マイクロアレイ解析手法あれこれ

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

自己紹介

Since 2002

東大・院農　応用生命工学専攻
生物情報工学研究室（6号館）
（清水謙多郎教授）

学位論文：「cDNAマイクロアレイを用いた遺伝子発現解析手法の開発」

2002/4

2003/11

2005/2

産総研・生命情報科学研究センター

放医研・先端遺伝子発現研究センター

東大・院農・アグリバイオ

マイクロアレイ

cDNA-AFLP (HiCEP)

「トランスクリプトーム解析」

Contents

1. マイクロアレイ解析手法
   - 組織特異的発現パターン検出法
   - 二群間比較用の手法
2. cDNA-AFLP(HiCEP)
   - Kadota *et al.*, *BMC Bioinformatics*, 2005
   - Kadota *et al.*, *submitted*

本日のお題

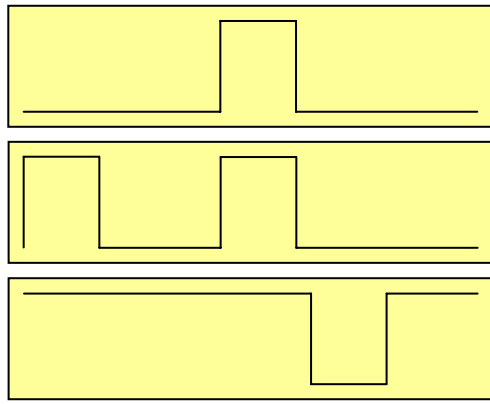
- 「マイクロアレイ解析手法あれこれ」
  - 組織特異的(選択的)発現遺伝子検出法
  - 二群間で発現の異なる遺伝子検出法

組織特異的発現遺伝子検出法

- ランキングに基づく方法
  - **Dixon test**（Greller and Tobin, *Genome Res.*, 1999）
  - **Pattern matching**（Pavlidis and Noble, *Genome Biol.*, 2001）
  - **Tissue specificity Index**（Yanai *et al.*, *Bioinformatics*, 2005）
  - **Entropy**（Schug *et al.*, *Genome Biol.*, 2005）
  - **Tukey-Kramer's Honest Significance Difference (HSD) test**（Liang *et al.*, *Physiol. Genomics*, 2006）
- 外れ値検出に基づく方法
  - **Akaike's Information Criterion (AIC)**（Kadota *et al.*, *Physiol. Genomics*, 2003）
  - **Sprent's non-parametric method**（Ge *et al.*, *Genomics*, 2005）
- 組み合わせ（AIC + a modified entropy）
  - **ROKU** (Kadota *et al.*, *BMC Bioinformatics*, 2006)

様々な組織（条件）

	S1	S2	S3	S4	…
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 *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



Greller and Tobin, *Genome Res.*, 1999

ランキングに基づく方法1

- Dixon test
  - Dixon WJ, *Biometrics*, 1953.
  - 一組織のみで高発現（低発現）しているパターンを検出

x

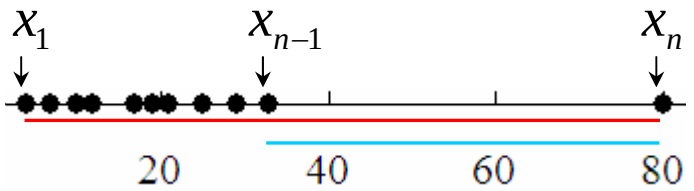
組織	発現量
大脳	10
延髄	19
骨	7
心臓	21
肺	4
肝臓	80
胃	25
皮膚	17
膵臓	33
小腸	29

sort →

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：$D(\boldsymbol{x}) = \dfrac{x_2 - x_1}{x_n - x_1}$）

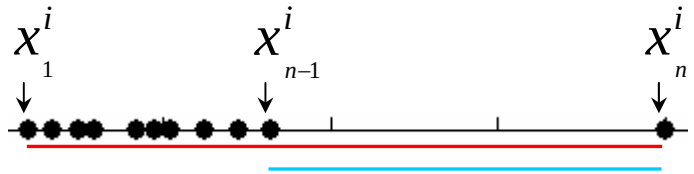
統計量Dの大きい遺伝子を抽出

上田太一郎, *応用統計学*, 1996

Dixon test（とその変法）の欠点

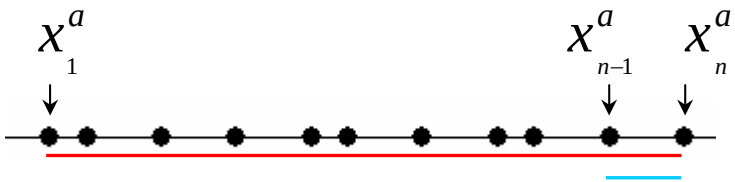
- 複数外れ値（二組織以上で高発現）に対応不可
  - 複数の外れ値が互いに外れ値をかばいあう効果（**マスク効果**）の影響を受ける

遺伝子i



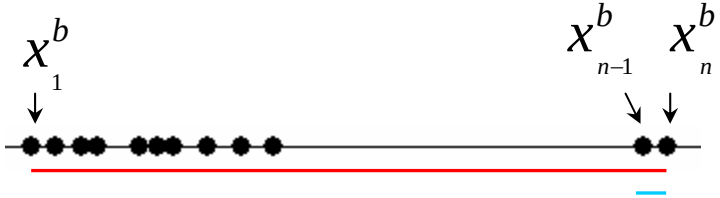
$$D(\boldsymbol{x}) = \frac{x_n - x_{n-1}}{x_n - x_1} \approx 0.6$$

遺伝子a



$$D(\boldsymbol{x}) = \frac{x_n - x_{n-1}}{x_n - x_1} \approx 0.1$$

遺伝子b



$$D(\boldsymbol{x}) = \frac{x_n - x_{n-1}}{x_n - x_1} \approx 0.03$$

ランキングに基づく方法2

- Tissue specificity index τ
  - 遺伝子発現ベクトル $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$に対し、

$$\tau = \frac{\sum_{i=1}^{n}(1-p_i)}{n-1}, \quad \textbf{where}\ p_i = x_i / \max(\boldsymbol{x})$$

  - 例: $\boldsymbol{x} = (0, 8, 0, 0, 0, 2, 0, 2, 0, 0, 0, 0)$

    $\boldsymbol{p} = (0, 1, 0, 0, 0, 0.25, 0, 0.25, 0, 0, 0, 0)$

    $\tau(\boldsymbol{x}) = (1+0+1+1+1+0.75+1+0.75+1+1+1+1)/(12-1) = 0.95$
  - $\tau(\boldsymbol{x})$のとりうる範囲: $0 \leqq \tau \leqq 1$

Housekeeping gene **Tissue-specific gene**

統計量 τ の大きい遺伝子を抽出

ランキングに基づく方法3

- エントロピー H
  - 遺伝子発現ベクトル $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$に対し、そのエントロピー$H(\boldsymbol{x})$は以下の式で表される:

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

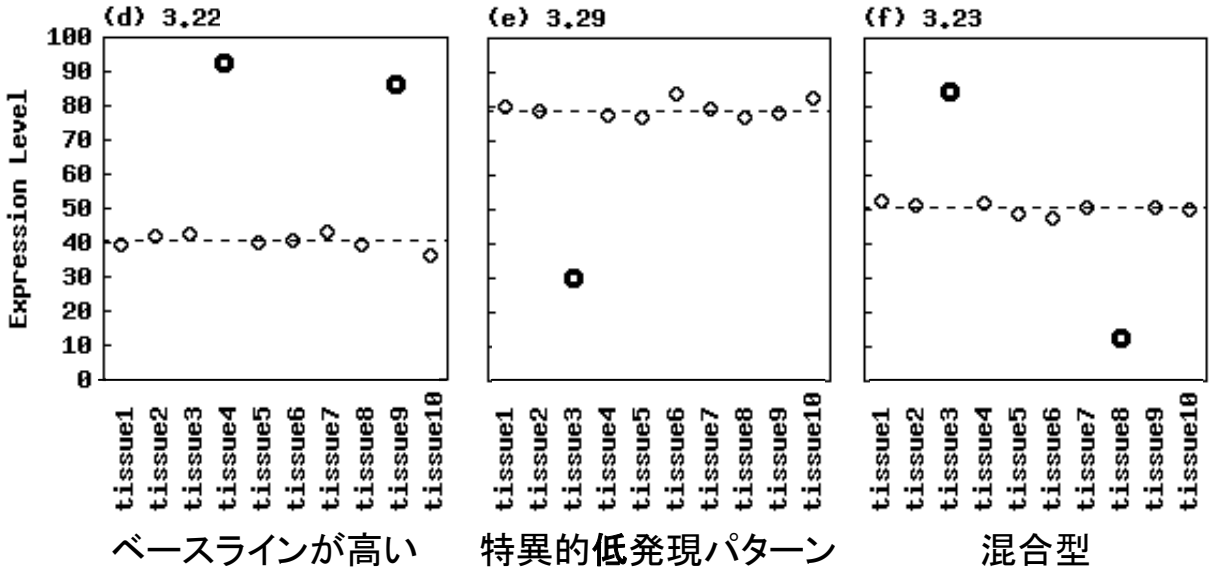
  - $H(\boldsymbol{x})$のとりうる範囲: $0 \leqq H(\boldsymbol{x}) \leqq \log_2(n)$

