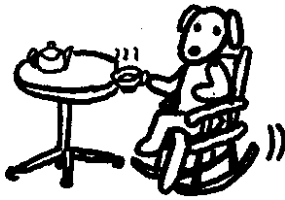
農業生物のトランスクリプトーム解析における情報処理

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

2009年ごろの私

- 次世代シーケンサー（NGS）解析についての認識
  - 単に短い塩基配列が沢山あるだけでしょ
  - 得られる配列データって、multi-FASTA形式のもので、単にそれをリファレンス配列にマッピングしてカウントするだけでしょ
  - それ以降の解析はマイクロアレイと同じなんじゃないのー



- 私について
  - **マイクロアレイ**を中心としたデータ解析手法の開発
  - 主に遺伝子発現行列の**数値データのみ**を取り扱ってきた
  - 配列解析系のスキルはほぼゼロで、用語がまるでわかっていない
  - 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活動の一環でsmallRNAのNGS（Illumina）解析をやりはじめた
  - 自分の研究テーマとして主体的にやり始めたのは2011年～

11.農学生命情報科学特論I

概要

次世代シーケンサーの普及により、以前は主にゲノム解析系で必要とされていた（塩基）配列解析のためのスキルがトランスクリプトーム解析においても要求される時代になりつつあります。様々な局面で応用可能な配列解析系のスキルアップを目指した実習を含む講義を行います。

Contents

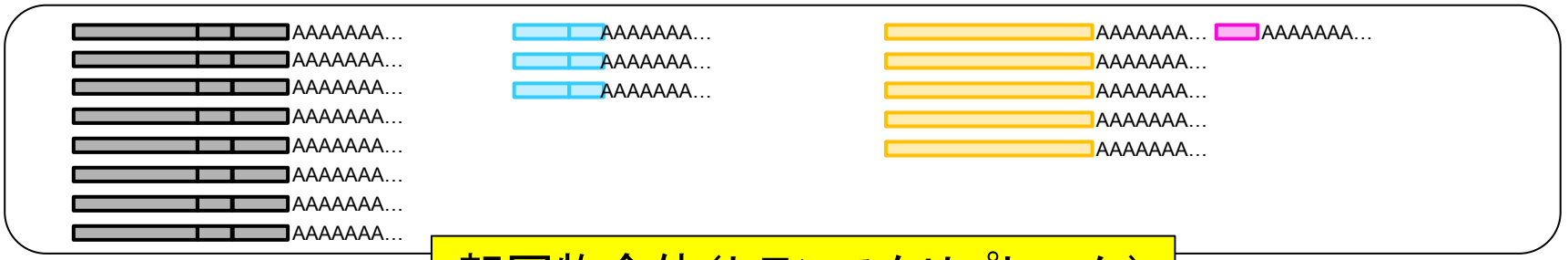
- トランスクリプトーム解析の概要
- RNA-Seq vs. マイクロアレイ
  - 長所・短所
  - 実データの比較
- RNA-Seqデータの正規化（の基礎）
  - マイクロアレイと異なる点（配列長補正が余分に必要）
  - 基本的な考え（RPKM）
  - RPKM正規化（正確にはRPM）の問題点
  - 門田正規化法

トランスクリプトームとは

- ある状態のあるサンプル（例：目）のあるゲノムの領域

遺伝子全体（ゲノム）

・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）



転写物全体（トランスクリプトーム）

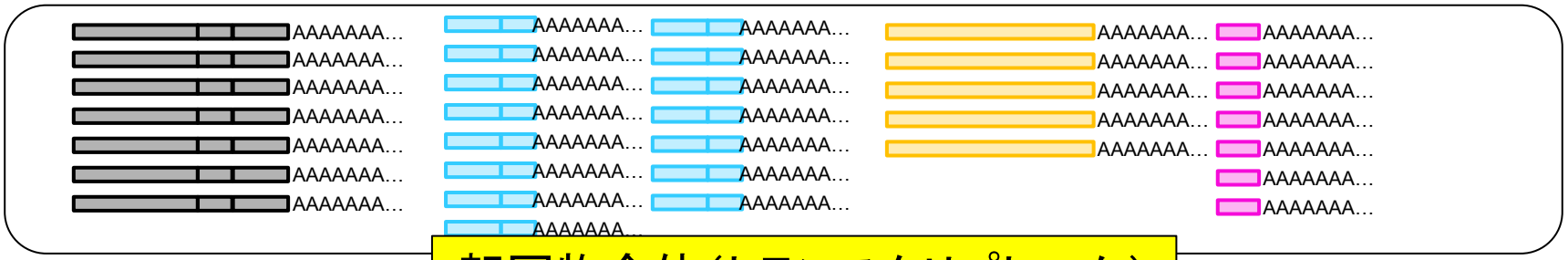
・遺伝子1は沢山転写されている（発現している）
・遺伝子4はごくわずかしか転写されてない
・…

