

Analysis of ChIP-seq data in Galaxy

November, 2012

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

Joint project between BaRC and IT

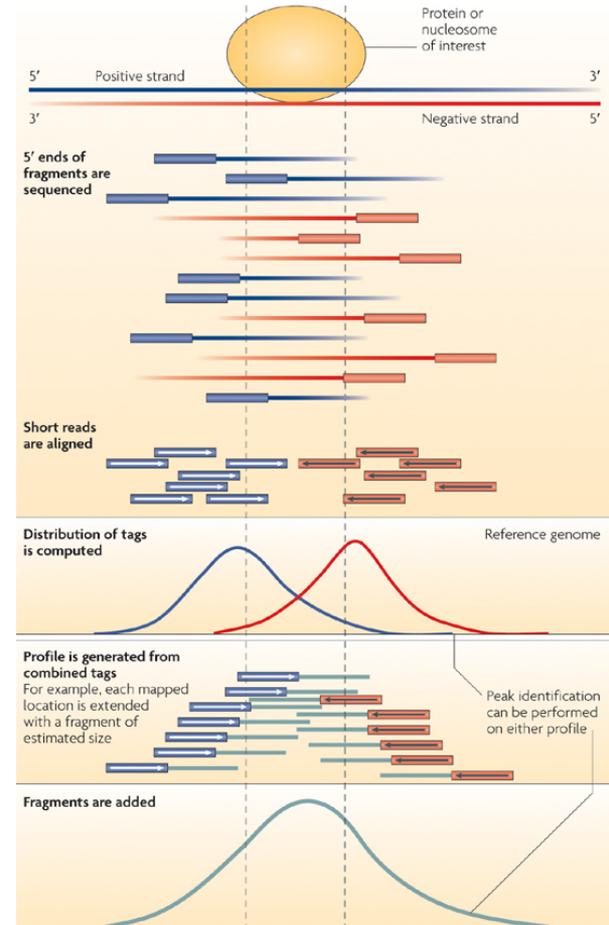
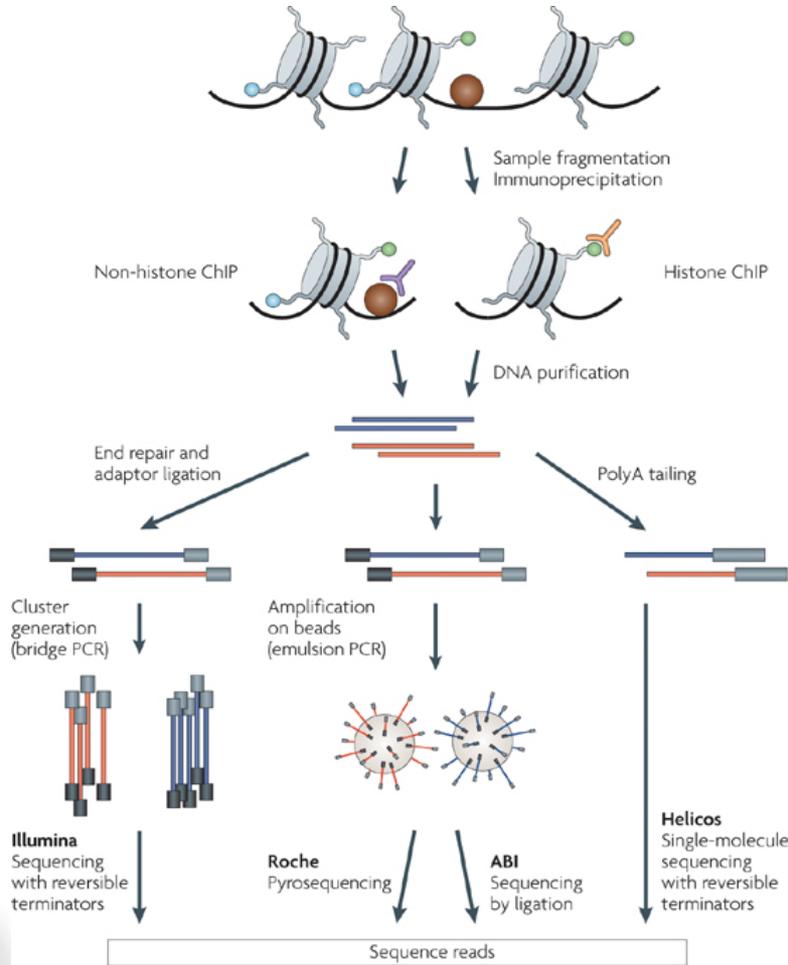
Main site: <http://main.g2.bx.psu.edu/>



