# Using deepTools within Galaxy

We have a publicly available deepTools installation embedded within the Galaxy framework: **deeptools.ie-freiburg.mpg.de**

Galaxy is a tremendously useful platform developed by the Galaxy Team at Penn State. This platform is meant to offer access to a large variety of bioinformatics tools that **can be used without computer programming experiences**.

We have compiled several information that will hopefully get you started with your data analysis quickly. If you have never worked with Galaxy before, we recommend to start from the top.

If you just need a refresher of **how to upload data into Galaxy**, please have a look down below.

If you do not know the difference between a BAM and a BED file, that's fine - just make sure you have a look at this brief overview here before starting your analysis as high-throughput sequencing data relies on several specific **data formats**.

If you would like to dive right into the analysis of BAM or bigWig files using deepTools Galaxy, here is a **manual that covers the basic functions of deepTools**, starting from an overview of a typical workflow of NGS data analysis and ending with several different examples to demonstrate the power of heatmaps and summary plots.

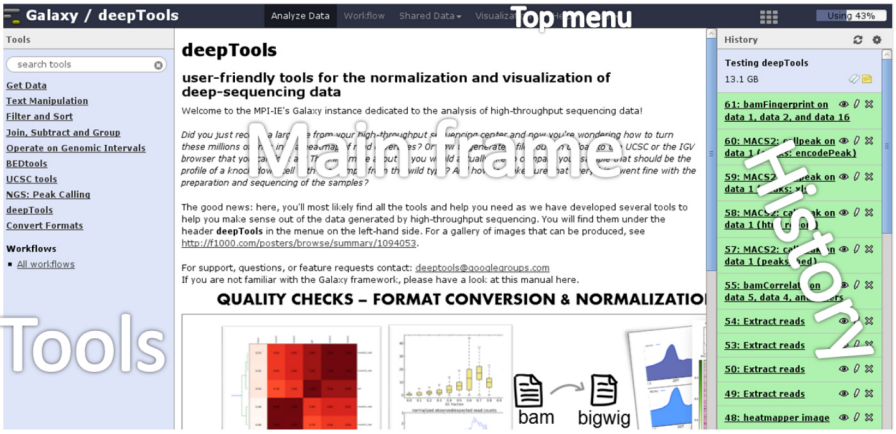
## Table of Content

# Basic features of Galaxy

The Galaxy team develops the platform, but since it is impossible to meet all bioinformatics needs (that can range from evolutionary analysis to mass spec data to high-throughput DNA sequencing (and way beyond)) with one single web server, many institutes have installed their own versions of Galaxy tuned to their specific needs. Our deepTools Galaxy is such a specialized server dedicated to the analysis of high-throughput DNA sequencing data. The overall makeup of this web server, however, is the same as for any other Galaxy installation, so if you've used Galaxy before, this section will probably not give you any new insights.

## The start site

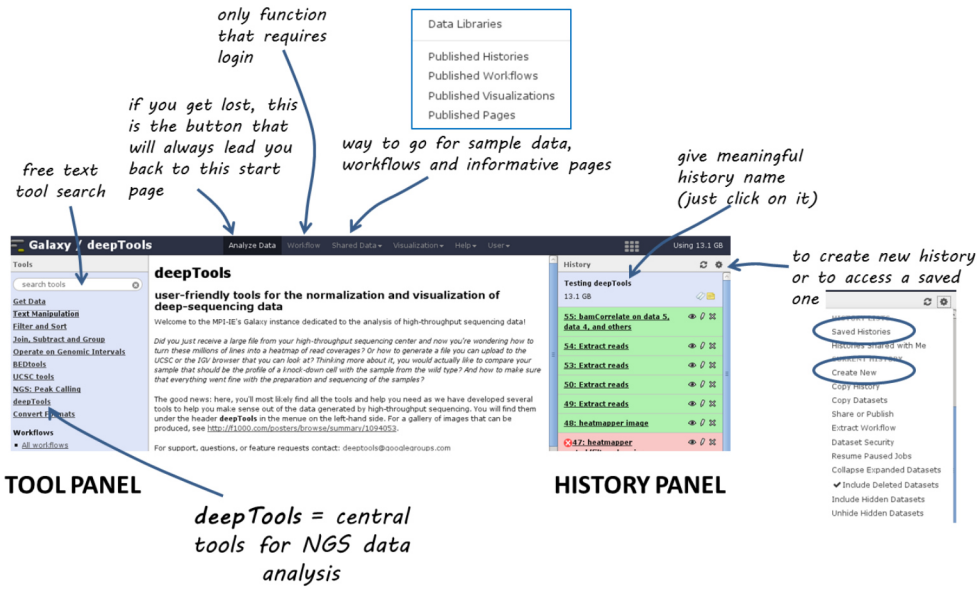
Here is a screenshot of what you should see at deeptools.ie-freiburg.mpg.de:




The start site contains 4 main features:

- **Top menu**: will lead you to other sections of Galaxy (away from the actual analysis part), such as workflows (registered users only) and content shared with you by other users such as sample data sets, pages and workflows
- **Tool panel** "What can be done": via this menu you can find all the *tools* installed in this Galaxy instance
- **Main frame** "What am I doing?": the center frame is your main working space where input will be required from you once you use a tool. In addition, you will always find general information about the tool here
- **History panel** "What did I do?": here you can find all *files* that one produces or uploads
    - the history is like a log book: everything you ever did is recorded here (unless you deleted things permanently)
    - histories can be shared with other users, they can also be downloaded
    - for each file that was produced, you will find all kinds of useful information such as the tool that was used to create the file, the tool's parameters etc.

Here's an annotated screenshot:

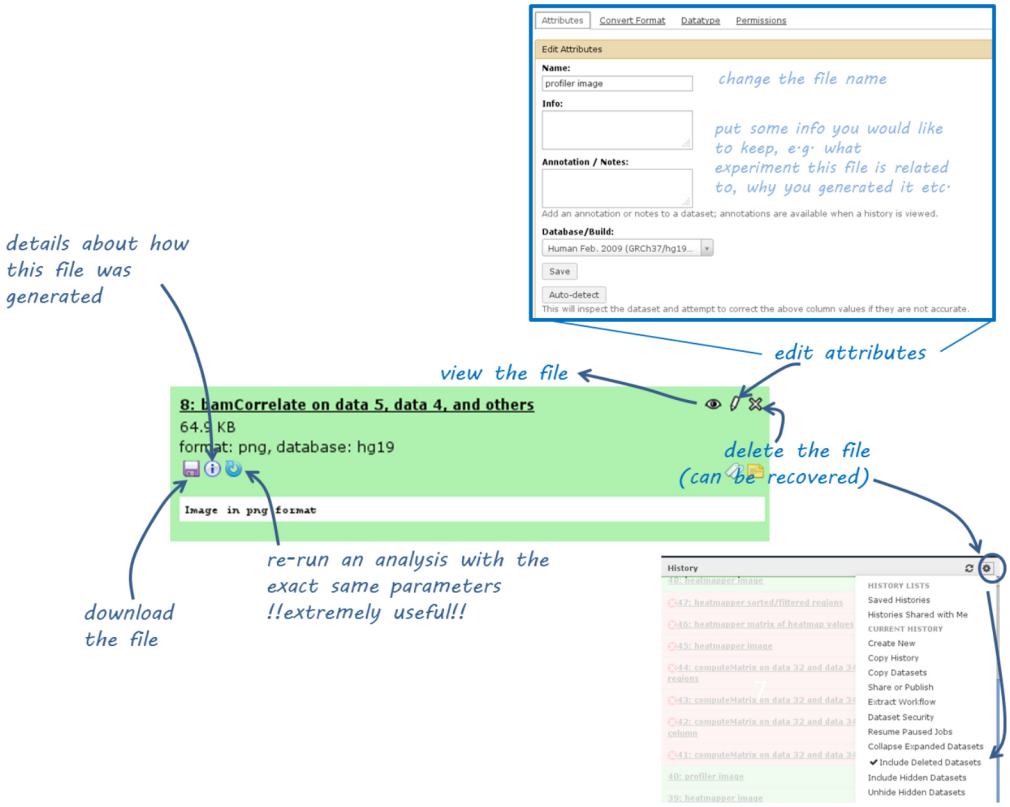



In the default state of the tool panel you see the **tool categories**, e.g. "Get Data". If you click on them, you will see the **individual tools** belonging to each category, e.g. "Upload File from your computer", "UCSC Main table browser" and "Biomart central server" in case you clicked on "Get Data". To use a tool such as "Upload File from your computer", just click on it.

