Galaxy

Web based platform for bioinformatics analysis

June 21, 2012

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

Joint project between BaRC and IT.

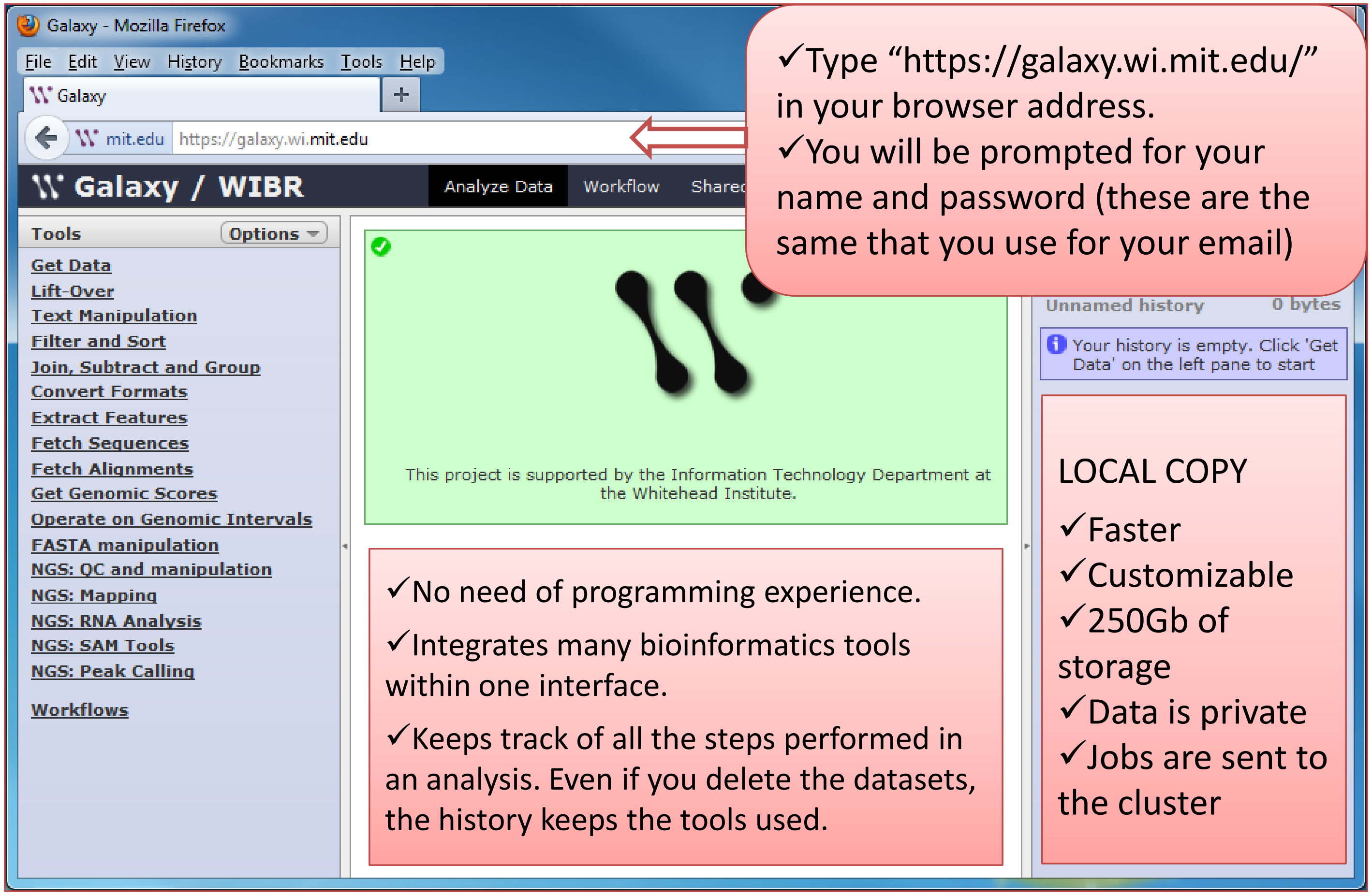
Main site: http://main.g2.bx.psu.edu/

Talk Outline

- The Galaxy interface
- Getting data into Galaxy
- Overview of the tools
- The Next Generation Sequencing tool box:
  - Preprocessing and quality control
  - Analysis of ChIP-seq
  - Analysis of RNA-seq
- Visualizing data on a genome browser and workflows available for analysis

Galaxy Interface:

A web based platform for analysis of large genomic datasets



This project is supported by the Information Technology Department at the Whitehead Institute.

