Vol.2013-BIO-33 No.9
2013/3/21

RNA-seq 解析における多群間比較に対応した正規化法

田中道廣[†1]　　藤渕航[†1]

RNA-Seq 解析は細胞中に存在する RNA 配列を直接定量するために，高精度な発現プロファイルを取得できることが期待されている．一般的に，RNA の総量はサンプル間で異なるため，複数のサンプルに由来する発現プロファイルを比較する場合には，サンプル間での発現値の正規化が不可欠である．これまで，二群間比較における Trimmed mean of M values (TMM)正規化法（Robinson et al., 2010）は提案されていたが，多群間比較に拡張することは難しかった．今回我々は, TMM 法を多群間比較に拡張可能な手法を開発した．

1. はじめに

細胞中に発現している RNA 配列を直接定量する RNA シーケンス（RNA-Seq）解析は，シグナル強度比を使うマイクロアレイでは定量が難しかった低発現遺伝子の検出を可能にする広いダイナミックレンジを実現する．また，一回の実験で多数のサンプルを同時にシーケンス解析するマルチプレックス法[1]の登場により、マイクロアレイと比べて 10 分の 1 のコストで発現プロファイルを得ることも可能になった [2].

遺伝子の発現量を定量する際にマイクロアレイ解析がシグナル強度比を使うのに対して，RNA−seq 解析ではゲノムにマッピングされたシーケンスリードの数を使用する．サンプルの総カウント数をそろえる Reads Per Million mapped reads (RPM), それに加えて遺伝子長をそろえる Read per kilobase of exon model per million mapped reads (RPKM)[3]といった正規化法が提案された．しかし，これらの正規化法は総リード数を使った補正法（遺伝子発現量を総リード数あたりの割合に変換）で計算しているために、一部の高発現している遺伝子が存在する場合, 他の発現遺伝子における発現差を過剰に見積もる可能性が指摘されている[3]. Robinson らはこの問題を解決するために Trimmed mean of M values (TMM)法[3]を提案しているが，二群間での比較を想定しており，コントロールサンプルを想定しない組織特異性を解析する際に必要な多群間での比較解析には, 適応が難しかった.

そこで本論文では 2 群間比較を前提とした TMM 法拡張し多群間比較に対しても有効な解析手法として提案する．

2. 提案手法

2.1 TMM 法

TMM 法では正規化定数を以下の手順で算出する．（１）すべての遺伝子について発現比と発現量を表す M 値と A 値を計算する．（２）すべての M 値について，上位 30%と下位 30%の遺伝子を取り除く．（３）すべての A 値について上位 5%と下位 5%の遺伝子を取り除く．（４）残った遺伝子の発現値から正規化係数を計算する．ただし、TMM法は 2 群間比較を前提としているために，コントロール実験を仮定しない多群間比較の場合には M 値とＡ値が算出できない問題があった.

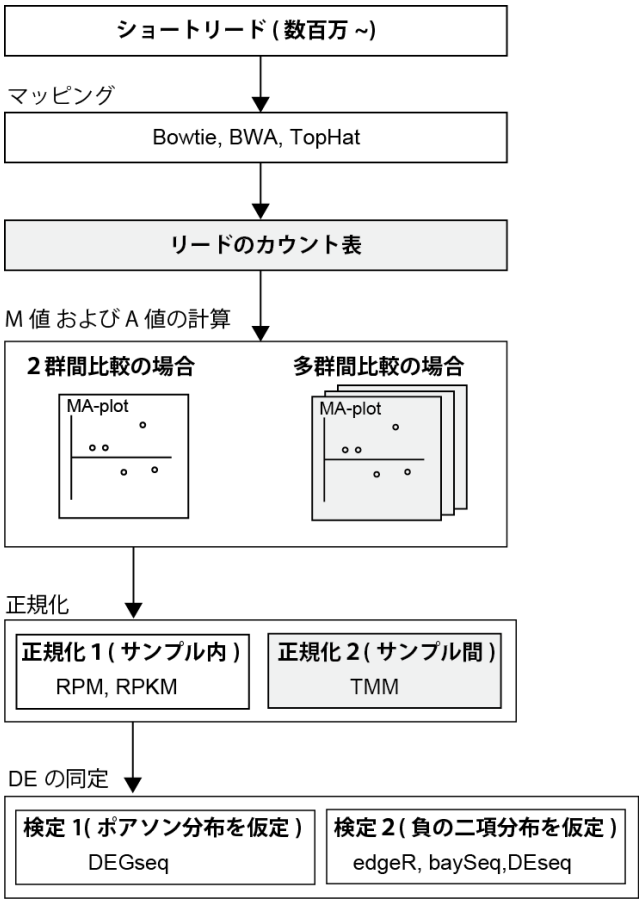

図 1　提案手法の流れ

2.2 提案手法

本研究では，二群間比較を前提に設計された TMM 法[4]を多群間比較に適応させるために，ライブラリ k についての重み付き正規化係数を計算する際に，ライブラリ k を除く残りのライブラリ（対照ライブラリ）それぞれについて M 値を計算して正規化係数の算出を行う手法を提案する．図１に提案手法の一連の流れを示す.

