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）
ウェブ資料

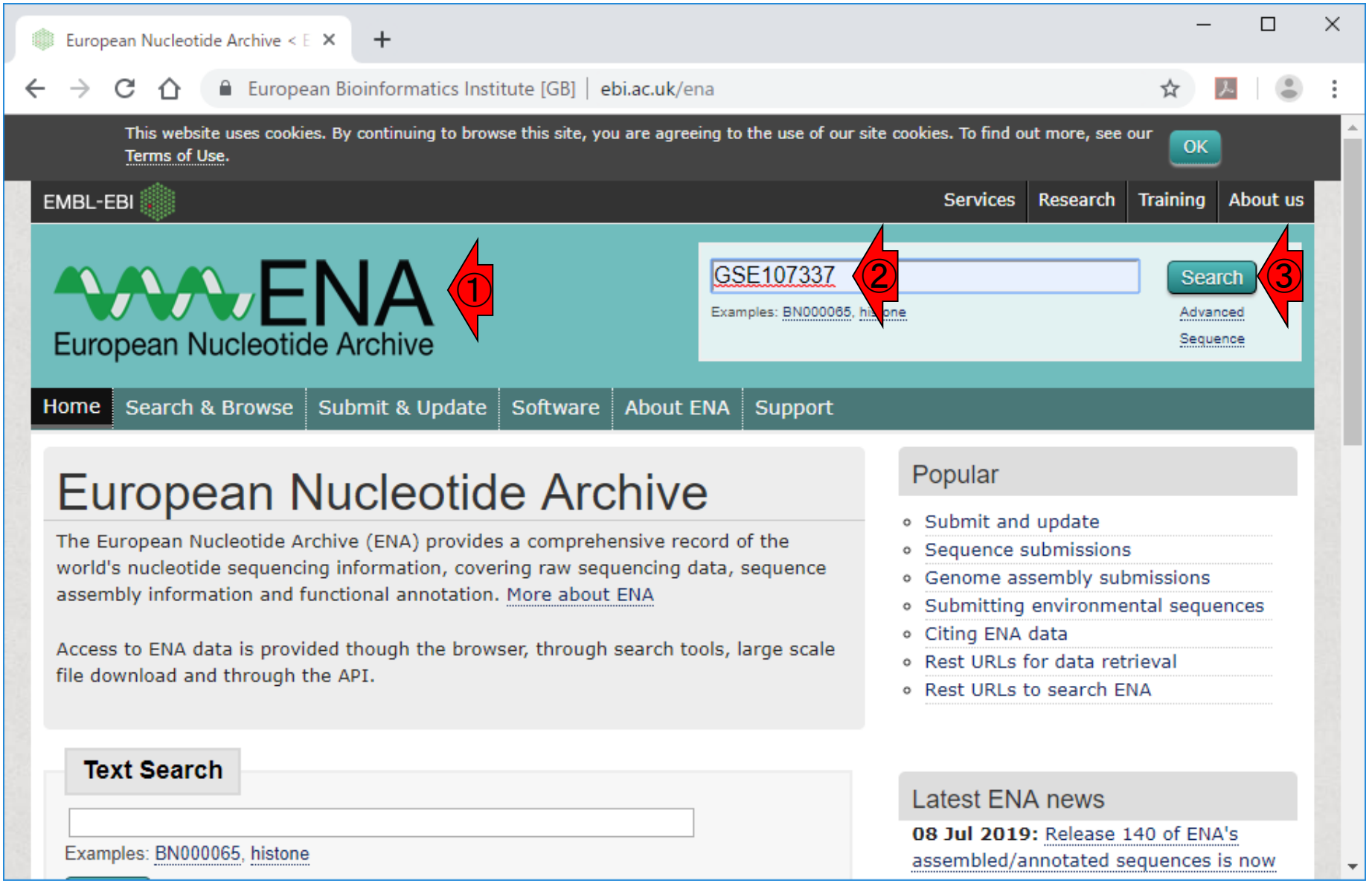
寺田朋子、清水謙多郎、門田 幸二*

Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

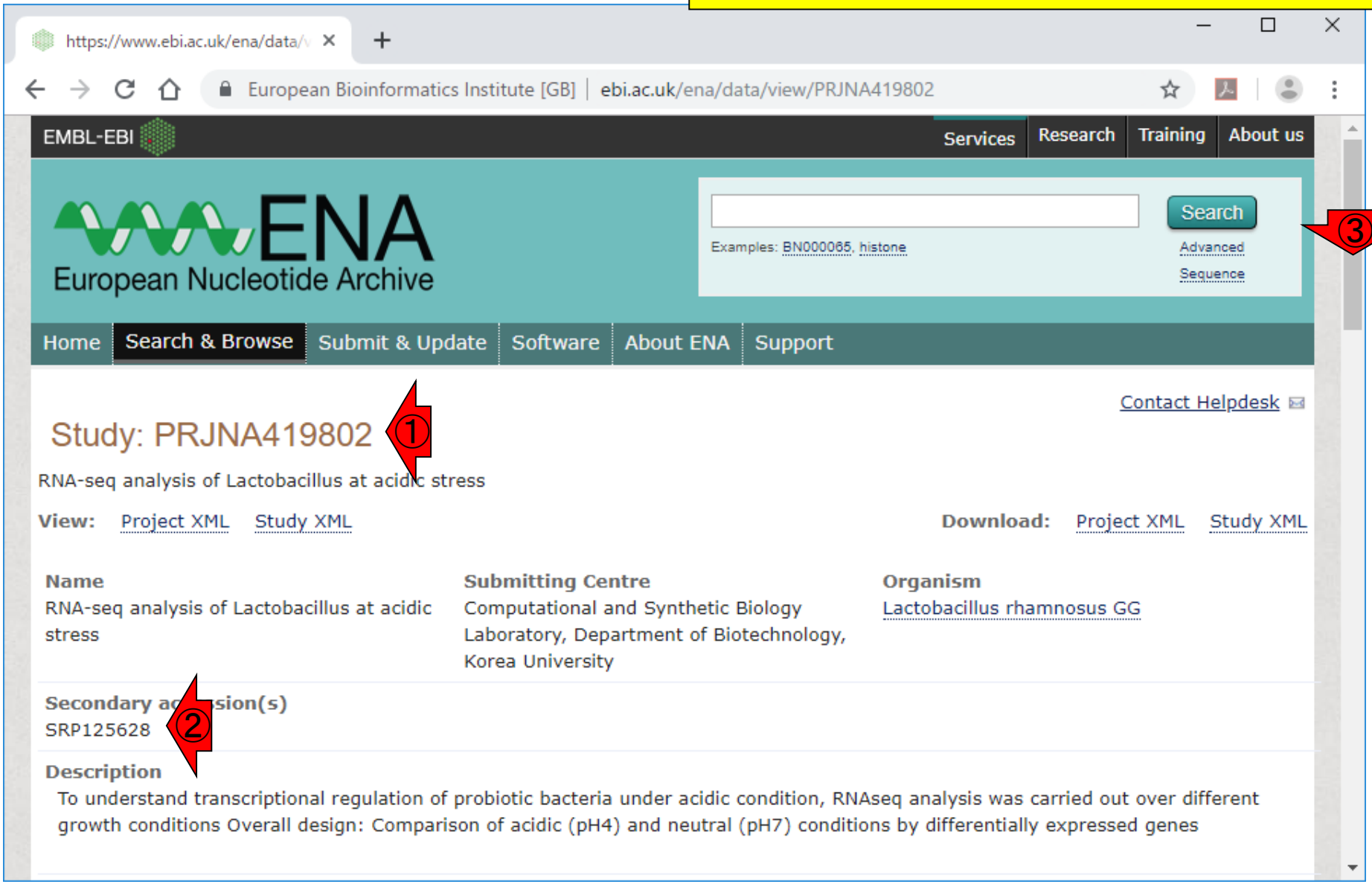
W1：ENA

①ENA上で、②GSE107337を、③検索。第13回のW18と同じです。



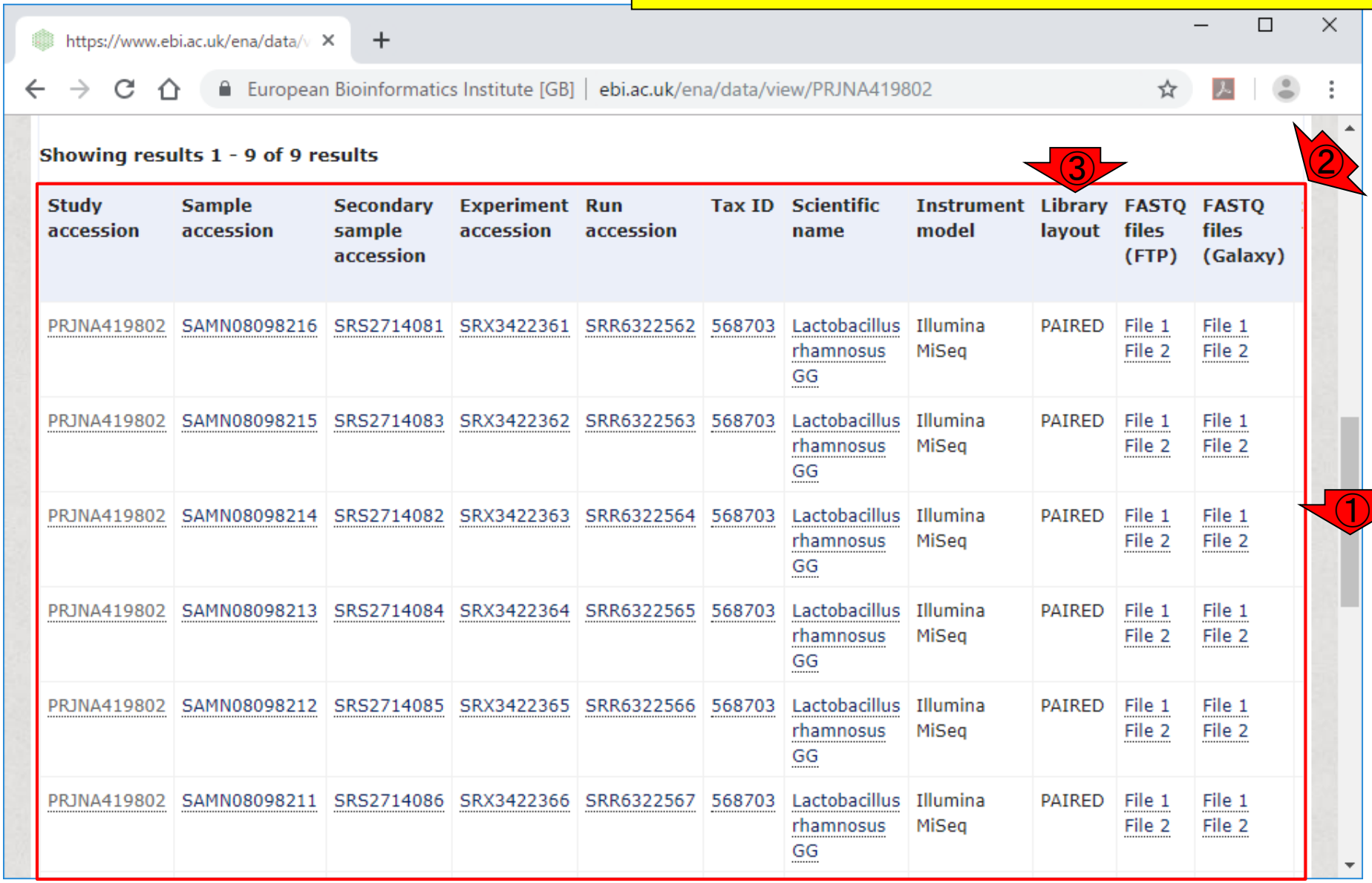
W1：ENA

GSE107337で検索しているはずですが、検索結果は①PRJNS419802や②SRP125628となっていることがわかります。結論としては問題ない。③ページ下部に移動



W1：ENA

①このあたりまで移動すると、②全9サンプルに付随する様々なID情報を一覧できる。原著論文には明記されていないが、③このデータがpaired-endであることもわかる。



Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)
PRJNA419802	SAMN08098216	SRS2714081	SRX3422361	SRR6322562	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098215	SRS2714083	SRX3422362	SRR6322563	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098214	SRS2714082	SRX3422363	SRR6322564	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098213	SRS2714084	SRX3422364	SRR6322565	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098212	SRS2714085	SRX3422365	SRR6322566	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098211	SRS2714086	SRX3422366	SRR6322567	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2

W1：ENA

①今回はここを利用して、ENAから直接GalaxyにFASTQファイルをアップロードする。ウェブ資料では、冗長な説明を回避すべく、②SRR6322564、③SRR6322567、およびここでは見えていないがSRR6322569の3サンプルに限定して行う。

Showing results 1 - 9 of 9 results

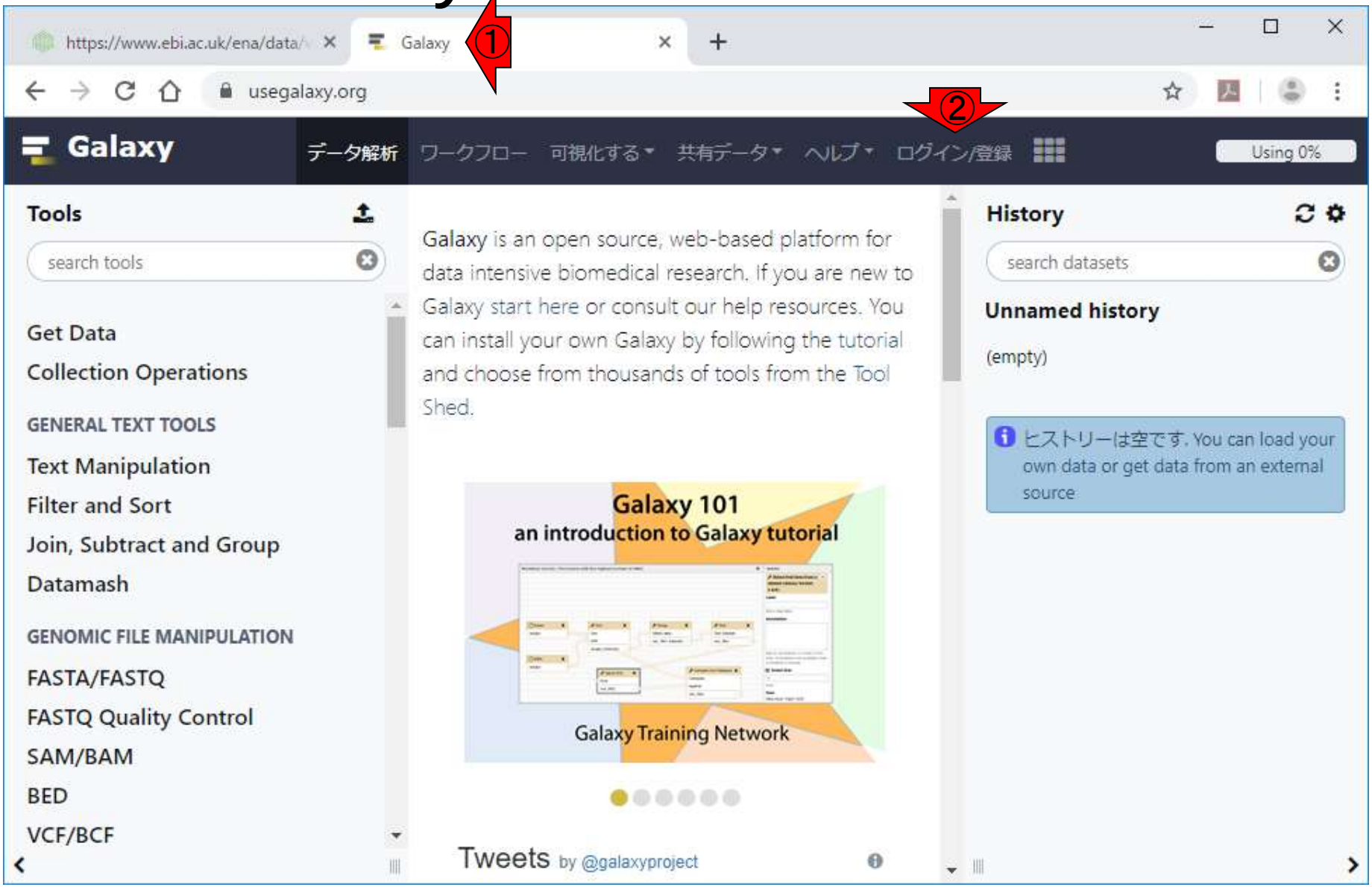
Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)
PRJNA419802	SAMN08098216	SRS2714081	SRX3422361	SRR6322562	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098215	SRS2714083	SRX3422362	SRR6322563	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098214	SRS2714082	SRX3422363	SRR6322564	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098213	SRS2714084	SRX3422364	SRR6322565	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098212	SRS2714085	SRX3422365	SRR6322566	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098211	SRS2714086	SRX3422366	SRR6322567	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2

Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

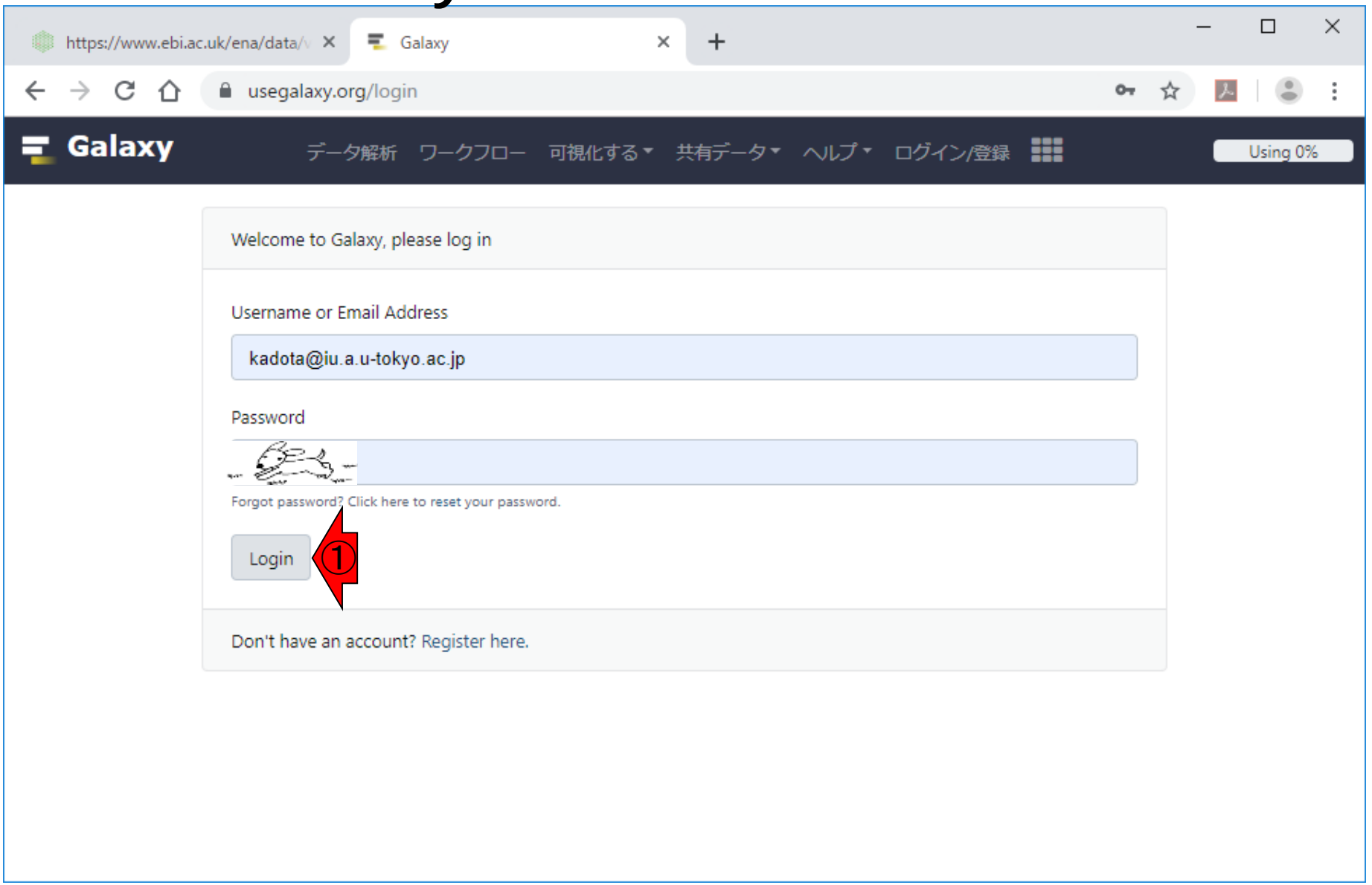
①新規にタブを開いて、Galaxy上で、②ログインしておく。

W2：Galaxy



ユーザ名とパスワードを入力して、①Login。

W2：Galaxy

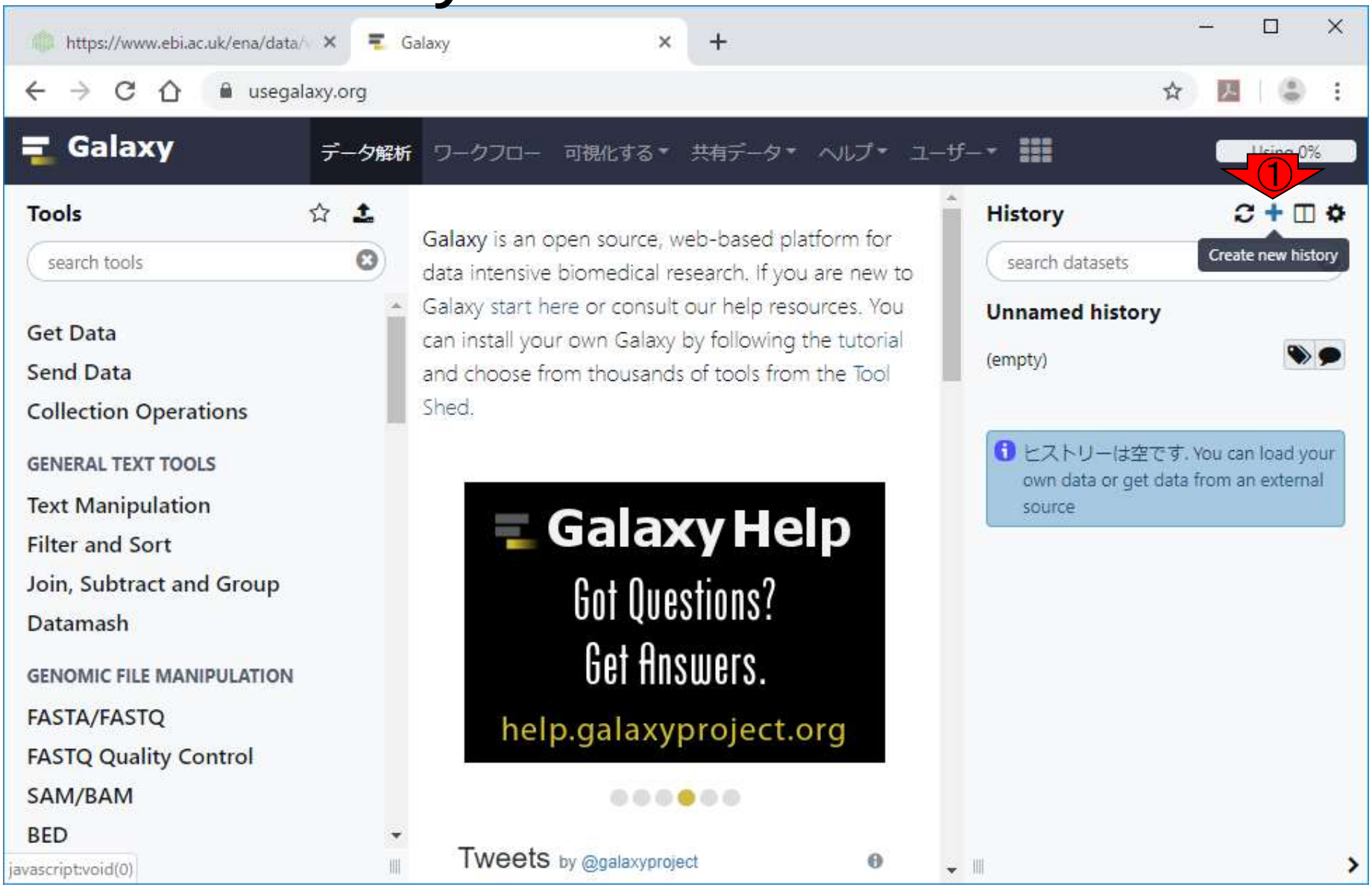


W2：Galaxy

私のログイン直後の状態。①予めこれまでの複数のヒストリーを全部消去してあったのでこんな感じになっているが、ヒトによってはここに直前に行ったヒストリーが見えていると思います。例えば、第12回のW11-2。

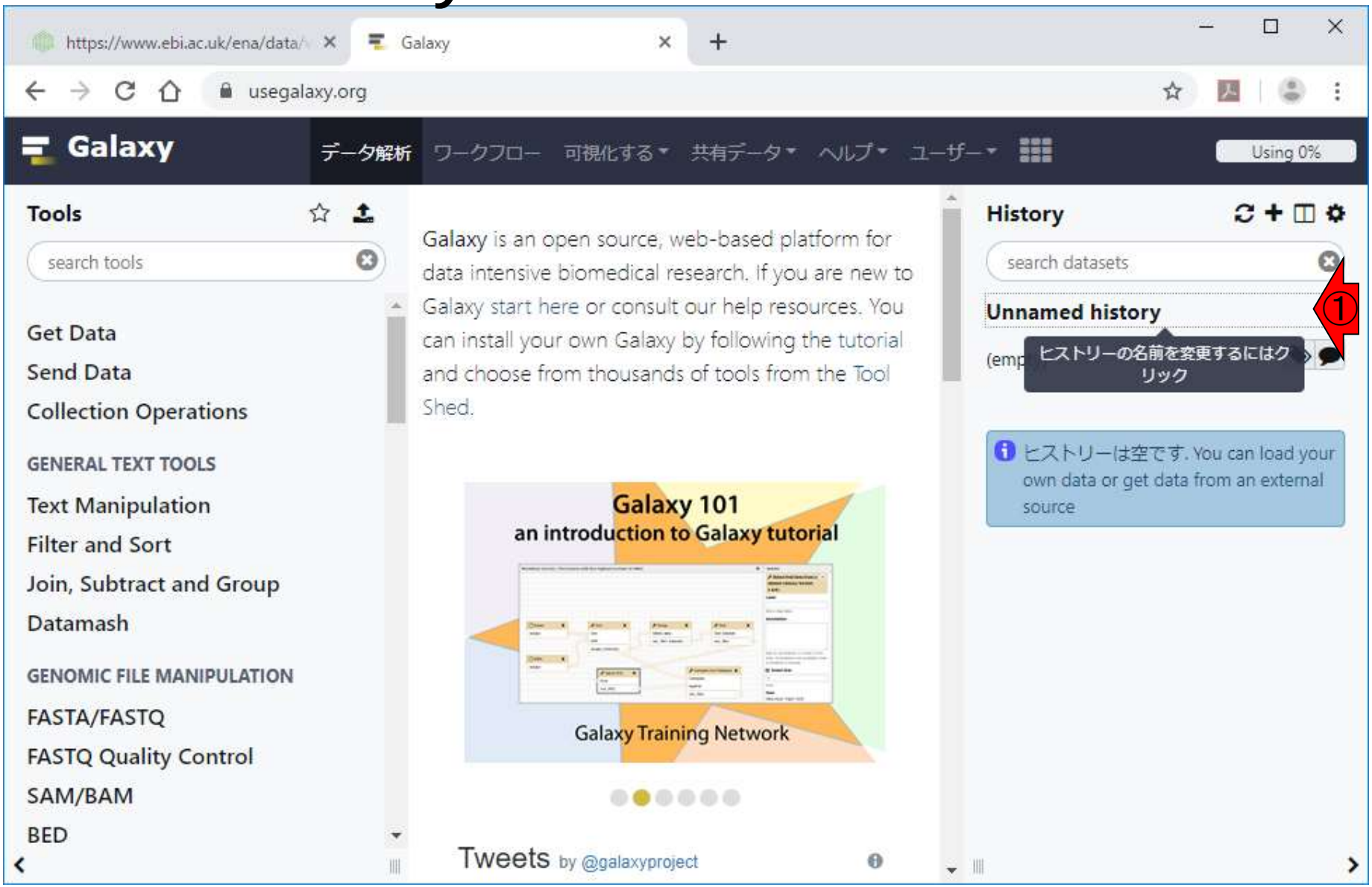
条件を揃えるべく、①Create new history。

W2：Galaxy



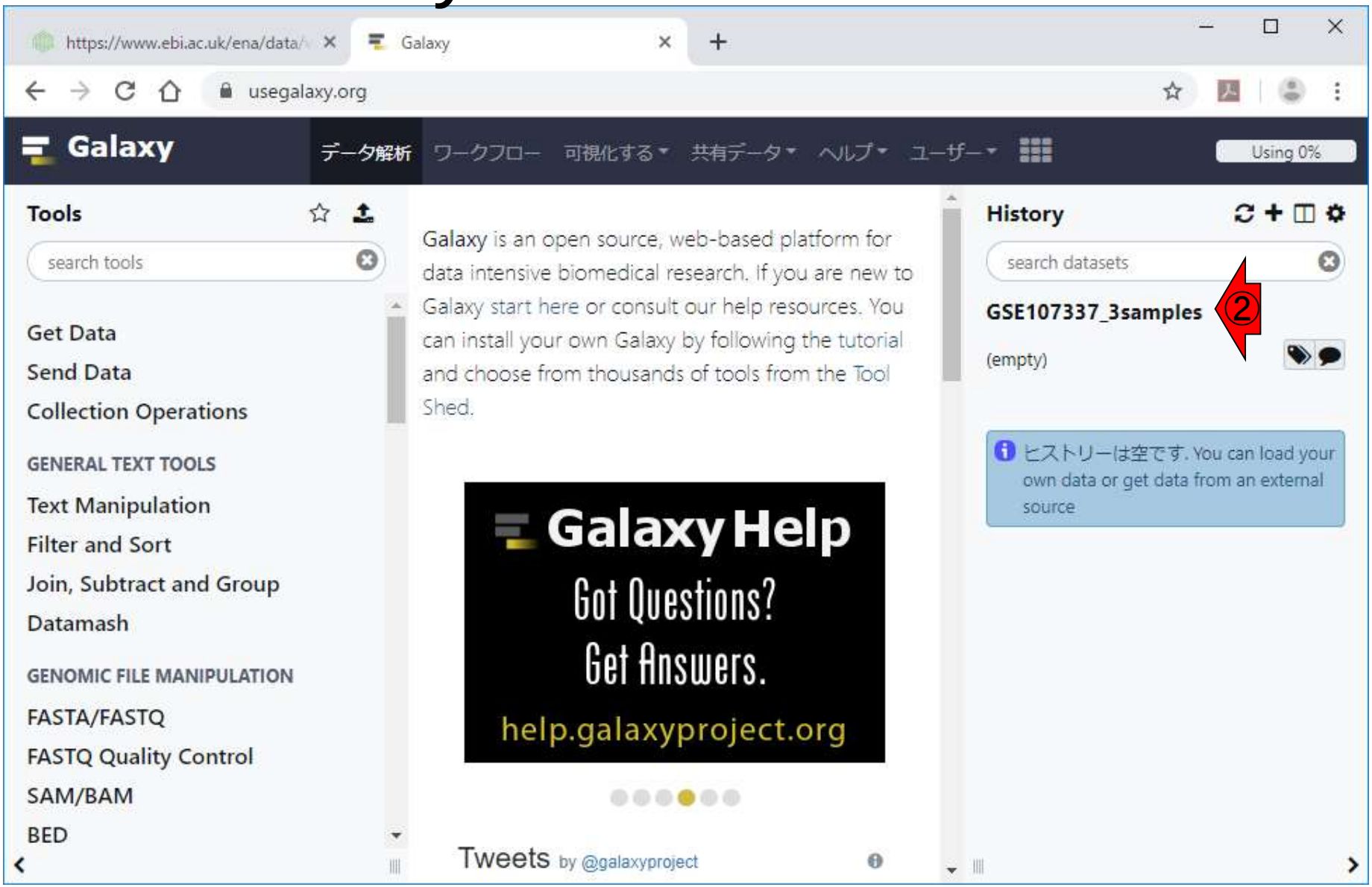
とりあえず①ヒストリー名をGSE107337_3samplesと変更。

W2：Galaxy



W2：Galaxy

とりあえず①ヒストリー名をGSE107337_3samplesと変更。
②変更完了。第12回のW5-2のおさらいでした。

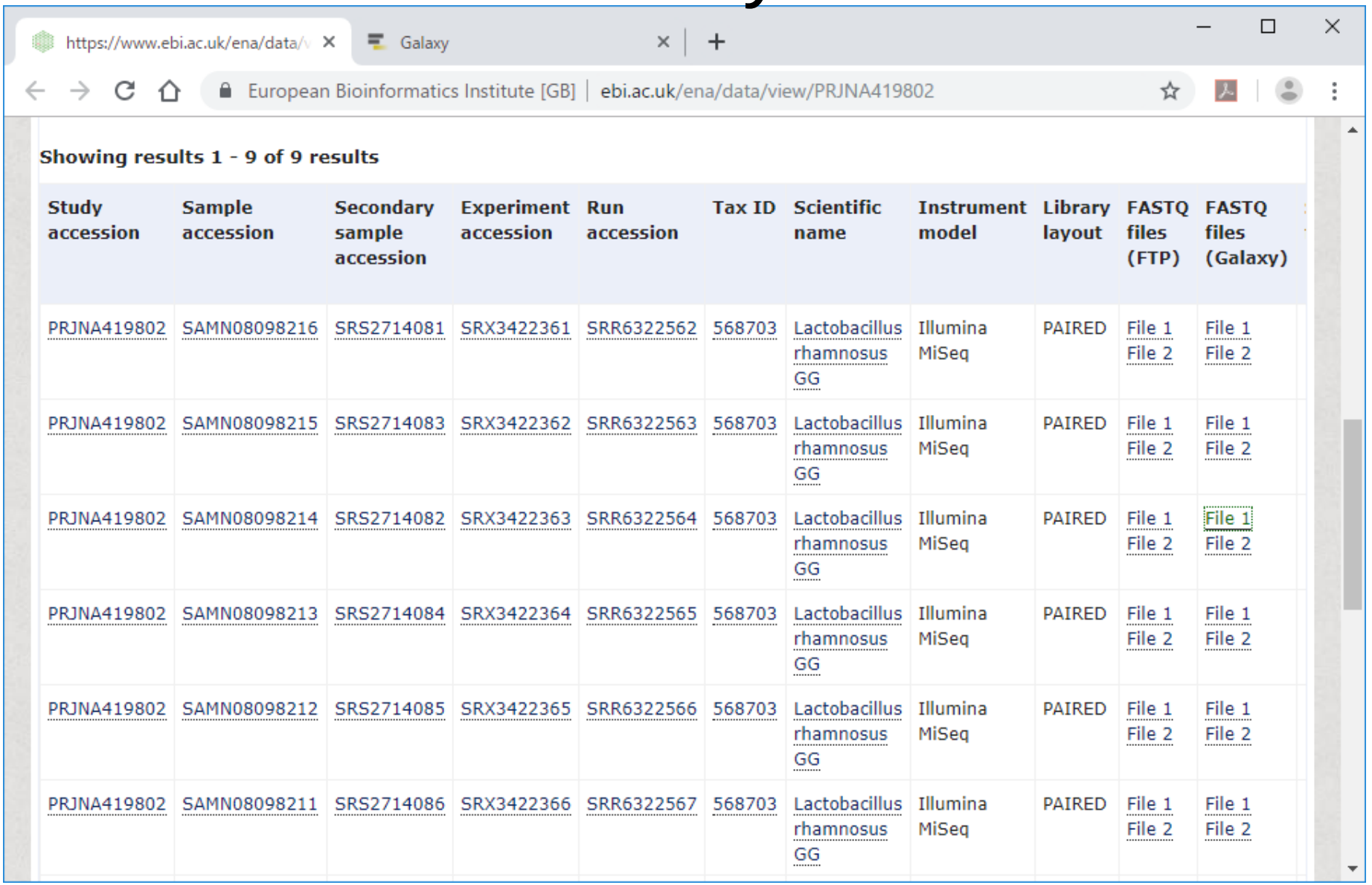


Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

ブラウザのタブをENAのほうに変更した状態。

W3-1：ENA→Galaxy



Showing results 1 - 9 of 9 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)
PRJNA419802	SAMN08098216	SRS2714081	SRX3422361	SRR6322562	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098215	SRS2714083	SRX3422362	SRR6322563	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098214	SRS2714082	SRX3422363	SRR6322564	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098213	SRS2714084	SRX3422364	SRR6322565	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098212	SRS2714085	SRX3422365	SRR6322566	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098211	SRS2714086	SRX3422366	SRR6322567	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2

W3-1：ENA→Ga

とりあえず、1つめの解析対象サンプルである、①SRR6322564の、②File 1（paired-endの1つ目；forward側とする）をGalaxyに送る。クリックするのは②のみ。

https://www.ebi.ac.uk/ena/data/ × Galaxy × +

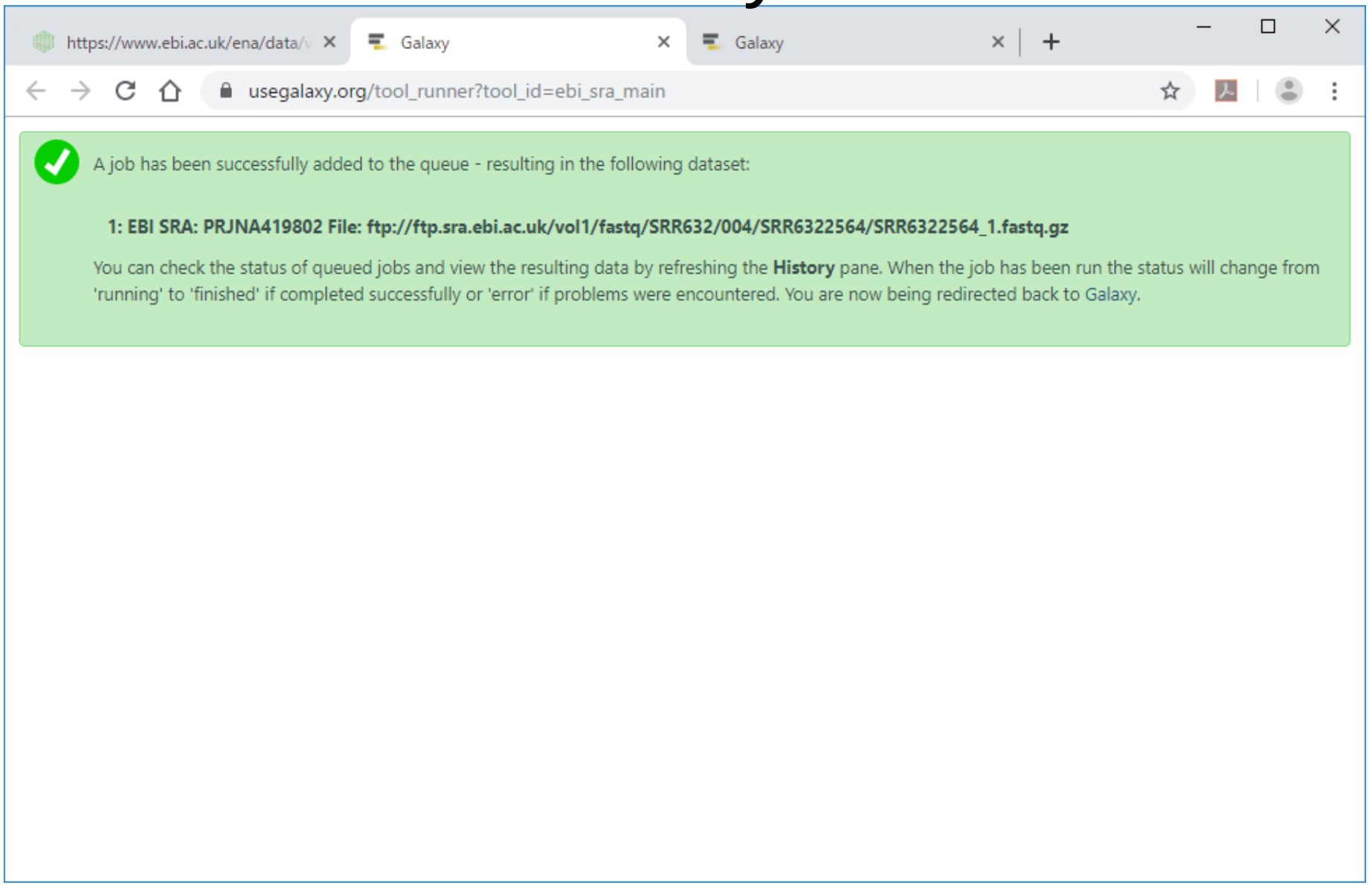
European Bioinformatics Institute [GB] | ebi.ac.uk/ena/data/view/PRJNA419802

Showing results 1 - 9 of 9 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)
PRJNA419802	SAMN08098216	SRS2714081	SRX3422361	SRR6322562	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098215	SRS2714083	SRX3422362	SRR6322563	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098214	SRS2714082	SRX3422363	SRR6322564	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098213	SRS2714084	SRX3422364	SRR6322565	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098212	SRS2714085	SRX3422365	SRR6322566	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098211	SRS2714086	SRX3422366	SRR6322567	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2

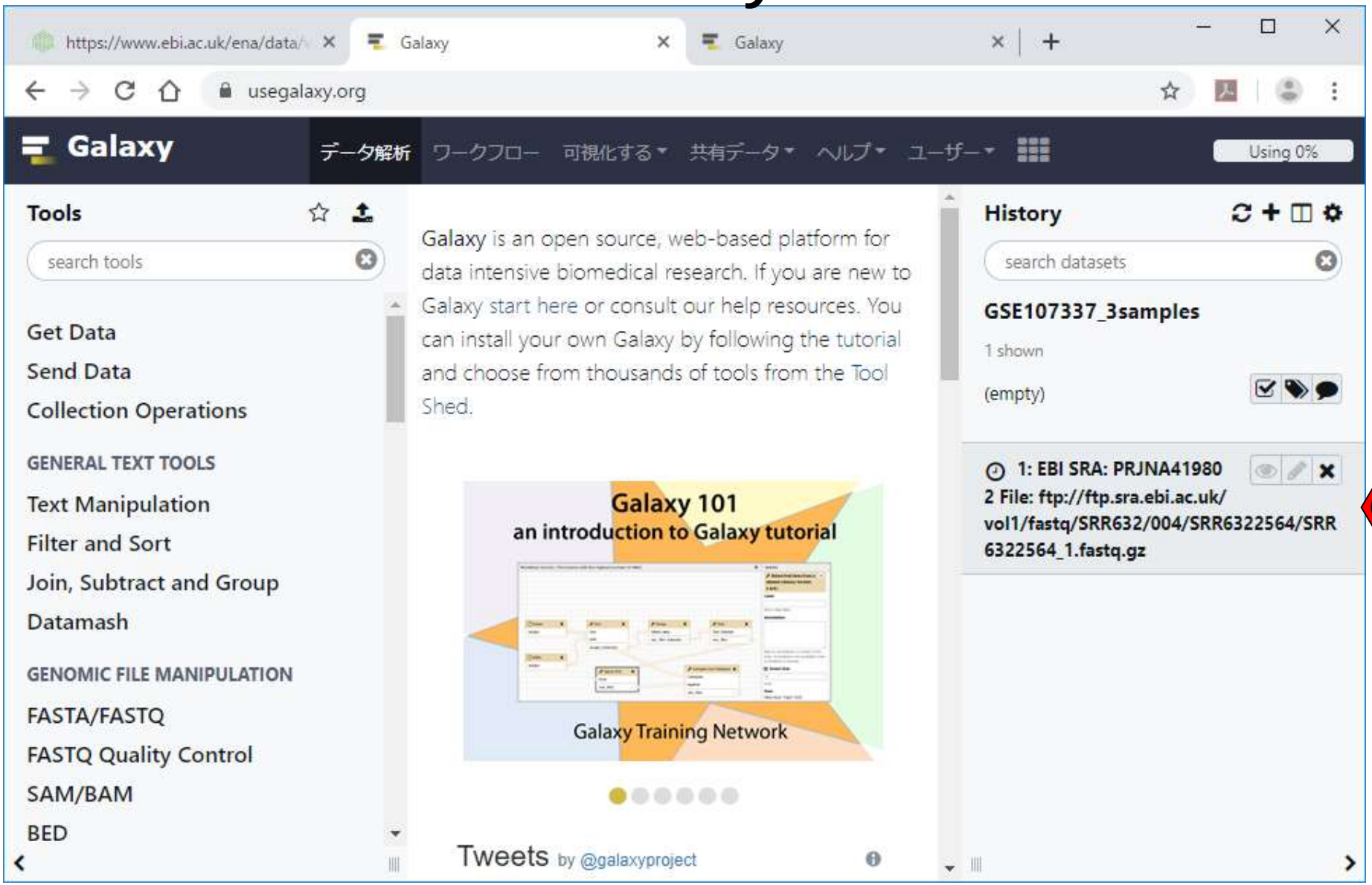
すぐにこんな感じになる。

W3-1：ENA→Galaxy



すぐにこんな感じになる。そしてまたすぐにこんな感じの画面に切り替わり、①灰色の実行待ち状態になる。

W3-1：ENA→Galaxy

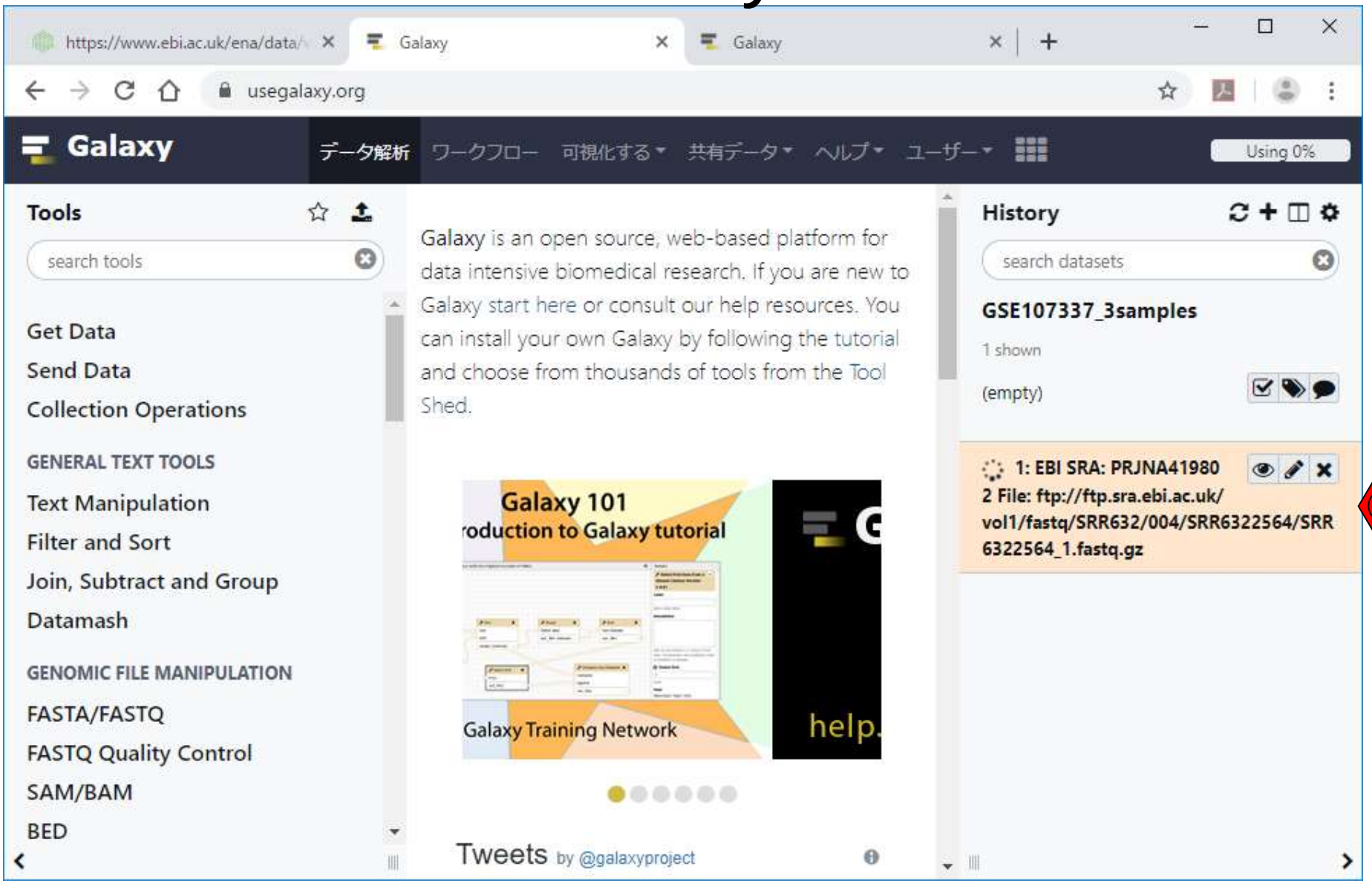


W3-1：ENA→Ga

すぐにこんな感じになる。そしてまたすぐにこんな感じの画面に切り替わり、①灰色の実行待ち状態になる。②さきほど作成したGSE107337_3samplesというヒストリー上で行われています。

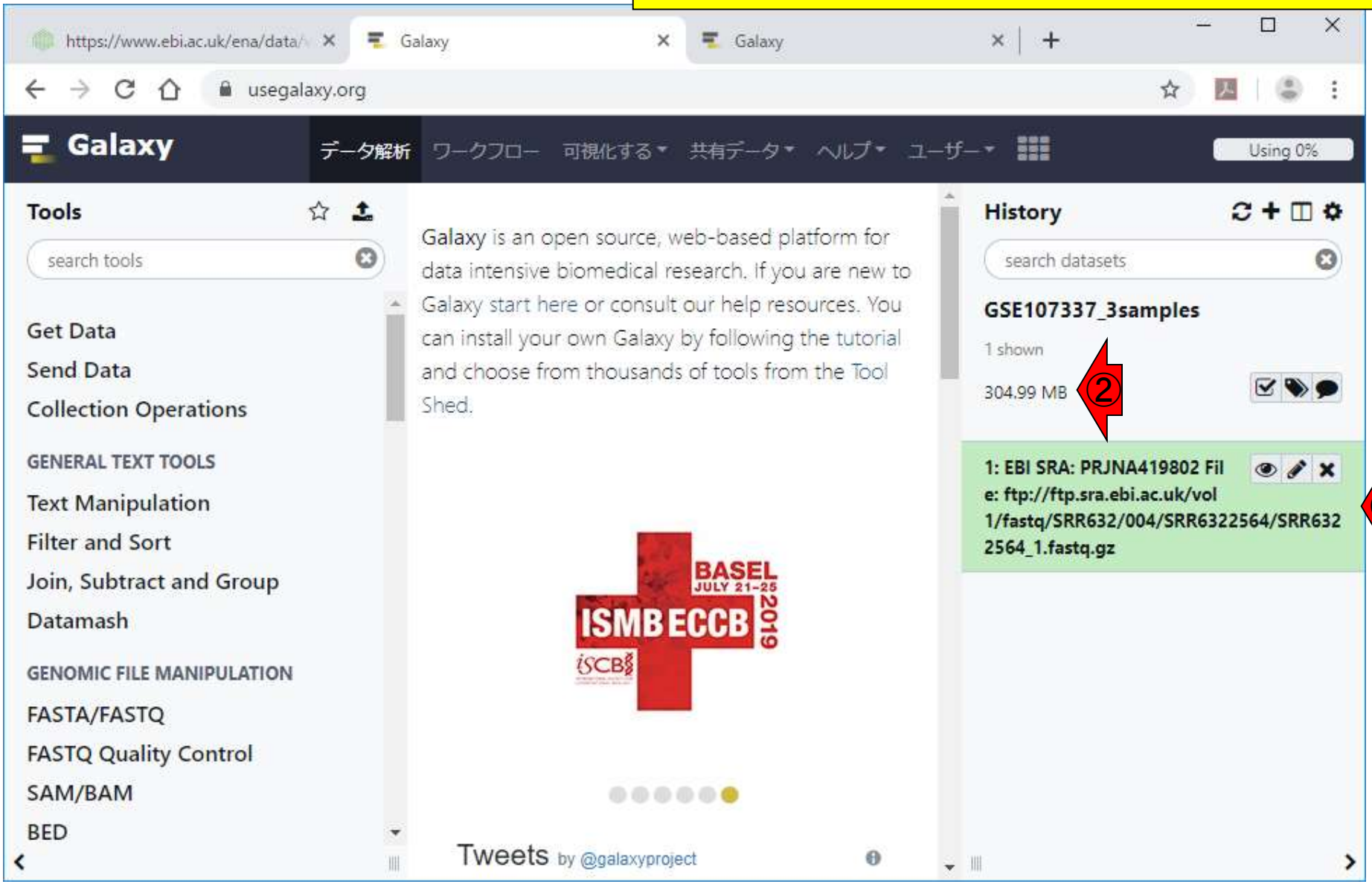
①取り込み作業中…。

W3-1：ENA→Galaxy



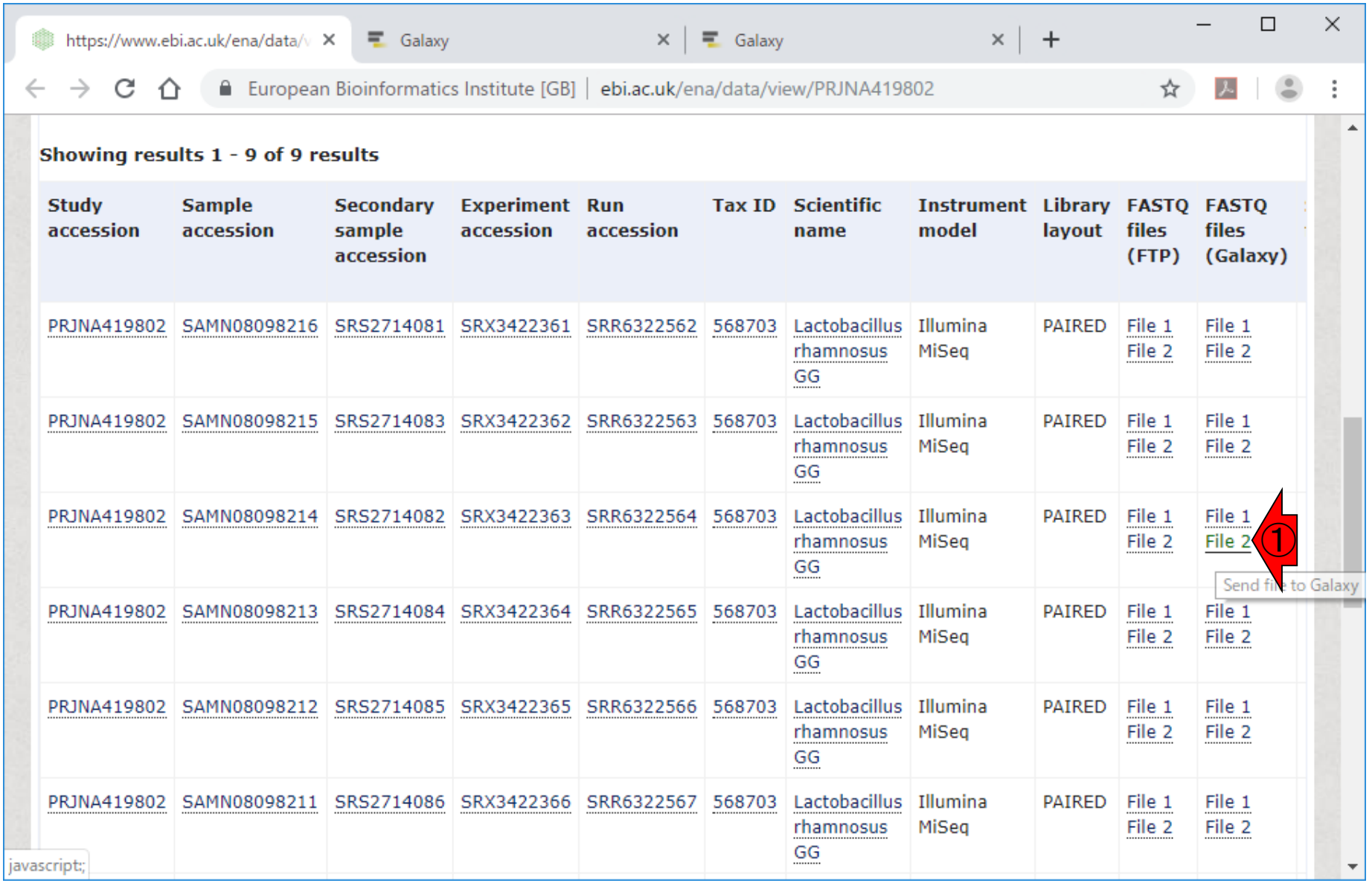
W3-1：ENA→Ga

①取り込み完了。②サイズがemptyから304.99 MBに増加しました。第13回原稿の表1の結果（319.8 MB）と似ており、妥当ですね。



①次はSRR6322564のFile 2（paired-endのreverse側）。

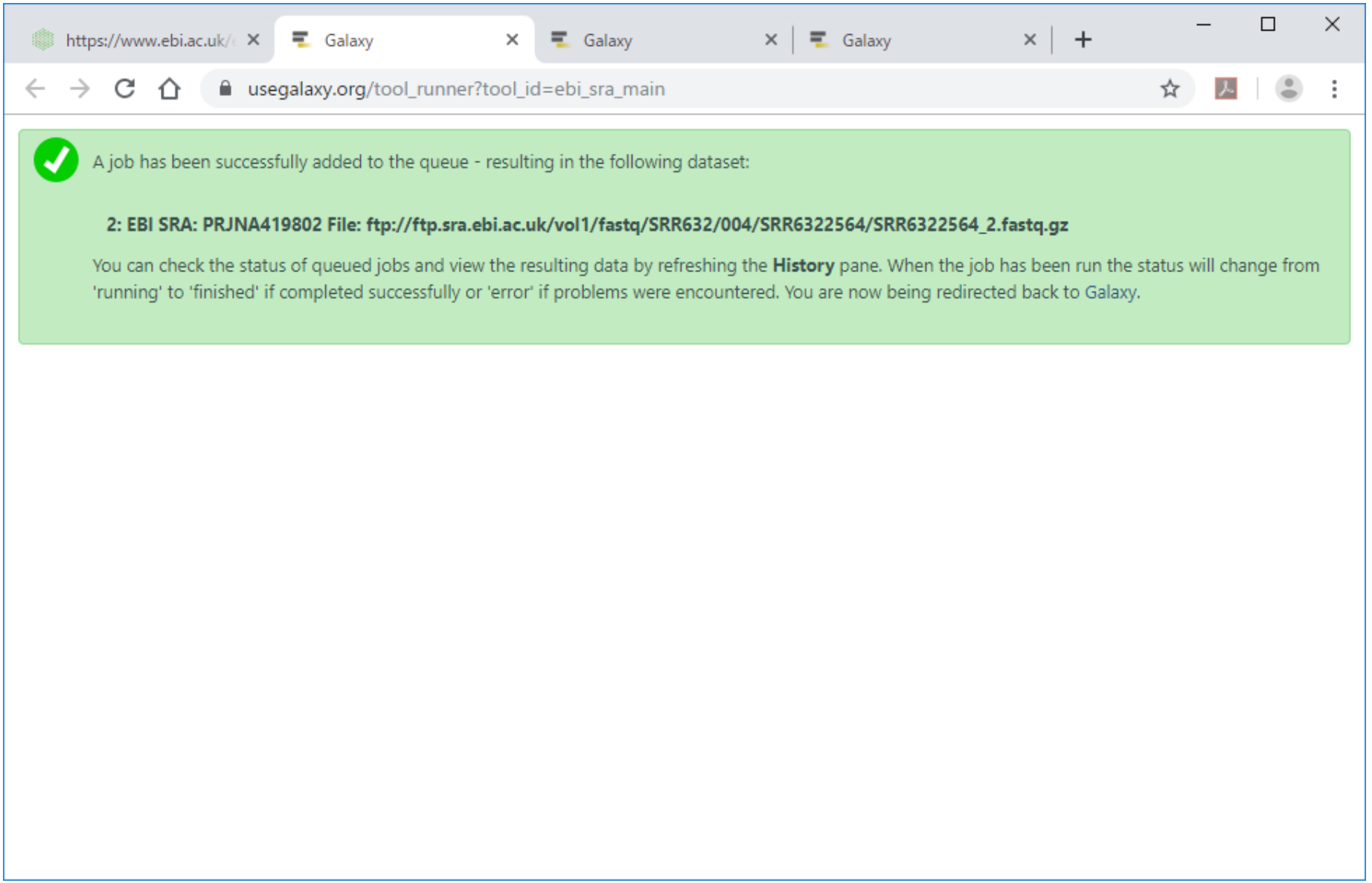
W3-2：SRR6322564のFile2



Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)
PRJNA419802	SAMN08098216	SRS2714081	SRX3422361	SRR6322562	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098215	SRS2714083	SRX3422362	SRR6322563	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098214	SRS2714082	SRX3422363	SRR6322564	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098213	SRS2714084	SRX3422364	SRR6322565	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098212	SRS2714085	SRX3422365	SRR6322566	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098211	SRS2714086	SRX3422366	SRR6322567	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2

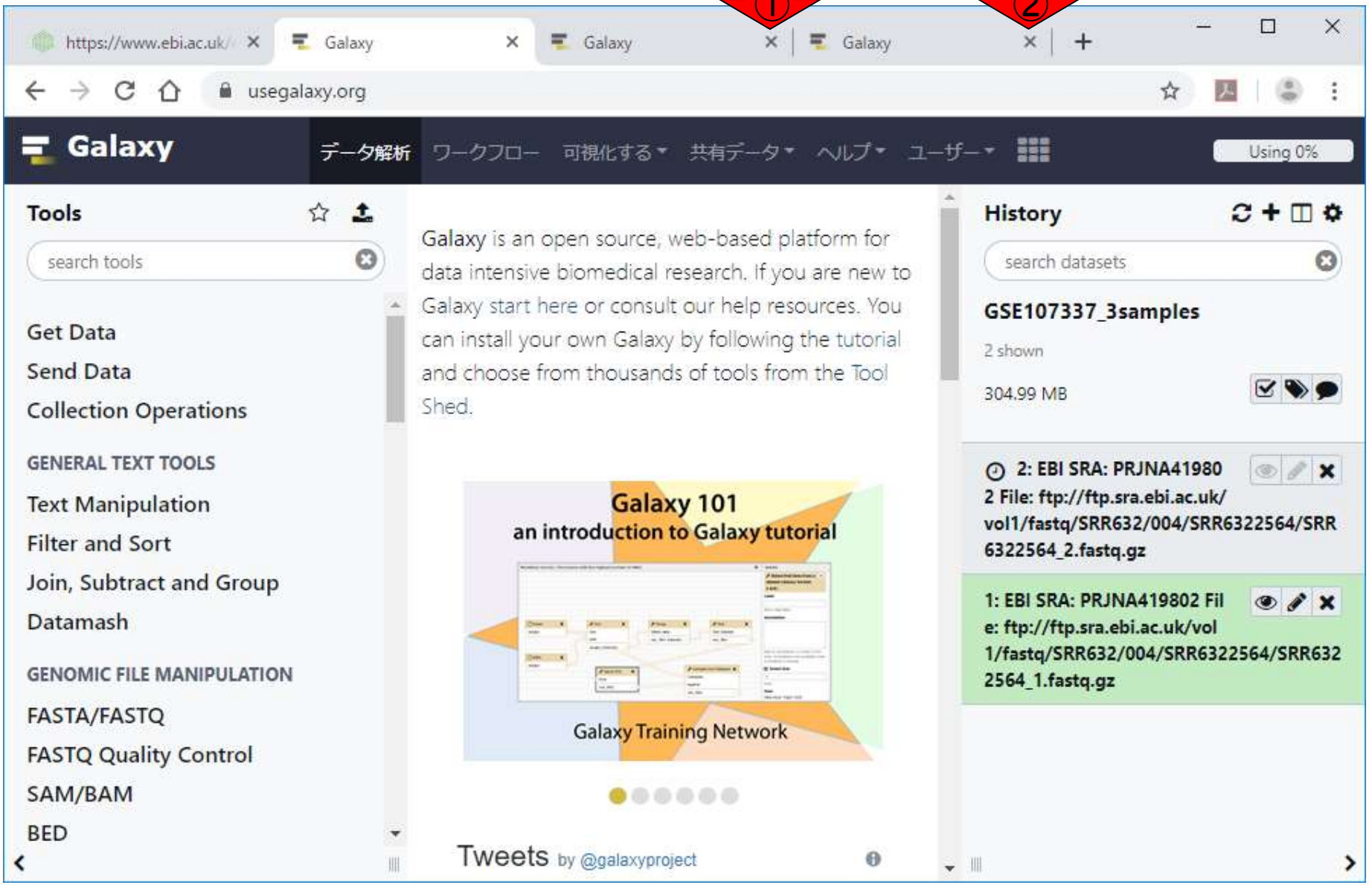
さきほどと同様に、すぐにこんな感じになる。ここまでで、どんどんタブが追加されていくのだと学習する。

W3-2：SRR6322564のFile2



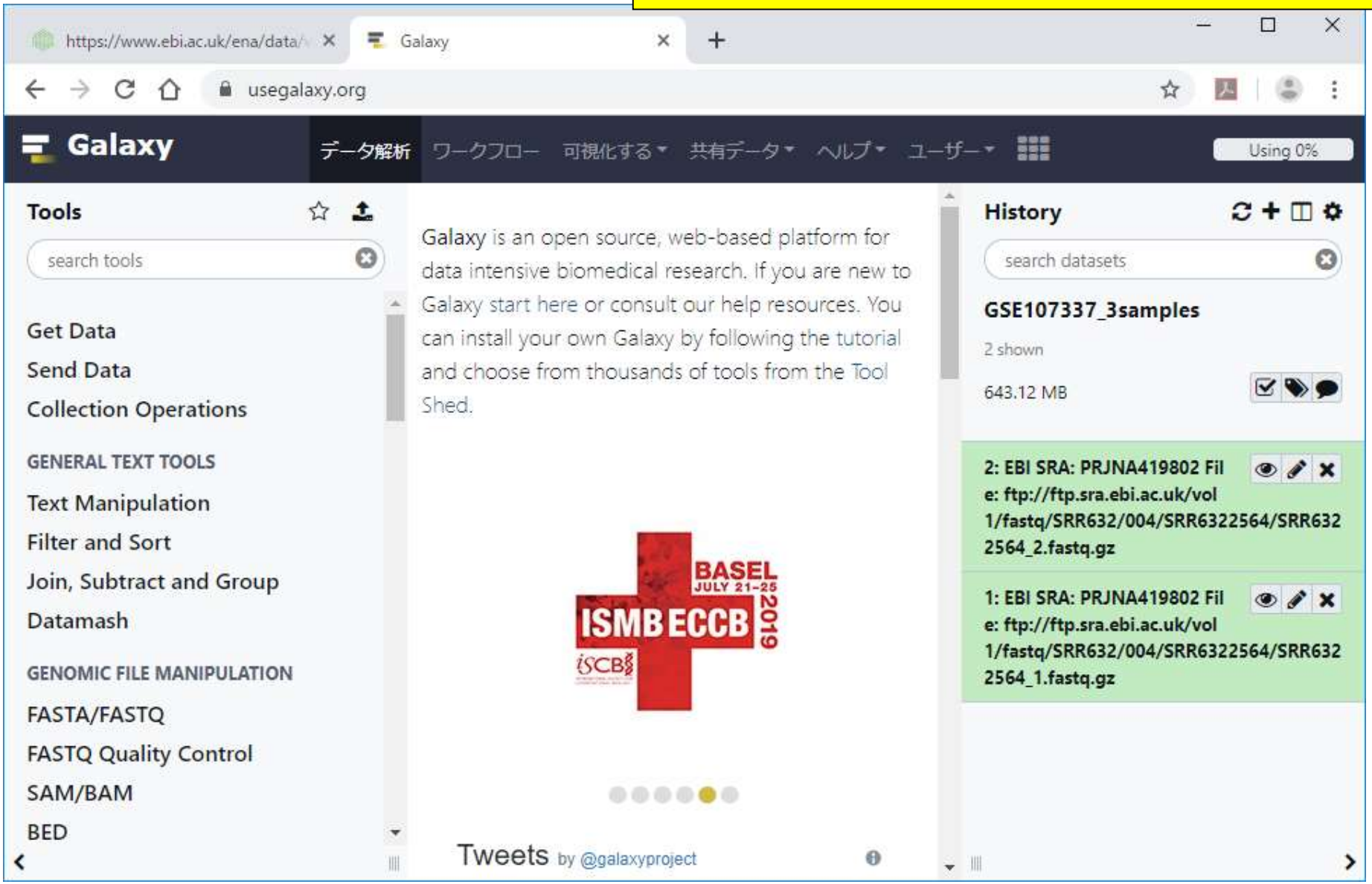
W3-2：SRR6322564のFile2

ほどなくして、こんな感じになる。実行待ち状態。①や②のタブは不要なので消してよい。



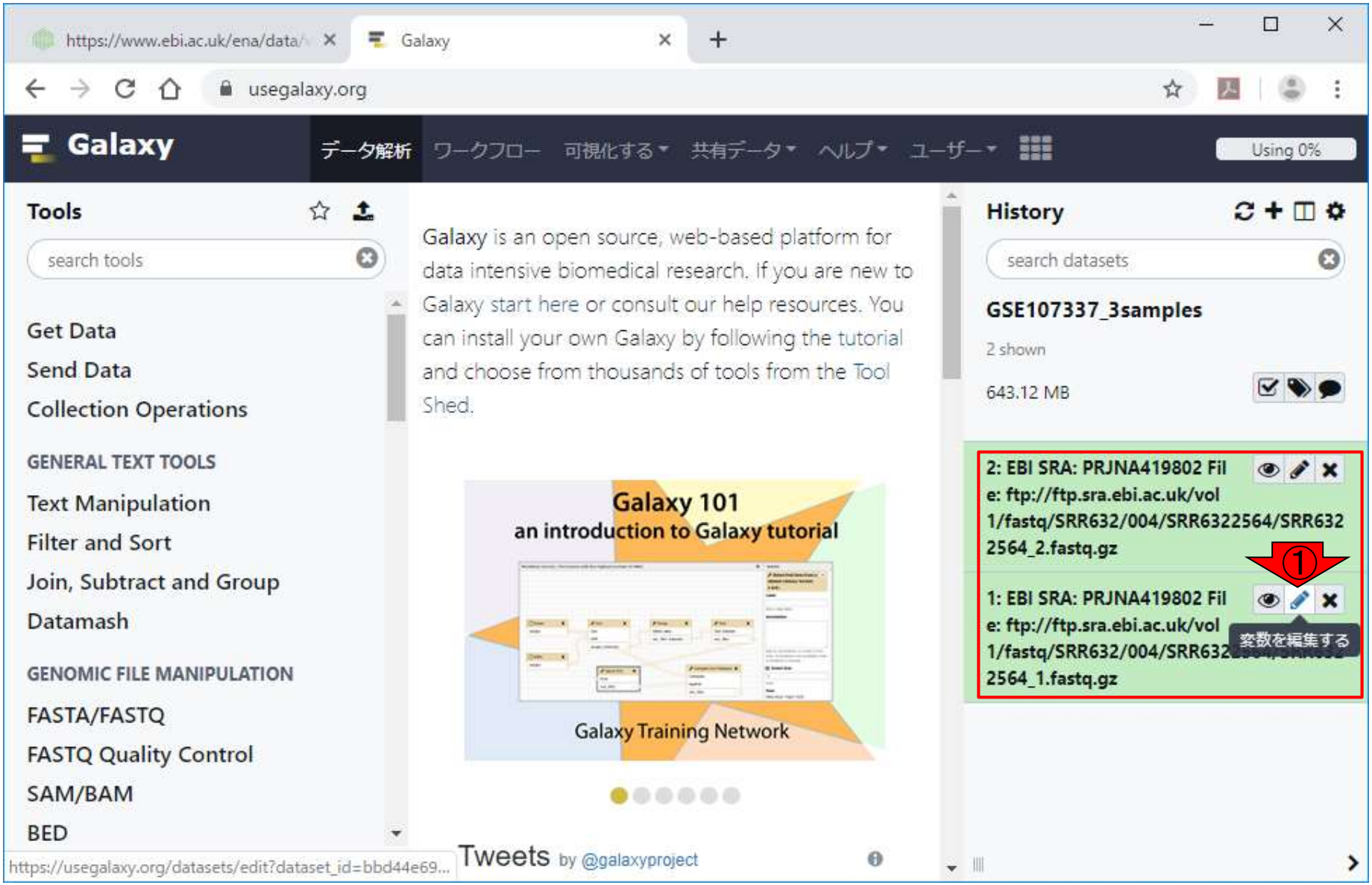
W3-2：SRR6322

取り込み（インポート）完了を意味する緑色になった。約3分。ここまでで、SRR6322564のpaired-endの2つのファイルが解析可能状態となった。



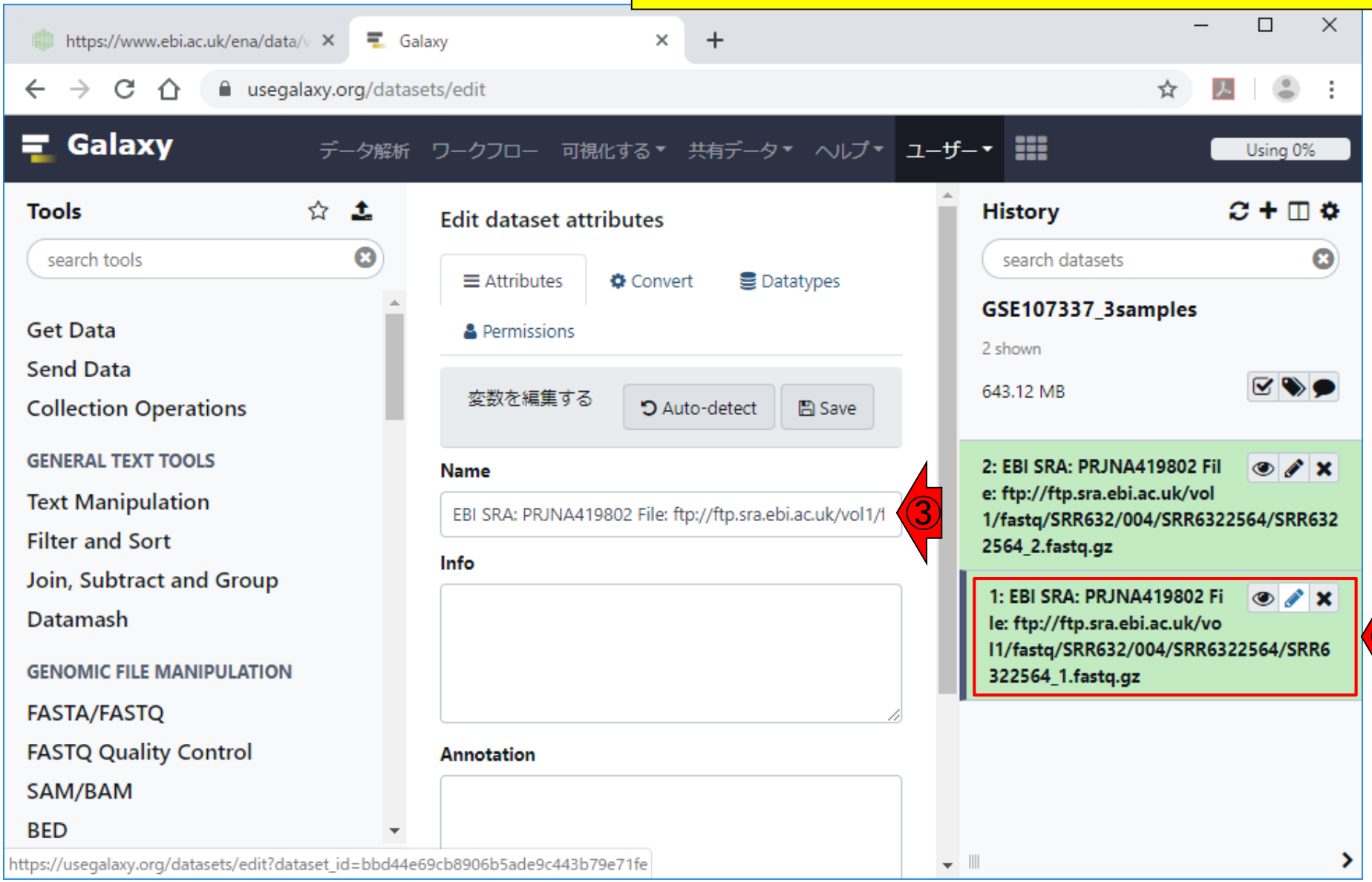
W3-3：rename

赤枠で示す部分の名前が長いので、①編集ボタンを押して名前を変更する。



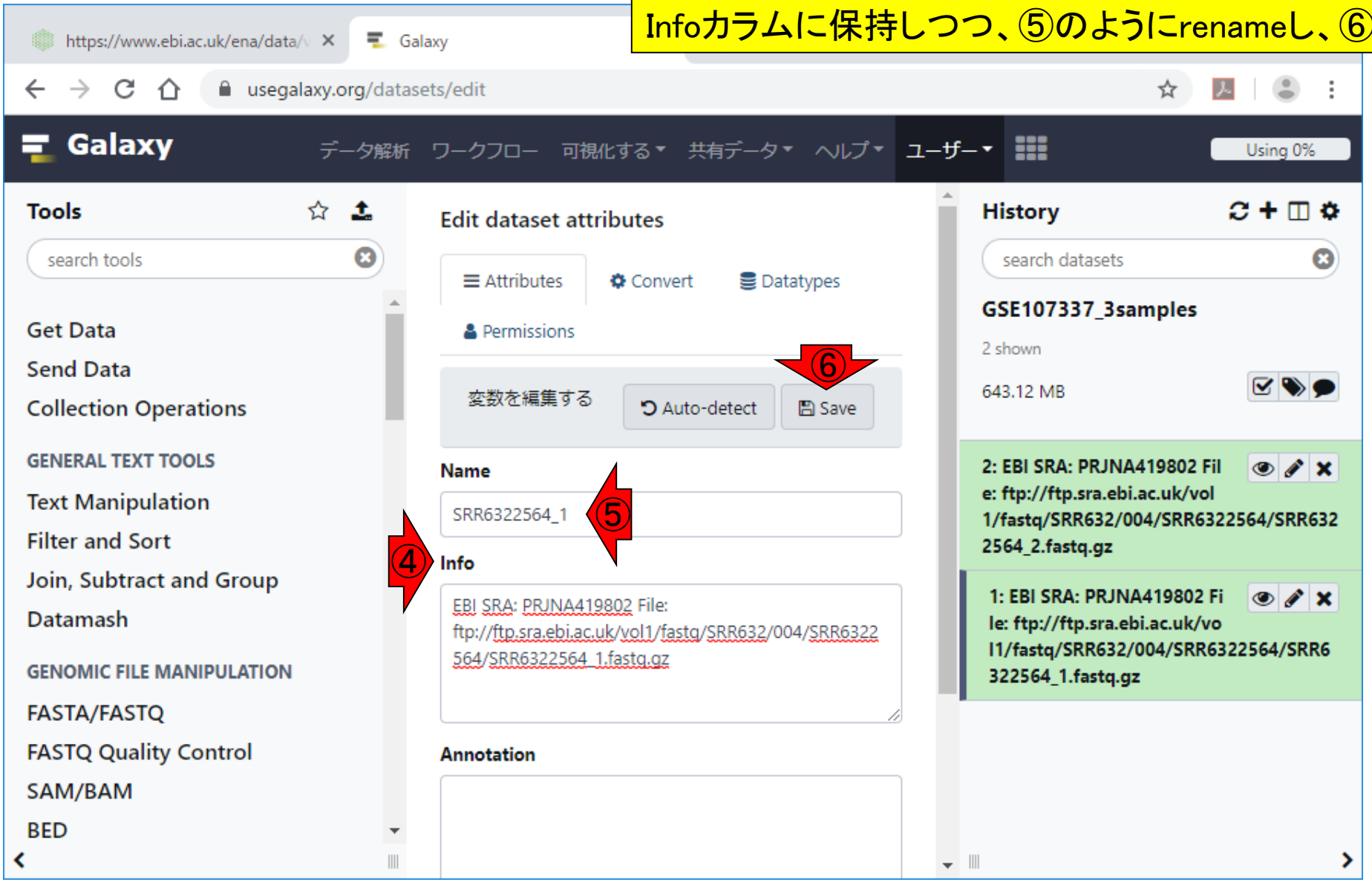
W3-3：rename

赤枠で示す部分の名前が長いので、①編集ボタンを押して名前を変更する。こんな感じになります。②赤枠で見えているのが、③の部分で編集可能。



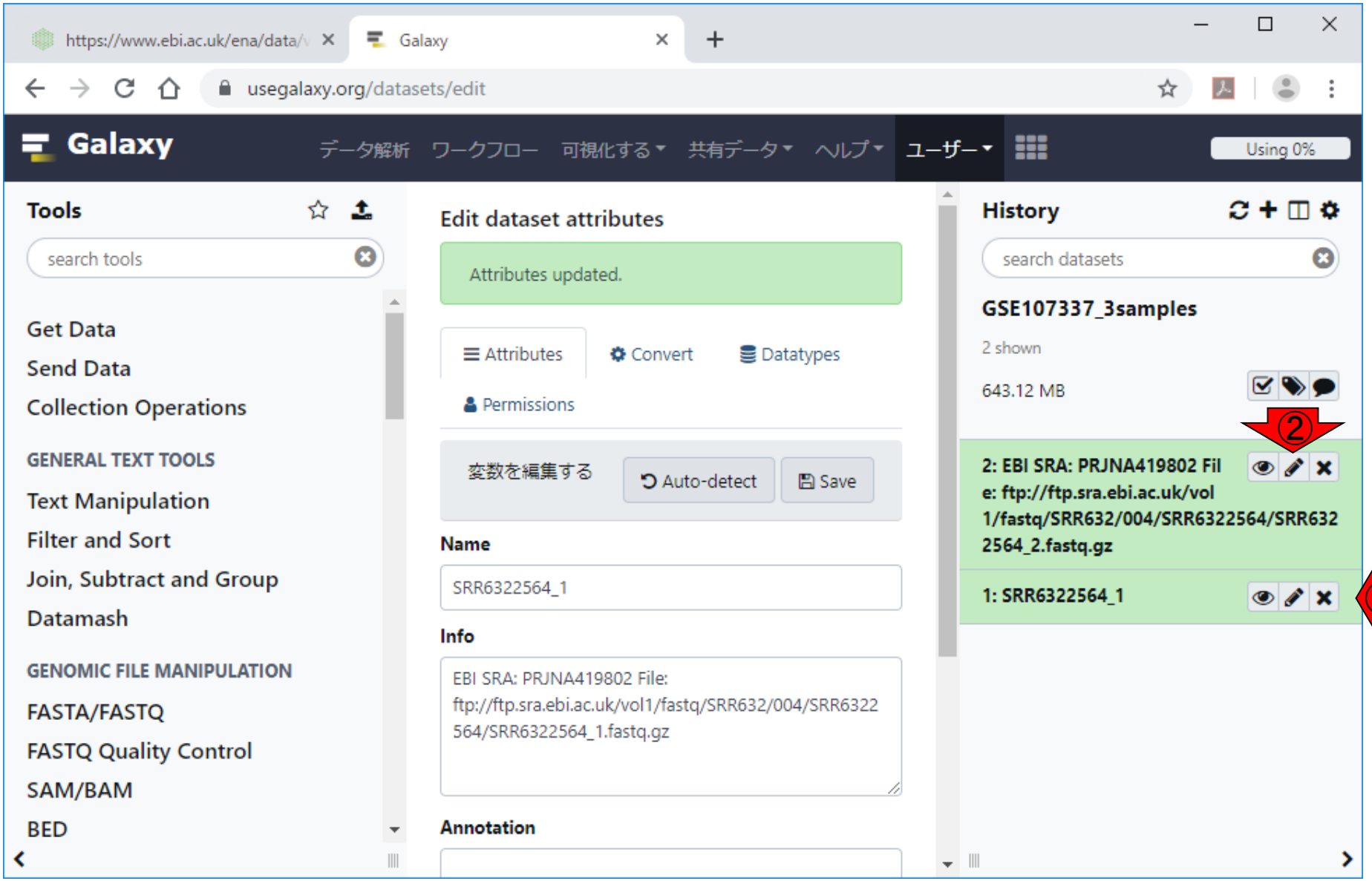
W3-3：rename

赤枠で示す部分の名前が長いので、①編集ボタンを押して名前を変更する。こんな感じになります。②赤枠で見えているのが、③の部分で編集可能。ここでは、元の情報を④Infoカラムに保持しつつ、⑤のようにrenameし、⑥Save。



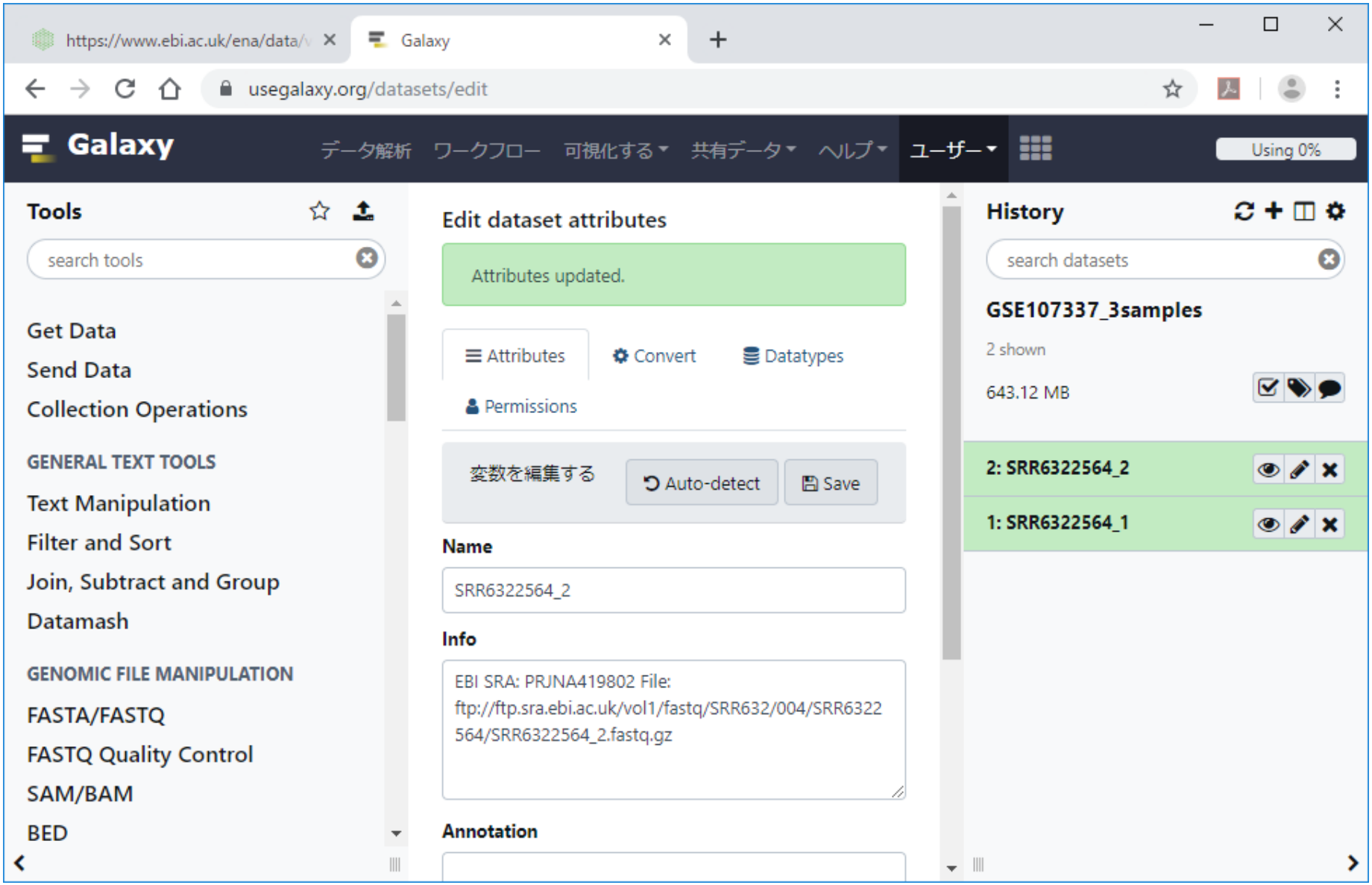
W3-3：rename

①forward側ファイルのrename完了。②reverse側についても同様に行う。



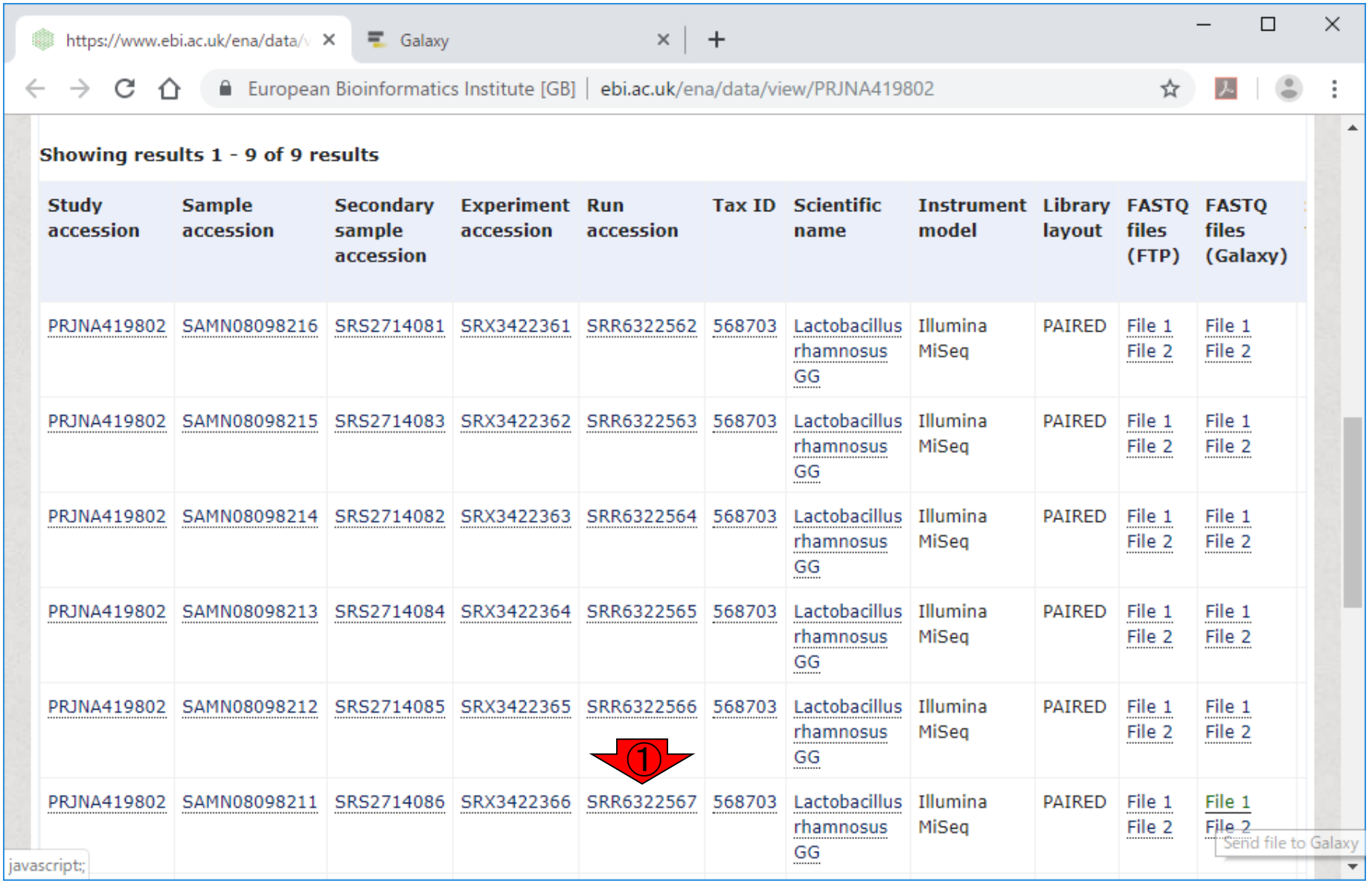
W3-3：rename

①forward側ファイルのrename完了。②reverse側についても同様に行う。作業完了。



再びENAのタブ。今度は①SRR6322567。クリックしない。

W3-4：SRR6322567



W3-4：SRR6322

再びENAのタブ。今度は①SRR6322567。クリックしない。Galaxyへの送付作業は、ファイル1つの送付完了を待つ必要はない。例えば、②File1をクリックして、0.2秒後にFile2をクリックしてもよい。やってみる。以前示したようにすぐに新規タブに飛ばされるが、すぐにENAに戻ってFile2をクリックすると次のスライドのようになる。

Showing results 1 - 9 of 9 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)
PRJNA419802	SAMN08098216	SRS2714081	SRX3422361	SRR6322562	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098215	SRS2714083	SRX3422362	SRR6322563	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098214	SRS2714082	SRX3422363	SRR6322564	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098213	SRS2714084	SRX3422364	SRR6322565	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098212	SRS2714085	SRX3422365	SRR6322566	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098211	SRS2714086	SRX3422366	SRR6322567	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2

①

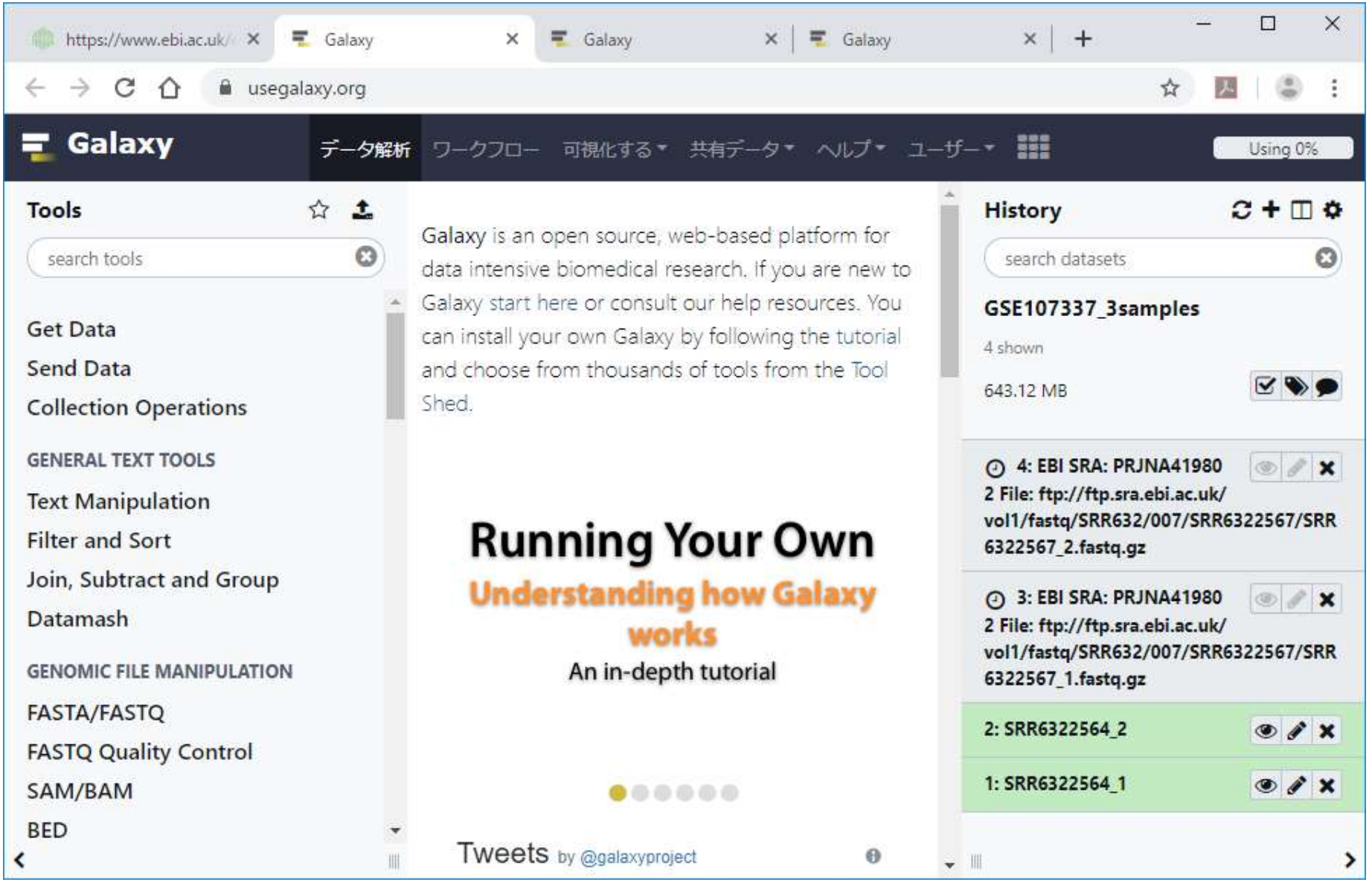
②

Send file to Galaxy

javascript:;

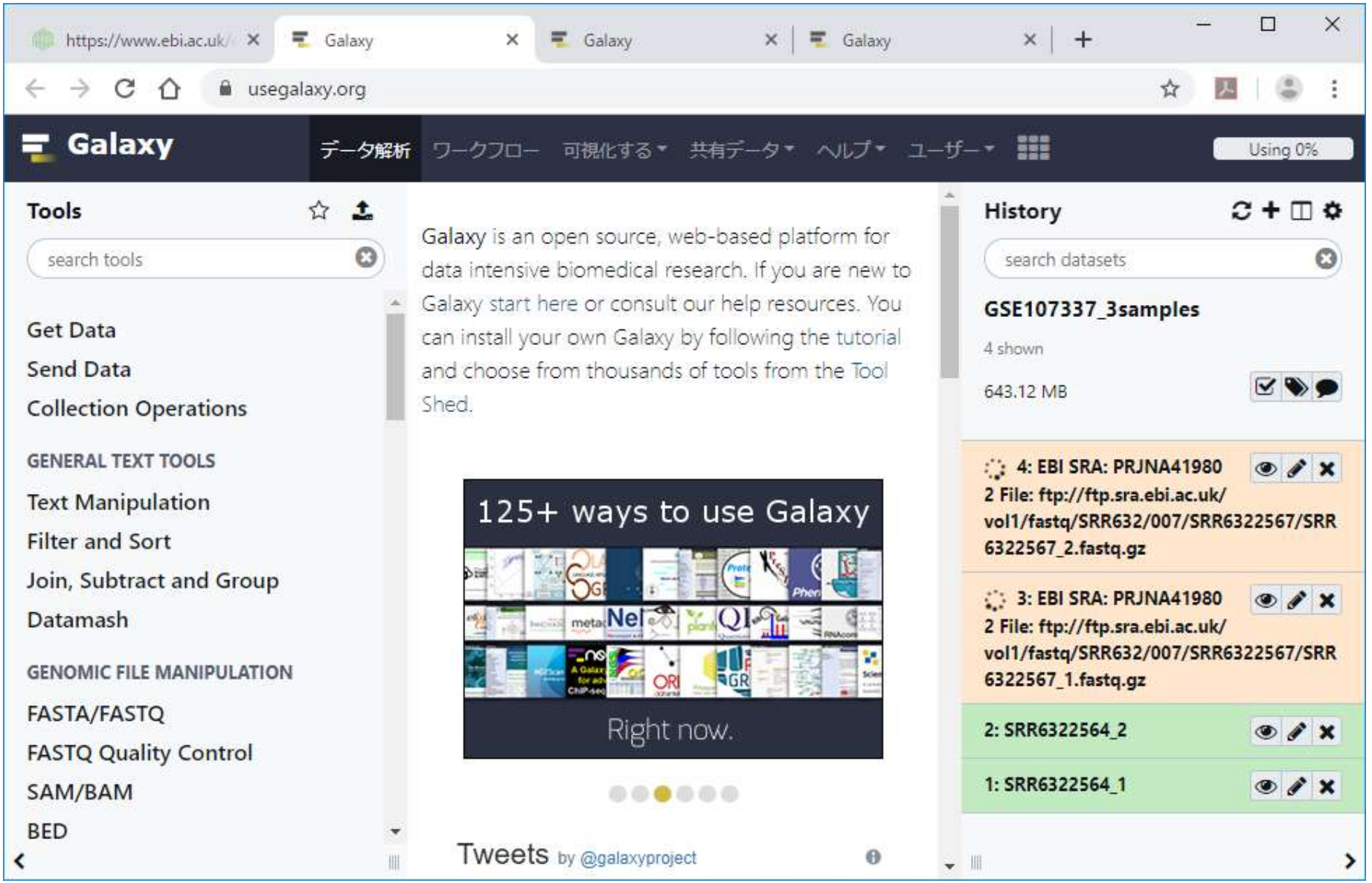
こんな感じ。

W3-4：SRR6322567



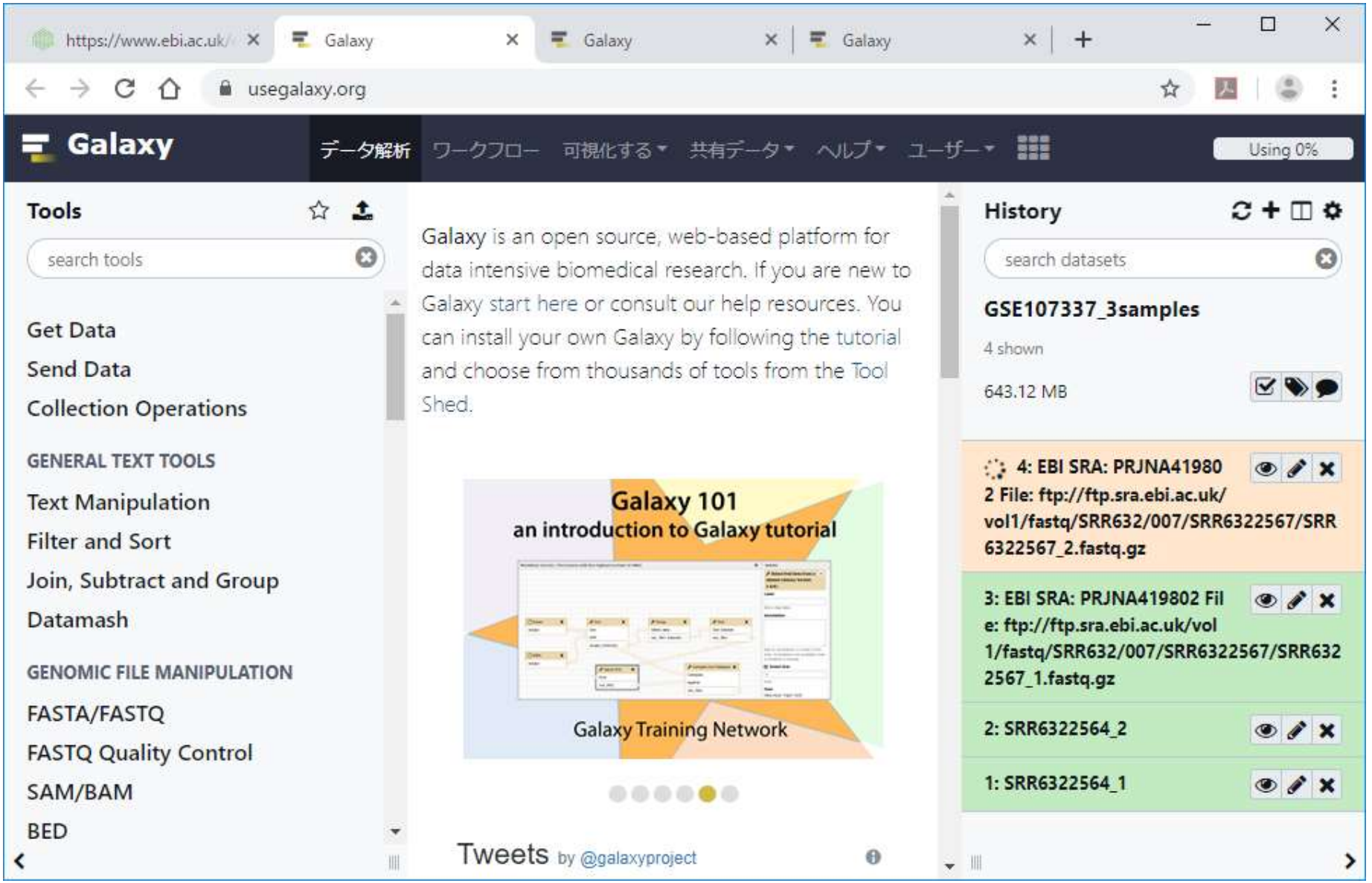
取り込み中…。

W3-4：SRR6322567



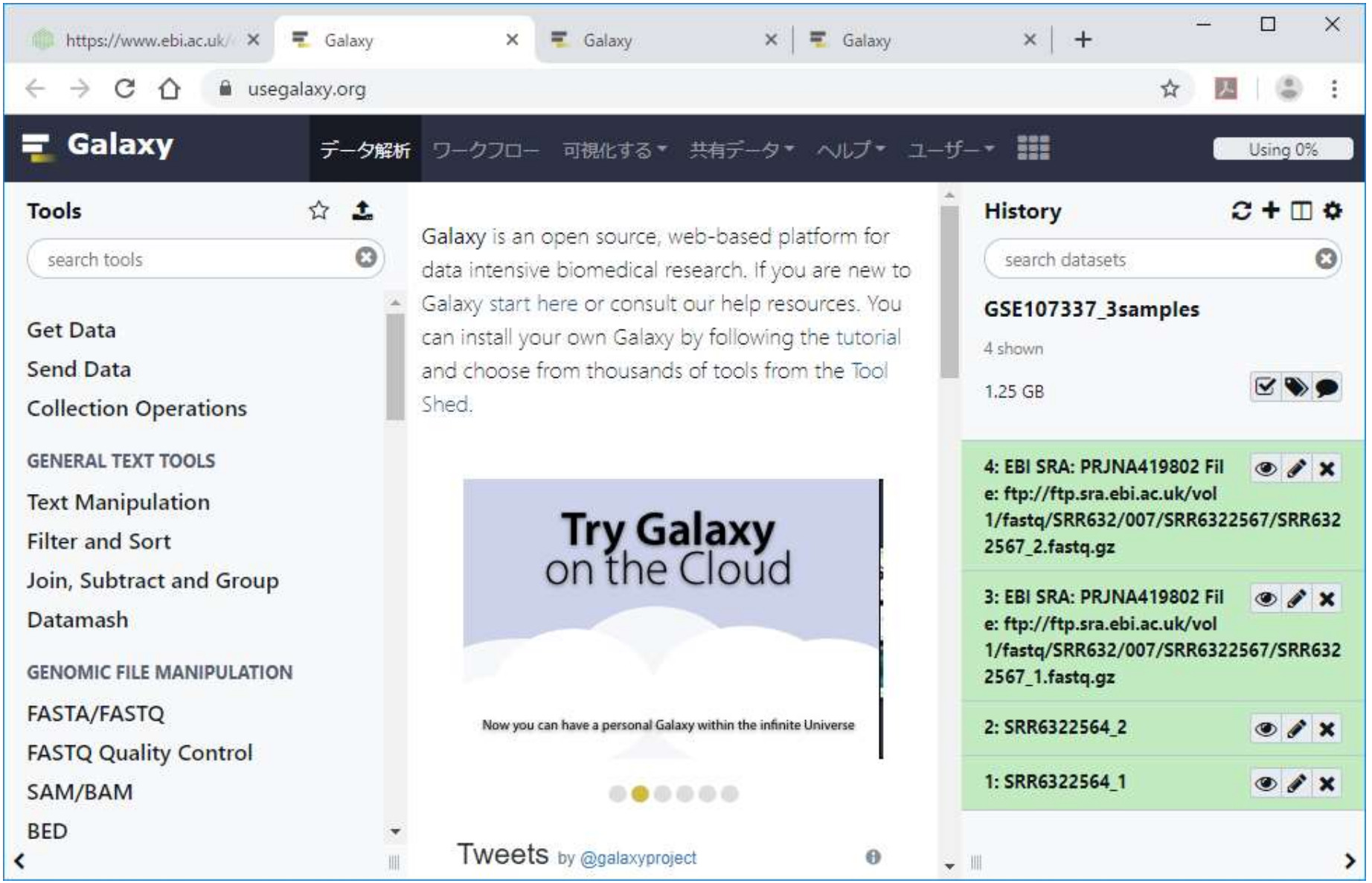
1つ完了。

W3-4：SRR6322567



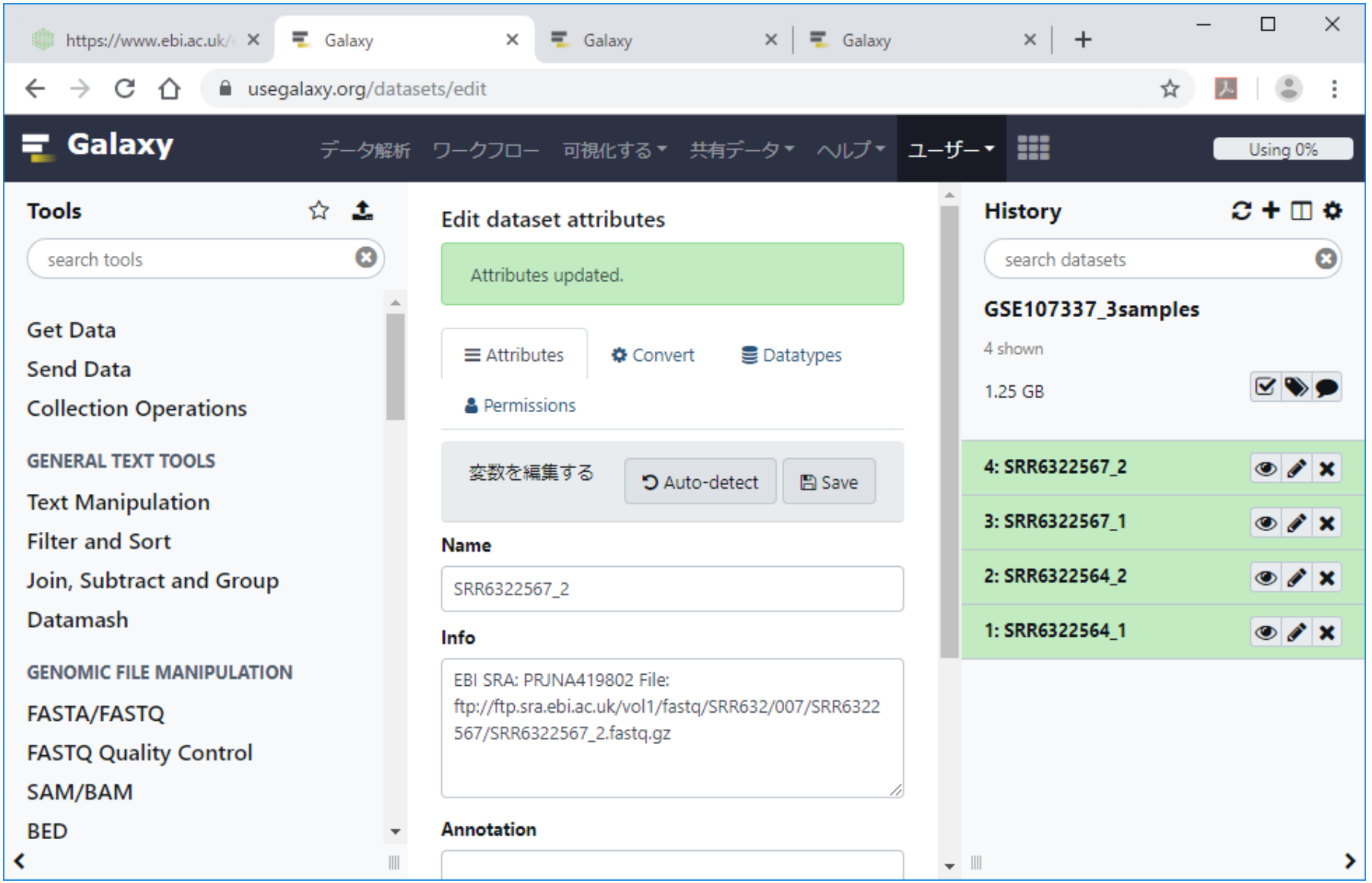
2つとも完了。

W3-4：SRR6322567



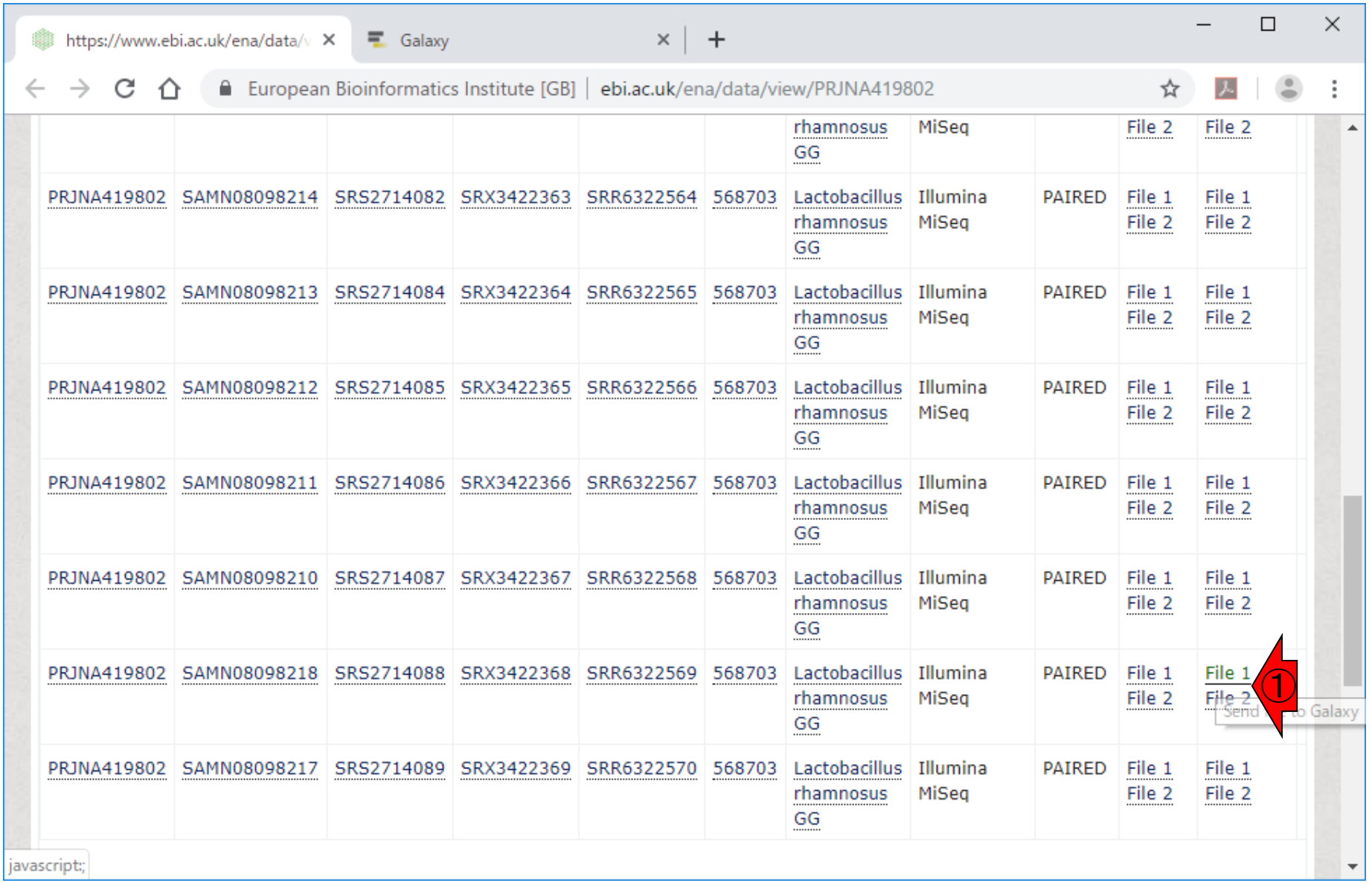
Rename（名前の変更）も完了。

W3-5：rename



最後は、①SRR6322569。

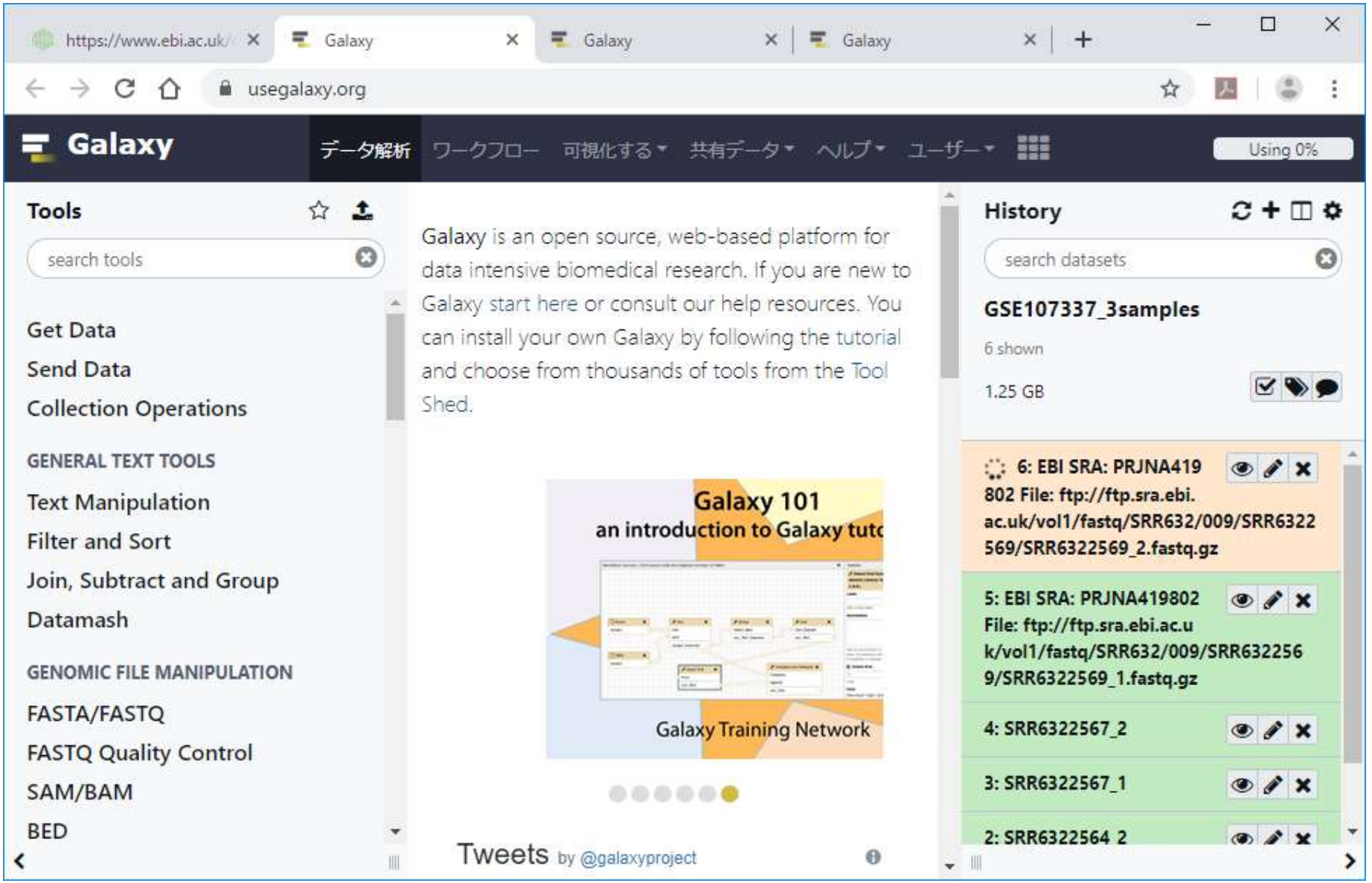
W3-6：SRR6322569



						rhamnosus GG	MiSeq		File 2	File 2
PRJNA419802	SAMN08098214	SRS2714082	SRX3422363	SRR6322564	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098213	SRS2714084	SRX3422364	SRR6322565	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098212	SRS2714085	SRX3422365	SRR6322566	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098211	SRS2714086	SRX3422366	SRR6322567	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098210	SRS2714087	SRX3422367	SRR6322568	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098218	SRS2714088	SRX3422368	SRR6322569	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098217	SRS2714089	SRX3422369	SRR6322570	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2

