

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

2016-03-22

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_sschatmeier.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/#Igalaxy-intro/) (<http://sschmeier.github.io/bioinf-workshop/#Igalaxy-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](#)).
- Understand how to call ChIP-peaks based on the mapped reads ([1.9](#)).
- Understand how to gather additional information about your 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*):

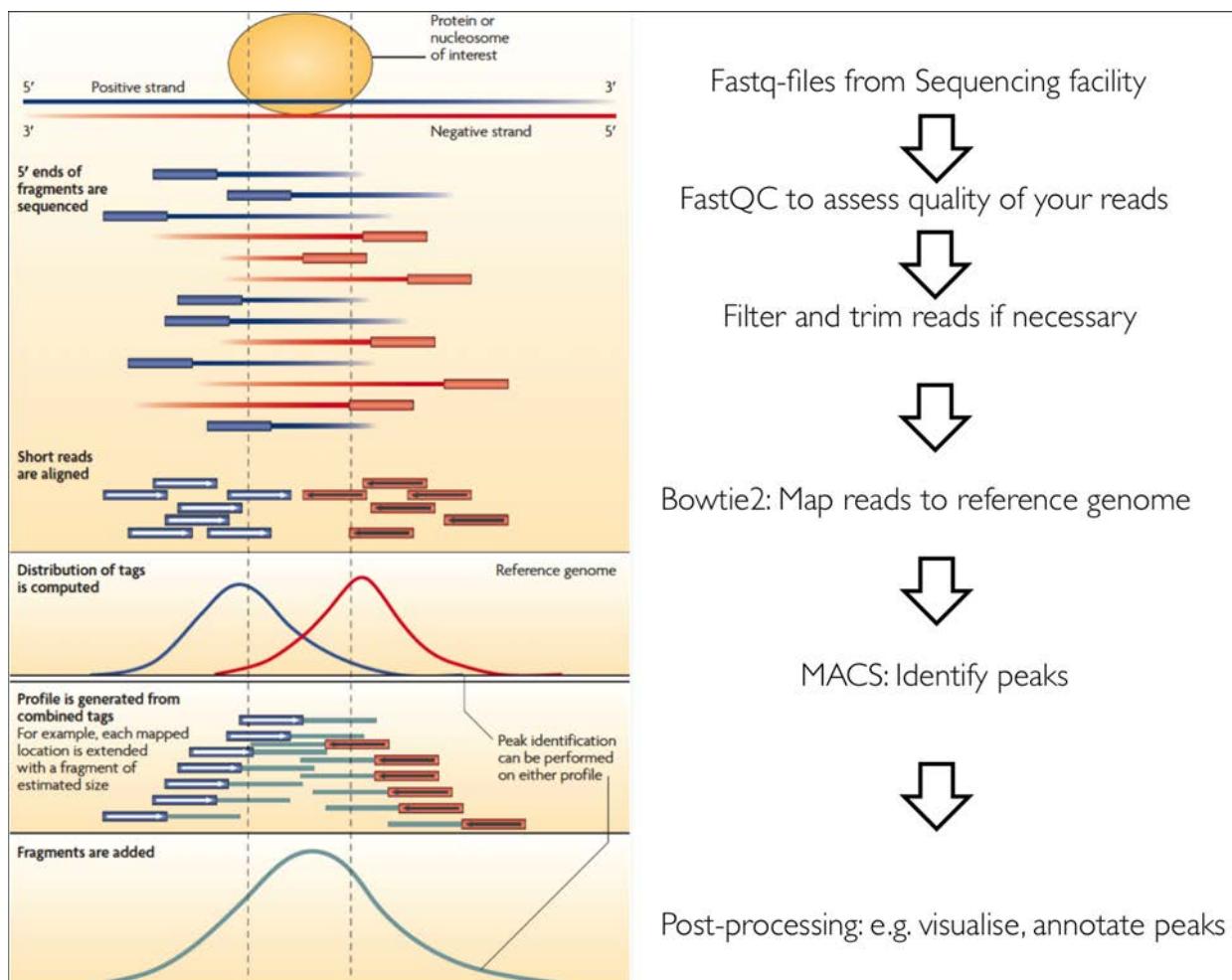


Figure 1: ChIP-workflow (adjusted from Park2009).

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).

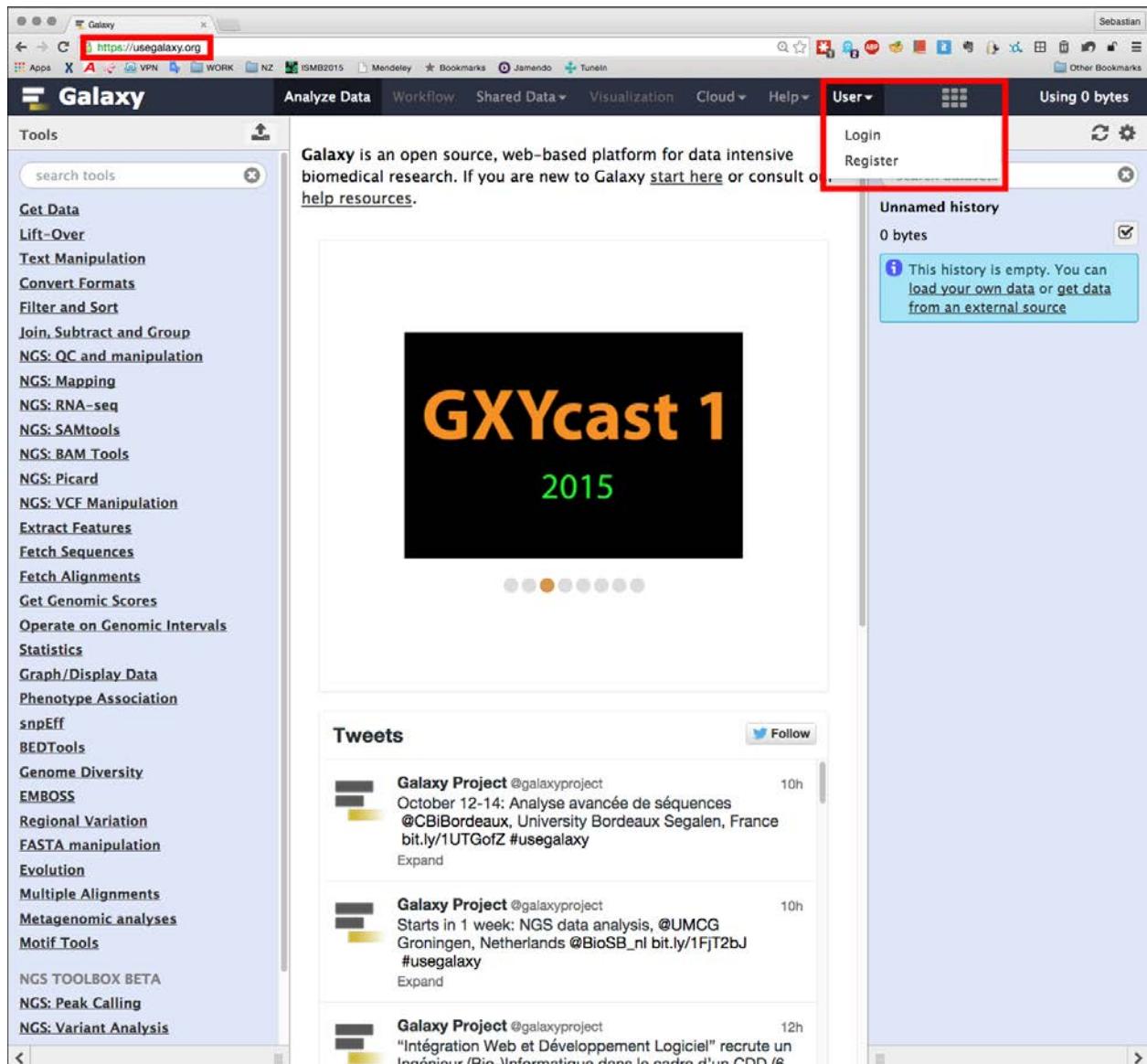


Figure 2: Log into your Galaxy account.

