マイクロアレイデータ解析結果の正しい?!解釈について

東京大学・大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
門田幸二（かどた　こうじ）
http://www.iu.a.u-tokyo.ac.jp/~kadota/
kadota@iu.a.u-tokyo.ac.jp

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

講義内容

- アレイデータの正規化（前処理）
  - 生データ　→　遺伝子発現行列
- クラスタリング
- 発現変動遺伝子（DEG）の同定
  - 二群間比較
    - 評価基準、評価法、および（Affymetrixチップの）ガイドライン
  - 多サンプル間比較
    - 組織特異的遺伝子
- 機能解析（GSEA解析）
  - Gene Ontology解析
  - パスウェイ解析

(Rで)マイクロアレイデータ解析 Microarray data analysis using R (last modified 2009/11/17)

What's new?
・2009年11月20, 24日13:00-マイクロアレイデータ解析講習会を開催します。11/24開催のほうで1名のみ(先着順で)追加募集します。(2009/11/17, 14:14掲載)
・発現変動遺伝子でないものの割合をざっと調べるための手段などを掲載しました。(2009/11/6)
・Bioconductorのversion2.5がリリースされたので、そこにリンクを張りなおしました。(2009/11/4)
・入力データ中に「'」があるとちゃんと読み込めなかった(例:「2'-PDE」というGene symbolの行で読み込みがストップしてしまう)のでread.table関数での読み込み時に「quote="'"」というオプションを追加しました。(2009/10/16)
・全体的に必要に応じて変更すべき箇所を前半部分に移動させて、エラーがより起こりにくくするなどしています。(2009/7/10-9/10)
・Affymetrix GeneChipデータ解析を行う上での推奨ガイドラインを掲載しました(2009/4/24)
・二群間比較用の私の最新手法を掲載しました(2008/6/26)

- はじめに (last modified 2009/8/7)
- Rのインストールと起動(last modified 2009/11/4) NEW
- Rの昔のバージョンのインストール (last modified 2009/8/21)
- 使用例(初心者向け)(last modified 2009/7/9)
- サンプルマイクロアレイデータ (last modified 2009/8/4)
- 発現データ取得 | Affymetrix data全体 | Celsius (Day 2007) (last modified 2007/11/13)
- 発現データ取得 | Gene Expression Omnibus (GEO)から | GEOquery (Davis 2007) (last modified 2009/8/5)
- 発現データ取得 | ArrayExpressから | ArrayExpress (last modified 2009/5/28)
- アノテーション情報取得 | Rのパッケージから (last modified 2009/8/5)
- アノテーション情報取得 | GEOから (last modified 2009/8/5)
- 正規化(cDNA or two-color or 二色法) について(last modified 2008/3/31)
- 正規化 | Stanford型 (or cDNA)マイクロアレイ (package: limma)
- 正規化 | Stanford型 (or cDNA)マイクロアレイ (package: marray)
- 正規化 | Stanford型 (or cDNA)マイクロアレイ GPA (Xiong 2008) (last modified 2008/3/10)
- 正規化(Affymetrix) について(last modified 2009/7/9)
- 正規化 | Affymetrix GeneChip | RMA++, Extrapolation Averaging (Harbron 2007) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | RMA+, Extrapolation Strategy, refRMA (Harbron 2007) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | DFW (Chen 2007) (last modified 2009/11/2) NEW
- 正規化 | Affymetrix GeneChip | FARMS (Hochreiter 2006) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | multi-mgMOS (Liu 2005) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | GLA (Zhou 2005) (last modified 2007/4/20)
- 正規化 | Affymetrix GeneChip | GCRMA (Wu 2004) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | PLIER (Affymetrix 2004) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | PDNN (Zhang 2003) (last modified 2009/8/6)
- 正規化 | Affymetrix GeneChip | VSN (Huber 2002) (last modified 2009/8/6)

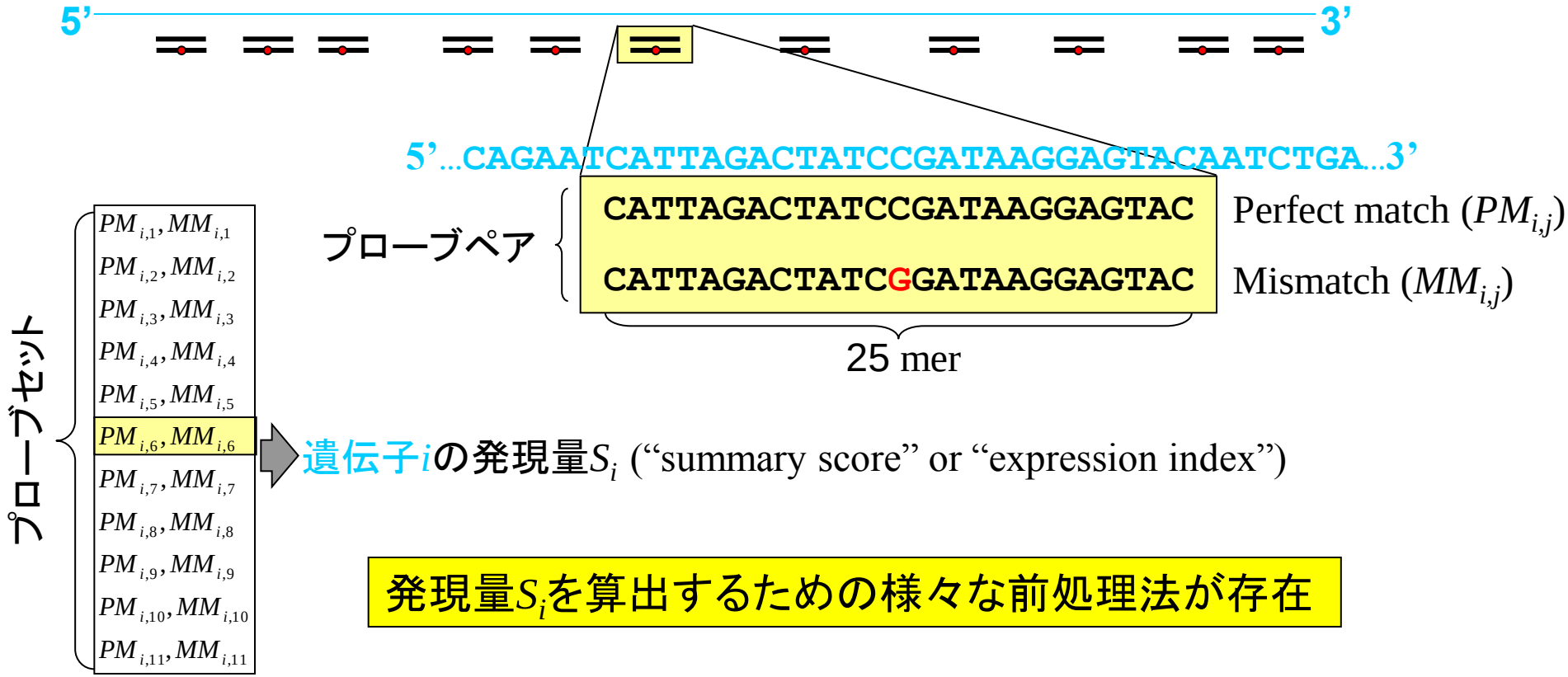
アレイデータの正規化（前処理）

- 実験によって得られた生のシグナル強度をそのまま利用することは普通はやりません
  - 二色法：蛍光色素（Cy3 and Cy5）の取り込み効率補正
  - 一色法：シグナルゲイン?!の補正

「こうであるべき！」という仮定を置いて、それを満たすような正規化を行った後のデータを利用する

Affymetrix製チップ解析戦略

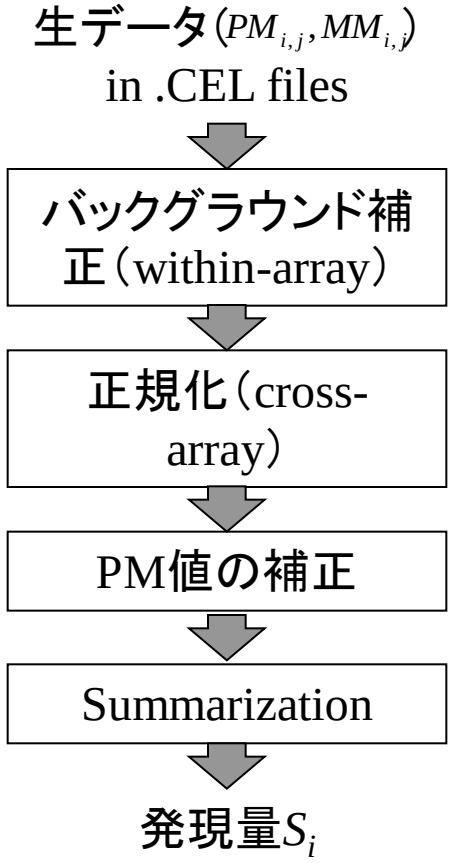
- 遺伝子iの発現量S_iをn_i(n_i=11~20)種類のプローブペアのシグナル強度をもとに計算



発現量S_iを算出するための様々な前処理法が存在

Affymetrix製チップ解析戦略（様々な前処理法）

- 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)



Availability: The R code for DFW is available upon request.

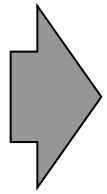
グローバル正規化

- 仮定：各サンプルから測定されたmRNAの全体量は一定

  チップ上の遺伝子数が少ない場合は非現実的だが、数千～数万種類の遺伝子が搭載されているので妥当（だろう）

	sample1	sample2	Ratio (sample1/sample2)	$\log_2$(Ratio)
gene1	10.5	12.4	0.84	-0.243
gene2	6.4	7.1	0.91	-0.141
gene3	8.0	8.5	0.94	-0.086
gene4	10.8	11.4	0.95	-0.075
gene5	5.6	6.7	0.83	-0.262
gene6	8.4	8.9	0.94	-0.090
gene7	6.2	7.0	0.90	-0.159
gene8	6.1	6.8	0.90	-0.145
gene9	6.6	6.5	1.01	0.010
gene10	5.1	5.8	0.89	-0.165
				-0.136

nomalization



$\log_2$(Ratio)
-0.107
-0.005
0.049
0.061
-0.126
0.045
-0.023
-0.010
0.145
-0.030
0.000

Quantile正規化

- 仮定：順位が同じならシグナル強度も同じ

正規化前

sample1	sample2
10.5	12.4
6.4	7.1
8.0	8.5
10.8	11.4
5.6	6.7
8.4	8.9
6.2	7.0
6.1	6.8
6.6	6.5
5.1	5.8

列ごとにソート →

sample1	sample2
5.1	5.8
5.6	6.5
6.1	6.7
6.2	6.8
6.4	7.0
6.6	7.1
8.0	8.5
8.4	8.9
10.5	11.4
10.8	12.4

行ごとの平均を算出 →

Average
5.4
6.1
6.4
6.5
6.7
6.8
8.3
8.6
10.9
11.6

対応する行の要素の元の位置に平均値を代入 →

正規化後

sample1	sample2
10.9	11.6
6.7	6.8
8.3	8.3
11.6	10.9
6.1	6.4
8.6	8.6
6.5	6.7
6.4	6.5
6.8	6.1
5.4	5.4

データセット中のサンプル数が変わると結果が変わる

正規化 → 遺伝子発現行列

二群間比較

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

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*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

時系列データ

	T1	T2	T3	T4	…
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*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

様々な解析が可能な状態

手順1:前処理法の適用

- Affymetrix GeneChipの場合
  - 様々な前処理法を適用し、複数の遺伝子発現行列データを得る
  - 例）**MAS, RMA**, qFARMS or DFW
- その他のメーカーの場合
  - メーカー推奨のやり方に従って、遺伝子発現行列データ（基本的に一つのみ）を得る

理由:どの前処理法を使うかでサンプル間クラスタリング（後述）の結果が大きく異なりうるから

クラスタリング

- サンプルの属性情報（癌 or 正常など）を使わずに、発現情報のみを用いて発現パターンの類似した遺伝子（またはサンプル）をクラスター（群）にしていく手法（Unsupervised learning

二群間比較

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

多サンプル

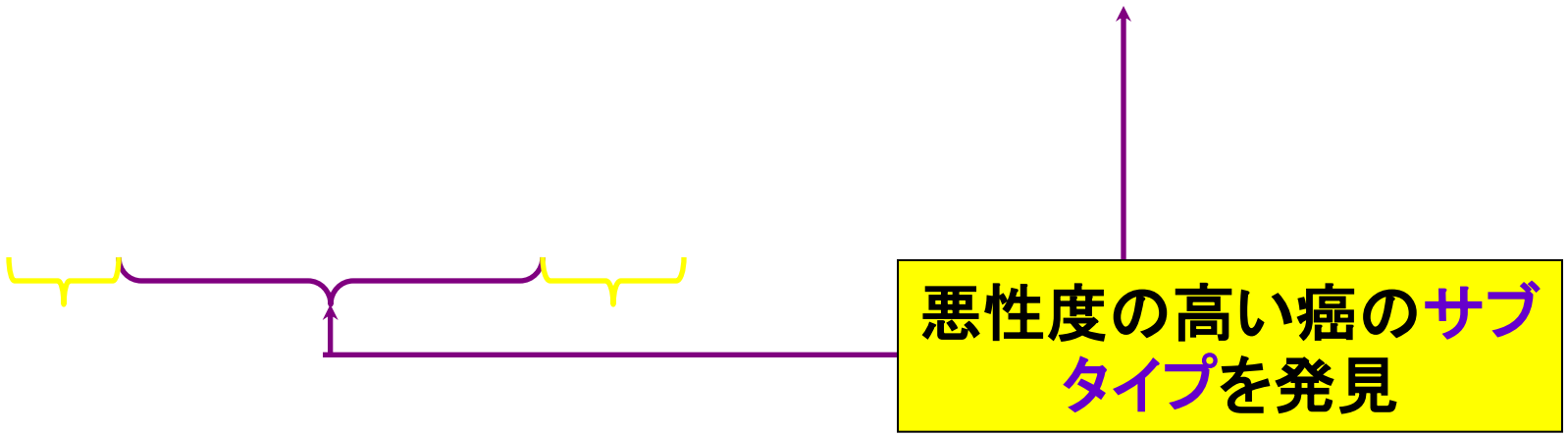
	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

時系列解析

	T1	T2	T3	T4	…
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*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

サンプル間クラスタリングの例

- メラノーマサンプル



クラスタリング

- 階層的クラスタリング
  - 発現パターンの類似した遺伝子を集めて系統樹を作成
- 非階層的クラスタリング
  - *K*-meansクラスタリング
    - 「*K*個のクラスターに分割（*K*の数は主観的に決定）する」と予め指定し、各クラスター内の遺伝子（サンプル）間の距離の総和が最小になるような*K*個のクラスターを作成
  - 自己組織化マップ（SOM）
  - 主成分分析（PCA）

距離(類似度)の定義

- 遺伝子(or サンプル)xとyの発現パターンの距離D

相関係数 $r_{xy} = \dfrac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ $\quad (-1 \le r_{xy} \le 1)$

i	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
…	…	…
n	x_n	y_n

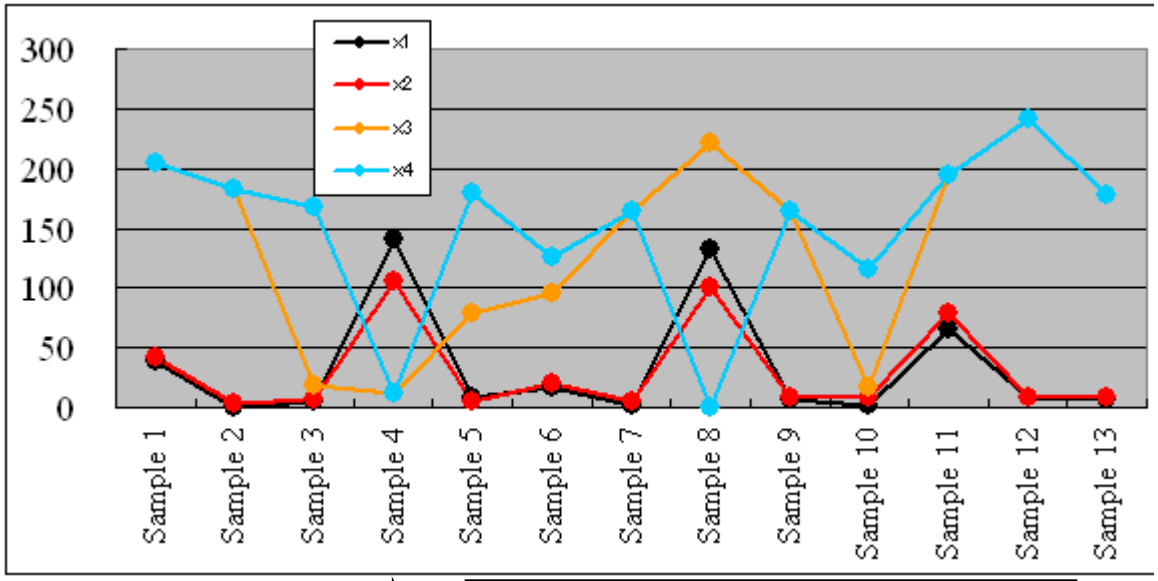
- xとyの発現パターンが酷似 $\rightarrow$ $r \approx 1$
- xとyの発現パターンがばらばら $\rightarrow$ $r \approx 0$
- xとyの発現パターンがほぼ正反対 $\rightarrow$ $r \approx -1$

