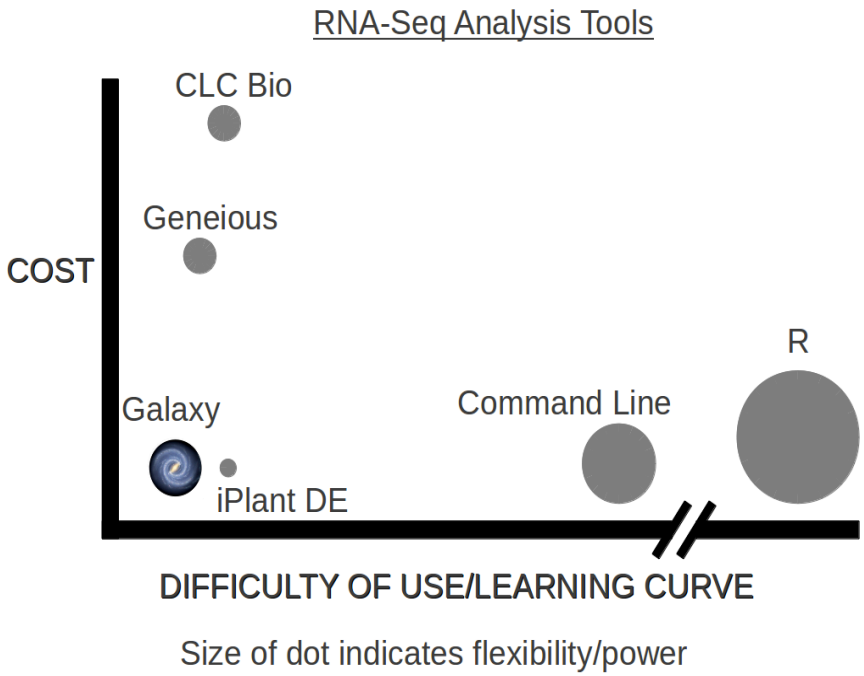
GALAXY
INTRODUCTION AND TUTORIAL

Harvard Medical School
Moazed Lab

Introduction

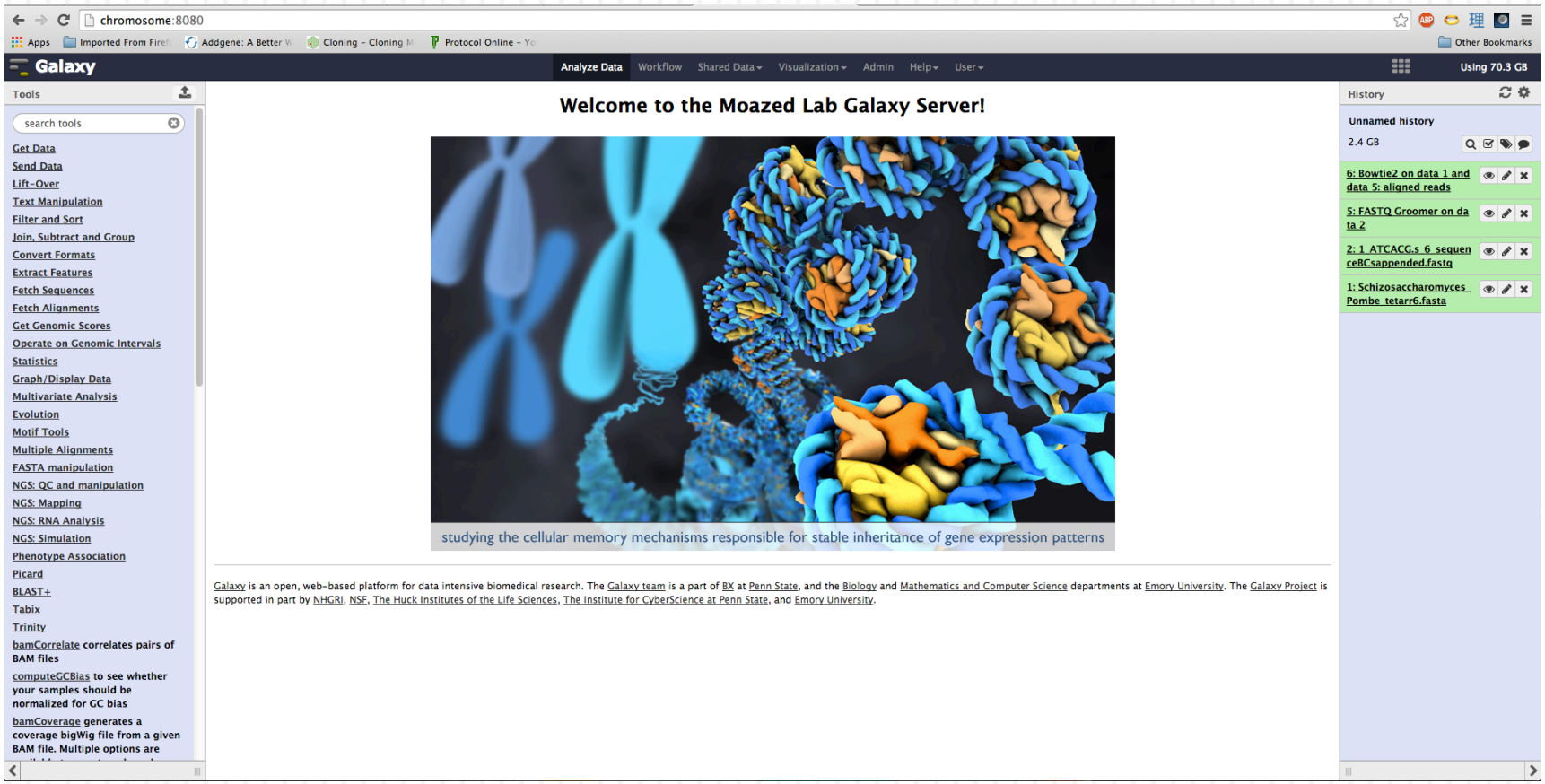
Galaxy is a free, easy to use resource for bioinformatics analysis.
It is not as flexible as other systems, but is a good starting point for the average biologist with a minimal learning curve compared to other methodologies.



The Galaxy interface

- Point your web browser to
  - http://chromosome:8080
- There are three main panes in the interface.
  - The left pane is the list of software tools you can use.
  - The middle pane is used to launch software and view contents of files.
  - The right pane is a history of your commands and results.

