Vol.2014-BIO-40 No.4
2014/12/18

発現変動遺伝子抽出の統計〜レビュー

瀬々 潤[1,a)]

概要：遺伝子網羅的に発現量を観測するマイクロアレイが超並列シーケンサを利用する RNA-seq になり，遺伝子発現の解析に新たな統計手法が必要となっている．要因の一例として，マイクロアレイでは相対的な発現量として採取されいた遺伝子発現量に関し，計数による計量が行える定量性の高さが挙げられる．また，反復実験が一般的となり，発現量の変化を統計的指標に基づいて検出する傾向が高まっている．このような実験の変化に対して，新たに考慮すべき統計解析が存在し，RNA-seq の発現量解析には，Cuffdiff, edgeR, DESeq など独特のソフトウエアが利用されている．本発表では，これらで利用されている統計解析についてレビューを行う．

Sese Jun[1,a)]

Abstract: Frequently used biological experiment technique for observing comprehensive gene expression has been changed from microarray using cDNA hybridization to RNA-seq using high-throughput sequencers so called NGS, which allow us to use statistical model to analyze the changes of gene expression levels of each gene. For example, while microarrays use the brightness of the spots, RNA-seqs measure the number of fragments on each gene, giving us more quantitative values. It is also important that biological replicates are generally required, but the number of really performed experiments is limited because of reducing the experimental cost. To handle these data, several statistical methods to find genes whose expression levels are statistically changed between two different conditions have been introduced, such as Cuffdiff, edgeR and DESeq. We here introduce the statistical methods.

1. はじめに — RNA-seq

遺伝子網羅的な発現解析が，マイクロアレイなどの DNA が相補対を作る現象を利用した手法から，配列自身を読む超並列シーケンサ（次世代シーケンサ，高速シーケンサ，High-throughput sequencer，Next-generation sequencer などとも呼ばれる．以下では，NGS と呼ぶ）を利用するようになったことで，ランダムサンプリングに基づくモデルを利用した統計的な解析が導入可能になった．その結果，環境間や刺激間で発現の変動した発現変動遺伝子について，統計的有意性に基いて発見することが可能となっている [1], [2]．この章ではまず，RNA-seq による遺伝子発現量の実験的な定量法に関して述べることで，発現変動遺伝子同定のためのモデル化に繋げる．

本稿では主としてモデル生物を始めとしたゲノム配列が既に読まれている種に対する解析を念頭において話を進めるが，ゲノムが決まっていない非モデル生物種からも RNA-seq が行われている．これらの解析に関しても，本稿で説明する事項とほぼ同様の解析が可能であるが，変化する点，及び，注意すべき点に関して，本稿の最後にまとめて述べる．

一般的な RNA-seq では，細胞群から抽出した RNA を精製し配列を読む（図 1）．この際，NGS では読める長さに制限がある．RNA が 1,000 塩基を超えるものがあるのに対し，1 度に読める塩基数は 100 塩基あるいは 150 塩基などである [*1]．このため，RNA は超音波などで断片化した上で，各断片が読まれる．RNA の抽出から配列を読むシークエンスの一連の過程において，各 RNA がとそれに対応する遺伝子との対応付けは行われない．そのため，読んだ配列を基に，その配列がどの遺伝子（ゲノム領域）由来のものであるかを判断する必要がある．

読まれた断片配列は，その配列と同一の配列をゲノム，あるいは既知の遺伝子の中から検索することで，どの遺伝

[1] 産総研 CBRC
CBRC, AIST, Koto, Tokyo 135–0064, Japan
[a)] sese.jun@aist.go.jp

[*1] イルミナ社の HiSeq あるいは MiSeq などのシーケンサを想定．例えば Pacific Biosciences 社のシーケンサでは，より長い配列を読むことが可能であるが，本稿で述べる発現変動遺伝子の抽出に関しては同一の手順が適用可能である．

