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14回 RNA-seq解析（その2）

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

Lactobacillus rhamnosus GGの酸ストレス応答を調べたRNA-seqデータ（SRP125628 or GSE107337）をGalaxy上で解析する一連の手順を解説する。具体的には、公共データベースENAからのGalaxyへのFASTQファイルのインポート、Trimmomaticを用いた前処理、Bowtie2を用いたゲノム配列へのマッピング、そしてhtseq-countを用いたカウントデータ取得まで述べる。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

本稿では、第13回[1]に引き続き、計9サンプルからなる*L. rhamnosus* GG[2]のRNA-seqデータ[3]を取り扱う。今回取り扱うRNA-seqデータ（GSE107337 or SRP125628）のサイズは、gzip圧縮状態で約6GBとそれほど大きくはない（第13回の表1）。それでも、公共データベースENA[4]上のFASTQファイルをそのままGalaxy[5]上にアップロードできれば便利である。そこで本稿は、第13回の図3と同様、ENA上でSRP125628を眺めている状態からスタートする。Galaxy上での作業が中心となるため、第11-12回[6-7]の内容をざっと復習しておくとよいだろう。

今回のウェブ資料では、SRR6322564（pH4.5_1h_rep3）、SRR6322567（pH4.5_24h_rep3）、SRR6322569（pH7_CCG_rep2）の計3サンプル分の作業のみ示す。主な理由は、9サンプル分全ての作業を一つ一つ掲載するのは冗長であり、かつGalaxyの画面が見辛くなるためである。これら3つのサンプルは、比較する3条件間（pH4.5_1h vs. pH4.5_24h vs. pH7_CCG）で、リード数の違いに起因する問題を可能な限り排除しうるものとして選ばれた。実際のリード数は、第13回の表1にも示されているように、それぞれ1,760,461個（pH4.5_1h_rep3）、1,795,874個（pH4.5_24h_rep3）、2,570,876個（pH7_CCG_rep2）である。尚、推奨ウェブブラウザは、Google ChromeまたはFirefox（Internet Explorerは非推奨）である。

GalaxyへのFASTQファイルのインポート

ENAはGalaxyとの連携がよく、GSE107337の検索結果画面から、GalaxyにFASTQデータを送ることができる（第13回の図3；第13回のW18）[W1]。実用上は、予めGalaxyにログインし、新規ヒストリーを作成しておくほうが便利である[W2]。ここでは、GSE107337_3samplesという名前の新規ヒストリーを作成した。後は、ENA画面右端の「FASTQ files（Galaxy）」という列の中で、解析したいサンプルのリンク先を順番にクリックしていけば、Galaxyにデータがインポートされていく[W3]。**図1**は、計6ファイルのインポート完了後の状態である。ヒストリー1-6が計6ファイルの各々に対応している。

*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

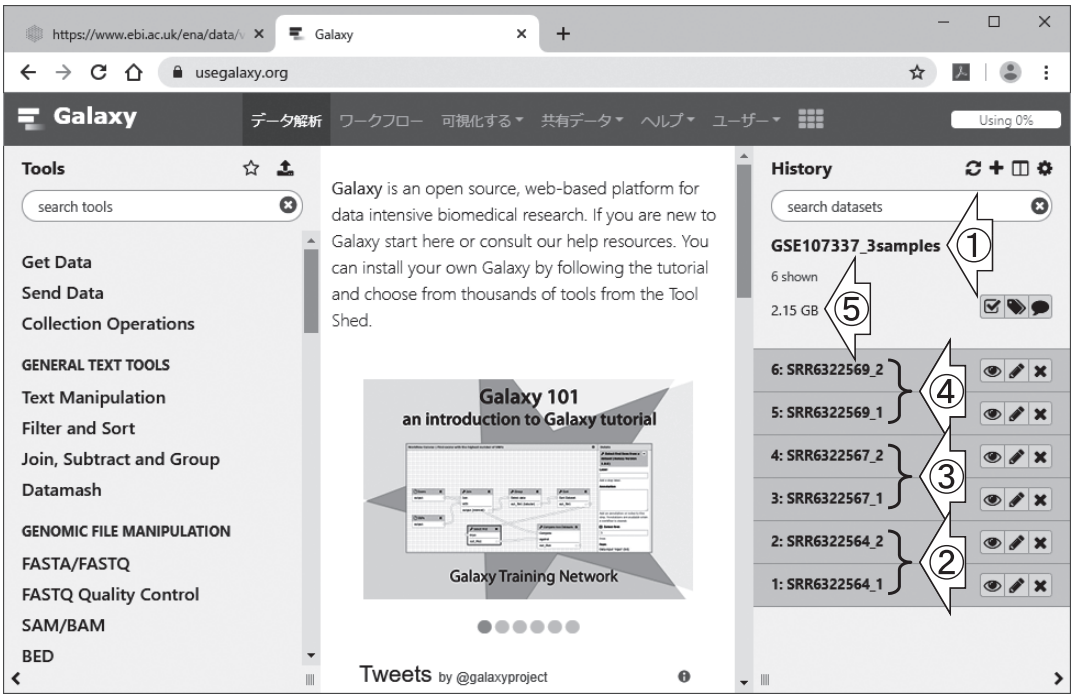

図 1. 解析データ取り込み完了後の Galaxy 画面
①ヒストリー名は GSE107337_3samples、② SRR6322564 (pH4.5_1h_rep3)、③ SRR6322567 (pH4.5_24h_rep3)、④ SRR6322569 (pH7_CCG_rep2)。計 6 ファイルからなり、⑤サイズは 2.15GB であることがわかる。W3-7 と同じ。

