Introduction to Galaxy - Practical

Galaxy allows experimental biologists without any programming experience to easily manipulate sequencing data using a point and click interface. For this practical, all the necessary tools and software are pre-installed.

Section 1. The Galaxy interface

Go to http://184.73.191.220/ [*group 1*] or http://54.225.103.130/ [*group 2*] or http://50.17.254.21/ [*group 3*].

The left-hand panel is the *tool panel*, which contains our tools. The *main panel* is the interface between tools and data and is located in the middle. The Menu appears across the top of the main panel. The *history panel* is on the right.

The tool panel

The tool panel contains all the tools available to Galaxy. The tools are categorised into groups of similar function (e.g. Text Manipulation). Clicking on a group reveals an expanded list of tools available under that category.

The main panel

The main panel is where you set the parameters for the current tool in use. Clicking on a tool will load it into the main panel, where you will see a range of parameters (most with default settings) applicable to that tool. This is also where you select which datasets to load into the tool.

The history panel

The history panel contains details of all the datasets that we have imported or created, plus the order in which they were created. All imported data will appear as a history item in the panel and the output from each tool will appear as at least one history item (depending on the number of output files the tool creates).

A history item will usually be one of 4 colours:

- **Green** - the task has completed successfully
- **Yellow** - the task is currently running
- **Grey** - the task is queued to run
- **Red** - the task failed

History items

Each consecutively numbered history item has a number of parts to it. In the top right corner of each history item are three icons - an eye, a pen and a cross.

- **Eye** - clicking on the eye icon will show the data for that item in the main panel.
- **Pen** - this will show all the attributes for this history item (history item name, datatype, number of comment lines, etc.). It can be used to rename the history item or change information about it (e.g. to specify that a fastq file is actually in fastq sanger format).

- **Cross** - clicking on this will delete the history item.

By clicking on the history item name the item will expand to reveal a number of extra items. It will also show the first few lines of the dataset.

The Menu

The Menu appears across the top of the main panel. It contains a number of menu items:

- **Analyze Data** - this brings you back to the Home Screen.
- **Workflow** - this allows you to use workflows - pre-written pipelines of analysis.
- **Shared Data** - this is where you will obtain data shared with you on this Galaxy instance.
- **Visualization** - this is where you will visualise dataset or retrieve stored ones.
- **Help** - this is where you'll find the Galaxy wiki, tutorials and help from the Galaxy community.
- **User** - Saved Datasets: this is where you'll be able to see all your datasets from each of your histories. If you want to move data between histories, this is where you can do it. You can also see all your current histories and any Pages (notes and instructions that you can write and share with others).

Setting up Galaxy account

Go to http://184.73.191.220/ [*group 1*] or http://54.225.103.130/ [*group 2*] or http://50.17.254.21/ [*group 3*].

Go to the **User** link at the top of Galaxy interface and choose **Register** (unless of course you already have an account). Then enter your information and you're in!

This will allow all the work that you create to be saved between sessions and allow you to name, save, share, and publish Galaxy histories, workflows, datasets and pages in a limited space set by your administrator of usually of 250GB.

The little indication bar at the top right corner of your Galaxy page **'Using ...MB'** will show your current usage. If you exceed your quota, no further jobs will be run until you have (permanently) deleted some of your datasets.

Section 2. Getting started

Getting data

We shall start by obtaining some data, inspecting it and running a simple job on it.

1. Click on **'Shared Data'** from the menu and select **'Data Libraries'**.
2. In the list of Data Libraries you will see one called **'GalaxyCam Training'**. Click on this and you will see a couple of sub folders.
3. Next to the folder **'getting_started'** click on the blue arrowhead and check the box next to **'Test.fasta'**. Make sure **'Import to current history'** is selected in the box below and click on **'Go'**.
4. A green bar signalling that your task has been performed successfully will appear at the top

of your screen.
5. Click on **'Analyze Data'** in the menu to return to the main interface. The Test.fasta sequence should now be in your history.

Renaming the History

Instead of working with an unnamed history, we shall call our history something meaningful.

1. At the top of the History column click on the Options gear and select **'Saved Histories'**. Any histories that are currently available will be shown in the central panel. You should see your **'Unnamed history'** there.
2. Click on your 'Unnamed history' from the right hand side panel, to rename it. A box will appear for you to create a new name for your history. Rename it by calling it something meaningful. In this case we'll call it **'Getting Started'**. Hit enter to save it.

Examining the data

We'll now have a good look at our data and all of its attributes.

1. Click on the dataset name **'Test.fasta'**.
2. The dataset will expand and we will see lots of information about the dataset, including a quick peek at the data itself and what format Galaxy thinks the data is (fasta).
3. Click on the **eye icon**. The data connected with this dataset will appear in the Main Panel. We can see that the data is in fasta format. If the dataset was very big, then only the first few hundred lines would be displayed.
4. Click on the **pen sign** (attributes). You can now see all the attributes associated with this dataset divided into sections. The name, annotation, data type, etc. are all editable.
5. We shall **rename the dataset**. In the 'Name' field, remove 'Test.fasta' and call it something else, e.g. **'Galaxy_sequence.fasta'**. Click on **'Save'** at the bottom of the current section. You should see that your history item now has a new name.

