マイクロアレイ解析の話：発現変動遺伝子検出あたりを中心に

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人材養成ユニット
門田幸二（かどた　こうじ）

2009/02/26

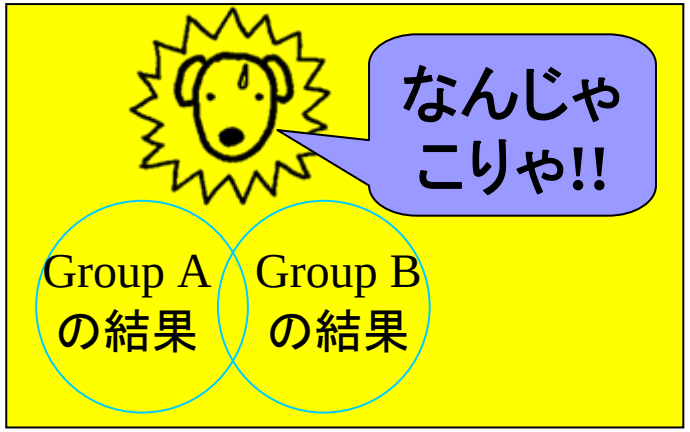
実用化にむけた取り組み

- 国外
  - MicroArray Quality Control (MAQC)プロジェクト (2005/2-2006/9)
  - External RNA Control (ERC) Consortium
  - MAQC-II (2006/9-2009/3)
- 国内
  - バイオチップコンソーシアム(JMAC)
    - 2007年10月に設立
    - バイオ産業分野の業界団体

解決すべき課題

- 再現性は本当にあるのか？
  - プラットフォーム間（メーカーの違い）
  - プラットフォーム内（実験場所の違い）
- どの解析手法がいいか？
  - 前処理（正規化）法：MAS5, RMA, MBEI, …
  - 発現変動遺伝子検出法
    - 組織特異的遺伝子：Dixon test, ROKU, …
    - 二群間比較（癌 vs. 正常）：*t*-test, SAM, …
- 重視すべき評価基準は？
  - 「感度・特異度」重視派
  - 「再現性（MAQCプロジェクト提唱）」重視派

「感度・特異度」と「再現性」は両立しない？！



話の内容

- 組織特異的遺伝子検出法
  - Dixon test, ROKU, …
- Affymetrix GeneChipデータ解析
  - PM-MM戦略
  - 様々な前処理（正規化）法
  - 様々な二群間での発現変動遺伝子ランキング法
  - 重視すべき評価基準は？
    - 感度・特異度
    - 再現性
  - 推奨ガイドライン（MAQC vs. 門田）

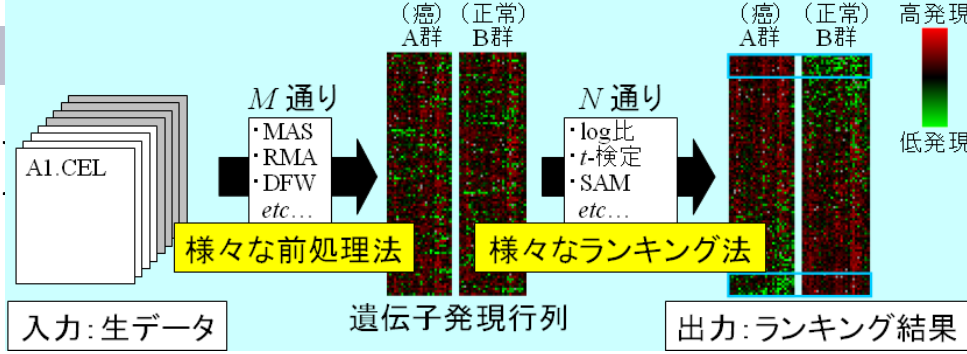
結論(組織特異的遺伝子検出法)

方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
Dixon test	×	×	?	○	-
Pattern matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
Entropy (*H*)	○	×	×	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
AIC	○	○	○	×	○
Sprent's method	○	○	○	×	×
ROKU (AIC+a modified *H*)	○	○	○	○	○

ROKUがおすすめ

Kadota K, Ye J, Nakai Y, Terada T, and Shimizu K, *BMC Bioinformatics*, **7**: 294, 2006
Kadota K, Konishi T, and Shimizu K, *Gene Regulation Systems Biol.*, **1**: 9-15, 2007

結論（Affymetrix Ge



- 「感度・特異度」が高い方法（組合せが重要である！）

前処理法	MAS5	multi-mgMOS	RMA	VSN	GCRMA	MBEI	PLIER	FARMS	DFW
ランキング法	WAD	WAD	RP	RP	RP	RP	RP	RP	RP

WAD: Weighted Average Difference
RP: Rank Products

Fold Changeに基づく方法

↕

従来：*t*-統計量に基づく方法

- （発現変動遺伝子リストの）「再現性」が高い方法
  - （前処理法によらず）WAD ↔ 従来： Average Difference (AD)法

Kadota K, Nakai Y, and Shimizu K, *Algorithm Mol. Biol.*, **3**: 8, 2008
Kadota K, Nakai Y, and Shimizu K, *submitted*

話の内容

- 組織特異的遺伝子検出法

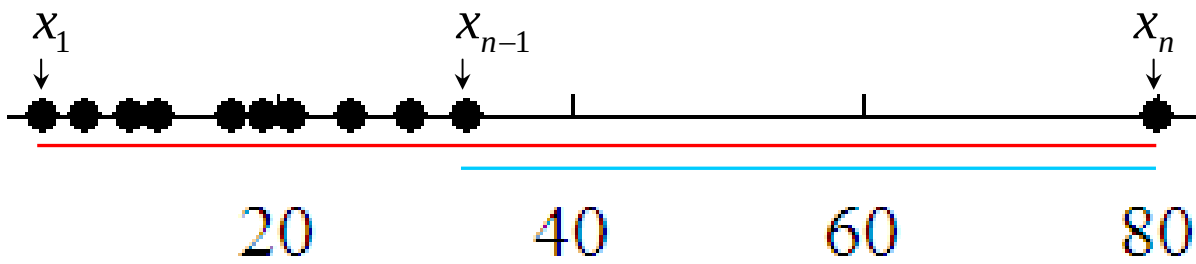
	方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
①	Dixon test	×	×	?	○	-
	Pattern matching	○	○	×	○	-
	Tissue specificity index	○	×	-	○	-
②	Entropy (*H*)	○	×	×	○	-
	Tukey-Kramer's HSD test	○	×	?	○	-
④	AIC	○	○	○	×	○
	Sprent's method	○	○	○	×	×
③	ROKU (AIC+a modified *H*)	○	○	○	○	○

組織特異的遺伝子検出法

- ①Dixon test（$0 \leqq D \leqq 1$）
  - 一組織のみで高発現（低発現）しているパターンを検出

x	
組織	発現量
肺	4
骨	7
脳	10
皮膚	17
延髄	19
心臓	21
胃	25
小腸	29
膵臓	33
肝臓	80

一般化	
組織	発現量
Tissue 1	x_1
Tissue 2	x_2
…	…
Tissue i	x_i
…	…
…	…
Tissue n-1	x_{n-1}
Tissue n	x_n



高発現の場合：$D(\boldsymbol{x}) = \dfrac{x_n - x_{n-1}}{x_n - x_1} = \dfrac{80-33}{80-4} = 0.618$

（低発現の場合：$D(\boldsymbol{x}) = \dfrac{x_2 - x_1}{x_n - x_1}$）

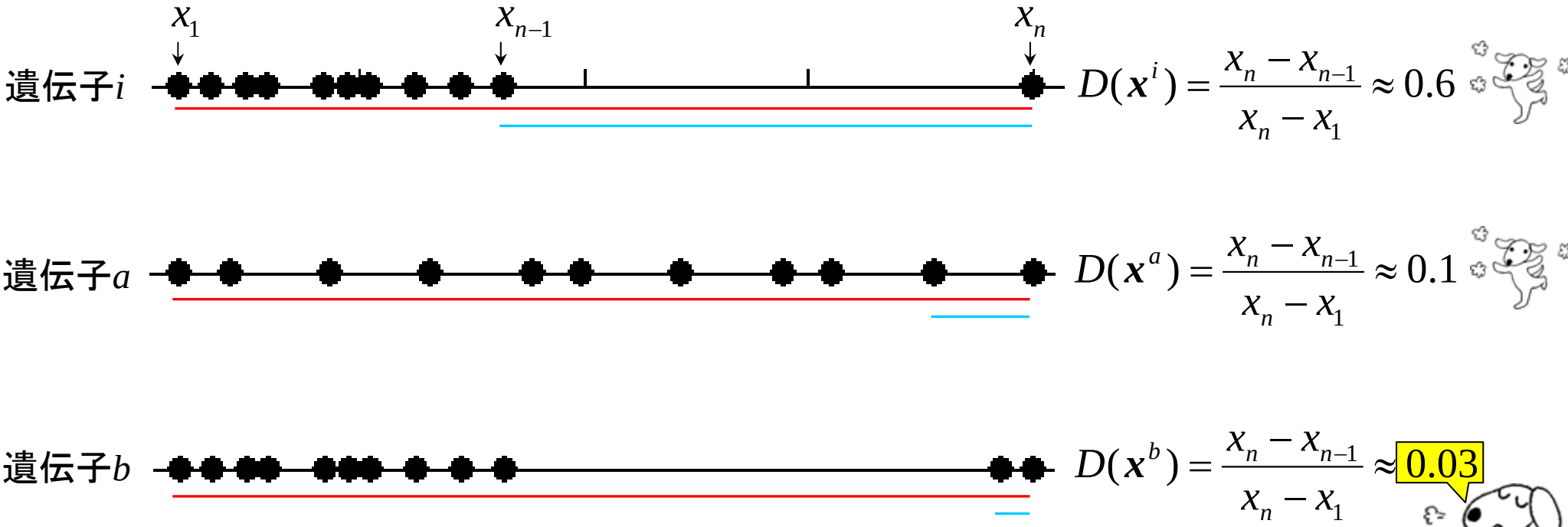
統計量Dの大きい遺伝子を抽出

Dixon WJ, *Biometrics*, 1953

組織特異的遺伝子検出法

方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
①Dixon test	×	×	?	○	-
[illegible] matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
②Entropy (H)	○	×	×	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
④AIC	○	○	○	×	○
Sperli's method	○	○	○	×	×
③ROKU (AIC + a modified H)	○	○	○	○	○

- ①Dixon testの欠点（$0 \leqq D \leqq 1$）
  - 複数の外れ値が互いに外れ値をかばいあう効果（**マスク効果**）の影響を受ける



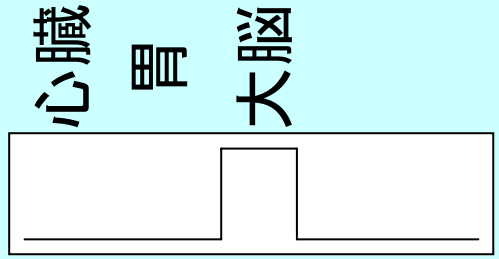
Dixon統計量によるランキングでは複数外れ値に対応不可

組織特異的遺伝子

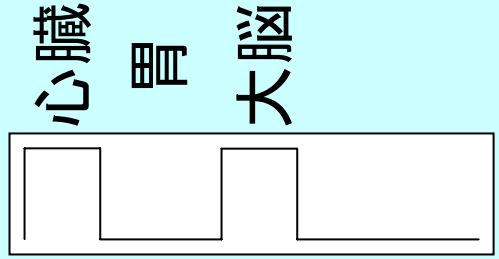
- やりたいこと

方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
① Dixon test	×	×	?	○	-
Pattern matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
② Entropy (H)	○	×	×	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
④ AIC	○	○	○	×	○
Sprent's method	○	○	○	×	×
③ ROKU (AIC+a modified H)	○	○	○	○	○

	心臓	肝	大脳	
gene 1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$ …
gene 2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$ …
…	…	…	…	… …
gene i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$
…	…	…	…	… …
gene p	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$ …



「大脳」特異的
高発現遺伝子



「心臓と大脳」特異的
高発現遺伝子

入力：遺伝子発現行列

出力：任意の組織特異的遺伝子

様々な特異的発現パターンを組織特異性の
度合いで統一的にランキングしたい

Schug *et al.*, *Genome Biol.*, 2005

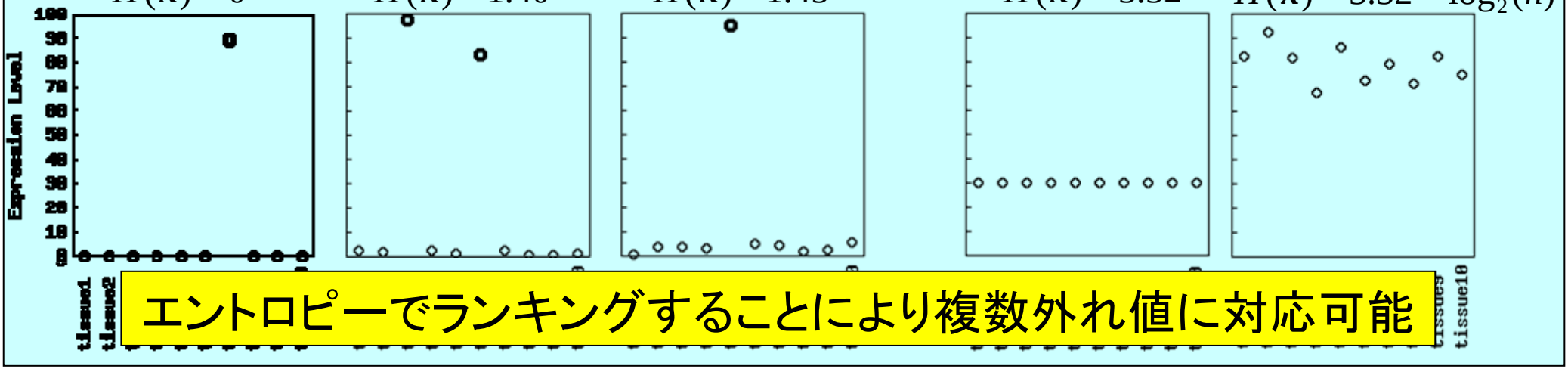
組織特異的遺伝子

方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
① Dixon test	×	×	?	○	-
Pattern matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
② Entropy (H)	○	×	×	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
④ AIC	○	○	○	×	○
Sprent's method	○	○	○	×	×
③ ROKU (AIC+a modified H)	○	○	○	○	○

