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 ウェブ資料

大田達郎*、寺田朋子、清水謙多郎、門田 幸二*

W1：Google Scholar

①Galaxyの2005年の論文の引用回数は、②1,467回

①

Giardine et al., *Genome Res.*, **15**: 1451–1455, 2005

W1：Google Scholar

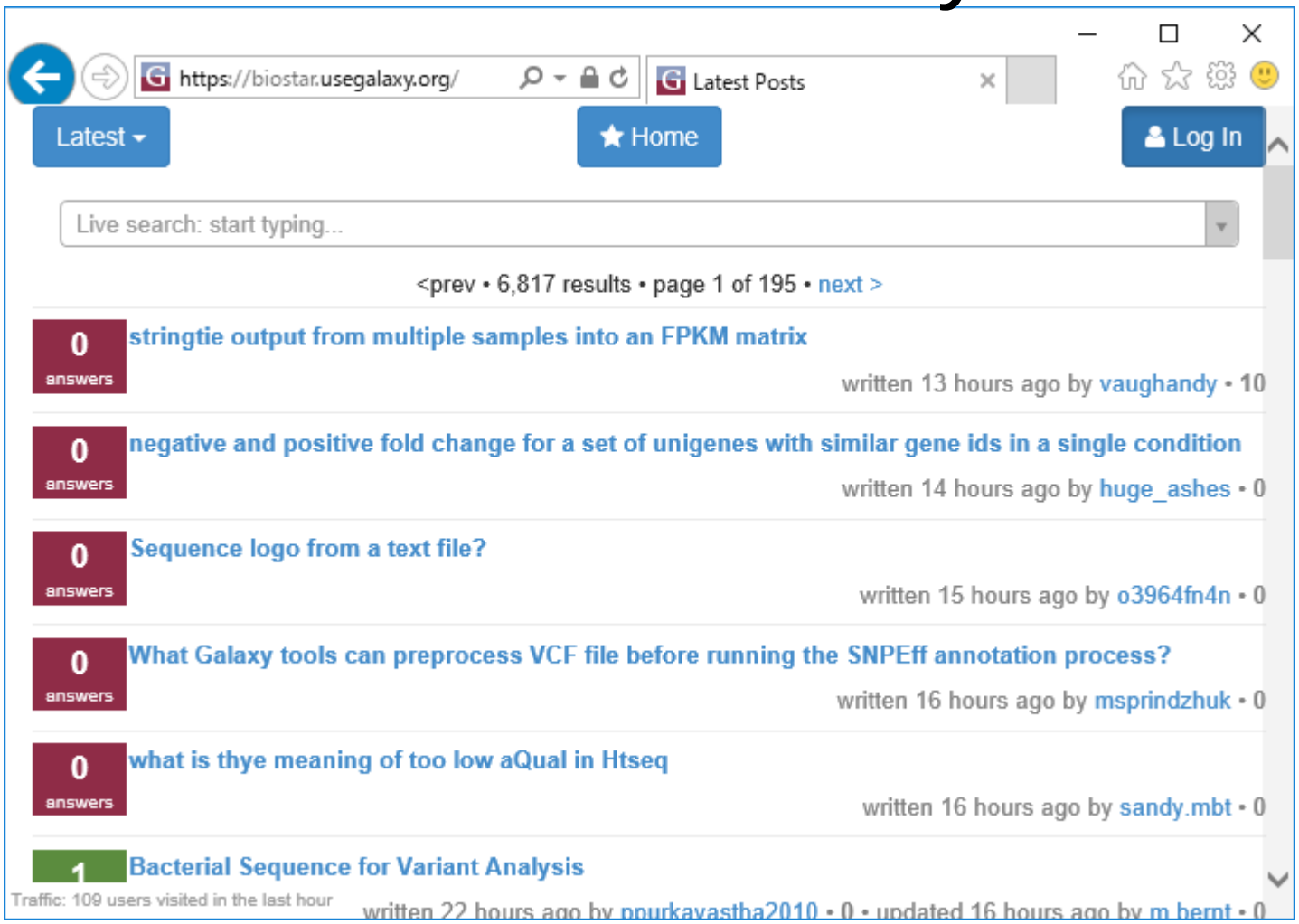
①Galaxyの2010年の論文の引用回数は、②2,517回

①

Goecks et al., *Genome Biol.*, **11**: 128, 2010

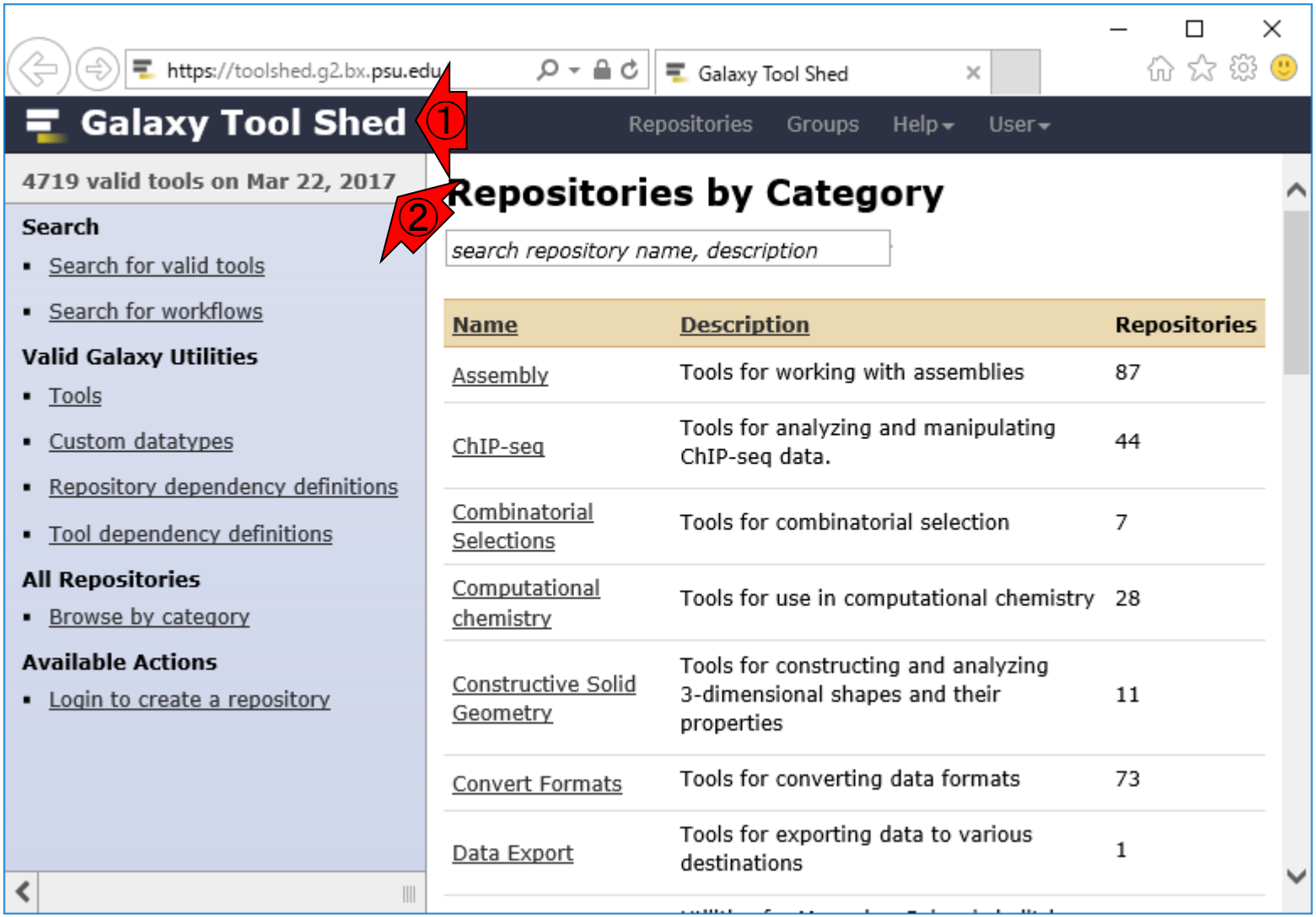
W2：BioStarのGalaxy版

ライフサイエンスQAを知っているヒトには馴染みのある見た目ですね



W3-1：ToolShed

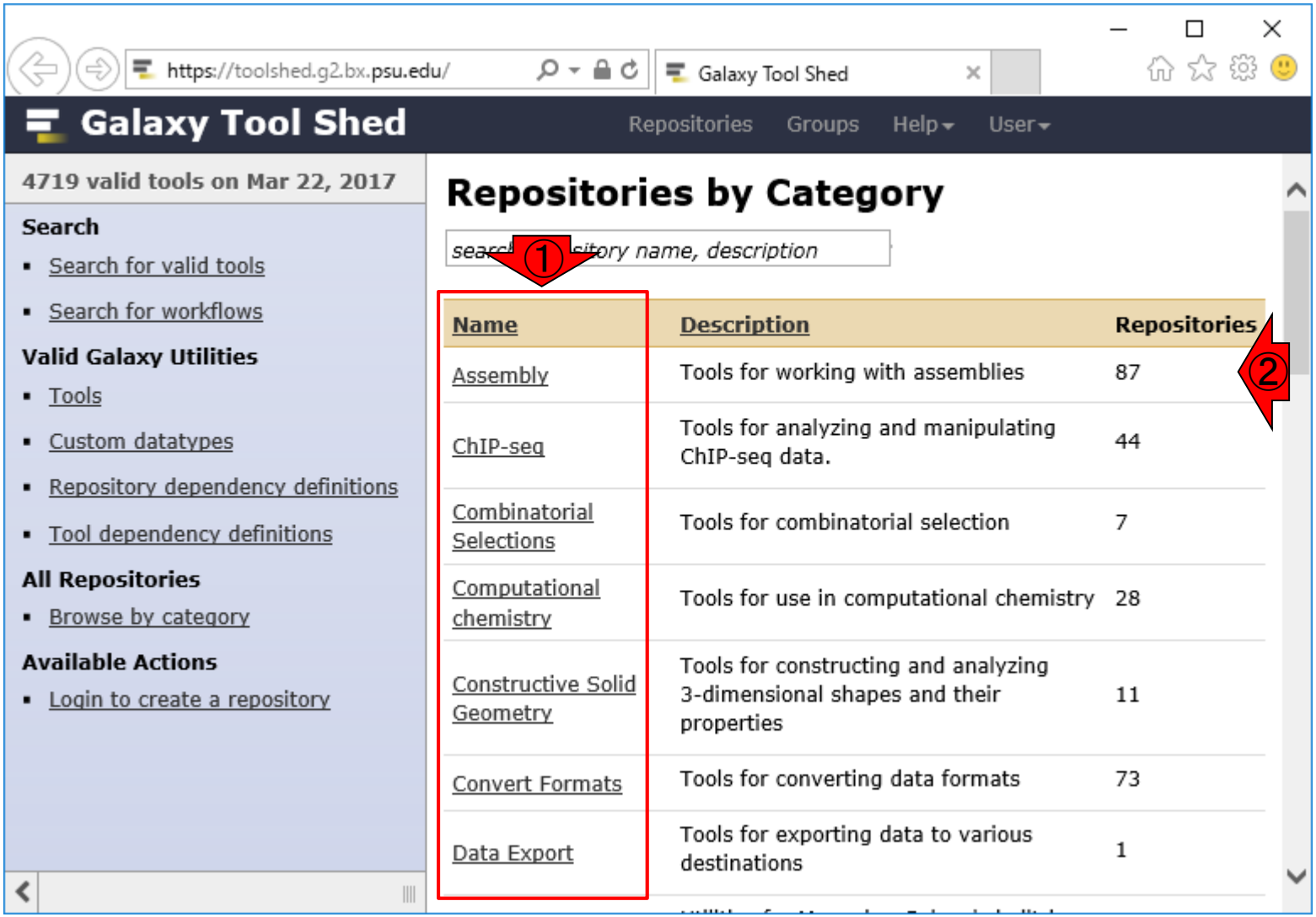
①ToolShedのトップ画面。②2017年3月22日現在、4,719個のツールを利用可能



Blankenberg et al., *Genome Biol.*, **15**: 403, 2012

W3-1：ToolShed

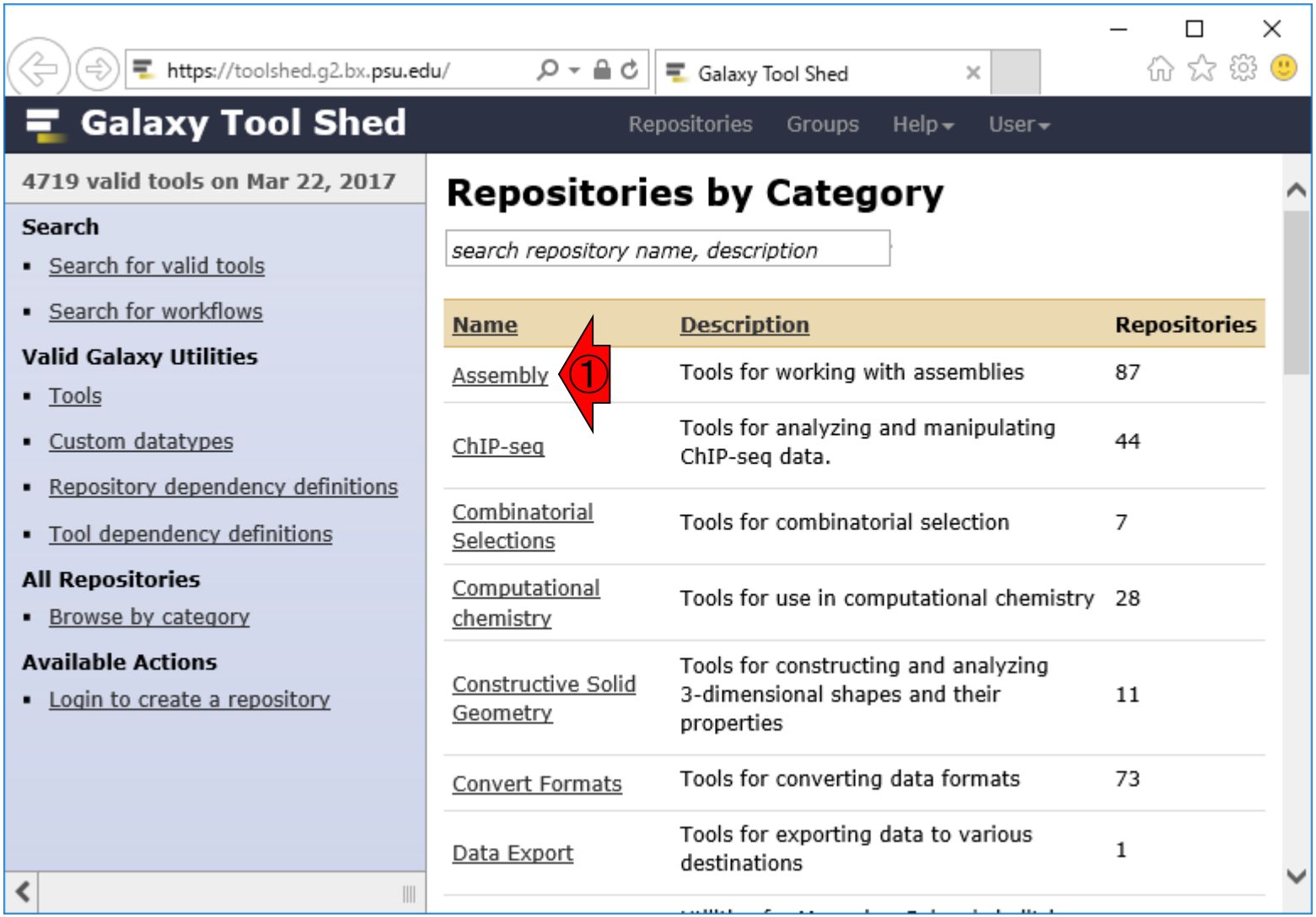
①カテゴリーごとに分かれているようだ。たとえば②Assemblyプログラムは87個あるのだろう



Blankenberg et al., *Genome Biol.*, **15**: 403, 2012

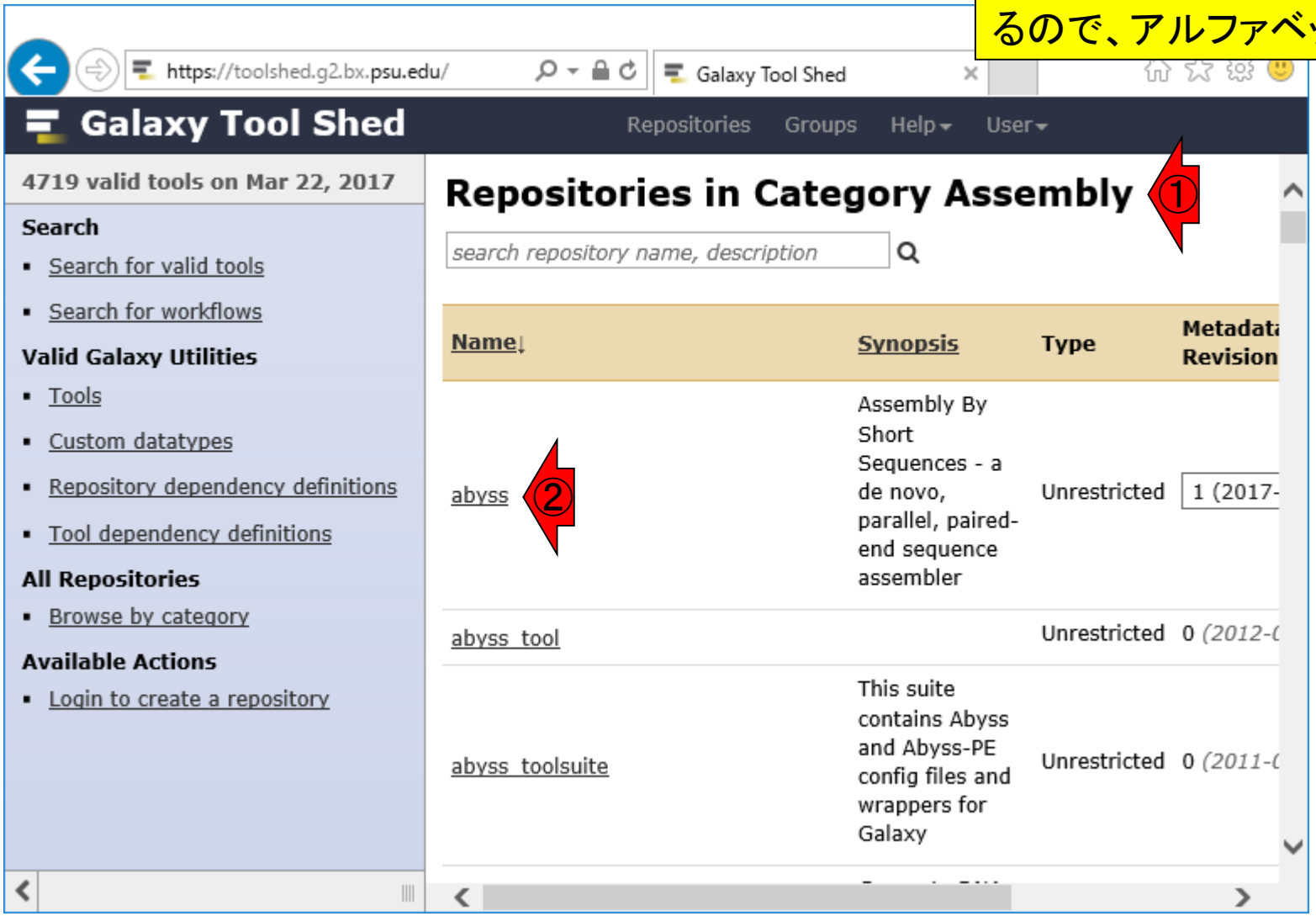
W3-1：ToolShed

①Assemblyをクリック。反応がない、、、が10秒くらい待つとページが切り替わる



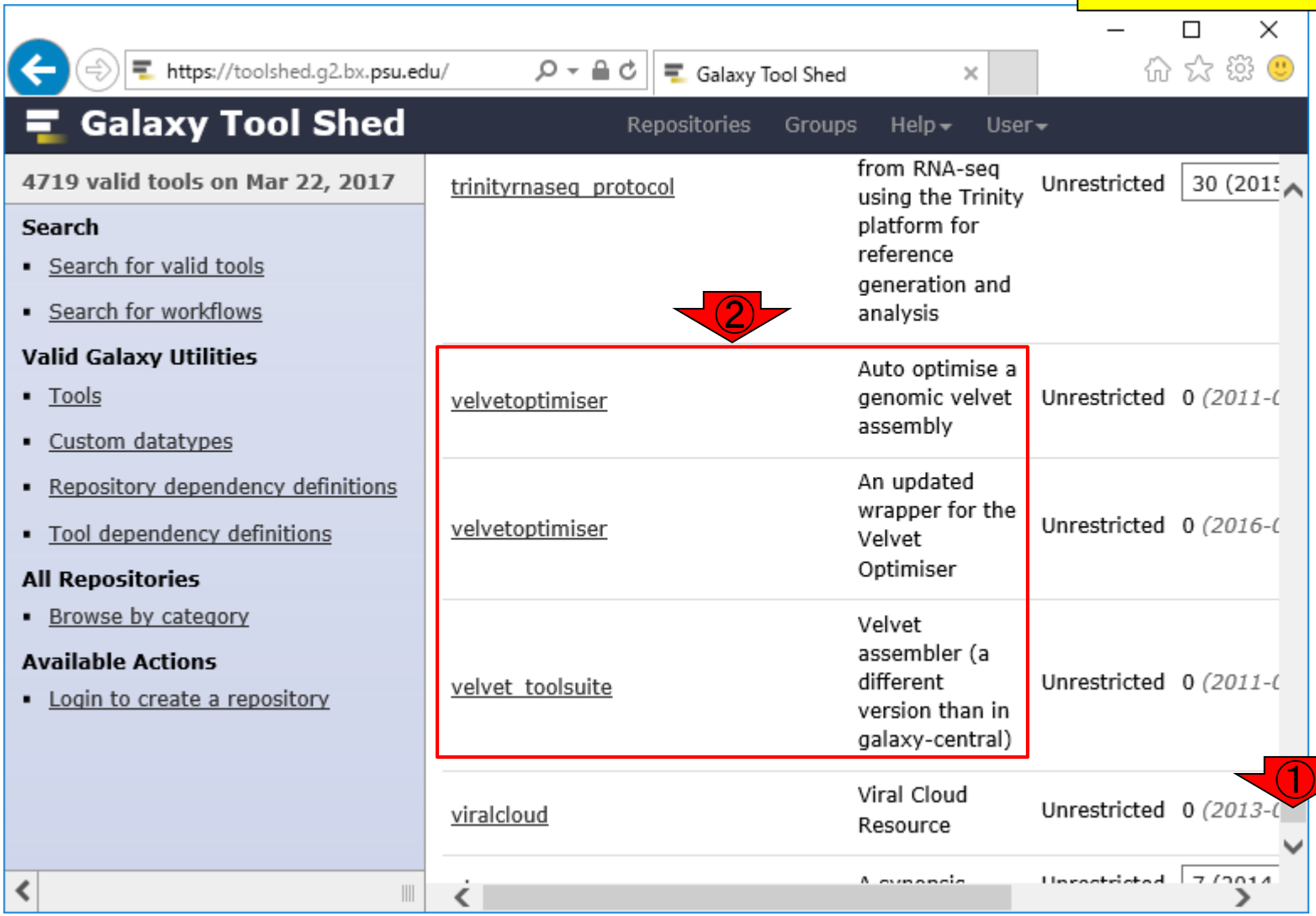
W3-1：ToolShed

①Assemblyのカテゴリーに属する87個のプログラムがリストアップされているのだろう。
②abyssというアセンブラから表示されているので、アルファベット順なのだろうと妄想



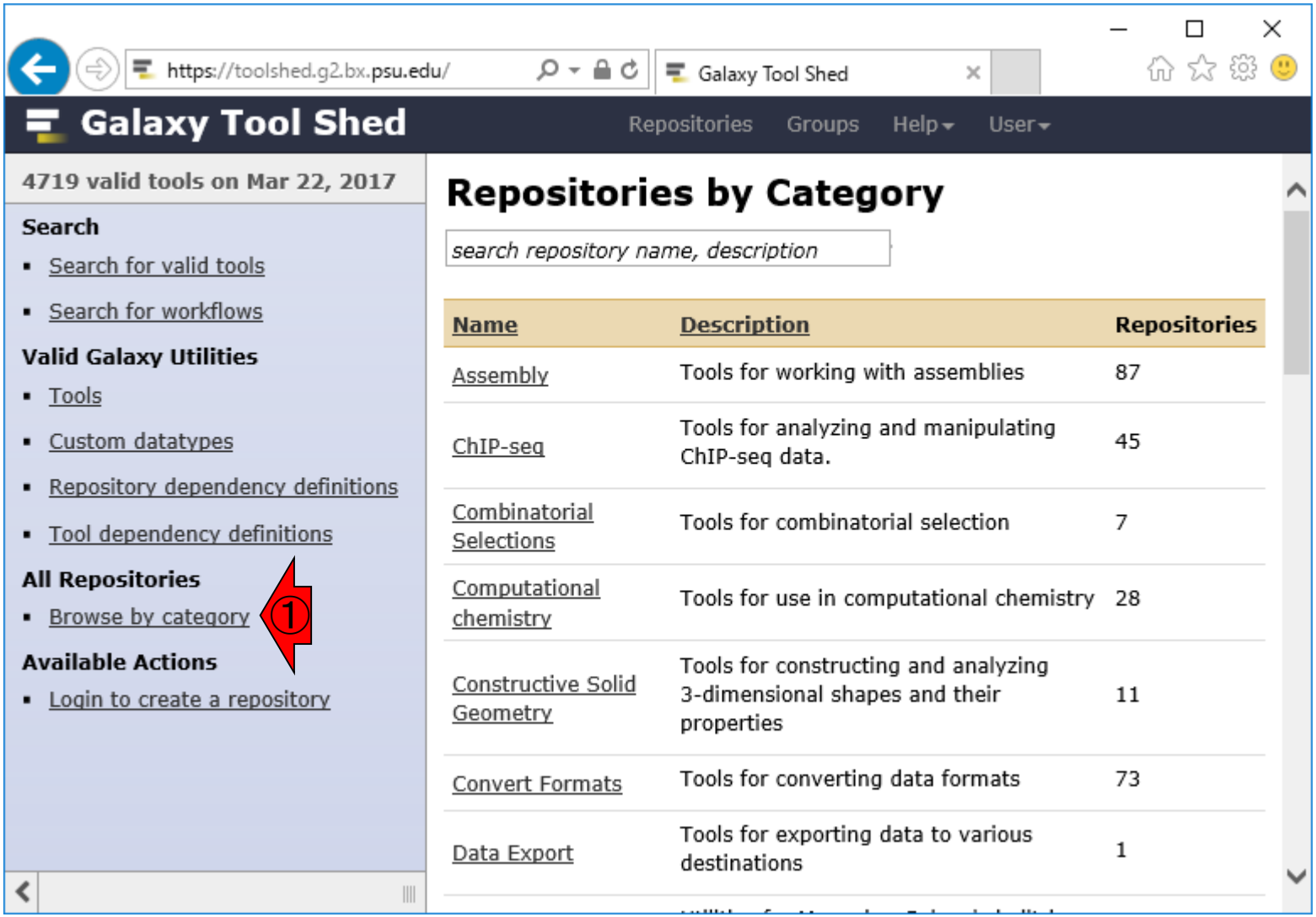
W3-1：ToolShed

Velvetアセンブラがあるかなと妄想しながら①ページ下部に移動。確かに②velvet関連プログラムが存在する



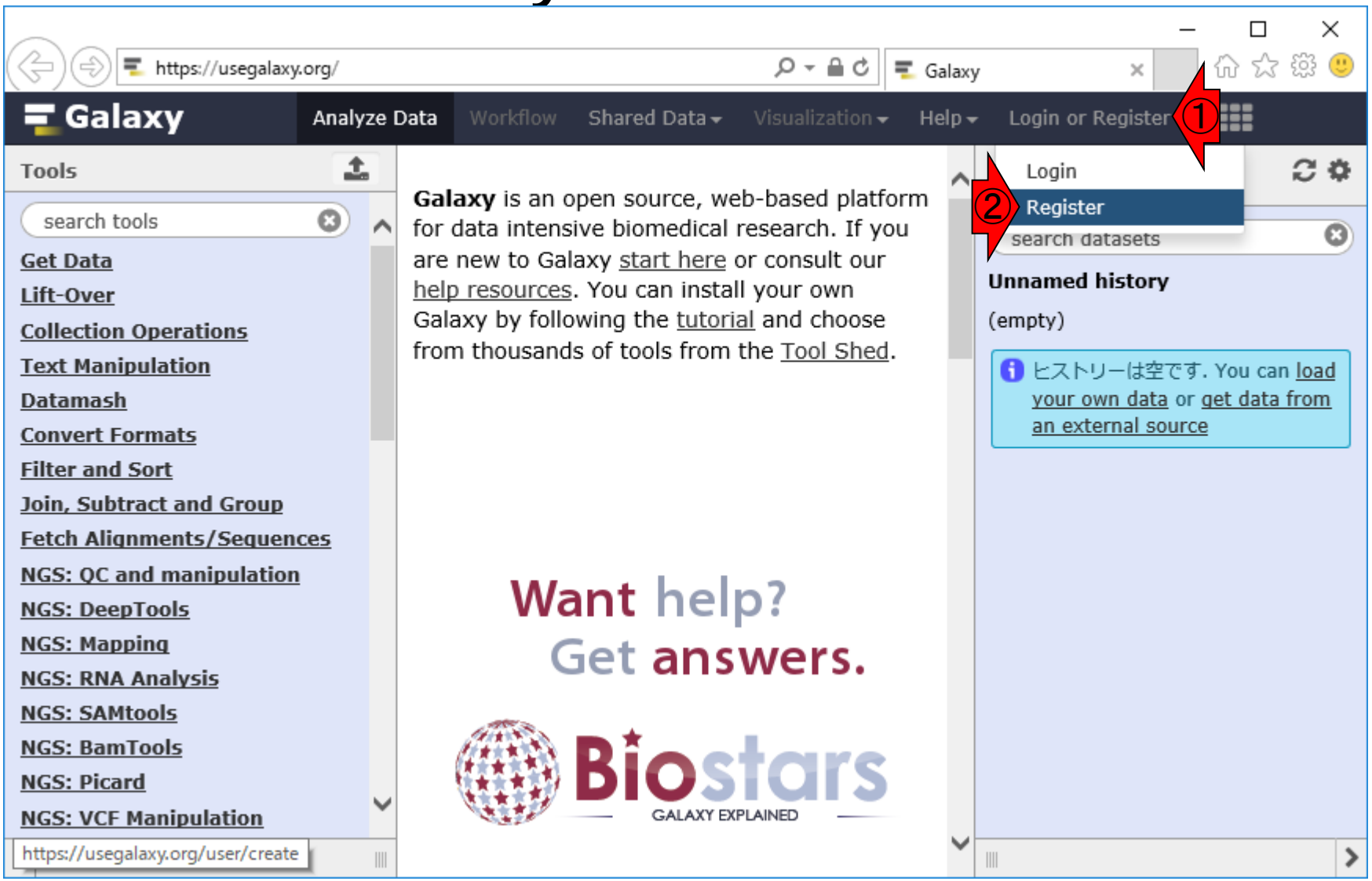
W3-1：ToolShed

①Browse by categoryを押すと、元の画面に戻ります



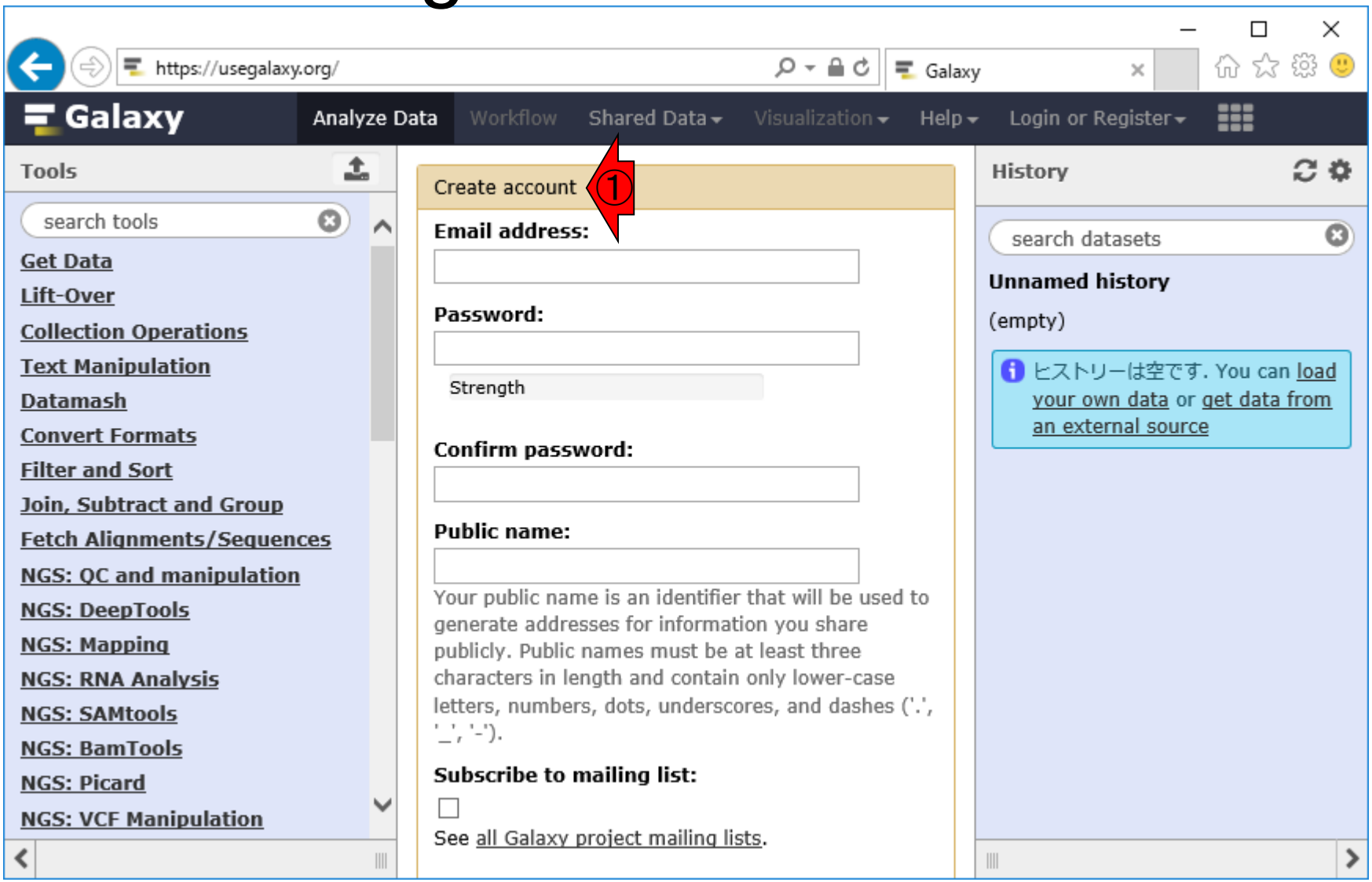
W4-1：Galaxy main

Galaxy mainのトップ画面。①Login or Register、②Registerを選択して、ユーザ登録する



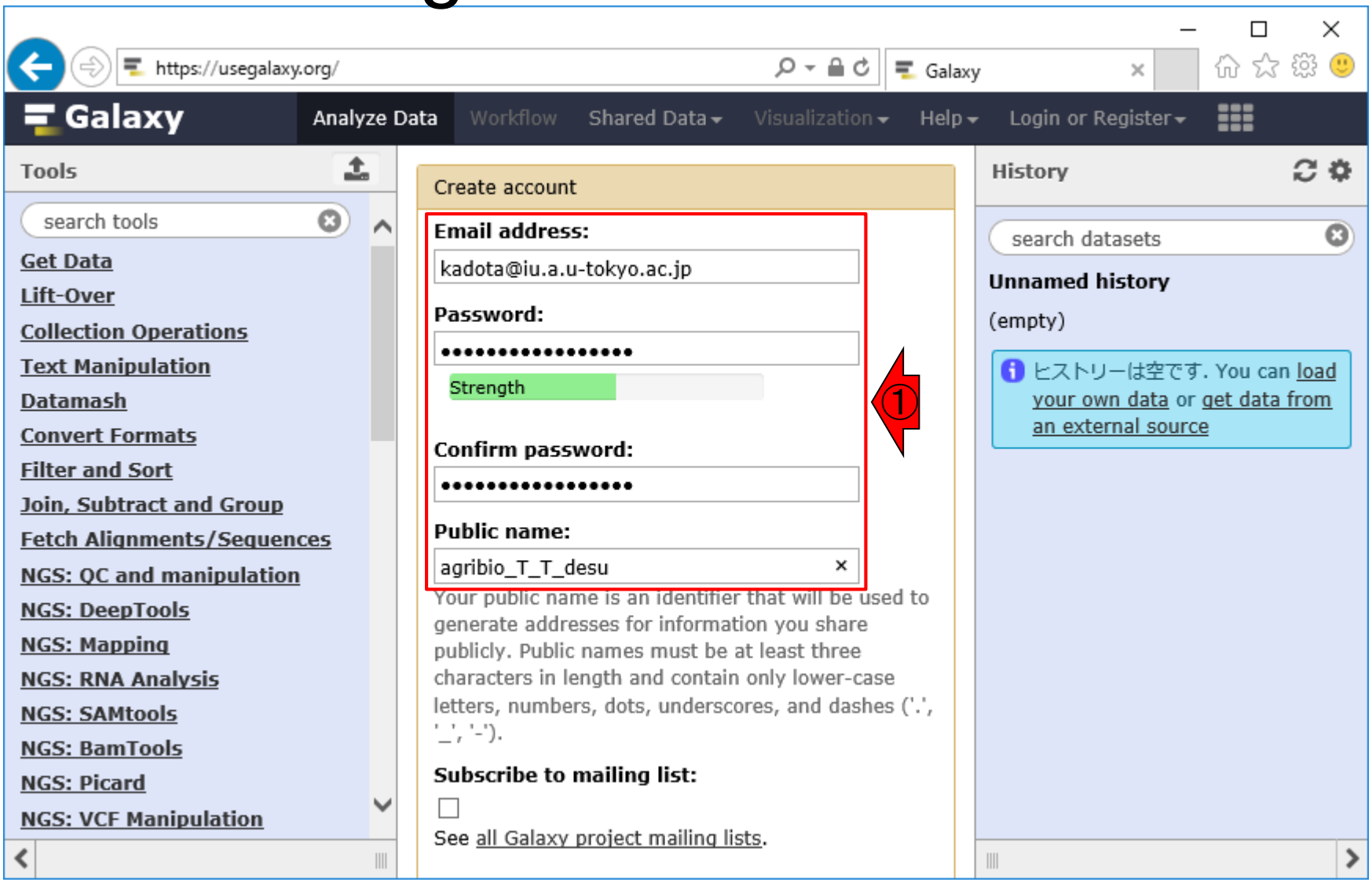
①Create account画面

W4-2：Register



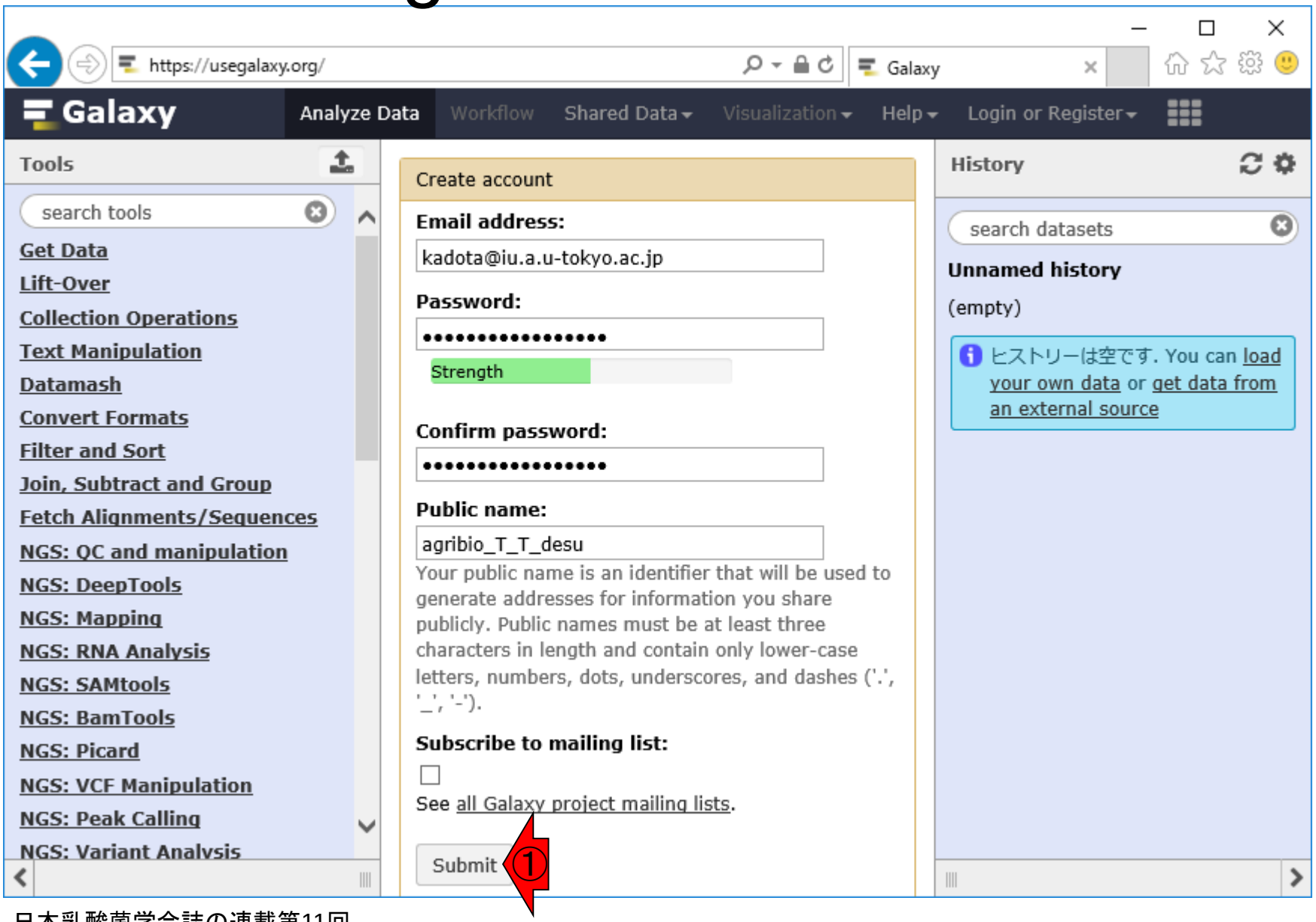
W4-2：Register

①必要事項を入力して…画面下部に移動できないので画面サイズを大きくする



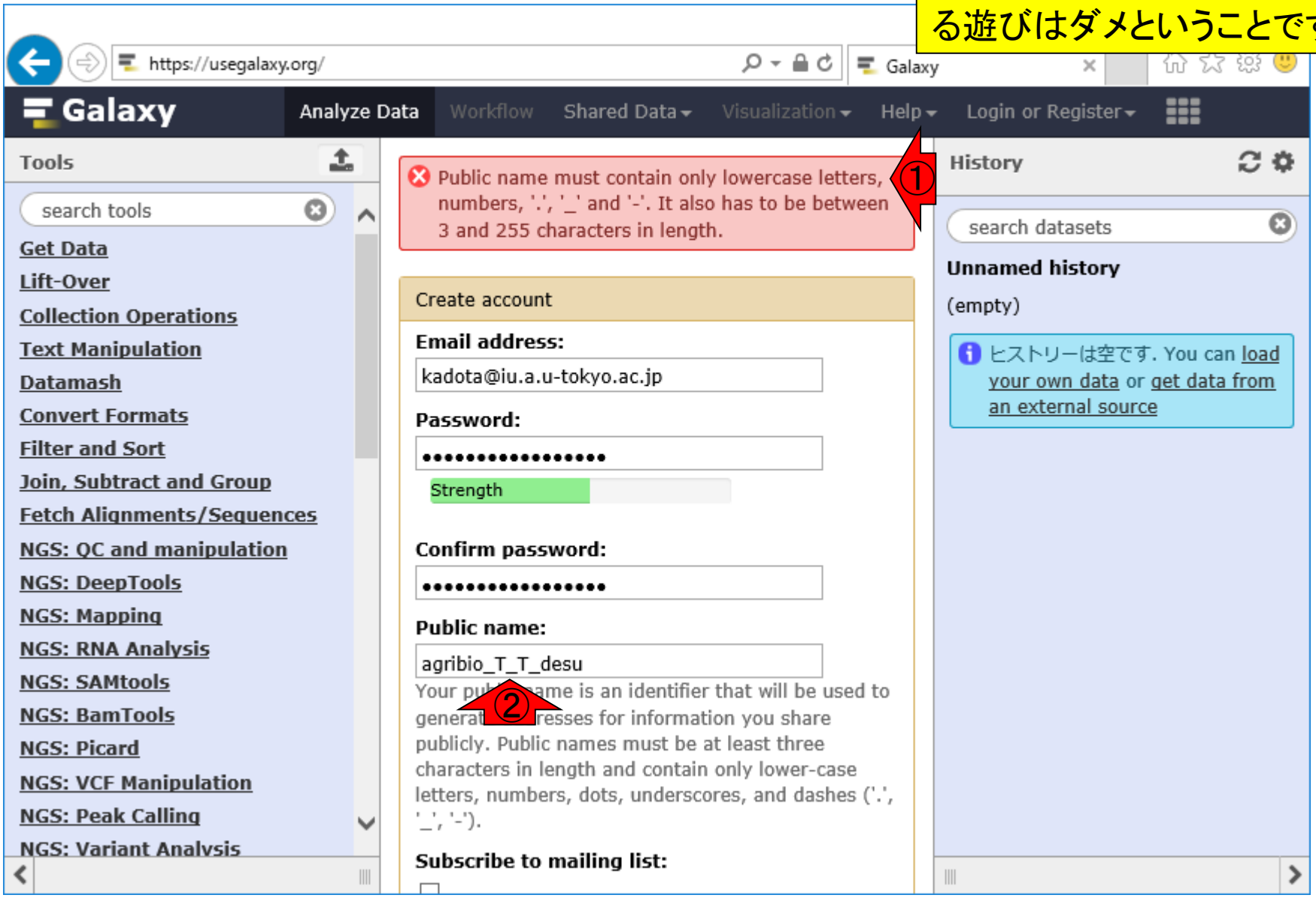
Submitボタンを発見したので、①Submit

W4-2：Register



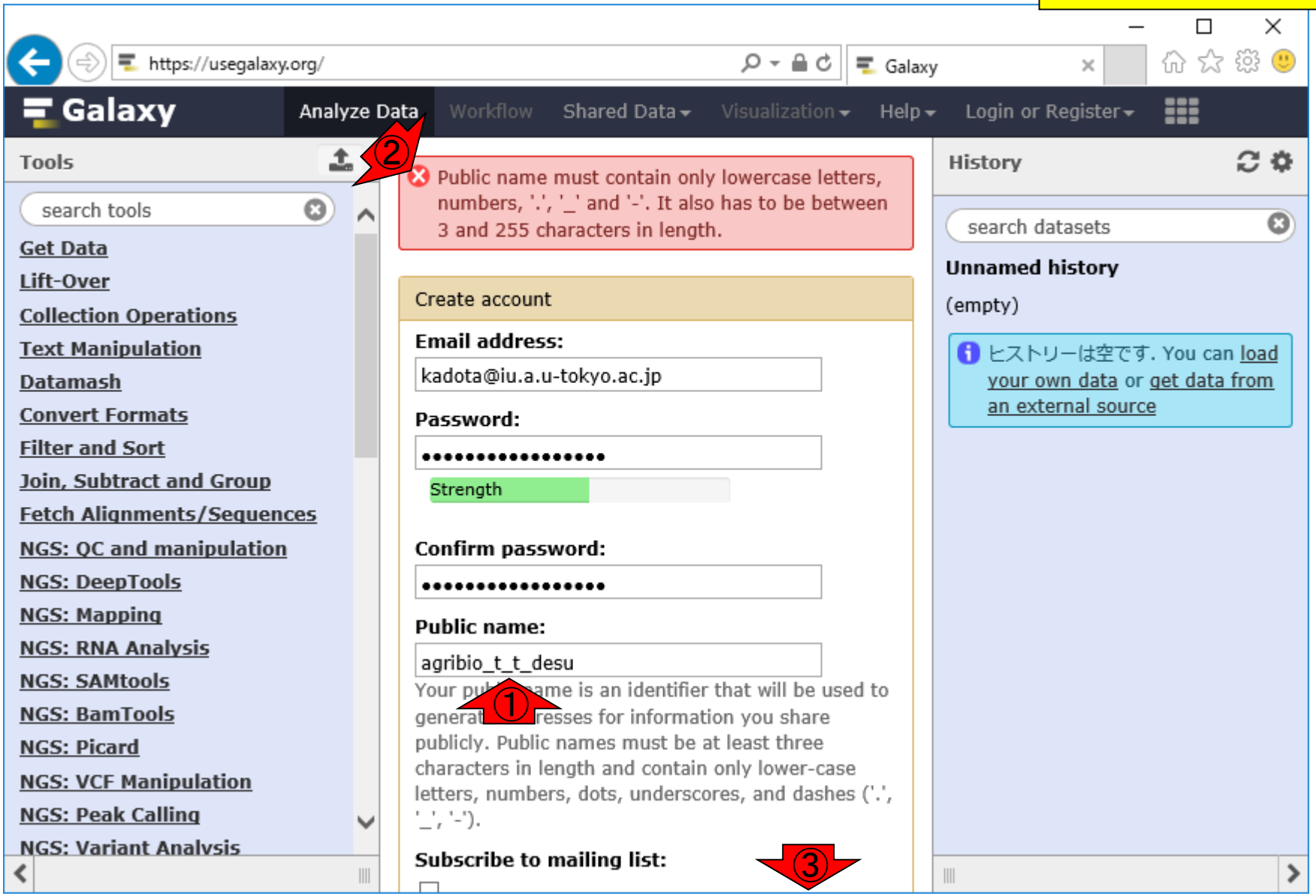
W4-3：Public name

①Public nameはlowercase letters（小文字)で書けと言われていたのを忘れていましたm(_ _)m。②泣き顔（T_T）を入れる遊びはダメということですねw



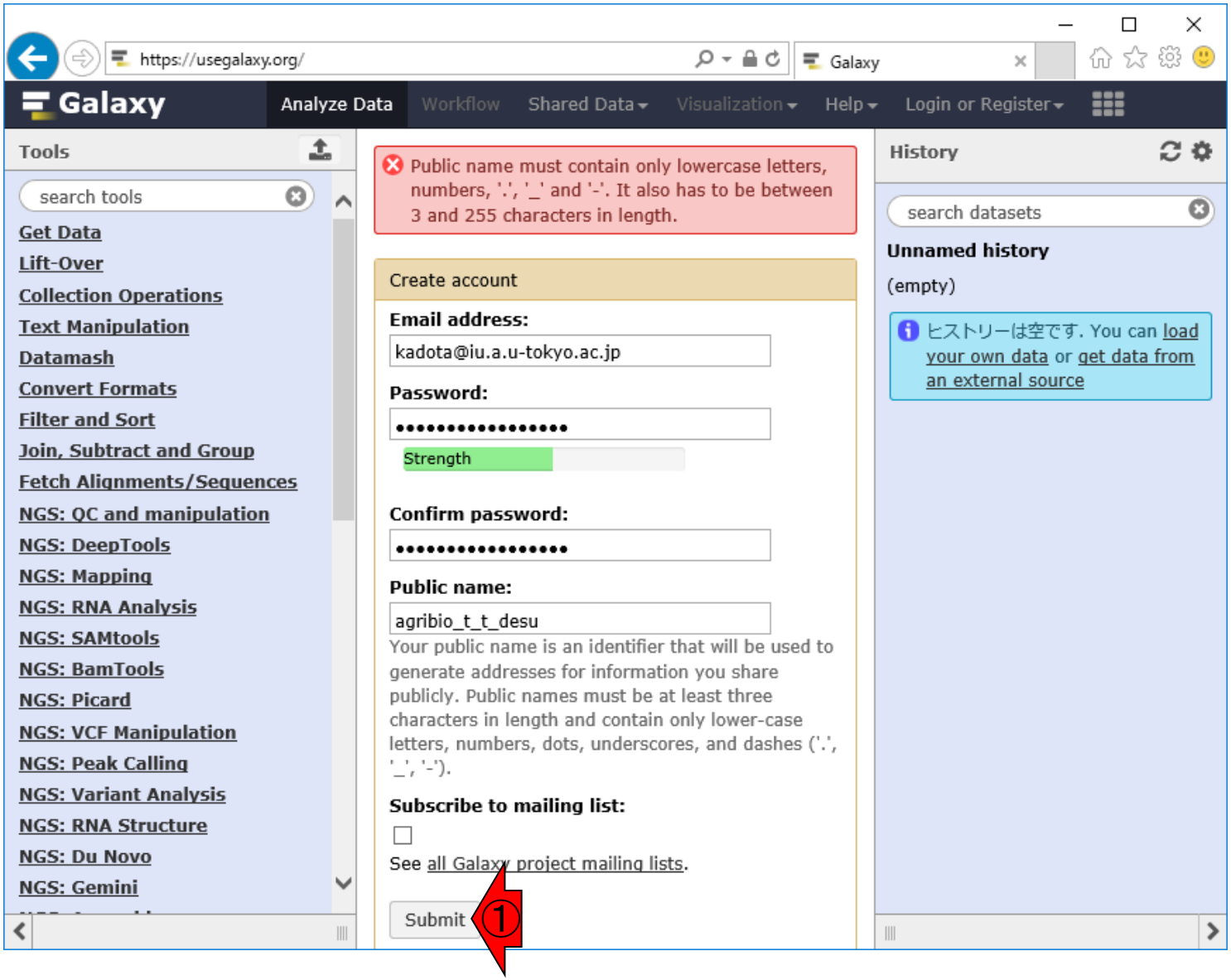
W4-3：Public name

①該当部分を小文字に変えて、
②×を押…せないようなので…
③画面サイズをさらに広げる



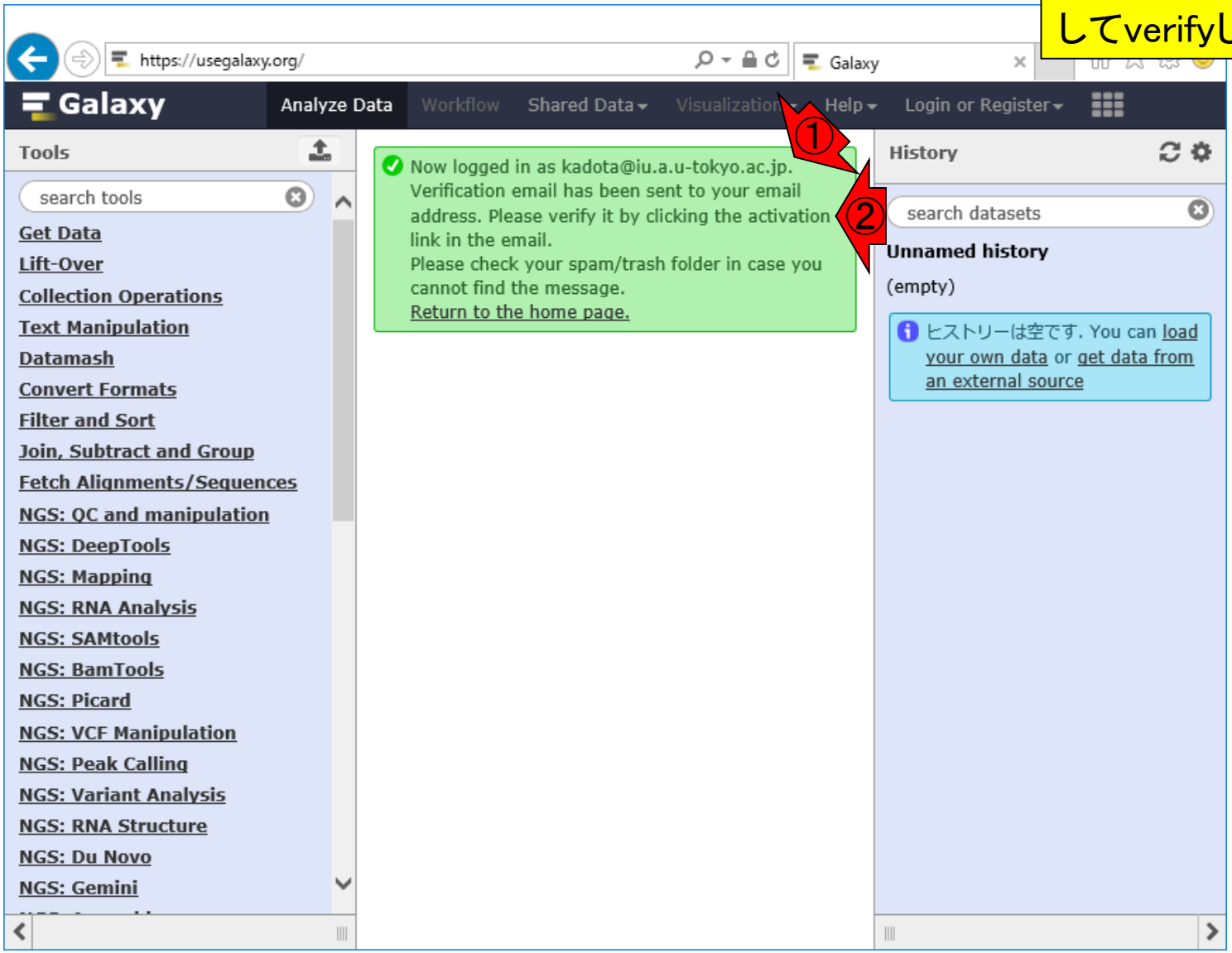
再度①Submit

W4-3：Public name



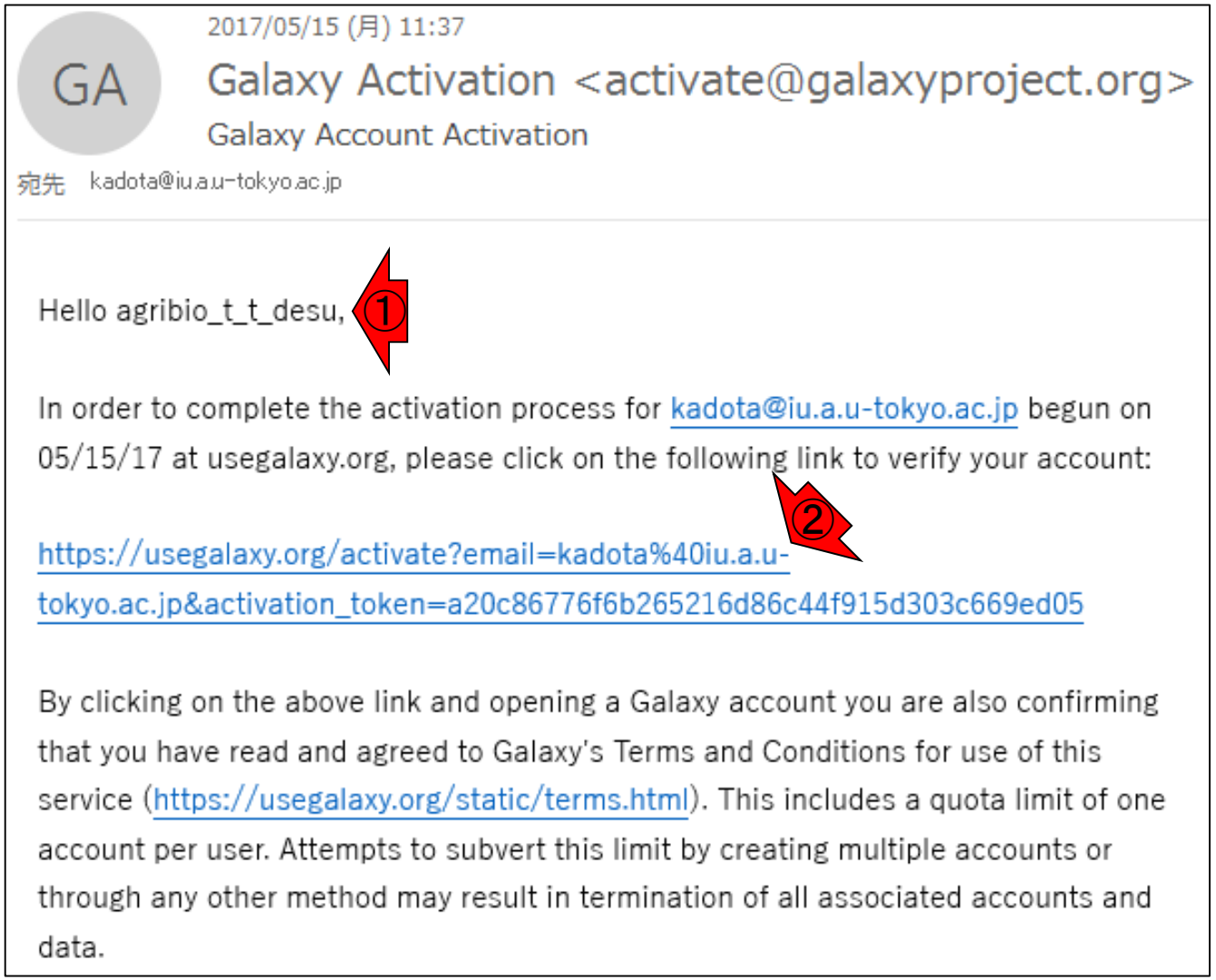
W4-4：ログインできた

①kadota@iu.a.u-tokyo.ac.jpさんとしてログイン状態となったようだ。②e-mailが送られたようなので、それをクリックしてverifyしないといけないようですね



W4-5：メールがすぐきた

確かに①ユーザagribio_t_t_desuさんのアドレス宛にメールがきた。②をクリック



2017/05/15 (月) 11:37

GA

Galaxy Activation <activate@galaxyproject.org>

Galaxy Account Activation

宛先 kadota@iu.a.u-tokyo.ac.jp

Hello agribio_t_t_desu, ①

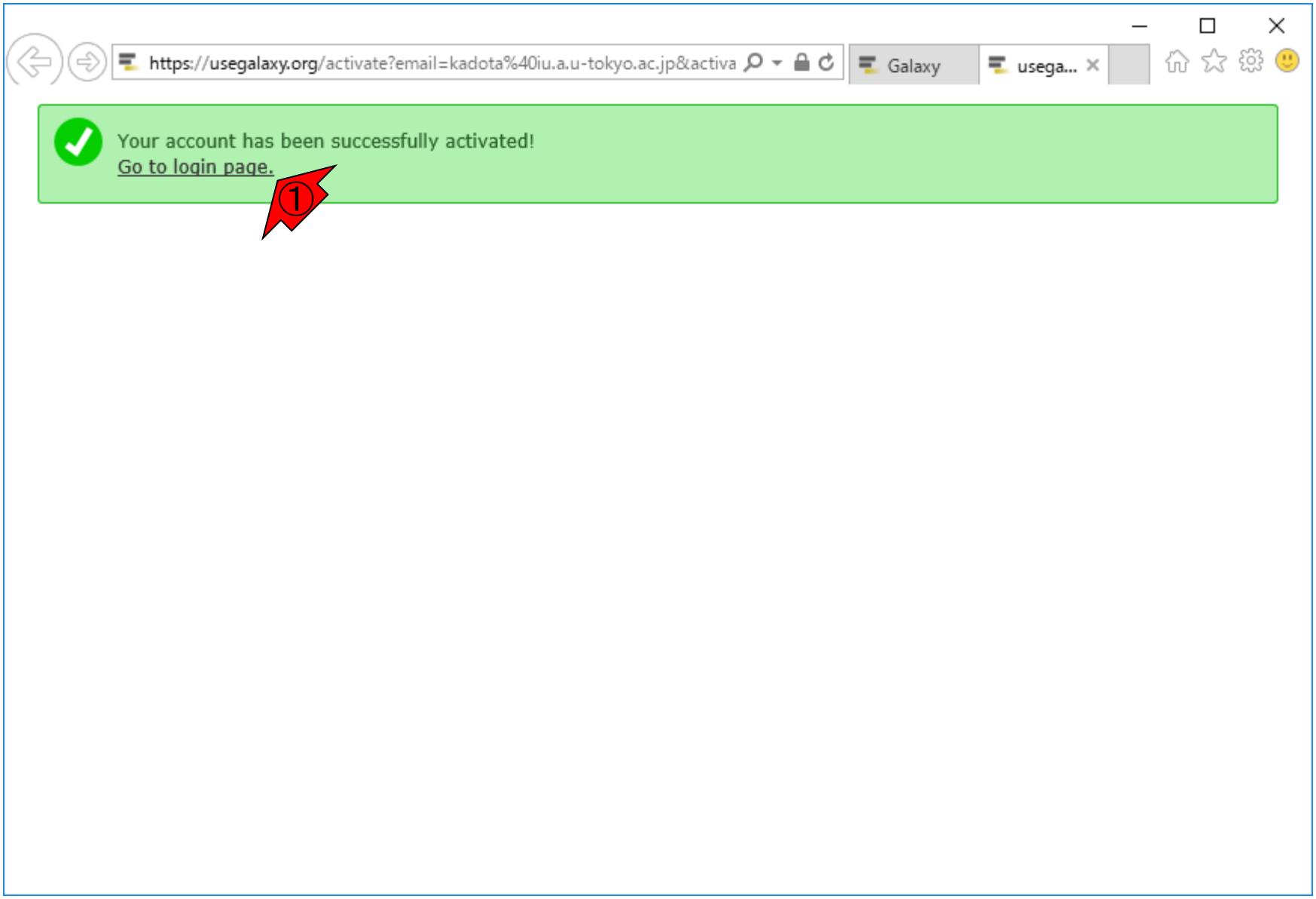
In order to complete the activation process for kadota@iu.a.u-tokyo.ac.jp begun on 05/15/17 at usegalaxy.org, please click on the following link to verify your account: ②

https://usegalaxy.org/activate?email=kadota%40iu.a.u-tokyo.ac.jp&activation_token=a20c86776f6b265216d86c44f915d303c669ed05

By clicking on the above link and opening a Galaxy account you are also confirming that you have read and agreed to Galaxy's Terms and Conditions for use of this service (https://usegalaxy.org/static/terms.html). This includes a quota limit of one account per user. Attempts to subvert this limit by creating multiple accounts or through any other method may result in termination of all associated accounts and data.

