Using Galaxy to analyze RNA-Seq data

Yating Liu

07/2017

# 1. Introduction

The goal of this tutorial is to provide an introduction on how to use Galaxy to analyze RNA-Seq data. This tutorial assumes that the reader has some basic familiarity with Galaxy [e.g*.*, has completed the Galaxy 101 tutorial (<https://galaxyproject.org/tutorials/g101/>)]. From this tutorial, you will learn how to:

* Get data from the Data Library
* Prepare dataset collections
* Find and use Galaxy tools to analyze RNA-Seq data
* Extract your analysis workflow and run it on a different sample

# 2. Register and login

## 2.1 Create an account

First of all, you need a Galaxy account in order to access the full functionality of Galaxy. Open a web browser and navigate to the G-OnRamp instance at <http://cloud5.galaxyproject.org>. (Note that this instance will only be available during the G-OnRamp workshop.)If you have already created an account on the G-OnRamp instance, you can skip this section and jump to section 2.2 Login.

To create an account, click on “User” at the menu bar and then click on “Register” (Figure 1).

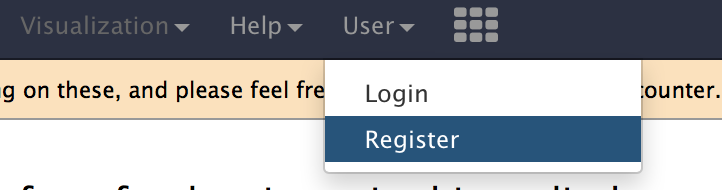


Figure 1: Create a user account on the G-OnRamp Galaxy instance.

The workspace will display the registration form (Figure 2). In addition to entering your email address and your password, you will also need to specify a public name. The public name is an identifier that will be assigned to all the Galaxy resources (e.g*.*, Histories, Workflows) that you share with other users. The public name must be at least three characters in length and contains only lower-case letters, numbers, and the '-' character. Click on the “Submit” button to create your account. If the new account has been created successfully, you will see a green message box indicating that you are now logged into your new account. You can click on “Return to the home page” link to return to the home page (Figure 3).

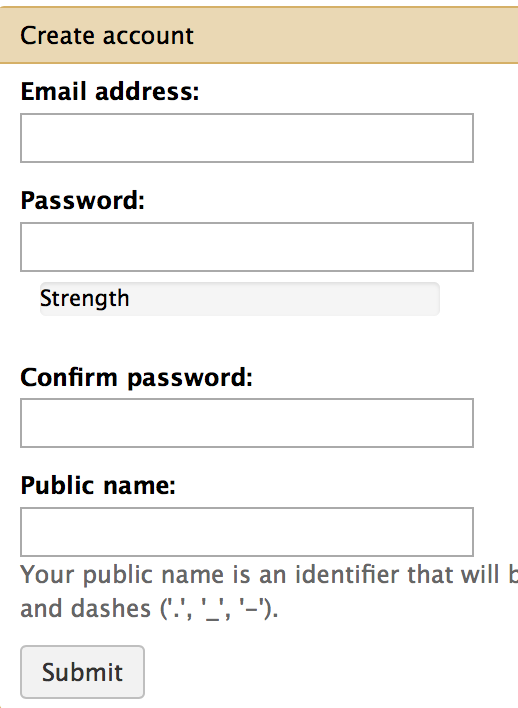


Figure 2: Registration form for creating a new Galaxy account.

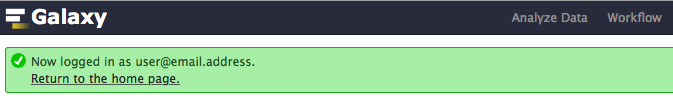


Figure 3: After you click on the “Submit” button, a green message box will appear indicating that you are now logged into your account. Click on the "Return to the home page" link to go back to the home page.

## 2.2 Login

When you return to G-OnRamp later, you can log in with your email address and password. First click on the “User” menu item at the menu bar, and then choose “Login” (Figure 4).

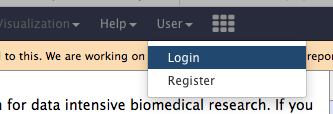


Figure 4: Open the login form from the “User” menu item at the menu bar.

Then enter your email address and password (Figure 5).

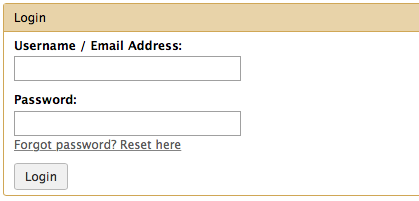


Figure 5: Enter the email address you have previously used to create your Galaxy account and your account password to log into Galaxy.

# 3. Get data

The RNA-Seq data that we will use in this tutorial are from the “Discovering and quantifying new transcripts” tutorial available through the “Learn Galaxy” page of the Galaxy web site (<https://galaxyproject.org/learn/>). These RNA-Seq datasets were produced as part of a study of the transcription factor TAL1 in mouse by Wu W *et al.* (DOI:10.1101/gr.164830.113). For training purposes, the RNA-Seq datasets were down-sampled to only include reads that mapped to chromosome 19 and to loci with relevance to hematopoiesis (i.e. creation of new blood cells) identified by Wu and colleagues. The genomic datasets associated with this study is available for download through the NCBI Gene Expression Omnibus (GEO) database under the accession number GSE51338.

To reduce the amount of time required to analyze these RNA-Seq datasets, we have further down-sampled them to only 1% of their original size. The datasets that we will use in this walkthrough are available through the Data Libraries. Click on the “Shared Data” menu item on the top menu bar and then select “Data Libraries” in the drop-down menu to navigate to Data Libraries (Figure 6). Click on the “RNA-Seq reads” folder. You should see eight test datasets in this folder. The data is structured in the following way:

* There are two samples: G1E cells and megakaryocytes (G1E and Mk)
* There are two biological replicates per sample (R1 and R2)
* Each replicate has forward and reverse reads (forward and reverse)

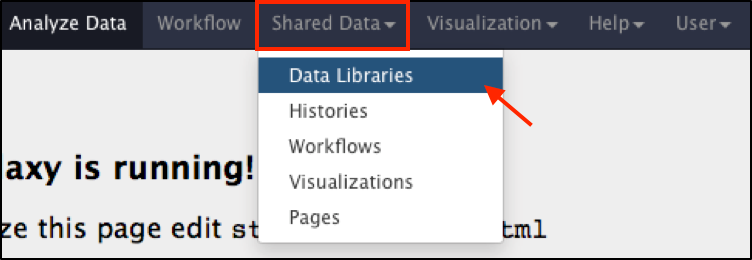


Figure 6: Click on the "Shared Data" menu item on the top menu bar and then click on "Data Libraries" in the drop-down menu.

Select all the datasets by selecting the checkbox before the “name” column header. Click on the “to History” button to copy all the files to a History (Figure 7). You will see a pop-up window titled “Import into History”. You can select a History from the dropdown menu (purple arrow) and then click on the “Import” button to import the datasets into the selected History (Figure 8). If you want to import data into a new History, type the name of the new History into the box next to “or create new” and then click on the “Import” button.