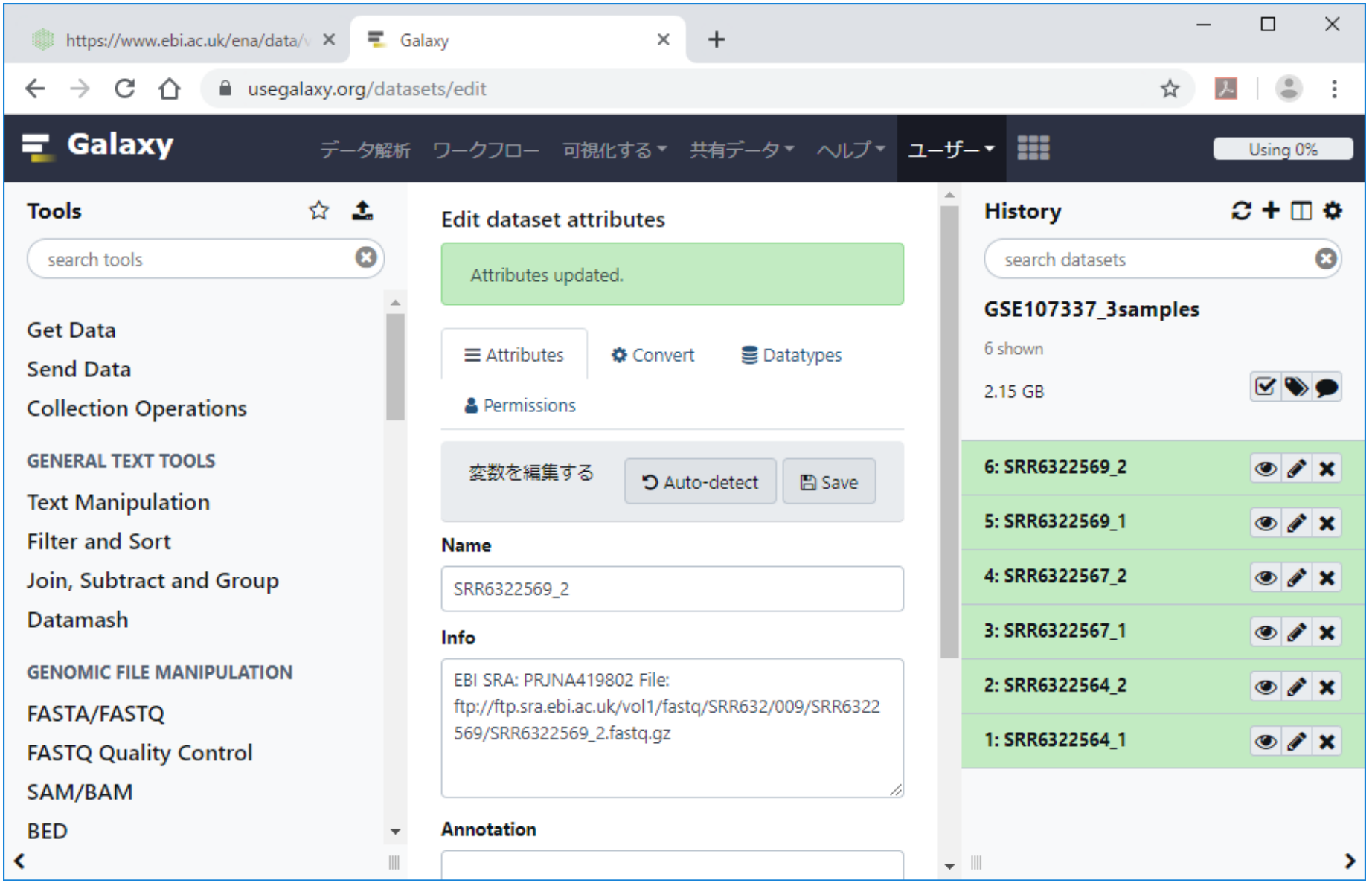
取り込み中…。

W3-6：SRR6322569



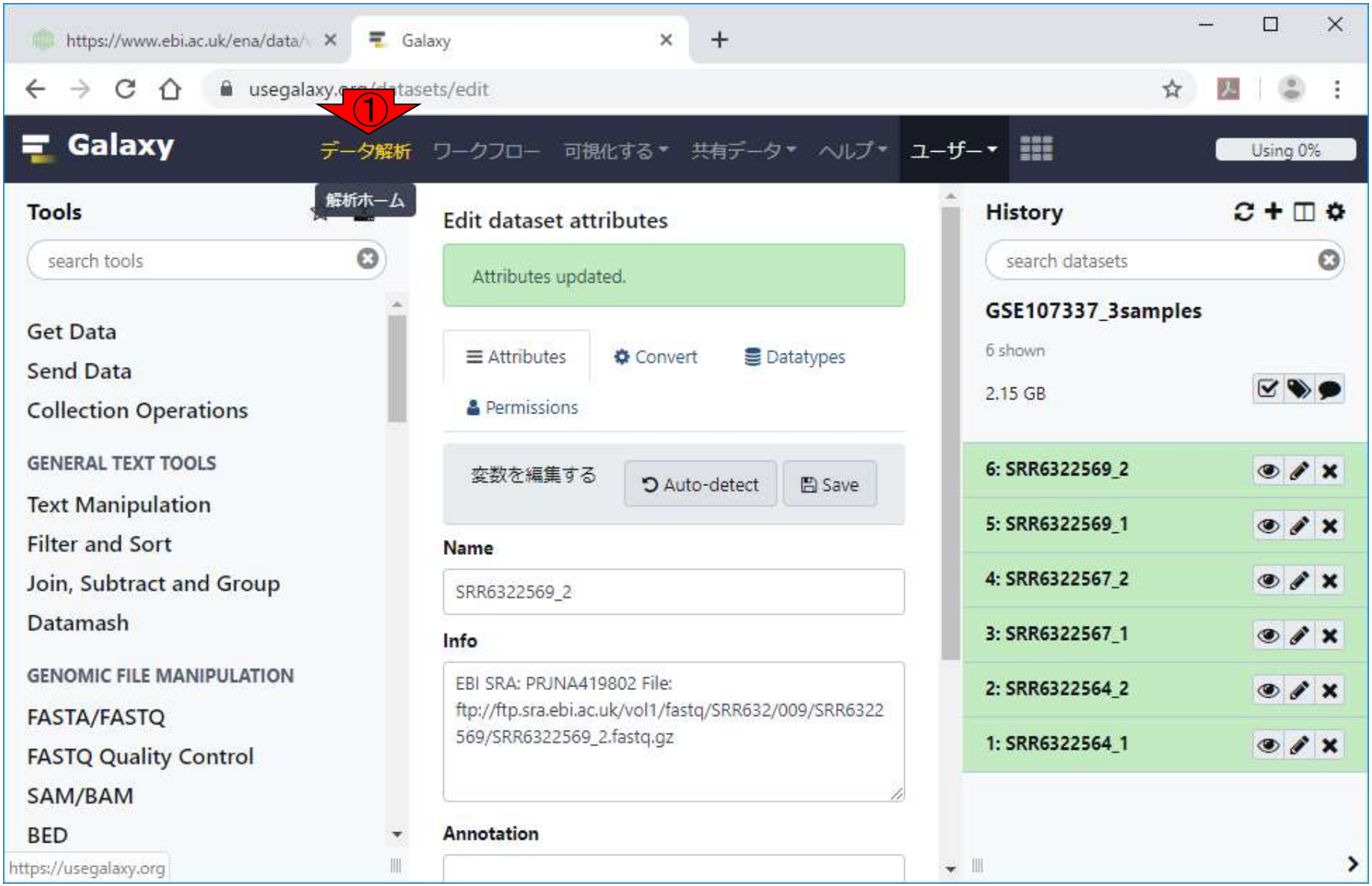
Renameまで完了。

W3-6：SRR6322569



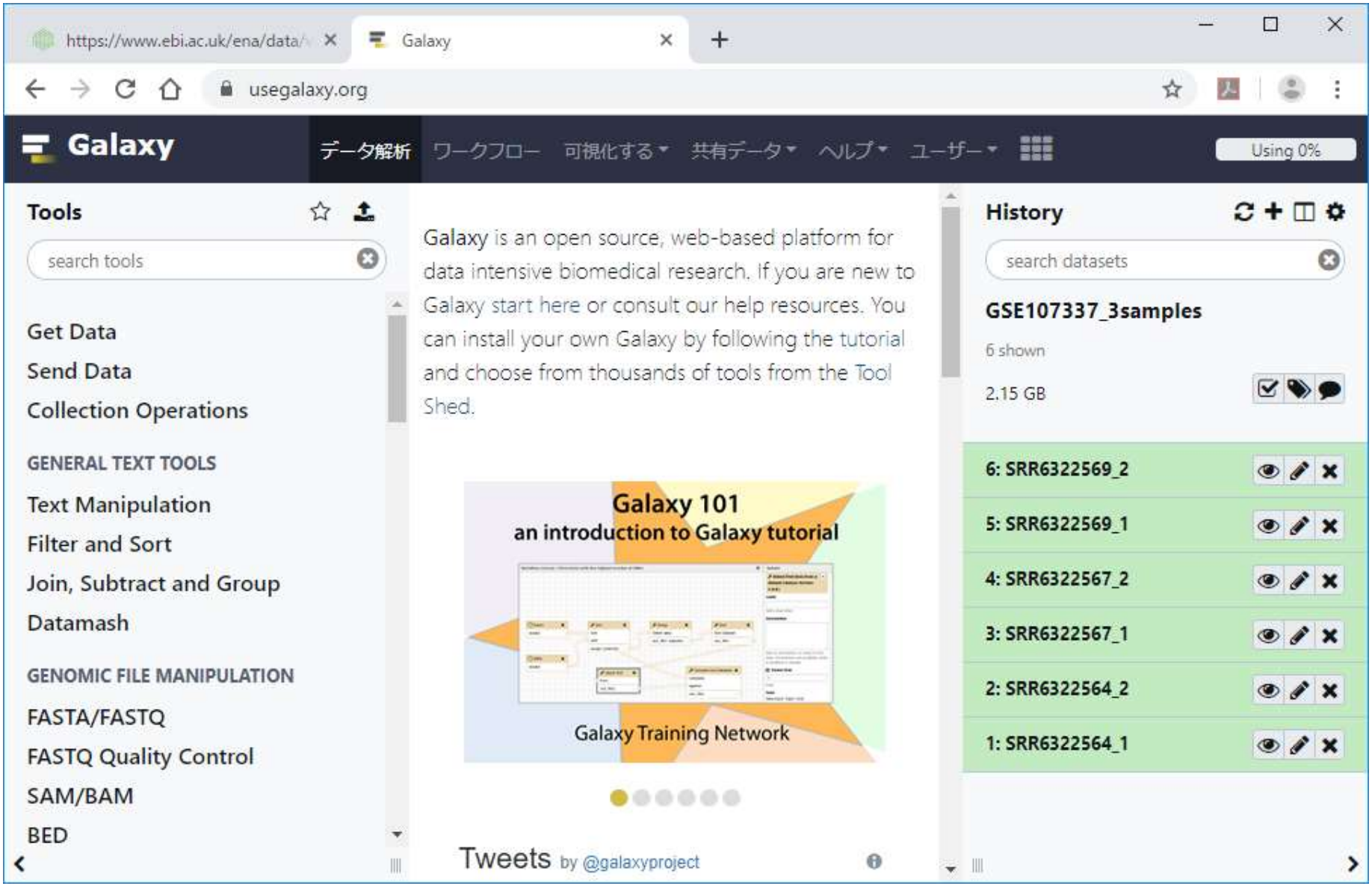
W3-6：SRR6322569

①データ解析のところを押すと、中央パネルをデフォルト画面に戻すことができます。



W3-7：解析準備完了

①データ解析のところを押すと、中央パネルをデフォルト画面に戻すことができます。こんな感じ。

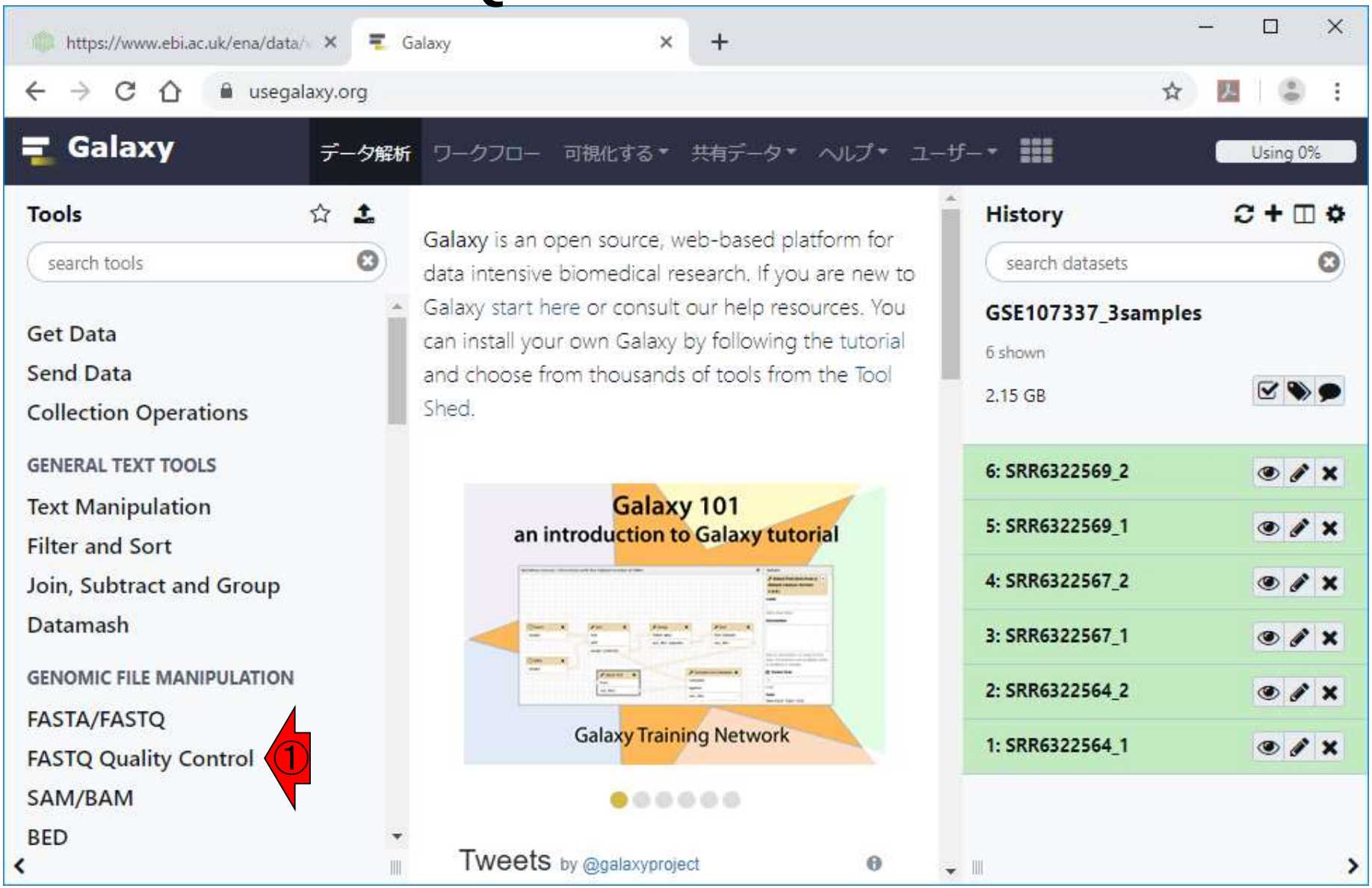


Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- **W4：クオリティチェック（FastQC）**
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

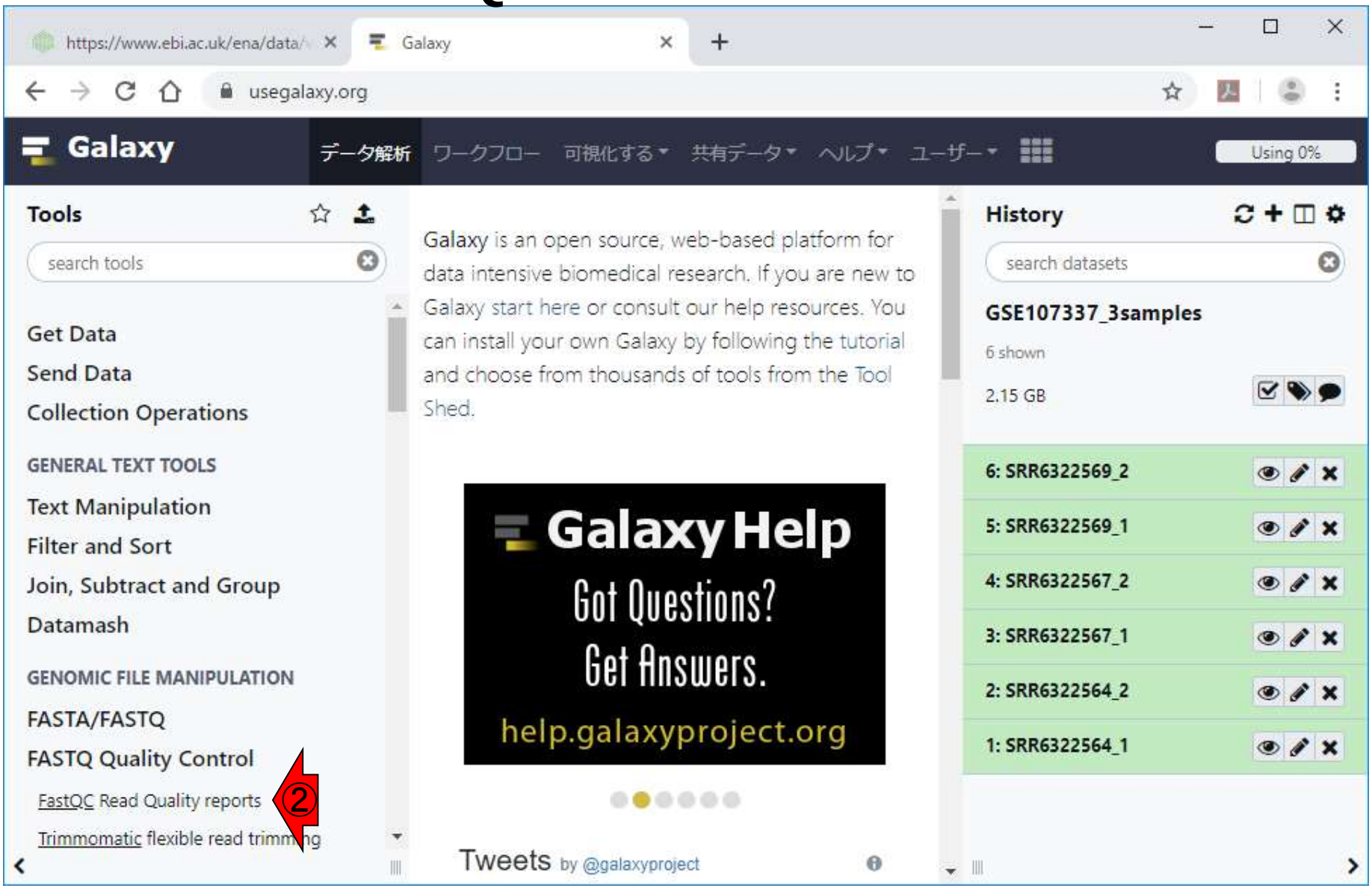
次はFastQC。①の中にFastQCがあるので、クリック。

W4-1：FastQC



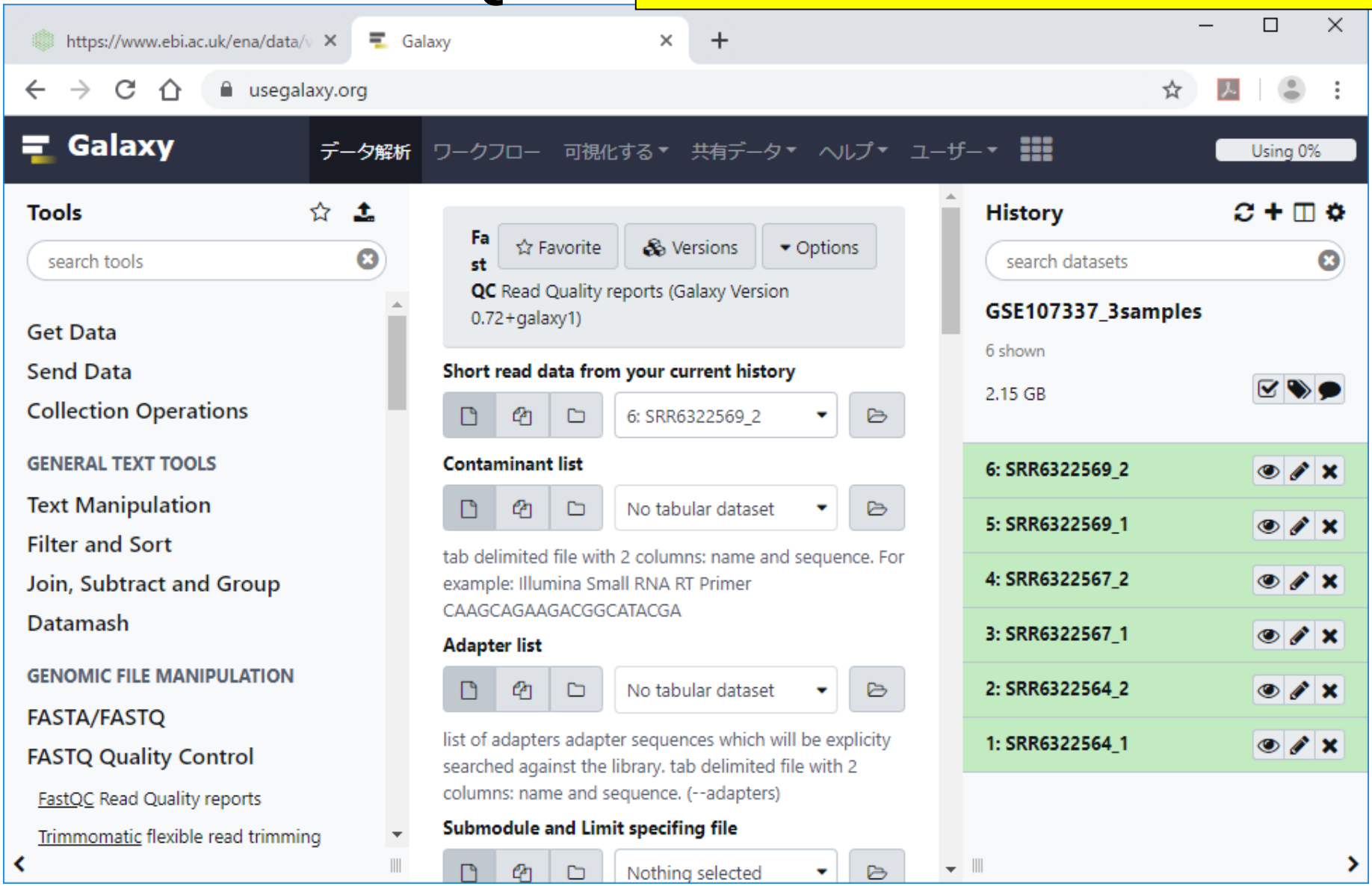
W4-1：FastQC

次はFastQC。①の中にFastQCがあるので、クリック。②FastQC。



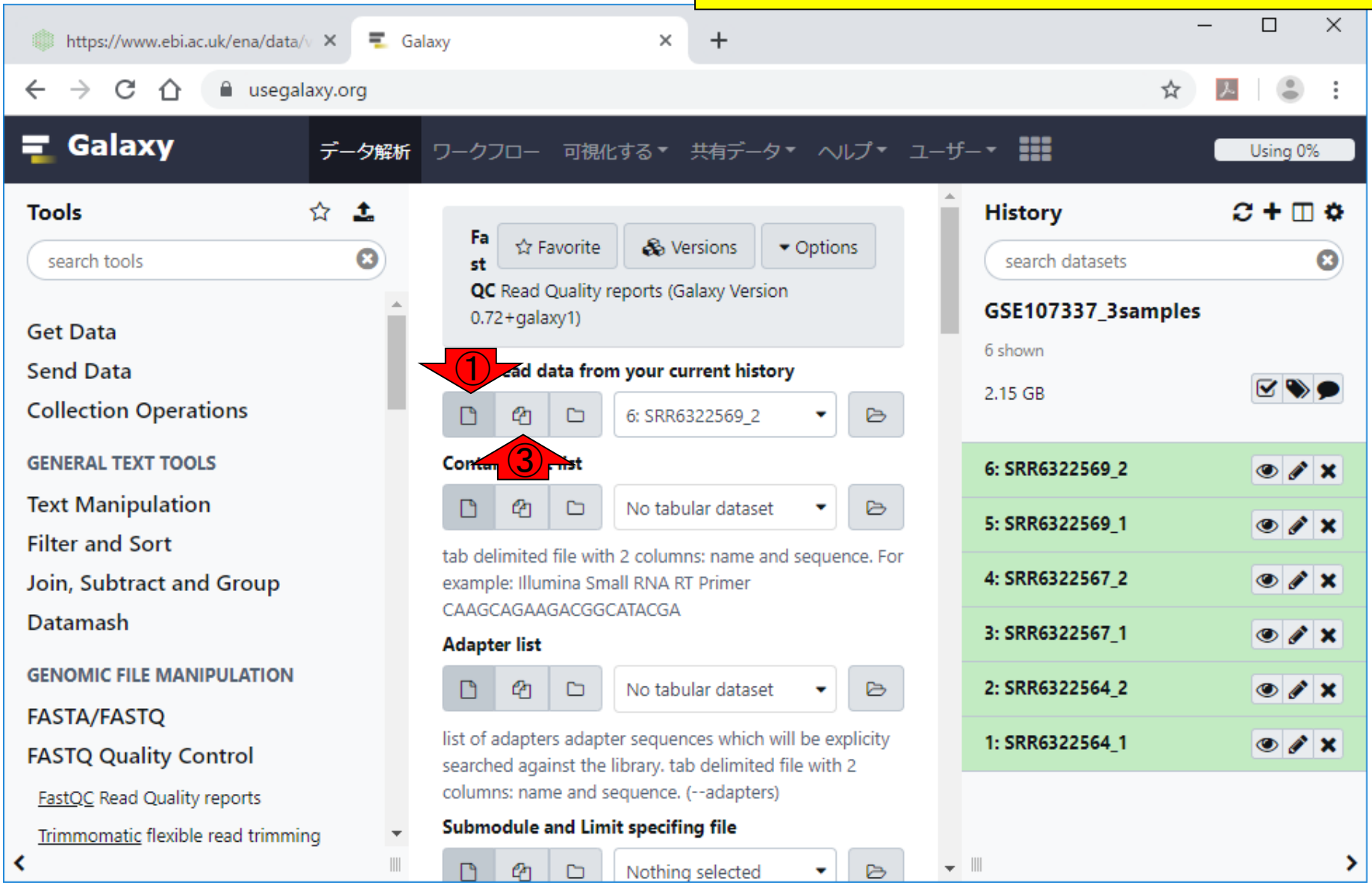
W4-1：FastQC

次はFastQC。①の中にFastQCがあるので、クリック。②FastQC。こんな感じになる。第11回のW9でもFastQCをやっています。



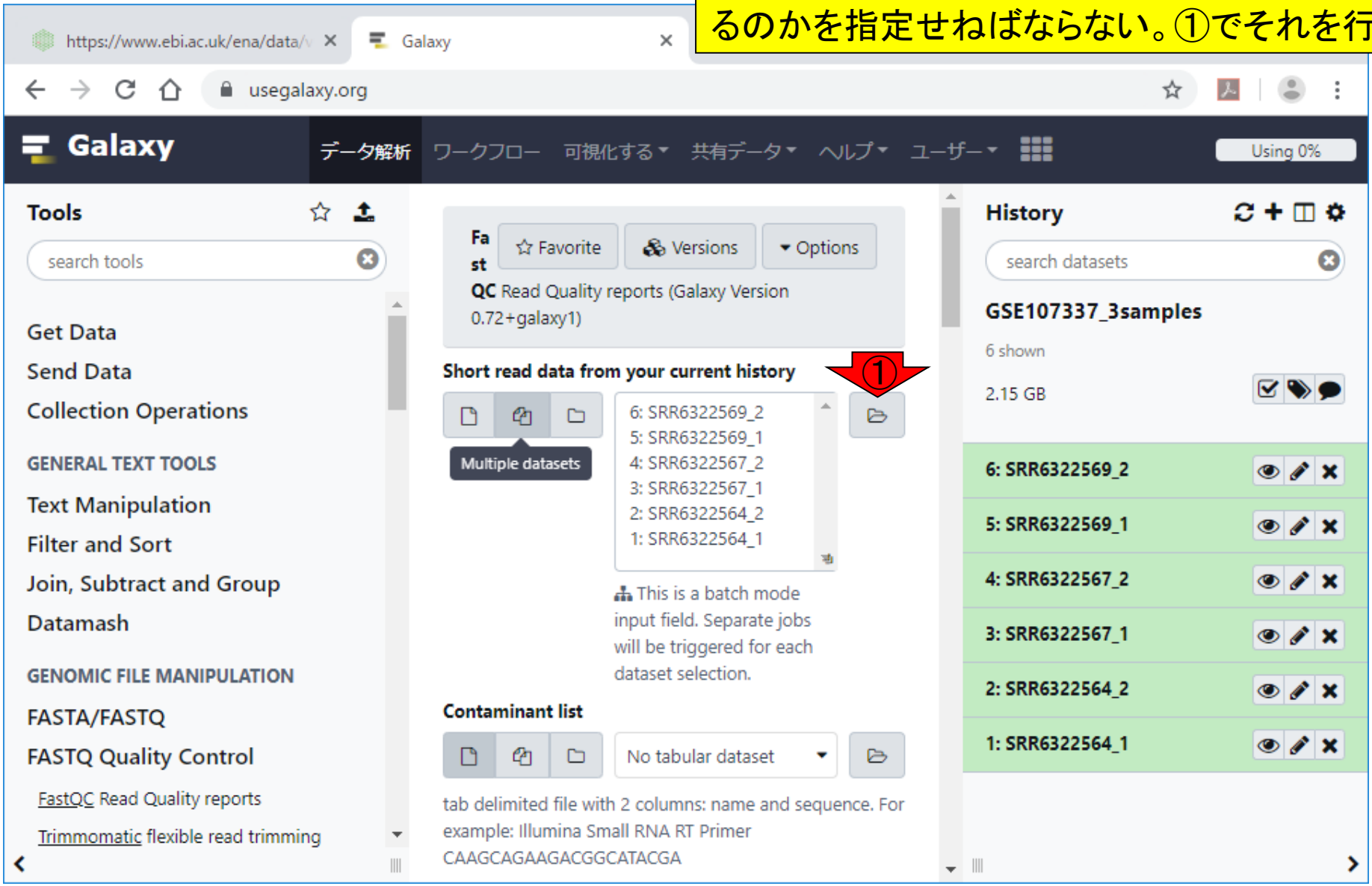
W4-2：複数個指定

デフォルトは①Single datasetだが、ここでは②計6ファイル全てについてFastQCを実行したいので、③Multiple datasetsに切り替える。



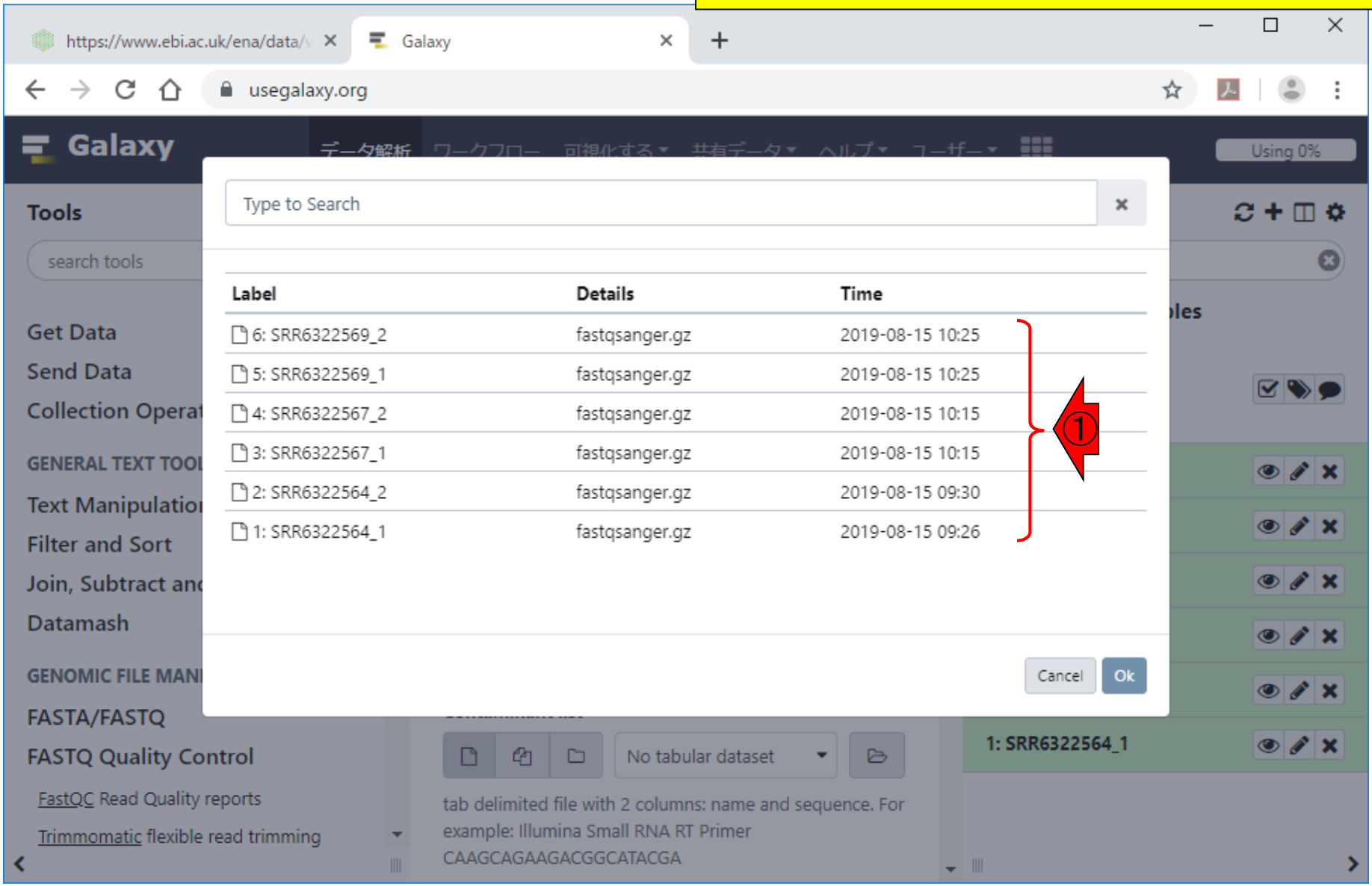
W4-2：複数個指定

こんな感じになる。現状では、まだヒストリーパネル上で実行可能な入力データがリストアップされているだけなので、この中からどれをFastQCの入力として実行するのかを指定せねばならない。①でそれを行う。



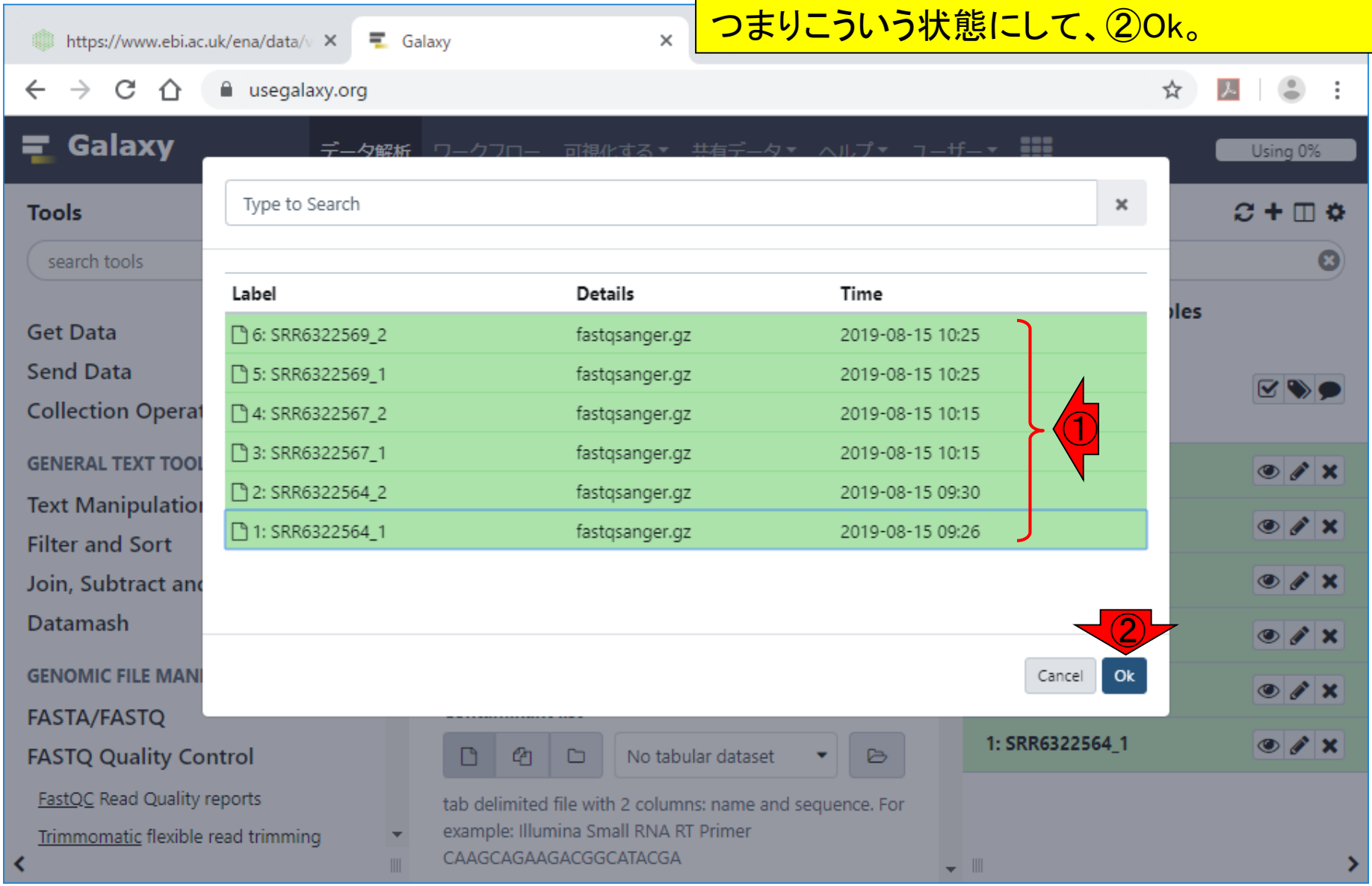
W4-2：複数個指定

こんな感じになって、ヒストリーパネル上で実行可能なFASTQファイルがリストアップされる。実行したいのは全てなので①全てのファイルをクリックして選択する。



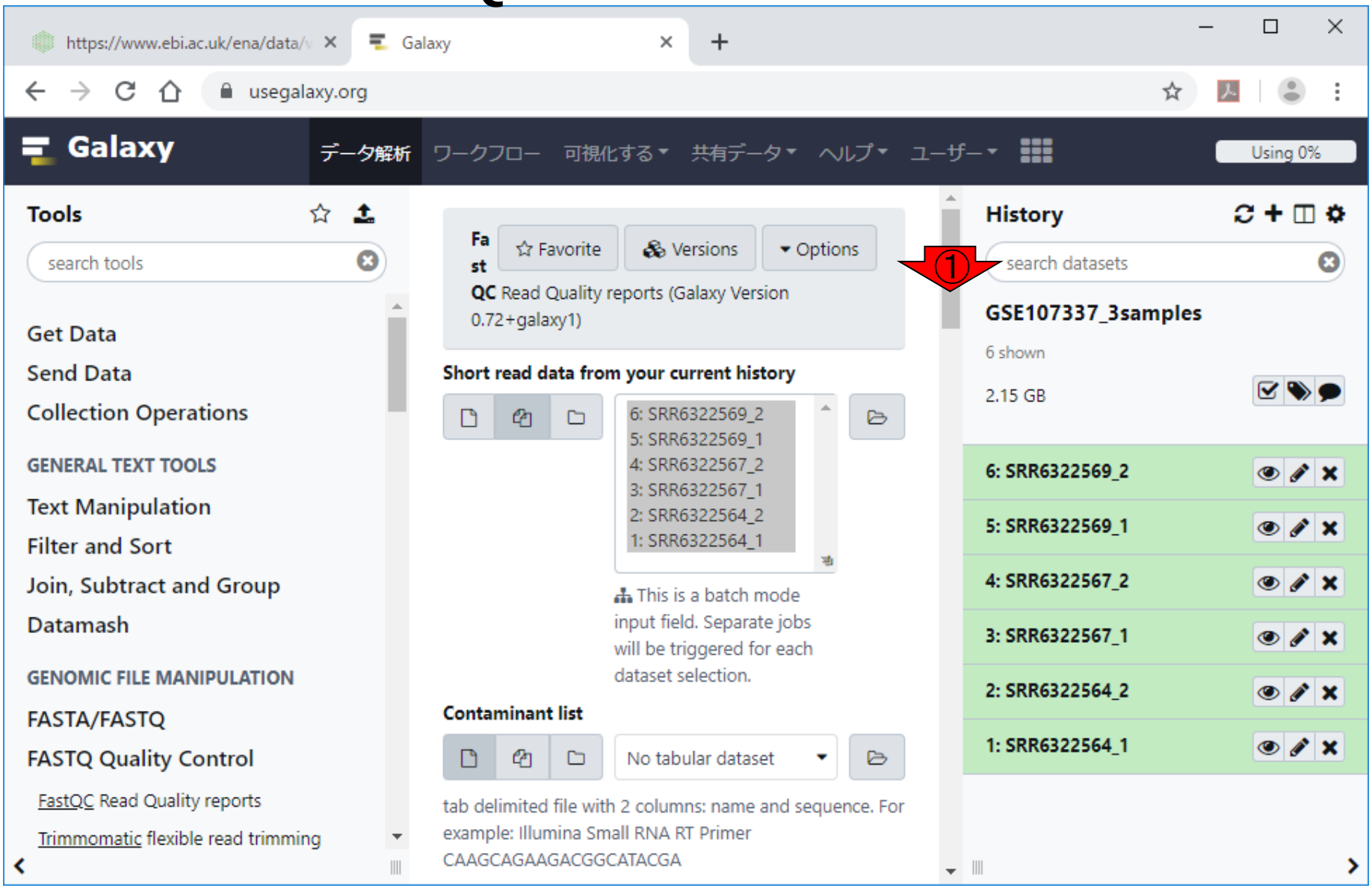
W4-2：複数個指定

こんな感じになって、ヒストリーパネル上で実行可能なFASTQファイルがリストアップされる。実行したいのは全てなので①全てのファイルをクリックして選択する。
つまりこういう状態にして、②Ok。



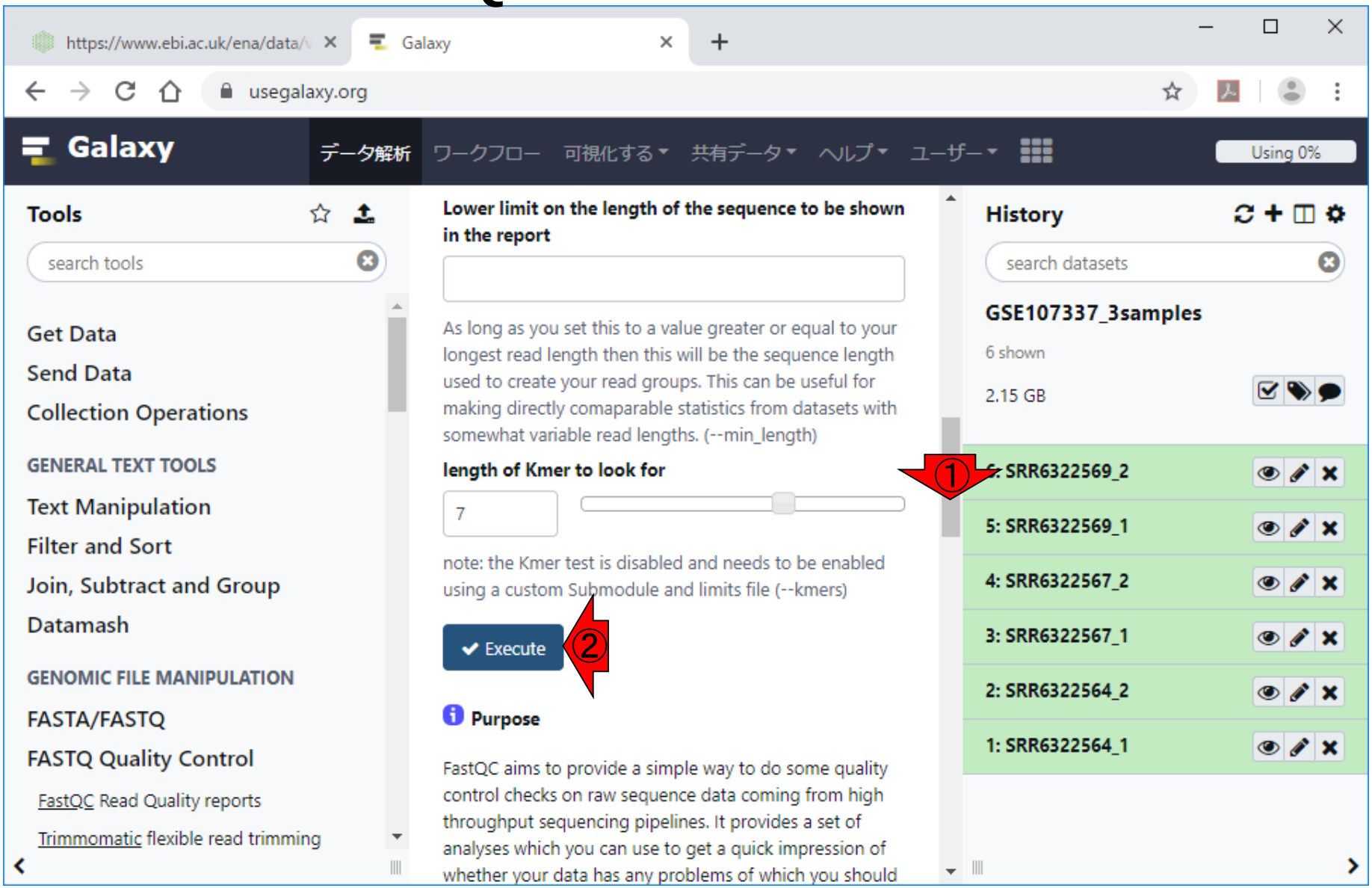
W4-3：FastQC実行

こんな感じになる。①下部に移動して、実行ボタンを探す。



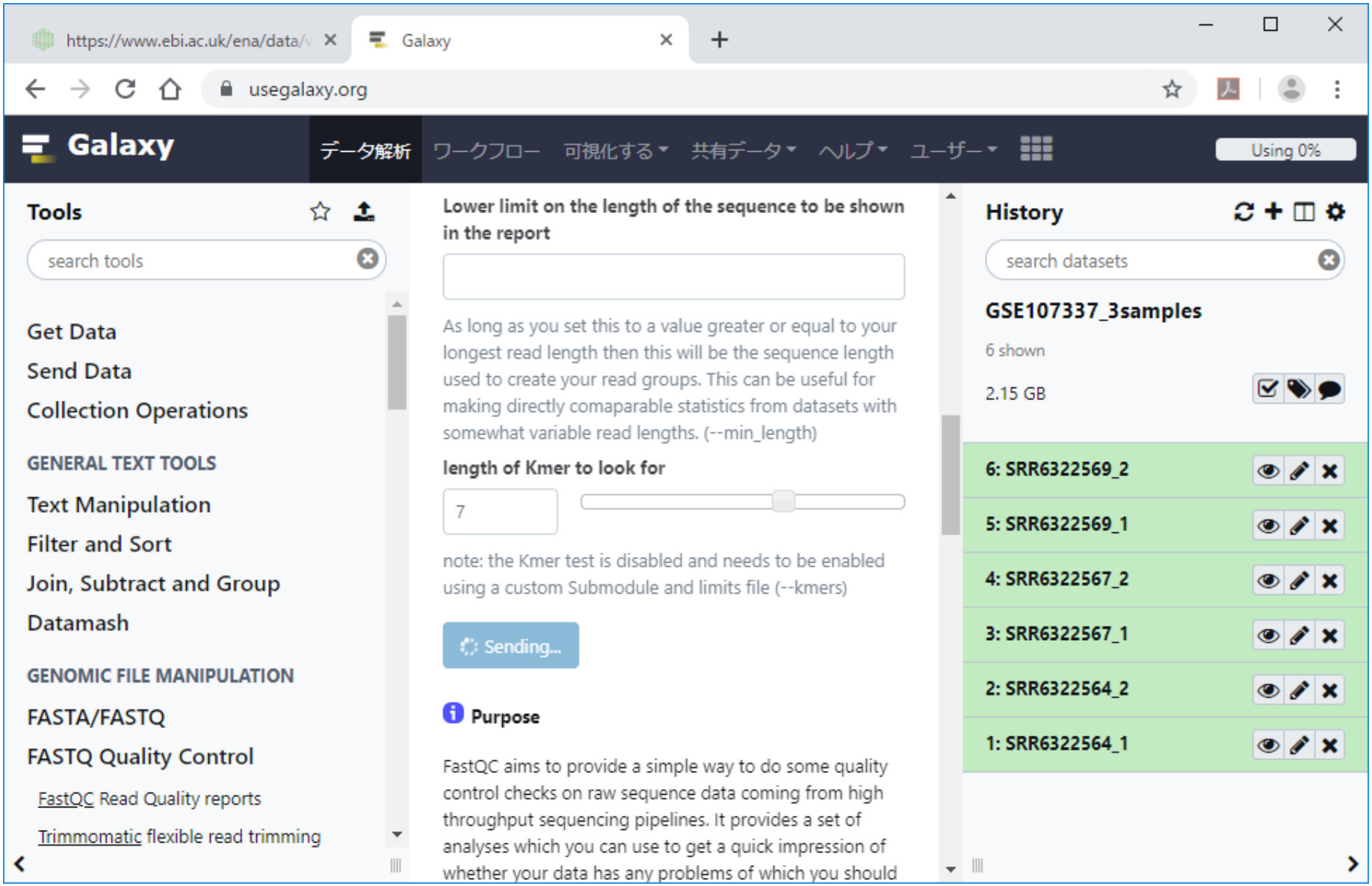
W4-3：FastQC実行

こんな感じになる。①下部に移動して、実行ボタンを探す。②Execute。



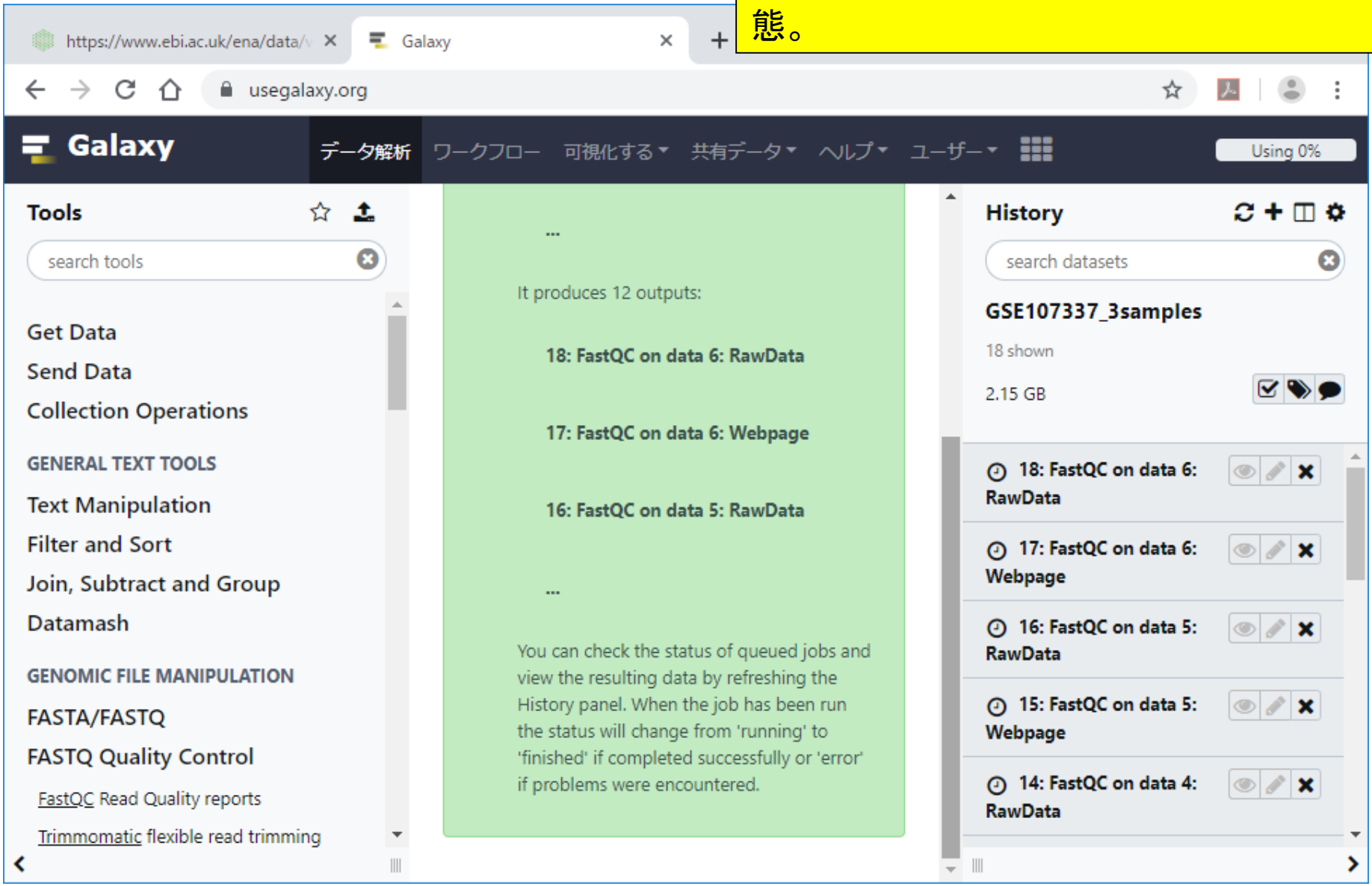
W4-4：実行中

こんな感じになる。①下部に移動して、実行ボタンを探す。②Execute。ボタンを押した直後。



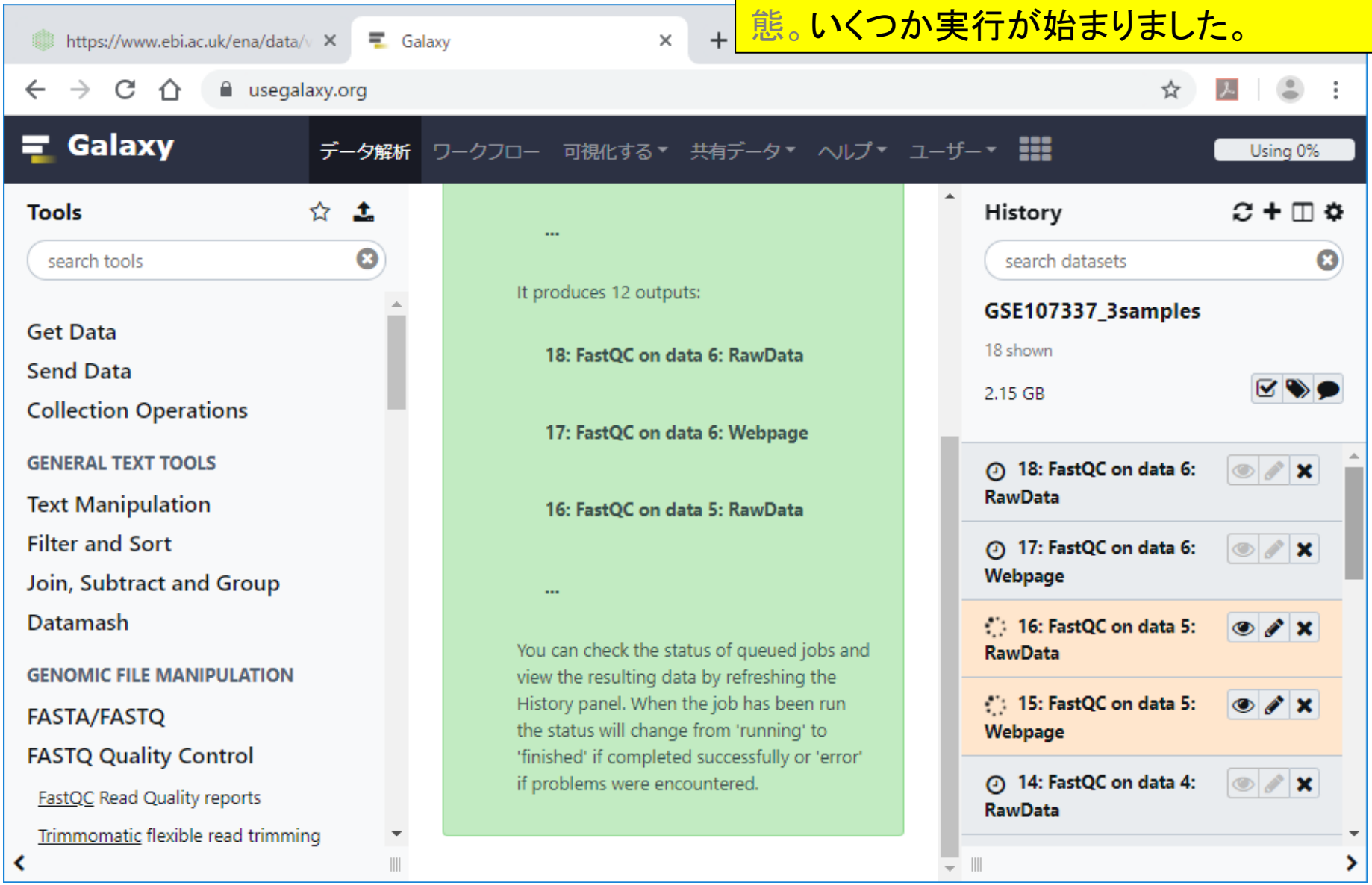
W4-4：実行中

こんな感じになる。①下部に移動して、実行ボタンを探す。②Execute。ボタンを押した直後。すぐにこんな感じに画面が切り替わる。灰色なので実行待ち状態。



W4-4：実行中

こんな感じになる。①下部に移動して、実行ボタンを探す。②Execute。ボタンを押した直後。すぐにこんな感じに画面が切り替わる。灰色なので実行待ち状態。いくつか実行が始まりました。

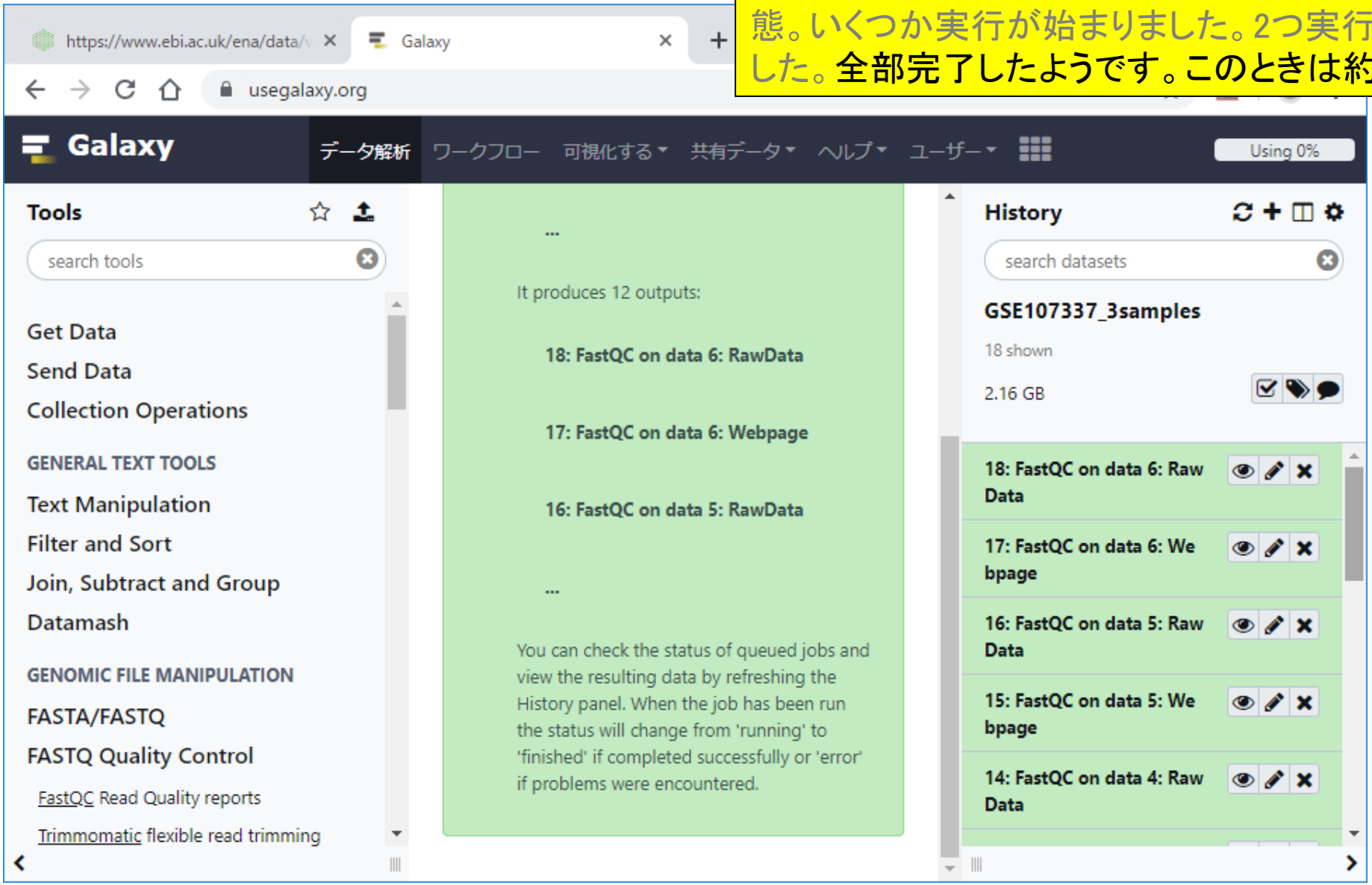


W4-4：実行中

こんな感じになる。①下部に移動して、実行ボタンを探す。②Execute。ボタンを押した直後。すぐにこんな感じに画面が切り替わる。灰色なので実行待ち状態。いくつか実行が始まりました。2つ実行完了しました。

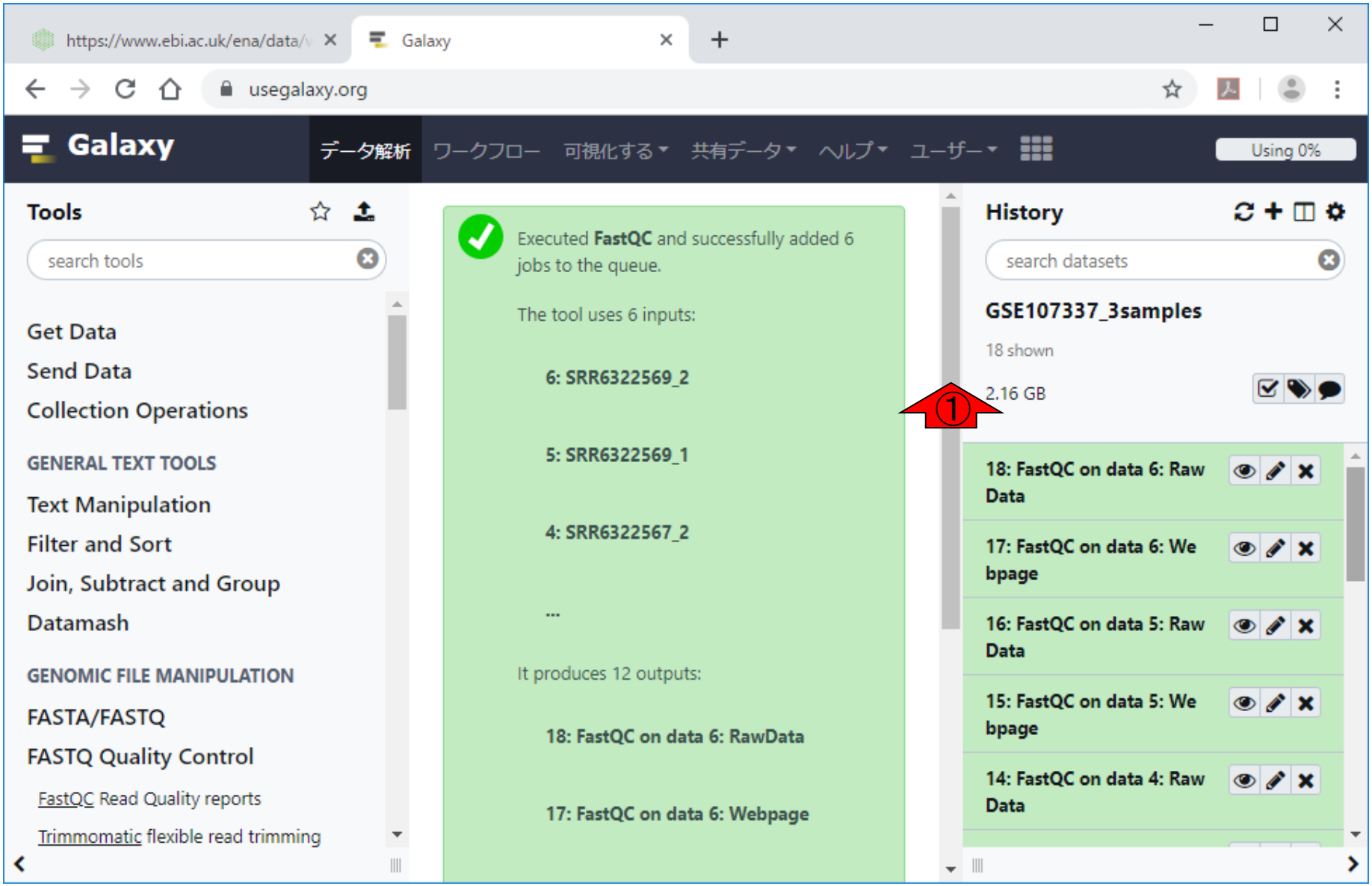
W4-5：実行完了

こんな感じになる。①下部に移動して、実行ボタンを探す。②Execute。ボタンを押した直後。すぐにこんな感じに画面が切り替わる。灰色なので実行待ち状態。いくつか実行が始まりました。2つ実行完了しました。全部完了したようです。このときは約5分。



①中央のパネルを一番上まで移動させただけ。

W4-5：実行完了

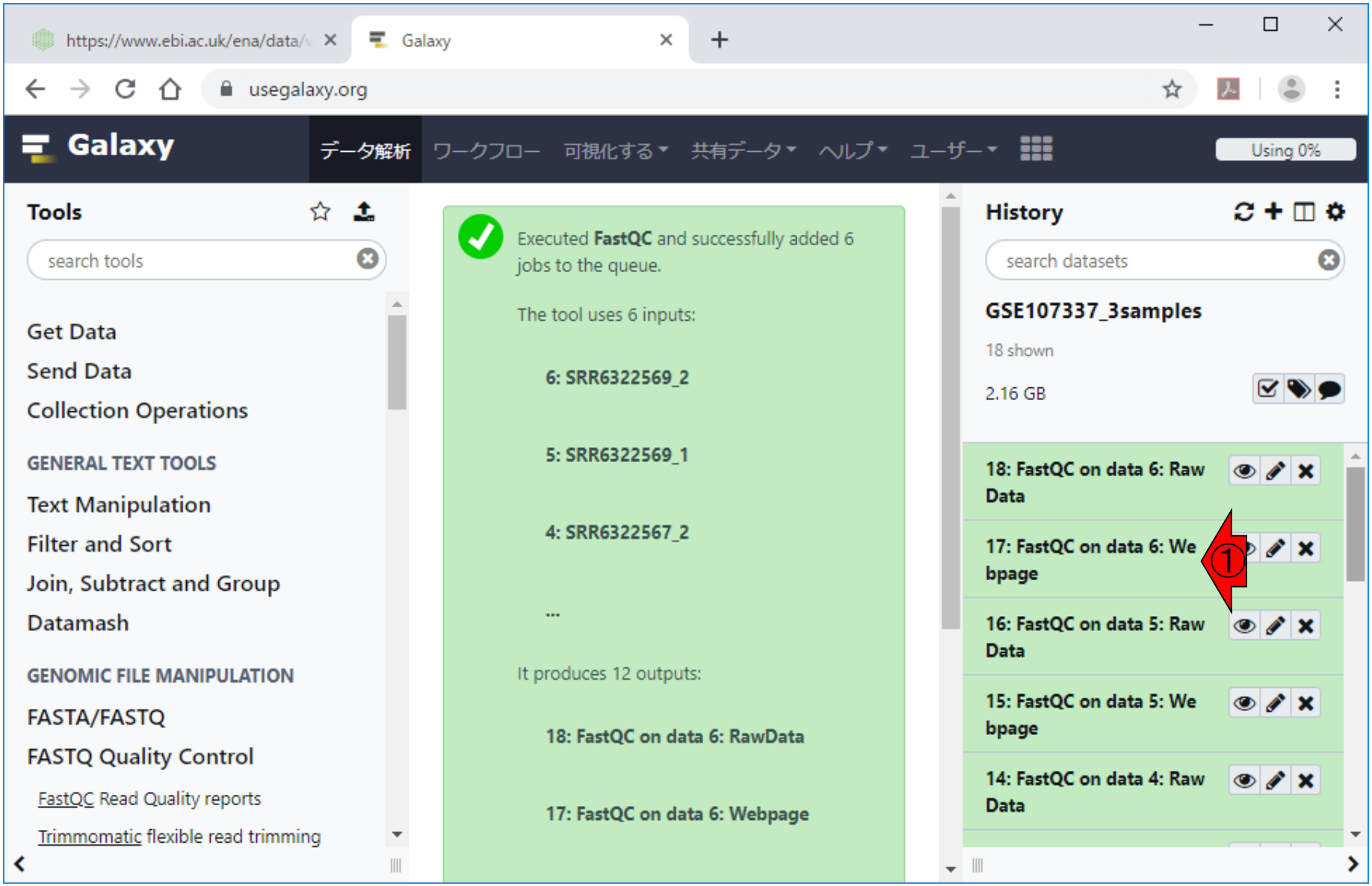


W4-6：解説

①FastQCの入力は、ヒストリー番号1-6のFASTQファイル。②出力は計12個。例えば画面上で見えているヒストリー17と18は、③data 6（つまりヒストリー6）の実行結果であることがわかる。④ヒストリー17はhtmlファイル版で、⑤ヒストリー18は生データ版だと読み解く。

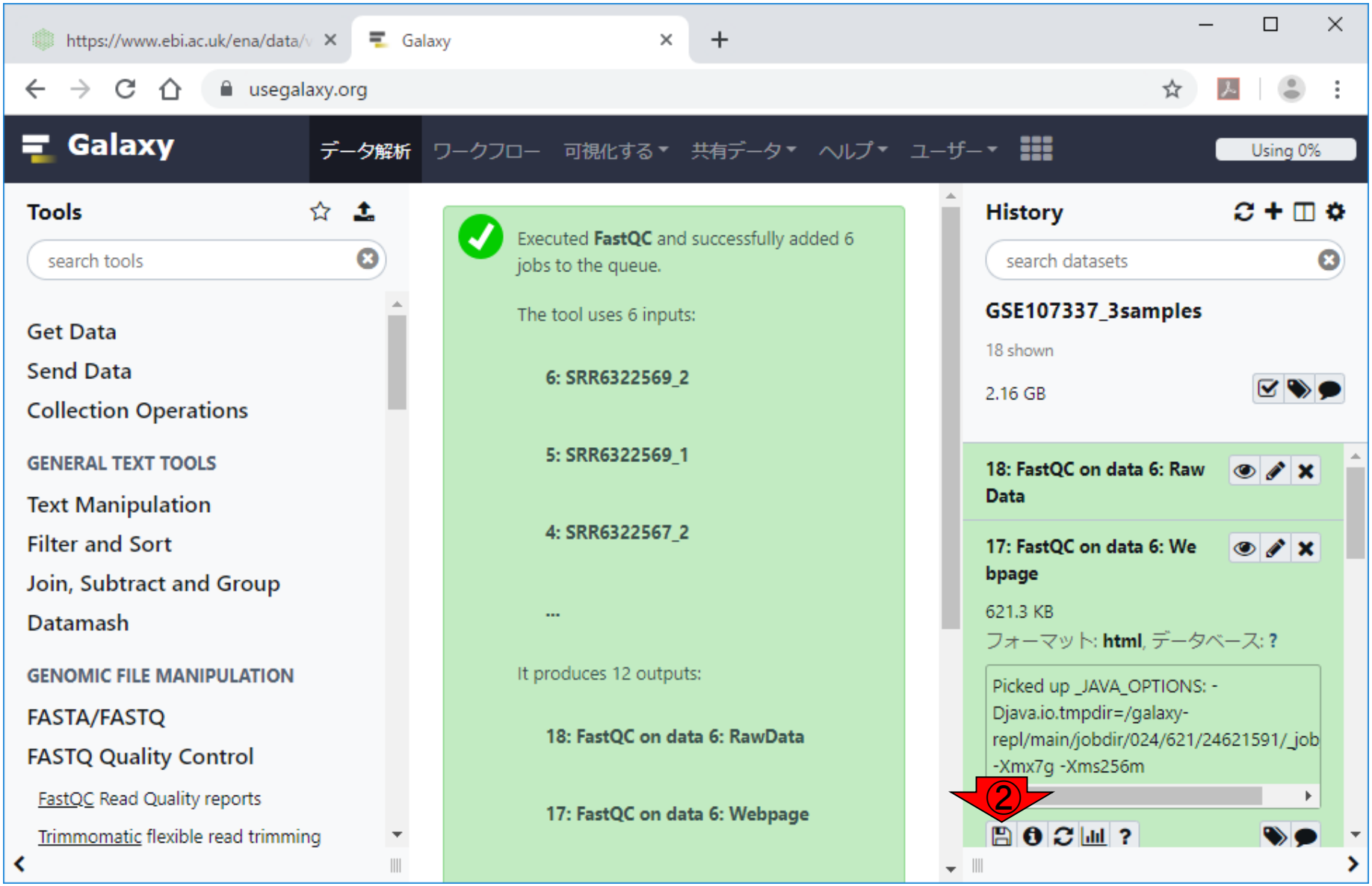
W4-7：htmlを保存

①を例として、Webpageのほうを保存。第11回W10-4と同じです。



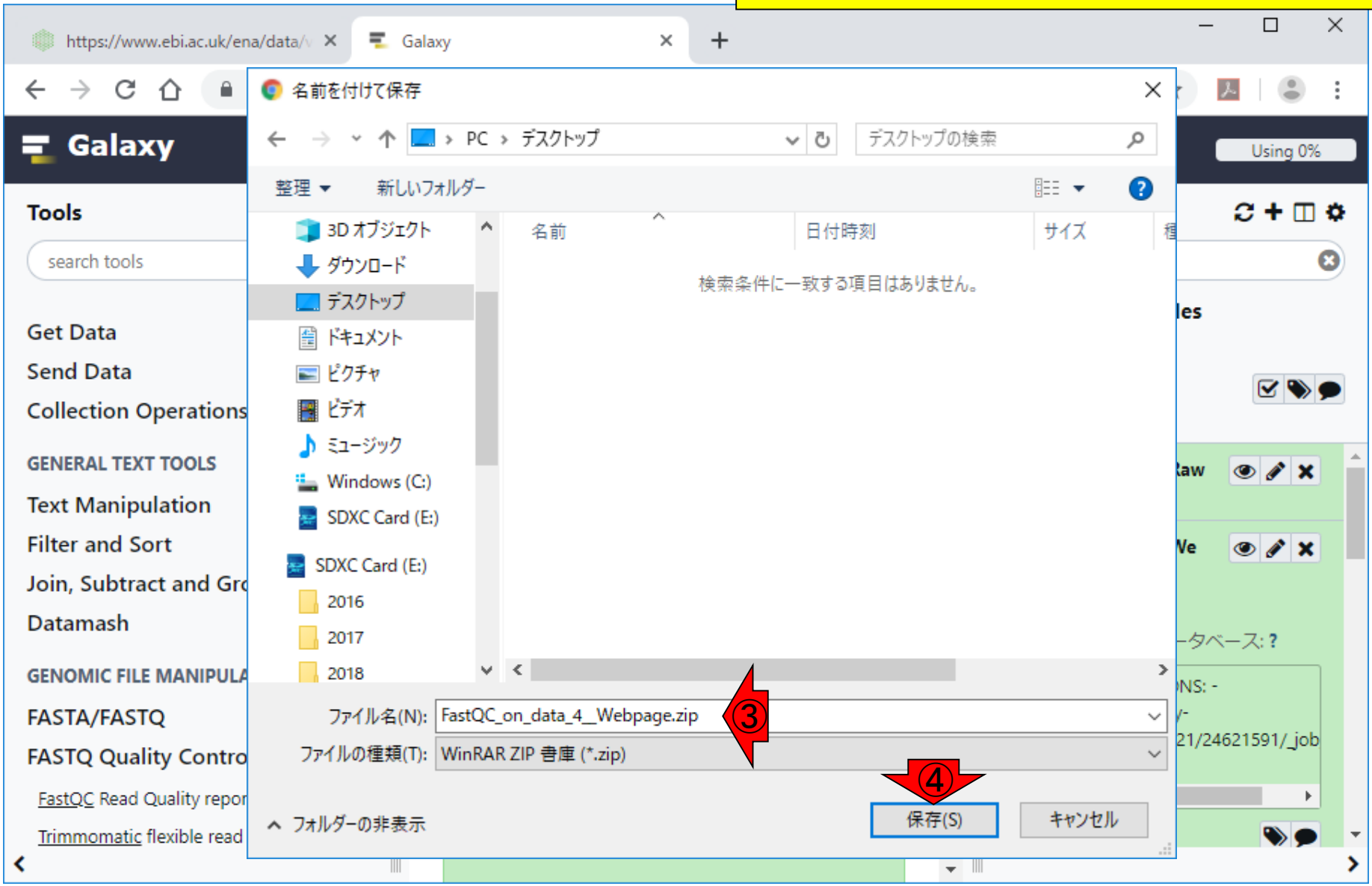
W4-7：htmlを保存

①を例として、Webpageのほうを保存。第11回W10-4と同じです。②を押す。



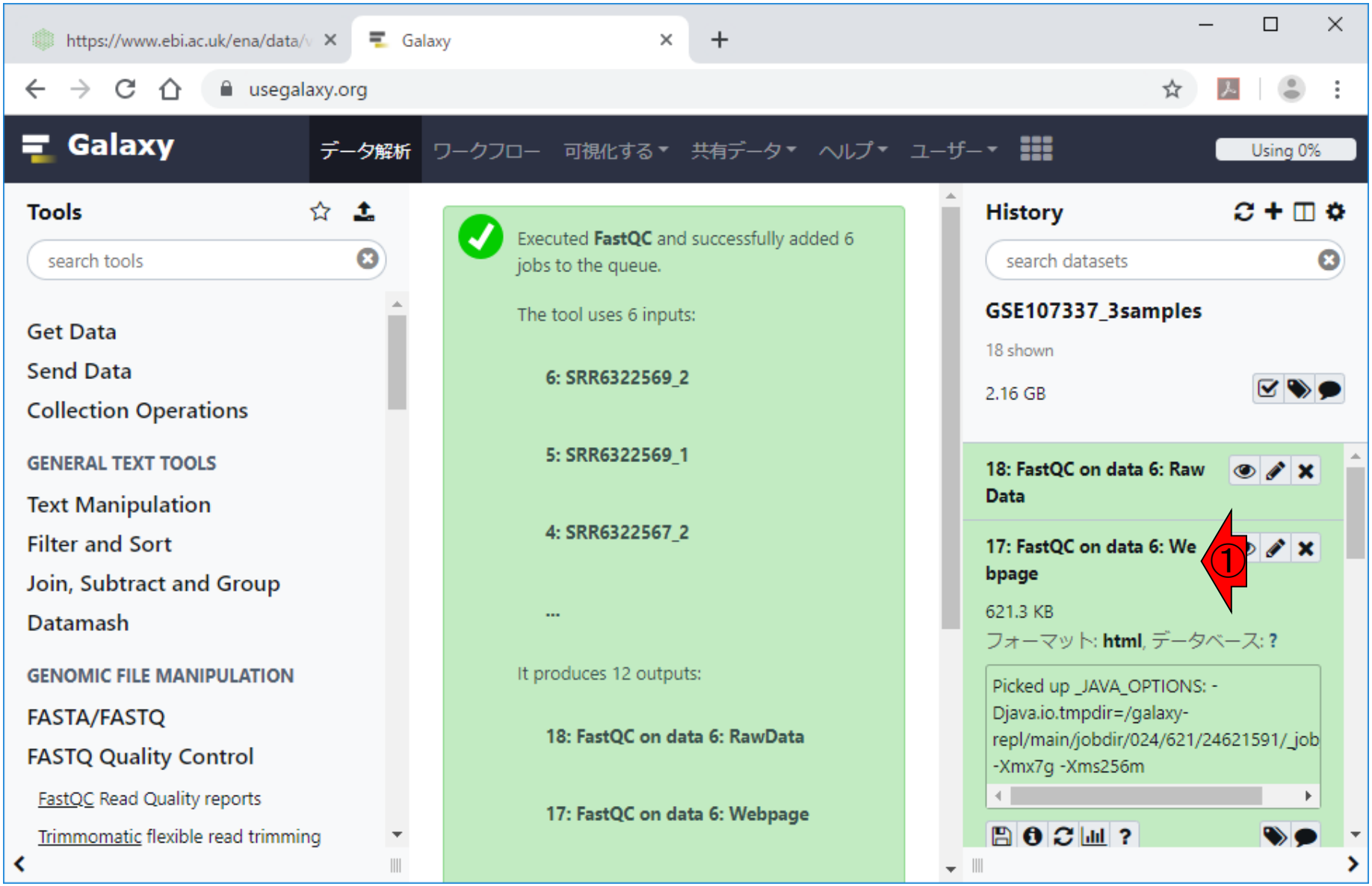
W4-7：htmlを保存

①を例として、Webpageのほうを保存。第11回W10-4と同じです。②を押す。③のような感じのzip圧縮ファイルを、④保存できます。



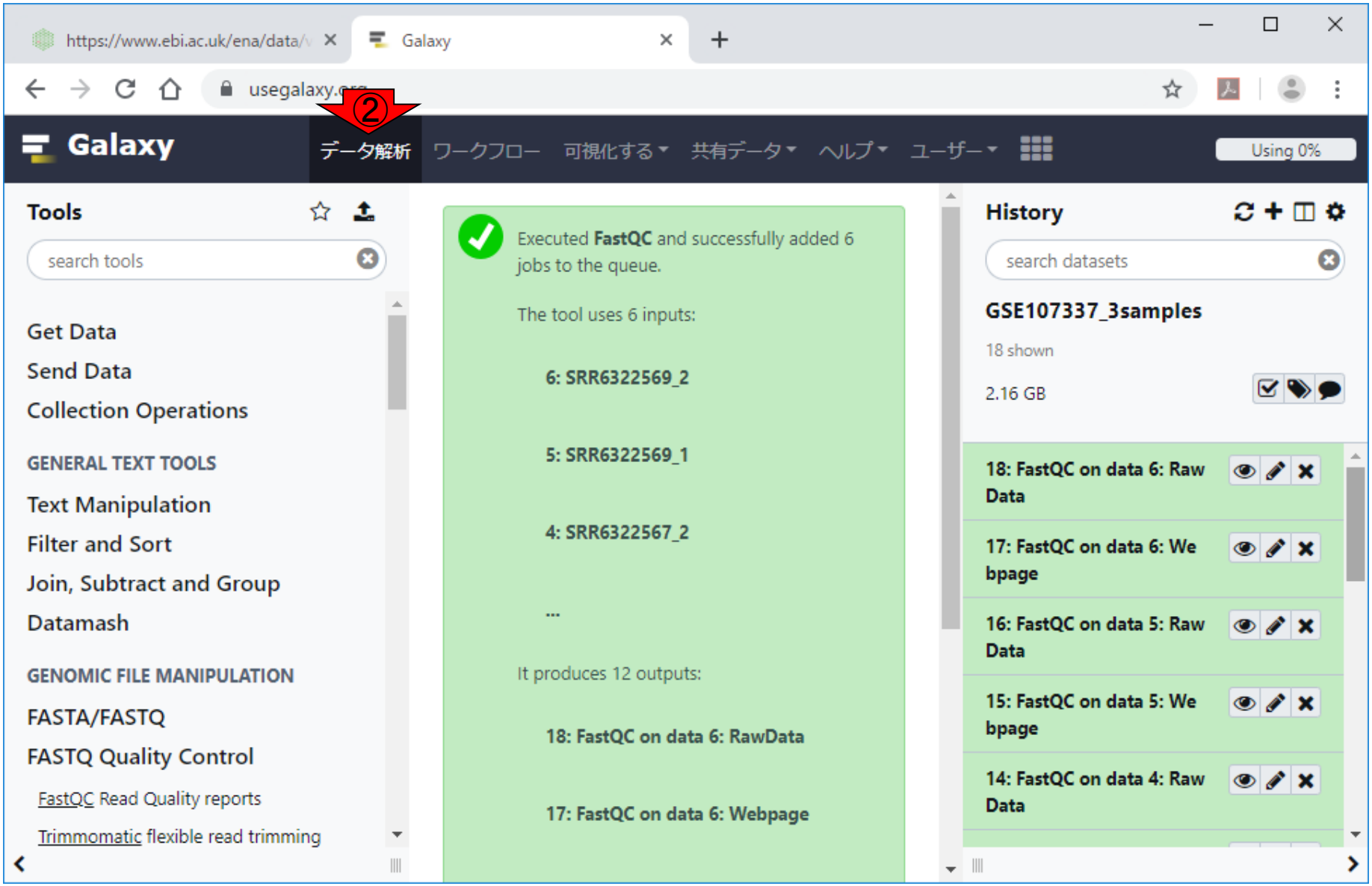
①を押す。

W4-8：元に戻す



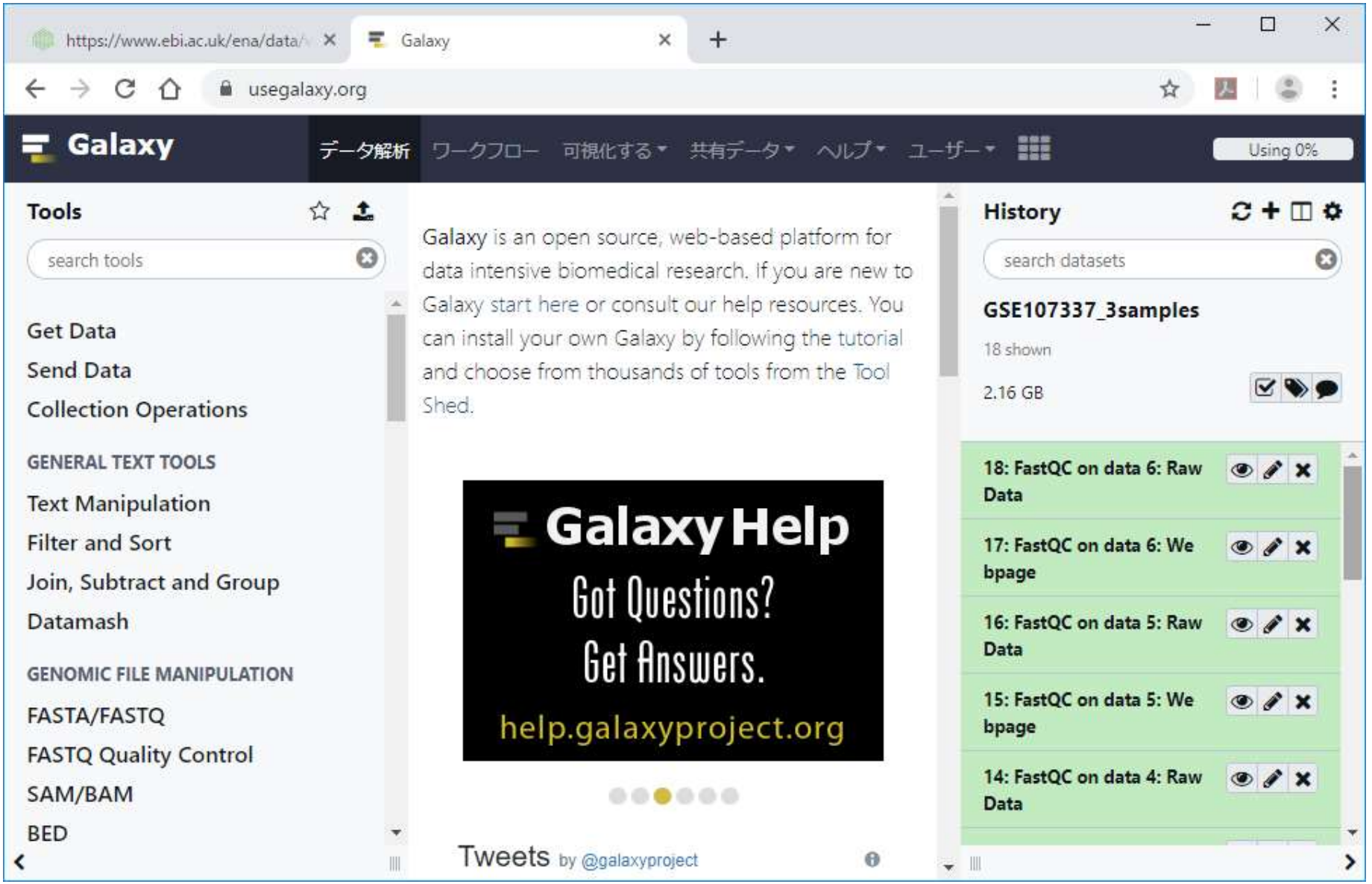
①を押す。②を押す。

W4-8：元に戻す



W4-8：元に戻す

①を押す。②を押す。このあたりはただの趣味なので馬鹿正直にトレースしなくてもよいw。

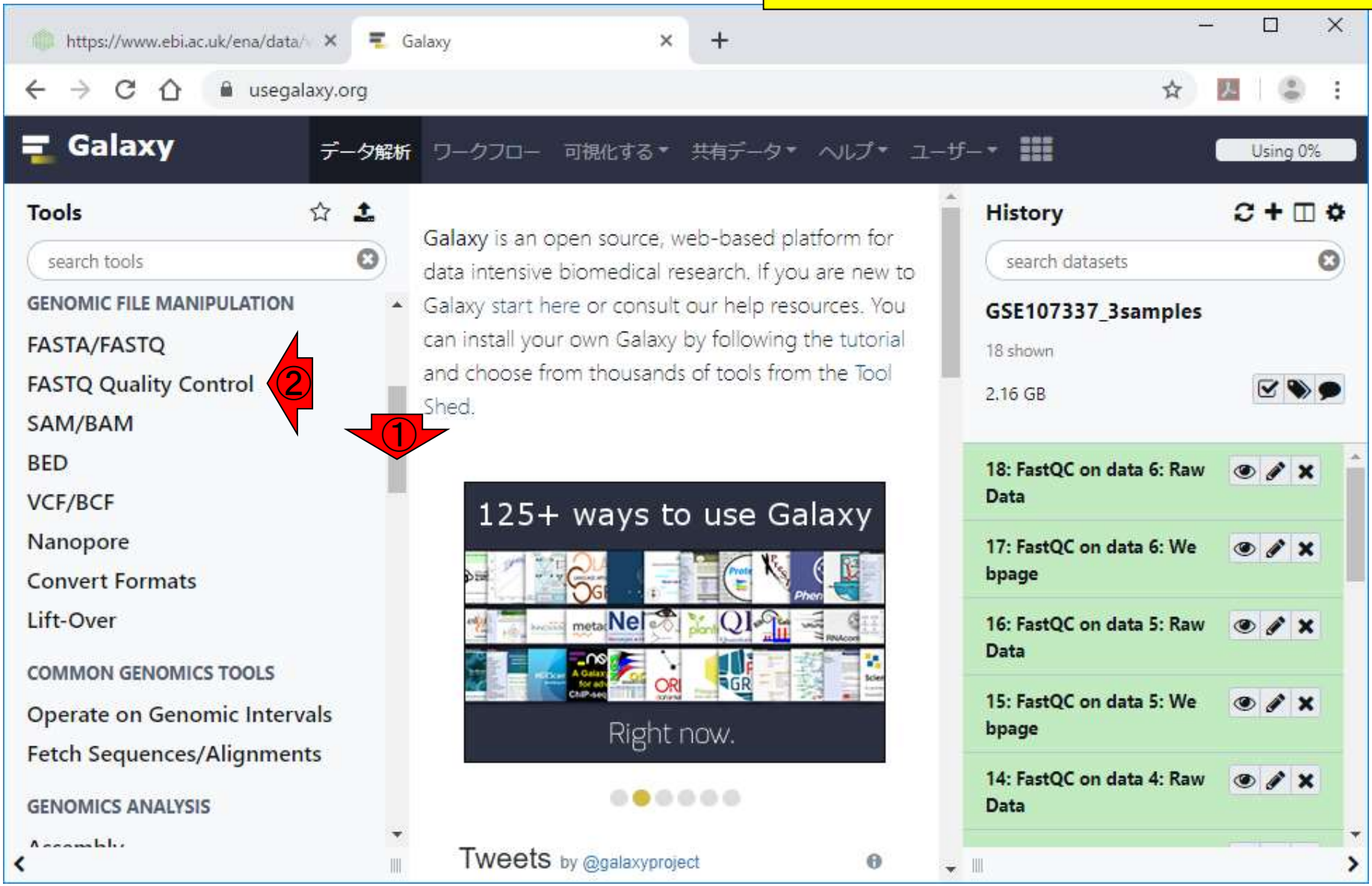


Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

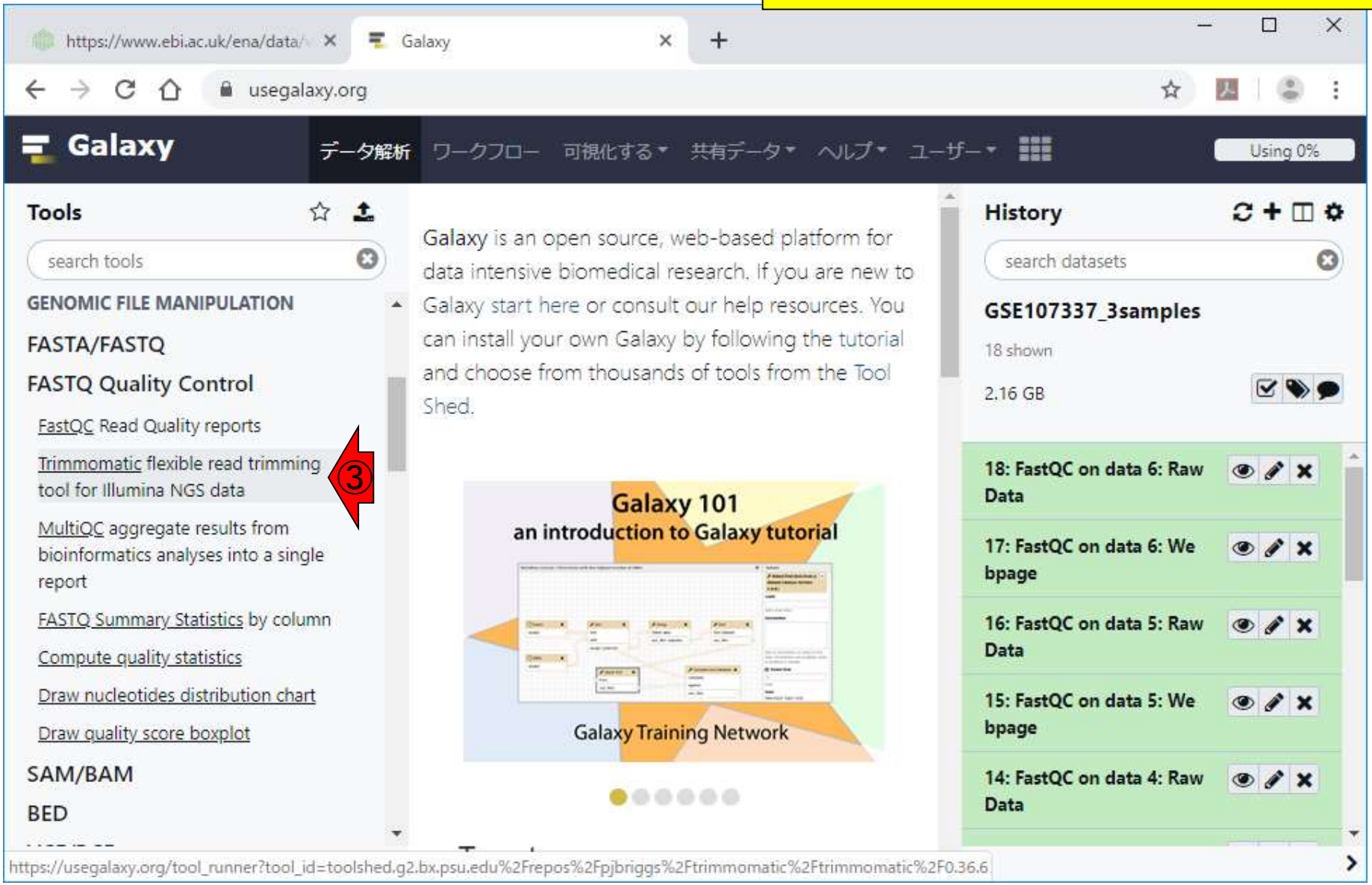
W5-1：Trimmomatic

次は、Trimmomatic。画面左側のツール選択パネル上で、①少し下部に移動したのち、②FASTQ Quality Control。



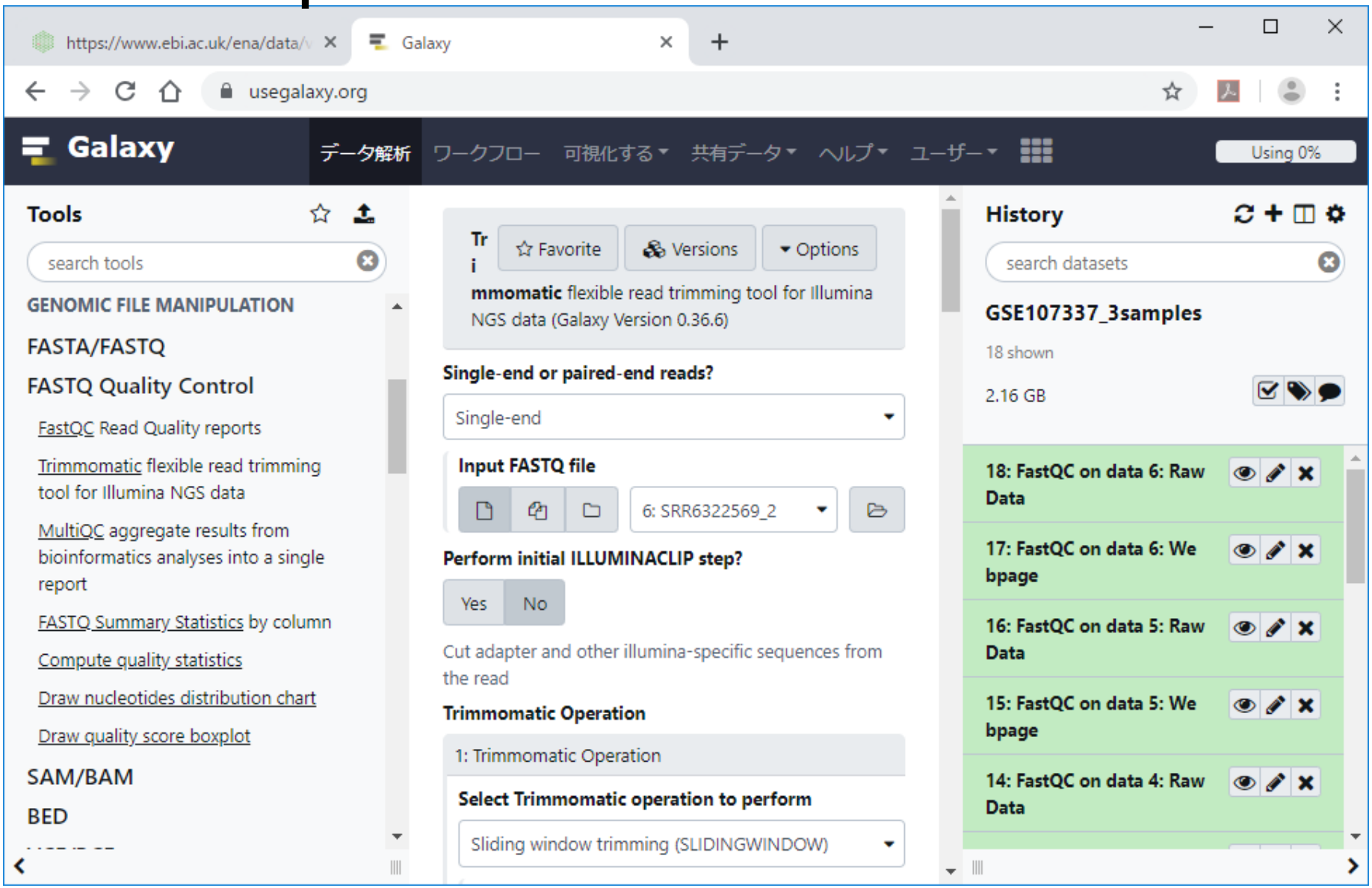
W5-1：Trimmomatic

次は、Trimmomatic。画面左側のツール選択パネル上で、①少し下部に移動したのち、②FASTQ Quality Control。③Trimmomaticを選択。



