

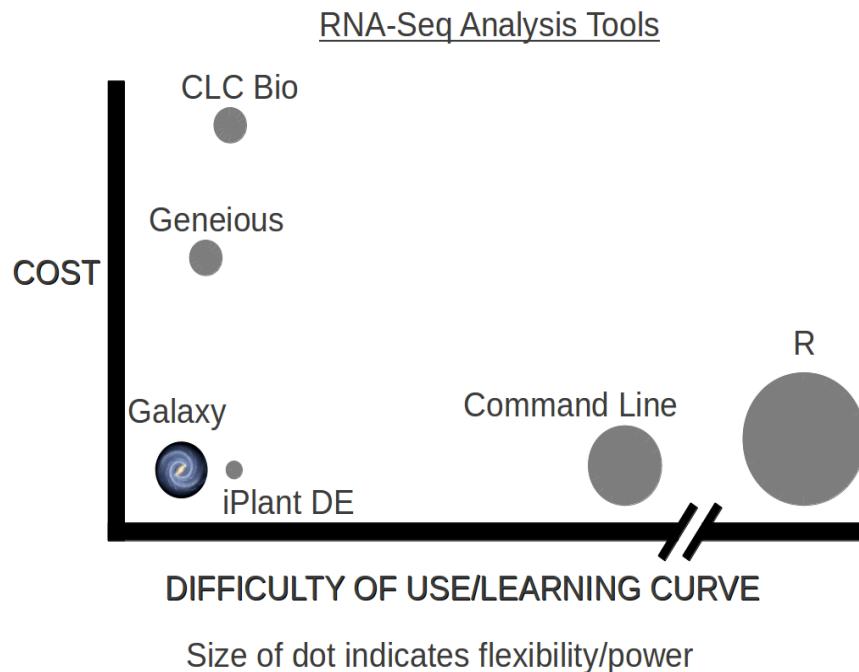


GALAXY INTRODUCTION AND TUTORIAL

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

The screenshot shows the Galaxy web interface running on a server at chromosome:8080. The title bar reads "Welcome to the Moazed Lab Galaxy Server!". The main content area features a 3D molecular model of chromatin or RNA, with text below it stating "studying the cellular memory mechanisms responsible for stable inheritance of gene expression patterns". The left sidebar contains a comprehensive list of tools categorized under "Tools", including Get Data, Send Data, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, FASTA manipulation, NGS: QC and manipulation, NGS: Mapping, NGS: RNA Analysis, NGS: Simulation, Phenotype Association, Picard, BLAST+, Tabix, Trinity, bamCorrelate, computeGCBias, and bamCoverage. The right sidebar displays the "History" panel, which lists six entries: 1. Schizosaccharomyces Pombe tetarr6.fasta, 2. 1 ATCACG.s 6 sequence ceBCs appended.fasta, 3. FASTQ Groomer on data2, 4. Bowtie2 on data1 and data5 aligned reads, 5. Other Bookmarks, and 6. Using 70.3 GB.

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

The screenshot shows the Galaxy web interface at chromosome:8080. The main header includes tabs for Analyze Data, Workflow, Shared Data (highlighted), Visualization, Admin, Help, and User. A sub-menu for 'Shared Data' is open, showing options: Data Libraries Beta, Published Histories, Published Workflows, Published Visualizations, and Published Pages. The central content area features a large image of a protein structure composed of blue and orange subunits, with the text "Galaxy Server!" above it and a subtitle below stating "studying the cellular memory mechanisms responsible for stable inheritance of gene expression patterns". To the left is a sidebar titled "Tools" with a search bar and a long list of tool categories. On the right is a "History" panel titled "Unnamed history" showing six recent jobs: Bowtie2, FASTQ Groomer, ATCACCs, Schizosaccharomyces, and Pombe_tetarr6. The bottom status bar shows the URL chromosome:8080/library/index.