トランスクリプトームとは

- ある状態のあるサンプル（例：目）のあるゲノムの領域

遺伝子全体（ゲノム）

・どの染色体上のどの領域にどの遺伝子があるかは調べる個体（例：ヒト）が同じなら不変（目だろうが心臓だろうが…）

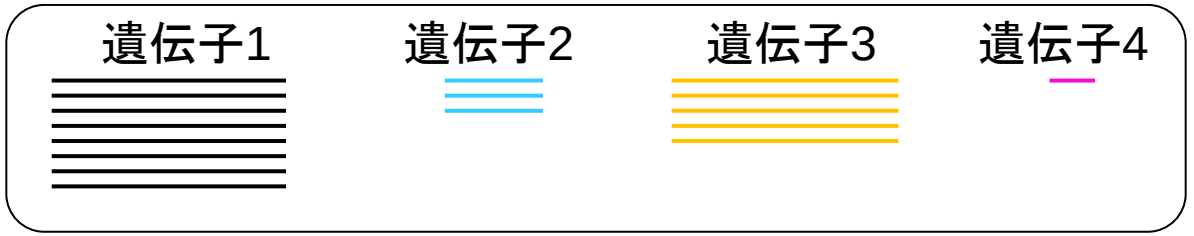


転写物全体（トランスクリプトーム）

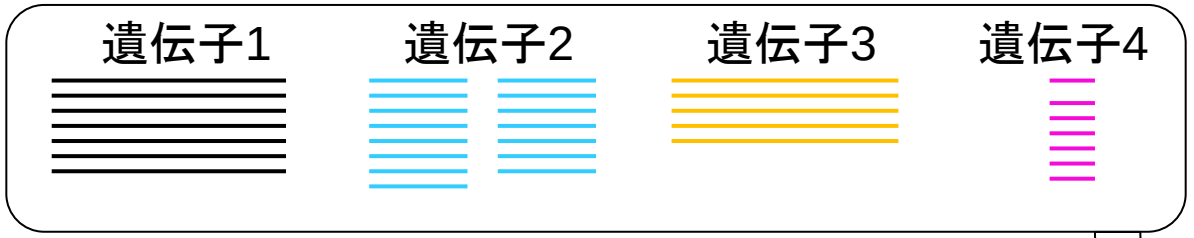
・遺伝子2は光刺激に応答して発現亢進
・遺伝子4も光刺激に応答して発現亢進

トランスクリプトーム情報を得る手段

- 光刺激前（T1）の目のトランスクリプトーム



- 光刺激後（T2）の目のトランスクリプトーム



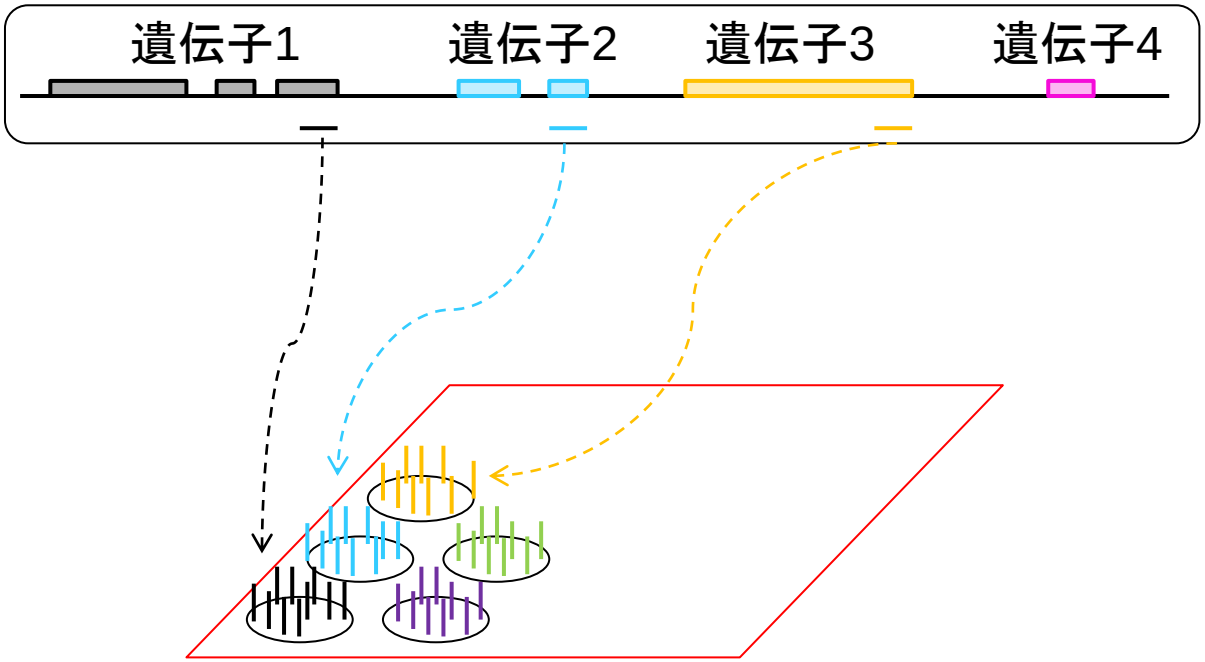
これがいわゆる
「遺伝子発現行列」

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

- マイクロアレイ
- RNA-Seq
- SAGE
- …

トランスクリプトーム取得（マイクロアレイ）

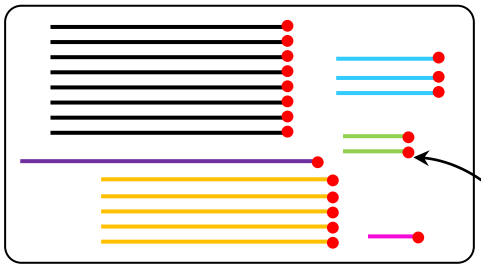
- よく研究されている生き物は多数の遺伝子（の配列情報）がわかっている



わかっている遺伝子（の配列の相補鎖）を搭載した”チップ”

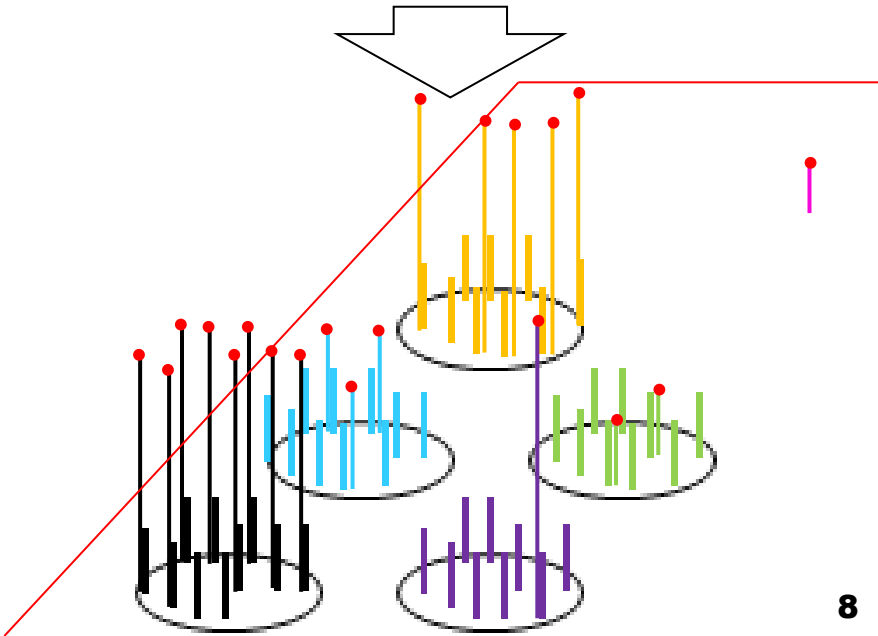
・メーカーによって搭載されている遺伝子の種類が異なる
→搭載されていない遺伝子（未知遺伝子含む、例：遺伝子4）の発現情報は測定不可…

光刺激前（T1）の目のトランスクリプトーム



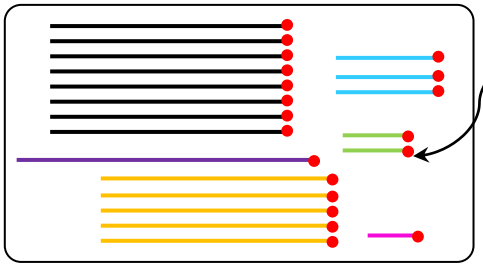
蛍光標識

ハイブリダイゼーション（二本鎖形成）

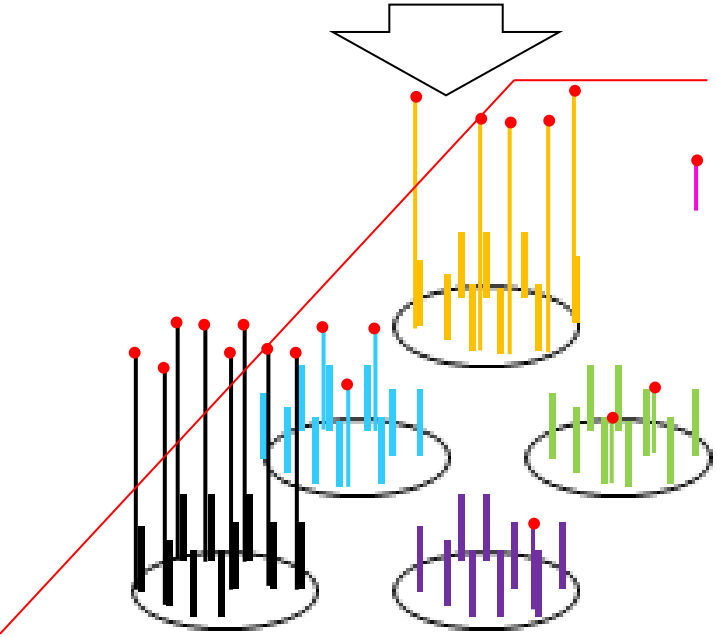


マイクロアレイデータ → 遺伝子発現行列

- 光刺激前（T1）の目のトランスクリプトーム

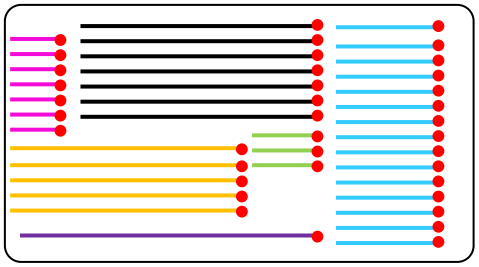


ハイブリダイゼーション
（二本鎖形成）



専用の検出器で各遺伝子に対応する領域の蛍光シグナル強度を測定

光刺激後（T2）の目の
トランスクリプトーム



ハイブリダイゼーション
と
シグナル検出

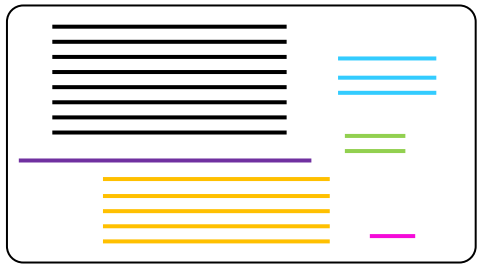
	T1	T2
遺伝子1	8	7
遺伝子2	3	15
遺伝子3	5	5
遺伝子4	?	?
遺伝子5	…	…
…	…	…

