シングルセル解析を用いた細胞状態遷移初期情報の検出手法の開発

Development of a method for detecting initial information on cell state transitions using single cell analysis

大坪 拓帆[1]・小倉 淳[1,2,3]

[1]長浜バイオ大学バイオサイエンス学部バイオサイエンス学科

[2]長浜バイオ大学バイオサイエンス学部アニマルバイオサイエンス学科

[3]長浜バイオ大学ゲノム編集研究所

Takuho Ohtsubo[1], Atsushi Ogura[123]

[1]Department of Bio-Science, Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology

[2]Department of Animal Bio-Science, Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology

[3]Genome Editing Research Institute, Nagahama Institute of Bio-Science and Technology

要旨

疾患を早期に発見するためには、正常細胞が炎症状態に移行する前の未病細胞を識別することが重要である。また、細胞分化の初期状態を検出することは、さまざまな発生メカニズムの解明にも重要である。そこで、近年、有用性が明らかになったシングルセルシーケンス解析をもとに、細胞状態遷移初期情報を検出する手法の開発をおこなった。まず、全トランスクリプトーム解析に適した解析パイプラインを構築した。次に、疾患進行に関するテストデータを用いた統合解析を実行した。まず細胞タイプ分類として、B 細胞と T 細胞、肺胞上皮組織 AT1、肺胞上皮組織 AT2 および M1 マクロファージと M2 マクロファージの分類が可能であった。疾患初期状態特異的遺伝子として、Mrvi1 と Ldb3 と Dysf と Plin4 が見つかった。これらの結果から、細胞状態遷移初期情報を抽出できることがわかった。

For early detection of disease, it is important to identify unaffected cells before normal cells transition to an inflammatory state. Detecting the initial state of cell differentiation is also important for elucidating various developmental mechanisms. Therefore, we developed a method to detect the initial information of cell state transition based on single cell sequencing analysis, which has recently shown its usefulness. First, we established an analysis pipeline suitable for whole transcriptome analysis. Next, we performed an integrated analysis using test data on disease progression. First, we were able to classify cell types as B and T cells, alveolar epithelial tissue AT1, alveolar epithelial tissue AT2, and M1 and M2 macrophages. Mrvi1 and Ldb3, Dysf and Plin4 were found as early disease state specific genes. These results indicate that it is possible to extract information on the initial cell state transition.

１．はじめに

シングルセル解析は、2013 年に Nature 誌の Technology of the year に掲載され、近年では毎週のよう

にその技術を用いた論文が nature や science に掲載されるほど注目を浴びる技術である。シングルセル解析では、ある組織の中で新たな特徴を持つ細胞の同定とその特徴を決定する遺伝子を解析可能である。組織単位での解析では解明できなかった細胞レベルでの分子メカニズムを解明する重要な技術だ。例えばガン細胞のシングルセル解析では、ガン細胞を摘出し、一細胞ごとに mRNA にセルバーコードを付加する。その後、次世代シークエンサーを用いてシングルセル発現データの形にする。既知の遺伝子の役割からガン細胞のより詳細な分類分けを行う。疾患を早期に発見するためには、正常細胞が炎症状態に移行する前の未病細胞を識別することが重要であり、実際に未病状態の細胞をシングルセル解析により同定しようという試みも多数行われている[1)]。また、シングルセル解析は発生学分野でも非常に強力であり、細胞間の不均一性の理解、分化経路の追跡、特定の対立遺伝子からの遺伝子発現の定量化、細胞系統の追跡など、さまざまな発生メカニズムの解明に重要な技術となっている[2)]。シングルセル解析は、最新技術であり様々な実験・解析手順が存在する[3)]。ビーズと一細胞を一つの液滴として処理する Drop-seq[4)] という手法やビーズの代わりにハイドロゲルを用いる InDrop という手法、マイクロウェル一つ一つに一細胞ずつ滴下して処理する Nx1-Seq[5)] などである。プロトコルに応じて、独自の分子識別子（UMI）使用の有無や長さにも違いがある[6)]。これらの技術は、今後さらなる発展が見込まれている。したがって、シングルセル技術のプロトコルごとに適した解析手法の構築や改良が必要不可欠である。

