トランスクリプトーム解析手法の開発

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

2008/12/08

トランスクリプトーム（transcriptome）とは

- 細胞中に存在する転写物全体（transcript + ome）
- トランスクリプトーム解析技術
  - DNAマイクロアレイ
    - Affymetrix GeneChip®, cDNAアレイ, …
  - 電気泳動に基づく方法
    - Differential Display, (cDNA-)AFLP, …
  - Sequenceに基づく方法
    - SAGE, CAGE, MPSS, …

農学生命科学分野への貢献

- トランスクリプトーム解析手法の開発
  - DNAマイクロアレイ
    - **1. 組織特異的遺伝子検出法：ROKU**
    - **2. 発現変動遺伝子検出法：WAD**
      - 応用例：ニュートリゲノミクス研究など
  - 電気泳動に基づく方法
    - **3. 波形補正およびピークアラインメント法：GOGOT**
      - 応用例：農産物の品種識別など

	マイクロアレイ	電気泳動
解析対象	△	○
アノテーション	○	△
データ解析	○	△

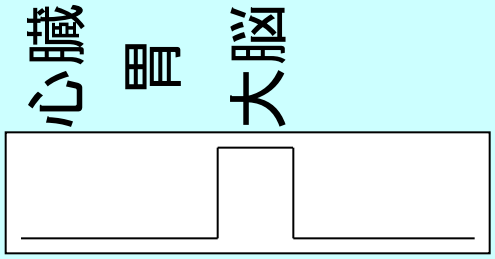
1. 組織特異的遺伝子検出法：ROKU

- 平成18年度修了生（葉 佳臻氏）の研究成果

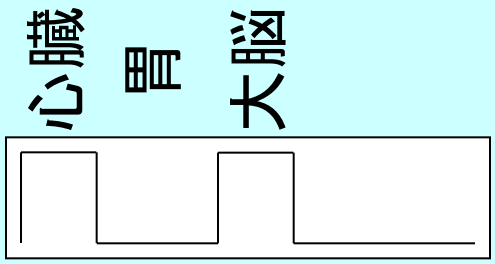
	心臓	肝	大脳		
gene 1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene 2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene n	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

入力：遺伝子発現行列

「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



出力：任意の組織特異的遺伝子

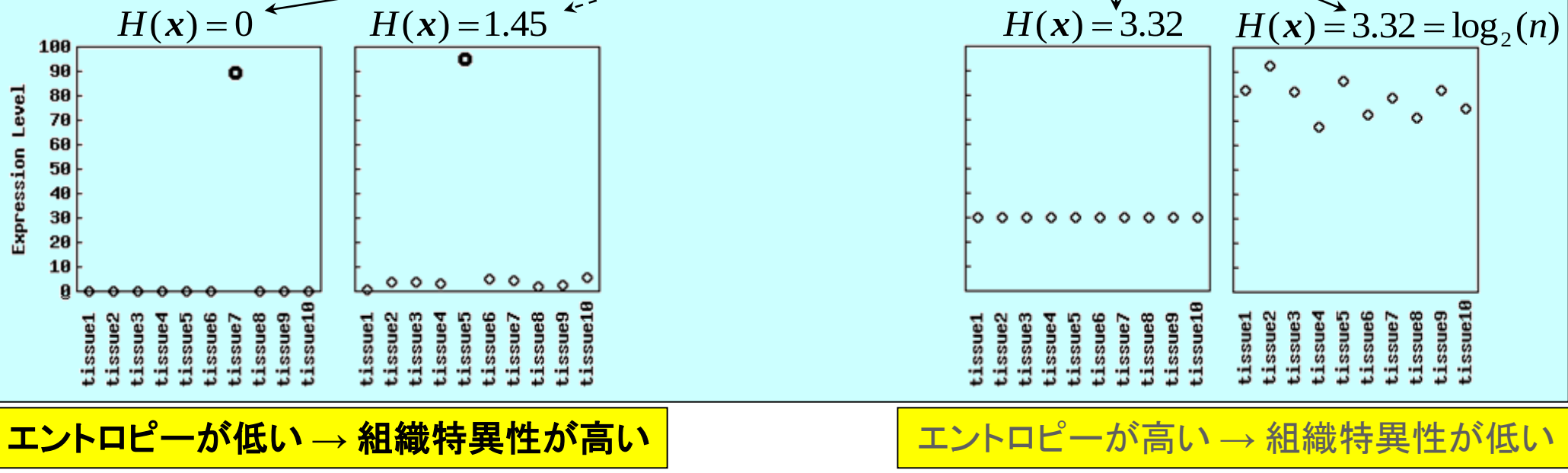
アイデア：エントロピーHを組織特異性度合いの指標にしよう

1. 組織特異的遺伝子検出法：ROKU

- エントロピーHの採用
  - 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： $0 \leqq H(\boldsymbol{x}) \leqq \log_2(n)$