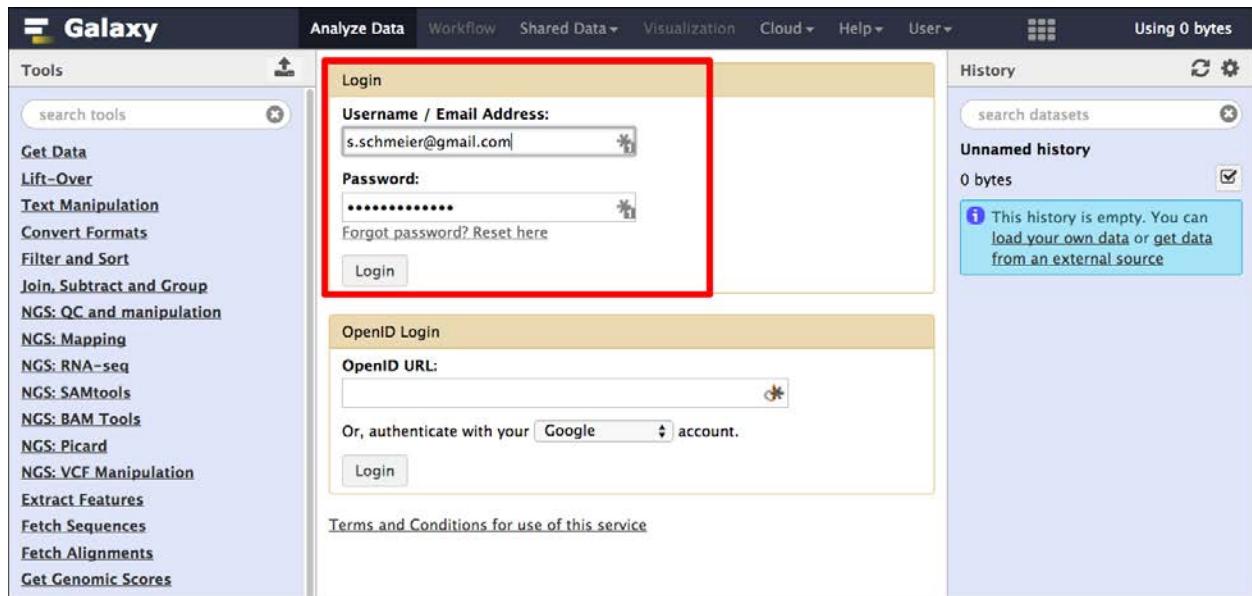


Figure 3: Log into your Galaxy account with your credentials.

1.4 Create a new history

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

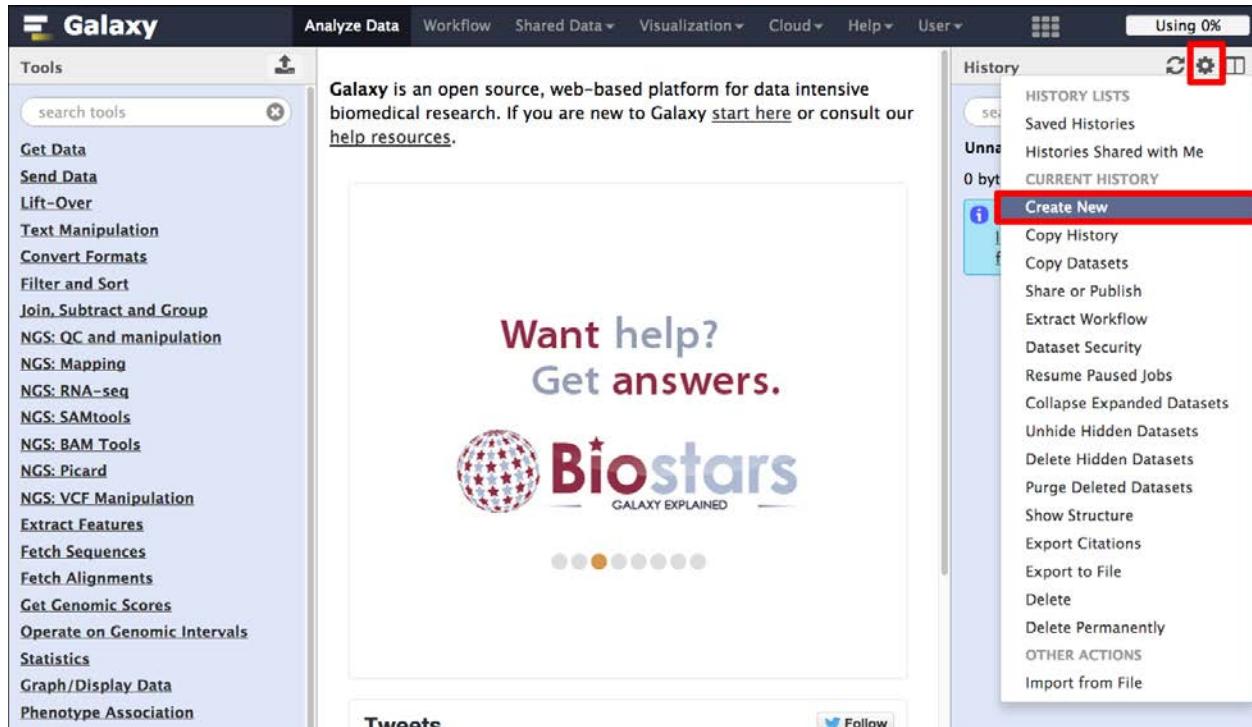


Figure 4: Log into your Galaxy account.

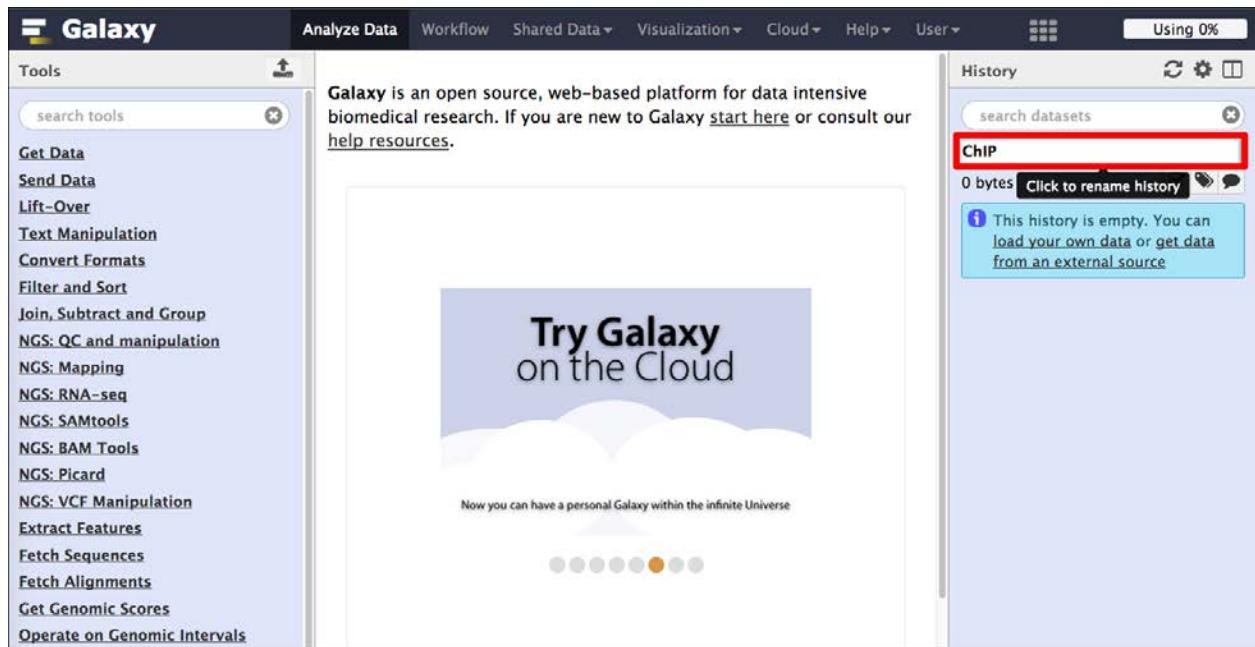


Figure 5: Log into your Galaxy account with your credentials.

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*)

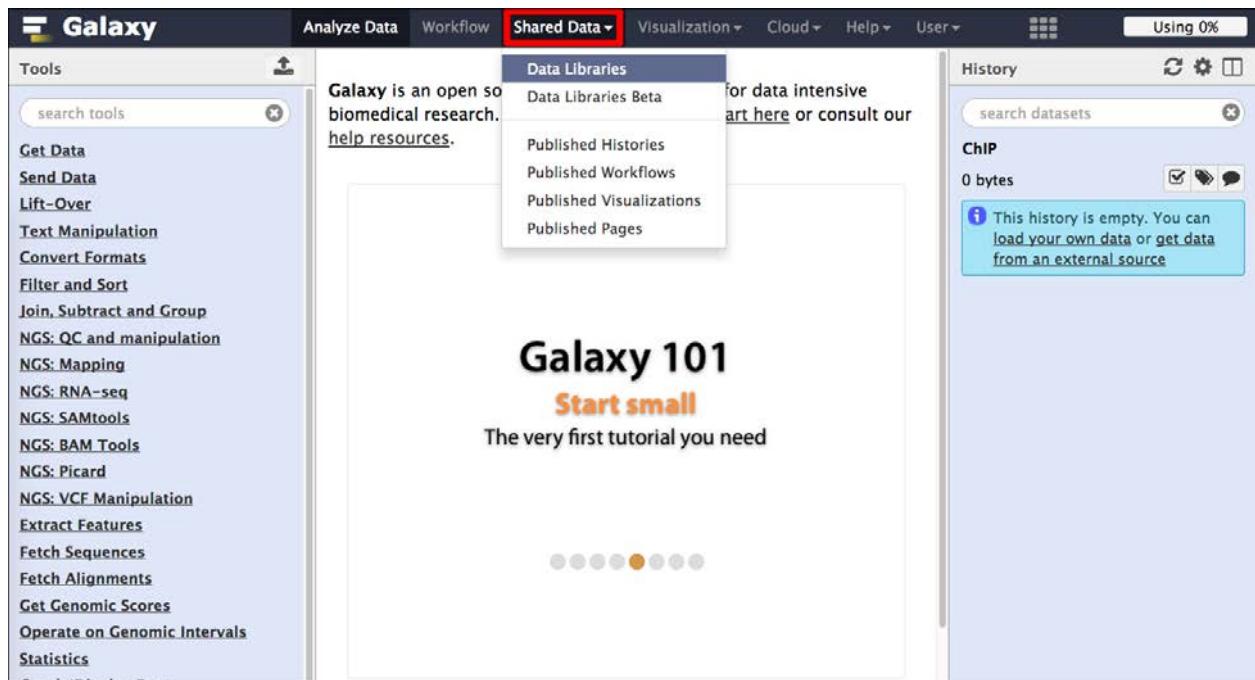


Figure 6: Load shared data tab.