距離 $D = 1 - r \quad (0 \le D \le 2)$

- $r = 1 \rightarrow D = 1 - 1 = 0$
- $r = 0 \rightarrow D = 1 - 0 = 1$
- $r = -1 \rightarrow D = 1 - (-1) = 2$

階層的クラスタリング

1. 遺伝子間距離を計算



例：4遺伝子の場合

	x^1	x^2	x^3	x^4
Sample 1	38	42	204	204
Sample 2	0	3	182	182
Sample 3	6	6	19	168
Sample 4	141	106	11	11
Sample 5	8	5	79	179
Sample 6	16	20	96	126
Sample 7	2	5	164	164
Sample 8	132	101	222	0
Sample 9	7	9	164	164
Sample 10	2	8	16	116
Sample 11	66	79	195	195
Sample 12	8	9	241	241
Sample 13	6	8	177	177

距離 $D = 1 - r \quad (0 \le D \le 2)$ ⇒ 距離 $D = \dfrac{1-r}{2} \quad (0 \le D \le 1)$

相関係数 $r_{1,2} = 0.98 \rightarrow$ 距離 $D_{1,2} = \dfrac{1-0.98}{2} = 0.01$

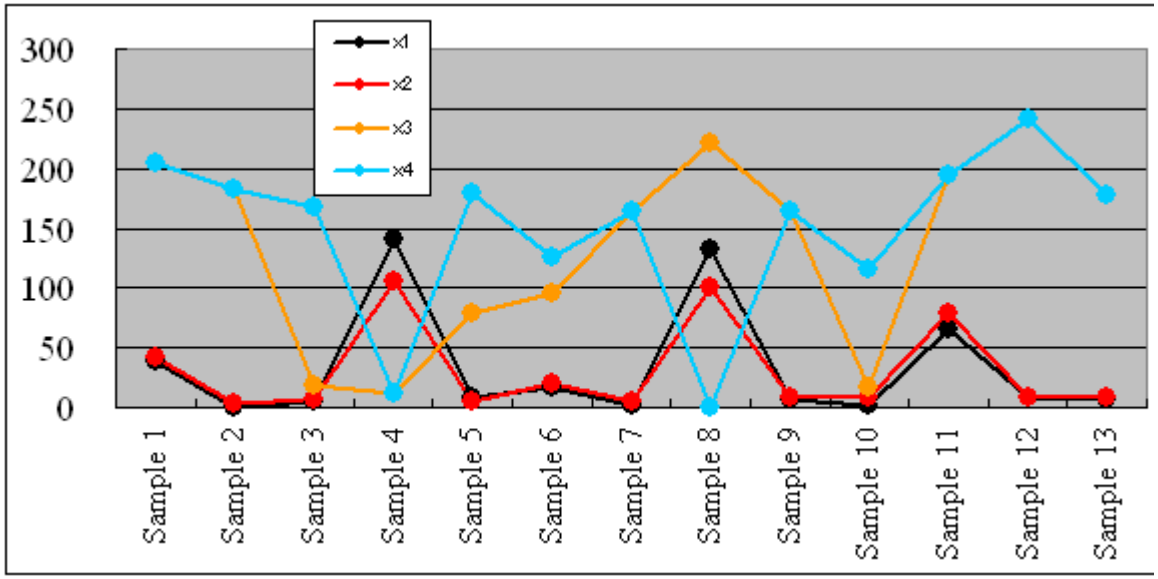
相関係数 $r_{1,3} = -0.01 \rightarrow$ 距離 $D_{1,3} = \dfrac{1-(-0.01)}{2} = 0.50$

相関係数 $r_{1,4} = -0.78 \rightarrow$ 距離 $D_{1,4} = \dfrac{1-(-0.78)}{2} = 0.89$

…

階層的クラスタリング

2. 距離行列を作成



距離 $D_{1,2} = \frac{1-0.98}{2} = 0.01$

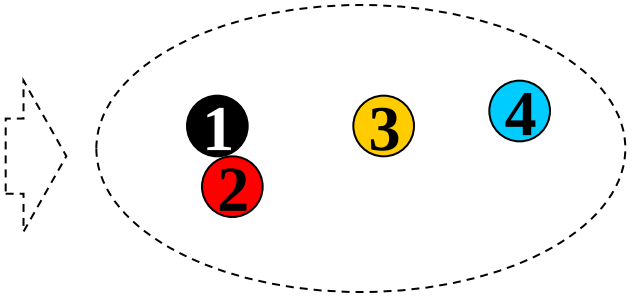
距離 $D_{1,3} = \frac{1-(-0.01)}{2} = 0.50$

距離 $D_{1,4} = \frac{1-(-0.78)}{2} = 0.89$

…

	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32

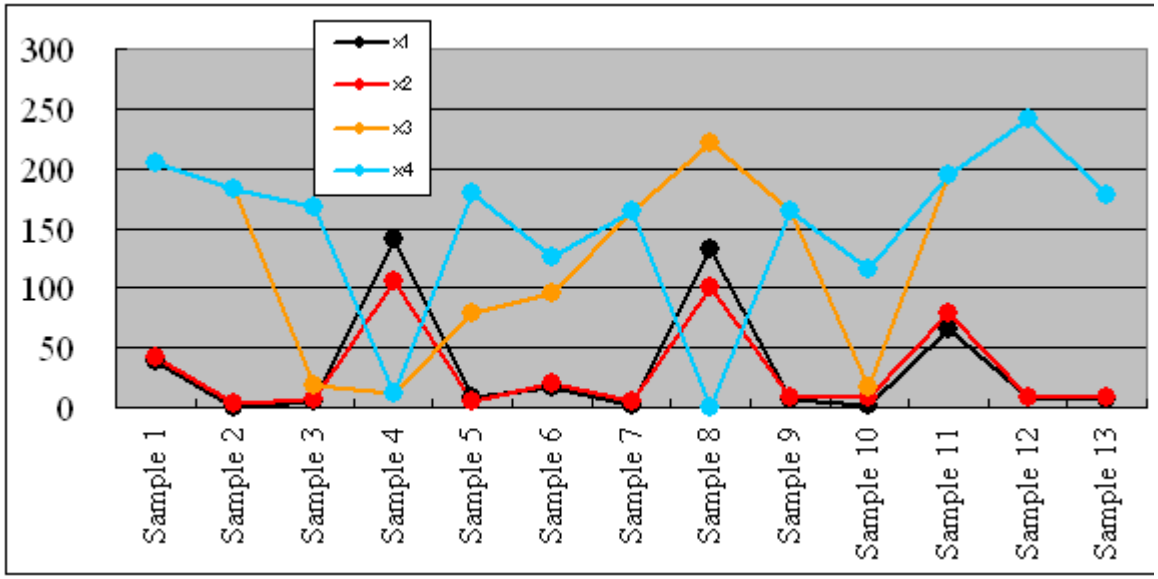
距離行列



イメージ

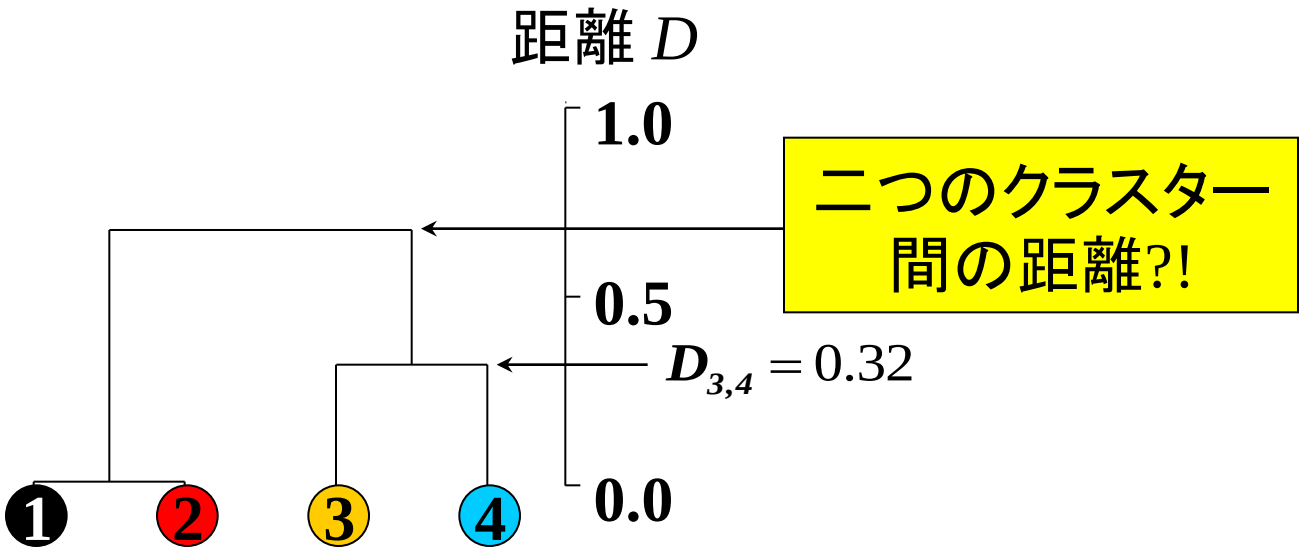
階層的クラスタリング

3. 樹形図を作成



	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32

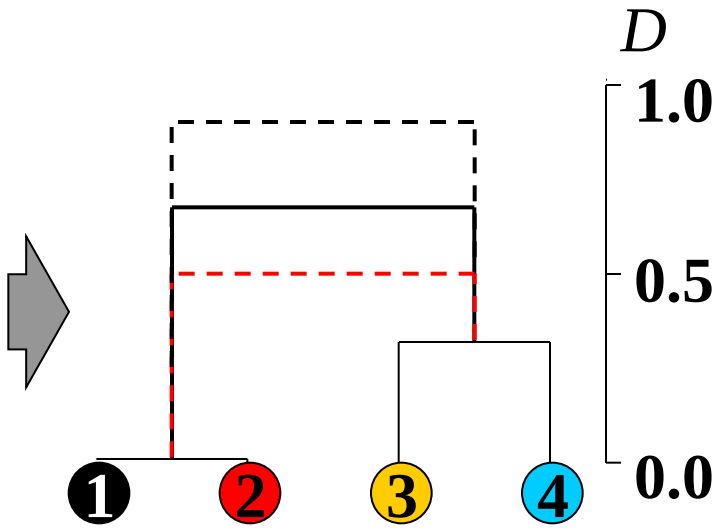
距離行列



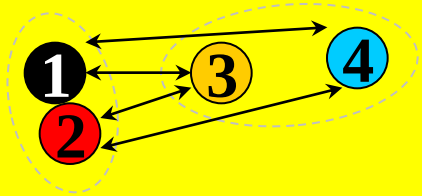
階層的クラスタリング

3. 樹形図を作成

	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32



平均連結法の場合



$$(D_{1,3}+D_{1,4}+D_{2,3}+D_{2,4})/4$$
$$=(0.50+0.89+0.47+0.84)/4$$
$$=0.68$$

単連結法の場合

$$\min(D_{1,3},D_{1,4},D_{2,3},D_{2,4})$$
$$=0.47$$

完全連結法の場合

$$\max(D_{1,3},D_{1,4},D_{2,3},D_{2,4})$$
$$=0.89$$

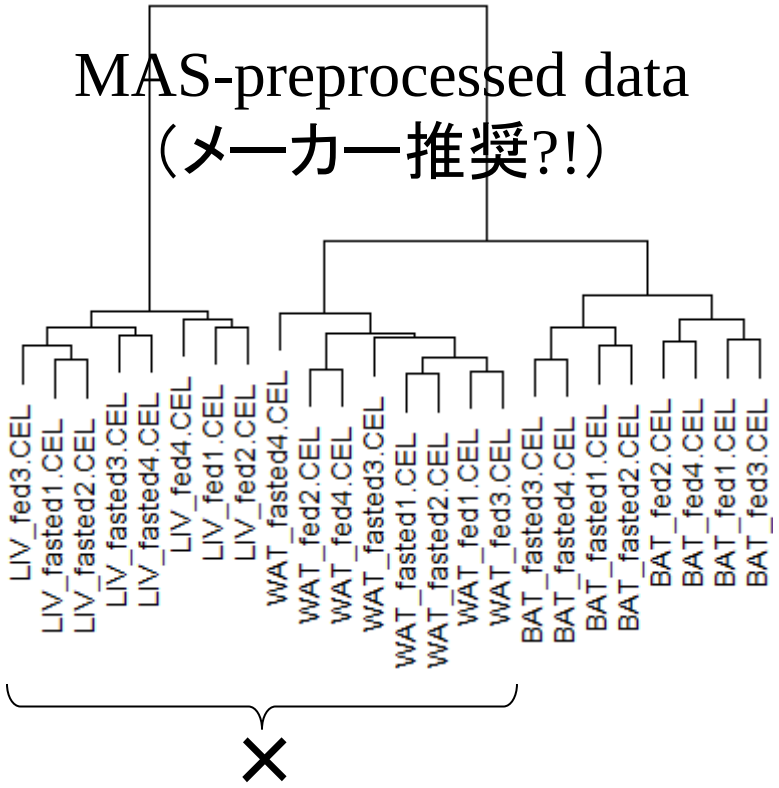
手順2：サンプル間クラスタリング

- Affymetrix GeneChipの場合
  - 様々な前処理法を適用して得られた遺伝子発現行列データごとに行う
  - 結果を眺めて、反復実験結果が同一クラスターに含まれる前処理法のデータを採用
    - RMAがよかった場合：それだけを採用でよし
    - それ以外の場合で残りの二つの前処理法の結果のトポロジーが同じ場合：二つのデータを同時並行で解析（論文では一つのみ）
    - 三つの結果がいずれも異なっていた場合：ご愁傷さまです…
- その他のメーカーの場合
  - メーカー推奨のやり方に従って、遺伝子発現行列データ（基本的に一つのみ）を得る

クラスタリング結果の解釈例

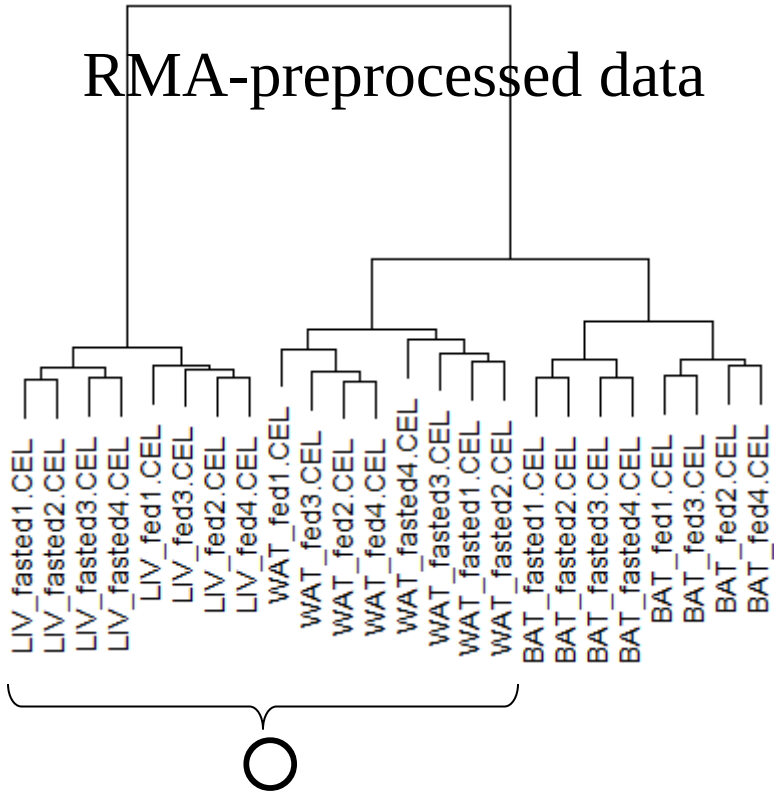
- 肝臓（LIV）、白色脂肪（WAT）、褐色脂肪（BAT）
  - 通常（fed）vs. 24時間絶食（fasted）