Data Libraries

Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User Using 70.3 GB

Data Library "Moazed Lab Published Data"

Pubically Available Published Data

Name	Message	Data type	Date uploaded	File size
Pombe Strains	S. Pombe Genomes			
GSE52534_PombeGenome_1491A.fasta	S. Pombe Genomes	fasta	Thu Jun 12 18:09:20 2014 (UTC)	12.3 MB
GSE52534_PombeGenome_1491C.fasta	S. Pombe Genomes	fasta	Thu Jun 12 18:09:34 2014 (UTC)	12.3 MB
SRP000417	TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the S. pombe siRNA pathway			
SRR006811.fasta	View information Edit permissions Import this dataset into selected histories Download this dataset	fastq	Tue Jun 24 14:51:52 2014 (UTC)	60.1 MB
SRR006812.sam	TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the S. pombe siRNA pathway	fastq	Tue Jun 24 14:51:52 2014 (UTC)	66.9 MB
SRP001978	Argonaute Surveillance and Dicer-Independent priRNAs TriggerRNAi and Heterochromatin Formation	fastq	Tue Jun 24 14:51:52 2014 (UTC)	58.9 MB
SRP033087	Effects of RNAi protein overexpression in mutant cells on H3K9 dimethylation in fission yeast cells	fastq	Thu Jun 12 17:59:09 2014 (UTC)	58.3 MB
SRP033132	Generation of RdP1-independent primary siRNAs	fastq	Thu Jun 12 17:59:09 2014 (UTC)	60.1 MB
SRP033133	Direct sequencing of polyadenylation sites in fission yeast S. pombe	fastq	Thu Jun 12 17:59:09 2014 (UTC)	66.9 MB
SRP033134	Role of 3' UTR in spreading of siRNAs	fastq	Thu Jun 12 17:59:09 2014 (UTC)	58.9 MB

For selected datasets: Import to current history Go

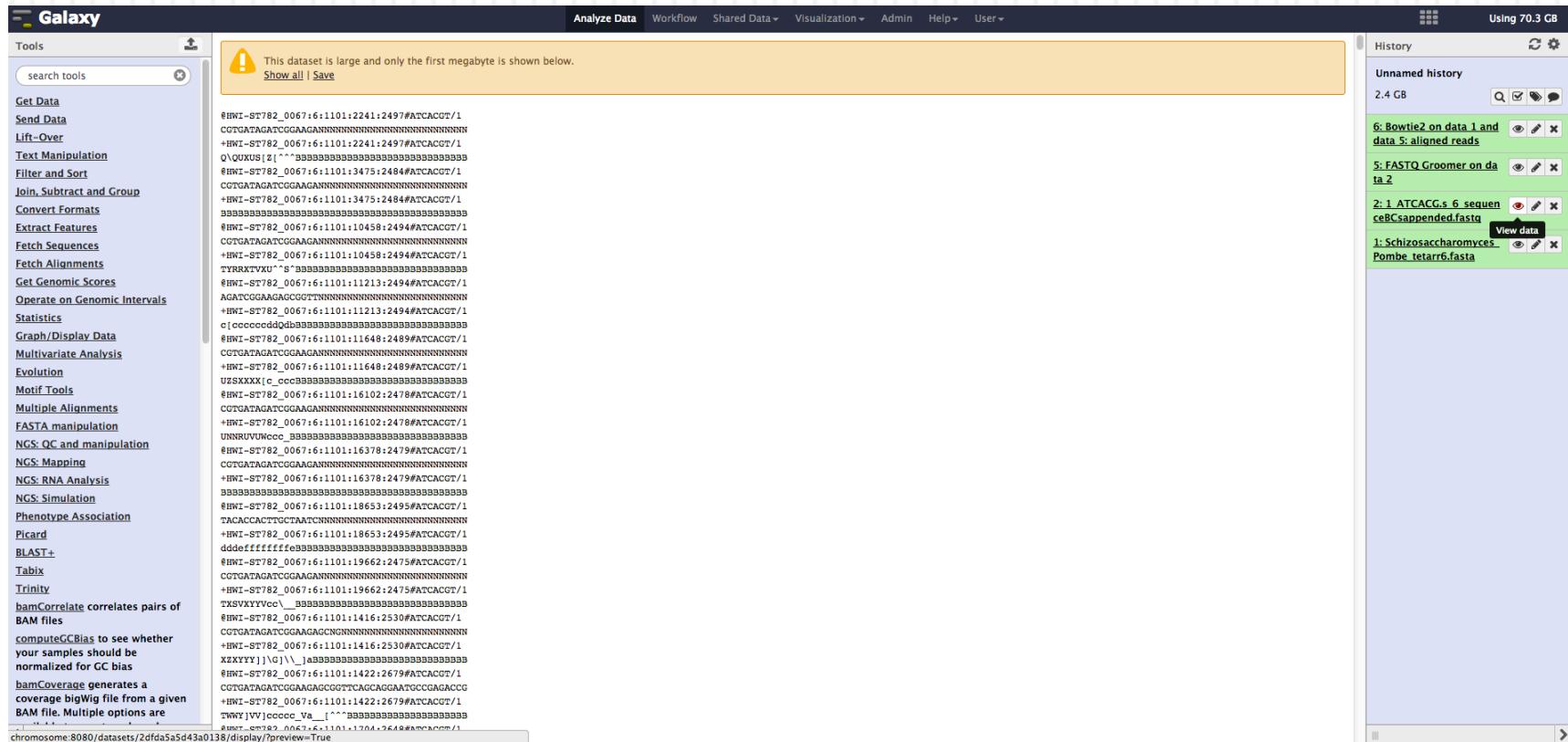
TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.