- ②エントロピーによるラ
  - 遺伝子$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$のエントロピー $H(\boldsymbol{x})$

$$H(\boldsymbol{x}) = -\sum_{i=1}^{n} p_i \log_2(p_i), \textit{where } p_i = x_i / \sum x_i$$

- $H(\boldsymbol{x})$のとりうる範囲： $0 \leqq H(\boldsymbol{x}) \leqq \log_2(n)$



エントロピーでランキングすることにより複数外れ値に対応可能

エントロピーが低い → 組織特異性が高い

エントロピーが高い → 組織特異性が低い

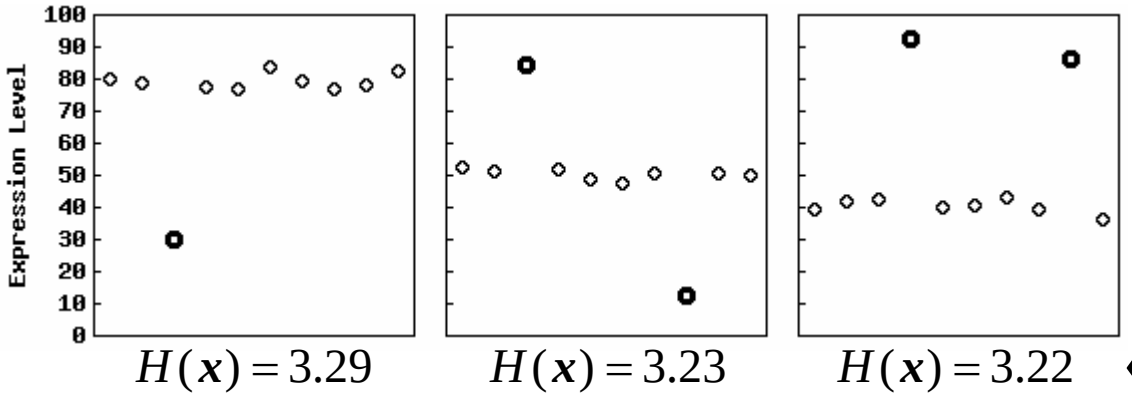
Schug *et al.*, *Genome Biol.*, 2005

組織特異的遺伝子

方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
① Dixon test	×	×	?	○	-
Pattern matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
② Entropy (H)	○	×	×	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
④ AIC	○	○	○	×	○
Sprent's method	○	○	○	×	×
③ ROKU (AIC+a modified H)	○	○	○	○	○

- ②エントロピーの短所

1. 組織特異的低発現パターンなどの検出が不可能

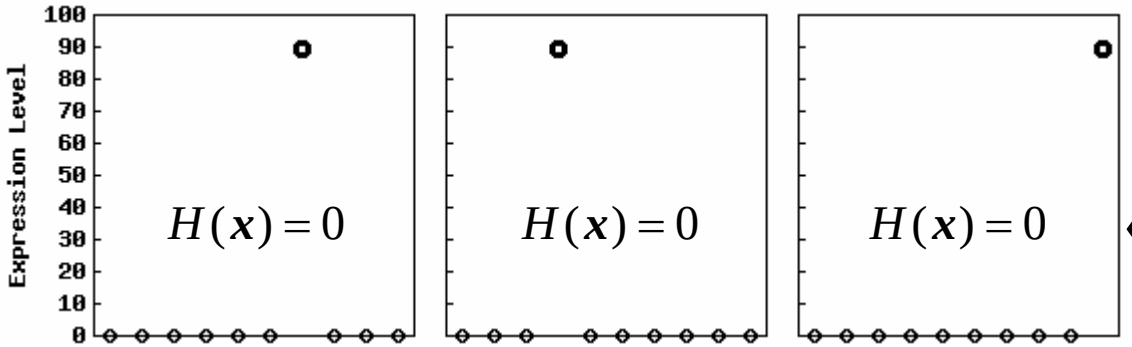


$H(x) = 3.29$　$H(x) = 3.23$　$H(x) = 3.22$

$$0 \leqq H(x) \leqq \log_2(n) = 3.32$$

上位にランキングされない

2. 特異的組織の同定が不可能



$H(x) = 0$　$H(x) = 0$　$H(x) = 0$

どの組織で特異的なのか分からない

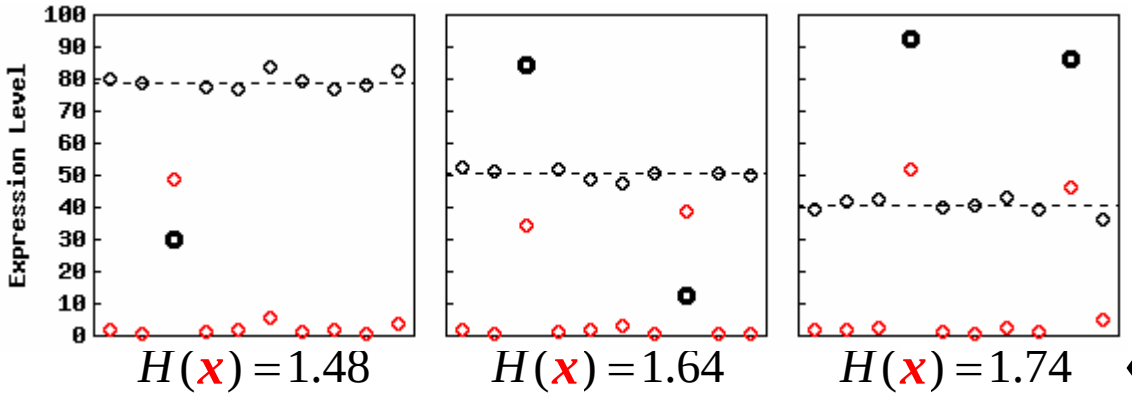
Kadota *et al.*, *BMC Bioinformatics*, 2006

組織特異的遺伝子

方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
① Dixon test	×	×	?	○	-
Pattern matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
② Entropy (H)	○	×	×	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
④ AIC	○	○	○	×	○
Sprent's method	○	○	○	×	×
③ ROKU (AIC+a modified H)	○	○	○	○	○

- ③ROKU

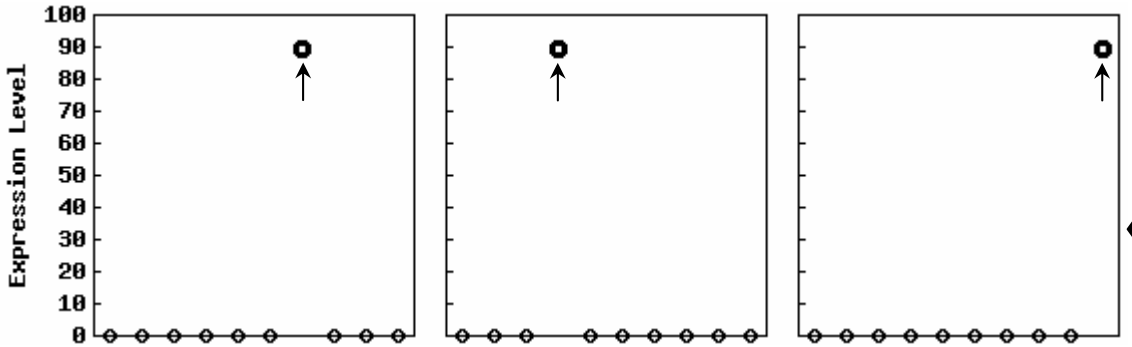
1. 遺伝子発現ベクトルxを変換: $\boldsymbol{x} \rightarrow \boldsymbol{x}$ by $x_i = |x_i - T_{bw}|$



$$0 \leqq H(\boldsymbol{x}) \leqq \frac{\log_2(n)}{\| \\ 3.32}$$

上位にランキングされる

2. AICに基づく外れ値検出法を採用



どの組織で特異的なのか分かる

上田太一郎, *応用統計学*, 1996
Kadota *et al.*, *Physiol. Genomics*, 2003

組織特異的遺伝子検出法

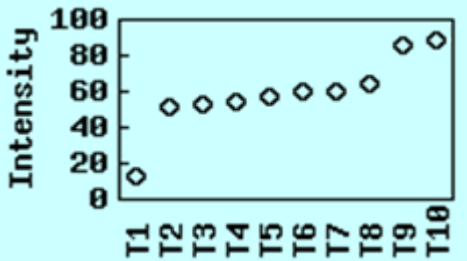
④AICに基づく外れ値検出法

- Akaike's Information Criterion (AIC)
- 様々な外れ値の組み合わせモデルから*AIC*が最小の組み合わせ(MAICE)を探索

$$AIC = n_n \log \hat{\sigma} + \sqrt{2} \times n_o \times \frac{\log n_n!}{n_n}$$

$n_n + n_o (= n)$: サンプル数
n_o : *Outlier*（外れ値）の数
n_n : *Non-outlier*の数
$\hat{\sigma}$: 標準偏差

計算例:



入力

組織	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
発現量	12	51	52	54	57	59	60	63	85	88

出力

出力結果	-1	0	0	0	0	0	0	0	1	1

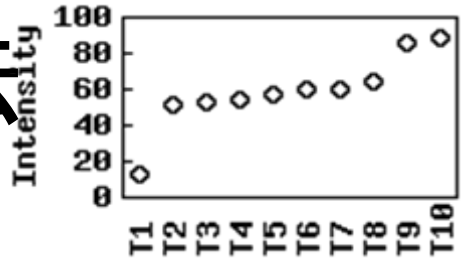
低発現側の外れ値 : -1, 高発現の～ : 1, それ以外 : 0

組織特異的遺伝子検出法

- ④AICに基づく外れ値検出法
  - 様々な外れ値の組み合わせモデルから*AIC*が最小の組み合わせ(MAICE)を探索
  - 様々な外れ値の組み合わせモデル最大探索範囲$Nmax = n/2 = 5$

$$AIC = n_n \log \hat{\sigma} + \sqrt{2} \times n_o \times \frac{\log n_n!}{n_n}$$

$n_n + n_o (= n)$: サンプル数
n_o : *Outlier*（外れ値）の数
n_n : *Non-outlier*の数
$\hat{\sigma}$: 標準偏差



(i)Mean-SD scaling

(ii)Calculate AIC

Nmax = 2　　Nmax = 5

outliers(high)

outliers(low)	none	T10	T9-10	T8-10	T7-10	T6-10
none	-0.53	0.68	1.27	3.14	4.67	5.67
T1	-2.22	-1.97	-6.19	-4.66	-2.91	
T1-2	-0.01	0.19	-4.18	-2.91		
T1-3	1.82	1.94	-2.91			
T1-4	3.27	3.24				
T1-5	4.31					

(iii)Detect outliers

Expression Data

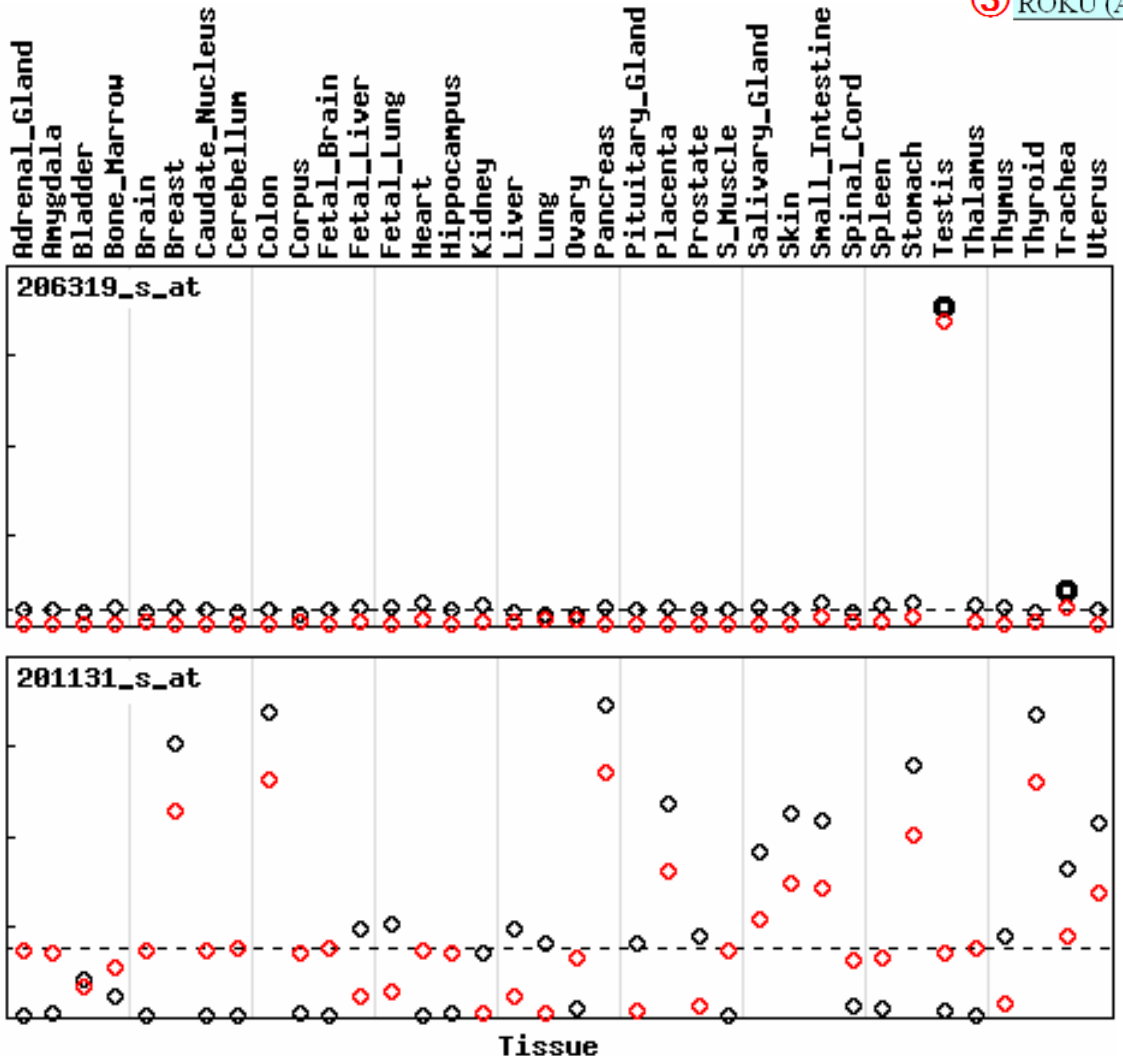
	12	51	52	54	57	59	60	63	85	88
5	-1	0	0	0	0	0	0	0	1	1

