様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　５　月　２６　日現在

機関番号：　１２６０１
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５００３５９

研究課題名（和文）シークエンスに基づく比較トランスクリプトーム解析のためのガイドライン構築

研究課題名（英文）Guidelines for differential expression analysis from RNA-seq data

研究代表者
門田　幸二（Kadota, Koji）

東京大学・農学生命科学研究科・特任准教授

研究者番号：６０３９２２２１

交付決定額（研究期間全体）：（直接経費）　　4,100,000円

研究成果の概要（和文）：RNA-seqは、生体内で発現しているRNA（トランスクリプトーム）を網羅的に調べる次世代の実験技術である。本研究では、旧世代ではあるものの成熟した実験技術であるマイクロアレイ用に開発された手法や知見をもとに、RNA-seqデータ解析精度向上を目指し研究を行った。得られたガイドラインや知見を手軽に利用できるよう、ウェブサイト上で公開するとともに、体系的にまとめた書籍を刊行した。

研究成果の概要（英文）：RNA-seq is a powerful tool for obtaining ribonucleic acids (RNA) sequence data expressed in a target sample. The purpose of this work is to provide guidelines for differential expression analysis from RNA-seq count data. The results were systematically summarized as a book and websites (http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html and http://www.iu.a.u-tokyo.ac.jp/~kadota/r.html).

研究分野：　バイオインフォマティクス

キーワード：　バイオインフォマティクス　トランスクリプトーム　RNA-seq　正規化　発現変動

１．研究開始当初の背景

トランスクリプトームデータ解析における感度・特異度・再現性の高い解析ガイドラインの構築は、基礎研究レベルのデータ解析の質を高めるだけでなく、臨床診断や農作物の品種識別など各種検査の実用化を加速する重要な取り組みである。マイクロアレイについては、これまでに感度・特異度の高い手法は t 検定のような統計的な手法であり、再現性の高い手法は何倍発現が変化したかという倍率変化に基づく手法であるという報告が多くなされてきた。これらの結果を踏まえ、米国食品医薬品局（FDA）主導で行われたマイクロアレイ品質管理（MAQC）プロジェクトは、これらの二つの手法を組み合わせたものをマイクロアレイ解析の効率的な発現変動遺伝子（DEG）検出のための推奨ガイドラインとして 2006 年に提案した。

研究代表者はこれまで、マイクロアレイ解析の感度・特異度・再現性を飛躍的に向上させるデータ解析戦略に関する研究を行い、MAQC 推奨ガイドラインを全ての評価基準において凌駕するガイドラインの提案、およびその有用性を補強する研究成果を発表してきた。特に、「感度・特異度の高い DEG 検出法はこれまで常識とされた統計的な方法ではなく倍率変化に基づく方法（特に WAD 法）である」という結論は、他の複数の研究グループによってもその後確認されている [1]。

RNA-seq については、リファレンス配列へのマップ後のカウントデータが負の二項分布（NB 分布）でモデル化できることを利用した統計的な DEG 検出法がいくつか提案されている。また、以下の二つの問題に対処するためのデータ正規化法の開発も進められている：発現レベルのダイナミックレンジが広いがゆえに少数の高発現の DEG がデータの正しい正規化を妨げる（sequence depth 関連問題）、転写物の配列長が長いほど沢山シークエンスされるという影響をなくすための効果的な配列長の定義が難しい（transcript length 関連問題）。RNA-seq の現況は、研究代表者によるマイクロアレイ解析用ガイドラインが提案される以前のマイクロアレイ分野が歩んできた道と酷似している。つまり、RNA-seq の分野では、DEG 同定を目的とした統計的な検出法のみが注目され、上記 sequence depth 関連問題を解決するための既存の正規化法には改良の余地が多分にあり、マイクロアレイ分野で実証した DEG 検出法と正規化法の組合せの重要性はほとんど認知されていない。これらの現状を鑑み、上記に着目した RNA-seq データ解析手法の開発を行えば、マイクロアレイ分野で申請者が以前に実現した飛躍的な精度向上が見込めるのではないかという着想を得た。

