

An introduction to ChIP-seq analysis with Galaxy

Sebastian Schmeier

*Institute of Natural and Mathematical Sciences
Massey University Auckland, New Zealand*

<http://sschmeier.com>

s.schmeier@gmail.com

2016-03-23

Contents

An introduction to ChIP-seq analysis with Galaxy	3
1.0 Preface	3
1.1 Overview	3
1.2 The task at hand	3
1.3 Log into Galaxy	4
1.4 Create a new history	6
1.5 Loading the data	7
1.6 Investigate the data	9
1.7 Quality assessment	11
1.7.1 FastQC	11
1.7.2 Read filtering	12
1.7.3 Quality trimming	14
1.8 Mapping reads	15
1.8.1 Bowtie2	16
1.8.2 Post-mapping processing	18
1.9 Peak calling	21
1.9.1 MACS	21
2.0 Post-processing	24
2.1 Overlap peaks with promoter regions	24
2.1.1 Get genes	24
2.1.2 Get promoter	25
2.1.3 Join	26
2.2 Enrichment analysis (genes) with Enrichr	27
2.3 Enrichment analysis (peaks) with GREAT	31
2.4 Visualisation	33
2.5 Motif finding	35
2.5.1 Find the peak center	36
2.5.2 Get flanking regions	38
2.5.3 Extract fasta-sequence	39
2.5.4 Run MEME-ChIP	39
2.6 References	41
2.7 Web links	41

An introduction to ChIP-seq analysis with Galaxy

1.0 Preface

In this brief tutorial we will learn how to use the excellent tool [Galaxy](http://galaxyproject.org/) (<http://galaxyproject.org/>) to analyse data from a chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiment. It is part of a series of introductory tutorials that can be found at <http://sschmeier.github.io/bioinf-workshop/>.

A PDF-version of this tutorial can be downloaded [here](#) or at http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/doc/Galaxy-ChIPseq-Introduction_sschmeier.pdf

Two accompanying lectures for this tutorial are available:

1. ChIP-seq introduction (<http://dx.doi.org/10.6084/m9.figshare.1545468>).
2. ChIP-seq data processing (<http://dx.doi.org/10.6084/m9.figshare.1554130>).

1.1 Overview

In this brief tutorial we will learn how to use the excellent tool [Galaxy](#) to analyse ChIP-seq data. If you are new to [Galaxy](#), you might want to start with the [Galaxy introductory tutorial](http://sschmeier.github.io/bioinf-workshop/#!/galaxy-intro/) (<http://sschmeier.github.io/bioinf-workshop/#!/galaxy-intro/>).

1.2 The task at hand

The overall purpose in this tutorial is to:

- Understand better the [Galaxy](#) system (1.3-1.4).
- Understand how to get your data of interest into the system (1.5).
- Understand how to quality control your sequencing data (1.7).
- Understand how to map sequence reads to a reference genome (1.8).
- Post-mapping processing to clean the data before peak calling (1.8.2)
- Understand how to call ChIP-peaks based on the mapped reads (1.9).
- Understand how to gather additional information about you data (2.0).

In order to develop an understanding of the points above, you will run through the workflow to analyse ChIP-seq data (see *Figure 1*):

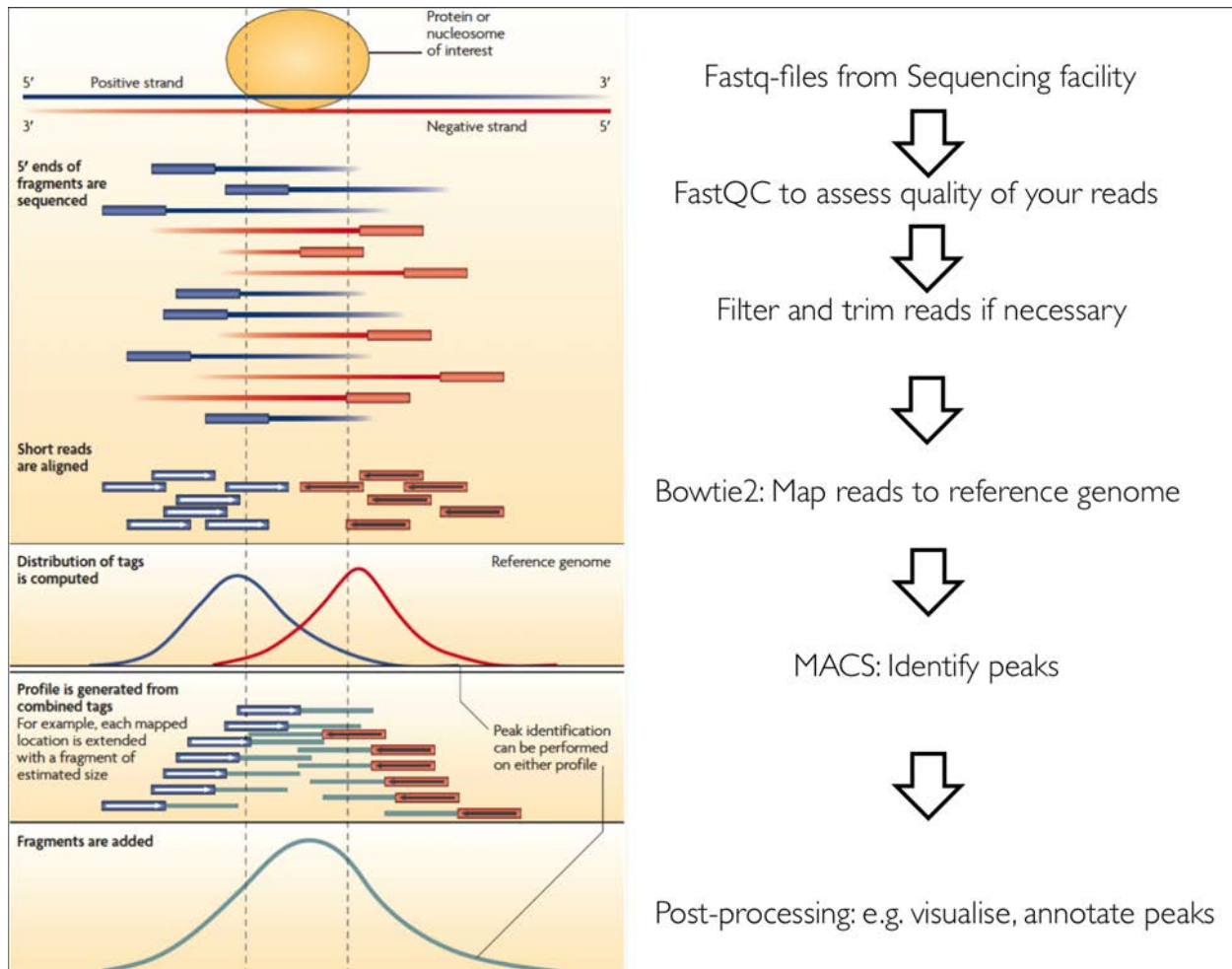


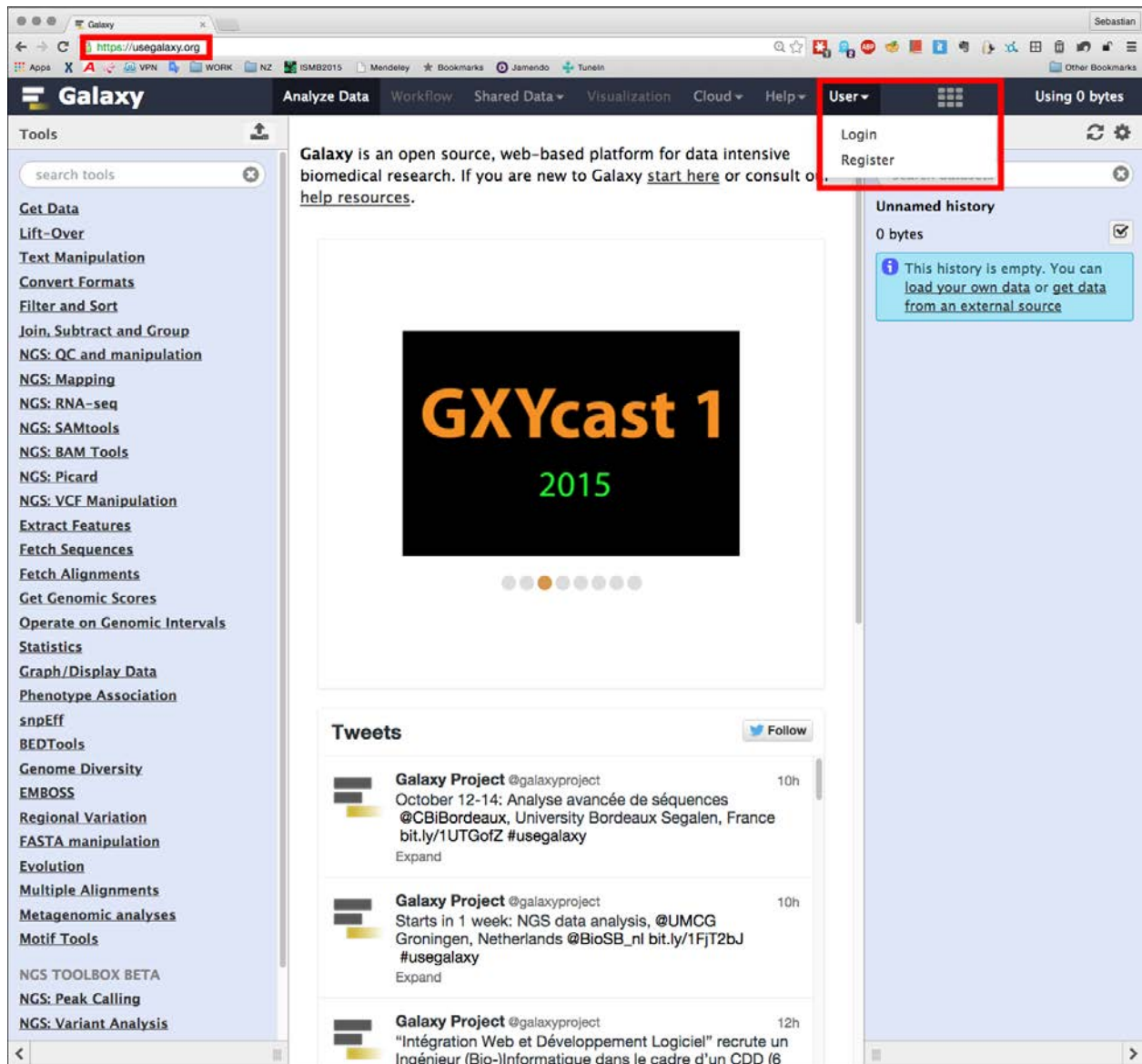
Figure 1: ChIP-workflow (adjusted from Park2009).

The individual tasks are:

1. Load the dataset.
2. Quality assess the reads.
3. Map the reads to the genome using Bowtie2.
4. Call peaks using MACS.
5. Run Enrichr with genes and GREAT with the peak regions to find enriched annotations.
6. Visualise the peaks in UCSC browser.
7. Prepare peak data and use MEME to find TFBS motifs.

1.3 Log into Galaxy

First, go to <https://usegalaxy.org/> and log into your Galaxy account (see Figure 2 and Figure 3).



The screenshot shows the Galaxy web interface in a browser window. The address bar displays <https://usegalaxy.org>. The main navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Cloud", "Help", and "User". The "User" menu is open, showing "Login" and "Register" options. The left sidebar lists various tools under categories like "Get Data", "Text Manipulation", "NGS: QC and manipulation", etc. The main content area features a "GXYcast 1 2015" banner and a "Tweets" section with three tweets from the Galaxy Project. The right sidebar shows "Unnamed history" with 0 bytes and a message: "This history is empty. You can load your own data or get data from an external source".

Figure 2: Log into your Galaxy account.

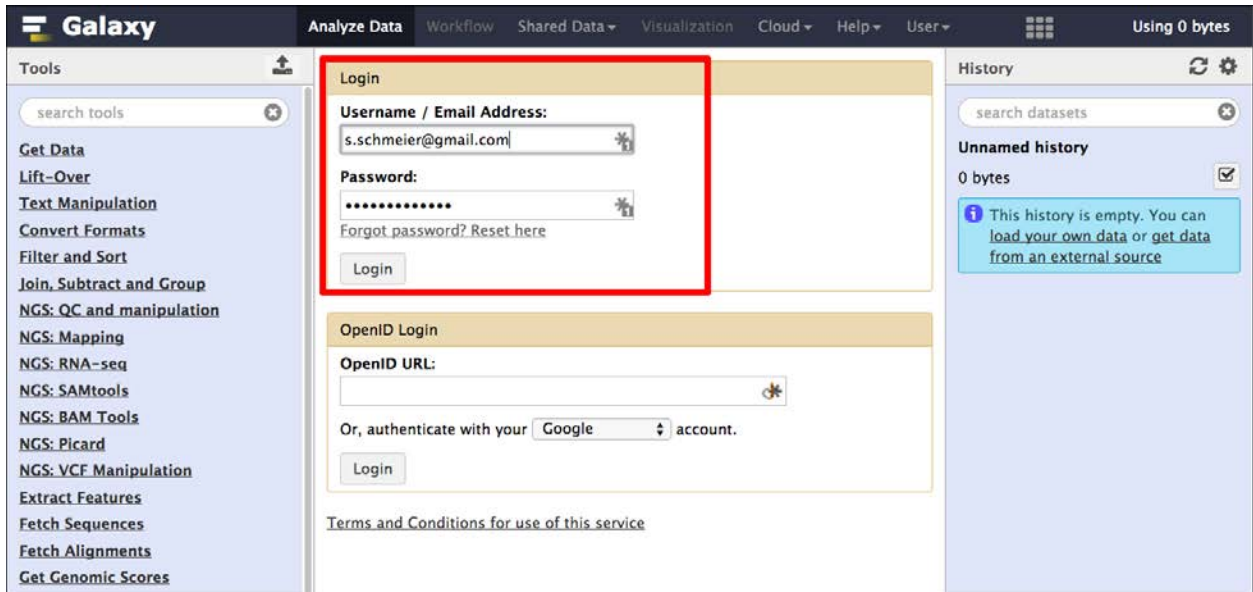


Figure 3: Log into your Galaxy account with your credentials.

1.4 Create a new history

Create a new history (see Figure 4) and rename it to something useful (see Figure 5).

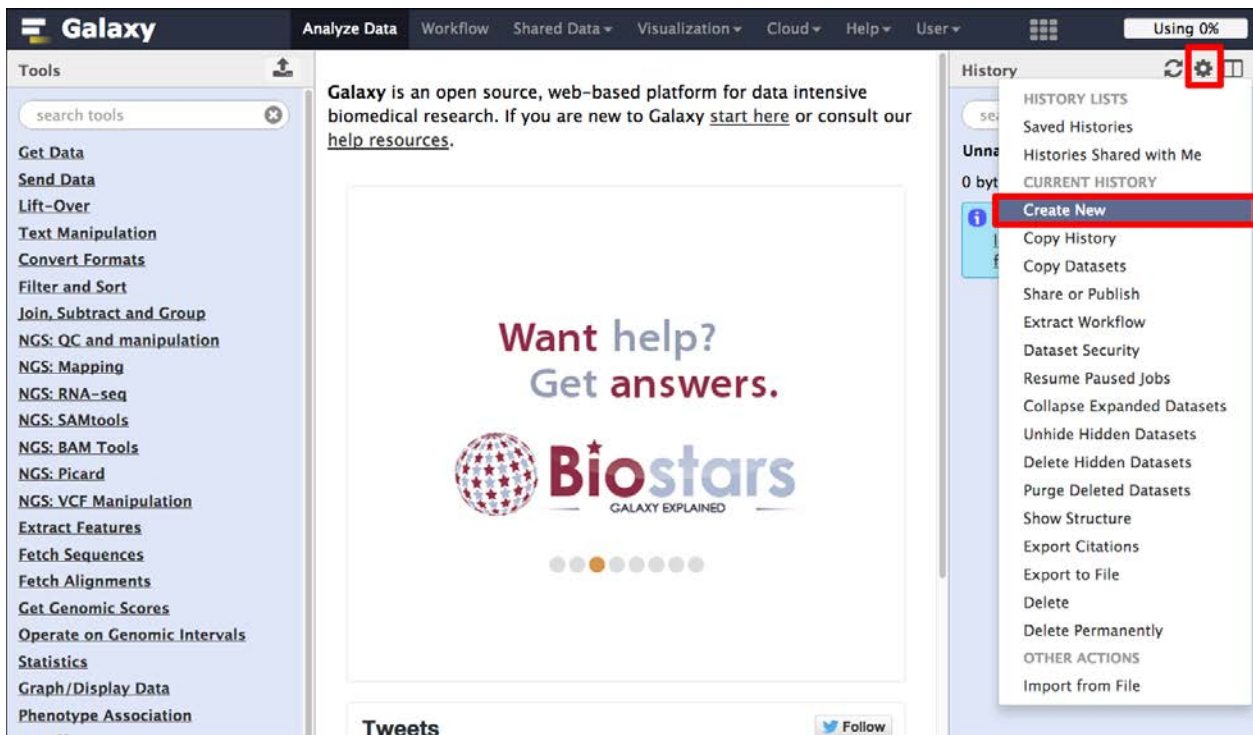


Figure 4: Log into your Galaxy account.

