

An introduction to Galaxy

Sebastian Schmeier

*Institute of Natural and Mathematical Sciences
Massey University Auckland, New Zealand*
<http://sebscientific.org>
s.schmeier@gmail.com

2015-08-11

Contents

Galaxy Introduction	3
1.1 Overview	3
1.2 How to get access to Galaxy	3
1.3 The user interface	4
1.3.1 Basics	4
1.3.2 User accounts	5
1.4 A word on tools	6
2.1 A simple example	6
2.2 Loading your own data	6
2.3 Loading data from the web	10
2.4 Loading shared data	11
2.5 Working with data	13
2.5.1 Renaming files	14
2.5.2 Removing unwanted information	15
2.5.3 Creating flanking regions	16
2.5.4 Filter data	17
2.5.5 Joining/intersecting data sets	18
2.6 Visualising data sets	20
2.7 Another word on the history	22
2.7.1 Saved histories	22
2.7.2 Sharing a history	23
2.8 Workflows	24
2.8.1 Creating workflows	24
2.8.2 Editing workflows	26
2.8.3 Applying workflows to your data	30
2.9 References	31
3.0 Web links	31

Galaxy Introduction

1.1 Overview

In this brief tutorial we will learn how to use the excellent tool [Galaxy](http://galaxyproject.org/) (<http://galaxyproject.org/>) to analyze biological data. We will see how it [Galaxy](#) allows you to make use of a number of tools in a simple to use graphical interface (more on that in a moment). A user is thus not required to use any of the tools on the command-line (even though many of the integrated tools were developed for the command-line in the first place) but can fully use and control the integrated tools with the mouse pointer. In addition, it also allows developers of tools to easily integrate them into a graphical user interface system that is already known to many scientists and thus make the tools available for the research community.

Another big advantage of [Galaxy](#) is that every step of the analysis is monitored and accessible via a history. This makes reproducible research not only a possibility but also easy to facilitate. Steps from the history can be packaged into work-flows, which can be reused with different data or shared with other scientists.



Figure 1: Galaxy Community Conference 2015

[Galaxy](#) enjoys a large and growing user and developer base, which is evident by its own yearly [conference](#) (see *Figure 1*, <http://gcc2015.tsl.ac.uk/>) and participation in [Google Summer of Code](#). It is relatively easy to find help should one need it, e.g. through their [mailing list](#) or [wiki](#) (<http://wiki.galaxyproject.org/>). Also, many commercial companies that provide next-generation sequencing services, provide Galaxy instances to analyze your data (e.g. we at [New Zealand Genomics Limited](#) (<http://nzgenomics.co.nz>) have a full fledged installation on our infrastructure ready for scientist to be used).

1.2 How to get access to Galaxy

There many option available to either give [Galaxy](#) a test run or do a full analysis with it. There is a ever growing list of public servers [available](#), some of which might have certain restrictions, e.g. maximum data-file size, etc. The standard server is accessible at: <https://usegalaxy.org/>

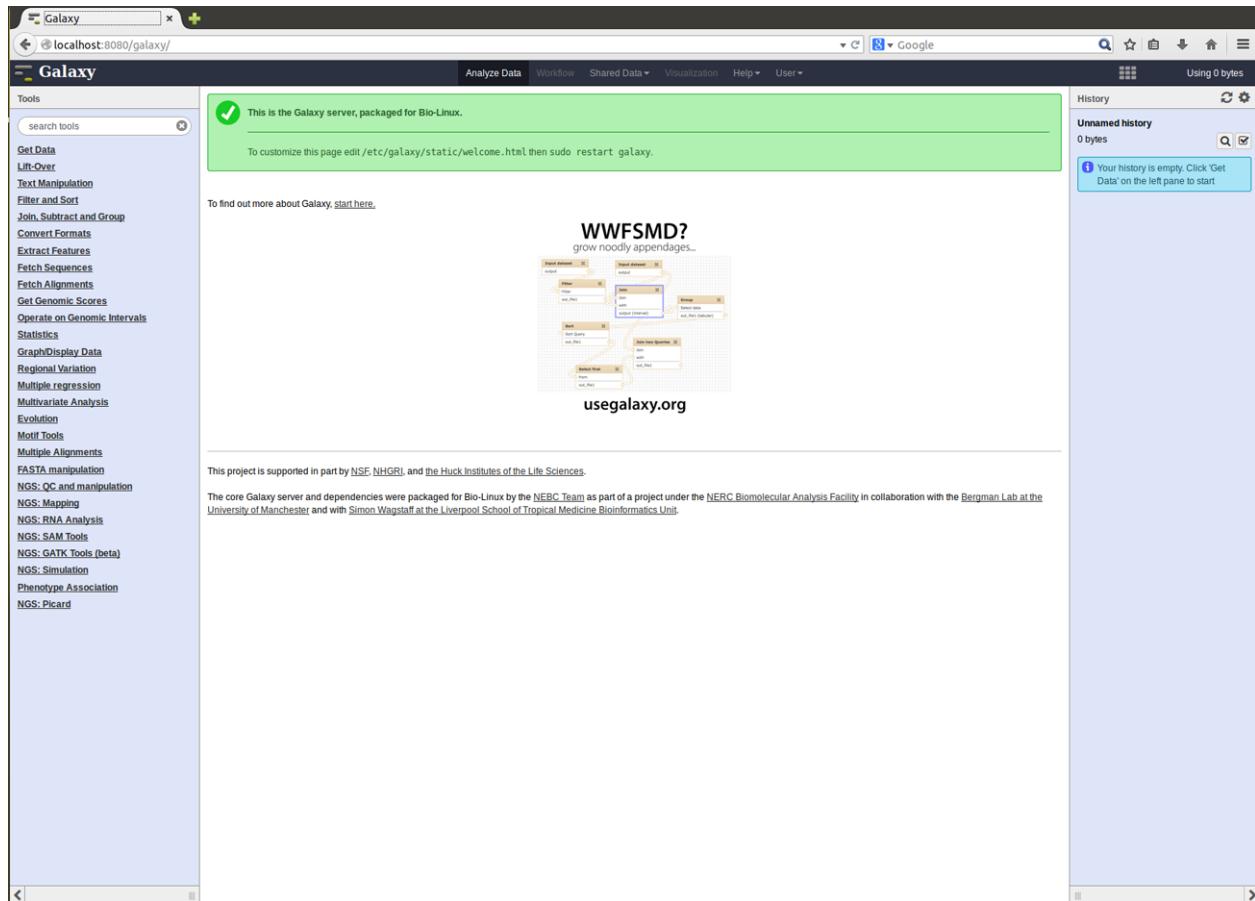
You can start your own [Galaxy](#) instances on [Cloud](#) infrastructure, e.g. [Amazon Cloud Services](#), should you have bigger analysis needs that you want to perform in the cloud.

You can [download](#) and install [Galaxy](#) on your own machine or server, even integrating a computer cluster on the back-end.

You can install [BioLinux](#) on your own machine or run [BioLinux](#) as a virtual machine and you are set as well, as [Galaxy](#) comes pre-installed on [BioLinux 8](#).