The screenshot shows the Galaxy Data Libraries interface. At the top, there is a search bar and a navigation menu with links like 'Analyze Data', 'Workflow', 'Shared Data', etc. Below the search bar is a link to 'Advanced Search'. The main area displays a table of datasets:

Data library name	Data library description
1000 Genomes	Data from the 1000 Genomes Project FTP site
AC-exome	
Bushman	Data for two papers about the Khoisan and other populations.
Charts Example Data	
ChIP-Seq Mouse Example	Data used in examples that demonstrate analysis of ChIP-Seq data
Chobi	
CloudMap	Contains userguide, reference files, and configuration files for the Cloudmap WGS analysis pipeline
Codon Usage Frequencies	
Coleman	IonPGM
DataImport-00107dec-86f4-44f6-af87-cb20d69c2fe9@createprivate.libraryexample.com	
DataImport-ff2b1cbd-ded0-41b8-afdb-741f1534ceb8@createprivate.libraryexample.com	
Demonstration Datasets	Demonstration datasets collected from various Galaxy tutorials
Denisovan sequences	Files from 'A high-coverage genome sequence from an archaic Denisovan Individual' Meyer et al. Science 2012 and basic processed data.
Erythroid Epigenetic Landscape	Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration

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).

The screenshot shows the 'Data Library "Demonstration Datasets"' page. The table lists datasets under the 'Mouse ChIP-seq: G1E CTCF binding' category. Four specific datasets are selected and highlighted with a red box:

Name	Message	Data type	Date uploaded	File size
G1E CTCF (chr19)	Sample datasets from Hardison lab for ChIP-seq analysis	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

At the bottom, there is a button labeled 'For selected datasets: Import to current history' with a red box around it, and a 'Go' button next to it.

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

For selected datasets: Import to current history Go

Figure 9: Load the datasets.

Figure 10: Loaded data in history panel

Alternatively, you can download the 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/>)).

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

The FastQ format consists of four parts:

- Sequence id:** @EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
- Sequence:** GATTGGGTTCAAAGCAGTATCGATCAAATAGTAATCCATTGTTC
- Phred quality of the corresponding nucleotide (ASCII code):** +'*(((((***+))%%++)(%%%)).1***-+*'')**55CCF>>>
- Phred quality:**

• One ASCII character per nucleotide.

• Encodes for a quality $Q = -10 \log_{10}(P)$, 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 (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 Ctcf has been ChIP-ed and sequenced. The "input" samples denote samples where the DNA was fragmented but before the immunoprecipitation against Ctcf 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**.

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.

Figure 14: Filtering reads of bad quality.

Furthermore, I edited the dataset name (1) to keep track 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*).

Figure 15: Rename the dataset to keep track.

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 shows the Galaxy web interface. On the left, a sidebar lists various tools for sequence manipulation. The main area displays the 'Filter by quality' tool. The 'Library to filter' dropdown is set to '4: G1E input (chr19)'. The 'Quality cut-off' dropdown is set to '20'. The 'Percent of bases' dropdown is set to '90'. A red box highlights the '4: G1E input (chr19)' dropdown. Below these settings, a 'What it does' section explains that the tool filters reads based on quality scores. It includes two bullet points: one for a percent of 100 and another for a percent of 50. An example sequence is shown, followed by notes about filtering based on percent and cut-off values. To the right, the 'History' panel shows a workflow: '4: G1E input (chr19)' leads to '10: G1E input trimmed', which then leads to '9: G1E input filtered'. This is followed by '3: G1E_ER4 input (chr19)', '2: G1E_ER4 CTCF (chr19)', and finally '1: G1E CTCF (chr19)'. The '10: G1E input trimmed' dataset is highlighted in green and has a size of 149.0 MB and a format of fastqsanger, database mm9. Below the history, a sequence preview shows a portion of the FASTQ file.

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 off the ends on both sides.

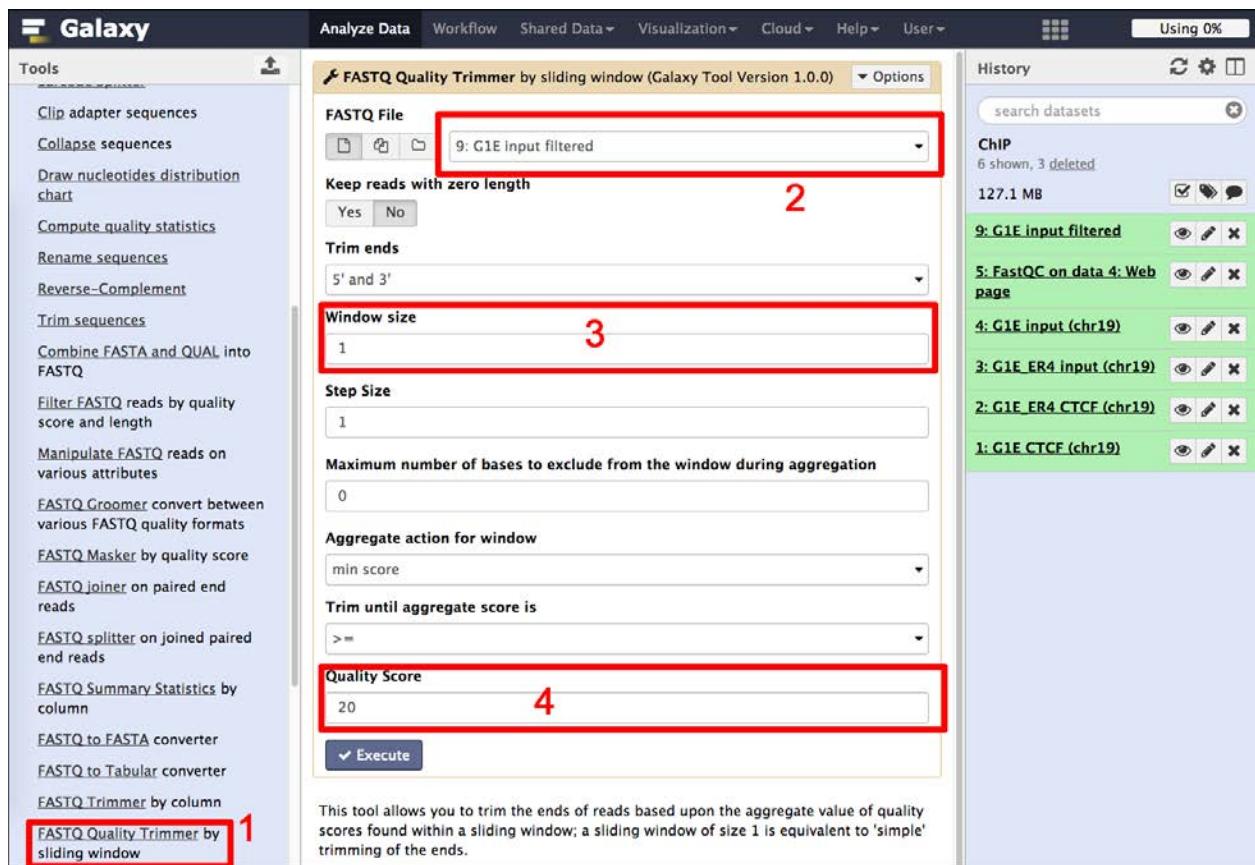


Figure 19: Filtering reads of bad quality.

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. On the left, there's a sidebar titled 'Tools' with a search bar and a list of categories like 'Get Data', 'Send Data', 'Lift-Over', etc. In the center, there's a banner for 'Galaxy 101' with the text 'Start small' and 'The very first tutorial you need'. Below the banner is a 'Tweets' section. On the right, there's a 'History' panel titled 'ChIP' showing 12 datasets. Some datasets are highlighted with red boxes: '16: G1E_ER4 input trimmed', '15: G1E_ER4 CTCF trimmed', '14: G1E CTCF trimmed', '13: G1E CTCF filtered', '10: G1E input trimmed', and '9: G1E input filtered'. The datasets are listed with their names and sizes (e.g., 264.6 MB).

Dataset ID	Dataset Name	Size
16	G1E_ER4 input trimmed	264.6 MB
15	G1E_ER4 CTCF trimmed	
14	G1E CTCF trimmed	
13	G1E CTCF 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).

