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12回 Galaxy：ヒストリーとワークフロー

寺田 朋子[1]、大田 達郎[2*]、清水 謙多郎[1, 3]、門田 幸二[1, 3*]

[1] 東京大学 大学院農学生命科学研究科
[2] 情報・システム研究機構 データサイエンス共同利用基盤施設
ライフサイエンス統合データベースセンター
[3] 東京大学 微生物科学イノベーション連携研究機構

Galaxyは、ウェブブラウザ上でマウスを操作して行うGUIベースの（Linuxコマンドを覚える必要のない）データ解析環境である。前回は、Galaxyの概要と公共サーバGalaxy Mainの基本的な利用法について述べた。第12回は、前回Galaxy Main上で行った解析結果（**ヒストリー**）の他の研究者との共有や、以前行った解析手順（**ワークフロー**）を他のデータに適用する手段を中心に述べる。また、ヒストリー間のデータのコピーや、自身のPCを経由しないGalaxy Mainへのデータの取り込みなど、関連した便利な機能についても紹介する。今回の内容は、**ウェブブラウザの違いに起因する不具合**を避けるため、Google ChromeまたはFirefox（Internet Explorerは非推奨）を用いてほしい。ウェブサイト（**Rで）塩基配列解析**（URL: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html）中に本連載をまとめた項目（URL: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_book_JSLAB）が存在する。ウェブ資料（以下、W）や関連ウェブサイトなどを効率的に活用してほしい。

Key words：NGS, Galaxy, workflow, history

はじめに

本連載では、これまで主にウェブブラウザInternet Explorer（IE）を用いて動作確認を行ってきた。しかしながら、今回初めてGalaxy[1)]の一部機能（ヒストリーの一覧表示やワークフロー）の実行時に、**IEの使用に起因する不具合**に遭遇した。2018年6月現在の最新OSであるWindows 10は、推奨ウェブブラウザがMicrosoft Edgeである。しかしながら、使い慣れたIEを継続的に利用しているWindowsユーザはまだ相当いると思われる。この問題は、おそらくIEユーザに限定されるものの、原因特定に要する労力および判明時の落胆が非常に大きいため注意してほしい。我々は今回の動作確認をChromeで行ったが、Firefoxでも問題ないと思われる［W1］。

今回も、前回（第11回）に引き続いてGalaxy Main上で作業を行う。必ずしも予め前回の内容を一通り行っておく必要はない。次項で述べる**ヒストリーの共有**によって、前回までの解析結果（ヒストリー）を共有することができるからである。但し、最低限第11回のW4を参考にしてGalaxy Mainのユーザ登録を済ませておくことを推奨する。Galaxy Main に**ログインした状態で作業を行う**ことで、ヒストリーやワークフローに関する今回の内容を十分に活かすことができるからである。尚、ここではGalaxy Mainにユーザ登録済のメインユーザをkadota@iu.a.u-tokyo.ac.jp（以下、**kadota_registered**）とする［W2］。また、ヒストリーを共有する登録済のユーザとしてtomoko.terada@iu.a.u-tokyo.ac.jp（以下、**terada_registered**）を、そして未登録ユーザをkoji.kadota@gmail.com（以下、

*To whom correspondence should be addressed.
Phone ：+81-3-5841-2395
Fax ：+81-3-5841-1136
E-mail：t.ohta@dbcls.rois.ac.jp
kadota@bi.a.u-tokyo.ac.jp