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

✓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

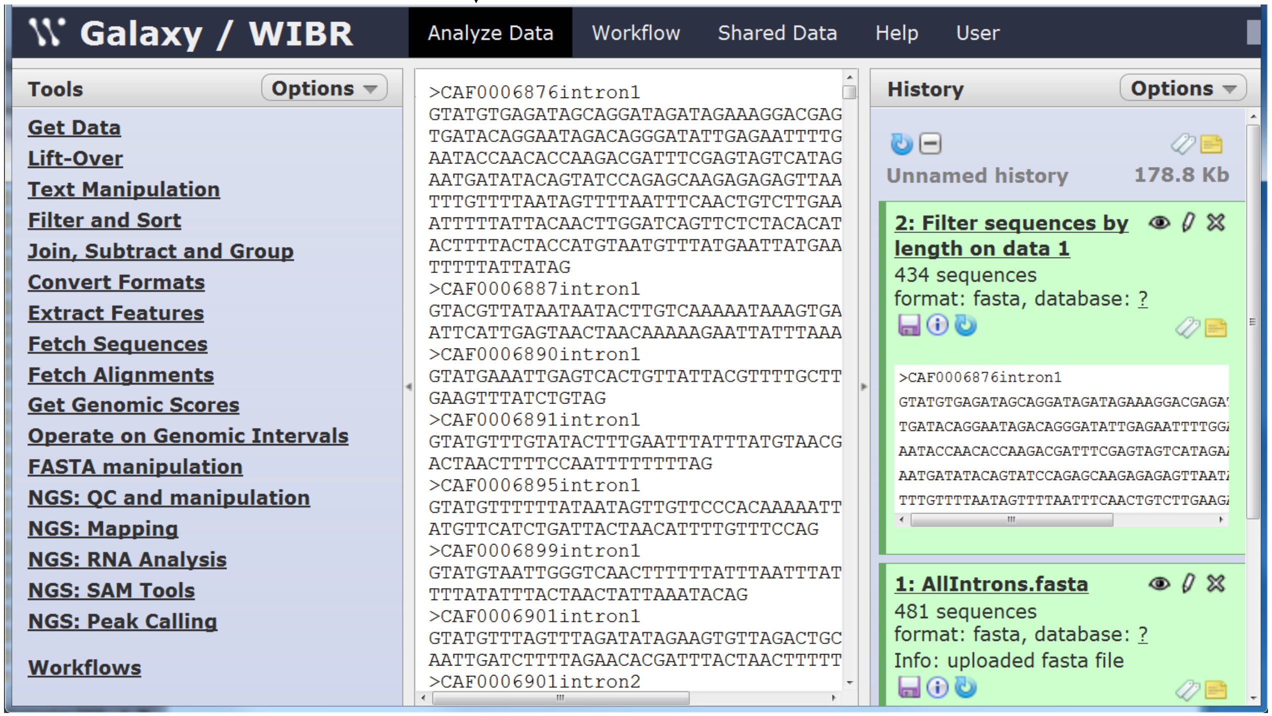
✓250Gb of storage

✓Data is private

✓Jobs are sent to the cluster

Galaxy Interface: Analyze Data

Data analysis



```
>CAF0006876intron1
GTATGTGAGATAGCAGGATAGATAGAAAGGACGAG
TGATACAGGAATAGACAGGGATATTGAGAATTTTG
AATACCAACACCAAGACGATTTCGAGTAGTCATAG
AATGATATACAGTATCCAGAGCAAGAGAGAGTTAA
TTTGTTTTAATAGTTTTAATTTCAACTGTCTTGAA
ATTTTTATTACAACTTGGATCAGTTCTCTACACAT
ACTTTTACTACCATGTAATGTTTATGAATTATGAA
TTTTTATTATAG
>CAF0006887intron1
GTACGTTATAATAATACTTGTCAAAAATAAAGTGA
ATTCATTGAGTAACTAACAAAAAGAATTATTTAAA
>CAF0006890intron1
GTATGAAATTGAGTCACTGTTATTACGTTTTGCTT
GAAGTTTATCTGTAG
>CAF0006891intron1
GTATGTTTGTATACTTTGAATTTATTTATGTAACG
ACTAACTTTTCCAATTTTTTTTAG
>CAF0006895intron1
GTATGTTTTTTATAATAGTTGTTCCCACAAAAATT
ATGTTCATCTGATTACTAACATTTTGTTTCCAG
>CAF0006899intron1
GTATGTAATTGGGTCAACTTTTTTATTTAATTTAT
TTTATATTTACTAACTATTAAATACAG
>CAF0006901intron1
GTATGTTTAGTTTAGATATAGAAGTGTTAGACTGC
AATTGATCTTTTAGAACACGATTTACTAACTTTTT
>CAF0006901intron2
```

Unnamed history 178.8 Kb

2: Filter sequences by length on data 1
434 sequences
format: fasta, database: ?

```
>CAF0006876intron1
GTATGTGAGATAGCAGGATAGATAGAAAGGACGAGA
TGATACAGGAATAGACAGGGATATTGAGAATTTTGG
AATACCAACACCAAGACGATTTCGAGTAGTCATAGA
AATGATATACAGTATCCAGAGCAAGAGAGAGTTAAT
TTTGTTTTAATAGTTTTAATTTCAACTGTCTTGAAG
```

1: AllIntrons.fasta
481 sequences
format: fasta, database: ?
Info: uploaded fasta file

Processed data

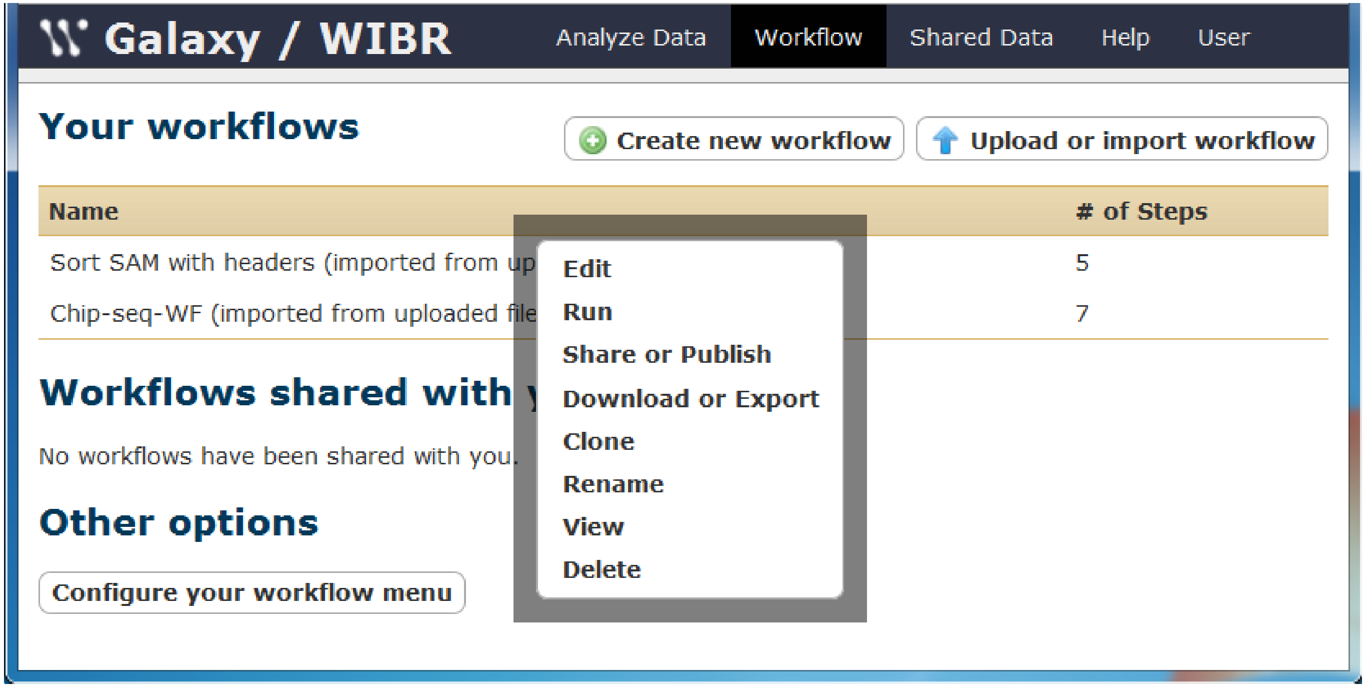
Green:
job is finished
Yellow:
job is running
Gray:
job is in queue
Red: there is a
problem