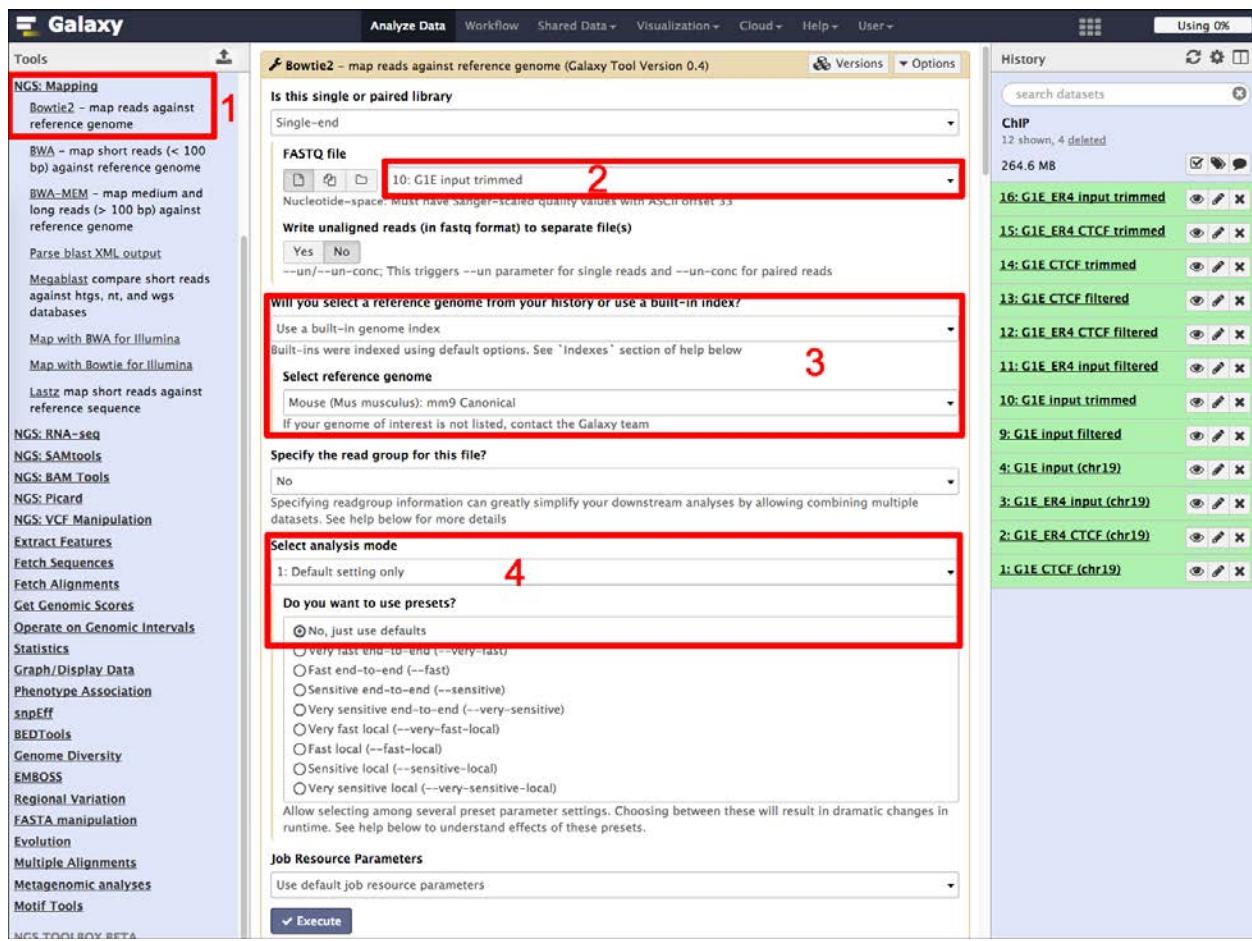


Figure 19: Bowtie2.

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*)

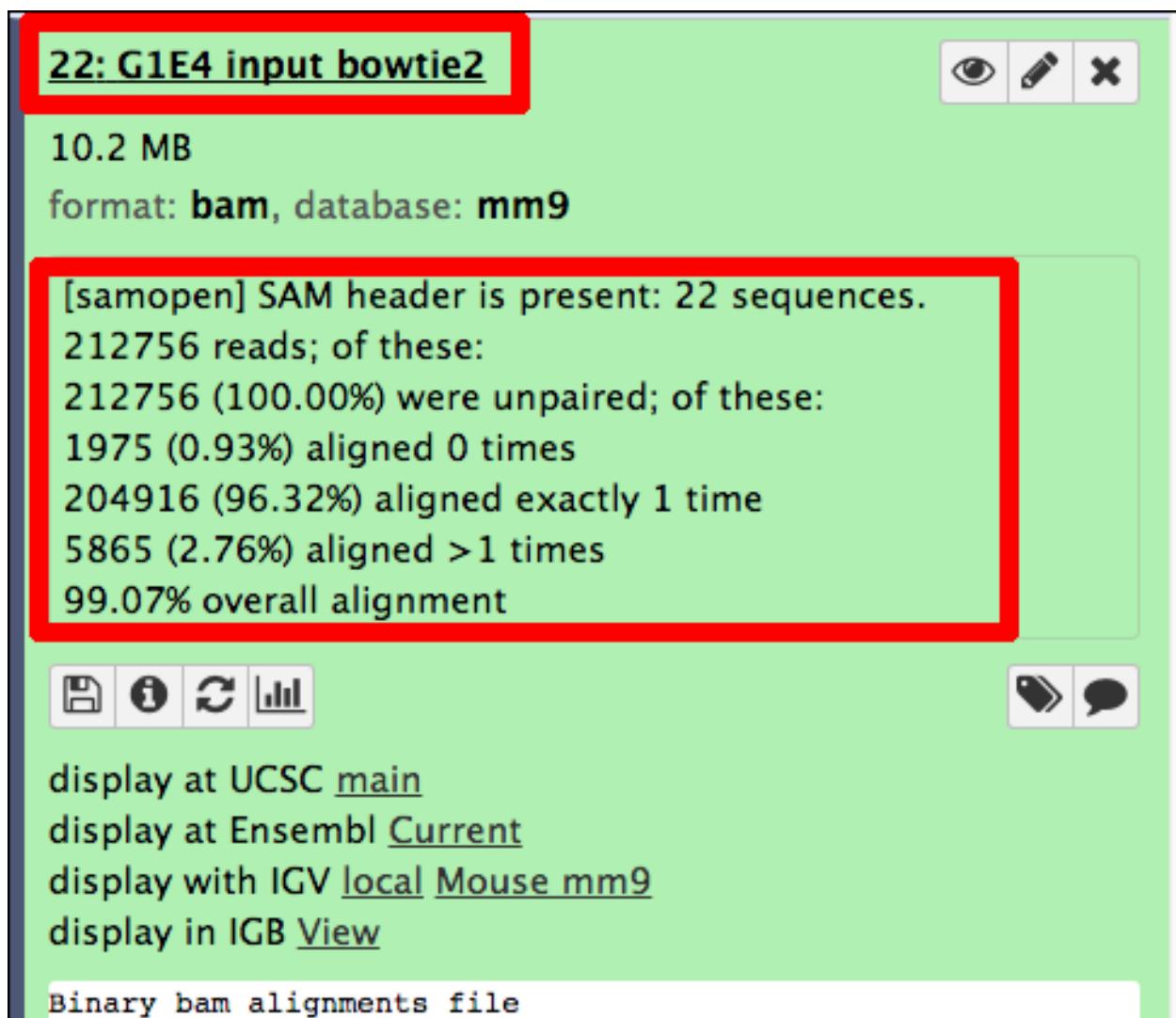


Figure 20: Bowtie2 mapping information.

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 with the 'Tools' sidebar open. The 'Tools' sidebar lists various NGS tools, with 'NGS: SAMtools' expanded. Under 'NGS: SAMtools', several options are listed, including 'Filter mapped reads on MD tag string', 'Merge BAM Files merges BAM files together', 'SAM-to-BAM convert SAM to BAM', 'Pileup-to-Interval condenses pileup format into ranges of bases', 'MPileup call variants', 'bcftools view Converts BCF format to VCF format', 'Reheader copy SAM/BAM header between datasets', 'Split BAM dataset on readgroups', 'Stats generate statistics for BAM dataset', 'BAM-to-SAM convert BAM to SAM', 'Sort BAM dataset', 'CalMD recalculate MD/NM tags', 'BedCov calculate read depth for a set of genomic intervals', 'IdxStats tabulate mapping statistics for BAM dataset', 'Flagstat tabulate descriptive stats for BAM dataset', 'Slice BAM by genomic regions', 'RmDup remove PCR duplicates', 'Filter_pileup on coverage and SNPs', 'Convert_SAM to interval', 'Filter_SAM on bitwise flag values', 'Generate_pileup from BAM dataset', and 'Filter_SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region'. The 'Filter_SAM or BAM, output SAM or BAM files on FLAG MAPQ RG LN or by region' option is highlighted with a red box. The main panel shows the configuration for this tool, including fields for 'Header in output' (set to 'Include Header'), 'Minimum MAPQ quality score' (set to '1'), 'Filter on bitwise flag' (set to 'no'), 'Select alignments from Library' (empty), 'Select alignments from Read Group' (empty), 'Output alignments overlapping the regions in the BED FILE' (empty), 'Select regions (only used when the input is in BAM format)' (empty), 'Select the output format' (set to 'BAM (-b)'), and a 'Execute' button.

Figure 20b: Samtools filtering.