Tissue-specific gene **Housekeeping gene**

統計量の小さい遺伝子を抽出

エントロピー*H*(とτ)の欠点

- 単純にエントロピーの低いもの(or τ の高いもの)を検出する方法だと、以下のような特異的発現パターンを検出するのが困難(上位にランクインしないから)。



$H(x)$のとりうる範囲:
$0 \leqq H(x) \leqq \log2(n)$

$n = 10$
$\rightarrow 3.32$

ランキングの改良

- 組織特異的低発現パターンなど様々な特異的発現パターンも含めて統一的にランキング可能にしたい
  - 遺伝子発現ベクトルxを変換（x → x'）

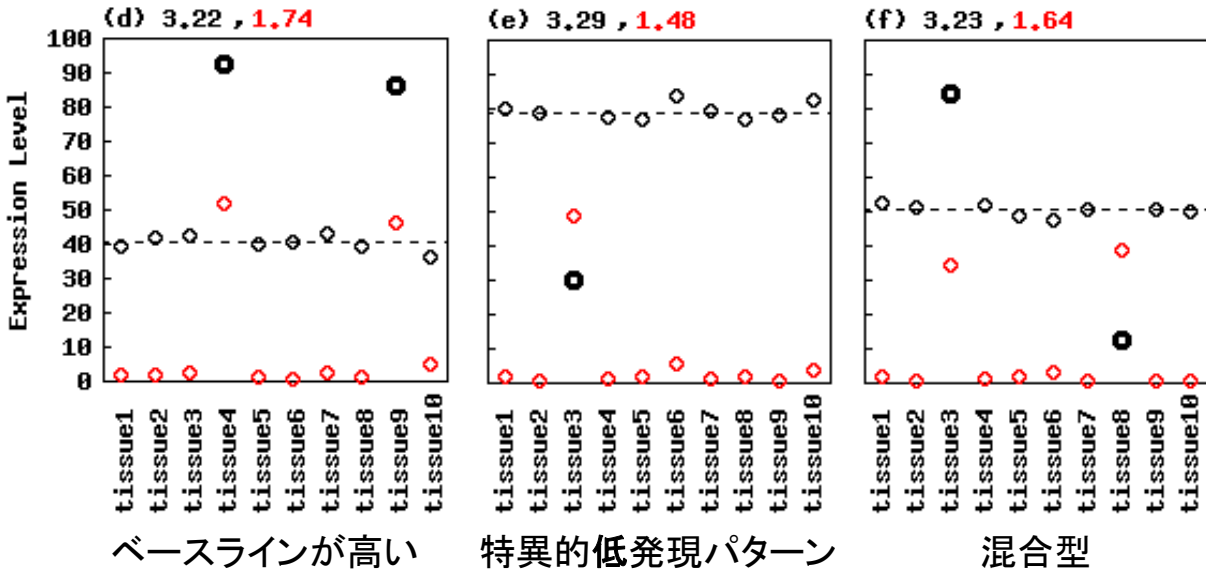
$$x_i' = \left| x_i - T_{bw} \right|$$

  - $H(x')$でランキング

→一応目的達成

点線：T_{bw}の値

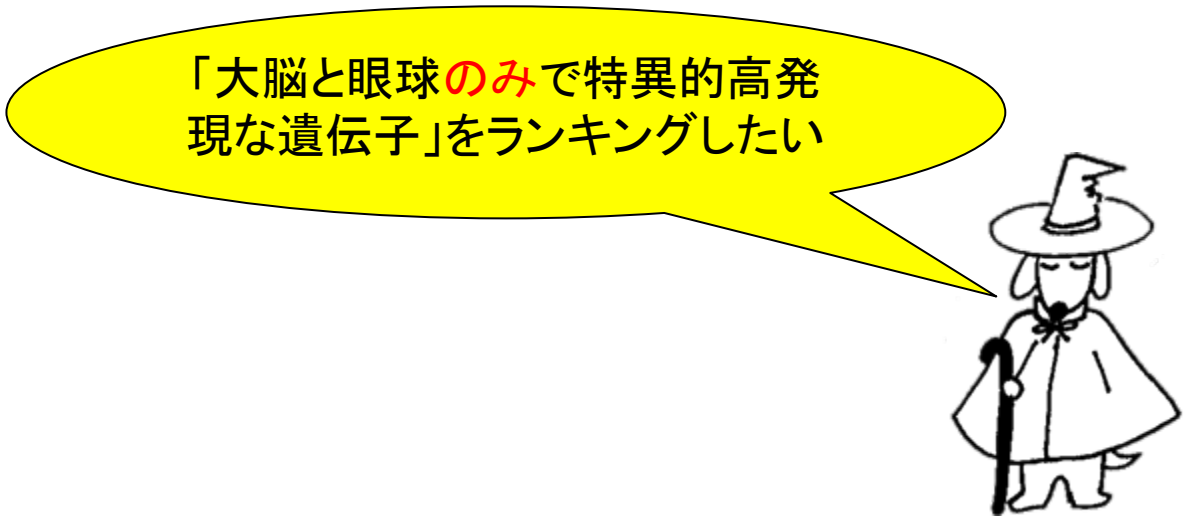
赤丸：変換後の各要素（x_i'）の値



しかし！

$H(x')$の問題点（$H(x)$と$\tau(x)$含む）

- ランク上位　→　様々なタイプの組織特異的（選択的）遺伝子
- どの組織で高発現（and/or 低発現）かは、教えてくれない



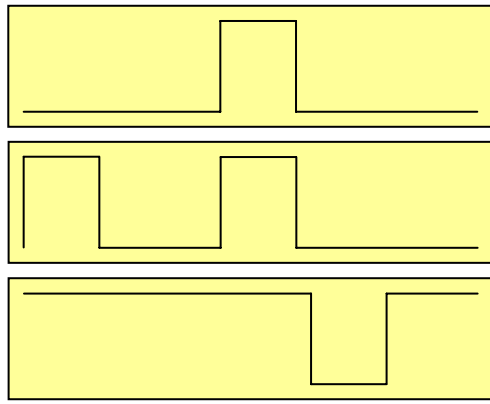
ランキングに基づく方法のみでは不十分

組織特異的発現遺伝子検出法

- ランキングに基づく方法
  - **Dixon test**（Greller and Tobin, *Genome Res.*, 1999）
  - **Pattern matching**（Pavlidis and Noble, *Genome Biol.*, 2001）
  - **Tissue specificity Index**（Yanai *et al.*, *Bioinformatics*, 2005）
  - **Entropy**（Schug *et al.*, *Genome Biol.*, 2005）
  - **Tukey-Kramer's Honest Significance Difference (HSD) test**（Liang *et al.*, *Physiol. Genomics*, 2006）
- 外れ値検出に基づく方法
  - **Akaike's Information Criterion (AIC)**（Kadota *et al.*, *Physiol. Genomics*, 2003）
  - **Sprent's non-parametric method**（Ge *et al.*, *Genomics*, 2005）
- 組み合わせ（AIC + a modified entropy）
  - **ROKU** (Kadota *et al.*, *BMC Bioinformatics*, 2006)

様々な組織（条件）

	S1	S2	S3	S4	…
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 *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



外れ値検出に基づく方法

- **Akaike's Information Criterion (AIC)**（Kadota *et al.*, *Physiol. Genomics*, 2003）
- **Sprent's non-parametric method**（Ge *et al.*, *Genomics*, 2005）
  - 長所：目的組織**のみ**で発現の異なる遺伝子群を一意に抽出可能

http://www.aist.go.jp/aist_j/aistinfo/aist_today/vol03_06/vol03_06_p28.pdf

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

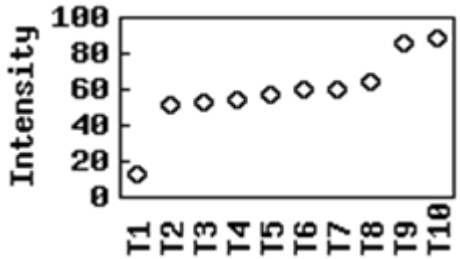
外れ値検出に基づく方法1

- 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

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
↓										
出力結果	-1	0	0	0	0	0	0	0	1	1



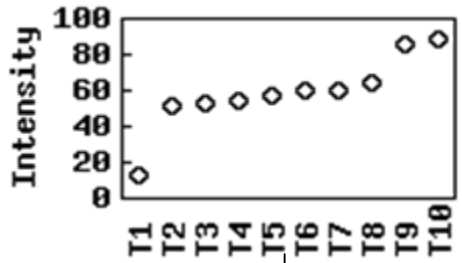
Kadota *et al.*, *Physiol. Genomics*, 2003

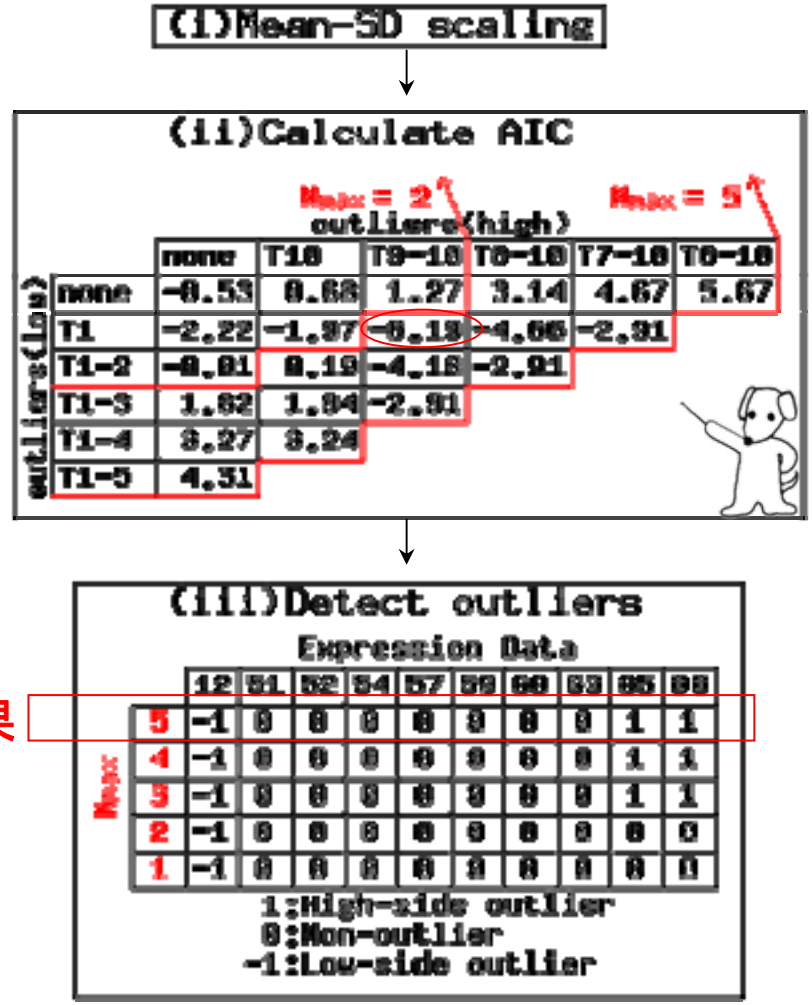
AIC計算例

- x = (12, 51, 52, 54, 57, 59, 60, 63, 85, 88)
- 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$Nmax = n/2 = 5$ (原著論文)

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

$n_n + n_o (= n)$: 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





デフォルトの結果

Kadota *et al.*, *Physiol. Genomics*, 2003

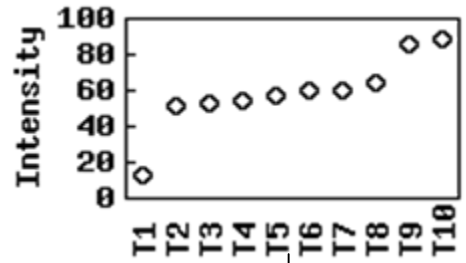
AIC計算例

- $\boldsymbol{x}$ = (12, 51, 52, 54, 57, 59, 60, 63, 85, 88)
- 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　(原著論文)

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

$n_n + n_o (= n)$: 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

*Nmax*が変わると得られる結果が異なることには論文中では触れられていない



(i)Mean-SD scaling

(ii)Calculate AIC

N_{max} = 2　　N_{max} = 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	-1.37	-0.13	-4.66	-2.91	
T1-2	-0.01	0.19	-4.16	-2.91		
T1-3	1.62	1.84	-2.91			
T1-4	3.27	3.24				
T1-5	4.31					

(iii)Detect outliers

Expression Data

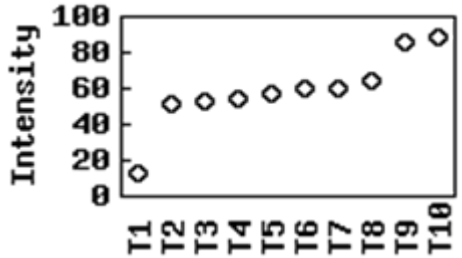
Nmax	12	51	52	54	57	59	60	63	85	88
5	-1	0	0	0	0	0	0	0	1	1
4	-1	0	0	0	0	0	0	0	1	1
3	-1	0	0	0	0	0	0	0	1	1
2	-1	0	0	0	0	0	0	0	0	0
1	-1	0	0	0	0	0	0	0	0	0

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

Nmax = 2

Ge *et al.*, *Genomics*, 2005

外れ値検出に基づく方法2

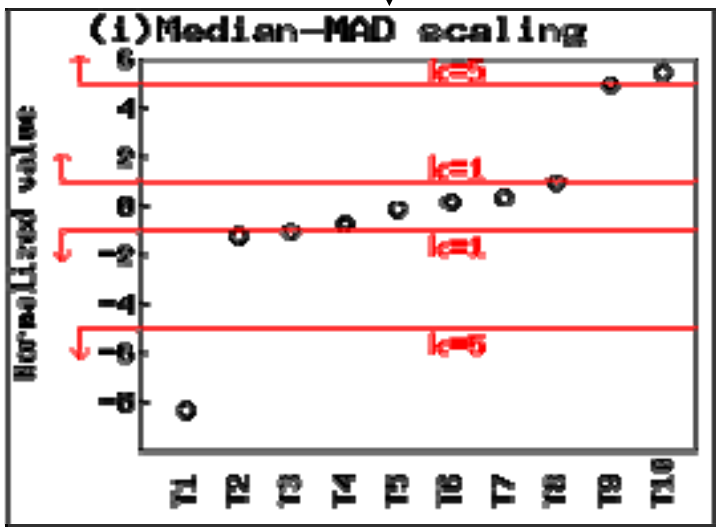


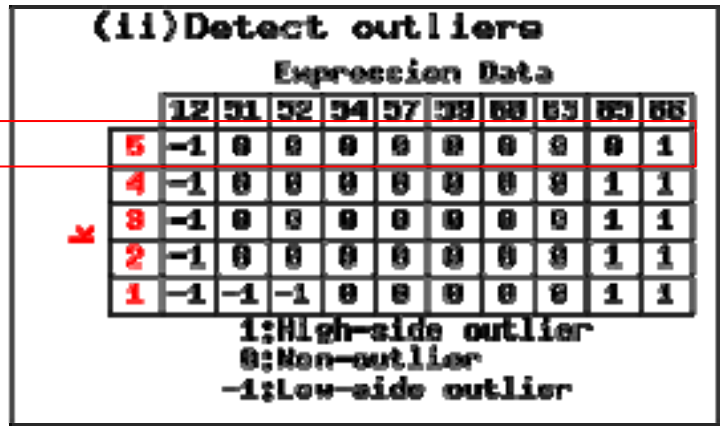
- Sprent's non-parametric method
  - 遺伝子発現ベクトル$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$に対して、

    $x_i < \mathrm{median}(\boldsymbol{x}) - k \times \mathrm{MAD}(\boldsymbol{x})$ and

    $x_i > \mathrm{median}(\boldsymbol{x}) + k \times \mathrm{MAD}(\boldsymbol{x})$

    を満たすx_iを外れ値とする
  - $k = 5$ (原著論文)

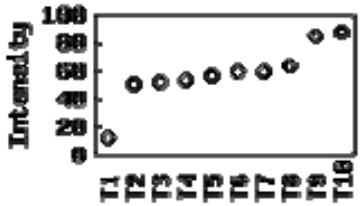




デフォルトの結果

kが変わると得られる結果が異なることには論文中では触れられていない

どっちが正しい？！



(a) AIC

(i)Mean-SD scaling

(ii)Calculate AIC

Max = 2

Max = 5

outliers(high)

outliers(low)	none	T10	T9-10	T8-10	T7-10	T6-10
none	-0.53	0.09	1.27	3.14	4.87	5.07
T1	-2.22	-1.87	-0.19	-4.66	-2.81	
T1-2	-0.01	0.13	-4.18	-2.31		
T1-3	1.82	1.34	-2.31			
T1-4	3.27	3.24				
T1-5	4.31					

(iii)Detect outliers

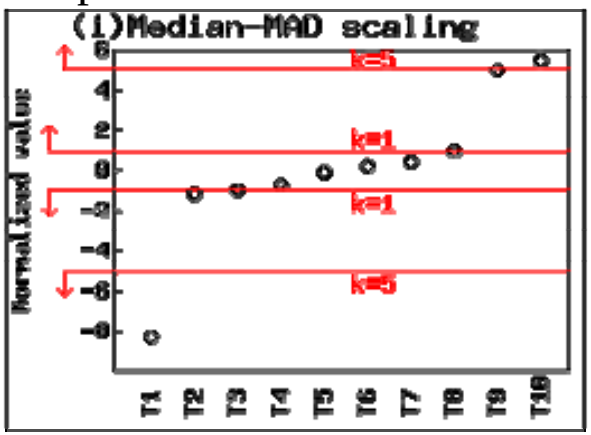
Expression Data

Max	12	51	52	54	57	58	60	63	65	68
5	-1	0	0	0	0	0	0	0	1	1
4	-1	0	0	0	0	0	0	0	1	1
3	-1	0	0	0	0	0	0	0	1	1
2	-1	0	0	0	0	0	0	0	0	0
1	-1	0	0	0	0	0	0	0	0	0