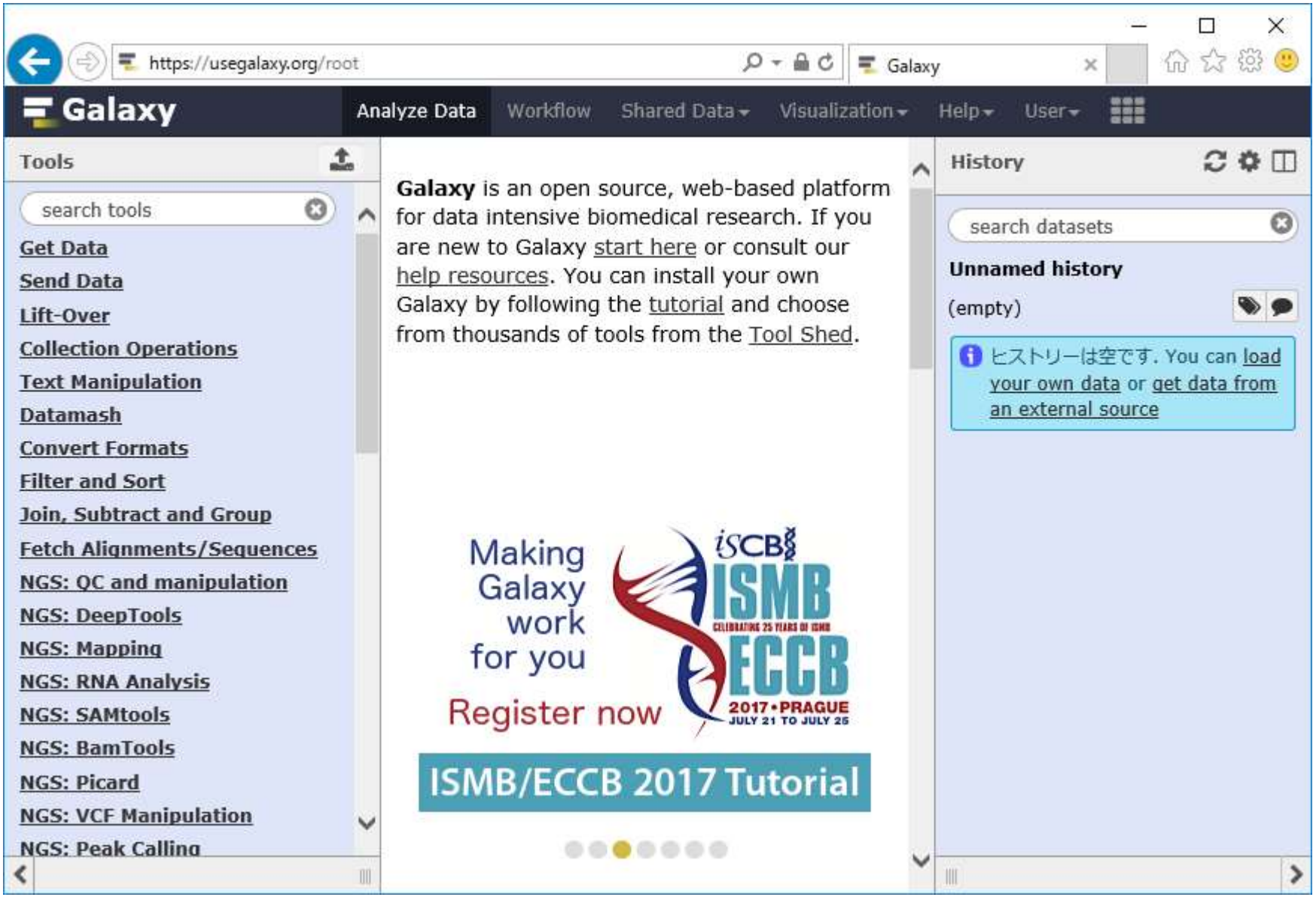
W4-6：アカウントactivated

いわれるがままに、①ログインページに行ってみよう



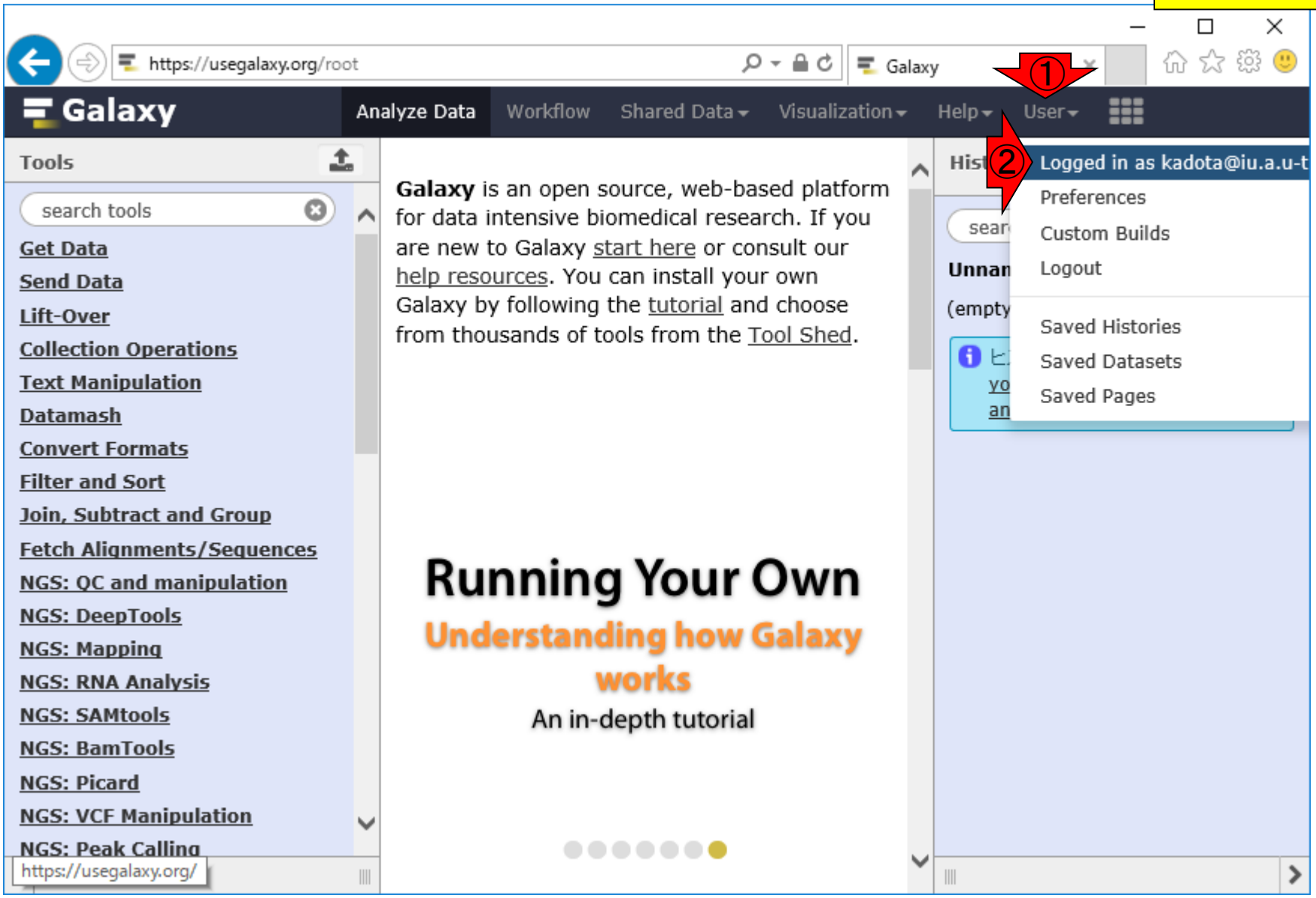
このページに飛ばされました

W4-7：ログインページ?!



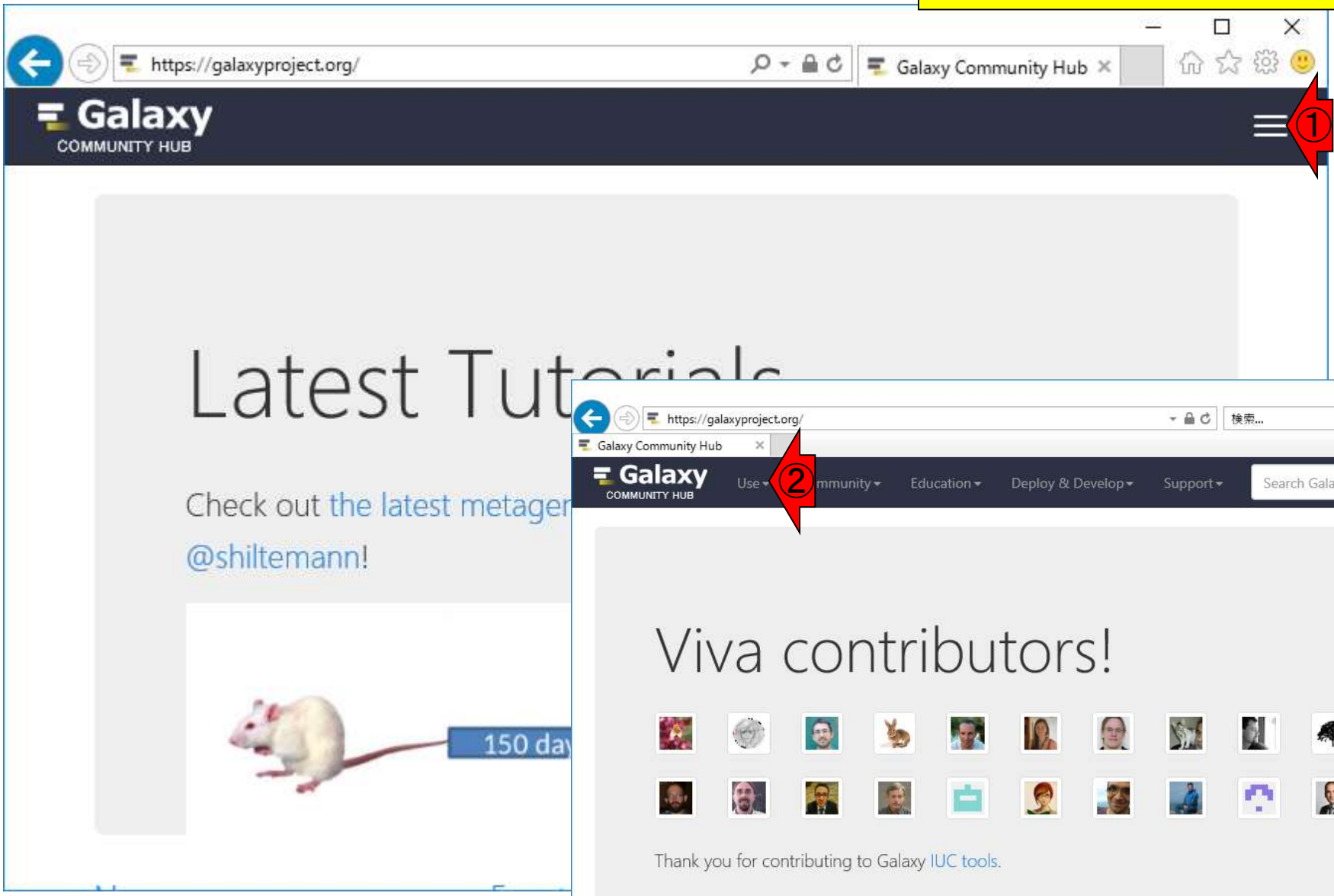
W4-8：ログインしてるのね

①Userをクリックすると、
②Logged in as …となっているからですかね



W5-0：大元から…

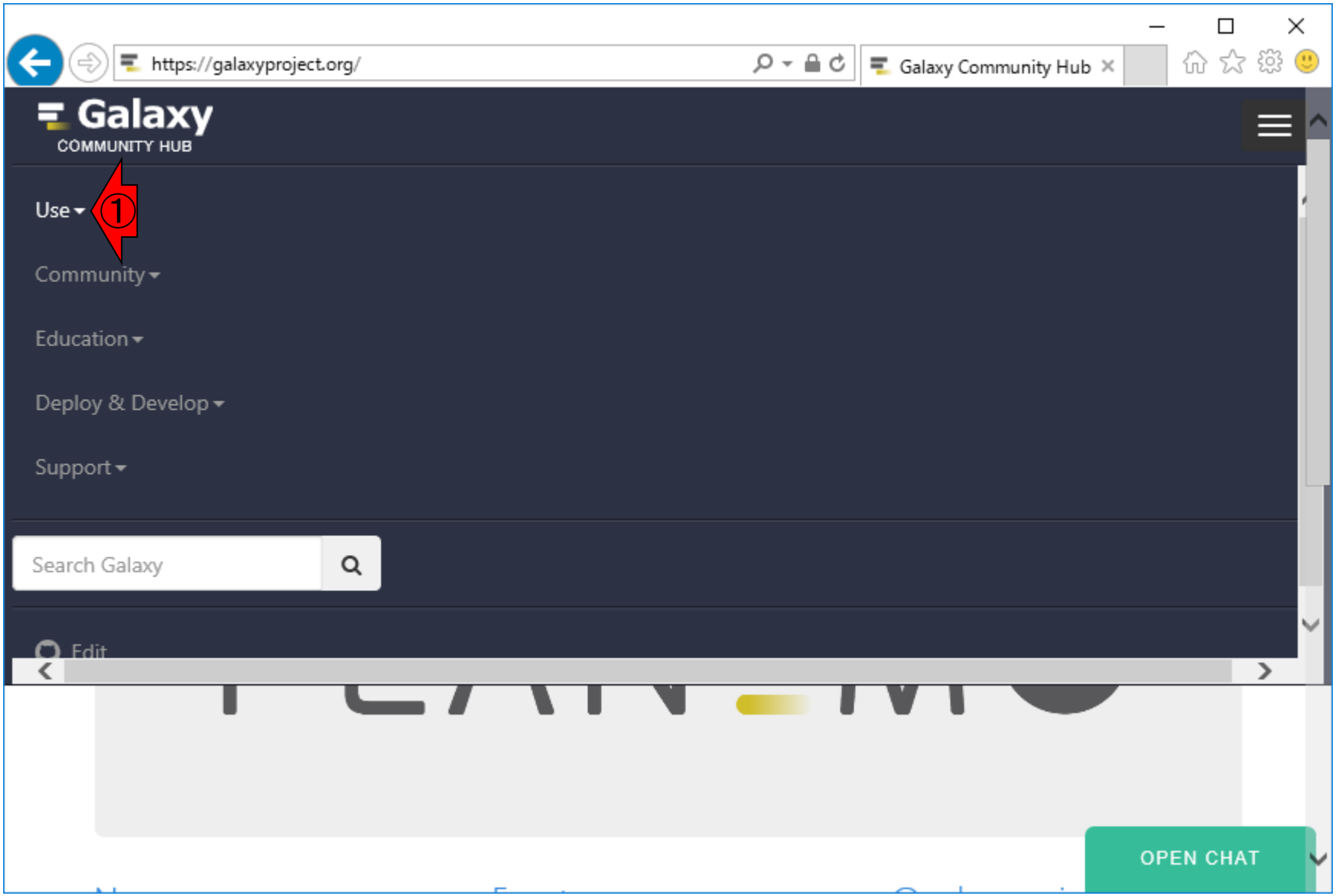
Galaxyの大元のページ。①をクリック。右下のように見えている場合は、②Useを押すと次のスライドとほぼ同じになります



https://galaxyproject.org/

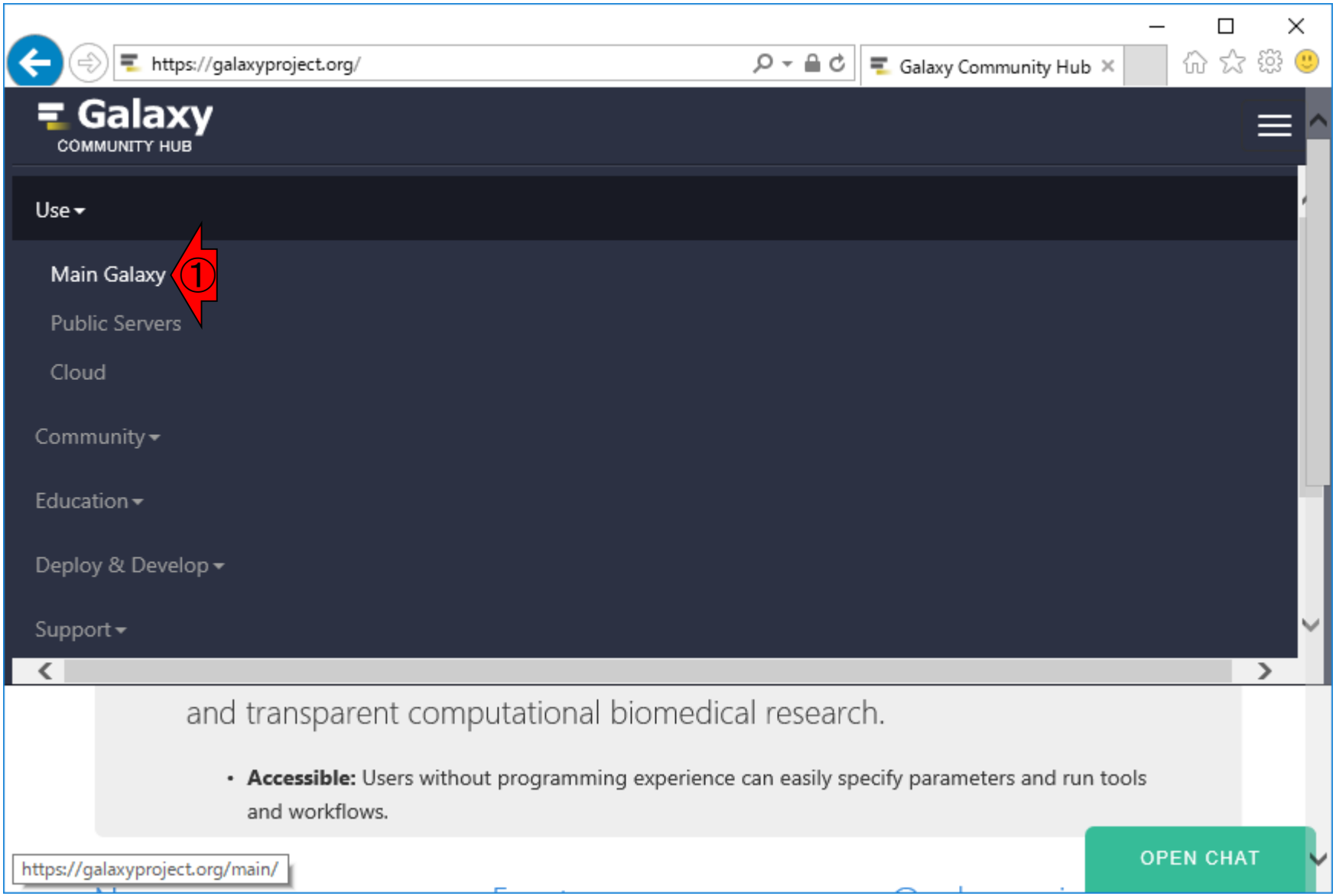
①Use

W5-0：大元から…



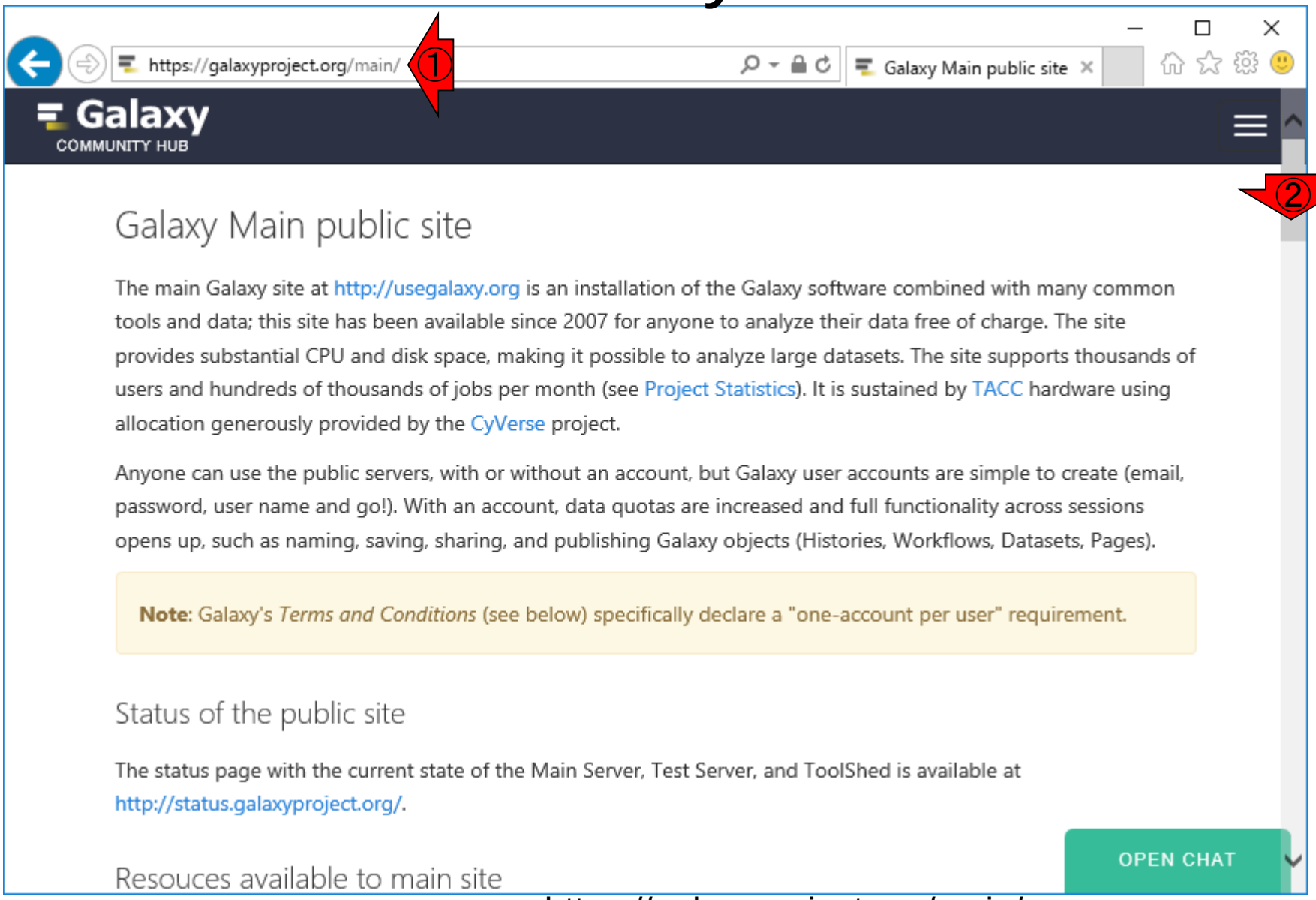
①Main Galaxy

W5-0：大元から…



W5-1：main Galaxy

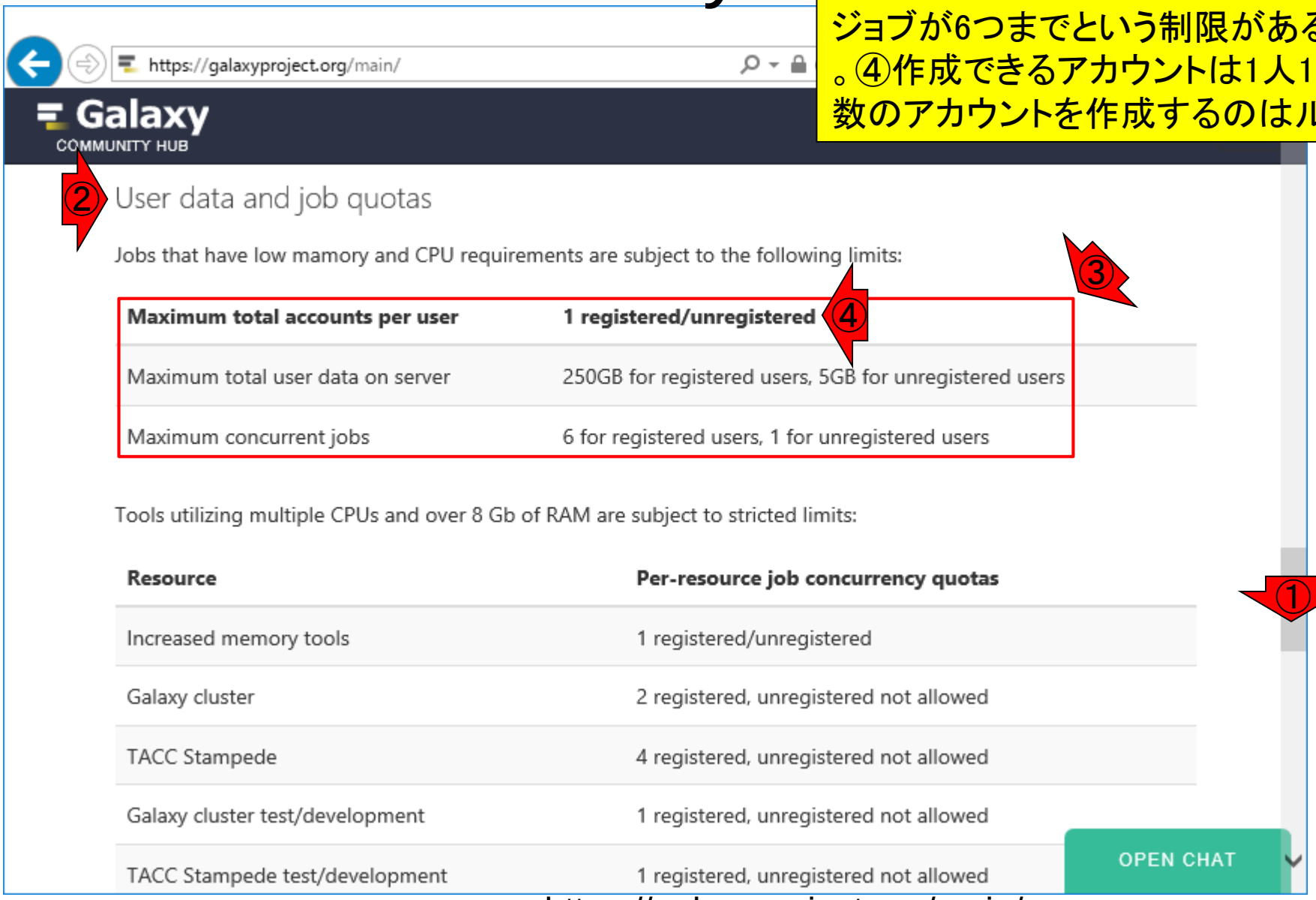
main Galaxyにたどり着きます。②ページ下部に移動



https://galaxyproject.org/main/

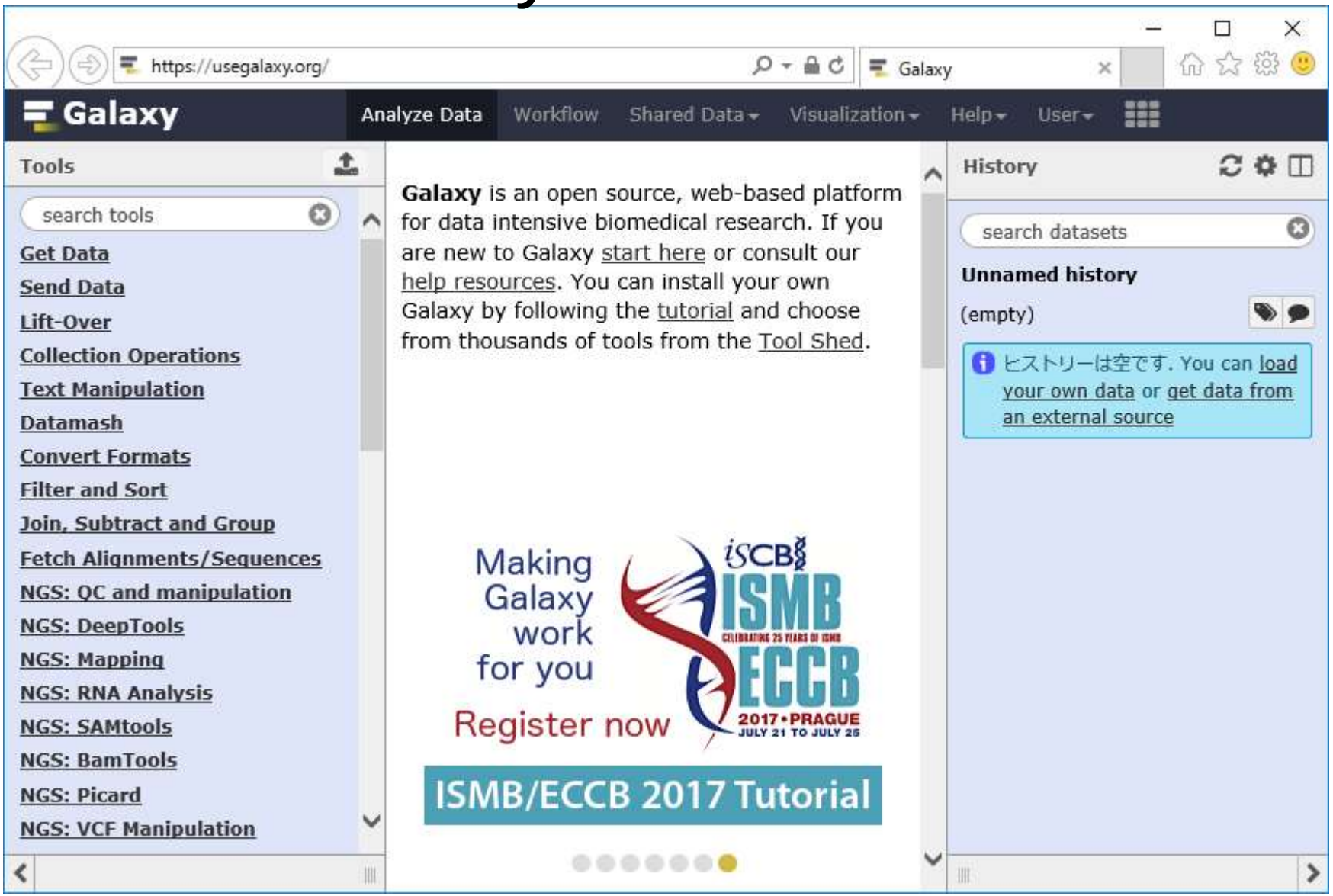
W5-1：main Galaxy

①のあたりまで移動して、②User data and job quotasを眺める。③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ことがわかる。④作成できるアカウントは1人1つなので、複数のアカウントを作成するのはルール違反



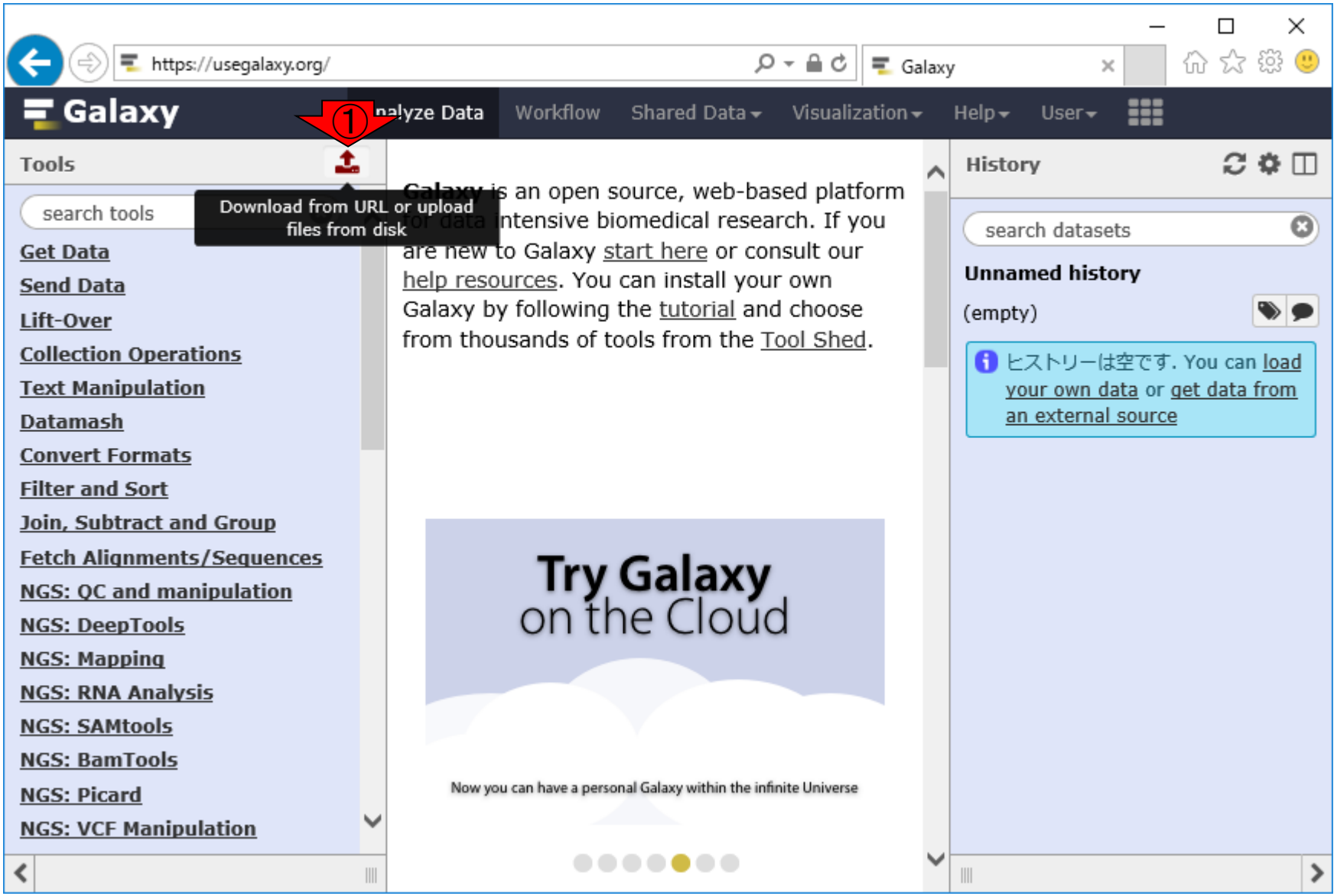
W5-2：Galaxyの画面構成

Galaxy mainの基本操作画面は、縦に3分割された構成



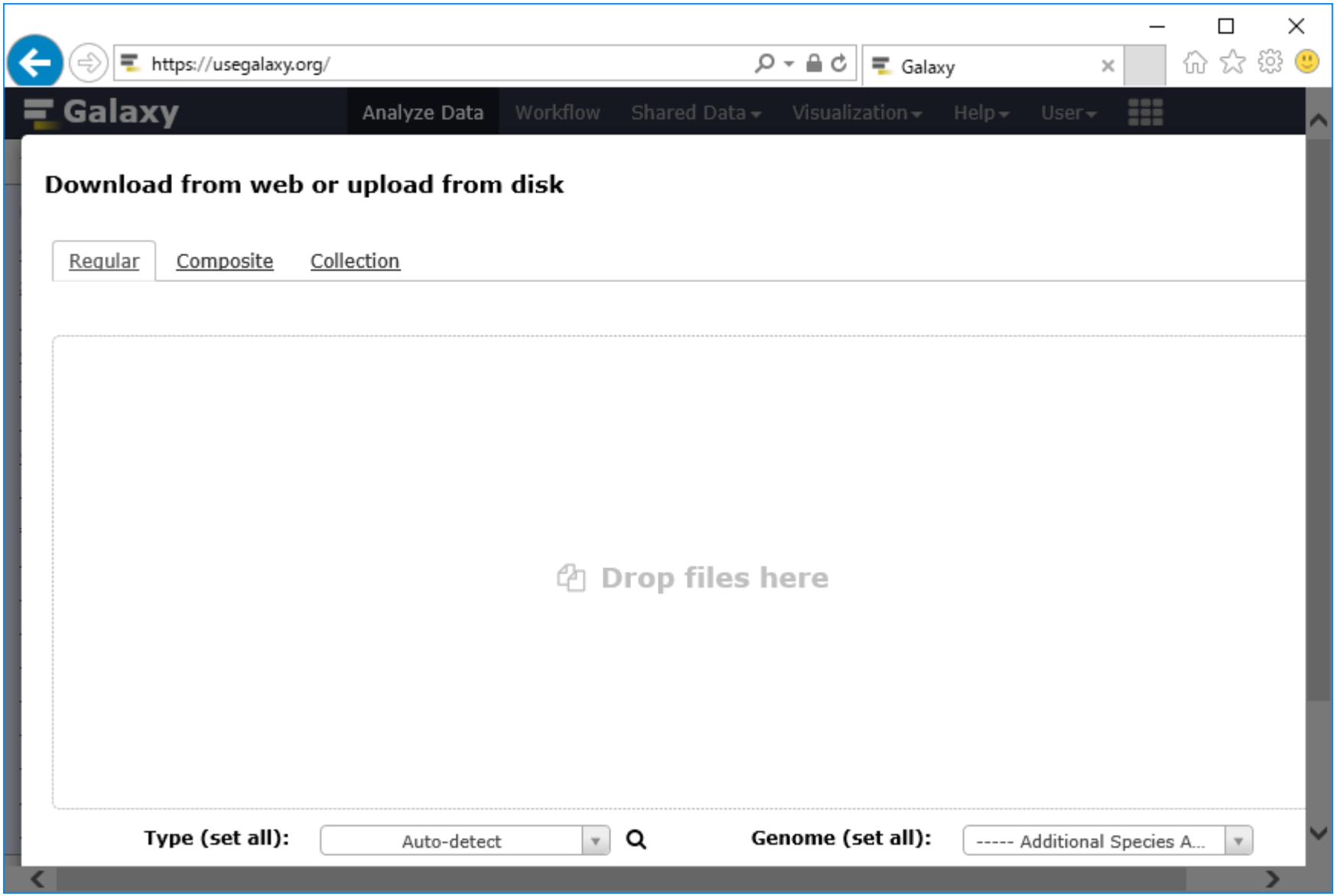
W5-3：アップロード1

①を押し、デスクトップ上のgzip圧縮FASTQファイル（DRR024501sub_1.fastq.gz）のアップロードを行う



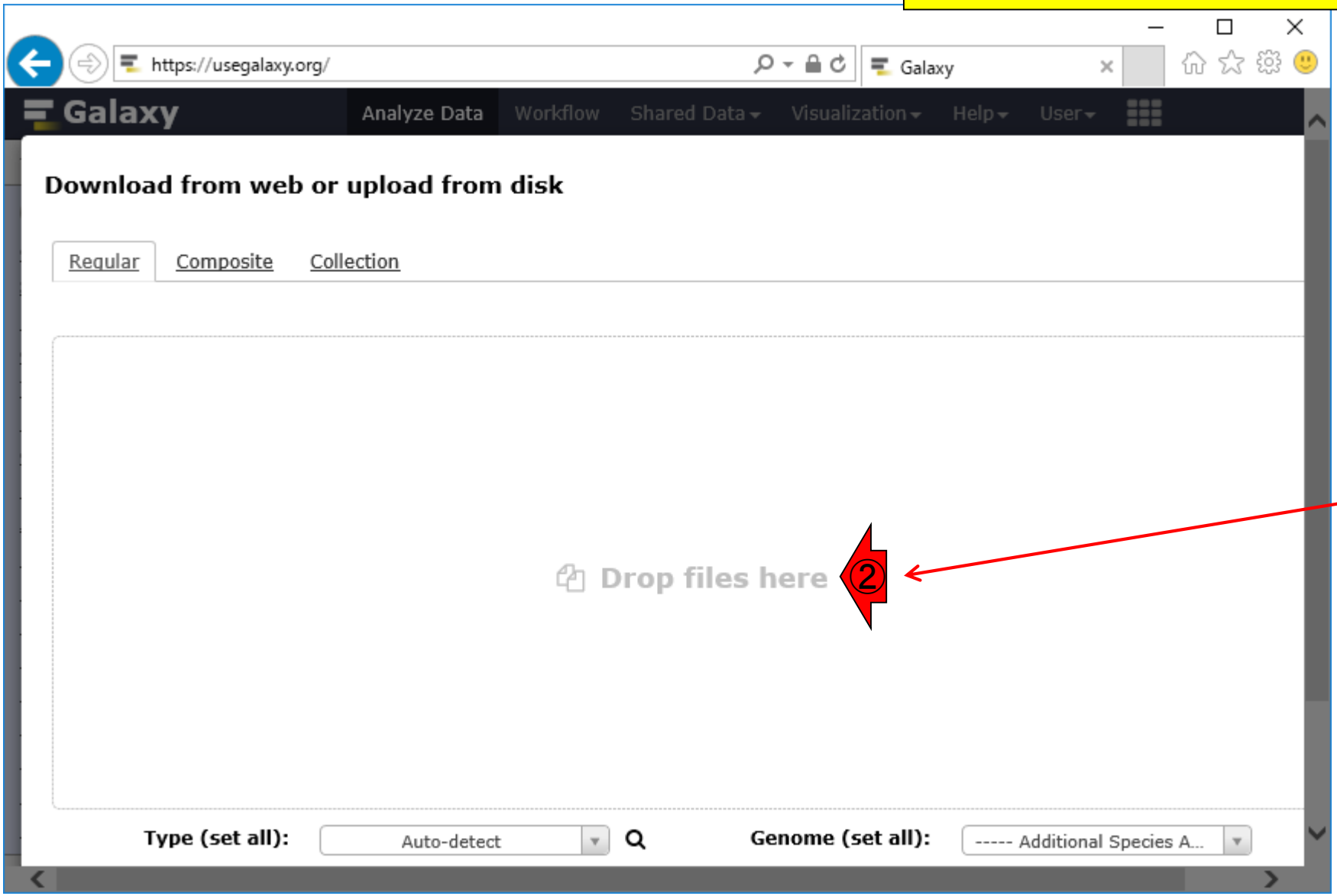
こんな画面になります

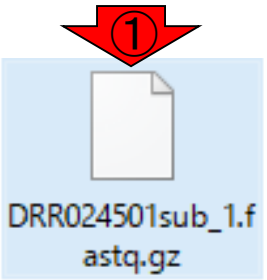
W5-3：アップロード2



W5-3：アップロード3

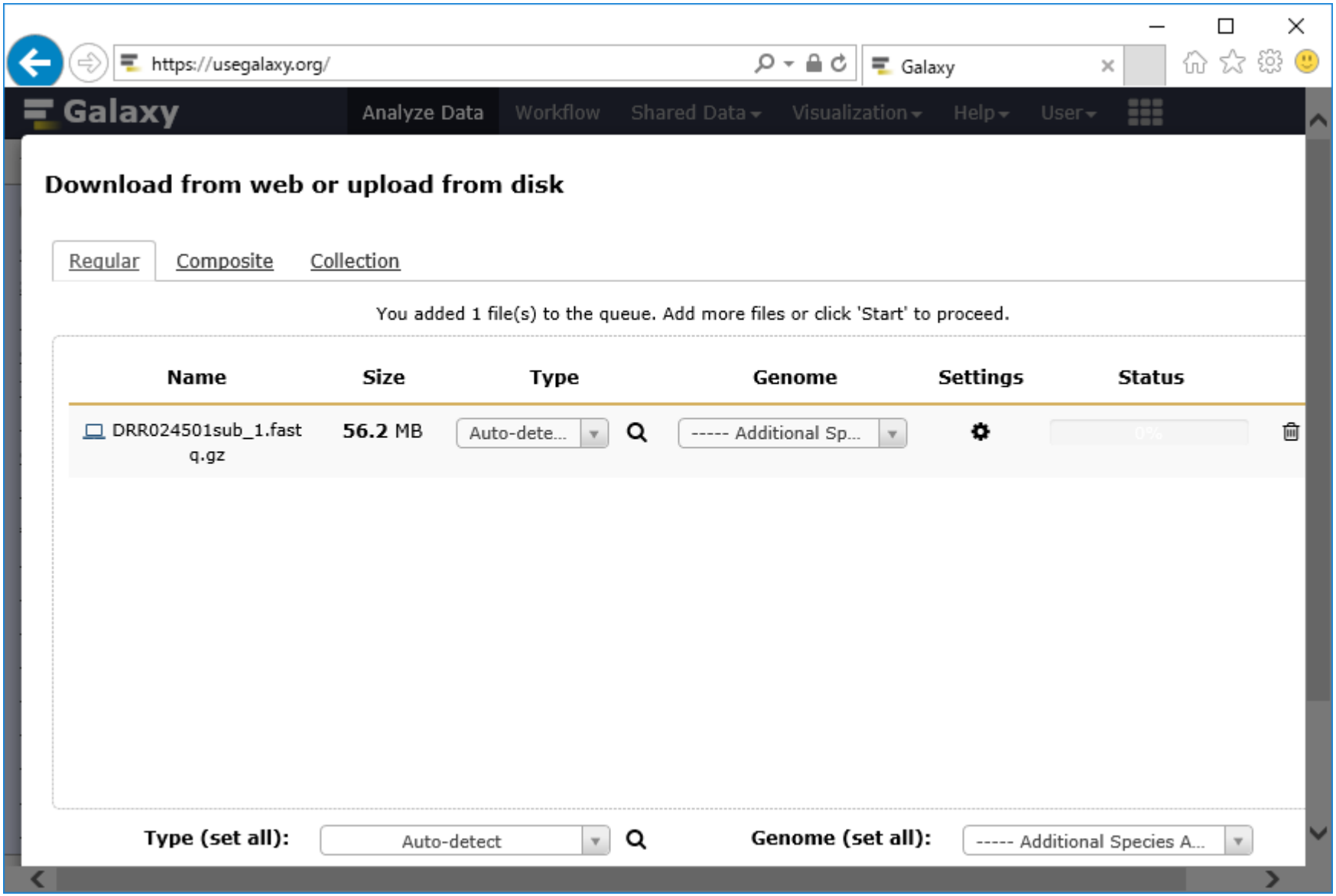
デスクトップ上にある①gzip圧縮FASTQファイル（DRR024501sub_1.fastq.gz）を②Drop files hereのところにドラッグ&ドロップ





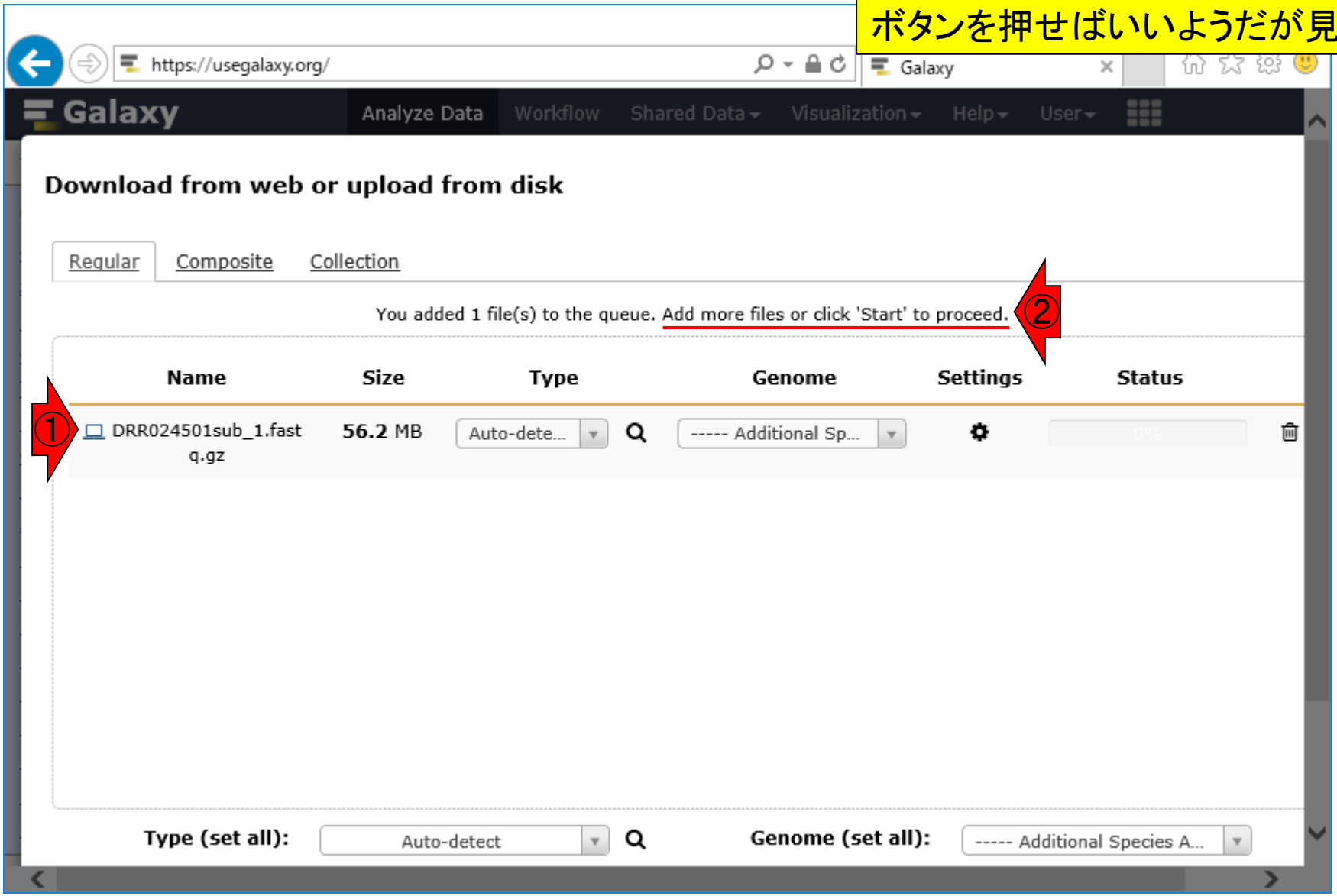
すぐにこんな感じの画面になります

W5-3：アップロード4



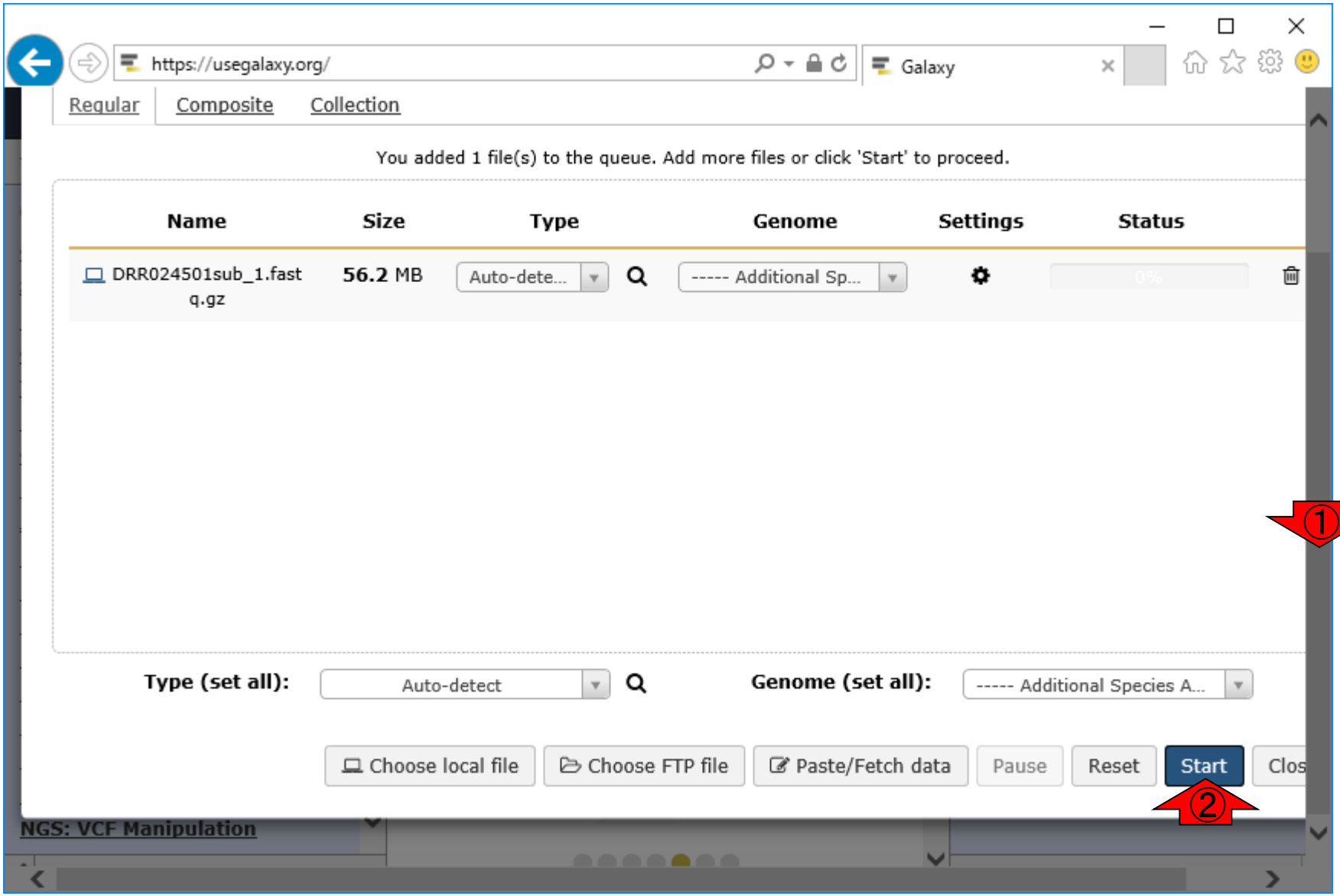
W5-3：アップロード5

追加で他のファイルもアップロードしたい場合は、同様な手順で行うのだろう。ここでは①のファイルのみアップロードしたいので、②Startボタンを押せばいいようだが見当たらない



