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18回 遺伝子発現データのクラスタリング

牧野 磨音[1]、清水 謙多郎[1, 2, 3]、門田 幸二[1, 2, 3*]

[1] 東京大学 大学院農学生命科学研究科
[2] 東京大学 大学院情報学環・学際情報学府
[3] 東京大学 微生物科学イノベーション連携研究機構

遺伝子発現データは、各行に遺伝子、各列にサンプルが配置された数値行列のことを指し、各要素には対応するサンプル中の遺伝子がどれだけ働いているかを表す発現量の数値情報が格納されている。本稿では、第15回で得た2,949遺伝子×9サンプルからなる *Lactobacillus rhamnosus* GGの酸ストレス応答を調べた数値行列データを用いる。まず、データ解析環境RStudioの基本的な利用法として、パッケージのインストールやロードといった基礎的な事柄を述べる。次に、似た発現パターンを示すサンプルのクラスタリングについて、その意義や目的、そして結果の解釈について述べる。遺伝子のクラスタリングについては、RNA-seq用の代表的なパッケージであるMBCluster.Seqについて、その概要と入出力形式を述べる。最後に、我々が最近開発したMBCluster.Seqの改良版という位置づけのMBCdeg法について紹介する。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）中のウェブ資料（以下、W）を併用してほしい。

Key words：clustering, RNA-seq, *Lactobacillus rhamnosus* GG, RStudio

はじめに

RStudioは、フリーソフトウェアRの統合開発環境である。大まかには、内部的にRを動かして、R単体での利用よりもさらに便利で使いやすくしたものという理解でよい。連載第1回[1)]の図2では、R単体の実行結果のみを示している。第17回[2)]では、アグリバイオインフォマティクス教育研究プログラムでR（RStudioを含む）を学べる科目やRの位置づけについて述べている。しかしRStudioを用いた実際のデータ解析については、本連載の枠組みではこれまで解説してこなかった。従って本稿では、**今回初めてRとRStudioをインストールし終えたばかりのRStudio初心者が、管理者としてRStudioを起動した状態を想定してスタート**する（W01）。インストール済みのRStudioがあればそれで実行してみてもよいが、R ver. 4.1.0以前のもので不具合が生じた場合は、最新版をインストールして再度試してほしい。

今回用いるのは、数値行列データ（JSLAB18.xlsx）とRスクリプト（JSLAB18.R）の2つのファイルである（W02）。本稿では**デスクトップ上に作成したhogeフォルダ内にこれらのファイルが存在するという前提**で話を進めるが、RStudioに慣れている読者は適宜変更してもらっても構わない。

作業ディレクトリの変更

RStudio起動後に、通常最初に行うのは作業ディレクトリの変更である（W03）。この作業は、JSLAB18.xlsxといったような名前のみで指定したファイルを「デスクトップ上にあるhogeフォルダ」内に限定して探索せよという宣言をRStudioに対して行っているのだと解釈すればよい。こうすることで、中身が異なる同じ名前の複数のファイルがPC内の他のフォルダ内に存在する場合でも、問題

*To whom correspondence should be addressed.
Phone ：+81-3-5841-2395
Fax ：+81-3-5841-1136
E-mail ：koji.kadota@gmail.com

