Analysis of ChIP-seq data in Galaxy

November, 2012

Local copy: https://galaxy.wi.mit.edu/

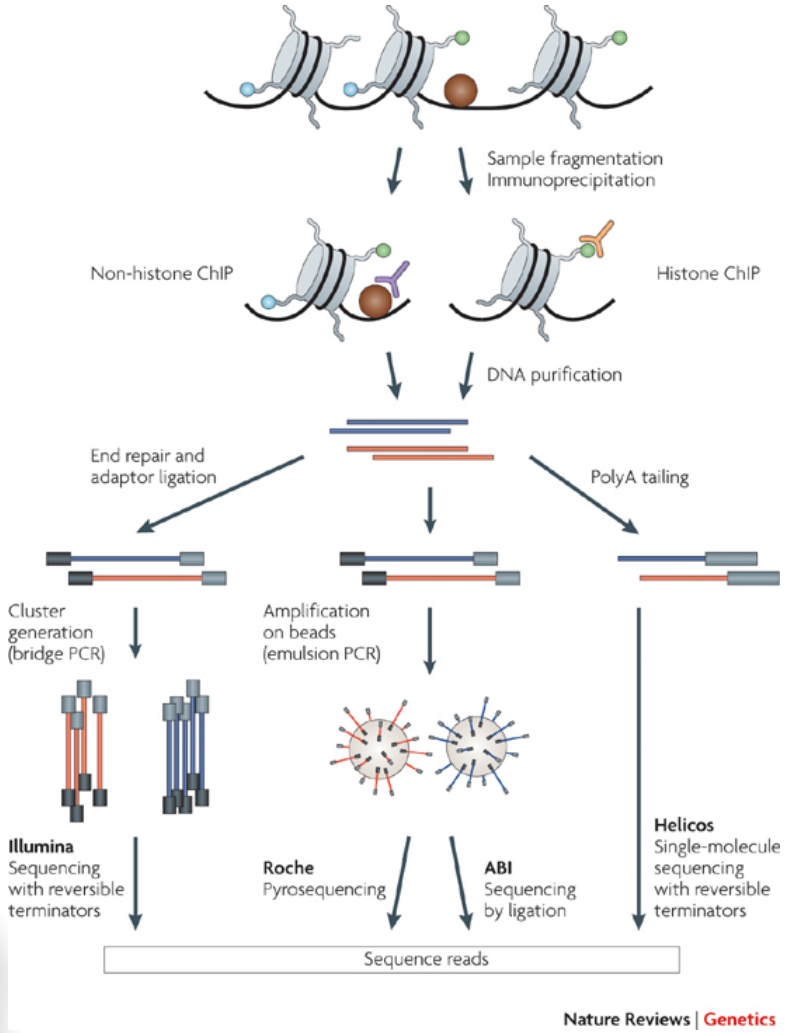
Joint project between BaRC and IT

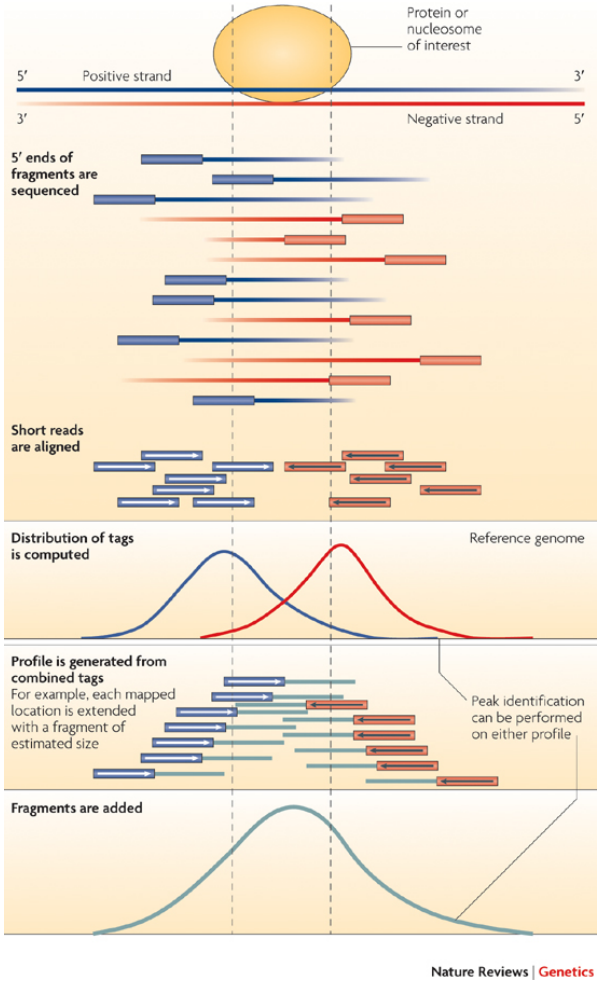
Main site: http://main.g2.bx.psu.edu/

Font Conventions

- **Bold and blue** refers to tools on the left hand window
- **Bold and green** refers to tabs and menus on the top (Analyze data, Shared Data, etc)
- Slides with Red Headers describe the hands-on exercises
- *Red and italic* refers to menus and history names used on the hands-on

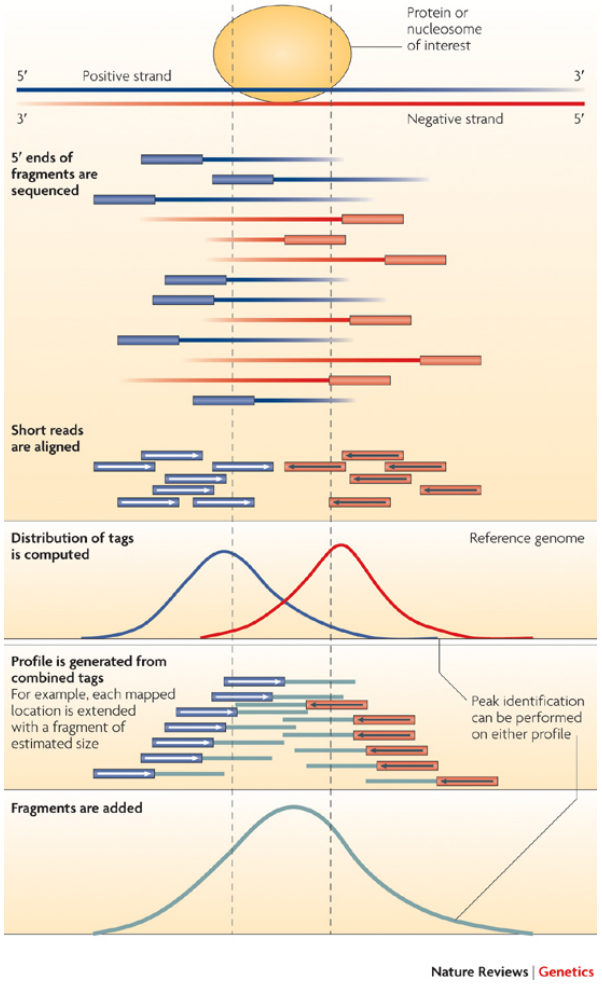
ChIP-seq





Park, P. J., *ChIP-seq: advantages and challenges of a maturing technology, Nat Rev Genet. Oct;10(10):669-80 (2009)*

General workflow for ChIP-seq analysis



Fastq files from the sequencing facility

↓

Check the quality of your reads
(NGS: QC and manipulation -> FastQC)

↓

Step 1: Map the reads to the genome
(BOWTIE)

↓

Step 2: Identify peaks
(MACS)

↓

Step 3: Post processing: Annotate peaks
i.e. find genes overlapping or close to the peaks

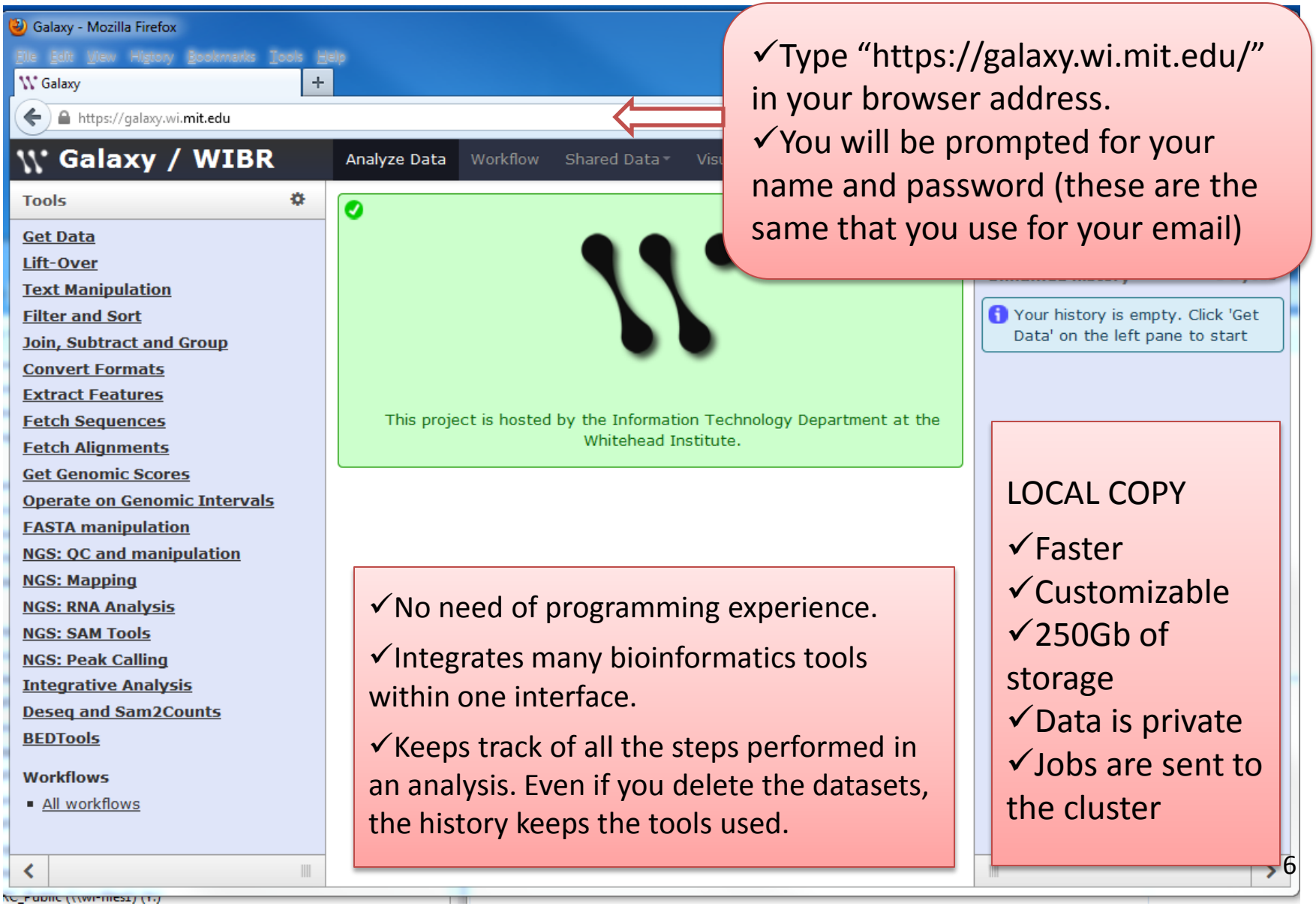
Park, P. J., *ChIP-seq: advantages and challenges of a maturing technology, Nat Rev Genet. Oct;10(10):669-80 (2009)*

Hands-On Exercises

- Data upload (get files needed for analysis)
  - Raw data: fastq files (ChIP and WCE)
  - Intermediate files: output files of the first analysis steps
  - Annotation files: genes and upstream regions, we will use them to get a set of genes that overlap or are close to the peaks
- ChIP-seq analysis.
  - Map with bowtie
  - Identify peaks bound with MACS
  - Find genes that overlap or are close to the peaks