© 2014 Information Processing Society of Japan

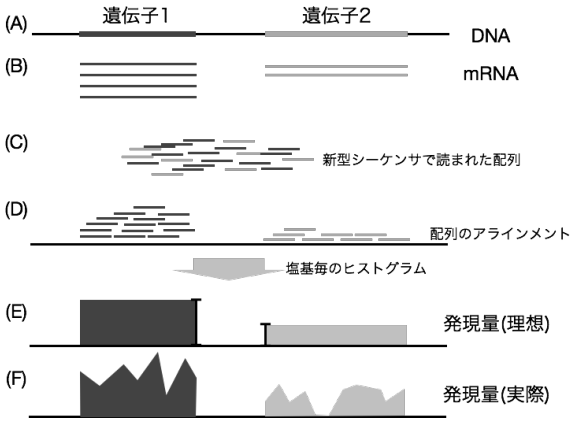

図 1 RNA-seq 解析の流れ

子に由来していたかを調べる（図 1(D)）．断片化される配列は，細胞内の RNA 配列全体と考えられ，また，読まれる配列も断片化した配列全体と考えられるので，十分大きな断片数を読むことで，各遺伝子から得られる断片数比が遺伝子間の発現量比に収束していくと考えられる．

ここで注意しなければいけないことが 2 つある．1 つ目は，細胞内に存在する mRNA の数は，5000 遺伝子の観測結果で，20 万配列程度と推定されているが [3]，その個数は確定していない上，時々刻々と細胞の状態ごとに異なると考えられるので，絶対的な量は分からない．推定可能な事は，存在している個数ではなく，各遺伝子間の比率である．これは，一般的な復元・非復元抽出において推定する青玉・赤球の比率と同様である．もう 1 つは，実際の遺伝子発現の比率は，各遺伝子からの断片数の比率ではなく，観測された断片数を遺伝子の長さで割ったものである．遺伝子の長さは，遺伝子毎に異なっているため，同じ個数が細胞内に存在しても，長さが長い遺伝子からは多くの断片が，短い遺伝子からはほとんど断片が観測されない．発現量を考える場合は，これらの正規化が必要となる．

しかし，断片の数が正しく推定することができるとすれば，遺伝子領域の長さは既知のことが多いため，RNA の総数を仮定することで，実際の発現量を推定することは可能となる．このため，各遺伝子から抽出される断片の比率を推定することが重要である．

2. 2 サンプル間の発現変動遺伝子

NGS における発現変動遺伝子に関して定式化をしよう．

問題 1 2 つのサンプル A, B の対照実験を考える．それぞれのサンプルから，NGS で $n(A), n(B)$ 本の有効な [*2] 断片が得られたとする．また，そのうち着目する遺伝子 g からは，$n(A,g), n(B,g)$ 本の断片が対応したとする．この時，$n(A,g)$ と $n(B,g)$ の間に，統計的に有意な差異がある場合 g はサンプル A,B 間の発現変動遺伝子と呼ぶ．■

[*2] NGS の配列には読み取りエラー等により遺伝子に対応付かない断片が存在するため，ここではゲノム配列に対応づいた断片に絞っている

表 1 発現変動遺伝子を考える際の分割表

サンプル	遺伝子 g 由来	それ以外	合計
A	$n(A,g)$	$n(A,\bar{g}) = n(A) - n(A,g)$	$n(A)$
B	$n(B,g)$	$n(A,\bar{g}) = n(B) - n(B,g)$	$n(B)$

サンプル A, B は，片方を基準（コントロール）として，もう一方での変化を観測する実験であることが多い．薬剤の投与前と投与後，刺激を与える前後，健常細胞とがん細胞などの 2 サンプルを想定している．

2.1 二項検定とポアソン分布を用いた検定

実験手法より，$n(A)$ は十分大きな値であるので，二項分布を用いた以下の定式化が考えられる．

定式化 1 遺伝子 g 由来の断片が抽出される確率が $n(A,g)/n(A)$ のサンプルから，$n(B)$ 回の独立な試行を行ったとき，$n(B,g)$ 回 g 由来の断片が抽出されたとする．$y = n(B,g)$ とすると，Y を確率変数とし，y 回断片が抽出される確率 $P(Y=y)$ は，

$$P(Y=y) = \binom{n(B)}{y} p^y (1-p)^{n(B)-y}$$

で表される．ただし，$p = n(A,g)/n(A)$ である．この確率が有意水準 α 以下（両側検定の場合，$1-\alpha$ 以上も含む）の場合，発現が有意に異なると考えられる．■

