「感度・特異度・再現性」高く発現変動遺伝子を検出するための推奨ガイドライン

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）

自己紹介

- 2002年3月
  - 東京大学・大学院農学生命科学研究科　博士課程修了
  - 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」（指導教官:清水謙多郎教授）
- 2002/4/1~
  - 産総研・生命情報科学研究センター　産総研特別研究員
- 2003/11/1~
  - 放医研・先端遺伝子発現研究センター　研究員
- 2005/2/16~
  - 東京大学・大学院農学生命科学研究科　特任助手
- 2007/4/1~現在
  - 東京大学・大学院農学生命科学研究科　特任助教

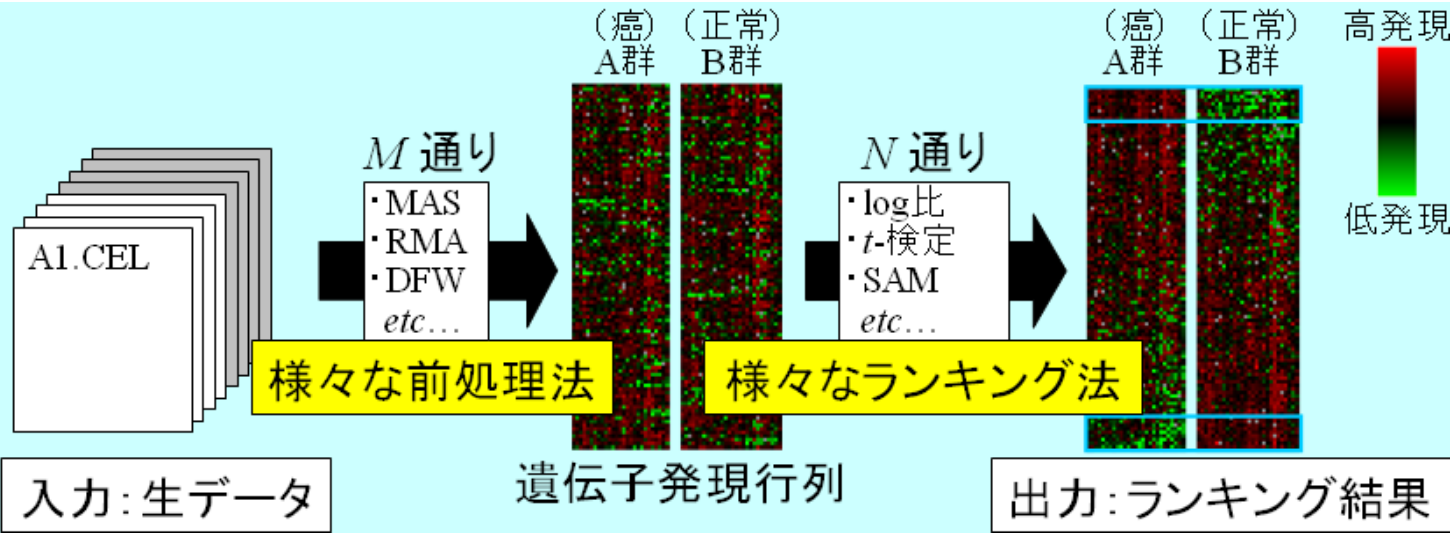
アグリバイオインフォマティクスプログラム

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

Affymetrix GeneChip解析の流れ

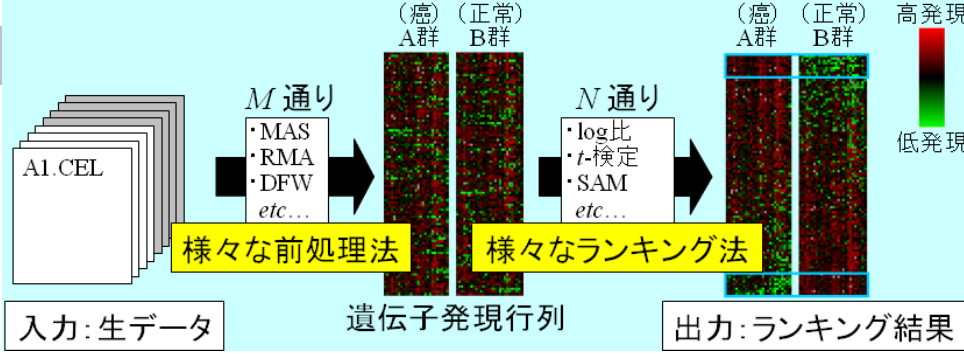
- 二群間比較の場合



- •Gene Ontology解析
- •(未知サンプルの)分類
- •モチーフ解析
- •パスウェイ解析

- 最近の知見
  - 前処理法とランキング法の組合せが重要

結論



- 「感度・特異度」が高い方法(組合せが重要!)

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」が高いランキング法
  - (前処理法によらず)WAD ↔ 従来: Average Difference (AD)法

Kadota K, Nakai Y, Shimizu K, *AMB*, **3**: 8, 2008
Kadota K, Nakai Y, Shimizu K, *AMB*, **4**: 7, 2009