正規化

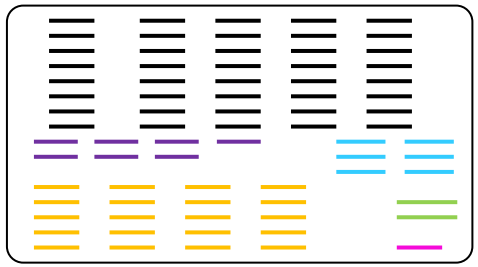
トランスクリプトーム取得（RNA-Seq）

- 次世代シーケンサー（Illumina社の場合）

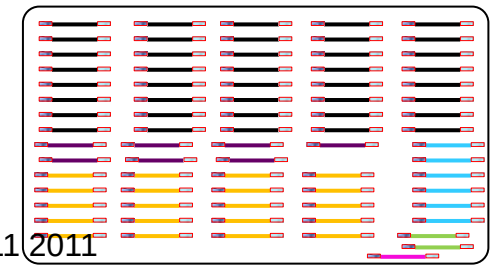
光刺激前（T1）の目のトランスクリプトーム



数百塩基程度
に断片化



二種類のアダプター
配列を両末端に付加

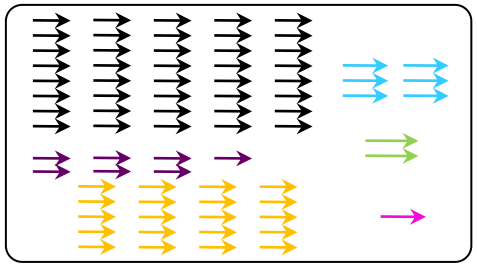


配列決定

・ペアードエンド法
断片配列の両末端が数百塩基以内の対の二種類の配列が得られる

・シングルエンド法

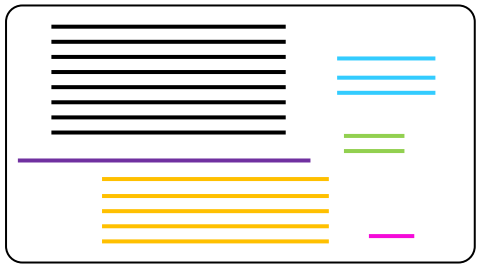
シングルエンド法
の場合



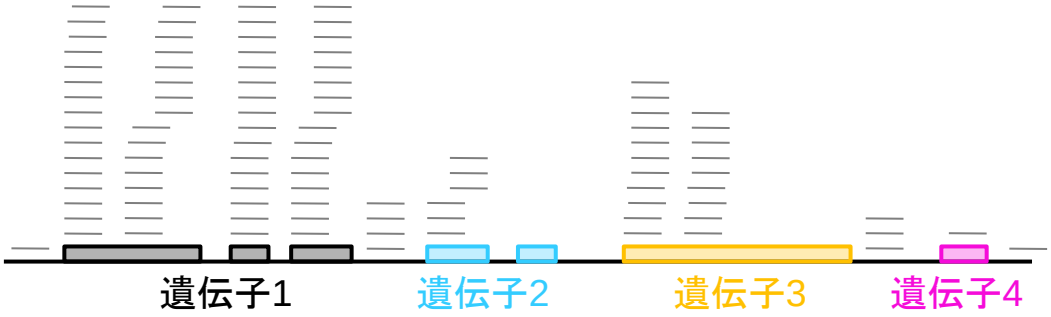
RNA-Seqデータ → 遺伝子発現行列

- RNA-seq

光刺激前（T1）の目のトランスクリプトーム



ゲノム配列にマッピング



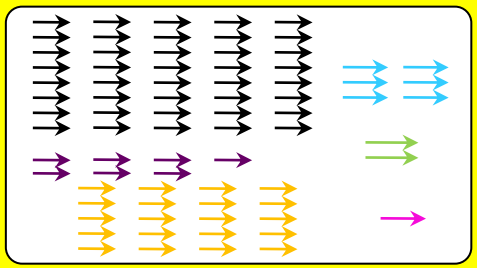
定量化(例：生のリード数をカウント)