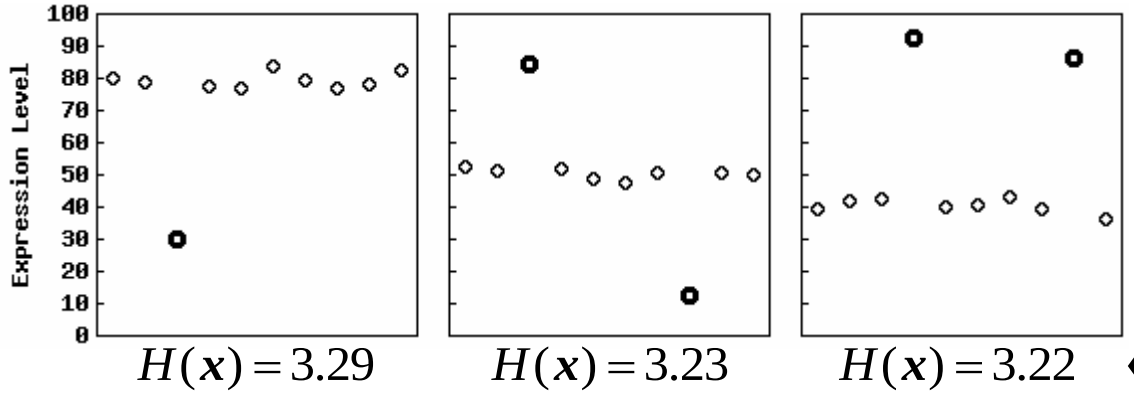


エントロピーが低い → 組織特異性が高い

エントロピーが高い → 組織特異性が低い

1. 組織特異的遺伝子検出法：ROKU

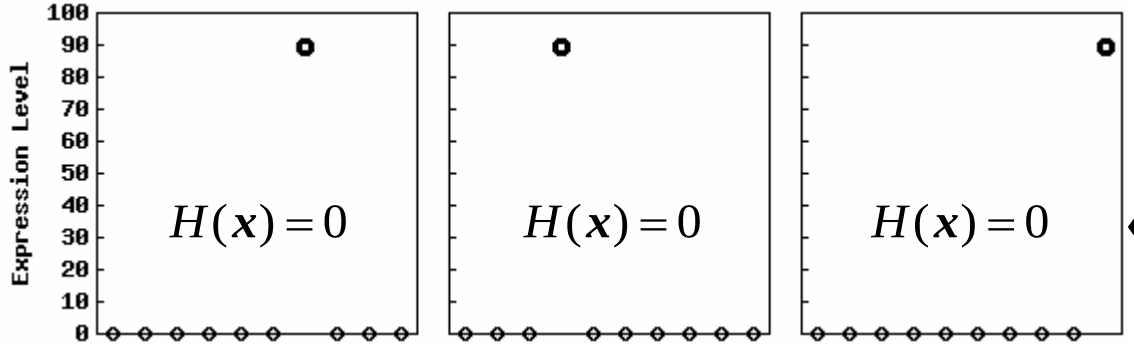
- エントロピー$H(x)$の短所
  1. 組織特異的低発現パターンなどの検出が不可能