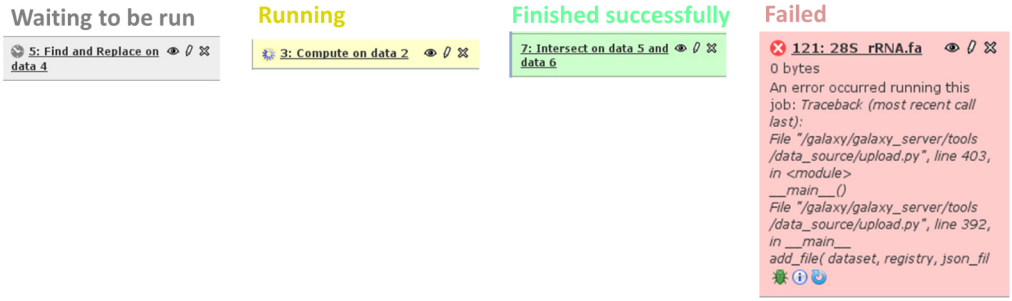
The tool search panel is extremely useful as it allows you to enter a key word (e.g. "BAM") that will lead to all the tools mentioning the key word in their tool name.

Once you've uploaded any kind of data, you will find the history on the right hand side filling up with green tiles. Each tile corresponds to one data set that you either uploaded or created. The data sets can be images, raw sequencing files, text files, tables, virtually anything. The content of each data set cannot be modified - everytime you want to change something *within* a data file (e.g. you would like to sort the values or add a line or cut a column), you will have to use a Galaxy tool that will lead to a *new* data set being produced. Every data set can be downloaded to your computer.

Have a look at the following screenshot to get a feeling for how many information Galaxy keeps for you (which makes it very feasible to reproduce any given analysis):




Each data set can have 4 different states that are intuitively color-coded:




If you encounter a failure after you've run a tool, please follow those steps (in this order):

1. click on the center button on the lower left corner of the failed data set ("i"): now check whether you chose the **correct data files**
2. if you're sure that you chose the correct files, hit the re-run button (blue arrow in the lower left corner) - check again whether your files had the **correct file format**
    - if you suspect that the format might be incorrectly assigned (e.g. a file that should be a bed-file is labelled as a tabular file), click the edit button of the input data file - ther you can change the corresponding attributes
3. if you've checked your input data and the error is persisting, click on the green bug (lower left corner of the failed data set) and send the **bug report** to us.

