RNA-Seqデータ解析における正規化法の選択：RPKM値でサンプル間比較は危険?!

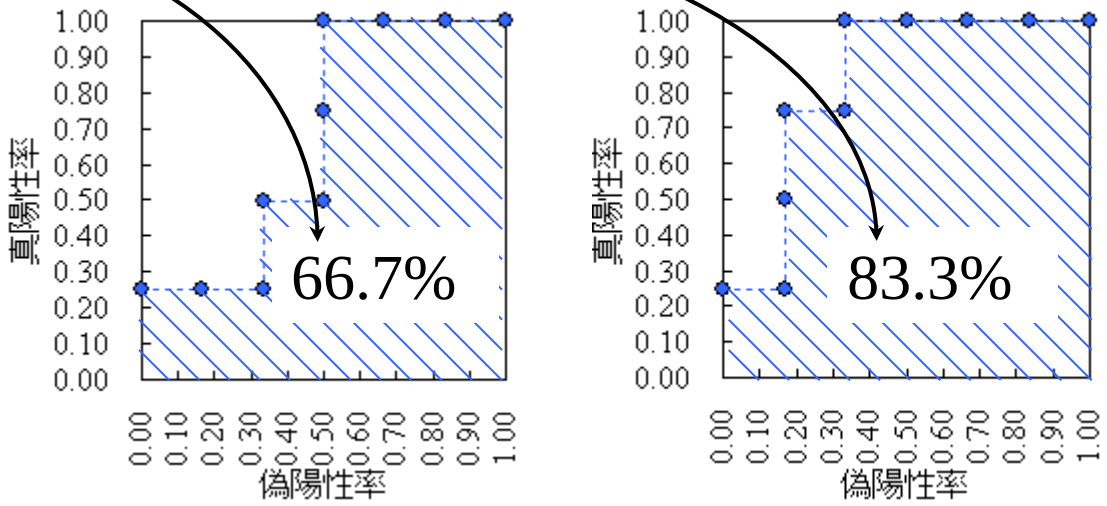
東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
門田 幸二（かどた　こうじ）
http://www.iu.a.u-tokyo.ac.jp/~kadota/
kadota@iu.a.u-tokyo.ac.jp

よりよい正規化法とは？

- その正規化法によって得られたデータ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	正規化法1		正規化法2	
1	gene8	真	gene8	真
2	gene5	偽	gene5	偽
3	gene4	偽	gene3	真
4	gene3	真	gene2	真
5	gene7	偽	gene7	偽
6	gene1	真	gene1	真
7	gene2	真	gene4	偽
8	gene9	偽	gene9	偽
9	gene10	偽	gene10	偽
10	gene6	偽	gene6	偽



Area Under the ROC Curve (ROC曲線の下部面積：AUC)

出発点

- （マイクロアレイと同じく）マップされたリード数情報を含む遺伝子発現行列データ

SupplementaryTable2_changed.txt

	A	B	C	D	E	F	G	H	I	J	K
1	EnsemblGeneID	R1 L1 Kidne	R1 L3 Kidne	R1 L7 Kidne	R2 L2 Kidne	R2 L6 Kidne	R1 L2 Liver	R1 L4 Liver	R1 L6 Liver	R1 L8 Liver	R2 L3 Liver
2	ENSG00000146556	0	0	0	0	0	0	0	0	0	0
3	ENSG00000197194	0	0	0	0	0	0	0	0	0	0
4	ENSG00000197490	0	0	0	0	0	0	0	0	0	0
5	ENSG00000205292	0	0	0	0	0	0	0	0	0	0
6	ENSG00000177693	0	0	0	0	0	0	0	0	0	0
7	ENSG00000209338	0	0	0	0	0	0	0	0	0	0
8	ENSG00000196573	0	0	0	0	0	0	0	0	0	0

SupplementaryTable2_changed

代表的な正規化法：RPKM

- RPM正規化（マイクロアレイなどと同じところ）
  - Reads per million mapped reads
  - サンプルごとにマップされた総リード（塩基配列）数が異なる。
    →各遺伝子のマップされたリード数を「総read数が100万（one million）だった場合」に補正

「raw counts : all reads= RPM : 1,000,000 」
A1BGの場合は「744 : 5,087,097 = RPM : 1,000,000」

$$\mathrm{RPM} = \text{raw counts} \times \frac{1{,}000{,}000}{\text{all reads}} = 744 \times \frac{1{,}000{,}000}{5{,}087{,}097} = 146.3$$

geneName	raw counts	RPKM	all reads	gene length	RPM
A1BG	744	82.9	5087097	1764	146.3
A1CF	159	13.7	5087097	2278	31.3
A2BP1	1	0.0	5087097	5415	0.2
A2LD1	4	0.6	5087097	1226	0.8

- RPKM正規化（RNA-seq特有）
  - Reads per kilobase of exon per million mapped reads
  - 遺伝子の配列長が長いほど配列決定(sequence)される確率が上昇
    →各遺伝子の配列長を「1000塩基（one kilobase）の長さだった場合」に補正

