Japanese Journal of Lactic Acid Bacteria

総　説

次世代シーケンサーデータの解析手法 第20回 RNA-seq カウントデータの性質と統計モデル

牧野 磨音[1]、坂本 光央[2]、清水 謙多郎[1, 3, 4]、門田 幸二[1, 3, 4*]

[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
[3] 東京大学大学院 情報学環・学際情報学府
[4] 東京大学 微生物科学イノベーション連携研究機構

RNA-seq 解析の目的の多くは、比較する状態または群間で発現の異なる遺伝子（DEG）の同定である。ほとんどのプログラムは R のパッケージとして提供されており、その入力は、カウントデータとよばれる各行が遺伝子、各列がサンプルからなる数値行列である。本稿では、なぜ負の二項分布とよばれる統計モデルが DEG 検出目的でよく用いられるのかについて、数式を交えて解説する。また、このカウントデータの性質を説明する手段としてよく用いられる平均－分散プロットについて、データの前処理から ggplot2 による描画まで述べる。

Key words：negative binomial distribution, RNA-seq, R Markdown, ggplot2

はじめに（W1）

前回[1]に引き続き、今回のウェブ資料（以下、W）も R Markdown（JSLAB20.Rmd）で作成している。レンダリングによって作成した R スクリプトと実行結果から構成される html ファイル（JSLAB20.html）も含めて、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）で提供している。今回は特に、JSLAB20.html を併用してほしい。なお、R パッケージのインストールやロード、RStudio の基本的な利用法については、第 18 回[2]でも解説している。

入力として用いる RNA-seq カウントデータは、第 18 回から利用している *Lactobacillus rhamnosus* GG（LGG）の酸ストレス応答を調べた 2,949 遺伝子×9 サンプルの数値行列である[3]。第 18 回の図 5 でも示されているように、（行名の列を除く）最初の 1 ～ 3 列目が酸ストレス短期暴露群（pH4.5_1h）、4 ～ 6 列目が酸ストレス長期暴露群（pH4.5_24h）、そして 7 ～ 9 列目が対照群（pH7_CCG）である。各群につき 3 反復ずつ取得したこの実験デザインにおいて、同一群内の反復データ間のばらつきがどのようになっているかを、要約統計量・数式・散布図などを用いて解説するのが本稿の主な内容である。

DEG 同定は多くの場合、「比較する群間に差がない」という帰無仮説を（形式的に）立て、統計的手法を用いて行われる。この枠組みにおいて、我々は非常に低い p 値が得られることを期待するが、これは「同一群内のばらつきの範囲内」という許容範囲を超えた結果が得られるのを望むことと同義である。本稿のトピックである平均－分散プロットは、同一群内のばらつき度合いを把握する描画手段としてよく用いられる散布図であり、これが帰無仮説に従う分布（つまり帰無分布）である。この分布は同一群内の反復の種類に大きく左右されるため、本稿は反復の種類の説明からスタートする。

生物学的な反復と技術的な反復

反復データには、生物学的な反復（biological replicates）と技術的な反復（technical replicates）の 2 種類が存在する。たとえばヒトの筋肉組織を塊として調べる bulk RNA-seq

*To whom correspondence should be addressed.
Phone ： +81-3-5841-8155
Fax ： +81-3-5841-1136
E-mail ： koji.kadota@gmail.com

の場合、「A さん、B さん、C さんという異なる個体からそれぞれデータを取得して 3 反復としたものが生物学的な反復」に、そして「同一個体（たとえば A さん）の筋肉組織を 3 分割して 3 反復としたものが技術的な反復」に相当する（**図 1a**）。また、筋肉組織を構成する細胞ごとの情報を調べる scRNA-seq の場合、「A さん、B さん、C さんという異なる個体からそれぞれ 1 つの細胞の発現データを取得して3反復としたものが生物学的な反復」に、そして「同一個体（たとえば A さん）の筋肉組織を構成する 3 つの筋肉細胞の発現データを取得して 3 反復としたものが技術的な反復」に対応する。もちろん実際の scRNA-seq では、異なる個体から 1 つだけ、あるいは同一個体から数個の細胞だけのデータを取得することはなく、ある個体の細胞塊に含まれる数百～数千細胞のデータを取得するのが基本形である。

ここまでで直感的に理解できることは、**データのばらつきは技術的な反復よりも生物学的な反復のほうが大きい**ということであろう。これは、発現データ取得手段がマイクロアレイの頃から議論されている実験デザインに関する事柄である[4]。技術的な反復データ間のばらつき（technical variation）は、実験機器や実験者の手技の安定性を計測する一種の指標である。それゆえ、一般にこのばらつきの度合いは、小さければ小さいほどよい。その一方で、生物学的な反復データ間のばらつき（biological variation）は、小さかろうが大きかろうがそれをそのまま受け止めるのが基本である。**重要なのは、どのような別個体のデータを取得しているのか**である。