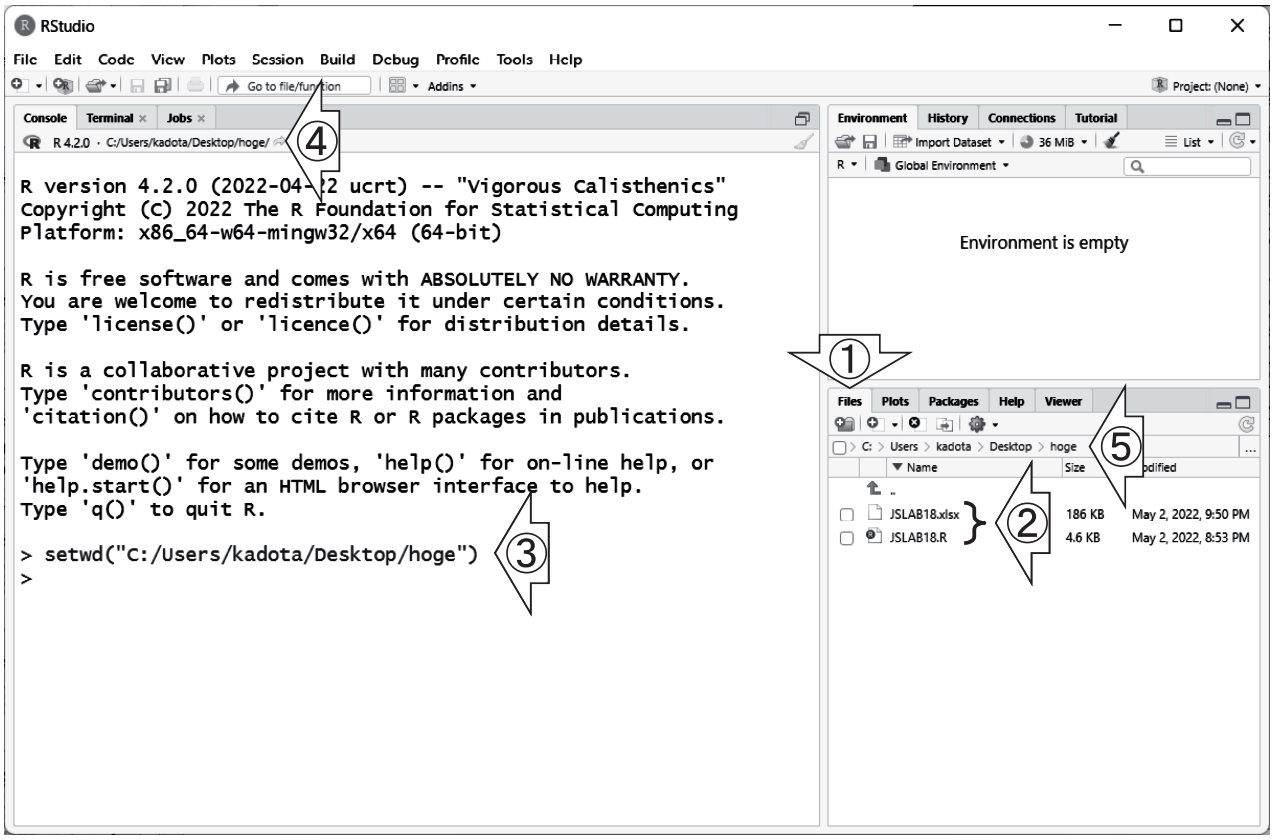


図 1. 作業ディレクトリ変更後の RStudio 画面

なく目的のファイルを一意に定めることができる。

図 1 は、作業ディレクトリ変更後の RStudio の画面である。① Files タブ上で見えている② 2 つのファイルが今回用いるものである。作業ディレクトリの変更で行った一連の操作は、③のコマンド実行と同義である。③の実行と同時に④の部分も変更され、④の場所が作業ディレクトリである。④と⑤は実質的に同じ場所であるが、⑤は単に RStudio 経由で任意のフォルダを指定して表示させているだけである。しかし、**初心者が遭遇するエラーは、「作業ディレクトリの変更忘れ」と「解析したいファイルが作業ディレクトリ内にそもそも存在しない（JSLAB18.xlsx.txt のような感じでファイルのダウンロード時に余分な拡張子がついてしまう不幸な事例を含む）」が原因**であることが多い。そのため、④と⑤が一致しているか、そして解析したいファイルが②の場所にリストアップされているかに多くの注意を払うことを勧める。

エディタの起動

R スクリプトとは、R 上で実行可能な一連のコマンドを 1 つ 1 つ逐次に実行できる形でまとめたものである。多くの場合、行ごとに実行可能であり、一定のまとまりごとに、そのまとまりがどのような役割を果たすのかを説明するコメントがつけられている。シャープ（#）より右側の文字は全て無視されるので、例えば # から始まる行は実行してもしなくてもよい。スクリプトはソースコードとも呼ばれ、拡張子として .R がつけられる場合がほとんどである。中身自体はテキストファイルであるため拡張子を .txt にしても特に問題はない。しかし RStudio 上で中身が R スクリプトであることが明示されている .R ファイルとして開くことで、文法的に問題がある箇所を警告してもらえるなどのメリットがある。従って、**R スクリプトファイルは RStudio などの R 専用のエディタ上で開くことを強く推奨**する。

図 2 は、R スクリプトファイル（JSLAB18.R）を RStudio 上で開いた状態である（W04）。作業自体は① Files タブ上で見えている②当該ファイルをクリックするだけである。③ JSLAB18.R という名前のタブが開かれ、ファイルの中身が左側の領域の上半分を使って表示されていることがわかる。これが R のエディタ画面である。一方、**図 1** で左側の領域全体を占めていて、**図 2** で左側の下半分になった部分は④コンソール（Console）と呼ばれるもの（正確にはペインとよぶ）である。**③スクリプトファイル中の実行したい行または領域を選択して、④ Console 画面上で実行し結果を眺めるというのが基本形**である。新規に出力ファイルを生成するスクリプトを実行した場合は、① Files タブ上で正しく生成されたかどうかを確認することもできる。

