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13回 RNA-seq 解析（その1）

寺田 朋子[1]、坂本 光央[2]、清水 謙多郎[1,3]、門田 幸二[1,3*]

[1] 東京大学 大学院農学生命科学研究科
[2] 理化学研究所 バイオリソース研究センター 微生物材料開発室
[3] 東京大学 微生物科学イノベーション連携研究機構

Lactobacillus rhamnosus は、健康なヒトの胃腸粘膜からしばしば単離されるプロバイオティクス乳酸菌である。今回は、まず *L. rhamnosus* GG に関する2つのゲノム配列決定論文の内容を簡単に紹介する。そして、Gepard によるゲノムスケールのドットプロットを用いて、一部領域が反転している状況を確認する。次に公共データベース上で検索する際の Tips を述べ、この菌株の酸ストレス応答を調べた RNA-seq データ（SRP125628 or GSE107337）を取得したのち、実験デザインの解説を行う。今回の内容は、次回以降解説予定のマッピング・カウントデータ取得・発現変動解析を含む一連の RNA-seq データ解析を行うために必要なファイルの取得や全体像を把握する準備編という位置づけである。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*

はじめに

本連載では、乳酸菌の RNA-seq データを第3～5回で取り扱っている[1-3]。具体的には、Illumina HiSeq 2000 で取得した *L. casei* 12A 株[4] の stranded paired-end データ（SRR616268）である。これが選ばれた主な理由は、第1～2回[5-6]で取り扱ったゲノム配列と同じ株由来のデータであるため、RNA-seq リードのリファレンスゲノムへの高いマップ率が期待されたからである。しかしながら実際には、第5回の**図2**で示されたように、forward 側の 100-107 bp 領域が原因で全リードの約 0.4% しかマップされなかった[3]。また、Illumina のリード長は forward と reverse ともに同じはずだが、このデータはなぜか異なっていた（forward 側が 107 bp、reverse 側が 93 bp）。これらの状況証拠を鑑み、SRR616268 は例題として不適切だと判断した。結果的に第3～5回まで適切とはいえないデータを用いたことに対し、率直に読者にお詫び申し上げる。

本連載開始（2014年）から5年が経過し、公共データベース（以下、DB）に登録された乳酸菌の RNA-seq データも増加している。本稿では、*L. rhamnosus* GG[7] の酸ストレス応答を調べた RNA-seq データ（SRP125628 or GSE107337）のダウンロードや実験デザインを中心に解説する。前回同様、予期しない不具合を避けるため、ウェブブラウザは Google Chrome または Firefox（Internet Explorer は非推奨）を用いてほしい。

リファレンスゲノム配列

リファレンスゲノムとして用いる *L. rhamnosus* GG は、健康なヒトの胃腸粘膜からしばしば単離されるヘテロ発酵型の乳酸菌である。このゲノム配列決定に関する原著

*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

論文によると、1本の環状染色体のみから構成され（ゲノムサイズは3.01MB）、プラスミドはもたない[7]［W1］。遺伝子数は2,944個であり、DDBJ/GenBank/EMBLのAccession番号はFM179322である。原著論文から辿れるFM179322のリンク先は「GenBank形式」で表示されているが、例えば「FASTA（text）形式」に切り替えることで、リファレンス配列の一般的な形式であるFASTA形式ファイル（FM179322.fasta）として保存することができる［W2］。また、GenBank形式の情報を眺めることで、正確なゲノムサイズ（3,010,111 bp）やこの菌株がAmerican Type Culture Collectionの53103に相当するもの（ATCC 53103）であることなどがわかる［W3］。Taxonomy IDが568703であることもわかる。

特記事項として、このTaxonomy IDのリンク先を辿っていくと、別グループの論文[8]がReferenceとして提示される点を挙げておく［W4］。このゲノム配列決定論文では、ゲノムサイズは3,005,051 bpと報告されている［W5］。また、2,834のタンパク質コード遺伝子が予測されており、その内訳は1,939個（68%）の既知機能遺伝子、および895個（32%）の仮想遺伝子となっている。Accession番号はAP011548と記載されており、前述のFM179322とは異なる。この理由はAP011548の論文で言及されており、FM179322のほうが登録および論文公開が早かったということである。AP011548論文では、両者のゲノム配列の違いについても言及されている。AP011548のゲノム配列は、FM179322中のものよりも約5 KB（3,010,111－3,005,051＝5,060 bp）短く、約8.9 KBの領域［618415, 627294 bp］が反転している。後で利用するため、AP011548のFASTA形式ファイル（AP011548.fasta）も作成しておく［W6］。