1.3 The user interface

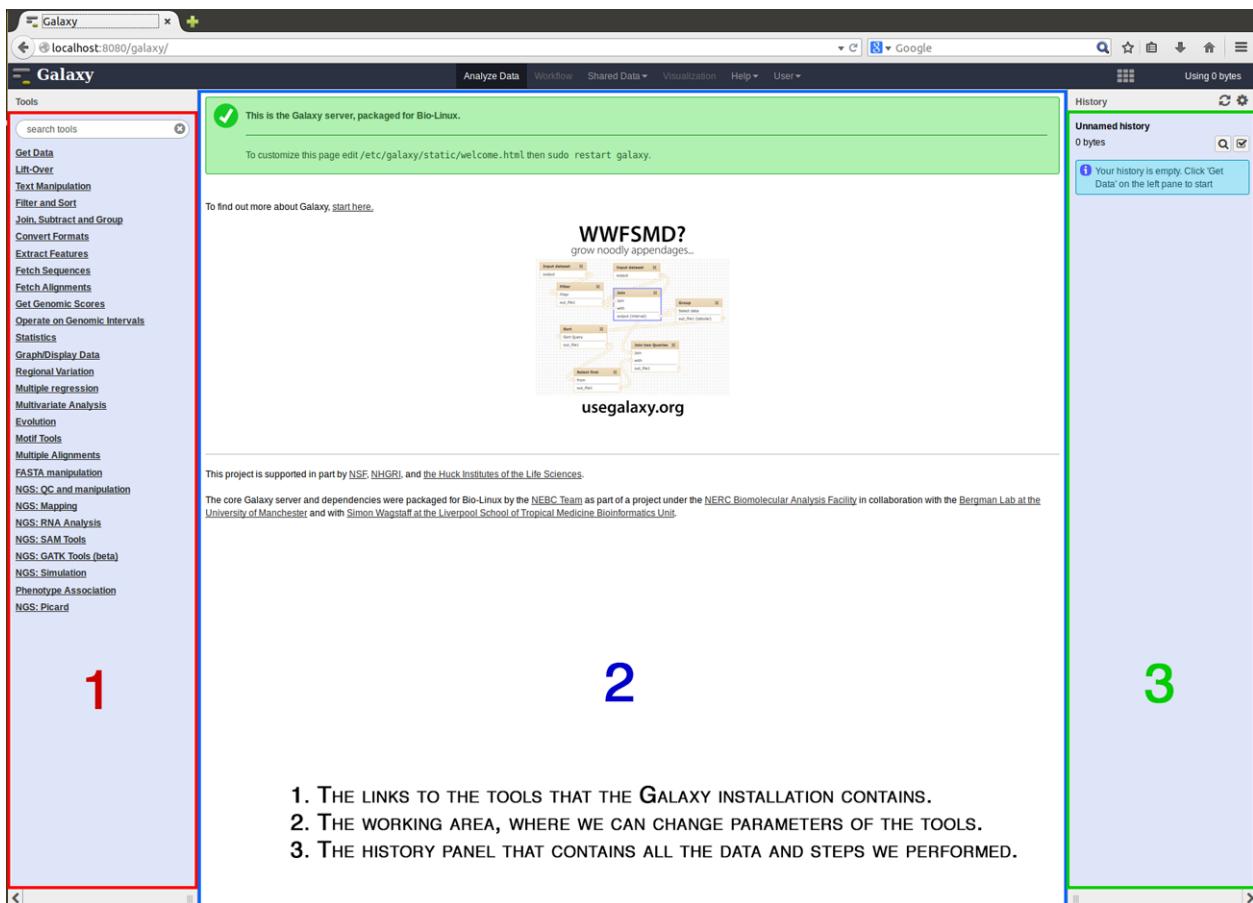
1.3.1 Basics



Hint! Click on the [Galaxy](#) screenshots to get a bigger version!

There are 3 areas of interest for now:

1. The links to the tools that the [Galaxy](#) installation contains (this can vary from [Galaxy](#) instance to instance).
2. The working area, where we can change parameters of the tools that we want to use for some of our data.
3. The history panel that contains all the data and steps we performed on the data.



1. THE LINKS TO THE TOOLS THAT THE GALAXY INSTALLATION CONTAINS.
2. THE WORKING AREA, WHERE WE CAN CHANGE PARAMETERS OF THE TOOLS.
3. THE HISTORY PANEL THAT CONTAINS ALL THE DATA AND STEPS WE PERFORMED.

1.3.2 User accounts

If you plan to use the public available Galaxy instance at <https://usegalaxy.org/>, it is a good idea to create a user account. This is relatively straight forward, just click on **User** in the top panel and then **Register** (1). This will allow you, amongst other things, to save histories, but more on this in later (2.7).

The screenshot shows the Galaxy web interface with the "User" dropdown menu open. The "Register" option is highlighted with a red box and labeled "1". Other options in the menu include "Login" and "Galaxy start here".

1.4 A word on tools

The tools that you find in the tools area of the [Galaxy](#) instance are nothing else than programs that were originally written for the command-line. As long as you have/write a program that expects a input-file and out-put-file as command-line arguments, it is quite easy to [integrate a tool](#) into an local [Galaxy](#) installation.

Attention! The tools that you find in your [Galaxy](#) instance might differ depending on where you access the particular [Galaxy](#) installation/instance., e.g. you might find a different toolset at the standard online [Galaxy](#) instance at <https://usegalaxy.org/>, than on your local installation.

2.1 A simple example

The purpose in this example is not to find anything of biological relevance but rather to:

1. Understand the [Galaxy](#) system
2. Understand how to get your data of interest into the system
3. Understand how to do simple data manipulation tasks
4. Understand how the [Galaxy](#) History system works
5. Understand how to set up a workflow and run your data through it

In order to develop the understanding of the five points above, we are going through a simple example:

"We want to find the mouse chromosome X genes that have single nucleotide polymorphism in their upstream regions"

The tasks required to find those mutations are:

1. Get single nucleotide polymorphism (SNP) data for chromosome X
 2. Get all gene locations on chromosome X
 3. Get upstream regions of the genes
 4. Overlap the SNPs with the genic upstream regions
 5. Visualise results in a genome browser
-

2.2 Loading your own data

Download the following file to your computer: [*mm9_chrX_SNP128_set.bed*](#). The file is in [bed-format](#), a simple tab-separated format containing 6 columns: **chromosome, start, stop, name, score, strand**.

Hint! Bed-format files can have more or less columns. However, the first three columns are the bare minimum.

1. On your [Galaxy](#) window go to the upper left in the tools area and click on **Get Data**. A subsection of **Get Data** will open and show available option for you to get data into the [Galaxy](#) system.
2. Choose **Upload File from your computer**.

The screenshot shows the Galaxy web interface. On the left, a sidebar titled 'Tools' contains a list of services: UCSC Main table browser, UCSC Archaea table browser, EBI SRA ENA SRA, BioMart Central server, GrameneMart Central server, Flymine server, modENCODE fly server, modENCODE modMine server, MouseMine server, Ratmine server, YeastMine server, modENCODE worm server, WormBase server, ZebrafishMine server, EuPathDB server, and GenomeSpace import from file browser. A red box labeled '1' highlights the 'Get Data' link, and another red box labeled '2' highlights the 'Upload File from your computer' button. The main content area features a large banner for 'Galaxy 101' with the subtext 'Start small' and 'The very first tutorial you need'. Below the banner is a series of small circular icons. At the bottom right of the main area, there is a link to 'Tweets by @galaxyproject'.

1. An additional window should open that allows you to select the your file.
2. You can specify the species, given that we are looking at mouse data from mm9 set it to the same.

The screenshot shows the 'Get Data' tool dialog box. It has a title bar 'Download data directly from web or upload files from your disk'. Below the title bar is a large dashed rectangular area with the placeholder text 'You can Drag & Drop files into this box.'. To the left of this area is a vertical list of services identical to the one in the main Galaxy interface. At the bottom of the dialog are several buttons: 'Type (set all):' with a dropdown menu ('Auto-detect'), a search icon, '1 Choose local file' (highlighted with a red box), 'Genome (set all):' with a dropdown menu ('Additional Species ...') (highlighted with a red box), and a row of buttons for 'Paste/Fetch data', 'Start', 'Pause', 'Reset', and 'Close'.

Once you hit the **Start** button, your data/analysis will be uploaded. In your history your data goes through three stages indicated by three different colors:

1. Grey: Scheduled for uploading/running
2. Yellow: Currently running
3. Green: Dataset/analysis is ready

The screenshot shows the Galaxy web interface at localhost:8080/galaxy/. The history panel on the right contains three entries:

- 1** 19: mm9_chrX_SNP12_8_set.bed (Grey background)
- 2** 19: mm9_chrX_SNP12_8_set.bed (Yellow background)
- 3** 19: mm9_chrX_SNP128_set.bed (Green background)

A legend on the left side of the interface defines the colors:

- Grey: This is the Galaxy server, packaged for Bio-Linux.
- Yellow: To customize this page edit /etc/galaxy/static/welcome.html then sudo restart galaxy.
- Green: To find out more about Galaxy, [start here](#).

The legend also includes a small icon of a DNA double helix.

