様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３１　年　　４　月　２６　日現在

機関番号：　１２６０１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2018
課題番号：　１５Ｋ０６９１９

研究課題名（和文）ロングリード時代に対応したトランスクリプトームデータ解析ガイドラインの構築

研究課題名（英文）Development of guidelines for transcriptome data analysis with long-reads

研究代表者
門田　幸二（Kadota, Koji）

東京大学・大学院農学生命科学研究科（農学部）・准教授

研究者番号：６０３９２２２１

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：生命科学分野では、生物を構成するサンプル内で働いている数多くの転写物（トランスクリプトームと呼ばれる）の塩基配列を調べたり、その働いている量（発現量）を網羅的に調べる作業が行われる。具体的には、様々な条件間で取得されたサンプル同士が発現量の観点でどの程度似ているかや、どの転写物（遺伝子と同じ意味）の発現量が比較する条件間で異なっているかを調べる作業が行われる。
本研究では、任意の条件間でデータの類似度をこれまで使われてこなかったスコアを用いて客観的に示せること、その数値が条件間で発現が異なる転写物の割合と大まかな相関があること、そのやり方をウェブツールTCC-GUI内に実装して提供した。

研究成果の学術的意義や社会的意義
サンプル間の全体的な類似傾向を眺めるクラスタリングは、トランスクリプトームデータ解析において、必ずといっていいほどよく行われる作業である。しかしながら、得られる結果を都合よく主観的に評価することもできるため、任意のグルーピングにおける全体的な類似度を客観的に評価する必要性やその枠組みの提供は重要である。本研究で提案したシルエットスコアは、クラスタリング結果だけでなく、発現変動解析結果の解釈にも援用できるものであり意義深いものと考える。

研究成果の概要（英文）：In the life sciences field, researchers have investigated nucleotide sequences and expression levels of transcripts (called transcriptome) working in a sample that constitutes an organism. They include the measurement of expression similarities between samples and the identification of genes differentially expressed between conditions of interest. In this research, we proposed to use Silhouette scores to objectively estimate the degree of separation between groups of interest. We confirmed that silhouettes is useful for exploring data with predefined group labels. It would help provide both an objective evaluation of the sample clustering results and insights into the differential expression results with regard to the compared groups.

研究分野：　バイオインフォマティクス

キーワード：　RNA-seq　発現変動解析

様　式　C－19、F－19－1、Z－19、CK－19（共通）

１．研究開始当初の背景

生体内で発現している転写物配列や発現量を網羅的に調べるトランスクリプトーム解析は、次世代シークエンサ(NGS)を用いた RNA-seq と呼ばれる手段が主流である。RNA-seq は、(i)全転写物の塩基配列決定、転写物レベルの(ii)発現量推定や(iii)発現変動解析など、モデルや非モデル生物を問わず様々な目的で利用されている。(i)については、これまでの第二世代 NGS 機器から得られる、150 塩基程度からなるリードを用いたアセンブリ（転写物配列の再構築）で四苦八苦する時代は終焉を迎えつつある。第三世代と呼ばれる Pacific Biosciences (PacBio)社の NGS 機器から得られる数千塩基長のロングリード情報、および各種エラー補正アルゴリズムによって得られる高解像度のパーソナルトランスクリプトーム配列取得後を見据えた、(ii)や(iii)の発現解析を効率的に行うためのガイドライン構築は重要な取り組みである。

RNA-seq に基づく発現解析の基礎データは、転写物または遺伝子ごとにマップされたリード数を数え上げた、カウントデータと呼ばれるものである。このデータは、(iii)の発現変動解析を統計的手法で実行する際の推奨入力データであり、比較するサンプル間で発現の異なる転写物(DEG)の同定がなされる。その一方で、同一サンプル内で転写物間の発現レベルの大小関係を調べたい場合には、配列長が長いほど沢山シークエンスされるという影響をなくすべく、配列長補正を基本とした RPKM または FPKM 値をその転写物の(ii)発現量とみなして比較する。つまり、解析目的によって入力データを変更する必要があるといわれており、研究代表者らもこの枠組みで研究を進めてきた。

