メタトランスクリプトーム解析：RNA-seqで環境を診る

佐藤　由也[1]*・小池　英明[2]

はじめに

RNA-seq（RNA sequencing）解析とは，主に次世代シーケンサーを用いたトランスクリプトーム解析を指す．高活性の代謝系を一網打尽で俯瞰できる本手法は，対象生物がそこで何をしているかといった，個体の挙動を知るための強力なツールである．しかし，この強力な解析手法にはコンピュータを使ったシーケンスデータ解析，いわゆるバイオインフォマティクスの知識・技術が必要になり，多くの実験科学者から，なんとなく敷居が高く思われがちだと思う．しかし，実際のところはどうであろうか……？

本稿の筆頭著者はここ数年でデータ解析を始めたばかりでビギナーに近い．そのため，多くの研究者がRNA-seq解析の何を難しそうに感じるかをよく理解できると考えている．本稿は，主に微生物を研究対象に扱う読者を想定しているが，ビギナー目線でなるべく噛み砕いた表現によりRNA-seqを紹介し，幅広い読者の「なんとなく難しそう」という先入観の払拭に貢献したい．また，データ解析は今後の生命科学研究において，より重要で一般的な技術になると思われる．その基礎の基礎ではあるが，誰かの役に立つ情報提供になれば嬉しい．

微生物コミュニティ自体を対象とする必要性

たった3種の微生物でさえ共在すると予測不能な振る舞いを示す[1]．一方，実環境において微生物が単一種で存在することは稀で，ほとんどの場合は複数種が混在し，微生物コミュニティを形成している．そして，それら多種の微生物活動の総和が環境現象や生態系のアウトプットとして表れる．たとえば，動物体内であれば共生細菌群の活動によってホストの体調が左右され，バイオフィルム構成微生物群によって金属表面の腐食が生じることが知られている．このように，微生物が実環境でおよぼす影響を考える時，どうしても複雑な微生物コミュニティそのものを対象に扱わなければならない．この課題に対し，メタゲノム解析，メタプロテオーム解析，メタメタボローム解析など有効なメタオミクス解析が開発されているが，本稿ではそれらのうち，メタトランスクリプトーム解析について紹介する．

メタトランスクリプトーム解析

微生物の挙動を知る強力なツールとしてトランスクリプトーム解析があげられる．トランスクリプトーム解析では対象生物のゲノム情報を基に各遺伝子のRNA発現量を解析することで，ある環境/細胞においてどの遺伝子や代謝系が活性化しているかを知ることができる．RNAの発現量を調べる手法はこれまでいくつか提案されており，10年ほど前はマイクロアレイ解析が主流であったが，現在は次世代シーケンサーを使ったRNA-seqによるものが一般的になっている．近年は受託解析を行う企業も増え，RNA-seqは以前よりずっと身近な選択肢になりつつある．しかしながらRNA-seq解析には大前提として，対象微生物のゲノム情報が必要であるため，多くの未培養微生物，ゲノム未解読微生物を含む実環境サンプルを対象に扱う（つまりメタトランスクリプトーム解析）には何らかの工夫が必要になる．一つはメタゲノム解析を並行して行う方法で，おそらくこれがオーソドックスなメタトランスクリプトーム解析方法である．この方法では，環境サンプル中の複数の微生物のゲノムを同時に解読し，取得したメタゲノムデータを参照配列としてRNA-seq解析を行う．一方，近年その代替策として，次世代シーケンサーから得られたRNA配列のみをつなぎ合わせて，自前で発現遺伝子のリスト（*de novo* assembly）を作る手法が開発されたため，こちらについて紹介する．

RNA-seq解析手順の概要

本稿では特に明示しない限り，現在もっとも一般的であるIllumina社の次世代シーケンサーを用いた，mRNAを対象としたRNA-seqについて記述する．また，バイオインフォマティクス一般に関する内容や，解析の詳細，各プログラムの入手方法や使い方については誌上スペースの都合上省略するが，筆者が参考にした書籍を数点紹介する（表1）．特に参考文献2と3は，学生やポスドクを主な読者層として想定し，読者に苦手意識を持たれないようにやさしくわかりやすく書かれておりお勧めであ