Using our data in a tool

FASTA manipulation

Calculate the length of the sequence

1. Click on the **'FASTA manipulation'** tools
2. Select **'Compute sequence length'**. This tool counts the length of each fasta sequence in the file.
3. From the dropdown menu select the FASTA sequence file you wish to get its length, here should be 'Galaxy_sequence.fasta'. Leave '0' to keep the whole thing. Click on **'Execute'.**
4. Click on the dataset newly created in your history **'Getting Started'** to see the result. The length of the sequence is 444 bases.

Change the width of the output FASTA

1. Select **'FASTA Width'** from under **'FASTA manipulation'** tools. This tool reformats a FASTA file, changing the width of the nucleotide lines. Outputting a single line (with width = 0) can be useful for scripting (with grep, awk, and perl). Every odd line is a sequence identifier, and every even line is a nucleotides line.
2. From the dropdown menu select the FASTA sequence file you wish to change the width of,

this should be 'Galaxy_sequence.fasta'. Click on **'Execute'.**

Text manipulation

Galaxy contains a number of text manipulation tools to help us work with files. For example, if we had lots of files all with a single fasta file in each, we could create a single multi-fasta file from them. We'll do that now.

1. Click on **'Shared Data'** and go back to the **'GalaxyCam Training'** data library.
2. Under **'getting_started'** click on the '**Pinfestans**' folder. You will see three files, each one a P. infestans effector gene. Select all three files (you can select all the files for a folder by clicking in the box next to the folder name) and import them into your current history.
3. Click on **'Analyze Data'** in the menu to return to the main interface. The three sequences should now appear in your history as separate items.
4. We will now use one of Galaxy's tools to create a multi-fasta file. Select the tool **'Text Manipulation'** > **'Concatenate datasets'**
5. This will load the tool interface into the main panel. We have 3 P. infestans fasta sequences in our history, select from the drop down list **'PinfestansAVH9.fasta'** for Concatenate Dataset , then use the **'Add new dataset'** button twice to add the other two fasta files. Select as Dataset 1; **'PinfestansAVRblb1.fasta'** and as Dataset 2 **'PinfestansAVRblb2.fasta'**. Hit **'Execute'**.
6. Wait for the tool to finish and then click on the eye symbol of the new history item. You should have all 3 sequences in just one file.

Filtering data

The Galaxy way to do things is to put different tools together in pipelines to accomplish more complex tasks. This allows us to have a few generic tools that can be used in various analysis steps, rather than having a lot of tools that only do a specific task.

Our last task in this part of the tutorial will be to take two small datasets and use the information in both of them to filter out data we're not interested in and drill down to data that we are interested in.

1. Go to the top of the History panel, click on Options and select 'Create New' to create a new history. Name the history **'NBLRR Filter'**.
2. Go to **'Shared Data'** > **'GalaxyCam Training'** > **'getting_started'** > **'NBARC'**, and import the two datasets **'SequenceInfo.tabular'** and **'NB_domains.tabular'**.
3. **EXERCISE 2.1. Select only the NB-ARC sequences that are longer than 1300 nts and have a GC content over 0.5.** Take a look through the tools in Galaxy and try and figure out how to combine the information in both files and then filter for the information that you're interested in.

Hint: You'll need to look at tools in the 'Join, Subtract and Group' section and also in the 'Filter and sort' section.

Section 3. Loading data

Create a new history

1. At the top of the History column click on the Options gear and select **'Create new'**.
2. A new **'Unnamed history'** is created. Rename it to be **'Loading Data'**. Hit enter to save it.

Upload from your Computer Illumina TruSeq chr22 exome targeted regions.

1. On the left Tools panel, click on **'Get Data'** > **'Upload File'** from your computer.
2. In the middle panel, select File Format **'bed'**, click on **'Choose file'** and select from your desktop the file **'truseq_exome_targeted_regions.hg19.chr22.bed'** set Genome: **'Human Feb. 2009 (GRCh37/hg19) (hg19)**' and click **'Execute'**.
3. On the right panel, click on the file name to see the file attributes.
4. Click on the **eye icon** to view the contents of the file in the middle panel.
5. Click on the **pen icon** to edit the attributes of the file like its name. Rename the file to **'illumina_truseq_exome_targeted_regions_hg19_chr22'** and click on **'Save'**.
6. Click on **'display at UCSC main'** to view the target regions as User Track.

Download the genomic intervals of the TAF1 binding sites.

1. Go to **'Get Data'** > **'Upload File'**
2. Copy/Paste in URL/Text: '**http://galaxyproject.org/CPMB/TAF1_ChIP.txt**', Select Genome: **'Human Feb. 2009 (GRCh37/hg19) (hg19)**' and click **'Execute'**.
3. In the History, click on the pencil to edit attributes; change the name of the file to **'TAF1_ChIP'**; and click on the **'Save'** button.

Retrieve all coding exons on human chromosome 22 from UCSC.

