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11回 統合データ解析環境 Galaxy

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

[1] 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
[2] 東京大学大学院農学生命科学研究科

次世代シーケンサー（以下、NGS）データの解析手段は多様である。本連載ではこれまでキーボード入力をベースとしたコマンドライン環境での解析手段を中心に解説してきたが、マウス操作をベースとしたGUI環境でのNGS解析手段の需要も根強い。第11回は、GUI環境でのデータ解析手段として特に海外で広く普及しているGalaxyを解説する。Galaxyはウェブブラウザを起動して各種解析を行うため、位置づけとしてはウェブツールである。しかしながら、ワークフローやヒストリー管理といった独特の用語およびGUI画面（見栄え）ゆえ、慣れるまでが大変であることもまた事実である。本稿では、Galaxyの概要、公共サーバの基本的な利用法について述べる。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, reproducibility

## はじめに

2014年からスタートしたこれまでの本連載の取組みは、主にBio-Linux[1]のコマンドライン環境でデータ解析を行うための本格的な手段やノウハウを提供するものであった。本稿で紹介するGalaxy[2-5]は、**GUI環境で解析を行うための便利な手段**という位置づけとなる。連載第1回[6]でも述べたように、バイオインフォマティクス中級～上級を目指す場合には、NGSハンズオン講習会の受講資格でもあるLinux環境構築や基本的なLinuxコマンドの習得が最低限の心得であろう。しかし今日では、NGSデータ解析の一般的な手順が確立されつつある。例えばゲノム解析の場合は、クオリティコントロール（以下、QC）、アセンブリ、アノテーション（遺伝子領域や機能予測）といったデータ解析の大枠が存在する。連載第6～9回では[7-10]、QCとして、FastQC[11]でデータの全体像を眺めるクオリティチェックを行ったのち[12]、FaQCs[13]を用いてクオリティスコアの低いリードやアダプター配列の除去を行った。その後、Velvet[14]を用いた*de novo*アセンブリ、DFAST[15]を用いたゲノムアノテーションを行った。この場合は、「FastQC－FaQCs－Velvet－DFAST」というプログラムの組み合わせでゲノム解析を行っていることに相当し、このような作業の流れを**ワークフロー**（Workflow）と呼ぶ。

ワークフローは、用いるプログラムやバージョンの違いによって異なる。例えば、第6回では*de novo*アセンブラの1つであるPlatanus（ver. 1.2.2）[16]を用いるやり方も示した。この場合は、ワークフローを「FastQC－FaQCs－Platanus（ver. 1.2.2）－DFAST」のように記述することができる。ワークフローはまた、データを取得したNGS機

*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

器によっても異なりうる。例えば、上記2つのワークフローは、Illumina MiSeqデータの解析に使用したものであった。第7回で解説したPacBio[17]データの場合は、HGAP[18]（Protocol3; ver. 2.2.0）を用いた*de novo*アセンブリからスタートしたので「HGAP - DFAST」と表現することができる。

Galaxyは、様々なNGS解析（QC、ジェノタイピング、変異解析、モチーフ同定、RNA-seq、ChIP-seq、メタゲノム解析など）に対応している[5]。様々なファイル形式（FASTQ、bam、bedなど）の読み込みや変換も可能である。定番のワークフローが充実しているのはもちろんのこと、ワークフローを構成する既存プログラムのバージョンアップや新しいプログラムのインストールが簡単に行えるToolShed[19]という仕組みも用意されている。GalaxyはLinuxコマンドに不慣れなヒトだけではなく、Linuxコマンドを使いこなせるバイオインフォマティクス中上級者の一部も日常的に利用している。Galaxyを使いこなせれば、バイオインフォマティクス中上級者が日常的に行っているコマンドライン環境と同等の解析が可能といっても過言でない。

## Galaxyプロジェクト

近年、データ解析もできる実験系研究者が増えてきており、かつて明確に役割分担されていた実験のみ行う（wet）研究者とデータ解析のみ行う（dry）研究者の垣根はなくなりつつある。しかしながら、実験の合間にLinuxコマンドを駆使してコマンドライン環境で各種データ解析を行う「dryもできるwetなヒト」はまだまだ少数派であり、マウス操作がベースのGUI環境で必要なデータ解析の一部を行うことのできるwetなヒトが多数派である。Linuxコマンドを覚える暇がない、コマンドライン環境は敷居が高い、というのが主な理由であろう。統合データ解析環境と称されることの多いGalaxyの開発プロジェクトは、当初そのような研究者のためにデータ解析の敷居を下げることを目的として2005年頃にスタートした[2]。

