Japanese Journal of Lactic Acid Bacteria

解　説

次世代シーケンサーデータの解析手法 第15回 RNA-seq解析（その3）

寺田 朋子[1]、清水 謙多郎[1,2]、門田 幸二[1,2*]

[1]東京大学 大学院農学生命科学研究科
[2]東京大学 微生物科学イノベーション連携研究機構

Galaxyは、ウェブブラウザ上でマウスを操作して行うGUIベースのデータ解析環境である。今回も前回に引き続いて、Galaxy上で行うRNA-seqデータ（GSE107337）の発現定量に関する解説を行う。まず、アノテーション情報を含むGFFファイルの前処理（フィルタリング）を行い、遺伝子領域に対応するゲノム中の塩基配列情報を抽出する。次に、得られた塩基配列群をリファレンス配列としてKallisto quantプログラムを実行し、遺伝子ごとのカウント値や発現量に相当するTPM値を得る。カウント値と周辺情報からCPM、CPK、FPKM、そしてTPMといった様々な補正値を導き出す考え方について述べる。最後に、今回得られた結果を、GSE107337の原著論文および第14回で得られたものと比較・検証する。ウェブサイト（Rで）塩基配列解析のサブ（URL: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq2.html）中に本連載をまとめた項目（URL: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq2.html#about_book_JSLAB）が存在する。ウェブ資料（以下、W）や関連ウェブサイトなどを効率的に活用してほしい。

Key words：NGS, Galaxy, RNA-seq, *Lactobacillus rhamnosus* GG

はじめに

第14回[1]に引き続き、*Lactobacillus rhamnosus* GG[2]のRNA-seqデータ[3]をGalaxy[4]上で取り扱う。前回はマッピング時のリファレンスとしてゲノム配列を用いたが、トランスクリプトーム配列にマップして高速にカウントデータを得るやり方も存在する。これは主にヒトやマウスなどの高等生物を対象とした戦略ではあるが、ゲノムサイズの小さい乳酸菌にも適用可能であることを示す。本稿では、事前準備（ゲノム配列とアノテーションファイルを用いたトランスクリプトーム配列取得）や後処理の解説を通じたGalaxyのスキルアップ、そして解析結果として得られるカウントデータ・配列長情報・補正後のTPM値[5]の導出を通じたRNA-seqデータの正規化についても議論する。

*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

今回のスタート地点は、前回（第14回）のW9-7のGalaxy main画面（ヒストリー名：GSE107337_3samples）である［W1］。しかし、必ずしも前回の内容を一通り行っておく必要はない。第12回[6]でも述べたように、ヒストリーの共有を我々にリクエストすればよいからである。ここでは新規ヒストリー（ヒストリー名：trans_map）を作成し［W2］、第12回のW13-3でも解説したヒストリー間のデータコピーによって事前準備を行う。具体的には、ゲノム配列ファイル（ASM2650v1.fa）、アノテーションファイル（ASM26501v1.gff3）、そしてTrimmomatic[7]実行後の計6個のFASTQファイル（SRR6322564, SRR6322567, and SRR6322569の計3サンプル分）をヒストリーtrans_map上に配置して以降の作業を行う［W3］。一通りのコピー作業が完了すれば、通常の3画面からなる状態に戻して解析準備完了である［W4］。