1. Click on the filename and you get some information about the data.
2. Here you will see information like how many regions (lines) are in the file, the format and genome
3. Here you can download the data, get even more information about the data and run the job again (here it would reload the data)

The screenshot shows the Galaxy web interface. On the left, there's a sidebar with a 'Tools' section containing various bioinformatics tools like 'Get Data', 'Lift-Over', 'Text Manipulation', etc. The main area displays a 'Galaxy 101' tutorial slide with the title 'Start small' and the subtitle 'The very first tutorial you need'. To the right is the 'History' panel, which lists datasets. One dataset, '19: mm9_chrX_SNP128_set.bed', is highlighted with a red box and labeled '1'. Below it, another dataset, '20,000 regions format: bed, database: mm9', is also highlighted with a red box and labeled '2'. At the bottom of the history panel, there are three buttons: a magnifying glass icon (labeled '3'), a 'View' link, a 'Current' link, and a 'Main' link.

Within the history panel and your data set there are several buttons of importance. The first one which looks like an eye will display you data in the working area.

This screenshot shows the Galaxy interface with a data table in the center. The table has columns labeled 1 through 7 and contains genomic data. To the right is the 'History' panel, which shows the same dataset '19: mm9_chrX_SNP128_set.bed' as in the previous screenshot. A red box highlights the eye icon in the history panel, labeled '1'. Below the history panel, there are three buttons: a magnifying glass icon (labeled '2'), a 'View' link, and a 'Current' link.

1. The second button will allow you to edit your data
2. You can change the file-name
3. Change the assignment of column numbers to particular properties
4. and finally save your changes.

The screenshot shows the Galaxy web interface. On the left, there's a sidebar with various tools like 'Get Data', 'UCSC Main table browser', and 'BioMart Central server'. The main area has tabs for 'Attributes', 'Convert Format', 'Datatype', and 'Permissions'. Under 'Attributes', there's a form for 'Edit Attributes' with fields for 'Name' (set to 'mm9_chrX_SNP128_set.bed'), 'Info' (set to 'uploaded bed file'), 'Database/Build' (set to 'Mouse July 2007 (NCBI37/mm9) (mm9)'), and 'Number of comment lines'. A large red box highlights the 'Chrom column' dropdown set to '1', the 'Start column' dropdown set to '2', the 'End column' dropdown set to '3', and the 'Strand column' dropdown set to '6'. Another red box highlights the 'Score column for visualization' dropdown with '1', '2', '3', and '4' checked. At the bottom right of the dialog is a 'Save' button with a red box around it. To the right, the 'History' panel shows a list of datasets: 'Unnamed history' (1 shown, 18 deleted, 1 hidden), '29.1 MB', and '19: mm9_chrX_SNP128_set.bed' (highlighted with a green box). The 'mm9_chrX_SNP128_set.bed' entry in the history has a red box around its delete icon.

The last button can delete your data/analysis again from the history panel.

This screenshot is similar to the previous one but shows the changes after saving. The 'Edit Attributes' dialog now shows the saved name 'mm9_chrX_SNP128_set.bed' in the 'Name' field. The history panel on the right still shows the dataset '19: mm9_chrX_SNP128_set.bed' with a red box around its delete icon.

2.3 Loading data from the web

Now we are focusing on getting some data from the [UCSC table browser](#). Many people UCSC were quite busy integrating lots of data and there is plenty of data available especially for mammalian model systems.

1. On your [Galaxy](#) window go to the upper left in the tools area and click on **Get Data**. A subsection of **Get Data** will open and show available option for you to get data into the [Galaxy](#) system.

2. Click on **UCSC Main table browser**. This will open the [UCSC table browser](#) in your [Galaxy](#) working area.
3. Here you can choose the genome that you want the data from, we will choose mm9
4. Here you can choose the kind of data that you which to download from the particular genome, we will choose here the **Genes and Gene Prediction group** and the **UCSC Genes** as well as the **knownGene** table. The **describe table schema** button will get you to another webpage that describes the data within the **knownGene** table. Feel free to explore.
5. Here you can chose if you which to download data from the whole genome or a subportion of it. We will choose here only data from **chrX** type this in the field and hit **lookup** button which will complete the start and stop coordinates of the genome.
6. Here we can specify the output-format. It is important here to make sure that the **Send output to Galaxy** choice is selected . Also, we want BED-format again.
7. After we are finsihed we can hit the **get output** button, after which our requested data will be loaded into the [Galaxy](#) interface.

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

3 clade: Mammal genome: Mouse assembly: July 2007 (NCBI37/mm9)
 4 group: Genes and Gene Predictions track: UCSC Genes
 5 table: knownGene
 6 region: position chrX:1-166650296 lookup
 7 output format: BED – browser extensible data Send output to Galaxy GREAT
 identifiers (names/accessions): paste list upload list
 filter: create
 intersection: create
 correlation: create
 output file: (leave blank to keep output in browser)
 file type returned: plain text gzip compressed
 get output summary/statistics

Finally, your data should appear in the right hand side history panel.

2.4 Loading shared data

Another way of loading data into your history panel is by loading data that was shared with you through [Galaxy](#). On the upper panel click on **Shared Data** and then on **Data Libraries**.

The screenshot shows the Galaxy web interface. The 'Shared Data' menu is open, with 'Data Libraries' selected. The search bar at the top has 'mouse' typed into it. The history panel on the right shows several datasets related to mouse analysis, including 'UCSC Main on Mouse' and 'mm9_chrX_SNP128_s et.bed'.

Here you will find a search field to search for available datasets. Search for mouse because currently we are working with mouse data.

The screenshot shows the 'Data Libraries' page in Galaxy. A search bar at the top has 'mouse' typed into it. Below the search bar, there is a table with two columns: 'Data library name' and 'Data library description'. The first row shows '1000 Genomes' with a description of 'Data from the 1000 Genomes Project FTP site'. The second row shows 'AC-exome' with a description of 'Data for two papers about the Khoisan and other populations.' The third row is partially visible with 'Bushman'.

Data library name	Data library description
1000 Genomes	Data from the 1000 Genomes Project FTP site
AC-exome	
Bushman	Data for two papers about the Khoisan and other populations.

Choose the **ChIP-Seq Mouse Example** dataset from the ENCODE project. This is data of chromatin immunoprecipitation followed by sequencing to find regions in the genome where transcription factors bind.

The screenshot shows the 'Data Libraries' page in Galaxy. A search bar at the top has 'mouse' typed into it. Below the search bar, there is a table with two columns: 'Data library name' and 'Data library description'. The first row shows 'ChIP-Seq Mouse Example' with a description of 'Data used in examples that demonstrate analysis of ChIP-Seq data'.

Data library name	Data library description
ChIP-Seq Mouse Example	Data used in examples that demonstrate analysis of ChIP-Seq data

Here you see an overview of the datasets available. You can choose the dataset, select **Import to current history**, and hit **Go**.

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

Data Library “ChIP-Seq Mouse Example”

Use this data to test out and learn Galaxy's ChIP-Seq capabilities. It has been scaled down to relatively small sizes. These files are from this mouse ChIP-SEQ experiment in the ENCODE project. These data were generated and analyzed by the labs of Michael Snyder at Stanford University and Sherman Weissman at Yale University. The original files from ENCODE were too large to use in teaching examples, so they have been reduced to contain only data that corresponds to chromosome 19 (the shortest). These files were created by, well, cheating. We first processed the entire dataset, mapping it to MM9. When went back and extracted from the original datasets only those records that eventually mapped to chromosome 19.

Name	Message	Data type	Date uploaded	File size
Mouse ChIP-Seq example Control Data, chr19, mm9	Control file for mouse ChIP-Seq example. An ungroomed Illumina FASTQ file, it contains only reads that map to chr19, mm9.	fastq	Mon Sep 19 20:01:54 2011 (UTC)	84.1 MB
Mouse ChIP-Seq Example Experimental Data, chr19, mm9	Experimental results for mouse ChIP-Seq example. An ungroomed Illumina FASTQ file that contains only reads that map to chr19, mm9.	fastq	Mon Sep 19 20:07:43 2011 (UTC)	47.4 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

Once the data is loaded in your history [Galaxy](#) will inform you. You can get back to your working area by clicking on [Analyze Data](#).

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

Data Library “ChIP-Seq Mouse Example”

1 dataset imported into 1 history: Unnamed history

Use this data to test out and learn Galaxy's ChIP-Seq capabilities. It has been scaled down to relatively small sizes. These files are from this mouse ChIP-SEQ experiment in the ENCODE project. These data were generated and analyzed by the labs of Michael Snyder at Stanford University and Sherman Weissman at Yale University. The original files from ENCODE were too large to use in teaching examples, so they have been reduced to contain only data that corresponds to chromosome 19 (the shortest). These files were created by, well, cheating. We first processed the entire dataset, mapping it to MM9. When went back and extracted from the original datasets only those records that eventually mapped to chromosome 19.

