

解 説

次世代シーケンサーデータの解析手法 第16回 なぜ次から次へと新規手法が開発されるのか？

門田 幸二^{1, 2, 3*}、清水 謙多郎^{1, 2, 3}

¹ 東京大学 大学院情報学環・学際情報学府

² 東京大学 大学院農学生命科学研究科

³ 東京大学 微生物科学イノベーション連携研究機構

我々はよく、「同じ目的のプログラムが多く存在し、どれを使えばいいのかよくわからない。」という感想を聞く。考えられる理由としては、多様な評価基準が挙げられる。例えば、同程度以下の性能であっても、従来法よりも高速であれば存在意義が認められる。また、同一群内の反復数が多い（または少ない）といった特定の条件下で効果的な発現変動解析手法もおそらく存在する。しかしながら、本当の意味で存在意義が疑われる手法も現実には存在する。おそらくこれは、比較対象が代表的な手法に偏っているためである。本稿では、なぜ疑わしい手法が存在しうることについて、RNA-seq 発現変動解析分野を例として筆者らの経験を交えて解説する。

Key words : RNA-seq, research integrity, normalization, symmetry/asymmetry

はじめに

世の中には無数の論文が存在する。そのため、今回の対象分野に限ったとしても、全てを網羅するのは事実上不可能である。実際、新規手法に関する論文投稿時に、全ての既存手法を含めた性能評価が要求されることはない（要求されたことはないし、査読者の立場でも要求したことはない）。ただし、edgeR¹⁾やDESeq2²⁾のようなこの分野でよく用いられている代表的なRパッケージは、手法比較論文³⁾などでもベンチマークとして含められるのが通例である。それゆえ、例えば後発の新規手法Aの投稿論文は、上記2つの手法(+a)と同程度以上の優れた何かがあることが認められれば受理される。これは別の新規手法BやCについても成立するため、結果として代表的な手法よりも優れていると結論付けられた3つの新規手法AとBとCが存在しうる。もし同じ解析条件なら、直接的な比較を通じてABC間の優劣を決めることができる。しかし往々

にして、それぞれの手法開発論文では、その長所を示せるようなデータで比較される。これは、仮に利害関係のない別の研究グループによる手法比較論文であったとしても、例えば手法Cの論文で用いられた枠組みに近い条件で比較を行えば、Cのほうが優れているといった結論になりがちだということを意味する。

次に、手法Cと似た特徴をもつ、さらに後発の新規手法C'について考えてみる。我々の感覚では、新規手法C'の開発グループは、彼らの責任において類似した特徴をもつ論文がないかを丁寧に探さねばならない。そして競合する既存手法Cを見つけ出し、C'が代表的な既存手法だけでなくCよりも優れていることを示さねばならない（さもなくばC'の存在価値はない）。万が一C'の存在に気づかずC'の論文が投稿されたとしても、査読者によってC'の存在が適切に指摘されれば、現行の査読システムがうまく機能したといえよう。リバイス時にC'が適切に引用され、CにはないC'の存在意義が示されれば、投稿先の基準に従って可否が判断されるであろう。今回は、Viethらによる単一細胞RNA-seq (scRNA-seq) のデータ解析パイプラインの性能比較論文⁴⁾に対して重大な疑義をもち、事実を列挙して痛烈に批判した我々の批評論文⁵⁾の内容を中心に解説する。大まかな関係性としては、Vieth論文がC'

*To whom correspondence should be addressed.

Phone : +81-3-5841-2395

Fax : +81-3-5841-1136

E-mail : koji.kadota@gmail.com

に、そして我々が以前開発した RNA-seq データ正規化法 DEGES⁶⁾ が C に相当する。

ステレオタイプなイメージの形成

今回のタイトルで示した問題へのユーザ側の対処法としては、様々な手法の特徴を把握し、自身のデータの性質・デザイン・解析目的に合致した手法を利用するのが理想である。一方で現実には、インストールのしやすさ、マニュアルの豊富さや丁寧さ、そして不明点や疑問点を問い合わせた際に誠実に回答してくれるかなどが重要である。ストレスのない実行時間や、インストール不要のウェブツールとしての提供などもまた、利用しやすさや R に馴染みのないユーザへの門戸を広げるという意味で重要といえる。実際、以前我々が edgeR 開発者にメールで問い合わせた際には丁寧な回答をいただいた。また DESeq2 は、前回⁷⁾も紹介した Galaxy⁸⁾ 上で (インストール不要で) 利用可能である。これら 2 つのパッケージは、(DESeq2 の前身である DESeq⁹⁾ を含め) RNA-seq の普及初期から存在し、今日でも継続的なアップデートがなされている。論文引用回数も 1 万回超 (2021 年 7 月 26 日に PubMed で調べた結果) であり、「勝者が全てを得る (Winners take all)」の典型例といえよう。