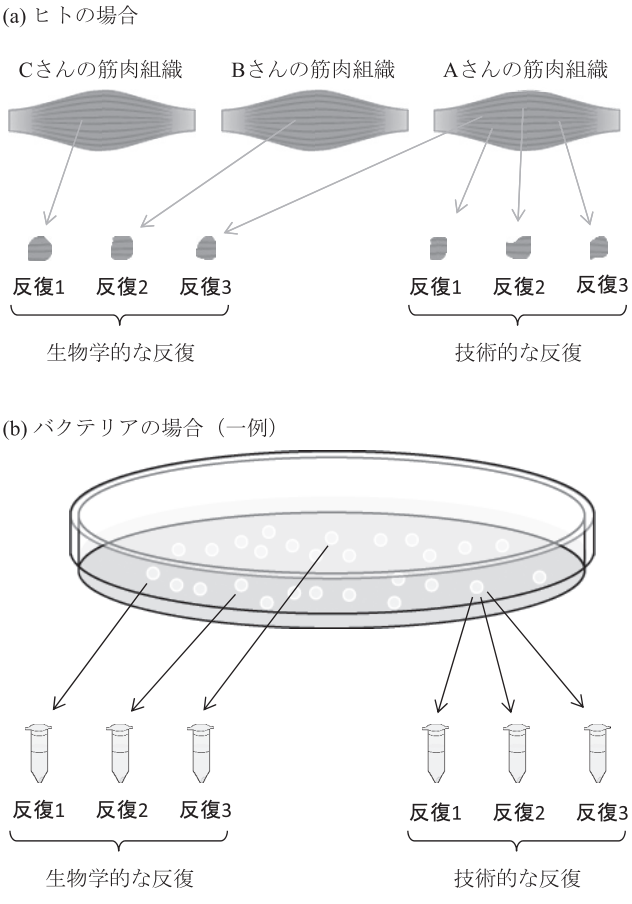

図 1. 生物学的な反復と技術的な反復
(a) ヒトの場合、(b) バクテリアの場合。TogoPictureGallery のイラスト（CC-BY-4.0）をそのままあるいは一部改変して利用。

具体的には、たとえば A さん、B さん、C さんがきょうだいの場合と赤の他人の場合で考えてみるとよいだろう。生物学的なばらつきは、後者よりも前者のほうがより小さいのは明らかである。また、当該の 3 人が同じ赤の他人同士でも、たとえば同じ県内に住んでいる場合（例：全員沖縄県）と地理的に遠い別の都道府県（例：A さんが北海道、B さんが石川県、C さんが佐賀県）の場合でも、後者よりも前者のほうがより小さなばらつきとなるであろう。これを突き詰めていくと、たとえば性別や人種の異なる別個体で反復をとれば、ヒトという生物種により普遍的な現象を捉えうることを意味する。

バクテリアの場合は、たとえば培地上の異なるコロニーをそれぞれ異なる細胞ペレット（cell pellets）として回収したのち、ペレットごとに処理を行ってデータを取得したものが生物学的な反復に相当する。また、培地上の同一コロニーを異なる細胞ペレットとなるように分割して回収し、ペレットごとにデータを取得したものが技術的な反復に相当する（**図 1b**）。今回用いた乳酸菌データが実際にどのような手順で処理がなされたのかを原著論文[3]のみから正確に読み解くのは困難であるが、後述するデータの性質より、生物学的な反復に相当するものだと我々は判断している。

解析データの記号表現（W2）

冒頭で言及した平均－分散プロットは、遺伝子ごとに同一群内の反復データの平均と分散の値を算出し、横軸に平均、縦軸に分散の値をプロットした散布図のことである。RNA-seq カウントデータだけでなく、マイクロアレイデータの性質を説明する手段としても利用されている[5-7]。ただし、基礎情報である平均と分散の算出に先立ち、サンプル（反復または列）間の総カウント数の違いによる影響を排除する必要がある[7]。**表 1** は、サンプルごとの総カウント数を、他の基本情報と合わせて示したものである。第 13 回[8]ではサンプル間のリード数が 10 倍以上異なることに言及したが、総カウント数についても同様の傾向になっていることがわかる。つまり、総カウント数の最大と最小の間には、2,641,162 / 233,195 = 11.33 倍もの違いがある（W2.2）。

入力であるカウントデータ行列のうち、総カウント数が最大のサンプルは 5 列目、最小のサンプルは 1 列目である。学術論文では、このような「総カウント数が最大または最小のサンプル」を関数や記号でうまく表現せねばならない局面がある。この目的のためによく利用されるの

表 1. 解析データ（乳酸菌 RNA-seq カウントデータ）の基本情報
一番右の列がサンプルごとの総カウント数（W2.1 の実行結果と同じ）、それ以外の情報は第 13 回の表 1 と同じである。

サンプル名	SRR ID	リード数（片側のみ）	総カウント数
pH4.5_1h_1	SRR6322562	301,126	233,195
pH4.5_1h_2	SRR6322563	1,470,602	840,602
pH4.5_1h_3	SRR6322564	1,760,461	1,072,851
pH4.5_24h_1	SRR6322565	1,375,368	938,788
pH4.5_24h_2	SRR6322566	3,869,088	2,641,162
pH4.5_24h_3	SRR6322567	1,795,874	1,176,816
pH7_CCG_1	SRR6322568	3,095,834	2,381,981
pH7_CCG_2	SRR6322569	2,570,876	1,758,968
pH7_CCG_3	SRR6322570	846,623	569,947

は argmax や argmin であるが、現実にはこの**記号を見た時点で思考停止してしまうヒトが多い**のではないだろうか。我々は、この原因はほとんど全て数式や記号のみで話が展開されていくためであると認識している。このような経験のある多くのヒトは、「たぶんこういうことだろうが…」という理解を確固たるものにするための分岐点（正しい理解のルートと曖昧な理解のルート）が読み進めていくうちにどんどん増えていき、最終的に論点整理しきれなくなり撃沈という経過を辿ってきたであろう。しかし今一度、本稿を足掛かりにして数式の克服（記号を用いた説明に慣れること）に再チャレンジしてみてほしい。

たとえば、まず行列を A、各行を個々の遺伝子（$i=1, 2, \cdots, G$）、各列を個々のサンプル（$j=1, 2, \cdots, S$）、そして i 行 j 列目の要素 $a_{i,j}$ をカウント数と定義する。行列とその要素（元ともよばれる）などをまとめて、$A = [a_{i,j}]_{G\times S,(1\leq i\leq G,1\leq j\leq S)}$ のように表してもよい。ここで、G は遺伝子数、そして S はサンプル数である。今我々が取り扱っているカウントデータの場合は、2949 遺伝子×9 サンプルの数値行列なので、$G=2949$、$S=9$ である。各要素の具体的な数値としては、たとえば $a_{3,1}=6579$、$a_{2947,2}=212$、$a_{2949,9}=6157$ のように解釈する（**図 2**）。