Name	Message	Data type	Date uploaded	File size
Mouse ChIP-Seq example Control Data, chr19, mm9	Control file for mouse ChIP-Seq example. An ungroomed Illumina FASTQ file, it contains only reads that map to chr19, mm9.	fastq	Mon Sep 19 20:01:54 2011 (UTC)	84.1 MB
Mouse ChIP-Seq Example Experimental Data, chr19, mm9	Experimental results for mouse ChIP-Seq example. An ungroomed Illumina FASTQ file that contains only reads that map to chr19, mm9.	fastq	Mon Sep 19 20:07:43 2011 (UTC)	47.4 MB

For selected datasets: Import to current history Go

You can get rid of the dataset again in your history as it will not be used anymore in theis tutorial.

2.5 Working with data

The aim here is to get understand how [Galaxy](#) can help you to prepare your data to be able to analyze it further. We will perform some easy tasks like removing redundant information, renaming new datasets, sub-selecting regions of interest, extending our genomic regions to look at promoters upstream of genes, finding the SNPs from our set that overlap the promoter regions.

2.5.1 Renaming files

You should aim at naming your files in a manner that they are easily recognizable. This is especially important once we manipulate them and create new files. You should make it a habit of renaming a file after it was created to keep track of what they are.

1. Click on the **edit icon** of the file you wish to change.
2. Type a new filename in the **Name** field.
3. Click on the **Save** button

The screenshot shows the Galaxy web interface. On the left, the 'Tools' sidebar is visible with various bioinformatics tools listed. In the center, the 'Edit Attributes' dialog is open for a dataset named 'mm9_knownGene_chrX'. A red box labeled '2' highlights the 'Name' input field. To the right, the 'History' panel displays a list of datasets. One dataset, '24: UCSC Main on Mouse knownGene (chrX:1-166650296)', is highlighted with a red box labeled '1'. Below it, the dataset's details are shown, including its format (bed) and database (mm9). A preview of the dataset's content is provided, showing genomic coordinates for chromosomes X and Y.

Attention! I also renamed the data ***Mouse ChIP-Seq example Control Data, chr19, mm9*** to -> ***mm9_ChIP_chr19_control*** and the data ***mm9_chrX_SNP128_set.bed*** to -> ***mm9_chrX_SNP128***.

The screenshot shows the Galaxy web interface. The main page features a large banner with the text 'Running Your Own'. The 'History' panel on the right lists three datasets: '26: mm9_ChIP_chr19_control', '24: mm9_knownGene_chrX', and '19: mm9_chrX_SNP128'. These datasets are highlighted with a red box. The 'History' panel also includes links to 'display in IGB View', 'display at Ensembl Current', and 'display at UCSC main'.

Attention! The numbering of the datasets here might be different from yours depending on how many datasets you have been working on before. The image above shows **24: mm9_knownGene_chrX**, however, this may vary for you (and might vary in what follows here as I might have done this tutorial in multiple sessions.). This is one reason why it is a good idea to rename the dataset.

2.5.2 Removing unwanted information

Our gene BED-file that we retrieved from **UCSC table browser** is in BED 12 format, e.g. it contains 12 columns, but only the first 6 are necessary for our purposes. Thus, we aim at removing the extra columns to make the file more readable. Let's do this by

1. Clicking on the **Text manipulation** tools section
2. Selecting the **Cut** tool.
3. Insert the columns you want to retain. We want the first 6 columns.
4. Choose the right file to do the manipulation on
5. **Execute** the tool

The screenshot shows the Galaxy web interface. On the left, the 'Tools' sidebar is open, showing various categories like 'Get Data', 'Send Data', 'Lift-Over', and 'Text Manipulation'. The 'Text Manipulation' category is highlighted with a red box and labeled '1'. Below it, the 'Cut columns from a table' tool is selected and highlighted with a red box, labeled '2'. In the main workspace, the 'Cut columns from a table' tool is displayed. The 'Cut columns' input field contains 'c1,c2,c3,c4,c5,c6' and is highlighted with a red box, labeled '3'. The 'From' dropdown menu shows '24: mm9_knownGene_chrX' and is highlighted with a red box, labeled '4'. At the bottom of the tool panel is a blue 'Execute' button, which is also highlighted with a red box, labeled '5'. To the right of the tool panel is the 'History' panel, which lists several datasets including 'Bioinf-course1', '26: mm9_ChIP_chr1_9_control', '24: mm9_knownGen_e_chrX', '19: mm9_chrX_SNP1', and '28'. The entire interface has a light blue background.

You should see a new file in the history. Here it is being scheduled for execution and should be green once the job is finished. Please rename the resulting dataset to $\rightarrow \text{mm9_knownGene_chrX_short}$.

The screenshot shows the Galaxy web interface. On the left, the 'Tools' sidebar lists various genomic manipulation tools. In the center, a success message box indicates a job has been added to the queue. Below it, a detailed description of the job '27: Cut on data 24' is provided. On the right, the 'History' pane shows a list of datasets, with the most recent one, '27: Cut on data 24', highlighted with a red border.

Dataset ID	Dataset Name	Actions
26:	mm9_ChIP_chr1_9_control	
24:	mm9_knownGene_chrX	
19:	mm9_chrX_SNP1_28	

2.5.3 Creating flanking regions

Because we are interested to look in the promoter regions of our genes we need to extract those. We here define the promoter as upstream regions from the transcription start site.

1. Find the **Operate on Genomic Intervals** sections
2. Select the **Get flanks** tool
3. Choose the right dataset: **mm9_knownGene_chrX_short**
4. The region we are interested in is **Around Start**
5. We want the **Upstream** region
6. We want **5000** bases upstream
7. **Execute**

Step 1: Select data (highlighted in red)

Step 2: Region (highlighted in red)

Step 3: Location of the flanking region/s (highlighted in red)

Step 4: Length of the flanking region(s) (highlighted in red)

Step 5: Execute (highlighted in red)

Attention! I renamed the resulting dataset -> **mm9_chrX_promoter**

2.5.4 Filter data

Filtering data can be done in many different ways, however, here we use the **filter** tool.

1. Find the **Filter and Sort** tool section
2. Select the **Filter** tool
3. Select our promoter dataset: **mm9_chrX_promoter**
4. We only want promoter within the first 8000000 bases, the start position of genes is specified in the second column (**c2**)
5. **Execute**

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User Using 0%

Tools

search tools

Get Data

Send Data

Lift-Over

Text Manipulation

Convert Formats

Filter and Sort 1

Histogram of a numeric column

Filter data on any column using simple expressions 2

Sort data in ascending or descending order

Select lines that match an expression

GFF

Extract features from GFF data

Filter GFF data by attribute using simple expressions

Filter GFF data by feature count using simple expressions

Filter GTF data by attribute values_list

Join, Subtract and Group

NGS: QC and manipulation

NGS: Mapping

NGS: BAM Tools

NGS: Picard

NGS: VCF Manipulation

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Filter data on any column using simple expressions (Galaxy) Tool Version 1.1.0

Filter

15: mm9_chrX_promoter 3

Dataset missing? See TIP below.

With following condition

c2<8000000| 4

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

0

Execute 5

TIP: Double equal signs, ==, must be used as "equal to" (e.g., c1 == 'chr22')

TIP: Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

TIP: If your data is not TAB delimited, use Text Manipulation->Convert

Syntax

The filter tool allows you to restrict the dataset using simple conditional statements.

Columns are referenced with c and a number. For example, c1 refers to the first column of a tab-delimited file

Make sure that multi-character operators contain no white space (e.g., <= is valid while < = is not valid)

When using 'equal-to' operator double equal sign '==' must be used (e.g., c1=='chr1')

Non-numerical values must be included in single or double quotes (e.g., c6=='+')

Filtering condition can include logical operators, but make sure operators are all lower case (e.g. (c1=='chrX' and c1=='chrY') or not c6=='+')

History

search datasets

Bioinf-course 1 5 shown, 11 deleted 203.4 MB

15: mm9_chrX_promoter 2,021 regions Edit attributes

format: interval, database: mm9

Location: Upstream, Region: start, Flank-length: 5000, Offset: 0

display at Ensembl Current display at UCSC main

1.Chrom	2.Start	3.End	4.Name	5.6.Str
chrX	3243629	3248629	uc009skj.1	0 -
chrX	3405667	3410667	uc009skk.1	0 +
chrX	3463320	3468320	uc009skl.1	0 -
chrX	3547091	3552091	uc012hdv.1	0 -
chrX	3667437	3672437	uc009skm.1	0 -
chrX	3743193	3748193	uc009skn.1	0 +

14: mm9_knownGene_chrX_s

13: mm9_ChIP_chr19_control

11: mm9_knownGene_chrX

10: mm9_chrX_SNP128

Attention! I renamed the resulting dataset -> **mm9_chrX_promoter_8000000**

Hint! If you click on the dataset name it will also tell you how many lines where extracted from the original dataset.