MAS-preprocessed data
（メーカー推奨?!）



×

RMA-preprocessed data

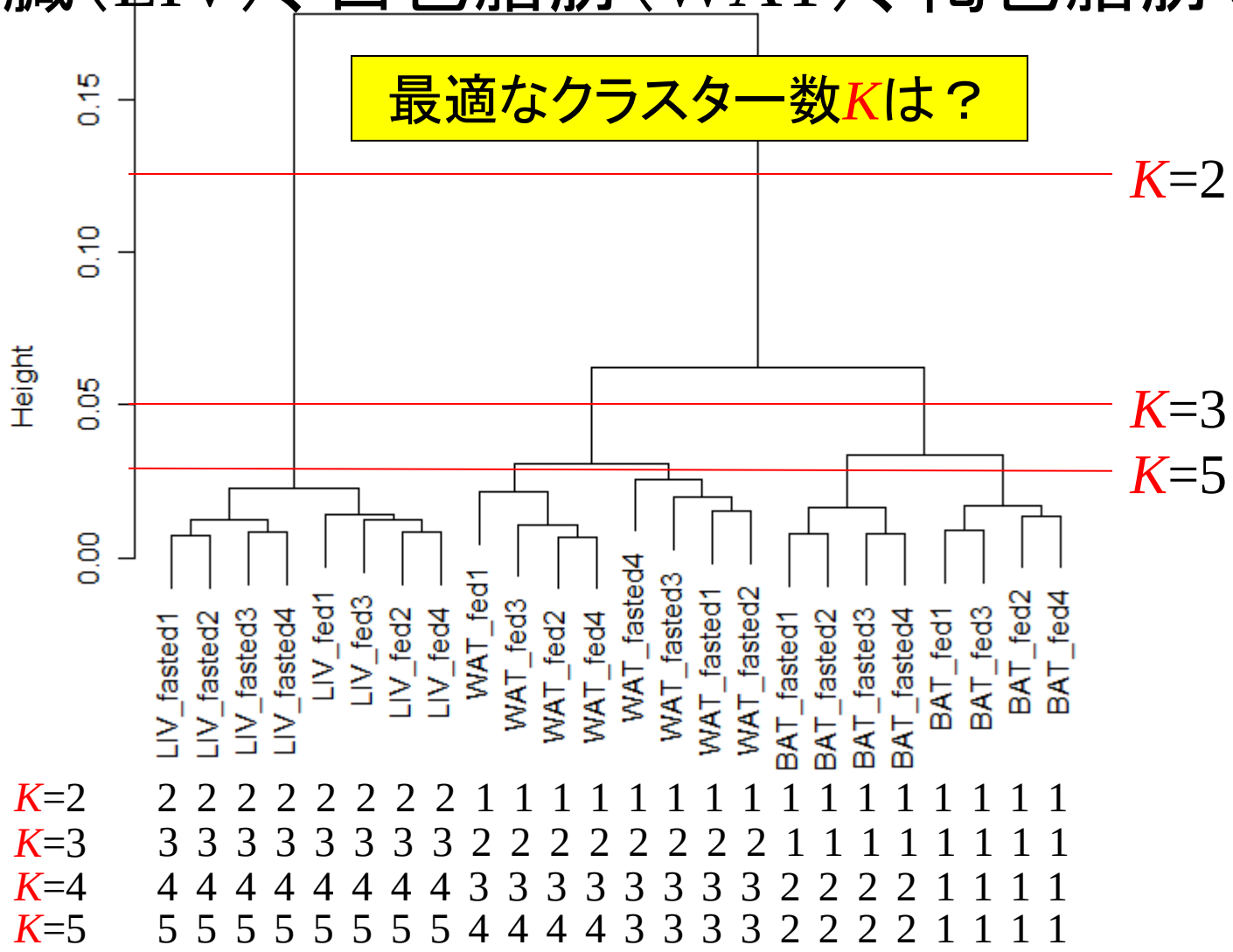


○

RMAがいいと判断（この場合）

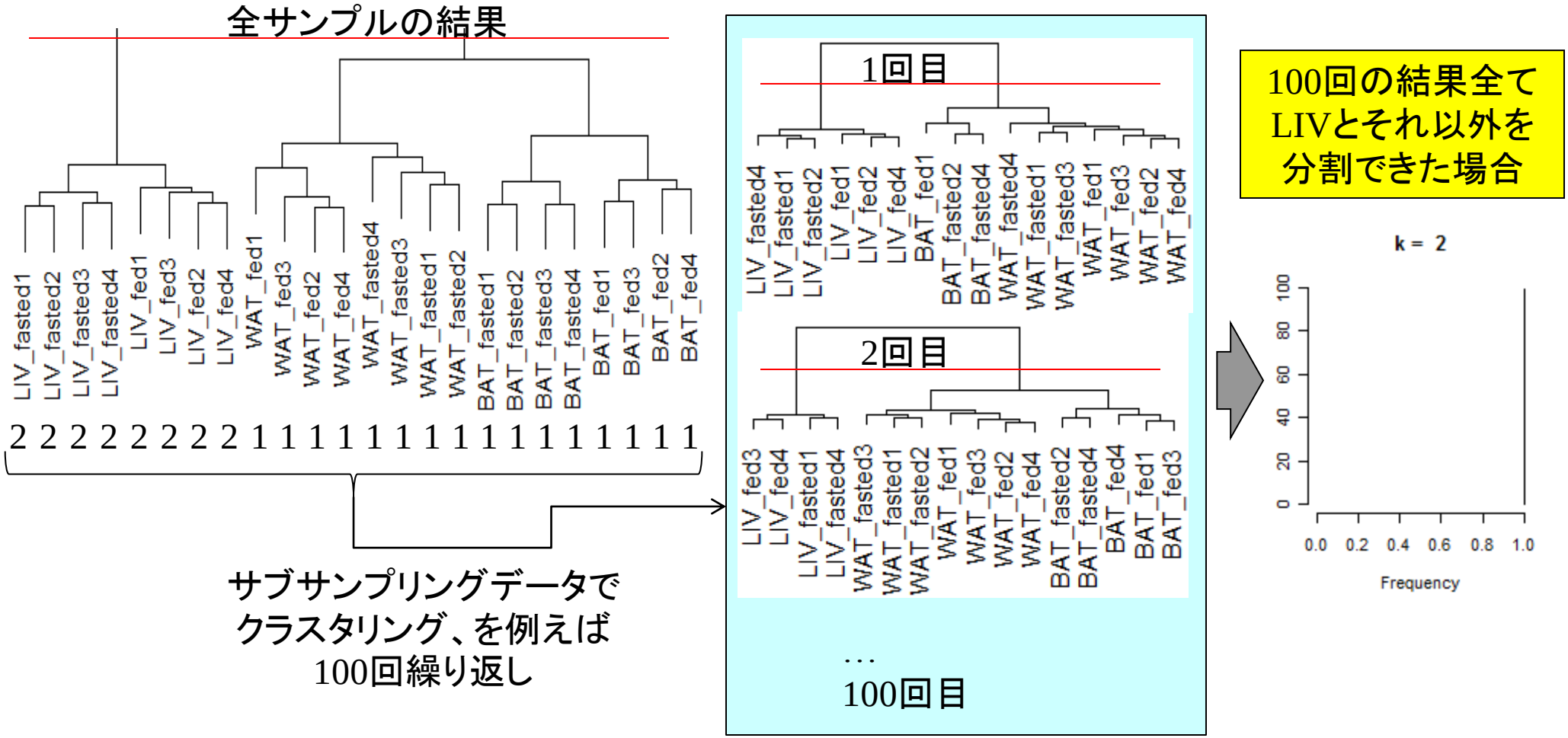
階層的クラスタリング例

- 肝臓(LIV)、白色脂肪(WAT)、褐色脂肪(BAT)



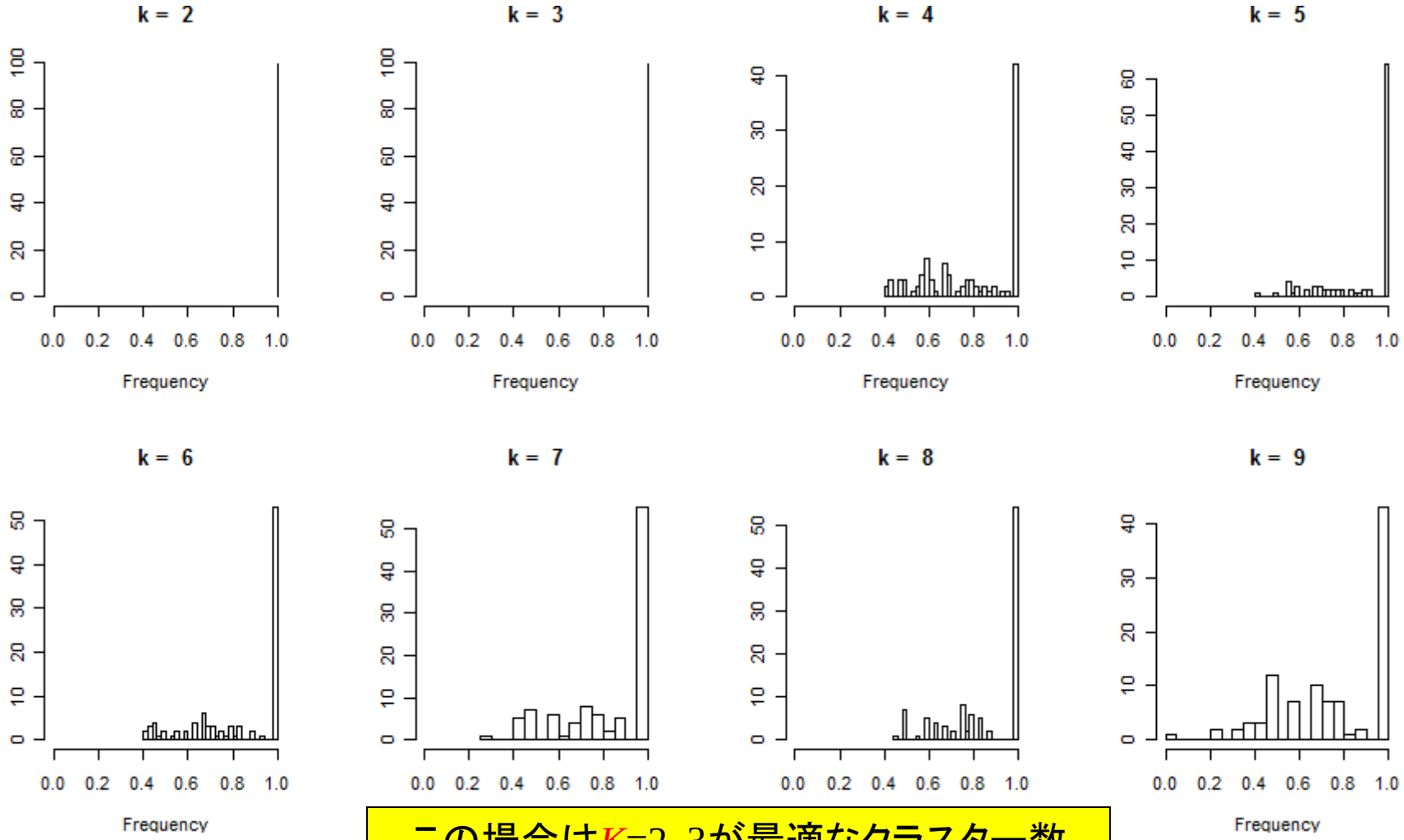
最適なクラスター数を見積もる方法

- 様々なKについて（例えばK=2）全サンプル（n）のクラスタリング結果をK個に分割した結果とサブサンプル（例えばn*0.7）のクラスタリング結果をK個に分割した結果の類似度を計算



手順2’：クラスター数をチェック

- *K*の値をいくつか試して（例では2~9）、最適な*K*の値を同定



この場合は*K*=2, 3が最適なクラスター数

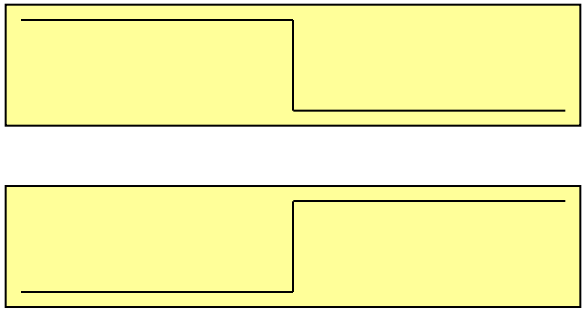
言いたいことと同じだったらラッキー

二群間比較

- 例1）
  - A群：癌サンプル
  - B群：正常サンプル

  →癌と正常で発現の異なる遺伝子

	A群			B群		
	A1	A2	…	B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	



二群間比較解析

- 例）急性白血病
  - A群：リンパ性（27 サンプル）
  - B群：骨髄性（11サンプル）

白血病のタイプで発現の
異なる遺伝子群を同定

二群間比較（解析手法）

- 倍率変化（Fold change; FC）に基づくランキング法
  - **2-fold, 3-fold (FC)**
  - The limit fold change model (Mutch *et al.*, *BMC Bioinformatics*, 2002)
  - **Rank product** (RP; Breitling *et al.*, *FEBS Lett.*, 2004)
  - **WAD** (Kadota *et al.*, *Algorithm. Mol. Biol.*, 2008)
  - …
- *t*-統計量に基づくランキング法
  - a signal-to-noise statistic (Golub *et al.*, Science, 1999)
  - **Student's (or Welch) *t*-test**
  - **SAM** (**samT**; Tusher *et al.*, *PNAS*, 2001)
  - **Samroc** (Broberg, P., *Genome Biol.*, 2003)
  - **a moderated *t* statistic** (Smyth, GK., *Stat. Appl. Genet. Mol. Biol.*, 2004)
  - **Intensity-based moderated *t* statistic** (IBMT; Sartor *et al.*, *BMC Bioinformatics*, 2006)
  - **Shrinkage *t* statistic** (Opgen-Rhein and Strimmer, *Stat. Appl. Genet. Mol. Biol.*, 2007)
  - …
- その他
  - Probability of Positive LogRatio (PPLR; Liu *et al.*, *Bioinformatics*, 2006)
  - FCPC (Qin *et al.*, *Bioinformatics*, 2008)