遺伝子名やサンプル名についても、$g_i=(g_1, g_2, \cdots, g_G)$ や $s_j=(s_1, s_2, \cdots, s_S)$ のように記号で表現することができる。この場合は、$g_1=$EBG00001128470、$g_{2949}=$LGG_02944、$s_1=$pH4.5_1h_1、そして $s_9=$pH7_CCG_3 だと解釈する。なお、i や j は添え字またはインデックスとよばれるものである。前述の「5 列目にあるカウント数が最大のサンプル」は s_5 であるが、この $j=5$ というインデックス情報で当該サンプルを指し示すことができる。

前述の argmax や argmin は、総カウント数情報が格納されたサンプル数分の要素からなる数値ベクトル（233195, 840602, …, 569947）を入力として実行した場合、総カウント数の最大値や最小値のインデックスを返す関数だという理解でもよい。以下に示すように、入力ベクトル中の要素そのものを返す max や min の実行結果と見比べると、両者の違いがよくわかるだろう。

- max(233195, 840602, …, 569947) = 2641162
- min(233195, 840602, …, 569947) = 233195
- argmax(233195, 840602, …, 569947) = 5
- argmin(233195, 840602, …, 569947) = 1

本稿の表記法と完全一致しているわけではないが、argmin については第 18 ～ 19 回で用いた MBCdeg 法の原著論文[9] で実際に利用しているので解読を試みてもよいだろう。なお、argmax や argmin に対応する R の関数としては、which.max や which.min が用意されている（W2.4）。

総カウント数の計算（W3）

さきほど定義した記号を用いると、j 列目のサンプル s_j の総カウント数 T_j は以下のように表すことができる。

$$T_j = \sum_{i=1}^{G} a_{i,j} \tag{1}$$

具体例として、1 列目のサンプル s_1（つまり pH4.5_1h_1）の総カウント数 $T_1=233195$ を得る手順を以下に示す。5 列目や 9 列目など、他の計算例を見比べるとより理解が深まるであろう（W3.1）。

	1	2	…	j	…	8	9
	pH4.5_1h_1	pH4.5_1h_2	…	s_j	…	pH7_CCG_2	pH7_CCG_3
1	EBG00001128470	21	262	…	$a_{1,j}$	…	319	124
2	EBG00001128476	8	96	…	$a_{2,j}$	…	85	27
3	EBG00001128500	6579	20948	…	$a_{3,j}$	…	67686	23262
…	…	…	…	…	…	…	…	…
i	g_i	$a_{i,1}$	$a_{i,2}$	…	$a_{i,j}$	…	$a_{i,8}$	$a_{i,9}$
…	…	…	…	…	…	…	…	…
2947	LGG_02942	28	212	…	$a_{2947,j}$	…	167	56
2948	LGG_02943	121	3160	…	$a_{2948,j}$	…	1698	672
2949	LGG_02944	1585	6737	…	$a_{2949,j}$	…	31179	6157
	総カウント数	233195	840602		T_j		1758968	569947

図 2. 解析データの概要と記号表現

$$
\begin{aligned}
T_1 &= \sum_{i=1}^{2949} a_{i,1} \\
&= a_{1,1} + a_{2,1} + \cdots + a_{2948,1} + a_{2949,1} \\
&= 21 + 8 + \cdots + 121 + 1585 \\
&= 233195
\end{aligned}
$$

それぞれのサンプルの総カウント数に対して T_j という記号を割り当てることで、前述の「総カウント数情報が格納されたサンプル数分（つまり $S=9$）の要素からなる数値ベクトル（233195, 840602, …, 569947）」は、（T_1, T_2, …, T_S）または T_j（$j=1, 2, \cdots, S$）のように表すことができる（T_S は T_9 でもよい）。記号を用いた表現手段に慣れてくると、たとえば上述の数値ベクトルで表現した argmax(233195, 840602, …, 569947) のほうが、むしろ違和感があるような気がしてくるのではないだろうか。具体的な使用例としては、$u=$argmax(T_1, T_2, …, T_S) とおくことで、総カウント数が最大となるサンプルを s_u と表現することができる。

CPM の計算（W4）

今は、平均－分散プロットの基礎情報である平均と分散を得るための前処理段階にある。ここでは、サンプル（反復または列）間の総カウント数の違いによる影響を排除する手段の 1 つとして、一般によく用いられる CPM 正規化を実行する。CPM は Counts per million の略であり、正規化後の総カウント数が 100 万に揃うような係数を掛ける正規化作業および正規化後の値を指す言葉である。i 行 j 列目の CPM 正規化後のカウント値 $a_{i,j}^{norm}$ は、対応する正規化前のカウント値 $a_{i,j}$ と j 列目のサンプル s_j の総カウント数 T_j を用いて以下のように表すことができる。

$$
a_{i,j}^{norm} = a_{i,j} \times \frac{1000000}{T_j} \tag{2}
$$

具体例として、2 列目のサンプル s_2（つまり pH4.5_1h_2）の総カウント数（T_2=840602）を用いて、3 行 2 列目（i=3 と j=2）の CPM 正規化後のカウント値 $a_{3,2}^{norm}$ を得る手順を以下に示す。他のいくつかの計算例を見比べると、より理解が深まるであろう（W4.1）。

$$
\begin{aligned}
a_{3,2}^{norm} &= a_{3,2} \times \frac{1000000}{T_2} \\
&= 20948 \times \frac{1000000}{840602} \\
&= 24920.236
\end{aligned}
$$

