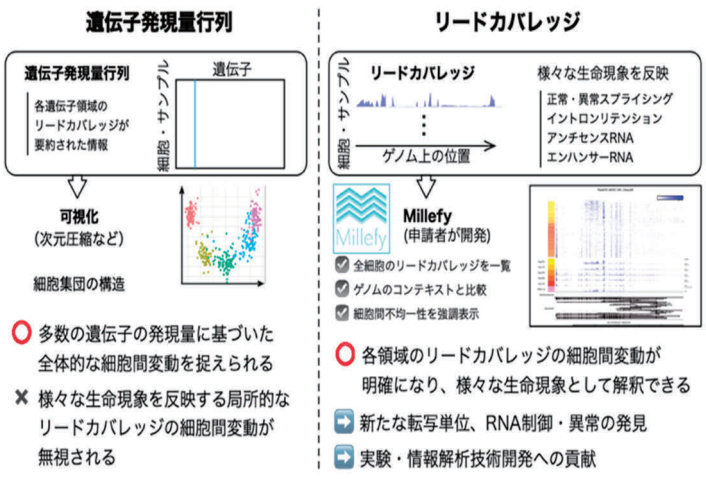
223　がんの細胞間不均一性を発見するリードカバレッジ解析	尾崎　遼

【目的】がんの多くは遺伝的変異により引き起こされるが、これまでバルク RNA-seq（非 1 細胞解像度）を用いた研究から、遺伝子発現量の変動に加え、スプライシング異常や新規転写単位など多様な RNA プロセシングイベントが、浸潤性や転移性といったがんの性質の変化、ステージ判別、予後予測に重要であることが示されている。最近、1 細胞 RNA-seq が、遺伝子発現量に基づいて腫瘍内・腫瘍間不均一性を調べる手段となりつつある。しかし、既存のデータ解析ツールでは遺伝子発現量のみを用いるため、遺伝子発現量に直接現れない異常スプライシングなど RNA レベルの現象が腫瘍や周辺細胞に存在しても見逃してしまうという問題があった。この問題の解決策になるのがリードカバレッジ解析である。リードカバレッジはゲノム上の各領域から転写された RNA 量を表すシグナルの分布であり、多様な RNA プロセシングイベントを反映する。そのため、リードカバレッジを 1 細胞レベルで解析することで、がんの進行や腫瘍不均一性に関連するものの遺伝子発現量解析では見逃される RNA レベルの様々な現象の動態を調べることができる。そこで本研究は、がん 1 細胞 RNA-seq データに対してリードカバレッジ解析を適用する方法論を確立することを目的とした。

【方法】細胞型や亜型、サンプルの性質を考慮しつつ、がん 1 細胞 RNA-seq データからリードカバレッジが局所的に変動する領域（ODEGR ; Overlooked Differentially Expressed Gene Regions）を検出する手法 multiODGERfinder を開発した。本手法ではまず各遺伝子座のリードカバレッジ行列に対して非負値行列因子分解を適用し、リードカバレッジの基本パターンを抽出した。次に、細胞集団（細胞型・細胞亜集団など）間で、抽出されたリードカバレッジの基本パターンのシグナルの値が変動するかを調べるため、シグナルの値に対して一元配置分散分析を適用した。さらに、一元配置分散分析で有意になった遺伝子座に対し、全ての細胞集団ペアの間でシグナルの値が変動するかを t 検定で確認することで、有意差を示す細胞集団ペアを抽出した。既報のトリプルネガティブ乳がん患者由来の 1 細胞 RNA-seq データを用いて ODEGR を人工的に導入したシミュレーションデータを構築し、multiODGERfinder の精度検証を行った。