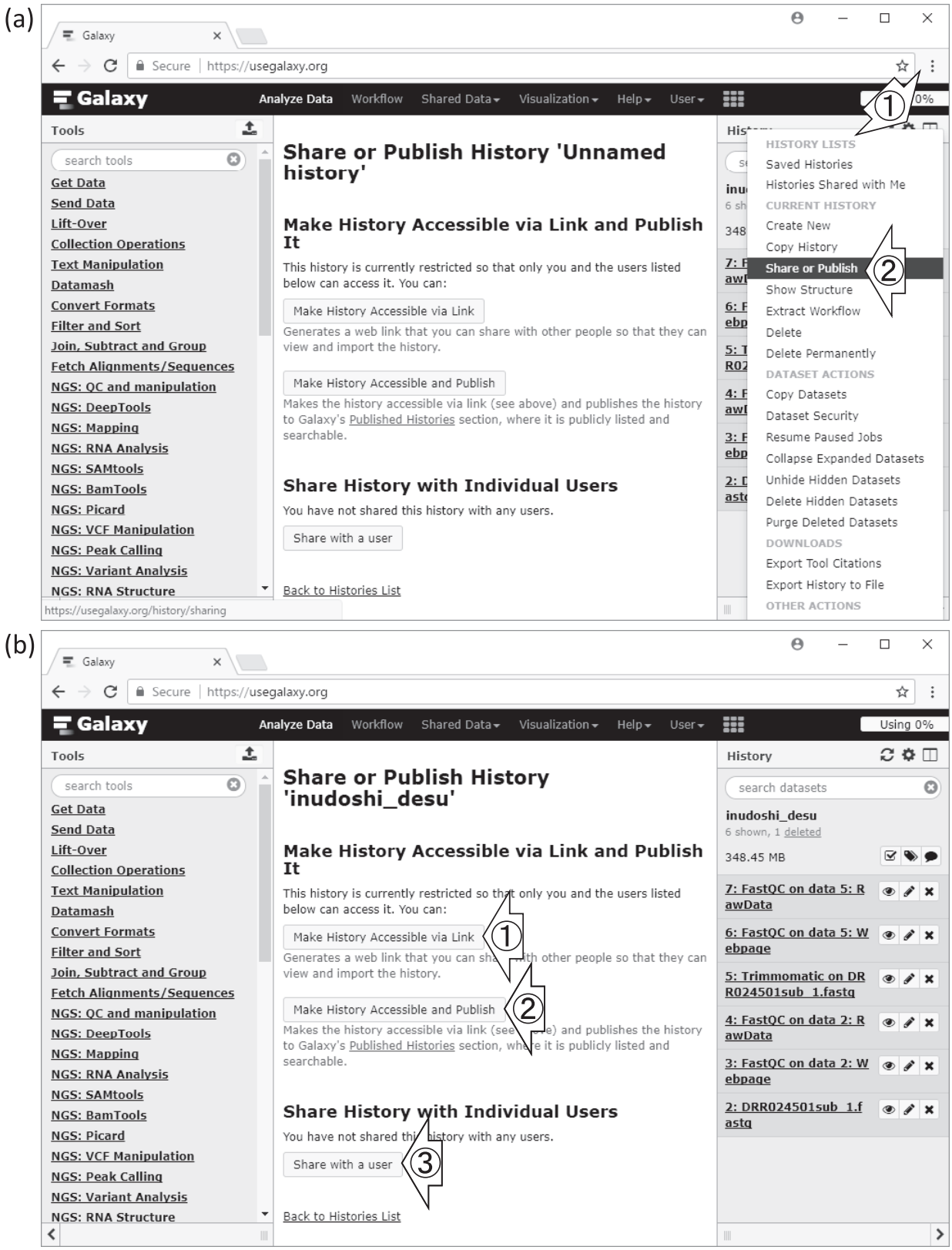

図 1. ヒストリーの共有（その 1）
(a) は W5-3 と、そして (b) は W5-4 と基本的に同じである。(a) でヒストリー名を “inudoshi_desu” に変更後、① History options – ② Share or Publish を実行すると、(b) のように変更が反映される。(b) で① Make History Accessible via Link をクリックした結果が図 2a である。

kadota_unregistered）として話を進める。

ヒストリーの共有

前回までに**kadota_registered**が一通り行った作業は、Illumina MiSeqを用いて得られた30万リードからなるgzip圧縮FASTQファイル（DRR024501sub_1.fastq.gz）のアップロード、FastQCの実行、Trimmomaticの実行、そしてTrimmomatic実行後のファイルに対するFastQCの実行であった［W3］。Galaxy GUI画面の右側にあるヒストリーパネル上で「History options」−「Share or Publish」を選ぶと、中央のパネル上に3つのヒストリー共有手段が提示される［W4-3］（**図1**）。

① Make History Accessible via Link
リンク先URLを知っているヒトのみでヒストリーを共有
② Make History Accessible and Publish
Galaxy MainのPublished Historiesというサイト上で公開
③ Share with a user
特定のユーザのみと共有

①と③の違いが分かりづらいかもしれないが、①のほうはリンク先のURLを知っていれば誰でもヒストリー情報にアクセスできるため、③よりも制限が緩いということになる。しかしながら、そのURLを知っているのは教えられたヒトのみであるため、通常利用の局面においては実質的に同じであろう。②を選択すると、Galaxy MainのPublished Historiesというサイトで公開される。このサイトでは様々なヒストリー名のものが見られるが、たとえ同じ研究グループ内のヒトであったとしても、ヒストリー内部の理解は簡単ではないだろう［W4-5］。尚、このヒストリー名（デフォルトは“Unnamed history”）は、右側のヒストリーパネル上で任意の名前に変更することができる［W5-2］。変更後のヒストリー名“inudoshi_desu”で、改めて「History options」−「Share or Publish」を選ぶと、中央のパネルのほうも反映される［W5-3］。

ここでは、連載第11回で行った**kadota_registered**の解析結果（ヒストリー）を、未登録ユーザである**kadota_unregistered**と共有する①のやり方を示す。**図1b**において①をクリックすると、ヒストリーinudoshi_desuのリンク先URLが現れる［W6-1］（**図2a**）。ヒストリー情報を提供する側である**kadota_registered**の仕事は、基本的にこのURL情報を共有したいヒトにメール送信するだけである［W6-2］。**図2b**は、情報を提供される側の**kadota_unregistered**がこのURLにアクセスした結果である［W6-4］。提供する側（**kadota_registered**）のヒストリーパネル情報に相当するものが、左のほうに見えていることがわかる。提供する側と見栄えが異なるのは、編集権限を持たない閲覧専用のような状態でこのヒストリー情報を眺めているからである。この状態を解除するには、Import historyボタンを押してヒストリーinudoshi_desuを取り込めばよい（インポートすればよい）［W7-1］。これは、Galaxy Mainにログインしていない**kadota_unregistered**のGalaxy環境に、ヒストリー情報inudoshi_desuをコピーしていることに相当する。

