様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年５月１１日現在

機関番号：１２６０１
研究種目：若手研究（B）
研究期間：2009～2011
課題番号：２１７１０２０８
研究課題名（和文）　マイクロアレイ解析の再現性・感度・特異度を飛躍的に向上させるデータ解析手法の開発
研究課題名（英文）　Methods for detecting differentially expressed genes with reproducibility, sensitivity, and specificity from microarray data
研究代表者
門田　幸二（KADOTA KOJI）
東京大学・大学院農学生命科学研究科・特任准教授
研究者番号：６０３９２２２１

研究成果の概要（和文）：実験によって得られたマイクロアレイデータから比較する二群間で発現変動遺伝子（DEG）を効率的に検出するためには、①正規化と②検出法の組合せが重要であることを明らかにした。また、研究代表者らが開発したDEG検出法であるWAD法を取り入れることが感度・特異度だけでなく再現性の観点からも重要であることを示した。

研究成果の概要（英文）：In order to increase sensitivity, specificity, and reproducibility in microarray analyses, researchers need to select suitable combinations of preprocessing algorithms and gene ranking methods. We recommend the use of WAD for the purpose.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,500,000	450,000	1,950,000
２０１０年度	500,000	150,000	650,000
２０１１年度	300,000	90,000	390,000
年度			
年度			
総　計	2,3000,000	690,000	2,990,000

研究分野：複合新領域
科研費の分科・細目：ゲノム科学・ゲノム情報科学→システムゲノム科学
キーワード：バイオインフォマティクス

１．研究開始当初の背景

DNAマイクロアレイは、遺伝子発現解析の基盤技術であり、これまでに即座の臨床応用を期待させる基礎研究の成果が多数報告されてきた。しかしながらわが国では、これらの成果の臨床応用が遅々として進まず、国民にとって有益な遺伝子発現解析検査が行われる体制になっていないのが現状である。これは主に、マイクロアレイ解析結果の再現性が低いために、全国どこでも同じ検査結果が得られないためである。

国内の遺伝子発現解析検査の実用化に向けた取り組みは、主に経済産業省と厚生労働省により、また産業界が中心となって設立されたバイオチップコンソーシアムによってなされている。しかしながらこれらは、検体の収集や実験プロトコルの標準化に向けた取り組みが中心であり、その後のデータ解析の標準化に向けた取り組みは手付かずなのが現状である。

同一検体の検査結果の再現性を保証することは、国民の望む全国一律の基準による医療の選択を可能にする上での必須事項である。これを踏まえ、国外ではマイクロアレイ解析結果の再現性に関する大規模な評価研究がなされている。米国食品医薬品局（FDA）は、遺伝子発現解析検査の実用化を目指したマイクロアレイ品質管理（MAQC）プロジェ

クトを行い、再現性に関する一定の肯定的な結論およびその結論を導いた推奨するデータ解析手法を提案している。しかしながら、MAQC プロジェクトの結論はごく一部の解析手法を評価した結果に基づくものであり、MAQC 推奨ガイドラインは疑わしいという指摘がなされていた。

２．研究の目的

マイクロアレイを用いた遺伝子発現解析検査（臨床診断）の迅速な実用化を目指し、マイクロアレイ解析の再現性・感度・特異度を飛躍的に向上させるデータ解析手法の開発および推奨ガイドラインの提案を具体的な目的として研究を行った。

３．研究の方法

研究代表者らは本研究の着想に至る背景として、マイクロアレイデータ解析結果の良し悪しは、数多く提案されている①正規化法と②候補遺伝子検出法の組み合わせにより大きく左右されることを指摘する論文を発表した（Kadota et al., AMB, 2008）。さらにこの報告の中で研究代表者は、感度・特異度の点で優れる②候補遺伝子検出法（WAD 法）を開発するとともに、WAD 法が①正規化法の違いに対しても頑健であることを示した。これらのパイロットスタディの結果を踏まえ、研究代表者の開発した発現変動遺伝子（DEG）検出法（WAD 法）や MAQC プロジェクトが最も再現性が高いと主張する AD 法を含む計 8 種類の DEG 検出法（WAD, AD, FC, RP, modT, samT, shrink, and ibmT）について、利用可能な公開データに応じて以下に示す解析を行った。