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

(b) Sprent's method

(i)Median-MAD scaling



(ii)Detect outliers

Expression Data

k	12	51	52	54	57	58	60	63	65	68
5	-1	0	0	0	0	0	0	0	0	1
4	-1	0	0	0	0	0	0	0	1	1
3	-1	0	0	0	0	0	0	0	1	1
2	-1	0	0	0	0	0	0	0	1	1
1	-1	-1	-1	0	0	0	0	0	1	1

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

デフォルトの結果

実際のマイクロアレイデータに対していくら適用しようとも…

AIC
Kadota *et al.*, *Physiol. Genomics*, 2003

48組織

14,610 clones

	S1	S2	S3	S4	…
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 *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

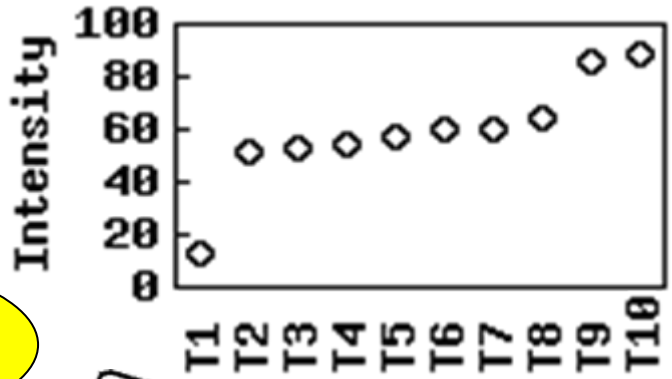
Sprent's non-parametric method
Ge *et al.*, *Genomics*, 2005 (GSE2361)

36組織

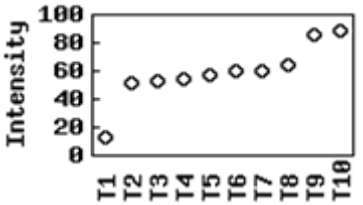
< 22,283 probesets

	S1	S2	S3	S4	…
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 *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

答えは分かりません（たぶん）！



ユーザー側の希望1



- パラメータの変更（Nmax or k）に対して頑健

(a) AIC

(i)Mean-SD scaling

(ii)Calculate AIC

Nmax = 2

Nmax = 3

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	-1.07	-0.19	-4.66	-2.01	
T1-2	-0.01	0.19	-4.18	-2.01		
T1-3	1.22	1.94	-2.01			
T1-4	3.27	3.24				
T1-5	4.31					

(iii)Detect outliers

Expression Data

Nmax	12	51	52	54	57	59	60	63	65	68
5	-1	0	0	0	0	0	0	0	1	1
4	-1	0	0	0	0	0	0	0	1	1
3	-1	0	0	0	0	0	0	0	1	1
2	-1	0	0	0	0	0	0	0	0	0
1	-1	0	0	0	0	0	0	0	0	0

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

(b) Sprent's method

(i)Median-MAD scaling

Normalized value

k=3
k=1
k=-1
k=-3

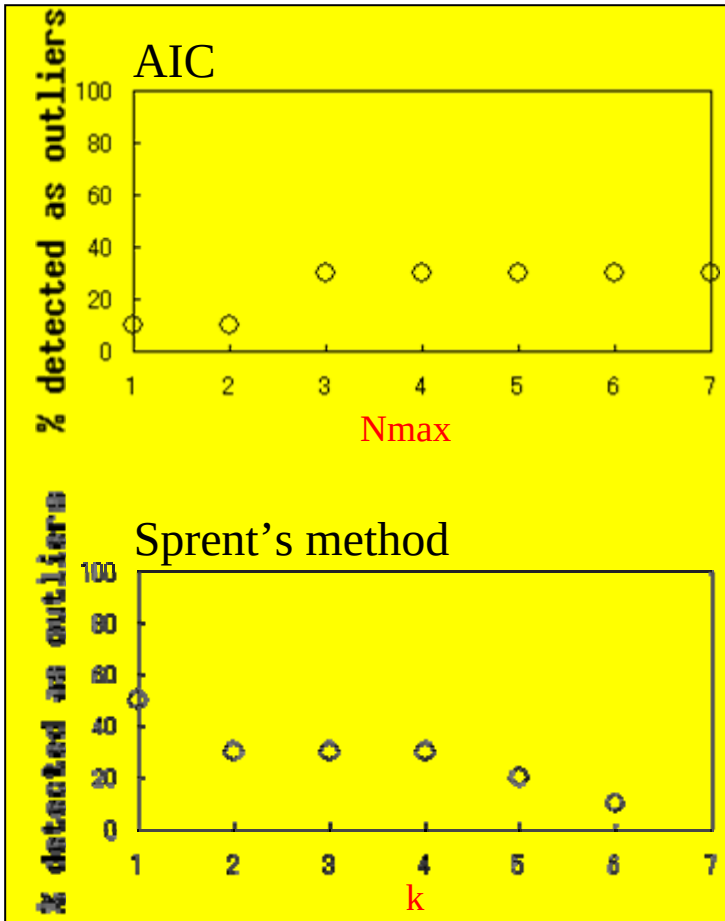
T1 T2 T3 T4 T5 T6 T7 T8 T9 T10

(ii)Detect outliers

Expression Data

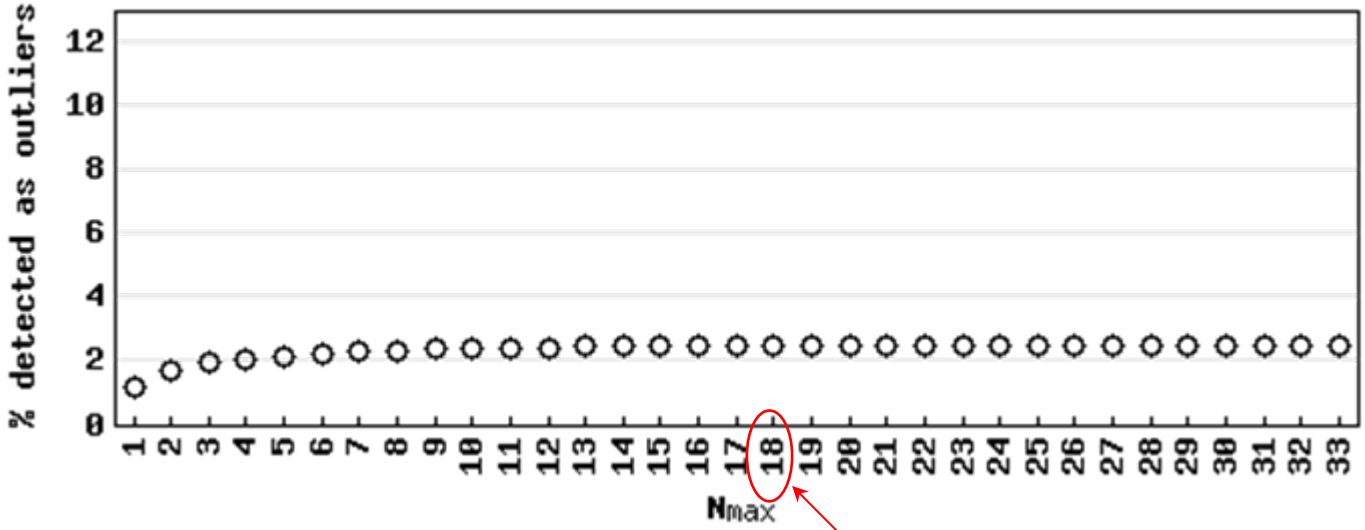
k	12	51	52	54	57	59	60	63	65	68
5	-1	0	0	0	0	0	0	0	0	1
4	-1	0	0	0	0	0	0	0	1	1
3	-1	0	0	0	0	0	0	0	1	1
2	-1	0	0	0	0	0	0	0	1	1
1	-1	-1	-1	0	0	0	0	0	1	1

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



実データ(36組織×22283 clones)比較結果

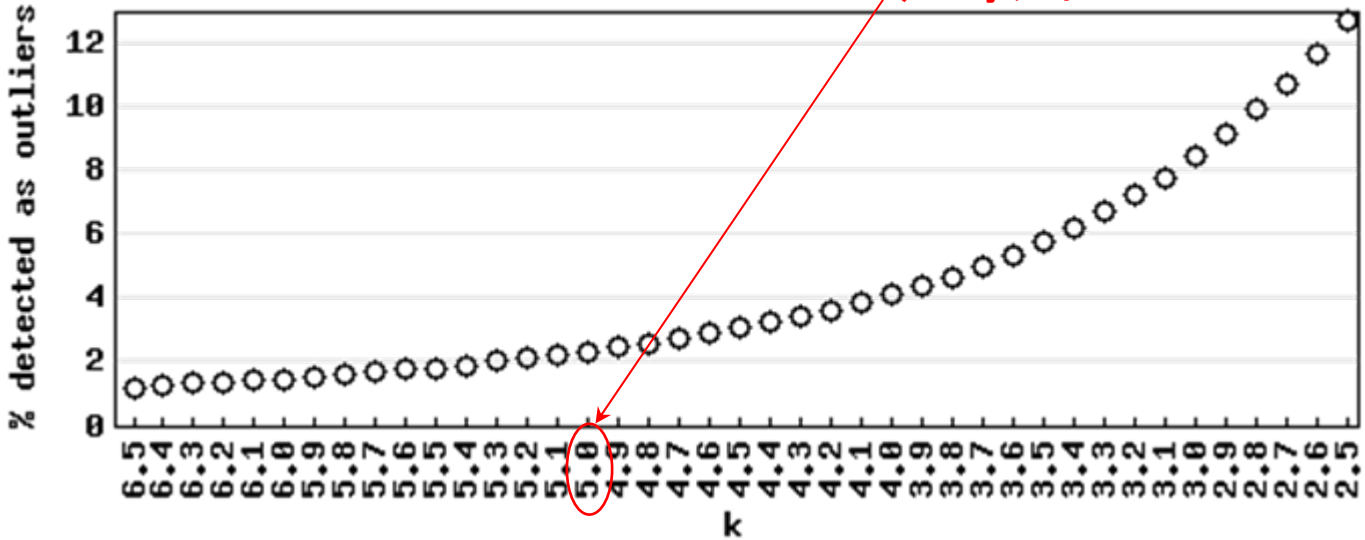
(a) AIC



AIC

Nmaxが変わっても結果「外れ値の割合（縦軸）」があまり変わらない

(b) Sprent's method

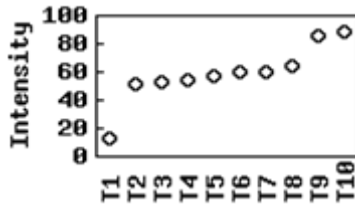


デフォルト

Sprent's method

kが変わると得られる結果もどんどん変わってしまう

ユーザー側の希望2



- 実験データの（追加や）削除に対して頑健

(a) AIC

Tissue	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Intensity	12	51	52	54	57	59	59	59	85	86

Output using N (= 10) observations

-1	0	0	0	0	0	0	0	1	1

Output using (N-1) observations

T1:		0	0	0	0	0	0	0	1	1
T2:	-1		0	0	0	0	0	0	1	1
T3:	-1	0		0	0	0	0	0	1	1
T4:	-1	0	0		0	0	0	0	1	1
T5:	-1	0	0	0		0	0	0	1	1
T6:	-1	0	0	0	0		0	0	1	1
T7:	-1	0	0	0	0	0		0	1	1
T8:	-1	0	0	0	0	0	0		1	1
T9:	-1	0	0	0	0	0	0	0		1
T10:	-1	0	0	0	0	0	0	0	1	

True-Positive(TP)
False-Negative(FN)
False_Positive(FP)
True-Negative(TN)

Accuracy = 100%
MCC = 100%

(b) Sprent's method

Tissue	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Intensity	12	51	52	54	57	59	60	63	85	86

Output using N (= 10) observations

-1	0	0	0	0	0	0	0	0	1

Output using (N-1) observations

T1:		0	0	0	0	0	0	0	1	1
T2:	-1		0	0	0	0	0	0	1	1
T3:	-1	0		0	0	0	0	0	1	1
T4:	-1	0	0		0	0	0	0	0	0
T5:	-1	0	0	0		0	0	0	0	0
T6:	-1	0	0	0	0		0	0	0	1
T7:	-1	0	0	0	0	0		0	0	1
T8:	-1	0	0	0	0	0	0		1	1
T9:	-1	0	0	0	0	0	0	0		1
T10:	-1	0	0	0	0	0	0	0	1	

True-Positive(TP)
False-Negative(FN)
False_Positive(FP)
True-Negative(TN)