*著者紹介　産業技術総合研究所環境管理研究部門（主任研究員）　E-mail: yuya-satou@aist.go.jp
[1]産業技術総合研究所環境管理研究部門，[2]産業技術総合研究所生物プロセス研究部門

表1．インフォマティクスおよびRNA-seq関連書籍

文献	内容
2, 3	バイオインフォマティクス全般
4	RNA-seqの実験から解析の流れ
5, 6	Perl言語入門と詳細な使い方

る．文献2はコンパクトな説明書といった感じで読破しやすく頭の整理に役立つ．文献3は詳細な実践編で，演習が多く実際の解析が体験できる．

解析環境のセットアップ　シーケンスデータ解析ではUNIXやLinuxと呼ばれるオペレーティングシステム（OS）で動くコンピュータが必要になる．それらは基本的にはコマンドラインといって，文字列で命令文を打ち込むことでコンピュータを動かすが，このような方式をCUI（character user interface）という．一方，Windows PCのようにアイコンのクリックなどで動かすシステムをGUI（graphical user interface）という．もちろんGUIでできることはGUIで問題ないが，シーケンスデータ解析では扱うファイルの容量が大きく複雑なため，GUIではファイル開封さえできないことがある．一方，多数のファイルに対しての繰り返し作業が得意だったり，その一連の作業命令をファイルとして保存して後日再利用できたりと，CUIにはデータ解析に向いた性質が多い．UNIXやLinuxというと身構えてしまいそうだが，多くの生物系研究者に汎用されているMac PCはUNIX系であり，Macを持っていればシーケンス解析を始められる[3]．なお，データ解析には各ステップで必要なプログラム（ソフトウェア）のインストールが必要になるが，ほとんどすべて無料で公開されているため，PCさえ準備できれば追加でかかる費用はほとんどない[2,3]．ニーズに合わせて多様な解析を行いたい場合は，簡単なプログラミング言語（Perl，Python，Rubyなど）を習得する必要がある．ちなみに筆者は重い計算用にはLinux系のOS（Ubuntu）を入れたデスクトップコンピュータを，細かな作業用にはノートPCのMacbook Proを用い，言語は主にPerlを使っている．

次世代シーケンサーの生データ　次世代シーケンサーから出力されるアウトプットファイルはFASTQという形式になっている．FASTQファイルは，簡単に言うと試料中の各DNA鎖についての配列情報（A，T，G，Cなど）とそのクオリティ値（Q値）から構成される．生物系の研究者の多くはFASTA形式と呼ばれる塩基配列ファイルを扱った経験があると思うが，そこに塩基ごとのクオリティが付記されたイメージである．なお，このQ値はシーケンサーのエラー確率を意味し，「Q値＝$-10\log_{10}$［エラー率］」で表され，たとえば，Q値が20ならばエラーが生じる率は1.0%であり，読み取られた塩基の信頼度は99%となる．同様にQ値が30ならば信頼度は99.9%である．

配列データ解析の流れ　配列データ解析で最初に行うことは，上述のQ値を基にした低クオリティ配列データの除去である．これにはたとえばTrimmomaticなど公開されているプログラムを使う[7]．次に，得られた高クオリティ配列データを使い，RNA発現量を解析する（詳しくは後述）．解析対象微生物のゲノム情報が公開されていれば手間は少なく，GenBankなどのデータベース（DB）からゲノム情報をダウンロードし[8]，それを参照配列として発現量解析を行うことができる．しかし前述の通り，対象微生物がゲノム未解読である場合，自前で発現量解析用の参照配列を作成する必要がでてくる．

自前で発現遺伝子リストを作る：*de novo* assembly

シーケンス解析において，“assemble”とは相同な複数の配列をつなぎ合わせてより長い配列を生成する工程を言う．次世代シーケンサーから出力されるDNA配列長は一般的に短いため，自分の知りたい遺伝子の全長配列を得るためにはこのようなつなぎ合わせの作業が必要になる．また，“*de novo*”は「初めから」を表すので，*de novo* assemblyとは既知ゲノムなどのテンプレートを使わずに，短い配列のつなぎ合わせのみによって新たに作製された長鎖DNA配列（のリスト）を指す．