TIP: Several compression options are available for downloading multiple library datasets simultaneously:

- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats

chromosome:8080/library_common/import_datasets_to_histories?library_id=6a286753f88bc7d1&show_deleted=False&ldda_ids=3da1741975b89cb3&controller=library&use_panels=False

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 screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Tools' (with a search bar), 'Attributes' (which is selected), 'Convert Format', 'Datatype', and 'Permissions'. Below the navigation is a toolbar with 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. A status bar at the top right indicates 'Using 70.3 GB'.

The main content area is titled 'Edit Attributes' for a dataset named '1_ATCAGG.s_6_sequenceBCsappende'. The form fields include:

- Name:** 1_ATCAGG.s_6_sequenceBCsappende
- Info:** uploaded fastq file
- Annotation / Notes:** (empty text area)
- Database/Build:** unspecified (?)

Buttons for 'Save' and 'Auto-detect' are present. A note below the 'Auto-detect' button states: "This will inspect the dataset and attempt to correct the above column values if they are not accurate."

The right side of the interface shows a 'History' panel titled 'Unnamed history' containing several items:

- 6: Bowtie2 on data 1 and data 5: aligned reads
- 5: FASTQ Groomer on data 2
- 2: 1_ATCAGG.s_6_sequenceBCsappende.fasta
- 1: Schizosaccharomyces Pombe tetarr6.fasta

The bottom left of the interface displays the URL: 'chromosome:8080/datasets/2dfda5a5d43a0138/edit'

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

History

Unnamed history
2.4 GB

6: Bowtie2 on data 1 and data 5: aligned reads

5: FASTQ Groomer on data 2

2: 1 ATCACG.s 6 sequences appended.fasta

1: Schizosaccharomyces Pombe tetarr6.fasta

Delete

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!