Accuracy = 92.2%
MCC = 77.5%

1.それぞれのデフォルトの結果を真実と仮定

2.一サンプルづつ除いた場合の結果と比較

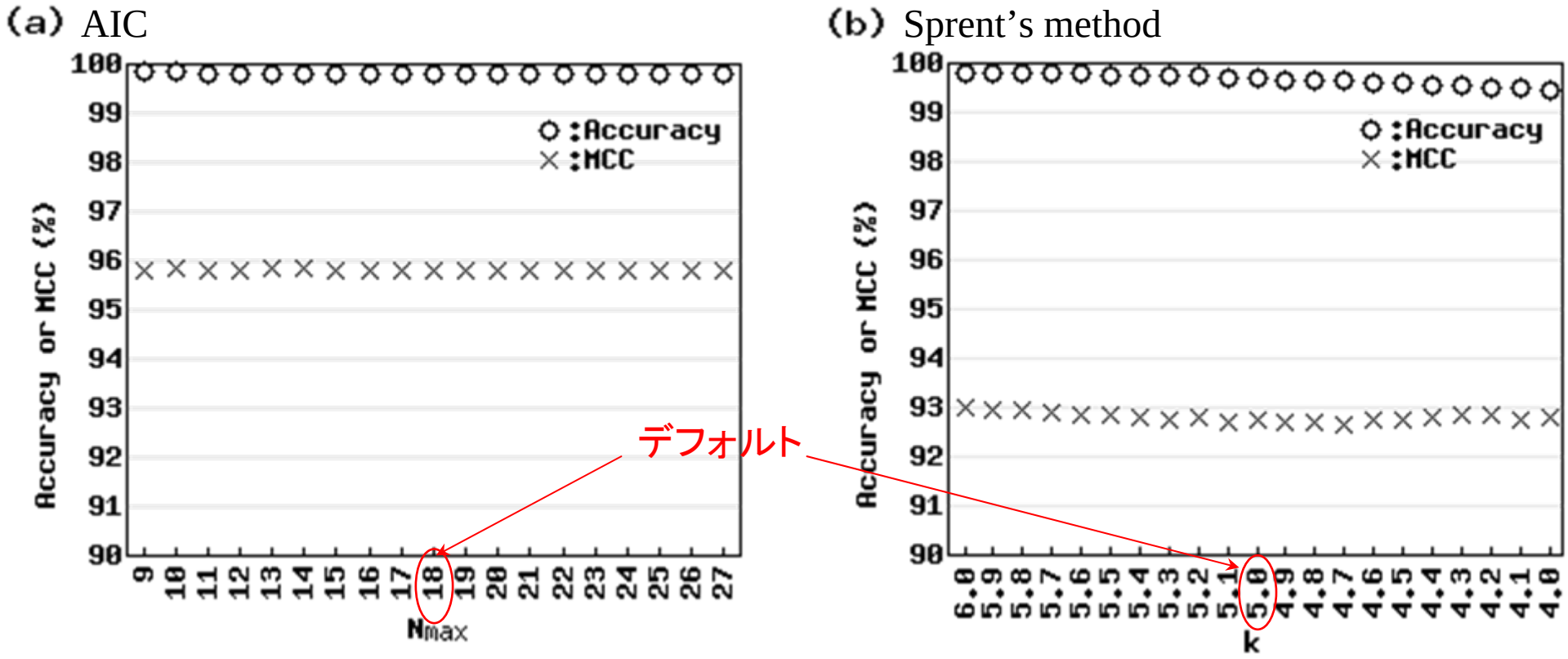
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

Kadota *et al.*, *Gene Regulation and Systems Biol.*, 1: 9-15, 2007

実験データの（追加や）削除に対して頑健

- 実データ(36組織×22283 clones)比較結果



結論：AICのほうがベター

外れ値検出に基づく方法

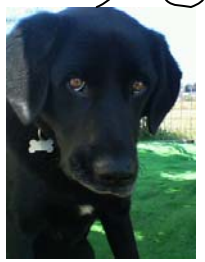
- **Akaike's Information Criterion (AIC)**（Kadota *et al.*, *Physiol. Genomics*, 2003）
- **Sprent's non-parametric method**（Ge *et al.*, *Genomics*, 2005）
  - 長所：目的組織のみで発現の異なる遺伝子群を一意に抽出可能
  - 短所：選ばれたサブセット内のランキングが難しい

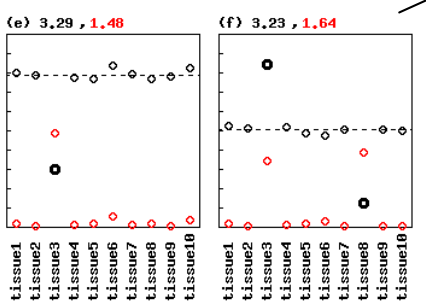
http://www.aist.go.jp/aist_j/aistinfo/aist_today/vol03_06/vol03_06_p28.pdf

組織特異的発現遺伝子検出法（まとめ）

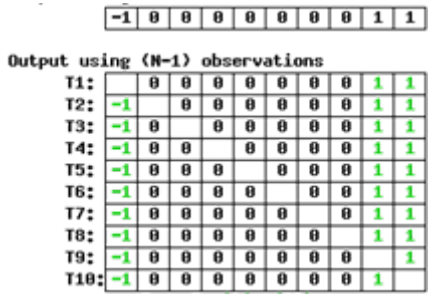
方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
Dixon test	No	No	?	Yes	-
Pattern matching	Yes	Yes	No	Yes	-
Tissue specificity index	Yes	No	-	Yes	-
Entropy (H)	Yes	No	-	Yes	-
Tukey-Kramer's HSD test	Yes	No	?	Yes	-
AIC	Yes	Yes	Yes	No	Yes
Sprent's method	Yes	Yes	Yes	No	No
ROKU (AIC+a modified H)	Yes	Yes	Yes	Yes	Yes

ロク





http://www.aist.go.jp/aist_j/aistinfo/aist_today/vol03_06/vol03_06_p28.pdf



本日のお題

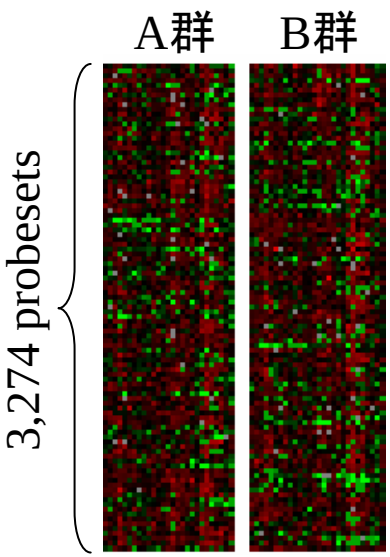
- 「マイクロアレイ解析手法あれこれ」
  - 組織特異的発現遺伝子検出法
  - 二群間で発現の異なる遺伝子検出法

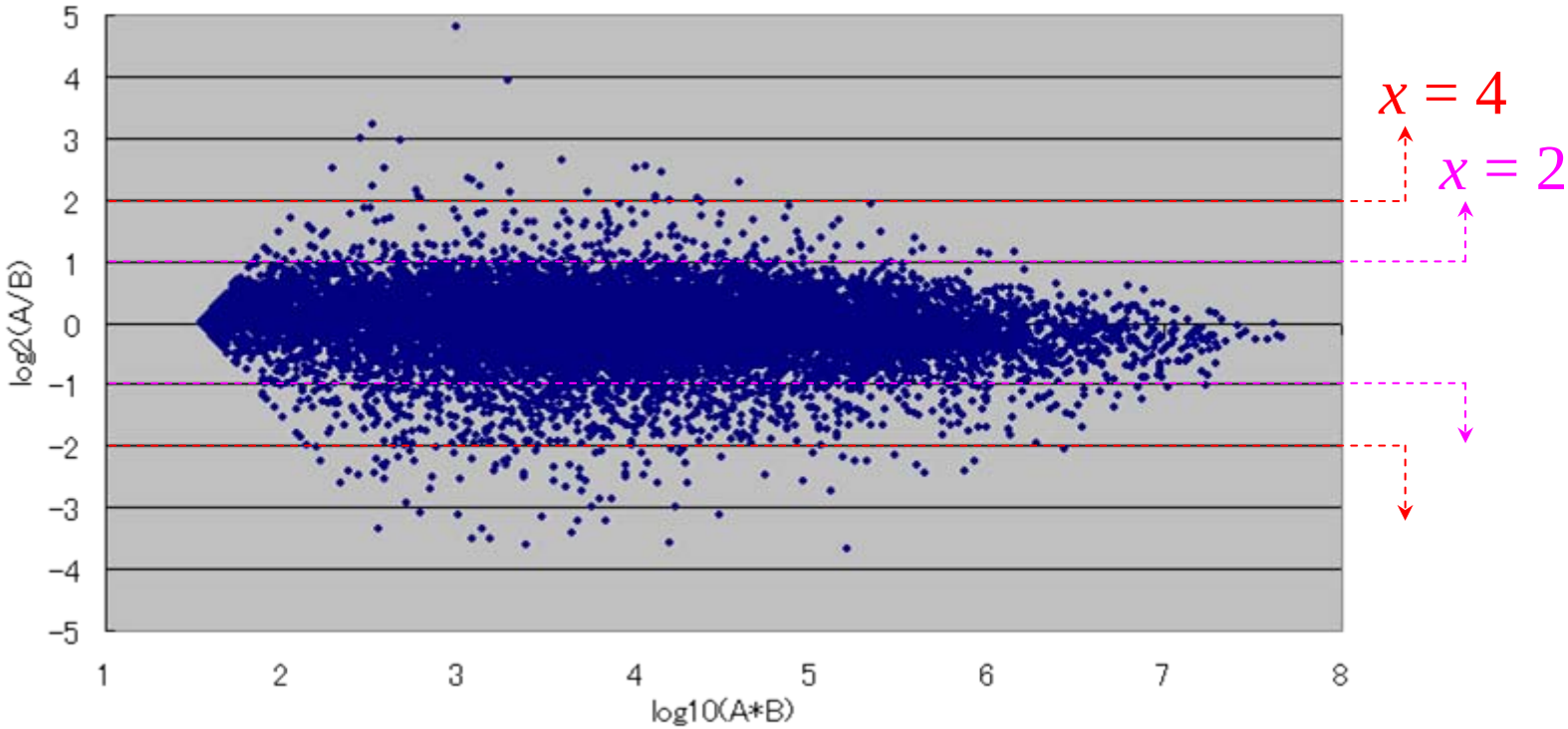
二群間で発現の異なる遺伝子検出法

- 倍率変化（Fold change; FC）に基づく方法
  - 2-fold, 3-fold
  - The limit fold change model (Mutch *et al*., *BMC Bioinformatics*, 2002)
  - Rank product (Breitling *et al*., *FEBS Lett*., 2004)
  - …
- *t*-statisticsに基づく方法
  - Student’s (or Welch) *t*-test
  - SAM (Tusher *et al*., *PNAS*, 2001)
  - Samroc (Broberg, P., *Genome Biol*., 2003)
  - Empirical bayes (Smyth, GK., *Stat. Appl. Genet. Mol. Biol*., 2004)
  - …
- その他
  - Rank difference analysis of microarrays (RDAM; Martin *et al*., *BMC Bioinformatics*, 2004)
  - Correspondence analysis (CA; Yano *et al*., *Nucleic Acids Res*., 2006)
  - …

二群間で発現の異なる遺伝子検出法

- 倍率変化（Fold change; FC）に基づく方法の問題
  - *x*-foldの*x*の根拠が希薄
  - 発現量の低い遺伝子群（S/N比が低い）が検出される傾向
    - Relative error increases at lower intensities (Quackenbush J., *Nat. Genet.*, 2002)
    - A greater inherent error in their measured levels (Mutch *et al.*, *BMC Bioinformatics*, 2002)



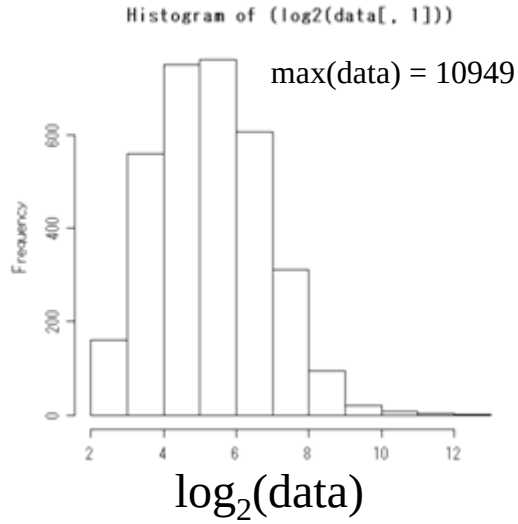


二群間で発現の異なる遺伝子検出法

- 「倍率変化（Fold change; FC）に基づく方法の問題」に対する対策
  - 発現強度依存の偏りを補正
    - The limit fold change model (Mutch *et al.*, *BMC Bioinformatics*, 2002)
    - Intensity-dependent *z*-score (Quackenbush J., *Nat. Genet.*, 2002)
    - …

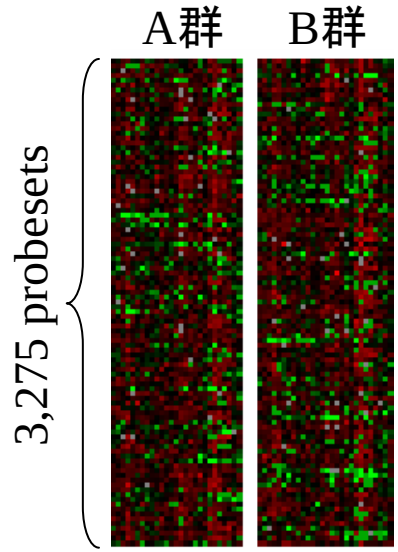
二群間で発現の異なる遺伝子検出法

- 「(classical) *t*-statisticsに基づく方法」の問題
  - 発現量の低い遺伝子群（S/N比が低い）が検出される傾向



rank(| *t* |)