また，二項分布はポアソン分布を用いることで，よい近似が可能である．

定式化 2 $y = n(B,g)$ 回 g 由来の断片が抽出される確率は，ポアソン分布を用いると

$$P(Y=y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

と表せる．ここで $\lambda = p \cdot n(B)$ である．この分布は期待値，分散共に λ である．■

NGS を用いた配列の抽出では，各サンプルから 1 千万を超える断片を取得する事が多い．また，全発現に占める遺伝子 g の存在確率は，多くの遺伝子では非常に小さい値となるため，p は一般に非常に小さな値となる．

2.2 カイ二乗検定

異なる定式化として，分割表を考え，フィッシャーの正確確率検定やカイ二乗検定を利用することで，2 つのサンプル間で独立性の検定を行う方法が考えられる．NGS ではサンプル数が大きく，フィッシャーの正確確率検定では計算が大規模になるので，ここではカイ二乗検定の利用を考える．

定式化 3 サンプル A と B から，表 1 で表される数のサンプルを得た．この時のカイ二乗値は，

$$\chi^2 = \sum_{i \in \{A,B\}, j \in \{g,\bar{g}\}} (n(i,j) - E(i,j))^2 / E(i,j)$$

© 2014 Information Processing Society of Japan

ここで，$E(i,j) = (n(A,j) + n(B,j)) \cdot n(i)/(n(A) + n(B))$ である．
カイ二乗検定では，この値が近似的に自由度 1 のカイ二乗値に従う事を利用する．数表より，この値を読み取り，有意水準以下であれば，サンプルが独立である仮説が棄却され，発現が異なると考えられる．■

ポアソン分布，カイ二乗検定共に非常に頻繁に用いられる分布や検定方法であるが，生物サンプルを扱うために必要な複製実験が直接的には扱えないことが問題となる可能性がある．特にカイ二乗検定では，サンプル数増大が検出力の増加につながり，発現が多い遺伝子は多少の発現変動でも軒並み有意になる可能性があり注意を要する．

3. 2 群間の発現変動遺伝子

前節で刺激の前後の様な 2 サンプル間の比較を扱ったが，遺伝子発現量は，常に一定の値をとっているわけではなく，細胞周期，実験時の環境など，様々な状態を反映しているため，実験を行う毎に完全に一致した値になることは無いと考えられる．そこで，RNA-seq を利用した多くの 2 群間比較解析においては，各群から複数回のサンプル（生物学的複製サンプル）を取り，2 群間の比較が行われる．各遺伝子に着目した場合，対応の無い 2 群間比較の問題と考えることができるが，各群で行われる実験の回数は，実験費用の問題，サンプルの用意の難しさから，各群の実験が 3 回から 5 回程度と非常に少ないことも多い．この少ない実験回数が，検定の際に問題となる．

3.1 t 検定の利用

2 群間比較で頻繁に利用される検定として t 検定（スチューデントの t 検定）が挙げられる．t 検定では，2 群間の平均が同一の母集団由来かを検定する．

問題 2 群 A から a 回の RNA-seq を行ったとする．それぞれの実験を $A_1, A_2, ..., A_a$ とする．同様に群 B から b 回の RNA-seq を行い，それぞれ $B_1, B_2, ..., B_b$ とする．これら 2 群の間の平均に差がないことを，t 検定を用いて検定する．
$n(i,g)$ を実験 i における遺伝子 g 由来の断片の数とする．統計量 T を以下の式で計算する

$$T = \frac{\bar{A} - \bar{B}}{\sqrt{\left(\frac{1}{a} - \frac{1}{b}\right) U_{AB}}}.$$

ここで，

$$\bar{A} = \frac{1}{a}\sum_{i=1}^{a} n(A_i, g),$$
$$\bar{B} = \frac{1}{b}\sum_{i=1}^{b} n(B_i, g),$$
$$U_{AB} = \frac{\sum_{i=1}^{a}(n(A_i,g) - \bar{A})^2 + \sum_{i=1}^{b}(n(B_i,g) - \bar{B})^2}{m + n - 2}$$

である．■

ただし，スチューデントの t 検定は，各群から得られた分布が正規分布に従う．また，各群の分散が等分散であることが仮定されており，本来はこれらを判断した上で，検定する必要がある．しかし，先に述べた通りサンプルの数は 3 つあるいは 5 つなどと，分布や分散を知るために十分な数は無く，現実的には困難である．