	T1
遺伝子1	40
遺伝子2	6
遺伝子3	20
遺伝子4	1
遺伝子5	…
…	…

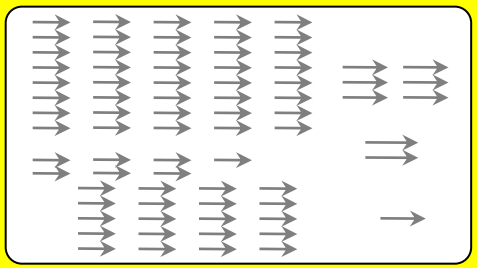
正規化

T1
8
3
5
1
…
…

－イメージ－
50-125塩基程度からなる配列が沢山ある



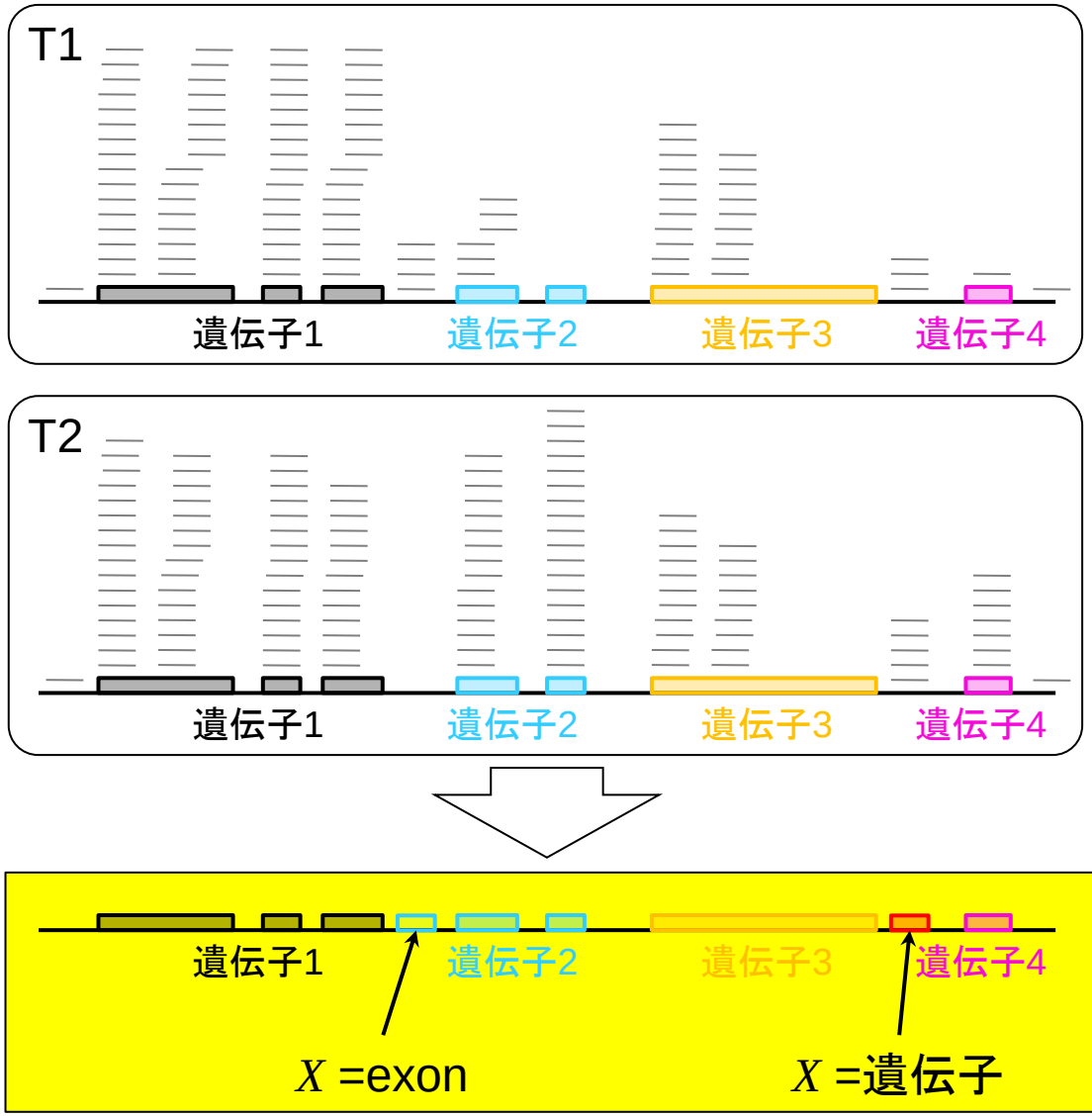
－実際－
数百万個の配列があり、どの遺伝子に対応するか不明



（短い）配列を読んだものという意味で（ショート）リードなどと呼ばれる

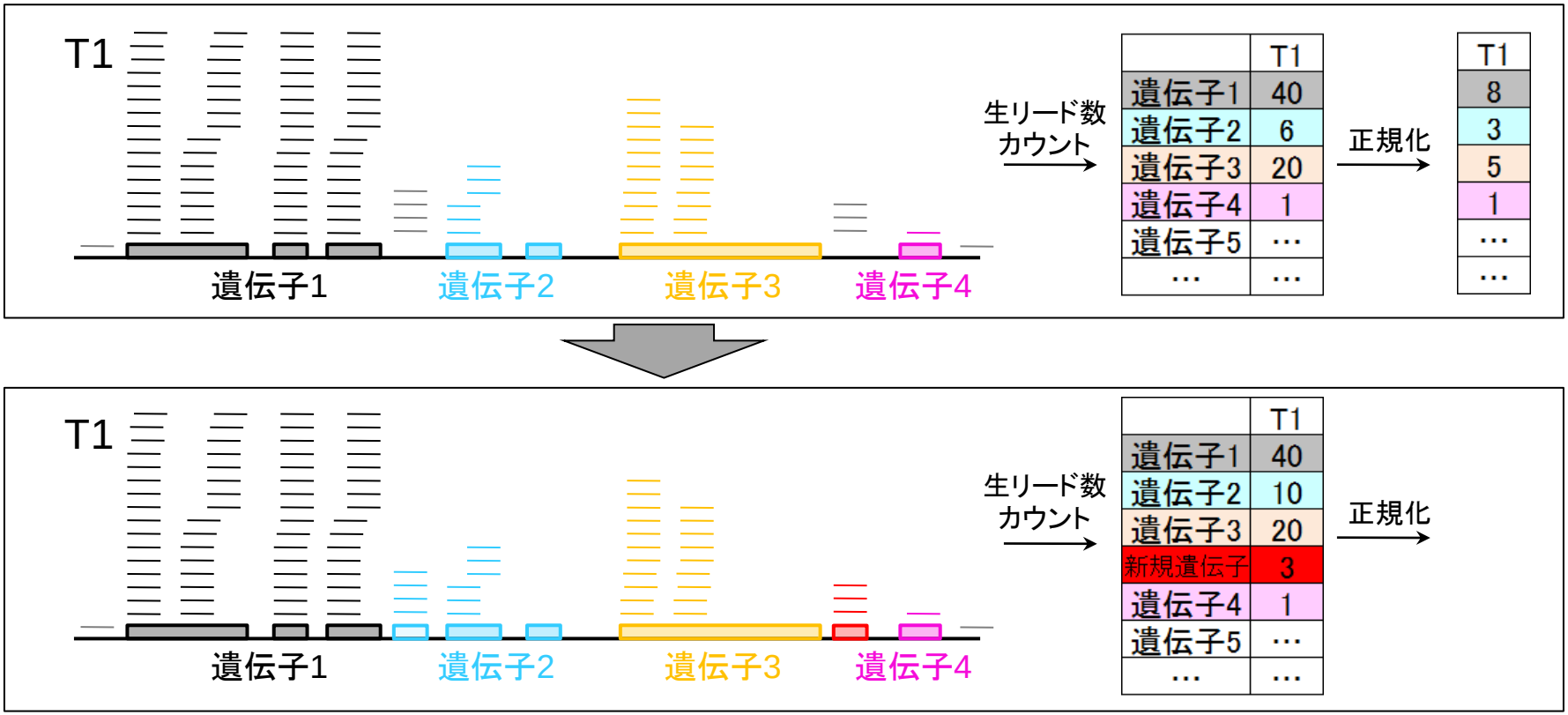
RNA-Seqの長所

- 新規Xの同定
  - X =exon, 遺伝子, …



RNA-Seqの長所

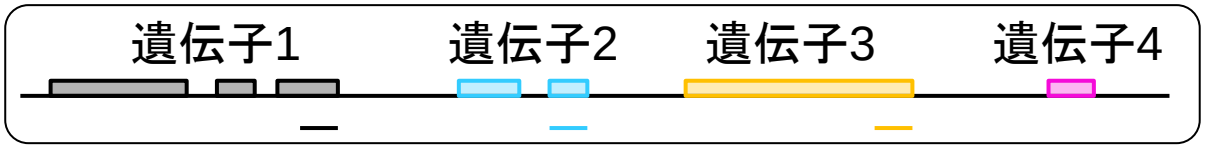
- 新規Xの同定



・“トランスクリプトーム（転写物の全体像）”の理解への一番の近道
・よりよい**遺伝子発現行列**を得るための基礎情報充実に貢献

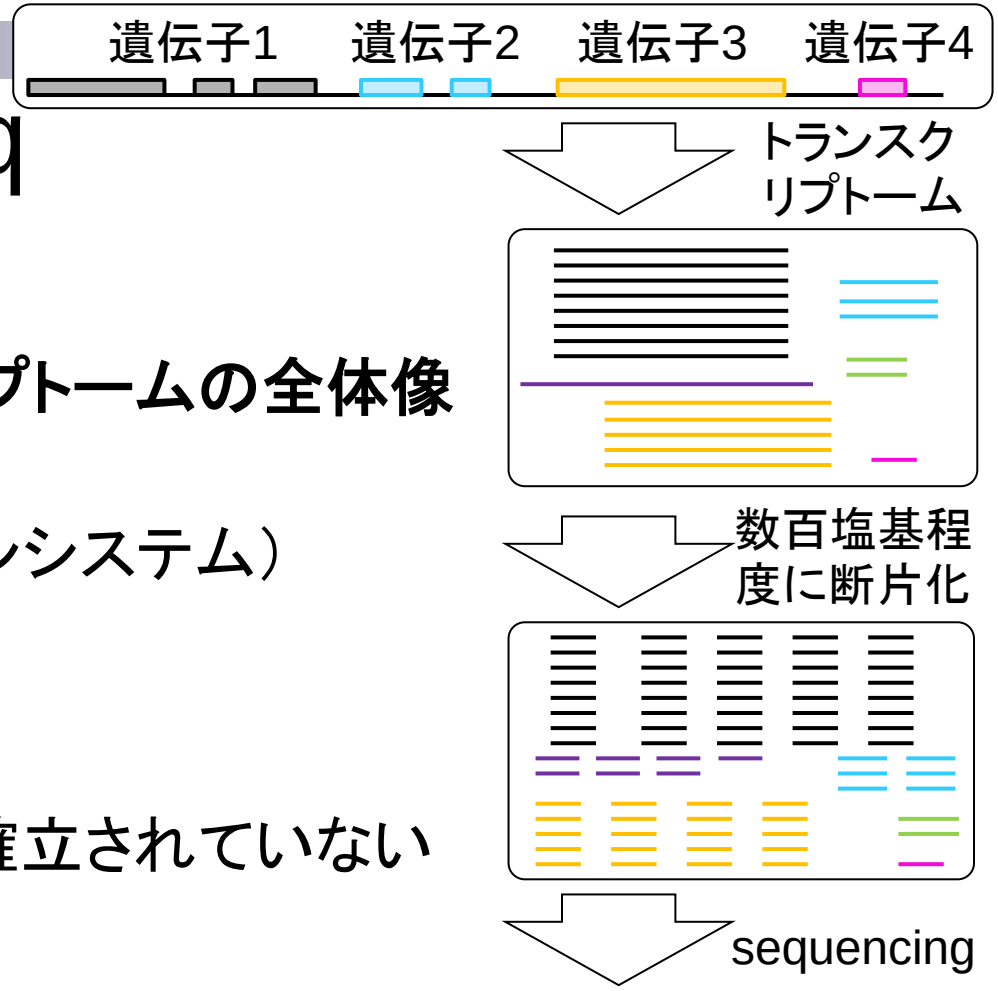
長所・短所：（発現解析用）マイクロアレイ

- 長所
  - すでに診断用マイクロアレイが市販されているなど**長年の実績**
  - お手軽、各種データ解析ツールが豊富
- 短所
  - （プローブ搭載のために）解析対象の塩基配列情報を予め知っておく必要がある。（クローズドシステム）
  - プローブが搭載されていない遺伝子の発現レベルは測定不可能（未知遺伝子も当然対象外）



- 主なユーザー
  - 主な解析対象が（アノテーション情報が豊富な）モデル生物で、既知遺伝子のみでいい、という研究者

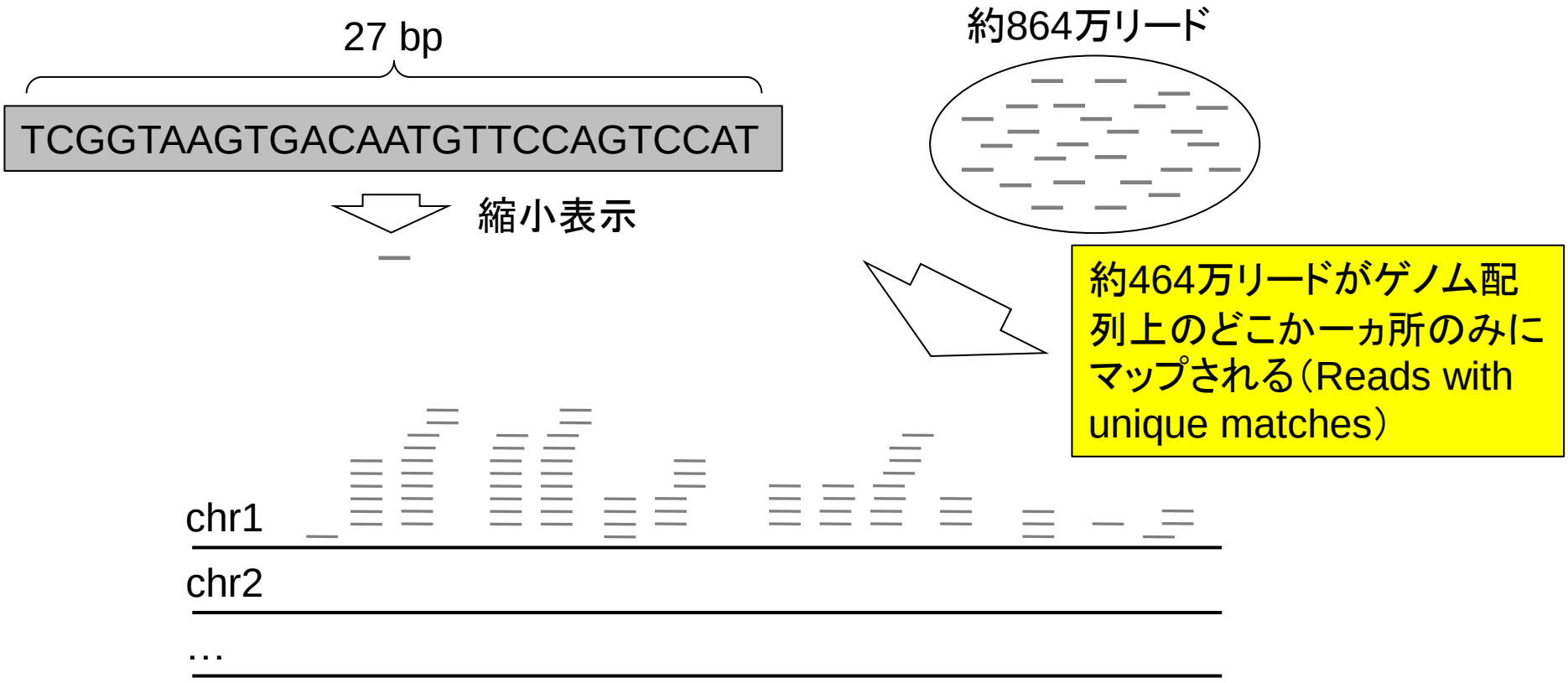
長所・短所：RNA-Seq



- 長所
  - （未知遺伝子を含む）**トランスクリプトームの全体像を理解することが原理的に可能**
  - 事前情報を必要としない（オープンシステム）
  - ダイナミックレンジが広い
- 短所
  - データ解析が大変、解析手法が確立されていない
- 主なユーザー
  - 無制限（モデル生物・非モデル生物を問わない）
  - （お金持ち…）