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

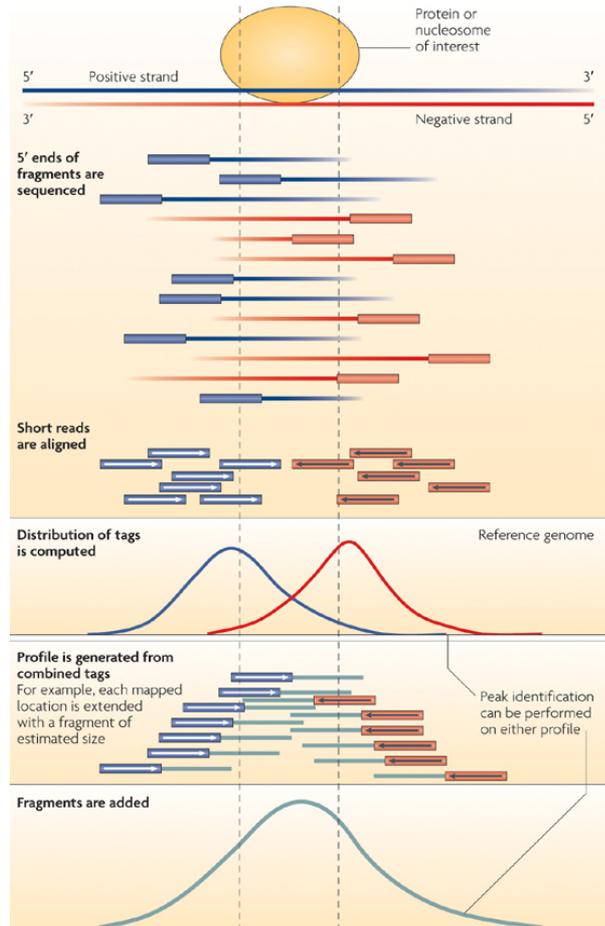
ChIP-seq



Nature Reviews | Genetics

Nature Reviews | Genetics

General workflow for ChIP-seq analysis



Nature Reviews | Genetics

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

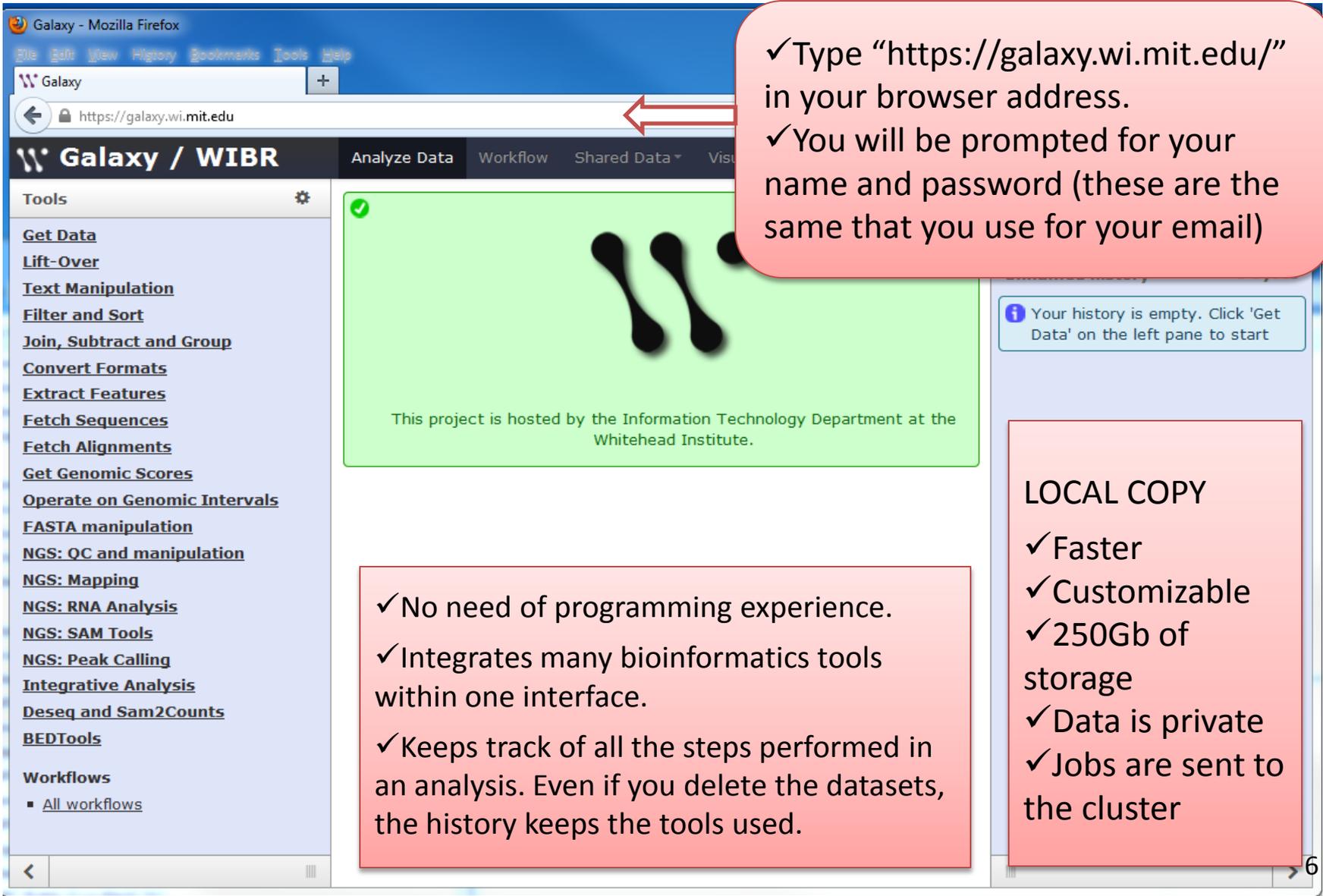
Park, P. J., *ChIP-seq: advantages and challenges of a maturing technology*, *Nat Rev Genet.* Oct;10(10):669-80 (2009)

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)

Your history is empty. Click 'Get Data' on the left pane to start

LOCAL COPY

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

✓ 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.

Galaxy Interface: Analyze Data

Data analysis

The screenshot shows the Galaxy / WIBR interface. At the top, there are navigation tabs: 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main area is divided into three sections:

- Tools window:** A sidebar on the left containing various tool categories such as 'Get Data', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'FASTA manipulation', 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: RNA Analysis', 'NGS: SAM Tools', 'NGS: Peak Calling', 'Integrative Analysis', 'Deseq and Sam2Counts', 'BEDTools', and 'Workflows'.
- Data display and tool's dialog window:** A central table displaying genomic data. The table has columns for chromosome (chr1), start coordinates, end coordinates, tool names (e.g., MACS_peak_1 to MACS_peak_28), and numerical values (e.g., 693.45, 112.82, etc.).
- History window:** A panel on the right showing a list of analysis jobs. The top job is 'imported: InputForMACS_mm9' (1.7 GB). Below it are several MACS-related jobs, each with a status icon (eye, pencil, or X) and a description like '8: MACS diagnosis report on Filter SAM on data 4'.