個々の遺伝子の発現変動の度合いを調べる研究

二群間比較（t-統計量に基づくランキング法）

- 「二群間の平均の差が大きく」、「群内のばらつきが小さい」遺伝子iを抽出
- a signal-to-noise(S2N)統計量

$$R(i) = \frac{\overline{A^i} - \overline{B^i}}{U_{A^i} + U_{B^i}}$$

← 二群間の平均の差

U_{A^i}：A群内のばらつき　U_{B^i}：B群内のばらつき

標本平均　$\overline{A^i} = \frac{1}{n_A}\sum_{j=1}^{n_A} A_j^i$

標本分散　$S_{A^i}^2 = \frac{1}{n_A}\sum_{j=1}^{n_A}(A_j^i - \overline{A^i})^2$

不偏分散　$U_{A^i}^2 = \frac{1}{n_A - 1}\sum_{j=1}^{n_A}(A_j^i - \overline{A^i})^2$

$n_A = 6, n_B = 5, n = n_A + n_B$

対数変換（log2変換）後のデータ

		A群						B群				
i		A_1^i	A_2^i	A_3^i	A_4^i	A_5^i	A_6^i	B_1^i	B_2^i	B_3^i	B_4^i	B_5^i
1	gene1	6.4	6.3	6.5	6.4	6.5	6.4	3.6	4.4	4.2	3.7	4.1
2	gene2	5.8	6.9	6.7	5.6	6.4	6.6	2.8	5.5	1	3.5	4.2
3	gene3	3.9	4.8	5	3.2	4.8	5.4	5.6	5.5	5.7	5.6	5.6

$$R(1) = \frac{6.42 - 4.00}{0.08 + 0.35} = \frac{2.41}{0.43} = 5.64$$

$$R(2) = \frac{6.34 - 3.38}{0.54 + 1.65} = \frac{2.96}{2.20} = 1.35$$

$$R(3) = \frac{4.51 - 5.61}{0.81 + 0.07} = \frac{-1.11}{0.88} = -1.26$$

統計量の絶対値が大きい → 候補発現変動遺伝子

二群間比較（倍率変化に基づくランキング法）

- log比：（対数変換後のデータなので）*t*検定系の数式の分子のみに相当

$$R(i) = \log(FC) = \overline{A^i} - \overline{B^i}$$ ←二群間の平均の差

Average Difference（AD）統計量と私は呼んでいる

対数変換（log2変換）後のデータ

		A群						B群				
i		$A_1{}^i$	$A_2{}^i$	$A_3{}^i$	$A_4{}^i$	$A_5{}^i$	$A_6{}^i$	$B_1{}^i$	$B_2{}^i$	$B_3{}^i$	$B_4{}^i$	$B_5{}^i$
1	gene1	6.4	6.3	6.5	6.4	6.5	6.4	3.6	4.4	4.2	3.7	4.1
2	gene2	5.8	6.9	6.7	5.6	6.4	6.6	2.8	5.5	1	3.5	4.2
3	gene3	3.9	4.8	5	3.2	4.8	5.4	5.6	5.5	5.7	5.6	5.6

$$R(1) = 6.42 - 4.00 = 2.41$$
$$R(2) = 6.34 - 3.38 = 2.96$$
$$R(3) = 4.51 - 5.61 = -1.11$$

統計量の絶対値が大きい → 候補発現変動遺伝子

参考資料

二群間比較（倍率変化に基づくランキング法）

- WAD：log比を基本としつつ、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

二群間比較（倍率変化に基づくランキング法）

- Rank products (RP)：A群 vs. B群の総当たりの比を計算し、その順位の相乗平均を統計量とする

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

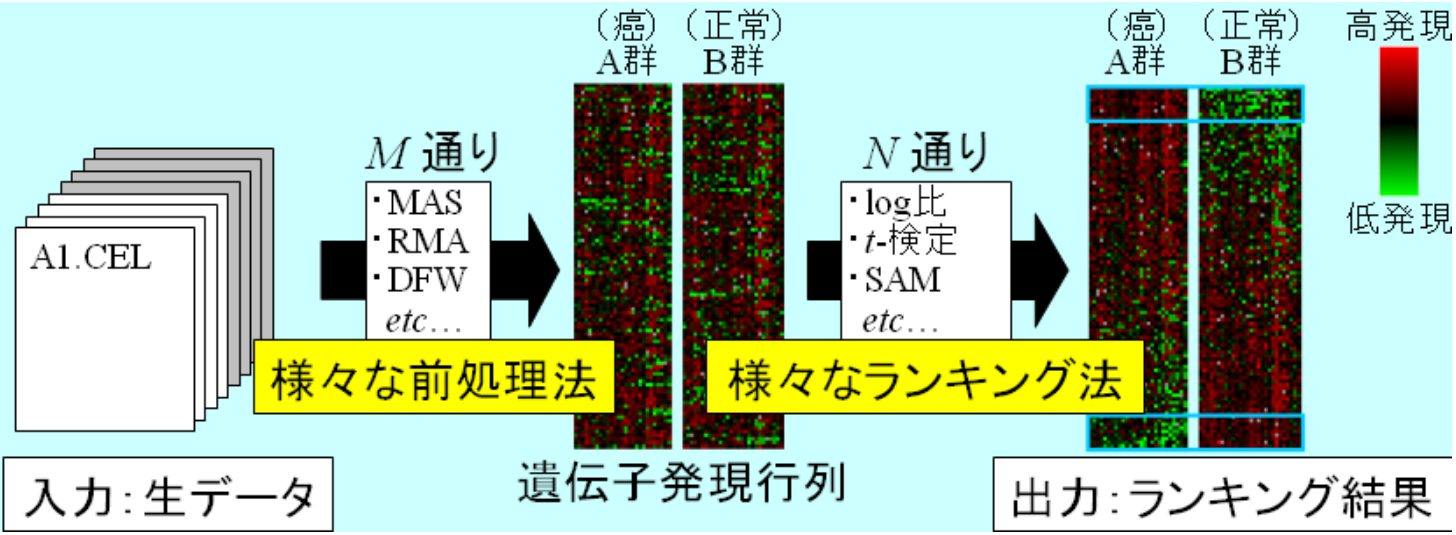
A1/B1	A1/B2	A1/B3	A2/B1	A2/B2	A2/B3	A3/B1	A3/B2	A3/B3
a11/b11	a11/b12	a11/b13	a12/b11	a12/b12	a12/b13	a13/b11	a13/b12	a13/b13
ai1/bi1	ai1/bi2	ai1/bi3	ai2/bi1	ai2/bi2	ai2/bi3	ai3/bi1	ai3/bi2	ai3/bi3
an1/bn1	an1/bn2	an1/bn3	an2/bn1	an2/bn2	an2/bn3	an3/bn1	an3/bn2	an3/bn3

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

評価の実際

- 例：Affymetrixの二群間比較（←最もよく研究されている）



- 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

- 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？

- 再現性

  同じサンプルの比較結果（発現変動遺伝子リスト）が場所間でどれだけ一致しているか？

「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**）

  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				

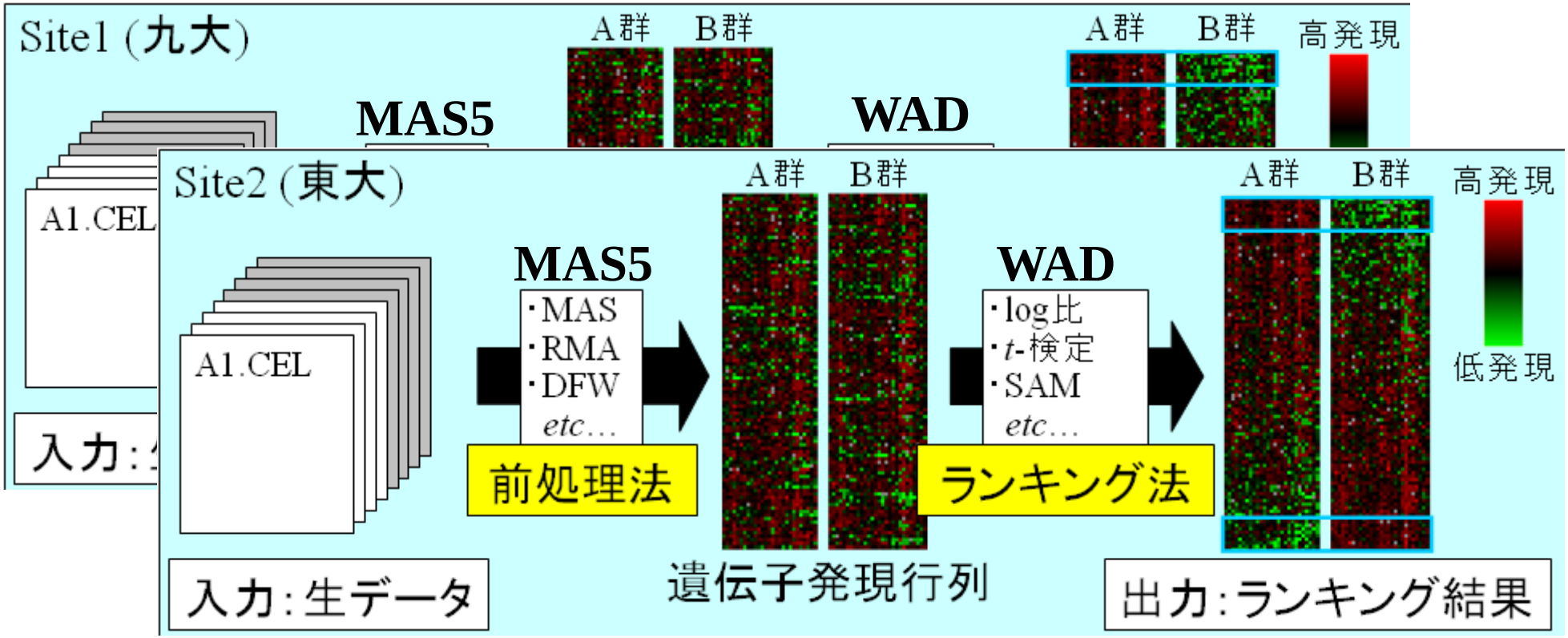
$\lvert\log_2(C/D)\rvert$
0.8
1.9
0.5
1.3
1.4

$\lvert\log_2(C/D)\rvert$	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%

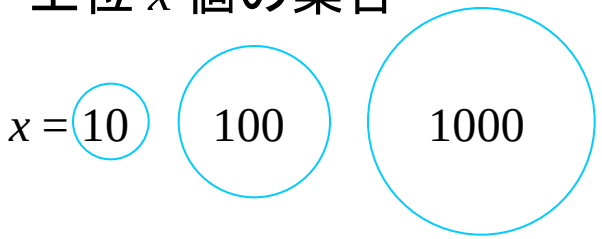
「再現性」を一致度で評価

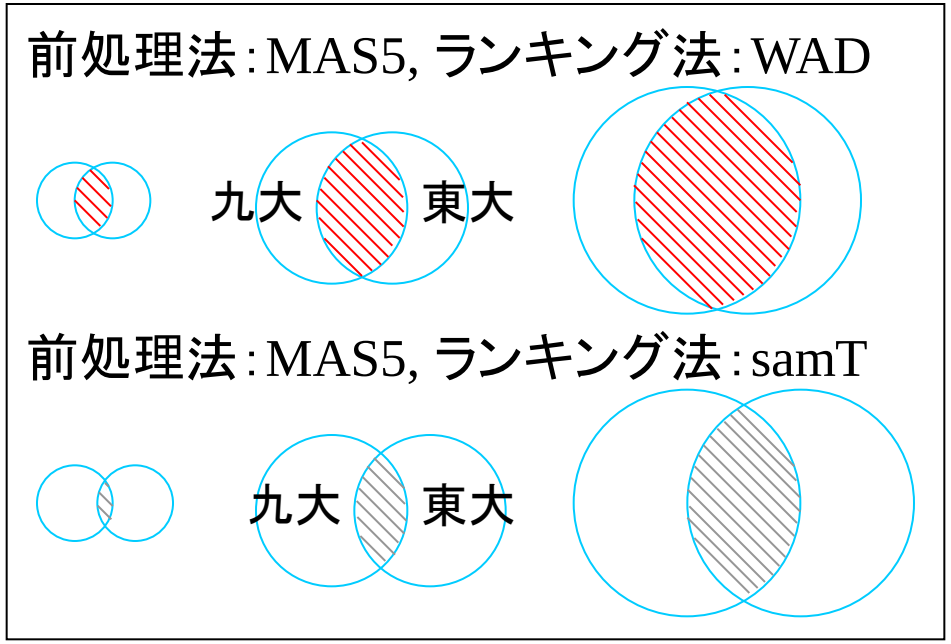
- MicroArray Quality Control (MAQC) プロジェクトで提唱（0≦POG≦100%）
- POG値が高い → ランキング結果の頑健性（再現性）が高い方法