Tools
window

Data display and tool's dialog
window

History window:
datasets for each
analysis are kept here

Galaxy Interface: Workflow



Your workflows

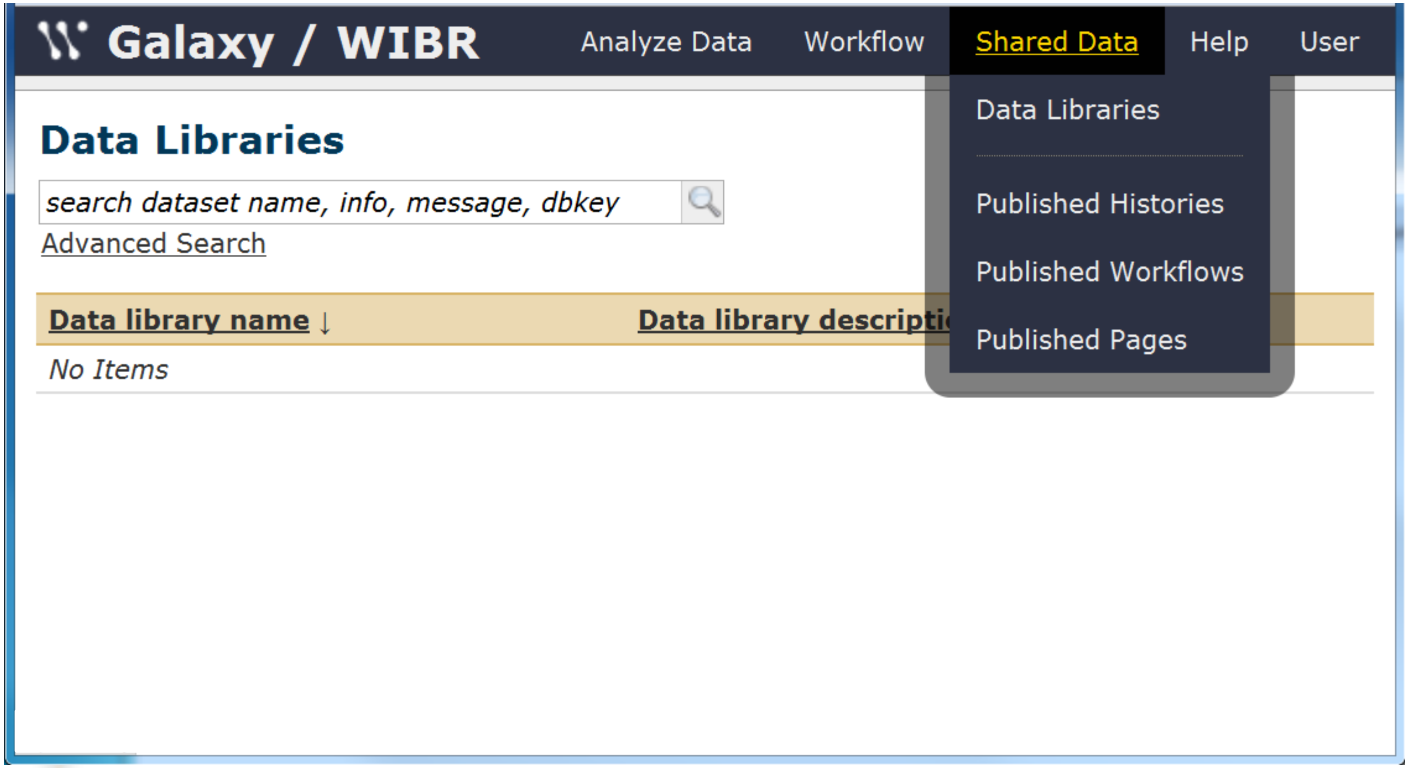
Name	# of Steps
Sort SAM with headers (imported from up	5
Chip-seq-WF (imported from uploaded file	7

Workflows shared with y

No workflows have been shared with you.