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

Sort BAM dataset (Galaxy Version 2.0)

BAM File: 73: Filter SAM or BAM, output SAM or BAM on data 18: bam

Sort by: Chromosomal coordinates

Execute

What it does

This tool uses `samtools sort` command to sort BAM datasets in coordinate or read name order.

Citations Show BibTeX

Definition of SAM/BAM format. [Link]

Li, H., and Handsaker, B., and Wysoker, A., and Fennell, T., and Ruan, J., and Homer, N., and Marth, G., and Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. In *Bioinformatics*, 25 (16), pp. 2078–2079. doi:10.1093/bioinformatics/btp352| [Link]

Li, H. (2011). Improving SNP discovery by base alignment quality. In *Bioinformatics*, 27 (8), pp. 1157–1158. doi:10.1093/bioinformatics/btr076| [Link]

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. In *Bioinformatics*, 27 (21), pp. 2987–2993. doi:10.1093/bioinformatics/btr509| [Link]

Danecek, P., Schiffels, S., Durbin, R.. Multiallelic calling model in bcftools (-m). [Link]

Durbin, R.. Segregation based metric for variant call QC. [Link]

Li, H.. Mathematical Notes on SAMtools Algorithms. [Link]

SAMTools GitHub page. [Link]

History

ChIP
34 shown, 40 deleted
310.33 MB
74: Sort on data 73
73: Filter SAM or BAM, output SAM or BAM on data 18: bam
10.1 MB
format: bam, database: mm9
display at UCSC main
display at Ensembl Current
display with IGV local Mouse_mm9
display in IGB View
Binary bam alignments file
72: Extract Genomic DNA on data 71
71: Get flanks on data 70
70: Compute on data 6
63: Cut on data 62
60: Build custom track on data 27, data 25, and others
58: Compare two Data sets on data 55 and data 57

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).

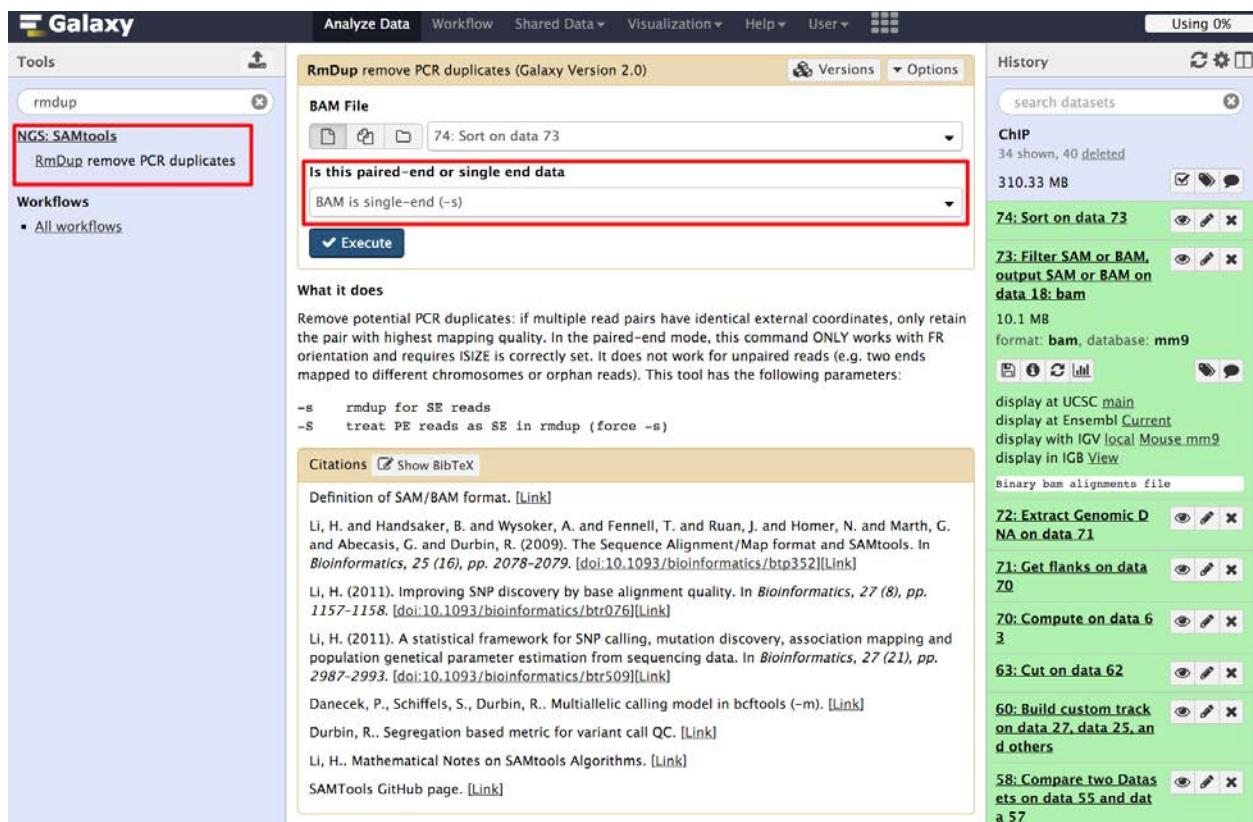


Figure 20d: Samtools rmdup.

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](#) 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.

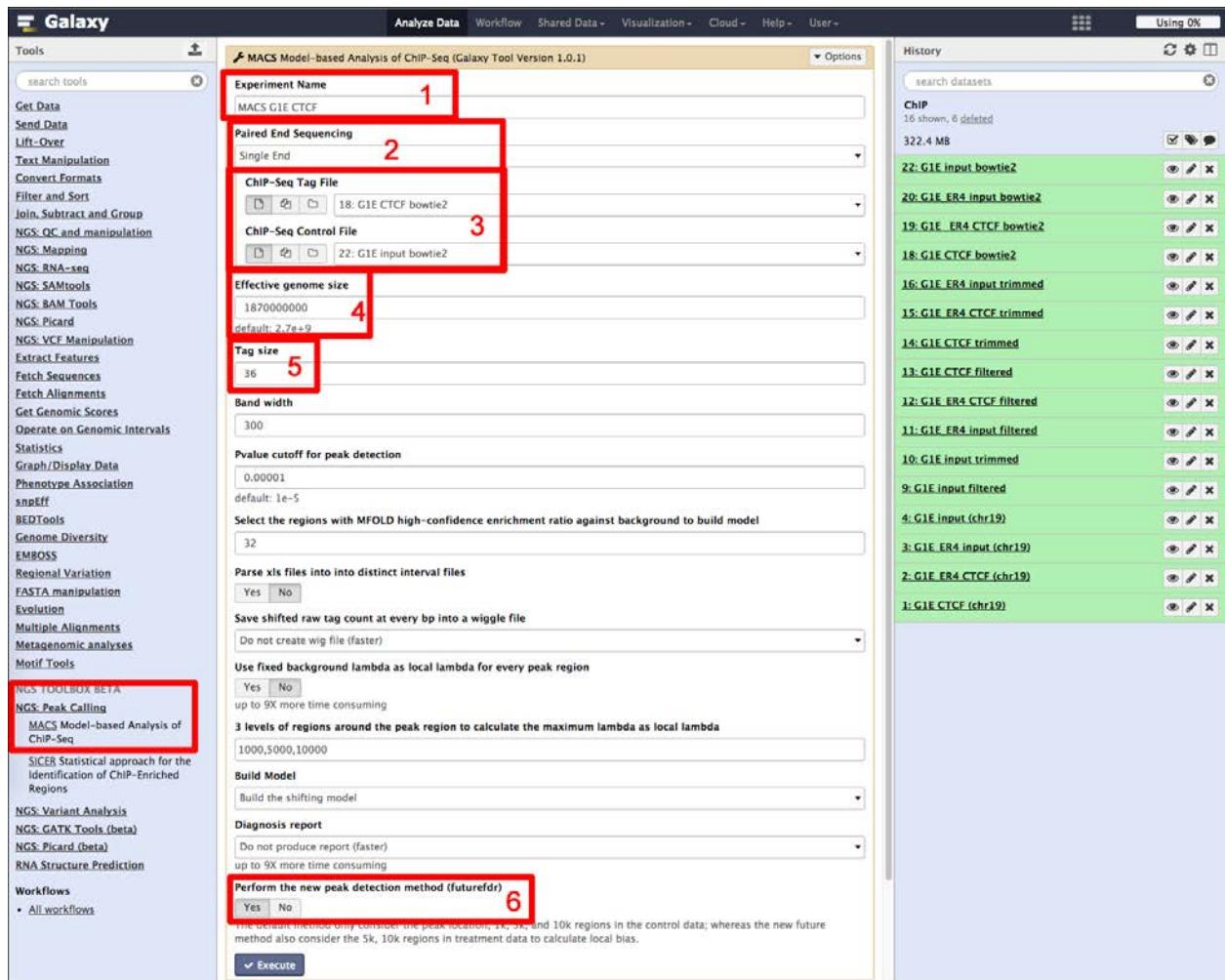


Figure 21: MACS peak calling.

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*).

Figure 22: MACS peak calling results.