Processed data

Green:
job is finished

Yellow:
job is running

Gray:

job is in queue
Red: there is a problem

Tools window

Data display and tool's dialog window

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



How to find your previous histories

The screenshot shows the Galaxy web interface in a Mozilla Firefox browser window. The browser address bar displays <https://galaxy.wi.mit.edu>. The main navigation bar includes 'Galaxy / WIBR', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 30%' indicator is visible in the top right. On the left, a 'Tools' sidebar lists various analysis tools such as 'Get Data', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'FASTA manipulation', 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: RNA Analysis', 'NGS: SAM Tools', 'NGS: Peak Calling', 'Integrative Analysis', 'Deseq and Sam2Counts', and 'BEDTools'. The central workspace contains a green box with a checkmark and the text: 'This project is hosted by the Information Technology Department at the Whitehead Institute.' On the right, the 'History' panel is open, displaying a 'History Lists' menu with options like 'Saved Histories', 'Histories Shared with Me', 'CURRENT HISTORY', 'Create New', 'Clone', 'Copy Datasets', 'Share or Publish', 'Extract Workflow', 'Dataset Security', 'Show Deleted Datasets', 'Show Hidden Datasets', 'Purge Deleted Datasets', 'Show Structure', 'Export to File', 'Delete', 'Delete Permanently', and 'OTHER ACTIONS' including 'Import from File'. A blue arrow points from the right edge of the image to the 'History' menu.

History menu

Getting Data: Upload File

Galaxy / WIBR | Analyze Data | Workflow | Shared Data | Help | User

Tools ⚙️

Get Data

- Upload File
- UCSC Main
- BioMart Central server
- modENCODE fly server
- Flymine server
- YeastMine server
- modENCODE worm server
- Wormbase server

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

FASTA manipulation

NGS: QC and manipulation

NGS: Mapping

NGS: RNA Analysis

NGS: SAM Tools

NGS: Peak Calling

Workflows

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:

TIP: Due to browser limitations, uploads are limited to 100MB. To upload large files from Tak, copy them to your upload folder in /nfs/galaxy/uploads/username@wi.mit.edu

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
Your FTP upload directory contains no files.		

To upload some files from your desktop, log in with an SFTP client to galaxy.wi.mit.edu using your LDAP credentials. Your username is username@wi.mit.edu

Convert spaces to tabs:
 Yes
Use this option if you are entering intervals by hand.

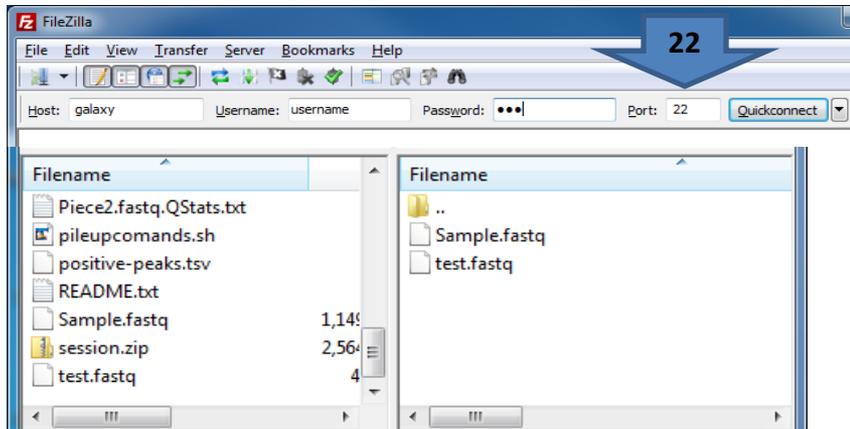
Genome:
unspecified (?)

File Format List:
Auto-detect
ab1
axt
bed
binseq.zip
blastxml
fasta
fastqsolexa
gff
gff3
interval
lav
maf
qual
scf
tabular
taxonomy
txt
txtseq.zip
wig

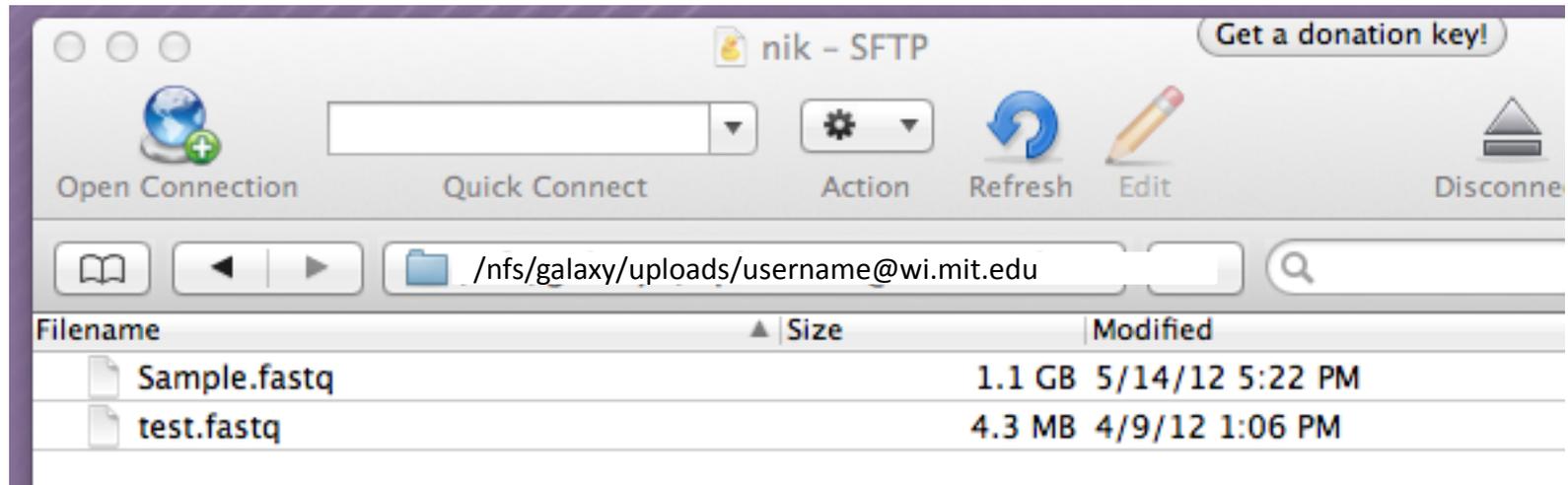
FTP File List:
unspecified (?)
A. thaliana (TAIR9)
C. elegans Jan. 2010 (WS210) (WS210)
D. melanogaster Apr. 2006 (BDGP R5/dm3) (dm3)
Human Feb. 2009 (GRCh37/hg19) (hg19)
Human Mar. 2006 (NCBI36/hg18) (hg18)
Mouse Dec. 2011 (GRCh38/mm10) (mm10)
Mouse July 2007 (NCBI37/mm9) (mm9)
Mouse Feb. 2006 (NCBI36/mm8) (mm8)
S. cerevisiae June 2008 (SGD/sacCer2) (sacCer2)
X. tropicalis Aug. 2005 (JGI4.1xenTro2)