提供された側（**kadota_unregistered**）のインポート後の画面は、提供した側（**kadota_registered**）のGalaxy Mainの通常画面と同じである［W7-2］。共有されたヒストリー情報をインポートすることで、基本的に左側のツール選択パネルから自由にプログラムを選択して、独立に解析を行うことができるようになる。もちろんGalaxy Mainにログインしていない状態のため、インポートしたヒストリーやその後独立して行った解析結果は、ウェブブラウザ上の開いているタブまたはブラウザそのものを閉じれば消える［W7-3］。これをどう捉えるかはヒトそれぞれであるが、例えば次のような利用法が考えられる。研究グループ内にGalaxyを使えるヒト（例：**kadota_registered**）がおり、一通りのデータ解析結果がGalaxy Main上に存在するとしよう。データ解析担当者が、Galaxy Main上でのデータ解析結果（のヒストリーのURL情報）をグループ内全員に知らせることで、グループ内の他の研究者（例：**kadota_unregistered**）が解析手順にミスがないかなどをチェックすることができる。

必要に応じてインポートを行うことで、例えばTrimmomatic[2]実行時にデフォルトの平均クオリティスコアを20ではなく30にすればどうなっていたかなどを独自に調べることもできる［W9-6］。但し、Galaxy Mainにログインしていない状態では機能が制限されるため、**kadota_registered**のヒストリーinudoshi_desuに対して**kadota_unregistered**が追加作業を行ったヒストリーを、さらに他者と共有することはできない［W10-1］。もちろん、この追加作業情報を保存することはできる。追加作業を行ったブラウザ上で、アカウントがなければ新規作成するなどしてGalaxy Mainにログインすればよいのである［W11-1］。

ヒストリーの操作

図3aは、**kadota_unregistered**がヒストリーinudoshi_desuをインポートし、追加作業を行った（オプションを変更してTrimmomaticおよびFastQCを実行した）結果のブラウザ上で、**kadota_registered**がログインした直後の画面である［W11-2］。**kadota_registered**が前回のログイン時に行った作業は、ヒストリーinudoshi_desuの作成までであった［W6-1］。**kadota_registered**でログイン後のヒストリーパネルは、ログイン前に**kadota_unregistered**