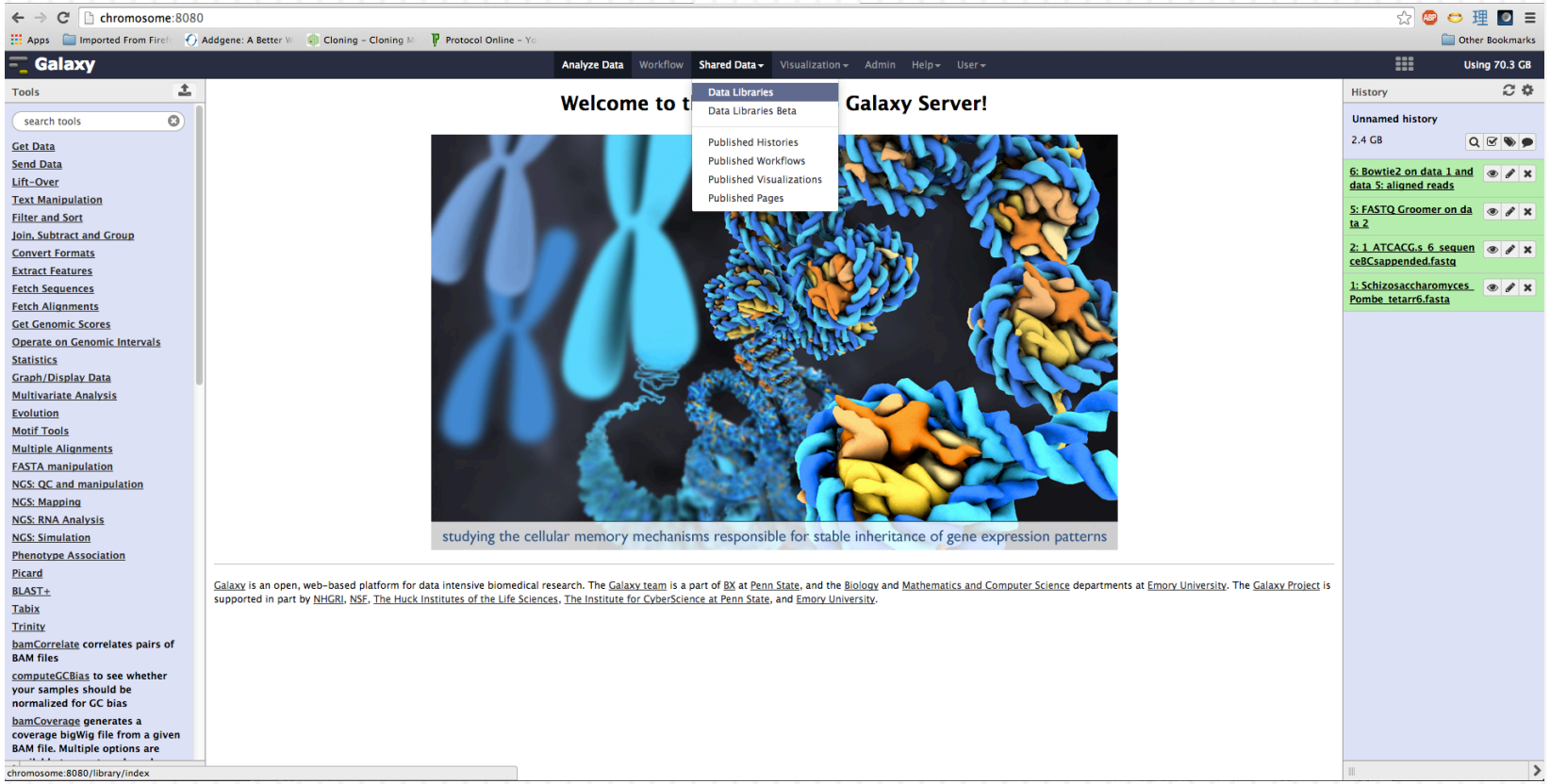
Example of the interface



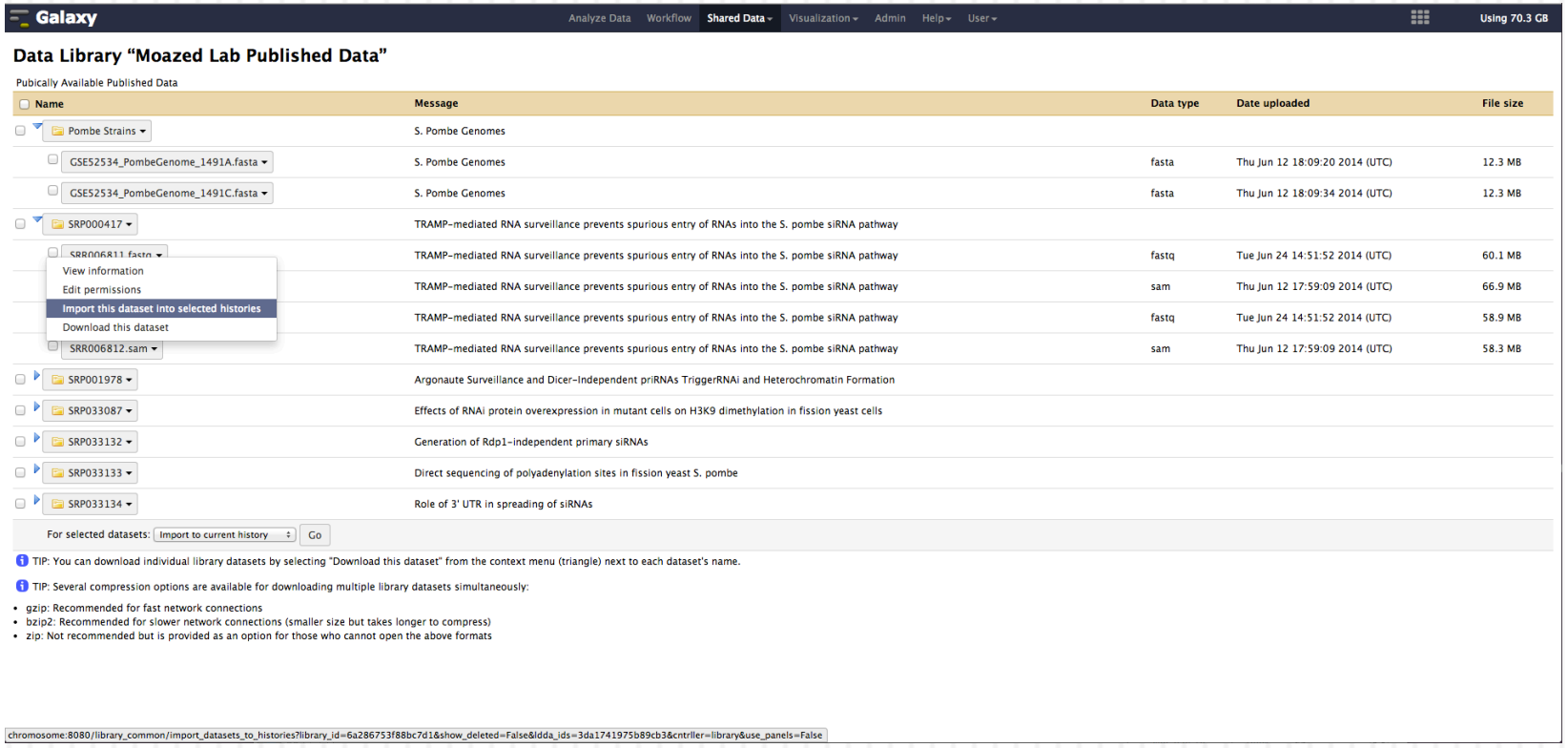
Getting data into Galaxy

- Files can be uploaded into Galaxy by selecting **Get Data** in the left pane and then choosing **Upload File**
- Large, re-useable data sets can be uploaded to the shared **Data Library**. Speak with Cary or Dan to have files stored in the shared Data Library.
- Files in the Data Library can be imported into your Work History by selecting **Shared Data** and then **Data Libraries**. You will then be able to select files you wish to import into your work history.