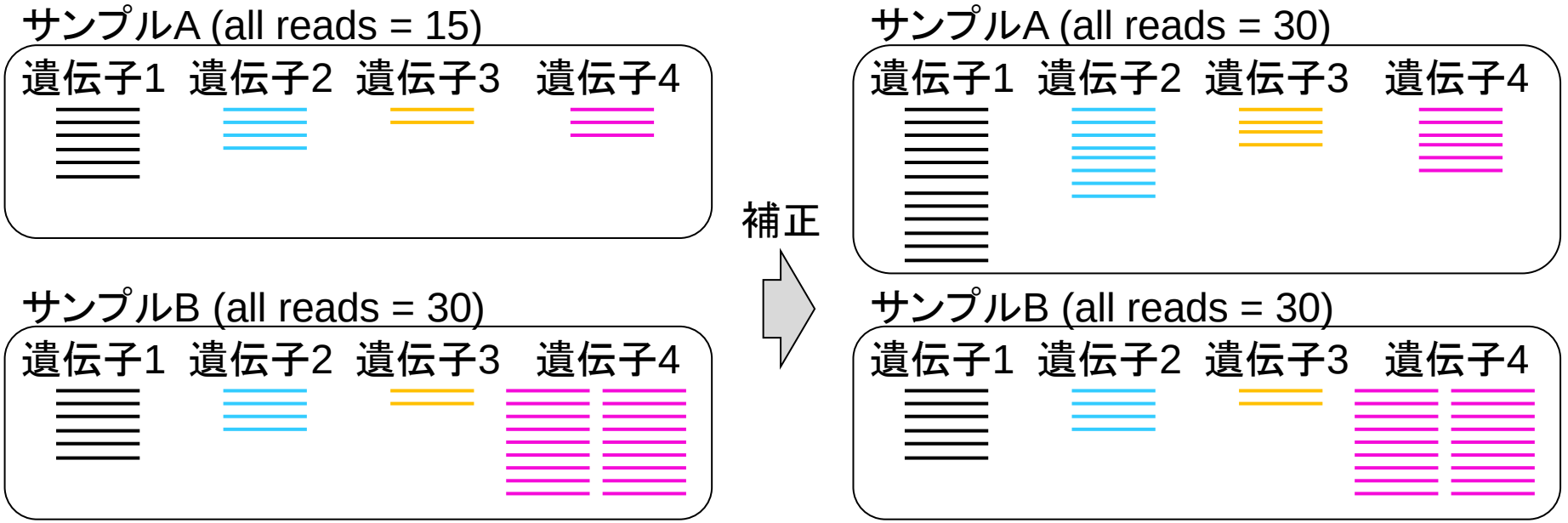
$$\mathrm{RPKM} = \text{raw counts} \times \frac{1{,}000{,}000}{\text{all reads}} \times \frac{1{,}000}{\text{gene length}} = \text{raw counts} \times \frac{1{,}000{,}000{,}000}{\text{gene length} \times \text{all reads}}$$

$$\mathrm{A1BG} = 744 \times \frac{1{,}000{,}000{,}000}{1{,}764 \times 5{,}087{,}097} = 82.9$$

RPKM（正確にはRPM）の問題点

- 仮定
  - 全4遺伝子
  - 長さが同じ（gene lengthの項を無視できるので）
  - 遺伝子4だけが発現変動遺伝子(DEG)

$$\text{raw counts} \times \frac{\text{定数}}{\cancel{\text{gene length}} \times \text{all reads}}$$

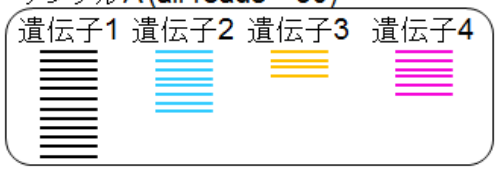


補正後の解析結果：Aで高発現が3個, Bで高発現が1個

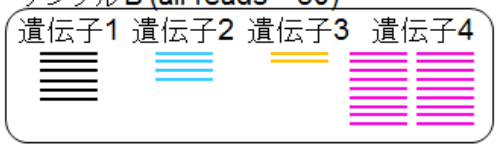
TMM正規化法

RPM補正後のデータ

サンプルA (all reads = 30)



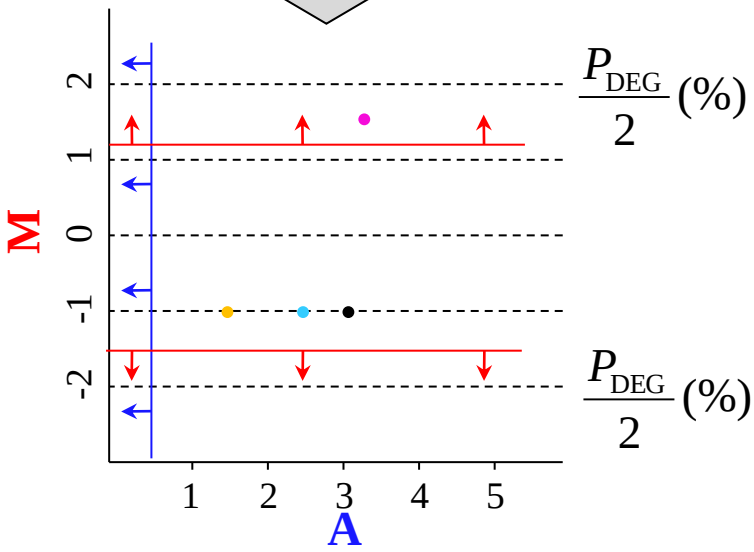
サンプルB (all reads = 30)



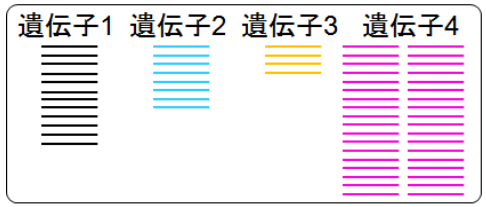
	A	B
遺伝子1	12	6
遺伝子2	8	4
遺伝子3	4	2
遺伝子4	6	18

$\log_2$(A)	$\log_2$(B)
3.58	2.58
3.00	2.00
2.00	1.00
2.58	4.17

縦軸（発現比） **R**: $\log_2$(B/A)	横軸（全体的な発現レベル） **I**: log2(sqrt(A×B))
M: $\log_2$(B)-$\log_2$(A)	**A**: ($\log_2$(A)+$\log_2$(B))/2
-1.00	3.08
-1.00	2.50
-1.00	1.50
1.58	3.38



TMM補正後のサンプルBのデータ



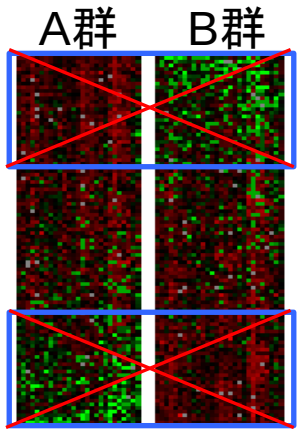
	$\log_2$(A)	$\log_2$(B) - TMM
遺伝子1	3.58	3.58
遺伝子2	3.00	3.00
遺伝子3	2.00	2.00
遺伝子4	2.58	5.17