GFFファイルの前処理

図1は、解析準備が完了した、①ヒストリーtrans_mapのGalaxy画面である［W5］。中央パネルには、②ア

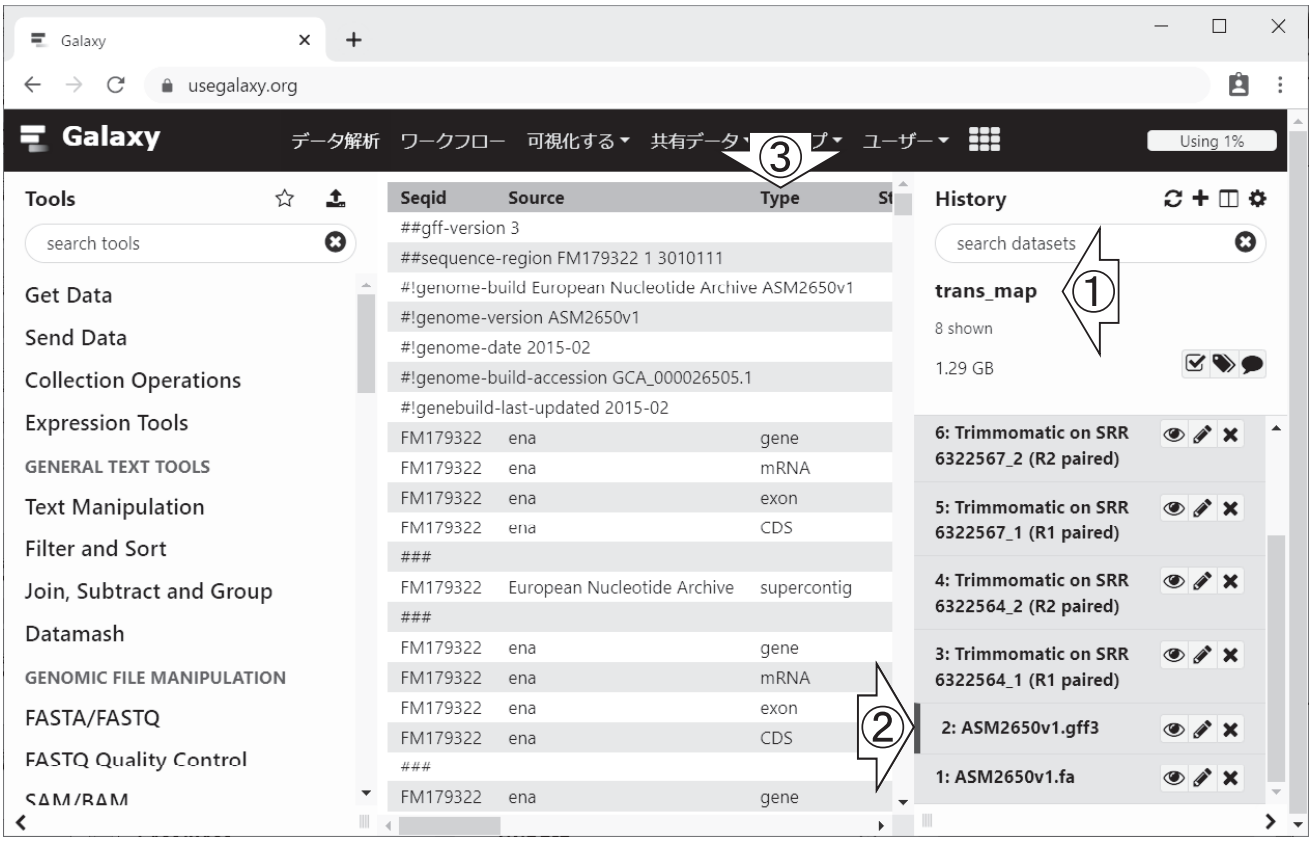

図 1. 解析準備完了後の Galaxy 画面
①ヒストリー名は trans_map。右側のヒストリーパネルには、ゲノム配列ファイル（ヒストリー 1）、アノテーションファイル（ヒストリー 2）、そして Trimmomatic 実行後の paired-end RNA-seq リードファイル（計 6 ファイル；ヒストリー 3-8）が含まれている。中央パネルには、② GFF ファイルの中身が表示されている。③ Type 列（3 列目に相当）には、様々な種類の features が含まれていることがわかる。

ノテーションファイルの中身が表示されている。今回は、このファイル中の gene 領域を抽出してリファレンス配列（マップされる側の配列）として利用する。理由については次項で述べるが、Galaxy 上で最初に行う作業は、この GFF ファイルを入力とした gene 行情報のみの抽出（他の features 行情報の除外）である。この作業は、Galaxy 画面左側のツール選択パネルの **Filter and Sort** というカテゴリに含まれる、**Extract features** というプログラムを用いて行う［W6-1］。

中央パネルの Extract features 操作画面上では、まず Select GFF data で、入力として与える GFF ファイルを指定する［W6-3］。次に、どの列の情報を用いて目的の features の行を抽出するかを指定すべく、「3 列目が gene の行を抽出したい」ということを認識させる。ここではまず 3 列目がターゲットであることを指定すべく、From のところを、デフォルトの「Column 1 / Sequence name」から、「Column 3 / Feature」に変更する［W6-5］。次に、指定した列の中から抽出したい行情報を指定すべく、Extract features のところで「gene」を選択する［W6-6］。Execute ボタンを押すと、数分程度で実行結果がヒストリー 9（Extract features on data 2）に作成される［W6-10］。**図 2** は、④ Extract features 実行結果を⑤中央パネル上に表示した Galaxy 画面である。**図 1** では Type 列（3 列目に相当）に様々な features が見られるが、**図 2** では⑥ Type 列が gene のみになっていることがわかる［W6-12］。この gene 領域の座標情報のみからなるヒストリー 9 が、ヒストリー 1 のゲノム配列ファイルと合わせてトランスクリプトーム配列取得に利用される。

トランスクリプトーム配列取得

前項の前処理の必要性をどのようにして認識したのかについて、実際の作業と戦略立案の関係性を述べる。まず我々は、ゲノム上の座標情報を取り扱う有名なプログラムとして BEDTools [8] が存在することを経験的に知っている。BEDTools は、ゲノム上の座標を表す BED（Browser Extensible Data）という形式ファイルを取り扱うためのプログラム群（なので Tools）を 1 つのパッケージとしてまとめたものである。今手元にあるのは BED 形式ファイルではなく GFF 形式ファイルではあるものの、ファイル形

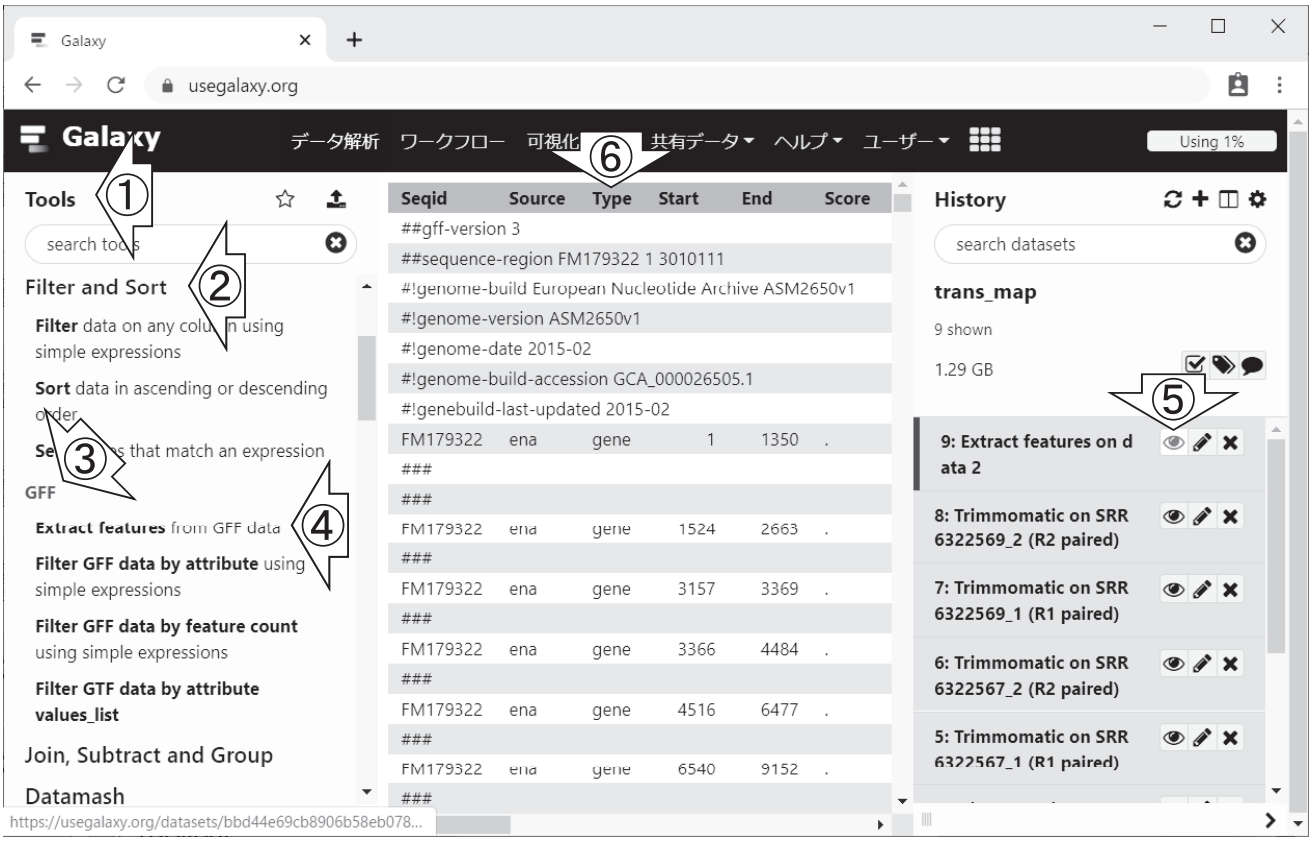

図 2. Extract features 実行後の Galaxy 画面
① Tools パネルの、② Filter and Sort カテゴリ内にある、③ GFF の、④ Extract features プログラム実行後の Galaxy 画面。⑤で中央パネル上にプログラム実行結果を表示させている。⑥ Type 列（3 列目に相当）には、gene のみの feature が含まれていることがわかる。