Galaxyプロジェクトは、主にペンシルベニア州立大学（Penn state University）とジョンホプキンス大学（Johns Hopkins University）のメンバーによって開発・維持されている。2005年の最初の論文[2]の引用回数が1,467回、2010年の論文[4]が2,517回（いずれもGoogle Scholar上の数値；2017年5月12日調べ）と世界中で広く利用されている[W1]。Galaxyは、**オープンソース**のアプリケーションソフトウェア（以下、ソフト）である。これは、プログラムのソースコードが公開されており、一定の条件下での使用、複製、改変、再頒布が認められているソフトであることを意味する。お金を払ってライセンスを購入するソフト（シェアウェアと呼ぶ）とは異なり、「開かれた」ソフトであることが特徴である。実際、Galaxyを開発するために雇用されているコア・デベロッパー以外にも、世界中の人々が様々な面で開発をサポートしている。

オープンソースのメリットは、無料で利用できること、開発に参加できることなどが挙げられる。デメリットは、無保証であること、シェアウェアのような手厚いユーザーサポートはないため問題に遭遇しても自力で解決しなければならないことなどが挙げられる。そのため、Galaxyプロジェクトは、開発者だけでなくユーザを交えたコミュニティの運営やミーティングの開催に積極的である。日本でも筆頭著者がco-chairを務めるGalaxy Community Japan（Pitagora-Galaxy Project）が2014年より活動をスタートし、ユーザ間での情報共有や問題解決に協力している。第2回[20]でも紹介した、バイオインフォマティクス全般についての質問投稿サイトBioStar[21]のGalaxy版（https://biostar.usegalaxy.org）も存在する［W2］。これらを利用すれば、他のユーザからの回答によって問題が解決することもある。

## 解析ソフトとしての特徴

一般的なゲノム解析用GUIソフトは、アセンブリやマッピング、ファイル形式の変換、データの可視化など一通りの解析ツールが揃っている。また、データの入力から結果の出力までに複雑な操作を必要としないことなども特徴として挙げられる。Galaxyもまた、基本的な塩基配列の操作やNGS解析に必要な一通りのソフトが揃っている。もちろん全てが標準でインストールされているわけではないため、上述のToolShedという仕組みを用いて、様々なオープンソースのソフトをインストールして利用する。ToolShedは、スマートフォンにおけるアプリケーションストアGoogle Playのようなものである。リストアップされているプログラムの中から自由にインストールすることができる［W3］。具体的なやり方は、次回以降で述べる予定である。

Galaxyの一番の長所は、データ解析の再現性を担保するために必要な仕組みが整っている点である。Galaxyプロジェクトは、当初プログラミングなどのデータ解析スキルのないヒトが簡単に使えることを目指して開発されていた。しかし最近では、**一度行った解析を繰り返し実行することができ、それを共有し、別のユーザが再現できる**機能の提供も重視している[5]。NGSデータ解析では、「FastQC - FaQCs - Velvet - DFAST」のような特定のワークフローを繰り返し実行することが多い。ワークフロー中の各ツールを毎回手動で実行すると効率が悪いため、一度定めたワークフローを自動的に実行できるようにしておくのが基本である。同じワークフローを別のデータに対して実行したり、ワークフローを構成するプログラム内部のパラメータ（またはオプション）を変えて結果の違いを調べたりするなどの作業が容易になるからである。

プログラムの入力と出力を繋いだワークフローは、コマンドライン環境の場合、第4回[12]および第8回[9]で述べた**シェルスクリプト**でも実現可能である。しかしながら、正しく動作するシェルスクリプトの記述や保管、実行履歴の管理や再実行は、中上級の多くのdry研究者にとっても面倒で煩雑な作業である。特に、共同研究などで他の研究者とデータや解析結果およびワークフローを適切に管理・共有する場合は、時間的・技術的・精神的な面で大きな負担がかかる。Galaxyは、ワークフローの構築、保存、履歴管理、再実行、そしてそれらの共有を容易に行えるという点で非常に優れている。これらについては、次回以降で述べる予定である。

## Galaxyの動作原理

ここでは、Galaxyの動作原理について述べる。ソフトにも様々な種類があるが、ここでは実際のプログラムがどこで動いているかに着目して整理する。例えば、本連載のウェブサイトを眺める際に用いるウェブブラウザはソフトであり、具体的にはGoogle Chrome、Firefox、Safari、Microsoft Edgeなどが相当する。そしてこれらのソフト自体は、読者自身のPCの中（開発者的な感覚ではPC上）で動いている。第2回[20]で解説したWindowsのコマンドプロンプト（最近ではWindows PowerShellやBash on Ubuntu on Windowsも利用可能）やMac/Linuxのターミナルで実行されるプログラムについても同様である。仮想化ソフトVirtualBoxを起動してBio-Linuxを立ち上げ、その中でFastQCを実行する場合も同じである。仮想環境自体が自身のPC内に構築したものであるため、FastQCというプログラムは手元にある自身のPC上で実行されている。プログラム実行に要する時間は、当然ながら実行した場所の性能（CPU、メモリ、そしてディスク容量）によって左右される。つまり、自身のPCおよび仮想環境（Bio-Linux）の性能に依存する。Galaxy利用時もまた、プログラムがどこで実行されるかを正しく認識することが重要である。