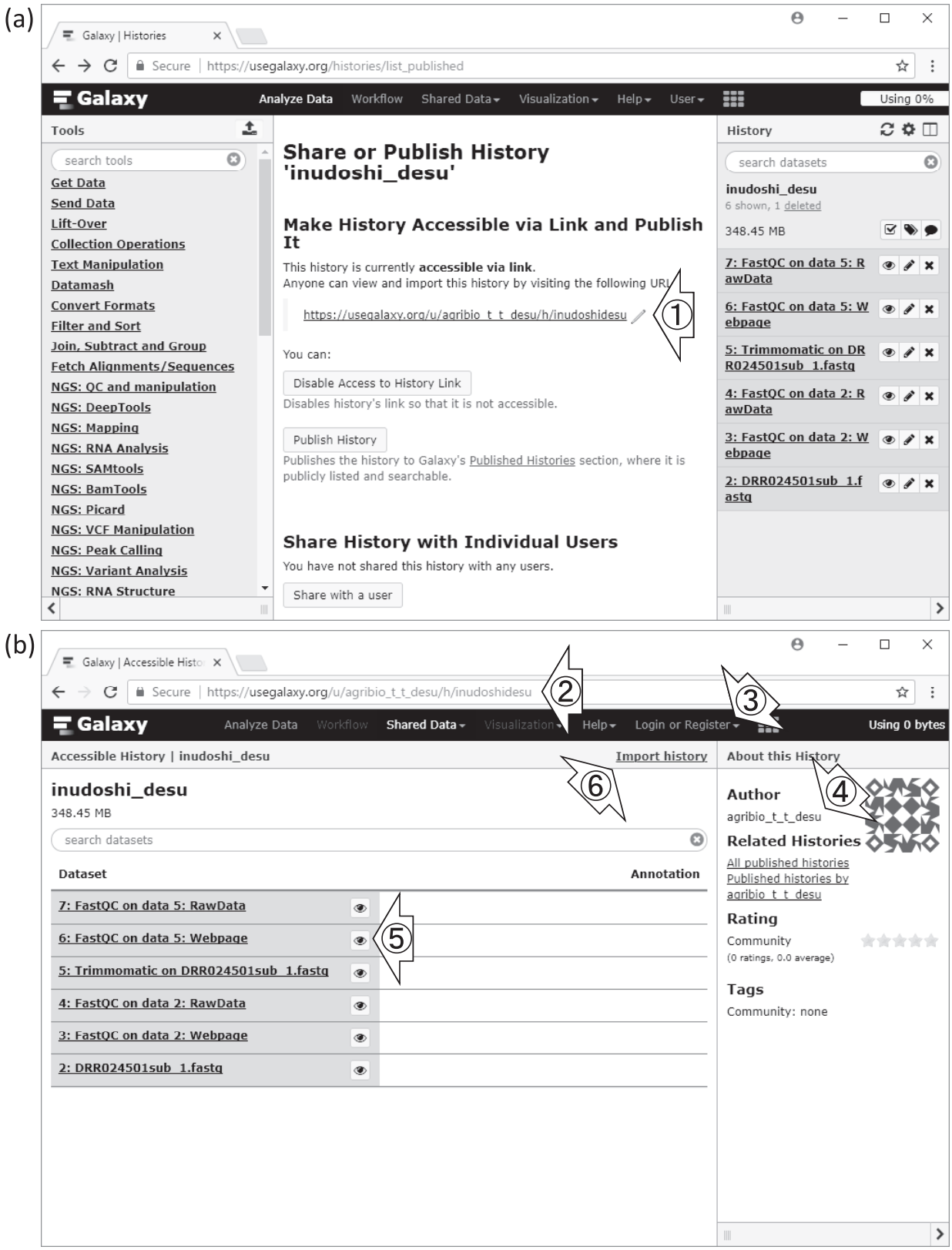

図 2. ヒストリーの共有（その 2）
(a) は W6-1 と、そして (b) は W6-4 から 6-6 と同じ。(a) はヒストリー情報を提供する側の画面、(b) は提供された側がリンク先にアクセスした結果の画面。①はこのヒストリーの URL 情報で、②と同じ。③確かにログインしていないことがわかる。④はヒストリー提供者 **kadota_registered** の Public name。⑤をクリックすると FastQC 実行結果が見られる。⑥ Import history でヒストリー情報を取り込むこともできる。

が追加作業を行ったW9-4までと同じである。**kadota_unregistered**によるヒストリーinudoshi_desuのインポート前後の違いからもわかるように［W7］、**kadota_registered**のログインによって追加作業情報を取り込めていることを意味する。

もちろん、**図3a**の①ヒストリー情報は、W6-1までで作成したヒストリーinudoshi_desuとは別物である。ヒストリーパネルの② View all historiesを実行すると、①現在のヒストリー情報（Current History）が左側に、そして③ W6-1までで作成したヒストリーinudoshi_desuが右側に表示されていることがわかる（**図3b**）。ヒストリーの削除・切替・新規作成・反映などの基本操作［W12］以外の有効なテクニックとしては、ヒストリー間でのデータコピーが挙げられる［W13-3］。他のヒストリーから現在のヒストリーへの一方向のコピーしかできないが、上記の基本操作（ヒストリーの切替と反映）を組み合わせれば問題ない。具体的な利用法としては、例えば新規ヒストリー作成後に手元のデータを改めてアップロードすることなく、既存のヒストリー内にあるデータをコピーして任意の解析をスタートさせることなどが挙げられる［W13-7］。

ワークフローの作成

ワークフローは、以前行った解析手順を別のデータに対して実行したい場合に有効な手段である。ヒストリーinudoshi_desuは、Illumina MiSeqを用いて得られた30万リードからなるpaired-endのforward側データ（DRR024501sub_1.fastq.gz）に対して解析を行ったものである［W14-1］。ここでは、inudoshi_desuをベースにワークフローを作成し、第6回[3]のW3-2で作成したreverse側のファイル（DRR024501sub_2.fastq.gz）に対して独立に行うやり方を説明する。

ワークフローは、作業の流れを示したものであるため[4]、白紙のノートに利用するプログラムを1つずつ選択して手順書を新規作成するように、ワークフローを1から作成することはもちろんできる。しかしながら、作成したいワークフローと似た既存のヒストリーを利用するのがおそらく一般的である。ここでは、「FastQC ➡ Trimmomatic ➡ FastQC」の3ステップからなるオリジナルのinudoshi_desuに対して、1ステップ（平均クオリティスコアの閾値を30に変更したTrimmomatic）追加して実行したもの［W13-3］をベースとして用いるべくCurrent Historyにした［W14-2］。

現在ヒストリーパネル上に表示されているものをワークフローのベースとして利用するには、「History options」−「Extract Workflow」を選択すればよい［W14-3］（**図4a**）。中央パネルに右側のヒストリーパネルに対応したワークフローが表示されるので、ワークフロー名などを任意に変更して編集する［W14-4］。このヒストリーに対応したワークフロー（作業の流れ）は、「FastQC ➡ Trimmomatic（defaultで実行）➡ FastQC ➡ Trimmomatic（オプション変更して実行）」の4ステップからなる。ステップごとに「Include “*プログラム名*” in workflow」のチェックのオンオフを行えるので［W14-8］、ここでは4ステップ目のTrimmomaticのチェックをオフにした、3ステップからなるワークフローを作成した［W14-9］（**図4b**）。左側のツール選択パネルから利用できるようにしておくと、使い勝手がよいだろう［W15-4］。