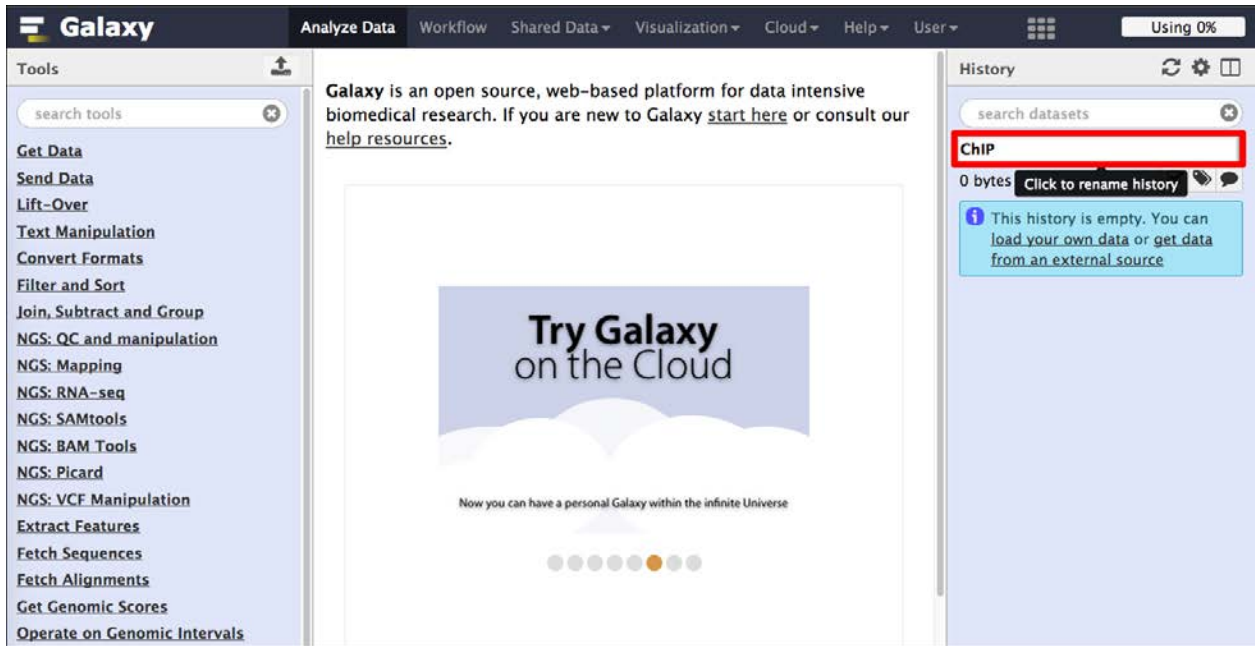


Figure 5: Log into your Galaxy account with your credentials.

1.5 Loading the data

We are going to use some *Shared Data* from the Galaxy Demonstration dataset.

1. Click on the *Shared Data* tab (see Figure 6).
2. Search for the *Demonstration Datasets*. (see Figure 7)

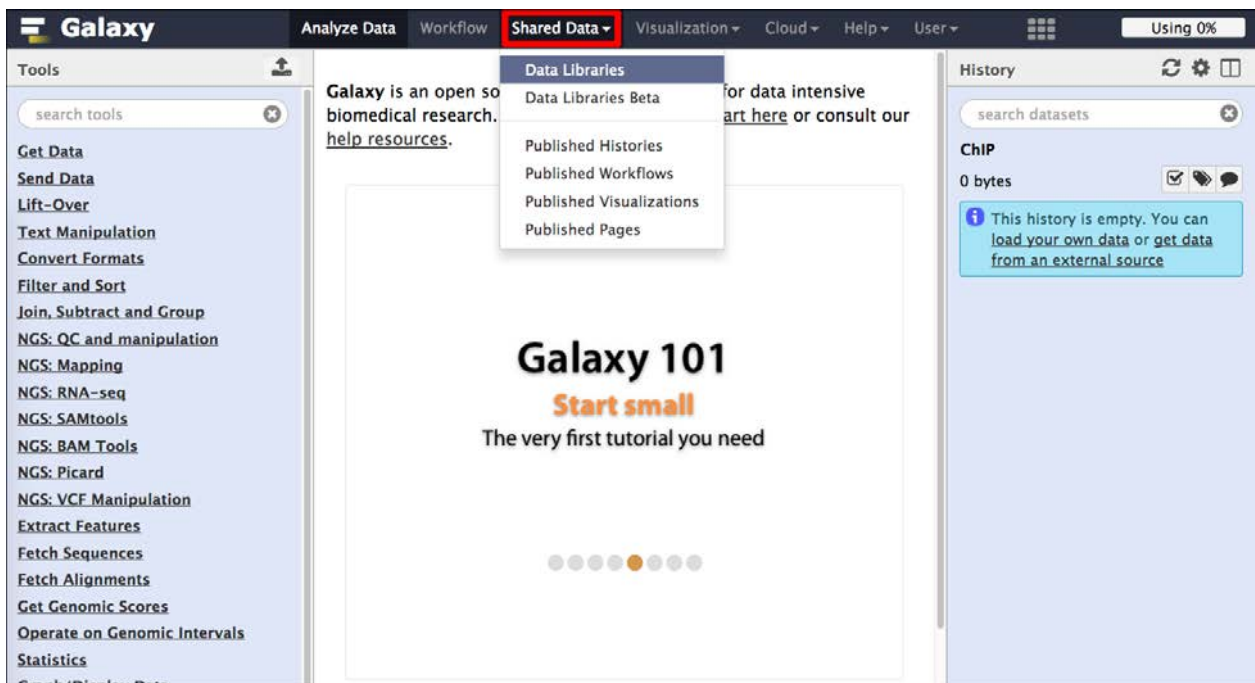


Figure 6: Load shared data tab.

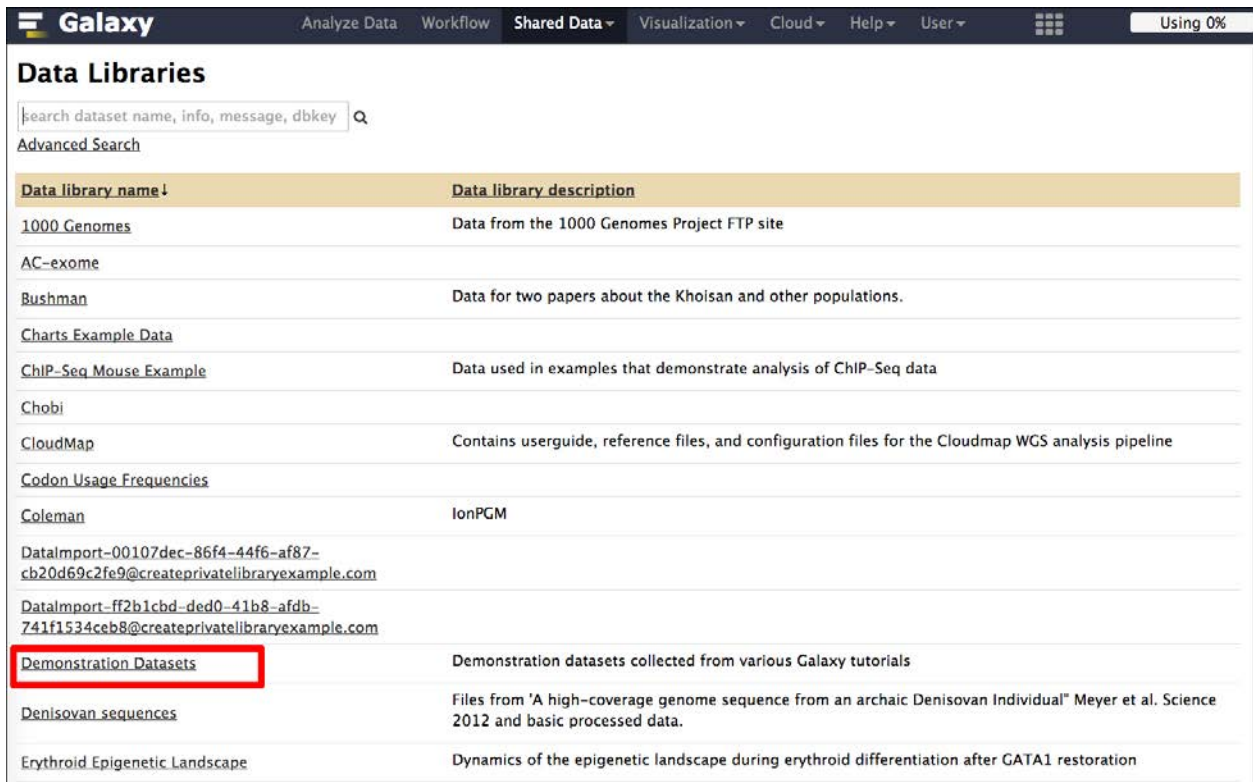


Figure 7: Look for the Demonstration Datasets.

Load the following 4 files: G1E CTCF, G1E_ER4 CTCF, G1E ER4 input, G1E input (see Figure 8).

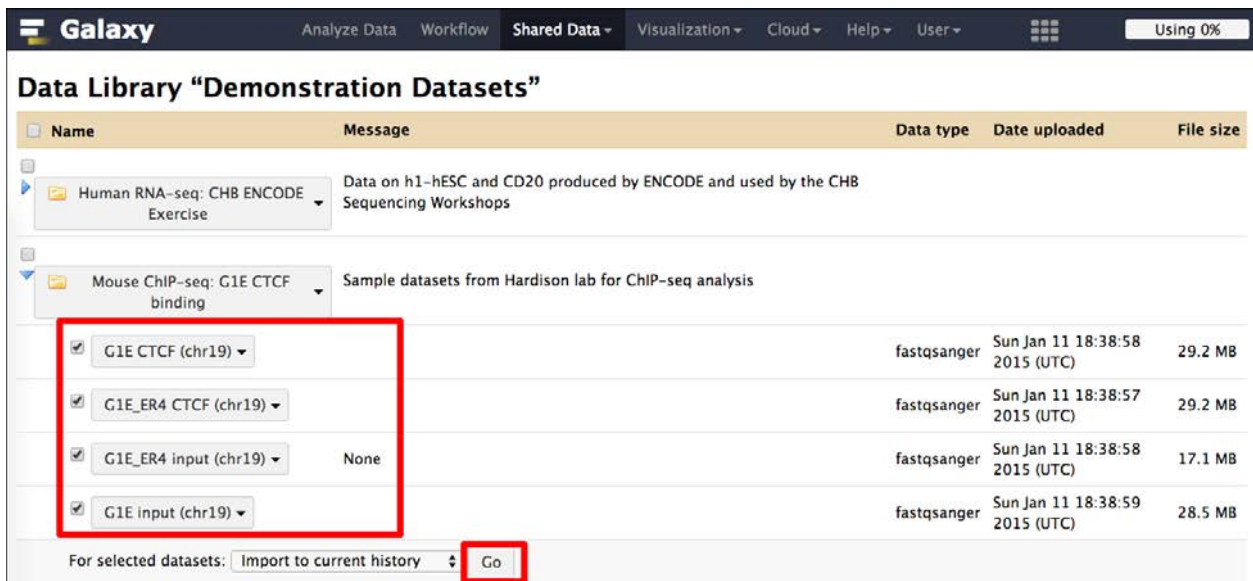


Figure 8: Load the datasets.

Once the files are loaded we can switch back to the analysis window by clicking *Analyze Data* tab (see Figure 9). We should find four datasets in the history panel (see Figure 10).

Name	Message	Data type	Date uploaded	File size
Human RNA-seq: CHB ENCODE Exercise	Data on h1-hESC and CD20 produced by ENCODE and used by the CHB Sequencing Workshops			
Mouse ChIP-seq: G1E CTCF binding	Sample datasets from Hardison lab for ChIP-seq analysis			
G1E CTCF (chr19)		fastqsanger	Sun Jan 11 18:38:58 2015 (UTC)	29.2 MB
G1E_ER4 CTCF (chr19)		fastqsanger	Sun Jan 11 18:38:57 2015 (UTC)	29.2 MB
G1E_ER4 input (chr19)	None	fastqsanger	Sun Jan 11 18:38:58 2015 (UTC)	17.1 MB
G1E input (chr19)		fastqsanger	Sun Jan 11 18:38:59 2015 (UTC)	28.5 MB

Figure 9: Load the datasets.

Figure 10: Loaded data in history panel

Alternatively, you can download the data [chipdata.zip](http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/data/chipdata.zip) or (~40MB) from <http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/data/chipdata.zip>, unzip it and upload the files to the Galaxy history.

Hint! Should you need to refresh how to upload data to Galaxy, have a look at the [Galaxy introductory tutorial \(http://sschmeier.github.io/bioinf-workshop/#!galaxy-intro/\)](http://sschmeier.github.io/bioinf-workshop/#!galaxy-intro/).

1.6 Investigate the data

The four files that we have now in our history are: G1E CTCF, G1E_ER4 CTCF, G1E ER4 input, G1E input. A closer look reveals that they are in fastq-sanger format (see [Figure 11](#) and [Figure 12](#)).

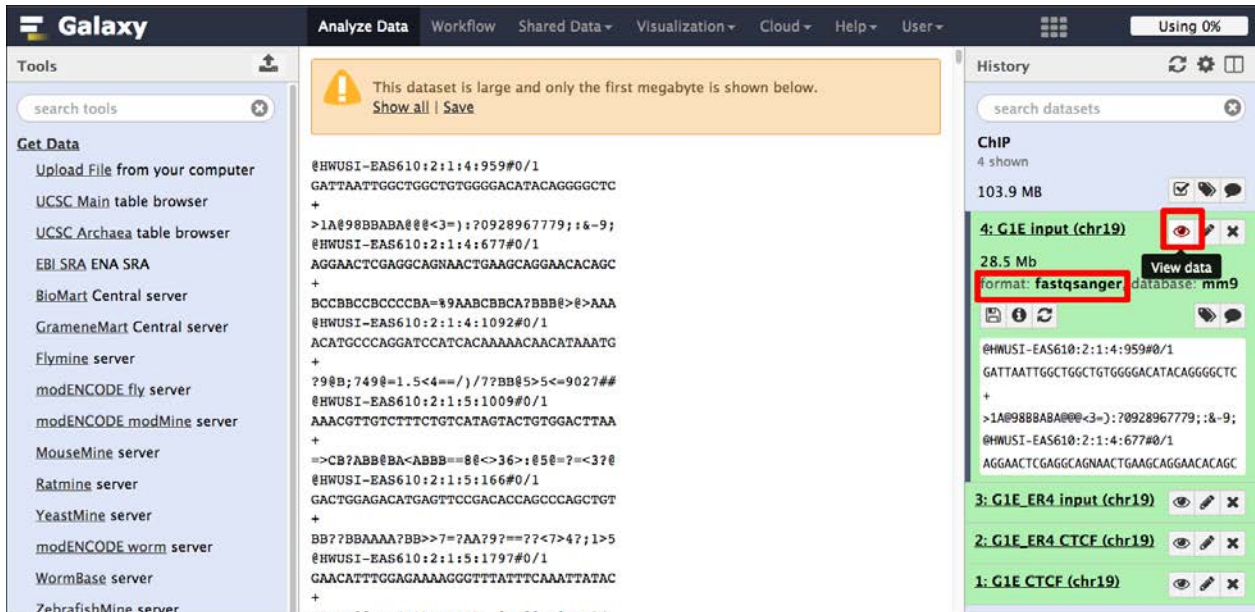


Figure 11: Information about the data.