【結果】ODEGR を人工的に導入したシミュレーションデータに multiODGERfinder を適用することで、リードカバレッジが局所的に変動する領域を抽出できるかを検証した。その結果、multiODGERfinder は ODEGR を有する遺伝子座と ODEGR を有しない遺伝子座を判別するタスクにおいて、AUROC が 0.8 以上という高い精度を示した。さらに、複数のトランスクリプト（アイソフォーム）を含む遺伝子座を判別できるかを検証した結果、0.7 以上の AUROC を示した。このことから、multiODGERfinder によるリードカバレッジ解析が、複数の細胞集団を含むがん 1 細胞 RNA-seq データからの ODEGR の発見に寄与できることが示唆された。

1 細胞 RNA-seq データ解析の 2 つの潮流



223. がんの細胞間不均一性を発見するリードカバレッジ解析

尾崎　遼

筑波大学 医学医療系 生命医科学域 バイオインフォマティクス研究室

Key words：がん、1 細胞 RNA シーケンシング，RNA プロセシング異常，リードカバレッジ解析

緒　言

がんの多くは遺伝的変異に起因するが、これまでバルク RNA-seq（非 1 細胞解像度）を用いた研究から、mRNA などの遺伝子発現量に加え、スプライシング異常、イントロンリテンション、エンハンサーRNA などの新規転写単位といった多様な RNA プロセシングイベントが、浸潤性や転移性といったがんの性質の変化、ステージ判別、予後予測に重要であることが明らかにされている［1〜5］。

最近、1 細胞解像度のトランスクリプトームを計測する 1 細胞 RNA-seq が、遺伝子発現量に基づいて腫瘍内・腫瘍間不均一性を調べる手段としても用いられつつある。しかしながら、既存の 1 細胞 RNA-seq 用解析ツールでは遺伝子発現量のみを用いるため、遺伝子発現量に直接現れない異常スプライシングなどの RNA レベルの現象が腫瘍や周辺細胞に存在しても発見できないという問題があった。このことは、基礎研究面のみならず新規創薬ターゲット候補を見逃すという応用面においても重大な損失となりうる。

遺伝子発現量のみの 1 細胞 RNA-seq データ解析を補完するのが、リードカバレッジ解析である。リードカバレッジはゲノム上の各領域から転写された RNA 量を表すシグナルの分布であり、既知および新規の転写領域、正常・異常スプライシング、イントロンリテンション、UTR 長変化、アンチセンス RNA やエンハンサーRNA などの非コード RNA の転写といった、多様な RNA プロセシングイベントを反映する。そのため、リードカバレッジを 1 細胞レベルで解析することで、がんの進行や腫瘍不均一性に関連するものの遺伝子発現量解析では見逃される RNA レベルの様々な現象のダイナミクスを調べることができる。

国内外では 1 細胞 RNA-seq の遺伝子発現量の解析が盛んに行われている状況であるが、リードカバレッジに着目した研究事例は少ない。既存研究としてがん 1 細胞 RNA-seq でのスプライシング解析などがあるが、スプライシングのみに特化しており、新規アイソフォームや新規転写単位、UTR 長変化などは考慮されていないといった問題があった［6］。筆者は、リードカバレッジの 1 細胞解析に着目して情報解析技術を開発してきた。本研究計画に直接関係する成果として、1 細胞解像度でのエンハンサーRNA や繰り返しスプライシング現象の検出手法の開発［7］、非線形次元圧縮を用いたリードカバレッジの細胞間変動の可視化手法の開発［8］、非負値行列因子分解を用いた二つの細胞集団間でのリードカバレッジの変動検出法の開発［9］（長崎大 松本拡高博士との共同研究）を行なってきた。これらの開発を通じ、これまでトリプルネガティブ乳がん（TNBC）患者由来の 1 細胞 RNA-seq において浸潤性に関連する c-JUN および NSRS 遺伝子の 3’ UTR 領域短縮化現象が細胞間不均一性を示すこと、ヒト iPS 細胞集団と神経前駆細胞集団の間でイントロンリテンションなどのリードカバレッジ変動が起こることなどを見出してきた。

筆者はリードカバレッジ解析というアンバイアスな方法をがん 1 細胞 RNA-seq データ解析に適用することで、事前知識無しにステージや予後、細胞型・亜型に関連した様々な RNA レベルの現象のダイナミクスを発見できるようにする。