第7回[8]では、PacBioデータの*de novo*アセンブリ実行には数百GB程度のメモリが必要であることを述べた。このときの主な作業は、自身のPC上でウェブブラウザを起動し、DDBJ Pipeline[22]上でHGAP[18]の実行命令を出すことであった。HGAPプログラム自体は、静岡県三島市にある国立遺伝学研究所のスーパーコンピュータシステム上で実行されている。**Galaxyは自身のPC上でウェブブラウザを通して実行命令を出すソフト**であり、ウェブブラウザを通して眺めているその**Galaxyが実在する場所で命令したプログラムが実行される**。例えば、第1回[6]で述べたDBCLS Galaxyを利用する場合には、DBCLSの計算サーバ上でプログラムが実行される。ここまでは、DDBJ Pipelineを利用した経験があれば容易に理解できるであろう。

多くのGalaxy初心者にとって難解な事柄は、ウェブブラウザを通して眺めているそのGalaxyが実在する「どこかの場所」が**自身のPC以外のコンピュータである必要はない（自身のPCでもよい）**という点であろう。具体的には、自身のPC上に実在するGalaxyのプログラムを自身のウェブブラウザを通して実行命令を出して実行させることもできる、ということである。「自身のPC以外の場所（ウェブサイト）」を訪れるのが一般的なウェブブラウザの利用法であるが、手元にあるhtmlファイルをダブルクリックすれば中身を閲覧可能であることと似たようなものだと思えばよい。もちろん、自身のPCの性能を超えたプログラムを実行させることはできない。例えば、メモリ8GB程度のPC上で、数百GB程度のメモリを要するHGAPプログラムを実行することは実質的に不可能である。このように、外部のGalaxyサーバを利用する動機付けとしては、*de novo*アセンブリなど比較的大きなメモリを要するプログラムの利用が挙げられる。もちろんDDBJ Pipelineでも目的のプログラムが利用可能な場合は、基本的にどちらでもよい。

Galaxyの動作原理を簡単にまとめると以下の通りとなる：① Galaxyに対する操作は自身のPC上のウェブブラウザを介して行う、② Galaxyが実在する場所自体はどこでもよい（自身のPC上でもよいし外部のGalaxyサーバでもよい）が、③どこでGalaxyを動かしているのかは正しく把握しておかねばならない。自身のPCにGalaxyをインストールして利用するメリットは、管理者としてGalaxyを操作できるので自由に解析ツールを追加可能であること、（予算が許す限り）自由に計算機資源を使えることが挙げられる。デメリットは、インストールの手間がかかること、遭遇するトラブルは自分で解決せねばならないことが挙げられる。外部の公共Galaxyサーバ（Public Galaxy Server）を利用するメリットは、Galaxyのインストールやトラブルシューティングの手間がかからないことなどが挙げられる。デメリットは、多くの場合利用可能なデータ量の制限（quota）が設けられていること、自分で新しく解析ツールをインストールする権限が与えられていないため管理者にその都度依頼する必要があることなどである。

## Public Galaxy Serverの基本的な利用法

Galaxy初心者は、手始めにGalaxy Projectが運用するPublic Galaxy Server（https://usegalaxy.org；通称“Galaxy main”）を利用してみるとよい。ユーザ数の多いGalaxy mainには、解析ツールが豊富に揃っている（インストール済みである）からである。アカウントの作成は、Galaxy mainの画面から“Register”を選択して、メールアドレス、パスワード、ユーザ名（Public name）などを入力し、

“Submit” をクリックすればよい［W4-3］。登録したメールアドレス宛に、“Galaxy Account Activation” というタイトルのメールが届くので、本文中にある URL をクリックすればアカウント作成の完了である［W4-8］。Galaxy main サーバでは、利用可能なデータ量が 1 アカウントにつき 250GB まで、同時に実行できるジョブが 6 つまでという制限がある［W5-1］。他にも、1 人で複数のアカウントを作成してはならない（作成できるアカウントは 1 人 1 つ）というルールがある。