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

Galaxy / WIBR Analyze Data Workflow Shared Data Visualization Help User Using 26%

Tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- BioMart Central server
- modENCODE fly server
- Flymine server
- modENCODE worm server
- WormBase server

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

FASTA manipulation

NGS: QC and manipulation

NGS: Mapping

NGS: RNA Analysis

NGS: SAM Tools

NGS: Peak Calling

Integrative Analysis

Deseq and s

BEDTools

Upload File (version 1.1.3)

Upload File

File:

Browse...

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files from Tak, copy them to your upload folder in /nfs/galaxy/uploads/username@wi.mit.edu

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
<input checked="" type="checkbox"/> WCE.fastq	803.7 MB	10/25/2012 02:37:46 PM
<input checked="" type="checkbox"/> Nanog.fastq	1.4 GB	10/25/2012 02:37:46 PM

To upload some files from your desktop, log in with an SFTP client to **galaxy.wi.mit.edu** using your LDAP credentials. Your upload folder is username@wi.mit.edu

Convert spaces to tabs:

Yes
Use this option if you are entering intervals by hand

Genome:

unspecified (?)

Execute

History

MapWithBowtie_mm10 0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

unspecified (?)

- A. thaliana (TAIR9)
- C. elegans Jan. 2010 (WS210) (WS210)
- D. melanogaster Apr. 2006 (BDGP R5/dm3) (dm3)
- Human Feb. 2009 (GRCh37/hg19) (hg19)
- Human Mar. 2006 (NCBI36/hg18) (hg18)
- Mouse Dec. 2011 (GRCh38/mm10) (mm10)
- Mouse July 2007 (NCBI37/mm9) (mm9)
- Mouse Feb. 2006 (NCBI36/mm8) (mm8)
- S. cerevisiae June 2008 (SGD/sacCer2) (sacCer2)
- X. tropicalis Aug. 2005 (JGI4.1xenTro2)

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*

The screenshot shows the Galaxy web interface. On the left, the 'Upload File (version 1.1.3)' tool is visible. In the center, the 'History' panel is open, showing a list of history items and a menu. A blue arrow labeled '1' points to the 'Create New' option. Another blue arrow labeled '2' points to the 'Saved Histories' option. On the right, the 'Saved Histories' table is displayed, showing a table with columns for Name, Datasets, Tags, Sharing, and Size. A context menu is open over the 'Unnamed history' row, and a blue arrow labeled '3' points to the 'Rename' option.

Name	Datasets	Tags	Sharing	Size
Unnamed history		0 Tags	0 b	
MapV		0 Tags	0 b	

Hands-on: Data Upload

stay in “*mapWithBowtie*” history

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

Galaxy / WIBR Analyze Data Workflow Shared Data Visualization Help User Using 26%

Tools

- Get Data**
 - Upload File from your computer
 - UCSC Main table browser
 - BioMart Central server
 - modENCODE fly server
 - Flymine server
 - modENCODE worm server
 - WormBase server
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- FASTA manipulation
- NGS: OC and manipulation
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- Integrative Analysis
- Deseq and Sam2Counts
- BEDTools

Upload File (version 1.1.3)

File Format:
fastq
Which format? See help below

File:
Browse...

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files from Tak, copy them to your upload folder in /nfs/galaxy/uploads/username@wi.mit.edu

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

	File	Size	Date
<input checked="" type="checkbox"/>	WCE.fastq	803.7 MB	10/25/2012 02:37:46 PM
<input checked="" type="checkbox"/>	Nanog.fastq	1.4 GB	10/25/2012 02:37:46 PM

To upload some files from your desktop, log in with an SFTP client to **galaxy.wi.mit.edu** using your LDAP credentials. Your upload folder is `username@wi.mit.edu`

Convert spaces to tabs:
 Yes
Use this option if you are entering intervals by hand.

Genome:
unspecified (?)

Execute

History

MapWithBowtie 0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

Hands-on: Data Upload

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

The screenshot shows the Galaxy / WIBR interface. The 'Shared Data' menu is open, and 'Published Histories' is selected. A table of published histories is visible below the menu.