式の変換（BED ⇔ GFF）を行うプログラムは Galaxy 上で提供されているだろうという想定のもと、Galaxy のツール選択パネル上で BEDTools を探すところから実際の作業を開始した。そして、ツール選択パネルで BED というカテゴリを発見し［W7-1］、その中にリストアップされているプログラム群を眺めることで、**bedtools GetFastaBed** というプログラムが目的を達成する上で最も近いものであると判断した［W7-2］。中央パネル上でこのプログラムの操作画面を眺め、入力形式として BED 以外に GFF も取り扱えると判断した［W7-4］。しかしながら、ヒストリー 2 に相当するオリジナルの GFF ファイル（ASM26501v1.gff3）を入力としても gene 領域の塩基配列情報を取得できないことを実体験し、試行錯誤を経て前項の Extract features プログラムの利用に至った。以降では、前処理後のヒストリー 9 のデータを入力とするが、ヒストリー 2 を指定してエラーが出ることを確認してもよいだろう［W7-5］。

前述したように、bedtools GetFastaBed プログラムの入力は、ヒストリー 9 の前処理後の GFF ファイルとヒストリー 1 のゲノム配列ファイルである。出力は、gene ごとの座標領域の塩基配列を抽出した multi-FASTA ファイルであり、ヒストリー 10 に作成される［W7-10］。**図 3** は、bedtools GetFastaBed 実行結果を中央パネル上に表示した Galaxy 画面である［W7-11］。この multi-FASTA ファイル中の配列数は 2,949 個であり［W7-12］、第 14 回の W9-8 で得られたカウント行列の行数と一致する。前回と同じ feature（GFF ファイル中の gene 情報）の塩基配列情報をリファレンスとして利用することで、次項で得られる数値行列との比較が可能となる。

発現定量

トランスクリプトーム配列をリファレンスとして発現定量を行う代表的なプログラムとしては、Kallisto [9] や Salmon [10] が挙げられる。これらは、大まかには「マッピングからカウント情報取得や RPKM [11] や TPM [5] のような補正後の発現量情報の取得」までを 1 つのプログラム内で行うものだと解釈すればよい。今回は、ツール選択パネル上で RNA-seq というカテゴリ内に存在する **Kallisto quant** というプログラムの利用例を紹介する［W8-1］。

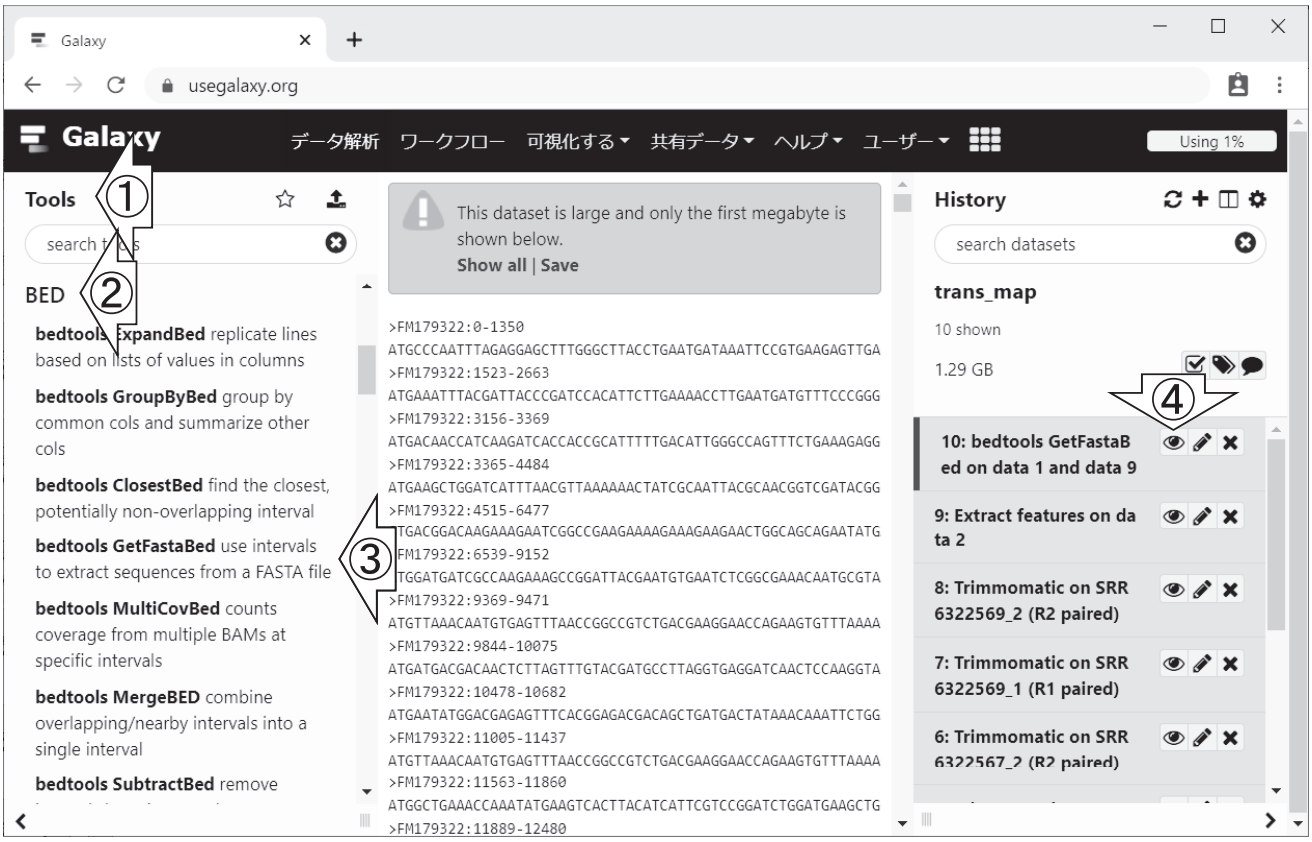

図 3. bedtools GetFastaBed 実行後の Galaxy 画面
① Tools パネルの、② BED カテゴリ内にある、③ bedtools GetFastaBed プログラム実行後の Galaxy 画面。④で中央パネル上にプログラム実行結果（multi-FASTA ファイルの中身）を表示させている。どの塩基配列も ATG から始まっており妥当と判断できる。

