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

令和　４ 年　 ５ 月 ３０ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１５２１

研究課題名（和文）ウェブサイト「(Rで)塩基配列解析」の情報更新・拡充

研究課題名（英文）Update and expansion of the website "nucleotide sequence analysis (with R)"

研究代表者
門田　幸二（KADOTA, Koji）

東京大学・大学院情報学環・学際情報学府・准教授

研究者番号：６０３９２２２１

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：ウェブサイト「(Rで)塩基配列解析」は、主に塩基配列データや遺伝子発現データ解析をフリーソフトウェアRで効率的に行うための包括的な情報サイトである。本課題は、このウェブサイトの①情報更新および②情報拡充を行うことを目的としたものである。本課題期間では、インストール手順や基本的な利用法の更新といった基礎的な事柄だけではなく、解析例のアップデートや追加を行った。また、学術論文を批判的に読み解く重要性について、英語と日本語の両方の論文としてまとめることができた。

研究成果の学術的意義や社会的意義
本ウェブサイトの情報更新・拡充を通じて、実験系研究者自身によるデータ解析を助け、日本全体の研究力向上に貢献できたと考えている。また、「なぜ次から次へと新規手法が開発されるのか？」という問いに対する明快な解答とともに、ただ論文を出せばよいわけではないという研究者としてのあるべき姿（研究倫理）を後進に示すことができた。

研究成果の概要（英文）：The website "nucleotide sequence analysis (with R)" is a comprehensive information site for efficient analysis of nucleotide sequence data and gene expression data using R, an data analysis environment. The purpose is to (1) update and (2) expand the information on this website.
During the term, we not only updated basic matters such as installation procedures and basic usage, but also updated and added analysis examples. We also summarized the importance of critical reading of academic papers.

研究分野：バイオインフォマティクス

キーワード：R　研究倫理　発現変動解析　RNA-seq

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

１．研究開始当初の背景

「(R で)塩基配列解析」は、主に NGS データ解析に関する日本最大規模の情報量を誇るウェブサイトである。基本的な塩基配列解析から、クオリティコントロール、アセンブル、マッピング、発現変動解析、作図など解析例のみでも 1,000 を超える膨大な情報を含んでおり、多くのユーザに利用されている。代表者は、2010 年の本サイト初公開から現在まで、主にエンドユーザからの要望に応えて新規項目の追加（情報拡充）に取り組んできた。理想的には、適切な人員およびエフォートを費やして、リンク切れへの対応など定期的な既存項目の情報更新を行いウェブサイト全体の質を継続的に維持すべきである。しかしながら、これまで全体を通したメンテナンスが追いついておらず、情報の劣化が深刻な課題となっていた。

当時、本サイトはメインページ内のみでも莫大な情報量（306 の項目数、1,000 個以上のフリーソフト R のスクリプト、5,000 以上のリンク先、約 100 万の単純文字カウント数）を含んでいた。さらにそのリンク先には、代表者の提供情報だけでも(a) R のインストール法や基本的な利用法の詳細情報、そして(b) R 以外の各種補足情報（NGS ハンズオン講習会や学会誌に連載中の NGS 関連資料、Linux の基本的な利用法や解析環境の構築手順）などの膨大な情報を含んでいた。本サイトの主要コンテンツである R スクリプトの多くは、概ね半年毎に更新される Bioconductor パッケージの関数群を内部的に利用している。このため、本サイトの提供開始初期のバージョンで動作確認済みの項目であっても、最新バージョンで同じスクリプトを実行するとエラーになる（あるいはその逆）という事例が少なからず存在していた。また、メインページの読み込みのみでもストレスを感じるデータ量となっており、改善を求める声が寄せられていた。

２．研究の目的

本研究の目的は、幅広く利用されているウェブサイト「(R で)塩基配列解析」の情報更新・拡充である。

３．研究の方法

本課題は方法と成果を切り分けで書くことが難しいため、成果のほうに一括して記載する。

４．研究成果

以下では、①～⑤の計 5 つに分けて述べる。

①ウェブサイトの分割

本サイトは、代表者の教育研究に関するほぼ全ての情報を含んでいる。中核となる情報は、R 上で動作する様々な統計解析用プログラムの豊富な解析例である。これらの多くは Bioconductor と呼ばれるサイトから提供されているパッケージを利用している。しかしながら、一定の形式に沿って作成され定評のある Bioconductor パッケージでさえ、パッケージ内の記述内容やその質は開発者ごとに大きく異なり玉石混交である。本サイトの特色は、パッケージ（開発者）間の違いを吸収し、複数の例題とともに多種多様な解析法を入出力の関係が一目でわかる独自形式で提供している点である。それゆえ、R と比較的関係性の低い一部の項目を本サイトから分離し、「(R で)塩基配列解析のサブ」に移行した（図 1）。これにより、本サイトの表示にかかる時間を減らすことができた。