2.5.5 Joining/intersecting data sets

Lets find those mutations that overlap our promoter subset.

1. Find the **Operate on genomic Intervals** tool section
2. Select the **Join** tool
3. Select our SNP data **mm9_chrX_SNP128** and the promoter dataset **mm9_chrX_promoter_8000000**
4. **INNER JOIN**
5. **Execute**

1

2

3

4

5

Attention! I renamed the resulting dataset -> ***SNPs_at_promoter***.

If you temporarily close the history tab we can have a closer look at the resulting dataset.

1	2	3	4	5	6	7	8	9	10	11	12
chrX	3243722	3243723	rs52395861	0	+	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3244443	3244444	rs46254379	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3244471	3244472	rs50874688	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3244489	3244490	rs33264252	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3244489	3244490	rs46879180	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3244555	3244556	rs48315292	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3244566	3244567	rs46735700	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3245984	3245985	rs51980349	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3246069	3246070	rs50772248	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3248007	3248008	rs51574865	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3248008	3248009	rs47066278	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3248012	3248013	rs49511429	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3248026	3248027	rs50458919	0	+	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3248050	3248051	rs51752810	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3248122	3248123	rs45894481	0	-	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3248431	3248432	rs51916321	0	+	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3248433	3248434	rs47111988	0	+	chrX	3243629	3248629	uc009skj.1	0	-
chrX	3248500	3248501	rs50501061	0	+	chrX	3243629	3248629	uc009skj.1	0	-

We see that we have 2,218 SNPs overlapping promoter regions in the genes in the first 8,000,000 base pairs. The **Join** tool put the overlapping elements right next to each other.

Note! that for one particular promoter we can have several SNPs (1).

2.6 Visualising data sets

Now that we basically have what we are looking for we want to visualise our found SNPs and the promoter that have mutations in an intuitive manner. Here, Genome Browsers come in that are helpful in getting an overview. In this section we prepare the data we would like to visualise and prepare a custom track for the [UCSC Genome Browser](#). First, what data do we want to visualise:

1. All SNPs
2. The SNPs that overlap our promoter regions
3. The promoter regions

To create a new track that we can visualise in USCS, do the following:

1. Find the **Graph/Display Data** tool section
2. Select the **Build custom track** tool
3. Click on insert track and select our promoter data ***mm9_chrX_promtoer_8000000***.
4. Give it a unique name
5. Insert more tracks for data like ***SNPs_at_promoter*** and ***mm9_chrX_SNP128***.
6. **Execute**

Attention! Make sure to use **unique names** for each track, because if you use the same name twice the last track overwrites the one from before.

The screenshot shows the Galaxy web interface with the following details:

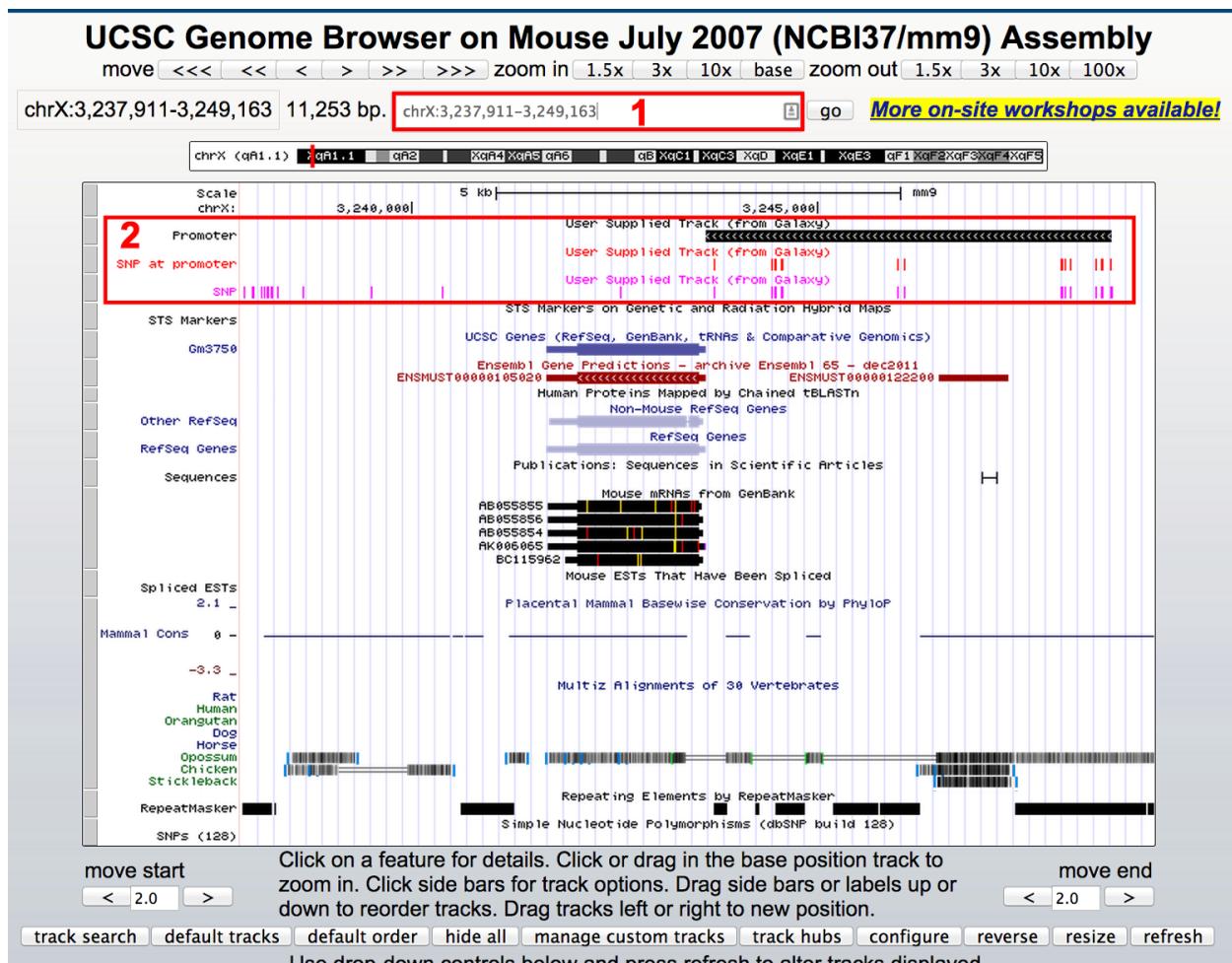
- Tools Sidebar:** Shows various genomic analysis tools like Convert Formats, Filter and Sort, Join, Subtract and Group, NGS: QC and manipulation, NGS: Mapping, NGS: BAM Tools, NGS: Picard, NGS: VCF Manipulation, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data (highlighted with red box 1), Build custom track for UCSC genome browser (highlighted with red box 2), Scatterplot of two numeric columns, Histogram of a numeric column, Plotting tool for multiple series and graph types, GMAI Multiple Alignment Viewer, Boxplot of quality statistics, VCF to MAF Custom Track for display at UCSC, and Phenotype Association.
- Tool Panel:** The 'Build custom track for UCSC genome browser' tool is selected. It has the following fields:
 - Dataset:** A dropdown menu showing '19: mm9_chrX_promtoer_8000000' (highlighted with red box 3).
 - Name:** A text input field containing 'Promoter' (highlighted with red box 4).
 - Description:** A text input field containing 'User Supplied Track (from Galaxy)'.
 - Color:** A dropdown menu set to 'Black'.
 - Visibility:** A dropdown menu set to 'Dense'.
- Action Buttons:** Two buttons at the bottom: '+ Insert Track' (highlighted with red box 5) and 'Execute' (highlighted with red box 6).
- Help and Notes:**
 - A note: 'This tool allows you to build custom tracks using datasets in your history for the UCSC genome browser. You can view these custom tracks on the UCSC genome browser by clicking on display at UCSC main/test link in the history panel of the output dataset.'
 - A warning: 'Please note that this tool requires all input datasets(tracks) to have the same genome build. The tool throws an error when this requirement is not met. You may then have to choose a valid dataset or remove invalid tracks.'
- History Panel:** Shows a list of datasets in the history, including '203.7 MB' (21: SNPs_at_promoter, 19: mm9_chrX_promtoer_8000000, 18: mm9_chrX_promoter, 14: mm9_knownGene_chrX_short, 13: mm9_ChIP_chr19_control, 11: mm9_knownGene_chrX, 10: mm9_chrX_SNP128).