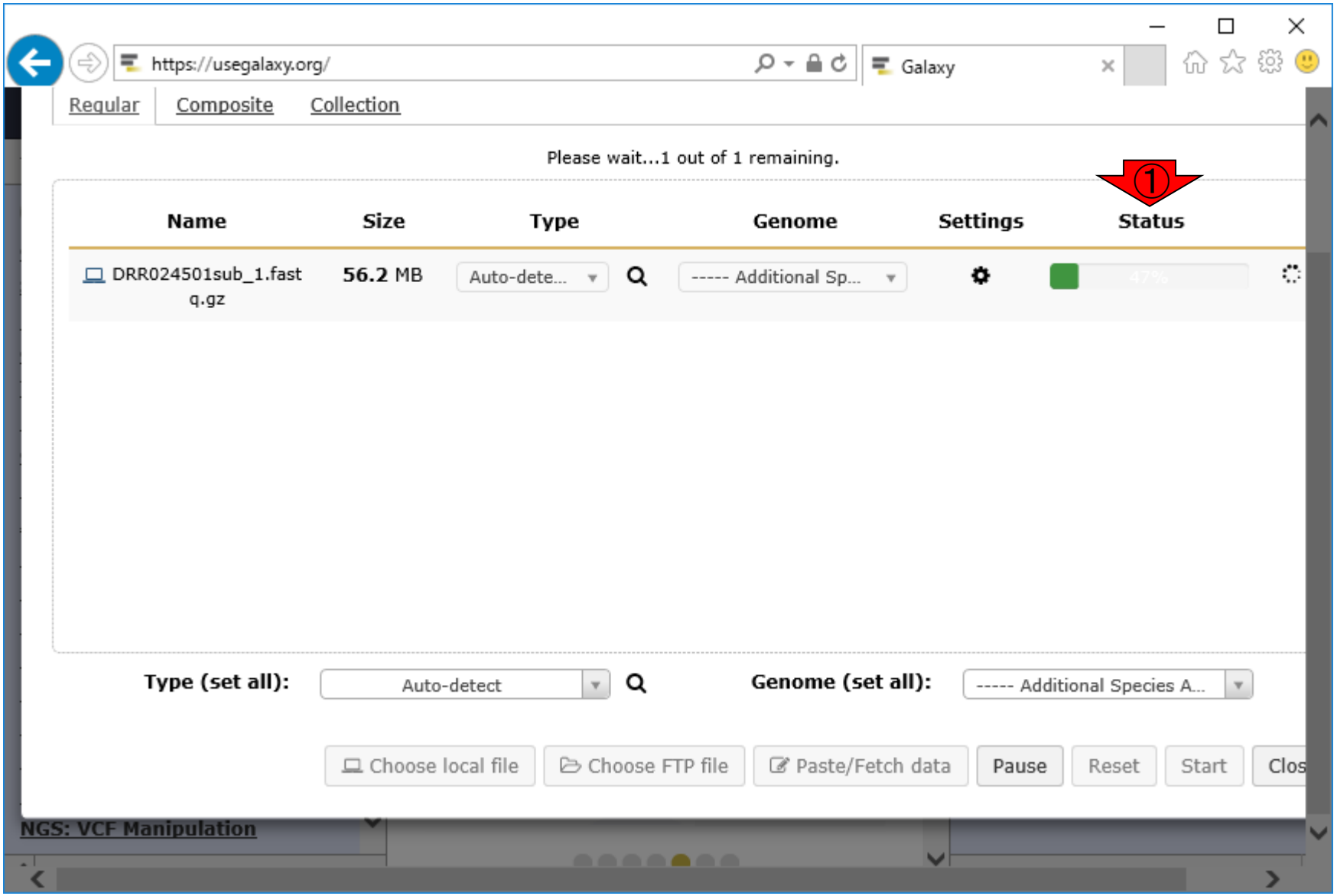
W5-3：アップロード6

①ページ下部に移動したら②
Startボタンが見つかったので押す



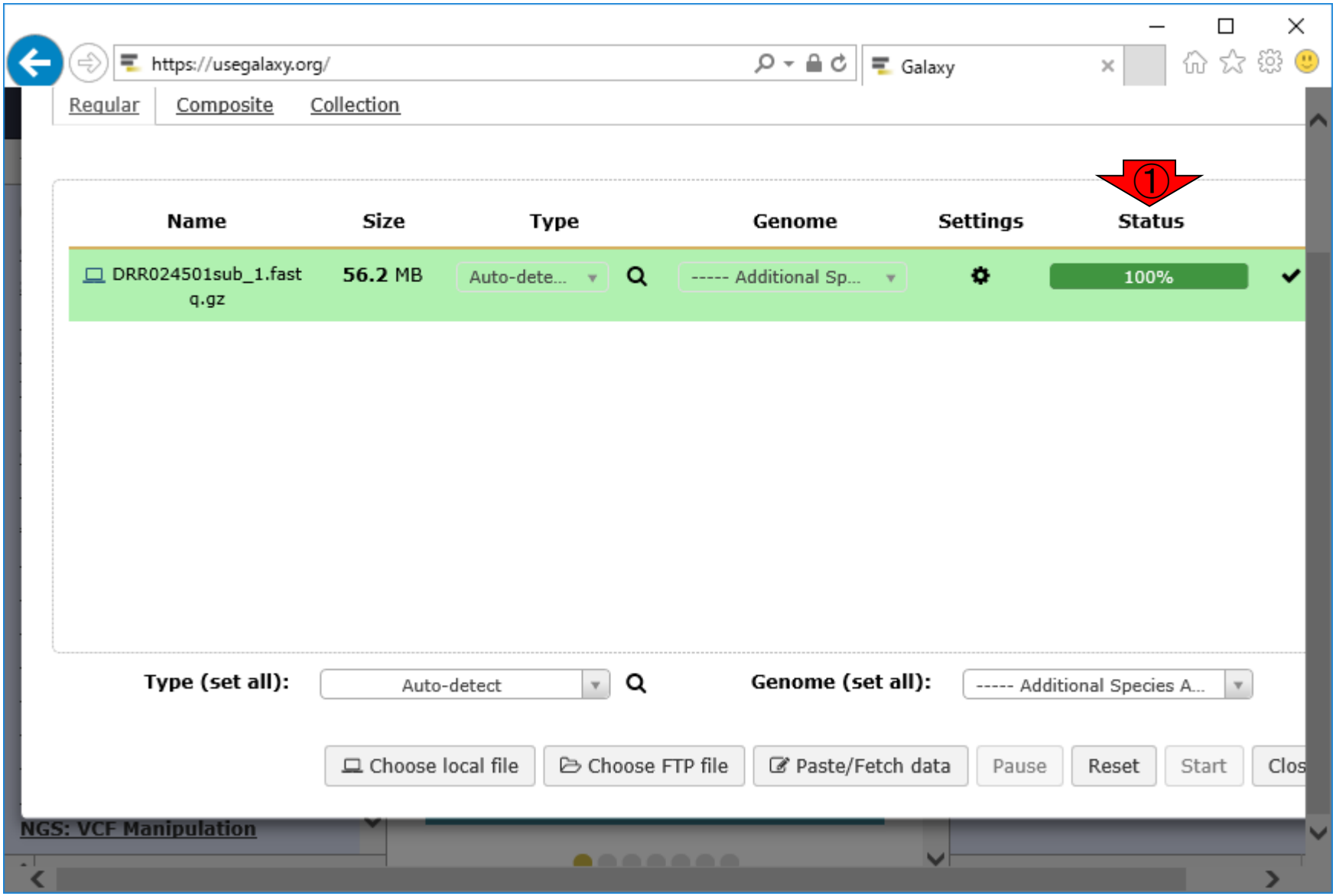
①アップロード中です…

W5-4：アップロード中…



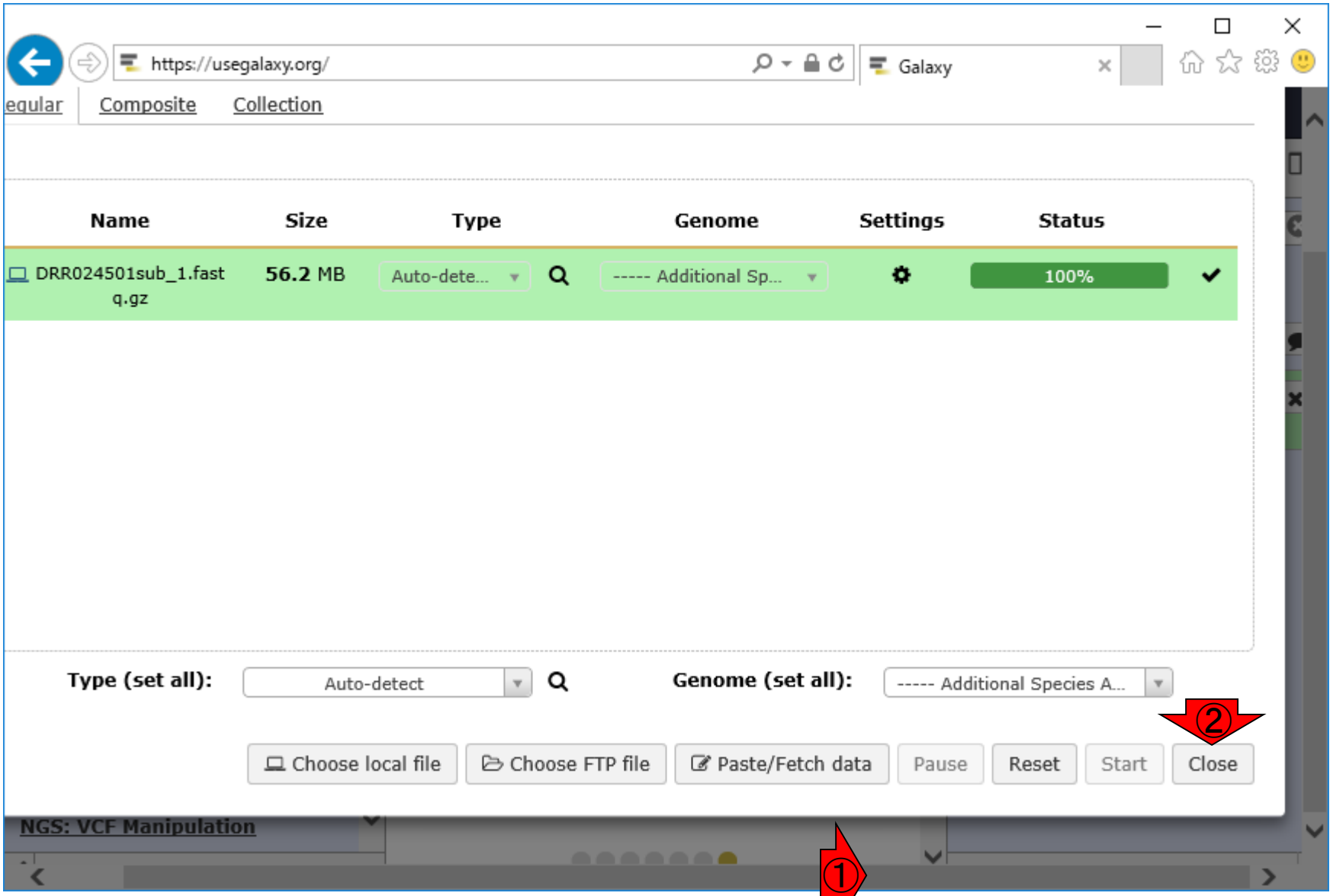
①アップロード完了です

W5-5：アップロード完了



①ページ右側に移動して、②Close

W5-6：Close

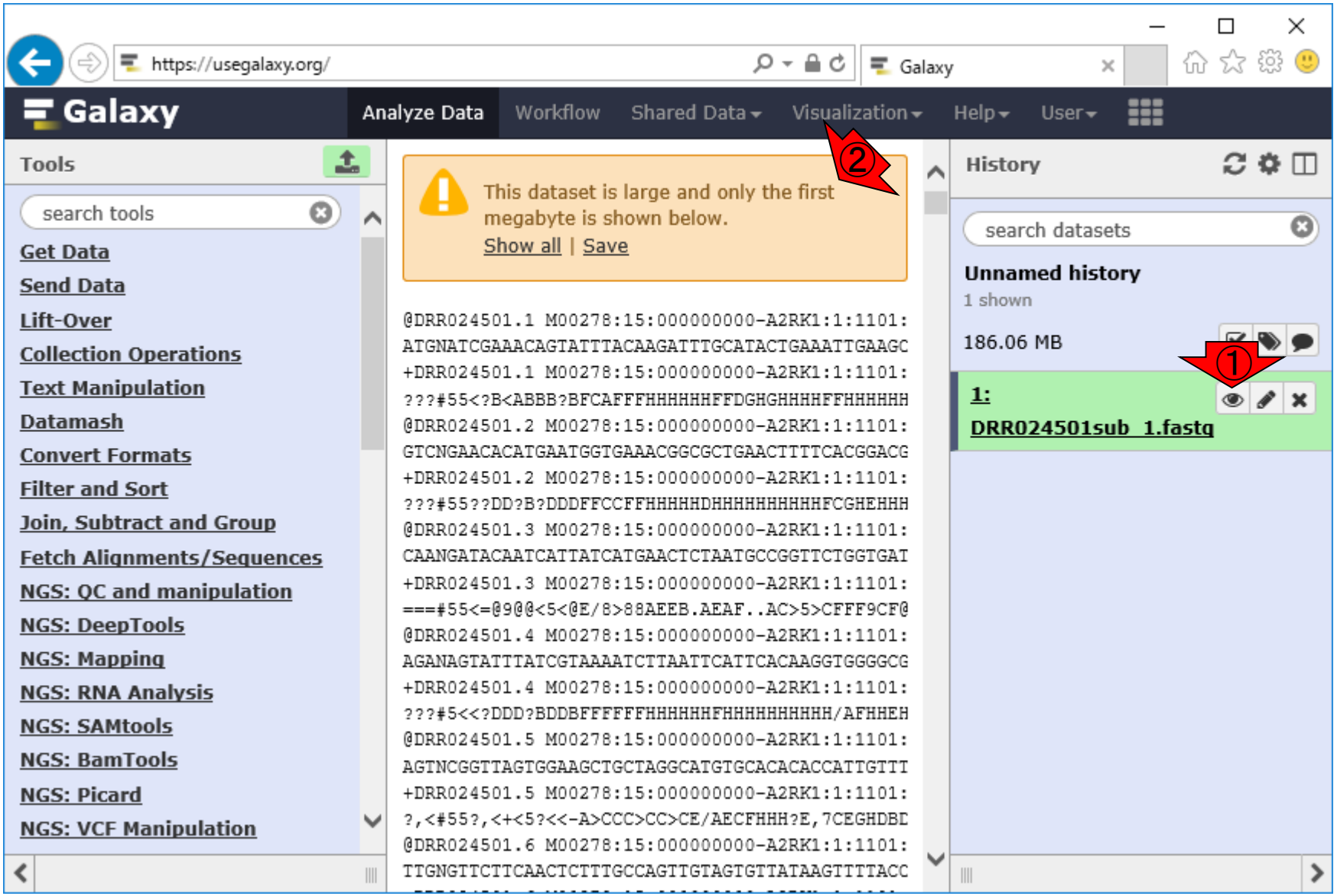


W6-1：アップロード後

アップロード後。gzip圧縮状態でアップロードしたオリジナルのファイルサイズは56.1MBであったが、①186.06MBとなっていることがわかる。②gzip非圧縮状態になった（展開された）ようだ

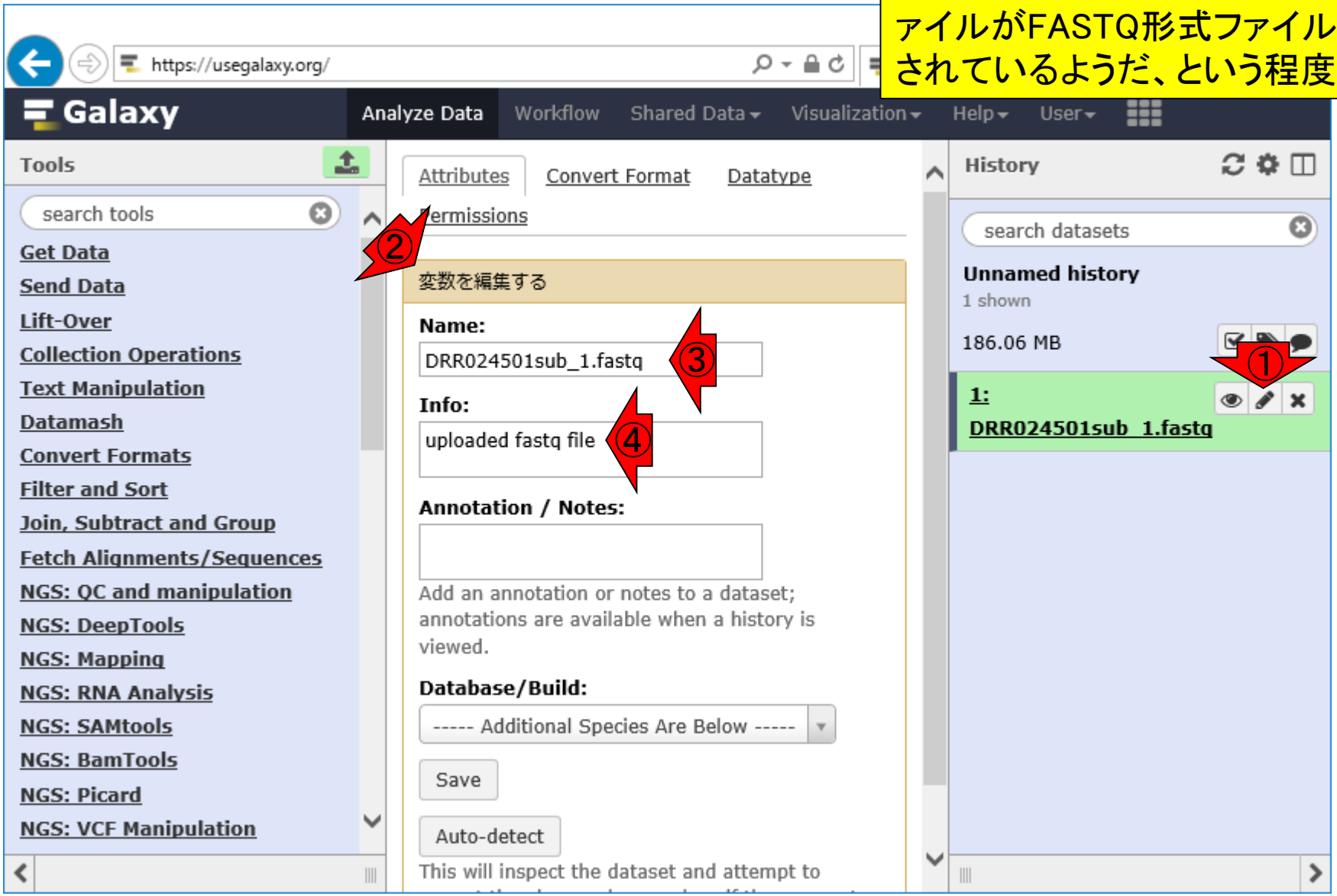
W6-2：データを表示

①がデータ表示ボタン（クリック）。②デカすぎるので最初の1MB分しか表示されないようだ



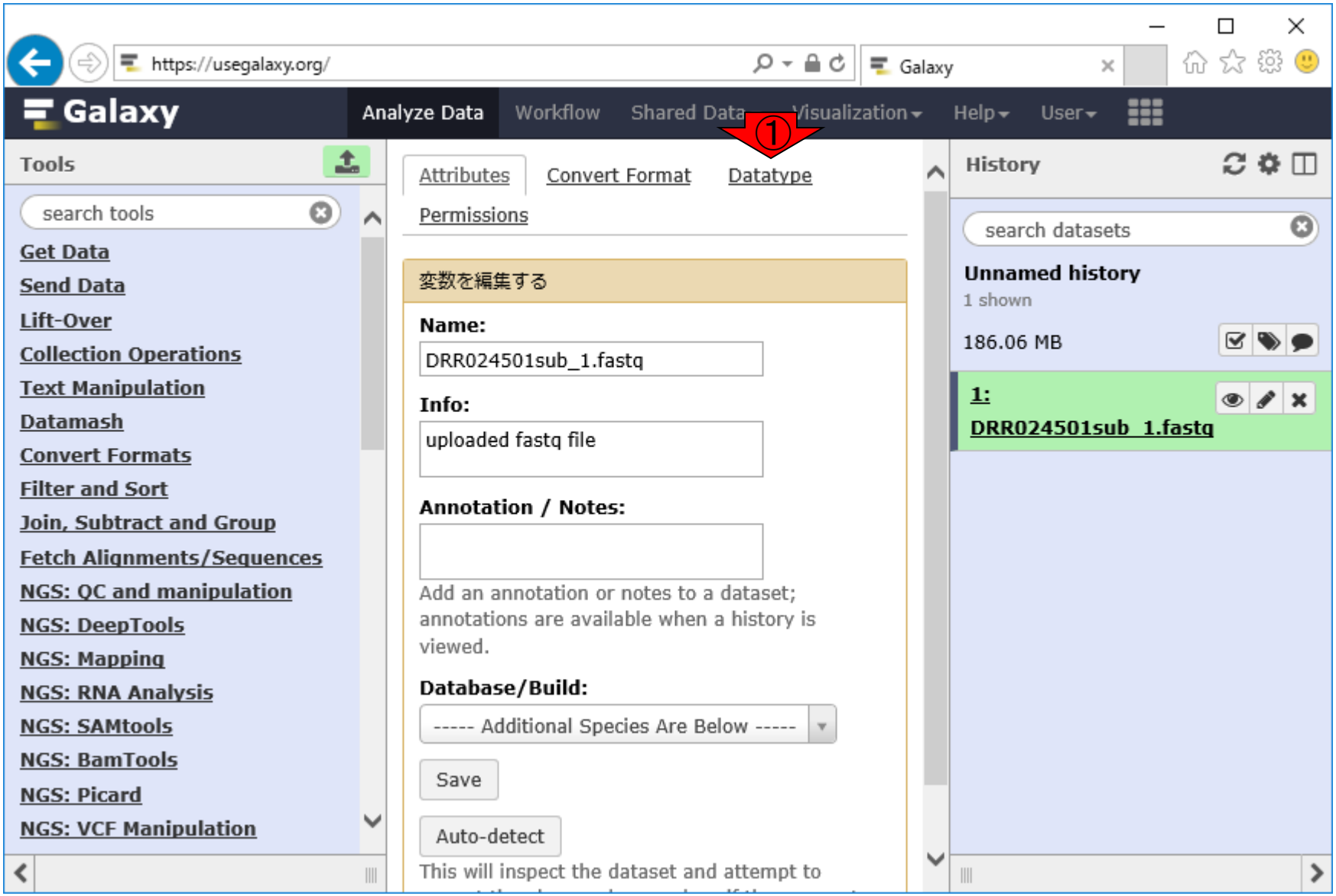
W6-3：編集1

①が編集ボタン。②変数を編集する、の意味が分かりづらいかもしれないが、とりあえずは③ファイル名を変更できたり、④このファイルがFASTQ形式ファイルだと正しく認識されているようだ、という程度でいいだろう



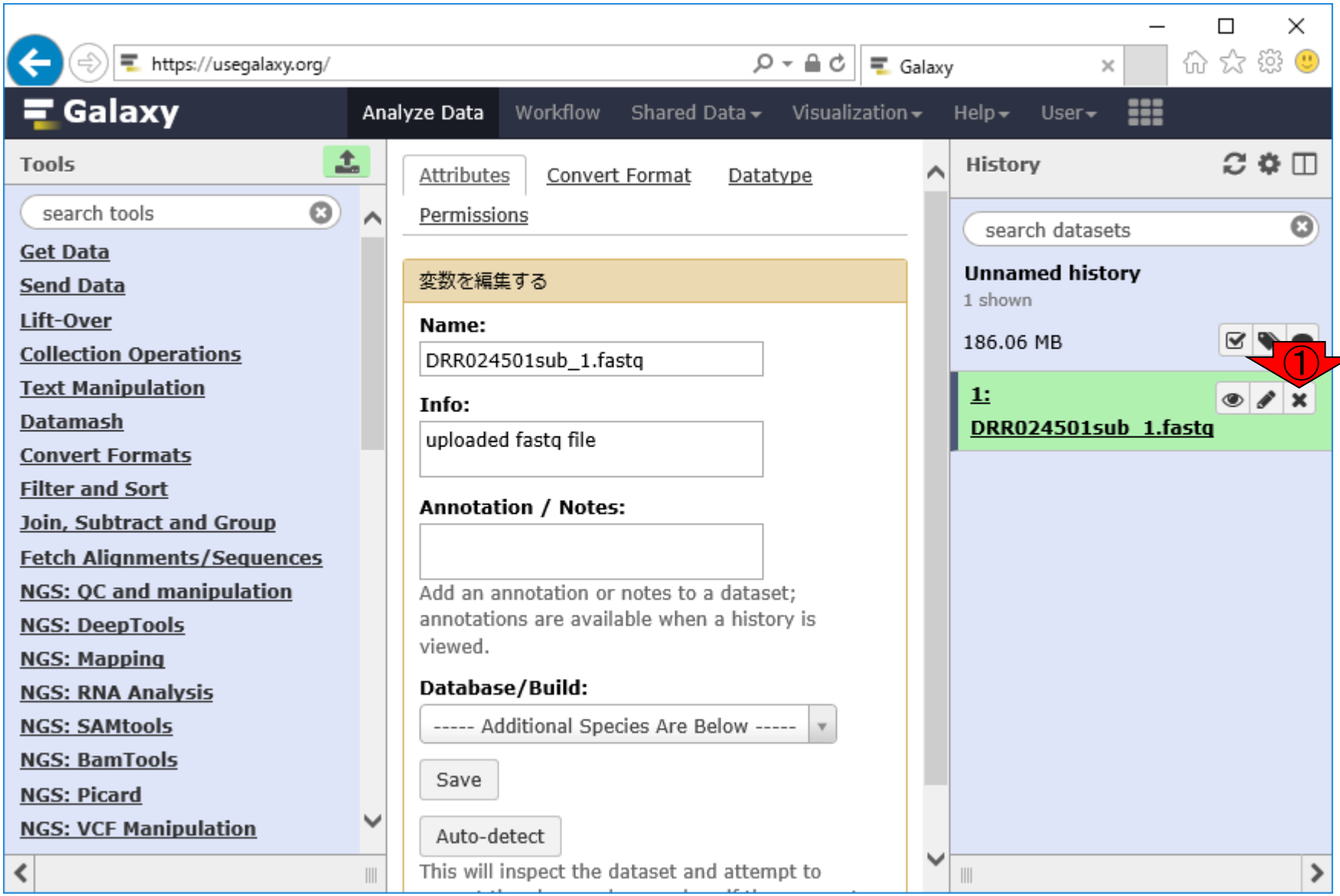
W6-3：編集2

もしファイル形式を間違って認識していたら、①Datatypeのところで手動で修正する



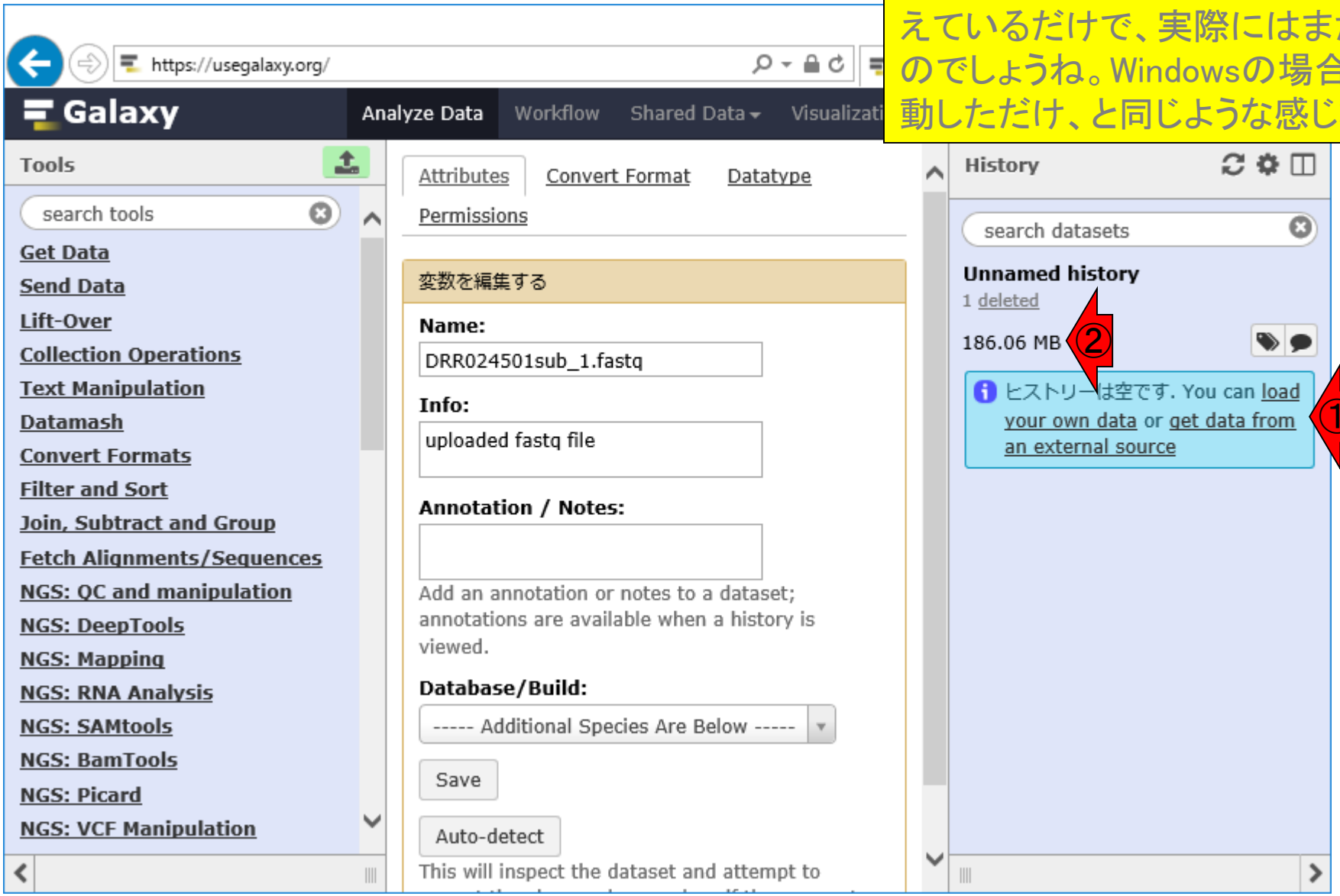
W6-4：削除

①×ボタンで削除することができるようだ。やってみる



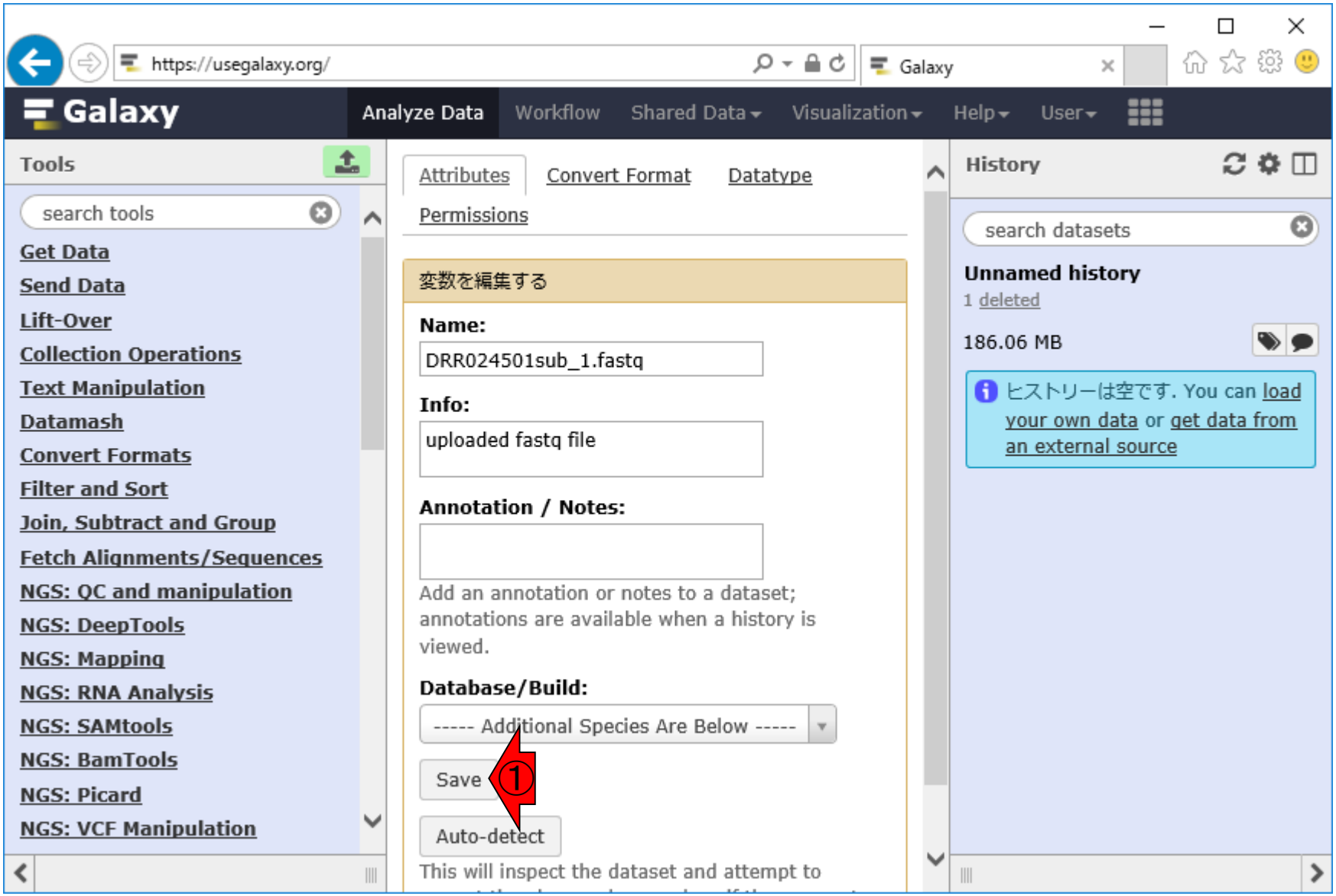
W6-5：削除されました

①ほんとに消していいかという確認もなしに消えるんですね…。しかし②ディスク容量は不変なので、ヒストリーパネル上からは消えているだけで、実際にはまだ残っているのでしょうね。Windowsの場合はゴミ箱に移動しただけ、と同じような感じなのでしょう



