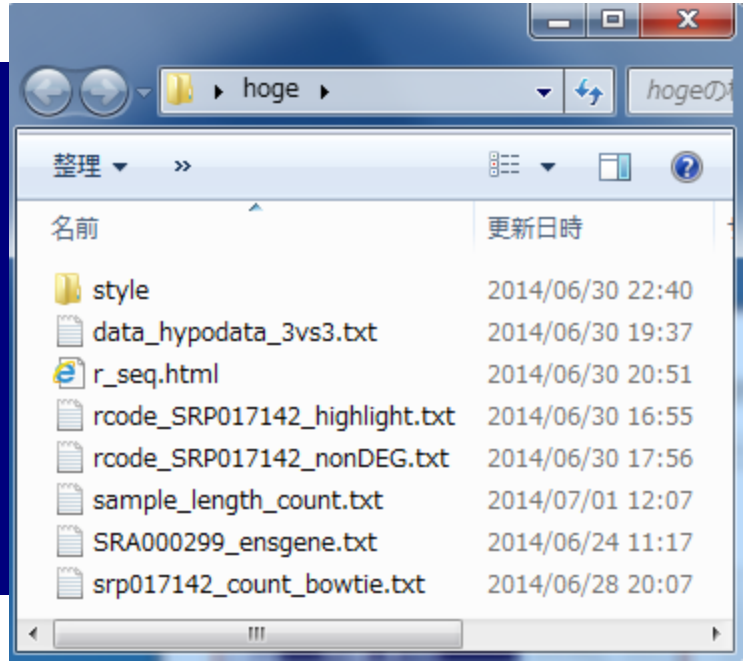
講義室後ろにあるUSBメモリ中のhogeフォルダをデスクトップにコピーしておいてください。



農学生命情報科学特論I 第4回

東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
門田幸二
kadota@iu.a.u-tokyo.ac.jp

前回の課題と正答

- アダプター配列除去前後のsmall RNA-seqデータをカイコゲノムにマップし、マップ率（マップされたリード数）を比較する
  1. マッピング**前**の総リード数を示せ
     - アダプター配列除去前のSRR609266.fastq.gz： 11,928,428 リード
     - アダプター配列除去後のhoge4.fastq.gz： 11,928,428 リード
  2. マッピング**後**の「**マップされたリード数**」を示せ
     - アダプター配列除去前のSRR609266.fastq.gz： 2,257 リード
     - アダプター配列除去後のhoge4.fastq.gz： 1,308,126 リード
  3. 結果の考察。

マッピング後の**総リード数**で
はなく、**マップされたリード数**
が正解ですね、失礼しました。

Contents（第4回）

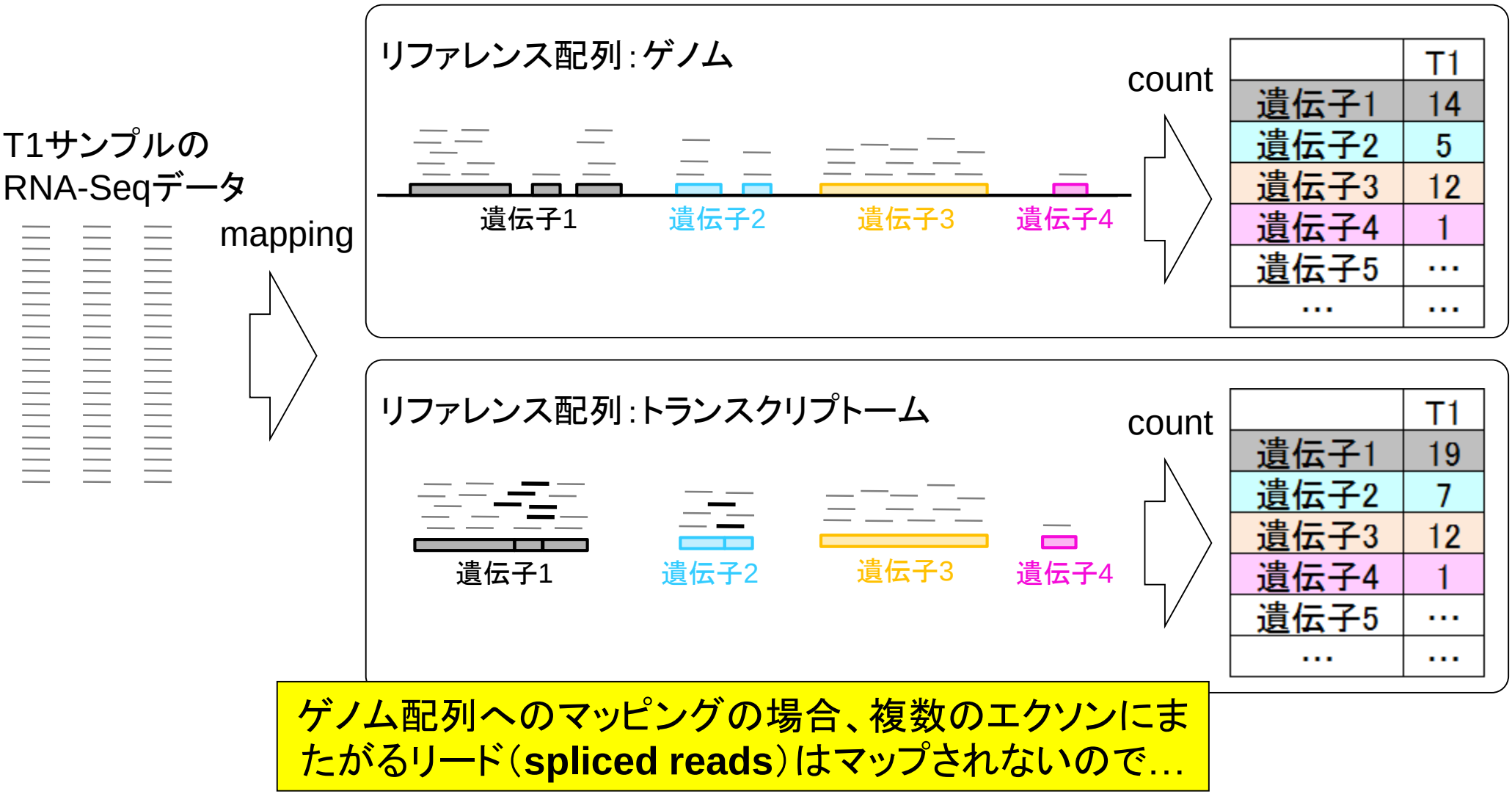
- 新規転写物同定（ゲノム情報を利用）
  - 基本的な考え方
  - Tophat-Cufflinksパイプライン
  - 可視化（ゲノムブラウザやViewer）
- 発現量推定（遺伝子レベルと転写物レベル）
  - RPKMの基本的な考え方
  - 計算時間短縮戦略（トランスクリプトーム情報のみを利用）
- カウントデータを用いたサンプル間比較解析
  - イントロ（カウントデータ取得まで）
  - サンプル間クラスタリング
  - 発現変動遺伝子検出
  - 分布やモデル
  - 課題

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などを用いて遺伝子構造推定（アノテーション）
  - 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
- 遺伝子または転写物(isoform)ごとの発現量の正確な推定
  - RSEM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 biasなどの各種補正がポイント
- 比較するサンプル間で発現変動している遺伝子または転写物の同定
  - *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
  - （GO解析など）DEG結果を用いる多くの下流解析結果に影響を及ぼす

マッピングの基本的なイメージ

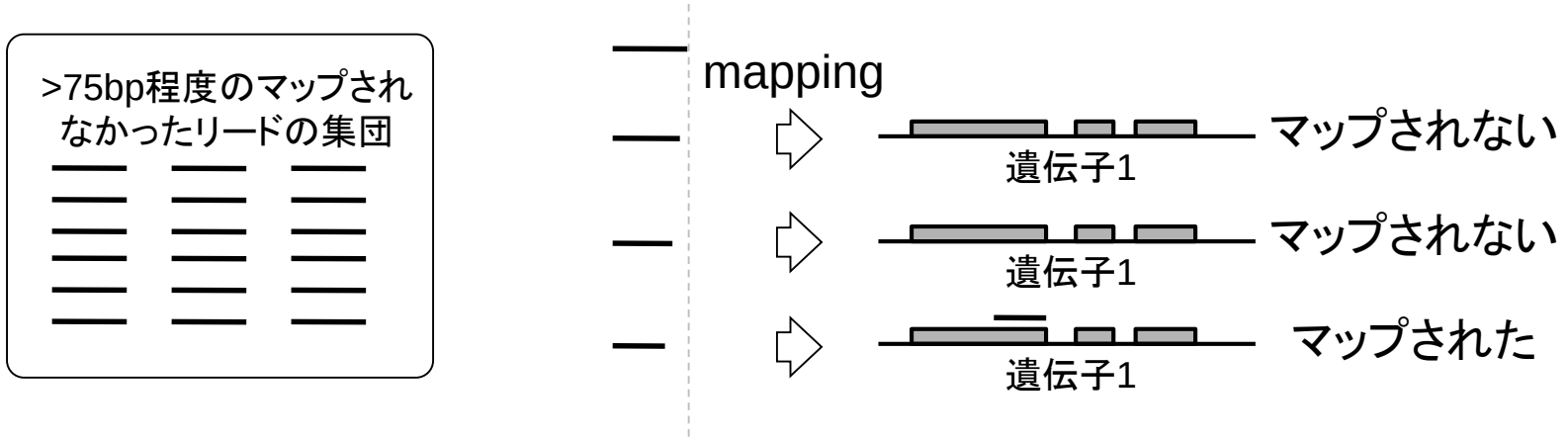
- 基本的なマッピングプログラム（bowtieなど）を用いた場合



ゲノム配列へのマッピングの場合、複数のエクソンにまたがるリード（**spliced reads**）はマップされないので…

対策（リード長が75bp程度以上の現在）

- 再帰的にマッピングする戦略（recursive mapping strategy）
  - 通常のマッピングプログラムでマップされなかったものに対して、リードを短くしてマップされるかどうかを繰り返すというイメージ



splice-aware aligner (spliced aligner)を用いることで新規転写物の同定も可能。理由は既知遺伝子構造情報を参照しなくてもどうにかなるから。

Splice-aware alignerの様々な戦略

Garber et al., *Nat. Methods*, **8**: 469-477, 2011のFig. 1

exon-first系は高速だがアルゴリズム的にprocessed pseudogene存在下で正確な構造推定が困難になる

Basic aligner (unspliced aligner)

- アセンブル | ゲノム用(last modified 2014/06/15) NEW
- アセンブル | トランスクリプトーム(転写物)用(last modified 20
- マッピング | について (last modified 2014/06/24) NEW
- マッピング | basic aligner (last modified 2014/06/24) NEW
- マッピング | splice-aware aligner (last modified 2014/06/24) N
- マッピング | Bisulfite sequencing用(last modified 2014/06/24)
- マッピング | (ESTレベルの長さの)contig(last modified 2014/
- 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(

マッピング | について NEW

リファレンス配列にマッピングを行うプログラム達です。basic aligner (unspliced aligner)はsplice-aware aligner (spliced aligner)内部で使われていたりします。

R用:

- Rsubread(Windows版なし): Liao et al., Nucleic Acids Res., 2013
- QuasR(Windows版あり): 原著論文はまだみたいです
- HTSeqGenie(Windows版なし): 原著論文はまだみたいです

R以外(basic aligner; unspliced aligner):

- SSAHA2: Ning et al., Genome Res., 2001
- RMAP: Smith et al., BMC Bioinformatics, 2008
- MAQ: Li et al., Genome Res., 2008
- MOM: Eaves and Gao, Bioinformatics, 2009
- Bowtie: Langmead et al., Genome Biol., 2009
- BWA: Li and Durbin, Bioinformatics, 2009(BWA-shortの論
- SHRiMP: Rumble et al., PLoS Comput. Biol., 2009
- SOAP2: Li et al., Bioinformatics, 2009
- RazerS: Weese et al., Genome Res., 2009
- PerM: Chen et al., Bioinformatics, 2009
- BFAST: Homer et al., PLoS One, 2009
- BWA: Li and Durbin, Bioinformatics, 2010(BWA-SWの論文)
- Novoalign: Krawitz et al., Bioinformatics, 2010
- GASSST: Rizk and Lavenier, Bioinformatics, 2010
- Stampy: Lunter and Goodson, Genome Res., 2011
- SHRiMP2: David et al., Bioinformatics, 2011
- SOAP3: Liu et al., Bioinformatics, 2012
- FANSe: Zhang et al., Nucleic Acids Res., 2012
- Bowtie 2: Langmead and Salzberg, Nat. Methods, 2012
- CUSHAW2: Liu and Schmidt, Bioinformatics, 2012
- CUDASW++ 3.0: Liu et al., BMC Bioinformatics, 2013
- Subread: Liao et al., Nucleic Acids Res., 2013
- SOAP3-dp: Luo et al., PLoS One, 2013

Windowsでマッピング可能なRパッケージ。内部的にbasic alignerのbowtieとsplice-aware alignerのSpliceMapを利用可能

比較的よく使われているもの

Splice-aware aligner (spliced aligner)

- アセンブル | ゲノム用(last modified 2014/06/15) NEW
- アセンブル | トランスクリプトーム(転写物)用(last modified 2014/06/20) NEW
- マッピング | について(last modified 2014/06/24) NEW
- マッピング | basic aligner(last modified 2014/06/24) NEW
- マッピング | splice-aware aligner(last modified 2014/06/24) NEW
- マッピング | Bisulfite sequencing用(last modified 2014/06/24) NEW
- マッピング | (ESTレベルの長さの)contig(last modified 2014/06/24) NEW
- 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)

R以外(splice-aware aligner; spliced aligner):

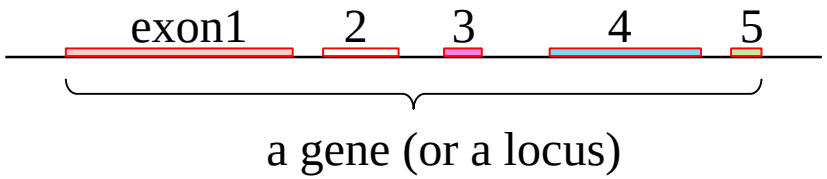
- BLAT: Kent WJ, Genome Res., 2002
- QPALMA: De Bona et al., Bioinformatics, 2008
- TopHat: Trapnell et al., Bioinformatics, 2009
- RNA-MATE: Cloonan et al., Bioinformatics, 2009
- GSNAP: Wu et al., Bioinformatics, 2010
- SpliceMap: Au et al., Nucleic Acids Res., 2010
- MapSplice: Wang et al., Nucleic Acids Res., 2010
- HMMSplicer: Dimon et al., PLoS One, 2010
- X-MATE: Wood et al., Bioinformatics, 2011
- RNASEQR: Chen et al., Nucleic Acids Res., 2012
- PASSion: Zhang et al., Bioinformatics, 2012
- ContextMap: Bonfert et al., BMC Bioinformatics, 2012
- STAR: Dobin et al., Bioinformatics, 2013
- TrueSight: Li et al., Nucleic Acids Res., 2013
- OLego: Wu et al., Nucleic Acids Res., 2013

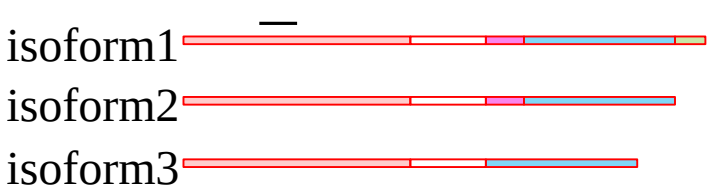
比較的よく使われているもの。Tophatは内部的にBowtieを利用(今はBowtie 2かも…)

Windowsでマッピング可能なRパッケージ。内部的にbasic alignerのbowtieとsplice-aware alignerのSpliceMapを利用可能

Reference-based strategy

- Splice-aware aligner出力結果をもとに遺伝子構造推定
  - Scripture (Guttman et al., *Nat. Biotechnol.*, **28**: 503-510, 2010)
  - **Cufflinks (Trapnell et al., *Nat. Biotechnol.*, 28: 511-515, 2010)**
  - STM (Surget-Groba and Montoya-Burgos, *Genome Res.*, **20**: 1432-1440, 2010)
  - ALEXA-seq (Griffith et al., *Nat. Methods*, **7**: 843-847, 2010)
  - ARTADE2 (Kawaguchi et al., *Bioinformatics*, **28**: 929-937, 2012)
  - …
- このtranscriptome reconstruction作業は結構大変
  - 理由1：広いダイナミックレンジ（低発現のものとノイズとの区別）
  - 理由2：off-targetの存在（mature mRNA以外のprecursor RNAなど）
  - 理由3：一つの遺伝子から複数のisoforms（どのisoform由来のリードか?!）





遺伝子構造推定のイメージ

Martin and Wang, *Nature Reviews Genet*., **12**: 671-682, 2011のFig. 2

- 解析 | 基礎 | 平均-分散プロット | Technical replicates(last modified 2014/02/18)
- 解析 | 基礎 | 平均-分散プロット | Biological replicates(last modified 2014/02/21)
- 解析 | 新規転写物同定(ゲノム配列を利用)(last modified 2014/06/23) NEW
- 解析 | 発現量推定(トランスクリプトーム配列を利用)(last modified 2014/06/23) NEW

- 解析 | 新規転写物同定(ゲノム配列を利用)

解析 | 新規転写物同定(ゲノム配列を利用) NEW

reference-based methodsというカテゴリに含まれるものたちです。ゲノム配列にRNA-seqデータのマッピングを行ってどこに遺伝子領域があるかなどの座標(アノテーション)情報を取得する遺伝子構造推定用です。GeneScissorsというのは、TopHat/Cufflinks系の有名なプログラムの出力結果をもとに間違いを修正するなどして精度を高めるもののようです。実質的にde novo transcriptome assemblyと目指すところは同じですが、やはりゲノムというリファレンス配列を用いるほうがより正確であるため、ゲノム配列が利用可能な場合は利用するのが一般的です。一般にこの種のプログラムは遺伝子構造推定だけでなく、発現量推定まで行ってくれます。2014年6月に調べた結果

R用:

- Solas(Windows版はなさそう; 2010年以降アップデートなし): Richard et al., Nucleic Aci
- NSMAP(Windows版はなさそう;アノテーションファイルなしでの実行に特化): Suo et al
- Sequgio: Suo et al., Bioinformatics, 2014
- spliceR(Cufflinksの出力をインプットにしているようだ): Vitting-Seerup et al., BMC Bioi

R以外:

- Scripture: Guttman et al., Nat Biotechnol., 2010
- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- rQuant: Bohnert and Ratsch, Nucleic Acids Res., 2010
- STM: Surget-Groba and Montoya-Burgos, Genome Res., 2010
- ALEXA-seq: Griffith et al., Nat Methods, 2010
- MISO: Katz et al., Nat. Methods, 2010
- MMSEQ: Turro et al., Genome Biol., 2011
- IsoEM: Nicolae et al., Algorithms Mol. Biol., 2011
- IsoformEx: Kim et al., BMC Bioinformatics, 2011
- RSEM: Li and Dewey, BMC Bioinformatics, 2011
- SLIDE: Li et al., PNAS, 2011
- BitSeq: Glaus et al., Bioinformatics, 2012
- ARTADE2: Kawaguchi et al., Bioinformatics, 2012
- RD: Wan et al., Biostatistics, 2012
- Plntron: Pirola et al., BMC Bioinformatics, 2012
- CEM: Li and Jiang, Bioinformatics, 2012
- eXpress: Roberts and Pachter, Nat Methods, 2013
- iReckon: Mezlini et al., Genome Res., 2013
- TrueSight: Li et al., Nucleic Acids Res., 2013
- PASTA: Tang et al., BMC Bioinformatics, 2013
- GeneScissors: Zhang et al., Bioinformatics, 2013
- TIGAR: Nariai et al., Bioinformatics, 2013
- PSGInfer: LeGault et al., Bioinformatics, 2013
- NURD: Ma et al., BMC Bioinformatics, 2013
- MITIE: Behr et al., Bioinformatics, 2013
- UnSplicer: Burns et al., Nucleic Acids Res., 2014
- eXpress-D: Roberts et al., BMC Bioinformatics, 2013
- PennSeq: Hu et al., Nucleic Acids Res., 2014
- Parseq: Mirauta et al., Bioinformatics, 2014
- FineSplice: Gatto et al., Nucleic Acids Res., 2014

