様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ５ 月 ２８ 日現在

機関番号：８２４０１
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６１５２

研究課題名（和文）1細胞RNA-Seqデータ内に含まれる細胞型を特定する解析手法の確立

研究課題名（英文）Establishment of the analysis method to identify cell type contained in single-cell RNA sequencing data

研究代表者

露崎　弘毅（Tsuyuzaki, Koki）

国立研究開発法人理化学研究所・生命機能科学研究センター・特別研究員

研究者番号：７０７６９５２０

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では、細胞集団に含まれる細胞型（Celltype）を検出する汎用的なデータ解析手法を確立する。近年、1細胞レベルでの全遺伝子発現量を計測する1細胞RNA-Seq法が登場し、細胞の不均質性や異質性を精度良く計測できるようになった。しかし、1細胞RNA-Seqデータ中である細胞が何者なのかを特定するのは困難である。なぜなら混入している細胞の種類は事前にわからないことが多いからである。そこで本研究では1細胞RNA-Seqデータから細胞型を特定する事を、"Celltyping"と呼び、そのためのデータ解析手法を確立した。

研究成果の学術的意義や社会的意義
1細胞RNA-Seqデータ解析において、Celltypingで決定された細胞型ラベルは、その下流のあらゆるデータ解析で利用されるため、Celltypingはその後の研究の進退に関わる重要な解析ステップである。しかしながら、現状のCelltypingの方法は時間がかかり、かつ解析者の事前知識に依存する主観的な作業であった。そのため、このCelltypingを自動的、かつ客観的に行えるための方法論を構築した。本研究の提案手法によるCelltypingは、再生医療・創薬、発生生物学、疾患関連細胞など、1細胞RNA-Seqを利用した生命科学研究全てに貢献する。

研究成果の概要（英文）：In this study, we will establish a versatile data analysis method to detect cell types contained in cell populations. In recent years, the single-cell RNA-Seq has been introduced to measure the gene expression level at the single-cell level, and it has become possible to accurately measure the heterogeneity of the cell population. However, it is difficult to identify which data point of single-cell RNA-Seq data corresponds what kind of cell type because the cell type is often unknown in advance. Therefore, in this study, we try to establish a data analysis method to identify the cell type of single-cell RNA-Seq data.

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　生命情報学　single-cell RNA-Seq　情報検索　次世代シーケンサー　1細胞RNA-Seq　機械学習　オミックス

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

本研究では、細胞集団に含まれる細胞型（Celltype）を検出する汎用的なデータ解析手法を確立する。細胞分取技術と DNA シーケンサーの高出力化が進み、近年、1 細胞レベルでの全遺伝子発現量を計測する 1 細胞 RNA-Seq 法が登場し、細胞の不均質性や異質性を精度良く計測できるようになった。これにより、様々な生命科学のより高解像度な理解が期待できる。しかし、1 細胞 RNA-Seq データに含まれる各データ点が、どの細胞の種類に対応するのかを特定するのは困難である。なぜなら混入している細胞の種類は事前にわからないことが多いからである。

２．研究の目的

(1) 本研究では 1 細胞 RNA-Seq データから細胞型を特定する事を、"Celltyping"と呼び、そのためのデータ解析手法の確立を試みた。Celltyping を行う一般的な解析方法としては、次元圧縮、クラスタリングなどの教師なし学習をデータに適用し、検出されたクラスターごとの発現変動遺伝子や変動した機能ターム・パスウェイ、既知のマーカー遺伝子の発現量を確認することで、クラスター単位で、何の細胞型に対応するのかを解析者が手作業でアノテーションしていくやり方である。Celltyping で決定された細胞型ラベルは、その下流のあらゆるデータ解析で利用されるため、Celltyping はその後の研究の進退に関わる重要な解析ステップである。

(2) しかしながら、上記のような方法は時間がかかり、かつ解析者の事前知識に大きく依存する主観的な作業であった。そのため、本研究では、この Celltyping を自動的、かつ客観的に行えるための方法論の開発に着手した。

３．研究の方法

(1) 既に Celltyping が行われた既報の 1 細胞 RNA-Seq データは、GEO、ArrayExpress などの公共データベースに生データ（FASTQ ファイル）として蓄積されつつある。また、近年では、Single Cell Portal など、1 細胞 RNA-Seq データの解析結果や解析済みファイルを公開した 2 次データベースも開発されつつある。そのため、手元のデータ（クエリ）の Celltyping　　に、このようなデータベース上のデータを利活用することを考えた。すなわち、クエリとデータベース間で発現プロファイルの類似度計算を行い、類似度が高い場合に、データベース上のデータの細胞型ラベルを、クエリに割り当てるのである。このアプローチにより、Celltyping の作業は自動化される。

(2) ただし、このアプローチには、以下の 2 つの問題があった。一つ目は、異なる実験や研究組織によって計測されたデータ同士には、例え同じ細胞を計測したものであっても、人為的な違いが出てしまう"バッチエフェクト"が見られることである。特に、1 細胞 RNA-Seq データは、Smart-Seq、Quartz-Seq など、異なる実験手法で計測されたデータが多く、それらデータ間で、細胞型が同じデータ同士の類似度が高くなるのかは自明ではなかった。