TMM = -1

縦軸で上位下位合わせてP_{DEG} %をTrim
→ 残りのデータでMのweighted Mean（TMM）を計算

TMM補正の有無で結論が異なることも…

- 得られたDEGセット中の割合
  - TMM補正なし（Marioni et al., *Genome Res.*, 2008）
    - サンプルA（Kidney）: 78%
    - サンプルB（Liver）: 22%
  - TMM補正あり（Robinson and Oshlack, 2010）
    - サンプルA（Kidney）: 53%
    - サンプルB（Liver）: 47%
- TMM法で使用されているパラメータ(一部)
  - $\log_2$(B/A)で発現変動順にランキングし、全体で全遺伝子数の60%分をTrim (P_{DEG} = 60%)。その内訳は、サンプルA側とサンプルB側で高発現なものを各50%とする(P_A = 50%) 。



Trim 後に残ったデータのみを用いて正規化係数を決定

http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html

参考URL

(Rで)塩基配列解析(主に次世代シーケンサーのデータ) by 門田幸二 (last modified 2011/08/26)

What's new?
・2011年9月以降、次世代シーケンサー解析周辺の話をいくつかやります。初心者向けのが9/8と11/17, 私の最新の手法の話が9/29と11/11の予定です。定員に限りがあるようですので、詳細は私のホームページの「講演など」の項目をご覧ください。(2011/08/16)NEW
・アノテーション情報取得(BioMart and biomaRt)のところで配列長情報取得時の誤りに気づきましたm(_ _)m 2011年8月16日14:20までに一通り修正してあります。(2011/08/10-16)NEW
・FASTQ形式ファイル周辺の記述を追加しています。(2011/08/1-4)
・Bioconductorのリンク先をver. 2.7 --> 2.8に変更しました。(2011/07/20)
・R2.13.1がリリースされていたのでこれに変更しました。(2011/07/14)
・DEGseqパッケージ関連のパラメータ指定ミスを修正しました。具体的には例えば「expCol1=1」→「expCol1=2」でしたm(_ _)m(2011/06/09)

- はじめに (last modified 2011/07/19)
- Rのインストールと起動(last modified 2011/08/18) NEW
- サンプルデータ(last modified 2011/02/03)
- イントロダクション | NGS | 各種覚書(last modified 2010/12/10)
- イントロダクション | NGS | 様々なプラットフォーム(last modified 2011/07/15)
- イントロダクション | NGS | リファレンス配列取得(マップされる側)(last modified 2011/02/03)
- イントロダクション | NGS | リファレンス配列取得後の各種情報抽出(特にRefSeq)(last modified 2011/03/20)
- イントロダクション | NGS | リファレンス配列取得後の各種情報抽出2(readFASTA関数の利用)(last modified 2011/04/07)
- イントロダクション | NGS | アノテーション情報取得(refFlatファイル)(last modified 2010/12/07)
- イントロダクション | NGS | アノテーション情報取得(BioMart and biomaRt)(last modified 2011/08/26) NEW
- イントロダクション | 一般 | 配列取得(last modified 2010/7/7)
- イントロダクション | 一般 | 指定した範囲の配列を取得(last modified 2011/07/27)
- イントロダクション | 一般 | 翻訳配列(translate)を取得(last modified 2011/07/27)
- イントロダクション | 一般 | 相補鎖(complement)を取得(last modified 2011/07/27)
- イントロダクション | 一般 | 逆相補鎖(reverse complement)を取得(last modified 2011/07/27)
- イントロダクション | 一般 | 逆鎖(reverse)を取得(last modified 2011/07/27)
- イントロダクション | 一般 | 二連続塩基の出現頻度情報を取得(last modified 2011/07/25)

TMM-normalized dataの作成法

• **前処理 | TMM正規化(Robinson_2010)**

前処理 | についてでも述べていますがNGSデータはマイクロアレイに比べてダイナミックレンジが広いという利点はあるとは思いますが、RPMやRPKMで実装されているいわゆるグローバル正規化に基づく方法はごく少数の高発現遺伝子の発現レベルの影響をもろに受けます。そしてこれらが比較するサンプル間で発現変動している場合には結論が大きく変わってしまいます。なぜなら総リード数に占める発現変動遺伝子のリード数の割合が大きいからです。
Robinsonらは参考文献1の「腎臓 vs. 肝臓」データの比較において実際にこのような現象が起きていることを(housekeeping)遺伝子の分布を真として)示し、少数の高発現遺伝子の影響を排除するためにtrimmed mean of M values (TMM)という正規化法を提案しています(参考文献2)。
この方法はRのedgeR (参考文献3)というパッケージ中にcalcNormFactorsという名前の関数で存在しますので、ここではこの関数を用いてTMM正規化した後のデータを得ることを目的としたやり方を紹介します。
尚、TMM正規化は実質的にRPM(or RPKM)正規化とセットで行われますので、第一段階でRPM正規化、第二段階でRPM正規化後のデータをもとにTMM正規化を実行、という流れで示します。
ここではサンプルデータ2のSupplementaryTable2_changed.txtの「A群 5サンプル vs. B群 5サンプル」のraw counts(特定の遺伝子領域にいくつリードがマップされたかをただカウントした数値データ)の二群間比較データを入力としてTMM補正後のファイルを出力するやり方を示します。
実際に掛っている正規化係数についての説明をしておくと、例えばR1L1Kidneyはマップされた総リード数が1804977です。RPM補正は総リード数を1,000,000にすることに相当しますので、1,000,000/1804977 = 0.5540237という正規化係数が掛っています。全サンプルの正規化係数はnorm_factor1で見られます。TMM正規化係数はnorm_factor2で見ることができます。もしこの値が全て1付近にあればRPKM補正後のデータを解析した結果と似た結果になることを意味し、例えばR1L1Kidneyに掛っている0.8222570のように1から離れた値になっていればTMM正規化後のデータ解析結果はRPKMのそれといくぶん異なる結果になることを意味します。