Gepardでドットプロット

上述した一部領域の反転状況は、第7～8回[9-10]でも解説したドットプロット[11]で確認することができる。この場合は、FM179322とAP011548とのゲノム配列をx軸とy軸にそれぞれ並べ、同一塩基部分をハイライトさせることに相当する（**図1**）。第7回[9]では、数万塩基程度までの比較的短い塩基配列同士のドットプロット作成手段としてdotterプログラム[12]を用いた。そして第8回[10]では、約230万塩基（2.3MB）の配列を入力とした場合に、dotterが実質的に実行不可能であることを示した。今回の約300万塩基の入力ファイルに対してもdotterは適用できないため、ここでは平成29年度NGSハンズオン講習会でも紹介したGepard[13]を利用する［W7］。

GepardはJavaプログラムであり、FM179322とAP011548の全ゲノム同士の比較でも、10秒程度でドットプロットを描画できる（**図1a**）［W8］。プログラムによって原点の位置は異なるが、①Gepardの場合もdotterと同じく左上となっていることが分かる。ここではx軸にFM179322.fasta、y軸にAP011548.fastaを配置するように入力ファイルを指定しているので、配列長を示すドットプロットの②右上端が3010110、③左下端が3005050となっているのは妥当である。実際の塩基数よりも1少ないのは、始点が0だからだと解釈すればよい。このドットプロットから読み取れることは、対角線上に位置する塩基が同じだという点である。つまり、同一配列間のドットプロットの特徴と酷似しているということである。

しかしながら、全ゲノム同士のドットプロットでは、AP011548のゲノム配列のほうが約5 KB短く、約8.9 KBの領域が反転しているという状況を読み取ることは当然できない。理由は、全ゲノム中に占める当該領域の割合が（5,060＋8,880）／ 3,005,051＝0.46%と極めて小さいからである。もちろんGepardは、注目する領域間のドットプロットを描画する機能がある。**図1b**は、比較する2つ配列ともに、領域［610000, 640000 bp］に限定して再描画した結果である。この領域は反転領域［618415, 627294 bp］を約1/3の割合で中央付近に含んでいる。④若干右上側にずれてはいるものの、左上端から右下端にプロットされている直線部分が同一領域に相当する。そして、⑤左下方向から右上方向にプロットされている直線部分が、反転領域に相当する［W9］。このように、メインの直線と**垂直**に交わるような形でプロットされるのが反転領域の典型的な特徴である。

対比として、第7回の**図1b**と比較してもよいだろう。このときに着目していた重複配列領域は、メインの左上端から右下端の直線と**平行**にプロットされていた。尚、**図1b**の領域［610000, 640000 bp］に限定したドットプロットにおいて、④メインの直線が若干右上側にずれている理由は、FM179322にはあるがAP011548にはない領域が存在するためだと解釈すればよい。例えば、領域［238000, 244000 bp］のドットプロットを眺めると理解できるであろう［W10］。

seqinrでドットプロット

ここでは、反転の具体的なイメージを掴んでもらうべく、以下の2つの仮想塩基配列を用いて示す。

配列*k*：ACTCGTAGTCTATCATACGA
配列*l*：ACTCGACTATCTGATTACGA

*l*の下線部分は、*k*の下線部分の左右を入れ替えたものに相当し、これが**反転された状態**である。それゆえ、配列*k*と*l*は下線部分が異なっている。**図2a**は、配列*k*同士のドットプロットである［W11］。この図は第7回でも利用したseqinrパッケージ[14]中のdotPlot関数を用いて作成しているため、**図1**と異なり原点の位置が①左下端になっている点に注意してほしい。同一配列同士の比較であるため、左下端から右上端にかけて、キレイに直線状のドットが描