Once you hit the **Execute** button you should have a new track created which is visible in the history panel (1). Click on the name of that track and click **display at UCSC main** (2).

The screenshot shows the Galaxy web interface with a focus on building a custom track. On the left, a sidebar titled 'Tools' lists various bioinformatics tools under categories like 'Convert Formats', 'Filter and Sort', and 'Graph/Display Data'. In the center, a table displays genomic data with columns 1, 2, 3, and 4. To the right, the 'History' panel shows a dataset named 'Bioinf-course 1' containing 204.5 MB of data. A specific entry in the history is highlighted with a red box and labeled '1', showing details about a custom track named '22: Build custom track on data 10, data 21, and data 19'. Below this, another red box labeled '2' highlights the 'display at UCSC main' link. The bottom part of the history panel shows the raw track definition code.

1	2	3	4
chrX	3243629	3248629	0
chrX	3405667	3410667	1
chrX	3463320	3468320	2
chrX	3547091	3552091	3
chrX	3667437	3672437	4
chrX	3743193	3748193	5
chrX	3902010	3907010	6
chrX	3995573	4000573	7
chrX	4069963	4074963	8
chrX	4441526	4446526	9
chrX	5046383	5051383	10
chrX	5241184	5246184	11
chrX	5619624	5624624	12
chrX	5660538	5665538	13
chrX	5660538	5665538	14
chrX	5750109	5755109	15
chrX	5897841	5902841	16
chrX	5972262	5977262	17
chrX	6149403	6154403	18
chrX	6351431	6356431	19
chrX	6618745	6623745	20
chrX	6618745	6623745	21

If you do so, a new window at the UCSC Genome browser will open. Put `chrX:3,237,911-3,249,163` in the search bar (1) and you will see a postion that shows what is going on. Right on top should be your three tracks located (2). You can scroll left and right, zoom in and out to get to other promoter regions. You can also change the resolution at which your features will be shown. Many other tracks from UCSC are also shown automatically and ad the bottom of the page you can chose to show or hide other tracks of interest.



2.7 Another word on the history

2.7.1 Saved histories

You are able to create an account on the public Galaxy [web-server](#). Once done, you will be able to save histories and fetch your old histories back. In this manner you are also able to save whole work-flows but more on that later.

For now you can look at your **Saved Histories** by clicking the config button in the upper right.

The screenshot shows the Galaxy web interface. On the left, there's a sidebar titled "Tools" with a search bar and a list of tool categories. The main area displays a welcome message about Galaxy and a title card for a tutorial: "Running Your Own Understanding how Galaxy works An in-depth tutorial". On the right, the "History" panel is open, showing a list of histories. A red box highlights the "Saved Histories" link under the "HISTORY LISTS" section.

You will see only one history the one we are currently working on. You can rename the history by clicking the name in the history panel or by doing a rename in the working area.

This screenshot shows the "Saved Histories" panel. It lists a single history named "Bioinf-course1". A red box highlights the history name "Bioinf-course1" in the list. The panel also includes a search bar, a table for managing histories, and a note about deleted histories.

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated
Unnamed history	3	0 Tags		209.8 MB	Apr 15, 2015	~21 min ago

2.7.2 Sharing a history

It is easy to share a saved history with colleagues or make them public (1). Several options are available.

The first screenshot shows the 'Saved Histories' page. A context menu is open over a history named 'Bioinf-course 1'. The menu items are: Switch, View, Share or Publish (highlighted with a red box), Copy, Rename, Delete, and Delete Permanently. A red number '1' is placed next to the 'Share or Publish' option.

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated↑	Status
Bioinf-course 1	0 Tags			204.5 MB	~1 day ago	~20 hours ago	current history

The second screenshot shows the 'Share or Publish History' page for 'Bioinf-course 1'. It has two main sections: 'Make History Accessible via Link and Publish It' and 'Share History with Individual Users'. Under 'Make History Accessible via Link', there is a button 'Make History Accessible via Link' which generates a shareable link. Under 'Share History with Individual Users', there is a button 'Share with a user'.

2.8 Workflows

2.8.1 Creating workflows

It is possible to create workflows out of histories to analyse similar type of data again with the same procedure and minimal costs. If you look into the history you can see that we still have all the steps present that were needed to come to our final result. Thus, you can convert this history into a workflow by clicking the history **Options** button (1) and choosing the **Extract Workflow** option (2)

The screenshot shows the Galaxy web interface with the 'History' menu open. The 'Extract Workflow' option is highlighted with a red box and the number 2. A red box labeled 1 highlights the 'History' icon in the top right corner of the interface.

We focus on the center pane in the next screenshot. Here, we are able to choose which steps to include/exclude and how to name the newly created workflow. Do not focus on the naming of the individual datasets, we need to edit this afterwards in any case. The importance is that all of the analysis steps are included, we can shuffle them around later.

1. You want to give the workflow a proper name
2. We need to realize that the data upload can unfortunately not be part of the workflow, the workflow can only on datasets already in our history. However, we only need two datasets, so deselect the third.
3. We do not include the filter step as we are really interested in finding all SNPs in **all** promoter regions not only in the first 8,000,000 base pairs.
4. Once this is done we can click **Create Workflow**.

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

Bioinf-course 1 - Intro

Create Workflow

Tool

- Upload File** *This tool cannot be used in workflows*
- UCSC Main** *This tool cannot be used in workflows*
- Unknown** *This tool cannot be used in workflows*
- Cut** *Include "Cut" in workflow*
- Get flanks** *Include "Get flanks" in workflow*
- Filter** *Include "Filter" in workflow*
- Join** *Include "Join" in workflow*
- Build custom track** *Include "Build custom track" in workflow*

History items created

- 10: mm9_chrX_SNP128 Treat as input dataset
- 11: mm9_knownGene_chrX Treat as input dataset
- 13: mm9_ChIP_chr19_control Treat as input dataset
- 14: mm9_knownGene_chrX_short
- 18: mm9_chrX_promoter
- 19: mm9_chrX_promoter_8000000
- 21: SNPs_at_promoter
- 22: Build custom track on data 10, data 21, and data 19

2.8.2 Editing workflows

Now we can see that Galaxy created our workflow. Click on the **Workflow** button in the top pane (1) to get to the workflow overview page.

Workflow "Bioinf-course 1 - Intro" created from current history. You can edit or run the workflow.

Tools

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Convert Formats

History

- search datasets
- Bioinf-course 1
 - 8 shown, 14 deleted
 - 204.5 MB
- 22: Build custom track on data 10, data 21, and data 19

On the workflow overview page click on the workflow and on the **Edit** option (1).

Your workflows

Name	# of Steps
Bioinf-course 1 - Intro	6

Workflow
No workflow

Other
Configure

Create new workflow **Upload or import workflow**

The next window will show you the workflow editor. You will see two areas that are of importance: 1 is the graphical representation of our workflow in form of a flow-diagram, and 2 is the area where we can see/change attributes of individual steps.