「感度・特異度」をAUC値で評価

- 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				

|log比|を計算

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

|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

AUC値=100%

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				

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

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

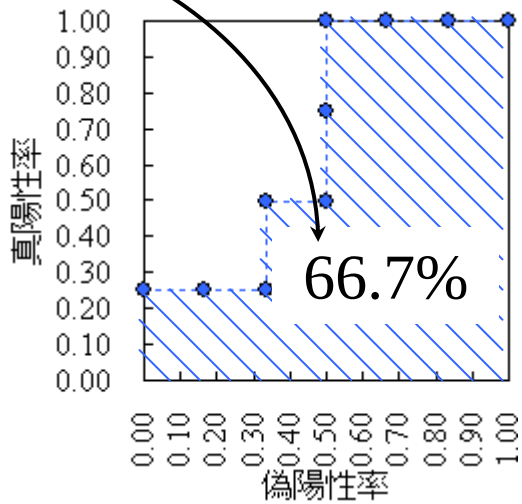
AUC値=83.3%

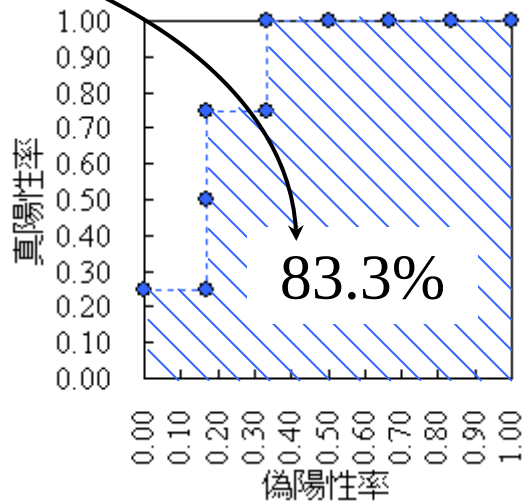
「感度・特異度」をAUC値で評価

- 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