$$0 \leqq H(x) \leqq \underbrace{\log_2(n)}_{\| \; 3.32}$$

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

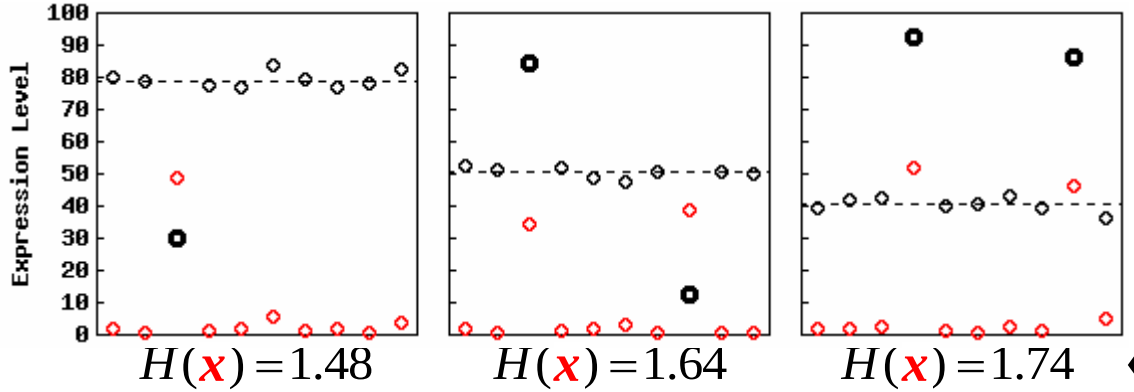


どの組織で特異的なのか分からない

1. 組織特異的遺伝子検出法：ROKU

- 改良点

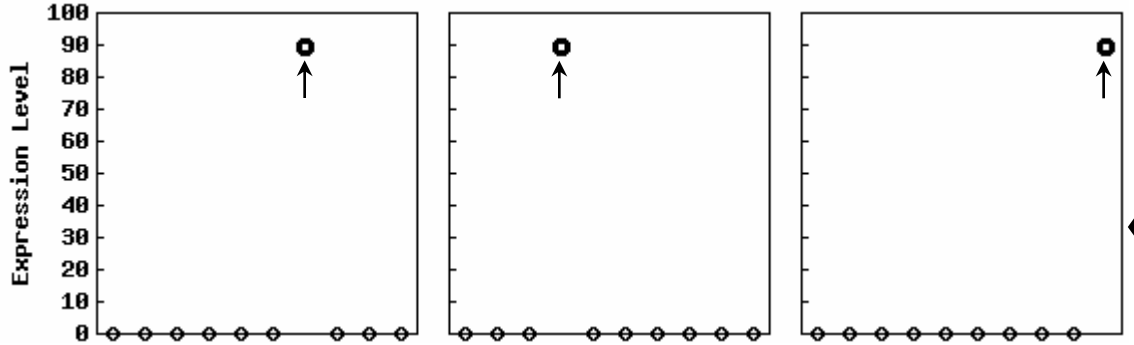
1. 遺伝子発現ベクトルxを変換： $\boldsymbol{x} \rightarrow \boldsymbol{x}$ by $x_i = |x_i - T_{bw}|$



$$0 \leqq H(\boldsymbol{x}) \leqq \underbrace{\log_2(n)}_{\| \; 3.32}$$

上位にランキングされる

2. AICに基づく外れ値検出法の採用



どの組織で特異的なのか分かる

1. 組織特異的遺伝子検出法：ROKU

- 組織特異的遺伝子検出法の比較

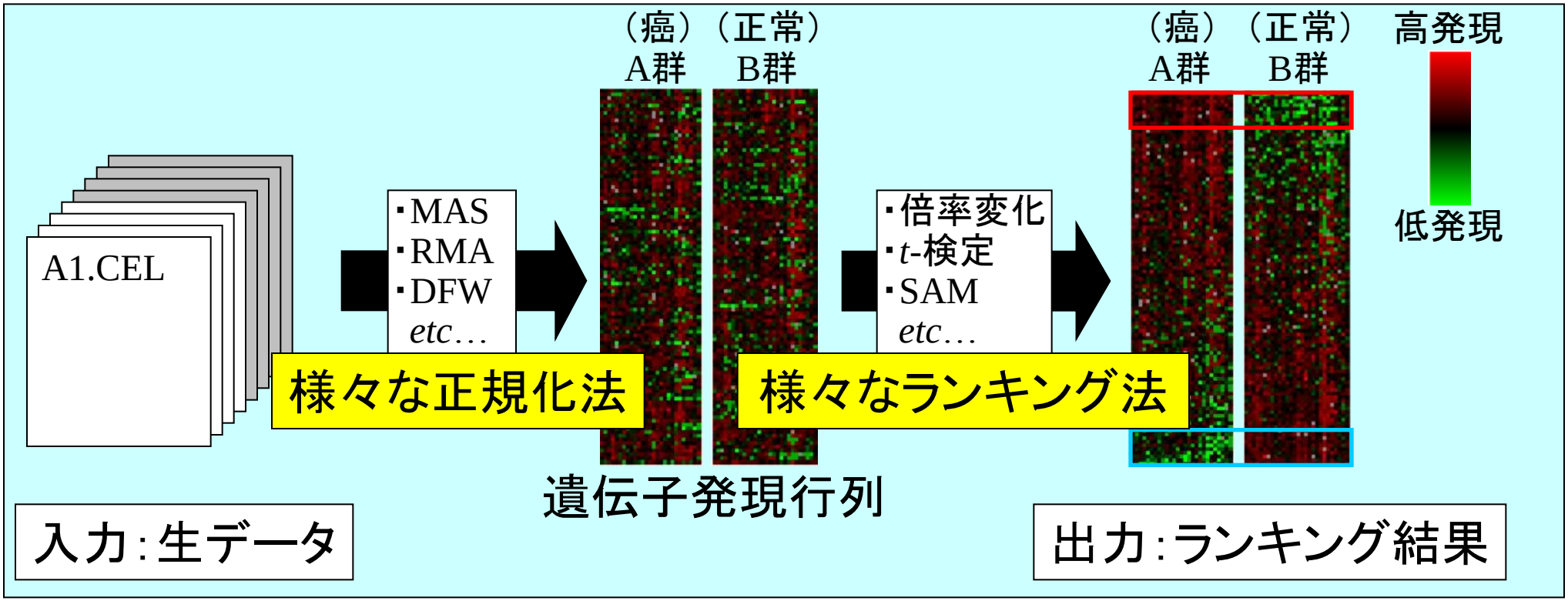
方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
Dixon test	×	×	?	○	-
Pattern matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
Entropy (*H*)	○	×	-	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
AIC	○	○	○	×	○
Sprent's method	○	○	○	×	×
ROKU (AIC+a modified *H*)	○	○	○	○	○