「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. 配列長情報なし(RPMとTMMを明示的に分離):

```
------    ここから    ------
in_f <- "SupplementaryTable2_changed.txt"                    #読み込みたい発現データファイルを指定してin_f1に格納
out_f <- "SupplementaryTable2_changed_TMM1.txt"              #出力ファイル名を指定
param1 <- 5                                                  #A群のサンプル数を指定
param2 <- 5                                                  #B群のサンプル数を指定
param3 <- 1000000                                            #補正後の総リード数を指定(RPMにしたい場合はここの数値はそのまま)

library(edgeR)                                               #パッケージの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="¥t", quote="")#発現データファイルの読み込み
data <- as.matrix(data)                                      #データの型をmatrixにしている

#第1段階：RPM正規化
norm_factor1 <- param3/colSums(data)                         #各列に対して掛ける正規化係数を計算してnorm_factor1に格納
```

利用可能なRパッケージたち

- *DEGseq* (Wang et al., *Bioinformatics*, **26**: 136-138, 2010)
  - ポワソン分布（variance = mean）を仮定しているためばらつきを過少評価
- *edgeR* (Robinson et al., *Bioinformatics*, **26**: 139-140, 2010)
  - 正規化法：TMM法
  - 負の二項分布（variance > mean）を仮定。meanのみのパラメータを用いて現実のばらつきを表現
- *DESeq* (Anders and Huber, *Genome Biol*., **11**: R106, 2010)
  - 正規化法：RLE法(relative log expression)
  - edgeRのモデルをさらに拡張（しているらしい）
- *baySeq* (Hardcastle and Kelly, *BMC Bioinformatics*, **11**:422, 2010)
  - 正規化法：RPM (たぶん)
  - 配列の長さ情報を与えるオプションがある
  - **データセット中に占めるDEGの割合（P_{DEG}）を一意に返す**
- *NBPSeq* (Di et al., *SAGMB*, **10**:24, 2011)

アルゴリズムの詳細ついては不明です

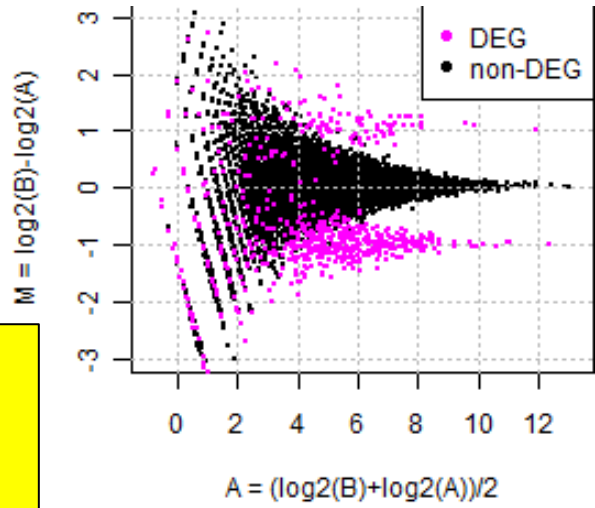
TMM法と門田法

- TMM法で用いられているパラメータ
  - P_{DEG} = 60%（全遺伝子数の60%分をTrimした残りのデータで正規化係数を決定）
  - P_{A} = 50%（発現変動遺伝子の内訳:「A群 > B群」=「A群 < B群」）

- 門田法
  - *baySeq*で自動推定されたP_{DEG}に相当しないデータを用いて正規化係数を決める

シミュレーションデータ

- TMM paper の**Fig. 3**で用いられたものと同じ関数を使用
  - ポアソン分布
  - 発現変動遺伝子（DEG）の倍率変化: 2
  - Number of common genes: 20,000
  - sample Aのみで発現している遺伝子数: **2,200** → 0
  - sample Bのみで発現している遺伝子数: 0
  - DEGの割合（P_{DEG}）: 5%
  - DEG中に占めるsampleA > Bの割合（P_A）: 80%