Other options

Galaxy Interface: Shared Data



Data Libraries

search dataset name, info, message, dbkey

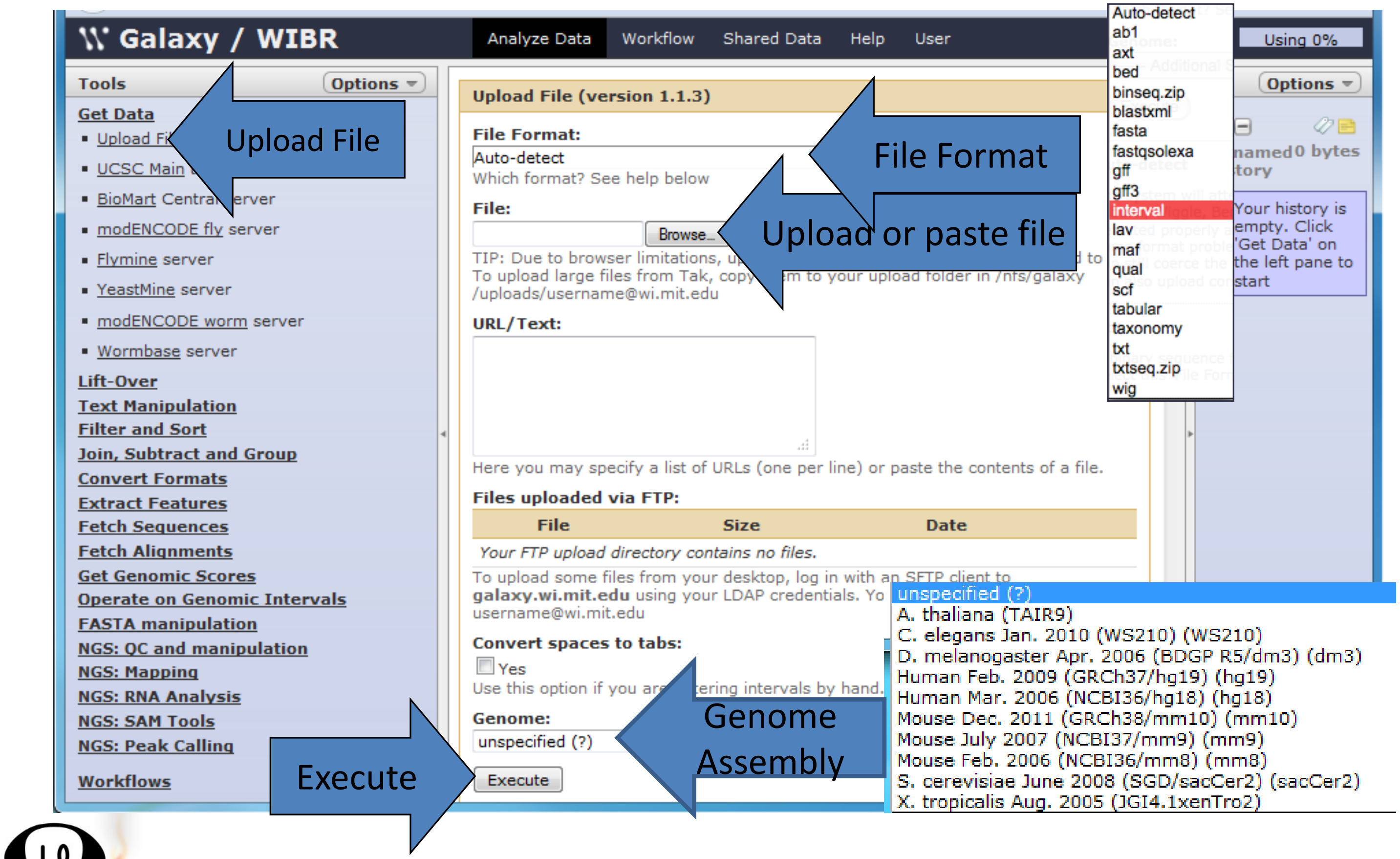
Advanced Search

Data library name ↓	Data library descripti
No Items	

Talk Outline

- The Galaxy interface
- Getting data into Galaxy
- Overview of the tools
- The Next Generation Sequencing tool box
  - Preprocessing and quality control
  - Analysis of ChIP-seq
  - Analysis of RNA-seq
- Visualizing data on a genome browser and workflows available for analysis

Getting Data: Upload File



Upload File (version 1.1.3)

File Format:

Auto-detect

Which format? See help below

File:

TIP: Due to browser limitations, u ... d to
To upload large files from Tak, copy ... m 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. Yo
username@wi.mit.edu

Convert spaces to tabs:

□ Yes

Use this option if you are ... ering intervals by hand.

Genome:

unspecified (?)

Upload File

File Format

Upload or paste file

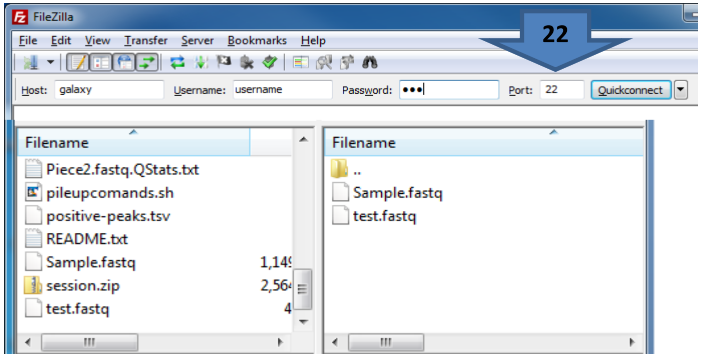
Genome Assembly

Execute

Getting Data: Uploading Large Files

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

FileZilla

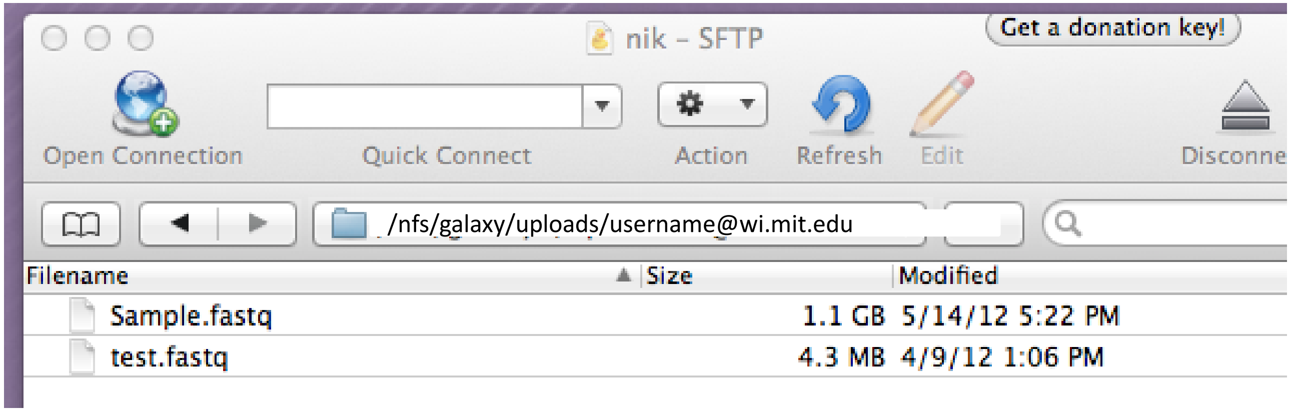


Host: galaxy Username: username Password: ••• Port: 22

Filename	
Piece2.fastq.QStats.txt	
pileupcomands.sh	
positive-peaks.tsv	
README.txt	
Sample.fastq	1,14
session.zip	2,56
test.fastq	4