ROC曲線が左上にあるほどよい方法





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

「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	

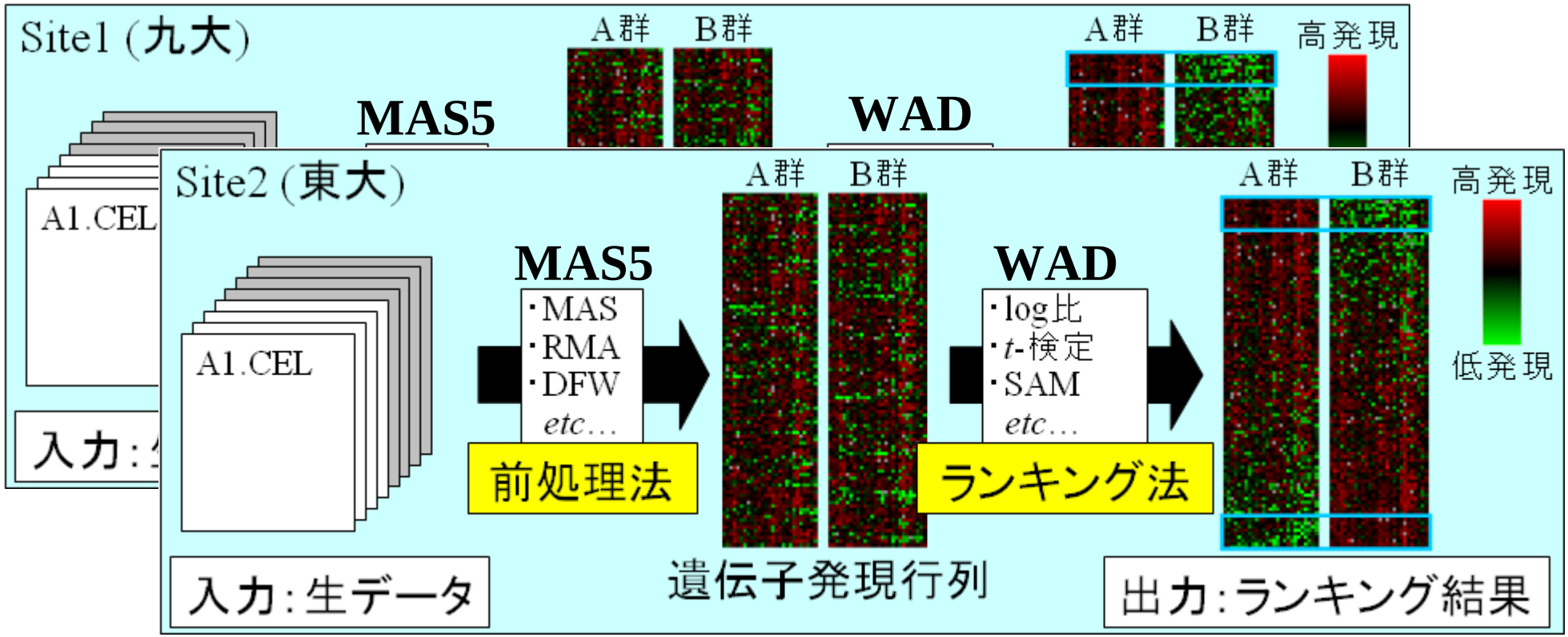
--- データセット3-26（原著論文の解析手段） ---

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

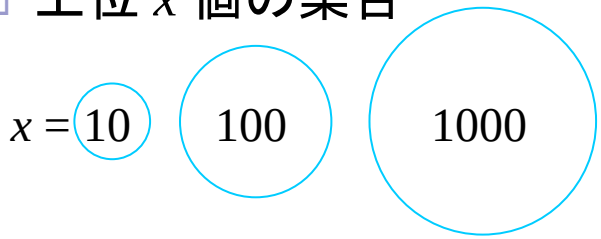
再現性で評価

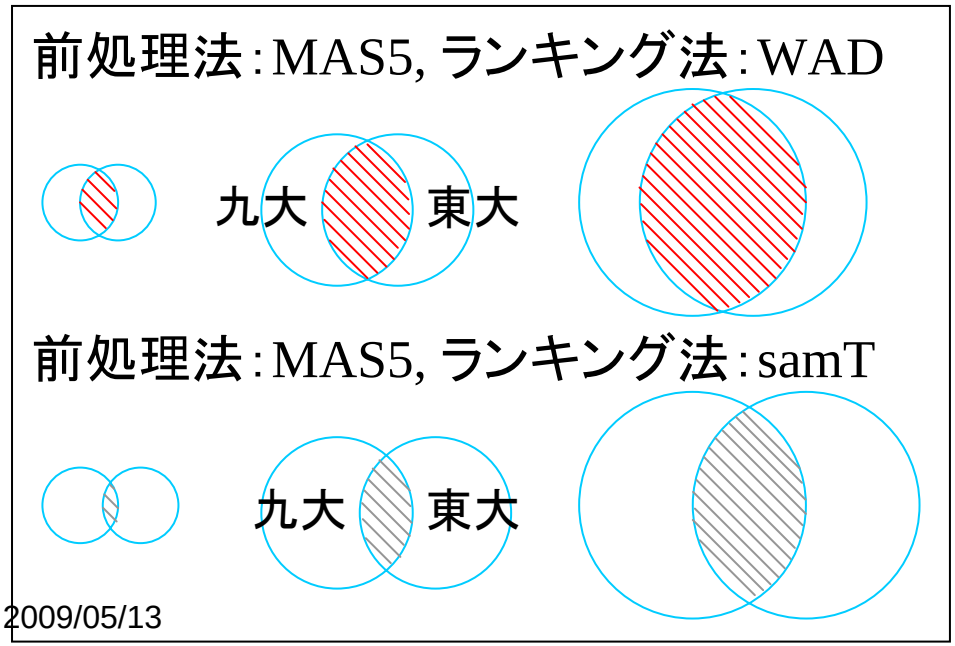
- **評価基準2(再現性)**: POG (**P**ercentage of **O**verlapping **G**enes)値の高さ
  - MicroArray Quality Control (MAQC) プロジェクトで提唱(0≦POG≦100%)
  - POG値が高い → ランキング結果の頑健性(再現性)が高い方法