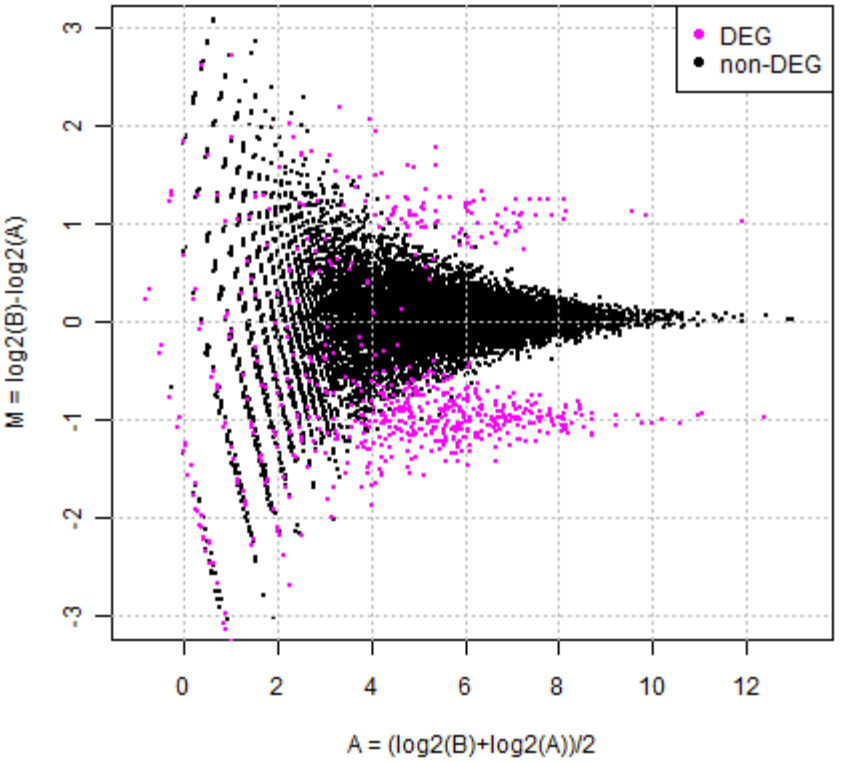


全遺伝子数が20,000個
そのうち5%（1,000個）がDEG（P_{DEG} = 5%）
80%（800個）がsampleAで高発現（P_A = 80%）

門田法のイメージ

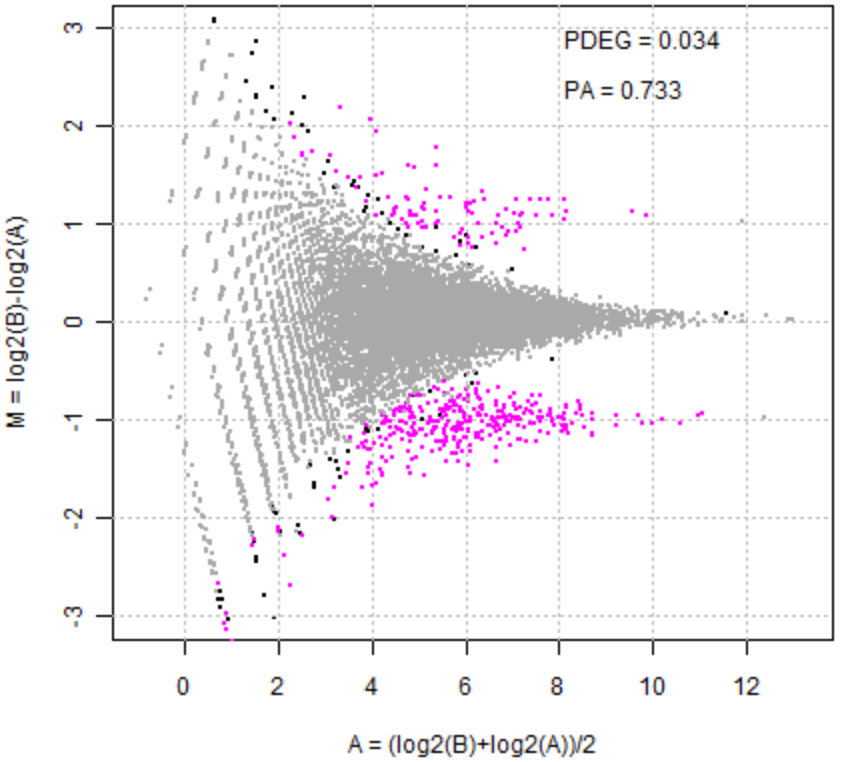
真実

P_{DEG}= 5%, P_A = 80%



*baySeq*による推定値

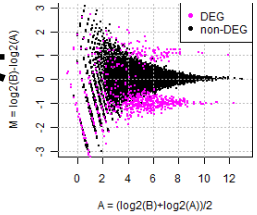
P_{DEG}= 3.4%, P_A = 73.3%



*baySeq*でDEGと判定されなかった（灰色部分に相当）データのみを用いて正規化係数を決定

評価基準

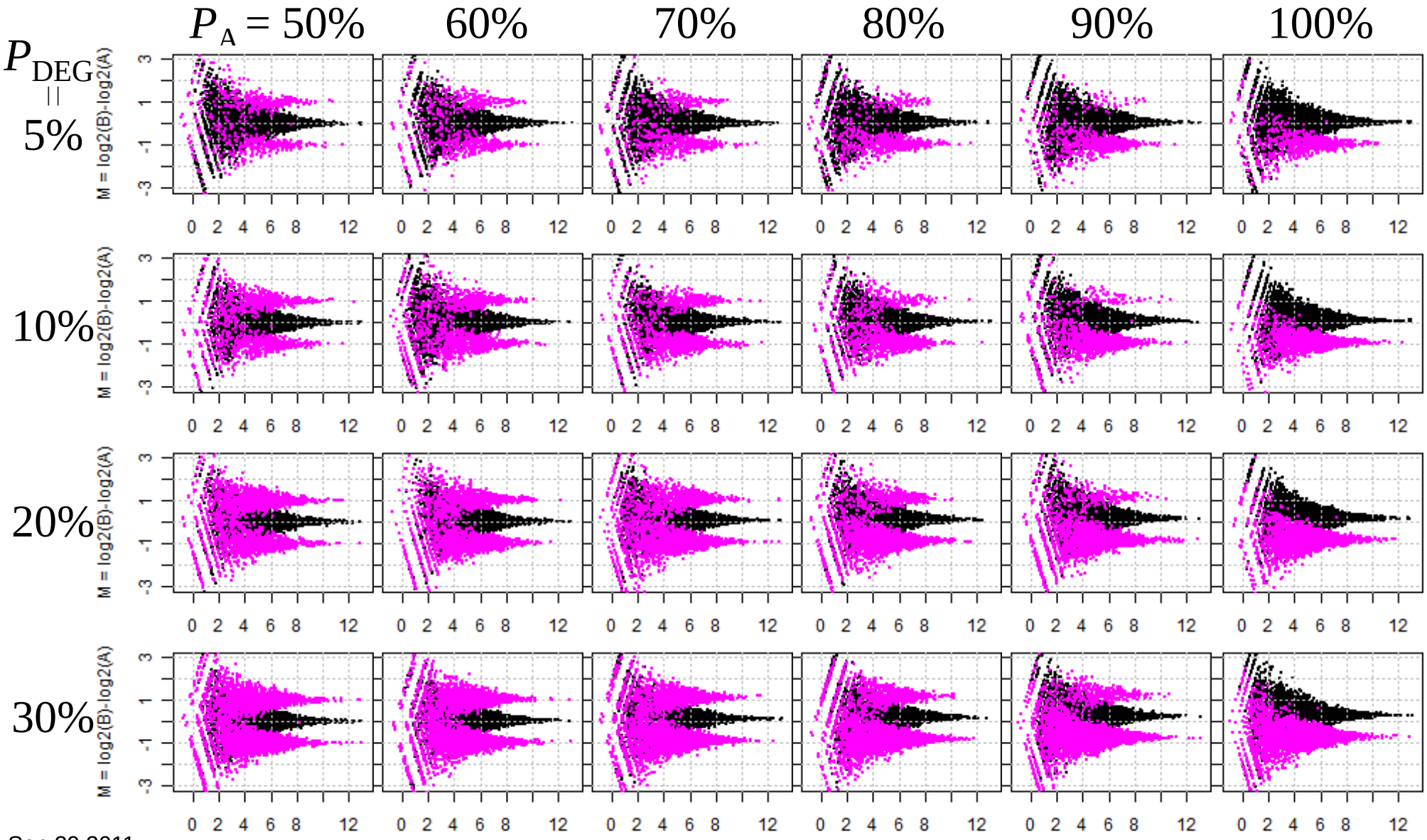
- AUC値（高いほど感度・特異度が高いことを意味する）
- あるシミュレーション条件（P_{DEG}=5%, P_A=80%）の結果