２．研究の目的

本研究では、次世代シーケンサーから得られた RNA-seq データの解析精度を飛躍的に向上させることを目指し、(1) RNA-seq 分野で比較検討されていないもののマイクロアレイ分野で高い精度を誇る DEG 検出法 WAD の性能評価、および新規正規化法の開発を通して、適宜改良を加えながら(2) 比較トランスクリプトームデータ解析の推奨ガイドライン構築を行うことを目的とした。

３．研究の方法

(1) RNA-seq 分野で一般によく用いられている正規化法（RPM や TMM）を適用したデータに対して、既存の統計的な検出法と WAD 法の性能評価を行う。本研究では、主に実験デザイン上重要な NB 分布に従うシミュレーションデータを用いて解析を行った。具体的なシミュレーション条件は、過去の研究で実際に用いられた「2 群間比較（G1 群 vs. G2 群）を目的とし、DEG が全遺伝子数に占める割合を P_{DEG}%、そのうち G1 群で高発現のものが P_{G1}%を占め（*i.e.,* G2 群で高発現のものは 100 - P_{G1}%）、発現変動の度合いは全て定数倍（x 倍）となるようにしたデータ」を基本とし、その他のパラメータについても既報のものを用いる。過去の研究では、発現変動遺伝子の割合（P_{DEG}%）や偏り（P_{G1}%）について、きわめて少数の条件での比較しか行われていないため、できるだけ多くのシナリオ（P_{DEG}=5, 10, 20, 30% ×P_{G1}%=50, 60, …, 100%）について解析を行った。

現在 RNA-seq データ正規化法で最も高い評価を受けているのは、Robinson らの開発した TMM 法である。この方法は、どのデータに対しても「G1 群で高発現のもの上位 30%、G2 群で高発現のもの上位 30%を除いた残りのデータを用いて正規化係数を決める」戦略をとっている。つまり、実際のデータセット中の発現変動遺伝子の割合（P_{DEG}%）や偏り（P_{G1}%）に関わらず、「P_{DEG}=60%, P_{G1}=50% と固定しているのが TMM 法」といえる。

本研究で開発する正規化法の中核をなす戦略は、「実際のデータ中に含まれる真の DEG を候補としてできるだけ正確に推定（i.e., P_{DEG} とそれに付随する P_{G1} を自動的に推定）し、それらを除いた残りのデータを用いて正規化係数を決める」である。これを実現するために必要なステップは「a. データ正規化（1 回目）、b. DEG 検出法を用いた DEG 候補の推定、そして c. DEG 候補以外のデータのみを用いて 2 回目のデータ正規化」である。データ解析の一般的な流れは、「①データ正規化→②DEG 検出」の 2 ステップで完結する。一方、本研究のデータ解析戦略は「①a.正規化→b.DEG 検出→c.正規化→②DEG 検出）」に相当し、一般的な解析戦略を 2 回繰り返すことと本質的には同じといえる。この「DEG 除去戦略(DEG elimination strategy; DEGES)と WAD 法を組み合わせた方法の開発および評

価を行った。

WAD 法は、倍率変化を用いたときと同様に DEG 数を合理的に見積もる手段が存在しないため、客観的な閾値を設定しづらいという弱点がある。そのため、当初は統計的な RNA-seq 用 DEG 検出法である baySeq をステップ b で採用し、baySeq から得られる DEG 数の閾値分だけ、WAD の DEG ランキング結果上位から除き、残りの non-DEG のみでステップ c の正規化へと供するという戦略を立てた。また、個別の方法として評価の高い TMM 正規化法をステップ a と c で採用することにより、既存の RNA-seq 用解析手法と WAD 法を組み合わせたマルチステップのハイブリッド正規化法の開発が本研究の中核であった。