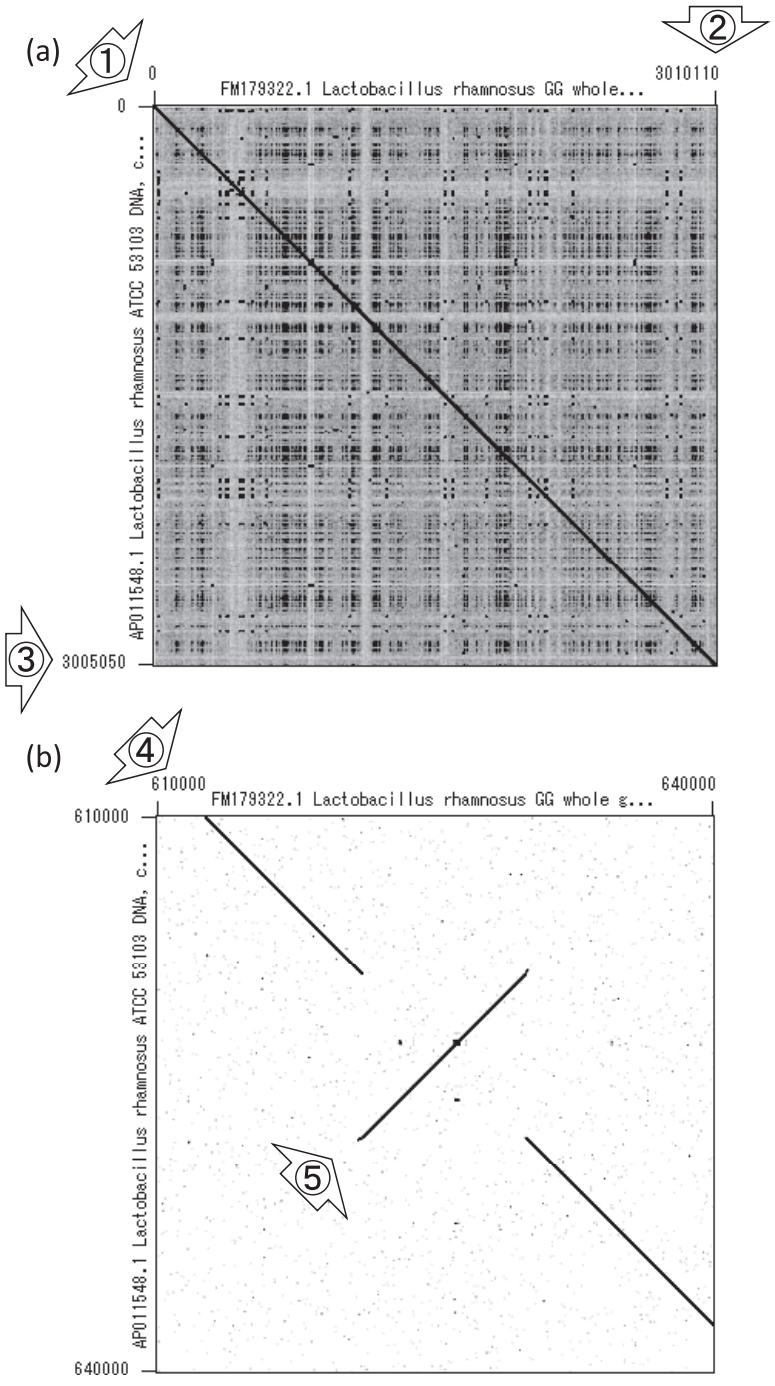


図 1. Gepard による FM179322 と AP011548 のゲノムスケールのドットプロット。
(a) 全ゲノム同士の比較。 (b) 領域 [610000, 640000 bp] の比較。

かれていることがわかる。**図 2b** は、配列 *k* と *l* のドットプロットである [W12]。枠で囲った反転領域のみ、メインの直線と**垂直**に交わるような形で、②左上方向から右下方向にかけて直線状のドットが描かれていることがわかる。

ここまでは、RNA-seq 解析を行う際に用いるリファレンスゲノム配列について述べた。今回は FM179322 と AP011548 の両方を偶然発見したため、疑問点の解消や違いを確認する目的で、ドットプロットの復習や Gepard の紹介を兼ねて解説した。どちらのゲノム配列も間違いではないと思われるが、Ensembl Bacteria [15] 中の *L. rhamnosus* GG の情報は FM179322 と同じである。また、FM179322 の原著論文のほうが先に公開されていることから、以後は Ensembl Bacteria から提供されている ASM2650v1 のゲノム (ASM2650v1.fa) およびアノテーション情報 (ASM2650v1.gff3) を利用する [W13]。