中央パネルの Kallisto quant 操作画面上では、まずリファレンスとして利用するトランスクリプトーム配列の指定を行う。デフォルトでは、Galaxy 内で用意されているトランスクリプトーム配列を利用する（Use a built-in transcriptome）オプションになっているため、これをヒストリーから選択できる（Use a transcriptome from history）オプションに変更する［W8-2］。変更後、FASTA reference transcriptome の候補としてヒストリー 10（デフォルト）とヒストリー 1 がリストアップされるので、目的のトランスクリプトーム配列情報からなるヒストリー 10 を選択する。次に、Trimmomatic 実行結果であるリードデータが paired-end であることを認識させ［W8-3］、forward 側と reverse 側に相当するヒストリー情報をそれぞれ指定して Kallisto quant を実行する［W8-6］。

後にこの解析結果を入力として sleuth [12] というプログラムを用いて発現変動解析まで行う際に重要になる事柄として、Kallisto quant 実行前に眺めたオプションのところに存在する Number of bootstrap samples という項目（デフォルトは 0）について少し触れておく［W8-6］。まず、Kallisto はヒトやマウスを対象とした転写物ごとの発現定量を目的としている。これらの生物種では、異なる転写物間で同じエクソンが共有されうる。このため、このような共有エクソン（shared exon）上にマップされたリード群（正確には k-mer と呼ばれるリード由来の部分配列）をどの転写物に対してどれだけ分配するかという問題に取り組む必要がある。Kallisto は、ブートストラップ（bootstrap）と呼ばれるテクニックを用いて信頼度の評価を行う枠組みを提供している。このテクニックは、データをリサンプリング（re-sampling）して同じ解析を行う作業を繰り返すため、一般に**繰り返しの回数を増やすほど定量結果のばらつきの推定精度は向上する。しかしながら、その分だけ計算時間が増える**だけでなく、試行ごとの計算結果の情報も増えていくことを正しく認識しておく必要がある。尚、W8-6 では Number of bootstrap samples オプションを変更せずに 0 のままで実行したので、定量結果の推定精度に関する情報は持たない。

今回はリファレンス配列と paired-end の 3 サンプル分のリードデータを与えたので、入力は計 7 つであった

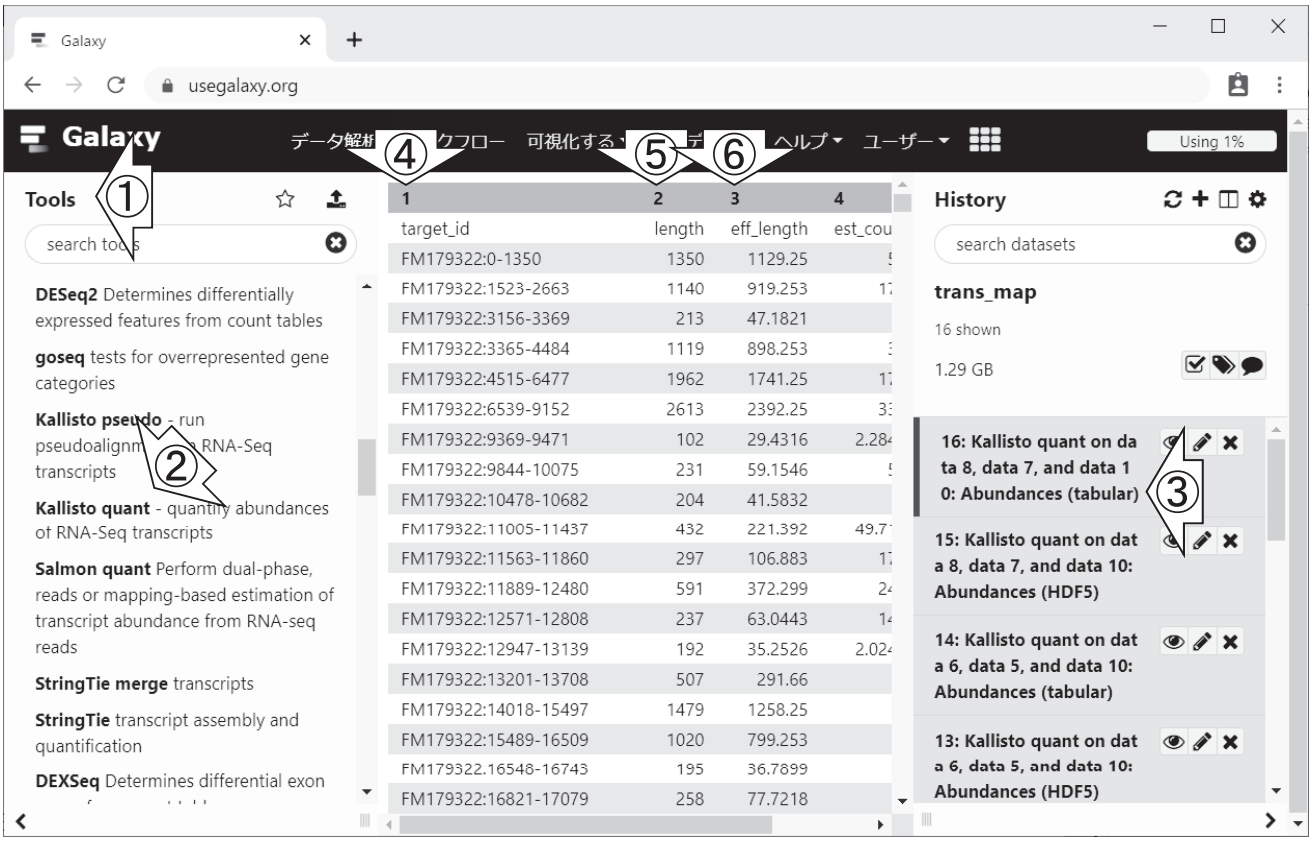

図 4. Kallisto quant 実行後の Galaxy 画面

① Tools パネルの RNA-seq カテゴリ内にある、② Kallisto quant プログラム実行後の Galaxy 画面。③ tabular の実行結果を中央パネル上に表示させている。④ 1 列目（target_id）は配列名、⑤ 2 列目はリファレンス配列（この場合は gene）の長さ、⑥ 3 列目は有効配列長（effective length）、4 列目（est_counts）はカウントの推定値［W10-5］、そして 5 列目（tpm）は TPM と呼ばれる補正後の発現量［W11-1］である。