Filename
..
Sample.fastq
test.fastq

CyberDuck

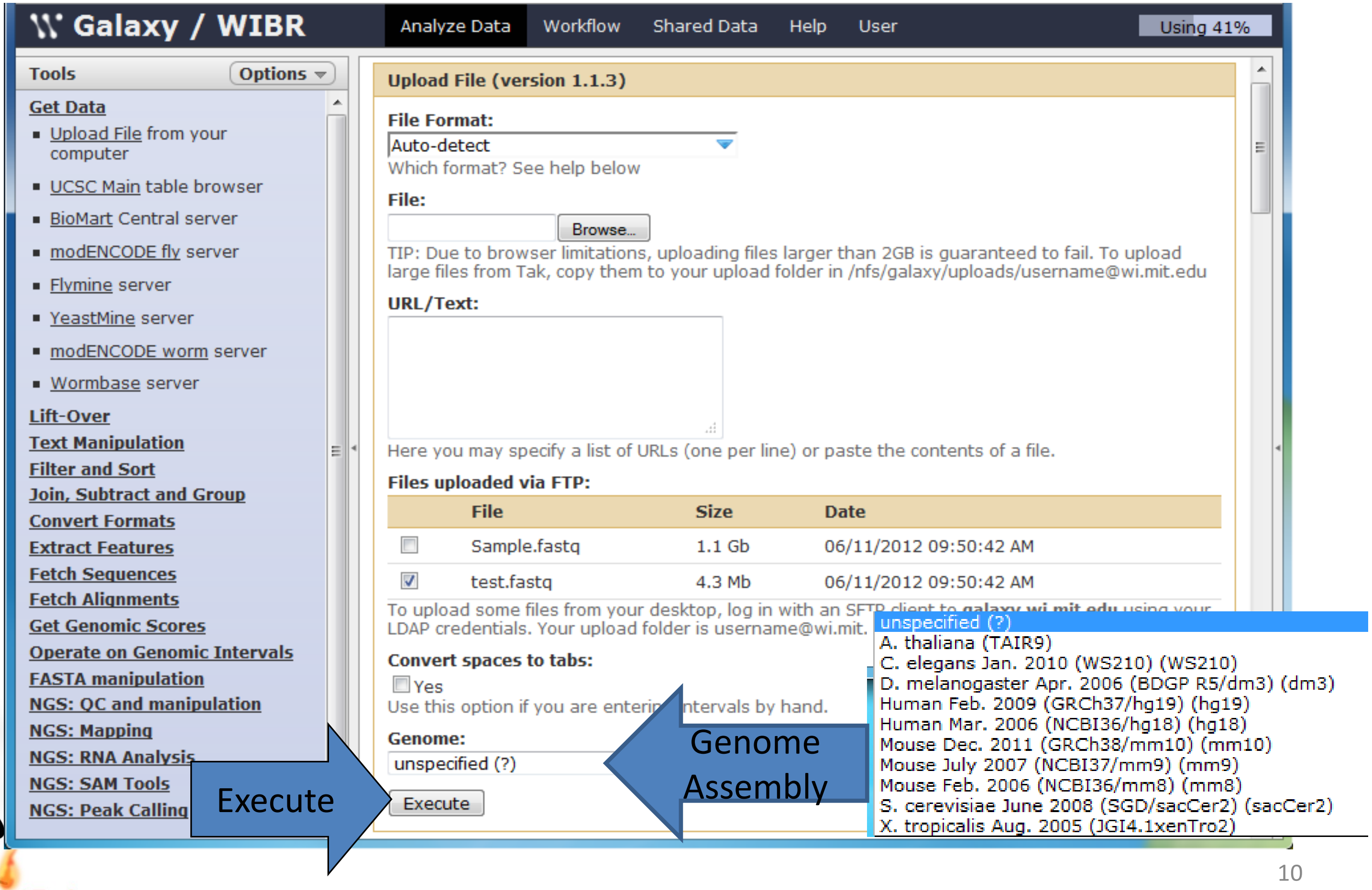


/nfs/galaxy/uploads/username@wi.mit.edu

Filename	Size	Modified
Sample.fastq	1.1 GB	5/14/12 5:22 PM
test.fastq	4.3 MB	4/9/12 1:06 PM

Getting Data: Uploading Large Files

Step 2: Select and upload the file within galaxy



Upload File (version 1.1.3)

File Format: Auto-detect

Which format? See help below

File:

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
□	Sample.fastq	1.1 Gb	06/11/2012 09:50:42 AM
☑	test.fastq	4.3 Mb	06/11/2012 09:50:42 AM

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.

Convert spaces to tabs:

□ Yes

Use this option if you are enteri ntervals by hand.

Genome: unspecified (?)

- 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)

Execute

Genome Assembly

Getting Data from UCSC (local copy)