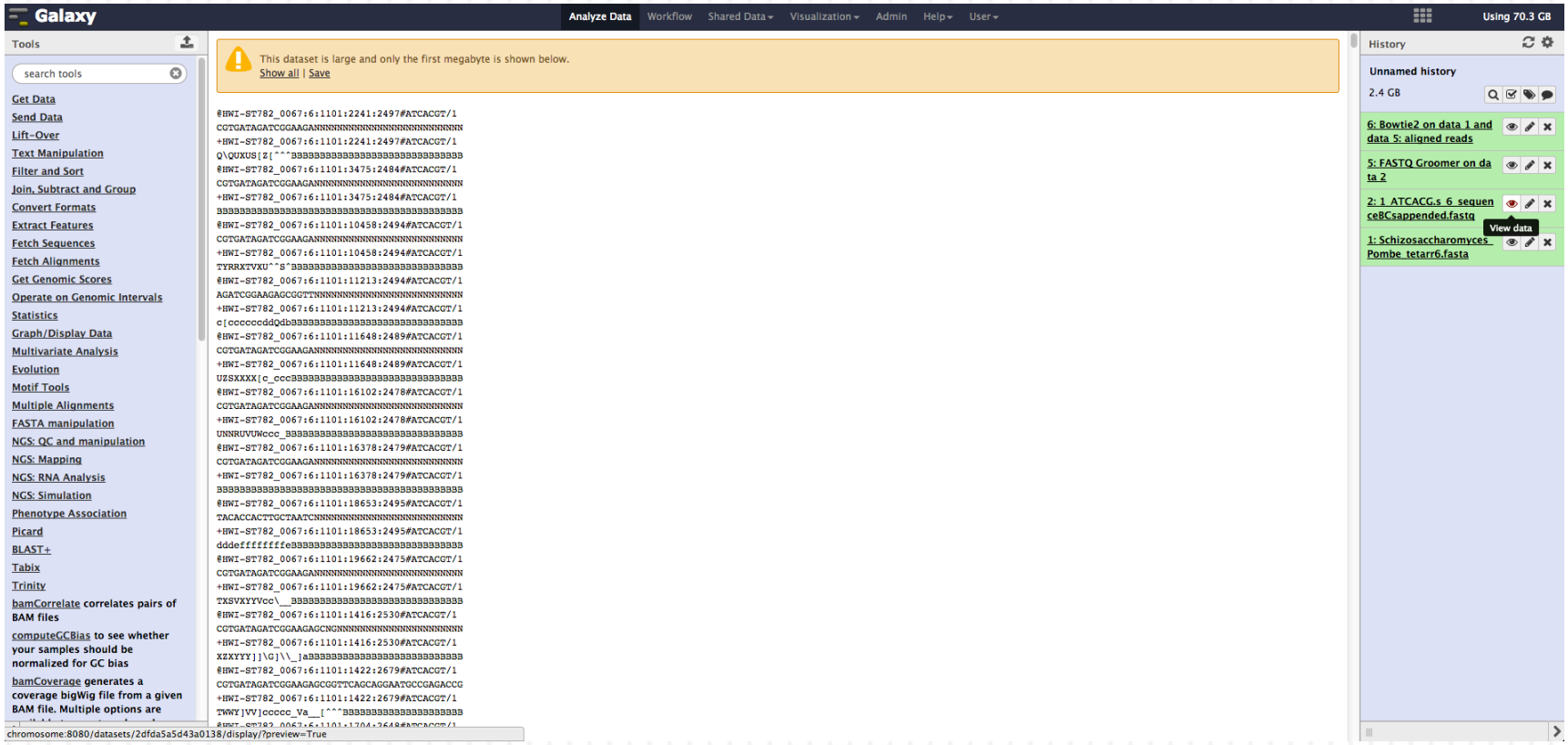
Data Libraries



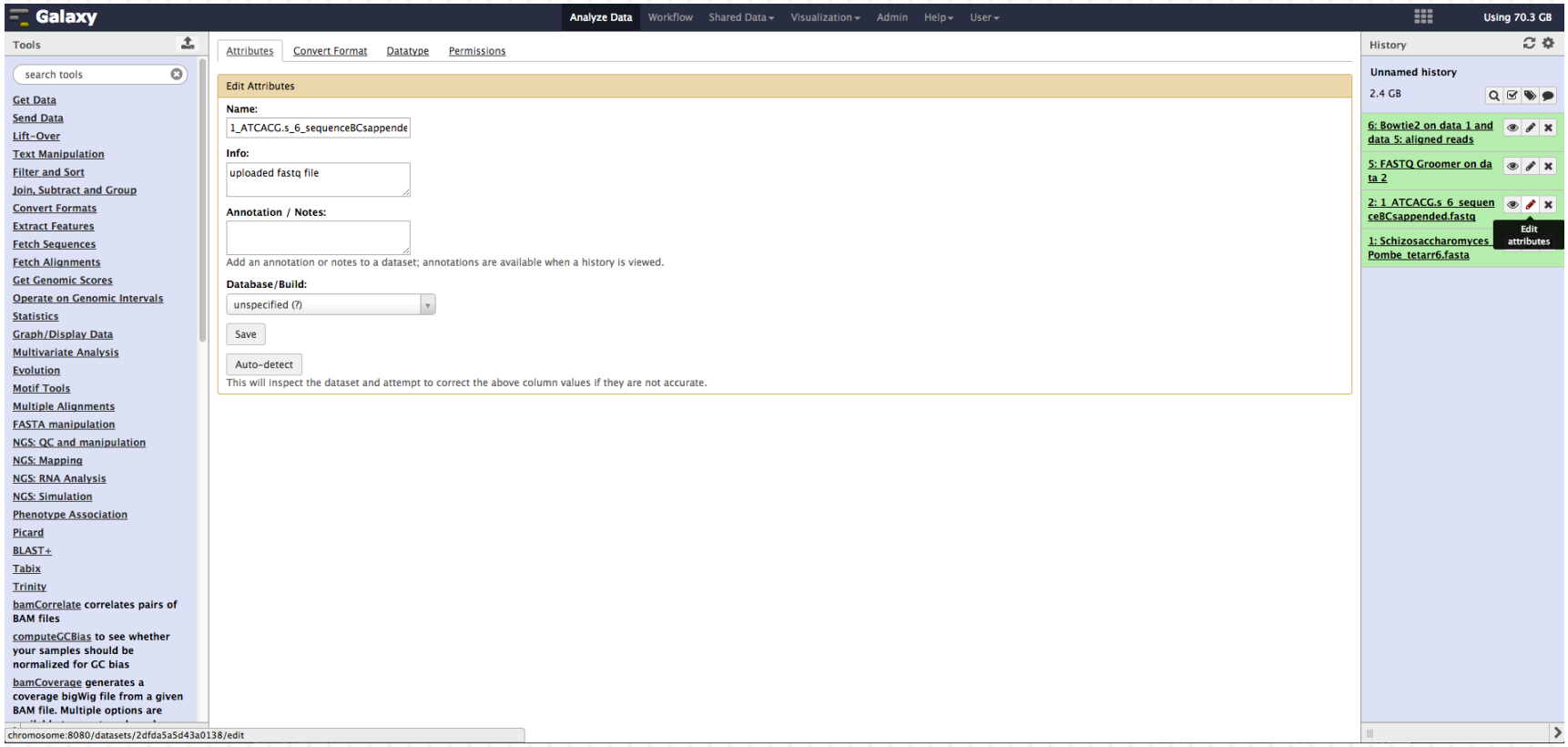
Data Libraries



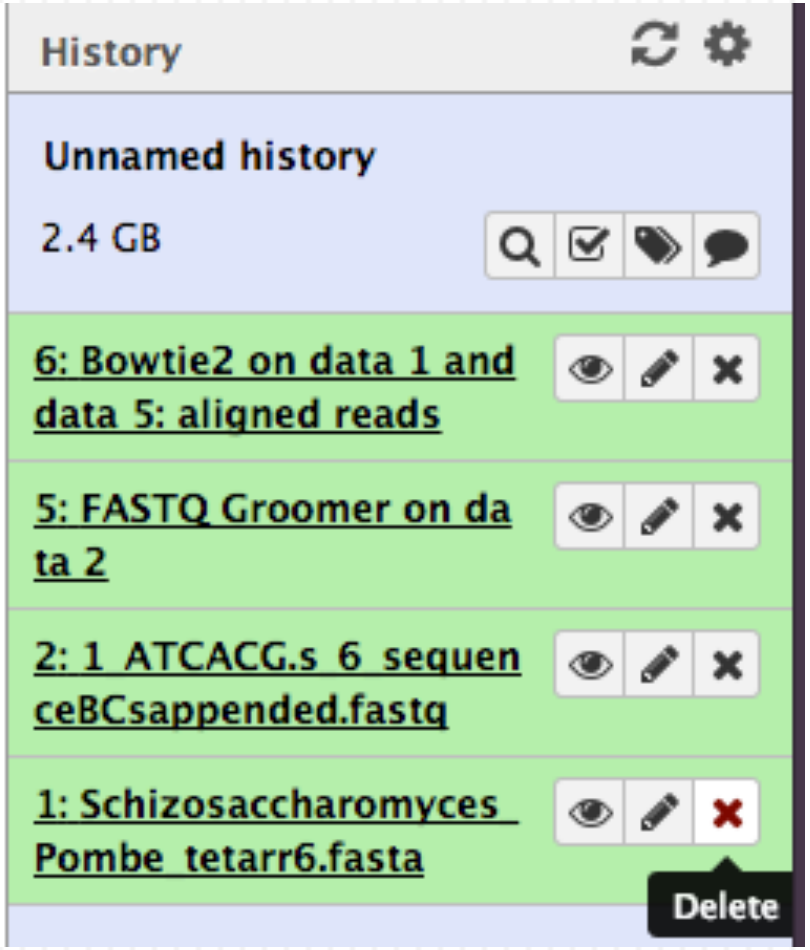
The eye icon may be used to examine the contents of data



The pencil icon can be used to edit attributes associated with the data



The “X” icon will delete data from your work history



Quick walkthrough of data analysis

- The next few slides will walk you through performing a DNA-seq mapping in Galaxy.
- From there, you can export your analysis to visualize the data in IGV.
- Let's get started!

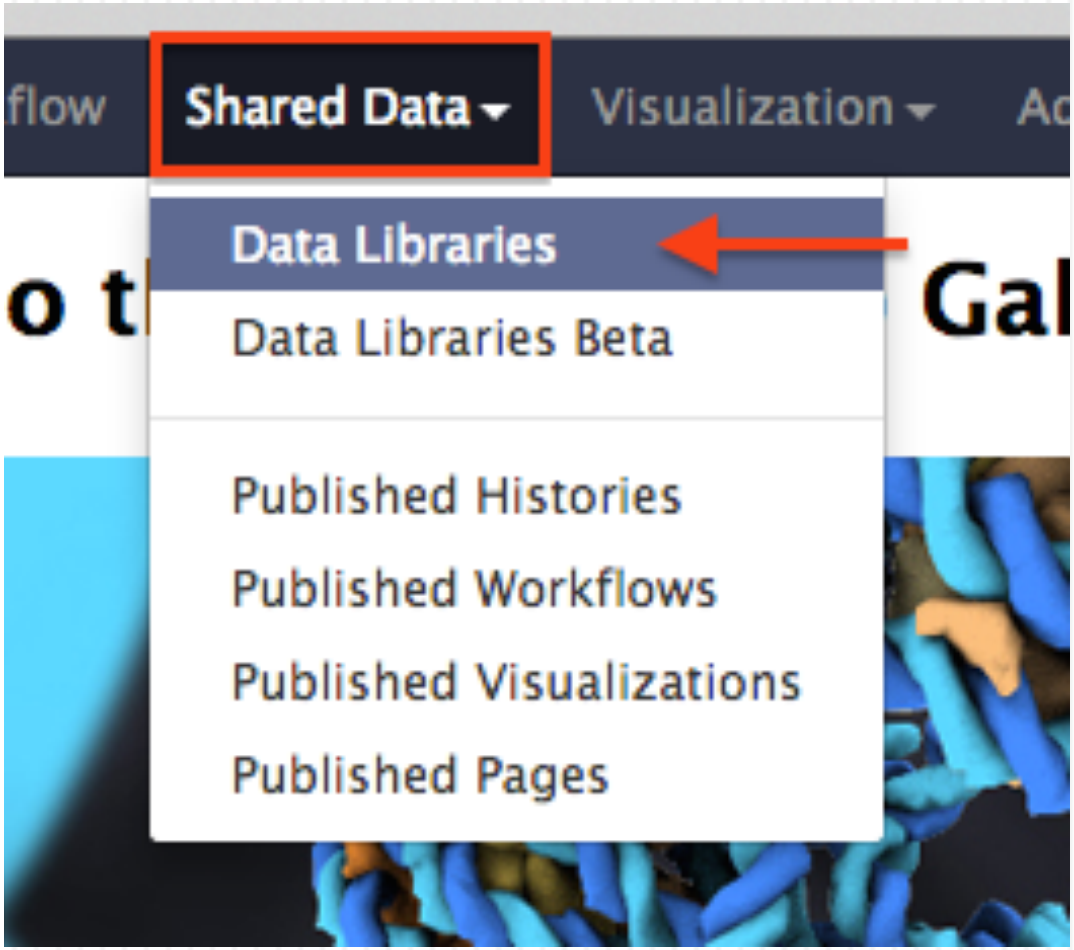
Galaxy bowtie tutorial

- This short tutorial will illustrate how to run bowtie within Galaxy for mapping DNA-Seq reads against a reference genome and how to construct a workflow from the finished analysis.

Importing the data from the Data Library

- The data required has been supplied in a Data Library entitled “Galaxy Tutorial Data Sets”
- Once you have logged in you can select “Shared Data” → “Data Libraries” from the top menu in the middle pane of the interface.