The Galaxy Interface

A web based platform for analysis of large genomic datasets

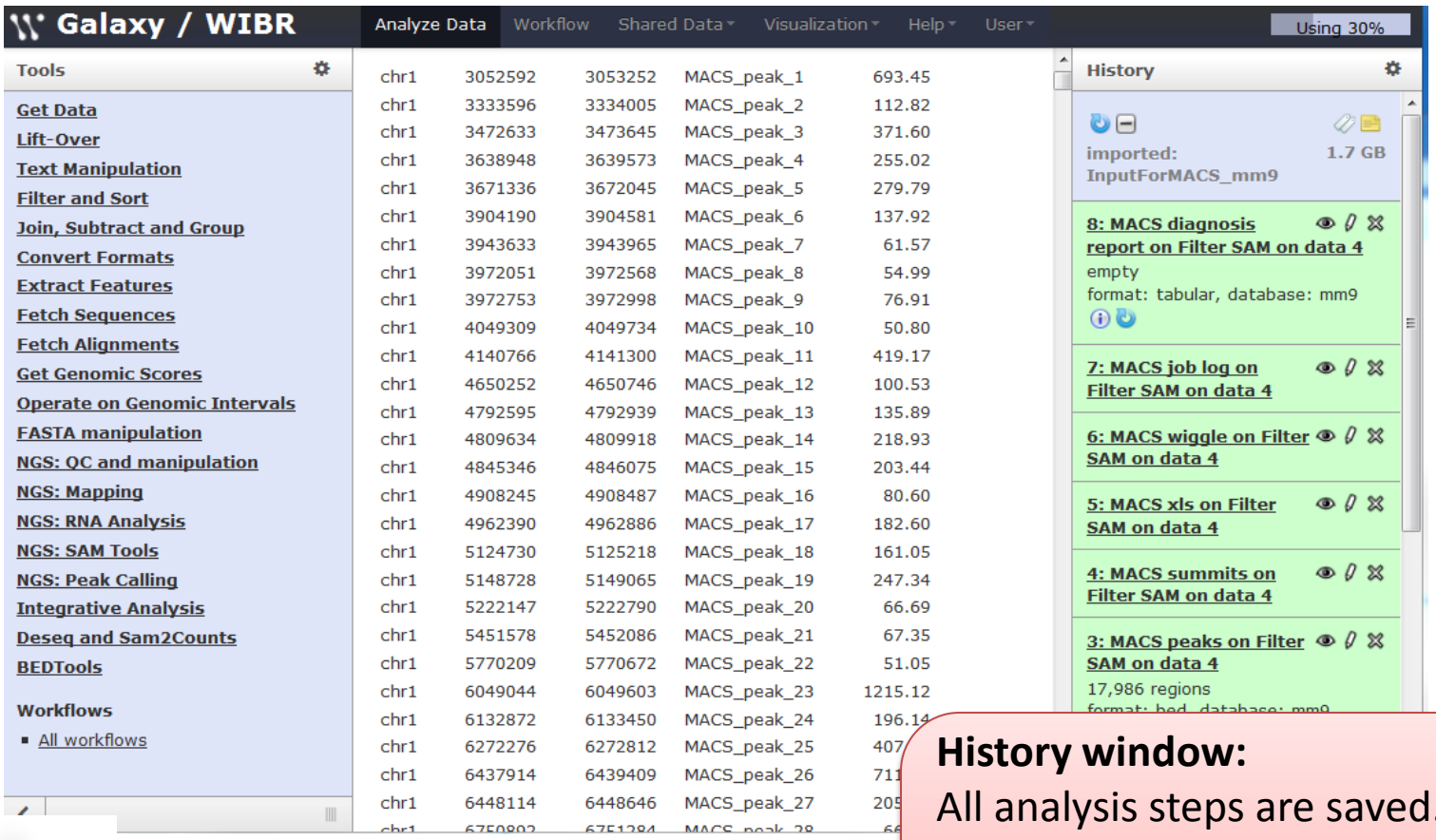


✓Type “https://galaxy.wi.mit.edu/” in your browser address.
✓You will be prompted for your name and password (these are the same that you use for your email)

✓No need of programming experience.
✓Integrates many bioinformatics tools within one interface.
✓Keeps track of all the steps performed in an analysis. Even if you delete the datasets, the history keeps the tools used.

LOCAL COPY
✓Faster
✓Customizable
✓250Gb of storage
✓Data is private
✓Jobs are sent to the cluster

Galaxy Interface: Analyze Data



Processed data

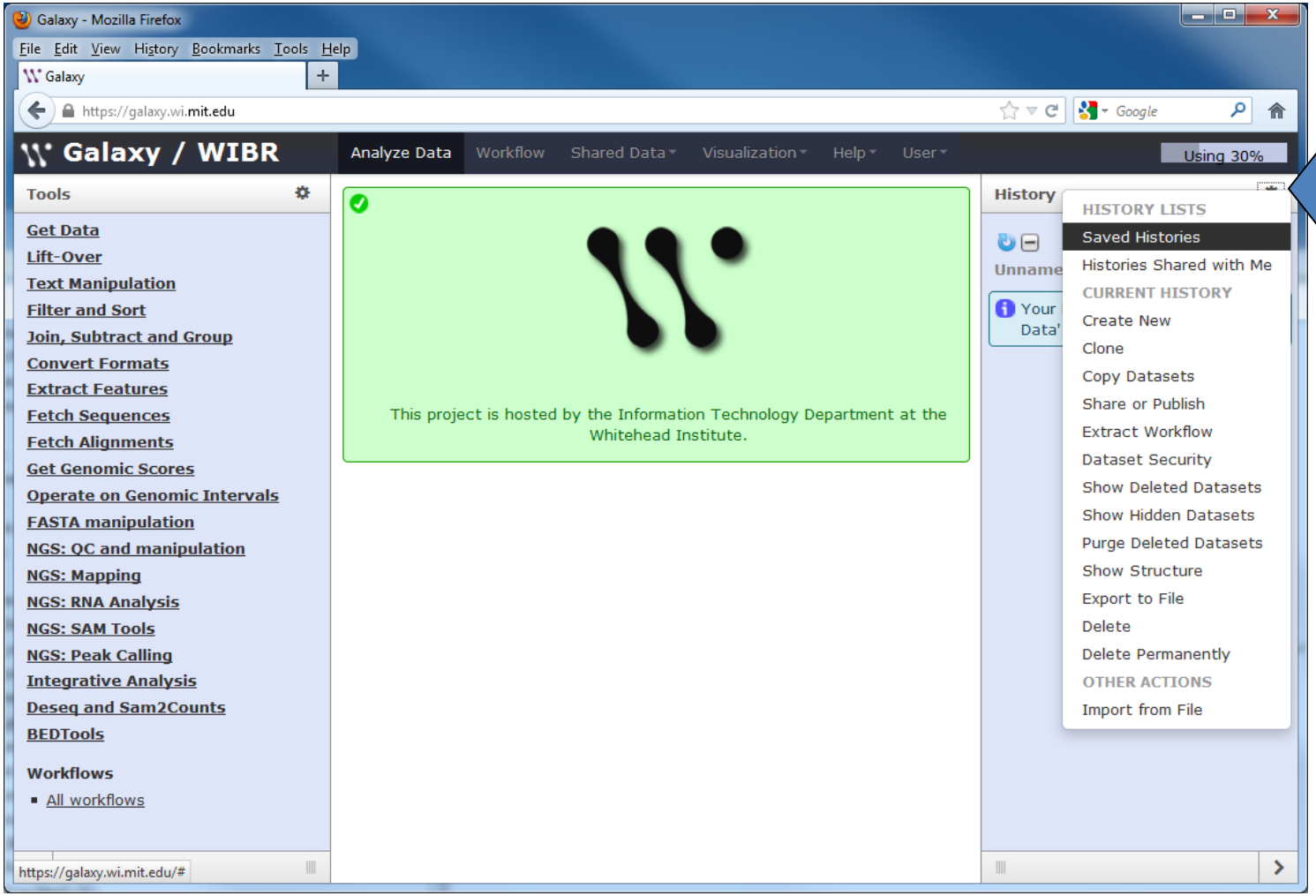
Green:
job is finished

Yellow:
job is running
Gray:

job is in queue
Red: there is a problem

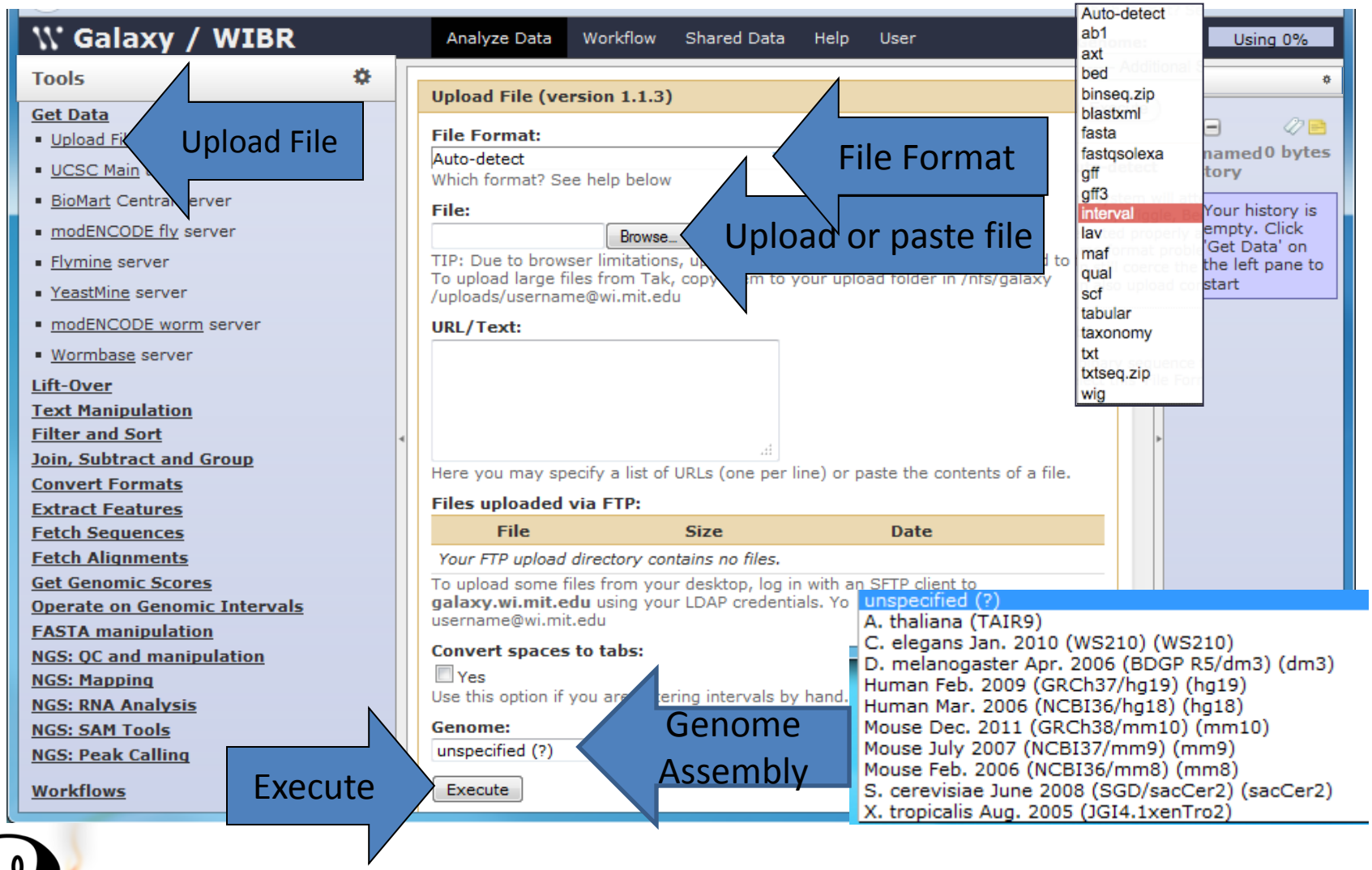
History window:
All analysis steps are saved.
Data is not overwritten.
Can create workflow to repeat an analysis.

How to find your previous histories



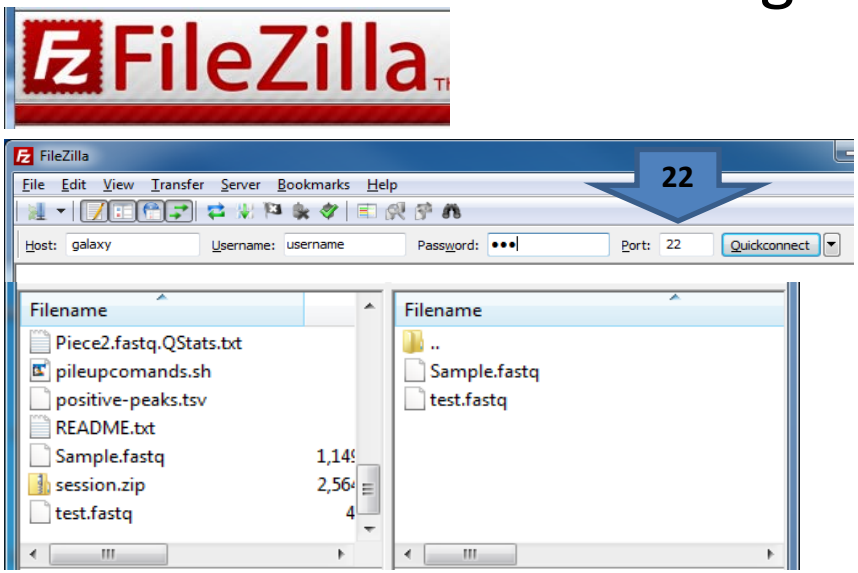
History menu

Getting Data: Upload File

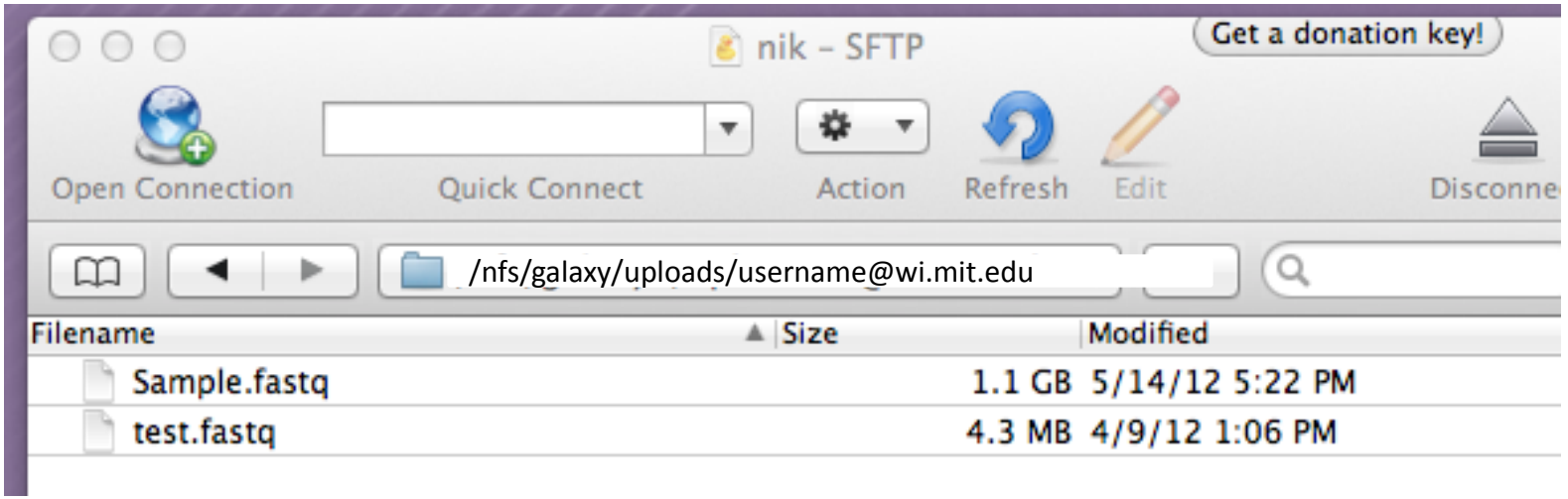


Getting Data: Uploading Large Files

Step 1: copy your file to /nfs/galaxy/uploads/username@wi.mit.edu using a sftp client

