

Galaxy Introduction

Sebastian Schmeier

2 May 2015

Contents

Galaxy Introduction	2
1.1 Overview	2
1.1.1 Important links	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	29

Galaxy Introduction



1.1 Overview

In this brief tutorial we will learn how to use the excellent tool [Galaxy](#) 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.



[Galaxy](#) enjoys a large and growing user and developer base, which is evident by its own yearly [conference](#) 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](#). 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](#) have a full fledged installation on our infrastructure ready for scientist to be used).

1.1.1 Important links

- [Wiki](#)
 - [Mailing lists](#)
 - [Other learning material](#)
-

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/>

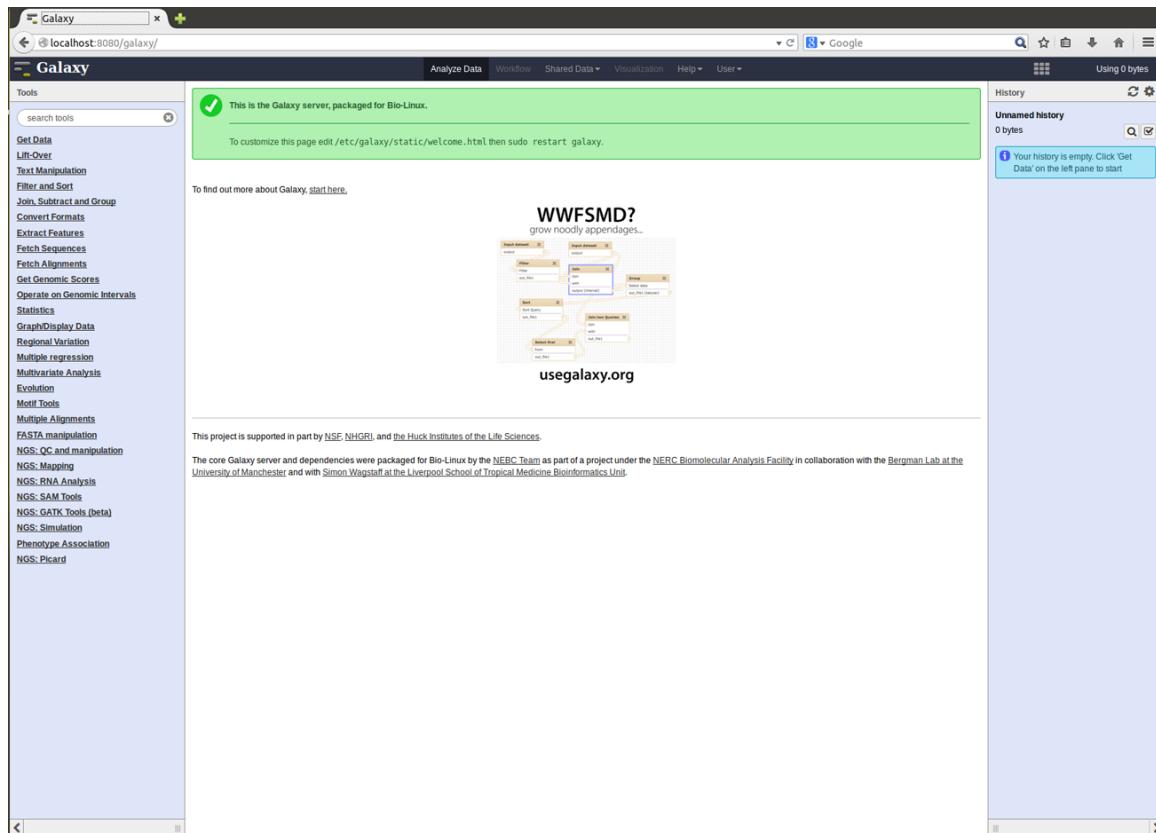
You can start your own [Galaxy](#) instances on [Cloud](#) infrastructure, e.g. [Amazon Cloud Services](#), should you have bigger analysis needs that you want to perform in the cloud.

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

You can install [BioLinux](#) on you 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

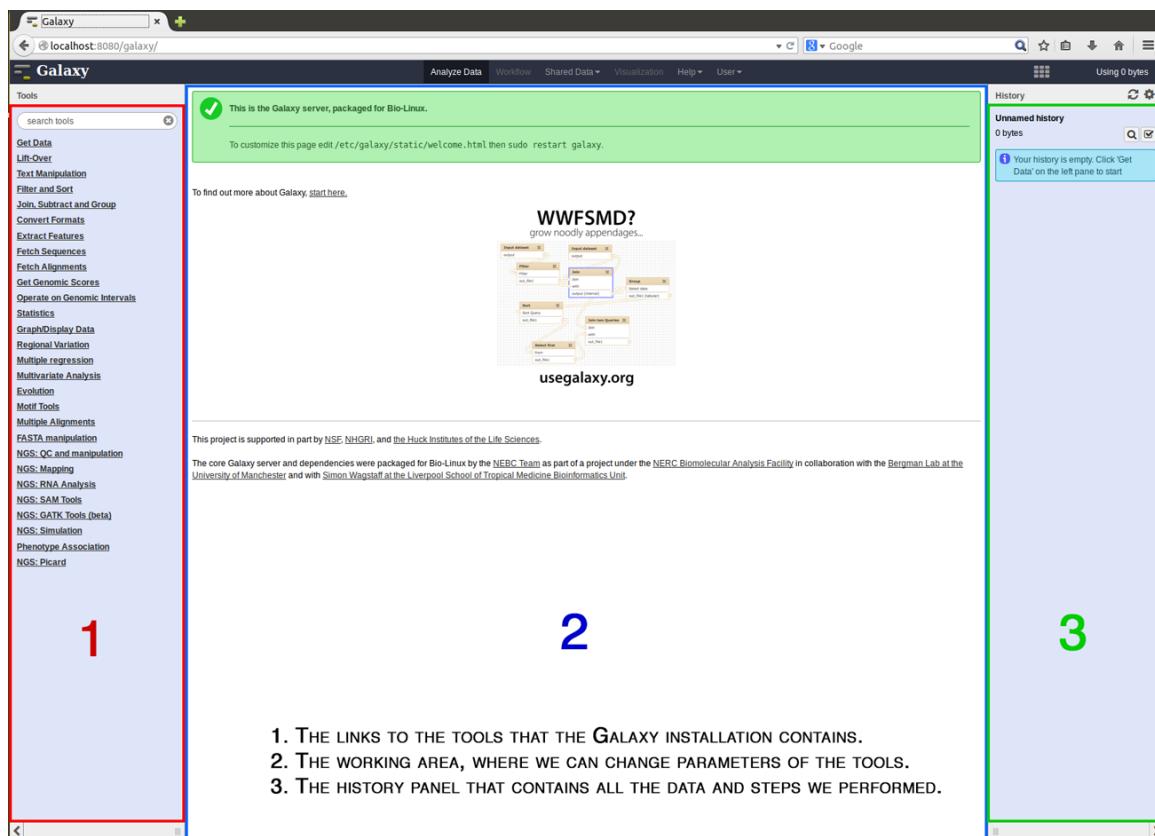
1.3.1 Basics



Hint! Click on the Galaxy screenshots to get a bigger version!