クオリティコントロール

第 11 回[6] の W9 では、クオリティチェックを行う FastQC[8] を 1 つのファイルに対して実行するやり方を述べた。ここでは、計 6 ファイルを一度に指定して実行する。Galaxy 画面左側のツール選択パネルの見栄えが第 11 回当時とは異なっているが、FASTQ Quality Control をクリックして、FastQC を選択する [W4-1]。中央パネルの FastQC 操作画面上で Multiple datasets に切り替えて、FastQC を実行したいヒストリー番号 (この場合はヒストリー 1-6) を指定したのち、Execute ボタンを押す [W4-3]。このときは、約 5 分で計 6 ファイルの FastQC 実行が完了し、12 個の出力結果 (ヒストリー 7-18 に相当) が得られた [W4-5]。**図 2** は、FastQC 実行後の Galaxy 画面である。中央パネル上でも、①入力が 6 個、②出力が 12 個だと示されている。右側のヒストリーパネルの名前を見比べることで、③例えばヒストリー 15 と 16 は、ヒストリー 5 (SRR6322569_1) の FASTQ ファイルを入力として FastQC を実行した結果だと読み解けばよい。

次に、低クオリティリードやアダプター配列の除去を行う Trimmomatic[9] を実行する [W5-1]。最初に、入力ファイルが forward 側と reverse 側に 2 分割された paired-end リードであることを認識させる [W5-2]。次に、forward 側 (Trimmomatic では R1/first of pair 側) のファイルがヒストリー 1, 3, 5 に相当する SRR6322564_1, SRR6322567_1, SRR6322569_1 であり [W5-3]、reverse 側 (Trimmomatic では R2/second of pair 側) のファイルがヒストリー 2, 4, 6 に相当する SRR6322564_2, SRR6322567_2, SRR6322569_2 であることを指定する [W5-4]。また、第 6 回[10] でも述べたように、Illumina MiSeq データの場合は多くのリードがアダプター配列を含んでいるため、アダプター配列の除去は無条件で実行したほうがよい [W5-6]。**図 3** は、Trimmomatic 実行後の Galaxy 画面である [W5-9]。尚、第 11 回 (の W11) ではファイルの型について議論を行ったが、約 2 年経過した現在は特段気にする必要はないようである。

現実には、プログラム実行時にどのオプションを採用すべきか判断がつかないこともある。例えば、今回我々はアダプター配列として「TruSeq3 (paired-ended, for MiSeq and HiSeq)」を指定して Trimmomatic を実行したが [W5-6]、他の有力な候補として「TruSeq3 (additional seqs) (paired-ended, for MiSeq and HiSeq)」も存在した。判断に迷う場合は、第 11 回でも述べたようにまずは Trimmomatic を実行し、得られたリードに対して再度 FastQC を実行した結果を眺めればよい。そして、アダプター配列の除去結果を眺め満足のいく結果が得られれば、その選択肢は間違いではなかったと判断すればよい [W6]。**図 4** は、2 回目の FastQC 実行後の Galaxy 画面である [W6-5]。

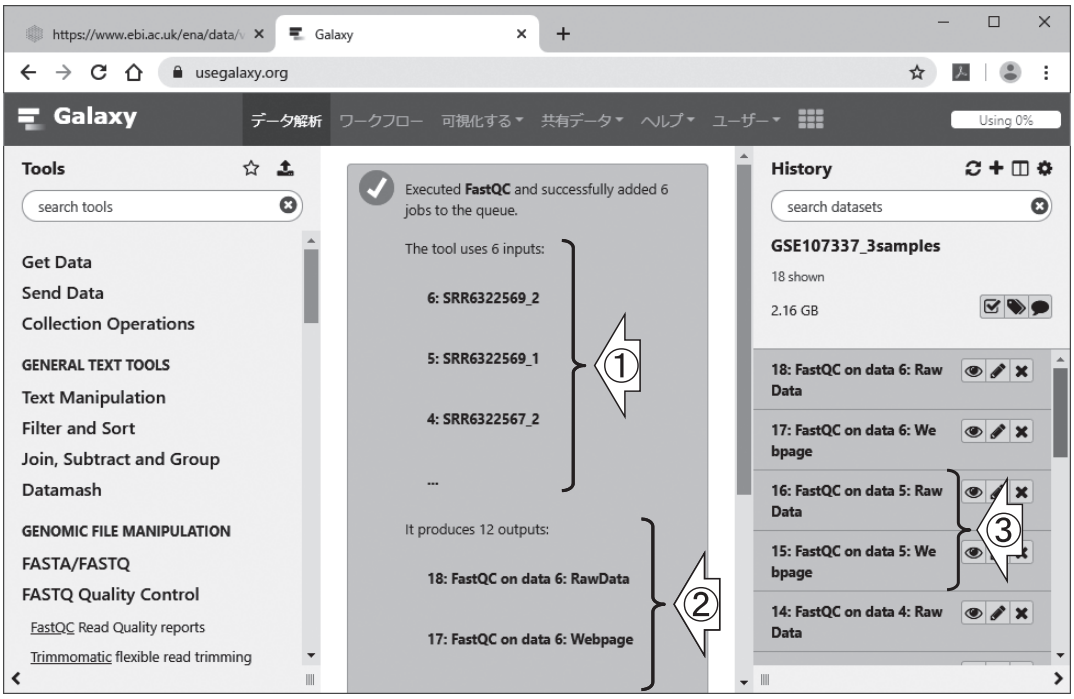

図 2. 1 回目の FastQC 実行後の Galaxy 画面

①入力は計 6 ファイル（ヒストリー 1–6）。②出力は計 12 ファイル（ヒストリー 7–18）。入力ファイル 1 つにつき、2 つの出力が得られることがわかる。例えば、③ FastQC の出力結果であるヒストリー 15 と 16 は、ヒストリー 5 のデータを入力としていることがヒストリーパネルから読み取れる。ヒストリー 15 が html 版、ヒストリー 16 が生データ版である。