BowtieやTophatが多く引用されるのはCufflinksなど他のソフトウェア上でもよく実装されているためであろう

Bowtie-Tophat-Cufflinksパイプライン

Fig. 1

basic aligner

splice-aware aligner

Fig. 2

RNA-seqデータとリファレンス配列情報を入力として、遺伝子構造推定から発現量、発現変動解析、描画までの一連の解析を提供

Bowtie-Tophat-Cufflinksパイプライン

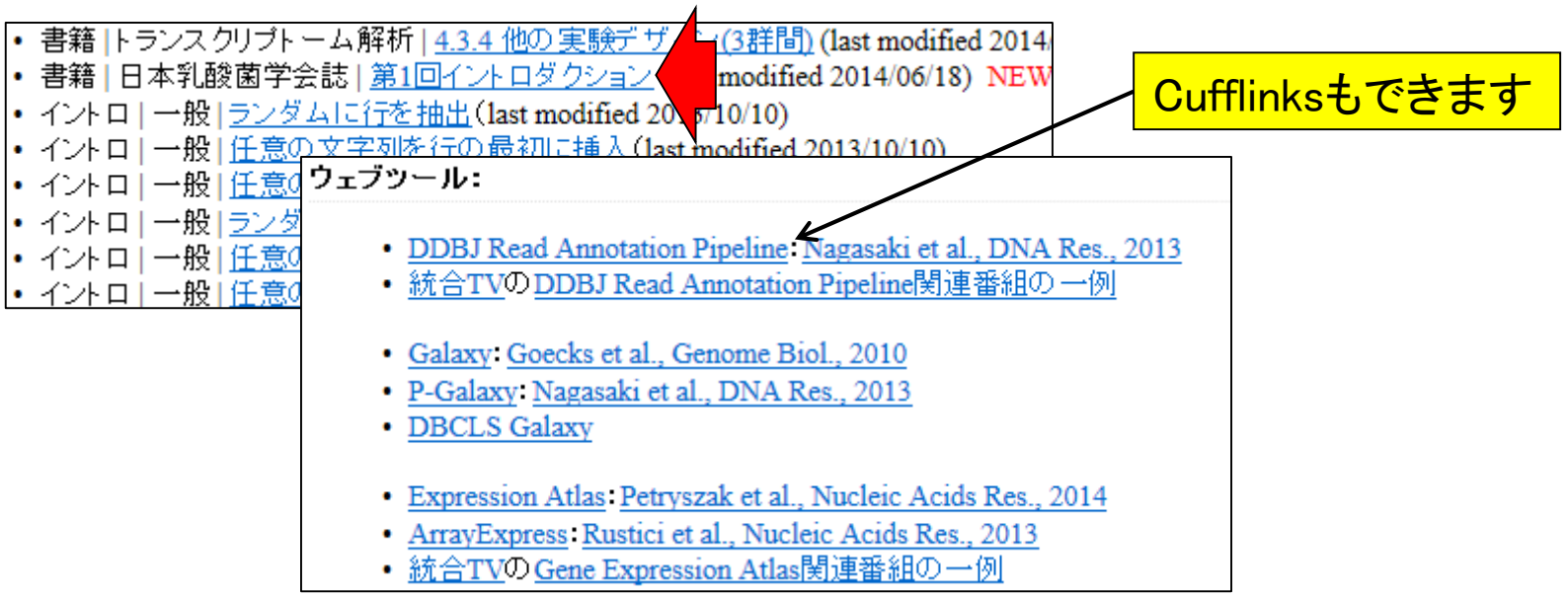
Fig. 2

RNA-seqデータとリファレンス配列情報を入力として、遺伝子構造推定から発現量、発現変動解析、描画までの一連の解析を提供

Fig. 3

NGSデータ解析手段

- 自前で大容量メモリ計算サーバ（Linux）を購入し、必要なソフトのインストールからスタート
  - 難易度は高いが思い通りの解析が可能
- Linuxサーバをもつバイオインフォ系の人にお願いする
  - 気軽に頼める知り合いがいればいいが、その人次第
- DDBJ Read Annotation Pipelineを利用
  - 一番お手軽な選択肢であり、有名どころはカバーされている



可視化(ゲノムブラウザやViewer)

- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck 2009)
- イントロ | NGS | 様々なプラットフォーム(last modified 2014/06/10) NEW
- イントロ | NGS | qPCRやmicroarrayなどとの比較(la modified 2014/06/05) N
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) modified 2014/06/25) N
- イントロ | NGS | 配列取得 | FASTQ or SRALite | 公共DBから(last modified 20
- イントロ | N
- イントロ | N
- イントロ | N
- イントロ | N
- イントロ | N

イントロ | NGS | 可視化(ゲノムブラウザやViewer) NEW

可視化ツールも結構あります。

R用:

- TileQC: Dolan and Denver, BMC Bioinformatics, 2008
- GenomeGraphs: Durinck et al., BMC Bioinformatics, 2009
- HilbertVis: Anders S., Bioinformatics, 2009
- rtracklayer: Lawrence et al., Bioinformatics, 2009
- genoPlotR: Guy et al., Bioinformatics, 2010
- ggbio: Yin et al., Genome Biol., 2012

R以外(ウェブベースのゲノムブラウザ; server-side):

Review(Wang et al., Brief Bioinform., 2013)によるserver-sideと client-sideの分類分けのうち、server-sideに相当するものたちだと思われます。

- VISTA: Frazer et al., Nucleic Acids Res., 2004
- Genome Projector(バクテリア系): Arakawa et al., BMC Bioinformatics, 2009
- Gramene(植物系): Jaiswal P, Methods Mol Biol., 2011
- Phytozome(植物系): Goodstein et al., Nucleic Acids Res., 2012
- The UCSC Genome Browser: Kuhn et al., Brief. Bioinform., 2013
  - 統合TVのUCSC Genome Browserの使い方～配列取得編～ 2013
- Ensembl: Flicek et al., Nucleic Acids Res., 2013
- ChromoZoom: Pak et al., Bioinformatics, 2013
- Genome Maps: Medina et al., Nucleic Acids Res., 2013
- CCDS: Farrell et al., Nucleic Acids Res., 2014

私は(数値解析系なので)可視化ツールは全く使いません

比較的よく使われているもの

可視化(ゲノムブラウザやViewer)

- イントロ | 一般 | 配列取得 | トランスクリプトーム配列 | biomaRt(Durinck 2009)
- イントロ | NGS | 様々なプラットフォーム(last modified 2014/06/10) NEW
- イントロ | NGS | qPCRやmicroarrayなどとの比較(la modified 2014/06/05) N
- イントロ | NGS | 可視化(ゲノムブラウザやViewer) modified 2014/06/25) N
- イントロ | NGS | 配列取得 | FASTQ or SRALite | 公DBから(last modified 20
- イントロ | N
- イントロ | N
- イントロ | N
- イントロ | N
- イントロ | N

私は(数値解析系なので)可視化ツールは全く使いません

イントロ | NGS | 可視化(ゲノムブラウザやViewer) NEW

可視化ツールも結構あります。

R用:

R以外(ゲノムブラウザ; client-side):

Review(Wang et al., Brief Bioinform., 2013)によるserver-sideと client-sideの分類分けのうち、client-sideに相当するものたちだと思われます。AJAX-based browsersともいうらしいです。

- GBrowse: Stein et al., Genome Res., 2002
- JBrowse: Skinner et al., Genome Res., 2009
- ABrowse: Kong et al., BMC Bioinformatics, 2012

比較的よく使われているもの

R以外(stand-alone系):

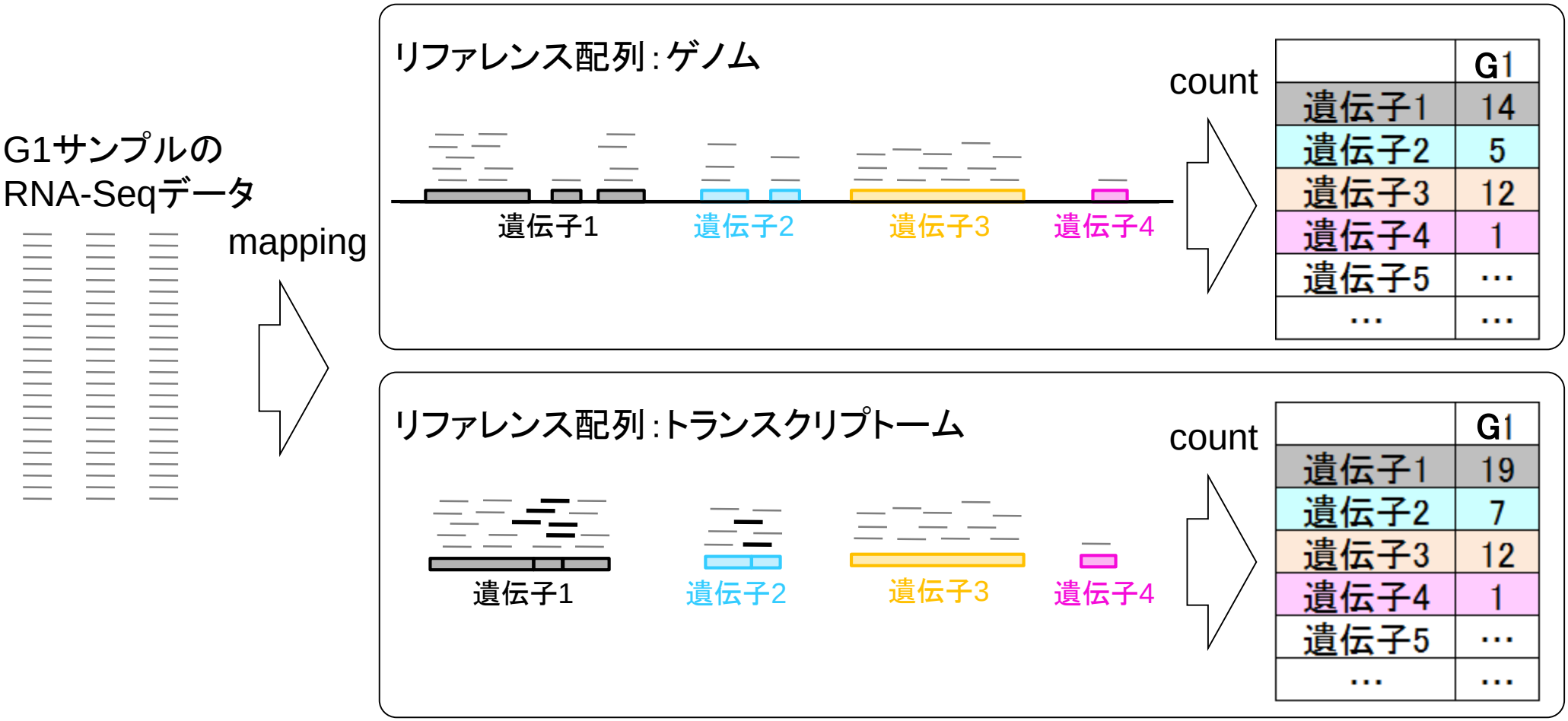
- EagleView: Huang and Marth, Genome Res., 2008 (Linux, Windows, and Mac)
- MagicViewer: Hou et al., Nucleic Acids Res., 2010 (Linux, Windows, and Mac)
- MapView: Bao et al., Bioinformatics, 2009 (Linux and Windows)
- NGSView: Arner et al., Bioinformatics, 2010 (Linux)
- Tablet: Milne et al., Bioinformatics, 2010 (Linux, Windows, and Mac)
- ZOOM Lite: Zhang et al., Nucleic Acids Res., 2010 (Linux and Windows)
- IGV: Thorvaldsdóttir et al., Brief. Bioinform., 2013 (Linux, Windows, and Mac)
  - 統合TVの Integrative Genomics Viewer IGVを使い倒す ～基本編～
- GenomeVISTA: Poliakov et al., Bioinformatics, 2014

Contents（第4回）

- 新規転写物同定（ゲノム情報を利用）
  - 基本的な考え方
  - Tophat-Cufflinksパイプライン
  - 可視化（ゲノムブラウザやViewer）
- 発現量推定（遺伝子レベルと転写物レベル）
  - **RPKMの基本的な考え方**
  - 計算時間短縮戦略（トランスクリプトーム情報のみを利用）
- カウントデータを用いたサンプル間比較解析
  - イントロ（カウントデータ取得まで）
  - サンプル間クラスタリング
  - 発現変動遺伝子検出
  - 分布やモデル
  - 課題

マップされたリード数 = 発現量ではないが…

- 基本的なマッピングプログラム（bowtieなど）を用いた場合



マップされたリード数のカウント情報は、発現量推定の基本情報です

研究目的別留意点：遺伝子間比較

- 発現量補正の基本形：$\text{カウント数} \times \frac{\text{定数}}{\text{配列長} \times \text{総リード数}}$
  - 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
    - Filloux et al., *BMC Bioinformatics*, **15**: 188, 2014

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

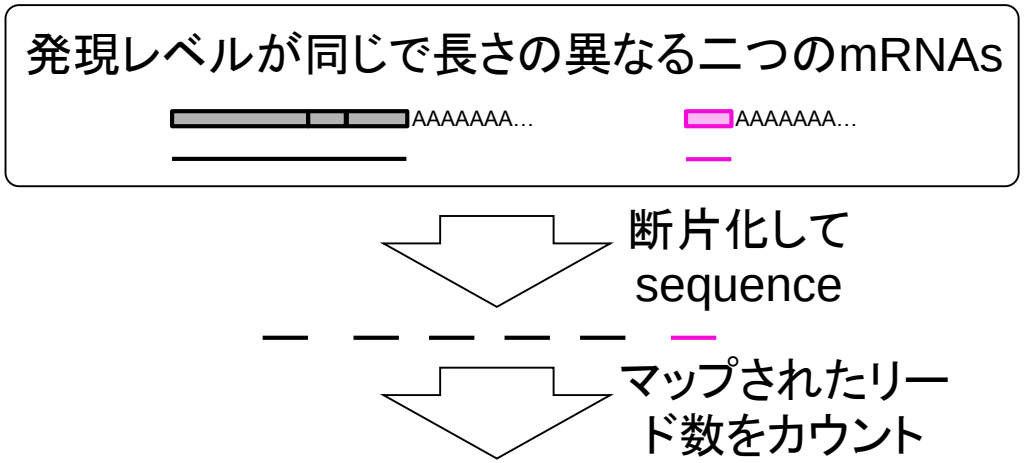
研究目的別留意点：サンプル間比較

- 発現量補正の基本形：カウント数$\times \frac{\text{定数}}{\text{配列長}\times\text{総リード数}}$
  - 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)
    - DEGESに基づく正規化法(Sun et al., *BMC Bioinformatics*, **14**: 219, 2013)

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

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



mRNA	リード数
AAAAAAA…	5
AAAAAAA…	1

1つのサンプル内で異なる遺伝子間の発現レベルの大小関係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

Garber et al., *Nat. Methods*, **8**: 469-477, 2011のFig. 3a

・長さが同じならリード数の多い方が発現量高い（gene 1 対 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 対 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小を比較したい場合には配列長を考慮すべきである

配列長とカウント数の関係を眺める

- 書籍 | トランスクリプトーム解析 | 2.3.4 マッピング(準備) (last modified 2014/06/20) NEW
- 書籍 | トランスクリプトーム解析 | 2.3.5 マッピング(本番) (last modified 2014/06/21) NEW
- 書籍 | トランスクリプトーム解析 | 2.3.6 カウントデータ取得 (last modified 2014/04/21)
- 書籍 | トランスクリプトーム解析 | 3.3.1 解析目的別留意点 (last modified 2014/04/20)
- 書籍 | トランスクリプトーム解析 | 3.3.2 データの正規化(基礎編) (last modified 2014/06/23) NEW

書籍 | トランスクリプトーム解析 | 3.3.1 解析目的別留意点

シリーズ Useful R 第7巻トランスクリプトーム解析のp129-132のRコードです。ここでは
クトップ上のhumanというフォルダ上で作業を行うという前提です。いくつかコメントしてお
17日にR ver. 3.1.0とBioconductor ver. 2.14をインストールしたノートPCで確認した結果で
方についてはRのインストールと起動をごらんください。いくつかの数値の違いは、Rの
るでしょうし、取得先のデータベース側のバージョンにもよると思います。したがって、「
力ファイルのリンク先のもの」と「手持ちの入力ファイル」が多少異なっていても気にする
「ファイル」-「ディレクトリの変更」でデスクトップ上のhumanに移動し以下をコピペ。

p130の網掛け部分:

書籍中では作業ディレクトリがデスクトップ上の"human"という前提になっていますが、
を含むカウントデータファイル(SRA000299_ensgene.txt; 約1.5MB)を置いてあるディレ
構いません。

```
in_f <- "SRA000299_ensgene.txt"
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
### 配列長 vs. リード数の両対数プロット ###
plot(data[,1:2], log="xy", xlab="Length", ylab="Count")
grid(col="gray", lty="dotted")
```

	width	Kidney	Liver
ENSG00000000003	2968	225	228
ENSG00000000005	1610	0	0
ENSG00000000419	1207	100	60
ENSG00000000457	6876	67	124
ENSG00000000460	6354	30	58
ENSG00000000938	3474	43	73
ENSG00000000971	8144	189	4250
ENSG00000001036	3119	331	122
ENSG00000001084	8463	99	460
ENSG00000001167	3811	62	81
ENSG00000001460	8514	63	12
ENSG00000001461	9404	348	60
ENSG00000001497	5589	133	63
ENSG00000001561	4651	213	53
ENSG00000001617	4826	259	18
ENSG00000001626	8099	8	34
ENSG00000001629	7130	256	83

入力ファイル読み込み時にrow.names=1としているので、dataオブジェクトの1列目がwidth列(配列長情報)、2列目がKidney列(腎臓サンプルのカウント情報)となる

配列長とカウント数の関係を眺める

p130の網掛け部分:

書籍中では作業ディレクトリがデスクトップ上の"human"という前提になっていますが、p92で作成した
を含むカウントデータファイル(SRA000299_ensgene.txt; 約1.5MB)を置いてあるディレクトリであればと
構いません。

```
in_f <- "SRA000299_ensgene.txt"
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
### 配列長 vs. リード数の両対数プロット ###
plot(data[,1:2], log="xy", xlab="Length", ylab="Count")
grid(col="gray", lty=
```

	width	Kidney	Liver
ENSG00000000003	2968	225	228
ENSG00000000005	1610	0	0
ENSG00000000419	1207	100	60
ENSG00000000457	6876	67	124
ENSG00000000460	6354	30	58
ENSG00000000938	3474	43	73
ENSG00000000971	9144	180	4250

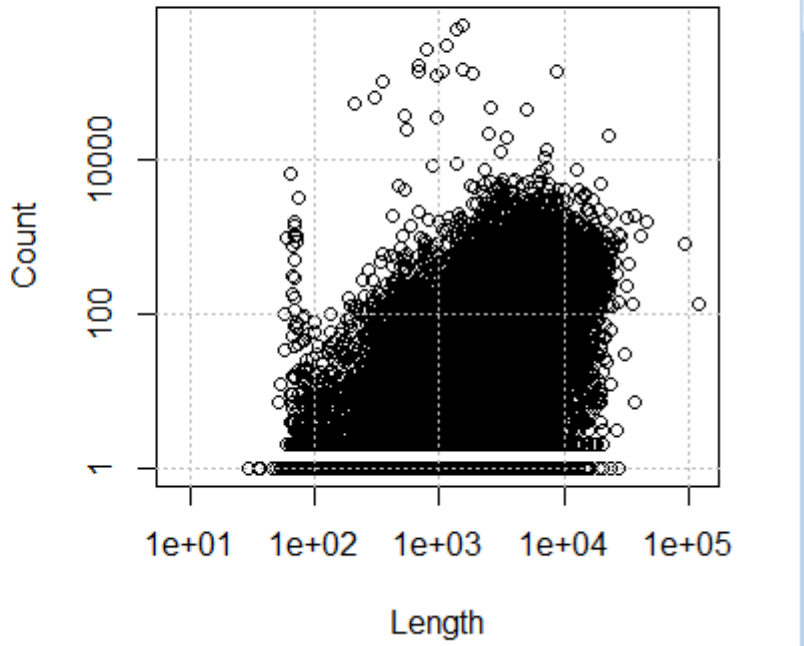
数値のダイナミックレンジが広いのでx軸y軸ともにlog10変換してプロットしている。0カウントのものはlogをとれない関係上、プロットできないという警告が出ています。

R Console

```
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$
 'licence()'は
形式につい$
みられます。
 'q()' と入力すれば R を終了します。

> in_f <- "SRA000299_ensgene.txt"
> data <- read.table(in_f, header=TRUE, row.names=1,
> ### 配列長 vs. リード数の両対数プロット ###
> plot(data[,1:2], log="xy", xlab="Length", ylab="Co
 警告メッセージ:
In xy.coords(x, y, xlabel, ylabel, log) :
  31234 y values <= 0 omitted from logarithmic plot
> grid(col="gray", lty="dotted")
>
```

R Graphics: Device 2 (ACTIVE)

確かに水平ではなく全体的に右斜め上になっている傾向が見られます



配列長とカウント数の関係を眺める

p130の網掛け部分:

書籍中では作業ディレクトリがデスクトップ上の"human"という前提になっていますが、p92で作成した
を含むカウントデータファイル(SRA000299_ensgene.txt; 約1.5MB)を置いてあるディレクトリであれば
構いません。

```
in_f <- "SRA000299_ensgene.txt"
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")
### 配列長 vs. リード数の両対数プロット ###
plot(data[,1:2], log="xy", xlab="Length", ylab="Count")
grid(col="gray", lty="dotted")
```

	width	Kidney	Liver
ENSG00000000003	2968	225	228
ENSG00000000005	1610	0	0
ENSG00000000419	1207	100	60
ENSG00000000457	6876	67	124
ENSG00000000460	6354	30	58
ENSG00000000938	3474	43	73
ENSG00000000971	8144	189	4250
ENSG00000001036	3119	331	122
ENSG00000001084	8463	99	460
			81
			12
			60
			63
			53
			18
			34
			83

ただの検証ですが、ゼロカウントデータが相当数存在することが分かります。

R Console

```
> dim(data)
[1] 60234     3
> obj <- as.logical(data[,2] == 0)
> sum(obj)
[1] 31234
> head(data[obj,])
                 width Kidney Liver
ENSG00000000005  1610      0     0
ENSG00000002079  6387      0     1
ENSG00000002586  3284      0     0
ENSG00000002745  3261      0     2
ENSG00000004809  3791      0     1
ENSG00000004846  9257      0     0
>
```

- 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
- 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 m
- マップ後 | 配列長とカウント数の関係 (last modified 2014/06/24) NEW
- 正規化 | について(last modified 2014/0[illegible]22) NEW

- マップ後 | 配列長とカウント数の関係

マップ後 | 配列長とカウント数の関係 NEW

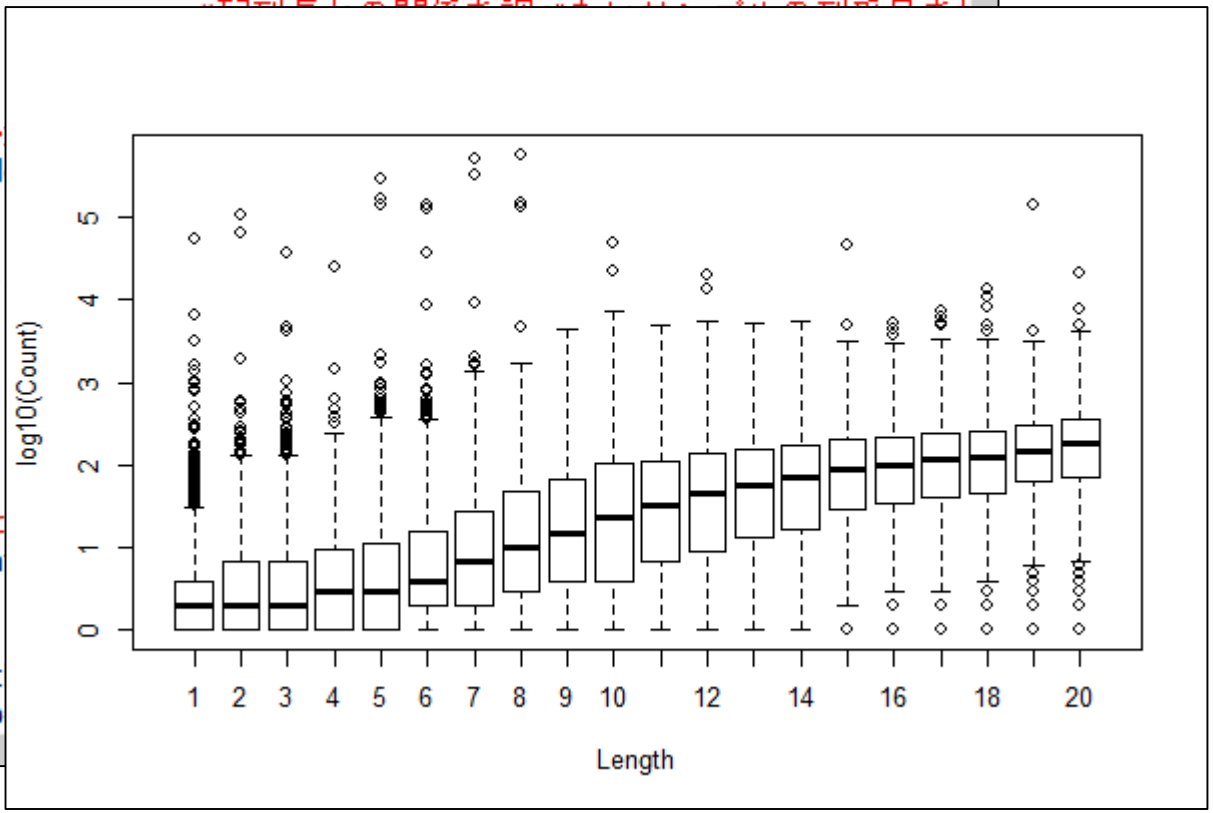
RNA-seqデータは、原理的に配列長が長い転写物ほどその断片配列のリード数が多い傾向にあります

8. 配列長を含むカウントデータファイル(SRA000299_ensgene.txt)の場合:

書籍 | トランスクリプトーム解析 | 2.3.6 カウントデータ取得のp92のコードを実行して得られたファイルです。
(Gene ID列を除く)1列目の配列長と2列目の"Kidney"のカウントデータとの関係を調べています。
横軸:配列長、縦軸:カウント数のboxplot(箱ひげ図)をpng形式ファイルで保存したい場合です。

```
in_f <- "SRA000299_ensgene.txt"          #入力ファイル名を指定してin_fに格納
out_f <- "hoge8.png"                     #出力ファイル名を指定してout_fに格納
param1 <- 20                             #boxplotを描くときの水準数(配列長順でソート
param2 <- c(1, 2)
param_fig <- c(600, 400)

#入力ファイルの読み込みとサブセ
data <- read.table(in_f, head
data <- data[, param2]

#前処理(ゼロカウントデータのフィ
data <- data[data[,2]>0,]

#前処理(配列長の短い順にソート)
data <- data[order(data[,1]),

#前処理(param1で指定した数にdat
f <- gl(param1, floor(nrow(da

#本番(ファイルに保存)
png(out_f, pointsize=13, widt
plot(f, log10(data[,2]), xlab
```

配列長順にソートし、カウント数を20分割したものをboxplotで示したもの。様々な表現手段があります。

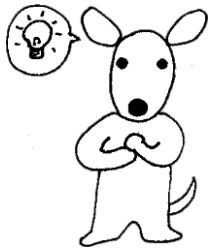


配列長の補正

mRNA	リード数	配列長(in bp)
	5	1500
	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」の計算に相当
  - 「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)」に相当

Reads Per Kilobase (RPK)
Counts Per Kilobase (CPK)

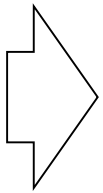


マイクロアレイデータの正規化

- 各サンプルから測定されたシグナル強度の和は一定
  - アレイ上の遺伝子数が少ない場合は非現実的だが、数千～数万種類の遺伝子が搭載されているので妥当という思想

	sample1	sample2
gene1	10.5	12.4
gene2	6.4	7.1
gene3	8.0	8.5
gene4	10.8	11.4
gene5	5.6	6.7
gene6	8.4	8.9
gene7	6.2	7.0
gene8	6.1	6.8
gene9	6.6	6.5
gene10	5.1	5.8
総和	73.7	81.1

グローバル正規化

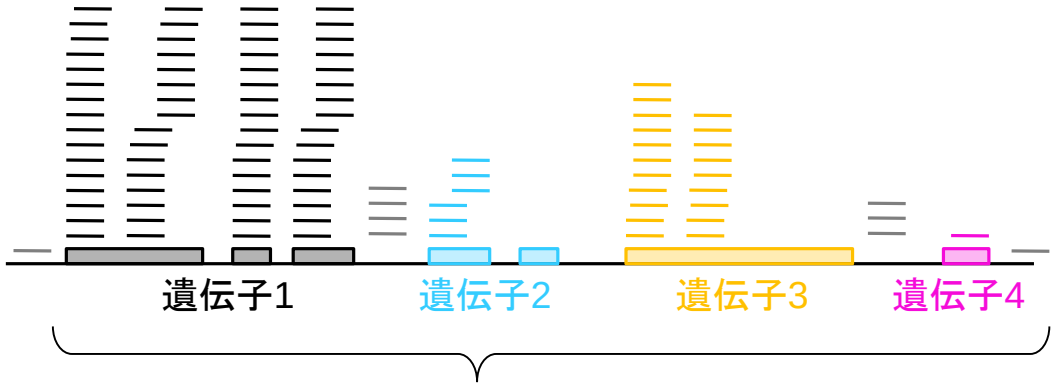


	sample1	sample2
gene1	14.2	15.3
gene2	8.7	8.8
gene3	10.9	10.5
gene4	14.7	14.1
gene5	7.6	8.3
gene6	11.4	11.0
gene7	8.4	8.6
gene8	8.3	8.4
gene9	9.0	8.0
gene10	6.9	7.2
総和	100.0	100.0

背景：サンプルごとにシグナル強度の総和は異なる
対策：総和が任意の値（例では100）になるような正規化係数を掛ける
例：sample1の正規化係数= 100 / 73.7

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1,000 bp、かつ総リード数が100万だったときのカウント数

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

sample_length_count.txt

ID	Length	Count
NM_203348.1	3543	3
NM_001008737.1	1897	19
NM_001037228.1	537	7
NM_033183.2	886	0
NM_138368.3	4443	56
NM_152833.2	2844	85
NM_001100111.1	682	0
NM_001102659.1	1376	0
NM_001104548.1	888	3

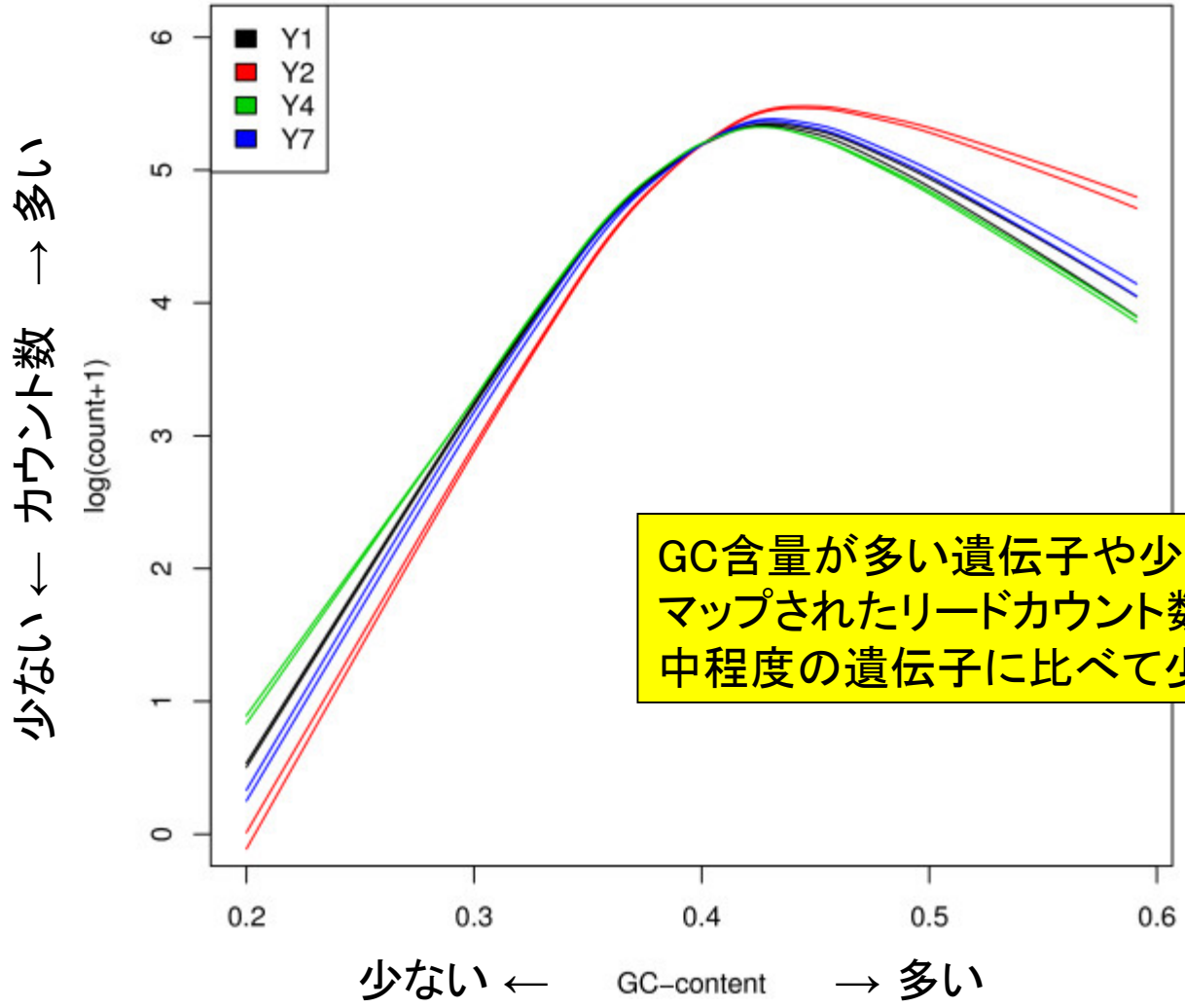
総リード数 = 2385273

hoge1.txt

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

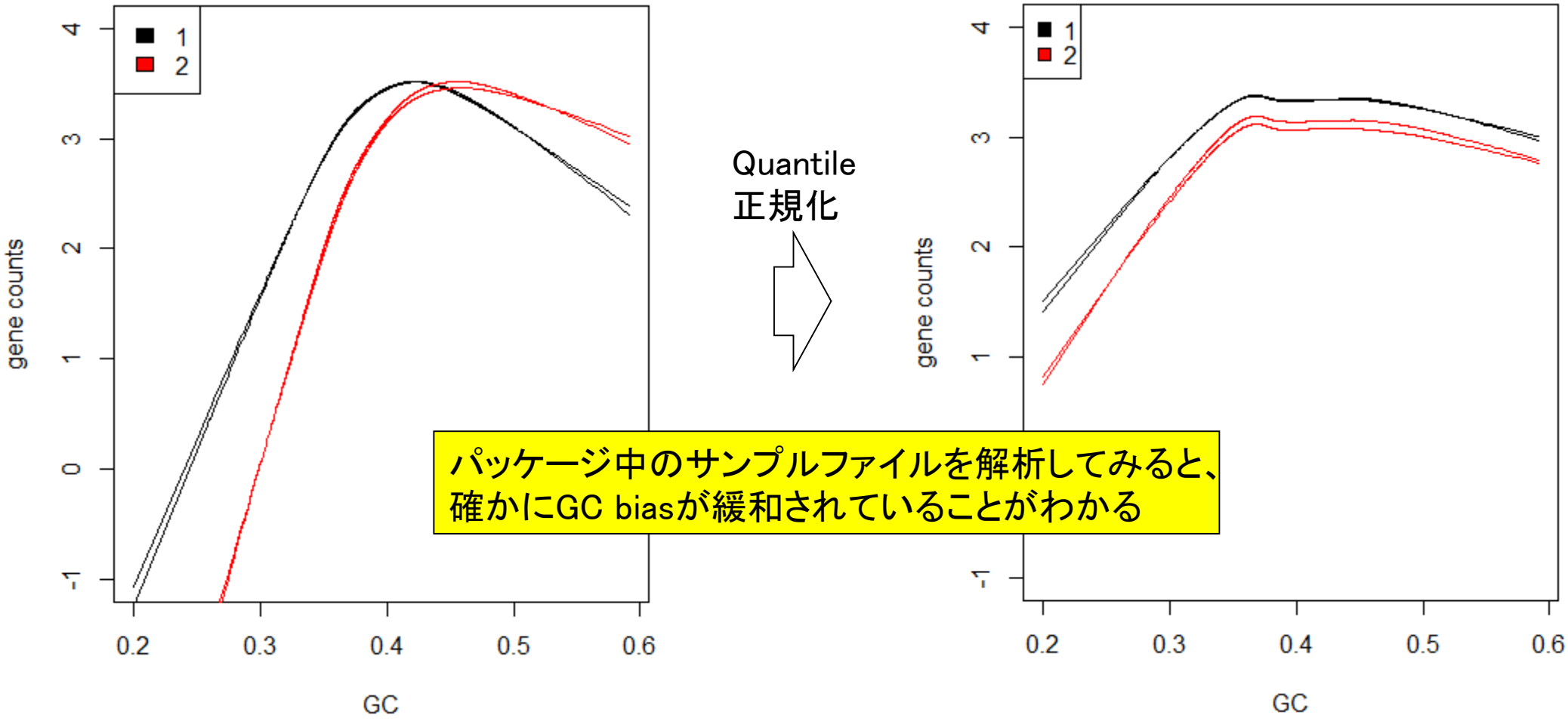
教科書の説明もみながら、RPK, RPM, RPKMの例題を実行しておきましょう

GC bias補正の必要性も提唱されている



GC含量が多い遺伝子や少ない遺伝子上にマップされたリードカウント数は、GC含量が中程度の遺伝子に比べて少ない傾向にある

GC bias補正の必要性も提唱されている



パッケージ中のサンプルファイルを解析してみると、
確かにGC biasが緩和されていることがわかる

Contents（第4回）

- 新規転写物同定（ゲノム情報を利用）
  - 基本的な考え方
  - Tophat-Cufflinksパイプライン
  - 可視化（ゲノムブラウザやViewer）
- 発現量推定（遺伝子レベルと転写物レベル）
  - RPKMの基本的な考え方
  - **計算時間短縮戦略（トランスクリプトーム情報のみを利用）**
- カウントデータを用いたサンプル間比較解析
  - イントロ（カウントデータ取得まで）
  - サンプル間クラスタリング
  - 発現変動遺伝子検出
  - 分布やモデル
  - 課題

高速に発現量推定するための様々な戦略

- ゲノム配列を利用するが、アノテーション情報も同時に読み込んで発現量を得たい特定の領域のみにマッピングして高速化：Cufflinks
- トランスクリプトーム転写物配列にマッピング：NEUMA, IsoEM, RSEM
- k-merを用いたalignment-freeな方法：Salifish, RNA-Skim

解析 | 発現量推定(トランスクリプトーム配列を利用) NEW

新規転写物(新規isoform)の発見などが目的でなく、既知転写物の発現量を知りたいだけの場合には、やたらと時間がかかるゲノム配列へのマッピングを避けるのが一般的です。有名なCufflinksも一応GTF形式のアノテーションファイルを与えることでゲノム全体にマップするのを避けるモードがあるらしいので、一応リストアップしています。転写物へのマッピングの場合には、splice-aware alignerを用いたジャンクションリードのマッピングを行う必要がないので、高速にマッピング可能なbasic alignerで十分です。但し、複数個所にマップされるリードは考慮する必要があり、確率モデルのパラメータを最尤法に基づいて推定するexpectation-maximization (EM)アルゴリズムがよく用いられます。マッピングを行わずに、k-merを用いてalignment-freeで行う発現量推定を行うSailfishやRNA-Skimは従来法に比べて劇的に高速化がなされているようです。2014年6月に調べた結果をリストアップします：

プログラム：

- Cufflinks: Trapnell et al., Nat Biotechnol., 2010
- NEUMA: Lee et al., Nucleic Acids Res., 2011
- IsoEM: Nicolae et al., Algorithms Mol. Biol., 2011
- RSEM: Li and Dewey, BMC Bioinformatics, 2011
- Sailfish: Patro et al., Nat Biotechnol., 2014
- RNA-Skim: Zhang and Wang, Bioinformatics, 2014

トランスクリプトーム配列へのマッピングはbowtieのようなbasic alignerで必要十分。しかしマッピングが律速であるため、alignment-freeな方法が注目されはじめている。

転写物配列にマップして高速に発現量推定

Bioinformatics. 2014 Jun 15;30(12):i283-i292. doi: 10.1093/bioinformatics/btu288.

RNA-Skim: a rapid method for RNA-Seq quantification at transcript level.

Zhang Z, Wang W.

Author information

Abstract

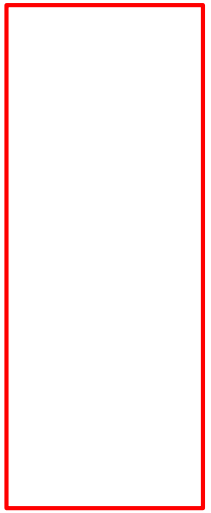
MOTIVATION: RNA-Seq technique has been demonstrated as a revolutionary means for exploring transcriptome because it provides deep coverage and base pair-level resolution. RNA-Seq quantification is proven to be an efficient alternative to Microarray technique in gene expression study, and it is a critical component in RNA-Seq differential expression analysis. Most existing RNA-Seq quantification tools require the alignments of fragments to either a genome or a transcriptome, entailing a time-consuming and intricate alignment step. To improve the performance of RNA-Seq quantification, an alignment-free method, Sailfish, has been recently proposed to quantify transcript abundances using all k-mers in the transcriptome, demonstrating the feasibility of designing an efficient alignment-free method for transcriptome quantification. Even though Sailfish is substantially faster than alternative alignment-dependent methods such as Cufflinks, using all k-mers in the transcriptome quantification impedes the scalability of the method.