(1) Affymetrix 社製 GeneChip データを用いた感度・特異度の高い①正規化法と②DEG 検出法の組み合わせ探索

数あるマイクロアレイメーカーの中で最もよく利用されている Affymetrix HG-U133A array を用い、RT-PCR によって候補遺伝子の発現変動を確認した複数の遺伝子を含む 36 個の二群間比較データセットのプローブレベルデータを取得し、9 種類の正規化法（MAS, RMA, DFW, PLIER, VSN, FARMS, mmgMOS, MBEI, and GCRMA）と上記 8 種類の DEG 検出法の全ての組合せについて評価を行った。評価は感度・特異度を同時に評価できる AUC 値（真の DEG を正しく上位にランキングできている組合せは 1、その逆は 0 となるので 1 に近い数値ほど優れている）に基づいて行った。

(2) MAQC 公開データを用いた感度・特異度の高い DEG 検出法の評価

MAQC 公開データは、メーカー推奨の正規化法を適用した後のものであるため、8 種類の DEG 検出法についてのみ AUC 値で評価を行った。また、MAQC は同一の二サンプル（Sample A vs. B と Sample C vs. D）間比較を異なる三つの場所（Site 1, 2, and 3）で取得したデータを提供しているため、例えば Affymetrix (AFX)だけでも 6 種類のデータセット（AB_1, AB_2, AB_3, CD_1, CD_2, CD_3）が存在する。本研究では、AFX 以外にも Applied Biosystems (ABI), Agilent (AG1), GE Healthcare (GEH), Illumina (ILM)の計 5 メーカーの MAQC データについて解析を行った。

(3) MAQC 公開データを用いた再現性の高い DEG 検出法の評価

MAQC プロジェクトでは、二群間比較マイクロアレイデータをいくつかの DEG 検出法を用いて発現変動順にランキングし、異なる三つの場所由来の DEG リストの再現性を Percentage of Overlapping Genes (POGs)という評価基準で評価した。本研究では、同じ MAQC 公開データをもとに同じ評価基準で解析を行い（MAQC のオリジナル研究時には publish されていなかった WAD 法を含む）計 8 つの DEG 検出法の「DEG リストレベルでの再現性評価」を行った。また、「たとえ DEG リストレベルでの再現性が低くても Gene Ontology や KEGG パスウェイといった機能レベルでの再現性の高さが確保されていれば実用上問題ないのではないか？」といった議論が研究者コミュニティでなされていた。そのため、本研究では 186 個の KEGG パスウェイ関連遺伝子セットの「機能レベルでの再現性評価」も行った。

４．研究成果

(1) 感度・特異度の高い正規化法と DEG 検出法の組合せ探索では、正規化法として最もよく用いられている MAS を採用するときには DEG 検出法 WAD を用いる（正規化法/DEG 検出法 : MAS/WAD）と高い AUC 値が得られることがわかった。また、二番目によく用いられている RMA 正規化法を採用するときには Rank products (RP)を DEG 検出法として用いるとよいことが分かった。

表 1 に調査を行った 9 つの正規化法ごとの推奨 DEG 検出法を示す。注意点としては、上記 9 つの中から一番よい組合せを提示することはできない、ということである。なぜなら、評価に用いたデータセットのそれぞれが、なんらかの正規化法と DEG 検出法の組合せを用いて候補遺伝子のランキングを行い、その結果に基づいて RT-PCR で発現変動の確認を行ったものを「真の DEG」としている。このため、評価に用いた 36 個のデータセット中で最も多く用いられた正規化法（や DEG 検出法）が不当に有利にならないようなバイ

アスのかからない評価をする必要がある。表 1 の推奨ガイドラインは、上記のバイアスを考慮に入れて評価を行った結果である。

表 1. 感度・特異度の高い組合せ

正規化法	DEG 検出法
PLIER	RP
VSN	RP
FARMS	RP
mmgMOS	WAD
MBEI	RP
GCRMA	RP
MAS	WAD
RMA	RP
DFW	RP

(2) マイクロアレイメーカーごとの感度・特異度の高い DEG 検出法の評価は、MAQC プロジェクト内で取得された RT-PCR アッセイの結果を用い、さまざまな基準で定義した「真の DEG」をどれだけ上位にランキングできるか（どの DEG 検出法の AUC 値が高いか？）に基づいて行われた。MAQC では二つのアッセイテクノロジー（TaqMan (TAQ)と StaRT-PCR (GEX)）を用いてデータを取得している。また、アッセイ後の生データから DEG を検出する方法も様々な方法が考えられる。本研究では、比較したい 8 つの DEG 検出法に含まれない *t*-test を用いて得られた False Discovery Rate (FDR)と最も一般的な倍率変化に基づく方法（Average Difference (AD)）で定義した「真の DEG」を用いて解析を行った。