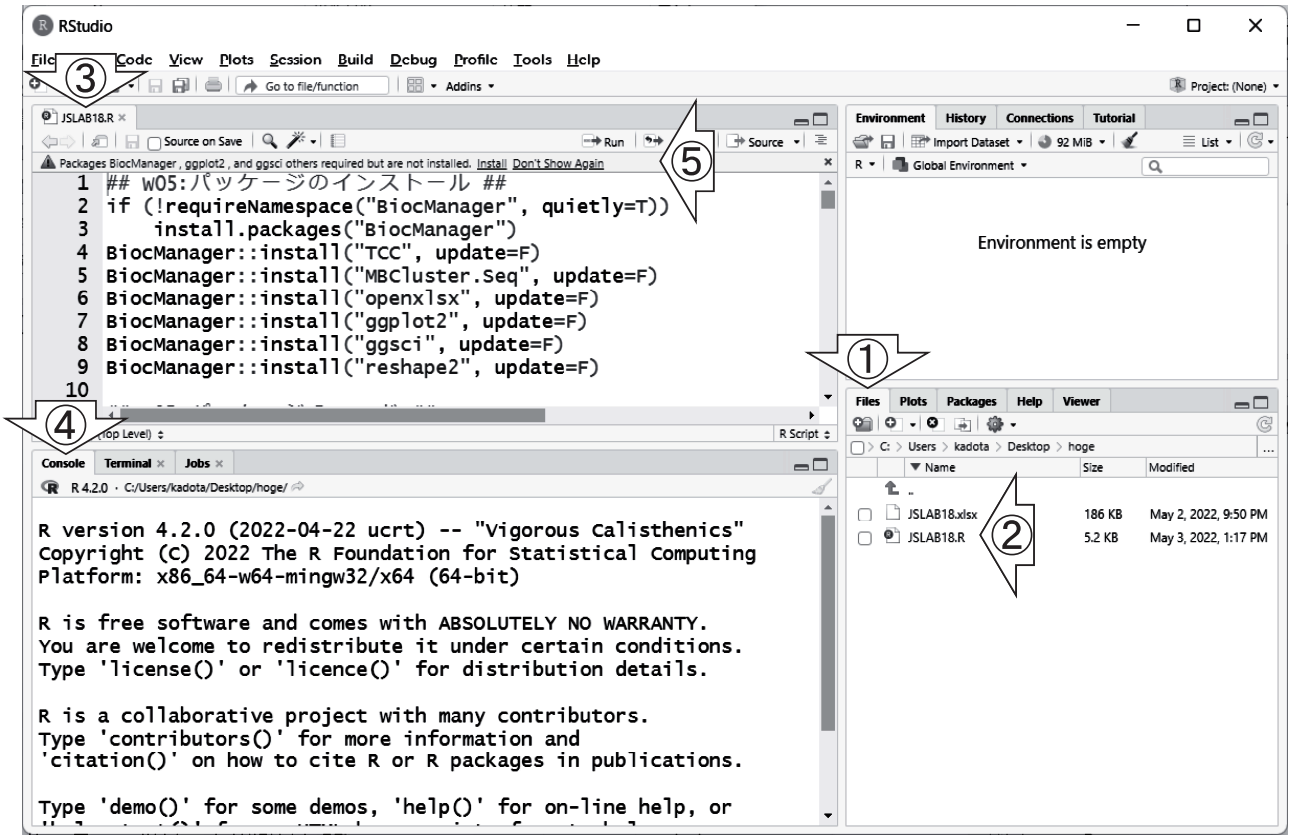


図 2. スクリプトファイル（JSLAB18.R）を開いた状態の RStudio 画面

パッケージのインストール

第 1 回[1]のおさらいを一部含むが、R と RStudio をインストールしただけの状態では機能が限られている。これは、新しい PC を購入しただけの状態と同じだと考えればよい。PC に Zoom や Slack などをインストールして利用するのと同様、RStudio の機能を有効利用するためには、CRAN や Bioconductor[3]など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。具体的には、今回入力ファイルとして用いる xlsx 形式のファイルを RStudio で読み込むためには、例えば CRAN から提供されている openxlsx というパッケージを予めインストールしておかねばならない。

一部の読者は、PDF ファイルを開くためのソフトウェアがインストールされていない状態で、手元にある PDF ファイルを開こうとした経験があるのではないだろうか。そのような場合、Acrobat Reader をインストールしますか？のようなメッセージが出て、ほぼ無意識に「はい」を選択するようなステップを経たはずである。最近の RStudio は高機能化が進んでおり、開いたスクリプトファイルを調べてインストールが必要なパッケージを提示し、インストールするかどうかを問い合わせてくれるようになっている。それが、**図 2** の⑤の部分に出現しているメッセージである。

図 2 の左上に見えているスクリプトファイル（JSLAB18.R）中の 2 ～ 9 行目のコマンドは、計 7 個のパッケージ（BiocManager, TCC[4], MBCluster.Seq[5], openxlsx, ggplot2, ggsci, and reshape2）のインストール作業に相当する。前述のとおり、これらのコマンドの大部分、具体的には Bioconductor から提供されている TCC 以外の 6 パッケージのインストール作業は、⑤のポップアップメッセージに対して Install を選択することでまかなえる。しかし、初心者の視点で見ると、どのパッケージがどのリポジトリ（CRAN や Bioconductor のようなパッケージ提供サイトのこと）から提供されているかを把握するのは難しい。また、この **2 ～ 9 行目のコマンドは、リポジトリに依存せず統一的にパッケージをインストール可能なやり方**であるため、今回はこのやり方で行う（⑤のポップアップは Don't Show Again をクリックする）。

図 3 は、パッケージのインストール作業終了後の状態である（W05）。① JSLAB18.R 中の 2 ～ 9 行目を反転させて ② Run ボタンを押す作業は、③コンソール画面上でのこれら一連のコマンド実行に相当する。終了したかどうかは、④の部分がコマンド入力待ち状態（コマンドプロンプトと呼ばれる > が出ている）かどうかで判断できる。⑤のフレーズ（「reshape2 は無事に…」）から、⑥の最後のコマンドまで成功裏に終わったことが示唆される。⑥のコマンドは、「BiocManager パッケージが提供する install という関数を