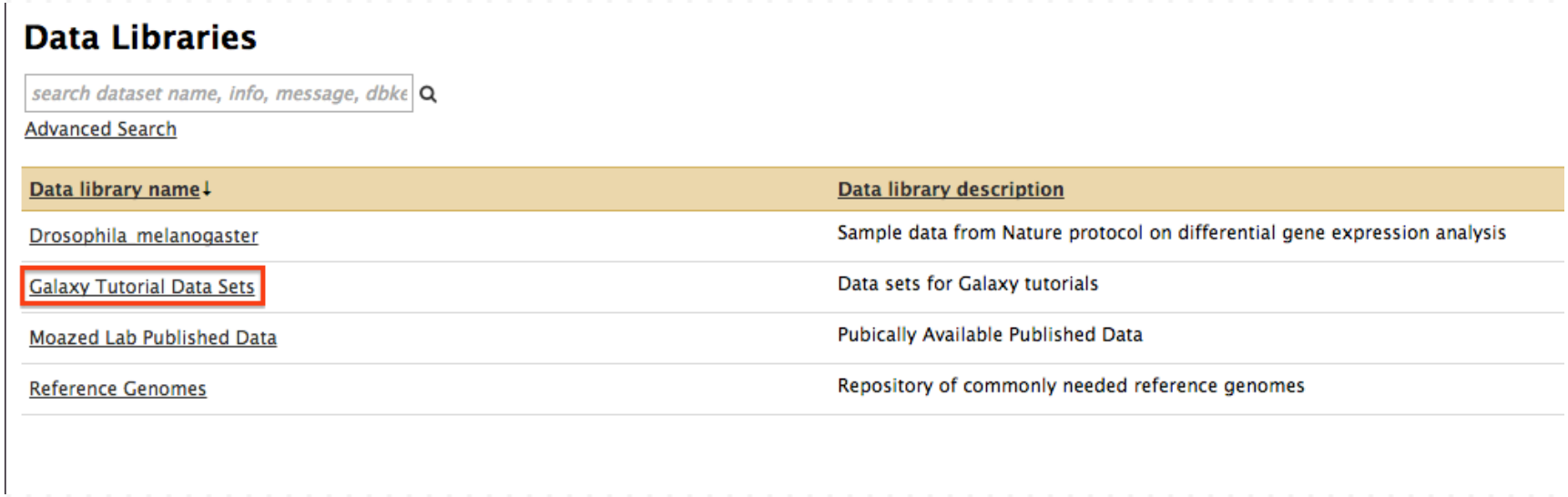
Navigating to a Data Library



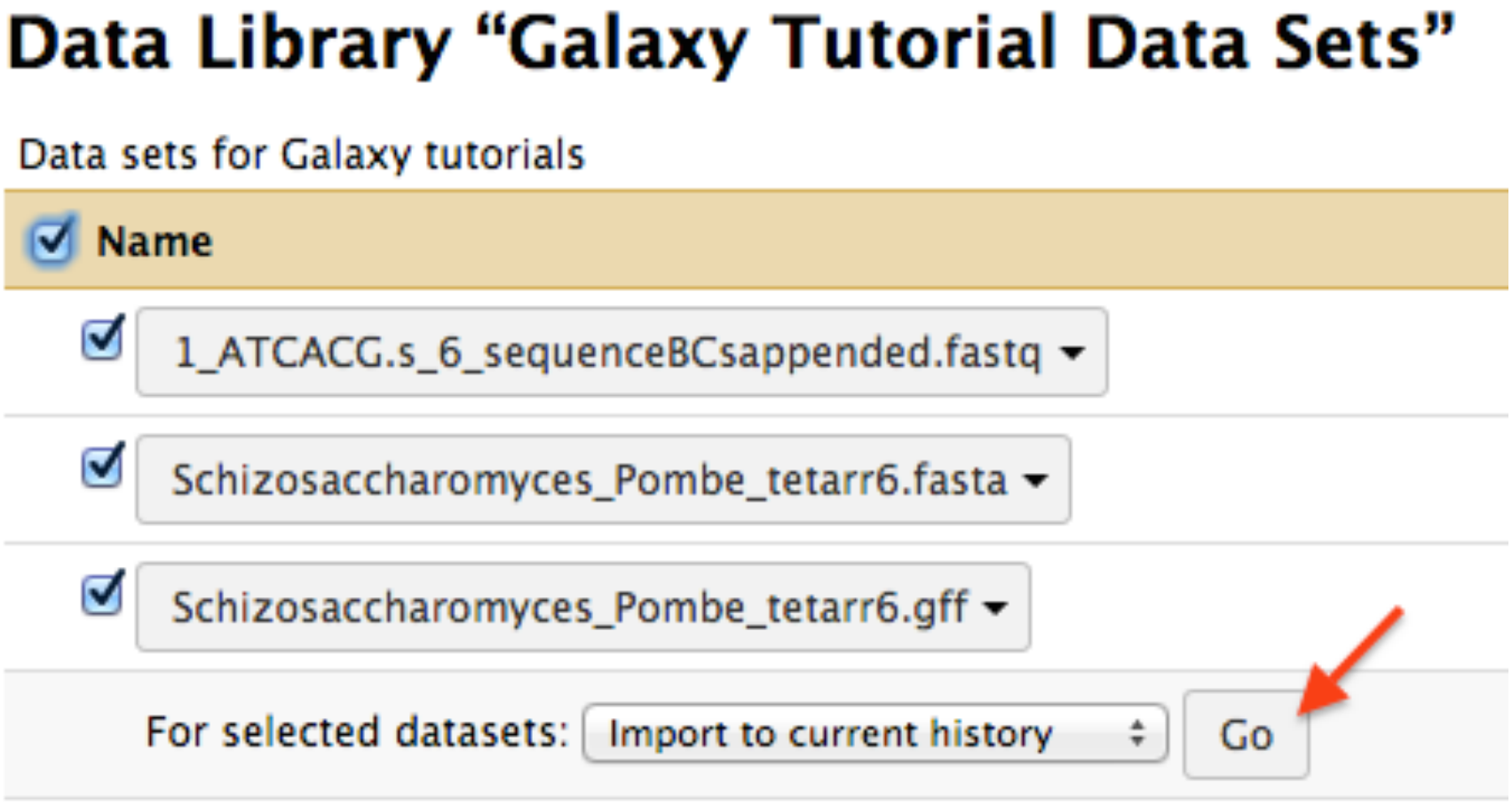
Data Libraries

- Here you will be presented with a list of the available data libraries.
- Select the “Galaxy Tutorial Data Sets” library

Available Data Libraries



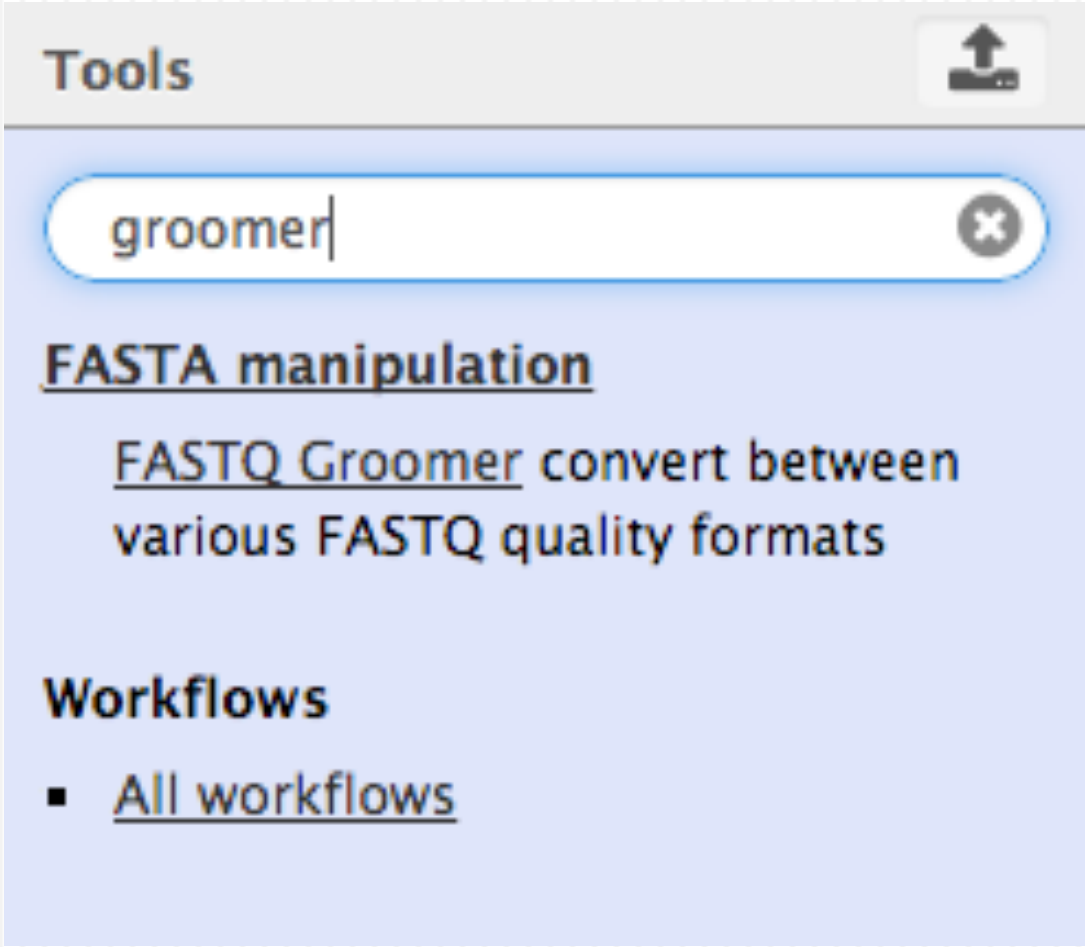
Importing the data to your history



Grooming FASTQ data

- Sequence data has quality scores associated with reads contained in the FASTQ file.
- We will need to use the data groomer to convert the scores into values that Galaxy can utilize to further process our data.
- If you type “groomer” into the search tools textbox in the left pane, you can narrow down the tools available to find the appropriate tool needed.