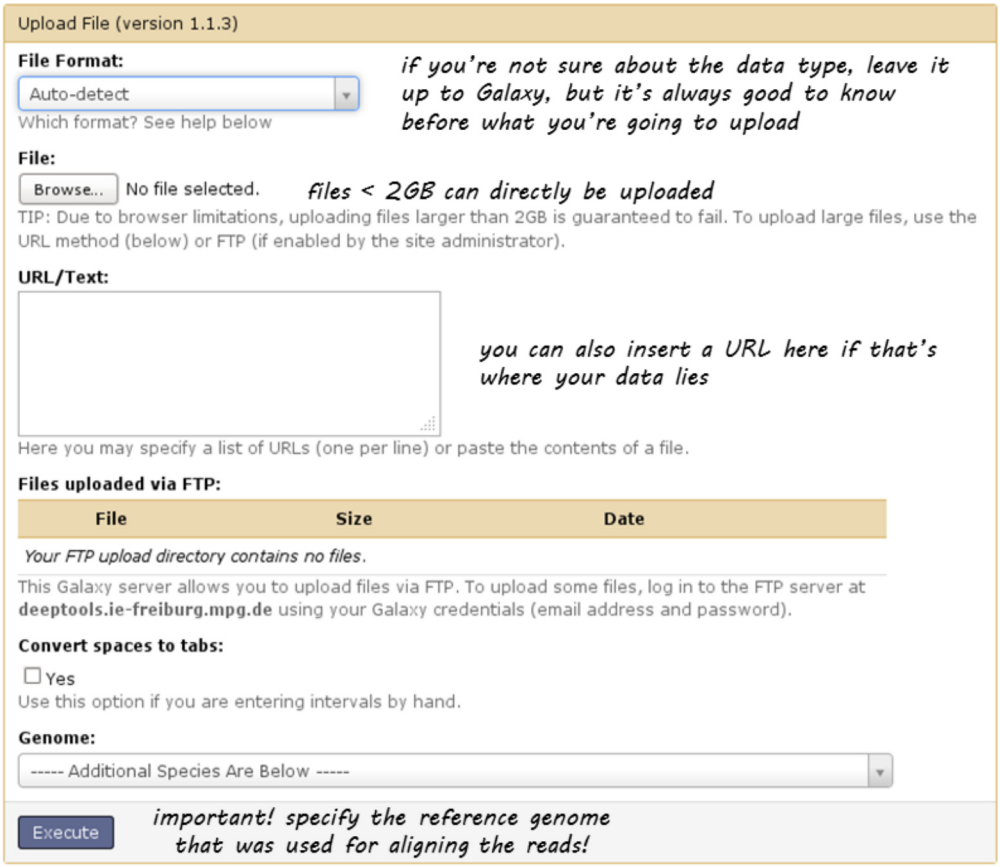
# Data import into Galaxy

There are three main ways to populate your Galaxy history with data files:

1. Data upload from your computer
2. Import a shared data set from the Galaxy data library
3. Download annotation data from public servers
4. **For registered users only**: Copy data sets between histories

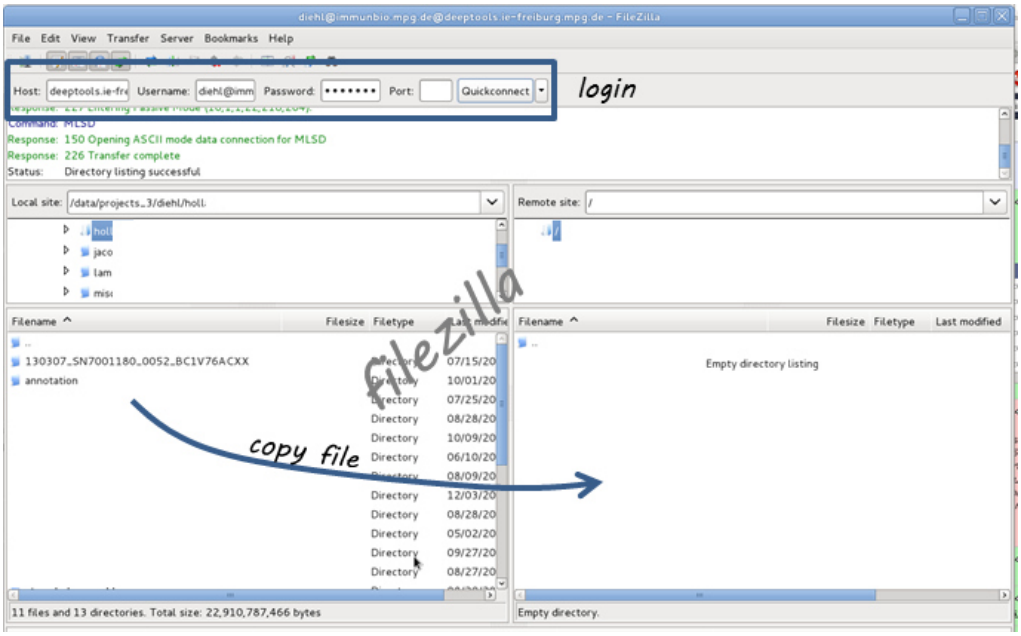
### Upload files from your computer

The data upload of files <2 GB that lie on your computer is fairly straight-forward: click on the category "Get data" and choose the tool "Upload file". Then select the file via the "Browse" button.