ワークフローの実行

reverse側のファイル（DRR024501sub_2.fastq.gz）を入力として、作成したワークフローを実行する手順を説明する。入力ファイルとして用いることができるのはヒストリーパネル上で見られるものに限定されるため、まずはその作業を行う［W16-2］。ローカル環境（自分のPC内）にあるファイルを**アップロード**するやり方については、第11回のW5-3で示した。ここでは、指定したURL上にあるファイルをGalaxy上に**ダウンロード**するやり方を示す。アップロードとダウンロードの言葉の違いは本質的ではなく、Galaxyへの他の便利なデータ取り込み手段を知るのが目的である。URLを指定するやり方の一番のメリットは、自分のローカル環境にデータを一旦保存する必要がない点である［W17-8］。W17-5で示したURLは所属機関のものであるが、第6回W3-1で用いたDDBJ SRA[5]のftpサイトのURLなどで置き換えて考えれば、その便利さが想像できるかもしれない。

図5は、ワークフローの（a）実行直前と（b）実行後の画面である。新規ヒストリーを作成し［W17-1］、reverse側のファイル（DRR024501sub_2.fastq.gz）のみをGalaxy上にダウンロードした状態にしておくことで［W17-10］、ワークフローの実行をスムーズに行うことができる。計3ステップからなるワークフローのうち、Step1とStep2は独立に実行可能である一方、Step3（2回目のFastQC）はStep2（Trimmomatic）の実行結果を入力データとして利用する関係となっている。このようなStep2とStep3の依存関係についても、ワークフロー中に正しく記載されているので心配ない［W18-6］。ワークフロー実行ボタンを押すと、3ステップ分のプログラムが文字通り流れ作業的に実行される。このときはStep2がStep1よりも先に終了したが、上記の理由（独立に実行可能）により全く問題ない［W18-10］。このワークフローの場合は、Step3がStep2よりも後に実行されるはずであり、実際そうなっていることがわかる［W18-12］。