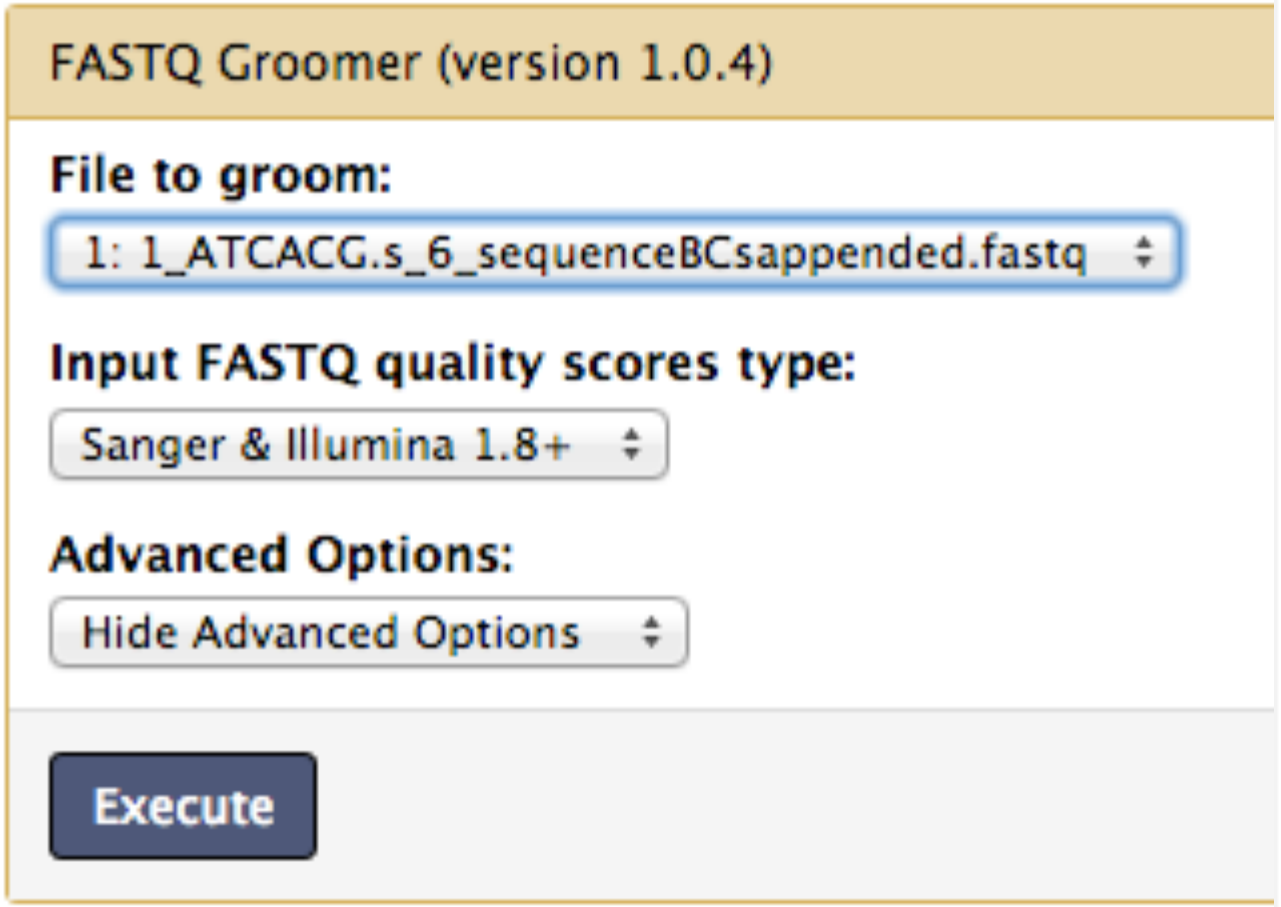
The FASTQ Groomer



Using the groomer

- Selecting the groomer will bring up a dialog in the center pane. Galaxy will try to fill in as many values as it can based on the contents of your history.
- The next screenshot illustrates this behavior.

FASTQ Groomer Dialog



Running the groomer

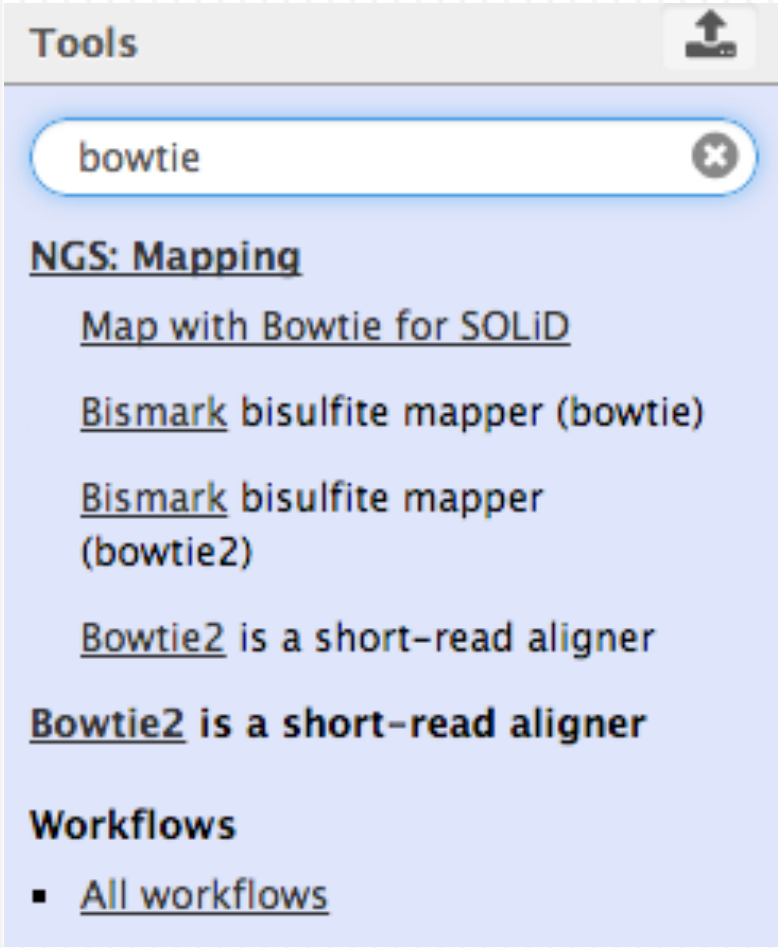
- The groomer has pre-populated the fields we require to begin, so we can select the execute button.
- The history will be populated with a new entry detailing the status of our run.
  - Gray with a little stop watch means the job has been queued for execution
  - Yellow means the job is currently running
  - Green indicates successful completion
  - Red indicates there was a problem with the execution of the task

Newly added and running FASTQ Groomer task

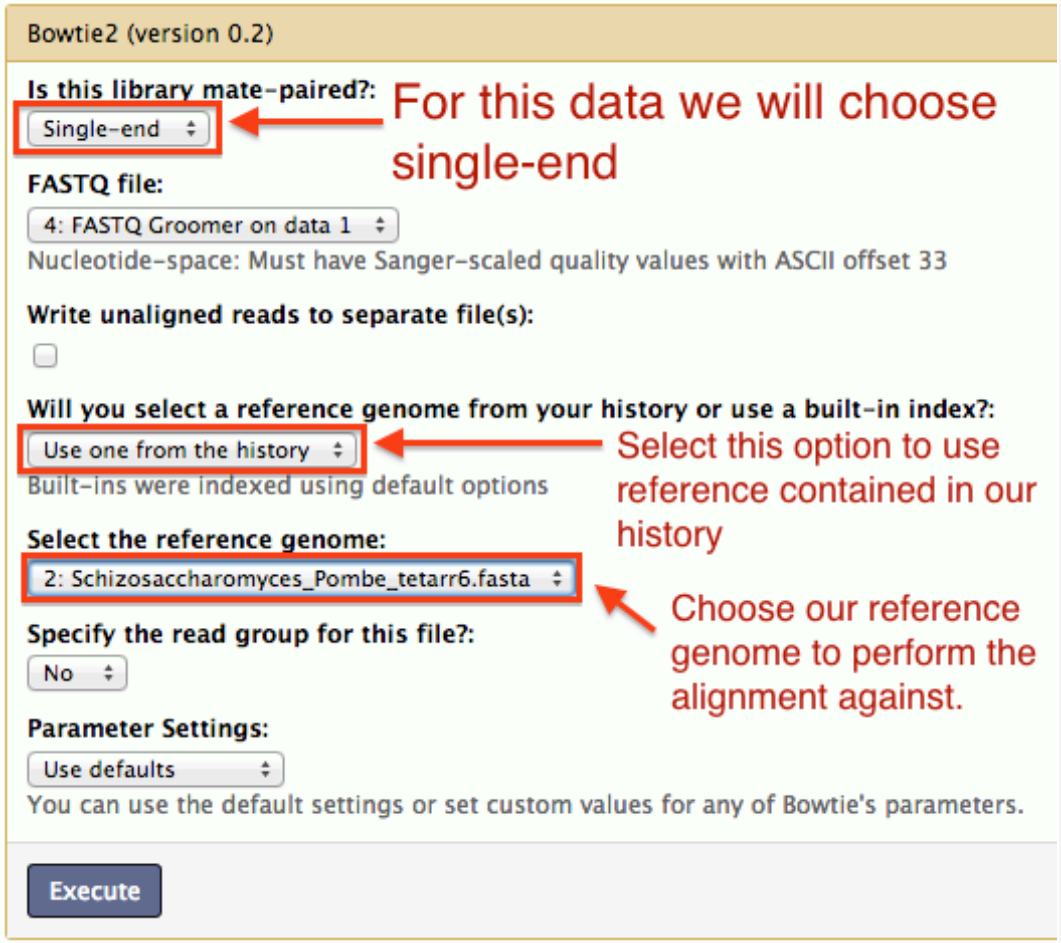
Running Bowtie to align reads

- While the groomer is executing, we can start to stage the next step of the analysis.
- Typing bowtie into the tools search box will allow us to select bowtie2 and begin the edit the bowtie run dialog box

