プレスリリース

慶應義塾大学

2022 年 11 月 18 日

報道関係者各位

慶應義塾大学医学部

シングルセル RNA シークエンスデータ間での普遍的な比較解析法を開発
ー疾患解析データの本質を捉えることが可能にー

慶應義塾大学医学部生理学教室の岡野雄士（医学部 5 年生）、加瀬義高助教、岡野栄之教授らの研究グループは、未解決であったシングルセル RNA シークエンス（scRNA-seq）（注 1）のデータ解析の技術的な課題を克服するモデルを開発し、scRNA-seq データ間の普遍的な特徴を抽出することに成功しました。

細胞の遺伝情報を網羅的に収集、解析することのできる RNA-seq（注 2）、特に細胞ひとつひとつの情報を読み取ることのできる scRNA-seq は現在の医学研究において、もはや必須のものとなっています。ただし、急速に発展してきた技術であるが故に、その解析手法にはいくつかの課題があり、それから得られた結果は果たして本当に生物学的な本質を写し出しているのか疑問視されていました。

また、scRNA-seq から細胞種の情報を推定するアノテーション（注 3）という処理は、細胞の形態学的な情報がない scRNA-seq データにおいて最も重要な処理工程でありますが、従来のアノテーション方法では、発現変動遺伝子解析（注 4）によって、データセットに含まれるサンプル間での相対的な比較を行い、有意に発現量が上昇している遺伝子群から恣意的に解釈可能なもの選んで細胞種を推定するため、異なる個体間で共有されている細胞種の「普遍的な特性」を抽出することができていませんでした。さらに、アノテーションは、scRNA-seq データに対して行われるさまざまな示唆的なデータ解析に先立って行われるため、scRNA-seq データ解析の結果が、どれだけ一般化可能な真実を写した結果であるのかが問題となっていました。

本研究では、scRNA-seq データから細胞種の「普遍的な特性」を遺伝子制御ネットワークとして可視化して、そのネットワークの類似性によって細胞特性の近さを評価する指標の開発に成功しました。さらにその指標を体系的な手法でアノテーションに応用することにより、異なる個体由来の scRNA-seq データを効果的に統合することができるようになりました。実際に、複数個体のヒト胎児脳由来 scRNA-seq データにて本研究成果のモデルを実装してみると、従来法と比較し、よりデータ横断的な特徴を反映したアノテーションが可能であることがわかりました。

この成果により、疾患解析などの研究でこれまで行われてきたscRNA-seqで見落とされていた重要かつ本質的な結果が得られることが期待されます。さらに、希少疾患の解析など、複数の個体や複数の研究機関から取得されたサンプルのデータを統合して解析する必要がある研究テーマにおいて、生データに恣意的な加工を加えずに普遍的特性を確認することができるため、さまざまな分野における応用が期待されます。本研究成果は、2022年11月17日11時（米国東部標準時）に国際幹細胞学会公式ジャーナルである*Stem Cell Reports*の特集号に掲載されました。

1. 研究の背景と概要

一細胞生物学（シングルセル・バイオロジー）の技術は生物学研究において観測可能な情報の量と解像度を大きく改善してきました。中でも scRNA-seq は遺伝子発現の状態を一細胞レベルで観察するために用いられます。RNA-seq データは探索的データ解析からモデルの検証に至るまでさまざまな分野のさまざまな用途で分析対象となっていますが、データそのものに細胞種のラベル情報（注 5）が付随していないため、アノテーションによってデータから細胞種を推定することが必要となります。

従来のアノテーション方法では、発現変動遺伝子解析によって、データセットに含まれるサンプル間での相対的な比較を行い、有意に発現量が上昇している遺伝子群から細胞種を推定するため、サンプル特性や個体差の影響を強く受けるというデメリットがありました。また、細胞種の概念は古典的に形態と機能によって定義されてきたため、形態と機能情報がない scRNA-seq において、マーカー遺伝子（注 6）の発現量と細胞種の対応関係が成立しているかどうかは明らかになっていません。そもそも得られる遺伝子発現量の数値は、サンプルの特性や定量手法、データ処理方法などの影響を受けて変動するので、アノテーションの結果が古典的な定義での細胞種として、どの程度妥当であるか可視化することは困難でした。さらに、異なるデータセットにおいて同じ細胞種としてアノテーションされた細胞群がどの程度同一の集団として扱えるのかという点も不透明でした。