［W8-7］。このときは 10 分足らずで計算が終了し、計 6 個の出力（ヒストリー 11 ～ 16）が得られた［W8-8］。サンプルごとに 2 種類の結果（tabular 形式と HDF5 形式）が出力され、tabular はタブ区切りテキストファイル（拡張子は .tabular）、そして HDF5（Hierarchical Data Format 5）は HDFView などの専用のビューワでしか見られないバイナリファイル（拡張子は .h5）である［W8-10］。この HDF5 ファイルのみ、定量結果の推定精度に関する情報（bootstrap estimates）を含んでいる。

もちろん我々は、Galaxy 上で得られる情報のみから Kallisto の全貌を把握しているわけではない。Kallisto のマニュアルも合わせて読むことで、注意すべき点や多少の違いを気にしなくてもよい点を学習している［W9-1］。例えば、我々は今回初めて HDF5 形式を知ったが、Kallisto マニュアル中の「abundances.h5 is a HDF5 binary file …」を眺めることで、これがバイナリファイルだと認識した［W9-2］。また、「This file can be read in by sleuth.」という記述から、sleuth を用いた発現変動解析を行いたいときに入力として利用するものだと判断した。同時に、今回ブートストラップは行っていないので［W8-6］、sleuth の原著論文タイトルにもある read mapping uncertainty の情報を現段階では持っていない。それゆえ、今回の場合は実質的に .h5 ファイルは無用であり、.tsv ファイルのみを眺めればよいと判断した［W9-2］。尚、tsv は tab-separated values の略であり、タブ区切りテキストファイルであることを明示した拡張子である。第 8 回[13] の DFAST[14] 実行結果中にも出現しており、第 9 回[15] のところでも簡単に触れてはいるものの、意外に知らないヒトが多いので今回改めて解説した。従って、Galaxy からダウンロードしたファイルの拡張子は .tabular になってはいるものの、.tsv や .txt に拡張子を変更しても問題はない。

定量結果の解説

Kallisto quant 実行結果の 1 つである、タブ区切りテキストファイル（ヒストリー 12, 14, and 16）の中身を解説する［W10-1］。**図 4** は、ヒストリー 16 の中身を中央パネル上に表示させたものである。ダウンロードしたファイルを Excel

で眺めるとより分かりやすいが、このデータは（ヘッダー行を除くと）2,949 行×5 列からなる。④1 列目（target_id 列）は配列名である。ヒストリー 10 の中身を中央パネルで表示させた、multi-FASTA ファイルの description 情報に相当する［W7-11］。ヒストリー 10 は、ヒストリー 1 のゲノム配列ファイル（ASM2650v1.fa）とヒストリー 9 の gene 領域の座標情報を入力として、gene 領域ごとの塩基配列を抜き出したものである。このゲノムの Genbank accession 番号は FM179322 であり、ASM2650v1.fa 中の 1 行目にも記載されている。一部の読者は、例えばヒストリー 16 の（ヘッダー行を除く）2 行目の配列名が「FM179322:1523-2663」であり、ヒストリー 9 の対応する座標は終点が 2663 と一致しているものの、始点が 1524 となっている点に違和感を持つかもしれない。しかし、これはゲノムの座標情報から配列長を算出する際の利便性に関する些細な事柄である。これの真の配列長は 1140 塩基である。そして、ヒストリー 9 の始点情報を用いて計算すると 2663－1524＋1＝1140 とせねばならないが、ヒストリー 16 の配列名のように予め始点側の数値を－1 しておけば、終点－始点＝2663－1523＝1140 と単純化できる。数値の違いはほぼないに等しいため実害を被ることは実質的にはないが、注意しておくに越したことはない。

ヒストリー 16 の 2 列目（length 列）は転写物配列の長さ $l_{original}$、そして 3 列目（eff_length 列）はその有効長（effective length; $l_{effective}$）である［W10-3］。配列によって多少のばらつきはあるものの、概ね有効長のほうが 200 塩基ほど短くなっている（つまり $l_{original}-l_{effective} \fallingdotseq 200$ だということ）ことがわかる。有効長の計算手順を完全に理解できるほどの情報がマニュアルに記されているわけではないが、この約 200 塩基という数値はマップされた paired-end リードの平均フラグメント長（mean of the fragment length distribution）、つまりマニュアル中の μ_{FLD} に相当する［W10-4］。フラグメント長という表現が難解な読者は、まず single-end リードで考えるとよい。長さが全て一定の single-end リードで Kallisto quant を実行した場合で考えると、このリード長が μ_{FLD} に相当する。マニュアル中の有効長を表現する式に対応させて書くと $l_{effective}=l_{original}-\mu_{FLD}+1$ となる。これの意味するところは、「$l_{original}$ という長さをもつ転写物配列の中から、μ_{FLD} という長さのリードを何個生成可能か考えたときに、その答えは $l_{original}-\mu_{FLD}+1$ という式で表すことができる」ということと本質的に同じである。例えば、「TCATGGAATG という転写物配列（$l_{original}=10$）の中から 5 塩基の長さ（$\mu_{FLD}=5$）をもつリードを何個生成可能か」という問題で考えた場合、$l_{effective}=l_{original}-\mu_{FLD}+1=10-5+1=6$ が答えとなる。つまり、6 個のリード（TCATG, CATGG, ATGGA, TGGAA, GGAAT, and GAATG）を生成可能ということである。自由度（degree of freedom）と似た概念だと思えばよいだろう。