しかしながら、これらの知見およびガイドラインは、50 塩基程度未満のショートリード時代のデータに基づいている。その一方で、発現解析用 NGS 機器の Illumina HiSeq 2500 は、2014 年度中に両端 250 塩基ずつ読める見込みであり、(ii)における配列長補正に関連した前処理の重要性は低下していくものと思われる。また、(iii)の発現変動解析で用いられる統計的手法は、かつては転写物レベルの解析時に、異なる転写物間で共有される exon (shared exon)に対するカウント数の割り振り（カウントデータ作成手順）が不明瞭であった。しかし、複数 exon にまたがるリード長になりつつあり、2015 年頃より曖昧性が大幅に軽減されたトランスクリプトーム配列へのミディアムリードのマッピングが本格化すると考えられる。

研究代表者はこれまで、(iii)の発現変動解析を行うためのショートリード時代のモデル（負の二項分布）を基本とした統計的手法 TCC を開発し、解析可能な実験デザインの拡張を継続的に行ってきた。本格化するミディアム～ロングリード時代への対応が急務であるものの、カウントデータの性質（モデル）が類似していれば、原理的にミディアム～ロングリード由来カウントデータへもそのまま適用可能だと思われる。また、「リード長 ≈ 転写物配列長」であれば、「カウント数 ≈ 相対発現レベル」と解釈できる。これらを鑑み、これまで解析目的別に分けられていた従来のデータ解析ガイドラインは、カウントデータを入力とする統計的手法の枠組みで統一化されていくのではないかという着想を得た。

２．研究の目的

本研究の目的は、2015 年以降本格化するミディアム～ロングリード時代の NGS 解析に対応すべく、統一的なトランスクリプトーム解析のためのガイドライン構築である。具体的には、研究代表者らがこれまで開発してきた比較トランスクリプトーム解析手法の適用可能範囲の拡張および改良を目的とした。

３．研究の方法

(1) 多群間比較用の発現変動解析における推奨パイプライン構築を行った。本研究では 3 群間比較に焦点を絞り、シミュレーションデータおよびリアルデータを用いて感度・特異度・計算時間の評価を行った。リアルデータは、Blekhman ら(2010)の 3 生物種間比較用カウントデータを用いた。シミュレーション解析は、R パッケージ TCC が提供する simulateReadCounts 関数を用いて行った。シミュレーション解析における感度・特異度の評価は、「どこかの群間で発現変動している順にソート」した結果に基づいて、ROC 曲線の下部面積（AUC）を用いて行った。シミュレーション条件は下記の通りである：遺伝子数は 10,000 個で固定、発現変動遺伝子(DEG)の割合は 5 または 25%の 2 種類（P_{DEG} = 0.05 or 0.25）、各群への DEG の割り当ては 7 通り、反復数（Nrep）は 4 通り(1, 3, 6, and 9)。計 12 個の手法（解析パイプライン）を比較した。