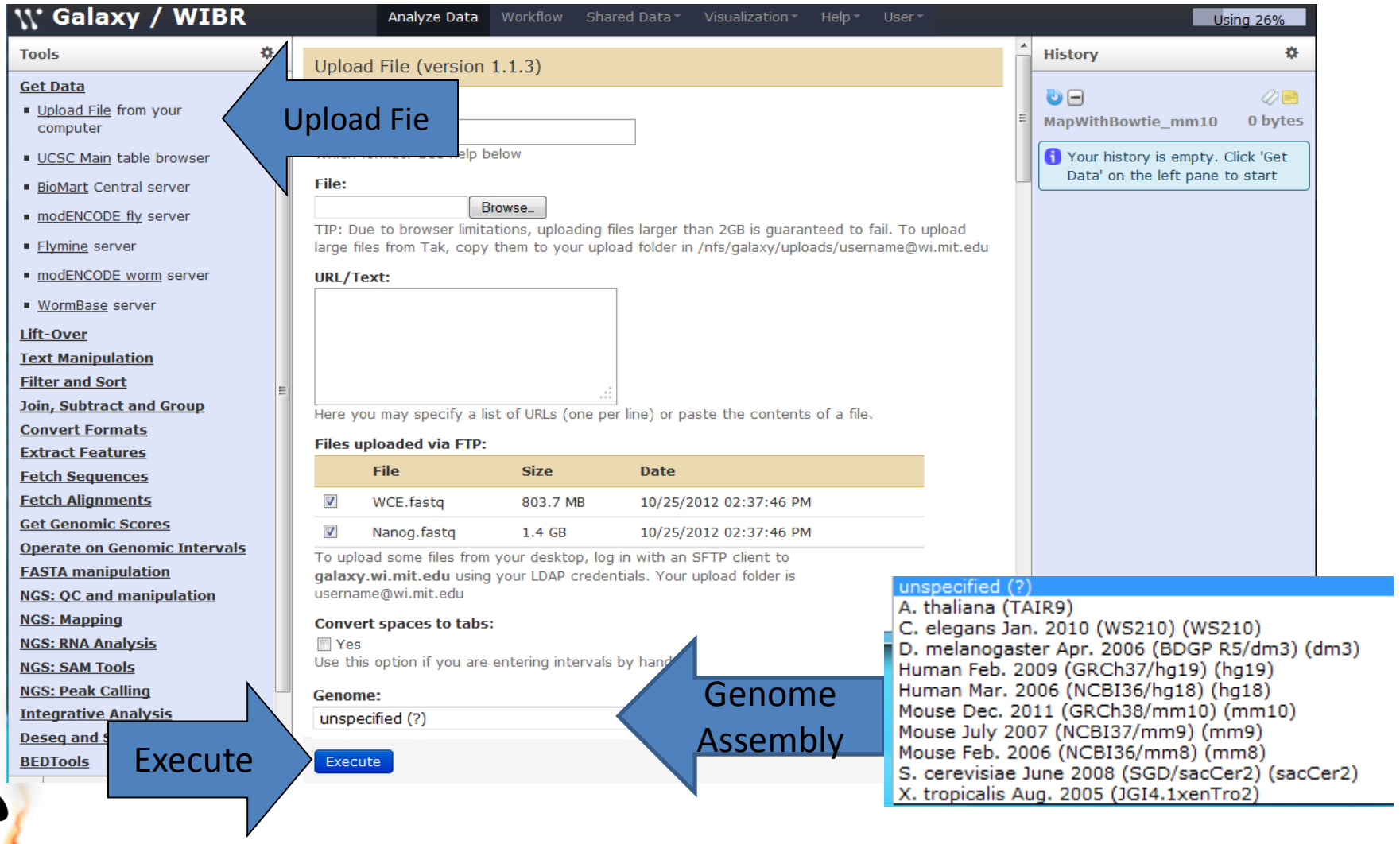


CyberDuck



Getting Data: Uploading Large Files

Step 2: Select and upload the file within galaxy



Hands-on: Data Upload

This is an schematic of the data we need to upload for each step.

Step 1. Map reads

History: mapWithBowtie

Input files:

WCE.fastq

Nanog.fastq

Step 2. Call peaks

History: InputForMACS_mm9

Input files:

Filter SAM on data 3_WCE

Filter SAM on data 4_Nanog

Step 3. Post processing

History: InputFor_annotatePeaks

Input files:

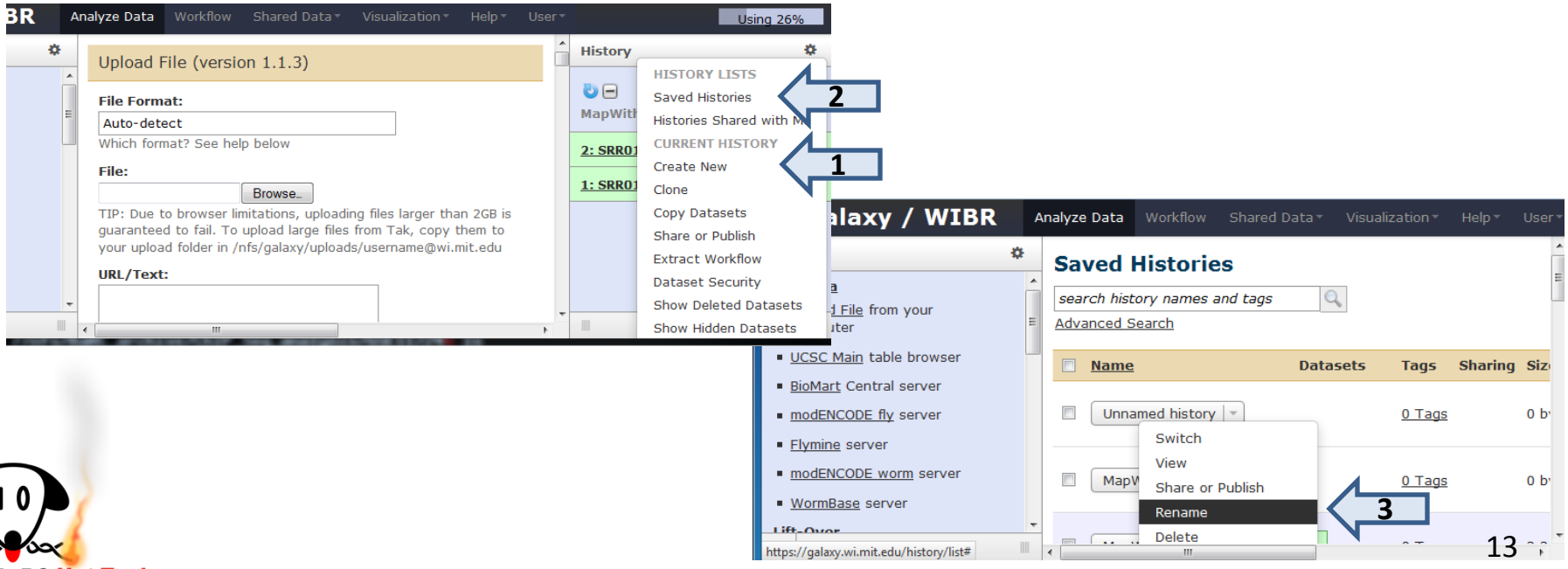
Peaks

Refseq Genes

3Kb Upstream of Refseq Genes

Hands-on: Data Upload

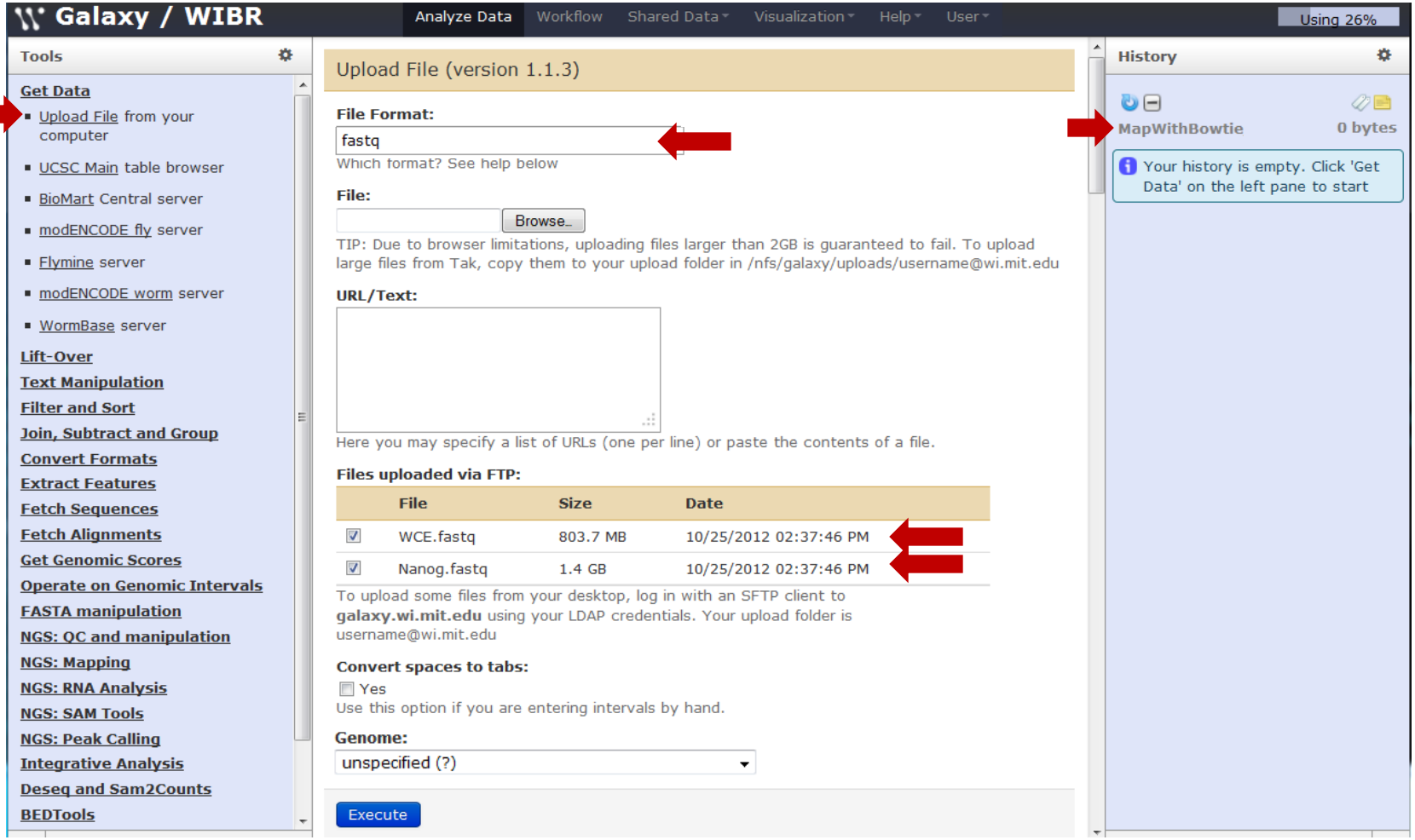
- Create a new history and name it "*mapWithBowtie*"
  1) On the history MENU select *Create New*
  2) On the history MENU select *Saved Histories*
  3) Once you see your histories on the middle window click on the "*Unnamed history*" drop down menu and select *Rename*