1. Go to **'Get Data'** > **'UCSC Main'**
2. Select
    a. genome: 'Human';
    b. assembly: 'Feb. 2009 (GRCh37/hg19)';
    c. group: 'Genes and Gene Predictions';
    d. track: **'RefSeq Genes'**;
    e. table: 'refGene';
    f. region: select **'position'**; enter **'chr22'**
    g. output format: 'BED - browser extensible data';
    h. tick Send output to 'Galaxy'.
    i. Make sure that your settings are exactly the same (in particular, region should be set to **'position' 'chr22'**, output format should be set to **'BED - browser extensible data'**, and **'Galaxy'** should be checked by Send output to option). Click **'get output'**.
3. On next page make sure under Create one BED record per is set to **'Coding Exons'** and click **'Send Query to Galaxy'**.
4. On the right hand side panel, in the history, once the newly created file "UCSC Main on Human knownGene (genome)" turns green, click the pencil **'Edit Attribute'** to change the file name to be '**UCSC_Human_refGene_chr22_Exons**', and click **'Save'**.

Examine the output bed file, it should contain 8,379 regions.

Retrieve Ensembl genes 75 on human chromosome 22 from BioMart

EXERCISE 3.1. Retrieve all the genes (associated gene names, start and end positions) on human chromosome 22 from Ensembl Genes 75 from BioMart using database query mode.

Hint: You'll need to look at tools in the 'Get Data' section. You should get a result file containing 4,460 genes.

Upload large files (>2GB) using FTP

To upload large files into Galaxy, you could use FTP instead of the browser. Files need to be transferred to the Galaxy FTP address and then uploaded into your history. Here is a clear step by step tutorial for this function:

http://wiki.bits.vib.be/index.php/Galaxy_beginner's_tutorial#Upload_big_files_.28.3E_2GB.29_using_FTP

Download data from shared data library

Alternatively, you could import into your history all the data files from a shared data library prepared for this part of the practical.

1. Create a new empty history called **'Loading Data from library'.**
2. Click on the menu **'Shared Data'** from the top panel, then **'Data Libraries'.**
3. Click on **'GalaxyCam Training'** from the list of shared data libraries.
4. Check the box next to **'loading_data'**. Make sure **'Import to current history'** is selected in the box below and click on **'Go'**.
5. A green bar signalling that your task has been performed successfully will appear at the top of your screen.
6. Click on **'Analyze Data'** in the menu to return to the main interface. Four files should now be be in your history.
    a. 4: UCSC_Human_refGene_chr22_exons
    b. 3: illumina_truseq_exome_targeted_regions_hg19_chr22
    c. 2: TAF1_ChIP
    d. 1: BioMart_Homo_sapiens_genes_GRCh37p13_chr22

Don't delete these files! They will be used in the next section.

Section 4. Interval Operations

Basic operations

Create a new history

1. At the top of the History column click on the Options gear and select **'Create new'**.
2. A new **'Unnamed history'** is created. Rename it to be **'Interval Operations'**. Hit enter to save it.

Obtain the UCSC repeat regions on human chromosome 22 from UCSC.

1. Go to **'Get Data'** > **'UCSC Main'**
2. Select
    a. genome: 'Human';
    b. assembly: 'Feb. 2009 (GRCh37/hg19)';
    c. group: **'Repeats'**;
    d. track: 'RepeatMasker'. RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences.;
    e. table: 'rmsk';
    f. region: select **'position'**; enter **'chr22'**
    g. output format: 'BED - browser extensible data';
    h. tick Send output to 'Galaxy'.
    i. Make sure that your settings are exactly the same (in particular, region should be set to **'position' 'chr22'**, output format should be set to **'BED - browser extensible data'**, and **'Galaxy'** should be checked by Send output to option). Click **'get output'**.
5. On next page make sure under Create one BED record per is set to **'Whole Gene'** and click **'Send Query to Galaxy'**.
6. On the right hand side panel, in the history, once the newly created file "UCSC Main on Human knownGene (genome)" turns green, click the pencil **'Edit Attribute'** to change the file name to be '**UCSC_Human_rmsk_chr22**', and click **'Save'**.

Load all coding exons on human chromosome 22 from shared library.

1. Click on the menu **'Shared Data'** from the top panel, then **'Data Libraries'.**
2. Click on **'GalaxyCam Training'** and expend **'loading_data'**.
3. Check the box next to '**UCSC_Human_refGene_chr22_Exons'**. Make sure **'Import to current history'** is selected in the box below and click on **'Go'**.

Practice basic interval operations.

1. Click on **'Operate on Genomic Intervals'** > **'Intersect'**.
2. Select Return: 'Overlapping intervals' of '**UCSC_Human_refGene_chr22_Exons'** that intersect: '**UCSC_Human_rmsk_chr22**', for at least: '1' (bp). This operation will return the exons which overlap with repeats regions. Click **'Execute'**.
3. Rename the returned file as: **'UCSC_Human_chr22_Intersect_on_Exons_Repeats'**, and click **'Save'**. You should see 266 regions.
4. Load **'UCSC_Human_chr22_Intersect_on_Exons_Repeats'** into UCSC browser, and turn on RefGenes track under Genes and Gene Predictions and RepeatMasker track under Repeats.
5. Zoom into individual intersect regions, e.g.: chr22:22,293,855-22,294,135 and

chr22:20,918,588-20,921,366, and compare the relationships between the intersect region as User Track with the exons and repeats regions.