式（2）の右辺の $a_{i,j}$ に掛かっている $1000000/T_j$ は、正規化係数とよばれるものに相当する。添え字（インデックス）が j（$=1, 2, \cdots, S$）であることからも明らかではあるが、この値はサンプル数分（つまり $S=9$）の要素からなる数値ベクトルである（W4.2）。正規化係数の値が 1 より大きい（小さい）と、正規化後のカウント値は正規化前に比べて大きく（小さく）なる。正規化後のカウント値 $a_{i,j}^{norm}$ を要素とする行列は、$A^{norm} = [a_{i,j}^{norm}]_{G\times S,(1\leq i\leq G, 1\leq j\leq S)}$ と表すことができる。但し、CPM の具体的な計算例を含めて第 15 回[10] でも述べているが、**正規化後のデータは群間比較時の入力として用いてはいけない**ので注意してほしい。ここでは**あくまでも同一群内の平均と分散を算出するための基礎情報として得ただけ**である。

平均と分散を算出（W5）

平均と分散は群ごとに算出するため、ここであらためてどのサンプル（どの列）がどの群に属するかを示すクラスラベル $L=(l_1, l_2, \cdots, l_S)$ を作成する。一般には、たとえば群数を K、群 k（$=1, \cdots, K$）の反復数を n_k として、様々な実験デザインを表現可能としておく。今入力として用いる行列 A^{norm} は、最初の 1 ～ 3 列目が酸ストレス短期暴露群（pH4.5_1h）、4 ～ 6 列目が酸ストレス長期暴露群（pH4.5_24h）、そして 7 ～ 9 列目が対照群（pH7_CCG）である。このため、$K=3$、$k=1$ が pH4.5_1h 群でその反復数が $n_1=3$、$k=2$ が pH4.5_24h 群でその反復数が $n_2=3$、そして $k=3$ が pH7_CCG 群でその反復数が $n_3=3$ だということを意味する。この場合、クラスラベル L の実体は、計 9 つの要素からなる整数ベクトル（1, 1, 1, 2, 2, 2, 3, 3, 3）のように解釈する（W5.1）。つまり、クラスラベル L の j 番目（$j=1, 2, \cdots, S$）の要素はそれぞれ、$l_1=l_2=l_3=1$、$l_4=l_5=l_6=2$、そして $l_7=l_8=l_9=3$ である。

クラスラベル L を用いることで、任意の群 k に属する L 中の要素（元）を $\{l_j \in L \,|\, l_j = k\}$ のように表現することができる。たとえば、群 2（つまり $k=2$）のクラスラベル情報は $\{l_j \in L \,|\, l_j = 2\}$ と表現することができる。縦棒（|）は条件を表す記号であり、縦棒より右側に書かれているのが具体的な条件である。L 自体は 9 つの要素からなるベクトルであるが、$l_j = 2$ という条件を満たす要素はそのうちの 3 つ、つまり $\{l_4, l_5, l_6\}$ である。縦棒の左側にある $l_j \in L$ は、「l_j は L に属する」という意味である。難解だと感じた読者は、たとえば $L=(l_1, l_2, \cdots, l_S)$ は $L = \{l_j \,|\, 1 \leq j \leq S, j$ は自然数$\}$ や $\{l_j \in L \,|\, 1 \leq j \leq S, j$ は自然数$\}$ とも表現できることを思い出せばよいだろう。この場合は、縦棒の右側でオリジナルの 9 つの要素全てを満たす条件を記載しているのだと解釈すればよい。

実用上は、$\{l_j \in L \,|\, l_j = k\}$ のような特定の群 k に対応するクラスラベル L 中の要素だけではなく、そのインデックス情報を利用したい局面もある。1 つの表現方法としては、クラスラベル L 自体に添え字をつけて、任意の群 k に属するクラスラベルのサブセットを $L^{Gk} = \{l_j \in L \,|\, l_j = k\}$、そのインデックスを $ind(L^{Gk})$ のように定義することである。たとえば群 2（つまり $k=2$）の場合は、$L^{G2} = \{l_j \in L \,|\, l_j = 2\} = \{l_4, l_5, l_6\}$、そのインデックスは $ind(L^{G2}) = \{4, 5, 6\}$ となる。こうすることで、たとえば行列 A^{norm} の中から群 k に対応する列

のみ取り扱いたい場合に、その列のインデックス情報を $j \in ind(L^{Gk})$ と指定できる。

クラスラベルを用いる一番のメリットは、群 k における i 行目の遺伝子の平均カウント値を $m_{i,k}$、そして $m_{i,k}$ を要素とする行列を $M = [m_{i,k}]_{G\times K,(1\leq i\leq G,1\leq k\leq K)}$ のように記号のみで過不足なく説明できる点である。たとえば、計 3 群 (K=3) のラベル情報をもつ A^{norm} からの場合は、2949 行×3 列の平均カウントの行列 $M = [m_{i,k}]_{2949\times 3,(1\leq i\leq 2949,1\leq k\leq 3)}$ が得られるのだと解釈する。$m_{i,k}$ は以下のように定式化できる。

$$m_{i,k} = \frac{1}{n_k} \sum_{j \in ind(L^{Gk})} a_{i,j}^{norm} \tag{3}$$

たとえば行列 M の 6 行 2 列の要素である $m_{6,2}$ には、6 行目の遺伝子の群 2 の平均が格納されている。つまり、g_6= LGG_00001 のカウントのうち、pH4.5_24h 群のみのカウントの平均が格納されている。これを式 (3) に近い形で表現すると、行列 A^{norm} 中の 6 行目の要素 $a_{6,j}^{norm}$ のうち、クラスラベルが 2 ($l_j = 2$) という条件を満たす j (l_4=l_5=l_6= 2 なので j=4, 5, 6) で構成される 3 つの値 (つまり $a_{6,4}^{norm}$ と $a_{6,5}^{norm}$ と $a_{6,6}^{norm}$) の平均として 412.648 が得られている。以下に示すように、式 (3) に具体的な数値を入れて確かめるとよいだろう (W5.2)。

$$\begin{aligned} m_{6,2} &= \tfrac{1}{n_2} \sum_{j \in \{4,5,6\}} a_{6,j}^{norm} \\ &= \tfrac{1}{3}(a_{6,4}^{norm} + a_{6,5}^{norm} + a_{6,6}^{norm}) \\ &= \tfrac{1}{3}(382.408 + 422.163 + 433.373) \\ &= 412.648 \end{aligned}$$