For files >2GB there's the option to upload via an FTP server. If your data is available via an URL that links to an FTP server, you can simply paste the URL in the empty text box.

If you do not have access to an FTP server, you can directly upload to our Galaxy's FTP. * first register with deeptools.ie-freiburg.mpg.de (via “User” --> “register”; registration requires an email address and is free of charge) * You will also need an FTP client, e.g. filezilla. * Then login to the **FTP client** using your **deepTools Galaxy user name and password** (host: deeptools.ie-freiburg.mpg.de). Down below you see a screenshot of how that looks like with filezilla. * Copy the file you wish to upload to the remote site (in filezilla, you can simply drag the file to the window on the right hand side) * Go back to deepTools Galaxy * Click on the tool "Upload file" (--> "Files uploaded via FTP") - here the files you just copied over via filezilla should appear. Select the files you want and hit “execute”. They will be moved from the FTP server to your history.


diehl@immunbio.mpg.de@deeptools.ie-freiburg.mpg.de - FileZilla
File Edit View Transfer Server Bookmarks Help
Host: deeptools.ie-fre Username: diehl@imm Password: ••••••• Port: Quickconnect
login
Response: 150 Opening ASCII mode data connection for MLSD
Response: 226 Transfer complete
Status: Directory listing successful
Local site: /data/projects_3/diehl/holl
Remote site: /
Filename Filesize Filetype Last modified
130307_SN7001180_0052_BC1V76ACXX
annotation
Empty directory listing
FileZilla
copy file
11 files and 13 directories. Total size: 22,910,787,466 bytes
Empty directory.

### Import data sets from the Galaxy data library

If you would like to play around with sample data, you can import files that we have saved within the general data storage of the deepTools Galaxy server. Everyone can import them into his or her own history, they will not contribute to the user's disk quota.

You can reach the data library via "Shared Data" in the top menu, then select "Data Libraries".

Within the Data Library you will find a folder called "Sample Data" that contains data that we downloaded from the Roadmap project and UCSC. More precisely, we donwloaded the FASTQ files and mapped the reads to the human reference genome (version hg19) to obtain the BAM files you see. In addition, you will find signal tracks of DNase-seq data from UCSC, bigWig files with GC content for flies and mice and some annotation files.

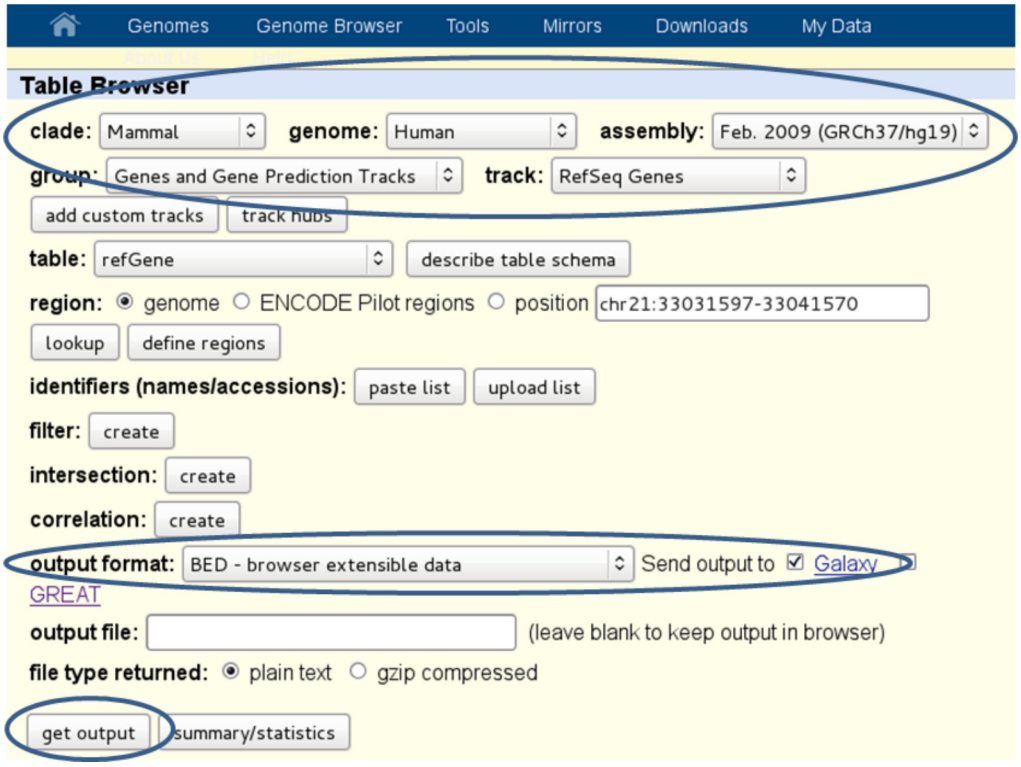
### Download annotation files from public data bases