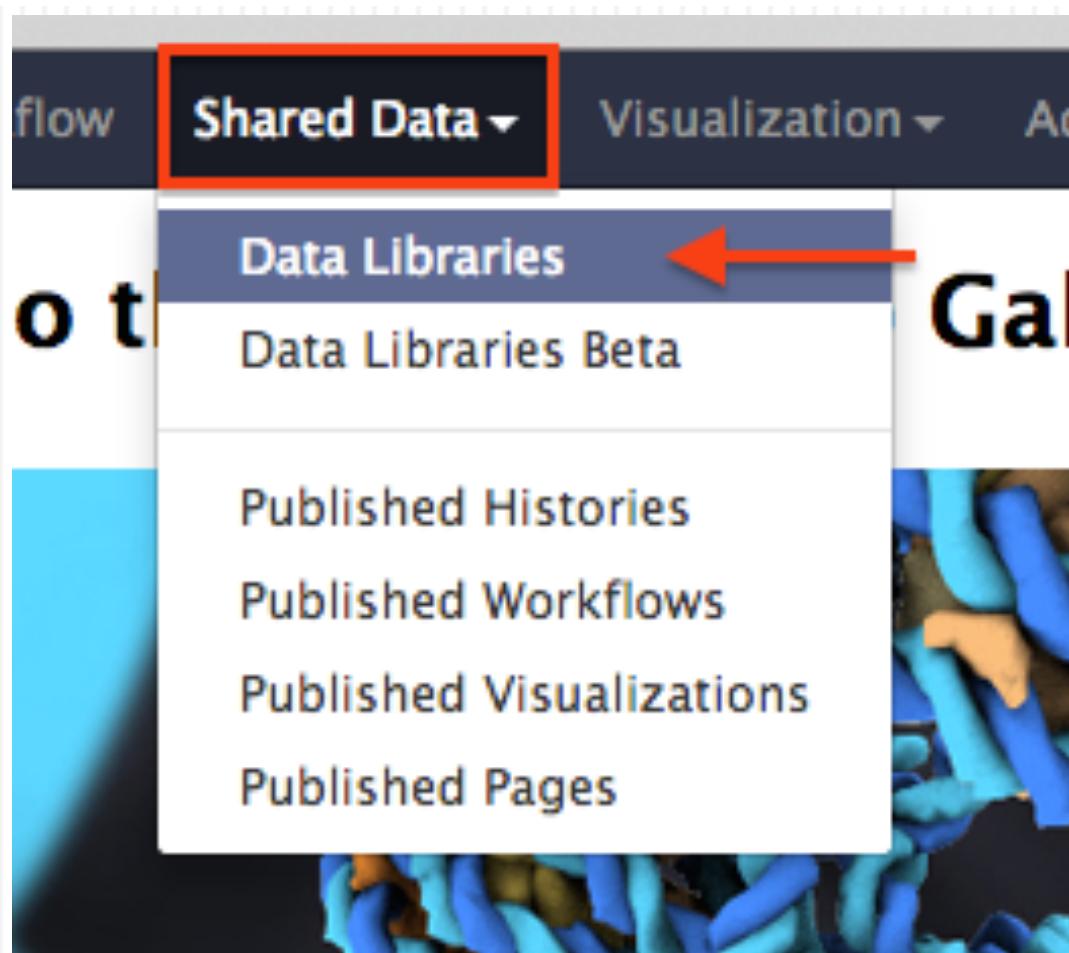
Galaxy bowtie tutorial

- 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

Data Libraries

search dataset name, info, message, dbke 

[Advanced Search](#)

<u>Data library name</u> ↓	<u>Data library description</u>
Drosophila melanogaster	Sample data from Nature protocol on differential gene expression analysis
Galaxy Tutorial Data Sets	Data sets for Galaxy tutorials
Moazed Lab Published Data	Pubically Available Published Data
Reference Genomes	Repository of commonly needed reference genomes

Importing the data to your history

Data Library “Galaxy Tutorial Data Sets”

Data sets for Galaxy tutorials

Name



1_ATCACG.s_6_sequenceBCsappended.fastq ▾



Schizosaccharomyces_Pombe_tetarr6.fasta ▾



Schizosaccharomyces_Pombe_tetarr6.gff ▾

For selected datasets: Import to current history ▾

Go



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

The screenshot shows a web interface for a bioinformatics tool. At the top, there's a header bar with an orange square on the left and a green bar on the right. Below the header, the word "Tools" is displayed in a brown font next to an upload icon (a folder with an upward arrow). A search bar contains the text "groomer". Below the search bar, the text "FASTA manipulation" is underlined in brown. To its right, the text "FASTQ Groomer convert between various FASTQ quality formats" is displayed. Further down, the section "Workflows" is underlined in brown, followed by a bullet point and the link "All workflows".

Tools

groomer

FASTA manipulation

FASTQ Groomer convert between various FASTQ quality formats

Workflows

- All workflows

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

FASTQ Groomer (version 1.0.4)

File to groom:

1: 1_ATCACCG.s_6_sequenceBCsappended.fastq ▾

Input FASTQ quality scores type:

Sanger & Illumina 1.8+ ▾

Advanced Options:

Hide Advanced Options ▾

Execute

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

Tools 

bowtie 

[NGS: Mapping](#)

[Map with Bowtie for SOLiD](#)

[Bismark bisulfite mapper \(bowtie\)](#)

[Bismark bisulfite mapper
\(bowtie2\)](#)

[Bowtie2 is a short-read aligner](#)

[Bowtie2 is a short-read aligner](#)

Workflows

- [All workflows](#)

Bowtie2 dialog

Bowtie2 (version 0.2)

Is this library mate-paired?: **Single-end** ← For this data we will choose single-end

FASTQ file: 4: FASTQ Groomer on data 1
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Write unaligned reads to separate file(s):

Will you select a reference genome from your history or use a built-in index?: **Use one from the history** ← Select this option to use reference contained in our history
Built-ins were indexed using default options