分散についても同様である。群 k における i 行目の遺伝子のカウント値の分散を $v_{i,k}$、そして $v_{i,k}$ を要素とする行列を $V = [v_{i,k}]_{G\times K,(1\leq i\leq G,1\leq k\leq K)}$ と定義できる。たとえば行列 V の 6 行 2 列の要素である $v_{6,2}$ には、6 行目の遺伝子の群 2 の分散が格納されている。分散の具体的な計算自体は、R の var 関数を用いて実行すればよい。一般的な不偏分散の式を本稿の記号に置き換えれば、以下のように $v_{i,k}$ を定式化できる。さきほどの平均の具体的な計算同様、i に 6、j に 4, 5, 6、k に 2 を代入して確かめてみるとよい (W5.3)。

$$v_{i,k} = \frac{1}{n_k - 1} \sum_{j \in ind(L^{Gk})} (a_{i,j}^{norm} - m_{i,k})^2 \tag{4}$$

平均と分散の関係性 (W6)

図 3 は、①平均 M および②分散 V の数値行列の概要である (W6.1)。数値行列を入力として R の summary 関数を実行すると、列ごとの各種統計量 (最小値、中央値、平均値、最大値など) が返される。この場合は、V1 という列が群 1 (pH4.5_1h 群) に、V2 が群 2 (pH4.5_24h 群) に、そして V3 が群 3 (pH7_CCG 群) の結果に相当する。①群ごとの平均カウント数の結果において、Mean の値が 339.1 に揃っているのは、総カウント数を 100 万に揃える CPM 正規化後のデータ (A^{norm}) に基づいているためである ($339.1 \times 2949 \approx 1000000$)。バクテリアのような数千程度の遺伝子数であれば、ここで示されているような数百程度の Mean の値となる。その一方で数万程度の遺伝子数であれば、数十程度の Mean の値となる (たとえば遺伝子数 2 万の場合は 1000000/20000 = 50)。一見何気ない部分ではあるが、意図通りの前処理ができているかを真っ先に確認するチェックポイントである。

次に最大値 (③と④) を眺める。③平均の最大値は 5 桁、

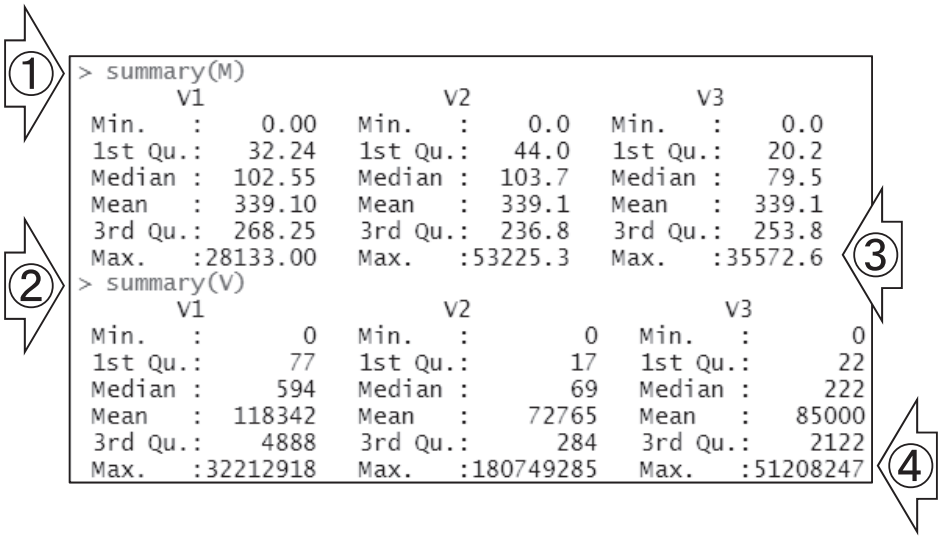

図 3. summary 関数実行結果のスクリーンショット

④分散の最大値はそれよりもさらに3～4桁大きいことがわかる。一般にダイナミックレンジが広いデータは対数軸で全体像を把握することが多く、この場合も**対数変換後のデータで平均－分散プロットを描画する必要がある**とこの段階で判断できる。また、群ごとの平均と分散の関係性は、第一四分位数（1st Qu.）、中央値（Median）、そして第三四分位数（3rd Qu.）を見比べることで大まかに把握可能である。たとえば群1はいずれも分散のほうが平均よりも大きな値になっていることがわかる。群3も群1ほどではないが同様の傾向になっている。しかし群2については、平均と分散が似たような分布になっている。具体的には、1st Qu. と Median で平均のほうが分散よりも大きく（44.0＞17と103.7＞69）、3rd Qu. で逆転している（236.8＜284）。しかし、いずれも他の群の違いに比べれば誤差範囲といえるレベルである。

この summary 関数実行結果は、平均－分散プロットによって得られる散布図のイメージそのものである。たとえば最も平均と分散が似ている群2の散布図では、2949遺伝子分の点が $y=x$ の直線近辺に散らばっている（つまり $v_{i,2}=m_{i,2}$）ことが容易に想像できる。その一方で、**群1の散布図では、ほとんどの点が $y=x$ の直線の左上側に偏って存在する**であろう。これは $v_{i,1}>m_{i,1}$ と表現してもよいが、たとえば $v_{i,1}=m_{i,1}+\phi_{i,1}\times m_{i,1}^2$（$\phi_{i,1}$ は定数）のように等号を用いて表現することができる。これは、一般的な多項式 $f(x)=a_0+a_1x+a_2x^2+a_3x^3+\cdots$ と似たようなものである。2次の多項式 $f(x)=a_0+a_1x+a_2x^2$ において、定数部分を $a_0=0$、$a_1=1$、そして $a_2=\phi$ とおけば、$v_{i,1}=m_{i,1}+\phi_{i,1}\times m_{i,1}^2$ と見栄えが同じ $f(x)=x+\phi x^2$ となる。