遺伝子 (転写物) の働き具合は、個体・組織・状態・細胞ごとに異なる。転写物の全貌 (トランスクリプトーム) を網羅的に調べる実験技術は、「臓器を塊 (bulk) として平均的な情報を調べる bulk RNA-seq」から「細胞ごと (single cell) の情報を調べる scRNA-seq」に移行している。edgeR や DESeq2 は元々 bulk 用に開発されたものであるため、これらが bulk 用の代表的な手法という位置づけになる。したがって、edgeR や DESeq2 の特徴が bulk 用の発現変動解析手法に対するステレオタイプなイメージということになる。内部的に用いられている正規化法についても同様である。edgeR は TMM 法¹⁰⁾ を、そして DESeq2 は MR 法⁹⁾ を利用しているが、これらは①大部分の遺伝子は発現変動遺伝子 (DEG) ではない (つまり non-DEG である) というシナリオを想定している¹¹⁻¹²⁾。必然的に、② bulk 用の正規化法はこのような古い発想に基づいているというのがステレオタイプなイメージということになる。

「古い発想」というのは、実際には全遺伝子の半分以上が DEG だと記された論文が存在することを根拠としている。例えば③ Zeisel らは、scRNA-seq の異なる細胞型間の比較で ~60% の DEG を同定している¹³⁾。また、④酵母の 2 群間比較用 bulk RNA-seq データ¹⁴⁾ でも同様の結果が得られている¹⁵⁻¹⁶⁾。ここで注意してもらいたいのは、実際によく用いられているのは edgeR と DESeq2 であることは間違いのないものの、⑤他にも数多くの (bulk) RNA-seq 用の発現変動解析手法 (正規化法含む) が存在し、必

ずしもその全てが①のシナリオを想定しているわけではないということである。例えば、RNA-seq 用として開発した我々の DEGES 正規化法は、①のようなシナリオを想定しているとは一言も書いていない⁶⁾。また、⑥仮に①のシナリオを想定していたとしても、そのことが DEG が多数を占めるデータに対応できないことに直結するわけではない。しかしながら、④・⑤・⑥に触れることなく①のみを示すような論文が続出すれば、②のイメージを定着させることができってしまうのである。

Vieth 論文の概要

scRNA-seq 解析分野では、ウェットとドライ両面で次々と新しい方法が開発されている。Vieth らは、前者としてライブラリ調製プロトコル、後者としてリードのマッピングから発現変動解析 (データ正規化含む) までを調査対象としている⁴⁾。特に後者は複数のステップからなるため、前者と合わせた計 2,979 通りの可能な組合せ (パイプライン) を、計 999 通りのシミュレーション条件を用いて性能評価している。主なシミュレーション条件は、2 群間比較 (A 群 vs. B 群)、遺伝子数は 10,000 個、DEG の割合に相当する Effect size が 5%・20%・60% の 3 通り、そして DEG の発現変動パターン (DE-patterns) が下記の 3 通りである。

- ・対称 (DEG の 50% が A 群で高発現 ; DEG の 50% が B 群で高発現)
- ・非対称 (DEG の 75% が A 群で高発現 ; DEG の 25% が B 群で高発現)
- ・完全非対称 (DEG の 100% が A 群で高発現 ; DEG の 0% が B 群で高発現)

例えば Effect size が 20% で DEG が非対称なら、 $10,000 \times 0.2 = 2,000$ 個の DEG のうち A 群で高発現のものが $2,000 \times 0.75 = 1,500$ 個、そして B 群で高発現のものが $2,000 \times 0.25 = 500$ 個となる。また、Effect size が 60% で完全非対称なら、 $10,000 \times 0.6 = 6,000$ 個の DEG の全てが A 群で高発現となる。他にも 3 通りのサンプルサイズ (96 vs. 96, 384 vs. 384, and 50 vs. 200 細胞) や発現変動の度合いなどがあるものの、性能に直結する主な事柄としては DEG の発現変動パターンが挙げられている。具体的には、DEG が対称の場合にはライブラリ調製が、そして非対称の場合には正規化が発現変動解析結果に大きなインパクトを与えると結論づけている。特に⑦正規化の影響が最も大きく、非対称の度合いが高くなるほど「scRNA-seq 用の正規化法である scran¹⁷⁾ や SCnorm¹⁸⁾」が「代表的な bulk RNA-seq 用の正規化法 (や他の scRNA-seq 用の正規化法である Linnorm¹⁹⁾)」よりもよい性能を示すと述べている。

