

次世代シーケンサーデータの解析手法 第1回イントロダクション

門田 幸二^{1*}、孫 建強²、湯 敏²、西岡 輔¹、清水 謙多郎^{1, 2}

東京大学大学院農学生命科学研究科

¹ アグリバイオインフォマティクス教育研究ユニット

² 応用生命工学専攻

次世代シーケンサー（以下、NGS）は、モデル生物や非モデル生物を問わずゲノム解析やトランスクリプトーム解析（以下、RNA-seq）、そして菌叢（microbiome）解析など幅広く利用されている。データ解析手法も多数提案されており、目的にもよるが乳酸菌程度のゲノムサイズであればノート PC で自在に解析できる環境を構築可能である。しかし現実には、データ解析環境の構築自体が多くのユーザにとって乗り越えることのできない壁である。また、たとえその壁を乗り越えて解析できたとしても、利用しているプログラムをよく理解しないまま実行ボタンを押し、得られた結果の解釈で戸惑う研究者も多い。そこで本連載では、NGS データ解析を最小限の労力で自在に行えるようになりたい実験系研究者向けの全般的な情報提供を目的とし、筆者らが平成 16 年度より実施している大学院教育プログラムの中から、NGS 解析周辺の講義内容を中心に基礎から応用まで幅広く述べる。第 1 回は、全体のイントロダクションを行う。ウェブサイト（R で）塩基配列解析（URL：http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html）中に本連載で述べるリンク先を掲載してあるので効率的に活用してほしい。

Key words : NGS, RNA-seq, Bioinformatics, R, Linux

情報収集先（講義、講習会、ウェブサイトなど）

筆者らの所属機関であるアグリバイオインフォマティクス教育研究プログラム（以下、アグリバイオ）では、数年前より NGS 関連ハンズオン講義（ノート PC を用いた実習を含む講義のこと）の割合を徐々に高めている。大学院講義ではあるものの、東京大学以外の学生、一般企業の社会人、ポスドクも受講可能である。アグリバイオの主な目標は、高度なバイオインフォマティクス養成ではなく、手元にあるデータを自在に解析する技術を身につけたい実験系研究者の養成である。例年 20% 程度の受講生が東京大学以外であり、受講費用もかからないため、本誌読者向けの講義プログラムといえる。

NGS データ解析手法に関わるバイオインフォマティクシヤンの多くは、日本バイオインフォマティクス学会（JSBi）か NGS 現場の会に所属している。これらの年会や研究会への参加を通じた情報収集も有意義であろう。例えば、2014 年度の JSBi 年会は 10 月に仙台で開催されるが、いくつか講習会も企画されている。HPCI 人材養成プログラムも比較的規模の大きな活動を実施している。お台場にある産業技術総合研究所・CBRC を拠点として、NGS に特化した内容ではないものの、セミナー、ワークショップ、チュートリアル、e-learning などが精力的に実施されている。

e-learning 系の代表格としては統合 TV¹⁾ が挙げられる。文字通り様々なウェブツールやデータベースなどの使い方を紹介する番組である。“NGS”というキーワード検索でリストアップされるものは全 776 番組中 9 番組と意外に少ない（2014 年 5 月 4 日調べ）ものの、“次世代シーケンサー”で検索すると 20 番組程度がヒットする。また、バイオインフォマティクシヤンの多くが利用している Linux（リ

*To whom correspondence should be addressed.

Phone : +81-3-5841-2395

Fax : +81-3-5841-1136

E-mail : kadota@bi.a.u-tokyo.ac.jp

なつくす、と読む) 環境を構築する方法や代表的なプログラミング言語である Perl の使い方などの番組も提供されており、バイオインフォマティクスを本格的に学びたいヒトにとっても有意義なウェブサイトである。NGS 解析にほぼ特化したものとしては、イルミナ社のウェビナーシリーズやアメリエフ社がスライドホスティングサービスの SlideShare に公開した講義スライドも比較的良好にヒットする。

NGS は、医療やライフサイエンス分野の諸課題を効率的に解決するための道具の一つであり、急速かつ広範に利用されている。これまで NGS 解析に特化したカリキュラムは未整備であったが、2014 年 3 月にバイオサイエンスデータベースセンター (NBDC) によって NGS 用のバイオインフォマティクス人材育成カリキュラムが策定された。このカリキュラムは、NGS データを扱うにあたり最低限必要とされる事柄を 2 週間程度で身につけることを想定した「速習」と、時間をかけて習得することを想定した「速習以外」に分かれている。2014 年 9 月に 2 週間の「速習」コースが試行的に開催され、多くの項目について講義資料も公開される予定である。この取り組みは、文字通りバイオインフォマティクス人材養成に関するものであるため、Linux 導入やプログラミング言語も含まれている。そのため、手元にあるデータをすぐに解析したい実験系研究者にとっては敷居が高いかもしれないが、このカリキュラム中に記載されている習得技術が本来執筆すべき事柄といっても過言ではない。