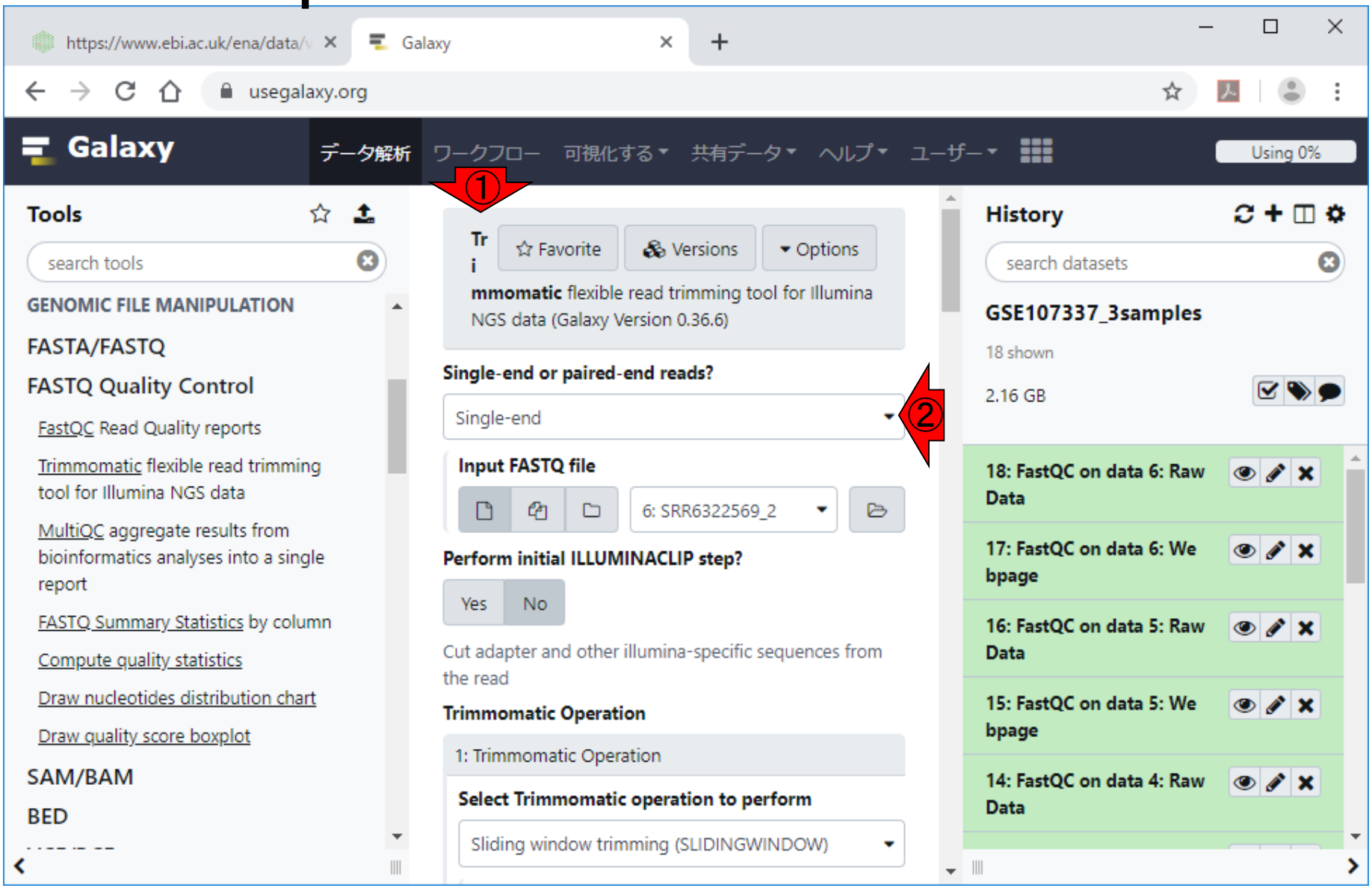
こんな感じになる。

W5-2：paired-end



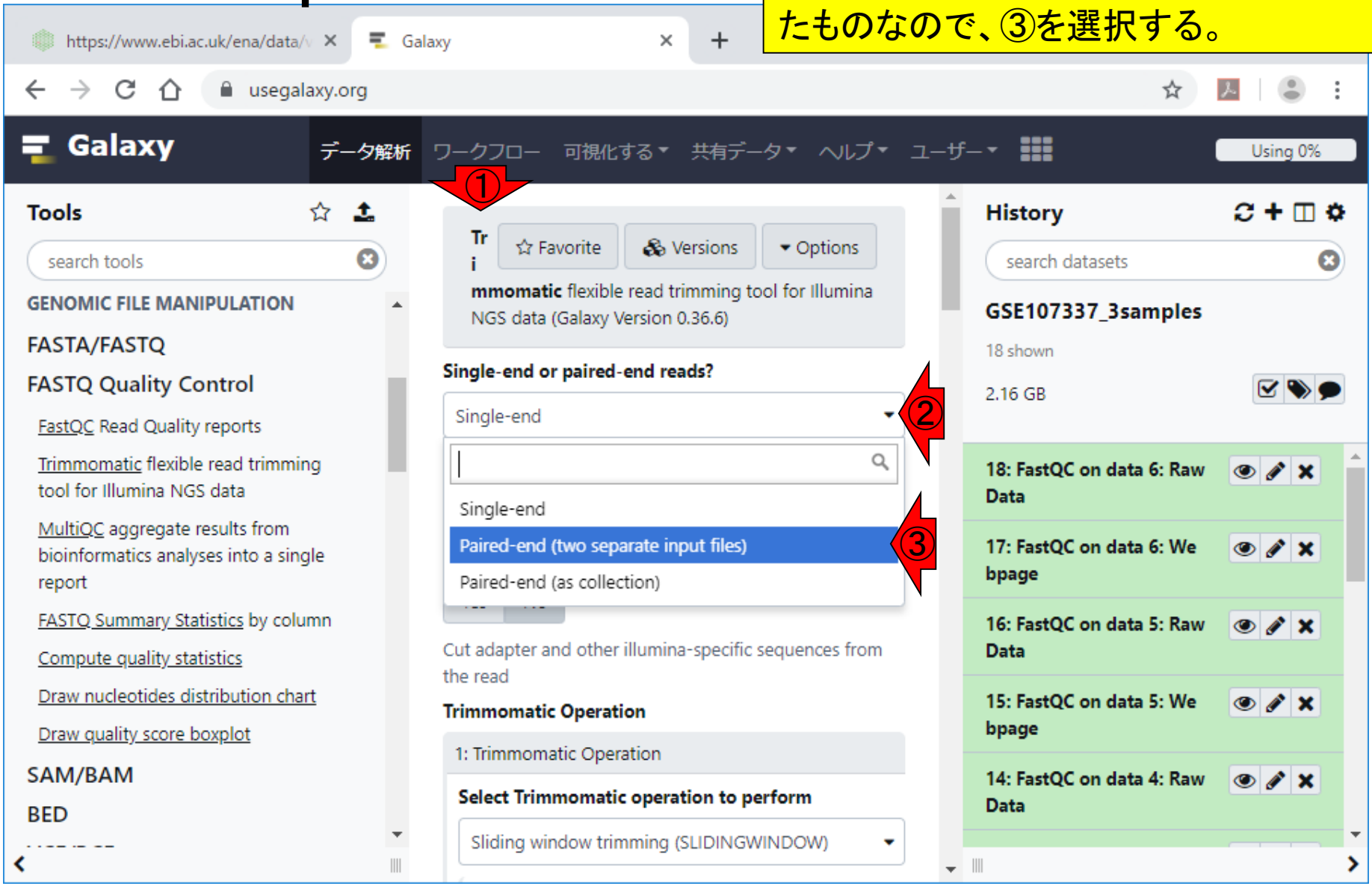
W5-2：paired-end

わかりづらいが、①Trimmomaticと書いています。
Paired-endなので、②のところを変更する。



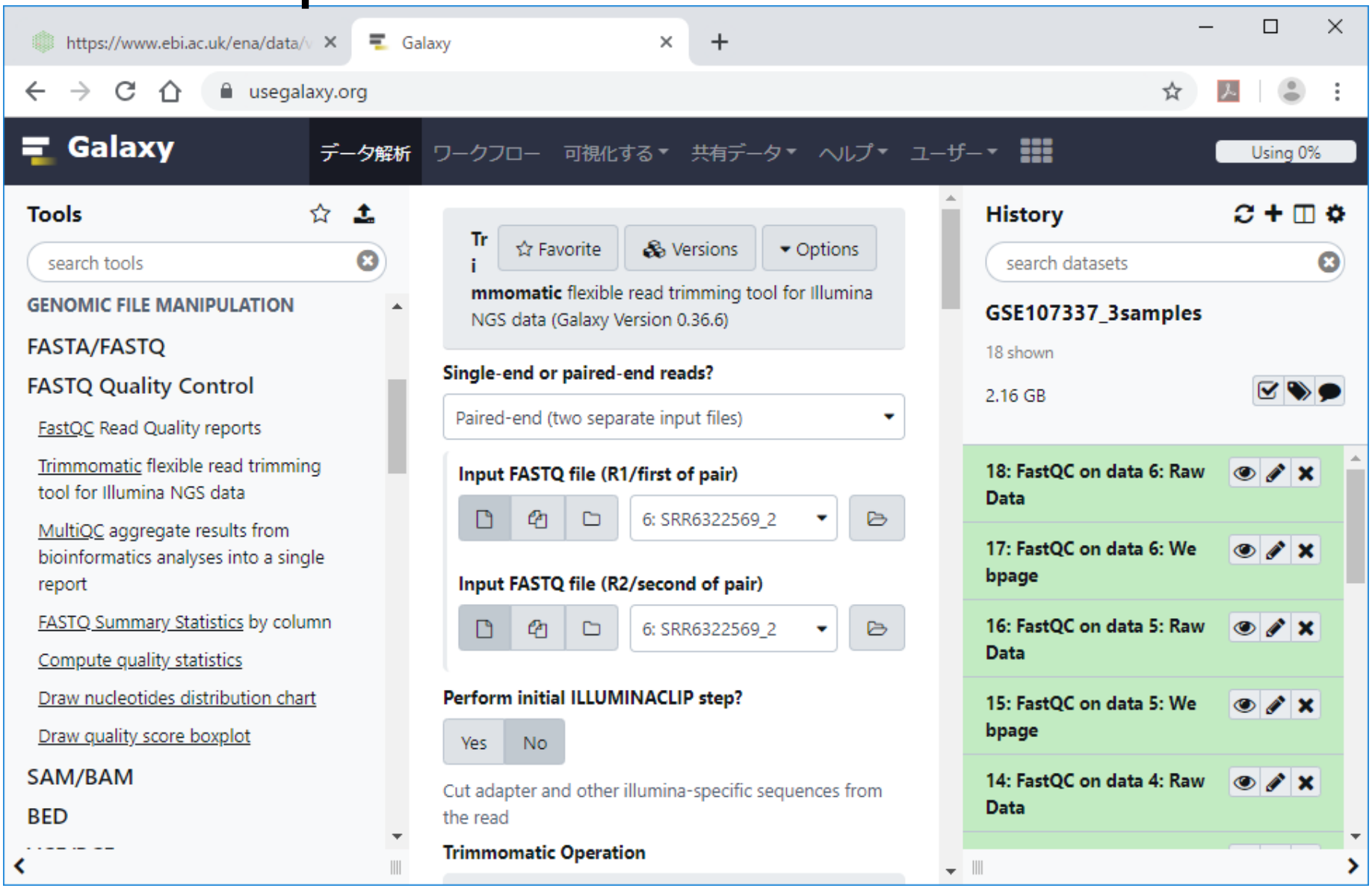
W5-2：paired-end

わかりづらいが、①Trimmomaticと書いています。Paired-endなので、②のところを変更する。今取り扱っているのはforward側とreverse側に2分割されたものなので、③を選択する。



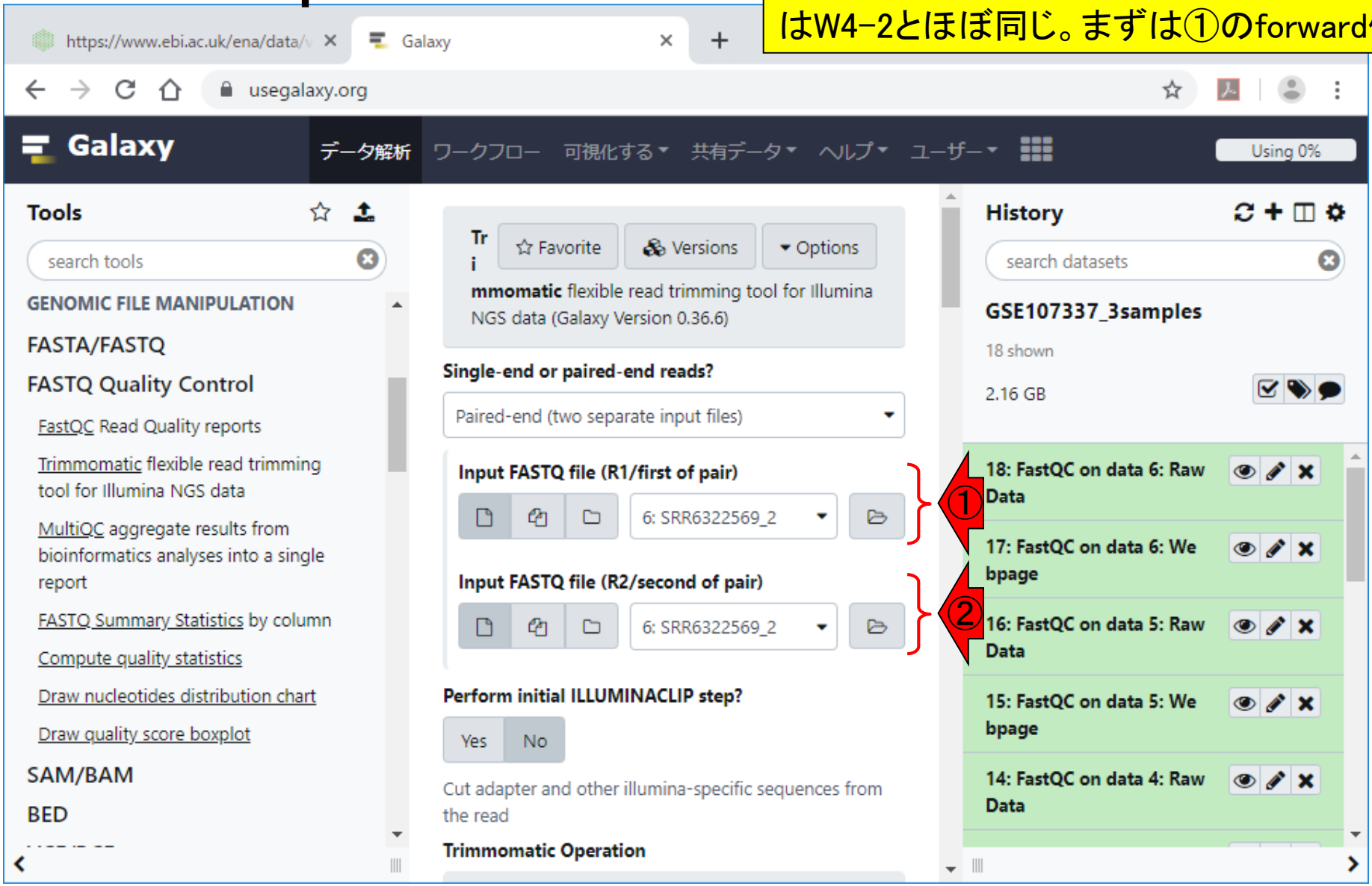
こんな感じになる。

W5-2：paired-end



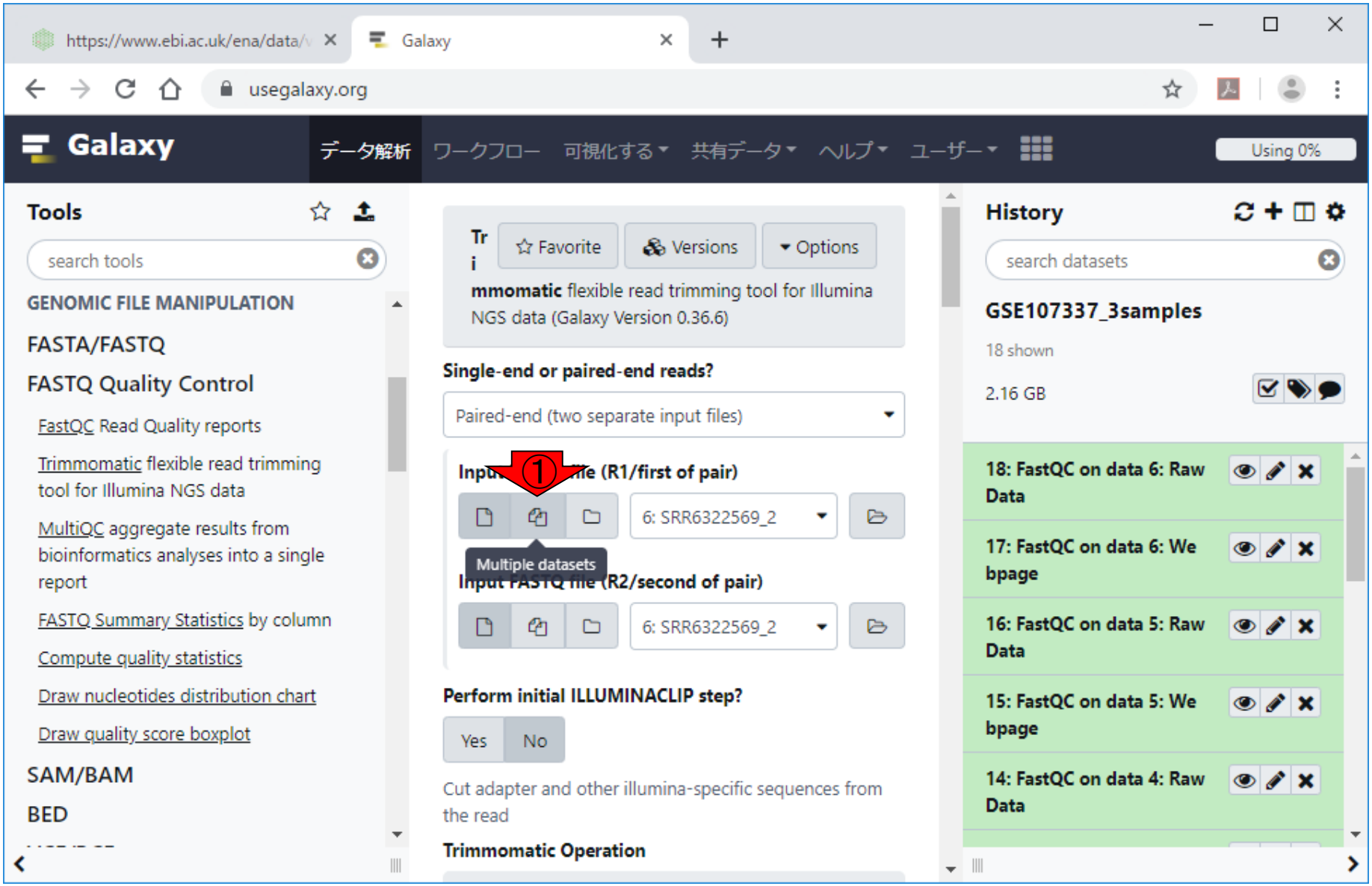
W5-2：paired-end

こんな感じになる。①の部分でforward側（R1側）のファイル群を、そして②の部分でreverse側（R2側）のファイル群をそれぞれ指定する。作業のノリはW4-2とほぼ同じ。まずは①のforward側から。



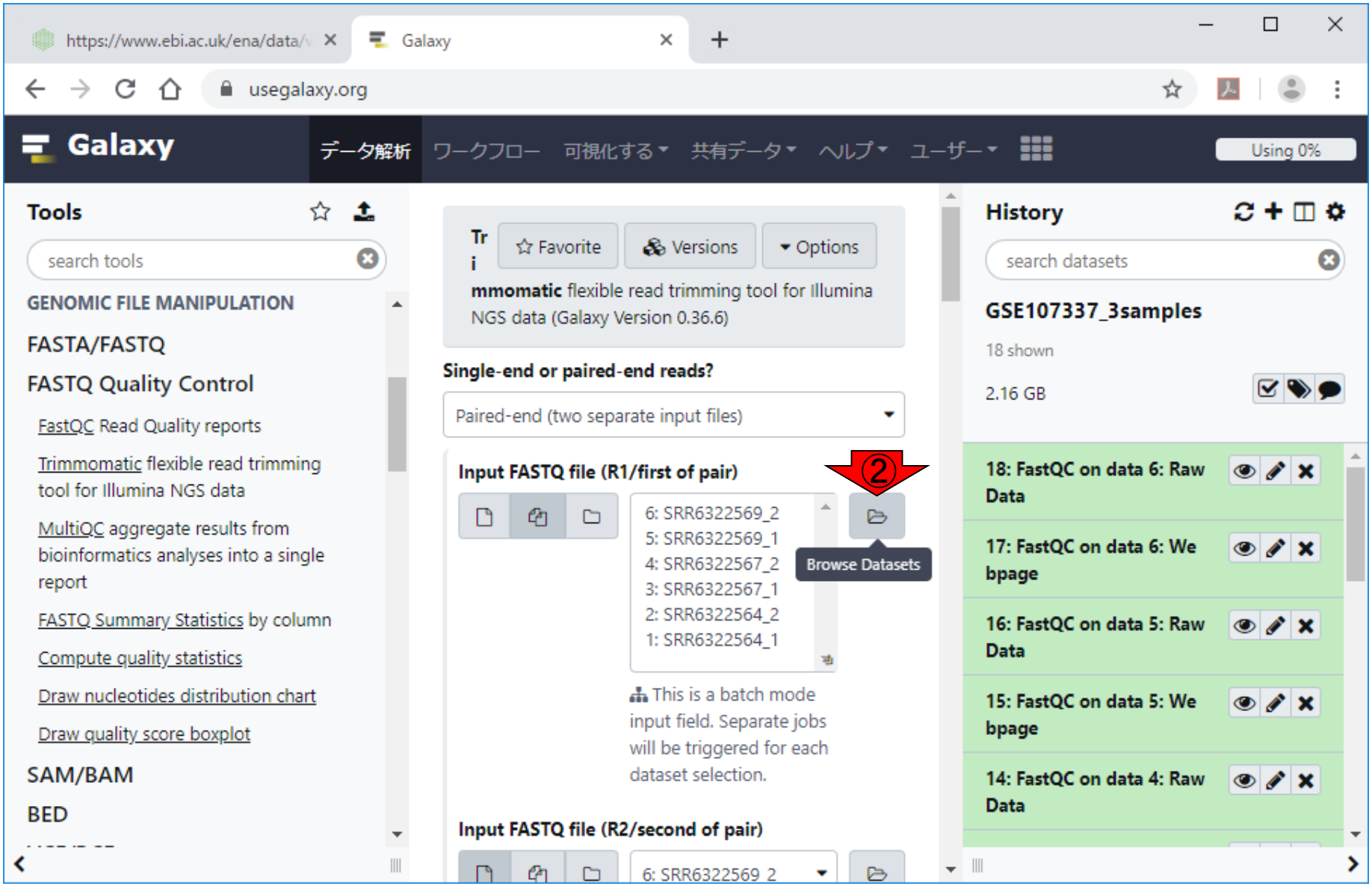
①Multiple datasets。

W5-3：forward側



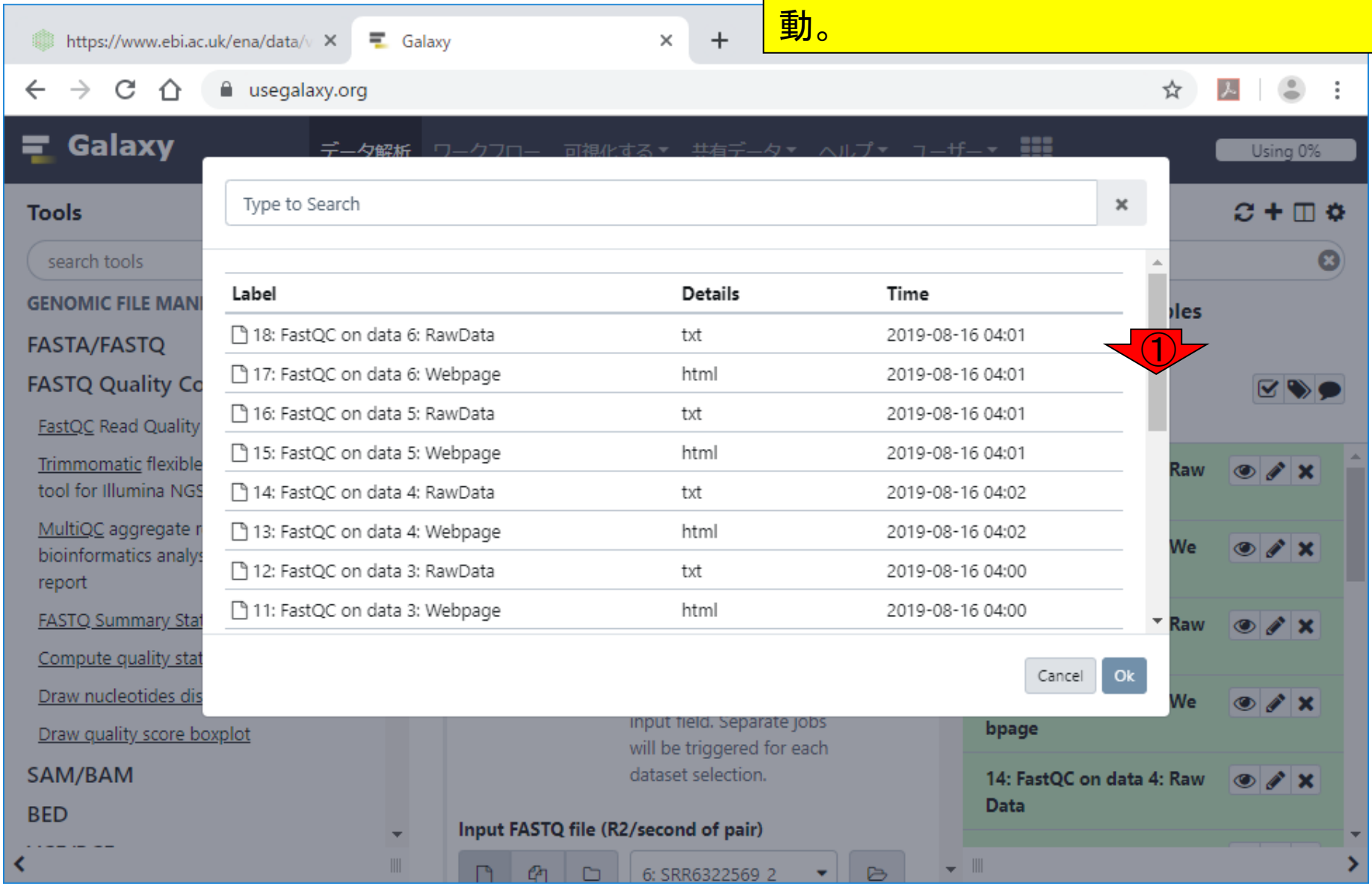
①Multiple datasets。②Browse Datasets。

W5-3：forward側



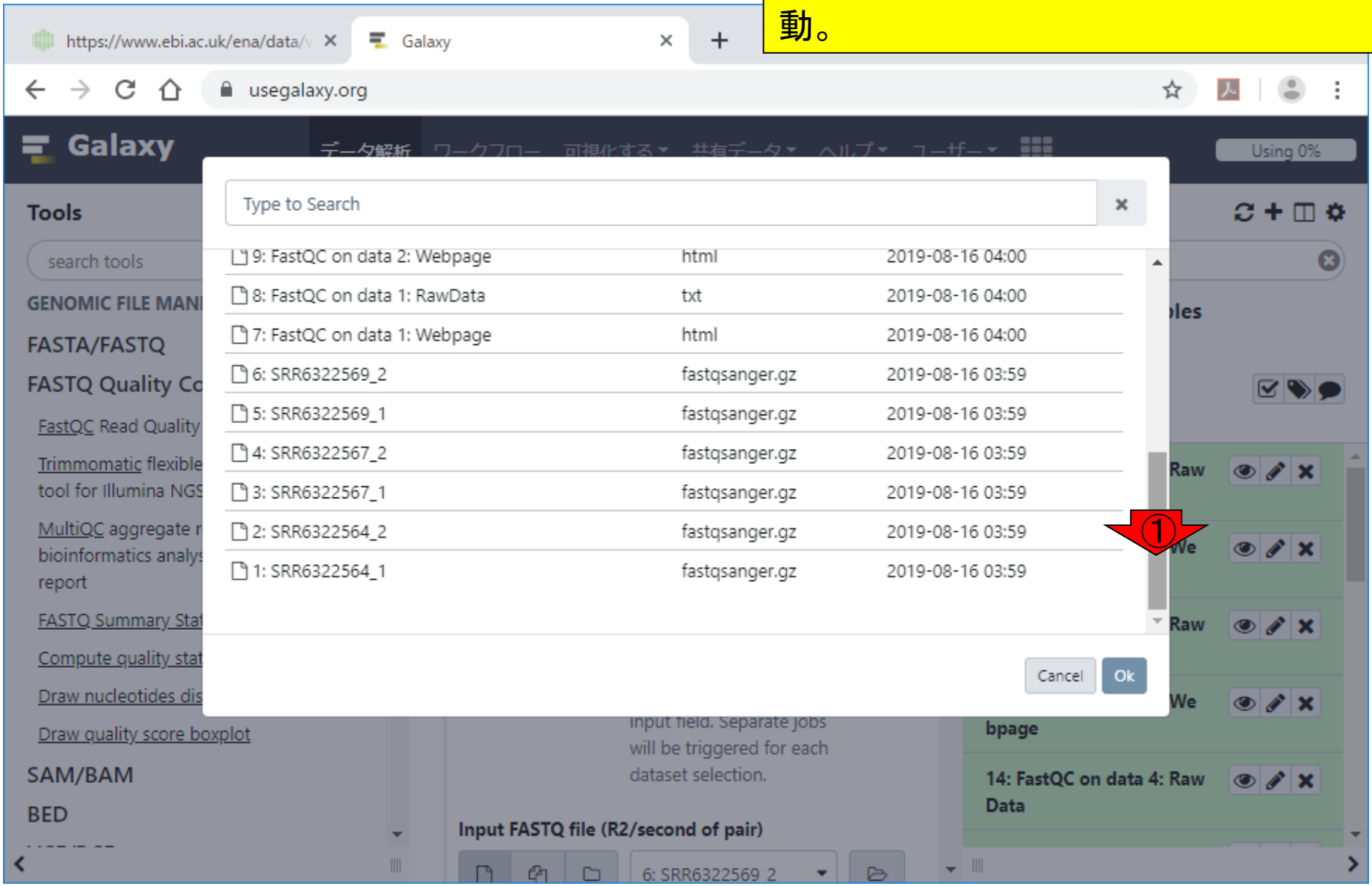
W5-3：forward側

こんな感じになる。ヒストリーの最新のものから表示されるようですね。入力として用いたいのはforward側のFASTQファイルなので、①下部に移動。



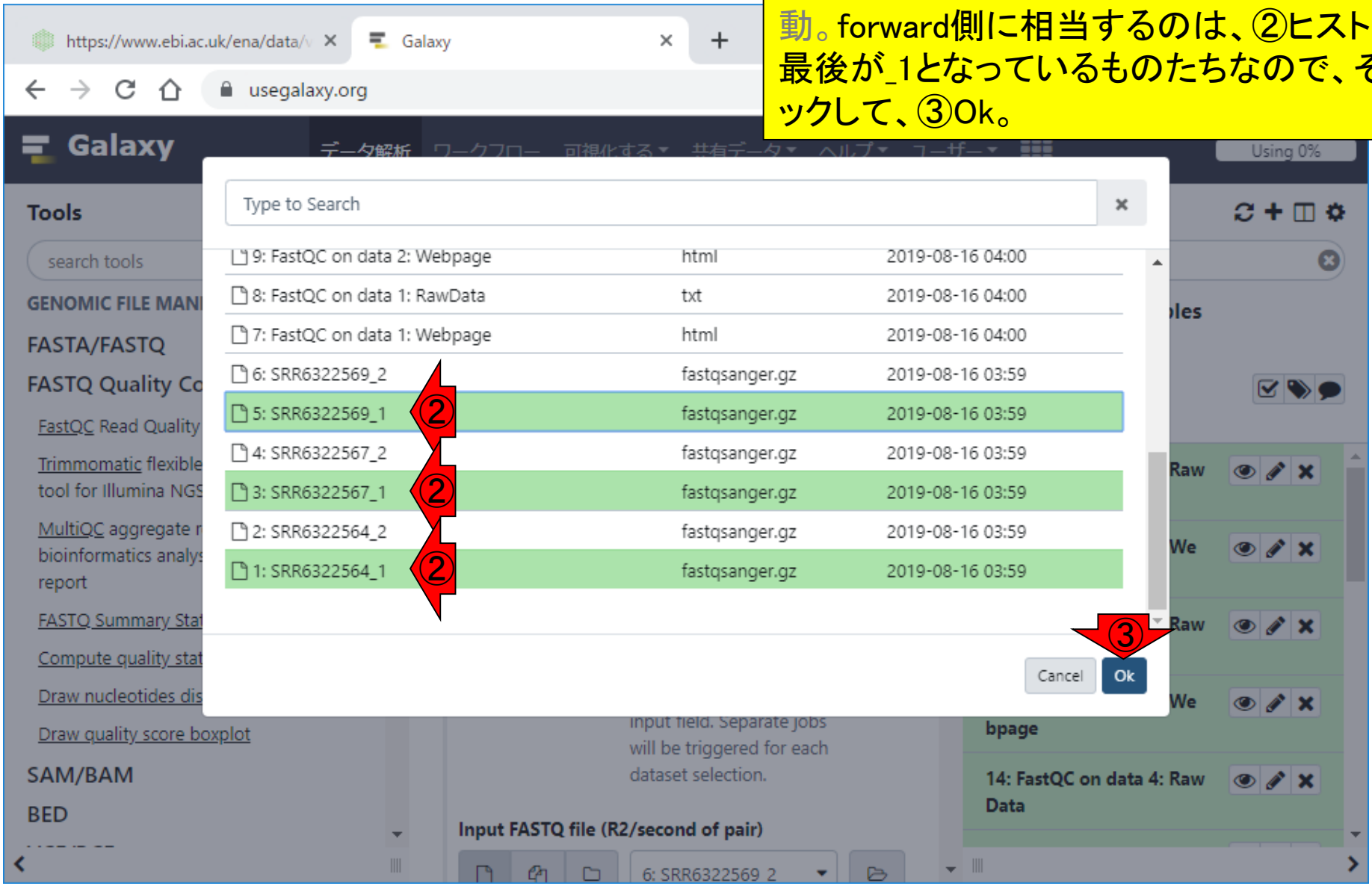
W5-3：forward側

こんな感じになる。ヒストリーの最新のものから表示されるようですね。入力として用いたいのはforward側のFASTQファイルなので、①下部に移動。



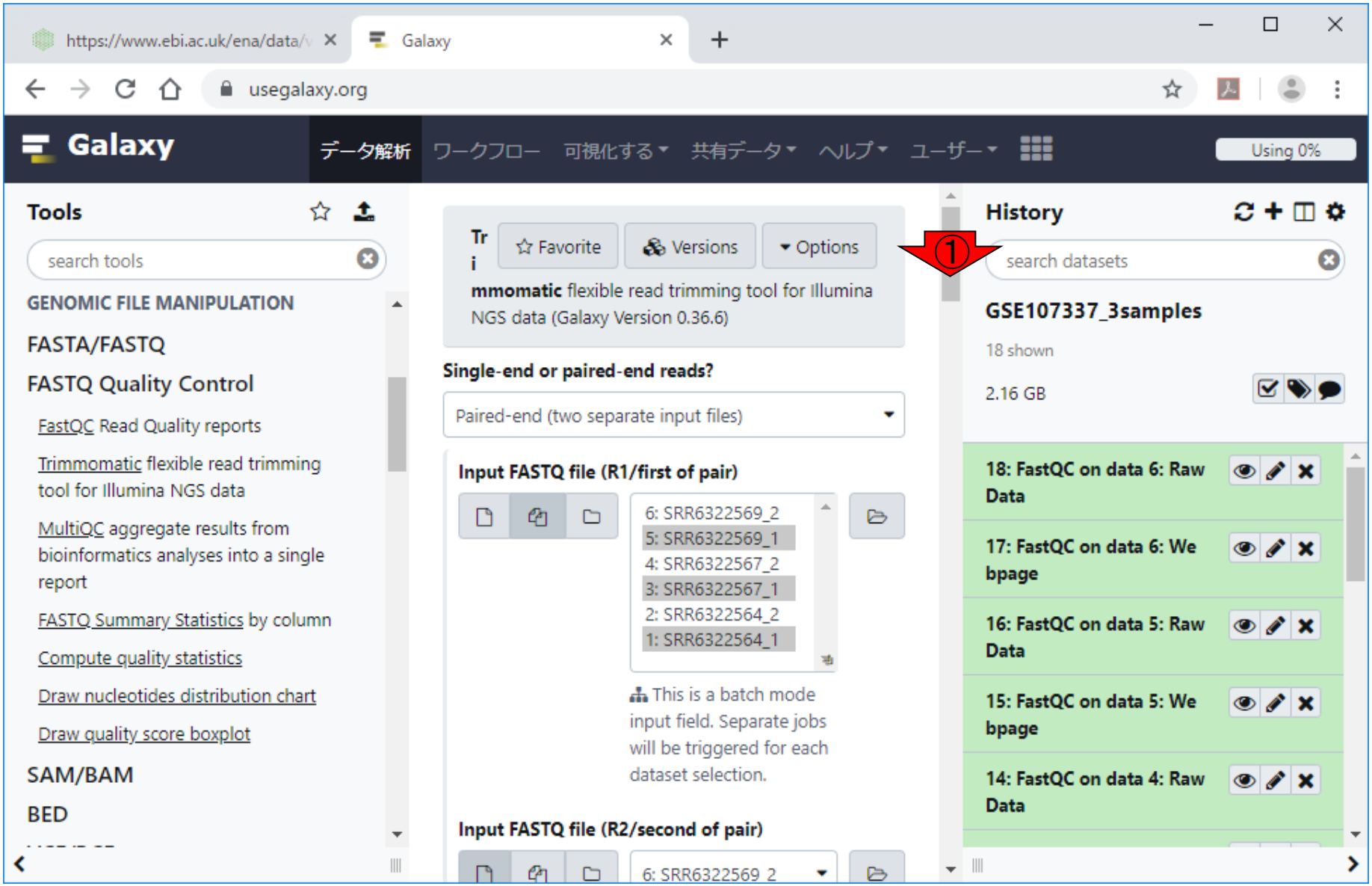
W5-3：forward側

こんな感じになる。ヒストリーの最新のものから表示されるようですね。入力として用いたいのはforward側のFASTQファイルなので、①下部に移動。forward側に相当するのは、②ヒストリー名の最後が_1となっているものたちなので、それをクリックして、③Ok。



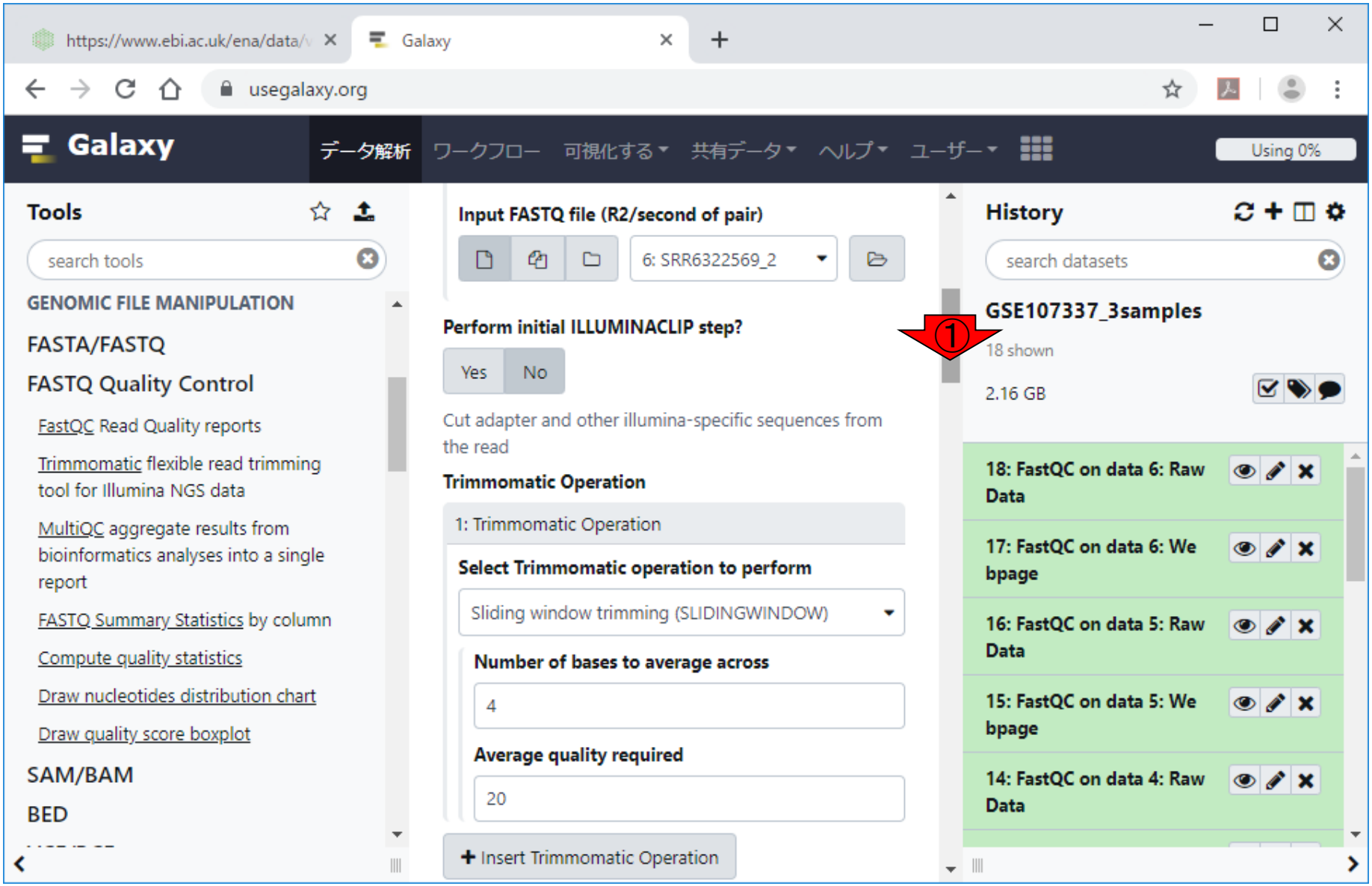
W5-3：forward側

こんな感じになる。次はreverse側なので、①中央パネルを少し下部に移動。



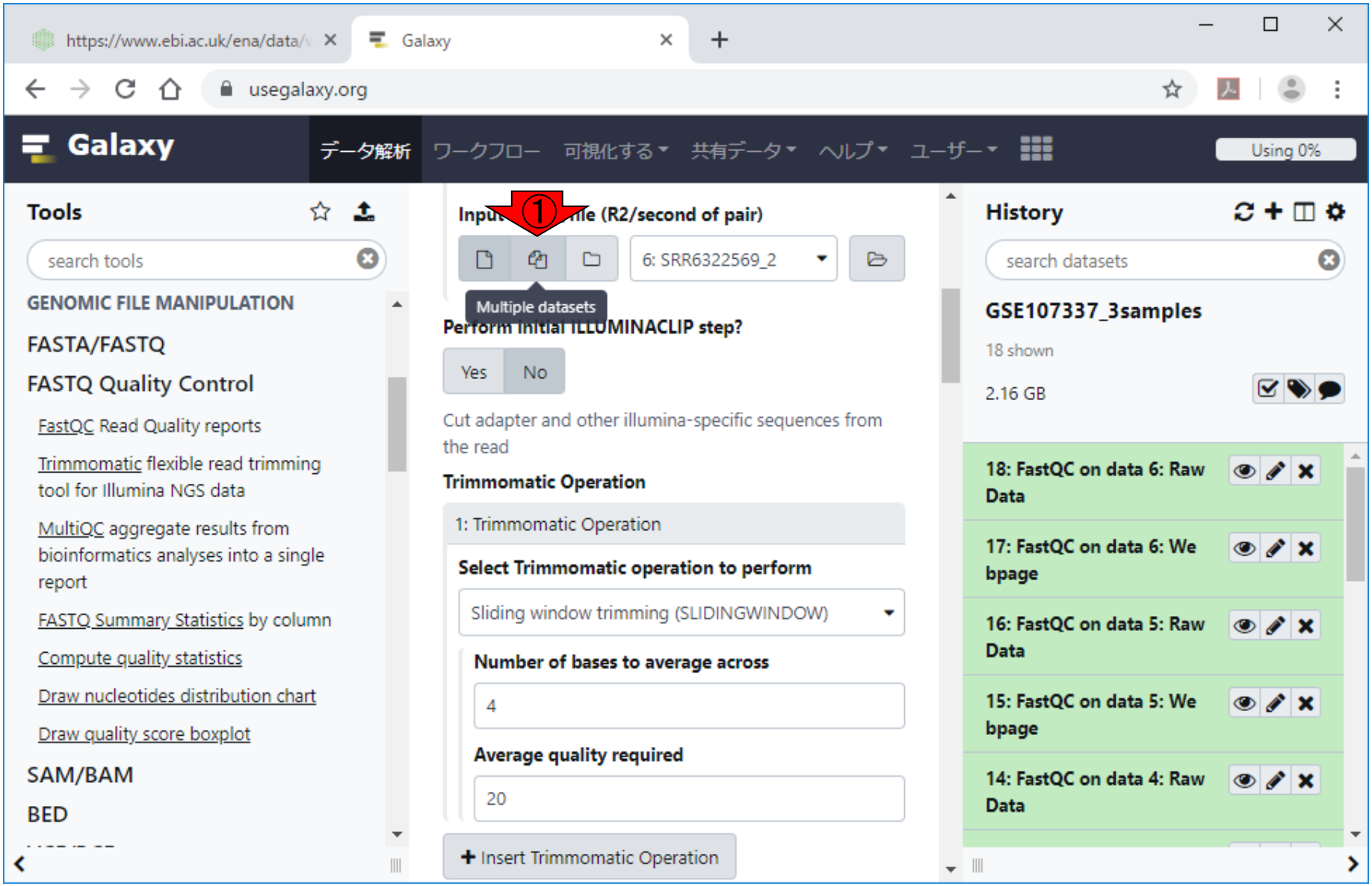
W5-4：reverse側

こんな感じになる。次はreverse側なので、①中央パネルを少し下部に移動。こんな感じ。



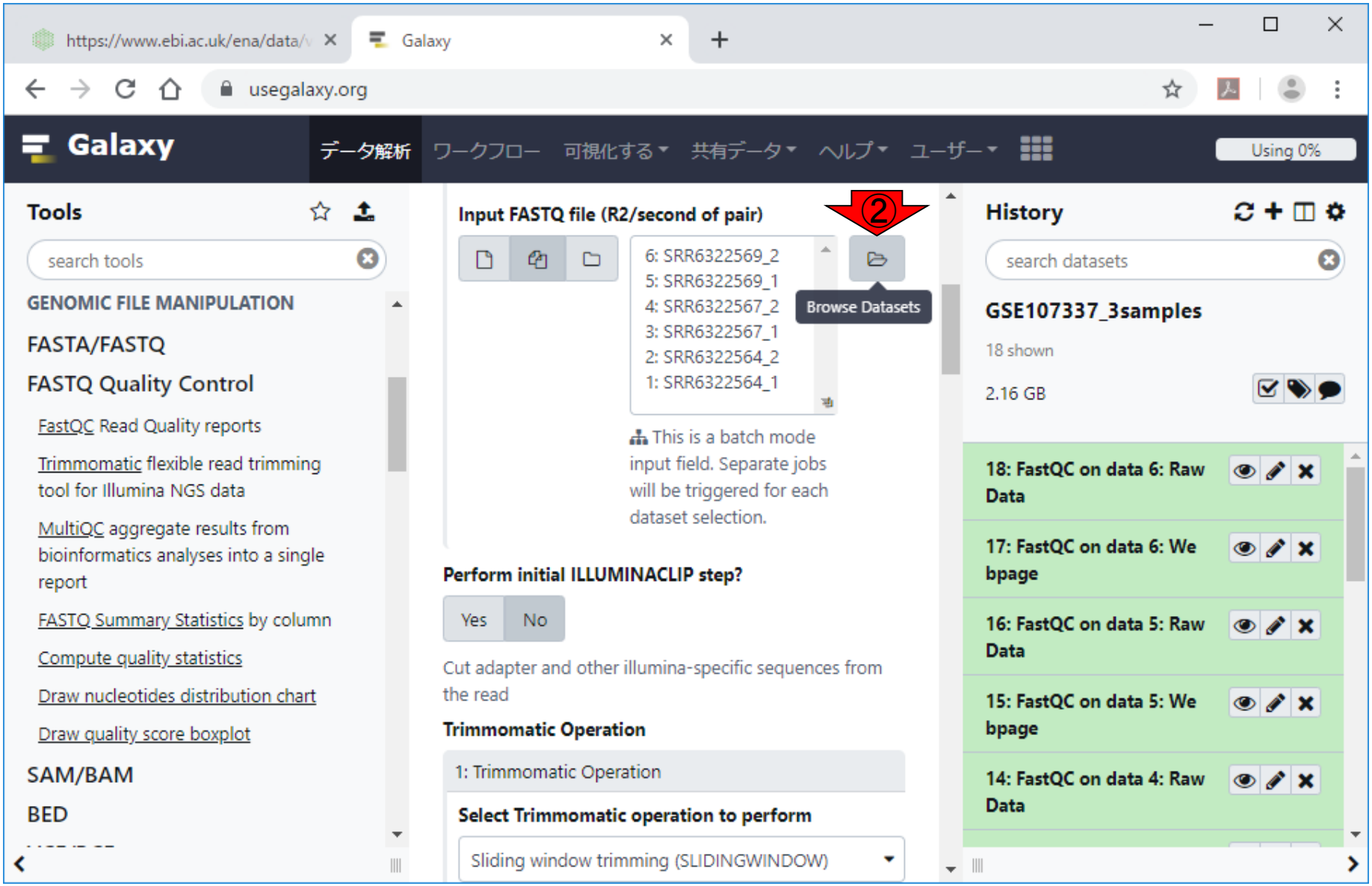
①Multiple datasets。

W5-4：reverse側



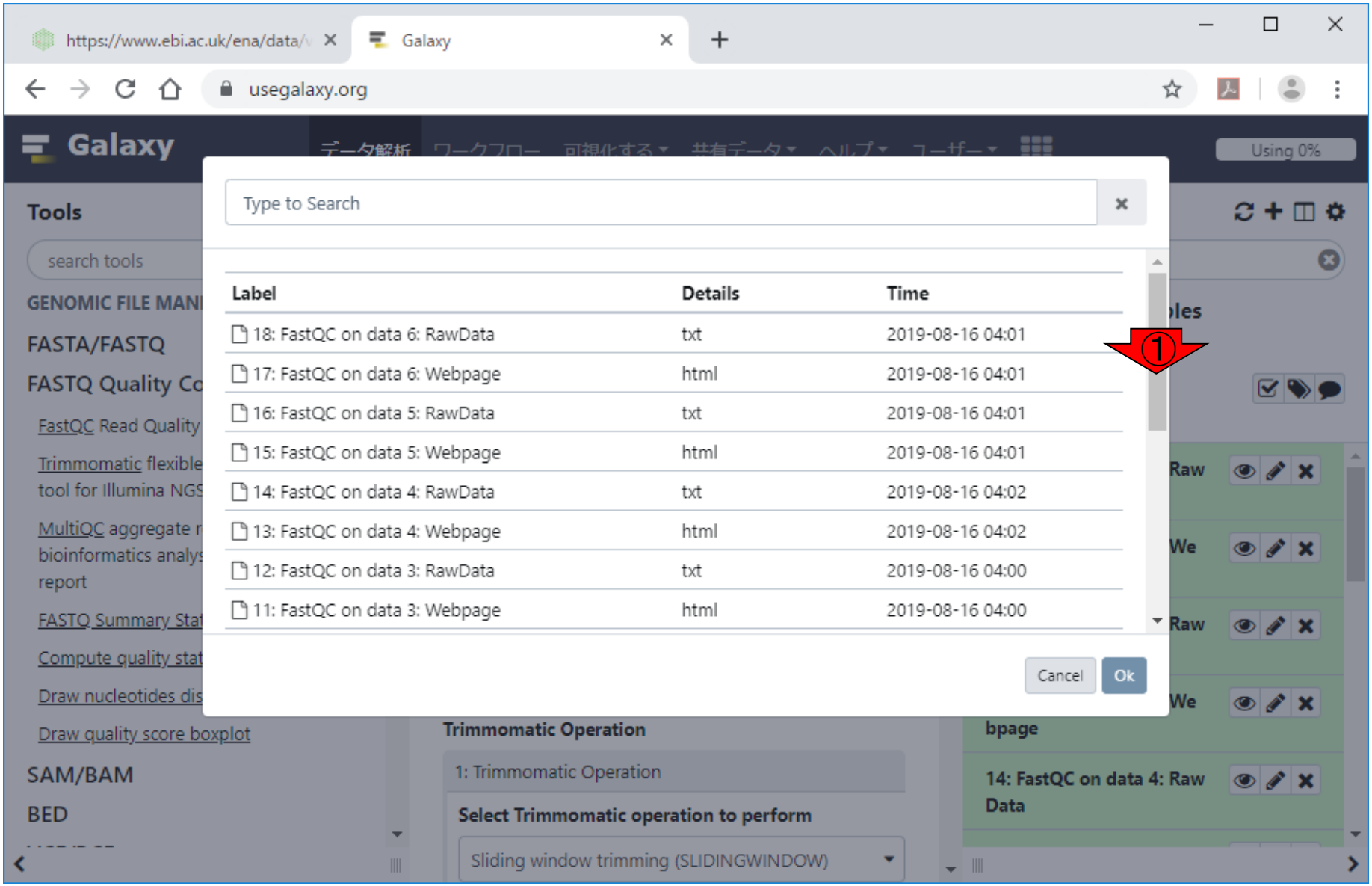
①Multiple datasets。②Browse Datasets。

W5-4：reverse側



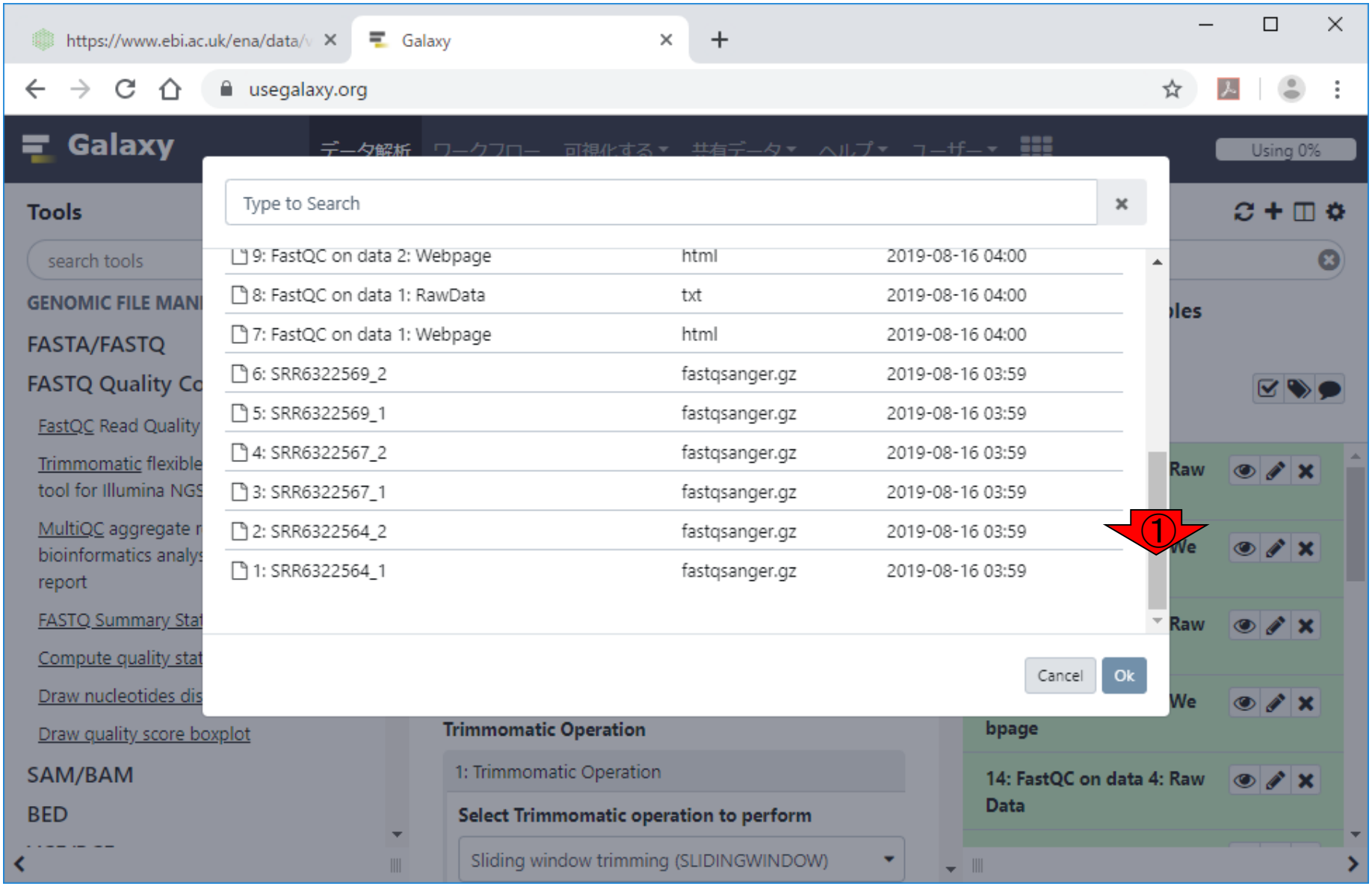
W5-4：reverse側

入力として用いたいのはreverse側のFASTQファイルなので、①下部に移動。



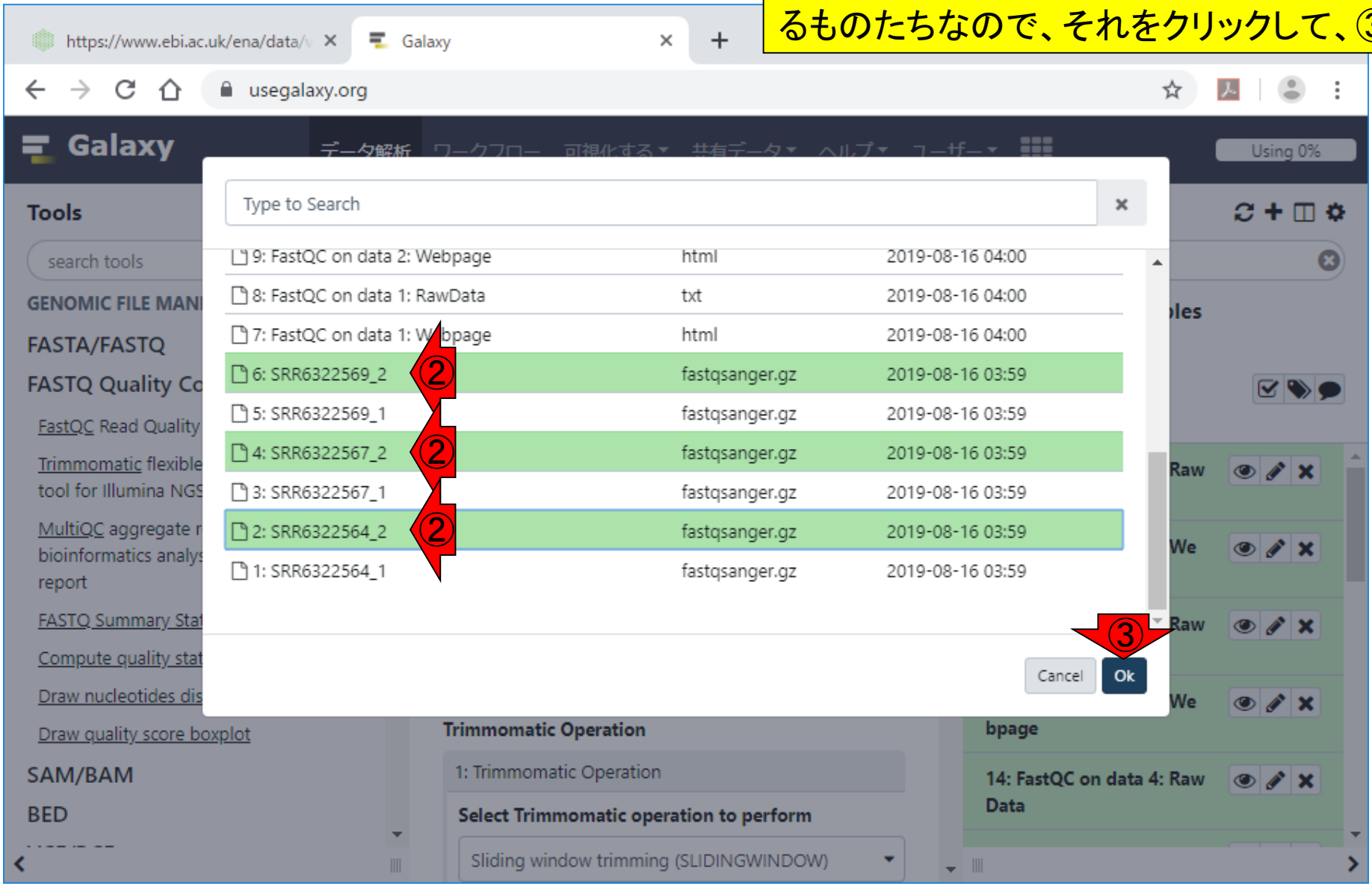
W5-4：reverse側

入力として用いたいのはreverse側のFASTQファイルなので、①下部に移動。移動後。



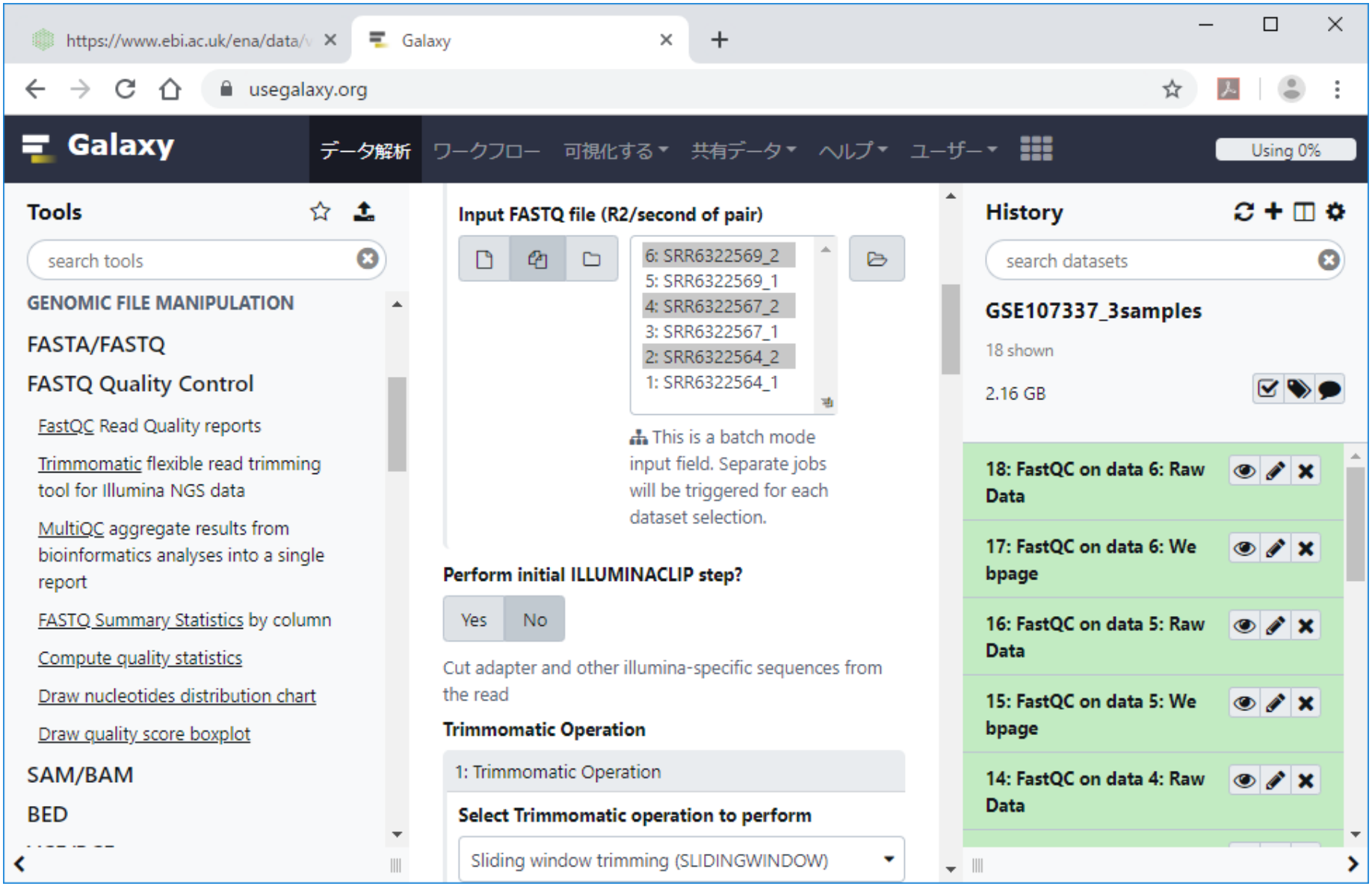
W5-4：reverse側

入力として用いたいのはreverse側のFASTQファイルなので、①下部に移動。移動後。reverse側に相当するのは、②ヒストリー名の最後が_2となっているものたちなので、それをクリックして、③Ok。



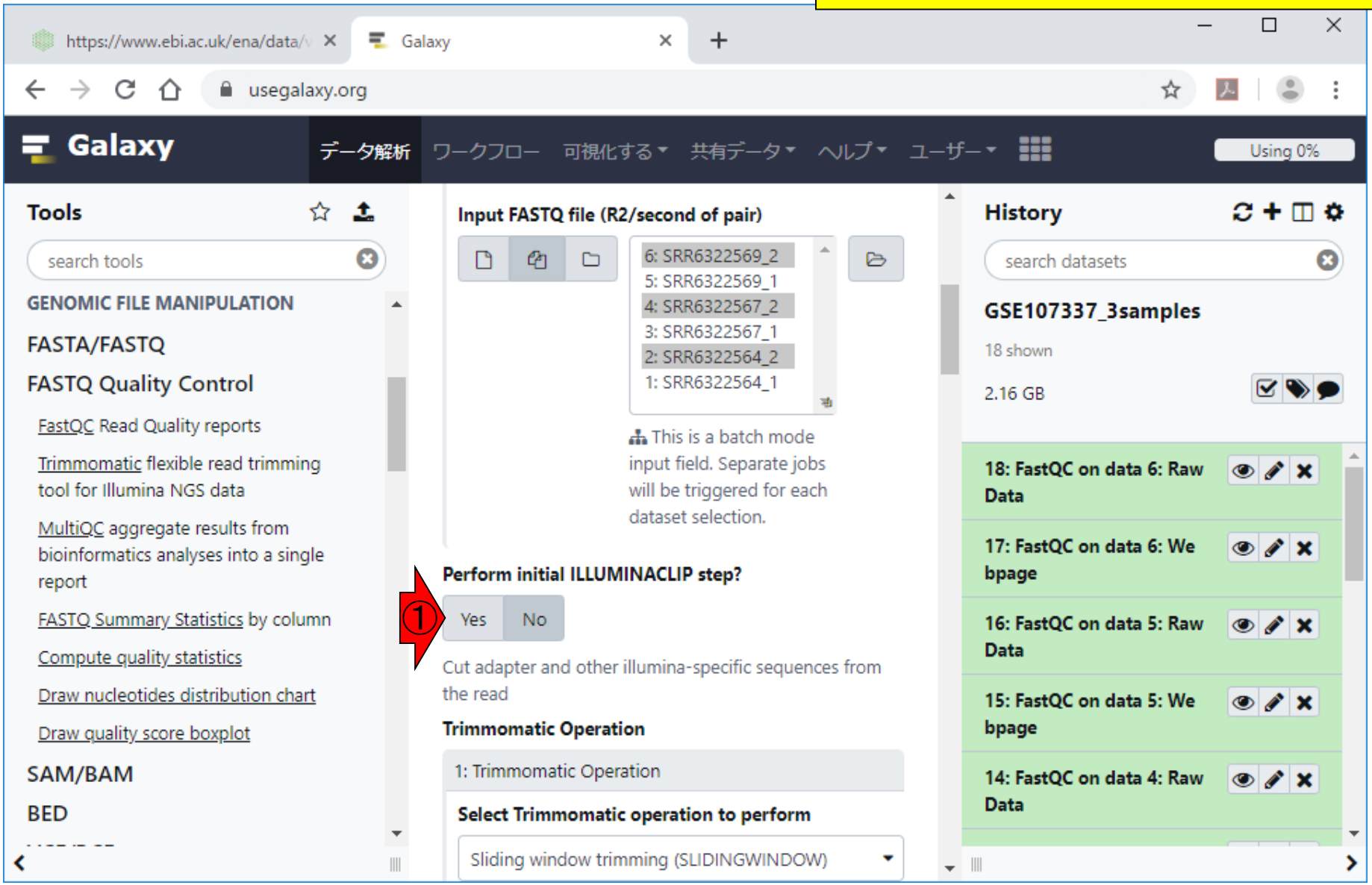
こんな感じになる。

W5-4：reverse側



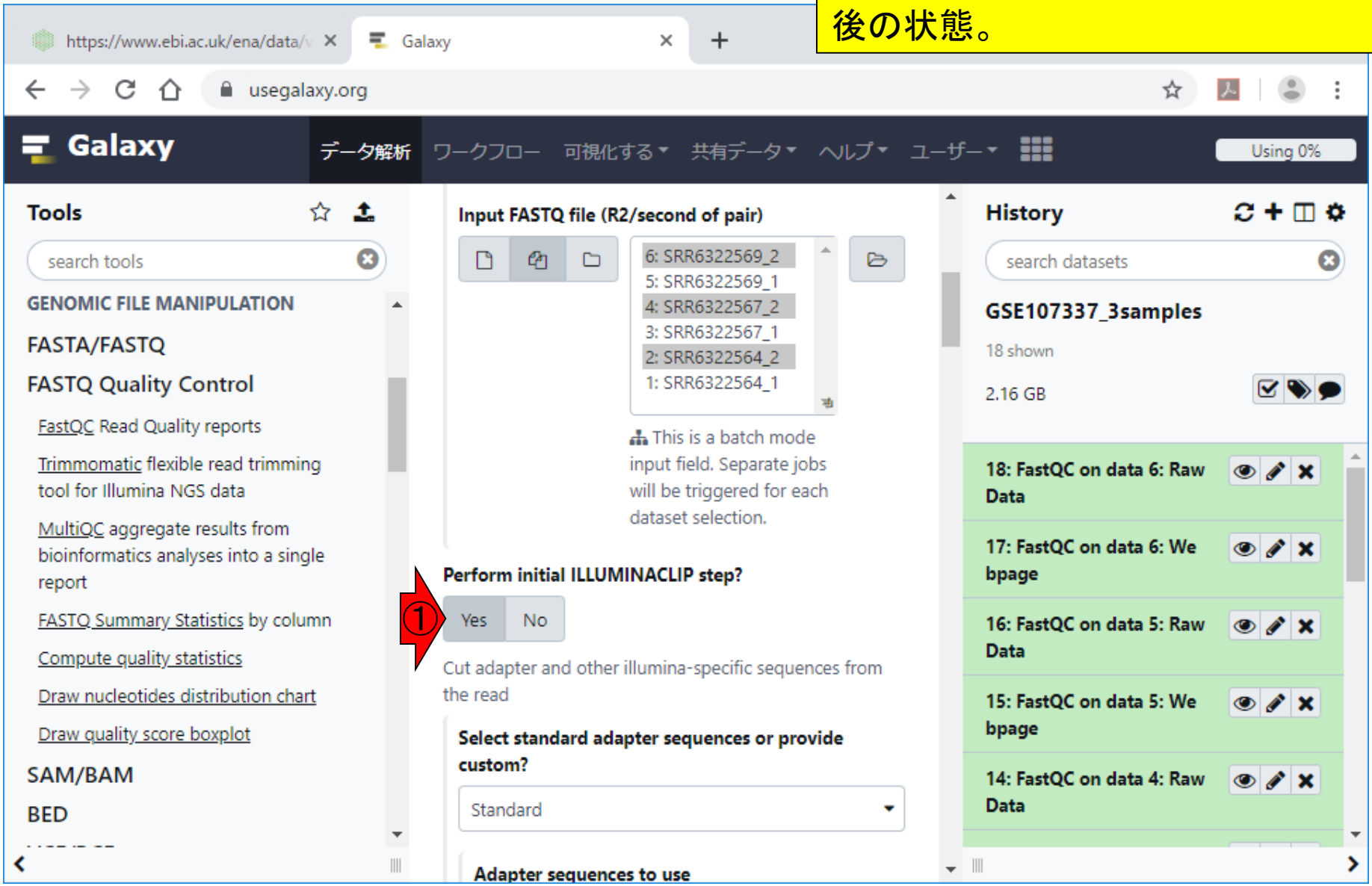
W5-5：アダプター除去

入力データはIlluminaのRNA-seqリードなので、そのアダプター配列除去も行ってもらいたい。デフォルトはNoなので、①Yesに変更。



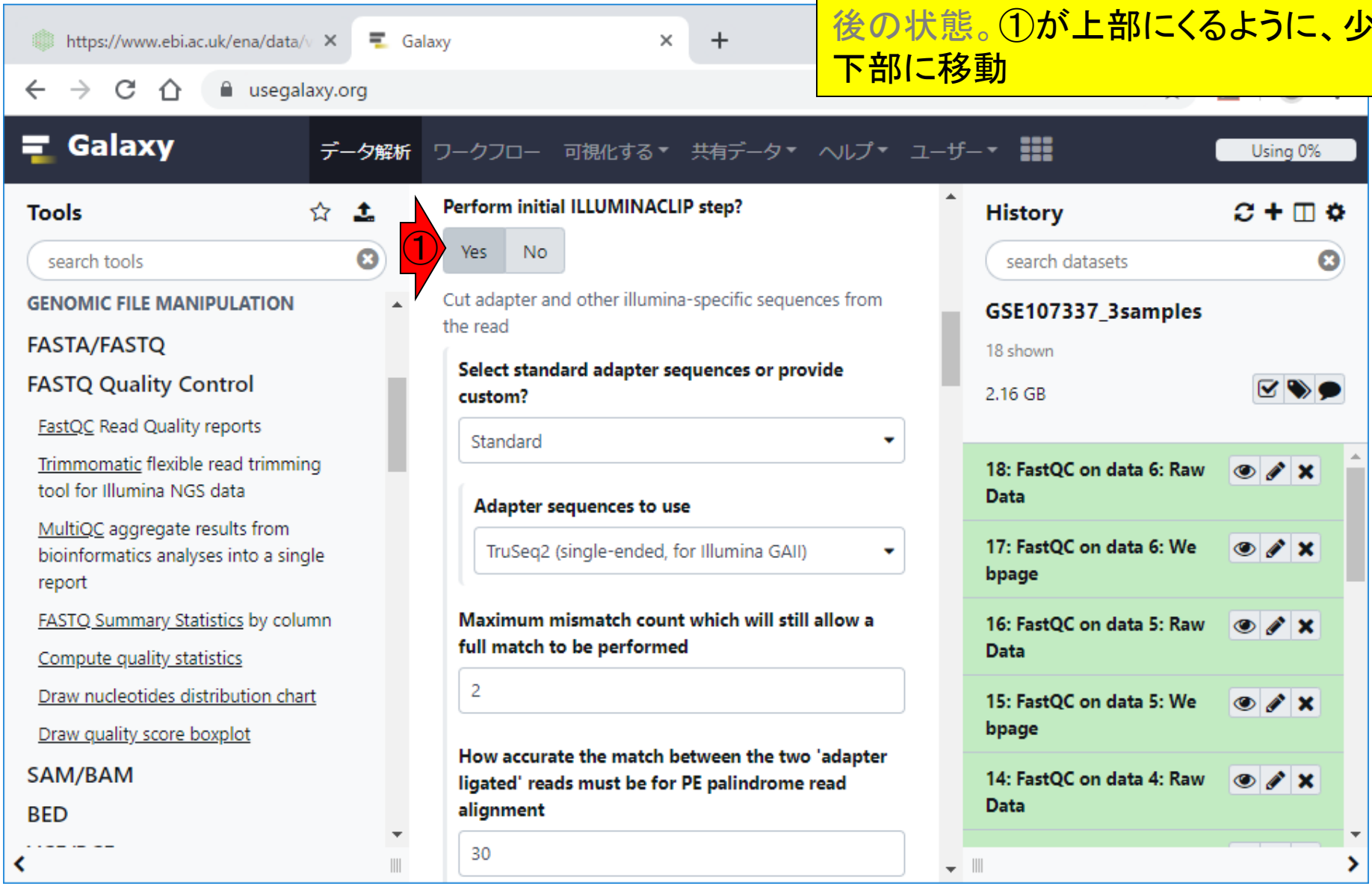
W5-5：アダプター除去

入力データはIlluminaのRNA-seqリードなので、そのアダプター配列除去も行ってもらいたい。デフォルトはNoなので、①Yesに変更。変更後の状態。



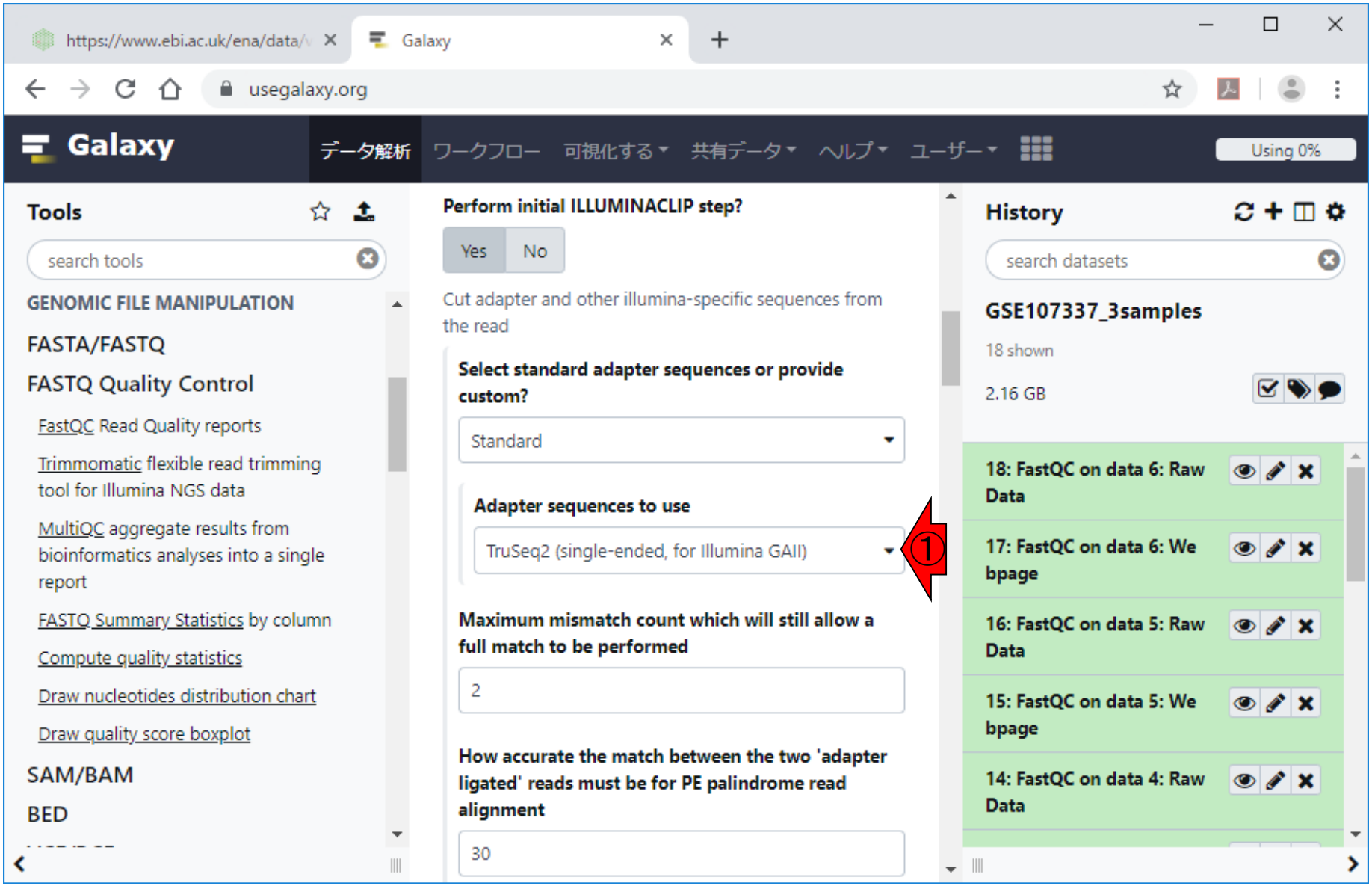
W5-5：アダプター除去

入力データはIlluminaのRNA-seqリードなので、そのアダプター配列除去も行ってもらいたい。デフォルトはNoなので、①Yesに変更。変更後の状態。①が上部にくるように、少しページ下部に移動



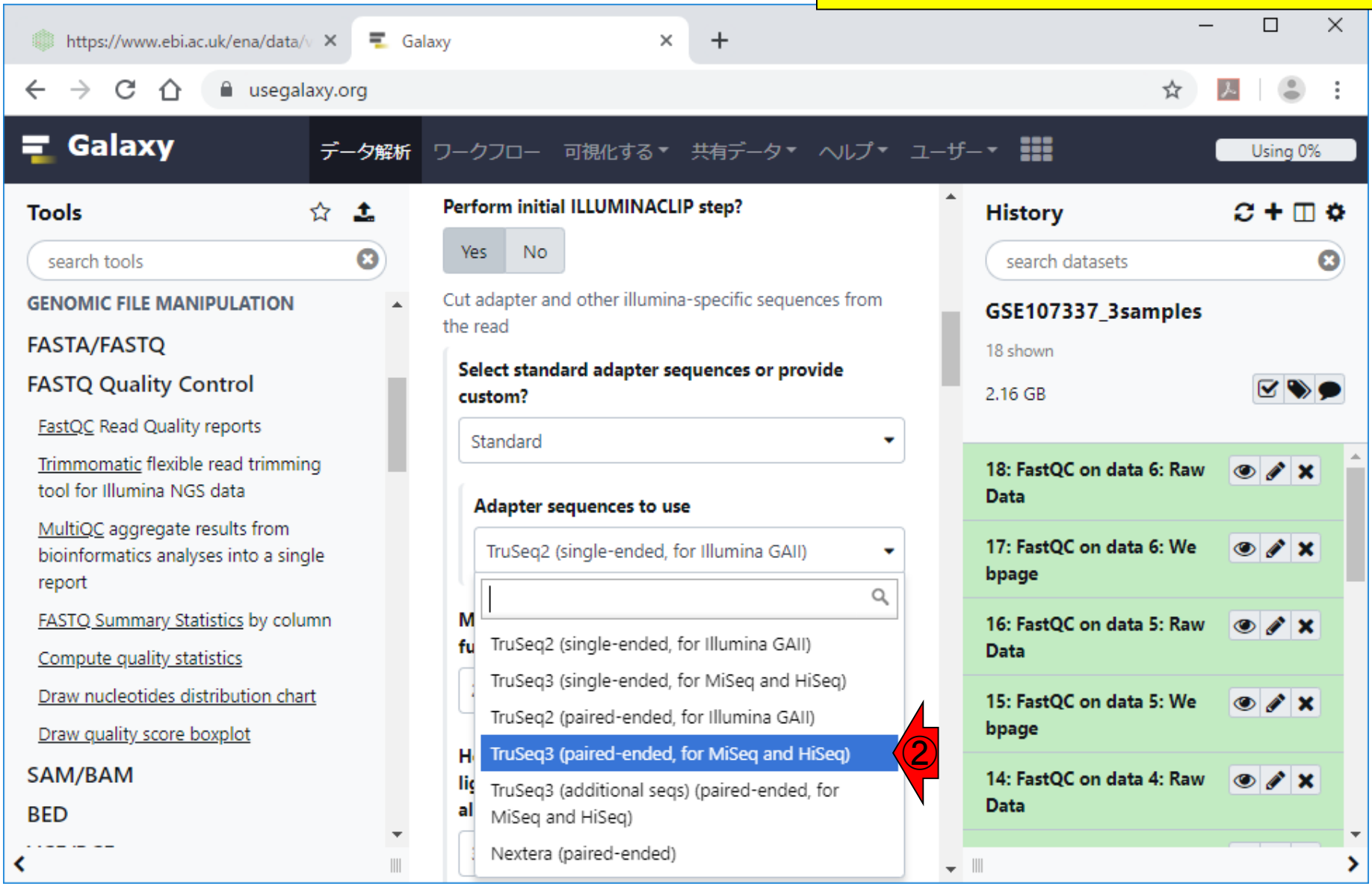
W5-6：配列を指定

アダプター配列のデフォルトは①single-endのTruSeq2になっている。



W5-6：配列を指定

アダプター配列のデフォルトは①single-endのTruSeq2になっている。入力データはMiSeqのpaired-endだと分かっているので、②を選択。



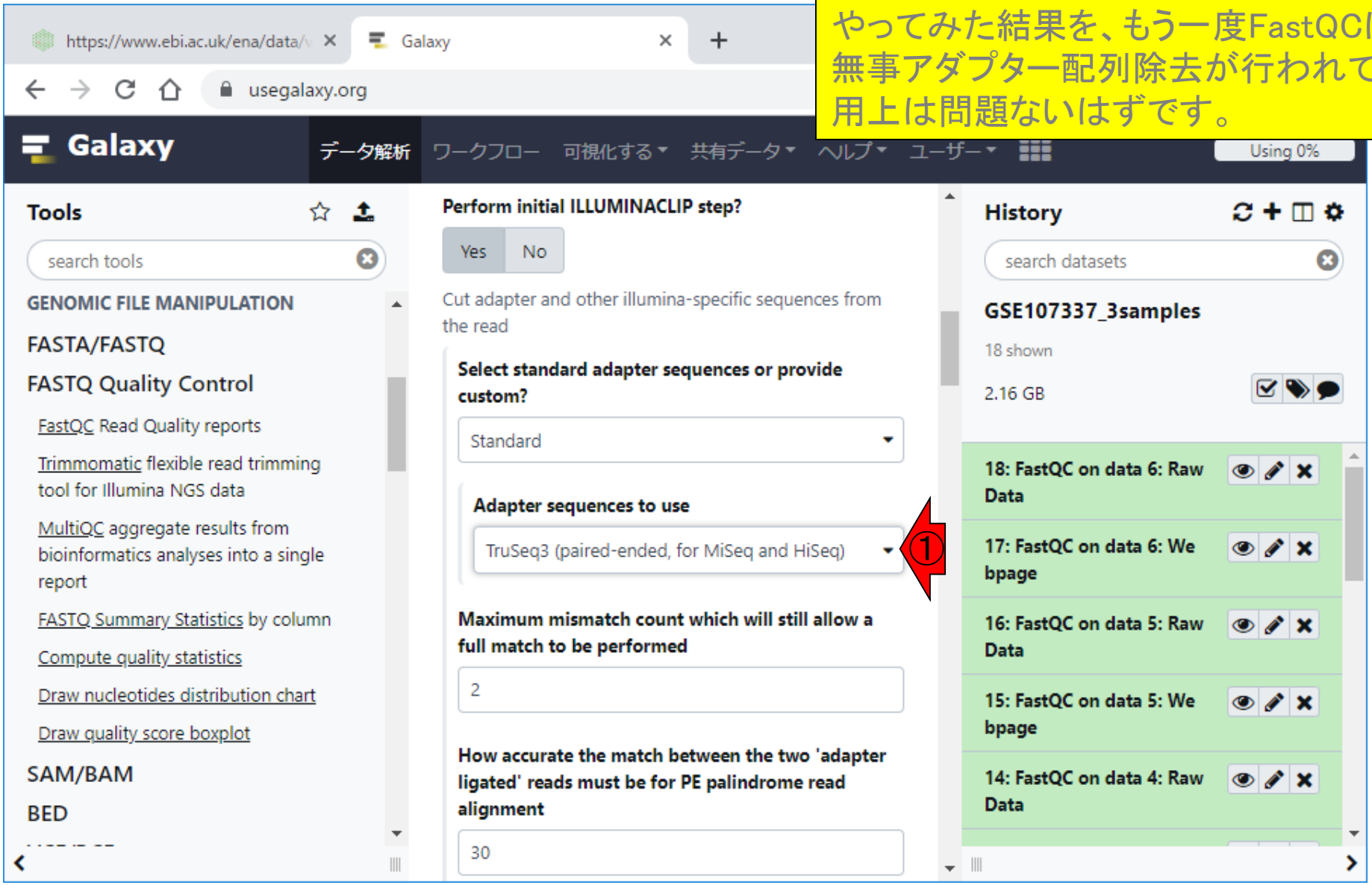
W5-6：補足説明

補足説明。①ENAに戻っています。②Illumina MiSeqの、③paired-endデータです。例えば④のリンク先を眺めると、バージョン番号までは分かりませんがTruSeqが用いられていることがわかります。

PRJNA419802	SAMN08098215	SRS2714083	SRX3422362	SRR6322563	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098214	SRS2714082	SRX3422363	SRR6322564	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098213	SRS2714084	SRX3422364	SRR6322565	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098212	SRS2714085	SRX3422365	SRR6322566	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098211	SRS2714086	SRX3422366	SRR6322567	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098210	SRS2714087	SRX3422367	SRR6322568	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098218	SRS2714088	SRX3422368	SRR6322569	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2
PRJNA419802	SAMN08098217	SRS2714089	SRX3422369	SRR6322570	568703	Lactobacillus rhamnosus GG	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2

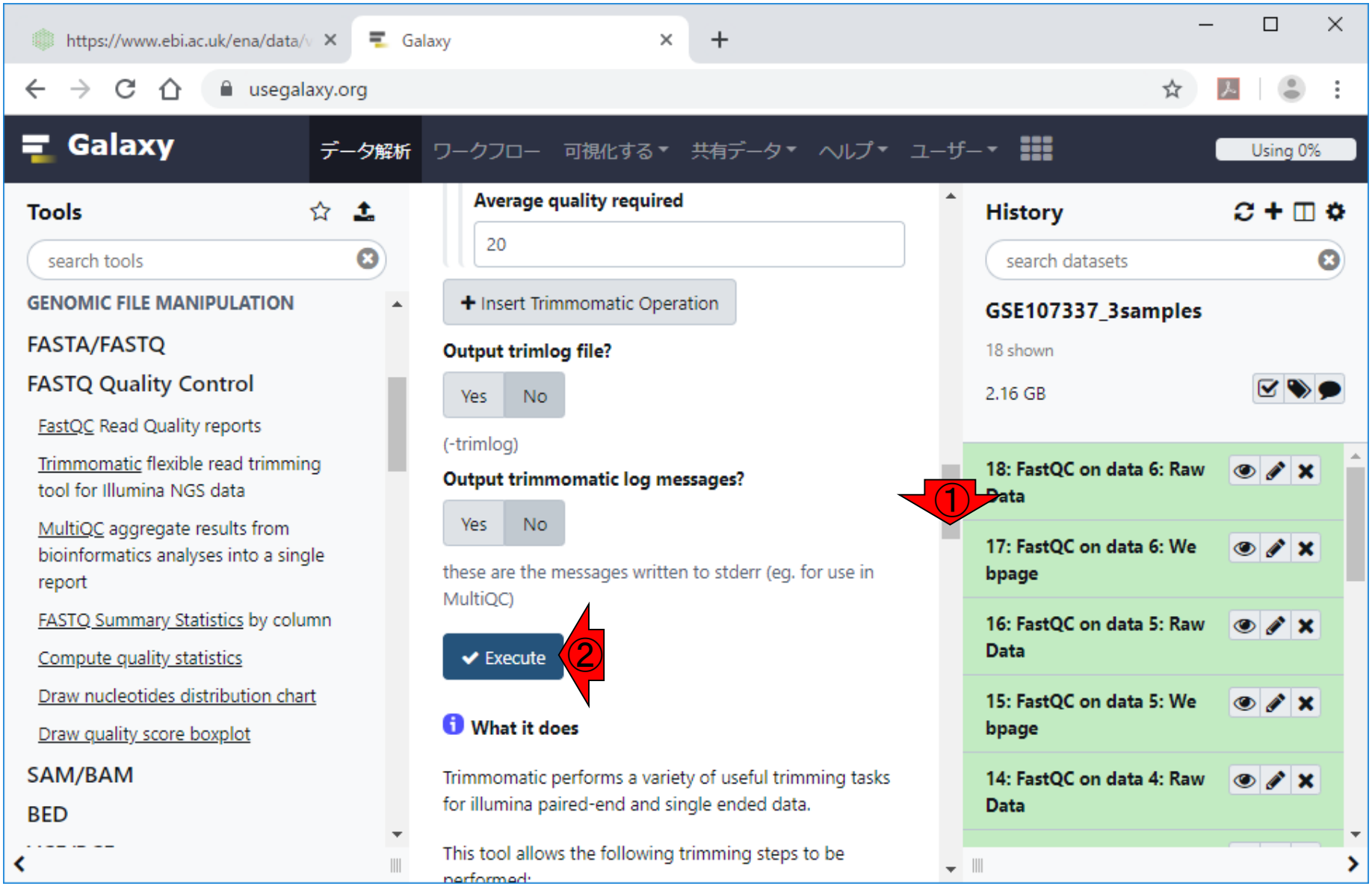
W5-6：配列を指定

①無事変更されたことを確認。実際問題としては、どれを選択していいかわかりかねる場合があります。しかし、とりあえずTrimmomaticをやってみた結果を、もう一度FastQCにかけて、無事アダプター配列除去が行われていれば実用上は問題ないはずです。



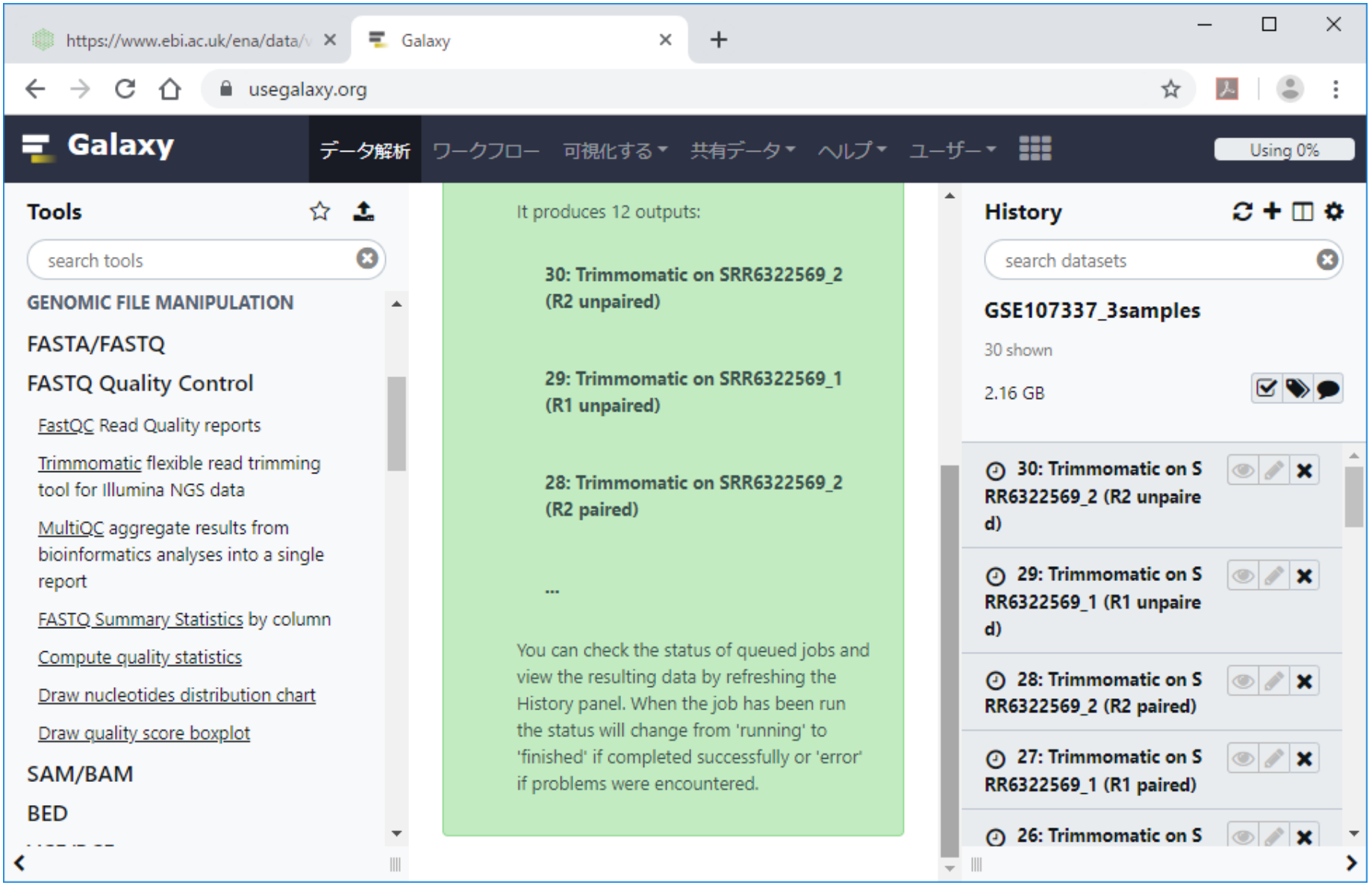
W5-7：実行

残りのオプションは特に変更せず、①下部に移動して、②実行。



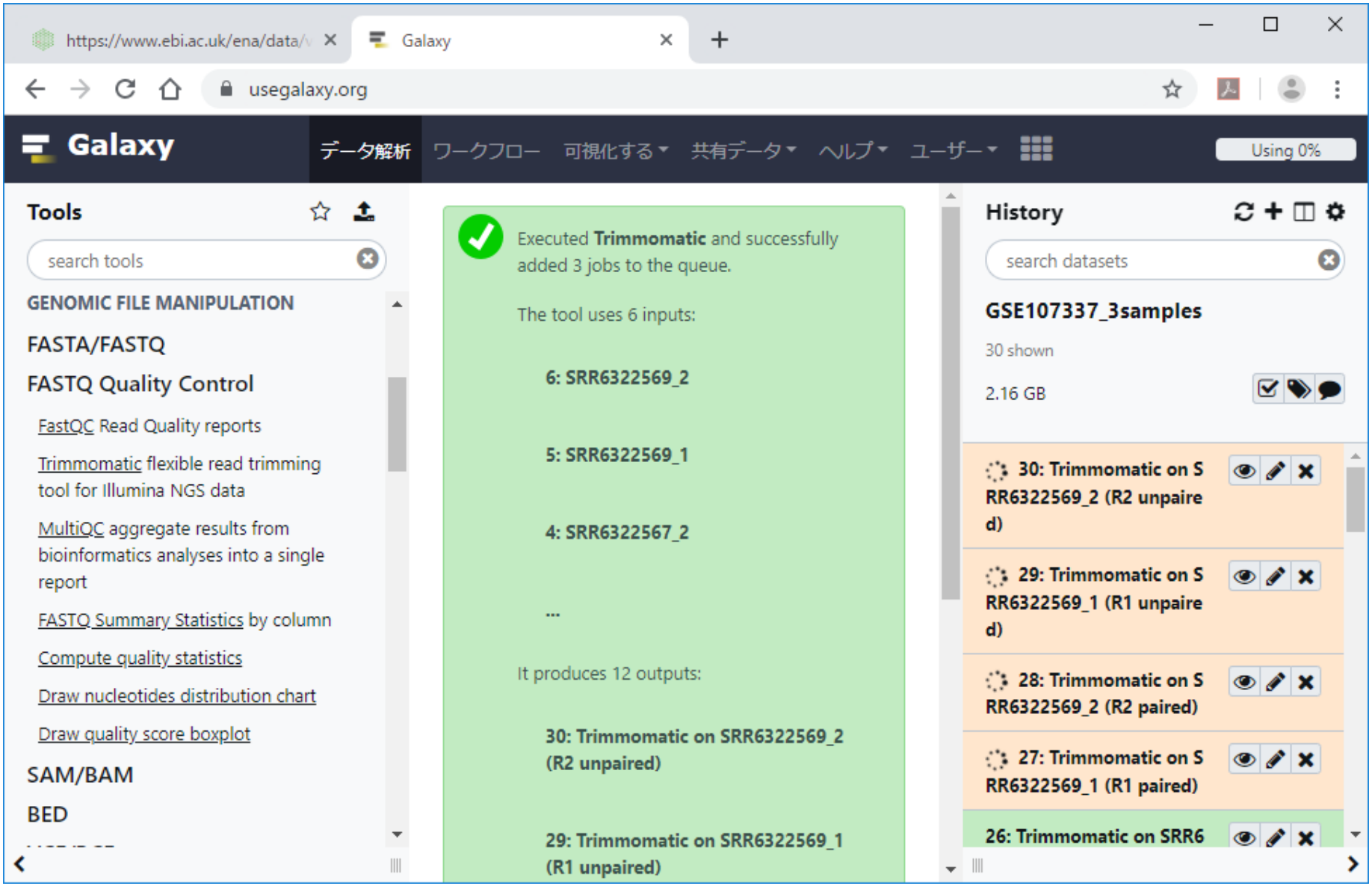
実行中…。

W5-8：実行中


https://www.ebi.ac.uk/ena/data/
Galaxy
usegalaxy.org
Galaxy
データ解析
ワークフロー
可視化する
共有データ
ヘルプ
ユーザー
Using 0%
Tools
search tools
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
FastQC Read Quality reports
Trimmomatic flexible read trimming tool for Illumina NGS data
MultiQC aggregate results from bioinformatics analyses into a single report
FASTQ Summary Statistics by column
Compute quality statistics
Draw nucleotides distribution chart
Draw quality score boxplot
SAM/BAM
BED
It produces 12 outputs:
30: Trimmomatic on SRR6322569_2 (R2 unpaired)
29: Trimmomatic on SRR6322569_1 (R1 unpaired)
28: Trimmomatic on SRR6322569_2 (R2 paired)
...
You can check the status of queued jobs and view the resulting data by refreshing the History panel. 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
GSE107337_3samples
30 shown
2.16 GB
30: Trimmomatic on SRR6322569_2 (R2 unpaired)
29: Trimmomatic on SRR6322569_1 (R1 unpaired)
28: Trimmomatic on SRR6322569_2 (R2 paired)
27: Trimmomatic on SRR6322569_1 (R1 paired)
26: Trimmomatic on S

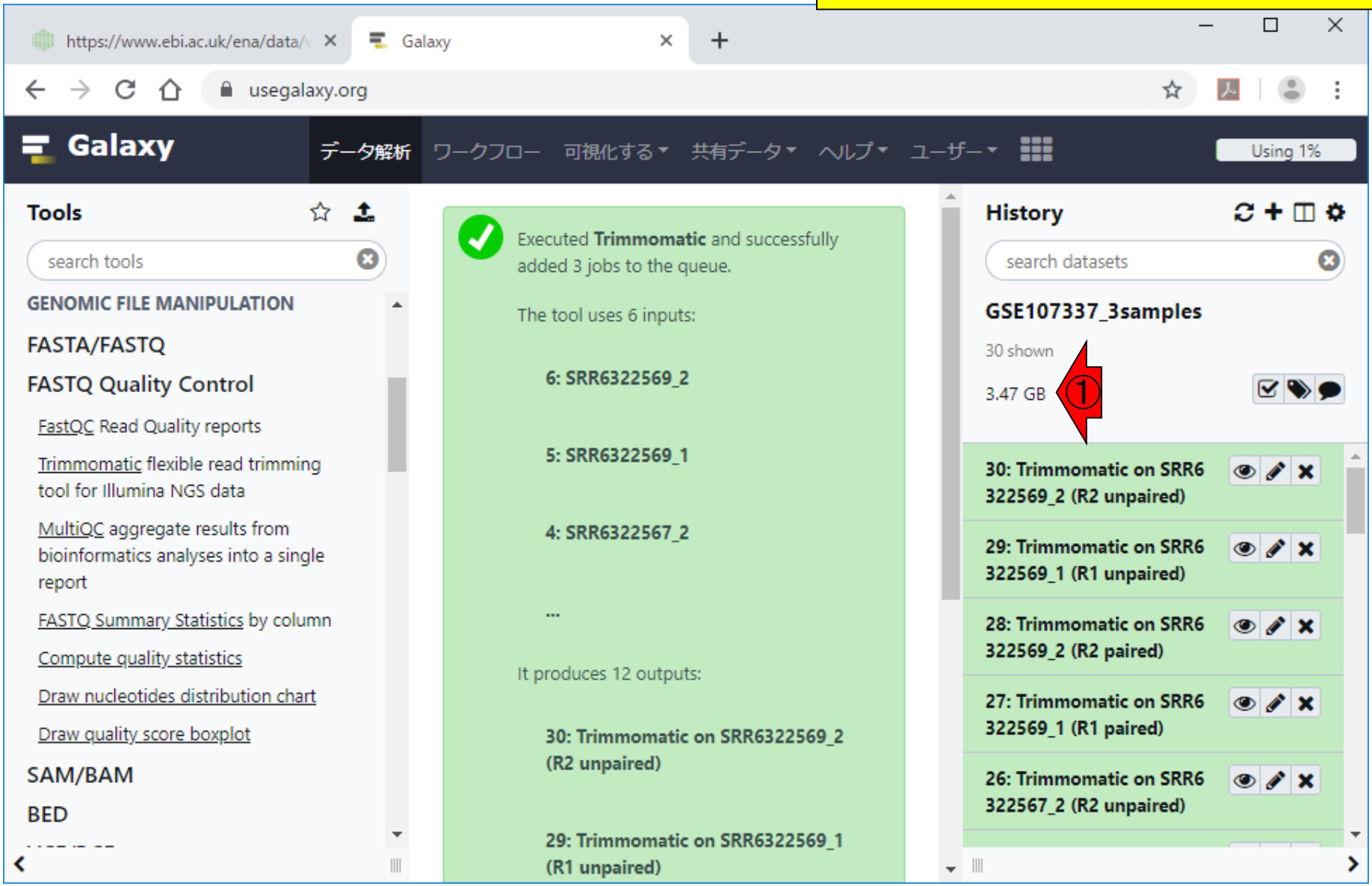
実行中…。

W5-8：実行中


https://www.ebi.ac.uk/ena/data/
Galaxy
usegalaxy.org
Galaxy
データ解析
ワークフロー
可視化する
共有データ
ヘルプ
ユーザー
Using 0%
Tools
search tools
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
FastQC Read Quality reports
Trimmomatic flexible read trimming tool for Illumina NGS data
MultiQC aggregate results from bioinformatics analyses into a single report
FASTQ Summary Statistics by column
Compute quality statistics
Draw nucleotides distribution chart
Draw quality score boxplot
SAM/BAM
BED
Executed Trimmomatic and successfully added 3 jobs to the queue.
The tool uses 6 inputs:
6: SRR6322569_2
5: SRR6322569_1
4: SRR6322567_2
...
It produces 12 outputs:
30: Trimmomatic on SRR6322569_2 (R2 unpaired)
29: Trimmomatic on SRR6322569_1 (R1 unpaired)
History
search datasets
GSE107337_3samples
30 shown
2.16 GB
30: Trimmomatic on SRR6322569_2 (R2 unpaired)
29: Trimmomatic on SRR6322569_1 (R1 unpaired)
28: Trimmomatic on SRR6322569_2 (R2 paired)
27: Trimmomatic on SRR6322569_1 (R1 paired)
26: Trimmomatic on SRR6

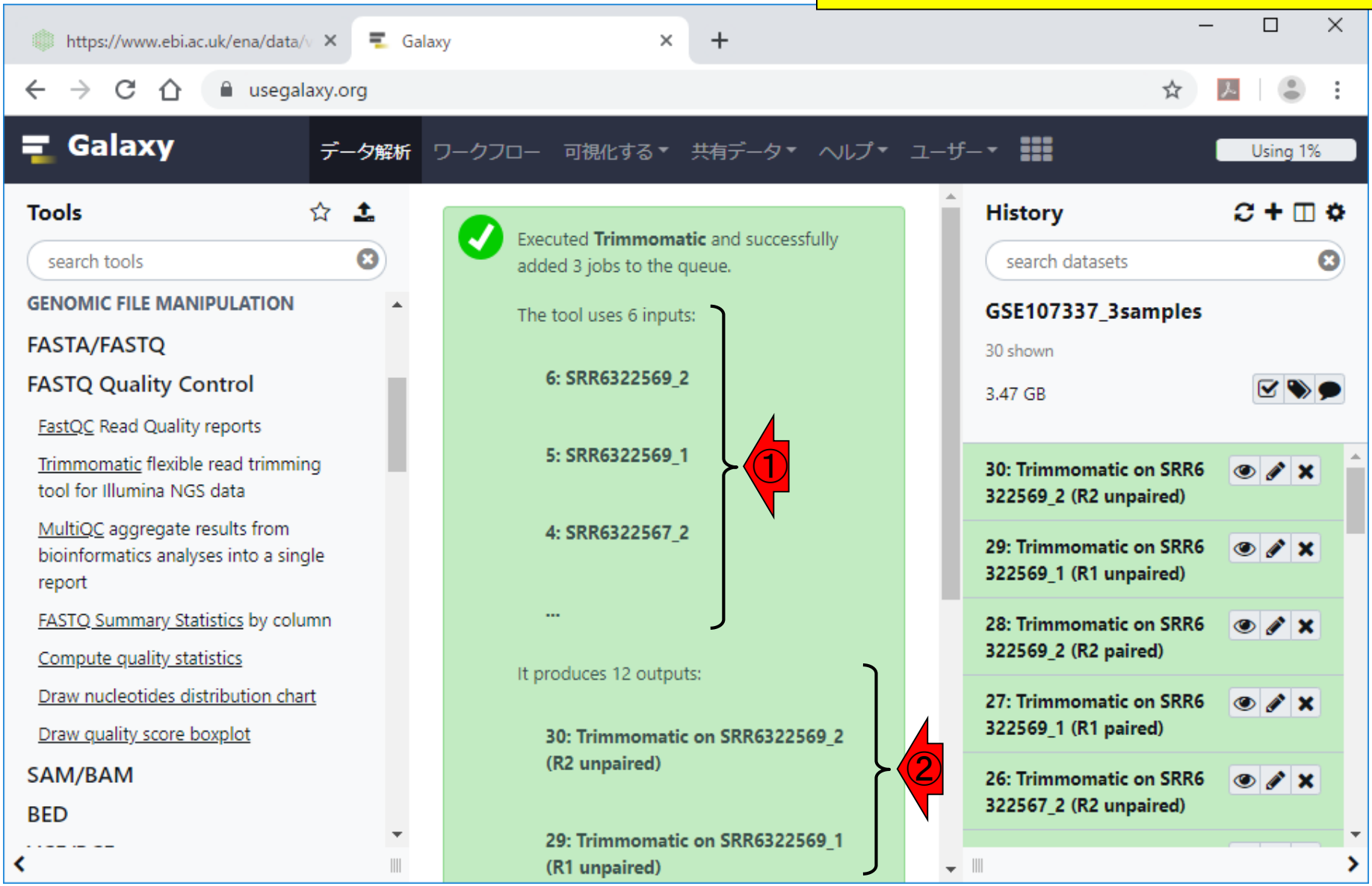
W5-9：実行完了

無事実行完了したようです。このときは約10分かかりました。①サイズが3.47 GBに増えていますね。



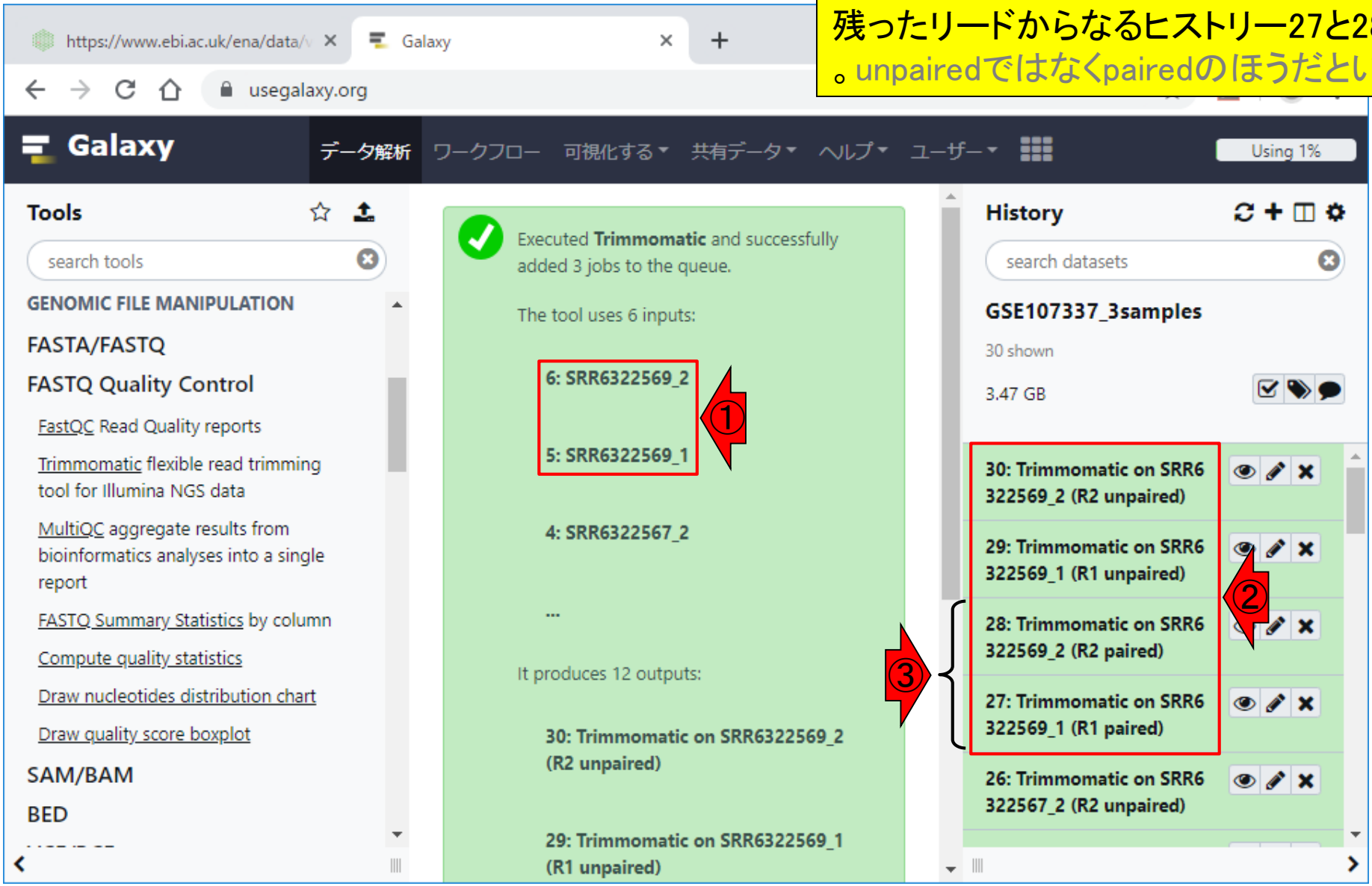
W5-10：解説

①Trimmomaticの入力は、ヒストリー1から6のFASTQファイル。②出力は計12個。ヒストリー19から30に相当します。



W5-10：解説

例えば、①SRR6322569の入力はヒストリー5と6、②出力はヒストリー27から30。このうち、その後の解析に用いるのは通常、③ペアで生き残ったリードからなるヒストリー27と28のデータ。unpairedではなくpairedのほうだということ。

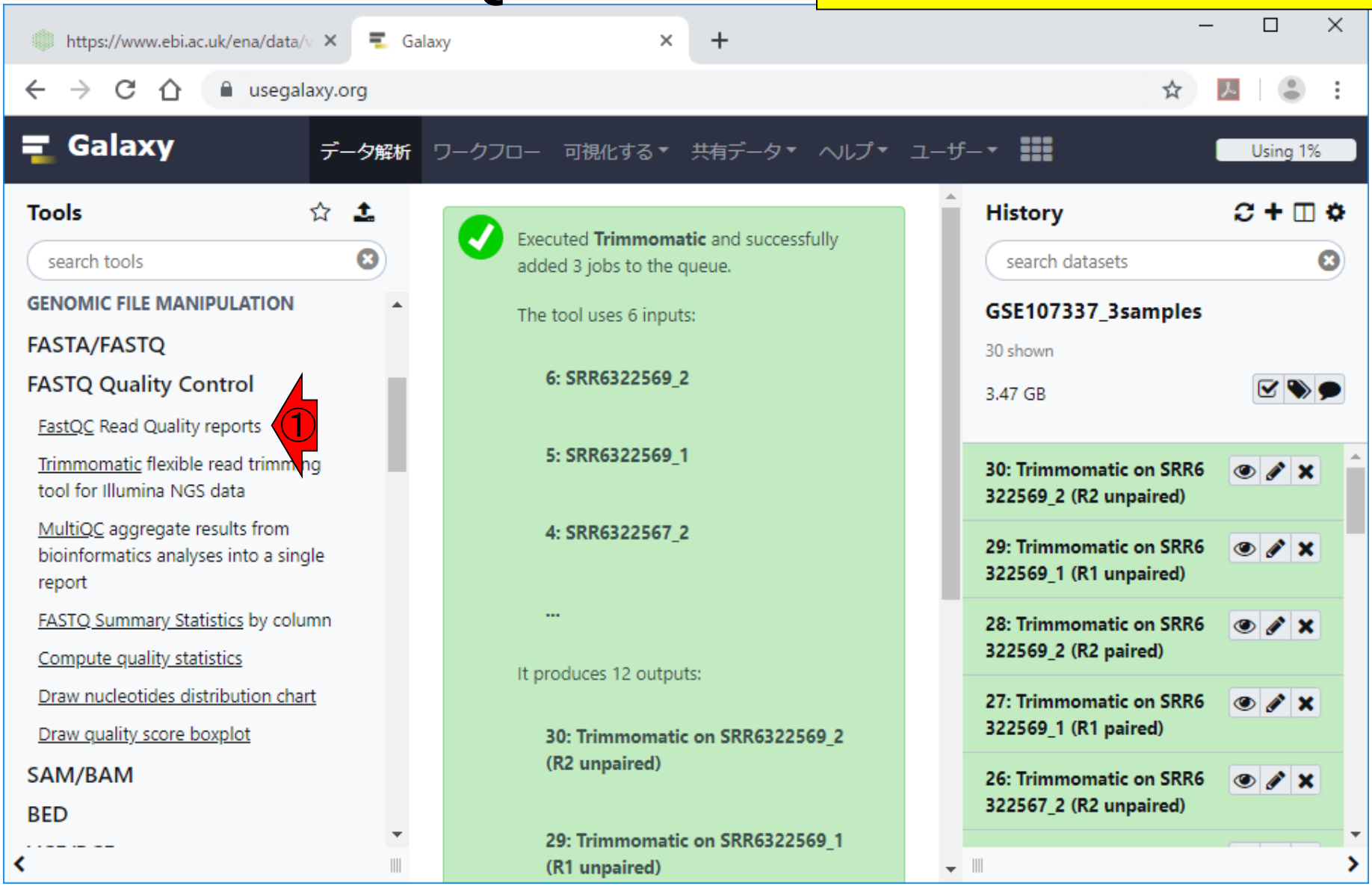


Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- **W6：クオリティチェック（FastQC）**
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