RESULTS: We propose a novel RNA-Seq quantification method, RNA-Skim, which partitions the transcriptome into disjoint transcript clusters based on sequence similarity, and introduces the notion of sig-mers, which are a special type of k-mers uniquely associated with each cluster. We demonstrate that the sig-mer counts within a cluster are sufficient for estimating transcript abundances with accuracy comparable with any state-of-the-art method. This enables RNA-Skim to perform transcript quantification on each cluster independently, reducing a complex optimization problem into smaller optimization tasks that can be run in parallel. As a result, RNA-Skim uses <4% of the k-mers and <10% of the CPU time required by Sailfish. It is able to finish transcriptome quantification in <10 min per sample by using just a single thread on a commodity computer, which represents >100 speedup over the state-of-the-art alignment-based methods, while delivering comparable or higher accuracy. Availability and implementation: The software is available at http://www.csbio.unc.edu/rs.

CONTACT: weiwang@cs.ucla.edu Supplementary information: Supplementary data are available at Bioinformatics online.

・Bowtie + eXpressで高精度な結果を追求（~days）
・RNA-Skimで超高速にそこそこの精度で定量化（~min）
1 day = 60*60*24 = 86,400 seconds

Zhang and Wang, *Bioinformatics*, **30**: i283-i292, 2014のTable 3



Contents（第4回）

- 新規転写物同定（ゲノム情報を利用）
  - 基本的な考え方
  - Tophat-Cufflinksパイプライン
  - 可視化（ゲノムブラウザやViewer）
- 発現量推定（遺伝子レベルと転写物レベル）
  - RPKMの基本的な考え方
  - 計算時間短縮戦略（トランスクリプトーム情報のみを利用）
- **カウントデータを用いたサンプル間比較解析**
  - **イントロ（カウントデータ取得まで）**
  - サンプル間クラスタリング
  - 発現変動遺伝子検出
  - 分布やモデル
  - 課題

カウントデータを用いたサンプル間比較解析

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

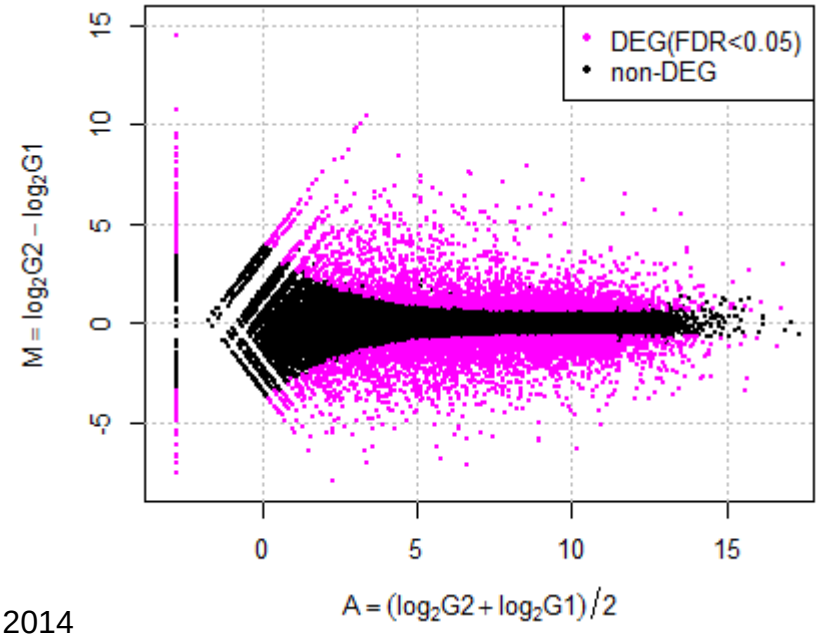
カウントデータ

59,857 genes

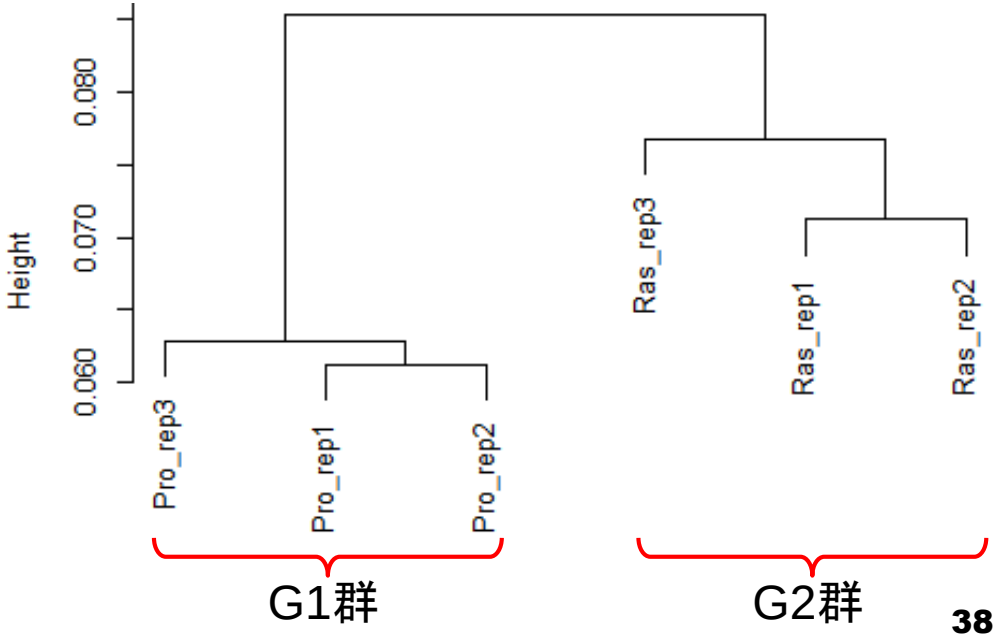
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

データ解析の基本イメージ

発現変動遺伝子(DEG)同定



サンプル間クラスタリング



イントロ（カウントデータ取得まで）

- 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
    - 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	

1つの論文中でChIP-seqもやっており、RNA-seqデータのみダウンロードする際にちょっと困る例を紹介。

イントロ（カウントデータ取得まで）

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

- パイプライン | 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) (last modified 2014/03/27)
- パイプライン | ゲノム | 機能解析 | 2群間 | 対応なし | 複製あり | SRP017142(Neyret-Kahn_2013) (last modified 2014/04/01)
- リンク集

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)

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54 bpと比較的長いリードであり、原著論文中でもsplice-aware alignerの一つであるTopHat (Trapnell et al., Bioinformatics, 200

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

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

ます。

もちろん主観ですが、ENA (ArrayExpress)よりもGEOのほうがわかりやすいという特殊事例です。

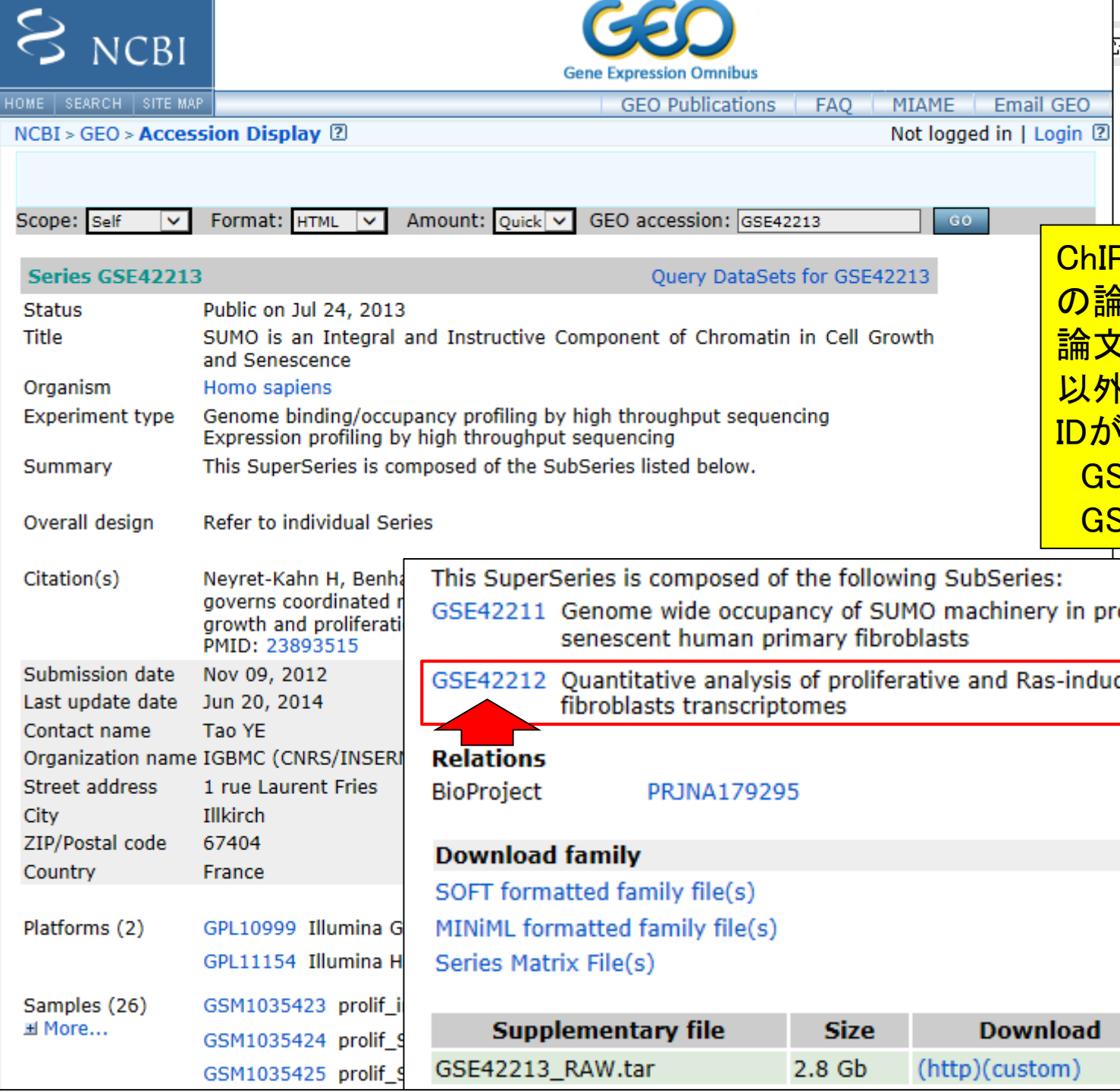
Step1. RNA-seqデータ 圧縮済みのFASTQファイルをダウンロード:

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
```



なし | 複製あり | SRP017142(Neyret-Kahn 2013)

NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Not logged in | Login

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE42213 | GO

Series GSE42213 Query DataSets for GSE42213

Status	Public on Jul 24, 2013
Title	SUMO is an Integral and Instructive Component of Chromatin in Cell Growth and Senescence
Organism	Homo sapiens
Experiment type	Genome binding/occupancy profiling by high throughput sequencing Expression profiling by high throughput sequencing
Summary	This SuperSeries is composed of the SubSeries listed below.
Overall design	Refer to individual Series
Citation(s)	Neyret-Kahn H, Benha governs coordinated r growth and proliferati PMID: 23893515
Submission date	Nov 09, 2012
Last update date	Jun 20, 2014
Contact name	Tao YE
Organization name	IGBMC (CNRS/INSERM
Street address	1 rue Laurent Fries
City	Illkirch
ZIP/Postal code	67404
Country	France
Platforms (2)	GPL10999 Illumina G GPL11154 Illumina H
Samples (26)	GSM1035423 prolif_i GSM1035424 prolif_S GSM1035425 prolif_S

More...

This SuperSeries is composed of the following SubSeries:

GSE42211 Genome wide occupancy of SUMO machinery in proliferative and Ras-induced senescent human primary fibroblasts

GSE42212 Quantitative analysis of proliferative and Ras-induced senescent human primary fibroblasts transcriptomes

Relations

BioProject PRJNA179295

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE42213_RAW.tar	2.8 Gb	(http)(custom)	TAR (of BED, WIG)

ChIP-seqとRNA-seq両方を1つの論文中でやっている場合には、論文と「1対1」対応のGSE42213以外に、さらに下の階層のGSE IDが付与されている。

GSE42211：ChIP-seqデータ

GSE42212：RNA-seqデータ

ENA (ArrayExpress)の場合は…

ArrayExpressで眺めると、サブシリーズのGSE ID (GSE42211とGSE42212)が見当たらない。

ENA (ArrayExpress)の場合は…

ArrayExpress > Experiments > E-GEOD-42213

E-GEOD-42213 - SUMO is an Integral and Instructive Component of Chromatin in Cell Growth and Senescence

Status	*Released on 24 July 2013, last updated on 2 June 2014*
Organism	Homo sapiens
Samples (26)	Click for detailed sample information and links to data
Protocols (4)	Click for detailed protocol information

ChIP-seqデータとRNA-seqデータ(GSE42211とGSE42212)をサブシリーズに分割せずに一覧可能にしたのはいいと思うが、なぜ26サンプルが34になっているのか不明。

ArrayExpress > Experiments > E-GEOD-42213 > Samples and Data*

E-GEOD-42213 - SUMO is an Integral and Instructive Component of Chromatin in Cell Growth and Senescer

Page 1 2 Showing 1 - 25 of 34 rows Page size 25 50

		Sample Characteristics				
Source Name	Sample_source_name	cell line	cell type	chip antibodies	genotype	organism
GSM1035423 1	prolif_input_DNA	WI38	primary fibroblasts	none (input)	infected with pBABE-puro-empty	Homo sapi
GSM1035424 1	prolif_SUMO1_ChIPSeq	WI38	primary fibroblasts	SUMO1 (non-commercial)	infected with pBABE-puro-empty	Homo sapi
GSM1035424 1	prolif_SUMO1_ChIPSeq	WI38	primary fibroblasts	SUMO1 (non-commercial)	infected with pBABE-puro-empty	Homo sapi
GSM1035425 1	prolif_SUMO1_ChIPSeq	WI38	primary fibroblasts	SUMO1 (non-commercial)	infected with pBABE-puro-empty	Homo sapi
GSM1035426 1	prolif_SUMO2_ChIPSeq	WI38	primary fibroblasts	SUMO2 (non-commercial)	infected with pBABE-puro-empty	Homo sapi
GSM1035426 1	prolif_SUMO2_ChIPSeq	WI38	primary fibroblasts	SUMO2 (non-commercial)	infected with pBABE-puro-empty	Homo sapi
GSM1035427 1	prolif_Ubc9_ChIPSeq	WI38	primary fibroblasts	(BD transduction)	infected with pBABE-puro-empty	Homo sapi
GSM1035427 1	prolif_Ubc9_ChIPSeq	WI38	primary fibroblasts	(BD transduction)	infected with pBABE-puro-empty	Homo sapi
GSM1035428 1	prolif_PolII_ChIPSeq	WI38	primary fibroblasts	anti-PolII (sc-9001)	infected with pBABE-puro-empty	Homo sapi

イントロ（カウントデータ取得まで）

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

#前処理
#sqlfile <- "SRAmetadb.sqlite"              #最新でなくてもよく、手元
sqlfile <- getSRAdbFile()                   #最新のSRAmetadb SQLite
sra_con <- dbConnect(SQLite(), sqlfile)#おまじない

#前処理(実験デザインの全体像を表示)
hoge <- sraConvert(param, sra_con=sra_con)#paramで指定したSRA I
hoge                                        #hogeの中身を表示
apply(hoge, 2, unique)                      #hoge行列の列ごとにユニー

#前処理(FASTQファイルサイズを表示)
k <- getFASTQinfo(in_acc=hoge$run)          #「hoge$run」で指定したS
k                                           #kの中身を表示
hoge2 <- cbind(k$library.name,              #ライブラリ名と、
               k$run.read.count,            #総リード数と、
               k$file.name,                 #ファイル名と、
               k$file.size)                 #ファイルサイズ、の順番で列方向で結合した結果をhoge2に格納
```

計6ファイル（2群間比較用）

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	

無事ダウンロードが終了すると、作業ディレクトリ（「デスクトップ」上の「SRP017142」フォルダ）中に7つのファイルが存在するはずです。4Gb程度ある"SRAmetadb.sqlite"ファイルは無視して構いません。残りの"SRR"からはじまる6つのファイルがダウンロードしたRNA-seqデータです。オリジナルのサンプル名(の略称)で対応関係を表すとsrp017142_samplename.txtのようになっていることがわかります。尚このファイルはマッピング時の入力ファイルとしても用います。

イントロ（カウントデータ取得まで）

- Step2: QuasRを用いたヒトゲノムへのマッピング
  - リファレンス配列としてBSgenome.Hsapiens.UCSC.hg19というRパッケージを利用

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"にマップ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にマップした結果であり、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")を指定しているので、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。マップ後 | カウント情報取得 | ゲノム | アノテーション有 | QuasR(Lerch_XXX)の記述内容と基本的に同じです。

```
in_f1 <- "srp017142_samplename.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 <- "BSgenome.Hsapiens.UCSC.hg19" #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 <- "srp017142_QC_bowtie.pdf"     #出力ファイル名を指定してout_f1に格納
out_f2 <- "srp017142_count_bowtie.txt" #出力ファイル名を指定してout_f2に格納
out_f3 <- "srp017142_genelength.txt"    #出力ファイル名を指定してout_f3に格納
out_f4 <- "srp017142_RPKM_bowtie.txt"  #出力ファイル名を指定してout_f4に格納
out_f5 <- "srp017142_transcript_seq.fa"#出力ファイル名を指定してout_f5に格納
out_f6 <- "srp017142_other_info1.txt"   #出力ファイル名を指定してout_f6に格納
param1 <- "hg19"                        #TranscriptDbオブジェクト作成用のリファレンスゲノムを指定(「ucscGenomes(
param2 <- "ensGene"                     #TranscriptDbオブジェクト作成用のtable名を指定(「supportedUCSCtables()
param3 <- "gene"                        #カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "junct

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

約18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：BSgenome.Athaliana.TAIR.TAIR9
ショウジョウバエ：BSgenome.Dmelanogaster.UCSC.dm3

ゲノム配列のRパッケージがあります

- イントロ | 一般 | Tips | 拡張子は同じで任意の文字を追加して保存(last modified 2013/09/26)
- イントロ | 一般 | 配列取得 | ゲノム配列 | 公共DBから(last modified 2014/05/28)
- イントロ | 一般 | 配列取得 | ゲノム配列 | BSgenome(last modified 2014/04/22)
- イントロ | 一般 | 配列取得 | プロモーター配列 | 公共DBから(last modified 2014/04/02)

イントロ | 一般 | 配列取得 | ゲノム配列 | BSgenome NEW

BSgenomeパッケージを用いて様々な生物種のゲノム配列を取得す
タザオ (A. lyrata)、セイヨウミツバチ (A. mellifera)、シロイヌナズナ
線虫(C.elegans)、犬(C.familiaris)、キイロショウジョウバエ(D.mela
(D.rerio)、大腸菌(E.coli)、イトヨ(G.aculeatus)、セキショクヤケイ(
ゲザル(M.mulatta)、マウス(M.musculus)、チンパンジー(P.troglod
芽酵母(S.cerevisiae)、トキソプラズマ(T.gondii)と実に様々な生物
ります。getSeq関数はBSgenomeオブジェクト 中の「single sequences
ているchr...というものを全て抽出しています。したがって、例えばマ
「chr1_random」や「chrUn_random」なども等価に取扱っている点に
「ファイル」-「ディレクトリの変更」でファイルを保存したいディレクト

1. 利用可能な生物種とRにインストール済みの生物種をリストアッ

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

#本番（利用可能なリストアップ；インストール済みとは限
available.genomes()                  #このパッケ
```

RおよびBioconductorの最新版をインストールしたヒトが、mm10などゲノム配列の最新版も利用できます。定期的なバージョンアップの意義。

R Console

```
[48] "BSgenome.Mmusculus.UCSC.mm10"
[49] "BSgenome.Mmusculus.UCSC.mm10.masked"
[50] "BSgenome.Mmusculus.UCSC.mm8"
[51] "BSgenome.Mmusculus.UCSC.mm8.masked"
[52] "BSgenome.Mmusculus.UCSC.mm9"
[53] "BSgenome.Mmusculus.UCSC.mm9.masked"
[54] "BSgenome.Osativa.MSU.MSU7"
[55] "BSgenome.Ptroglodytes.UCSC.panTro2"
[56] "BSgenome.Ptroglodytes.UCSC.panTro2.masked"
[57] "BSgenome.Ptroglodytes.UCSC.panTro3"
[58] "BSgenome.Ptroglodytes.UCSC.panTro3.masked"
[59] "BSgenome.Rnorvegicus.UCSC.rn4"
[60] "BSgenome.Rnorvegicus.UCSC.rn4.masked"
[61] "BSgenome.Rnorvegicus.UCSC.rn5"
[62] "BSgenome.Rnorvegicus.UCSC.rn5.masked"
[63] "BSgenome.Scerevisiae.UCSC.sacCer1"
[64] "BSgenome.Scerevisiae.UCSC.sacCer2"
[65] "BSgenome.Scerevisiae.UCSC.sacCer3"
[66] "BSgenome.Sscrofa.UCSC.susScr3"
[67] "BSgenome.Sscrofa.UCSC.susScr3.masked"
[68] "BSgenome.Tgondii.ToxoDB.7.0"
[69] "BSgenome.Tguttata.UCSC.taeGut1"
[70] "BSgenome.Tguttata.UCSC.taeGut1.masked"
> |
```