1:High-side outlier
0:Non-outlier
-1:Low-side outlier

22,263遺伝子×36組織のデータ(Ge *et al.*, *Genomics*, 2005)

実データで比較

- 全体的な組織特異性の度合いで正しくランキングできるのは？



方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
Dixon test	×	×	?	○	-
Pattern matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
② Entropy (H)	○	×	×	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
AIC	○	○	○	×	○
Sprent's method	○	○	○	×	×
③ ROKU (AIC+a modified H)	○	○	○	○	○

②	③
Schug et al., *Genome Biol.*, 2005	Kadota et al., *BMC Bioinfo.*, 2006

$H(\boldsymbol{x}) = 4.235$　$H(\boldsymbol{x}) = 1.950$

③のほうが正しくランキング可能

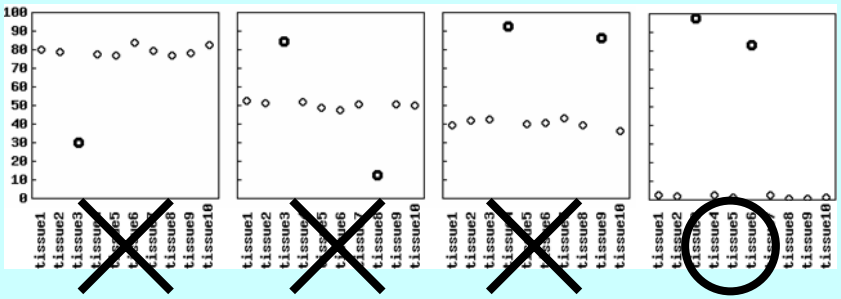
$H(\boldsymbol{x}) = 4.228$　$H(\boldsymbol{x}) = 4.729$

目的組織特異性が高いのは？

② Schug et al., *Genome Biology*, 2005

1) 遺伝子 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$の全体的な組織特異性度合いを表す統計量

$$H(\boldsymbol{x}) = -\sum_{i=1}^{n} p_i \log_2(p_i), \textbf{ where } p_i = x_i \Big/ \sum x_i$$



2) 組織tにおける特異性度合いを表す統計量

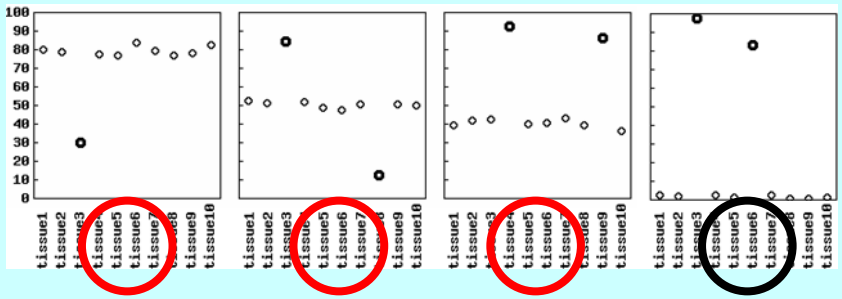
$$Q_t(\boldsymbol{x}) = H(\boldsymbol{x}) - \log_2(p_t)$$

全遺伝子について統計量を計算し、最低の統計量をもつものが最もt組織特異的高発現遺伝子

③ Kadota et al., *BMC Bioinformatics*, 2006

1)遺伝子$\boldsymbol{x}$を変換($x_i = |x_i - T_{bw}|$)し、変換後のベクトル$\boldsymbol{x}$のエントロピーを利用

$$H(\boldsymbol{x}) = -\sum_{i=1}^{n} p_i \log_2(p_i), \textbf{ where } p_i = x_i \Big/ \sum x_i$$



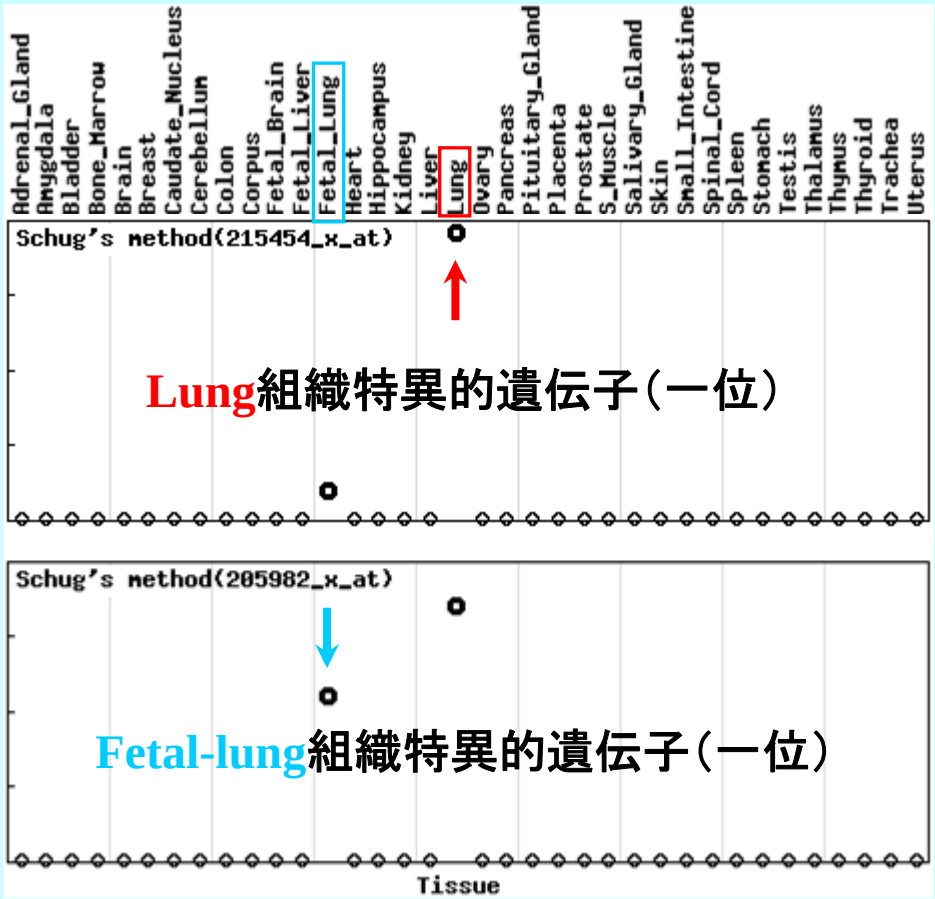
2) AICに基づく外れ値検出法の適用

入力	2.1	1.3	9.0	1.9	1.1	2.0	0.8	0.5	1.2
				t					
出力	0	0	1	0	0	0	0	0	0

組織tのみで1、それ以外で0の遺伝子群を抽出。その中で最低の$H(\boldsymbol{x})$をもつものが最もt組織特異的高発現遺伝子

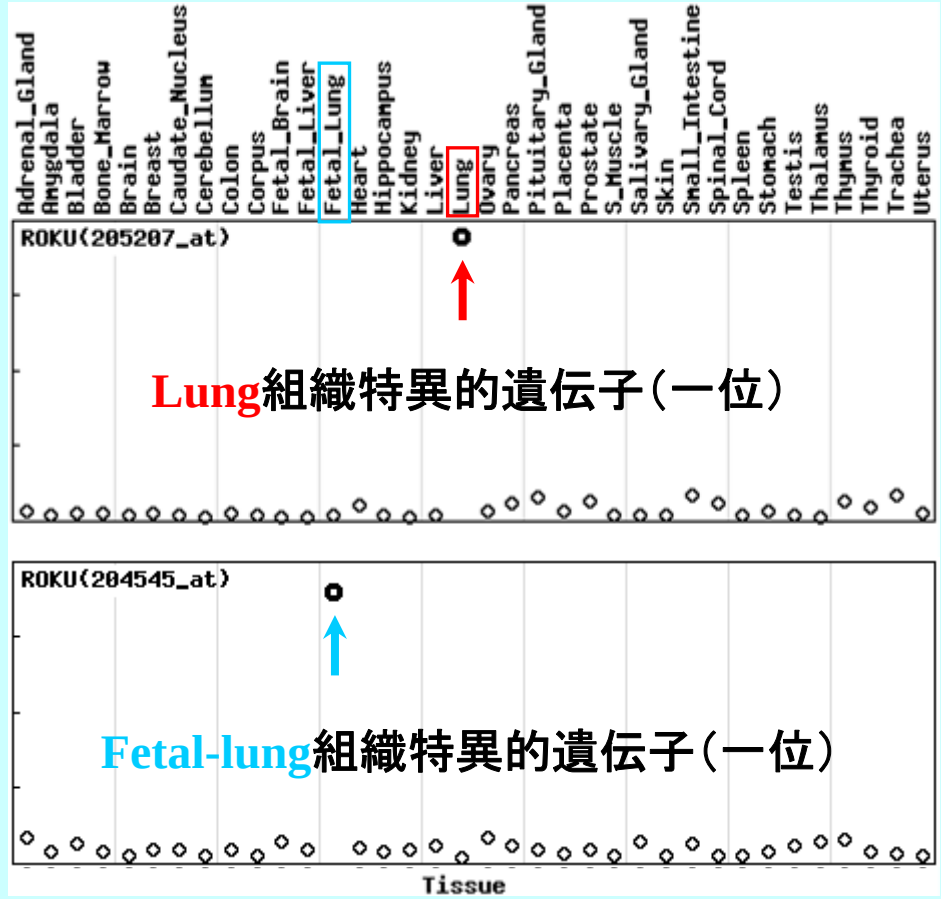
目的組織特異性が高いのは？

② Schug et al., *Genome Biology*, 2005



目的組織以外でも特異的：×

③ Kadota et al., *BMC Bioinformatics*, 2006



目的組織のみで特異的：○

結論(組織特異的遺伝子検出法)

方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
Dixon test	×	×	?	○	-
Pattern matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
Entropy (*H*)	○	×	×	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
AIC	○	○	○	×	○
Sprent's method	○	○	○	×	×
ROKU (AIC+a modified *H*)	○	○	○	○	○

ROKUがおすすめ

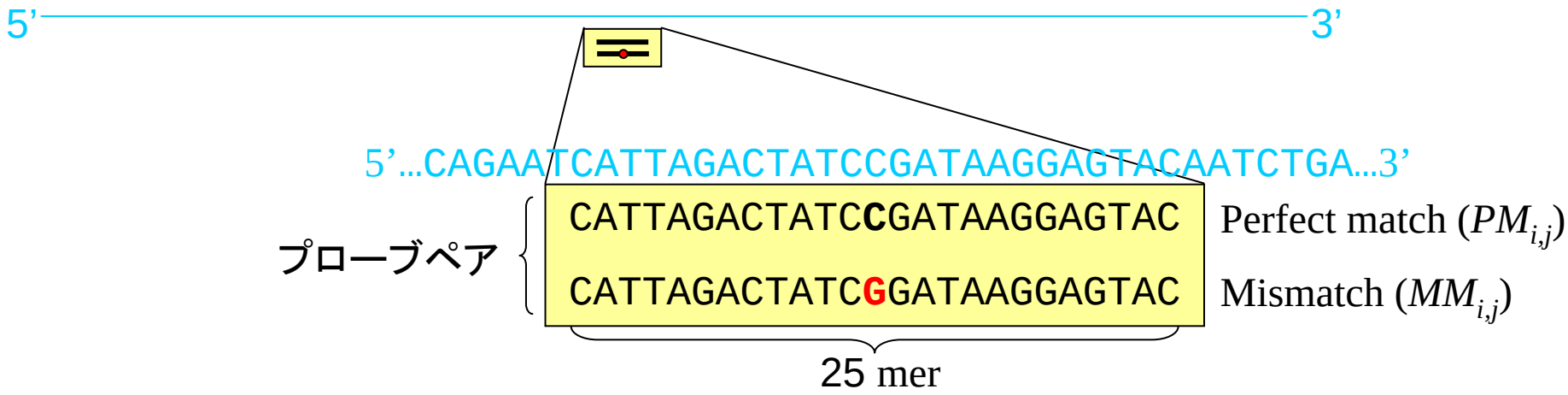
Kadota K, Ye J, Nakai Y, Terada T, and Shimizu K, *BMC Bioinformatics*, **7**: 294, 2006
Kadota K, Konishi T, and Shimizu K, *Gene Regulation Systems Biol.*, **1**:9-15, 2007

話の内容

- 組織特異的遺伝子検出法
  - Dixon test, ROKU, …
- Affymetrix GeneChipデータ解析
  - **PM-MM戦略**
  - 様々な前処理（正規化）法
  - 様々な二群間での発現変動遺伝子ランキング法
  - 重視すべき評価基準は？
    - 感度・特異度
    - 再現性
  - 推奨ガイドライン（MAQC vs. 門田）