Galaxy main の基本操作画面は、縦に 3 分割された構成となっている（図 1；W5-2）。左側がツール選択パネル、中央が操作や結果表示パネル、そして右側が実行されたジョブと履歴のパネルである。ここでは、Galaxy の操作に慣れることを目的として、Bio-Linux 環境で行った Illumina MiSeq データのクオリティチェック[7] を Galaxy main で行う。具体的には、30 万リードからなる gzip 圧縮 FASTQ ファイル（DRR024501sub_1.fastq.gz［W5-3］）のアップロードと FastQC を実行する。Galaxy main サーバへのアップロード作業はドラッグ & ドロップを基本としているため、DDBJ Pipeline 利用時の FTP 経由のアップロード（第 6 回の W13）よりも簡単である。

図 2 は、アップロード後の Galaxy main 画面である［W6-1］。約 56MB からなる gzip 圧縮 FASTQ ファイルをアップロードしたが、右側の履歴パネルの表示は① 186.06MB からなる② DRR024501sub_1.fastq になっていることがわかる。このことから、Galaxy にはアップロー

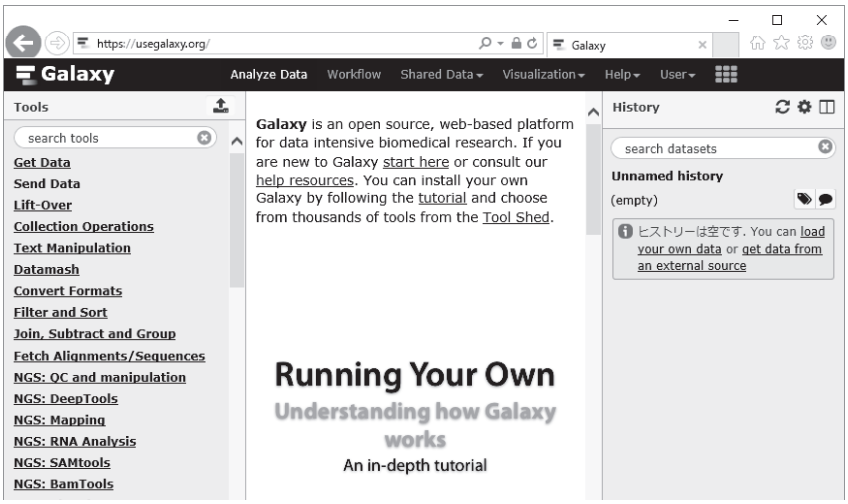


図 1. Galaxy main の基本操作画面
縦に 3 分割された構成となっており、左がツール選択パネル、中央が操作や結果表示パネル、そして右側がヒストリーパネルである［W5-2］。

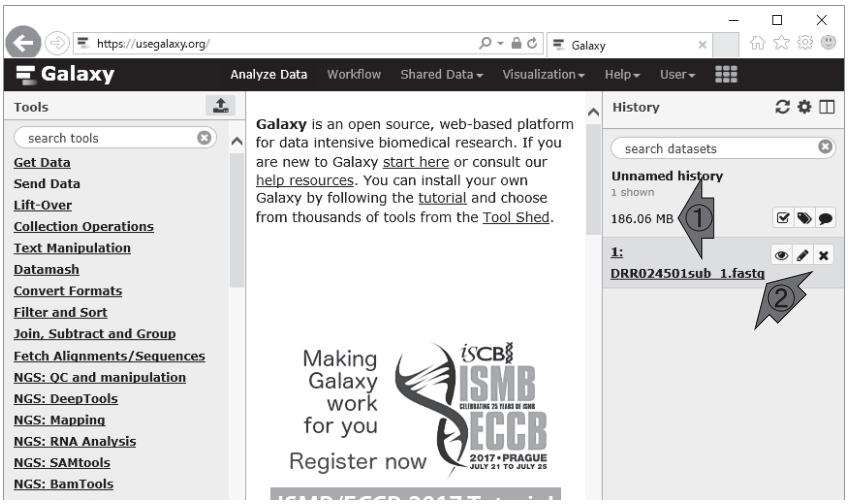


図 2. FASTQ ファイルのアップロード
gzip 圧縮した 56.1MB の DRR024501sub_1.fastq.gz ファイルをアップロードしたあとの状態。①アップロード後は 186.06MB となっており、② gzip 非圧縮状態になっていることがわかる［W6-1］。

ドしたファイルが標準的な圧縮形式であれば自動で展開する機能がついていることがわかる。ファイルの中身の表示［W6-2］、ファイル名の変更、自動認識されたファイル形式の確認や修正［W6-3］などの操作も独特である。右側パネルのヒストリーからの削除［W6-4］がディスク（Galaxy main サーバ）からのデータセットの削除［W7-7］とイコールではない点や、削除されたという情報がいつまでも右側のヒストリーパネルに表示されたままになっている点［W7-8］も最初は戸惑うかもしれないが、全ては慣れである。