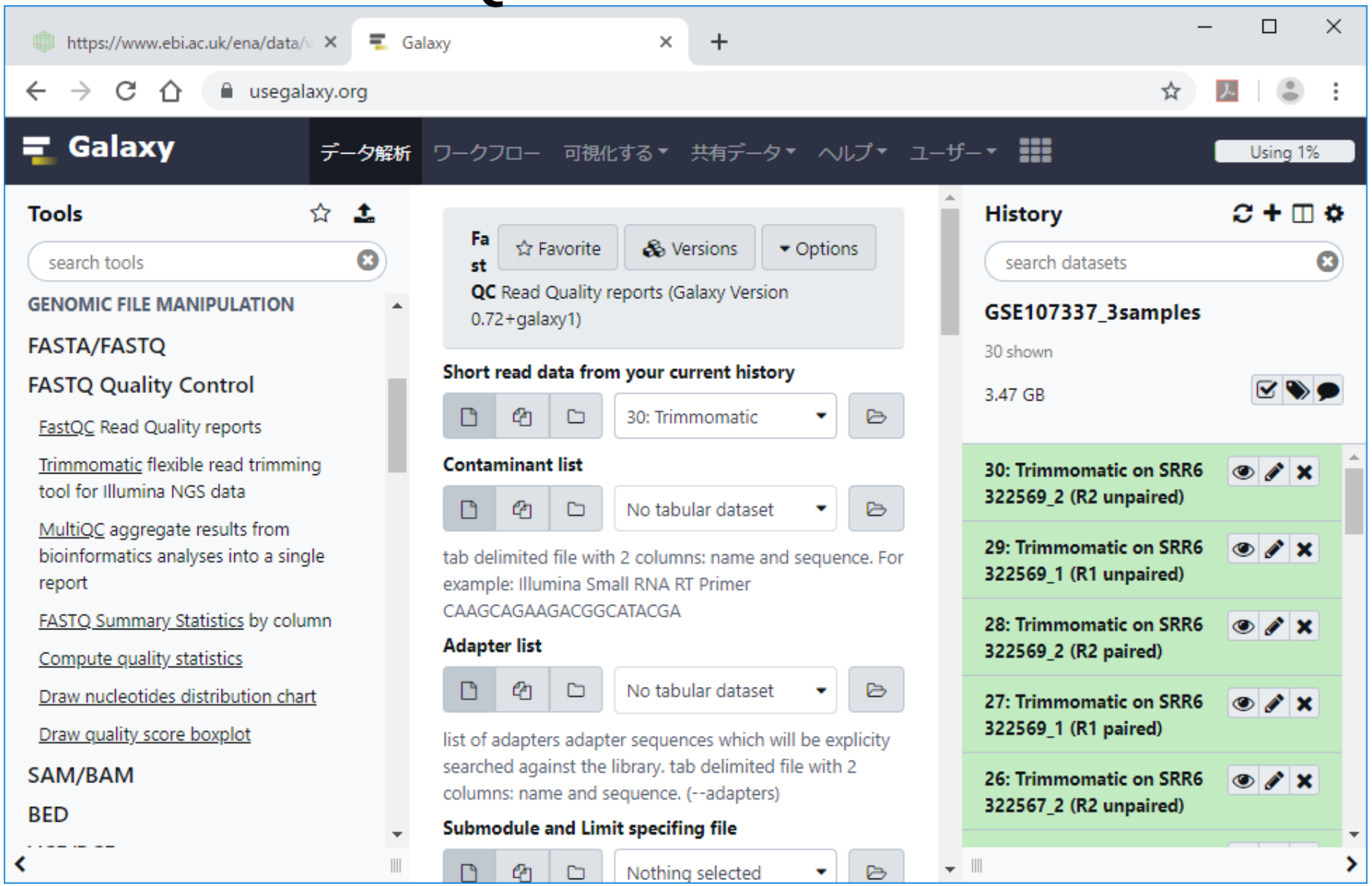
W6-1：FastQC

Trimmomatic実行結果の妥当性を確認すべく、再度①FastQCを実行。第11回のW13とノリは同じ。



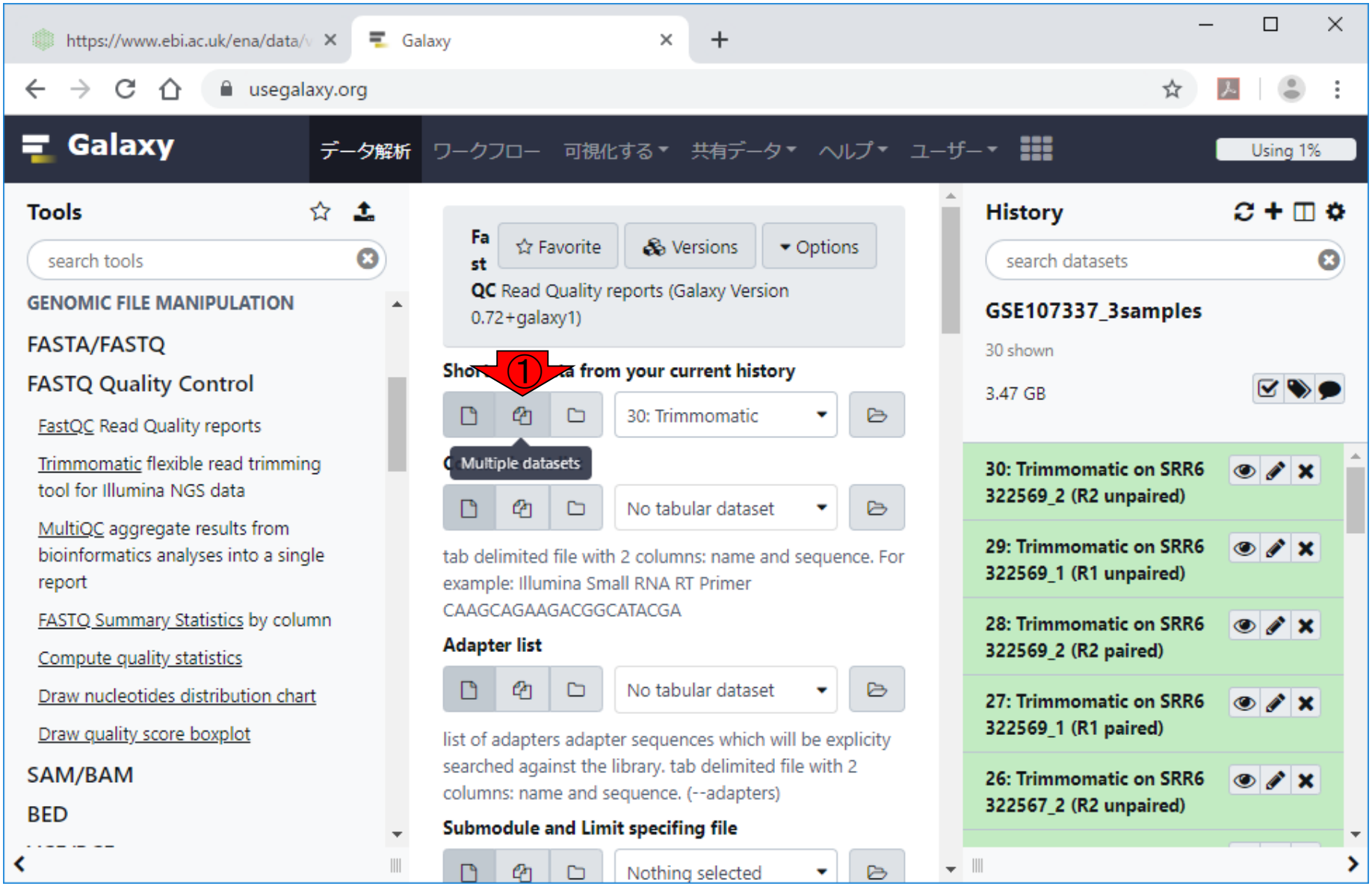
こんな感じになる。

W6-1：FastQC



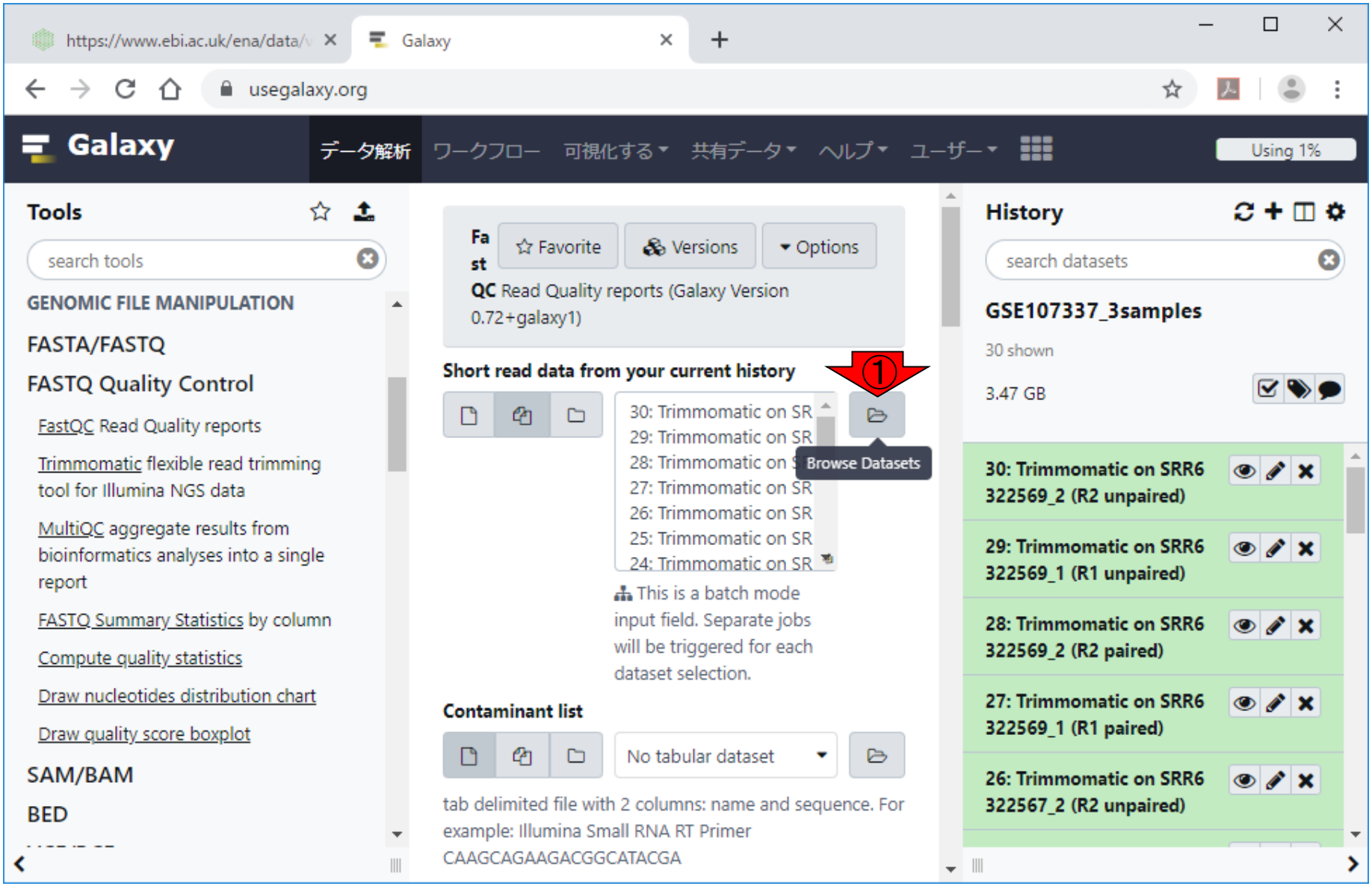
こんな感じになる。①Multiple datasets。

W6-2：複数個指定



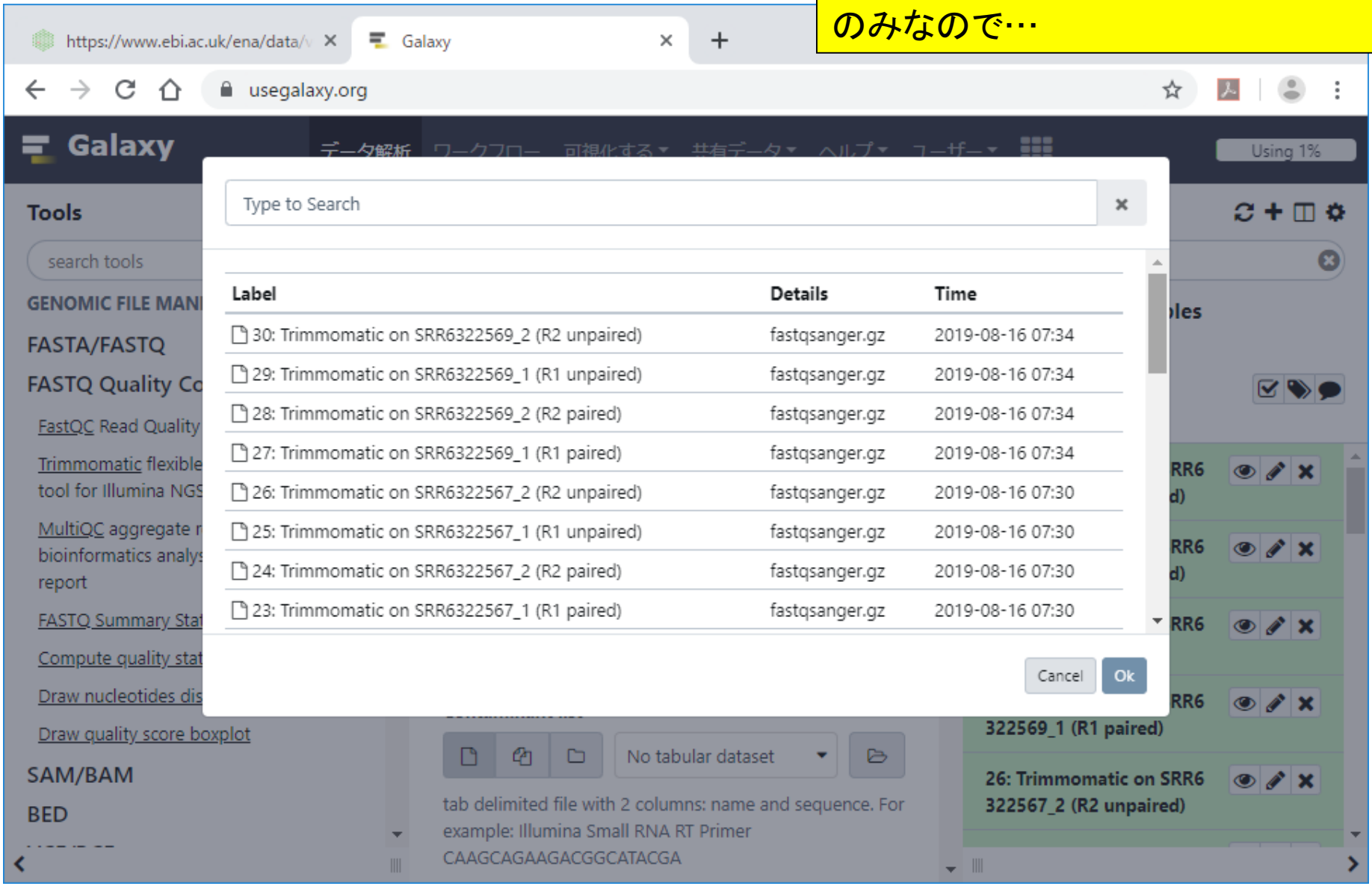
W6-2：複数個指定

こんな感じになる。①Multiple datasets。②Browse Datasets。



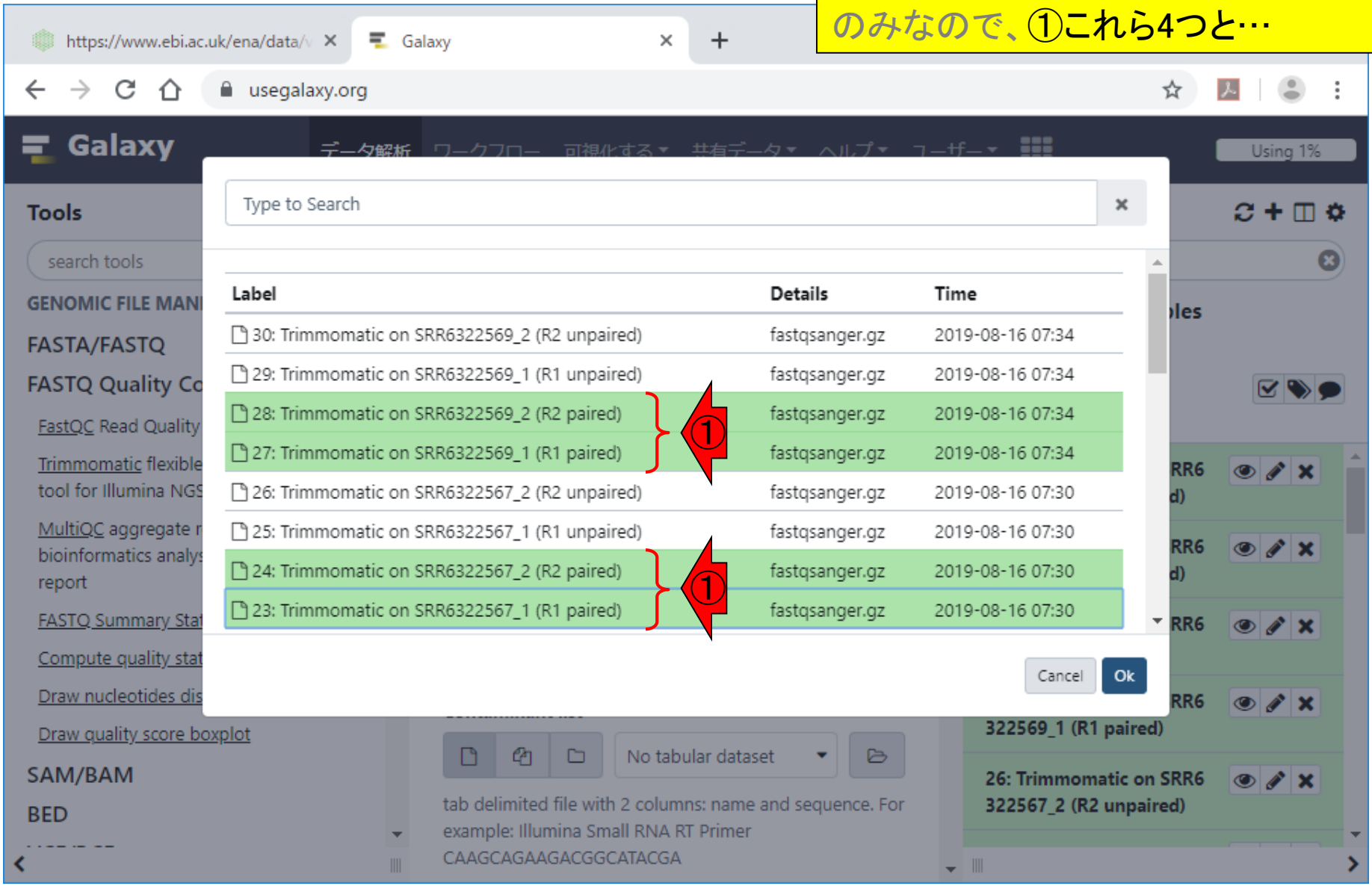
W6-2：複数個指定

こんな感じになる。①Multiple datasets。②Browse Datasets。こんな感じになる。実行したいのはTrimmomatic実行後のpairedのデータのみなので…



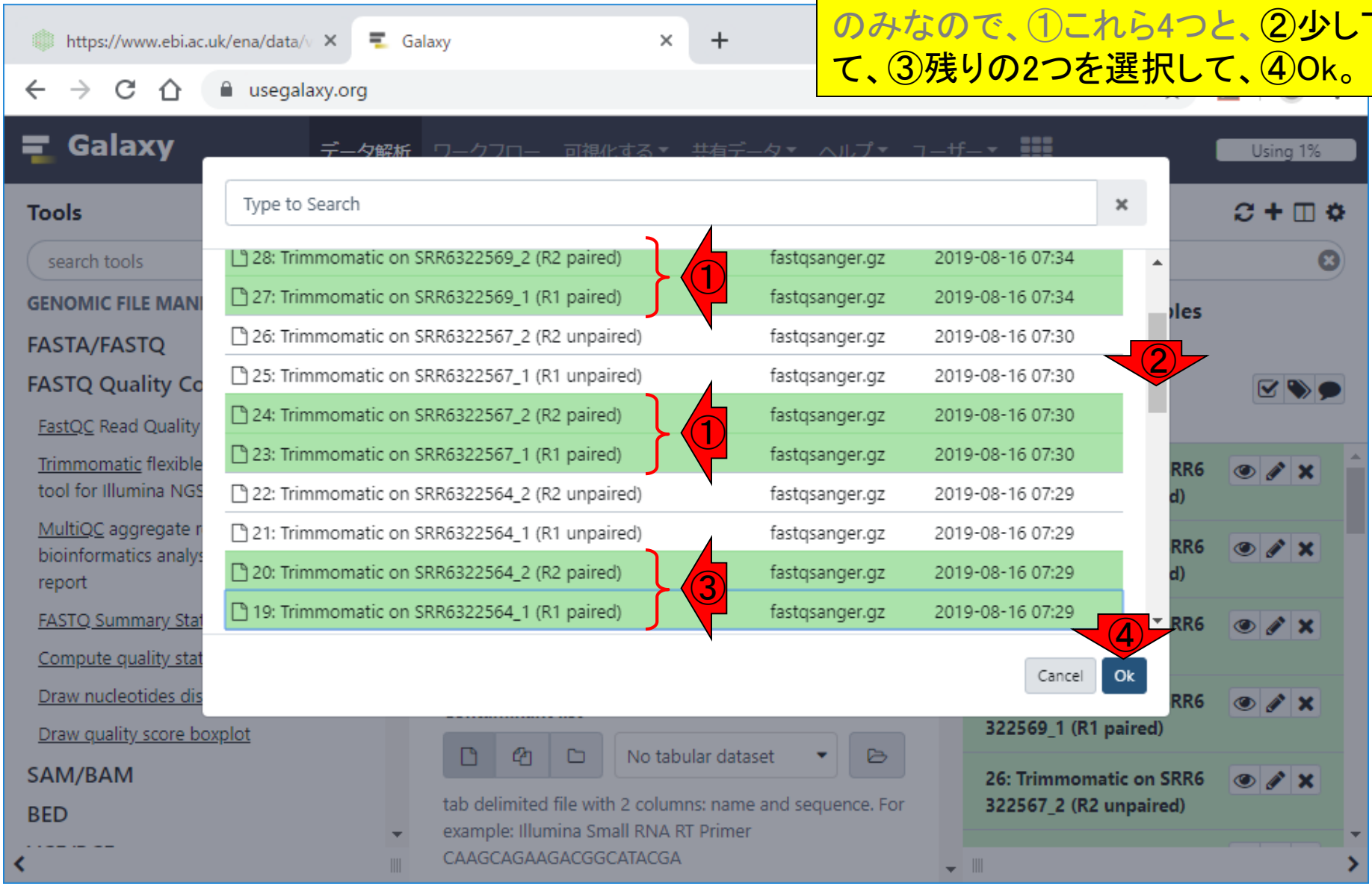
W6-2：複数個指定

こんな感じになる。①Multiple datasets。②Browse Datasets。こんな感じになる。実行したいのはTrimmomatic実行後のpairedのデータのみなので、①これら4つと…



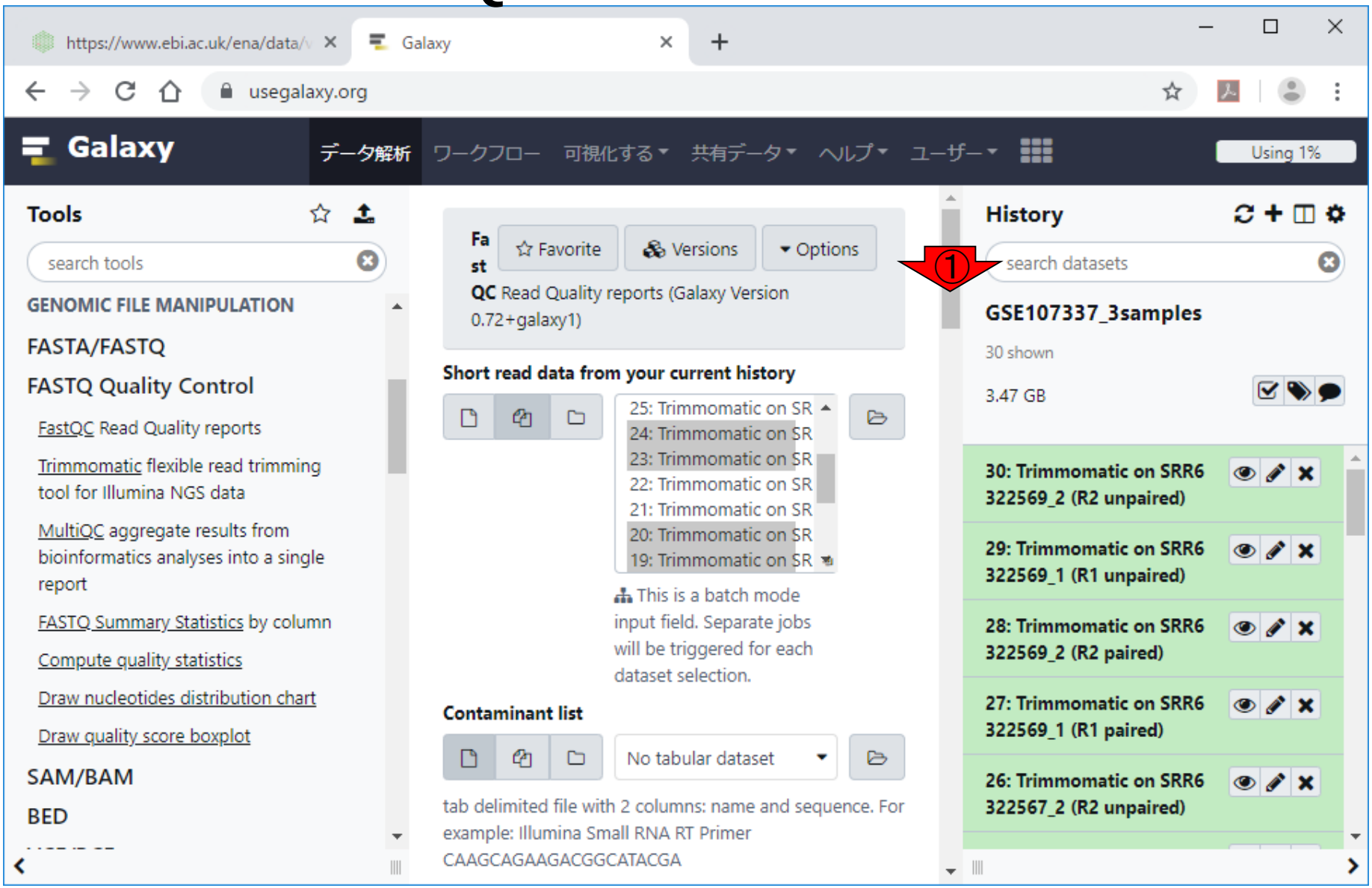
W6-2：複数個指定

こんな感じになる。①Multiple datasets。②Browse Datasets。こんな感じになる。実行したいのはTrimmomatic実行後のpairedのデータのみなので、①これら4つと、②少し下に移動して、③残りの2つを選択して、④Ok。



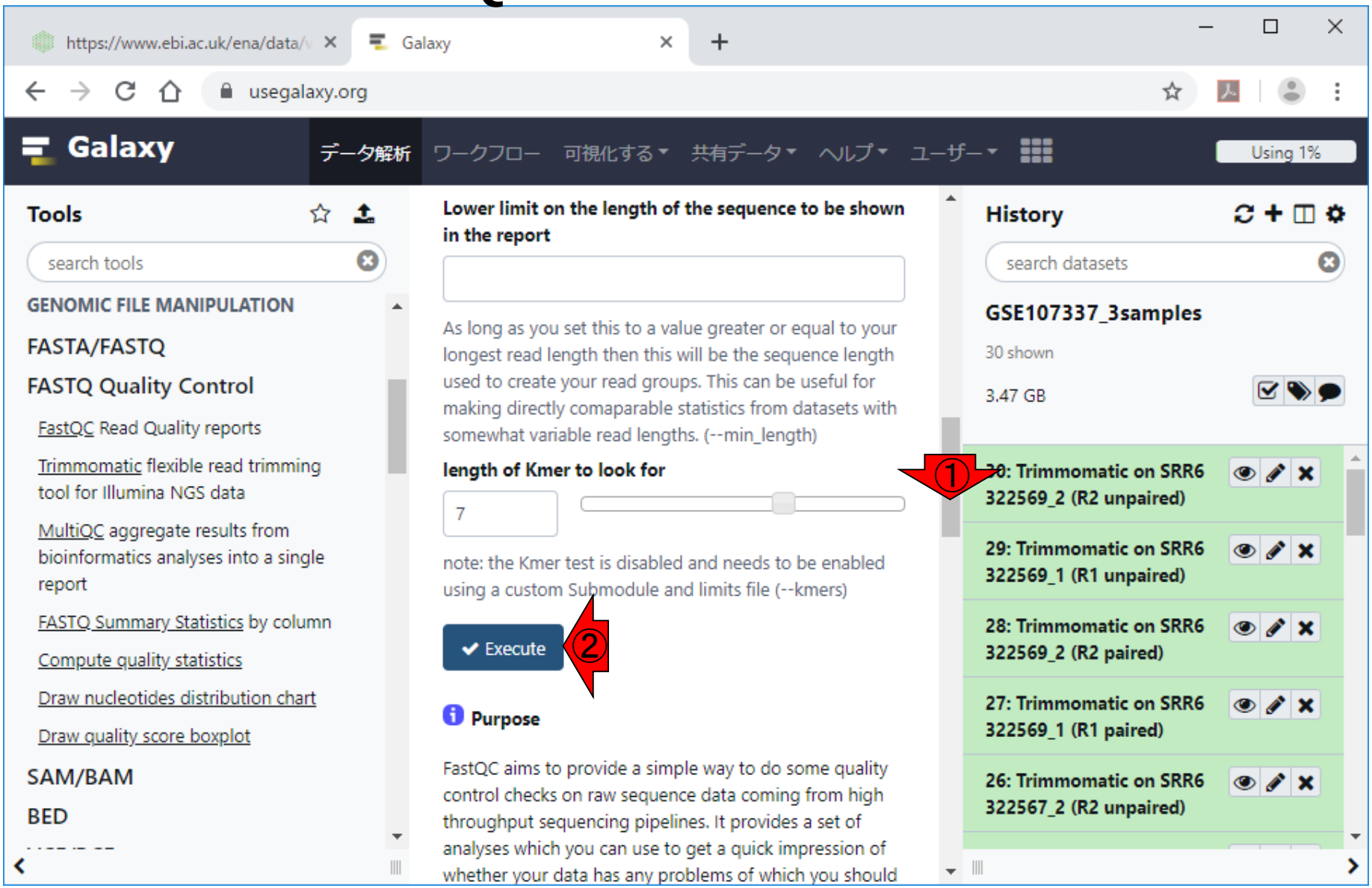
W6-3：FastQC実行

こんな感じになる。①下部に移動して、実行ボタンを探す。



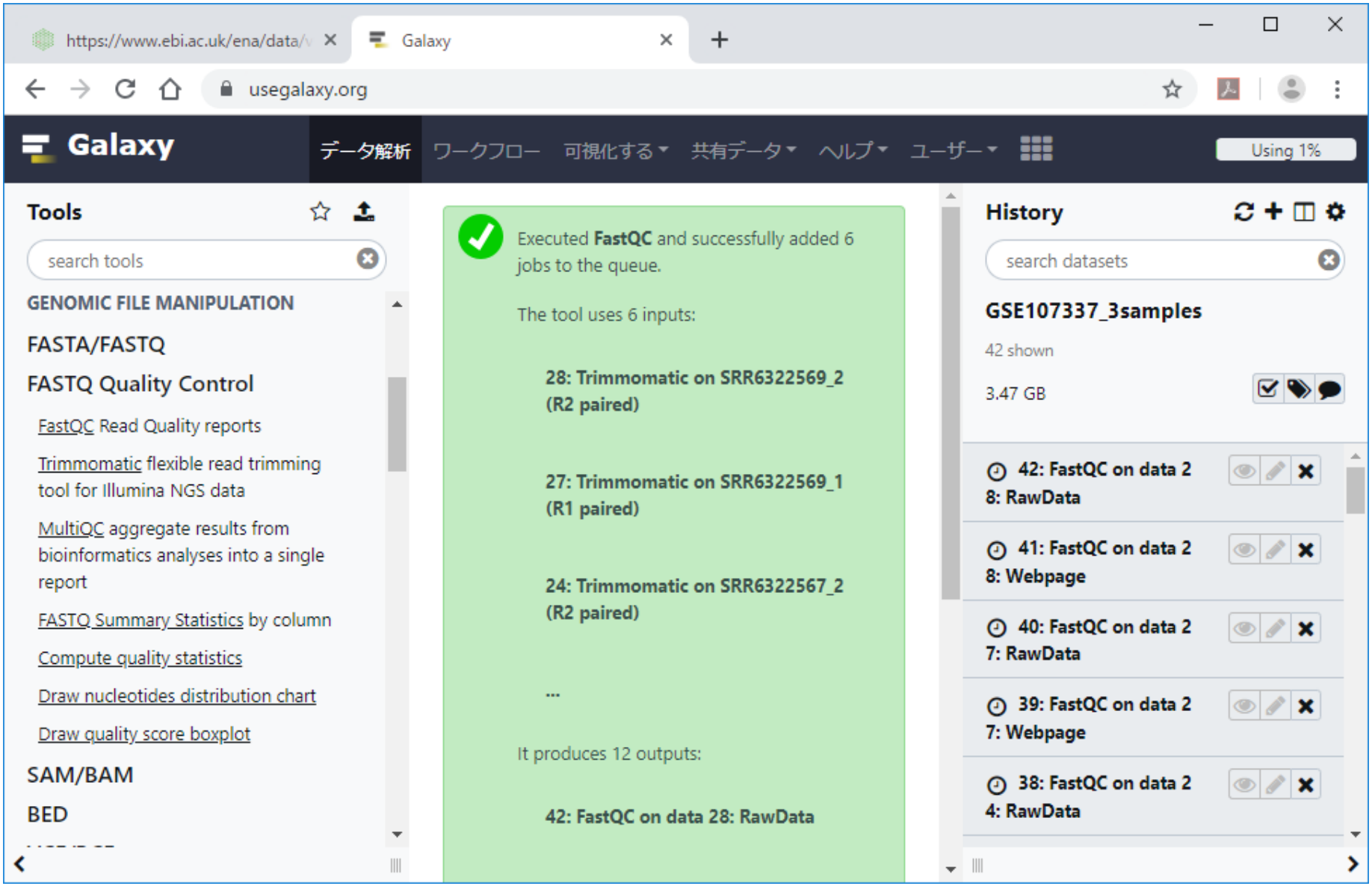
W6-3：FastQC実行

こんな感じになる。①下部に移動して、実行ボタンを探す。②Execute。



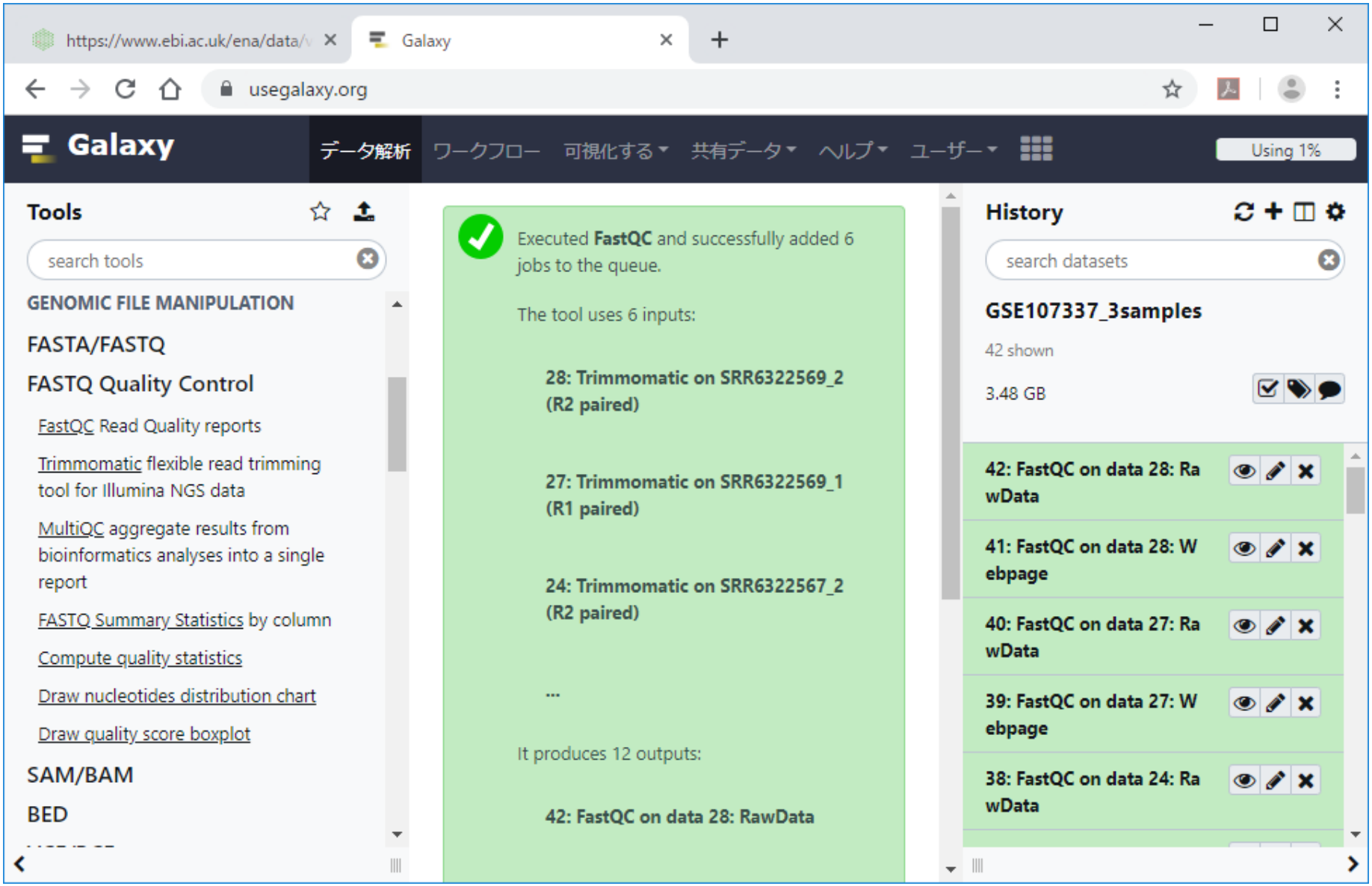
実行中。

W6-4：実行中


https://www.ebi.ac.uk/ena/data/
Galaxy
usegalaxy.org
Galaxy
データ解析
ワークフロー
可視化する
共有データ
ヘルプ
ユーザー
Using 1%
Tools
search tools
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
FastQC Read Quality reports
Trimmomatic flexible read trimming tool for Illumina NGS data
MultiQC aggregate results from bioinformatics analyses into a single report
FASTQ Summary Statistics by column
Compute quality statistics
Draw nucleotides distribution chart
Draw quality score boxplot
SAM/BAM
BED
Executed FastQC and successfully added 6 jobs to the queue.
The tool uses 6 inputs:
28: Trimmomatic on SRR6322569_2 (R2 paired)
27: Trimmomatic on SRR6322569_1 (R1 paired)
24: Trimmomatic on SRR6322567_2 (R2 paired)
...
It produces 12 outputs:
42: FastQC on data 28: RawData
History
search datasets
GSE107337_3samples
42 shown
3.47 GB
42: FastQC on data 2 8: RawData
41: FastQC on data 2 8: Webpage
40: FastQC on data 2 7: RawData
39: FastQC on data 2 7: Webpage
38: FastQC on data 2 4: RawData

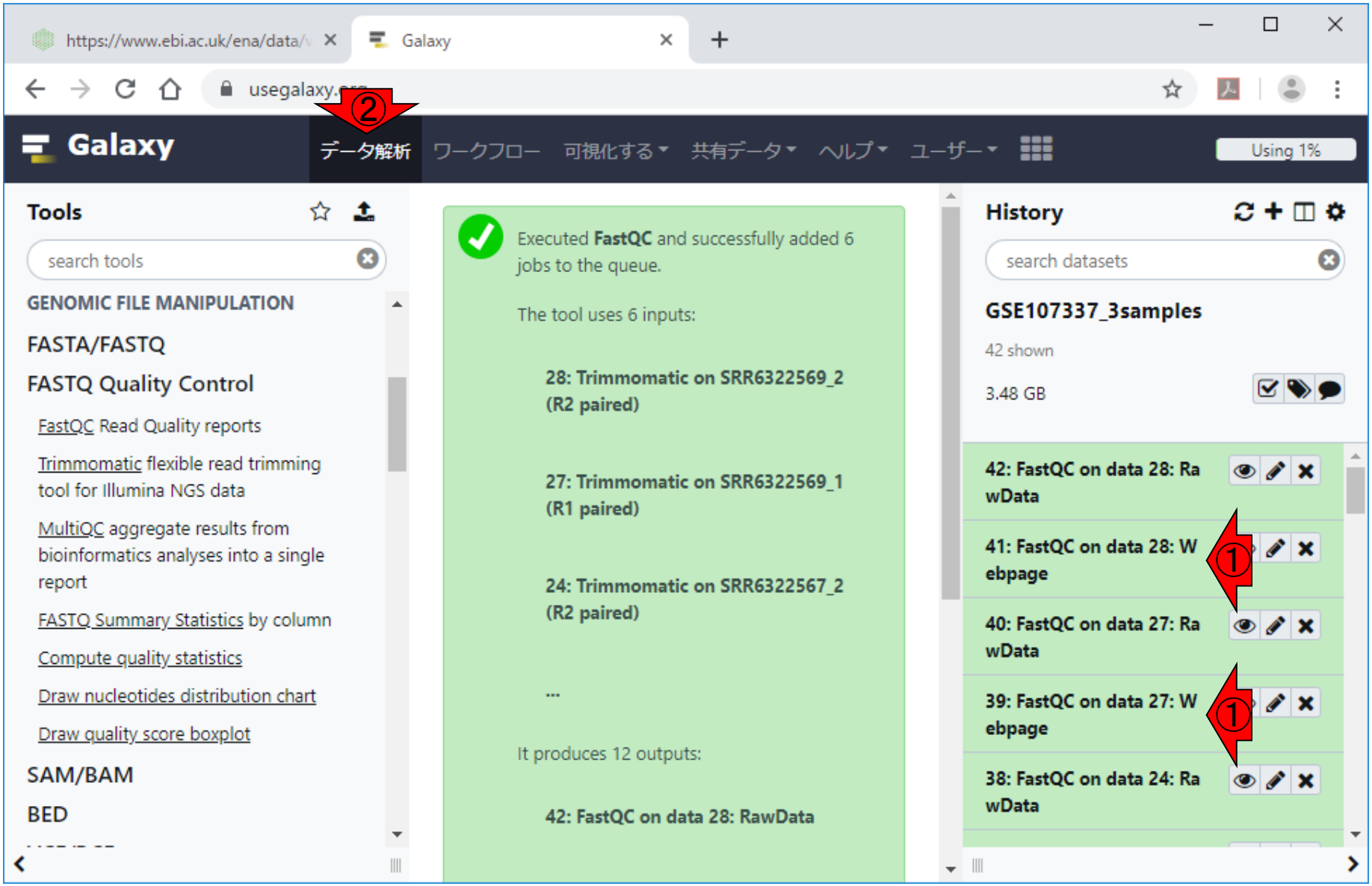
実行完了。

W6-5：実行完了



W6-6：htmlを保存

W4-7と同様の手順で、①htmlを保存。②中央パネルをリフレッシュ。

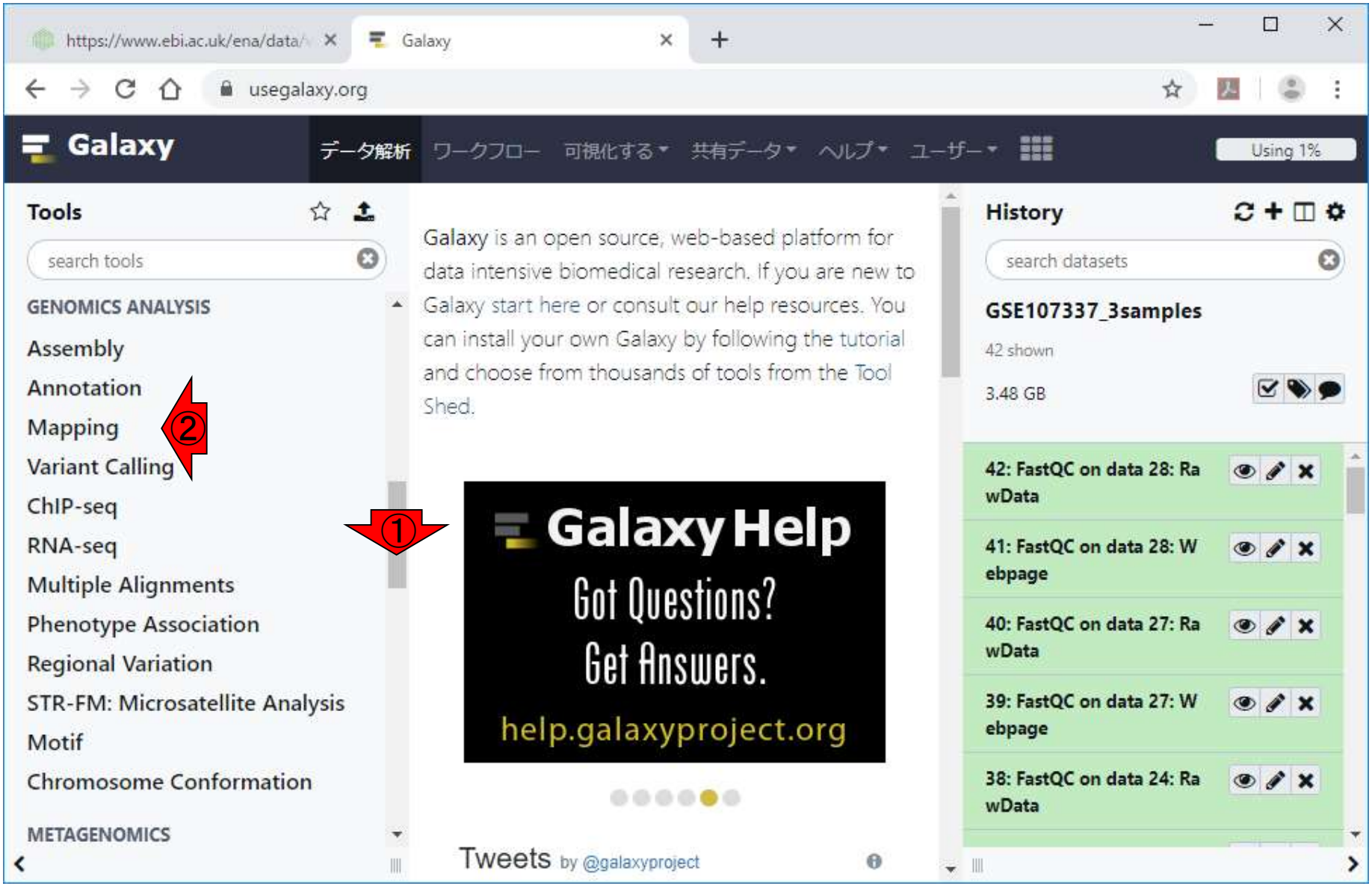


Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- **W7：ゲノム配列へのマッピング（Bowtie2）**
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

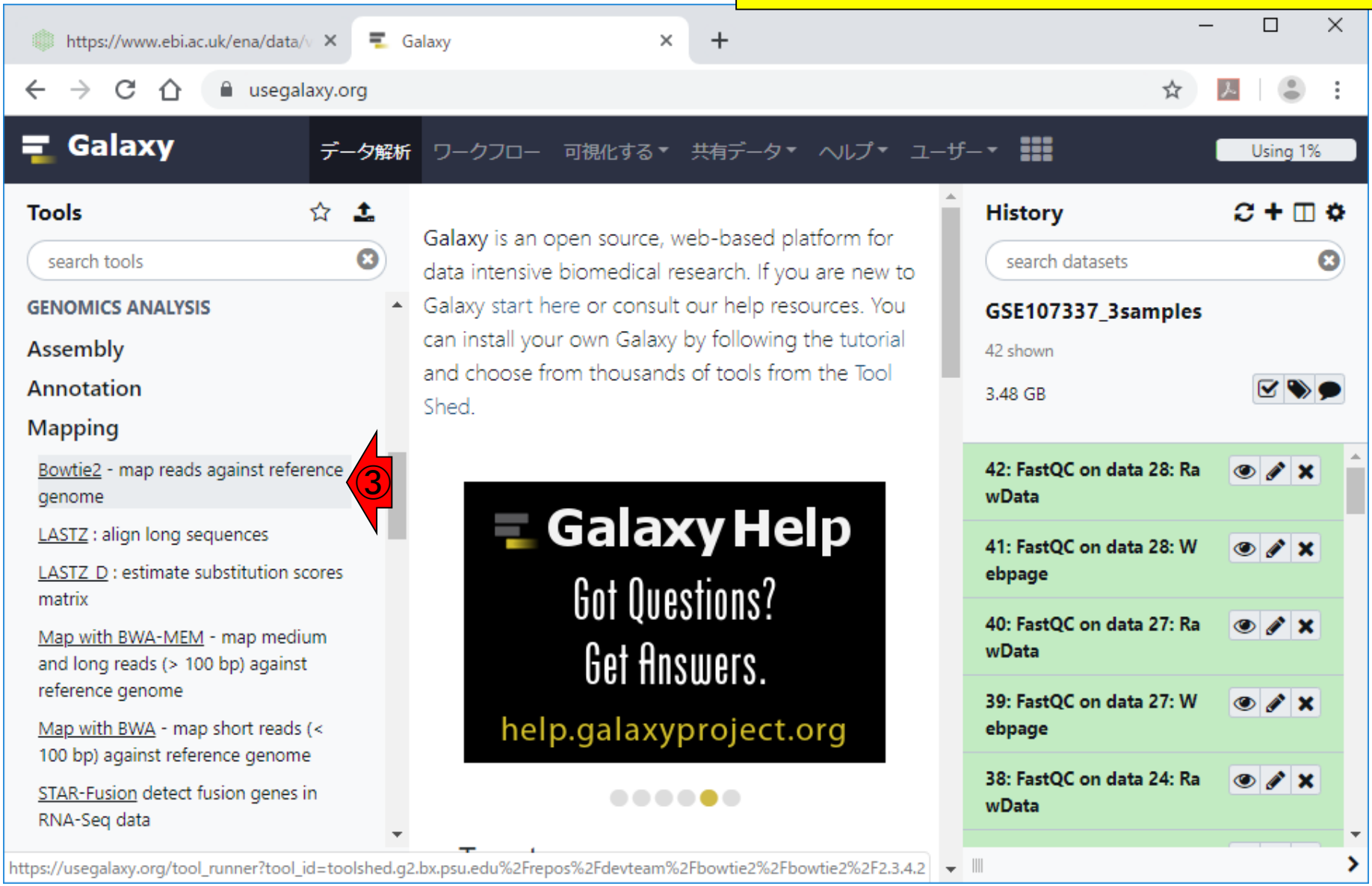
W7-1：Bowtie2

次は、Bowtie2プログラムを用いてマッピング。①ツール選択パネル下部に移動して、②Mapping。



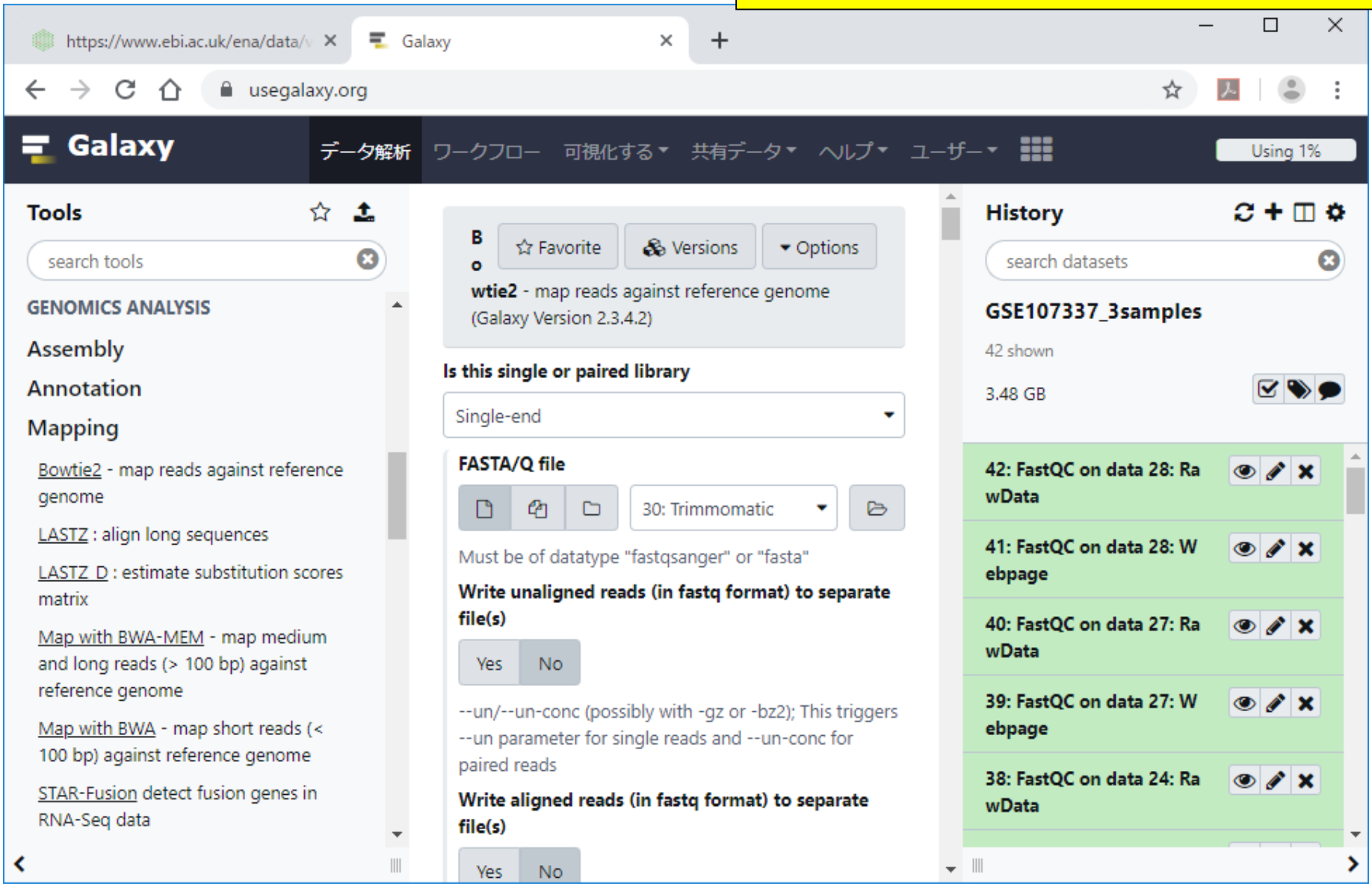
W7-1：Bowtie2

次は、Bowtie2プログラムを用いてマッピング。①ツール選択パネル下部に移動して、②Mapping。③Bowtie2。



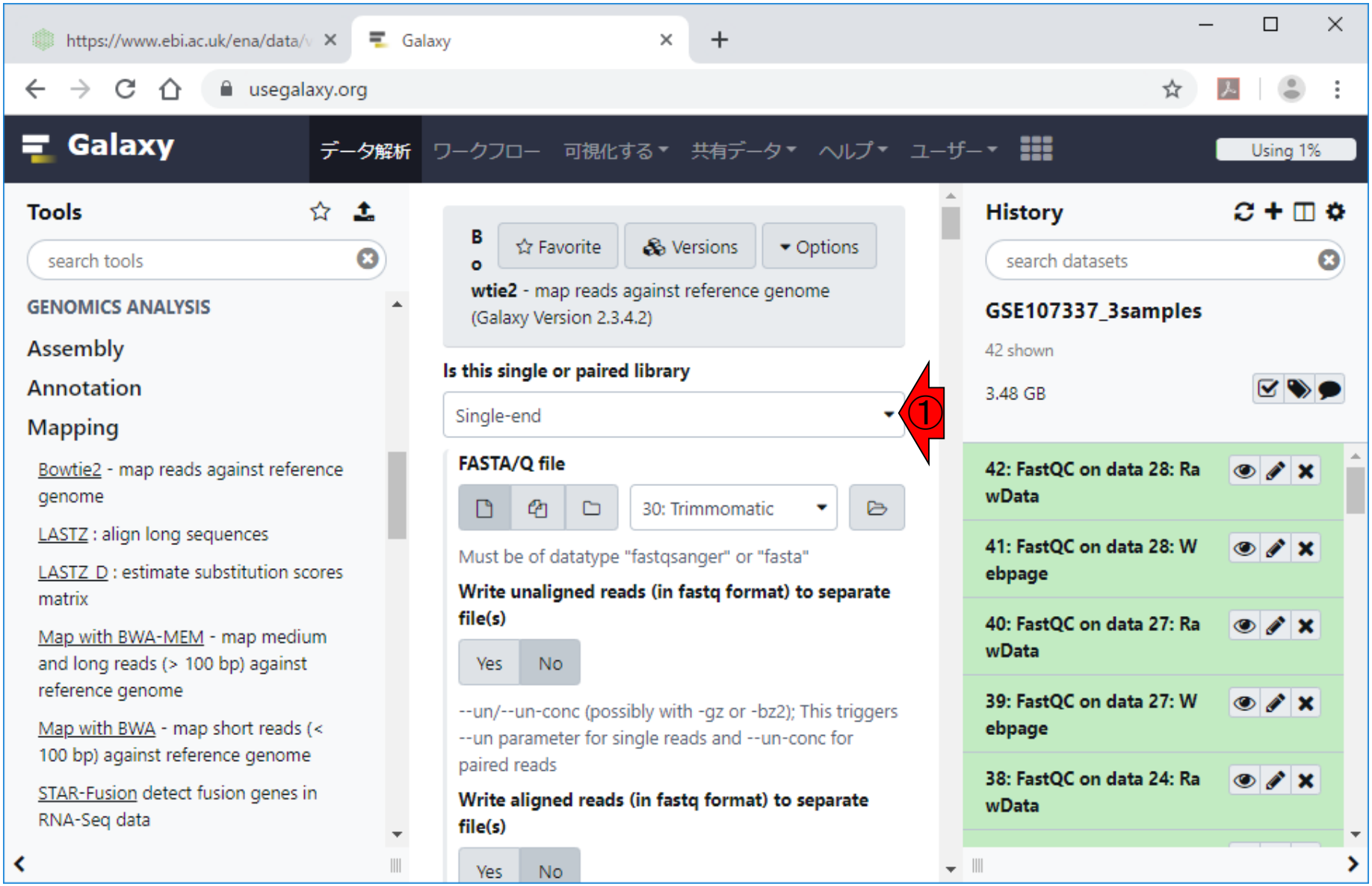
W7-1：Bowtie2

次は、Bowtie2プログラムを用いてマッピング。①ツール選択パネル下部に移動して、②Mapping。③Bowtie2。こんな感じになる。



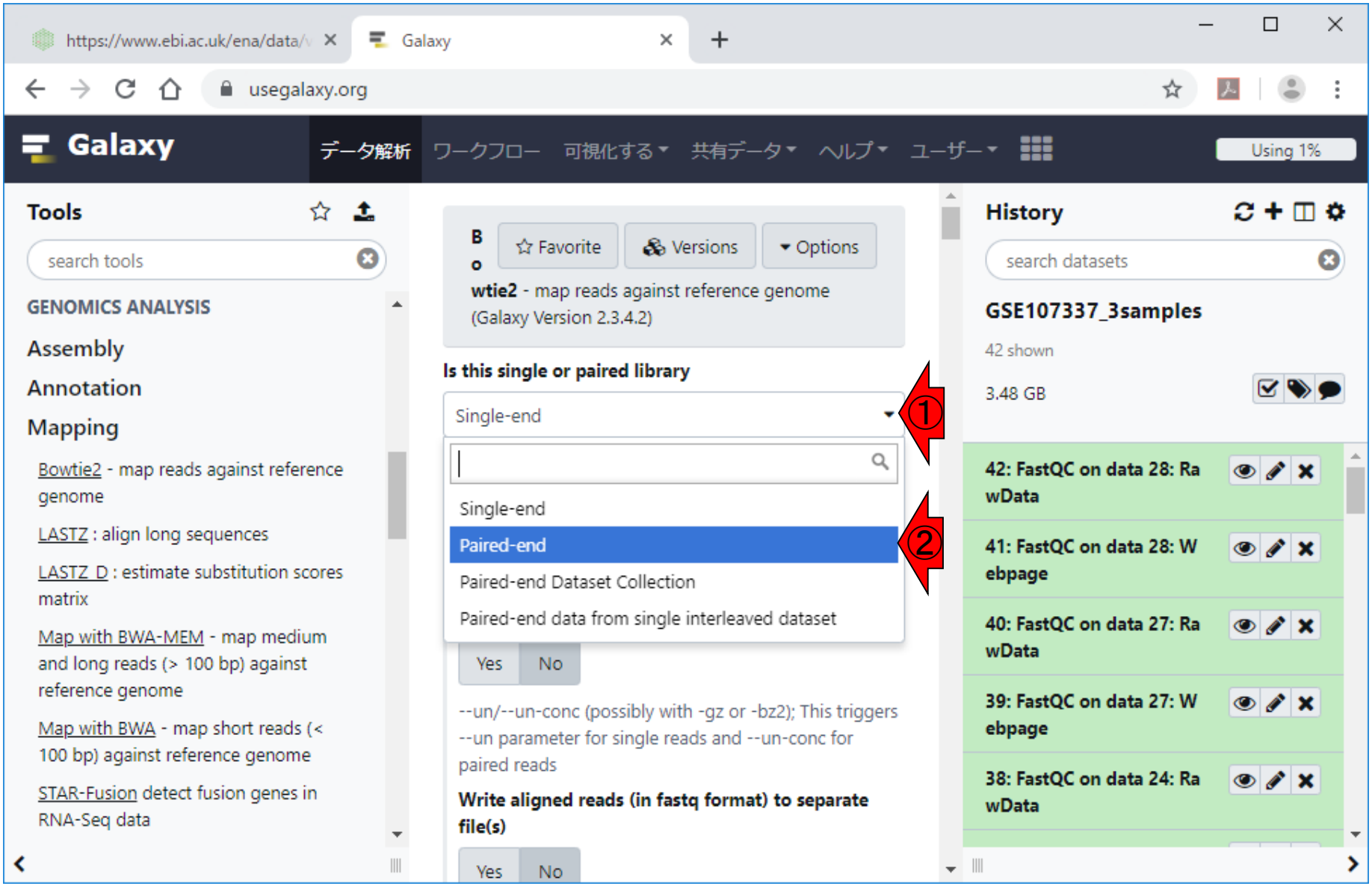
①デフォルトはSingle-endなので…。

W7-2：オプション



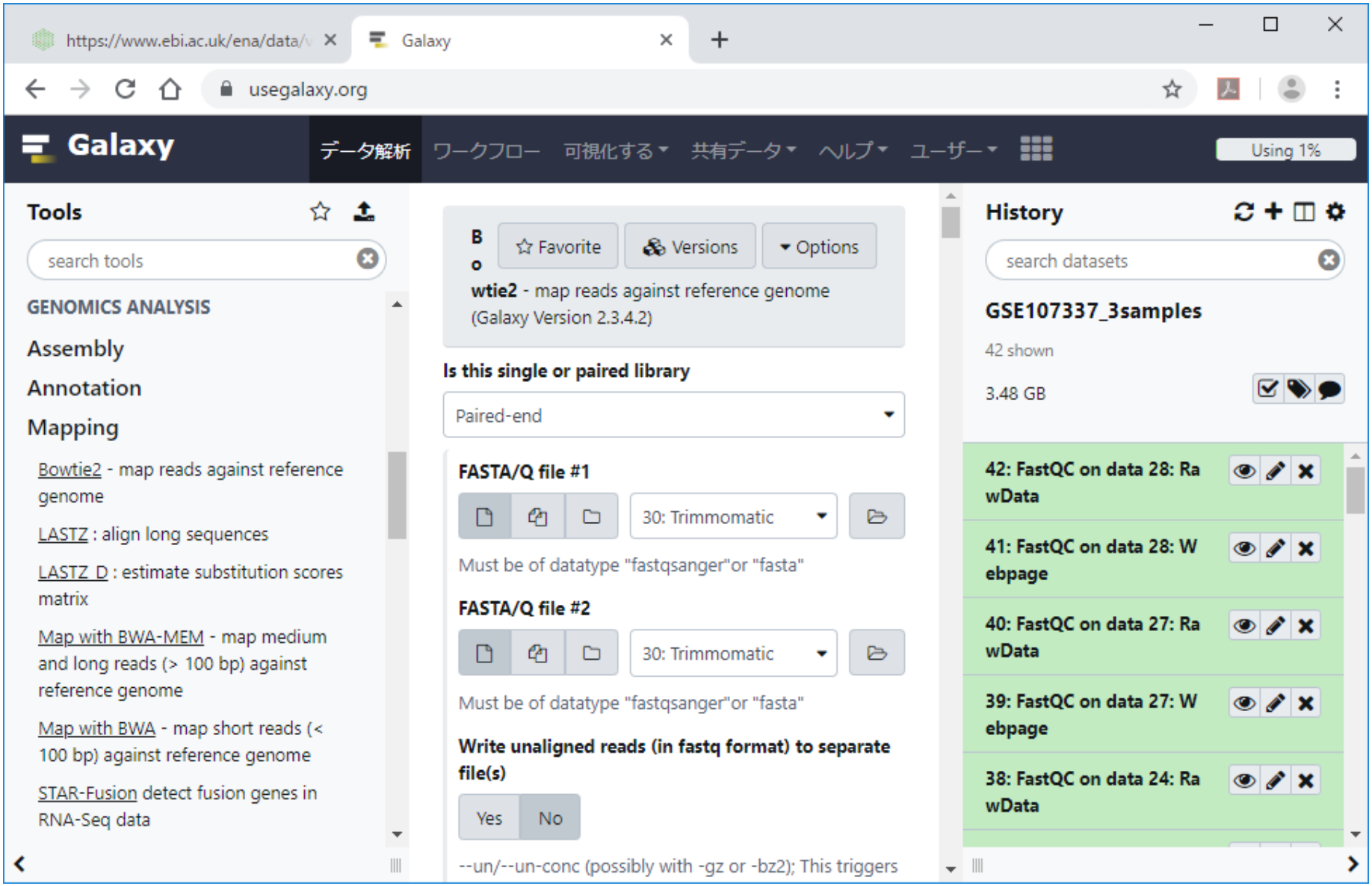
W7-2：オプション

①デフォルトはSingle-endなので･･･②Paired-endに変更。



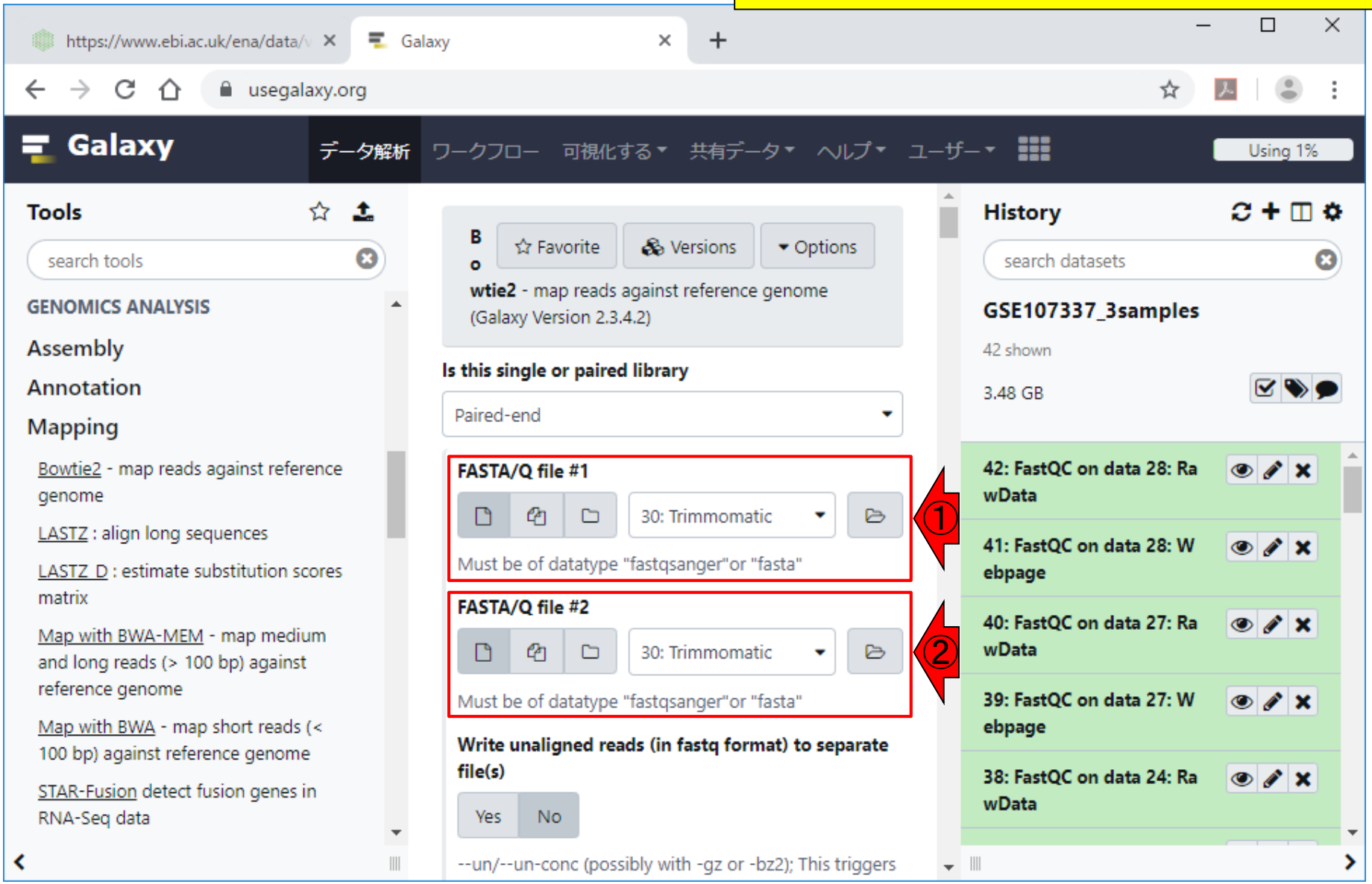
W7-2：オプション

①デフォルトはSingle-endなので･･･②Paired-endに変更。こんな感じになる。



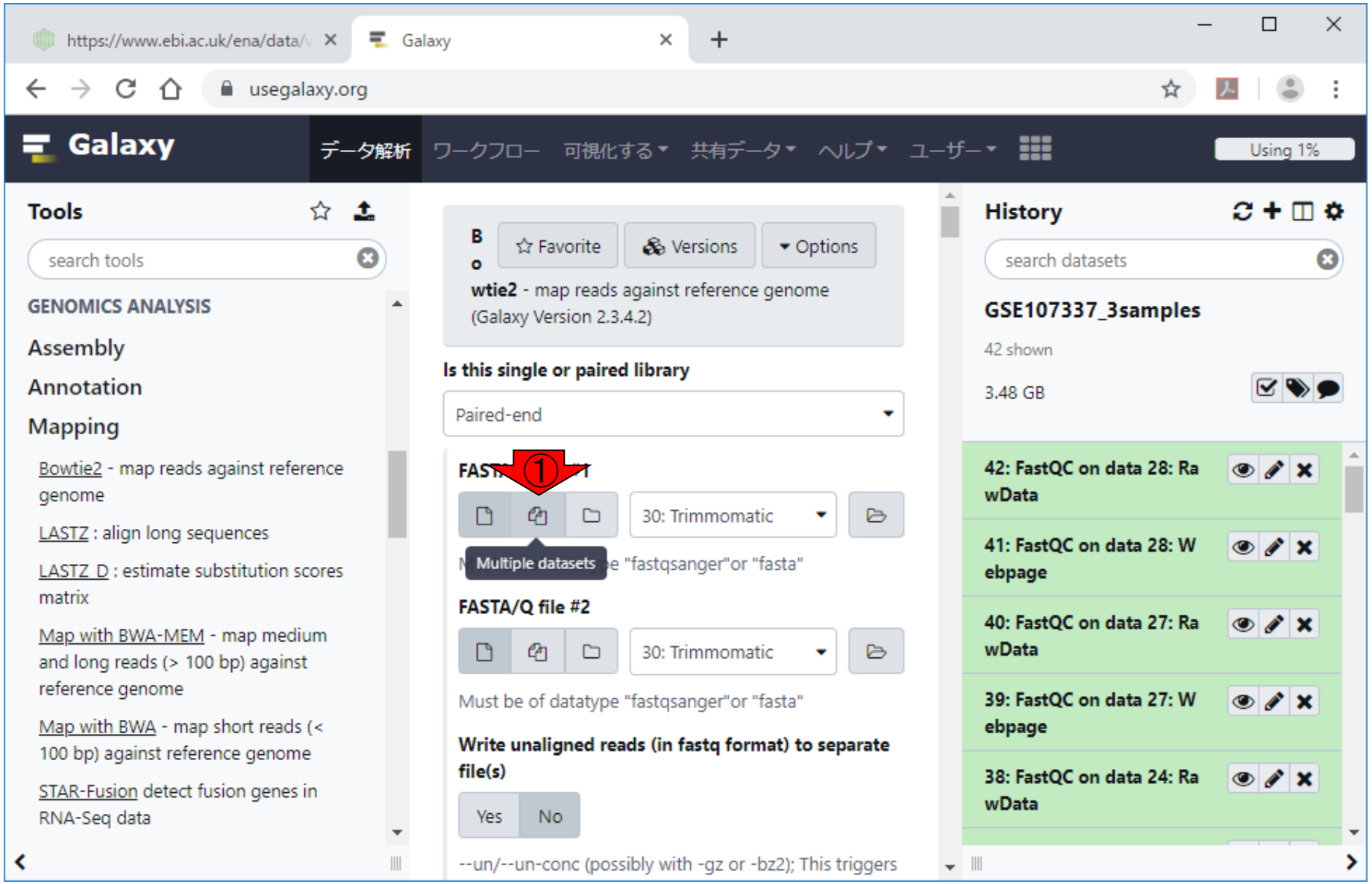
W7-2：オプション

①がpaired-endのforward側（1つめ）、②がreverse側（2つめ）のリードに相当するデータのヒストリー番号を指定するところなので。1つ1つやっていく。



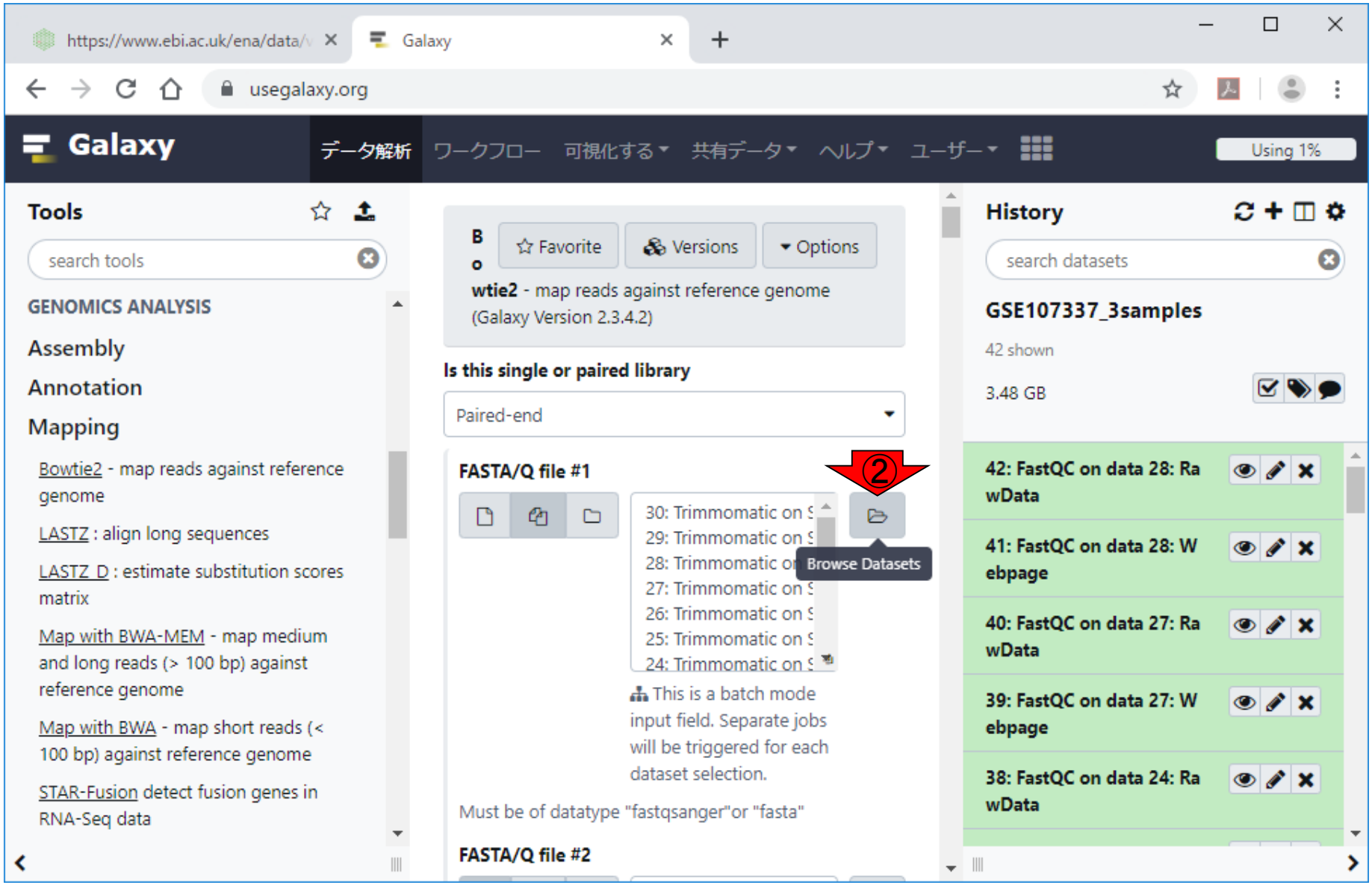
①Multiple datasets。

W7-3：forward側



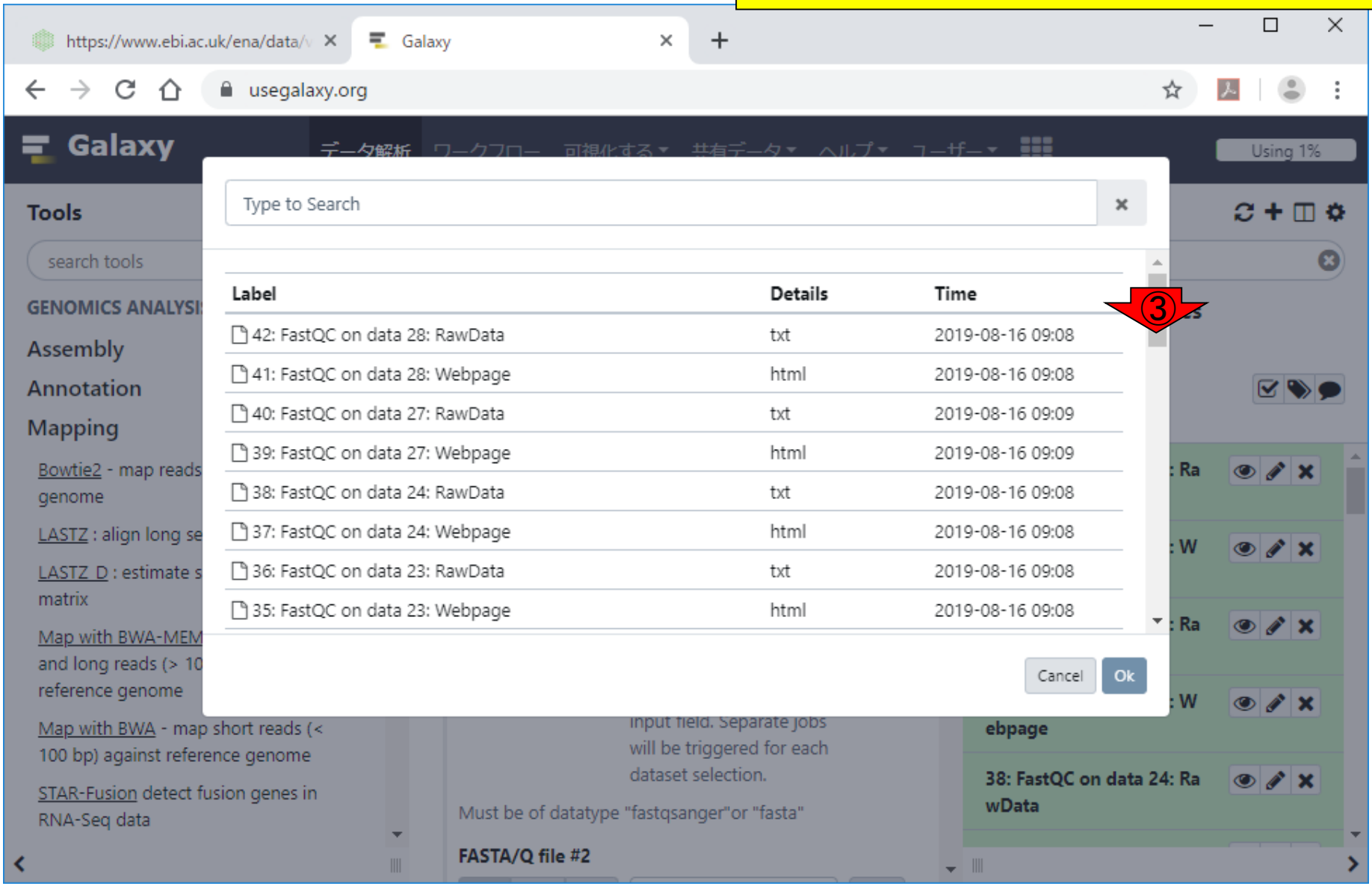
①Multiple datasets。②Browse Datasets。

W7-3：forward側



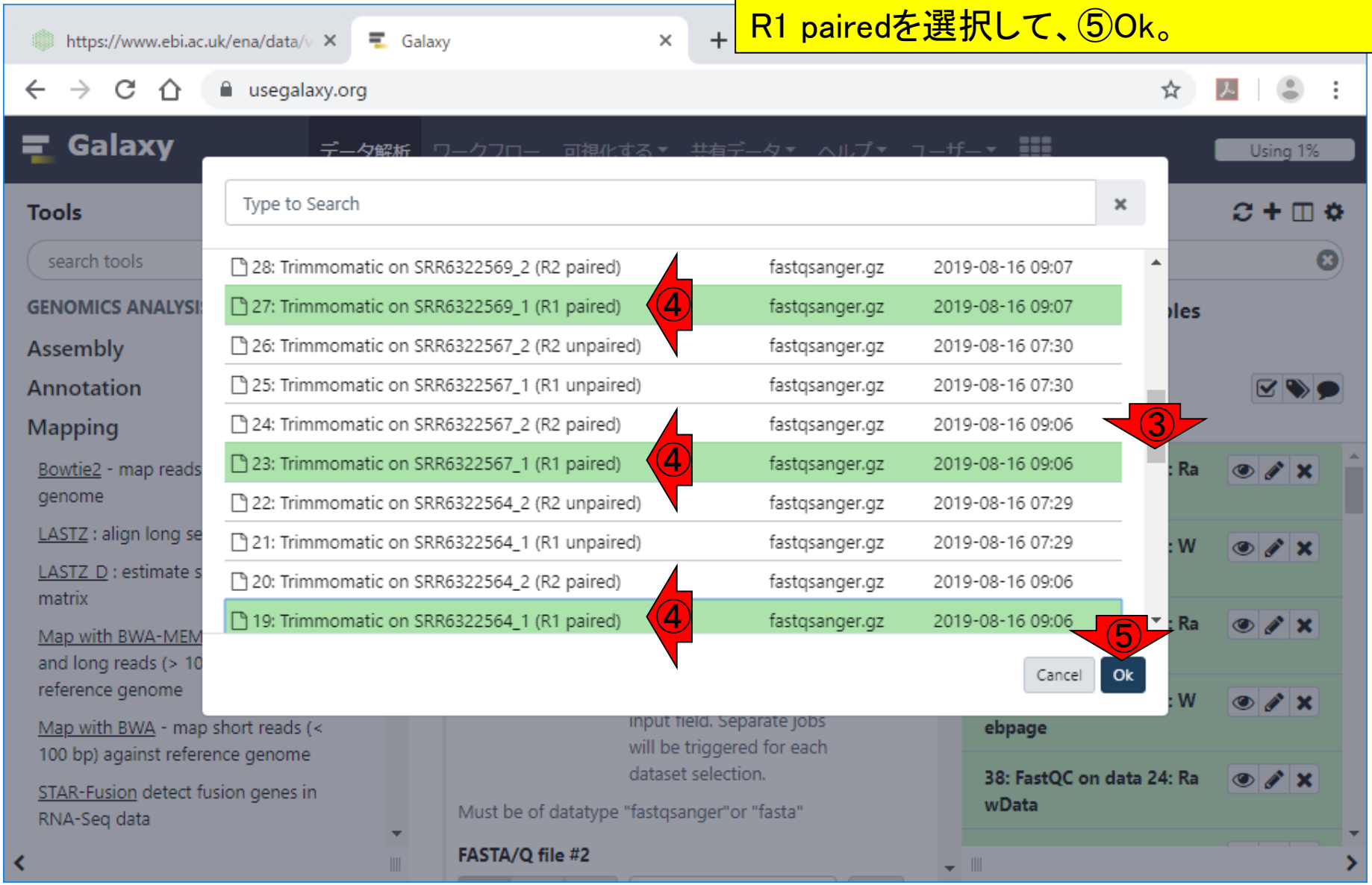
W7-3：forward側

①Multiple datasets。②Browse Datasets。こんな感じになる。③下部に移動して、入力として与えたいTrimmomatic実行後のデータを探す。



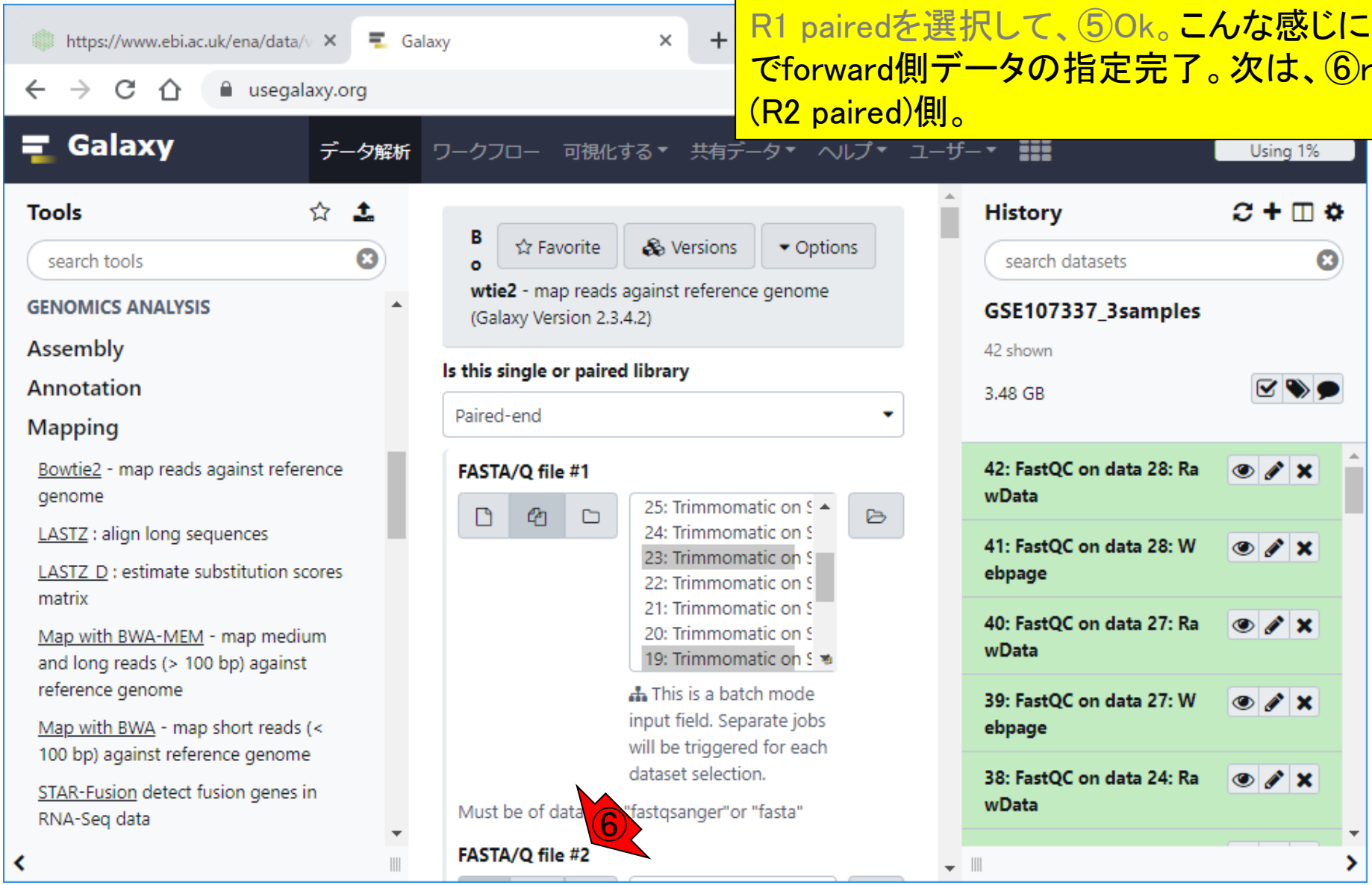
W7-3：forward側

①Multiple datasets。②Browse Datasets。こんな感じになる。③下部に移動して、入力として与えたいTrimmomatic実行後のデータを探す。④forward側のR1 pairedを選択して、⑤Ok。



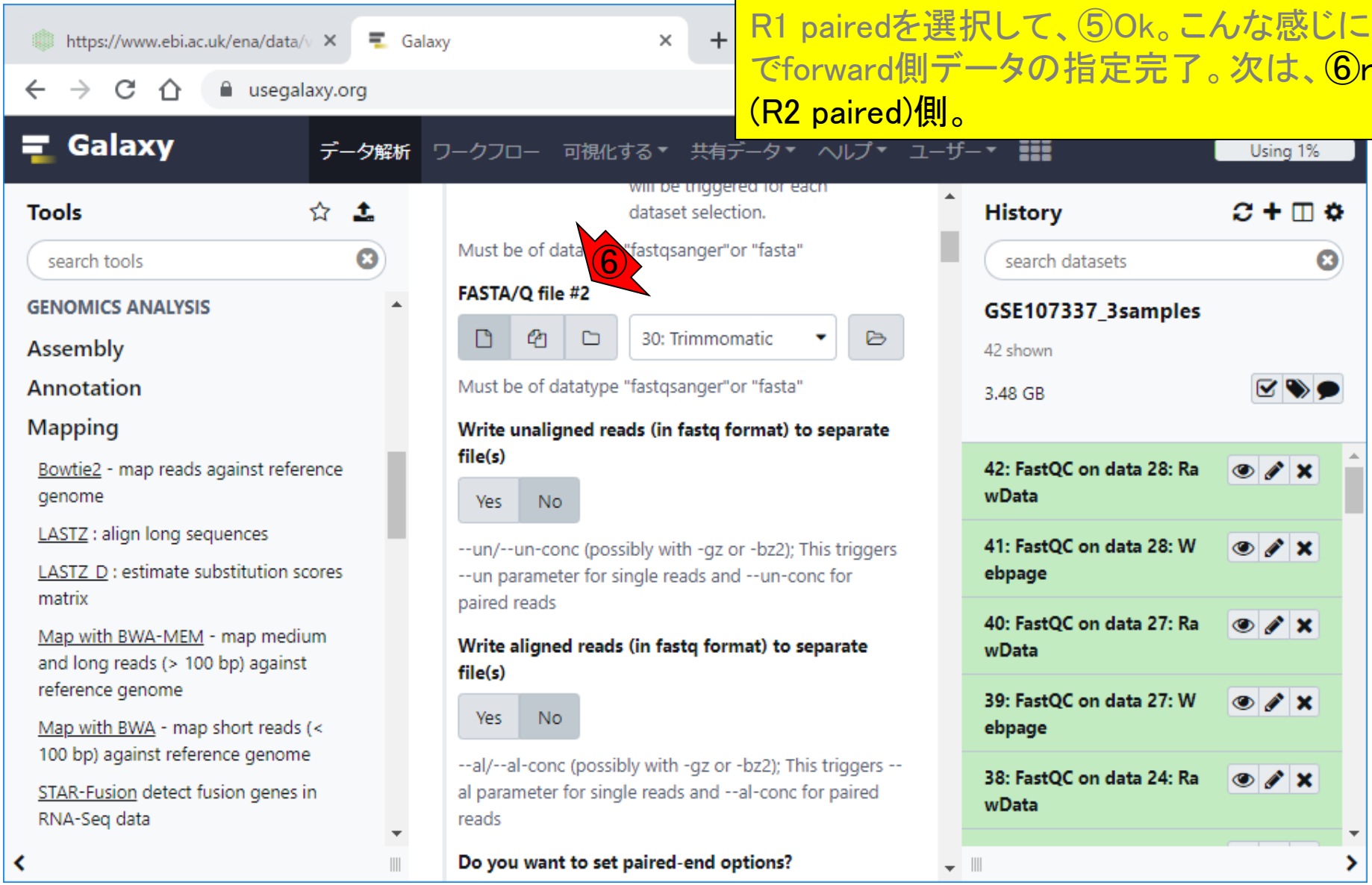
W7-3：forward側

①Multiple datasets。②Browse Datasets。こんな感じになる。③下部に移動して、入力として与えたいTrimmomatic実行後のデータを探す。④forward側のR1 pairedを選択して、⑤Ok。こんな感じになる。これでforward側データの指定完了。次は、⑥reverse(R2 paired)側。



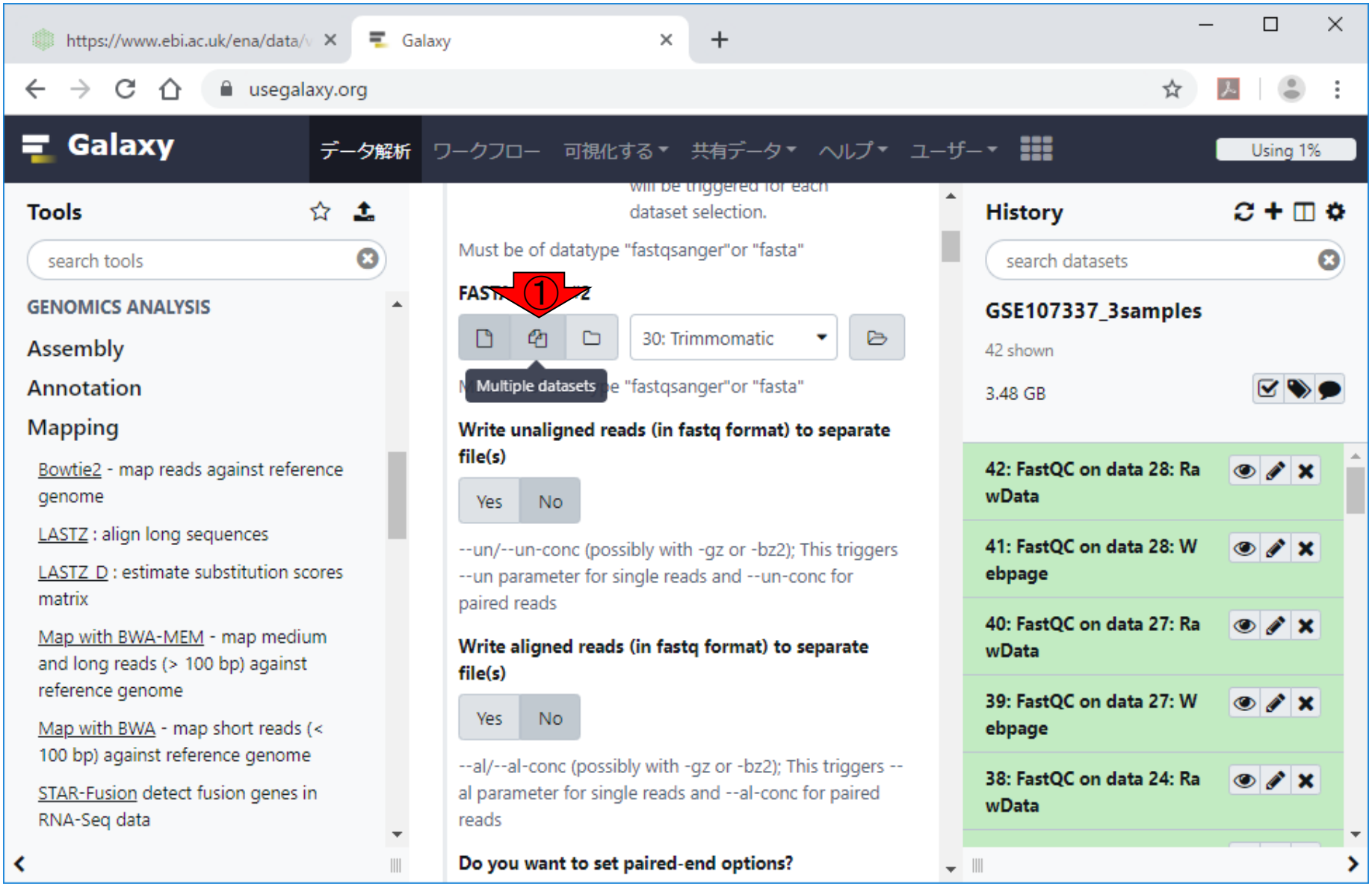
W7-4：reverse側

①Multiple datasets。②Browse Datasets。こんな感じになる。③下部に移動して、入力として与えたいTrimmomatic実行後のデータを探す。④forward側のR1 pairedを選択して、⑤Ok。こんな感じになる。これでforward側データの指定完了。次は、⑥reverse(R2 paired)側。



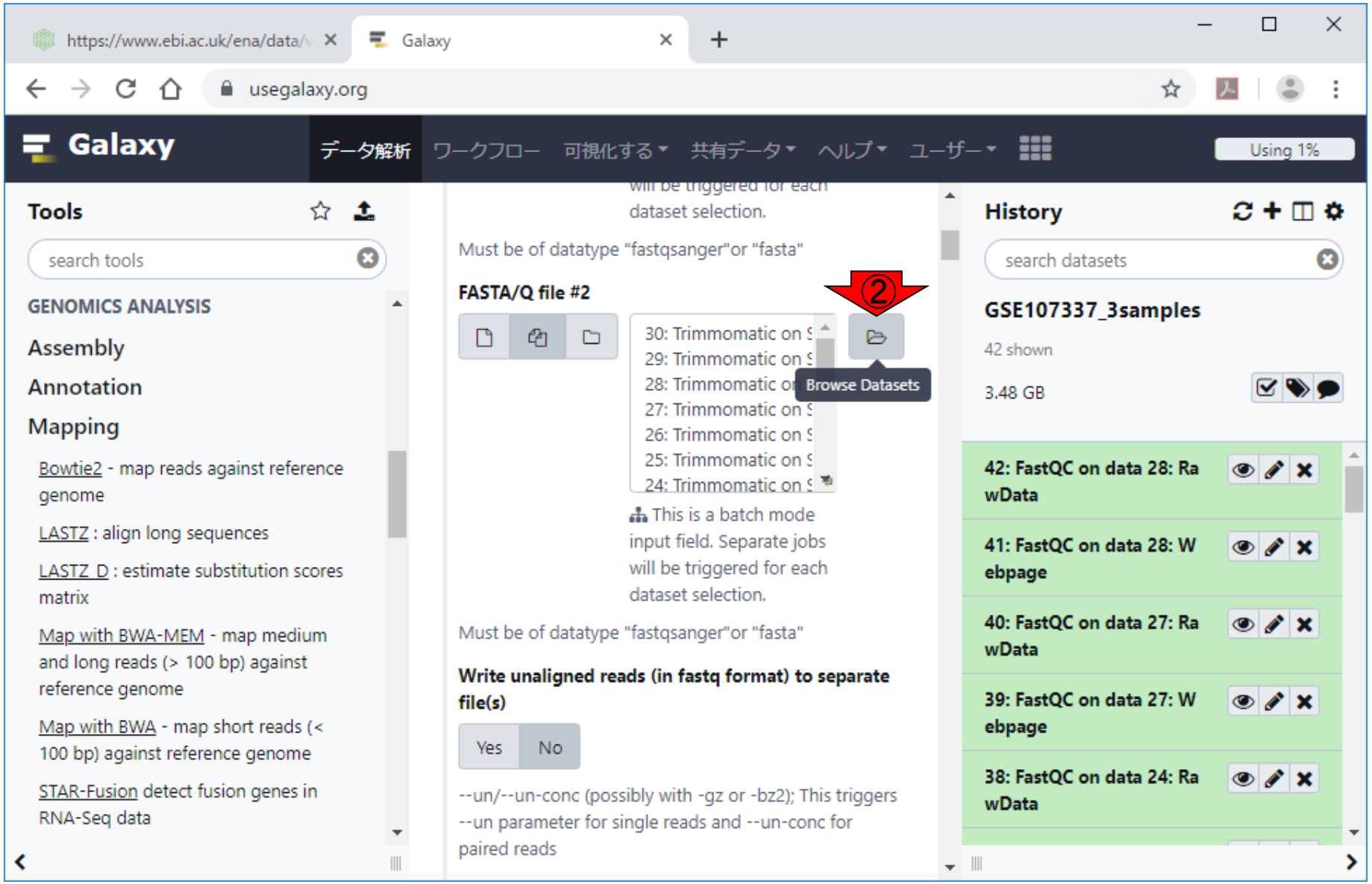
①Multiple datasets。

W7-4：reverse側



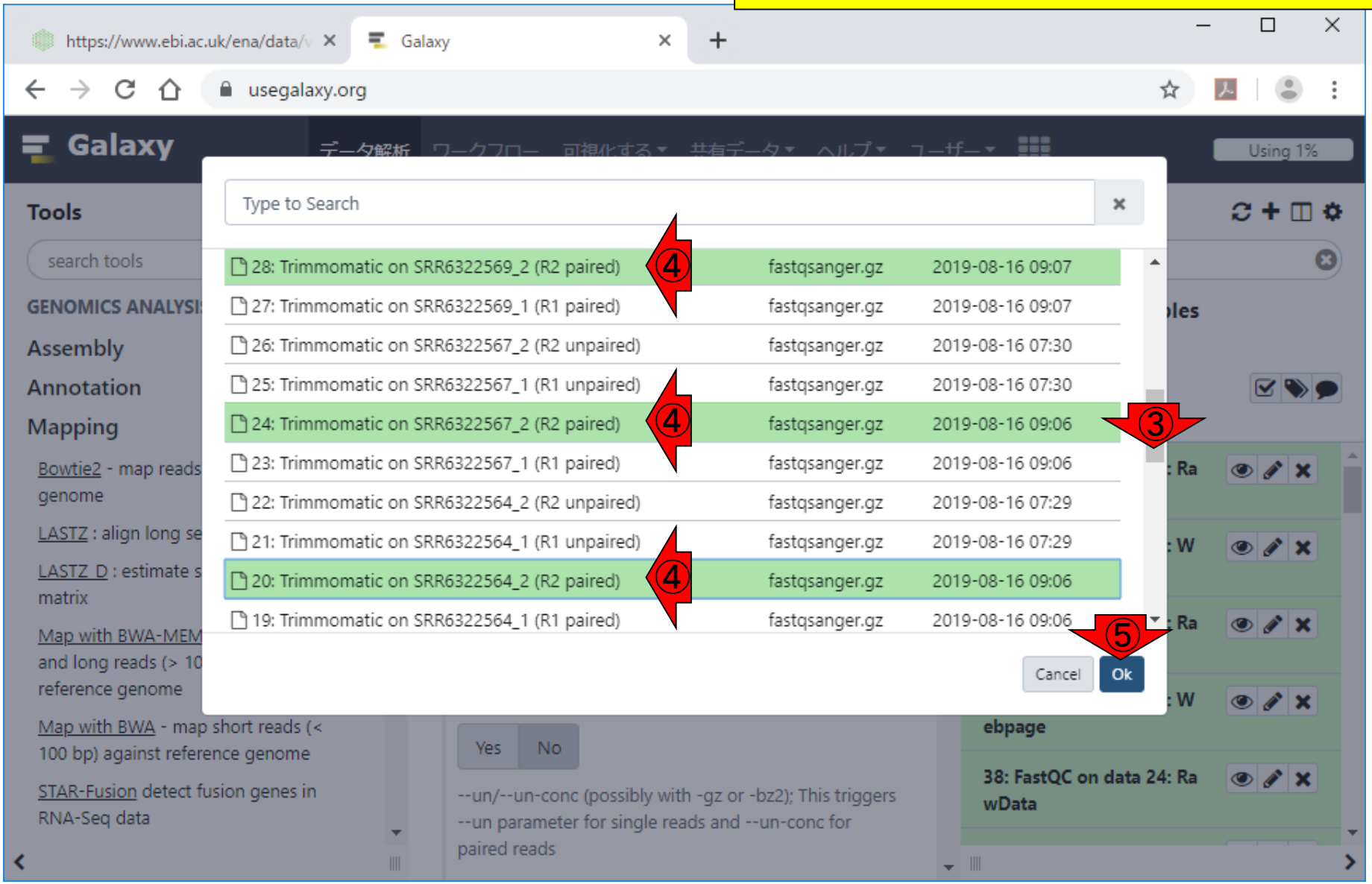
①Multiple datasets。②Browse Datasets。

W7-4：reverse側



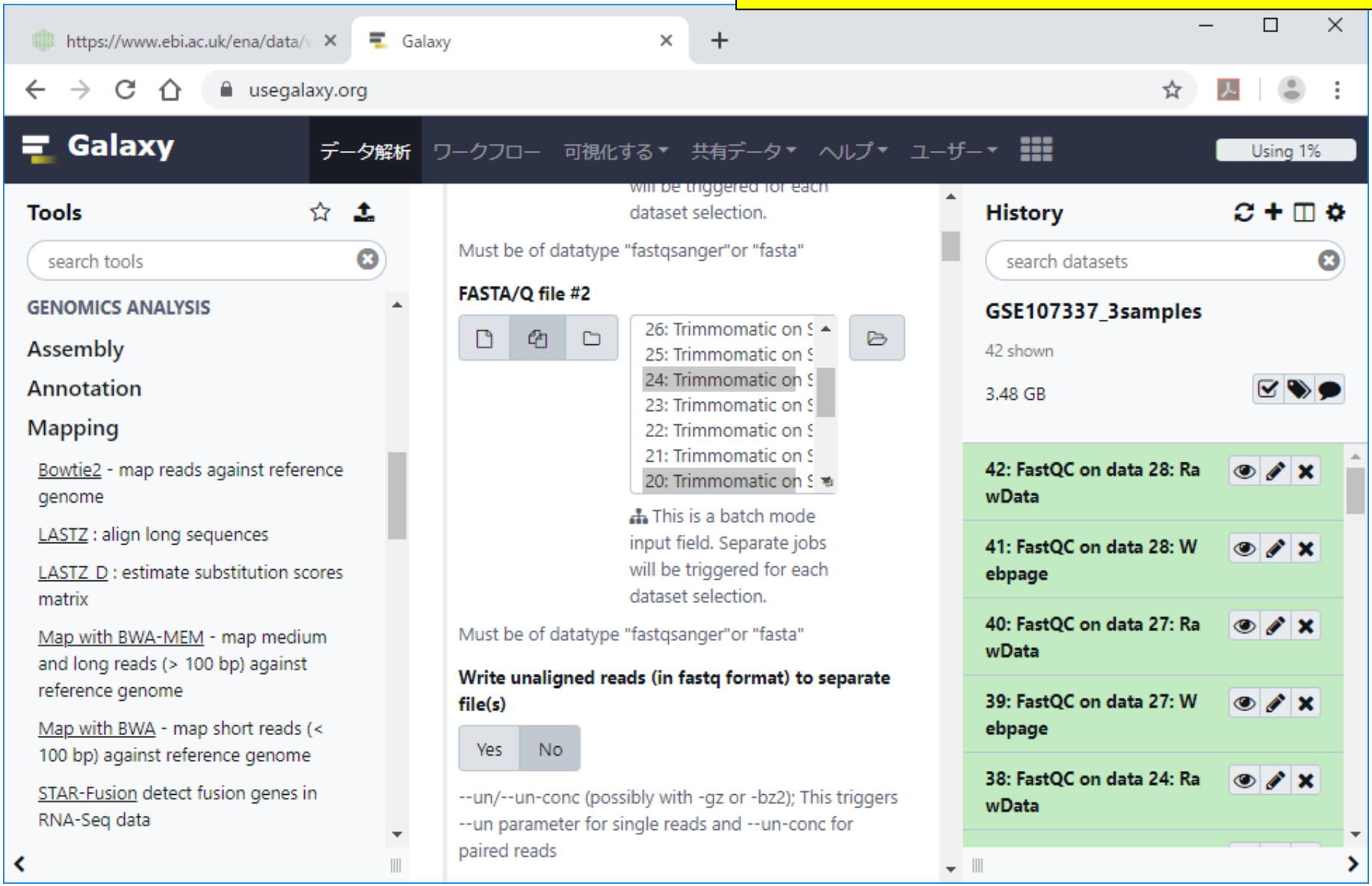
W7-4：reverse側

①Multiple datasets。②Browse Datasets。③下部に移動して、④reverse側のR2 pairedを選択して、⑤Ok。



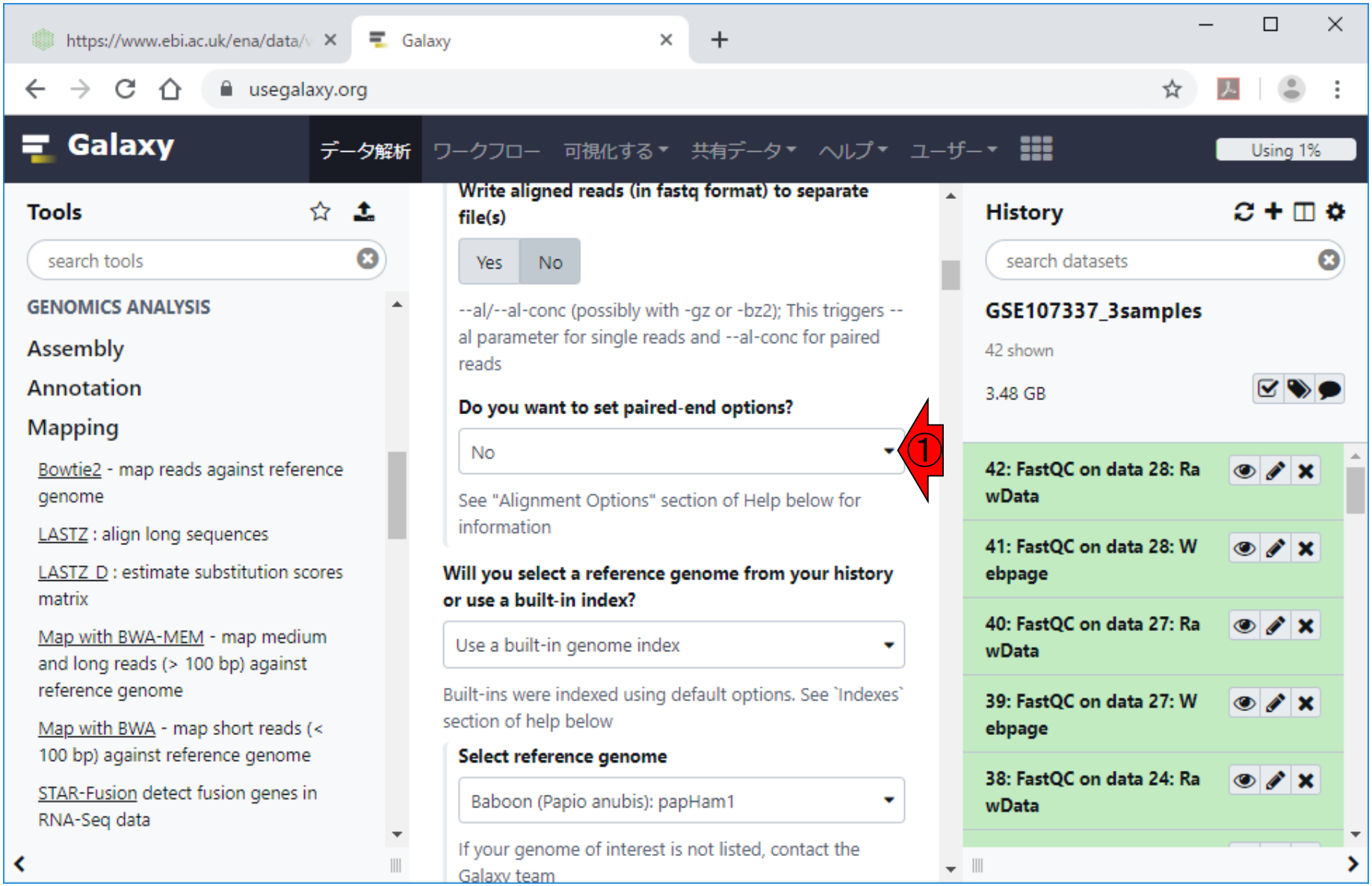
W7-4：reverse側

①Multiple datasets。②Browse Datasets。③下部に移動して、④reverse側のR2 pairedを選択して、⑤Ok。こんな感じになる。