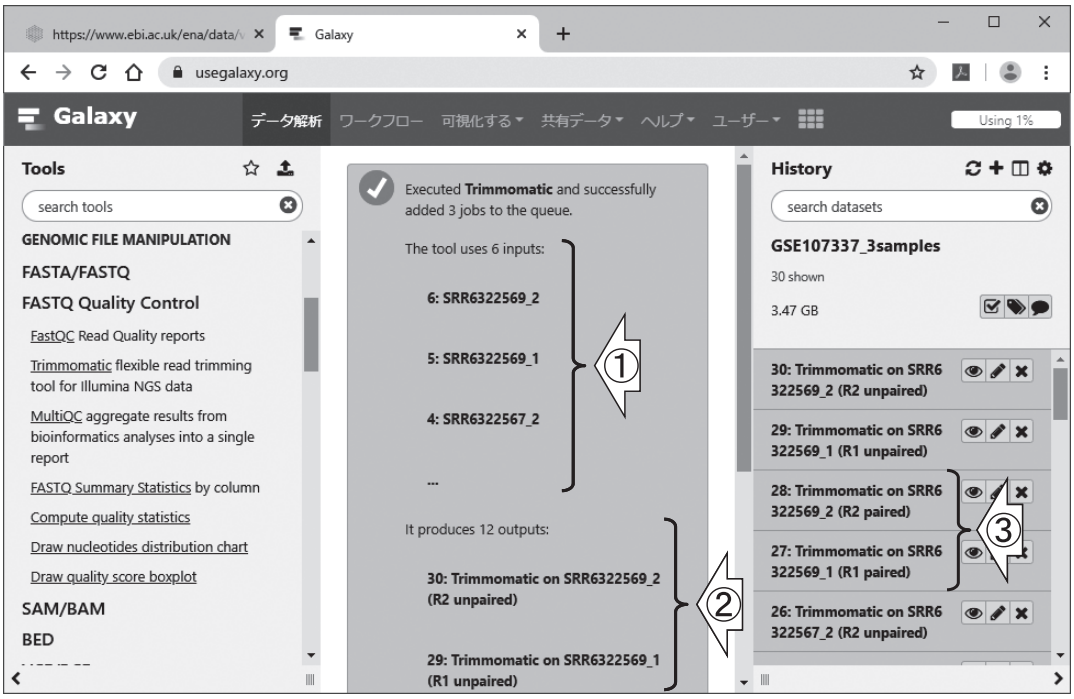

図 3. Trimmomatic 実行後の Galaxy 画面

①入力は計 6 ファイル（ヒストリー 1–6）。②出力は計 12 ファイル（ヒストリー 19–30）。例えば、ヒストリー 27 から 30 が SRR6322569 の出力結果である。このデータの入力は、ヒストリー 5（forward 側 or R1 側）と 6（reverse 側 or R2 側）である。その後の解析に用いるのは、ペアでリードが生き残った③ヒストリー 27（R1 paired）と 28（R2 paired）のほうである［W5–10］。