２．シングルセル解析手法の開発

２．１　一次解析

解析パイプラインは、イルミナシークエンサーから生成された R1 および R2 ペアエンド FASTQ ファイルで動作する。必要な最小リード長は、ライブラリー分子の各末端で 75 bp である。 R1 リードにはセルラベルと分子識別子（=umi）に関する情報が含まれ、R2 リードには遺伝子に関する情報が含まれる。

・ステップ 1.リード品質でのフィルタリング：シーケンスの品質が低いリードペアが最初に削除される。このステップにより、シーケンス品質の低下の影響が軽減される。

・ステップ 2.アダプターのトリミング：cutadapt を用いて、R2 リードから 5prime Tagging Adapters を削除する。

・ステップ 3.ファイルの 16 分割：解析時間を短縮させる目的で、seqkit を用いて指定したディレクトリに fastq ファイルを 16 分割して出力している。以降のステップは、GNU parallel を用いて 16 分割されたファイルに対して並列でコマンドを実行している。

・ステップ 4. R1 リードに注釈付け：フィルター処理された R1 リードを分析して、セルラベルセクションシーケンス（CLS）、共通シーケンス（L）、一意の分子識別子（UMI）シーケンス、および poly（T）テールを識別する。セルラベルの情報は、各 R1 リードに沿って 3 つのセクション（CLS1、CLS2、CLS3）に分かれる。2 つの一般的な配列（L1、L2）は 3 つの CLS の分離のために用いられている。セルラベルは、3 つの CLS の定義済みシーケンスの 1 つに確定した組み合わせによって定義される。最初に、予想される位置である CLS1：位置 1-9、CLS2：位置 22-30、および CLS3：位置 44-52 の 3 つすべての事前に設計された CLS シーケンスの完全一致についてリードをチェックする。完全に一致するリードは保持される。UMI とは CLS3 のすぐ下流にある 8 個のランダムな塩基配列であり、CLS に完全一致または塩基置換がある場合、UMI シーケンスは 53〜60 の位置にある。

・ステップ 5. R2 リードに注釈付け：Bowtie2 を使用して、フィルター処理された R2 リードをリファレンスゲノムである GRCm38 (mm10) にマッピングする。
・ステップ 6. R1 および R2 注釈からの情報を 1 つに統合：上記を行ない、有効と判断された R1 リードと R2 リードのペアは、さらなる分析のために保持する。有効な R1 リードには、識別された CLS、N 以外の塩基を持つ UMI シーケンス、およびポリ T テールが必要である。有効な R2 リードには、リファレンスゲノムの遺伝子に一意にマッピングされたリードが必要であり、長さが 60 塩基を超えるアライメントが必要である。
・ステップ 7.分子に注釈を付け：同じセルラベル、同じ UMI シーケンス、および同じ遺伝子を持つリードは、一つにまとめられる。
・ステップ 8.UMI の圧縮：細胞内の遺伝子の分子数を過剰に見積もらないために、 分子カウントに対する UMI エラーの影響を除去する。置換エラーである UMI エラーは識別され、親 UMI バーコードに合わせて調整される。類似配列を持つ UMI を 1 つに圧縮し、正確な UMI 数とする。
・ステップ 9.ニープロット：シーケンスデータを取得し検出されたセルバーコードを抽出しても、これらすべてが利用できるセルラベルではない。一部のセルラベルにはシーケンスエラーまたは PCR エラーが含まれる。さらに、一部のセルラベルは、細胞に由来しない（=エラー）場合がある。したがって、利用できる細胞数を確認する必要がある。変曲点、つまり「膝」を探す方法を用いている。変曲点の下のほとんどの「細胞」は実際には細胞を含まず、単に「周囲」RNA を含む液滴である。今回は、ニープロットを用いて真の細胞数を推定し、利用できない細胞をフィルタリングしている。

２．２　二次解析

二次解析の主な内容は、転写プロファイルに基づく細胞タイプの新規発見と注釈付けである。これは教師なしクラスタリングに相当する。今回は、未病クラスタを検知するため、タイムコース全てを用いてのデータセットの統合解析を実行した。次元圧縮解析に関しては、教師なし学習の一つであり、多様体学習によるデータの可視化が可能な UMAP を用いた。クラスタリング手法に関しては、Louvain アルゴリズムを用いた。