FastQ-format

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCAGG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTT
+
' '*((( (**+))%%%+)(%%%) .1***-+*' '))*55CCF>>>>
```

Sequence id

Sequence

Phred quality of the corresponding nucleotide (ASCII code)

- Phred quality
- One ASCII character per nucleotide.
- Encodes for a quality $Q = -10 \cdot \log_{10}(P)$, where P is the error probability

The Relationship Between Quality Score and Base Call Accuracy		
Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
$-10 \cdot \log_{10}(0.1) = 10$ (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

Sebastian Schmeier

Figure 12: The fastq-format.

What we are looking at is data from the [G1E mouse cell-line](#) from [Gata1-null mouse embryonic stem cells](#). We are looking at two conditions, normal G1E cell-lines and G1E-ER4 cell-lines, where the effect of Gata1 deletion is restored. Under both conditions Ctf has been ChIP'ed and sequenced. The "input" samples denote samples where the DNA was fragmented but before the immunoprecipitation against Ctf and can thus be used as controls. We

are also only looking at a subset of the full dataset, only chr19. **Thus, we can compare the CTCF occupancy between G1E and G1E-ER4 cell-lines.**

Note!

TODO:

1. Find out what **Ctcf** is.
2. Find out why studying **Gata1** in mouse embryonic stem cells is of interest?

Hint! You can use [NCBI gene](#) or [wikigenes](#) or even [wikipedia](#) to find out about **Ctcf** and **Gata1**.

Attention! **Before you go further:** Step 1.7 up to step 1.9 is the same process for each fastq-file individually. Thus, it might be a good idea to do the steps for **ONE** file and then create a workflow out of the steps that allows you to do all of the steps for each of the other three files in one go. This will significantly speed up the processing time. Should you not remember how to create a workflow, see [Galaxy introductory tutorial \(http://sschmeier.com/bioinf-workshop/#!galaxy-intro/\)](http://sschmeier.com/bioinf-workshop/#!galaxy-intro/)..

1.7 Quality assessment

Now we need to assess the quality of the reads in each sample and filter and quality trim the reads if necessary.

1.7.1 FastQC

First, we run FastQC on each sample to get a feel for the overall quality of the data (see [Figure 13](#)).

Figure 13: FastQC.

Have a look at the HTML result page. Depending on what the results are you might want to do some filtering and quality trimming.

Note!

TODO: Run FastQC on all four files and investigate the quality. Note for each sample the nucleotide number where the quality markedly drops.

1.7.2 Read filtering

Here, we want to get rid of all reads that are of low quality. This strongly depends on your definition for “low quality”. In the figure below the default values are used (see *Figure 14*). The *Quality cut-off* value is 20 and 90% of all nucleotides of the read need to be equal or above this cut-off value to be accepted.

Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools

NGS: QC and manipulation

- FastQC Read Quality reports
- Select high quality segments
- Build base quality distribution
- Draw quality score boxplot
- Quality format converter (ASCII-Numeric)
- Filter by quality**
- FASTQ to FASTA converter
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Draw nucleotides distribution chart
- Compute quality statistics
- Rename sequences

Filter by quality (Galaxy Tool Version 1.0.0) Options

Library to filter: 4: G1E input (chr19)

Quality cut-off value: 20

Percent of bases in sequence that must have quality equal to / higher than cut-off value: 90

Execute

What it does

This tool filters reads based on quality scores.

- Using **percent = 100** requires all cycles of all reads to be at least the quality cut-off value.
- Using **percent = 50** requires the median quality of the cycles (in each read) to be at least the quality cut-off value.

Figure 14: Filtering reads of bad quality.

Furthermore, I edited the dataset name (1) to keep track of the kind of data (see *Figure 15*). The original name I copied into the notes field (2), however it is not strictly necessary as the information from which dataset this one was derived is still available when clicking the info button (see *Figure 16*). Finally, I renamed the dataset to something useful (3, see *Figure 15*).

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

Tools

NGS: Mapping

- Bowtie2 - map reads against reference genome
- BWA - map short reads (< 100 bp) against reference genome
- BWA-MEM - map medium and long reads (> 100 bp) against reference genome
- Parse blast XML output
- Megablast compare short reads against htgs, nt, and wgs databases
- Map with BWA for Illumina
- Map with Bowtie for Illumina
- Lastz map short reads against reference sequence

NGS: RNA-seq

- NGS: SAMtools
- NGS: BAM Tools
- NGS: Picard
- NGS: VCF Manipulation

Edit Attributes

Name: G1E input filtered 3

Info:

Quality cut-off: 20
Minimum percentage: 90

Annotation / Notes: Filter by quality on data 4 2

Database/Build: Mouse July 2007 (NCBI37/mm9) (mm9)

Save

Auto-detect

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

History

search datasets

Copy of 'ChIP' (active items only)
6 shown, 4 deleted
138.6 MB

10: Filter by quality on data 4
21.9 MB
format: fastqsanger, database: mm9

Quality cut-off: 20
Minimum percentage: 90
Input: 276733 reads.
Output: 212756 reads.
discarded 63977 (23%) low-quality reads.

@HWUSI-EAS618-Z:1:4:677#0/1
AGGAAGCTCGAGGCGAAGCTGAAGGAGAACACAGC
+
BCCBCCBCCBCCBA-NBAACBCCA7888B-B-AAA
@HWUSI-EAS618-Z:1:5:1009#0/1
AAACGTTGCTTCTGTCATAGTACTGTGGACTTAA

Figure 15: Rename the dataset to keep track.

The screenshot shows the Galaxy web interface. The main panel displays the configuration for the 'Filter by quality' tool. The tool parameters are as follows:

Input Parameter	Value	Note for rerun
Library to filter	4: G1E input (chr19)	
Quality cut-off value	20	
Percent of bases in sequence that must have quality equal to / higher than cut-off value	90	

The history panel on the right shows the execution details for the tool. The tool name is '10: G1E input filtered'. The input is '21.9 MB' and the output is '138.6 MB'. The tool parameters are listed as 'format: fastqsanger, database: mm9'. The quality cut-off is 20, and the minimum percentage is 90. The input is 276733 reads, and the output is 212756 reads. 63977 (23%) low-quality reads were discarded. The history panel also shows the input and output files, and the number of reads discarded.

Figure 16: Detailed information about a dataset can be gathered by clicking the info button.

Note!

TODO: Run the filtering on all four files and note how many reads got excluded for each sample (see the next section on how to speed this process up by re-running analyses).

Re-running an analysis

Click on the re-run button of the analysis (1) you would like to re-run (see Figure 17). The parameter window pops up with all the original parameters used. Now you can select a different dataset (2) and run the original analysis with the same parameters (see Figure 17).

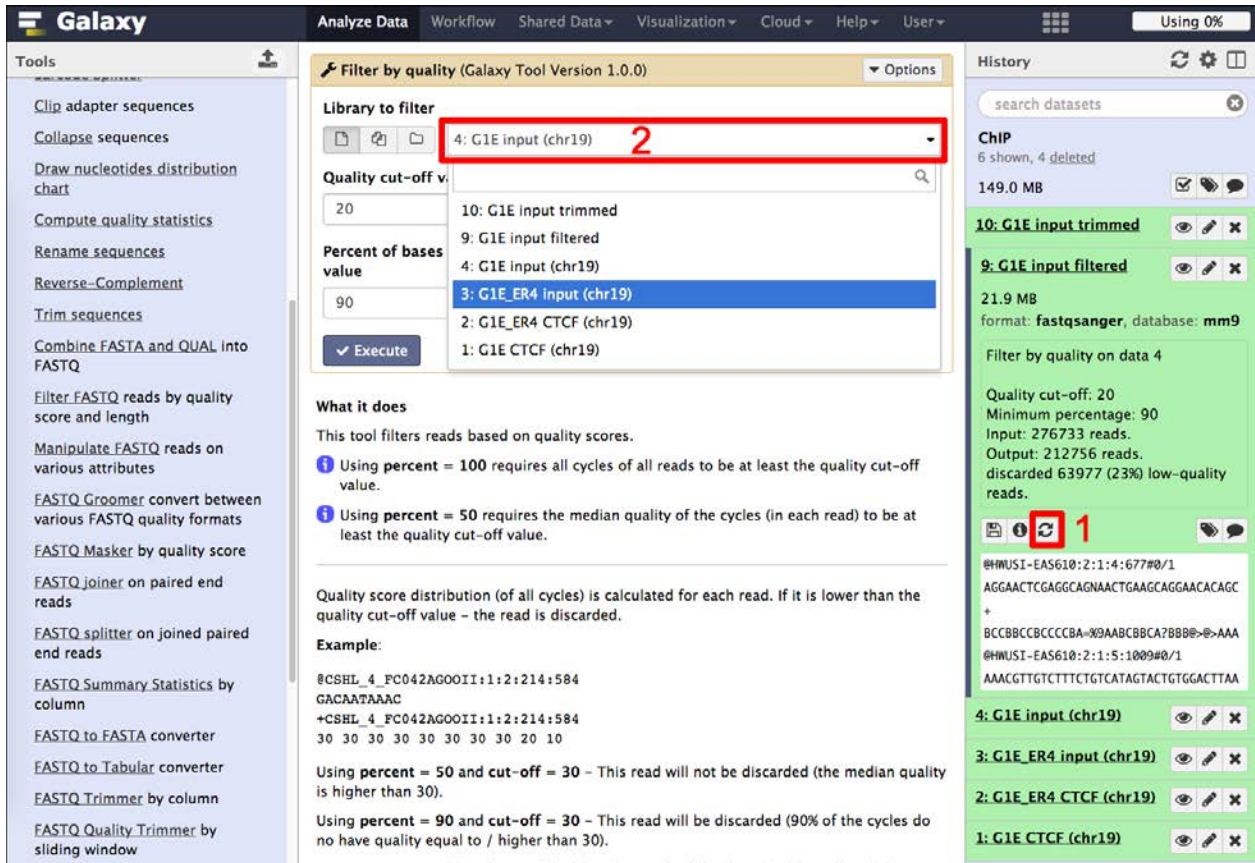


Figure 17: Re-run button to re-run the same analysis.

1.7.3 Quality trimming

Finally, we can use a quality trimmer to get rid of bad starts and ends of reads (see Figure 19). To do so, select the *FASTQ Quality Trimmer* (1). Choose the **filtered** dataset from the step before (2). In Figure 19 I use a simple window size of 1 (3) and a quality score of 20 (4) to just trim of the ends on both sides.

Figure 19: Filtering reads of bad quality.

Note!

TODO: Run the quality trimmer on all *filtered* datasets and rename the sets to something meaningful.

Attention! Trimming reads is not always necessary or desired. Here, we do it to see how the trimming process works in Galaxy. However, in a real situation we might decide not to trim at all.

1.8 Mapping reads

By now we should have 4 sets of filtered and trimmed reads with a meaningful name (see Figure 19). These form the basis for the subsequent analyses. Now we are going to map the reads to the reference genome.

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. The left sidebar contains a 'Tools' panel with a search bar and various tool categories like 'Get Data', 'Text Manipulation', and 'NGS: Mapping'. The main content area displays 'Galaxy 101 Start small' with a progress indicator. The right sidebar shows the 'History' panel with a search bar and a list of datasets. The datasets are numbered 1 to 16, with 10, 14, 15, and 16 highlighted in red boxes. The datasets are:

Dataset ID	Dataset Name	View	Edit	Delete
16	G1E_ER4 input trimmed	👁	✏	✖
15	G1E_ER4 CTCF trimmed	👁	✏	✖
14	G1E CTCF trimmed	👁	✏	✖
13	G1E CTCF filtered	👁	✏	✖
12	G1E_ER4 CTCF filtered	👁	✏	✖
11	G1E_ER4 input filtered	👁	✏	✖
10	G1E input trimmed	👁	✏	✖
9	G1E input filtered	👁	✏	✖
4	G1E input (chr19)	👁	✏	✖
3	G1E_ER4 input (chr19)	👁	✏	✖
2	G1E_ER4 CTCF (chr19)	👁	✏	✖
1	G1E CTCF (chr19)	👁	✏	✖

Figure 19: The datasets for mapping.

1.8.1 Bowtie2

We can now map the trimmed data to the reference genome using Bowtie2. Select Bowtie2 in the tools panel under section *NGS: Mapping* (1, see Figure 19). We select the *trimmed* dataset we want to map (2) and select an appropriate reference genome (3).

Figure 19: Bowtie2.

Using an inbuilt index choose the same genome built as your data is (here we are looking at mouse mm9 data). Choose the canonical index, **mm9 Canonical**.

From the Galaxy Bowtie tool description:

A Note on Built-in Reference Genomes

The default variant for all genomes is “Full”, defined as all primary chromosomes (or scaffolds/contigs) including mitochondrial plus associated unmapped, plasmid, and other segments. When only one version of a genome is available in this tool, it represents the default “Full” variant. Some genomes will have more than one variant available. The “Canonical Male” or sometimes simply “Canonical” variant contains the primary chromosomes for a genome. For example a human “Canonical” variant contains chr1-chr22, chrX, chrY, and chrM. The “Canonical Female” variant contains the primary chromosomes excluding chrY.

Finally, we just use the default parameters of Bowtie2 (4) and execute the analysis (see Figure 19).

We can not look at the resulting data in detail, as the output is in a format called *bam* which is a binary version of the Sequence Alignment/Map (SAM) format (see <http://genome.ucsc.edu/goldenpath/help/bam.html> and <http://samtools.sourceforge.net/> for an explanation). However, by clicking the dataset-name, we get more detailed information about the mapping (see Figure 20)

22: G1E4 input bowtie2

10.2 MB
format: **bam**, database: **mm9**

[samopen] SAM header is present: 22 sequences.
212756 reads; of these:
212756 (100.00%) were unpaired; of these:
1975 (0.93%) aligned 0 times
204916 (96.32%) aligned exactly 1 time
5865 (2.76%) aligned >1 times
99.07% overall alignment

display at UCSC [main](#)
display at Ensembl [Current](#)
display with IGV [local](#) [Mouse mm9](#)
display in IGB [View](#)

Binary bam alignments file

Figure 20: Bowtie2 mapping information.

Note!

TODO: Run Bowtie2 on each of the four trimmed datasets. Note for each sample the number of reads that could be aligned exactly once to the genome and the overall alignment percentage.

1.8.2 Post-mapping processing

First, we need to filter out multi-mapping reads. We will use samtools to do this. The important parameter here is the **Minimum MAPQ quality score** which should be set to **1**, which will remove multi-mapping reads, as reads that multi-map will get a score of 0 (see *Figure 20b*).

The screenshot shows the Galaxy web interface for the tool "Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region (Galaxy Version 1.1.1)". The tool configuration panel includes the following options:

- SAM or BAM file to filter:** 18: G1E CTCF bowtie2
- Header in output:** Include Header
- Minimum MAPQ quality score:** 1 (highlighted with a red box)
- Filter on bitwise flag:** no
- Select alignments from Library:** (empty field)
- Select alignments from Read Group:** (empty field)
- Output alignments overlapping the regions in the BED FILE:** Nothing selected
- Select regions (only used when the input is in BAM format):** (empty field)
- Select the output format:** BAM (-b)

The sidebar on the left shows the "Tools" menu with "NGS: SAMtools" selected, and "Filter SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region" highlighted with a red box.

What it does
This tool uses the samtools view command in SAMtools toolkit to filter a SAM or BAM file on the MAPQ (mapping quality), FLAG bits, Read Group, Library, or region.

Input
Input is either a SAM or BAM file.

Output
The output file will be SAM or BAM (depending on the chosen option), filtered by the selected options.

Options
Filtering by read group or library requires headers in the input SAM or BAM file.
If regions are specified, only alignments overlapping the specified regions will be output. An alignment may be given multiple times if it is overlapping several regions. A region can be presented, for example, in the following format:

Figure 20b: Samtools filtering.

Second, sort the output from the former step (see Figure 20c).

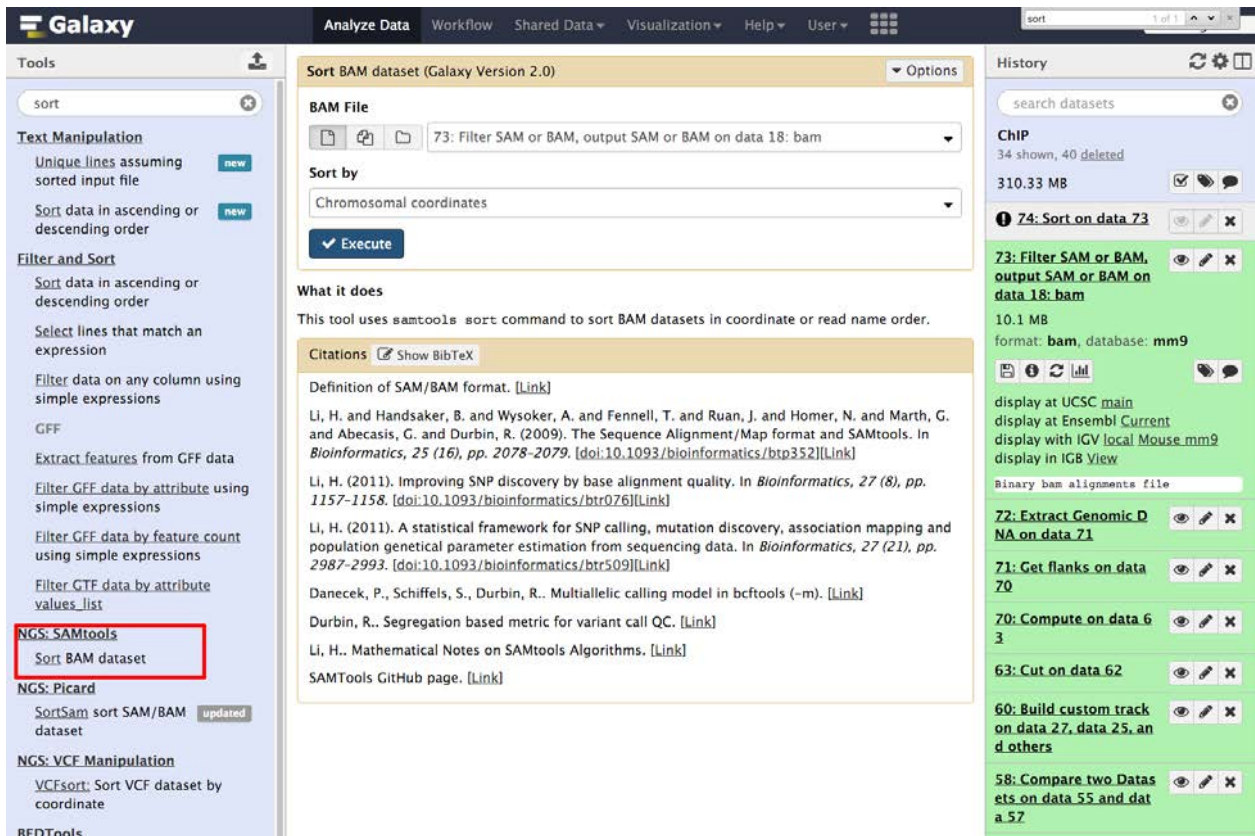


Figure 20c: Samtools filtering.

Third, remove duplicate reads with samtools. Here you need to specify that we are dealing with single-end reads (see Figure 20d).

The screenshot displays the Galaxy interface for the 'RmDup remove PCR duplicates' tool. The tool's configuration is shown in the center, with a red box highlighting the 'Is this paired-end or single end data' dropdown menu, which is currently set to 'BAM is single-end (-s)'. The 'Execute' button is visible below the dropdown. To the left, the 'Tools' panel shows 'NGS: SAMtools' and 'RmDup remove PCR duplicates' under the 'Workflows' section. To the right, the 'History' panel shows a list of datasets, including '74: Sort on data 73', '73: Filter SAM or BAM, output SAM or BAM on data 18: bam', and '72: Extract Genomic DNA on data 71'. The 'What it does' section provides a detailed description of the tool's function and its parameters.

Figure 20d: Samtools rmdup.

1.9 Peak calling

Hint! You should have 4 bowtie2-generated bam-files in your history. If Galaxy did not run your bowtie2 tasks it could be that the queues are full. In this case, please download the Bowtie2 bam-files [here](http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/data/bowtie2-results-bam.zip) or at <http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/data/bowtie2-results-bam.zip>. Unzip the files and upload all files to your Galaxy history and go to 1.8.2 and finally, continue to 1.9.1.

1.9.1 MACS

Select the MACS tool in the **NGS Peak Calling** section:

1. Once you have the tool open (see *Figure 21*), give it a useful name.
2. We are dealing with single-end reads, so select this option.
3. We give it the Bowtie mapped file of the CTCF-ChIP'ed experiment and the "input" of the same cell-line as a control-file.
4. We need to adjust the genome size to that of mm9 Canonical which is 1.87e+9.
5. We also change the tag-size to 36.
6. Finally, we adjust the peak detection method to the "new" one.

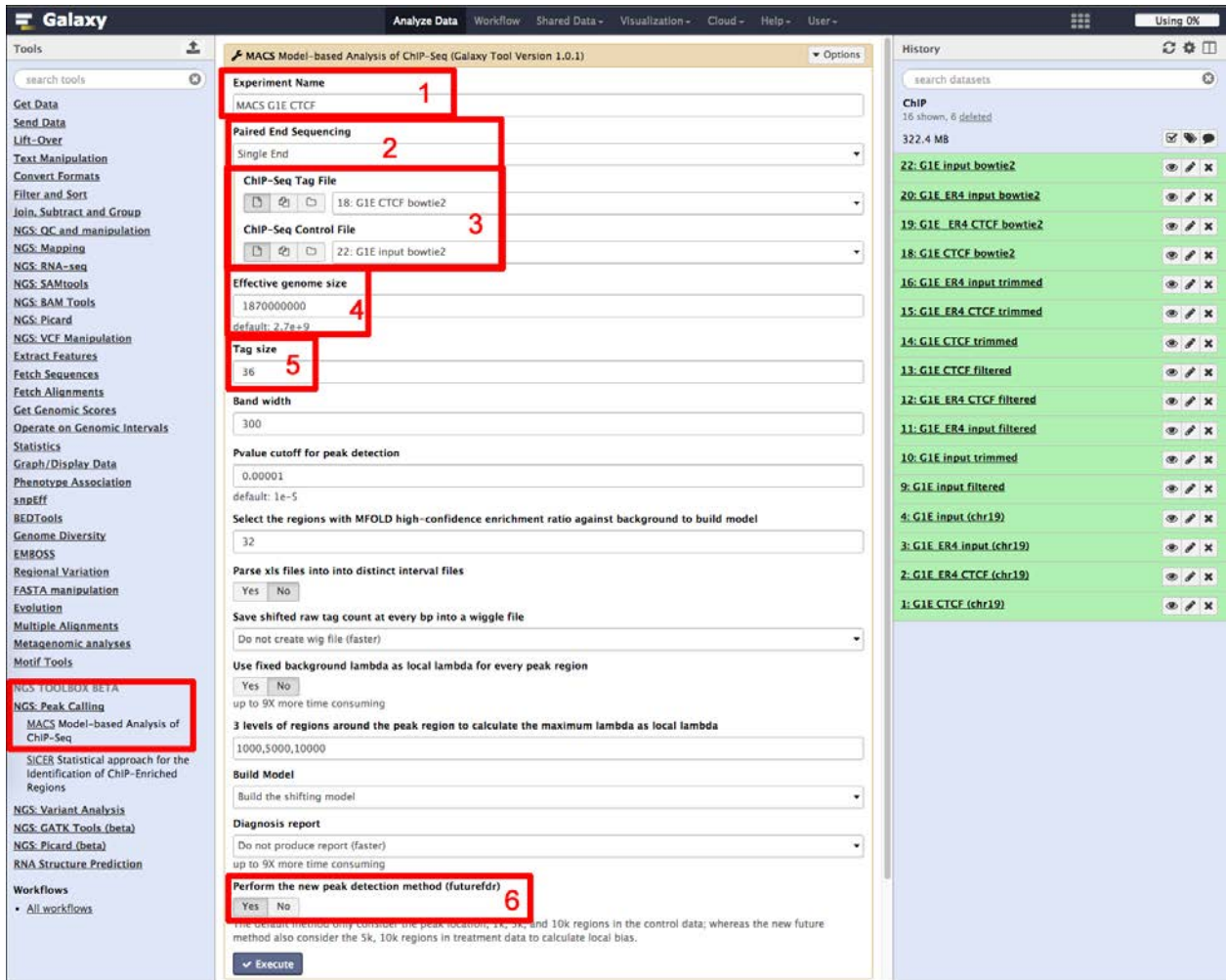


Figure 21: MACS peak calling.

The results of the MACS run are two datasets (see Figure 22). One bed-file that contains the enriched regions and a html-file that provides more information about the MACS run, e.g. we can have a look at the estimated peak model (see Figure 23) or get more information about the peaks in the created xls-file (see Figure 24).

Additional Files:

- MACS_GIE_CTCF_model.pdf
- MACS_GIE_CTCF_model.r
- MACS_GIE_CTCF_model.r.log
- MACS_GIE_CTCF_negative_peaks.xls
- MACS_GIE_CTCF_peaks.xls

Messages from MACS:

```

INFO @ Mon, 21 Sep 2015 23:36:23:
# ARGUMENTS LIST:
# name = MACS_GIE_CTCF
# format = BAM
# ChIP-seq file = /galaxy-repl/main/files/012/526/dataset_12526748.dat
# control file = /galaxy-repl/main/files/012/526/dataset_12526910.dat
# effective genome size = 1.87e+09
# tag size = 36
# band width = 300
# model fold = 32
# pvalue cutoff = 1.00e-05
# Ranges for calculating regional lambda are : peak_region,1000,5000,10000
INFO @ Mon, 21 Sep 2015 23:36:23: #1 read treatment tags...
INFO @ Mon, 21 Sep 2015 23:36:23: #1 read treatment tags...
WARNING @ Mon, 21 Sep 2015 23:36:26: NO records for chromosome chr13, minus str
WARNING @ Mon, 21 Sep 2015 23:36:26: NO records for chromosome chr10, minus str
WARNING @ Mon, 21 Sep 2015 23:36:26: NO records for chromosome chr18, minus str
INFO @ Mon, 21 Sep 2015 23:36:26: #1.2 read input tags...
WARNING @ Mon, 21 Sep 2015 23:36:28: NO records for chromosome chr16, minus str
WARNING @ Mon, 21 Sep 2015 23:36:28: NO records for chromosome chr15, plus str
INFO @ Mon, 21 Sep 2015 23:36:28: #1 Background Redundant rate: 0.01
INFO @ Mon, 21 Sep 2015 23:36:28: #1 finished!
INFO @ Mon, 21 Sep 2015 23:36:28: #2 Build Peak Model...
INFO @ Mon, 21 Sep 2015 23:36:34: #2 number of paired peaks: 9733
INFO @ Mon, 21 Sep 2015 23:36:34: #2 finished!
    