Contents（第4回）

- 新規転写物同定（ゲノム情報を利用）
  - 基本的な考え方
  - Tophat-Cufflinksパイプライン
  - 可視化（ゲノムブラウザやViewer）
- 発現量推定（遺伝子レベルと転写物レベル）
  - RPKMの基本的な考え方
  - 計算時間短縮戦略（トランスクリプトーム情報のみを利用）
- カウントデータを用いたサンプル間比較解析
  - イントロ（カウントデータ取得まで）
  - **サンプル間クラスタリング**
  - 発現変動遺伝子検出
  - 分布やモデル
  - 課題

サンプル間クラスタリング

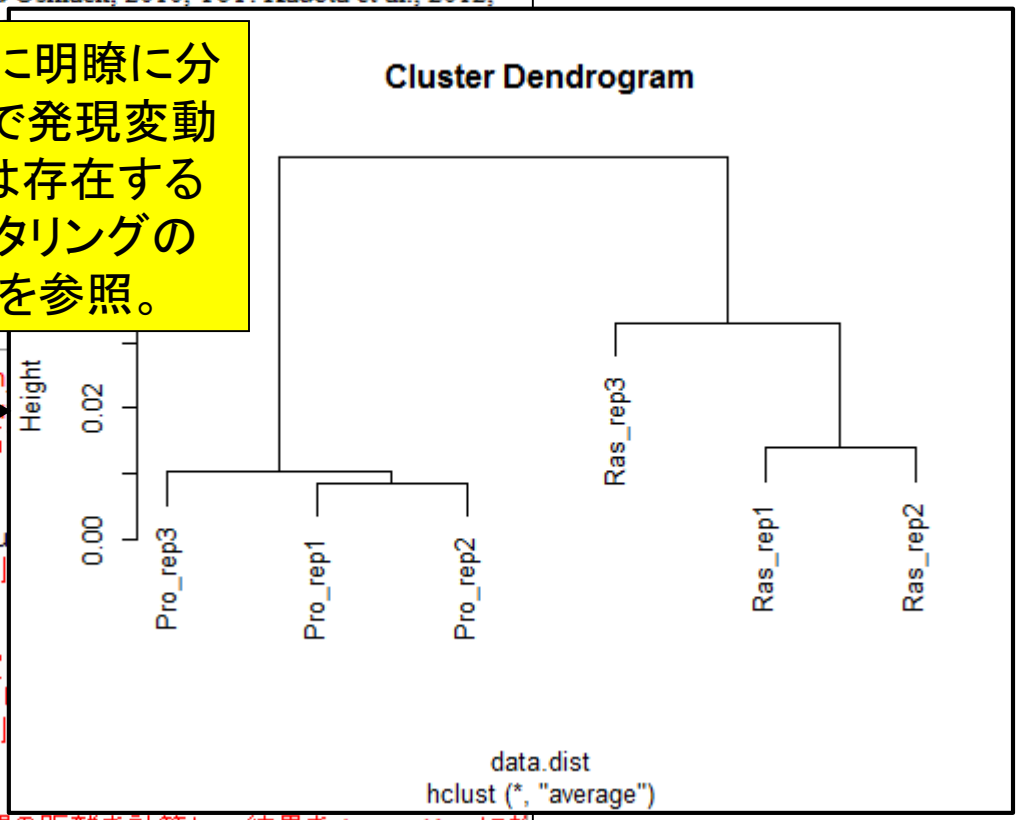
Step3. サンプル間クラスタリング：

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
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(S
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。Pe
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いいだろうとい
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
も関わらずRPKMデータを用いると、RPKM補正後の値が1未満のものがかなり存在
ウントデータの処理が必要ですがやりかた次第で結果がころころかわりうるという状況が嫌なので、RNA-seqデータの場合に
は私はSpearman相関係数にしています。また、ベクトルの要素間の差を基本とするdistance metrics (例:ユークリッド距離やマ
ンハッタン距離など)は、比較的最近のRNA-seqデータ正規化法 (TMM: Robinson and Oshlack, 2010, TbT: Kadota et al., 2012,
TCC; Sun et al., 2013)論文の重要性が理解できれば、その
ンプルごとに転写物の組成比が異なるため、RPMやCPMの
数値の差」に基づいて距離を定めるのはいかがなものか？
グを行った結果と比較することで、転写物の組成比に関する
が低いものを予めフィルタリングしておく必要もあるのだろう
ちのデータにこの類似度を適用したときの理論上の短所をき
す。ここではカウントデータでクラスタリングをしていますが、
(srp017142_RPKM_bowtie.txt) でも得られる樹形図のトポロ
います。配列長補正の有無で、サンプル間の相関係数の値
復実験間でそれほど違わないので、多少順位に変動があっ

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

Pro群とRas群に明瞭に分かれているので発現変動遺伝子(DEG)は存在すると判断。フィルタリングの思想は教科書を参照。

```
in_f3 <- "srp017142_count_bowtie.txt"   #入力ファイル名を指定してin
out_f6 <- "srp017142_count_cluster.png" #出力ファイル名を指定してo
param_fig <- c(500, 400)                #ファイル出力時の横幅と縦幅

#入力ファイルの読み込み
data <- read.table(in_f3, header=TRUE, row.names=1, sep="\t", qu
dim(data)                               #オブジェクトdataの行数と列

#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
data <- unique(data[obj,])              #objがTRUEとなる行のみ抽出
dim(data)                               #オブジェクトdataの行数と列

#クラスタリングおよび結果の保存
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
```



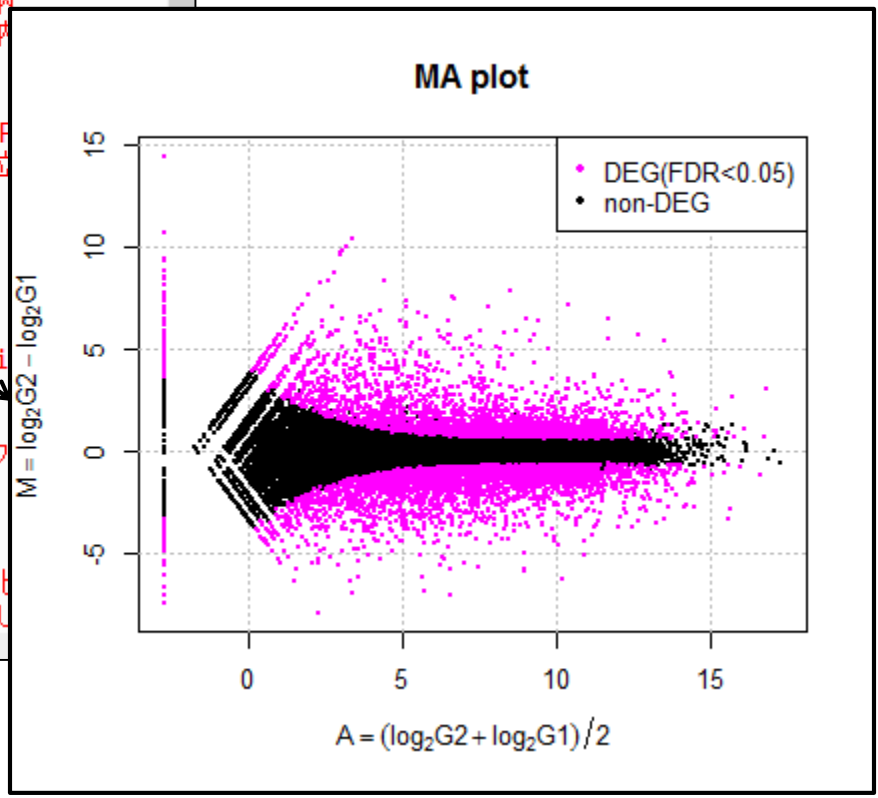
発現変動遺伝子検出

Step4. 発現変動遺伝子(DEG)同定:

カウントデータファイル(srp017142_count_bowtie.txt)を入力として2群間で発現の異な
す。このデータはbiological replicatesありのデータなので、TCCパッケージ(Sun et al.,
に従って、iDEGES/edgeR正規化(Sun et al., 2013; Robinson et al., 2010; Robinson an
and Smyth, 2008)を行ったのち、edgeRパッケージ中のan exact test (Robinson and Sm
検出を行っています。解析 | 発現変動 | 2群間 | 対応なし | 複製あり | iDEGES/edgeR-
正規化 | サンプル間 | 2群間 | 複製あり | iDEGES/edgeR(Sun_2013)の記述内容と基本

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

```
in_f4 <- "srp017142_count_bowtie.txt"   #入力ファイル名を指定してin_f4に格納
out_f7 <- "srp017142_DEG_bowtie.txt"    #出力ファイル名を指定してout_f7に格納
out_f8 <- "srp017142_MAplot_bowtie.png" #出力ファイル名を指定してout_f8に格納
out_f9 <- "srp017142_other_info2.txt"   #出力ファイル名を指定してout_f9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 (F
param_fig <- c(430, 390)                #MA-plot描画時の横幅と縦幅を指定(単位

#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_f4, header=TRUE, row.names=1, sep="\t", quote="")#i

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

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



発現変動遺伝子(DEG)と判定されたものが多数存在することがわかる

発現変動遺伝子検出

Step4. 発現変動遺伝子(DEG)同定:

カウントデータファイル(srp017142_count_bowtie.txt)を入力として2
す。このデータはbiological replicatesありのデータなので、TCCパ
に従って、iDEGES/edgeR正規化(Sun et al., 2013; Robinson et al.,
and Smyth, 2008)を行ったのち、edgeRパッケージ中のan exact tes
検出を行っています。解析 | 発現変動 | 2群間 | 対応なし | 複製あ
正規化 | サンプル間 | 2群間 | 複製あり | iDEGES/edgeR(Sun_2013

```
in_f4 <- "srp017142_count_bowtie.txt"   #入力ファイ
out_f7 <- "srp017142_DEG_bowtie.txt"    #出力ファイ
out_f8 <- "srp017142_MAplot_bowtie.png"#出力ファイ
out_f9 <- "srp017142_other_info2.txt"   #出力ファイ
param_G1 <- 3                           #G1群のサン
param_G2 <- 3                           #G2群のサン
param_FDR <- 0.05                       #DEG検出時
param_fig <- c(430, 390)                #MA-plot描

#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_f4, header=TRUE, row.names=

#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)        #TCCクラス

#本番(正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", te
                       iteration=3, FDR=0.1, floo
```

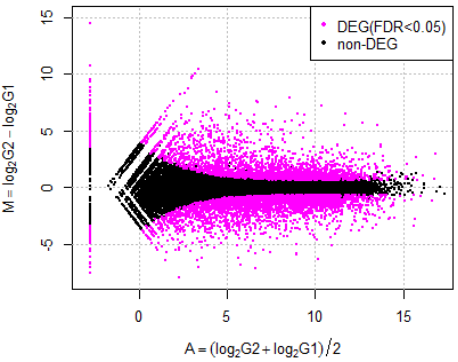
```
1. Numbers of DEGs satisfying several FDR thresholds.
FDR < 0.05:[1] 5669
FDR < 0.10:[1] 6679
FDR < 0.20:[1] 8110
FDR < 0.30:[1] 9151


2. Session info.
R version 3.1.0 (2014-04-10)
Platform: x86_64-w64-mingw32/x64 (6

locale:
[1] LC_COLLATE=Japanese_Japan.932
[3] LC_MONETARY=Japanese_Japan.932 LC_NUMERIC=C
[5] LC_TIME=Japanese_Japan.932

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] TCC_1.4.0               ROC_1.40.0              baySeq_1.18.0
 [4] edgeR_3.6.0             limma_3.20.1            DESeq2_1.4.0
 [7] RcppArmadillo_0.4.200.0 Rcpp_0.11.1             GenomicRanges_1.16.1
[10] GenomeInfoDb_1.0.2      IRanges_1.22.3          DESeq_1.16.0
[13] lattice_0.20-29         locfit_1.5-9.1          Biobase_2.24.0
[16] BiocGenerics_0.10.0

loaded via a namespace (and not attached):
 [1] annotate_1.42.0      AnnotationDbi_1.26.0 DBI_0.2-
 [4] genefilter_1.46.0    geneplotter_1.42.0   grid_3.1
 [7] RColorBrewer_1.0-5   RSQLite_0.11.4       samr_2.0
[10] splines_3.1.0        stats4_3.1.0         survival_2.37-7
[13] XML_3.98-1.1         xtable_1.7-3         XVector_0.4.0
```

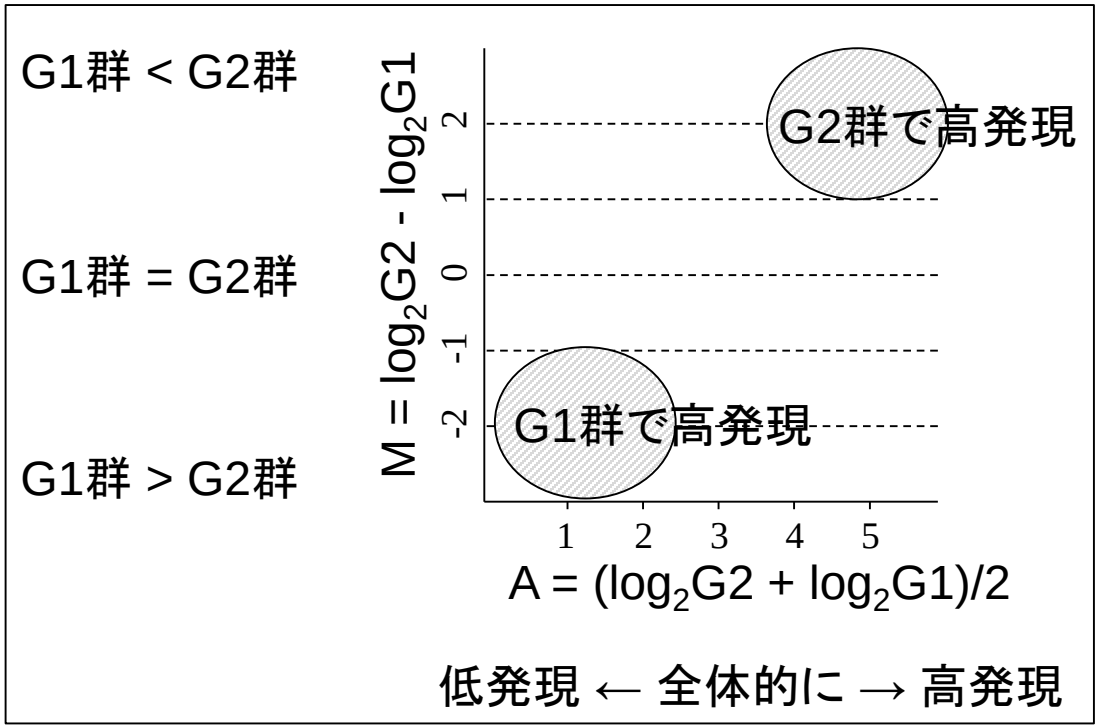
5%偽物を含むのを許容するとDEG数は5,669個。20%の偽物混入を許容すると8,110 DEGs。FDR閾値が30%の場合は9,151個。このデータセット中に存在する本物のDEGは9,151×0.7 = 6,405.7個程度だと判断できる。

論文に記載すべきデータ解析環境の情報

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）





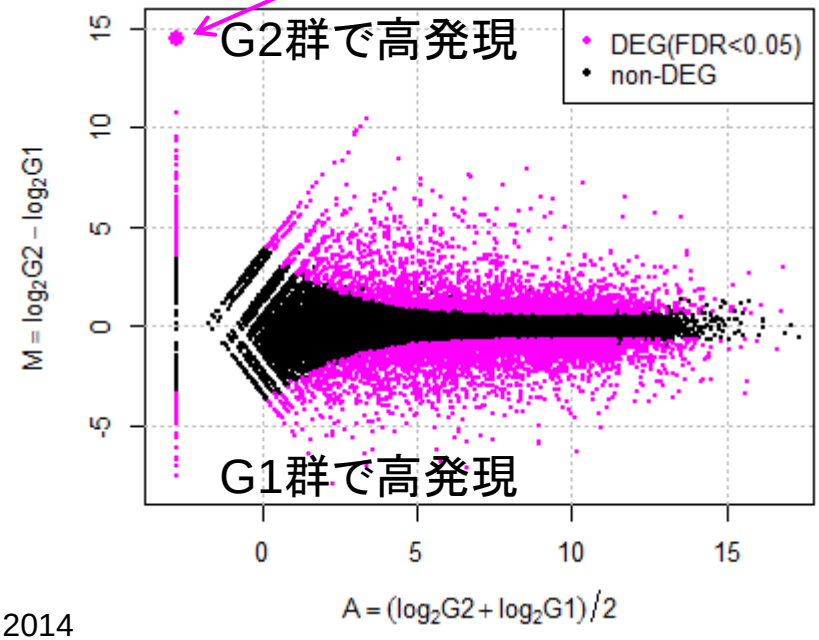
DEGが存在しないデータのM−A plotを眺めることで、縦軸の閾値のみに相当する**倍率変化**を用いたDEG同定の危険性が分かります

発現変動遺伝子検出結果

- *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。

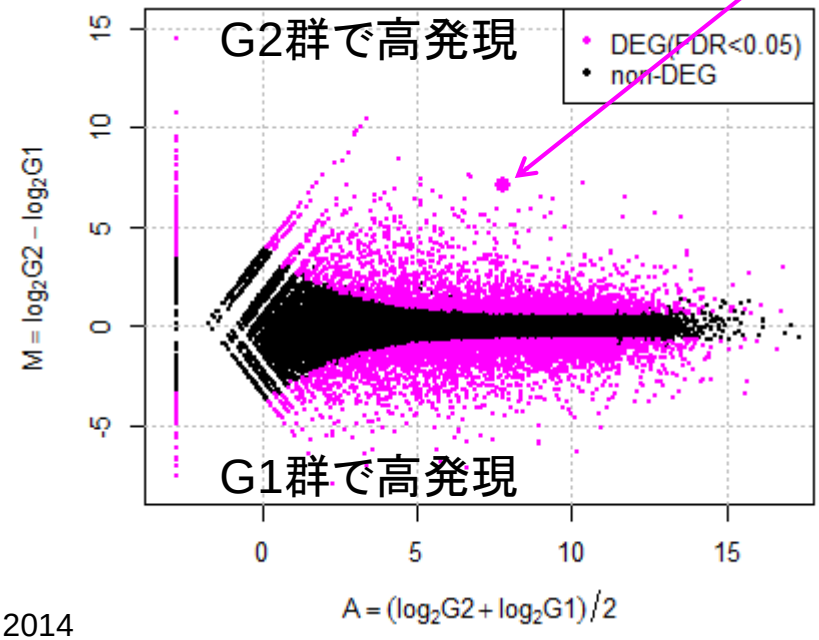
基本的には、これらが解析結果です
1位はRas群（G2群）で高発現のDEG

発現変動遺伝子検出結果

- *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。

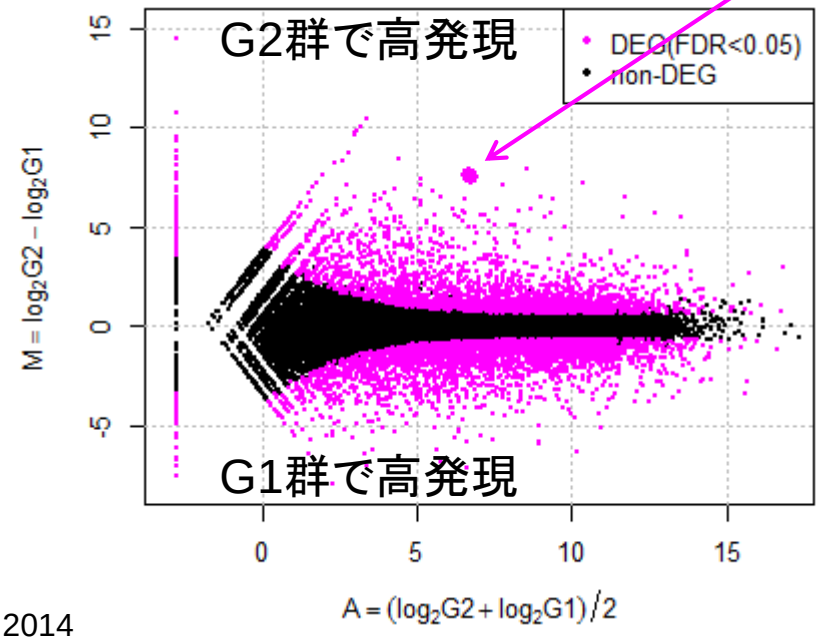
2位もRas群（G2群）で高発現のDEG

発現変動遺伝子検出結果

- *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。

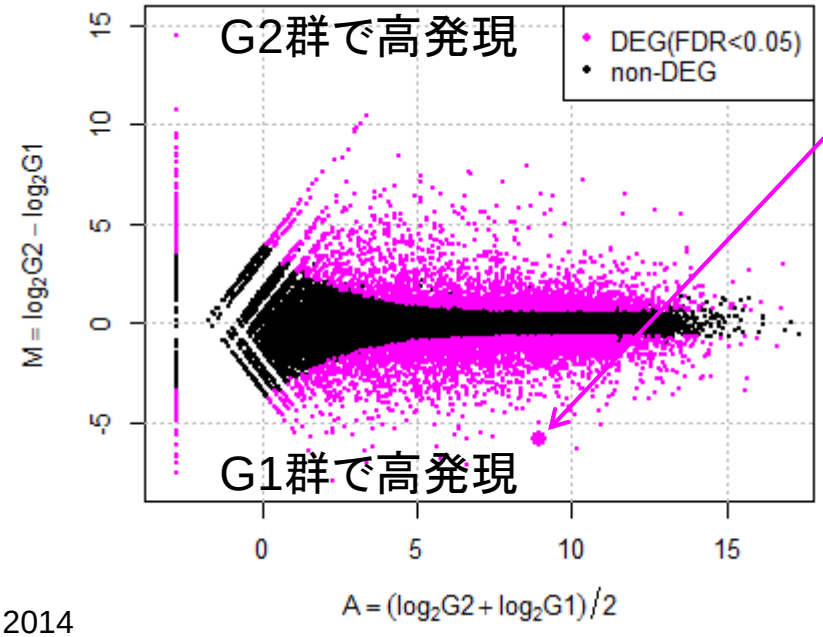
3,4位もRas群（G2群）で高発現のDEG

発現変動遺伝子検出結果

- *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。

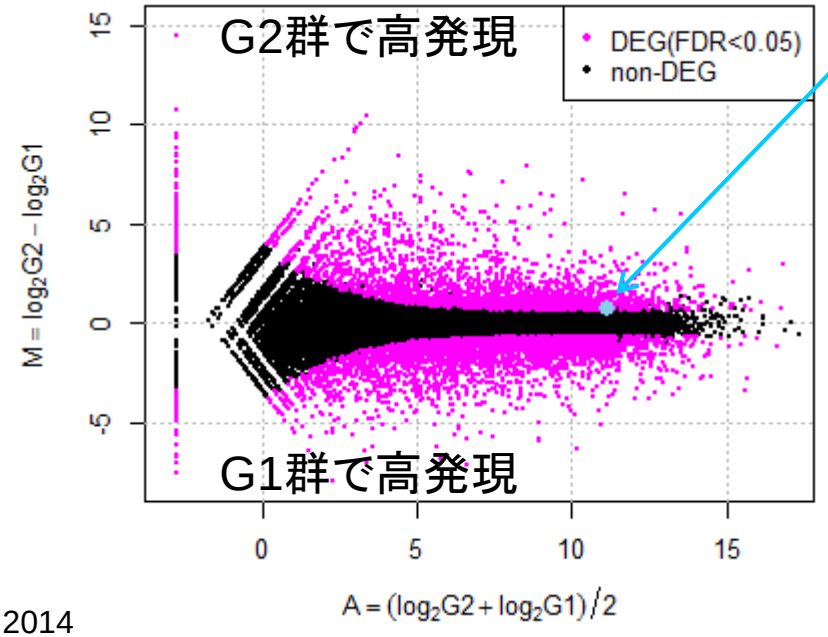
5位はPro群（G1群）で高発現のDEG

発現変動遺伝子検出結果

- *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
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。

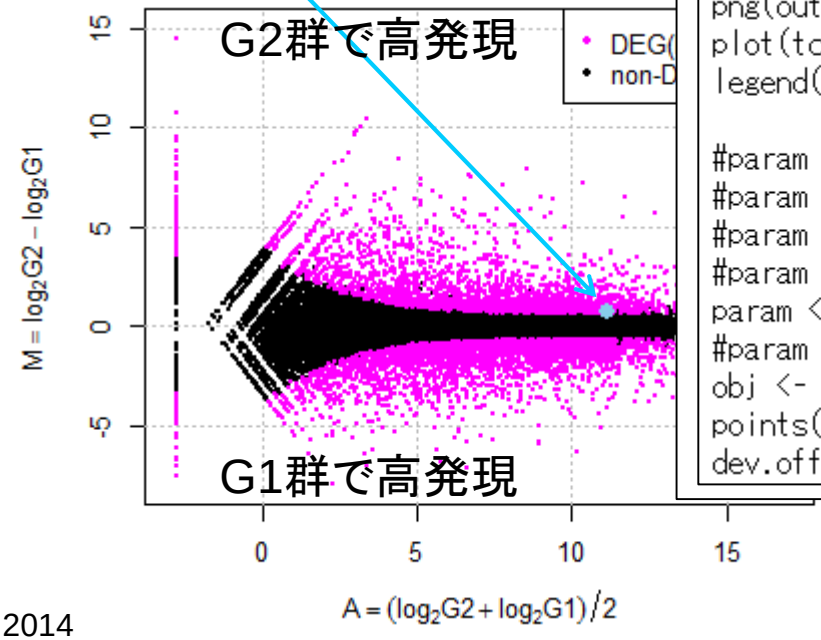
指定したFDR閾値（0.05）をギリギリ満たす5,669位の遺伝子

発現変動遺伝子検出結果

- *TCC*を用いたDEG同定結果ファイル

ハイライトさせたいGene IDの位置情報を論理値ベクトルobjとして取得後、points関数を用いてobjがTRUEとなる要素のみ、pch, cex, colオプションを駆使して追加で描画している。

rownames(tcc$coun	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3	gene_id						
ENSG00000148848	286.8	327.2	262.0	486.4	475.5	419.5	ENSG00000148848						
ENSG00000186603	16.3	13.2	10.1	23.8	16.7	69.3	ENSG00000186603						
ENSG00000168556	161.8	142.9	146.1	218.7	257.7	236.6	ENSG00000168556						
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									
ENSG00000105778	900.5	1027.8	984.6	1529.2									
ENSG00000246451	46.2	91.7	73.6	46.3									



rcode_SRP017142_highlight.txt(の一部)