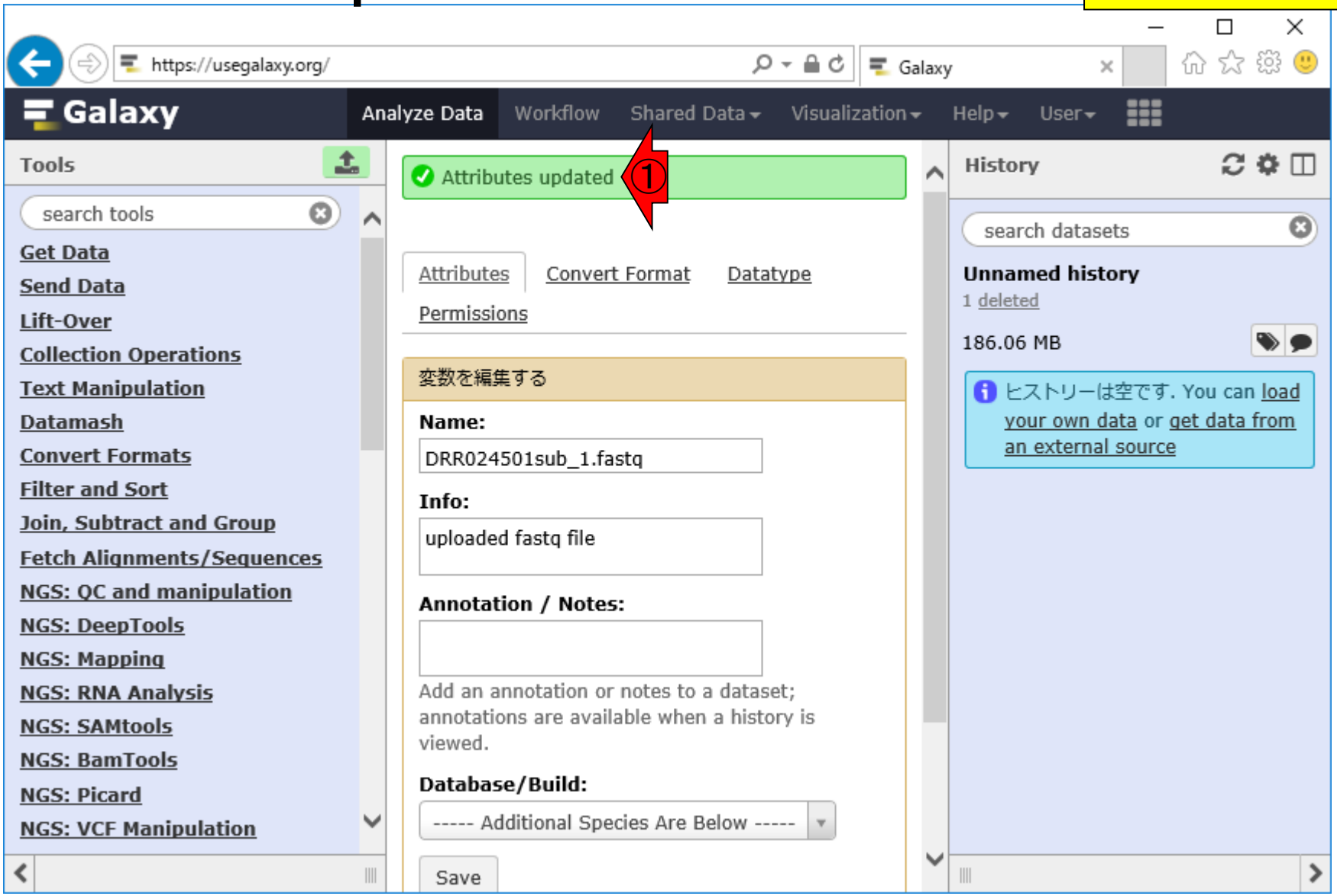
W6-7：Saveしてみる

ファイルが存在しないはずだが、①Saveボタンを押してみる



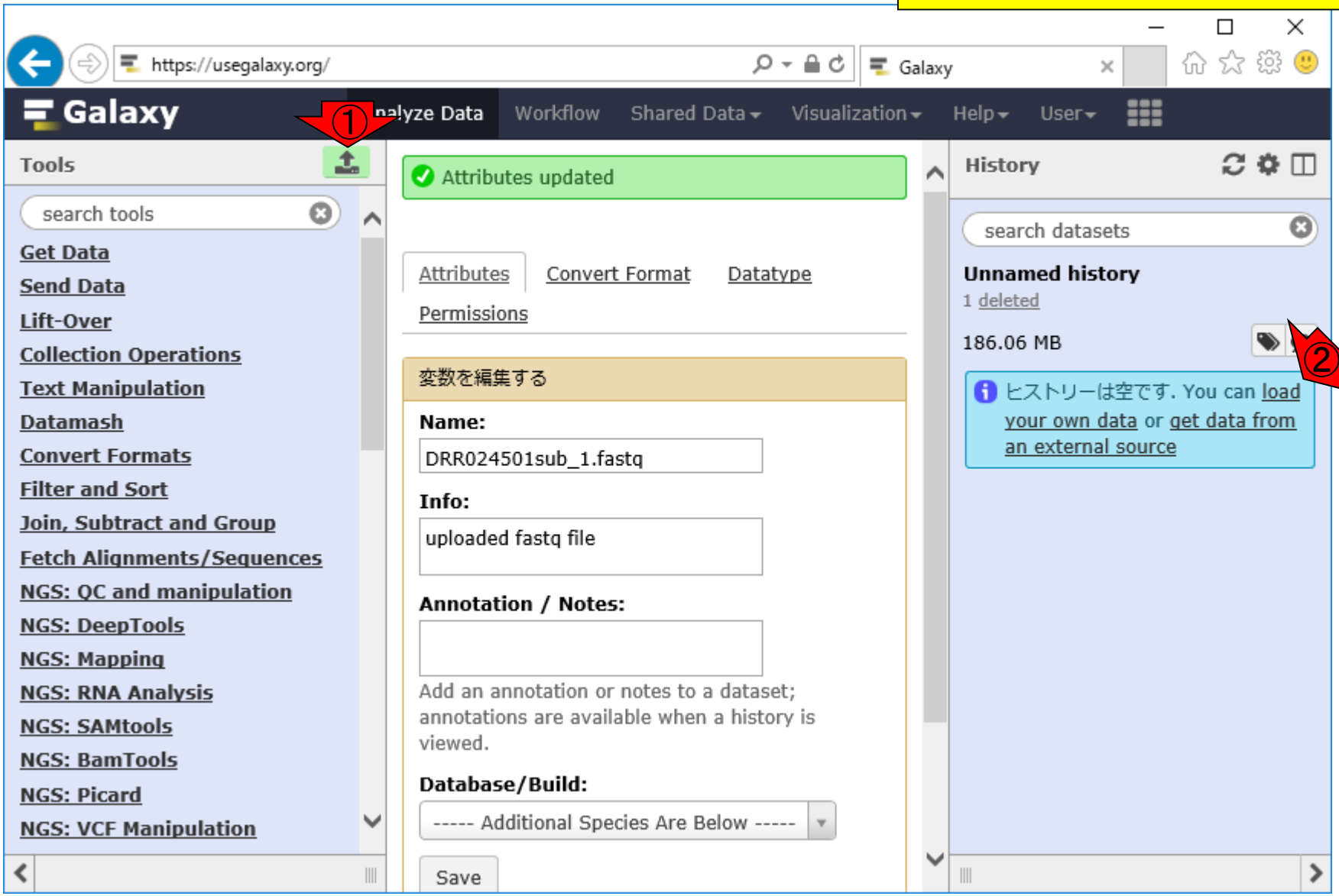
W6-8：updated…

①Attributes updatedなので、
やはりヒストリーパネル上では
消えているだけなのでしょう



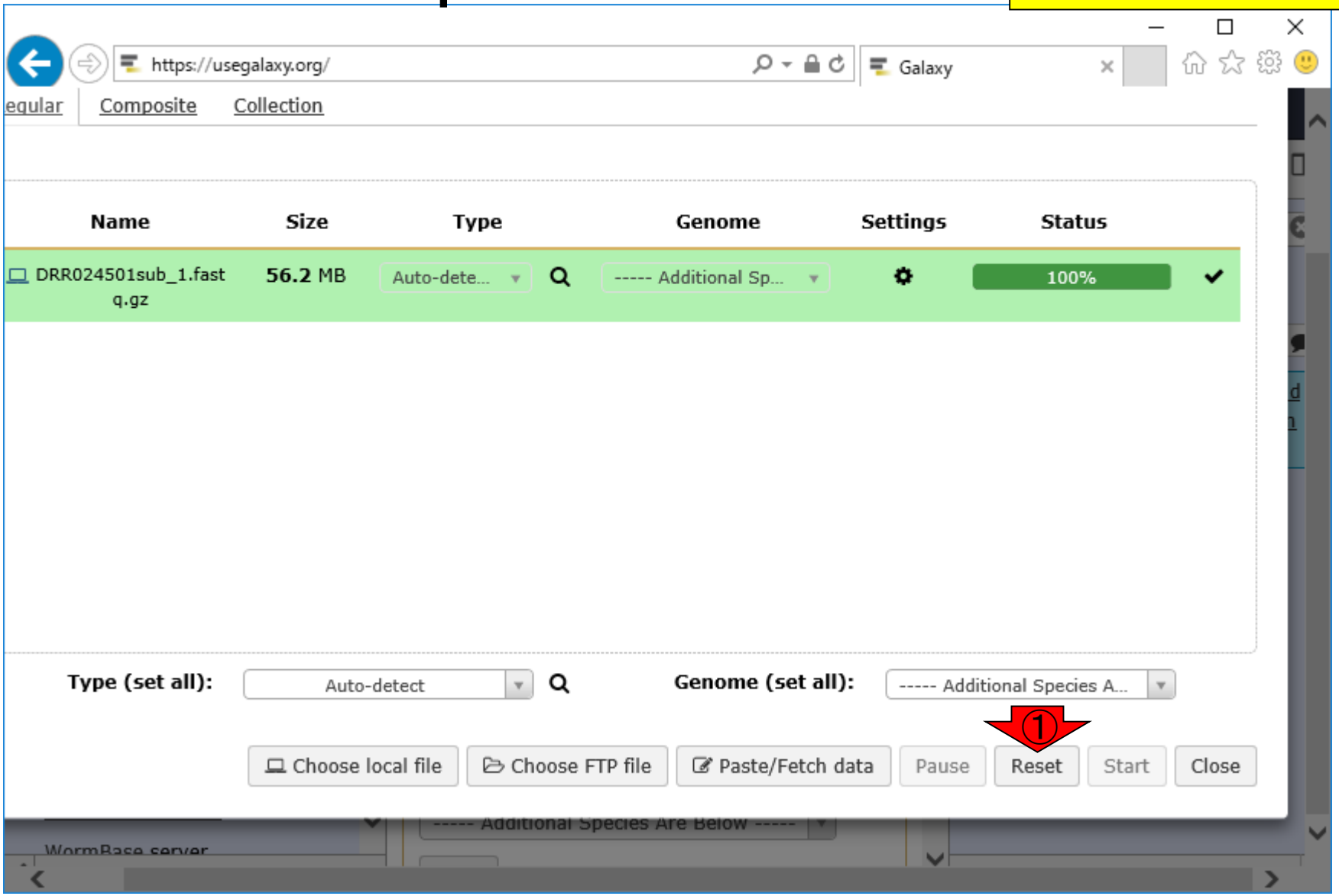
W7-1：再アップロード

W5-3で利用した①アップロードボタンではなく、②load your own dataのところを押してもおそらくいいのだろう。クリックしてみる



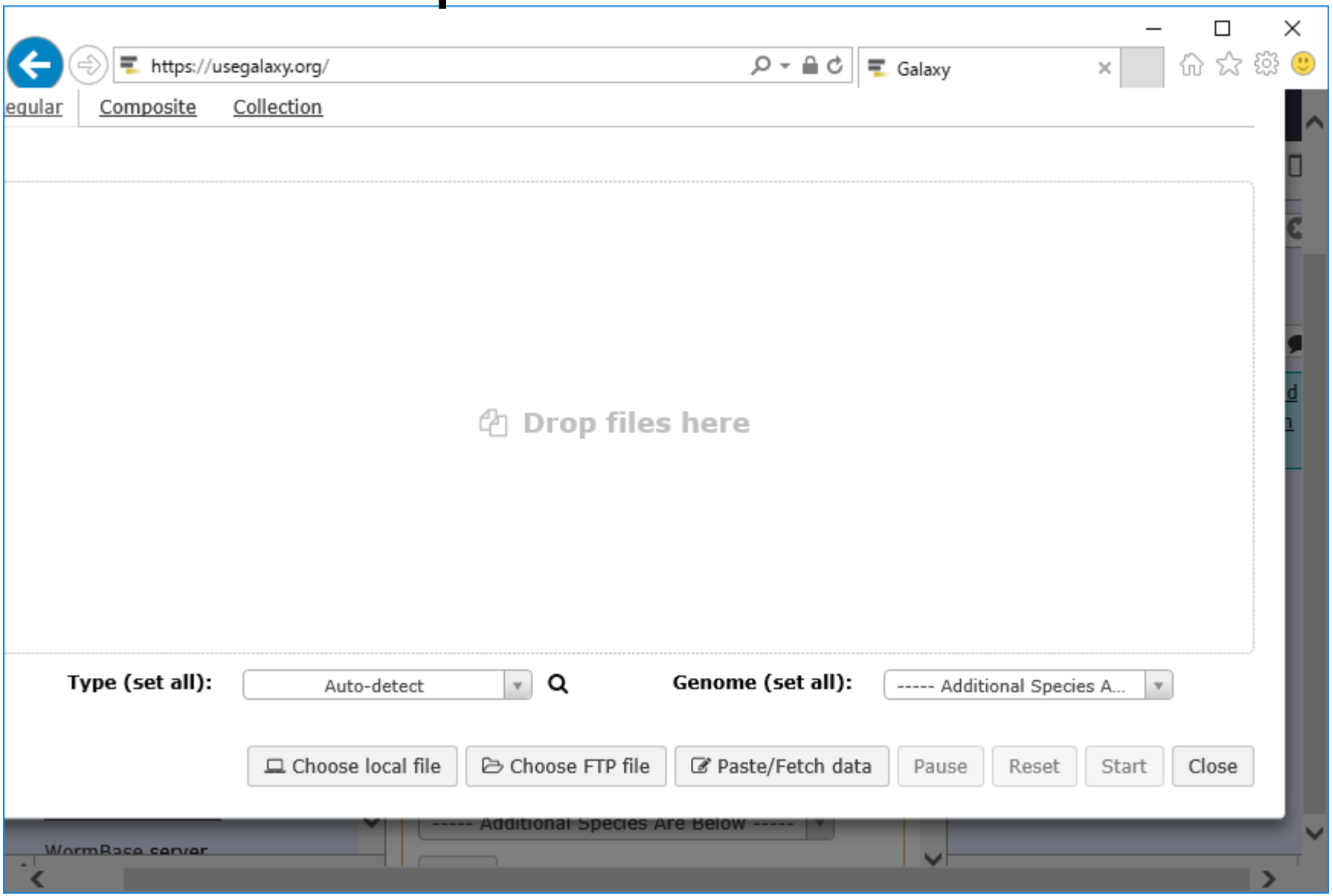
W7-2：Drop files here

W5-6と同じような画面になりました。このままでは追加アップロードできないようなので、一旦①Reset



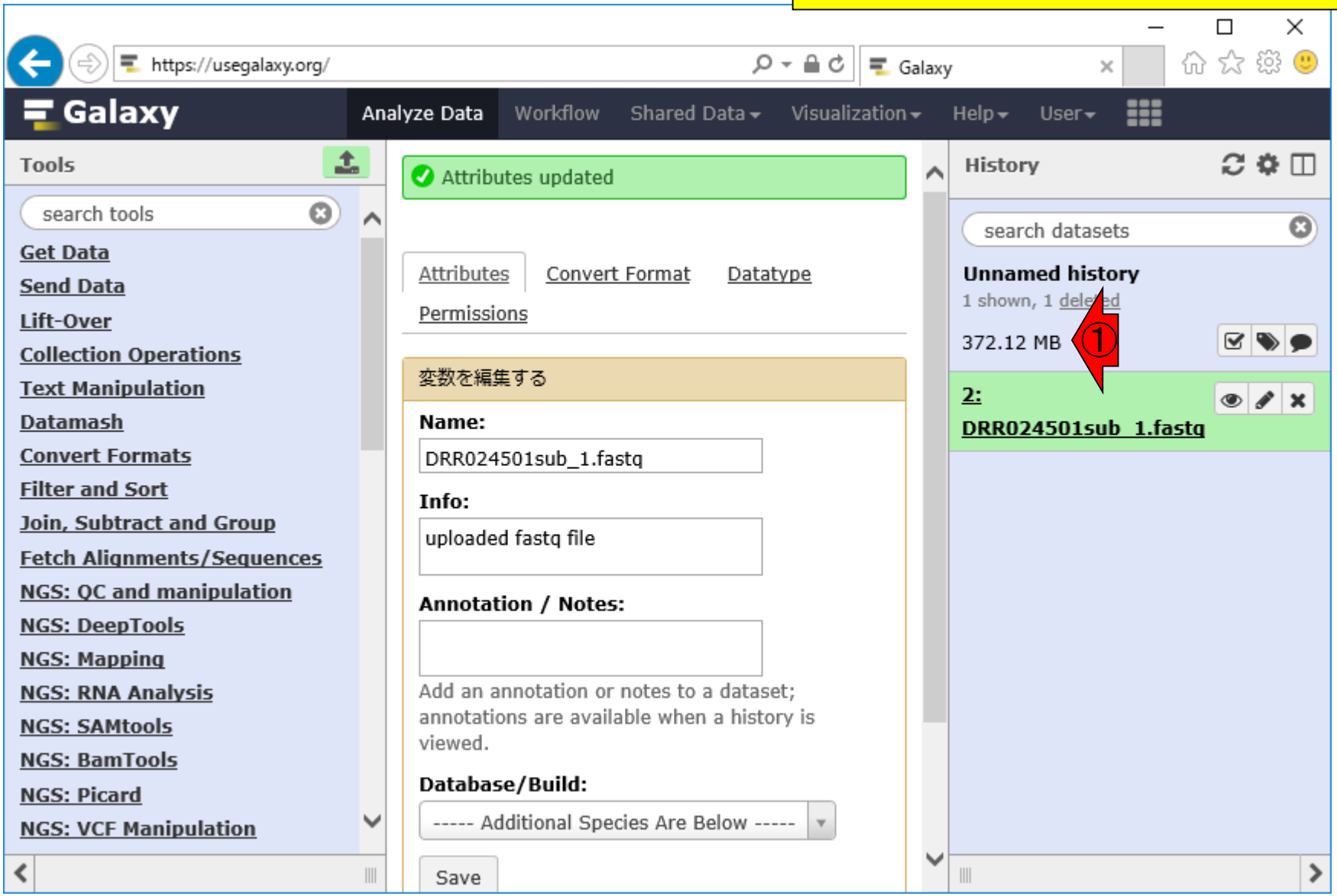
W7-2：Drop files here

Reset後の状態。この状態から、W5-3と同じ手順でもう一度同じファイルをアップロード



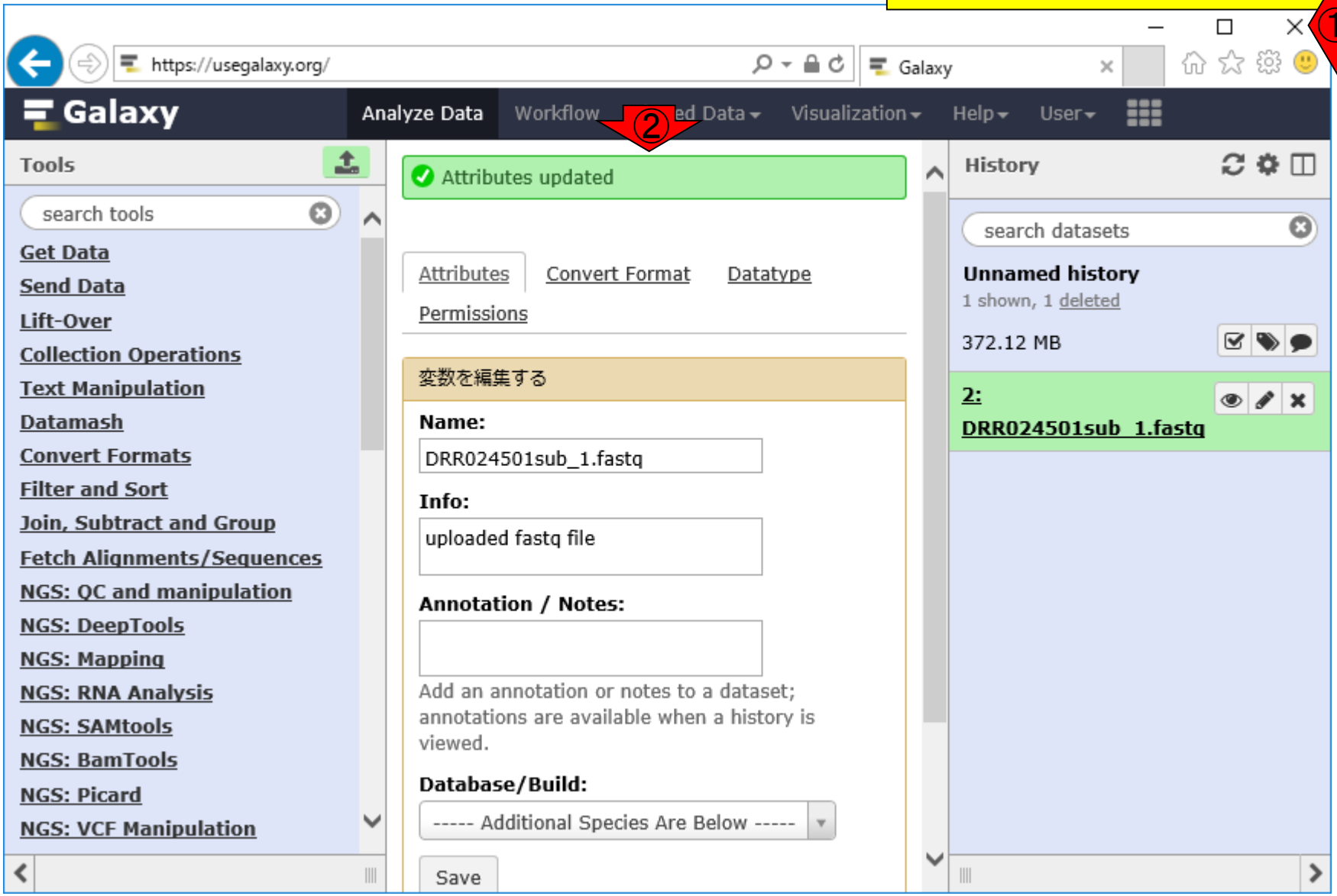
W7-3：アップロード後

①186.06 × 2 = 372.12MBと2つ分のファイルサイズになりました。同じファイルですが、上書きではなく別個のものとして認識されているようですね



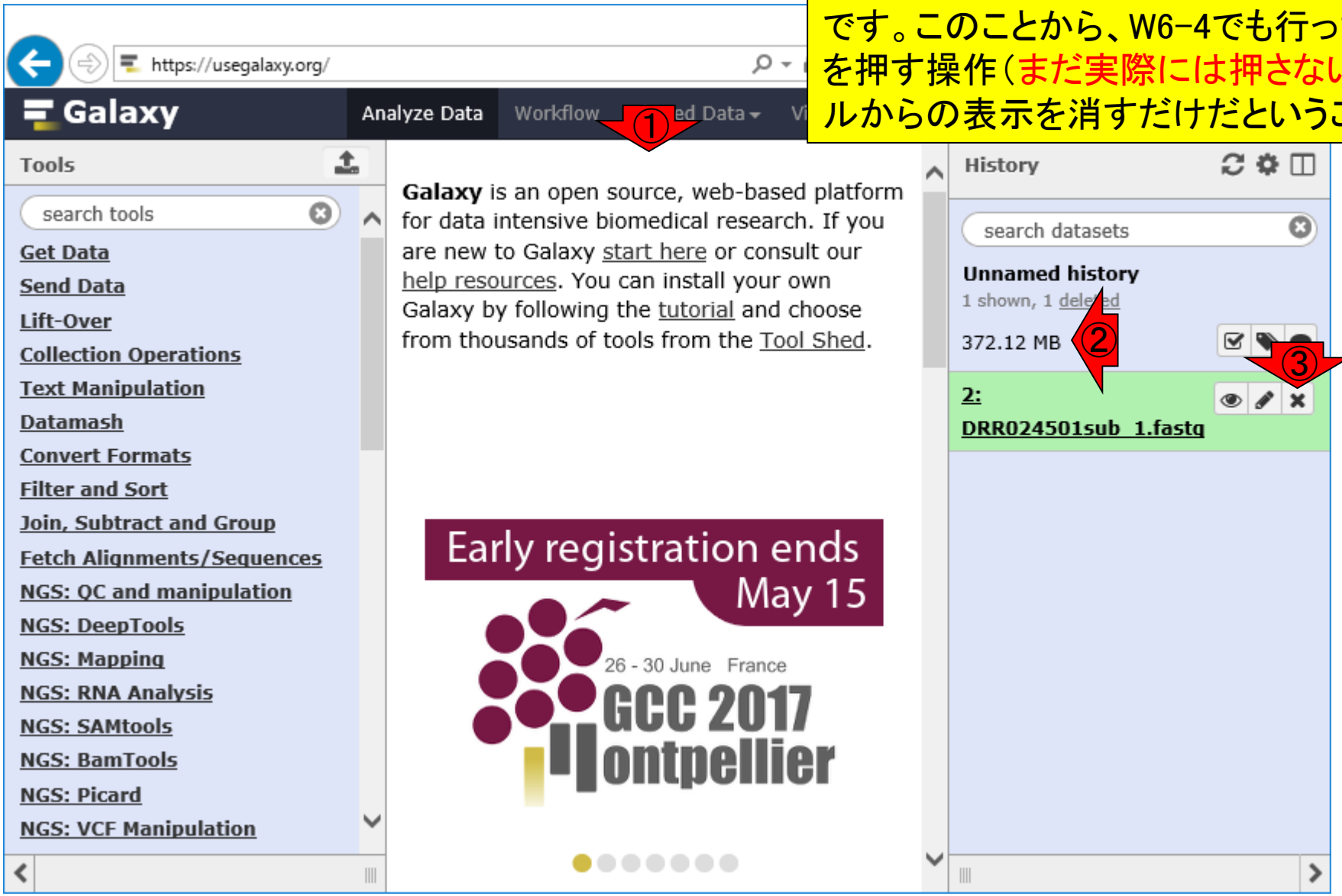
W7-4：一旦終了

①を押して一旦終了させてみる。目的は、②真ん中のパネル表示が再度ログインすると消えるのではないかという想像の確認



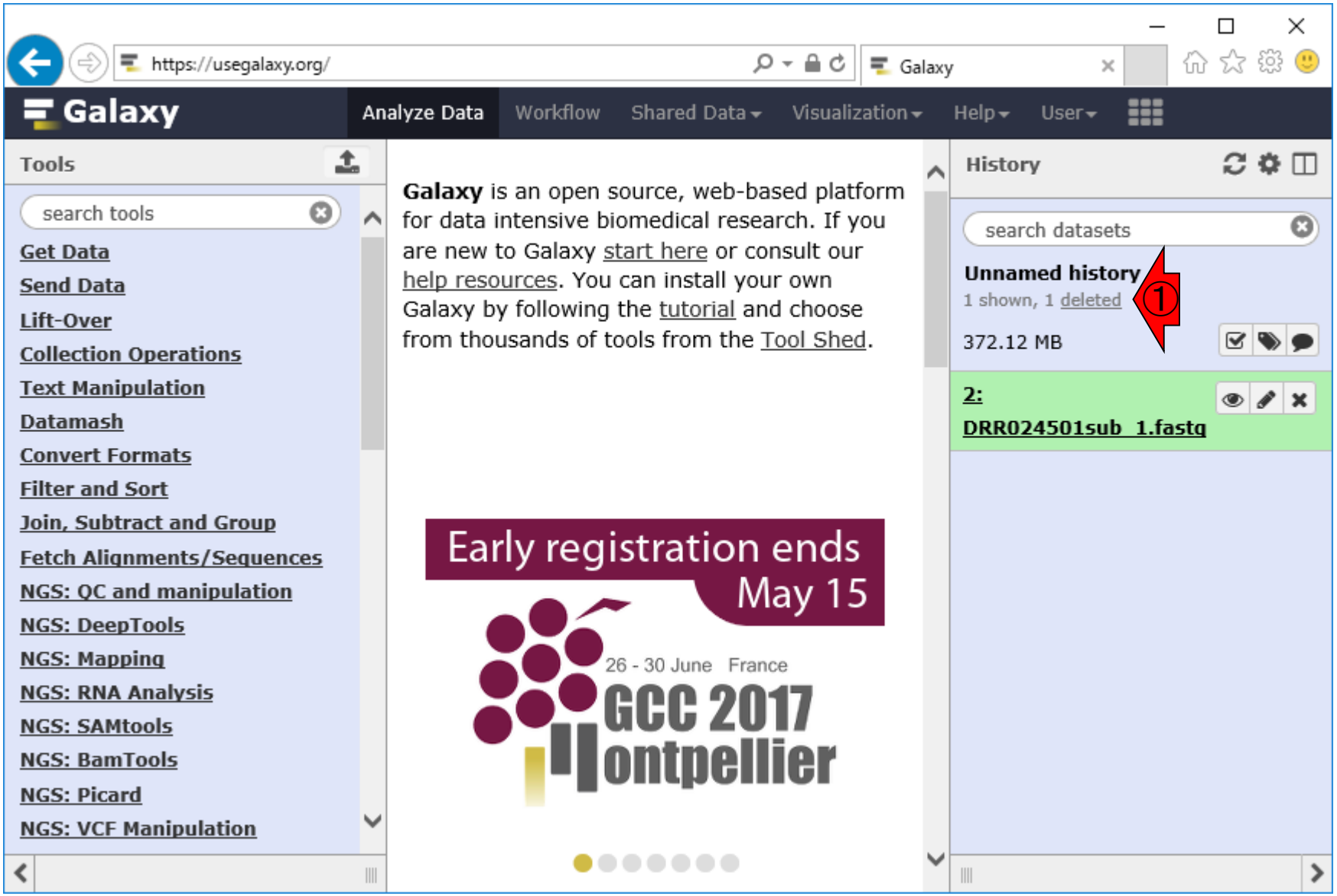
W7-5：再度アクセス

①真ん中のパネル表示が消えてます。②右側のヒストリ（履歴）パネルは、186.06×2＝372.12MBと2つ分のファイルサイズになったままです。このことから、W6-4でも行った③×ボタンを押す操作（まだ実際には押さない）は履歴パネルからの表示を消すだけだということを実感



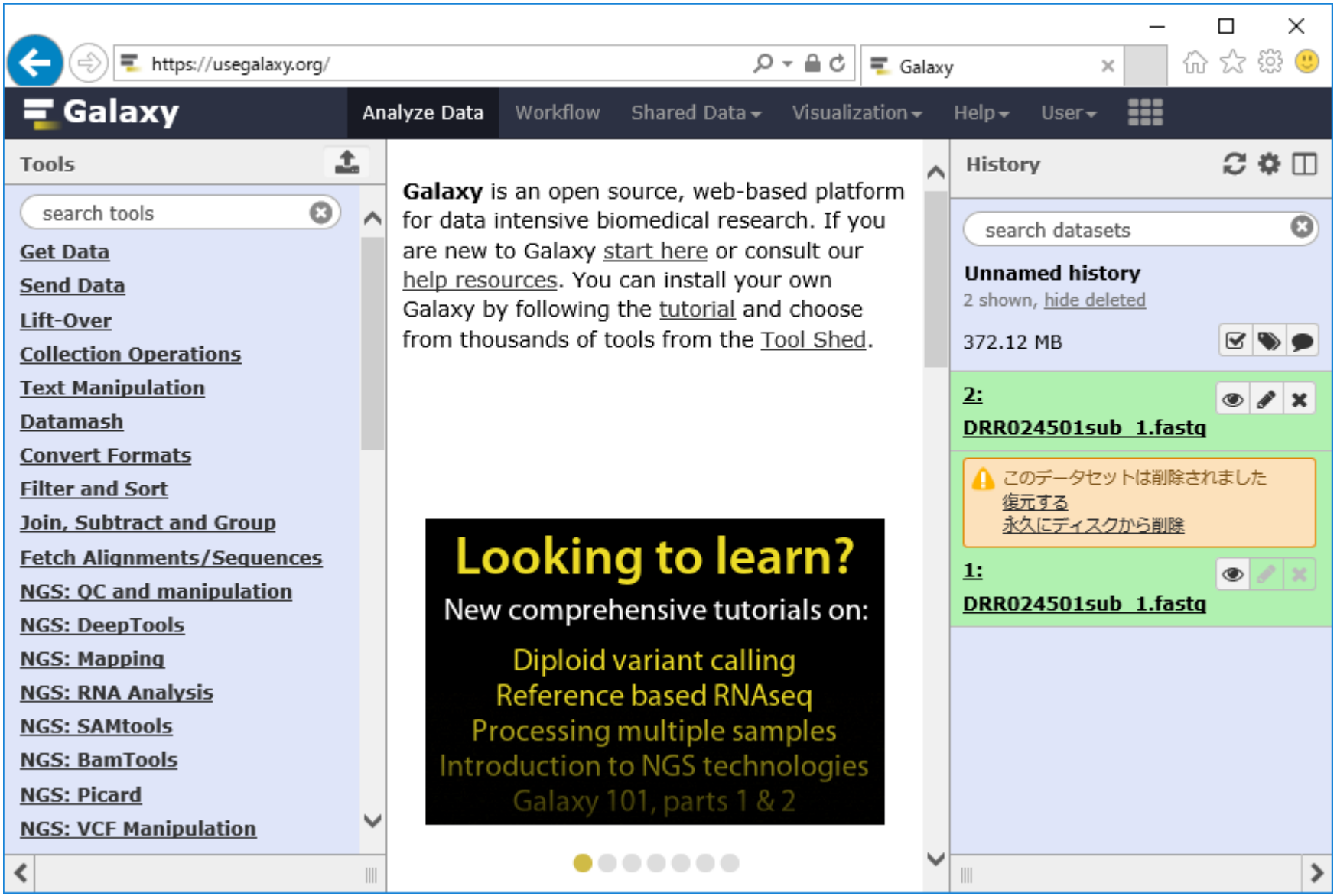
①deletedのところを押してみると…

W7-6：deletedを押してみる



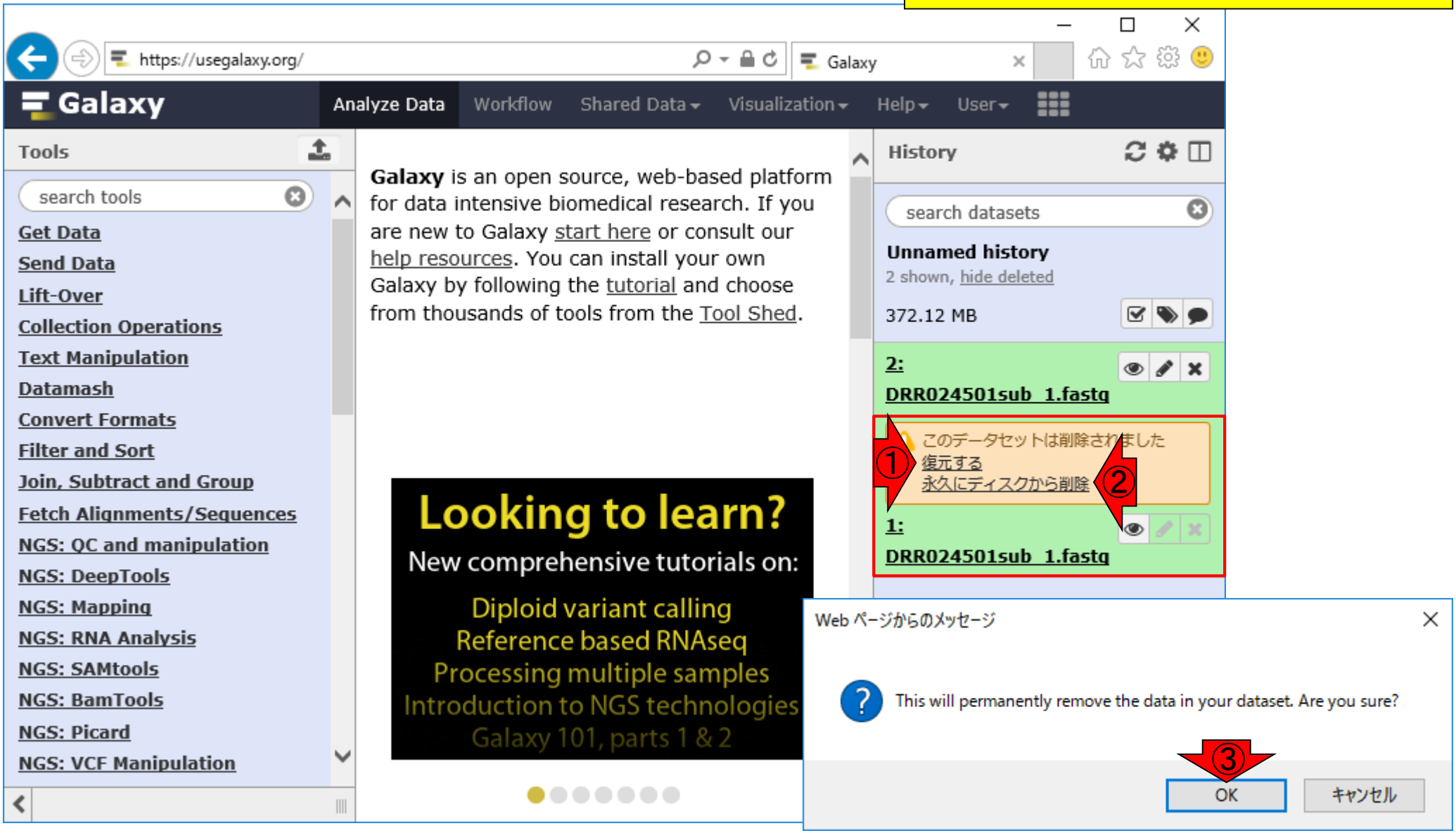
こんな感じになります

W7-6：deletedを押してみる



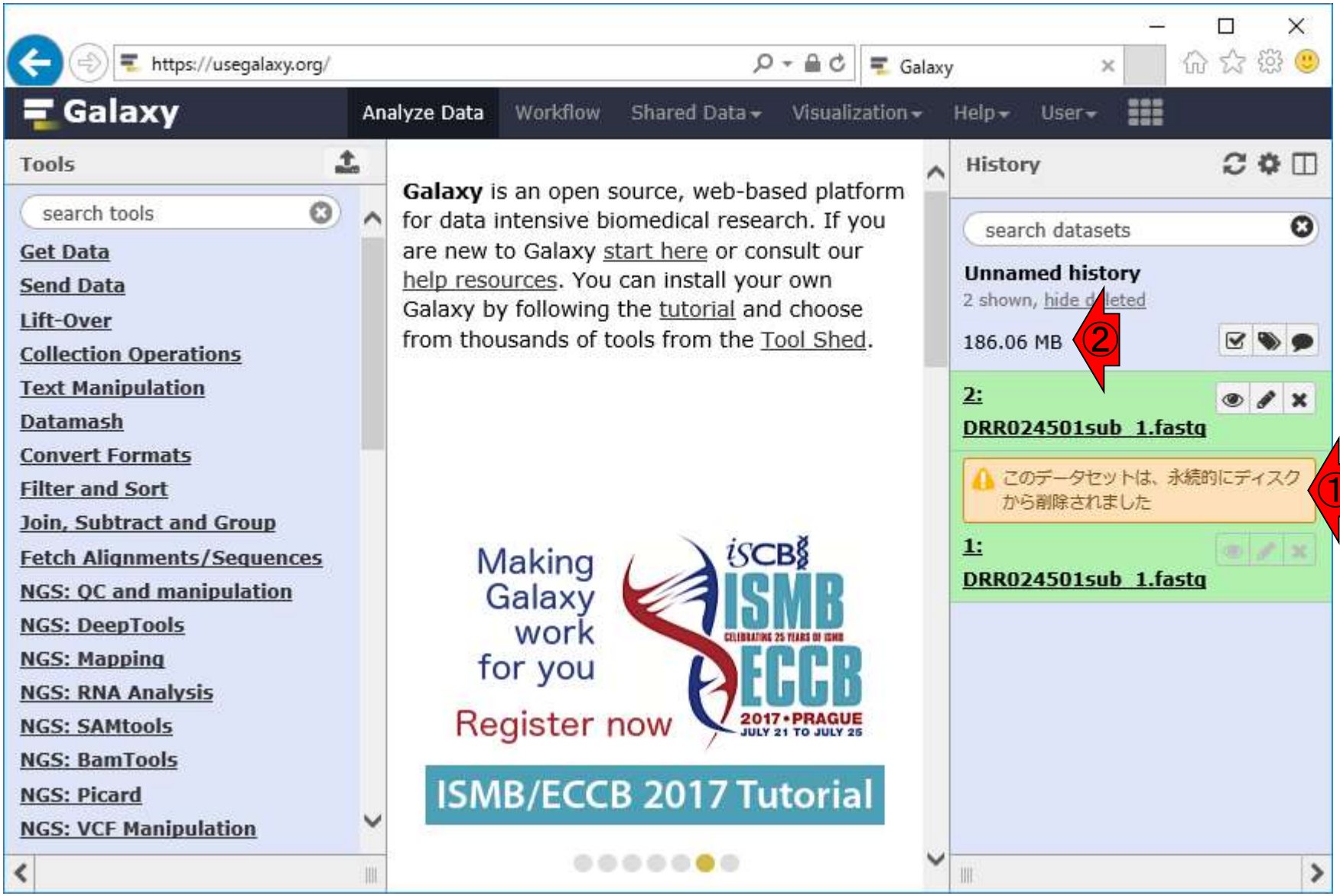
W7-7：永久に削除

①復元することも、②永久にディスクから削除することもできるのですね。ここでは②削除してみます。③OK



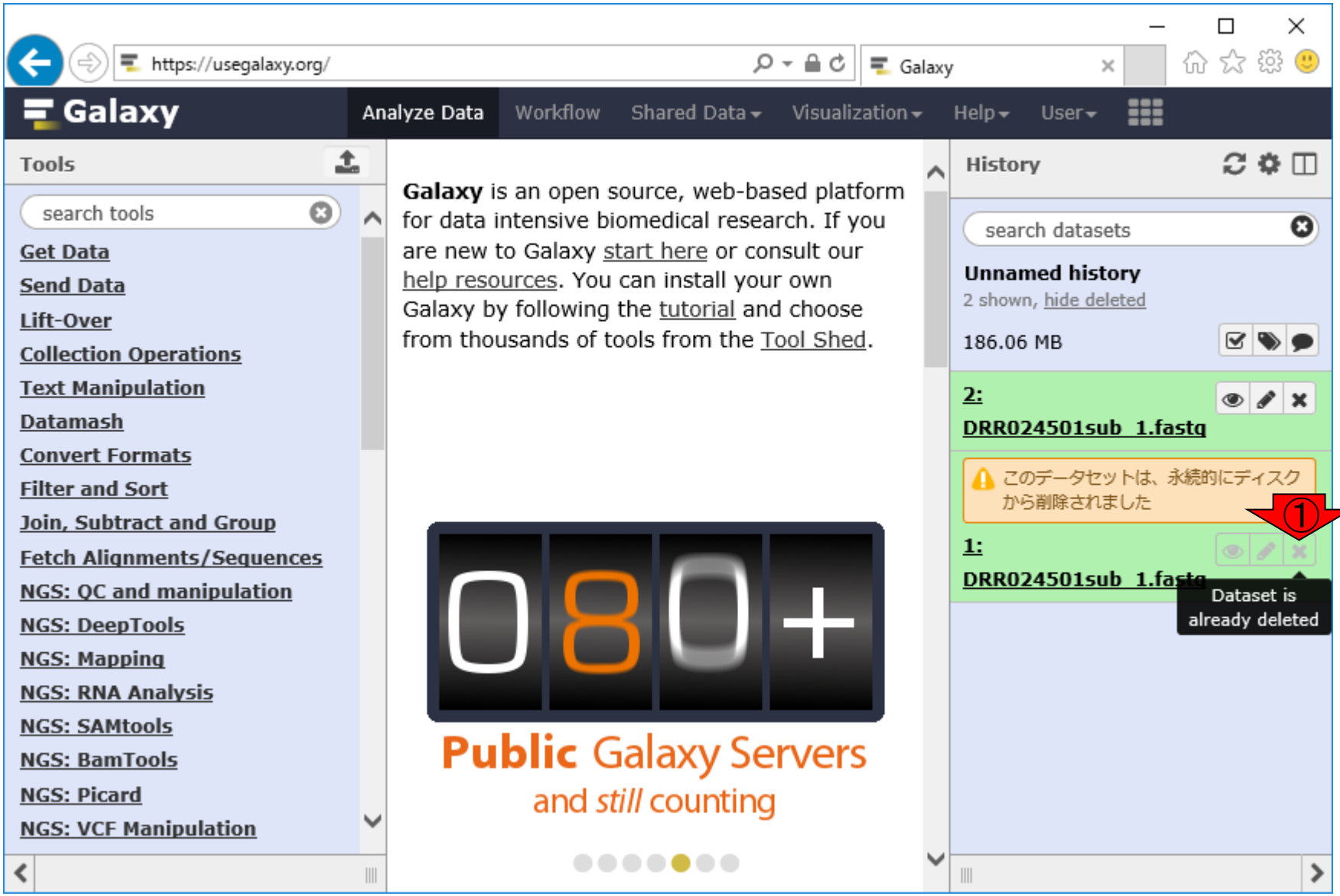
W7-7：永久に削除

①永続的に削除されたようです。②ディスク使用量も確かに減りました



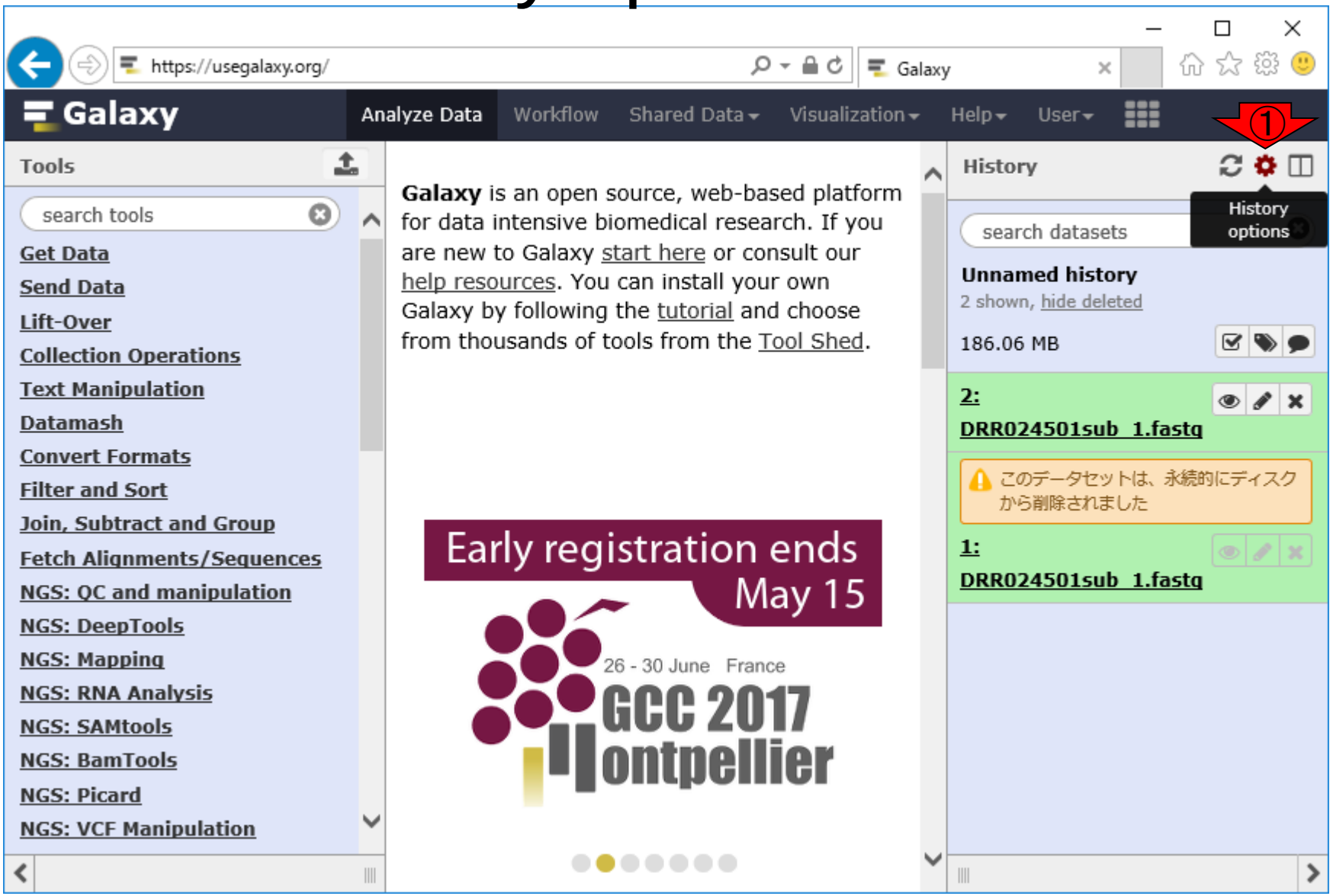
W7-8：お前は既に…

①消えているので、消すことはできません



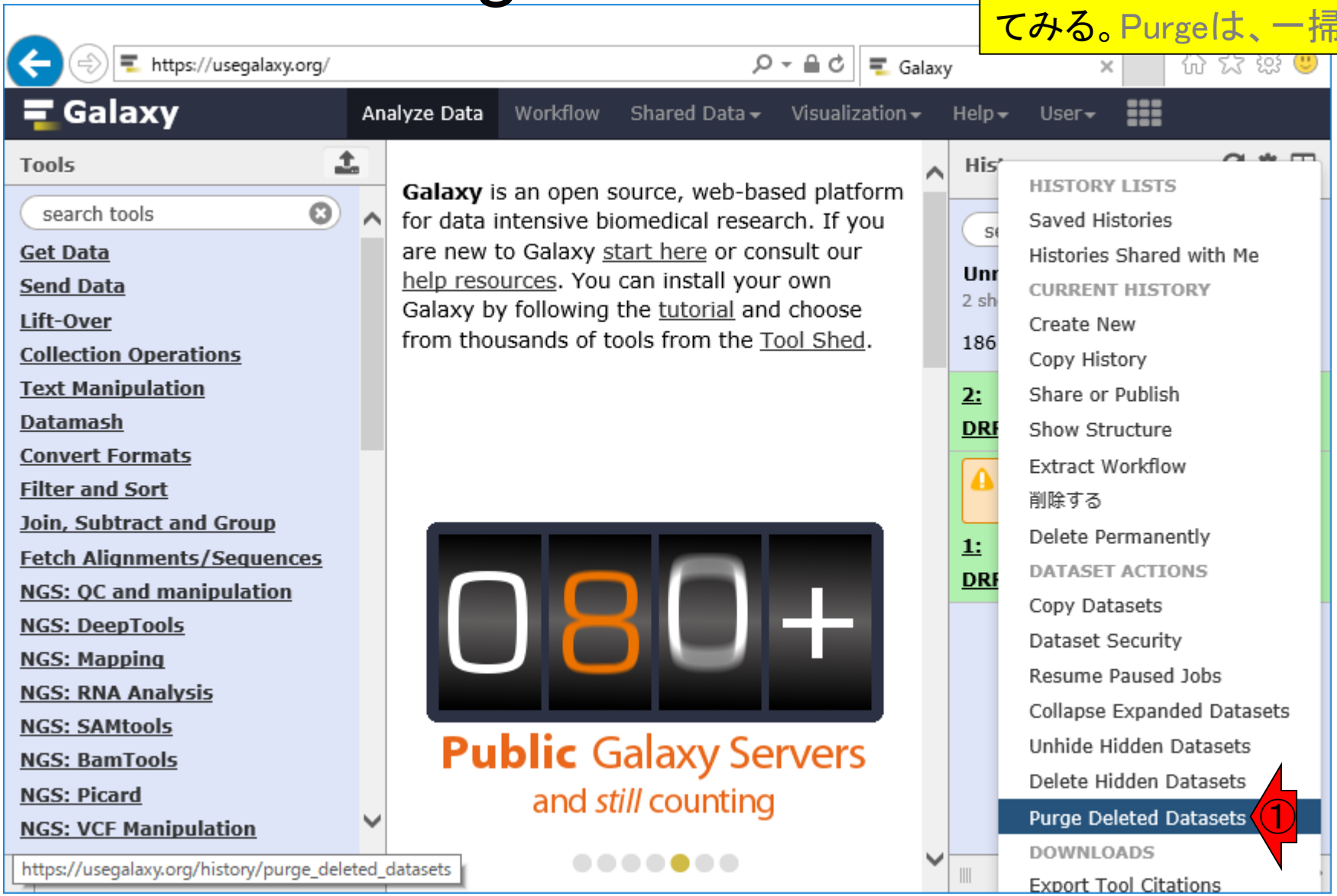
W7-9：History options

①この歯車マークのところがHistory optionsです



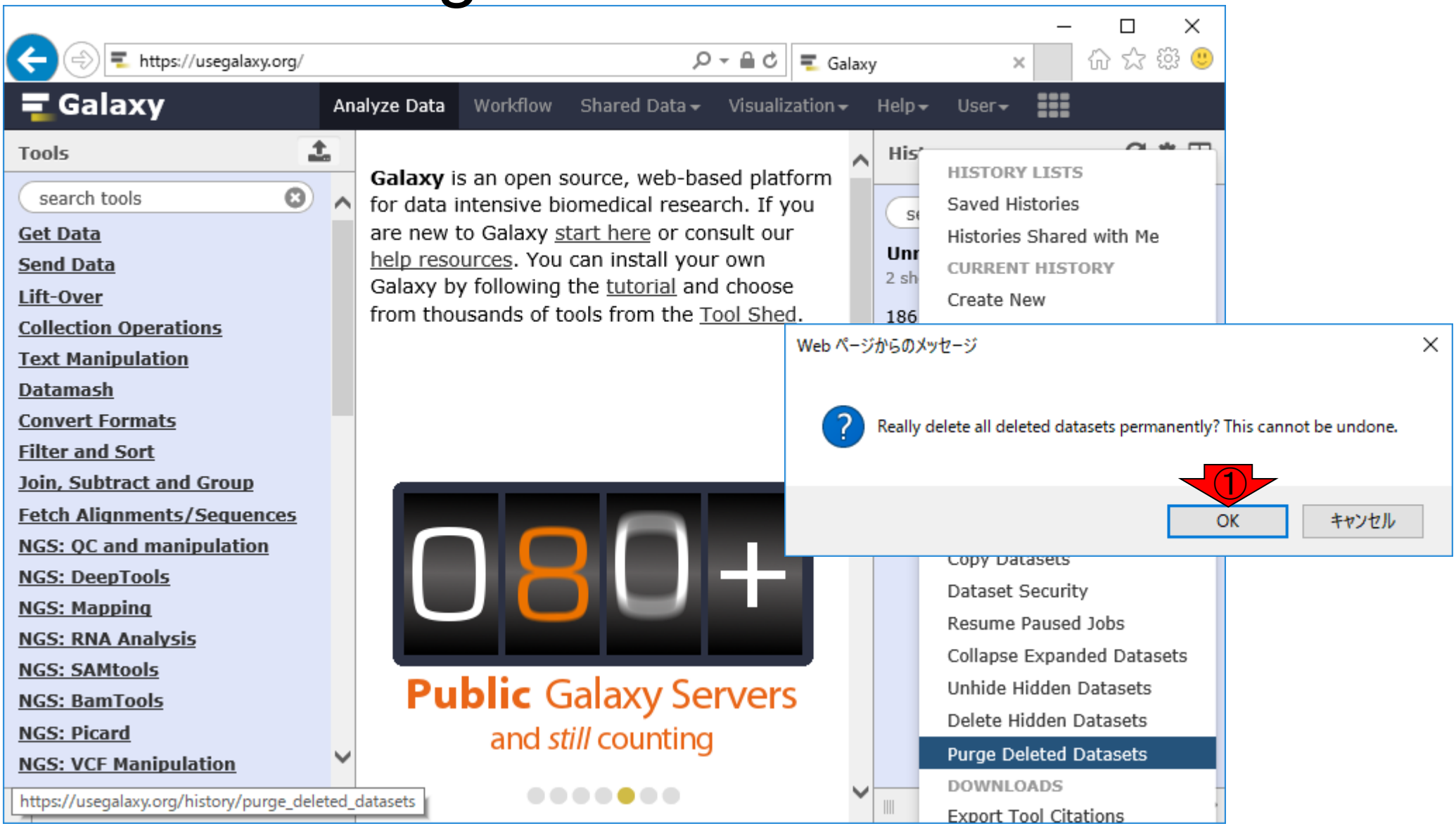
W7-10：Purge Deleted …

①Purge Deleted Datasetsは、W7-7で行った「永久にディスクから削除」と同じ作業なのだろうと思いつつ、やってみる。Purgeは、一掃する、の意味



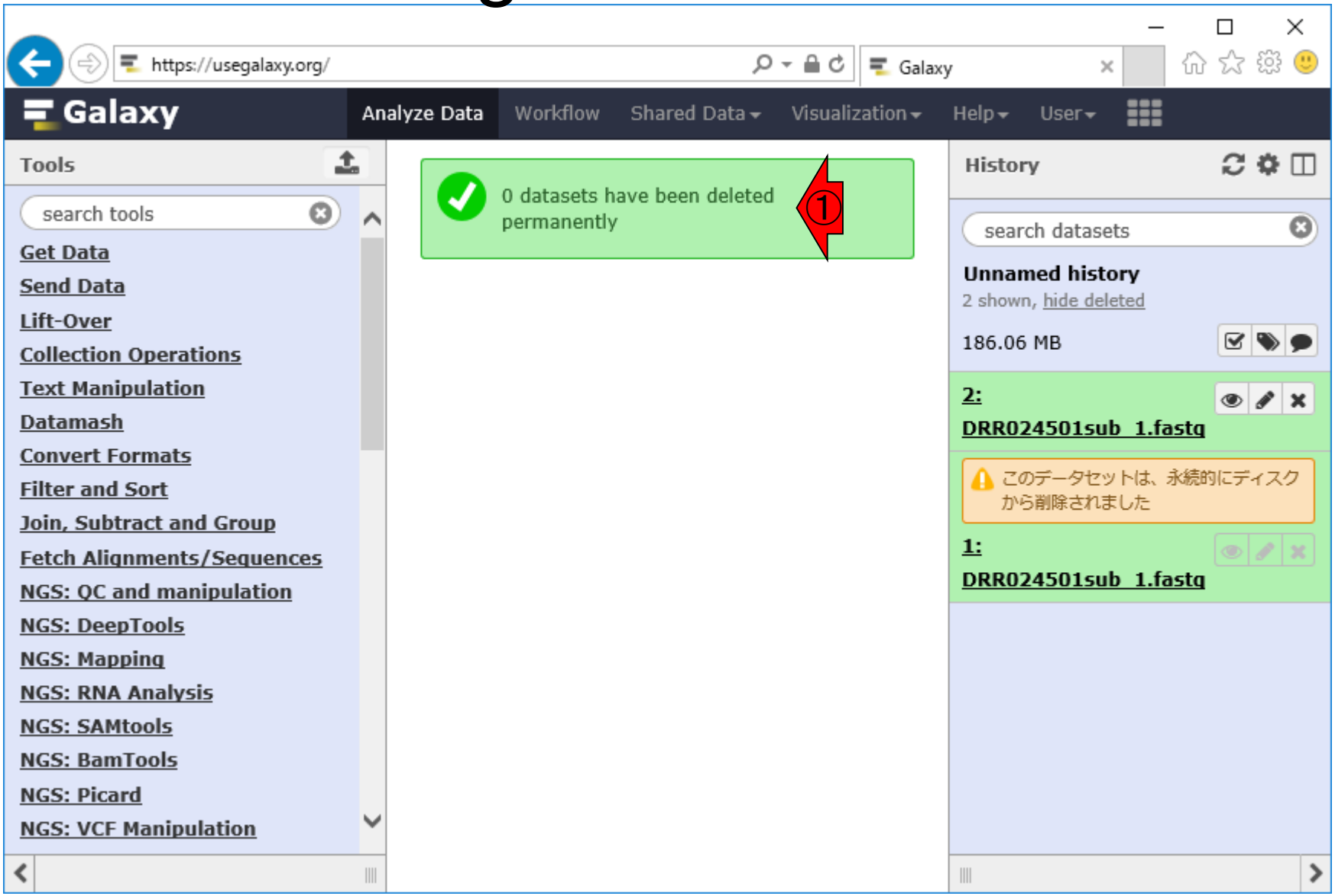
①OK

W7-10：Purge Deleted …



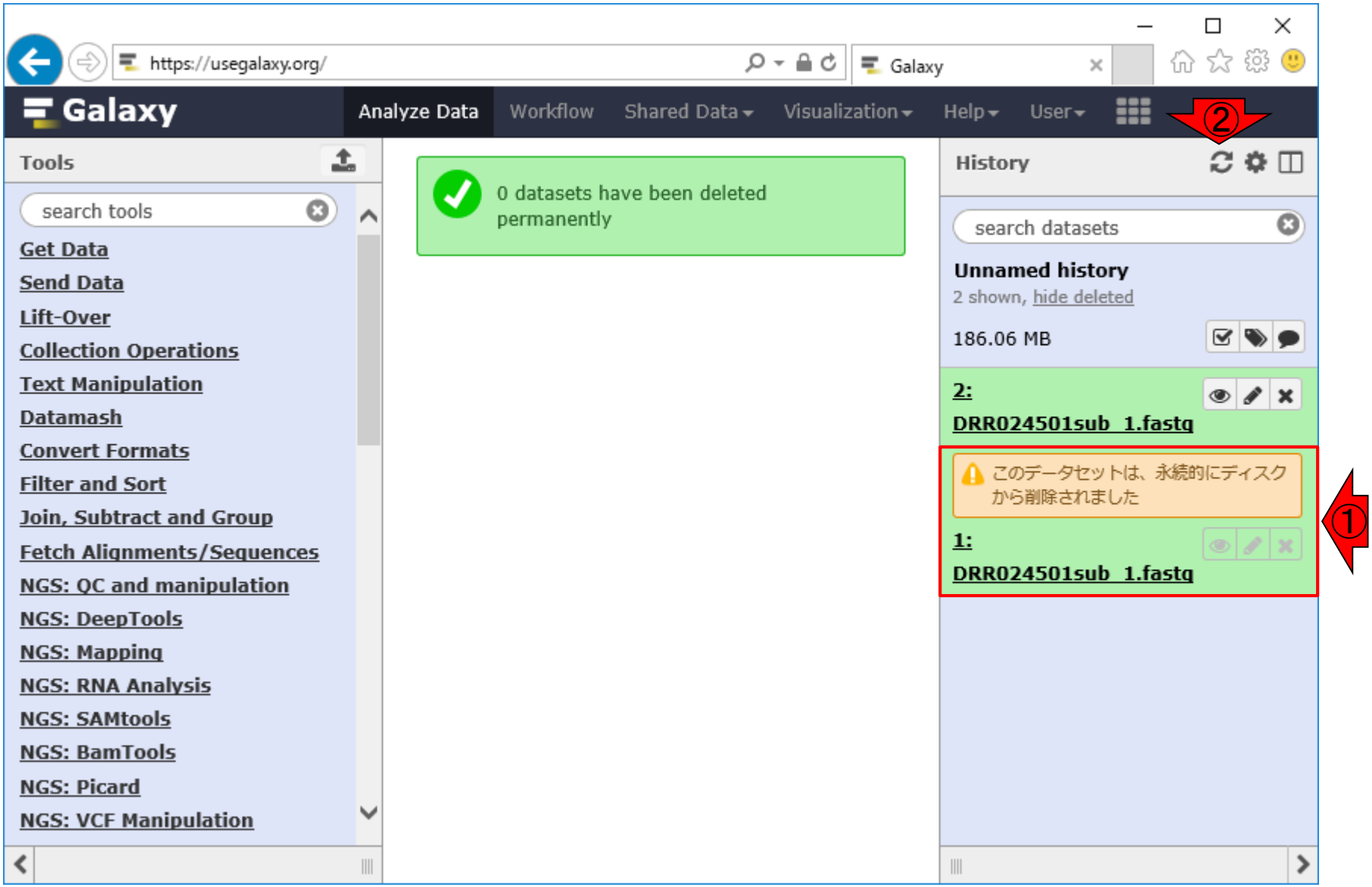
W7-10：Purge Deleted …

①永続的に削除されたデータセットは0です。妥当ですね



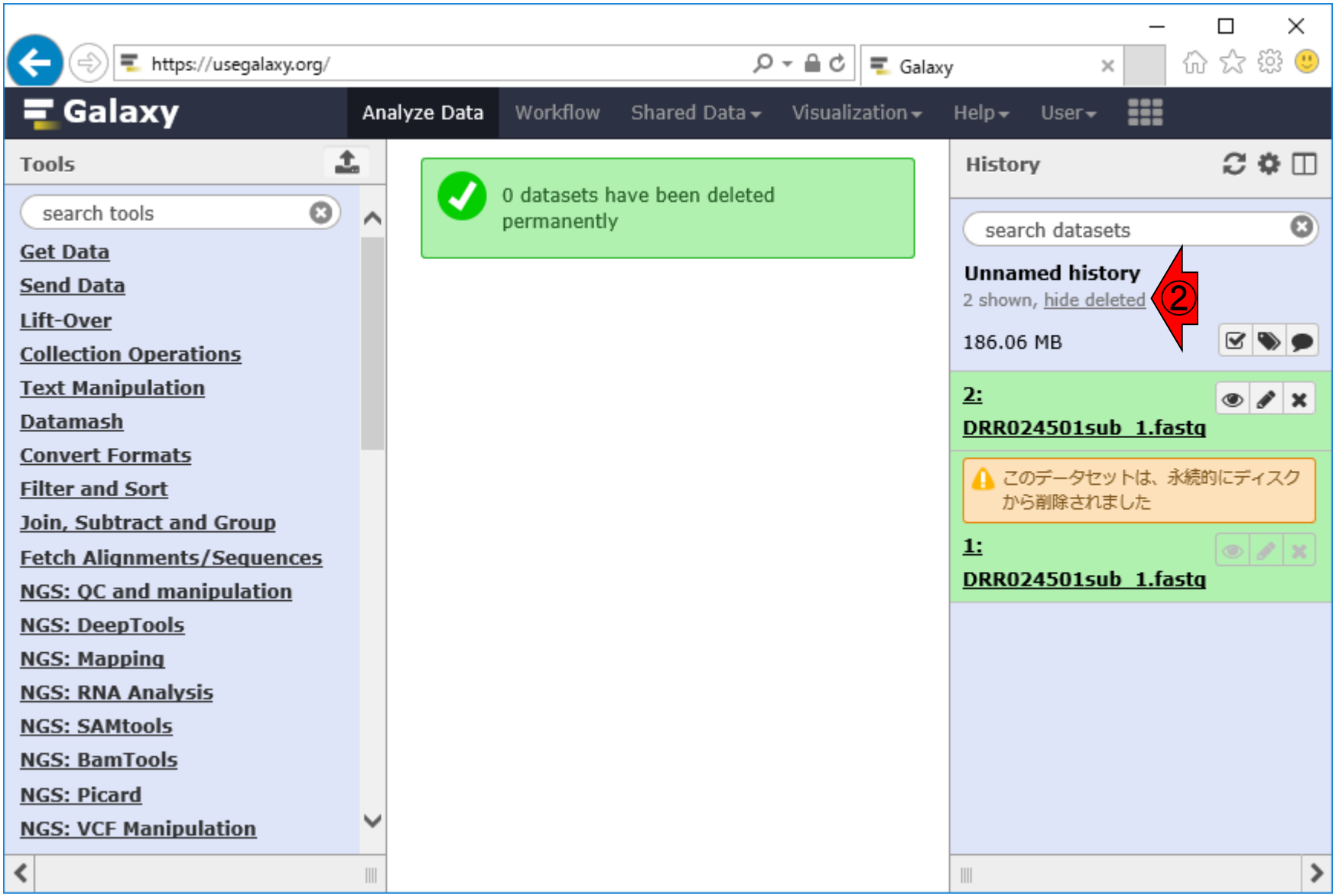
W8-1：ヒストリー画面

①が消えることを期待して、②のRefresh historyを押してみる



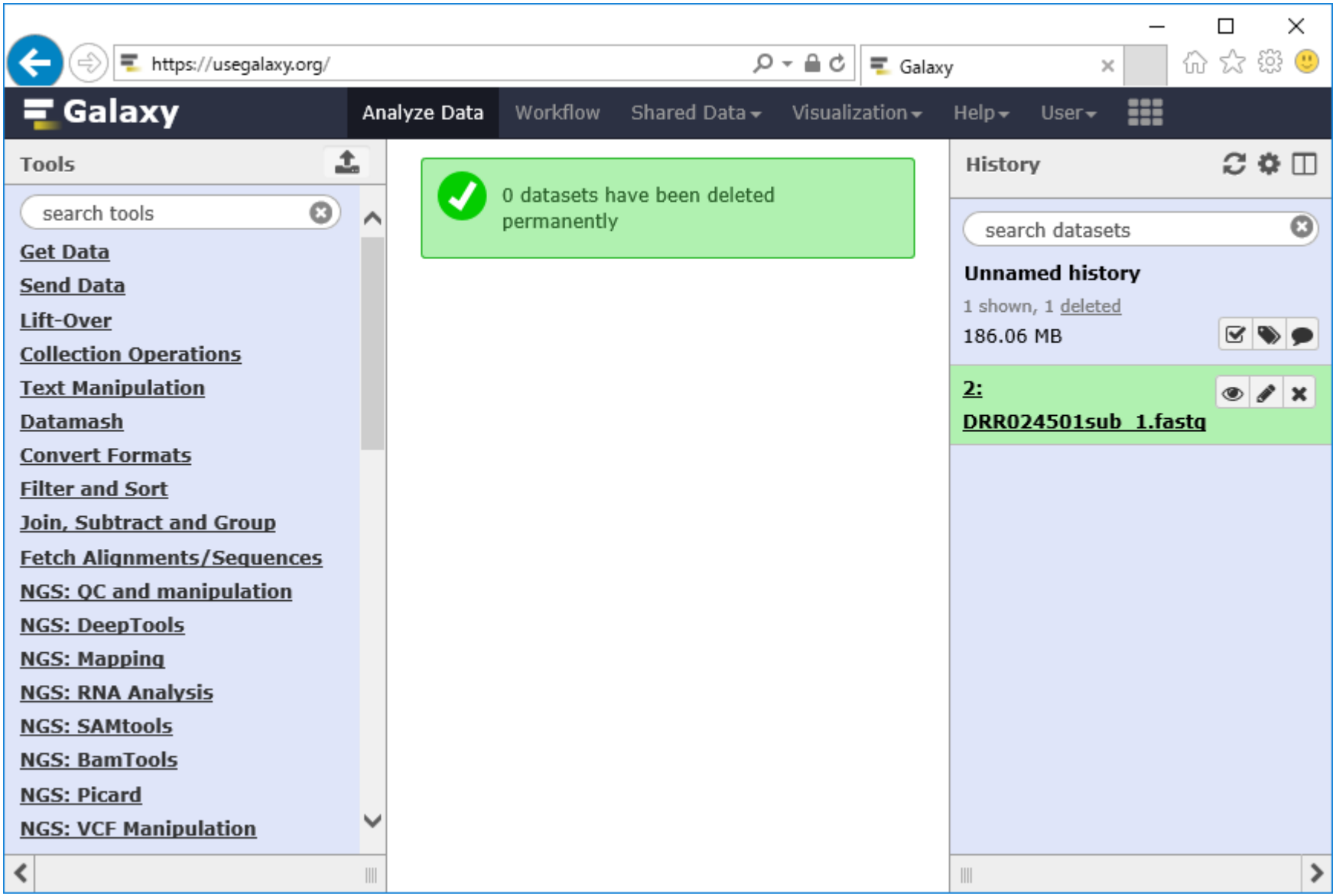
W8-1：ヒストリー画面

変化なし。①hide deletedで、削除されたものを非表示にすることはできます



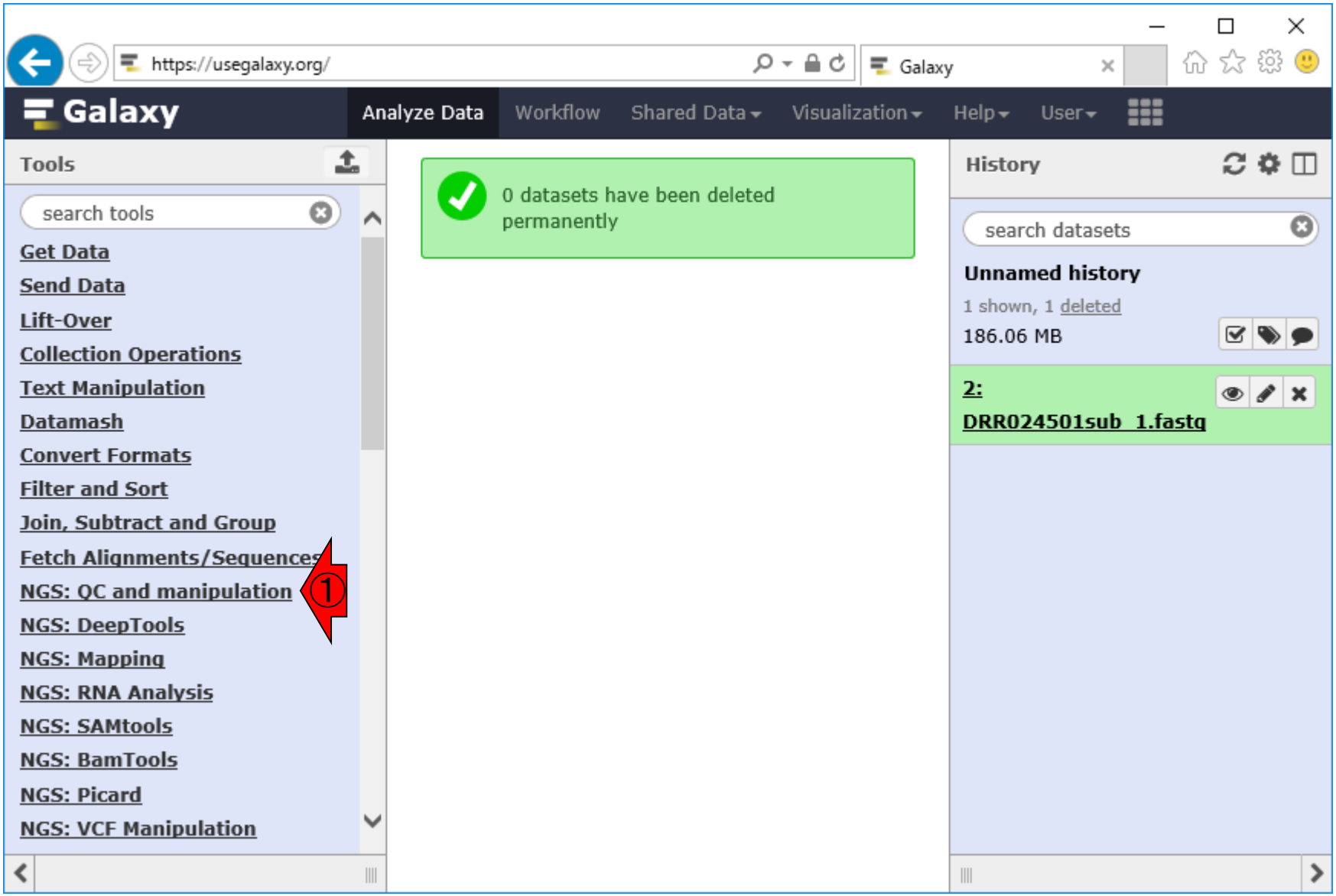
こんな感じになります

W8-2：非表示後



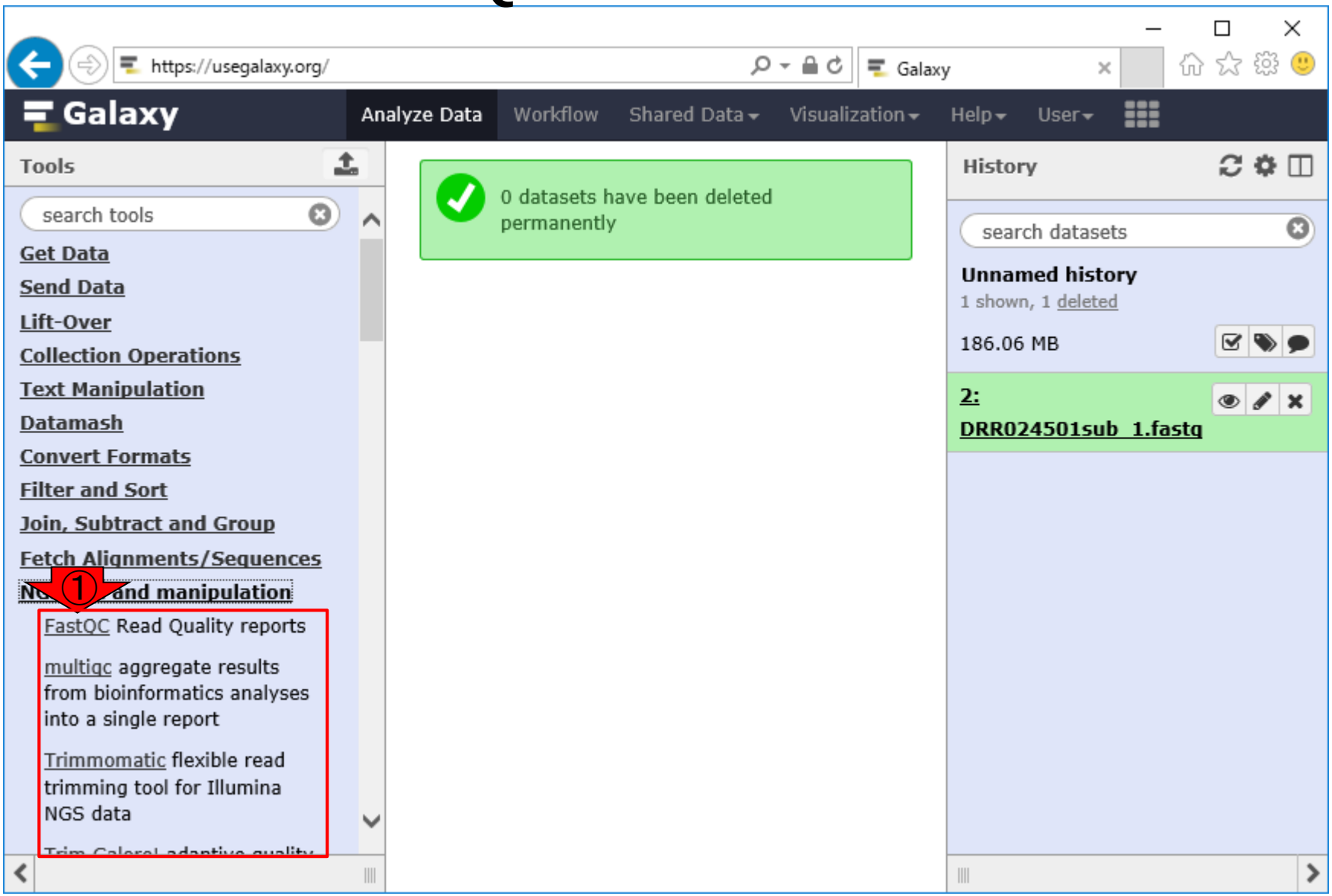
W9-1：クオリティチェック

FastQCを用いてクオリティチェックをやってみましょう。①NGS: QC and manipulation



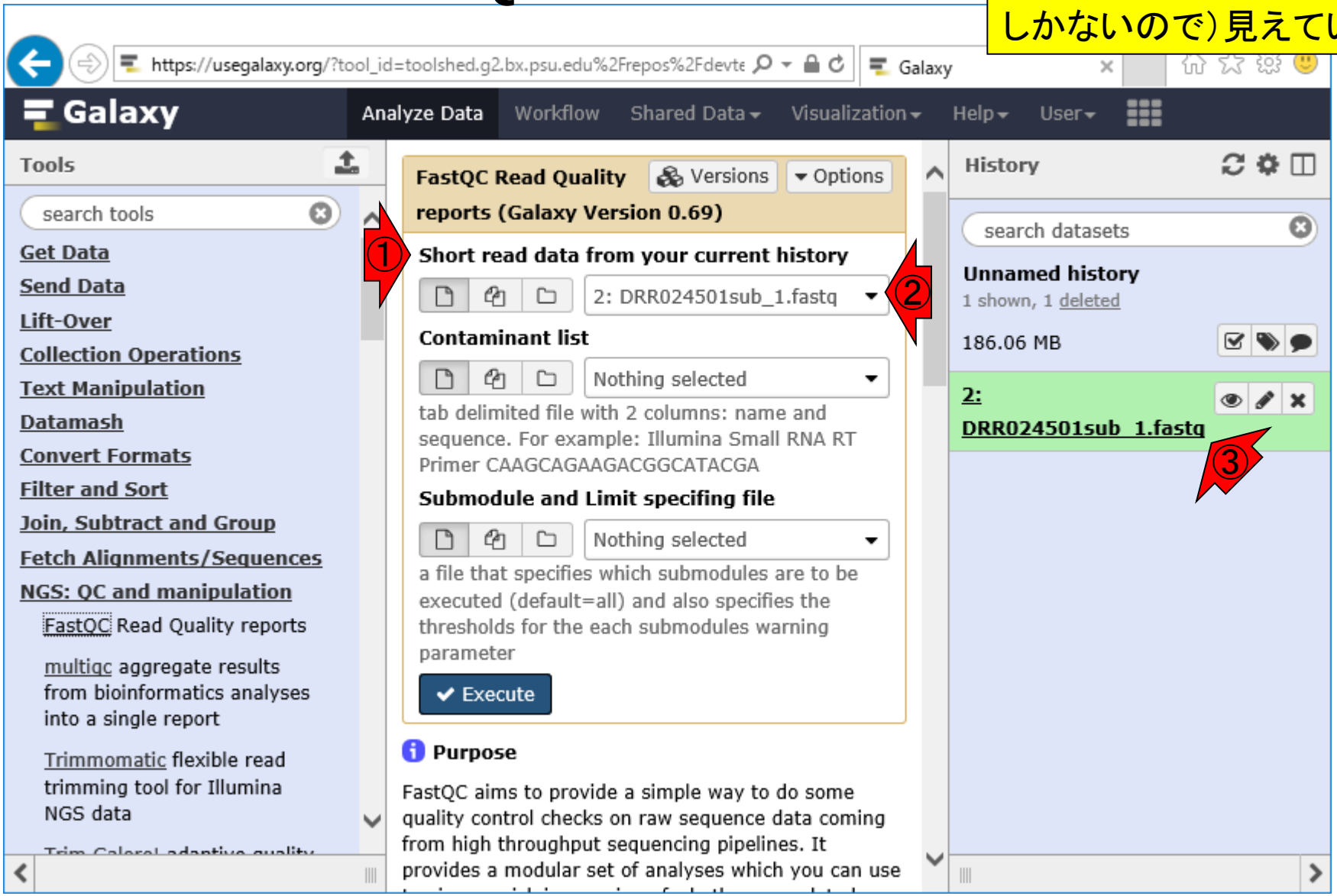
W9-2：FastQC

①FastQC。赤枠内のプログラムの表示順は、見る時期によって結構違いますのでご注意ください



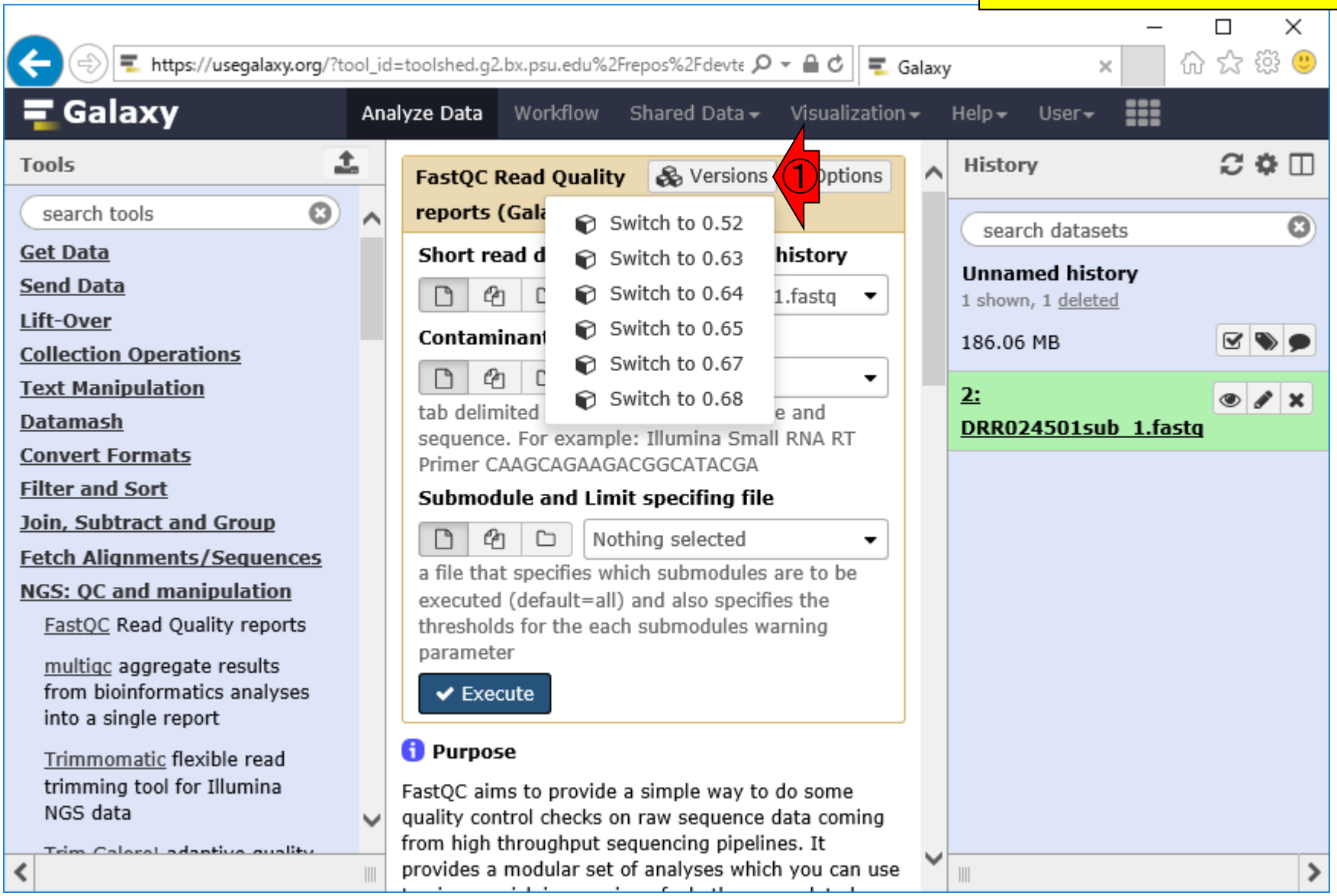
W9-3：FastQCの操作画面

①Short read data from your current historyの、②入力ファイルのところに、③アップロードしたファイルが（1つしかないので）見えているはず