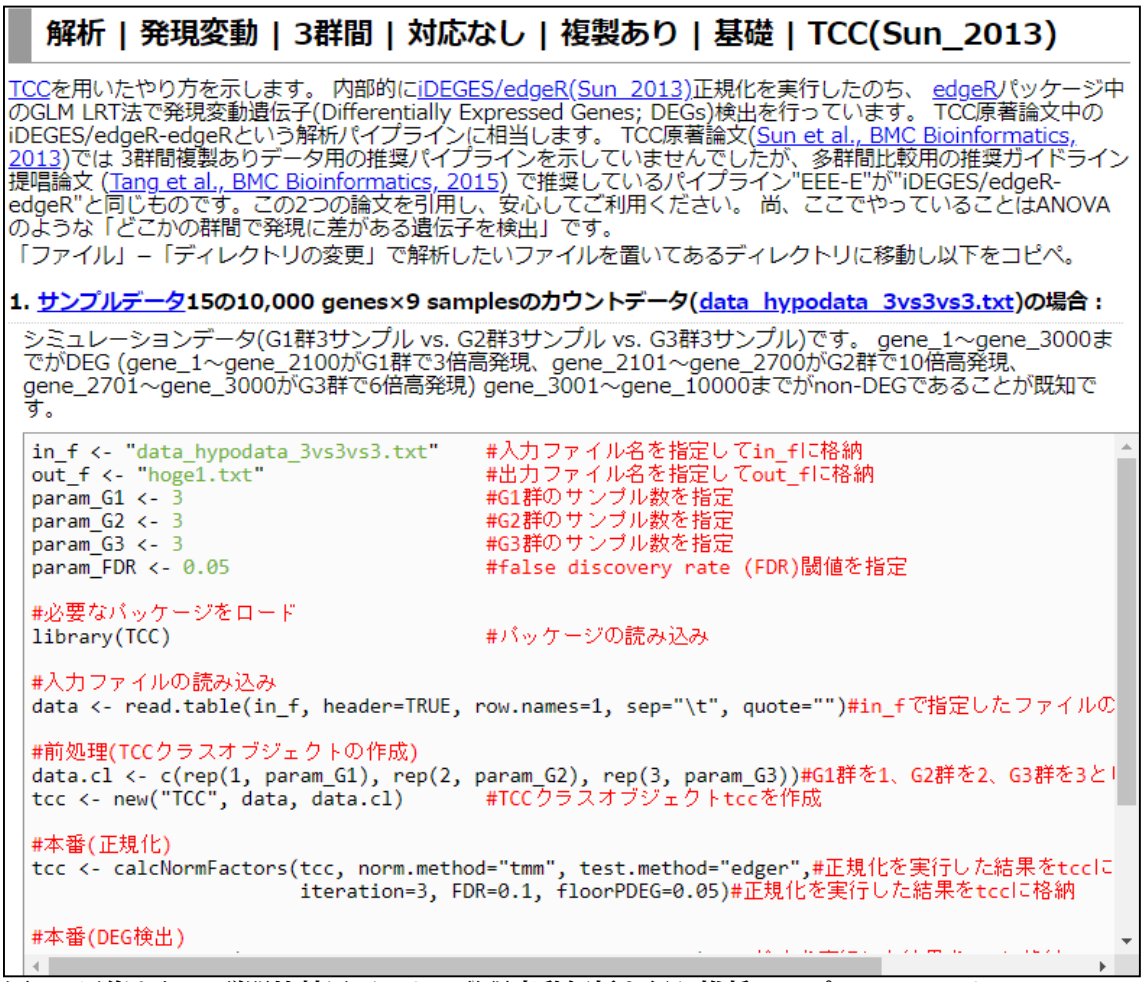
(2) ロングリードの代表格である PacBio データを用いた発現解析について検討を行った。結果として、PacBio の生データ形式（bax/bas.h5）は一般的な形式(FASTQ)とは異なっていること、FASTQ 形式に変換した後のデータだとリード数が大幅に少なくなること、公共データベースでは PacBio の生データが提供されていないことなど、データ解析環境に関する情報収集を中心に行った。

(3) サンプル間クラスタリング結果と発現変動解析結果（特に発現変動遺伝子の割合 P_{DEG}）の関連について調査した。具体的には、比較する群間での P_{DEG} 値が大きいほど、クラスタリング結果において群が明瞭に分かれているという正の相関を、客観的な指標であるシルエットスコア（Silhouette scores）で示し得るというアイデアの確認を行った。評価はシミュレーショ

ンおよびリアルデータを用いて行った。シミュレーションデータについては、(1)と同様 TCC パッケージ中の関数を用いて様々なシナリオに基づくカウントデータを生成した。Blekhman ら(2010)の 3 生物種間比較用カウントデータを用いて予備的な解析を行い、反復の多い 2 群間比較用リアルカウントデータを用いて、より詳細な評価を行った。具体的には、ReCount から得た Bottomly らのデータセットおよび Cheung らのデータセットを用いて、様々な反復数における P_{DEG} とシルエットスコアの相関を調べた。

４．研究成果

(1) 多群間比較用の発現変動解析手法の比較結果として、推奨ガイドラインを構築することができた。具体的には、反復ありデータの場合は *EEE-E* という解析パイプラインが推奨となった。*EEE-E* は、TCC 内部で edgeR というパッケージを繰り返し実行して頑健性を高めたパイプラインである。また、反復なしデータの場合は *SSS-S* という解析パイプラインが推奨となった。*SSS-S* は、TCC パッケージ内部で DESeq2 というパッケージを繰り返し実行して頑健性を高めたパイプラインである。いずれも TCC パッケージが提供する解析パイプラインであり、反復の有無を自動判定して推奨パイプラインをデフォルトとして実行する仕様となっている。エンドユーザは、実用上は反復の有無を気にすることなく TCC パッケージをデフォルトで実行すればよい。本研究内容は、査読付き論文として発表済みである（Tang et al., *BMC Bioinformatics*, 2015）。また、ウェブサイト「（R で）塩基配列解析」上でエンドユーザが使いやすい形で提供している（図 1）。



解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 基礎 | TCC(Sun_2013)