データ解析環境 (Linux)

NGS データ解析を高速かつ効率的に実行するプログラムの多くは、Linux というデータ解析環境上でのみ動作する。それゆえ、バイオインフォマティクスを目指す場合には、しばしば Linux 環境構築が最初の課題として与えられる。最近では、普段利用している Windows または Macintosh マシン上で Linux 環境を同時に利用するための手軽な手段も提供されている。この種の勉強を本格的に始めようとする、慣れない用語に戸惑う読者は意外に多い。まず Linux とは何か? Linux は、Windows や Mac OS などと同じオペレーティングシステム (以下、OS) の一種である。多少の誤解を恐れずに説明する。Windows や Macintosh が安定して使いやすい大衆車だとすると、Linux は乗りこなせると速いスポーツカーのようなものである。次に戸惑うのは Linux の種類 (ディストリビューション) に関する記述である。Windows に Windows 7 や Windows 8.1 があるように、Linux にも様々な種類がある。Linux (スポーツカー) に関する情報収集をしているつもりでも、いつのまにか説明内容が Ubuntu (うぶんつ、と読む) とか LinuxMint とか CentOS などに切り替わっていてだんだん混乱してくる場合が多いだろう。Ubuntu と

か LinuxMint とか CentOS などは Linux の一種である。様々な意見を総合すると、初心者は Ubuntu がお薦めである。

ここでは、Windows PC 上で Ubuntu 環境を構築するという前提で話を進める (Mac ユーザは Windows を Mac に読み替えるだけでよいが、Linux と Mac は互換性が非常に高いので仮想化マシン導入は必要ないかもしれない)。手順にしたがって Ubuntu をインストールする際に面食らうのは、一見意味不明な用語である仮想マシンまたは仮想化ソフトのインストールを要求されることである。手順としては、Windows PC に「仮想化ソフト」をインストールして「仮想の PC 本体」を作り、仮想の PC 本体上で Ubuntu を構築する。感覚的には、Windows という OS の上で Microsoft Word や Excel のようなアプリケーションソフトウェアをインストールして動かすのと同様に、Windows 内で Ubuntu を動かすのである。ではなぜ仮想化ソフトのインストールという一見余分な手順を要求されるのか? それは Ubuntu がアプリケーションソフトウェアではなく OS という、本来ならハードウェアの上で直接動かすべきソフトウェアだからである。異なる OS 間では、プログラムの内部構造が異なるため、Ubuntu の上で動くプログラムはそのまま Windows の上で動かすことはできない。そこで仮想化ソフトを使って、Windows OS (ホスト OS) 上で Ubuntu という Linux OS (ゲスト OS) を同時通訳的に実行し、あたかも Ubuntu の PC があるような環境を作って、その上で Ubuntu の上で動くプログラムを実行するのである。仮想化ソフトは、単純に Windows という車と Linux という車を同時に操作するために必要な一種のコントローラと解釈してもよいだろう。通常は 1 台の車しか運転できないが、仮想化ソフトというコントローラのおかげで 2 台の車を同時に操作することができる。しばしば動作が不安定になるのは 2 台の車を同時に操作するという複雑なことをしているためであると解釈すればよい。

代表的な無料の仮想化ソフトとしては、VMware 社の VMware Player とオラクル社の VirtualBox の 2 つが挙げられる。VMware Player は非営利に限りフリーであり、歴史が古く安定している。VirtualBox は比較的最近開発されたソフトウェアで利便性が高い一方、やや動作が不安定だという声をきく。アグリバイオの講義で使用する PC には VMware Player をインストールしている。理由は、ウェブ上で収集可能な情報量が多くトラブル対応が比較的容易だったからである (図 1)。

アセンブルなど多くの NGS 解析手法を自在に操りたいバイオインフォマティクス中級～上級を目指したい場合には、是非 Linux 環境構築に挑戦してほしい。最初は手持ちのノート PC で十分である。cd, ls, grep などの独特なコマンド操作、見慣れない GUI 画面、Windows 向けのソフトウェアなどとは異なりダブルクリックでプログラムのイ