実際には、このデータの場合は Trimmomatic によって前処理済みであるため、リード長は一定ではない。また、一定の割合で同一配列をもつリードが生成されうるため、ユニークなリードの個数（つまり種類数）で考えると若干減る。例えば、上記転写物配列の中から 3 塩基の長さ（$\mu_{FLD}=3$）をもつリードを何個生成可能か」という問題で考えた場合、$l_{effective}=l_{original}-\mu_{FLD}+1=10-3+1=8$ 個生成可能ということにはなる。しかし、ユニークなリードの個数で考えた場合、7 種類（TCA, CAT, ATG, TGG, GGA, GAA, and AAT）となり、このような事柄も実際には考慮せねばならない。これは第 6 回[16]で述べた *k*-mer 解析によるゲノムサイズ推定の戦略と基本的に同じである。第 6 回では $\mu_{FLD}=251$ 塩基の長さをもつリード群から、141 塩基の長さをもつリードの部分配列（つまり *k*-mer with *k*＝141）を生成し、得られた *k*-mer の種類数を推定ゲノムサイズとしている（実際には出現頻度が低いものを除いている）。この例のように *k*-mer の長さが一定以上あれば、*k*-mer の種類数をゲノムサイズの推定値として利用可能である。

話を元に戻す。ヒストリー 16 の 4 列目（est_counts 列）はカウント数の推定値（estimated counts）である。これは、single-end の場合は転写物配列上にマップされたリード数、paired-end の場合はマップされたフラグメント数に相当する。“estimated” がついているのは、Kallisto が実際にはリードそのものではなく、*k*-mer に分割したもので疑似的なマッピングを行っていることに由来するためだと解釈すればよい。

CPM, CPK, FPKM, and TPM

ヒストリー 16 の 5 列目（tpm）は、いわゆる TPM（Transcripts per million）値[5]と呼ばれるものである［W11-1］。マニュアル中の記述に従うと、解析サンプルである RNA pool 中の転写物の割合を測定した値（a measurement of the proportion of transcripts in your pool of RNA）である［W11-2］。4 列目（est_counts）はカウント情報であり、発現情報ではない。5 列目の数値データ（TPM 値）が、一般に発現データと呼ばれるものに相当する。2014 年出版の関連書籍[17]では（CPM/RPM, CPK/RPK, and FPKM/RPKM には言及しているものの）TPM に言及していないが、2019 年出版の書籍[18]では頻繁に出力結果の一部として登場していることがわかる。以下では、カウント値から周辺情報を用いて TPM 値を算出する手順を示す。具体的には、入力として 3-4 列目の情報のみを用い、5 列目の情報を出力として得る。

例として、（ヘッダー行を除く）2 行目の転写物である FM179322:1523-2663 のカウント値（est_counts 列の 1703 という数値）から、CPM・CPK・FPKM の算出を経て最終的に 306.542 という TPM 値を導く［W11-3］。まず CPM

(Counts Per Million) は、サンプル間で総カウント数の違いを補正することを目的としている。具体的には、Per Million からも予想できるが「総カウント数を 100 万 (one million) に揃える」だけである。このサンプル (ヒストリー 16 の SRR6322569) のカウント数の総和は 1,758,959 であり、4 列目の数値を全て足したものに相当する [W11-4]。CPM という補正後の値にする作業は、以下のように行う：

$$CPM = \text{カウント値} \times \frac{1000000}{\text{総カウント数}} = 1703 \times \frac{1000000}{1758959} = 968.19$$

解釈としては、例えば「このサンプルの総カウント数は 100 万より 1.759 倍ほど高いので、カウント値を 1.759 で割ってやれば、総カウント数が 100 万だったときのカウント数になる」のように理解してもよいだろう。全遺伝子に対して同じ係数が掛かる (1.759 で割る) ので、このサンプルにおける CPM 値と元のカウント値の関係は以下のようになる：

$$1.759 \times CPM = \text{カウント値}$$

もちろん係数はサンプルごとに異なる。Galaxy 上で実行済みの他のサンプルとして、ヒストリー 14 (SRR6322567) とヒストリー 12 (SRR6322564) の総カウント数は、それぞれ 1,176,813 と 1,072,837 である。これは、両サンプルともに、CPM 値を得る際に掛ける係数がヒストリー 16 のものに比べてより 1 に近いことを意味する。特に、ヒストリー 12 のサンプルにおいては、CPM 値が元のカウント値よりもほんのわずかに小さくなる (1.073 で割る) 程度だということである。ここまでが、連載第 13 回 [19] の最後のほうで述べた「同一遺伝子の発現レベルの大小関係を異なるサンプル間で大まかに比較したい場合」に行う作業に相当する。

次に、「同一サンプル内で異なる遺伝子間の発現レベルの大小関係を大まかに比較したい場合」に行う CPK 補正について述べる。これは、遺伝子の長さ情報を用いた補正に相当するものであり、Kallisto quant 上では 3 列目の eff_length 情報 ($l_{effective}$) を用いて TPM 値が算出されている [W11-5]。CPK は、Counts Per Kilobase の略であり、作業的には「$l_{effective}$ を 1,000 (one Kilobase) に揃える」だけである。CPK という補正後の値にする作業は、FM179322:1523-2663 のカウント値である 1,703 と、$l_{effective}$ =919.253 の情報を利用して以下のように行う：

$$CPK = \text{カウント値} \times \frac{1{,}000}{l_{effective}} = 1{,}703 \times \frac{1{,}000}{919.253} = 1{,}852.591$$

FPKM は、CPM と CPK の両方の係数を組み合わせることで、サンプル間補正と遺伝子間補正の両方を同時に行うことを目的とした補正値である。FM179322:1523-2663 の FPKM 値は、以下のように計算する：

$$FPKM = \text{カウント値} \times \frac{1{,}000{,}000}{\text{総カウント数}} \times \frac{1{,}000}{l_{effective}}$$
$$= 1{,}703 \times \frac{1{,}000{,}000}{1{,}758{,}959} \times \frac{1{,}000}{919.253} = 1{,}053.232$$

$l_{effective}$ の値は遺伝子ごとに異なるため、当然ながらカウント値に掛ける係数も遺伝子ごとに異なる。これは FPKM の総和 (または平均値) がサンプルごとに異なるという効果をもたらす。実際、ヒストリー 16 (SRR6322569)、14 (SRR6322567)、12 (SRR6322564) の FPKM 値の総和は、それぞれ 3,435,847.75、4,005,980.92、5,275,181.61 となる [W11-6 ～ W11-8]。この値 (FPKM 値の総和) が 100 万に揃うように補正を行ったものが TPM である。FM179322:1523-2663 の TPM 値は、以下のように計算する：

$$TPM = FPKM\text{値} \times \frac{1{,}000{,}000}{FPKM\text{の総和}}$$
$$= 1{,}053.232 \times \frac{1{,}000{,}000}{3{,}435{,}847.75} = 306.54$$