Hands-on: Data Upload

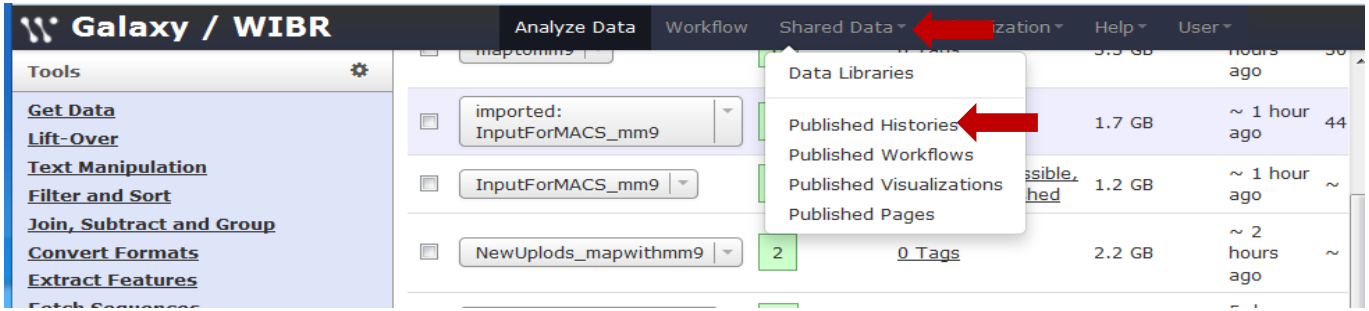
stay in "*mapWithBowtie*" history

Upload the files that I have copied for to your uploads directory

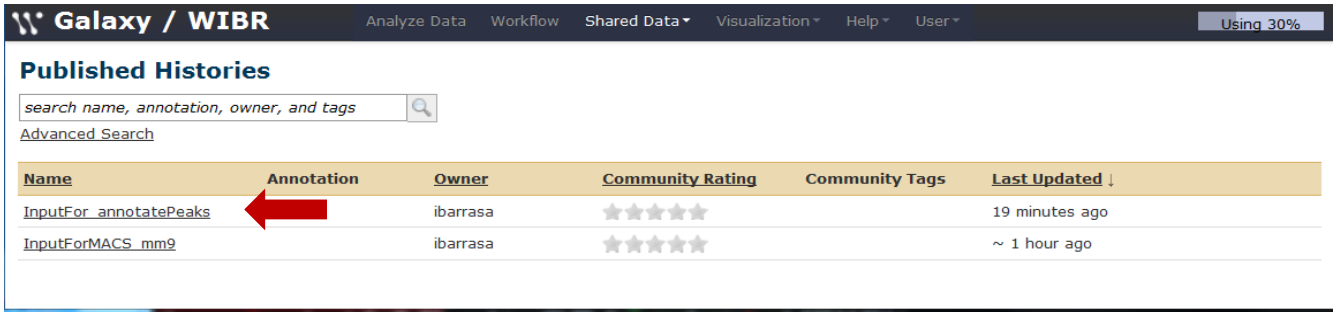


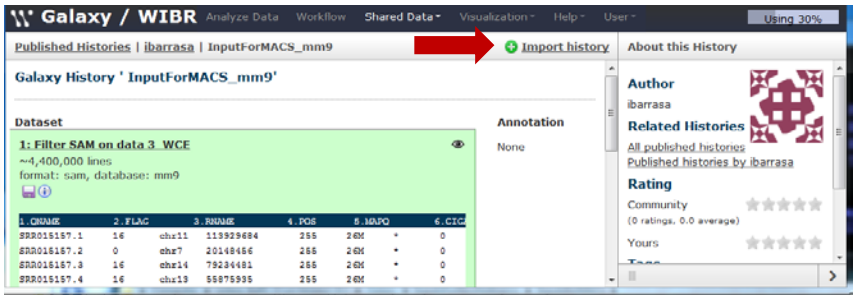
Hands-on: Data Upload

- Click on the **Shared Data** Tab and select **Published Histories**



- Select *InputForMACS_mm9*

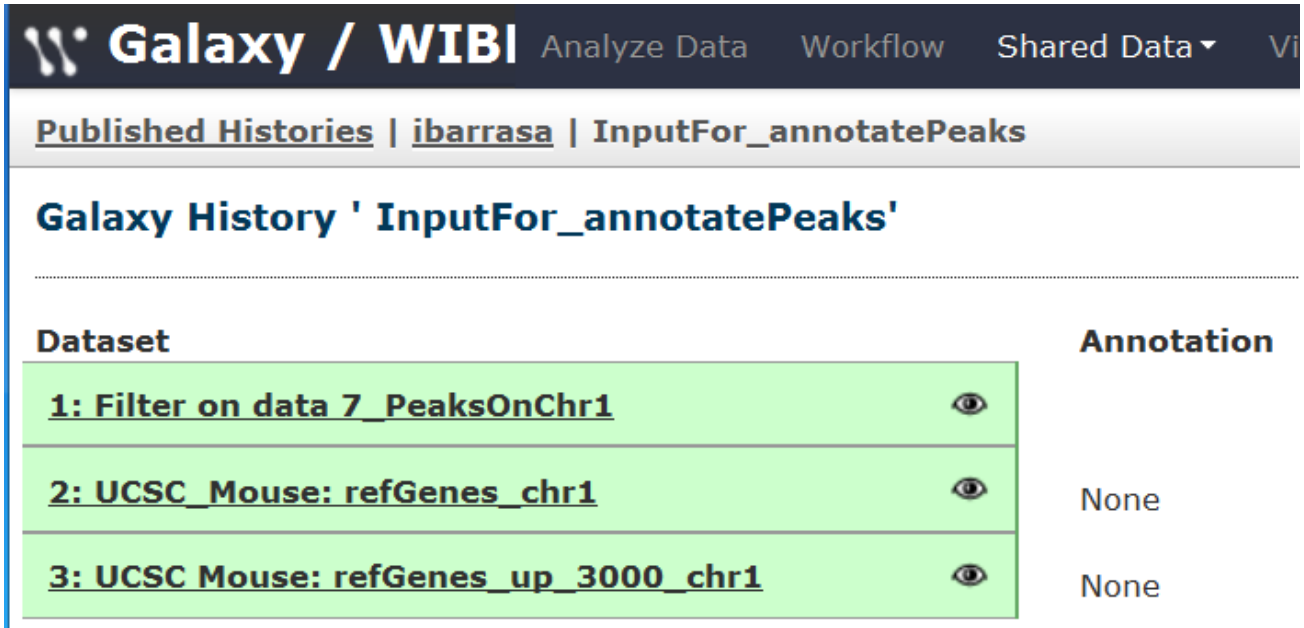




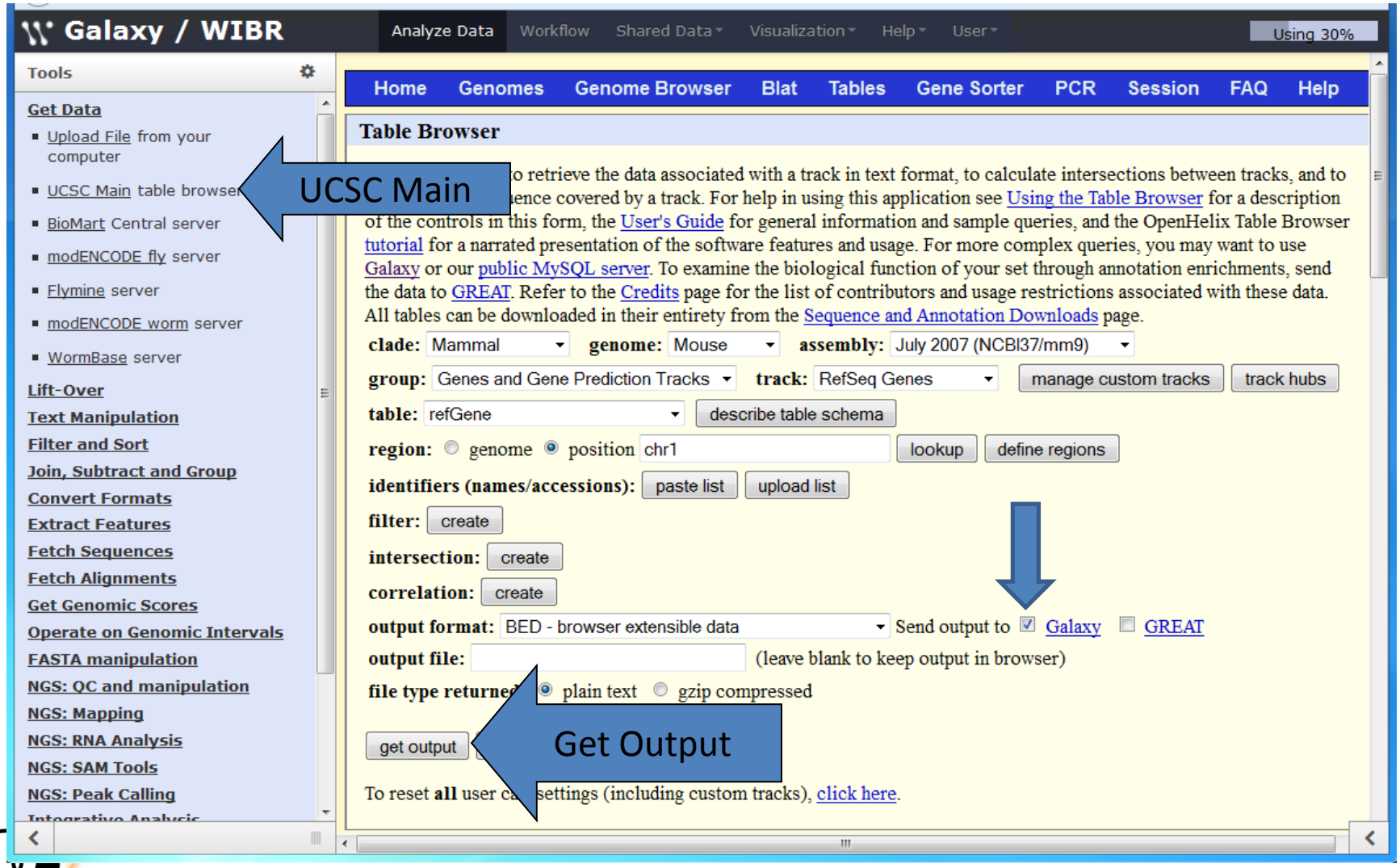
Click on **Import history**

Hands-on: Data Upload

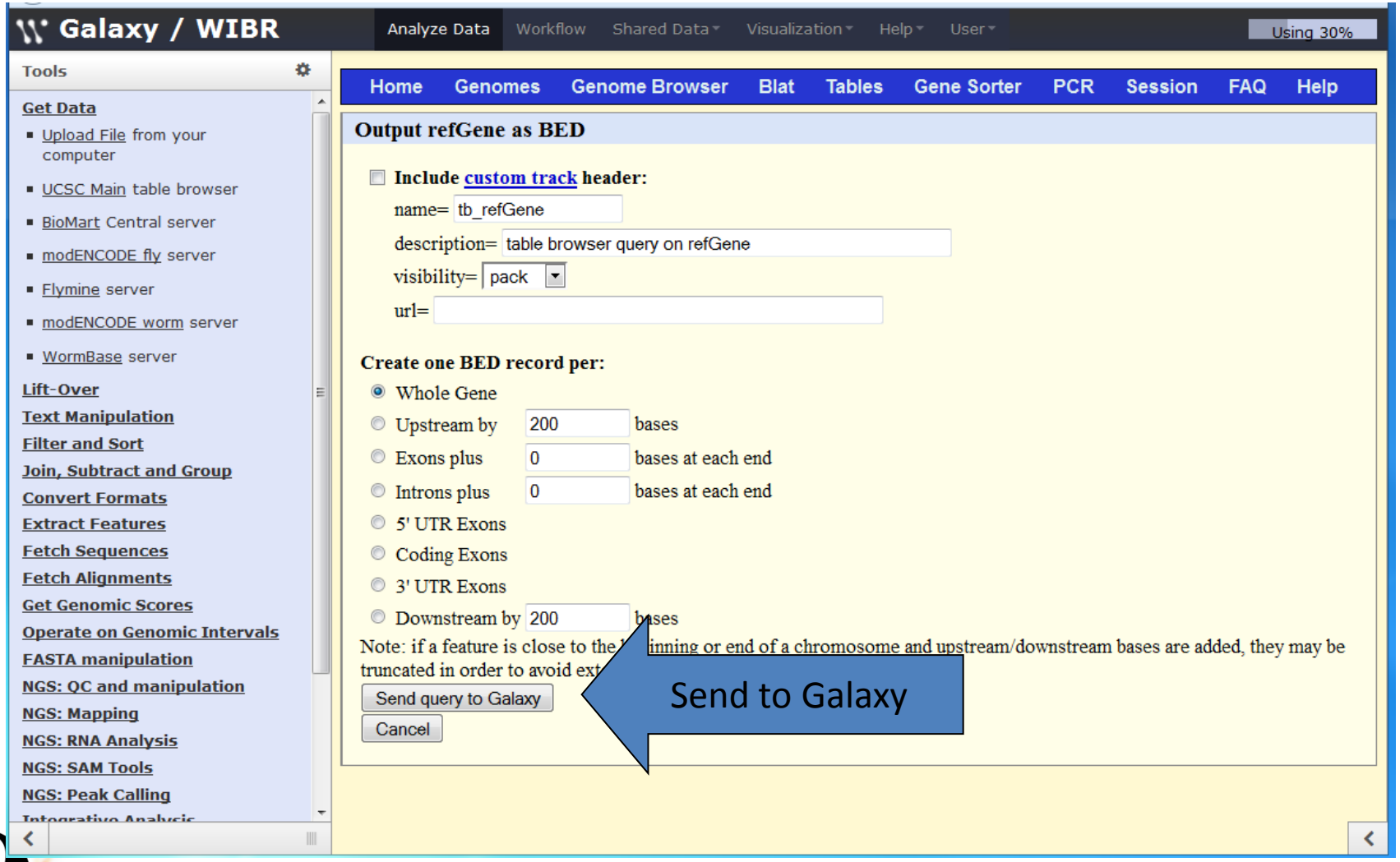
- Click on the **Shared Data** Tab and select **Published Histories**
- Select *InputFor_annotatePeaks*
- Click on **Import history**