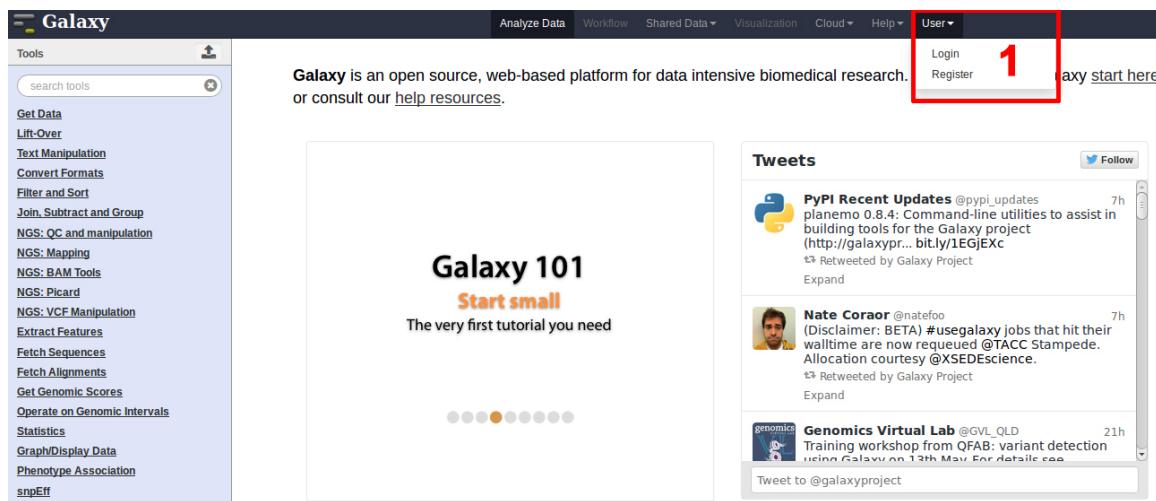
There are 3 areas of interest for now:

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.



1.3.2 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).



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 accepts 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:

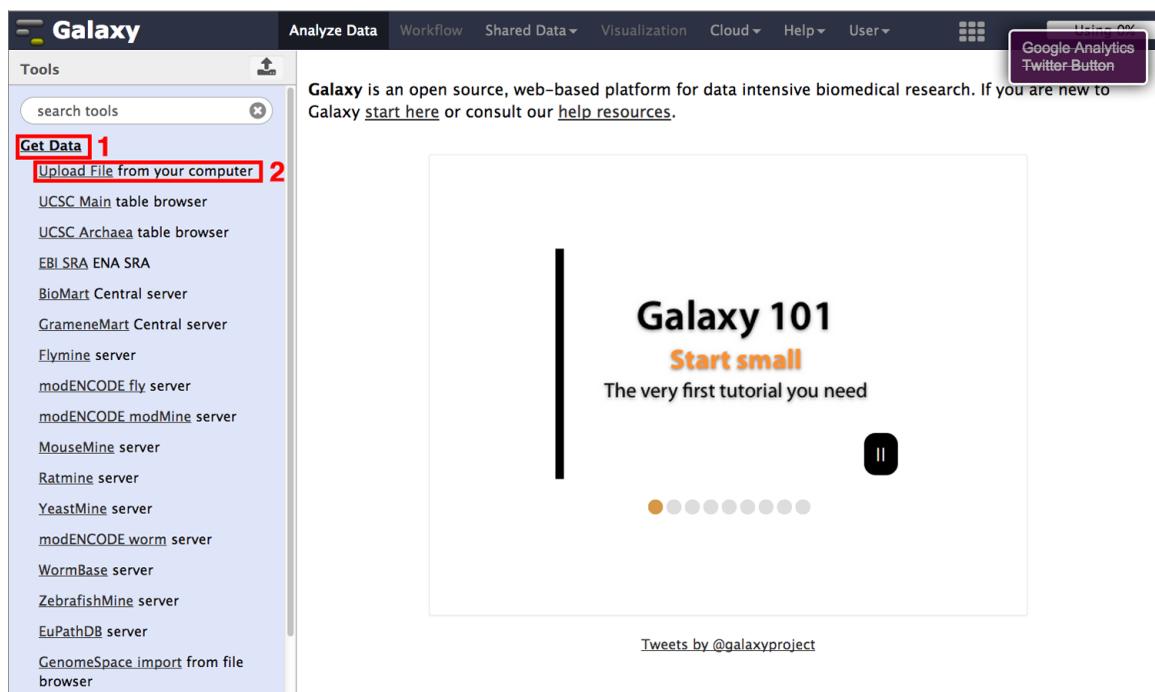
1. Get single nucleotide polymorphism (SNP) data for chromosome X
 2. Get all gene locations on chromosome X
 3. Get upstream regions of the genes
 4. Overlap the SNPs with the genic upstream regions
 5. Visualise results in a genome browser
-