もともと*de novo* assembleはゲノム解析などで，シーケンサーから得られた短鎖配列をつないでゲノム全長配列を生成する工程を指したが，近年は同じ技術がトランスクリプトーム解析に応用されている．すなわち，解析対象サンプル中で発現しているRNAの配列を次世代シーケンサーで網羅的に解読し，得られた短鎖配列をつなぎ合わせると，そのサンプル中で発現していた転写産物（RNA）の全長配列になる．環境サンプルを対象にする場合，そこにいる微生物群全種の全遺伝子配列は到底わからないが，少なくとも一定量発現している遺伝子については配列を知ることができる．

RNA配列アセンブラーの特徴　従来型と次世代型シーケンサーでは*de novo* assemble用のプログラム（アセンブラー）が異なる．アセンブラーは生データの配列長が長ければ長いほど信頼性の高いassemblyを作ることができる．これは直感的に理解しやすく，一配列が長いほど他の配列と重なる部分が長くなり，つなぎ合わせがしやすい．しかし，Illumina社のシーケンサーから得

られる配列データは，鎖長は短い（～300 bp）がトータルの配列データ量が圧倒的に多いことが特徴であり，従来のアセンブラーではつなぎ合わせが難しい．そのうえ転写産物は，そもそもその全長が短い（微生物遺伝子の平均鎖長は1 kbp前後）ことからも，従来のアセンブラーは適用しづらかった．そこで登場したのが短鎖用に特化したアセンブラーであり，RNA-seqに強いTrinityなどがある[9]．このプログラムは興味深く，生データ配列を一度短く区切ってからつなぎ直すというアルゴリズム（de Bruijnグラフというグラフ理論に基づいた方法）を採用しており，あえて短くすることでコンピュータでの計算が高効率化し，シーケンサーのエラー率や，真核生物で観察されるスプライシングの影響などを反映しやすくなっているらしい．図1に示す通り，高クオリティのFASTQファイルをインプットとしてTrinityで計算をすると，発現遺伝子のリスト（assembly）がFASTA形式で出力される．

各配列の機能推定　次に，生成した*de novo* assemblyの各遺伝子配列について機能推定を行うが，ここは非常に重要なステップである．最終的には機能遺伝子をコードする領域（coding sequence: CDS）の解析を行うが，その前にrRNAなどの非CDS配列をassemblyリストから除外しなければならない．もちろんウェット実験でもrRNA除去の工程があるが，それでも結構な割合でrRNAなどがシーケンスデータに残っている．ここでは一例として，筆者らが行った一連の非CDS除去手順を記載する．まず，rRNAの除去にはSilva rRNA DBを参照にRiboPickerというプログラムを使用した[10,11]．ただ，これでは5S rRNAが除去できなかったため，5S rRNA DBを参照にFasta36というアライメントプログラムを使ってさらに配列除去を行った[12,13]．次に，意外にかなりの量が含まれていたのがtmRNA（transfer messenger RNA）で，その除去にはtmRNA DBを参照にFasta36を使用した[14]．最後にtRNAの除去を行ったが，ここではtRNA scanという，tRNA配列予測プログラムを使用した[15]．これでやっと，CDSがメインのassemblyリストが出来上がり，各遺伝子の機能推定に進む．CDSの機能予測にはBlastというプログラムを用いた[16]．これは，一般的に用いられるWeb上のBlast検索と同じプログラムだが，Web上で解析をするのではなく，プログラム自体を自分のコンピュータにインストールし，自分のコンピュータ内でBlast検索を行う．また，参照するDBは多様な遺伝子を保存している必要があるが，ある程度の信頼性が必要であるため，NCBIが提供しているRefSeqを用いた[17]．このようにして，環境サンプル中で発現している遺伝子のDNA配列とその機能情報のリストを作り，これを参照配列に発現量解析を行う．