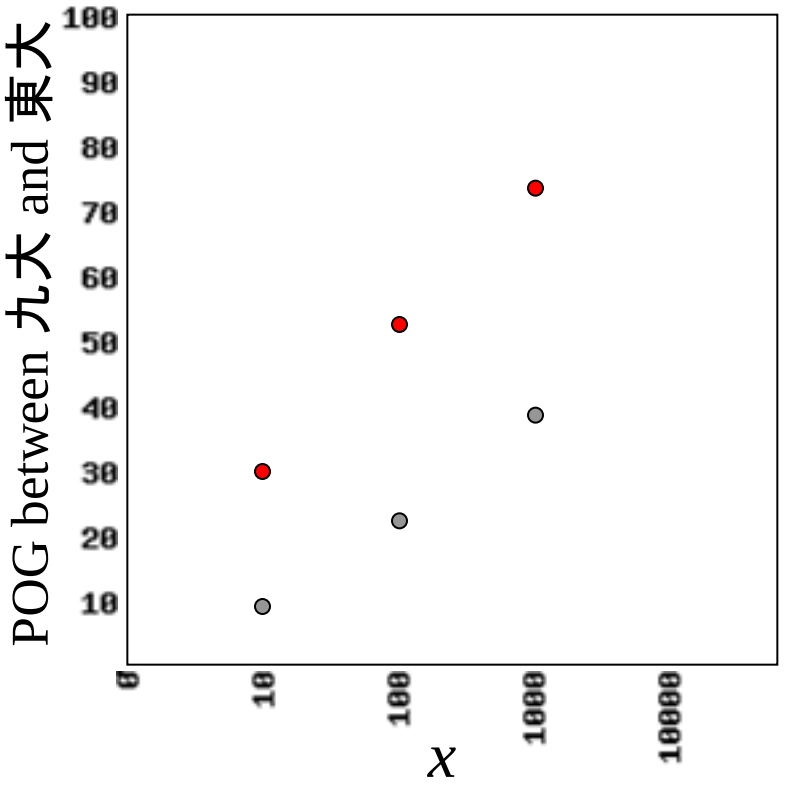


「再現性」を一致度で評価

- 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



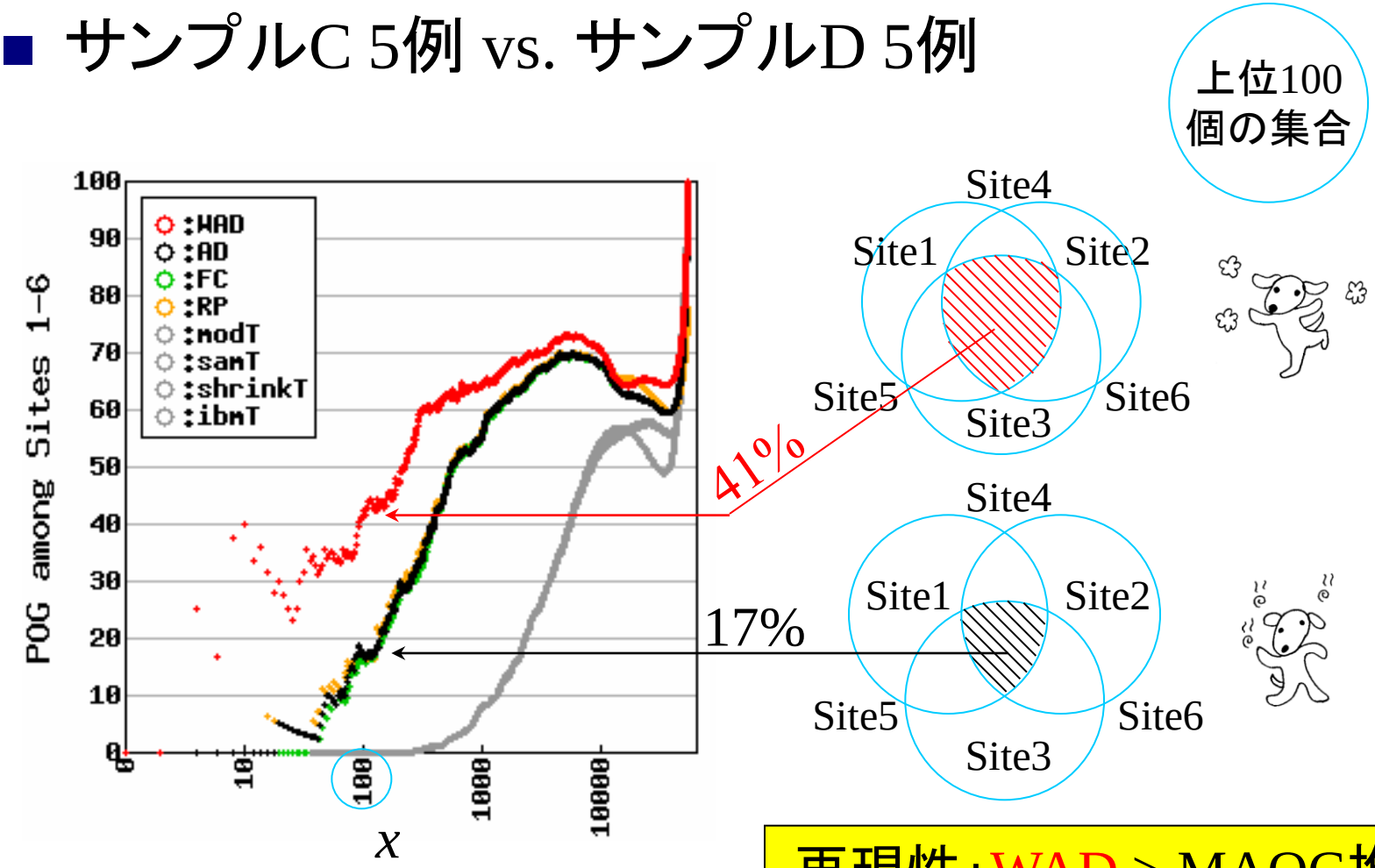




再現性：WAD > samT

「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）

結論（Affymetrixデータ; 二群間比較）

- 「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

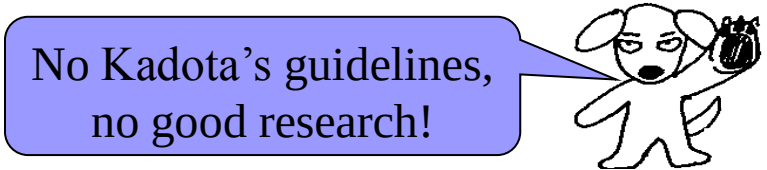
Fold Changeに基づく方法

↕

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

MAQC Consortium, *Nat. Biotechnol.*, **24**:1151-1161, 2006



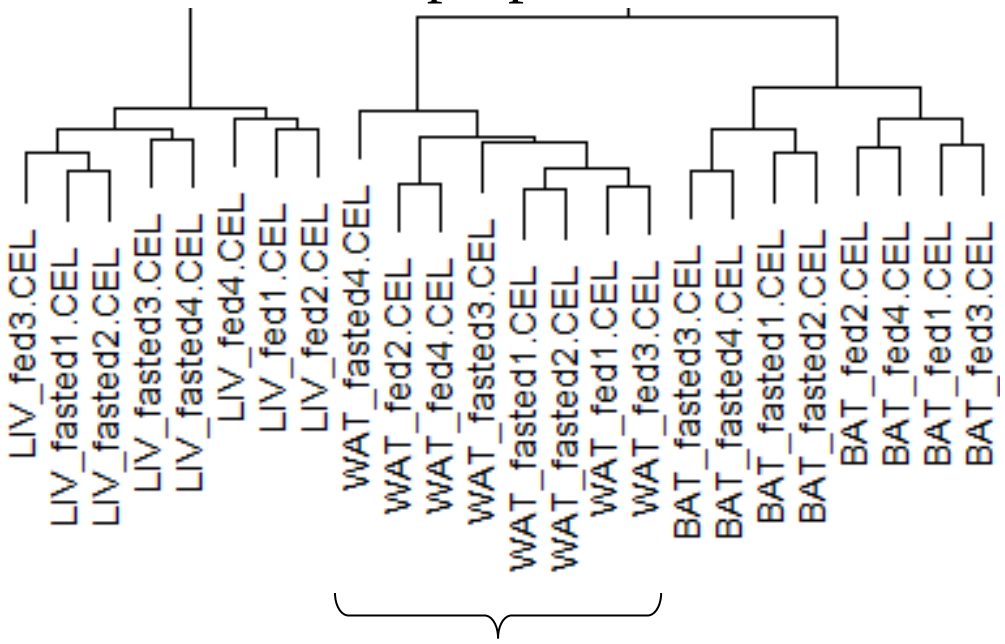
手順3：発現変動遺伝子のランキング

- Affymetrix GeneChipの場合
  - 推奨の組み合わせのものを利用
  - RMAデータの場合はRank productsを利用、など
- その他のメーカーの場合（今のところ根拠なし）
  - チップごとに正規化したデータ
    - WAD
  - 全サンプルのデータをQuantile正規化したようなデータ
    - Rank products

クラスタリング結果を眺めることで…

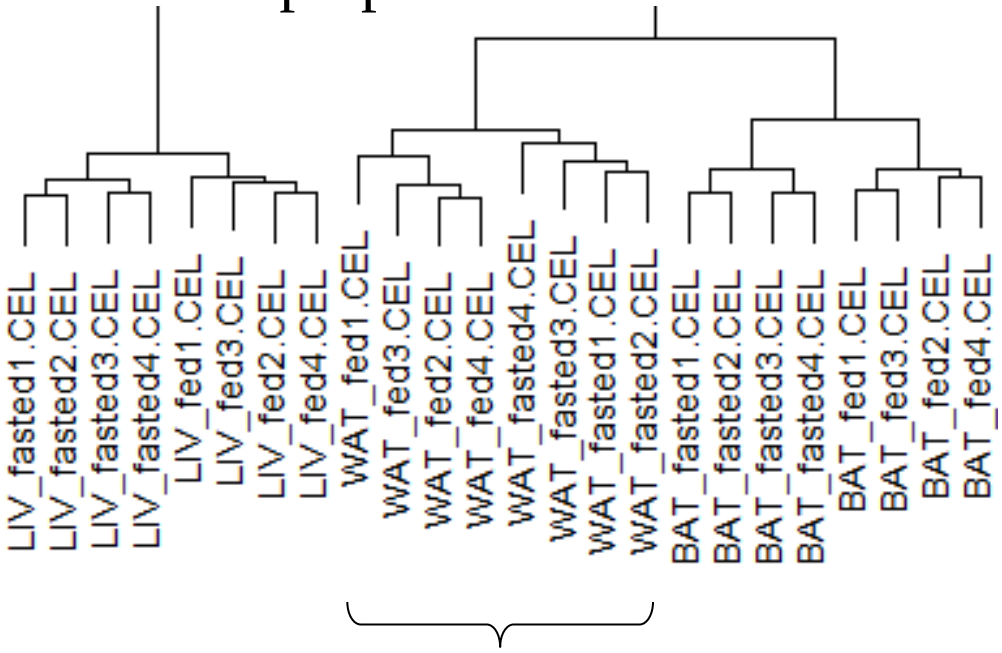
- 本物（真の発現変動遺伝子）があるかどうかの検討がつきます。

MAS-preprocessed data



発現変動遺伝子なさそう…

RMA-preprocessed data

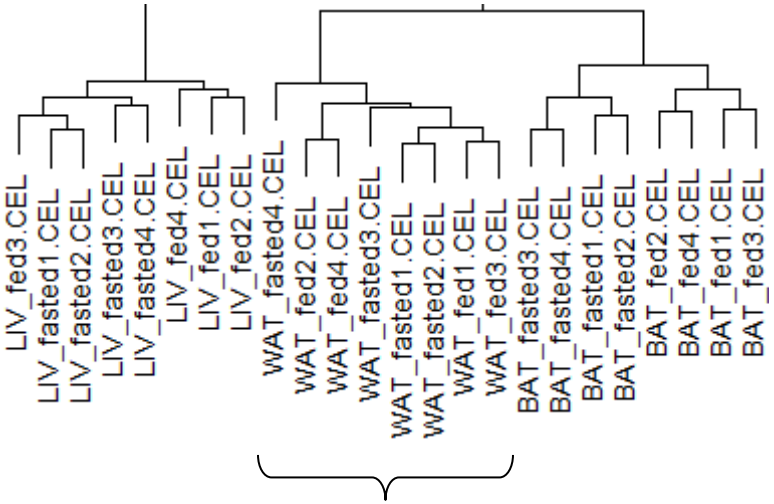


発現変動遺伝子沢山ありそう

RMA-quantified dataなので...

- Rank products法を適用
- WATサンプルの4 fed vs. 4 fasted samplesのデータの解析結果

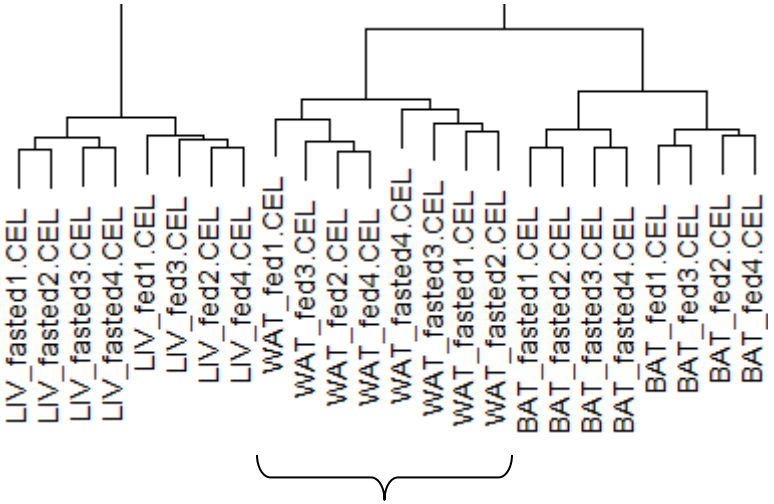
MAS-preprocessed data



FDRの閾値

0.01以下：4個（fasted < fed），45個（fasted > fed）
0.10以下：90個（fasted < fed），198個（fasted > fed)

RMA-preprocessed data



FDRの閾値

0.01以下：359個（fasted < fed），278個（fasted > fed）
0.10以下：970個（fasted < fed），928個（fasted > fed)

二群間比較解析戦略

- 発現変動遺伝子（マーカー遺伝子）の同定
  - 個々の遺伝子について統計量を算出し、ランキング
  - 手法選択のガイドライン（Kadota *et al.*, *AMB*, 2009）
    - 感度・特異度重視の場合
    - 再現性重視の場合
- Gene Set Enrichment Analysis (GSEA)
  - アノテーション情報が豊富な生物種用の解析手段
    - 同じセットに属する遺伝子をひとまとめにして解析
      - 例１：酸化的リン酸化に関係する遺伝子セット（KEGG: hsa00190）
      - 例２：脂肪酸β酸化に関係する遺伝子セット（ GO:0006635）
    - 比較する二群間でその遺伝子セットが動いたかどうかを評価
      - 帰無仮説：動いてない
      - 対立仮説：動いた
    - 沢山の遺伝子セットについて解析を行い、動いた遺伝子セットを列挙
      - positional gene sets
      - pathway gene sets
      - motif gene sets
      - GO gene sets

      etc...

      } 様々な視点での解析が可能

様々な遺伝子セットはMSigDBからゲット

- 例：KEGG Pathway遺伝子セット

	A	B	C	D	E	F	G	H	I	
1	HSA00010_GLYCOLYSIS_AND_GLUCONEOGENESIS	LDHC	LDHB	LDHA	ADH1C	PGAM1	ADH1B	PGAM2	ADH1A	A
2	HSA00020_CITRATE_CYCLE	OGDHL	OGDH	CLYBL	IDH3G	LOC28339	IDH2	IDH1	SUCLA2	F
3	HSA00030_PENTOSE_PHOSPHATE_PATHWAY	ALDOA	TALDO1	ALDOC	ALDOB	PGD	TKTL2	TKTL1	DERA	F
4	HSA00031_INOSITOL_METABOLISM	ALDH6A1	TPI1							
5	HSA00040_PENTOSE_AND_GLUCURONATE_INTERCONVERSION	UGDH	UGT1A7	UGT1A6	UGT1A9	UGT1A8	UGT1A3	UGT1A5	UGT1A4	U
6	HSA00051_FRUCTOSE_AND_MANNOSE_METABOLISM	ALDOA	SORD	PFKFB4	HSD3B7	PFKFB3	ALDOC	PFKFB2	ALDOB	F
7	HSA00052_GALACTOSE_METABOLISM	LALBA	HSD3B7	HK2	HK1	G6PC2	GLB1	GALK2	GALK1	H
8	HSA00053_ASCORBATE_AND_ALDARATE_METABOLISM	ALDH7A1	ALDH1B1	ALDH1A3	MIOX	UGDH	ALDH2	ALDH3A2	ALDH9A1	A
9	HSA00061_FATTY_ACID_BIOSYNTHESIS	OLAH	MCAT	ACACA	FASN	ACACB	OXSM			
10	HSA00062_FATTY_ACID_ELONGATION_IN_MITOCHONDRIA	HSD17B1C	ACAA2	PPT2	ECHS1	PPT1	HSD17B4	HADH	MECR	H
11	HSA00071_FATTY_ACID_METABOLISM	ACOX1	HSD17B1C	ACADSB	CPT2	ADHFE1	EHHADH	ADH5	ADH1C	A
12	HSA00072_SYNTHESIS_AND_DEGRADATION_OF_KETONE_BODI	HMGCS2	OXCT1	HMGCS1	OXCT2	BDH2	ACAT2	BDH1	ACAT1	H
13	HSA00100_BIOSYNTHESIS_OF_STEROIDS	TM7SF2	GGCX	EBP	MVD	CYP51A1	HMGCR	FDPS	LSS	F
14	HSA00120_BILE_ACID_BIOSYNTHESIS	ADHFE1	HSD3B7	ADH5	ADH1C	ADH6	ADH1B	ADH7	ADH1A	A
15	HSA00130_UBIQUINONE_BIOSYNTHESIS	ND1	NDUFB11	ND4	ND5	ND2	ND3	NDUFA13	COQ7	C
16	HSA00140_C21_STEROID_HORMONE_METABOLISM	HSD3B2	CYP17A1	HSD3B1	AKR1C4	CYP11A1	CYP21A2	CYP11B1	CYP11B2	H
17	HSA00150_ANDROGEN_AND_ESTROGEN_METABOLISM	ARSD	ARSE	CYP11B1	CYP11B2	SULT2B1	PRMT3	AKR1C4	PRMT2	F
18	HSA00190_OXIDATIVE_PHOSPHORYLATION	ATP6AP1	NDUFAB1	COX5A	COX5B	ATP8	ATP6	UQCR	COX6C	A
19	HSA00220_UREA_CYCLE_AND_METABOLISM_OF_AMINO_GROUI	SAT1	ALDH18A1	SRM	NAGS	ASS1	SAT2	AGMAT	ASL	A
20	HSA00230_PURINE_METABOLISM	ADCY3	FHIT	ADCY4	GDA	ADCY1	ADCY2	GMPR2	ADCY7	A
21	HSA00232_CAFFEINE_METABOLISM	XDH	CYP2A13	NAT1	NAT2	CYP2A6	CYP2A7	CYP1A2		
22	HSA00240_PYRIMIDINE_METABOLISM	CTPS	DTYMK	CAD	CANT1	PRIM1	NT5M	NT5C3	PRIM2	N
23	HSA00251_GLUTAMATE_METABOLISM	GCLC	GLUD2	GLUD1	GNPNAT1	CAD	QARS	NAGK	GCLM	F

Pathway ID　　Name　　Gene symbols

1行につき1セット

様々なGSEA系の解析手法

- GSEA (Subramanian *et al*., *PNAS*, 2005)
- PAGE (Kim and Volsky, *BMC Bioinformatics*, 2005)
- Hotelling's T^2-test (Kong *et al*., *Bioinformatics*, 2006)
- GSA (Efron and Tibshirani, *Ann. Appl. Stat*., 2007)
- GeneTrail (Backes *et al*., *NAR*, 2007)
- SAM-GS (Dinu *et al*., *BMC Bioinformatics*, 2007)
- GSEA-P (Subramanian *et al*., *Bioinformatics*, 2007)
- GlobalANCOVA (Hummell *et al*., *Bioinformatics*, 2008)
- …

PAGE法

- Parametric Analysis of Gene set Enrichmentの略
  1. 各遺伝子iについて対数変換後のデータのAverage Difference (AD^i)を計算　$AD^i = \overline{A^i} - \overline{B^i}, \qquad (i = 1,2,...,a)$
  2. AD^iの平均μと標準偏差σを計算
  3. 興味ある遺伝子セット（例：i=5,89, 684, 2543, …に相当する計m個の遺伝子）のADの平均S_mを計算