†1 京都大学 iPS 細胞研究所
Center for iPS Cell Research and Application, Kyoto University

2.2.1 M値およびA値の計算

ライブラリ k を構成している遺伝子 g=1, 2, 3, …, G について、対照ライブラリ r を過程する時, 発現比 $M^r_{g,k}$ と発現量 $A^r_{g,k}$ を以下の様に定義する。

$$M^r_{g,k} = \log_2 \frac{\left(\frac{Y_{g,k}}{N_k}\right)}{\left(\frac{Y_{g,r}}{N_r}\right)} \quad (1)$$

$$A^r_{g,k} = \frac{1}{2}\log_2\left(\frac{Y_{g,k}}{N_k}\frac{Y_{g,r}}{N_r}\right) \quad (2)$$

2.2.2 トリム処理する遺伝子の同定

ライブラリ数 n の多群(2<n)のデータについて、ライブラリ k の正規化係数を算出する場合，ライブラリ k 除く残りのライブラリを対照ライブラリ l_j (j=1, 2, …, n)と定義する. ライブラリ k と全ての対照ライブラリの間で M 値および A 値を計算する．１つ以上の対照ライブラリで条件(１) M 値について, 上位 30%と下位 30%および条件(２) A 値について上位 5%と下位 5%を満たす遺伝子をトリム対象とする.

TMM 係数の計算

ライブラリ k の重み付き TMM 係数を以下のように定義する。

$$\log_2(TMM_k) = \frac{\sum_{g=1}^{G} w_g \sum_{C=1}^{n} M^{l_j}_{g,k}}{\sum_{g=1}^{G} w_g} \quad (3)$$

ライブラリ C に関する遺伝子 g のカウント数が $Y_{g,C}$，ライブラリ C を構成する遺伝子の総カウント数が N_C である場合、遺伝子 g の重み係数 w_g を以下のように定義する.

$$w_g = \sum_{g=1}^{G} \frac{N_C - Y_{g,C}}{N_C Y_{g,C}} \quad (4)$$

カウント数の正規化

ライブラリ C に関する遺伝子 g のカウント数 E^C_g の正規化は TMM 法と RPM 法と併用して以下のように計算する.

$$normalization\ E^C_g = E^C_g \frac{1000000}{\sum_{g=1}^{G} Y_{g,C}\, TMM_C} \quad (5)$$

3. 実験

3.1 実験方法

本実験では、R package TCC から提供されている関数 generateSimulationData を使用して作成したデータセットを実験に用いた。データセットの構成は以下の通りである.

3.1.1 発現プロファイル

Biological variations が負の二項分布に従う[5]と仮定し，同分布から，A 群 B 群 C 群について遺伝子数が 10,000 からなる発現プロファイルを生成した.発現値はリード配列のカウント数，それぞれ個体について全遺伝子数の 30%を発現変動遺伝子とした。発現変動遺伝子は A 群, B 群, C 群にそれぞれ 70%, 20%, 10%, Fold change はそれぞれ 3 , 10, 6 として構成した.

3.1.2 評価指標

TMM 補正後の発現プロファイルからトリム対象外となった遺伝子を収集し, TMM 正規化の効果を検証した.

3.2 結果

シュミレーションデータセットを用いた実験結果を図２に示した.

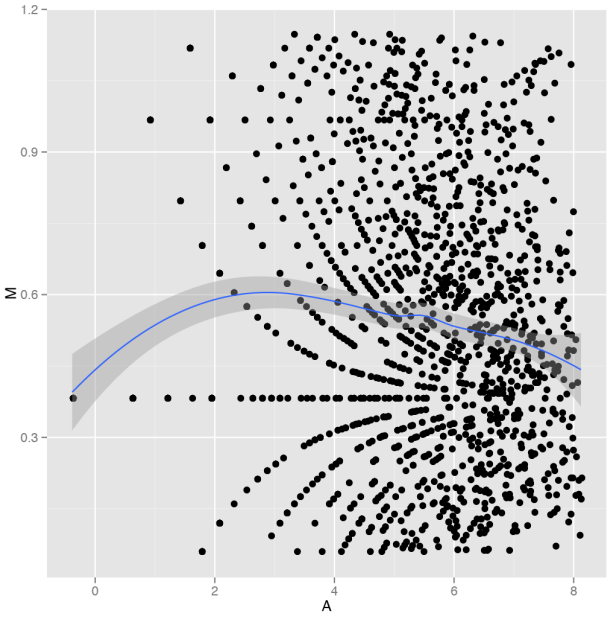

図 2 ライブラリ A と対照ライブラ B に関する MA プロット（トリム処理対象外となった 1,506 遺伝子について）

4. 考察

図２からは発現量が低い遺伝子についての正規化処理の効果が十分でないことがわかる．これは TMM 係数を形成している重み w_g の影響と考えられる．また本手法では多群間比較を適応させるために各遺伝子はトリム処理される確率は 2 群間比較の場合と比べて高くなる．この場合，正規化係数の計算に使用できる遺伝子が少なくなる.したがって，正規化係数の計算に使う遺伝子数を確保しつつ，トリム処理の対象となる基準をいかに，向上させていくかが今後の課題である.

参考文献

1) Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P and Linnarsson S.: Highly multiplexed and strand-specific single-cell RNA 5’ end sequencing. Nature Protocol, Vol.7, No.5, pp.813-828 (2012).
2) Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P and Linnarsson S.: Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Biology, Vol.21, No.7, pp.1160-1167 (2011).

3) Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods, Vol.5, pp.621-628 (2008).
4) Robinson, MD and Oshlack, A.: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, Vol.11, No.3 (2010).
5) Robinson, MD and Smyth, GK.: Moderated statistical tests for assessing differences in tag abundance. Bioinformatics, Vol.2, No.10, pp.2881-2887 (2007).