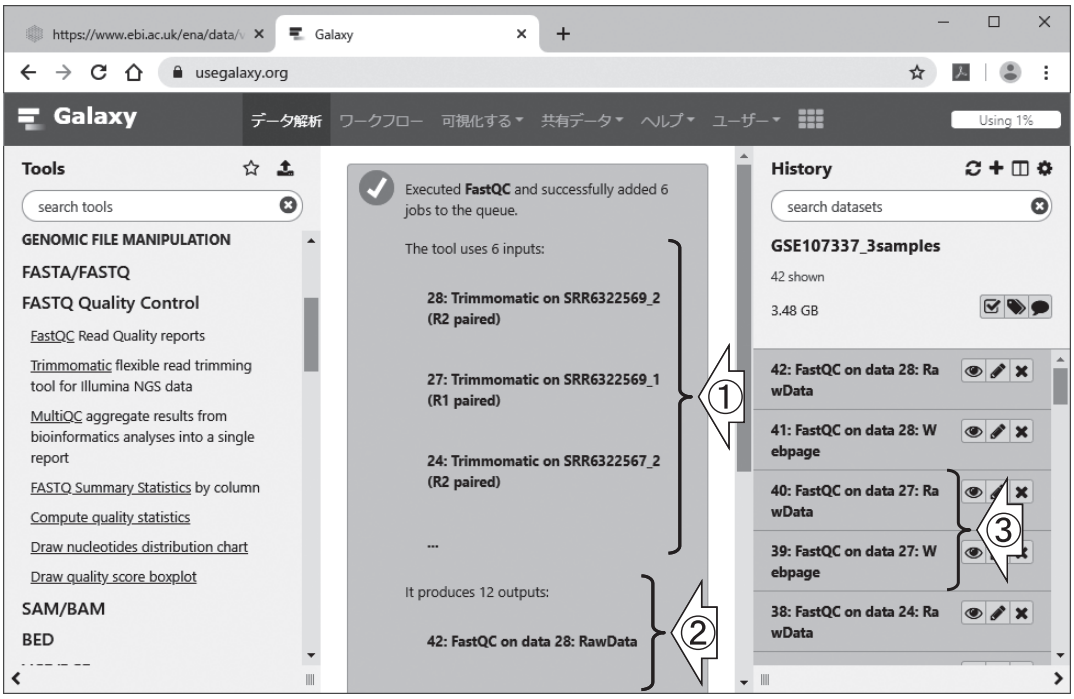

図 4. 2 回目の FastQC 実行後の Galaxy 画面
①入力は計 6 ファイル（ヒストリー 19, 20, 23, 24, 27, 28）。②出力は計 12 ファイル（ヒストリー 31–42）。入力ファイル 1 つにつき、2 つの出力が得られることがわかる。例えば、③ FastQC の出力結果であるヒストリー 39 と 40 は、ヒストリー 27 のデータを入力としていることがヒストリーパネルから読みとれる。ヒストリー 39 が html 版、ヒストリー 40 が生データ版である。

マッピング

Trimmomatic 処理後のリードデータを入力として、マッピングを行う。マッピングプログラムは、Galaxy 上で利用可能でもあることから、原著論文[3]と同じく Bowtie2[11]を用いる。Galaxy 画面左側のツール選択パネルから Mapping をクリックして、Bowtie2 を選択する［W7-1］。最初に、マップする側のリードが paired-end データであることを認識させる［W7-2］。次に、forward 側（Bowtie2 では #1 側）のファイルがヒストリー 19, 23, 27 に相当する SRR6322564_1（R1 paired）, SRR6322567_1（R1 paired）, SRR6322569_1（R1 paired）だと認識させる［W7-3］。同様にして、reverse 側（Bowtie2 では #2 側）のファイルがヒストリー 20, 24, 28 に相当する SRR6322564_2（R2 paired）, SRR6322567_2（R2 paired）, SRR6322569_2（R2 paired）だと認識させる［W7-4］。他のオプションとして、paired-end options をデフォルトの No から Yes に変更する［W7-5］。

次に、マップされる側であるリファレンスゲノ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ゲノムを利用する（Use a built-in genome index）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 genome from the history and build index）オプションに変更する［W7-6］。但し、現時点ではリファレンスゲノム配列がヒストリー上に存在しないので、Ensembl Bacteria[12]から提供されている *L. rhamnosus* GG のゲノム情報（ASM2650v1.fa）、および後のカウント情報取得時に必要なアノテーション情報（ASM2650v1.gff3）を Galaxy のサーバ上にアップロードする［W7-7］。これらのファイルは第 13 回で利用を宣言したものであり、第 13 回の W13 からも取得可能である。

アップロードが完了すると、これらのファイルがヒストリーから見られるようになる［W7-8］。同時に、ゲノム情報に相当するヒストリー 43 の FASTA ファイルがリファレンスゲノム配列の候補として認識されるようになる。ここまでで一通りの準備完了であり、あとは実行ボタンを押すだけである［W7-9］。**図 5** は、Bowtie2 実行後の Galaxy 画面である［W7-11］。①入力は、マップする側のリード情報（ヒストリー 19, 20, 23, 24, 27, 28）、およびマップされる側のリファレンスゲノム情報（ヒストリー 43）である。②出力は BAM 形式ファイルであり、ヒストリー 45 が SRR6322564（pH4.5_1h_rep3）、46 が SRR6322567（pH4.5_24h_rep3）、47 が SRR6322569（pH7_CCG_rep2）のマッピング結果である。