本人材養成プログラムで開発した手法（ROKU）が最も優れている

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

2. 発現変動遺伝子検出法：WAD

- 発現変動遺伝子検出の流れ（Affymetrixデータの場合）

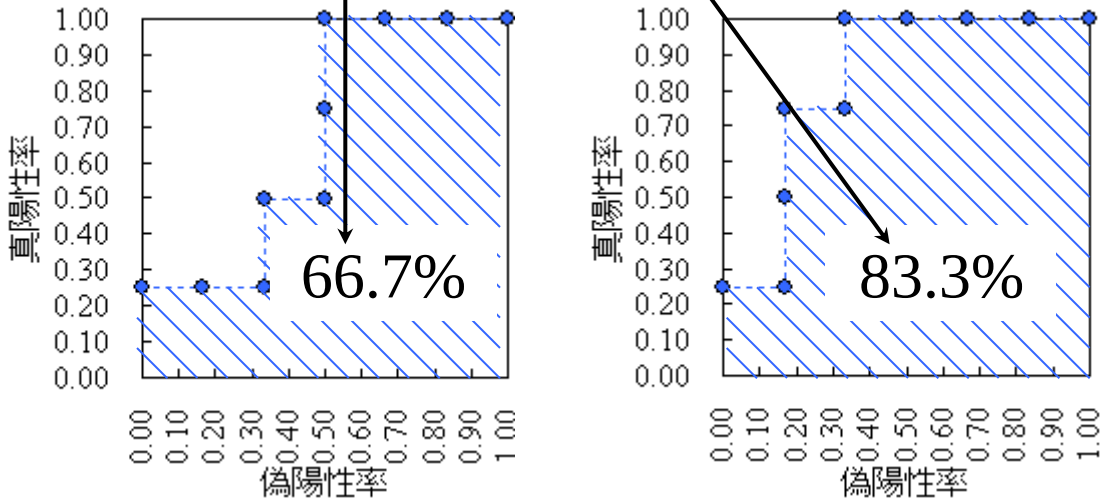


問題：用いる手法によって結果がかなり異なる
手法選択のガイドラインなし

2. 発現変動遺伝子検出法：WAD

- **評価基準1（感度・特異度）**：AUC値の高さ（0≦AUC≦1）
  - 真の発現変動遺伝子をどれだけ上位にランキングできるか？
  - AUC値が高い → 感度・特異度の高い方法
  - 比較例：10遺伝子の発現データを*t*-検定と倍率変化でランキングした結果を比較

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**)

2. 発現変動遺伝子検出法：WAD

- **評価基準1（感度・特異度）**：AUC値の高さ（0≦AUC≦1）
  - 24データセットの平均%AUC値

ランキング法	正規化法 MAS	正規化法 RMA	正規化法 DFW	
WAD	**96.7 (1)**	91.4 (6)	91.4 (5)	FC系
AD	93.8 (6)	93.1 (2)	92.2 (2)	
FC	93.6 (7)	**93.1 (1)**	92.2 (2)	
RP	91.5 (8)	92.5 (3)	**92.5 (1)**	
modT	95.7 (5)	91.4 (5)	90.1 (7)	*t*検定系
samT	96.0 (3)	91.2 (8)	90.0 (8)	
shrinkT	95.7 (4)	91.3 (7)	91.5 (4)	
ibmT	96.3 (2)	91.8 (4)	90.3 (6)	