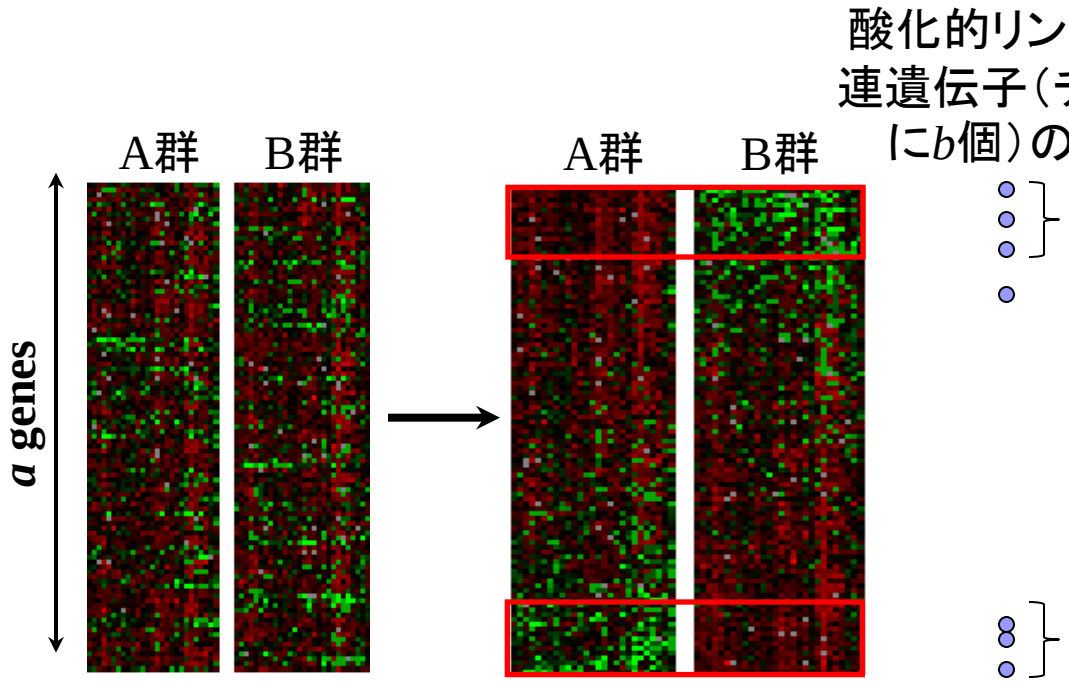
$$S_m = (AD^5 + AD^{89} + AD^{684} + AD^{2543} + ...)/m$$

  4. Zスコアを計算 $Z = (S_m - \mu) \times \sqrt{m} / \sigma$

Zスコアの絶対値が大きい遺伝子セットほど二群間でより発現変動している、と解釈

GSEA以前の解析手段

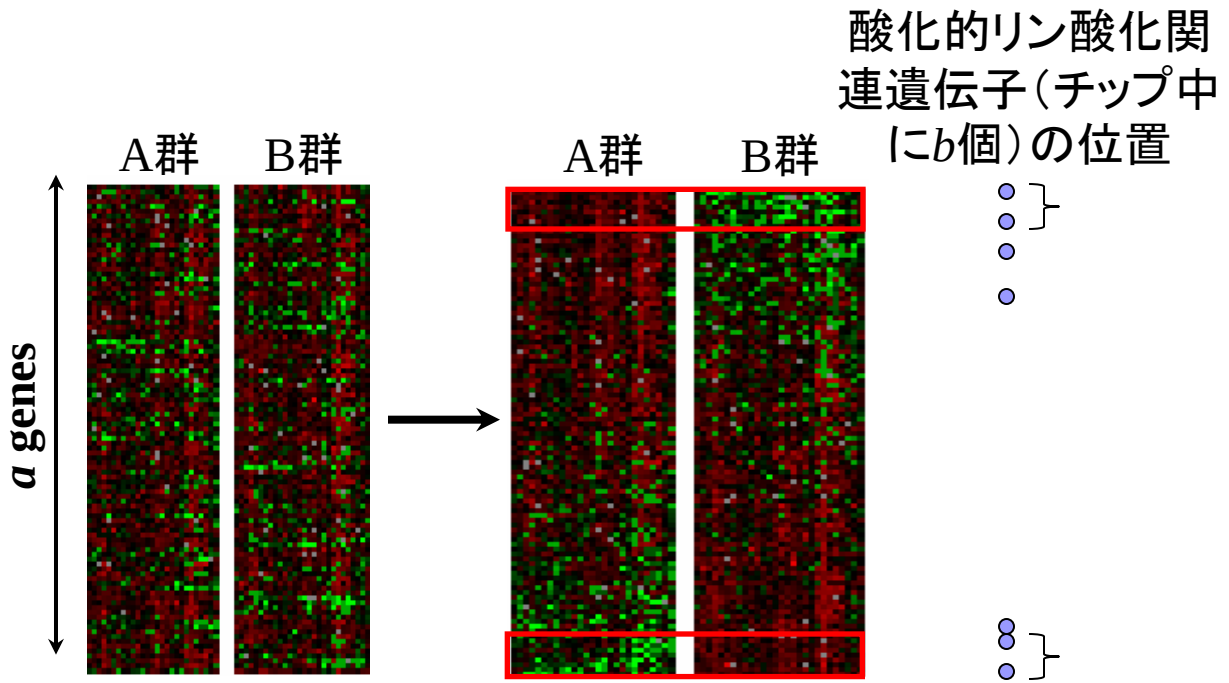
- 例：酸化的リン酸化関連遺伝子セット
  1. Average Differenceのような統計量を各遺伝子について算出
  2. 上位x個を抽出し、**酸化的リン酸化関連遺伝子群**のバックグラウンド（b/a）に対する濃縮度合い（c/x）を評価



帰無仮説：
「チップ中の全遺伝子数(a)に対する酸化的リン酸化関連遺伝子数(b)の割合（b/a）」
と
「酸化的リン酸化関連遺伝子数(b) に対する上位x個の中に占める酸化的リン酸化関連遺伝子数(c) の割合(c/x)」
は等しい

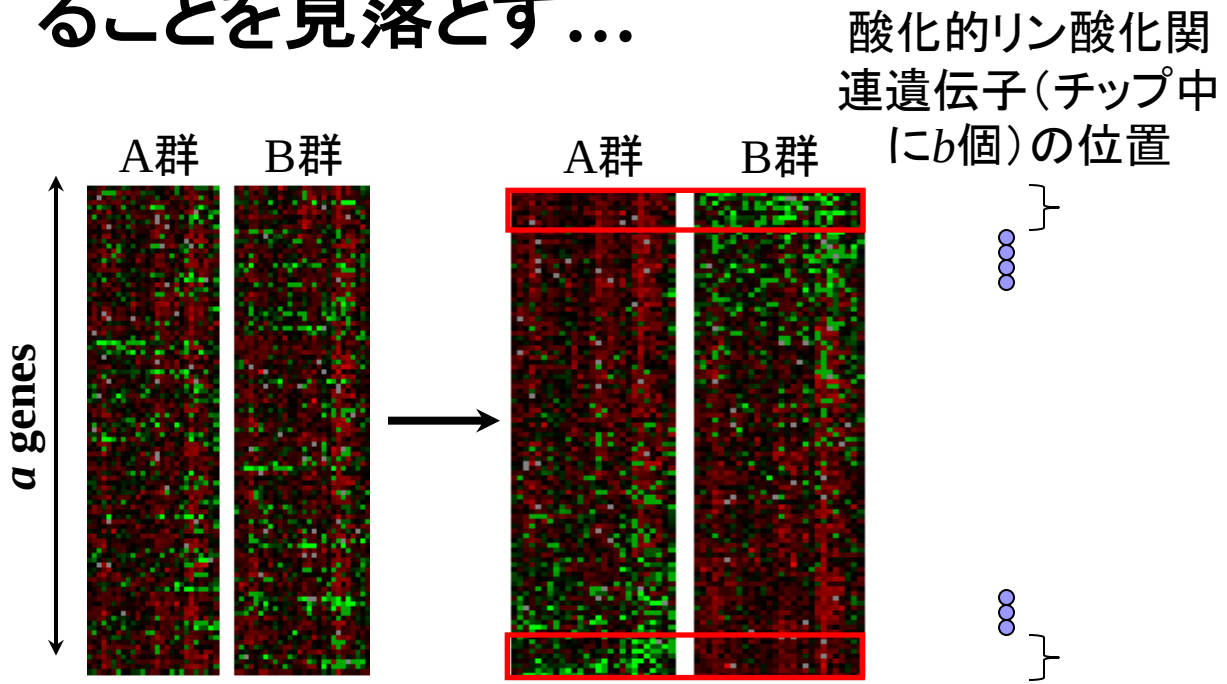
GSEA以前の解析手段の問題点1

- 上位x個のx次第で結果が変わる



GSEA以前の解析手段の問題点2

- 下図のように、全体としては**酸化的リン酸化関連遺伝子セット**が有意差があるといえるような場合でも、上位x個の中に一つも含まれないので有意差があるといえなくなる…。
- 現実の解析では**酸化的リン酸化関連遺伝子セットが動いていることを見落とす…**



様々なGSEA系手法

- なぜ次々と提案されるのか？
  - Ans.1：発現変動遺伝子のランキング法（gene-level statistics）はいくらでもある
    - PAGE：Average Difference (*AD*) ← 倍率変化そのもの
    - GSEA：S2N統計量など
    - その他：Rank products, WAD, SAMなど
  - Ans.2：興味ある遺伝子セットの偏り度合い（濃縮度）を見積もる統計量（gene set statistics）はいくらでもある
    - PAGE：Z検定
    - GSEA：Enrichment Score
    - その他：平均%順位, AUC, medianなど
  - Ans.3：有意性を評価する手段もいくつか考えられる
    - sample label permutation
    - gene resampling

極論：論文になっていない組合せを“新規手法だ！”とすることも可能…

手法選択のガイドラインはない（に等しい）

- どの遺伝子セットが動いている・いないという正解情報（“地上の真実”）を知るすべがない
  - 論文でありがちなプレゼンテーション
    - 既知の遺伝子セットはちゃんと上位にあった。我々はさらに他に動いている遺伝子セットを見つけた。（感度の高さをアピール）
    - “感度の高さ”という点については正しいのかもしれないが、“特異度”は低いのかも…。（本当は動いていない遺伝子セットまで動いていると判断してしまうこと）
  - シミュレーションで本当は動いていないデータセットを作成することはできるが、その結果と現実の結果には相当のギャップがある

GSEA系手法を使えるのはごく一部の生物種

- アノテーション情報が豊富な生物種はGene Ontologyやパスウェイの情報が豊富
  →多くの遺伝子セットを用意できる→GSEA系手法を適用可能
- それ以外の生物種は、まずは様々な発現変動遺伝子をひたすら同定しまくるなどして地道にアノテーション情報を増やしていく以外にない（のではないだろうか）

手順4-1：GSEAを実行

- 重複したGene symbol名のものをまとめたファイルを作成

data_rma.txt

	A	B	C	D	E	F
1		BAT_fastec	BAT_fastec	BAT_fastec	BAT_fastec	BAT_fed1 .C
2	1367452_at	10.19567	10.19738	9.745476	10.07931	10.52457
3	1367453_at	9.656576	9.671801	9.865965	9.908294	9.664242
4	1367454_at	9.79646	9.624494	9.727593	9.782829	9.653699
5	1367455_at	10.54806	10.49293	10.78856	10.75429	10.76457
6	1367456_at	11.32595	11.37424	11.3108	11.47578	11.71273
7	1367457_at	8.764598	8.656652	8.388112	8.458402	8.962444
8	1367458_at	7.518224	7.399644	7.851105	7.981798	8.281839
9	1367459_at	11.68005	11.63014	11.55972	11.59776	11.8071
10	1367460_at	11.53232	11.60598	11.62443	11.64675	11.63179
11	1367461_at	9.069431	9.173958	9.01567	9.100487	9.375125
12	1367462_at	11.71362	11.67918	11.49388	11.75273	11.93639
13	1367463_at	12.11401	11.99826	11.77689	11.9625	12.38199
14	1367464_at	9.363433	9.382455	9.332388	9.53715	9.477257
15	1367465_at	10.17004	10.09226	9.667496	9.803292	10.37527
16	1367466_at	9.888797	9.843451	9.841109	9.882572	10.19688
17	1367467_at	11.88144	11.94542	11.51734	11.83313	12.01797
18	1367468_at	8.983567	8.890969	8.753483	9.049001	9.048987

31,099 行（data_rma.txt）

data_rma_nr.txt

	A	B	C	D	E	F
1		BAT_fastec	BAT_fastec	BAT_fastec	BAT_fastec	BAT_fed1 .C
2	Sumo2	10.19567	10.19738	9.745476	10.07931	10.52457
3	Cdc37	9.656576	9.671801	9.865965	9.908294	9.664242
4	Copb2	9.79646	9.624494	9.727593	9.782829	9.653699
5	Vcp	10.54806	10.49293	10.78856	10.75429	10.76457
6	Ube2d3	10.46601	10.41727	10.50646	10.60873	10.73912
7	Becn1	8.764598	8.656652	8.388112	8.458402	8.962444
8	Lypla2	7.518224	7.399644	7.851105	7.981798	8.281839
9	Arf1	11.16933	11.15347	11.07006	11.11323	11.32332
10	Gdi2	10.82737	10.85763	11.07692	10.99968	10.92015
11	Copb1	9.069431	9.173958	9.01567	9.100487	9.375125
12	Capns1	11.71362	11.67918	11.49388	11.75273	11.93639
13	Phb2	8.784034	8.334072	8.667951	8.693247	8.894173
14	Siahbp1	9.363433	9.382455	9.332388	9.53715	9.477257
15	Dad1	10.17004	10.09226	9.667496	9.803292	10.37527
16	Prpf8	9.888797	9.843451	9.841109	9.882572	10.19688
17	Iscu	11.88144	11.94542	11.51734	11.83313	12.01797
18	Scand1	8.983567	8.890969	8.753483	9.049001	9.048987

14,140 行（data_rma_nr.txt）

理由1：変なバイアスを除きたいから

理由2：遺伝子セットがGene symbolで与えられているから

手順4-1：GSEAを実行

- 必要なファイルをMSigDBからダウンロード

c1: positional gene sets	C1 gene sets file	c1.all.v2.5.symbols.gmt
c2: curated gene sets	C2 gene sets file	c2.all.v2.5.symbols.gmt
	canonical pathway gene sets gene sets file	c2.cp.v2.5.symbols.gmt
	chemical and genetic perturbations gene sets file	c2.cgp.v2.5.symbols.gmt
	BioCarta gene sets file	c2.biocarta.v2.5.symbols.gmt
	GenMAPP gene sets file	c2.genmapp.v2.5.symbols.gmt
	KEGG gene sets file	c2.kegg.v2.5.symbols.gmt
c3: motif gene sets	C3 gene sets file	c3.all.v2.5.symbols.gmt
	transcription factor targets gene sets file	c3.tft.v2.5.symbols.gmt
	microRNA targets gene sets file	c3.mir.v2.5.symbols.gmt
c4: computational gene sets	C4 gene sets file	c4.all.v2.5.symbols.gmt
	cancer gene neighborhoods gene sets file	c4.cgn.v2.5.symbols.gmt
	cancer modules gene sets file	c4.cm.v2.5.symbols.gmt
c5: gene ontology gene sets	C5 gene sets file	c5.all.v2.5.symbols.gmt
	GO biological process gene sets file	c5.bp.v2.5.symbols.gmt
	GO cellular component gene sets file	c5.cc.v2.5.symbols.gmt
	GO molecular function gene sets file	c5.mf.v2.5.symbols.gmt

GSEA実行例（手順4-2）

- LIVサンプルの4 fed vs. 4 fasted samplesデータの Gene Ontology（Biological Process）解析結果
  - 上位10遺伝子セット

絶対値の大きいものほど偏り度合いが高いことを表す。符号はその方向。
正の値 → A群 > B群

*p*値

data_rma_nr_LIV_BP_sorted.txt

	A	B	C	D	E	F
1	Geneset_name	GO_ID	Member_num	Member_num_thischip	z_page	p_page
2	FATTY_ACID_METABOLIC_PROCESS	GO:0006631	63	52	7.98	1.55E-15
3	FATTY_ACID_BETA_OXIDATION	GO:0006635	11	11	6.15	7.71E-10
4	CELLULAR_RESPONSE_TO_NUTRIENT_LEVELS	GO:0031669	10	10	-6.12	9.31E-10
5	CELLULAR_RESPONSE_TO_STRESS	GO:0033554	10	10	-5.87	4.31E-09
6	CELLULAR_RESPONSE_TO_EXTRACELLULAR_STIMULUS	GO:0031668	12	12	-5.57	2.61E-08
7	FATTY_ACID_OXIDATION	GO:0019395	18	18	5.12	3.07E-07
8	MONOCARBOXYLIC_ACID_METABOLIC_PROCESS	GO:0032787	88	75	5.03	4.84E-07
9	CELLULAR_LIPID_METABOLIC_PROCESS	GO:0044255	255	197	4.88	1.08E-06
10	BIOSYNTHETIC_PROCESS	GO:0009058	470	376	-4.81	1.54E-06
11	CELLULAR_BIOSYNTHETIC_PROCESS	GO:0044249	321	263	-4.50	6.92E-06