6. Repeat step 2 with alternative return option Overlapping pieces of Intervals, what changed in the intersection regions this time?
7. Click on **'Operate on Genomic Intervals'** > **'Coverage'**.
8. Select What portion of **'UCSC_Human_chr22_Intersect_on_Exons_Repeats'**, is covered by '**UCSC_Human_rmsk_chr22**'. This returns the counts and percentage of bases for the intervals in the first dataset which are also covered by the intervals in the second dataset. Click **'Execute'**.
9. Rename the result data set to **'UCSC_Human_chr22_Coverage_Intersect_by_rmsk'**, and click **'Save'**.

Retrieve shared data.

The history of this section can be found in **'Shared Data'** > **'Published Histories'**. It is called 'Interval Operations'.

Identify promoter regions containing TAF1 binding sites

Create a new history

3. At the top of the History column click on the Options gear and select **'Create new'**.
4. A new **'Unnamed history'** is created. Rename it to be **'Promoters with TAF1'**. Hit enter to save it.

Extract only the TAF1 binding sites on chromosome 11 from TAF1

In the Loading Data section, we already obtained the genomic intervals of the TAF1 binding sites. Copy it to the current history. Now let's extract all the human gene coordinates.

1. From the new history, click on the gear icon 'History Options' and then **'Copy Datasets'**
2. Select **'Loading Data from library'** as Source History and tick the box next to **'TAF1_ChIP'** and select **'Promoters with TAF1'** as Destination History. Click on **'Copy History Items'**.
3. Then click on the Refresh history icon to see the dataset in your current history.

First, let's focus on chromosome 11 and extract only the TAF1 binding sites on chromosome 11 from TAF1.

1. Select in Tools **'Filter and Sort'** > **'Filter'**
2. Select Filter: **'TAF1_ChIP'**; with following conditions: **'c2=='chr11''**; leave Number of header lines to skip as 0; and Click **'Execute'**.
3. Rename the output interval file to **'TAF1_ChIP_chr11'**, and click **'Save'**.

Obtain the coordinates of all the genes on Human chromosome 11.

1. Go to **'Get Data'** > **'UCSC Main'**
2. Select
    a. genome: 'Human';
    b. assembly: 'Feb. 2009 (GRCh37/hg19)';
    c. group: **'Genes and Gene Predictions'**;
    d. track: **'RefSeq Genes'**;

e. table: 'refGene';
f. region: select **'position'**; enter **'chr11'**
g. output format: 'BED - browser extensible data';
h. tick Send output to 'Galaxy'.
i. Make sure that your settings are exactly the same (in particular, region should be set to **'position' 'chr11'**, output format should be set to **'BED - browser extensible data'**, and **'Galaxy'** should be checked by Send output to option). Click **'get output'**.

7. On next page make sure under Create one BED record per is set to **'Whole Gene'** and click **'Send Query to Galaxy'**.
8. On the right hand side panel, in the history, once the newly created file "UCSC Main on Human knownGene (genome)" turns green, click the pencil **'Edit Attribute'** to change the file name to be '**UCSC_Human_refGene_chr11**', and click **'Save'**.
9. Examine the output bed file, it should contain 2,946 regions.

Retrieve 1000 bp upstream regions of each gene.

1. Select in Tools **'Operate on Genomic Intervals'** > **'Get flanks'**
2. Select data '**UCSC_Human_refGene_chr11**' and set the Length of the flanking regions to **'1000'**. Click **'Execute'**.
3. Rename the output table to be '**UCSC_Human_RefGene_chr11_Promoters**', and click **'Save'**.

Identify the promoter regions containing TAF1 binding sites.

1. Change datatype of **'TAF1_ChIP_chr11'** from tabular to interval by editing the attributes of this dataset (pencil icon in history), click on tab **'Datatype'** , select for New Type: **'interval'** and click on **'Save'** button. Click on tab **'Attributes'**, and enter Number of comment lines: **'1'**, select Chrom column: **'2'**, Start column **'3'**, End column **'4'**. Database/Build should be set to 'Human Feb. 2009 (GRCh37/hg19) (hg19)'. Click **'Save'**.
2. Select in Tools **'Operate on Genomic Intervals'** > **'Join'**
3. Join '**UCSC_Human_RefGene_chr11_Promoters**' with **'TAF1_ChIP_chr11'** with min overlap set to '1', and Return set to 'Only records that are joined (INNER JOIN)'. Then click **'Execute'**.
4. Rename the output to **'Join_Promoters_TAF1s_chr11'**, and click **'Save'**.
5. It returns 13 regions.

Examine these intervals in UCSC genome browser.