## クオリティチェック（FastQC）

アップロードした FASTQ ファイルに対して、FastQC によるクオリティチェックを行う［W9］。Galaxy の基本操作画面上で、左側のツール選択パネルから① **NGS：QC and manipulation** をクリックし目的の② **FastQC** を選ぶと、中央パネルに FastQC の操作画面が現れる（**図 3**）。③入力ファイルが④アップロードしたものと同じになっていることを確認して、⑤実行（Execute）ボタンを押す［W9-8］。但し、ボタンを押せば直ちに実行されるわけではない［W9-9］。DDBJ Pipeline 利用時と同じく、公共サーバを多くのヒトが利用しているからである。Galaxy の場合は、右側のヒストリーパネルが**実行待ち状態のときは灰色、実行中は黄色**、実行が**無事終了すると緑色、失敗すると赤色**で表される［W9-10］。

ヒストリーパネル上で緑色に変わった FastQC 実行結果（Webpage と書いてあるほう）の目玉アイコンをクリックすれば、第 6 回の W4-2 で眺めたものと同じような html レポートが中央パネルに表示される［W10-2］。ヒストリーパネル上でフロッピーディスクアイコンをクリックすることで、html レポートファイルをダウンロードすることができる［W10-4］。Galaxy の中央パネル上で眺めるよりも、ダウンロードした html ファイルをダブルクリックしてローカル環境で眺めるほうが全体像を把握しやすいだろう。ここでは、Galaxy 上で実行した FastQC（ver. 0.11.5）のクオリティスコア分布［W10-6］が、Bio-Linux 上で実行した FastQC（ver. 0.11.4）の結果［W10-7］と酷似していることを確認した。

## ファイルの型（Trimmomatic を例に）

この MiSeq データにはアダプター配列が含まれており［W11-1］、第 6 回は FaQCs [23] を用いてアダプター除去を行った。Galaxy main では FaQCs を選択できないため、ここでは Trimmomatic [24] を利用してアダプター除去を行う［W11-2］。しかしながら、大抵の Galaxy 初心者は Trimmomatic を選択したところまでで行き詰まるであろう［W11-3］。FastQC 実行時にはヒストリーパネルの DRR024501sub_1.fastq ファイルが中央パネル上で見られていたが［W9-3］、Trimmomatic 実行時にはそれを指定できないからである。

Galaxy は、使用するプログラムによって同じ FASTQ 形式でもより詳細に型（エンコーディングの方式）の指定を行う必要がある。2010 年頃から NGS 解析に携わっているヒトであればある程度理解できる部分だとは思われるが、FASTQ 形式にもいくつかの方式が存在するからである [25]。例えば、FastQC 実行結果の Basic Statistics の項目において、Encoding が “Sanger / Illumina 1.9” と表記されている。これは、FastQC プログラムが

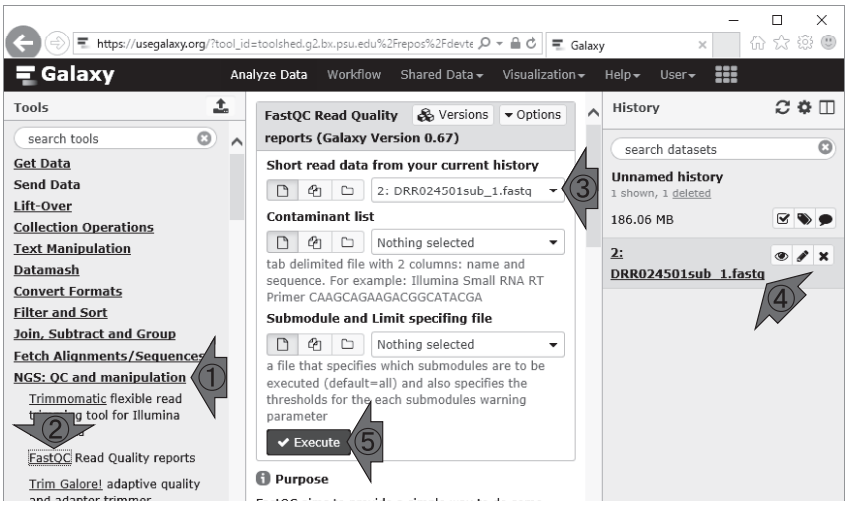


図 3. FastQC の実行
① NGS：QC and manipulation で見られる② FastQC を選択すると、中央パネル上で FastQC の操作画面が見られる［W9-3］。③入力ファイルのところに、④アップロードしたファイルが見られることを確認して、⑤ Execute ボタンを押せば FastQC が実行される［W9-8］。