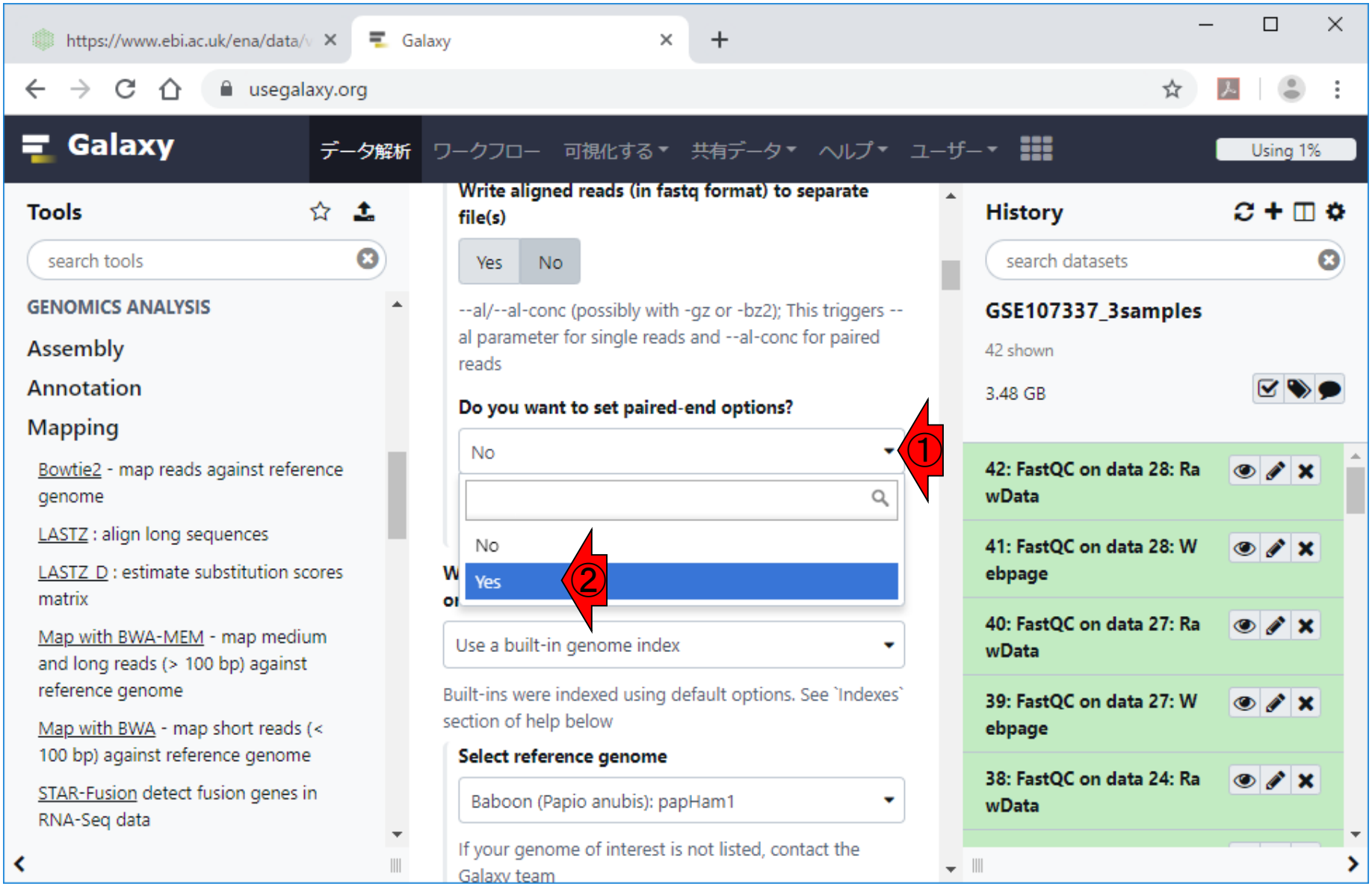
W7-5：オプション

他のオプションとして、①paired-end optionsがNoになっているので…



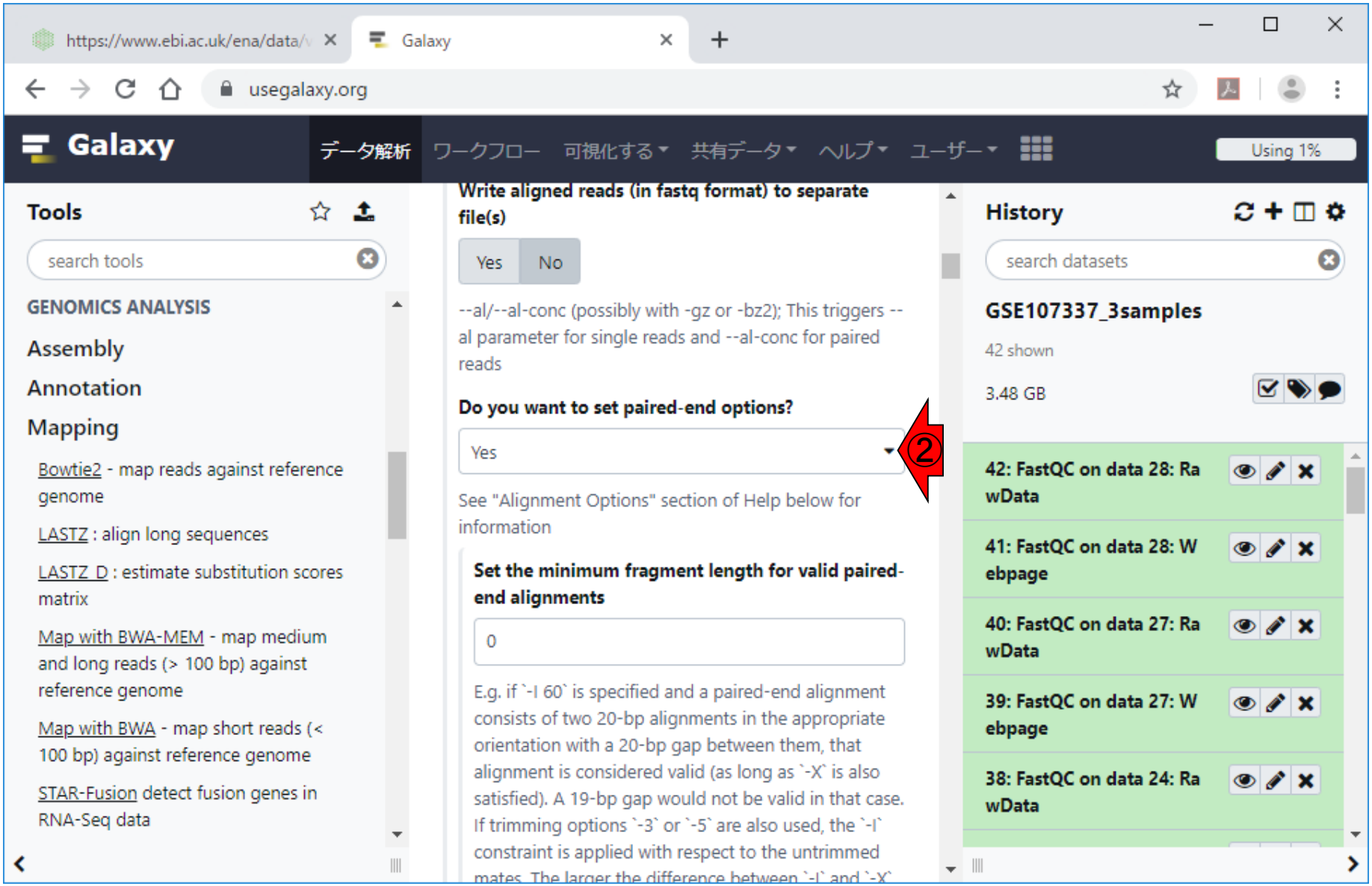
W7-5：オプション

他のオプションとして、①paired-end optionsがNoになっているので…②Yesに変更しておく。



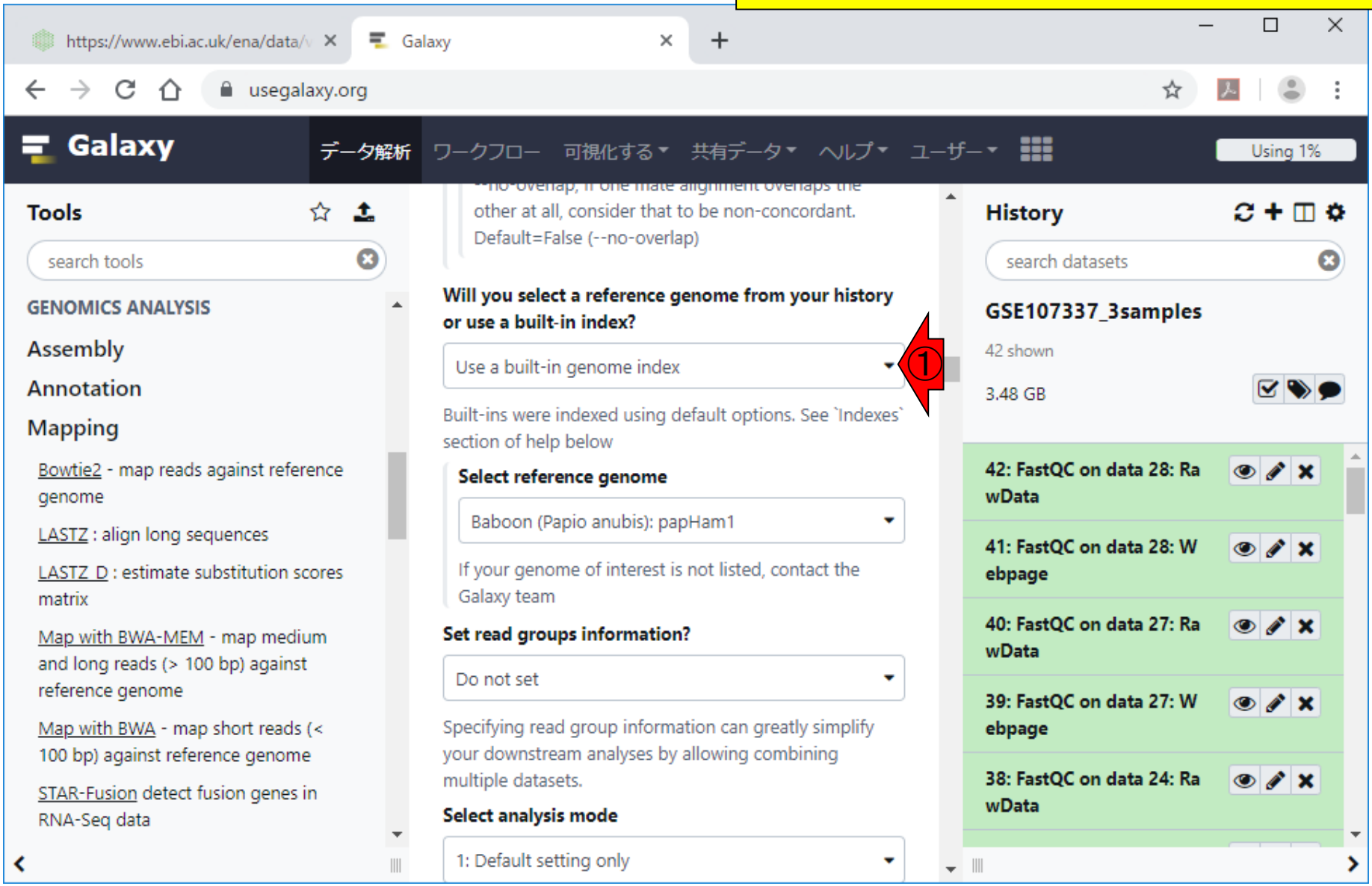
W7-5：オプション

他のオプションとして、①paired-end optionsがNoになっているので…②Yesに変更しておく。変更後。



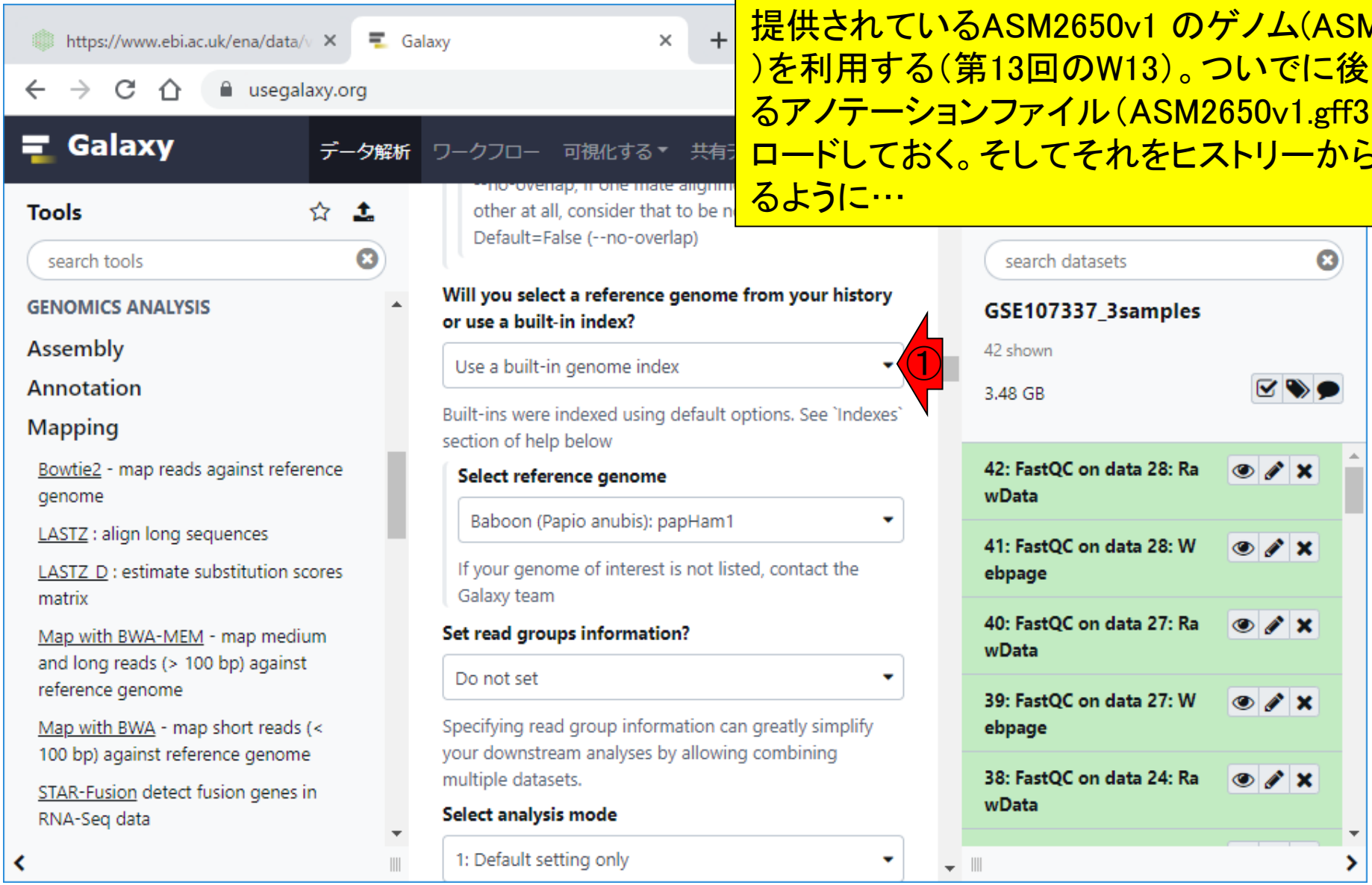
W7-6：リファレンス

中央パネル下部に移動して、リファレンスゲノム配列を指定するところ。①デフォルトはGalaxy上で指定可能なもののみ。



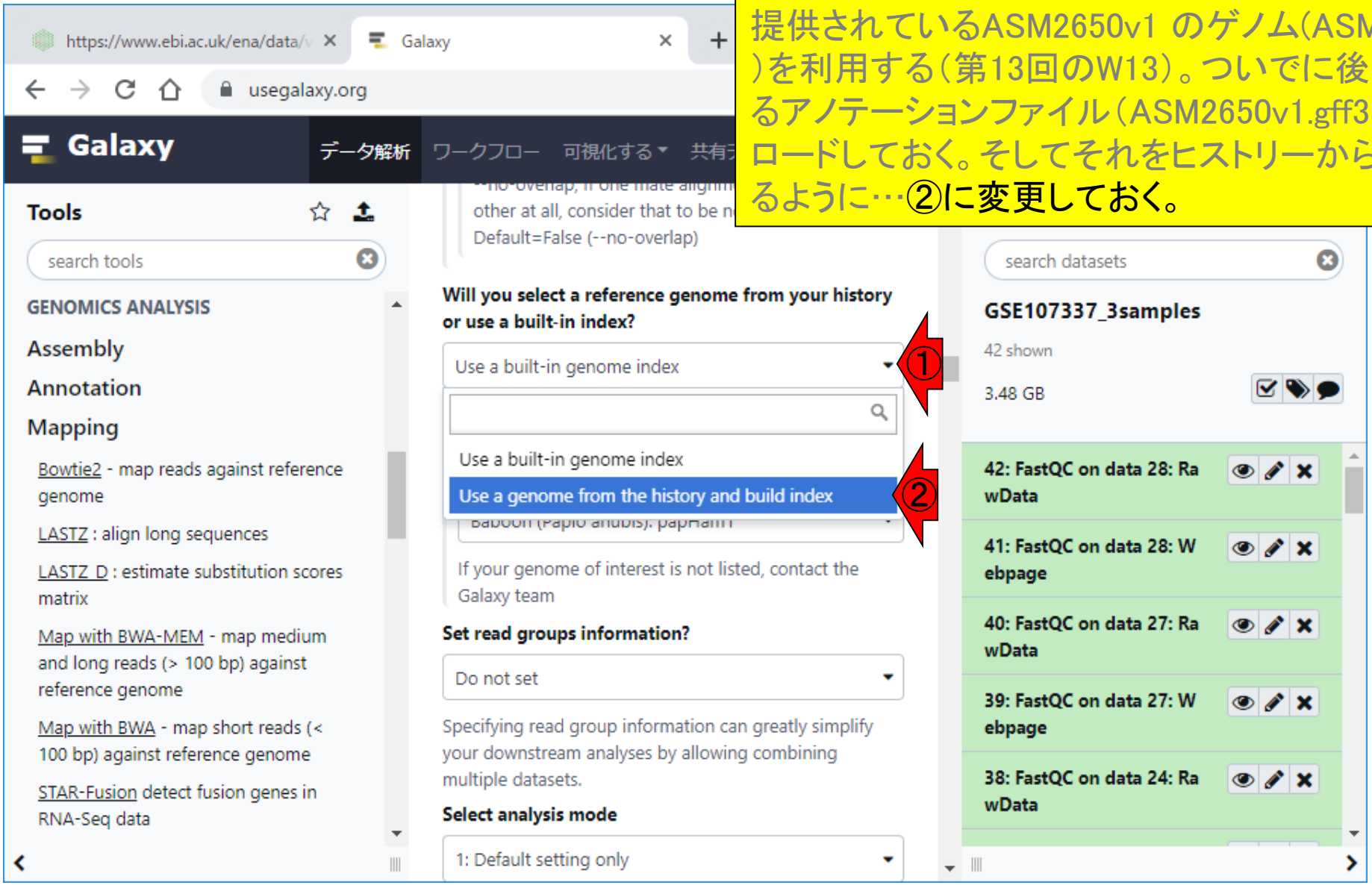
W7-6：リファレンス

中央パネル下部に移動して、リファレンスゲノム配列を指定するところ。①デフォルトはGalaxy上で指定可能なもののみ。ここでは、Ensembl Bacteria から提供されているASM2650v1 のゲノム(ASM2650v1.fa)を利用する（第13回のW13）。ついでに後に利用するアノテーションファイル（ASM2650v1.gff3）もアップロードしておく。そしてそれをヒストリーから選択できるように…



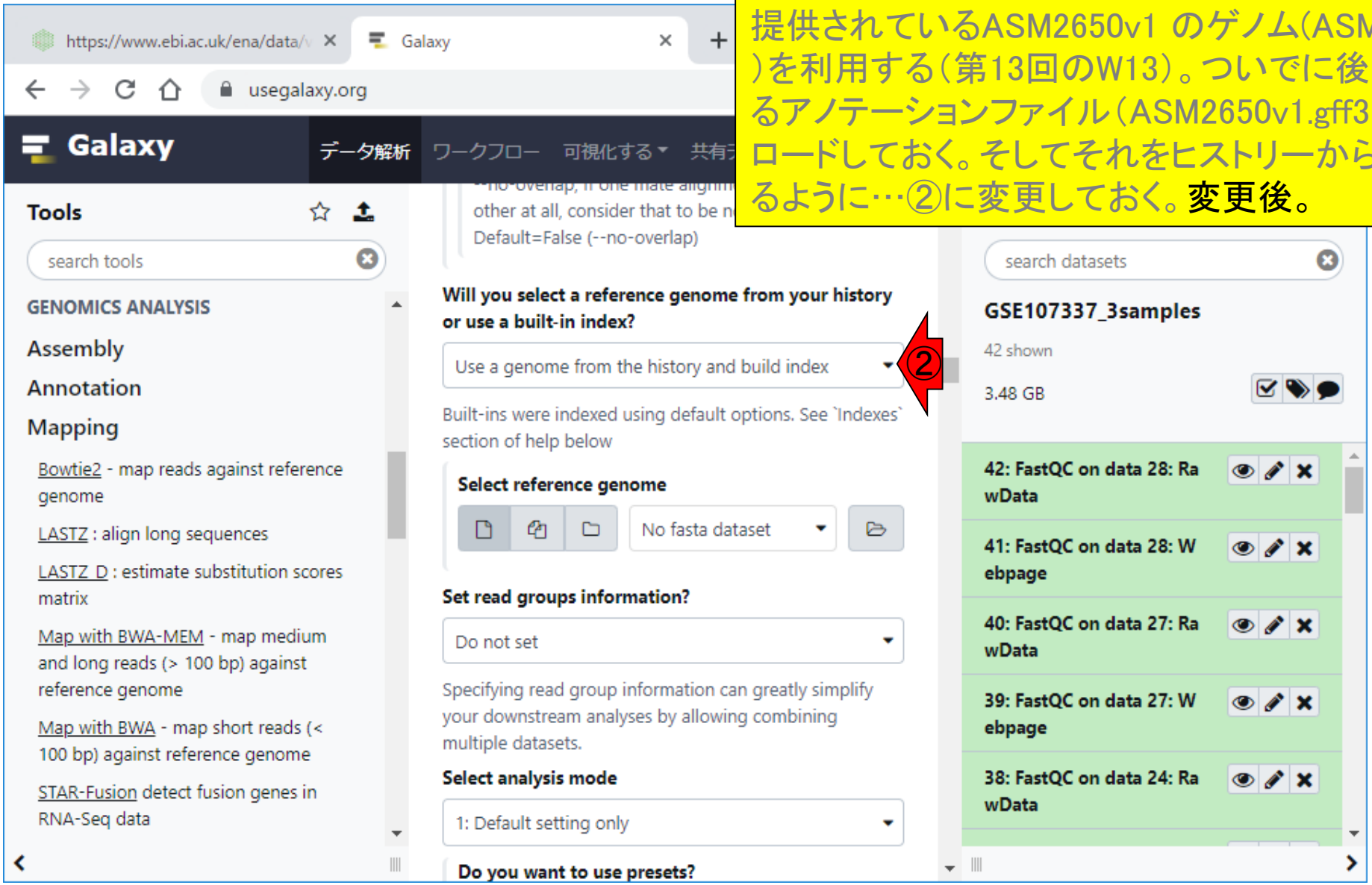
W7-6：リファレンス

中央パネル下部に移動して、リファレンスゲノム配列を指定するところ。①デフォルトはGalaxy上で指定可能なもののみ。ここでは、Ensembl Bacteria から提供されているASM2650v1 のゲノム(ASM2650v1.fa)を利用する（第13回のW13）。ついでに後に利用するアノテーションファイル（ASM2650v1.gff3）もアップロードしておく。そしてそれをヒストリーから選択できるように…②に変更しておく。



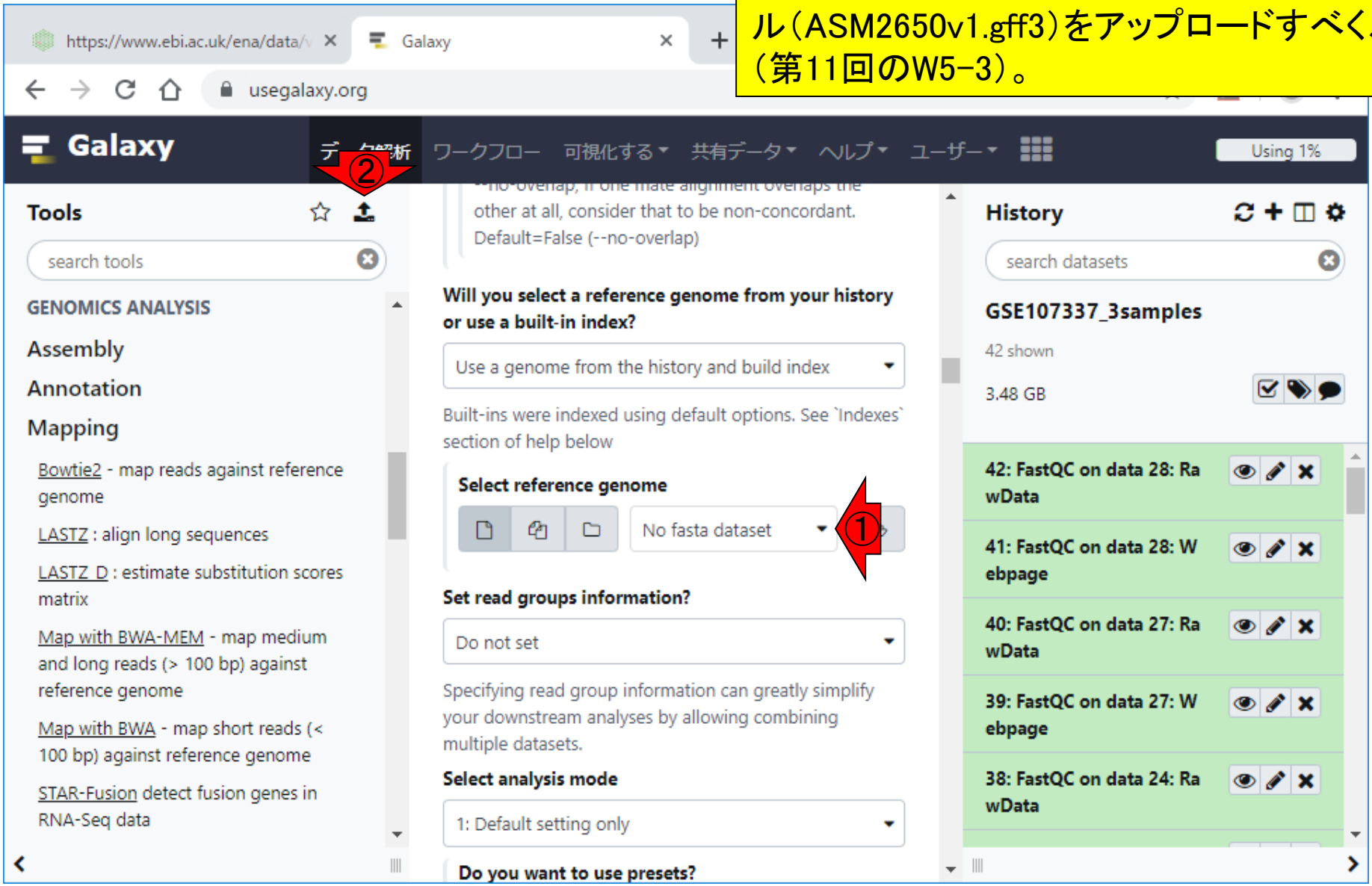
W7-6：リファレンス

中央パネル下部に移動して、リファレンスゲノム配列を指定するところ。①デフォルトはGalaxy上で指定可能なもののみ。ここでは、Ensembl Bacteria から提供されているASM2650v1 のゲノム(ASM2650v1.fa)を利用する（第13回のW13）。ついでに後に利用するアノテーションファイル（ASM2650v1.gff3）もアップロードしておく。そしてそれをヒストリーから選択できるように…②に変更しておく。**変更後。**



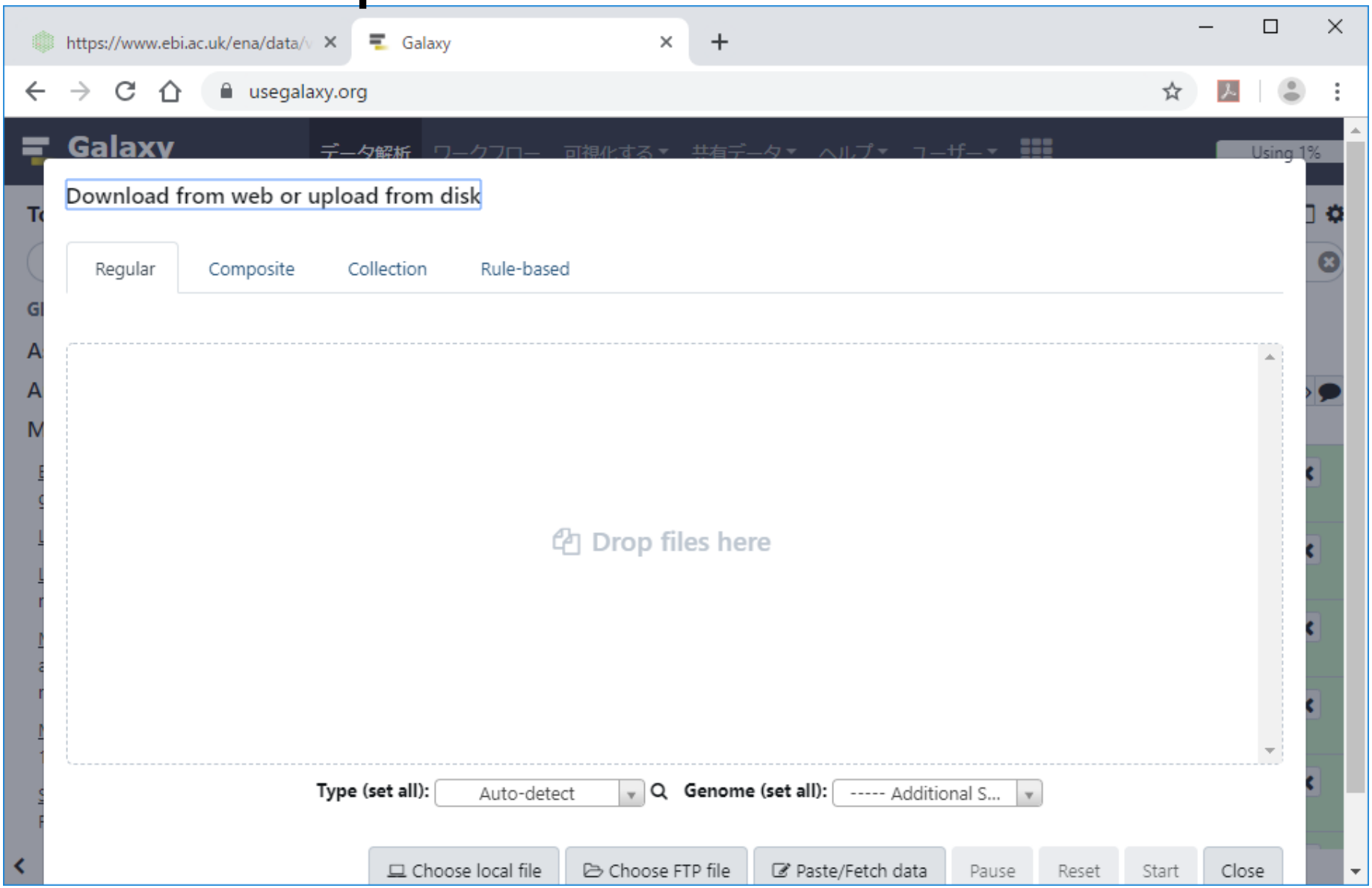
W7-6：リファレンス

まだヒストリーから選択可能な状態ではないので、①No fasta datasetとなっている。ASM2650v1 のゲノムファイル(ASM2650v1.fa)とアノテーションファイル（ASM2650v1.gff3）をアップロードすべく、②を押す（第11回のW5-3）。



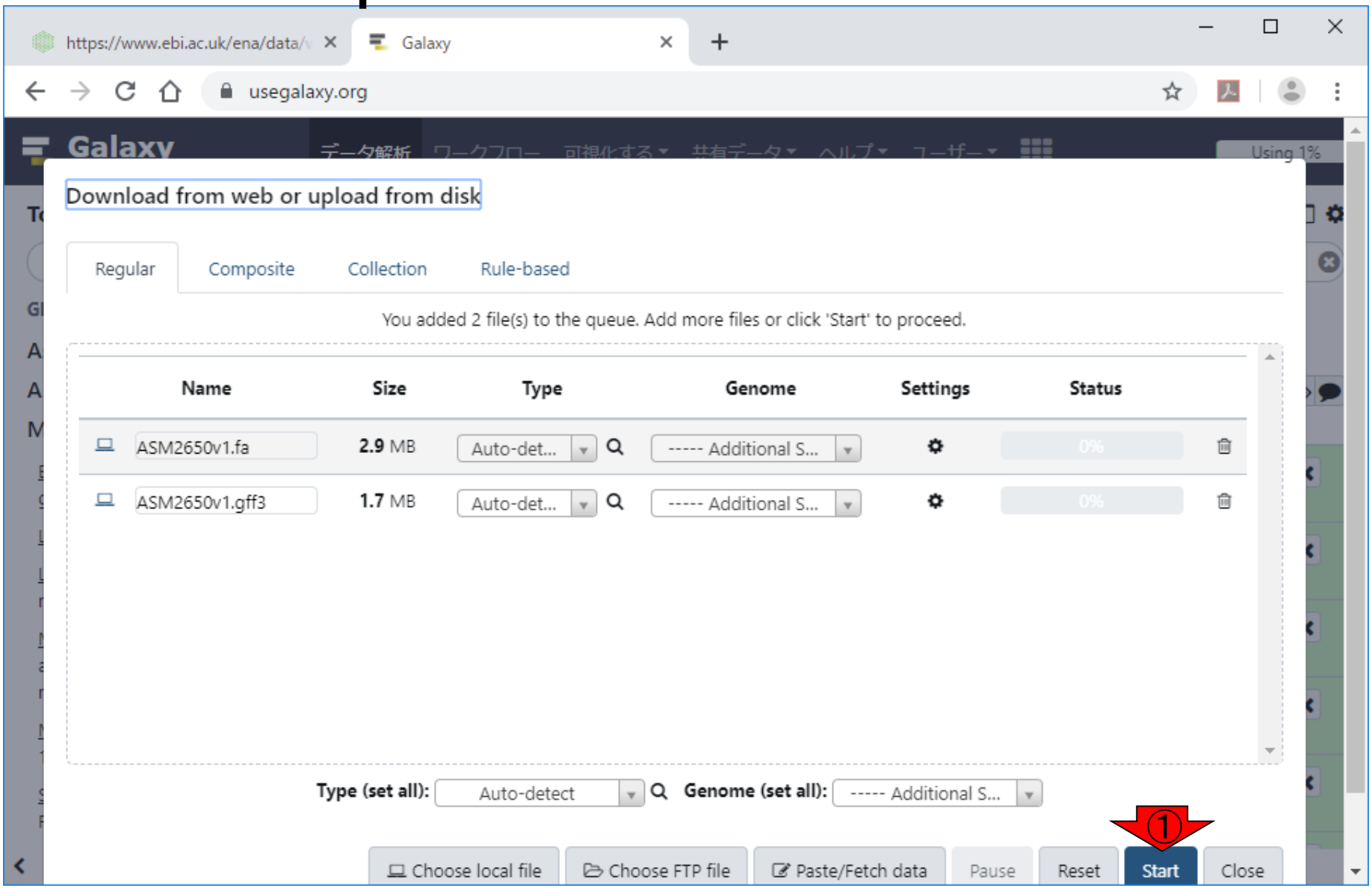
こんな感じになる。

W7-7 : Upload



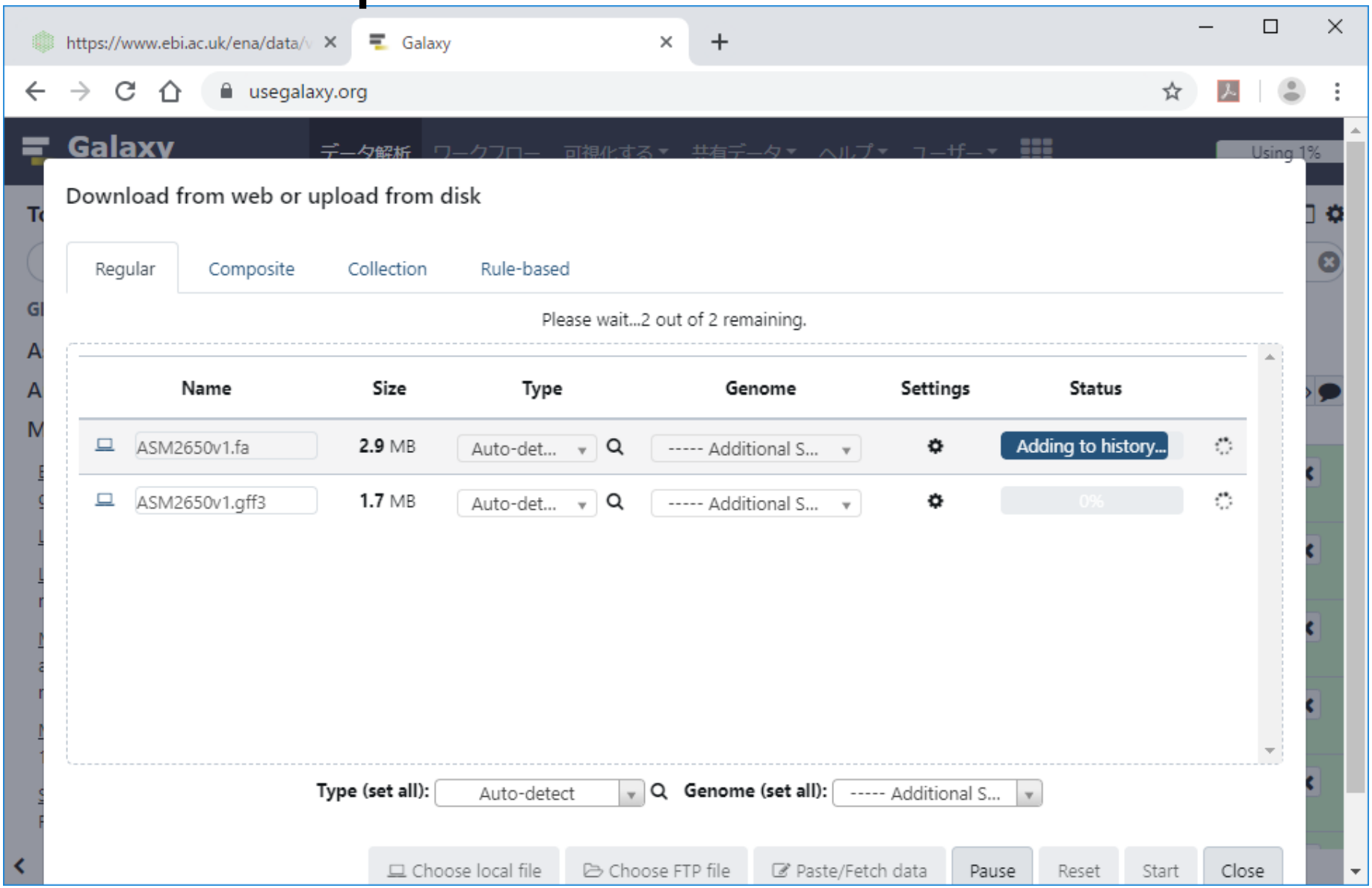
W7-7 : Upload

こんな感じになる。2つのファイルをドラッグ&ドロップで置いたところ。①Start。



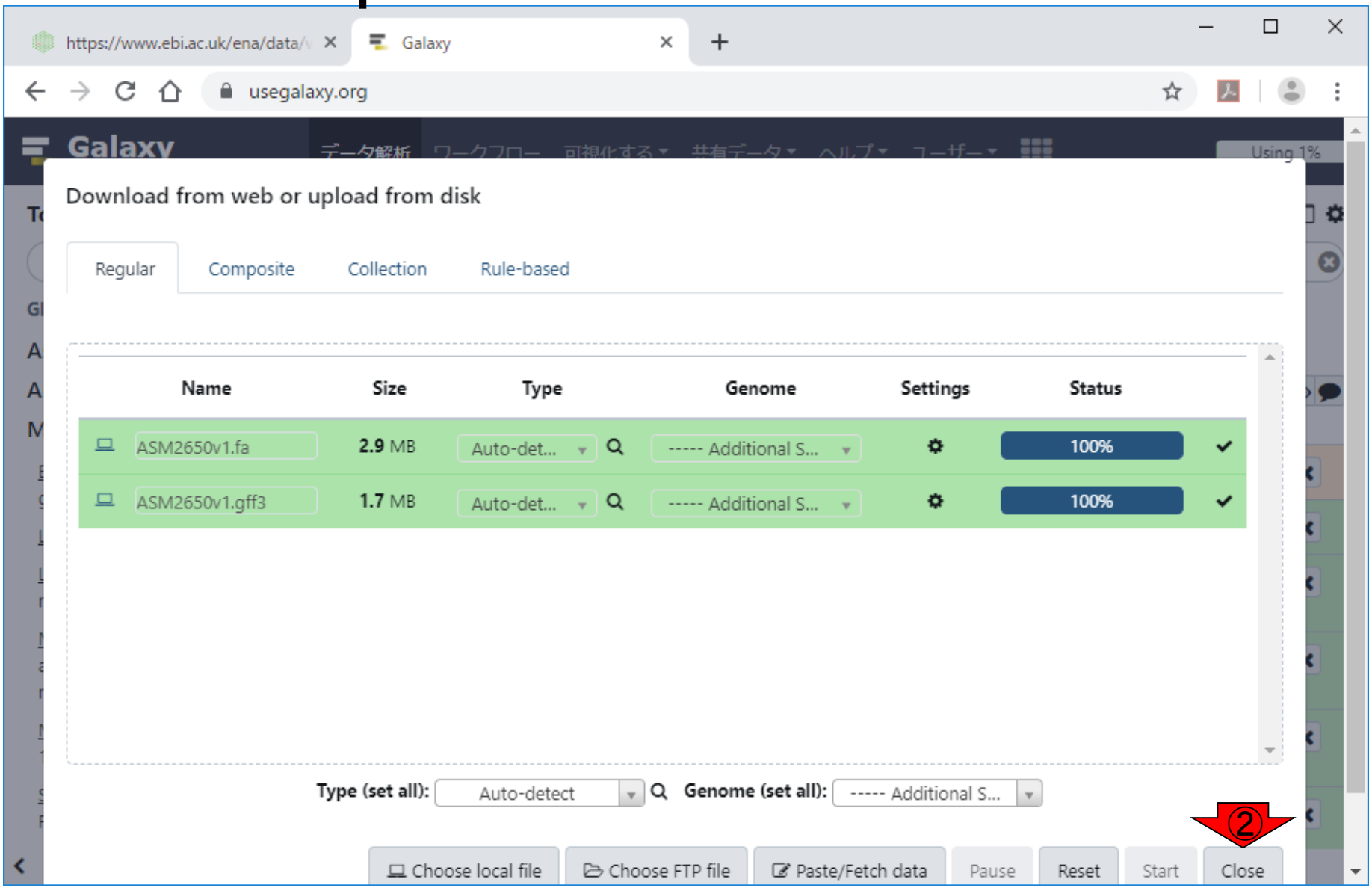
W7-7 : Upload

こんな感じになる。2つのファイルをドラッグ&ドロップで置いたところ。①Start。アップロード中…。



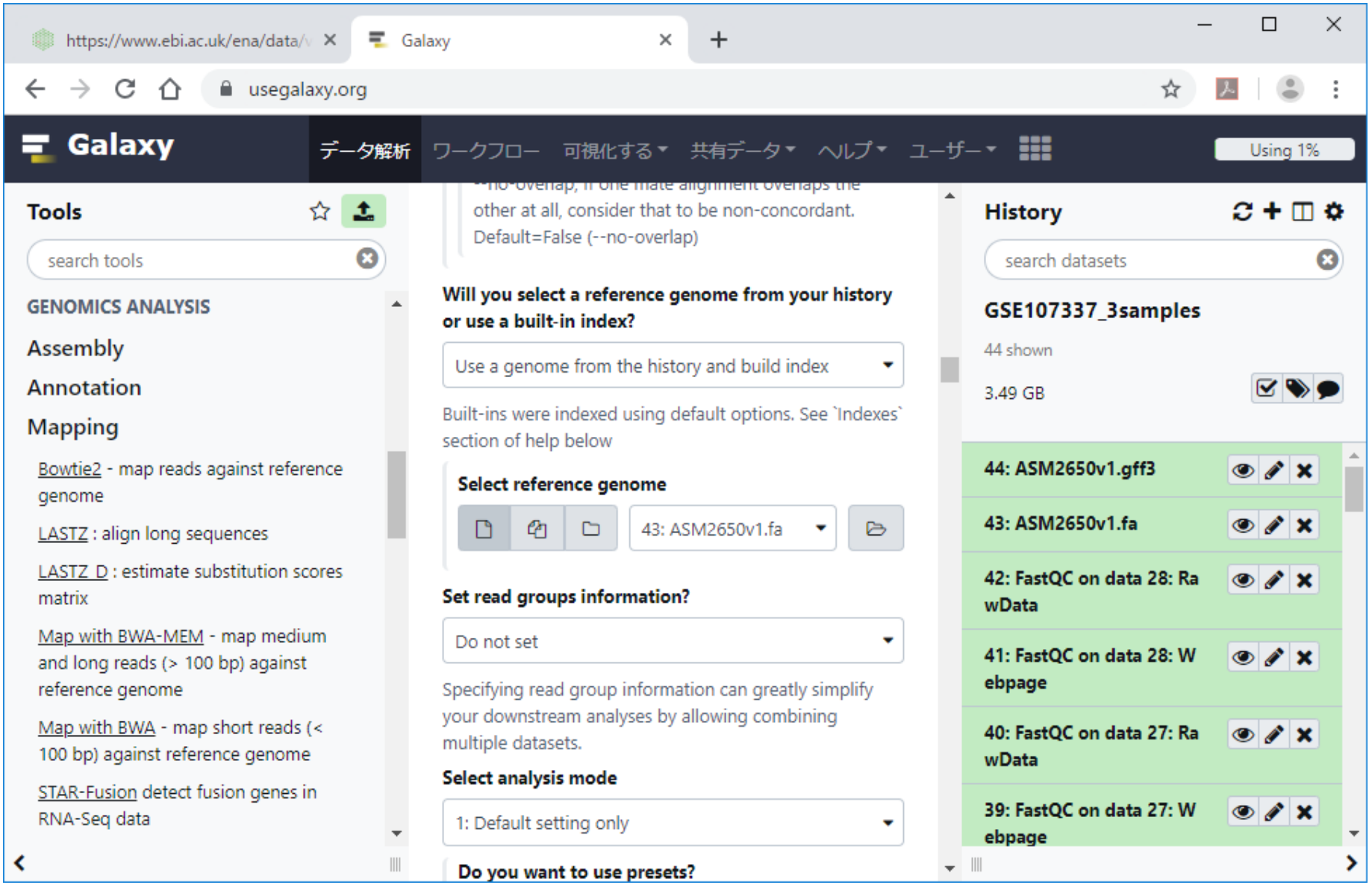
W7-7 : Upload

こんな感じになる。2つのファイルをドラッグ&ドロップで置いたところ。①Start。アップロード完了。②Close



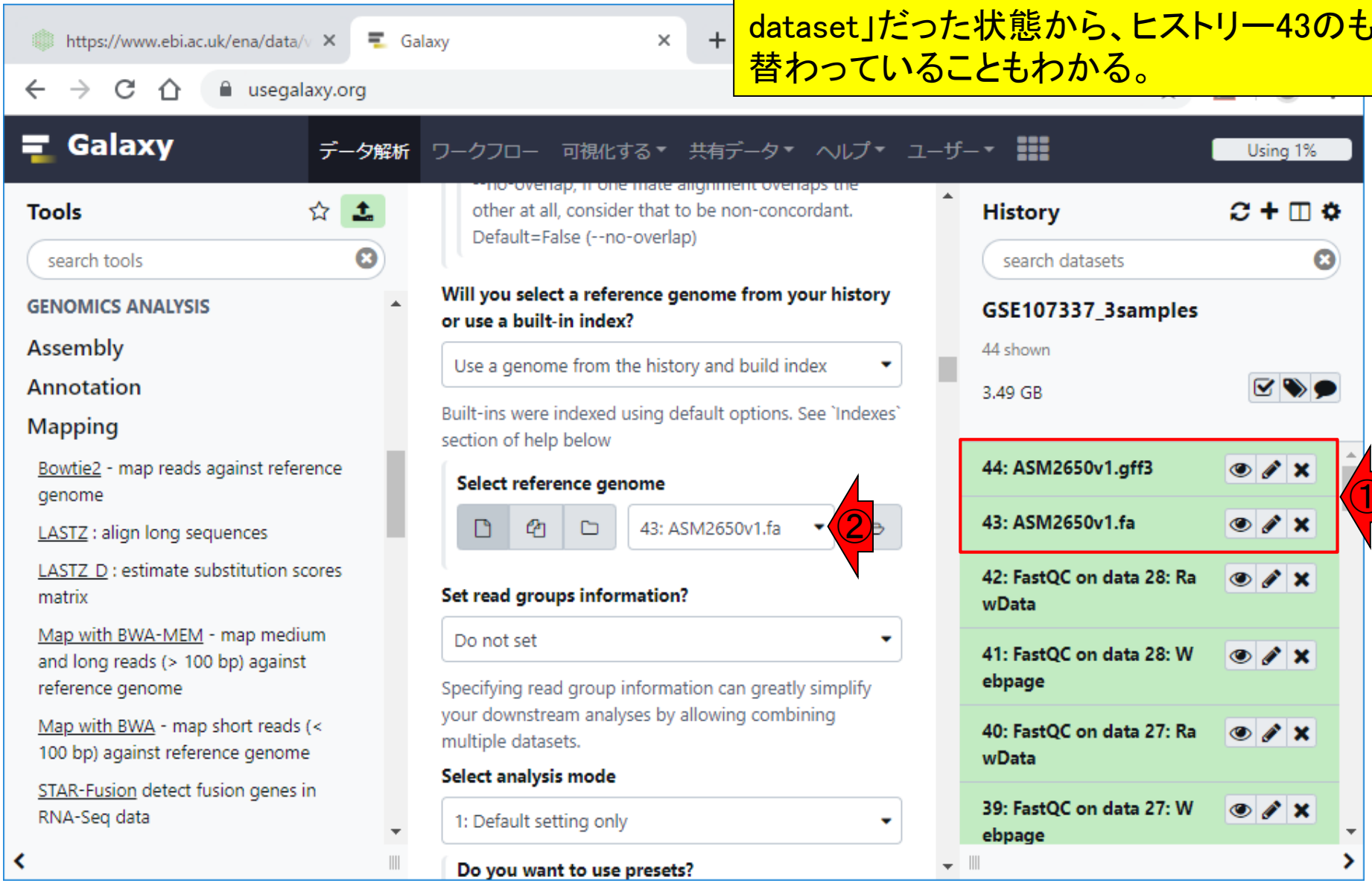
こんな感じになる。

W7-8：確認



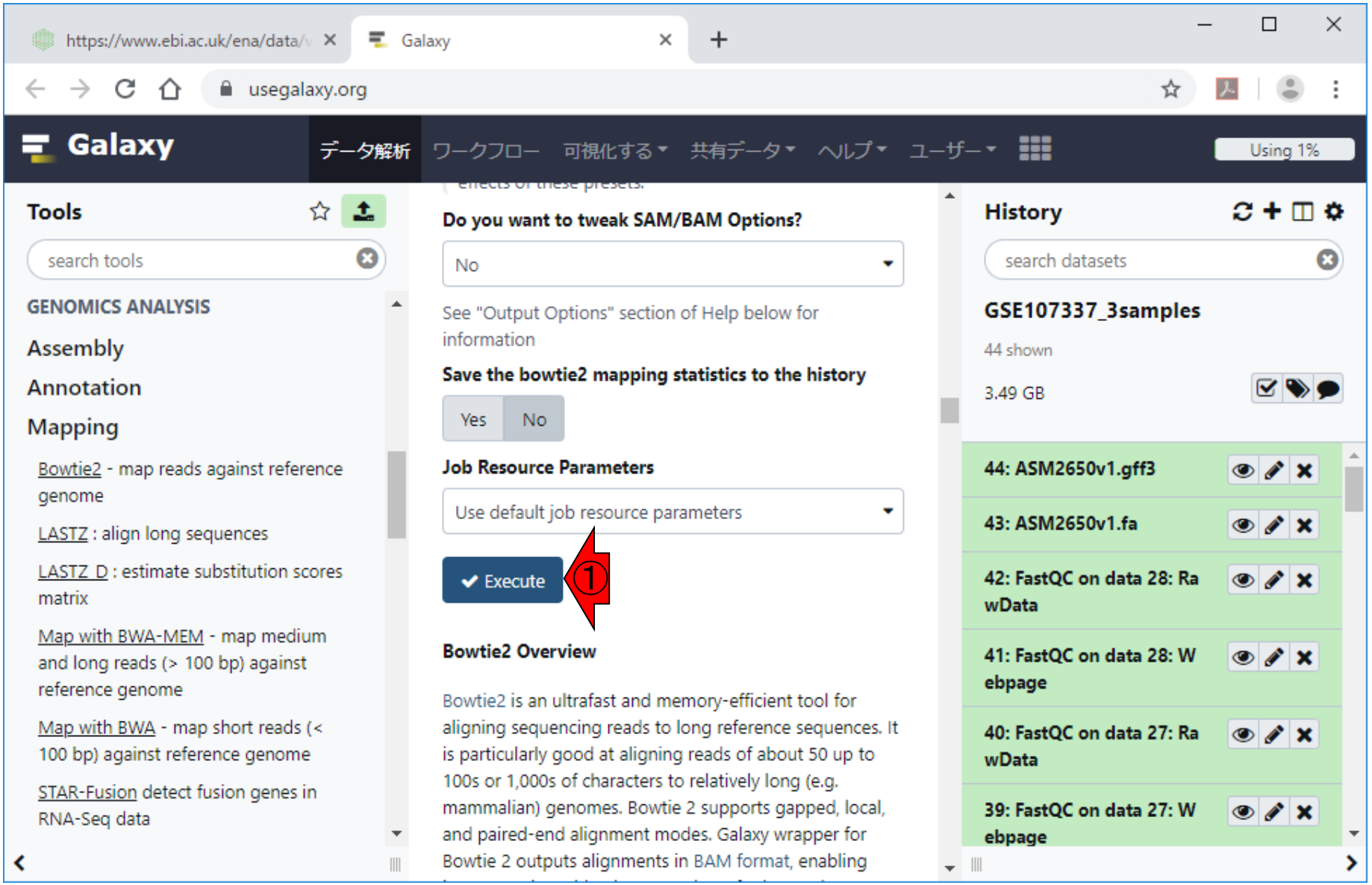
W7-8：確認

こんな感じになる。①アップロードした2つのファイルが、ヒストリーに追加されているのがわかる。②リファレンスゲノムの候補として、W7-6の「No fasta dataset」だった状態から、ヒストリー43のものに切り替わっていることもわかる。



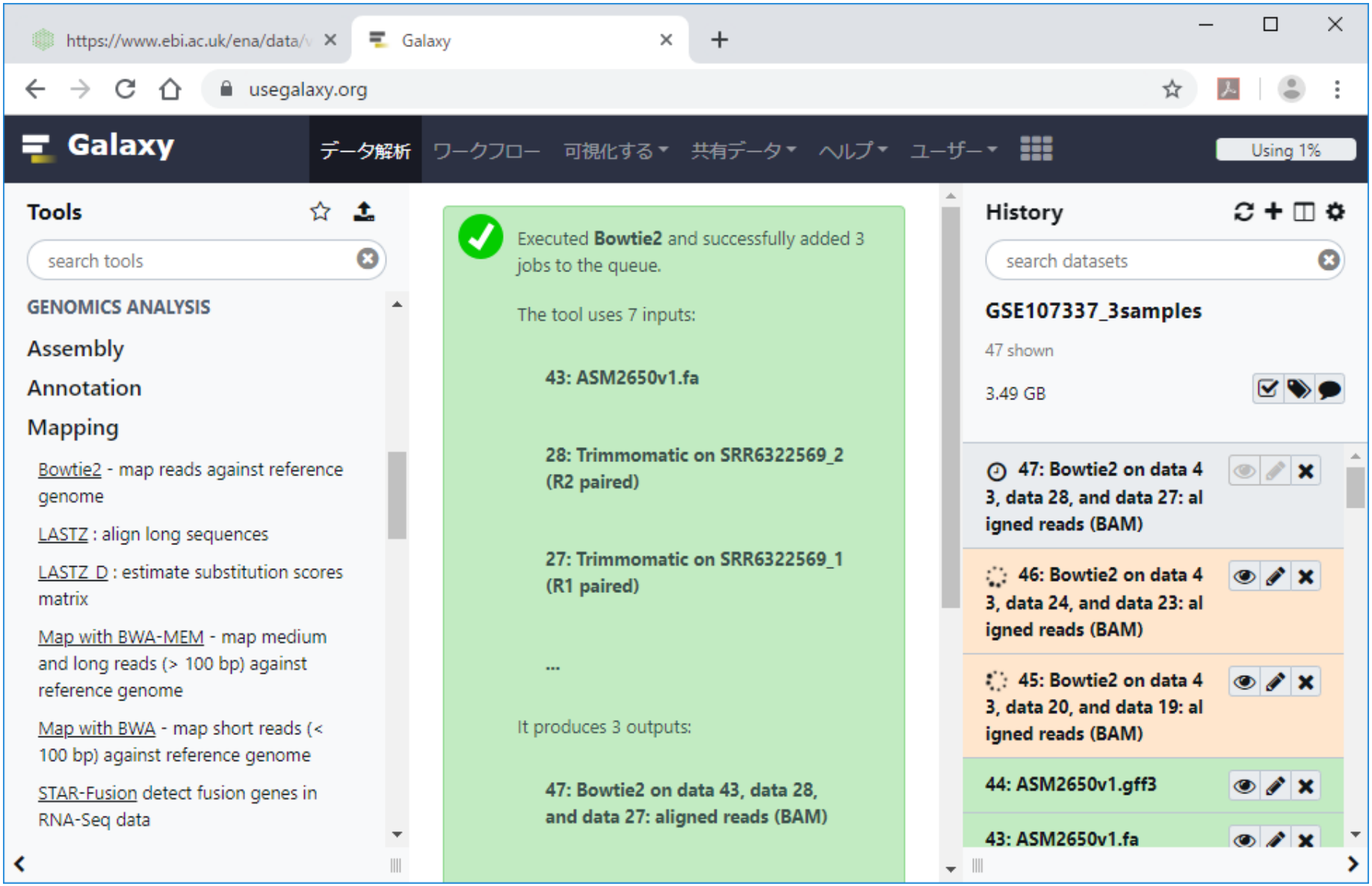
中央パネル下部に移動して、①実行。

W7-9：実行



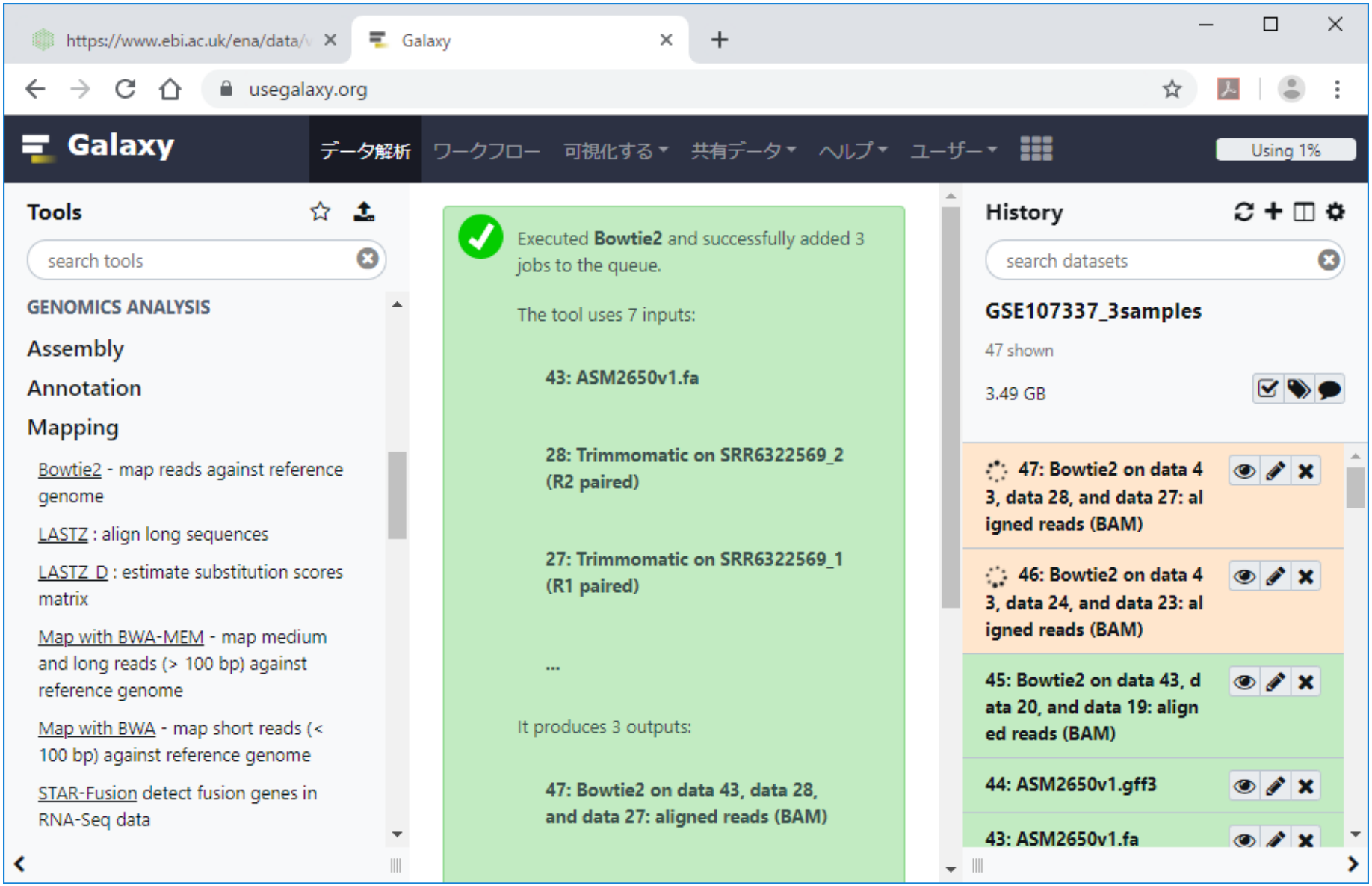
中央パネル下部に移動して、①実行。実行中…。

W7-10：実行中



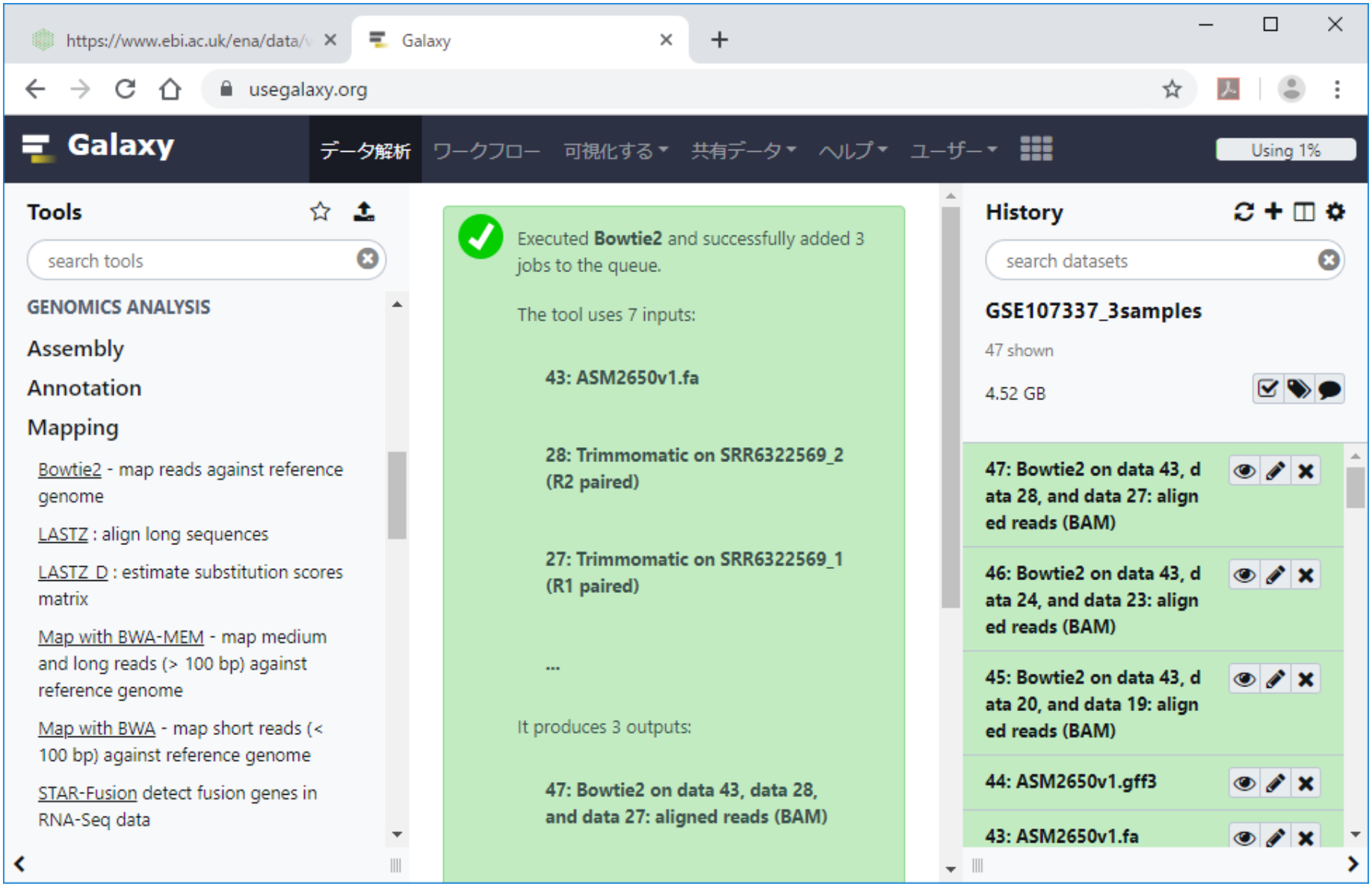
中央パネル下部に移動して、①実行。実行中…。

W7-10：実行中



実行完了。このときは、約10分かかりました。

W7-11：実行完了

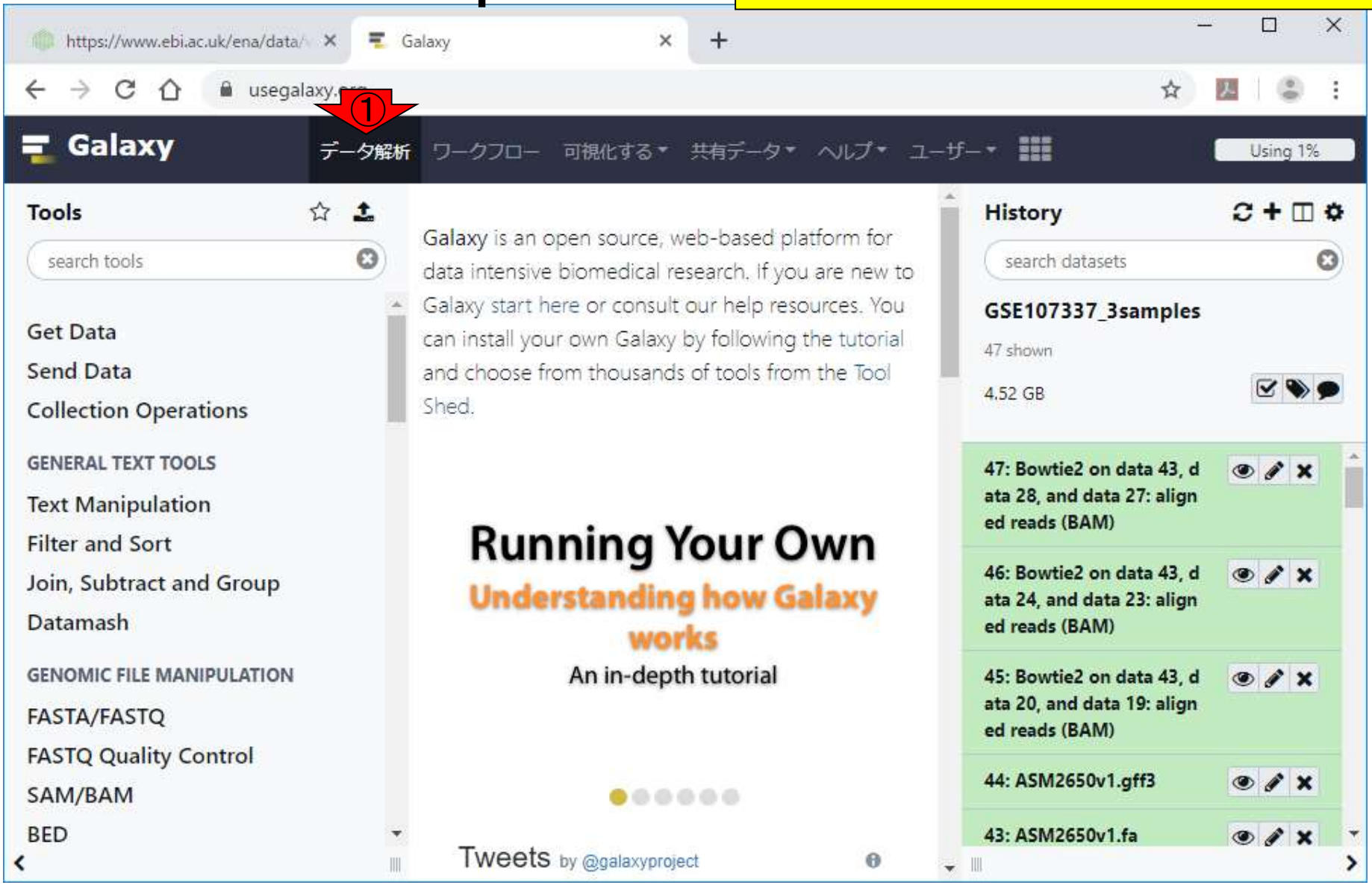


Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

W8-1：htseq-count

次はカウント情報取得。htseq-countというプログラムを用いる。①を押してヒストリーパネル以外をリフレッシュした後の状態。



W8-1：htseq-count

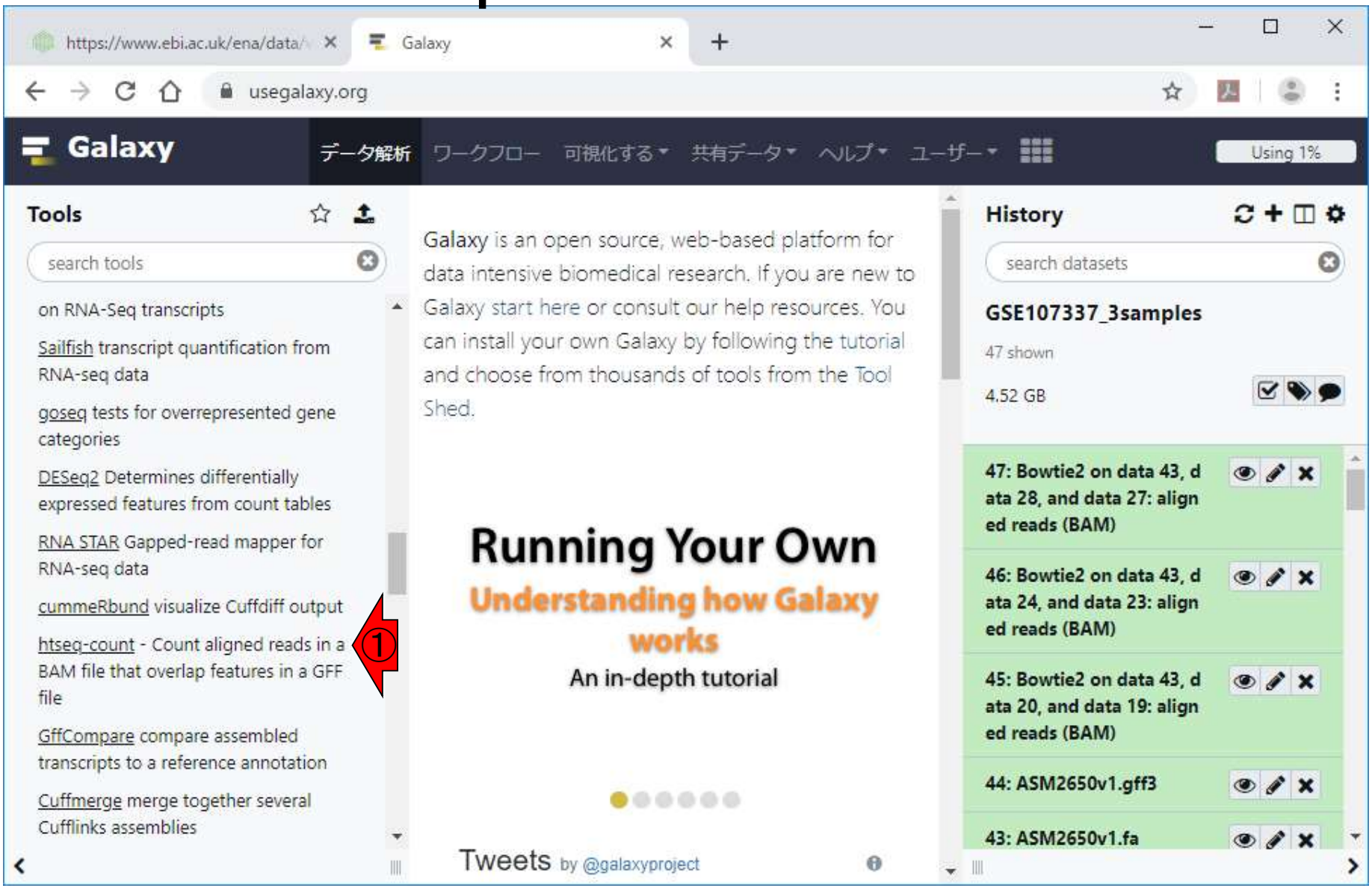
W7で行っていたのは、①Mapping。カウント情報取得を行うhtseq-countプログラムは、②RNA-seqのカテゴリー内に存在する。そのことを知っているからそうしています。もしプログラム名しかわからない場合は、③の検索窓でhtseq-countで探してもよい。

W8-1：htseq-count

②をクリックした結果。実は③featureCountsというプログラムでもカウント情報取得という目的を達成することはできる。がここではhtseq-countを利用すべく、④下部を少しずつ眺めながらhtseq-countを探す。

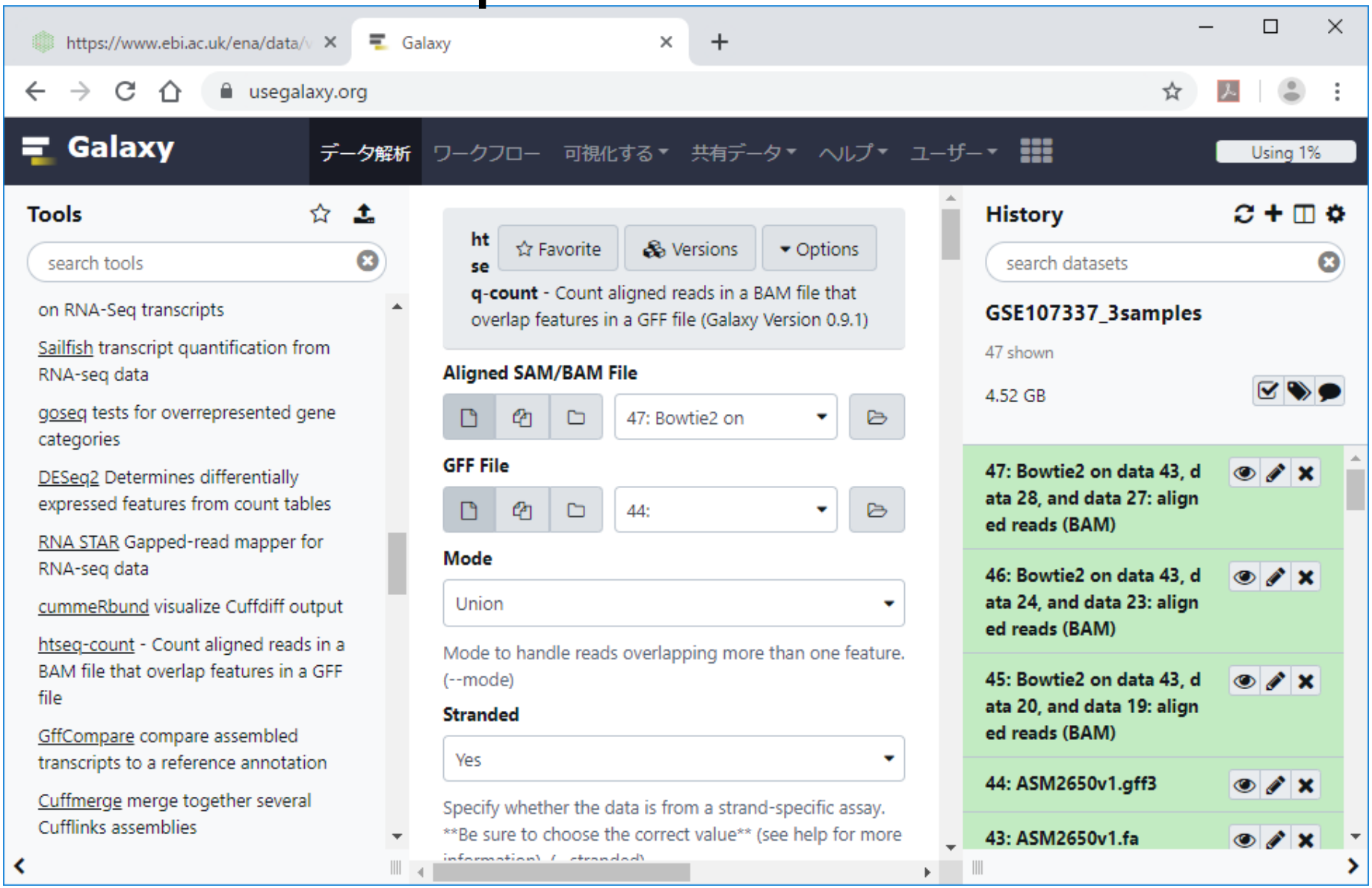
①htseq-countを発見したのでクリック。

W8-1：htseq-count



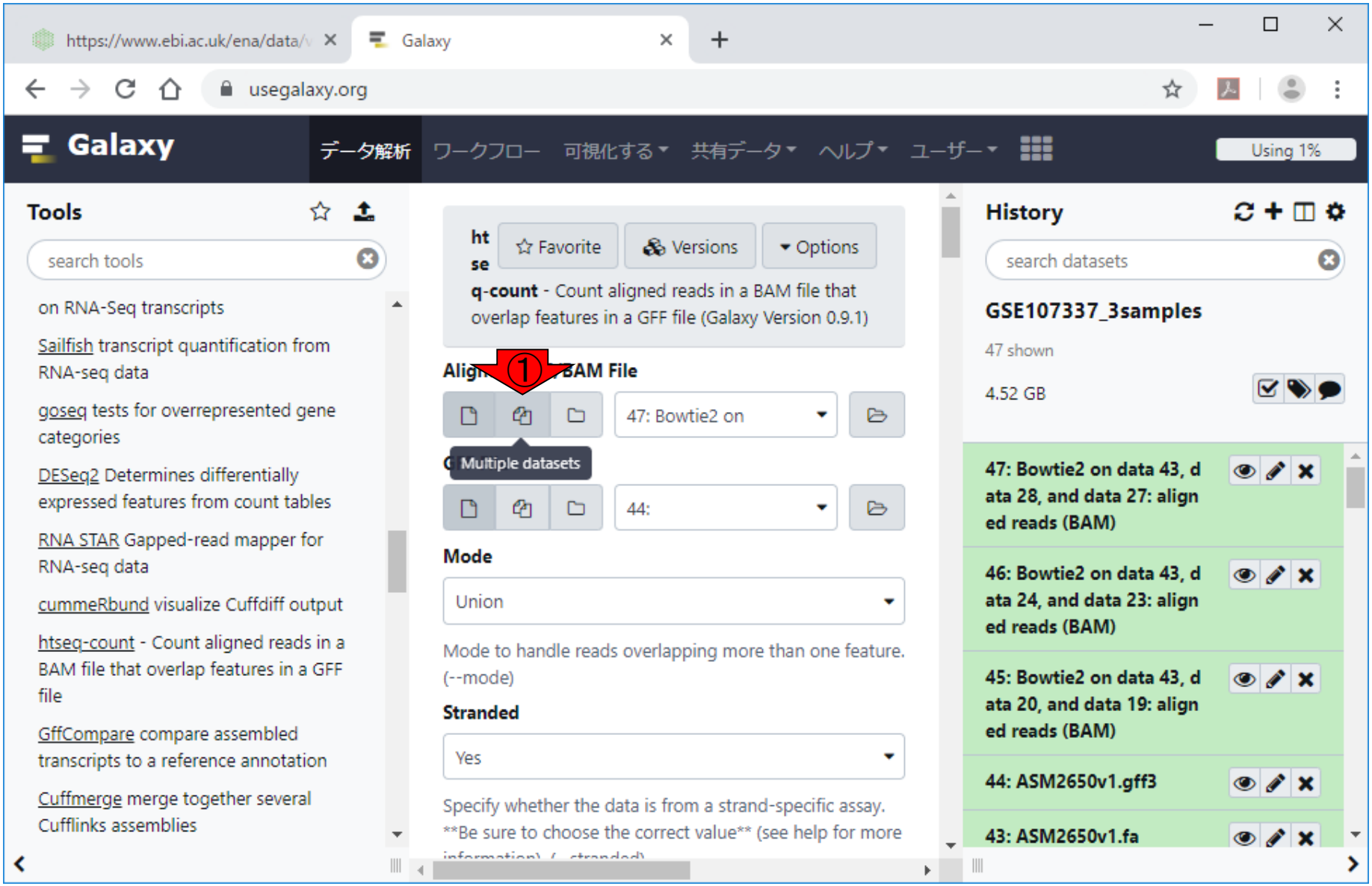
こんな感じになる。

W8-1：htseq-count



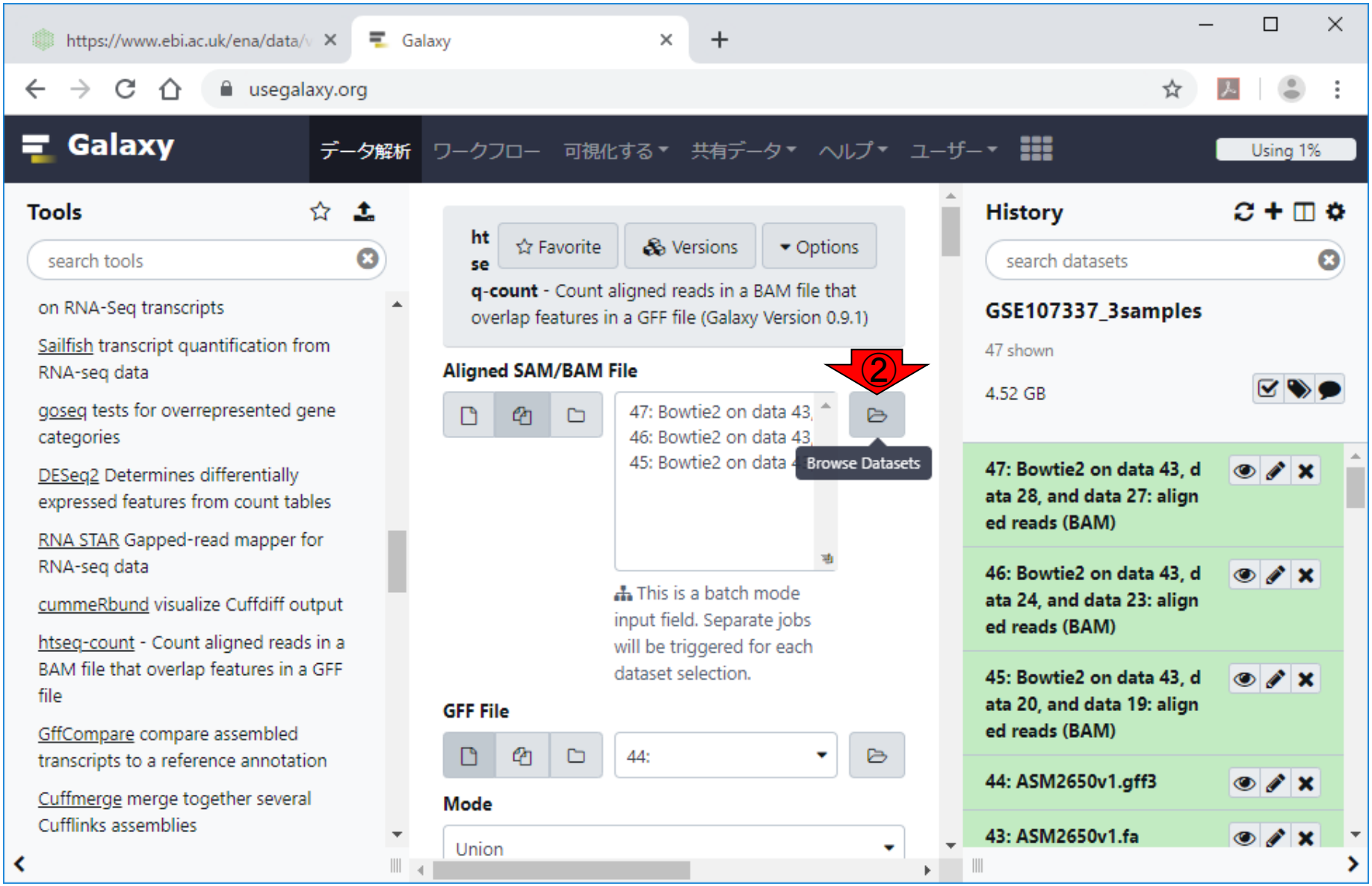
W8-2：複数個指定

計3つのBAMファイルを一度に取り扱いたいので、
①Multiple datasets。



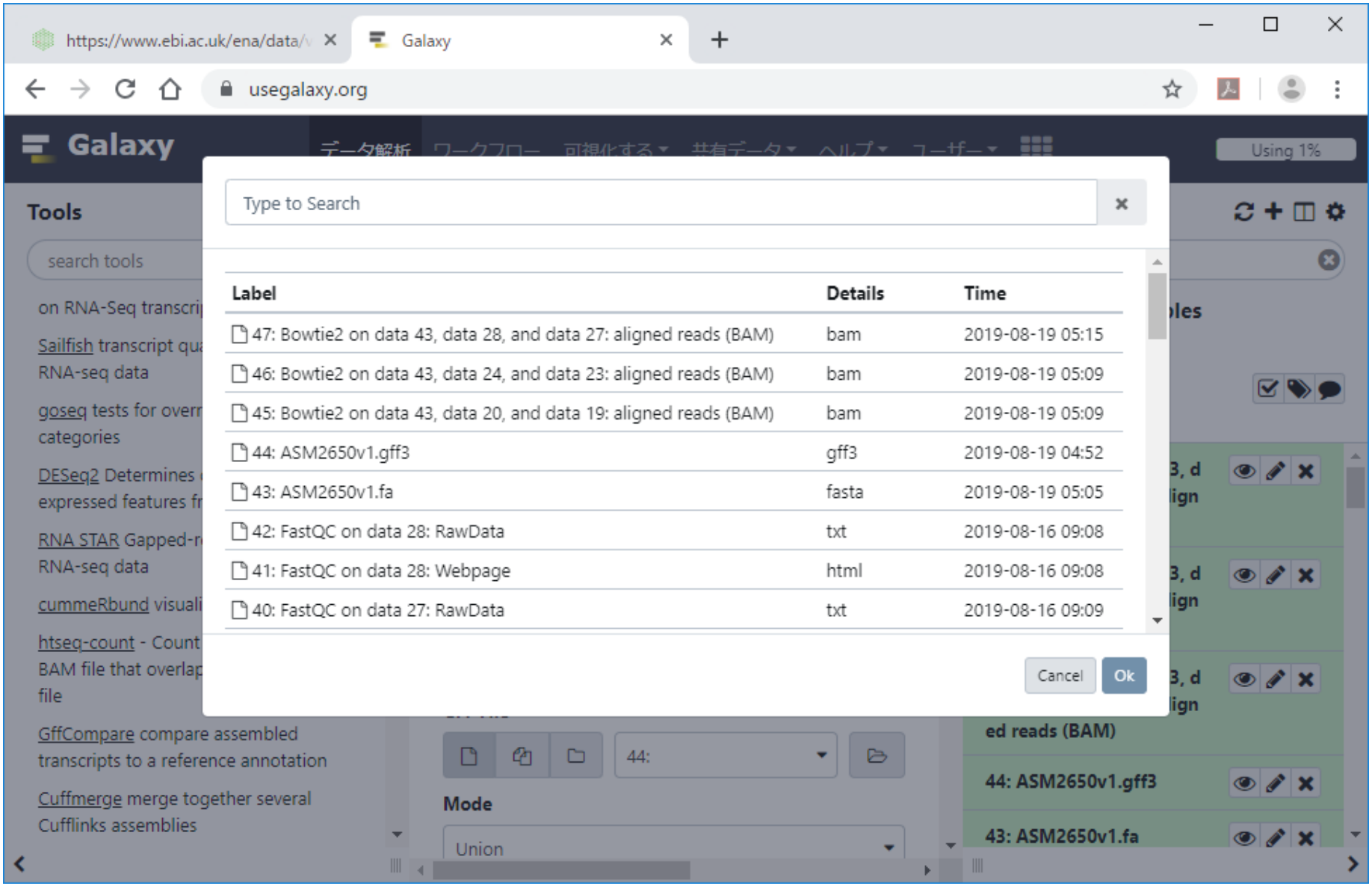
W8-2：複数個指定

計3つのBAMファイルを一度に取り扱いたいので、
①Multiple datasets。②Browse Datasets。



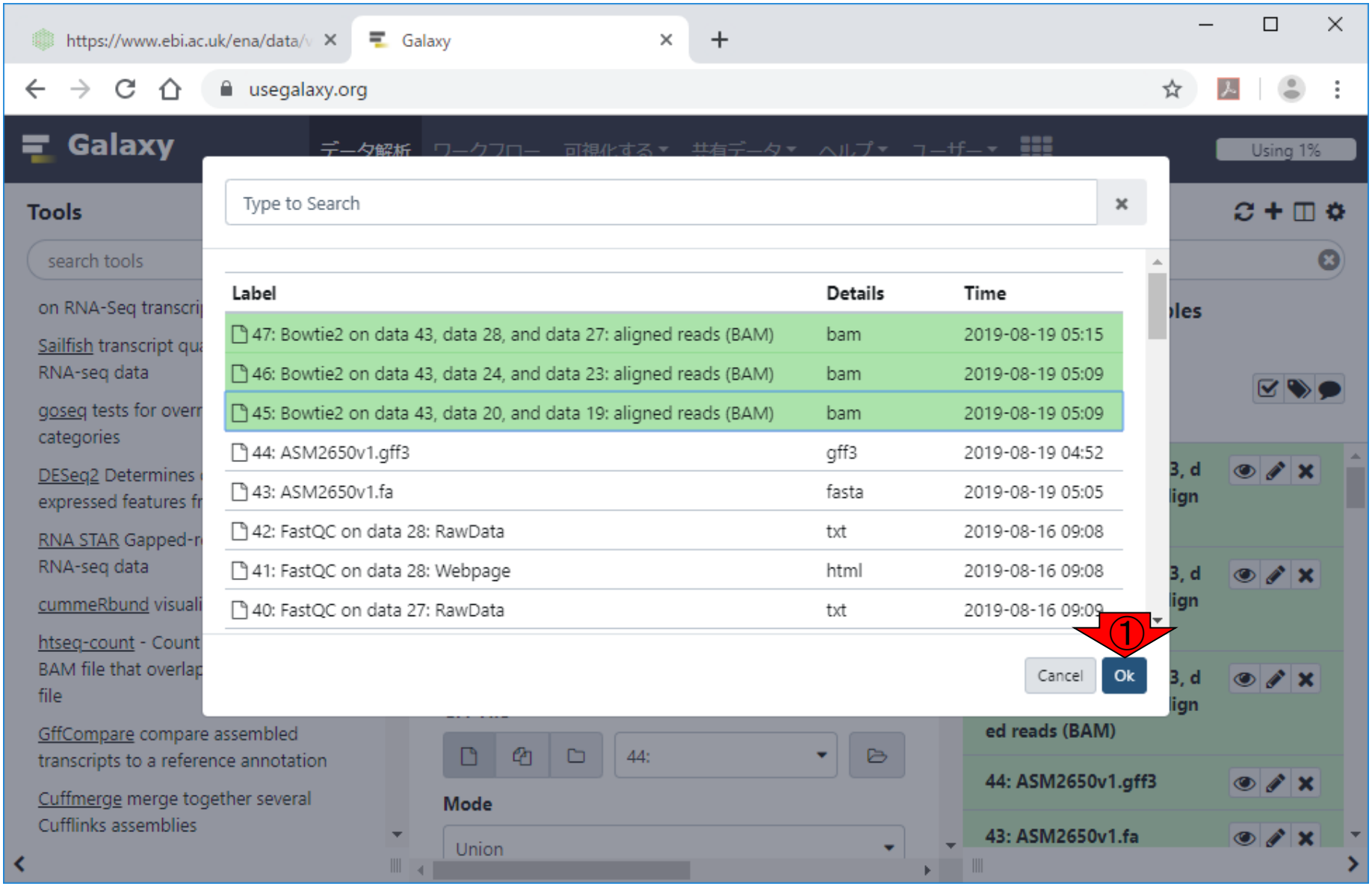
こんな感じになる。入力はBAMファイルなので…

W8-2：複数個指定



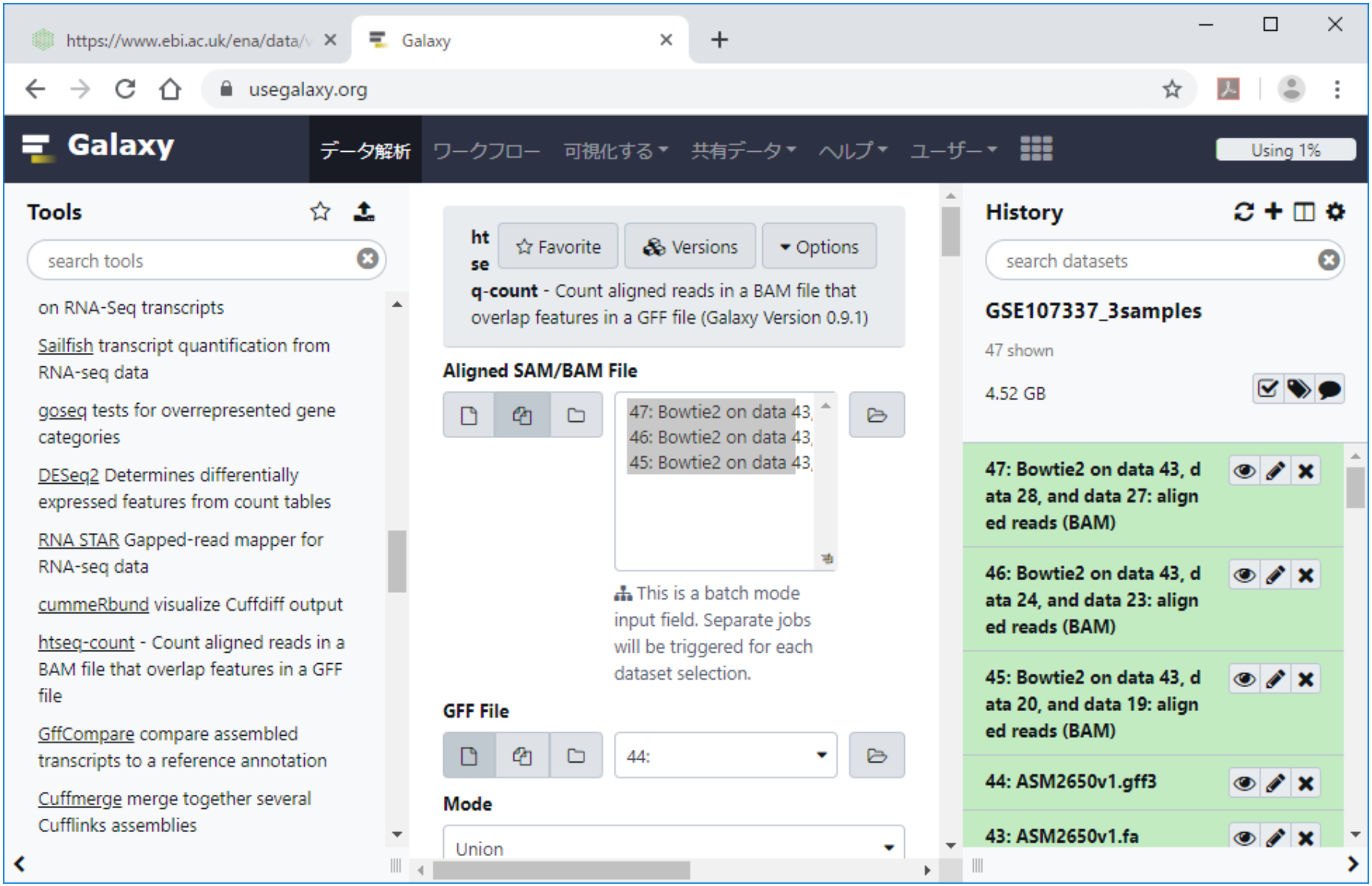
W8-2：複数個指定

こんな感じになる。入力はBAMファイルなので…それらを選択して、①Ok。



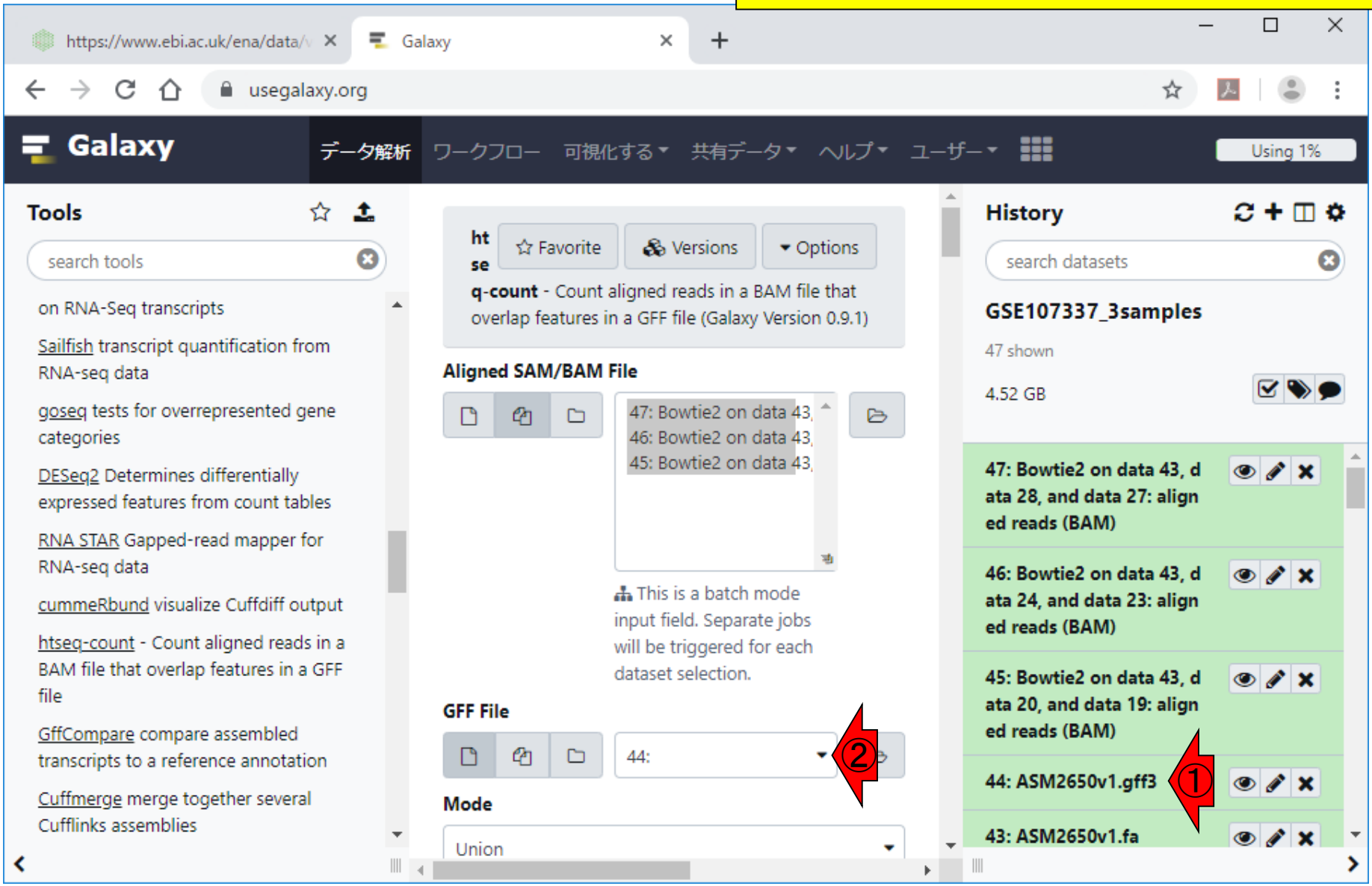
W8-2：複数個指定

こんな感じになる。ここまででマッピング結果ファイルの指定は完了。



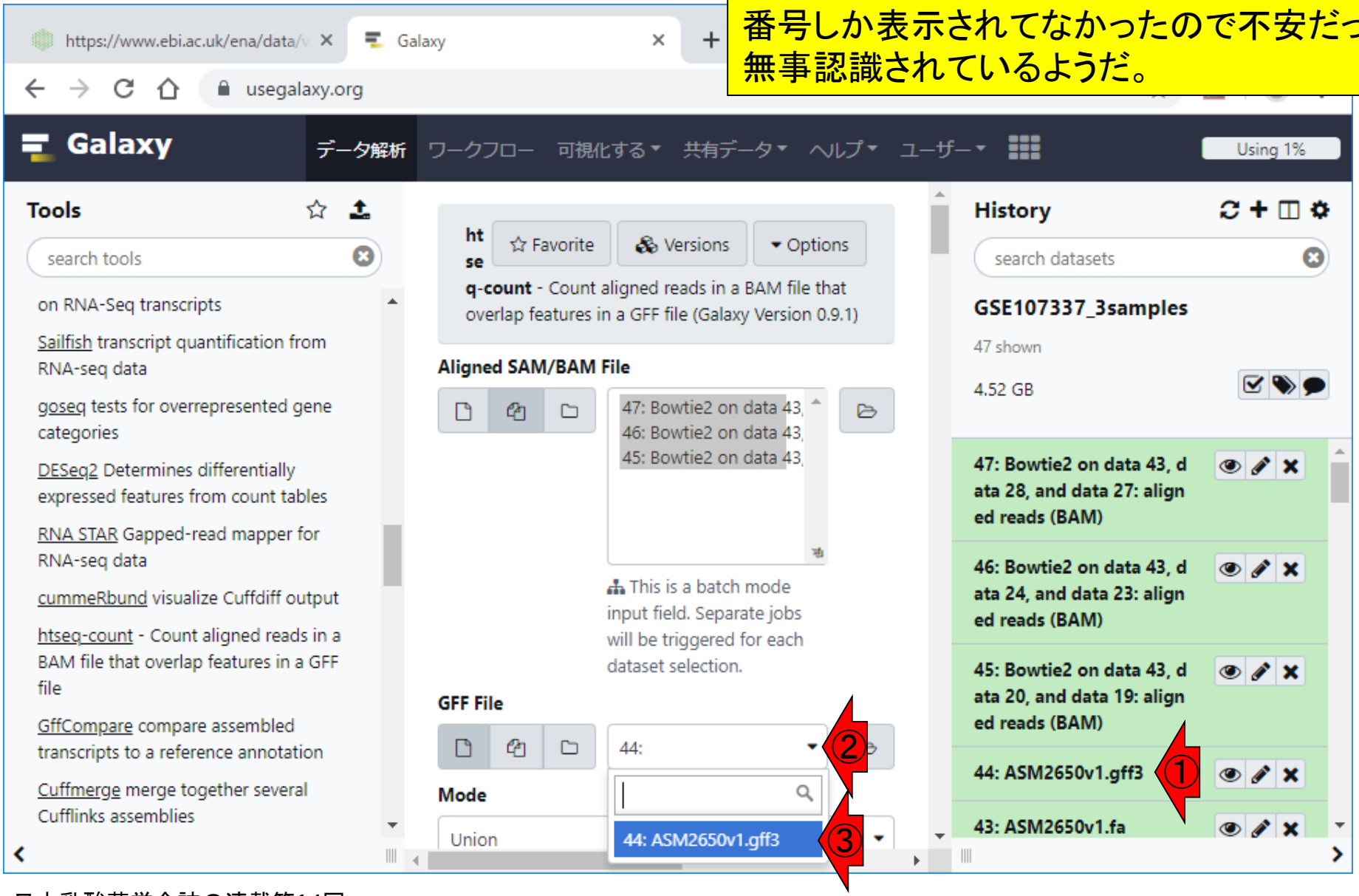
W8-3：GFFファイル

次は、どの領域のカウント情報を取得するかを指定する。①で見えているgff3ファイルを、②の部分で認識させるべく、②をクリック。



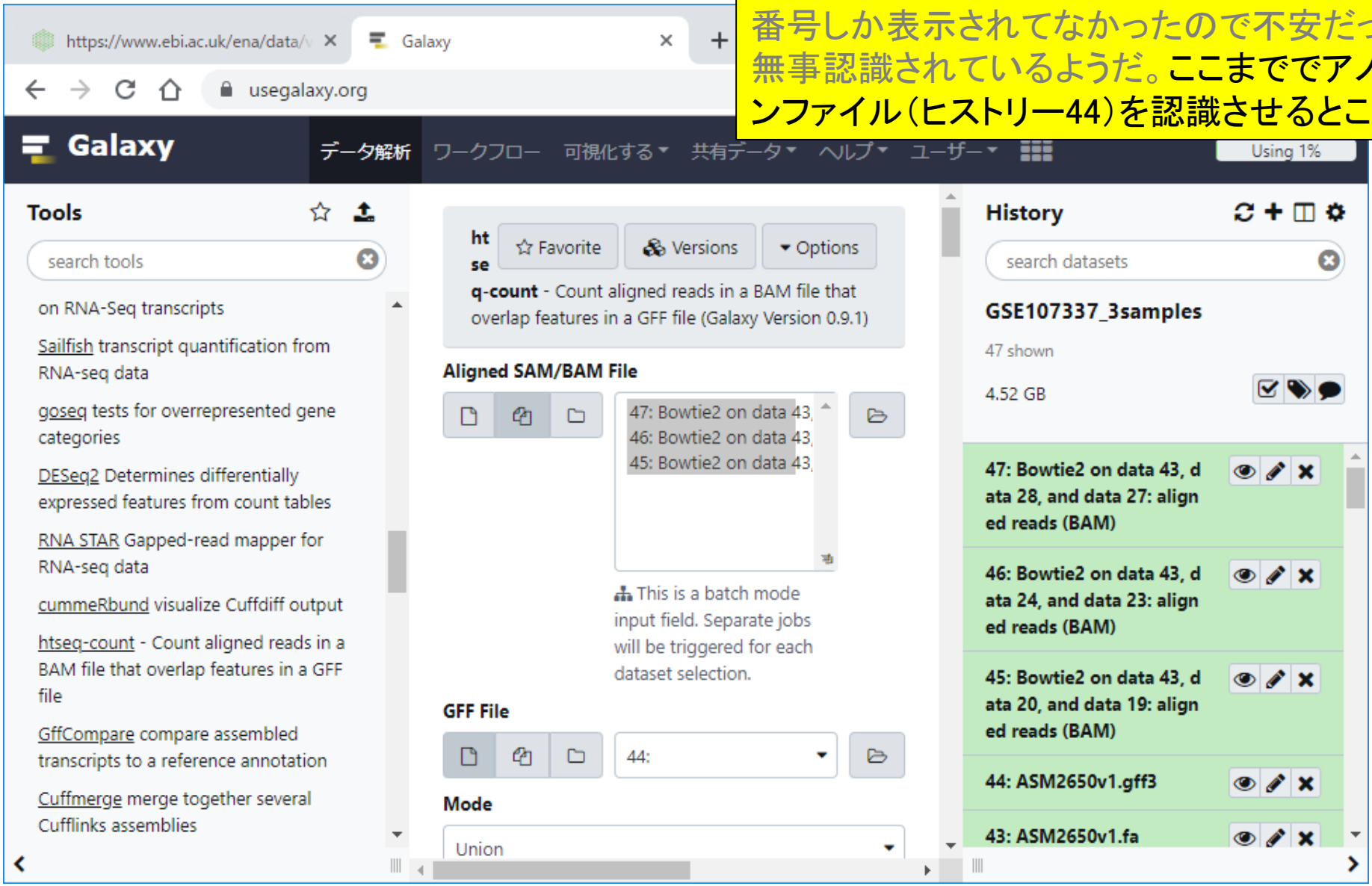
W8-3：GFFファイル

次は、どの領域のカウント情報を取得するかを指定する。①で見えているgff3ファイルを、②の部分で認識させるべく、②をクリック。②の部分ではヒストリー番号しか表示されてなかったので不安だったが、③無事認識されているようだ。



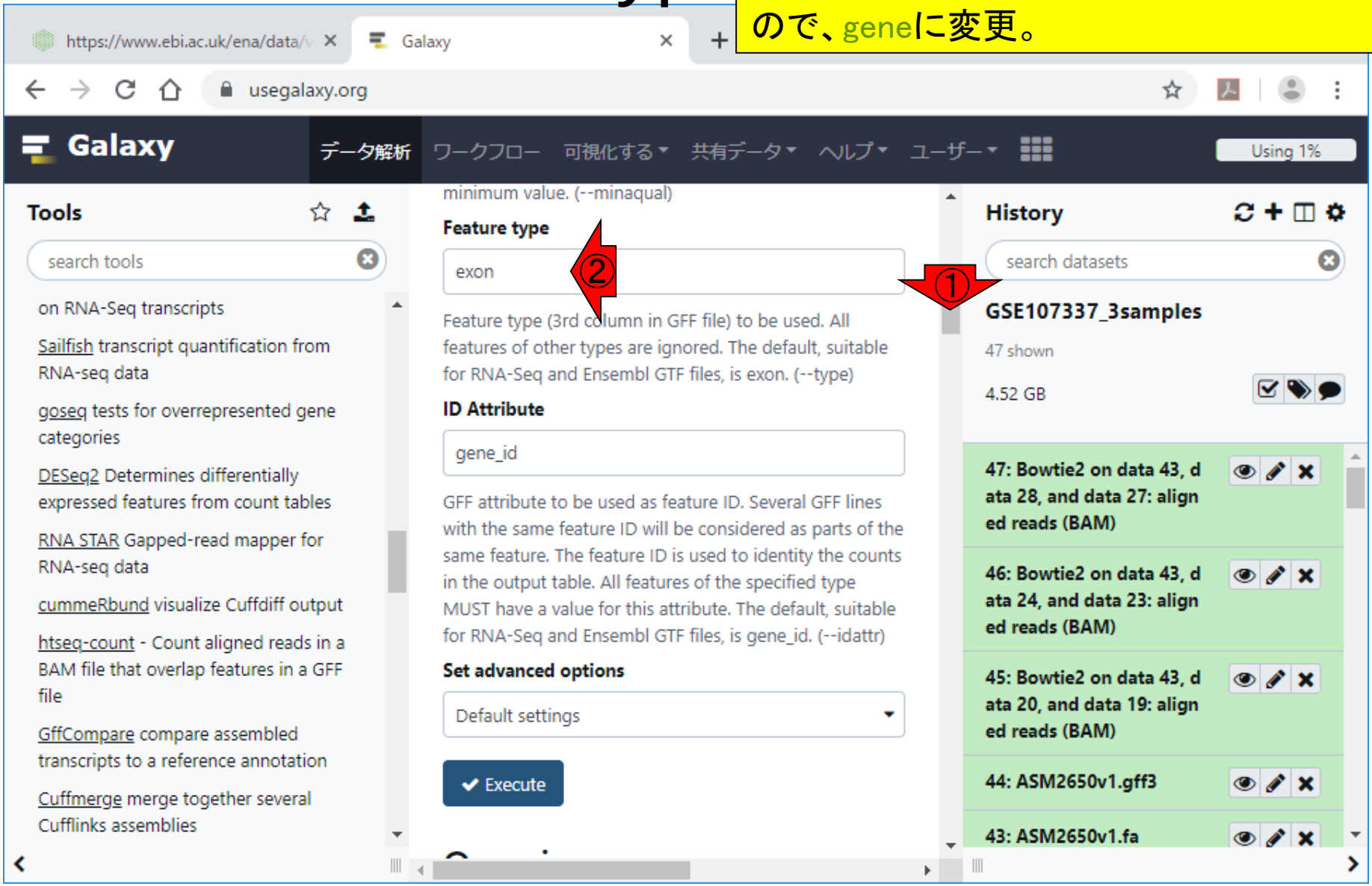
W8-3：GFFファイル

次は、どの領域のカウント情報を取得するかを指定する。①で見えているgff3ファイルを、②の部分で認識させるべく、②をクリック。②の部分ではヒストリー番号しか表示されてなかったので不安だったが、③無事認識されているようだ。ここまででアノテーションファイル（ヒストリー44）を認識させるところが完了