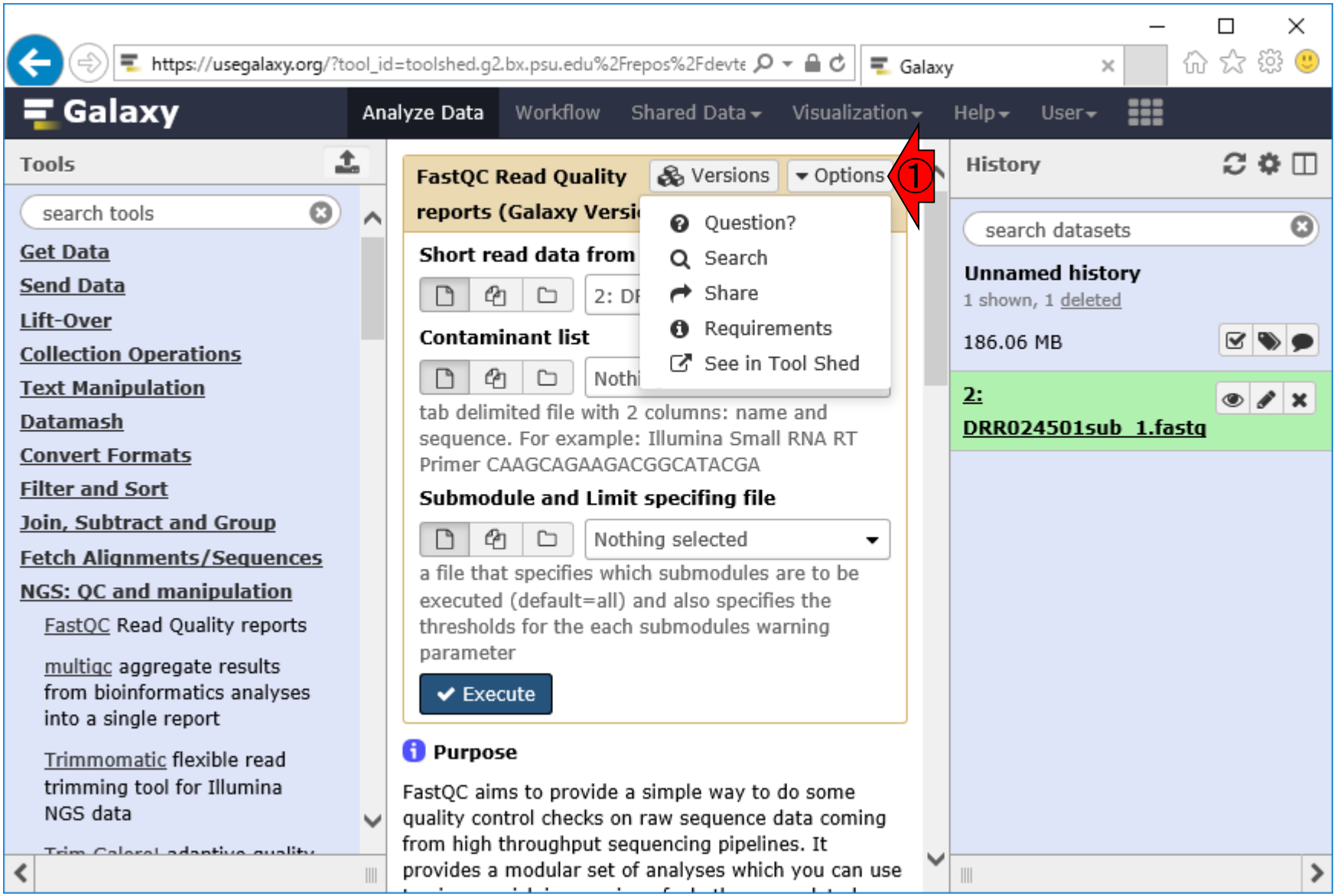
W9-4：バージョンの切替え

FastQCのバージョン(2017.07.05現在のデフォルトはver. 0.69)も①Versionsを押すことで切替えられるようです



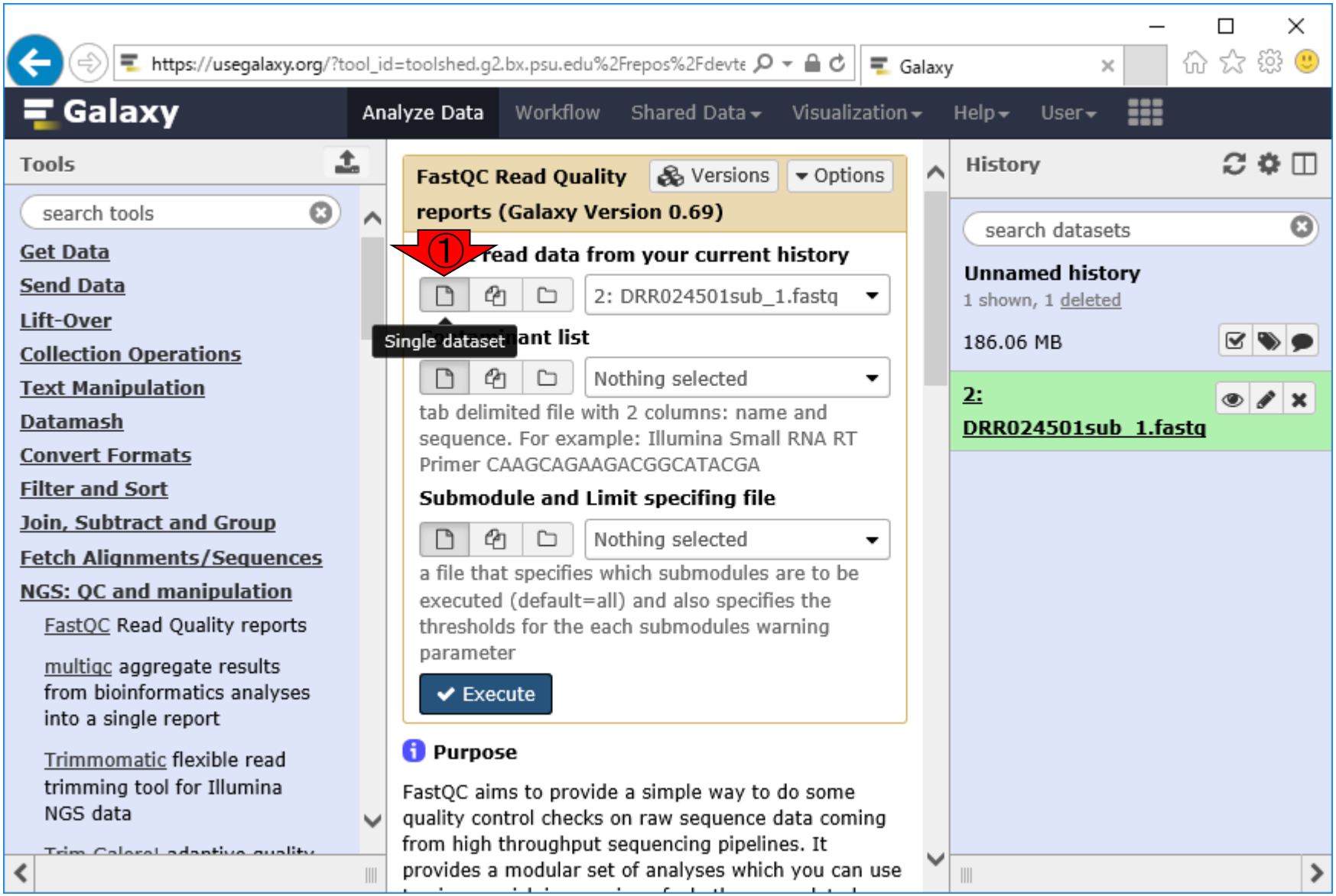
W9-5：オプション

①オプションもいろいろ選べそうですが…
よくわからないので見なかったことにする



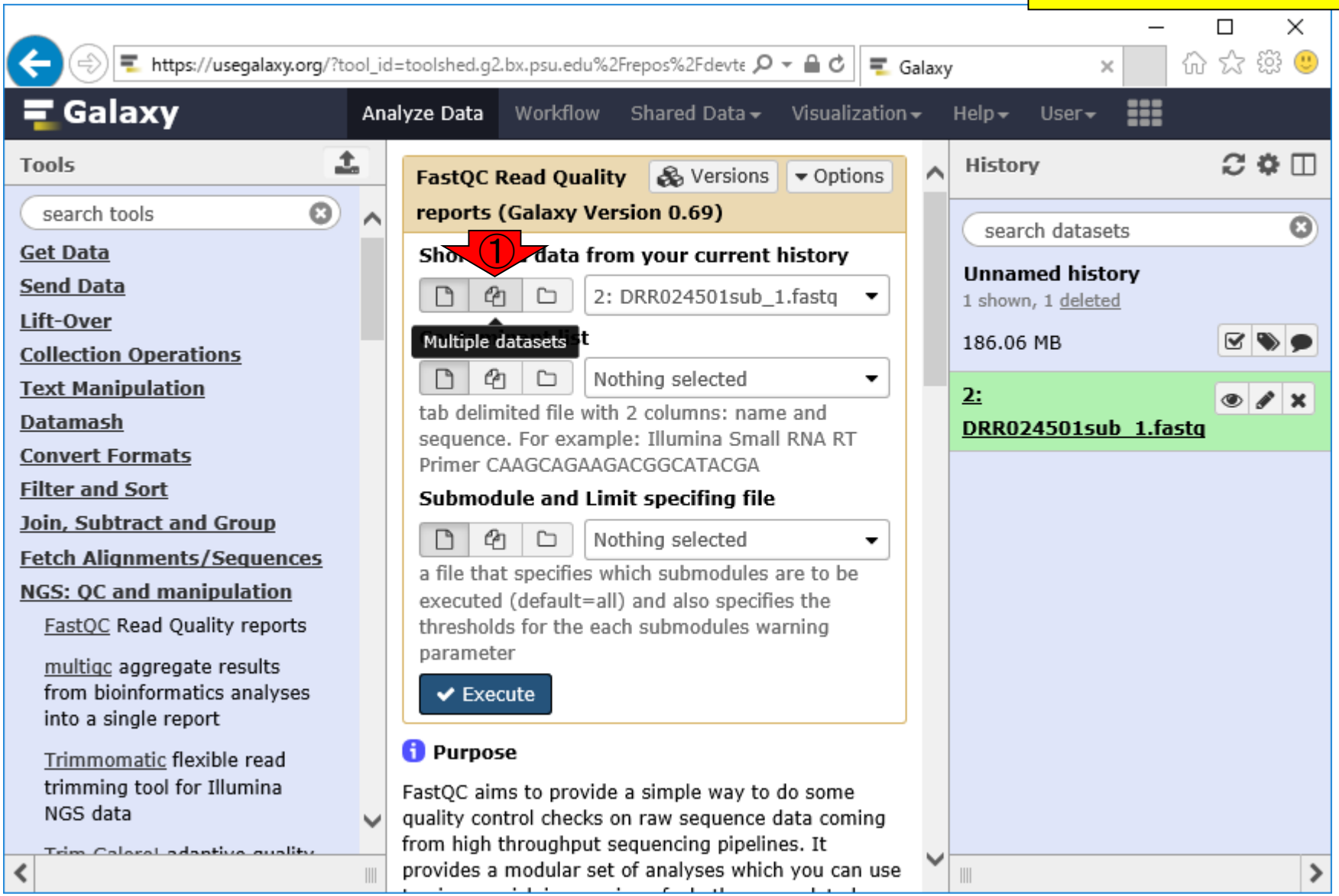
W9-6：1つのファイルなので

①の部分が反転されているのでしょう



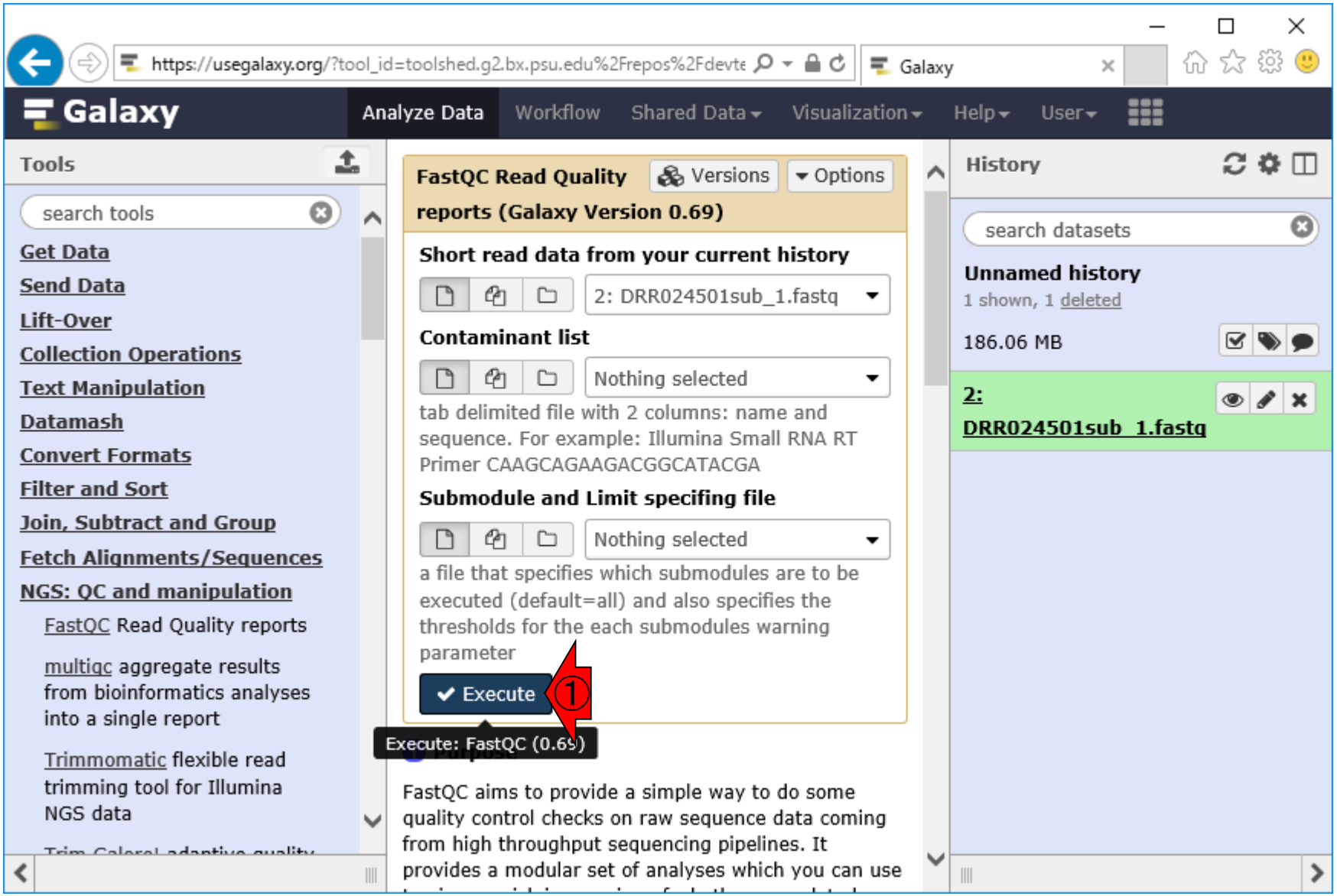
W9-7：複数のファイルの場合

複数ファイルを同時に実行したい場合は①Multiple datasetsをクリックして指定すればいいのでしょう



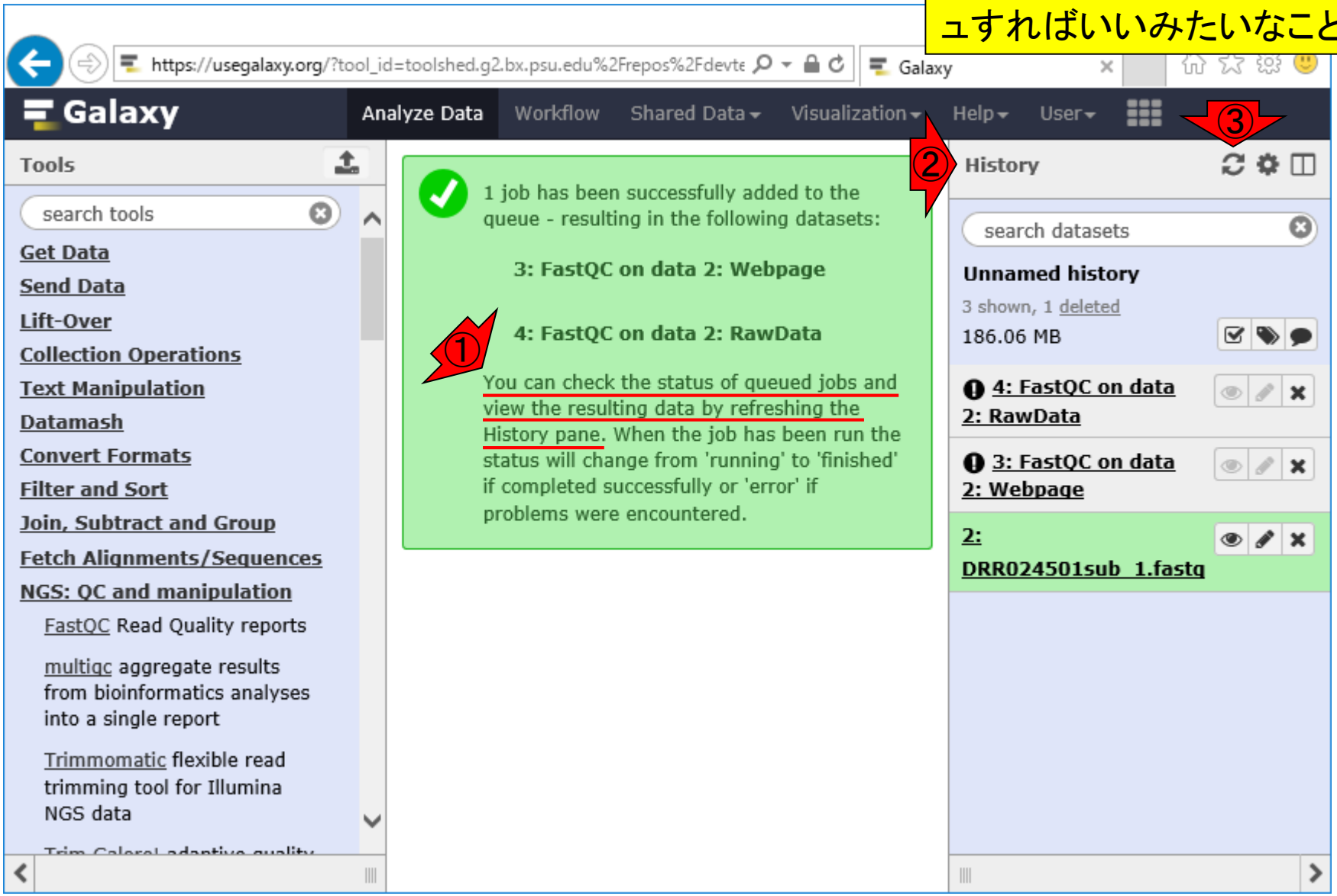
W9-8：Execute

とりあえずデフォルトのままで実行すべく、①Execute



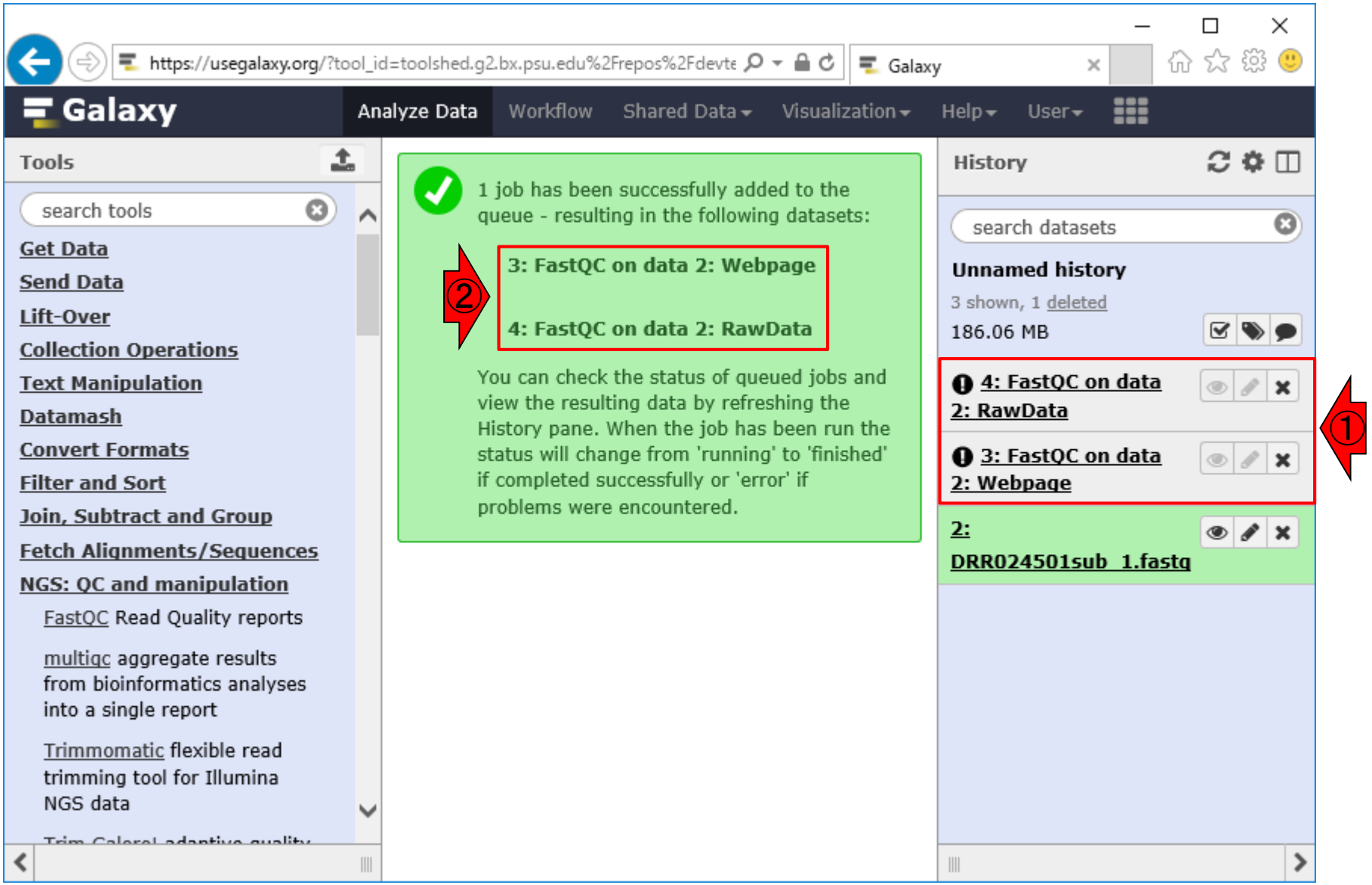
W9-9：すぐに画面が

こんな感じで切り替わりました。よくわからないが、①の赤下線部分をよくみると、②右側のヒストリーパネルで③リフレッシュすればいいみたいなことが書いてある



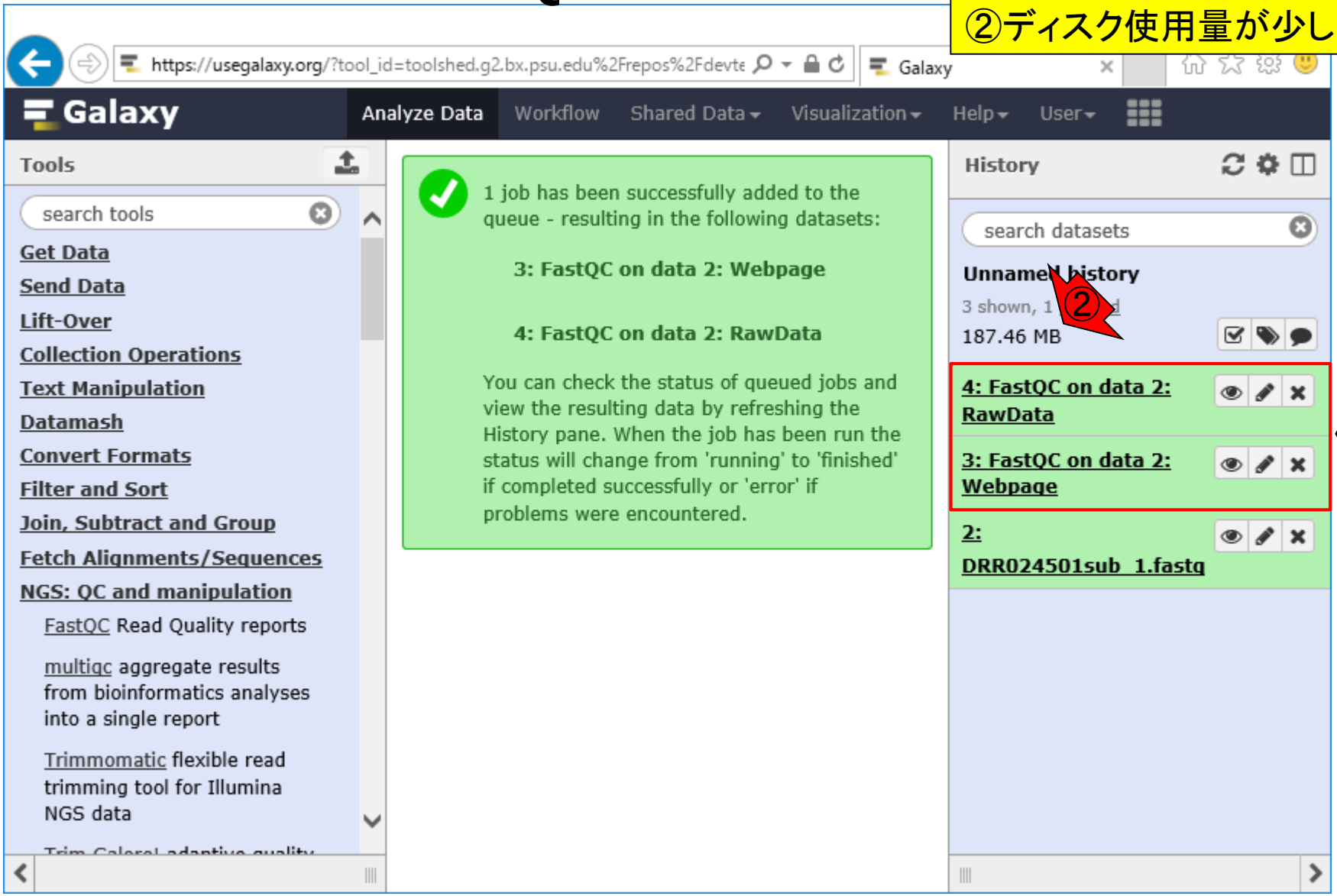
W9-9：すぐに画面が

①は②で見えているものと同じですね。ちなみに①の部分が灰色のときは実行待ち状態です



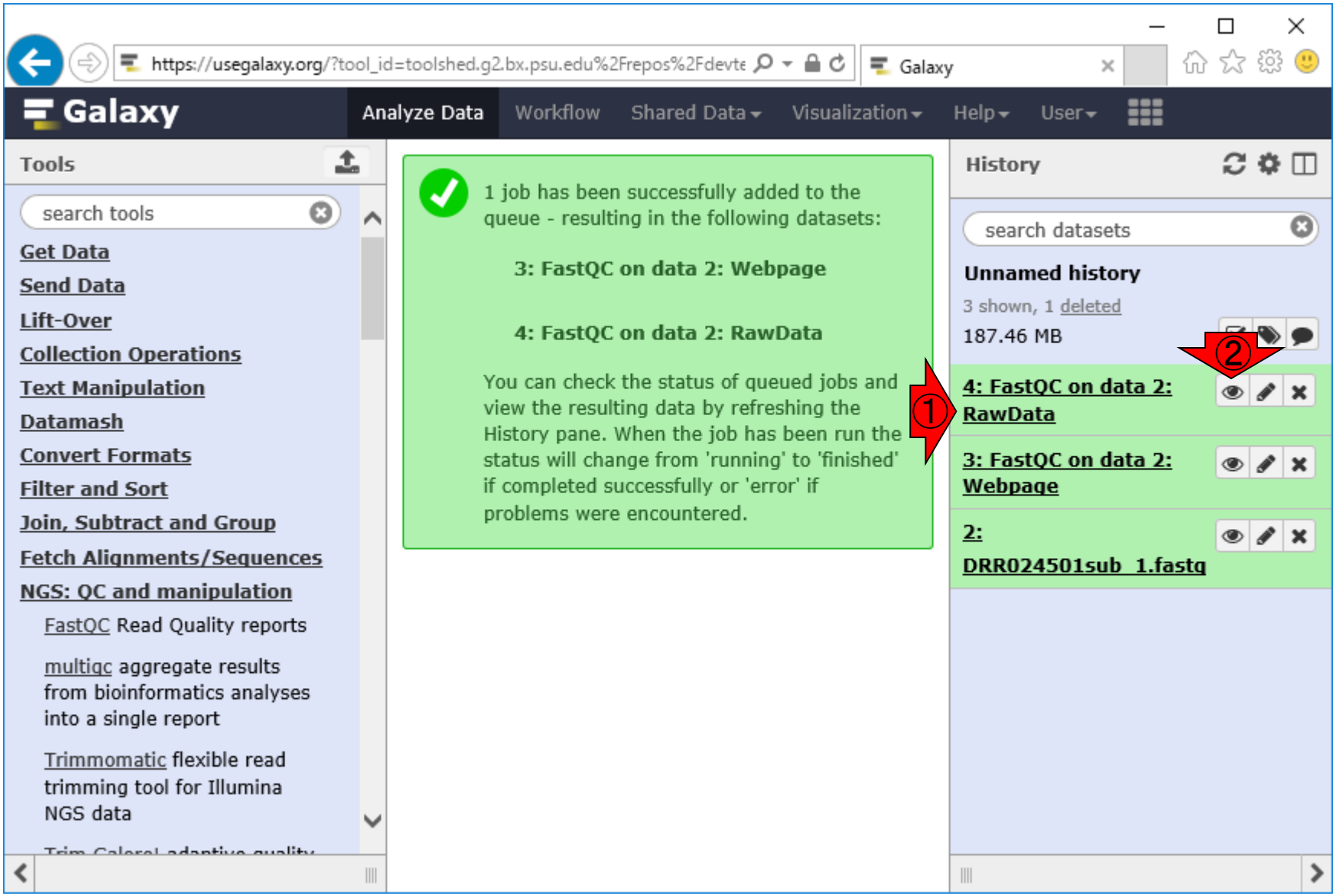
W9-10：FastQC実行終了

実行が無事終了すると①の赤枠部分が緑色になります。実行中は黄色、実行に失敗すると赤色になるらしいです。
②ディスク使用量が少し増加しました



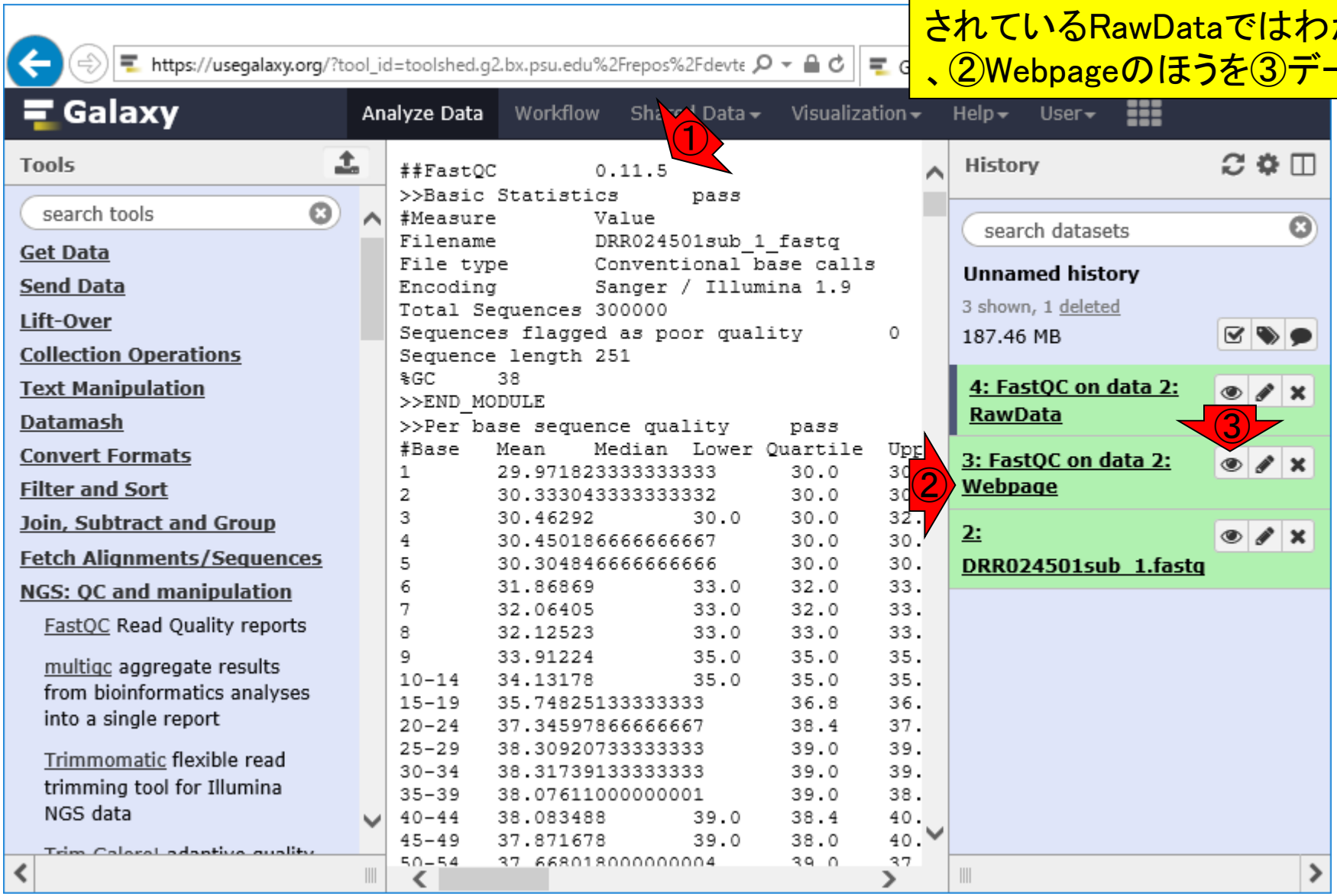
W10-1：実行結果を眺める

①RawDataと書いてあるほうを②データ表示



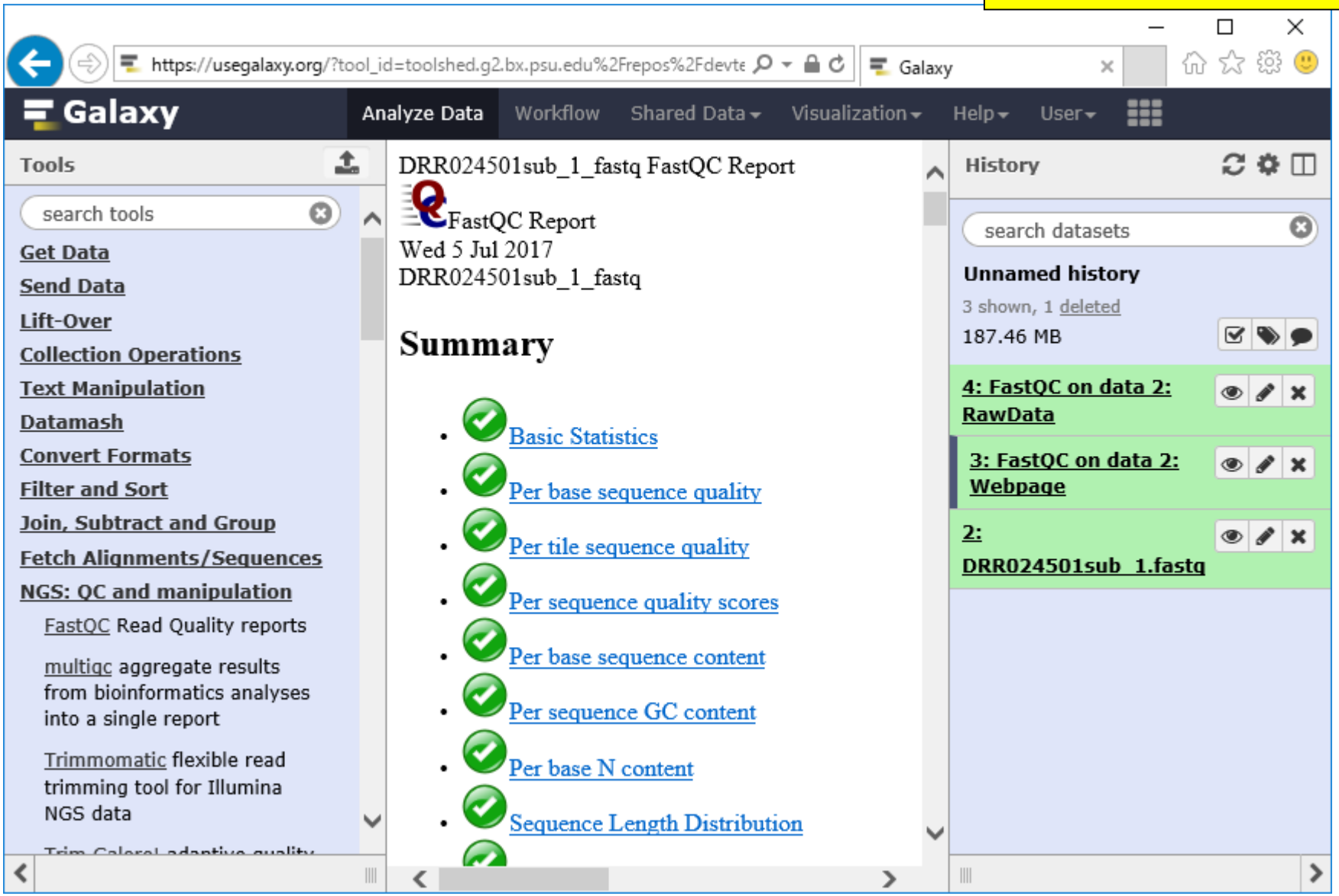
W10-1：実行結果を眺め

こんな感じになります。Galaxy Versionは0.69でしたが、内部的に動いているのは、①0.11.5のようですね。中央パネルに表示されているRawDataではわかりにくいので、②Webpageのほうを③データ表示



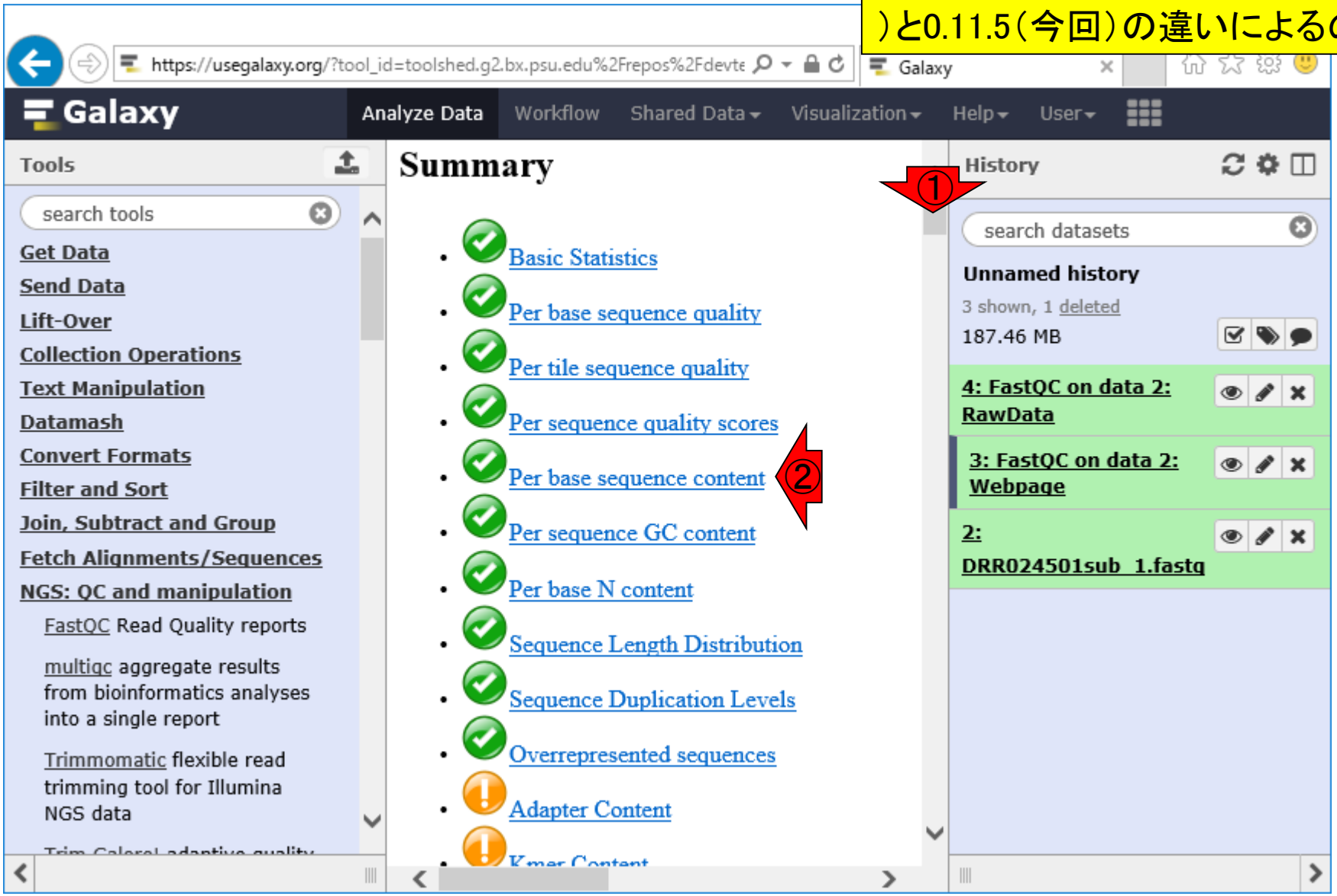
W10-2：htmlレポート

中央パネルが見たことのある
FastQCのhtmlレポートになりました。
第6回W4-2と同じような結果です



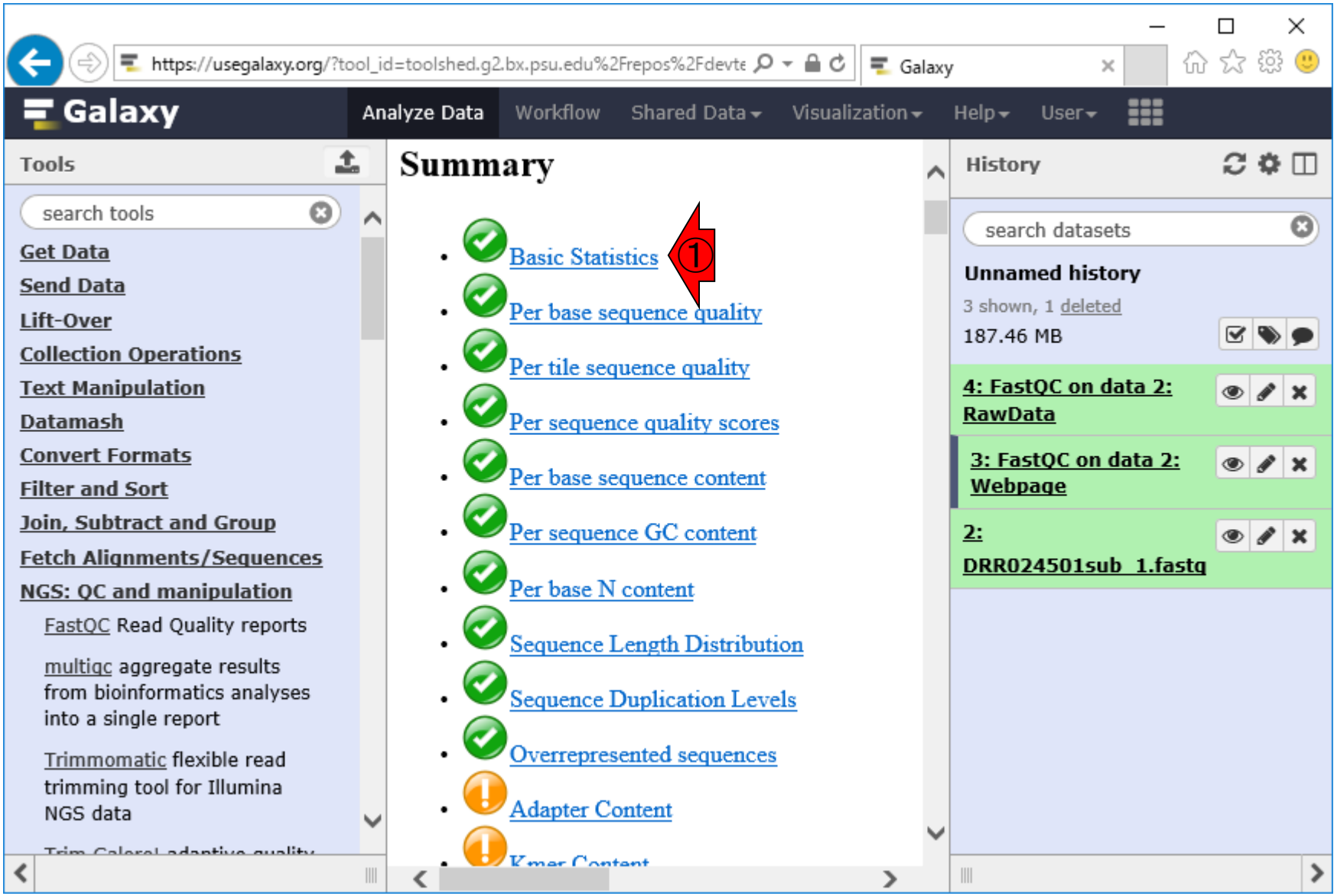
W10-2：htmlレポート

①ちょっとページ下部に移動。②が第6回W4-2では黄色だったが、緑色になっている点が異なる。FastQC ver. 0.11.4（第6回W4-2）と0.11.5（今回）の違いによるのかもしれない