Workflow Canvas | Bioinf-course 1 - Intro

1

2

Details

Edit Workflow Attributes

Name: Bioinf-course 1 - Intro

Tags:

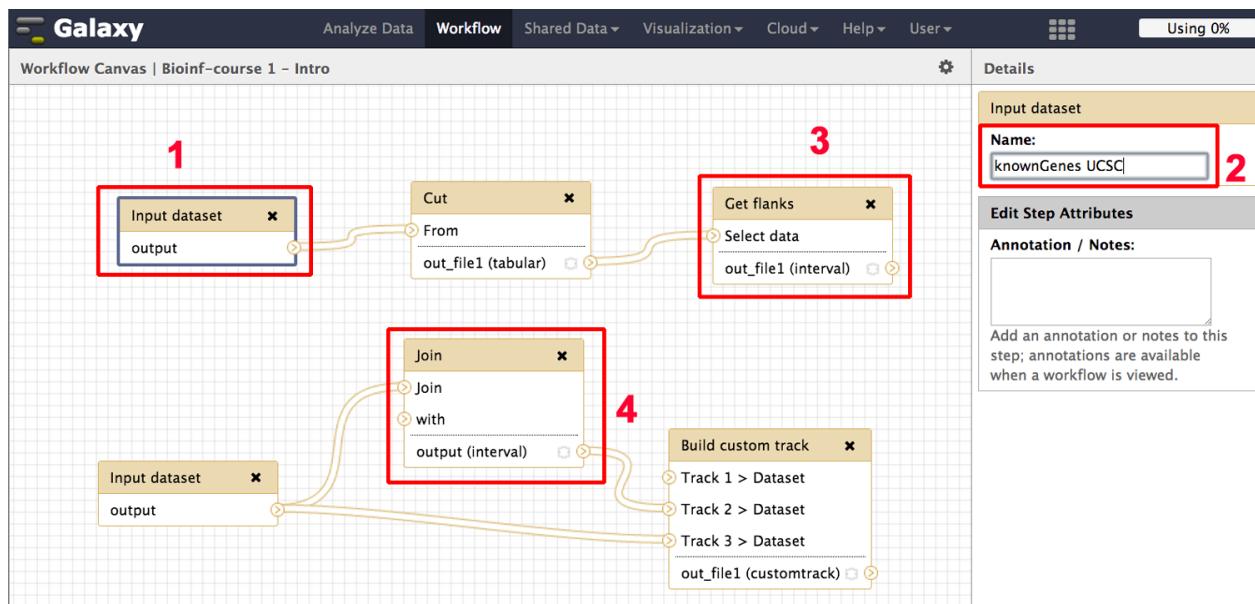
Apply tags to make it easy to search for and find items with the same tag.

Annotation / Notes:

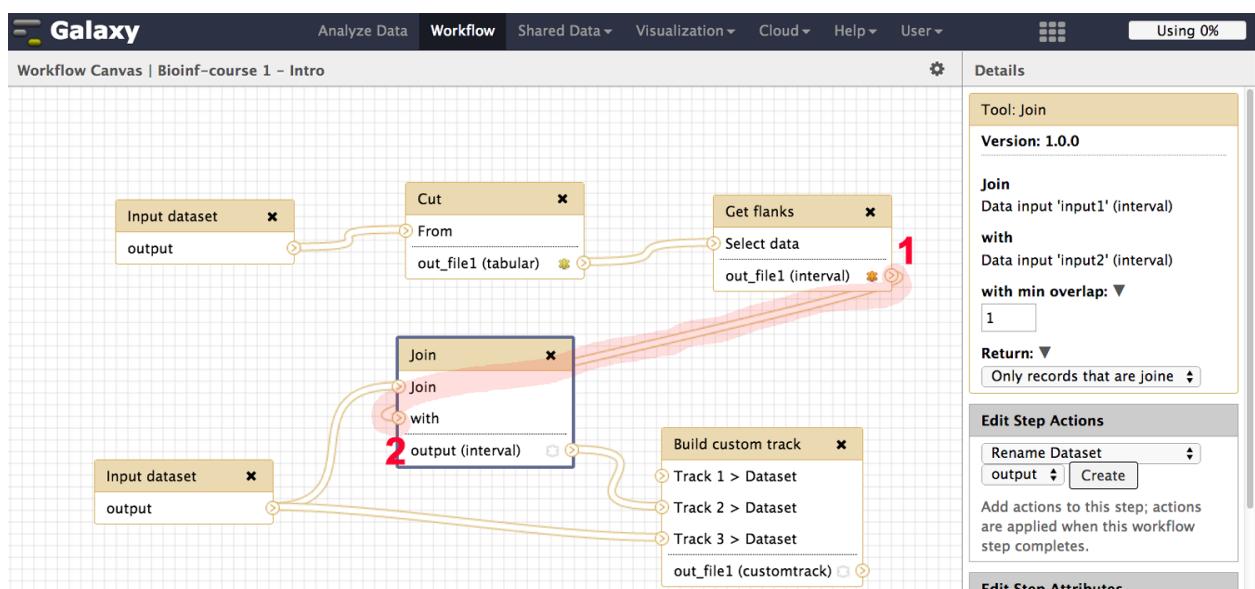
Describe or add notes to workflow
Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.

In the next picture I pulled apart the two input data fields to disentangle the view a bit. We recognise that our workflow is a bit messed up and we need to fix it, e.g. the two input datasets are not connected at the **Join** tool.

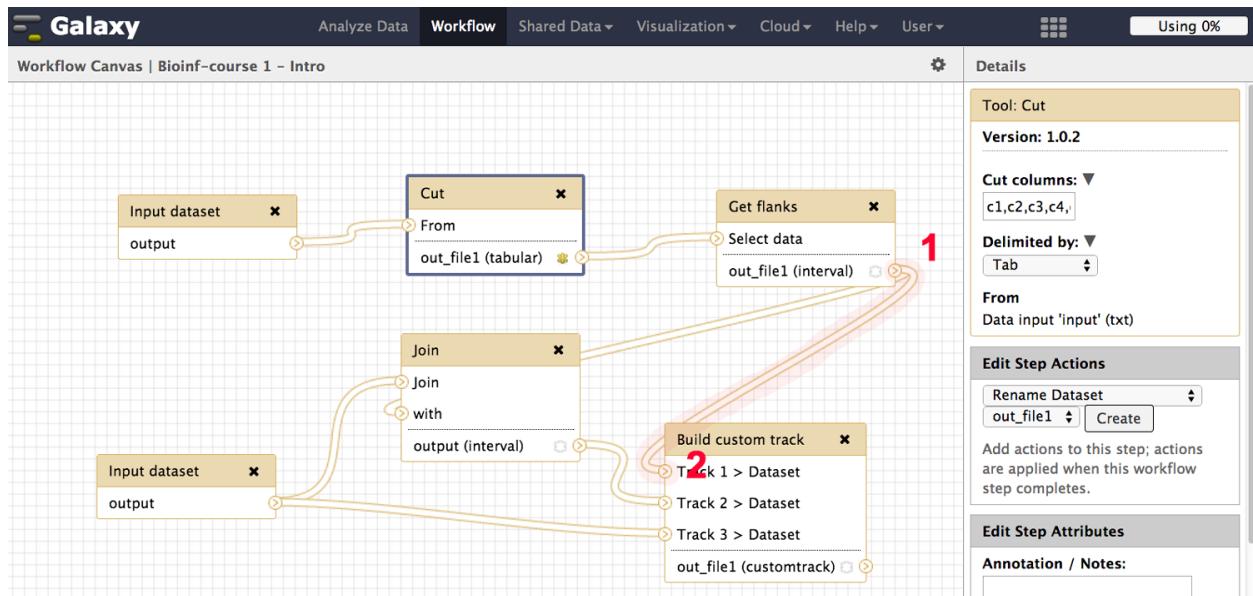
In 1 we find the **knownGenes** input dataset as we remember it needs to be **cut** and we need to extract flanking regions for the genes (**Get flanks**). The first thing to do is to rename this dataset to **knownGenes UCSC** (2), so that we later know what this dataset is. We realise that the results of the flanking regions from 3 (**out_file1 (interval)**) is not joined to the SNP data in 4.