図 1. アグリバイオ講義風景。講義時に貸与する 100 台弱の Windows PC に VMware Player や R がインストールされている。

インストールができない理不尽さに嫌気がさすこともあるだろう。しかし、NGS データ解析を最も効率的に行えるのは Linux 環境であることや、以下に述べる R も Linux 上で実行可能である。「NGS 講習会」などでウェブ検索すれば、初心者向け Linux 講習会もどこかで開催されていることに気づくであろう。Linux 系の講習会は手厚いサポートを要するため、多くても 40 名程度以下に受講人数が制限されている場合が多いが積極的に参加してみるとよい。

注意すべきは、Linux 自体はデータ解析環境にすぎないという点である。つまり、Linux をインストールしただけではアセンブルやマッピングなどの NGS 解析用プログラムを実行することはできない。それらのプログラムは独立にインストールする必要があるが、一般に Linux 環境上でのプログラムのインストールは苦行である。もちろん、NGS 解析用プログラムを含む様々な解析ソフトが一通り組み込まれた Bio-Linux²⁾ というものが存在する。これは、Ubuntu をもとにして、バイオインフォマティクス解析用にカスタマイズされた OS である。Bio-Linux をインストールすれば、Ubuntu の操作感で FastQC や Picard による NGS データのクオリティコントロールやフィルタリング、ABYSS³⁾ や Velvet⁴⁾ によるアセンブル、Bowtie2⁵⁾ や BWA⁶⁾ によるマッピング、Cufflinks⁷⁾ による遺伝子構造推定、MEME⁸⁾ によるモチーフ解析、MAFFT⁹⁾ や T-Coffee¹⁰⁾ による多重配列アラインメント、WebLogo¹¹⁾ による sequence logos¹²⁾ の実行、Cytoscape¹³⁾ によるネットワーク解析、blast2 による BLAST¹⁴⁾ の実行など、実に多様な解析を行う環境が整っている。尚、プログラムやソフトウェアの他に、パッケージ、ライブラリ、ツールなど多様な呼び方が存在する。微妙なニュアンスの違いや包含

関係などはあるものの、FastQC はプログラムであり、ソフトウェアであり、ツールでもある、などと解釈すればよいだろう。

データ解析環境 (R)

R¹⁵⁾ は Microsoft Word や Excel のようなソフトウェアの一つである。Windows, Macintosh, Linux のどの OS にもインストール可能であり、NGS データ解析分野においても幅広く利用されている。統計解析ソフト R、R 言語、R 環境など様々な呼び方が存在する。これは、視点を変えると様々な捉え方ができるためであること、R だけで検索すると無関係のものが多くヒットしてしまうため、それを避けたいという理由もある。R は、NGS データ解析を行う上でも非常に強力なツールとして近年急速に普及が進んでいる。実際、アグリバイオの NGS 関連講義でも採用されているほか、上記の NGS 用カリキュラムにも組み込まれている。

ただし、R のインストールを行っただけの状態では事実上 NGS 解析はできない。PC の機能をフル活用すべく Word や Excel のような各種ソフトウェアをインストールするのと同様、R の機能をフル活用して NGS 解析を行うためには CRAN や Bioconductor¹⁶⁾ から配布されているパッケージと呼ばれるものをインストールする必要がある。Excel のアドインや Java プラグインの概念がわかるヒト向けの説明としては、パッケージはそれらと同じようなものという理解でよい。特に、NGS 解析を行う上で Bioconductor から配布されている各種パッケージのインストールは必須である。筆者らは、一部を除く CRAN お

よび Bioconductor から配布されている全てのパッケージをインストールすることを勧めている¹⁷⁾。R パッケージのインストールは Linux に比べ格段に容易ではあるものの、個別のパッケージのインストールで済まそうとした場合、パッケージ間の依存関係の問題に悩まされることが経験上多いためである。具体的な推奨インストール手順については、(R で) 塩基配列解析 (URL: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html) 中の「R のインストールと起動」を参考にされたい。研究所や大学内の有線 LAN 経由でインストールする場合に、ときどきプロキシの問題でパッケージのインストールがうまくいかないという相談を受ける。この場合、「R プロキシ インストール」などのキーワードでインターネット検索することによって問題を解決できるであろう。また、R-Tips や RjpWiki などのウェブサイトも有用である。

(R で) 塩基配列解析は、ゲノムやトランスクリプトームなどの NGS 解析を R で行うためのスクリプト集である。上記推奨手順に従って必要な各種パッケージをインストール済みであるという前提のもとに記述されているが、この前提条件さえクリアしていればコピー & ペーストで手軽に解析可能である。また、自分の実験デザインやデータに適合したスクリプト探しが容易であること、例題も豊富であることから、テンプレートと解析目的が若干異なる場合でも必要な変更箇所を特定しやすいという特徴をもつ。「自分の入力ファイルを実行してもエラーが出てうまくいかない」と「自分のデータの場合にどこをどう変更すればいいかわからない」は、実験系ユーザがデータ解析でつまづく二大原因である。このウェブページが多くユーザに利用されるのは、データ解析でつまづくことがほとんどないという点に尽きる。