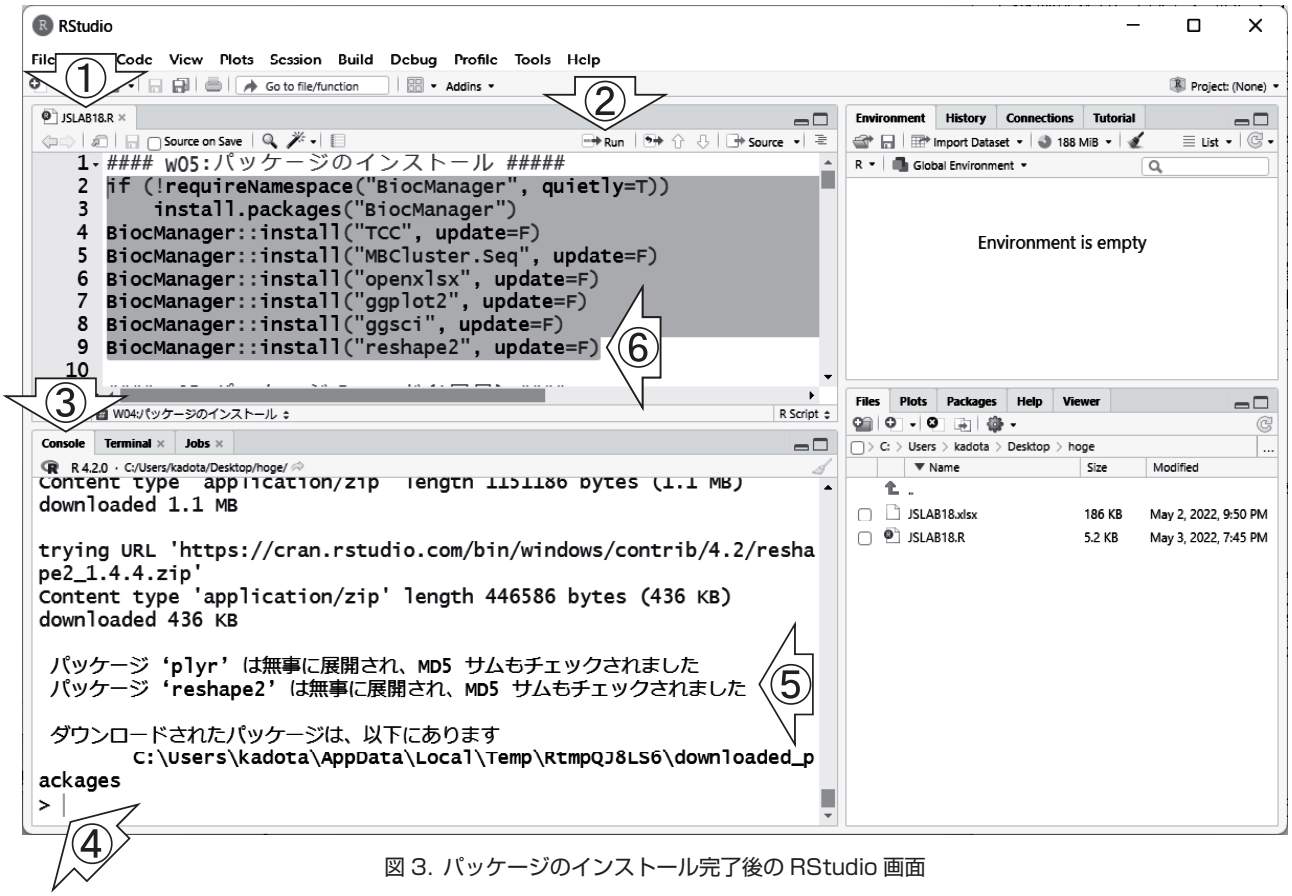


図 3. パッケージのインストール完了後の RStudio 画面

用いて、reshape2 というパッケージを update＝F というオプションで実行せよ」という意味である。尚、⑤の MD5 サムは、第 3 回[6] でも解説しているが、ダウンロードしたファイル（この場合は reshape2）が提供元と完全に一致しているかどうかを確認するためのものである。

パッケージのロード

パッケージのロードは、インストールしたパッケージを読み込んで、使える状態にする作業である（W07）。大まかには、PC にインストールしたソフトウェアの起動のようなものだという理解でよい。R の場合は、複数個のパッケージを一度にインストールすることが多いため、**図 3** 左下のコンソール画面の結果のみでは全てのパッケージが無事インストールできているか確認しきれない。それゆえ、我々は通常、当該パッケージ群のロードを通じてインストールの成否を確認する。

図 4 は、スクリプトファイル（JSLAB18.R）中の 12 ～ 25 行目のコマンド実行結果である。① 11 行目と② 19 行目のコメントからも想像できるが、同じ作業を 2 回行っていることがわかる。この理由は、① 1 回目（12 ～ 17 行目）の実行のみでは、大量のメッセージが表示されて判読しづらいこと、実用上問題ない警告メッセージが表示されることがある点が挙げられる。特に後者はそのまま進めて問題ないかどうかの判断自体が初心者には難しいが、② 2 回目（20 ～ 25 行目）の実行結果として、③のように**2 回目で何もメッセージが表示されていなければ基本的に問題ない**と判断してよい。ここまでで RStudio の解析環境構築は完了である。

データの概要

今回入力データとして用いるファイル（JSLAB18.xlsx）は、第 13 ～ 15 回で解説した *L. rhamnosus* GG（LGG）の酸ストレス応答を調べたカウントデータである[7–9]。この研究では、以下に示す計 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[10]。他分野の読者は、例えば「低濃度抗がん剤投与群、高濃度抗がん剤投与群、対照群」のように解釈してもよいし、よりシンプルに「G1 群、G2 群、G3 群」の 3 群間比較だとみなしてもよい。

- 酸ストレス短期暴露群（pH4.5_1h）
- 酸ストレス長期暴露群（pH4.5_24h）
- 対照群（pH7_CCG）

図 5 は、JSLAB18.xlsx の概要である。1 行目は列名の情報からなるヘッダー行、① 1 列目は行名（遺伝子 ID）情報から構成されていることがわかる。これが典型的な入力

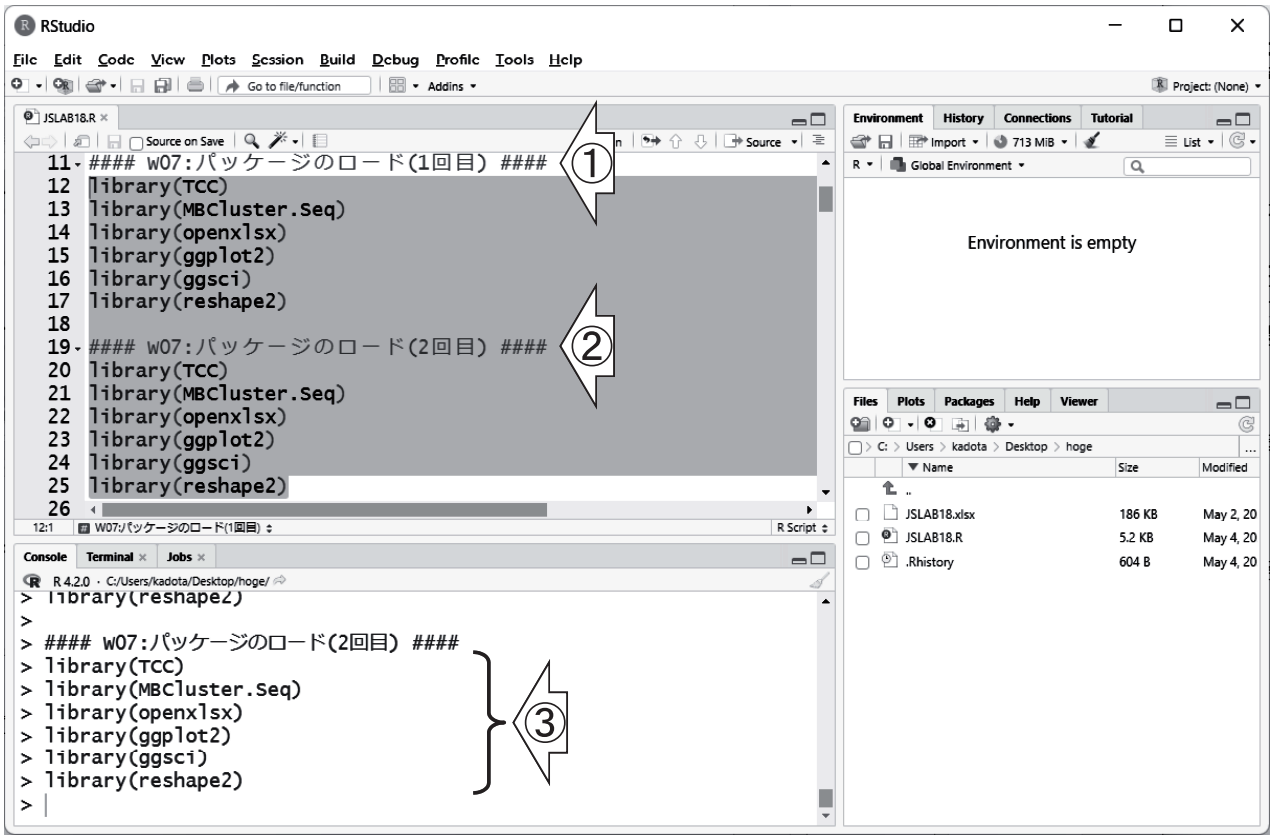
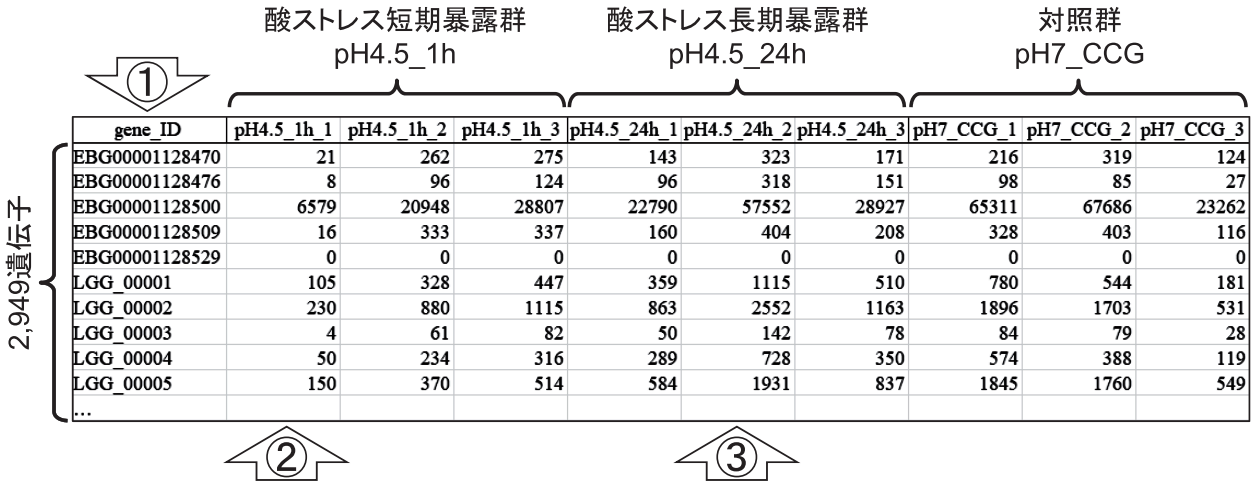