1. Select in Tools **'Graph/Display Data'** > **'Build custom track'**.
2. Click **'Add new Track'**, select Track 1: **'TAF1_ChIP_chr11'**, name: **'TAF1'**, color: **'Blue'**;
3. Click **'Add new Track'**, select Track 2: '**UCSC_Human_RefGene_chr11_Promoters**'; name: **'Promoters'**; color: **'green'**;
4. Click **'Add new Track'**, select Track 3: **'Join_Promoters_TAF1s_chr11'**, name: **'Overlaps'**; color: **'Red'**.
5. Click **'Execute'**. This returns a Build Custom Track.
6. Click on the returned dataset to expend it. Then click Display at UCSC main.
7. View the custom track file in the UCSC genome browser. Zoom into region **chr11:1,856,968-1,861,997** to check which gene's promoter region contain TAF1 sites.

EXERCISE 4.1. Repeat above steps with Intersect function and compare the difference between the outputs.

Retrieve shared data.

The history of this section can be found in **'Shared Data'** > **'Published Histories'**. It is called 'Promoters with TAF1'.

Finding coding exons with highest SNP density

The goal here is to list all the exons on chromosome 22, sort them descendingly by the number of single nucleotide polymorphism they contain, and name the gene symbols associated with the top 100 exons in the list.

Outline:

- Retrieve all Human coding exons on chr22 from UCSC.
- Retrieve all Human SNP data on chr22 from UCSC.
- Joining exons with SNPs to find the intersections between these two files.
- Count the number of SNP falling in each exon and rank the exons by their associated SNP density.

Create a new history

1. At the top of the History column click on the Options gear and select **'Create new'**.
2. A new **'Unnamed history'** is created. Rename it to be **'Exons with SNPs Density'**. Hit enter to save it.

Retrieve all Human coding exons on chr22 from UCSC.

1. From the new history, click on the gear icon 'History Options' and then **'Copy Datasets'**
2. Select **'Loading Data from library'** as Source History and tick the box next to '**UCSC_Human_refGene_chr22_Exons**' and select **'Exons with SNPs Density'** as Destination History. Click on **'Copy History Items'**.
3. Then click on the Refresh history icon to see the dataset in your current history.

Examine the output bed file, it should contain 8,379 regions.

Retrieve all Human SNP data on chr22 from UCSC.

1. Go to **'Get Data'** > **'UCSC Main'**
2. Select
    a. genome: 'Human';
    b. assembly: 'Feb. 2009 (GRCh37/hg19)';
    c. group: **'Variation'**;
    d. track: **'Common SNPs(138)'**;
    e. table: **'snp138Common'**;
    f. region: position type **'chr22'**;
    g. output format: **'BED - browser extensible data'**;
    h. tick Send output to **'Galaxy'**.
    i. Make sure that your settings are exactly the same then click **'get output'**.
4. On next page make sure under Create one BED record per is set to **'Whole Gene'** and click **'Send Query to Galaxy'**.
5. Rename the output dataset to '**UCSC_Human_chr22_SNPs**', and click **'Save'**.
6. Click the eye logo to view the returned file. It contains ~180,000 records.

Joining exons with SNPs

1. Go to **'Operate on Genomic Intervals'** > **'Join'**
2. Select Join: '**UCSC_Human_chr22_Exons**' with: '**UCSC_Human_chr22_SNPs**' with min overlap: '**1**', Return: **'Only records that are joined'**, and click **'Execute'**.
3. Rename the returned file '**UCSC_Human_chr22_Exons_SNPs_join**'. It contains 4,406 regions.

Counting the number of SNPs per exon

1. This can be easily done with the **'Join, Subtract, and Group'** > **'Group'** tool.
2. Select data '**UCSC_Human_chr22_Exons_SNPs_join**'.
3. Choose column 4 by selecting **'c4'** in Group by column.
4. Then click on **'Add new Operation'**
5. Operation 1 Type: **'Count'**, On column: **'c4'**, and click **'Execute'**.
6. Rename the result dataset to **'UCSC_Human_chr22_Exons_SNPs_join_count'** and click **'Save'**.
7. The result dataset contains two columns and 2,555 lines. The first contains the exon name while the second shows the number of times this name has been repeated in the dataset '**UCSC_Human_chr22_Exons_SNPs_join**'.

Sorting Exons by SNPs counts

To see which exon has the highest number of SNPs we can simply sort the dataset **'UCSC_Human_chr22_Exons_SNPs_join_count'** on the second column in descending order.

1. This is done with **'Filter and Sort'** > **'Sort'** tool.
2. Select Dataset: **'UCSC_Human_chr22_Exons_SNPs_join_count'**, on column **'c2'**, with flavor 'Numerical sort', everything in: 'Descending order', the click on **'Execute'**
3. You can now see that the highest number of SNPs per exon is 30.
4. Rename the dataset to **'UCSC_Human_chr22_Exons_SNPs_join_count_sort'**

Selecting top one hundred

Now let's select top 100 exons with the highest number of SNPs.

1. For this we will use **'Text Manipulation'** > **'Select First'** tool.
2. Select first: **'100'** from **'UCSC_Human_chr22_Exons_SNPs_join_count_sort'**
3. Rename the dataset to '**UCSC_Human_chr22_Exons_SNPs_join_count_sort_top100'**

Recovering exon information and displaying data in genome browsers

Now we know that in this dataset the top one hundred exons contain between 5 and 30 SNPs. But what else can we learn about these? To know more we need to get back the positional information (coordinates) of these exons. This information was lost at the grouping step and now all we have is just two columns.