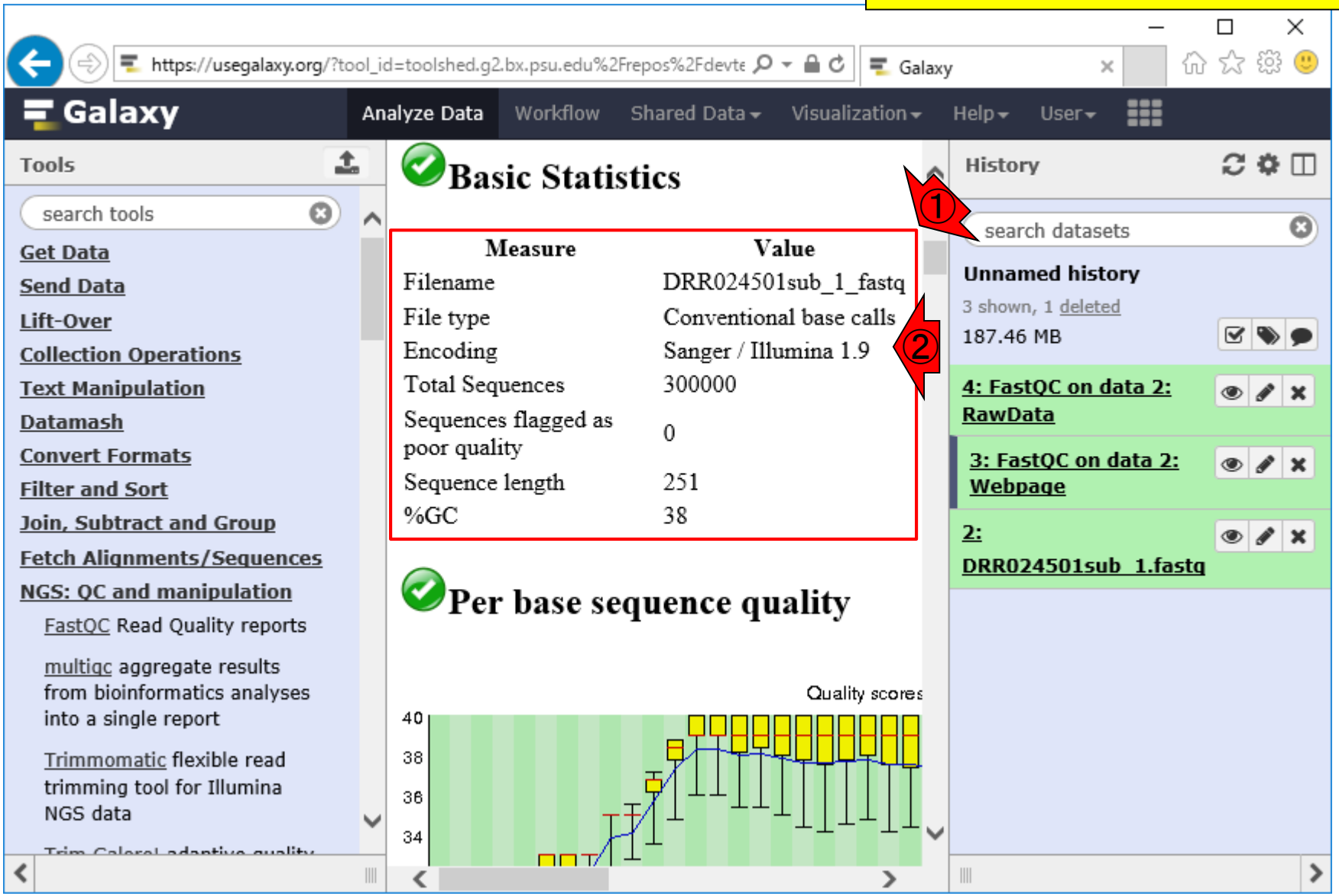
①FastQC実行結果の基本情報を眺める

W10-3：Basic Statistics



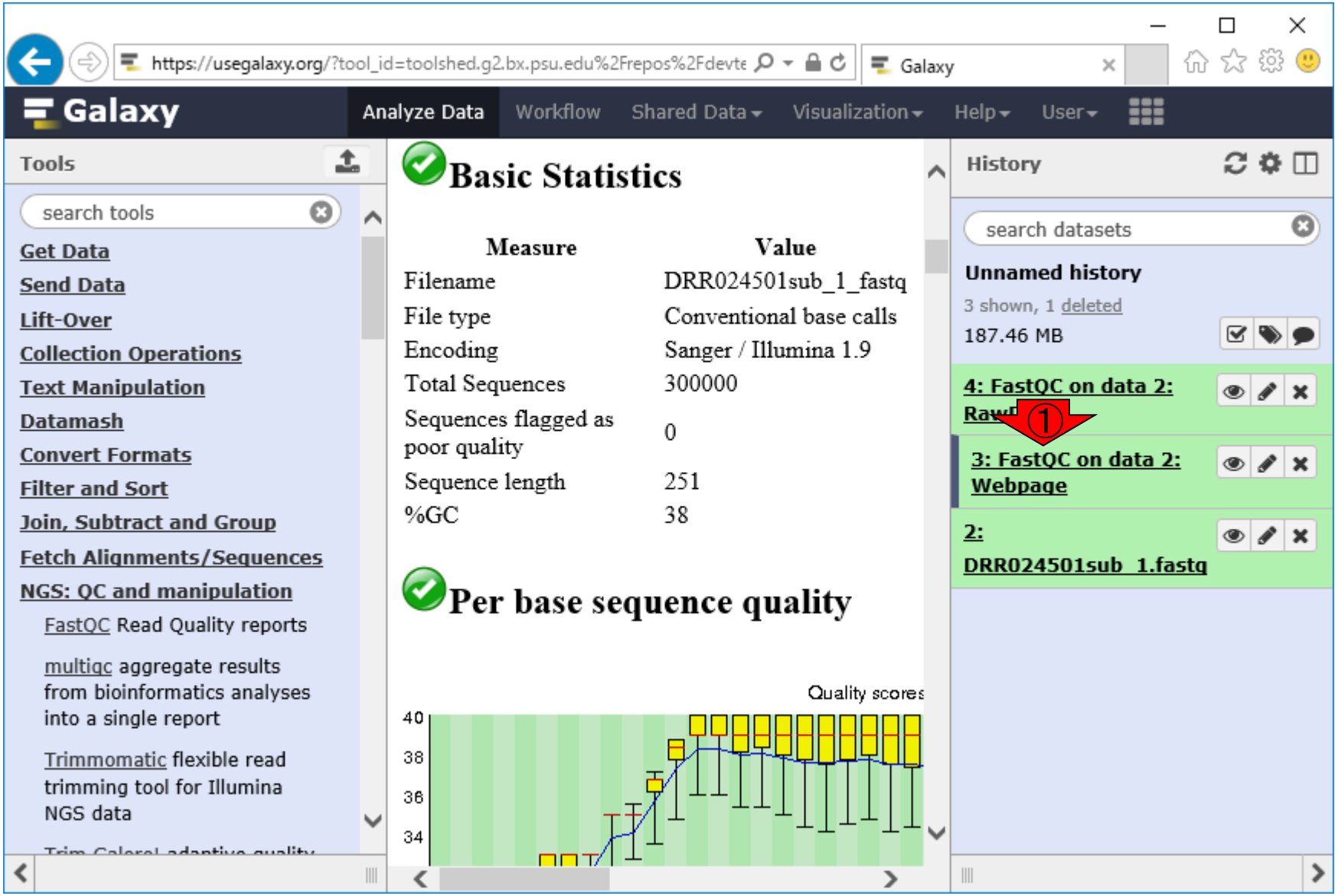
W10-3：Basic Statistics

①このあたりの結果は第6回W4-2と同じですね。②同じFASTQ形式でもいくつかの亜流があり、これはSanger/Illumina 1.9というもの



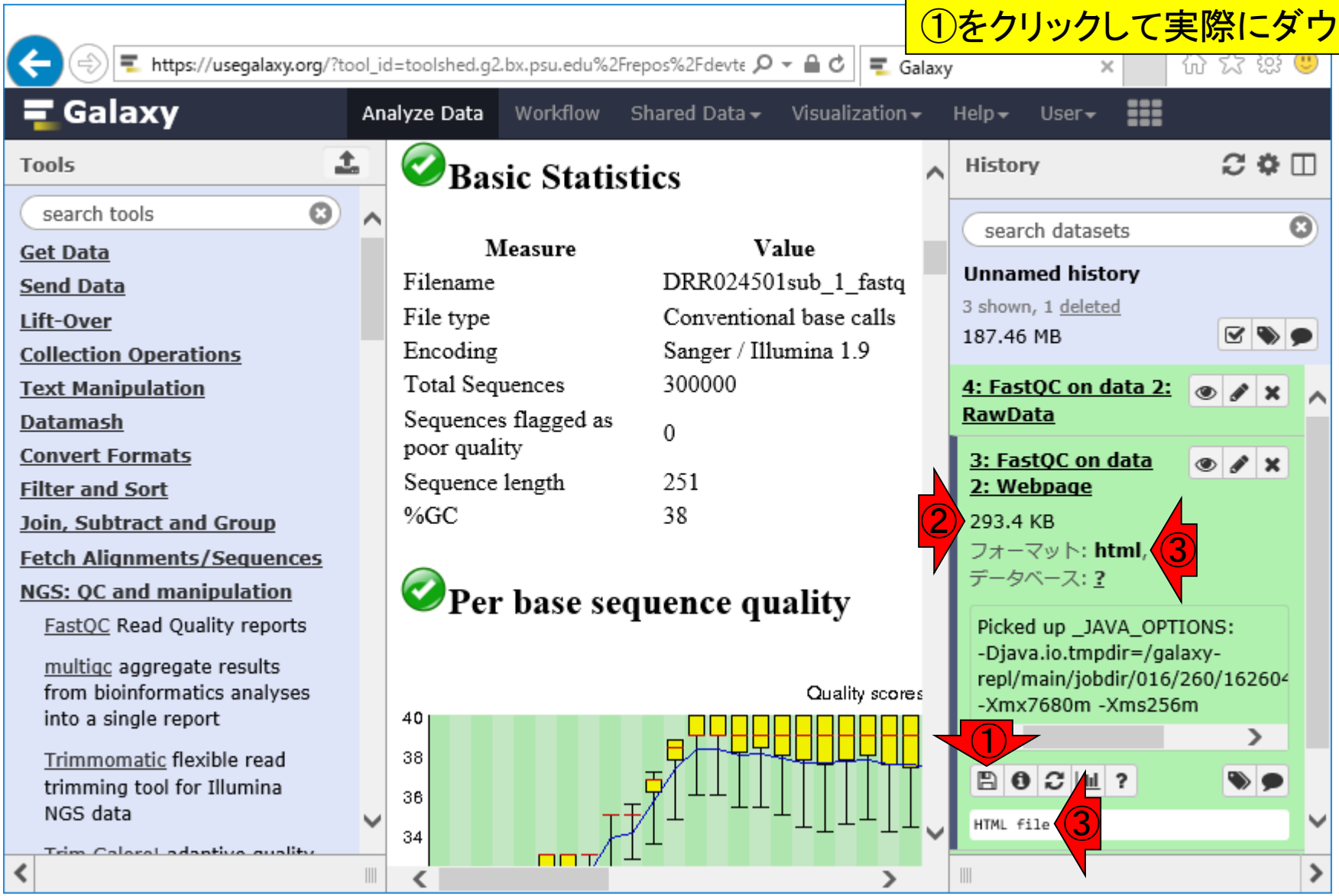
①のところをクリック

W10-4：結果のダウンロード



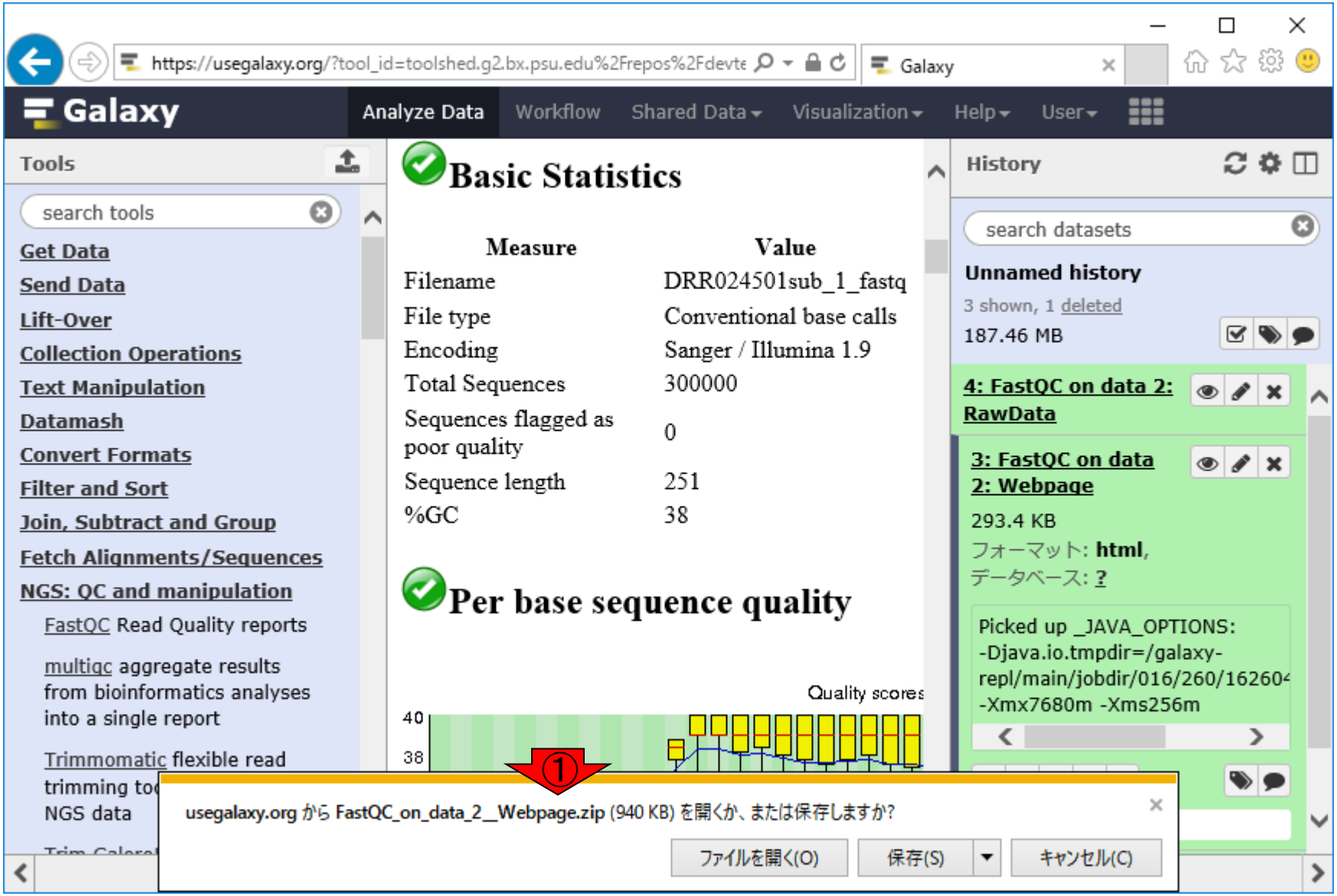
W10-4：結果のダウンロー

こんな感じになります。①フロッピーディスクアイコンをクリックすれば、②293.4 KBの③htmlファイルをダウンロードできます。①をクリックして実際にダウンロードします



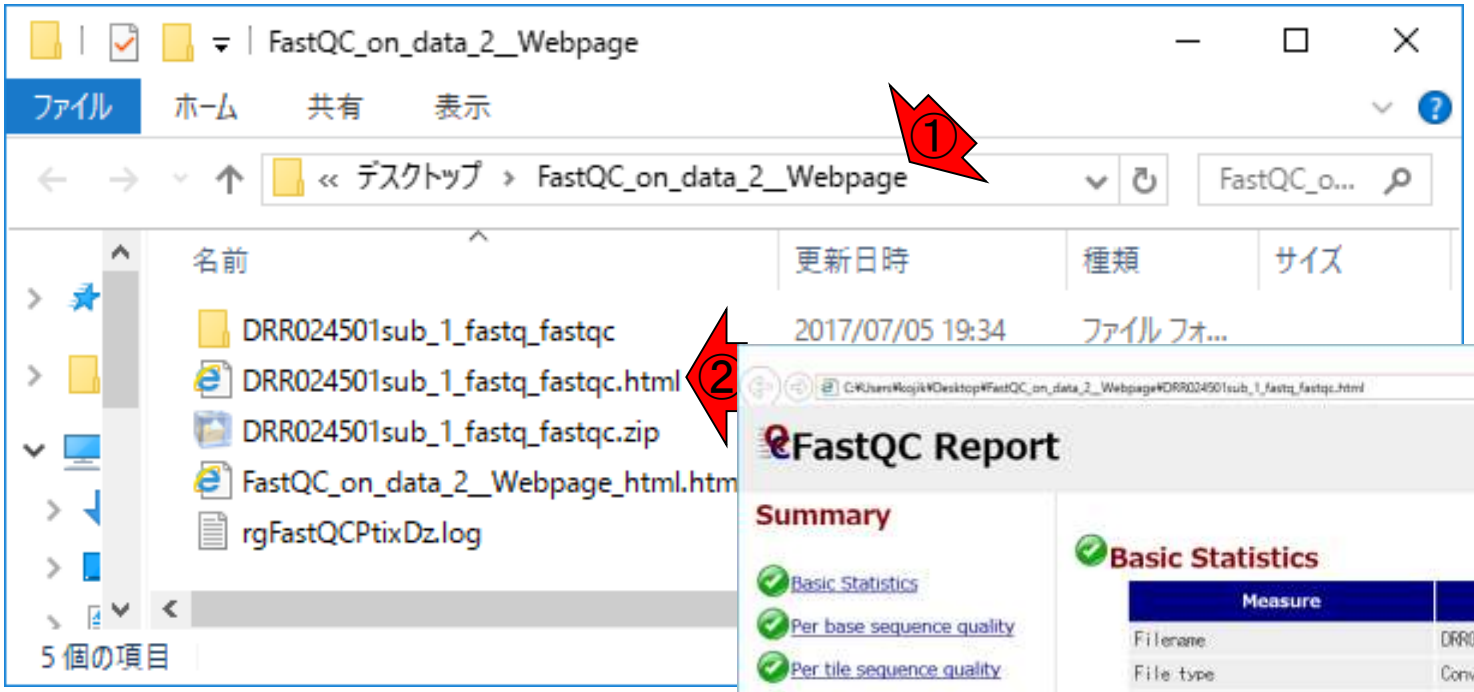
①で示すzip圧縮ファイルで保存されます

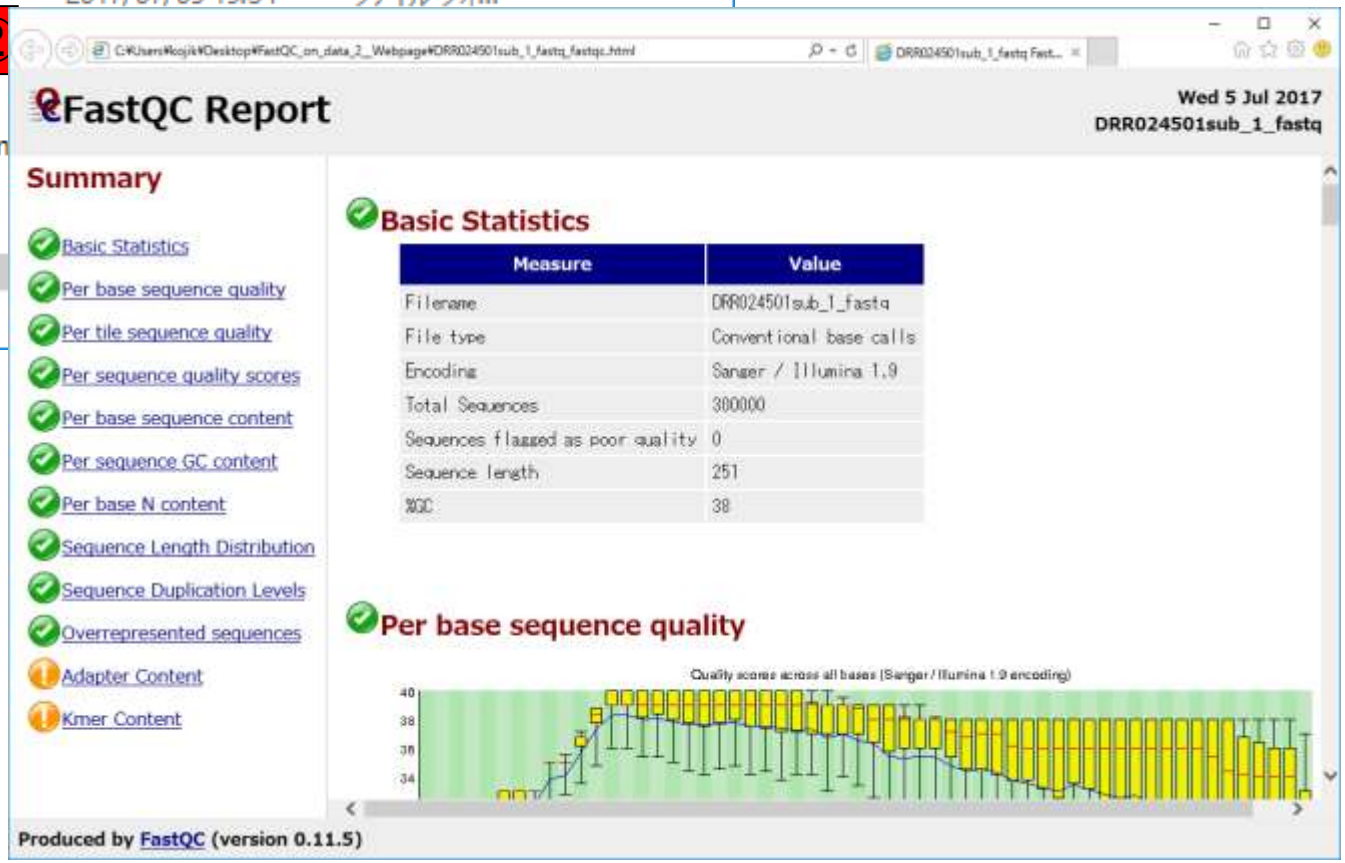
W10-4：結果のダウンロード



W10-5：解凍すると…

①のような感じになり、②をダブルクリックすると第6回W4-2とほぼ同じ感じになります





W10-6：Per base sequence…

①Per base sequence quality。
第6回W4-2とほぼ同じですね

file:///C:/Users/kojik/Desktop/FastQC_on_data_2_Webpage/DRR024501sub_1_fastq_fastqc.html#M1

DRR024501sub_1_fastq Fast...

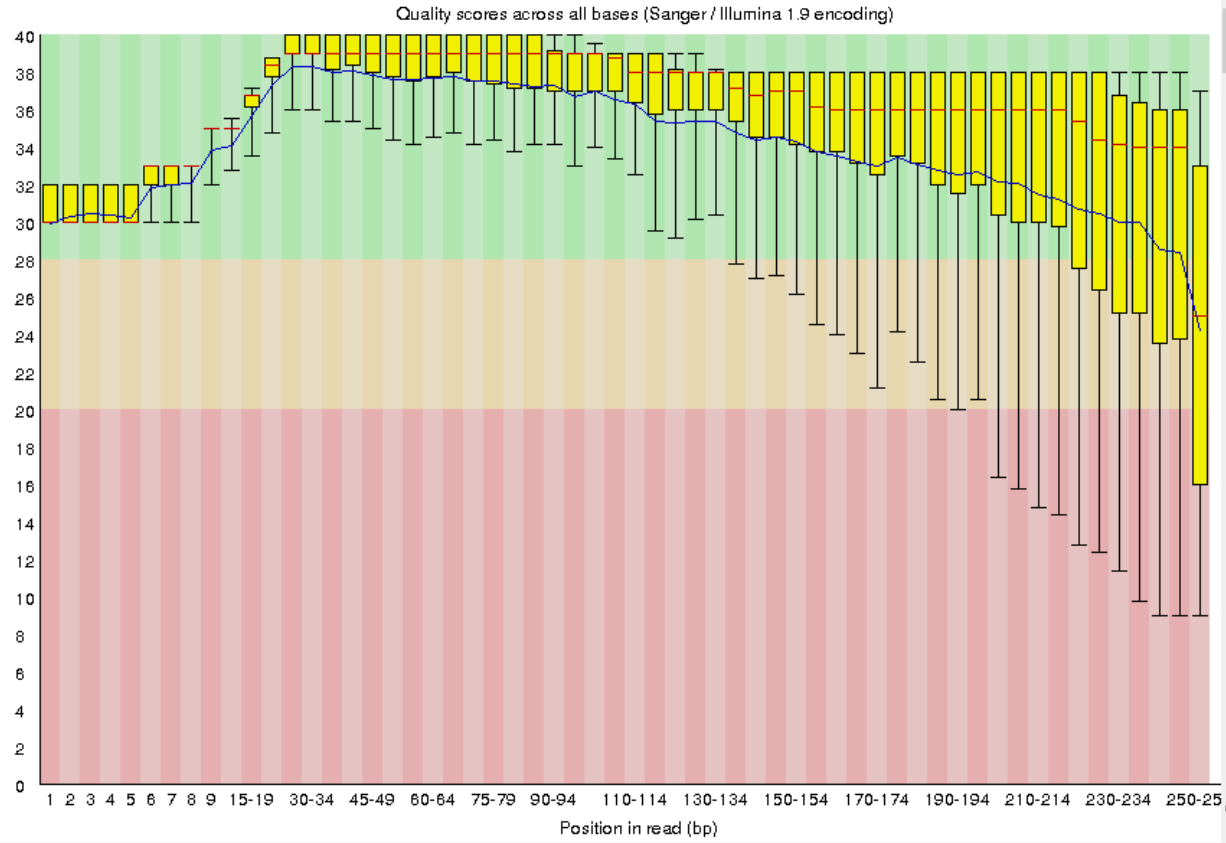
FastQC Report

Wed 5 Jul 2017
DRR024501sub_1_fastq

Summary

- Basic Statistics
- Per base sequence quality ← ①
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Per base sequence quality



Produced by FastQC (version 0.11.5)

W10-7：Per base sequence…

①第6回のhtmlレポートの結果です。同じであることがわかります

file:///C:/Users/kojik/Documents/2015/執筆/乳酸菌学会誌/第6回/DRR024501/DRR024501sub_1_fastqc.html#M1

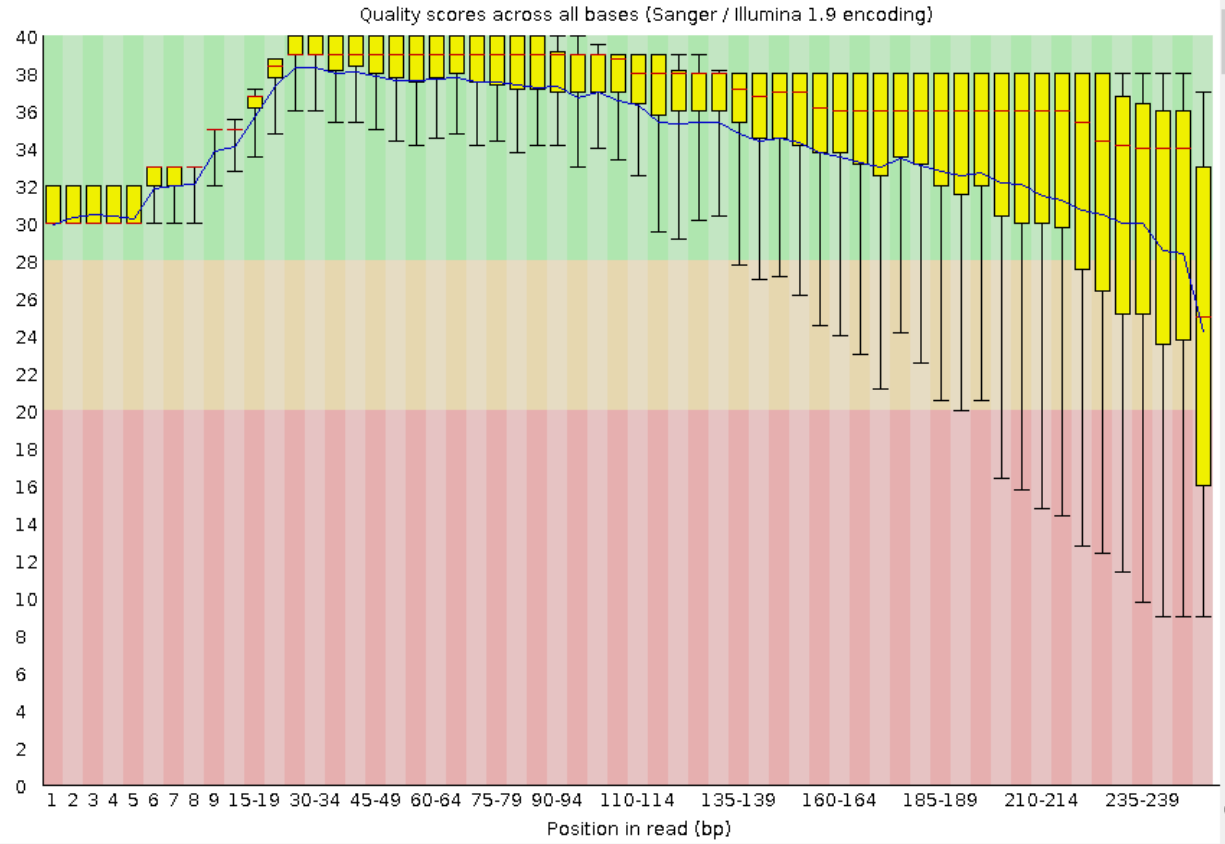
FastQC Report

Tue 29 Dec 2015
DRR024501sub_1.fastq.gz

Summary

- Basic Statistics
- Per base sequence quality ← ①
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Per base sequence quality



Produced by FastQC (version 0.11.4)

W11-1：Adapter Content

①Adapter Contentの結果として、②Illumina Universal Adapterが③含まれていたことを思い出そう（第6回W4-4）

file:///C:/Users/kojik/Desktop/FastQC_on_data_2__Webpage/DRR024501sub_1_fastq_fastqc.html#M10

DRR024501sub_1_fastq Fast...

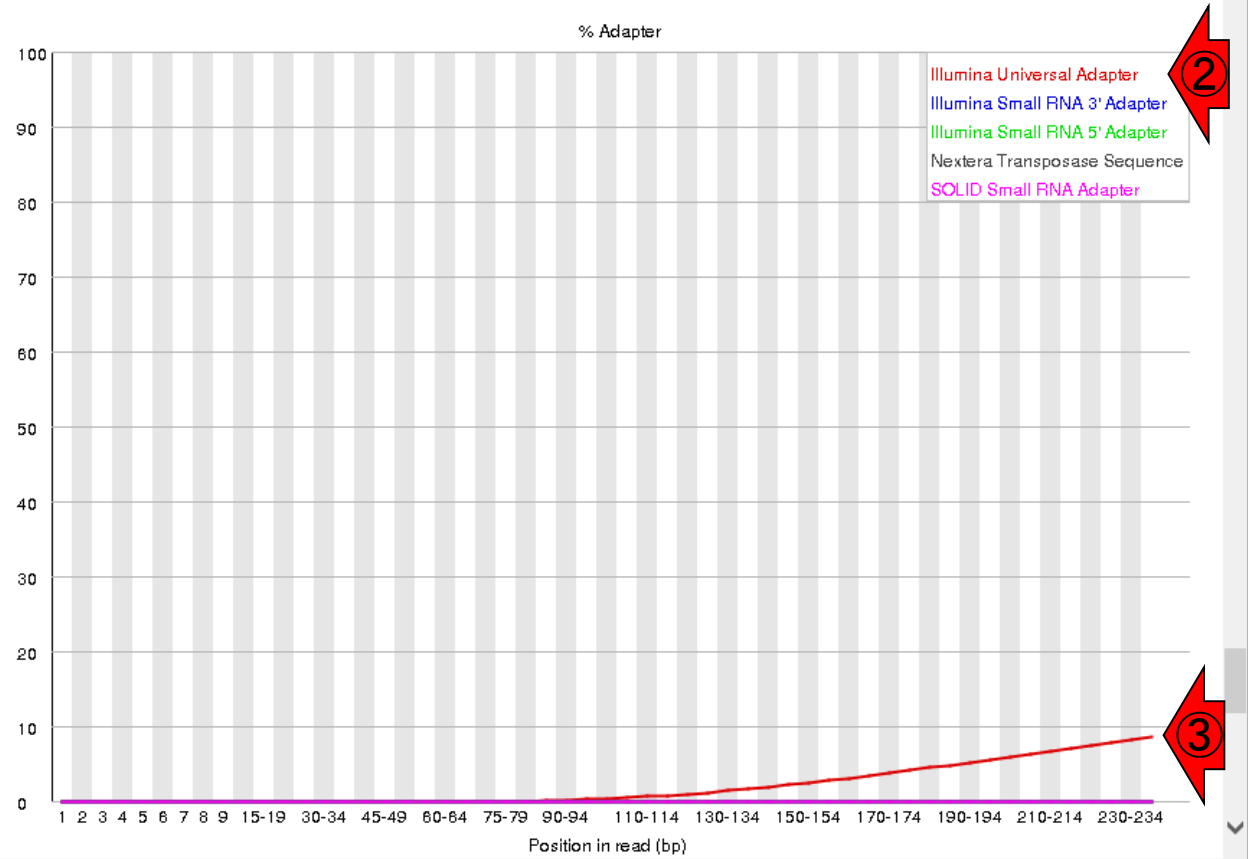
FastQC Report

Wed 5 Jul 2017
DRR024501sub_1_fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content ①
- Kmer Content

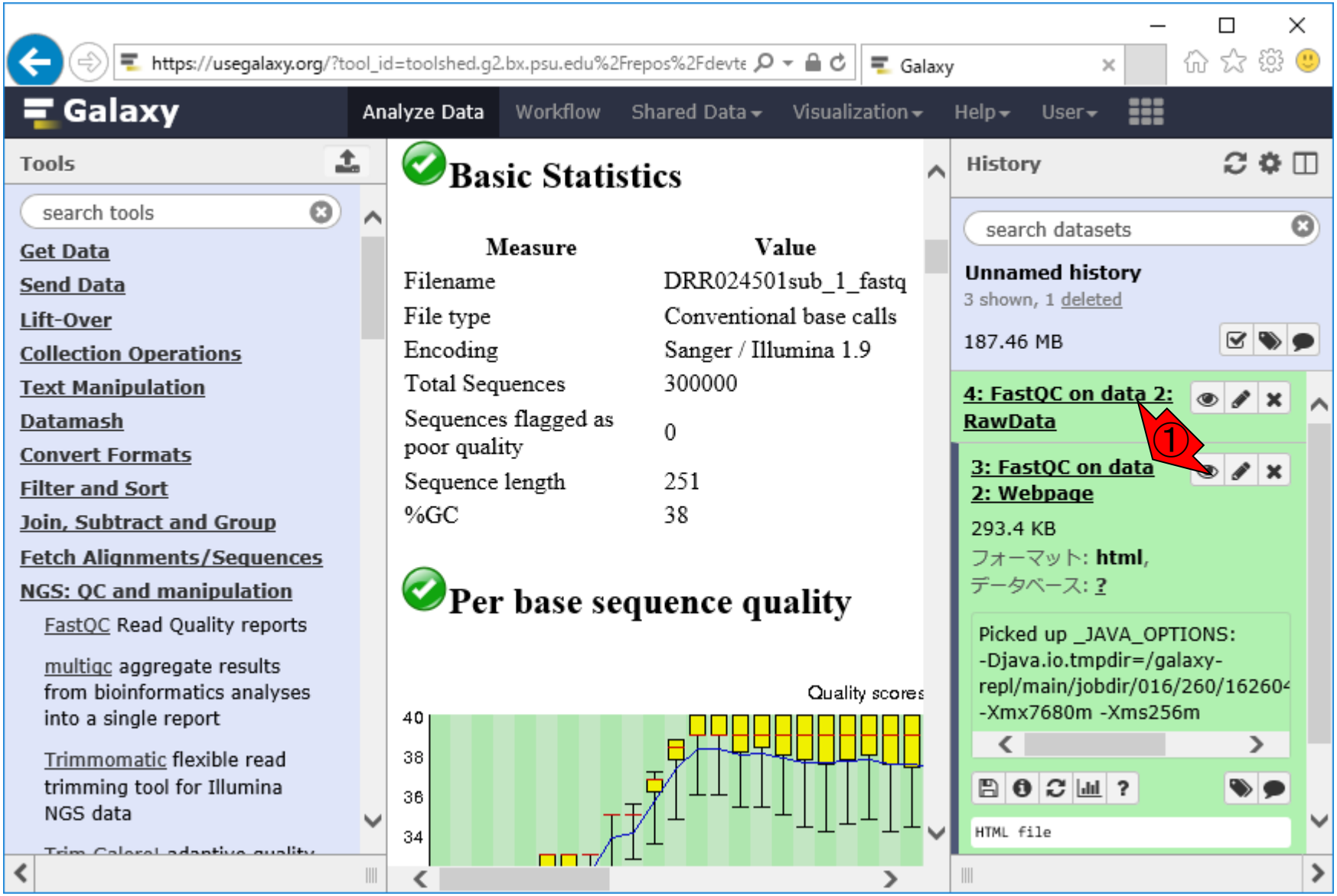
Adapter Content



Produced by FastQC (version 0.11.5)