図 4. パッケージのロード後の RStudio 画面



gene_ID	pH4.5_1h_1	pH4.5_1h_2	pH4.5_1h_3	pH4.5_24h_1	pH4.5_24h_2	pH4.5_24h_3	pH7_CCG_1	pH7_CCG_2	pH7_CCG_3
EBG00001128470	21	262	275	143	323	171	216	319	124
EBG00001128476	8	96	124	96	318	151	98	85	27
EBG00001128500	6579	20948	28807	22790	57552	28927	65311	67686	23262
EBG00001128509	16	333	337	160	404	208	328	403	116
EBG00001128529	0	0	0	0	0	0	0	0	0
LGG_00001	105	328	447	359	1115	510	780	544	181
LGG_00002	230	880	1115	863	2552	1163	1896	1703	531
LGG_00003	4	61	82	50	142	78	84	79	28
LGG_00004	50	234	316	289	728	350	574	388	119
LGG_00005	150	370	514	584	1931	837	1845	1760	549
…									

図 5. 入力データ（JSLAB18.xlsx）の概要

ファイルの形式であり、実際のデータは 2,949 行×9 列の数値行列である。② pH4.5_1h_1 というサンプル名のデータが他に比べて全体的に数値が小さいことに気づくが、これは第 13 回の表 1 を見れば納得できる。つまり、このデータは、生のリード数の点で他よりも明らかに少ないためであろう。逆に、③ pH4.5_24h_2 が全体的に大きな数値となっているのは、リード数やファイルサイズが最大値であることから妥当である。**重要な点は、本格的なデータ解析を進めていく際に、これらの事象が何らかの悪影響を与えている可能性を排除しない**ことであろう。

サンプルのクラスタリング

我々は通常、探索的データ解析の一環として、まずサ

図 6. サンプルクラスタリング実行後の RStudio 画面

ンプルのクラスタリングを実行する（W08）。これは、全体的なサンプル間の類似度を評価する枠組みであり、前述のような全体的な数値の大小の違いが同一群内の反復データ間の類似度に悪影響を与えていないかなどの観点で結果を眺める。ここでは、TCC パッケージが提供する clusterSample 関数を用いてサンプルのクラスタリングを行う。サンプル間の類似度を順位尺度である Spearman 相関係数で定めているため、全体的な数値の大小に影響されないのが特徴である。

図 6 は、28 ～ 36 行目のコマンド実行結果である。コメント行（28, 30, 33 行目）の記載内容からもわかるが、① 29 行目で入力ファイル（JSLAB18.xlsx）を読み込み、② 34 行目で出力ファイル名（fig7.png）を定めていることがわかる。③実行時にエラーメッセージが出ていない（null device というのは特に気にしなくてもよい）こと、④作業ディレクトリに fig7.png が作成されていることもわかる。

入力ファイル読み込み時に利用している read.xlsx 関数は、**図 4** でロードした openxlsx パッケージが提供しているものである（から利用できる）。①で読み込んだ情報は、左向きの矢印の先にある data という名前のもの（これをオブジェクトという）に格納されている。この中身の概要は、⑤ Environment タブで見えている⑥ data のところに示されている（W09）。「2949 obs. of 9 varia…」の意味は、このデータが 2,949 行×9 列の数値行列であることが既知の状態で見れば容易に理解できるだろう。

図 7 は、サンプルのクラスタリング実行結果ファイル（fig7.png）の中身である。大きさがわかるように外枠を示しているが、**図 6** の②で指定した縦横比（500×400 ピクセル）になっていることがわかる。①酸ストレス長期暴露群（pH4.5_24h）は 3 つの反復データのみで 1 つのクラスタを形成している一方で、②対照群（pH7_CCG）のクラスタ中に③ pH4.5_1h_1 が紛れ込んでいるのがわかる。我々は、**もし③のデータが他と比べて特に変わったことがなければ、図 7 の結果を素直に受け止めていた**であろう。