	A群					B群					
ProbeSetID	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	*t*
31850_at	14	12	12	14	14	16	18	17	17	16	1
1647_at	19	16	16	21	20	25	37	35	31	39	2
1513_at	109	96	92	101	103	121	164	164	133	164	3
34280_at	31	28	30	30	30	27	20	20	24	23	4
37543_at	17	16	19	16	16	14	13	14	14	14	5
38312_at	32	27	29	31	35	23	18	20	26	19	6
1228_s_at	42	47	51	57	51	62	76	73	85	65	7
36090_at	15	15	15	15	15	16	17	17	19	17	8
40767_at	17	16	21	21	20	13	13	10	8	15	9
1954_at	17	15	17	16	16	15	12	13	14	13	10
38342_at	97	85	97	96	105	87	76	79	82	78	11
38042_at	12	11	13	13	12	11	10	10	10	11	12

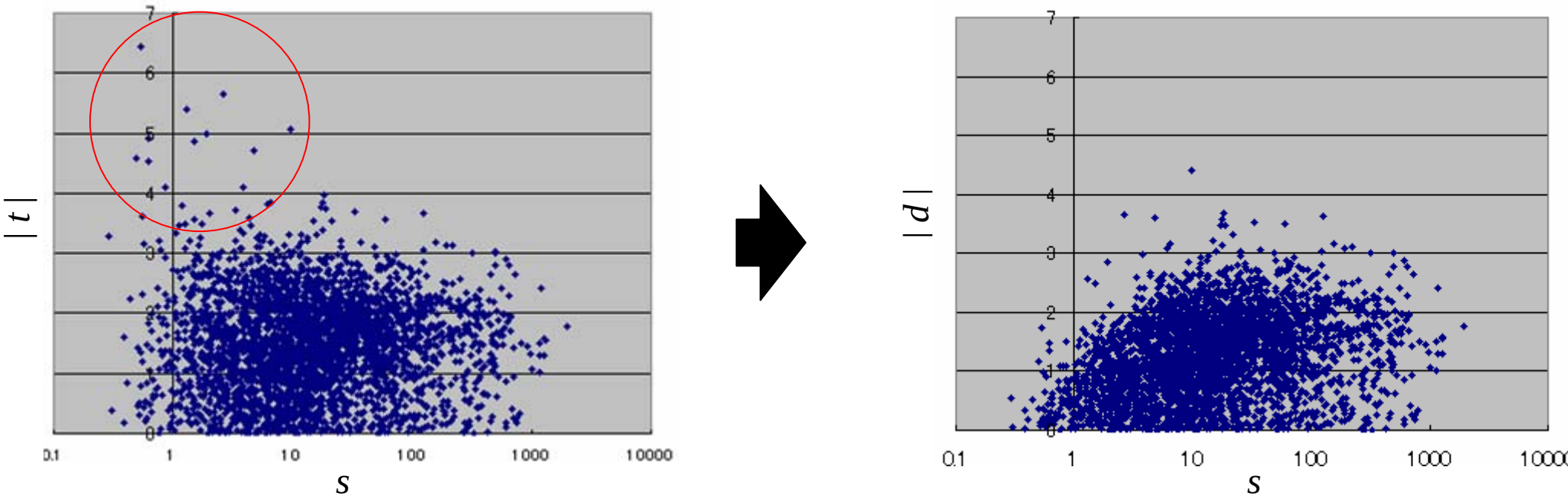


二群間で発現の異なる遺伝子検出法

- 改良版*t*-statistic
  - Significance Analysis of Microarrays (SAM)

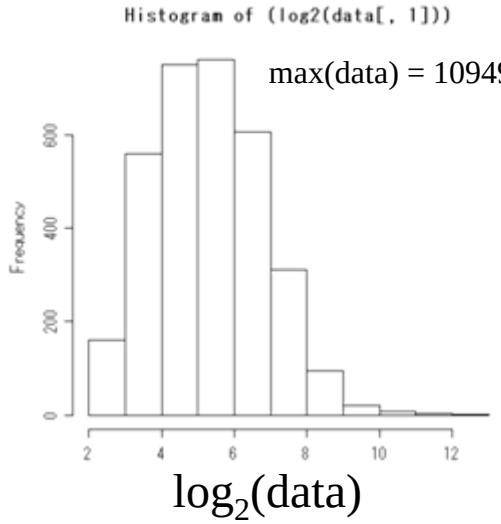
    分母に定数項(fugde factor) s_0を付加

$$d = \frac{\overline{A} - \overline{B}}{s + s_0} \qquad \left[s_0 = 0 \rightarrow \text{Student's } t \text{ test} \quad t = \frac{\overline{A} - \overline{B}}{s} \right]$$



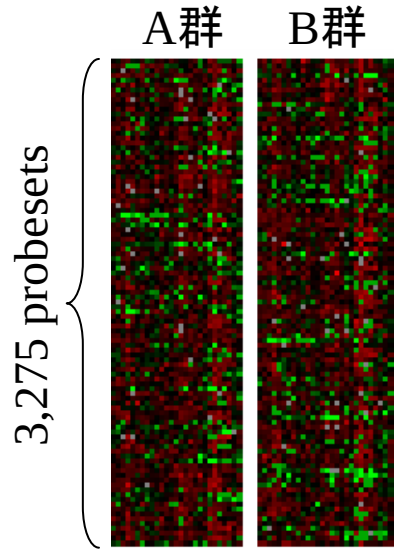
二群間で発現の異なる遺伝子検出法

- SAMの効果
  - 「発現量の低い遺伝子群（S/N比が低い）が検出される傾向」が緩和



rank(| *d* |)

	A群					B群						
ProbeSetID	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	*t*	*d*
31850_at	14	12	12	14	14	16	18	17	17	16	1	854
1647_at	19	16	16	21	20	25	37	35	31	39	2	6
1513_at	109	96	92	101	103	121	164	164	133	164	3	2
34280_at	31	28	30	30	30	27	20	20	24	23	4	146
37543_at	17	16	19	16	16	14	13	14	14	14	5	1337
38312_at	32	27	29	31	35	23	18	20	26	19	6	46
1228_s_at	42	47	51	57	51	62	76	73	85	65	7	5
36090_at	15	15	15	15	15	16	17	17	19	17	8	1834
40767_at	17	16	21	21	20	13	13	10	8	15	9	205
1954_at	17	15	17	16	16	15	12	13	14	13	10	1495
38342_at	97	85	97	96	105	87	76	79	82	78	11	26
38042_at	12	11	13	13	12	11	10	10	10	11	12	2136



あやしげ？！な候補の順位を下げてくれる

二群間で発現の異なる遺伝子検出法

- SAMの効果
  - 「発現量の低い遺伝子群（S/N比が低い）が検出される傾向」が緩和

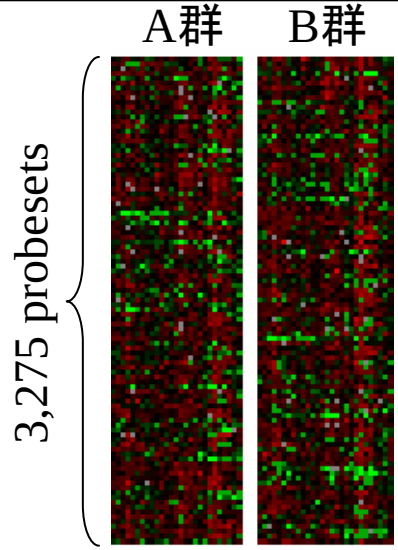
$$d = \frac{\overline{A} - \overline{B}}{s + s_0}$$

ProbeSetID	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	*t*	*d*
omake	5000	5000	5000	5000	5000	10	10	10	10	10	3275	1
1513_at	109	96	92	101	103	121	164	164	133	164	3	2
36378_at	209	249	191	186	227	195	127	138	127	118	13	3
37639_at	245	216	217	190	133	231	769	562	949	801	21	4
1228_s_at	42	47	51	57	51	62	76	73	85	65	7	5
1647_at	19	16	16	21	20	25	37	35	31	39	2	6
692_s_at	226	276	252	291	279	238	160	181	214	185	15	7
37669_s_at	275	292	303	345	425	249	152	201	183	250	22	8
40327_at	106	100	91	103	161	96	345	400	384	403	25	9
37742_at	105	97	99	111	140	116	215	182	201	193	19	10
37730_at	87	85	90	98	105	103	166	139	187	195	18	11
40140_at	131	114	128	127	125	123	223	226	190	244	29	12

副次的な効果:

「分母が0 → $|t| = \infty$ となって計算不能だった遺伝子」

に対しても適切？！な統計量を与えることが可能に



二群間で発現の異なる遺伝子検出法

- 倍率変化（Fold change; FC）に基づく方法
  - 2-fold, 3-fold
  - The limit fold change model (Mutch *et al*., *BMC Bioinformatics*, 2002)
  - **Rank product (Breitling *et al*., *FEBS Lett*., 2004)**
  - …
- *t*-statisticsに基づく方法
  - Student’s (or Welch) *t*-test
  - SAM (Tusher *et al*., *PNAS*, 2001)
  - Samroc (Broberg, P., *Genome Biol*., 2003)
  - Empirical bayes (Smyth, GK., *Stat. Appl. Genet. Mol. Biol*., 2004)
  - …
- その他
  - Rank difference analysis of microarrays (RDAM; Martin *et al*., *BMC Bioinformatics*, 2004)
  - Correspondence analysis (CA; Yano *et al*., *Nucleic Acids Res*., 2006)
  - …

二群間で発現の異なる遺伝子検出法

- 倍率変化（Fold change; FC）に基づく方法
  - Rank products

入力データ

	A1	A2	A3	B1	B2	B3
gene1	a11	a12	a13	b11	b12	b13
…	…	…	…	…	…	…
genei	ai1	ai2	ai3	bi1	bi2	bi3
…	…	…	…	…	…	…
genen	an1	an2	an3	bn1	bn2	bn3

$n_A = 3$　$n_B = 3$

総当りの発現比を計算

$(n_A \times n_B) = 9$通り

A1/B1	A1/B2	A1/B3	A2/B1	A2/B2	A2/B3	A3/B1	A3/B2	A3/B3
a11/b11	a11/b12	a11/b13	a12/b11	a12/b12	a12/b13	a13/b11	a13/b12	a13/b13
ai1/bi1	ai1/bi2	ai1/bi3	ai2/bi1	ai2/bi2	ai2/bi3	ai3/bi1	ai3/bi2	ai3/bi3
an1/bn1	an1/bn2	an1/bn3	an2/bn1	an2/bn2	an2/bn3	an3/bn1	an3/bn2	an3/bn3

列ごとにRankを計算した後、各行に対して相乗平均値（RPs）を計算

	RP
gene1	RP1
…	
genei	RPi
…	
genen	RPn

二群間で発現の異なる遺伝子検出法

- 倍率変化（Fold change; FC）に基づく方法
  - Rank products (RP)

SAM

t-test

ProbeSetID	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	*t*	*d*	*FC*	*RP*
omake	5000	5000	5000	5000	5000	10	10	10	10	10	3275	1	1	1
1513_at	109	96	92	101	103	121	164	164	133	164	3	2	1320	1830
36378_at	209	249	191	186	227	195	127	138	127	118	13	3	1285	180
37639_at	245	216	217	190	133	231	769	562	949	801	21	4	64	88
1228_s_at	42	47	51	57	51	62	76	73	85	65	7	5	1385	1896
1647_at	19	16	16	21	20	25	37	35	31	39	2	6	713	846
692_s_at	226	276	252	291	279	238	160	181	214	185	15	7	1699	384
37669_s_at	275	292	303	345	425	249	152	201	183	250	22	8	1123	143
40327_at	106	100	91	103	161	96	345	400	384	403	25	9	111	242
37742_at	105	97	99	111	140	116	215	182	201	193	19	10	999	1772
37730_at	87	85	90	98	105	103	166	139	187	195	18	11	908	1481
40140_at	131	114	128	127	125	123	223	226	190	244	29	12	1062	1866