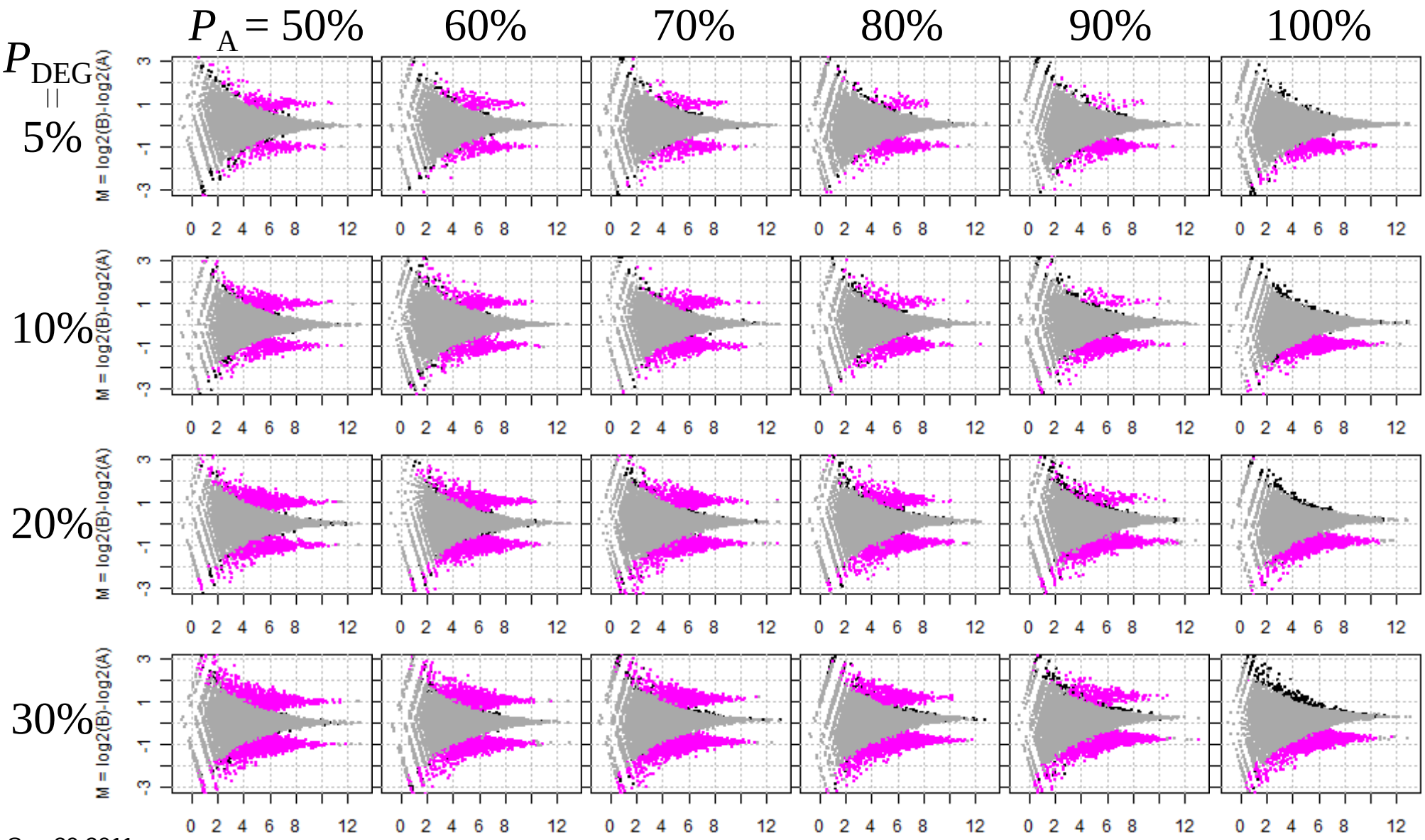
Rパッケージ	*edgeR*			*DESeq*			*baySeq*			*NBPSeq*		
正規化法	RAW	TMM	門田法	RAW	TMM	門田法	RAW	TMM	門田法	RAW	TMM	門田法
Trial1	78.75	79.43	79.26	79.61	80.37	80.34	87.61	88.13	88.03	76.66	77.03	77.06
Trial2	77.13	76.76	76.82	78.85	78.90	79.04	88.27	88.76	88.79	76.08	76.44	76.37
Trial3	78.42	78.57	77.63	78.07	78.53	78.55	89.41	89.71	89.64	77.37	77.75	77.71
Trial4	79.45	79.48	79.40	77.78	77.78	77.78	87.59	87.69	87.66	75.99	76.14	76.04
Trial5	77.91	77.88	77.54	78.03	78.06	78.18	89.00	89.17	89.12	77.01	76.94	77.12
Average	78.33	**78.43**	78.13	78.47	78.73	**78.78**	88.37	**88.69**	88.65	76.62	76.86	**76.86**