ワークフローの編集と共有

ワークフロー内のステップ間の依存関係については、

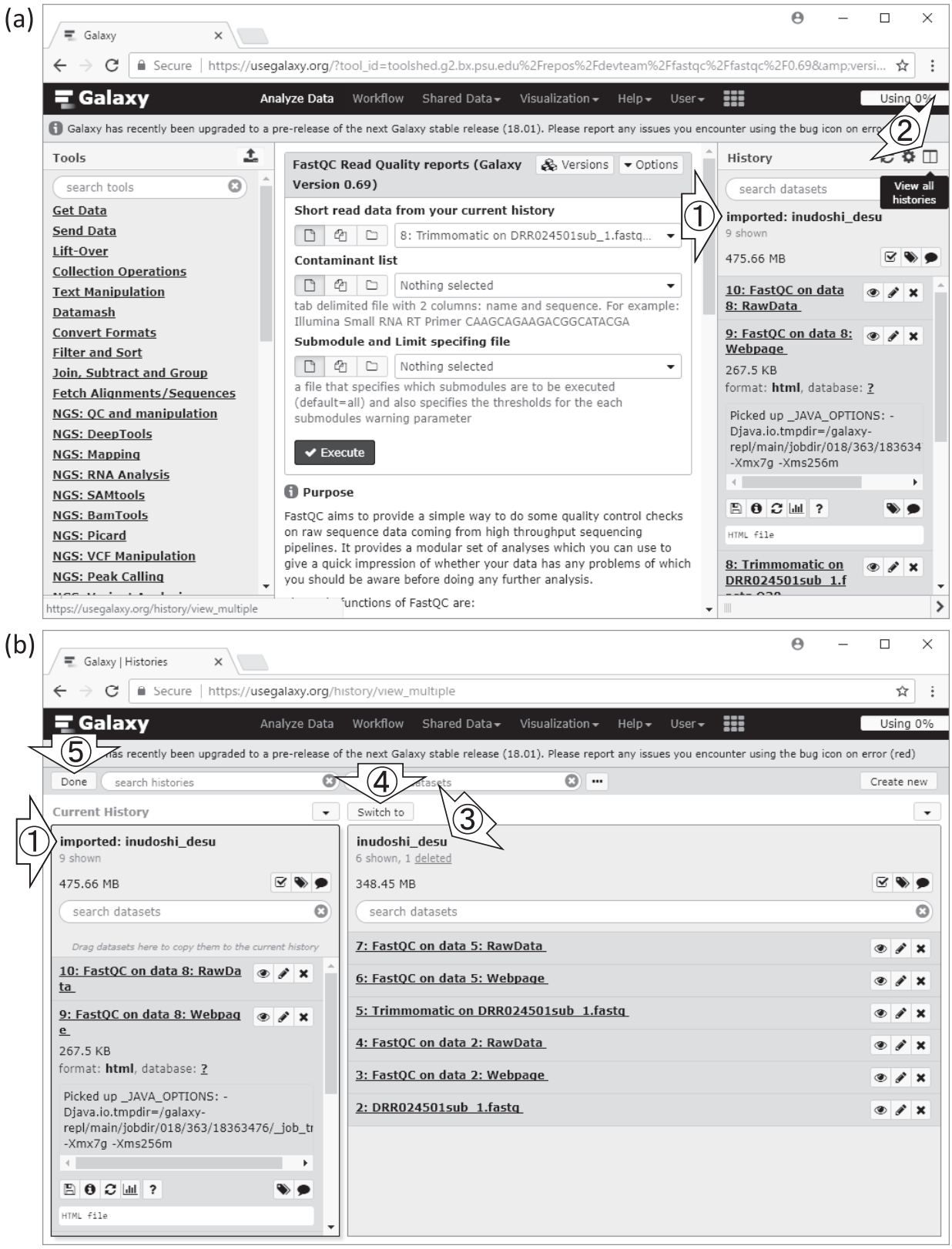

図 3. ヒストリーの操作
(a) は W11-2 と、そして (b) は W11-4 と同じ。①が現在のヒストリー情報。② View all histories をクリックすると、(b) の画面に切り替わり、**kadota_registered** の全ヒストリー情報が見られる。③ W6-1 までで作成したヒストリー inudoshi_desu を現在のヒストリーに切り替えたい場合は、④ Switch to、⑤ Done をクリックすればよい。

