「公共データベースからシングルセルRNA-seqデータを取得する」（講師：理化学研究所 粕川 雄也氏）質疑応答
AJACS「シングルセルRNA-seqを知って・学んで・使う」（2024年12月23日開催）

No	質問	回答
1	ウエットの研究者がシングルセルデータの解析を行う現実的な方法（道筋）はどのようなものがあるでしょうか？ ■本日の発表を聞くと、ウエットの研究者が一からデータ解析をするハードルは高いと感じました ■データ解析を行ってくれる共同研究者を見つけるのが難しい場合もあります	今回の発表が比較的解析の応用編的な内容になっていたため、難しいと感じられたようにも思えます。しかし一細胞RNA-seqの解析の基礎的な部分についてはかなりツールが整備されており、10x Genomics社のプロトコルでは、CellRangerというソフトウェアを使うことで、シーケンサーの出力（FASTQファイル）から品質評価、発現プロファイルの作成までの一連の処理を一度に実行することができるようになっていたり、Seurat、scanpy などのフリーのパッケージを使うことでクラスタリング、マーカー遺伝子探索やその後の解析手法をまとめて処理できるようになっており、いわゆるバルクRNA-seqよりも簡単に解析できるようになっています。実際にハードルになるところとしては、（1）データ解析環境の構築、（2）品質評価法や解析手法の選択（例えば、どのクラスタリング方法がいいのか）、（3）論文化後の最終的なデータ登録方法になるかと思います。こういったところをドライの共同研究者にやってもらえると助かるのですが、実際にそういった共同研究者を探すのは難しいのが現状です（こういう解析ができるドライ解析者は基本的に忙しいというのもあります）。現実的な解決方法としては、ラボ内でコンピューターの操作や統計処理に明るい方に勉強して習得してもらう、ドライの研究者に相談できるようにしておく（可能なら誰かが習いに行く）、ドライ解析もセットになっているような施設にシーケンシングからお願いする、などが行われているのではないかと思います。
2	UMAP等クラスタリングの公開データの再現が、ソフトのバージョンなどもあり完全にはできないこともあるかと思うのですが、そう言った場合に再現できているという基準がぜひあれば伺いたいです	本来はMethodに書かれている通りに再解析すれば同じ結果になるべきではありますが、実際は難しいのではないかと思います。バージョンの違いもありますし、解析に乱数を使うアルゴリズムを使っている場合は、乱数列も揃えないと同じ結果にならないなどが主な理由になるかと思います。おそらく元論文の結論と矛盾しない結果が出ていることが再現と考えるのが限界のようにも思います。ただ最近は論文の法に図の元データやスクリプトを添付することを求める風潮も出てきていますので、そちらがあればそれを見てみたり、再解析してみたりする方法も今後はできるようになるかと期待します。
3	「公共データをマージして利用したい」という質問に関する疑問です。「異なるサンプルで取られた細胞ごとのデータをマージする」ということに、どういう目的があるのか、またそのようにしてマージされた異なるサンプル由来のシングルセルデータがどういう意味を持つのでしょうか。	マージ（integration）したいケースというのはいくつか考えられ、同じ細胞種が含まれる別実験の結果を合わせて観測・解析対象となる細胞数を増やしたい場合、健常者由来のサンプルと患者由来のサンプルを合わせて疾患時の一細胞レベルに遺伝子発現の差を見たい場合、様々な細胞種をカバーしたリファレンスセットを作りたい場合等があります。適切に integartion された結果は単一の場合に比べてより多様な解析ができる可能性があり、手法開発もホットに行われている分野です。
4	シングルセルはデータサイズが大きいので、ダウンロードの際にどのように工夫されているか伺えたら幸いです。	ダウンロードスピードについては、普通にダウンロードしても時間がかかりそうな場合はいくつか代替法があります。講義でも説明した fasterq-dump というコマンドはファイルを分割してダウンロードすることで速度アップを図ります。またNCBIであれば Aspera という商用の高速ダウンロードソフトが使えるようになっていたりします。またダウンロード用のミラーサイトを用意している場合もあり、例えば Amazon Web Service（AWS）のS3オブジェクトからダウンロードできたりもします。またダウンロード時のマナーとして各データベースサイトで制限をつけている場合がありますので（例えば1秒間の接続数を3件以内するように等）、注意してダウンロードしてください。参考: https://www.ncbi.nlm.nih.gov/books/NBK25497/
5	もし公開データのコードなどにミスがあった場合、論文は訂正になるのでしょうか？ そういったことを避けるためにできる対策をお伺いしたいです	コードに誤りがあるような場合は、errata/correction を出して修正したり、論文の結論が変わってしまう場合は retraction するのが理想です。ただ、いざ間違っていてもそこまで対応するかどうかは著者の良心に任されてしまっているのが現状でしょうか。ミスを避けるためには、コードをしっかり確認することが基本だと思いますが、可能なら別の人にレビューしてもらったり解析の再現をしてもらうなどすれば可能性は減らせるかもしれません。