Affymetrix GeneChipデータ解析

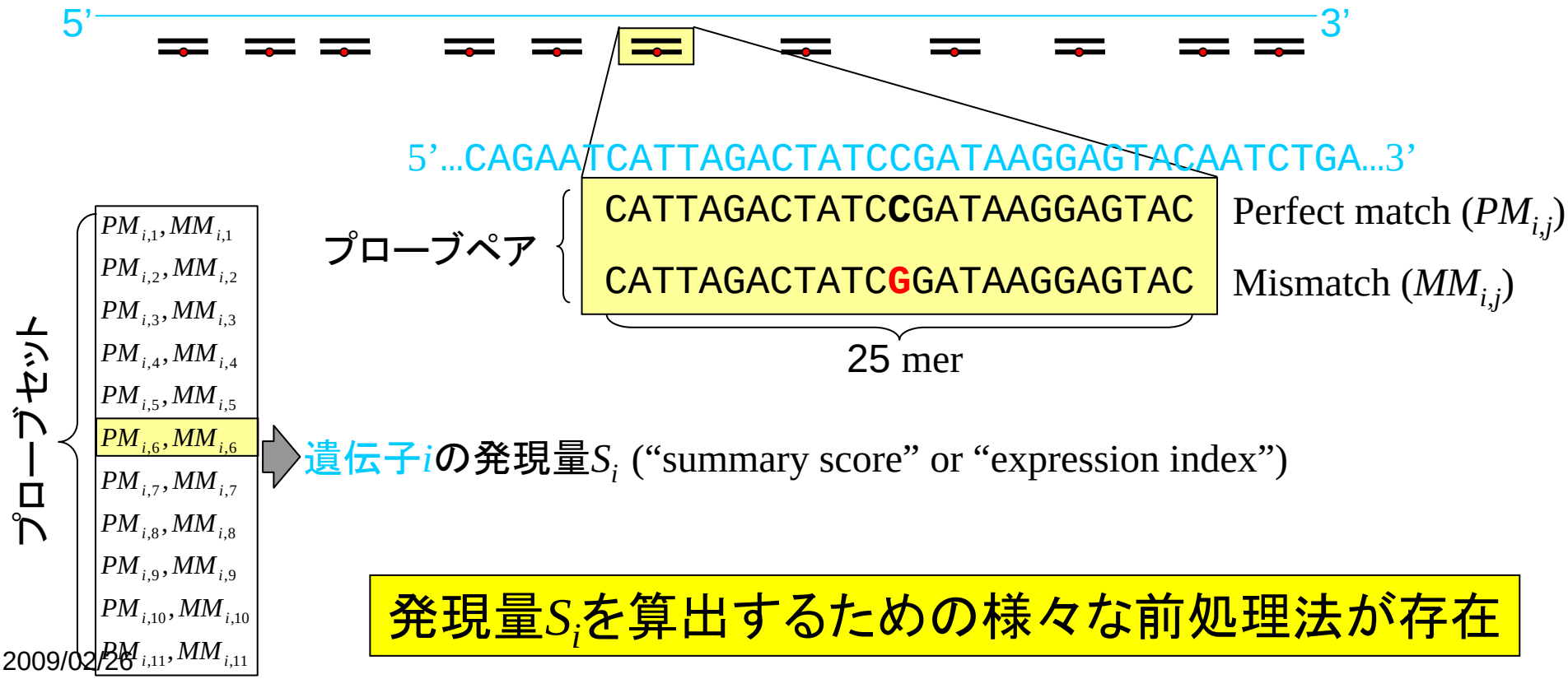
- PM-MM戦略
  - 遺伝子iの発現量S_iを正確に知るために



特異的なハイブリダイゼーションと非特異的なハイブリダイゼーションを区別すべく、目的遺伝子配列に対してPMと一塩基MMがペアになっている

Affymetrix GeneChipデータ解析

- PM-MM戦略
  - 遺伝子iの発現量S_iを正確に知るためにn_i(n_i=11~20)種類のプローブペアのシグナル強度をもとに計算



発現量S_iを算出するための様々な前処理法が存在

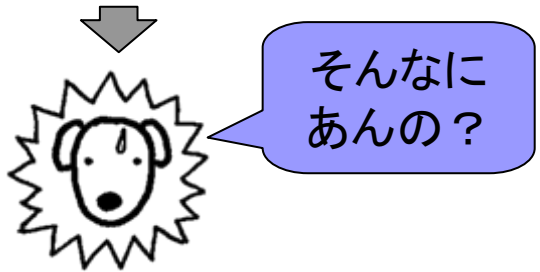
話の内容

- 組織特異的遺伝子検出法
  - Dixon test, ROKU, …
- Affymetrix GeneChipデータ解析
  - PM-MM戦略
  - **様々な前処理（正規化）法**
  - 様々な二群間での発現変動遺伝子ランキング法
  - 重視すべき評価基準は？
    - 感度・特異度
    - 再現性
  - 推奨ガイドライン（MAQC vs. 門田）

Affymetrix GeneChipデータ解析

- 様々な前処理法（発現量S_iを算出するための）
  - MBEI (Li and Wong, *PNAS*, **98**, 31-36, 2001)
  - MAS5 (Hubbell *et al.*, *Bioinformatics*, **18**, 1585-92, 2002)
  - RMA (Irizarry *et al.*, *Biostatistics*, **4**, 249-64, 2003)
  - GCRMA (Wu *et al.*, *Tech. Rep., John Hopkins Univ.*, 2003)
  - PDNN (Zhang *et al.*, *Nat. Biotechnol.*, **21**, 818-21, 2003)
  - PLIER (Affymetrix, 2004)
  - SuperNorm (Konishi, T., *BMC Bioinformatics*, **5**, 5, 2004)
  - multi-mgMOS (Liu *et al.*, *Bioinformatics*, **21**, 3637-3644, 2005)
  - GLA (Zhou and Rocke, *Bioinformatics*, **21**, 3983-3989, 2005)
  - FARMS (Hochreiter *et al.*, *Bioinformatics*, **22**, 943-949, 2006)
  - DFW (Chen *et al.*, *Bioinformatics*, **23**, 321-327, 2007)
  - Hook (Binder *et al.*, *AMB*, **3**, 11, 2008)

どの前処理法がいい？



生データ（$PM_{i,j}, MM_{i,j}$）
in .CEL files

↓

バックグラウンド補正（within-array）

↓

正規化（cross-array）

↓

PM値の補正

↓

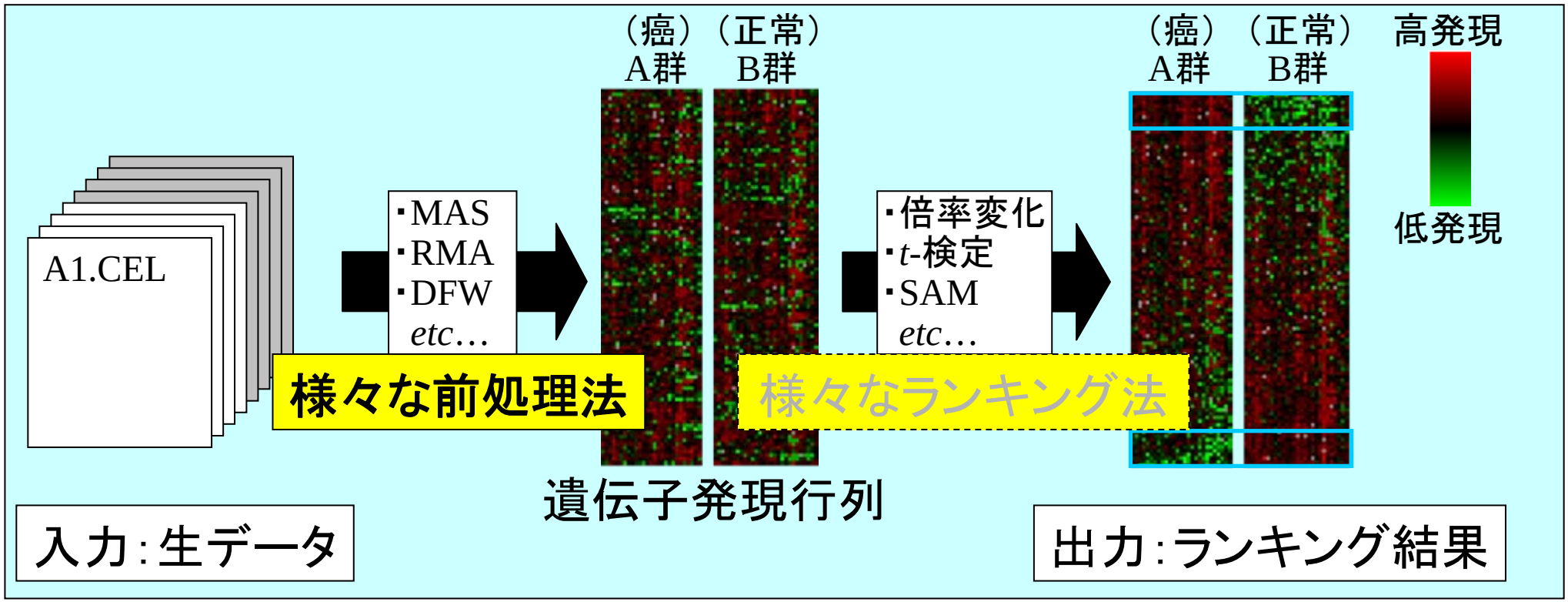
Summarization

↓

発現量S_i

何が問題？

- 発現変動遺伝子検出が目的の場合



問題：用いる手法によって結果がかなり異なる
手法選択のガイドラインなし

Affymetrix GeneChipデータ解析

- どの前処理法がいい？（比較例：**MAS5** vs. **RMA**）
  - 評価基準1（感度・特異度）：既知の発現変動遺伝子（spiked-in genes）をどれだけ上位にランキング可能か？（AUC値の高さ）

MAS5の遺伝子発現行列

Gene	sample C1	sample C2	sample D1	sample D2
gene 1				
gene 2				
gene 3				
gene 4				
gene 5				

RMA の遺伝子発現行列

Gene	sample C1	sample C2	sample D1	sample D2
gene 1				
gene 2				
gene 3				
gene 4				
gene 5				

|log比|を計算

$\lvert\log_2(C/D)\rvert$
0.4
3.0
0.2
2.0
0.7

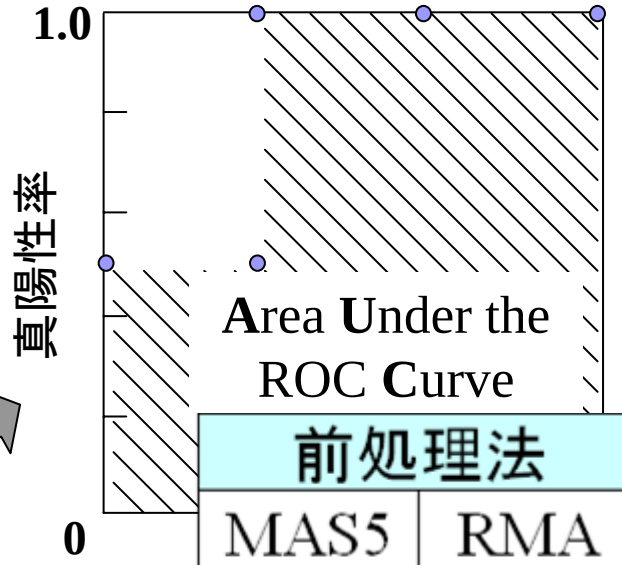
$\lvert\log_2(C/D)\rvert$
0.8
1.9
0.5
1.3
1.4

|log比|でランキング

$\lvert\log_2(C/D)\rvert$	Gene
3.0	gene 2
2.0	gene 4
0.7	gene 5
0.4	gene 1
0.2	gene 3

$\lvert\log_2(C/D)\rvert$	Gene
1.9	gene 2
1.4	gene 5
1.3	gene 4
0.8	gene 1
0.5	gene 3

ROC曲線が左上にあるほどよい方法



前処理法	MAS5	RMA
ランキング法	FC (log比)	1.000	0.833

Affymetrix GeneChipデータ解析

- どの前処理法がいい？
  - 評価基準1（感度・特異度）：既知の発現変動遺伝子（spiked-in genes）をどれだけ上位にランキング可能か？（AUC値の高さ）
    - MBEI (Li and Wong, *PNAS*, **98**, 31-36, 2001)
    - MAS5 (Hubbell *et al.*, *Bioinformatics*, **18**, 1585-92, 2002)
    - RMA (Irizarry *et al.*, *Biostatistics*, **4**, 249-64, 2003)
    - GCRMA (Wu *et al.*, *Tech. Rep., John Hopkins Univ.*, 2003)
    - PDNN (Zhang *et al.*, *Nat. Biotechnol.*, **21**, 818-21, 2003)
    - PLIER (Affymetrix, 2004)
    - SuperNorm (Konishi, T., *BMC Bioinformatics*, **5**, 5, 2004)
    - multi-mgMOS (Liu *et al.*, *Bioinformatics*, **21**, 3637-3644, 2005)
    - GLA (Zhou and Rocke, *Bioinformatics*, **21**, 3983-3989, 2005)
    - **FARMS (Hochreiter *et al.*, *Bioinformatics*, 22, 943-949, 2006)**
    - **DFW (Chen *et al.*, *Bioinformatics*, 23, 321-327, 2007)**
    - Hook (Binder *et al.*, *AMB*, **3**, 11, 2008)