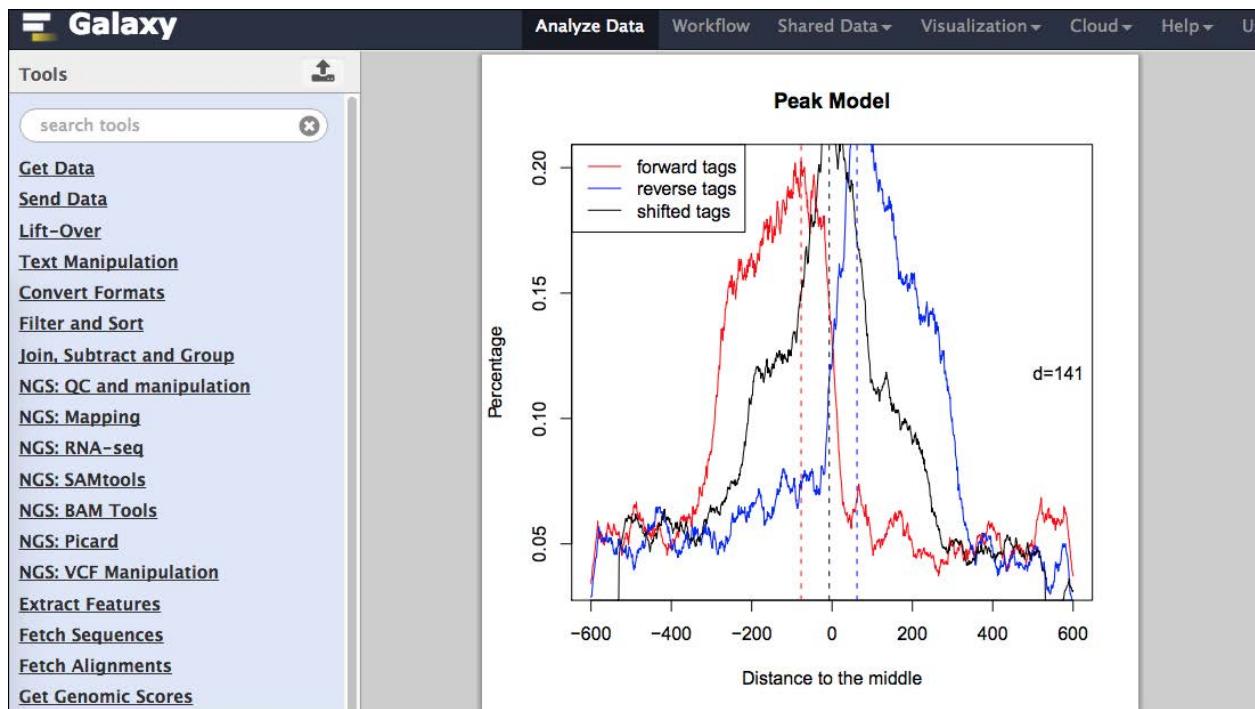


Figure 23: MACS peak model.

	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.

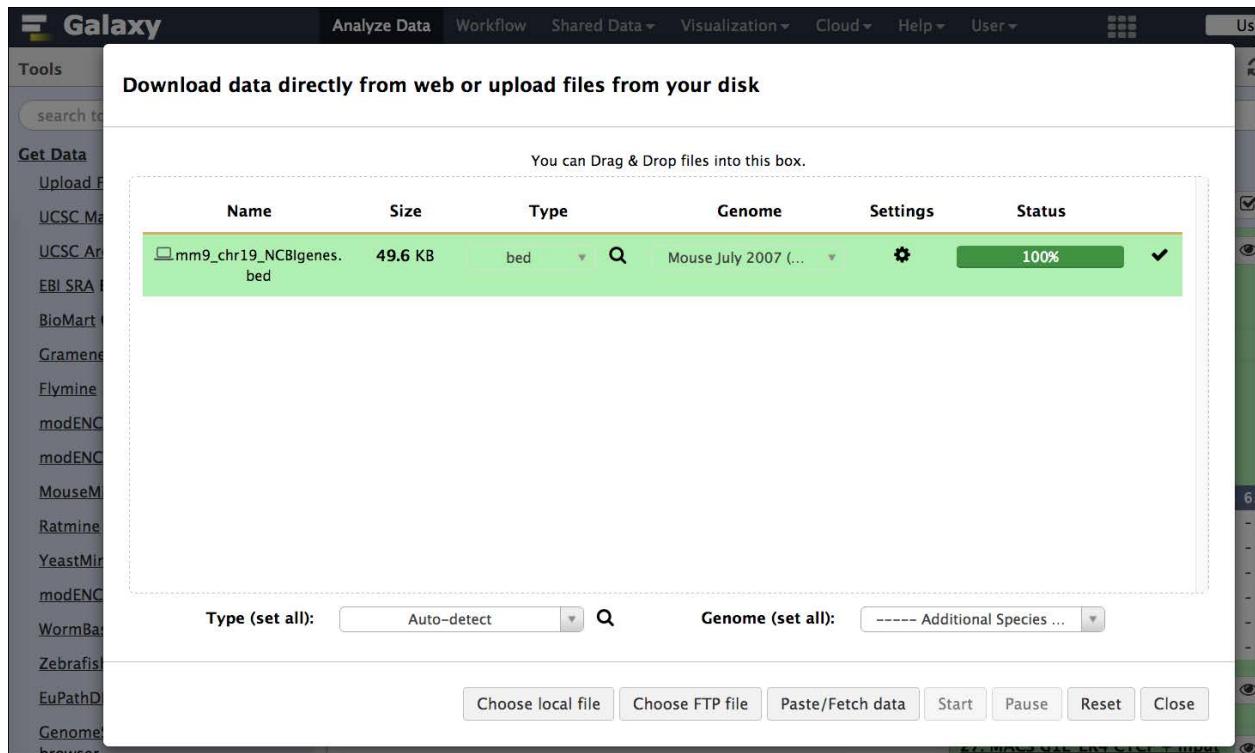


Figure 25: Upload the gene bed-file.

The screenshot shows the Galaxy web interface with a table of genomic data and a history panel. The table has columns labeled 1 through 6. The history panel shows a dataset named "51:mm9_chr19_NCBigenes.bed" with 1,428 regions, format bed, database mm9, and notes "uploaded bed file". It also lists display options: "display in IGB View", "display at Ensembl Current", and "display at UCSC main". A preview of the data table is shown at the bottom.

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	Mrp121	1	+
chr19	3283046	3292837	Mrp121	1	+
chr19	3323300	3385733	Cpt1a	1	+
chr19	3388868	3398168	Mtl5	1	+
chr19	3388868	3407785	Mtl5	1	+
chr19	3389400	3407785	Mtl5	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.

The screenshot shows the Galaxy web interface with the 'Get flanks' tool selected. The tool configuration is as follows:

- Select data:** Input dataset is '51: mm9_chr19_NCBigenes.bed'.
- Region:** Around Start.
- Location of the flanking region/s:** Upstream.
- Offset:** 0.
- Length of the flanking region(s):** 10000.

The history panel on the right shows the output dataset '51: mm9_chr19_NCBigenes.bed' with 1,428 regions, format bed, database mm9. It also lists other datasets like 'ChIP' and 'MACS G1E_ER4 CTCF + input'.

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.

The screenshot shows the Galaxy web interface with the 'Join' tool selected. The tool configuration is as follows:

- Join:** First dataset is '52: Gene promoters chr19'.
- with:** Second dataset is '25: MACS G1E CTCF + input (peaks)'.
- with min overlap:** 1 (bp).
- Return:** Only records that are joined (INNER JOIN).

The history panel on the right shows the output dataset '52: Gene promoters chr19' with 1,428 regions, format interval, database mm9. It also lists other datasets like 'ChIP' and 'MACS G1E_ER4 CTCF + input'.

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 Ctcf peaks for both peak-files.

2.2 Enrichment analysis (genes) with Enrichr

Now lets take the genes with Ctcf 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 web interface with the 'Group' tool selected. The left sidebar lists various tools under the 'Tools' category. The main panel shows the 'Group data by a column and perform aggregate operation' tool configuration. In the 'Select data' section, a dataset named '53: Overlap promoter + G1E CTCF' is chosen. In the 'Group by column' section, 'Column: 4' is selected. Below these, there are sections for ignoring case while grouping (Yes or No), ignoring lines beginning with specific characters (a list of symbols like >, @, +, <, *, -, =, |, ?, \$, ., :, &, %, ^, #), and an 'Operation' section with a 'Execute' button. To the right, the 'History' panel shows the results of the previous step, 'ChIP', which produced a dataset '53: Overlap promoter + G1E CTCF' containing 228 regions. This dataset is displayed as an interval database in mm9 format. Below the history are several other datasets listed in the history.

Figure 29: Aggregate the gene symbol column.

The screenshot shows the Galaxy web interface. On the left, the 'Tools' panel is open, displaying a list of available tools categorized under 'Text Manipulation' and 'Convert Formats'. In the center, a list of gene symbols is shown, starting with '1810006K21Rik' and '4930579J09Rik'. On the right, the 'History' panel shows a dataset named '55: Group on data 53' which contains 156 lines and is in tabular format for mm9. The dataset is grouped by column 4. Below this, another dataset is listed: '53: Overlap_promoter + G1E CT CE'.

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 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.

 [Login | Register](#)
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](#).