2.2 Loading your own data

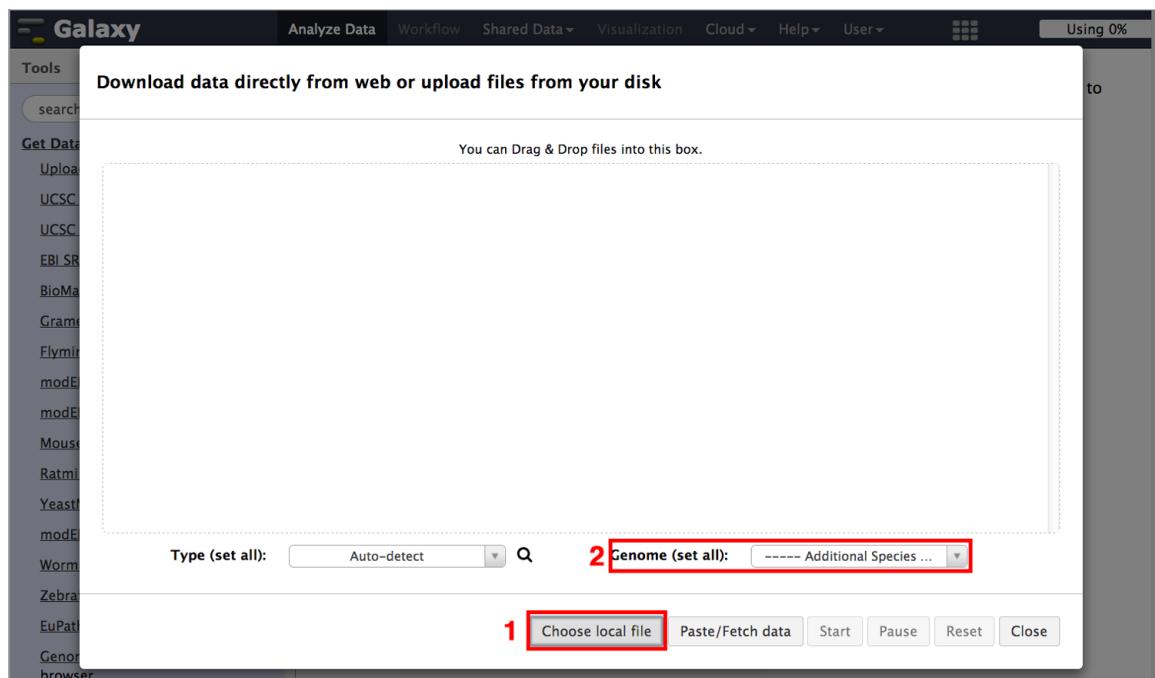
Download the following file to your computer: [mm9_chrX_SNP128_set.bed](#). The file is in [bed-format](#), a simple tab-separated format containing 6 columns: **chromosome, start, stop, name, score, strand**.

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**.



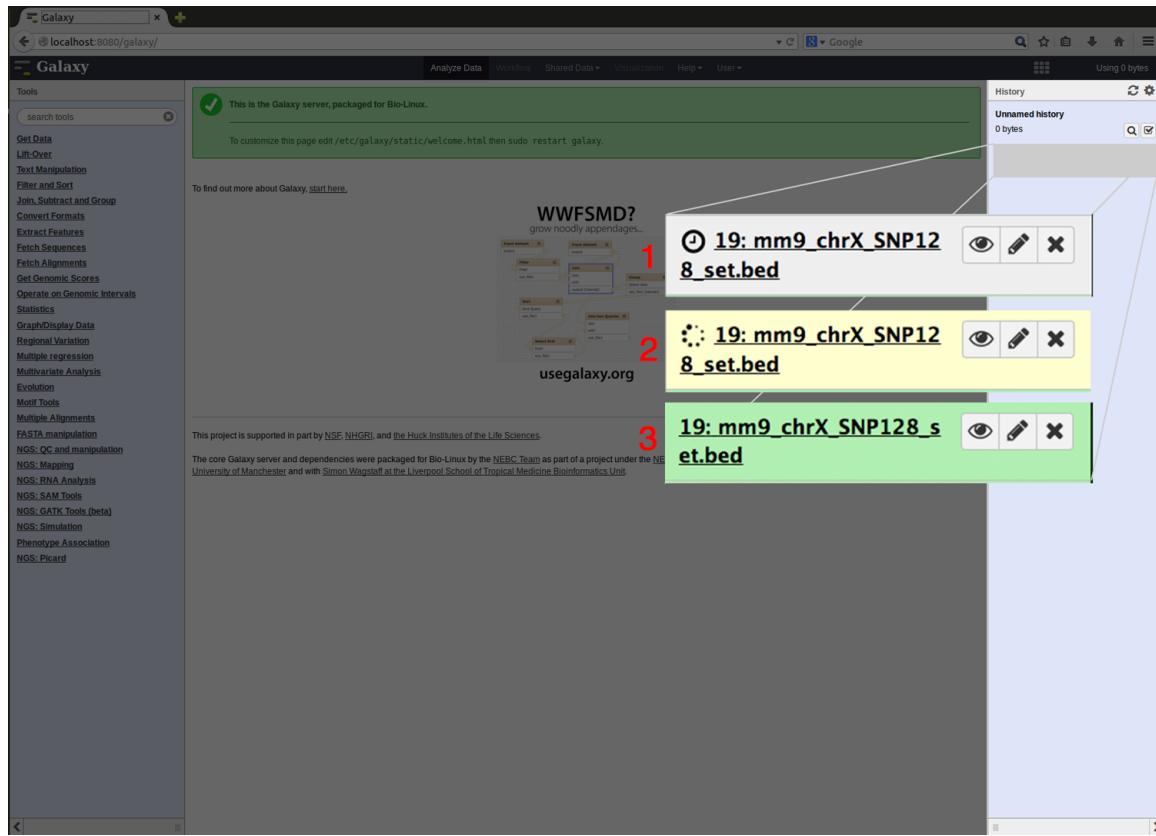
1. An additional window should open that allows you to select the your file.
2. You can specify the species, given that we are looking at mouse data from mm9 set it to the same.



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



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)

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#).

Galaxy 101

Start small

The very first tutorial you need

19: mm9_chrX_SNP128_set.bed
20,000 regions
format: bed, database: mm9

uploaded bed file

display in IGB View
display at Ensembl Current
display at UCSC main

1. Chrom	2. Start	3. End	4. Name	5	6	7
chrX	3242568	3242569	rs51257154	0	-	1
chrX	3242572	3242573	rs49693543	0	-	1
chrX	3242573	3242574	rs45795462	0	-	1
chrX	3749157	3749158	rs45795462	0	+	1
chrX	3749158	3749159	rs49693543	0	+	1
chrX	3749162	3749163	rs51257154	0	+	1
chrX	3907318	3907319	rs48647149	0	+	1
chrX	3907321	3907322	rs48584752	0	+	1
chrX	3907739	3907740	rs45858970	0	+	1
chrX	3907803	3907804	rs48529475	0	+	1
chrX	3907824	3907825	rs46088235	0	+	1

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.

1	2	3	4	5	6	7
chrX	3242568	3242569	rs51257154	0	-	1
chrX	3242572	3242573	rs49693543	0	-	1
chrX	3242573	3242574	rs45795462	0	-	1
chrX	3749157	3749158	rs45795462	0	+	1
chrX	3749158	3749159	rs49693543	0	+	1
chrX	3749162	3749163	rs51257154	0	+	1
chrX	3907318	3907319	rs48647149	0	+	1
chrX	3907321	3907322	rs48584752	0	+	1
chrX	3907739	3907740	rs45858970	0	+	1
chrX	3907803	3907804	rs48529475	0	+	1
chrX	3907824	3907825	rs46088235	0	+	1

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, the 'Tools' sidebar is visible with various options under 'Get Data'. The main area is titled 'Edit Attributes' with tabs for 'Attributes', 'Convert Format', 'Datatype', and 'Permissions'. The 'Attributes' tab is selected. The 'Name' field contains 'mm9_chrX_SNP128_set.bed' (labeled 2). The 'Info' field contains 'uploaded bed file'. Under 'Database/Build', it says 'Mouse July 2007 (NCBI37/mm9) (mm9)'. The 'Number of comment lines:' dropdown is set to 0. The 'Chrom column:' dropdown is set to 1. The 'Start column:' dropdown is set to 2. The 'End column:' dropdown is set to 3. The 'Strand column (click box & select)' dropdown is set to 6. The 'Name/Identifier column (click box & select)' dropdown is set to 4. The 'Score column for visualization:' dropdown has items 1, 2, 3, and 4 checked. A red box highlights the 'Score column for visualization:' dropdown and the 'Save' button (labeled 4). Below the 'Save' button is an 'Auto-detect' button and a note: 'This will inspect the dataset and attempt to correct the above column values if they are not accurate.' On the right, the 'History' panel shows an unnamed history with one dataset: '19: mm9_chrX_SNP128_set.bed' (labeled 1). The dataset is 29.1 MB.