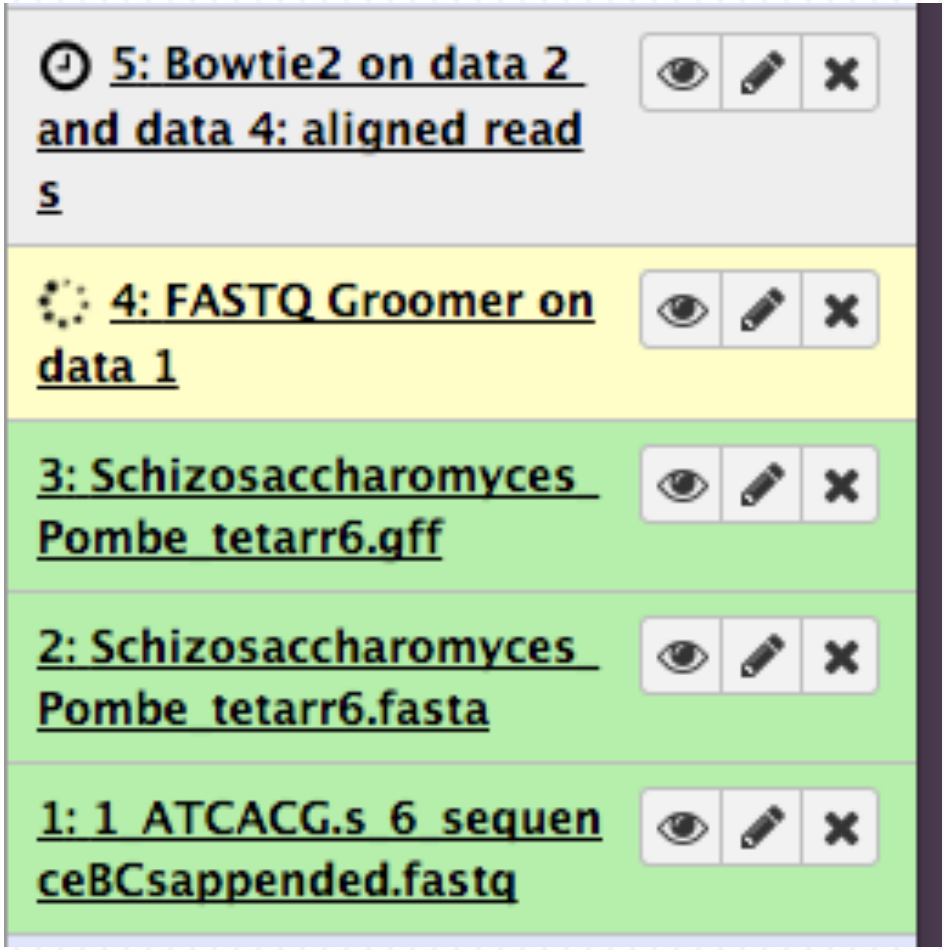
Search results for bowtie



Bowtie2 dialog



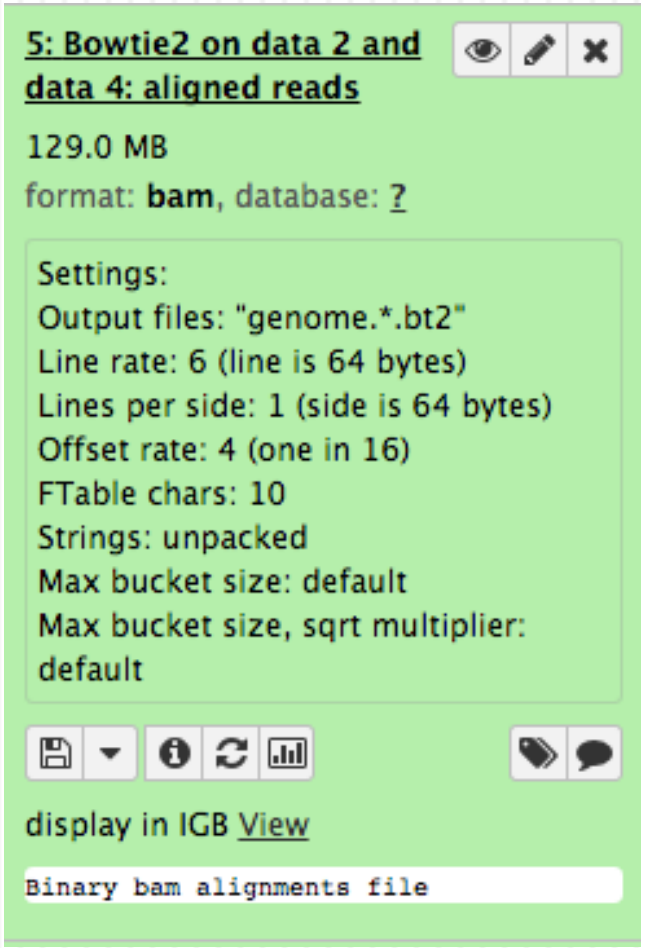
Queuing bowtie to run when groomer completes



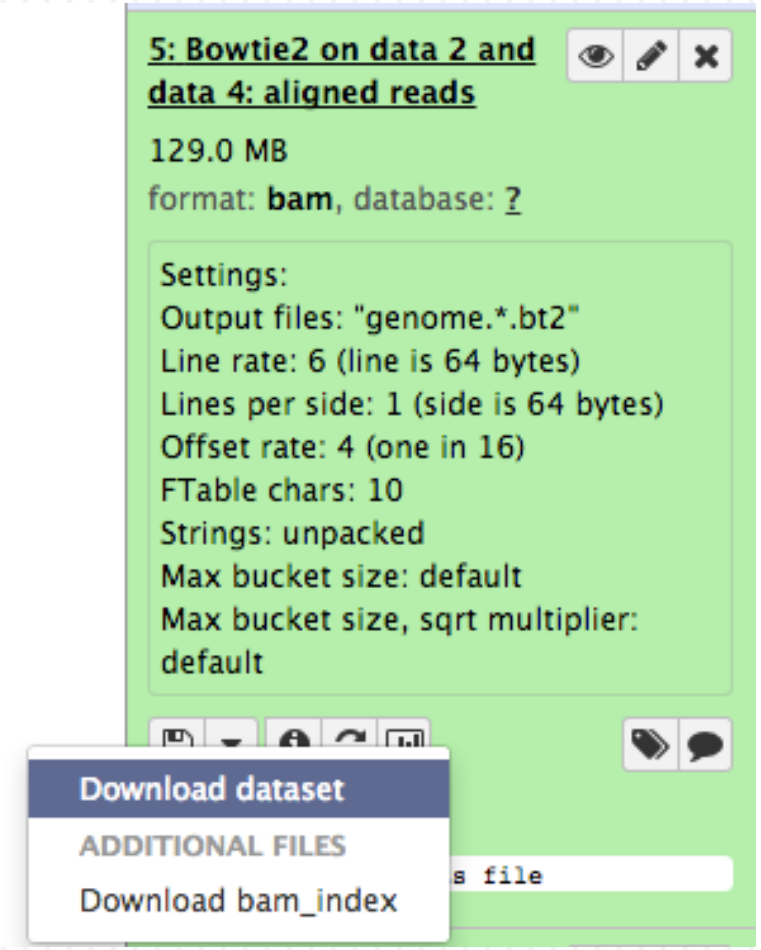
Retrieving the data for further analysis

- Once the alignment completes, you will have an entry in your history that contains the BAM file and the BAI (BAM Index) file.
- You can download these for use in IGV or other downstream visualization tools by selecting the floppy disk icon.