３．テストデータを利用した結果

３．１　データについて

マウスの一個体に対してブレオマイシン投与による肺線維症誘導を行なった後 0, 3, 7, 10, 14, 21 日後のタイムコースで BD Rhapsody システムを用いてライブラリを作成し[7]、シーケンスした生リードを用いた。深麻酔下での口からの吸引により、ブレオマイシン硫酸塩（2 mg / kg を滅菌生理食塩水 50 μ L に溶解したもの）を単回投与した。マウス肺細胞は、ブレオマイシン投与後 0、3、7、10、14、および 21 日目各 1 サンプルを分離し、シーケンスを実施した。イルミナ社の Novaseq6000 シーケンサーを使用し、シーケンスサイズは、101 bp である。サンプルのデータ量は、各 200 Gbp ~ 300 Gbp である。

３．２　一次解析の結果

３．２．１　各ステップで品質が確認されたデータに関して

一次解析は、シーケンスされた生リードを二次解析に使用できる細胞ごとの遺伝子の発現マトリクス（cell × gene matrix）の形にするため行なっている。上に示した一次解析におけるステップ 1 から

ステップ 9 までをテストデータに実施し、その過程において品質が有効と確認されたデータに関して、図 1 に示している。

sample name	Day00	Day03	Day07	Day10	Day14	Day21
reads (original)	1,138 M	1,120 M	923M	1,640 M	1,096 M	1,028 M
1.Quality Control	92.69%	92.73%	93.09%	93.32%	90.51%	92.64%
2.Mapping rate	87.11%	89.45%	88.97%	89.63%	88.30%	89.37%
3.cell BC and UMI assignment	74.83%	73.80%	77.36%	78.23%	71.97%	74.57%
4.UMI compression	6.78%	9.95%	8.67%	7.50%	11.81%	11.29%
5.Number of cells used	4048	5513	4427	5499	6250	6882

図 1. 一次解析の結果

３．２．２　細胞クラスタリング

全タイムコース 6 サンプル（=0,3,7,10,14,21day）の時系列統合クラスタリングを実施し、統合ではアンカーシステムを用いた。アンカーは、別々で解析したデータセットの情報を保持した状態で結びつける役割を持つ。アンカーペアごとに、スコアを割り当てた。この時、アンカーペアの距離が近いほど、スコアは高くなる。アンカースコアが低いアンカーは不正なアンカーとして識別されフィルタリングされた。スコアの高いアンカーは、解析した各データセットを結びつけ、同じ座標にプロットされた。アンカーされなかった情報は、統合できないものとして違う座標にプロットされた (図 2)。

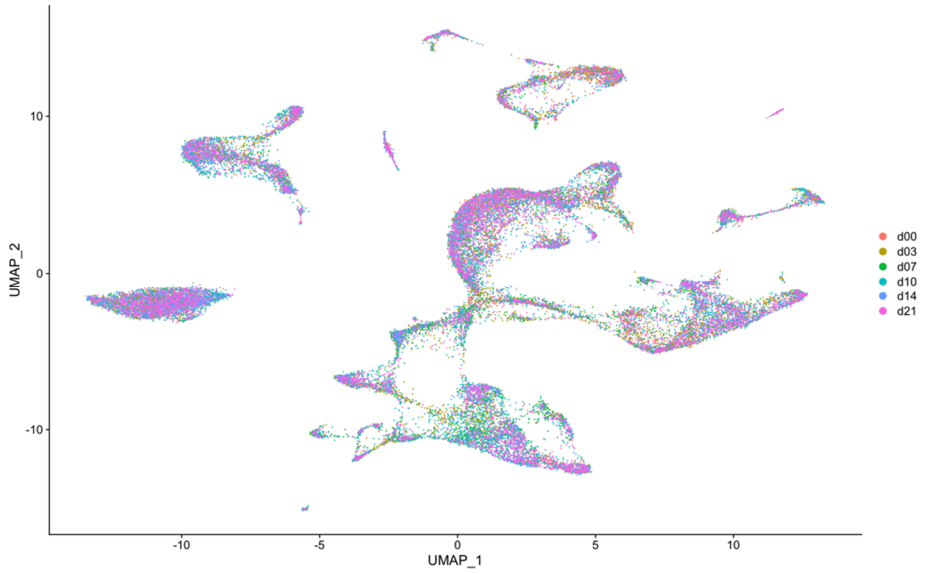
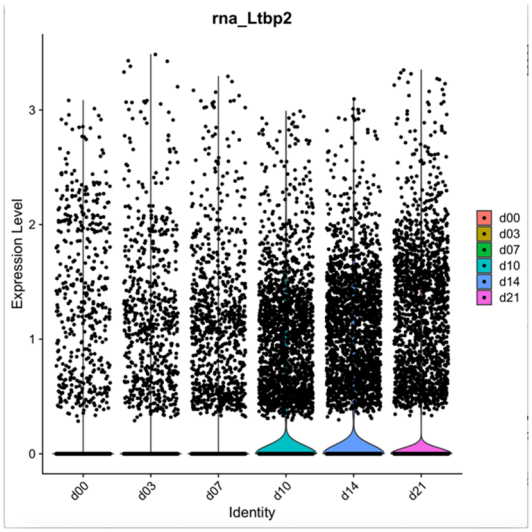