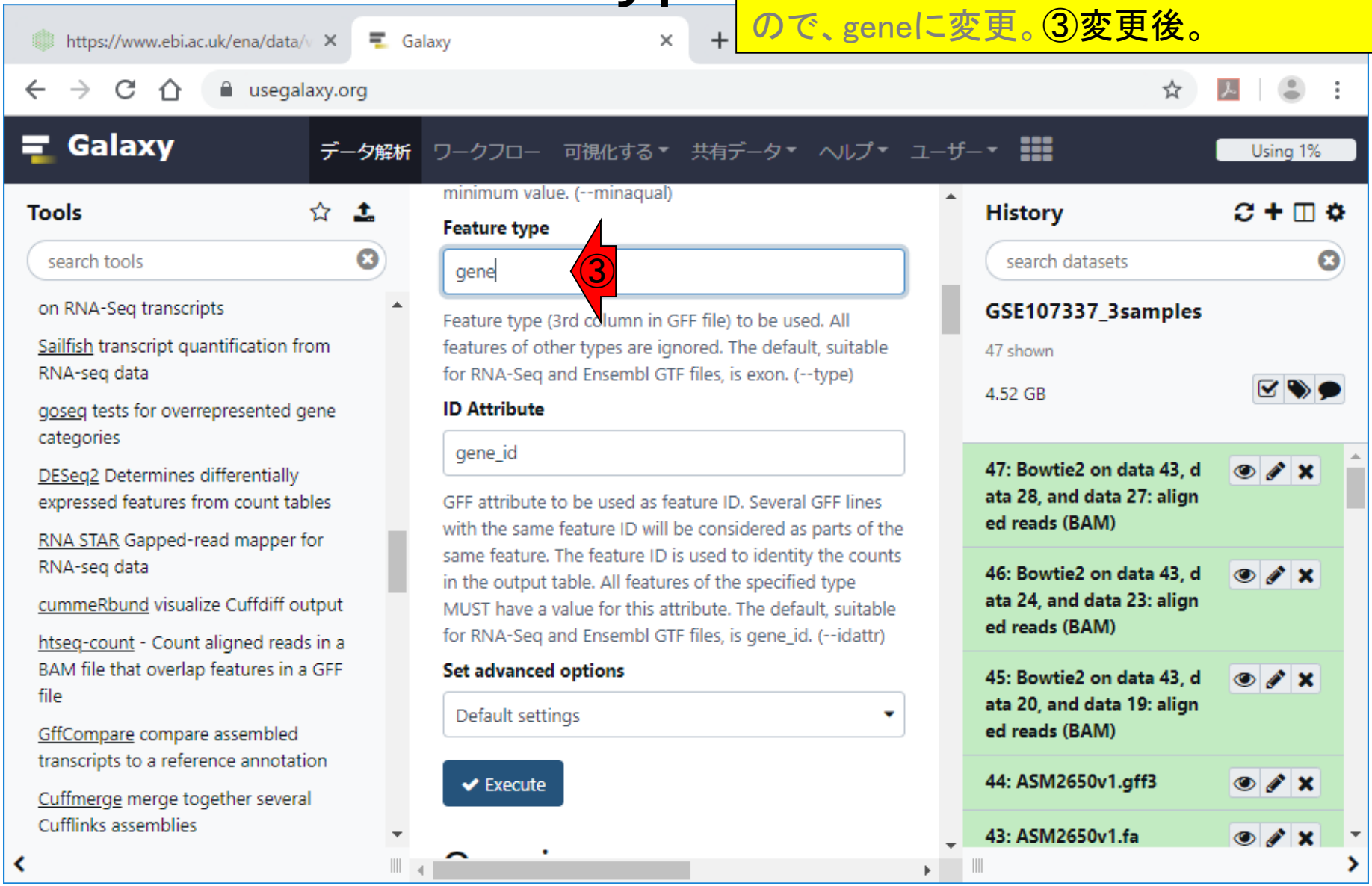
W8-4：Feature type

①少し下部に移動。次は何の情報をカウントしたいのかを指定する。②デフォルトはexonだが、ここでは原著論文と同じくgeneのカウント情報を取得したいので、geneに変更。



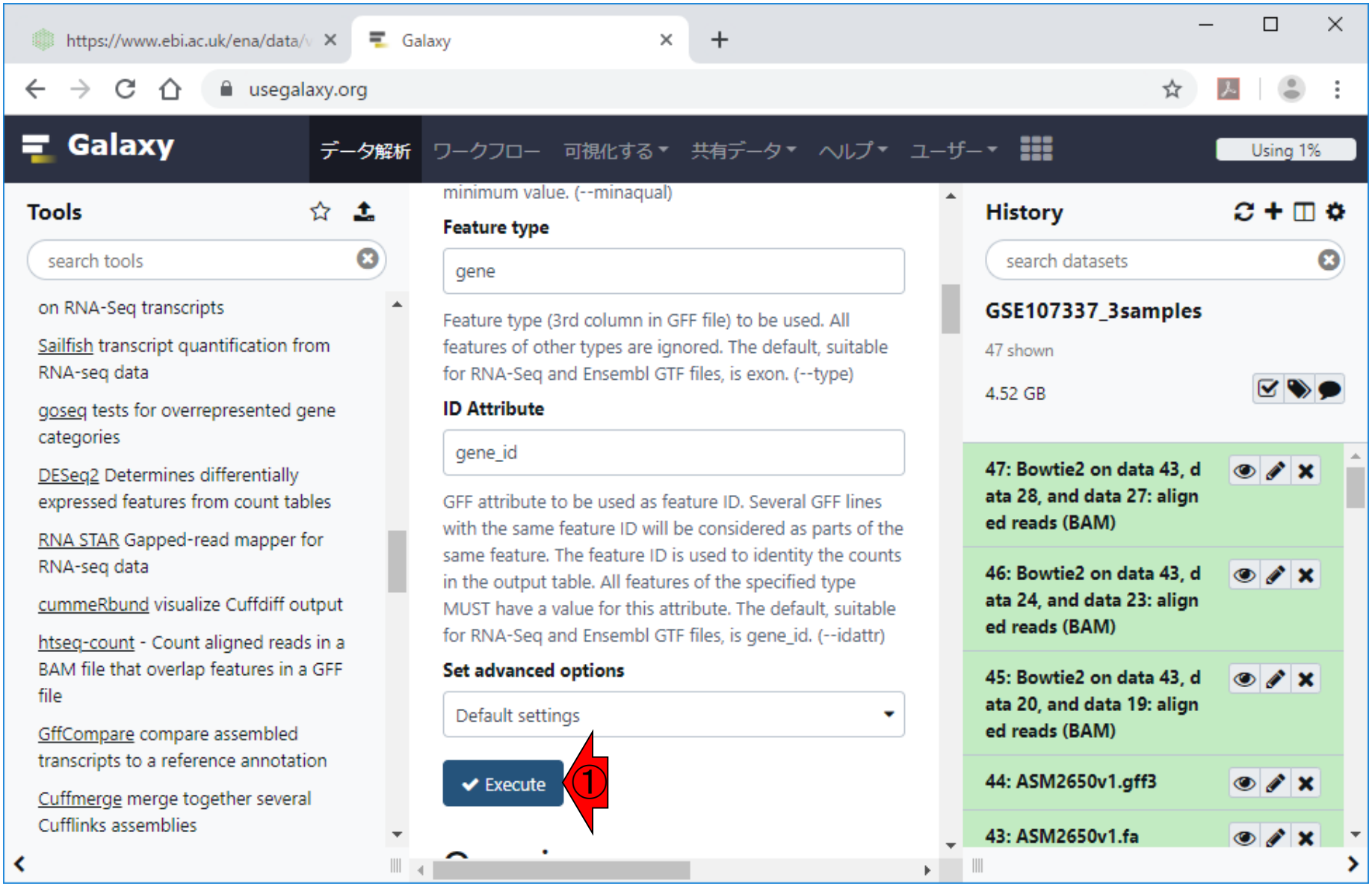
W8-4：Feature type

①少し下部に移動。次は何の情報をカウントしたいのかを指定する。②デフォルトはexonだが、ここでは原著論文と同じくgeneのカウント情報を取得したいので、geneに変更。③変更後。



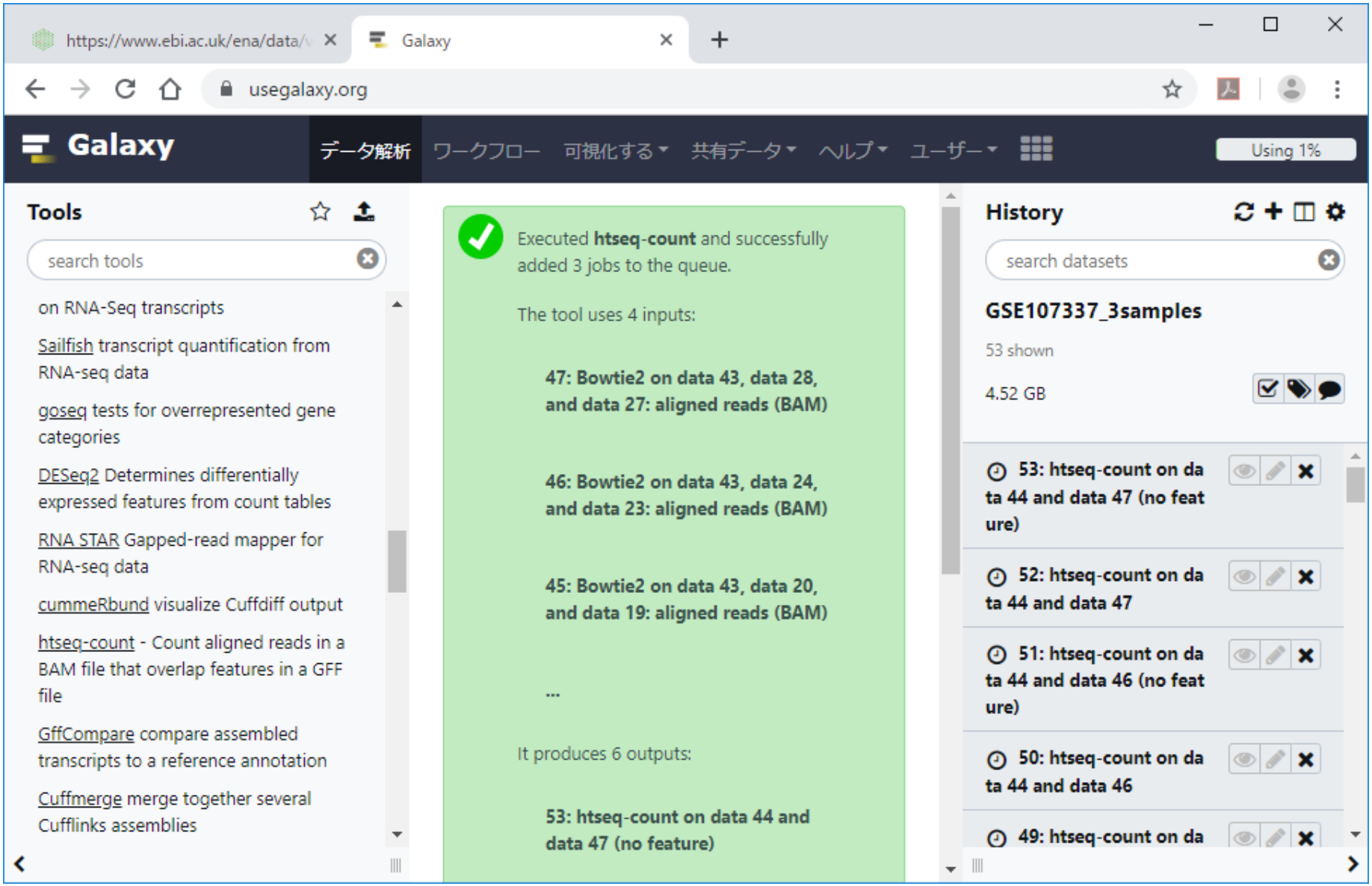
それ以外は特に変更せずに、①Execute。

W8-5：実行



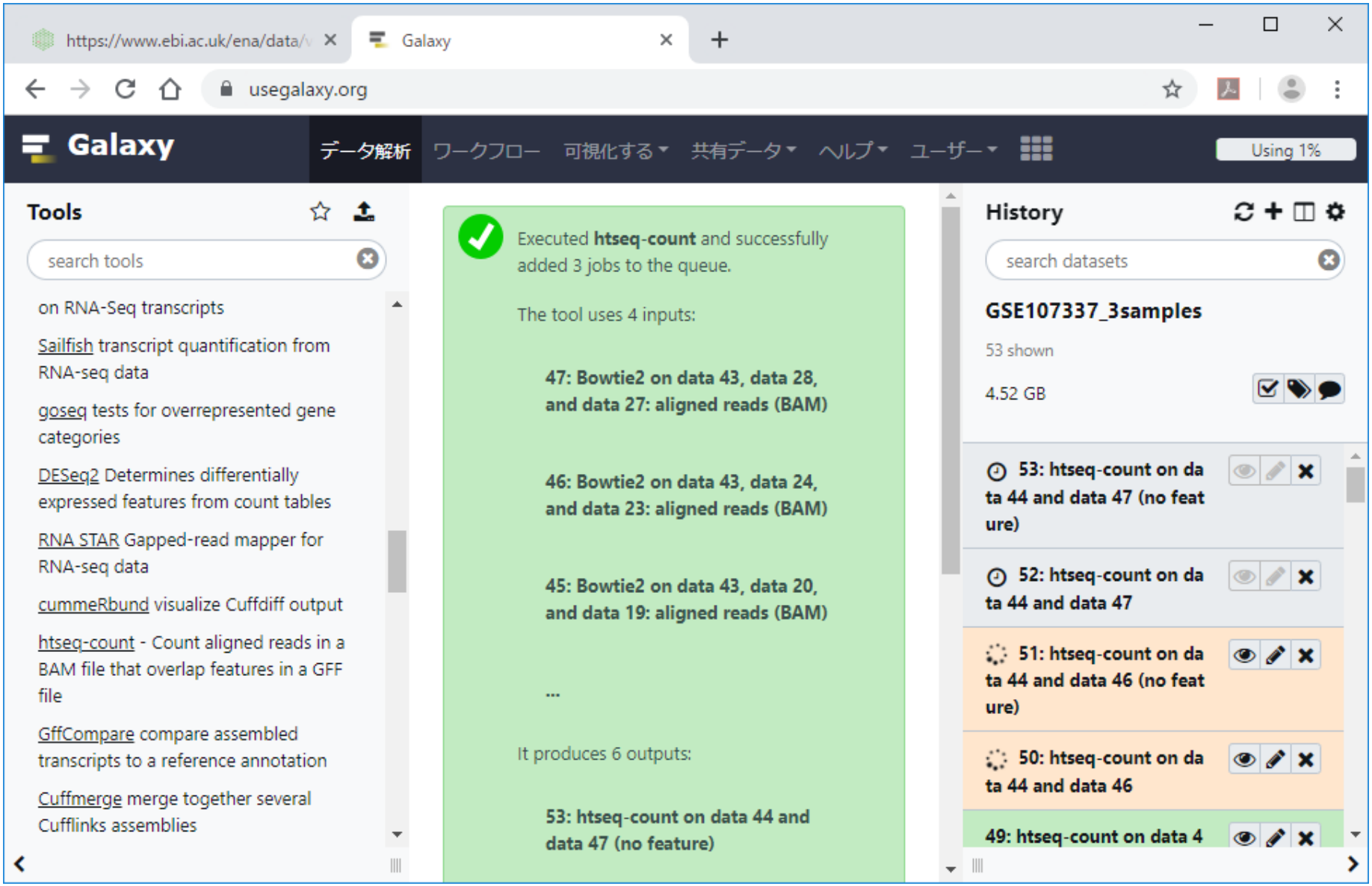
実行中…。

W8-6：実行中


https://www.ebi.ac.uk/ena/data/
Galaxy
usegalaxy.org
Galaxy
データ解析
ワークフロー
可視化する
共有データ
ヘルプ
ユーザー
Using 1%
Tools
search tools
on RNA-Seq transcripts
Sailfish transcript quantification from RNA-seq data
goseq tests for overrepresented gene categories
DESeq2 Determines differentially expressed features from count tables
RNA STAR Gapped-read mapper for RNA-seq data
cummeRbund visualize Cuffdiff output
htseq-count - Count aligned reads in a BAM file that overlap features in a GFF file
GffCompare compare assembled transcripts to a reference annotation
Cuffmerge merge together several Cufflinks assemblies
Executed htseq-count and successfully added 3 jobs to the queue.
The tool uses 4 inputs:
47: Bowtie2 on data 43, data 28, and data 27: aligned reads (BAM)
46: Bowtie2 on data 43, data 24, and data 23: aligned reads (BAM)
45: Bowtie2 on data 43, data 20, and data 19: aligned reads (BAM)
...
It produces 6 outputs:
53: htseq-count on data 44 and data 47 (no feature)
History
search datasets
GSE107337_3samples
53 shown
4.52 GB
53: htseq-count on data 44 and data 47 (no feature)
52: htseq-count on data 44 and data 47
51: htseq-count on data 44 and data 46 (no feature)
50: htseq-count on data 44 and data 46
49: htseq-count on da

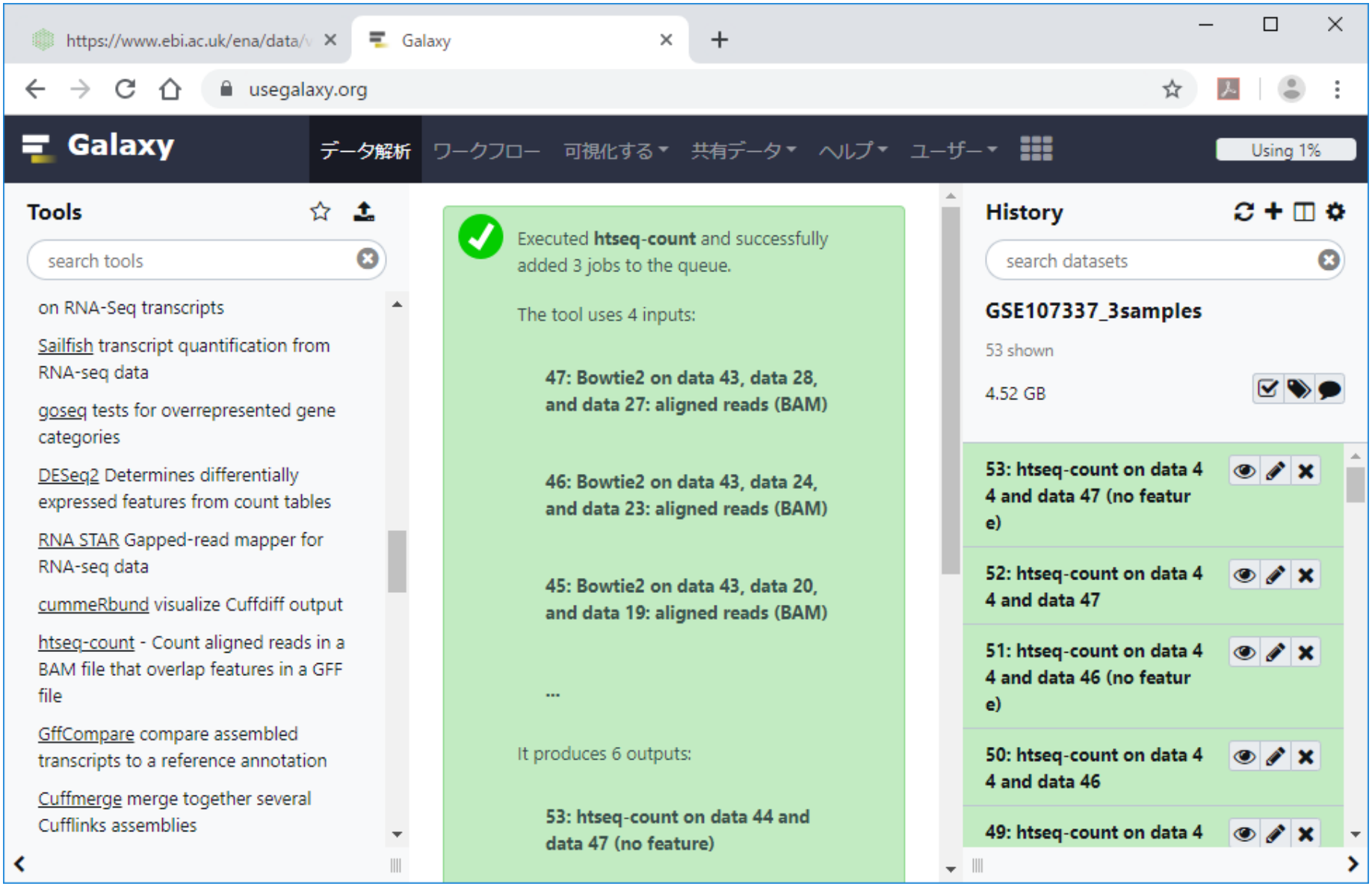
実行中…。

W8-6：実行中


https://www.ebi.ac.uk/ena/data/
Galaxy
usegalaxy.org
Galaxy
データ解析
ワークフロー
可視化する
共有データ
ヘルプ
ユーザー
Using 1%
Tools
search tools
on RNA-Seq transcripts
Sailfish transcript quantification from RNA-seq data
goseq tests for overrepresented gene categories
DESeq2 Determines differentially expressed features from count tables
RNA STAR Gapped-read mapper for RNA-seq data
cummeRbund visualize Cuffdiff output
htseq-count - Count aligned reads in a BAM file that overlap features in a GFF file
GffCompare compare assembled transcripts to a reference annotation
Cuffmerge merge together several Cufflinks assemblies
Executed htseq-count and successfully added 3 jobs to the queue.
The tool uses 4 inputs:
47: Bowtie2 on data 43, data 28, and data 27: aligned reads (BAM)
46: Bowtie2 on data 43, data 24, and data 23: aligned reads (BAM)
45: Bowtie2 on data 43, data 20, and data 19: aligned reads (BAM)
...
It produces 6 outputs:
53: htseq-count on data 44 and data 47 (no feature)
History
search datasets
GSE107337_3samples
53 shown
4.52 GB
53: htseq-count on data 44 and data 47 (no feature)
52: htseq-count on data 44 and data 47
51: htseq-count on data 44 and data 46 (no feature)
50: htseq-count on data 44 and data 46
49: htseq-count on data 4

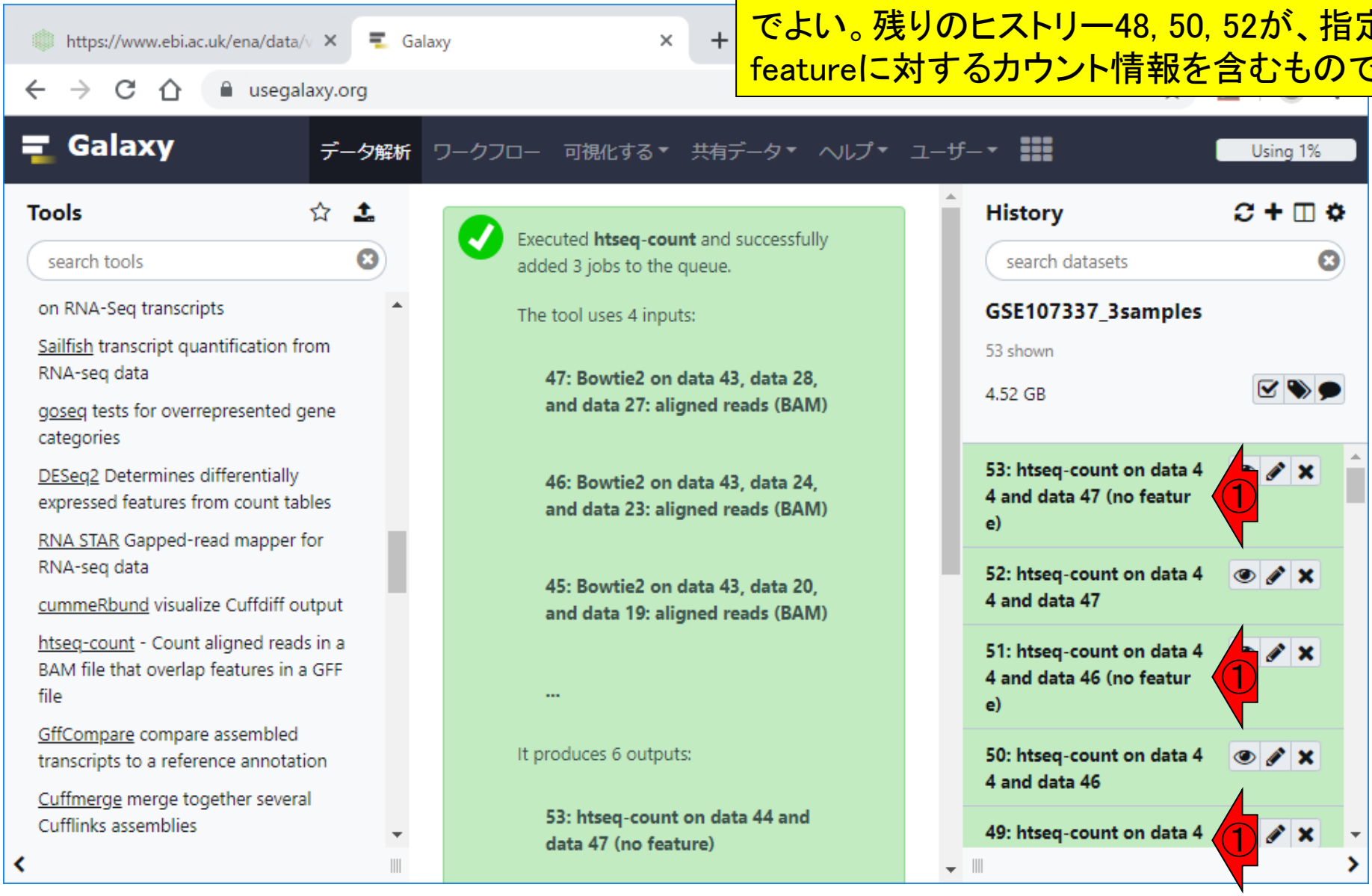
実行完了。約30分かかりました。

W8-7：実行完了



W8-8：解説

計6ファイルの結果のうち、①no featureが付加されたもの（ヒストリー49, 51, 53）は、ほしいカウントデータ以外の統計情報に相当するものなので基本無視でよい。残りのヒストリー48, 50, 52が、指定したfeatureに対するカウント情報を含むものです。

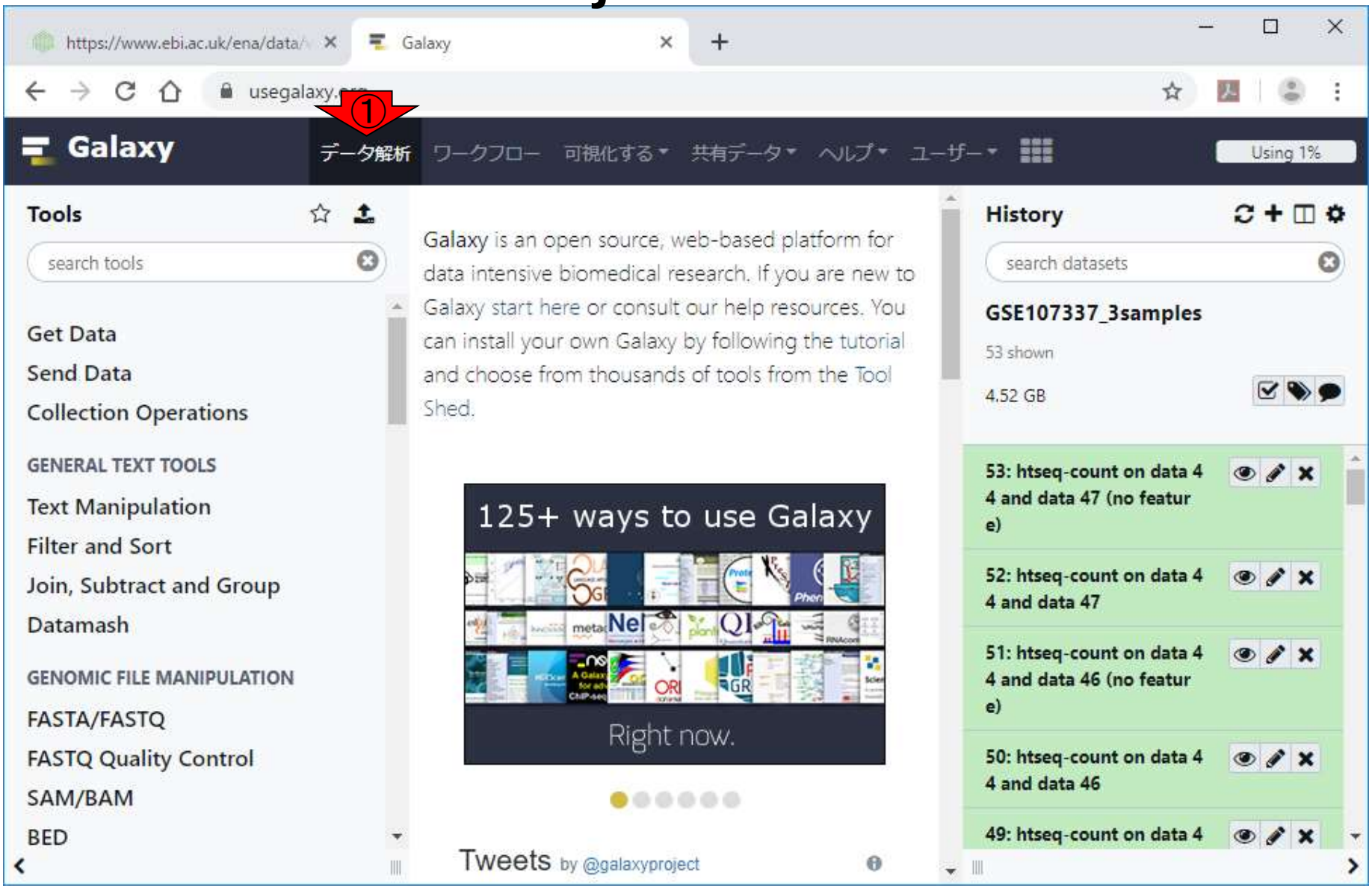


Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

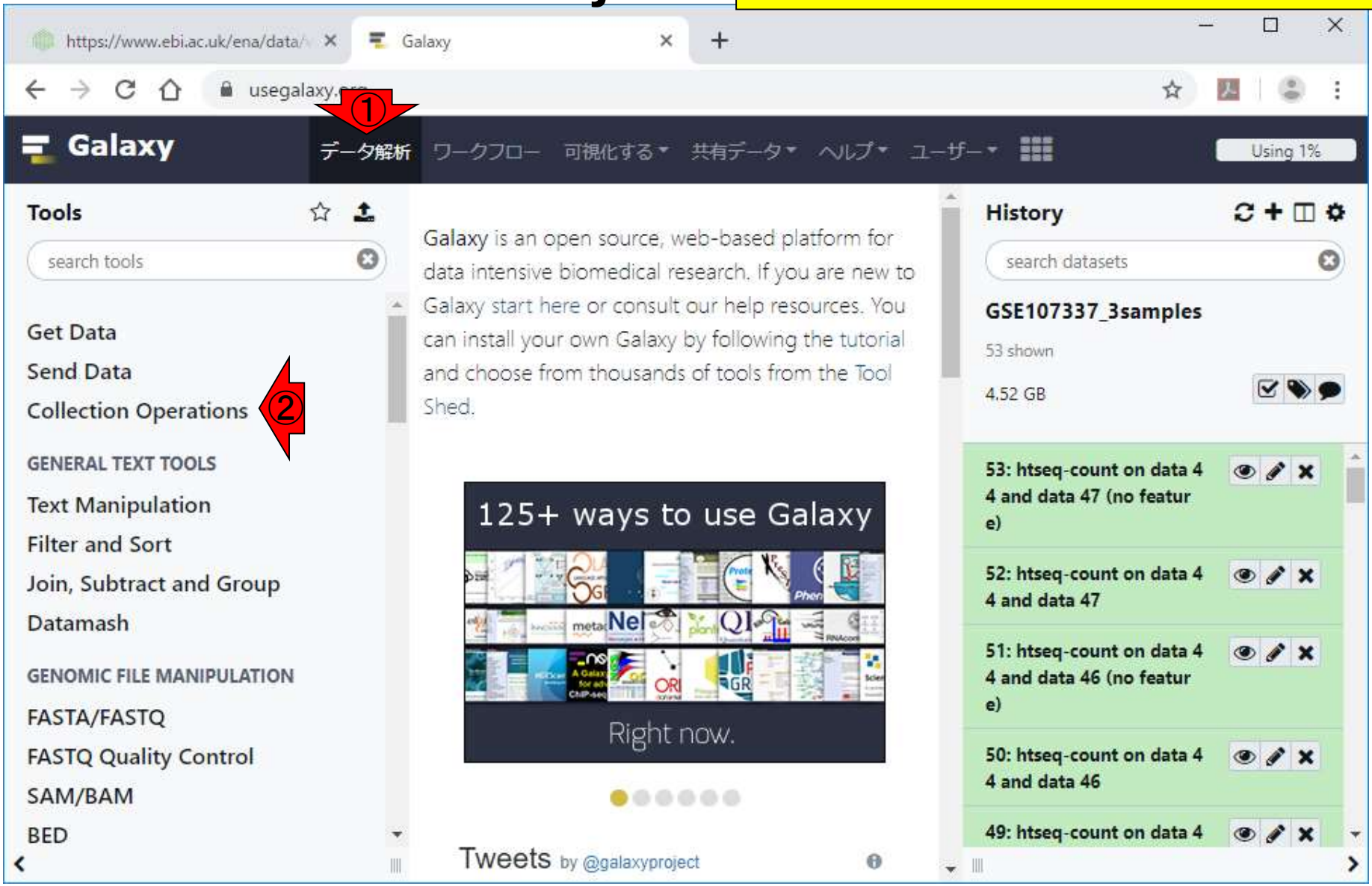
W9-1：Column join

次はカウントデータ取得結果の連結。①を押してヒストリーパネル以外をリフレッシュした後の状態。



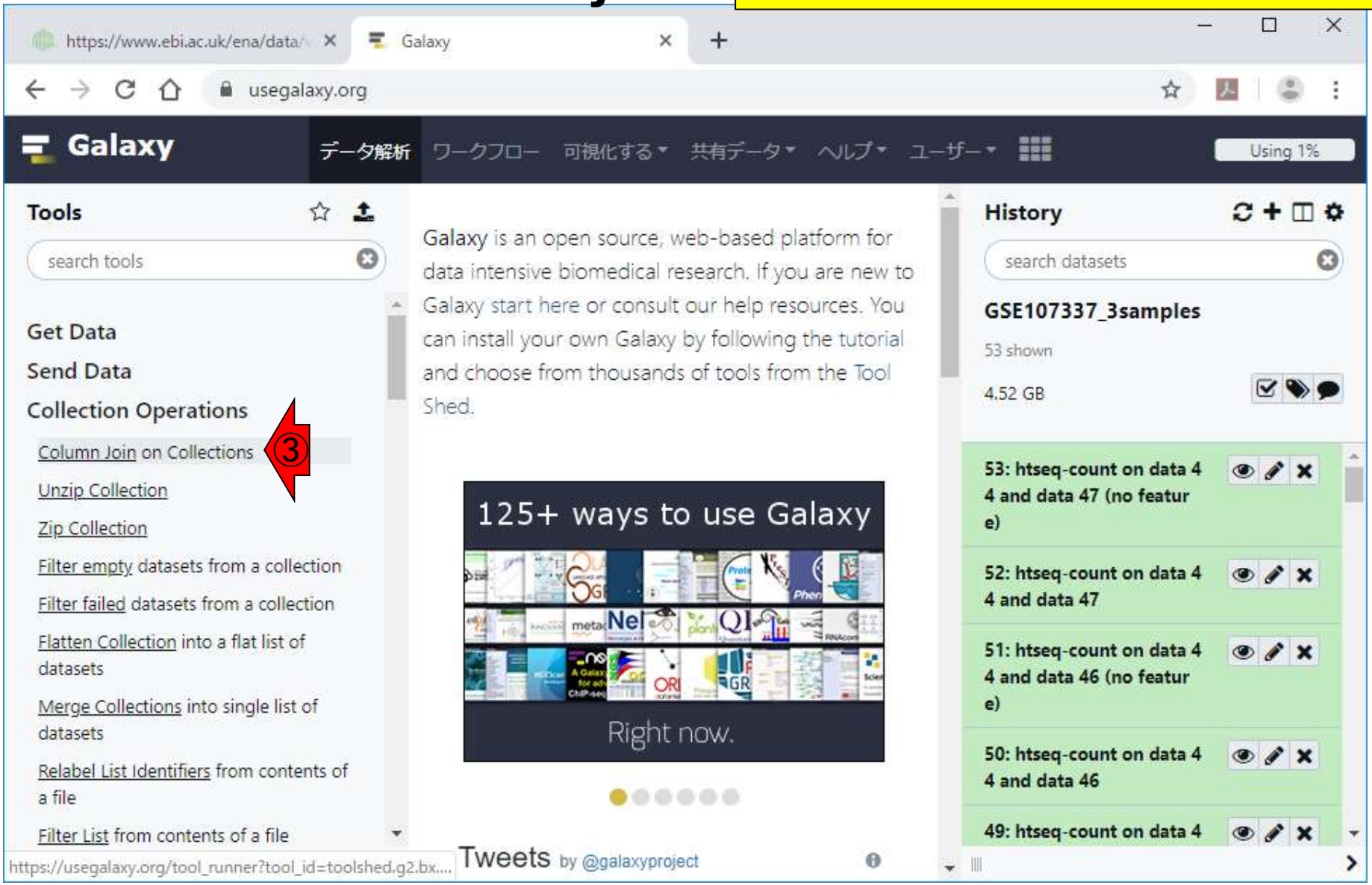
W9-1：Column join

次はカウントデータ取得結果の連結。①を押してヒストリーパネル以外をリフレッシュした後の状態。②Collection Operations。



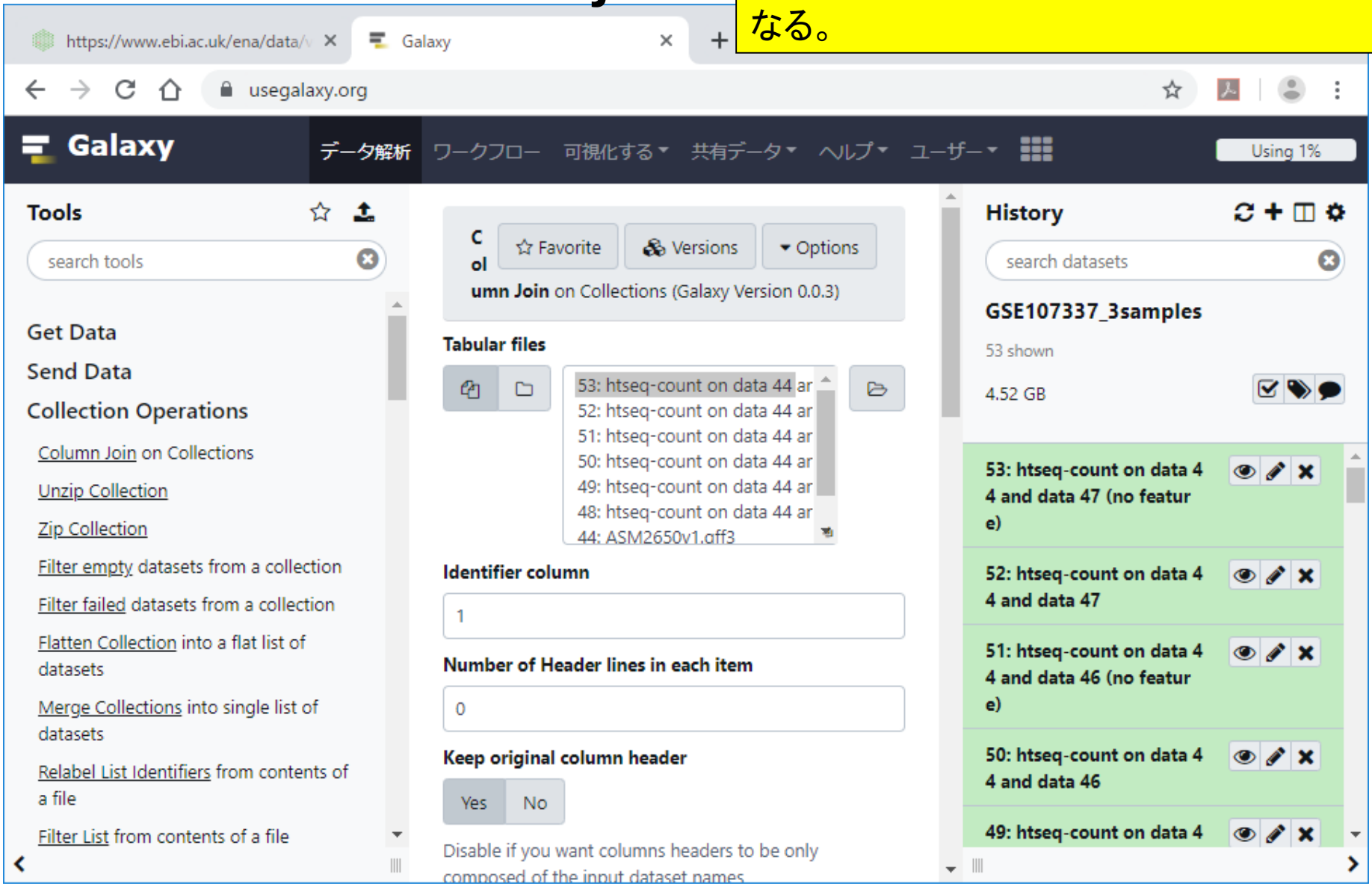
W9-1：Column join

次はカウントデータ取得結果の連結。①を押してヒストリーパネル以外をリフレッシュした後の状態。②Collection Operations。③Column join。



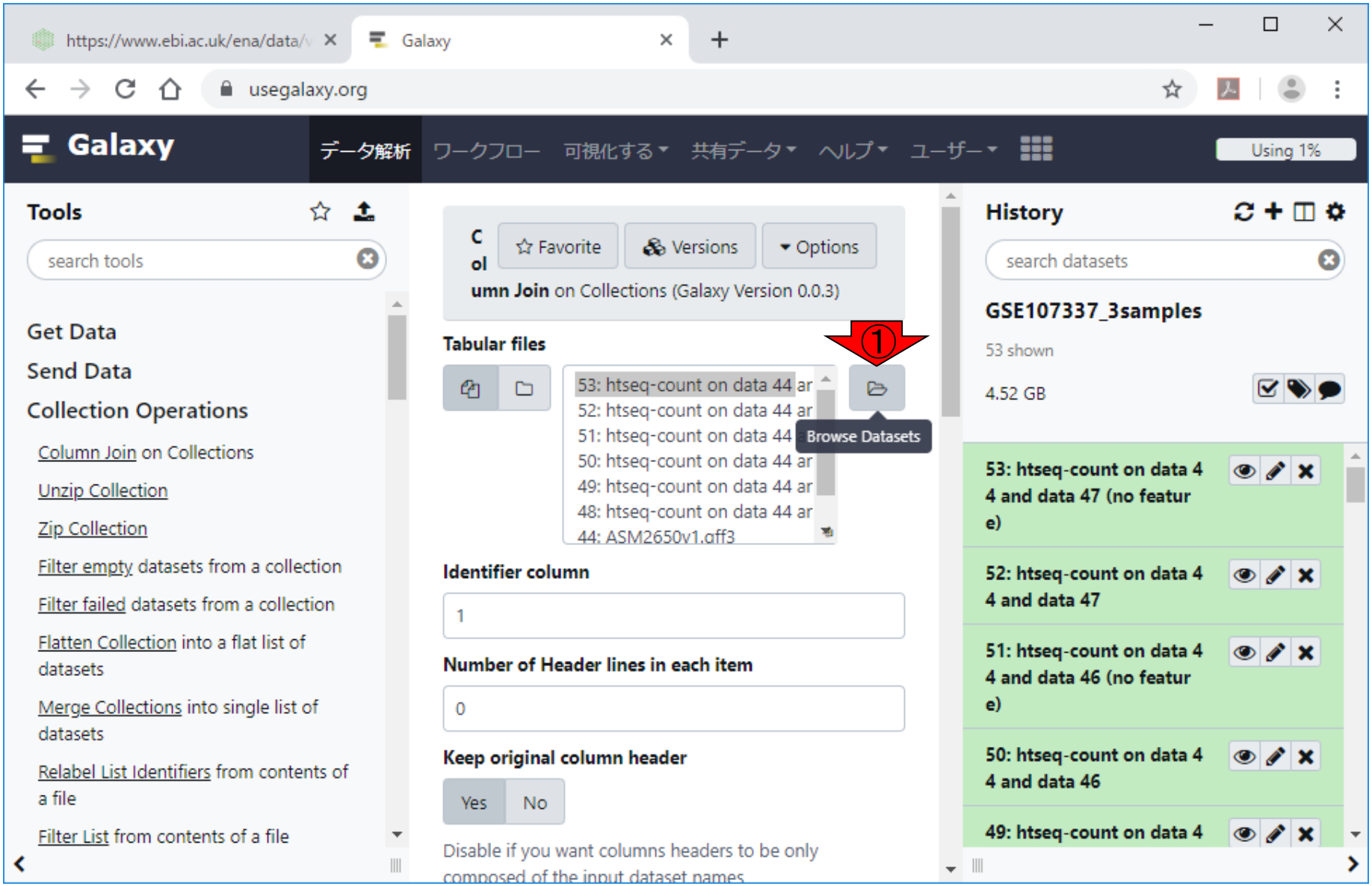
W9-1：Column join

次はカウントデータ取得結果の連結。①を押してヒストリーパネル以外をリフレッシュした後の状態。②Collection Operations。③Column join。こんな感じになる。



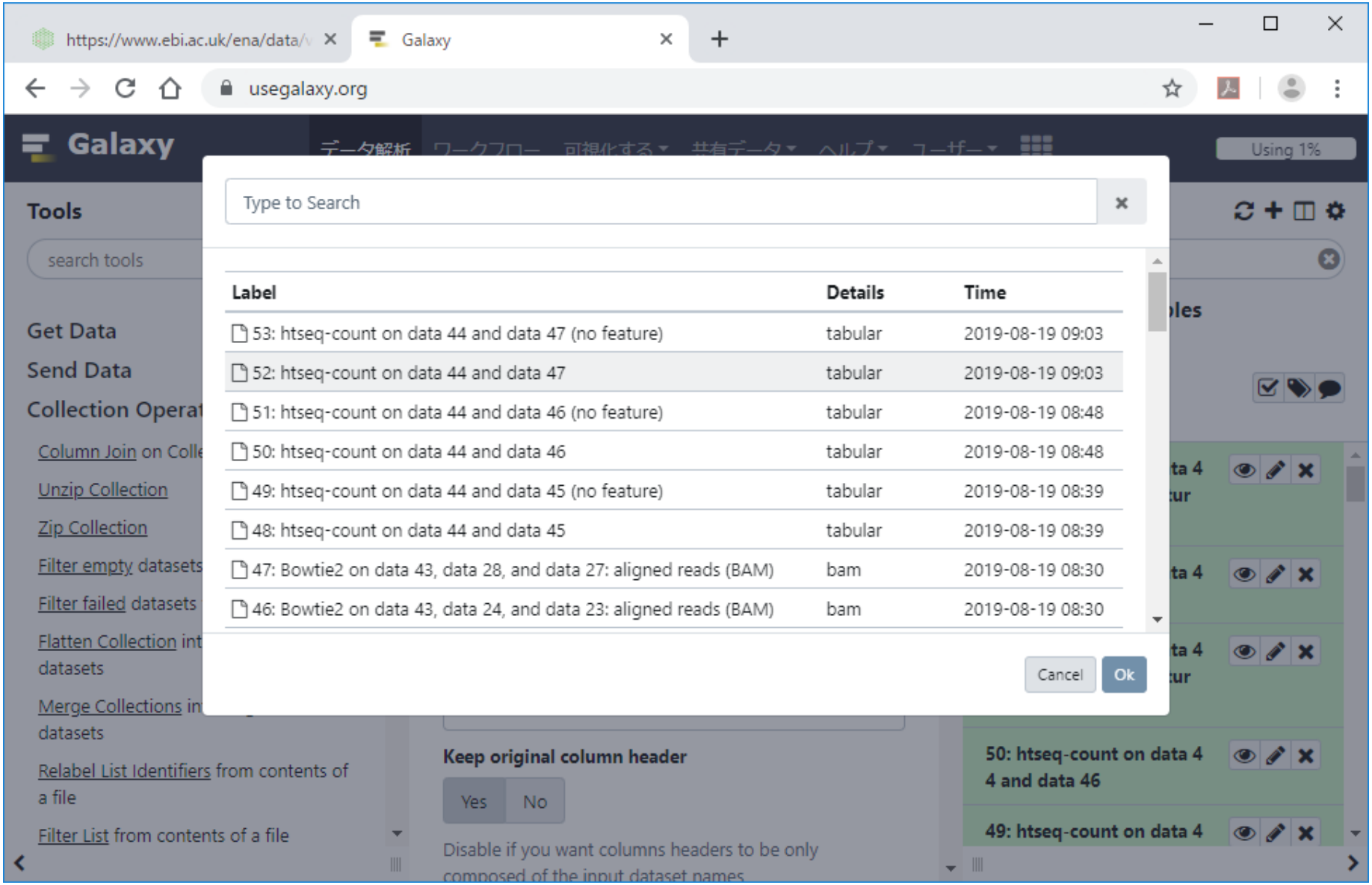
①Browse Datasets。

W9-2：連結対象指定



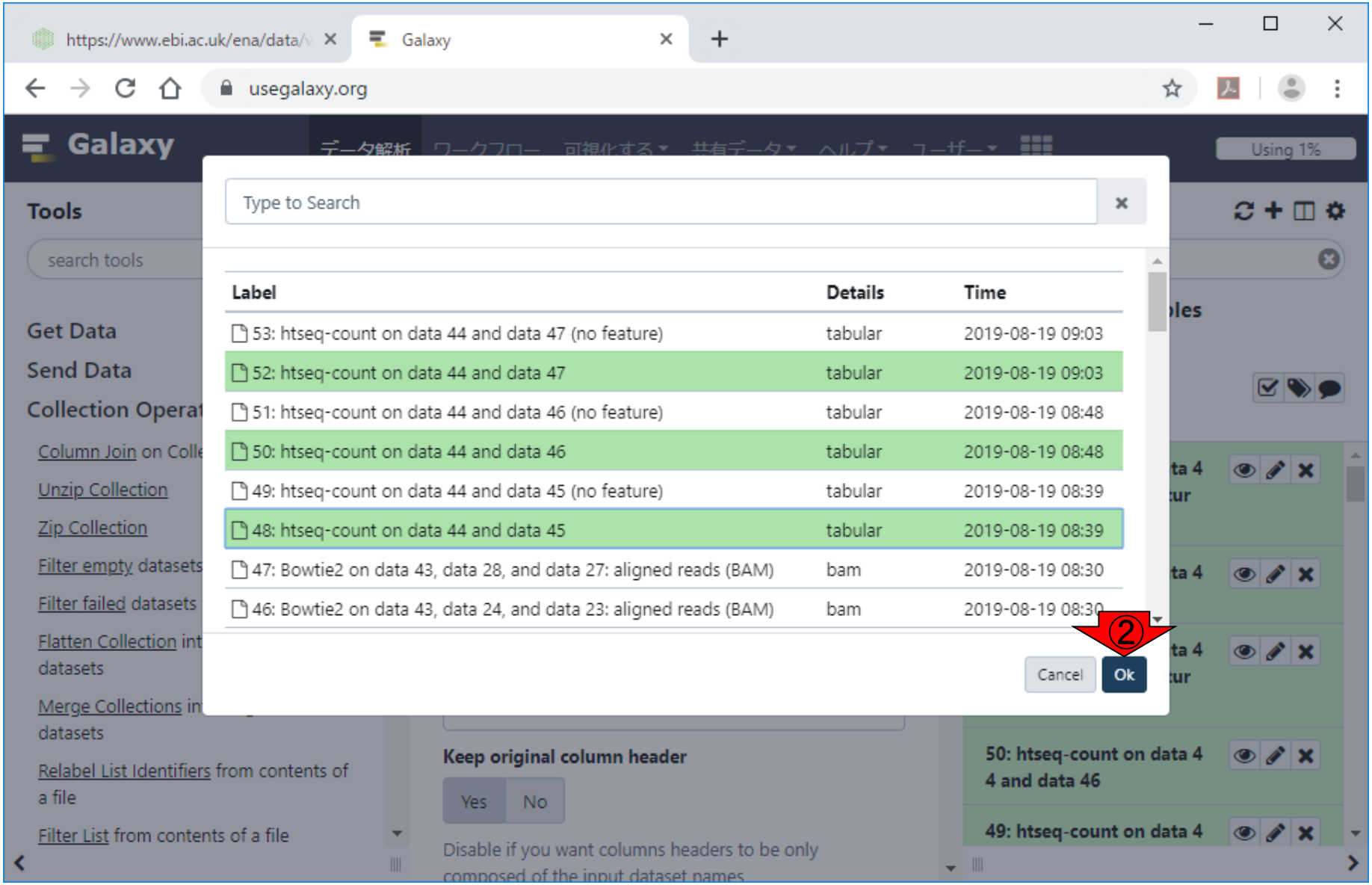
W9-2：連結対象指定

①Browse Datasets。連結したいものはヒストリー48, 50, 52なので…。



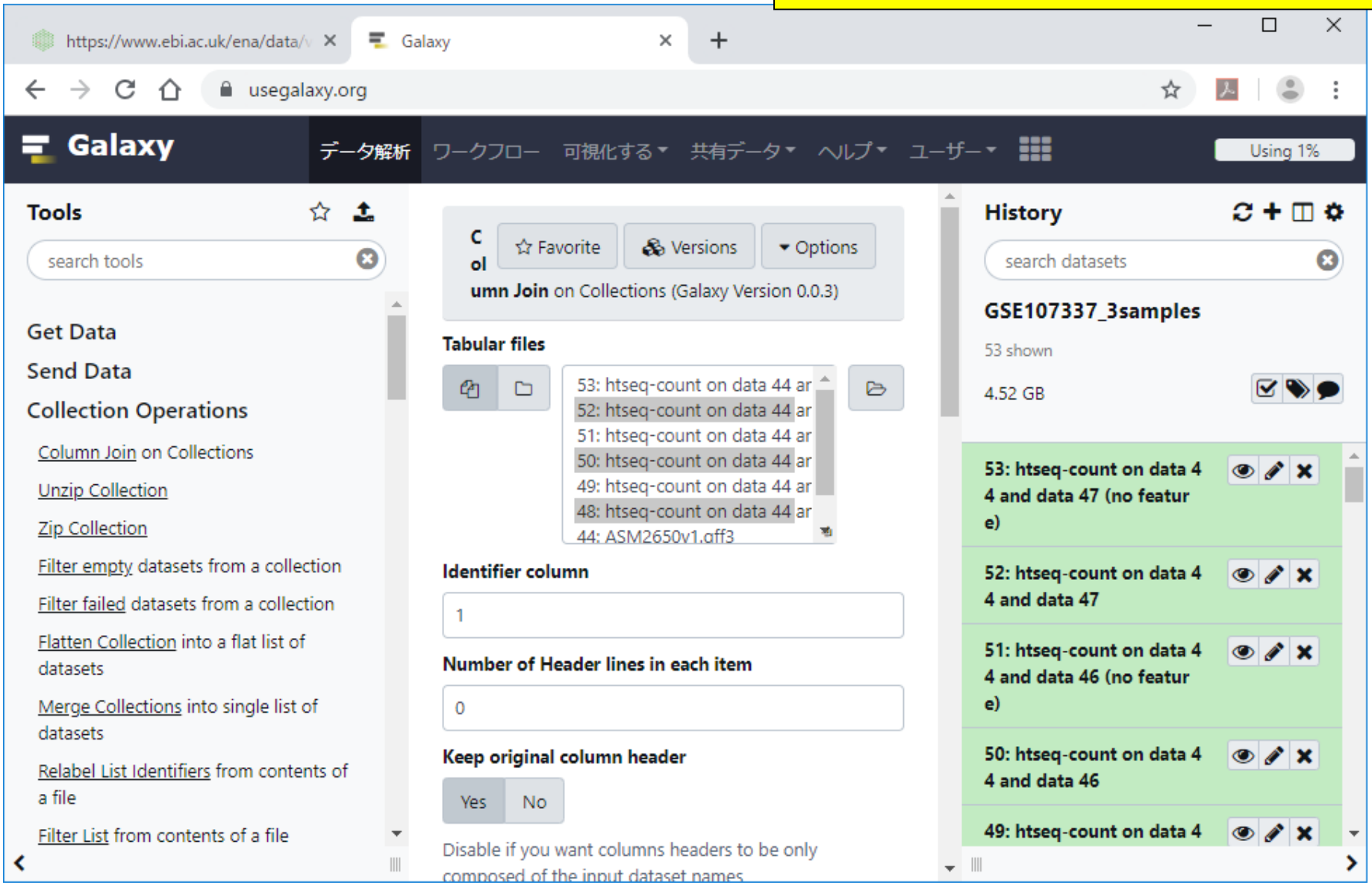
W9-2：連結対象指定

①Browse Datasets。連結したいものはヒストリー48, 50, 52なので…それらを指定して、②Ok。



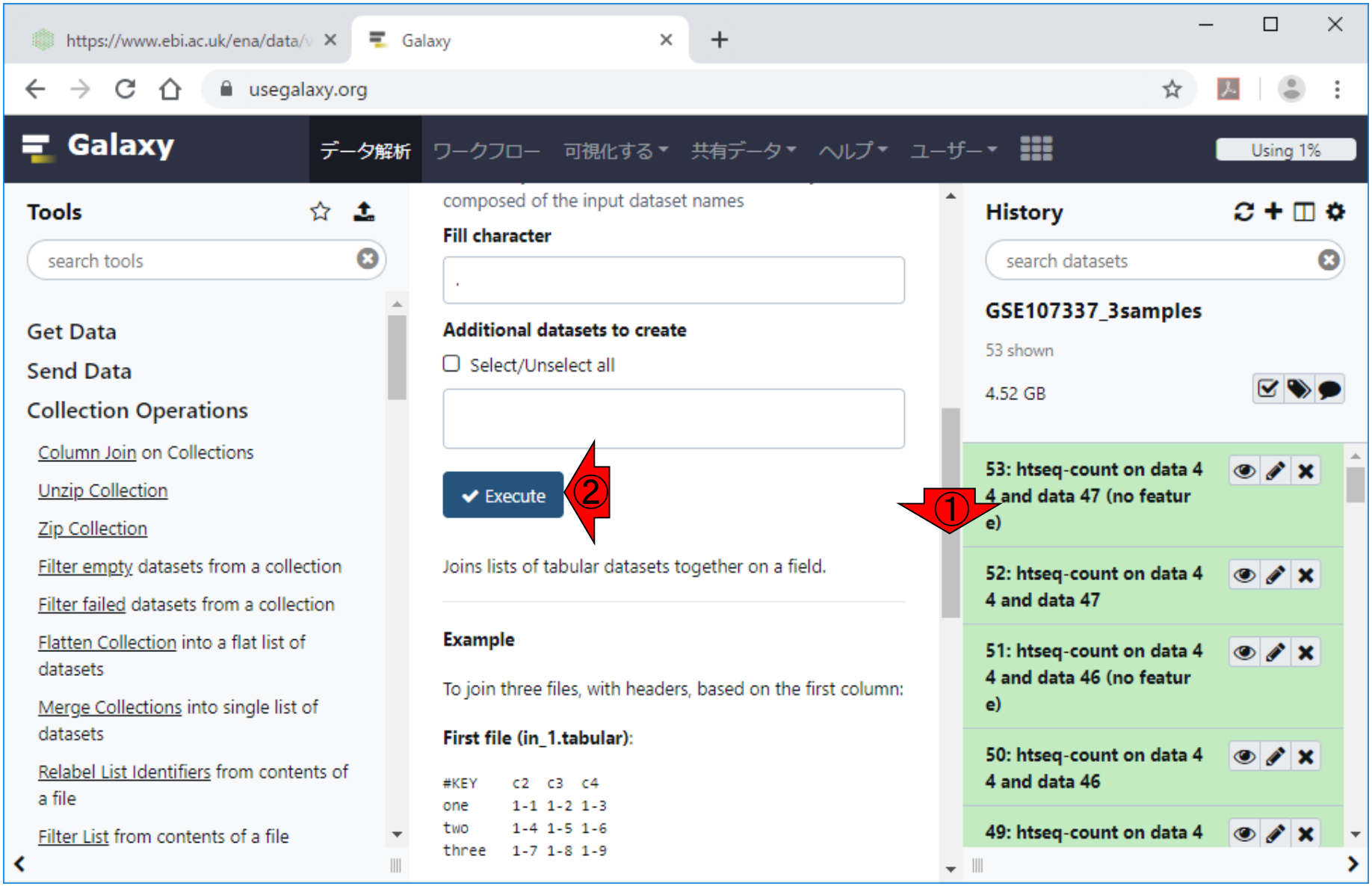
W9-2：連結対象指定

①Browse Datasets。連結したいものはヒストリー48, 50, 52なので…それらを指定して、②Ok。こんな感じになる。



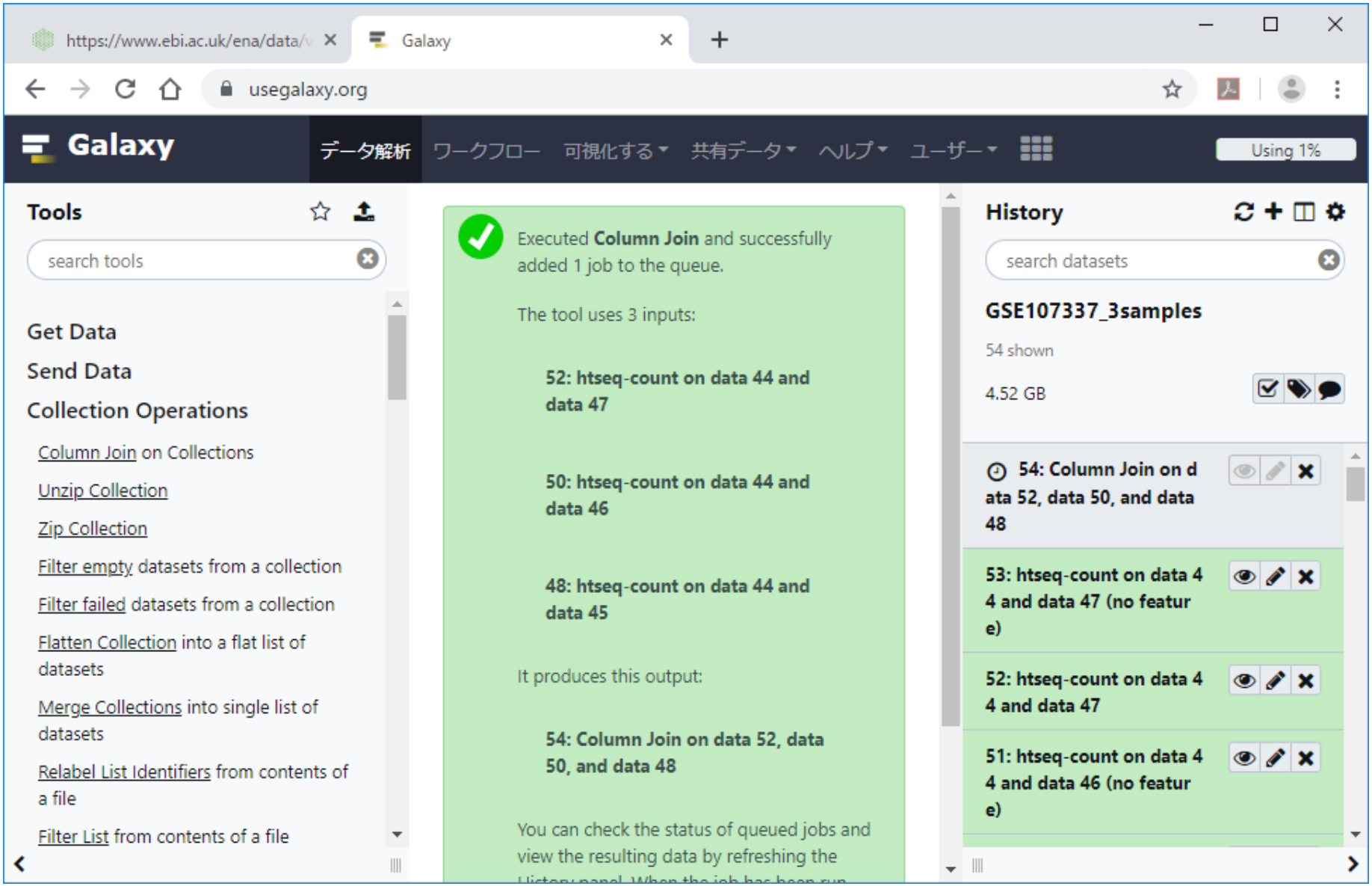
①下部に移動して、②実行。

W9-3：実行



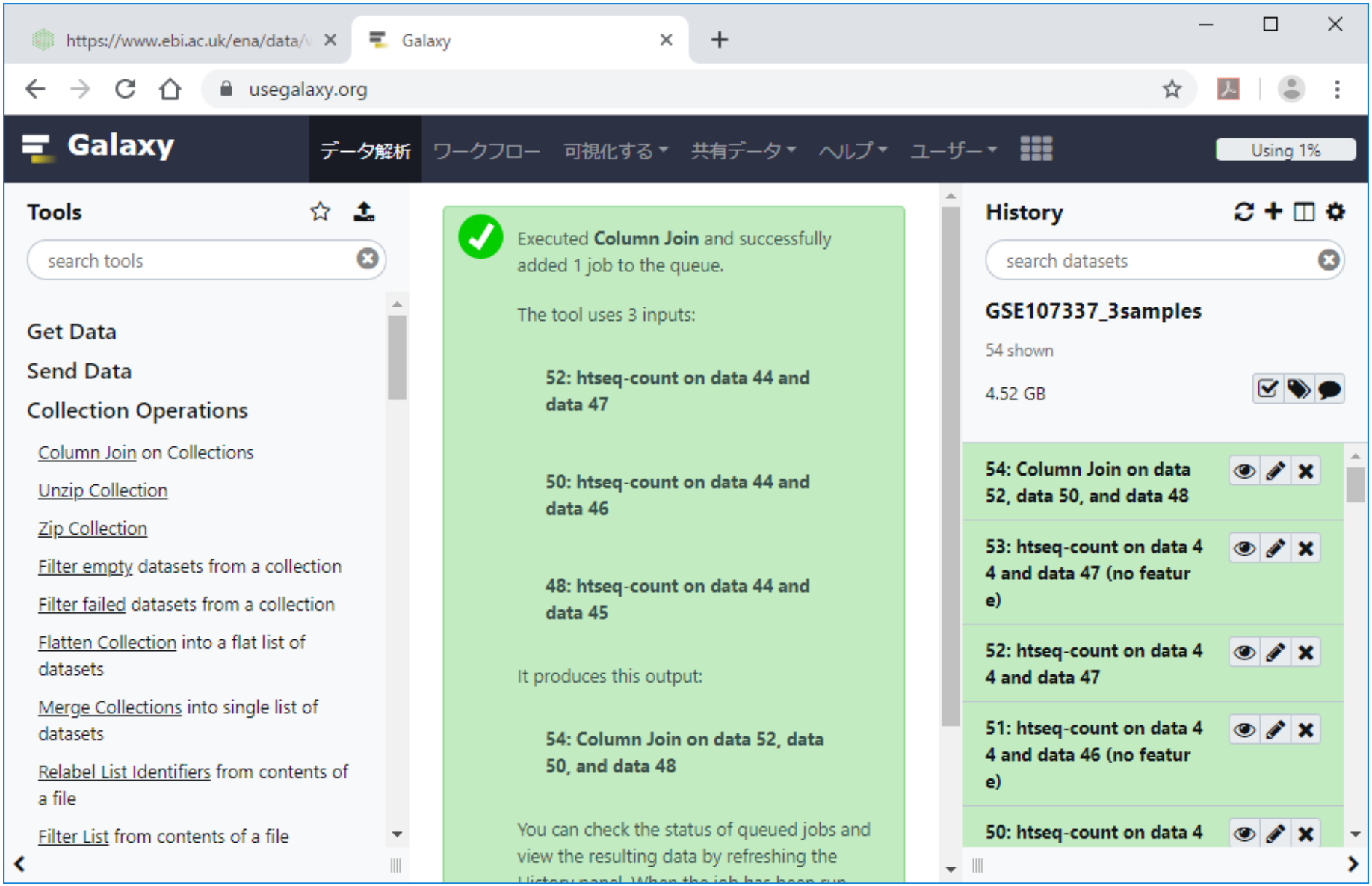
①下部に移動して、②実行。実行中。

W9-4：実行中



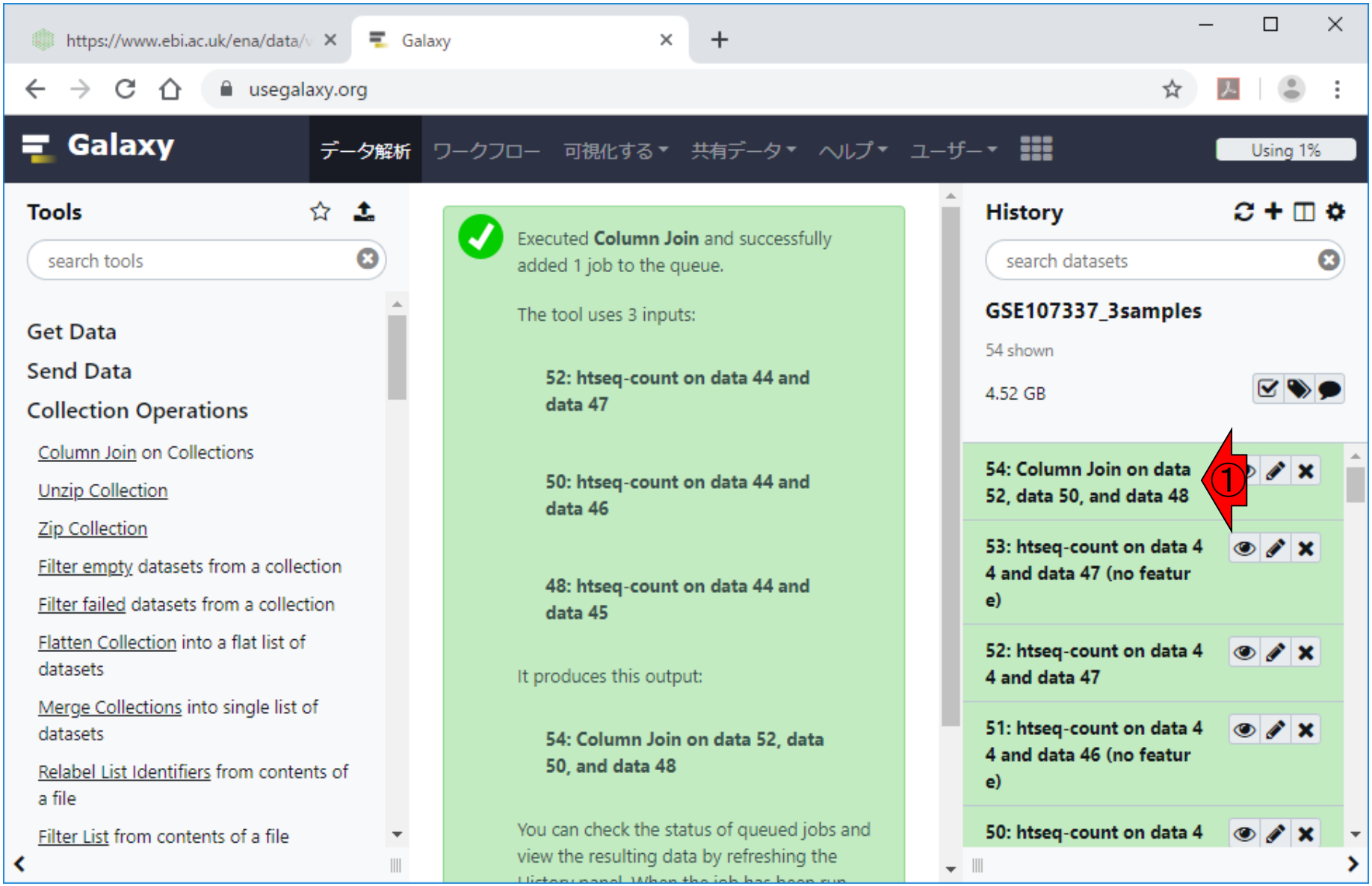
①下部に移動して、②実行。実行中。実行完了。

W9-5：実行完了



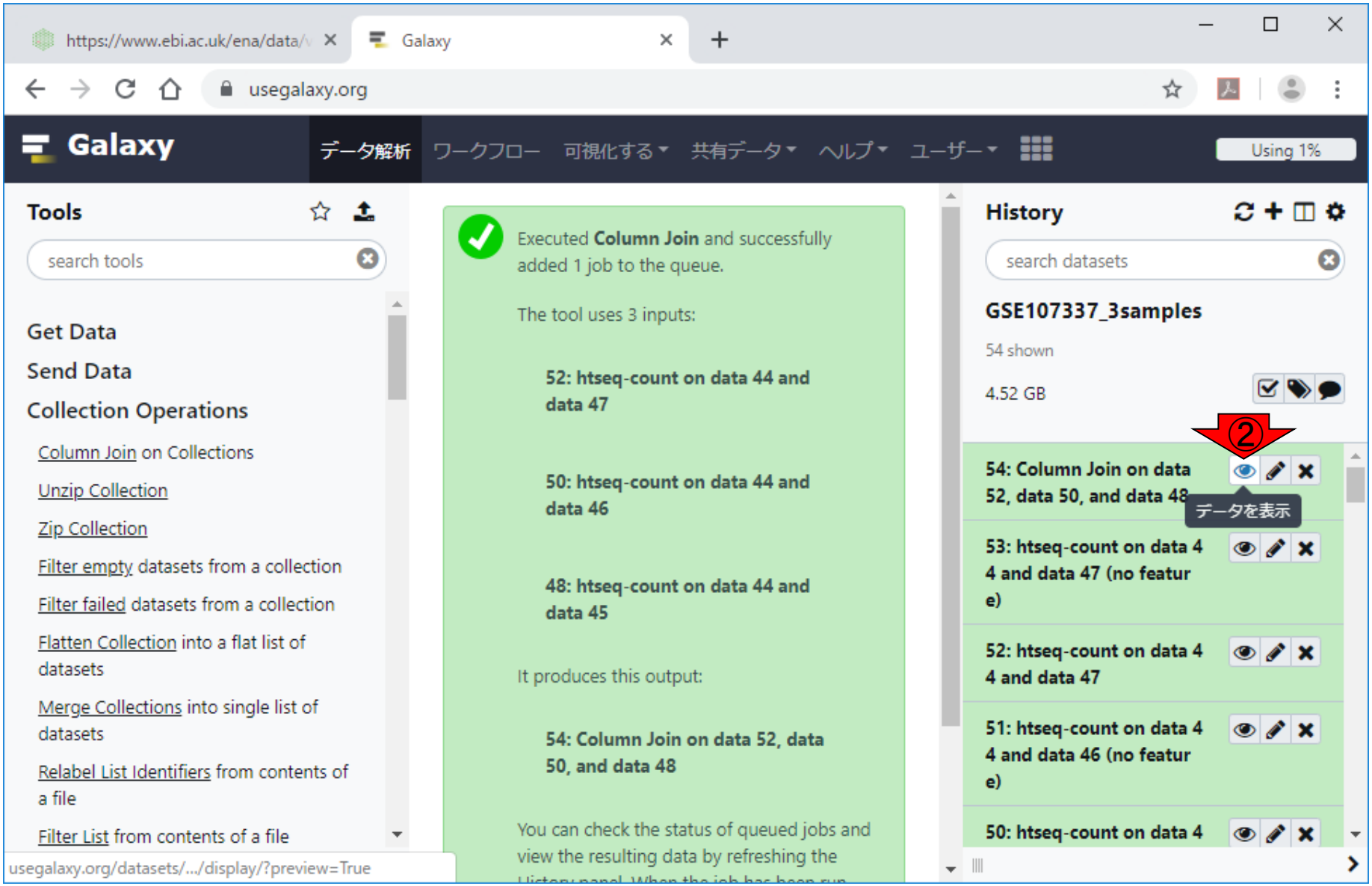
出力結果は、①1つ（ヒストリー54）。

W9-6：解説



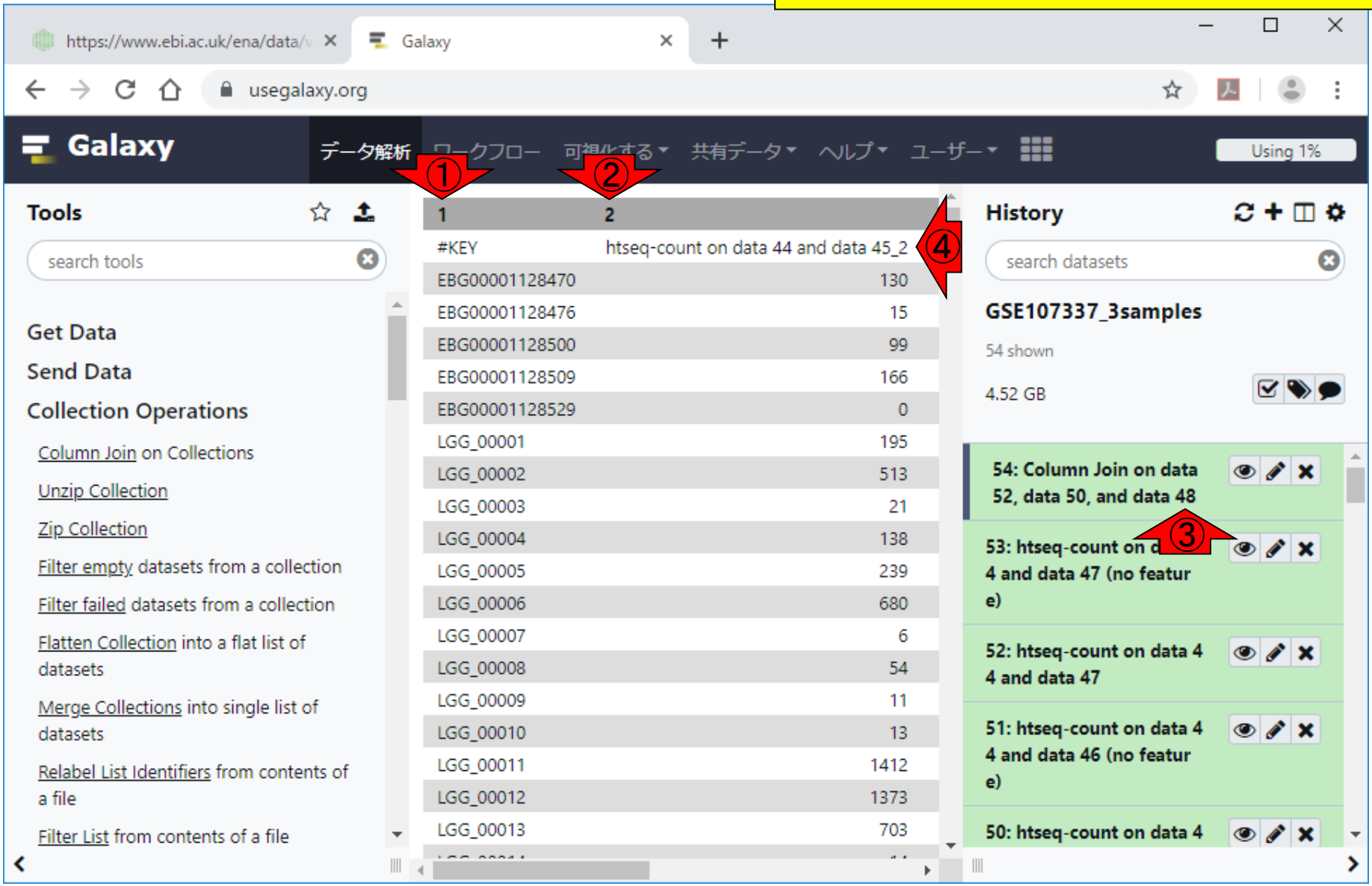
W9-6：解説

出力結果は、①1つ（ヒストリー54）。②データを中央パネル上で表示させて、大まかに眺める。



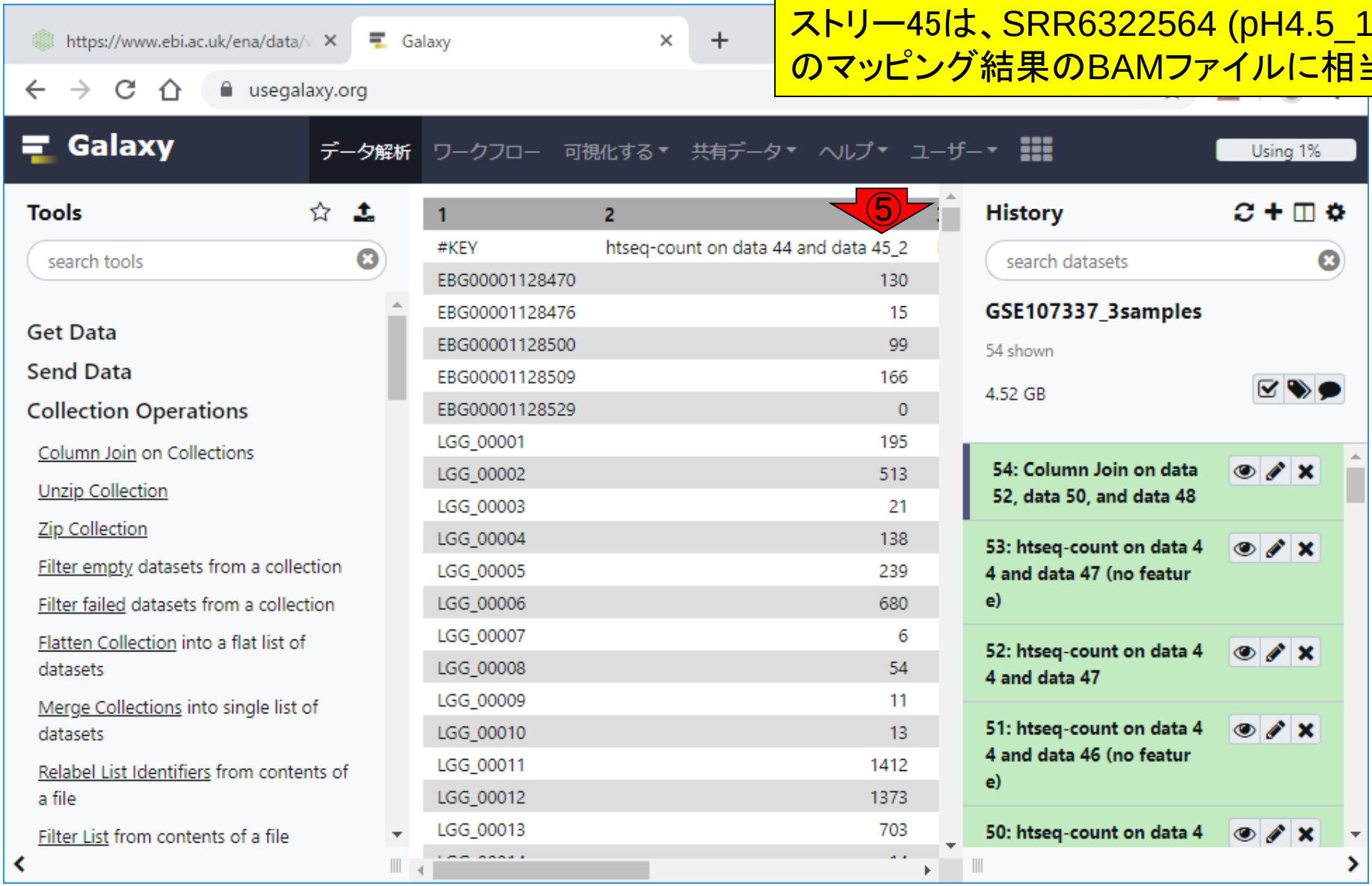
W9-7：行列

こんな感じになる。①1列目はgene ID情報。②2列目は、③ヒストリー48の情報。ヒストリー48の名前が④で見えているものになります。



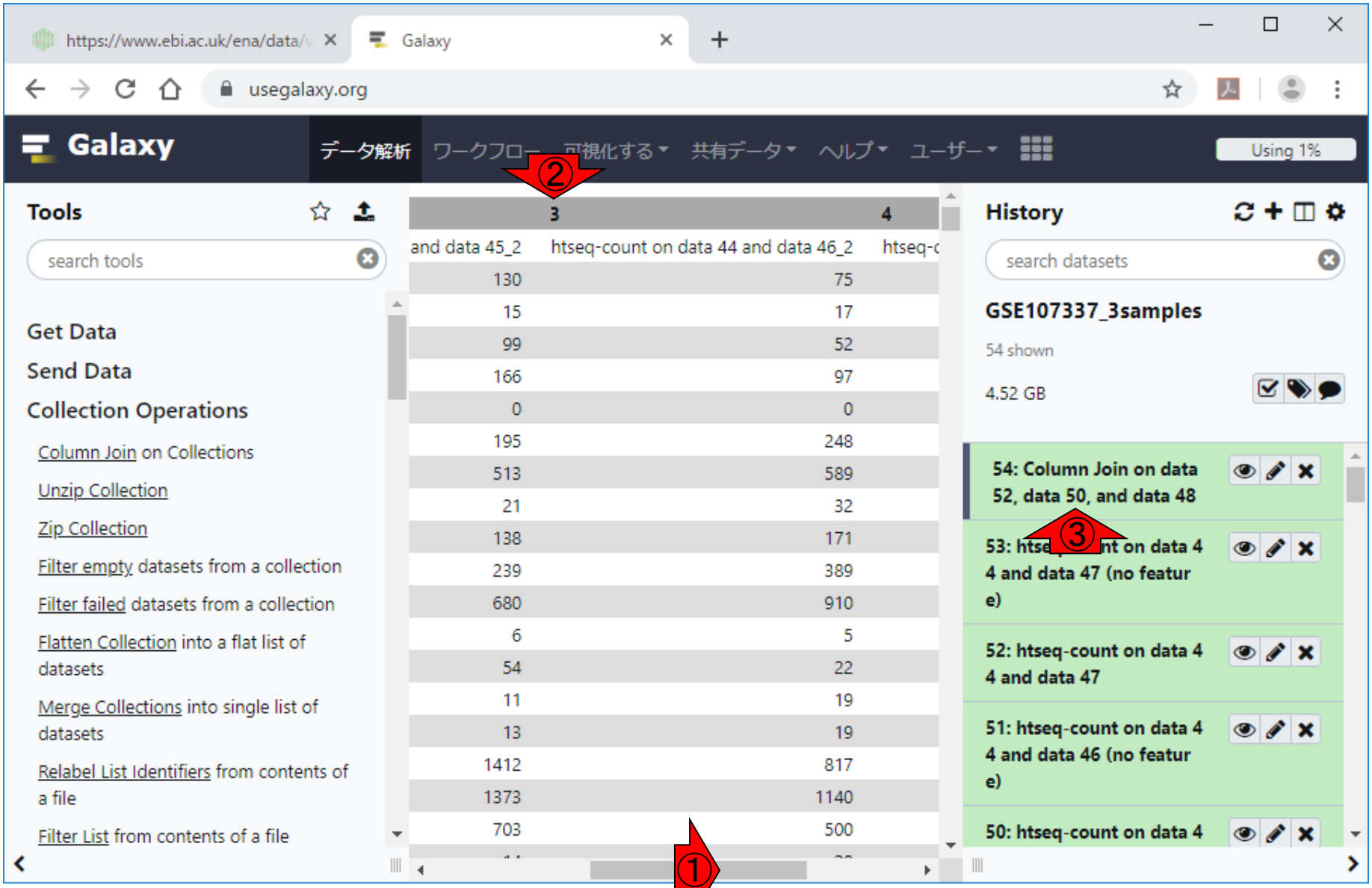
W9-7：行列

こんな感じになる。①1列目はgene ID情報。②2列目は、③ヒストリー48の情報。ヒストリー48の名前が④で見えているものになります。ちなみに、⑤ヒストリー45は、SRR6322564 (pH4.5_1h_rep3)のマッピング結果のBAMファイルに相当します。



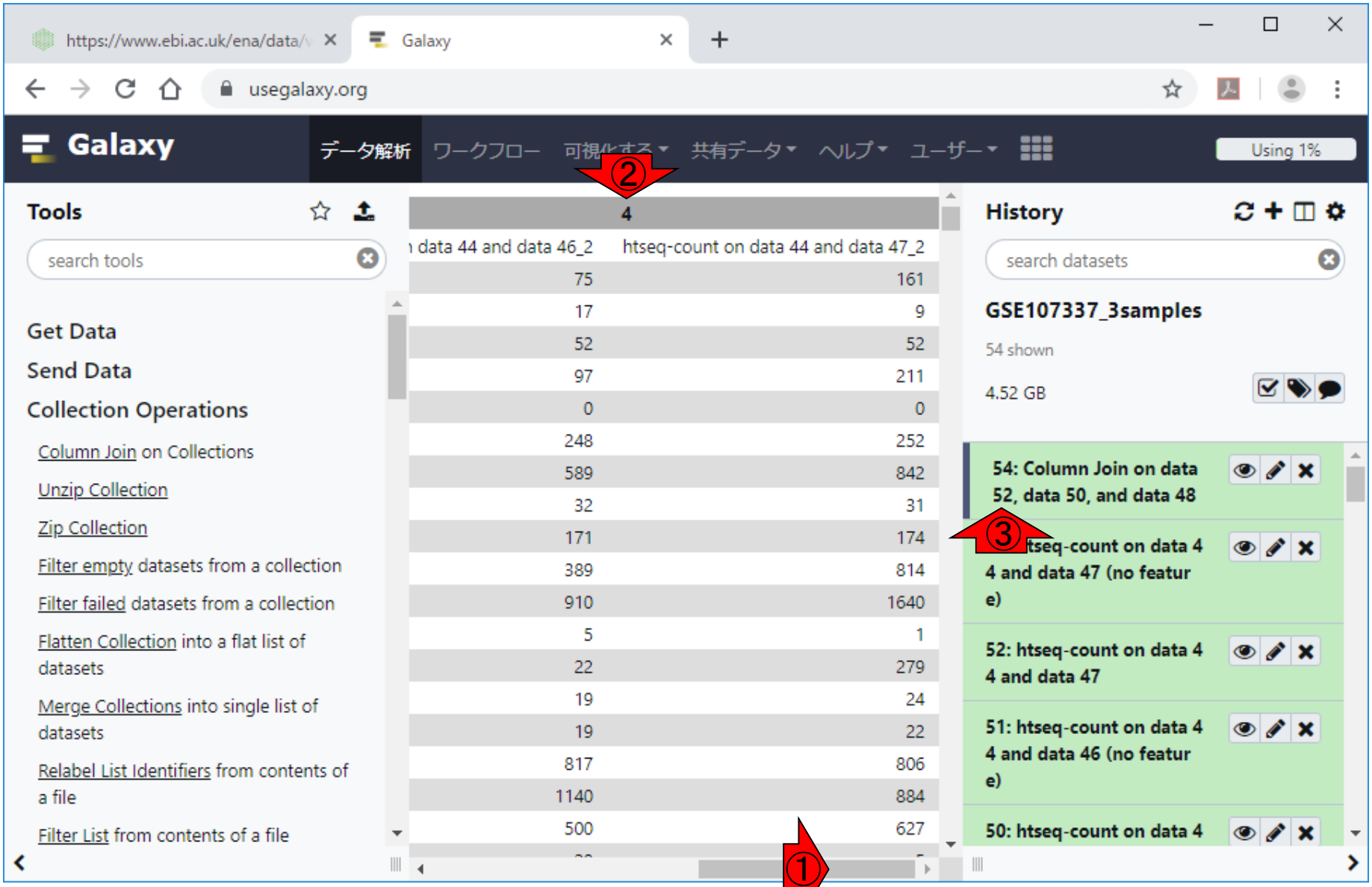
W9-7：行列

①中央パネルを右のほうに移動させたところ。②3列目が、③ヒストリー50の情報。



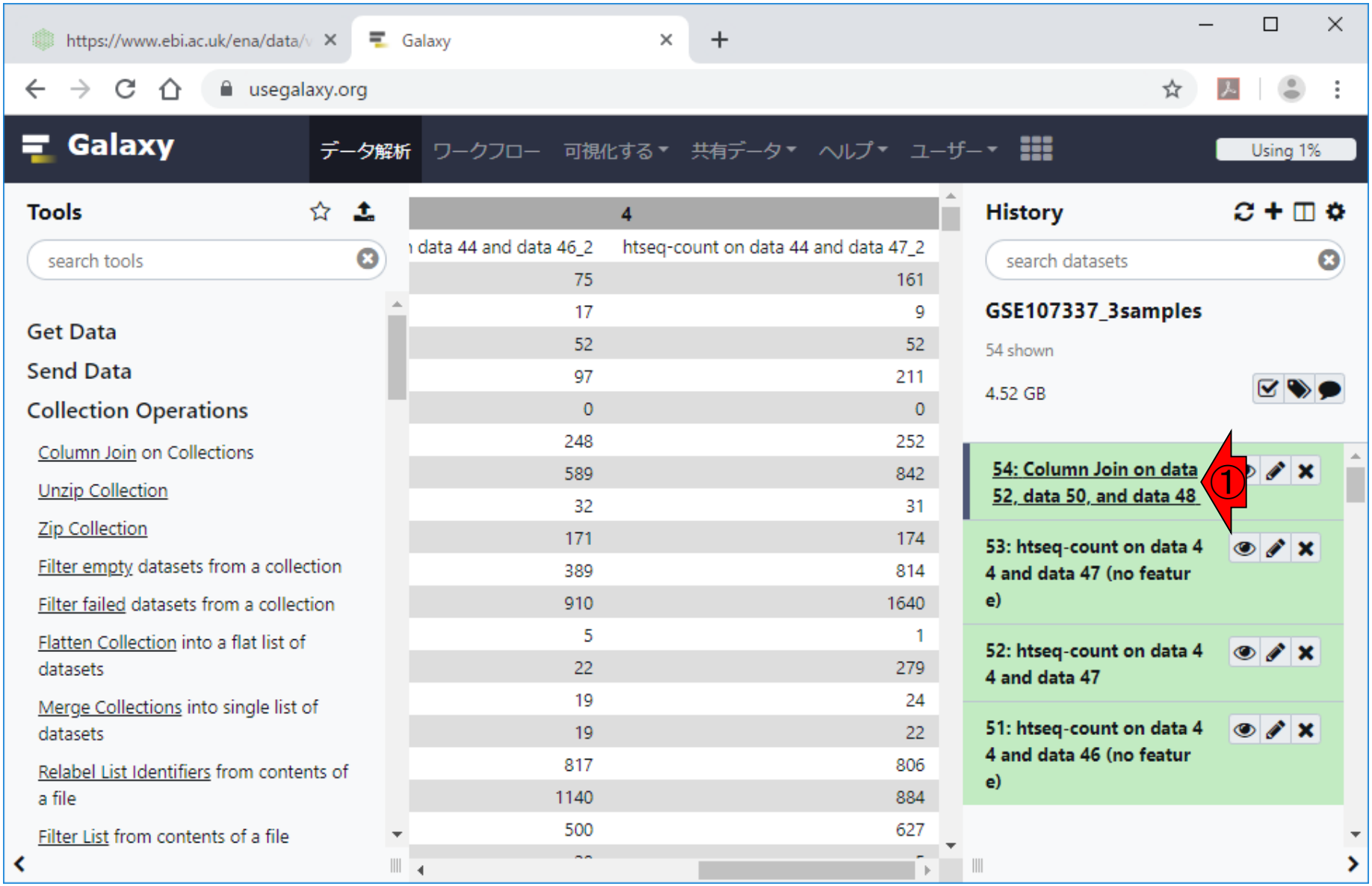
W9-7：行列

①中央パネルを一番右まで移動させたところ。②4列目が、③ヒストリー52の情報。



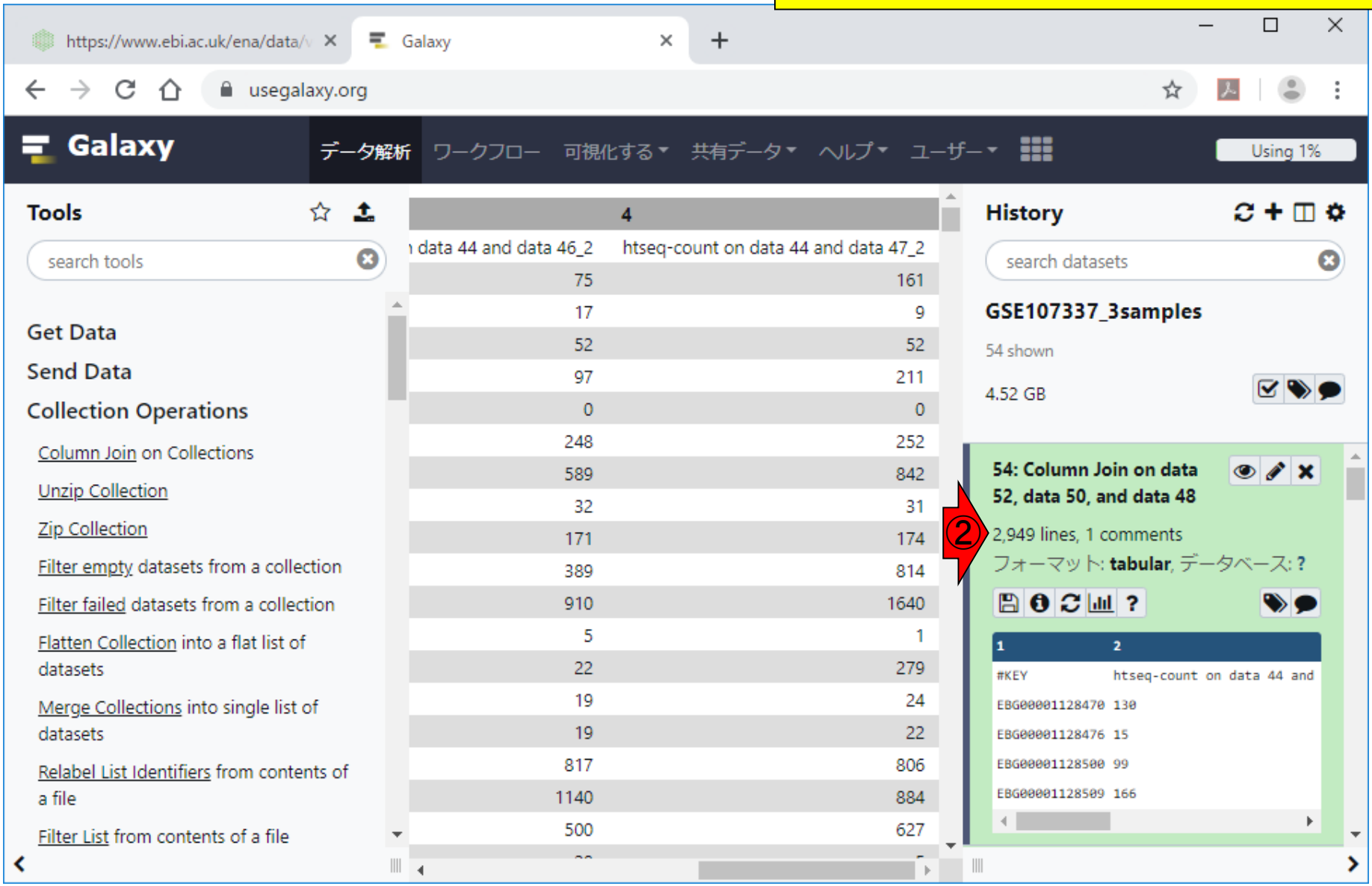
①をクリックして保存する。

W9-8：保存



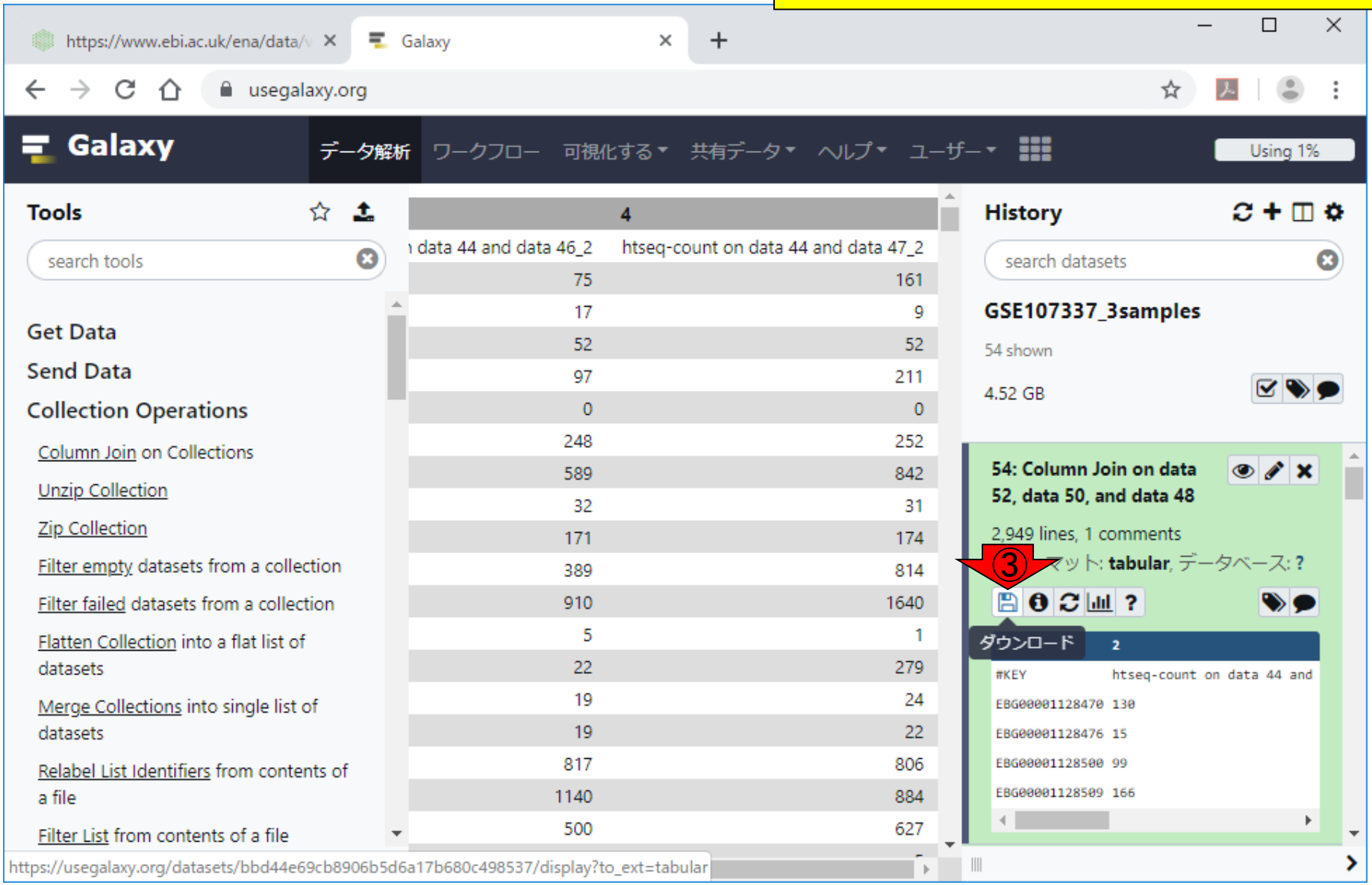
W9-8：保存

①をクリックして保存する。②（ヘッダー行を除く）行数は2,949。つまり、2,949遺伝子からなるカウントデータだということがわかる。



W9-8：保存

①をクリックして保存する。②（ヘッダー行を除く）行数は2,949。つまり、2,949遺伝子からなるカウントデータだということがわかる。③ダウンロード。



W9-9：Excelで概観

①Excelで眺めるとこんな感じになります。これは最初のほう。

#KEY	htseq-count on data 44 and data 45_2	htseq-count on data 44 and data 46_2	htseq-count on data 44 and data 47_2
EBG00001128470	130	75	161
EBG00001128476	15	17	9
EBG00001128500	99	52	52
EBG00001128509	166	97	211
EBG00001128529	0	0	0
LGG_00001	195	248	252
LGG_00002	513	589	842
LGG_00003	21	32	31
LGG_00004	138	171	174
LGG_00005	239	389	814
LGG_00006	680	910	1640
LGG_00007	6	5	1
LGG_00008	54	22	279
LGG_00009	11	19	24
LGG_00010	13	19	22
LGG_00011	1412	817	806

W9-9：Excelで概観

①Excelで眺めるとこんな感じになります。これは最後のほう。

#KEY	htseq-count on data 44 and data 45_2	htseq-count on data 44 and data 46_2	htseq-count on data 44 and data 47_2
LGG_02928	51	75	131
LGG_02929	6	5	3
LGG_02930	6	6	5
LGG_02931	40	50	159
LGG_02932	16	8	16
LGG_02933	34	39	53
LGG_02934	111	27	84
LGG_02935	9	9	3
LGG_02936	36	52	49
LGG_02937	156	252	289
LGG_02938	69	131	82
LGG_02939	120	80	141
LGG_02940	386	197	342
LGG_02941	201	112	224
LGG_02942	120	69	72
LGG_02943	2111	499	879
LGG_02944	4255	2668	14963

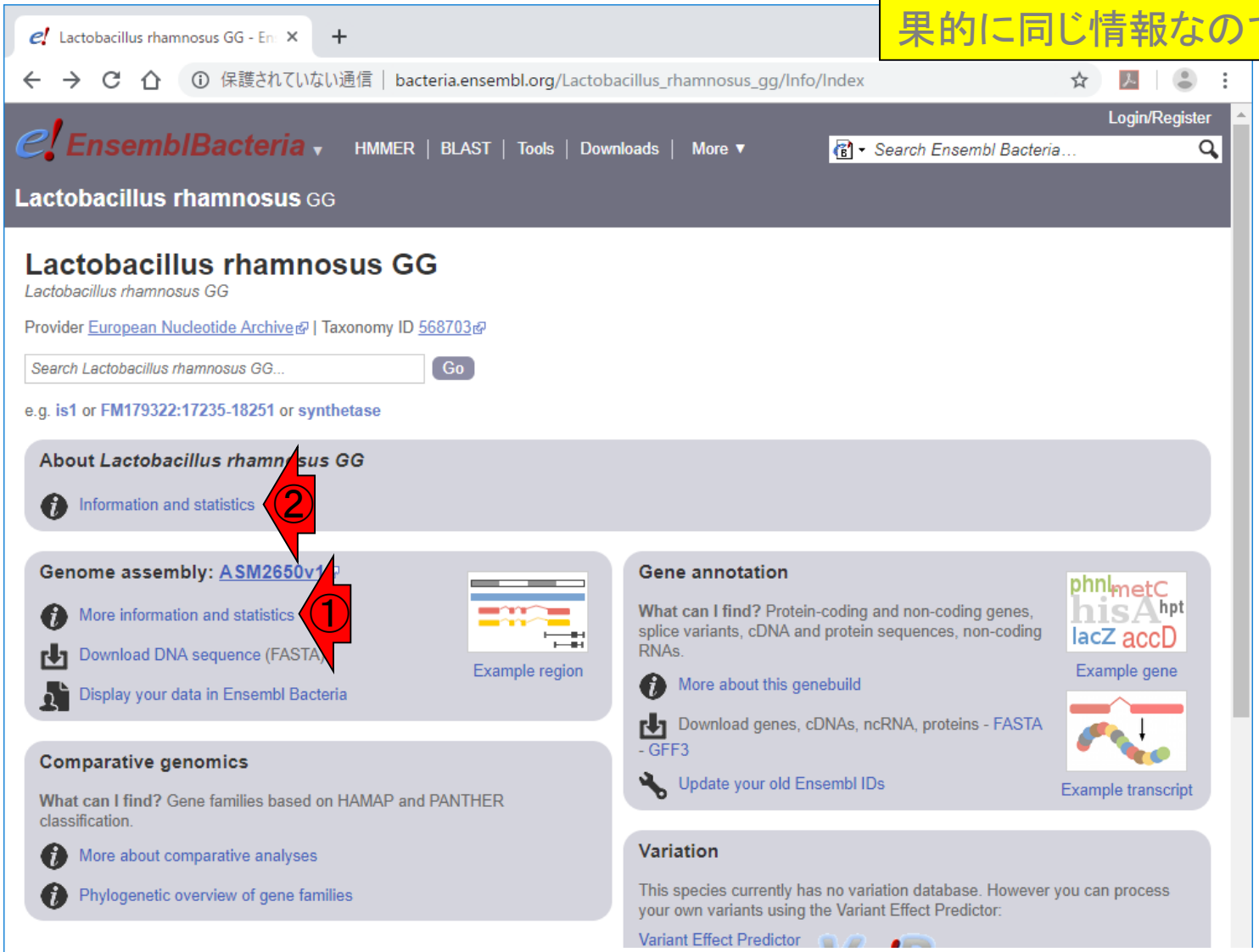
W9-10：サンプル名

①私はこの段階で、後の解析結果を解釈しやすくするため、サンプル名などを手動で変更しておく。

gene_ID	pH4.5_1h_rep3	pH4.5_24h_rep3	pH7_CCG_rep2 ①
EBG00001128470	130	75	161
EBG00001128476	15	17	9
EBG00001128500	99	52	52
EBG00001128509	166	97	211
EBG00001128529	0	0	0
LGG_00001	195	248	252
LGG_00002	513	589	842
LGG_00003	21	32	31
LGG_00004	138	171	174
LGG_00005	239	389	814
LGG_00006	680	910	1640
LGG_00007	6	5	1
LGG_00008	54	22	279
LGG_00009	11	19	24
LGG_00010	13	19	22
LGG_00011	1412	817	806

W9-11：確認

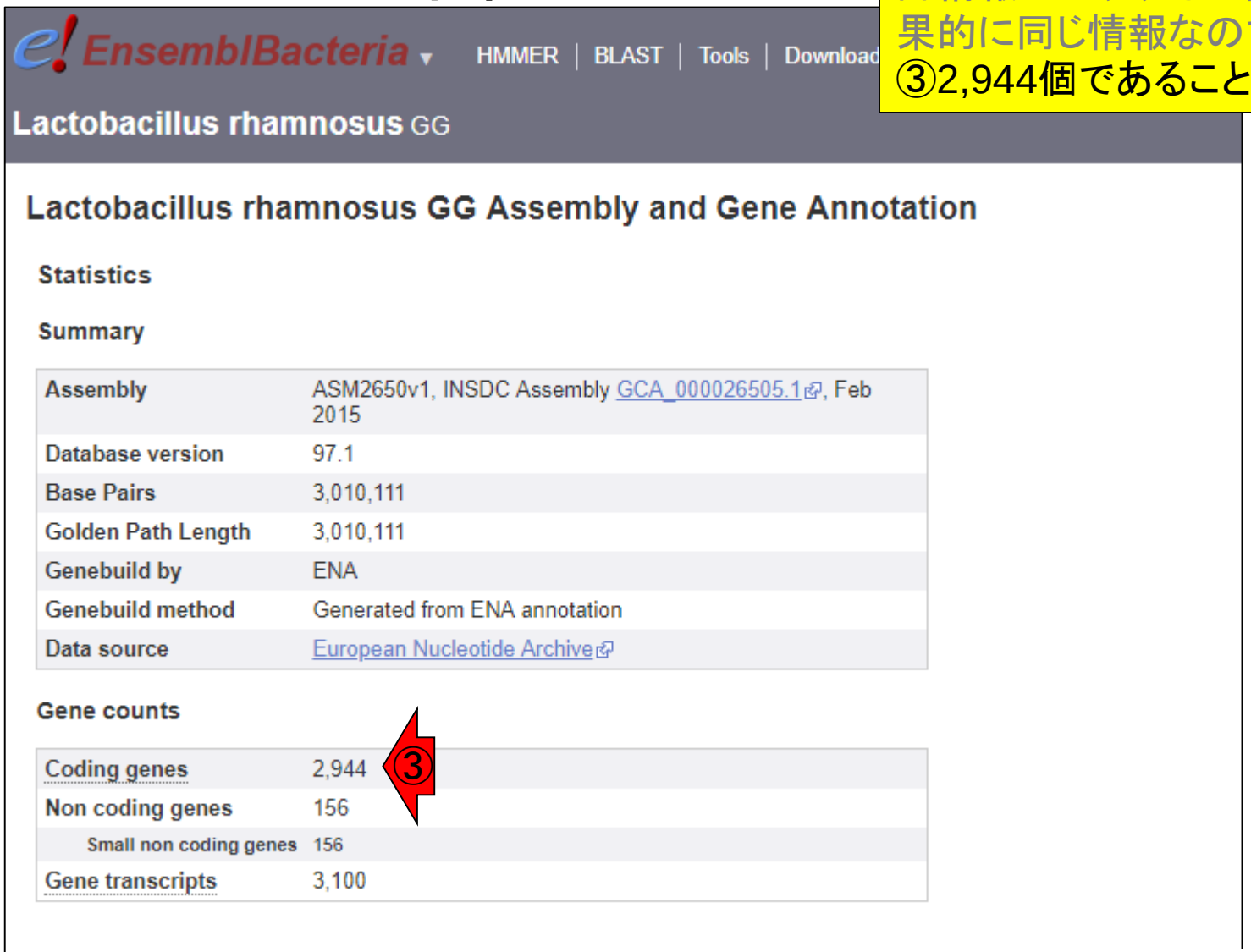
得られたカウントデータ行列の行数（= 2,949）が、Ensembl Bacteria中の*L. rhamnosus GG*の統計情報と一致するか調べるべく、①をクリック。結果的に同じ情報なので、②をクリックしてもよい。



http://bacteria.ensembl.org/Lactobacillus_rhamnosus_gg/Info/Index

W9-11：確認

得られたカウントデータ行列の行数（= 2,949）が、Ensembl Bacteria中の*L. rhamnosus GG*の統計情報と一致するか調べるべく、①をクリック。結果的に同じ情報なので、②をクリックしてもよい。
③2,944個であることがわかります。