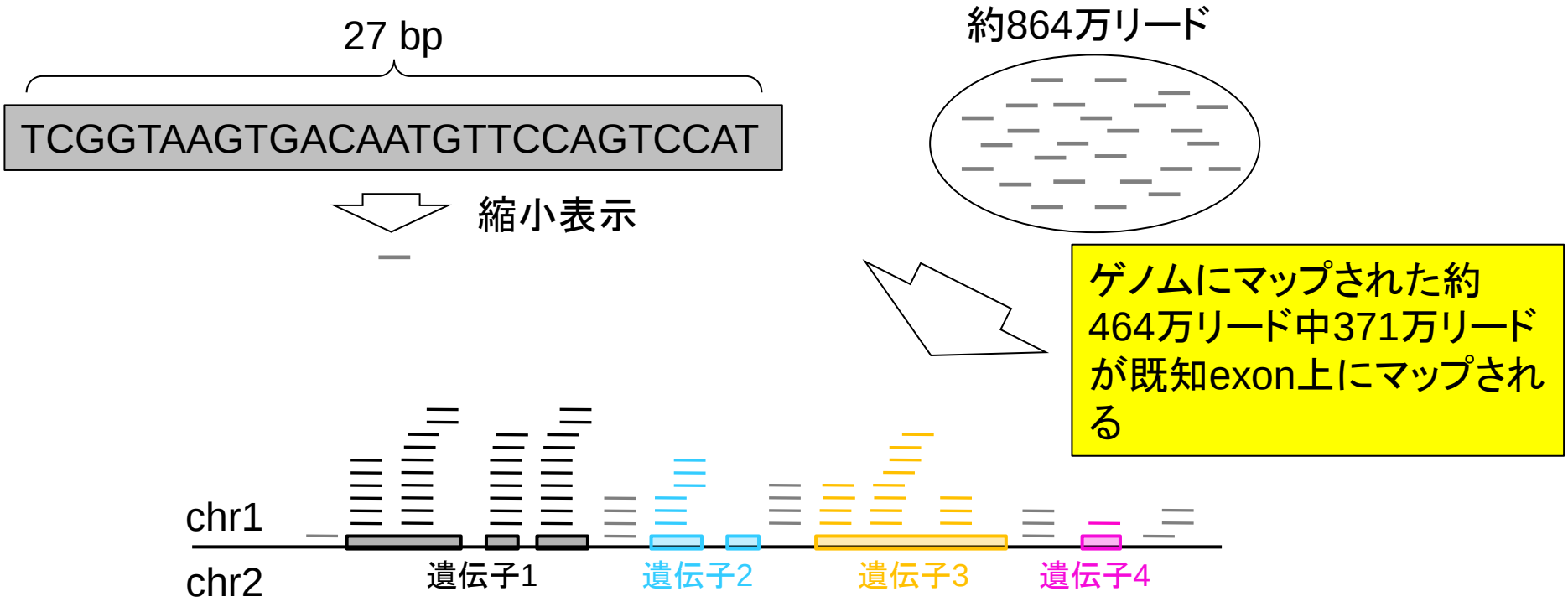
実データの比較（RNA-Seq vs. マイクロアレイ）

- Human embryonic kidney (HEK) 293T cells（とB cells）
  - **マイクロアレイ**：Illumina HumanRef8 V2.0 BeadChips
  - **RNA-Seq**: Illumina 1G Genome Analyzer



実データの比較（RNA-Seq vs. マイクロアレイ）

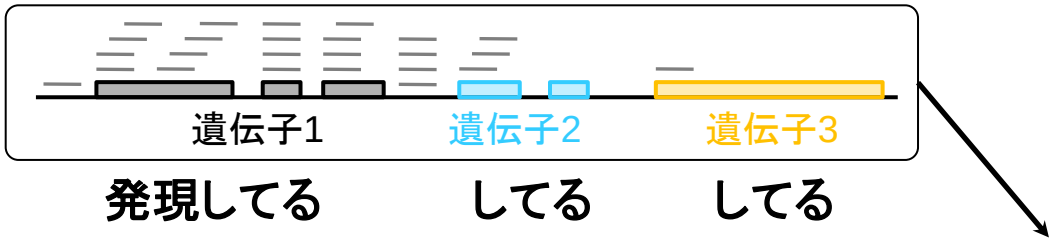
- Human embryonic kidney (HEK) 293T cells（とB cells）
  - **マイクロアレイ**：Illumina HumanRef8 V2.0 BeadChips
  - **RNA-Seq**: Illumina 1G Genome Analyzer



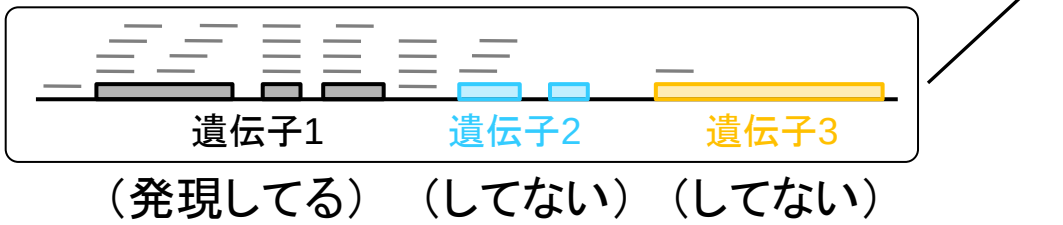
・既知エクソン領域以外にマップされたものは新規exonの可能性！
・大抵のマイクロアレイとの比較はアレイ上に搭載されている既知遺伝子についてのみ!

実データの比較（RNA-Seq vs. マイクロアレイ）

- マイクロアレイ上に搭載されている13,118遺伝子について、「発現している」とされた遺伝子数の比較
  - 閾値緩め（≧1 read）の場合



  - 閾値厳しめ（≧5 read）の場合



RNA-seqでのみ発現している遺伝子数 >> マイクロアレイでのみ

実データの比較（RNA-Seq vs. マイクロアレイ）

- 「HEK cells versus B cells」のlog ratio分布の比較

  7,043 genes

発現している：○, 発現してない：×

	マイクロアレイ			RNA-Seq			
	B cells	HEK cells	log比計算	B cells	HEK cells	log比計算	Plot
gene1	○	○	○	○	○	○	○
gene2	○	○	○	○	×	×	×
gene3	×	○	×	○	○	○	×
gene4	○	○	○	○	○	○	○
gene5	×	○	×	×	○	×	×
gene6	○	○	○	○	○	○	○
gene7	○	○	○	○	○	○	○
gene8	×	×	×	×	×	×	×
gene9	○	○	○	○	○	○	○
gene10	×	×	×	○	○	○	×
gene11	○	○	○	○	○	○	○
…							

全体として高発現側の遺伝子群の発現レベルは似ている

他の比較結果（RNA-Seq vs. マイクロアレイ）

- 発現量レベルの比較
  - LiverサンプルのRNA-Seqデータ vs. マイクロアレイデータ

RPKM?

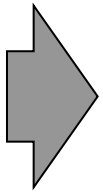
マイクロアレイ

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」と仮定
  - チップ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（だろう）

	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

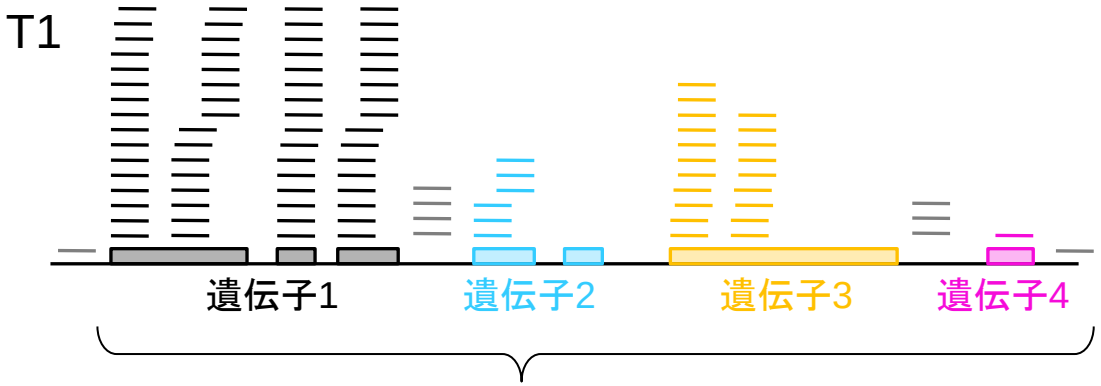


	sample1	sample2
gene1	14.2	15.3
gene2	8.7	8.8
gene3	10.9	10.5
gene4	14.7	14.1
gene5	7.6	8.3
gene6	11.4	11.0
gene7	8.4	8.6
gene8	8.3	8.4
gene9	9.0	8.0
gene10	6.9	7.2
総和	100.0	100.0