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

This screenshot is similar to the previous one but shows the result of a deletion. The 'History' panel now shows the same dataset with a red box around the delete icon (labeled 1). The 'Save' button in the 'Edit Attributes' dialog is also highlighted with a red box (labeled 4).

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.

The screenshot shows the Galaxy web interface with the 'Table Browser' application selected. The interface is divided into two main sections: a left sidebar with a list of available tools and a right panel for configuring the table browser.

- 1** In the sidebar, 'Get Data' is highlighted.
- 2** 'UCSC Main table browser' is selected from the list.
- 3** In the main panel, 'clade: Mammal', 'genome: Mouse', and 'assembly: July 2007 (NCBI37/mm9)' are set.
- 4** 'group: Genes and Gene Predictions' and 'track: UCSC Genes' are selected.
- 5** 'table: knownGene' is selected, and the 'describe table schema' button is highlighted.
- 6** 'output format: BED – browser extensible data' is selected, and the 'Send output to Galaxy' checkbox is checked.
- 7** The 'get output' button is highlighted.

At the bottom of the panel, there is a note: 'To reset all user cart settings (including custom tracks), [click here](#)'.

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 top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data' (with a dropdown menu), 'Visualization', 'Cloud', 'Help', 'User', and a grid icon. A search bar for datasets is present. On the left, a 'Tools' sidebar lists various bioinformatics tools. The main content area displays a message about Galaxy being an open source tool for biomedical research, followed by a digital clock-style counter showing '060+' and a 'History' panel listing datasets. The 'Shared Data' dropdown is highlighted with a red box, and the 'Data Libraries' option is also highlighted.

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' search results. The search bar contains 'mouse'. The results table has two columns: 'Data library name' and 'Data library description'. The first result is '1000 Genomes' with the description 'Data from the 1000 Genomes Project FTP site'. The second result is 'AC-exome' with the description 'Data for two papers about the Khoisan and other populations.' A link 'Charts Example Data' is visible at the bottom.

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

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' search results for 'mouse'. The search bar contains 'mouse'. The results table has two columns: 'Data library name' and 'Data library description'. The first result is 'ChIP-Seq Mouse Example' with the description 'Data used in examples that demonstrate analysis of ChIP-Seq data'. This row is highlighted with a red box.

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](http://bit.ly/QmD6Nk). These data were generated and analyzed by the labs of [Michael Snyder at Stanford University](http://snyderlab.stanford.edu/) and [Sherman Weissman at Yale University](http://info.med.yale.edu/bcmm/SMW/SMWhome2.html). 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](http://bit.ly/QmD6Nk). These data were generated and analyzed by the labs of [Michael Snyder at Stanford University](http://snyderlab.stanford.edu/) and [Sherman Weissman at Yale University](http://info.med.yale.edu/bcmm/SMW/SMWhome2.html). 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 this 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 easy 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

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***.

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, with the 'Text Manipulation' section highlighted (1). Within this section, the 'Cut columns from a table' tool is selected (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' (3). The 'From' dropdown is set to '24: mm9_knownGene_chrX' (4). The 'Execute' button is highlighted with a red box (5). A warning message below the tool states: 'WARNING: This tool breaks column assignments. To re-establish column assignments run the tools and click on the pencil icon in the latest history item.' Below the tool, there is an example showing the input 'apple,is,good windows,is,bad' and the output 'apple good windows bad'. The 'What it does' section provides a detailed description of the tool's function.

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 -> ***mm9_knownGene_chrX_short***.

The screenshot shows the Galaxy web interface. On the left, a sidebar titled 'Tools' lists various genomic manipulation tools. In the center, a message box indicates that a job has been successfully added to the queue, resulting in the following datasets: '27: Cut on data 24'. Below this, instructions explain how to check the status of queued jobs and view the resulting data by refreshing the History pane. On the right, the 'History' pane displays the dataset '27: Cut on data 24' (highlighted with a red border), along with other datasets: '26: mm9_ChIP_chr1_9_control', '24: mm9_knownGene_chrX_short', and '19: mm9_chrX_SNP1_28'. The total size of the history is listed as 209.8 MB.

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**

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools

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 1

Profile Annotations for a set of genomic intervals

Merge the overlapping intervals of a dataset

Fetch closest non-overlapping feature for every interval

Concatenate two datasets into one dataset

Subtract the intervals of two datasets

Join the intervals of two datasets side-by-side

Intersect the intervals of two datasets

Get flanks returns flanking region/s for every gene 2

Coverage of a set of intervals on second set of intervals

Complement intervals of a dataset

Cluster the intervals of a dataset

Base Coverage of all intervals

Statistics

Get flanks returns flanking region/s for every gene (Galaxy Tool) Version 1.0.0 Options

Select data
27: mm9_knownGene_chrX_short 3

Region
Around Start 4

Location of the flanking region/s
Upstream 5

Offset
0

Use positive values to offset co-ordinates in the direction of transcription and negative values to offset in the opposite direction.

Length of the flanking region(s)
5000 6

Use non-negative value for length

Execute 7

This tool finds the upstream and/or downstream flanking region(s) of all the selected regions in the input file.

Note: Every line should contain at least 3 columns: Chromosome number, Start and Stop co-ordinates. If any of these columns is missing or if start and stop co-ordinates are not numerical, the tool may encounter exceptions and such lines are skipped as invalid. The number of invalid skipped lines is documented in the resulting history item as a "Data issue".