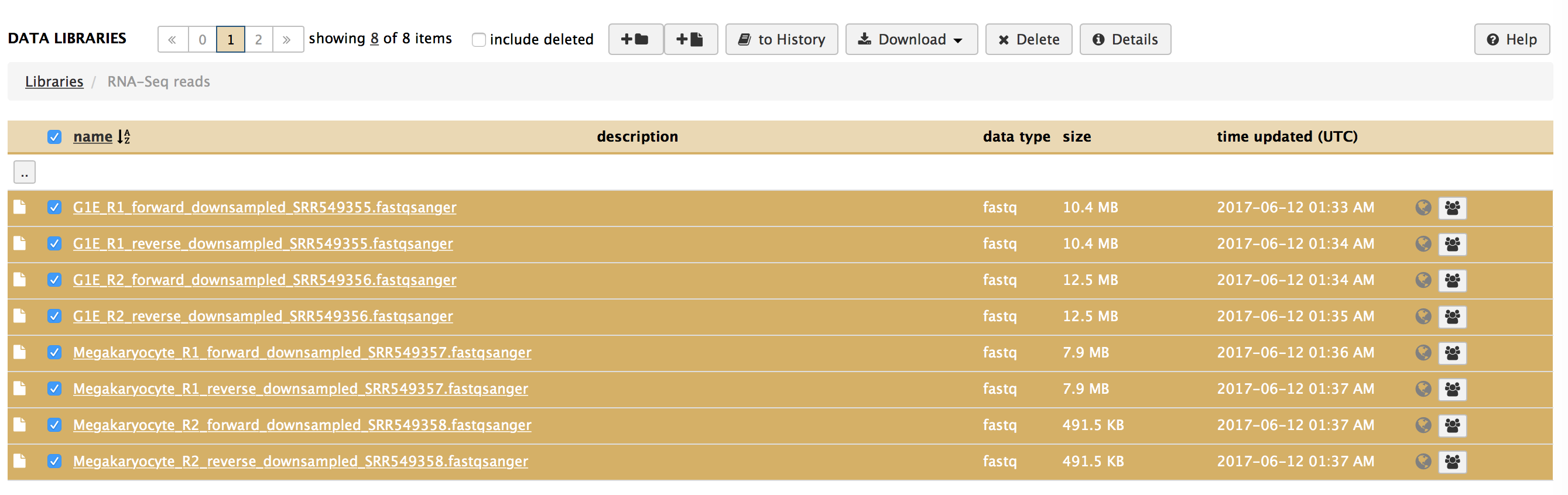


Figure 7: Select the checkbox next to the “name” column header to select all the datasets. Alternatively, you can select the checkbox next to each dataset individually.

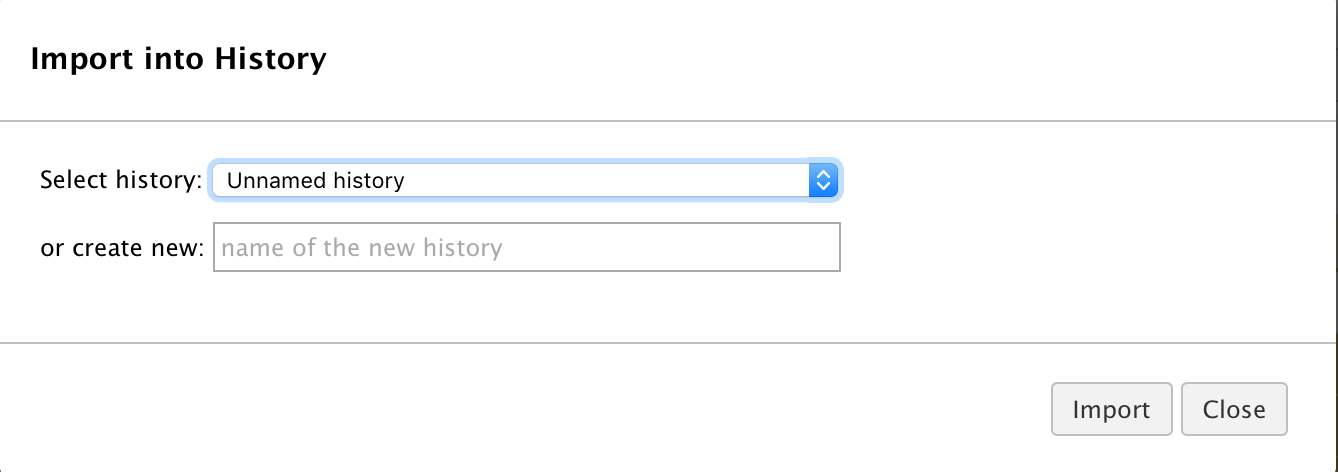


Figure 8: After clicking on the “to History” button, you will see the "Import into History" window. Click on the "Import" button to import the selected datasets into the current History called “Unnamed history”.

A green message box will appear once the datasets have been added to the current History. After you see the green message box, click on the “Analyze Data” menu item on the top menu bar to navigate back to Home page. The eight RNA-Seq datasets will appear in the History panel.

There are multiple quality encodings for the fastq file format [see the “Encoding” section of the “FASTQ format” page on Wikipedia (<https://en.wikipedia.org/wiki/FASTQ_format#Encoding>) for details]. The quality scores in the RNA-Seq read files we have just imported are encoded using the Sanger format (i.e. the **fastqsanger** datatype). However, the auto-detect functionality in Galaxy might assign the **fastq** datatype to these files instead. Because tools such as Trimmomatic (which we will use later in this walkthrough) can only process files that are in the fastqsanger datatype, we need to change the datatype of the imported RNA-Seq read files.

Click on the pencil icon of each file to open the data attributes editor. Click on the “Datatype” tab and select **fastqsanger** (be carefully not to select fastqcsanger) in the drop-down menu for the “New Type” (Figure 9). Click on the “Save” button to save the change in datatype.

Note that changing the datatype does not convert the data within a dataset. Consequently, if the fastq quality scores were encoded using a different scheme (e.g*.*, Illumina 1.3), then you will need to use tools such as FASTQ Groomer (available on the Galaxy Main server) to convert the quality encoding of the RNA-Seq read files to Sanger format.

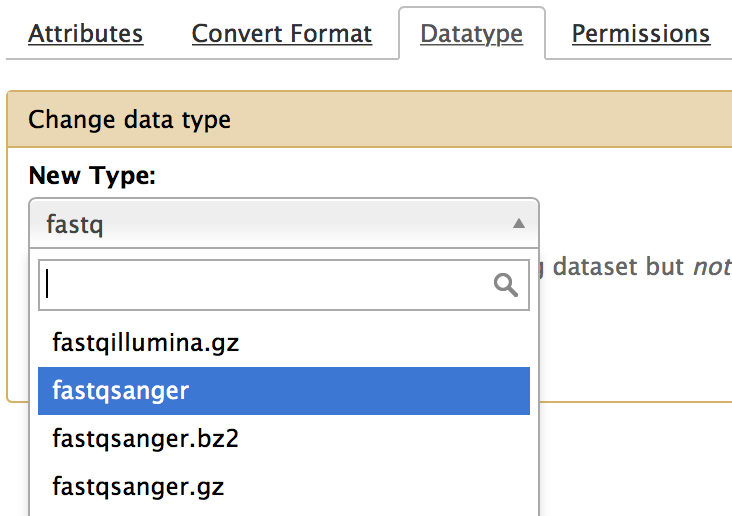


Figure 9: Click on the "Datatype" tab, and select "**fastqsanger**" in the drop-down menu as the new datatype. Then click on the "Save" button.

# 4. Preparing dataset collections

Because we have two biological replicates for each sample and two sets of reads for each replicate, we would like to run Galaxy tools on all the datasets for each sample at once. We can use the dataset collection feature in Galaxy to group multiple datasets together. In this case, we will create two dataset collections: one containing the datasets for G1E and the other containing the datasets for Mk.

We will create the G1E collection first. Type “G1E” in the “search datasets” search box in the History panel and press Enter. The four RNA-Seq datasets from the G1E sample will be listed in the History panel (Figure 10). Click on the checkbox icon on the History panel and then click on the “All” button to select the four files currently listed in the History (Figure 11). Click on the “For all selected…” button and select “Build List of Dataset Pairs” in the drop-down menu (Figure 12). This will open the “Create a collection of paired datasets” window where we can specify how to pair the RNA-Seq datasets in the G1E sample.