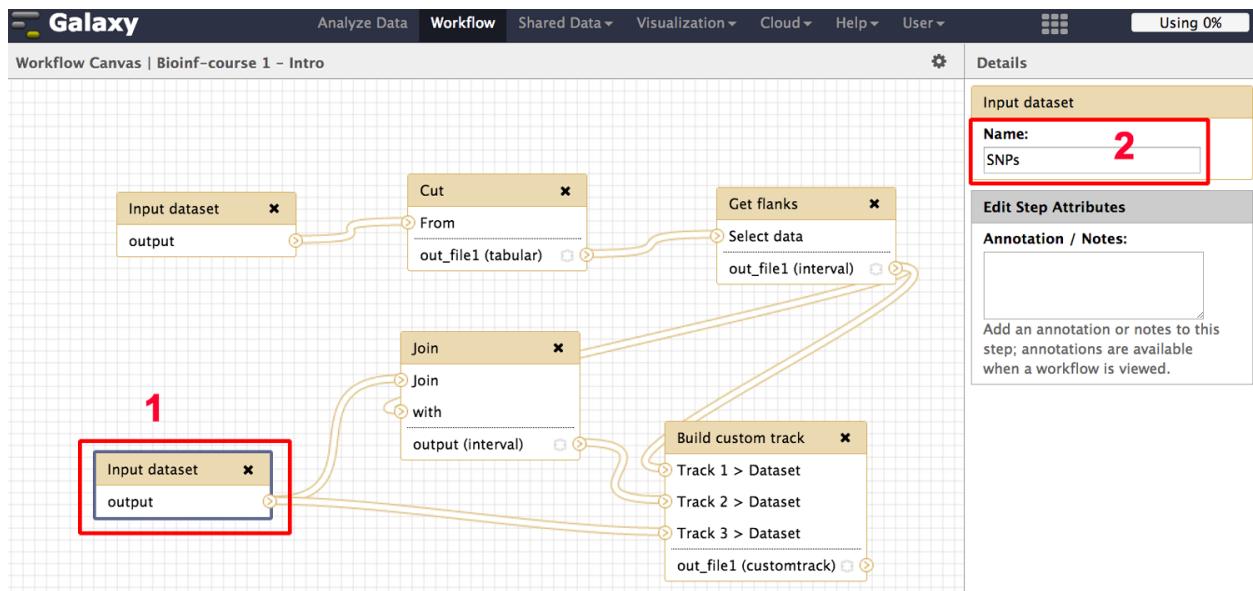
We connect the output of **Get flanks (out_file1 (interval))** (1) to the input of the **Join** tool 2 by dragging a connector from 1 to 2.



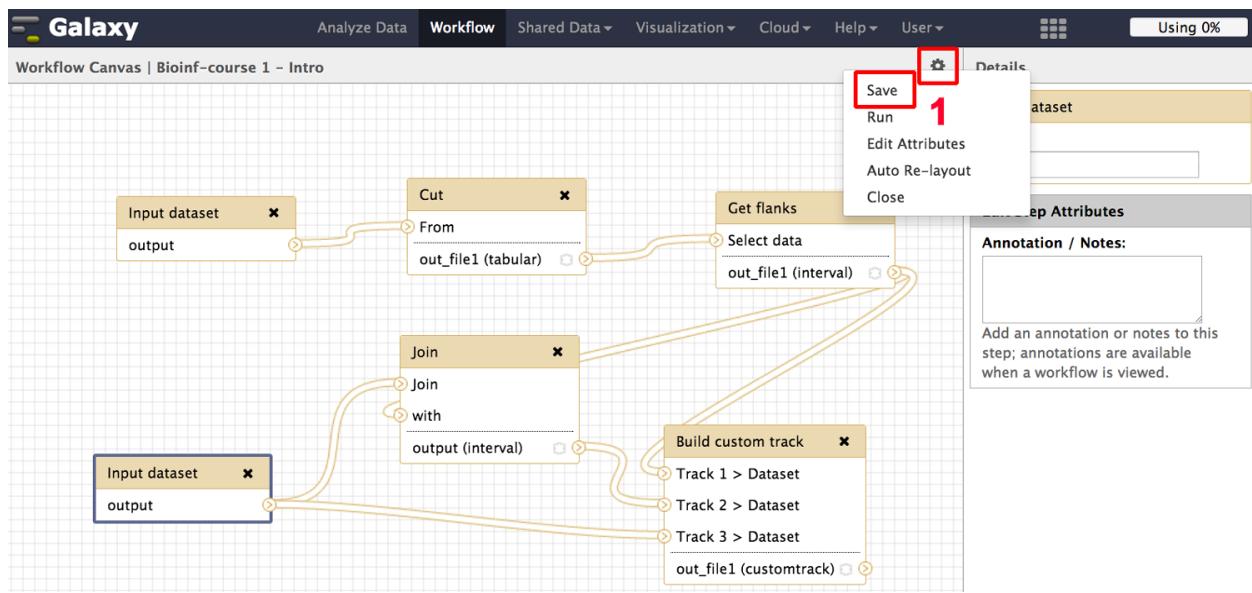
We also want to show our promoters in the output UCSC track that we create as a result, but it is not connected to it either. We fix that by dragging a connector file from the output dataset of the **Get flanks** step 1 to the **Build custom track** input 2.



Next we rename the second input dataset into the workflow in 1 to **SNPs** in the **Details** pane (2).

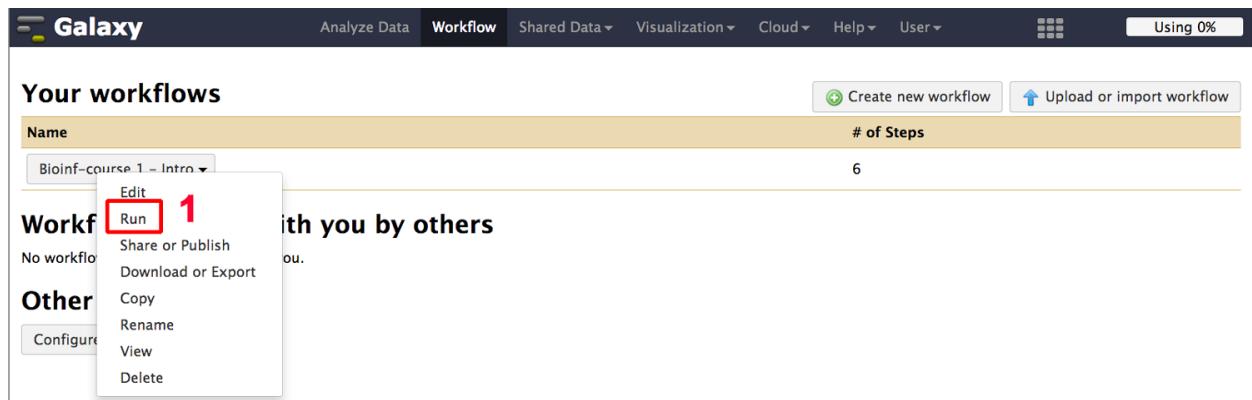


Finally, we save the workflow (1)



2.8.3 Applying workflows to your data

Now that we have the workflow let's run it. First go to the **Workflow** panel and select the workflow and hit **Run** (1).



Now we see the workflow and we can expand each section by clicking on the headers. We choose an appropriate dataset for the **knownGenes UCSC** (1) and the **SNPs** (2). We can see in that the dataset of **Step 1** (knownGenes) is used in **Step 3** and that the output from **Step 3** is used in **Step 4**, exactly what we want (3). We also see that we join the results from **Step 4** with our **SNPs** input dataset from **Step 2** (4). Just specify your geneset and SNPs and click the **Run workflow** button.

Running workflow "Bioinf-course 1 - Intro"

1 Step 1: Input dataset
knownGenes UCSC
18: mm9_chrX_promoter
type to filter

2 Step 2: Input dataset
SNPs
10: mm9_chrX_SNP128
type to filter

3 Step 3: Cut (version 1.0.2)
Cut columns
c1,c2,c3,c4,c5,c6
Delimited by
Tab
From
Output dataset 'output' from step 1

4 Step 4: Get flanks (version 1.0.0)
Select data
Output dataset 'out_file1' from step 3
Region
Around Start
Location of the flanking region/s
Upstream
Offset
0
Length of the flanking region(s)
5000

Step 5: Join (version 1.0.0)
Join
Output dataset 'output' from step 2
with
Output dataset 'out_file1' from step 4
with min overlap
1
Return
Only records that are joined (INNER JOIN)

Step 6: Build custom track (version 1.0.0)

Send results to a new history named: Bioinf-course 2

Run workflow

This concludes the introduction. You can find more advanced bioinformatics tutorials [here](#).

2.9 References

3.0 Web links

This tutorial: <http://sschmeier.github.io/bioinf-workshop/galaxy-intro/>

Galaxy: <http://galaxyproject.org/>

Galaxy Wiki: <http://wiki.galaxyproject.org/>

Galaxy mailing lists: <http://wiki.galaxyproject.org/MailingLists>

Galaxy learning material: <https://wiki.galaxyproject.org/Learn>