表 2 に「真の DEG」を上記四つの基準（TAQ_FDR, TAQ_AD, GEX_FDR, and GEX_AD）で定義して得られた結果（推奨 DEG 検出法）を示す。

表 2. プラットフォームごとの推奨検出法

	TAQ FDR	TAQ AD	GEX FDR	GEX AD
AFX	WAD	RP	WAD	WAD
ABI	ibmT	RP	ibmT	ibmT
AG1	WAD	RP	WAD	ibmT
GEH	WAD	RP	WAD	WAD
ILM	WAD	RP	WAD	WAD

表 1 の結果と同様、MAQC データセットの解析結果においても、Affymetrix GeneChip (AFX)を用いてデータを取得する場合には DEG 検出法として WAD または RP がよい（高い AUC 値を示す）ことがわかる。プラットフォームが ABI のもののみ ibmT 法が推奨されているが、この方法は AFX データの様々な感度・特異度解析結果においても WAD に次ぐパフォーマンスを示していたことからリーズナブルである。全体的には、どのメーカー（プラットフォーム）のものを用いても WAD または RP がよいという結果が得られた。

(3) 再現性の高い DEG 検出法は、三つの実験場所で独立に得られた実験データを、比較する 8 つの DEG 検出法をそれぞれ用いて発現変動順にランキングした結果得られる三つのベクトルに基づいて、上位 X 個中何個が共通しているかという POG 値で評価された。

表 3 に Sample C vs. D の二群間比較によって得られた上位 100 遺伝子の POG 値を示す。

表 3.実験場所間での POG 値

	AFX	ABI	AG1	GEH	ILM
WAD	64	76	54	39	66
AD	23	11	2	5	18
FC	23	0	0	3	19
RP	20	8	2	7	9
modT	20	26	3	1	13
samT	32	31	5	10	21
shrT	21	40	2	1	12
ibmT	29	39	10	1	22

結論は単純で「どのプラットフォームを用いた場合でも WAD 法が抜群の再現性の高さを示す」である。表 3 は上位 100 遺伝子（$X = 100$）の結果のみであるが、どの X の値についても全体的に WAD 法が優れていることを確認済みである。

WAD 法の再現性の観点での優位性は、Gene Ontology 解析や KEGG Pathway 解析といった遺伝子セットに基づく機能解析レベルの評価においても確認済みである。WAD の高再現性（Inter-site reproducibility or Intra-platform reproducibility）は、プラットフォーム非依存であるばかりでなく、評価を行った 5 つのプラットフォーム間（Inter-platform reproducibility）でも高い再現性を保持していることを確認している。

以上の結果から、MAQC 推奨ガイドラインを完全に凌駕する新たなガイドラインの提案を行うことができた。本研究成果に基づく推奨ガイドラインは以下に要約できる：

- 現在でも最もよく用いられている Affymetrix GeneChip を用いて解析を行う場合には、本研究で得られた正規化法と DEG 検出法の最適な組合せを用いるとよい。
- その他のプラットフォームは利用可能な正規化法が限られているために評価が難しいが、DEG 検出法に限っていえば WAD か RP に集約される。再現性まで考慮に入れたトータルでは WAD 法がお勧めである。

５．主な発表論文等

〔雑誌論文〕（計６件全て査読有）

① Kadota K* and Shimizu K., Evaluating methods for ranking differentially expressed genes applied to MicroArray Quality Control data. *BMC Bioinformatics*, **12**:227, 2011. doi:10.1186/1471-2105-12-227

② Ishikawa K, Yoshida S*, Kadota K, Nakamura T, Niiro H, Arakawa S, Yoshida A, Akashi K, Ishibashi T, Gene Expression Profile of Hyperoxic/Hypoxic Retinas in Mouse Model of Oxygen-induced Retinopathy. *Investigative Ophthalmology & Visual Science*, **51**(8):4307-4319, 2010. doi: 10.1167/iovs.09-4605

