

Working with BED Files in Galaxy and Conducting GO Enrichment Analysis

MMG1001 Assignment 4 – Group Heather (originally developed by Laura Campitelli)

Corresponding Reading

- Visel *et al.* 2009. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457(7231): 854–858

Tools

- **UCSC Table Browser:** <https://genome.ucsc.edu/cgi-bin/hgTables>
- **Galaxy** (online server to execute **bedtools**): <https://usegalaxy.org>
- The **Gene Ontology (GO)** resource: <http://geneontology.org>

Additional Files

- **forebrain_peaks_p300.bed** and **limb_peaks_p300.bed** (from Visel *et al.*'s supplemental tables 2 and 4, modified to remove extra columns and headers)

Overview

The authors of the paper found that forebrain p300 peaks were particularly enriched 10kb up- or downstream of genes expressed in E11.5 forebrain tissue. We will follow a workflow to identify what genes are within this 10kb region of the p300 peaks and perform a GO enrichment analysis to see what molecular functions these genes have.

Step 1: Download BED file of mouse mm9 genes

- Go to the [UCSC Table Browser](https://genome.ucsc.edu/cgi-bin/hgTables) and select the following options to download a BED file of all the mouse genes with **RefSeq** annotations under the **mm9** reference genome, provide an informative name for the output file (such as **mm9_refseq_genes.bed**), then click *get output*:

The screenshot shows the UCSC Table Browser interface. The following options are selected and highlighted with red boxes:

- clade:** Mammal
- genome:** Mouse
- assembly:** July 2007 (NCBI37/mm9)
- group:** Genes and Gene Predictions
- track:** RefSeq Genes
- table:** refGene
- region:** genome
- output format:** BED - browser extensible data
- output file:** mm9_refseq_genes.bed
- file type returned:** gzip compressed

Red annotations on the right side of the form:

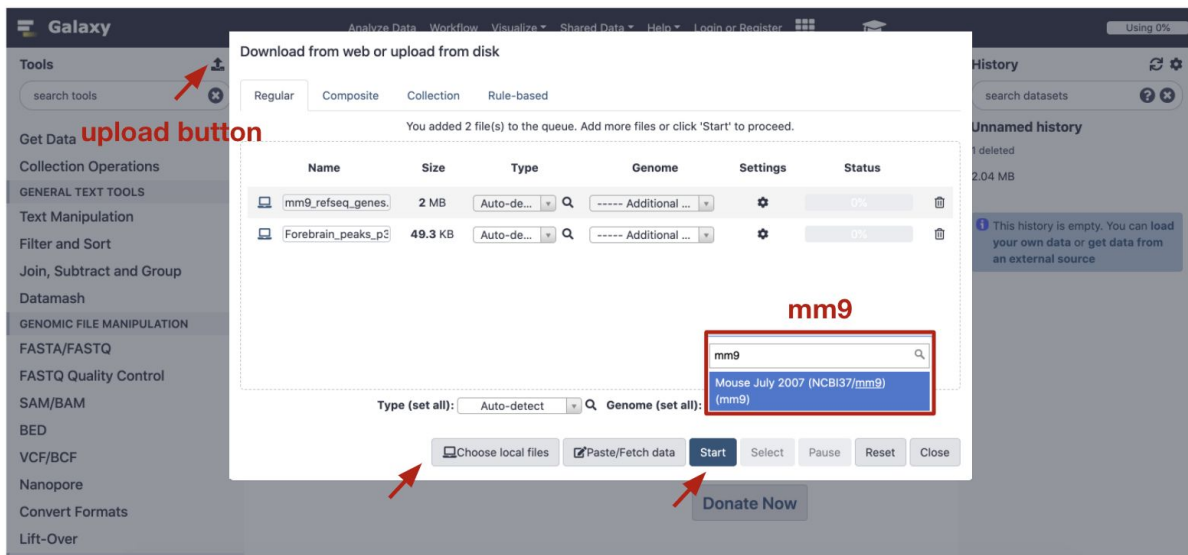
- Mouse mm9 RefSeq Genes** (pointing to the track selection)
- genome** (pointing to the region selection)
- BED file name** (pointing to the output file name)
- gzip compressed** (pointing to the file type selection)

A red arrow points to the **get output** button. Below the form, a link is provided: "To reset all user cart settings (including custom tracks), [click here](#)."