想定される様々なシナリオに対する頑健性はどうか？

様々なシナリオ



門田法のイメージ



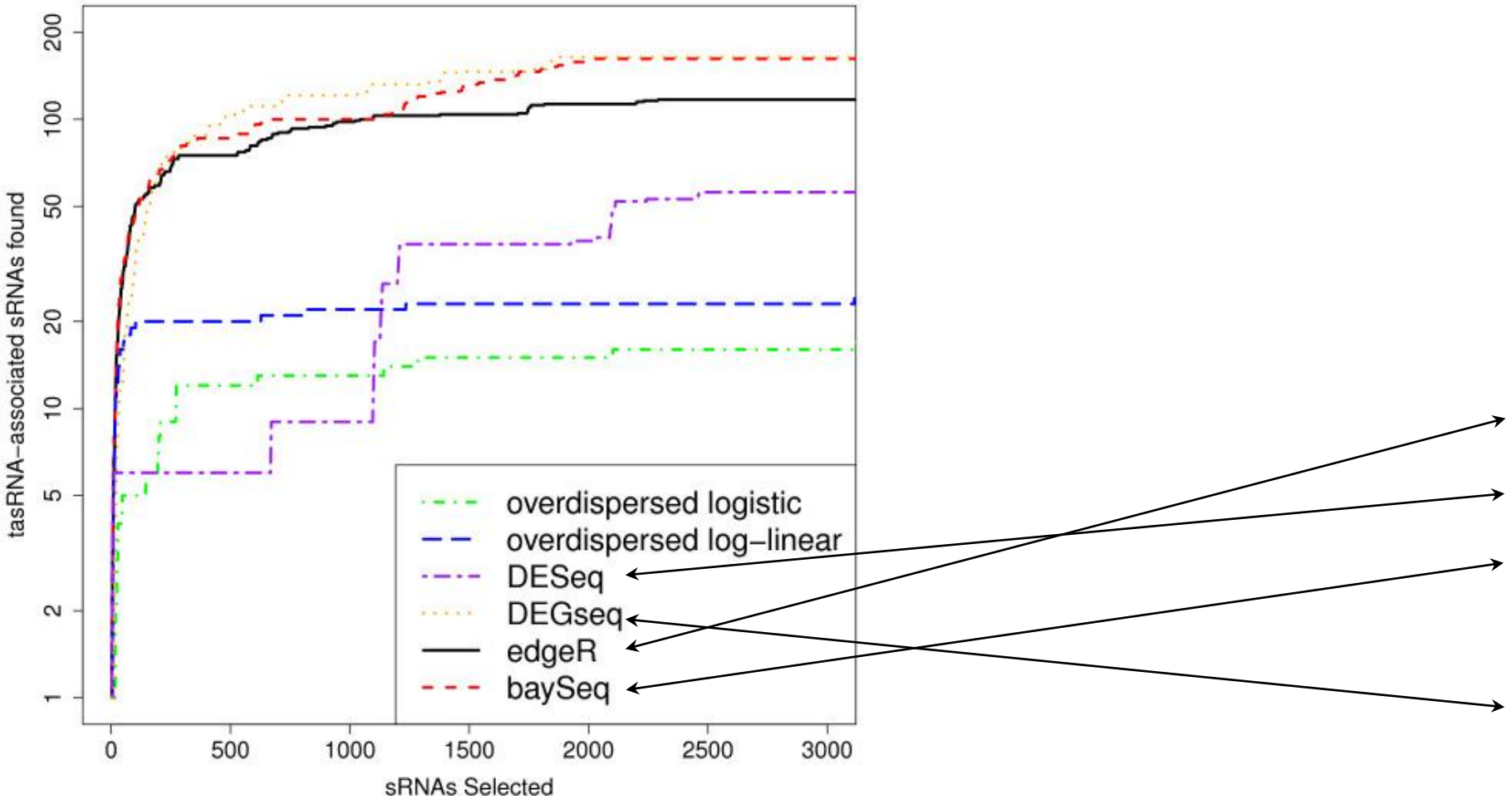
リアルデータ

- *baySeq*論文 のFig. 5のデータ
  - 公共データベース（GEO）: GSE16959
  - *Arabidopsis thaliana*のleaf samplesの20-24塩基のsmall RNAs (sRNAs)
  - two wild-type (WT) samples vs. two RDR6 (RNA-dependent RNA polymerase 6) knockout (KO) samples
  - RDR6はtasRNAs (trans-acting sRNAs)生成に必要であることが既知
  - 70,619 unique small RNA sequencesが*Arabidopsis thaliana*ゲノムにヒット
  - tasRNA lociのみに100%マッチし、発現がWT > KOとなる657 potentially true positivesを同定(P_A = 100%)

70,619行×4列のsRNA発現行列中に657個の真の発現変動sRNAsを含むデータ

解析結果 (ROC曲線の一部)

Fig. 5 in *baySeq* paper



まとめ

- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | について (last modified 2011/03/02)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | baySeq (Hardcastle 2010) (last modified 2011/09/14) NEW
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | DESeq (Anders 2010) (last modified 2011/02/10)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | Fisher's exact test (FET) by DEGseq (last modified 2011/06/09)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | Likelihood ratio test (LRT) by DEGseq (last modified 2011/06/09)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | Fold change (FC) by DEGseq (last modified 2011/06/09)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | MARS (Wang 2010) by DEGseq (last modified 2011/06/09)
- 解析 | NGS(RNA-seq) | 発現変動遺伝子 | 二群間 | edgeR (Robinson 2010) (last modified 2010/11/24)

- 性能評価結果
  - シミュレーション: *baySeq* > *DESeq* > *edgeR* > *NBPSeq*
  - あるリアルデータ: *baySeq* ≒ *edgeR* > *NBPSeq* > *DESeq*
- 正規化法はよりよいものを使うべし
  - RPM法, TMM法, 門田法など
  - オリジナルの手順よりも門田法と組み合わせるとよりよい結果に…
- 計算環境
  - Windows 7, Intel Xeon 3.33GHz (2 processor), 96GBメモリ

謝辞

共同研究者

西山智明　博士（金沢大学）

清水謙多郎　博士（東京大学）

グラント

- 若手研究(B)（H21-23年度）:「マイクロアレイ解析の再現性・感度・特異度を飛躍的に向上させるデータ解析手法の開発」(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)

解析データ提供

- Dr. Thomas J. Hardcastle（*baySeq*著者）

その他（苦労した点など）

- *baySeq*実行時のリサンプリング回数を当初1000回にしていた。→結果がころころ変わって困った。よくよくマニュアルを見ると推奨は10,000回だった…*baySeq*原著論文の精度はひょっとして…