「感度・特異度」の高い前処理法

--- 「Affycomp II 」の問題 ---

- ベンチマークデータセットのみ → 現実のデータ解析精度を保証しない
- ランキング法が「log比」のみ → 他のランキング法については？

話の内容

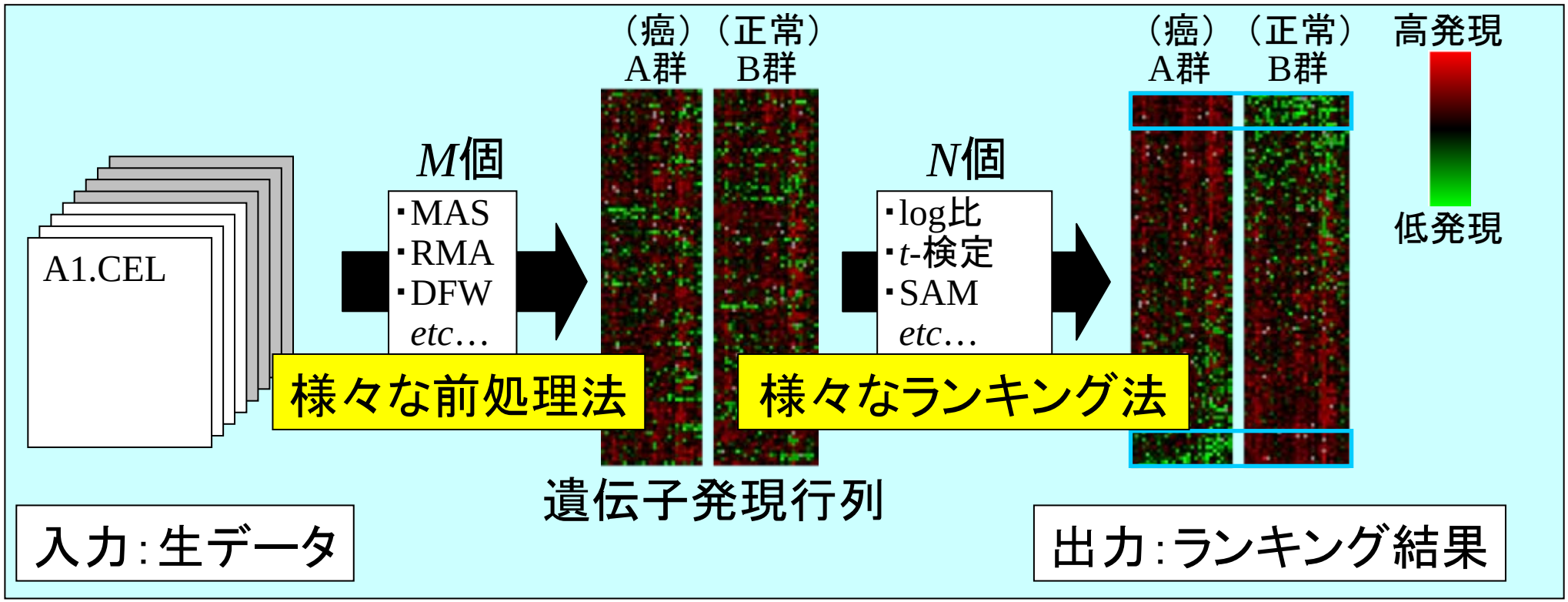
- 組織特異的遺伝子検出法
  - Dixon test, ROKU, …
- Affymetrix GeneChipデータ解析
  - PM-MM戦略
  - 様々な前処理（正規化）法
  - **様々な二群間での発現変動遺伝子ランキング法**
  - 重視すべき評価基準は？
    - 感度・特異度
    - 再現性
  - 推奨ガイドライン（MAQC vs. 門田）

Affymetrix GeneChipデータ解析

- 様々な二群間での発現変動遺伝子ランキング法
  - log比
  - SAM (Tusher *et al*., *PNAS*, **98**, 5116-5121, 2001)
  - CyberT (Baldi and Long, *Bioinformatics*, **17**, 509-519, 2001)
  - samroc (Broberg P., *Genome Biol*., **4**, R41, 2003)
  - a moderated *t* (Smyth GK, *SAGMB*, **3**, Article 3, 2004)
  - RDAM (Martin *et al*., *BMC Bioinformatics*, **5**, 148, 2004)
  - Rank products (Breitling *et al*., *FEBS Lett*., **573**, 83-92, 2004)
  - DEDS (Yang *et al*., *Bioinformatics*, **21**, 1084-1093, 2005)
  - IBMT (Sartor *et al*., *BMC Bioinformatics*, **7**, 538, 2006)
  - PPLR (Liu *et al*., *Bioinformatics*, **22**, 2107-2113, 2006)
  - a shrinkage *t* (Opgen-Rhein *et al*., *SAGMB*, **6**, 9, 2007)
  - Layer ranking (Chen *et al*., *BMC Bioinformatics*, **8**, 74, 2007)
  - Combined P (Hess and Iyer, *BMC Genomics*, **8**, 96, 2007)
  - WAD (Kadota *et al*., *AMB*, **3**, 8, 2008)

何が問題？

- 発現変動遺伝子検出が目的の場合



(M×N)通りの組み合わせ

Weighted Average Difference (WAD)

- 全体的にシグナル強度の高い遺伝子が上位にくるように重みをかけた統計量

unlogged data

Gene	A1	A2	A3	B1	B2
gene1	128	64	128	128	64
gene2	1024	1024	1024	1024	1024
gene3	512	1024	1024	2048	246
gene4	1024	1024	2048	256	256
gene5	2	2	2	32	32
gene6	2	4	4	64	128
gene7	16	8	32	64	8

$\log_2$-transformed data

Gene	A1	A2	A3	B1	B2
gene1	7	6	7	7	6
gene2	10	10	10	10	10
gene3	9	10	10	11	8
gene4	10	10	11	8	8
gene5	1	1	1	5	5
gene6	1	2	2	6	7
gene7	4	3	5	6	3

Average Difference（AD）統計量 $AD_i = \left|\overline{B_i} - \overline{A_i}\right|$

AD	rank
0.17	6
0.00	7
0.20	5
2.33	3
4.00	2
4.83	1
0.50	4

平均シグナル強度 $x_i = \left(\overline{B_i} + \overline{A_i}\right)/2$

xを(0~1)の範囲に規格化 $w_i = \dfrac{x_i - \min(x)}{\max(x) - \min(x)}$

WAD統計量 $WAD_i = AD_i \times w_i$

x	w	WAD	rank
6.58	0.51	0.09	5
10.00	1.00	0.00	6
9.57	0.94	0.18	3
9.17	0.88	2.06	1
3.00	0.00	0.00	6
4.08	0.15	0.75	2
4.25	0.18	0.09	4

$AD_i = \left|\overline{B_i} - \overline{A_i}\right|$ より

$$AD_{gene6} = \left|(6+7)/2 - (1+2+2)/3\right| = 4.83$$

$x_i = \left(\overline{B_i} + \overline{A_i}\right)/2$ より

$$x_{gene6} = \left((6+7)/2 + (1+2+2)/3\right)/2 = 4.08$$

$w_i = \dfrac{x_i - \min(x)}{\max(x) - \min(x)}$ より

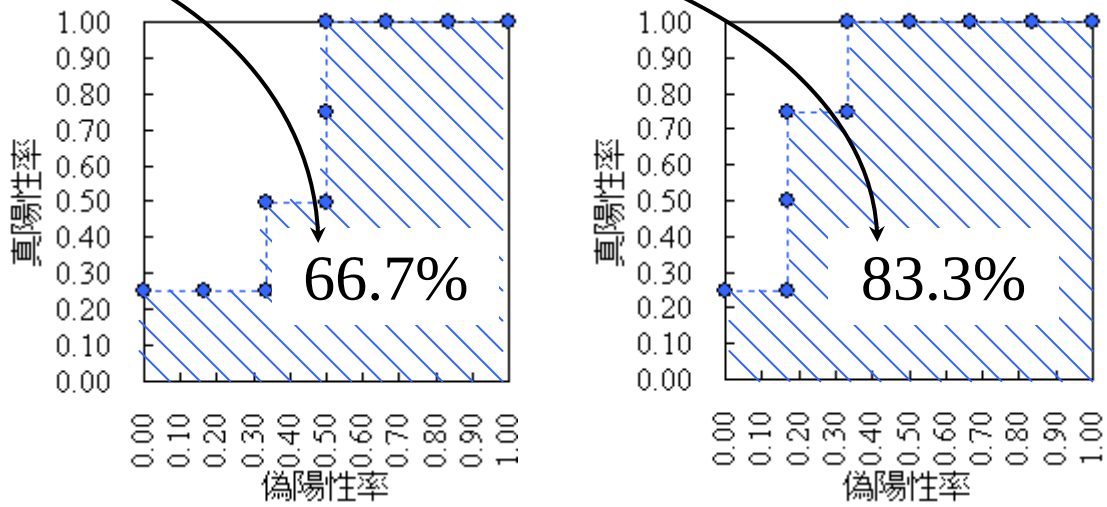
$$w_{gene6} = \frac{4.08 - 3.00}{10.00 - 3.00} = 0.15$$

WADの一位：gene4, ADの一位：gene6

Affymetrix GeneChipデータ解析

- どのランキング法がいい？（比較例：*t*-検定 vs. 倍率変化）
  - 評価基準1（感度・特異度）：既知の発現変動遺伝子（spiked-in genes）をどれだけ上位にランキング可能か？（AUC値の高さ）

ランキング法				
rank	*t*-検定		倍率変化	
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

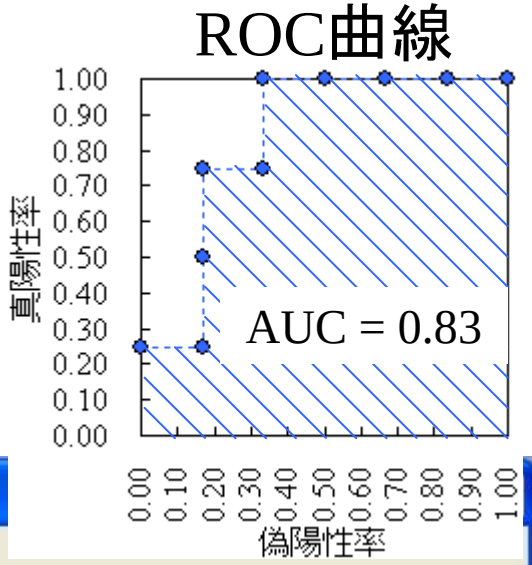


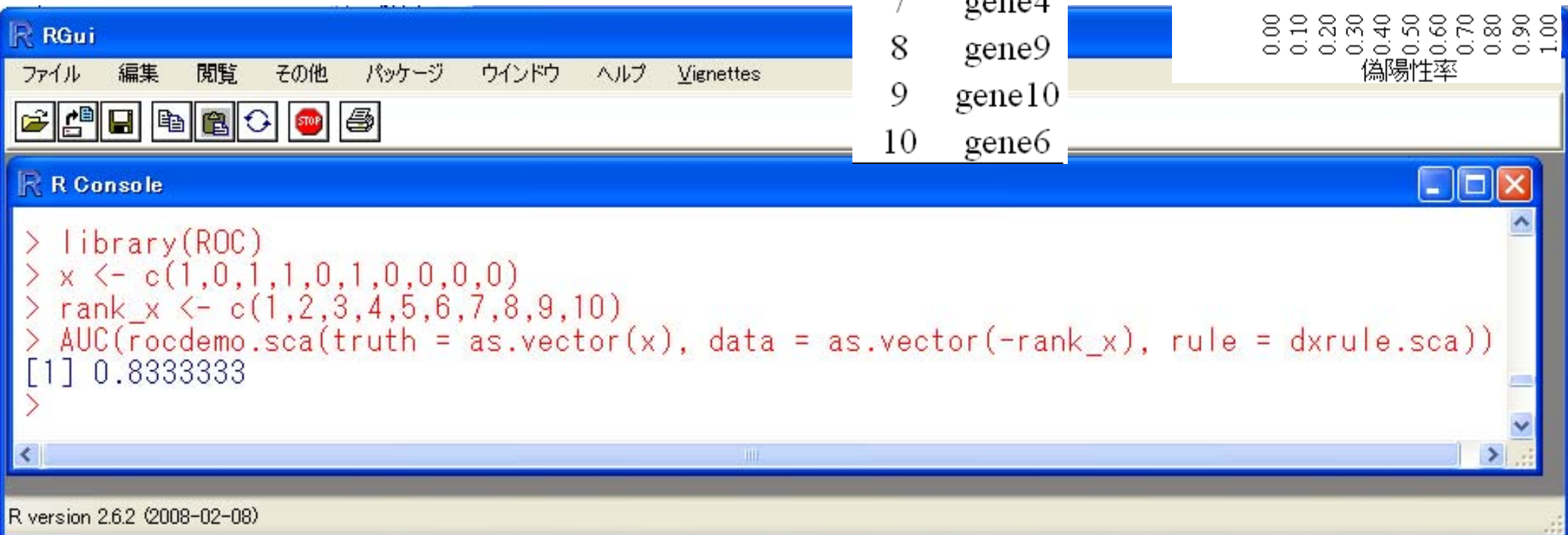
Area **U**nder the ROC **C**urve (ROC曲線の下部面積：**AUC**)

Affymetrix GeneChipデータ解析

- AUC値はRで簡単に計算できます

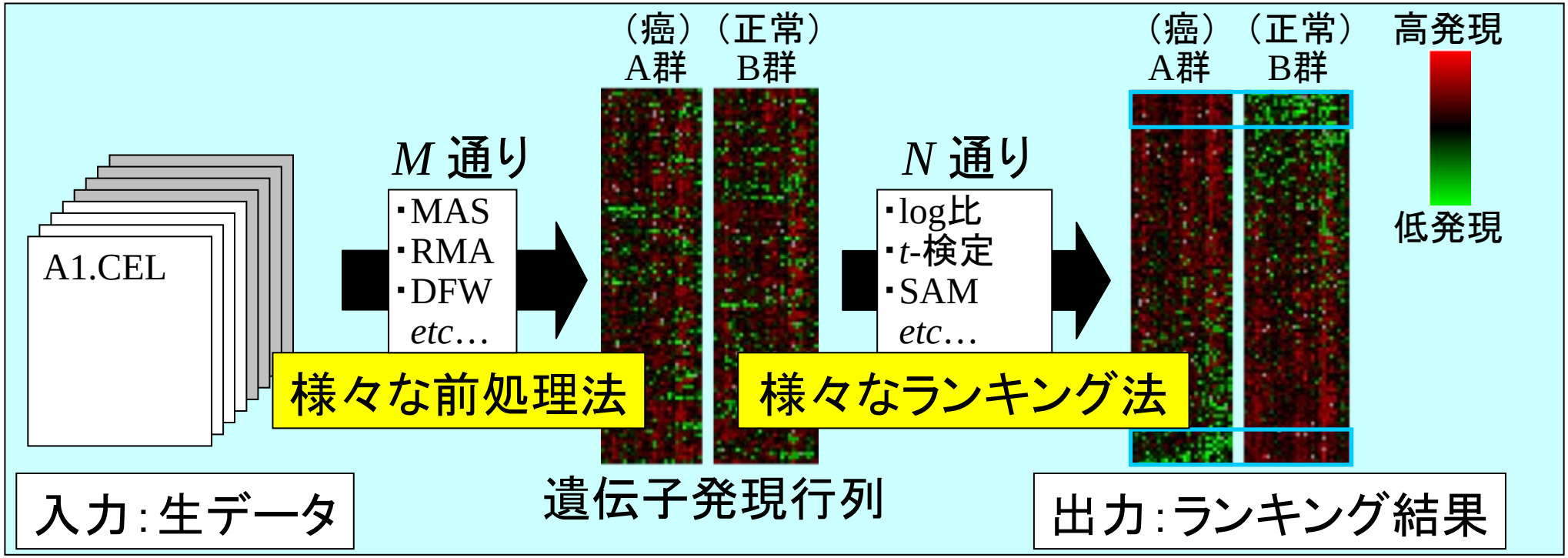
	真実
Rank	Gene
1	gene8
2	gene5
3	gene3
4	gene2
5	gene7
6	gene1
7	gene4
8	gene9
9	gene10
10	gene6