TCCを用いたやり方を示します。 内部的にiDEGES/edgeR(Sun 2013)正規化を実行したのち、 edgeRパッケージ中のGLM LRT法で発現変動遺伝子(Differentially Expressed Genes; DEGs)検出を行っています。 TCC原著論文中のiDEGES/edgeR-edgeRという解析パイプラインに相当します。 TCC原著論文(Sun et al., BMC Bioinformatics, 2013)では 3群間複製ありデータ用の推奨パイプラインを示していませんでしたが、多群間比較用の推奨ガイドライン提唱論文 (Tang et al., BMC Bioinformatics, 2015) で推奨しているパイプライン"EEE-E"が"iDEGES/edgeR-edgeR"と同じものです。この2つの論文を引用し、安心してご利用ください。 尚、ここでやっていることはANOVAのような「どこかの群間で発現に差がある遺伝子を検出」です。
「ファイル」－「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1. サンプルデータ15の10,000 genes×9 samplesのカウントデータ(data_hypodata_3vs3vs3.txt)の場合：

シミュレーションデータ(G1群3サンプル vs. G2群3サンプル vs. G3群3サンプル)です。 gene_1～gene_3000までがDEG (gene_1～gene_2100がG1群で3倍高発現、gene_2101～gene_2700がG2群で10倍高発現、gene_2701～gene_3000がG3群で6倍高発現) gene_3001～gene_10000までがnon-DEGであることが既知です。

```
in_f <- "data_hypodata_3vs3vs3.txt"     #入力ファイル名を指定してin_fに格納
out_f <- "hoge1.txt"                    #出力ファイル名を指定してout_fに格納
param_G1 <- 3                           #G1群のサンプル数を指定
param_G2 <- 3                           #G2群のサンプル数を指定
param_G3 <- 3                           #G3群のサンプル数を指定
param_FDR <- 0.05                       #false discovery rate (FDR)閾値を指定

#必要なパッケージをロード
library(TCC)                            #パッケージの読み込み

#入力ファイルの読み込み
data <- read.table(in_f, header=TRUE, row.names=1, sep="\t", quote="")#in_fで指定したファイルの

#前処理(TCCクラスオブジェクトの作成)
data.cl <- c(rep(1, param_G1), rep(2, param_G2), rep(3, param_G3))#G1群を1、G2群を2、G3群を3と
tcc <- new("TCC", data, data.cl)        #TCCクラスオブジェクトtccを作成

#本番(正規化)
tcc <- calcNormFactors(tcc, norm.method="tmm", test.method="edger",#正規化を実行した結果をtccに
                       iteration=3, FDR=0.1, floorPDEG=0.05)#正規化を実行した結果をtccに格納

#本番(DEG検出)
```

図 1. 反復あり 3 群間比較用データの発現変動解析を行う推奨パイプラインのスクリーンショット。エンドユーザは、本サイトの推奨手順通りに R および関連パッケージをインストールしておけば、サンプルデータの発現変動解析をコピペで実行可能である。

(2) ロングリードデータを用いた発現解析の可能性については、「困難である（時期尚早である）」という結論が得られた。公共 DB では PacBio の生データが提供されていないことが主な理由である。他の理由としては、PacBio 用 *de novo* アセンブラである HGAP は bax.h5 ファイルのみ入力として受け付け、256GB 程度のメモリを搭載した Linux マシンが必要であることなどが挙げられる。つまり、エンドユーザが気軽に取り扱えない条件が揃っているということである。これらの知見については、解説記事としてまとめた（谷澤ら, *日本乳酸菌学会誌*, 2016a）。

(3) サンプル間クラスタリング結果と発現変動解析結果の関係性については、(1)の推奨ガイド

ラインに関する研究から着想を得たものである。比較する群間での P_{DEG} 値が大きいほど、クラスタリング結果において比較する群が明瞭に分かれているという正の相関を、客観的な指標であるシルエットスコア（Silhouette scores）で示すことができた。シルエットスコアが 0 に近いほど P_{DEG} 値も 0 に近づき、クラスタリング結果において比較する群が入り混じっている点が特に重要である。なぜ発現変動遺伝子が存在しないのかを、サンプル間クラスタリング結果とともに客観的な数値（シルエットスコア）で表現することができるからである。本研究内容は、査読付き論文として発表済みである（Zhao et al., *Biol Proced Online*, 2018）。任意のグルーピングに対してシルエットスコアを計算する R コードは、エンドユーザが気軽に利用できるようにウェブサイト「（R で）塩基配列解析」上で提供している（図 2）。また、別プロジェクトで開発したウェブツール TCC-GUI 上でも、任意のグルーピングに対するシルエットスコアの表示機能を実装している（Su et al., *BMC Res. Notes*, 2019）。