3.2 負の二項分布を用いたモデル化

現在，RNA-seq 解析で多く用いられている手法が負の二項分布を用いた検定である．ポアソン分布を基本とし，パラメータを一つ増やすことで，多サンプルを得た場合に明らかになる過分散を考慮にいれることが可能である．

RNA-seq によるランダムサンプリングの分布が定式化 2 の通りポアソン分布に従うと仮定すると，分散は λ となることが知られている．一方，実際にデータを調べると，λ が大きい所では，分散が λ より大きな値を取っている事が知られている ([4] の Figure 1，あるいは [2] の Supplemnetal Text Figure 2.)．このため，ポアソン分布を用いて検定を行うと，特に発現量が大きい遺伝子に対して，本来の値以上に低い P 値を算出する可能性がある．

過分散が起きた場合に，適用されるモデルが負の二項分布である．負の二項分布を用いた検定は，以下のように定式化される

定式化 4 確率変数を Y として，パラメータ p と r を用いると，負の二項分布は

$$P(Y = y) = \binom{y + r - 1}{r} p^y (1 - p)^r$$

と表せる．また，ガンマ関数 $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$ が，x が自然数の時，$\Gamma(x) = (x-1)!$ である事を用いると，

$$P(Y = y) = \frac{\Gamma(y + r)}{\Gamma(r)\Gamma(y + 1)} p^y (1 - p)^r$$

となる．期待値は $pr/(1-p)$，分散は $pr/(1-p)^2$ である．
■

系 1 変数 r を無限に飛ばすと，負の二項分布はポアソン分布に近似できる．

証明 1 期待値を表す新たな変数として $\lambda = \frac{pr}{1-p}$ を導入すると，$p = \frac{\lambda}{r+\lambda}$ である．これを，負の二項分布の式に代入して，変形する．

$$\begin{aligned} f(y; k, r) &= P(Y = y) \\ &= \frac{\Gamma(y + r)}{\Gamma(r)\Gamma(y + 1)} p^y (1 - p)^r \\ &= \frac{\lambda^y}{y!} \cdot \frac{\Gamma(y + r)}{\Gamma(r)(r + \lambda)^r} \cdot \frac{1}{\left(1 + \frac{\lambda}{r}\right)^r} \end{aligned}$$

ここで r を無限に飛ばすと，第 1 項は r に依存せず，第 2 項は 1，第 3 項は指数関数に収束するので，

$$\lim_{r \to \infty} f(y; k, r) = \frac{\lambda^y}{y!} \frac{1}{e^\lambda}$$

© 2014 Information Processing Society of Japan

これは，期待値 λ のポアソン分布である．■
よって，ポアソン分布にパラメータを 1 つ加え拡張した分布として利用可能である．

遺伝子発現量のモデル化に話を戻すと，m 個のサンプル全てから同じ数の断片を観測したと仮定する．この時の遺伝子 g の平均 $\mu(g)$, 分散 $\sigma(g)^2$ が

$$\mu(g) = pr/(1-p)$$
$$\sigma(g)^2 = pr/(1-p)^2$$

となるように，p と r を決めれば，各遺伝子の負の二項分布をモデル化することが可能である．

一般に，サンプル間で得られる断片の数は異なるので，遺伝子 g がサンプル i 中で含まれている確率 $s(i,g) = n(i,g)/n(i)$ を設定し，この値の平均や分散（あるいは，この値に断片数に関連した定数を掛けた値）を利用する．

3.3 分散のモデル化

少数のサンプルから求めた分散は外れ値に弱く，正しい値から離れることも多い．そのため，平均値 $\mu(g)$ に対し，分散 $\sigma(g)^2$ を平均値の関数 $f(\mu(g))$ として表すことで，外れ値に対処し，より正確な発現量のモデル化が行われている．現在，頻繁に利用されている発現変動遺伝子抽出ソフトウエアの edgeR [5], [6]，DESeq [4]，Cuffdiff [2] の何れにおいても負の二項分布によるモデル化が行われているが，分散の推定が多少異なる．

edgeR の例

edgeR では，平均 μ とパラメータ ϕ を用意し，負の二項分布を以下のように表す．

$$P(Y=y|\mu,\phi) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}\left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}}\left(\frac{\mu}{\phi^{-1}+\mu}\right)^{y}$$

