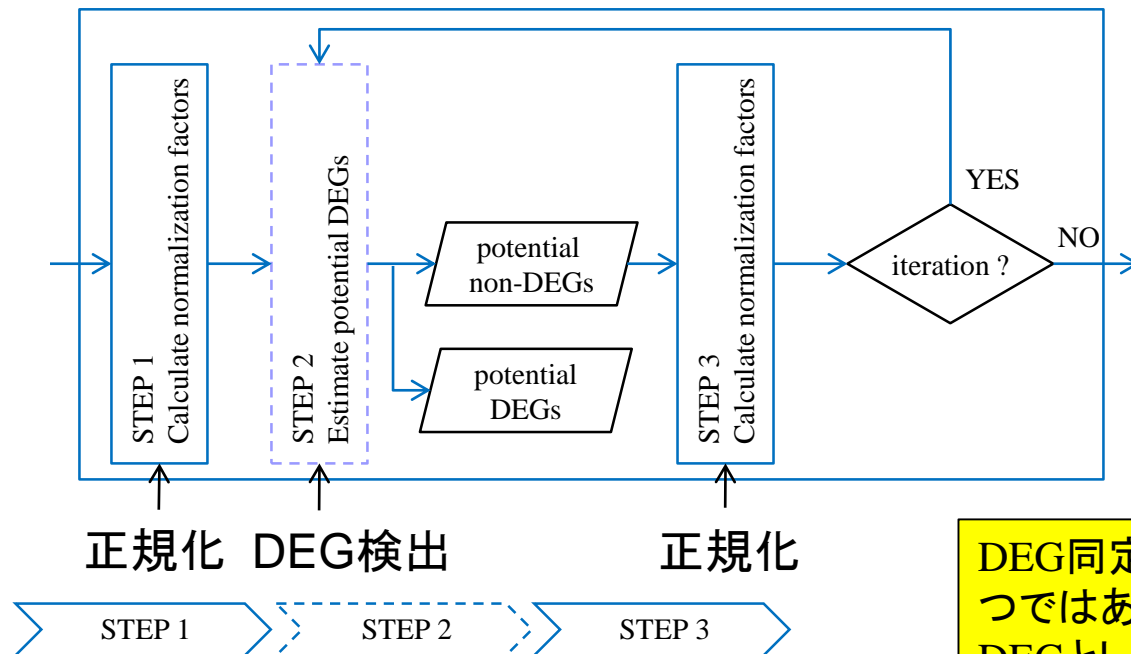
西山:「DEGESはいかが？」

門田:「面白くないので却下！」

西山:「左様デゲスカ...DEGESって何デゲス？」

門田:「採用！」

■ 概念図



RNA-seqなどから得られるタグカウントデータの正規化をmulti-stepで行う概念の総称

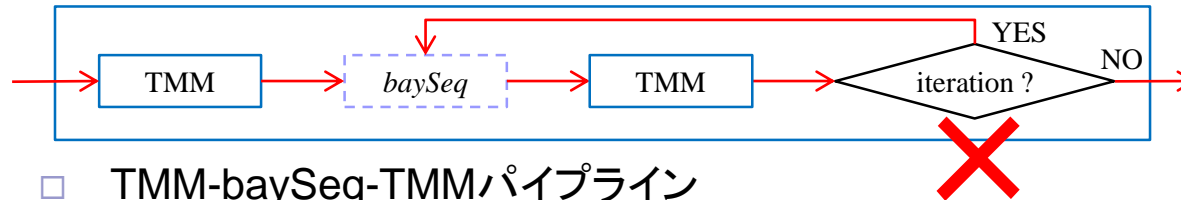
DEG同定を正確に行うのが正規化の目的の一つではあるが、正規化時にDEGの存在自体がDEGとして同定されるのを阻むことがわかった(自爆テロ)。それゆえ、正規化時にDEGの検出を行って、non-DEGのみ利用するのがポイント