- On the new page, leave the default settings as they are (no custom header, one BED record per whole gene) and click *get BED*. Save the file in a folder you'll remember

Step 2: Upload files to Galaxy

- Go to the [Galaxy](#) main page. The left side lists the available tools and the right side will hold the files and command history. At the top left, next to *Tools*, click the upload button
- Select *Choose local files*, then add **mm9_refseq_genes.bed.gz** and **forebrain_peaks_p300.bed**
- For *Genome (set all)*, search for **mm9** and click *Start*. When the files turn green, click *Close* to exit the window
- When your files have finished uploading to Galaxy, they will turn green on the right side of the screen



Step 3: Use bedtools slop to add 10kb flanks to p300 peaks

- On the left side of Galaxy under *BED*, select **bedtools SlopBed**
- Specify the BED file to be **forebrain_peaks_p300.bed** and select **mm9** for the *Genome file*
- Keep all default parameters except for *Number of base pairs*, which should be changed to **10000**. Click *Execute*

bedtools Slopbed

The screenshot shows the Galaxy web interface for the **bedtools SlopBed** tool. The tool description on the left states: "bedtools SlopBed adjust the size of intervals". The configuration panel shows the following settings:

- BED/bedGraph/GFF/VCF file:** 2: Forebrain_peaks_p300.bed
- Genome file:** mm9
- Define -l and -r as a fraction of the feature's length:** Yes
- Define -l and -r based on strand:** Yes
- Choose what you want to do:** Increase the BED/bedGraph/GFF/VCF entry by the same number base pairs in each direction.
- Number of base pairs:** 10000
- Print the header from the A file prior to results:** Yes

The **Execute** button is highlighted with a red arrow.

- A new file will appear in your history on the right, named something like *bedtools SlopBed on data 2*. We can change this to a more informative name like **forebrain_peaks_p300_10k_flanks.bed** by clicking the pencil next to the new file, editing the name, and clicking *Save*. This new file contains the locations of the p300 peaks in forebrain tissue with coordinates extended by 10kb in both directions

The screenshot shows the **Edit dataset attributes** dialog for the file **3: bedtools SlopBed on data 2**. The **Name** field is set to **forebrain_peaks_p300_10k_flanks.bed**. The **Save** button is highlighted with a red arrow.

Step 4: Use bedtools intersect to identity mouse genes within 10kb of p300 peaks

- On the left side of Galaxy under *BED*, select **bedtools Intersect intervals**
- Specify *File A* to be the newly created **forebrain_peaks_p300_10k_flanks.bed**
- Specify *File B* to be **mm9_refseq_genes.bed.gz**
- For *What should be written to the output file*, select **Write the original entry in B for each overlap...**
- Keep the remaining default parameters and click *Execute*. Rename the new file **forebrain_peaks_p300_mm9_refseq_genes.bed**. This new file filtered down the original bed file of mouse genes to only those genes that were found within 10kb of the forebrain p300 peaks

bedtools Intersect intervals

The screenshot displays the Galaxy web interface for the 'bedtools Intersect intervals' tool. The left sidebar shows the 'Tools' section with 'bedtools Intersect intervals' highlighted. The main panel shows the tool configuration for 'bedtools Intersect intervals find overlapping intervals in various ways (Galaxy Version 2.29.0)'. The configuration includes:

- File A to intersect with B:** 3: forebrain_peaks_p300_10k_flanks.bed
- File B to intersect with A:** 1: mm9_refseq_genes.bed.gz
- BAM/BED/bedGraph/GFF/VCF format:** BAM/BED/bedGraph/GFF/VCF format
- Combined or separate output files:** One output file per 'input B' file
- Calculation based on strandedness?** Overlaps on either strand
- What should be written to the output file?** ☒ Write the original entry in B for each overlap. Useful for knowing what A overlaps. Restricted by the fraction- and reciprocal option (-wb)
- Treat split/spliced BAM or BED12 entries as distinct BED intervals when computing coverage.** ☐ Yes ☒ No
- Required overlap:** Default: 1bp
- Report only those alignments that **do not** overlap with file(s) B:** ☐ Yes ☒ No

The right sidebar shows the 'History' section with 'Unnamed history' and a list of files: 3: forebrain_peaks_p300_10k_flanks.bed, 2: Forebrain_peaks_p300.bed, and 1: mm9_refseq_genes.bed.gz.