それらの課題を克服すべく、本研究では、データセット内での相対的な発現量の比較とは異なったフレームワークで細胞群の特色を評価するための手法と背景理論の構築を行いました。細胞の形態やマーカーを用いた細胞種の古典的な定義をヒントに、遺伝子発現の統計学的従属性を RNA-seq データで見られる細胞の機能として捉え、それらの関係を遺伝子制御ネットワーク（Gene Regulatory Network : GRN）として可視化し、GRN の類似性によって細胞特性の近さを評価する指標を開発しました。GRN の設計に関して、データには現れないような生物学的に既知の情報を実験者が事前にモデルに組み込むことで適切な解像度での比較が行えるように、GRN に組み込む遺伝子を自由に選択できるような設計をしました。この際に、特定の遺伝子を選択する操作がもたらす影響についても、ネットワーク構造への親和性が高い代数構造であるトロピカル半環（注 7）を用いて説明しました。また、RNA-seq のデータ解析において同一視される細胞群（細胞のクラスター：注 8）について、その構成方法によっては単一細胞と同様の評価を行うことが困難であることを示し、GRN を構成するアルゴリズムの特性を踏まえた上で、細胞群同士の比較においては細胞特性の近さは半擬距離空間（注 9）で表現されることを示しました。

さらに、本研究で細胞群の類似性の比較尺度をアノテーションに応用し、従来法との比較を先行研究によって取得されたオープンソースの scRNA-seq データに対して実施したところ、従来法よりもデータ横断的な特徴を反映したアノテーションに成功しました。

具体的には、アノテーション結果が付随した豊富なサンプル数を有するヒト成人脳由来 scRNA-seq データを、生物学的に自然な細胞種の分類を反映していることを期待して参照データとし、胎生時期の異なる 4 個体からなるヒト胎児脳由来 scRNA-seq データに含まれるサンプルが、興奮性神経細胞・抑制性神経細胞・神経細胞以外の細胞のいずれかに帰属するようなアノテーションを行いました。これにより、本研究で開発された比較尺度は任意の構成方法で区分された細胞集団に対して適応することができたことから、中枢神経系の細胞に関する情報を効果的に分類に組み込むために、先行研究で明らかにされた中枢神経系マーカー遺伝子群に対する因子解析（注 10）を行い、因子得点が形成する座標空間上での分布をもとに細胞群の分割を行いました。

また、参照データに付随したアノテーション情報を教師データ（注 11）として、勾配ブースティ

ング決定木モデル（注 12）を作成することで重要な特徴量を抽出し、実験者による選択バイアスの影響を緩和する実装を行いました。このようにして決定された細胞集団に対して、選択された遺伝子群による GRN を算出し、類似性を参照データとクエリデータ間で比較することによって、半疑距離関数の値を最小化するラベルを付随することでアノテーションを行いました。このような手法で実施されたアノテーションの結果は、異なるデータセット間での結果の生物学的意味づけ（細胞特性の共通点や相違点）を GRN の構造によって可視化できる点や、アノテーションにおいて最も決定的なプロセスである、細胞群に名称を付随する際の判断が擬距離関数によって客観的に与えられる点がメリットとして示されました。

2. 研究の成果と意義・今後の展開

今回の研究では、細胞種の古典的定義方法とRNA-seqデータ解析における細胞腫の決定の仕方のミスマッチを指摘しました。また、RNA-seqデータにおいて普遍的に観測可能な細胞機能の特性を、発現量の数値そのものとは異なった方法で表現するための理論構築を古典的定義方法に着想を得た方法で行いました。さらに、そのように新たに定式化された細胞特性の表現方法に、古典的細胞種に対して蓄積された生物学的知見を自然に導入する手法を開発しました。

細胞の「普遍的な特性」を表現することは、分析対象がどれだけ自然で一般化された集団であるかを確認する上で力を発揮するため、本研究で用いた、胎生時期の異なる4個体からなるヒト胎児脳由来scRNA-seq データのように、モデル動物や培養細胞といった、妥当性を事前に検証すべき発生学的研究や病態解析研究などの解析において有用です。