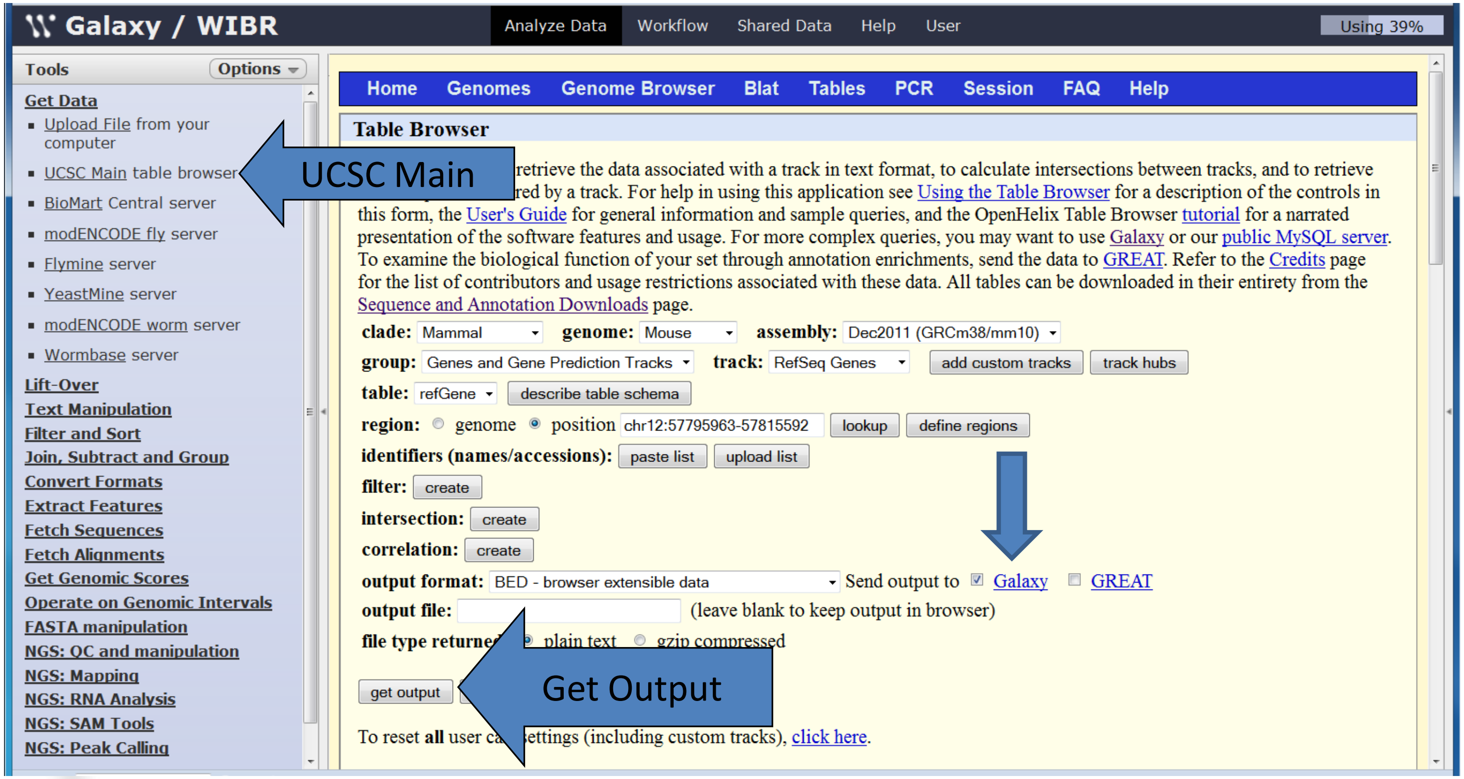


Table Browser

retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve red 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:** Dec2011 (GRCm38/mm10)

group: Genes and Gene Prediction Tracks **track:** RefSeq Genes [add custom tracks] [track hubs]

table: refGene [describe table schema]

region: ○ genome ● position chr12:57795963-57815592 [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)

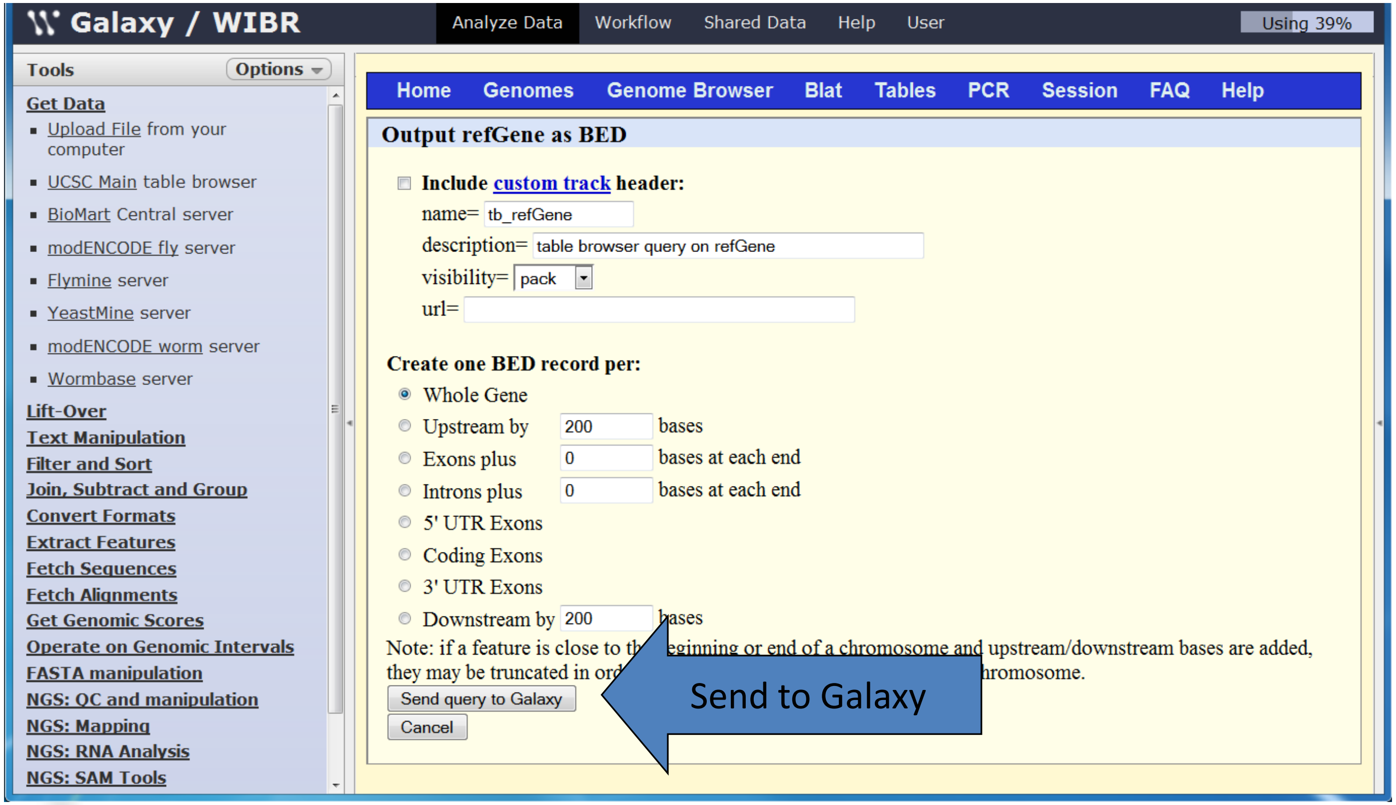
file type returne ● plain text ○ gzip compressed

[get output]

To reset **all** user ca ettings (including custom tracks), click here.