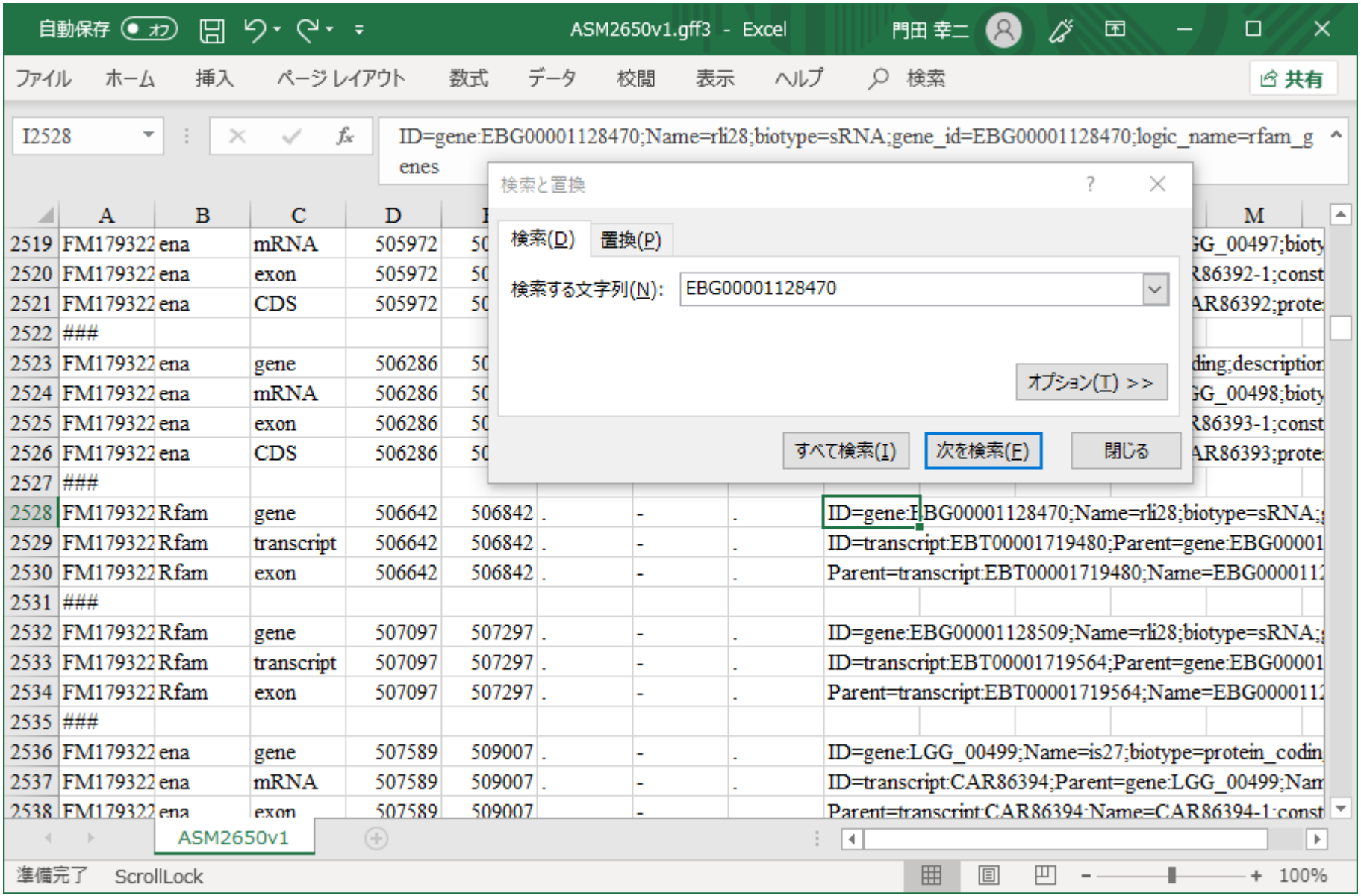
http://bacteria.ensembl.org/Lactobacillus_rhamnosus_gg/Info/Annotation/#assembly

Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

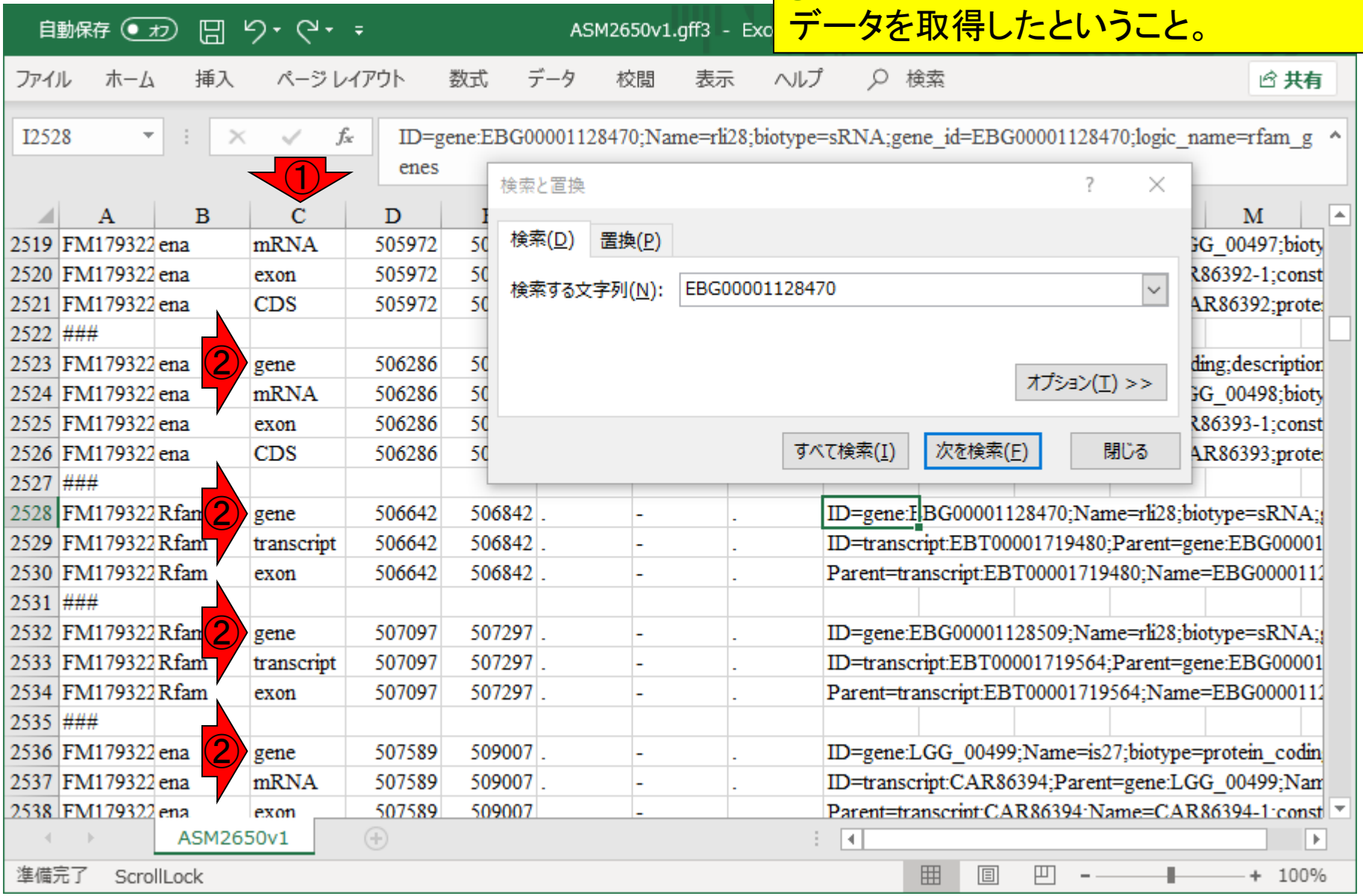
W10-1：行名情報

GFFファイル（ASM2650v1.gff3）をExcelで開いて、EBG00001128470で検索した結果。



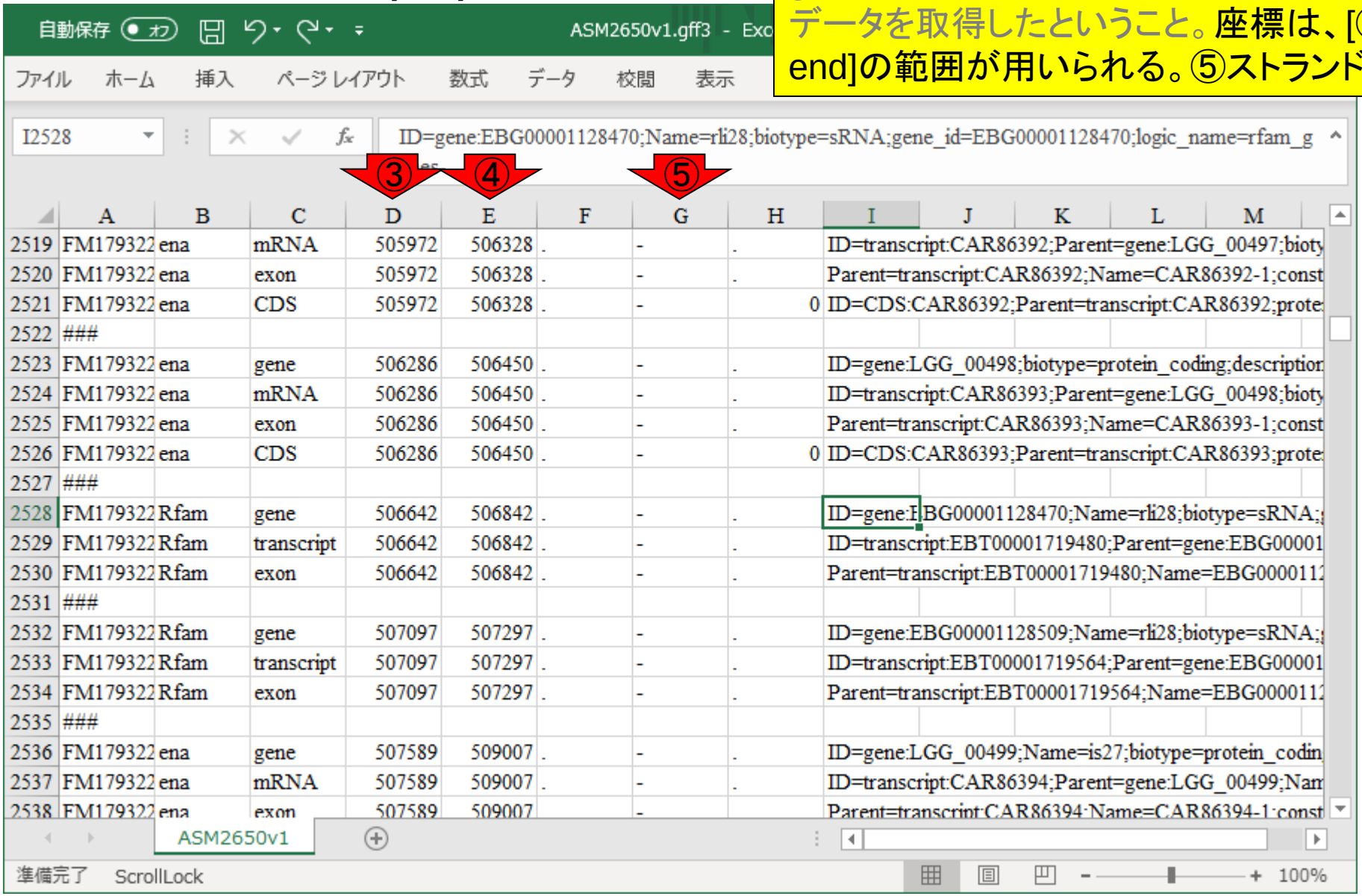
W10-1：行名情報

W8-4で見られるFeature typeをgeneに変更したということは、GFFファイル中の①3列目が②geneとなっている行の座標情報を用いてカウントデータを取得したということ。



W10-2：4,5,7列目

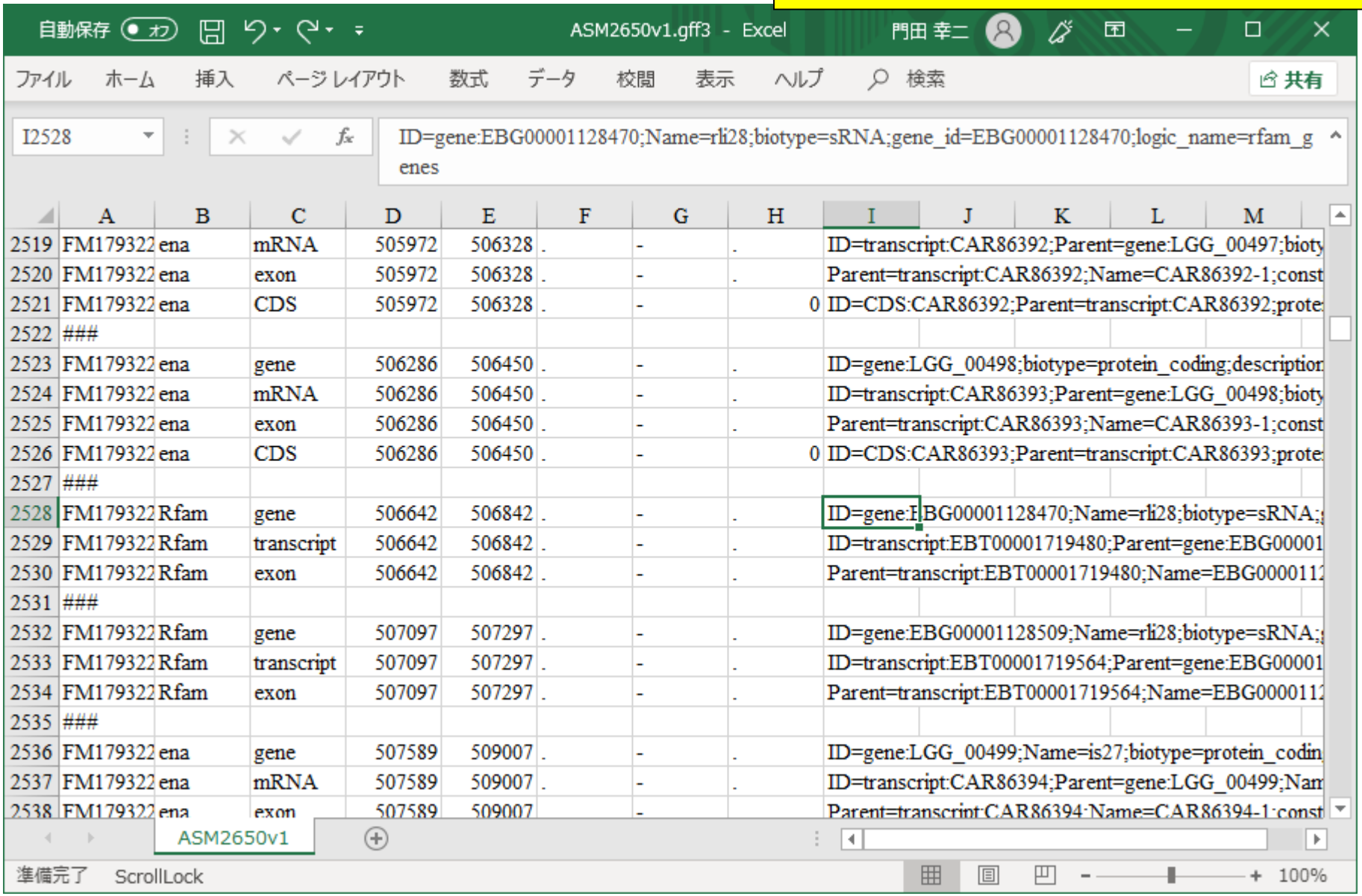
W8-4で見られるFeature typeをgeneに変更したということは、GFFファイル中の①3列目が②geneとなっている行の座標情報を用いてカウントデータを取得したということ。座標は、[③start, ④end]の範囲が用いられる。⑤ストランド情報も。



	A	B	C	D	E	F	G	H	I
2519	FM179322	ena	mRNA	505972	506328	.	-	.	ID=transcript:CAR86392;Parent=gene:LGG_00497;bioty
2520	FM179322	ena	exon	505972	506328	.	-	.	Parent=transcript:CAR86392;Name=CAR86392-1;const
2521	FM179322	ena	CDS	505972	506328	.	-	0	ID=CDS:CAR86392;Parent=transcript:CAR86392;prote
2522	###								
2523	FM179322	ena	gene	506286	506450	.	-	.	ID=gene:LGG_00498;biotype=protein_coding;description
2524	FM179322	ena	mRNA	506286	506450	.	-	.	ID=transcript:CAR86393;Parent=gene:LGG_00498;bioty
2525	FM179322	ena	exon	506286	506450	.	-	.	Parent=transcript:CAR86393;Name=CAR86393-1;const
2526	FM179322	ena	CDS	506286	506450	.	-	0	ID=CDS:CAR86393;Parent=transcript:CAR86393;prote
2527	###								
2528	FM179322	Rfam	gene	506642	506842	.	-	.	ID=gene:EBG00001128470;Name=rli28;biotype=sRNA;
2529	FM179322	Rfam	transcript	506642	506842	.	-	.	ID=transcript:EBT00001719480;Parent=gene:EBG00001
2530	FM179322	Rfam	exon	506642	506842	.	-	.	Parent=transcript:EBT00001719480;Name=EBG0000112
2531	###								
2532	FM179322	Rfam	gene	507097	507297	.	-	.	ID=gene:EBG00001128509;Name=rli28;biotype=sRNA;
2533	FM179322	Rfam	transcript	507097	507297	.	-	.	ID=transcript:EBT00001719564;Parent=gene:EBG00001
2534	FM179322	Rfam	exon	507097	507297	.	-	.	Parent=transcript:EBT00001719564;Name=EBG0000112
2535	###								
2536	FM179322	ena	gene	507589	509007	.	-	.	ID=gene:LGG_00499;Name=is27;biotype=protein_codin
2537	FM179322	ena	mRNA	507589	509007	.	-	.	ID=transcript:CAR86394;Parent=gene:LGG_00499;Nam
2538	FM179322	ena	exon	507589	509007	.	-	.	Parent=transcript:CAR86394;Name=CAR86394-1;const

W10-3：9列目

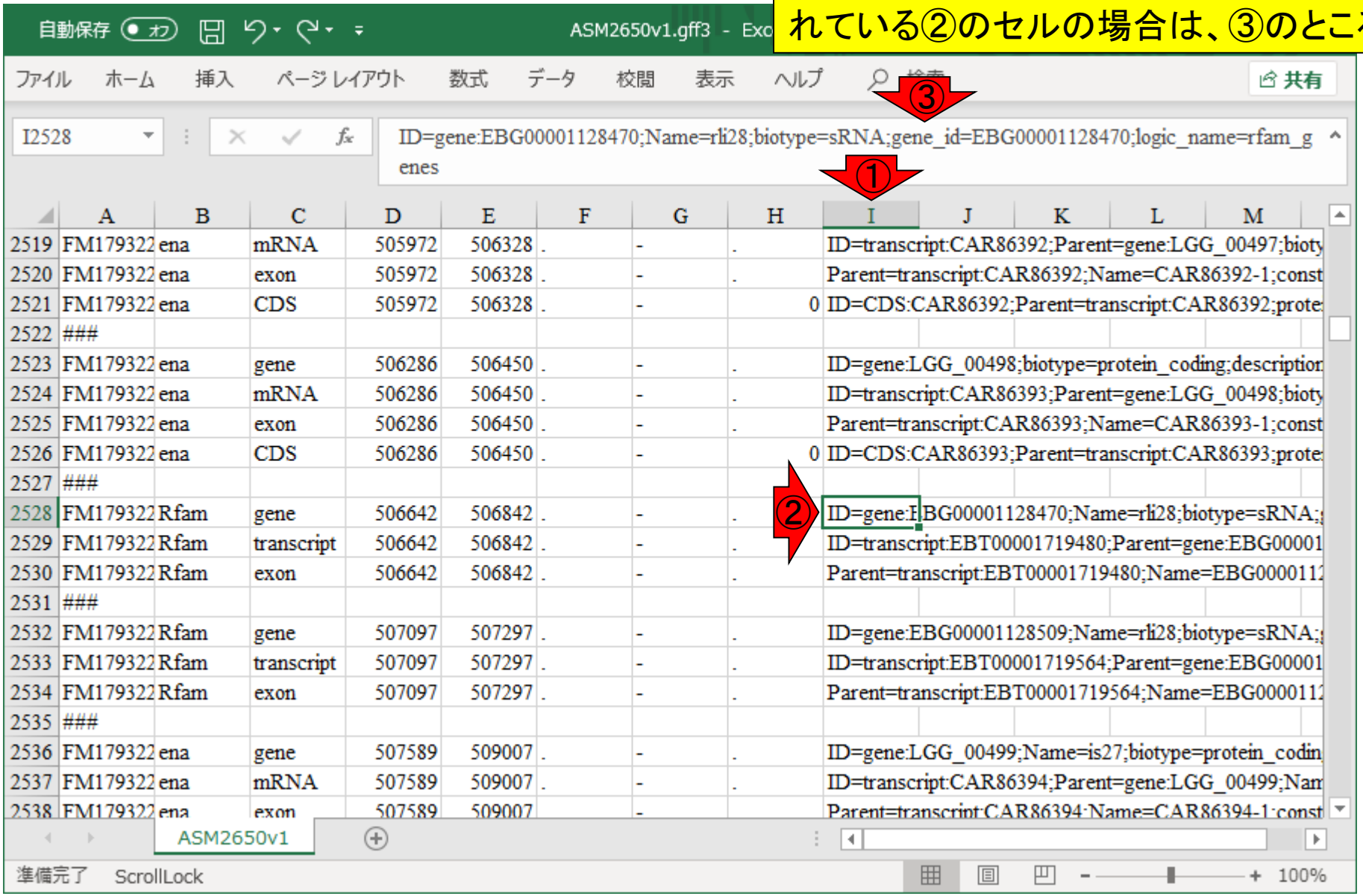
W8-4の画面上でID Attributeのところで指定したgene_idが、カウントデータを取得した際に、用いる行名情報に相当。



	A	B	C	D	E	F	G	H	I	J	K	L	M
2519	FM179322	ena	mRNA	505972	506328	.	-	.	ID=transcript:CAR86392;Parent=gene:LGG_00497;bioty				
2520	FM179322	ena	exon	505972	506328	.	-	.	Parent=transcript:CAR86392;Name=CAR86392-1;const				
2521	FM179322	ena	CDS	505972	506328	.	-	0	ID=CDS:CAR86392;Parent=transcript:CAR86392;prote				
2522	###												
2523	FM179322	ena	gene	506286	506450	.	-	.	ID=gene:LGG_00498;biotype=protein_coding;description				
2524	FM179322	ena	mRNA	506286	506450	.	-	.	ID=transcript:CAR86393;Parent=gene:LGG_00498;bioty				
2525	FM179322	ena	exon	506286	506450	.	-	.	Parent=transcript:CAR86393;Name=CAR86393-1;const				
2526	FM179322	ena	CDS	506286	506450	.	-	0	ID=CDS:CAR86393;Parent=transcript:CAR86393;prote				
2527	###												
2528	FM179322	Rfam	gene	506642	506842	.	-	.	ID=gene:EBG00001128470;Name=rli28;biotype=sRNA;				
2529	FM179322	Rfam	transcript	506642	506842	.	-	.	ID=transcript:EBT00001719480;Parent=gene:EBG00001				
2530	FM179322	Rfam	exon	506642	506842	.	-	.	Parent=transcript:EBT00001719480;Name=EBG0000112				
2531	###												
2532	FM179322	Rfam	gene	507097	507297	.	-	.	ID=gene:EBG00001128509;Name=rli28;biotype=sRNA;				
2533	FM179322	Rfam	transcript	507097	507297	.	-	.	ID=transcript:EBT00001719564;Parent=gene:EBG00001				
2534	FM179322	Rfam	exon	507097	507297	.	-	.	Parent=transcript:EBT00001719564;Name=EBG0000112				
2535	###												
2536	FM179322	ena	gene	507589	509007	.	-	.	ID=gene:LGG_00499;Name=is27;biotype=protein_codin				
2537	FM179322	ena	mRNA	507589	509007	.	-	.	ID=transcript:CAR86394;Parent=gene:LGG_00499;Nam				
2538	FM179322	ena	exon	507589	509007		-		Parent=transcript:CAR86394;Name=CAR86394-1;const				

W10-3：9列目

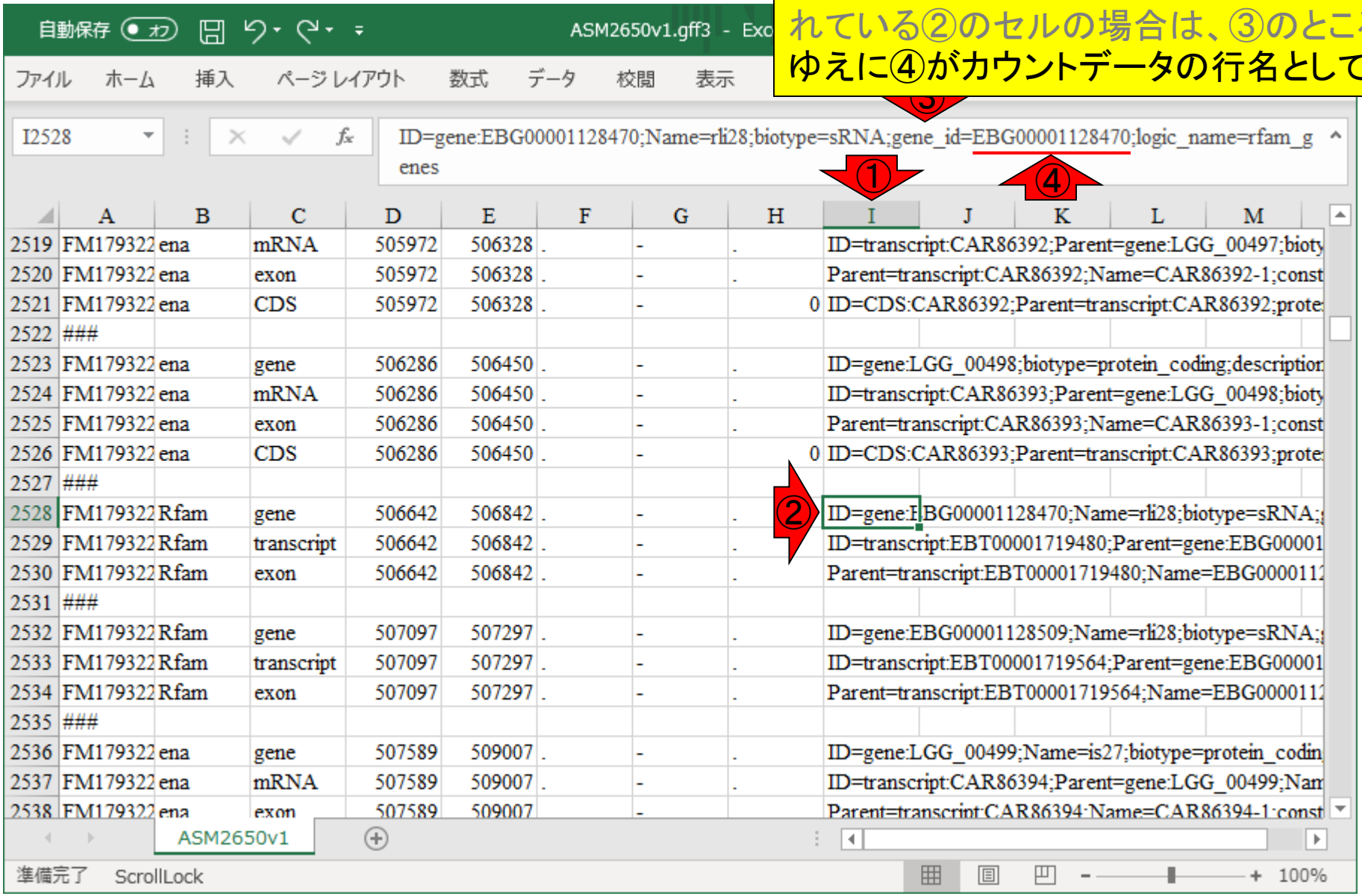
指定したgene_idは、GFFファイル中の①9列目（attribute列と呼ばれる）に存在します。例えば、EBG00001128470の検索結果としてハイライトされている②のセルの場合は、③のところにある。



	A	B	C	D	E	F	G	H	I	J	K	L	M
2519	FM179322	ena	mRNA	505972	506328	.	-	.	ID=transcript:CAR86392;Parent=gene:LGG_00497;bioty				
2520	FM179322	ena	exon	505972	506328	.	-	.	Parent=transcript:CAR86392;Name=CAR86392-1;const				
2521	FM179322	ena	CDS	505972	506328	.	-	0	ID=CDS:CAR86392;Parent=transcript:CAR86392;prote				
2522	###												
2523	FM179322	ena	gene	506286	506450	.	-	.	ID=gene:LGG_00498;biotype=protein_coding;description				
2524	FM179322	ena	mRNA	506286	506450	.	-	.	ID=transcript:CAR86393;Parent=gene:LGG_00498;bioty				
2525	FM179322	ena	exon	506286	506450	.	-	.	Parent=transcript:CAR86393;Name=CAR86393-1;const				
2526	FM179322	ena	CDS	506286	506450	.	-	0	ID=CDS:CAR86393;Parent=transcript:CAR86393;prote				
2527	###												
2528	FM179322	Rfam	gene	506642	506842	.	-	.	ID=gene:EBG00001128470;Name=rli28;biotype=sRNA;				
2529	FM179322	Rfam	transcript	506642	506842	.	-	.	ID=transcript:EBT00001719480;Parent=gene:EBG00001				
2530	FM179322	Rfam	exon	506642	506842	.	-	.	Parent=transcript:EBT00001719480;Name=EBG0000112				
2531	###												
2532	FM179322	Rfam	gene	507097	507297	.	-	.	ID=gene:EBG00001128509;Name=rli28;biotype=sRNA;				
2533	FM179322	Rfam	transcript	507097	507297	.	-	.	ID=transcript:EBT00001719564;Parent=gene:EBG00001				
2534	FM179322	Rfam	exon	507097	507297	.	-	.	Parent=transcript:EBT00001719564;Name=EBG0000112				
2535	###												
2536	FM179322	ena	gene	507589	509007	.	-	.	ID=gene:LGG_00499;Name=is27;biotype=protein_codin				
2537	FM179322	ena	mRNA	507589	509007	.	-	.	ID=transcript:CAR86394;Parent=gene:LGG_00499;Nam				
2538	FM179322	ena	exon	507589	509007		-		Parent=transcript:CAR86394;Name=CAR86394-1;const				

W10-3：9列目

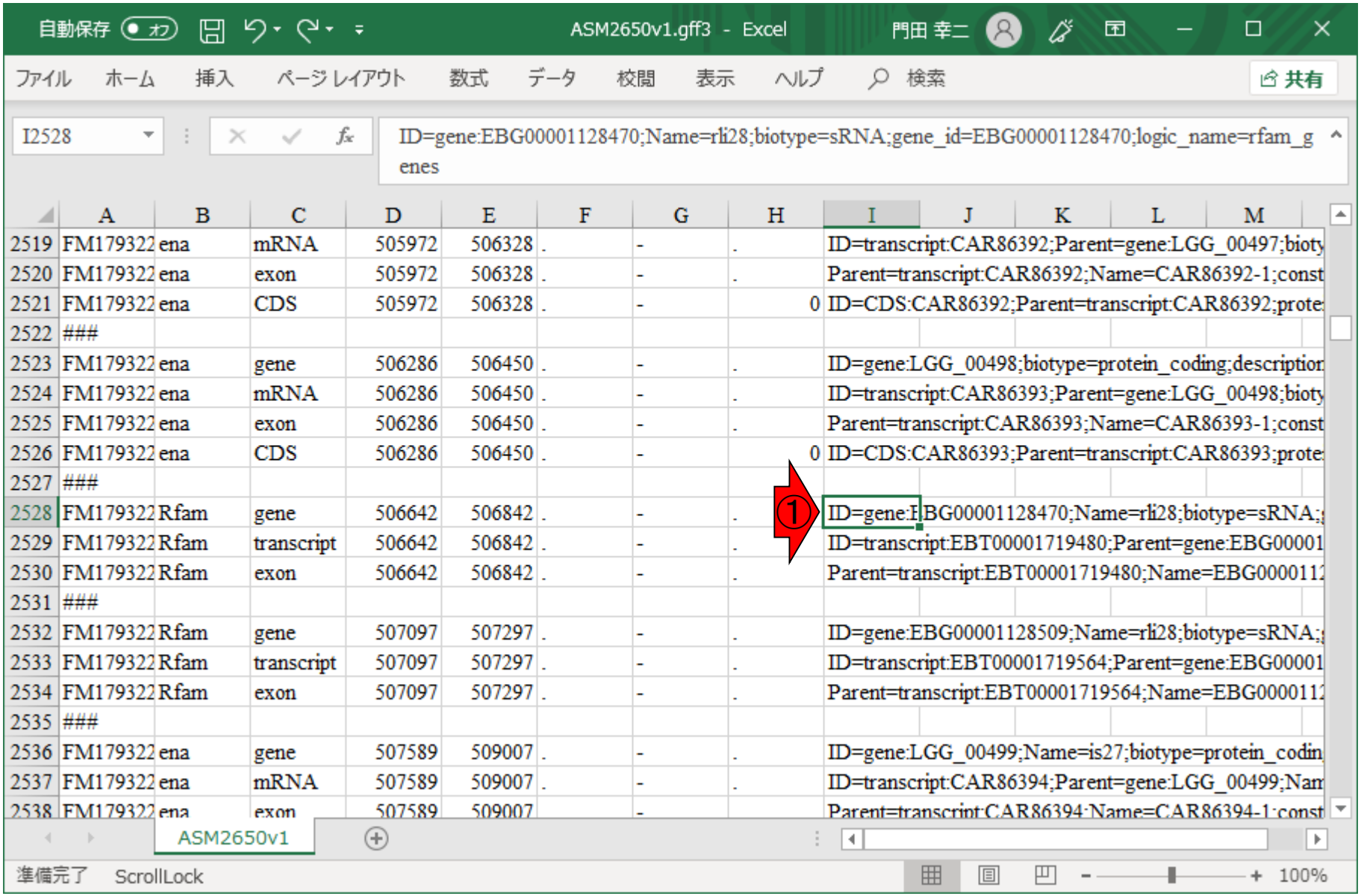
指定したgene_idは、GFFファイル中の①9列目（attribute列と呼ばれる）に存在します。例えば、EBG00001128470の検索結果としてハイライトされている②のセルの場合は、③のところにある。
ゆえに④がカウントデータの行名として使われる



	A	B	C	D	E	F	G	H	I	J	K	L	M
2519	FM179322	ena	mRNA	505972	506328	.	-	.	ID=transcript:CAR86392;Parent=gene:LGG_00497;bioty				
2520	FM179322	ena	exon	505972	506328	.	-	.	Parent=transcript:CAR86392;Name=CAR86392-1;const				
2521	FM179322	ena	CDS	505972	506328	.	-	0	ID=CDS:CAR86392;Parent=transcript:CAR86392;prote				
2522	###												
2523	FM179322	ena	gene	506286	506450	.	-	.	ID=gene:LGG_00498;biotype=protein_coding;description				
2524	FM179322	ena	mRNA	506286	506450	.	-	.	ID=transcript:CAR86393;Parent=gene:LGG_00498;bioty				
2525	FM179322	ena	exon	506286	506450	.	-	.	Parent=transcript:CAR86393;Name=CAR86393-1;const				
2526	FM179322	ena	CDS	506286	506450	.	-	0	ID=CDS:CAR86393;Parent=transcript:CAR86393;prote				
2527	###												
2528	FM179322	Rfam	gene	506642	506842	.	-	.	ID=gene:EBG00001128470;Name=rli28;biotype=sRNA;				
2529	FM179322	Rfam	transcript	506642	506842	.	-	.	ID=transcript:EBT00001719480;Parent=gene:EBG00001				
2530	FM179322	Rfam	exon	506642	506842	.	-	.	Parent=transcript:EBT00001719480;Name=EBG0000112				
2531	###												
2532	FM179322	Rfam	gene	507097	507297	.	-	.	ID=gene:EBG00001128509;Name=rli28;biotype=sRNA;				
2533	FM179322	Rfam	transcript	507097	507297	.	-	.	ID=transcript:EBT00001719564;Parent=gene:EBG00001				
2534	FM179322	Rfam	exon	507097	507297	.	-	.	Parent=transcript:EBT00001719564;Name=EBG0000112				
2535	###												
2536	FM179322	ena	gene	507589	509007	.	-	.	ID=gene:LGG_00499;Name=is27;biotype=protein_codin				
2537	FM179322	ena	mRNA	507589	509007	.	-	.	ID=transcript:CAR86394;Parent=gene:LGG_00499;Nam				
2538	FM179322	ena	exon	507589	509007		-		Parent=transcript:CAR86394;Name=CAR86394-1;const				

この画面上では、①EBG00001128470と…

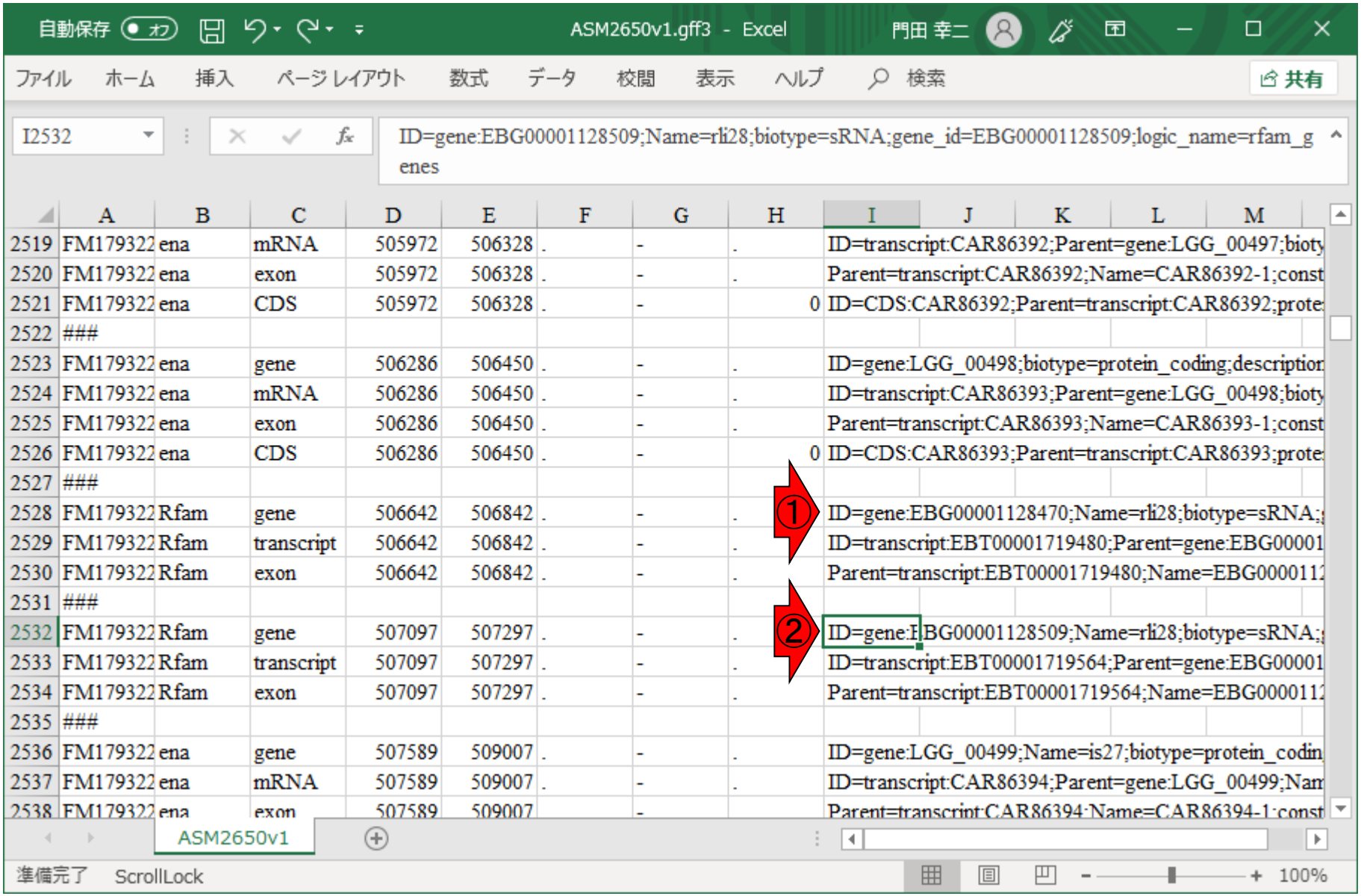
W10-4：情報提供元



	A	B	C	D	E	F	G	H	I	J	K	L	M
2519	FM179322	ena	mRNA	505972	506328	.	-	.	ID=transcript:CAR86392;Parent=gene:LGG_00497;bioty				
2520	FM179322	ena	exon	505972	506328	.	-	.	Parent=transcript:CAR86392;Name=CAR86392-1;const				
2521	FM179322	ena	CDS	505972	506328	.	-	0	ID=CDS:CAR86392;Parent=transcript:CAR86392;prote				
2522	###												
2523	FM179322	ena	gene	506286	506450	.	-	.	ID=gene:LGG_00498;biotype=protein_coding;description				
2524	FM179322	ena	mRNA	506286	506450	.	-	.	ID=transcript:CAR86393;Parent=gene:LGG_00498;bioty				
2525	FM179322	ena	exon	506286	506450	.	-	.	Parent=transcript:CAR86393;Name=CAR86393-1;const				
2526	FM179322	ena	CDS	506286	506450	.	-	0	ID=CDS:CAR86393;Parent=transcript:CAR86393;prote				
2527	###												
2528	FM179322	Rfam	gene	506642	506842	.	-	.	ID=gene:EBG00001128470;Name=rli28;biotype=sRNA;				
2529	FM179322	Rfam	transcript	506642	506842	.	-	.	ID=transcript:EBT00001719480;Parent=gene:EBG00001				
2530	FM179322	Rfam	exon	506642	506842	.	-	.	Parent=transcript:EBT00001719480;Name=EBG0000112				
2531	###												
2532	FM179322	Rfam	gene	507097	507297	.	-	.	ID=gene:EBG00001128509;Name=rli28;biotype=sRNA;				
2533	FM179322	Rfam	transcript	507097	507297	.	-	.	ID=transcript:EBT00001719564;Parent=gene:EBG00001				
2534	FM179322	Rfam	exon	507097	507297	.	-	.	Parent=transcript:EBT00001719564;Name=EBG0000112				
2535	###												
2536	FM179322	ena	gene	507589	509007	.	-	.	ID=gene:LGG_00499;Name=is27;biotype=protein_codin				
2537	FM179322	ena	mRNA	507589	509007	.	-	.	ID=transcript:CAR86394;Parent=gene:LGG_00499;Nam				
2538	FM179322	ena	exon	507589	509007		-		Parent=transcript:CAR86394;Name=CAR86394-1;const				

W10-4：情報提供元

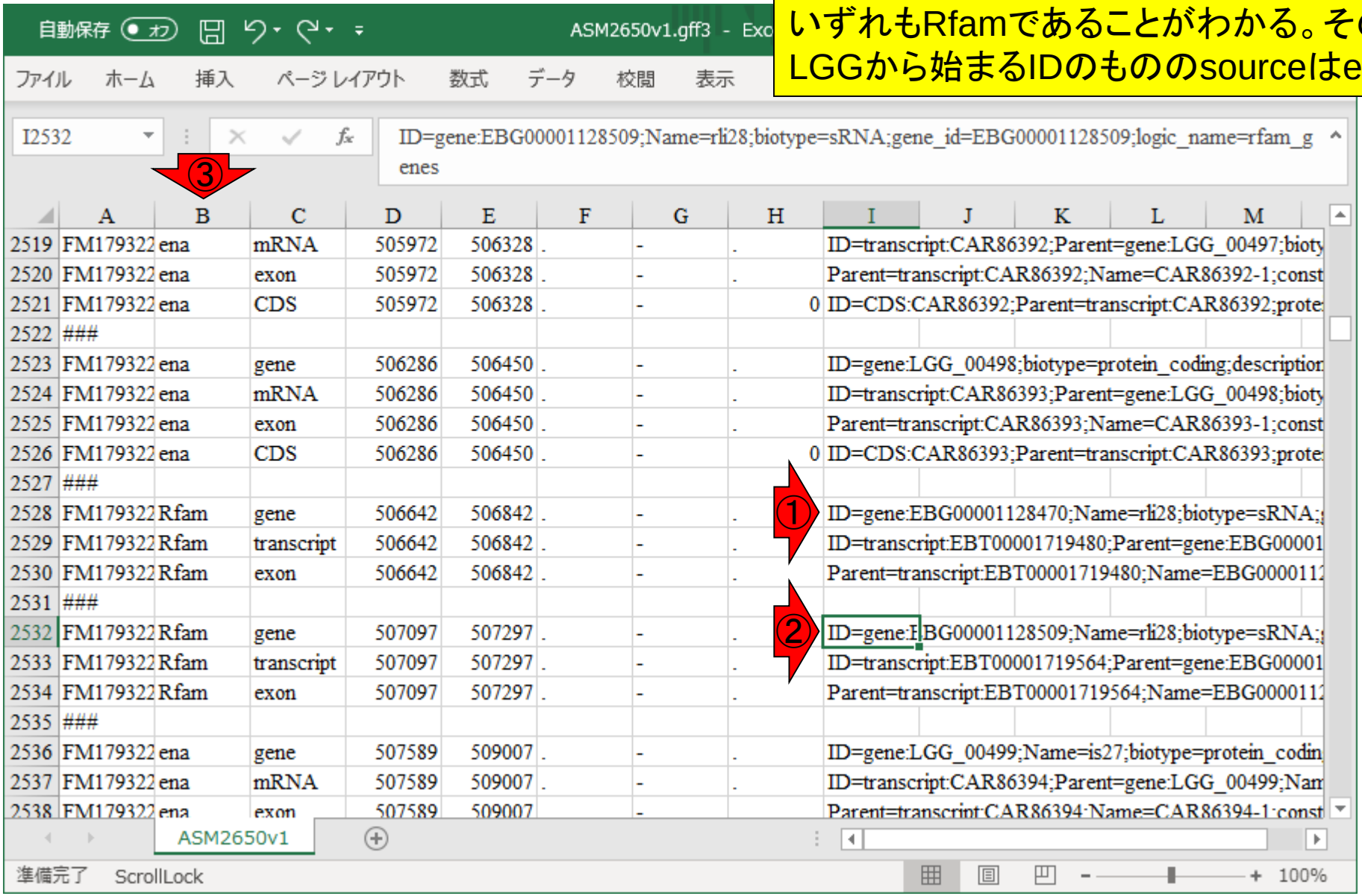
この画面上では、①EBG00001128470と②EBG00001128509が見られる。



	A	B	C	D	E	F	G	H	I	J	K	L	M
2519	FM179322	ena	mRNA	505972	506328	.	-	.	ID=transcript:CAR86392;Parent=gene:LGG_00497;bioty				
2520	FM179322	ena	exon	505972	506328	.	-	.	Parent=transcript:CAR86392;Name=CAR86392-1;const				
2521	FM179322	ena	CDS	505972	506328	.	-	0	ID=CDS:CAR86392;Parent=transcript:CAR86392;prote				
2522	###												
2523	FM179322	ena	gene	506286	506450	.	-	.	ID=gene:LGG_00498;biotype=protein_coding;description				
2524	FM179322	ena	mRNA	506286	506450	.	-	.	ID=transcript:CAR86393;Parent=gene:LGG_00498;bioty				
2525	FM179322	ena	exon	506286	506450	.	-	.	Parent=transcript:CAR86393;Name=CAR86393-1;const				
2526	FM179322	ena	CDS	506286	506450	.	-	0	ID=CDS:CAR86393;Parent=transcript:CAR86393;prote				
2527	###												
2528	FM179322	Rfam	gene	506642	506842	.	-	.	ID=gene:EBG00001128470;Name=rli28;biotype=sRNA;				
2529	FM179322	Rfam	transcript	506642	506842	.	-	.	ID=transcript:EBT00001719480;Parent=gene:EBG00001				
2530	FM179322	Rfam	exon	506642	506842	.	-	.	Parent=transcript:EBT00001719480;Name=EBG0000112				
2531	###												
2532	FM179322	Rfam	gene	507097	507297	.	-	.	ID=gene:EBG00001128509;Name=rli28;biotype=sRNA;				
2533	FM179322	Rfam	transcript	507097	507297	.	-	.	ID=transcript:EBT00001719564;Parent=gene:EBG00001				
2534	FM179322	Rfam	exon	507097	507297	.	-	.	Parent=transcript:EBT00001719564;Name=EBG0000112				
2535	###												
2536	FM179322	ena	gene	507589	509007	.	-	.	ID=gene:LGG_00499;Name=is27;biotype=protein_codin				
2537	FM179322	ena	mRNA	507589	509007	.	-	.	ID=transcript:CAR86394;Parent=gene:LGG_00499;Nam				
2538	FM179322	ena	exon	507589	509007		-		Parent=transcript:CAR86394;Name=CAR86394-1;const				

W10-4：情報提供元

この画面上では、①EBG00001128470と②EBG00001128509が見られる。2列目に記載されている、③これらの情報提供元（source）は、いずれもRfamであることがわかる。その一方で、LGGから始まるIDのもののsourceはena。



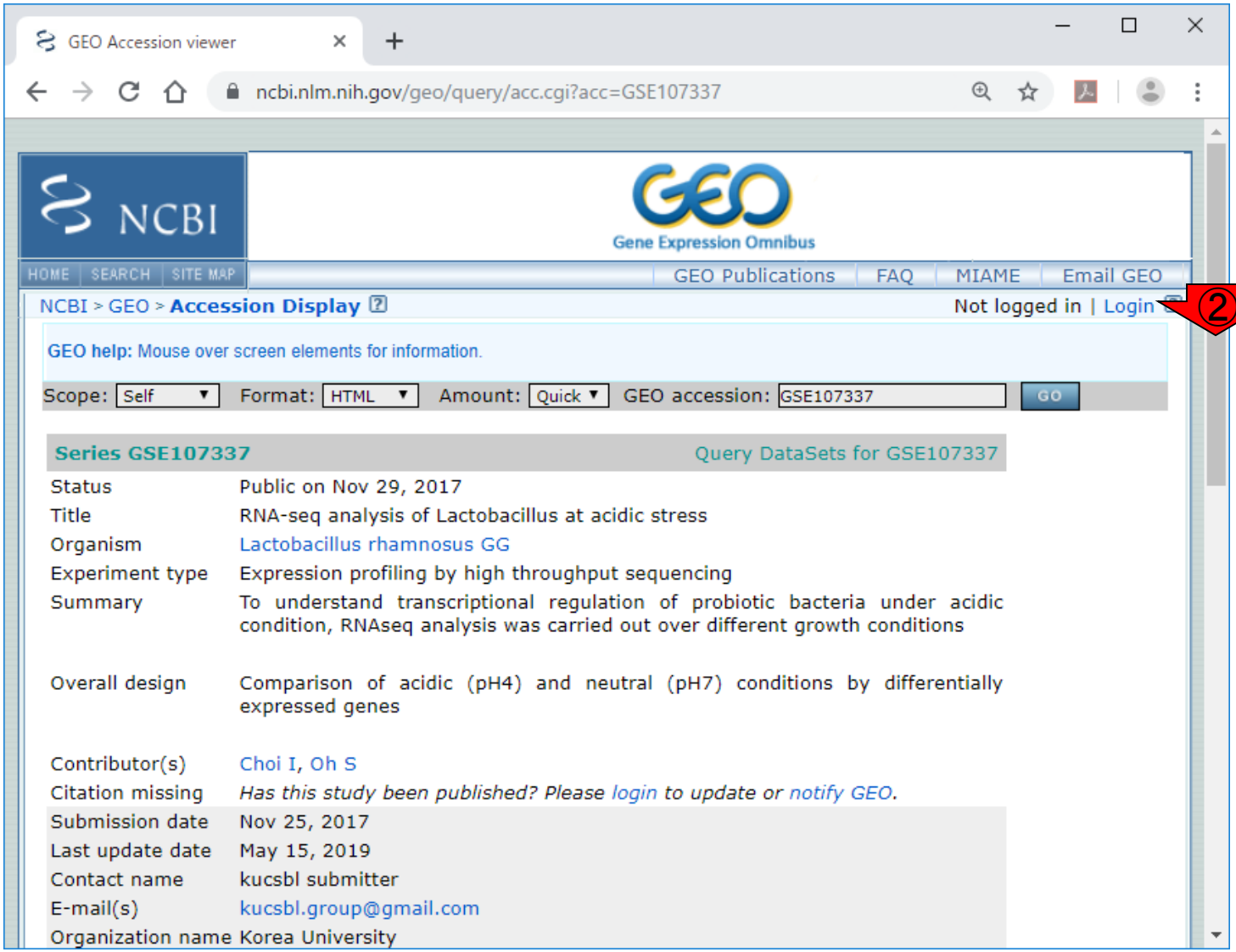
	A	B	C	D	E	F	G	H	I	J	K	L	M
2519	FM179322	ena	mRNA	505972	506328	.	-	.	ID=transcript:CAR86392;Parent=gene:LGG_00497;bioty				
2520	FM179322	ena	exon	505972	506328	.	-	.	Parent=transcript:CAR86392;Name=CAR86392-1;const				
2521	FM179322	ena	CDS	505972	506328	.	-	0	ID=CDS:CAR86392;Parent=transcript:CAR86392;prote				
2522	###												
2523	FM179322	ena	gene	506286	506450	.	-	.	ID=gene:LGG_00498;biotype=protein_coding;description				
2524	FM179322	ena	mRNA	506286	506450	.	-	.	ID=transcript:CAR86393;Parent=gene:LGG_00498;bioty				
2525	FM179322	ena	exon	506286	506450	.	-	.	Parent=transcript:CAR86393;Name=CAR86393-1;const				
2526	FM179322	ena	CDS	506286	506450	.	-	0	ID=CDS:CAR86393;Parent=transcript:CAR86393;prote				
2527	###												
2528	FM179322	Rfam	gene	506642	506842	.	-	.	ID=gene:EBG00001128470;Name=rli28;biotype=sRNA;				
2529	FM179322	Rfam	transcript	506642	506842	.	-	.	ID=transcript:EBT00001719480;Parent=gene:EBG00001				
2530	FM179322	Rfam	exon	506642	506842	.	-	.	Parent=transcript:EBT00001719480;Name=EBG0000112				
2531	###												
2532	FM179322	Rfam	gene	507097	507297	.	-	.	ID=gene:EBG00001128509;Name=rli28;biotype=sRNA;				
2533	FM179322	Rfam	transcript	507097	507297	.	-	.	ID=transcript:EBT00001719564;Parent=gene:EBG00001				
2534	FM179322	Rfam	exon	507097	507297	.	-	.	Parent=transcript:EBT00001719564;Name=EBG0000112				
2535	###												
2536	FM179322	ena	gene	507589	509007	.	-	.	ID=gene:LGG_00499;Name=is27;biotype=protein_codin				
2537	FM179322	ena	mRNA	507589	509007	.	-	.	ID=transcript:CAR86394;Parent=gene:LGG_00499;Nam				
2538	FM179322	ena	exon	507589	509007		-		Parent=transcript:CAR86394;Name=CAR86394-1;const				

Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

W11-1：GSE107337

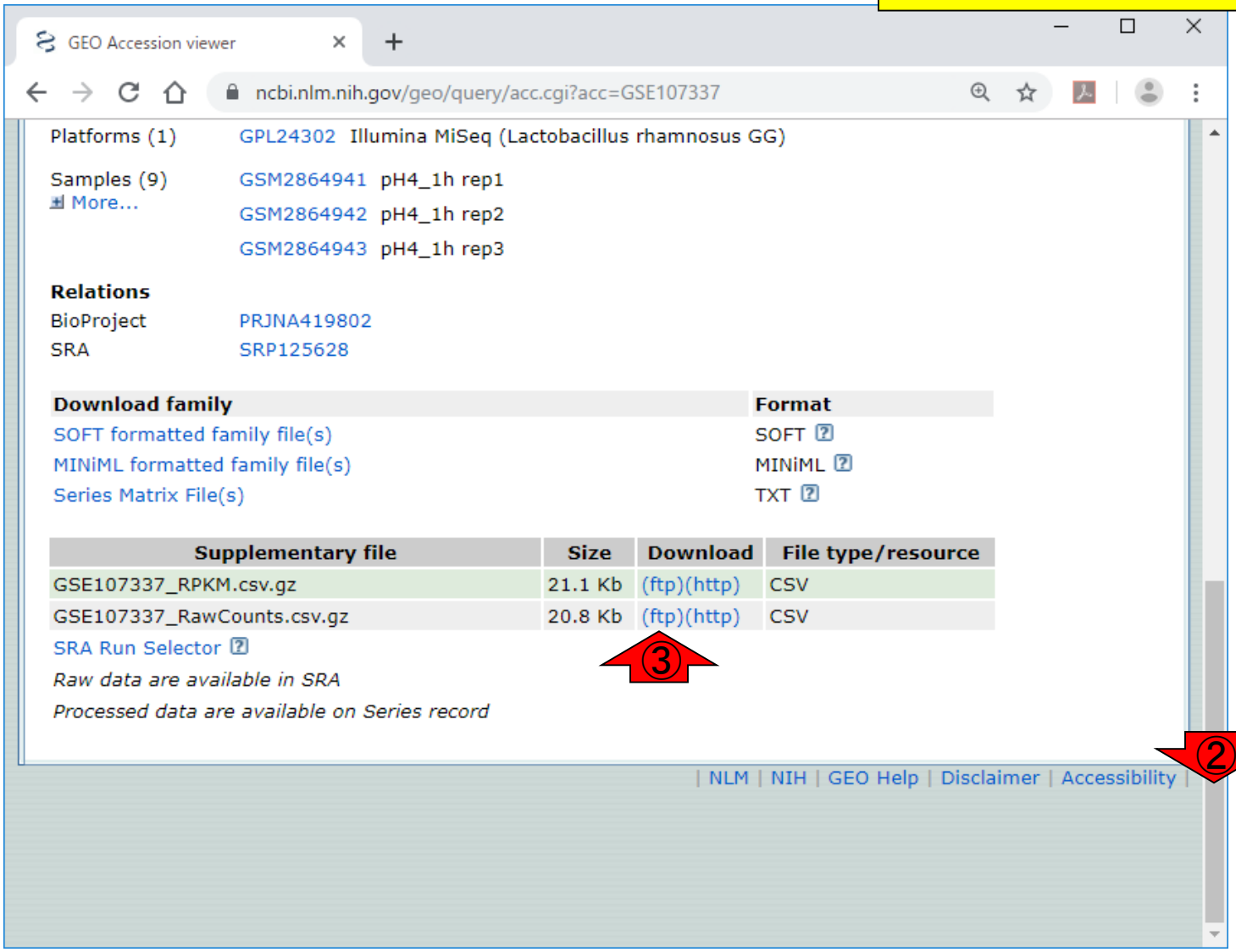
原著論文著者らがGEO上で公開している、①GSE107337のページ。②下部に移動。



https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107337

W11-1：GSE107337

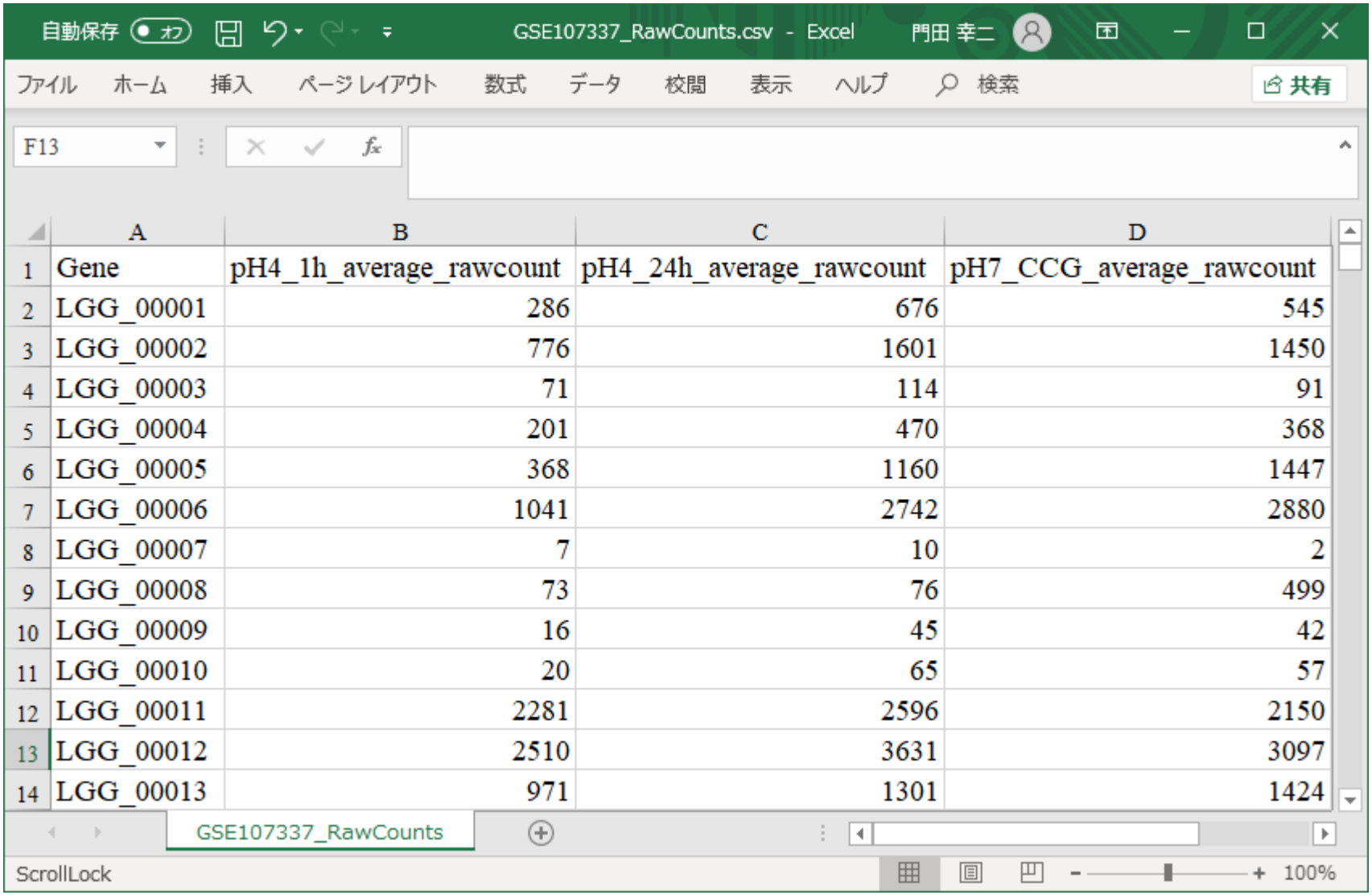
原著論文著者らがGEO上で公開している、①GSE107337のページ。②下部に移動。③をクリックして得られたのが、次のスライドのファイル。



https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107337

W11-1：GSE107337

原著論文著者らがGEO上で公開している、①GSE107337のraw countデータ。



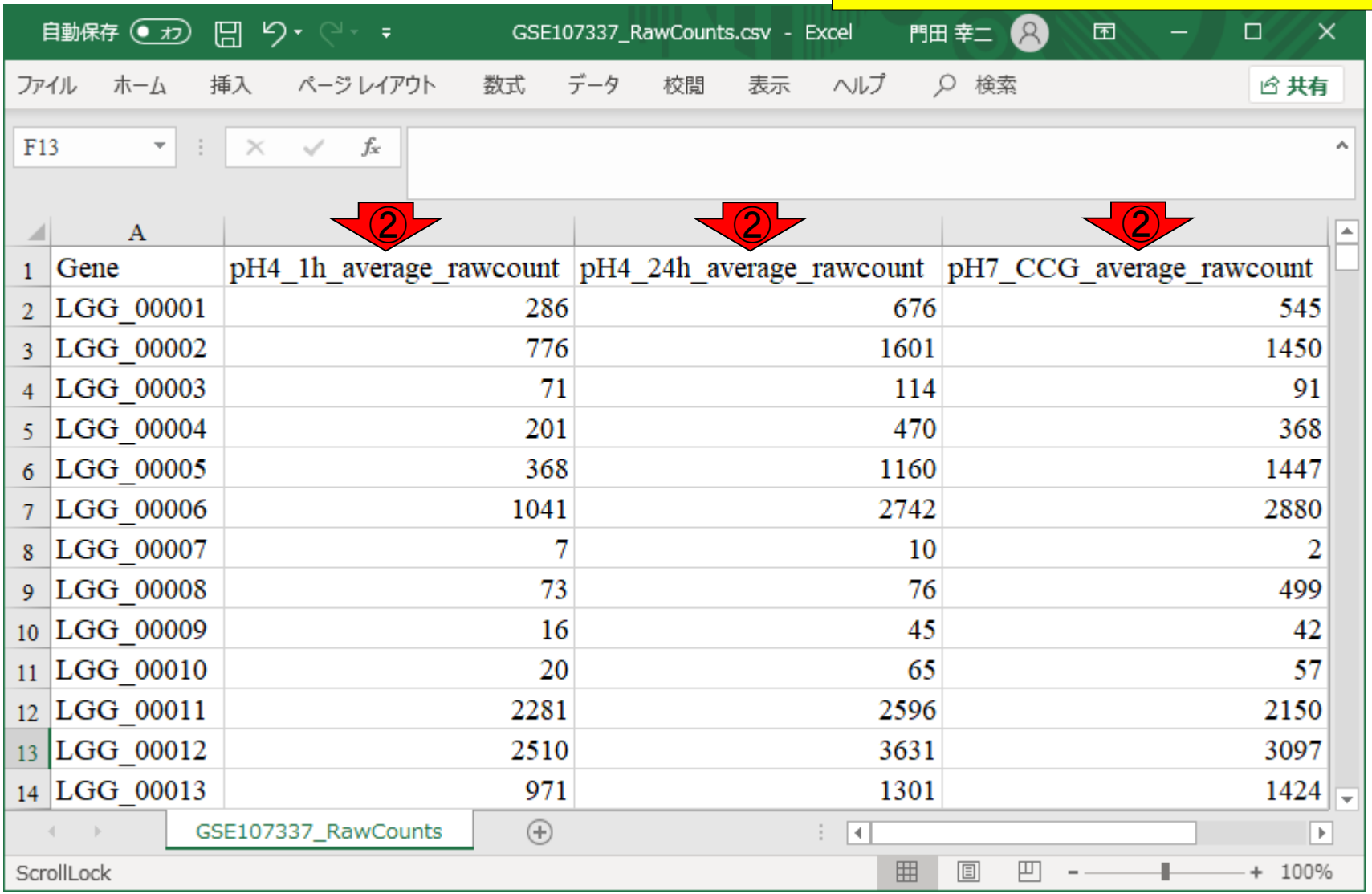
	A	B	C	D
1	Gene	pH4_1h_average_rawcount	pH4_24h_average_rawcount	pH7_CCG_average_rawcount
2	LGG_00001	286	676	545
3	LGG_00002	776	1601	1450
4	LGG_00003	71	114	91
5	LGG_00004	201	470	368
6	LGG_00005	368	1160	1447
7	LGG_00006	1041	2742	2880
8	LGG_00007	7	10	2
9	LGG_00008	73	76	499
10	LGG_00009	16	45	42
11	LGG_00010	20	65	57
12	LGG_00011	2281	2596	2150
13	LGG_00012	2510	3631	3097
14	LGG_00013	971	1301	1424

①

ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE107nnn/GSE107337/suppl/GSE107337_RawCounts.csv.gz

W11-2：平均値

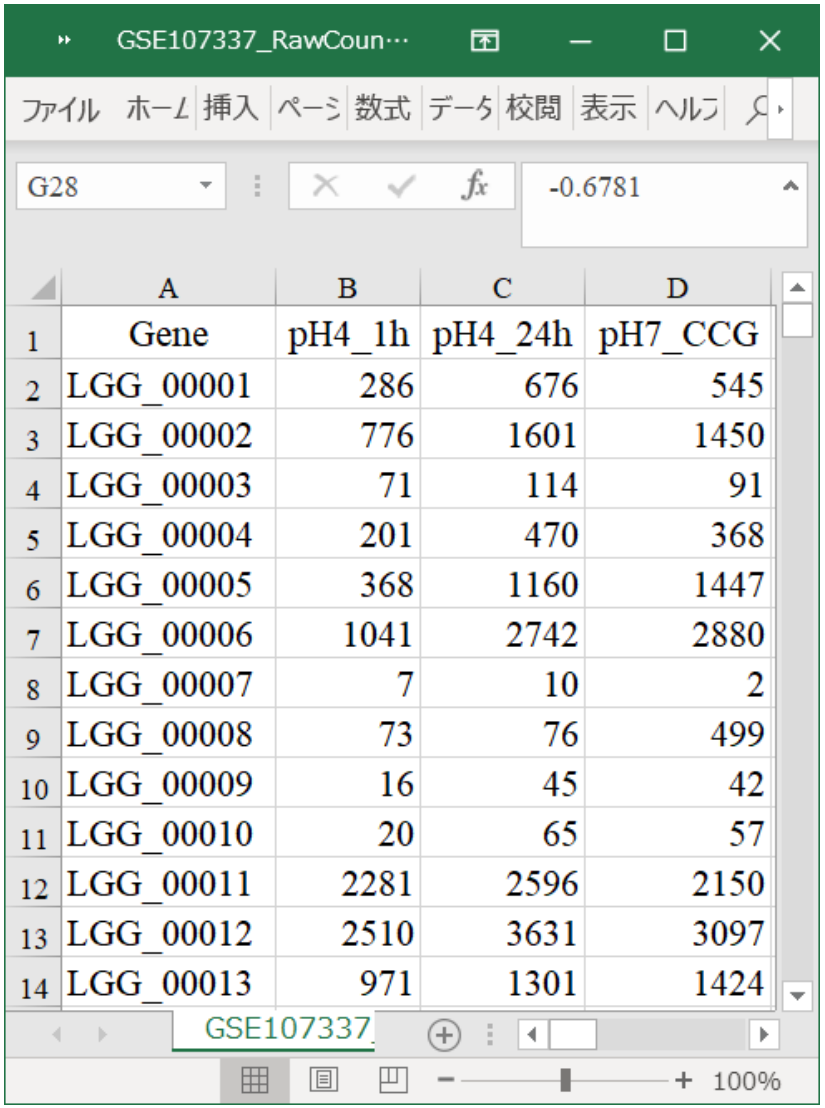
原著論文著者らがGEO上で公開している、①GSE107337のraw countデータ。②各群について3反復の平均値だということがわかる。



	A			
1	Gene	pH4_1h_average_rawcount	pH4_24h_average_rawcount	pH7_CCG_average_rawcount
2	LGG_00001	286	676	545
3	LGG_00002	776	1601	1450
4	LGG_00003	71	114	91
5	LGG_00004	201	470	368
6	LGG_00005	368	1160	1447
7	LGG_00006	1041	2742	2880
8	LGG_00007	7	10	2
9	LGG_00008	73	76	499
10	LGG_00009	16	45	42
11	LGG_00010	20	65	57
12	LGG_00011	2281	2596	2150
13	LGG_00012	2510	3631	3097
14	LGG_00013	971	1301	1424

W11-2：列名変更

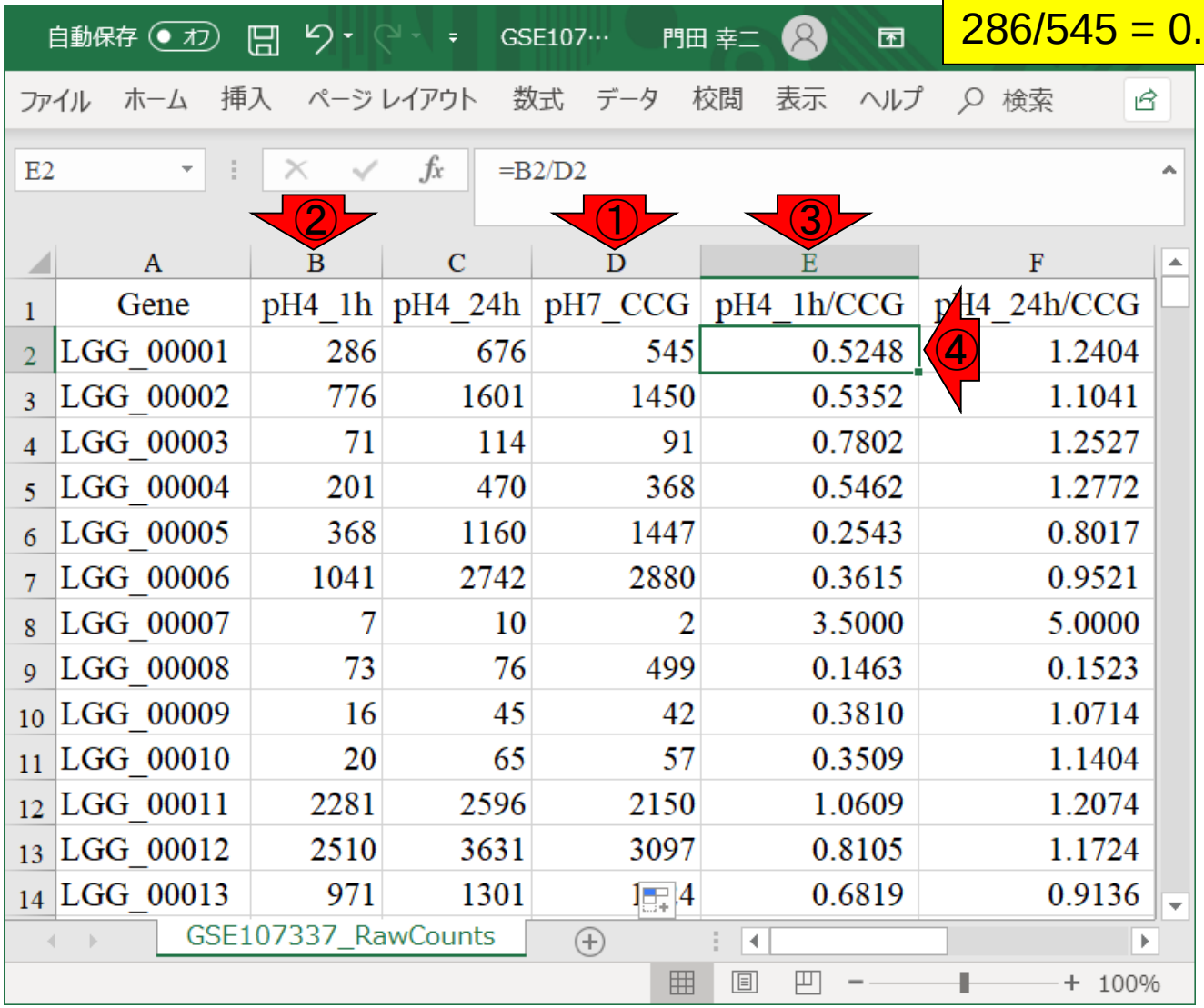
原著論文著者らがGEO上で公開している、①GSE107337のraw countデータ。②各群について3反復の平均値だということがわかる。平均値だということはわかったので、列名を短くしておく。



	A	B	C	D
1	Gene	pH4_1h	pH4_24h	pH7_CCG
2	LGG_00001	286	676	545
3	LGG_00002	776	1601	1450
4	LGG_00003	71	114	91
5	LGG_00004	201	470	368
6	LGG_00005	368	1160	1447
7	LGG_00006	1041	2742	2880
8	LGG_00007	7	10	2
9	LGG_00008	73	76	499
10	LGG_00009	16	45	42
11	LGG_00010	20	65	57
12	LGG_00011	2281	2596	2150
13	LGG_00012	2510	3631	3097
14	LGG_00013	971	1301	1424

W11-3：倍率変化

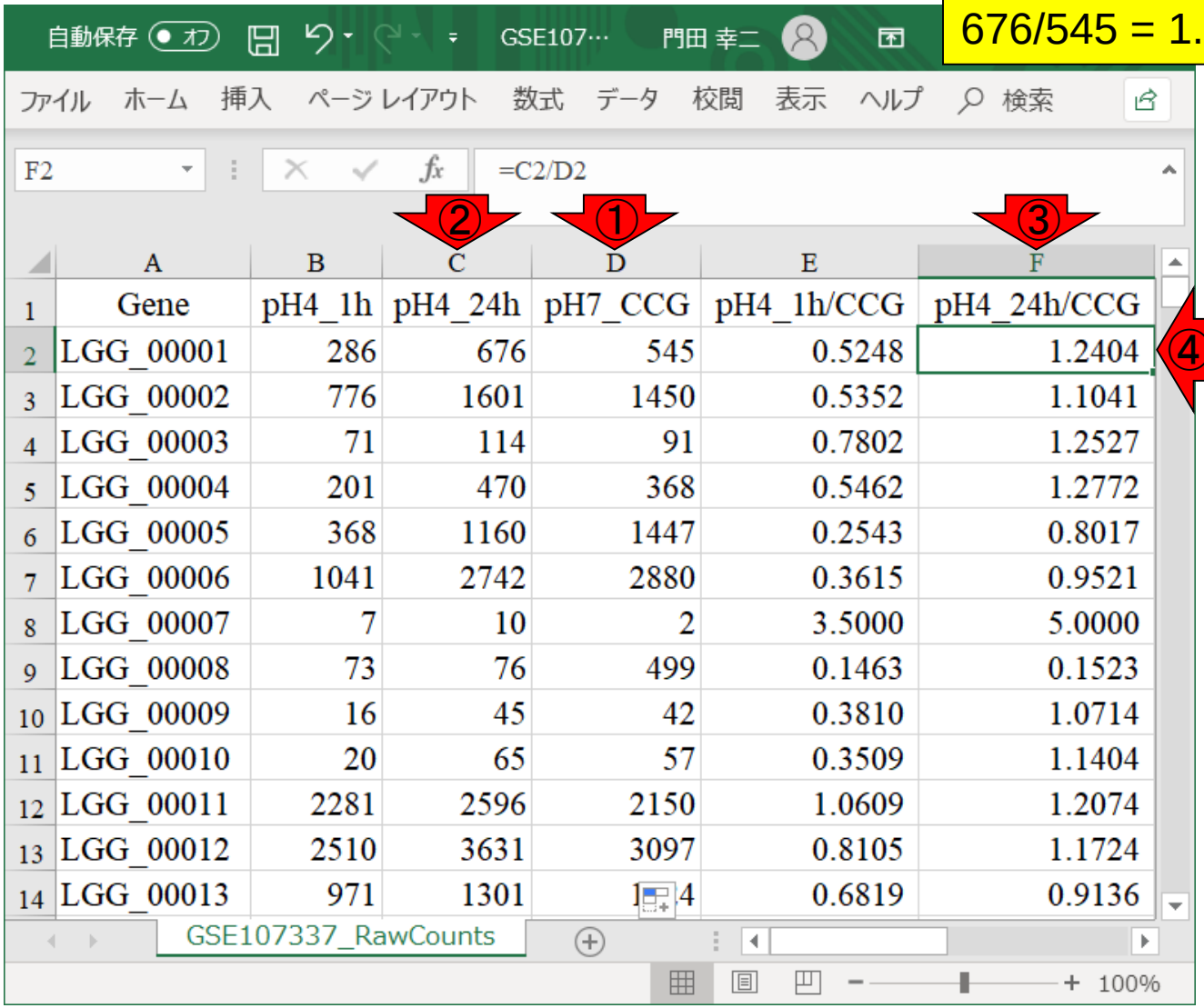
①コントロール（中性状態）に対する、②短期（1hなのでshort-term）の酸性ストレスの倍率変化を計算したのが③。④LGG_00001の場合は、286/545 = 0.5248。



	A	B	C	D	E	F
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG
2	LGG_00001	286	676	545	0.5248	1.2404
3	LGG_00002	776	1601	1450	0.5352	1.1041
4	LGG_00003	71	114	91	0.7802	1.2527
5	LGG_00004	201	470	368	0.5462	1.2772
6	LGG_00005	368	1160	1447	0.2543	0.8017
7	LGG_00006	1041	2742	2880	0.3615	0.9521
8	LGG_00007	7	10	2	3.5000	5.0000
9	LGG_00008	73	76	499	0.1463	0.1523
10	LGG_00009	16	45	42	0.3810	1.0714
11	LGG_00010	20	65	57	0.3509	1.1404
12	LGG_00011	2281	2596	2150	1.0609	1.2074
13	LGG_00012	2510	3631	3097	0.8105	1.1724
14	LGG_00013	971	1301	1424	0.6819	0.9136

W11-3：倍率変化

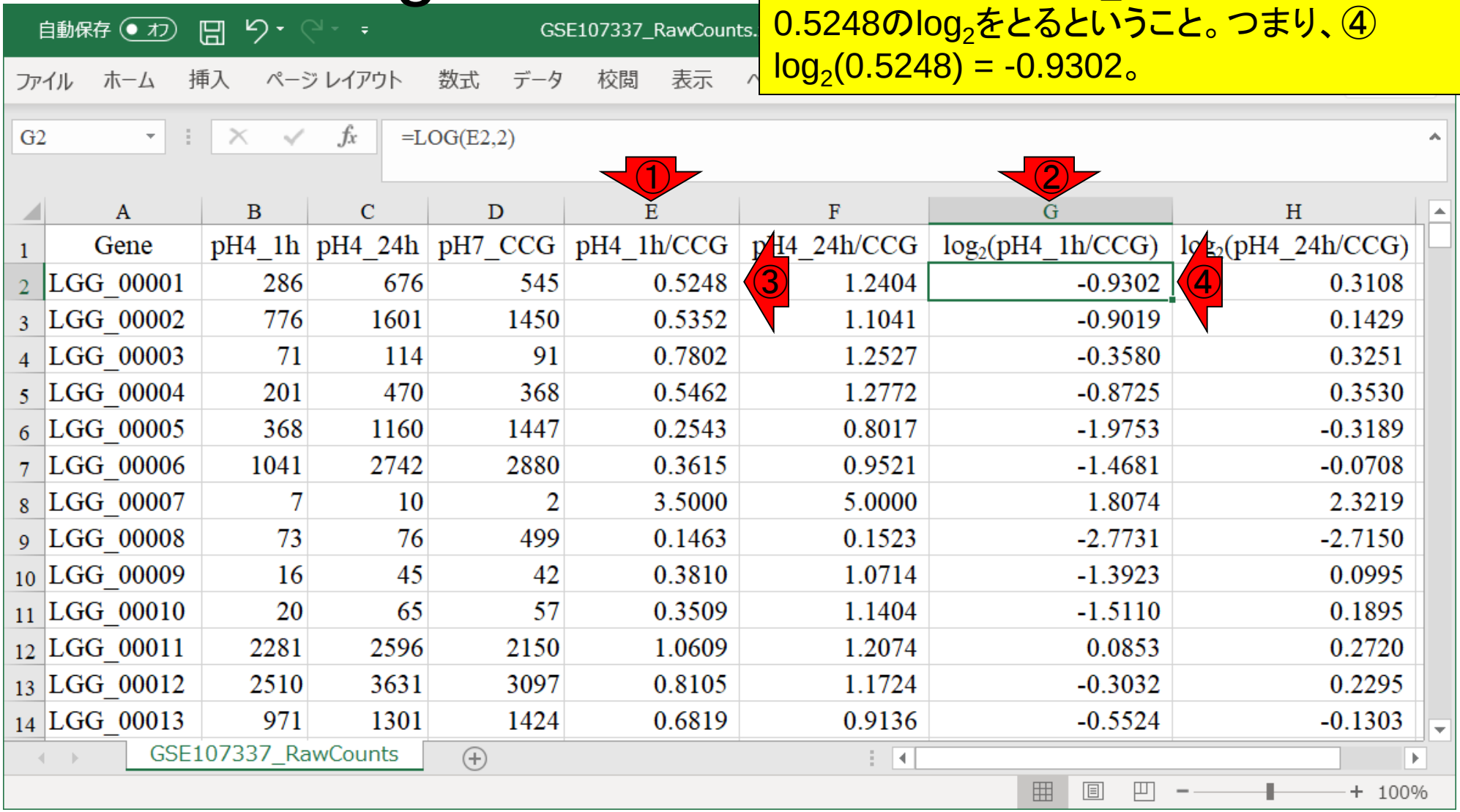
①コントロール（中性状態）に対する、②長期（24hなのでlong-term）の酸性ストレスの倍率変化を計算したのが③。④LGG_00001の場合は676/545 = 1.2404。



	A	B	C	D	E	F
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG
2	LGG_00001	286	676	545	0.5248	1.2404
3	LGG_00002	776	1601	1450	0.5352	1.1041
4	LGG_00003	71	114	91	0.7802	1.2527
5	LGG_00004	201	470	368	0.5462	1.2772
6	LGG_00005	368	1160	1447	0.2543	0.8017
7	LGG_00006	1041	2742	2880	0.3615	0.9521
8	LGG_00007	7	10	2	3.5000	5.0000
9	LGG_00008	73	76	499	0.1463	0.1523
10	LGG_00009	16	45	42	0.3810	1.0714
11	LGG_00010	20	65	57	0.3509	1.1404
12	LGG_00011	2281	2596	2150	1.0609	1.2074
13	LGG_00012	2510	3631	3097	0.8105	1.1724
14	LGG_00013	971	1301	1[illegible]4	0.6819	0.9136

W11-4：log比

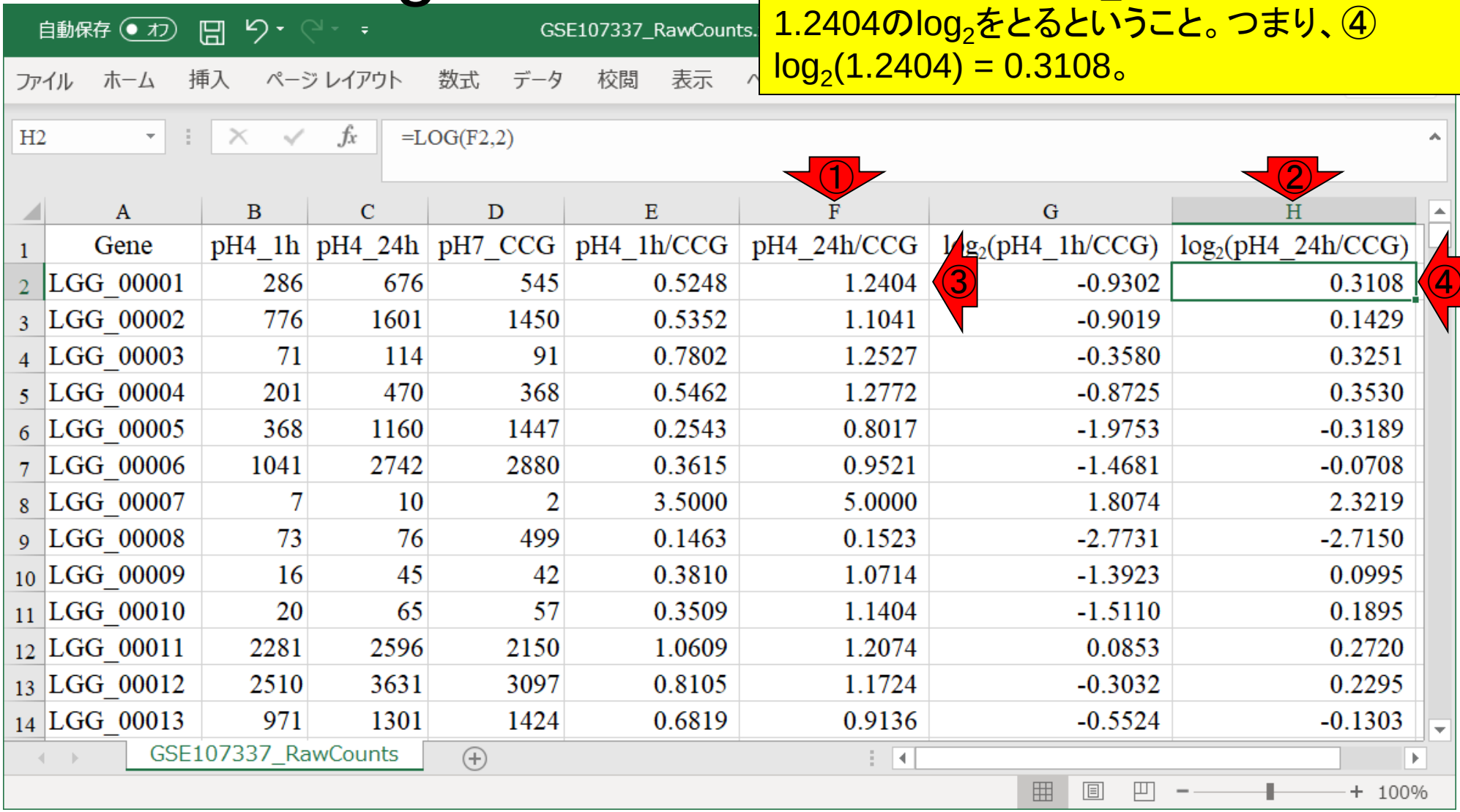
「①コントロール（中性状態）に対する短期（1hなのでshort-term）の酸性ストレスの倍率変化」の、②対数（底は2）。LGG_00001の場合は、③0.5248の$\log_2$をとるということ。つまり、④$\log_2(0.5248) = -0.9302$。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
2	LGG_00001	286	676	545	0.5248	1.2404	-0.9302	0.3108
3	LGG_00002	776	1601	1450	0.5352	1.1041	-0.9019	0.1429
4	LGG_00003	71	114	91	0.7802	1.2527	-0.3580	0.3251
5	LGG_00004	201	470	368	0.5462	1.2772	-0.8725	0.3530
6	LGG_00005	368	1160	1447	0.2543	0.8017	-1.9753	-0.3189
7	LGG_00006	1041	2742	2880	0.3615	0.9521	-1.4681	-0.0708
8	LGG_00007	7	10	2	3.5000	5.0000	1.8074	2.3219
9	LGG_00008	73	76	499	0.1463	0.1523	-2.7731	-2.7150
10	LGG_00009	16	45	42	0.3810	1.0714	-1.3923	0.0995
11	LGG_00010	20	65	57	0.3509	1.1404	-1.5110	0.1895
12	LGG_00011	2281	2596	2150	1.0609	1.2074	0.0853	0.2720
13	LGG_00012	2510	3631	3097	0.8105	1.1724	-0.3032	0.2295
14	LGG_00013	971	1301	1424	0.6819	0.9136	-0.5524	-0.1303

W11-4：log比

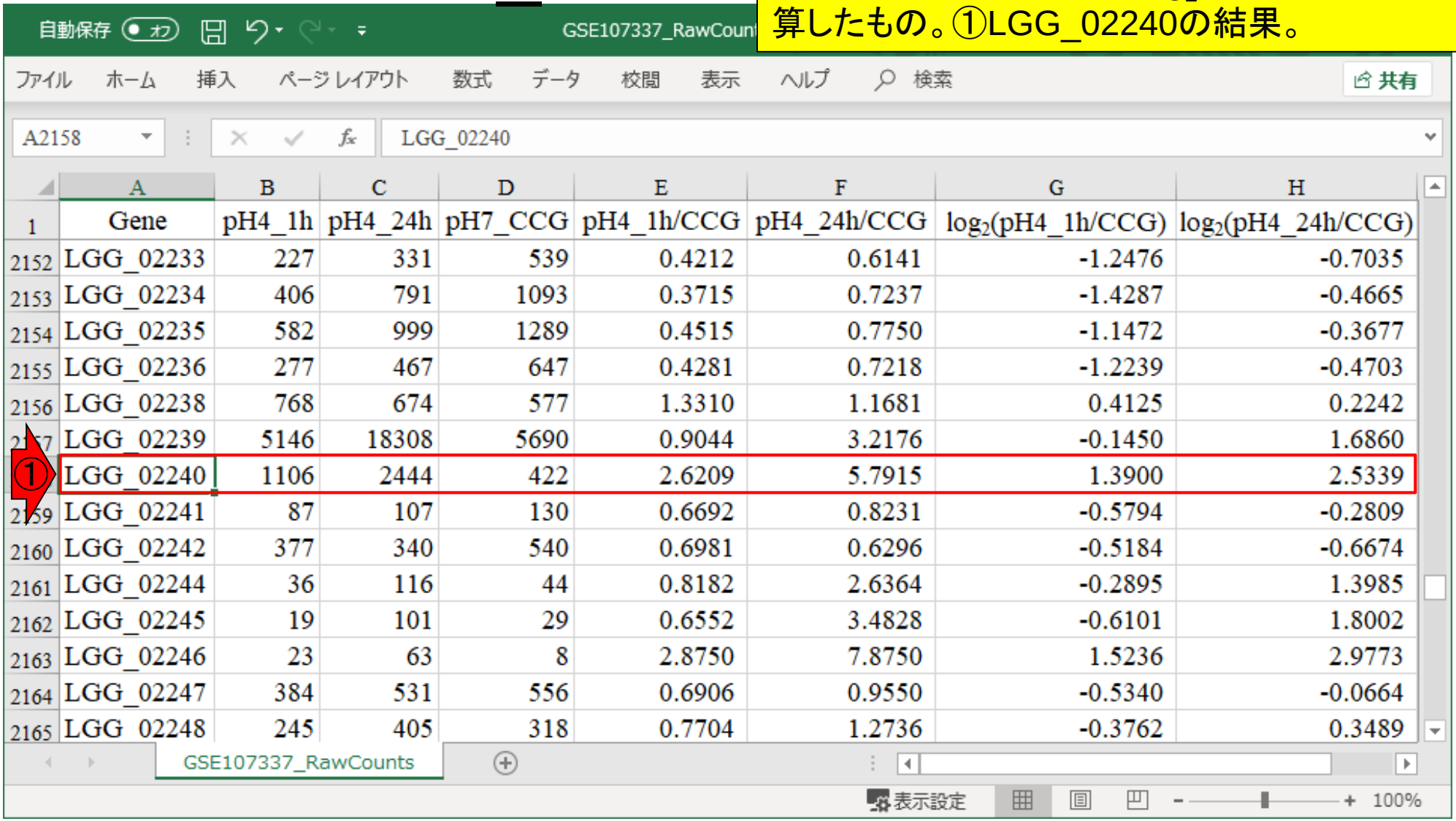
「①コントロール（中性状態）に対する長期（24hなのでlong-term）の酸性ストレスの倍率変化」の、②対数（底は2）。LGG_00001の場合は、③1.2404の$\log_2$をとるということ。つまり、④$\log_2(1.2404) = 0.3108$。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
2	LGG_00001	286	676	545	0.5248	1.2404	-0.9302	0.3108
3	LGG_00002	776	1601	1450	0.5352	1.1041	-0.9019	0.1429
4	LGG_00003	71	114	91	0.7802	1.2527	-0.3580	0.3251
5	LGG_00004	201	470	368	0.5462	1.2772	-0.8725	0.3530
6	LGG_00005	368	1160	1447	0.2543	0.8017	-1.9753	-0.3189
7	LGG_00006	1041	2742	2880	0.3615	0.9521	-1.4681	-0.0708
8	LGG_00007	7	10	2	3.5000	5.0000	1.8074	2.3219
9	LGG_00008	73	76	499	0.1463	0.1523	-2.7731	-2.7150
10	LGG_00009	16	45	42	0.3810	1.0714	-1.3923	0.0995
11	LGG_00010	20	65	57	0.3509	1.1404	-1.5110	0.1895
12	LGG_00011	2281	2596	2150	1.0609	1.2074	0.0853	0.2720
13	LGG_00012	2510	3631	3097	0.8105	1.1724	-0.3032	0.2295
14	LGG_00013	971	1301	1424	0.6819	0.9136	-0.5524	-0.1303

W11-5：LGG_02240

おさらい。今眺めているものは、著者らがGEO上で公開しているGSE107337のraw countデータから、我々が倍率変化や$\log_2$（倍率変化）を再計算したもの。①LGG_02240の結果。



自動保存 オフ　GSE107337_RawCount

ファイル　ホーム　挿入　ページ レイアウト　数式　データ　校閲　表示　ヘルプ　検索　共有

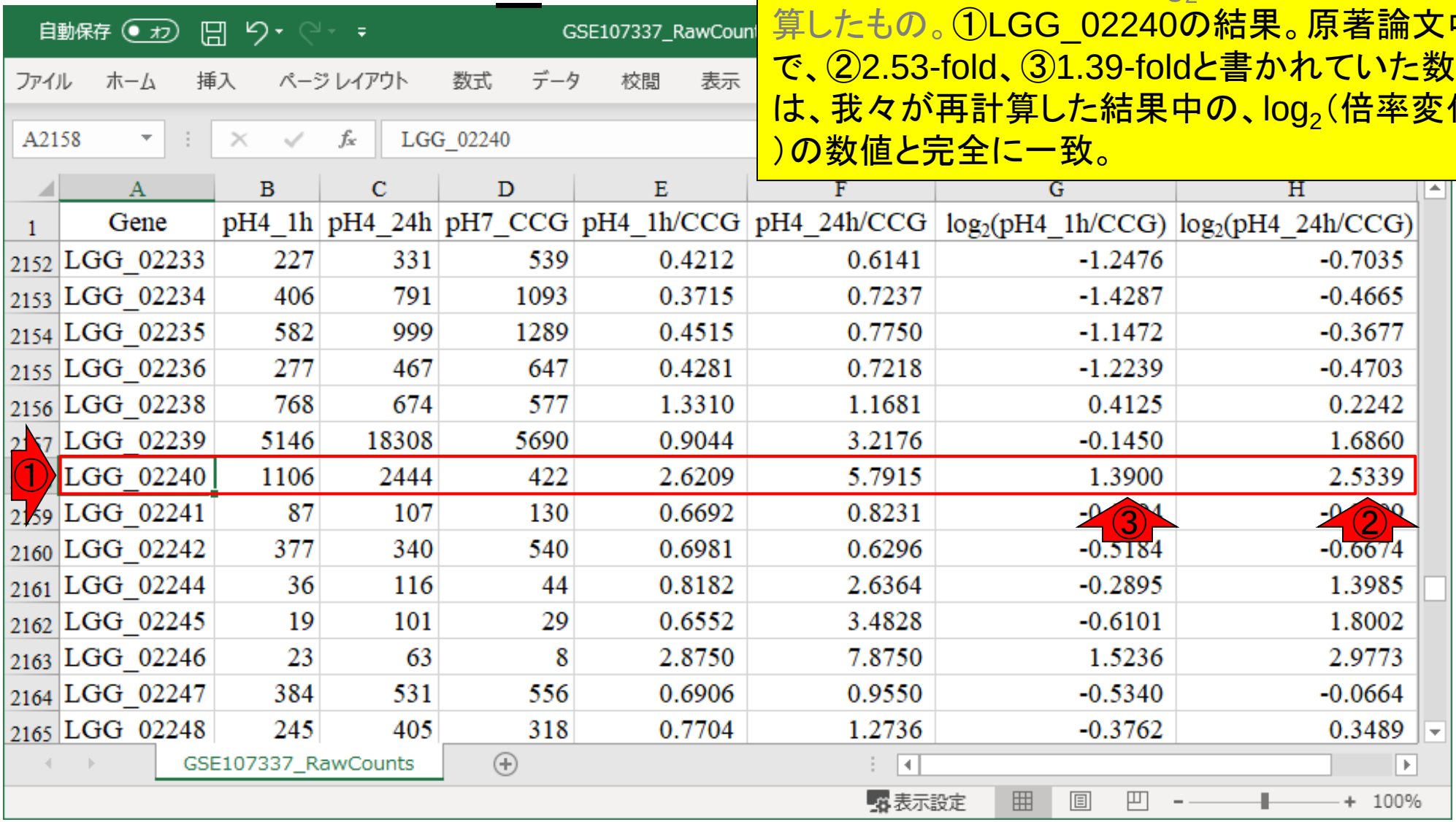
A2158　LGG_02240

	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	log2(pH4_1h/CCG)	log2(pH4_24h/CCG)
2152	LGG_02233	227	331	539	0.4212	0.6141	-1.2476	-0.7035
2153	LGG_02234	406	791	1093	0.3715	0.7237	-1.4287	-0.4665
2154	LGG_02235	582	999	1289	0.4515	0.7750	-1.1472	-0.3677
2155	LGG_02236	277	467	647	0.4281	0.7218	-1.2239	-0.4703
2156	LGG_02238	768	674	577	1.3310	1.1681	0.4125	0.2242
2157	LGG_02239	5146	18308	5690	0.9044	3.2176	-0.1450	1.6860
①	LGG_02240	1106	2444	422	2.6209	5.7915	1.3900	2.5339
2159	LGG_02241	87	107	130	0.6692	0.8231	-0.5794	-0.2809
2160	LGG_02242	377	340	540	0.6981	0.6296	-0.5184	-0.6674
2161	LGG_02244	36	116	44	0.8182	2.6364	-0.2895	1.3985
2162	LGG_02245	19	101	29	0.6552	3.4828	-0.6101	1.8002
2163	LGG_02246	23	63	8	2.8750	7.8750	1.5236	2.9773
2164	LGG_02247	384	531	556	0.6906	0.9550	-0.5340	-0.0664
2165	LGG_02248	245	405	318	0.7704	1.2736	-0.3762	0.3489

GSE107337_RawCounts　表示設定　100%

W11-5：LGG_02240

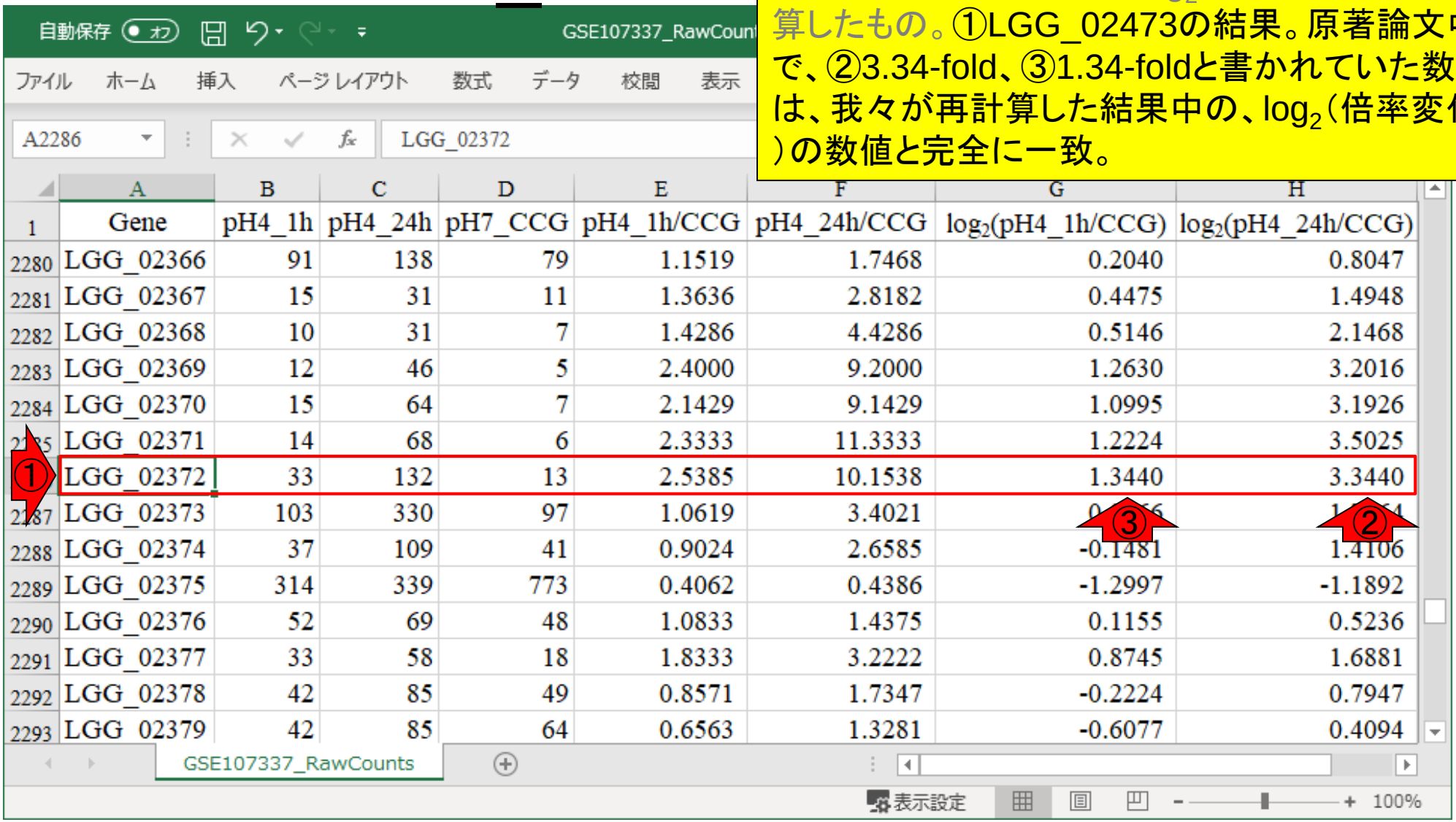
おさらい。今眺めているものは、著者らがGEO上で公開しているGSE107337のraw countデータから、我々が倍率変化や$\log_2$（倍率変化）を再計算したもの。①LGG_02240の結果。原著論文中で、②2.53-fold、③1.39-foldと書かれていた数値は、我々が再計算した結果中の、$\log_2$（倍率変化）の数値と完全に一致。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
2152	LGG_02233	227	331	539	0.4212	0.6141	-1.2476	-0.7035
2153	LGG_02234	406	791	1093	0.3715	0.7237	-1.4287	-0.4665
2154	LGG_02235	582	999	1289	0.4515	0.7750	-1.1472	-0.3677
2155	LGG_02236	277	467	647	0.4281	0.7218	-1.2239	-0.4703
2156	LGG_02238	768	674	577	1.3310	1.1681	0.4125	0.2242
2157	LGG_02239	5146	18308	5690	0.9044	3.2176	-0.1450	1.6860
2158	LGG_02240	1106	2444	422	2.6209	5.7915	1.3900	2.5339
2159	LGG_02241	87	107	130	0.6692	0.8231	-0.5184	-0.6074
2160	LGG_02242	377	340	540	0.6981	0.6296	-0.5184	-0.6674
2161	LGG_02244	36	116	44	0.8182	2.6364	-0.2895	1.3985
2162	LGG_02245	19	101	29	0.6552	3.4828	-0.6101	1.8002
2163	LGG_02246	23	63	8	2.8750	7.8750	1.5236	2.9773
2164	LGG_02247	384	531	556	0.6906	0.9550	-0.5340	-0.0664
2165	LGG_02248	245	405	318	0.7704	1.2736	-0.3762	0.3489

W11-5：LGG_02372

おさらい。今眺めているものは、著者らがGEO上で公開しているGSE107337のraw countデータから、我々が倍率変化や$\log_2$（倍率変化）を再計算したもの。①LGG_02473の結果。原著論文中で、②3.34-fold、③1.34-foldと書かれていた数値は、我々が再計算した結果中の、$\log_2$（倍率変化）の数値と完全に一致。



自動保存 オフ　GSE107337_RawCount

ファイル　ホーム　挿入　ページ レイアウト　数式　データ　校閲　表示

A2286　LGG_02372

	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
2280	LGG_02366	91	138	79	1.1519	1.7468	0.2040	0.8047
2281	LGG_02367	15	31	11	1.3636	2.8182	0.4475	1.4948
2282	LGG_02368	10	31	7	1.4286	4.4286	0.5146	2.1468
2283	LGG_02369	12	46	5	2.4000	9.2000	1.2630	3.2016
2284	LGG_02370	15	64	7	2.1429	9.1429	1.0995	3.1926
2285	LGG_02371	14	68	6	2.3333	11.3333	1.2224	3.5025
①	LGG_02372	33	132	13	2.5385	10.1538	1.3440	3.3440
2287	LGG_02373	103	330	97	1.0619	3.4021	③	②
2288	LGG_02374	37	109	41	0.9024	2.6585	-0.1481	1.4106
2289	LGG_02375	314	339	773	0.4062	0.4386	-1.2997	-1.1892
2290	LGG_02376	52	69	48	1.0833	1.4375	0.1155	0.5236
2291	LGG_02377	33	58	18	1.8333	3.2222	0.8745	1.6881
2292	LGG_02378	42	85	49	0.8571	1.7347	-0.2224	0.7947
2293	LGG_02379	42	85	64	0.6563	1.3281	-0.6077	0.4094

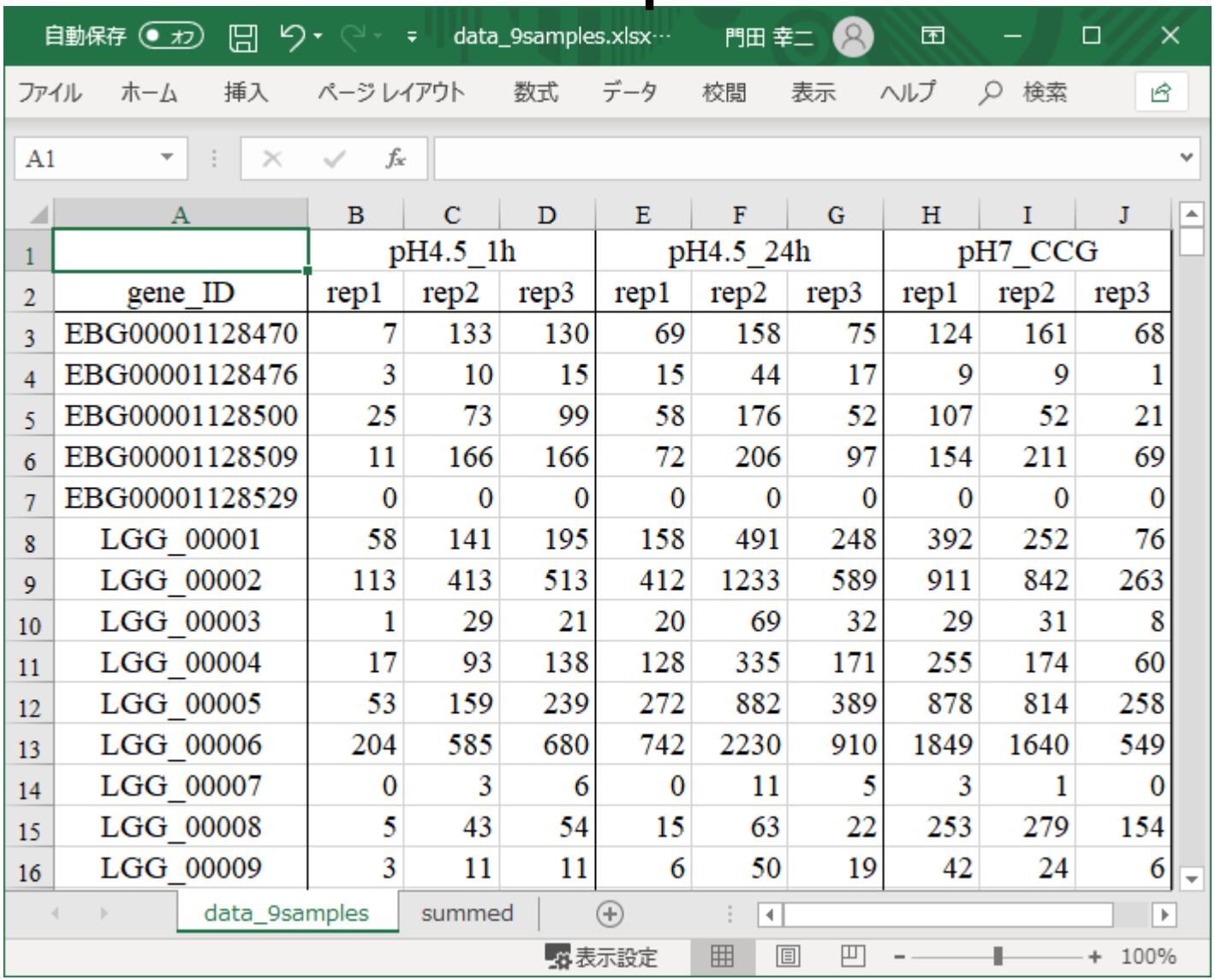
GSE107337_RawCounts　表示設定　100%

Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

W12-1：9 samples

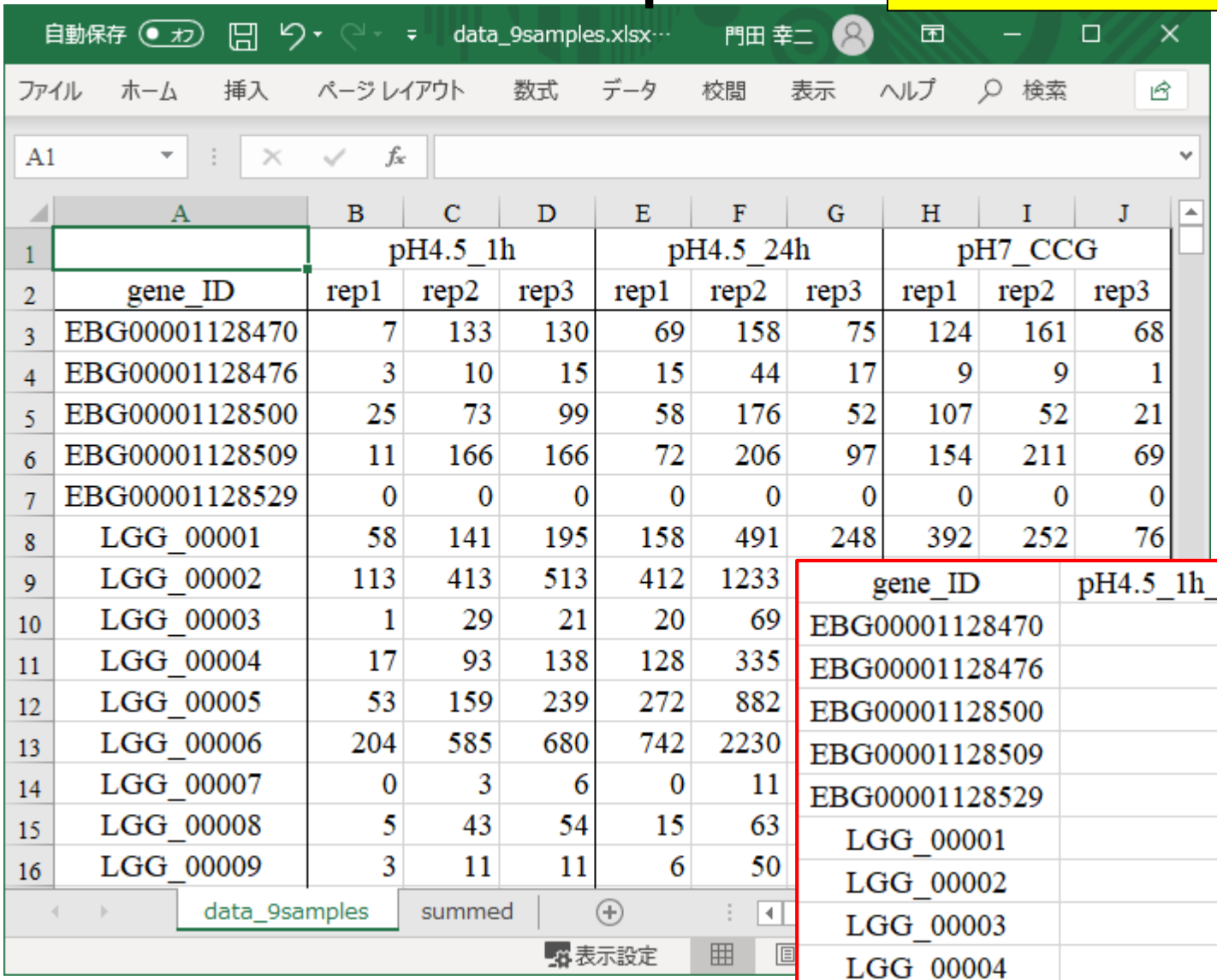
W9-10までと同様の手順で、9サンプル分のカウントデータ取得まで行った結果。



	A	B	C	D	E	F	G	H	I	J
1		pH4.5_1h			pH4.5_24h			pH7_CCG		
2	gene_ID	rep1	rep2	rep3	rep1	rep2	rep3	rep1	rep2	rep3
3	EBG00001128470	7	133	130	69	158	75	124	161	68
4	EBG00001128476	3	10	15	15	44	17	9	9	1
5	EBG00001128500	25	73	99	58	176	52	107	52	21
6	EBG00001128509	11	166	166	72	206	97	154	211	69
7	EBG00001128529	0	0	0	0	0	0	0	0	0
8	LGG_00001	58	141	195	158	491	248	392	252	76
9	LGG_00002	113	413	513	412	1233	589	911	842	263
10	LGG_00003	1	29	21	20	69	32	29	31	8
11	LGG_00004	17	93	138	128	335	171	255	174	60
12	LGG_00005	53	159	239	272	882	389	878	814	258
13	LGG_00006	204	585	680	742	2230	910	1849	1640	549
14	LGG_00007	0	3	6	0	11	5	3	1	0
15	LGG_00008	5	43	54	15	63	22	253	279	154
16	LGG_00009	3	11	11	6	50	19	42	24	6