背景：サンプル（or chip）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）

- 「各サンプルからsequenceされた総リード数は一定」と仮定



	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

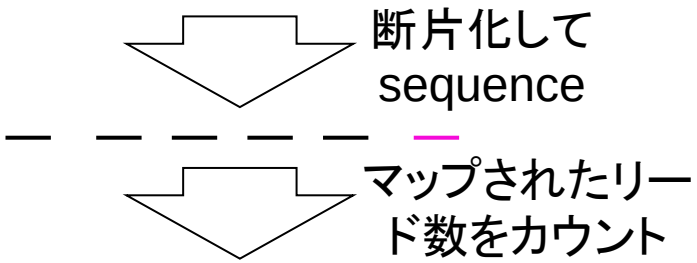
RPM正規化

	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**

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



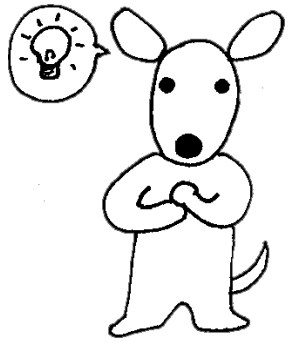
mRNA	リード数
AAAAAAA...	5
AAAAAAA...	1

一つのサンプル内での異なる遺伝子間の発現レベルの高低を(配列長を考慮せずに)比較することはできない

配列長の補正

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

- 前提条件：配列長が既知
- 補正の基本戦略：配列長で割る
  - 「1 / 配列長」を掛ける場合
    → 「塩基あたりの平均のリード数」を計算しているのと等価
  - 「1000 / 配列長」を掛ける場合
    → 「その遺伝子の配列長が1000bpだったときのリード数」と等価
    Reads Per Kilobase of exon



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」

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

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

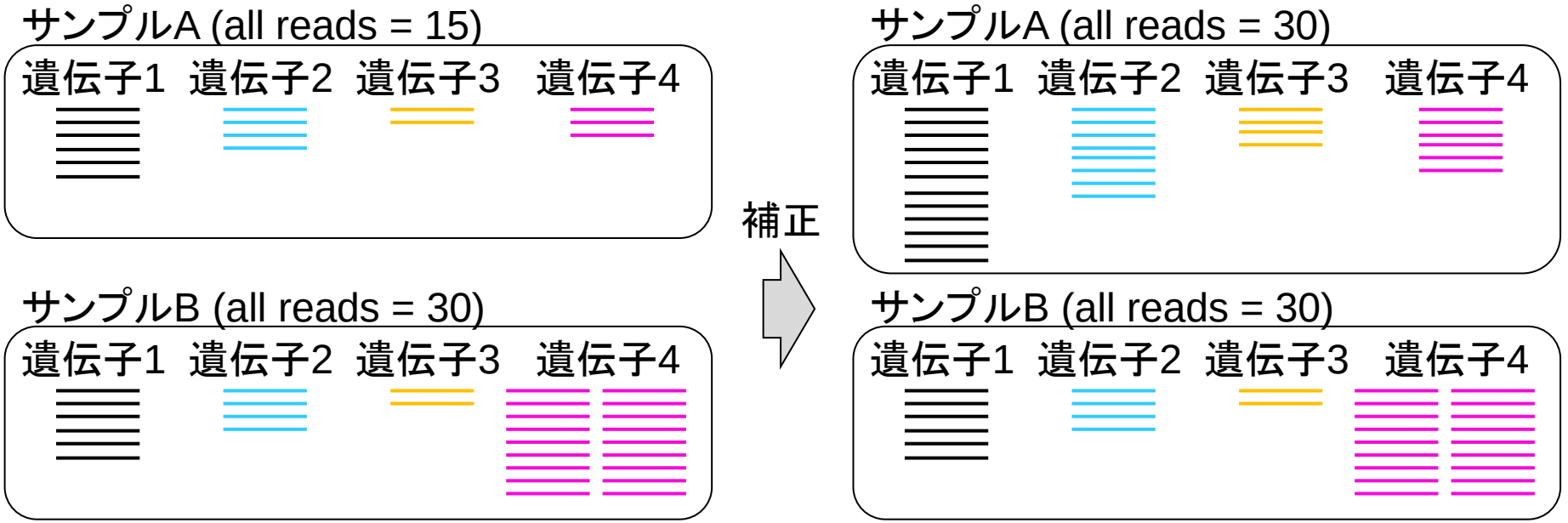
- RPKM正規化（RNA-Seq特有）
  - Reads per kilobase of exon per million mapped reads
  - 遺伝子の配列長が長いほど配列決定(sequence)される確率が上昇
    →各遺伝子の配列長を「1000塩基（one kilobase）の長さだった場合」に補正

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

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)

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

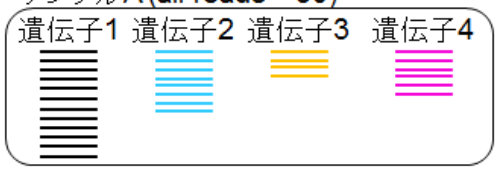


補正後の解析結果：Aで高発現が3個, Bで高発現が1個

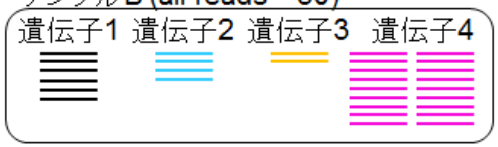
TMM正規化法

RPM補正後のデータ

サンプルA (all reads = 30)



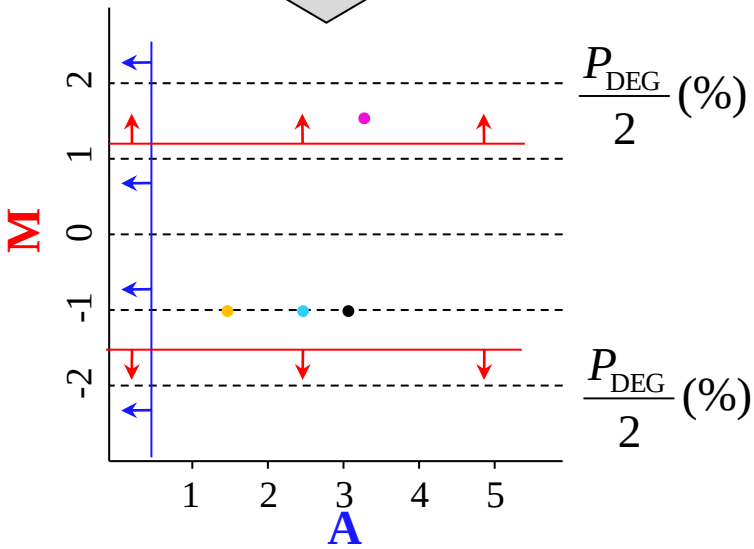
サンプルB (all reads = 30)



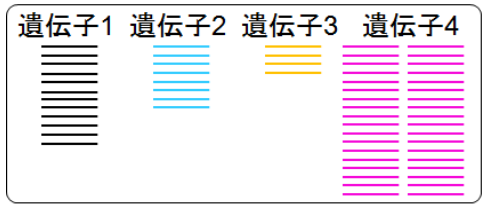
	A	B
遺伝子1	12	6
遺伝子2	8	4
遺伝子3	4	2
遺伝子4	6	18

$\log_2$(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



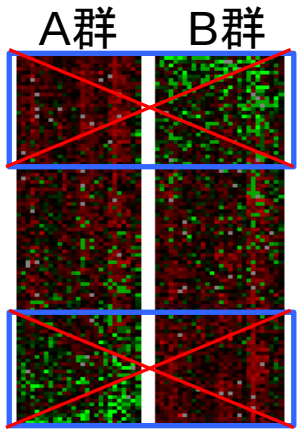
	$\log_2$(A)	$\log_2$(B) - TMM
遺伝子1	3.58	3.58
遺伝子2	3.00	3.00
遺伝子3	2.00	2.00
遺伝子4	2.58	5.17

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

利用可能なRパッケージたち

- *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
  - 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)

興味１：どれが高精度か？
興味2:個別の正規化法をほかのRパッケージ中の方法とどう組み合わせるのか？

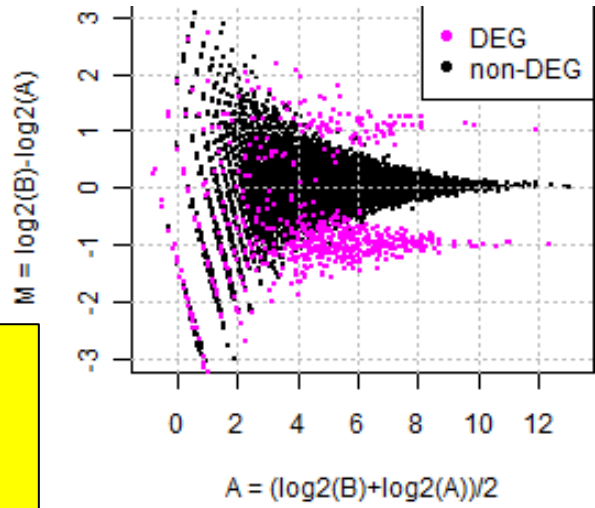
TMM法と門田（TbT）法

- 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（正規化法だが、内部的に発現変動検出を行う）
  - Step1：既存手法を適用して正規化
  - Step2：既存の発現変動遺伝子検出法を用いて、発現変動遺伝子候補を同定（**実データのP_A 値をうまく反映**）
  - Step3：得られた発現変動遺伝子以外のデータのみを用いて、再び既存手法を適用して正規化