トランスクリプトーム解析手段としてマイクロアレイ解析を行ってきた研究者の多くは、R および筆者のウェブページ (R で) マイクロアレイデータ解析の利用経験があるかもしれない。Windows や Macintosh といった普段利用する PC 環境上の R のみで、生データ取得、発現変動解析、Gene Ontology や Pathway 解析などの各種機能解析まで一通りのマイクロアレイ解析が可能である。したがって、NGS を用いたトランスクリプトーム解析 (RNA-seq) の場合になぜ Linux 環境が推奨されるのか理解し難い読者も多いかもしれない。この理由は明確で、マイクロアレイデータは数値であるのに対し、RNA-seq データは塩基配列だからである。1980 年代後半頃から活動していたバイオインフォマティクス分野の先駆者は、FASTA¹⁸⁾ や BLAST¹⁴⁾ など塩基配列データを効率的に解析する手法開発を Linux 環境 (当時は UNIX 環境) 下で行っていた。現在活躍している塩基配列解析系バイオインフォマティクスの多数派は、その流れを汲んでいる。開発してきたプログラム群の蓄積もあり、研究室単位や細分化された研究分野ごとに C 言語を代表とする実行速度が非常に速い

コンパイラ型言語、Perl、Python、Ruby など実行速度が遅いもののプログラム作成が比較的容易なインタプリタ型言語など様々な流派が存在するが、Linux 環境下で動作するプログラム開発を基本としている点では同じである。しかし、これらのプログラミング言語が担ってきた役割の多くを通常利用 PC 環境下で代替可能な R の機能をフル活用したいという要望は多い。それゆえ、本連載では可能な限り R 環境で解析するというスタンスをとる。

プログラミング言語

上述のように、プログラミング言語はコンパイラ型とインタプリタ型に大別可能である。NGS データ解析分野で利用される解析プログラムのうち、コンパイラ型言語 (C 言語や C++ 言語) で書かれているものの代表例はアセンブリやマッピングなどの計算量の多いプログラムである。例えばゲノムアセンブリプログラムの一つである Platanus¹⁹⁾ は、C++ で記述されている。一方、インタプリタ型言語は、プログラム作成が比較的容易という特徴を生かして、ファイル形式の変換のようなちょっとした作業に利用される場合が多い。代表的なものとしては、Perl、Python、Ruby などが挙げられる。Perl は 1980 年代後半に登場し、これら 3 つのプログラミング言語の中では最も歴史が古い。特徴としては、テキストなどの文字列処理を行う際に便利である。筆者の印象では、40 代以上の多くのバイオインフォマティシャンは Perl プログラミング経験がある。Python は、1990 年代前半に登場し、プログラミングの容易さなどの特徴から NGS 解析分野でよく利用されている。Ruby は 1990 年代中盤に登場し、Perl や Python の長所を継承している。しかし後発組であるがゆえに、すでに普及していた Perl や Python の牙城を崩すには至っていないという印象を受ける。例えば、RNA-seq データ解析プログラムの一つである Grape²⁰⁾ は、内部的に Perl、Java、R、そして Python を利用している。

三者ともに NGS 解析分野でも利用可能なプログラム群 (BioPerl、Biopython、BioRuby) が提供されており、このインタプリタ型言語でないといけないという明瞭な差別化はおそらく不可能である。また、これらのインタプリタ型言語でできることの多くが R でも可能である。ただし、筆者らの知る限り、R でマッピングを行うことは可能であるが、アセンブルはできない。それゆえ、Platanus のような最先端のアセンブルプログラムを利用したい場合には、簡便な R 環境ではなく Linux 環境を構築する必要がある点に注意されたい。

ウェブツール

Linux 環境を回避しつつ自力でアセンブルを含む計算コストのかかるデータ解析を行う手段としては、ウェブ

ツールの利用が挙げられる。特に DDBJ Read Annotation Pipeline²¹⁾ は、おそらく最もお手軽な解析手段だと思われる。Bio-Linux 同様、マッピングでは BWA や Bowtie などが、そしてゲノムアセンブリでは ABySS や Velvet などが利用可能である。トランスクリプトームアセンブリでは Trinity²²⁾ が利用可能である。DDBJ の名前から推察できるように、NGS データの登録と一体的にデータ解析まで行えることや日本語のマニュアルも存在する。また、手元の FASTQ ファイルをアップロードして解析することも可能である。統合 TV において DDBJ Read Annotation Pipeline で検索すると 4 つの番組がリストアップされるのでいくつか視聴するとよい。