BaRC Hot Topics *Galaxy*

Getting Data from UCSC (local copy)



Output refGene as BED

□ **Include custom track header:**

name= tb_refGene

description= table browser query on refGene

visibility= pack

url=

Create one BED record per:

● Whole Gene

○ Upstream by 200 bases

○ Exons plus 0 bases at each end

○ Introns plus 0 bases at each end

○ 5' UTR Exons

○ Coding Exons

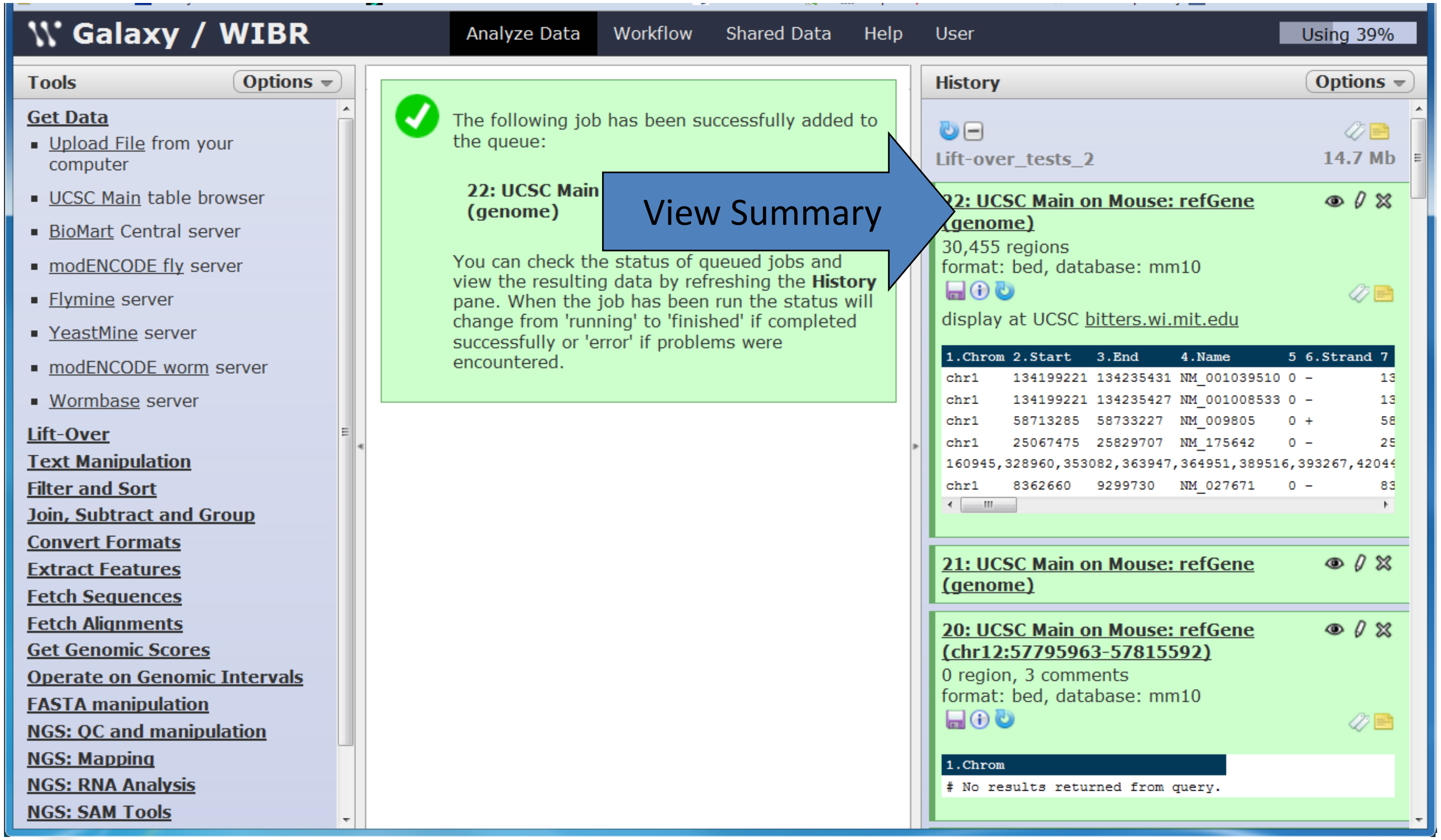
○ 3' UTR Exons

○ Downstream by 200 bases

Note: if a feature is close to th eginning or end of a chromosome and upstream/downstream bases are added, they may be truncated in ord hromosome.

Send to Galaxy

Data Uploaded



The following job has been successfully added to the queue:

22: UCSC Main (genome)

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

View Summary

Lift-over_tests_2 14.7 Mb

22: UCSC Main on Mouse: refGene (genome)

30,455 regions
format: bed, database: mm10

display at UCSC bitters.wi.mit.edu

```
1.Chrom 2.Start   3.End     4.Name       5 6.Strand 7
chr1    134199221 134235431 NM_001039510 0 -        13
chr1    134199221 134235427 NM_001008533 0 -        13
chr1    58713285  58733227  NM_009805    0 +        58
chr1    25067475  25829707  NM_175642    0 -        25
160945,328960,353082,363947,364951,389516,393267,42044
chr1    8362660   9299730   NM_027671    0 -        83
```

21: UCSC Main on Mouse: refGene (genome)