批評論文の概要

2つの重要な事柄が挙げられる。1つ目は、Vieth 論文ではまず①～③が述べられ、bulk RNA-seq用の正規化法として代表的な TMM および MR のみと比較した結果として、⑦が書かれているという点である（正確には Positive Count というものも比較しているようだが、詳細が全く説明されていない上に特筆すべき性能でもないためここでは無視する）。2つ目は、Vieth らが⑦と同趣旨の内容を Fig. 3b の考察として述べる際に「(bulk RNA-seq 用の正規化法を比較した) Evans らの結果と同様に」と記載している点である（原文：In line with Evans et al. (2018), …）。Evans 論文は、我々の DEGEGS 正規化法を含む様々な bulk RNA-seq 用の正規化法を、対称・非対称を含む様々なシミュレーションシナリオで比較したものである²⁰⁾。つまり、⑧ Vieth らは非対称なシナリオの bulk 解析結果が既に存在することを認識している。

Evans 論文は bulk 用のシミュレーション解析であるため、サンプルサイズ (2 vs. 2 and 5 vs. 5) が Vieth 論文に比べて圧倒的に少ないという違いは確かに存在する。しかし、Vieth 論文の結論に直結する事柄は、サンプルサイズが多いか少ないかではなく、DEG の発現変動パターンが対称か非対称かである。Vieth 論文では、おそらく①のシナリオが Effect size = 5% や 20% に相当し、③を反映したシナリオが Effect size = 60% なのだろう。Vieth らは、Introduction の 3 段落目で③を事例として挙げつつ、「非対称の問題は、概念的には scRNA-seq と bulk RNA-seq を区別する特徴の 1 つであり、これまで取り組まれてこなかった（原文：This issue of asymmetry is conceptually one of the characteristics that distinguishes single cell from bulk RNA-seq and has not been addressed so far.）」と述べている。この問題を Vieth 論文のモチベーションとして掲げていることを踏まえれば、⑦を主要な結論とするロジックに特段の違和感を感じないだろう。

しかし実際には、③と同様の Effect size を示す④ bulk データが存在するだけでなく、⑧の文章が Vieth 論文の Discussion 中に存在するのである。Evans 論文は、対称・非対称だけでなく、Effect size が 5-95% まで 5% 刻みで調べている。つまり Evans 論文は、Effect size の観点でいえば完全に Vieth 論文のシナリオを含むのである。したがって、この時点で Vieth らがモチベーションとして掲げた事柄に大きな疑問符がつく（つまり「has not been addressed so far」という主張は間違い）。さらに、⑨ Vieth 論文で用いられた Effect sizes (= 5, 20, and 60%) の範囲を Evans 論文の結果で眺めると、我々の DEGEGS が代表的な bulk 用の正規化法 (TMM と MR) よりも明らかに優れていることがわかる（例えば Evans 論文の Figure 7 や 8）。Vieth らは、Evans 論文の存在を認識している（つまり⑧）にもかかわらず、⑨の事実に触れることなく⑦を述べているのである。

批評論文の中では、我々は特にこの点を痛烈に批判している。

研究者としての誠実な姿勢

本稿で強調しておかねばならないのは、Vieth らの⑦の主張自体は決して捏造でも法に触れるものでもないということである。理由は、「正規化が結果に与えるインパクトが大きいことや、bulk 用の“代表的な”正規化法より優れているというのは、彼らの解析によって得られた事実関係を正確に述べている」からである。我々が批判しているのは、あくまでも⑩読者に誤解を与えぬよう誠実な姿勢で論文を執筆すべきだという研究者としての倫理観についてである。本稿や我々の批評論文を知ることなく Vieth 論文を読めば、おそらくほとんどの読者は「scRNA-seq は bulk RNA-seq の頃に比べて得られる DEG の割合 (Effect size) も非常に多く、データの性質自体も大きく異なる。得られる DEG の割合が少ないと仮定して設計された bulk RNA-seq 用の正規化法の性能は、scRNA-seq 用に比べて特に非対称なシナリオの場合に劣っている。よって scRNA-seq 用として開発された scran や SCnorm を使おう。」といった思考回路になるであろう。しかし事実は、④・⑤・⑥・⑨なのである。