In many cases you will want to query your sequencing data results for known genome annotation, such as genes, exons, transcription start sites etc. These information can be obtained via the two main sources of genome annotation, UCSC and BioMart. Please note that UCSC and BioMart will cater to different ways of genome annotation, i.e. genes defined in UCSC might not correspond to the same regions in a gene file downloaded from BioMart. (For a brief overview over the issues of genome annotation, you can check out Wikipedia, if you'd always wanted to know much more about those issues, this might be a good start.)

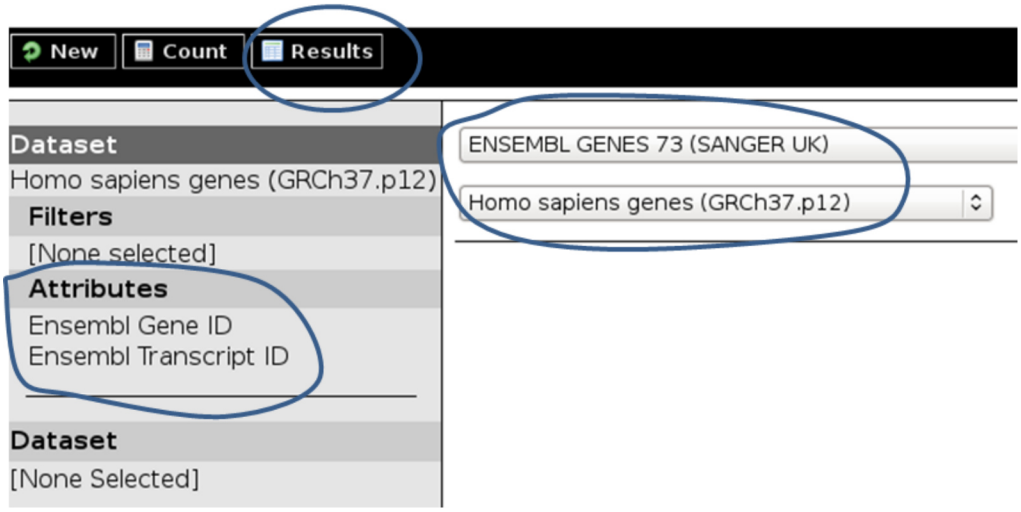
You can access the data stored at UCSC or BioMart conveniently through our Galaxy instance which will import the resulting files into your history. Just go to **"Get data"** --> "UCSC" or "BioMart".

The majority of annotation files will probably be in BED format, however, you can also find other data sets. UCSC, for example, offers a wide range of data that you can browse via the "group" and "track" menus (for example, you could download the GC content of the genome as a signal file from UCSC via the "group" menu ("Mapping and Sequencing Tracks"). Note, however, that the download through this interface is limited to 100,000 lines per file which might not be sufficient for some mammalian data sets).

Here's a screenshot from downloading a BED-file of all RefSeq genes defined for the human genome (version hg19):



And here's how you would do it for the BioMart approach:




Per default, **BioMart will not output a BED file** like UCSC does. It is therefore important that you make sure you get all the information you need (most likely: chromosome, gene start, gene end, ID, strand) via the "Attributes" section. You can click on the "Results" button at any time to check the format of the table that will be sent to Galaxy (Note that the strand information will be decoded as 1 for "forward" or "plus" strand and -1 for "reverse" or "minus" strand.)