そこで本研究は、がん 1 細胞 RNA-seq データに対してリードカバレッジ解析を適用する方法論を確立することを目的とした。特に、がん患者およびがんモデル由来の 1 細胞 RNA-seq に典型的な「複数の細胞型・亜型が内在する複数サンプル間での比較」という実験デザインにおいて、ステージや予後、細胞型・亜型に関連した様々

なリードカバレッジの変動を検出できるようにする。

方　法

1. 1細胞 RNA-seq データの前処理

既報のトリプルネガティブ乳がん患者由来の 1 細胞 RNA-seq データ［10］を用いた。米国 NCBI の Short Read Archive から各細胞のリードデータを FASTQ 形式でダウンロードした。リードのゲノムへのマッピング、発現量定量、BigWig ファイル（リードカバレッジの処理に適したファイル形式）への変換を、ramdaq パイプライン（https://github.com/rikenbit/ramdaq）を使用して実施した。細胞型の定義（B cell、endothelial、epithelial 1~5、macrophage、stroma、T cell）は既報のアノテーションに従った。細胞型が未定義である細胞を除き、1,093 細胞のデータを解析に使用した。

2. 1細胞 RNA-seq データのリードカバレッジ行列の取得

各細胞の BigWig ファイルから、各タンパク質コード遺伝子の遺伝子領域（遺伝子の 5'端から 3'端まで）領域におけるリードカバレッジを取得し、行数が細胞数、列数が遺伝子領域の bin 数（bin 幅は 100 bp）である行列を構成した。これをリードカバレッジ行列と称する。リファレンス遺伝子モデルには GENCODE を用いた。

3. multiODGERfinder

細胞型や亜型、サンプルの性質を考慮しつつ、がん 1 細胞 RNA-seq データからリードカバレッジが局所的に変動する領域（ODEGR : Overlooked Differentially Expressed Gene Regions）を検出する手法 multiODGERfinder を開発した。multiODGERfinder では、多群（3 種類以上の細胞型）間で変動する ODEGR を検出するために、各遺伝子座のリードカバレッジ行列に対して下記の操作を行う：1）リードカバレッジ行列 X に対して非負値行列因子分解　（NMF）を適用し、各遺伝子座のリードカバレッジの基本パターンを抽出する。細胞数を N、遺伝子領域の bin 数を M、基底数を K とすると、NMF は、N×M 非負値行列 X（リードカバレッジ行列）を、N×K 非負値行列 W と K×M の非負値行列 H の積へと分解する。2）細胞集団（細胞型・細胞亜集団など）の間で、抽出されたリードカバレッジの基本パターンのシグナルの値が変動するかを調べるため、W の各列の値に対して一元配置分散分析（ANOVA : Analysis of Variance）を適用する。この際、得られた最大の F 値を FNMF とする。3）並行して、リードカバレッジ行列 X を bin 方向（列方向）に平均し、平均リードカバレッジに対して ANOVA を行い、得られた F 値を Fmean とおく。4）FNMF と FNMF の差ΔF を計算する。5）ΔF の値が正の大きな値を示す遺伝子について、全ての細胞集団ペアの間でシグナルの値が変動するかを t 検定で確認することで、有意差を示す細胞集団ペアを抽出する。以上の過程を、プログラミング言語 R を用いて実装した。

4. シミュレーションデータによる精度検証

リードカバレッジが局所的に変動する領域（ODEGR : Overlooked Differentially Expressed Gene Regions）を人工的に導入したシミュレーションデータを構築した。まず、10,738 遺伝子の遺伝子領域に対応するリードカバレッジ行列それぞれについて、ゲノムの bin 方向にリードカバレッジを平均した。平均リードカバレッジについて実施し、−log10（p-value）について上位 100 遺伝子をリードカバレッジ高変動遺伝子とし、−log10（p-value）＜10 である遺伝子 6,485 遺伝子をリードカバレッジ低変動遺伝子とした。リードカバレッジ低変動遺伝子からランダムに遺伝子を選び、そのリードカバレッジ行列に対し、ランダムに選んだリードカバレッジ高変動遺伝子のリードカバレッジ行列を、領域長を狭めた上で ODEGR として結合し、正例（ODEGR を持つ遺伝子）とした。また、リードカバレッジ低変動遺伝子を負例（ODEGR を持つ遺伝子）とした。正例と負例を識別する性能について、AUROC（area under the receiver operating characteristic）にて評価した。