おさらい

- 選択肢が沢山ある。どの組み合わせがいい？



- **AUC値の高さ（0≦AUC≦1）で評価**

  →「感度・特異度」の高い組み合わせを探索

解析データ

- 発現変動遺伝子の全てが既知 or 一部がRT-PCRで確認された計38データセット

シリアル	Affymetrix platform	文献	GEO ID	ランキング	前処理法	真のDEG
1	U95A	Cope_2004	spike-in data			16
2	U133A	-	spike-in data			42
3	U133A	Crimi_2005	GSE1462	FC	MAS5	4
4	U133A	Manley_2007	GSE7819	FC	MAS5	11
5	U133A	Thalacker-Mercer_2007	GSE8441	FC	MAS5	9
6	U133A	Jin_2007	GSE9499	FC	MAS5	77
7	U133A	Hall_2004	GSE974	B	MAS5	3
8	U133A	Viemann_2006	GSE2638,2639	B	MAS5	13
9	U133A	Viemann_2006	GSE2638,2639	B	MAS5	16
10	U133A	Toruner_2004	GSE3524	B	MAS5	4
11	U133A	Csoka_2004	GSE3860	B	MAS5	8
12	U133A	Plager_2007	GSE5667	B	MAS5	3
13	U133A	Plager_2007	GSE5667	B	MAS5	3
14	U133A	Goh_2007	GSE6236	B	MAS5	7
15	U133A	Gumz_2007	GSE6344	B	MAS5	19
16	U133A	Reischl_2007	GSE6710	B	MAS5	7
17	U133A	Parikh_2007	GSE7146	B	MAS5	6
18	U133A	Hsu_2007	GSE7765	B	MAS5	13
19	U133A	Spira_2004	GSE1650	*t*	MAS5	8
20	U133A	Wood_2005	GSE1615	*t*	MAS5	8
21	U133A	Wood_2005	GSE1615	*t*	MAS5	8
22	U133A	Eckfeldt_2005	GSE2666	*t*	MAS5	5
23	U133A	Eckfeldt_2005	GSE2666	*t*	MAS5	6
24	U133A	Hyrcza_2007	GSE6740	*t*	MAS5	40
25	U133A	Hyrcza_2007	GSE6740	*t*	MAS5	62
26	U133A	Tripathi_2007	GSE9574	*t*	MAS5	5
27	U133A	Wu_2006	GSE4917	FC	RMA	5
28	U133A	Cole_2007	GSE7148	FC	RMA	10
29	U133A	Horwitz_2004	GSE5967	B	RMA	6
30	U133A	Pescatori_2007	GSE6011	B	RMA	10
31	U133A	Gomez_2007	GSE8562	B	RMA	8
32	U133A	Jaworski_2006	GSE1937	*t*	RMA	12
33	U133A	Raetz_2006	GSE1577	*t*	RMA	9
34	U133A	Barth_2005	GSE2240	*t*	RMA	3
35	U133A	Barth_2005	GSE2240	*t*	RMA	9
36	U133A	Burleigh_2007	GSE2531	*t*	RMA	17
37	U133A	Ryan_2006	GSE5389	*t*	RMA	6
38	U133A	Lockstone_2007	GSE5390	*t*	RMA	8

FC: 倍率変化（Fold Change）に基づく方法、*t*: *t*-検定に基づく方法、B: 両方（FC and *t*）

「感度・特異度」解析結果

- 平均%AUC値（AUCが高い →「感度・特異度」が高い）

前処理法

ランキング法	MAS5	multi-mgMOS	RMA	VSN	GCRMA	MBEI	PLIER	FARMS	DFW	
Datasets 3-26										
WAD	**96.74**	**94.84**	91.37	90.97	92.67	88.19	89.15	91.58	91.41	Fold Change (FC)系
AD	93.76	93.13	93.10	**92.96**	**93.42**	89.14	87.32	92.47	92.24	
FC	93.63	92.82	**93.12**	92.92	93.16	89.71	86.20	92.49	92.24	
RP	91.51	92.20	92.54	92.48	93.23	**90.07**	**92.01**	**93.06**	**92.53**	
modT	95.67	91.91	91.38	90.70	91.19	86.79	85.43	89.23	90.11	*t*検定系
samT	95.95	92.99	91.23	90.94	91.07	87.09	85.25	89.40	89.96	
shrinkT	95.73	92.56	91.32	90.62	92.12	86.89	84.65	-	91.45	
ibmT	96.34	93.11	91.77	91.02	91.06	87.10	86.27	90.04	90.25	

--- データセット3-26（原著論文の解析手段） ---

前処理法：**MAS5**

ランキング法：FC系が4, *t*検定系が8, 両方同時が12

「感度・特異度」解析結果

- 平均%AUC値（AUCが高い →「感度・特異度」が高い）

前処理法

ランキング法		MAS5	multi-mgMOS	RMA	VSN	GCRMA	MBEI	PLIER	FARMS	DFW	
Datasets 3-26: **MAS5**, FC系が4, *t* 検定系が8, 両方同時が12											
	WAD	**96.74**	**94.84**	91.37	90.97	92.67	88.19	89.15	91.58	91.41	Fold Change (FC)系
	AD	93.76	93.13	93.10	**92.96**	**93.42**	89.14	87.32	92.47	92.24	
	FC	93.63	92.82	**93.12**	92.92	93.16	89.71	86.20	92.49	92.24	
	RP	91.51	92.20	92.54	92.48	93.23	**90.07**	**92.01**	**93.06**	**92.53**	
	modT	95.67	91.91	91.38	90.70	91.19	86.79	85.43	89.23	90.11	*t*検定系
	samT	95.95	92.99	91.23	90.94	91.07	87.09	85.25	89.40	89.96	
	shrinkT	95.73	92					.65	-	91.45	
	ibmT	96.34	93					.27	90.04	90.25	
Datasets 27-38: **RMA**, FC系が2, *t* 検定系が7, 両方同時が3											
	WAD	**92.42**	**92.36**	96.73	95.41	95.75	93.39	91.30	93.55	94.09	Fold Change (FC)系
	AD	87.41	86.99	**96.77**	96.22	96.18	93.11	89.01	93.81	94.22	
	FC	88.23	85.92	96.73	96.20	96.06	92.94	88.82	93.81	94.22	
	RP	84.55	86.94	96.53	**96.53**	**96.25**	**93.53**	**91.57**	**94.76**	**94.67**	
	modT	90.90	89.61	95.28	94.53	93.33	90.50	89.28	91.62	92.36	*t*検定系
	samT	90.31	89.14	95.70	95.08	93.60	90.49	88.53	91.90	92.05	
	shrinkT	90.97	89.85	94.85	94.04	94.48	90.13	88.56	-	93.68	

前処理法の比較は困難

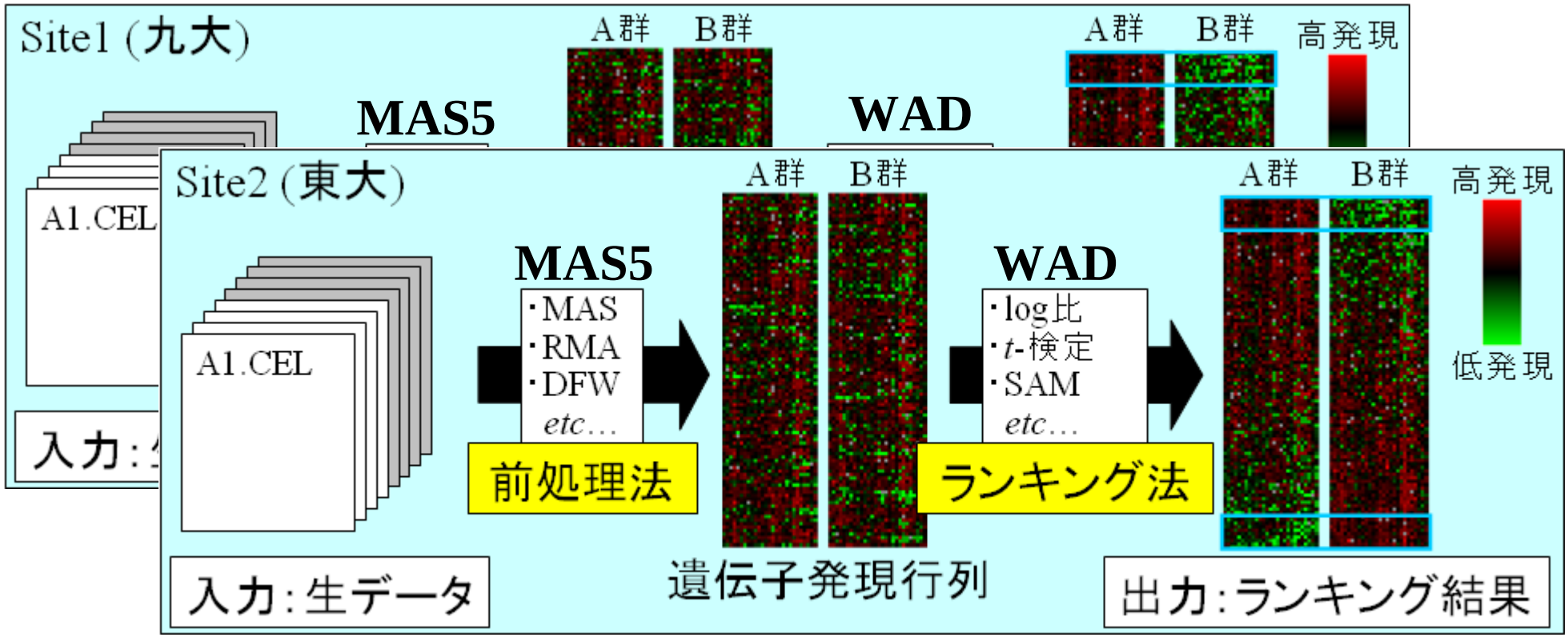
推奨ランキング法はFold Change系（WAD or RP）

話の内容

- 組織特異的遺伝子検出法
  - Dixon test, ROKU, …
- Affymetrix GeneChipデータ解析
  - PM-MM戦略
  - 様々な前処理（正規化）法
  - 様々な二群間での発現変動遺伝子ランキング法
  - **重視すべき評価基準は？**
    - 感度・特異度
    - **再現性**
  - 推奨ガイドライン（MAQC vs. 門田）

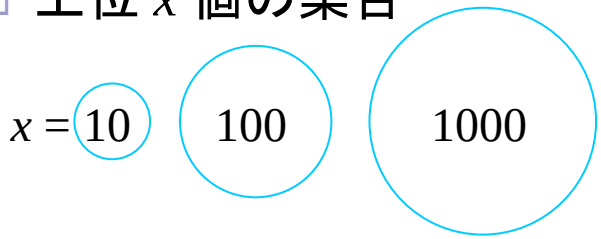
Affymetrix GeneChipデータ解析

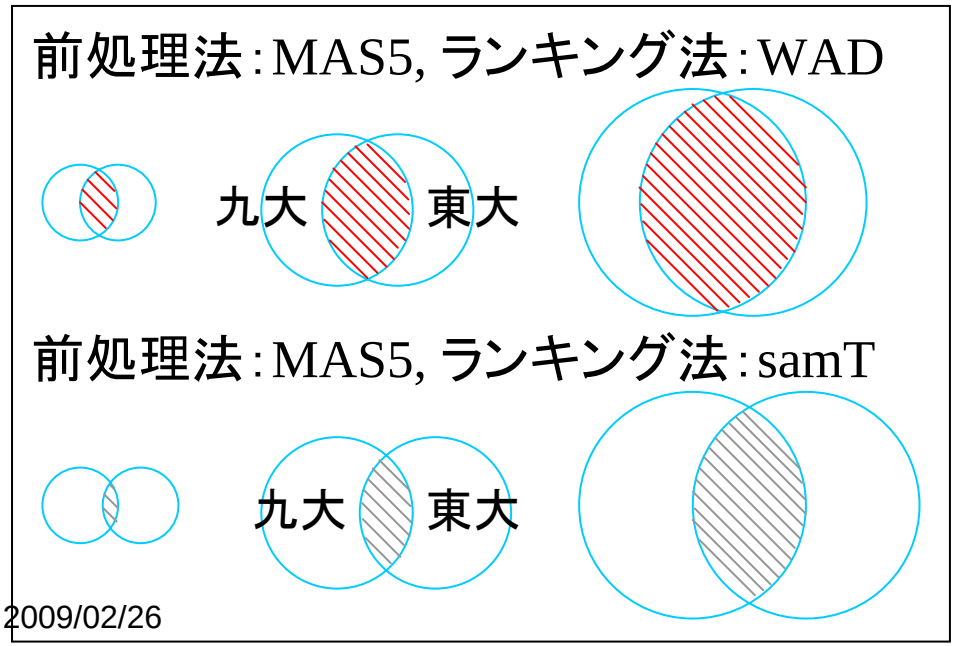
- **評価基準2(再現性)**: POG (**P**ercentage of **O**verlapping **G**enes)値の高さ
  - MicroArray Quality Control (MAQC) プロジェクトで提唱(0≦POG≦100%)
  - POG値が高い → ランキング結果の頑健性(再現性)が高い方法