DEGESって何デゲス？

- DEGESのstep1-3で内部的に用いる方法は実用上なんでも?!よい

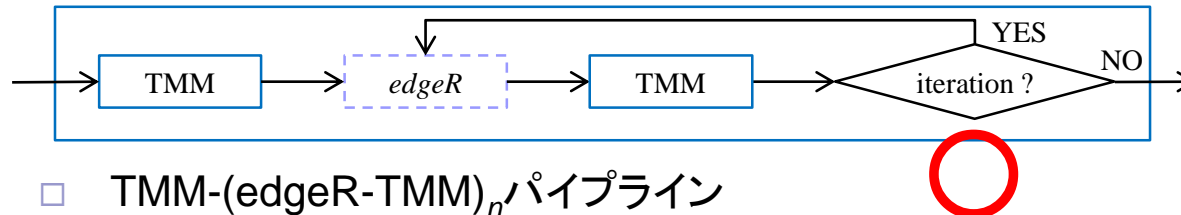


- TbT正規化法 (Kadota et al., 2012)



- TMM-baySeq-TMMパイプライン
- step2でbaySeqパッケージ中のDEG同定法(経験ベイズ)を利用しているため遅い...
- Iterative TbT(step2-3を繰り返してより頑健な正規化係数を得る)は非現実的

- iDEGES/edgeR正規化法 (Sun et al., 2013)



- TMM-(edgeR-TMM)_nパイプライン
- Step2でedgeRパッケージ中のDEG同定法(exact test)を利用しているため速い!
- DEGESをiterativeに行う頑健なiDEGES(愛デゲス)パイプラインを利用可能

TCCパッケージ(ver. 1.0.0)に実装済み

どういうデータのとときに有効デゲスか？

■ 仮想データ (10,000 genes × 6 samples)

□ 2,000 DEGs (20%がDEG)

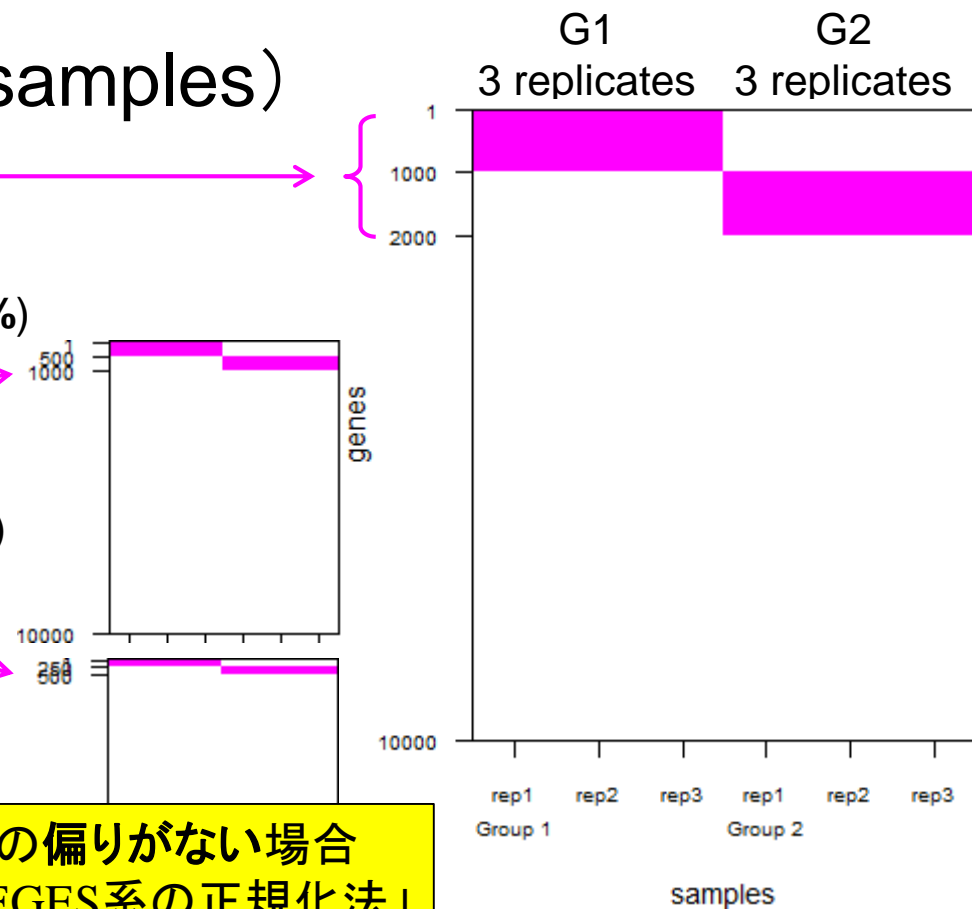
- Group1 (G1)で高発現: gene1~1000 (50%)
- Group2 (G2)で高発現: gene1001~2000 (50%)