```
#ファイルに保存(M-A plot)↓
png(out_f8, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイル
plot(tcc, FDR=param_FDR, xlim=c(-3, 17), ylim=c(-8, 15))#param_FDRで指定した閾
legend("topright", c(paste("DEG(FDR<", param_FDR, ")", sep=""), "non-DEG"),#凡
       col=c("magenta", "black"), pch=20)#凡例を作成している↓
#param <- "ENSG00000240386"                          # 1位↓
#param <- "ENSG00000128564"                          # 2位↓
#param <- c("ENSG00000188064", "ENSG00000101188") # 3,4位↓
#param <- "ENSG00000163431"                          # 5位↓
param <- "ENSG00000189159"                          # 5669位 skyblue↓
#param <- c("ENSG00000166359", "ENSG00000146676") # 24461, 24462位 skyblue↓
obj <- is.element(tcc$gene_id, param)↓
points(result$a.value[obj], result$m.value[obj], pch=20, cex=2, col="skyblue")
dev.off()                                           #おまじない↓
```

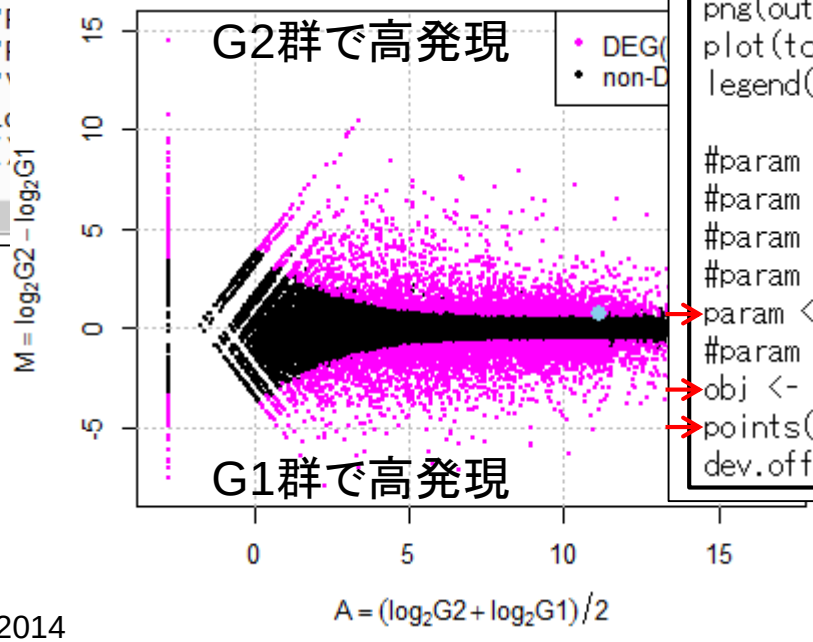
Step4. 発現変動遺伝子(DEG)同定:

カウントデータファイル(srp017142_count_bowtie.txt)を入力として2群間で発現の異なる遺伝子の検出を行います。このデータはbiological replicatesありのデータなので、TCCパッケージ(Sun et al., 2013)の推奨ガイドラインに従って、iDEGES/edgeR正規化(Sun et al., 2013; Robinson et al., 2010; Robinson and Oshlack, 2010; Robinson and Smyth, 2008)を行ったのち、edgeRパッケージ中のan exact test (Robinson and Smyth, 2008)を行って、DEG検出を行っています。解析 | 発現変動 | 2群間 | 対応なし | 複製あり | iDEGES/edgeR-edgeR(Sun_2013)および正規化 | サンプル間 | 2群間 | 複製あり | iDEGES/edgeR(Sun_2013)の記述内容と基本的に同じです。

対応なし | 複製あり | SRP017142(Neyret-Kahn_2013)

```
normalized <- getNormalizedData(tcc)    #正規化後のデータを取り出してnormalizedに格納
tmp <- cbind(rownames(tcc$count), normalized, result)#正規化後のデータの右側にDEG検出
write.table(tmp, out_f7, sep="\t", append=F, quote=F, row.names=F)#tmpの中身をout_f

#ファイルに保存(M-A plot)
png(out_f8, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイルの各種
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
legend("topright", c(paste("DEG(FDR<", param_FDR, ")", sep=""), "non-DEG"),#凡例を作
       col=c("magenta", "black"), pch=20)#凡例を作成している
dev.off()                                #おまじない

#ファイルに保存(各種情報)
sink(out_f9)                             #指定した
cat("1. Numbers of DEGs satisfying several FDR
cat("FDR < 0.05:");print(sum(tcc$stat$q.value
cat("FDR < 0.10:");print(sum(tcc$stat$q.value
cat("
cat("
cat("
sessi
sink(
```

テンプレートとの違いは赤矢印部分のみ



rcode_SRP017142_highlight.txt(の一部)