しかしながら実際には、③ pH4.5_1h_1 は他と比べて極端に全体的に数値が小さいことや、リード数が少ないということを確認している。また、クラスタリングアルゴリズムとしても、サンプル間の類似度を定義する数式として、ユークリッド距離のような数値ベクトル中の対応する要素間の差（の二乗や絶対値）を足していくような類のものではないため、手法選択が間違っているとも考えにくい。また、もし③を除いて考えると、大きく「対照群 vs. 酸ストレス暴露群」の 2 つのクラスタとして分かれることや、酸ストレス群の中でもさらに「④短期暴露群 vs. ①長期暴露群」に分かれている。これらの状況を鑑み、**我々は③を外れサンプル**[11] **であると判断**した。

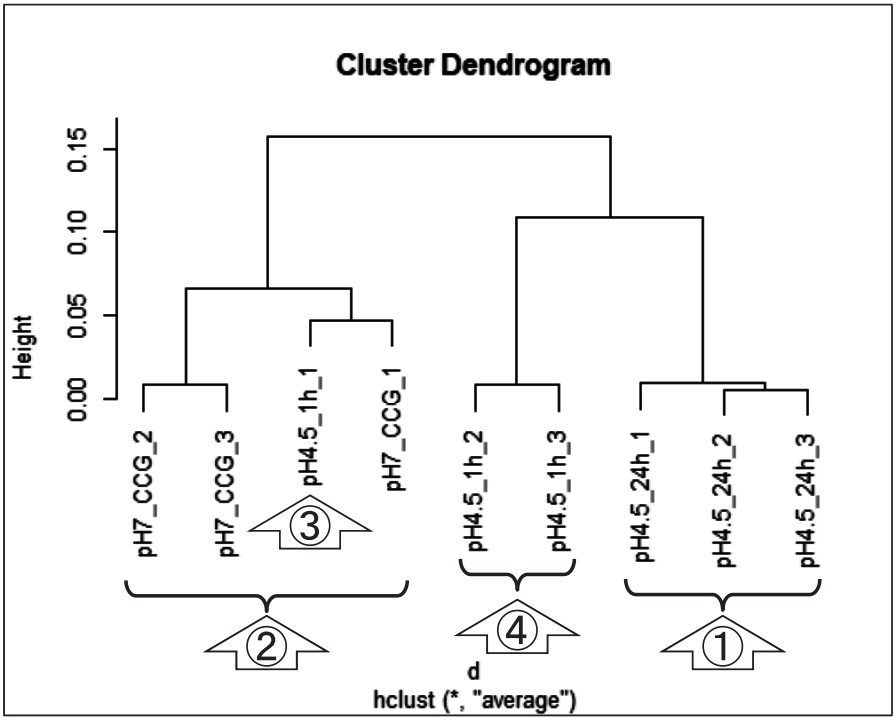


図 7. サンプルクラスタリング実行結果ファイル（fig7.png）

遺伝子のクラスタリング（MBCluster.Seq）

遺伝子のクラスタリングとは、似た発現パターンを示すグループに遺伝子を分類する作業のことである。類似した発現パターンを持つ遺伝子群は、機能的にも類似していることが期待される[12]。例えばある遺伝子の機能が未知である場合、同じクラスタに分類された機能既知の遺伝子群と類似した機能を持つのではないかという推測に利用することができる。

遺伝子発現解析の多くの実験デザインは 2 群間比較（G1 群 vs. G2 群）や 3 群間比較（G1 群 vs. G2 群 vs. G3 群）であり、その主目的は発現変動遺伝子（DEG）の検出である。特に 2 群間比較の場合は、DEG のとりうるパターンは、基本的に G1 群で高発現パターン（DEG_G1）か G2 群で高発現パターン（DEG_G2）のいずれかであるため、遺伝子のクラスタリングを行う概念自体がない。それゆえ**遺伝子クラスタリングは、3 群間比較や時系列データのような、とりうるパターンが複雑になりうる実験デザインの DEG 検出結果のパターン分類目的でしか利用されてこなかった**[13]。

代表的な RNA-seq カウントデータの遺伝子クラスタリング用パッケージは MBCluster.Seq[5] である。このパッケージは、入力として「DEG のカウントデータ」と「想定する発現パターン数に相当するクラスタ数 K」を与えてクラスタリングを実行する。出力は、「クラスタ中心の発現パターン」と「遺伝子ごとの各クラスタへの属しやすさを表す事後確率の数値行列」である。例えば 2 群間比較（G1 群 vs. G2 群）の結果として得られた 100 個の DEG のカウントデータを入力として、$K=3$ として MBCluster.Seq を実行すると、出力結果として「計 3 つのクラスタ中心の発現パターン（3 行×2 列のデータ）」と「100 遺伝子×3 列の事後確率の数値行列」が得られる。

DEG 検出を兼ねた遺伝子のクラスタリング（MBCdeg）

もちろん遺伝子クラスタリングの考え方を DEG 検出そのものに適用するアイデアもごく少数ではあるが存在する[13,14]。そのうちの 1 つが、我々が 2021 年に提唱した、MBCluster.Seq をベースとして DEG 検出を行う戦略（MBCdeg と命名）である[13]。簡単にいえば、（DEG 検出も目的に含まれるので）「全遺伝子のカウントデータ」と「想定する発現パターン数に相当するクラスタ数 K」を与えて MBCluster.Seq を実行するものである。

前述のとおり、MBCluster.Seq 自体は本来 DEG 検出を想定しておらず、DEG 検出後のパターン分類目的で用いるものである。しかし**全遺伝子のカウントデータを入力として MBCluster.Seq を実行すること自体は手続き的に可能**である。この場合、直感的には「DEG ではない遺伝子群（non-DEG）から構成される大きなクラスタ」が形成され、「そのクラスタ中心の発現パターンはフラット（比較する群間でほぼ同じ値）」であることが期待される。我々の MBCdeg は、この性質を利用して non-DEG クラスタを同定し、そのクラスタに属する事後確率が低い順に遺伝子をランキングするというのが基本戦略である。non-

