審 査 の 結 果 の 要 旨

氏　名　湯 敏

ハイスループットシークエンシング（HTS）の主な用途は、RNA 転写物の発現（RNA-seq）を測定することである。マイクロアレイは、既知の転写物の検出と転写変異体の識別に、その用途が限定されるのに対し、RNA-seq では、その配列、構造、および発現レベルに照らして、セル内の全ての転写物を検出することができる。異なるグループまたは条件で発現変動遺伝子（DEG）を同定することは、RNA-seq の重要な目的の一つである。DEG の解析では、まず、各遺伝子に、シークエンシングサンプルからマップされたリードの数を要素とするカウント行列が構築される。

これまで、RNA-seq のカウント行列をもとに DEG を検出するための多くの方法が提案されている。RNA-seq のようなダイナミックレンジの広いデータの正規化では、高発現の DEG の存在が正しいデータ正規化を阻み、結果として高精度の DEG 検出を阻むという問題があった。この問題に対して、申請者の研究室では、正規化法の内部で DEG 検出を行い、DEG 以外のデータのみで正規化を行う手法を開発し、既存のパッケージの手順よりも、これらのパッケージ中の関数を組合せた DEG Elimination Strategy（DEGES）のほうが有効であることを示した。さらに、頑強な DEGES 正規化の考え方を一般化し、これを改良した手法を開発し、R パッケージとして実装した。しかしながら、従来開発されてきた手法は、基本的に 2 群間比較にその目的が限定されている。実際の利用では、3 群以上の多群間比較あるいは多因子比較を利用する場合が多いが、基本的に 2 群間比較の積み重ねで実現する必要があり、これらの手法の最適な組み合わせについては、十分な評価が行われてこなかったのが現状である。本研究では、3 群間比較に対して、性能評価を行い、推奨ガイドラインを提案することを目的としている。

第 1 章で序論として本研究の位置づけについて述べた後、第 2 章では本研究で評価の対象としたデータ正規化および DEG 検出の手法について記述している。本研究では、これらの手法の組み合わせに対して、3 群間比較の性能評価を行った。

第 3 章では、シミュレーションデータによる性能評価について述べている。申請者は、データ正規化および DEG 検出の 9 種のソフトウェアパッケージ TCC（ver. 1.7.15）、edgeR（ver. 3.8.5）、DESeq（ver. 1.18.0）、DESeq2（ver. 1.6.3）、voom in limma（ver. 3.22.1）、

SAMseq in samr（ver. 2.0）、PoissonSeq（ver. 1.1.2）、baySeq（ver. 2.0.50）、EBSeq（ver. 1.6.0）の組み合わせで 12 種の解析パイプラインを構築した。TCC は、申請者の研究室で開発された独自のもので、多段の正規化手法 DEGES の枠組みに従ったもので、内部で edgeR、DESeq、DESeq2 を使用している。ここで、データ正規化を X、DEG 検出を Y で表すと、RNA-seq 全体の解析手順は、X-Y のパイプラインとして記述できる。TCC で実装されているパイプラインは、X-$(Y\text{-}X)_n$-Y と記載される。ここに n は、繰り返しの回数である。デフォルトは n = 3 であり、回数を省略して XYX-Y と記載することとし、さらに X=Y の場合は、単に all-X パイプラインと呼ぶこととする。シミュレーションデータは、10,000 個の遺伝子に対して、5～25% の遺伝子が DEG であるとした。性能指標としては、感度および特異性を同時に評価するため、AUC（area under the receiver operating characteristic curve）値を用いている。DEG の 遺伝子の 4 つの異なる値に対し、100 試行した結果の平均 AUC 値で評価を行ったとき、データ複製がある場合は、all-edgeR パイプラインが最も高い性能を示し、データ複製がない場合は、all-DESeq2 パイプラインが最も高い性能を示すことを明らかにしている。2 群間比較ではデータ複製がない場合、all-DESeq パイプラインが最も高い性能を示したが、DESeq2 は、その後に発表されたものであるため、2 群間比較においても、all-DESeq2 パイプラインが高い性能を示すことが予想される。

第 4 章では、ヒト、チンパンジー、アカゲザルの実験データを用いた性能評価の結果を示している。第 3 章と同じ 12 種のパイプラインのうち、all-edgeR、all-DESeq、all-DESeq2 パイプラインが類似の傾向を示すことを明らかにしている。さらに、サンプル間クラスタリング結果から DEG 検出結果のおおよその見積もりが可能であることも示している。

以上、本研究では RNA-seq の 3 群間比較の手法について、現在使用されているデータ正規化および DEG 検出の手法を組み合わせて網羅的な性能評価を行い、推奨されるパイプラインを提案した。その研究成果は、学術上応用上寄与するところが少なくない。よって、審査委員一同は本論文が博士（農学）の学位論文として価値あるものと認めた。