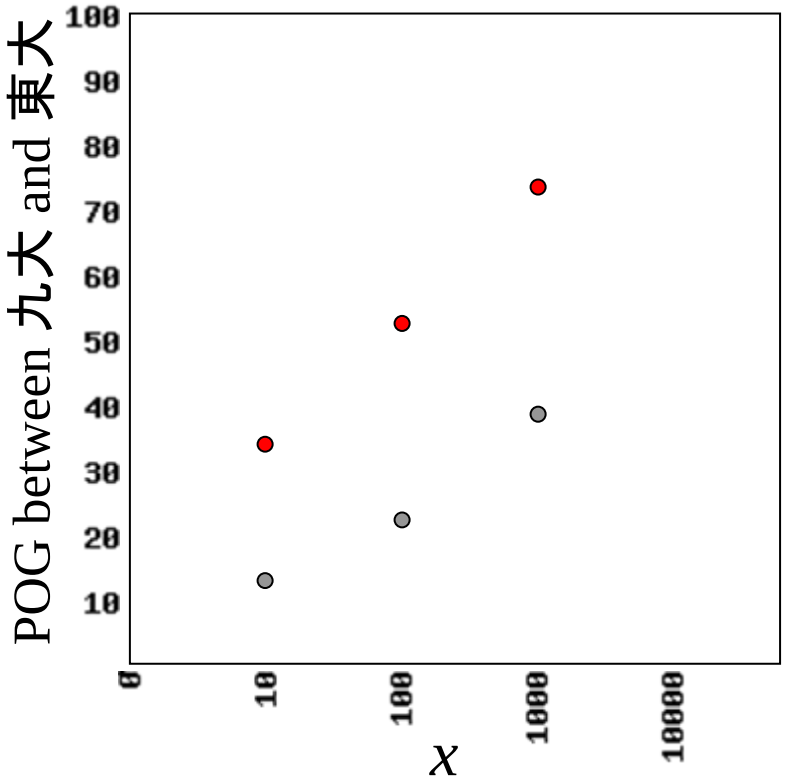


再現性で評価

- **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



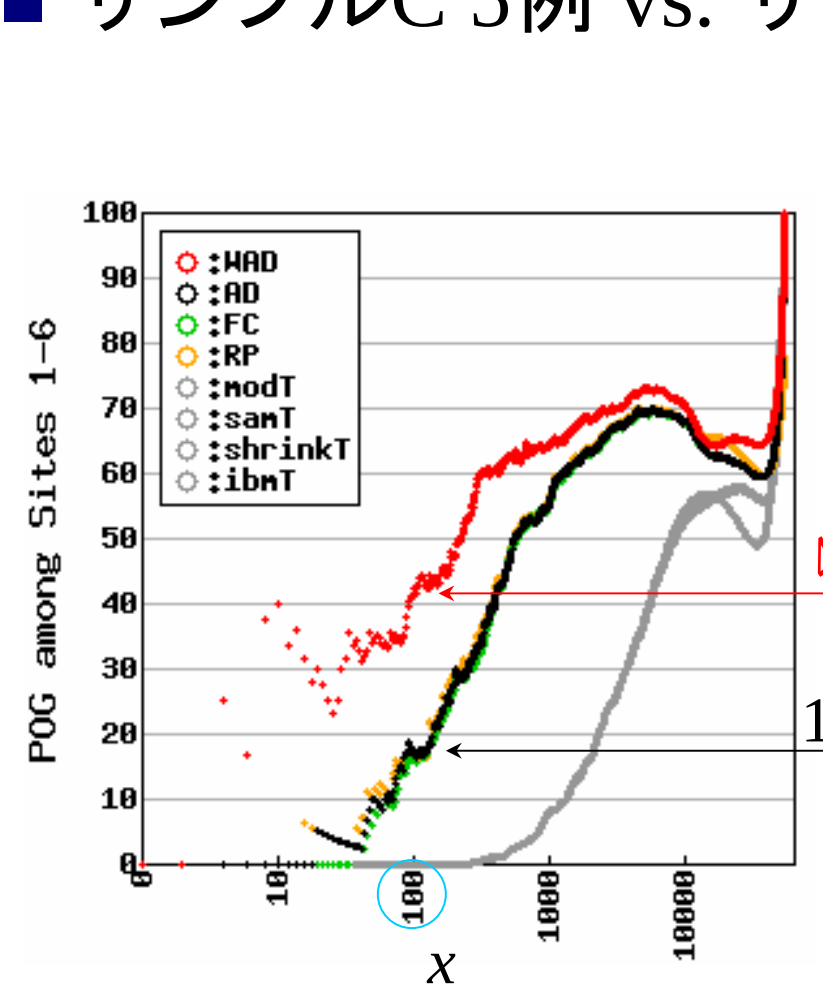


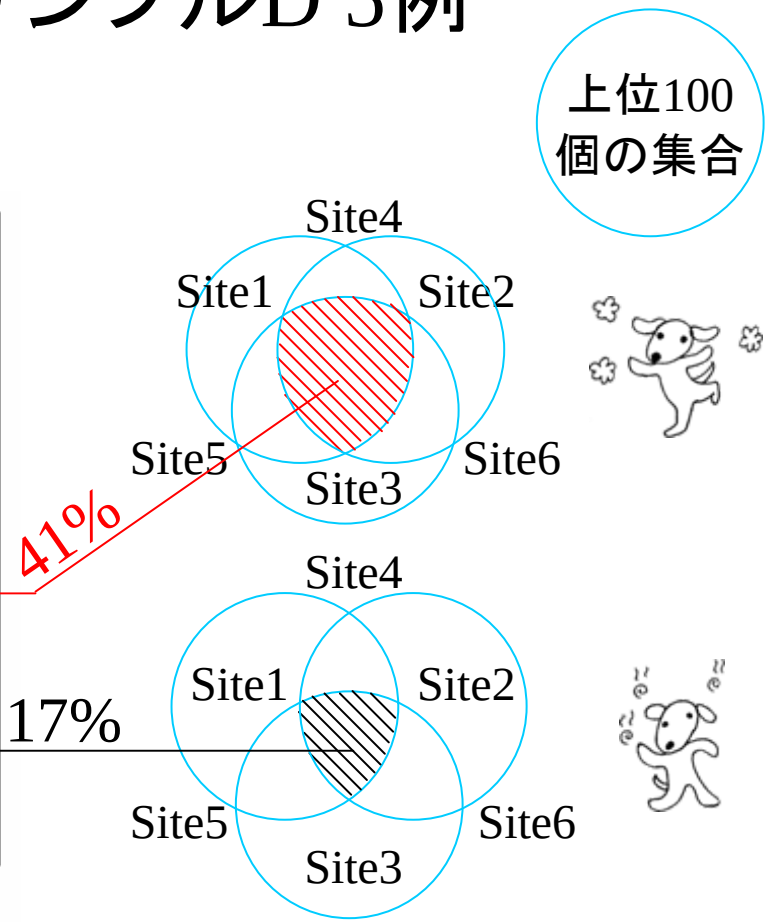