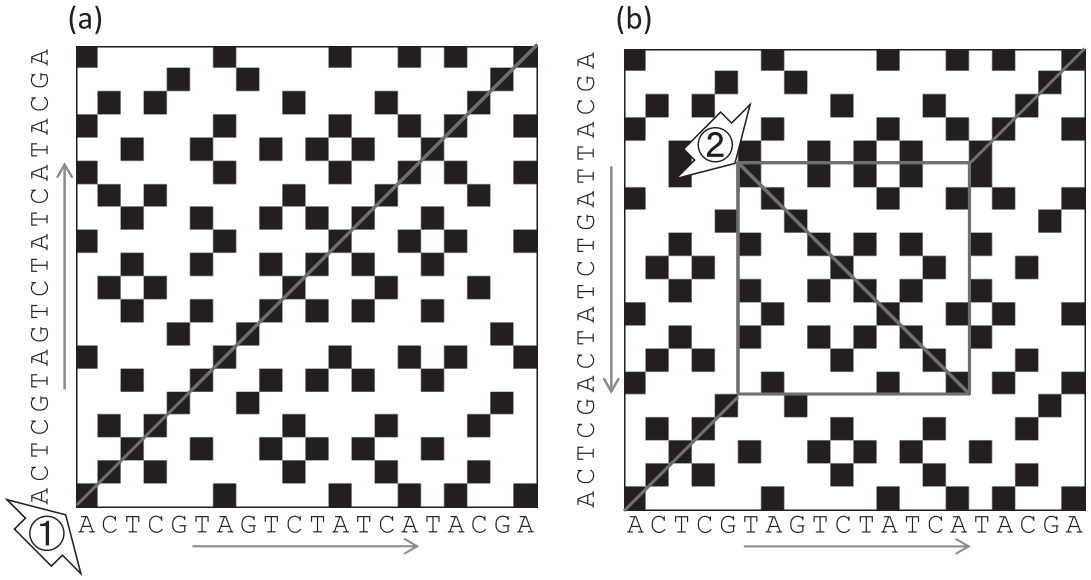


図 2. seqinr による仮想塩基配列のドットプロット。
(a) 同一配列（配列 k vs. k）の場合。 (b) 反転領域を含む場合（配列 k vs. l）。

RNA-seq データの取得

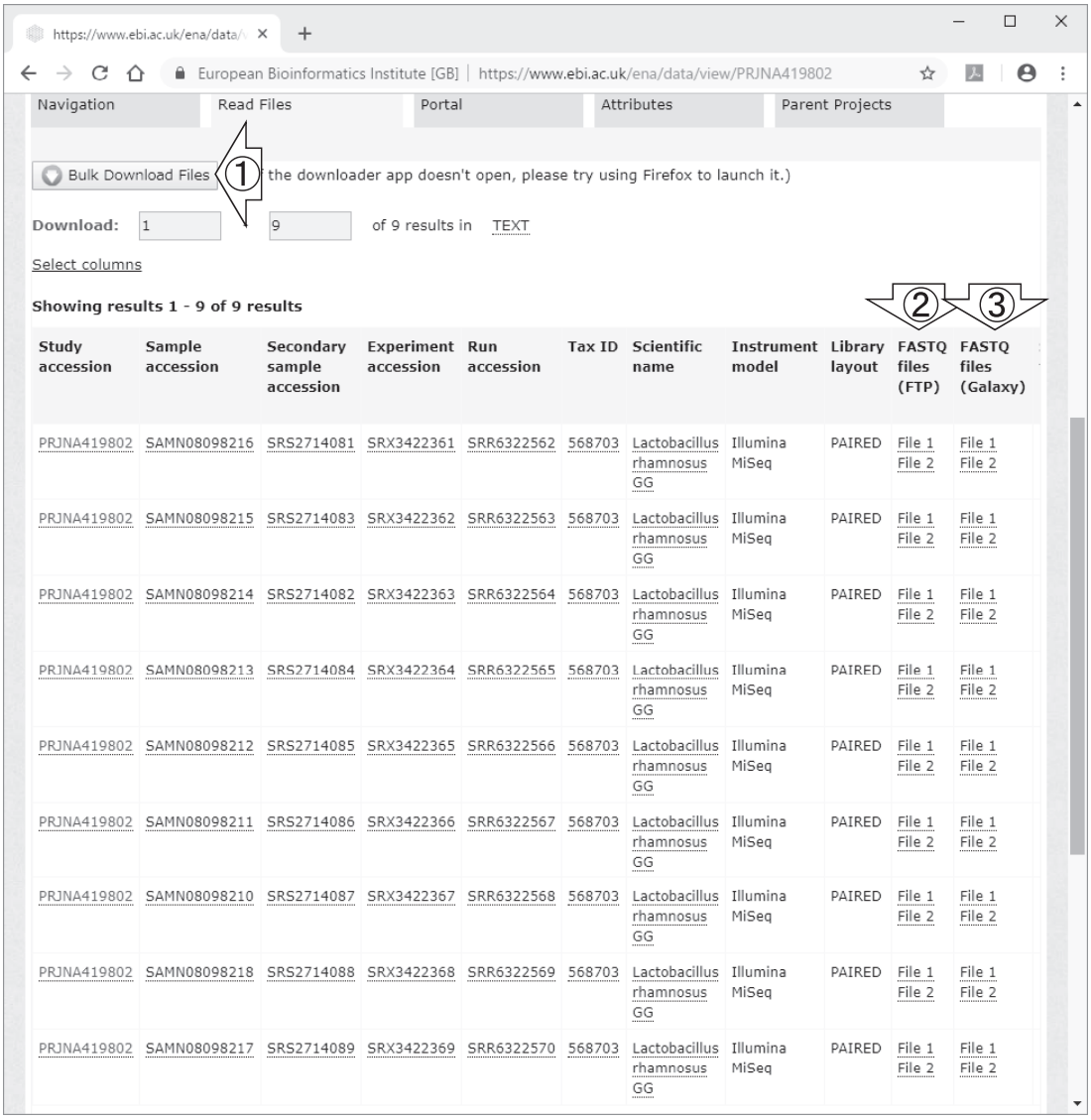
ここでは、公共 DB を眺めて *L. rhamnosus* GG の RNA-seq データ（SRP125628 or GSE107337）の発見、その原著論文[16]の同定、そしてデータ取得までを述べる。公共 DB の基礎知識や基本的な利用法については第 3 回で述べている。SRR・DRS・PRJ といった様々な ID の対応関係について復習しておくとよい。