(3) 二つ目は、データベースのサイズが膨大な場合における、計算量・メモリ使用量である。近年 Human Cell Atlas や Mouse Cell Atlas、Tabula Muris といった、アトラス級のデータが次々と公開されていく中で、類似度検索がどの程度現実的な時間で終わらない恐れがあった。そのため、本研究では、これらの問題を解決するための方法論を開発した。

４．研究成果

(1) バッチエフェクトに関しては、データベース側のデータに主成分分析（PCA）を適用し、データが大きくばらつく低次元空間上の方向（固有ベクトル）を計算し、クエリデータもこの低次元空間に射影することで、経験的に生物学的な遺伝子変動のみに着目し、バッチ間で変動するような人為的な変動は回避できることがわかった。この方法は、情報検索の分野では Folding-in 呼ばれる方法で、近年オミックス解析においても Reference Component Analysis（RCA）という名前で利用されるようになった。この手法のもう一つのメリットとしては、あらかじめデータベース側で PCA をしておくことで、類似度検索時に時間のかかる PCA の計算をやらなくて済み、またデータの次元が圧縮されるため、その後の検索速度が向上する点である。また検索速度に関しては他にも、Locality Sensitive Hashing（LSH）を利用して、高速にデータベース・クエリ間のビット列間の類似度を計算する検索ツールを共同研究者の佐藤建太氏が実装したため、これを利用した。

(2) テストデータである、Tabula Muris などのアトラス級大規模 1 細胞 RNA-Seq データに、提案手法を適用したところ、競合手法の scmap と比較しても、より高精度・かつ高速・低メモリで類似度検索が行えることがわかった。また類似したデータベース・クエリ間のペアでのわずかな発現プロファイルの違いに着目する機能も追加したため、これにより健常者-疾患患者間の細胞状態の違いなどにも応用できる可能性が示唆された。開発したソフトウェアは CellFishing.jl という Julia 言語のパッケージとして公開されており、あらゆる研究者が自由に利用可能である（https://github.com/bicycle1885/CellFishing.jl）。

(3) 今後の課題としては、アトラスレベルのデータベースが今後より一層増えることで、計算機のメモリにデータベースの全データが乗り切らない場合を想定し、CellFishing.jl の内部でも利用されている PCA を、より高速・低メモリな実装に変更することを検討している。既に、どのような PCA のアルゴリズム・実装が精度を落とさず、かつスケーラブルに計算できるのかをベンチマークしており、結果を論文に投稿中である（https://doi.org/10.1101/642595）。

(4) また、上記の既報のデータベースをリファレンスとしたアプローチは、クエリにこれまで計測されたことがない細胞型が含まれている場合に、取りこぼす可能性がある。そのような場合、手元のデータに、既存の知識をどのように割り当てるのかということが重要であることから（https://arxiv.org/abs/1712.08865）、解析者が細胞型の違いに気付きやすく、かつ解釈しやすい情報を提示するインターフェースを構築しようと考えている。具体的には、細胞間の相互作用や（https://doi.org/10.1101/566182）、事前の教師なし学習に依存せずに、機能ターム・パスウェイレベルでの発現変動に注目するソフトウェアを開発することで、リファレンスを使う場合、使わない場合の両方状況に対応していく。

(5) 本研究の提案手法による Celltyping は、1 細胞 RNA-Seq を利用した生命科学研究全てに貢献するものであり、再生医療・創薬（例 ： 幹細胞・ニッチ細胞の同定、オルガノイド内の異質性）、発生生物学（例 ： 分化状態の異なる細胞の同定）、疾患関連細胞（例 ：Cancer Associate Fibroblast）など、細胞型に注目するあらゆる研究に今後適用可能である。

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① Kenta Sato, Koki Tsuyuzaki, Kentaro Shimizu and Itoshi Nikaido, CellFishing.jl: an ultrafast and scalable cell search method for single-cell RNA-sequencing. BMC Genome Biology (peer-reviewed), 2019
② Koki Tsuyuzaki and Itoshi Nikaido, Biological Systems as Heterogeneous Information Networks: A Mini-review and Perspectives. WSDM2018 HeteroNAM'18 (peer-reviewed), 2018

〔学会発表〕（計 3 件）

① Koki Tsuyuzaki and Itoshi Nikaido, Biological Systems as Heterogeneous Information Networks: A Mini-review and Perspectives. WSDM2018 HeteroNAM'18, 2018
② 露崎弘毅、「データベースとデータ解析の融合～なぜデータベースは必要か～」、ワークショップ（1AW01）いかにして「使える」データベースを維持し続けるか、2017 年度生命科学系学会合同年次大会（招待講演）、2017
③ 露崎弘毅、「Heterogeneous Information Networks / Meta-path の紹介」、第 6 回生命医薬情報学連合大会、2017

〔図書〕（計 3 件）

① 露崎弘毅、二階堂愛、羊土社、シングルセル解析 実験ガイド（実験医学別冊）、第 2 章シングルセル解析の実際、1 細胞 RNA-Seq データ解析、2017
② 露崎弘毅、学研プラス、次世代シーケンサーDRY 解析教本（細胞工学別冊）、Level1（準備編）、R の使い方、2015
③ 露崎弘毅、二階堂愛、学研プラス、次世代シーケンサーDRY 解析教本（細胞工学別冊）、Level3（論文別・作図コマンド解説）、R + biomaRt + reshape2 + ggplot2 + grid + entropy、2015

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計 0 件）

名称：

発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ ： https://researchmap.jp/kokitsuyuzaki/

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：
ローマ字氏名：