さきほど群1に限定して定式化した $v_{i,1}=m_{i,1}+\phi_{i,1}\times m_{i,1}^2$ は、任意の群 k に対して $v_{i,k}=m_{i,k}+\phi_{i,k}\times m_{i,k}^2$ と表すことができる。一部の読者は、見慣れない ϕ（ふぁい）の出現によって、理解が追い付いていないかもしれない。しかし今我々は、任意の群 k における i 行目の遺伝子の平均 $m_{i,k}$ と分散 $v_{i,k}$ の具体的な数値情報が手元にある（W6.2）。したがって、$v_{i,k}=m_{i,k}+\phi_{i,k}\times m_{i,k}^2$ を簡単な式変形によって $\phi_{i,k}=(v_{i,k}-m_{i,k})/m_{i,k}^2$ としておき、具体的な $m_{i,k}$ と $v_{i,k}$ の値を代入すれば、任意の群 k における i 行目の遺伝子の $\phi_{i,k}$ を得ることができる。たとえば、$m_{4,3}=190.113$ と $v_{4,3}=2223.958$ を代入すれば、$\phi_{4,3}=(2223.958-190.113)/(190.113)^2=0.0563$ が得られる。$\phi_{i,k}$ を得る計算式の分子 $(v_{i,k}-m_{i,k})$ からも読み解けるが、$\phi_{i,k}$ は分散 $v_{i,k}$ と平均 $m_{i,k}$ の差が小さいほどゼロに近い値となり、マイナスの値も計算上は得られる。たとえば、$m_{6,3}=318.101$ と $v_{6,3}=82.894$ を代入すれば、$\phi_{6,3}=(82.894-318.101)/(318.101)^2=-0.00232$ が得られる。なお、この $\phi_{i,k}$ は dispersion パラメータとよばれる[11]。全ての $\phi_{i,k}$ を計算することで、平均 M や分散 V と同じサイズの行列 $\Phi=[\phi_{i,k}]_{G\times K,(1\leq i\leq G,1\leq k\leq K)}$ が得られる。

重要な点は、$v_{i,k}=m_{i,k}+\phi_{i,k}\times m_{i,k}^2$ のような式を用いれば、RNA-seq カウントデータの平均と分散の関係性をうまく捉えられるということである。今我々は、計3種類の同一群内の反復データから算出した、2949遺伝子×3群＝8847個の平均の $m_{i,k}$ と dispersion の $\phi_{i,k}$ という2種類の情報をもつ。この式は、$m_{i,k}$ と $\phi_{i,k}$ の情報があれば $v_{i,k}$ がわかるということであり、**RNA-seq カウントデータが負の二項分布（negative binomial distribution; NB 分布）に従う**ことと同義である。これは、統計学の分野では「データを特定の確率分布に従う確率変数とみなしている」ためである。この場合は、「RNA-seq カウントデータを NB 分布に従う確率変数とみなしている」ということになる。さらに具体的には、「RNA-seq カウントデータ（$a_{i,j\,|\,l_j=k}^{norm}$）は、平均（$m_{i,k}$）、dispersion（$\phi_{i,k}$）という2つのパラメータから構成される NB 分布に従う確率変数」だとみなせるということである。統計学の分野では、これを $a_{i,j\in ind(L^{Gk})}^{norm}\sim NB(m_{i,k},\phi_{i,k})$ のように表記する[11]。

$a_{i,j\in ind(L^{Gk})}^{norm}\sim NB(m_{i,k},\phi_{i,k})$ を i=6 および k=2 の具体例で説明すると、k=2（$l_j=2$）という条件を満たす j（$l_4=l_5=l_6=2$ なので j=4, 5, 6）で構成される3つの値（つまり $a_{6,4}^{norm}=382.408$ と $a_{6,5}^{norm}=422.163$ と $a_{6,6}^{norm}=433.373$）は、$m_{6,2}=412.648$ および $\phi_{6,2}=0.00179$ という2つのパラメータから構成される NB 分布から得られたデータ（NB 分布に従う確率変数）だとみなせるということである。確率変数という言葉がそれでも難解だと感じる読者は、これを「確率的に変化する数値」に置き換えると同時に、k=2 の群（つまり酸ストレス長期暴露群（pH4.5_24h））における i=6 行目の遺伝子に対して、4つ目や5つ目の反復データを得た結果を想像してみるとよい。ほとんどのヒトは、既に得られている3つの値（つまり $a_{6,4}^{norm}=382.408$ と $a_{6,5}^{norm}=422.163$ と $a_{6,6}^{norm}=433.373$）のいずれかと全く同じ値が出るとは思わず、これらの既出の値の範囲内か、範囲外だとしても既出の値からそれほどかけ離れない 380 や 438 のような値が得られると予想するであろう。このように、**手元にない仮想的なデータを思い浮かべれば、確率変数とみなすほうがむしろ適切だと思える**のではないだろうか。

R では NB 分布に従う乱数を発生させる rnbinom 関数が提供されているので、読者自身が実際に乱数を発生させ、その分散が期待値と一致するか確認できる（W6.3）。たとえば、$m_{6,2}=412.648$ および $\phi_{6,2}=0.00179$ を用いて8個乱数を発生させると、（400, 392, 446, 430, 377, 423, 406, 412）という結果が得られる。この分散は 484.786 であり、期待値（$v_{6,2}=717.253$）とかけ離れた結果だと思われるかもしれない。しかし、たとえば 172000 個の乱数から計算した分散は 716.481 となるように、発生させる乱数（試行回数）を増やせば期待値に収束していく。納得できるまで読者自身で検証してみるとよいだろう。