TPM 値を算出する他の考え方として、「CPK 値の総和を 100 万に揃えた補正を行えばサンプル間で比較可能になる」でもよい。具体的には、FM179322:1523-2663 の CPK 値 (=1,852.591) と、ヒストリー 16 (SRR6322569) の CPK 値の総和 (=6,043,515.33 [W11-6]) を用いて、以下のように計算する：

$$TPM = CPK\text{値} \times \frac{1{,}000{,}000}{CPK\text{の総和}}$$
$$= 1{,}852.591 \times \frac{1{,}000{,}000}{6{,}043{,}515.33} = 306.54$$

補足情報として、$l_{effective}$ の値はサンプル間で一定ではなく、若干異なる点にも注意してほしい。つまり、ヒストリー 16, 14, 12 の eff_length 列の数値が一定ではないということである。これはおそらく、リファレンストランスクリプトーム配列に RNA-seq リード (の部分配列) をマップした際に、複数配列にマップされる状況がサンプルごとに若干異なるためであろう。

原著論文と比較

第 14 回で議論した 9 つの遺伝子群の倍率変化および $\log_2$ (倍率変化) との値と比較すべく、我々も改めて全サンプル (計 9 サンプル) について同様の手順で Kallisto quant を実行した [W12-1]。**表 1** は、②各群 3 反復のカウントデータをマージし [W12-2]、③倍率変化と④ $\log_2$ (倍率変

表 1. 原著論文提供情報との比較結果

（a）原著論文提供のカウント情報、（b）第 14 回のパイプライン（bowtie2 → htseq-count）で得たカウント情報、（c）今回のパイプライン（Kallisto quant）で得たカウント情報。（a）と（b）は、第 14 回で述べた結果をまとめたものに相当する。①遺伝子名、②カウント情報、③倍率変化、④ $\log_2$（倍率変化）。（c）の遺伝子名は、リファレンスゲノム上の遺伝子領域（始点 _ 終点）で示している。例えば、LGG_02372 の遺伝子領域は［2446836, 2449787］と解釈する。

① 遺伝子名	② pH4_1h	② pH4_24h	② pH7_CCG	③ 24h/C	③ 1h/C	④ $\log_2$(24h/C)	④ $\log_2$(1h/C)
(a) 原著論文提供**のカウント情報**							
LGG_02240	1106	2444	422	5.7915	2.6209	2.5339	1.3900
LGG_02372	33	132	13	10.1538	2.5385	3.3440	1.3440
LGG_00113	176	277	1159	0.2390	0.1519	-2.0649	-2.7192
LGG_00115	136	296	860	0.3442	0.1581	-1.5387	-2.6607
LGG_00108	92	351	912	0.3849	0.1009	-1.3776	-3.3093
LGG_00109	119	402	1615	0.2489	0.0737	-2.0063	-3.7625
LGG_01064	27	268	11	24.3636	2.4545	4.6067	1.2955
LGG_02032	86	207	617	0.3355	0.1394	-1.5756	-2.8429
LGG_00418	156	443	1710	0.2591	0.0912	-1.9486	-3.4544
(b) 第14回で得た**カウント情報**							
LGG_02240	1231	2914	490	5.9469	2.5122	2.5721	1.3290
LGG_02372	56	188	15	12.5333	3.7333	3.6477	1.9005
LGG_00113	224	368	1561	0.2357	0.1435	-2.0847	-2.8009
LGG_00115	186	409	1192	0.3431	0.1560	-1.5432	-2.6800
LGG_00108	120	490	1226	0.3997	0.0979	-1.3231	-3.3529
LGG_00109	147	554	2275	0.2435	0.0646	-2.0379	-3.9520
LGG_01064	49	389	18	21.6111	2.7222	4.4337	1.4448
LGG_02032	116	290	831	0.3490	0.1396	-1.5188	-2.8407
LGG_00418	222	609	2233	0.2727	0.0994	-1.8745	-3.3304
(c) 第15回で得た**カウント情報**							
2310404_2310685	2790	6613	1109	5.9630	2.5158	2.5760	1.3310
2446836_2449787	96	375	34	11.0294	2.8235	3.4633	1.4975
120031_121410	469	774	3262	0.2373	0.1438	-2.0754	-2.7981
122428_123486	391	838	2448	0.3423	0.1597	-1.5466	-2.6464
113614_114381	262	1045	2728	0.3831	0.0960	-1.3843	-3.3802
114378_115283	319	1104	4160	0.2654	0.0767	-1.9138	-3.7050
1076321_1077787	81	764	27	28.2963	3.0000	4.8225	1.5850
2081384_2083315	236	595	1737	0.3425	0.1359	-1.5456	-2.8797
420195_420905	404	1159	4420	0.2622	0.0914	-1.9312	-3.4516

化）を算出した結果である［W13-1］。（a）原著論文提供のカウント情報と（b）第 14 回の手順で得られたカウント情報に基づく結果は、第 14 回の原稿中で述べた事柄をまとめたものに相当する。（c）が今回得られた Kallisto quant 実行結果の中から、①対応する遺伝子領域と②そのカウント情報を抽出した結果である。（c）のカウント値は全体的に大きいものの、その③倍率変化と④ $\log_2$（倍率変化）の傾向は、（a）や（b）とよく似ていることがわかる。これはおそらく、乳酸菌を含むバクテリアは、ほとんどの場合 1 つの遺伝子領域から転写されるものが 1 種類（遺伝子≒転写物）だからであろう。

おわりに

今回は、Trimmomatic で前処理した RNA-seq リードに対して、トランスクリプトーム配列をリファレンスとして Kallisto quant を実行し、カウント情報および TPM 値を得た。Galaxy 上での操作感は、第 14 回の W7 で紹介した Bowtie2 によるマッピングとほぼ同じである。作業の基本形を一度覚えておけば、Kallisto のような比較的最近開発されたプログラムであっても、Galaxy 上で簡単に実行できるのが非 Linux ユーザにとっての一番のメリットと言えるだろう。

乳酸菌 RNA-seq データの原著論文[3] の Supplement では、RPKM 値のファイル（GSE107337_RPKM.csv）が提供されている。なぜ TPM 情報を提供しないのかと疑問に思われるかもしれないが、そのこと自体は問題ではない。TPM 値が欲しければ、自分で RPKM 値の総和が 100 万になるように補正すればよいだけだからである。TPM か RPKM/FPKM かの違いは、サンプル間比較をより正確に行いたいという思想に基づくものであり、TPM のほうが戦略的にベターであることは間違いないだろう。し