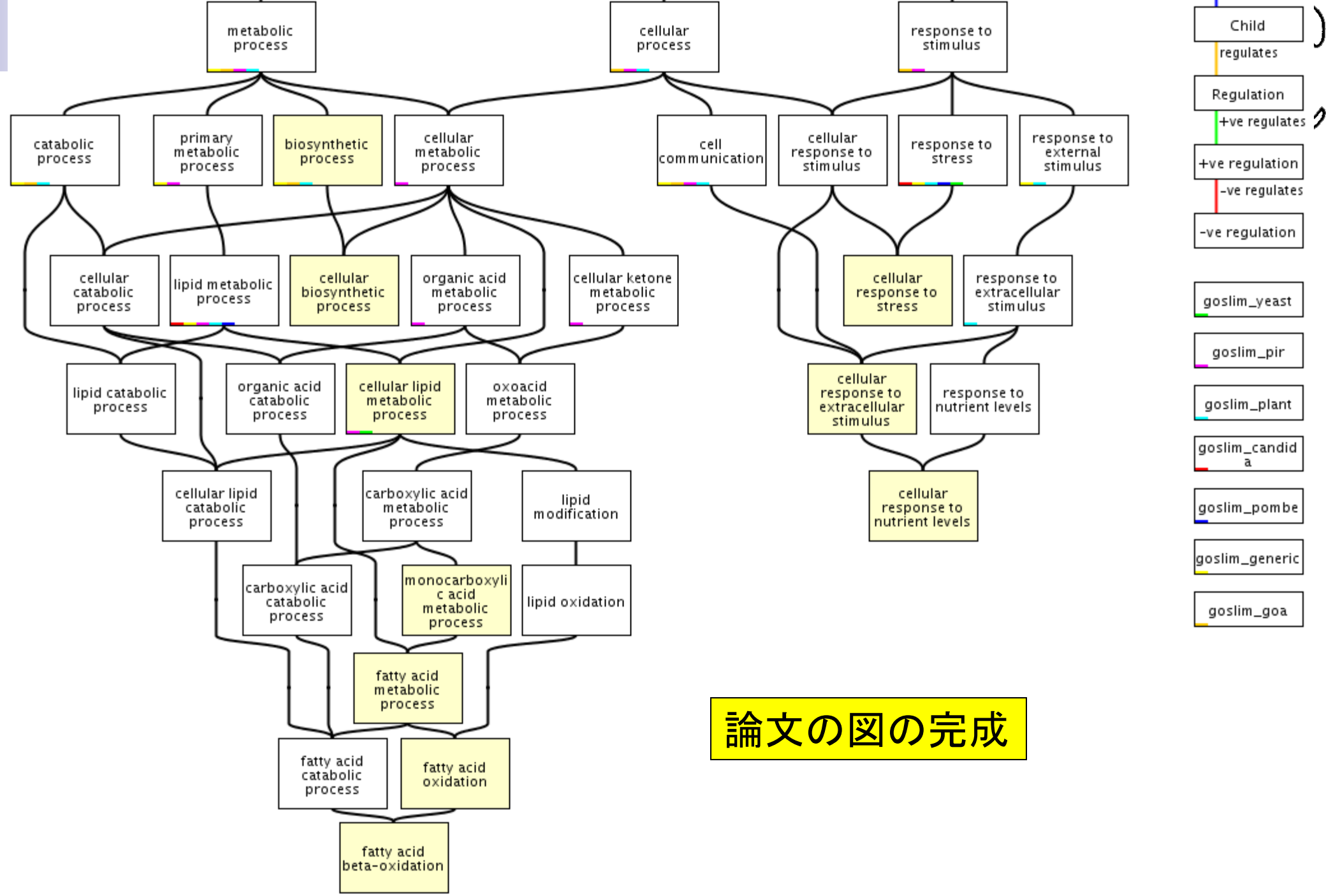
論文の表の完成

このGO IDに含まれる遺伝子セットのメンバー数

63個中52個が自分が用いたアレイ中に搭載されている

手順4-3：GOの階層構造にマップ

- LIVサンプルの4 fed vs. 4 fasted samplesデータのGene Ontology（Biological Process）解析結果
  - 上位*x*遺伝子セット（例：*x* = 10）のGO IDsをQuickGOにかける


metabolic process
cellular process
response to stimulus
catabolic process
primary metabolic process
biosynthetic process
cellular metabolic process
cell communication
cellular response to stimulus
response to stress
response to external stimulus
cellular catabolic process
lipid metabolic process
cellular biosynthetic process
organic acid metabolic process
cellular ketone metabolic process
cellular response to stress
response to extracellular stimulus
lipid catabolic process
organic acid catabolic process
cellular lipid metabolic process
oxoacid metabolic process
cellular response to extracellular stimulus
response to nutrient levels
cellular lipid catabolic process
carboxylic acid metabolic process
lipid modification
cellular response to nutrient levels
carboxylic acid catabolic process
monocarboxylic acid metabolic process
lipid oxidation
fatty acid metabolic process
fatty acid catabolic process
fatty acid oxidation
fatty acid beta-oxidation
Child
regulates
Regulation
+ve regulates
+ve regulation
-ve regulates
-ve regulation
goslim_yeast
goslim_pir
goslim_plant
goslim_candida
goslim_pombe
goslim_generic
goslim_goa

論文の図の完成

GSEA実行例（手順4-2’）

- LIVサンプルの4 fed vs. 4 fasted samplesデータの KEGG Pathway解析結果
  - 上位10遺伝子セット

data_rma_nr_LIV_KEGG_sorted.txt

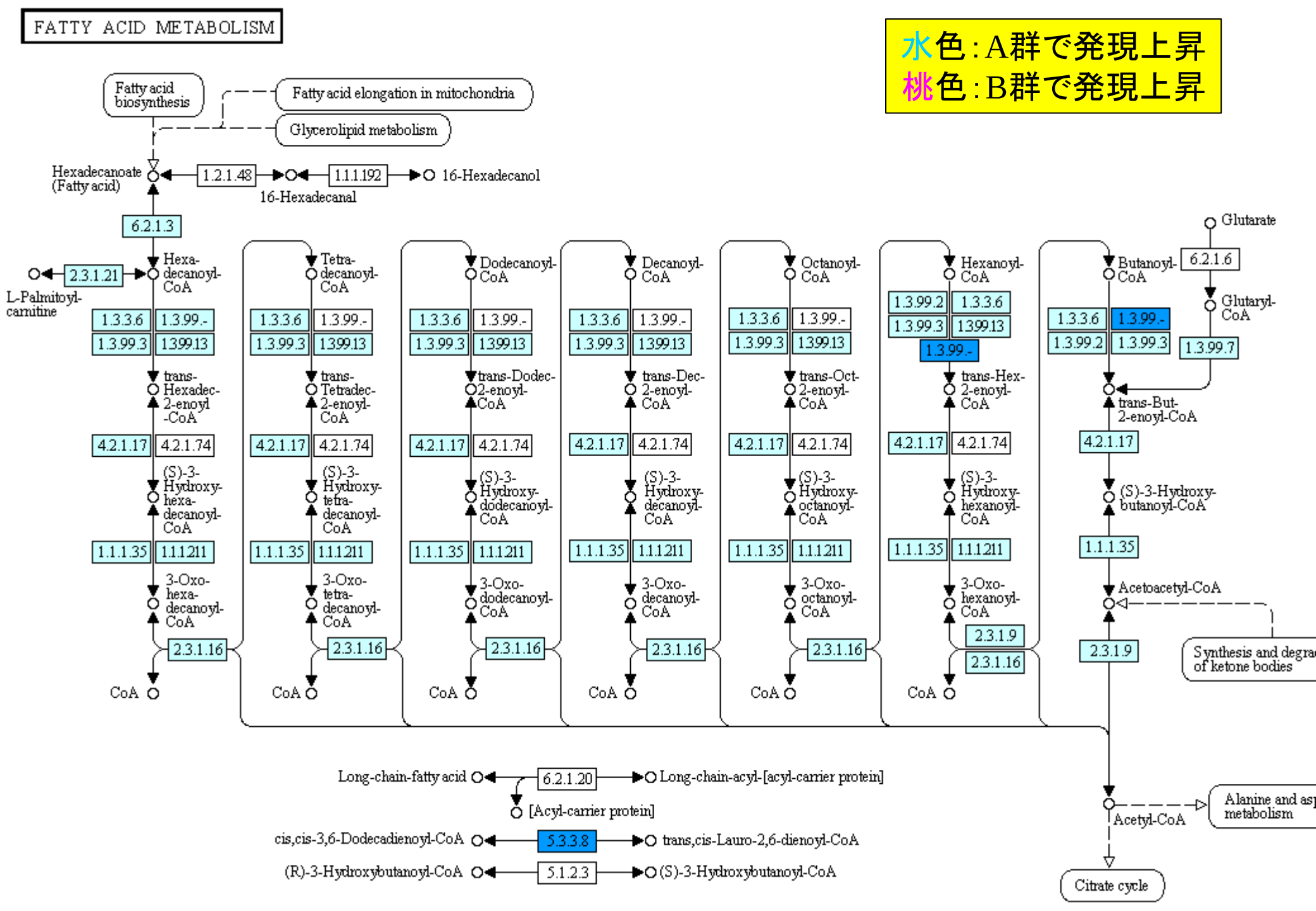
	A	B	C	D	E
1	Geneset_name	Member_num	Member_num_thischip	z_page	p_page
2	HSA00100_BIOSYNTHESIS_OF_STEROIDS	24	22	-9.86	0
3	HSA00061_FATTY_ACID_BIOSYNTHESIS	6	6	-5.43	5.62E-08
4	HSA00071_FATTY_ACID_METABOLISM	47	39	5.09	3.51E-07
5	HSA00900_TERPENOID_BIOSYNTHESIS	6	5	-4.89	9.99E-07
6	HSA01040_POLYUNSATURATED_FATTY_ACID_BIOSYNTHESIS	14	13	-4.84	1.28E-06
7	HSA04920_ADIPOCYTOKINE_SIGNALING_PATHWAY	72	67	4.62	3.90E-06
8	HSA03320_PPAR_SIGNALING_PATHWAY	70	61	4.51	6.43E-06
9	HSA00521_STREPTOMYCIN_BIOSYNTHESIS	10	8	-4.12	3.71E-05
10	HSA00710_CARBON_FIXATION	23	20	-3.90	9.45E-05
11	HSA03010_RIBOSOME	98	56	-3.80	0.000145

論文の表の完成

手順4-3’：パスウェイ上にマップ

- LIVサンプルの4 fed vs. 4 fasted samplesデータのKEGG Pathway解析結果から、第3位の”HSA00071”を構成する遺伝子メンバーの二群間（fed vs. fasted）での変動の程度を4諧調色で表示
  - logratio <= -1を水色（A群で発現上昇）
  - -1 < logratio < 0を薄水色（A群で発現上昇）
  - 0 < logratio < 1を薄ピンク色（B群で発現上昇）
  - logratio >= 1をピンク色（B群で発現上昇）

logratio = mean(B) – mean(A)


FATTY ACID METABOLISM
水色：A群で発現上昇
桃色：B群で発現上昇
Fatty acid biosynthesis
Fatty acid elongation in mitochondria
Glycerolipid metabolism
Hexadecanoate (Fatty acid)
1.2.1.48
16-Hexadecanal
1.1.1.192
16-Hexadecanol
6.2.1.3
2.3.1.21
L-Palmitoyl-carnitine
Hexa-decanoyl-CoA
Tetra-decanoyl-CoA
Dodecanoyl-CoA
Decanoyl-CoA
Octanoyl-CoA
Hexanoyl-CoA
Butanoyl-CoA
Glutarate
6.2.1.6
Glutaryl-CoA
1.3.99.7
1.3.3.6
1.3.99.-
1.3.99.3
1.3.99.2
1399.13
trans-Hexadec-2-enoyl-CoA
trans-Tetradec-2-enoyl-CoA
trans-Dodec-2-enoyl-CoA
trans-Dec-2-enoyl-CoA
trans-Oct-2-enoyl-CoA
trans-Hex-2-enoyl-CoA
trans-But-2-enoyl-CoA
4.2.1.17
4.2.1.74
(S)-3-Hydroxy-hexa-decanoyl-CoA
(S)-3-Hydroxy-tetra-decanoyl-CoA
(S)-3-Hydroxy-dodecanoyl-CoA
(S)-3-Hydroxy-decanoyl-CoA
(S)-3-Hydroxy-octanoyl-CoA
(S)-3-Hydroxy-hexanoyl-CoA
(S)-3-Hydroxy-butanoyl-CoA
1.1.1.35
1.1.1211
3-Oxo-hexa-decanoyl-CoA
3-Oxo-tetra-decanoyl-CoA
3-Oxo-dodecanoyl-CoA
3-Oxo-decanoyl-CoA
3-Oxo-octanoyl-CoA
3-Oxo-hexanoyl-CoA
Acetoacetyl-CoA
2.3.1.16
2.3.1.9
CoA
Synthesis and degra of ketone bodies
Long-chain-fatty acid
6.2.1.20
Long-chain-acyl-[acyl-carrier protein]
[Acyl-carrier protein]
cis,cis-3,6-Dodecadienoyl-CoA
5.3.3.8
trans,cis-Lauro-2,6-dienoyl-CoA
(R)-3-Hydroxybutanoyl-CoA
5.1.2.3
(S)-3-Hydroxybutanoyl-CoA
Acetyl-CoA
Alanine and asp metabolism
Citrate cycle

問題点

- EC番号とGene symbolが1対1対応ではない...
  - 例）"HSA00071"を構成する39 gene symbolsのうち、EC:2.3.1.21に対応するのは4つある...
  - 現状では最終的に反映されている色は、同一EC番号の一番最後に出てきたgene symbol (*i.e.*, CPT2)の発現レベル

1	DCI	−0.38	
2	HSD17B10	−0.26	
3	HSD17B4	−0.56	
4	ACOX1	−0.60	
5	HADHB	−0.64	
6	ACADM	−0.38	
7	ACADL	−0.26	
8	CPT1B	0.09	EC:2.3.1.21
9	ACAT1	−0.51	
10	ACADS	−0.57	
11	ECHS1	−0.20	
12	CPT1A	−3.51	EC:2.3.1.21
13	ACADVL	−0.33	
14	ALDH2	−0.20	
15	ALDH3A1	0.00	
16	ACSL3	1.20	
17	ACSL6	−0.14	
18	EHHADH	−1.03	
19	ALDH3A2	−0.15	
20	ADH7	−0.55	
21	ACADSB	−1.55	
22	ACOX3	−0.80	
23	ADH4	−0.94	
24	HADHA	−0.62	
25	HADH	−0.05	
26	ACSL1	−0.08	
27	ALDH7A1	−0.08	
28	ACAT2	1.75	
29	CPT1C	0.11	EC:2.3.1.21
30	ACAA2	−0.36	
31	GCDH	−0.40	
32	ALDH1A3	−0.56	
33	ALDH1B1	1.66	
34	ADHFE1	−0.58	
35	ACSL5	0.48	
36	CPT2	−0.54	EC:2.3.1.21
37	ACSL4	−0.11	
38	ALDH9A1	0.08	
39	PECI	−0.54	

アグリバイオインフォマティクス教育研究プログラムのフォーラム活動について

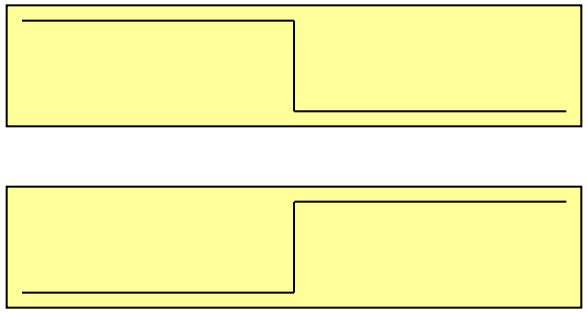
本プログラムでは、研究課題ごとにフォーラムを形成し、セミナー、シンポジウムの開催から、企業との共同研究、学位論文の指導などを行い、当該課題の研究・教育の活性化を図ります。フォーラムのメンバーは、本研究科の教員のほか、他大学、企業、試験研究機関の方々から構成されます。これらのメンバーから、「農学生命情報科学実習II」の受講を通して**学位論文の研究におけるバイオインフォマティクスに関係した研究の指導を受けることができます**。バイオインフォマティクスを利用した農学生命科学の研究、あるいは、バイオインフォマティクスそのものの研究を行って学位を取得した人には、「修了認定証」を発行します。修了の認定は、各専攻の学位審査とは別にフォーラムのメンバーが審査会を開いて行います。研究指導は、研究室の指導教員との合意に基づいて行いますので、**希望する人は、指導教員と相談の上、アグリバイオインフォマティクス教育研究プログラム事務局までご連絡下さい**。現在のところ、以下の4つのフォーラムが形成されています：