平均－分散プロット（W7 ～ W10）

図 4 は、横軸に平均 M、縦軸に分散 V の値をプロットした散布図である。群ごとにプロットした（a）～（c）については遺伝子数（＝2949）だけの点が、そして全データをまとめた（d）については 2949×3＝8847 個の点が存在する。左下から右上に伸びる $y=x$ の直線は、平均と分散が完全一致する点に相当する。図の左上（右下）にある数値は、$y=x$ の直線よりも左上（右下）側にある点の数、つまり $M<V$（$M>V$）の遺伝子数である。この $M<V$（$M>V$）は、$\phi_{i,k}=(v_{i,k}-m_{i,k})/m_{i,k}^2$ より、dispersion パラメータ（Φ）が正（負）の値になることと同義である。**図 3** に基づく議論と本質的に同じであるが、（a）と（c）では $M<V$ となる遺伝子数が $M>V$ に比べて 3 倍以上多いことがわかる。（b）では逆の傾向になっているものの、（d）全体としてみると、明らかにΦを含む NB 分布で平均と分散の関係を定式化するほうがよいことがわかる。

図 4b の群 2（pH4.5_24h 群）については、遺伝子ごとの dispersion パラメータ（$\phi_{i,2}$）の平均についても、$\frac{1}{G}\sum_{i=1}^{G}\phi_{i,2}=\mathrm{mean}(\phi_{1,2},\phi_{2,2},\ldots,\phi_{G,2})=-0.00263$ とマイナスの値になっていることがわかる（W10.1）。第一四分位数（1st Qu.）や第三四分位数（3rd Qu.）を眺めても、全体的にゼロ付近に位置していると判断できる。したがって、この群に限っていえば、NB 分布というよりはむしろポアソン分布に従うと解釈するほうが自然といえよう。ポアソン分布は、NB 分布の数式（$v_{i,k}=m_{i,k}+\phi_{i,k}\times m_{i,k}^2$）において、dispersion パラメータが全てゼロ（$\phi_{i,k}=0$）の場合に相当する分布だからである（つまり $v_{i,k}=m_{i,k}$）。技術的な反復データはポアソン分布に従うことも知られており[12]、このデータもそれに匹敵するレベルで群内類似度が高いのだと解釈すればよい（W10.2）。

統計モデルの意味

一部の読者は、「我々が興味あるのは G1 群 vs. G2 群のような群間比較であり、群内のばらつきではない」と思われているかもしれない。そのような読者は、統計的検定の意味や手続きを思い返してほしい。この場合の帰無仮説（null hypothesis）は「比較する群間に差がない」であるが、どの程度であれば差がないとするかを具体的に定めないと

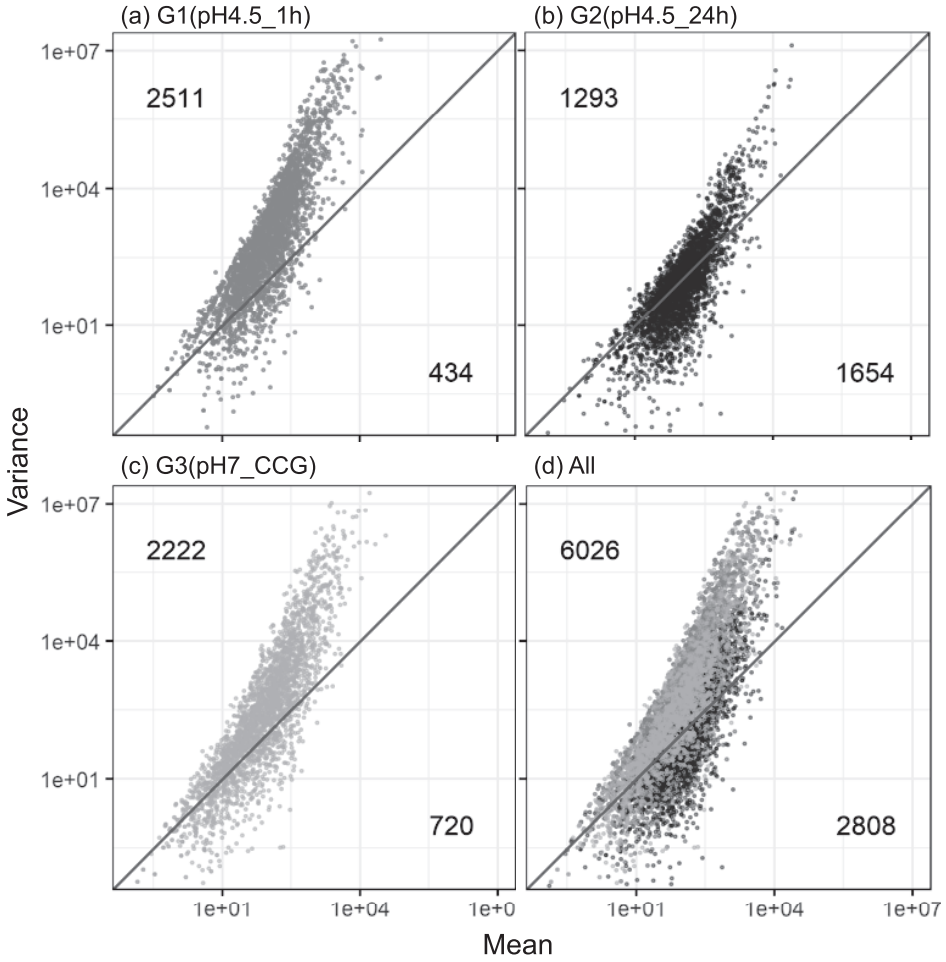

図 4. 平均－分散プロット

いけない。それが帰無仮説に従う分布、つまり帰無分布（null distribution）である。**図 4 で我々が見ているのは帰無分布そのもの**である。

この場合の統計的検定は、たとえば比較する群を同一群由来だとみなした場合に、検定したい遺伝子の平均と分散の値が帰無分布からどれだけ離れているかを *p* 値として返す作業だと理解すればよい。もちろん *p* 値を得るためには、「帰無分布をうまく数式で表現したもの」をあらかじめ構築しておく必要がある。これがいわゆる**モデル構築**（model construction）という作業に相当し、構築されたものは**統計モデル**（statistical model）とよばれる。統計モデルの実体はただの数式であり、この場合は「任意の平均と分散の値を入力として与えると、それが帰無分布上のどのあたりに位置するかを *p* 値で返してくれる数式」である。