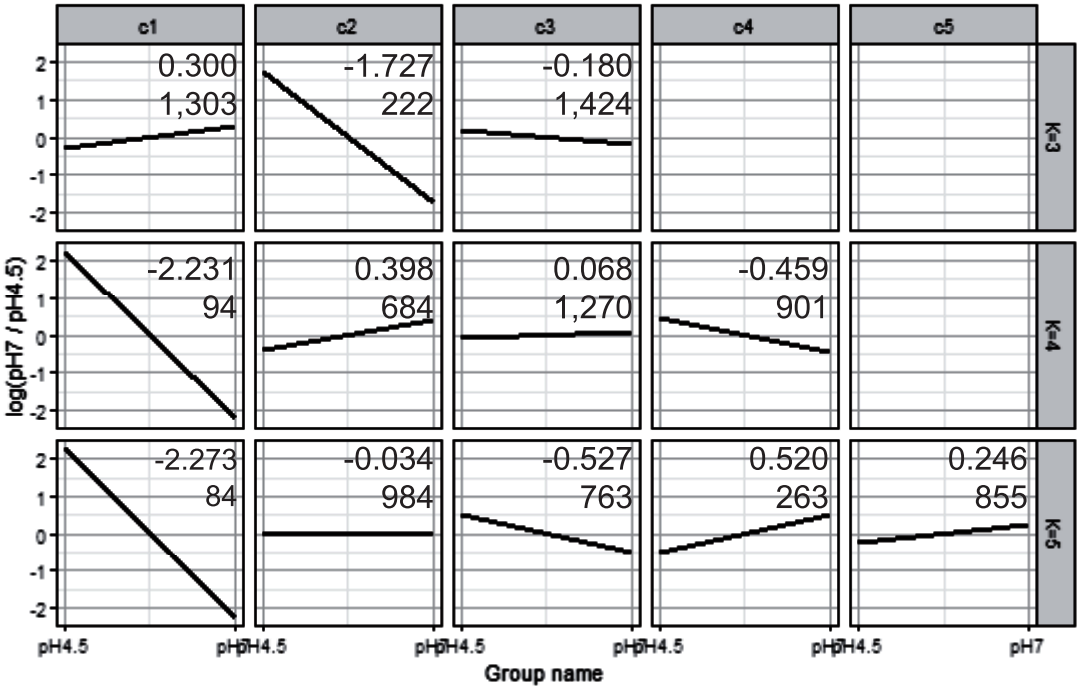


図 8. MBCdeg 実行結果
スクリプト実行結果として自動生成される fig8.png をベースとして、右上部分の数値情報を後から追加している。

DEG クラスタの事後確率は、統計的検定を行った結果として得られる p 値と似たようなものである。つまり事後確率の値が低い遺伝子ほど、non-DEG とみなすには無理がある（つまり DEG である）と解釈すればよい。

MBCluster.Seq は、遺伝子を事後確率が最も高いクラスタに割り当てることで発現パターン分類を行うのが本来の利用法である。non-DEG 以外のクラスタは、当然フラットではない何らかの発現変動パターンをもつため、MBCdeg を使うことで DEG 同定と発現パターン分類が同時にできるというメリットもある。例えば、2 群間比較データを $K=3$ で実行すると、直感的には得られた 3 つのクラスタが「non-DEG、DEG_G1、DEG_G2」のいずれかのパターンを示すことが期待される。それゆえ、MBCdeg の実行結果の段階で、得られた DEG が DEG_G1 と DEG_G2 のどちらのパターンに属するかまで得られることになる。

MBCdeg は、DEG 検出専用の有名なパッケージである edgeR [15] や DESeq2 [16] はもちろんのこと、我々が提供する TCC をも圧倒する性能を示している [13]。あくまでも TCC が生成するシミュレーションデータに基づく結論ではあるものの、データ中に占める DEG の割合や偏りといった様々なシナリオを用いた性能評価の結果として、TCC を上回る性能であったという点が重要である [17]。

MBCdeg の実行

MBCdeg には、データ正規化の部分で MBCluster.Seq のデフォルトの方法を実装した MBCdeg1 と、頑健な RNA-seq データ正規化法である DEGES [18] を実装した MBCdeg2 の 2 種類が存在する。ここでは、単純化のために「酸ストレス長期暴露群（pH4.5_24h）vs. 対照群（pH7_CCG）」という 2 群間比較を MBCdeg2 で実行する。クラスタ数については、I 群間比較の場合は、$K>I+1$ を推奨している。つまりこの場合は $K>3$ が推奨ということになるが、ここでは挙動確認の意味も込めて $K=3, 4, 5$ を独立に実行する（W10）。

図 8 は、クラスタごとのクラスタ中心の発現パターンを計 3 種類の K 値で実行した結果である。縦軸の値は log 比、横軸は左側が pH4.5_24h 群、右側が pH7_CCG 群である。右上の数値は、上が pH7_CCG 群側の log 比の値、下がそのクラスタに属する遺伝子数を表している。真ん中の $K=4$ の結果を例にとると、1 つめのクラスタ（c1）には 94 個の遺伝子が属しており、その代表パターンは pH7_CCG 群が pH4.5_24h 群に比べて $e^{2\times(-2.231)}=0.0115$ 倍高発現（86.66 倍低発現）だと解釈する。2 つめのクラスタ（c2）には 684 個の遺伝子が属しており、その代表パターンは $e^{2\times(0.398)}=2.217$ 倍高発現だと解釈する。3 つめのクラスタ（c3）には 1,270 個の遺伝子が属しており、その代表パターンは $e^{2\times(0.068)}=1.146$ 倍高発現だと解釈する。4 つめのクラスタ（c4）には 901 個の遺伝子が属しており、その代表パター

ンは $e^{2\times(-0.459)}=0.399$ 倍高発現だと解釈する。この場合は、c3 が最も変動の少ない non-DEG クラスタと定義される。

$K=4$ の結果の場合、pH7_CCG 群で低発現パターンを示すクラスタは c1 と c4、高発現パターンのクラスタは c2 だと解釈する。$K=5$ の場合は、non-DEG クラスタは 984 個の遺伝子から構成される c2、pH7_CCG 群で低発現パターンを示すクラスタは 84 個の c1 と 763 個の c3、高発現パターンのクラスタは 263 個の c4 と 855 個の c5 だと解釈する。