Results


5: Bowtie2 on data 2 and data 4: aligned reads
129.0 MB
format: bam, database: ?
Settings:
Output files: "genome.*.bt2"
Line rate: 6 (line is 64 bytes)
Lines per side: 1 (side is 64 bytes)
Offset rate: 4 (one in 16)
FTable chars: 10
Strings: unpacked
Max bucket size: default
Max bucket size, sqrt multiplier: default
display in IGB View
Binary bam alignments file

Download dataset and index options

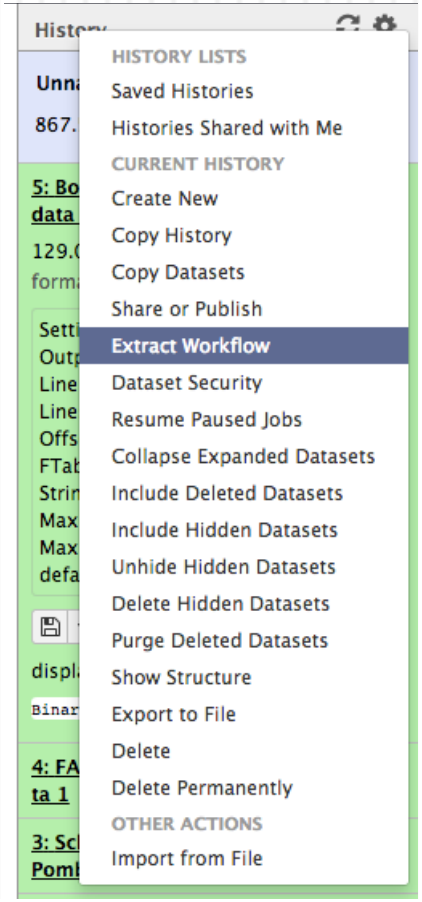


Creating a re-usable workflow

- Now that we have performed an alignment, we might like to run the same alignment again on new data sets.
- One way to accomplish this, is to create a workflow based on our analysis
- Let's walk through creating a workflow from our current analysis for later use

Extracting the workflow

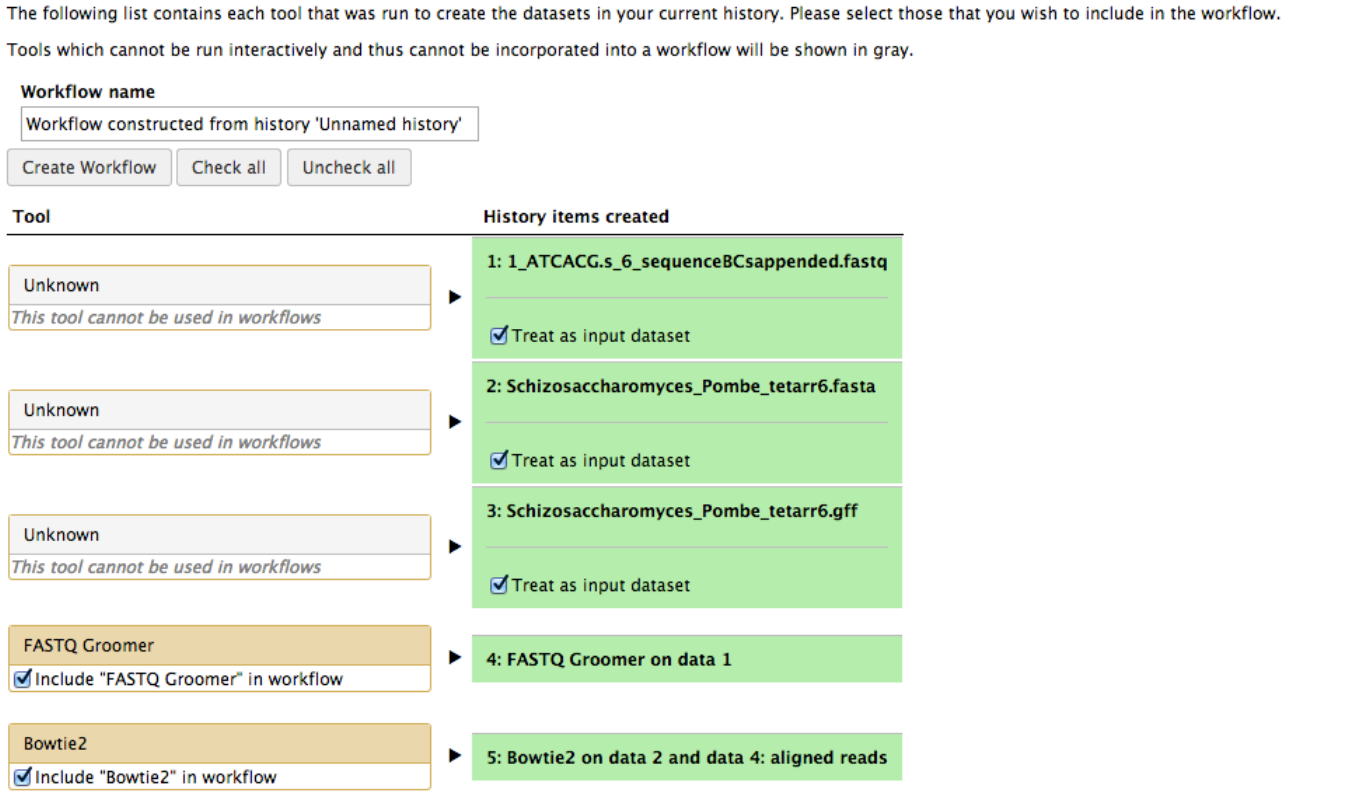
- To extract the workflow, click the gear icon in the history pane and select extract workflow

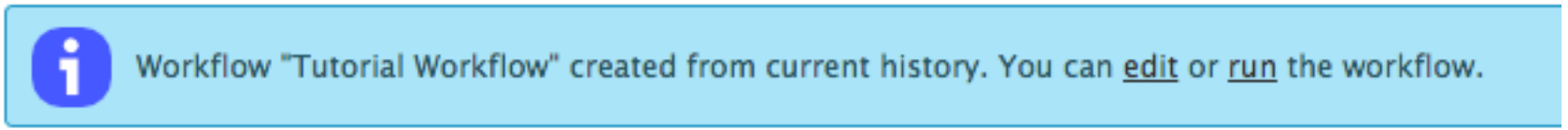


Edit the workflow

- The center pane will show the current workflow based upon your history.
- Choose a name for your workflow and select the “Create Workflow” button.

Workflow creation dialog

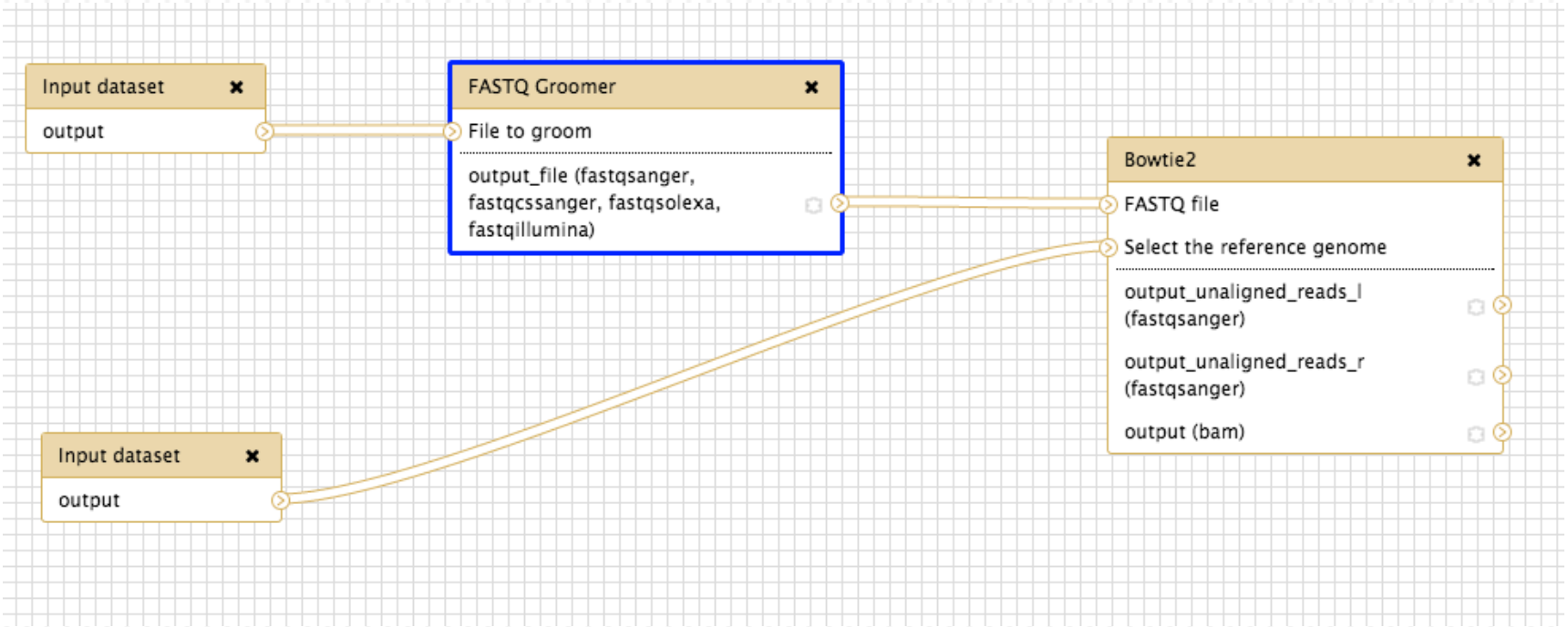




Visualization tool for workflow

- Now that we've created a workflow, we can edit it in the visual editing tool if we need to tweak it or make changes.
- You can select edit from the dialog that appears when you first save the workflow
- Or you can navigate to the workflow by selecting "Workflow" from the top menu bar. Select the workflow you want and choose Edit from the drop down menu that is presented.