現実には、**図 4** で見えている帰無分布を表現できる数式（つまり統計モデル）は複数存在しうる。また、この帰無分布は全て 3 反復のみの限られたデータから得られたものである点も忘れてはならない。つまり、今手元にはないが 4 つめや 5 つめの反復データが将来得られると、帰無分布の形状は変わりうる。このような事柄も鑑みると、単に手持ちデータへのあてはまりのよさだけを唯一無二の評価基準とすべきではないこともわかるであろう。このあたりが統計学分野でよく行われている、複数の候補の中からどの統計モデルを選択したほうがよいかという、**モデル選択**（model selection）とよばれる議論である。

最後のモデル云々で途端に難解だと感じた読者は、たとえば居住地における過去 10 年の気温のデータから統計モデルを構築し、3 日後・100 日後・1098 日後の気温をどうやって予測するかという問題を考えてみるとよいだろう。多項式と正弦波（sin カーブ）の 2 択の場合、どちらの数式を選択するだろうか？著者らは当然後者を採用する。もちろん正弦波はあくまでも基本形であり、これをベースとした改良版は複数存在しうる。この議論と同様、**RNA-seq カウントデータの統計モデルとしては NB 分布に従うモデル（つまり NB モデル）が基本形だというのがこの分野のコンセンサス**だということである。

おわりに

本稿では、RNA-seq カウントデータの性質を解説するという大枠の中で、バイオインフォマティクス分野の論文を読み解く上で必要なリテラシーの向上（数式や記号表現に対する苦手意識の払しょく）を意識した構成にした。本文中ではほとんど触れなかったが、ggplot2 を用いた描画の基礎から応用については W08 〜 W09 で詳細に解説している。**図 4** は、W9.8.3 で得た画像をベースとして、PowerPoint で微調整しながら作成したものである。作図は凝りだすとキリがないうえ（作図沼に落ちる）、そもそも ggplot2 内で最初から最後まで描画せねばならないものでもない。今回の**図 4** 作成手順のように、手作業の余地を残すくらいの心構えでもよいだろう。

我々は今でも「反復なしデータで統計的手法を実行するとエラーになってしまうが、*p* 値を得るにはどうすればよいか？」という質問をたまに受ける。本稿の内容を正しく理解できた読者は、そもそも反復なしだと帰無分布を得ようがない（つまり統計モデルの構築ができない）ことがわかるであろう。もちろん反復なしデータでもエラーを出さずに結果を返すプログラムは存在するが、それらの結果は決して過信しないよう注意してほしい。

次回は、今回得た乳酸菌 RNA-seq カウントデータの平均（*M*）と dispersion パラメータ（Φ）を用いて、シミュレーションデータの生成を行うトピックを中心に解説する予定である。バイオインフォマティクス系の論文でよく行われているシミュレーションデータ解析の流れを知ることで、表層的なノウハウだけでなく、バイオインフォマティシャンの思考回路にまで踏み込んでもらいたい。

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。図 1 のイラストは、TogoPictureGallery（CC-BY-4.0）の図を（そのままあるいは一部改変して）利用させていただきました。

利益相反（COI）

牧野磨音、坂本光央、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) 牧野磨音，清水謙多郎，門田 幸二 (2022) 次世代シーケンサーデータの解析手法：第 19 回 R Markdown，日本乳酸菌学会誌 **33**: 195–205.
2) 牧野磨音，清水謙多郎，門田 幸二 (2022) 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，日本乳酸菌学会誌 **33**: 87–94.
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**: 1604–13.
4) Zakharkin SO, Kim K, Mehta T, Chen L, Barnes S, Scheirer KE, et al. (2005) Sources of variation in Affymetrix microarray experiments. BMC Bioinformatics **6**: 214.
5) Archer KJ, Dumur CI, Ramakrishnan V (2004) Graphical technique for identifying a monotonic variance stabilizing transformation for absolute gene intensity signals. BMC Bioinformatics **5**: 60.
6) Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol **11**: R106.
7) 門田幸二 著 (金明哲 編) (2014) シリーズ Useful R ⑦ トランスクリプトーム解析，共立出版
8) 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.
9) 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.
10) 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.
11) Robinson MD, Smyth GK (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics **9**: 321–332.
12) Marioni JC, Mason CE, Mane SM, Stehpens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. **18**: 1509–1517.

Methods for analyzing next-generation sequencing data 20. Properties and statistical models of RNA-seq count data.

Manon Makino[1], Mitsuo Sakamoto[2], Kentaro Shimizu[1, 3, 4], Koji Kadota[1, 3, 4]

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

Abstract

RNA-seq is a tool for measuring gene expression and is commonly used for identifying differentially expressed genes (DEGs) under different conditions or groups. Most of the programs for DEGs has been provided as a free software environment called R. Users typically start the analysis with a numerical matrix called “count data”, where each row a gene, each column a sample (a group’s replicate), and each cell the number of counts. We describe the characteristics of the count data with a scatter plot (so-called “mean-variance plot”) displaying the relationship between the mean (x-axis) and variance (y-axis) within replicates. We explain why a statistical model called the negative binomial distribution (NB model) has been used for identifying DEGs.