```

24: MACS on data 22 and data 18 (html report)

322.6 MB

format: html, database: mm9

HTML file

23: MACS on data 22 and data 18 (peaks: bed)

405 regions, comments

format: bed, database: mm9

1.Chrom	2.Start	3.End	4.Name	5
chr19	3284482	3284776	MACS_peak_1	92.61
chr19	3291823	3292396	MACS_peak_2	132.00
chr19	3450651	3452121	MACS_peak_3	64.64
chr19	3587686	3588189	MACS_peak_4	103.41
chr19	3623513	3624226	MACS_peak_5	71.88

Figure 22: MACS peak calling results.

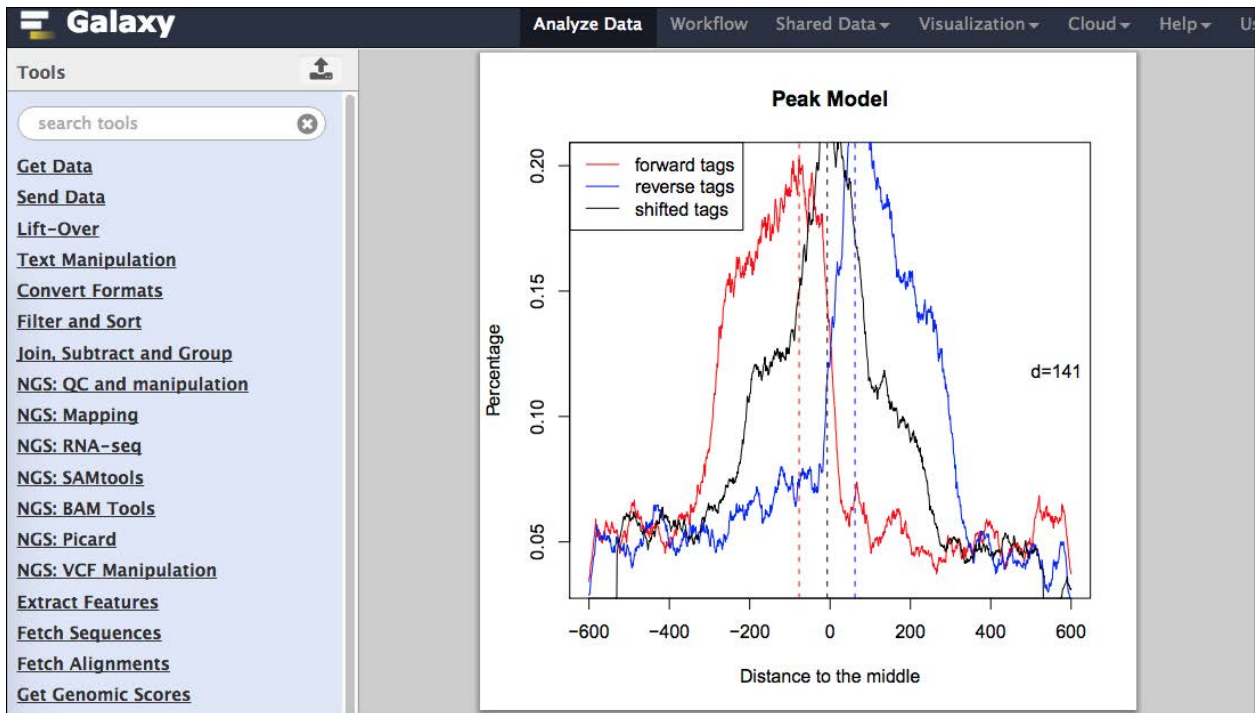


Figure 23: MACS peak model.

	A	B	C	D	E	F	G	H	I	
1	# This file is generated by MACS									
2	# ARGUMENTS LIST:									
3	# name = MACS_G1E_CTCF									
4	# format = BAM									
5	# ChIP-seq file = /galaxy-repl/main/files/012/526/dataset_12526748.dat									
6	# control file = /galaxy-repl/main/files/012/526/dataset_12526910.dat									
7	# effective genome size = 1.87e+09									
8	# tag size = 36									
9	# band width = 300									
10	# model fold = 32									
11	# pvalue cutoff = 1.00e-05									
12	# Ranges for calculating regional lambda are : peak_region,1000,5000,10000									
13	# unique tags in treatment: 213711									
14	# total tags in treatment: 214878									
15	# unique tags in control: 210164									
16	# total tags in control: 210781									
17	# d = 141									
18	chr	start	end	length	summit	tags	-10*LOG10(pvalue)	fold_enrichment	FDR(%)	
19	chr19	3204403	3204776	374	266	12	92.61	17.73	0	
20	chr19	3291824	3292396	573	337	35	132	10.3	0	
21	chr19	3450652	3452121	1470	765	44	64.64	10.69	0.34	
22	chr19	3587687	3588189	503	257	29	103.41	12.95	0	
23	chr19	3623514	3624226	713	381	29	71.88	11.44	0	
24	chr19	3723759	3725525	1767	1270	40	53.23	12.07	1.32	
25	chr19	3946485	3947973	1489	438	33	51.4	14.51	2.27	
26	chr19	3980149	3981146	998	393	49	156.53	12.41	0	
27	chr19	4012707	4013391	685	242	28	95.26	11.32	0	
28	chr19	4047685	4048485	801	365	22	56.6	6.3	0.84	
29	chr19	4098871	4099526	656	340	30	120	16.92	0	

Figure 24: MACS peak details.

Note!

TODO:

1. Do the MACS peak calling for both cell-lines. Look at both peak models and note the distance and differences between the models.
2. Rename the peak-files to something meaningful and while you are doing it change the **score**-column to **5**.
3. What do you expect in terms of called peaks if you would run G1E-CTCF without a control (the "input"-file)?
4. RUN G1E-CTCF without the input control. Note the differences.

2.0 Post-processing

Now that we established the peaks, we can do several different analyses to gain information about the genes they regulate or differences in peak abundance as well as functional association.

2.1 Overlap peaks with promoter regions

2.1.1 Get genes

Let's upload some genes and extract promoter information for them. Please download the following file ([mm9_chr19_NCBIgenes.bed](#) or from http://sschmeier.github.io/bioinf-workshop/galaxy-chipseq/data/mm9_chr19_NCBIgenes.bed) and upload to your Galaxy history (see *Figure 25*). the file contains 1428 gene regions in bed-format.

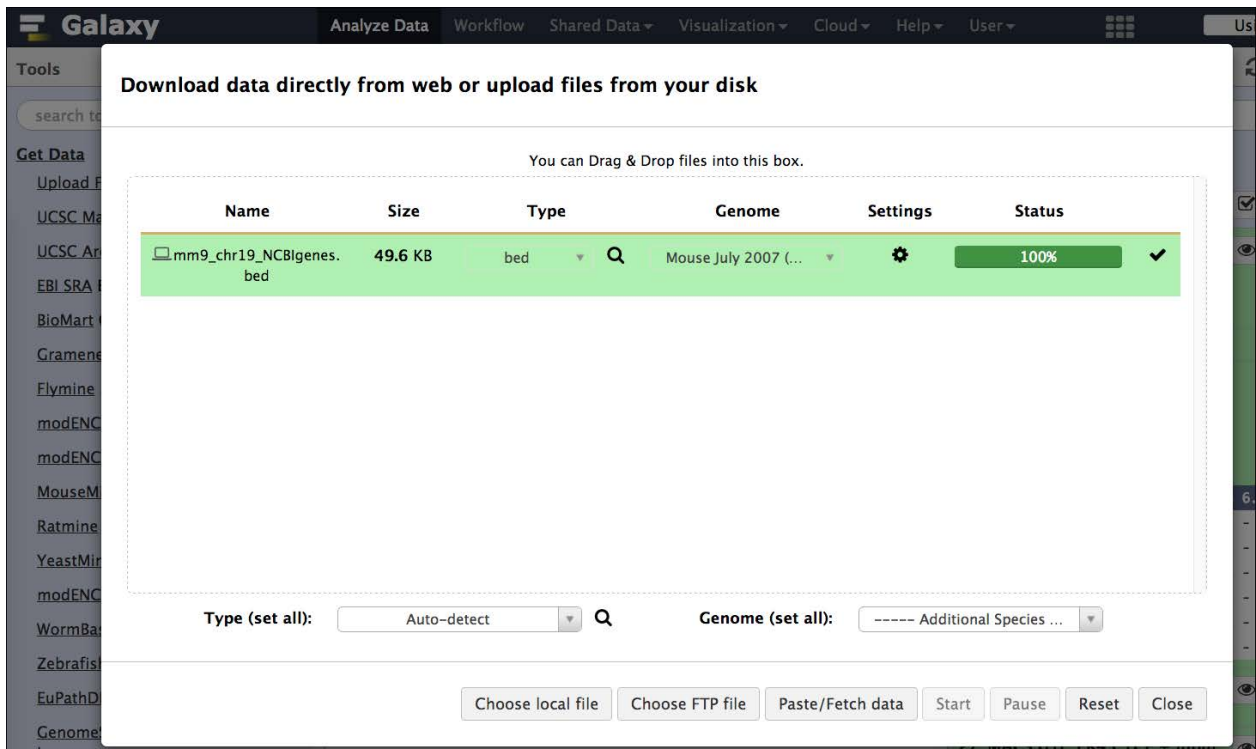


Figure 25: Upload the gene bed-file.

The screenshot shows the Galaxy interface with the main panel displaying a table of genomic intervals. The table has columns labeled 1 through 6. The right panel shows the file details for "51: mm9_chr19_NCBIgenes.bed", including the number of regions (1,428) and the format (bed, database: mm9). Below the details, there is a table with columns: 1.Chrom, 2.Start, 3.End, 4.Name, 5, 6.Strand. The table contains 21 rows of genomic intervals.

1	2	3	4	5	6
chr19	3065710	3197714	AK077035	1	-
chr19	3153210	3197714	AK006563	1	-
chr19	3153798	3197714	AK007025	1	-
chr19	3259075	3283010	Ighmbp2	1	-
chr19	3264810	3283010	Ighmbp2	1	-
chr19	3272720	3283010	Ighmbp2	1	-
chr19	3283046	3291197	Mrpl21	1	+
chr19	3283046	3292837	Mrpl21	1	+
chr19	3323300	3385733	Cpt1a	1	+
chr19	3388868	3398168	Mt15	1	+
chr19	3388868	3407785	Mt15	1	+
chr19	3389400	3407785	Mt15	1	+
chr19	3409916	3414457	Gal	1	-
chr19	3454927	3575749	Ppp6r3	1	-
chr19	3477775	3575749	Ppp6r3	1	-
chr19	3483527	3494038	mKIAA1558	1	-
chr19	3483527	3575749	Ppp6r3	1	-
chr19	3510945	3575749	mKIAA1558	1	-
chr19	3584824	3615879	Lrp5	1	-
chr19	3584824	3686564	Lrp5	1	-
chr19	3621680	3686564	Lrp5	1	-
chr19	3689686	3708168	AK144662	1	-

Figure 26: The file is in bed-format.

2.1.2 Get promoter

Get the promoter regions by using **Operate on Genomic Intervals** => **Get flanks**. Choose the upstream regions and 10,000 bases (see Figure 27). Rename the promoter-set to something meaningful.

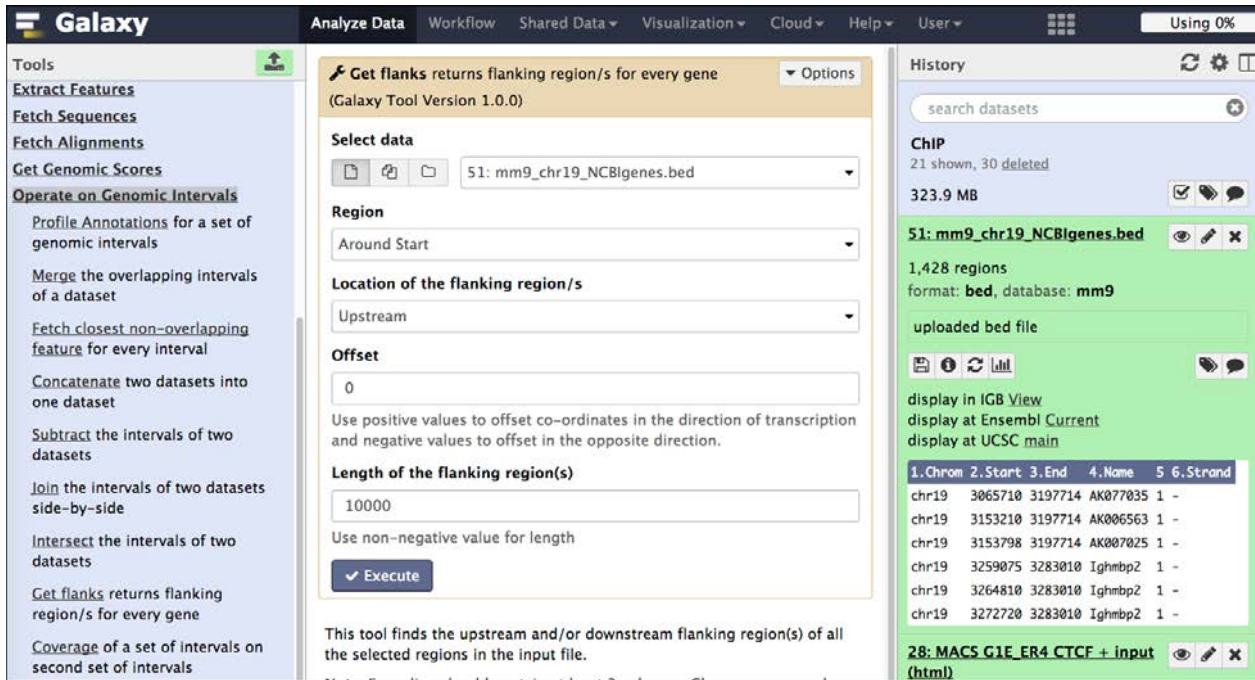


Figure 27: Get upstream flanking regions of the TSS of genes.

2.1.3 Join

Now we are going to join (overlap) the peaks with the promoter regions by choosing the tool: **Operate on Genomic Intervals** => **Join** (see Figure 28). Again rename the resulting dataset to something useful.

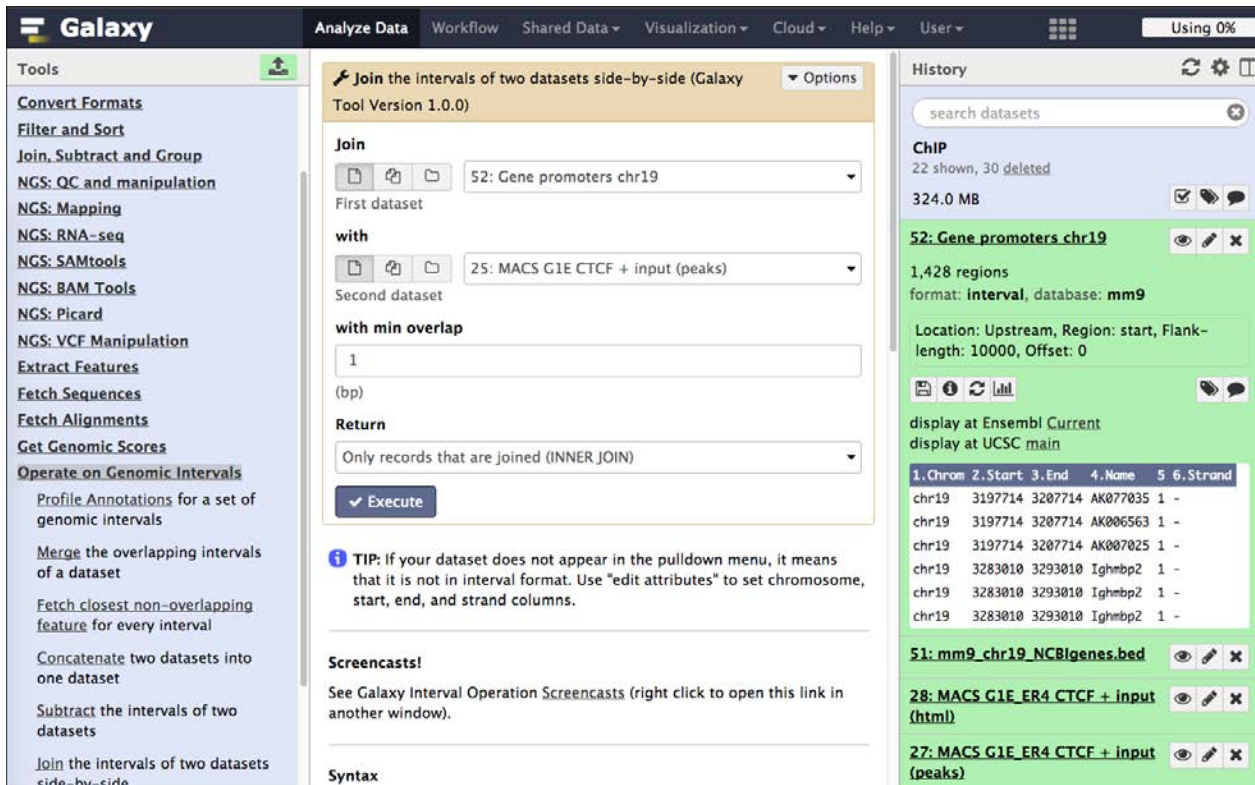


Figure 28: Overlap promoter and peaks with the join tool.

Note!

TODO: Join the peak file for G1E CTCF and G1E_ER4 CTCF with the gene promoter regions. Note the numbers and differences in promoter numbers that overlap Ctfc peaks for both peak-files.

2.2 Enrichment analysis (genes) with Enrichr

Now lets take the genes with Ctfc in their promoter regions and do some functional annotation. To do this, we need the unique genes from the overlap of peaks and promoters from the step before. We will be using the tool: **Join, Subtract and Group => Group** to do this. **Group** aggregates data in a certain column. We will use it to aggregate column 4, the gene symbol column (see *Figure 29*). Copy the resulting genes symbol (see *Figure 30*).

The screenshot shows the Galaxy interface with the 'Group data by a column and perform aggregate operation' tool selected. The tool configuration is as follows:

- Select data:** 53: Overlap promoter + G1E CTCF
- Group by column:** Column: 4
- Ignore case while grouping?:** No
- Ignore lines beginning with these characters:**
 - >
 - @
 - +
 - <
 - *
 -
 - =
 - |
 - ?
 - \$
 - .
 - :
 - &
 - %
 - ^
 - #
- Operation:** + Insert Operation
- Execute:** [Execute]

The History panel on the right shows a list of datasets:

- ChIP (23 shown, 31 deleted, 324.0 MB)
- 53: Overlap promoter + G1E CTCF (228 regions, format: interval, database: mm9)
- 52: Gene promoters chr19
- 51: mm9_chr19_NCBIGenes.bed
- 28: MACS G1E_ER4 CTCF + input (html)
- 27: MACS G1E_ER4 CTCF + input (peaks)
- 26: MACS on data 22 and data 18 (html report)
- 25: MACS G1E CTCF + input (peaks)
- 22: G1E input bowtie2

Figure 29: Aggregate the gene symbol column.

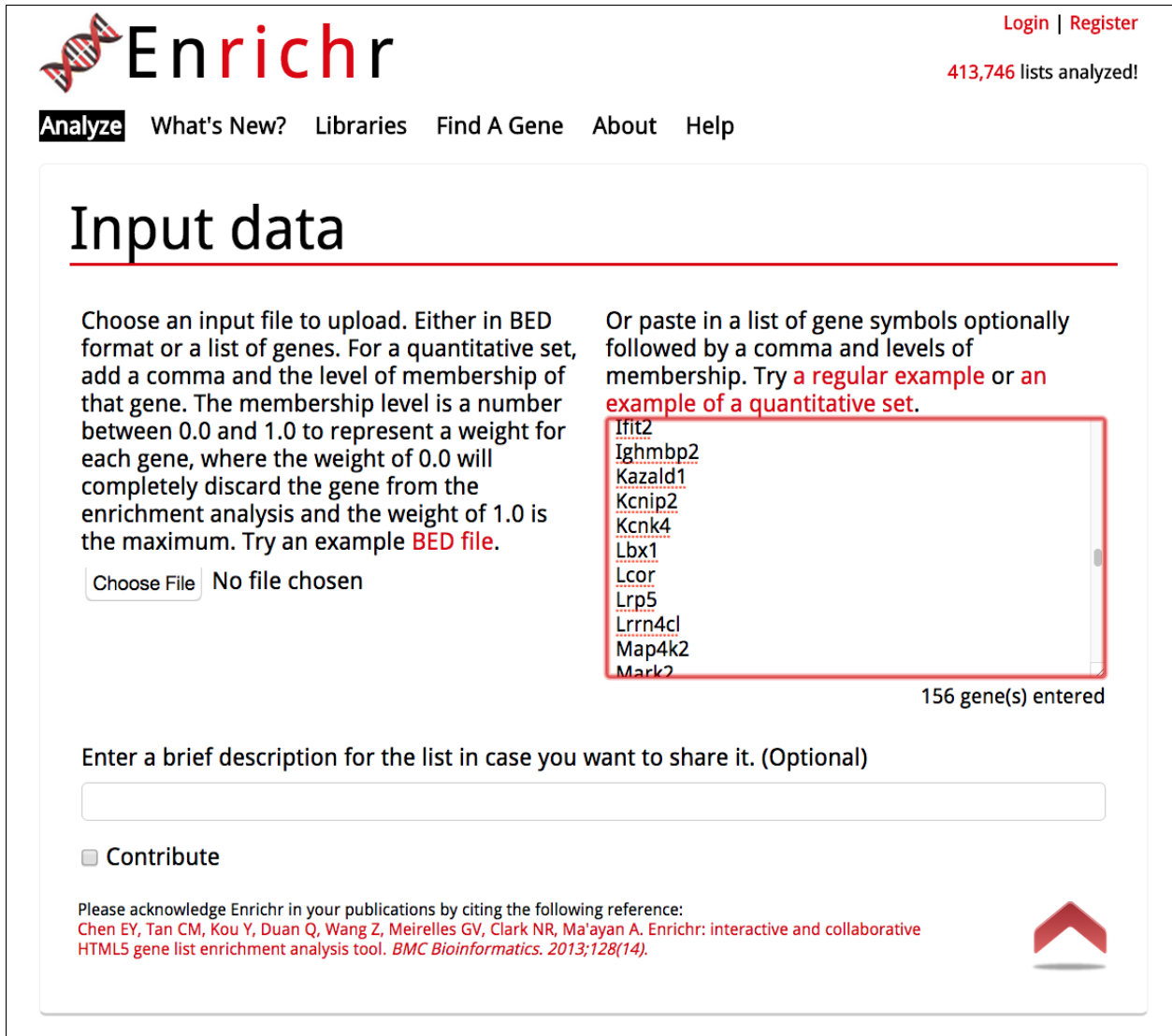
The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. The 'Using 0%' indicator is visible in the top right. The main workspace is divided into three panels:

- Tools:** A sidebar on the left containing various tool categories such as 'Get Data', 'Send Data', 'Text Manipulation', 'Convert Formats', 'Filter and Sort', 'Join, Subtract and Group', 'NGS: QC and manipulation', 'NGS: Mapping', and 'NGS: RNA-seq'.
- Workflow:** A central panel showing a workflow step labeled '1'. Below the step name, a list of gene symbols is displayed: 1810006K21Rik, 4930579J09Rik, AK006563, AK007025, AK008826, AK016444, AK028012, AK029443, AK036616, AK040197, AK040231, AK048429, AK049068, AK052572, AK052812, AK076978, AK077035, AK082813, AK148054, AK158434, AK168860, and AK170722.
- History:** A panel on the right showing a search for datasets. It displays a dataset named 'ChIP' with 24 shown and 31 deleted items, totaling 324.0 MB. Below this, a dataset named '55: Group on data 53' is highlighted in green, showing 156 lines in tabular format from the mm9 database. The content of this dataset is:


```
1
1810006K21Rik
4930579J09Rik
AK006563
AK007025
AK008826
AK016444
```

Figure 30: The aggregated gene symbols.

Now, go to the online tool [Enrichr](http://amp.pharm.mssm.edu/Enrichr/) (<http://amp.pharm.mssm.edu/Enrichr/>). [Enrichr](http://amp.pharm.mssm.edu/Enrichr/) provides a way to analyse mammalian gene lists to find enriched annotation terms to get a better understanding of the functions of the gene list under investigation. Go to **Analyze** tab and paste your gene list into the field (see [Figure 31](#)). Click on the arrow.



Enrichr Login | Register
413,746 lists analyzed!

Analyze What's New? Libraries Find A Gene About Help

Input data

Choose an input file to upload. Either in BED format or a list of genes. For a quantitative set, add a comma and the level of membership of that gene. The membership level is a number between 0.0 and 1.0 to represent a weight for each gene, where the weight of 0.0 will completely discard the gene from the enrichment analysis and the weight of 1.0 is the maximum. Try an example [BED file](#).

Or paste in a list of gene symbols optionally followed by a comma and levels of membership. Try [a regular example](#) or an [example of a quantitative set](#).

Choose File No file chosen

```
Ifit2
Ighmbp2
Kazald1
Kcnp2
Kcnk4
Lbx1
Lcor
Lrp5
Lrrn4cl
Map4k2
Mark2
```

156 gene(s) entered


Enter a brief description for the list in case you want to share it. (Optional)

Contribute

Please acknowledge Enrichr in your publications by citing the following reference:
Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;128(14).

Figure 31: The Enrichr tool.

On the result pages (see Figure 32) you will find several different categories (e.g. *Transcription*, *Pathways*, etc.) of with different databases where term-gene association information was extracted. Figure 32 for example shows the enriched pathways from the [Reactome](http://www.reactome.org/) (<http://www.reactome.org/>) database.


Login | Register

Transcription
Pathways
Ontologies
Disease/Drugs
Cell Types
Misc
Legacy
Crowd

Description No description available (156 genes)

KEGG 2015

WikiPathways 2015

Reactome 2015

Bar Graph
Table
Grid
Network
⚙️

Hover each row to see the overlapping genes.

10 entries per page
Search:

Index	Name	P-value	Z-score	Combined Score
1	Signaling by Wnt	0.03101	-2.24	1.74
2	Organelle biogenesis and maintenance	0.004934	-2.20	1.71
3	misspliced LRP5 mutants have enhanced beta-catenin-dependent signaling	0.02128	-2.19	1.70
4	RNF mutants show enhanced WNT signaling and proliferation	0.02128	-2.16	1.68
5	XAV939 inhibits tankyrase, stabilizing AXIN	0.02128	-2.16	1.68
6	TCF dependent signaling in response to WNT	0.02128	-2.16	1.68
7	Signaling by WNT in cancer	0.03189	-2.13	1.65
8	Polymerase switching on the C-strand of the telomere*	0.007373	-2.07	1.61
9	Telomere C-strand (Lagging Strand) Synthesis*	0.01758	-2.07	1.61
10	Lagging Strand Synthesis*	0.01499	-2.03	1.57

Showing 1 to 10 of 415 entries | [Export entries to table](#)
◀ Previous Next ▶

Terms marked with an * have an overlap of less than 5

Figure 32: The Enrichr results show enriched term associations to the input gene list.

Note!

TODO:

1. Find and note the top 5 enriched [Gene Ontology](#) process terms for both the G1E and G1E_ER4 genes that have Ctf in their promoters.
2. Now that you have unique gene lists for G1E and G1E_ER4, how many genes are in common, e.g. which genes in both cases have Ctf in their promoter region?

Hint! For point 2. you can use the **Join, Subtract and Group => Compare two Datasets** tool.

2.3 Enrichment analysis (peaks) with GREAT

Here we are going to use another tool called **GREAT** (<http://bejerano.stanford.edu/great/public/html/>). Great as opposed to **Enrichr** expects bed-regions directly, thus we do not need to get the genes that overlap our peak regions. Take the results from MACS, cut out the first 4 columns with **Text Manipulation** => **Cut** (as GREAT does not expect floats as scores and will produce errors), copy the regions and paste them into the **GREAT** interface.