> Be aware, that BED files from UCSC will have chromosomes labelled with “chr” while ENSEMBL usually returns just the number – this might lead to incompatibilities, i.e. when working with annotations from UCSC and ENSEMBL, you need to make sure to use the same naming!

### For registered users only: Copy data sets between histories

In case you have registered with deepTools Galaxy you can have more than one history. In order to minimize the disk space you're occupying we strongly suggest to **copy** data sets between histories when you're using the same data set in different histories. This can easily be done via the History panel's option button --> "Copy dataset". In the main frame, you should now be able to select the history you would like to copy from on the left hand side and the target history on the right hand side.

# Which tools can I find in the deepTools Galaxy?

As mentioned above, each Galaxy installation can be tuned to the individual interests. Our goal is to provide a Galaxy that enables you to **quality check, process and normalize and subsequently visualize your data obtained by high-throughput DNA sequencing**.

We provide the following kinds of tools:

1. deepTools - NGS data handling
2. peak calling (ChIP-seq specific)
3. operating on genomic intervals
4. working with text files and tables

### deepTools

The most important category is called **"deepTools"** that contains 8 major tools:

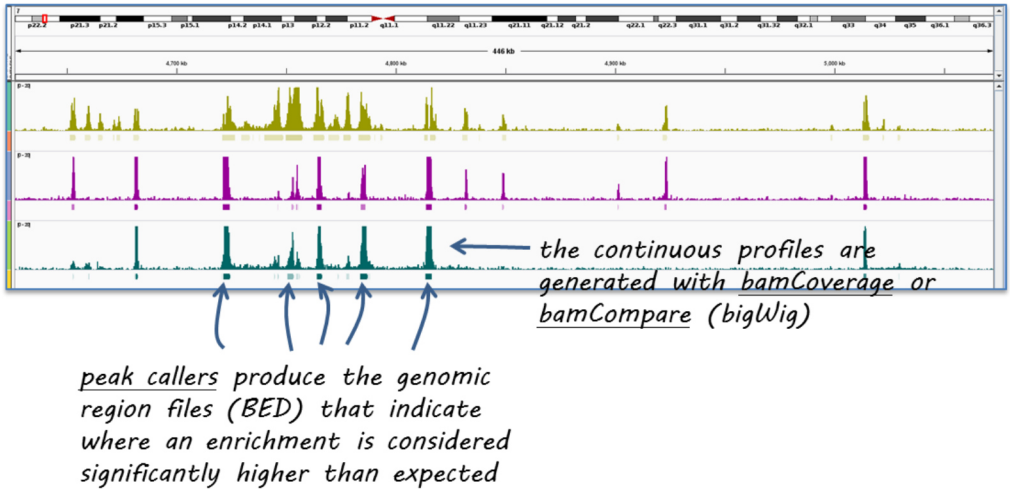
| tool | type | input files | main output file(s) | application |
|---|---|---|---|---|
| bamCorrelate | QC | 2 or more BAM | clustered heatmap | Pearson or Spearman correlation between read distributions |
| bamFingerprint | QC | 2 BAM | 1 diagnostic plot | assess enrichment strength of a ChIP sample |
| computeGCBias | QC | 1 BAM | 2 diagnostic plots | calculate the exp. and obs. GC distribution of reads |
| bamCoverage | normalization | BAM | bedGraph or bigWig | obtain the normalized read coverage of a single BAM file |
| bamCompare | normalization | 2 BAM | bedGraph or bigWig | normalize 2 BAM files to each other using a mathematical operation of your choice (e.g. log2ratio, difference) |
| computeMatrix | visualization | 1 bigWig, 1 BED | zipped file, to be used with heatmapper or profiler | compute the values needed for heatmaps and summary plots |
| heatmapper | visualization | computeMatrix output | heatmap of read coverages | visualize the read coverages for genomic regions |
| profiler | visualization | computeMatrix output | summary plot ("meta-profile") | visualize the average read coverages over a group of genomic regions |

1. General overview of how we use deep Tools; here is the pdf version of this overview
2. Each individual tool is described in more detail on separate pages - just follow the links in the table above
3. For each tool, you will find specific explanations within the deepTools Galaxy main frame, too.