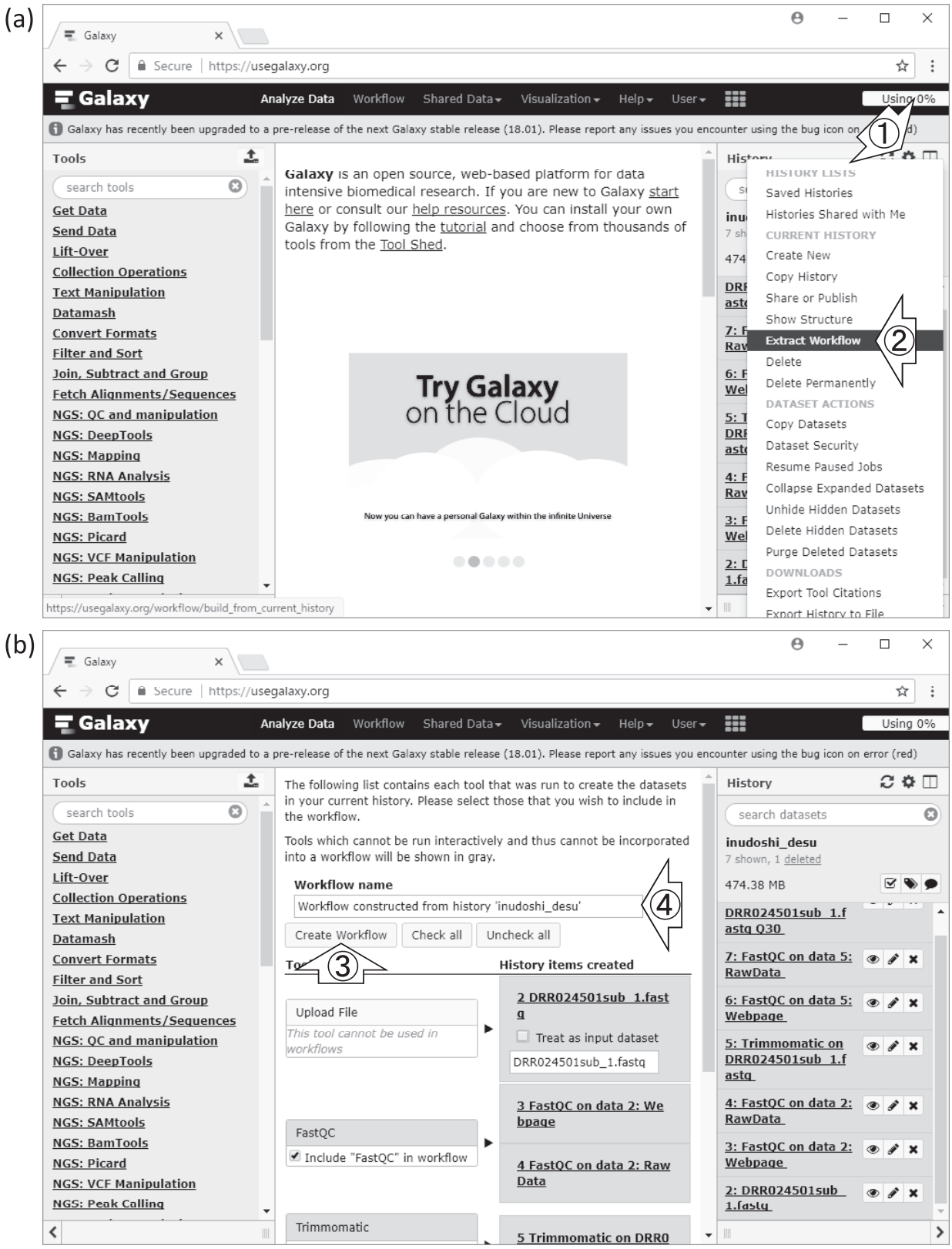

図 4. ワークフローの作成
(a) は W14-3 と、そして (b) は W14-9 と同じ。(a) 作成したいワークフローに近いヒストリーをひな形として利用すべく、Current History にした状態で① History options ➡ ② Extract Workflow。(b) 中央パネルの編集画面上で利用するプログラムを選択したのち、③ Create Workflow すれば④の名前のワークフローが作成される。

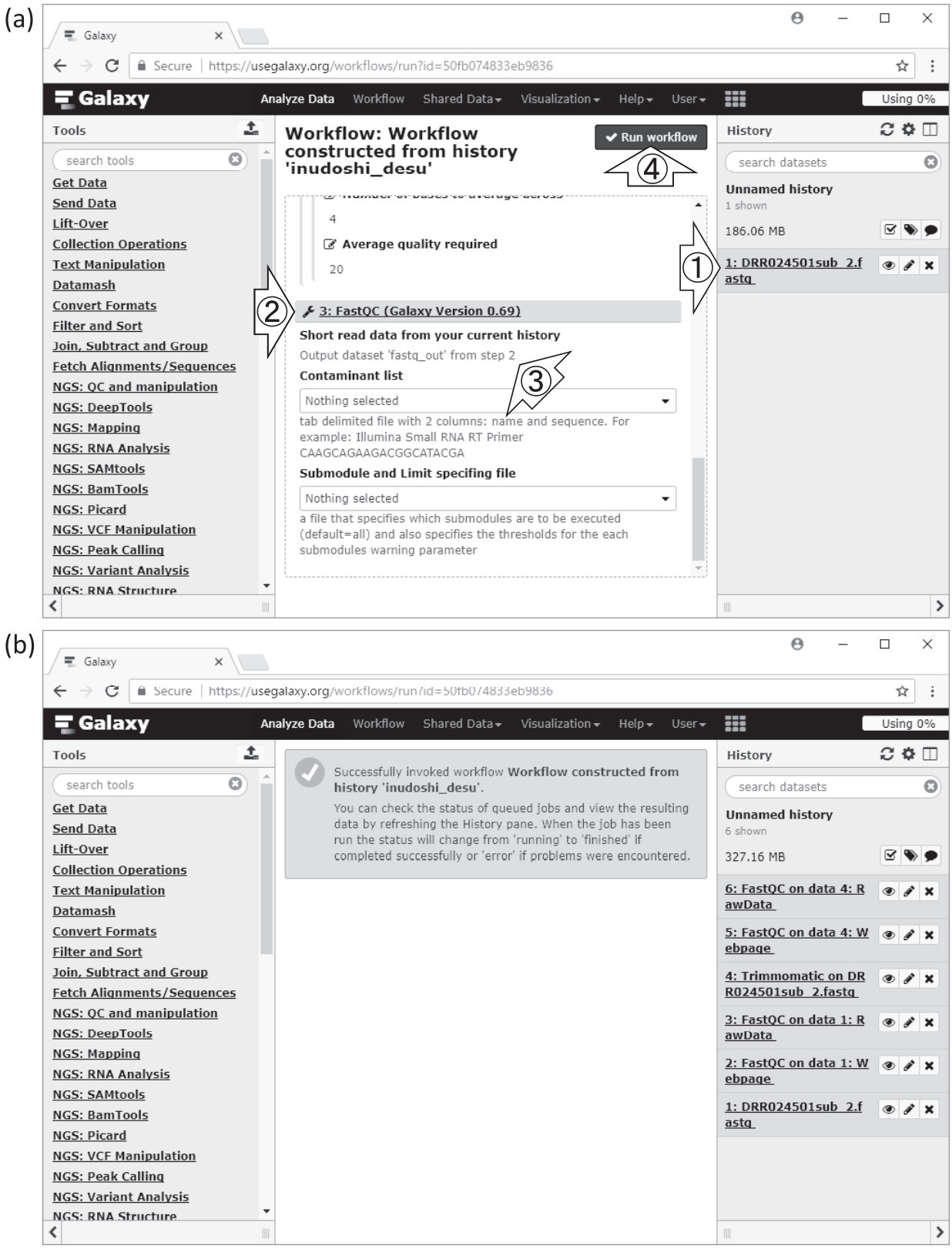

図 5. ワークフローの実行
(a) は W18-7 と、そして (b) は W18-12 と同じ。(a) 新規ヒストリーを作成し (W17-1)、① reverse 側のファイルを Galaxy 上に取り込んで (W17-10)、中央パネル上で② Step3 の入力ファイルが③ Step2 の実行結果となっていることを確認したのち、④ワークフローを実行。(b) 無事終了した状態。