もし非対称なシナリオを争点化したいのであれば、bulk 用の“代表的な”正規化法だけでなく、そのシナリオで“ベストな”正規化法（つまり DEGEGS）も正直に含めて比較すべきであろう。我々は、これが⑩の具体的な行動指針であると考えている。もし Vieth らが④・⑤・⑥・⑨を正しく認識し、正直に bulk 用の DEGEGS も含めて比較していたならば、彼らの論文の Introduction（特にモチベーション部分の説明）や⑦の結論は異なるものになっていたであろう。一部の研究者は、なぜその研究を行う必要があるかという論文のモチベーションに大きな疑問符がつけられる論文や、結論が異なりうるほどの事案を含む論文は撤回すべきだと思うかもしれない。また別の研究者は、今回我々が指摘したような事柄は、ただの業界の慣習であり黙認すべきだと考えているかもしれない。でなければ、論文を出しづらくなってしまふからである。どうあるべきか？については、それぞれの価値観に従って実践すればよいだろう。

今回のタイトルで示した疑問に対する、おそらく全員が納得できる理由付けは、本稿の要旨でも述べた多様な評価基準である。例えば計算時間がかかってもよいから精度を重視したい局面もあれば、たとえ同一研究者であったとしても、状況次第ではその逆を望む場面も十分にありうる。しかし本文中の冒頭でも述べたように、⑩自分の専門分野の全ての論文およびその内容を完全に把握していると言い切れる研究者はおそらく皆無である。それでも我々が Vieth 論文に対する批評を書くことを決めた大きな理由は、⑧だと判断したためである。もし彼らが Evans 論文を引用していなければ、我々は⑩を考慮して批評論文を書

くことはなかった。我々は私信で Vieth らとやりとりをしていないため真意は不明であるが、おそらく彼らは Evans 論文の詳細 (つまり⑨) までは把握しきれていなかったものと思われる。しかしもし仮にそうであったとしても、⑫論文の第一義的な責任の所在は著者らにある。研究者は⑩を常に意識すべきであろう。

査読者の果たすべき役割

論文の査読に対する考え方はヒトそれぞれだと思われるが、たとえ C のような存在意義の疑わしい論文であったとしても、ミスや疑念を建設的に指摘する使命感をもって査読が行われるのが理想である。例えば、我々の批評論文の核心部分である「④・⑤・⑥・⑨」は、本来査読の時点で適切に指摘がなされるのが理想である。また、⑬否定文とすべきところが肯定文になっているようなミスがもしあれば、それに気づき適切に指摘するのが査読者の役割であろう。しかしながら、Vieth 論文には⑬のミスが Introduction の第 3 段落中に存在する⁵⁾。この種のミスは、ある程度経験を積んだ研究者であれば、自己修正した上で読み進めることができる。しかし、学術論文の記載内容を読み解く能力がまだそれほど高くないものにとっては、理解を妨げる大きな混乱のもととなる。世界中の学術論文の内容を学習し知識の体系化を行うような取り組みを行う上でも、情報源側のミスをできる限り取り除くことが非常に重要である。

⑫の大前提のもとではあるが、今回 Vieth 論文の査読者全員が⑬のミスに気付かなかった点も非常に残念である。もし我々の批評論文の核心部分である「④・⑤・⑥・⑨」の一部でも査読者が適切に指摘し、Vieth らがそれに誠実に対応できていれば、今回のような痛烈な批判を受けずに済んだかもしれない。

おわりに

今回の内容は手法開発者側の視点で書いたが、大半の読者はエンドユーザー側である。また、一般には一読者としてかわる論文数のほうが、著者や査読者としてかわる論文数よりも圧倒的に多いだろう。しかしながら、今回紹介したような事例はどの分野においても多かれ少なかれ存在する。そのため、著者側に期待する高い倫理観や査読者側に期待する高い使命感だけではなく、(この原稿を含めて)読者として論文を批判的に読み解く姿勢も重要である。例えば、⑨は「Evans らによる性能比較結果で“DEGES が抜群の性能を示すギリギリの範囲”を Vieth らがたまたま採用していたおかげで、運よく非常にスッキリとしたロジックで批評論文を書くことができたわけではないのか？」と勘ぐってほしい。