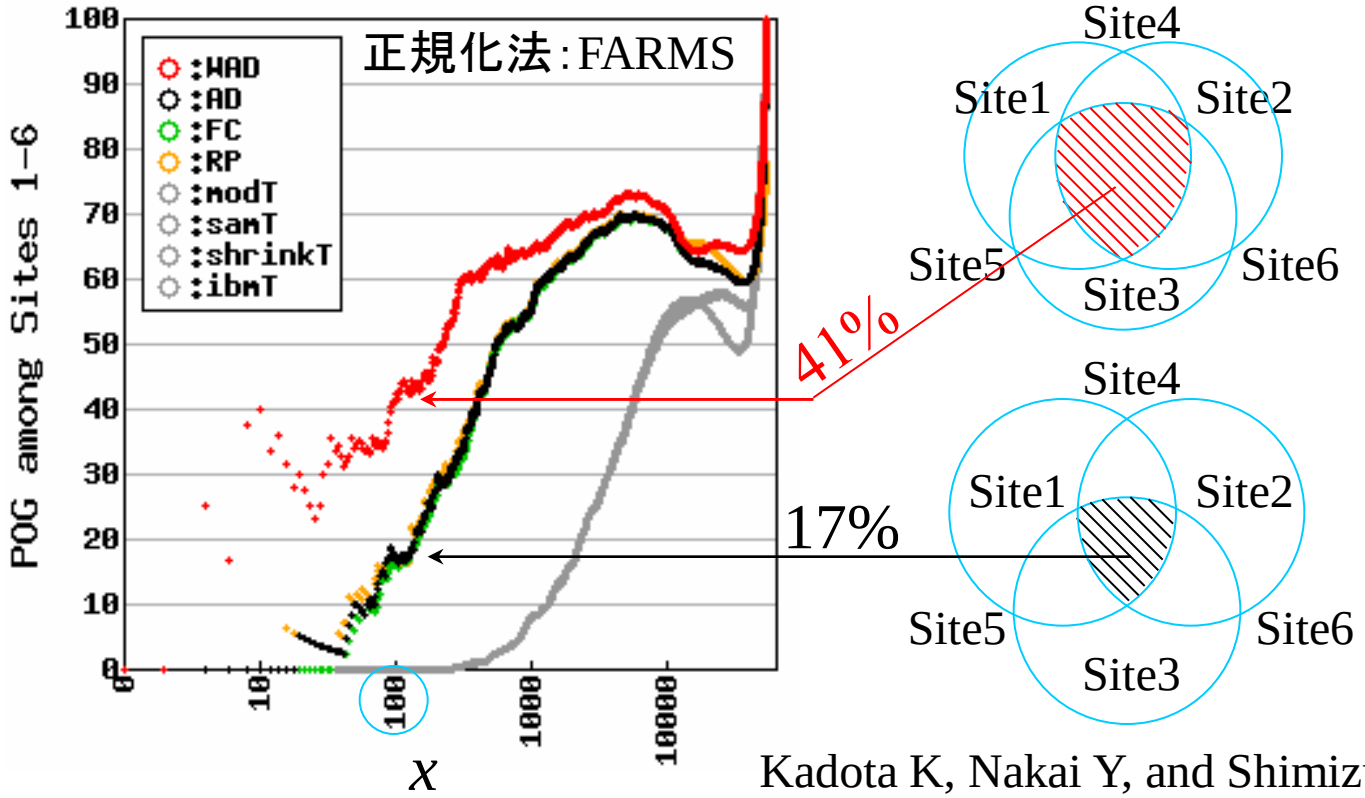
従来：
- 正規化法だけの比較
  → DFWがいいらしい…
- ランキング法だけの比較
  → ibmT（*t*検定系）がいいらしい…
- 感度・特異度の高い手法は*t*検定系

本人材養成プログラムの成果：
- WAD法の開発
- 正規化法とランキング法の相性
- 組み合わせのガイドライン策定
- 感度・特異度の高い手法はFC系

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

2. 発現変動遺伝子検出法：WAD

- **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≦1）
  - POG値が高い → ランキング結果の頑健性（再現性）が高い方法
  - 比較例：MAQCのサンプルC 5例 vs. サンプルD 5例の同じ実験を6 sitesで行い、Site間でランキング上位 *x*個（横軸）の共通遺伝子数（の割合；縦軸）を比較



上位100個の集合

最も再現性の高い手法はAD法である
→ WAD法である

「感度・特異度」と「再現性」は両立しない
→ 両立可能である

Kadota K, Nakai Y, and Shimizu K, *Algorithm Mol. Biol.*, **4**:7 2009

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

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

$= 4.08$

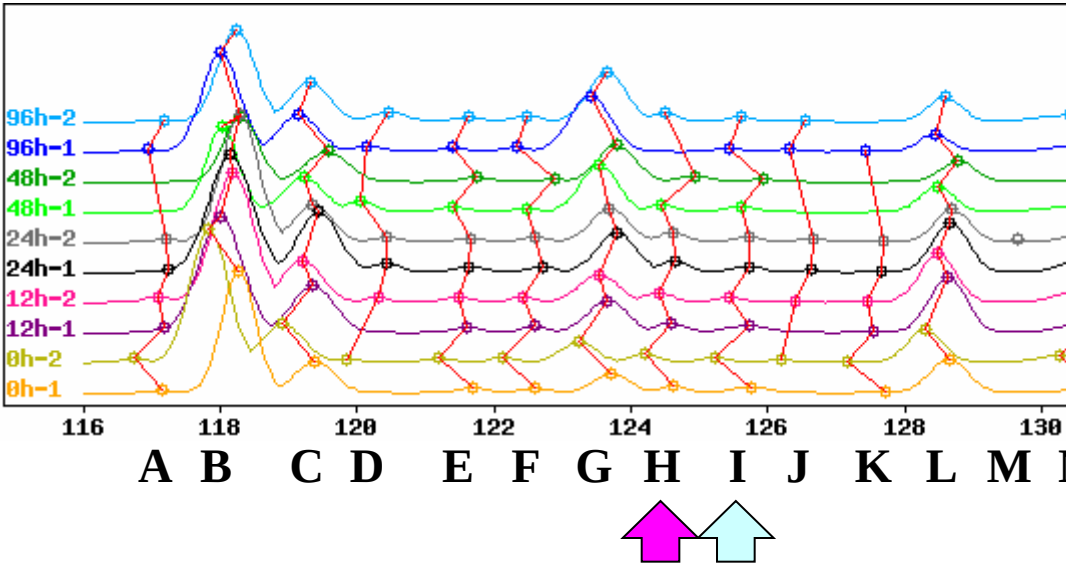
$w_i = \dfrac{x_i - \min(x)}{\max(x) - \min(x)}$ より