Step 5: Obtain the list of RefSeq genes found within 10kb of p300 forebrain peaks

- On the left side of Galaxy under *Text Manipulation*, select *Cut*
- Specify *File to cut* to be **forebrain_peaks_p300_mm9_refseq_genes.bed**
- When you click on **forebrain_peaks_p300_mm9_refseq_genes.bed** on the right side, you can see a snippet of the data it contains. If you scroll to the right you can see that the RefSeq names are in column 7. Back in the middle of the page, under *List of Fields*, select **Column: 7**
- Keep the remaining default parameters and click *Execute*. The new file will be just the list of RefSeq genes, but it might contain duplicate entries

Cut

Galaxy

Analyze Data Workflow Visualize Shared Data Help Login or Register

Tools

search tools

Join two files

Replace Text in entire line

UniProt ID mapping and retrieval

Replace column by values which are defined in a convert file

Text transformation with sed

Unfold columns from a table

Unique lines assuming sorted input file

Replace Text in a specific column

Multi-Join (combine multiple files)

Select last lines from a dataset (tail)

Cut columns from a table (cut)

Create text file with recurring lines

Unique occurrences of each record

Sort data in ascending or descending order

Search in textfiles (grep)

Sort a row according to their columns

Select first lines from a dataset (head)

tac reverse a file (reverse cat)

Concatenate datasets tail-to-head (cat)

Add column to an existing dataset

Advanced Cut columns from a table (cut) (Galaxy Version 1.1.0)

File to cut

5: forebrain_p300_peaks_mm9_refseq_genes.bed

Operation

Keep

Delimited by

Tab

Cut by

fields

List of Fields

Select/Unselect all

Column: 7

Execute

What it does

This tool runs the cut unix command, which extract or delete columns from a file.

Field List Example:

1,3,7 - Cut specific fields/characters.

History

search datasets

Unnamed history

4 shown, 1 deleted

5: forebrain_p300_peaks_mm9_refseq_genes.bed

3,736 regions

format: bed, database: mm9

display in IGB View

display at Ensembl Current

display with IGV local Mouse mm9

display at UCSC main

4: New 5: 6: Strand 7: 8

91 chr1 58983545 11293763 NM_00130525 0

26 chr1 11484185 11965983 NM_001318451 0

26 chr1 11484185 11965983 NM_001374634 0

26 chr1 11484185 11965983 NM_001374633 0

26 chr1 11484185 11965983 NM_001168369 0

3: forebrain_peaks_p300_10k_flanks.bed

2: Forebrain_peaks_p300.bed

1: mm9_refseq_genes.bed

- On the left side of Galaxy under *Text Manipulation*, select *Unique* (**important**: make sure you do **not** select *Unique lines*, which should also work in theory, but was giving weird results when testing)
- Specify *File to scan for unique values* to be the output of the last command (something like *Advanced Cut on data 4*)
- Keep the remaining default parameters and click *Execute*. The new file will be the sorted list of unique RefSeq genes (no duplicate entries)

Unique

Galaxy

Analyze Data Workflow Visualize Shared Data Help Login or Register

Tools

search tools

Join two files

Replace Text in entire line

UniProt ID mapping and retrieval

Replace column by values which are defined in a convert file

Text transformation with sed

Unfold columns from a table

Unique lines assuming sorted input file

Replace Text in a specific column

Multi-Join (combine multiple files)

Select last lines from a dataset (tail)

Cut columns from a table (cut)

Create text file with recurring lines

Unique occurrences of each record

Sort data in ascending or descending order

Search in textfiles (grep)

Sort a row according to their columns

Unique occurrences of each record (Galaxy Version 1.1.0)

File to scan for unique values

6: Advanced Cut on data 5

Ignore differences in case when comparing

Yes No

Column only contains numeric values

Yes No

Advanced Options

Hide Advanced Options

Execute

Syntax

This tool returns all unique lines using the 'sort -u' command. It can be used with unsorted files. If you need additional options, like grouping or counting your unique results, please use the 'Unique lines from sorted file' tool.