かしながら（実験デザインにもよるが）サンプル間比較をより正確に行いたいのであれば、TPMのような単に数値の総和を一定値に揃えるだけの正規化ではなく、おそらくedgeR[20]やsleuth[12]のような発現変動解析用プログラムを直接実行するほうがよい。一般によく利用される倍率変化情報であれば、大抵の解析結果に含まれているからである。また、2019年出版の書籍[18]でも言及しているが、DEGES正規化[21]というアルゴリズムを実装した我々のTCCパッケージ[22]、あるいはそのGUI版[23]を利用してもよい。これはedgeRを内部的に繰り返し実行することで、正規化に悪影響を与えうる発現変動遺伝子の影響をできるだけ排除するという思想で開発されたアルゴリズムである。

もちろん発現変動解析用プログラムは、同一サンプル内で異なる遺伝子間の発現レベルの大小関係を比較する目的では利用できない。しかし少なくとも我々の周辺では、このような目的を設定したヒトを見たことがない。サンプル間クラスタリングのような探索的な解析（Exploratory Analysis）を行う際にTPM情報を利用するという考え方もあるようだが、我々はカウントデータを利用している。長さを補正することで異なる遺伝子間の発現レベルの大小関係は変わりうるが、遺伝子数分だけの要素からなるサンプルベクトル間の類似度をスカラーで要約する際に、入力がTPMかカウントかの違いは実質的に問題にはならないというのが我々のスタンスである。

それでもなお、定量PCRのような他の手段で得られた結果と比較する際にTPMは有用だと思われるかもしれない。しかしこの場合でも、CPK・RPKM/FPKM・TPMの三者はともに比率尺度のものであり、直線性を評価する際の傾きが多少異なるだけという結果になる。つまり、同一サンプル内での評価という点では、これらは数学的に等価である。

謝　辞

本連載の一部は、JSPS科研費18K11521の助成を受けたものです。

参 考 文 献

1）寺田朋子，清水謙多郎，門田幸二（2019）次世代シーケンサーデータの解析手法：第14回RNA-seq解析（その2）．日本乳酸菌学会誌 **30**: 153–161.
2）Kankainen M, Paulin L, Tynkkynen S, von Ossowski I, Reunanen J, et al. (2009) Comparative genomic analysis of Lactobacillus rhamnosus GG reveals pili containing a human- mucus binding protein. Proc Natl Acad Sci U S A **106**: 17193–17198.
3）Bang M, Yong CC, Ko HJ, Choi IG, Oh S. (2018) Transcriptional Response and Enhanced Intestinal Adhesion Ability of Lactobacillus rhamnosus GG after Acid Stress. J Microbiol Biotechnol **28**: 1604–1613.
4）Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, et al. (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res **44**: W3–W10.
5）Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics **26**: 493–500.
6）寺田朋子，大田達郎，清水謙多郎，門田幸二（2018）次世代シーケンサーデータの解析手法：第12回Galaxy：ヒストリーとワークフロー．日本乳酸菌学会誌 **29**: 79–88.
7）Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120.
8）Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**: 841–842.
9）Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol **34**: 525–527.
10）Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods **14**: 417–419.
11）Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5**: 621–628.
12）Pimentel H, Bray NL, Puente S, Melsted P, Pachter L (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods **14**: 687–690.
13）谷澤靖洋，神沼英里，中村保一，遠野雅徳，寺田朋子，清水謙多郎，門田幸二（2016）次世代シーケンサーデータの解析手法：第8回アセンブリ後の解析．日本乳酸菌学会誌 **27**: 187–195.
14）Tanizawa Y, Fujisawa T, Kaminuma E, Nakamura Y, Arita M (2016) DFAST and DAGA: web-based integrated genome annotation tools and resources. Biosci Microbiota Food Health **35**: 173–184.
15）谷澤靖洋，真島 淳，藤澤貴智，李 慶範，中村保一，清水謙多郎，門田幸二（2017）次世代シーケンサーデータの解析手法：第9回ゲノムアノテーションとその可視化、DDBJへの登録．日本乳酸菌学会誌 **28**: 3–11.
16）谷澤靖洋，神沼英里，中村保一，清水謙多郎，門田幸二（2016）次世代シーケンサーデータの解析手法：第6回ゲノムアセンブリ．日本乳酸菌学会誌 **27**: 41–52.
17）門田幸二（2014）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.
18）坊農秀雅 編（2019）RNA-Seqデータ解析 WETラボのための鉄板レシピ（実験医学別冊），羊土社，東京.
19）寺田朋子，坂本光央，清水謙多郎，門田幸二（2019）次世代シーケンサーデータの解析手法：第13回RNA-seq解析（その1）．日本乳酸菌学会誌 **30**: 38–45.
20）Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–140.
21）Kadota K, Nishiyama T, Shimizu K (2012) A normalization strategy for comparing tag count data. Algorithms Mol Biol **7**: 5.
22）Sun J, Nishiyama T, Shimizu K, Kadota K (2013) TCC: an R package for comparing tag count data with robust normalization strategies. BMC Bioinformatics **14**: 219.
23）Su W, Sun J, Shimizu K, Kadota K (2019) TCC-GUI: a Shiny-based application for differential expression analysis of RNA-Seq count data. BMC Res Notes **12**: 133.

Methods for analyzing next-generation sequencing data XV. RNA-seq analysis (Part 3)

Tomoko Terada[1], Kentaro Shimizu[1, 2], and Koji Kadota[1, 2]

[1] *Graduate School of Agricultural and Life Sciences, The University of Tokyo.*
[2] *Collaborative Research Institute for Innovative Microbiology, The University of Tokyo.*

Abstract

Galaxy is an integrative data analysis environment run on the web browser. As before, we explain a topic to quantify RNA expression from an RNA-seq dataset (GSE107337) which examined the acid stress response of *Lactobacillus rhamnosus* GG. First, we describe a preprocessing of annotation file with GFF format and extract nucleotide sequence information of the gene regions on the genome. Next, we obtain both count and TPM values for individual genes using the Kallisto quant program. Lastly, we compare the current results with those obtained from the original study and our previous ones. Supplementary materials are available online at: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq2.html#about_book_JSLAB.