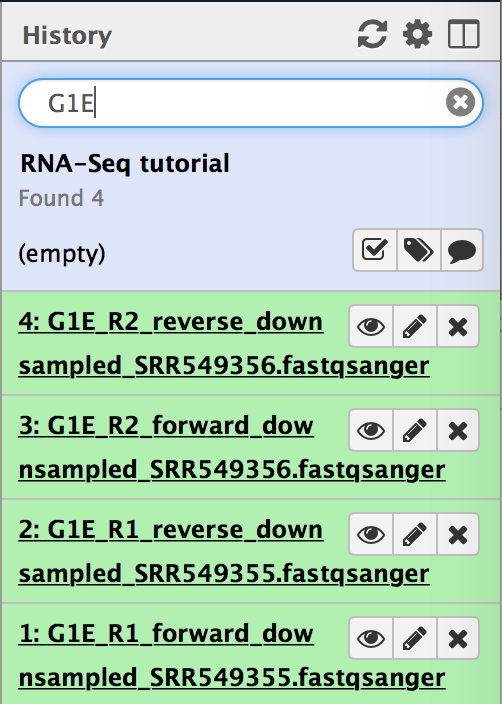


Figure 10: Find all of the G1E datasets from the History by typing "G1E" in the search box and then press "Enter"

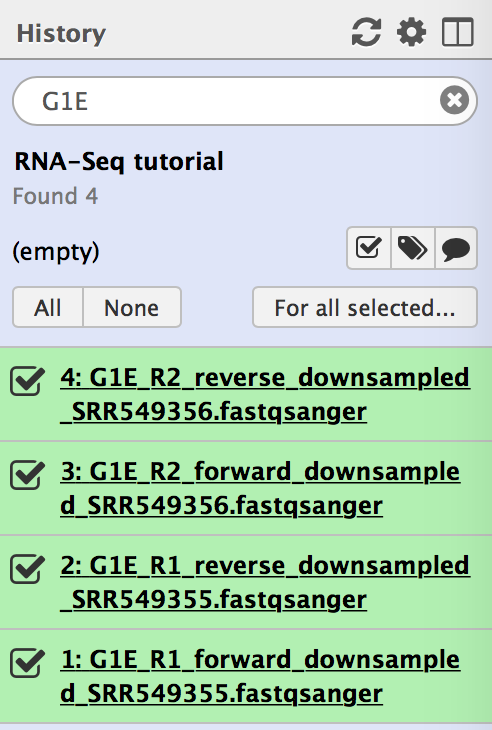


Figure 11: Click on the checkbox icon (purple arrow) on the History panel and then click on the “All” button (red arrow) to select all of the RNA-Seq datasets for the G1E sample.

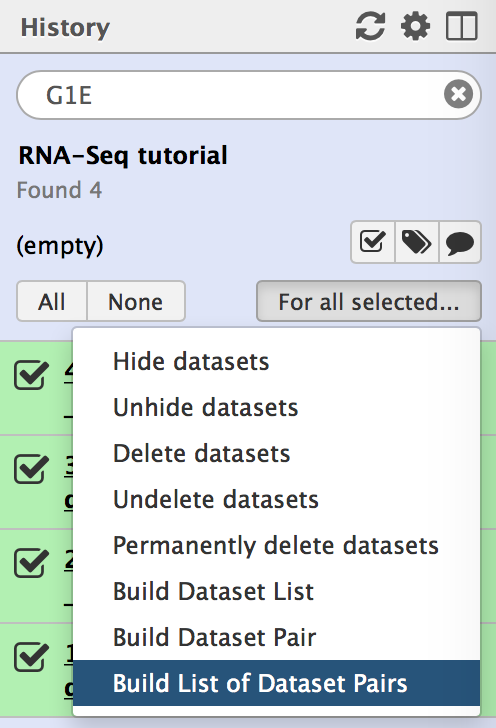


Figure 12: After selecting all the G1E datasets, click on the "For all selected" button and select the "Build List of Dataset Pairs" option to start the collection builder.

We will need to specify a filter in the two text boxes in order to produce the list of forward and reverse datasets, respectively. In the left search box, enter “\_forward” so that only the forward datasets are shown in the list below the search box. In the right search box, enter “\_reverse” so that only the reverse datasets are shown in the list below. Click on the “Pair these datasets” button to pair the forward reads with the reverse reads from each replicate (Figure 13).

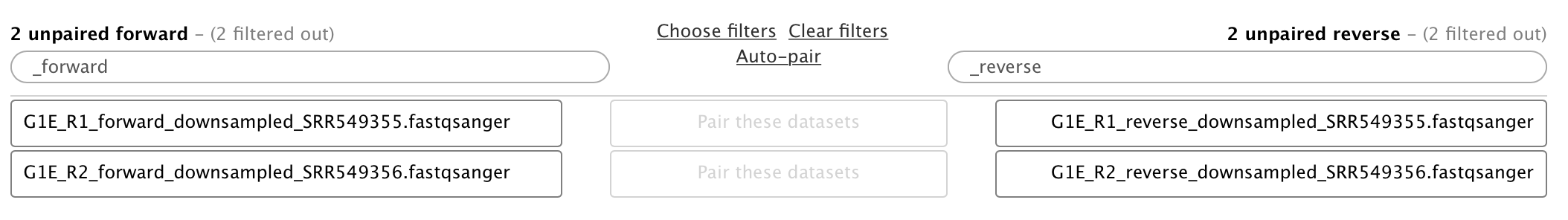


Figure 13: Enter "\_forward" and "\_reverse" in the two text boxes to create a list of the forward and reverse datasets. Then use the "Pair these datasets" button to pair the forward and reverse dataset from each biological replicate.

After all the datasets have been successfully paired (Figure 14), scroll down to the bottom of the window. Enter “G1E” into the Name field and then click on the “Create list” button (Figure 15). You will get a new item in the History representing the G1E dataset collection. Clear search term “G1E” in the search box and press enter in order to see all the datasets. Repeat these steps to create the Mk dataset collection.

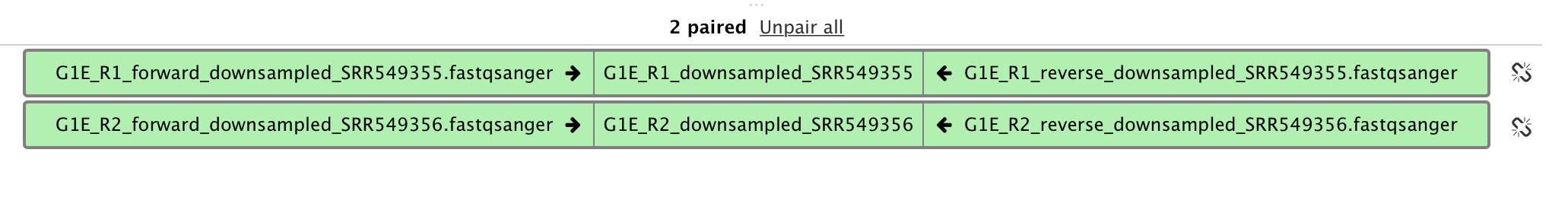


Figure 14: After the reads are paired, you will see two paired datasets listed in the middle column (i.e. G1E\_R1\_downsampled\_SRR549355 and G1E\_R2\_downsampled\_SRR549356).

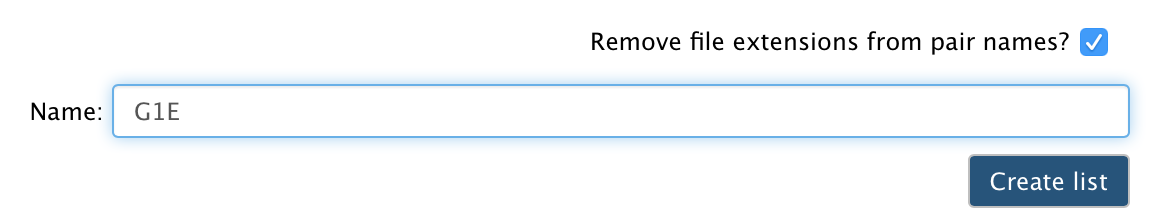


Figure 15: Type "G1E" into the Name field and click on "Create list".