DRR024501sub_1.fastq を “Sanger / Illumina 1.9” だと自動認識したことを意味する［W10-3］。

Galaxy main 上で Trimmomatic を実行する場合は、まず中央パネル上の Input FASTQ file のところを眺め、入力ファイルを fastqsanger という型に変更する必要性を認識する［W11-3］。そして、ヒストリーパネル上の入力ファイルを編集して、Datatype をデフォルトの fastq から fastqsanger に変更する［W11-7］（図 4）。こうすることで、Trimmomatic の Input FASTQ file 上で DRR024501sub_1.fastq が認識されるようになる［W11-9］。型変換の問題は、Trimmomatic 実行時に限った話ではない。プログラムごとに受け付けるファイルの型は、プログラムを Galaxy に登録するヒトによって決められる。厳密に型指定を行うことによって、想定外のバグを防ぐというのが Galaxy の方針なのであろう。

## アダプタートリミング（Trimmomatic）

Trimmomatic は、アダプター配列除去に特化したプログラムではない。このため、Perform initial ILLUMINACLIP step? で Yes を選択し［W12-1］、入力ファイルに合わせたアダプター配列（この場合は Illumina MiSeq 用）を選択し

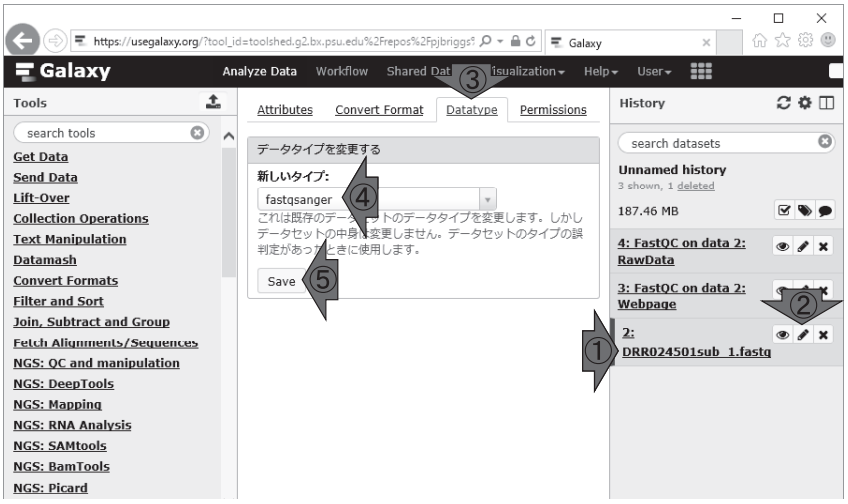


図 4. 入力ファイルの型変換
Galaxy は、プログラムごとに受け付けるファイルの型が決められている。Trimmomatic の場合は、同じ FASTQ ファイルでも fastq ではなく fastqsanger にしなければならない［W11-3］。このため、Trimmomatic 実行前に、① DRR024501sub_1.fastq を②編集して、③ Datatype タブ上で④ fastqsanger に変更して⑤ Save しておく必要がある［W11-7］。

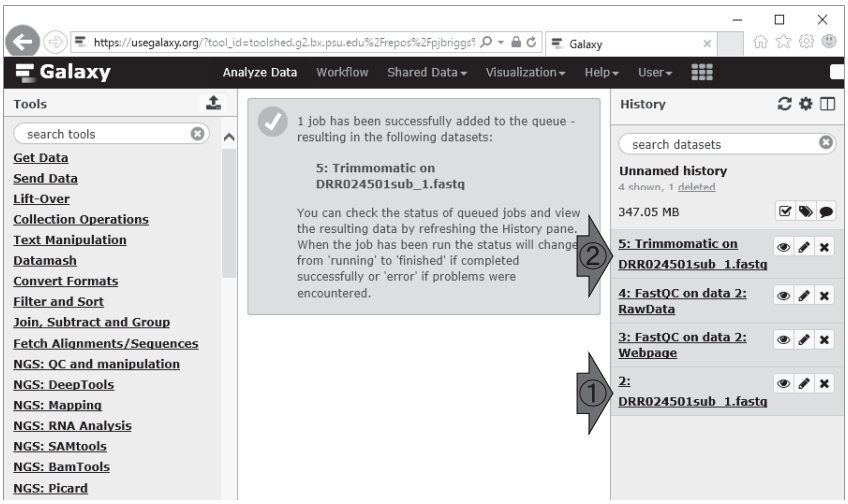


図 5. Trimmomatic 実行終了後の状態
Trimmomatic の入力と出力。ヒストリーパネル上で②「5: Trimmomatic on DRR024501sub_1.fastq」となっているので、①を入力として Trimmomatic を実行した結果が②であることがよくわかる［W12-8］。