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

3. 波形補正およびピークアラインメント法：GOGOT

- 電気泳動波形解析の問題
  - 比較する実験数が増えるほど、同一転写物（遺伝子）の認識（アラインメント）精度が下がる

理想的なアラインメント



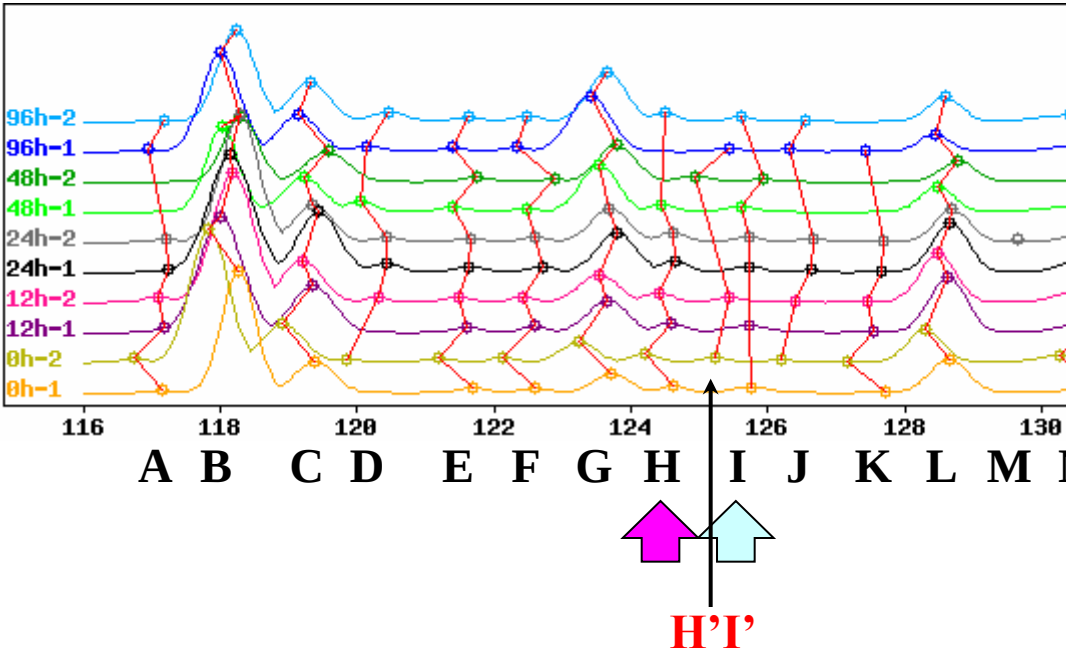
遺伝子発現行列

	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	**33**	**41**	**47**	**49**	**55**	**48**	**34**	**27**		**43**
I	**28**	**24**	**34**	**25**	**30**	**27**	**25**	**14**	**18**	**23**
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