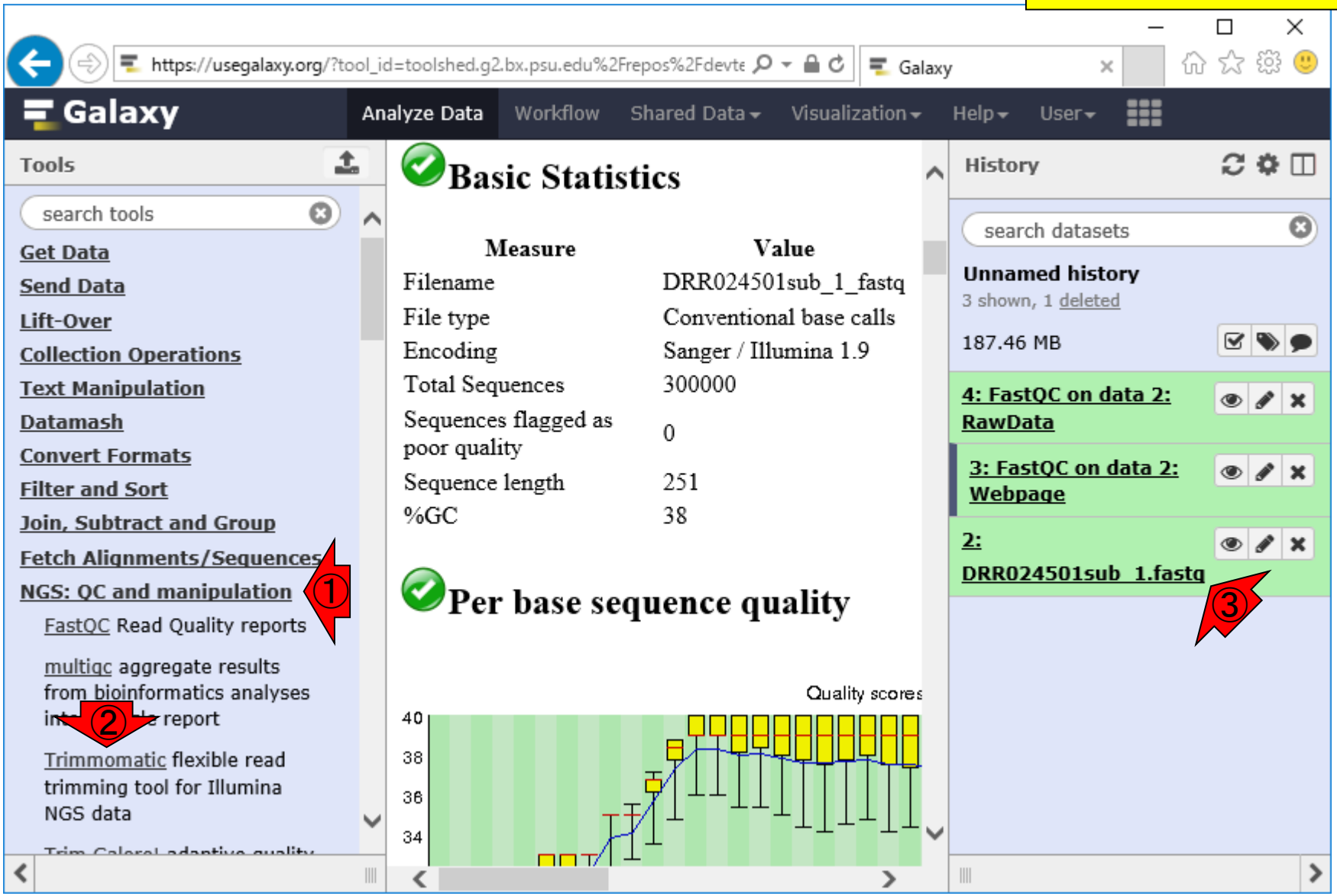
①を押して詳細表示を隠す

W11-2：Trimmomatic



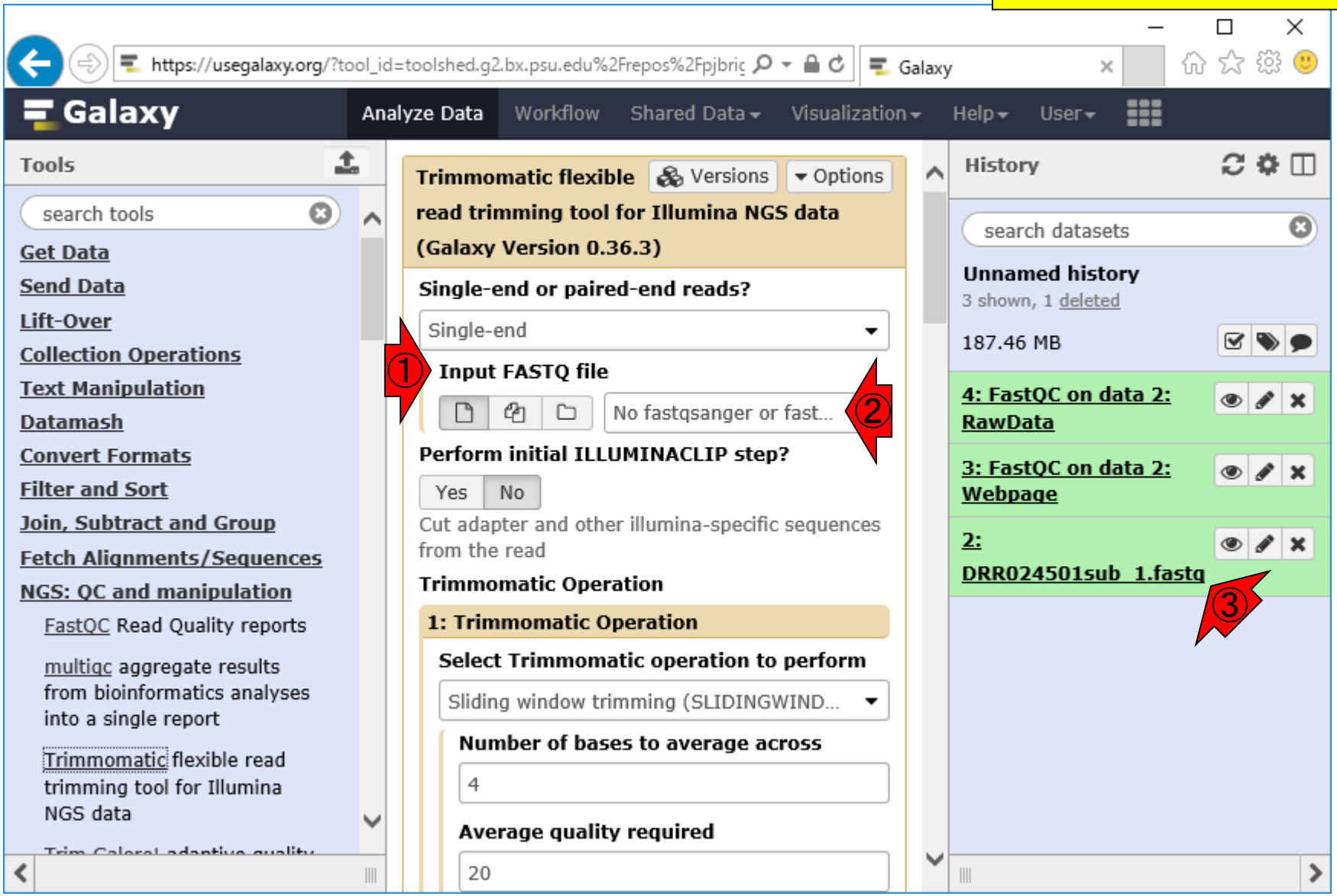
W11-2：Trimmomatic

①FastQCと同じ項目内にある、②Trimmomaticを使ってアダプターを除去してみる。③が入力ファイル



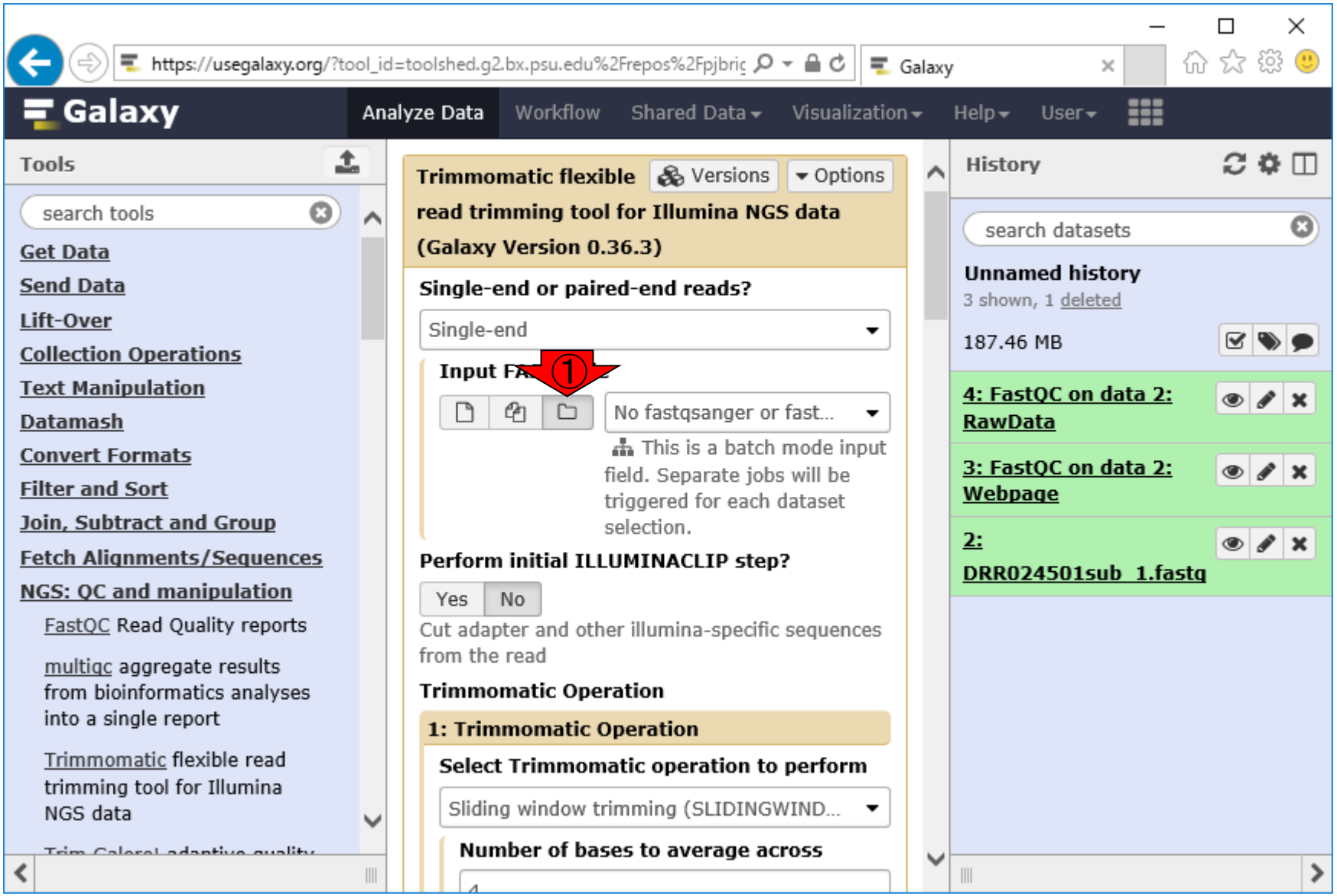
W11-3：入力ファイル形式

①Input FASTQ fileのところで、②No fastqsanger or …となり、③入力ファイルを認識させることができません



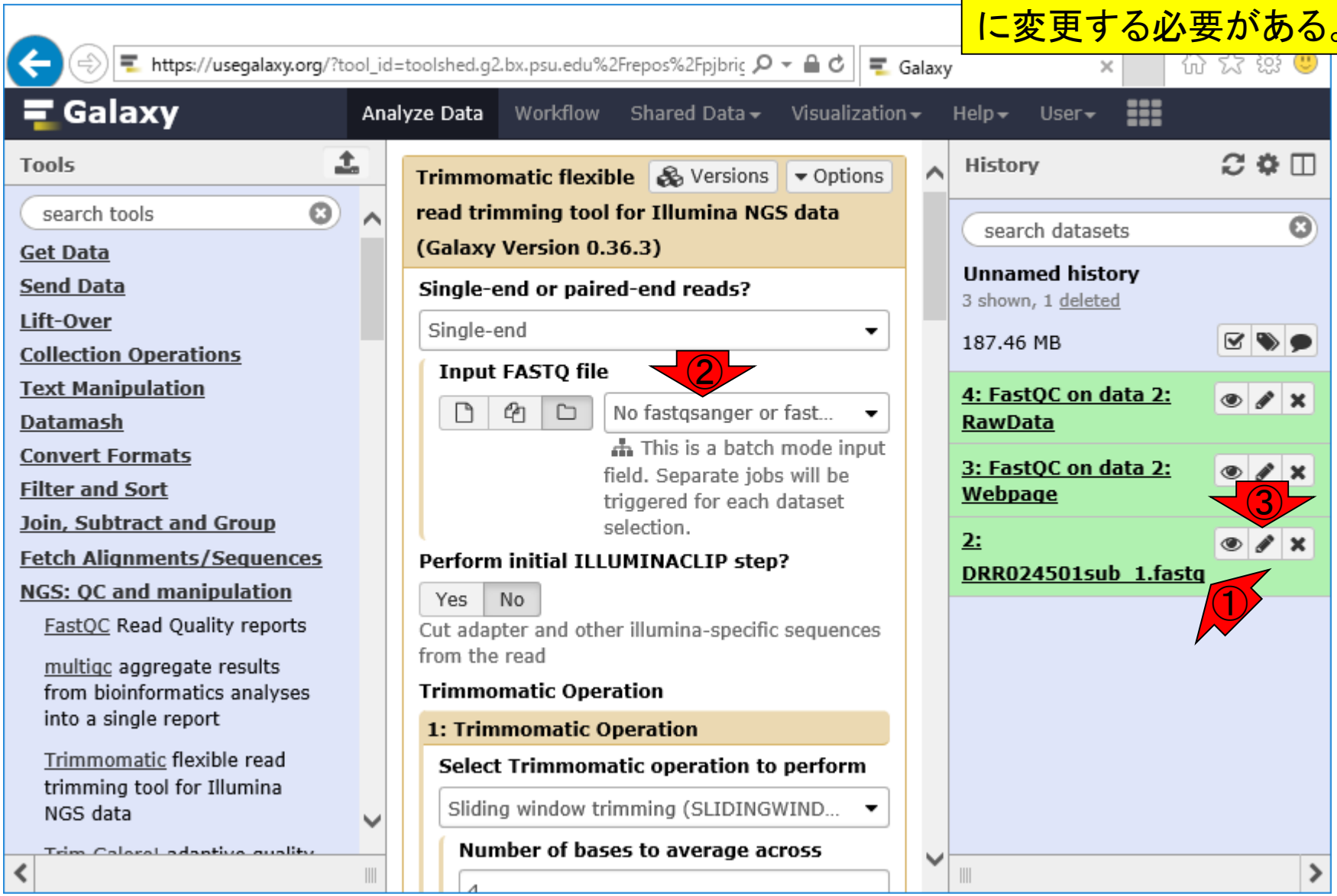
①ここを押してみてもよくわかりません

W11-3：入力ファイル形式



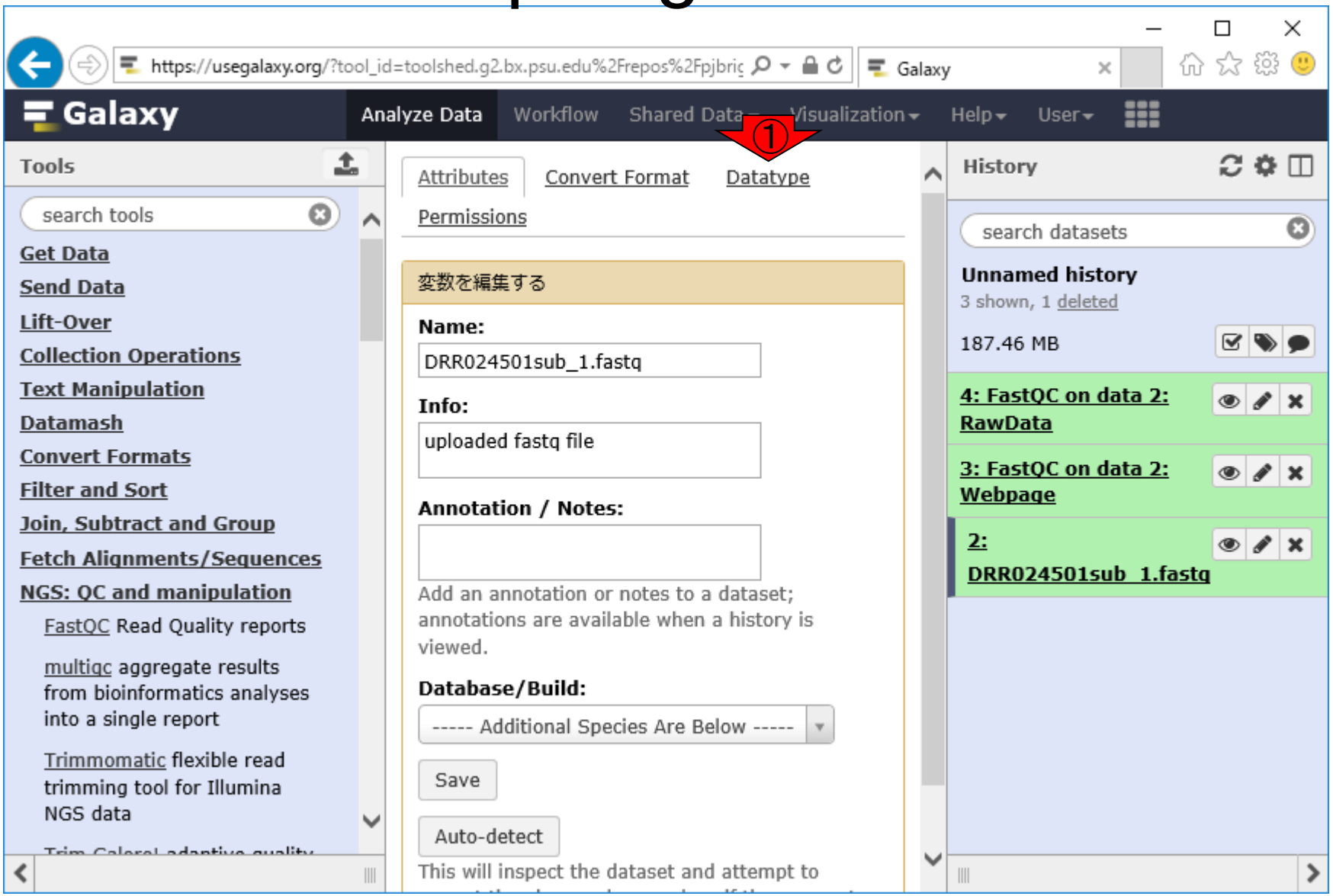
W11-3：入力ファイル形式

認識されない原因は、①のファイル形式が②fastqsangerではないため。③でファイル形式を明示的にfastqsangerに変更する必要がある。③を押す



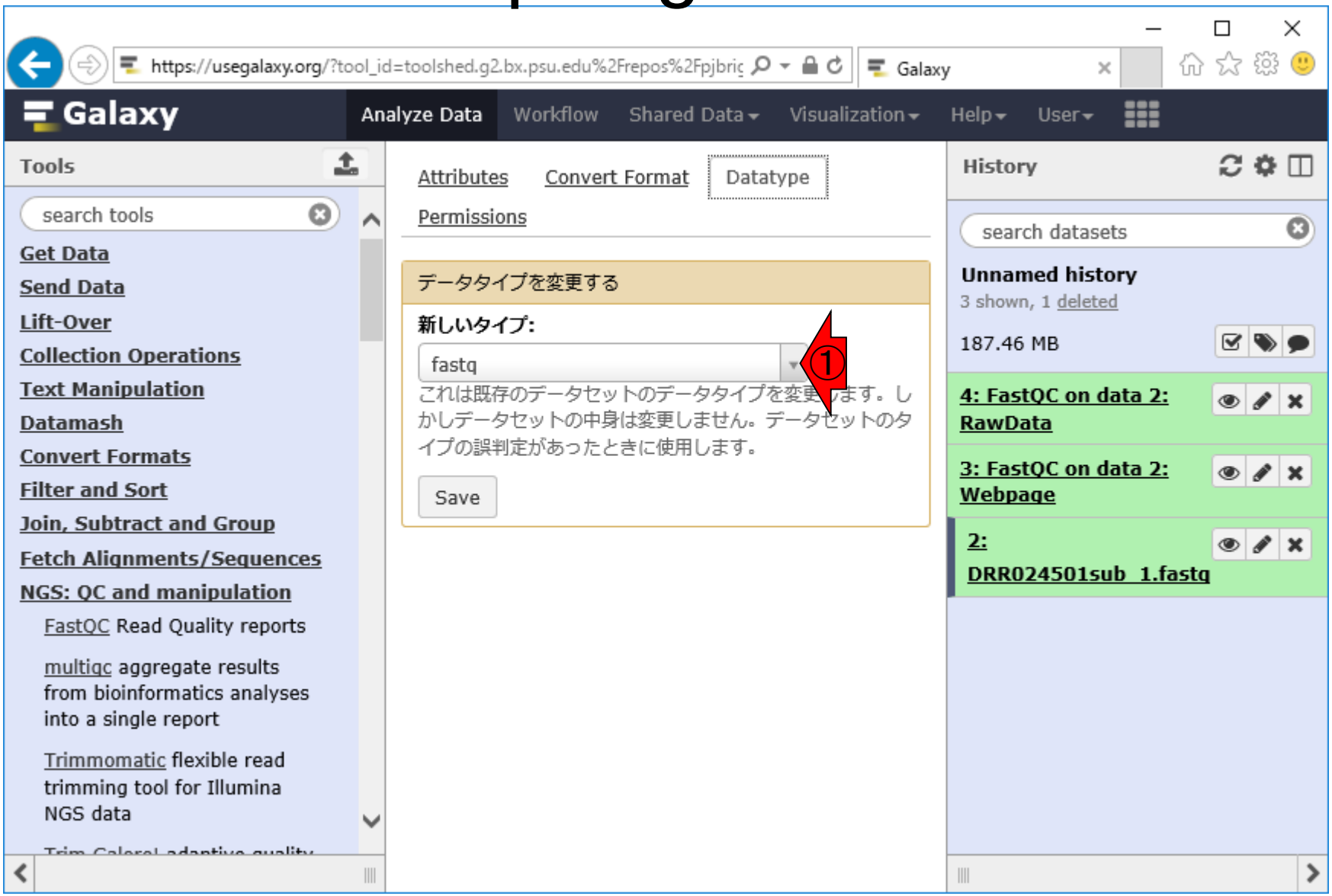
①Datatypeタブを選んで

W11-4：fastqsanger



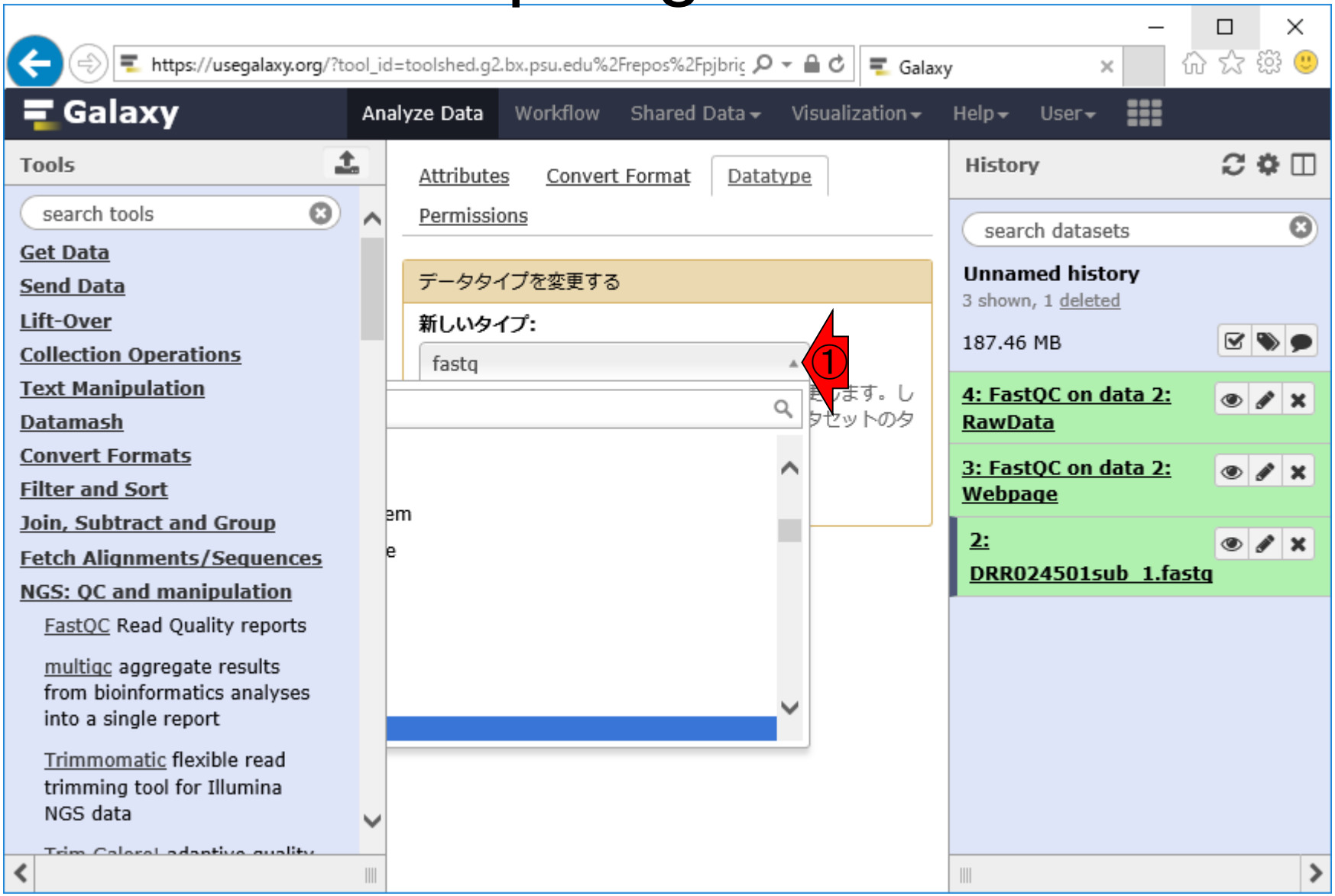
W11-4：fastqsanger

①デフォルトがfastqになっているので、これをfastqsangerに変える



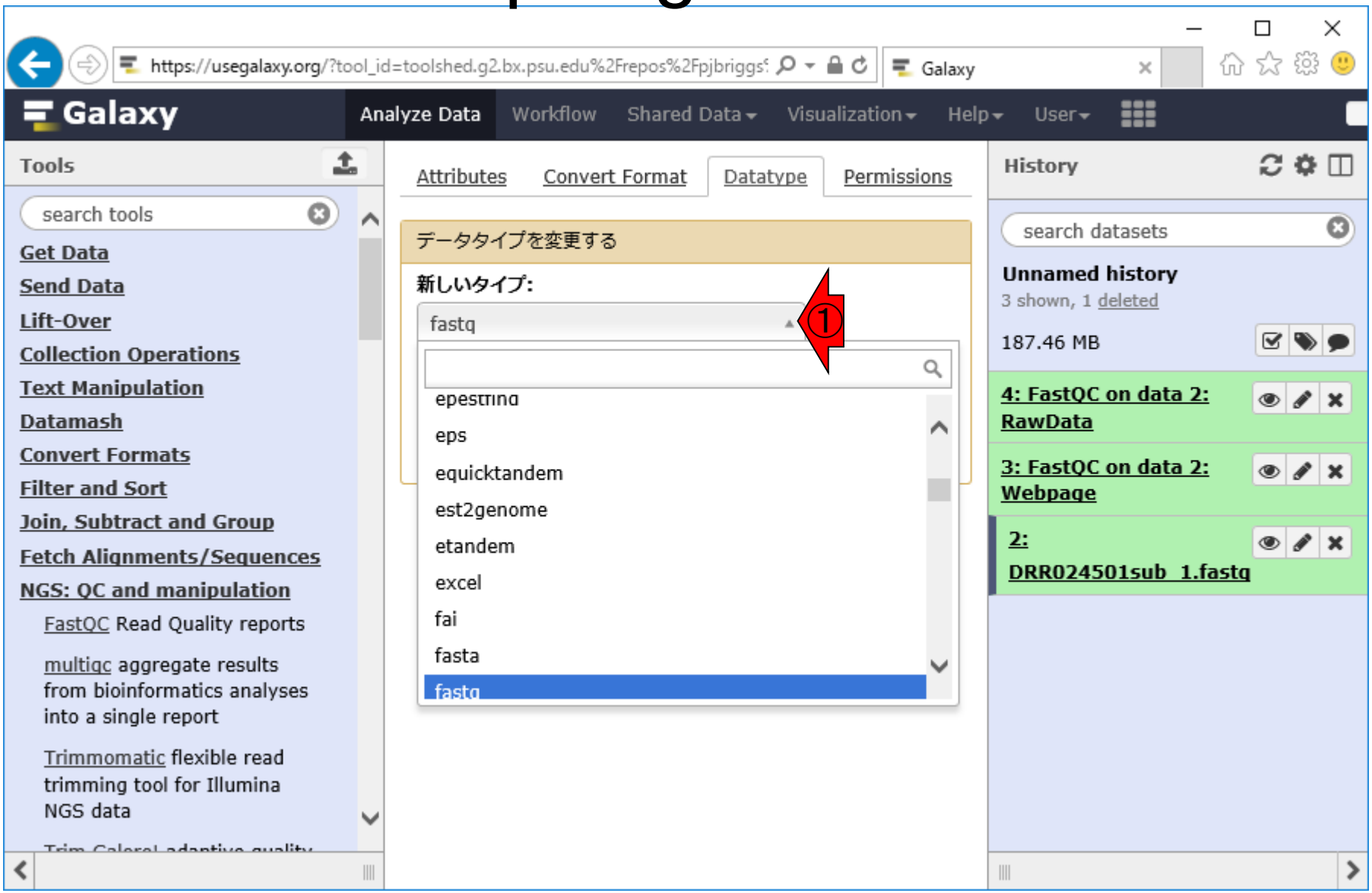
こんな風になってしまったら、横幅を広げる

W11-4：fastqsanger



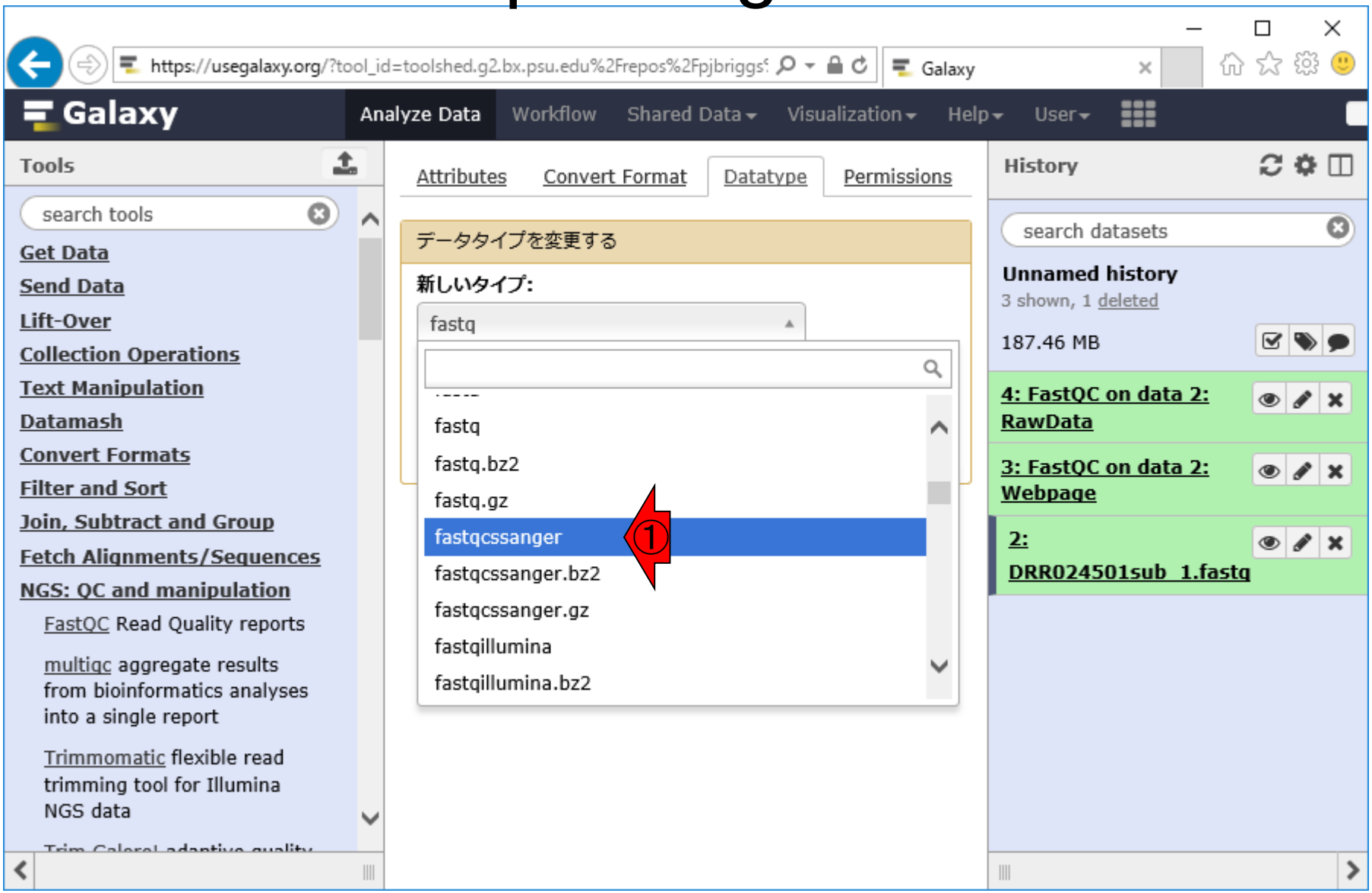
横幅を広げたところ

W11-4：fastqsanger



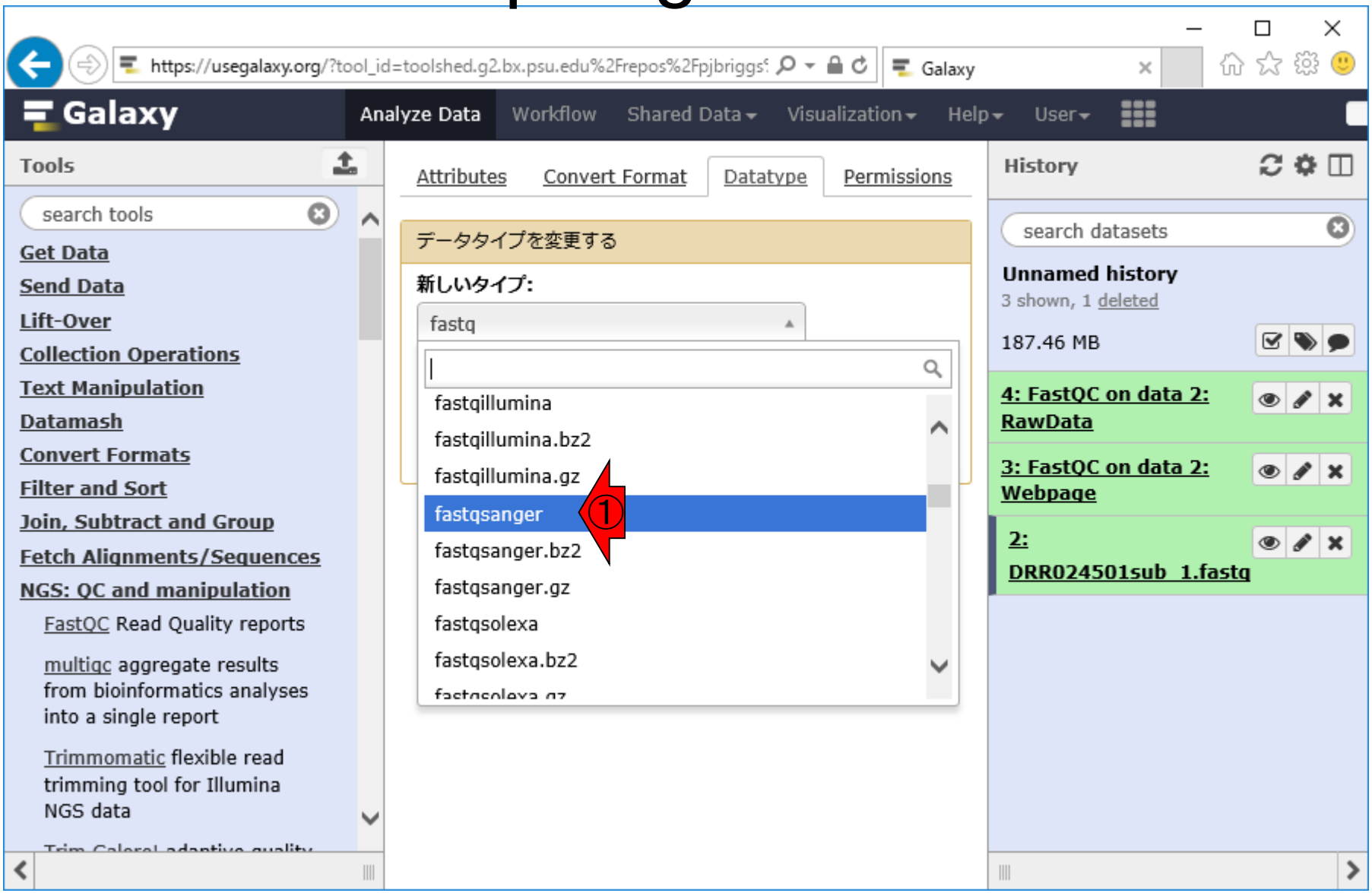
注意してください

W11-5：fastqcssangerではない



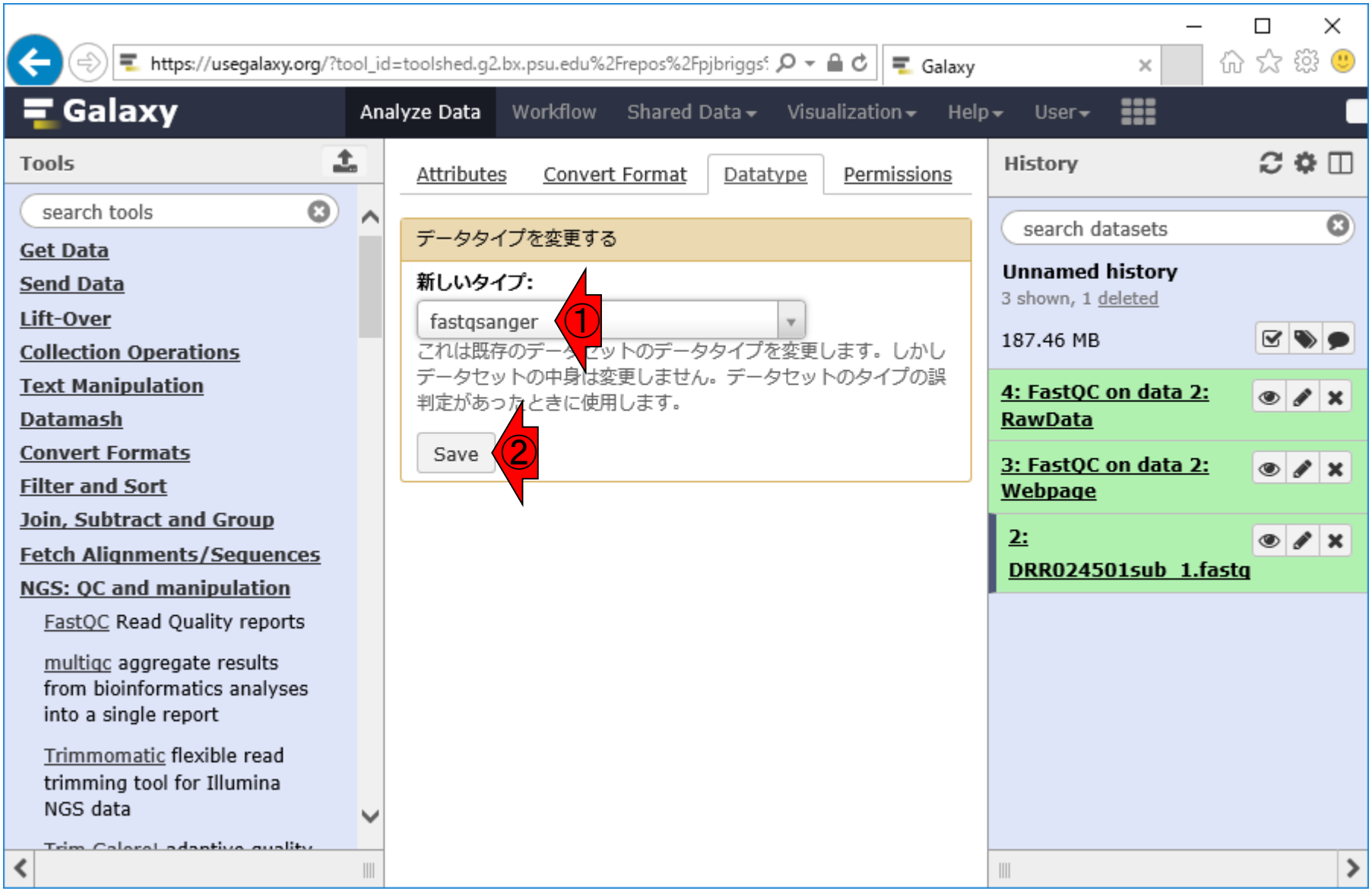
W11-6：fastqsanger

①fastqsangerがあるので、それを選択



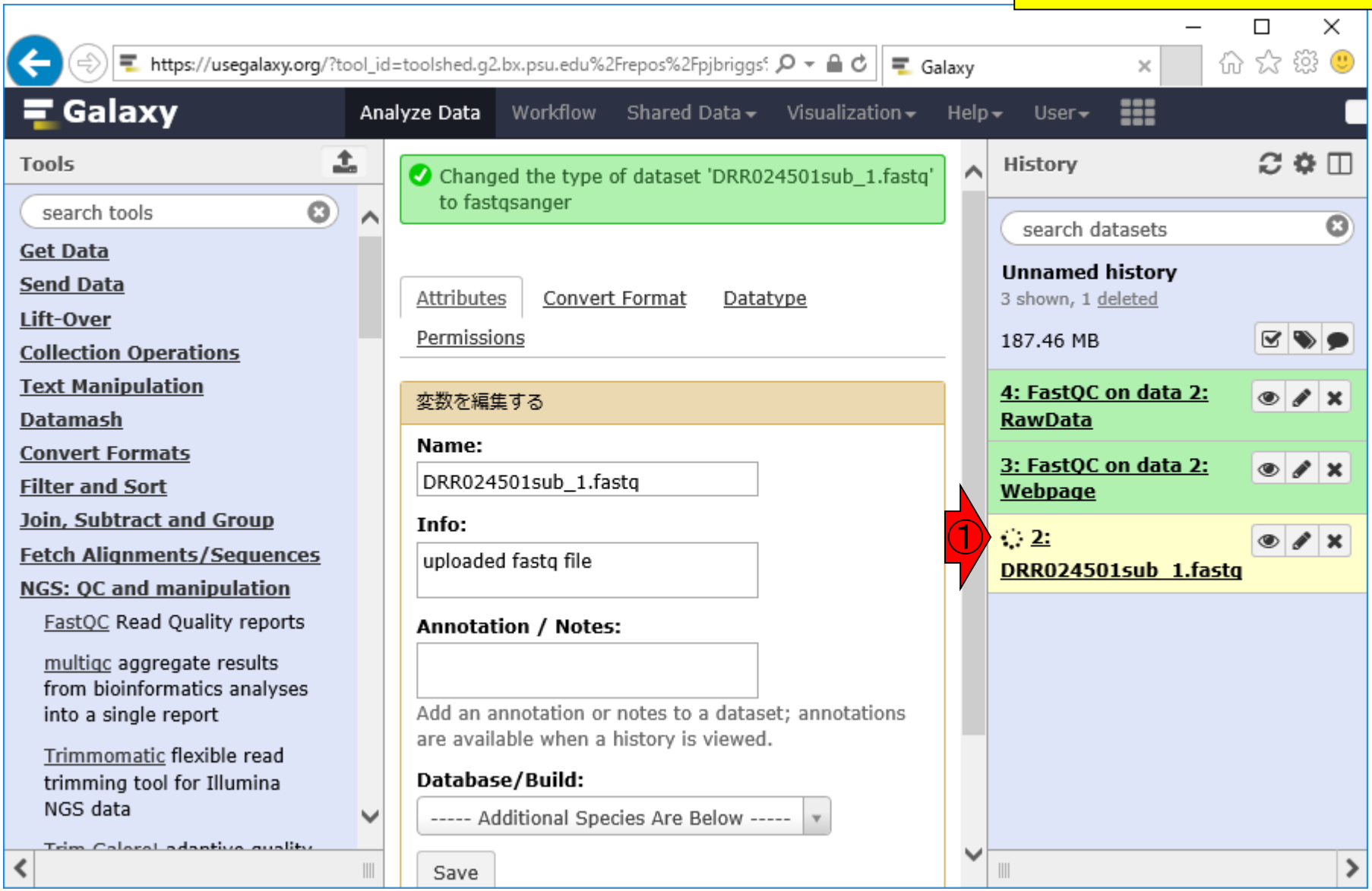
W11-7：Save

①fastqsangerになっていることを確認して、②Save



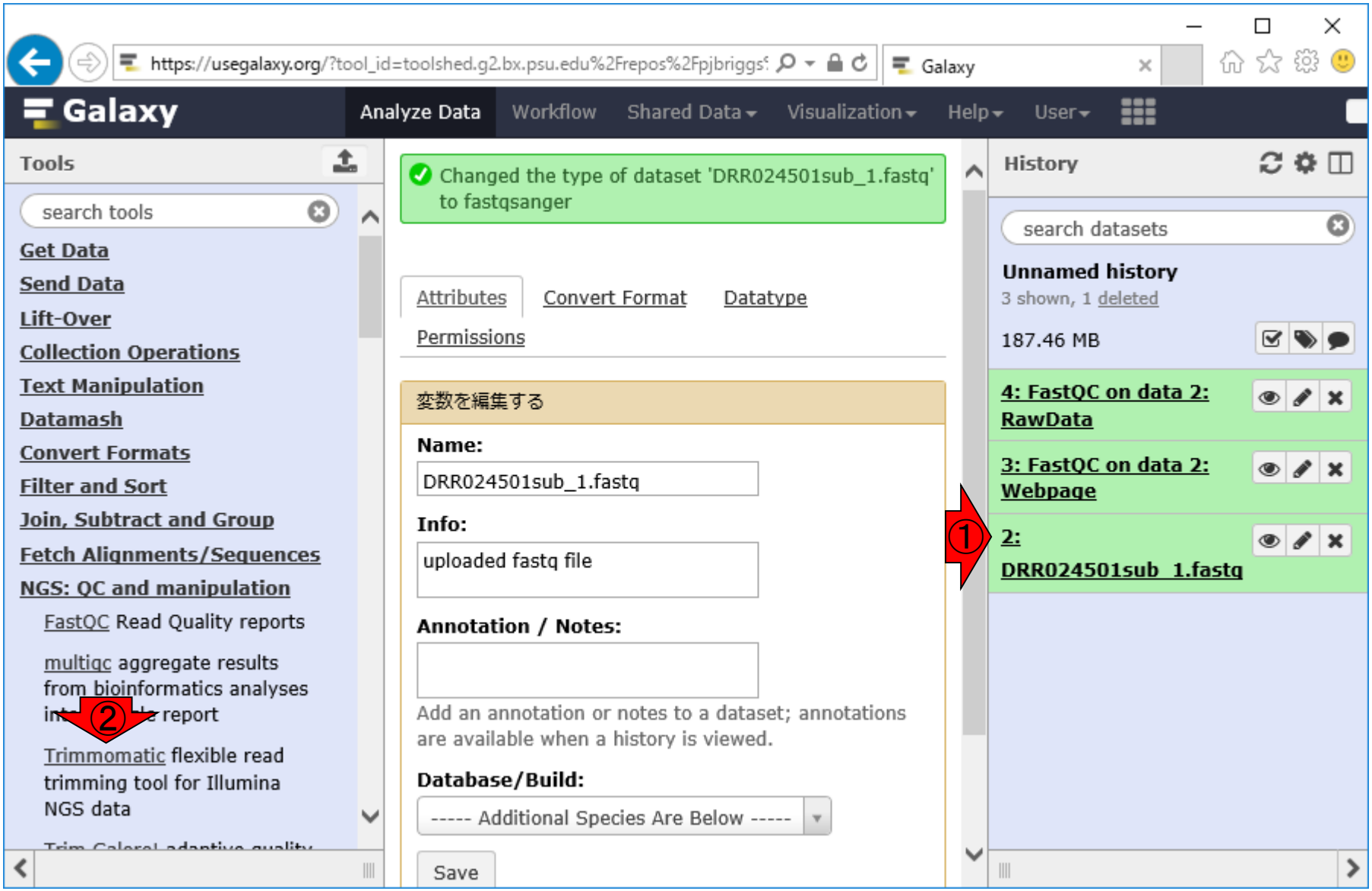
W11-8：Save中…

①Save実行中…。この状態が10分以上続くこともあります。以前やったときは1時間以上かかりました



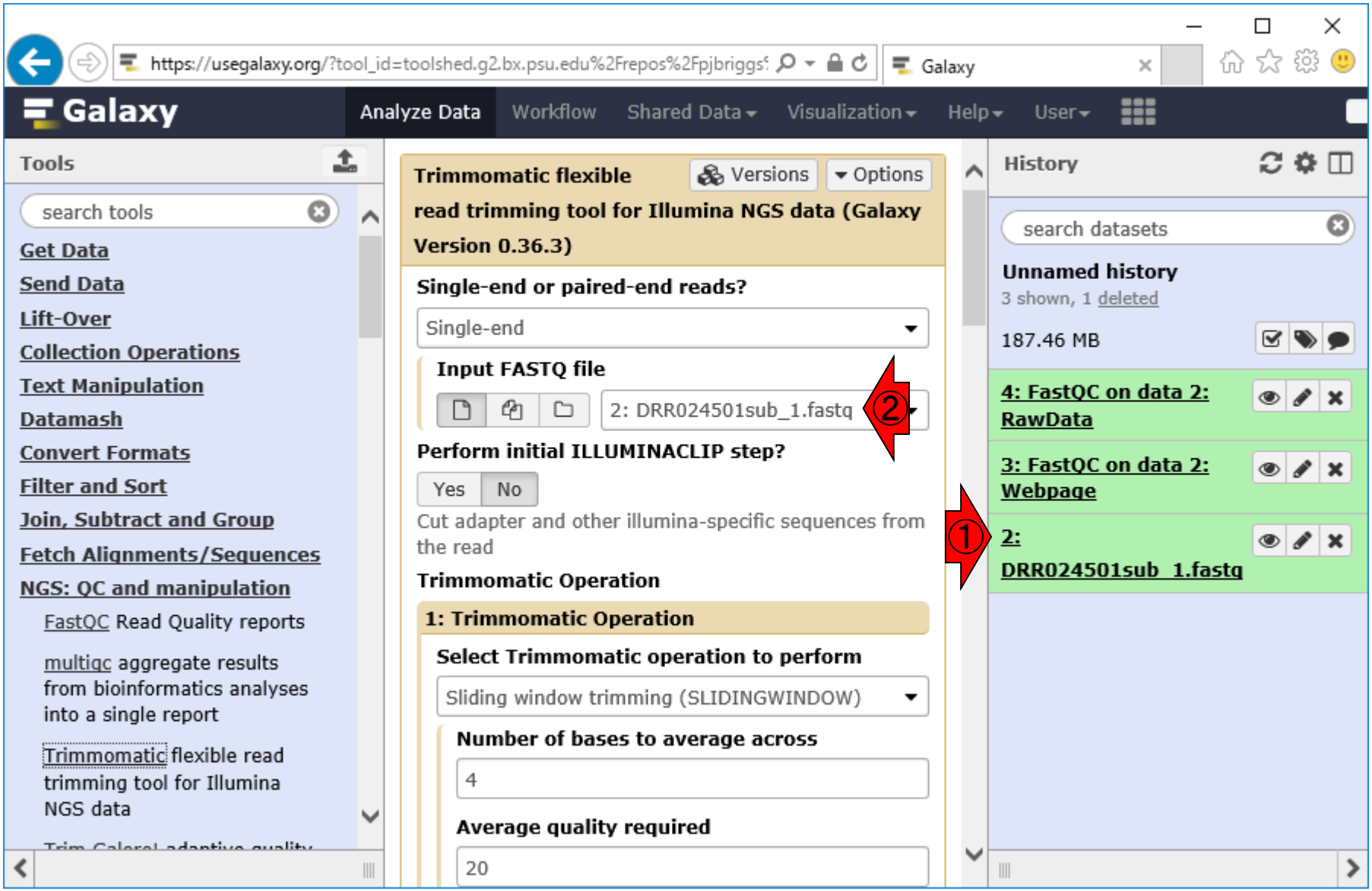
①Save終了。あらためて②Trimmomatic

W11-9：Save終了



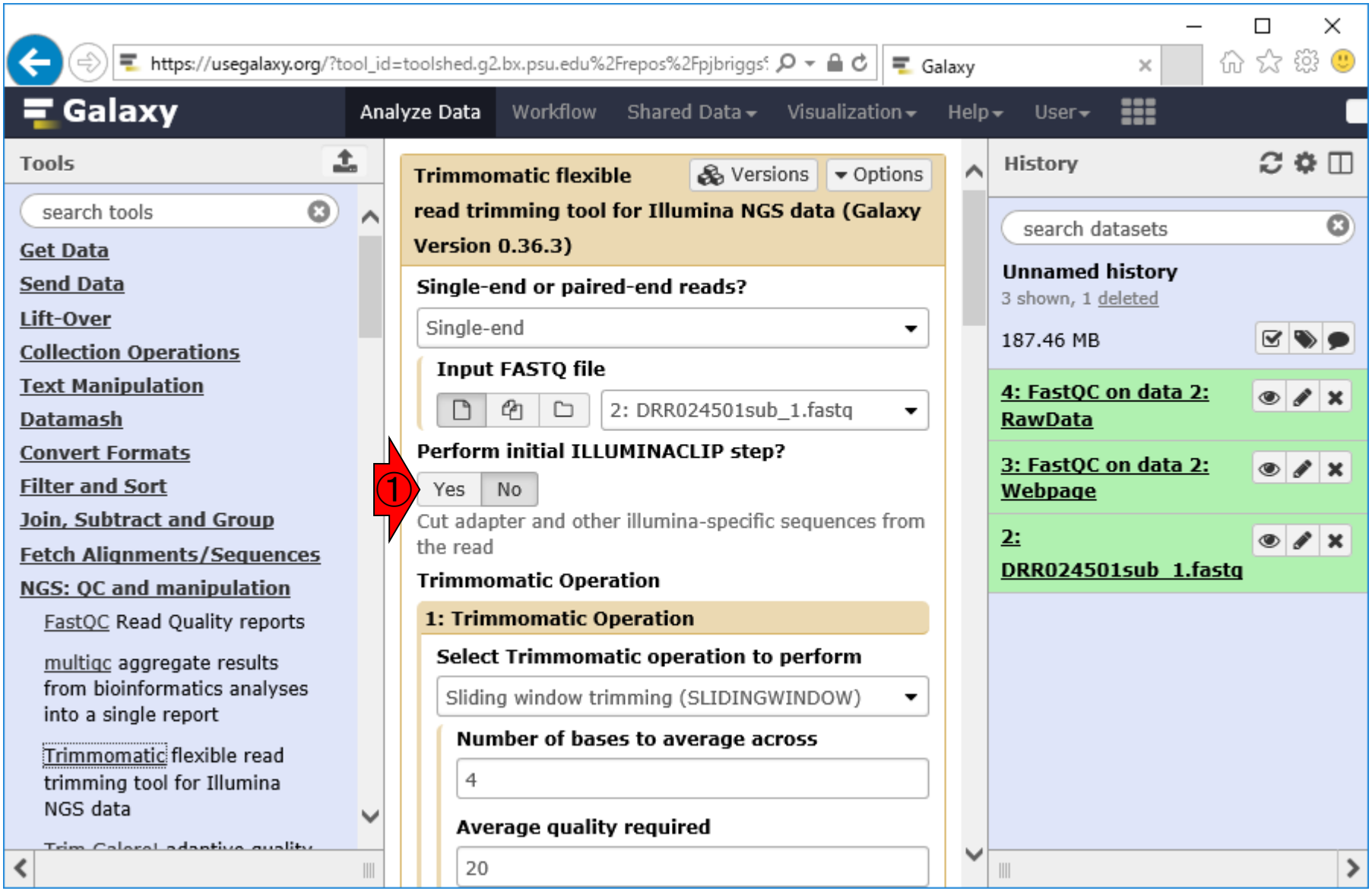
W12-1：Trimmomatic

①の入力ファイルが、②TrimmomaticのInput FASTQ file上で見られるようになりました



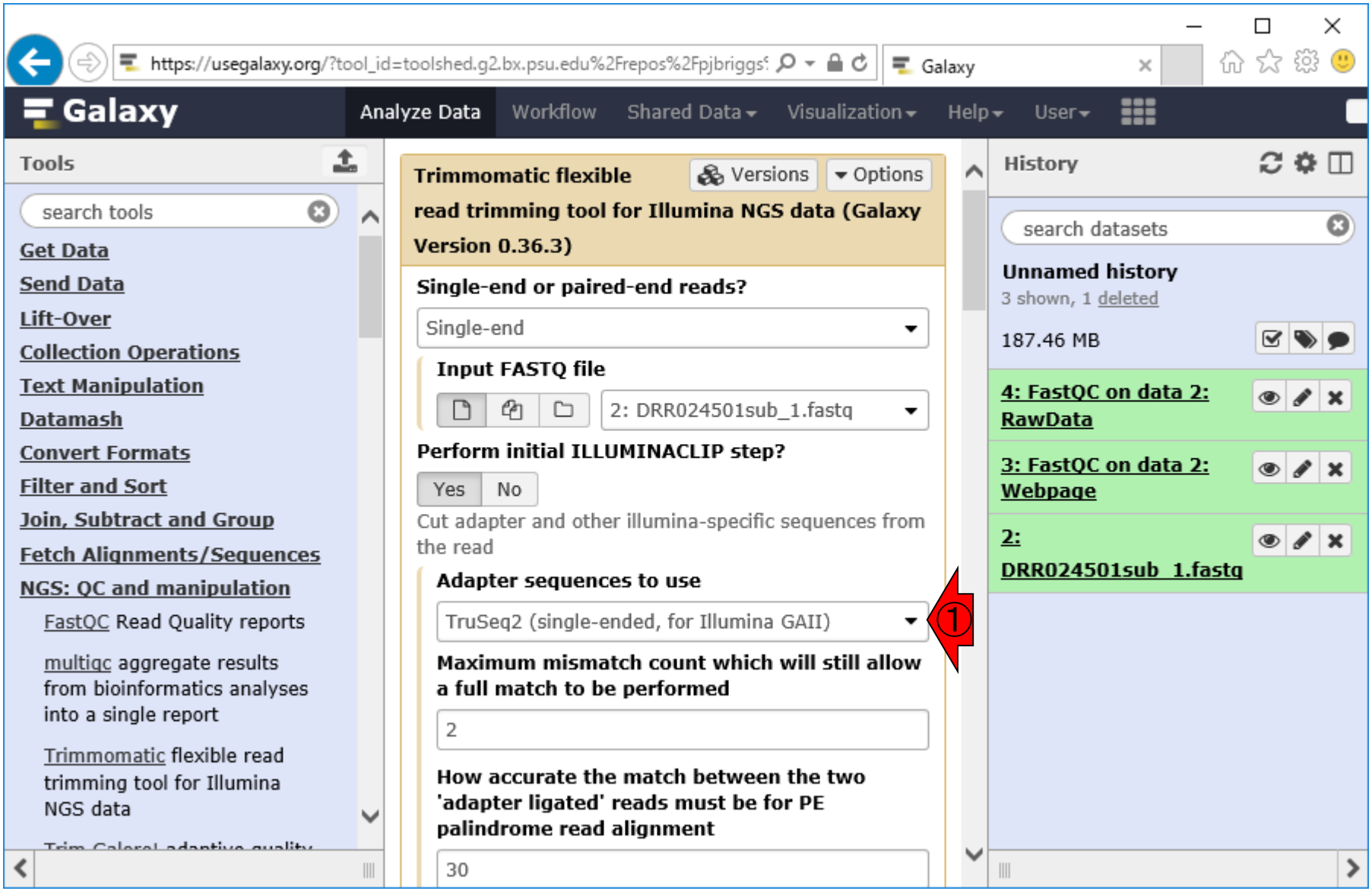
アダプター配列除去が目的なので、①Yes

W12-1：Trimmomatic



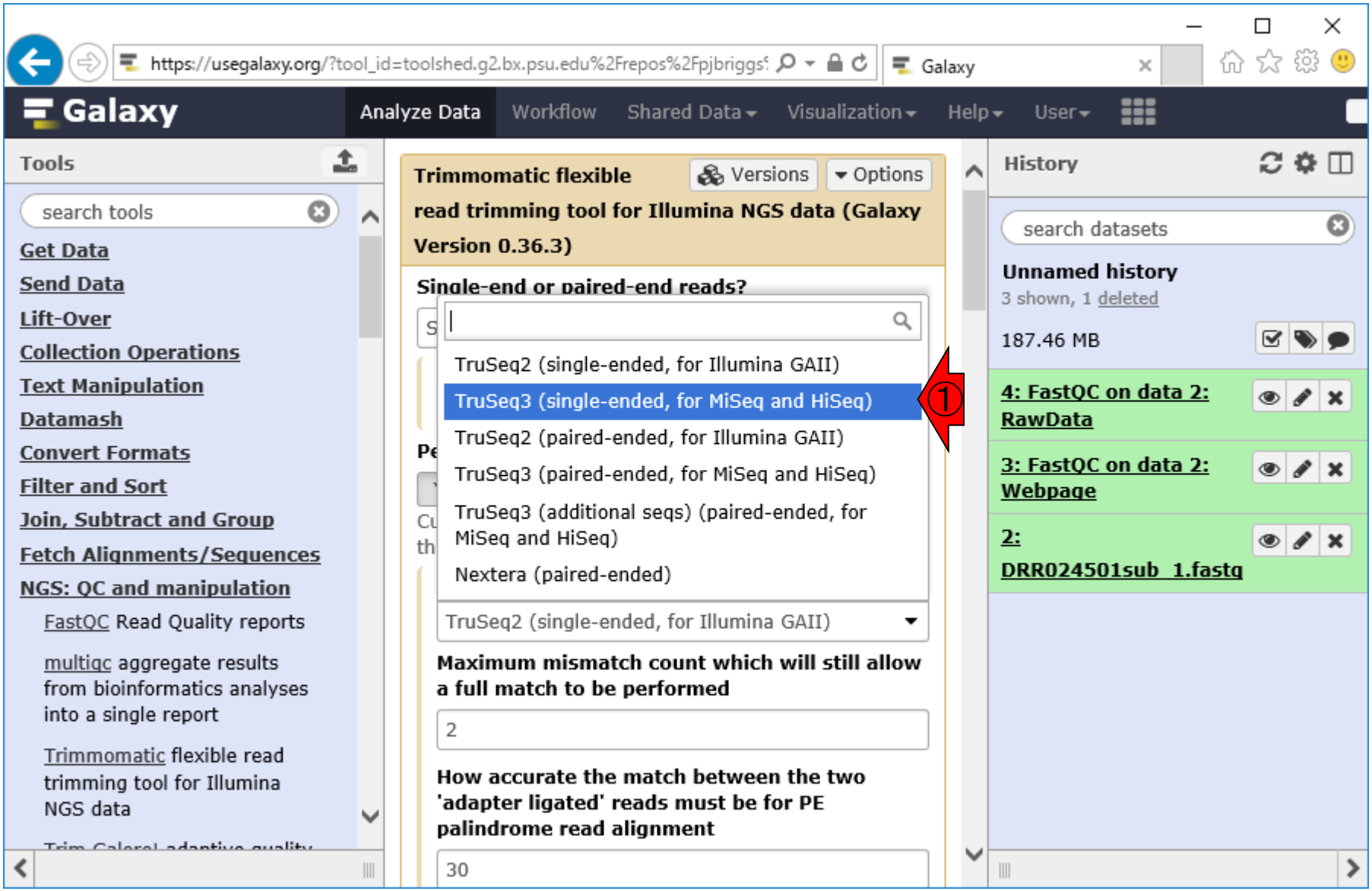
①でアダプター配列を指定

W12-2：アダプター配列を指定



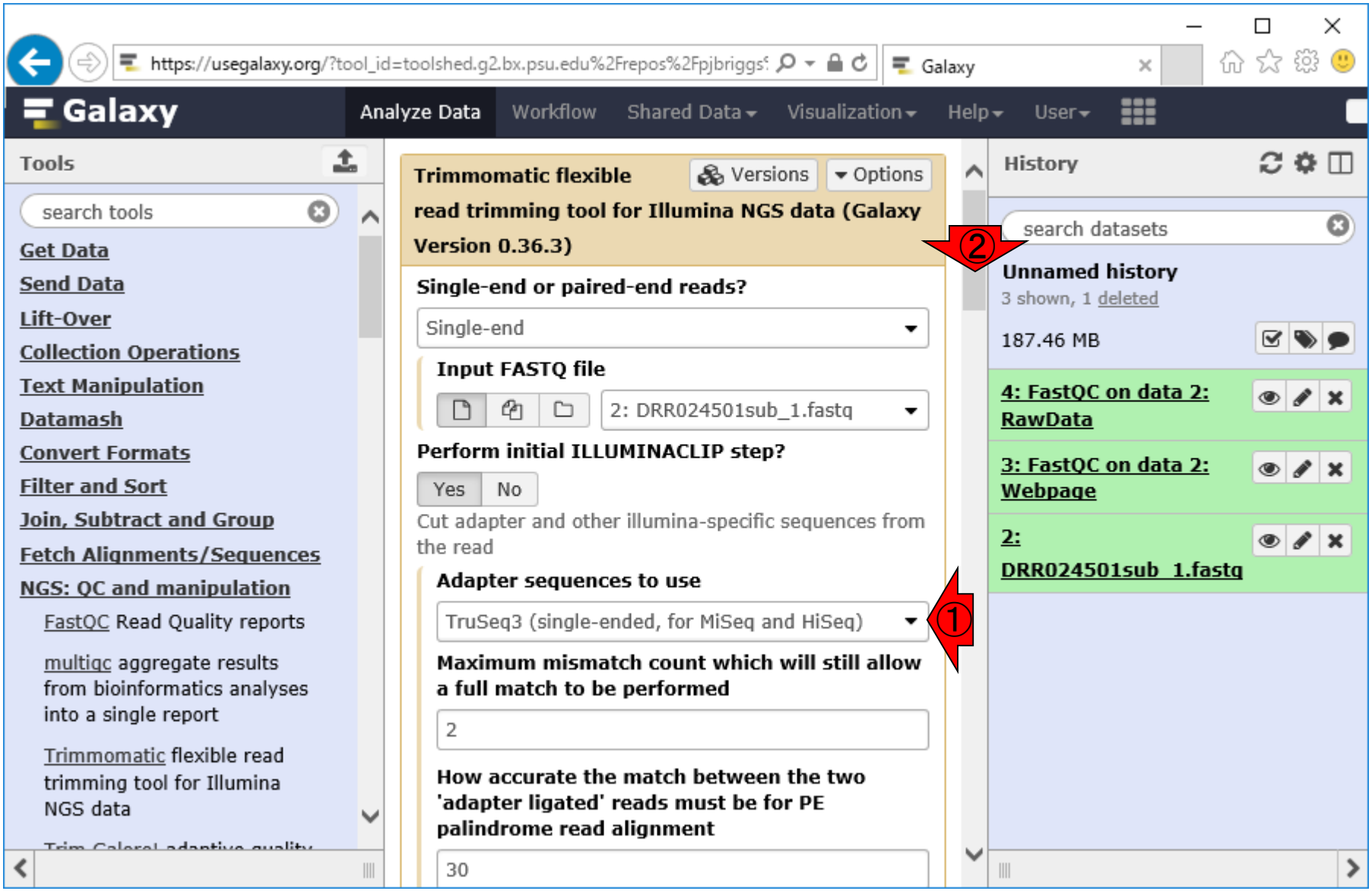
W12-2：アダプター配列を指定

入力ファイルはsingle-end MiSeqデータなので①



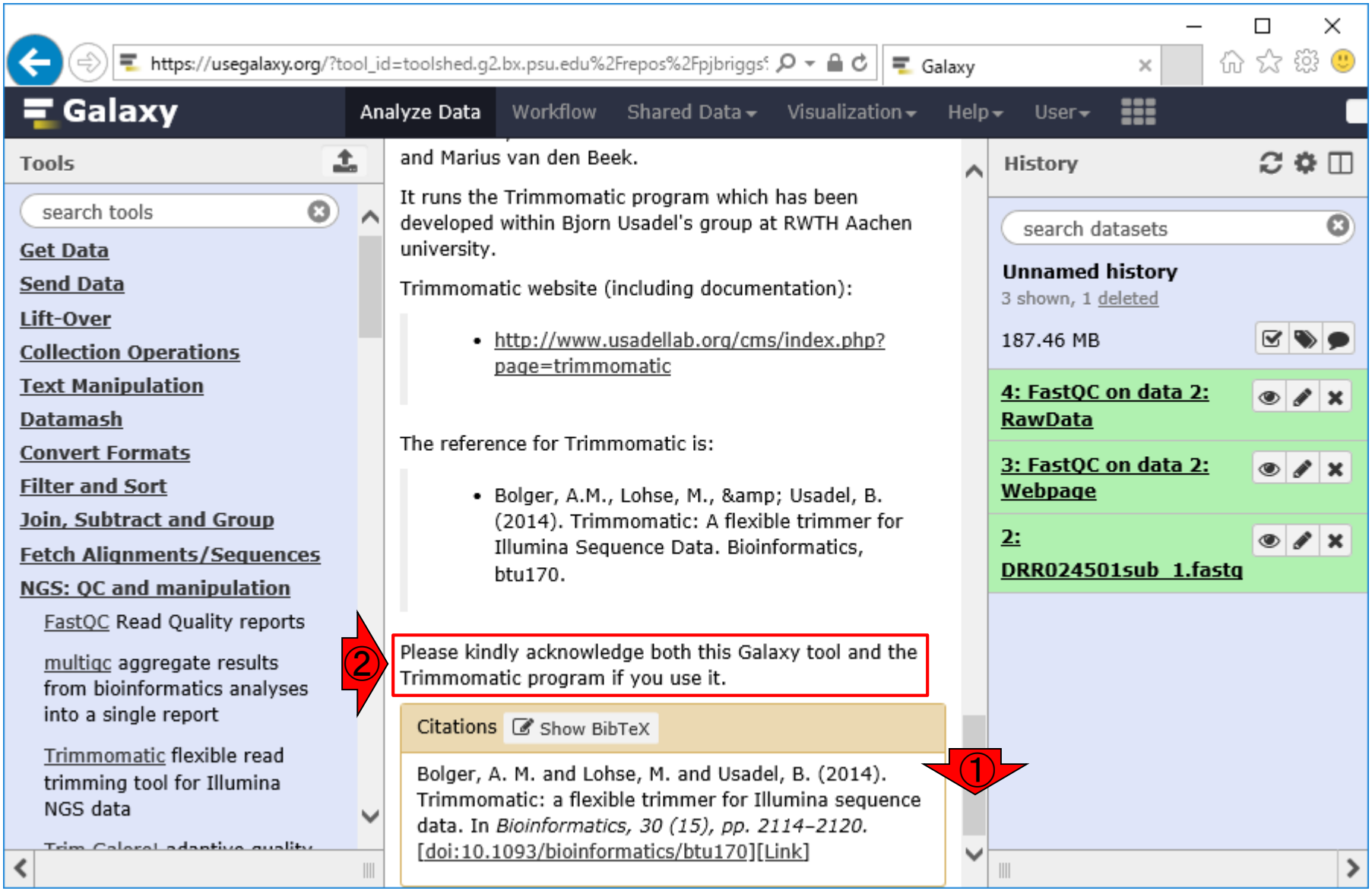
W12-3：ページ最下部へ

①無事TruSeq3になっていることを確認して、②ページ最下部に移動



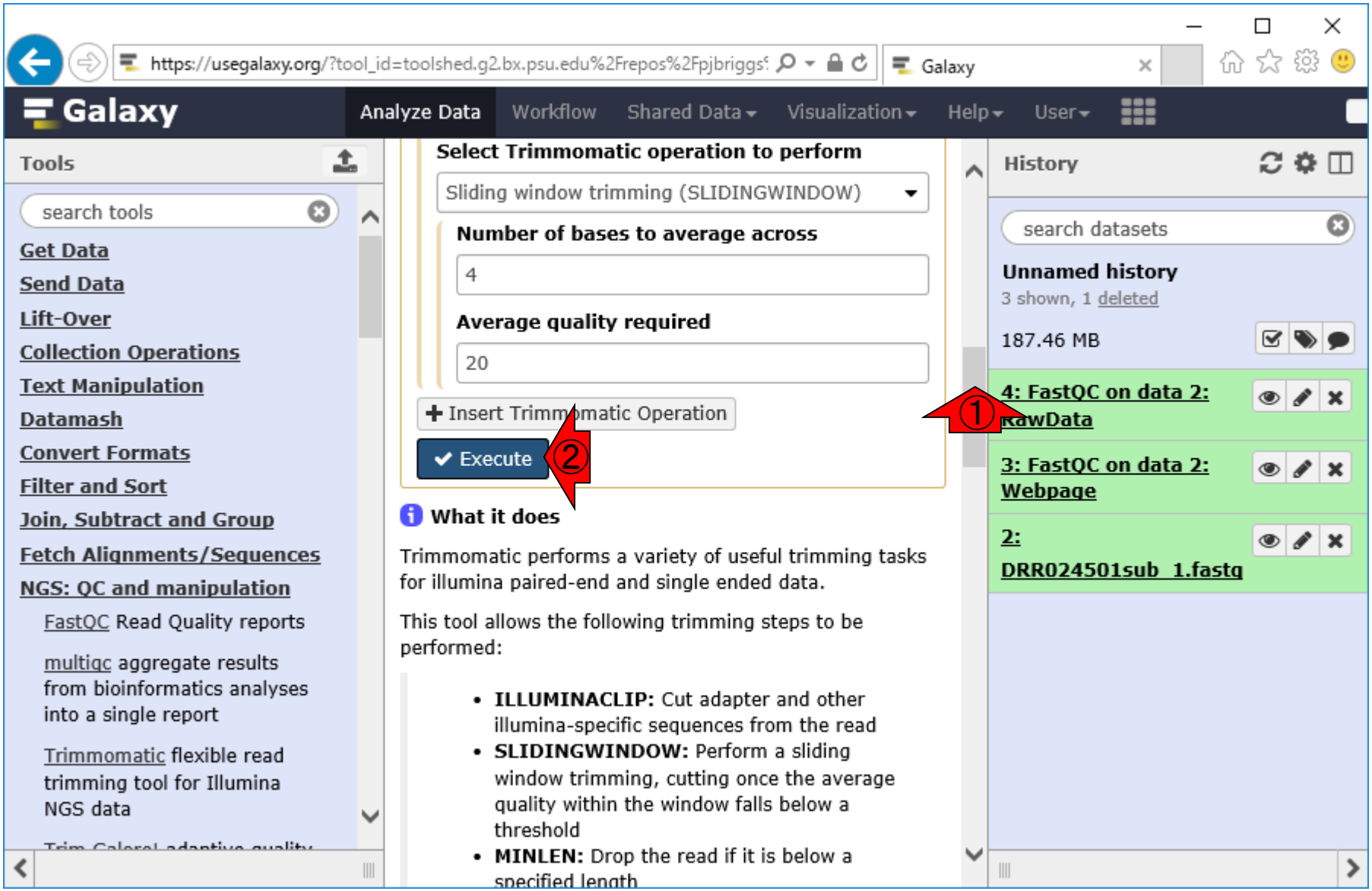
W12-4：Citations

①ページ最下部。②この場合の引用文献は、GalaxyとTrimmomaticです。お忘れなきよう



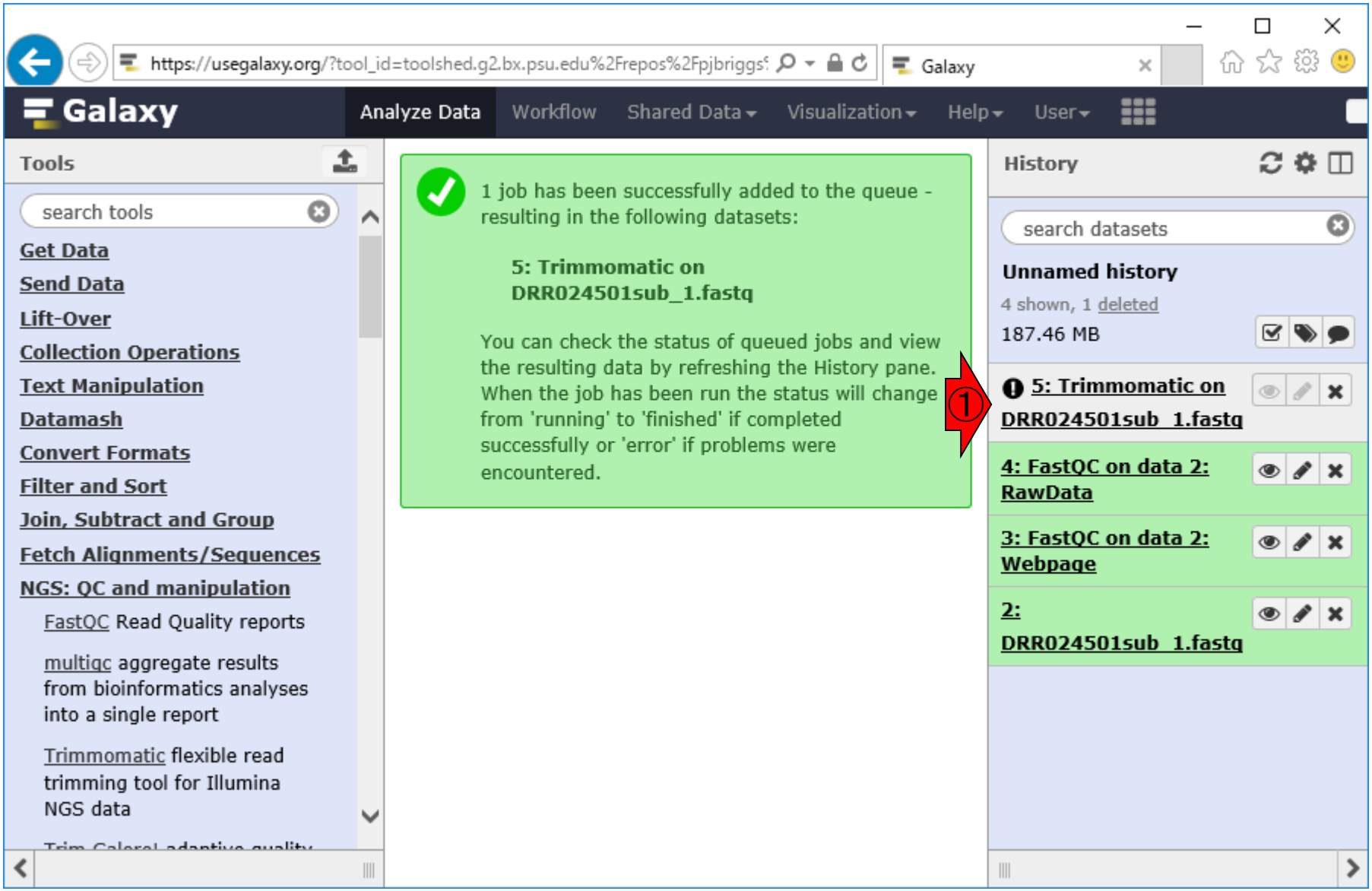
①のあたりまで戻って、②Execute

W12-5：Trimmomatic実行



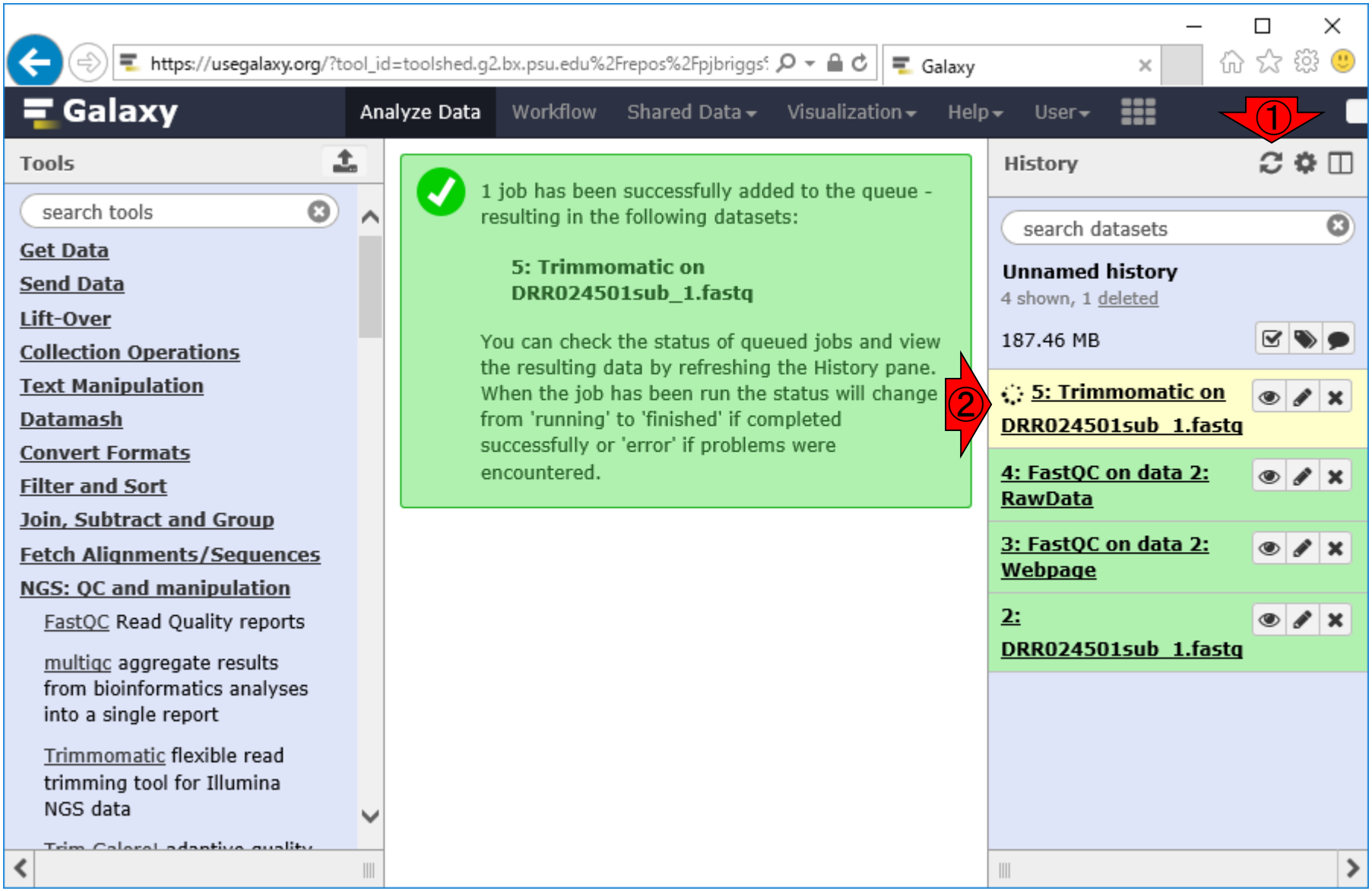
①実行待ち状態（灰色）

W12-6：実行待ち状態



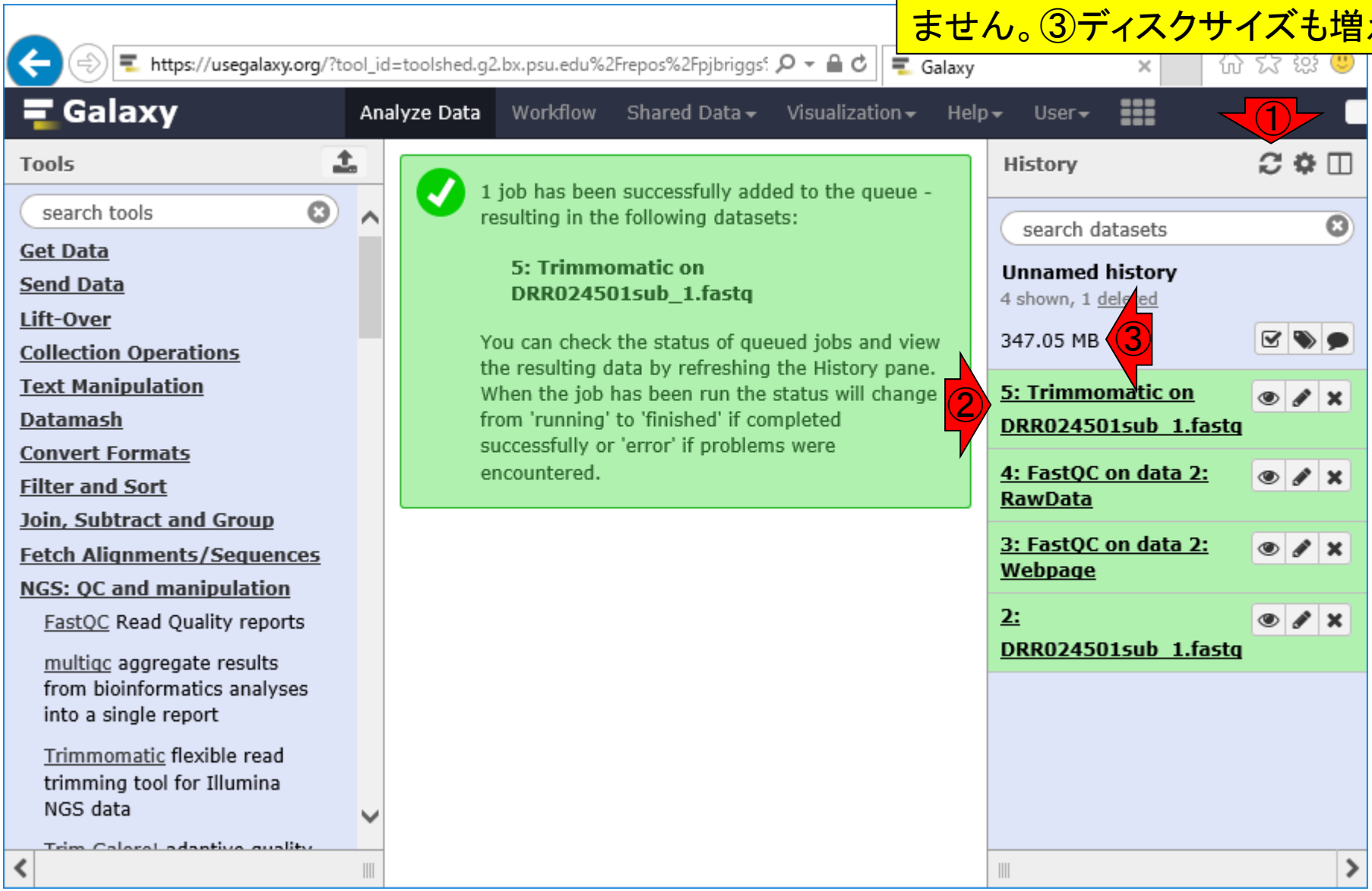
W12-7：実行中

①約1分後にリロードしてみると、
②実行中（黄色）になっていました



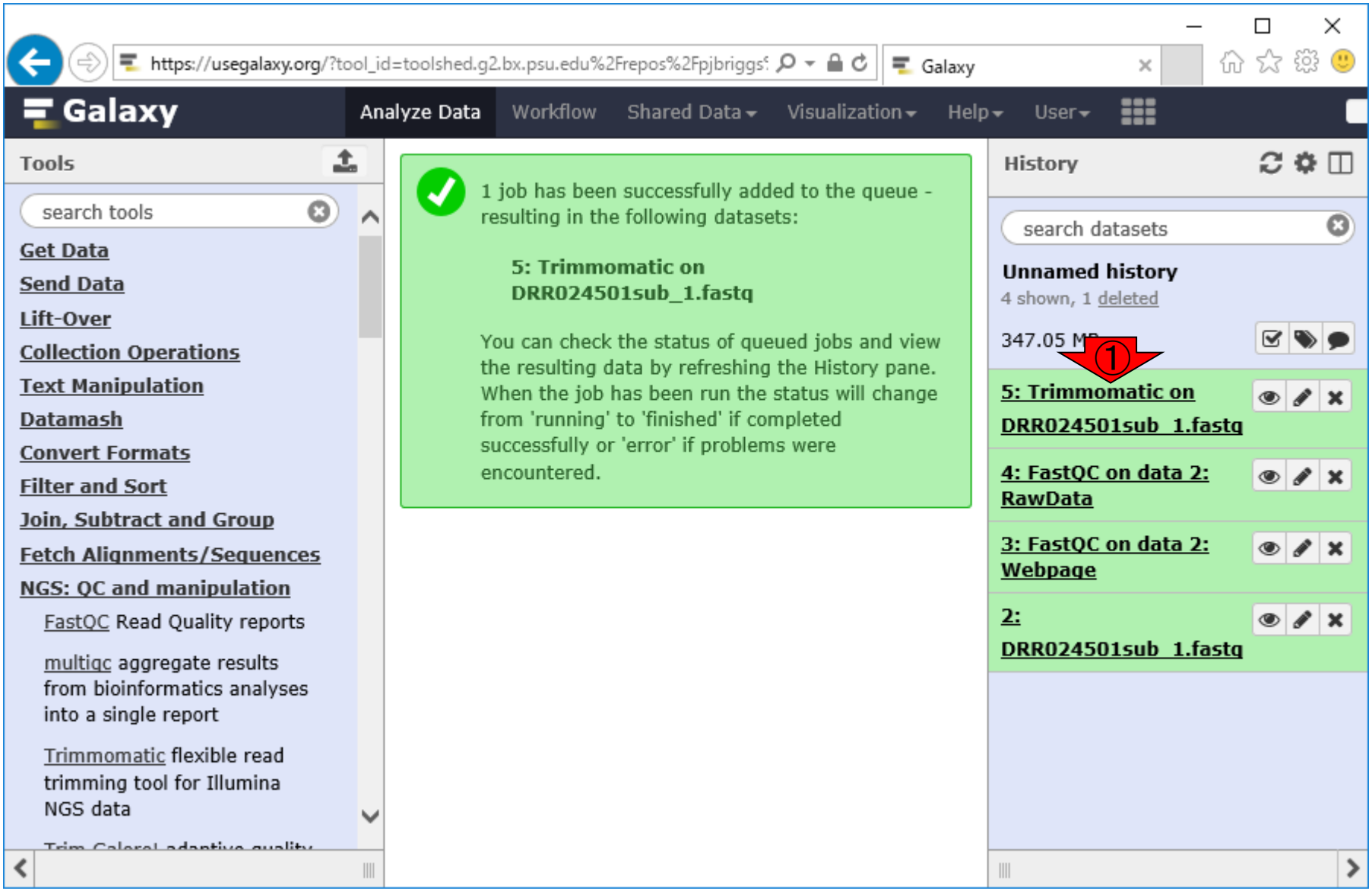
W12-8：実行終了

①さらに約1分後にリロードしてみると、②実行終了（緑色）になっていました。定期的なリロードによるチェックは重要かもしれません。③ディスクサイズも増えてますね



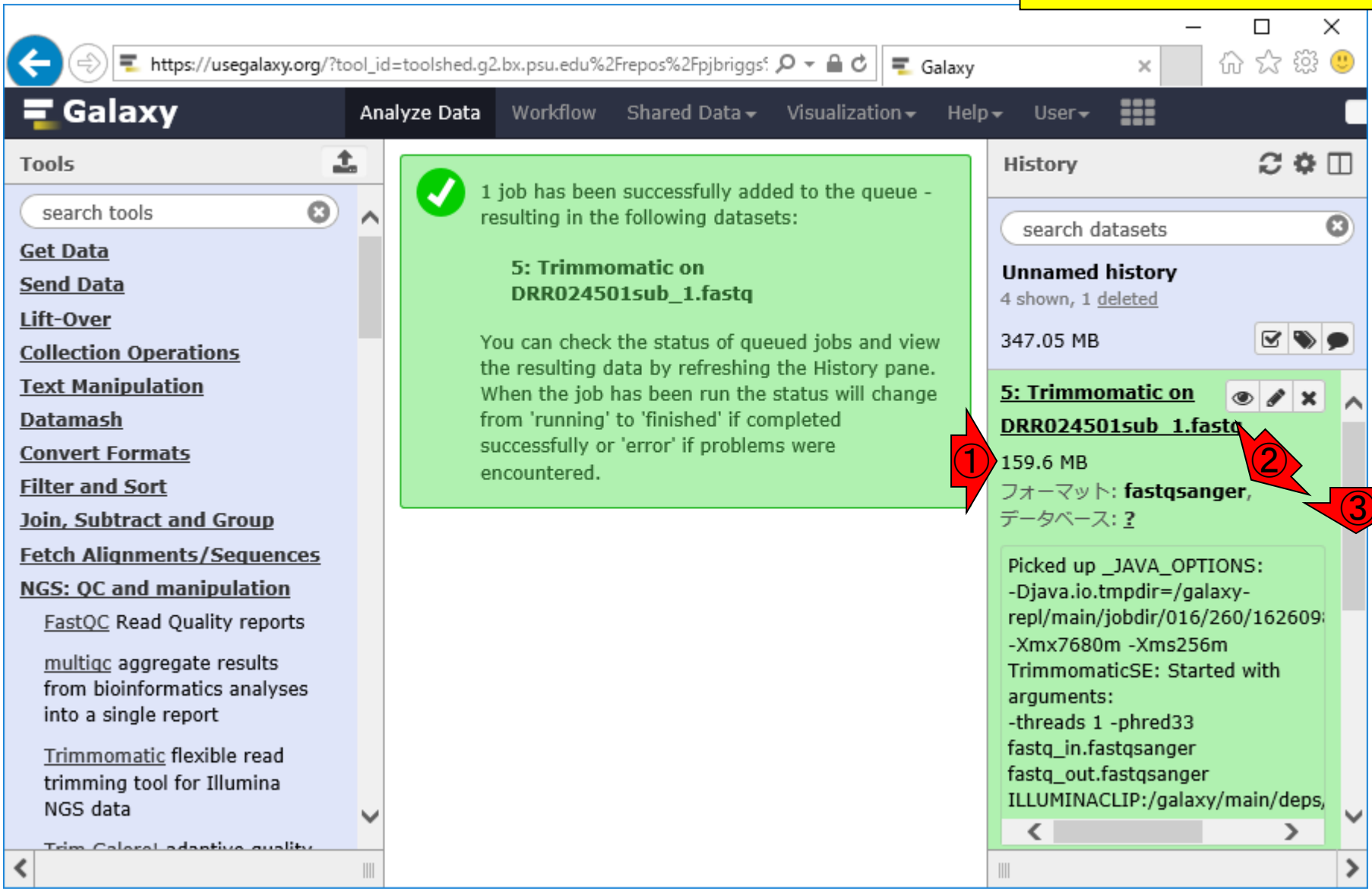
W12-9：結果のダウンロード

①のところをクリック。基本的にW10−4と同じ作業です



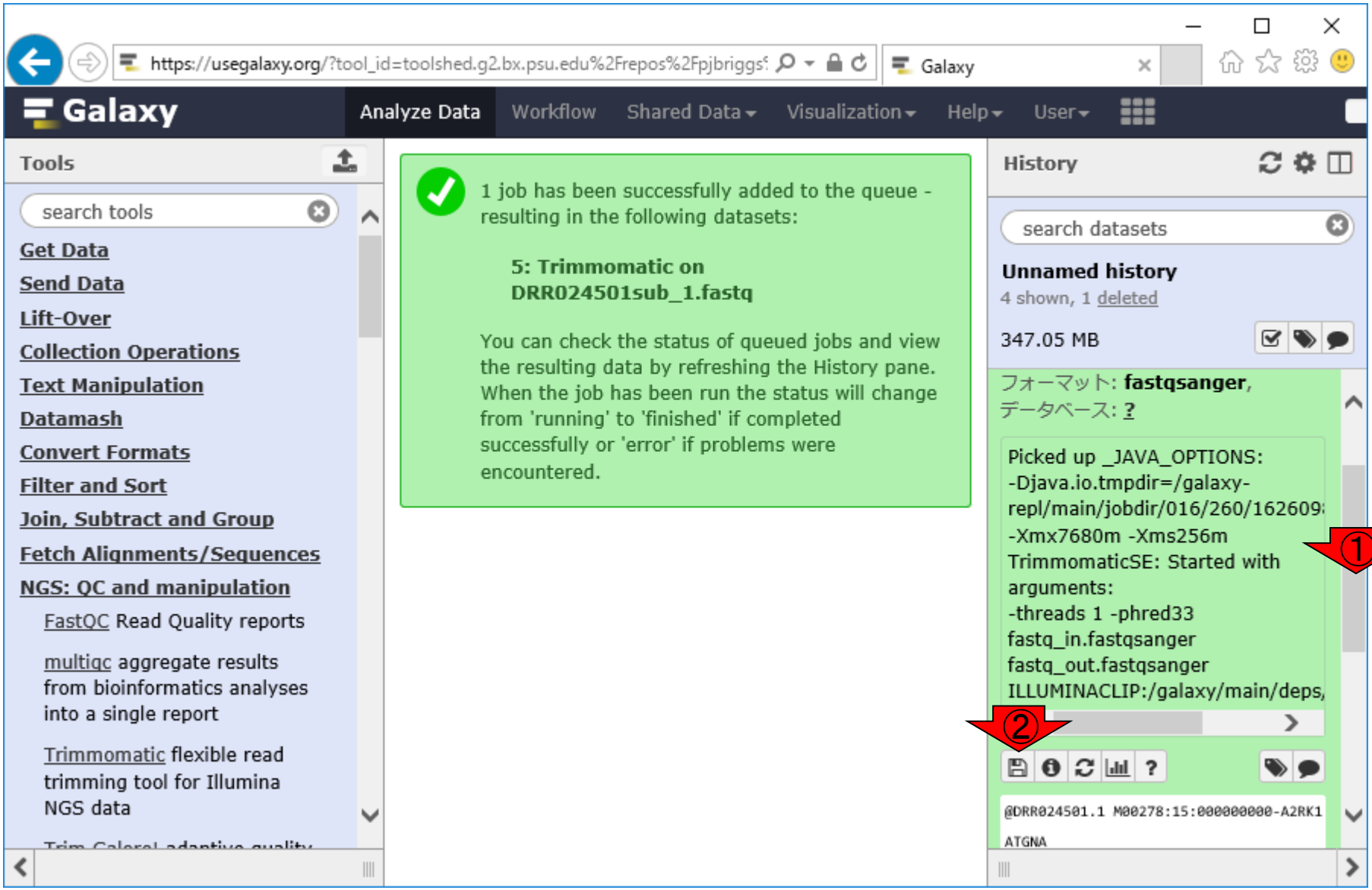
W12-9：結果のダウンロード

①ファイルサイズ。②Trimmomaticの出力ファイルはfastqsanger形式になってますね。③下部に移動



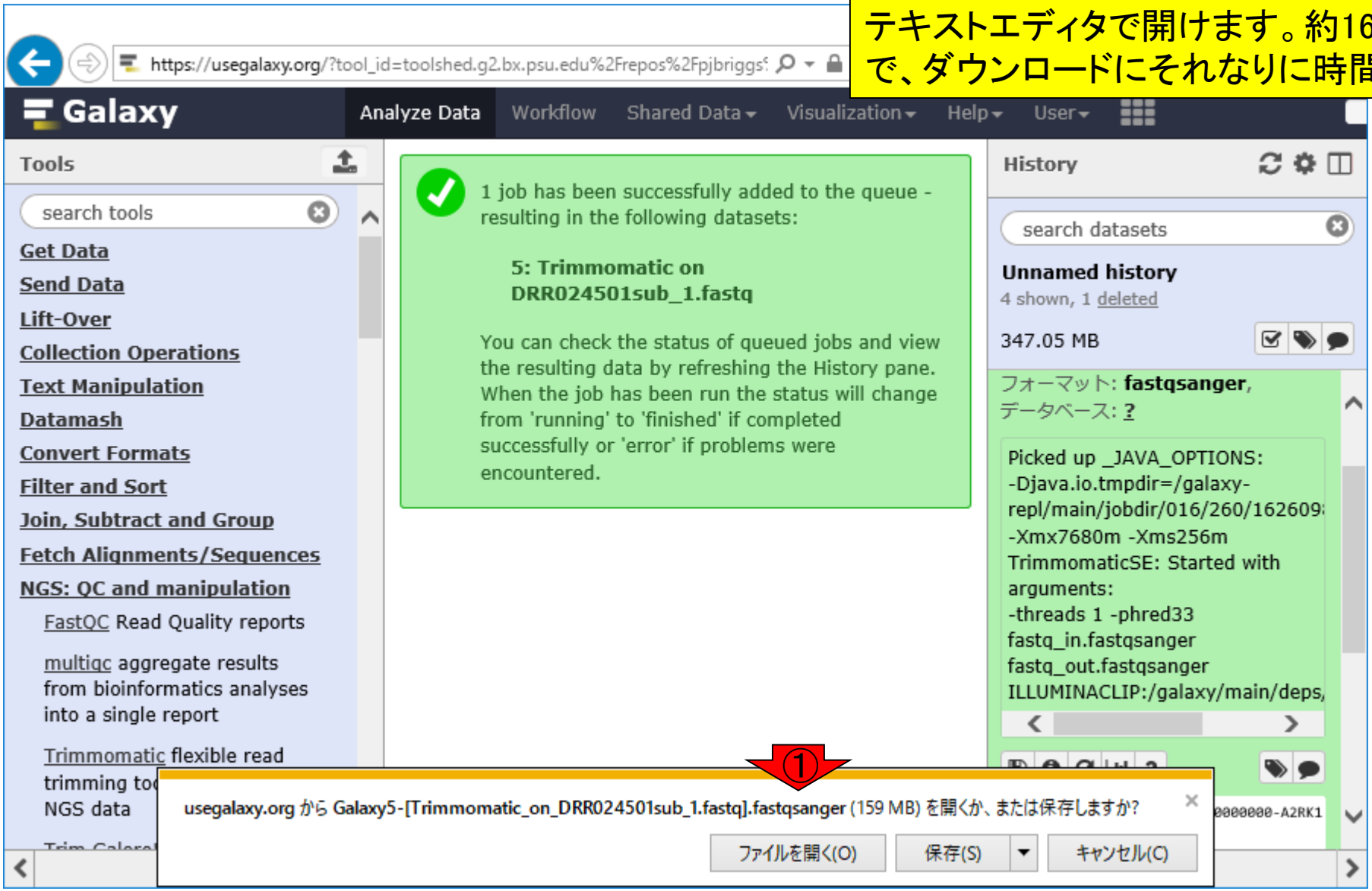
①このあたり。②Save

W12-9：結果のダウンロード

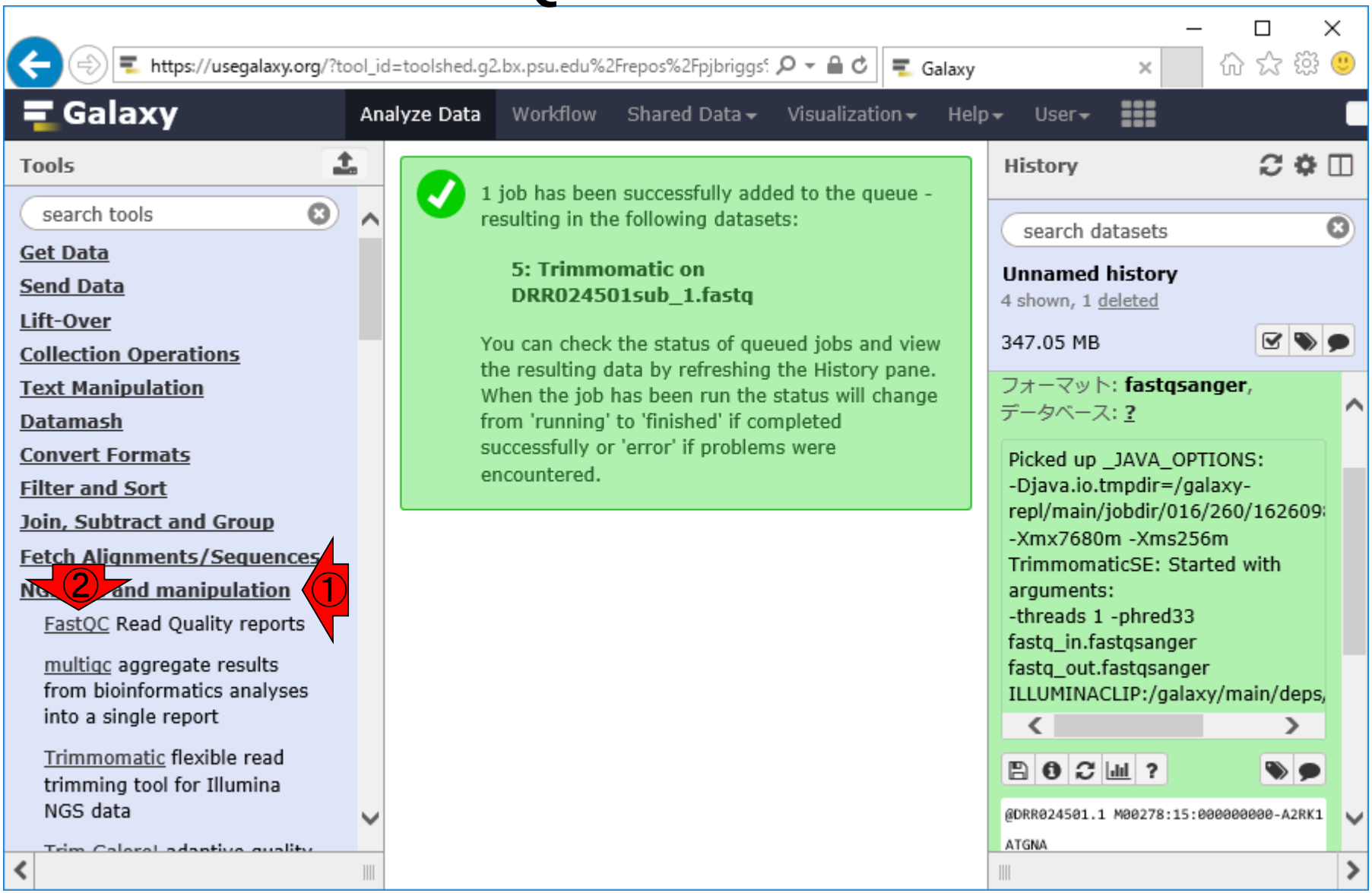


W12-9：結果のダウンロ

①で示すファイル名、およびfastqsangerという拡張子で保存されます。zip圧縮されているわけではないので、ダウンロード後は普通にテキストエディタで開けます。約160MBあるので、ダウンロードにそれなりに時間はかかる