この平均は μ，分散は $\mu+\phi\mu^2$ と表せる．本来平均と分散は遺伝子ごと決まる値であるが，ϕ を全ての遺伝子で共通にすることで，分散が不当な値になることを防いでいる．

DESeq の例

DESeq では，サンプル i, 遺伝子 g の分散 $\sigma(i,g)^2$ を各遺伝子，全サンプルに対して推定する．その際，分散を

$$\sigma(i,g)^2 = \mu(i,g) + t(i)^2\nu(j)$$

と分解する．ここで，$\mu(i,g)$ は，サンプル i における遺伝子 g の推定発現量（実際の発現量ではなく，全サンプルの平均発現比率にサンプル i から得られた総断片数 $t(i)$ を掛けたもの），$t(i)^2$ は総断片数，$\nu(j)$ は遺伝子毎に決まる値である．

この時，$\nu(j)$ は，発現量 $\mu(i,g)$ に対して滑らかな関数であることを仮定し，一般線形化モデル (Genelarized linear model; 以下 GLM) による近似をおこなっている．回帰は R の limma パッケージにより行われている．また，この $\nu(j)$ により，類似の発現量を持つ遺伝子同士は類似の分散を持つような制約が働く．（edgeR でも，第 2 項に発現量が現れているため，この効果は働いている）

Cuffdiff(Cuffdiff2)

Cuffdiff も DESeq 同様の回帰を用いて分散をモデル化し，GLM による回帰を行っている．利用しているパッケージは LOCFIT であり，局所線形回帰による回帰を行っている．

3.4 負の二項分布を用いた検定

負の二項分布を用いたモデル化を紹介したが，確率密度関数 $P(Y=y)$ の推定であり，このモデルを用いて，2 群間のサンプルが，同一の平均を有するか否かの検定を行う必要がある．

しかしながら，特に DESeq と Cuffdiff においてはモデルが複雑であり検定が容易ではない．また，サンプル間に依存関係があるため，互いに独立の仮定がおけず解析的な計算も困難である．そこで，ランダムサンプリング法を用いた P 値の近似が行われている．

基本的な考え方は以下のとおりである．

(1) 実データにおける P 値を，推定した分布に従って計算する．
(2) 推定した負の二項分布に従った乱数を発生させる．
(3) その値の P 値を計算する．
(4) 2,3 を繰り返し，1 で計算した値より小さい値が出る割合を計算する
(5) その割合が P 値となる

4. その他，気をつける必要のあること

4.1 多重検定補正

全遺伝子に対して検定を行った場合，多重検定による偽陽性確率の増大が問題となるため，これを補正する多重検定補正法が適用される．例えば，単一の遺伝子に対して有意水準 α で検定を行った場合，ランダムなデータで P 値が α 以下になる確率は α である．一方，10 の遺伝子に有意水準 α で検定を行うと，全ての遺伝子が有意でない確率は $(1-\alpha)^{10}$ なので，1 個以上の遺伝子が α 以下になる確率は $1-(1-\alpha)^{10}$ であり，$\alpha = 0.05$ の場合で 40%にのぼる．網羅的遺伝子解析では，1 万を超える遺伝子が検出されるので，ほぼ確実に偽陽性が生まれる．つまり，ランダムなデータからでも発現変動遺伝子が数個は見つかってしまうことになる．

多重検定補正法として，1 つ以上の偽陽性が生まれる確率 (Family-wise error rate; FWER) を α 以下に抑える方

© 2014 Information Processing Society of Japan

法と，発見された有意差のある遺伝子のうち，偽陽性が含まれる確率 (q-value と呼ばれる事もある) を α 以下に抑える方法 (False discovery rate; FDR) の 2 通りが存在し，遺伝子発現量解析では後者が利用される事が多い．FDR の細かい制御方法は，本稿の範囲を超えるので別稿に譲るが，FWER, FDR いずれの方法においても，統計的検定の値は変わらず，有意水準のみを補正することで，それぞれの値を制御する．つまり，全ての遺伝子に対して P 値を計算し，その後，FWER もしくは FDR を α 以下になるように制御した補正後の有意水準 δ を計算，そして，P 値が δ 以下の値を有意な差のある遺伝子群とみなす．