これは実際そうである。もし Vieth らが(対称はともか

く非対称のシナリオで) Effect size = 70 ~ 80% を採用していれば、たとえ批評論文を書いていたとしてもそのロジックに苦勞したであろう。理由は、Evans 論文の Figure 7 の非対称シナリオによれば、Effect size = 70 ~ 80% で最も性能が高いのは DEGES でも TMM でも MR でもなく Total Count だからである。これは、第 15 回でも紹介した CPM 補正と実質的に同じ基本的な正規化法に相当するものである。ただし、Total Count は対称のシナリオでは一貫して劣っている点にも注意せねばならない (Evans 論文の Figure 7 の下半分)。このように、Evans 論文の Figure 7 を眺めるだけでも、シナリオ次第で精度の高い方法が変わりうるのがわかる。

我々の DEGES 正規化法は、2010 年に出版された TMM 正規化法にヒントを得たものである。TMM のアルゴリズムは、DEG の発現パターンが対称の場合にはうまくいくが、非対称の場合には理論上うまくいかない(ことに気づいた)。それゆえ、対称のシナリオでの理論性能は同じだが、非対称の場合でもうまくいくように設計したのが DEGES のアルゴリズムである。しかしながら、DEGES 開発当時 (2010-2011 年) は今回の主要なトピックである非対称なシナリオが DEG の発現パターンとして存在することが想定されていなかった頃である。そのため、査読者から「本当に非対称となるような事例が存在するのか？」と指摘され、「そのようなシナリオの場合でも理論上うまくいくのが提案手法だ。」と反論したことを今でもよく覚えている。

DEGES 論文の Table 1 では、Effect size (論文中では P_{DEG} と表記) が 5%・10%・20%・30% の 4 通り、そして DEG の発現変動パターンが対称 ($P_A = 50\%$) から完全非対称 ($P_A = 100\%$) までの 6 通りのシナリオ (計 $4 \times 6 = 24$ 通り) で、(a) DEGES (論文中では edgeR/TbT と表記) と (b) TMM (論文中では edgeR/default と表記) の性能評価を行っている。DEGES はどのシナリオでも安定した性能を示す一方で、TMM は対称 (表の左側) から完全非対称 (表の右側) になるにつれて徐々に性能が低下しているのがわかる。そしてその低下の度合いは、Effect size が大きいほど顕著であることもわかる。つまり我々は、2012 年にはこのような知見を bulk のシミュレーションデータですでに得ていたということである。それゆえ、Vieth らのシミュレーション条件で性能評価を行えば、Effect size = 60% で非対称なシナリオの場合に「⑦正規化は結果に与えるインパクトが最も大きい」となるのはごく当たり前のこととして受け止められるのである。逆に、対称なシナリオの場合には正規化法の違いによる性能差が理論上出にくいいため、必然的に正規化法以外のものが議論の対象になるのも当たり前である。また、もし Effect size が 10% 程度以下で評価が行われれば、「対称・非対称であろうと正規化法の違いは性能にほとんど影響しない」という知見が得られるであろう。論文の結論の文章だけに着目している

と、矛盾する知見が得られうる。読者は、結論が導きだされた解析条件にも注意を払わねばならない。

Evans 論文の結果に着目すると、Effect size が 60% 程度以下であれば DEG の発現変動パターンによらず DEGES が最もよい性能を発揮すると解釈してよい。30% 以下の範囲に限定すれば、さらにそれを DEGES 論文が補強する形となる。ただし、2019 年に出版された書籍²¹⁾でも言及しているが、我々は 50% 以上の DEG を含むようなデータをそもそも想定していない。それゆえ、今でも「70~95% という異常に高い Effect size のシナリオまで含めて考えると DEGES はそれほどでもない (原文: DESeq and TMM generally perform well)」とも読み取れる Evans 論文の内容には同意しかねる部分も存在する。しかしながら実際には、③や④という Effect size の大きなものだけでなく、非対称なリアルデータも存在する²²⁻²³⁾。このことを踏まえれば、現在我々が非常識だと思っているシナリオに合致したリアルデータも既に存在するのかもしれない。

Vieth 論文は、scRNA-seq が主流になった現在においても依然としてデータの正規化が重要であることを示した点で意義はある。また、我々の批評論文はあくまでも

DEGES を比較対象に加えるべき点までであり、Vieth らが推奨する scran や SCnorm よりも優れているとまでは主張していない点に注意してほしい。Lieberman らが間接的に述べているように²⁴⁾、あくまでも scRNA-seq 用の正規化法が bulk 用のものよりも優れていると主張できなくなっただけにすぎない (DEGES is competitive with scRNA-seq normalization methods) というのが正しい現状認識であろう。データの正規化は、マイクロアレイの時代から今日まで継続的に議論されている。実際、我々の大学院生向けの講義においても、時間の関係上省略すると必ず不満が出るほどである。今回は、コロナ渦で激変した我々のバイオインフォマティクス教育の取り組み全般について紹介する予定である。