To get coordinates back we will match the names of exons in dataset '**UCSC_Human_chr22_Exons_SNPs_join_count_sort_top100'** (column 1) against names of the exons in the original dataset '**UCSC_Human_refGene_chr22_Exons**' (column 4).

1. This can be done with **'Join, Subtract and Group'** > **'Compare two Datasets'** tool
2. Compare: '**UCSC_Human_refGene_chr22_Exons**' Using column: **'c4'** against:

'**UCSC_Human_chr22_Exons_SNPs_join_count_sort_top100'** and column: **'c1'** To find **'Matching rows of 1st dataset'**, then click **'Execute'.**

3. Rename the result file to **'UCSC_Human_chr22_Exons_top100_coodinates'**.
4. The best way to learn about these exons is to look at their genomic surrounding. There is really no better way to do this than using genome browsers. Because this analysis was performed on 'standard' human genome, you have two choices: UCSC Genome Browser and Ensembl.
5. For example, clicking on **'display at UCSC main'** will show your regions, look at 'User Track' on top of browser image.

Retrieve shared data.

The history of this section can be found in **'Shared Data'** > **'Published Histories'**. It is called 'Exons with SNPs Density'.

Something To think about

Why not use the Intersect function here? What happens if you use Intersect here? What happens when you swap the order of the two files to be intersected?

The resulting table by the Join function Exons_SNPs_joined is a combination of 6 columns of the exon table and 6 columns of the SNP table. It returns n-to-n matches between the exon records and the SNP records, i.e. every exon can map to multiple SNPs and every SNPs can map to multiple exons. The resulting table by the "intersect" function returns 1-to-n matches, and every record of the first input file are compared against all the records of the second input file. Also, the results of Intersect do not contain information on both exons and SNPs.

EXERCISE 4.2. Apart from the above functions that we have introduced, there are also the following useful genomic interval functions: Subtract, Merge, Base coverage, Complement, Cluster etc... Practice these functions with your existing datasets.

Below are relevant tutorials for you to view in your own time:

- http://screencast.g2.bx.psu.edu/GOPS_Cluster/
- http://screencast.g2.bx.psu.edu/quickie5_join/flow.html

Section 5. Workflows

The Galaxy workflow system allows users to create pipelines of analysis that they can then use to re-run the same analysis steps as often as they require. They can also share the workflows between other Galaxy users (e.g. users in the same lab or collaborators at distant institutions) who can then import the workflow and then run the analyses using the same parameters. Galaxy workflows can also be published as supplementary data for papers, so making all analysis totally transparent.

Create workflow from history

One of the powerful functions of Galaxy is its ability to allow you to extract a workflow from the tasks you have completed and to re-apply the workflow to similar tasks repeatedly.

1. In History panel, click **'Options'** > **'Saved Histories'**. Select the history which you would like to extract the workflow process from. In this case, let's select the task we completed earlier from the history named **'Exons with SNPs Density'**.
2. Once the datasets of this history are all loaded and can be viewed from the History panel, click **'Options'** > **'Extract workflow'** to start the extraction process.
3. Rename the workflow to **'Finding_Exons_with_highest_SNP_density'**.
4. All the tools and history items used in your current history will appear in the main panel. It shows that the first steps of loading data from UCSC cannot be automated, and the output files from these steps need to be provided to the workflow. Following that, the steps which could be included are: Join, Group, Sort, Select first, and Compare two Datasets. Tick these functions by order to include them in the new workflow. If there are tools or datasets you don't want to use, uncheck them.
5. Click **'Create workflow'.**

Modify the existing workflow.

1. Click **'Workflow'** from the Menu to go to the workflow view, select the workflow **'Finding_Exons_with_highest_SNP_density'** to modify, and click on **'Edit'**.
2. In the workflow view, each box represents either an input/output file or a function in the process. The arrows between the boxes represent the input and the output directions and the order of the process. The relationship between two boxes can be modified or removed by dragging and adding or deleting the arrows between the boxes.
3. Select the function **'Select first'** to modify, in the right panel called Details, change the parameter 'Select first:' from '100' to **'10'** to have the top 10 records to be selected.
4. When finishing adjusting all the parameters of all the functions on the workflow canvas, do not forget to save it by clicking on the gear icon and select **'Save'**.

Apply the workflow on a new dataset.

Workflow can be run over and over again on any suitable datasets.

1. Create a new history called **'Run Exons with SNPs Density on chrX'**
2. Download all Human coding exons of chrX instead of chr22, and save it as **'UCSC_Human_refGene_chrX_Exons'** from UCSC.
3. Download all Human SNP dataset of chrX, and save it as '**UCSC_Human_chrX_SNPs**'.
4. Click **'Workflow'** in the Menu, select the workflow **'Finding_Exons_with_highest_SNP_density'** to be run, click on the drop down menu and select **'Run'**.

5. On the 'Running workflow' page select the input data files.
6. Select for input dataset **'UCSC_Human_refGene_chrX_Exons'** and for input dataset '**UCSC_Human_chrX_SNPs**'. Then click **'Run workflow'**. All the steps will be sequentially executed without any further interventions needed.
7. When all the jobs have finished, click on the last dataset in your history to see the result file.