まず NCBI の SRA[17]で「Lactobacillus RNA-seq」でキーワード検索を行うと、14 件のプロジェクトがヒットした（2019 年 2 月 1 日調べ）[W14]。後に原著論文[16]にたどり着いた際に確認に用いた GSE107337 は、この 14 件の中から「実験デザインが分かりやすく、それほどサンプル数が多くない」という条件を最初に満たした Accession 番号である [W15]。このデータセットは NGS 機器の 1 つである Illumina MiSeq を用いて得られており、全部で 9 サンプルからなる [W16]。このデータの原著論文は一見するとまだないように見えるが、Registration date が 25-Nov-2017 であることから、うまく探せば見つかるのではという視点で探す。我々は試行錯誤の末、NBCI が運営する発現 DB である GEO[18]の情報（正確には GSE107337 の著者情報）を用いることで、2018 年 10 月 28 日に公開されている GSE107337 の原著論文[16]にたどり着くことができた [W17]。

第 3 回でも述べたように、FASTQ ファイルのダウンロードは ENA[19]または DRA[20]で行う。ENA 上で GSE107337 を検索すると、PRJNA419802 や SRP125628 などのデータセット全体を指し示す ID や、全 9 サンプルに付随する様々な ID 情報を俯瞰できる（**図 3**）[W18]。ENA は①一括ダウンロードにも対応している [W19]。著者らのうち 2 名が試した限りでは、1 名はすんなり成功したが [W20]、後程動作確認したもう 1 名は一度失敗した [W21]。失敗した場合の対応策としては、ENA 上でファイルを②個別にダウンロードするか、R の SRAdb パッケージ[21]を利用した一括ダウンロードを試してもよいだろう [W22]。但し SRAdb を利用する際は、原著論文中に記載されている GSE107337 ではなく、その ID から辿れる SRP125628 を指定しなければならない。著者らの経験上、失敗するときは何をやっても失敗するが、翌日のリトライで何事もなかったかのように成功する場合が多い。

このデータセットは、全 18 ファイル（paired-end なので 1 サンプルにつき 2 ファイル）合わせても約 6GB と比較的サイズが小さいため、ダウンロードできないという事態は想像し難い。しかし万が一そういう事態に遭遇したとしても、自分の PC に一旦ダウンロードすることなく、Galaxy[22]で解析するやり方もある。それが**図 3** の右端に見えている③ FASTQ files（Galaxy）という名前の列を利用するやり方である。ENA 上の FASTQ ファイルを直接 Galaxy 上にアップロードする手段と理解すればよく、具体的なやり方については次回以降で述べる予定である。

尚、DRA 上で検索を行う場合には、検索する場所にも注意してほしい。エンドユーザから見れば GSE107337 は Accession 番号であるが、DRASearch の Accession という場所で GSE107337 を検索しても何も見つからない [W23]。しかし、Keyword という場所で GSE107337 を検索すると 3 件ヒットする [W24]。検索結果から辿れる SRP125628 を眺めることで、確かに SRA・SRX・SRS などのメタデータ情報の存在を確認することができるのである。この事実から予想できるように、DRA 上の SRP125628 の URL（http://ddbj.nig.ac.jp/DRASearch/



Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)
PRJNA419802	SAMN08098216	SRS2714081	SRX3422361	SRR6322562	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098215	SRS2714083	SRX3422362	SRR6322563	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098214	SRS2714082	SRX3422363	SRR6322564	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098213	SRS2714084	SRX3422364	SRR6322565	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098212	SRS2714085	SRX3422365	SRR6322566	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098211	SRS2714086	SRX3422366	SRR6322567	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098210	SRS2714087	SRX3422367	SRR6322568	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098218	SRS2714088	SRX3422368	SRR6322569	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098217	SRS2714089	SRX3422369	SRR6322570	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2

図 3. ENA 上の GSE107337 のページ

study?acc＝SRP125628）は、DRASearch の Accession という場所で SRP125628 を検索した結果としても得られる［W25］。原著論文から得られる情報は GSE107337 ではあるものの、現実にはそこから辿れる別の ID である SRP125628 を利用するほうが有意義なのである。また、第 3 回では DRA 上で FASTQ ファイルの提供が遅れている例を示したが、今回の SRP125628 は FASTQ と SRA ファイルの両方の提供がまだなされていない例でもある（2019 年 2 月 1 日調べ）。

RNA-seq データ（GSE107337）の詳細

GSE107337 の原著論文[16]では、残念ながら Illumina MiSeq を用いた塩基配列データ取得以降の詳細な解析手順が示されていない。論文中に記載されていることは “Data analysis was performed using EdgeR, Bioconductor components in R packages with LGG as the reference genome.” のみである（LGG は *L. rhamnosus* GG のこと）。LGG の酸ストレスの影響を調べたこの研究では、計 3 状態の RNA-seq データを各状態につき 3 反復ずつ取得して比較している［W26］。具体的には、酸ストレス短期暴露条件群（pH4.5_1h）と酸ストレス長期暴露条件群（pH4.5_24h）を、対照群（pH7_CCG）と比較する実験デザインである。**表 1** は、サンプルごとの SRR ID、リード数、そして gzip 圧縮状態でのファイルサイズ（単位はバイト）を示したものである［W27］。リード数とファイルサイズはほぼ比例関係となっており、リード数の最大と最小の間には 10 倍以上の違い（3,869,088 / 301,126＝12.85）があることもわかる。

原著論文[16]中では示されていないものの、論文著者らは GEO の GSE107337 において、データ解析手順に関するより詳細な情報を記載している。例えば、GEO 内で

表 1

サンプル名	SRR ID	リード数（片側のみ）	ファイルサイズ（forward 側）	ファイルサイズ（reverse 側）
pH4.5_1h_rep1	SRR6322562	301,126	53,833,728	59,662,336
pH4.5_1h_rep2	SRR6322563	1,470,602	265,838,592	296,910,848
pH4.5_1h_rep3	SRR6322564	1,760,461	319,807,488	354,557,952
pH4.5_24h_rep1	SRR6322565	1,375,368	241,553,408	268,341,248
pH4.5_24h_rep2	SRR6322566	3,869,088	682,504,192	759,697,408
pH4.5_24h_rep3	SRR6322567	1,795,874	315,613,184	353,099,776
pH7_CCG_rep1	SRR6322568	3,095,834	551,383,040	615,870,464
pH7_CCG_rep2	SRR6322569	2,570,876	451,117,056	510,308,352
pH7_CCG_rep3	SRR6322570	846,623	148,500,480	167,350,272

pH4.5_1h_rep1 のサンプル ID に相当する GSM2864941 を眺めると、リファレンスゲノム配列として NC_013198.1 を用いていることがわかる［W28］。また、リファレンスゲノムへのリードのマッピング（アラインメント）手段として Bowtie [23] が用いられており、**RPKM 値** [24] の作成には edgeR パッケージ [25] が利用されていることもわかる［W29］。発現変動遺伝子（以下、DEG）同定の入力データとして用いられる**カウントデータ**は、マッピング結果ファイルとアノテーション情報を入力として作成される。一般的なカウントデータの形式は、各行が遺伝子、各列がサンプルからなる数値行列である。カウントデータファイル中の行数は、アノテーションファイル中の遺伝子数に応じて変化する。例えば、遺伝子数が 2,944 個の場合は、2,944 行分の数値情報を含むカウントデータファイルとなる。アノテーションファイル中の遺伝子領域内にマップされたリード数をカウントしたデータであることが、カウントデータと呼ばれる所以である。