「Rank productsはSAMの結果とよく似ていた」と原著論文には書いてあったが…

二群間で発現の異なる遺伝子検出法

- 上位100遺伝子のオーバーラップ

60%以上重なっているので確かに似た結果を返す

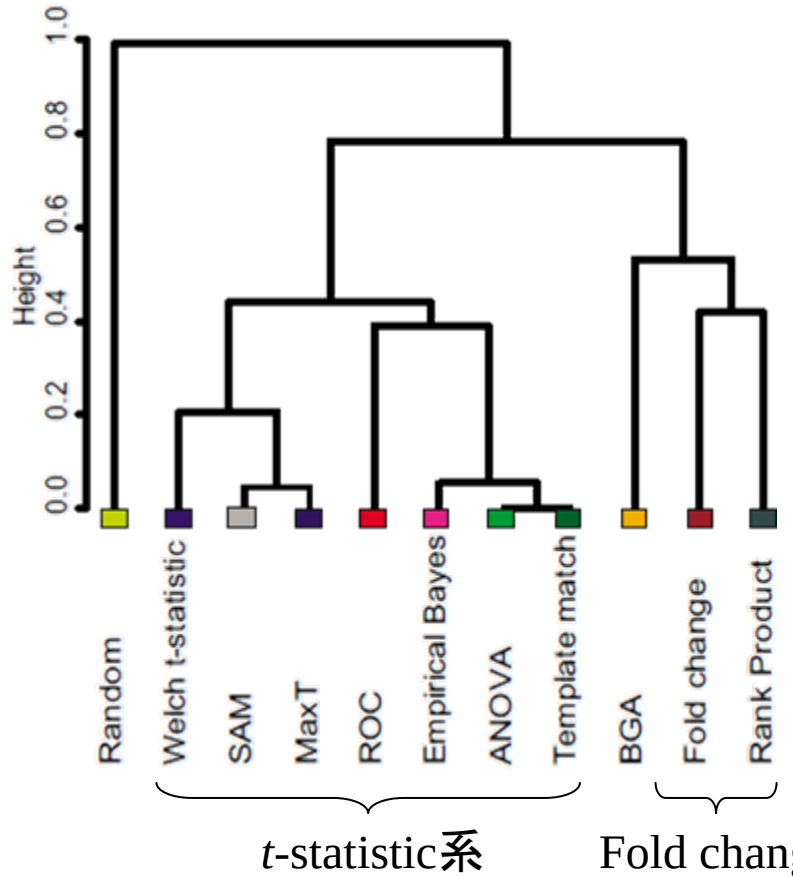
二群間で発現の異なる遺伝子検出法

- 上位100遺伝子のオーバーラップ

用いるデータセットによってはかなり違った結果を返す

二群間で発現の異なる遺伝子検出法

- 上位100個のoverlap
  - 9 datasetsの平均

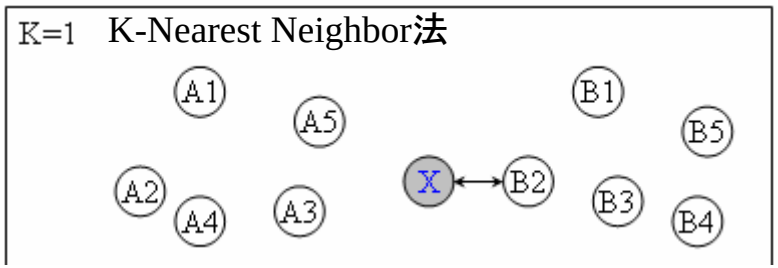


Jefferyらの結論と同じ

Jeffery *et al.*, *BMC Bioinformatics*, 2006
Breitling *et al.*, *FEBS Lett.*, 2004

二群間で発現の異なる遺伝子検出法

- 比較のための評価基準
  - 分類精度の高さ
    - SVM
    - K-NN
    - Naïve Bayes
    - *etc…*
  - 実験データの（追加や）削除に対して頑健かどうか



ProbeSetID	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	*RP*
omake	5000	5000	5000	5000	5000	10	10	10	10	10	1
1514_g_at	63	68	144	58	860	83	3739	1742	1650	2904	2
1805_g_at	32	16	412	659	1892	43	3338	3640	3032	2671	3
1894_f_at	3016	2181	492	444	38	2106	72	584	42	44	4
217_at	34	26	223	312	870	37	2832	1822	1997	2307	5
38604_at	88	35	51	77	937	52	1742	850	2183	3094	6
1008_f_at	4453	3064	1721	1911	177	3872	240	1488	175	176	7
37793_r_at	180	188	147	97	37	175	22	49	34	47	8
41025_r_at	73	59	56	33	17	59	11	16	13	17	9
36423_at	7	8	39	20	71	7	390	241	175	438	10

ProbeSetID	A1	A2	A3	B1	B2	B3	*RP*
omake	5000	5000	5000	10	10	10	1
1514_g_at	63	68	144	83	3739	1742	12
1805_g_at	32	16	412	43	3338	3640	2
1894_f_at	3016	2181	492	2106	72	584	5
217_at	34	26	223	37	2832	1822	10
38604_at	88	35	51	52	1742	850	19
1008_f_at	4453	3064	1721	3872	240	1488	15
37793_r_at	180	188	147	175	22	49	3
41025_r_at	73	59	56	59	11	16	4
36423_at	7	8	39	7	390	241	47

二群間で発現の異なる遺伝子検出法

1. SAM
2. ANOVA
3. Empirical bayes
4. Template matching
5. maxT
6. BGA
7. AUC
8. Welch *t* statistic
9. Fold change
10. Rank products

- 比較のための評価基準1
  - 分類精度の高さ

Empirical bayes

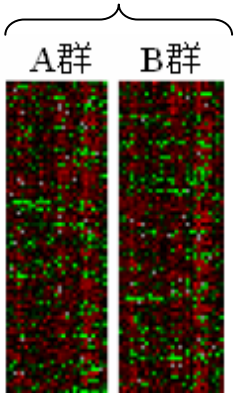
Rank products

Area under a ROC curve (AUC)

10 samples

20 samples

30 samples



Breitling *et al.*, *FEBS Lett.*, 2004

比較のための評価基準2-1

- 実験データの（追加や）削除に対して頑健かどうか
  - Complete setでのランキング上位*x*個の順位のmedian がsubsetでどれだけ保持されているか

Complete set

ProbeSetID	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	*RP*
omake	5000	5000	5000	5000	5000	10	10	10	10	10	1
1514_g_at	63	68	144	58	860	83	3739	1742	1650	2904	2
1805_g_at	32	16	412	659	1892	43	3338	3640	3032	2671	3
1894_f_at	3016	2181	492	444	38	2106	72	584	42	44	4
217_at	34	26	223	312	870	37	2832	1822	1997	2307	5
38604_at	88	35	51	77	937	52	1742	850	2183	3094	6
1008_f_at	4453	3064	1721	1911	177	3872	240	1488	175	176	7
37793_r_at	180	188	147	97	37	175	22	49	34	47	8
41025_r_at	73	59	56	33	17	59	11	16	13	17	9
36423_at	7	8	39	20	71	7	390	241	175	438	10

Subset (3 vs. 3)

ProbeSetID	A1	A2	A3	B1	B2	B3	*RP*
omake	5000	5000	5000	10	10	10	1
1514_g_at	63	68	144	83	3739	1742	12
1805_g_at	32	16	412	43	3338	3640	2
1894_f_at	3016	2181	492	2106	72	584	5
217_at	34	26	223	37	2832	1822	10
38604_at	88	35	51	52	1742	850	19
1008_f_at	4453	3064	1721	3872	240	1488	15
37793_r_at	180	188	147	175	22	49	3
41025_r_at	73	59	56	59	11	16	4
36423_at	7	8	39	7	390	241	47

$x = 10$ → median = 5.5 (complete set)
median = 7.5 (subset)

Subsetで低い値をもつ → よい方法

比較のための評価基準2-1

- 実験データの（追加や）削除に対して頑健かどうか
  - Complete setでのランキング上位**50**個の順位のmedian がsubsetでどれだけ保持されているか (medianの最小値 = 25.5)

既存の手法の比較ではRank productsが圧倒的によい

Kadota *et al.*, *in preparation*

比較のための評価基準2-2

- 実験データの（追加や）削除に対して頑健かどうか
  - Complete setでのランキング上位*x*個を、subsetでの上位*x*個中にどれだけ保持できているか

Complete set

ProbeSetID	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	*RP*
omake	5000	5000	5000	5000	5000	10	10	10	10	10	1
1514_g_at	63	68	144	58	860	83	3739	1742	1650	2904	2
1805_g_at	32	16	412	659	1892	43	3338	3640	3032	2671	3
1894_f_at	3016	2181	492	444	38	2106	72	584	42	44	4
217_at	34	26	223	312	870	37	2832	1822	1997	2307	5
38604_at	88	35	51	77	937	52	1742	850	2183	3094	6
1008_f_at	4453	3064	1721	1911	177	3872	240	1488	175	176	7
37793_r_at	180	188	147	97	37	175	22	49	34	47	8
41025_r_at	73	59	56	33	17	59	11	16	13	17	9
36423_at	7	8	39	20	71	7	390	241	175	438	10

Subset (3 vs. 3)

ProbeSetID	A1	A2	A3	B1	B2	B3	*RP*
omake	5000	5000	5000	10	10	10	1
1514_g_at	63	68	144	83	3739	1742	12
1805_g_at	32	16	412	43	3338	3640	2
1894_f_at	3016	2181	492	2106	72	584	5
217_at	34	26	223	37	2832	1822	10
38604_at	88	35	51	52	1742	850	19
1008_f_at	4453	3064	1721	3872	240	1488	15
37793_r_at	180	188	147	175	22	49	3
41025_r_at	73	59	56	59	11	16	4
36423_at	7	8	39	7	390	241	47

6/10*100 = 60%

Subsetで高い値をもつ → よい方法

比較のための評価基準2-2

- 実験データの（追加や）削除に対して頑健かどうか
  - Complete setでのランキング上位**50**個を、subsetでの上位**50**個中にどれだけ保持できているか

既存の手法の比較ではRank productsが圧倒的によい

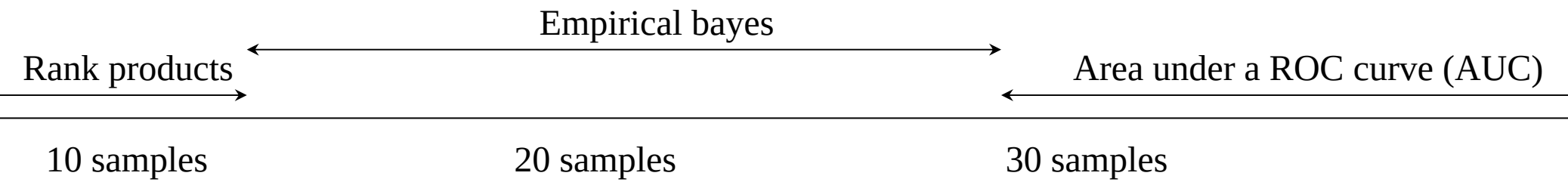
こんな最近の比較結果もありますが…

- Vardhanabhuti *et al.*, *OMICS*, 2006
  - (GCRMA coupled with) **Cyber-T** or **SAM** is the best.

Jeffery *et al.*, *BMC Bioinformatics*, 2006
Breitling *et al.*, *FEBS Lett.*, 2004
Kadota *et al.*, *in preparation*

まとめ

- 分類精度の高さ
  - Low levels of noise and large sample size → **Area under a ROC curve (AUC)**
  - High levels of noise and small sample size → **Rank products**
  - A range of sample sizes → **Empirical bayes *t*-statistic**



- 実験データの（追加や）削除に対して頑健かどうか
  - **Rank product**

「Rでマイクロアレイデータ解析」

- はじめに
- インストールと起動(last updated on 2006/7/11)
- 使用例(初心者向け)(last updated on 2006/5/30)
- サンプルマイクロアレイデータ (last updated on 2007/1/24)
- データ取得 --- GEO (package: GEOquery)(last updated on 2006/5/30)
- 正規化 --- dual channel (Stanford型マイクロアレイ) (package: limma)
- 正規化 --- dual channel (Stanford型マイクロアレイ) (package: marray)
- 正規化 --- single channel (Affymetrix型マイクロアレイ) (package: affy) (last updated on 2006/6/20)
- 正規化 --- single channel (Affymetrix型マイクロアレイ) (package: plier) (last updated on 2006/6/21)
- 正規化 --- single channel (Affymetrix型マイクロアレイ) (GLA; GLog Average) (last updated on 2007/3/8) New
- 正規化 --- single channel (Affymetrix型マイクロアレイ) (package: farms) (last updated on 2006/6/20)
- 基本的な解析 --- 遺伝子ごとの平均発現量など (last updated on 2006/2/28)
- 基本的な解析 --- 最大発現量を示す組織を調べたい (last updated on 2006/2/28)
- 前処理 --- 遺伝子のフィルタリング (package: genefilter; genefilter)
- 前処理 --- 遺伝子のフィルタリング (package: som; filtering)
- 前処理 --- スケール化 --- Mean-SD (Z) scaling (package: som; normalize)
- 前処理 --- スケール化 --- Mean-SD (Z) scaling (package: genefilter; genescale)
- 前処理 --- スケール化 --- Mean-SD (Z) scaling (package: base; scale)
- 前処理 --- スケール化 --- Mean-MAD scaling (package: base; scale)
- 前処理 --- スケール化 --- Median-SD scaling (package: base; scale)
- 前処理 --- スケール化 --- Median-MAD scaling (package: stats)
- 前処理 --- スケール化 --- Tukey-MAD scaling (package: stats)
- 前処理 --- スケール化 --- Range scaling (package: genefilter; genescale)
- 前処理 --- スケール化 --- Log-transformation (package: base; log)
- 解析 --- 似た発現パターンを持つ遺伝子の同定 (package: genefilter; genefinder)
- 解析 --- 自作統計量を用いて並べ替え検定(random permutation test)(last updated on 2006/7/31)
- 解析 --- 発現変動遺伝子の同定 --- 二群間比較 --- samroc (package: SAGx) (last updated on 2007/3/29) New
- 解析 --- 発現変動遺伝子の同定 --- 二群間比較 --- RankProd (package: RankProd) (last updated on 2007/3/15) New
- 解析 --- 発現変動遺伝子の同定 --- 二群間比較 --- Empirical bayes statistic (package: limma) (last updated on 2007/3/8) New
- 解析 --- 発現変動遺伝子の同定 --- 二群間比較 --- MTP (package: multtest; MTP)
- 解析 --- 発現変動遺伝子の同定 --- 二群間比較 --- ANOVA (package: genefilter)
- 解析 --- 発現変動遺伝子の同定 --- 二群間比較 --- Student's t-test(last updated on 2007/3/8) New
- 解析 --- 発現変動遺伝子の同定 --- 二群間比較 --- Welch t-test(last updated on 2007/3/8) New
- 解析 --- 発現変動遺伝子の同定 --- 二群間比較 --- SAM (package: samr) (last updated on 2007/3/15) New
- 解析 --- 発現変動遺伝子の同定 --- 二群間比較 --- SAM (package: siggenes) (last updated on 2007/2/16)
- 解析 --- 発現変動遺伝子の同定 --- 組織特異的(選択的)発現遺伝子 --- Kadota's AIC-based method (last updated on 2006/10/31)
- 解析 --- 発現変動遺伝子の同定 --- 組織特異的(選択的)発現遺伝子 --- Sprent's non-parametric method (last updated on 2006/10/31
- 解析 --- 発現変動遺伝子の同定 --- 組織特異的(選択的)発現遺伝子 --- Schug's H(x) statistic (last updated on 2006/7/10)
- 解析 --- 発現変動遺伝子の同定 --- 組織特異的(選択的)発現遺伝子 --- Schug's Q statistic (last updated on 2006/6/16)
- 解析 --- 発現変動遺伝子の同定 --- 組織特異的(選択的)発現遺伝子 --- ROKU (last updated on 2006/8/2)
- 解析 --- 発現変動遺伝子の同定 --- 組織特異的(選択的)発現遺伝子 --- Tukey-Kramer's HSD test (last updated on 2007/3/29) New
- 解析 --- 発現変動遺伝子の同定 --- 時系列データ --- Periodic genes (package: GeneTS) (last updated on 2006/7/11)
- 解析 --- 発現変動遺伝子の同定 --- 時系列データ --- Periodic genes (Lomb-Scargle periodograms) (last updated on 2006/7/11)
- 解析 --- 発現変動遺伝子の同定 --- 時系列データ --- non-periodic genes (maSigPro) (last updated on 2006/7/11)