Getting Data from UCSC (local copy)



Getting Data from UCSC (local copy)



Hands-on: Data Upload

Optional

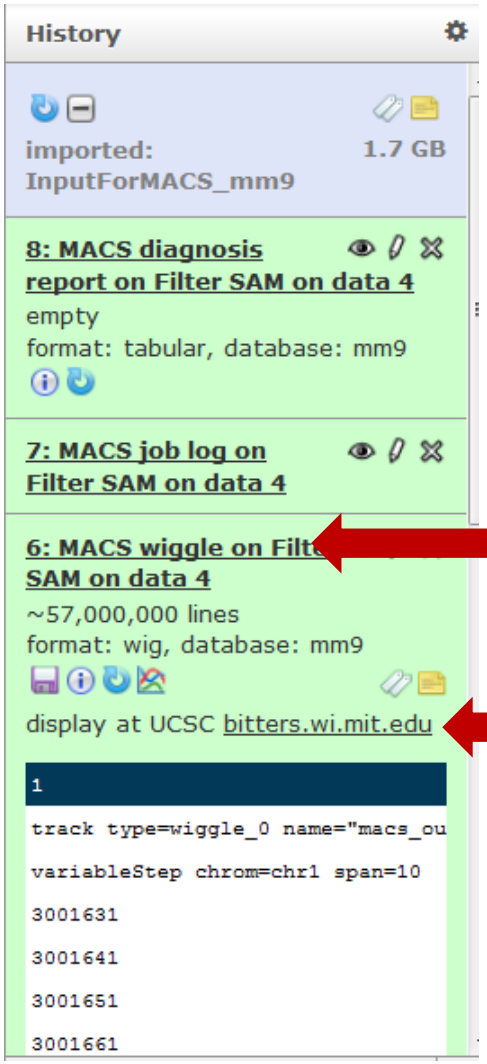
stay in "*imported: InputFor_annotatePeaks* " history

Use the link to the UCSC main table browser

1. Get all mouse refseq genes mm9 chr1
2. Get all 3Kb upstream regions from mouse refseq genes mm9 chr1

Now you have all the data we need for the hands-on exercises

Important Icons



Display data

Download

Edit attributes

View Details

Delete

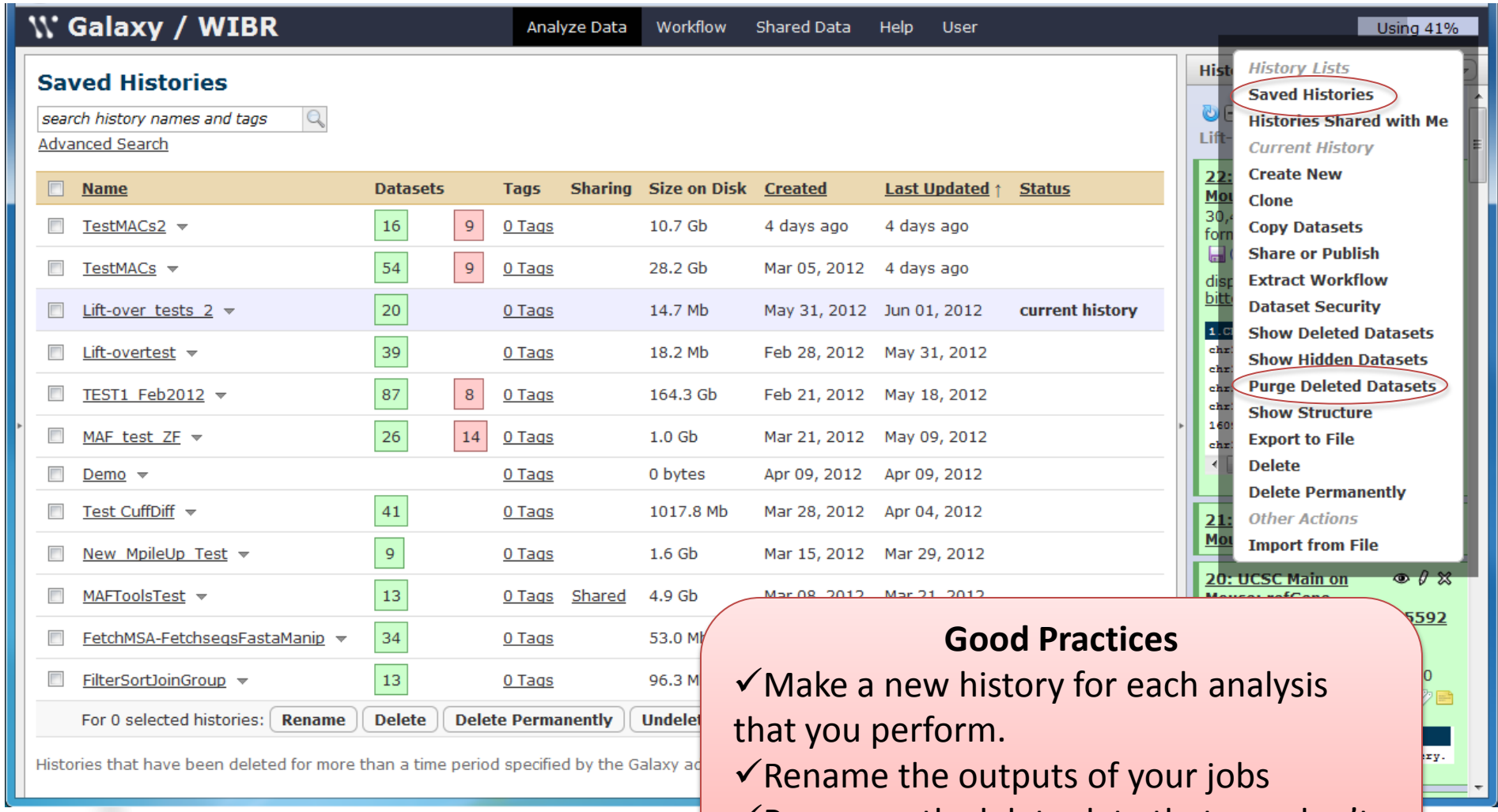
Run this job again

View or Report this error
Reporting an error will create a ticket

Clicking on the name of the dataset displays it bellow

Display data in local UCSC browser

History



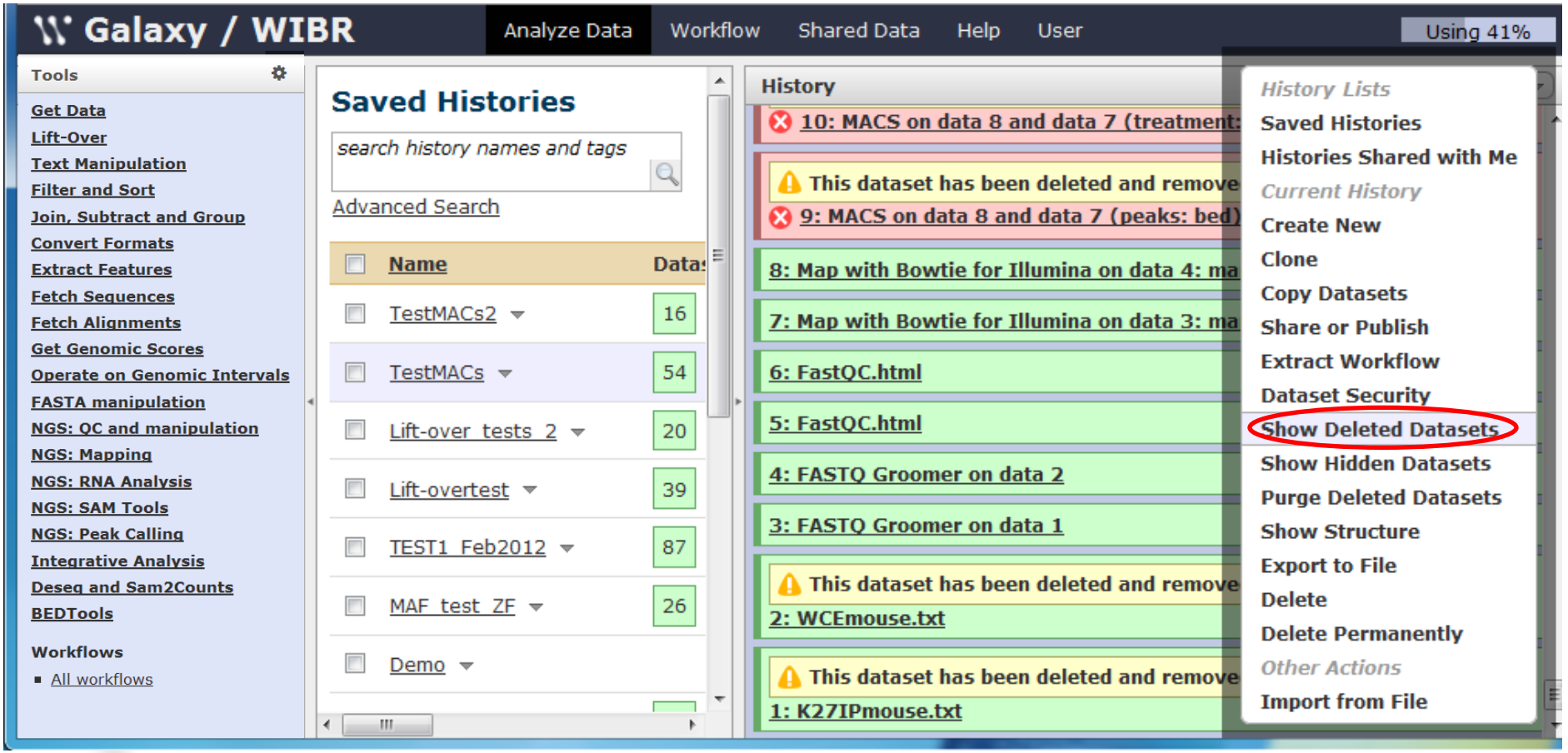
Good Practices

✓Make a new history for each analysis that you perform.

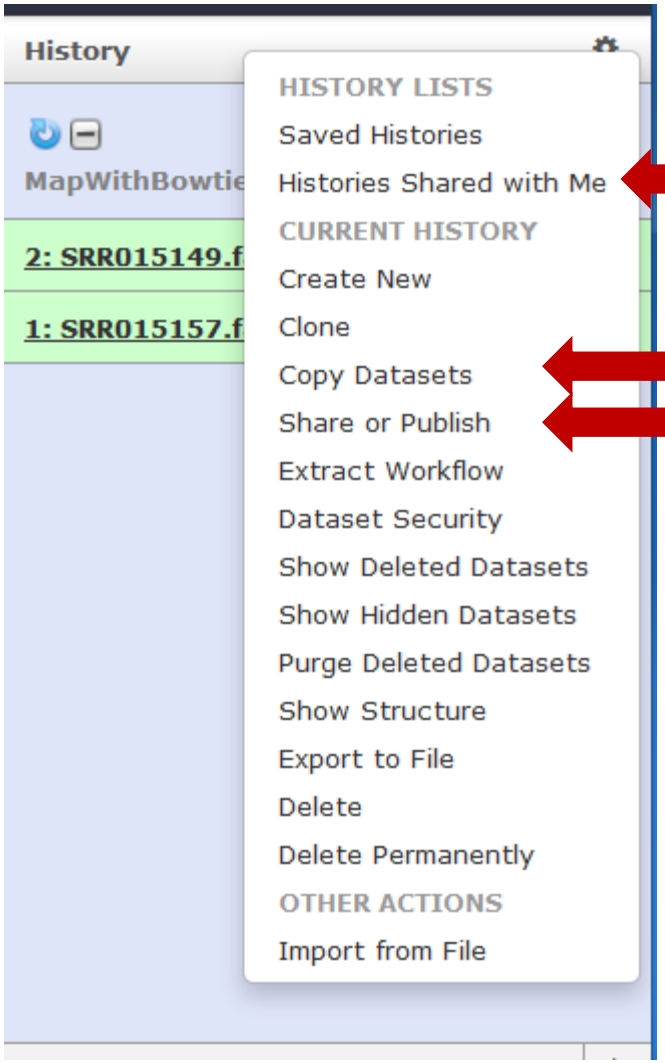
✓Rename the outputs of your jobs

✓Permanently delete data that you don't need (or you will reach your quota of 250Gb).