# 5. Pre-processing

## 5.1 Assess the quality of the RNA-Seq reads with FastQC

The first step of the analysis is to evaluate the quality of the RNA-Seq reads using **FastQC**. **FastQC** provides a simple way to do some quality control checks on sequence data from second and third generation high throughput sequencers.

Search for “FastQC” in the Tools panel. Click on the “FastQC” link to open the tool (Figure 16).

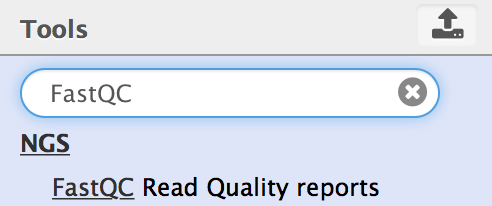


Figure 16: Search for the "FastQC" tool in the Tools panel, and open the tool by clicking on the "FastQC" link.

Click on the folder icon under the “Short read data from your current History” field to use a dataset collection as input, then select the G1E dataset collection in the drop-down menu. Click on the “Execute” button to launch FastQC (Figure 17).

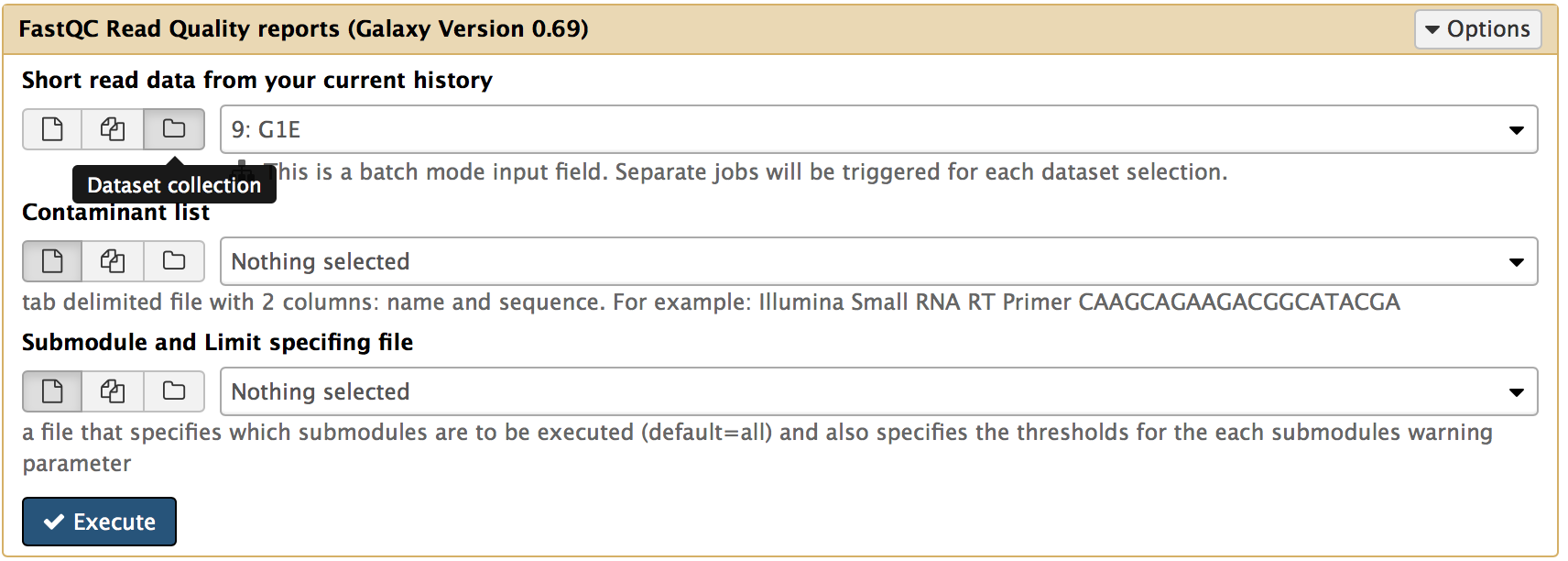


Figure 17: Click on the folder icon in order to select a dataset collection, then select the "G1E" dataset collection from the drop-down menu. Click on the "Execute" button to run the tool.

Two new History items will be created after the FastQC analysis is complete (Figure 18). Each item includes the FastQC reports for all the datasets in the G1E dataset collection. History item 19 contains the FastQC reports in Web page (HTML) format, while History item 20 contains the FastQC reports in plain text format.

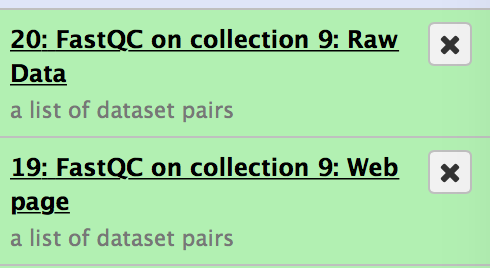


Figure 18: FastQC will create two new items in the History. Each one contains the FastQC results for all datasets within the collection (in Web page and plain text formats, respectively).

Click on History item 19, you will see the links to the two biological replicates of G1E (i.e. G1E\_R1\_downsampled\_SRR549355 and G1E\_R2\_downsampled\_SRR549356). Click on the link for the G1E\_R1\_downsampled\_SRR549355 replicate, and you can see the links to the FastQC reports for the forward and reverse reads, respectively. Click on the eye icon to view the reports (Figure 19). You will see a “Summary” table of FastQC report. Scroll down to view the “Per base sequence quality” report (Figure 20).

You can view the reports for G1E\_R2\_downsampled\_SRR549356 replicate by clicking on the “Back to FastQC on collection 9: Webpage” link at the top of the History panel, and then select the G1E\_R2\_downsampled\_SRR549356 replicate.

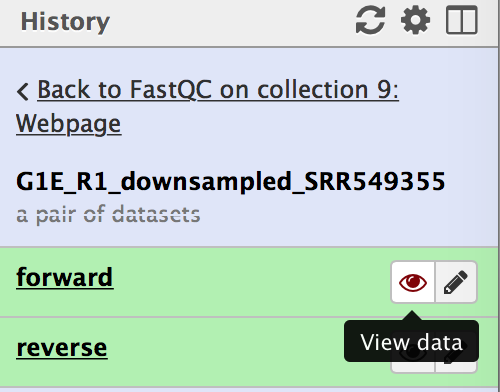


Figure 19: Click on the eye icon to view the FastQC report for each dataset.