Contents

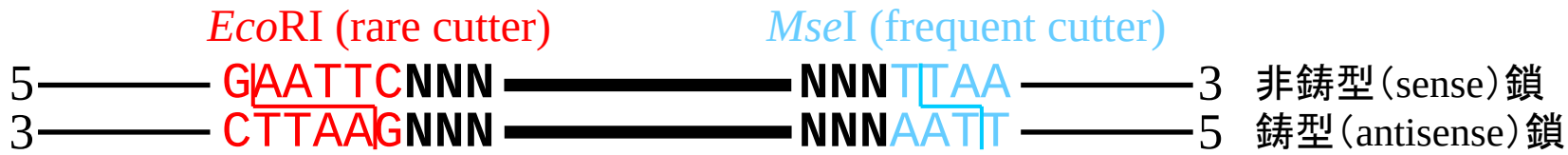
1. マイクロアレイ
   - 組織特異的発現パターン検出法
   - 二群間比較用の手法
2. cDNA-AFLP(HiCEP)
   - データ解析手法について

AFLP: Vos *et al.*, *Nucleic Acids Res.*, 1995

cDNA-AFLP: Bachem *et al.*, *Plant J.*, 1996

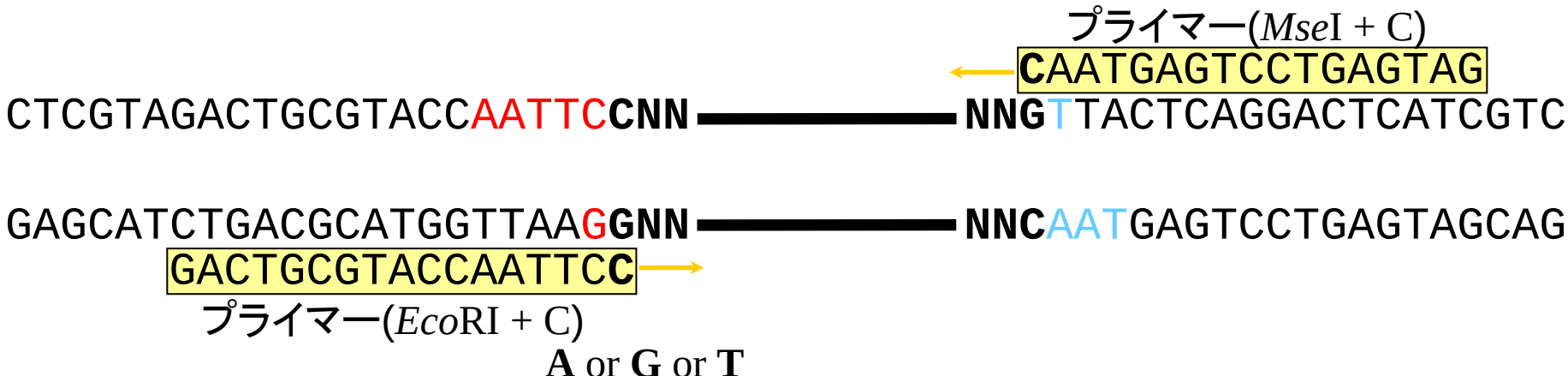
(cDNA-)AFLP法

1. 二種類の制限酵素によるゲノムDNA(cDNA)切断



2. アダプター配列の結合

3. 予備PCR増幅(解析するDNA断片のサブグループ化;4×4=16分割)

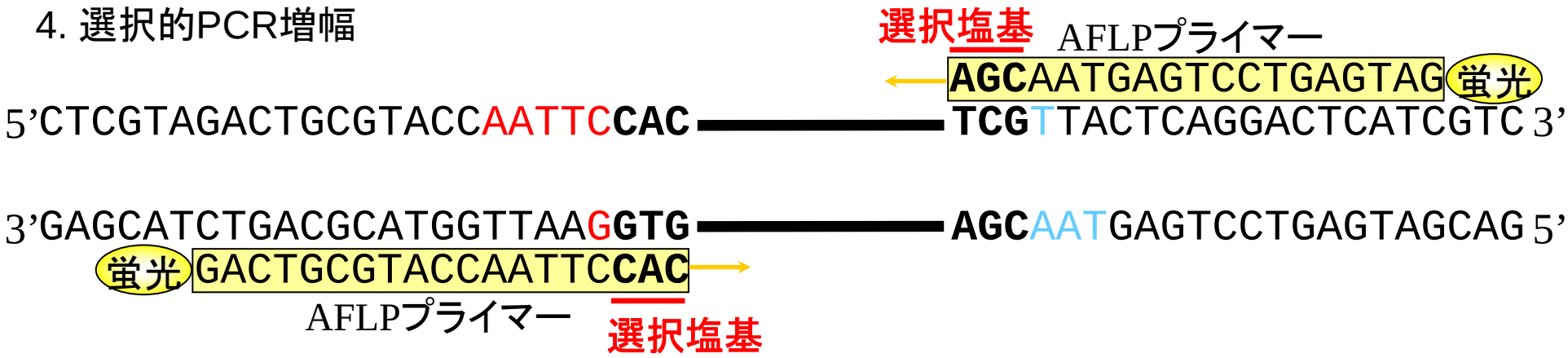


AFLP: Vos *et al.*, *Nucleic Acids Res.*, 1995

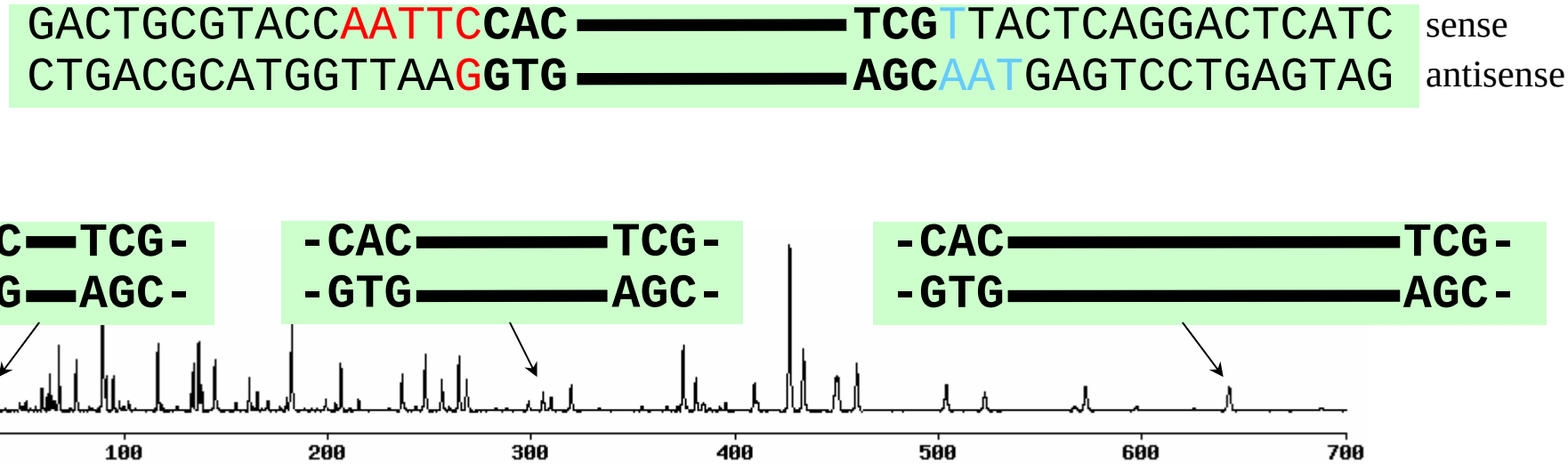
cDNA-AFLP: Bachem *et al.*, *Plant J.*, 1996