- *DESeq*と*baySeq*はTMMや門田法で正規化した後のデータをどうやって入力データとすればいいのか…。最終的にこの二つは同じやり方でできた。

- *edgeR*もしばらく上記二つと同じやり方でやっていたが、最近（9月）そのやり方ではだめだということに気付いた…。

- 一つのシミュレーション条件を100回で30時間かかります。

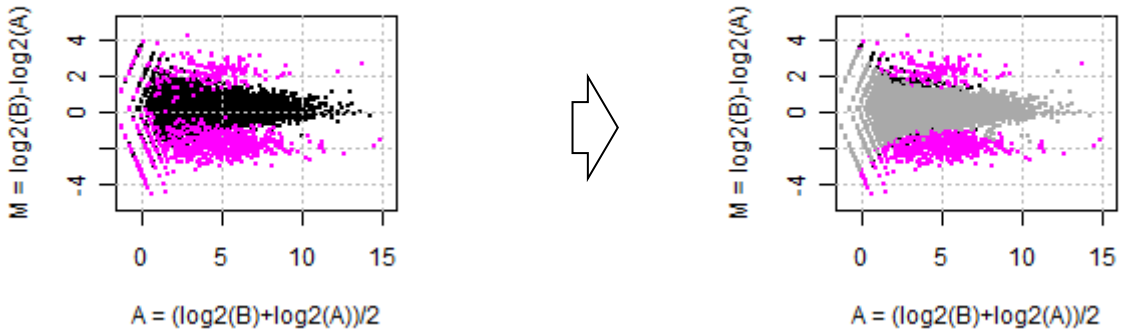
その他(実は一番重要かも…)

- 読み込ませる入力データ
  - ①:通常(生のリードカウント)
    - TMM正規化係数などは引数で与える
  - ②:TMM or 門田正規化後のデータ
    - プログラム内部で正規化されないように操作
  - ③:RPM正規化後のデータ
    - TMM正規化係数などは引数で与える

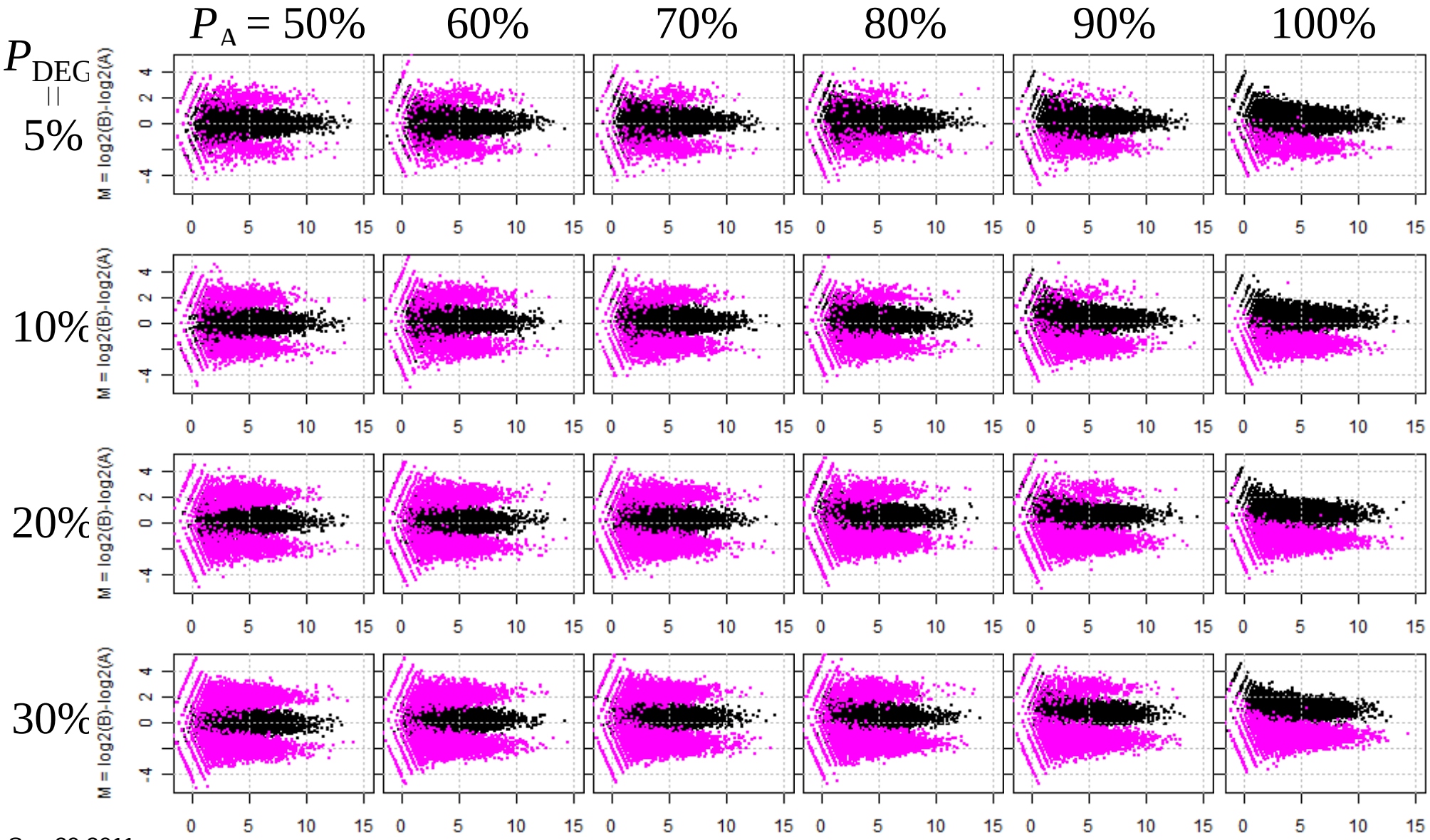
リアルデータ

シミュレーションデータ

- *NBPSeq*のデータを生データの（μの）経験分布として使用
  - 負の二項分布（平均=μ、分散=$\mu+\mu^2/size$；size=10とした）
  - 発現変動遺伝子（DEG）の倍率変化: 4
  - Number of common genes: 20,000
  - DEGの割合（P_{DEG}）: 5, 10, 20, 30%
  - DEG中に占めるsampleA > Bの割合（P_A）: 50, 60, 70, 80, 90, 100%



様々なシナリオ



様々なシナリオ