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%

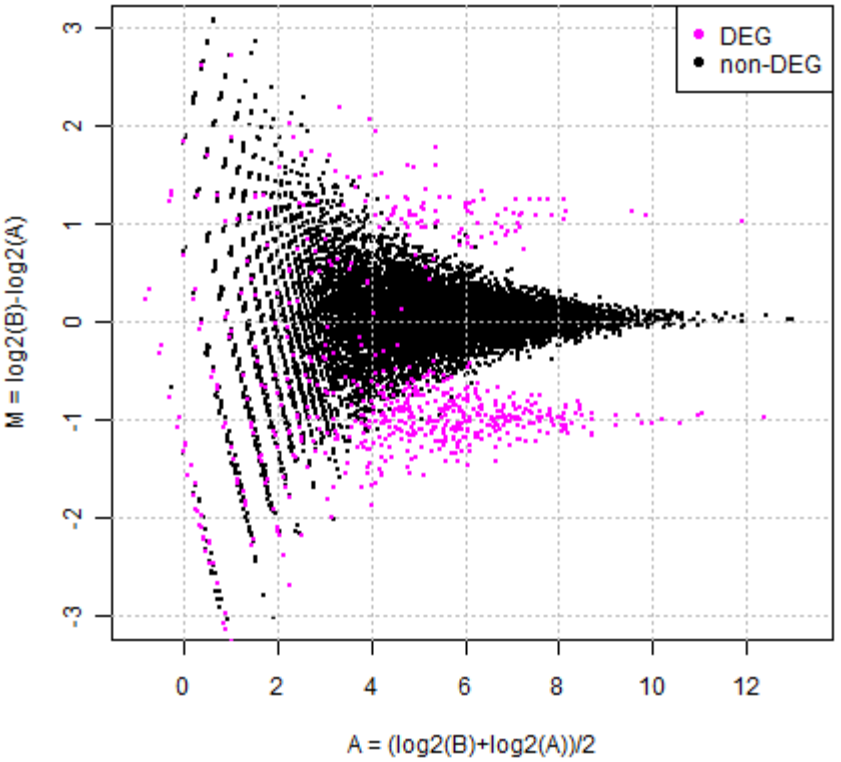


全遺伝子数が20,000個
そのうち5%（1,000個）がDEG（P_{DEG} = 5%）
80%（800個）がsampleAで高発現（P_A = 80%）

門田法のイメージ

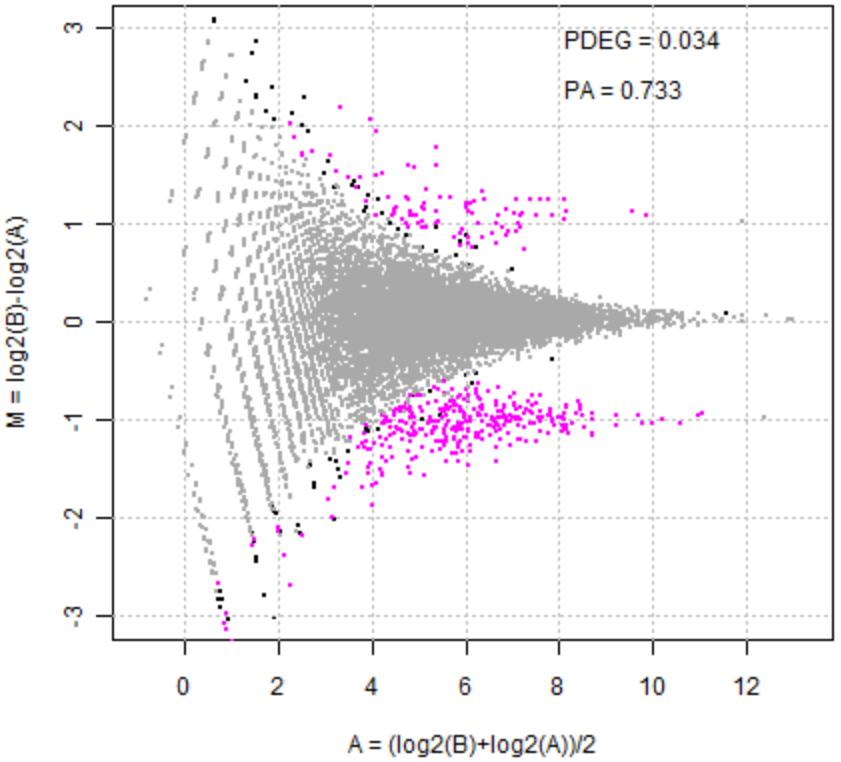
真実

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



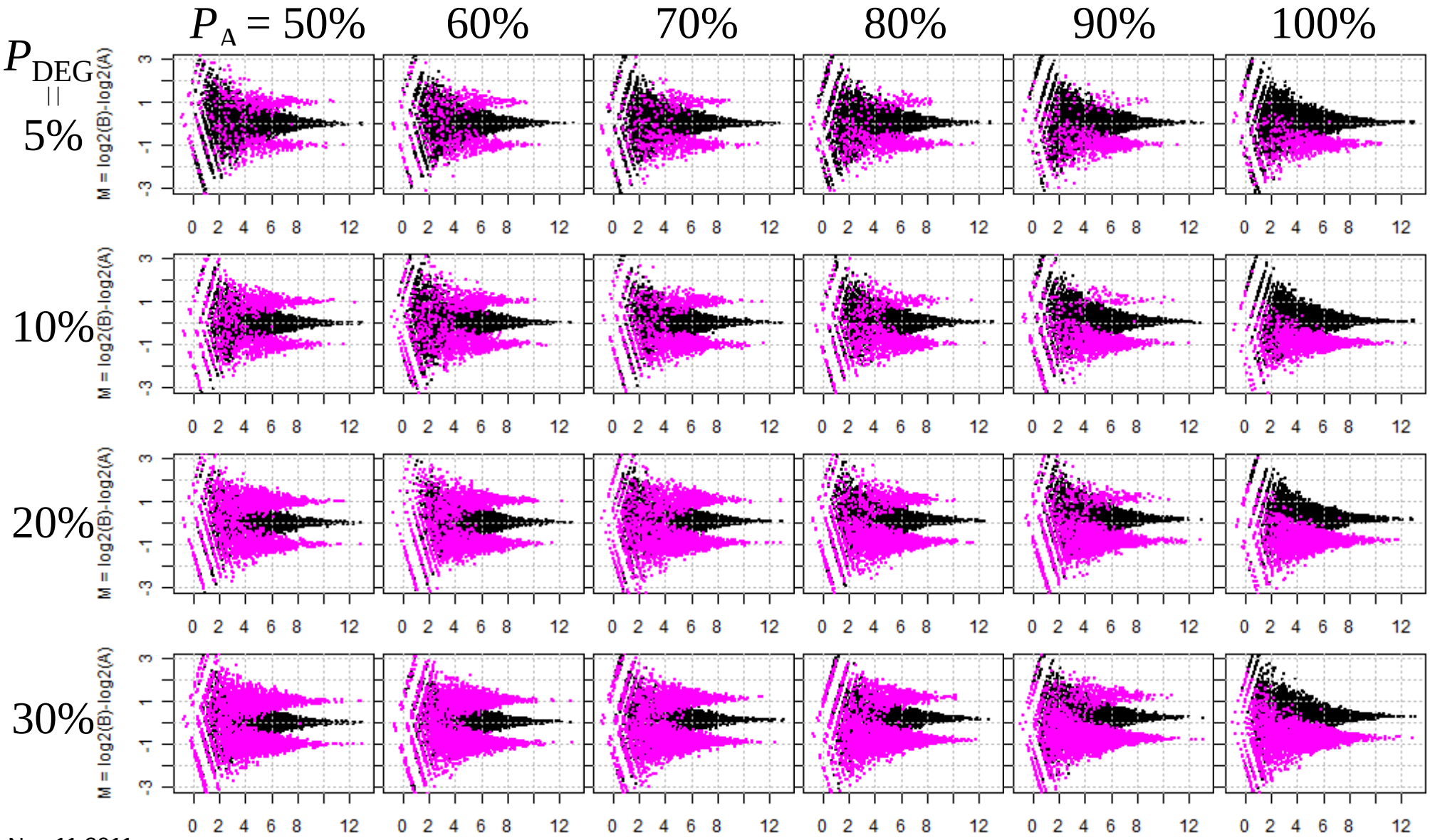
推定値

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

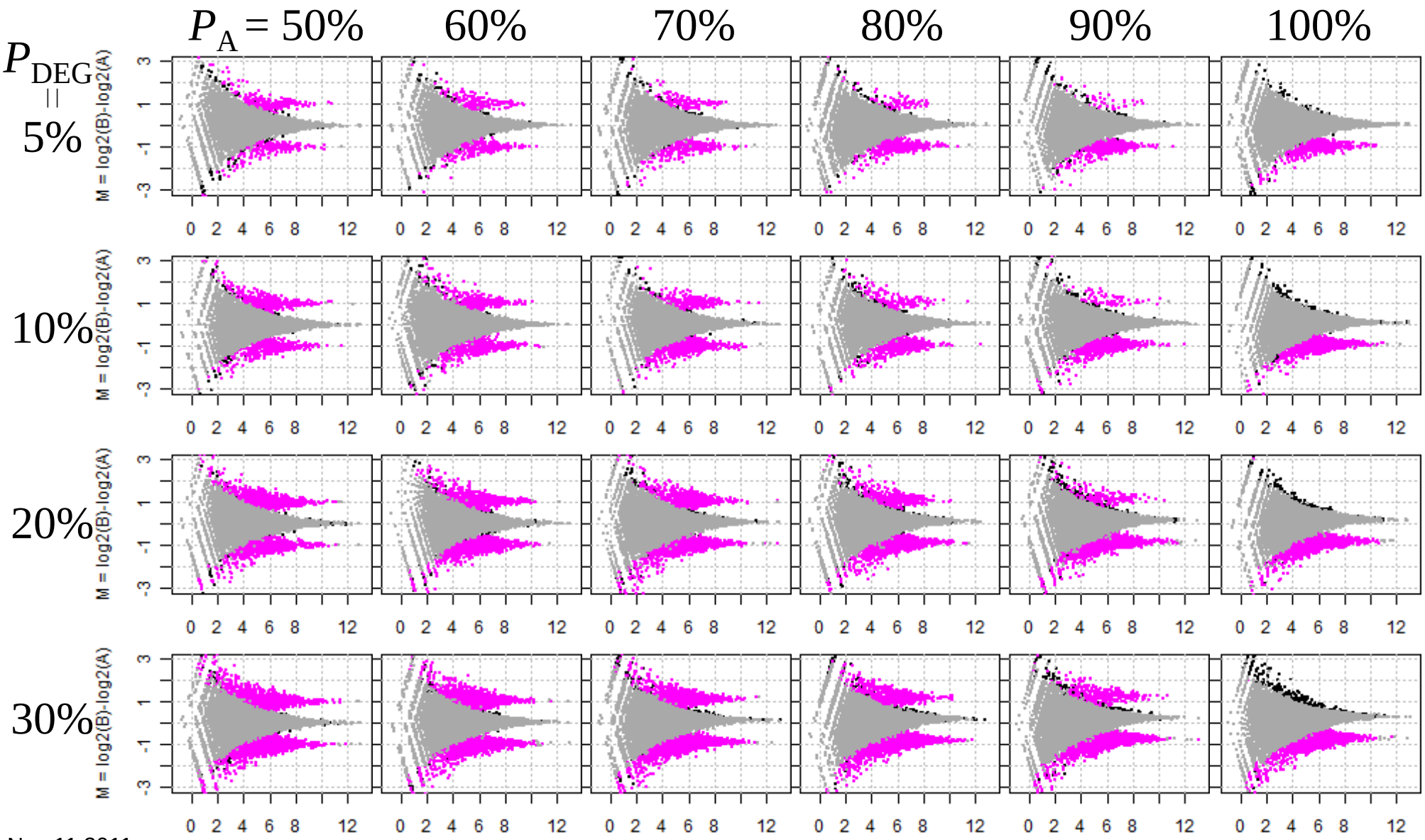


DEGと判定されなかった（灰色部分に相当）データのみを用いて正規化係数を決定

様々なシナリオ



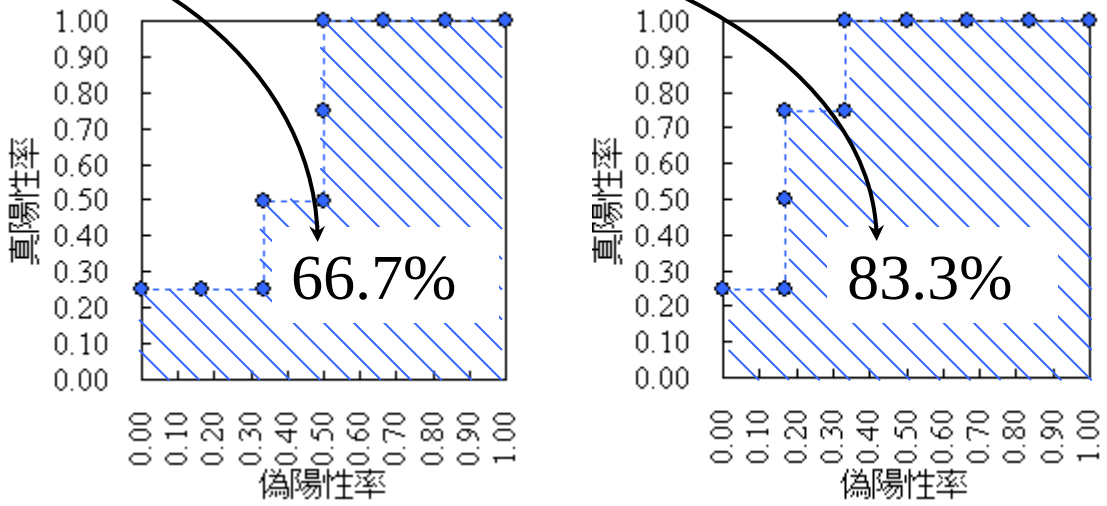