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
InputFor_annotatePeaks		ibarrasa	★★★★★		19 minutes ago
InputForMACS_mm9		ibarrasa	★★★★★		~ 1 hour ago

- Select *InputForMACS_mm9*

The screenshot shows the 'Published Histories' page. The 'InputForMACS_mm9' entry is selected, and a red arrow points to it.

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
InputFor_annotatePeaks		ibarrasa	★★★★★		19 minutes ago
InputForMACS_mm9		ibarrasa	★★★★★		~ 1 hour ago

The screenshot shows the 'Import history' button and the 'About this History' panel. The 'Import history' button is highlighted with a red arrow.

Import history

About this History

Author
ibarrasa

Related Histories
All published histories
Published histories by ibarrasa

Rating
Community
(0 ratings, 0.0 average)

Yours
★★★★★

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**

Galaxy / WIBI Analyze Data Workflow Shared Data ▾ Vi

[Published Histories](#) | [ibarrasa](#) | [InputFor_annotatePeaks](#)

Galaxy History ' InputFor_annotatePeaks'

Dataset	Annotation
1: Filter on data 7_PeaksOnChr1	
2: UCSC_Mouse: refGenes_chr1	None
3: UCSC Mouse: refGenes_up_3000_chr1	None

Getting Data from UCSC (local copy)

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

Tools

- Get Data**
 - Upload File from your computer
 - UCSC Main** table browser
 - BioMart Central server
 - modENCODE fly server
 - Flymine server
 - modENCODE worm server
 - WormBase server
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- Integrative Analysis

Table Browser

to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to reference covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the [OpenHelix Table Browser tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: Mammal genome: Mouse assembly: July 2007 (NCBI37/mm9)

group: Genes and Gene Prediction Tracks track: RefSeq Genes manage custom tracks track hubs

table: refGene describe table schema

region: genome position chr1 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: BED - browser extensible data Send output to Galaxy GREAT

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output

To reset all user cache settings (including custom tracks), [click here](#).

Getting Data from UCSC (local copy)

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

Tools

- Get Data
 - Upload File from your computer
 - UCSC Main table browser
 - BioMart Central server
 - modENCODE fly server
 - Flymine server
 - modENCODE worm server
 - WormBase server
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- Integrative Analysis

Output refGene as BED

Include [custom track header](#):

name=

description=

visibility=

url=

Create one BED record per:

- Whole Gene
- Upstream by bases
- Exons plus bases at each end
- Introns plus bases at each end
- 5' UTR Exons
- Coding Exons
- 3' UTR Exons
- Downstream by bases

Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in order to avoid extending beyond the chromosome boundaries.

Send to Galaxy

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

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

- Imported:** InputForMACS_mm9 (1.7 GB)
- 8: MACS diagnosis report on Filter SAM on data 4** (empty, tabular format, mm9 database)
- 7: MACS job log on Filter SAM on data 4**
- 6: MACS wiggle on Filter SAM on data 4** (~57,000,000 lines, wig format, mm9 database)

Red arrows point from the text on the right to the dataset name in entry 6 and the UCSC link below it.

- 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 below

Display data in local UCSC browser

```
1
track type=wiggle_0 name="macs_ou
variableStep chrom=chr1 span=10
3001631
3001641
3001651
3001661
```

History

The screenshot shows the Galaxy / WIBR interface. At the top, there are navigation tabs: 'Analyze Data', 'Workflow', 'Shared Data', 'Help', and 'User'. The 'Using 41%' indicator is in the top right. The main content area is titled 'Saved Histories' and includes a search bar for history names and tags, with a link to 'Advanced Search'. Below this is a table of saved histories with columns for Name, Datasets, Tags, Sharing, Size on Disk, Created, Last Updated, and Status. The 'Lift-over tests 2' history is highlighted as the 'current history'. A context menu is open over the table, listing various actions such as 'Create New', 'Clone', 'Copy Datasets', 'Share or Publish', 'Extract Workflow', 'Dataset Security', 'Show Deleted Datasets', 'Show Hidden Datasets', 'Purge Deleted Datasets', 'Show Structure', 'Export to File', 'Delete', 'Delete Permanently', and 'Import from File'. The 'Purge Deleted Datasets' option is circled in red. At the bottom of the table, there are buttons for 'Rename', 'Delete', 'Delete Permanently', and 'Undelete'.

<input type="checkbox"/>	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
<input type="checkbox"/>	TestMACs2 ▾	16	9	0 Tags	10.7 Gb	4 days ago	4 days ago	
<input type="checkbox"/>	TestMACs ▾	54	9	0 Tags	28.2 Gb	Mar 05, 2012	4 days ago	
<input type="checkbox"/>	Lift-over tests 2 ▾	20	0 Tags		14.7 Mb	May 31, 2012	Jun 01, 2012	current history
<input type="checkbox"/>	Lift-overtest ▾	39	0 Tags		18.2 Mb	Feb 28, 2012	May 31, 2012	
<input type="checkbox"/>	TEST1 Feb2012 ▾	87	8	0 Tags	164.3 Gb	Feb 21, 2012	May 18, 2012	
<input type="checkbox"/>	MAF_test_ZF ▾	26	14	0 Tags	1.0 Gb	Mar 21, 2012	May 09, 2012	
<input type="checkbox"/>	Demo ▾		0 Tags		0 bytes	Apr 09, 2012	Apr 09, 2012	
<input type="checkbox"/>	Test CuffDiff ▾	41	0 Tags		1017.8 Mb	Mar 28, 2012	Apr 04, 2012	
<input type="checkbox"/>	New MpileUp_Test ▾	9	0 Tags		1.6 Gb	Mar 15, 2012	Mar 29, 2012	
<input type="checkbox"/>	MAFToolsTest ▾	13	0 Tags	Shared	4.9 Gb	Mar 08, 2012	Mar 21, 2012	
<input type="checkbox"/>	FetchMSA-FetchseqsFastaManip ▾	34	0 Tags		53.0 Mb			
<input type="checkbox"/>	FilterSortJoinGroup ▾	13	0 Tags		96.3 Mb			

For 0 selected histories: **Rename** **Delete** **Delete Permanently** **Undelete**

Histories that have been deleted for more than a time period specified by the Galaxy administrator.