結果および考察

1. F_{NMF} と F_{mean} の比較

シミュレーションデータのうち、正例（リードカバレッジ低変動遺伝子に ODGER が埋め込まれている）および負例（リードカバレッジ低変動遺伝子）のデータを用い、提案手法の要である ΔF の値（F_{NMF} と F_{mean} の差）が ODEGR を表すのに有効かを検証した。

シミュレーションデータのうち、正例に属する遺伝子のリードカバレッジ行列に NMF を適用し、得られた W の各列について ANOVA を適用し、得られた最大の F 値を F_{NMF} とおく。並行して、正例 ODEGR として埋め込んだリードカバレッジ高変動遺伝子のリードカバレッジ行列を bin 方向に平均し、平均リードカバレッジに対して ANOVA を行い、得られた F 値を F_{mean} とおく。予め埋め込まれた ODEGR が NMF によって適切に抽出されているとすると、F_{NMF} と F_{mean} の値は正の相関を示すと期待される。実際、NMF の基底数が K=2、5、10 のいずれの時にも、それぞれ高いスピアマン相関係数を示した。同様に、シミュレーションデータのうち、正例に属する遺伝子のリードカバレッジ行列から計算された F_{NMF} と F_{mean} の値も K=2、5、10 のいずれの時にも、それぞれ高いスピアマン相関係数を示した（図 1）。これらの結果は、提案手法によって ODGER を抽出できることを示唆する。

さらに、リードカバレッジ行列の細胞（行）がランダムにシャッフルされた行列を作成し、シャッフルされたリードカバレッジ行列に対して NMF を適用し、F_{NMF} を計算した。シャッフルしたリードカバレッジ行列に対し、F_{NMF} はほぼ 0 の値を示した。この結果は、観測された F_{NMF} の値は偶然の結果でない可能性を示している。

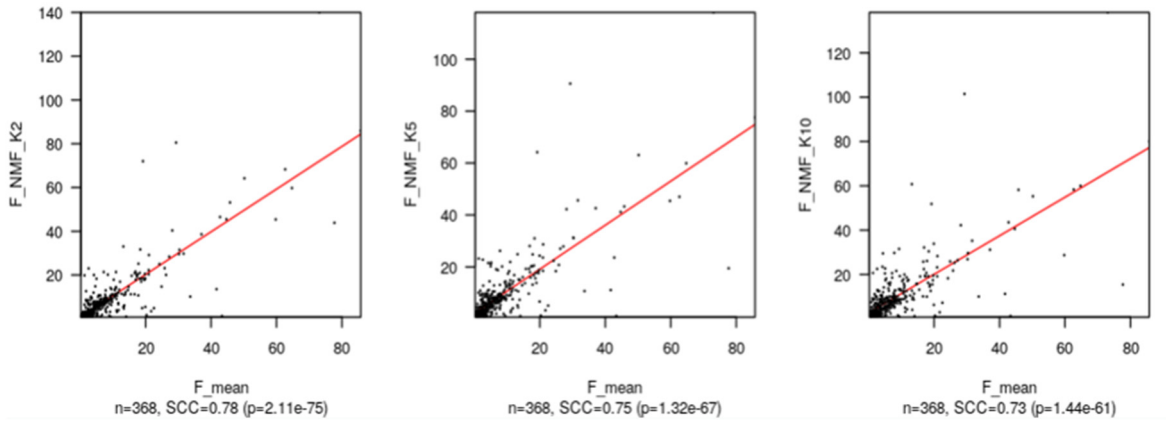


図 1. F_{mean} と F_{NMF} の値の比較
各点はシミュレーションデータセットの各遺伝子を表す。K=2、5、10 における、Fmean（x 軸）と FNMF（縦軸）を示している。スピアマン相関係数の p 値は t 分布への漸近近似によって計算された。