W12-1：9 samples

W9-10までと同様の手順で、9サンプル分のカウントデータ取得まで行った結果。赤枠の3サンプル分のみのカウントデータ（W9-10）も表示。



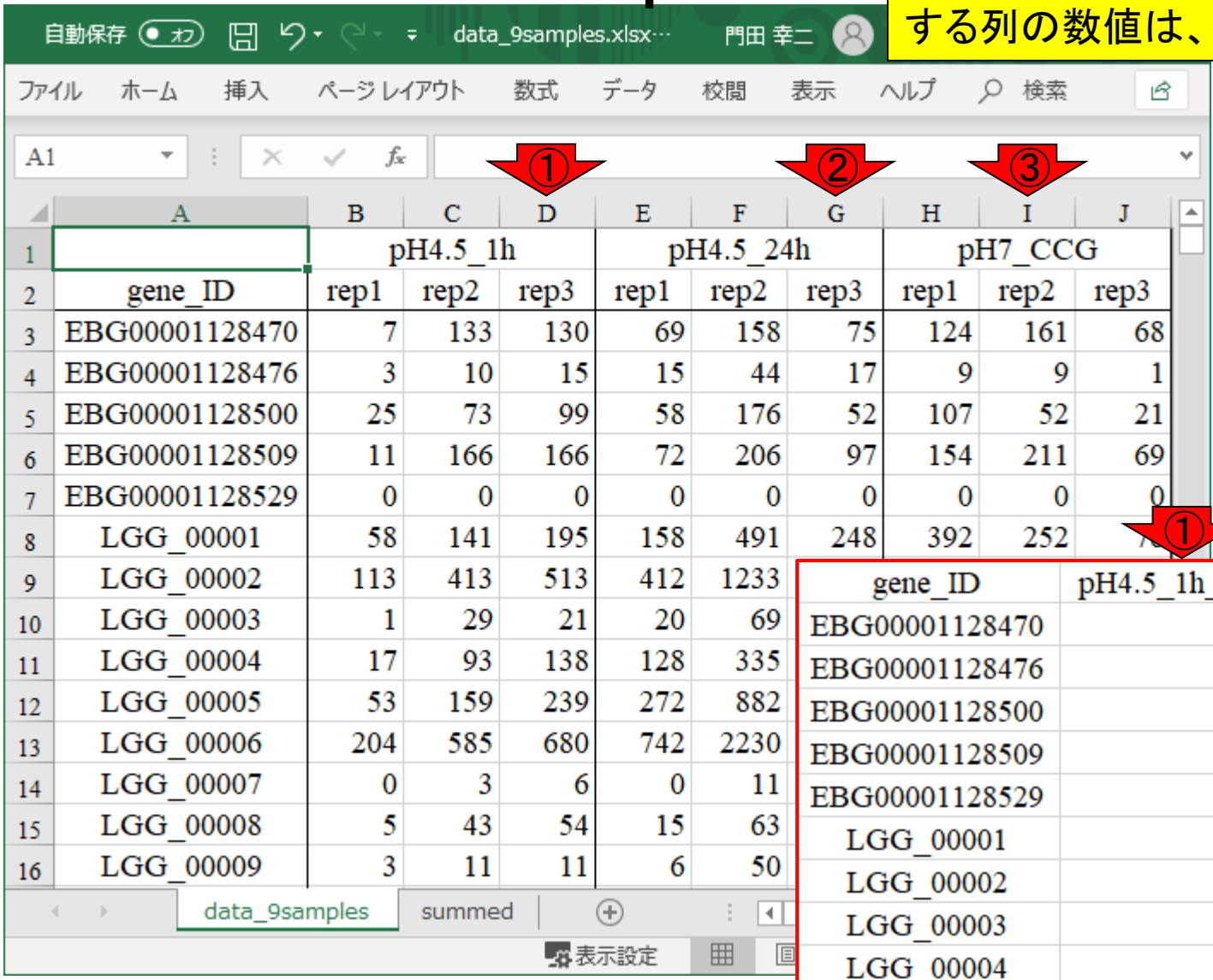
	A	B	C	D	E	F	G	H	I	J
1		pH4.5_1h			pH4.5_24h			pH7_CCG		
2	gene_ID	rep1	rep2	rep3	rep1	rep2	rep3	rep1	rep2	rep3
3	EBG00001128470	7	133	130	69	158	75	124	161	68
4	EBG00001128476	3	10	15	15	44	17	9	9	1
5	EBG00001128500	25	73	99	58	176	52	107	52	21
6	EBG00001128509	11	166	166	72	206	97	154	211	69
7	EBG00001128529	0	0	0	0	0	0	0	0	0
8	LGG_00001	58	141	195	158	491	248	392	252	76
9	LGG_00002	113	413	513	412	1233				
10	LGG_00003	1	29	21	20	69				
11	LGG_00004	17	93	138	128	335				
12	LGG_00005	53	159	239	272	882				
13	LGG_00006	204	585	680	742	2230				
14	LGG_00007	0	3	6	0	11				
15	LGG_00008	5	43	54	15	63				
16	LGG_00009	3	11	11	6	50				

data_9samples　summed　表示設定

gene_ID	pH4.5_1h_rep3	pH4.5_24h_rep3	pH7_CCG_rep2
EBG00001128470	130	75	161
EBG00001128476	15	17	9
EBG00001128500	99	52	52
EBG00001128509	166	97	211
EBG00001128529	0	0	0
LGG_00001	195	248	252
LGG_00002	513	589	842
LGG_00003	21	32	31
LGG_00004	138	171	174
LGG_00005	239	389	814

W12-1：9 samples

W9-10までと同様の手順で、9サンプル分のカウントデータ取得まで行った結果。赤枠の3サンプル分のみのカウントデータ（W9-10）も表示。対応する列の数値は、確かに一致している。



	A	B	C	D	E	F	G	H	I	J
1		pH4.5_1h			pH4.5_24h			pH7_CCG		
2	gene_ID	rep1	rep2	rep3	rep1	rep2	rep3	rep1	rep2	rep3
3	EBG00001128470	7	133	130	69	158	75	124	161	68
4	EBG00001128476	3	10	15	15	44	17	9	9	1
5	EBG00001128500	25	73	99	58	176	52	107	52	21
6	EBG00001128509	11	166	166	72	206	97	154	211	69
7	EBG00001128529	0	0	0	0	0	0	0	0	0
8	LGG_00001	58	141	195	158	491	248	392	252	
9	LGG_00002	113	413	513	412	1233				
10	LGG_00003	1	29	21	20	69				
11	LGG_00004	17	93	138	128	335				
12	LGG_00005	53	159	239	272	882				
13	LGG_00006	204	585	680	742	2230				
14	LGG_00007	0	3	6	0	11				
15	LGG_00008	5	43	54	15	63				
16	LGG_00009	3	11	11	6	50				

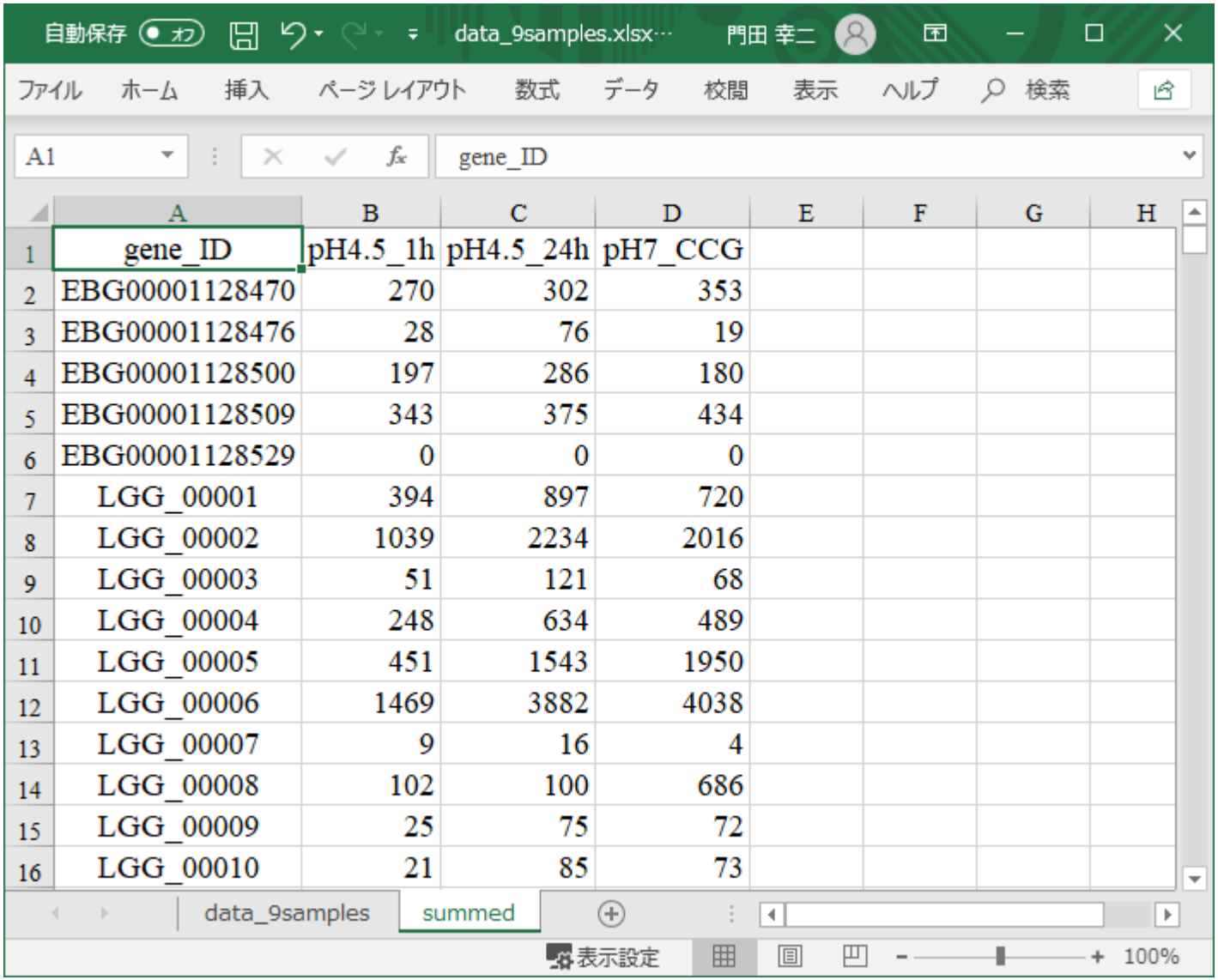
data_9samples　summed　表示設定

①　②　③

gene_ID	pH4.5_1h_rep3	pH4.5_24h_rep3	pH7_CCG_rep2
EBG00001128470	130	75	161
EBG00001128476	15	17	9
EBG00001128500	99	52	52
EBG00001128509	166	97	211
EBG00001128529	0	0	0
LGG_00001	195	248	252
LGG_00002	513	589	842
LGG_00003	21	32	31
LGG_00004	138	171	174
LGG_00005	239	389	814

同一群内の3反復の和（sum）をとった結果。

W12-2：sum



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG				
2	EBG00001128470	270	302	353				
3	EBG00001128476	28	76	19				
4	EBG00001128500	197	286	180				
5	EBG00001128509	343	375	434				
6	EBG00001128529	0	0	0				
7	LGG_00001	394	897	720				
8	LGG_00002	1039	2234	2016				
9	LGG_00003	51	121	68				
10	LGG_00004	248	634	489				
11	LGG_00005	451	1543	1950				
12	LGG_00006	1469	3882	4038				
13	LGG_00007	9	16	4				
14	LGG_00008	102	100	686				
15	LGG_00009	25	75	72				
16	LGG_00010	21	85	73				

W12-2：sum

同一群内の3反復の和（sum）をとった結果。赤枠の原著論文のほうはaverage countであったが（W11-2）、averageとsumは数学的に等価と考えてよいので、似た数値分布になるsumのほうを採用した。

data_9samples.xlsx

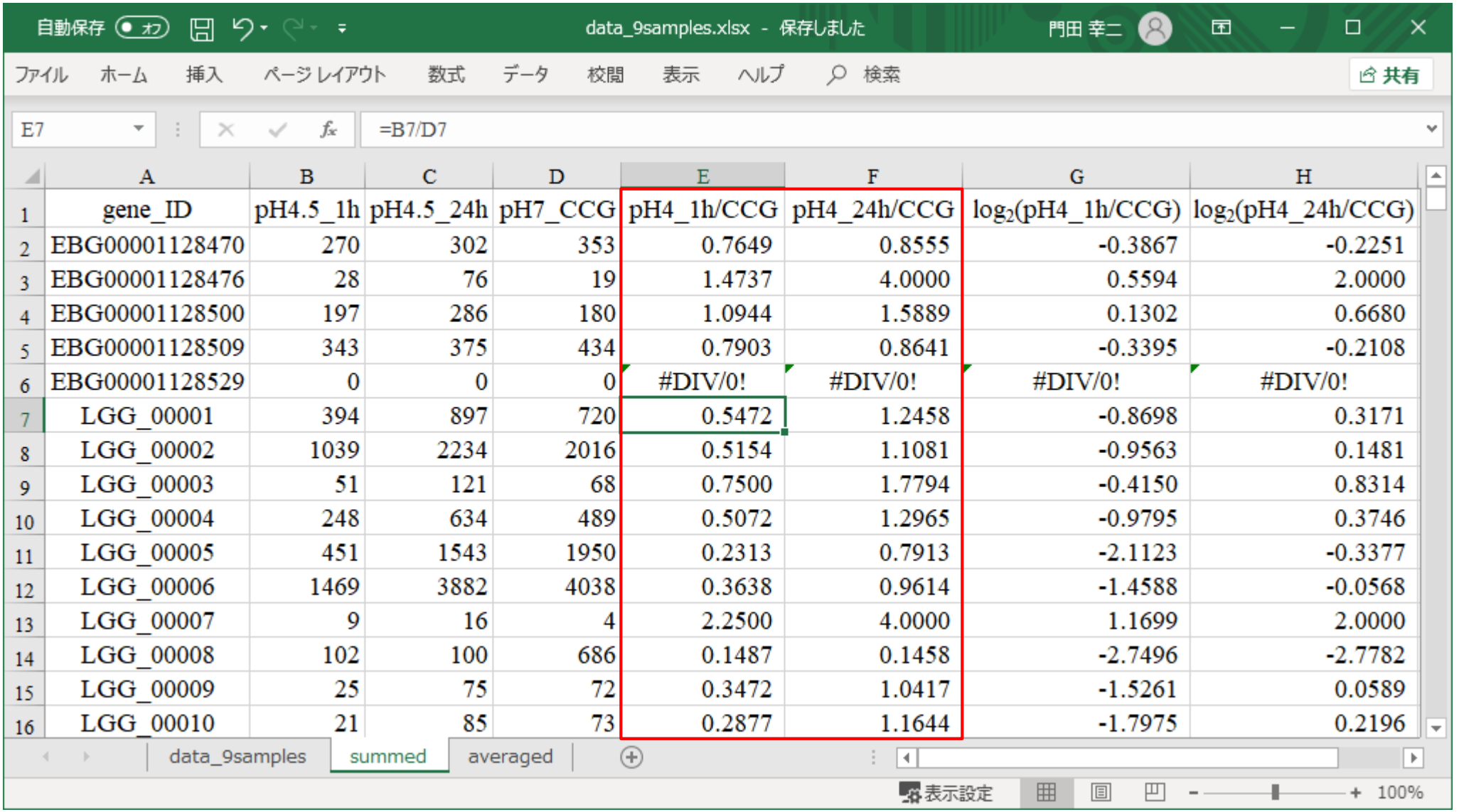
	A	B	C	D	E	F	G
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG			
2	EBG00001128470	270	302	353			
3	EBG00001128476	28	76	19			
4	EBG00001128500	197	286	180			
5	EBG00001128509	343	375	434			
6	EBG00001128529	0	0	0			
7	LGG_00001	394	897	720			
8	LGG_00002	1039	2234	2016			
9	LGG_00003	51	121	68			
10	LGG_00004	248	634	489			
11	LGG_00005	451	1543	1950			
12	LGG_00006	1469	3882	4038			
13	LGG_00007	9	16	4			
14	LGG_00008	102	100	686			
15	LGG_00009	25	75	72			
16	LGG_00010	21	85	73			

GSE107337_RawCoun…

	A	B	C	D
1	Gene	pH4_1h	pH4_24h	pH7_CCG
2	LGG_00001	286	676	545
3	LGG_00002	776	1601	1450
4	LGG_00003	71	114	91
5	LGG_00004	201	470	368
6	LGG_00005	368	1160	1447
7	LGG_00006	1041	2742	2880
8	LGG_00007	7	10	2
9	LGG_00008	73	76	499
10	LGG_00009	16	45	42
11	LGG_00010	20	65	57
12	LGG_00011	2281	2596	2150
13	LGG_00012	2510	3631	3097
14	LGG_00013	971	1301	1424

赤枠が倍率変化算出部分。

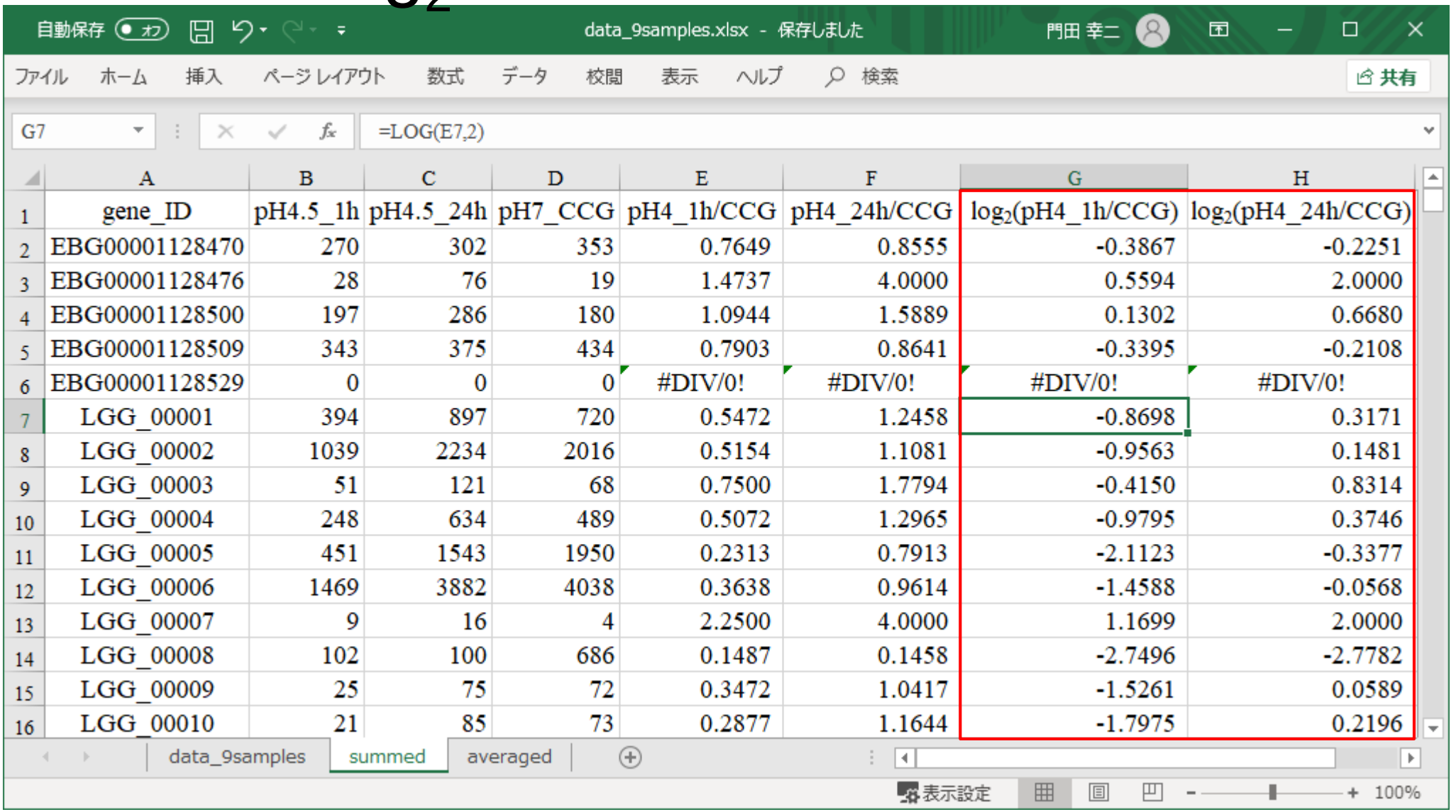
W12-3:倍率変化



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
2	EBG00001128470	270	302	353	0.7649	0.8555	-0.3867	-0.2251
3	EBG00001128476	28	76	19	1.4737	4.0000	0.5594	2.0000
4	EBG00001128500	197	286	180	1.0944	1.5889	0.1302	0.6680
5	EBG00001128509	343	375	434	0.7903	0.8641	-0.3395	-0.2108
6	EBG00001128529	0	0	0	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
7	LGG_00001	394	897	720	0.5472	1.2458	-0.8698	0.3171
8	LGG_00002	1039	2234	2016	0.5154	1.1081	-0.9563	0.1481
9	LGG_00003	51	121	68	0.7500	1.7794	-0.4150	0.8314
10	LGG_00004	248	634	489	0.5072	1.2965	-0.9795	0.3746
11	LGG_00005	451	1543	1950	0.2313	0.7913	-2.1123	-0.3377
12	LGG_00006	1469	3882	4038	0.3638	0.9614	-1.4588	-0.0568
13	LGG_00007	9	16	4	2.2500	4.0000	1.1699	2.0000
14	LGG_00008	102	100	686	0.1487	0.1458	-2.7496	-2.7782
15	LGG_00009	25	75	72	0.3472	1.0417	-1.5261	0.0589
16	LGG_00010	21	85	73	0.2877	1.1644	-1.7975	0.2196

赤枠がlog_2(倍率変化)算出部分。

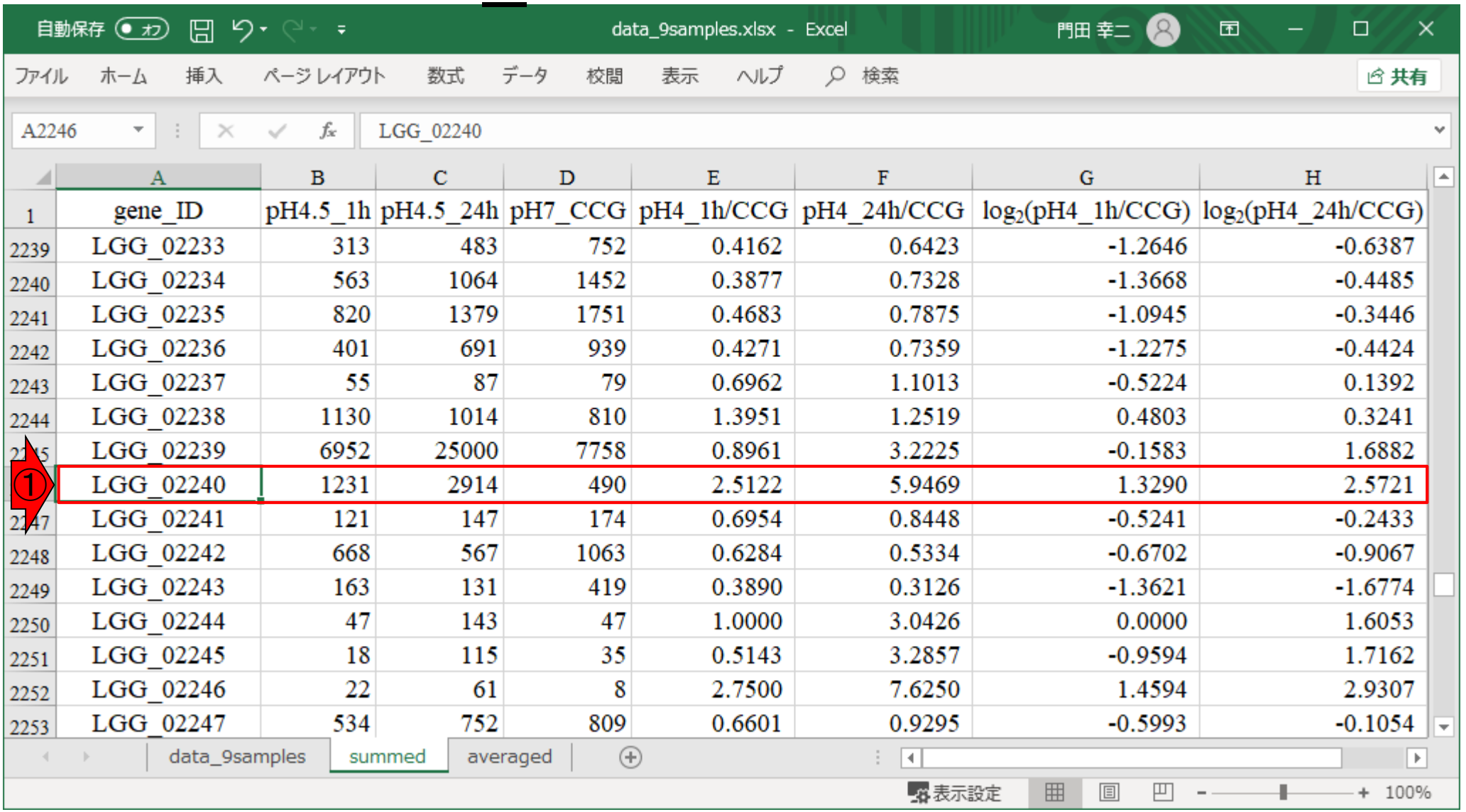
W12-4:log_2(倍率変化)



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	log_2(pH4_1h/CCG)	log_2(pH4_24h/CCG)
2	EBG00001128470	270	302	353	0.7649	0.8555	-0.3867	-0.2251
3	EBG00001128476	28	76	19	1.4737	4.0000	0.5594	2.0000
4	EBG00001128500	197	286	180	1.0944	1.5889	0.1302	0.6680
5	EBG00001128509	343	375	434	0.7903	0.8641	-0.3395	-0.2108
6	EBG00001128529	0	0	0	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
7	LGG_00001	394	897	720	0.5472	1.2458	-0.8698	0.3171
8	LGG_00002	1039	2234	2016	0.5154	1.1081	-0.9563	0.1481
9	LGG_00003	51	121	68	0.7500	1.7794	-0.4150	0.8314
10	LGG_00004	248	634	489	0.5072	1.2965	-0.9795	0.3746
11	LGG_00005	451	1543	1950	0.2313	0.7913	-2.1123	-0.3377
12	LGG_00006	1469	3882	4038	0.3638	0.9614	-1.4588	-0.0568
13	LGG_00007	9	16	4	2.2500	4.0000	1.1699	2.0000
14	LGG_00008	102	100	686	0.1487	0.1458	-2.7496	-2.7782
15	LGG_00009	25	75	72	0.3472	1.0417	-1.5261	0.0589
16	LGG_00010	21	85	73	0.2877	1.1644	-1.7975	0.2196

①LGG_02240の計算結果。

W12-5：LGG_02240



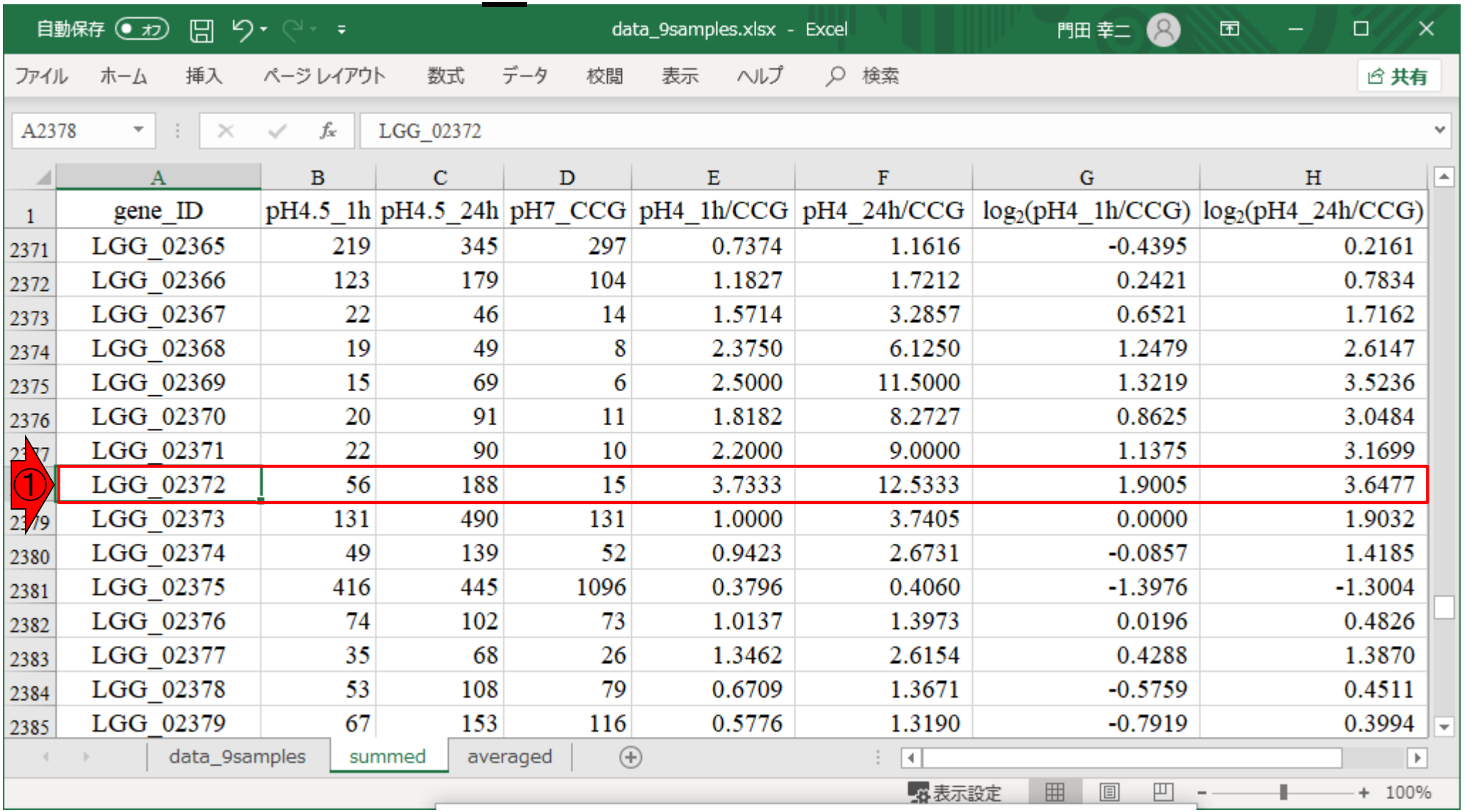
	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
2239	LGG_02233	313	483	752	0.4162	0.6423	-1.2646	-0.6387
2240	LGG_02234	563	1064	1452	0.3877	0.7328	-1.3668	-0.4485
2241	LGG_02235	820	1379	1751	0.4683	0.7875	-1.0945	-0.3446
2242	LGG_02236	401	691	939	0.4271	0.7359	-1.2275	-0.4424
2243	LGG_02237	55	87	79	0.6962	1.1013	-0.5224	0.1392
2244	LGG_02238	1130	1014	810	1.3951	1.2519	0.4803	0.3241
2245	LGG_02239	6952	25000	7758	0.8961	3.2225	-0.1583	1.6882
①	LGG_02240	1231	2914	490	2.5122	5.9469	1.3290	2.5721
2247	LGG_02241	121	147	174	0.6954	0.8448	-0.5241	-0.2433
2248	LGG_02242	668	567	1063	0.6284	0.5334	-0.6702	-0.9067
2249	LGG_02243	163	131	419	0.3890	0.3126	-1.3621	-1.6774
2250	LGG_02244	47	143	47	1.0000	3.0426	0.0000	1.6053
2251	LGG_02245	18	115	35	0.5143	3.2857	-0.9594	1.7162
2252	LGG_02246	22	61	8	2.7500	7.6250	1.4594	2.9307
2253	LGG_02247	534	752	809	0.6601	0.9295	-0.5993	-0.1054

data_9samples summed averaged

表示設定 100%

①LGG_02372の計算結果。

W12-5:LGG_02372



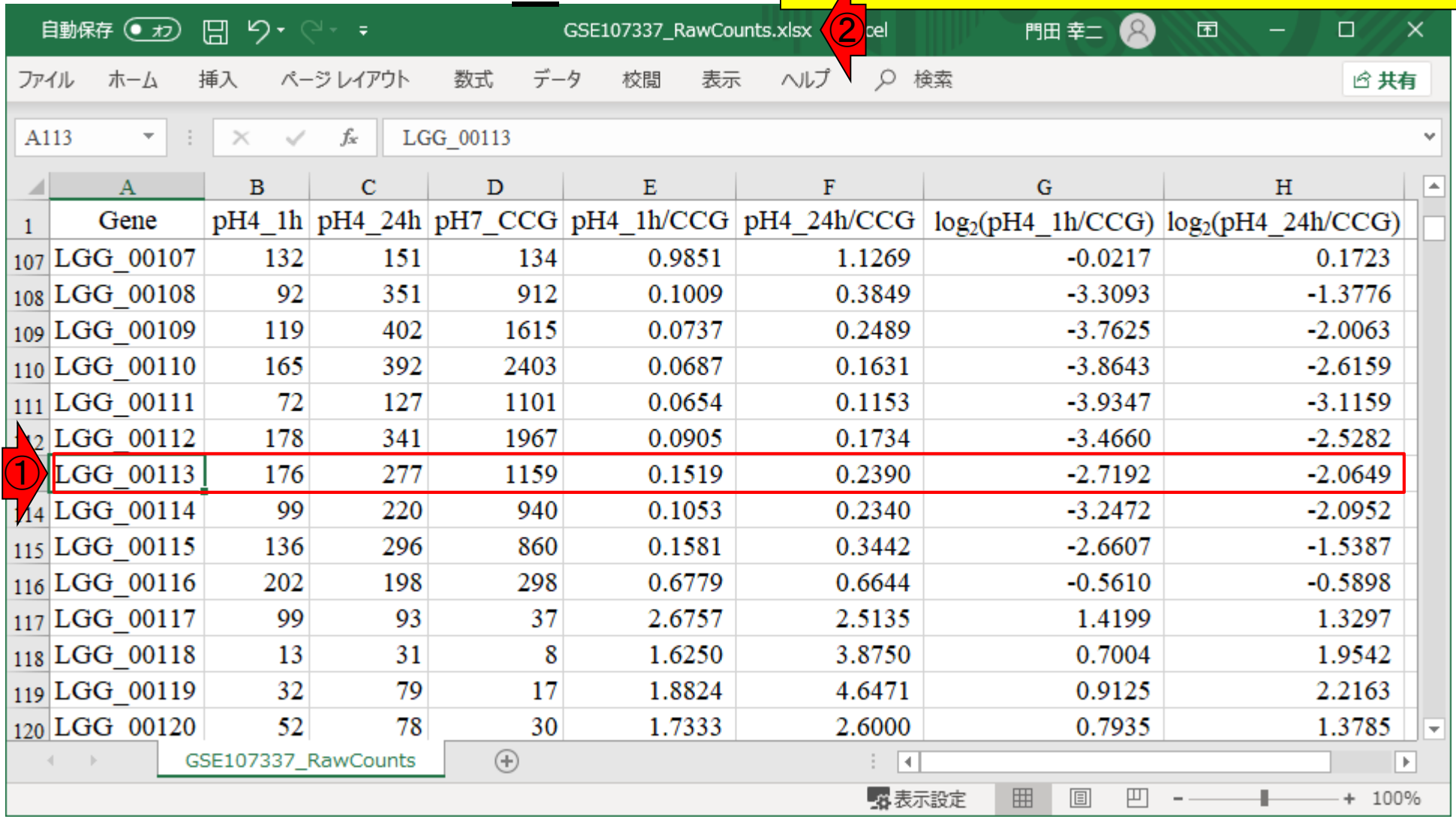
	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
2371	LGG_02365	219	345	297	0.7374	1.1616	-0.4395	0.2161
2372	LGG_02366	123	179	104	1.1827	1.7212	0.2421	0.7834
2373	LGG_02367	22	46	14	1.5714	3.2857	0.6521	1.7162
2374	LGG_02368	19	49	8	2.3750	6.1250	1.2479	2.6147
2375	LGG_02369	15	69	6	2.5000	11.5000	1.3219	3.5236
2376	LGG_02370	20	91	11	1.8182	8.2727	0.8625	3.0484
2377	LGG_02371	22	90	10	2.2000	9.0000	1.1375	3.1699
2378	LGG_02372	56	188	15	3.7333	12.5333	1.9005	3.6477
2379	LGG_02373	131	490	131	1.0000	3.7405	0.0000	1.9032
2380	LGG_02374	49	139	52	0.9423	2.6731	-0.0857	1.4185
2381	LGG_02375	416	445	1096	0.3796	0.4060	-1.3976	-1.3004
2382	LGG_02376	74	102	73	1.0137	1.3973	0.0196	0.4826
2383	LGG_02377	35	68	26	1.3462	2.6154	0.4288	1.3870
2384	LGG_02378	53	108	79	0.6709	1.3671	-0.5759	0.4511
2385	LGG_02379	67	153	116	0.5776	1.3190	-0.7919	0.3994

Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

W13-1：LGG_00113

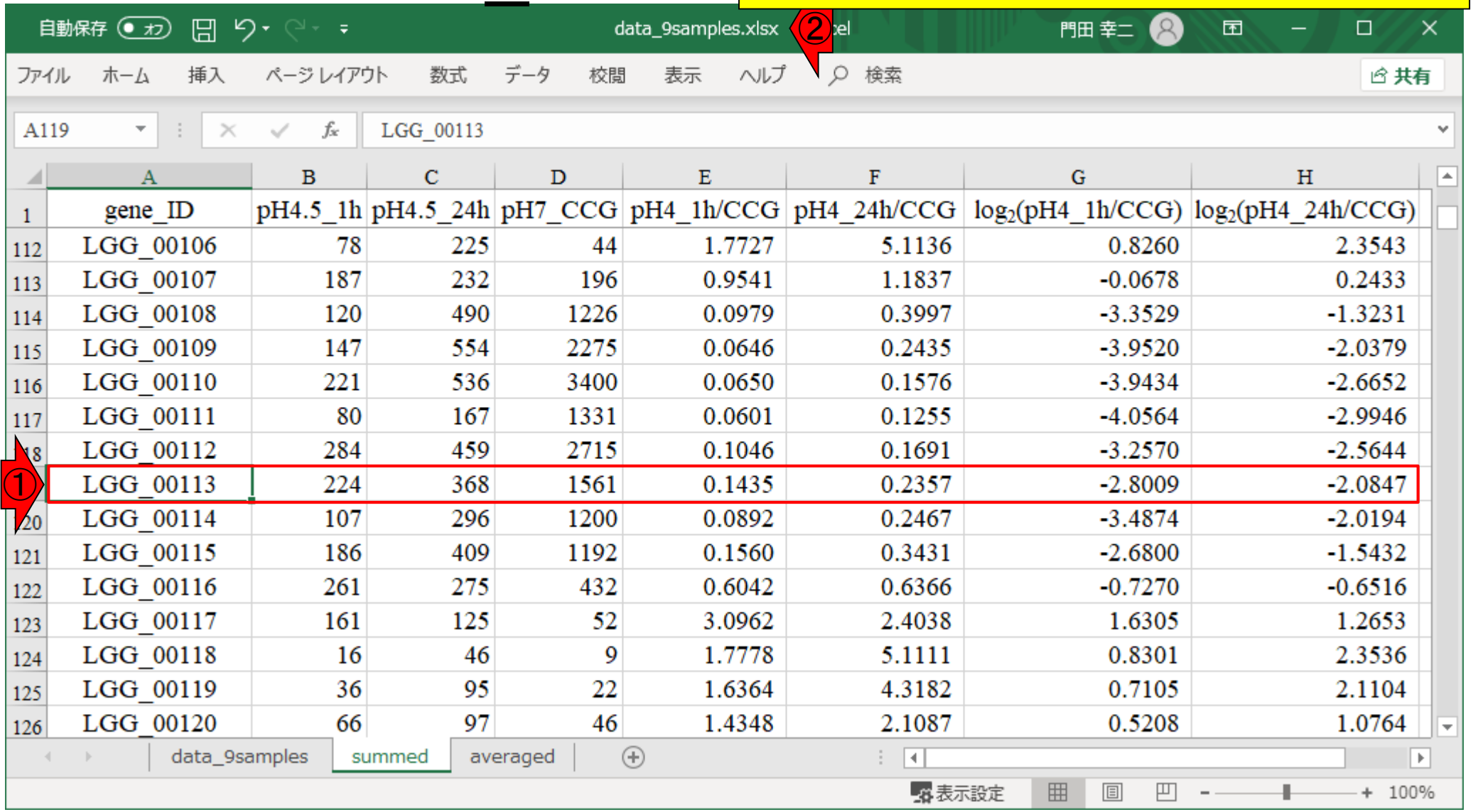
原著論文のp1608-1609にかけての文で述べられている、①LGG_00113の酸ストレス条件下における発現量低下を②公共データで確認。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
107	LGG_00107	132	151	134	0.9851	1.1269	-0.0217	0.1723
108	LGG_00108	92	351	912	0.1009	0.3849	-3.3093	-1.3776
109	LGG_00109	119	402	1615	0.0737	0.2489	-3.7625	-2.0063
110	LGG_00110	165	392	2403	0.0687	0.1631	-3.8643	-2.6159
111	LGG_00111	72	127	1101	0.0654	0.1153	-3.9347	-3.1159
112	LGG_00112	178	341	1967	0.0905	0.1734	-3.4660	-2.5282
113	LGG_00113	176	277	1159	0.1519	0.2390	-2.7192	-2.0649
114	LGG_00114	99	220	940	0.1053	0.2340	-3.2472	-2.0952
115	LGG_00115	136	296	860	0.1581	0.3442	-2.6607	-1.5387
116	LGG_00116	202	198	298	0.6779	0.6644	-0.5610	-0.5898
117	LGG_00117	99	93	37	2.6757	2.5135	1.4199	1.3297
118	LGG_00118	13	31	8	1.6250	3.8750	0.7004	1.9542
119	LGG_00119	32	79	17	1.8824	4.6471	0.9125	2.2163
120	LGG_00120	52	78	30	1.7333	2.6000	0.7935	1.3785

W13-1：LGG_00113

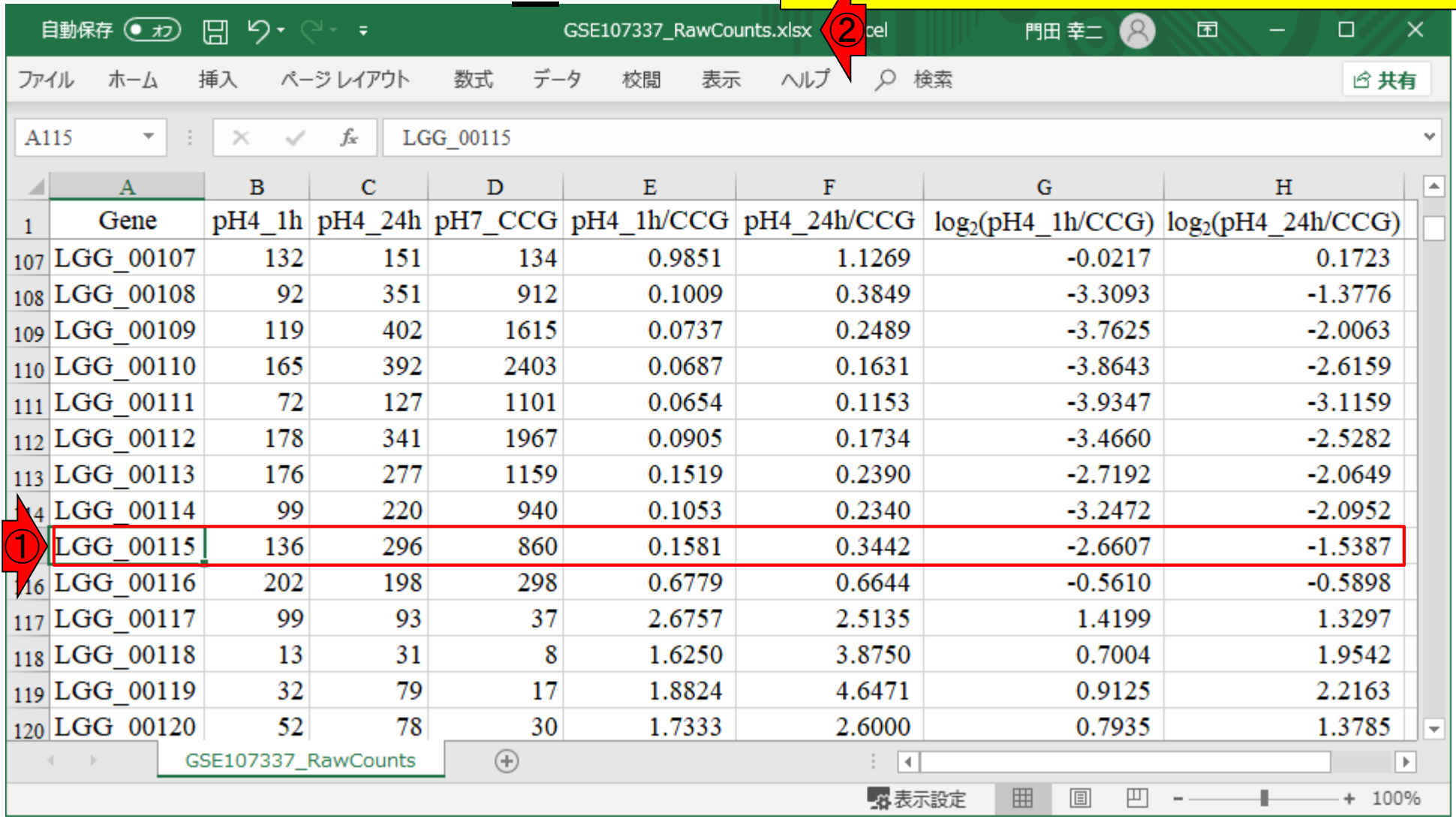
原著論文のp1608-1609にかけての文で述べられている、①LGG_00113の酸ストレス条件下における発現量低下を②自分らのデータで確認。



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
112	LGG_00106	78	225	44	1.7727	5.1136	0.8260	2.3543
113	LGG_00107	187	232	196	0.9541	1.1837	-0.0678	0.2433
114	LGG_00108	120	490	1226	0.0979	0.3997	-3.3529	-1.3231
115	LGG_00109	147	554	2275	0.0646	0.2435	-3.9520	-2.0379
116	LGG_00110	221	536	3400	0.0650	0.1576	-3.9434	-2.6652
117	LGG_00111	80	167	1331	0.0601	0.1255	-4.0564	-2.9946
118	LGG_00112	284	459	2715	0.1046	0.1691	-3.2570	-2.5644
119	LGG_00113	224	368	1561	0.1435	0.2357	-2.8009	-2.0847
120	LGG_00114	107	296	1200	0.0892	0.2467	-3.4874	-2.0194
121	LGG_00115	186	409	1192	0.1560	0.3431	-2.6800	-1.5432
122	LGG_00116	261	275	432	0.6042	0.6366	-0.7270	-0.6516
123	LGG_00117	161	125	52	3.0962	2.4038	1.6305	1.2653
124	LGG_00118	16	46	9	1.7778	5.1111	0.8301	2.3536
125	LGG_00119	36	95	22	1.6364	4.3182	0.7105	2.1104
126	LGG_00120	66	97	46	1.4348	2.1087	0.5208	1.0764

W13-2：LGG_00115

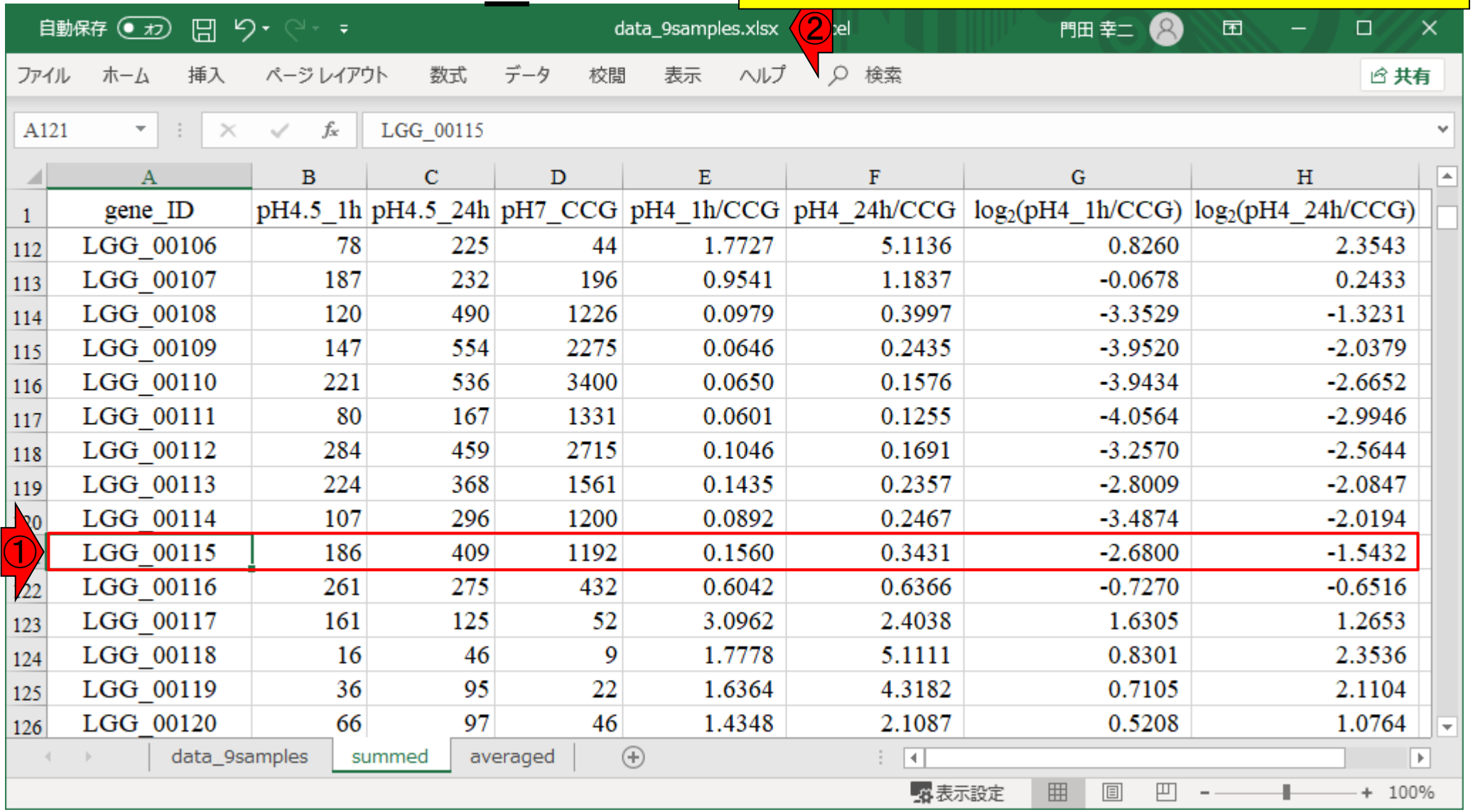
原著論文のp1608-1609にかけての文で述べられている、①LGG_00115の酸ストレス条件下における発現量低下を②公共データで確認。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
107	LGG_00107	132	151	134	0.9851	1.1269	-0.0217	0.1723
108	LGG_00108	92	351	912	0.1009	0.3849	-3.3093	-1.3776
109	LGG_00109	119	402	1615	0.0737	0.2489	-3.7625	-2.0063
110	LGG_00110	165	392	2403	0.0687	0.1631	-3.8643	-2.6159
111	LGG_00111	72	127	1101	0.0654	0.1153	-3.9347	-3.1159
112	LGG_00112	178	341	1967	0.0905	0.1734	-3.4660	-2.5282
113	LGG_00113	176	277	1159	0.1519	0.2390	-2.7192	-2.0649
114	LGG_00114	99	220	940	0.1053	0.2340	-3.2472	-2.0952
115	LGG_00115	136	296	860	0.1581	0.3442	-2.6607	-1.5387
116	LGG_00116	202	198	298	0.6779	0.6644	-0.5610	-0.5898
117	LGG_00117	99	93	37	2.6757	2.5135	1.4199	1.3297
118	LGG_00118	13	31	8	1.6250	3.8750	0.7004	1.9542
119	LGG_00119	32	79	17	1.8824	4.6471	0.9125	2.2163
120	LGG_00120	52	78	30	1.7333	2.6000	0.7935	1.3785

W13-2：LGG_00115

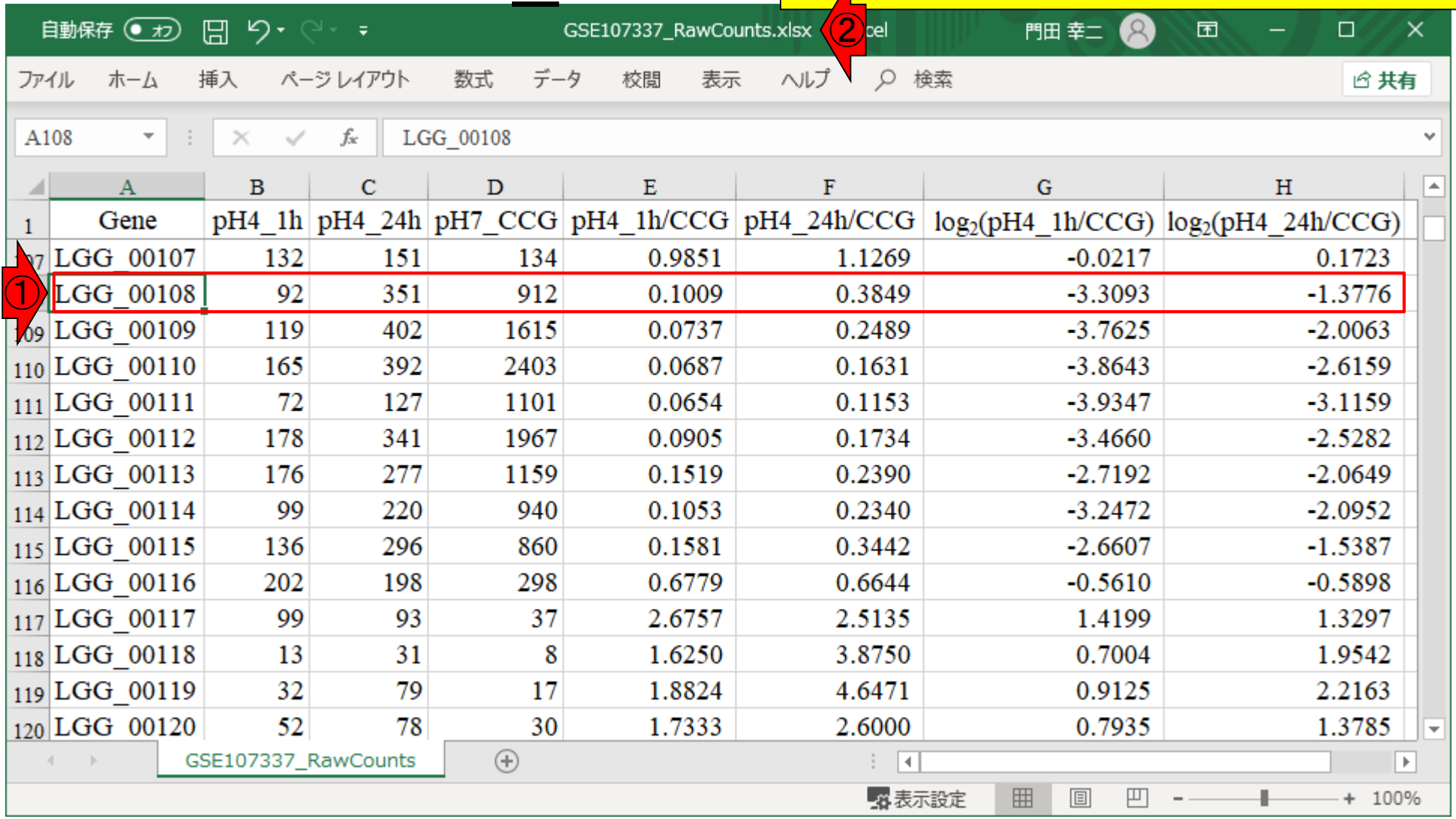
原著論文のp1608-1609にかけての文で述べられている、①LGG_00115の酸ストレス条件下における発現量低下を②自分らのデータで確認。



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
112	LGG_00106	78	225	44	1.7727	5.1136	0.8260	2.3543
113	LGG_00107	187	232	196	0.9541	1.1837	-0.0678	0.2433
114	LGG_00108	120	490	1226	0.0979	0.3997	-3.3529	-1.3231
115	LGG_00109	147	554	2275	0.0646	0.2435	-3.9520	-2.0379
116	LGG_00110	221	536	3400	0.0650	0.1576	-3.9434	-2.6652
117	LGG_00111	80	167	1331	0.0601	0.1255	-4.0564	-2.9946
118	LGG_00112	284	459	2715	0.1046	0.1691	-3.2570	-2.5644
119	LGG_00113	224	368	1561	0.1435	0.2357	-2.8009	-2.0847
120	LGG_00114	107	296	1200	0.0892	0.2467	-3.4874	-2.0194
121	LGG_00115	186	409	1192	0.1560	0.3431	-2.6800	-1.5432
122	LGG_00116	261	275	432	0.6042	0.6366	-0.7270	-0.6516
123	LGG_00117	161	125	52	3.0962	2.4038	1.6305	1.2653
124	LGG_00118	16	46	9	1.7778	5.1111	0.8301	2.3536
125	LGG_00119	36	95	22	1.6364	4.3182	0.7105	2.1104
126	LGG_00120	66	97	46	1.4348	2.1087	0.5208	1.0764

W13-3：LGG_00108

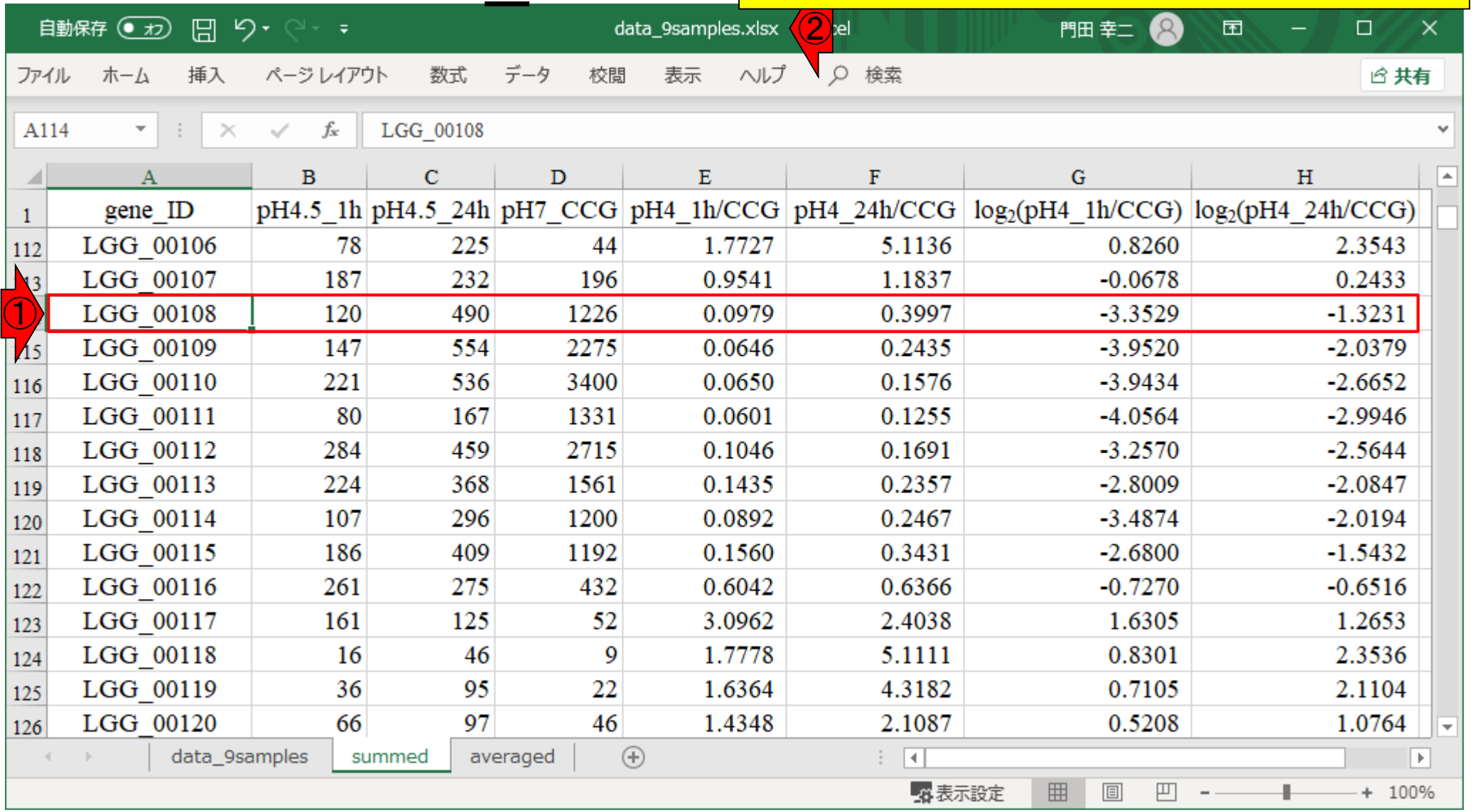
原著論文のp1608-1609にかけての文で述べられている、①LGG_00108の酸ストレス条件下における発現量低下を②公共データで確認。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
107	LGG_00107	132	151	134	0.9851	1.1269	-0.0217	0.1723
①	LGG_00108	92	351	912	0.1009	0.3849	-3.3093	-1.3776
109	LGG_00109	119	402	1615	0.0737	0.2489	-3.7625	-2.0063
110	LGG_00110	165	392	2403	0.0687	0.1631	-3.8643	-2.6159
111	LGG_00111	72	127	1101	0.0654	0.1153	-3.9347	-3.1159
112	LGG_00112	178	341	1967	0.0905	0.1734	-3.4660	-2.5282
113	LGG_00113	176	277	1159	0.1519	0.2390	-2.7192	-2.0649
114	LGG_00114	99	220	940	0.1053	0.2340	-3.2472	-2.0952
115	LGG_00115	136	296	860	0.1581	0.3442	-2.6607	-1.5387
116	LGG_00116	202	198	298	0.6779	0.6644	-0.5610	-0.5898
117	LGG_00117	99	93	37	2.6757	2.5135	1.4199	1.3297
118	LGG_00118	13	31	8	1.6250	3.8750	0.7004	1.9542
119	LGG_00119	32	79	17	1.8824	4.6471	0.9125	2.2163
120	LGG_00120	52	78	30	1.7333	2.6000	0.7935	1.3785

W13-3：LGG_00108

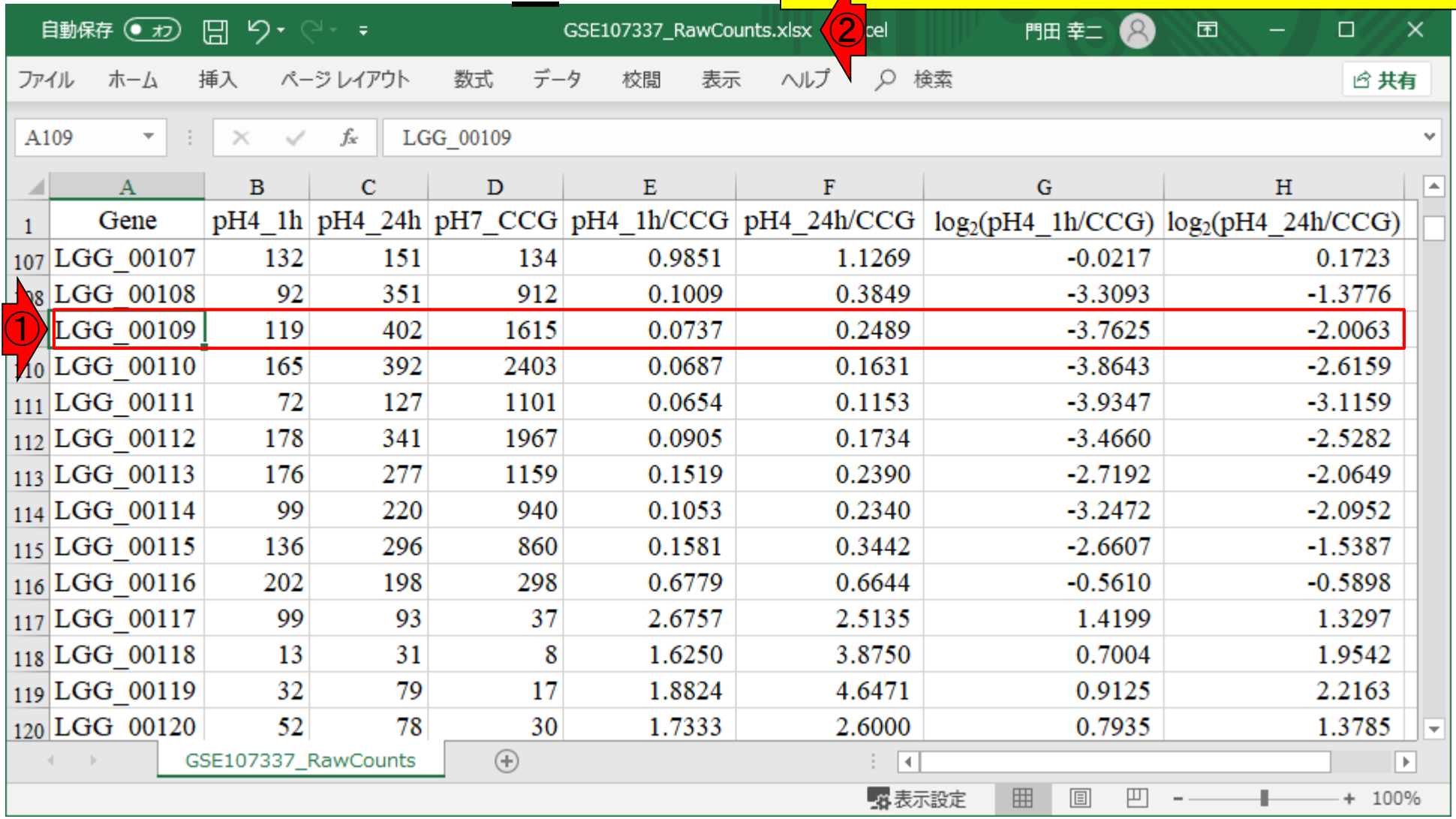
原著論文のp1608-1609にかけての文で述べられている、①LGG_00108の酸ストレス条件下における発現量低下を②自分らのデータで確認。



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
112	LGG_00106	78	225	44	1.7727	5.1136	0.8260	2.3543
113	LGG_00107	187	232	196	0.9541	1.1837	-0.0678	0.2433
114	LGG_00108	120	490	1226	0.0979	0.3997	-3.3529	-1.3231
115	LGG_00109	147	554	2275	0.0646	0.2435	-3.9520	-2.0379
116	LGG_00110	221	536	3400	0.0650	0.1576	-3.9434	-2.6652
117	LGG_00111	80	167	1331	0.0601	0.1255	-4.0564	-2.9946
118	LGG_00112	284	459	2715	0.1046	0.1691	-3.2570	-2.5644
119	LGG_00113	224	368	1561	0.1435	0.2357	-2.8009	-2.0847
120	LGG_00114	107	296	1200	0.0892	0.2467	-3.4874	-2.0194
121	LGG_00115	186	409	1192	0.1560	0.3431	-2.6800	-1.5432
122	LGG_00116	261	275	432	0.6042	0.6366	-0.7270	-0.6516
123	LGG_00117	161	125	52	3.0962	2.4038	1.6305	1.2653
124	LGG_00118	16	46	9	1.7778	5.1111	0.8301	2.3536
125	LGG_00119	36	95	22	1.6364	4.3182	0.7105	2.1104
126	LGG_00120	66	97	46	1.4348	2.1087	0.5208	1.0764

W13-4：LGG_00109

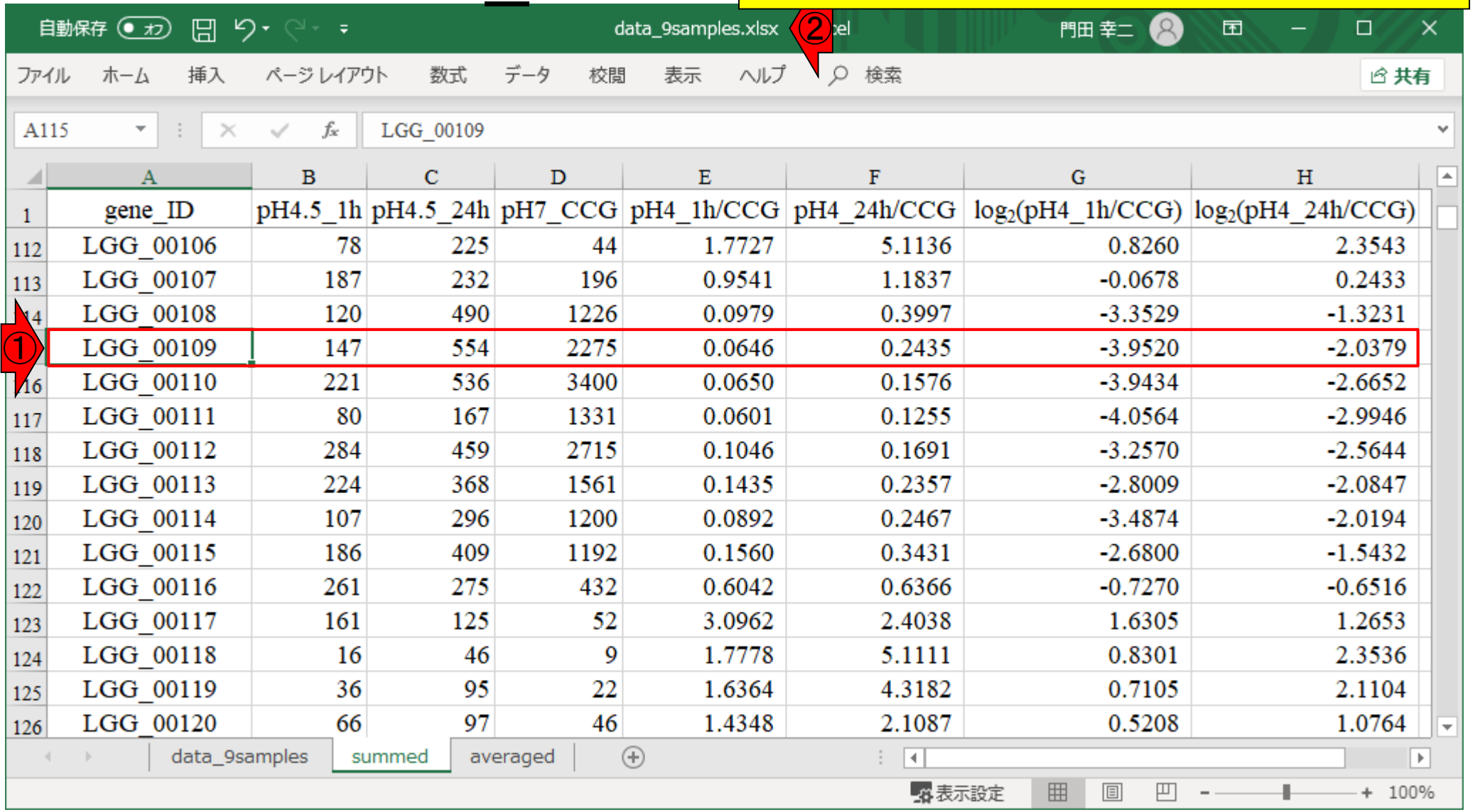
原著論文のp1608-1609にかけての文で述べられている、①LGG_00109の酸ストレス条件下における発現量低下を②公共データで確認。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
107	LGG_00107	132	151	134	0.9851	1.1269	-0.0217	0.1723
108	LGG_00108	92	351	912	0.1009	0.3849	-3.3093	-1.3776
109	LGG_00109	119	402	1615	0.0737	0.2489	-3.7625	-2.0063
110	LGG_00110	165	392	2403	0.0687	0.1631	-3.8643	-2.6159
111	LGG_00111	72	127	1101	0.0654	0.1153	-3.9347	-3.1159
112	LGG_00112	178	341	1967	0.0905	0.1734	-3.4660	-2.5282
113	LGG_00113	176	277	1159	0.1519	0.2390	-2.7192	-2.0649
114	LGG_00114	99	220	940	0.1053	0.2340	-3.2472	-2.0952
115	LGG_00115	136	296	860	0.1581	0.3442	-2.6607	-1.5387
116	LGG_00116	202	198	298	0.6779	0.6644	-0.5610	-0.5898
117	LGG_00117	99	93	37	2.6757	2.5135	1.4199	1.3297
118	LGG_00118	13	31	8	1.6250	3.8750	0.7004	1.9542
119	LGG_00119	32	79	17	1.8824	4.6471	0.9125	2.2163
120	LGG_00120	52	78	30	1.7333	2.6000	0.7935	1.3785

W13-4：LGG_00109

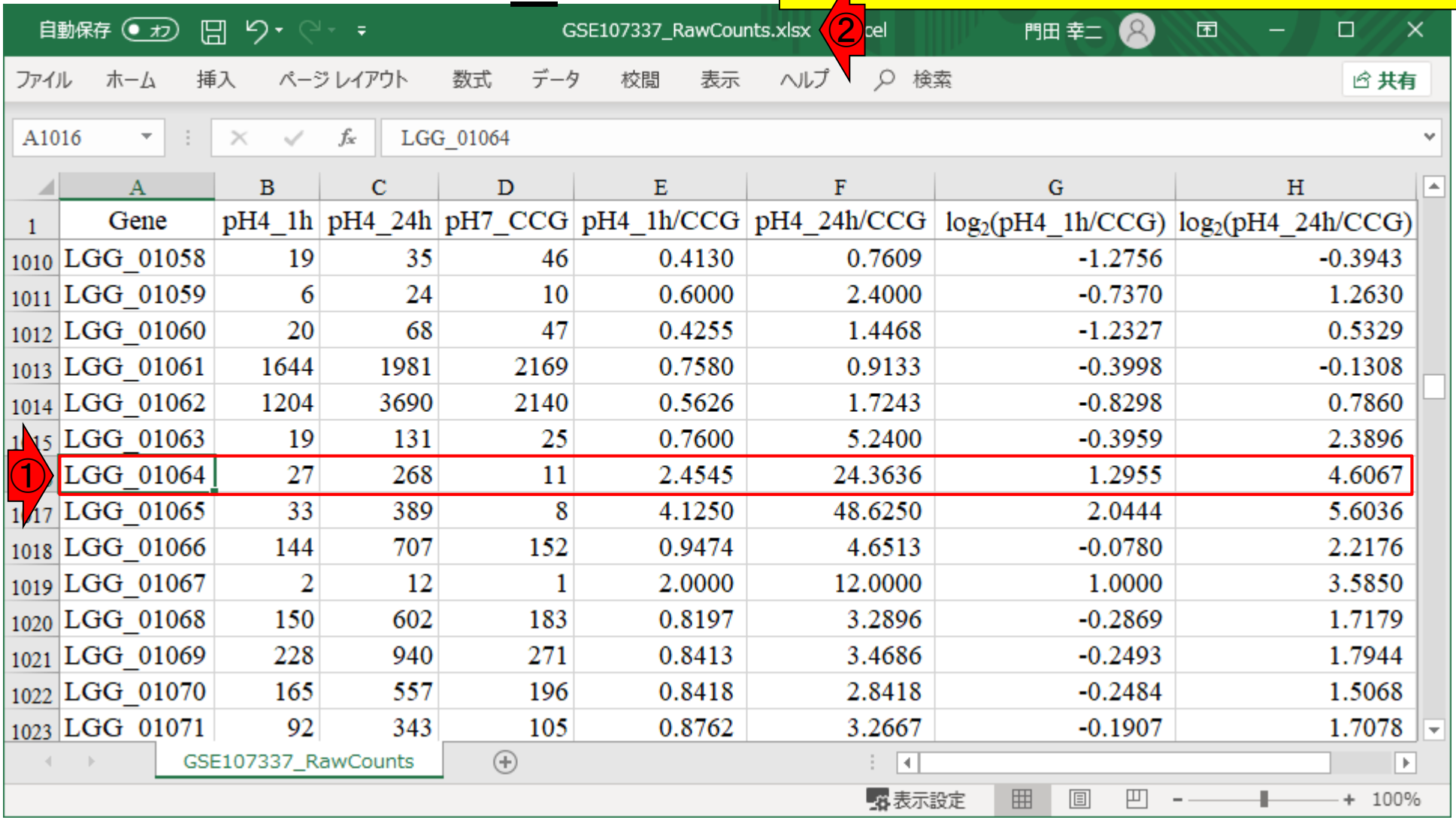
原著論文のp1608-1609にかけての文で述べられている、①LGG_00109の酸ストレス条件下における発現量低下を②自分らのデータで確認。



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
112	LGG_00106	78	225	44	1.7727	5.1136	0.8260	2.3543
113	LGG_00107	187	232	196	0.9541	1.1837	-0.0678	0.2433
114	LGG_00108	120	490	1226	0.0979	0.3997	-3.3529	-1.3231
115	LGG_00109	147	554	2275	0.0646	0.2435	-3.9520	-2.0379
116	LGG_00110	221	536	3400	0.0650	0.1576	-3.9434	-2.6652
117	LGG_00111	80	167	1331	0.0601	0.1255	-4.0564	-2.9946
118	LGG_00112	284	459	2715	0.1046	0.1691	-3.2570	-2.5644
119	LGG_00113	224	368	1561	0.1435	0.2357	-2.8009	-2.0847
120	LGG_00114	107	296	1200	0.0892	0.2467	-3.4874	-2.0194
121	LGG_00115	186	409	1192	0.1560	0.3431	-2.6800	-1.5432
122	LGG_00116	261	275	432	0.6042	0.6366	-0.7270	-0.6516
123	LGG_00117	161	125	52	3.0962	2.4038	1.6305	1.2653
124	LGG_00118	16	46	9	1.7778	5.1111	0.8301	2.3536
125	LGG_00119	36	95	22	1.6364	4.3182	0.7105	2.1104
126	LGG_00120	66	97	46	1.4348	2.1087	0.5208	1.0764

W14-1：LGG_01064

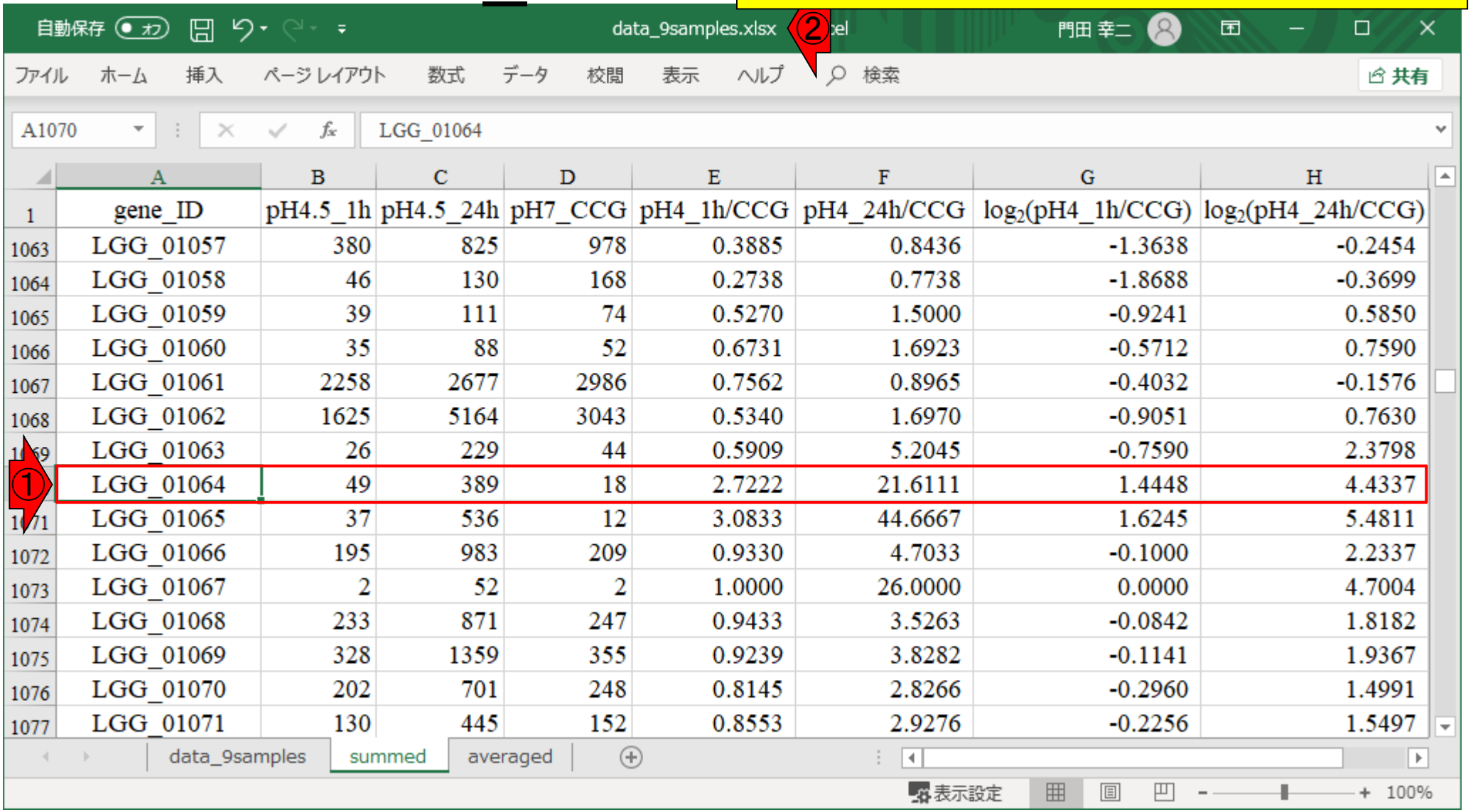
原著論文のp1611の左中央付近で述べられている、①LGG_01064の酸ストレス条件下における発現量上昇を②公共データで確認。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
1010	LGG_01058	19	35	46	0.4130	0.7609	-1.2756	-0.3943
1011	LGG_01059	6	24	10	0.6000	2.4000	-0.7370	1.2630
1012	LGG_01060	20	68	47	0.4255	1.4468	-1.2327	0.5329
1013	LGG_01061	1644	1981	2169	0.7580	0.9133	-0.3998	-0.1308
1014	LGG_01062	1204	3690	2140	0.5626	1.7243	-0.8298	0.7860
1015	LGG_01063	19	131	25	0.7600	5.2400	-0.3959	2.3896
1016	LGG_01064	27	268	11	2.4545	24.3636	1.2955	4.6067
1017	LGG_01065	33	389	8	4.1250	48.6250	2.0444	5.6036
1018	LGG_01066	144	707	152	0.9474	4.6513	-0.0780	2.2176
1019	LGG_01067	2	12	1	2.0000	12.0000	1.0000	3.5850
1020	LGG_01068	150	602	183	0.8197	3.2896	-0.2869	1.7179
1021	LGG_01069	228	940	271	0.8413	3.4686	-0.2493	1.7944
1022	LGG_01070	165	557	196	0.8418	2.8418	-0.2484	1.5068
1023	LGG_01071	92	343	105	0.8762	3.2667	-0.1907	1.7078

W14-1：LGG_01064

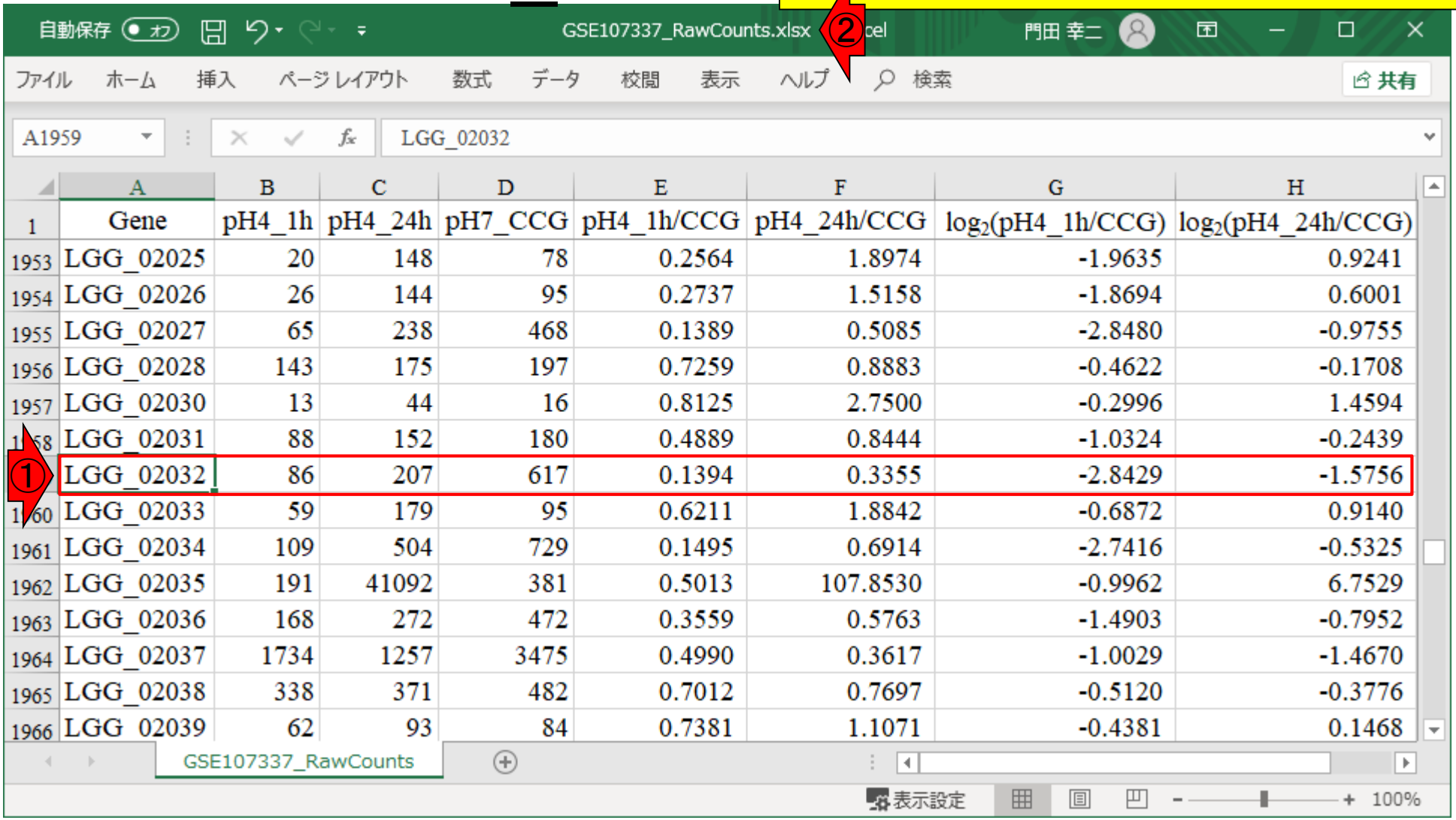
原著論文のp1611の左中央付近で述べられている、①LGG_01064の酸ストレス条件下における発現量上昇を②自分らのデータで確認。



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
1063	LGG_01057	380	825	978	0.3885	0.8436	-1.3638	-0.2454
1064	LGG_01058	46	130	168	0.2738	0.7738	-1.8688	-0.3699
1065	LGG_01059	39	111	74	0.5270	1.5000	-0.9241	0.5850
1066	LGG_01060	35	88	52	0.6731	1.6923	-0.5712	0.7590
1067	LGG_01061	2258	2677	2986	0.7562	0.8965	-0.4032	-0.1576
1068	LGG_01062	1625	5164	3043	0.5340	1.6970	-0.9051	0.7630
1069	LGG_01063	26	229	44	0.5909	5.2045	-0.7590	2.3798
1070	LGG_01064	49	389	18	2.7222	21.6111	1.4448	4.4337
1071	LGG_01065	37	536	12	3.0833	44.6667	1.6245	5.4811
1072	LGG_01066	195	983	209	0.9330	4.7033	-0.1000	2.2337
1073	LGG_01067	2	52	2	1.0000	26.0000	0.0000	4.7004
1074	LGG_01068	233	871	247	0.9433	3.5263	-0.0842	1.8182
1075	LGG_01069	328	1359	355	0.9239	3.8282	-0.1141	1.9367
1076	LGG_01070	202	701	248	0.8145	2.8266	-0.2960	1.4991
1077	LGG_01071	130	445	152	0.8553	2.9276	-0.2256	1.5497

W14-2：LGG_02032

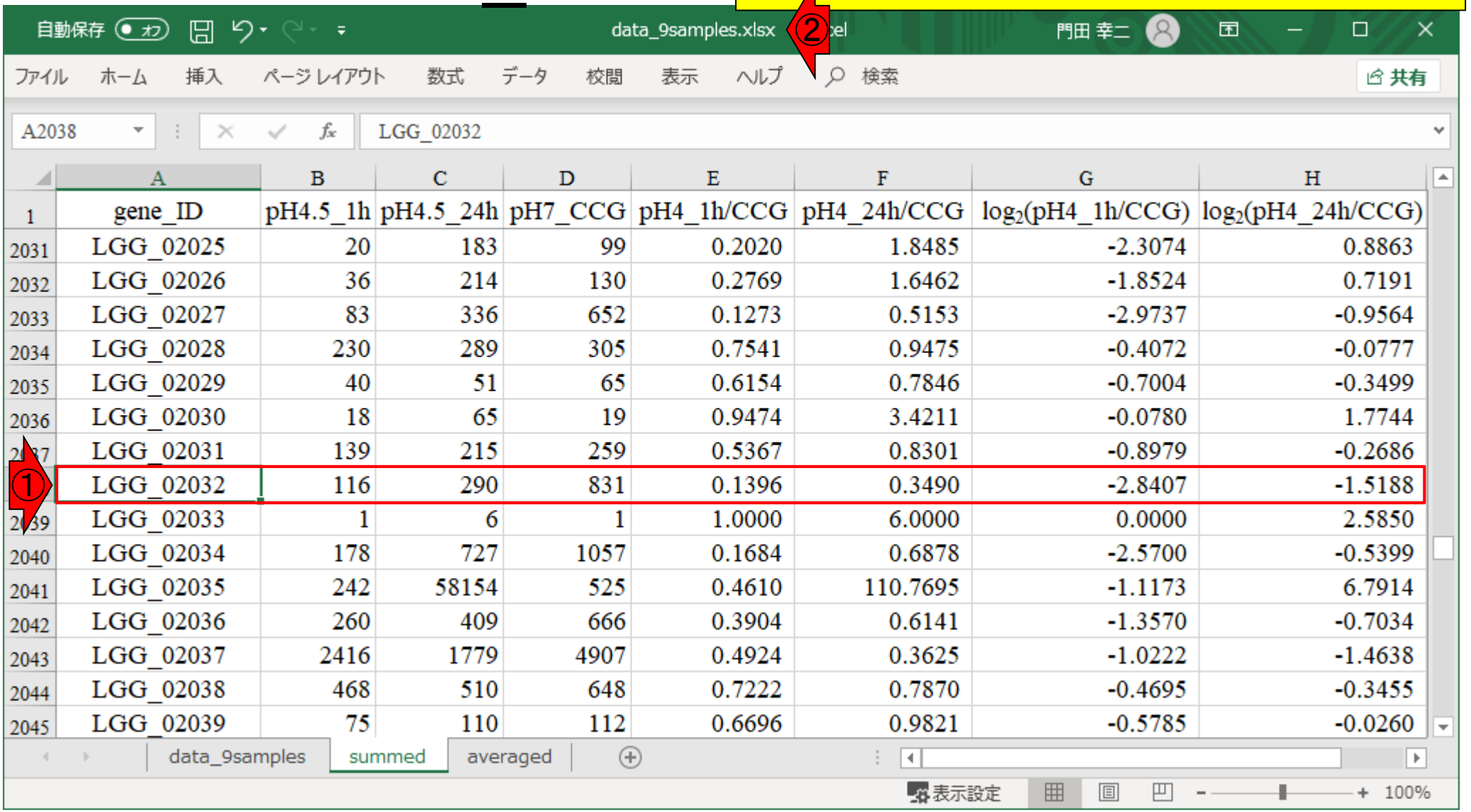
原著論文のp1611の左中央付近で述べられている、①LGG_02032の酸ストレス条件下における発現量低下を②公共データで確認。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
1953	LGG_02025	20	148	78	0.2564	1.8974	-1.9635	0.9241
1954	LGG_02026	26	144	95	0.2737	1.5158	-1.8694	0.6001
1955	LGG_02027	65	238	468	0.1389	0.5085	-2.8480	-0.9755
1956	LGG_02028	143	175	197	0.7259	0.8883	-0.4622	-0.1708
1957	LGG_02030	13	44	16	0.8125	2.7500	-0.2996	1.4594
1958	LGG_02031	88	152	180	0.4889	0.8444	-1.0324	-0.2439
1959	LGG_02032	86	207	617	0.1394	0.3355	-2.8429	-1.5756
1960	LGG_02033	59	179	95	0.6211	1.8842	-0.6872	0.9140
1961	LGG_02034	109	504	729	0.1495	0.6914	-2.7416	-0.5325
1962	LGG_02035	191	41092	381	0.5013	107.8530	-0.9962	6.7529
1963	LGG_02036	168	272	472	0.3559	0.5763	-1.4903	-0.7952
1964	LGG_02037	1734	1257	3475	0.4990	0.3617	-1.0029	-1.4670
1965	LGG_02038	338	371	482	0.7012	0.7697	-0.5120	-0.3776
1966	LGG_02039	62	93	84	0.7381	1.1071	-0.4381	0.1468

W14-2：LGG_02032

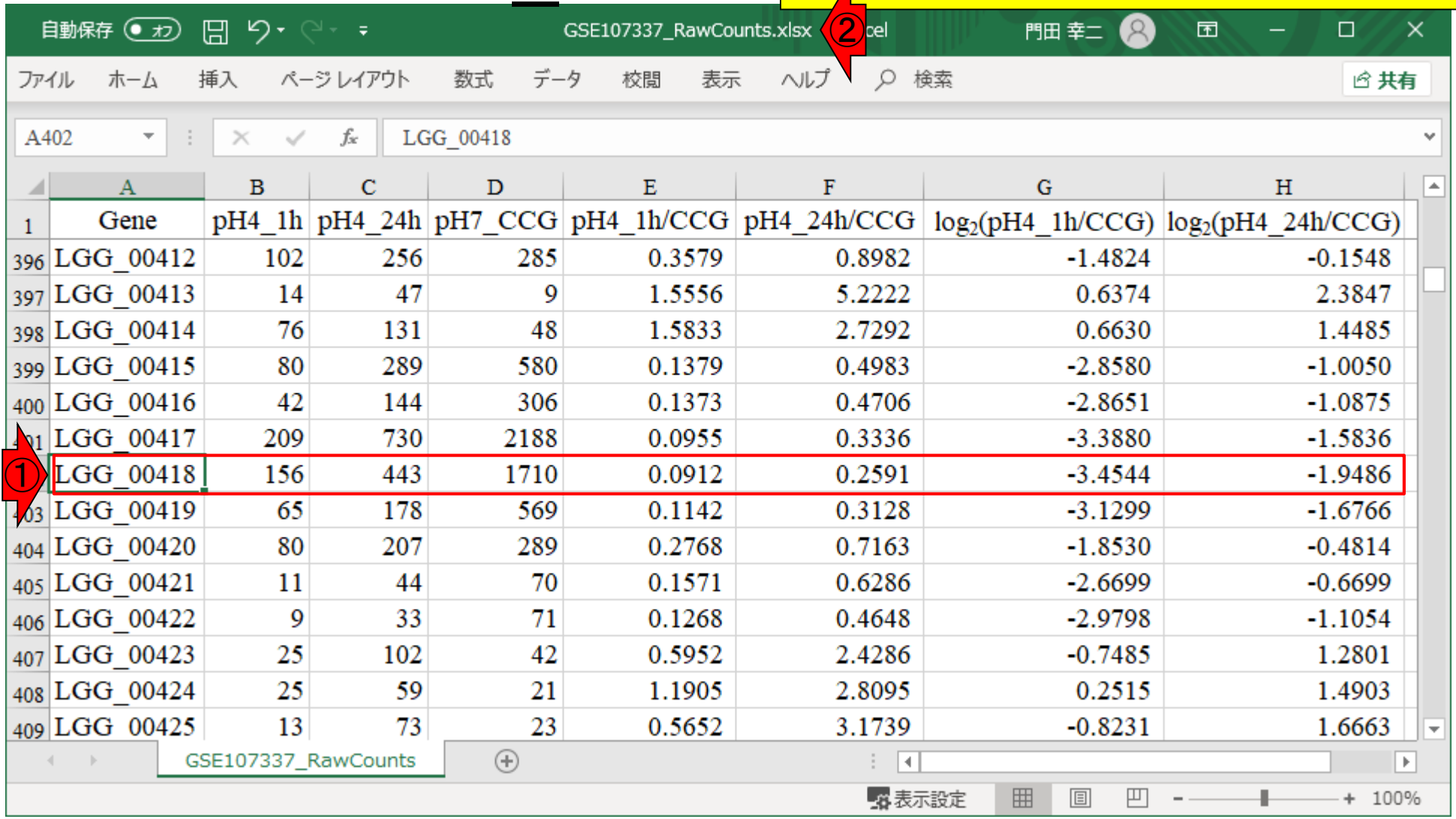
原著論文のp1611の左中央付近で述べられている、①LGG_02032の酸ストレス条件下における発現量低下を②自分らのデータで確認。



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
2031	LGG_02025	20	183	99	0.2020	1.8485	-2.3074	0.8863
2032	LGG_02026	36	214	130	0.2769	1.6462	-1.8524	0.7191
2033	LGG_02027	83	336	652	0.1273	0.5153	-2.9737	-0.9564
2034	LGG_02028	230	289	305	0.7541	0.9475	-0.4072	-0.0777
2035	LGG_02029	40	51	65	0.6154	0.7846	-0.7004	-0.3499
2036	LGG_02030	18	65	19	0.9474	3.4211	-0.0780	1.7744
2037	LGG_02031	139	215	259	0.5367	0.8301	-0.8979	-0.2686
2038	LGG_02032	116	290	831	0.1396	0.3490	-2.8407	-1.5188
2039	LGG_02033	1	6	1	1.0000	6.0000	0.0000	2.5850
2040	LGG_02034	178	727	1057	0.1684	0.6878	-2.5700	-0.5399
2041	LGG_02035	242	58154	525	0.4610	110.7695	-1.1173	6.7914
2042	LGG_02036	260	409	666	0.3904	0.6141	-1.3570	-0.7034
2043	LGG_02037	2416	1779	4907	0.4924	0.3625	-1.0222	-1.4638
2044	LGG_02038	468	510	648	0.7222	0.7870	-0.4695	-0.3455
2045	LGG_02039	75	110	112	0.6696	0.9821	-0.5785	-0.0260

W14-3：LGG_00418

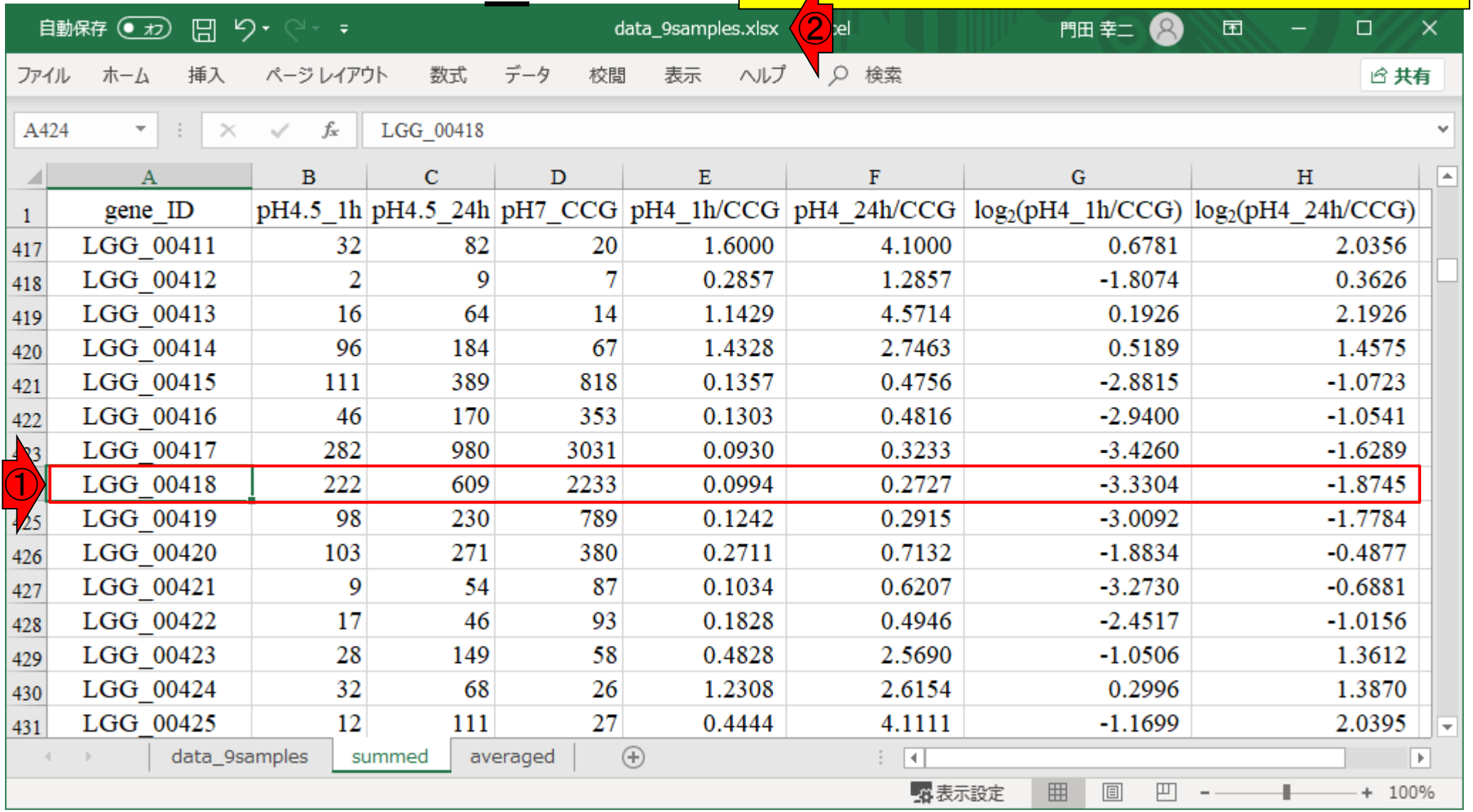
原著論文のp1611の左中央付近で述べられている、①LGG_00418の酸ストレス条件下における発現量低下を②公共データで確認。



	A	B	C	D	E	F	G	H
1	Gene	pH4_1h	pH4_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
396	LGG_00412	102	256	285	0.3579	0.8982	-1.4824	-0.1548
397	LGG_00413	14	47	9	1.5556	5.2222	0.6374	2.3847
398	LGG_00414	76	131	48	1.5833	2.7292	0.6630	1.4485
399	LGG_00415	80	289	580	0.1379	0.4983	-2.8580	-1.0050
400	LGG_00416	42	144	306	0.1373	0.4706	-2.8651	-1.0875
401	LGG_00417	209	730	2188	0.0955	0.3336	-3.3880	-1.5836
402	LGG_00418	156	443	1710	0.0912	0.2591	-3.4544	-1.9486
403	LGG_00419	65	178	569	0.1142	0.3128	-3.1299	-1.6766
404	LGG_00420	80	207	289	0.2768	0.7163	-1.8530	-0.4814
405	LGG_00421	11	44	70	0.1571	0.6286	-2.6699	-0.6699
406	LGG_00422	9	33	71	0.1268	0.4648	-2.9798	-1.1054
407	LGG_00423	25	102	42	0.5952	2.4286	-0.7485	1.2801
408	LGG_00424	25	59	21	1.1905	2.8095	0.2515	1.4903
409	LGG_00425	13	73	23	0.5652	3.1739	-0.8231	1.6663

W14-3：LGG_00418

原著論文のp1611の左中央付近で述べられている、①LGG_00418の酸ストレス条件下における発現量低下を②公共データで確認。



	A	B	C	D	E	F	G	H
1	gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG	pH4_1h/CCG	pH4_24h/CCG	$\log_2$(pH4_1h/CCG)	$\log_2$(pH4_24h/CCG)
417	LGG_00411	32	82	20	1.6000	4.1000	0.6781	2.0356
418	LGG_00412	2	9	7	0.2857	1.2857	-1.8074	0.3626
419	LGG_00413	16	64	14	1.1429	4.5714	0.1926	2.1926
420	LGG_00414	96	184	67	1.4328	2.7463	0.5189	1.4575
421	LGG_00415	111	389	818	0.1357	0.4756	-2.8815	-1.0723
422	LGG_00416	46	170	353	0.1303	0.4816	-2.9400	-1.0541
423	LGG_00417	282	980	3031	0.0930	0.3233	-3.4260	-1.6289
424	LGG_00418	222	609	2233	0.0994	0.2727	-3.3304	-1.8745
425	LGG_00419	98	230	789	0.1242	0.2915	-3.0092	-1.7784
426	LGG_00420	103	271	380	0.2711	0.7132	-1.8834	-0.4877
427	LGG_00421	9	54	87	0.1034	0.6207	-3.2730	-0.6881
428	LGG_00422	17	46	93	0.1828	0.4946	-2.4517	-1.0156
429	LGG_00423	28	149	58	0.4828	2.5690	-1.0506	1.3612
430	LGG_00424	32	68	26	1.2308	2.6154	0.2996	1.3870
431	LGG_00425	12	111	27	0.4444	4.1111	-1.1699	2.0395

Contents

- W1：公共データベースENA
- W2：クラウド解析環境Galaxy
- W3：ENAからGalaxyへのFASTQファイルの送付
- W4：クオリティチェック（FastQC）
- W5：クオリティコントロール（Trimmomatic）
- W6：クオリティチェック（FastQC）
- W7：ゲノム配列へのマッピング（Bowtie2）
- W8：カウント情報取得（htseq-count）
- W9：カウント情報の連結（Column join）
- W10：カウントデータの数値行列の行名について
- W11：公共カウントデータ（GSE107337）を用いて、倍率変化と$\log_2$(倍率変化)を算出
- W12：Galaxy上で得られたカウントデータについても同様に計算し、比較
- W13とW14：原著論文中の記載事項を確認
- W15：2つのカウントデータから共通遺伝子の情報を用いて類似度を計算

同一サンプル間の類似度が最も高いことを確認。

W15-1：data_original.txt

Gene	pH4_1h	pH4_24h	pH7_CCG
LGG_00001	286	676	545
LGG_00002	776	1601	1450
LGG_00003	71	114	91
LGG_00004	201	470	368
LGG_00005	368	1160	1447
LGG_00006	1041	2742	2880
LGG_00007	7	10	2
LGG_00008	73	76	499
LGG_00009	16	45	42
LGG_00010	20	65	57
LGG_00011	2281	2596	2150
LGG_00012	2510	3631	3097
LGG_00013	971	1301	1424
LGG_00014	17	78	16
LGG_00015	82	255	33
LGG_00016	30	149	21
LGG_00017	30	127	20

同一サンプル間の類似度が最も高いことを確認。

W15-2：data_galaxy.txt

gene_ID	pH4.5_1h	pH4.5_24h	pH7_CCG
EBG00001128470	270	302	353
EBG00001128476	28	76	19
EBG00001128500	197	286	180
EBG00001128509	343	375	434
EBG00001128529	0	0	0
LGG_00001	394	897	720
LGG_00002	1039	2234	2016
LGG_00003	51	121	68
LGG_00004	248	634	489
LGG_00005	451	1543	1950
LGG_00006	1469	3882	4038
LGG_00007	9	16	4
LGG_00008	102	100	686
LGG_00009	25	75	72
LGG_00010	21	85	73
LGG_00011	2536	3127	2407
LGG_00012	2667	4325	3536

同一サンプル間の類似度が最も高いことを確認。

W15-3：類似度計算

RGui (64-bit)

ファイル　編集　閲覧　その他　パッケージ　ウインドウ　ヘルプ　Vignettes

R Console

```
> data_ori <- read.table("data_original.txt", header=TRUE, sep="\t", row.names=1$
> data_gal <- read.table("data_galaxy.txt", header=TRUE, sep="\t", row.names=1) $
> dim(data_ori)                                  #行数と列数を表示
[1] 2838    3
> dim(data_gal)                                  #行数と列数を表示
[1] 2949    3
> common <- intersect(rownames(data_ori), rownames(data_gal))#共通した行名情報を$
> length(common)                                 #要素数(遺伝子数)を表示
[1] 2838
> data <- cbind(data_ori[common, ], data_gal[common, ])#列方向で結合した結果をda$
> cor(data, method="spearman")                   #総当たりのSpearman相関係数を計算
             pH4_1h   pH4_24h   pH7_CCG  pH4.5_1h pH4.5_24h   pH7_CCG
pH4_1h     1.0000000 0.9023052 0.8919116 0.9544823 0.8670647 0.8617587
pH4_24h    0.9023052 1.0000000 0.8183148 0.8583811 0.9544563 0.7889634
pH7_CCG    0.8919116 0.8183148 1.0000000 0.8533259 0.7849992 0.9662319
pH4.5_1h   0.9544823 0.8583811 0.8533259 1.0000000 0.9059912 0.8948864
pH4.5_24h  0.8670647 0.9544563 0.7849992 0.9059912 1.0000000 0.8279148
pH7_CCG    0.8617587 0.7889634 0.9662319 0.8948864 0.8279148 1.0000000
> |
```

W15-4：総カウント数

公開カウントデータのほうで、総カウント数の最大と最小の間には、①2.18倍もの差がある！

RGui (64-bit)

ファイル　編集　閲覧　その他　パッケージ　ウインドウ　ヘルプ　Vignettes

R Console

```
> dim(data_ori)                                  #行数と列数を表示
[1] 2838    3
> dim(data_gal)                                  #行数と列数を表示
[1] 2949    3
> common <- intersect(rownames(data_ori), rownames(data_gal))#共通した行名情報を$
> length(common)                                 #要素数(遺伝子数)を表示
[1] 2838
> data <- cbind(data_ori[common, ], data_gal[common, ])#列方向で結合した結果をda$
> cor(data, method="spearman")                   #総当たりのSpearman相関係数を計算
             pH4_1h    pH4_24h    pH7_CCG  pH4.5_1h pH4.5_24h    pH7_CCG
pH4_1h     1.0000000 0.9023052 0.8919116 0.9544823 0.8670647 0.8617587
pH4_24h    0.9023052 1.0000000 0.8183148 0.8583811 0.9544563 0.7889634
pH7_CCG    0.8919116 0.8183148 1.0000000 0.8533259 0.7849992 0.9662319
pH4.5_1h   0.9544823 0.8583811 0.8533259 1.0000000 0.9059912 0.8948864
pH4.5_24h  0.8670647 0.9544563 0.7849992 0.9059912 1.0000000 0.8279148
pH7_CCG    0.8617587 0.7889634 0.9662319 0.8948864 0.8279148 1.0000000
> colSums(data_ori)
 pH4_1h pH4_24h pH7_CCG
 774376 1670016 1690218
> max(colSums(data_ori))/min(colSums(data_ori))
[1] 2.182684
>
```

①