カウント情報取得

マッピング結果は、リファレンスゲノムにマップされたか、ゲノム上のどこにマップされたかという情報から構成

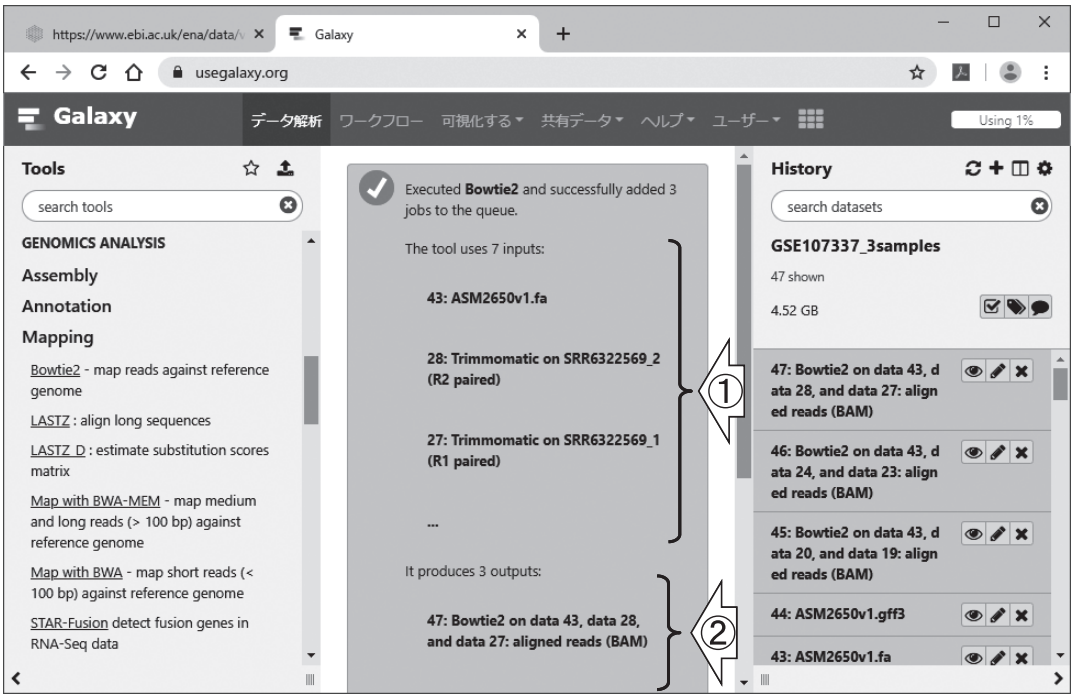

図 5. Bowtie2 実行後の Galaxy 画面
①入力は計 7 ファイル（ヒストリー 19, 20, 23, 24, 27, 28, 43）。②出力は BAM 形式の計 3 ファイル（ヒストリー 45–47）である。

される。そして GFF 形式のアノテーションファイルは、ゲノム上のどこに gene/mRNA/exon/CDS 領域が存在するのかという座標情報を含んでいる。ここでのカウント情報取得とは、マッピング結果と GFF ファイルを入力として与え、どの領域（gene or mRNA or exon or CDS）上にマップされたリード数を数え上げたいか（カウントしたいか）を指定して実行することである。例えば、1 つのマッピング結果（例えばヒストリー 45 のみ）を入力として、gene 領域ごとのカウントデータを取得したとしよう。この場合、出力として得られるカウントデータは、gene の数分だけの要素からなる整数ベクトルとなる。整数である理由は、リード数は 1 個 2 個とカウントしていくので、小数点を含む実数とはならないからである。もし exon 領域ごとのカウントデータを取得した場合は、exon 数は一般に gene 数よりも多いので、得られるカウントベクトルの要素数も多くなる。

Galaxy 上で利用可能なカウント情報取得プログラムとしては、featureCounts [13] と htseq-count [14] が挙げられる［W8-1］。プログラム名のうち、htseq-count の htseq は、High-Throughput Sequencing の意味だと直感的に理解できるであろう。しかしながら、featureCounts の feature は、一部の読者にとっては難解かもしれない。Feature とは、gene/mRNA/exon/CDS などのゲノム上の特定の領域を表す一般的な用語である。feature が gene や exon などを要素としてもつということが正しく理解できれば、上記段落で gene や exon を例とした表現が「feature ごとのカウントデータを取得した場合は、出力として得られるカウントデータは feature の数分だけの要素からなる整数ベクトルとなる」のように全てを含む形で言い換えられることが分かるであろう。尚、R でゲノム上の特定の領域を取り扱う代表的なパッケージである GenomicFeatures [15] の feature も同じ意味である。

ここでは、htseq-count を用いてカウント情報を取得する。Galaxy 画面左側のツール選択パネルから RNA-seq をクリックして、htseq-count を選択する［W8-1］。中央パネル上で、マッピング結果である 3 つの BAM ファイル（ヒストリー 45–47）を指定する［W8-2］。次に、GFF 形式のアノテーションファイル（ヒストリー 44）を指定する［W8-3］。これで入力ファイルの指定は完了である。次に、どの feature についてカウント情報を取得したいかを指定する。デフォルトの Feature type は exon になっているが［W8-4］、原著論文と同じく gene に変更して実行する［W8-5］。**図 6** は、htseq-count 実行後の Galaxy 画面である［W8-7］。出力は、計 6 ファイル（ヒストリー 48–53）である。このうち、no feature という名前が付加されたヒストリー 49, 51, 53 は、欲しいカウントデータ以外の統計情報が含まれている。したがって、残りのヒストリー 48, 50, 52 が feature 数分（この場合は gene 数分）の要素からなる整数ベクトルを含む SRR6322564（pH4.5_1h_rep3），SRR6322567（pH4.5_24h_rep3），SRR6322569（pH7_CCG_rep2）のカウント情報に相当する［W8-8］。