History is not removed when datasets are removed



Other useful commands on the History menu

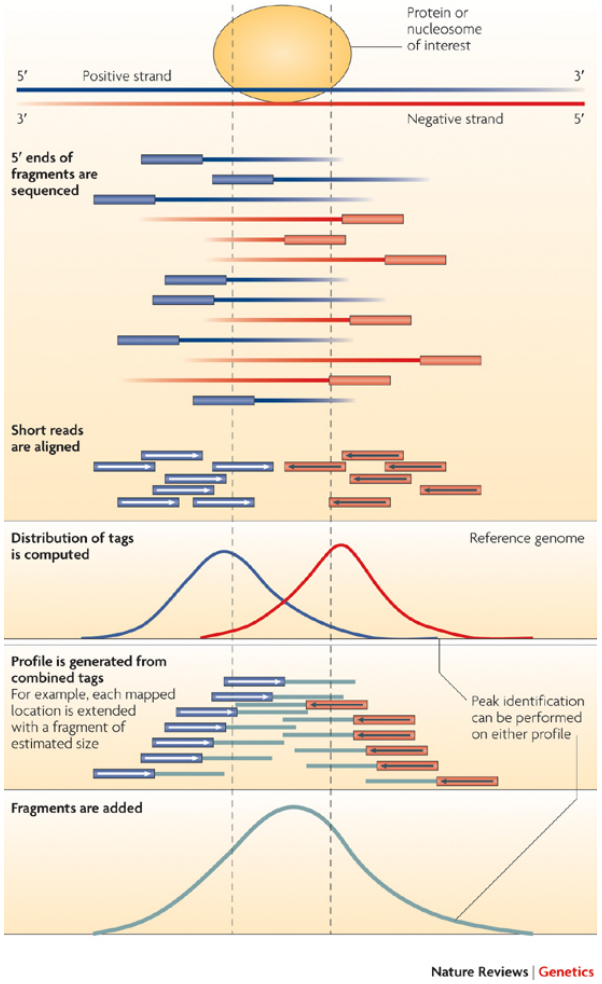


Access histories that other users shared with you

Transfer data between histories

Share your history with other users

General workflow for ChIP-seq analysis



Fastq files from the sequencing facility

↓

Check the quality of your reads
(NGS: QC and manipulation -> FastQC)

↓

← Step 1: Map the reads to the genome
(BOWTIE)

↓

← Step 2: Identify peaks
(MACS)

↓

Step 3: Post processing: Annotate peaks
i.e. find genes overlapping or close to the peaks

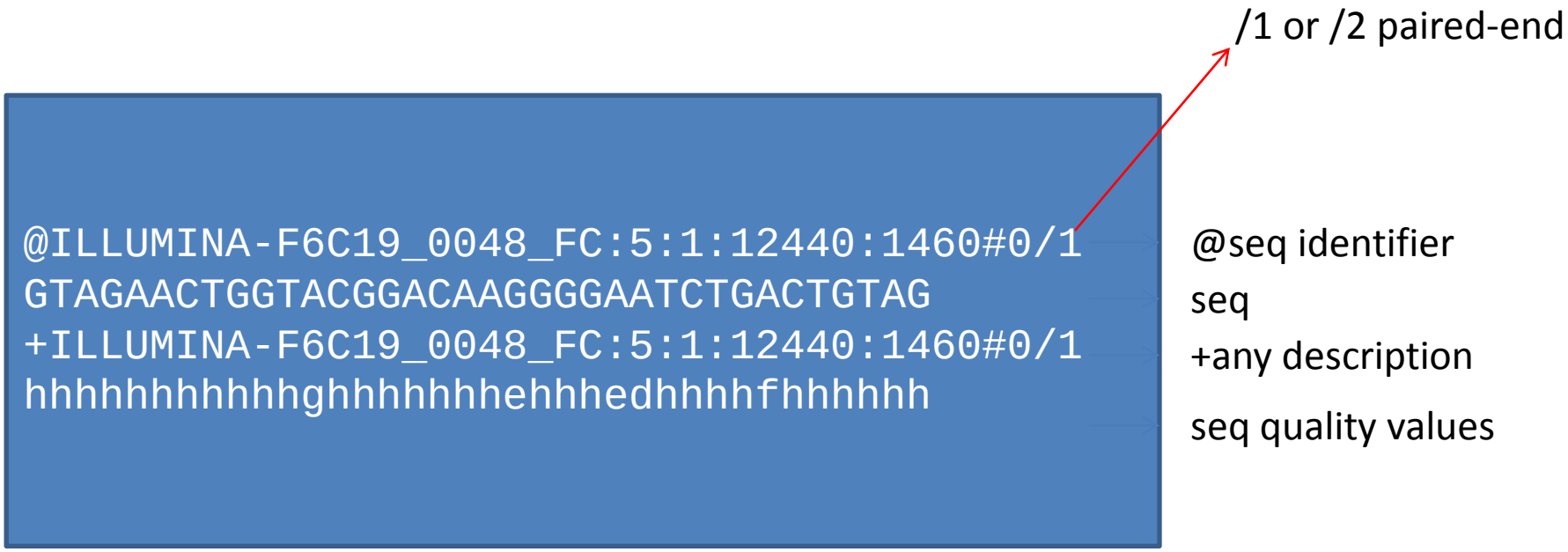
Park, P. J., *ChIP-seq: advantages and challenges of a maturing technology, Nat Rev Genet. Oct;10(10):669-80 (2009)*

Analysis of ChIP-seq data using Galaxy

1. History: *mapWithBowtie*
    1. Run FASTQ Groomer to convert fastq file to fastq Sanger format
    2. Map with bowtie
    3. Filter out unmapped reads
2. History: *imported: InputForMACS_mm9*
    1. Call peaks bound using MACS
    2. Select the peaks that are on chr1
3. History: *imported: InputFor_annotatePeaks.* Annotate peaks
    1. Annotate peaks using **Operate on Genomic Intervals** tools
    2. Annotate peaks using the **Integrative Analysis- > peak2gene** tool

Illumina data format

- Fastq format:



Sequence quality values on different FASTQ formats

http://en.wikipedia.org/wiki/FASTQ_format

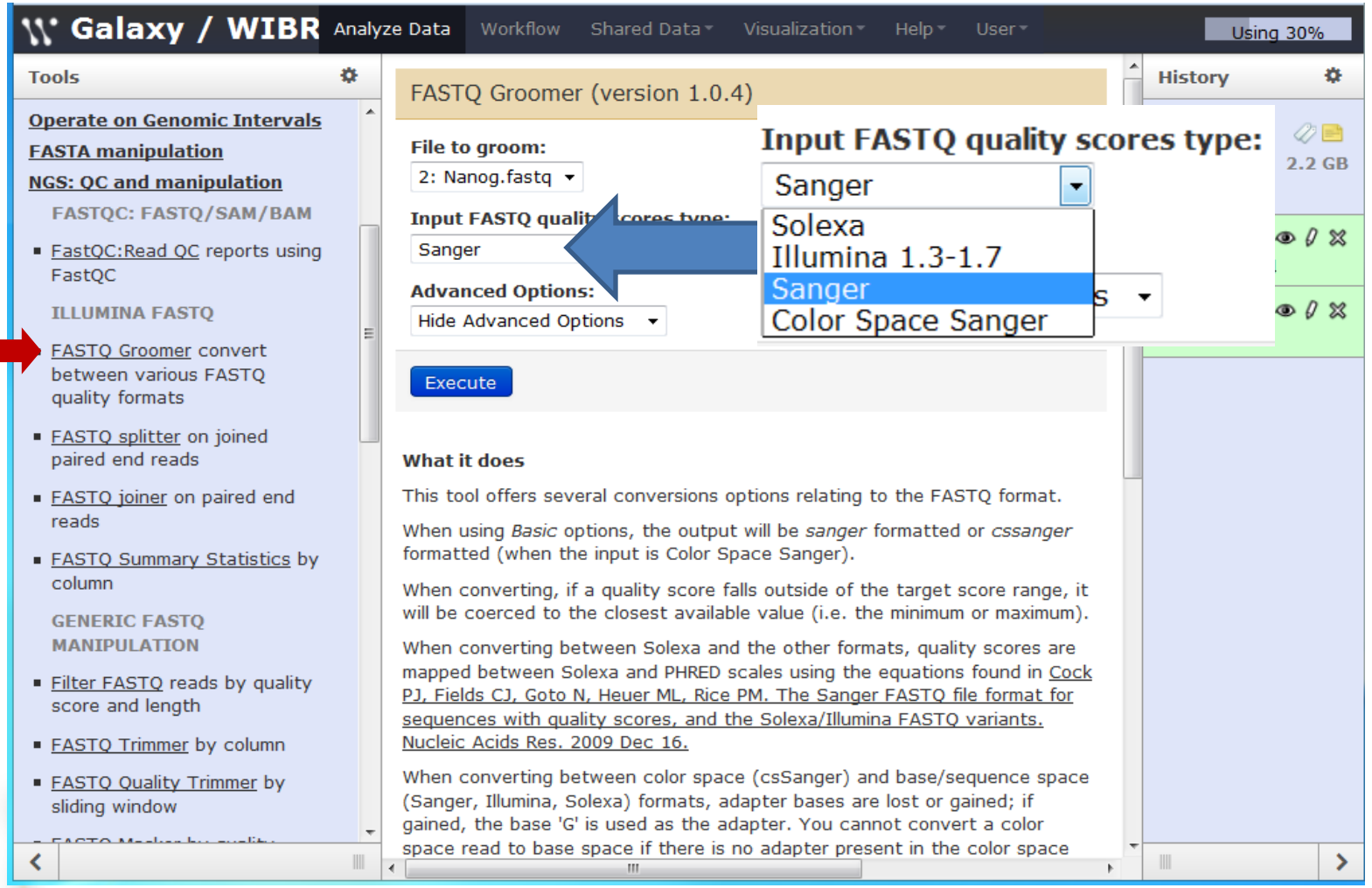
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.............................
 ......................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 .............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefgh
 |                         |    |        |                              |
33                       59   64       73                            104