図 2. 細胞クラスタリング

３．２．３　細胞クラスタごとのマーカー遺伝子の検出

肺線維症のマーカー遺伝子である、Ltbp2 のタイムコースでのバイオリンプロット解析を行った。タイムコースの後半ほど、Ltbp2 が多く発現している。ブレオマイシン投与後、日数が経つにつれて、肺



線維症が重症化していることがわかる (図 3)。

図 3. タイムコースごとの Ltbp2 遺伝子の発現細胞および発現量

３．２．４　細胞状態遷移初期に関わるマーカー遺伝子の検出

疾患状態を示す細胞集団のうち、肺繊維症のマーカー遺伝子である Ltbp2 の発現が低い細胞クラスタを探索したところ、肺胞上皮細胞 AT2 クラスタにおいて、疾患が進行していない細胞クラスタを発見した。この細胞クラスタに特異的な遺伝子発現を発現変動解析により解析したところ、Mrvi1 と Ldb3 と Dysf と Plin4 が見つかった。

図 4. Ltbp2 低発現細胞細胞集団

４．結論と考察

LTBP2 未発現細胞における肺線維症初期特異的遺伝子を確認することで、疾患初期状態を示している未病細胞群の推定が可能であった。Ltbp2 未発現細胞における肺線維症初期特異的遺伝子として、Mrvi1、Ldb3、Dysf、Plin4 が見つかった。これらの遺伝子は、正常な肺胞のマーカー遺伝子である可能性も否定できない。そこで、これらが正常な肺胞のマーカー遺伝子であるか確認したところ対応する機能を持つ遺伝子ではなかった。このようなシングルセル解析による細胞状態遷移初期の細胞集団を推定することで、細胞状態遷移の推定に使える新規マーカー遺伝子の推定が可能になることが示唆された。今回の新規マーカー遺伝子の推定には、既知の疾患マーカー遺伝子を用いたが、今後は教師なし学習データからこうしたマーカー遺伝子の推定を行なっていく必要がある。

参考文献

1) Tang, X., Huang, Y., Lei, J. *et al*. The single-cell sequencing: new developments and medical applications. *Cell Biosci* **9,** 53 (2019)
2) Jonathan A Griffiths, Antonio Scialdone, John C Marioni, Using single-cell genomics to understand developmental processes and cell fate decisions, Mol Syst Biol. (2018) 14: e8046
3) ZIEGENHAIN, Christoph, et al. Comparative analysis of single-cell RNA sequencing methods. Molecular cell, 2017, 65.4: 631-643. e4.
4) MACOSKO, Evan Z., et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 2015, 161.5: 1202-1214.
5) HASHIMOTO, Shinichi, et al. Comprehensive single-cell transcriptome analysis reveals heterogeneity in endometrioid adenocarcinoma tissues. Scientific reports, 2017, 7.1: 1-14.
6) SMITH, Tom; HEGER, Andreas; SUDBERY, Ian. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome research, 2017, 27.3: 491-499.
7) BD Single Cell Genomics Bioinformatics(Rhapsody et al., 2018)