- **微生物インフォマティクス・フォーラム**
- **基盤バイオインフォマティクス・フォーラム**
- **アグリ／バイオ・センシングと空間情報学フォーラム**
- **食品インフォマティクス・フォーラム**

遺伝子発現行列

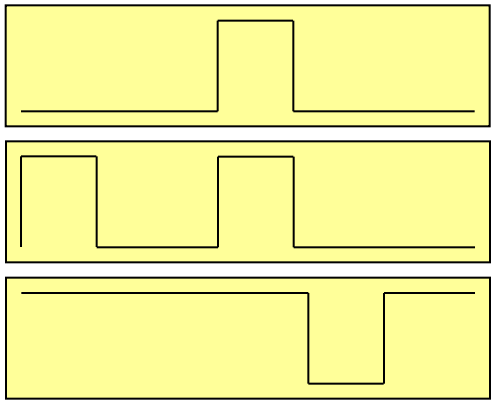
二群間比較

	A群 A1	A群 A2	…	B群 B1	B群 B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	



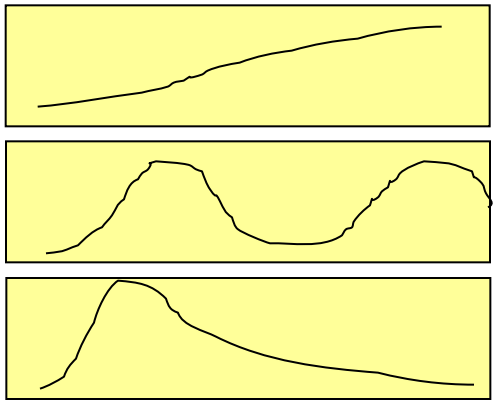
様々な組織（条件）

	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*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



時系列データ

	T1	T2	T3	T4	…
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*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



組織特異的遺伝子検出法

- ランキングに基づく方法
  - Dixon test（Greller and Tobin, *Genome Res.*, **9**, 282-296, 1999）
  - Pattern matching（Pavlidis and Noble, *Genome Biol.*, **2**, research0042, 2001）
  - Entropy（Schug *et al.*, *Genome Biol.*, **6**, R33, 2005）
  - Tissue specificity Index（Yanai *et al.*, *Bioinformatics*, **21**, 650-659, 2005）
- 外れ値検出に基づく方法
  - Akaike's Information Criterion (AIC)（Kadota *et al.*, *Physiol. Genomics*, **12**, 251-259, 2003）
  - Sprent's non-parametric method（Ge *et al.*, *Genomics*, **86**, 127-141, 2005）
- その他
  - Tukey-Kramer's Honest Significance Difference (HSD) test（Liang *et al.*, *Physiol. Genomics*, **26**, 158-162, 2006）
  - ROKU (Kadota *et al.*, *BMC Bioinformatics*, **7**, 294, 2006)

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*)	○	○	○	○	○

結論：おすすめはROKU

組織特異的遺伝子

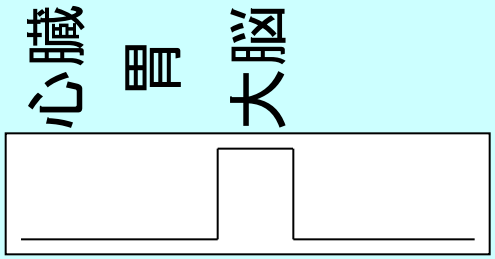
- やりたいこと1

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)	○	○	○	○	○

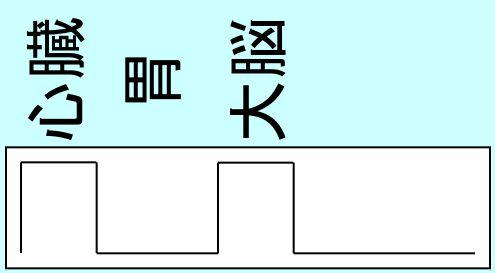
	心臓	肝	大脳	
gene 1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$ …
gene 2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$ …
…	…	…	…	… …
gene i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$
…	…	…	…	… …
gene p	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$ …

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

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

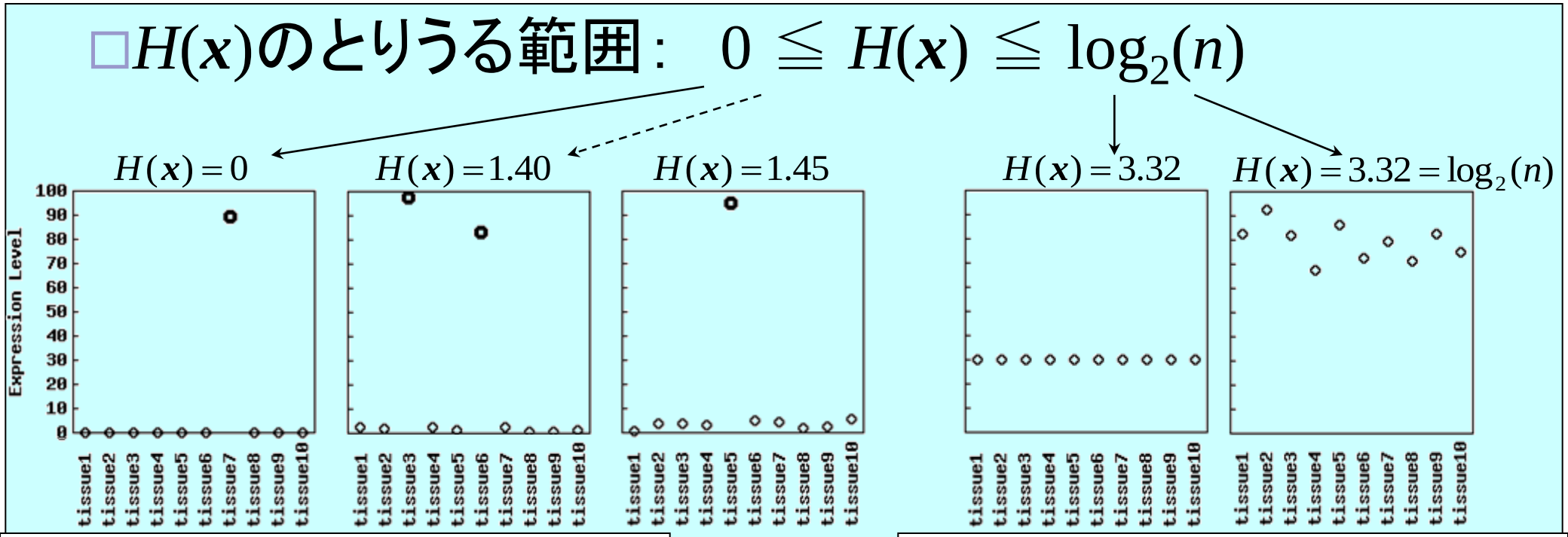
組織特異的遺伝子検出法

②エントロピーによるランキング

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

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

- $H(\boldsymbol{x})$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

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

②エントロピー計算例

- 遺伝子iのエントロピー $H(x_i)$

$$H(x_i) = -\sum_{j=1}^{N} p_{ij} \log_2(p_{ij}) \qquad p_{ij} = x_{ij} / \sum_{j=1}^{N} x_{ij}$$

$$0 \leqq H \leqq \log_2 N$$

gene	Tissue_1	…	Tissue_j	…	Tissue_N	$H(x)$
gene_1	x_{11}		x_{1j}		x_{1N}	$H(x_1)$
…						
gene_i	x_{i1}		x_{ij}		x_{iN}	$H(x_i)$
…						
gene_m	x_{m1}		x_{mj}		x_{mN}	$H(x_m)$

$0 \leqq H \leqq 2.32$

	x_i	p_i	$-p_{ij}\log_2(p_{ij})$
組織1→	$x_{i1} = 0.1$	$p_{i1} = 0.01$	$-p_{i1}\log_2(p_{i1}) = 0.06$
組織2→	$x_{i2} = 10.1$	$p_{i2} = 0.96$	$-p_{i2}\log_2(p_{i2}) = 0.05$
組織3→	$x_{i3} = 0.1$	$p_{i3} = 0.01$	$-p_{i3}\log_2(p_{i3}) = 0.06$
組織4→	$x_{i4} = 0.1$	$p_{i4} = 0.01$	$-p_{i4}\log_2(p_{i4}) = 0.06$
組織5→	$x_{i5} = 0.1$	$p_{i5} = 0.01$	$-p_{i5}\log_2(p_{i5}) = 0.06$
	sum = 10.5	sum = 1	$H(x_i) = 0.31$

特異的発現パターン

→低いエントロピー

	x_i	p_i	$-p_{ij}\log_2(p_{ij})$
組織1→	x_{i1} 40.1	$p_{i1} = 0.17$	$-p_{i1}\log_2(p_{i1}) = 0.44$
組織2→	x_{i2} 35.2	$p_{i2} = 0.15$	$-p_{i2}\log_2(p_{i2}) = 0.41$
組織3→	x_{i3} 60.8	$p_{i3} = 0.26$	$-p_{i3}\log_2(p_{i3}) = 0.50$
組織4→	x_{i4} 50.4	$p_{i4} = 0.21$	$-p_{i4}\log_2(p_{i4}) = 0.48$
組織5→	x_{i5} 48.7	$p_{i5} = 0.21$	$-p_{i5}\log_2(p_{i5}) = 0.47$
	sum = 235	sum = 1	$H(x_i) = 2.30$

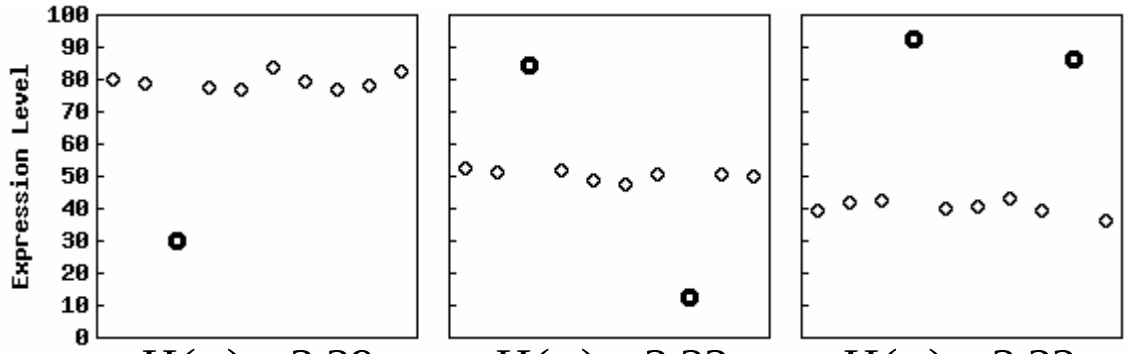
そうでないパターン

→高いエントロピー

組織特異的遺伝子検出法

②エントロピーの短所

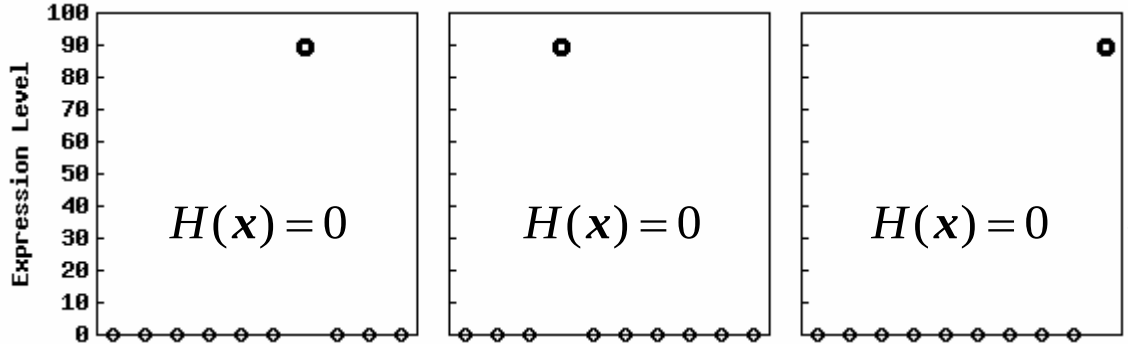
1. 組織特異的低発現パターンなどの検出が不可能



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

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

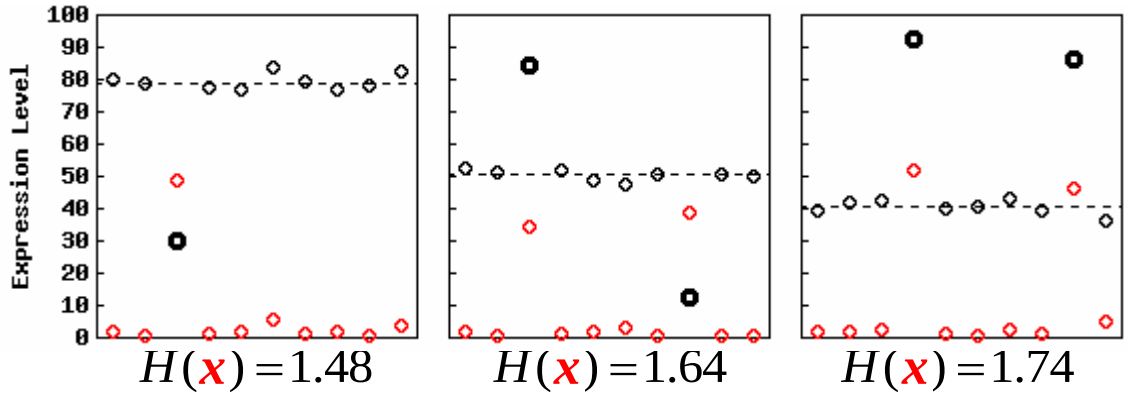


どの組織で特異的なのか分からない

組織特異的遺伝子検出法

- ③ROKU

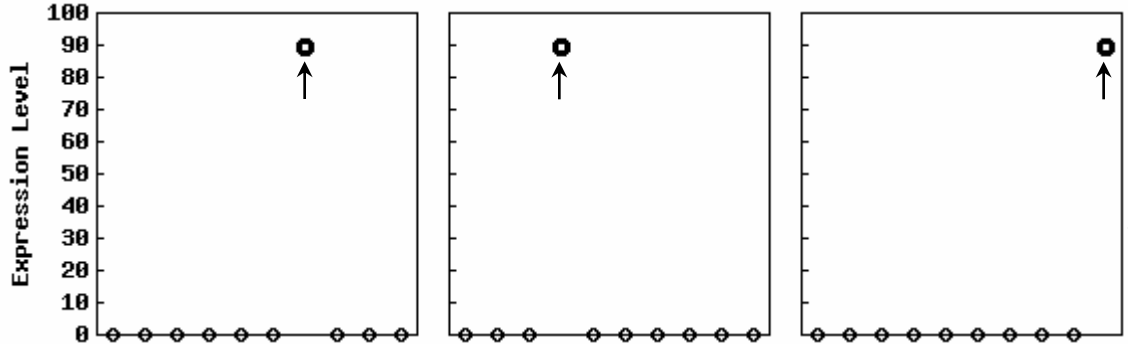
1. 遺伝子発現ベクトルxを変換: $x \rightarrow \boldsymbol{x}$ by $x_i = |x_i - T_{bw}|$



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

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

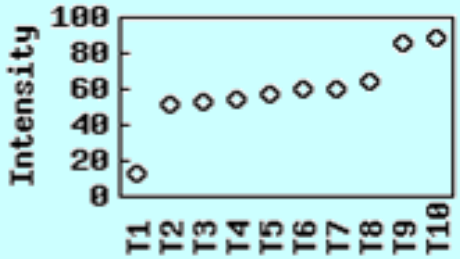
④AICに基づく外れ値検出法

- 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

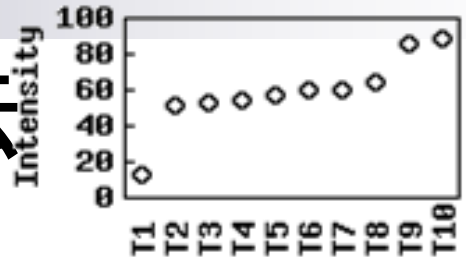
組織特異的遺伝子検出法

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