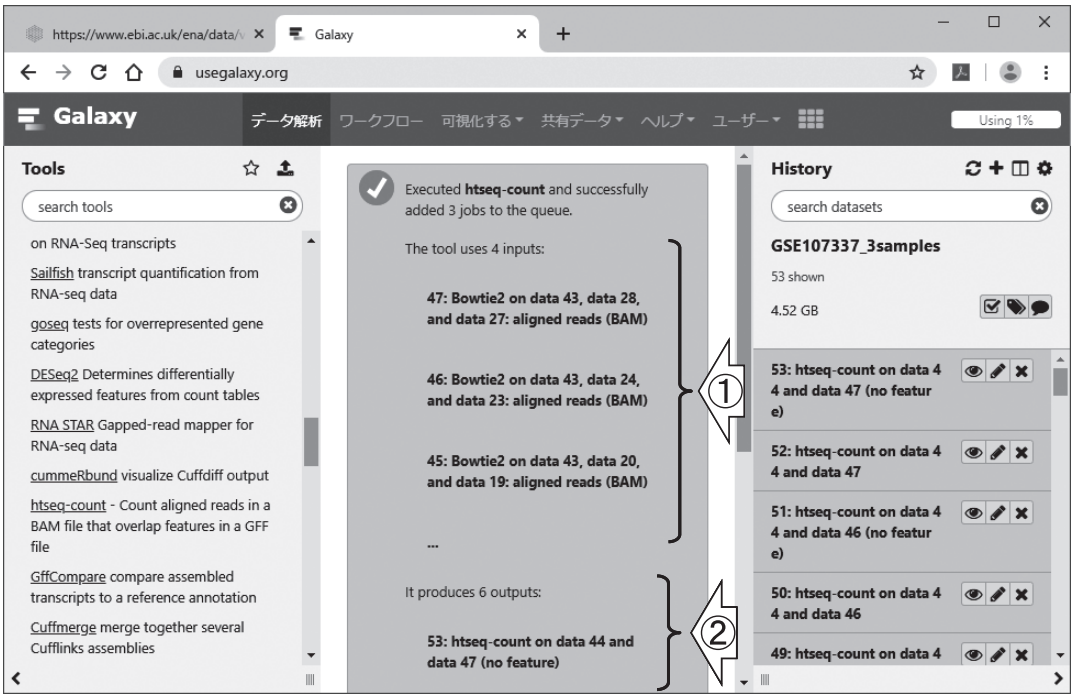

図 6. htseq-count 実行後の Galaxy 画面
①入力は計 4 ファイル（ヒストリー 44–47）。②出力は計 6 ファイル（ヒストリー 48–53）である。

カウント情報の連結

Galaxy は、独立に行われた結果を連結する各種機能を提供している。ここでは、Galaxy 画面左側のツール選択パネルから見られる Collection Operations 内の Column join という機能を利用して、別々のヒストリー内にあるカウントデータの連結を行う［W9-1］。**図 7** は、Column join 実行後の Galaxy 画面である［W9-5］。連結対象（ヒストリー 48, 50, 52）を指定して実行ボタンを押せば、3 サンプル分のカウントデータの整数ベクトルを列方向で連結した結果（ヒストリー 54）が得られる［W9-7］。得られたカウントデータ行列の、列数は 3（行名情報からなる最初の列を除く）［W9-7］、行数は 2,949（ヘッダー行を除く）である［W9-8］。我々は通常、ダウンロードしたカウントデータファイル（ファイル名：data_3samples.txt）を Excel で眺めて全体像を把握した段階で、列名部分を簡略化したサンプル名に変更しておく［W9-10］。サンプル間クラスタリング結果などの解釈を容易にするためである。

次に、カウントデータ中の 2,949 行という数値の妥当性について検証する。この数値は、Ensembl Bacteria のサイト上で見られる数値（2,944 genes）よりも 5 個多い［W9-11］。カウントデータファイルを眺めて 2,949 genes を調べると、LGG から始まる 2,944 個の gene ID と、EBG から始まる 5 個の gene ID からなることがわかる。実際には、カウント行列を Excel で確認した段階で、すぐに EBG から始まる異質な 5 個の ID が目に留まる［W9-9］。そして念のために、カウントデータ取得時に用いた GFF ファイル提供元である Ensembl Bacteria を再訪して、違いに気づくという流れとなる。我々は経験上、この LGG から始まるか EBG から始まるかといった ID の違いは、情報提供元の違いに起因する場合が多いことを知っている。GFF ファイルの 2 列目には source 情報、つまり情報提供元が記されている。この観点から、例えば EBG00001128470 に合致する行の source 情報を他と見比べて納得するのである［W10-1］。5 つの EBG から始まる ID の source は Rfam[16] であり、LGG から始まる ID の source は ena となっていることに気づくであろう［W10-4］。

5 つの EBG から始まる ID のカウント情報を削除すべきかどうかについて見解を述べる。もし原著論文中でこれらの取り扱いに関する言及があれば、それに従うのも一手であろう。もし言及されていない場合は、それらが後の解析結果に悪影響を与えそうかどうかを判断基準とすればよい。具体的には、これらのカウント数が他と比べて多いわけではなく、比較する 3 条件間（pH4.5_1h vs. pH4.5_24h vs. pH7_CCG）で極端にカウント数が異なっているようには見受けられなかった［W9-9］。これらの事実を鑑み、我々の下した結論は「残しておいても問題ない（残す）」である。