また、RNA-seqデータの時系列的解析や、希少疾患の解析など、複数の個体や複数の研究機関から取得されたサンプルのデータを統合して解析する必要がある研究テーマにおいて、生データに恣意的な加工を加えずに普遍的特性を確認することができるため、さまざまな分野における応用が期待されます。

今後の展望として、本研究の成果を疾患解析へ応用し、治療ターゲットの発見に役立てていくと共に、学術的には、本研究で集合論的に定式化された細胞集団とその特性の近さについて、代数的位相幾何学などの観点からさらなる考察を行なっていくことを検討しています。

3. 特記事項

本研究は、坂口光洋記念慶應義塾大学医学振興基金、日本損害保険協会交通事故医療特定研究助成、武田科学振興財団、JSPS 科研費 JP22K16696、Keio University Yagami Data Security Lab の支援によって行われました。

4. 論文

英文タイトル：A set-theoretic definition of cell types with an algebraic structure on gene regulatory networks and application in annotation of RNA-seq data

タイトル和訳：遺伝子制御ネットワーク上の代数構造による細胞種の集合論的定義と RNA-seq データのアノテーションへの応用

著者名：岡野雄士、加瀬義高、岡野栄之

掲載誌：*Stem Cell Reports*（特集号）

DOI：10.1016/j.stemcr.2022.10.015

【用語解説】

（注 1）シングルセル RNA シークエンス：検体の細胞を一細胞ごとに分けてから RNA-seq を行う手法。

（注 2）RNA-seq：細胞などの遺伝子発現量を転写レベルで網羅的に解析する手法。

（注 3）アノテーション：データが何を示しているのか分類し、命名する処理を指す。

（注 4）発現変動遺伝子解析：複数の細胞集団の間で発現している遺伝子の違いを網羅的に探索する手法。

（注 5）ラベル情報：データの取得された条件など、データに対応した情報のこと。

（注 6）マーカー遺伝子：特定の細胞種において特徴的に発現しているとされている遺伝子のこと。

（注 7）トロピカル半環：整数や実数などの集合と正の無限大の和集合に対して導入される代数構造で、加法は二項の最小値、乗法は二項の和として定義される。別名、min-plus 代数としても知られている。

（注 8）クラスター：データ上で近い値を持ったサンプルの集団のこと。

（注 9）半擬距離空間：距離関数の性質の一部を失った半擬距離関数が成立する空間を示す。半擬距離関数は距離関数と同様に、0 以上の値を示し、同一の点に対しては 0 を返し、三角不等式が成立するが、非対称関数であり（例：点 x から点 y までの半擬距離が点 y から点 x までの半擬距離と一致しない）、また同一でない 2 点間に対しても 0 を返す場合がある。

（注 10）因子解析：データで表現されている現象が、背景では少数の因子の影響を受けて決定されていることを仮定して、データを少数の潜在変数に分解するデータ解析手法。

（注 11）教師データ：機械学習モデルにおける予測対象の正しい値。予測結果と照らし合わせることでモデルの性能を評価することにも用いられる。

（注 12）勾配ブースティング決定木モデル：機械学習モデルの一つ。決定木と呼ばれる機械学習モデルを効果的に複数組み合わせることで精度を向上させることができるとされている。

※ご取材の際には、事前に下記までご一報くださいますようお願い申し上げます。

※本リリースは文部科学記者会、科学記者会、厚生労働記者会、厚生日比谷クラブ、各社科学部等に送信しております。

【本発表資料のお問い合わせ先】

慶應義塾大学医学部　生理学教室

教授　岡野　栄之（おかの　ひでゆき）

〒160-8582 東京都新宿区信濃町 35

TEL：03-5363-3747　FAX：03-3357-5445　E-mail：hidokano@keio.jp

【本リリースの配信元】

慶應義塾大学信濃町キャンパス総務課：山崎・飯塚・奈良

〒160-8582 東京都新宿区信濃町 35

TEL：03-5363-3611　FAX：03-5363-3612　E-mail：med-koho@adst.keio.ac.jp

http://www.med.keio.ac.jp