(2) RNA-seq データ解析用推奨ガイドラインの構築は、上記開発手法の評価以外にカウントデータ取得以降の頑健なサンプル間クラスタリングを行うためのフィルタリング法の開発を行った。RNA-seq カウントデータは、遺伝子など特定の領域にマップされたリード数からなる。そのため、マップされていない領域はゼロカウントとなる。また、NGS 機器から生み出されるリード数が数億リードレベルになっている。このため、従来の総リード数を 100 万に揃える RPM 補正を適用し、補正後のデータで低発現領域をフィルタリングすると、本当はフィルタリングしなくてもよいほど S/N 比が高いものまで除いてしまう可能性がある。もちろんプログラムのマニュアル中に書かれていることも多いが、エンドユーザはデフォルトで実行しがちである。このため、デフォルトでそのような可能性をできるだけ排除すべく、ユニークな発現パターンのみ残すようなフィルタリングを行い、Spearman 相関係数をサンプル間の類似性尺度として用いる R の関数を作成した。

４．研究成果

(1) RNA-seq データ正規化法内部に WAD 法を組み込んだ DEGES に基づく 3 ステップを基本とするマルチステップ正規化法と、通常の 1 ステップで行う正規化法を当初 Poisson 分布に従うシミュレーションデータを用いて評価した。結果として、WAD 法を組み込んだ方法の ROC 曲線下部面積（AUC 値）が高いことを確認した。しかし、論文投稿時に「DEG 数決定だけではなく DEG ランキング結果も含めて baySeq を利用するほうが直接的である」という査読者からの指摘を受け、結果的に DEGES 戦略提唱論文(Kadota et al., AMB, 2012)からは WAD 法に関する言及は消えた。

DEGES に基づく頑健な正規化戦略は、その後データ解析環境 R のパッケージ TCC として実装され、主に別の研究プロジェクトの取り組みとして既存の RNA-seq 用解析手法のみの組合せで高い精度を維持したまま劇的な高速化が達成されている(Sun et al., BMC Bioinformatics, 2013)。WAD 法は、TCC パッケージの中核であるマルチステップ正規化法の DEG 同定部分に実装済みである。前述のように、WAD 法は DEGEG の枠組みに組み込むことで、既存の non-DEGES 正規化法に比べて高精度であることはその後も確認済みである。しかし、劇的な高速化が達成された既存の RNA-seq 用解析手法のみの組合せに基づく DEGES に比べて劣っている。この主な原因は、WAD 法にとって明らかに不利なシミュレーション解析をやらざるを得ない事情による。つまり、NB 分布に基づくシミュレーションカウントデータを生成し、NB モデルに基づく統計的手法を実装した RNA-seq 用のパッケージとの比較を行っているためである。

RNA-seq データに基づく発現解析は、①同一サンプル内での異なる転写物間の発現レベルの大小関係を知る目的と、本研究の対象である②異なるサンプル間で DEG を同定する目的に大別される。現状では、前者は配列長補正を必要とし RPKM 値などが、そして後者は配列長補正を事実上行わないカウントデータが用いられる。しかし、本来であればこの両者は同じ入力データを用いて解析が行われるべきと考える。今後の展開として、NB 分布の枠組みから外れる配列長補正後のデータで比較することで、より公平な評価となることが期待される。

(2) RNA-seq カウントデータを入力として頑健なサンプル間クラスタリングを行う考え方を図書(門田幸二、共立出版、2014)中で示し、エンドユーザが簡便に利用できるよう TCC パッケージ中に clusterSample 関数として実装した。この関数をデフォルトオプションで実行すると、①全サンプルでゼロカウントの行を除き、②ユニークな発現パターンのみ抽出し、③1 - Spearman 相関係数を距離として定義し、④平均連結法で階層的クラスタリングを行う。一例として、ReCount データベースで提供されているヒトの様々な組織のカウントデータ(bodymap データセット)を入力として、①と②のフィルタリングの有無によるクラスタリング結果の違いを図 1 および 2 に示す。

図書の中では、組織特異的遺伝子が多く含まれていることが知られている testis の発現パターンが他の多くと似ていないことを論拠にしている。このデータセットの場合、step ①のゼロカウント行の寄与率(75%)が非常に高いが、②は実質的に①を含んでいる。また、一般に低発現遺伝子のフィルタリングは、DEG 同定前に行うことはあっても、サンプル間クラスタリング前に行うというコンセンサスはおそらくない。それゆえ、フィルタリングの有無でこれだけ結果が異なるという実例を示せたのは本研究の大きな成果と言える。