謝 辞

本連載の一部は、JSPS 科研費 18K11521 の助成を受けたものです。

利益相反 (COI)

門田幸二、清水謙多郎：本論文発表の内容に関連して開示すべき COI 状態はない。

参 考 文 献

- 1) Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.
- 2) Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**: 550.
- 3) Stupnikov A, McInerney CE, Savage KI, McIntosh SA, Emmert-Streib F., et al. (2021) Robustness of differential gene expression analysis of RNA-seq. *Comput Struct Biotechnol J.* **19**: 3470-3481.
- 4) Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I (2019) A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun.* **10**: 4667.
- 5) Kadota K, Shimizu K (2020) Commentary: A Systematic Evaluation of Single Cell RNA-Seq Analysis Pipelines. *Front Genet.* **11**: 941.
- 6) Kadota K, Nishiyama T, Shimizu K (2012) A normalization strategy for comparing tag count data. *Algorithms Mol Biol* **7**: 5.
- 7) 寺田朋子, 清水謙多郎, 門田幸二 (2020) 次世代シーケンサーデータの解析手法: 第 15 回 RNA-seq 解析 (その 3). *日本乳酸菌学会誌* **31**: 25-34.
- 8) Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, et al. (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* **44**: W3-W10.
- 9) Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol.* **11**: R106.
- 10) Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**: R25.
- 11) Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, et al. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform.* **14**: 671-683.
- 12) Zhuo B, Emerson S, Chang JH, Di Y (2016) Identifying stably expressed genes from multiple RNA-Seq data sets. *PeerJ* **4**: e2791.
- 13) Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, Manno GL, et al. (2015) Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**: 1138-1142.
- 14) Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, et al. (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**: 839-851.
- 15) Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, et al. (2015) Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics* **31**: 3625-30.
- 16) Zhao S, Sun J, Shimizu K, Kadota K (2018) Silhouette scores for arbitrary defined groups in gene expression data and insights into differential expression results. *Biol Proced Online* **20**: 5.
- 17) Lun ATL, Bach K, Marioni JC (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**: 75.
- 18) Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, et al. (2017) SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* **14**: 584-586.
- 19) Yip SH, Wang P, Kocher JA, Sham PC, Wang J (2017) Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* **45**: e179.
- 20) Evans C, Hardin J, Stoebel DM (2018) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform.* **19**: 776-792.
- 21) 坊農秀雅 編 (2019) RNA-Seq データ解析 WET ラボのための鉄板レシピ (実験医学別冊), 羊土社, 東京.

- 22) Lin CY, Lovén, J, Rahl PB, Paranal RM, Burge CB, et al. (2012) Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**: 56-67.
- 23) Nie Z, Hu G, Wei G, Cui K, Yamane A, et al. (2012) c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**: 68-79.
- 24) Lieberman B, Kusi M, Hung CN, Chou CW, He N, et al. (2021) Toward uncharted territory of cellular heterogeneity: advances and applications of single-cell RNA-seq. *J Transl Genet Genom.* **5**: 1-21.

Methods for analyzing next-generation sequencing data XVI. Why are new methods being developed one after the other?

Koji Kadota^{1, 2, 3}, Kentaro Shimizu^{1, 2, 3}

¹ *Interfaculty Initiative in Information Studies, The University of Tokyo.*

² *Graduate School of Agricultural and Life Sciences, The University of Tokyo.*

³ *Collaborative Research Institute for Innovative Microbiology,
The University of Tokyo.*

Abstract

We often hear people say, “There are so many programs out there with the same purpose, I don’t know which one to use.” One possible reason is that there are various evaluation criteria. For example, even if the performance of a method is less than or equal to that of a conventional method, it is still considered valuable if it is much faster than the conventional one. There are many differential expression methods that are more powerful under certain conditions, such as a large (or small) number of replicates in the same group. In reality, however, there are some methods whose value is questionable. This is probably due to the fact that the comparison has been biased toward representative methods. In this paper, we explain why such methods can exist, using our experience in the field of RNA-seq differential expression analysis as an example. Supplementary materials are available online at: http://www.iu.au-tokyo.ac.jp/kadota/r_seq2.html#about_book_JSLAB.