Example 1

For the following dataset:

```
chr22 1000 7000 NM_174568 0 +
```

running get flanks with Region: Around start, Offset: -200, Flank-length: 300 and Location: Upstream will return (Red: Dataset positive strand; Blue: Flanks output):

```
chr22 500 800 NM_174568 0 +
```

Query positive strand >>>>>>>>>>>>>>>>>>>>>

History

search datasets

Bioinf-course1
4 shown, 27 deleted, 1 hidden
210.0 MB

27: mm9_knownGene_c
hrX_short
2,021 regions
format: interval, database: mm9

display at Ensembl Current
display at UCSC main

1. Chrom	2. Start	3. End	4. Name	5
chrX	3241669	3243629	uc009skj.1	0
chrX	3410667	3412627	uc009skk.1	0
chrX	3461360	3463320	uc009skl.1	0
chrX	3546313	3547091	uc012hdv.1	0
chrX	3665477	3667437	uc009skm.1	0
chrX	3748193	3749684	uc009skn.1	0

26: mm9_ChIP_chr19_co
ntrol
24: mm9_knownGene_ch
rx
19: mm9_chrX_SNPs128

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 positionof genes is specified in the second column (**c2**)
 5. **Execute**

1 Filter and Sort

2 Filter data on any column using simple expressions

3 15: mm9_chrX_promoter

4 c2<8000000

5 Execute

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 +

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 were 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 'Build custom track for UCSC genome browser' tool open. The interface is organized into several panels:

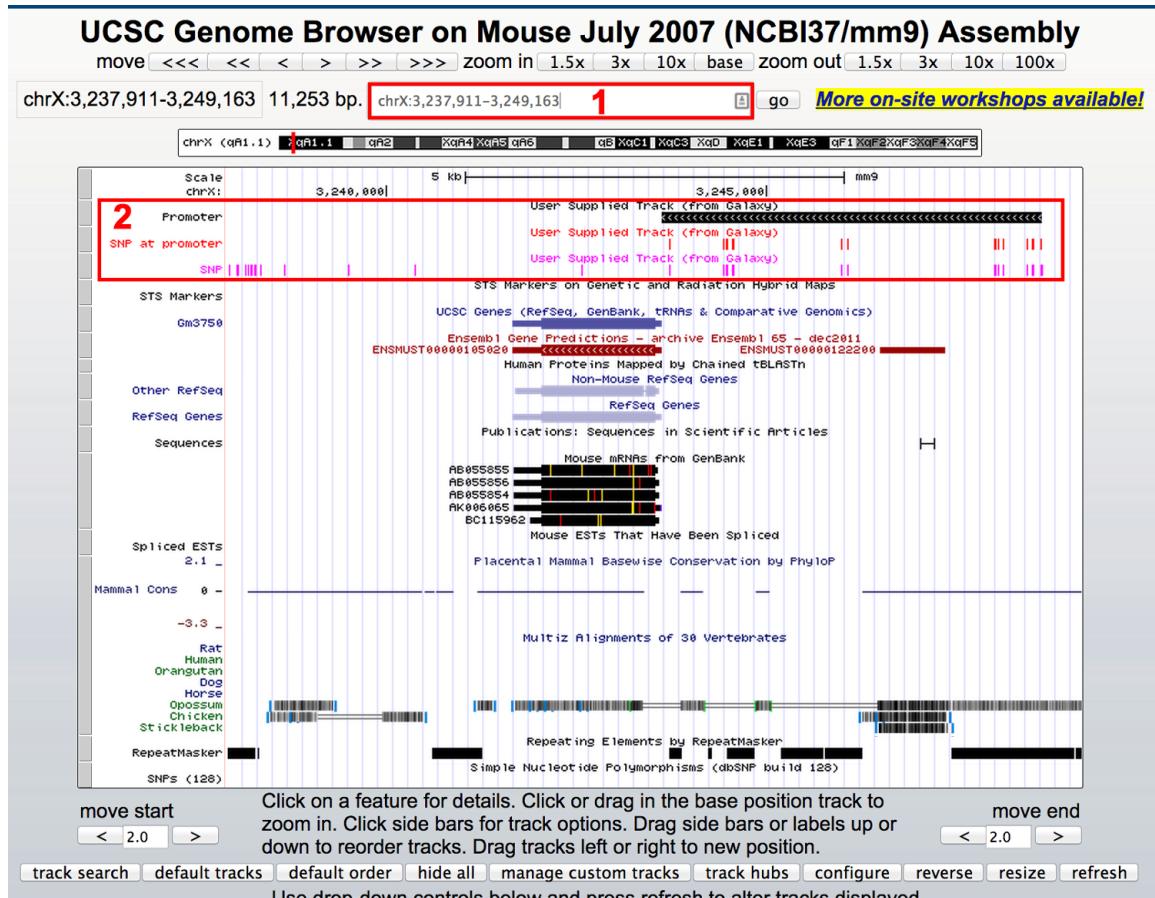
- Left Sidebar (Tools):** Contains a tree view of various genomic analysis tools, with 'Graph/Display Data' (step 1) and 'Build custom track for UCSC genome browser' (step 2) highlighted.
- Top Bar:** Includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User' menus.
- Central Panel (Tool Panel):** Displays the 'Build custom track for UCSC genome browser' tool version 1.0.0. It has sections for 'Track' (containing '1: Track'), 'Dataset' (with a dropdown menu showing '19: mm9_chrX_promtoer_8000000' (step 3)), 'name' (set to 'Promoter' (step 4)), 'description' (set to 'User Supplied Track (from Galaxy)'), 'Color' (set to 'Black'), 'Visibility' (set to 'Dense'), and buttons for '+ Insert Track' (step 5) and 'Execute' (step 6). A tooltip for '+ Insert Track' provides a brief description of the tool's function.
- Right Panel (History):** Shows a list of datasets in the history, including 'Bioinf-course 1' (7 shown, 14 deleted), '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', and '10: mm9_chrX_SNP128'. Each dataset entry includes a preview icon, edit, and delete buttons.

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. On the left, the 'Tools' sidebar is open, displaying various bioinformatics tools. In the center, a table lists genomic data with columns 1, 2, 3, and 4. To the right, the 'History' panel shows a dataset named 'Bioinf-course 1' containing 8 items, with a total size of 204.5 MB. A specific item in the history is highlighted with a red box and labeled '1'. This item is a 'customtrack' named '22: Build custom track on data 10, data 21, and data 19'. It contains 22,360 lines, 4 comments, and is formatted as 'customtrack, database: mm9'. Below the history, a preview of the track is shown in the UCSC Genome Browser, with a red box around the 'Display at UCSC main' link labeled '2'.

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 position 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 at 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 is a sidebar titled "Tools" with various links like "Get Data", "Text Manipulation", and "NGS: QC and manipulation". The main area displays a banner for "Running Your Own Understanding how Galaxy works" with 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 in 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" list. The "Bioinf-course1" history is highlighted with a red box. The list includes other histories such as "26: Mouse ChIP-Seq example Control Data.chr19.mm9", "24: UCSC Main on Mouse knownGene (chrX:1-66650296)", and "19: mm9_chrX_SNP128_subset.bed".

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.