ワークフローの編集画面を眺めると理解しやすいかもしれない［W19-1］。このワークフローの場合、Step2 と Step3 が線で結ばれており、Step2 の出力が Step3 の入力として利用されていることが視覚的にわかる［W19-7］。ワークフロー内で用いられるプログラム内部のオプションについても、ステップごとに変更可能である［W19-8］。また、このワークフローが内部的に FastQC や Trimmomatic を含んでいることがわかるように、任意のキーワードを属性（Attributes）情報としてタグ付けすることもできる［W19-9］。

ワークフローの共有手段は、ヒストリーの共有手段（**図 1b**）と基本的に同じである。ここでは、特に隠す必要もないので Make Workflow Accessible and Publish を選択して公開した［W20-1］。公開完了後は Galaxy Main の Published Workflows というサイトで公開される。当該ワークフローの URL 情報を他の共同研究者に知らせてもよいし、ワークフロー名の一部の文字でキーワード検索してもらってもよいだろう［W20-3］。

最後に、他の Galaxy Main 登録済ユーザへの公開したワークフロー情報の告知、およびワークフロー利用の実例を示す。ここでは、公開済ワークフローと動作確認用 FASTQ ファイルの URL 情報を **terada_registered** にメールし、**terada_registered** がログインするところから示した［W21-2］。Galaxy Main 登録済ユーザの場合は、一般にヒストリーパネルに過去の実行結果が残っている状態（**terada_registered** の場合はヒストリー “Trim”）からスタートする［W21-3］。ワークフロー情報の URL にアクセスし、Import workflow ボタンを押すことでワークフローが利用可能となる［W21-6］。

動作確認用データのダウンロードが完了すれば、下準備は完了である［W22-1］。ここでは新規ヒストリー “inudoshi_desu reverse_data” を作成して実行結果を保存すべく、ヒストリーオプションを変更して実行した（Send results to a new history を Yes にして Run workflow ボタンを押した）。これは、現在ヒストリーパネル上に表示されている “Trim” には保存されないようにしたことを意味する［W22-8］。正常にワークフローを実行し終えたかどうかを確認するためには、View all histories ボタンを押して、作成したヒストリー “inudoshi_desu reverse_data” の実行状況を確認する必要がある［W23-2］。W12 でも示したように、現在のヒストリー（Current History）に切り替えることで、ワークフロー実行結果を眺めることができる［W23-5］。

おわりに

今回は、Galaxy の主な特徴であるヒストリーとワークフローについて述べた。例示した内容は非常に簡単なものであったが、実際の解析現場で行われるものも基本的な枠組みは同じである。共有されたヒストリーの内容については、提供側から詳細な説明を受ければよい。自分がワークフローを構築する側になったとしても、公開されているワークフローや自分が以前作成したワークフローの中から目的に近いものを改変して利用すればよい[6]。また、それが複雑なワークフローの場合でも、step-by-step で組み立てていけばよい。

参 考 文 献

1） Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, et al. (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res **44**: W3-W10.
2） Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114-2120.
3） 谷澤靖洋，神沼英里，中村保一，清水謙多郎，門田幸二（2016）次世代シーケンサーデータの解析手法：第 6 回ゲノムアセンブリ．日本乳酸菌学会誌 **27**: 41-52.
4） 大田達郎，寺田朋子，清水謙多郎，門田幸二（2017）次世代シーケンサーデータの解析手法：第 11 回統合データ解析環境 Galaxy．日本乳酸菌学会誌 **28**: 167-175.
5） Mashima J, Kodama Y, Fujisawa T, Katayama T, Okuda Y, et al. (2017) DNA Data Bank of Japan. Nucleic Acids Res **45**: D25-D31.
6） 孫建強，湯敏，西岡輔，清水謙多郎，門田幸二（2014）次世代シーケンサーデータの解析手法：第 2 回 GUI 環境からコマンドライン環境へ．日本乳酸菌学会誌 **25**: 166-174.

Methods for analyzing next-generation sequencing data XII. Galaxy - Sharing histories and workflows

Tomoko Terada[1], Tazro Ohta[2], Kentaro Shimizu[1, 3], and Koji Kadota[1, 3]

[1] *Graduate School of Agricultural and Life Sciences, The University of Tokyo.*
[2] *Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems.*
[3] *Collaborative Research Institute for Innovative Microbiology, The University of Tokyo.*

Abstract

Galaxy is an integrative data analysis environment run on the web browser which users can use without using Linux command line. The previous article showed an introduction to the Galaxy system and the basic usage of the public Galaxy server "Galaxy Main." In this article, using the last article's results, we present the features to share the analysis results (history) with the other users, or ones to apply the analysis procedures (workflow) to the other data. We also show the useful Galaxy's features such as copying data across histories, or the direct data import from remote servers. We found a compatibility issue on Internet Explorer with the Galaxy system. Thus we recommend using Google Chrome or Firefox to try the procedures we show in this article. Supplementary materials are available online at: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_book_JSLAB.