History Lists
Saved Histories
Histories Shared with Me
Current History
Create New
Clone
Copy Datasets
Share or Publish
Extract Workflow
Dataset Security
Show Deleted Datasets
Show Hidden Datasets
Purge Deleted Datasets
Show Structure
Export to File
Delete
Delete Permanently
Other Actions
Import from File

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

The screenshot shows the Galaxy / WIBR interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Help', and 'User'. The 'Tools' panel on the left lists various genomic analysis tools. The 'Saved Histories' panel in the center displays a table of saved history items:

Name	Data
<input type="checkbox"/> TestMACs2	16
<input type="checkbox"/> TestMACs	54
<input type="checkbox"/> Lift-over tests_2	20
<input type="checkbox"/> Lift-overtest	39
<input type="checkbox"/> TEST1 Feb2012	87
<input type="checkbox"/> MAF test_ZF	26
<input type="checkbox"/> Demo	

The 'History' panel on the right shows a list of datasets. Some datasets are marked as deleted with a red 'X' icon and a warning message: 'This dataset has been deleted and removed'. The 'Show Deleted Datasets' option in the context menu is highlighted with a red circle.

Other useful commands on the History menu

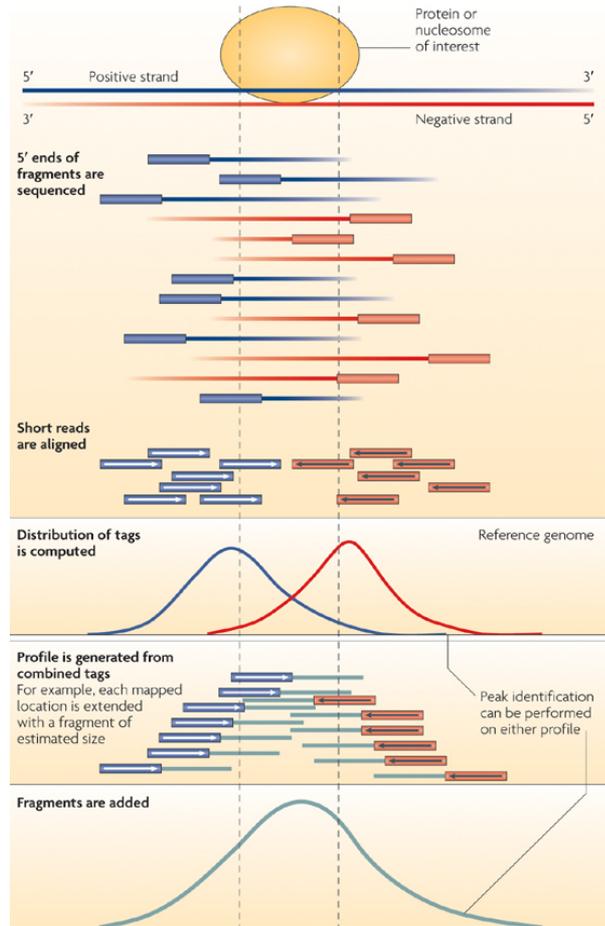


Access histories that other users shared with you

Transfer data between histories

Share your history with other users

General workflow for ChIP-seq analysis



Nature Reviews | Genetics

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:

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1  
GTAGAACTGGTACGGACAAGGGAATCTGACTGTAG  
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1  
hhhhhhhhhhghhhhhhhhehhhedhhhhfhhhhhh
```

/1 or /2 paired-end

@seq identifier
seq
+any description
seq quality values

FASTQ formats and FASTQ Groomer

ILLUMINA FASTQ

- [FASTQ Groomer](#) convert between various FASTQ quality formats

The screenshot shows the Galaxy / WIBR interface with the FASTQ Groomer tool (version 1.0.4) selected. The left sidebar contains a list of tools under the heading "Operate on Genomic Intervals". A red arrow points to the "FASTQ Groomer" tool description. The main panel shows the tool's configuration options: "File to groom:" is set to "2: Nanog.fastq", and "Input FASTQ quality scores type:" is set to "Sanger". A blue arrow points to the dropdown menu for "Input FASTQ quality scores type:", which lists "Sanger", "Solexa", "Illumina 1.3-1.7", "Sanger", and "Color Space Sanger". The "Execute" button is visible below the configuration options. The right sidebar shows the "History" panel with a job entry for "FASTQ Groomer" using 30% of the available resources.

FASTQ Groomer (version 1.0.4)

File to groom:

2: Nanog.fastq

Input FASTQ quality scores type:

Sanger

Advanced Options:

Hide Advanced Options

Execute

Input FASTQ quality scores type:

Sanger

Solexa

Illumina 1.3-1.7

Sanger

Color Space Sanger

What it does

This tool offers several conversions options relating to the FASTQ format.

When using *Basic* options, the output will be *sanger* formatted or *cssanger* formatted (when the input is Color Space Sanger).

When converting, if a quality score falls outside of the target score range, it will be coerced to the closest available value (i.e. the minimum or maximum).

When converting between Solexa and the other formats, quality scores are mapped between Solexa and PHRED scales using the equations found in [Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009 Dec 16.](#)

When converting between color space (csSanger) and base/sequence space (Sanger, Illumina, Solexa) formats, adapter bases are lost or gained; if gained, the base 'G' is used as the adapter. You cannot convert a color space read to base space if there is no adapter present in the color space

Mapping Reads with Bowtie

Galaxy / WIBR Analyze Data Workflow Shared Data Visualization Help User

Tools

- Get Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping**
- Map with Bowtie for Illumina
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Peak Calling
- Integrative Analysis
- Deseq and Sam2Counts
- BEDTools
- Workflows
 - All workflows

Map with Bowtie for Illumina (version 1.1.2)

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index

Built-ins were indexed using default options

Select a reference genome:
Mouse (mm9 Canonical)

if your genome of interest is not listed - co

Select a reference genome:

- Arabidopsis thaliana (TAIR9)
- C elegans (WS210)
- D melangogaster (dm3)
- Human (hg18 Canonical)
- Human (hg18 Full)
- Human (hg19 Canonical)
- Human (hg19 Full)
- Mouse (mm10 Canonical)
- Mouse (mm10 Full)
- Mouse (mm8 Canonical)
- Mouse (mm8 Full)
- Mouse (mm9 Canonical)**
- Mouse (mm9 Full)
- S cerevisiae (sacCer2)
- X tropicalis (xenTro2)
- Zebrafish (danRer7)
- phiX174

Is this library mate-paired?:
Single-end

FASTQ file:
2: FASTQ Groomer on Nanog

Must have ASCII encoded quality scores

Bowtie settings to use:
Full parameter list

For most mapping needs use Commonly use parameter list

Skip the first n reads (-s):
0

Only align the first n reads (-u):
-1
-1 for off

Trim n bases from high-quality (left) end of each read before alignment (-5):
0

Mapping Reads with Bowtie

Seed length (-l):

28

← Set it to the read length used in your experiment; for today's session leave it as the default "28"

Minimum value is 5

Whether or not to round to the nearest 10 and saturating at 30 (--nomaqround):

Round to nearest 10

Number of mismatches for SOAP-like alignment policy (-v):

-1

-1 for default MAQ-like alignment policy

Whether or not to try as hard as possible to find valid alignments when they exist (-y):

Do not try hard

Tryhard mode is much slower than regular mode

Report up to n valid alignments per read (-k):

1

Whether or not to report all valid alignments per read (-a):

Do not report all valid alignments

Suppress all alignments for a read if more than n reportable alignments exist (-m):

-1

-1 for no limit

Write all reads with a number of valid alignments exceeding the limit set with the -m option to a file (--max):

Write all reads that could not be aligned to a file (--un):

Whether or not to make Bowtie guarantee that reported singleton alignments are 'best' in terms of stratum and in terms of the quality values at the mismatched positions (--best):

Use best

←

Removes all strand bias. Only affects which alignments are reported by Bowtie. Runs slower with best option

NGS: SAM Tools -> Filter SAM

The screenshot shows the Galaxy / WIB web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. A 'Using 30%' indicator is visible in the top right. The left sidebar contains a 'Tools' menu with categories like 'Get Data', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'FASTA manipulation', 'NGS: QC and manipulation', 'NGS: Mapping', 'NGS: RNA Analysis', and 'NGS: SAM Tools'. A red arrow points to the 'Filter SAM on bitwise flag values' tool under 'NGS: SAM Tools'. The main panel displays the 'Filter SAM (version 1.0.0)' tool configuration. It includes a 'Select dataset to filter:' dropdown set to '3: Map with Bowtie f..apped reads', a 'Flags' section with 'Flag 1' set to 'The read is unmapped', and 'Set the states for this flag:' with 'No' selected. Below are 'Remove Flag 1', 'Add new Flag', and 'Execute' buttons. The 'What it does' section explains that the tool allows parsing of SAM datasets using bitwise flags. A 'Bit Info' table is shown below, with the following content:

Bit	Info
0x0001	the read is paired in sequencing, no matter wh
0x0002	the read is mapped in a proper pair (depends on inferred during alignment) 1

The right sidebar shows a 'History' panel with a list of jobs: '5: Filter SAM on data 3', '4: Map with Bowtie for Illumina on data 2: mapped reads', '3: Map with Bowtie for Illumina on data 1: mapped reads', '2: FASTQ Groomer on Nanog', and '1: FASTQ Groomer on WCE'. The '3: Map with Bowtie for Illumina on data 1: mapped reads' job is expanded, showing details: '~5,100,000 lines, 24 comments, format: sam, database: mm9, Info: Sequence file aligned.' Below this is a table with columns: 1. QNAME, 2. FLAG, 3. RNAME, 4. POS, 5. MAPQ, 6. SEQ. The table content is as follows:

1. QNAME	2. FLAG	3. RNAME	4. POS	5. MAPQ	6. SEQ
@HD	VN:1.0	SO:unsorted			
@SQ	SN:chr1	LN:197195432			
@SQ	SN:chr2	LN:181748087			
@SQ	SN:chr3	LN:159599783			
@SQ	SN:chr4	LN:155630120			
@SQ	SN:chr5	LN:152537259			

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

Galaxy / WIBR Analyze Data Workflow Shared Data Visualization Help User Using 30

Tools

- Get Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Peak Calling
 - MACS Model-based Analysis for ChIP-Seq
- Integrative Analysis
- Deseq and Sam2Counts
- BEDTools
- Workflows
 - All workflows

MACS (version 1.0.0)

Treatment file:
2: Filter SAM on data 4_Nanog

Input file:
1: Filter SAM on data 3_WCE

Format:
Auto

Effective Genome Size:
Mouse (mm9) ← mm9

Tag size (Optional):
25 ← Make this your read length (leave it as is for the hands on)

P-Value:
1e-05

Keep duplicate tags at the exact same location?:
 Keep ALL
 Auto by Binomial
 Keep Single

Use Model?:
True

small fold enrichment for model building:
10

large fold:
30

Advanced Options:
No

Diagnosis Report:
No

Execute

MACS

Filter and Sort: Filter data on any column

The screenshot shows the Galaxy/WIBR interface with the 'Filter' tool (version 1.1.0) selected. The tool configuration is as follows:

- Filter:** 7: MACS peaks on Fil..M on data 4
- With following condition:** c1=='chr1'

Two red arrows point to the 'Filter' dropdown and the 'With following condition' input field. The 'Execute' button is visible below the configuration.

The History panel on the right shows a list of jobs. Job 7, 'MACS peaks on Filter SAM on data 4', is highlighted with a red arrow. The job details show 18,625 regions in bed format for database mm9. Below the details is a table of the filtered data:

1. Chrom	2. Start	3. End	4. Name	5
chr1	3052590	3053252	MACS_peak_1	782.92
chr1	3078375	3078692	MACS_peak_2	50.79
chr1	3333596	3334007	MACS_peak_3	112.53
chr1	3472633	3473645	MACS_peak_4	378.27
chr1	3638946	3639575	MACS_peak_5	296.20
chr1	3671336	3672045	MACS_peak_6	326.30

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

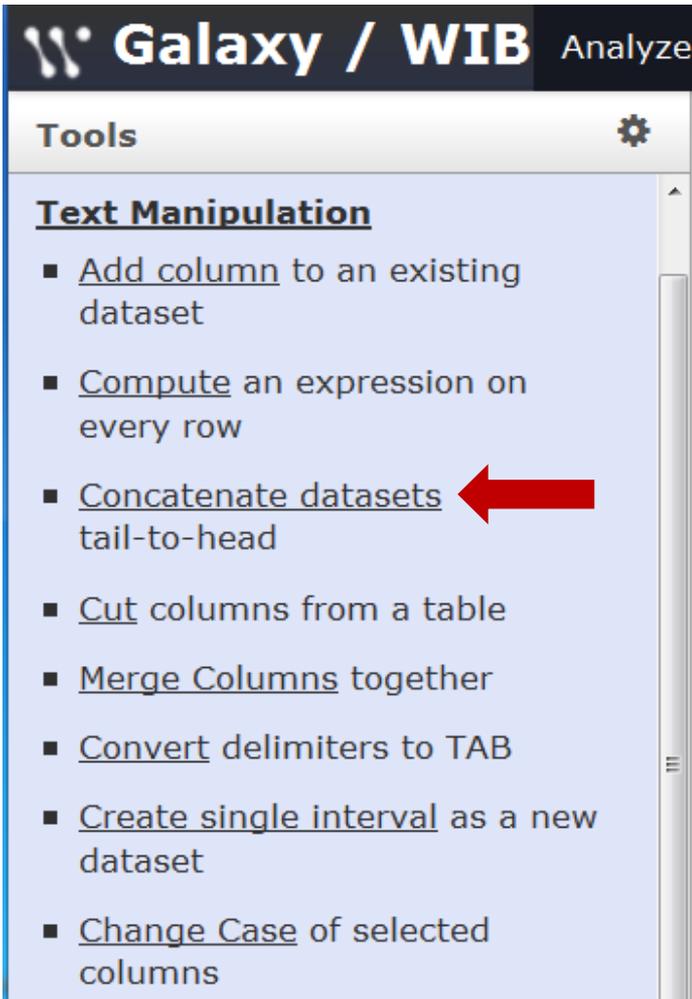
- Using the file that MACS generates “MACS peaks on Filter SAM on data 4 “ select only the peaks on chr1

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 $FC > \text{cut-off value}$
 - Get peaks with $-\log P > \text{cut-off value}$

Post processing

Text Manipulation



The screenshot shows the Galaxy / WIB Analyze interface. The 'Tools' section is expanded to show 'Text Manipulation' tools. A red arrow points to the 'Concatenate datasets tail-to-head' option.

Galaxy / WIB Analyze

Tools

Text Manipulation

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Cut columns from a table
- Merge Columns together
- Convert delimiters to TAB
- Create single interval as a new dataset
- Change Case of selected columns

- Paste two files side by side
- Remove beginning of a file
- Select random lines from a file
- Select first lines from a dataset
- Select last lines from a dataset
- Trim leading or trailing characters
- Line/Word/Character count of a dataset
- Secure Hash / Message Digest on a dataset
- Filter on ambiguities in polymorphism datasets
- Arithmetic Operations on tables

Operate on Genomic Intervals: Intersect the intervals of two datasets

The screenshot shows the Galaxy / WIB interface. The top navigation bar includes 'Galaxy / WIB', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. On the left, the 'Tools' panel is expanded to 'Operate on Genomic Intervals', with a red arrow pointing to the 'Intersect the intervals of two datasets' tool. The main panel displays the configuration for the 'Intersect (version 1.0.0)' tool. The 'Return:' dropdown is set to 'Overlapping Intervals'. The 'of:' dropdown is set to '4: Concatenate datas.. and data 3'. The 'that intersect:' dropdown is set to '1: Filter on data 7_PeaksOnChr1'. The 'for at least:' input field is set to '1'. An 'Execute' button is visible. A tip states: 'If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.' A 'Screencasts!' section provides a link to 'Galaxy Interval Operation Screencasts'.

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 intervals of a dataset
- [Join](#) the intervals of two datasets side-by-side
- [Get flanks](#) returns flanking region/s for every gene
- [Fetch closest non-overlapping feature](#) for every interval

Intersect (version 1.0.0)

Return:

Overlapping Intervals
(see figure below)

of:

4: Concatenate datas.. and data 3
First dataset

that intersect:

1: Filter on data 7_PeaksOnChr1
Second dataset

for at least:

1
(bp)

Execute

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!

See Galaxy Interval Operation [Screencasts](#) (right click to open this link in another window).

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