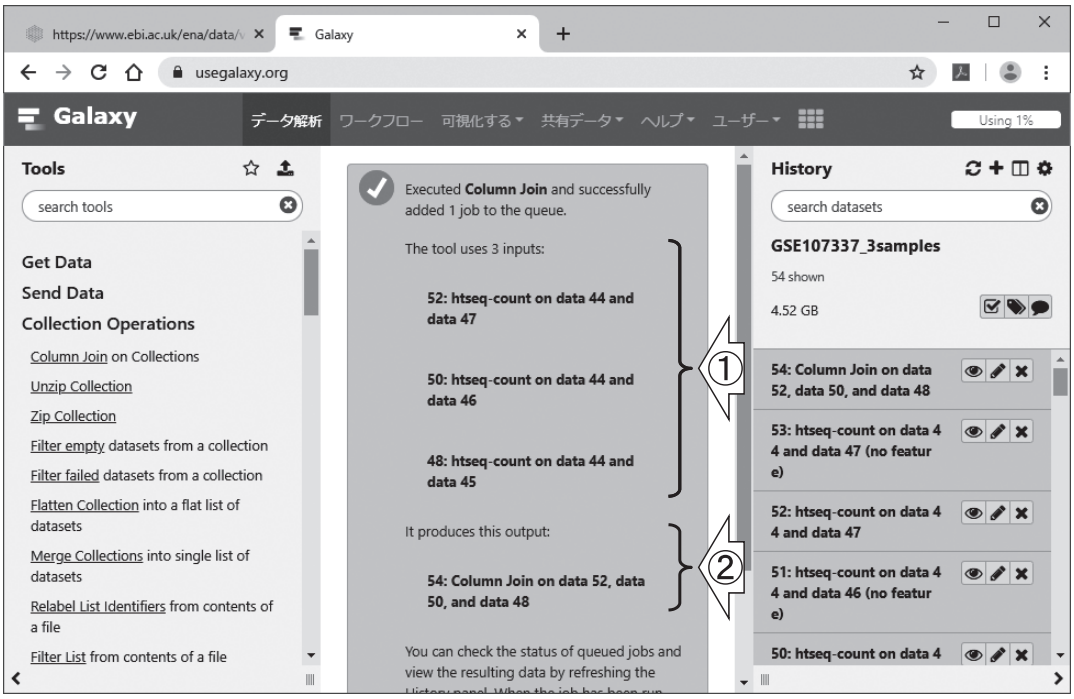

図 7. Column join 実行後の Galaxy 画面
①入力は計 3 ファイル（ヒストリー 48, 50, 52）。②出力は計 1 ファイル（ヒストリー 54）である。

GFF ファイルに慣れる

話が若干前後するが、htseq-count プログラム実行時に内部的にどのようなことが行われているかを、GFF ファイルの形式と絡めて説明する。GFF ファイルに対する EBG00001128470 の検索結果からもわかる通り［W10-1］、カウント行列の行名に相当する gene ID 情報は、GFF ファイル内に含まれている。まず、htseq-count を用いたカウント情報取得時に Feature type オプションを gene に変更して実行したことを思い出してほしい［W10-1］。これは GFF ファイル中の① 3 列目が② gene となっている行の座標情報（GFF ファイルの③ 4 列目と④ 5 列目）および⑤ストランド情報（7 列目）を用いて、その領域付近にマップされたリード数をカウントするよう指示したことを意味する［W10-2］。

GFF ファイル自体はタブ区切り（tab-separated）テキストファイルである。シャープ（#）から始まる行は、コメント行である。Excel で読み込んだ際の 1 列目は 1 番目のフィールド、2 列目は 2 番目のフィールドなどと呼ばれる。1 列目は、リファレンス配列の名前（seqname）が書かれている。今眺めている GFF ファイル（ASM2650v1.gff3）は、シャープ（#）から始まる行以外は全て FM179322 になっていることに気づくであろう。この理由は、元のゲノム配列が 1 本の環状染色体のみから構成され、その Accession 番号は FM179322 だからである。

GFF ファイル中の 9 番目のフィールド（9 列目）には、この feature 領域に対する付加的な属性（Attribute）情報が含まれている［W10-3］。このフィールド内は、複数の情報がセミコロン区切り（semicolon-separated）で格納されている。カウント行列取得時は、このフィールド内のどの情報を「行の名前」として利用するかを明確に指定せねばならない。W8-4 で行ったカウント情報取得時は、Feature type 以外はデフォルト値を利用したため特に触れなかったが、実際の行名には ID Attribute オプションのデフォルト（gene_id）が採用されている。つまり、GFF ファイル中の 9 番目のフィールドをセミコロンで分割し、分割されたもの（ID＝…, Name＝…, biotype＝…., gene_id＝…, logic_name＝…）の中から gene_id の情報を文字列検索で探し出し、その右辺側の文字列を行名として出力しているのである。この原理が分かれば、例えば W8-4 で ID Attribute オプションのところを gene_id から ID に変更して実行すると、得られるカウント行列の行名が EBG00001128470 から gene:EBG00001128470 に変更されるはずだと想像できるだろう。

但し、おそらく W8-4 で ID Attribute オプションのところを、例えば Name に変更して実行するとエラーとなる。理由は異なる feature 領域で同じ名前が使われてしまうことになるからである。例えば W10-4 の 2528 行 I 列のセル（領域［506642, 506842］の属性情報）と 2532 行 I 列のセル（領域［507097, 507297］の属性情報）を見比べると、両方とも Name＝rli28 になっていることが分かる。エラーを吐くことなく、行名を rli28 として 2 つの領域（［506642,

506842］と［507097, 507297］）上にマップされたリード数の総和をカウントしてほしいと思う読者はいるだろうか？賛否両論あると思われるが、少なくとも「重複した行名のものが存在する」的なことをユーザに気づかせてくれるプログラムは親切と言えるだろう。我々の経験上、GFF ファイルを入力とする際のエラー遭遇率は高く、そして厄介である。ここで述べた事例のみではどうにもならないこともあると思われる。それでも、手元の入力情報をどのような手順で処理すれば出力結果が得られるのか？どのようなエラーの可能性が存在しうるのかに思いを馳せながら GFF ファイルを眺めれば、いくつかの問題は解決できるであろう。

原著論文と比較

最後に、得られたカウントデータの数値を、原著論文[3]の著者らが GEO[17]上で公開している GSE107337 のカウントデータ（GSE107337_RawCounts.csv）と比較する［W11-1］。この公開データは、LGG という ID のみからなる 2,838 genes 分のカウント情報から構成されている。また、3 条件×3 反復で計 9 サンプル分のデータではなく、3 反復の平均値となっていることが、ヘッダー行部分の列名から読み取れる（average_rawcount と書かれているということ）［W11-2］。