Share the workflow.

You could share the workflow. Go to the workflow view and select the workflow you would like to share. Click the dropdown menu and select **'Sharing'**. Same as sharing the histories, you can share your workflow either through a public accessible link or with specific users.

Retrieve shared data.

The workflow used in this section can be found in **'Shared Data'** > **'Published Workflows'**. It is called 'Finding_Exons_with_highest_SNP_density'**.**

The history of this section can be found in **'Shared Data'** > **'Published Histories'**. It is called 'Run Exons with SNPs Density on chrX'.

Create a new workflow

Select 50 longest exons from a list of exons.

1. Go to the workflow view, click **'Create new workflow'** button, name it **'Select_50_longest_Exons'** and click **'create'**. This will take you to a blank workflow canvas, where we will create our workflow.
2. To get your first tool onto the canvas just click on it, so go to the tool section **'Text Manipulation'** and click on the tool **'Compute'** to calculate the length of exons.
3. A box entitled 'Compute' will appear on your canvas. The box should be surrounded in blue, which means that it's the currently selected box.
4. Put your cursor over the top part of the box and you should notice that it changes from a pointer to a cross. Click on the box and drag it to the top left corner of the canvas. You'll also notice that the box has a cross on it in the top right corner. Clicking on the cross will remove the tool from the canvas.
5. In the right-hand column (where your history usually is) are the details and parameters associated with the tool. We will edit some of these parameters.
6. Then, add the input dataset for this new workflow. On the left hand panel, at the bottom of all the tools under Workflow control, click **'Inputs'** and then **'Input dataset'** to select the input file for this workflow. A box entitled 'Input dataset' will appear on your canvas.
7. To connect the Input dataset and the Compute tool, simply click the outward arrow of the dataset, hold and drag to the inward arrow symbol of the tool. A green arrow indicates the input dataset datatype is compatible with the function. You can remove the connection by mousing over the connection and clicking on the cross.
8. Add next functions of this workflow in order: **'Filter and Sort'** > **'Sort'**, and **'Text manipulation'** > **'Select first'**. Now we have a workflow of three steps: Compute, Sort, and Select first. Connect the output file from each previous function to the next function as an input.
9. Connect the output of the Compute tool with the input of the Sort tool; and the output of the Sort tool to the input of the Select first tool.
10. Now we are going to adjust the parameters for each tools by clicking on each tool box.
11. Click on the **'Compute'** tool, the input file has two columns that indicates the starts of an

exon (c2) and its ends (c3). Therefore c3-c2 returns the length of that exon. In the right-hand column, under 'Add expression', you should have **'c3-c2'**. This will create a 7th column into the output dataset.

12. Click on the **'Sort'** tool, in the right-hand column under Sort Dataset on column enter **'7'** to order the length of the exon calculated at the previous step (be careful not to add a new column selection by hitting enter).
13. Click on the **'Select first'** tool, in the right-hand column under Select first enter **'50'** to only get the first 50 results.
14. Go to the options at the top of the canvas, and click on **'Save'** to save the entire workflow.

Congratulations - you've just created your first workflow and will save yourself lots of time!

Running a workflow

1. Go to **'Analyze Data'**
2. Create a new history called **'Run Select 50 Longest Exons'**
3. Retrieve all Human coding exons of chr22 from data library 'GalaxyCam Training' > 'loading_data', called '**UCSC_Human_refGene_chr22_Exons'**.
4. Go to the workflow view, select the workflow you've just created called **'Select_50_longest_Exons'**, click on the drop down menu and select **'Run'**.
5. Select '**UCSC_Human_refGene_chr22_Exons'** as the input dataset, then click on **'Run workflow'.**
6. When all the jobs have finished, click on the last dataset in your history to see the result file. Examine the output, the last column shows the exon length by descending order.

Retrieve shared data.

The workflow used in this section can be found in **'Shared Data'** > **'Published Workflows'**. It is called 'Select_50_longest_Exons'**.**

The history of this section can be found in **'Shared Data'** > **'Published Histories'**. It is called 'Run Select 50 Longest Exons'.

Importing a workflow

Please register and view this session in Galaxy main portal: https://usegalaxy.org

From the Galaxy main portal you can view and access all the published workflows and publish your own workflows there to share with others.

https://main.g2.bx.psu.edu/workflow/list_published

In addition to the Dataset, History, and Workflow objects, Galaxy also allows Page object to be published and shared. The Galaxy pages feature allows the creation of documents that integrate datasets, histories, and workflows, and it is often used as a complete document to record an entire analysis for a publication. This enables others to easily access and repeat the analysis. For example, Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study:

https://usegalaxy.org/u/aun1/p/heteroplasmy

Try to download one of the published workflow and view it.

1. Go to the project page https://usegalaxy.org/u/aun1/p/heteroplasmy, and find '3.1 workflows'. Select the first workflow which deals with PCR duplicate data, and click the green import workflow button.
2. When the import step completes, a message will show you the option to start using this workflow. Click **'Edit'** to view the graph of the entire workflow of this project. You can modify the parameters and apply this workflow to either the project dataset or your own dataset. Explore further options when you have the time.