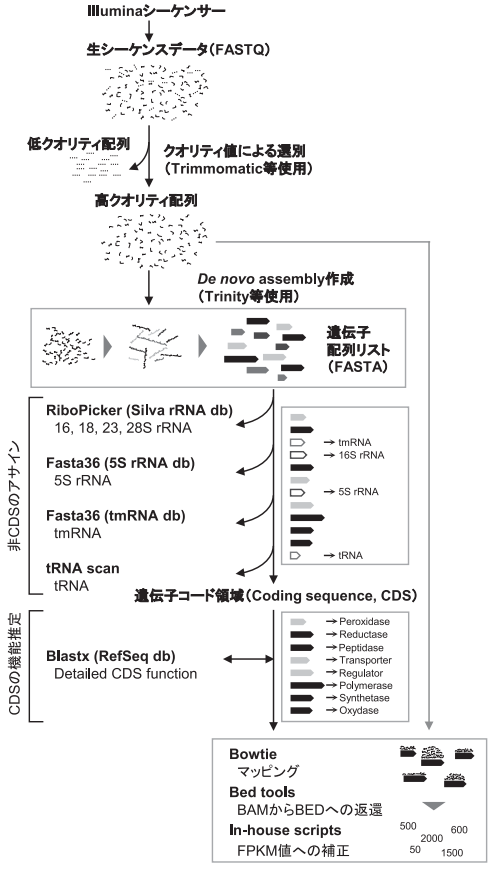

図1．*De novo* assemblyを用いたRNA-seq解析の手順

遺伝子発現量の定量　ここではマッピングと呼ばれる作業を行う．これは文字通り，対象微生物のゲノム配列などをテンプレートにして，次世代シーケンサーで読み込んだRNA配列を貼り合わせて行く作業である．ある遺伝子領域には多くのRNA配列が貼り付けられ，別の領域にはあまり貼りつかなかったりするが，それを定量することによって各遺伝子の発現量が計算される．このステップでは，たとえばBowtieなどのプログラムが使われる[18]．しかし，Bowtieから出力される結果ファイルは人の目で見ても解らない，bamというバイナリー形式になっている．そのため，得られた結果ファイルをすぐに別の形式に変換する．そこではたとえば，BEDToolsというプログラムパッケージ中のBamToBedというプログラムを使う[19]．これによって変換されるbedファイルはテキスト形式で，人間も読解可能である．

遺伝子発現量のノーマライズ　次に行う作業は，上記で算出された発現量値のノーマライズである．マッピングによって得られる遺伝子の発現量は一つの遺伝子領

域に対して何本のリード（次世代シーケンサーから出力される塩基配列ファイルのうちの，一本の塩基配列をリードと呼ぶ）がマップされたかという情報だが，長い遺伝子ほどマップされる確率は高いので，遺伝子の長さによる補正をかける必要がある．一般的には得られた値を遺伝子鎖長で割り，1 kbpあたりのリード数という値にする．さらには，得られる発現量の値は，そもそも次世代シーケンサーから出力される配列数の合計値によっても異なってしまう．次世代シーケンサーから出力されるデータ量というのは常に一定というわけではなく，解析の度にある程度変化する．たとえば，ある時には次世代シーケンサーの調子が良く，たくさんのデータが出力されたとすると，その時のサンプルの遺伝子発現量は全体的に高く評価されてしまう．このバイアスを解消するため，合計が1,000,000リードであった場合にはいくつになるか，ということで補正をかける．長さの補正と合計リード数の補正を合わせて，RPKM（Reads per Kilo-base per Million-read）値もしくはFPKM（Fragments per Kilo-base per Million-read）値としたものがRNA-seq解析で共通の遺伝子発現量の単位として使われている．なお，このステップではPerl言語により自身で作成した簡単なプログラムを使った．

De novo assemblyを用いたRNA-seqの利点と解析例

De novo assemblyを参照配列にしたRNA-seqの利点をいくつか列挙する．まず一つは，シーケンス回数を減らせることである．メタゲノムを参照配列とする場合は，メタゲノム用に環境サンプル中のDNAを対象に次世代シーケンス解析を行い，さらにそれとは別にRNAを対象にシーケンス解析をするが，*de novo* assemblyの場合はRNAのシーケンスだけで済む．さらに，環境微生物学および微生物生態学における重要な利点として，マイナー種の解析に強いという点があげられる．メタゲノム解析は改良が重ねられ，かなり高効率になっているが，環境中のDNAシーケンスを行うため，原理上，存在量の高い微生物種ほど検出されやすい．そのため，たとえば対象環境中で存在量は少ないが重要な働きをする微生物がいたとしても，メタゲノム解析で検出されなければ遺伝子発現解析の対象になることは難しい．