この論文中では、2 つの遺伝子（LGG_02240 と LGG_02372）についてコントロール（pH7_CCG）に対する酸ストレス応答の倍率変化（何倍発現変動したか）が明記されている。具体的には、LGG_02240 は 24 時間の酸ストレス条件下（pH4.5_24h）で 2.53 倍、1 時間の酸ストレス条件下（pH4.5_1h）で 1.39 倍の発現上昇と書かれている（原著論文中の 1,608 ページ目の左上）。また、LGG_02372 は pH4.5_24h で 3.34 倍、pH4.5_1h で 1.34 倍の発現上昇と書かれている。これらの倍率変化の数値は、GSE107337_RawCounts.csv のデータに基づいて**再計算した倍率変化の値とは一致しなかったが、$\log_2$（倍率変化）の値と完全一致**した［W11-5］。つまり、LGG_02240 は pH4.5_24h で $\log_2(2444/422)=2.5339$、pH4.5_1h で $\log_2(1106/422)=1.3900$ であった。また、LGG_02372 は pH4.5_24h で $\log_2(132/13)=3.3440$、pH4.5_1h で $\log_2(33/13)=1.3440$ だったということである。

Galaxy 上で我々が得たカウントデータは、計 3 サンプル分（pH4.5_1h_rep3, pH4.5_24h_rep3, pH7_CCG_rep2）であった［W9-10］。上記公開データとの比較のために、我々も改めて計 9 サンプル分について同様の手順でカウントデータ取得まで行い［W12-1］、倍率変化［W12-3］、および $\log_2$（倍率変化）の値を算出した［W12-4］。結果として、原著論文と近い値を示したのは $\log_2$（倍率変化）のほうであった。具体的には、LGG_02240 は pH4.5_24h で $\log_2(2914/490)=2.572$、pH4.5_1h で $\log_2(1231/490)=1.329$ であった。また、LGG_02372 は pH4.5_24h で $\log_2(188/15)=3.648$、pH4.5_1h で $\log_2(56/15)=1.901$ であった。LGG_02372 において公開データのものとの乖離が大きい理由は、カウント数の少なさに起因するゆらぎのためであろう。

原著論文では、ジアミノピメリン酸パスウェイに属する 4 遺伝子（LGG_00113［W13-1］, LGG_00115［W13-2］, LGG_00108［W13-3］, and LGG_00109［W13-4］）について、酸ストレス条件下で発現が低下したと述べられていた。我々も「公開版のカウントデータ」と「我々が Galaxy で得たカウントデータ」の両方で、同様の傾向になっていることを確認した［W13］。他にも、LGG_01064 の発現上昇［W14-1］、および LGG_02032［W14-2］と LGG_00418［W14-3］の発現低下についても確認した［W14］。さらに、2,838 遺伝子×3 サンプルからなる公開カウントデータ（data_original.txt［W15-1］）と 2,949 遺伝子×3 サンプルからなる我々のカウントデータ（data_galaxy.txt［W15-2］）の共通遺伝子を抽出し、計 6 サンプル間の総当たりの類似度（Spearman's correlation coefficient）を算出した結果、同一サンプル（条件）間の類似度が最も高いことを確認した［W15-3］。以上の結果より、今回我々が行ったカウントデータを得るまでの作業は、手順としては妥当だと言えよう。

おわりに

今回は、独立に行って得たカウントデータの妥当性までを議論した。しかしながら、実際には $\log_2$（倍率変化）であるにもかかわらず、倍率変化だと原著論文中で述べられている点は適切とはいえない。また、本来この種の議論は RPKM のような補正後の値で行うべきであるが、生のカウントデータに基づいて行っている点も不適切である。我々が公開カウントデータに対して調べた総カウント数（pH4.5_1h＝774,376、pH4.5_24h＝1,670,016、pH7_CCG＝1,690,218）の最大と最小の間には 2 倍以上の開きがあるからである［W15-4］。次回はこれらの点について議論する予定である。

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，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.
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) Toribio AL, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L, et al. (2017) European Nucleotide Archive in 2016. Nucleic Acids Res **45**: D32-D36.
5) 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.
6) 大田達郎，寺田朋子，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 11 回統合データ解析環境 Galaxy．日本乳酸菌学会誌 **28**: 167-175.
7) 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.
8) Andrews S. (2015) FastQC a quality control tool for high throughput sequence data, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
9) Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114-2120.
10) 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.
11) Langmead B, Salzberg SL. (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods **9**: 357-359.
12) Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, et al. (2019) Ensembl 2019. Nucleic Acids Res **47**: D745-D751.
13) Liao Y, Smyth GK, Shi W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**: 923-930.
14) Anders S, Pyl PT, Huber W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics **31**: 166-169.
15) Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, et al. (2013) Software for computing and annotating genomic ranges. PLoS Comput Biol **9**: e1003118.
16) Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, et al. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res **46**: D335-D342.
17) Clough E, Barrett T. (2016) The Gene Expression Omnibus Database. Methods Mol Biol **1418**: 93-110.

Methods for analyzing next-generation sequencing data XIV. RNA-seq analysis (Part 2)

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

Lactobacillus rhamnosus is a probiotic lactic acid bacterium frequently isolated from human gastrointestinal mucosa of healthy individuals. We describe an analysis procedure to RNA-seq data (SRP125628 or GSE107337) on the acid stress response of *L. rhamnosus* GG on an integrated data analysis environment called Galaxy. Specifically, we will describe importing FASTQ files from the public database ENA to Galaxy, preprocessing using Trimmomatic, mapping to reference genome sequence using Bowtie2, and obtaining count data using htseq-count. We will also discuss about similarities of count data between the original and ours. Supplementary materials are available online at: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq2.html#about_book_JSLAB.