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
```

To discriminate between Solexa and Illumina 1.3+ check if your sequences' quality scores have any of the characters ;<=>?

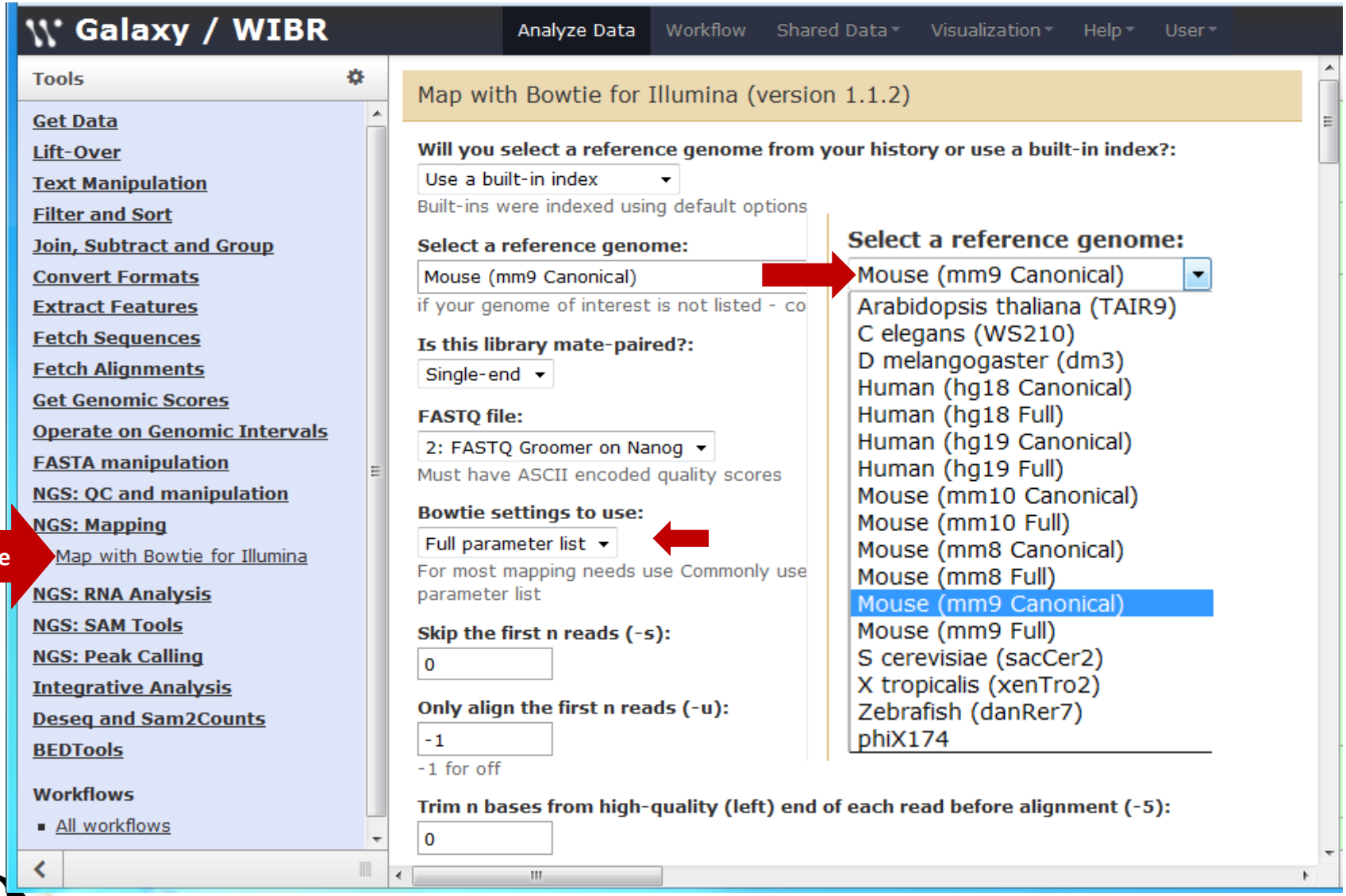
FASTQ formats and FASTQ Groomer

ILLUMINA FASTQ

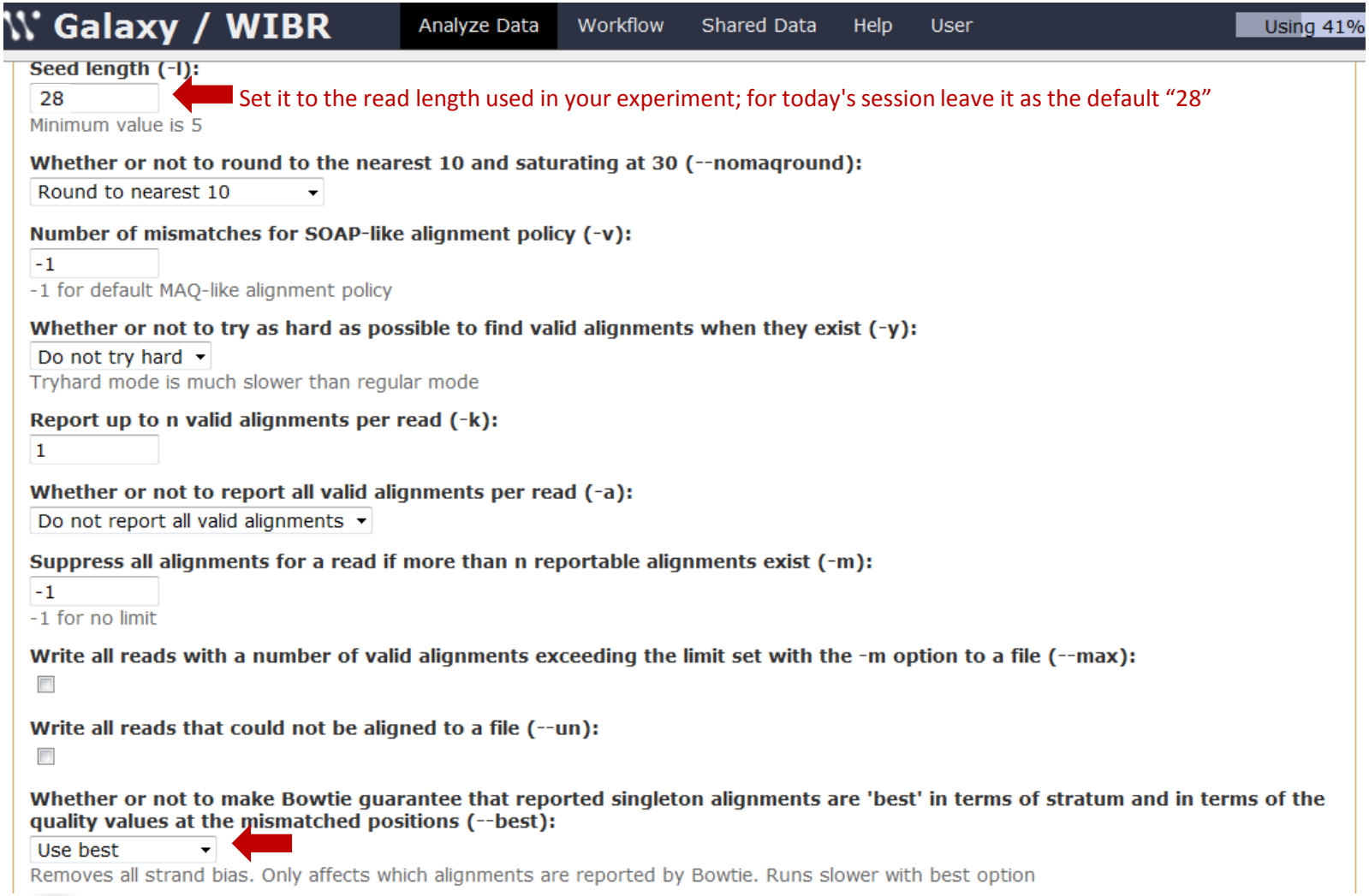
- FASTQ Groomer convert between various FASTQ quality formats



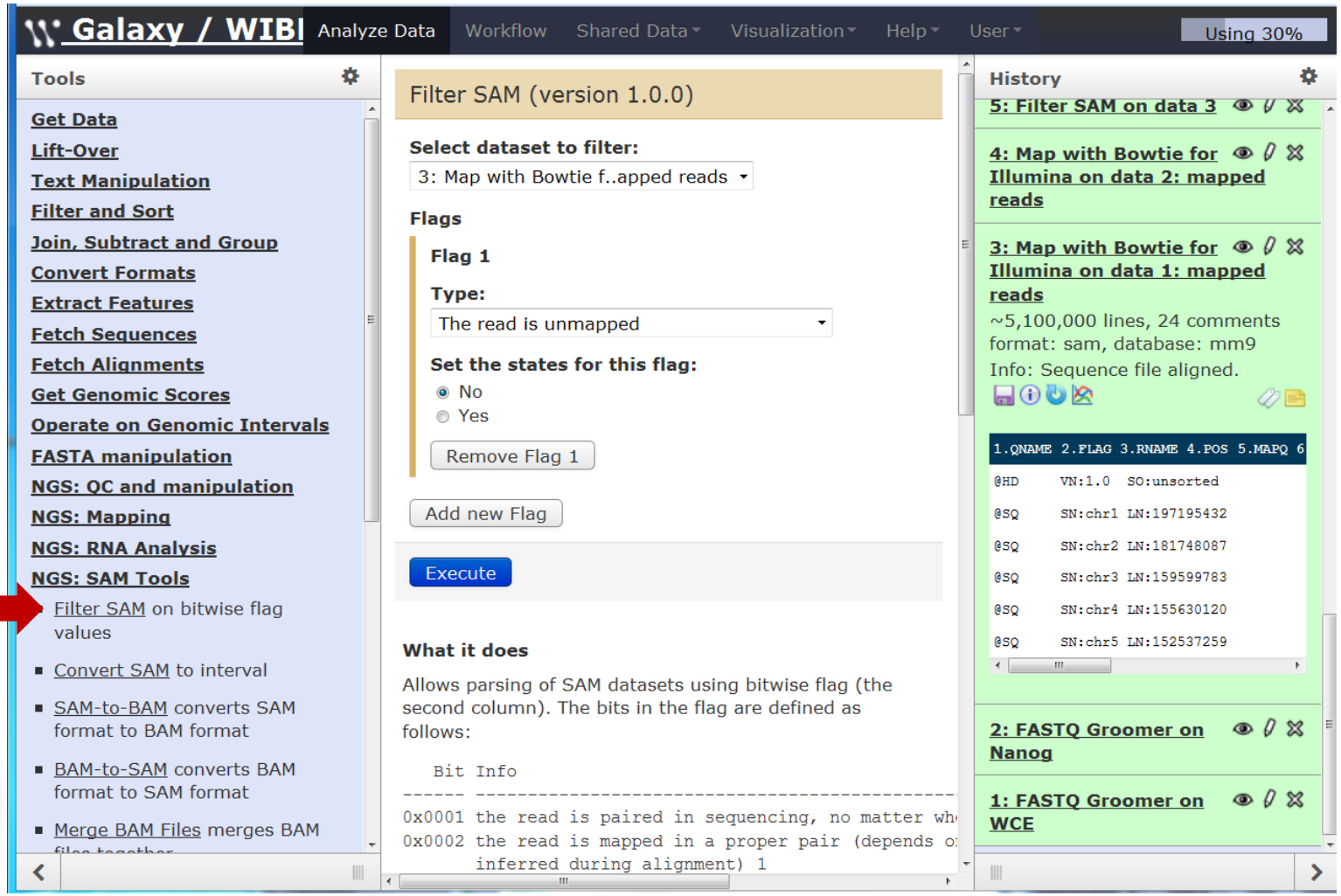
Mapping Reads with **Bowtie**



Mapping Reads with **Bowtie**



NGS: SAM Tools -> Filter SAM



Hands-on: Analysis of ChIP-seq data 1

History: *mapWithBowtie*

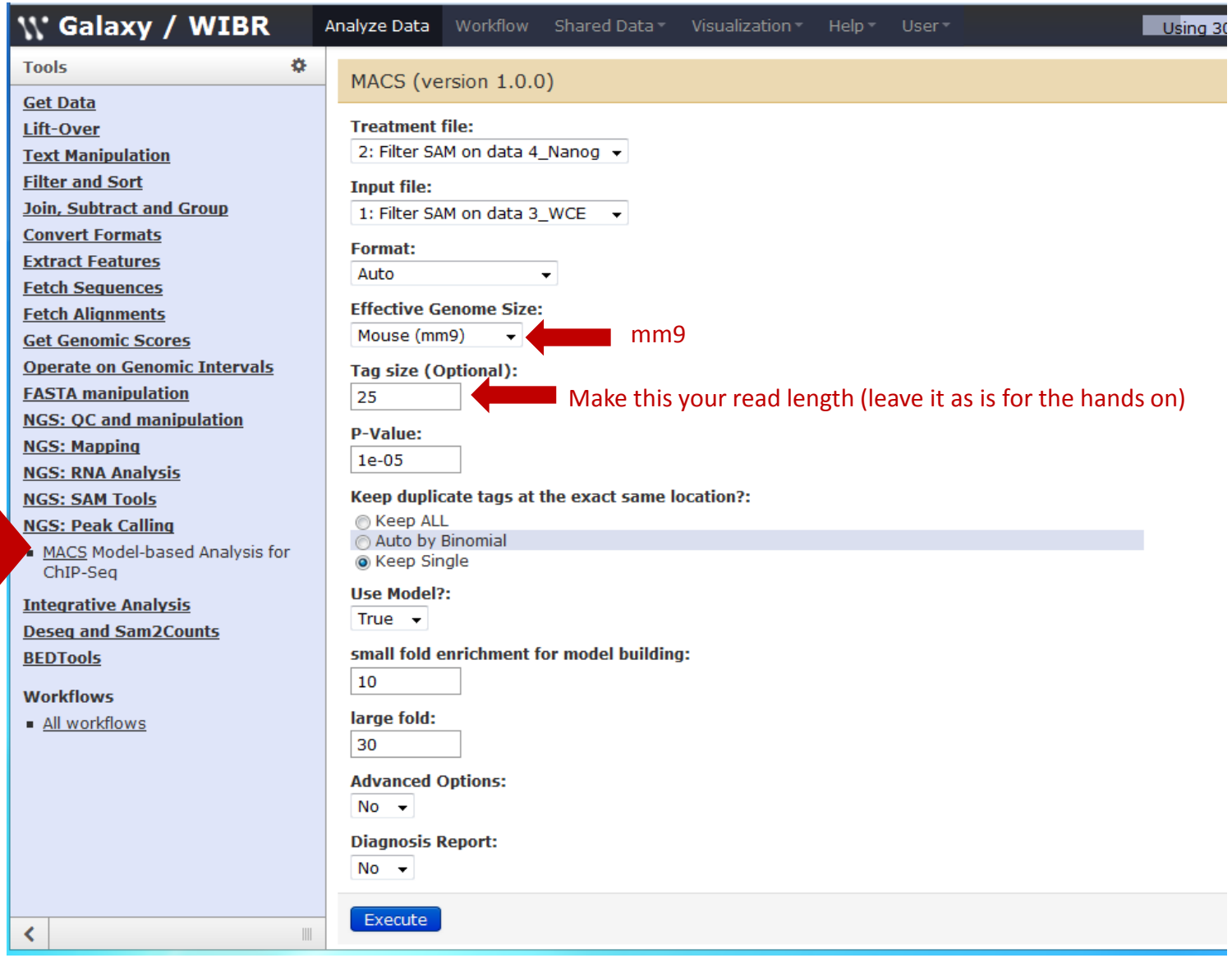
- Run **NGS: QC and manipulation -> FASTQ Groomer** on the 2 fastq files. The input files are Sanger format, but you still have to run fastq Groomer
- Map each of the fastq files with bowtie

  NGS: Mapping -> Map with Bowtie for Illumina
  - Genome to map to: mm9 canonical
  - Other parameters: use best
- Take the output from bowtie and filter out reads not mapped using:

  NGS: SAM Tools -> Filter SAM

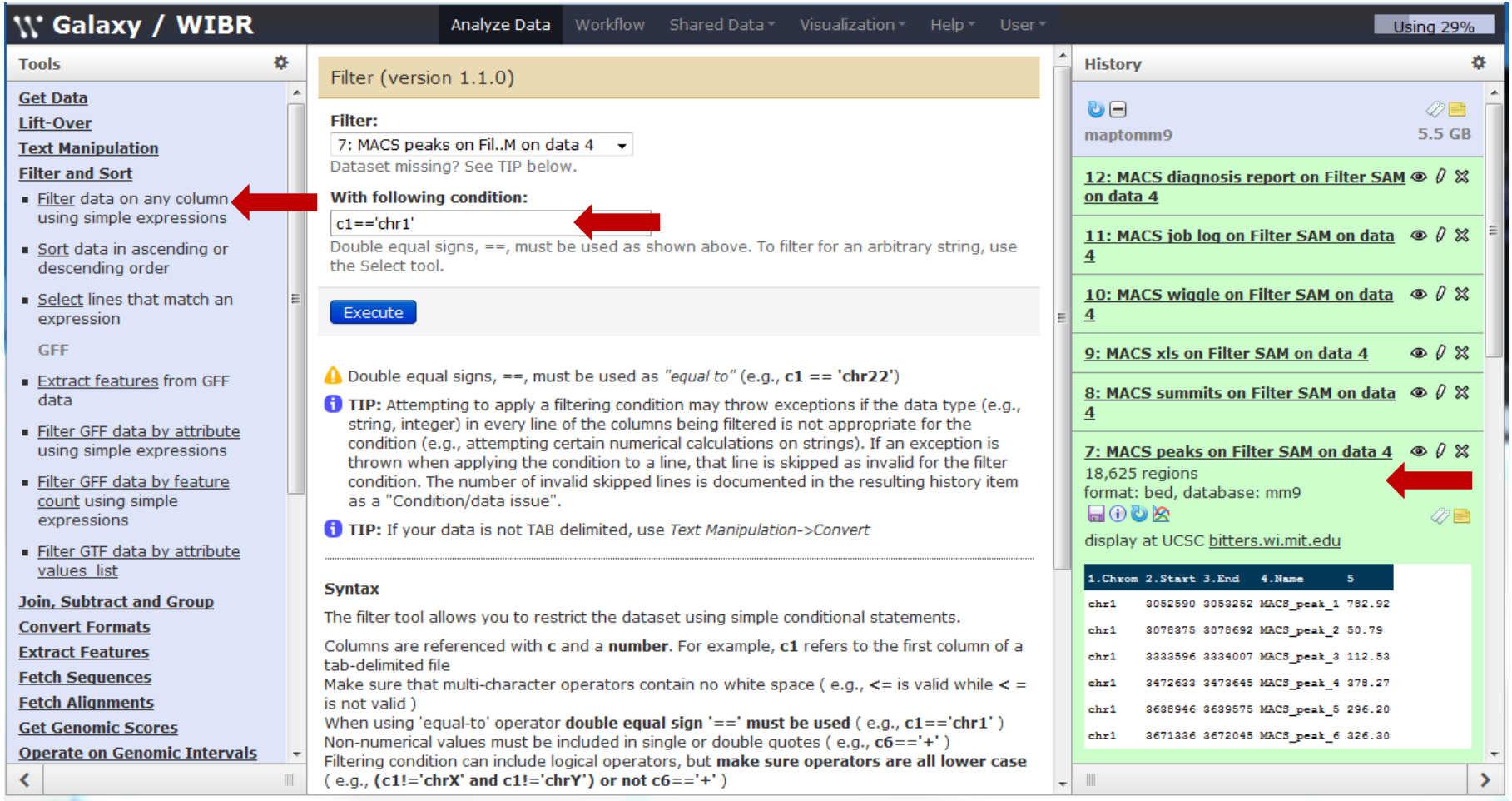
Tip: You don't have to wait for fastq groomer or bowtie to finish to send the next job

Analysis of ChIP-seq data: **MACS**



MACS

Filter and Sort:
Filter data on any column



Hands-on: Analysis of ChIP-seq data 2

History: *InputForMACS_mm9*

- Take the filtered mapped reads (uploaded by me in this history) and run MACS

  NGS: Peak Calling -> MACS Model-based Analysis for ChIP-Seq

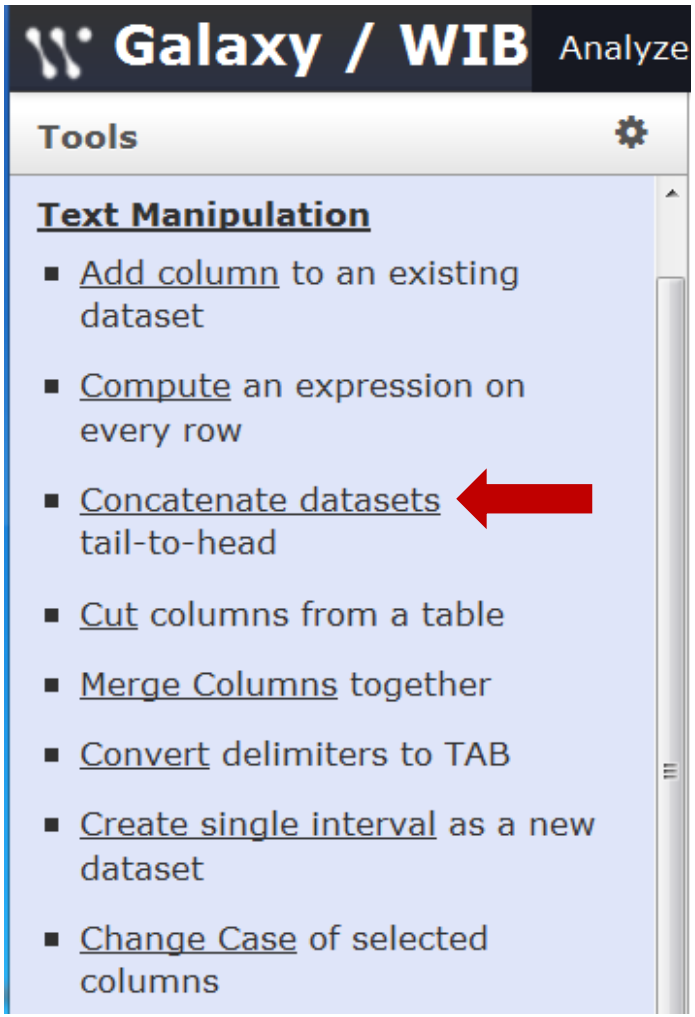
- Using the file that MACS generates “MACS peaks on Filter SAM on data 4 “ select only the peaks on chr1

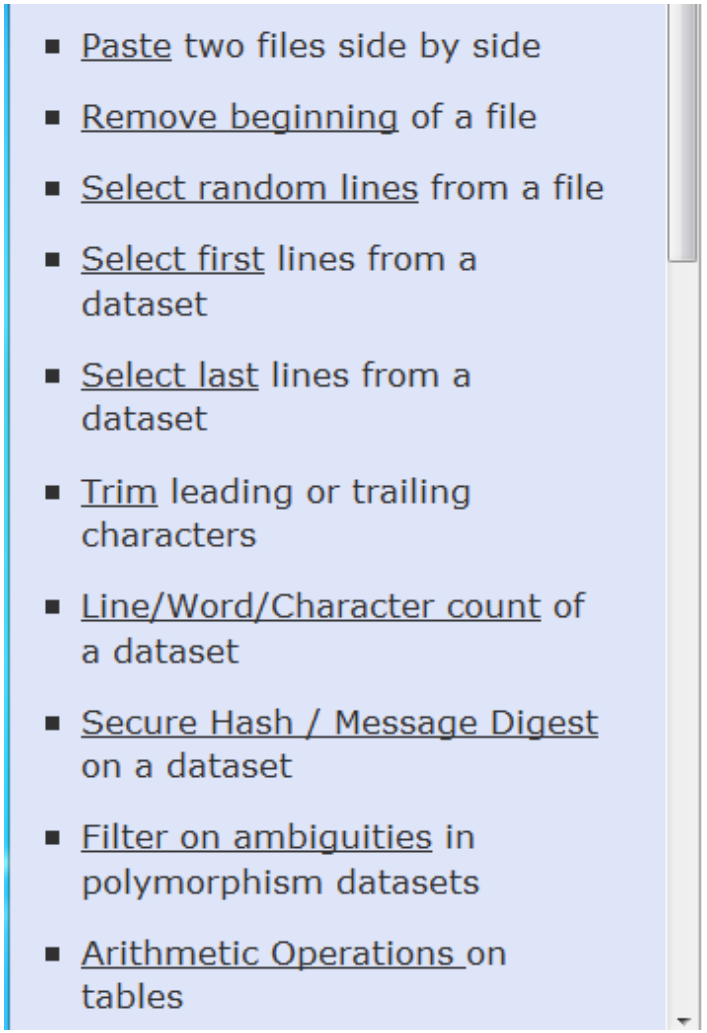
  Filter and Sort -> Filter data on any column using simple expressions

- Other filters you may want to use when you are running your analysis are:
  - Get the top 2000 peaks
  - Get peaks with FC > cut-off value
  - Get peaks with -log P > cut-off value

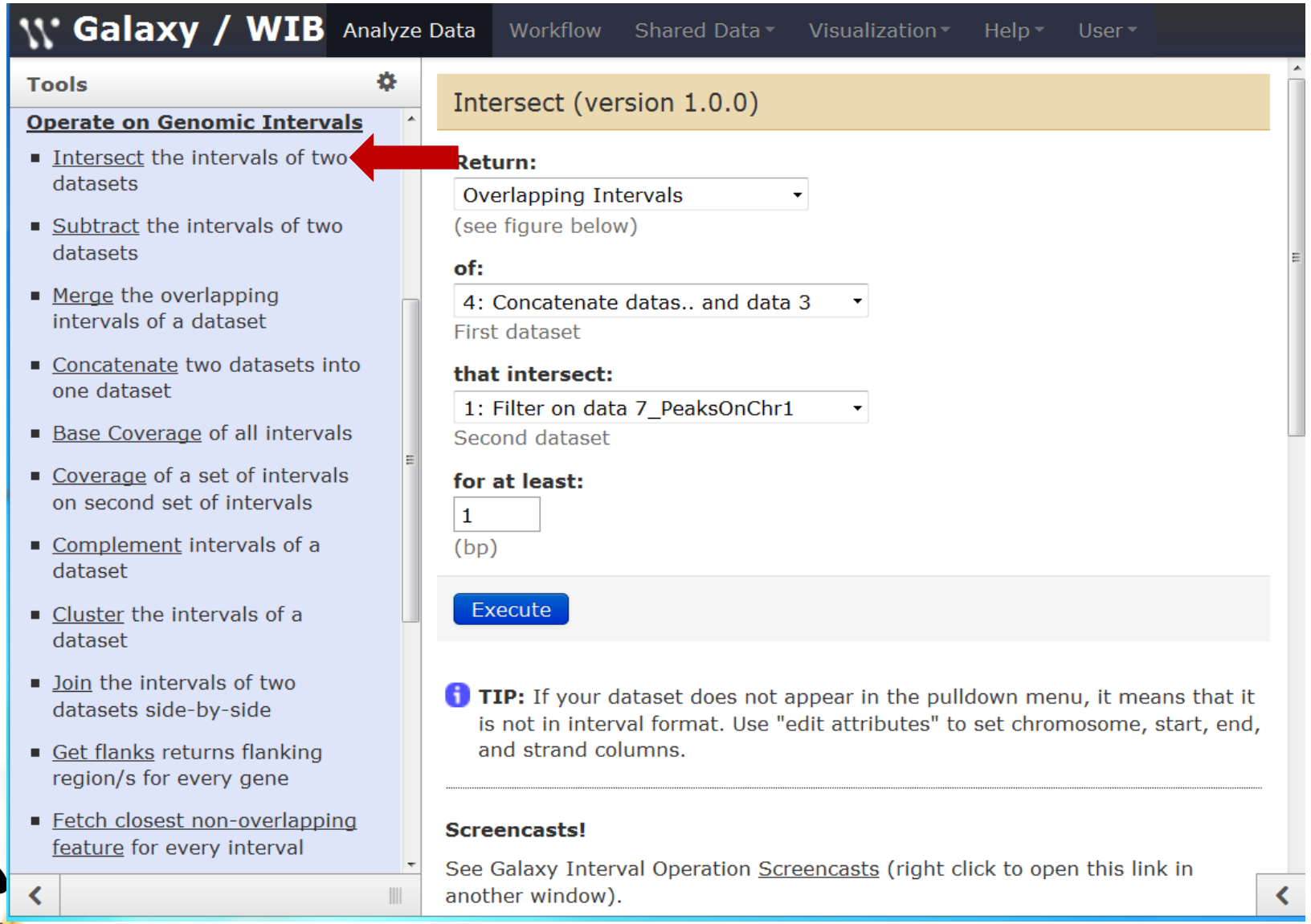
Post processing

Text Manipulation





Operate on Genomic Intervals: Intersect the intervals of two datasets



Operate on Genomic Intervals: Intersect the intervals of two datasets

Galaxy / WIBI | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 30%

Tools

Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a dataset
- Cluster the

Syntax