門田法のイメージ



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)

比較に用いた組合せは計８通り

- 既存の4つのRパッケージを評価
  - デフォルトの手順通りのやった場合
    - *edgeR*/default
    - *DESeq*/default
    - *baySeq*/default
    - *NBPSeq*/default
  - 門田正規化法（TbT）を組み込んだ場合
    - *edgeR*/TbT
    - *DESeq*/TbT
    - *baySeq*/TbT
    - *NBPSeq*/TbT

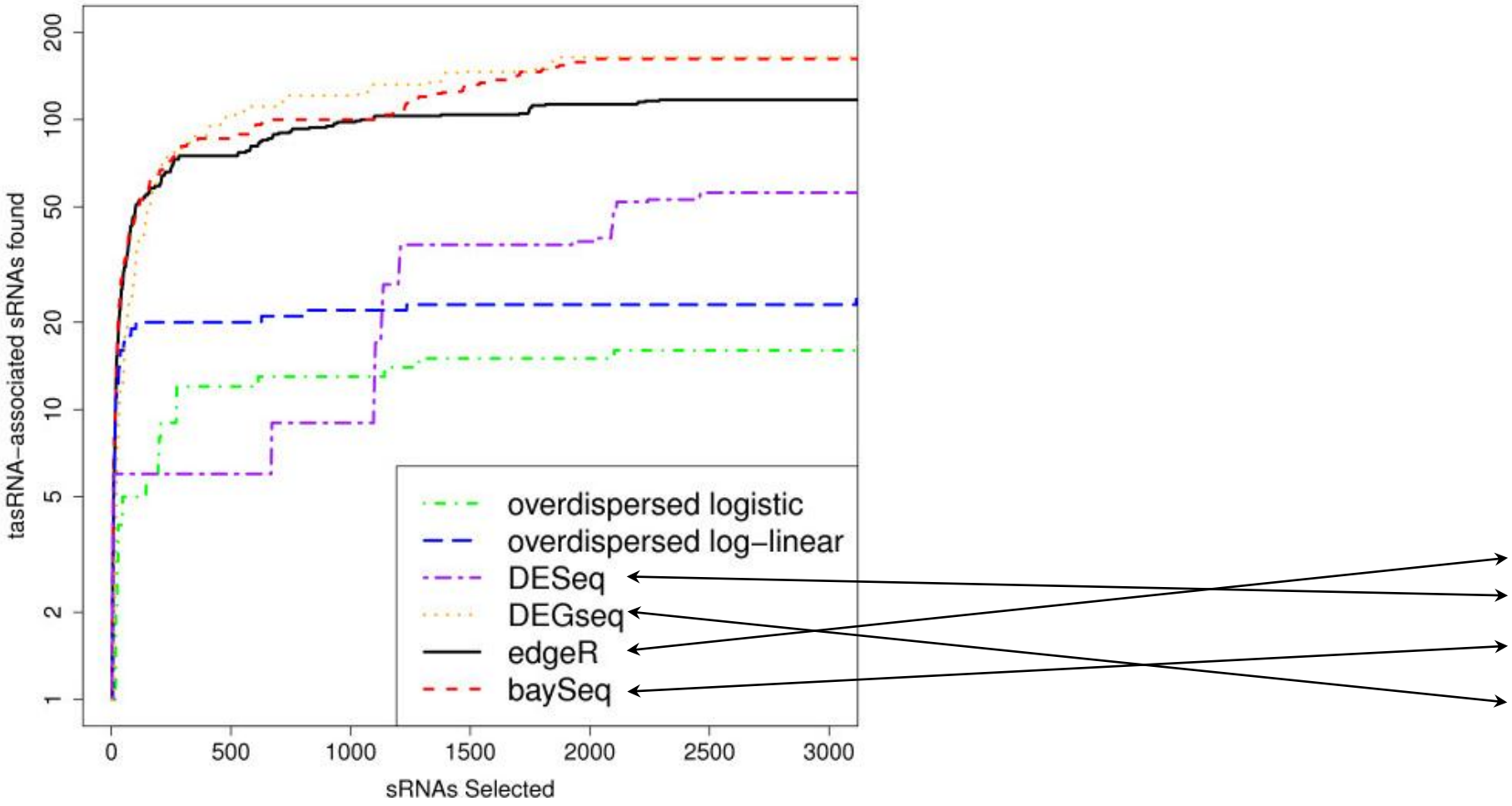
リアルデータ

- *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

- 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…

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）