③ Minamoto T, Hanai S, Kadota K, Oishi K, Matsumae H, Fujie M, Azumi K, Satoh N, Satake M, Ishida N*, Circadian clock in *Ciona intestinalis* revealed by microarray analysis and oxygen consumption. *Journal of Biochemistry*, **147**(2):175-184, 2010. doi: 10.1093/jb/mvp160

④ Oishi K*, Uchida D, Ohkura N, Doi R, Ishida N, Kadota K, Horie S, Ketogenic Diet Disrupts the Circadian Clock and Increases Hypofibrinolytic Risk by Inducing Expression of Plasminogen Activator Inhibitor-1. *Arteriosclerosis Thrombosis and Vascular Biology*, **29**(10):1571-1577, 2009. doi: 10.1161/ATVBAHA.109.190140

⑤ Ishii S*, Kadota K, Senoo K., Application of a clustering-based peak alignment algorithm to analyze various DNA fingerprinting data. *Journal of Microbiological Methods*, **78**(3):344-350, 2009.
doi: 10.1016/j.mimet.2009.07.005

⑥ Kadota K*, Nakai Y, Shimizu K., Ranking differentially expressed genes from Affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity. *Algorithms for Molecular Biology*, **4**:7, 2009. doi:10.1186/1748-7188-4-7

〔学会発表〕（計８件）

① 門田幸二, R でトランスクリプトーム解析, 2011 年度 HPCI チュートリアルセミナー, 2012 年 3 月 9 日, 生命情報工学研究センター(東京)

② 門田幸二, R によるトランスクリプトーム解析～NGS 由来塩基配列データを自在に解析する～, R でつなぐ次世代オミックス情報統合解析研究会, 2012 年 2 月 22 日, 理化学研究所横浜研究所(神奈川)

③ 門田幸二, RNAseq による定量的解析と qPCR、マイクロアレイなどとの比較, 新学術領域研究「複合適応形質進化の遺伝子基盤解明」・複合適応形質進化インフォマティクスオープンセミナー, 2010 年 12 月 28 日, 金沢大学(石川)

④ 門田幸二, トランスクリプトーム解析におけるバイオインフォマティクス要素技術～私の相場観～, 日本バイオインフォマティクス学会・第 2 回アグリバイオインフォマティクス研究会第 1 部, 2010 年 9 月 17 日, 琉球大学(沖縄)

⑤ 門田幸二, 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プログラム・マイクロアレイデータ解析講習会, 2009 年 11 月 20 日 (第 1 回), 2009 年 11 月 24 日 (第 2 回) , 東京大学(東京)

⑥ 門田幸二, トランスクリプトームデータの解析戦略とその周辺, 東京大学大学院農学生命科学研究科第 36 回アグリバイオインフォマティクスセミナー, 2009 年 10 月 21 日, 東京大学(東京)

⑦ 門田幸二, マイクロアレイを用いた遺伝子発現解析, 基礎生物学研究所 バイオインフォマティクス・トレーニングコース 2009, 2009 年 8 月 19-21 日 (第 1 回), 2009 年 9 月 8-10 日 (第 2 回) , 基礎生物学研究所(愛知)

⑧ 門田幸二, 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 ILSI Japan 寄付講座「機能性食品ゲノミクス」公開シンポジウム・食品の機能予測とニュートリゲノミクス, 2009 年 5 月 13 日, 東京大学(東京)

〔図書〕（計２件）

① Ishii S, Kadota K, Senoo K, Clustering-based peak alignment algorithm for objective and quantitative analysis of DNA fingerprinting data. *Handbook of Molecular Microbial Ecology I: Metagenomics and complementary approaches*, Edited by Frans J de Bruijn, Wiley-Blackwell, 67-73, 2011. ISBN: 978-0-470-64479-9

② 門田幸二, 1-4-1. データ解析総論, *バイオチップ実用化ハンドブック*, NTS, 103-114, 2010. ISBN: 978-4-86043-270-6

〔その他〕
研究代表者ホームページ
http://www.iu.a.u-tokyo.ac.jp/~kadota/

６．研究組織
(1)研究代表者
門田　幸二（KADOTA KOJI）
東京大学・大学院農学生命科学研究科・特任准教授
研究者番号：６０３９２２２１