□ 1,000 DEGs (10%がDEG)

- Group1 (G1)で高発現: gene1~500 (50%)
- Group2 (G2)で高発現: gene501~1000 (50%)

□ 500 DEGs (5%がDEG)

- Group1 (G1)で高発現: gene1~250 (50%)
- Group2 (G2)で高発現: gene251~500 (50%)



DEG数のGroup間での偏りがない場合
「TMM正規化法」と「DEGES系の正規化法」
の理論上の性能は互角デゲス。

どういうデータのとときに有効デゲスか？

■ 仮想データ (10,000 genes × 6 samples)

□ 2,000 DEGs (20%がDEG)

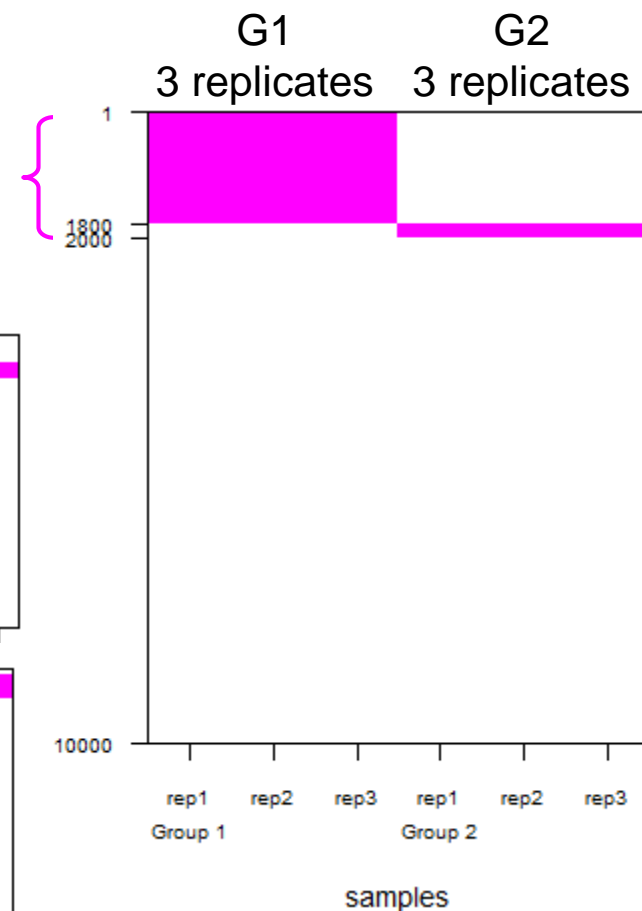
- Group1 (G1)で高発現: gene1~1800 (90%)
- Group2 (G2)で高発現: gene1801~2000 (10%)

□ 1,500 DEGs (15%がDEG)

- Group1 (G1)で高発現: gene1~900 (60%)
- Group2 (G2)で高発現: gene901~1500 (40%)

□ 1,000 DEGs (10%がDEG)

- Group1 (G1)で高発現: gene1~200 (20%)
- Group2 (G2)で高発現: gene201~1000 (80%)



DEGES系正規化法は、DEG数のGroup間での偏りが大きいほど有効なんデゲス！