History

search datasets

Unnamed history

5 shown, 1 deleted

6: Advanced Cut on data 5

5: forebrain_p300_peaks_mm9_refseq_genes.bed

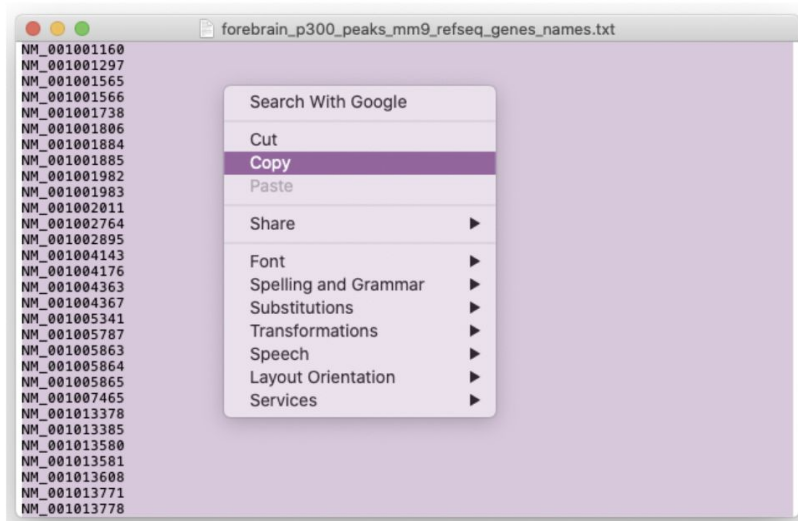
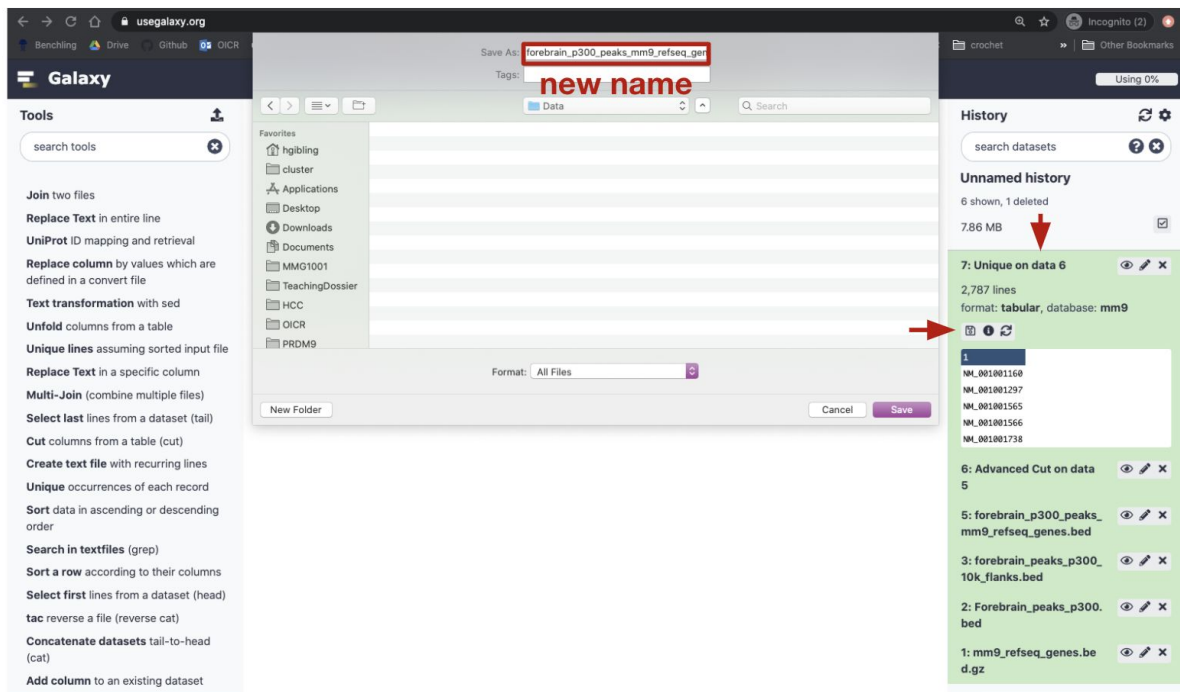
3: forebrain_peaks_p300_10k_flanks.bed