当然ながら、アノテーションファイル中の情報は、マッピング時に用いたリファレンスゲノムと完全に対応していなければならない。最初のほうでゲノム配列の違いを議論した FM179322 と AP011548 を例にとると、「FM179322 由来のリファレンスゲノム配列にマッピングした結果ファイル」と「AP011548 由来のアノテーションファイル」では同一遺伝子の座標情報が（一部については同じかもしれないが）異なる。そのため、当然ながら正しいカウント情報を取得することができない。大抵の場合は、カウントデータ取得時にアノテーションファイル中で示された遺伝子領域の座標情報がリファレンスゲノム中に存在しないなどの理由とともに正しくエラーを出してくれるものの、気を付けるべき大事なポイントである。

RPKM 値は、「カウント情報」と「遺伝子の長さ情報」を入力として、「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」と「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」の両方の目的を達成するために補正された数値である。これは、例えば前者の同一遺伝子のサンプル間比較を、**表 1** で示した「リード数最小の pH4.5_1h_rep1」と「リード数最大の pH4.5_24h_rep2」間で行う場合を考えるとよい。もしマップされたカウント数のみからなるカウントデータを補正なしで用いると、おそらくほとんど全ての遺伝子のカウント数が pH4.5_24h_rep2 ＞ pH4.5_1h_rep1 となるであろう。また、RNA-seq は転写物の断片配列をシークエンスしたものであるため、原理的に長い転写物ほどカウント数が多くなる傾向にある。それゆえ、遺伝子の長さ情報を用いた補正なしに遺伝子間のカウント数を比較すると、長い転写物ほど多く発現しているという誤った結果を導いてしまう。このような**基本的な補正の考え方を具体化したものが RPKM 値**であり、広く普及している。

今後の予定

GSE107337 論文 [16] の著者らは、GEO のサイト上でカウントデータと RPKM のファイルを提供している［W30］。両者ともに反復データの平均値しか示されていないものの［W31］、論文中の記載内容と一致した計算結果になることを確認済みである［W32］。次回以降は、この乳酸菌 RNA-seq データ（GSE107337）を用いて、Galaxy 上でクオリティコントロール→リファレンスゲノム配列へのマッピング→カウントデータ取得までを行う。また、RPKM 値の算出や DEG 同定も行い、原著論文の結果と比較検討する。さらに発展的な解析として、発現パターン分類 [26] やシルエットスコアの計算 [27] なども行う予定である。

おわりに

今回のストーリー展開は、実際に行った作業の流れ（RNA-seq データを取得したのち、同じ菌株のゲノム配列およびアノテーションファイルを取得）とは異なる。また、ターゲットとした *L. rhamnosus* GG のゲノム配列を取得する際に、我々は実際には先に AP011548 の原著論文 [8] を発見した。そして論文中に記載されていたゲノムサイズや遺伝子数が Ensembl Bacteria [15] 中の情報と明らかに異なっている点に疑問をもち、試行錯誤した末に FM179322

の原著論文[7]を発見した。ある程度の全体像を把握した後から振り返ってみれば、AP011548の原著論文[8]を通読すればよかっただけである。本連載初期（第3～5回）に用いていたRNA-seqデータ（SRR616268）が不適切である可能性については、第5回原稿執筆時に認識はしていた。しかし、そうだと確信したのはさらに数年後である。

次回以降本格的に利用予定の乳酸菌RNA-seqデータ（GSE107337）についても、一抹の不安は残っている。このデータはIllumina MiSeqで取得されているが、著者らの経験上MiSeqは配列解析用がほとんどであり、発現解析に用いられるのは稀だからである。著者らは既にGalaxy上で1サンプル分（SRR6322562）のカウントデータ取得まで終えており、GEOのサイト上で提供されているカウントデータ（GSE107337_RawCounts.csv）と似た傾向を示すところまでは確認した。しかし、**表1**からもわかるように、これまでよく発現解析に用いられてきたIllumina HiSeqシリーズの一般的なリード数（数千万～数億リード）よりも1～2桁少ない。もちろん遺伝子数（～3000遺伝子）もヒトなどと比べて1桁少ないが、今後新たな問題に遭遇する可能性がゼロでないことは明記しておきたい。

重要なお知らせとして、第7回で紹介したDDBJ Pipeline[28]のサービスが2019年2月中旬をもって終了する。DDBJ Pipelineユーザは、メールによるお知らせを2月初めに受け取っているであろう。また、本連載のウェブページを（Rで）塩基配列解析から（Rで）塩基配列解析のサブに変更している。もちろん主要な項目についてはこれまでのリンク先からも辿れるようにしているが、本連載開始（2014年）から5年が経過し、これまで紹介してきた膨大な情報の中にはリンク切れなど様々な不具合も生じはじめているであろう。我々も継続的に注意深くチェックしているものの、もし発見したら些細なことでもよいのでぜひ指摘してほしい。