```
#ファイルに保存(M-A plot)↓
png(out_f8, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファイル
plot(tcc, FDR=param_FDR, xlim=c(-3, 17), ylim=c(-8, 15))#param_FDRで指定した閾
legend("topright", c(paste("DEG(FDR<", param_FDR, ")", sep=""), "non-DEG"),#凡
       col=c("magenta", "black"), pch=20)#凡例を作成している↓
#param <- "ENSG00000240386"                          # 1位↓
#param <- "ENSG00000128564"                          # 2位↓
#param <- c("ENSG00000188064", "ENSG00000101188") # 3,4位↓
#param <- "ENSG00000163431"                          # 5位↓
param <- "ENSG00000189159"                           # 5669位 skyblue↓
#param <- c("ENSG00000166359", "ENSG00000146676") # 24461, 24462位 skyblue↓
obj <- is.element(tcc$gene_id, param)↓
points(result$a.value[obj], result$m.value[obj], pch=20, cex=2, col="skyblue")
dev.off()                                            #おまじない↓
```

Contents（第4回）

- 新規転写物同定（ゲノム情報を利用）
  - 基本的な考え方
  - Tophat-Cufflinksパイプライン
  - 可視化（ゲノムブラウザやViewer）
- 発現量推定（遺伝子レベルと転写物レベル）
  - RPKMの基本的な考え方
  - 計算時間短縮戦略（トランスクリプトーム情報のみを利用）
- カウントデータを用いたサンプル間比較解析
  - イントロ（カウントデータ取得まで）
  - サンプル間クラスタリング
  - 発現変動遺伝子検出
  - 分布やモデル
  - 課題

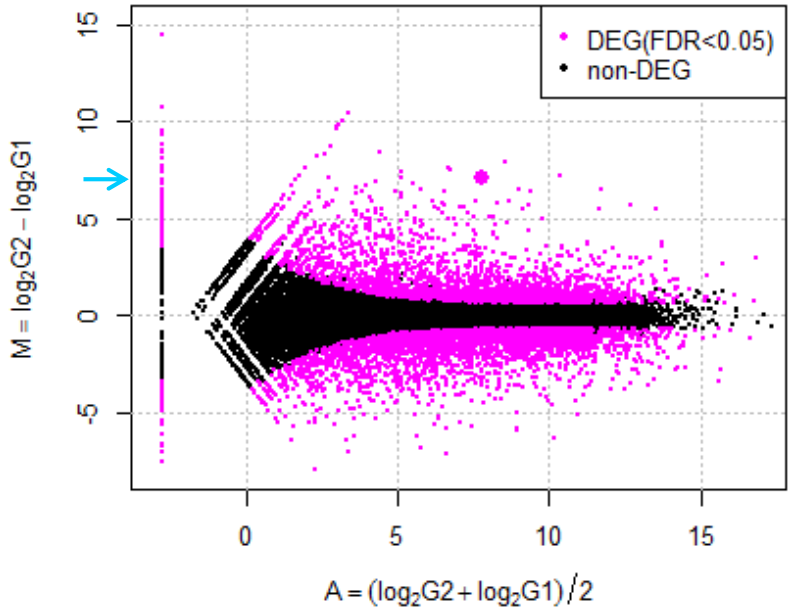
分布やモデルのイントロ

- *TCC*を用いたDEG同定

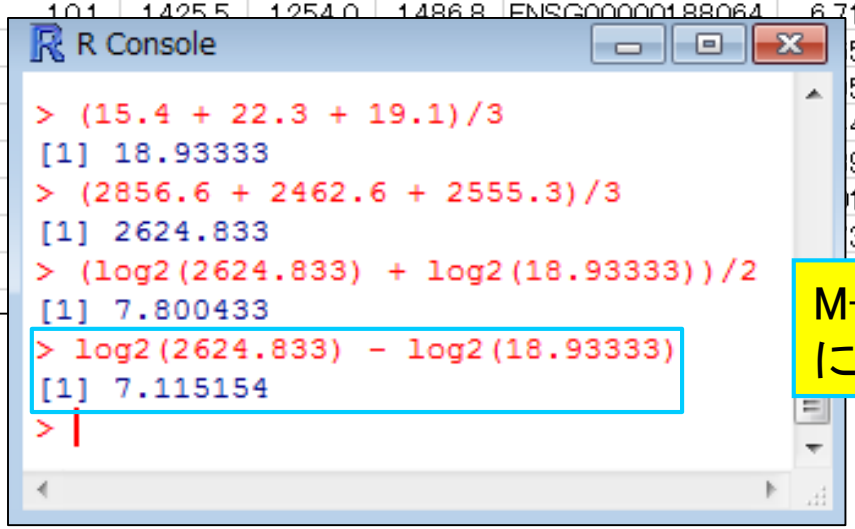
G1群　G2群

59,857 genes

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	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											

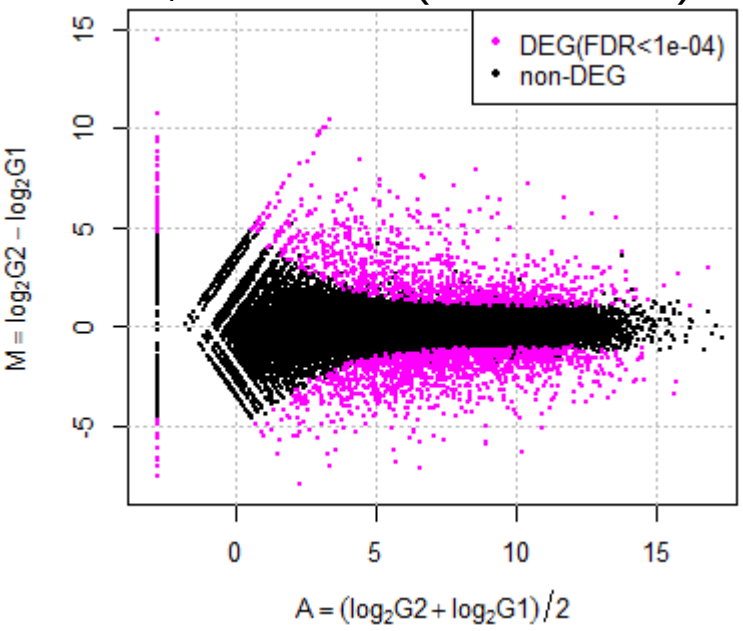


M−A plotのM値は倍率変化（log比）に相当（$2^{7.11}$ 倍G2群で高発現）

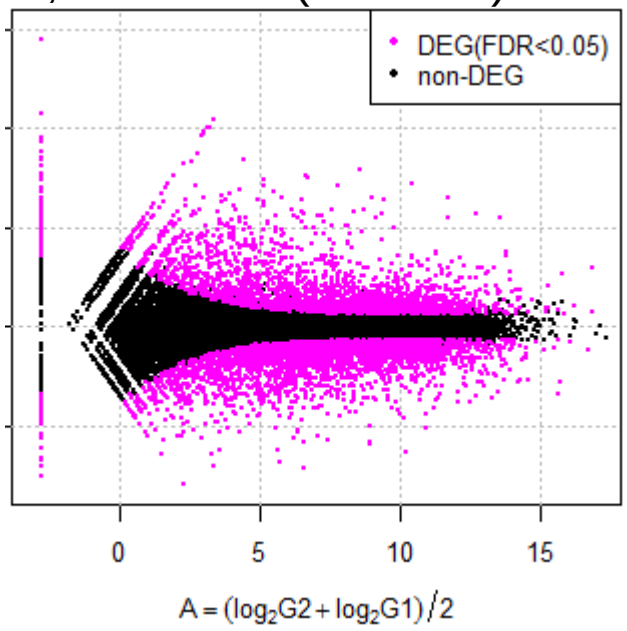
DEG同定結果：FDR閾値の違い

- *TCC*を用いたDEG同定

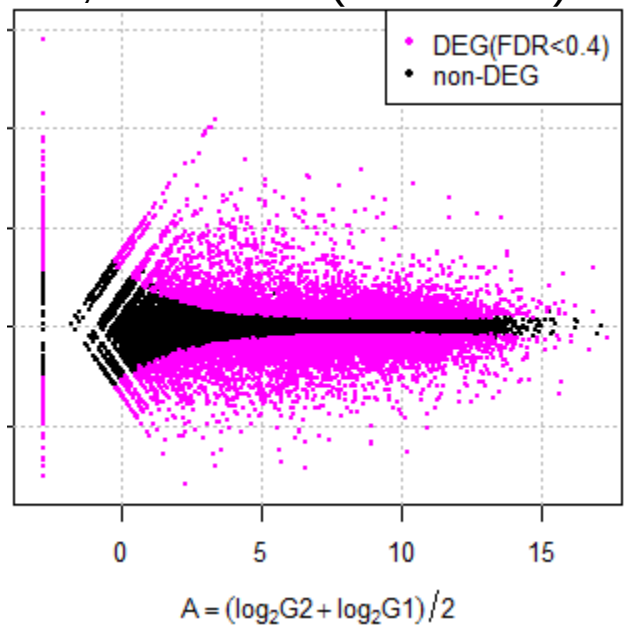
2,314 DEGs (FDR 0.01%)



5,669 DEGs (FDR 5%)



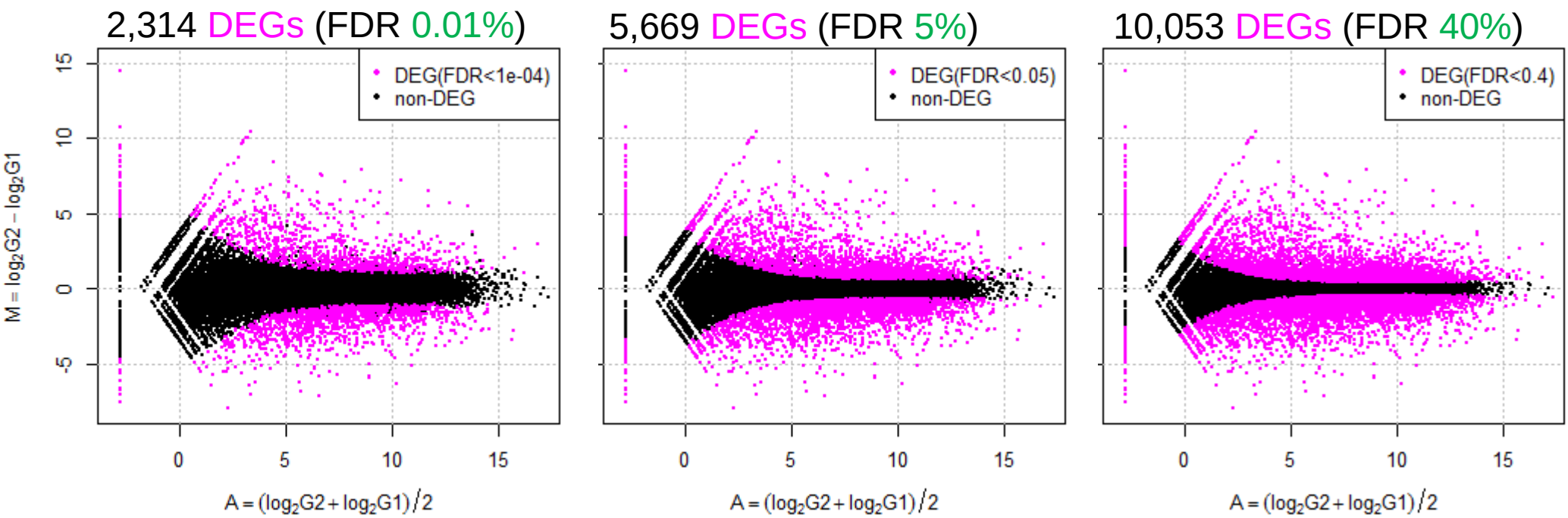
10,053 DEGs (FDR 40%)



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*を用いたDEG同定



黒の分布はnon-DEGの分布に相当

分布やモデル

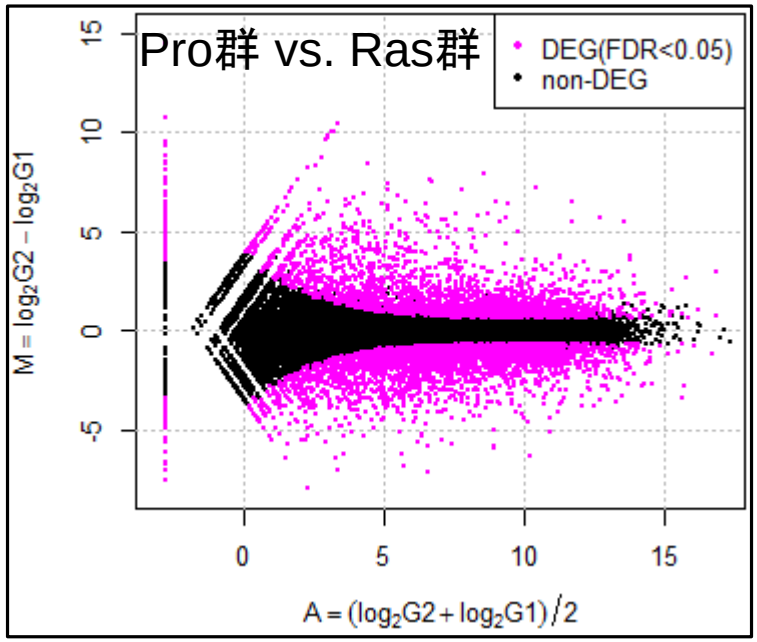
同一群（G1群）　同一群（G2群）

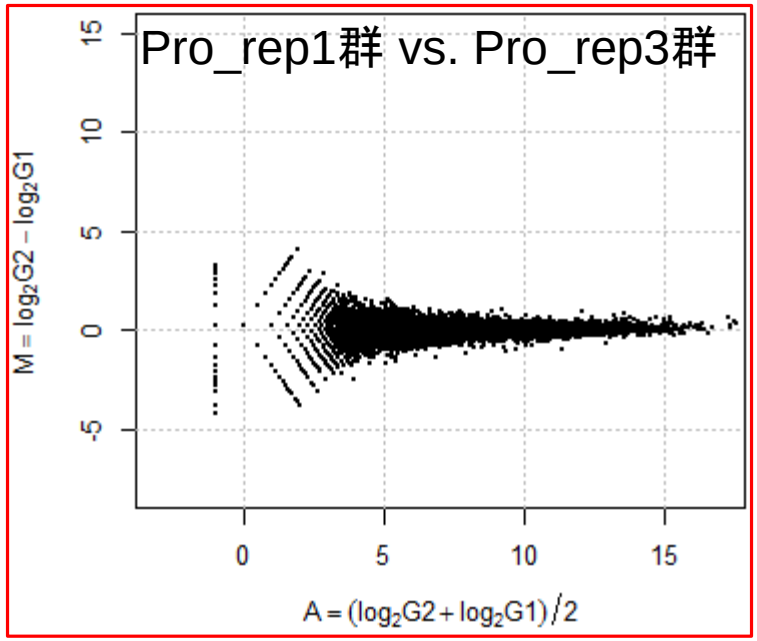
59,857 genes

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	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





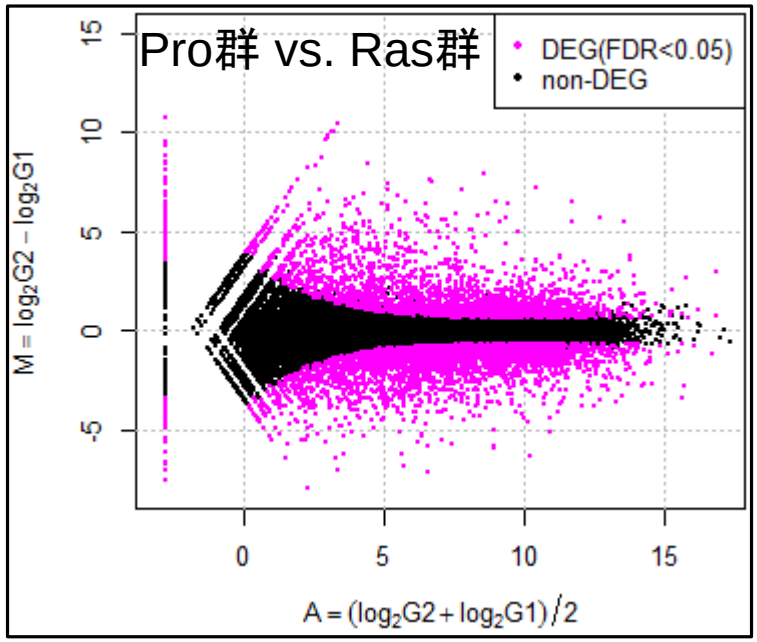
分布やモデル

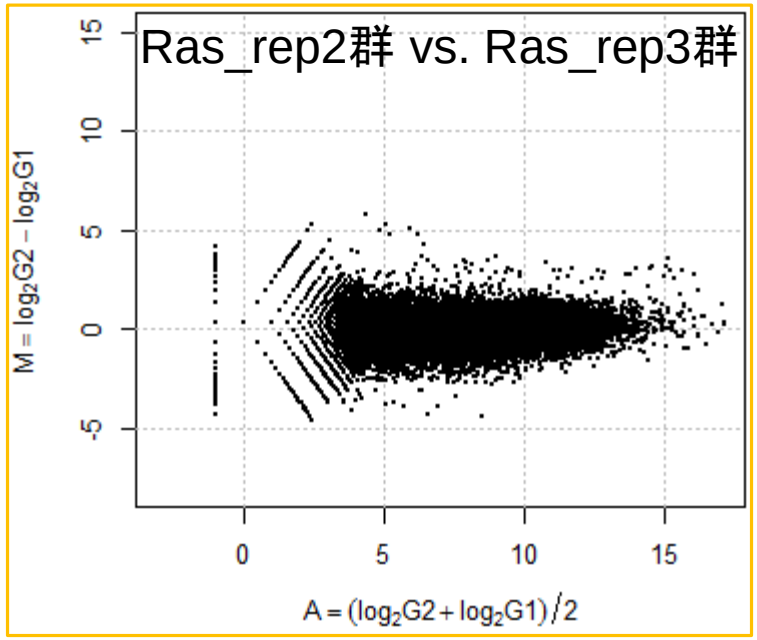
同一群（G1群）　同一群（G2群）

59,857 genes

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	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

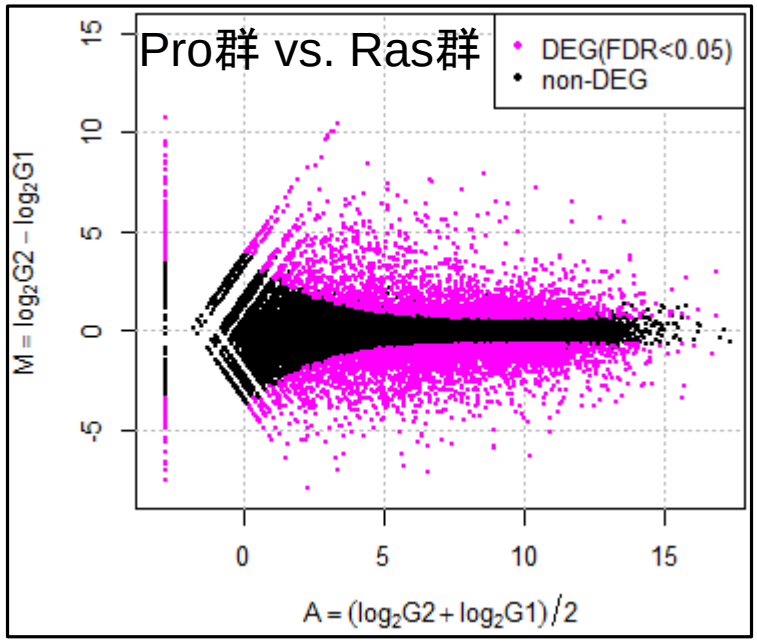
同一群（G1群）　同一群（G2群）

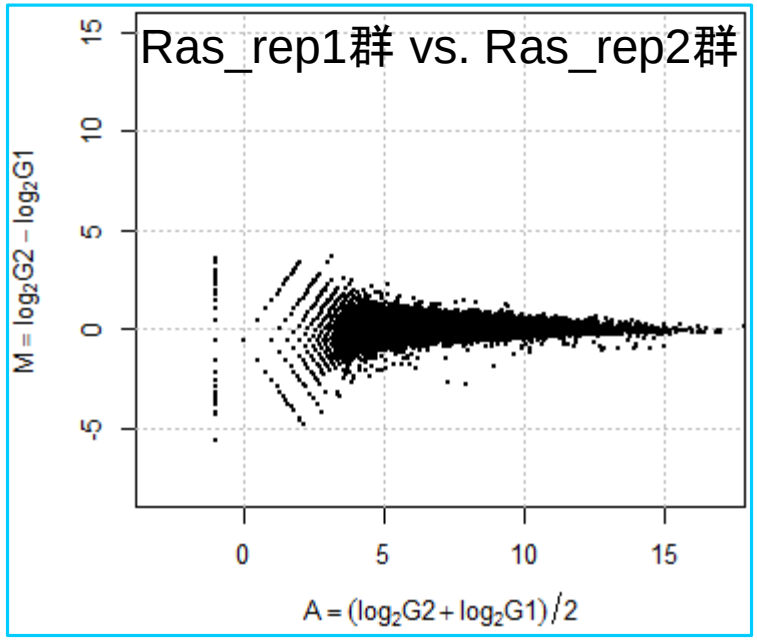
59,857 genes

	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	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

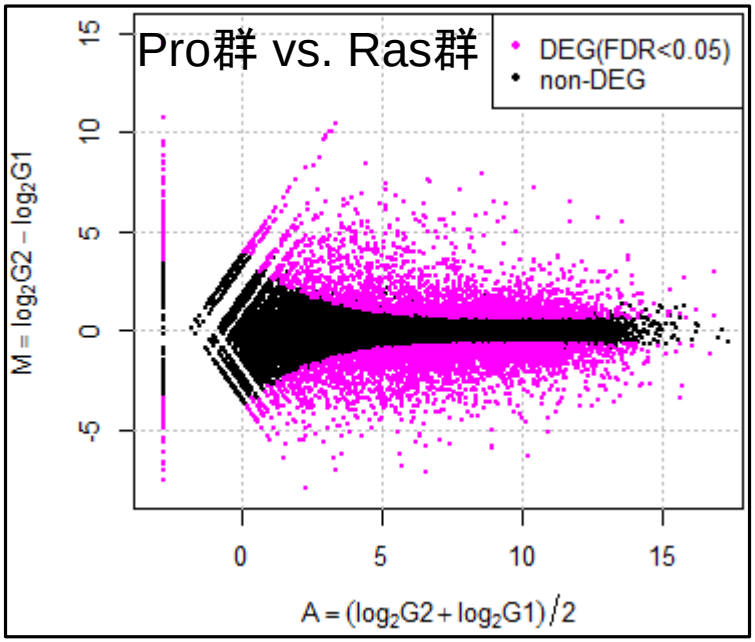
同一群（G1群）　同一群（G2群）

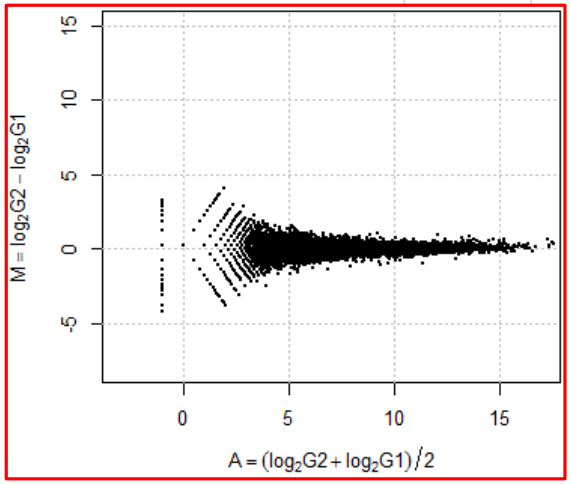
59,857 genes

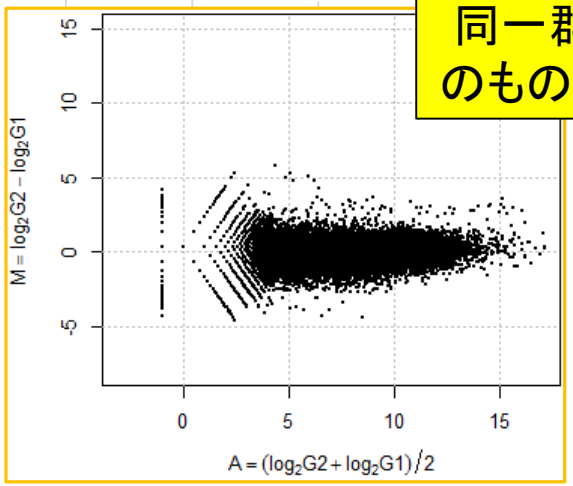
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

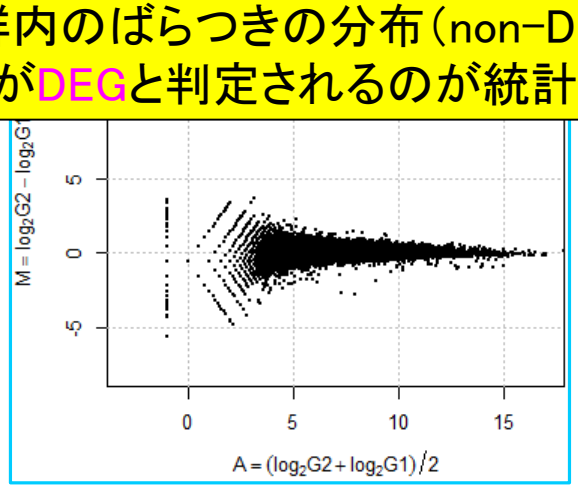
G1群　G2群　G1群　G2群

G1群　G2群







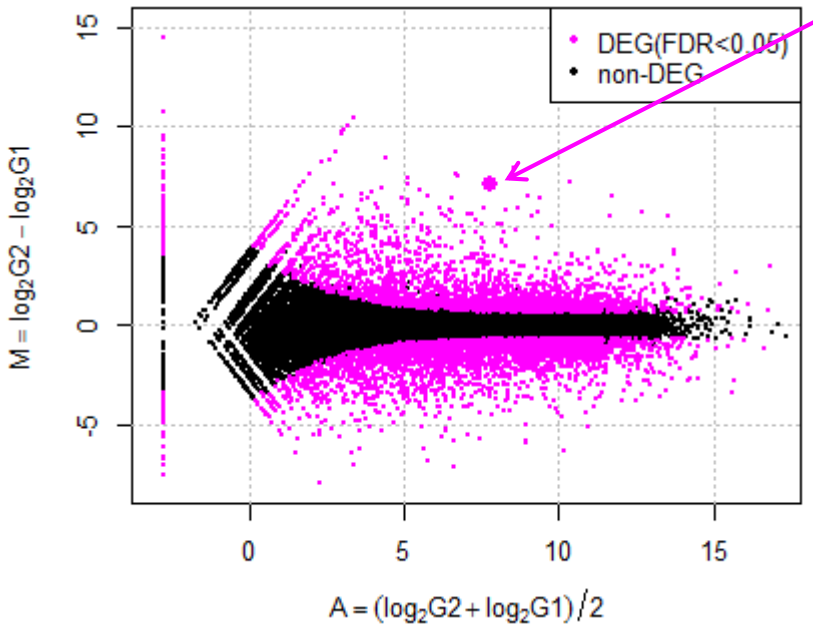


同一群内のばらつきの分布（non-DEG分布）以外のものがDEGと判定されるのが統計的手法の結果

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

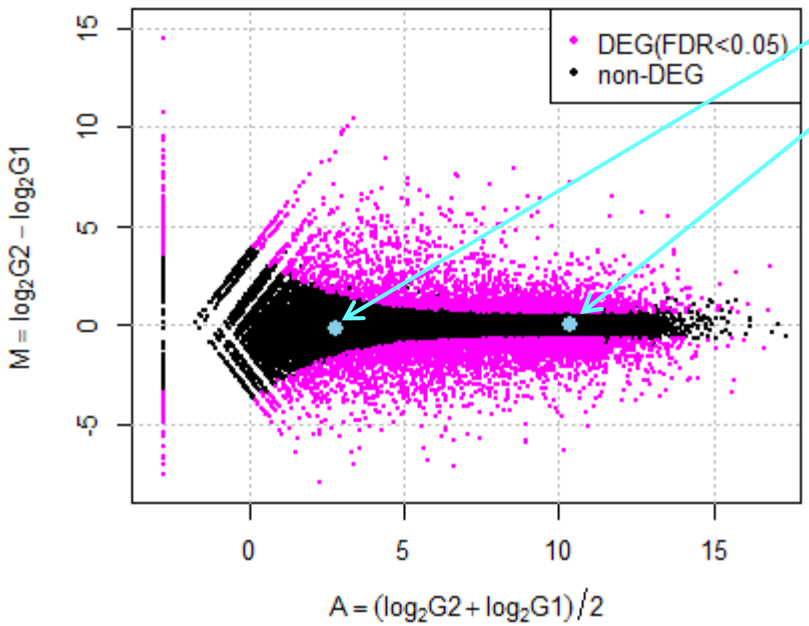


同一群内のばらつきの分布(non-DEG分布)から遠く離れたところに位置するものは0に近い***p***-value

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

倍率変化の結果

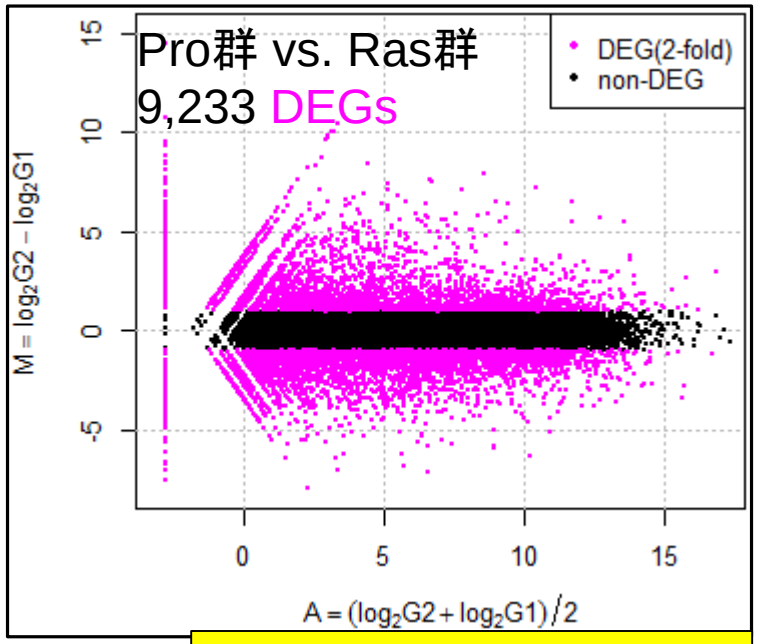
同一群（G1群）　同一群（G2群）

59,857 genes

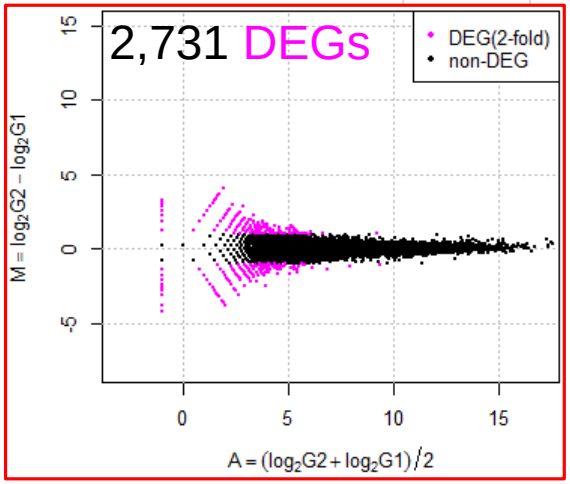
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

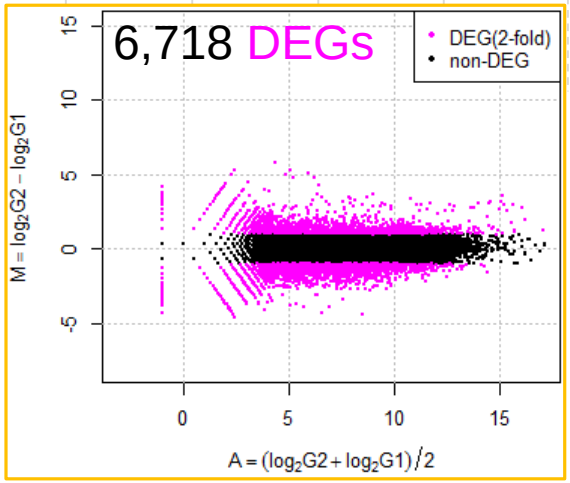
G1群　G2群　G1群　G2群

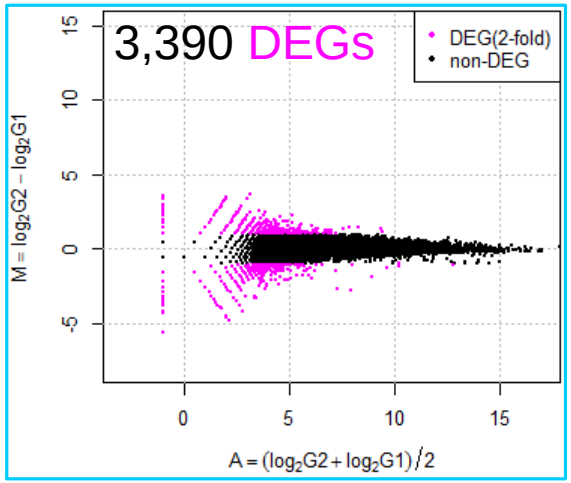
G1群　G2群



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

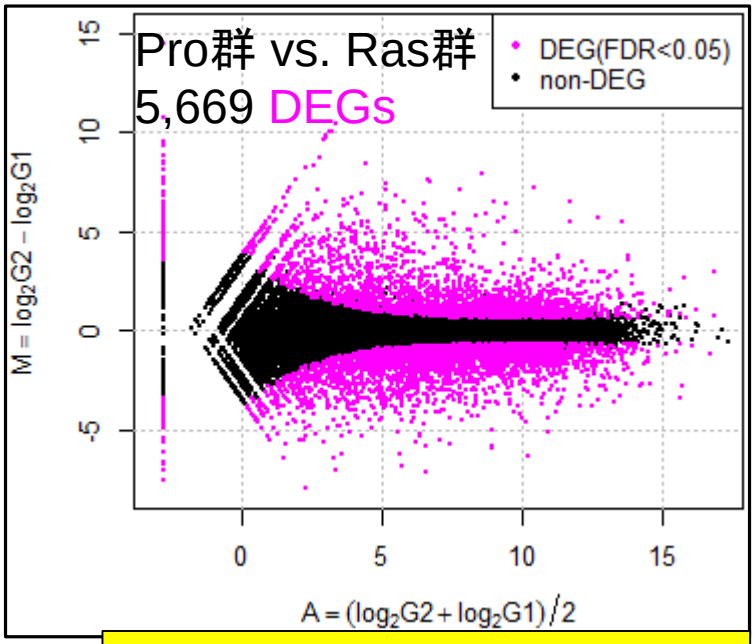
同一群（G1群）　同一群（G2群）

59,857 genes

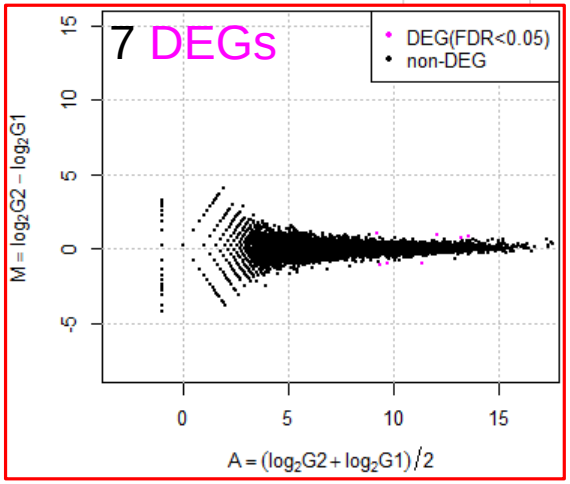
	Pro_rep1	Pro_rep2	Pro_rep3	Ras_rep1	Ras_rep2	Ras_rep3
ENSG00000000003	480	513	366	124	271	366
ENSG00000000005	0	0	0	1	0	0
...						
ENSG00000240386	0	0	0	4001	5500	6851
...						
ENSG00000128564	18	27	19	2038	2657	2138
...						

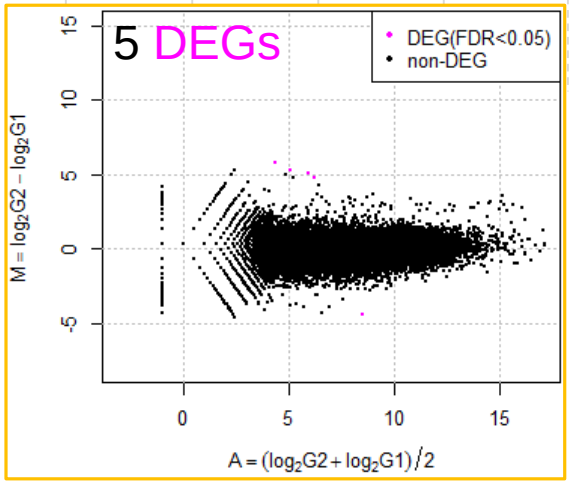
G1群　G2群　G1群　G2群

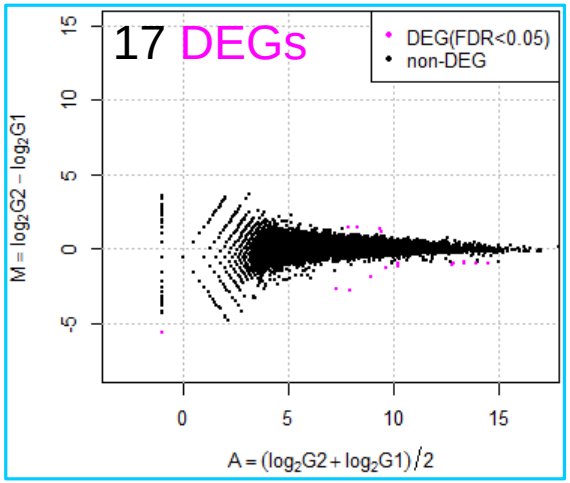
G1群　G2群



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





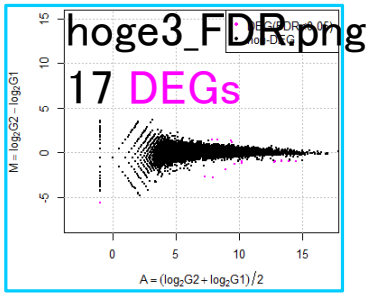


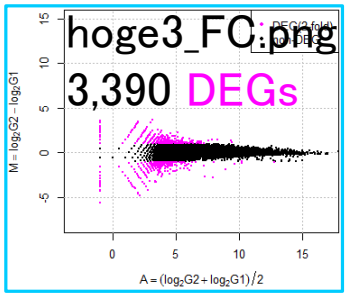
...