2: Forebrain_peaks_p300.bed

1: mm9_refseq_genes.bed

Step 6: Download files from Galaxy

- Download the new file by clicking on the name on the right side of the page and clicking on the floppy disc icon. Rename it **forebrain_peaks_p300_mm9_refseq_genes_names.txt** and save it in a folder you'll remember
- Open the file (in a plain text editor), do *control-a* or *command-a* to select all of the text, and then copy the text



Step 7: Perform GO enrichment analysis for the genes within 10kb of p300 forebrain peaks

- Go to the [Gene Ontology](#) page and paste in the list of RefSeq genes into the box on the right
- Select *biological process* and *Mus musculus*, then click *Launch*

THE GENE ONTOLOGY RESOURCE

GO Enrichment Analysis ?

Powered by PANTHER

Current release : GO terms | annotations
gene products | species (see statistics)

Search GO term or Gene Product in AmiGO ...

Any ● Ontology ● Gene Product

paste gene names

biological process

Mus musculus

Launch

- Panther software performs an enrichment analysis of the GO terms associated with each of the input genes, using *Fisher's Exact test* and *FDR correction* as defaults. The default display groups the GO terms by hierarchy, in decreasing order of fold enrichment (for the topmost term in the hierarchy)

GO biological process complete		Mus musculus (REF)	upload_1 (▼ Hierarchy, NEW! ?)					
		#	#	expected	Fold Enrichment	+/-	raw P value	FDR
hierarchy	sympathetic ganglion development	9	5	.27	18.71	+	3.16E-05	2.51E-03
	↳ ganglion development	16	6	.48	12.63	+	2.82E-05	2.26E-03
	↳ tissue development	1665	107	49.43	2.16	+	2.35E-13	1.24E-10
	↳ anatomical structure development	5143	263	152.68	1.72	+	6.38E-21	1.68E-17
	↳ developmental process	5533	275	164.26	1.67	+	2.44E-20	4.81E-17
	↳ animal organ development	3123	180	92.72	1.94	+	3.52E-18	4.62E-15
	↳ system development	4153	238	123.29	1.93	+	5.46E-25	4.31E-21
	↳ multicellular organism development	4769	252	141.58	1.78	+	8.61E-22	2.72E-18
	↳ multicellular organismal process	7309	297	216.99	1.37	+	2.01E-10	5.47E-08
	↳ nervous system development	2087	161	61.96	2.60	+	2.57E-28	4.06E-24
↳ sympathetic nervous system development	23	6	.68	8.79	+	1.51E-04	9.37E-03	

topmost term
(most specific)

fold enrichment

corrected p-value

Questions

- Which GO terms related to forebrain development are enriched, and what is the fold enrichment and corrected p-value? (topmost (most specific) term is fine)
- Why are there enriched GO terms not related to forebrain development?
- Why is looking at GO enrichment useful (compared to looking at lists of enriched genes)?

If you have time, repeat steps 2-7 with the limb p300 peaks. It might help to first delete all your Galaxy files except for mm9_refseq_genes.bed.gz (or make sure they are clearly named 'forebrain').

Which GO terms related to limb development are enriched? Are any GO terms enriched in both limb and forebrain?