再現性:WAD > samT

「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

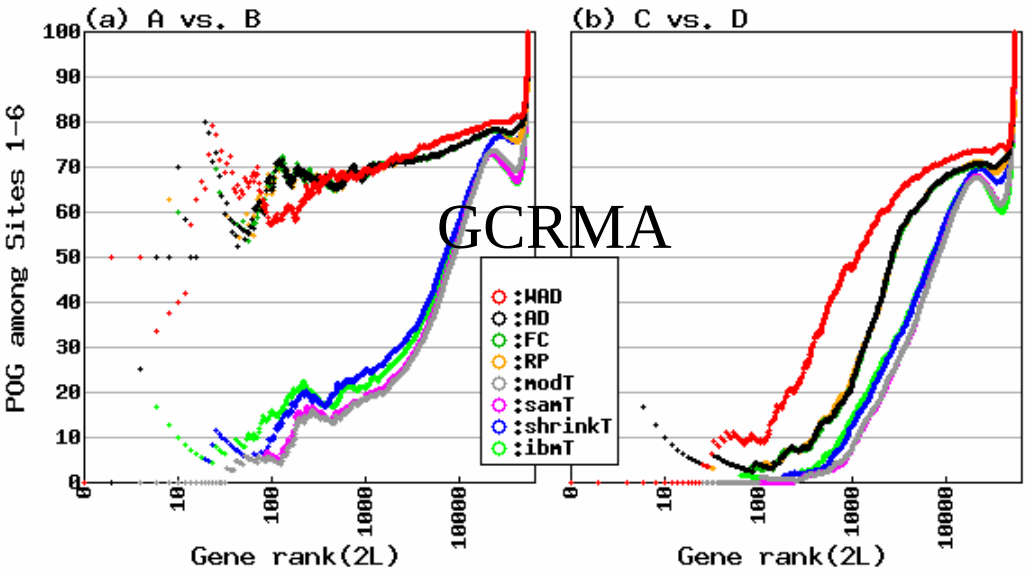




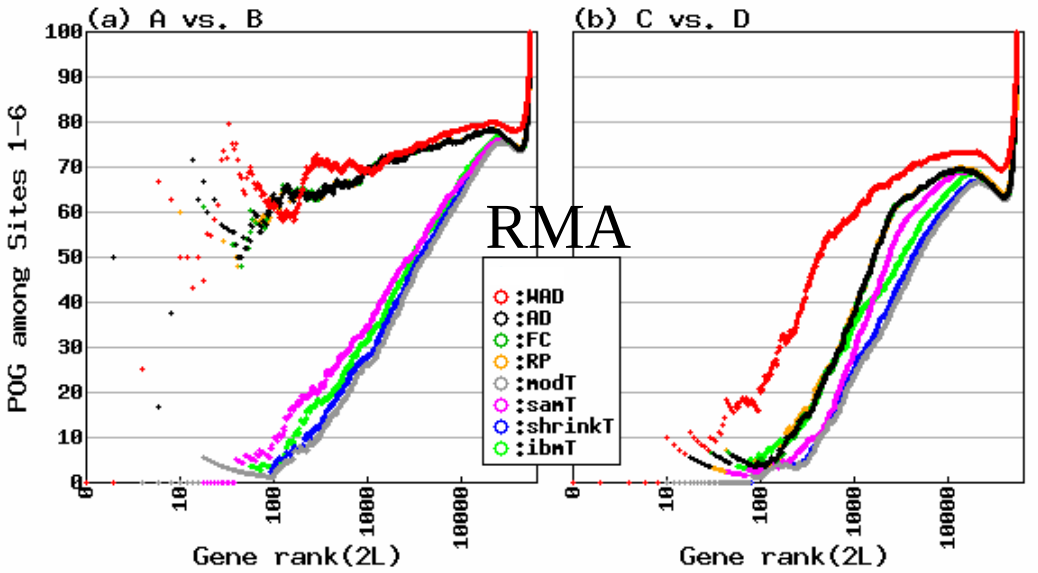