サブページのほうは、研究代表者が 2014 年から日本乳酸菌学会誌上で連載を継続している「NGS データの解析手法」の解説記事の原稿やウェブ資料の追加を中心に行った。連載第 14 回（図 2）と 15 回は、ウェブツール Galaxy を用いた RNA-seq 解析に関するものである。R や Linux 利用の敷居が高いヒト向けの指南書でもあるため、研究分野全体の裾野を広げるという意味で重要な貢献だと考えている。

図 1．分割後のウェブサイト。

図 2．連載第 14 回の項目。冒頭部分のスクリーンショットを示している。

②ウェブサイトの質向上

文言の表記ゆれやリンク切れの修正、消去されたパッケージの削除、新規項目や最新プログラムおよび原著論文の追加などを行った。また、令和 2 年度には、安定的な運用のためのウェブサーバ移行を行うとともに、関数内で使用するオプションを適切なものに修正する作業なども行った。

また、本サイト利用の前提である R、RStudio、そしてパッケージのインストール手順の内容をアップデートするとともに、「基本的な利用法」の解説内容も更新した。特に後者については、塩基配列を入力としてその翻訳配列を得る基礎的な項目において、そのスクリプトがうまく動作しない事象を確認したためである。利用者にとっては気にも留めない部分かもしれないが、地道に丁寧な動作確認を通じて発見できたことは、本課題の目的に完全に合致した仕事だと考えている。

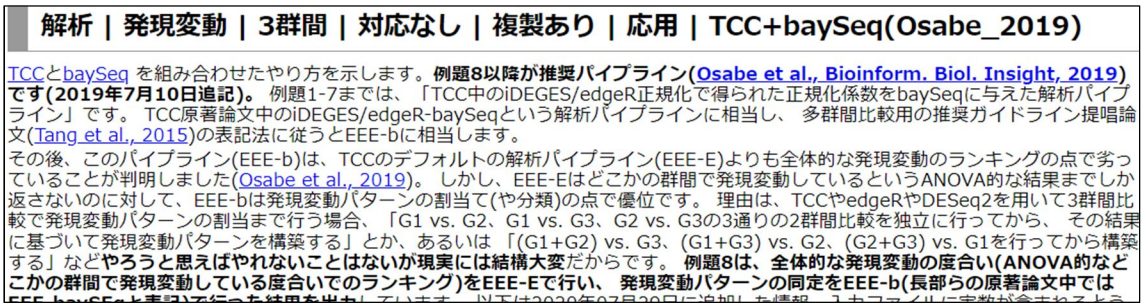
③研究倫理に関する注意喚起

上記の新規項目の追加に関連して、当初は single-cell RNA-seq (scRNA-seq)解析関連の項目を重点的に追加する方向で作業していた。しかしこの過程で、これまで bulk RNA-seq の論文中で報告済みのいくつかの事柄が無視されていることに気づいた。具体的には、「bulk RNA-seq 用に開発された"有名な"データ正規化法」を「発現変動遺伝子数やその群間での偏りが非常に大きい scRNAseq 用に開発されたデータ正規化法」と比較し、後者のほうがよいと結論付ける不適切な論文を発見した。一見まともそうなロジックに思えるが、実際には「発現変動遺伝子数やその群間での偏りが非常に大きい場合にも対応可能な bulk RNA-seq 用の頑健なデータ正規化法」は存在する（がそれとの比較がなされていない）。また、scRNA-seq を bulk RNA-seq と区別する大きな特徴として、ゼロカウントのデータの多さ（ゼロ過剰）もしばしば強調されている。しかしながら、おそらく最初にゼロ過剰の特徴について報告がなされたのは、実際には 2013 年の bulk RNA-seq 用カウントデータモデル論文である。これらの調べた限りの事実関係を論文にまとめた（Kadota and Shimizu, 2020）。これは、ただ論文を出せばよいわけではないという研究者としての倫理観を後進に示した重要な成果といえる。

上記論文は英語であるため、「なぜ次から次へと新規手法が開発されるのか?」というタイトルで、日本語の解説記事としてまとめた（門田 and 清水, 2021）。研究者レベルでも「なぜ次から次へと新規手法が開発されるのか？」という問いに対して明確に答えられるヒトは少ない。たとえ同程度以下の性能であっても、従来法よりも高速であれば存在意義が認められるといったような、多様な評価基準のためだという理由付けが一般的である。しかし現実には、自分にとって都合のよい論文のみを引用して都合のよい結論を導く不適切な論文が一定数存在する。これを原著論文から読み解くのは大学院生レベルでは容易ではない。そのため、本成果は学術論文を批判的に読み解く重要性について、日本語で論点整理した重要な貢献だと考えている。なお、このウェブ資料は、サブページのほうに掲載済みである。

④発現変動解析のガイドライン構築や解説

ユーザから要望が寄せられていた、3 群間比較での RNA-seq カウントデータを入力として行う発現パターン分類のガイドラインに関する論文発表（Osabe et al., 2019）を行うとともに、本サイト上の項目の 1 つとして示した（図 3）。また、本サイト上で解析例を示した。また、RNA-seq 解析分野で誤用が目立つ TPM 値の意味合いを改めて述べるとともに、サンプル間比較の際に本来使うべきものではないことを注意喚起する解説記事を公開した（寺田ら，2020）。