佐藤らのグループでは重油含有廃水を処理する活性汚泥（数千種以上から構成される水処理微生物群）リアクターを対象に*de novo* assemblyを用いたRNA-seq解析を行っているため紹介する[20]．この実験では二つのリアクターを同じ条件で運転したにもかかわらず，何らかの原因で重油処理性能に大きな差が生じた．RNA-seq解析を行ったところ重油分解微生物が特定されたが，分解酵素の発現量は二つのリアクター間で大差なく，処理能を二分する要因は他にあることが示唆された．さらに詳細に解析を進めると，興味深いことに，重油分解微生物の呼吸基質の供給を担う，存在量0.25%にも満たないマイナー種の活性に差異があり，それよって水処理システム（微生物コミュニティ）全体の重油分解活性が左右されることを見いだしている．この研究から，*de novo* assemblyによるRNA-seqが環境微生物群そのものを対象として機能し，一見すると原因不明な環境中での現象も，RNA-seqによって解析可能であることが示された．また，上述のように存在量が小さいマイナー種が生態系において重要な役割を担うという報告例は少ないが，これこそ*de novo* assemblyの強みを生かした結果といえる．

おわりに

「われわれはたいていの場合，見てから定義しないで，定義してから見る．」これは「世論（Walter Lippman著）」という書籍の，人間のステレオタイプについての一説である．我々は多くの場合，見聞きした情報に対して，すでに自分の中に蓄積したイメージを付加して捉える．情報が新しく馴染みがないほど，その独自イメージの部分が大きくなりがちだと思う．そしてたとえば，バイオインフォマティクス，データ解析，プログラミングなど，いかにも難しそうな単語を目にすると，「いかにも難しそう」と感じる人が多いかもしれない．しかし多くの物事に共通して，実際に取り組んでみると思ったより簡単と感じることがある．個人的には，データ解析のスキルは今後より必要性が高まると思うので，読者の方で躊躇している方がいたらぜひトライしていただきたい．インフォマティクスに関する教科書的な内容全部を理解することは非常に大変だと思うが，部分的に知っているだけでも情報解析は可能である．必要なことだけ，必要になった時に習得すれば良いと思う．このようなゆるい感覚でも，指導してくれる師匠に恵まれれば，充分にシーケンス解析技術は身につくはずである．

文　献

1) Becks, L. *et al.*: *Nature*, **435**, 1226 (2005).
2) 坊農秀雅：Dr. Bonoの生命科学データ解析，メディカル・サイエンス・インターナショナル (2017).
3) 清水厚志ら：次世代シークエンサーDRY解析教本，学研メディカル秀潤社 (2015).
4) 鈴木　穣ら：NGSアプリケーション RNA-Seq実験ハンドブック，羊土社 (2016).
5) 高橋順子：ゼロからわかるPerl言語超入門，技術評論

社 (2011).
6) Schwartz, R. L. ら 著, 近藤嘉雪 訳：初めてのPerl第6版, オライリー・ジャパン (2012).
7) Bolger, A. M. *et al.*: *Bioinformatics*, **30**, 2114 (2014).
8) Benson, D. A. *et al.*: *Nucleic Acids Res.*, **41**, D36 (2013).
9) Grabherr, M. G. *et al.*: *Nat. Biotechnol.*, **29**, 644 (2011).
10) Quast, C. *et al.*: *Nucleic Acids Res.*, **41**, D590 (2013).
11) Schmieder, R. *et al.*: *Bioinformatics*, **28**, 433 (2012).
12) Szymanski, M. *et al.*: *Nucleic Acids Res.*, **44**, D180 (2016).
13) Pearson, W. R. and Lipman, D. J.: *Proc. Natl. Acad. Sci. USA*, **85**, 2444 (1988).
14) Zwieb, C. *et al.*: *Nucleic Acids Res.*, **31**, 446 (2003).
15) Lowe, T. M. and Eddy, S. R.: *Nucleic Acids Res.*, **25**, 955 (1997).
16) Camacho, C. *et al.*: *BMC Bioinformatics*, **10**, 421 (2009).
17) O'Leary, N. A. *et al.*: *Nucleic Acids Res.*, **44**, D733 (2016).
18) Langmead, B. and Salzberg, S. L.: *Nat. Methods*, **9**, 357 (2012).
19) Quinlan, A. R. and Hall, I. M.: *Bioinformatics*, **26**, 841 (2010).
20) 佐藤由也ら：日本生物工学会大会講演要旨集，p. 234 (2017).