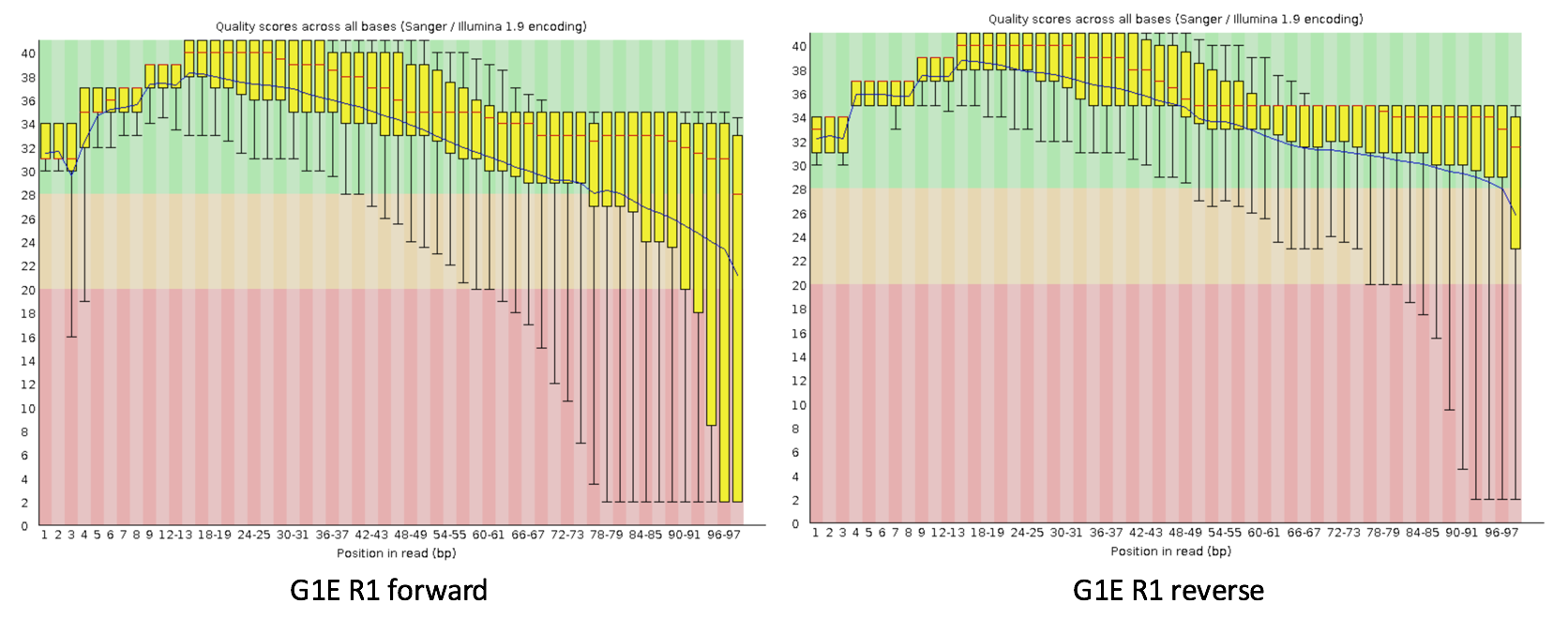


Figure 20: The FastQC Per base sequence quality reports for the G1E R1 forward (left) and reverse (right) reads.

## 5.1. Remove low quality bases from RNA-Seq reads using Trimmomatic

To increase the efficacy of read mapping, we can use **Trimmomatic** to remove low-quality bases from the ends of the reads. Search for “Trimmomatic” in the Tools panel, and then open the tool. Select the “Paired-end (as collection)” option from a drop-down menu under the “Single-end or paired-end reads?” field. Then select “G1E” as the FASTQ dataset collection (Figure 21). Execute the tool to remove the low-quality bases in all of the reads within the G1E dataset collection.

Trimmomatic will produce two dataset collections. The collection with the “paired” suffix contains the subset of reads where both the forward and reverse reads pass the quality filter. The collection with the “unpaired” suffix contains the reads where only one of the paired-end reads pass the quality filter. We will use the “paired” collection in the rest of this walkthrough.

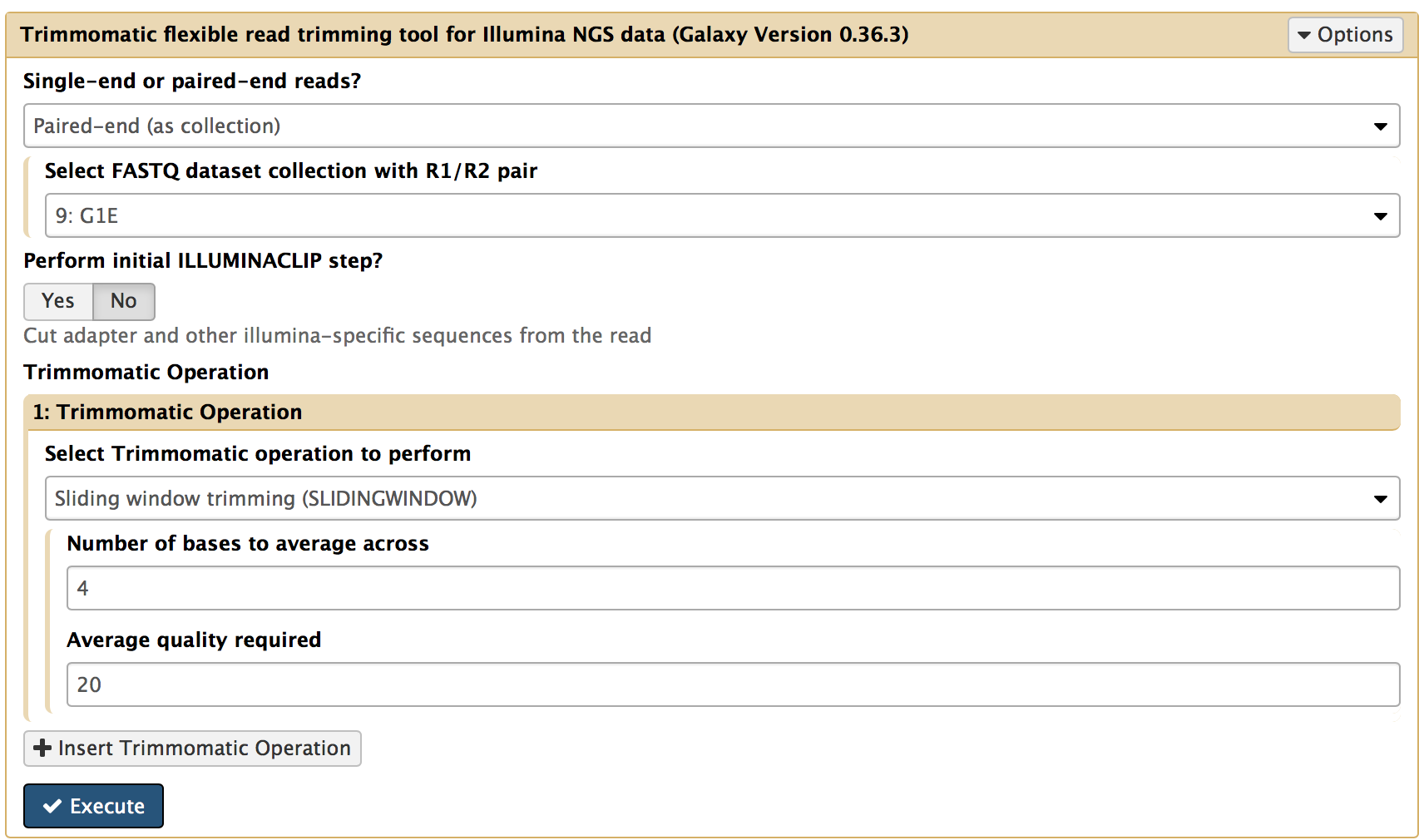


Figure 21: Select the "Paired-end (as collection)" option and then select the "G1E" dataset collection. Click on "Execute" to launch Trimmomatic.

# 6. RNA-Seq read mapping using HISAT2

Next, we would like to map the RNA-Seq reads to the reference (mouse) genome. **HISAT2** is an accurate and fast tool for mapping second generation sequencing reads to a genome. Enter “HISAT2” into the search box in the Tools panel and open the tool.

Select the “Collection of paired reads” option under the “Single end or paired reads?” field. Select the trimmed paired reads (i.e. “Trimmomatic on collection 9: paired”) under the “Paired reads field. We will use the built-in mouse genome assembly “Mouse (mm10)” as the reference genome. Scroll down to the bottom and execute the HISAT2 alignment program (Figure 22).