This screenshot shows the context menu for the "Bioinf-course1" history in the "Saved Histories" list. The "Share or Publish" option is highlighted with a red box and a number "1". The menu also includes "Switch", "View", "Copy", "Rename", "Delete", and "Delete Permanently".

The screenshot shows the Galaxy web interface with the title "Share or Publish History 'Bioinf-course 1'". On the left, there's a sidebar titled "Tools" containing a search bar and a list of tools categorized under "Get Data", "Text Manipulation", "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", and "Operate on Genomic Intervals". The main content area has two main sections: "Make History Accessible via Link and Publish It" and "Share History with Individual Users". Under "Make History Accessible via Link and Publish It", there are buttons for "Make History Accessible via Link" and "Make History Accessible and Publish". Under "Share History with Individual Users", there is a button for "Share with a user". At the bottom, there is a link to "Back to Histories List".

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 a history list on the left and a context menu open on the right. The history list contains several entries, each with a timestamp and a list of steps. The context menu is titled "History" and includes options like "HISTORY LISTS", "Saved Histories", "Histories Shared with Me", "CURRENT HISTORY", "Create New", "Copy History", "Copy Datasets", "Share or Publish", "Extract Workflow" (which is highlighted with a red box), "Dataset Security", "Resume Paused Jobs", "Collapse Expanded Datasets", "Unhide Hidden Datasets", "Delete Hidden Datasets", "Purge Deleted Datasets", "Show Structure", "Export Citations", "Export to File", "Delete", and "Delete Permanently". Other actions like "Import from File" are also listed at the bottom.

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 1

Create Workflow **Check all** **Uncheck all**

Tool

- Upload File** 2
This tool cannot be used in workflows
- UCSC Main** 2
This tool cannot be used in workflows
- Unknown** 2
This tool cannot be used in workflows
- Cut** 3
 Include "Cut" in workflow
- Get flanks** 3
 Include "Get flanks" in workflow
- Filter** 3
 Include "Filter" in workflow
- Join** 3
 Include "Join" in workflow
- Build custom track** 3
 Include "Build custom track" in workflow

History

search datasets

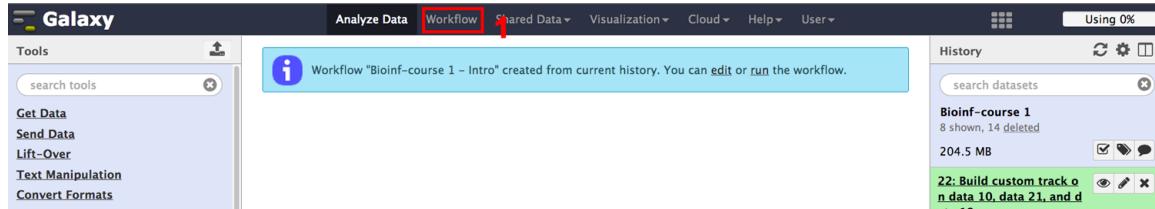
Bioinf-course 1
8 shown, 14 deleted

204.5 MB

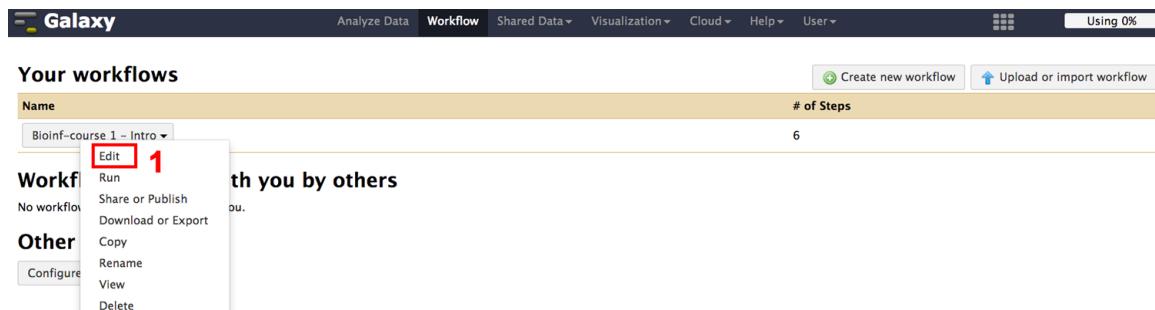
- 22: Build custom track on data 10, data 21, and data 19
- 21: SNPs_at_promoter
- 19: mm9_chrX_promoter_8000000
- 18: mm9_chrX_promoter
- 14: mm9_knownGene_chrX_short
- 13: mm9_ChIP(chr19)_control
- 11: mm9_knownGene_chrX
- 10: mm9_chrX_SNP128

2.8.2 Editing workflows

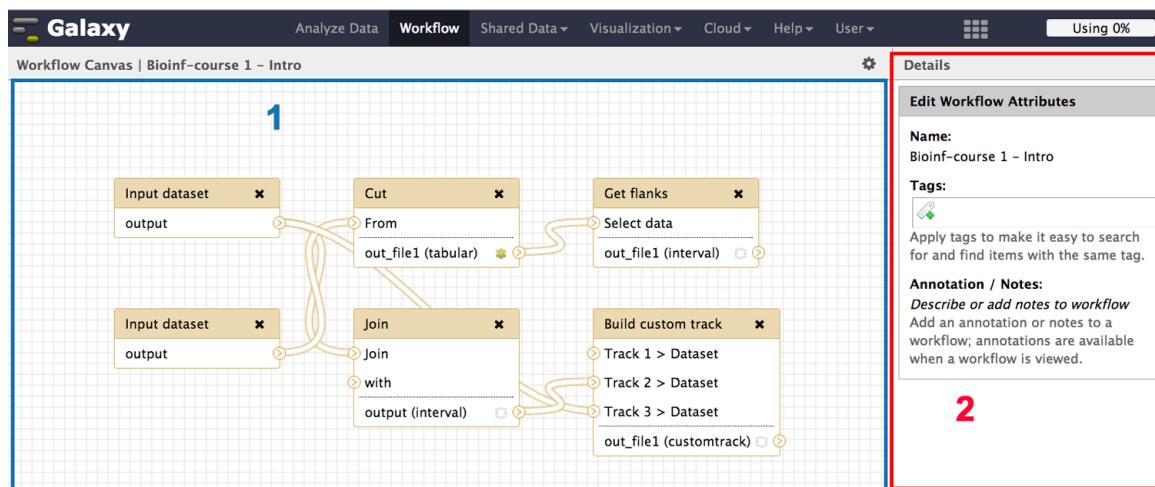
Now we can see that Galaxy created our workflow. Click on the **Workflow** button in the top pane (1) to get to the workflow overview page.