No file chosen

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](#).

Ift12
Ighmbp2
Kazald1
Kcnip2
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.

 **Enrichr**

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:

- Find and note the top 5 enriched Gene Ontology process terms for both the G1E and G1E_ER4 genes that have Ctcf in their promoters.
- 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 Ctcf 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** excepts 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 except floats as scores and will produce errors), copy the regions and paste them into the **GREAT** interface.

GREAT Overview News Use GREAT Demo Video How to Cite Help Forum

GREAT version 3.0.0 current (02/15/2015 to now)

GREAT predicts functions of cis-regulatory regions.

Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions. Cis-regulatory regions can be identified via both experimental methods (e.g. ChIP-seq) and by computational methods (e.g. comparative genomics). For more see our [Nature Biotech Paper](#).

News

- Feb 15, 2015: GREAT version 3.0 switches to Ensembl genes, adds the mouse mm10 assembly, and adds new ontologies.
- Apr 3, 2012: GREAT version 2.0 adds new annotations to human and mouse ontologies and visualization tools for data exploration.
- Feb 18, 2012: The [GREAT forums](#) are released, allowing increased user-to-user interaction

[More news items...](#)

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: [Choose 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: [Choose File](#) No file chosen
- BED data:

[When should I use a background set?](#)
[What should my background regions file contain?](#)

Association rule settings

Show settings »

Submit **Reset** **Help**

Figure 33: The GREAT website.

GREAT Overview News Use GREAT Demo Video How to Cite Help Forum Bejerano Lab, Stanford University

GREAT version 3.0.0 current (02/15/2015 to now)

Job Description

Region-Gene Association Graphs

Global Controls Global Export Which data is exported by each option?

GO Molecular Function (4 terms) Global controls

Table controls: Export Shown top rows in this table: 20 Set Term annotation count: Min: 1 Max: Inf Set Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water	1	2.6326e-16	9.1693e-13	78.0167	10	2.47%	1	8.8353e-6	37.2489	6	9	1.58%
iron ion binding	4	2.2764e-12	1.9822e-9	5.2559	28	6.91%	3	3.7623e-3	3.9732	16	225	4.22%
stearoyl-CoA 9-desaturase activity	6	1.7961e-11	1.0426e-8	69.6114	7	1.73%	2	2.5647e-3	37.2489	4	6	1.06%
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	17	6.7018e-8	1.3731e-5	3.9978	22	5.43%	4	2.1462e-2	3.7607	14	208	3.69%

The test set of 405 genomic regions picked 379 (2%) of all 21,176 genes. GO Molecular Function has 3,483 terms covering 15,735 (74%) of all 21,176 genes, and 181,165 term - gene associations. 3,483 ontology terms (100%) were tested using an annotation count range of [1, Inf].

GO Biological Process (1 term) Global controls

Table controls: Export Shown top rows in this table: 20 Set Term annotation count: Min: 1 Max: Inf Set Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
Wnt receptor signaling pathway	2	1.4738e-14	7.4104e-11	4.1678	42	10.37%	3	4.7691e-2	3.7249	15	225	3.96%

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).

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

- Left Sidebar (Tools):**
 - Graph/Display Data (highlighted with a red box)
 - Build custom track for UCSC genome browser
 - Scatterplot of two numeric columns
 - Histogram of a numeric column
 - Plotting tool for multiple series and graph types
 - GMAI Multiple Alignment Viewer
 - Boxplot of quality statistics
 - VCF to MAF Custom Track for display at UCSC
 - Phenotype Association
 - snpEff
 - BEDTools
- Top Bar:** Analyze Data, Workflow, Shared Data, Visualization, Cloud, Help, User
- Tool Panel (Build custom track for UCSC genome browser):**
 - Track 1:** Dataset 56: Overlap promoter + G1E_ER4 CTCF. Configuration: name = Gene promoter overlapping G1E_ER4 CTCF, description = User Supplied Track (from Galaxy), Color = Green, Visibility = Dense.
 - Track 2:** Dataset 53: Overlap promoter + G1E CTCF. Configuration: name = Gene promoter overlapping G1E CTCF, description = User Supplied Track (from Galaxy), Color = Black, Visibility = Dense.
 - Track 3:** Dataset 25: MACS G1E CTCF + input (peaks). Configuration: name = MACS G1E CTCF + input (peaks).
- History Panel:** Shows a list of datasets, including ChIP, MACS, and Bowtie2 results, each with edit and delete icons.

Figure 35: Building a custom UCSC track.

The screenshot shows the Galaxy web interface. On the left, a sidebar titled 'Tools' lists various bioinformatics tools under categories like 'Get Data', 'Text Manipulation', 'Convert Formats', etc. In the center, a table displays genomic data with columns labeled 1, 2, 3, and 4. The first column contains chromosome identifiers (chr19). The second column contains values such as 3197714, 3207714, etc. The third column contains values such as 0, 1, 2, etc. The fourth column contains values such as 3283010, 3293010, etc. To the right of the table is a 'History' panel. The top part of the history panel shows a list of datasets: 'ChIP' (28 shown, 33 deleted), '324.2 MB' (with a file icon), and a green box containing '60: Build custom track on data 27, data 25, and others'. This green box also indicates '1,641 lines, 5 comments' and 'format: customtrack, database: mm9'. Below this, it says 'Generated a custom track containing 4 subtracks.' and has a 'display at UCSC main' button. At the bottom of the history panel, there is another table with columns 1 and 2, matching the structure of the main table.

Figure 36: Visualising a Galaxy dataset/track.

The screenshot shows the JCS Genome Browser interface. At the top, a navigation bar includes links for 'Genomes', 'Genome Browser', 'Tools', 'Mirrors', 'Downloads', 'My Data', 'View', and 'Help'. Below the navigation bar, a search bar shows 'chr19:3,152,709-3,252,718 100,010 bp.' and a placeholder 'enter position, gene symbol or search terms'. A 'go' button is followed by 'hg38'. Below the search bar, a message reads 'replaces hg19 as default human assembly'. The main content area shows a genomic track for chromosome 19. The track is labeled 'User Supplied Track (from Galaxy)' and 'User Supplied Track'. It includes 'MACS peaks for MACS_G1E_CTCF' and 'MACS peaks for MACS_G1E_ER4_CTCF'. Other tracks visible include 'Scale', 'chr19:', 'ping G1E_ER4 CTCF', 'User Track', 'Mapping G1E CTCF', 'for MACS_G1E_CTCF', 'MACS_G1E_ER4_CTCF', 'ENCCFF447V0V', 'STS Markers', 'UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)', and 'AK0877035', 'AK086563', 'HK087025'. A zoom scale from 1.5x to 100x is at the top, and a 'zoom in' and 'zoom out' button is present.

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 Ctcf. 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 a 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*).

The screenshot shows the Galaxy web interface with the 'Compute an expression on every row' tool selected. The configuration panel on the right contains the following settings:

- Add expression:** c2+1
- as a new column to:** 63: Cut on data 62
- Round result?**: YES

A large blue 'Execute' button is at the bottom.

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).

The screenshot shows the Galaxy History panel with the following details:

- Dataset Name:** 64: Compute on data 6
- Description:** 3
- Size:** 324.2 MB
- Format:** tabular, database: mm9
- Notes:** Creating column 3 with expression c2+1 kept 100.00% of 405 lines.
- Content Preview:** A table with columns 1, 2, and 3, showing data for chromosomes 19 and 21.

Figure 41: Center peaks.