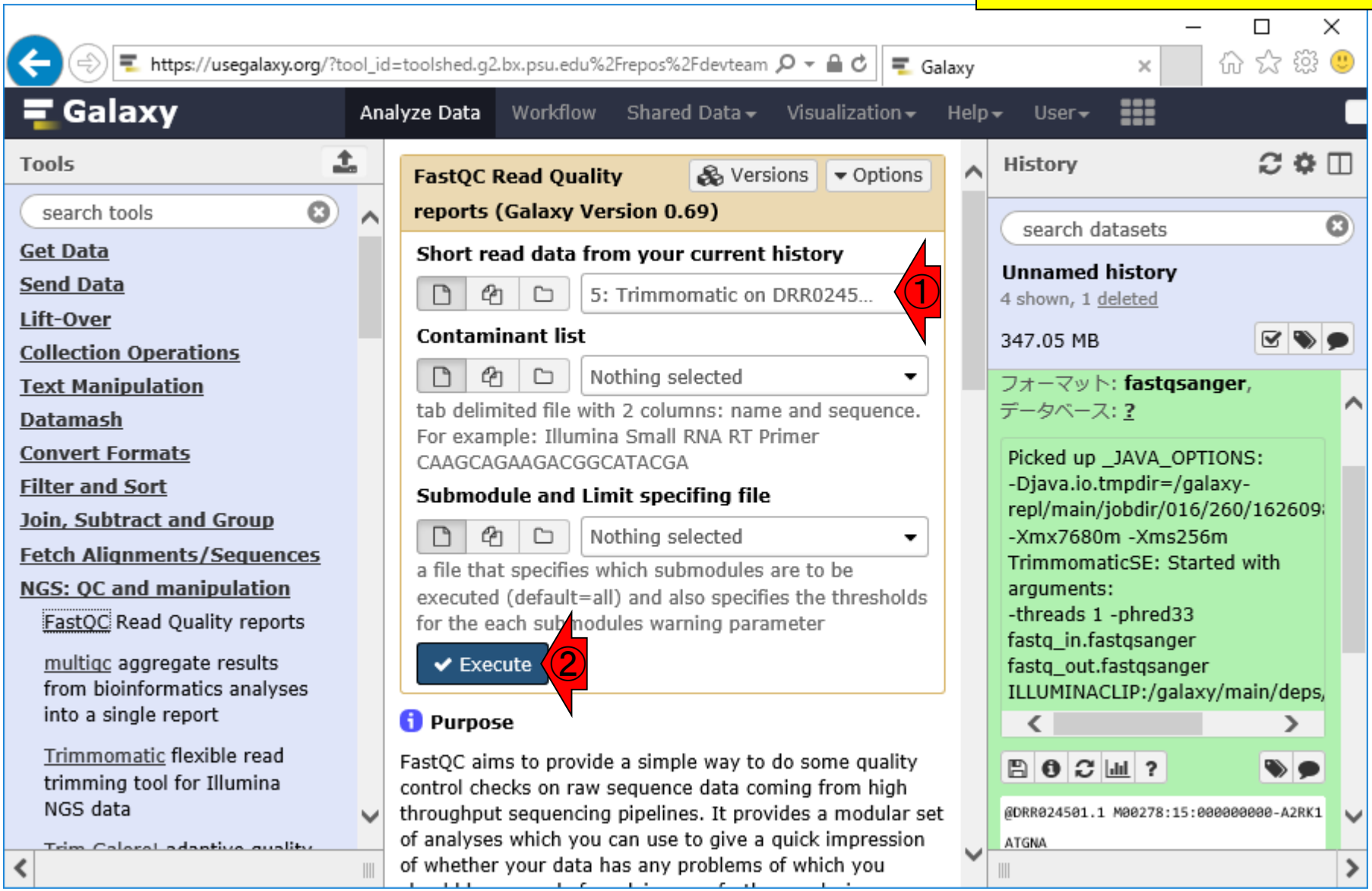


W13-1：FastQC


https://usegalaxy.org/?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fpjbriggs
Galaxy
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
search tools
Get Data
Send Data
Lift-Over
Collection Operations
Text Manipulation
Datamash
Convert Formats
Filter and Sort
Join, Subtract and Group
Fetch Alignments/Sequences
and manipulation
FastQC Read Quality reports
multiqc aggregate results from bioinformatics analyses into a single report
Trimmomatic flexible read trimming tool for Illumina NGS data
②
①
1 job has been successfully added to the queue - resulting in the following datasets:
5: Trimmomatic on DRR024501sub_1.fastq
You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.
History
search datasets
Unnamed history
4 shown, 1 deleted
347.05 MB
フォーマット: fastqsanger,
データベース: ?
Picked up _JAVA_OPTIONS:
-Djava.io.tmpdir=/galaxy-repl/main/jobdir/016/260/162609
-Xmx7680m -Xms256m
TrimmomaticSE: Started with arguments:
-threads 1 -phred33
fastq_in.fastqsanger
fastq_out.fastqsanger
ILLUMINACLIP:/galaxy/main/deps,
@DRR024501.1 M00278:15:000000000-A2RK1
ATGNA

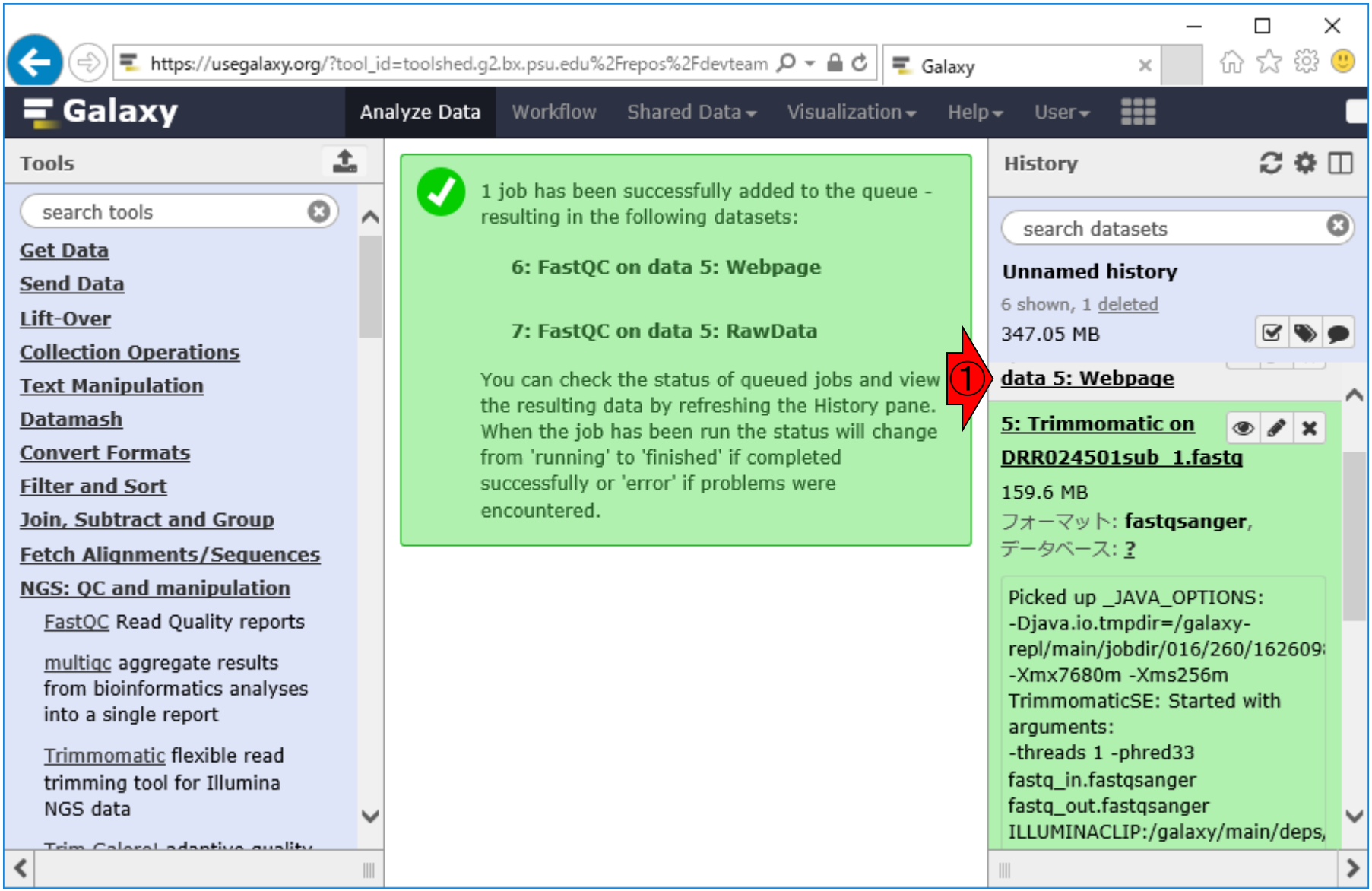
W13-2：Execute

①最後に作成したTrimmomatic実行後のFASTQファイルがデフォルトで見えているので、そのまま②Execute



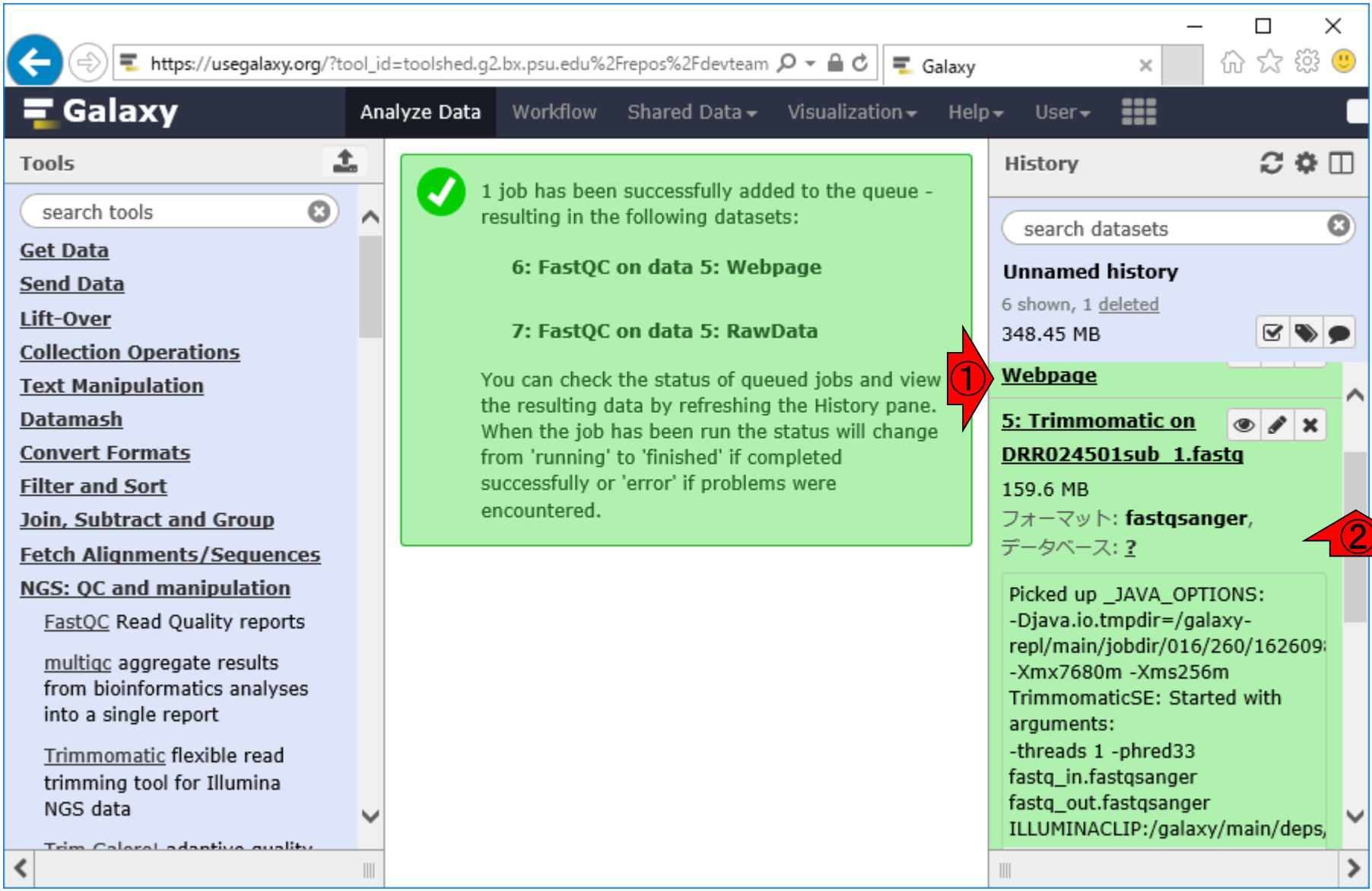
W13-3：実行待ち

①実行待ち状態（灰色）であることが、このあたりでわかります



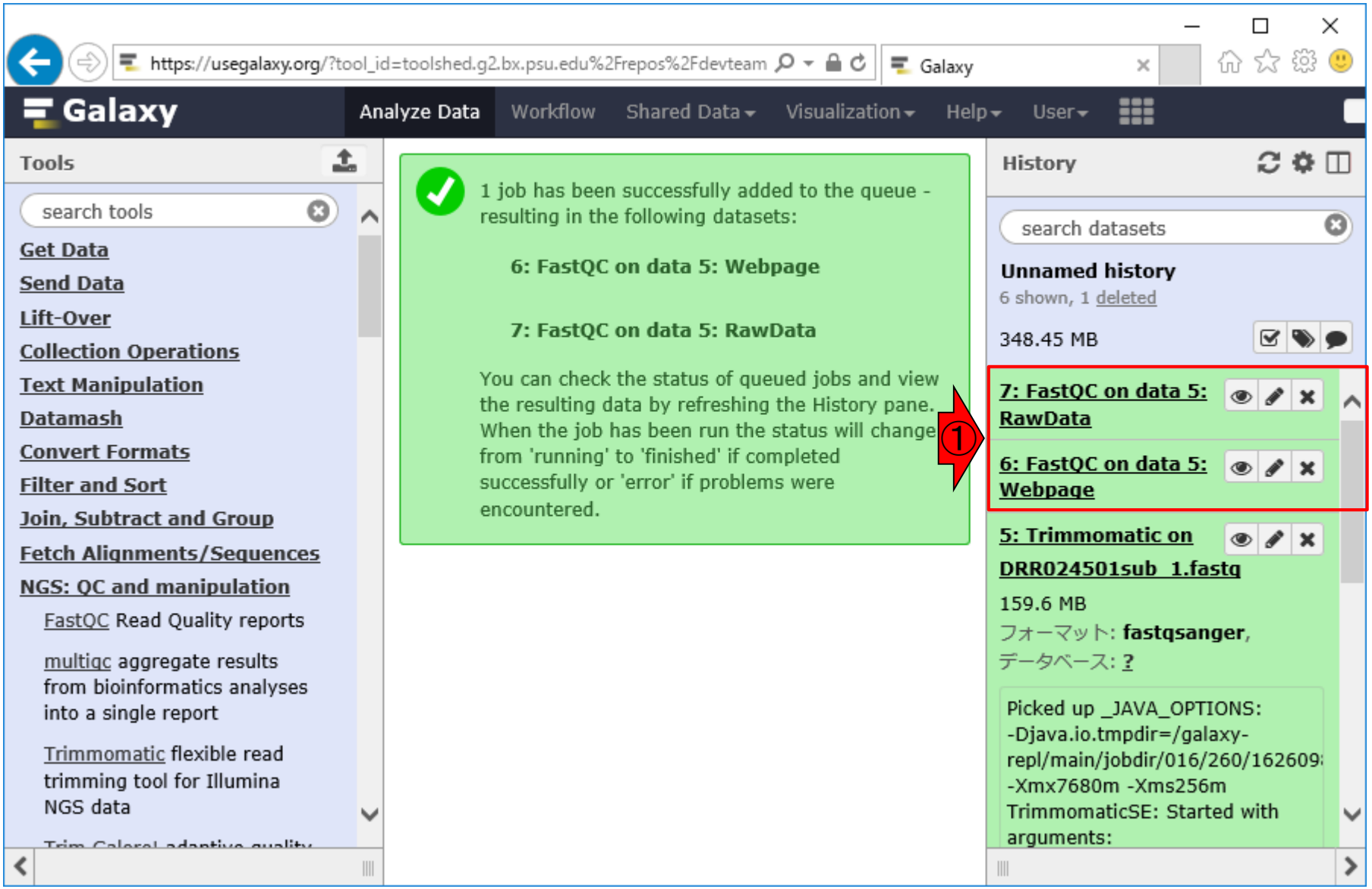
W13-4：実行終了

①緑色になり、FastQCの実行が終了したっぽいので、②ページ上部に移動



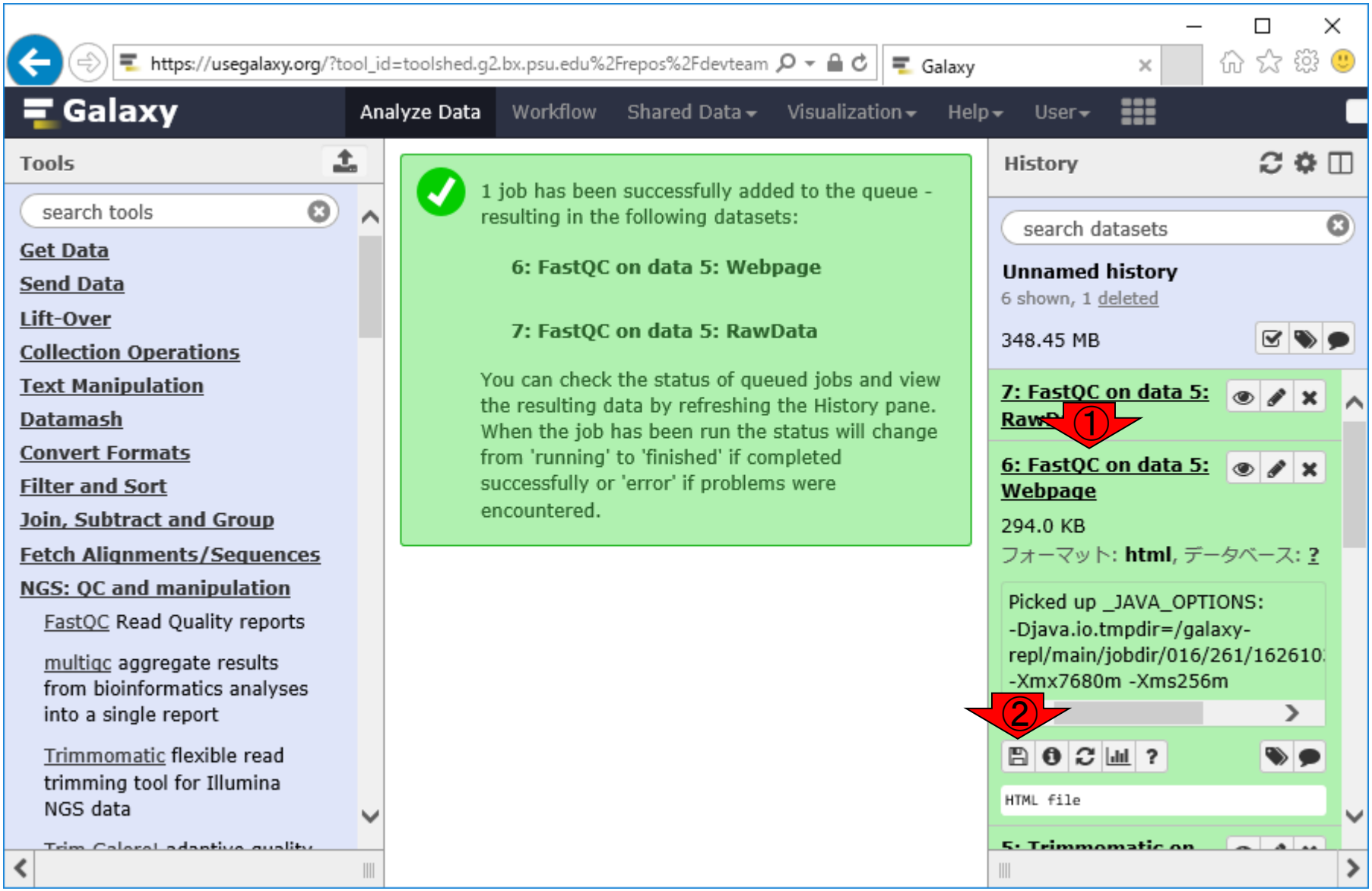
W13-4：実行終了

①こんな感じで無事終了していますね



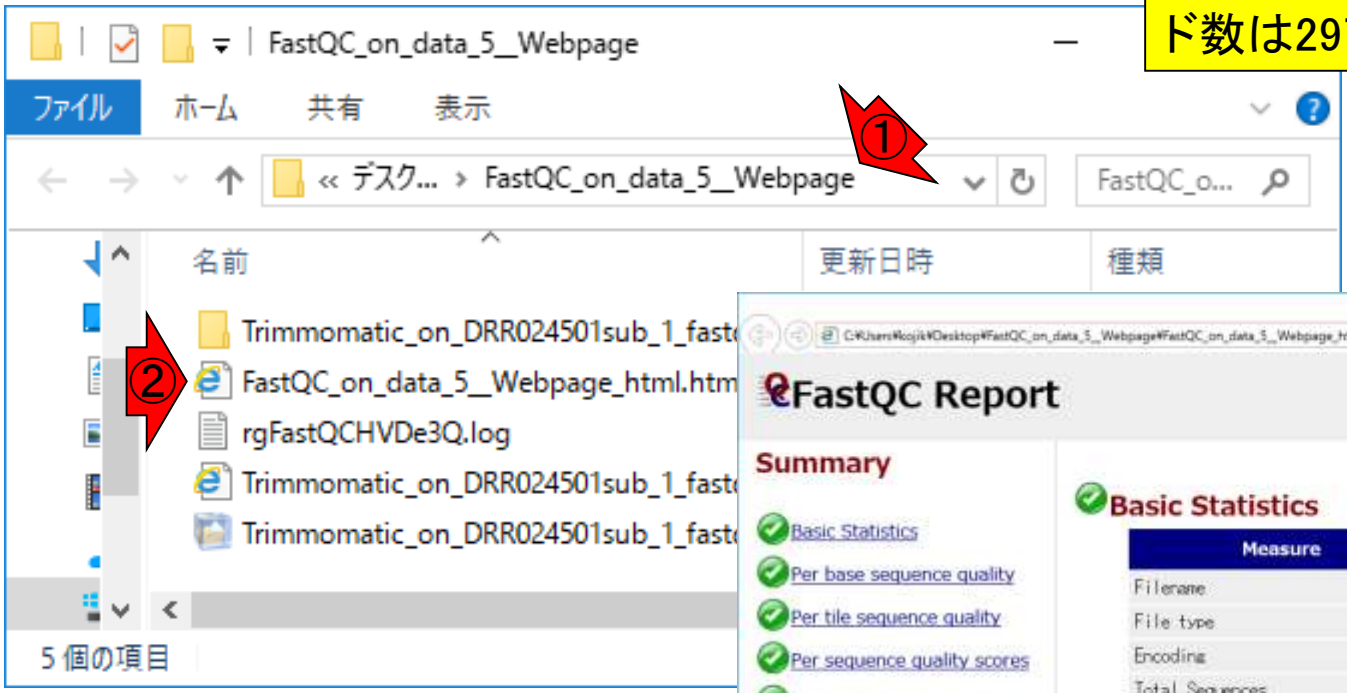
W13-5：結果のダウンロード

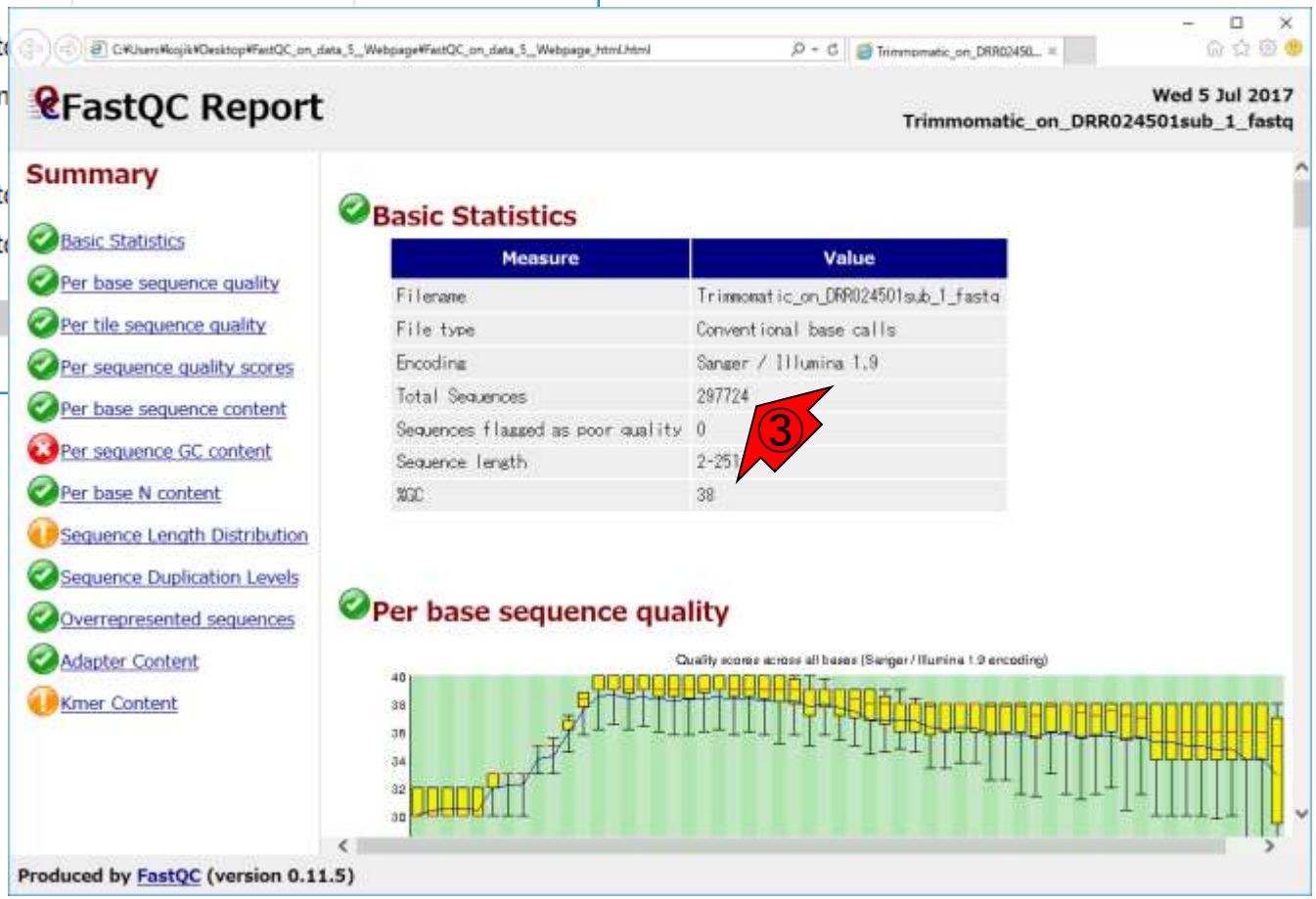
①をクリックして、②zip
ファイルをダウンロード



W13-6：解凍すると…

zipファイルを解凍すると、①のような感じになる。②をダブルクリックするとTrimmomatic実行後のFastQC結果が見られます。③リード数は297,724となったようです





W13-7：Per base sequence.

①Per base sequence qualityをクリックして、クオリティスコア分布を眺めている。綺麗ですね

file:///C:/Users/kojik/Desktop/FastQC_on_data_5_Webpage/FastQC_on_data_5_Webpage_html.html#M1

Trimmomatic_on_DRR02450...

FastQC Report

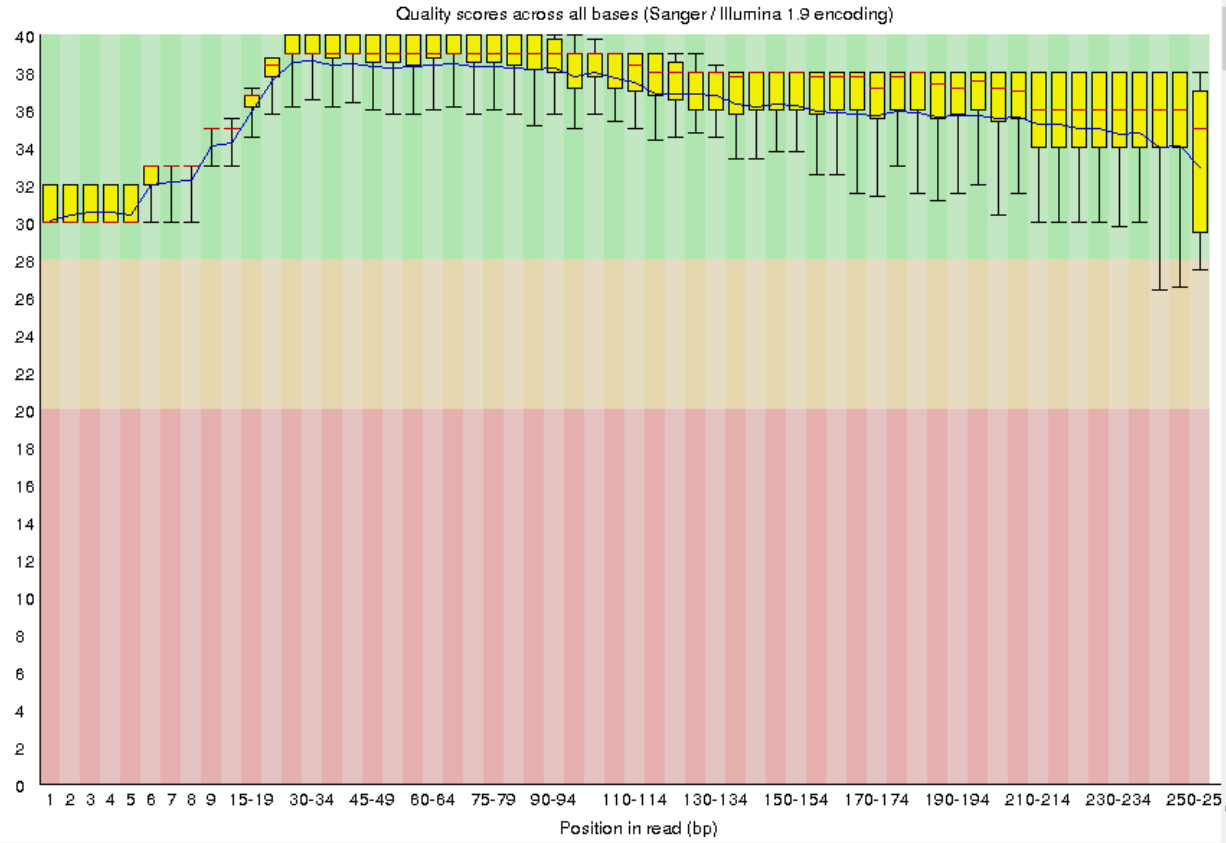
Wed 5 Jul 2017

Trimmomatic_on_DRR024501sub_1_fastq

Summary

- Basic Statistics
- Per base sequence quality ①
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Per base sequence quality



Produced by FastQC (version 0.11.5)

W13-8：Adapter Content

①Adapter Contentの結果として、W11-1で見えていた②Illumina Universal Adapterがなくなっていることがわかる。FaQCs実行後よりもキレイになくなっていることがわかる（第6回W6-2）

file:///C:/Users/kojik/Desktop/FastQC_on_data_5_Webpage/FastQC_on_data_5_Webpage_html.html#M10

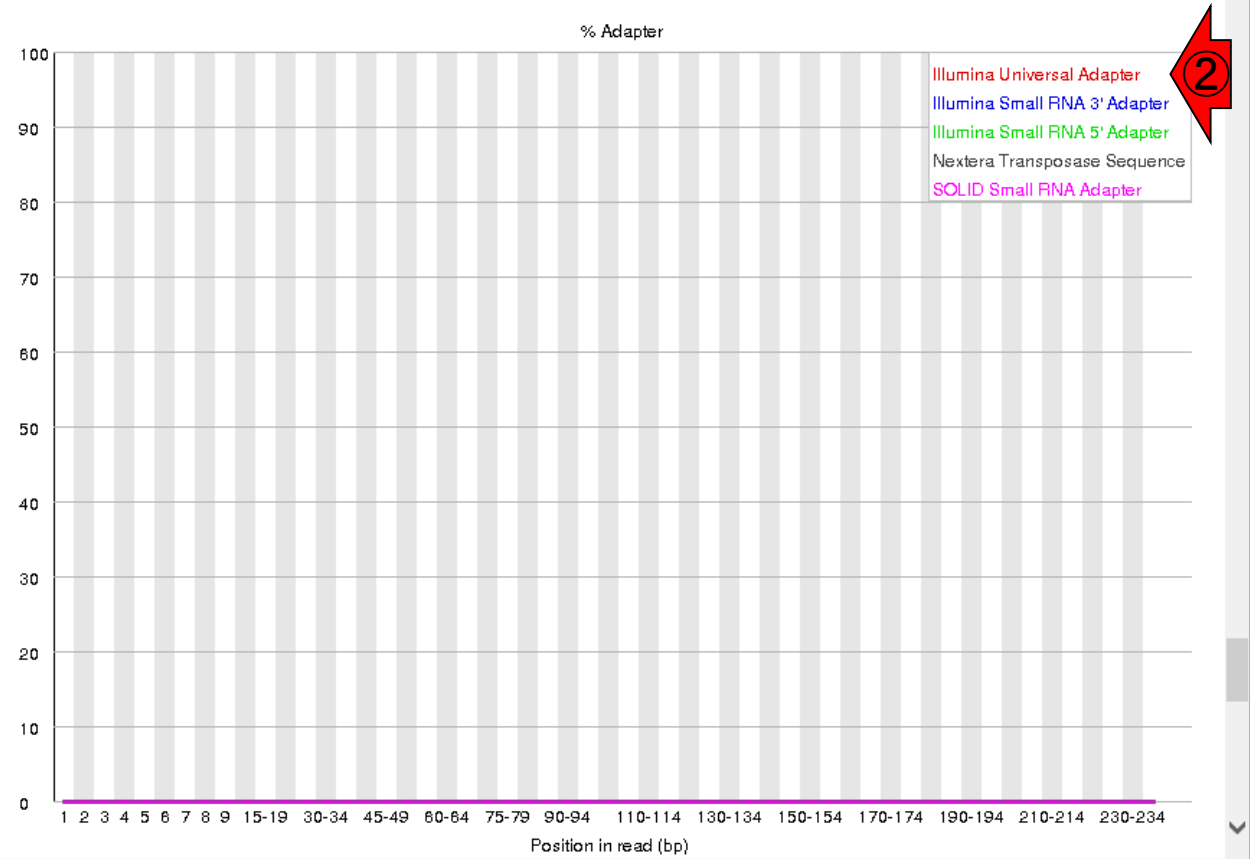
FastQC Report

Trimmomatic_on_DRR024501sub_1_fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content ①
- Kmer Content

Adapter Content



Produced by FastQC (version 0.11.5)

W14-1：NGSハンズオン講習会

①JST-NBDCが主導する②NGSハンズオン講習会の内容は、Linux環境での実習が中心