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

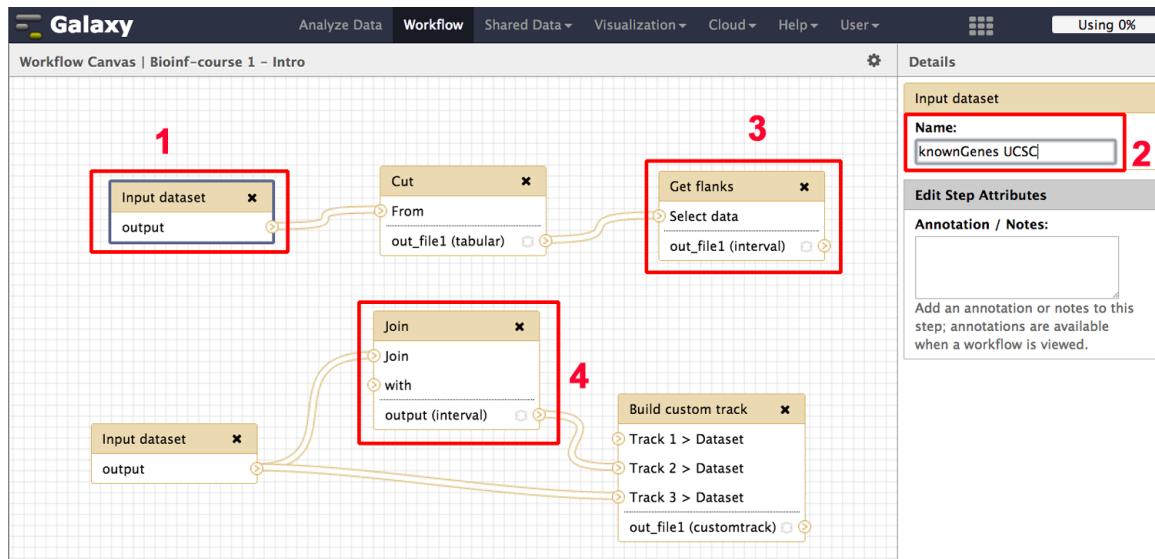


The next window will show you the workflow editor. You will see two areas that are of importance: 1 is the graphical representation of our workflow in form of a flow-diagram, and 2 is the area where we can see/change attributes of individual steps.

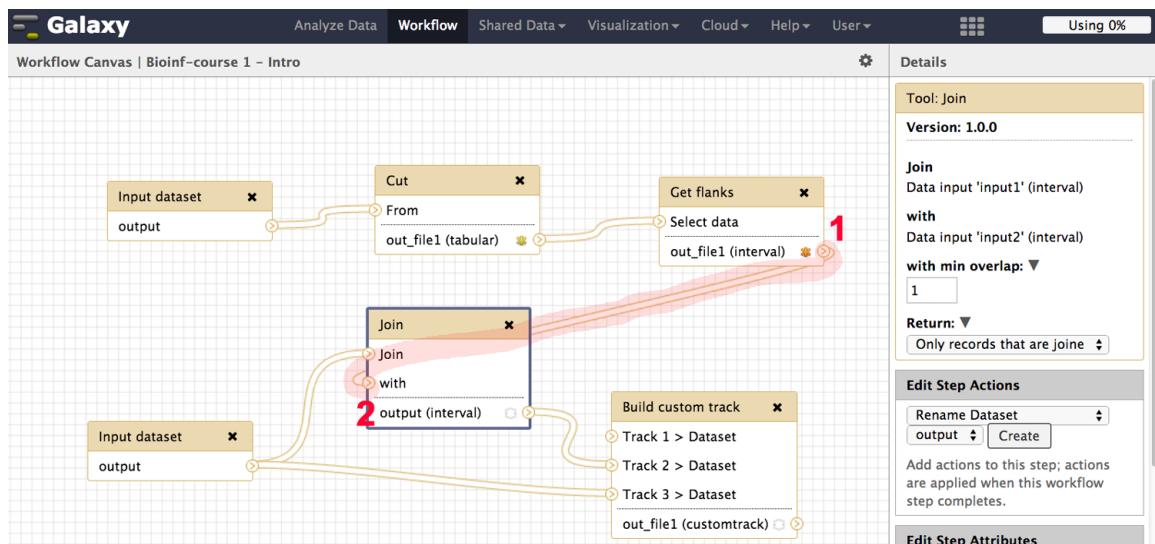


In the next picture I pulled apart the two input data fields to disentangle the view a bit. We recognise that our workflow is a bit messed up and we need to fix it, e.g. the two input datasets are not connected at the **Join** tool.

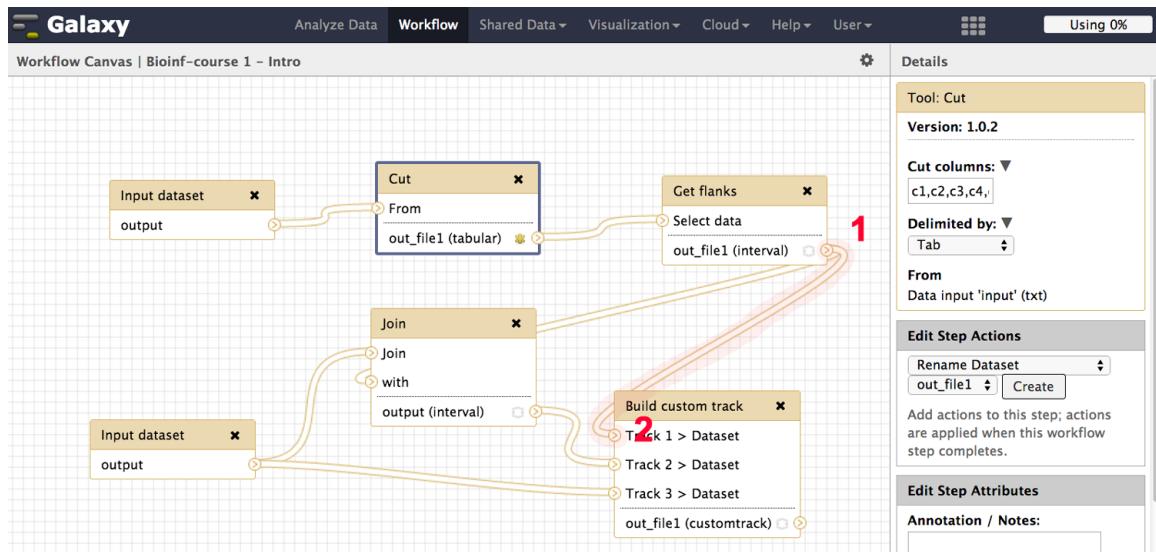
In 1 we find the **knownGenes** input dataset as we remember it needs to be **Cut** and we need to extract flanking regions for the genes (**Get flanks**). The first thing to do is to rename this dataset to **knownGenes UCSC** (2), so that we later know what this dataset is. We realise that the results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