て実行する［W12-5］。引用文献に関する情報は中央パネルの最下部で見られるので、指定された文献を正しく引用してほしい［W12-4］。我々は、得られた入力と出力のファイルサイズの関係（186MB［W6-1］と 159MB［W12-9］）から、アダプター除去およびデフォルト設定の各種フィルタリングがそれなりに正しく動作したのだろうと判断した（**図 5**）。

Trimmomatic のアダプター除去精度を調べるためには、Trimmomatic 実行後のデータに対して再度 FastQC を実行し、その結果を眺めればよい。ヒストリーパネル上には、Trimmomatic 実行前後の 2 つの FASTQ ファイルが存在する［W12-9］。Galaxy のプログラムは、そのプログラムが受け付けるファイルの型に合致したもののうち、最後に作成したものを入力の第一候補とする。FastQC の場合は Trimmomatic 実行後の FASTQ ファイルが入力の第一候補となるため、そうなっていることを確認して実行ボタンを押すだけでよい［W13-2］。

**図 6** は、FastQC 実行後の Galaxy main 画面である［W13-4］。Trimmomatic 実行後のリード数は 297,724 個であり［W13-6］、FaQCs 実行後のリード数（297,633 個；第 6 回の W5-2 および W6-2）よりも若干多かった。

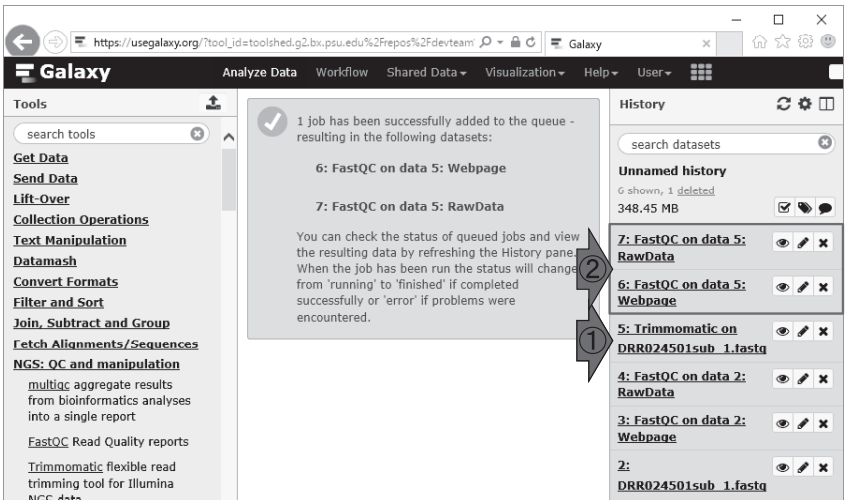


図 6. FastQC 実行終了後の状態
①入力は Trimmomatic 実行後のデータ。②の枠内が FastQC 実行結果。「6: FastQC on data 5…」などとなっていることから、①「5: Trimmomatic on …」のデータを入力としたことがわかる［W13-4］。

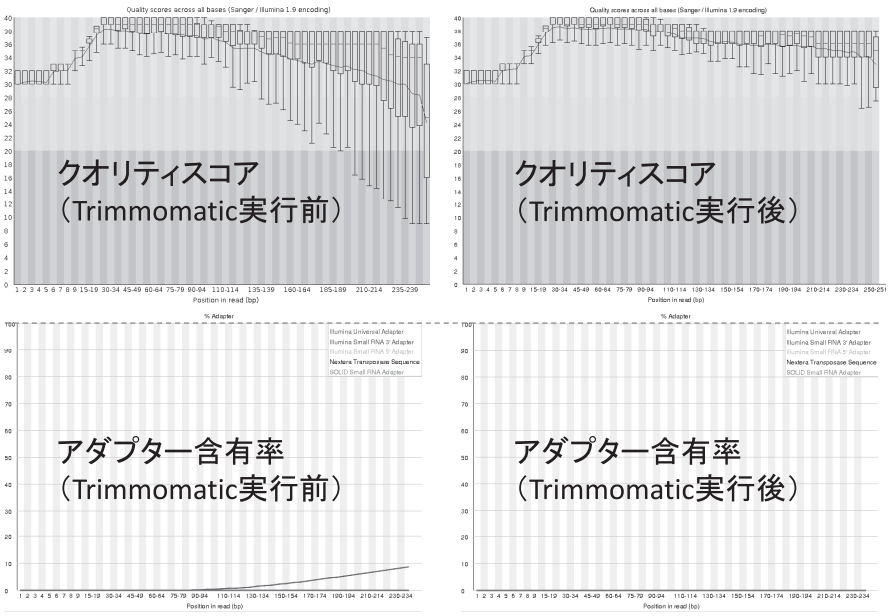


図 7. Trimmomatic 実行前後のクオリティスコアおよびアダプター含有率
左側の Trimmomatic 実行前は、第 6 回の図 1 の左側と同じ。Trimmomatic 実行後（右側）のクオリティスコアは明らかに上昇しており、アダプターはほぼ完全に除去できていることがわかる。