2. F 値および ΔF の精度検証

ODEGR を人工的に導入したシミュレーションデータに multiODGERfinder を適用することで、提案手法によってリードカバレッジが局所的に変動する領域を抽出できるかを検証した。そのために、F_{NMF} と F_{mean} の差である ΔF の値によって、シミュレーションデータの正例と負例の判別性能を AURO によって評価した。その結果、multiODGERfinder は ODEGR を有する遺伝子座と ODEGR を有しない遺伝子座を判別するタスクにおいて、AUROC が 0.80 以上という高い精度を示した。さらに、複数のトランスクリプト（アイソフォーム）を含む遺伝子座を判別できるかを検証した結果、0.7 以上の AUROC を示した。このことから、multiODGERfinder に

よるリードカバレッジ解析が、複数の細胞集団を含むがん 1 細胞 RNA-seq データからの ODEGR の発見に寄与できることが示唆された。

以上の検証結果に基づき、今後、提案手法を実際の複数のがん種由来の 1 細胞 RNA-seq データに対して適用し、医科学的な知見が得られるか確認する予定である。

共同研究者・謝辞

本研究の共同研究者は、長崎大学情報データ科学部准教授の松本拡高博士および情報・システム研究機構ライフサイエンス統合データベースセンター特任助教の大田達郎博士である。

文　献

1) Trincado JL, Sebestyen E, Pages A, Eyras E. The prognostic potential of alternative transcript isoforms across human tumors. Genome Med. 2016;8:14. PMID: 27473031, DOI: 10.1186/s13073-016-0330-8.
2) Shiraishi Y, Kataoka K, Chiba K, Okada A, Kogure Y, Tanaka H, et al. A comprehensive characterization of cis-acting splicing-associated variants in human cancer. Genome Res. 2018;28(8):1111-1125. PMID: 29907612, DOI: 10.1101/gr.233165.117.
3) Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2016;17(1):19-32. PMID: 26626313, DOI: 10.1038/nrg.2015.3.
4) Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. Genome Med. 2015;7:45. PMID: 25928073, DOI: 10.1186/s13073-015-0168-9.
5) Zhang Z, Lee JH, Ruan H, Ye Y, Krakowiak J, Hu Q, et al. Transcriptional landscape and clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer. Nat Commun. 2019;10(1):4562. PMID: 31601806, DOI: 10.1038/s41467-019-12520-8.
6) Manipur I, Granata I, Guarracino MR. Exploiting single-cell RNA sequencing data to link alternative splicing and cancer heterogeneity: A computational approach. Int J Biochem Cell Biol. 2019;108:51-60. PMID: 30633986 DOI: 10.1016/j.biocel.2018.12.015
7) Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun. 2018;9(1):619. PMID: 29434199 DOI: 10.1038/s41467-018-02866-0
8) Ozaki H, Hayashi T, Umeda M, Nikaido I. Millefy: visualizing cell-to-cell heterogeneity in read coverage of single-cell RNA sequencing datasets. BMC Genomics. 2020;21(1):1-10. PMID: 32122302 DOI: 10.1186/s12864-020-6542-z
9) Matsumoto H, Hayashi T, Ozaki H, Tsuyuzaki K, Umeda M, Iida T, et al. An NMF-based approach to discover overlooked differentially expressed gene regions from single-cell RNA-seq data. NAR Genomics Bioinformatics. 2020;2(1):lqz020. PMID: 34632380 DOI: 10.1093/nargab/lqz020
10) Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F, Ellisen LW. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun. 2018;9(1):3588. PMID: 30181541 DOI: 10.1038/s41467-018-06052-0