6	細胞のアノテーションについてです。どの程度に細かくクラスタリングされた状態でアノテーションをおこなうものでしょうか。	ご質問ありがとうございます。大切な点であると思います。データ解析を行う際は、研究目的も含め、データの中身を理解していることが前提となります。クラスタリング粒度を決定する決め手となるのは、その結果が生物学的に示唆的であるか、という点につきるかと思います。
7	「公共データをマージして利用したい」という点に関連して質問です。マージ可能かどうかの判断基準について伺いたく思います。例えば、異なるリード数・リード量・リード長のデータはマージは非推奨なのでしょうか？	マージ（integration）の条件ですが「これこれ以下であるべき」といったカットオフ条件のようなものはないように思います。一般的な傾向として integartionしようとしているデータセット間のプロトコル条件が異なるほどintegration は難しくなること、パラメータ設定やそもそも手法の傾向として、より強度に integration してしまうと、実際には違うものがあたかも類似した細胞のように間違って判断してしまう（アーティファクトの）可能性が出てくる点は注意が必要だと思います。

No	質問	回答
8	3’、5’など異なるプラットフォームで得られたデータをマージするのは問題ないのでしょうか？	個人的にはかなり難しいような印象もありお薦めはしないのですが、それでもマージ（integration）するような方法がない訳ではないので、実際のところはやってみて問題ないか評価するのが一般的かなと思います。ただ無理矢理 integration してもそれが本当に同じ細胞が近傍にくるようなクラスタリング結果・次元圧縮結果になっているのか（アーティファクトではないのか）といった点は注意をしておく必要があります。ちなみに、10x Genomics 社のサイトにも同じ質問がありましたが、どちらとも判断していないようでした。 参考：Can I combine gene expression data from 3’ and 5’ assay chemistries?	10X Genomics https://kb.10xgenomics.com/hc/en-us/articles/115003145272-Can-I-combine-gene-expression-data-from-3-and-5-assay-chemistries
9	公共データをマージして利用したい場面が多いです。ダウンロードする前に使える/使えないを判断する材料があれば教えてください。またどのサイトがおススメでしょうか？	実際に試してみないと本当にマージ（Integrationということが多いようです）できるか分からないのですが、プロトコルがまったく異なる等明らかに integration できないようなものは避ける、QC結果が提供されているような場合はそれを見てあらかじめ判断する、等ぐらいではないかと思います。
10	公共データから使用するにあたって、Quality Controlの方法が非常に重要になってくると思います。コンセンサスの得られている方法がありましたら教えて下さい。私は、Mt遺伝子の割合・発現遺伝子数に基づくフィルタリングと、doublet/emptyの除去を行なっていますが、十分でしょうか。	コンセンサスの得られた方法はないように思います。Seuratパッケージで使われるデフォルトのフィルタリング条件が結果的に使われていることが多いような印象もあります（少なくともその条件であれば論文等でレビュアーから変な指摘をされない安心感？）ちなみに我々もSkewCというQuality Controlチェックツールを自作していていますし、他にも色々提案されています。色々試すほかないというのが実情かと思います。
11	Single Cell Expression Atlasを使っていますが，公共DBを選ぶときにどこを見るか（Human Cell Atlasで生データの信頼性を担保するものは何か？）を教えてください	公共データベースに登録されているデータの信頼性については、基本的に自分でとったデータと同じような品質確認プロセスを行って評価するのが一般的だと思います。公共データベースを使うときはどのようなデータの収集や品質確認処理を行っているのかを今一度確認してみて自分が欲しい品質のデータがどの程度手に入るのかを見るのがいいように思います。例えば INSDCは論文発表に登録が必須とされている一次データベースである関係で、かなり多くのデータが手に入る一方、メタデータは登録者の手間に依存していたり品質の確認をしていない（データの整合性のみチェックする）ため、欲しい品質のデータが必ずしも手に入りません。Single-cell Expression Atlas等はINSDCのデータを再処理した二次データベースで、データ自体はINSDCのもののサブセットである一方、何らかの品質確認と選択を行っているため、提供されているデータの品質は高めと考えられます。こういった特性をよく知ったうえで使うデータベースを決めるのが重要であるように思います。