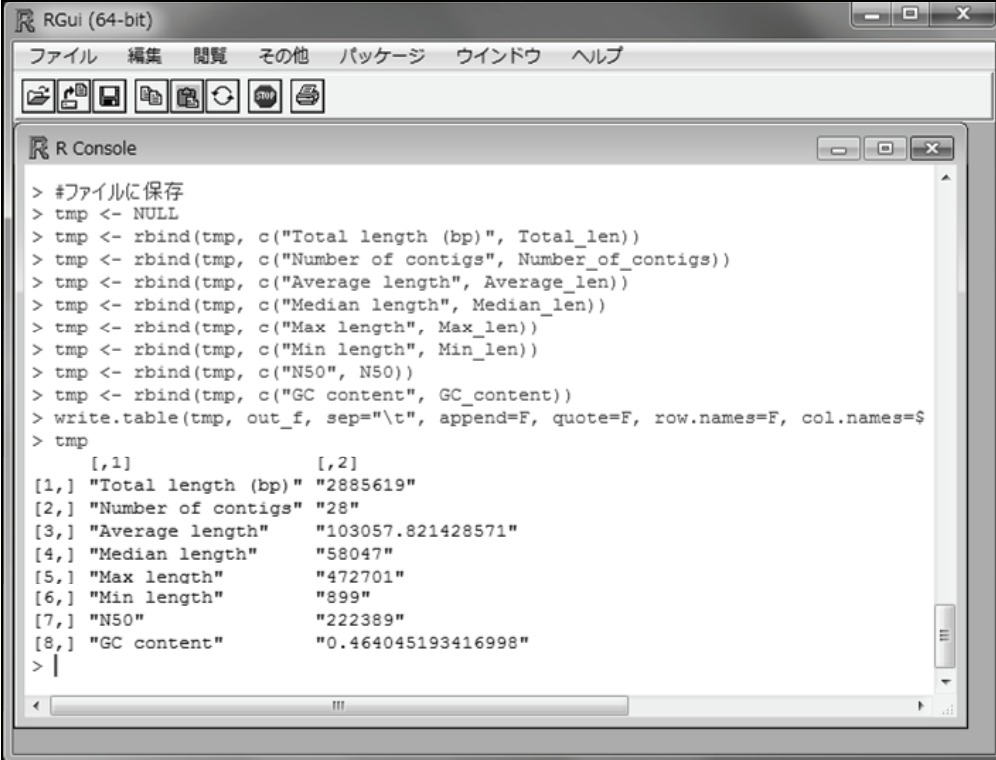
Galaxy²³⁾ の利用も DDBJ Read Annotation Pipeline と同様の解析が可能であり、検討に値する。ウェブツールであるがゆえに、アグリバイオのような 100 人規模のハンズオン講義で利用するのは（大人数でアクセスすると結果がなかなか返ってこないため）事実上不可能である。しかし、通常の個人利用のような状況であれば特に問題なく、実験系研究者にも比較的利用者が多い。DDBJ Read Annotation Pipeline は、アセンブリやマッピングなどの一次解析以降の高次解析ツールとして Galaxy を提供している（P-Galaxy, <https://p-galaxy.ddbj.nig.ac.jp/>）。本家の Galaxy (<https://usegalaxy.org/>) と基本的な見栄えは同じであるが、いくつか独自の機能が組み込まれている。初心者は、日本語情報の豊富な DBCLS Galaxy (<http://galaxy.dbcls.jp/>) から始めたほうが分かりやすいかもしれない。

Expression Atlas²⁴⁾ も今後の発展が大いに期待される。これは比較的最近まで Gene Expression Atlas²⁵⁾ と呼ばれていたものである。基本的には、二大発現データベース（以下、DB）の一つである ArrayExpress²⁶⁾ 中の多くのデータセットについて発現変動解析などの付加価値がつけられた二次的な DB という理解でよい。Gene Expression Atlas については統合 TV に番組が存在するのでそちらを参考にされたい。Gene という名前が消えたのは、NGS（特に RNA-seq）の出現により、ArrayExpress²⁶⁾ 中に占める RNA-seq データの割合が増加したためであろう。つまり、3' 発現アレイの頃の遺伝子レベルの解像度ではなく、トランスクリプトームアレイ²⁷⁾ や RNA-seq から得られる転写物レベルの解像度の発現データが増加しており、もはや遺伝子という言葉が死語になりつつあるからである。実際、アグリバイオの講義でも「遺伝子発現～」から「発現～」という表現方法に切り替えつつある。