3. 波形補正およびピークアラインメント法：GOGOT

- 電気泳動波形解析の問題
  - 比較する実験数が増えるほど、同一転写物（遺伝子）の認識（アラインメント）精度が下がる

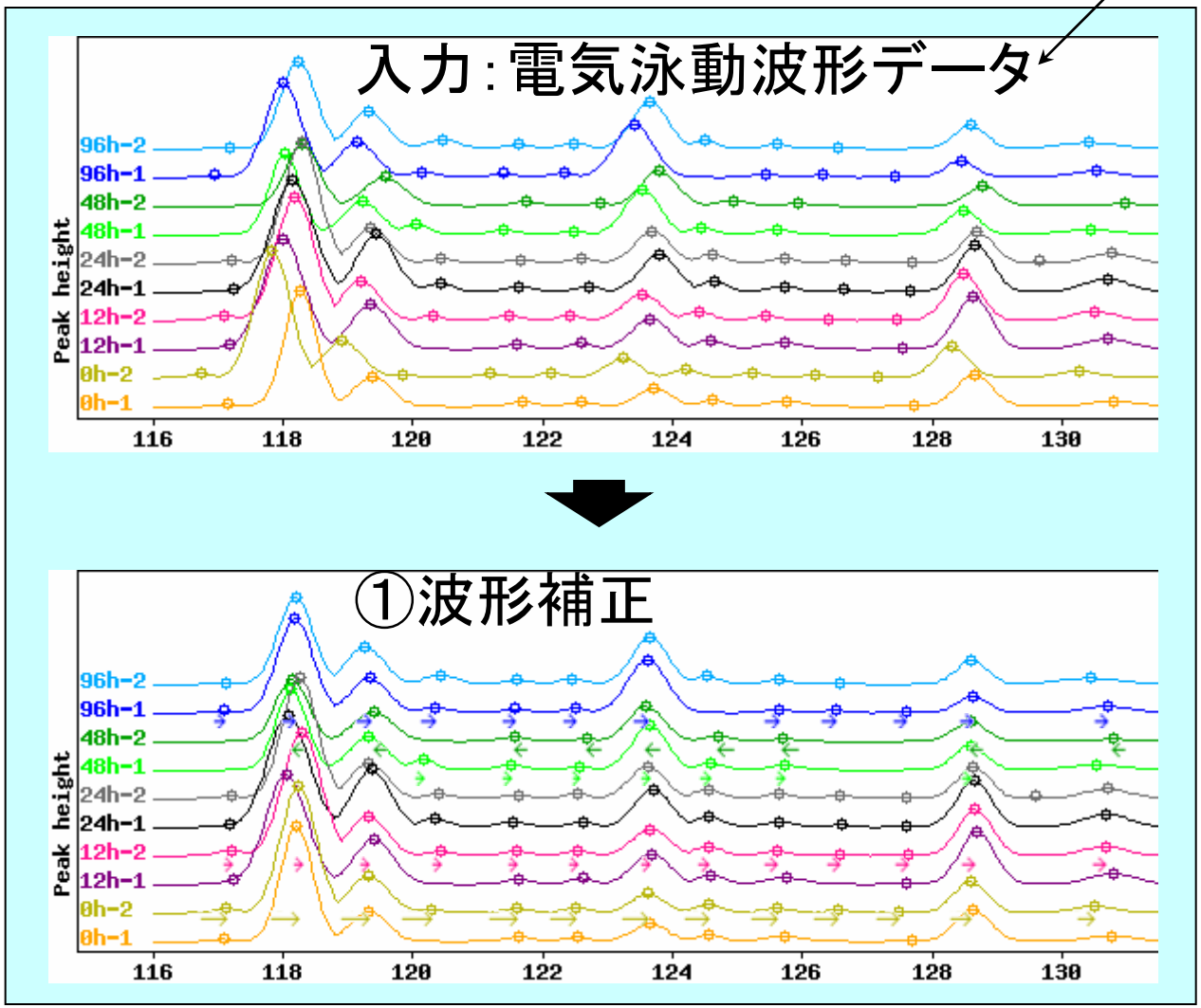
現実…



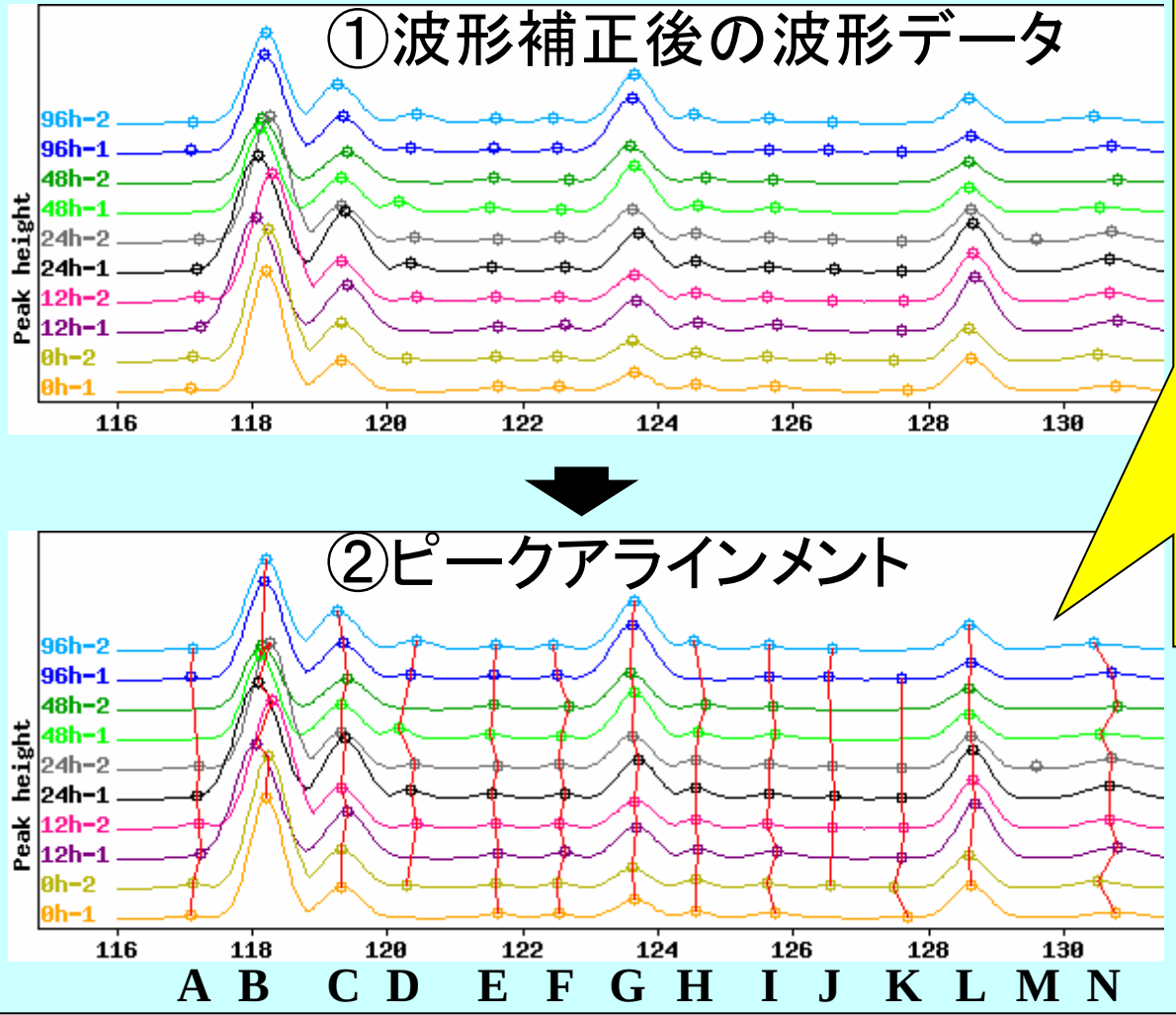
遺伝子発現行列

	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	**33**	**41**	**47**	**49**	**55**	**48**	**34**			**43**
H'I'		**24**		**25**				**27**	**18**	
I	**28**		**34**		**30**	**27**	**25**	**14**		**23**
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

3. 波形補正およびピークアラインメント法：GOGOT



3. 波形補正およびピークアラインメント法：GOGOT



出力：遺伝子発現行列

	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	33	41	47	49	55	48	34	27		43
I	28	24	34	25	30	27	25	14	18	23
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