解析 | 発現変動 | 3群間 | 対応なし | 複製あり | 応用 | TCC+baySeq(Osabe_2019)

TCCとbaySeq を組み合わせたやり方を示します。**例題8以降が推奨パイプライン(Osabe et al., Bioinform. Biol. Insight, 2019)です(2019年7月10日追記)。** 例題1-7までは、「TCC中のiDEGES/edgeR正規化で得られた正規化係数をbaySeqに与えた解析パイプライン」です。 TCC原著論文中のiDEGES/edgeR-baySeqという解析パイプラインに相当し、 多群間比較用の推奨ガイドライン提唱論文(Tang et al., 2015)の表記法に従うとEEE-bに相当します。

その後、このパイプライン(EEE-b)は、TCCのデフォルトの解析パイプライン(EEE-E)よりも全体的な発現変動のランキングの点で劣っていることが判明しました(Osabe et al., 2019)。 しかし、EEE-Eはどこかの群間で発現変動しているというANOVA的な結果までしか返さないのに対して、EEE-bは発現変動パターンの割当て(や分類)の点で優位です。 理由は、TCCやedgeRやDESeq2を用いて3群間比較で発現変動パターンの割当まで行う場合、「G1 vs. G2、G1 vs. G3、G2 vs. G3の3通りの2群間比較を独立に行ってから、 その結果に基づいて発現変動パターンを構築する」とか、あるいは 「(G1+G2) vs. G3、(G1+G3) vs. G2、(G2+G3) vs. G1を行ってから構築する」など**やろうと思えばやれないことはないが現実には結構大変**だからです。 **例題8は、全体的な発現変動の度合い(ANOVA的などこかの群間で発現変動している度合いでのランキング)をEEE-Eで行い、 発現変動パターンの同定をEEE-b(長部らの原著論文中では**

図 3．発現パターン分類の推奨ガイドラインに関する項目（冒頭部分のみ）。

⑤パッケージのバージョンの違いの影響に関する注意喚起

本サイトで提供する解析例の動作確認は、エラーメッセージが出た場合のみ対応というのが基本方針であった。しかしクローズドのハンズオン講習会の準備作業中に、偶然、発現変動解析分野で有名なパッケージである edgeR のバージョンの違いによる結果（具体的には FDR 閾値を満たす発現変動遺伝子数）の大きな違いを発見したため、注意喚起を行うことができた。

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　6件）

１．著者名 Kadota K and Shimizu K	４．巻 11
２．論文標題 Commentary: A systematic evaluation of single cell RNA-seq analysis pipelines.	５．発行年 2020年
３．雑誌名 Frontiers in Genetics	６．最初と最後の頁 941
掲載論文のDOI（デジタルオブジェクト識別子） 10.3389/fgene.2020.00941	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 寺田 朋子，清水 謙多郎，門田 幸二	４．巻 31
２．論文標題 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）	５．発行年 2020年
３．雑誌名 日本乳酸菌学会誌	６．最初と最後の頁 25-34
掲載論文のDOI（デジタルオブジェクト識別子） 10.4109/jslab.31.25	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Osabe T, Shimizu K, Kadota K	４．巻 13
２．論文標題 Accurate Classification of Differential Expression Patterns in a Bayesian Framework With Robust Normalization for Multi-Group RNA-Seq Count Data.	５．発行年 2019年
３．雑誌名 Bioinform Biol Insights	６．最初と最後の頁 正確な情報だとエラーになる
掲載論文のDOI（デジタルオブジェクト識別子） 10.1177/1177932219860817	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 寺田朋子，清水謙多郎，門田幸二	４．巻 30
２．論文標題 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）	５．発行年 2019年
３．雑誌名 日本乳酸菌学会誌	６．最初と最後の頁 153-161
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 寺田朋子，坂本光央，清水謙多郎，門田幸二	４．巻 30
２．論文標題 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）	５．発行年 2019年
３．雑誌名 日本乳酸菌学会誌	６．最初と最後の頁 38-45
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　1件／うち国際学会　0件）

１．発表者名 門田幸二
２．発表標題 トランスクリプトーム解析分野のバイオインフォマティクス系研究者の雑感
３．学会等名 数理生物学セミナー2020@TMDU（招待講演）
４．発表年 2020年

〔図書〕　計1件

１．著者名 門田幸二	４．発行年 2019年
２．出版社 羊土社	５．総ページ数 255
３．書名 RNA-Seqデータ解析　WETラボのための鉄板レシピ（坊農秀雅 編）	

〔産業財産権〕

〔その他〕

(Rで)塩基配列解析
http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html
(Rで)塩基配列解析のサブ
http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq2.html

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	寺田　朋子 (Terada Tomoko)	東京大学・大学院農学生命科学研究科・学術専門職員	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関