本研究で 1 - Spearman 相関係数を距離として定義したのは、順位尺度でデータを取扱う

ことでゼロカウントデータの取り扱いに悩まされず、またデータの正規性にも気を配らなくて済むからである。また、ユークリッド距離やマンハッタン距離などのサンプル間で要素の差分を足し込んで距離とする方法は、サンプル間でのデータの正規化にも気を配る必要がある。本研究で開発した DEGES に基づく正規化なども適用可能ではあるが、正規化の成否の保証はないことなどから、ユークリッド距離などを積極的に採用するメリットはないと考えた。今後は、より説得力のあるガイドラインとするため、これらの詳細な比較結果を原著論文としてまとめる。

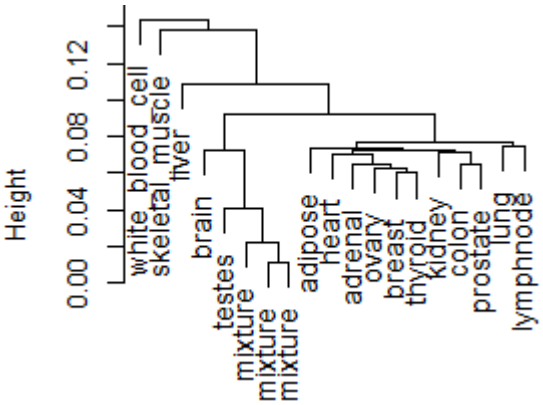

図 1．フィルタリングなし

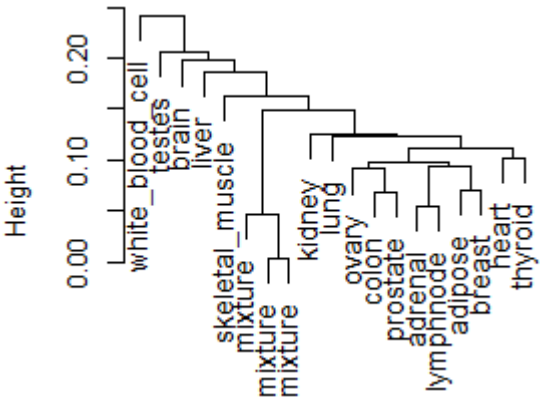

図 2．フィルタリングあり

＜引用文献＞

① Dembélé D1, Kastner P., Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinformatics*, **15**:14, 2014.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

① Sun J, Nishiyama T, Shimizu K, Kadota K, TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics*, 査読有, **14**:219, 2013. doi:10.1186/1471-2105-14-219

② Kadota K, Nishiyama T, Shimizu K., A normalization strategy for comparing tag count data. *Algorithms for Molecular Biology*, 査読有, **7**:5, 2012. doi:10.1186/1748-7188-7-5

〔学会発表〕（計 17 件）

① Sun J, Tang M, Shimizu K, Kadota K, TCC: an R/Bioconductor package for differential expression analysis of RNA-seq data, 8th AYRCOB, 2015 年 1 月 19~20 日, Tung Univ. (HSINCHU TAIWAN)

② Tang M, Sun J, Kadota K, Shimizu K, A comparison of methods for differential expression detection from multi-group RNA-Seq data, 8th AYRCOB, 2015 年 1 月 19~20 日, Tung Univ. (HSINCHU TAIWAN)

③ Tang M, Sun J, Shimizu K, Kadota K, Evaluation of methods for differential expression analysis from RNA-seq, 生命医薬情報学連合大会 2014, 2014 年 10 月 2～4 日, 仙台国際センター(宮城)

〔図書〕（計 2 件）

① 門田幸二著（金明哲 編），シリーズ Useful R 第 7 巻 トランスクリプトーム解析，共立出版，2014. ISBN: 978-4-320-12370-0

② 門田幸二,「トランスクリプトミクスの推奨データ解析ガイドライン」，ニュートリゲノミクスを基盤としたバイオマーカーの開発，シーエムシー出版，45-52, 2013. ISBN: 978-4-7813-0820-3

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

(R で)塩基配列解析
http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html

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