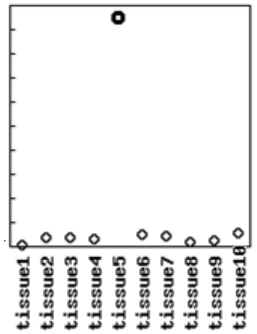
正確な遺伝子発現行列の自動作成が可能になった

→電気泳動波形データも各種マイクロアレイ解析手法の適用が可能

農学生命科学分野への適用例

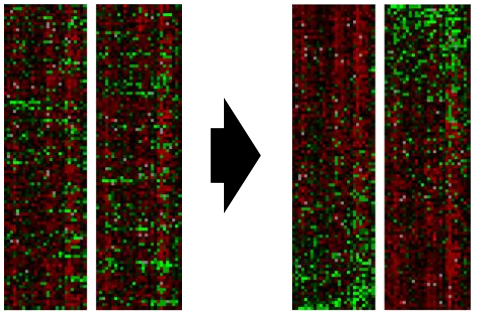
1. 組織特異的遺伝子検出法：ROKU

- 栄養条件特異的遺伝子の検出（農学生命科学研究科の博士論文研究）



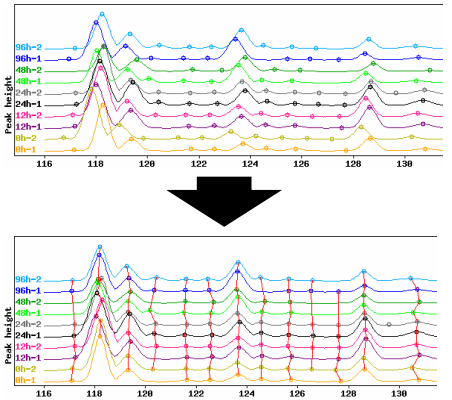
2. 発現変動遺伝子検出法：WAD

- 絶食状態 vs. 摂食状態での発現変動遺伝子検出（被養成者の修士論文研究）
- ゲノム情報を利用した分解プラスミドpCAR1の異種*Pseudomonas*属細菌内における挙動の理解（新谷先生の研究）



3. 波形補正およびピークアラインメント法：GOGOT

- 環境中から単離したバクテリアのrep-PCR DNAフィンガープリンティング（被養成者の研究）
- 微生物コミュニティを変性剤密度勾配電気泳動法（被養成者の研究）



（農学を含む）多くの生命科学研究で利用

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）

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

秋田県立大学　生物資源科学部

　小西 智一　准教授