Trimmomatic 実行後のクオリティスコア分布（**図 7** の右上）は、実行前（**図 7** の左上）に比べて明らかに上昇している。FaQCs 実行後のクオリティスコア分布およびアダプター除去精度（第 6 回の**図 2** の左側）と比べると、全体的に Trimmomatic のほうが優れていることがわかる。

## おわりに

NGS 解析の多くは Galaxy 上で行うことができる。科学技術振興機構 バイオサイエンスデータベースセンター（JST-NBDC）が主導する **NGS ハンズオン講習会**の内容は、Linux 環境での実習が中心である［W14-1］。Galaxy がこの講習会に含まれない理由は、単純に受講人数規模の点で難しいからである[6]。Linux コマンドを覚える必要のない（Linux-free な）NGS 解析の世界も存在する。Linux コマンドの壁を乗り越えられず NGS 解析を諦めていた読者が、本稿をきっかけに Galaxy による NGS 解析の世界に一人でも多く本格参入し、効率的な研究の推進に貢献できれば幸いである。

## 謝　辞

本連載の一部は、JST-NBDC との共同研究の成果によるものです。また、JSPS 科研費 JP15K06919 の助成を受けたものです。

## 参　考　文　献

1) Field D, Tiwari B, Booth T, Houten S, Swan D, et al. (2006) Open software for biologists: from famine to feast. Nat Biotechnol **24**: 801-803.
2) Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res **15**: 1451-1455.
3) Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, et al. (2010) Manipulation of FASTQ data with Galaxy. Bioinformatics **26**: 1783-1785.
4) Goecks J, Nekrutenko A, Taylor J; Galaxy Team. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol **11**: 128.
5) 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.
6) 門田幸二，孫建強，湯 敏，西岡 輔，清水謙多郎（2014）次世代シーケンサーデータの解析手法：第 1 回イントロダクション．日本乳酸菌学会誌 **25**: 87-94.
7) 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.
8) 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（2016）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**: 101-110.
9) 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（2016）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**: 187-195.
10) 谷澤靖洋，真島 淳，藤澤貴智，李 慶範，中村保一，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 9 回ゲノムアノテーションとその可視化、DDBJ への登録．日本乳酸菌学会誌 **28**: 3-11.
11) Andrews S. (2015) FastQC a quality control tool for high throughput sequence data, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
12) 孫 建強，湯 敏，清水謙多郎，門田幸二（2015）次世代シーケンサーデータの解析手法：第 4 回クオリティコントロールとプログラムのインストール．日本乳酸菌学会誌 **26**: 124-132.
13) Lo CC, Chain PS. (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinformatics **15**: 366.
14) Zerbino DR, Birney E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18**: 821-829.
15) Tanizawa Y, Fujisawa T, Kaminuma E, Nakamura Y., Arita M. (2016) DFAST and DAGA: Web-based integrated genome annotation tools and resources. Biosci Microbiota Food Health **35**: 173-184.
16) Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, et al. (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res **24**: 1384-1395.
17) Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, et al. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods **7**: 461-465.
18) Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods **10**: 563-569.
19) Blankenberg D, Von Kuster G, Bouvier E, Baker D, Afgan E, et al. (2012) Dissemination of scientific software with Galaxy ToolShed. Genome Biol **15**: 403.
20) 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.
21) Parnell LD, Lindenbaum P, Shameer K, Dall'Olio GM, Swan DC, et al. (2011) BioStar: an online question & answer resource for the bioinformatics community. PLoS Comput Biol. **7**: e1002216.
22) Nagasaki H, Mochizuki T, Kodama Y, Saruhashi S, Morizaki S, et al. (2013) DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. DNA Res **20**: 383-390.
23) Lo CC, Chain PS. (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinformatics **15**: 366.
24) Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114-2120.
25) Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res **38**: 1767-1771.

# Methods for analyzing next-generation sequencing data XI. Galaxy -an integrated data analysis environment

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

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

There are a variety of methods to analyze Next-Generation Sequencing (NGS) data. In this series, we have introduced data analysis methods based on the Command Line Interface (CLI) which uses keyboard input. However, there are still many users demanding an analysis method using the mouse input. Here, we introduce the Galaxy, an integrated data analysis environment which is one of the most popular data analysis methods. Galaxy can be categorized as a web tool as it is used via a web browser. However, it is not straightforward to get used to Galaxy, because of its looking and terms, for example, workflows or history management. This paper shows the overview of Galaxy and the basic usage of the public server. Supplementary materials are available at our web site, http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_book_JSLAB.