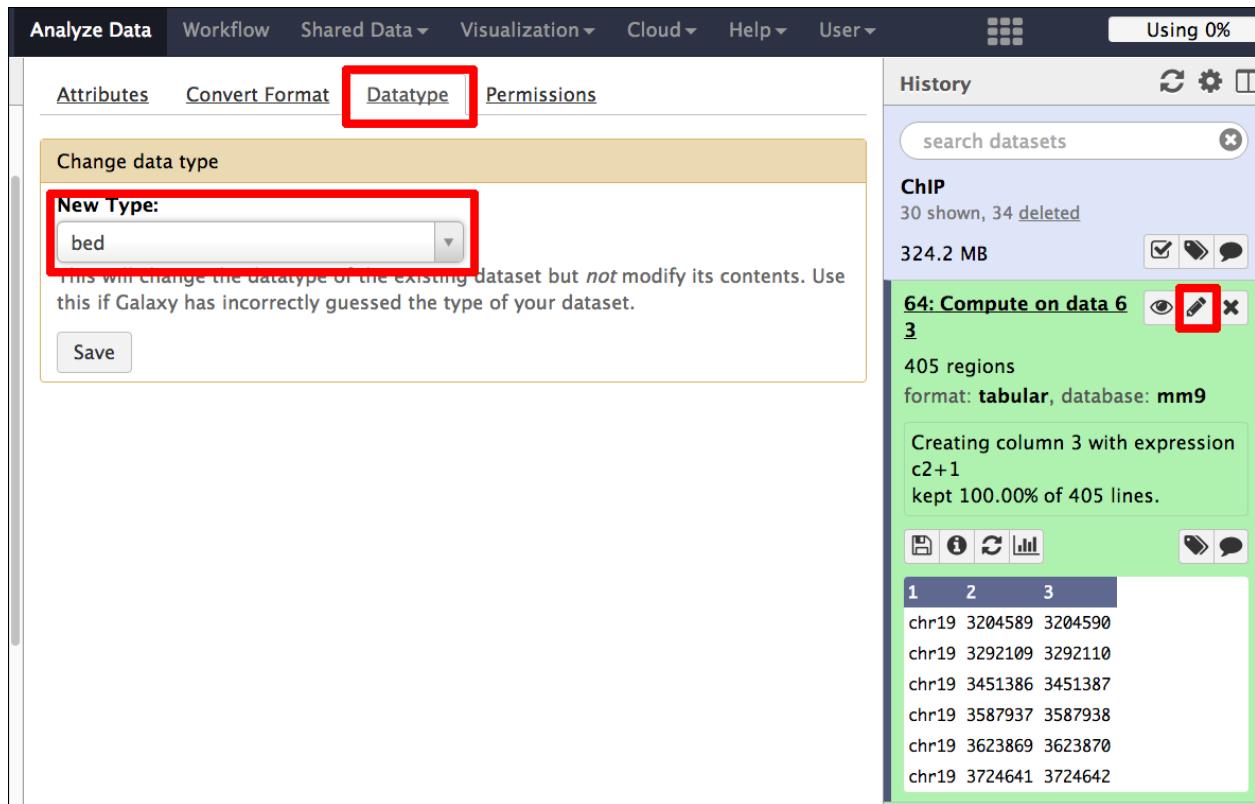


Figure 42: Change the data-format to bed.

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

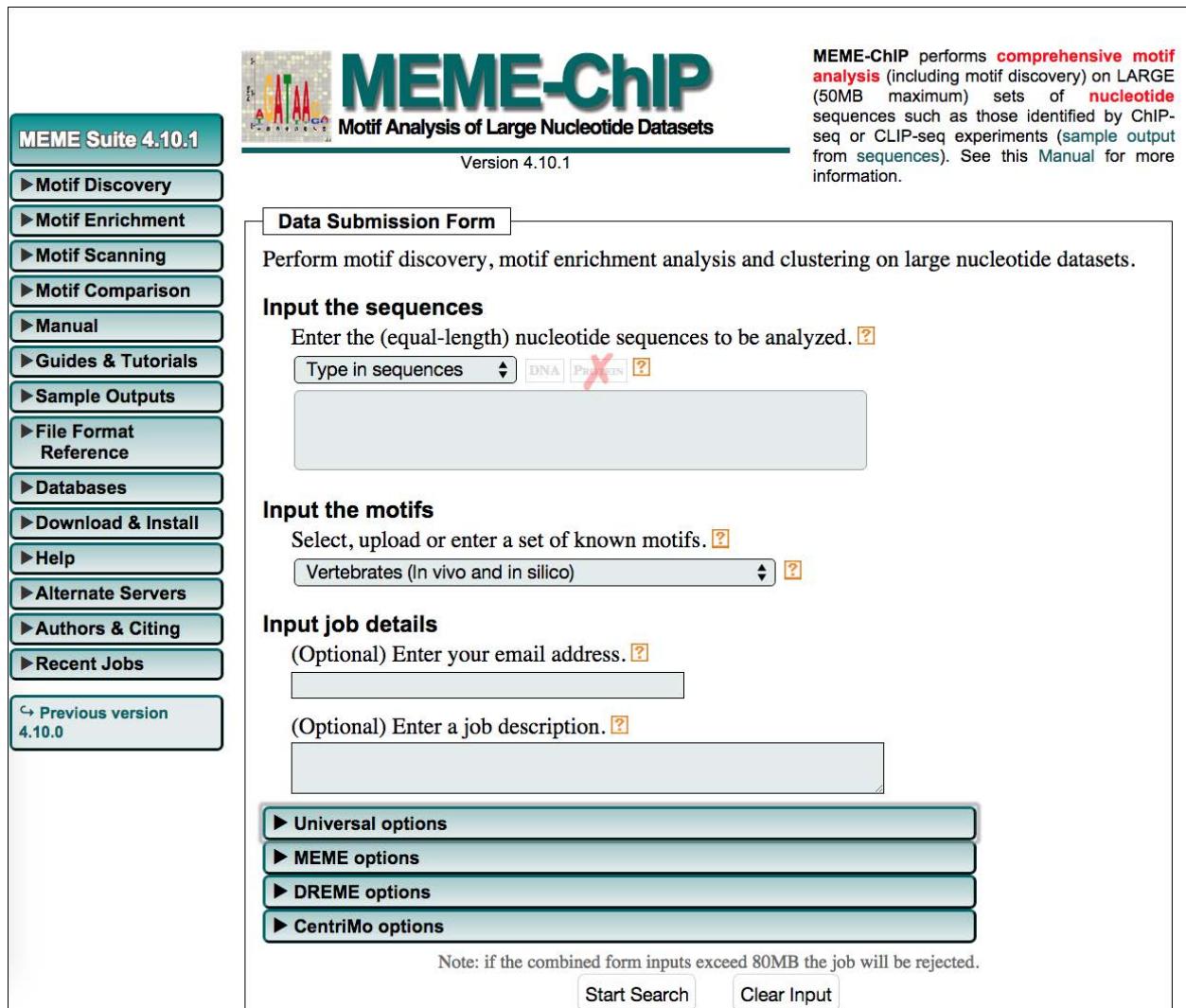
Use Fetch Sequences => Extract Genomic DNA to extract for the regions the genomic DNA (see Figure 44).

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) (<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 Analysis of Large Nucleotide Datasets

Version 4.10.1

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. [?](#)
 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.

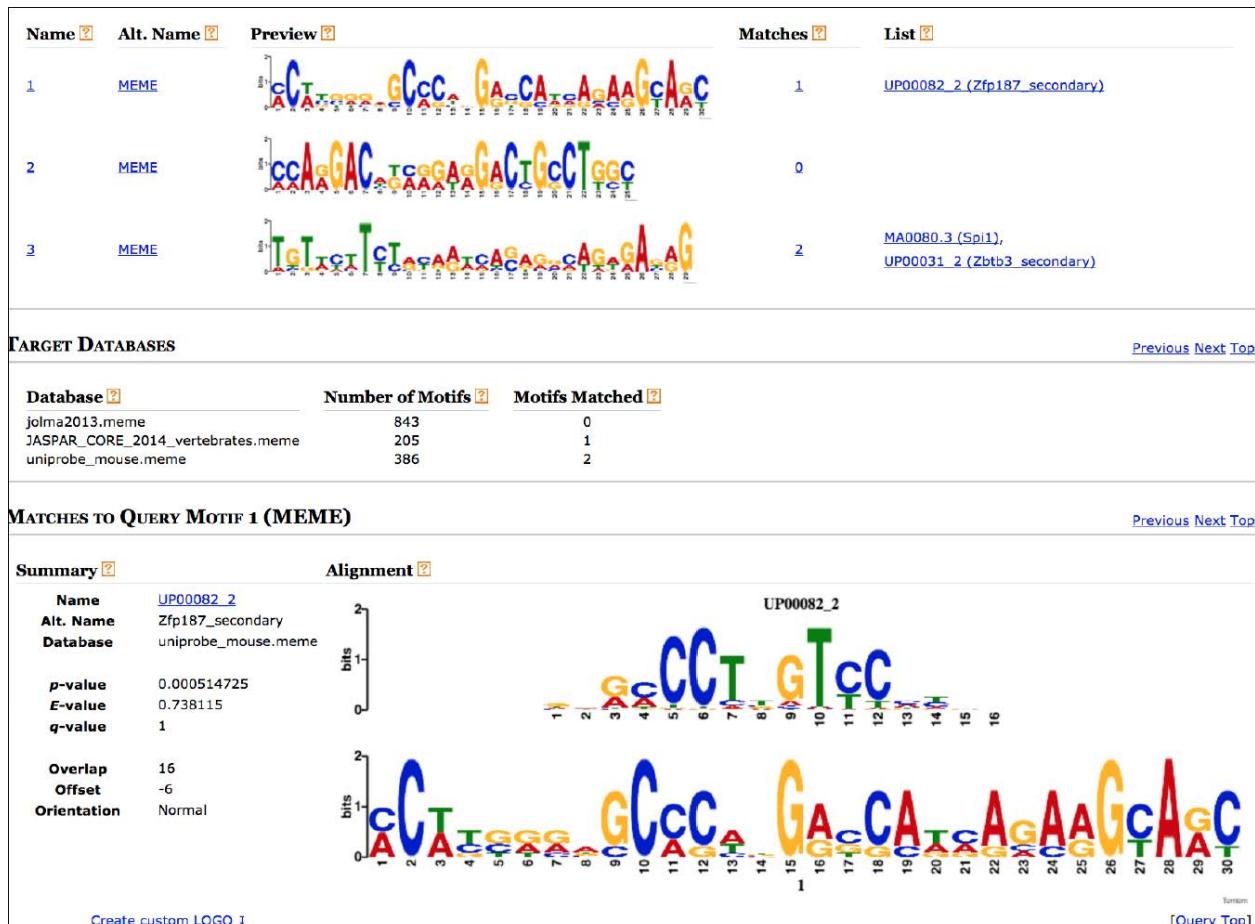


Figure 46: MEME-ChIP results.

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/>