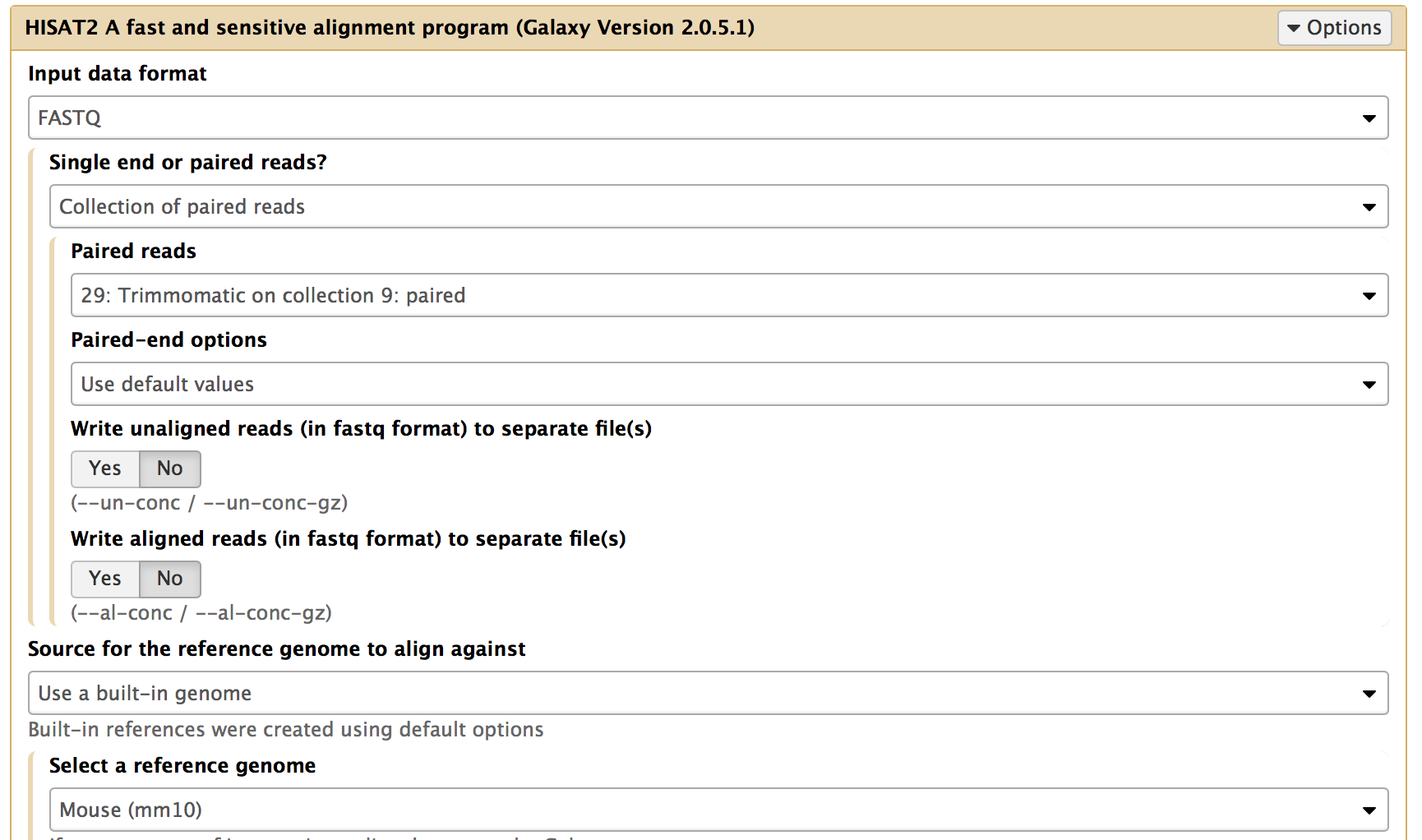


Figure 22: Configure HISAT2 to use the trimmed paired reads from Trimmomatic and the built-in “Mouse (mm10)” reference genome.

# 7. Post-processing

The HISAT2 tool produces BAM files that describe the alignments between the RNA-Seq reads and the mouse genome assembly. We can filter these BAM files to only retain the high-quality alignments using the **Filter SAM or BAM** tool.

Search for the “Filter SAM or BAM” tool in the Tools panel and open the tool. Click on the folder icon under the “SAM or BAM file to filter” field in order to select the “HISAT2 on collection 29” dataset collection as the input for this tool. Use “20” for the “Minimum MAPQ quality score” field. Select “yes” to enable the “Filter on bitwise flag”. Check the first two options to only keep the reads which are paired and are mapped in a proper pair. Finally, click on the “Execute” button to create the clean BAM files (Figure 23).

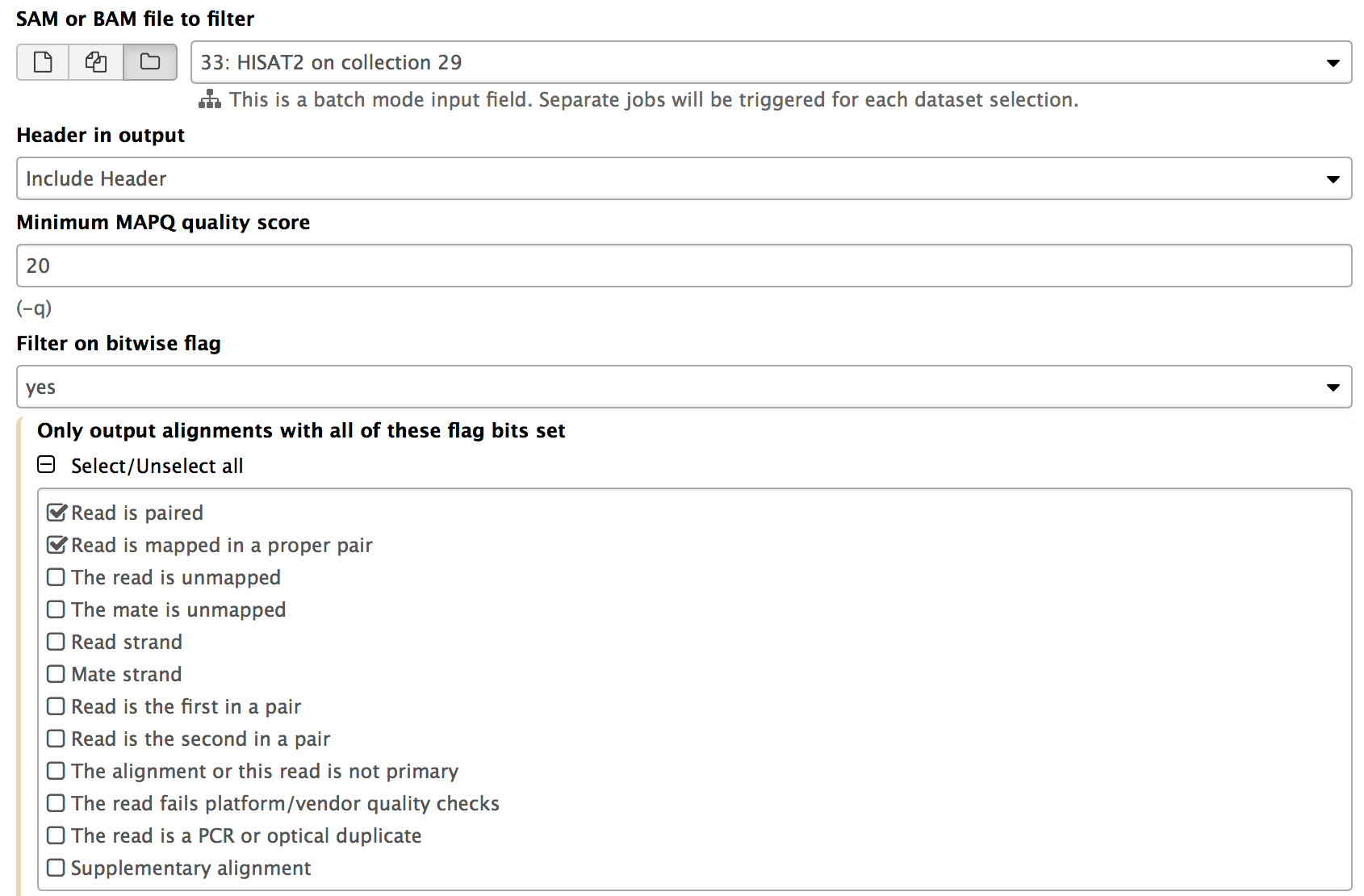


Figure 23: Click on the folder icon and select the HISAT2 collection as the “SAM or BAM file to filter”. Set the minimum MAPQ quality score to “20”. Select “yes” under the “Filter on bitwise flag” field and then specify the filter criteria (i.e. read is paired and is mapped in a proper pair).