4.2 正規化

マイクロアレイの際には，サンプル間で輝度のバラつきがありそれを補正する必要が生じていた．RNA-seq の場合，理想的な条件下では補正の必要は無いが，実際観測されたデータからは，ライブラリ間で遺伝子の発現量が全体的に大きく，あるいは，小さくなり，非常に多くの遺伝子の発現量に有意な発現差が生じるケースが少なくない．このため，RNA-seq でも，正規化の必要性が言われている [7]．

生物学的に考えて，二つの類似のサンプル（刺激の前後など）から得られた遺伝子の発現量は，大勢の遺伝子で変化が無く，一部の遺伝子において発現量の変化があると考えられる．この仮定のもと，正規化が行われる．

また，目視で確認する方法として，一般に横軸に平均発現量，縦軸にライブラリ間の発現差を取った MA プロットを利用する方法がある．

4.3 1 細胞 RNA-seq における統計

近年，1 細胞レベルでの遺伝子発現量が RNA-seq によって観測可能になってきた．通常の RNA-seq では，少なくとも数百細胞が混在した状況下で遺伝子発現量の観測を行っており，この場合，数百細胞の平均を観測していることになる．それに対し，1 細胞 RNA-seq では，単一細胞内の発現量を計測することになり，特に細胞間の違いを見たい場合に行われる実験となっている．現時点で，実験プロトコルが発展途上であることからくる定量性の問題は解決されつつあり，本稿で紹介した手法を利用することが可能である．また，1 細胞に特化した検定法の開発も行われている [8]．しかし，本稿で紹介した手法も含めて，細胞間の異種性を扱うことが困難であり，必ずしも 1 細胞 RNA-seq のモデルに則しているとは言いがたい．よって，今後数理的な手法の発展が望まれるところである．

4.4 非モデル生物における発現量解析

RNA-seq 等の遺伝子配列を読んだ配列をアセンブリ（配列をつなぎ合わせる事）によりゲノム配列が分からなくても，遺伝子の配列を知ることはできるようになってきた．これにより，ゲノム配列が非常に長い，あるいは，繰り返し配列が多いことでゲノムの解読が難しかった種であっても，遺伝子発現量を取ることは可能になっている．一度遺伝子配列が分かれば，本稿で利用した手法はそのまま応用が可能である．

5. まとめ

本稿では RNA-seq を利用した網羅的遺伝子発現解析における発現変動遺伝子抽出に関して，導入を行った．いずれの手法もモデル化手法が多少異なっており，全く同じ結果が抽出できるわけではない [9]．状況に応じて，適切な手法を使う必要がある．

参考文献

[1] Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W. and Robinson, M. D.: Count based differential expression analysis of RNA sequencing data using R and Bioconductor, *Nature Protocols*, Vol. 8, No. 9, pp. 1765–1786 (2013).

[2] Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L.: Differential analysis of gene regulation at transcript resolution with RNA-seq., *Nature Biotechnology*, Vol. 31, No. 1, pp. 46–53 (2013).

[3] Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M.: Global quantification of mammalian gene expression control., *Nature*, Vol. 473, No. 7347, pp. 337–342 (2011).

[4] Anders, S. and Huber, W.: Differential expression analysis for sequence count data, *Genome Biology*, Vol. 11, No. 10, p. R106 (2010).

[5] Robinson, M. D. and Smyth, G. K.: Small-sample estimation of negative binomial dispersion, with applications to SAGE data, *Biostatistics*, Vol. 9, No. 2, pp. 321–332 (2007).

[6] Robinson, M. D., McCarthy, D. J. and Smyth, G. K.: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, Vol. 26, No. 1, pp. 139–140 (2009).

[7] Sun, J., Nishiyama, T., Shimizu, K. and Kadota, K.: TCC: an R package for comparing tag count data with robust normalization strategies, *BMC Bioinformatics*, Vol. 14, No. 1, p. 219 (2013).

[8] Kharchenko, P. V., Silberstein, L. and Scadden, D. T.: Bayesian approach to single-cell differential expression analysis, *Nature Methods*, Vol. 11, No. 7, pp. 740–742 (2014).

[9] Soneson, C. and Delorenzi, M.: A comparison of methods for differential expression analysis of RNA-seq data., *BMC Bioinformatics*, Vol. 14, No. 1, p. 91 (2013).

© 2014 Information Processing Society of Japan