再現性：WAD > MAQC推奨法（AD）

「再現性」解析結果

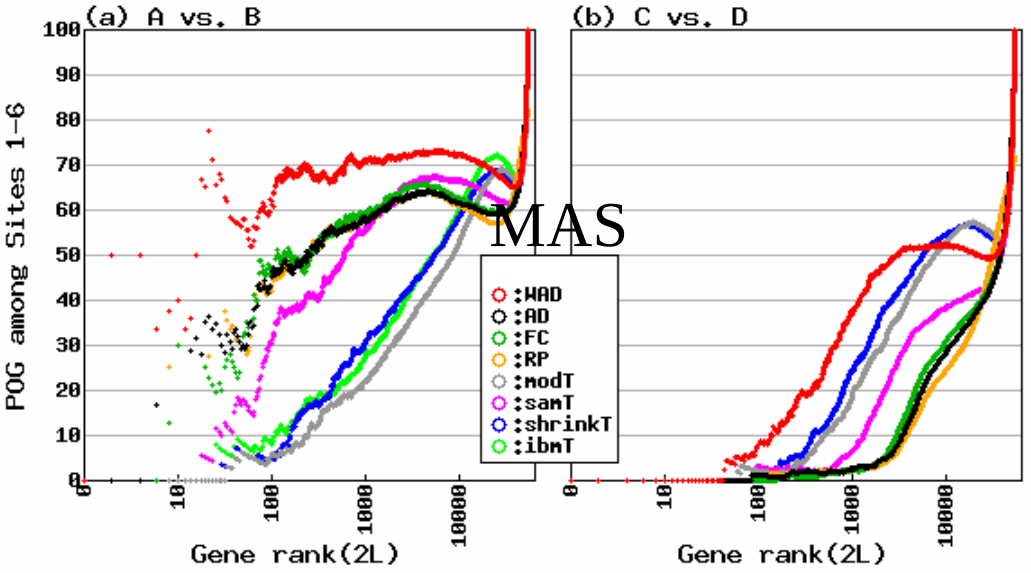
A: Universal Human Reference RNA
B: Human Brain Reference RNA
C: 75% A + 25% B
D: 25% A + 75% B