```
in_f <- "srp017142_count_bowtie.txt"    #入力ファイル名を指定してin_fに格納
out_f2 <- "hoge3_FDR.png"               #出力ファイル名を指定してout_f2に格納
out_f3 <- "hoge3_FC.png"                #出力ファイル名を指定してout_f3に格納
param_subset <- c(4, 5)                 #取り扱いたいサブセット情報を指定
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検出時のfalse discovery rate (FDR)閾値を指定
param_FC <- 2                           #fold-change閾値(param_FC倍)を指定
param_fig <- c(400, 390)                #MA-plot描画時の横幅と縦幅を指定(単位はピクセル)

library(TCC)                            #パッケージの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep
data <-data[,param_subset]              #param_subsetで
data.cl <- c(rep(1, param_G1), rep(2, param_G2))#G1群を
tcc <- new("TCC", data, data.cl)        #TCCクラスオブ
tcc <- calcNormFactors(tcc, norm.method="deseq", test.
                       iteration=3, FDR=0.1, floorPDEG
tcc <- estimateDE(tcc, test.method="deseq", FDR=param_
result <- getResult(tcc, sort=FALSE)    #p値などの結果を
head(result, n=3)                       #確認してるだけ
sum(tcc$stat$q.value < param_FDR)       #FDR < param_FD
```


rcode_SRP017142_nonDEG.txt。解析したいサンプルの列番号とサンプル数を指定。パッケージのバージョン次第で結果が変わりうるのは確認済み。

```
###################################
### ファイルに保存(M-A plot; FDR)
###################################
png(out_f2, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファ
plot(tcc, FDR=param_FDR, xlim=c(-3,17), ylim=c(-8, 15),#指定した閾値を満た
     normalize=T, median.lines=F)    #指定した閾値を満たすDEGをマゼンタ色
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

###################################
### ファイルに保存(M-A plot; fold-change; FC)
###################################
M <- getResult(tcc)$m.value              #M-A plotのM値を抽出
hoge <- rep(1, length(M))                #初期値を1にしたベクトルhogeを作成
hoge[abs(M) > log2(param_FC)] <- 2        #条件を満たす位置に2を代入
cols <- c("black", "magenta")            #色情報を指定してcolsに格納
png(out_f3, pointsize=13, width=param_fig[1], height=param_fig[2])#出力ファ
plot(tcc, col=cols, col.tag=hoge, xlim=c(-3, 17), ylim=c(-8, 15),#M-A plot
     normalize=T, median.lines=F)    #M-A plotを描画
legend("topright", c(paste("DEG(", param_FC, "-fold)", 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(abs(M) > log2(4))                    #4倍以上発現変動する遺伝子数を表示
sum(abs(M) > log2(2))                    #2倍以上発現変動する遺伝子数を表示
```





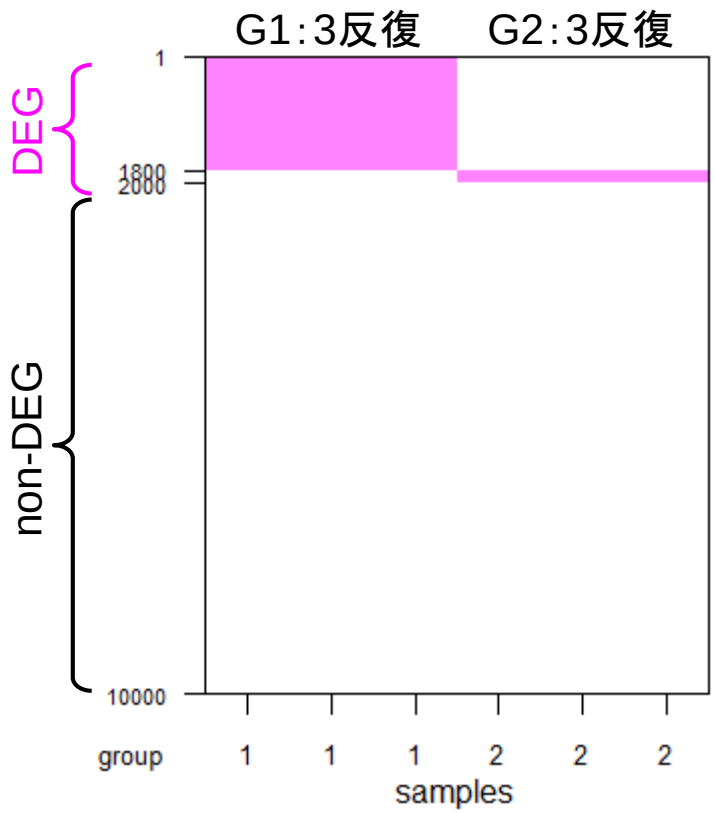
課題用シミュレーションデータ

- 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）

*TCC*パッケージを用いて、
複製あり2群間比較を行う

DEG
G1で高発現
G2で高発現
non-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	42	73	67	62	44	37
gene_9999	5	1	2	3	4	11
gene_10000	2	4	5	2	0	0



課題

- data_hypodata_3vs3.txtのサンプル間比較解析を行う。
  1. TCCパッケージを用いた発現変動遺伝子(DEG)検出を行い、FDR閾値が0.05、0.20、および0.40を満たす遺伝子数を示せ。また、このデータセット中の大まかなDEG数を示すとともにその根拠を簡単に述べよ。
     - FDR閾値0.05を満たす遺伝子数(q-value < 0.05):
     - FDR閾値0.20を満たす遺伝子数(q-value < 0.20):
     - FDR閾値0.40を満たす遺伝子数(q-value < 0.40):
     - このデータセット中に含まれる推定DEG数(偽物を差し引いた本物のDEG数):
     - 推定したDEG数の根拠:
  2. 結果の考察。シミュレーションデータ(data_hypodata_3vs3.txt)のサンプル間クラスタリング結果との比較や、実データ(srp017142_count_bowtie.txt)解析結果との比較など自由に述べてよい。

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 × 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 × 0.001/100 = 0.1
    - p < 0.01を満たすDEG数が400個の場合：FDR = 10,000 × 0.01/400 = 0.25
    - p < 0.05を満たすDEG数が926個の場合：FDR = 10,000 × 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-value < 0.05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$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

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 × 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-value < 0.05に相当
  - False discovery rate (FDR) 1%というのがq-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	estimatedDEG
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。
→4189×0.01 = 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$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
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$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
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×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!!

(Rで)塩基配列解析

~NGS、RNA-seq、ゲノム、トランスクリプトーム、正規化、発現変動、統計、モデル、バイオインフォマティクス~
(last modified 2014/06/30, since 2010)

参考

過去の講義や講演資料の
PDFはこちらから取得可能

What's new?

- このウェブページはフリーソフトRと利用可能なパッケージの多くをインストール済みである前提で記述していますので、Rのインストールと起動を参考にして必要なパッケージのインストールを行ってください。2014年5月14日に記述内容を若干変更しています。
- 2014年7月22日にイルミナウェビナーで話します。興味ある方はどうぞ。(2014/06/30) NEW
- 門田幸二 著シリーズ Useful R 第7巻トランスク
- 2014年9月1日~12日に「バイオインフォマティク
東大農で開催します。受講申込は6/24夕方に
(2014/06/25) NEW
- 参考資料(講義、講習会、本など)の項目を追

- はじめに (last modified 2014/01/30)
- 参考資料 (講義、講習会、本など) modifie
- 過去のお知らせ (last modified 2014/ 0/30) NE
- Rのインストールと起動 (last modified 2014/05/
- サンプルデータ (last modified 2014/06/21) NE
- 書籍 | トランスクリプトームについて (last modif
- 書籍 | トランスクリプトーム解析 | 2.3.1 RNA-se
- 書籍 | トランスクリプトーム解析 | 2.3.2 リファレ

参考資料(講義、講習会、本など) NEW

基本的に私門田の個人ページに記載してあるものです。かなり古い講演資料などの情報をもとに勉強されている方もいらっしゃるようですので、ここでは2013年夏以降の情報を載せておくとともに、大まかな内容についても述べておきます。講演予定のものについては、資料のアップは講演当日が基本です。50-100MB程度ありますがオリジナルのPowerPointファイルがほしい方はお気軽にリクエストしてください。講義資料としての利用などは事前連絡や謝辞も気にせずご自由にお使いください。

書籍

- **門田幸二**著(金明哲 編), シリーズ Useful R 第7巻トランスクリプトーム解析, 共立出版, 2014. ISBN: 978-4-320-12370-0
  内容:マイクロアレイとRNA-seq解析を例としてRを用いてトランスクリプトーム解析を行うための体系的な本としてまとめました。数式が苦手なヒト向けに、重みつき平均の具体的な計算例などを挙げてオプションの意味などがわかるような中身の理解に重点を置いた構成にしてあります。書籍中のRコードは「書籍 | トランスクリプトーム解析 | ...」をご覧ください。
- **門田幸二**,「トランスクリプトミクスの推奨データ解析ガイドライン」, *ニュートリゲノミクスを基盤としたバイオマーカーの開発*, シーエムシー出版, 45-52, 2013. ISBN: 978-4-7813-0820-3
  内容:マイクロアレイ解析の話がメインです。実験デザインの重要性を述べています。Affymetrix GeneChipデータの数値化と発現変動遺伝子(DEG)検出法の組合せの重要性の話や、サンプル間クラスタリングである程度DEGに関する情報がわかることを述べています。MAS5データを用いる場合は特に倍率変化で議論することも無意味であること、RMAのようなマルチアレイ正規化法を用いて得られたマイクロアレイデータの場合には なぜ倍率変化でうまくいく傾向にあるかなどの理由をM-A plotを用いて説明しています。

講習会、講義、講演資料

- **門田幸二**,「講義資料」, アグリバイオインフォマティクス教育研究プログラムの大学院講義科目:農学生命情報科学特論I, 東京大学(東京), 2014.06.25
  内容:教科書の2.3節が中心。デノボゲノムアセンブリ(de novo genome assembly)の大まかな手順を説明。基本的なテクニックとしてk-merの基本的な考え方をシミュレーションデータで解析するとともに、フィルタリング、エ

Jul 2, 2014

まとめ

- 書籍 | トランスクリプトーム解析 | 4.3.3 2群間比較 (last modified 2014/04/
- 書籍 | トランスクリプトーム解析 | 4.3.4 他の実験デザイン(3群間) (last mo
- 書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション modified 2014/0
- イントロ | 一般 | ランダムに行を抽出 (last modified 20 /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/

書籍 | 日本乳酸菌学会誌 | 第一回イントロダクション

日本乳酸菌学会誌の第1回分です。

要約:

- (Rで)塩基配列解析はここ

情報収集先:

- アグリバイオインフォマティクス教育研究プログラム
- 日本バイオインフォマティクス学会(JSBi)
- NGS現場の会
- 産総研CBRC
- HPCI人材養成プログラム
  - e-learning
  - tutorial(講習会)
  - セミナー
  - ワークショップ
- 統合TV: Kawano et al., Brief Bioinform., 2012
  - WindowsでUNIX! 1. Cygwin インストール編
  - WindowsでUNIX! 2. ファイル操作編
  - WindowsでUNIX! 3. ファイル操作応用編
  - WindowsでUNIX! 4. ファイル操作発展編
  - Perlの使い方 インストール編(Windows)
  - Perlの使い方 事例編(前編)

NGS解析に必要な全般的な解説記事

講演予定はこちら。リンク先などから芋づる式に情報収集

講演など(上記講義以外)(last modified: 2014.06.30)

32. 門田幸二,「Rでゲノム・トランスクリプトーム解析:CpG解析から機能解析まで」, HPCI講習会・バイオインフォマティクス実習コース, 産業技術総合研究所ゲノム情報研究センター(東京), 2015.03.05-06
31. 門田幸二,「フリーソフトRを用いたビッグデータ解析:塩基配列解析を中心に」, 生命医薬情報学連合大会2014, 中級者向けバイオインフォマティクス入門講習会, 東北大学(宮城), 10:50-12:20, 2014.10.04
30. 門田幸二,「ビッグデータ解析とR」, 生命医薬情報学連合大会2014, HPCIワークショップ「医療とビッグデータ解析」, 東北大学(宮城), 9:00-10:30, 2014.10.04
29. 門田幸二,「3. データ解析基礎」, バイオインフォマティクス人材育成カリキュラム(次世代シークエンサ)速習コース, 東京大学(東京), 2014.09.08-09
28. 門田幸二,「トランスクリプトームデータ解析戦略2014」, イルミナウェビナー・RNA-Seqシリーズ, イルミナ株式会社(東京), 2014.07.22