EXERCISE 5.1. Extract the genomic sequence of the 50 basepair region flanking each of the SNPs on human chromosome Y, and then package these steps into a reusable workflow.

Hint. Outline of steps:

- Extract coordinates of SNPs on Human chromosome Y.
- Extract coordinates of flanking regions around these SNPs.
- Extract genomic sequences of these flanking regions.
- Extract a workflow out of this task and share it.

Section 6. Data visualisation.

Biological visualisation

As we have already experienced, we can visualise data using UCSC genome browser but biological data can also be visualised in the Galaxy in-built genome browser called *Trackster*. Trackster is for the high-throughput sequencing era. It deals particularly well with very large data sets, and numerous simultaneous tracks.

Visualise in Trackster.

1. Go to **'Shared Data'** > **'Published Histories'**. Import in your history 'Promoters with TAF1'.
2. Click on data set **'TAF1_ChIP_chr11',** then **'Visualize'** chart icon and **'Trackster'**.
3. Enter the name **'TAF1_ChIP_chr11_Vis'**
4. Click on the plus icon 'Add tracks' and select '**UCSC_Human_RefGene_chr11_Promoters**' and **'Join_Promoters_TAF1s_chr11'.** Click **'Add'**.
5. Select **'chr11'** from the drop down menu.
6. Zoom into region **chr11:1,856,968-1,861,997.**
7. Track can be configured by clicking on the gear icon next to them when going over their names.
8. Do not forget to click **'Save'** to keep your visualisation within Galaxy for viewing it another time or sharing it with others.

Retrieve shared data.

The visualisation created in this section can be found in **'Shared Data'** > **'Published Visualizations'**. It is called 'TAF1_ChIP_chr11_Vis'**.**

Numerical visualisation

Tabular results can be visualised with Galaxy using Charts.

Unclustered Heatmap

1. Create new history named **'Charts'**
2. Get data from this URL: http://www.compsysbio.org/bacteriome/dataset/functional_interactions.txt
3. Click on the dataset newly uploaded to expend it. Click on the Visualize chart icon, and select **'Charts'** to make a new chart.
4. Give it a name **'Unclustered Heatmap'**
5. Double click on **'Heatmap'** type
6. Choose Column: 1 for Column labels; Column: 2 for Row labels; and Column: 3 for Observation.
7. Click on **'Draw'** in the right corner.

To retrieve your graphs, you can go to the top menu **'Visualization'** > **'Saved Visualizations'**, 'Unclustered Heatmap' should be listed**.**

Clustered Heatmap

1. Make a new chart from the same dataset called

http://www.compsysbio.org/bacteriome/dataset/functional_interactions.txt
2. Give it the name **'Clustered Heatmap'**
3. Double click on **'Clustered Heatmap'** type
4. Choose Column: 1 for Column labels; Column: 2 for Row labels; and Column: 3 for Observation
5. Click **'Draw'**
6. Use the mouse wheel or your touch pad to zoom into the bottom left hand side
7. Tooltips pop up if you move the mouse pointer over a box. Find the interaction between B4143 and B3295, you should be able to see it associated value of 0.271021.
8. Click on the **Editor** icon in the right corner to further customize this chart.
9. Go to the **Configuration** tab.
10. Paste a database URL into the template URL field and add the __LABEL__ tag. You may use http://www.ncbi.nlm.nih.gov/geoprofiles/?term=__LABEL__ or any other database. Click on **'Draw'** to redraw the chart. Data points are now linked to web sources.
11. Double click on a box and the browser will open two new tabs using the previously defined URL template.
12. Select one element, find the interaction between B4143 and B3295 again. What are the corresponding protein functions?

Analyse the score distribution with a histogram

1. Make a new chart from the same dataset called http://www.compsysbio.org/bacteriome/dataset/functional_interactions.txt
2. Give it a name **'Score Histogram'**
3. Double click on **'Histogram'** type
4. Choose Column: 3 for Observations
5. Click **'Draw'**
6. Click on **Screenshot** and select **'Save as PNG'** to save the chart onto your desktop as a PNG file.

Analyze the protein distribution with a discrete histogram

1. Make a new chart from the same dataset called http://www.compsysbio.org/bacteriome/dataset/functional_interactions.txt
2. Give it a name **'Discrete Histogram'**
3. Double click on **'Discrete Histogram'** type
4. Choose Column: 1 for Observations
5. Click **'Add Data'**, select Column: 2 for Observations
6. Click on **'Draw'**
7. Which proteins have most interactions?
    a. B4143 Chaperon (http://www.ncbi.nlm.nih.gov/geoprofiles/?term=B4143)
    b. B3295 RNA Polymerase (http://www.ncbi.nlm.nih.gov/geoprofiles/?term=B3295)

Retrieve shared data.

The visualisations created in these sections can be found in **'Shared Data'** > **'Published Visualizations'**. They are called 'Unclustered Heatmap', 'Clustered Heatmap', 'Score Histogram' and 'Discrete Histogram'**.**