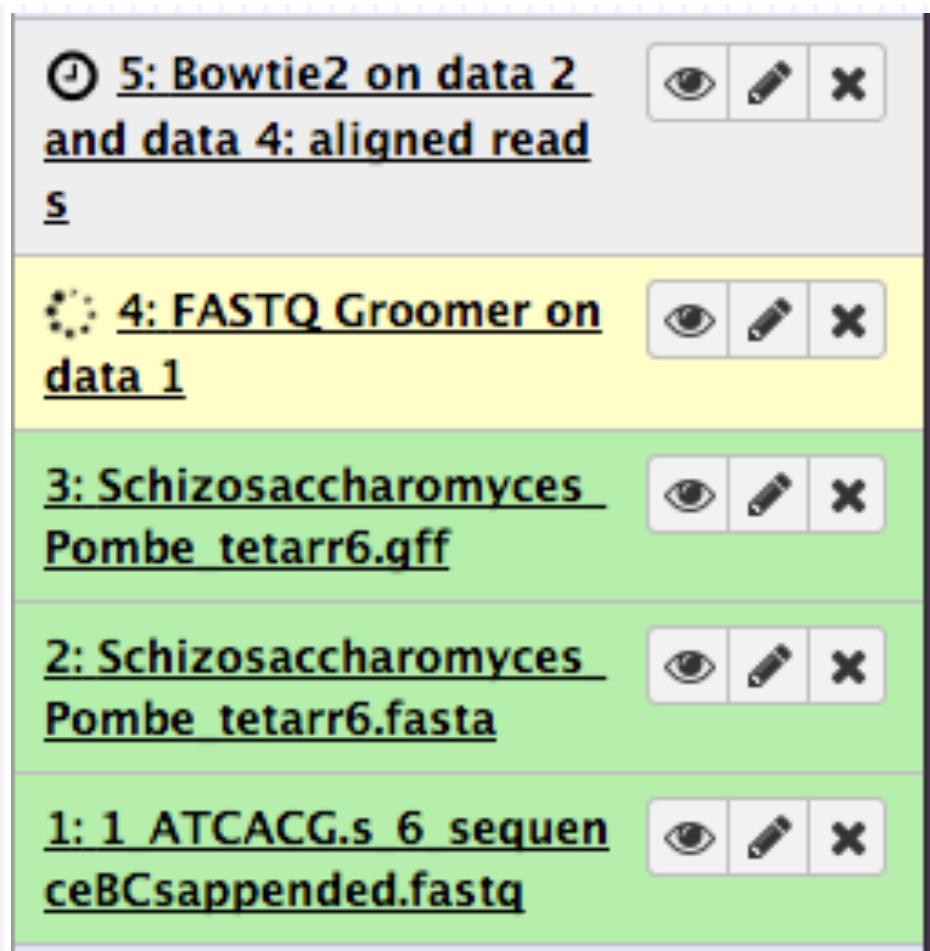
Select the reference genome: 2: Schizosaccharomyces_Pombe_tetarr6.fasta ← Choose our reference genome to perform the alignment against.

Specify the read group for this file?: No

Parameter Settings: Use defaults
You can use the default settings or set custom values for any of Bowtie's parameters.

Execute

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

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

Download dataset

ADDITIONAL FILES

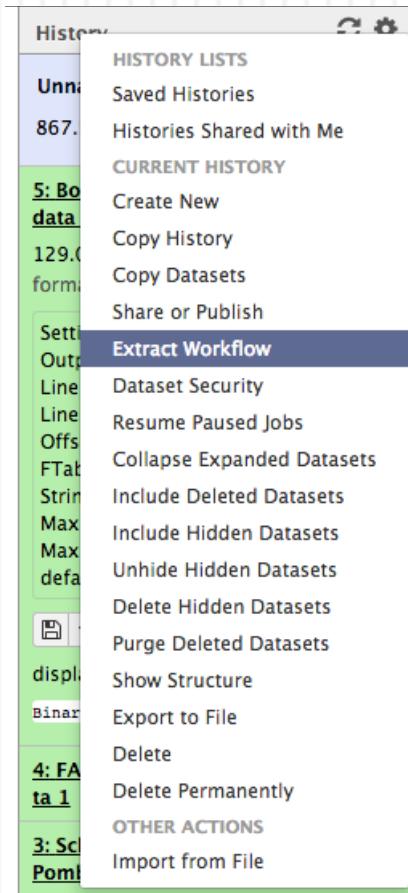
Download bam_index

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

The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name

Workflow constructed from history 'Unnamed history'

[Create Workflow](#)

[Check all](#)

[Uncheck all](#)