The screenshot shows the GREAT website interface. At the top, there is a navigation bar with links: Overview, News, Use GREAT, Demo, Video, How to Cite, Help, and Forum. Below this is a dropdown menu for the GREAT version, currently set to 'GREAT version 3.0.0 current (02/15/2015 to now)'. A prominent banner reads 'GREAT predicts functions of cis-regulatory regions.' Below the banner is a paragraph explaining that GREAT assigns biological meaning to non-coding genomic regions by analyzing nearby gene annotations. A 'News' section follows, listing updates from February 2015 to April 2012. The 'Species Assembly' section offers radio buttons for Human (GRCh37), Mouse (NCBI build 37 and 38), and Zebrafish (Wellcome Trust Zv9), with a link to 'Zebrafish CNE set' and a question 'Can I use a different species or assembly?'. The 'Test regions' section has radio buttons for 'BED file' (with a file selection button showing 'Galaxy25-...bed).bed') and 'BED data' (with a text area containing a table of coordinates and peak names). Below the table are two green links: 'What should my test regions file contain?' and 'How can I create a test set from a UCSC Genome Browser annotation track?'. The 'Background regions' section has radio buttons for 'Whole genome', 'BED file' (with a 'No file chosen' button), and 'BED data' (with an empty text area). Below this are two green links: 'When should I use a background set?' and 'What should my background regions file contain?'. The 'Association rule settings' section has a 'Show settings »' button. At the bottom, there are 'Submit' and 'Reset' buttons.

Species Assembly

- Human: GRCh37 (UCSC hg19, Feb/2009)
- Mouse: NCBI build 37 (UCSC mm9, Jul/2007)
- Mouse: NCBI build 38 (UCSC mm10, Dec/2011)
- Zebrafish: Wellcome Trust Zv9 (danRer7, Jul/2010) [Zebrafish CNE set](#)

Can I use a different species or assembly?

Test regions

- BED file: Galaxy25-...bed).bed
- BED data:

chr19	60968112	60968799	MACS_peak_401
chr19	61160078	61161098	MACS_peak_402
chr19	61173687	61174023	MACS_peak_403
chr19	61185950	61186399	MACS_peak_404
chr19	61275219	61276078	MACS_peak_405

What should my test regions file contain?
How can I create a test set from a UCSC Genome Browser annotation track?

Background regions

- Whole genome
- BED file: No file chosen
- BED data:

When should I use a background set?
What should my background regions file contain?

Association rule settings

Figure 33: The GREAT website.

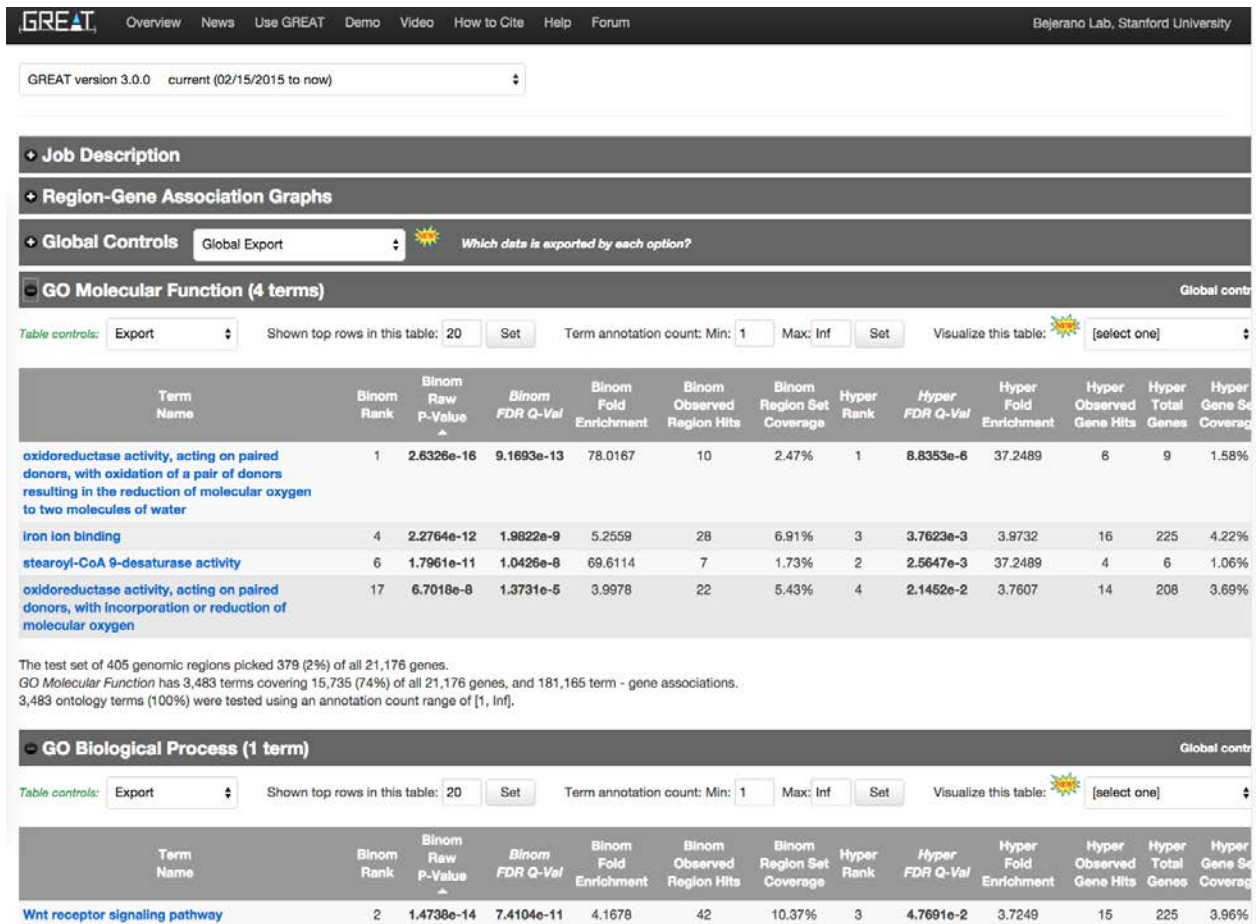


Figure 34: GREAT result page.

Note!

TODO: Run **GREAT** for both MACS result-files and note the top 5 **GO Biological processes**. Are they different to the ones from **Enrichr**?

2.4 Visualisation

Let us now create a visualisation track of the promoters that overlap G1E CTCF peaks and G1E_ER4 CTCF peaks. Use **Graph/Display Data => Build custom track** (see *Figure 33*). Also add the two MACS peak bed-files. Look at the track at UCSC (see *Figure 36* and *Figure 37*).

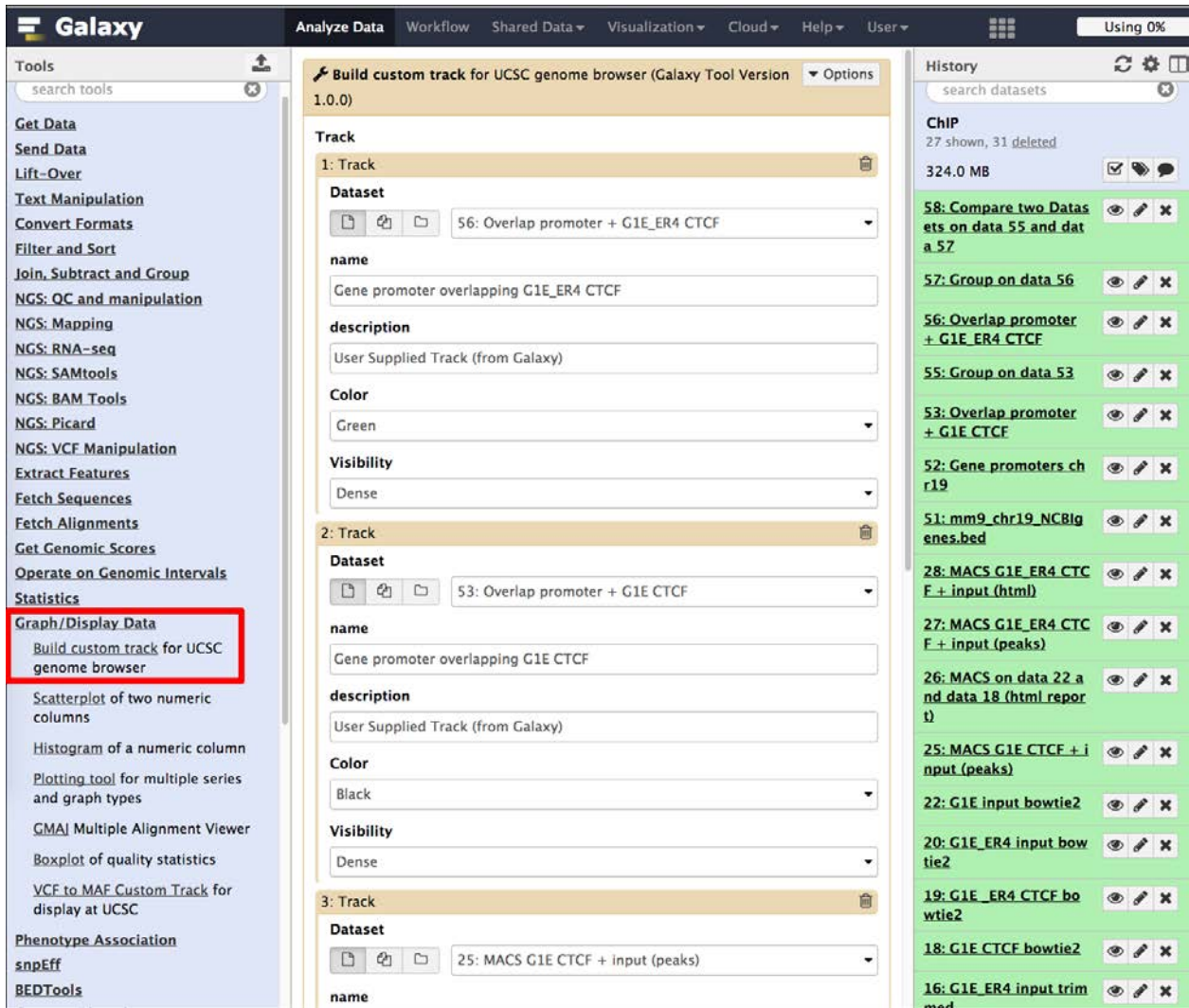


Figure 35: Building a custom UCSC track.

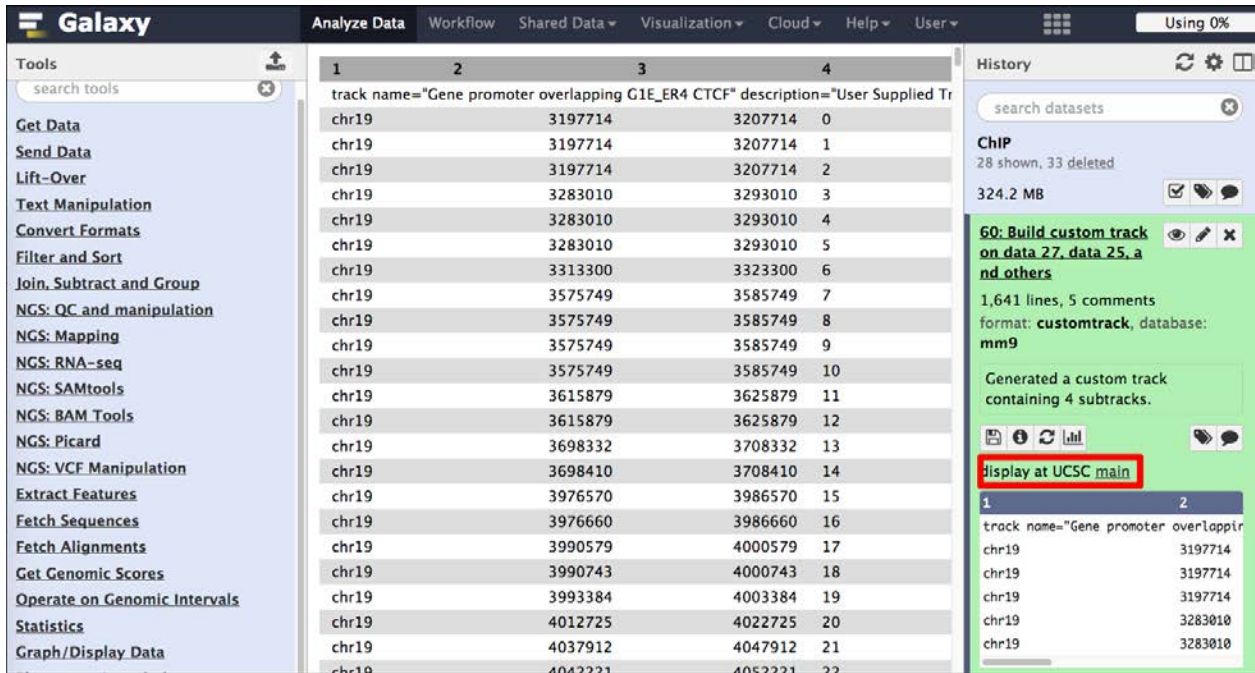


Figure 36: Visualising a Galaxy dataset/track.

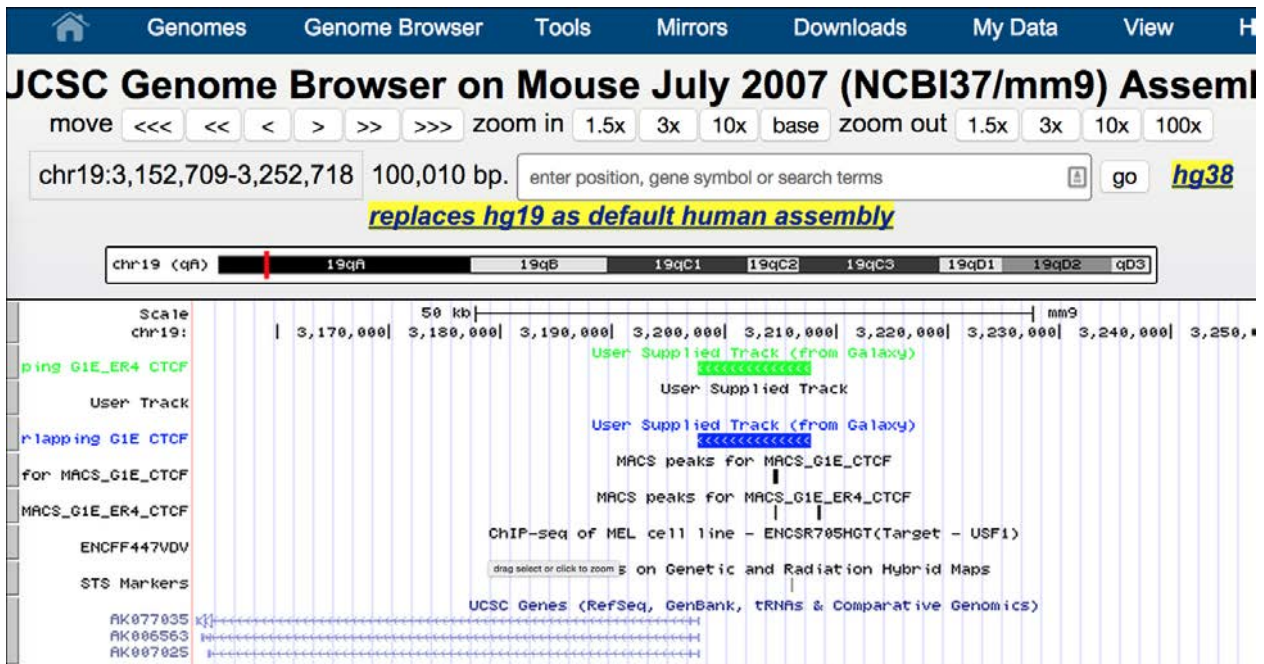


Figure 37: Custom UCSC track at the UCSC genome browser website.

2.5 Motif finding

Here we want to establish enriched sequence motifs in the peak regions to hypothesise on the actual binding site of Ctf. We are going to use MEME-ChIP (<http://meme.ebi.edu.au/meme/tools/meme-chip>) for this. However, MEME-ChIP expects fasta-sequence data as an input, not bed-files. So, we need to extract for our peak bed-files the actual sequence. Another restriction is, that MEME-ChIP expects regions of similar size, this is also not given in the MACS results. The workflow for this analysis looks like this:

1. Find the center of each MACS peak region.
2. Get the flanking region ± 250 bases (as the recommended region size for [MEME-ChIP](#) is 500bp).