(cDNA-)AFLP法

4. 選択的PCR増幅

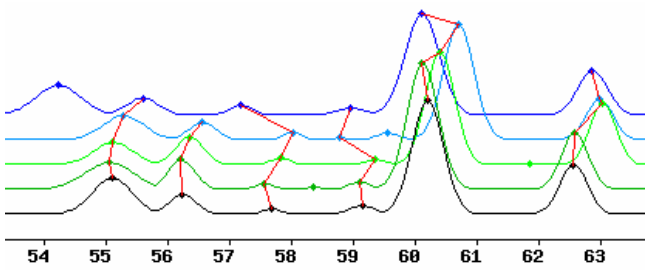


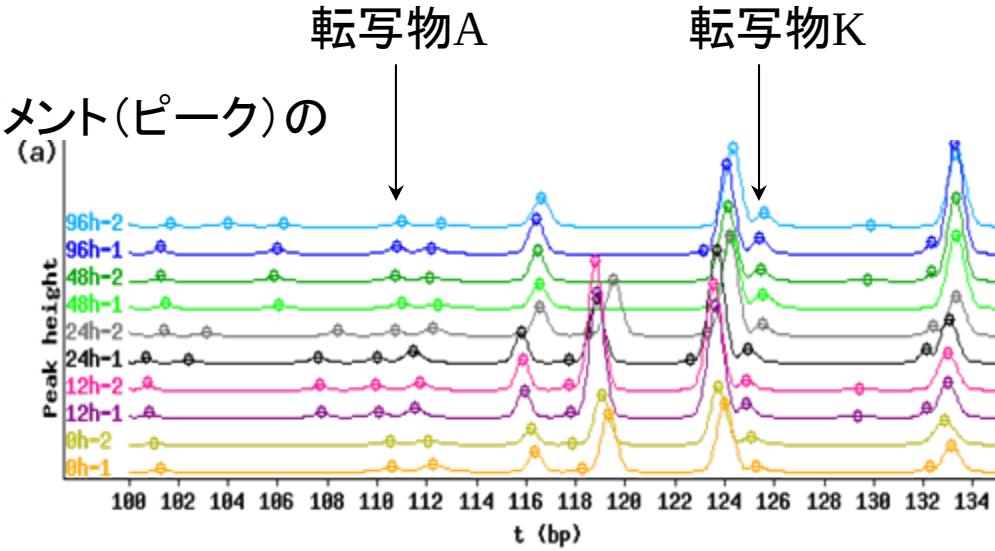
5. 電気泳動



電気泳動法に基づく遺伝子発現解析

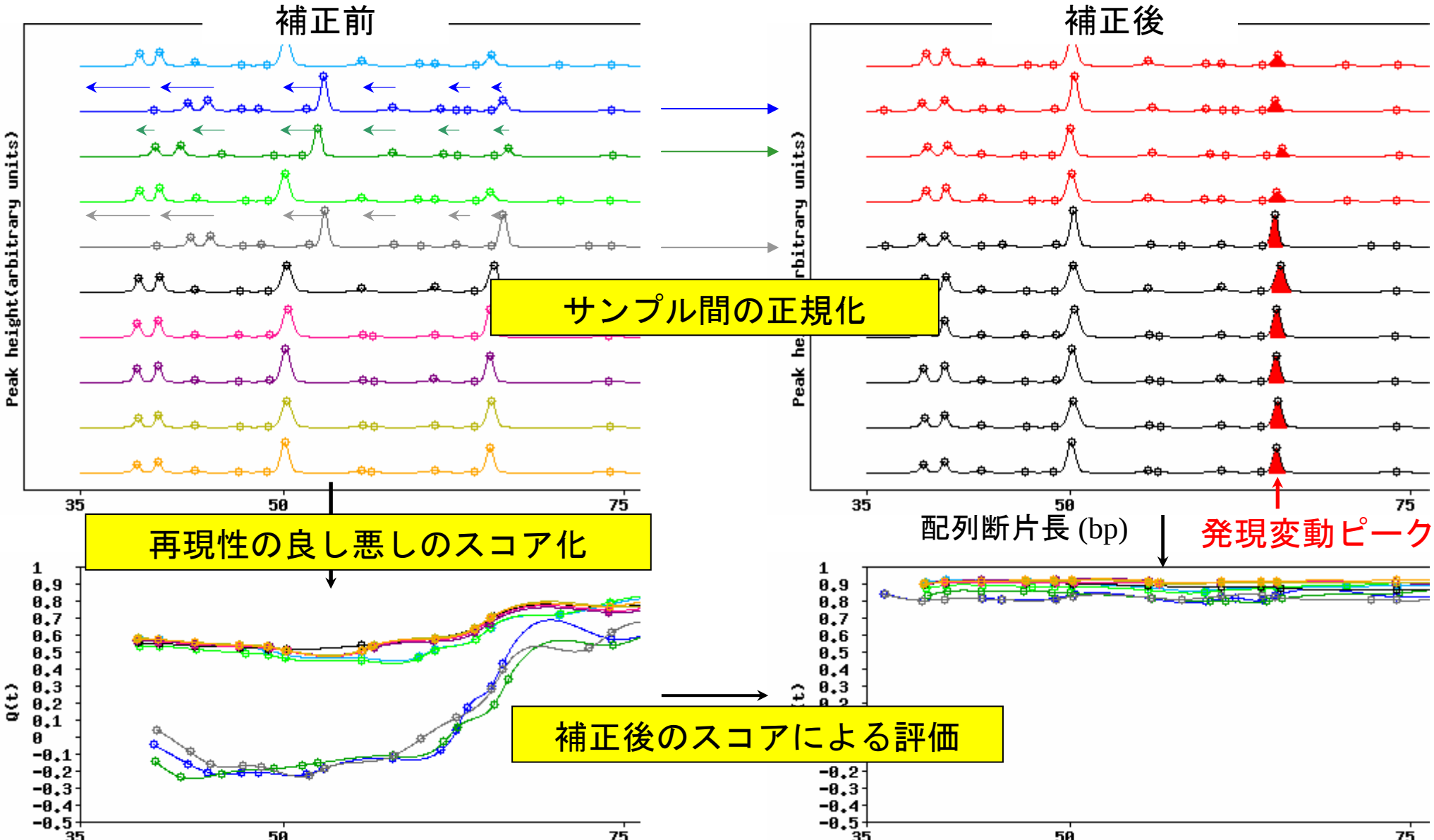
- 長所
  1. 解析対象となるゲノムの塩基配列情報を必要としない
  2. 少量のDNA量で解析可能（PCR増幅を行うため）
- 短所
  1. 比較する実験数が増えるほど、フラグメント（ピーク）のアラインメント精度が下がる





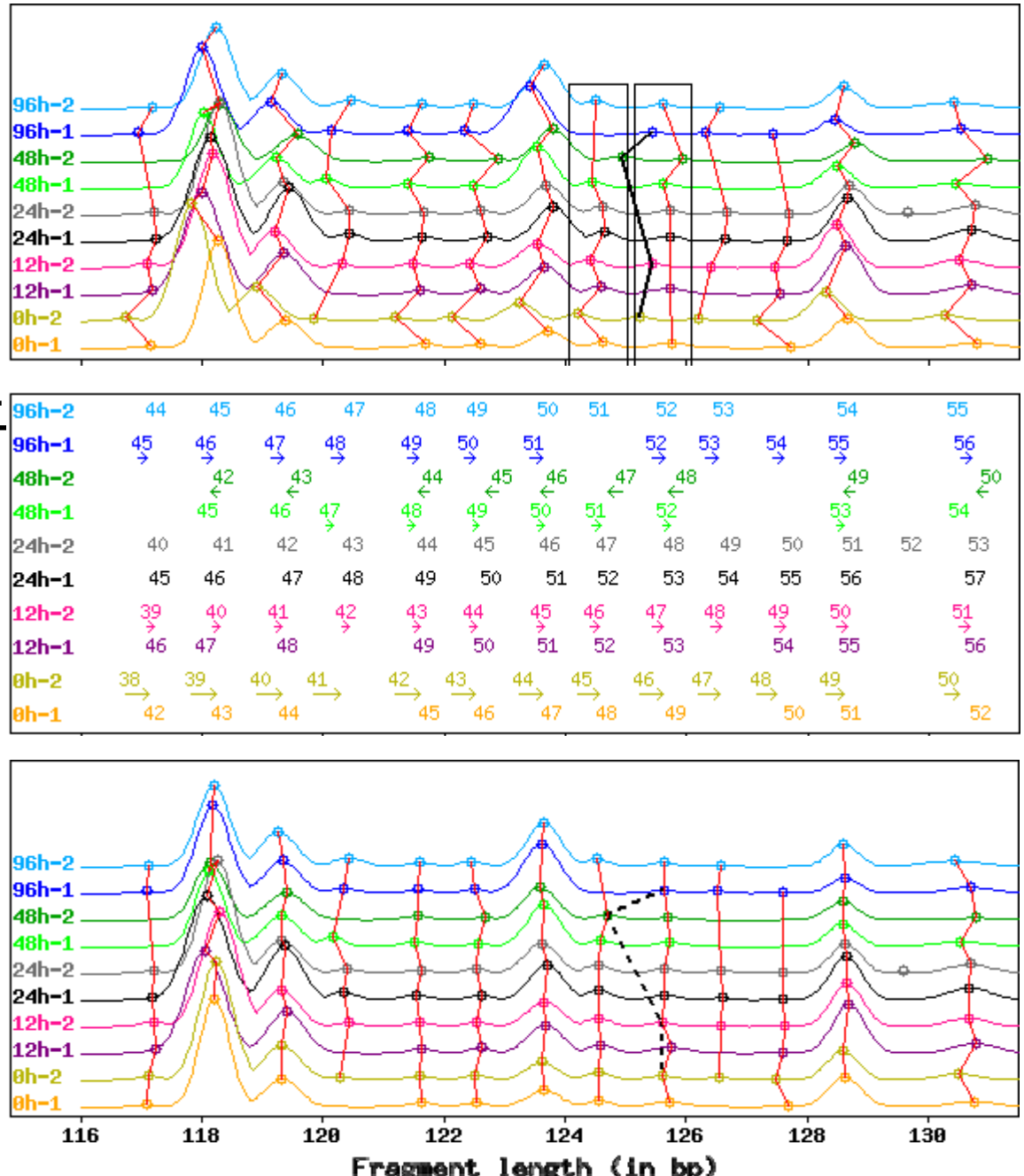
  2. 切り出したゲル片には、（多くの場合）複数のDNAフラグメントが含まれる
     →アノテーションが困難

門田法1


補正前
補正後
Peak height(arbitrary units)
35
50
75
サンプル間の正規化
再現性の良し悪しのスコア化
配列断片長 (bp)
発現変動ピーク
Q(t)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
補正後のスコアによる評価

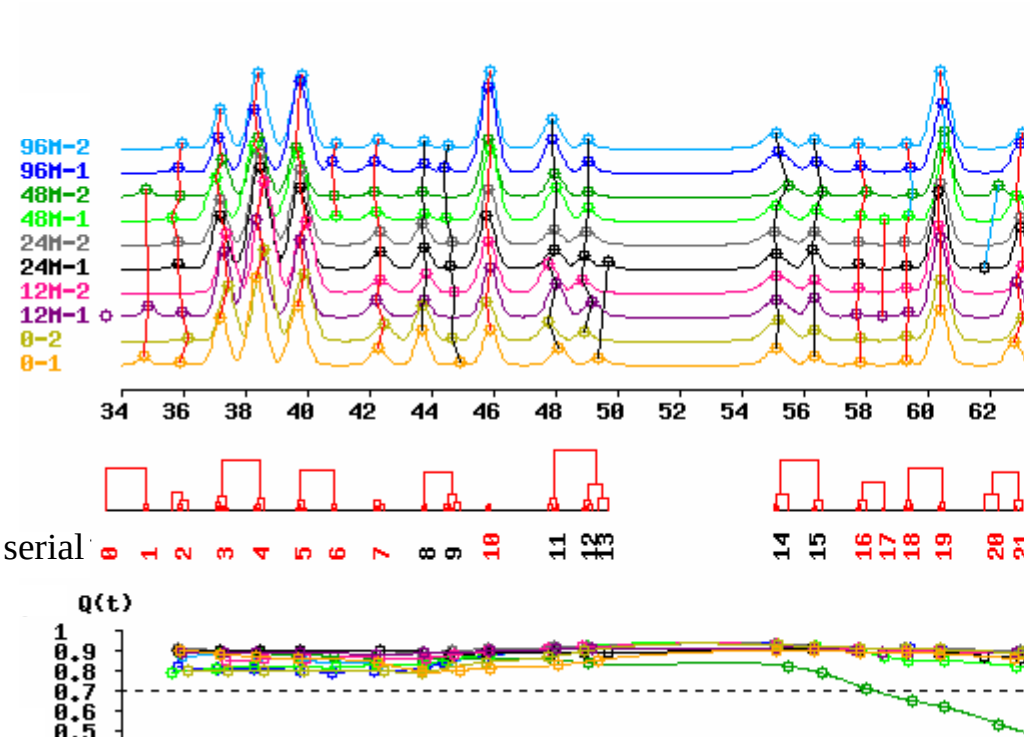
門田法2

- フラグメント長補正
  →アラインメント精度向上



門田法2(利点)

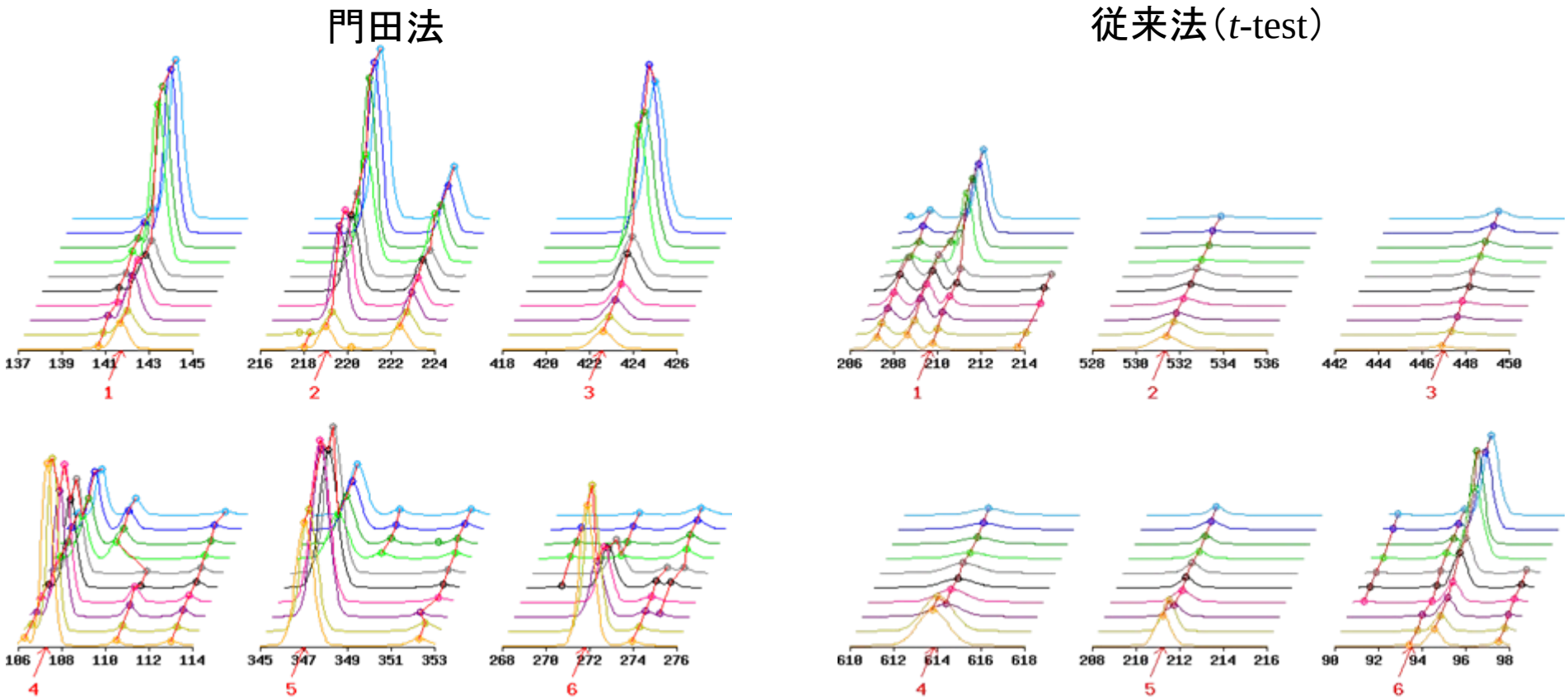
- 遺伝子発現行列を作成可能
  →既存のマイクロアレイ解析手法も適用可能



serial	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
0			5							
1	41		58					39		
2	16	17	26		21	16	16	8	24	23
3	235	276	330	297	266	222	213	185	179	199
4	437	456	480	554	507	455	379	300	320	375
5	300	333	381	357	409	371	332	249	452	369
6							27	8	52	25
7	81	82	86	63	83	64	49	27	51	41
8	172	174	84	91	108	105	30	23	41	31
9	11	16		6	11	13	15		22	17
10	173	191	245	255	266	274	359	289	424	384
11	84	91	169	144	96	71	164	114	163	144
12		44	73	64	62	64	60	30	53	43
13	39				37					
14	87	102	85	85	75	64	77	58	101	76
15	48	53	93	60	95	66	55	28	57	45
16	13	15	18	21	22	16	24	27	34	25
17			7				7			
18	21	23	24	28	18	18	23	14	24	27

門田法2（利点）

- 発現変動ピーク上位6個の比較



マイクロアレイ以外のトランスクリプトーム解析手法の開発もやってます

謝辞

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）

秋田県立大学　生物資源科学部

小西智一　准教授

放射線医学総合研究所　先端遺伝子発現研究G

安倍真澄　グループリーダー

荒木良子　チームリーダー

CBRC

高橋勝利　元親分

挿絵担当：門田雅世