Visual workflow editor



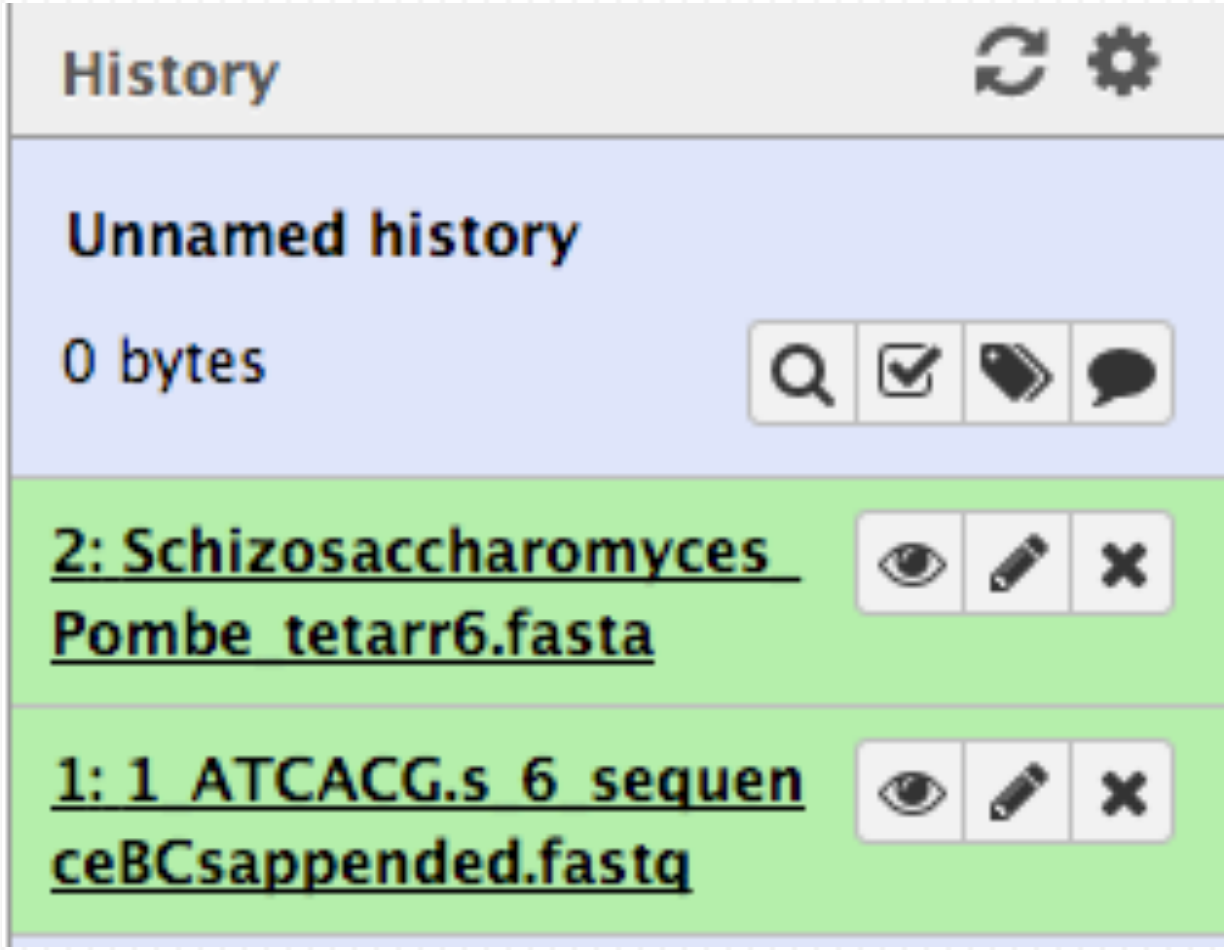
More on the workflow editor

- Tools, inputs and intermediate outputs to be passed as inputs to the next tool in a pipeline are all displayed as boxes
- Boxes have nodes associated with them labeled as inputs on the left side, and outputs on the right.
- We can now run this workflow multiple times.
- Let's create a new history and try running the workflow on the tutorial inputs again to illustrate how this can be useful.

Running a workflow

- To get back to our history, choose “Analyze Data” from the top menu.
- Select the gear on the right pane History menu, and choose “Create New”
- Import the fastq and reference genome file again from the Data Library

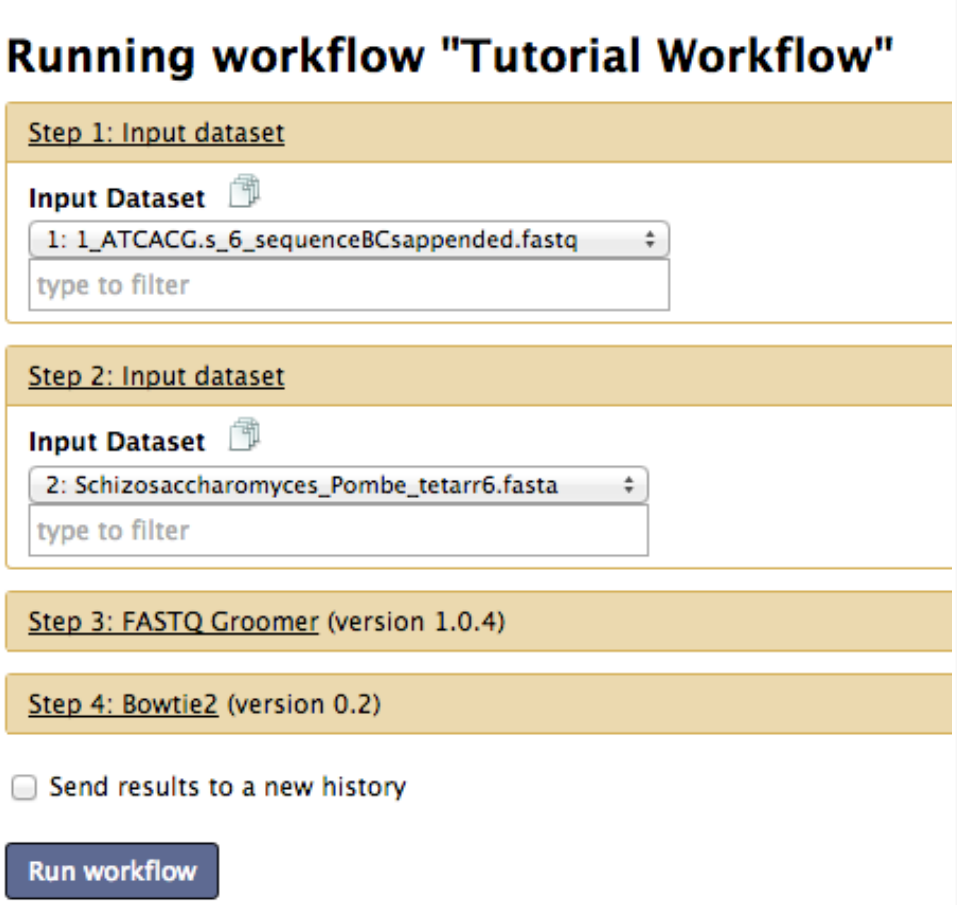
Your new history with data imported



Run the workflow

- Now select workflow from the top menu.
- Choose the workflow you created earlier and select “Run” from the dropdown menu presented.
- You will now have a dialog in the center pane based on your current working history as it relates to the workflow. Galaxy tries to figure out what files should be inputs into the workflow.
- Clicking on the individual steps will show the parameters associated with the step, based on the definitions of the workflow.

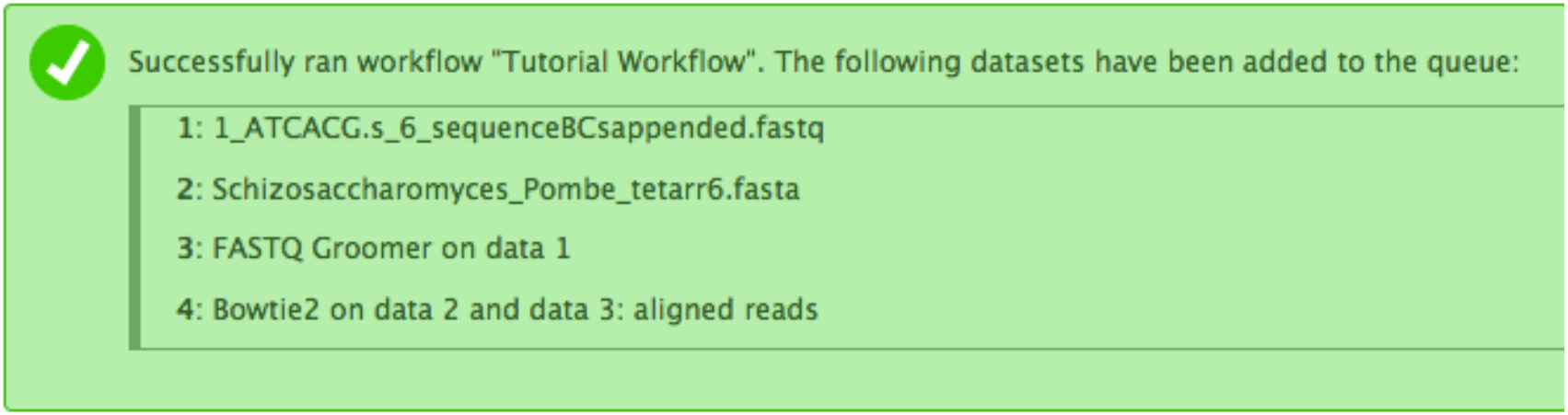
Running workflow dialog



More on workflow execution

- If the inputs look correct and the options are as you would like them to be, you can click the "Run workflow" button.
- The workflow will begin execution.
- Go have a nice cup of coffee while the computer does all the work! ☺
- When you return, your results will be in the current workflow history

Successful workflow run message



More tutorials and resources

- I am but a humble document, covering the very basics of Galaxy.
- For more information and tutorials on how to run more complex analyses, please check out the following references:
  - https://usegalaxy.org/u/aun1/p/galaxy101
  - https://usegalaxy.org/page/list_published

Resources, where to get help

- There are also webcasts and videos detailing how to interact with Galaxy. Look for them here on vimeo:
  - http://vimeo.com/channels/usegalaxy
- If you need assistance, the biostar galaxy forums is the place to go. The Galaxy development team runs these forums.
  - https://biostar.usegalaxy.org/


UAA