謝　辞

本連載の一部は、JSPS科研費JP15K06919および18K11521の助成を受けたものです。

参考文献

1) 孫建強，三浦文，清水謙多郎，門田幸二（2015）次世代シーケンサーデータの解析手法：第3回 Linux環境構築からNGSデータ取得まで．日本乳酸菌学会誌 **26**: 32–41.
2) 孫建強，湯敏，清水謙多郎，門田幸二（2015）次世代シーケンサーデータの解析手法：第4回クオリティコントロールとプログラムのインストール．日本乳酸菌学会誌 **26**: 124–132.
3) 孫建強，清水謙多郎，門田幸二（2015）次世代シーケンサーデータの解析手法：第5回アセンブル、マッピング、そしてQC．日本乳酸菌学会誌 **26**: 193–201.
4) Broadbent JR, Neeno-Eckwall EC, Stahl B, Tandee K, Cai H, et al. (2012) Analysis of the Lactobacillus casei supragenome and its influence in species evolution and lifestyle adaptation. BMC Genomics **13**: 533.
5) 門田幸二，孫建強，湯敏，西岡輔，清水謙多郎（2014）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．日本乳酸菌学会誌 **25**: 87–94.
6) 孫建強，湯敏，西岡輔，清水謙多郎，門田幸二（2014）次世代シーケンサーデータの解析手法：第2回 GUI環境からコマンドライン環境へ．日本乳酸菌学会誌 **25**: 166–174.
7) 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.
8) Morita H, Toh H, Oshima K, Murakami M, Taylor TD, et al. (2009) Complete genome sequence of the probiotic Lactobacillus rhamnosus ATCC 53103. J Bacteriol **191**: 7630–7631.
9) 谷澤靖洋，神沼英里，中村保一，遠野雅徳，大崎研，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7回ロングリードアセンブリ．日本乳酸菌学会誌 **27**: 101–110.
10) 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.
11) Maizel JV Jr, Lenk RP. (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc Natl Acad Sci U S A **78**: 7665–7669.
12) Sonnhammer EL, Durbin R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene **167**: GC1–10.
13) Krumsiek J, Arnold R, Rattei T. (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics **23**: 1026–1028.
14) Charif D, Thioulouse J, Lobry JR, Perrière G. (2005) Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics **21**: 545–547.
15) Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, et al. (2019) Ensembl 2019. Nucleic Acids Res **47**: D745–D751.
16) 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.
17) Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, et al. (2019) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **47**: D23–D28.
18) Clough E, Barrett T. (2016) The Gene Expression Omnibus Database. Methods Mol Biol **1418**: 93–110.
19) 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.
20) Kodama Y, Mashima J, Kosuge T, Kaminuma E, Ogasawara O, et al. (2018) DNA Data Bank of Japan: 30th anniversary. Nucleic Acids Res **46**: D30–D35.
21) Zhu Y, Stephens RM, Meltzer PS, Davis SR. (2013) SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinformatics **14**: 19.
22) 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.
23) Langmead B, Trapnell C, Pop M, Salzberg SL. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10**: R25.
24) Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.

(2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5**: 621-628.
25) Robinson MD, McCarthy DJ, Smyth GK. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139-140.
26) Tang M, Sun J, Shimizu K, Kadota K. (2015) Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. BMC Bioinformatics **16**: 361.
27) Zhao S, Sun J, Shimizu K, Kadota K. (2018) Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results. Biol Proced Online **20**: 5.
28) Nagasaki H, Mochizuki T, Kodama Y, Saruhashi S, Morizaki S, et al. (2013) DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. DNA Res **20**: 383-390.

Methods for analyzing next-generation sequencing data XIII. RNA-seq analysis (Part 1)

Tomoko Terada[1], Mitsuo Sakamoto[2], Kentaro Shimizu[1, 3], and Koji Kadota[1, 3]

[1] Graduate School of Agricultural and Life Sciences, The University of Tokyo.
[2] Microbe Division/Japan Collection of Microorganisms, RIKEN BioResource Research Center.
[3] 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 first introduce two reports regarding the complete genome sequence of *L. rhamnosus* GG. We next compare the two genome sequences by using dotplot with Gepard and confirm an inverted region. We obtain an RNA-seq dataset (SRP125628 or GSE107337) which examined the acid stress response of this strain and explain the experimental design. The content of this manuscript is necessary to perform the data analysis which will be explained next time onwards. Supplementary materials are available online at: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq2.html#about_book_JSLAB.