# 8. Visualization on UCSC Genome Browser

Click on the new History item created by the “Filter SAM or BAM”tool**.** You shouldsee two filtered BAM files. Click on the link to one of files to expand the item. You can view the BAM alignment on the UCSC Genome Browser by clicking on the “main” link. Because the RNA-Seq dataset only includes reads that mapped to chromosome 19, you need to enter “chr19” into the search box and then click on the “go” button to navigate to chromosome 19 of the mouse genome assembly (Figure 24). Now you can view the alignment tracks on the UCSC Genome Browser (Figure 25).

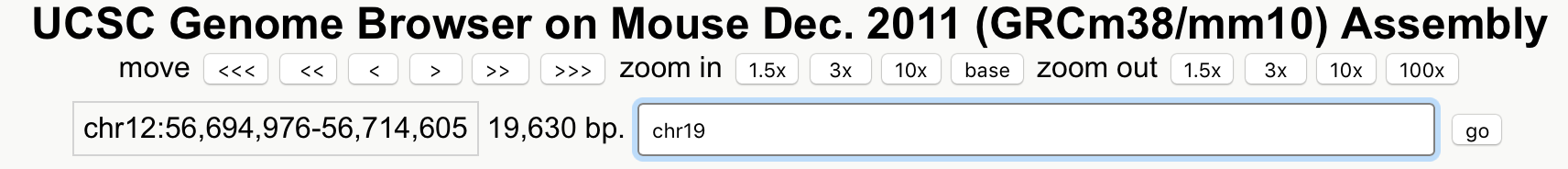


Figure : To navigate to mouse chromosome 19, enter "chr19" into the search box and then click on the "go" button.

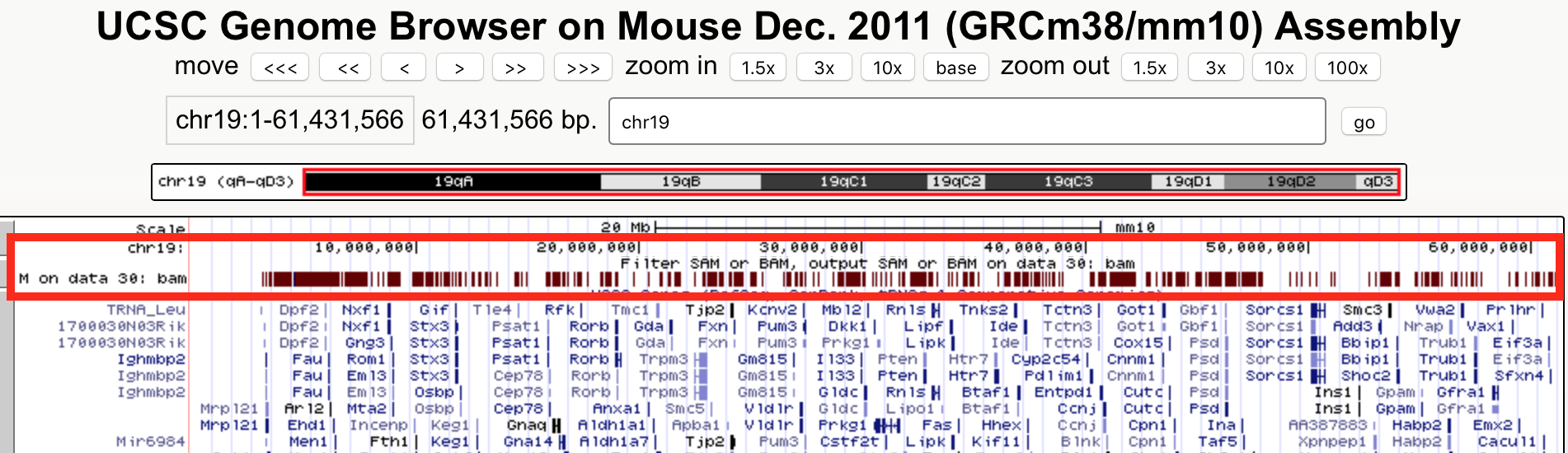


Figure : View the RNA-Seq BAM alignments for the G1E sample on the UCSC Genome Browser. The red rectangle in the screenshot highlights the custom track produced by the “Filter SAM or BAM”tool.

# 9. Extract the workflow and run it on the Mk sample

You can repeat the entire analysis procedure for the Mk sample. The simplest and more reusable solution is to create a workflow from the History of our analysis of the G1E sample.

Click on the gear (setting) icon on the History panel and select “Extract Workflow” from the drop-down menu (Figure 26). Uncheck the History items for the eight datasets with the RNA-Seq reads from each replicate, and the Mk dataset collection. **Make sure that the G1E dataset collection is still selected** (Figure 27). You can change the name of the workflow to “RNA-Seq mapping and data processing” under the “Workflow name” field (Figure 28).

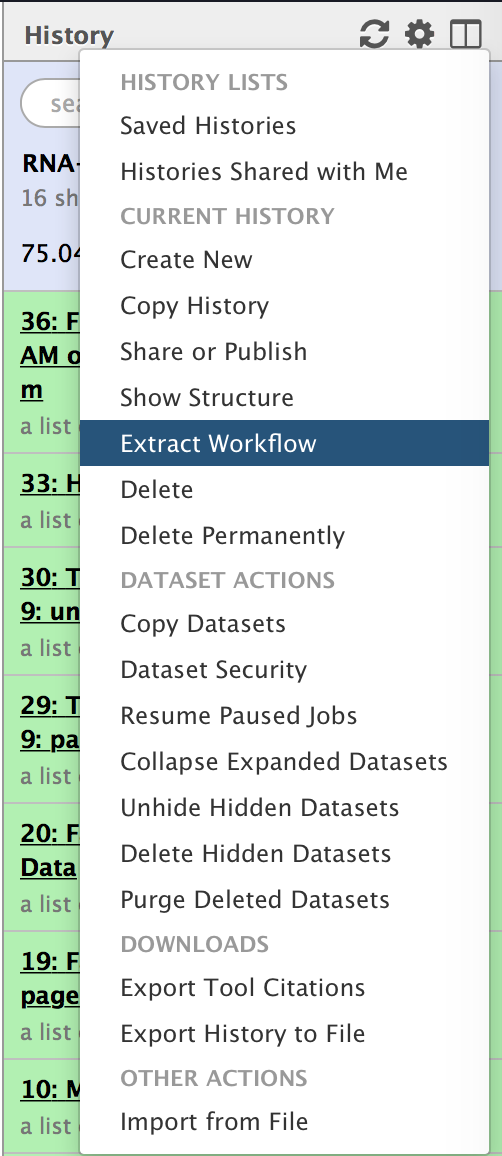


Figure 26: Click on the gear icon and then select the “Extract Workflow” option in the drop-down menu to create a workflow from History.

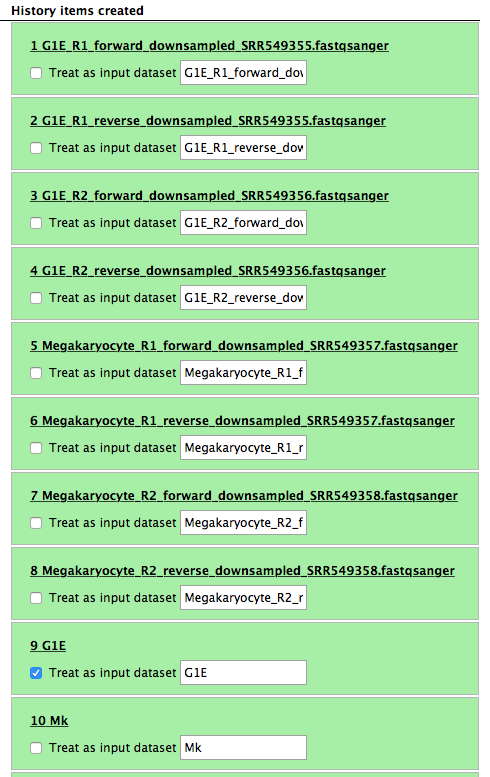


Figure 27: Unselect the eight RNA-Seq datasets and the Mk dataset collection. Leave the G1E dataset collection checked.