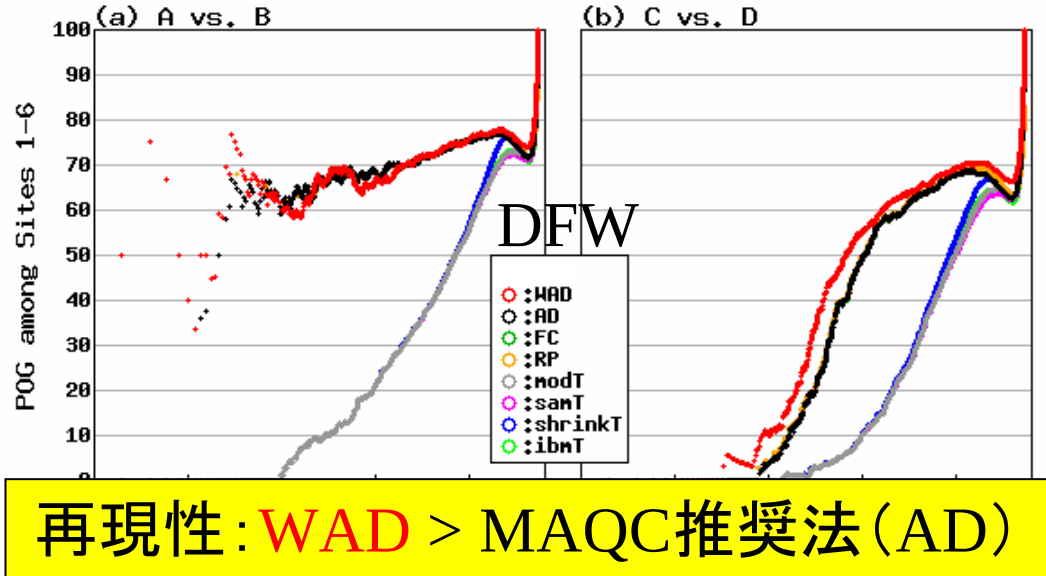
GCRMA



RMA



MAS



DFW

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

xを(0~1)の範囲に規格化

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

Average Difference (AD)統計量

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

平均シグナル強度

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

WAD統計量

$$WAD_i = AD_i \times w_i$$

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

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

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

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