Where overlap is at least sets the minimum length (in base pairs) of overlap between elements of the two datasets
Overlapping Intervals returns entire intervals from the first dataset that overlap the second dataset. The returned intervals are completely unchanged, and this option only filters out intervals that do not overlap with the second dataset.
Overlapping pieces of Intervals returns intervals that indicate the exact base pair overlap between the first dataset and the second dataset. The intervals returned are from the first dataset, and all fields besides start and end are guaranteed to remain unchanged.

Examples

Overlapping Intervals:

Overlapping Pieces of Intervals:

Hands-on: Analysis of ChIP-seq data 3

History: *inputFor_annotatePeaks*

Annotate peaks.

1. Combine the mm10 refseq genes file and the 3Kb upstream of refseq gene file
   - **Text Manipulation -> Concatenate datasets tail-to-head**

   Find the genes or upstream regions that overlap with peaks
   - **Operate on Genomic Intervals -> Intersect the intervals of two datasets**

2. Find genes located at 3 Kb or less from the peak center using
   - **Integrative Analysis -> peak2gene**

Tutorials and References

- Galaxy tutorials
  http://galaxy.psu.edu/screencasts.html
- Previous Hot Topics
  http://jura.wi.mit.edu/bio/education/hot_topics
- References

Giardine et al. (2005) Galaxy: a platform for interactive large-scale analysis. *Genome Research* 15:1451-5

Taylor et al. (2007) Using Galaxy to perform large-scale interactive data analyses. *Current Protocols in Bioinformatics* Chapter 10, unit 10.

Blankenberg et al. (2010) Manipulation of FASTQ data with Galaxy. *Bioinformatics* 26(14):1783-5

Park, P. J. (2009) ChIP-seq: advantages and challenges of a maturing technology *Nat. Rev. Genet.* 10(10):669-80

Pepke et al. (2009) Computation for ChIP-seq and RNA-seq studies. *Nature Methods* 6, S22 - S32

Wilbanks et al. (2010) Evaluation of Algorithm Performance in ChIP-Seq peak Detection. *PLoS ONE* 5(7)

Szalkowski et al. (2010) Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Brief Bioinform* doi: 10.1093/bib/bbq068

Rye et al. (2011) A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Res.* 39(4):e25

Zhang et al. (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* vol. 9 (9) pp. R137