### Peak calling

In ChIP-seq analysis, peak calling algorithms are essential downstream analysis tools to identify regions of significant enrichments (i.e. where the ChIP sample contained significantly more sequenced reads than the input control sample). By now, there must be close to 100 programs out there (see Wilbanks et al. for a comparison of peak calling programs).

In contrast to deepTools that were developed for handling and generating *continuous* genome-wide profiles, peak calling will result in a *list of genomic regions*. Have a look at the screenshot to understand the difference.




We have included the peak callers MACS, SICER and CCAT within our Galaxy instance with MACS being the most popular peak calling algorithm for the identification of localized transcription factor binding sites while SICER was developed for diffuse ChIP-seq signals.

### Working with genomic intervals

Galaxy has 2 file formats to store lists of genomic regions:

- INTERVAL
    - tab-separated
    - requirements:
        1. Column: chromosome
        2. Column: start position
        3. Column: end position
        4. all other columns can contain any value or character
- BED
    - very similar to INTERVAL, but stricter when it comes to what is expected to be kept in which column:
        - 1. to 3. Column: same as interval
        - Column 4: name
        - Column 5: score
        - Column 6: strand

In case you would like to work with several lists of genomic regions, e.g. generate a new list of regions that are found in two different files etc., there are two categories of tools dedicated to performing these tasks: * Operate on genomic intervals * BEDtools

Each tool's function is explained within Galaxy. Do browse those tools as they will give you a very good glimpse of the scope of possible analyses!

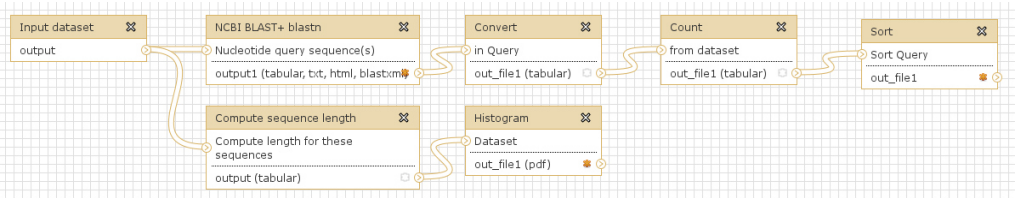
### Working with text files and tables

In addition to deepTools that were specifically developed for the handling of NGS data, we have incorporated several standard Galaxy tools that enable you to manipulate tab-separated files such as gene lists, peak lists, data matrices etc.

There are 3 main categories: * **Text manipulation** * unlike Excel where you can easily interact with your text and tables via the mouse, data manipulations within Galaxy are strictly based on commands. If you feel like you would like to do something to certain *columns* of a data set, go through the tools of this category * e.g. adding columns, cutting columns, pasting two files side by side, selecting random lines etc. * **Filter and Sort** * in addition to the common sorting and filtering, there's the very useful toool to __select lines that match an expression" * **Join, Subtract, Group** * this category is very useful if you have several data sets that you would like to work with, e.g. by comparing them

# Workflows

Workflows are Galaxy's equivalent of protocols. This is an extremely useful feature as it allows users to share their protocols and bioinformatic analyses in a very easy and transparent way. This is the graphical representation of a Galaxy workflow that can easily be modified via drag'n'drop within the workflows manual (you must be registered with deepTools Galaxy to be able to generate your own workflows).




# Where to get help?

Please check our deepTools Galaxy FAQs

- general Galaxy help: wiki.galaxyproject.org/Learn
- specific help with deepTools Galaxy: deeptools@googlegroups.com
- if you encounter a failing data set, please send a bug report via Galaxy and we will get in touch

This tool is developed by the Bioinformatics Facility at the Max Planck Institute for Immunobiology and Epigenetics, Freiburg.