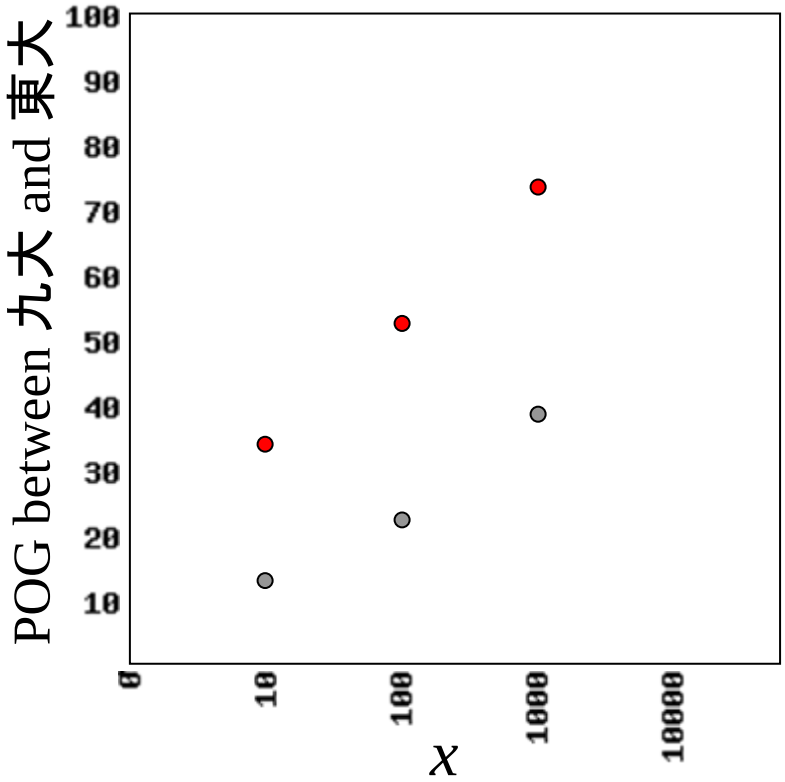


Affymetrix GeneChipデータ解析

- **評価基準2(再現性)**: POG (**P**ercentage of **O**verlapping **G**enes)値の高さ
  - MicroArray Quality Control (MAQC) プロジェクトで提唱(0≦POG≦100%)
  - POG値が高い → ランキング結果の頑健性(再現性)が高い方法
  - 上位 *x* 個の集合



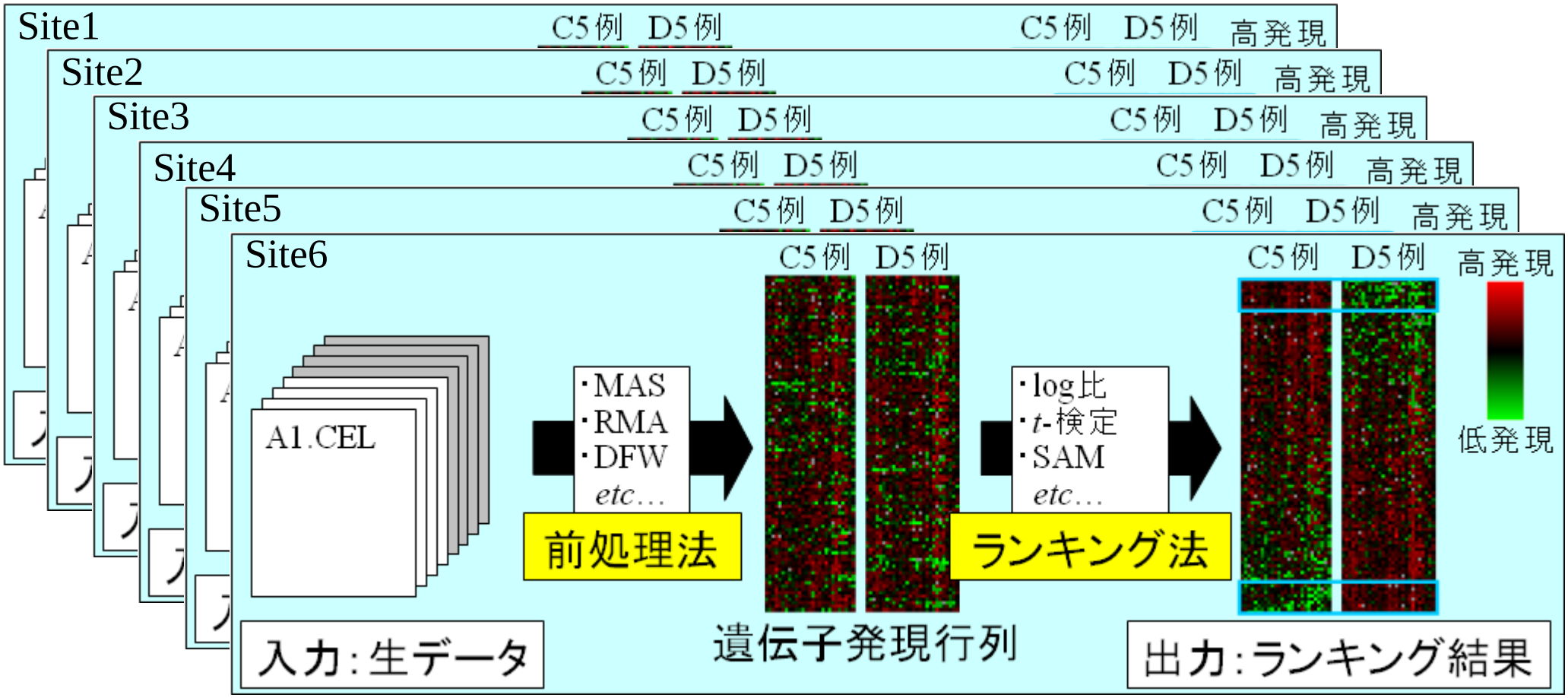




再現性:WAD > samT

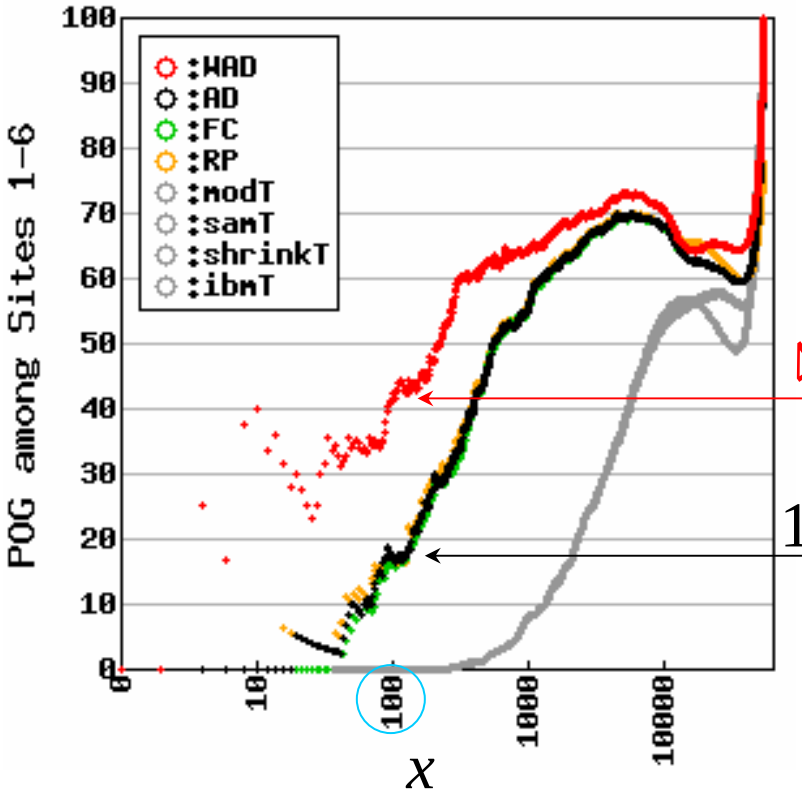
解析データ

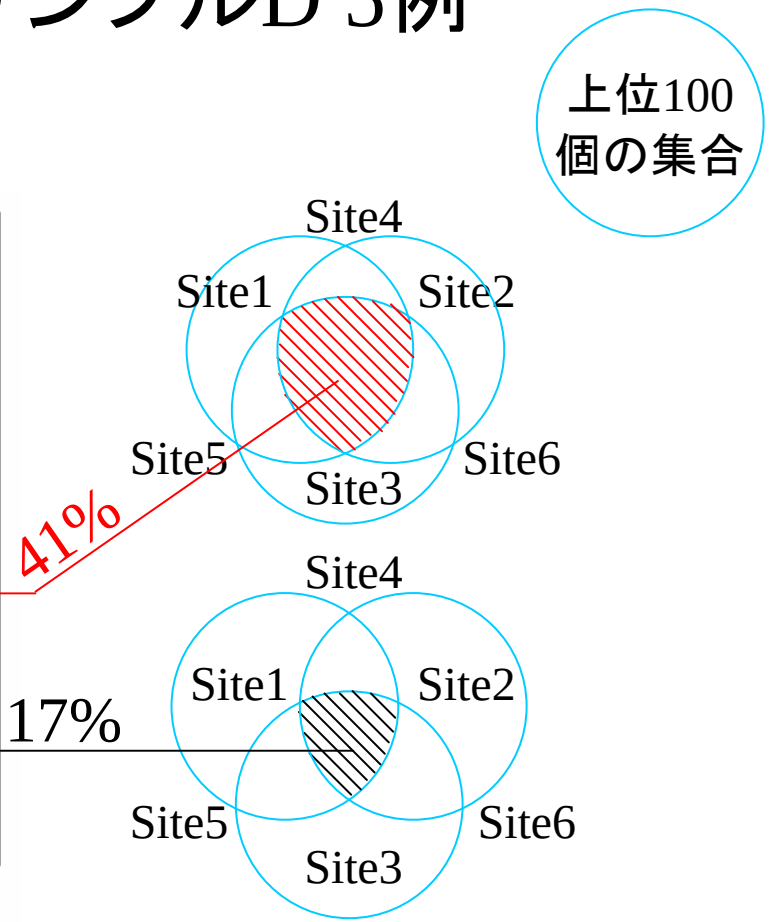
- MicroArray Quality Control（MAQC）projectのデータセット



「再現性」解析結果（前処理法：FARMS）

- サンプルC 5例 vs. サンプルD 5例





再現性：WAD > MAQC推奨法（AD）

Weighted Average Difference (WAD)

- 全体的にシグナル強度の高い遺伝子が上位にくるように重みをかけた統計量

unlogged data

Gene	A1	A2	A3	B1	B2
gene1	128	64	128	128	64
gene2	1024	1024	1024	1024	1024
gene3	512	1024	1024	2048	246
gene4	1024	1024	2048	256	256
gene5	2	2	2	32	32
gene6	2	4	4	64	128
gene7	16	8	32	64	8

$\log_2$-transformed data

Gene	A1	A2	A3	B1	B2
gene1	7	6	7	7	6
gene2	10	10	10	10	10
gene3	9	10	10	11	8
gene4	10	10	11	8	8
gene5	1	1	1	5	5
gene6	1	2	2	6	7
gene7	4	3	5	6	3

Average Difference（AD）統計量 $AD_i = \left|\overline{B_i} - \overline{A_i}\right|$

AD	rank
0.17	6
0.00	7
0.20	5
2.33	3
4.00	2
4.83	1
0.50	4

平均シグナル強度 $x_i = \left(\overline{B_i} + \overline{A_i}\right)/2$

xを(0~1)の範囲に規格化 $w_i = \dfrac{x_i - \min(x)}{\max(x) - \min(x)}$

WAD統計量 $WAD_i = AD_i \times w_i$

x	w	WAD	rank
6.58	0.51	0.09	5
10.00	1.00	0.00	6
9.57	0.94	0.18	3
9.17	0.88	2.06	1
3.00	0.00	0.00	6
4.08	0.15	0.75	2
4.25	0.18	0.09	4

$AD_i = \left|\overline{B_i} - \overline{A_i}\right|$ より

$AD_{gene6} = \left|(6+7)/2 - (1+2+2)/3\right|$

$= 4.83$

$x_i = \left(\overline{B_i} + \overline{A_i}\right)/2$ より

$x_{gene6} = \left((6+7)/2 + (1+2+2)/3\right)/2$

$= 4.08$

$w_i = \dfrac{x_i - \min(x)}{\max(x) - \min(x)}$ より

$w_{gene6} = \dfrac{4.08 - 3.00}{10.00 - 3.00} = 0.15$

WADの一位：gene4, ADの一位：gene6

話の内容

- 組織特異的遺伝子検出法
  - Dixon test, ROKU, …
- Affymetrix GeneChipデータ解析
  - PM-MM戦略
  - 様々な前処理（正規化）法
  - 様々な二群間での発現変動遺伝子ランキング法
  - 重視すべき評価基準は？
    - 感度･特異度
    - 再現性
  - **推奨ガイドライン**（MAQC vs. 門田）

MAQC推奨ガイドライン

- 「感度・特異度」の高いランキング法
  - *t*-検定系の方法（*P*値）
- 「再現性」の高いランキング法
  - Fold Change（FC）系の方法
    - Average Difference (AD)法

→「感度・特異度・再現性」の高い…
  - 緩めの*P*値の閾値＋AD（FC）によるランキング

Shi *et al.*, *BMC Bioinformatics.*, **9**(Suppl 9):S10, 2008

Conclusions and recommendations

1. A fundamental step of microarray studies is the identification of a small subset of DEGs from among tens of thousands of genes probed on the microarray. DEG lists must be concordant to satisfy the scientific requirement of reproducibility, and must also be specific and sensitive for scientific relevance. A baseline practice is needed for properly assessing reproducibility/concordance alongside specificity and sensitivity.

