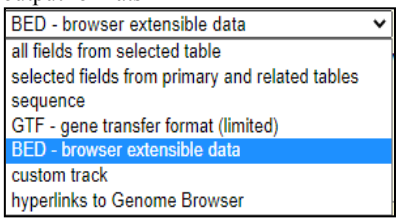
# #Introduction to Galaxy Analyses tools

## From peaks to genes tools

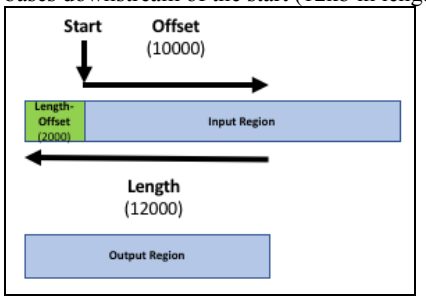
Goal to turn list of genomic regions into a list of possible target genes

| Tool | What for | What in | What out | *Example operation* |
|---|---|---|---|---|
| **UCSC Main** table browser | **Retrieve and export data from the Genome Browser** annotation track database | What in: **select** dataset, **define** region of interest (genome \| position + identifiers), **retrieve and display data**<br>**Send output to Galaxy:** displays results of query in Galaxy, a framework for interactive genome analysis.<br>**Send output to GREAT:** displays the functional enrichments of the query results in GREAT, a tool for analysis of the biological function of cis-regulatory regions.<br>**get output:** Submits a data query based on the specified parameters and returns the output.<br>**Summary/statistics:** Displays statistics about the data specified by the parameters. | output formats<br>  <br>file type returned<br>**plain text** - data is in ASCII format<br>***.gzip** compressed archieve format for Linux\|Unix | Retrieve list of genes of an animals, viruses, insects for i.e. mice etc.<br>*https://genome.ucsc.edu/cgi-bin/hgTables?GALAXY_URL=https%3A//usegalaxy.eu/tool_runner&tool_id=ucsc_table_direct1&sendToGalaxy=1&hgta_compressType=none&hgta_outputType=bed* |
| **Select last** lines from a dataset (tail) | **Keep the last X lines in a file**. It is a tool from text processing tools | **Keep last lines**\| Keep from this line on<br>Input File:<br>chr7 57134 57154 D17003_CTCF_R7 356 -<br>chr7 57247 57267 D17003_CTCF_R4 207 +<br>chr7 57314 57334 D17003_CTCF_R5 269 +<br>chr7 57341 57361 D17003_CTCF_R7 375 +<br>chr7 57457 57477 D17003_CTCF_R3 188 + | **Output**<br>Show last two lines of above file. The result is:<br>chr7 57341 57361 D17003_CTCF_R7 375 +<br>chr7 57457 57477 D17003_CTCF_R3 188 + | to compare the two files, to make sure that the chromosome names follow the same format<br>https://usegalaxy.eu/root?tool_id=toolshed.g2.bx.psu.edu/repos/bgruening/text_processing/tp_tail_tool/1.1.0 |
| **Replace Text in a specific column** | **Performs find & replace operation** on a specified column in a given file.<br>For more complex patterns, use the ***awk*** tool. | **Parameters**<br>**In column:** i.e. 5<br>**Find pattern:** i.e. any symbol or letter expression „hello“<br>**Replace with:** text, or & (ampersand) and \\1 \\2 \\3<br>**Insert Replacement** (add new replacement for a new column) | The same text file in a *.gzip with replacement or replacements | For i.e. to convert the chromosome names, to change 20 and 21 to X and Y<br>*https://usegalaxy.eu/root?tool_id=toolshed.g2.bx.psu.edu/repos/bgruening/text_processing/tp_replace_in_column/1.1.3* |
| **Get** | **This tool finds the** | Every line should contain at least | **Result** | Adding promoter regions, i.e. to get regions 2kb bases upstream of the start of the gene to 10kb |

# #Introduction to Galaxy Analyses tools

## From peaks to genes tools

Goal to turn list of genomic regions into a list of possible target genes

| | | | | |
|---|---|---|---|---|
| **flanks** re turns flanking region/s for every gene | **upstream and/or downstream flanking region(s)** of all the selected regions in the input file. | 3 columns: Chromosome number, Start and Stop co-ordinates.<br>**Parameters**<br>**Region:** Whole feature\| Around start \| Around end<br>**Location of the flanking region/s:**<br>**Offset:**<br>Use positive values to offset co-ordinates in the direction of transcription and negative values to offset in the opposite direction**.**<br>**Length of the flanking region(s)**<br>Use non-negative value | BED format file with flanking regions for every gene | bases downstream of the start (12kb in length)<br>  <br>https://useqalaxy.eu/root?tool_id=toolshed.g2.bx.psu.edu/repos/devteam/get_flanks/get_flanks1/1.0.0 |

# #Introduction to Galaxy Analyses tools

## From peaks to genes tools

Goal to turn list of genomic regions into a list of possible target genes

# Formats from Peaks and Genes

GNU zip (gzip) –compressed archive format for Linux and Unix systems

Txt. – plain text file

**Interval** format is a Galaxy format for representing genomic intervals. It is tab-separated, but has the added requirement that three of the columns must be:

- chromosome ID
- start position (0-based)
- end position (end-exclusive)

An optional strand column can also be specified, and an initial header row can be used to label the columns, which do not have to be in any special order. Unlike BED format (see below) arbitrary additional columns can also be present.

The **BED - Browser Extensible Data** format provides a flexible way to encode gene regions. BED lines have three required fields:

- chromosome ID
- start position (0-based)
- end position (end-exclusive)

  There can be up to and nine additional optional fields, but the number of fields per line must be consistent throughout any single set of data.