一口に RNA-seq データ解析とはいっても、その目的は多様である。ゲノム配列情報などが乏しい非モデル生物の解析の場合には、トランスクリプトームアセンブリによる転写物配列自体の決定が目的となることもある。これは一昔前の EST 解析のようなものである。また、（ドラフト）ゲノム配列が手元にある場合には、Cufflinks²⁸⁾ など

を用いたエクソンの位置や isoform 情報などを得る遺伝子構造および発現量推定が行われる。この種のプログラムは、GENSCAN²⁹⁾ などの予測ではなく、ゲノム配列上のどこからどの程度転写されているかを知るために RNA-seq リードがどこにどの程度マップされるかという情報を利用するものである。Cufflinks などから得られる RPKM 値や FPKM 値のような発現量情報を用いることで、目的のサンプル内でどの転写物がどの程度発現しているかという発現レベルの大小関係を把握することができる。これが Expression Atlas で提唱されている“baseline”情報の概念に相当するものである。また、Expression Atlas で提唱されているサンプル間での違いを解析するときの“contrasts”という概念に相当するものとして、比較するグループ間での発現変動解析が目的の場合には、入力データは上記発現量情報（つまり“baseline”情報）ではないという点にも注意されたい³⁰⁻³¹⁾。これは一種の解析目的別留意点であり、筆者による日本語の書籍でも詳述されている³²⁾。

Expression Atlas は、いくつかの主要なリファレンスとなるデータセットについて、Cufflinks 実行結果に相当する baseline 情報（つまり FPKM 値）を保持している。また、DESeq³³⁾ という発現変動解析用 R プログラム実行結果に相当する contrast 情報も調べることができる。一部の読者は、これらのサービスは公共 DB がマイクロアレイデータで満たされていた Gene Expression Atlas 時代とほとんど変わらず、RNA-seq 版になっただけだと思うかもしれない。しかし、これらの解析をエンドユーザが自力で行うのは非常に困難である。マイクロアレイデータであればどんなに大きなプロジェクトでも～数 GB 程度の生データ量であったのに対し、RNA-seq の場合は数百 GB 程度³⁴⁾ にもなる FASTQ 形式の生リードファイルのダウンロードからスタートしなければならない。そしてこの規模は、ノート PC 上に全ファイルを保存することすら困難なレベルである。それゆえ筆者らは、数年前に 18 データセットについて比較的小規模な計算量で済むマッピングからカウントデータ作成までを行って提供した ReCount³⁵⁾ という DB を今でも重宝している。Expression Atlas についても、現状ではその膨大な計算量のため計算済みのデータセットはそれほど多くはないものの、ArrayExpress などの公共 DB に登録されている RNA-seq データがマイクロアレイデータと同じような数値行列形式で提供されればエンドユーザの負担は劇的に軽減する。もちろん baseline 情報や contrast 情報は用いるプログラム次第で結果が変わるものの、Expression Atlas から得られる情報はベンチマークとして利用可能である。このため、自分で他のプログラムを用いて大変な労力をかけてデータ解析する前に、この種の二次 DB で利用可能な情報がないかどうか探するというのが効率的かもしれない。日本でも、DDBJ Read Annotation Pipeline で誰かが解析を行ったプログラムやパラメータ、およびその結果もみんなでも共有できれば、公



```

RGui (64-bit)
ファイル 編集 閲覧 その他 パッケージ ウィンドウ ヘルプ
R Console
> #ファイルに保存
> tmp <- NULL
> tmp <- rbind(tmp, c("Total length (bp)", Total_len))
> tmp <- rbind(tmp, c("Number of contigs", Number_of_contigs))
> tmp <- rbind(tmp, c("Average length", Average_len))
> tmp <- rbind(tmp, c("Median length", Median_len))
> tmp <- rbind(tmp, c("Max length", Max_len))
> tmp <- rbind(tmp, c("Min length", Min_len))
> tmp <- rbind(tmp, c("N50", N50))
> tmp <- rbind(tmp, c("GC content", GC_content))
> write.table(tmp, out_f, sep="\t", append=F, quote=F, row.names=F, col.names=$
> tmp
      [,1]           [,2]
[1,] "Total length (bp)" "2885619"
[2,] "Number of contigs"  "28"
[3,] "Average length"    "103057.821428571"
[4,] "Median length"     "58047"
[5,] "Max length"        "472701"
[6,] "Min length"        "899"
[7,] "N50"               "222389"
[8,] "GC content"        "0.464045193416998"
> |