3. Extract the fasta-sequence for the regions.
4. Download the fasta-file and upload to [MEME-ChIP](#).
5. Run [MEME-ChIP](#).

2.5.1 Find the peak center

We have the start and stop position in our bed-files, thus we can calculate the center point with **start + ((stop-start)/2)**. The tool we need is: **Text Manipulation** => **Compute** (see [Figure 38](#)). Make sure you round the results.

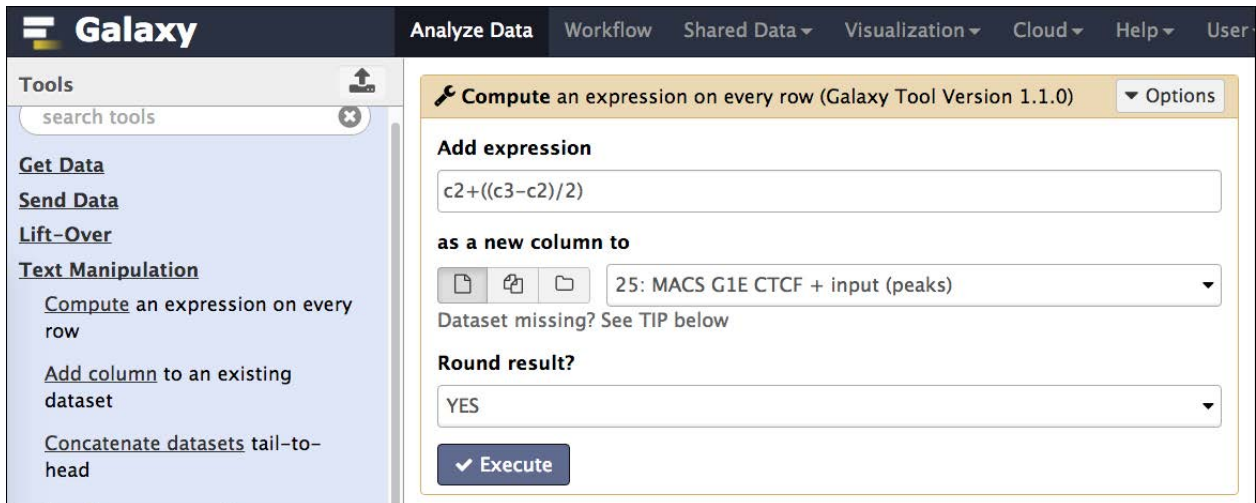


Figure 38: Calculate center peak position.

Now we cut out the first column (chromosome) and last column (center) with **Text Manipulation** => **Cut** (see [Figure 39](#)).

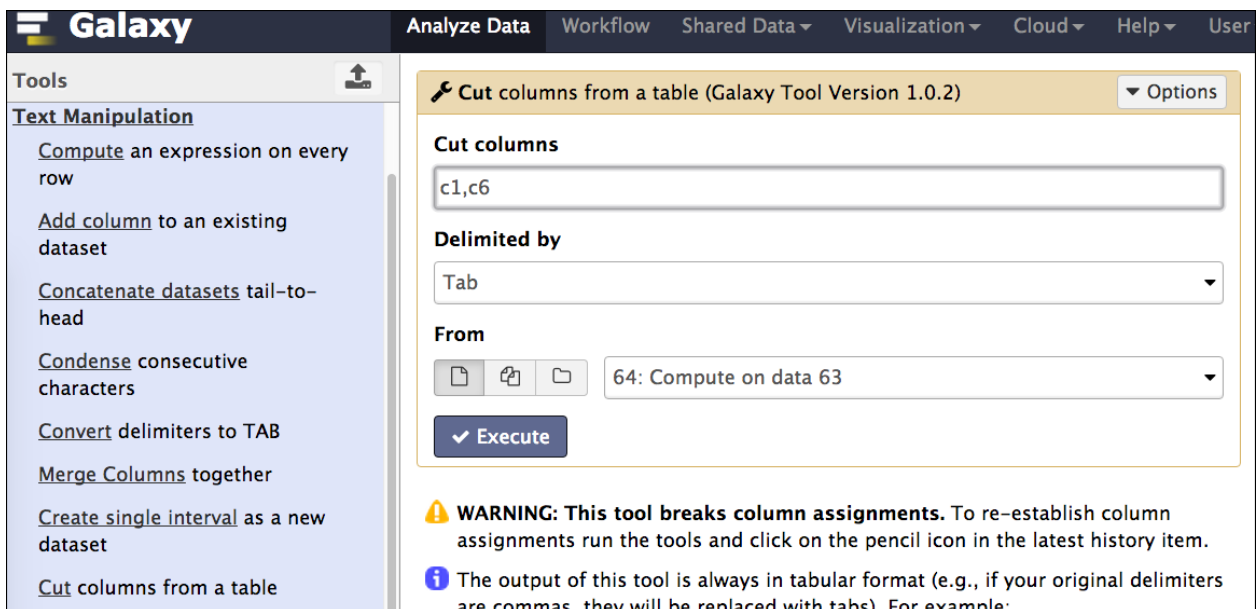


Figure 39: Cut columns.

We add another column to the result that will represent the stop-position with **Text Manipulation** => **Compute**. Make this **c2 + 1** (see [Figure](#)).

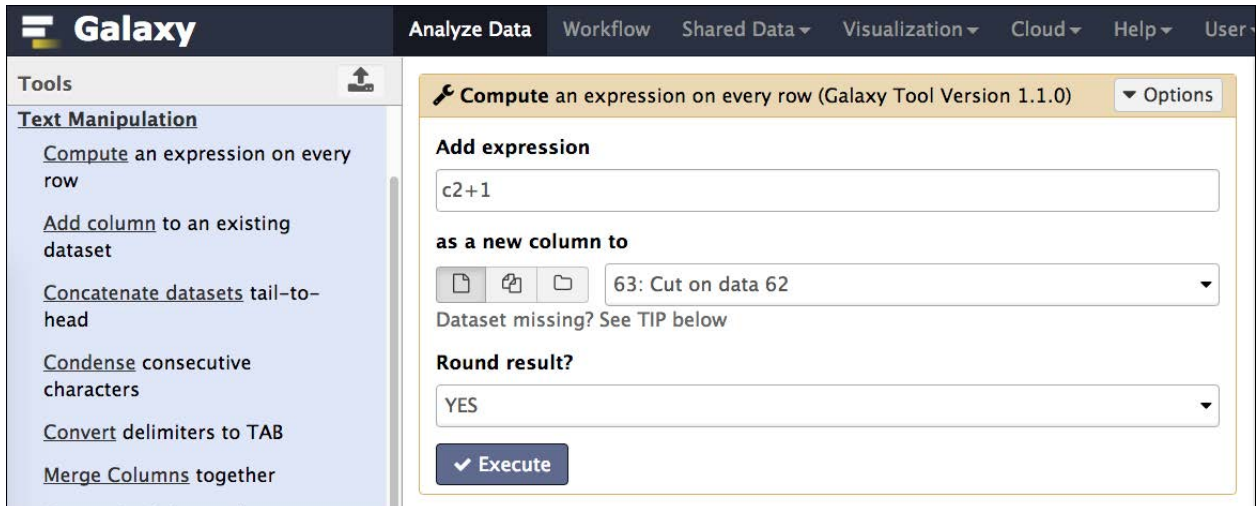


Figure 40: Compute the stop-position.

Now, the resulting data is in *tabular* format (see Figure 41) and we need to change it first to *bed*-format (see Figure 42).

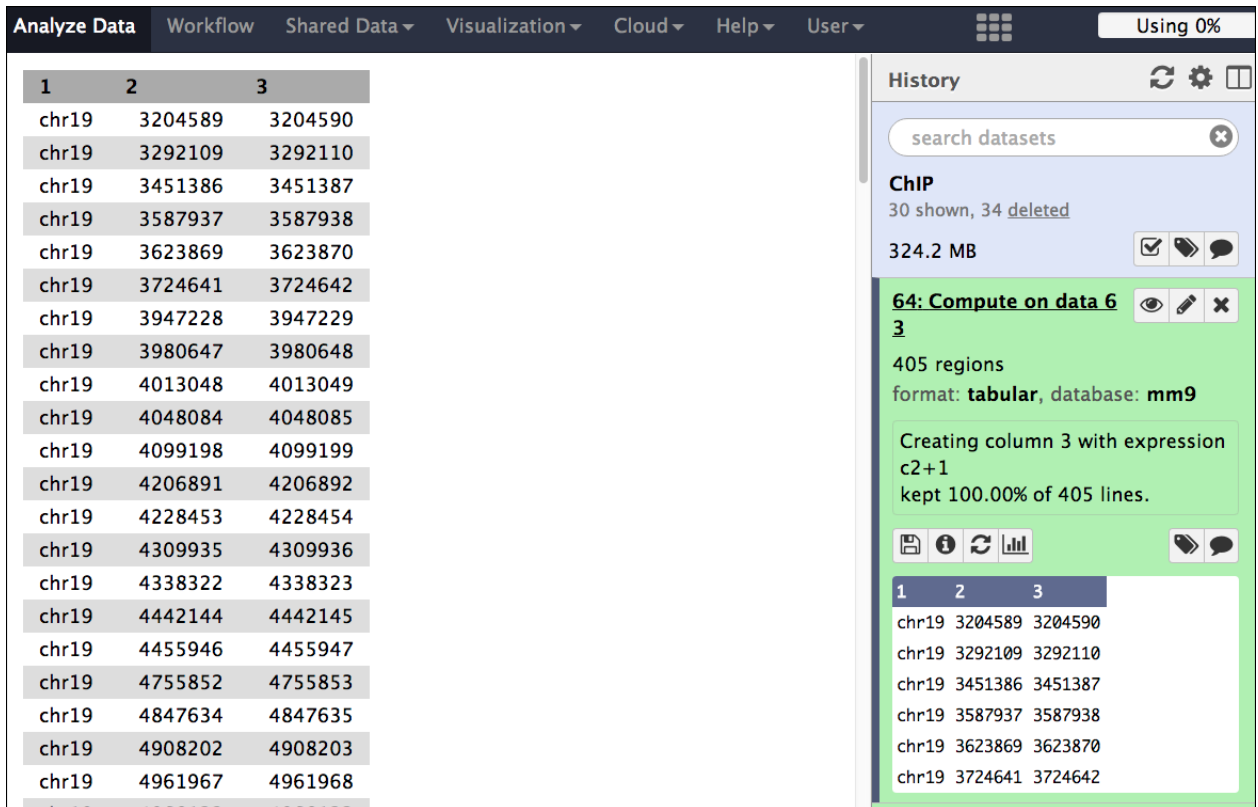


Figure 41: Center peaks.

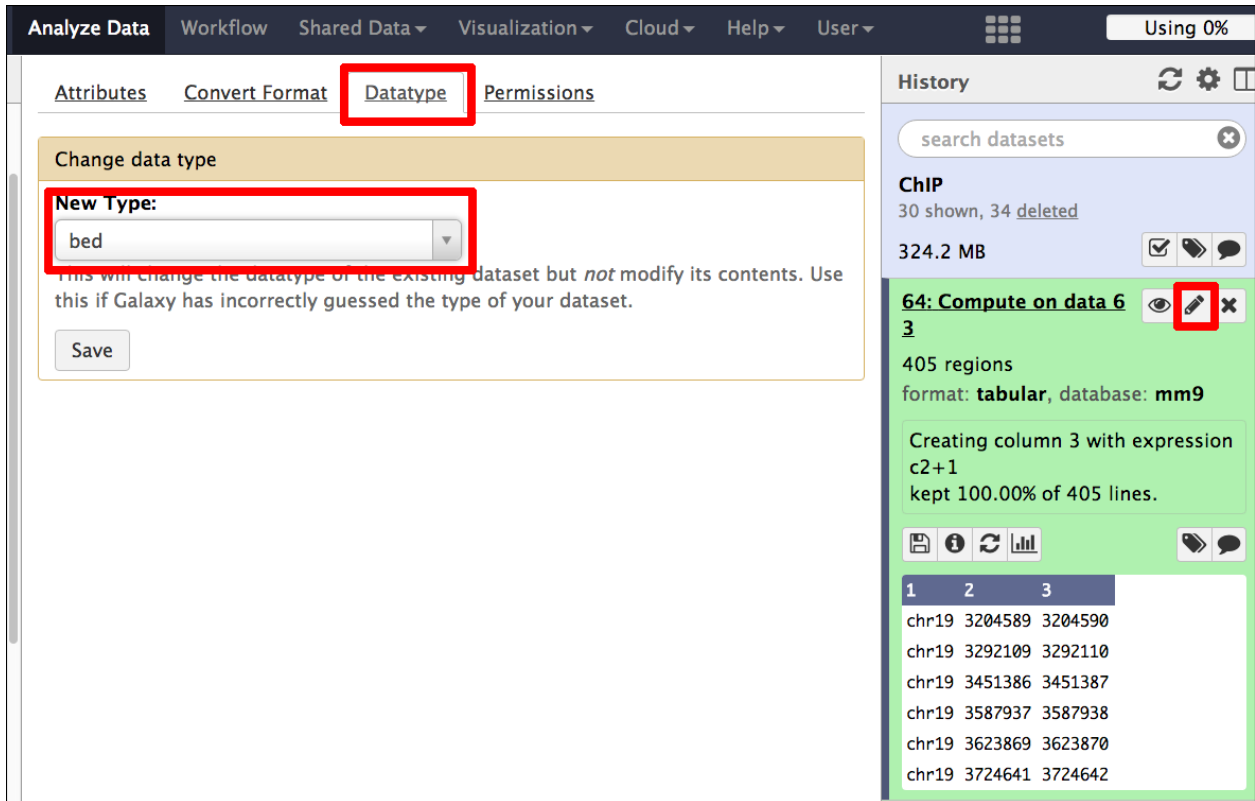


Figure 42: Change the data-format to bed.

2.5.2 Get flanking regions

Use **Operate on Genomic Intervals** => **Get flanks**. Extend **both** sides of the start position by **500** bases (see Figure 43).

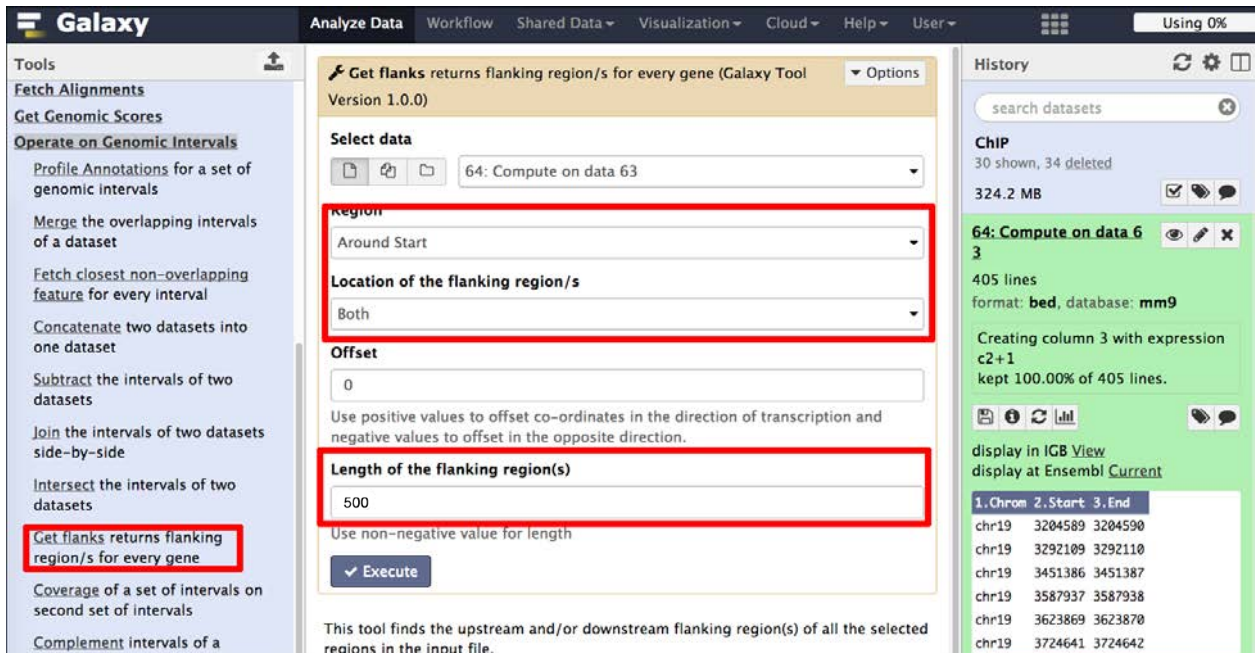


Figure 43: Get flanking regions.

2.5.3 Extract fasta-sequence

Use **Fetch Sequences** => **Extract Genomic DNA** to extract for the regions the genomic DNA (see Figure 44).

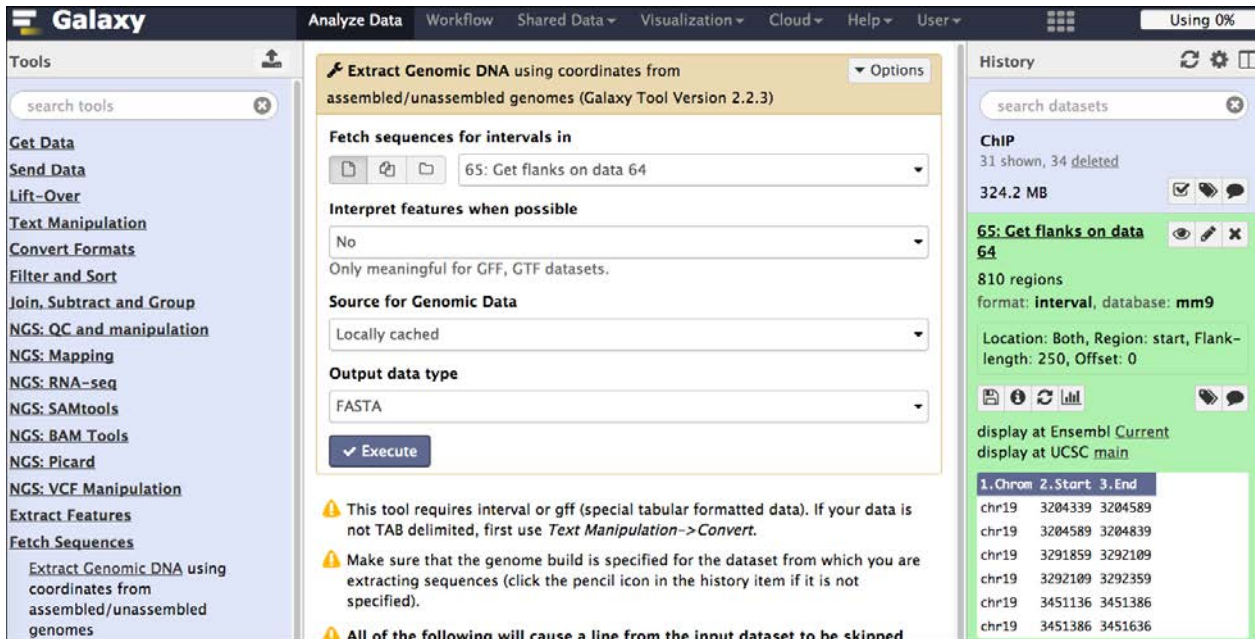


Figure 44: Extract DNA for regions.

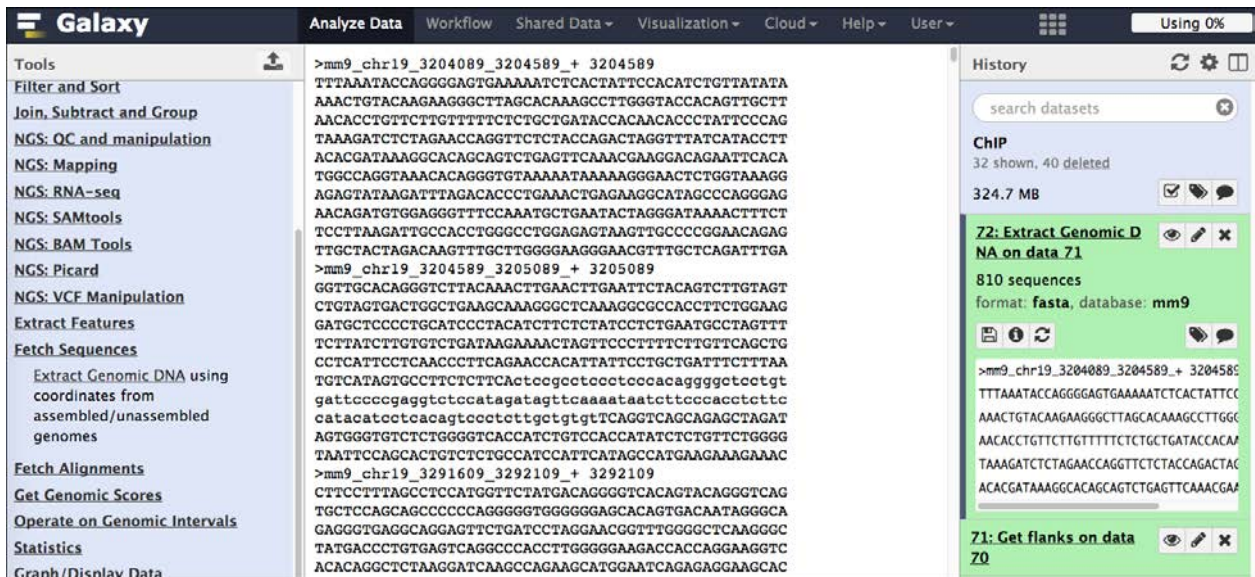


Figure 45: Region in fasta-format.

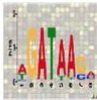
2.5.4 Run MEME-ChIP

Go to **MEME-ChIP** (<http://meme.ebi.edu.au/meme/tools/meme-chip>) and copy the fasta-sequences into the field and run the application (see Figure 45). This may result in enriched sequence motifs that were found in the uploaded sequences (see Figure 46).

MEME Suite 4.10.1

- ▶ Motif Discovery
- ▶ Motif Enrichment
- ▶ Motif Scanning
- ▶ Motif Comparison
- ▶ Manual
- ▶ Guides & Tutorials
- ▶ Sample Outputs
- ▶ File Format Reference
- ▶ Databases
- ▶ Download & Install
- ▶ Help
- ▶ Alternate Servers
- ▶ Authors & Citing
- ▶ Recent Jobs

↩ Previous version
4.10.0



MEME-ChIP

Motif Analysis of Large Nucleotide Datasets

Version 4.10.1

MEME-ChIP performs **comprehensive motif analysis** (including motif discovery) on LARGE (50MB maximum) sets of **nucleotide** sequences such as those identified by ChIP-seq or CLIP-seq experiments (sample output from sequences). See this Manual for more information.

Data Submission Form

Perform motif discovery, motif enrichment analysis and clustering on large nucleotide datasets.

Input the sequences
Enter the (equal-length) nucleotide sequences to be analyzed. ?

Type in sequences DNA Protein ?

Input the motifs
Select, upload or enter a set of known motifs. ?

?

Input job details
(Optional) Enter your email address. ?

(Optional) Enter a job description. ?

- ▶ Universal options
- ▶ MEME options
- ▶ DREME options
- ▶ CentriMo options

Note: if the combined form inputs exceed 80MB the job will be rejected.

Figure 45: MEME-ChIP interface.



Figure 46: MEME-ChIP results.

Note!

TODO: Note the enriched motif for the G1E CTCF and G1E_ER4 peak regions. Are there any differences?

2.6 References

Hawkins RD, Hon GC & Ren B. **Next-generation genomics: an integrative approach.** *Nature Reviews Genetics.* 2010; 11, 476-486

Park PJ. **ChIP-seq: advantages and challenges of a maturing technology.** *Nature Reviews Genetics.* 2009; 10, 669-680

2.7 Web links

Galaxy: <https://usegalaxy.org>

Enrichr: <http://amp.pharm.mssm.edu/Enrichr/>

GREAT: <http://bejerano.stanford.edu/great/public/html/>

Gene Ontology: <http://amigo.geneontology.org/>

MEME-ChIP: <http://meme.ebi.edu.au/meme/tools/meme-chip>

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