2. Reports of DEG list instability in the literature are often a direct consequence of comparing DEG lists derived from a simple t-statistic when the sample size is small and variability in variance estimation is large. Therefore, the practice of using *P* alone for gene selection should be discouraged.

3. A DEG list should be chosen in a manner that concurrently satisfies scientific objectives in terms of inherent tradeoffs between reproducibility, specificity, and sensitivity.

4. Using FC and *P* together balances reproducibility, specificity, and sensitivity. Control of specificity and sensitivity can be accomplished with a *P* criterion, while reproducibility is enhanced with an FC criterion. Sensitivity can also be improved by better platforms with greater dynamic range and lower variability or by increased sample sizes.

5. FC-ranking should be used in combination with a non-stringent *P* threshold to select a DEG list that is reproducible, specific, and sensitive, and a joint rule is recommended as a baseline practice.

推奨ガイドラインの比較

- 「感度・特異度」の高いランキング法
  - *t*-検定系の方法（*P*値） ⟷ FC系の方法（WAD or RP）
- 「再現性」の高いランキング法
  - Fold Change（FC）系の方法（AD法） ⟷ FC系の方法（WAD）

MAQC
- MAQC Consortium, *Nat. Biotechnol.*, 2006
- Shi *et al.*, *BMC Bioinformatics.*, 2008

門田ら
- Kadota *et al.*, *AMB.*, 2008
- Kadota *et al.*, *submitted*

「感度・特異度」の高いランキング法

- *t*-検定系の方法（MAQC推奨）⇔ FC系の方法（門田推奨）

「HG-U133A」を用いた論文（計364報）に対する
利用した**前処理法**の調査結果（上位5つ）

前処理法	2003	2004	2005	2006	2007	2008
MAS5 (2002)	8	34	53	42	47	16
RMA (2003)		8	15	29	20	9
MBEI (2001)	0	3	7	16	8	3
GCRMA (2004)			0	5	8	4
VSN (2002)	0	0	0	4	0	2

ランキング法の比較に用いられてきたデータの
多くは**MAS5**処理して得られた遺伝子発現行列

「感度・特異度」の高いランキング法

- *t*-検定系の方法（MAQC推奨）⇔ FC系の方法（門田推奨）

前処理法

ランキング法		MAS5	multi-mgMOS	RMA	VSN	GCRMA	MBEI	PLIER	FARMS	DFW	
	Datasets 3-26：**MAS5**, FC系が4, *t*検定系が8, 両方同時が12										
	WAD	**96.74**	**94.84**	91.37	90.97	92.67	88.19	89.15	91.58	91.41	Fold Change (FC)系
	AD	93.76	93.13	93.10	**92.96**	**93.42**	89.14	87.32	92.47	92.24	
	FC	93.63	92.82	**93.12**	92.92	93.16	89.71	86.20	92.49	92.24	
	RP	91.51	92.20	92.54	92.48	93.23	**90.07**	**92.01**	**93.06**	**92.53**	
	modT	95.67	91.91	91.38	90.70	91.19	86.79	85.43	89.23	90.11	*t*検定系
	samT	95.95	92.99	91.23	90.94	91.07	87.09	85.25	89.40	89.96	
	shrinkT	95.73	92.56	91.32	90.62	92.12	86.89	84.65	-	91.45	
	ibmT	96.34	93.11	91.77	91.02	91.06	87.10	86.27	90.04	90.25	
	Datasets 27-38：**RMA**, FC系が2, *t*検定系が7, 両方同時が3										
	WAD	**92.42**	**92.36**	96.73	95.41	95.75	93.39	91.30	93.55	94.09	Fold Change (FC)系
	AD	87.41	86.99	**96.77**	96.22	96.18	93.11	89.01	93.81	94.22	
	FC	88.23	85.92	96.73	96.20	96.06	92.94	88.82	93.81	94.22	
	RP	84.55	86.94	96.53	**96.53**	**96.25**	**93.53**	**91.57**	**94.76**	**94.67**	
	modT	90.90	89.61	95.28	94.53	93.33	90.50	89.28	91.62	92.36	*t*検定系
	samT	90.31									
	shrinkT	90.97									
	ibmT	91.92	90.60	95.49	94.80	93.67	90.66	89.89	89.95	92.43	

赤枠の中だけで評価すると*t*-検定系がよい

推奨ガイドラインの比較

- 「感度・特異度」の高いランキング法
  - *t*-検定系の方法（*P*値） ⟷ FC系の方法（WAD or RP）
- 「再現性」の高いランキング法
  - Fold Change（FC）系の方法（AD法） ⟷ FC系の方法（WAD）

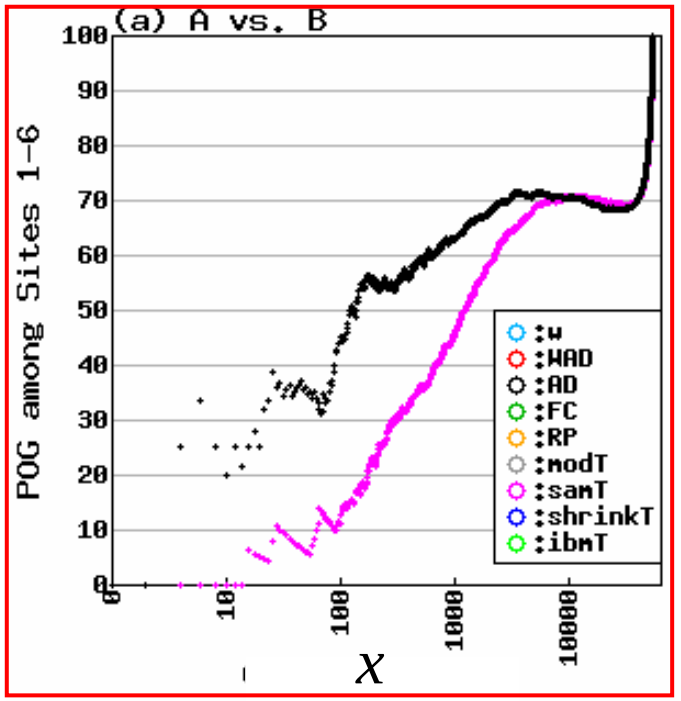
MAQC
- MAQC Consortium, *Nat. Biotechnol.*, 2006
- Shi *et al.*, *BMC Bioinformatics.*, 2008

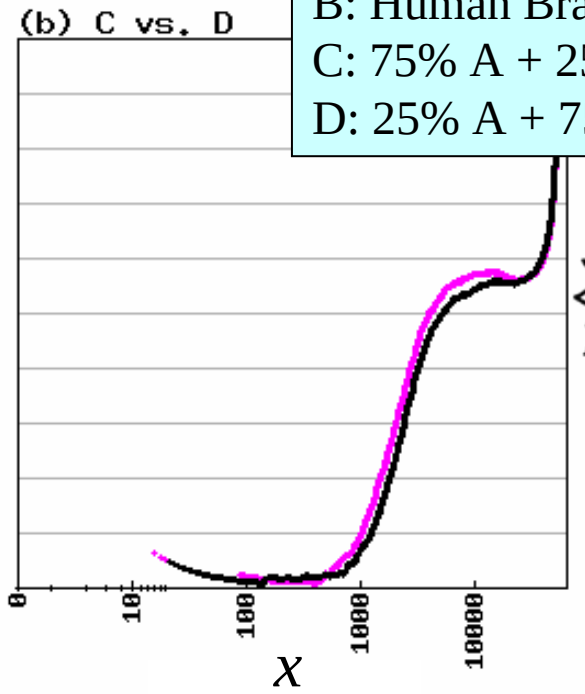
門田ら
- Kadota *et al.*, *AMB.*, 2008
- Kadota *et al.*, *submitted*

「再現性」の高いランキング法は“FC系”で一致

- AD（MAQC推奨）⇔ WAD（門田推奨）

A: Universal Human Reference RNA
B: Human Brain Reference RNA
C: 75% A + 25% B
D: 25% A + 75% B



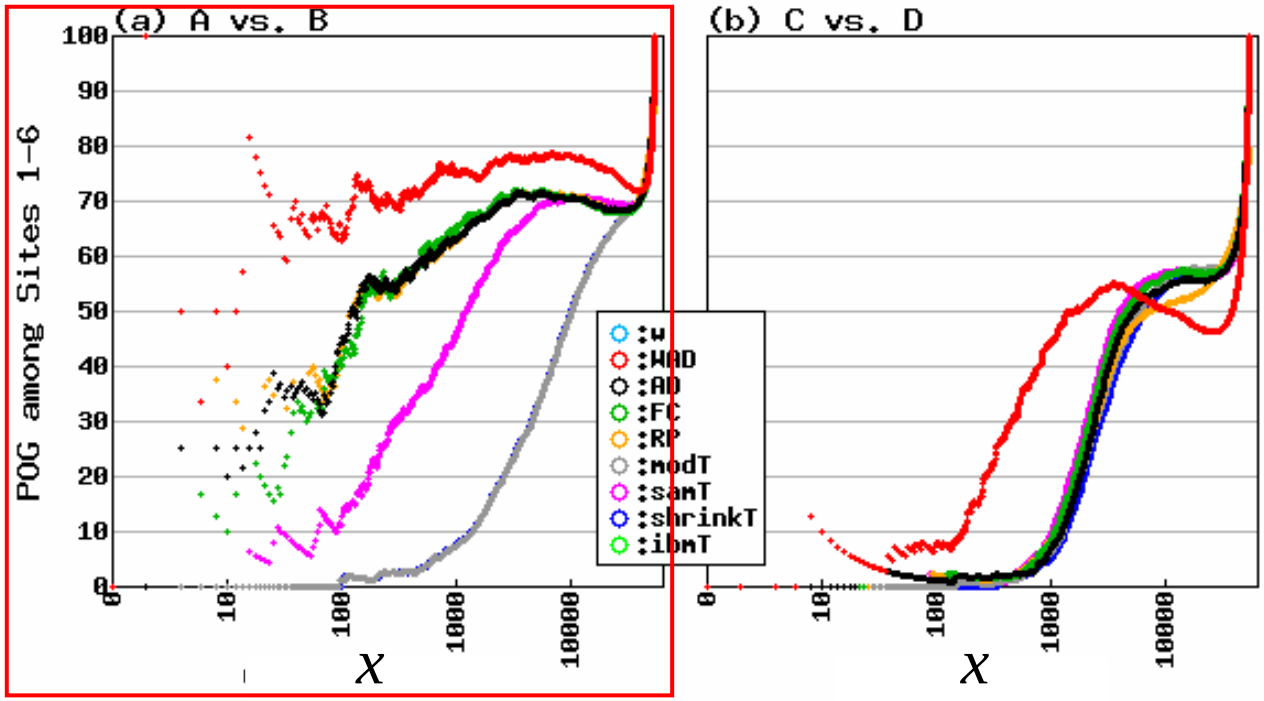


MAQCの解析は:

- 用いた前処理法が**PLIER**のみ
- 比較したランキング法がAD, samT, …のみ
- C vs. Dの比較結果にsamTが含まれてない

「再現性」の高いランキング法は“FC系”で一致

- AD（MAQC推奨）⇔ WAD（門田推奨）



MAQCの解析は:
- 用いた前処理法がPLIERのみ
- 比較したランキング法がAD, samT, …のみ
- C vs. Dの比較結果にsamTが含まれてない

門田らの解析は:
- 用いた前処理法は9種類
- 比較したランキング法は8種類

推奨ガイドラインの比較

- 「感度・特異度」の高いランキング法
  - *t*-検定系の方法（*P*値） ⟷ FC系の方法（WAD or RP）
- 「再現性」の高いランキング法
  - Fold Change（FC）系の方法（AD法） ⟷ FC系の方法（WAD）

MAQC
- MAQC Consortium, *Nat. Biotechnol.*, 2006
- Shi *et al.*, *BMC Bioinformatics.*, 2008

門田ら
- Kadota *et al.*, *AMB.*, 2008
- Kadota *et al.*, *submitted*

Kadota K, Nakai Y, and Shimizu K, *Algorithm Mol. Biol.*, **3**:8, 2008
Kadota K, Nakai Y, and Shimizu K, *submitted*

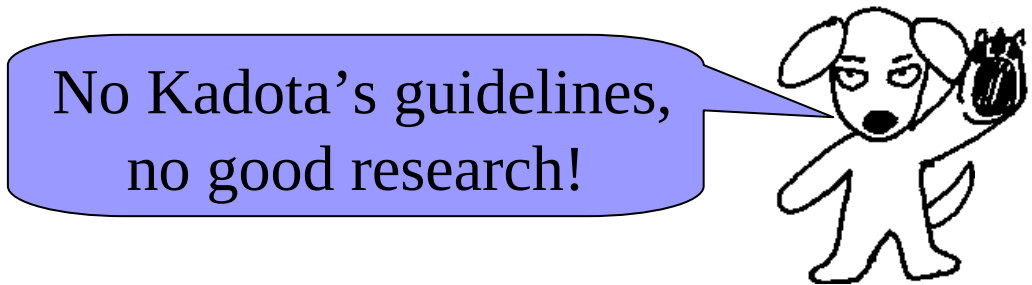
結論（Affymetrix GeneChipデータ解析）

- 「感度・特異度」が高い方法（組合せが重要である！）

前処理法	MAS5	multi-mgMOS	RMA	VSN	GCRMA	MBEI	PLIER	FARMS	DFW
ランキング法	WAD	WAD	RP	RP	RP	RP	RP	RP	RP

WAD: Weighted Average Difference
RP: Rank Products
} Fold Changeに基づく方法

- （発現変動遺伝子リストの）「再現性」が高い方法
  - （前処理法によらず）WAD



共同研究者

東京大学　大学院農学生命科学研究科

清水 謙多郎　教授

中井 雄治 特任准教授

葉 佳臻　氏（アグリバイオ人材養成プログラム修了生）