一部の読者は non-DEG だと判定されたクラスタを構成する遺伝子数が、$K=4$ の c3 が 1,270 個、$K=5$ の c2 が 984 個と大きく異なっていると思われたかもしれない。主な理由は、$K=4$ の c3 は pH7_CCG 群で若干の高発現を示す遺伝子を含んでいるからだと解釈すればよい。実際、$K=5$ では大きめの高発現パターンを示す c4 が 263 個、小さめの高発現パターンを示す c5 が 855 個となっており、$K=4$ で高発現パターンを示す c2 の 684 個よりも全体として多くなっている。$K=4$ の c3 にカテゴライズされていた一部の若干の高発現を示していた遺伝子群が、$K=5$ とクラスタ数を増やしたことによって、c5 に含められるようになったのだと解釈すればよい。

ただしこれらの解釈は、あくまでも**図 8** の結果を見比べて導き出しただけの想像にすぎない。原著論文レベルのクオリティに仕上げるためには、「In fact, …」で書けるような、実際のところどうだったかを丁寧に調べる必要がある。この場合は、例えば $K=4$ の c3 に属していた遺伝子のどれだけが $K=5$ の c2 や c5 に含まれているのか？　などは確認しておきたいところである。この種のちょっとした解析を自在に行うためには、出力結果のどこにどのような情報が格納されているかを把握したり、それらの情報を用いて集合演算を行うスキルが必要となる。

なお、$K=3$ の結果は、最も変動の少ない c3 が non-DEG クラスタとみなされるものの、その代表パターンは $e^{2\times(-0.180)}=0.698$ 倍 pH7_CCG 群で高発現（1.433 倍低発現）となっている。$K=4$ の c3 が $e^{2\times(0.068)}=1.146$ 倍高発現、そして $K=5$ の c2 が $e^{2\times(-0.034)}=0.934$ 倍高発現（1.070 倍低発現）という結果と比較すると、満足のいく結果とは言い難い。しかしそもそも 2 群間比較の場合は $K>3$ を推奨しており、この意味において原著論文のガイドラインは妥当といえる。

おわりに

本稿の前半は、RStudio の基本的な利用法、パッケージのインストールやロード、そして解析データの全体像を述べた。後半は、RNA-seq カウントデータ専用のパッケージを用いたクラスタリングについて述べた。最初のサンプルクラスタリングでは、パッケージと関数の関係性、ありがちなミスとその対処法、そして結果の解釈のしかたについて解説した。次に代表的な遺伝子クラスタリング用パッケージである MBCluster.Seq の概要を述べるとともに、このパッケージが本来想定していなかった DEG 検出にも拡張して応用可能であり、かつその精度が DEG 検出専用のパッケージをも凌駕することを述べた。最後に、MBCluster.Seq に基づく DEG 検出アルゴリズムである MBCdeg を実行し、指定したクラスタ数の違いによる結果の違いや解釈の仕方について述べた。

指定した K 値ごとの遺伝子の事後確率情報は、今回の W10 のスクリプト実行結果ファイルとして生成される MBCdeg_K ○ .xlsx（○の部分には 3, 4, 5 の数値が入る）中に含まれている。このスクリプトについては、どこにどのような出力結果が格納されているかや、それらを用いてどのように集合演算を行うのかなどを含め、次回以降丁寧に解説していく予定である。

謝　辞

本内容の一部は、JSPS 科研費 21K12120 の助成を受けたものです。

利益相反（COI）

牧野磨音、清水謙多郎、門田幸二：本論文発表の内容に関連して開示すべき COI 状態はない。

参 考 文 献

1) 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.
2) 門田幸二，大森良弘，寺田朋子，三浦文，寺田透，清水謙多郎（2021）次世代シーケンサーデータの解析手法：第 17 回バイオインフォマティクス教育の今後，日本乳酸菌学会誌 **32**: 129–37.
3) Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5**: R80.
4) 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
5) Si Y, Liu P, Li P, Brutnell TP (2014) Model-based clustering for RNA-seq data. Bioinformatics **30**: 197–205.
6) 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.
7) 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.
8) 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**: 153–61.
9) 寺田朋子，清水謙多郎，門田幸二（2020）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），日本

乳酸菌学会誌 **31**: 25–34.
10) 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**: 1604–13.
11) Kadota K, Tominaga D, Akiyama Y, Takahashi K (2003) Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification. Chem-Bio Informatics J **3**: 30–45.
12) Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95: 14863–8.
13) Osabe T, Shimizu K, Kadota K (2021) Differential expression analysis using a model-based gene clustering algorithm for RNA-seq data. BMC Bioinformatics **22**: 511.
14) Vavoulis DV, Francescatto M, Heutink P, Gough J (2015) DGEclust: differential expression analysis of clustered count data. Genome Biol **16**: 39.
15) Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–40.
16) Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15**: 550.
17) 門田幸二，清水謙多郎 (2021) 次世代シーケンサーデータの解析手法：第 16 回なぜ次から次へと新規手法が開発されるのか？, 日本乳酸菌学会誌 **32**: 123–8.
18) Kadota K, Nishiyama T, Shimizu K (2012) A normalization strategy for comparing tag count data. Algorithms Mol Biol **7**: 5.

Methods for analyzing next-generation sequencing data 18. Clustering for gene expression data.

Manon Makino[1], Kentaro Shimizu[1, 2, 3], Koji Kadota[1, 2, 3]

[1]Graduate School of Agricultural and Life Sciences, The University of Tokyo.
[2]Interfaculty Initiative in Information Studies, The University of Tokyo.
[3]Collaborative Research Institute for Innovative Microbiology, The University of Tokyo.

Abstract

Gene expression data refers to a numerical matrix in which genes are arranged in each row and samples in each column, and each element stores numerical information on the amount of expression. In this paper, we use the numerical matrix which measured the acid stress response of *Lactobacillus rhamnosus* GG, consisting of 2,949 genes × 9 samples. We first describe the basic usage of RStudio, an integrated development environment for R. Next, we discuss the significance of clustering samples having similar expression patterns, and the interpretation of the results. For gene clustering of RNA-seq data, we outline MBCluster.Seq, a representative package for the purpose. Finally, we introduce a modified version of the package (called MBCdeg) that can also be used for detecting differentially expressed genes (DEGs). Supplementary materials are available online at https://www.iu.a.u-tokyo.ac.jp/kadota/r_seq2.html.