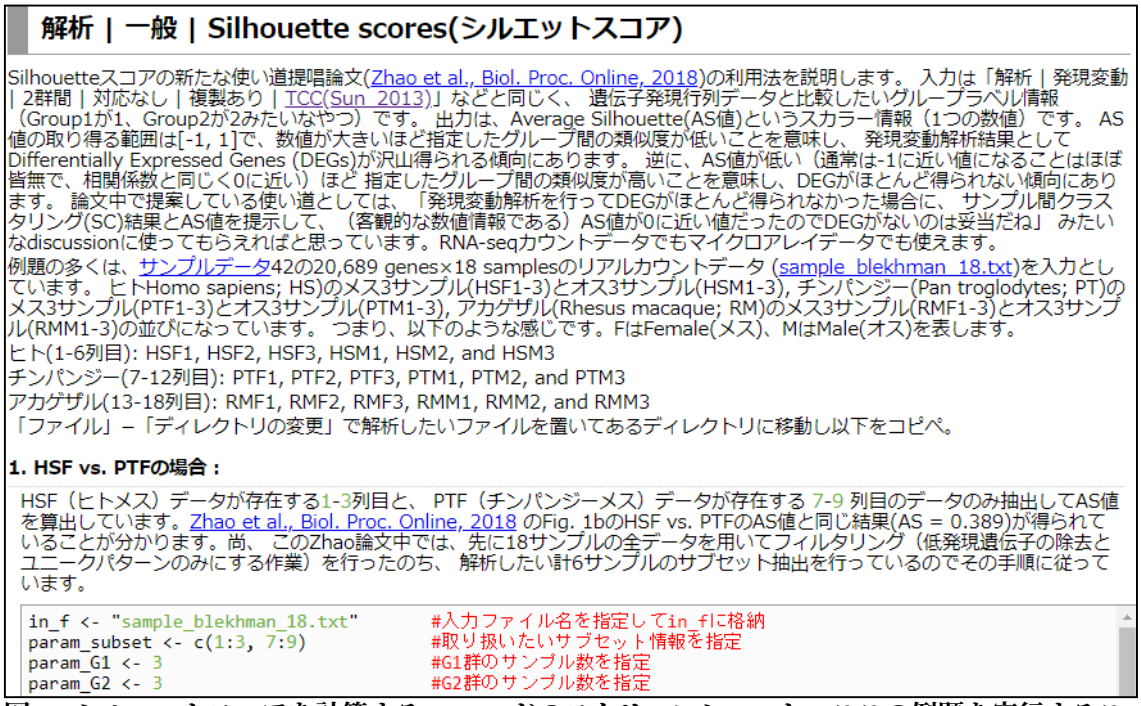


図 2. シルエットスコアを計算する R コードのスクリーンショット。ここの例題を実行することで、原著論文中の Fig. 1 の結果を再現できるようになっている。

５．主な発表論文等

〔雑誌論文〕（計　5 件）

1. Su W, Sun J, Shimizu K, Kadota K, TCC-GUI: a Shiny-based application for differential expression analysis of RNA-Seq count data. *BMC Res Notes*, **12**:133, 2019. **査読有**, https://doi.org/10.1186/s13104-019-4179-2
2. Zhao S, Sun J, Shimizu K, Kadota K, Silhouette scores for arbitrary defined groups in gene expression data and insights into differential expression results. *Biol Proced Online*, **20**:5, 2018. **査読有**, https://doi.org/10.1186/s12575-018-0067-8
3. **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, 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**(3):187-195, 2016b. **査読無**
4. **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, 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**(2):101-110, 2016a. **査読無**
5. Tang M, Sun J, Shimizu K, Kadota K, Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. *BMC Bioinformatics*, **16**:361, 2015. **査読有**, https://doi.org/10.1186/s12859-015-0794-7

〔学会発表〕（計　2 件）

1. **湯敏, 孫建強, 清水謙多郎, 門田幸二**, Benchmarking methods for simulation and analysis of RNA sequencing data. **生命医薬情報学連合大会 2015 年大会**, 2015 **年**
2. **孫建強, 湯敏, 清水謙多郎, 門田幸二**, DEGES-based method for estimating the gene dispersions of RNA-seq count data without replicates. **生命医薬情報学連合大会 2015 年大会**, 2015 **年**

〔その他〕

ホームページ（http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html）

６．研究組織
(2)研究協力者
研究協力者氏名：寺田　朋子
ローマ字氏名：Tomoko Terada