```

図2. Rでゲノム配列解析。Rコード実行結果のスクリーンショットを示している。

共データを再解析するという手間が軽減されることが期待される。Expression Atlasと似たウェブツールであるRefExを提供しているDBCLSの今後の活動に期待したい。

Rでゲノム解析

最後に、乳酸菌ゲノム配列を読み込んで、総塩基数、コンティグ数、GC含量などの簡単なデータ解析を行うやり方を示す。乳酸菌ゲノムはすでに解読済み³⁶⁾であり、公共DBの一つであるEnsembl³⁷⁾から取得可能である。(Rで)塩基配列解析中では、*Lactobacillus casei* 12A株のFASTA形式ファイル("Lactobacillus_casei_12a.GCA_000309565.1.22.dna.toplevel.fa")を読み込んで上記解析結果をファイル("result_JSLAB1.txt")に保存しているが、ここでは出力ファイルの内容をR画面上に示す(図2)。

原著論文³⁶⁾のTable 1の記載内容(コンティグ数:28、トータル塩基数:2,885,619 bp、%GC:46.4)と同じ結果が得られていることがわかる。また、最長コンティグ(472,701 bp)と最短コンティグ(899 bp)もEnsembl Bacteriaのウェブサイトと同じである。このように、Rでも塩基配列を自在に解析することができる。最低限必要なのは、自分が解析したいファイル名への変更、およびR起動後の作業ディレクトリの変更(つまり解析したいファイルを置いてあるフォルダの指定)のみである。ユーザは、参考ウェブページ中の項目の中から自分が行いたい解析に近いものを探し出し、テンプレートとして利用するだけである。他にもCpG解析、マッピング、発現変動解析、機能解析など様々なデータ解析が可能である。本稿が自力NGS解析の一助になれば幸いである。

参考文献

- 1) Kawano, S., Ono, H., Takagi, T., Bono, H.: Tutorial videos of bioinformatics resources: online distribution trial in Japan named TogoTV. *Brief. Bioinform.*, **13**, 258-268 (2012).
- 2) Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N., Thurston, M.: Open software for biologists: from famine to feast. *Nat. Biotechnol.*, **24**, 801-803 (2006).
- 3) Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I.: ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117-1123 (2009).
- 4) Zerbino, D.R. and Birney, E.: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821-829 (2008).
- 5) Langmead, B. and Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357-359 (2012).
- 6) Li, H. and Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760 (2009).
- 7) Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A.,

- Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511-515 (2010).
- 8) Bailey, T.L., Williams, N., Misleh, C., Li, W.W.: MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369-373 (2006).
 - 9) Katoh, K. and Standley, D.M.: MAFFT: iterative refinement and additional methods. *Methods Mol. Biol.*, **1079**, 131-146 (2014).
 - 10) Magis, C., Taly, J.F., Bussotti, G., Chang, J.M., Di Tommaso, P., Erb, I., Espinosa-Carrasco, J., Notredame, C.: T-Coffee: Tree-based consistency objective function for alignment evaluation. *Methods Mol. Biol.*, **1079**, 117-129 (2014).
 - 11) Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E.: WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188-1190 (2004).
 - 12) Schneider, T.D. and Stephens, R.M.: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097-6100 (1990).
 - 13) Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498-2504 (2003).
 - 14) Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410 (1990).
 - 15) R Development Core Team: R: A language and environment for statistical computing. In R Foundation for Statistical Computing. Vienna, Australia (2010).
 - 16) Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80 (2004).
 - 17) 門田幸二: 1.4 R および各種パッケージのインストール, p.27-31, シリーズ Useful R 第7巻 トランスクリプトーム解析, 金明哲 編, 共立出版, 東京 (2014).
 - 18) Pearson, W.R. and Lipman, D.J.: Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U S A*, **85**, 2444-2448 (1988).
 - 19) Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., Itoh, T.: Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.*, in press.
 - 20) Knowles, D.G., Roder, M., Merkel, A., Guigo, R.: Grape RNA-Seq analysis pipeline environment. *Bioinformatics*, **29**, 614-621 (2013).
 - 21) Nagasaki, H., Mochizuki, T., Kodama, Y., Saruhashi, S., Morizaki, S., Sugawara, H., Ohyanagi, H., Kurata, N., Okubo, K., Tagagi, T., Kaminuma, E., Nakamura, Y.: DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. *DNA Res.*, **20**, 383-390 (2013).
 - 22) Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedmann, N., Regev, A.: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644-652 (2011).
 - 23) Goecks, J., Nekrutenko, A., Taylor, J.; Galaxy Team: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86 (2010).
 - 24) Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N.A., Gonzalez-Porta, M., Hastings, E., Huber, W., Jupp, S., Keays, M., Kryvykh, N., McMurry, J., Marioni, J.C., Malone, J., Megy, K., Rustici, G., Tang, A.Y., Taubert, J., Williams, E., Mannion, O., Parkinson, H.E., Brazma, A.: Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42** (Database issue), D926-D932 (2014).
 - 25) Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., Holloway, E., Klebanov, A., Kryvykh, N., Kurbatova, N., Kurnosov, P., Malone, J., Melnichuk, O., Petryszak, R., Pultsin, N., Rustici, G., Tikhonov, A., Travillian, R.S., Williams, E., Zorin, A., Parkinson, H., Brazma, A.: Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **40** (Database issue), D1077-D1081 (2012).
 - 26) Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., Kurbatova, N., Malone, J., Mani, R., Mupo, A., Pedro Pereira, R., Pilicheva, E., Rung, J., Sharma, A., Tang, Y.A., Ternent, T., Tikhonov, A., Welter, D., Williams, E., Brazma, A., Parkinson, H., Sarkans, U.: ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res.*, **41** (Database issue), D987-D990 (2013).
 - 27) Furney, S.J., Pedersen, M., Gentien, D., Dumont, A.G., Rapinat, A., Desjardins, L., Turajlic, S., Piperno-Nemann, S., de la Grange, P., Roman-Roman, S., Stern, M.H., Marais, R.: SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer Discov.*, **3**, 1122-1129 (2013).
 - 28) Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511-515 (2010).
 - 29) Burge, C. and Karlin, S.: Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78-94 (1997).
 - 30) Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., Robinson, M.D.: Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.*, **8**, 1765-1786 (2013).
 - 31) Sun, J., Nishiyama, T., Shimizu, K., Kadota, K.: TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics*, **14**, 219 (2013).
 - 32) 門田幸二: 3.3.1 解析目的別留意点, p.129-132, シリーズ Useful R 第7巻 トランスクリプトーム解析, 金明哲 編, 共立出版, 東京 (2014).
 - 33) Anders, S. and Huber, W.: Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106 (2010).
 - 34) Yu, Y., Fuscoe, J.C., Zhao, C., Guo, C., Jia, M., Qing, T., Bannon, D.I., Lanchashire, L., Bao, W., Du, T., Luo, H., Su, Z., Jones, W.D., Moland, C.L., Branham, W.S., Qian, F., Ning, B., Li, Y., Hong, H., Guo, L., Mei, N., Shi, T., Wang, K.Y., Wolfinger, R.D., Nikolsky, Y., Walker, S.J., Duerksen-Hughes, P., Mason, C.E., Tong, W., Thierry-Mieg, J., Thierry-Mieg, D., Shi, L., Wang, C.: A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4

- developmental stages. *Nat. Commun.*, **5**, 3230 (2014).
- 35) Frazee, A.C., Langmead, B., Leek, J.T.: ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, **12**, 449 (2011).
- 36) Broadbent, J.R., Neeno-Eckwall, E.C., Stahl, B., Tandee, K., Cai, H., Morovic, W., Horvath, P., Heidenreich, J., Perna, N.T., Barrangou, R., Steele, J.L.: Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation. *BMC Genomics*, **13**, 533 (2012).
- 37) Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Kulesha, E., Martin, F.J., Maurel, T., McLaren, W.M., Murphy, D.N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S.J., Vullo, A., Wilder, S.P., Wilson, M., Zadissa, A., Aken, B.L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T.J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D.R., Searle, S.M.: Ensembl 2014. *Nucleic Acids Res.*, **42** (Database issue), D749–D755 (2014).

Methods for analyzing next-generation sequencing data

I. Introduction

Koji Kadota¹, Jianqiang Sun², Min Tang², Tasuku Nishioka¹
and Kentaro Shimizu^{1, 2}

¹*Agricultural Bioinformatics Research Unit,*

²*Department of Biotechnology, Graduate School of Agricultural and Life Sciences,
The University of Tokyo.*

Abstract

Next-generation sequencing (NGS) technology is a fundamental means of studying genome, transcriptome, and microbiome. We can analyze NGS data of bacteria such as *Lactobacillus* using laptop computer nowadays. However, for many non-bioinformaticians, it is difficult to construct an environment for analyzing NGS data. In this review, we describe educational programs, websites, application softwares, programming languages, and webtools. An example of genome analysis of *Lactobacillus casei* 12A is also provided.