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

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

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

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

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

WADの一位：gene4, ADの一位：gene6

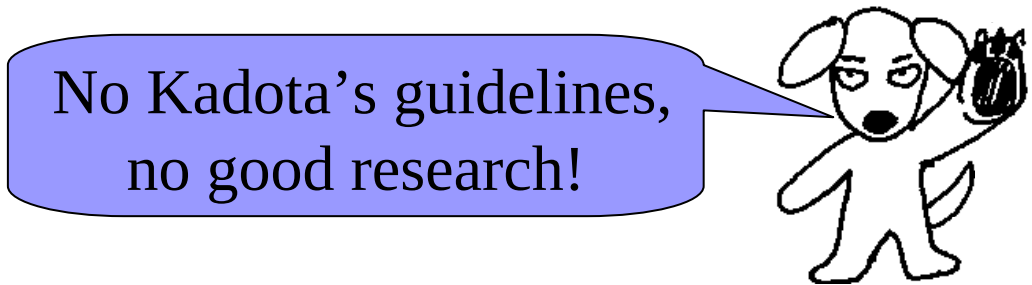
結論（Affymetrix GeneChipデータ解析）

- 「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



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）

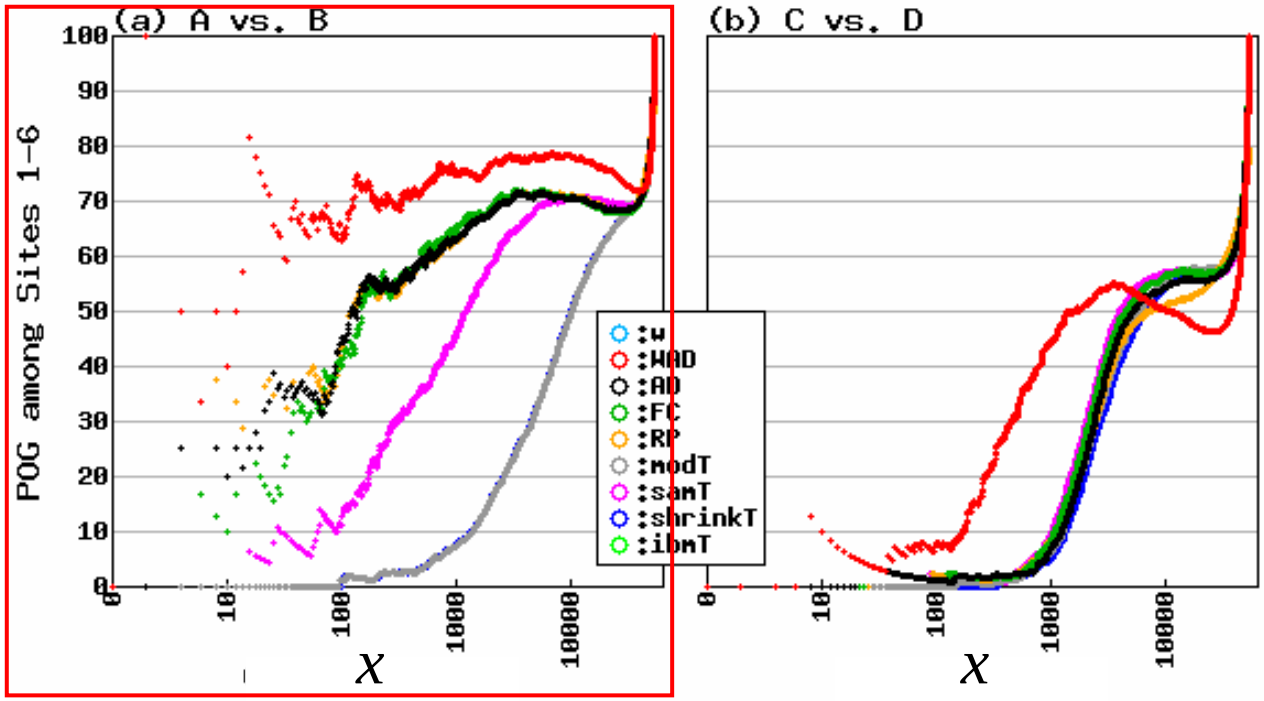
前処理法

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

「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