Tool

History items created

Unknown

This tool cannot be used in workflows

1: 1_ATCACG.s_6_sequenceBCsappended.fastq

Treat as input dataset

Unknown

This tool cannot be used in workflows

2: Schizosaccharomyces_Pombe_tetarr6.fasta

Treat as input dataset

Unknown

This tool cannot be used in workflows

3: Schizosaccharomyces_Pombe_tetarr6.gff

Treat as input dataset

FASTQ Groomer

Include "FASTQ Groomer" in workflow

4: FASTQ Groomer on data 1

Bowtie2

Include "Bowtie2" in workflow

5: Bowtie2 on data 2 and data 4: aligned reads

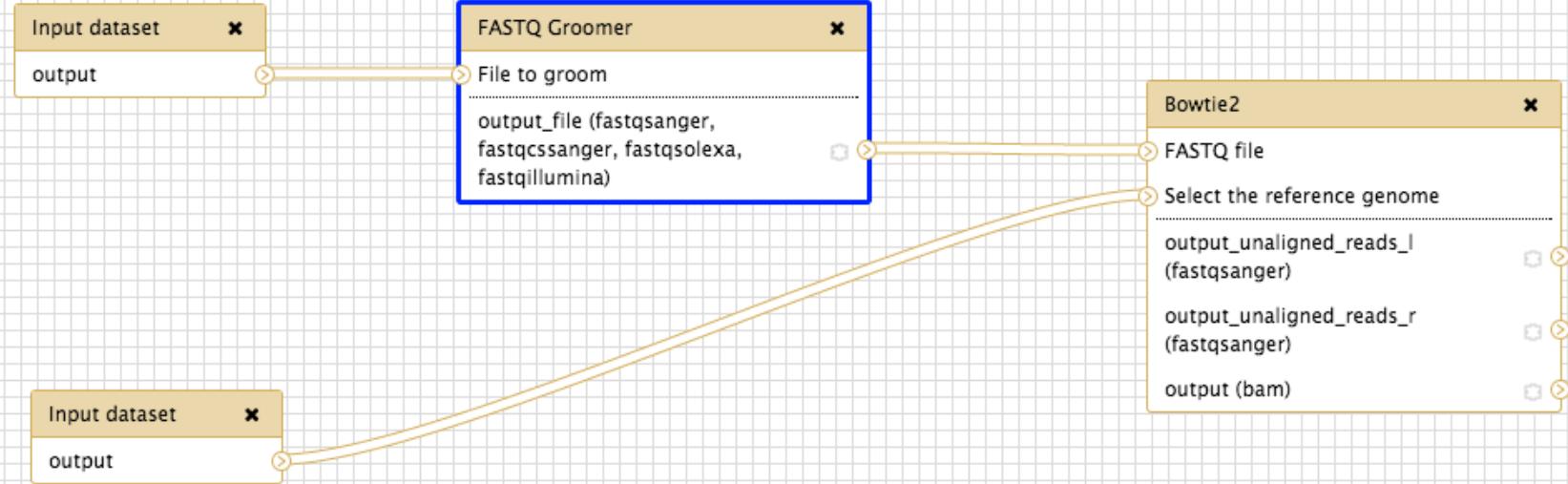


Workflow "Tutorial Workflow" created from current history. You can [edit](#) or [run](#) the workflow.

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

The screenshot shows a software interface with a top navigation bar featuring an orange square icon, a green bar, and a search bar with placeholder text "Search". Below this is a "History" section with the following details:

- Unnamed history**: 0 bytes. Includes a search icon and three other icons (checkmark, clipboard, speech bubble).
- 2: Schizosaccharomyces Pombe tetarr6.fasta**: Includes an eye icon, a pencil icon, and a delete icon.
- 1: 1 ATCACG.s 6 sequen ceBCs appended.fasta**: Includes an eye icon, a pencil icon, and a delete icon.

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

Running workflow "Tutorial Workflow"

Step 1: Input dataset

Input Dataset

1: 1_ATCACG.s_6_sequenceBCsappended.fastq

type to filter

Step 2: Input dataset

Input Dataset

2: Schizosaccharomyces_Pombe_tetarr6.fasta

type to filter

Step 3: FASTQ Groomer (version 1.0.4)

Step 4: Bowtie2 (version 0.2)

Send results to a new history

Run workflow

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



Successfully ran workflow "Tutorial Workflow". The following datasets have been added to the queue:

- 1: 1_ATCACG.s_6_sequenceBCsappended.fastq
- 2: Schizosaccharomyces_Pombe_tetarr6.fasta
- 3: FASTQ Groomer on data 1
- 4: Bowtie2 on data 2 and data 3: aligned reads

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/>