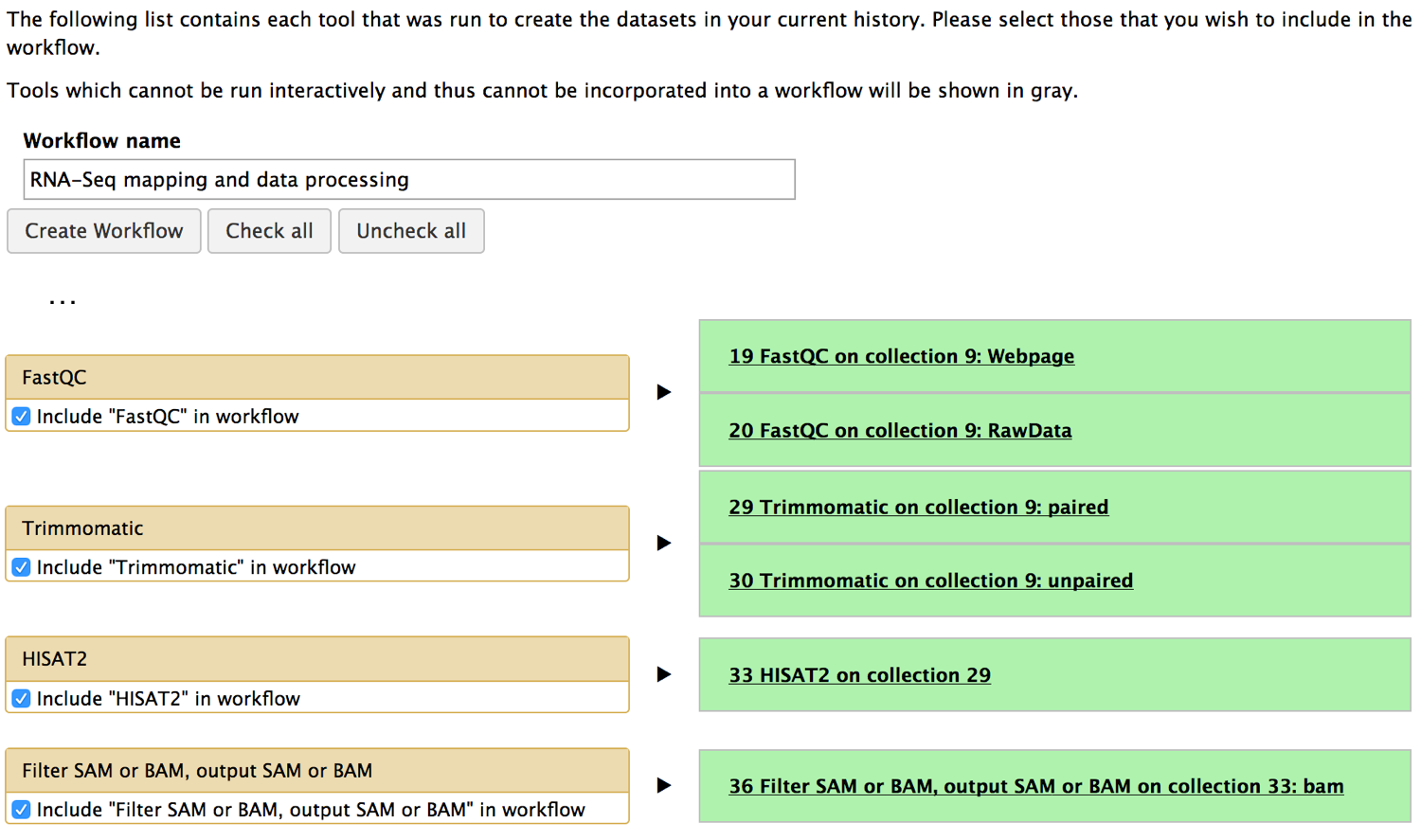


Figure 28: Rename the workflow to “RNA-Seq mapping and data processing” under the “Workflow name” field.

After you click on the “Create Workflow” button, you will see the message: Workflow “RNA-Seq mapping and data processing” created from current History. You can then edit or run the workflow. Let’s click on “edit” to see the workflow in the Workflow Canvas. You can verify that the workflow is correct by comparing it with the workflow shown in Figure 29. In particular, note the folder icons within each tool and the thicker connectors in the workflow. These symbols indicate that the workflow will operate on a dataset collection instead of an individual dataset. Run the new workflow on the Mk dataset collection.

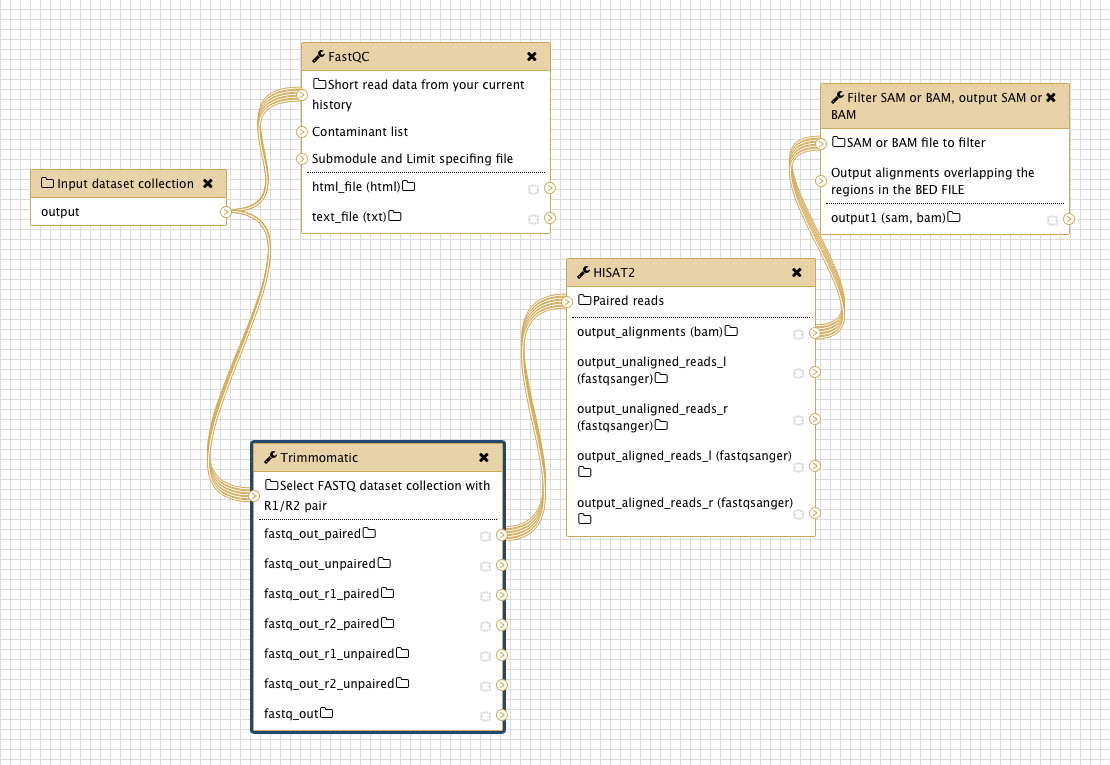


Figure 29: Open “RNA-Seq mapping and data processing” workflow in the Workflow Canvas. Make sure the input dataset collection is connected with the FastQC and Trimmomatic tools.