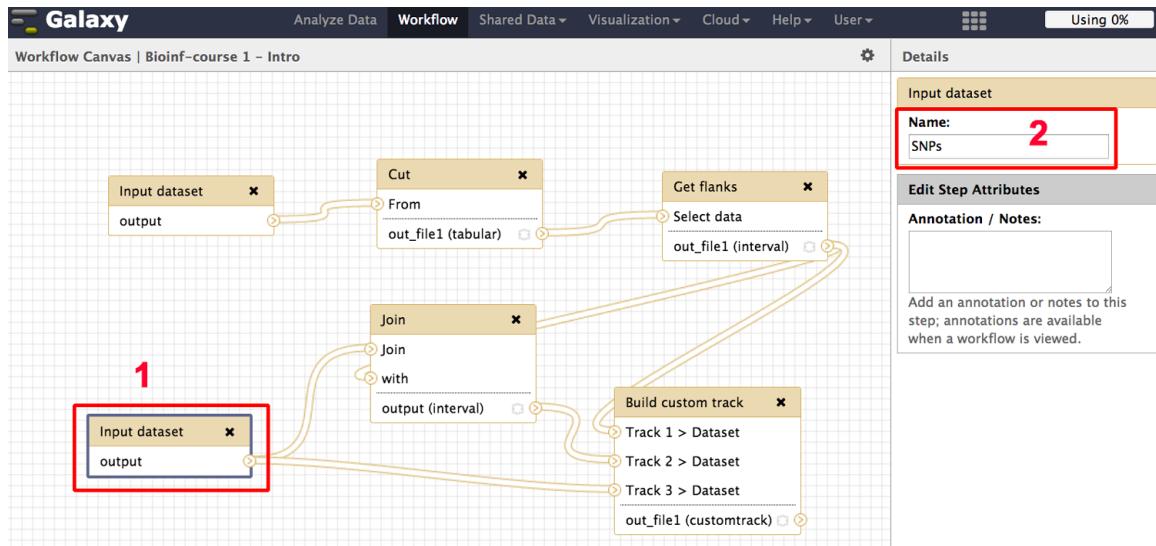
We connect the output of **Get flanks (out_file1 (interval))** (1) to the input of the **Join** tool 2 by dragging a connector from 1 to 2.



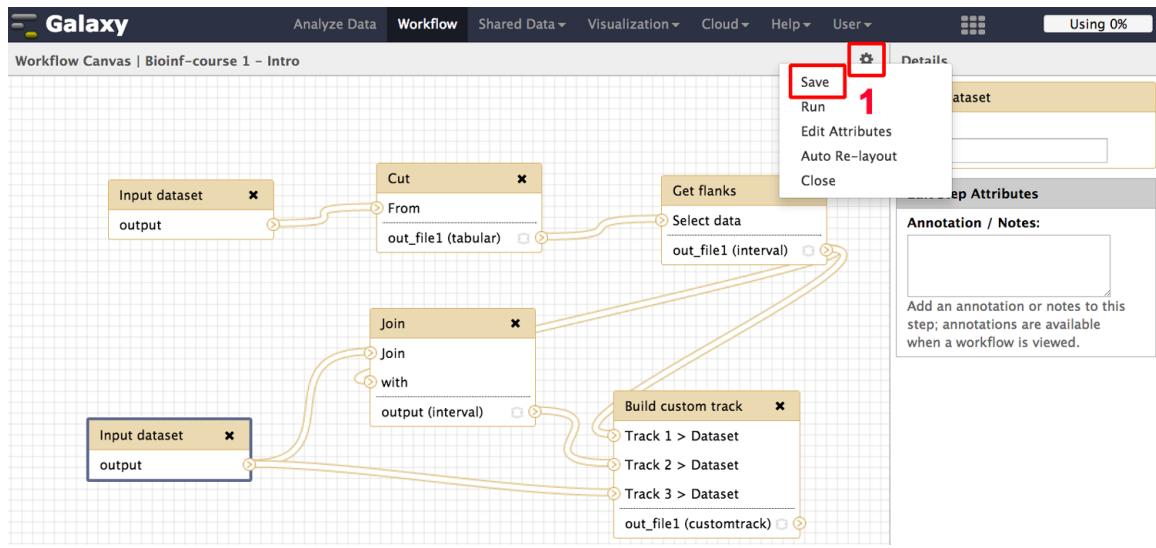
We also want to show our promoters in the output UCSC track that we create as a result, but it is not connected to it either. We fix that by dragging a connector file from the output dataset of the **Get flanks** step 1 to the **Build custom track** input 2.



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

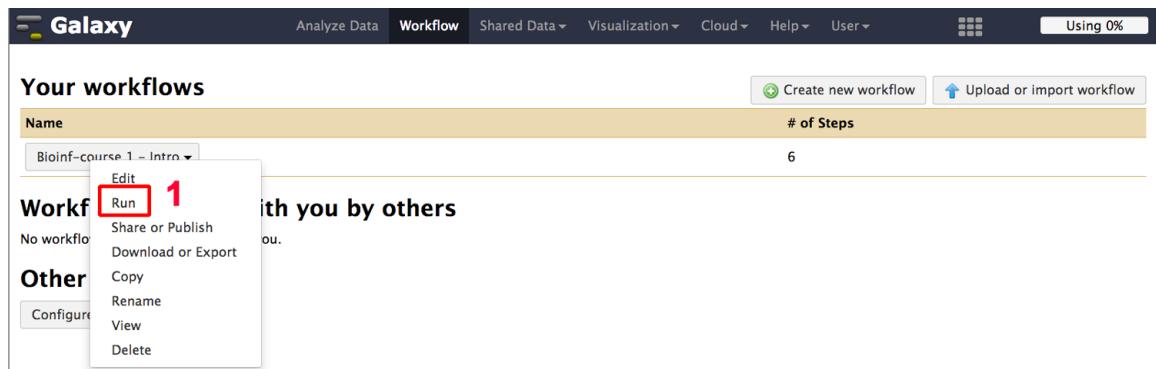


Finally, we save the workflow (1)



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"

Step 1: Input dataset

knownGenes UCSC
18: mm9_chrX_promoter
type to filter

Step 2: Input dataset

SNPs
10: mm9_chrX_SNP128
type to filter

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

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

History

search datasets

Bioinf-course 1
8 shown, 14 deleted
204.5 MB

22: Build custom track o
n data 10, data 21, and d
ata 19

21: SNPs at promoter

19: mm9_chrX_promote
r_8000000

18: mm9_chrX_promoter

14: mm9_knownGene_ch
rX_short

13: mm9_ChIP_chr19_co
ntrol

11: mm9_knownGene_ch
rX

10: mm9_chrX_SNP128

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