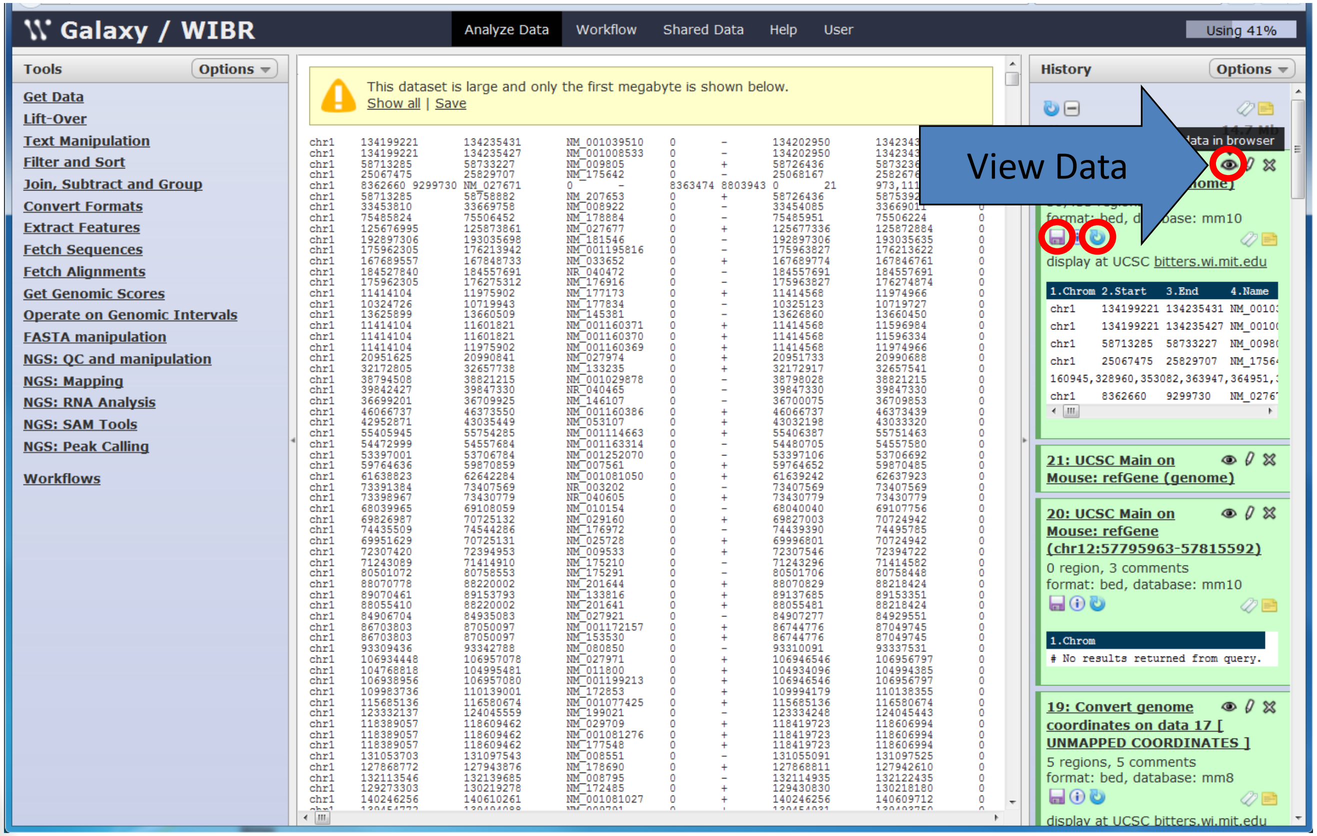
20: UCSC Main on Mouse: refGene (chr12:57795963-57815592)

0 region, 3 comments
format: bed, database: mm10

```
1.Chrom
# No results returned from query.
```

Adapted from OpenHelix tutorial

View Data



Galaxy / WIBR — Analyze Data | Workflow | Shared Data | Help | User — Using 41%

Tools — Options

- 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
- Workflows

This dataset is large and only the first megabyte is shown below.
Show all | Save

```
chr1    134199221       134235431       NM_001039510    0       -       134202950       1342343
chr1    134199221       134235427       NM_001008533    0       -       134202950       1342343
chr1    58713285        58733227        NM_009805       0       +       58726436        5873236
chr1    25067475        25829707        NM_175642       0       -       25068167        2582676
chr1    8362660 9299730 NM_027671       0       -       8363474 8803943 0       21      973,111
chr1    58713285        58758882        NM_207653       0       +       58726436        5875392
chr1    33453810        33669758        NM_008922       0       -       33454085        33669011        0
chr1    75485824        75506452        NM_178884       0       -       75485951        75506224        0
chr1    125676995       125873861       NM_027677       0       +       125677336       125872884       0
chr1    192897306       193035698       NM_181546       0       -       192897306       193035635       0
chr1    175962305       176213942       NM_001195816    0       -       175963827       176213622       0
chr1    167689557       167848733       NM_033652       0       +       167689774       167846761       0
chr1    184527840       184557691       NR_040472       0       -       184557691       184557691       0
chr1    175962305       176275312       NM_176916       0       -       175963827       176274874       0
chr1    11414104        11975902        NM_177173       0       +       11414568        11974966        0
chr1    10324726        10719943        NM_177834       0       -       10325123        10719727        0
chr1    13625899        13660509        NM_145381       0       -       13626860        13660450        0
chr1    11414104        11601821        NM_001160371    0       +       11414568        11596984        0
chr1    11414104        11601821        NM_001160370    0       +       11414568        11596334        0
chr1    11414104        11975902        NM_001160369    0       +       11414568        11974966        0
chr1    20951625        20990841        NM_027974       0       +       20951733        20990688        0
chr1    32172805        32657738        NM_133235       0       +       32172917        32657541        0
chr1    38794508        38821215        NM_001029878    0       -       38798028        38821215        0
chr1    39842427        39847330        NR_040465       0       -       39847330        39847330        0
chr1    36699201        36709925        NM_146107       0       -       36700075        36709853        0
chr1    46066737        46373550        NM_001160386    0       +       46066737        46373439        0
chr1    42952871        43035449        NM_053107       0       +       43032198        43033320        0
chr1    55405945        55754285        NM_001114663    0       +       55406387        55751463        0
chr1    54472999        54557684        NM_001163314    0       -       54480705        54557580        0
chr1    53397001        53706784        NM_001252070    0       -       53397106        53706692        0
chr1    59764636        59870859        NM_007561       0       +       59764652        59870485        0
chr1    61638823        62642284        NM_001081050    0       +       61639242        62637923        0
chr1    73391384        73407569        NR_003202       0       -       73407569        73407569        0
chr1    73398967        73430779        NR_040605       0       +       73430779        73430779        0
chr1    68039965        69108059        NM_010154       0       -       68040040        69107756        0
chr1    69826987        70725132        NM_029160       0       +       69827003        70724942        0
chr1    74435509        74544286        NM_176972       0       -       74439390        74495785        0
chr1    69951629        70725131        NM_025728       0       +       69996801        70724942        0
chr1    72307420        72394953        NM_009533       0       +       72307546        72394722        0
chr1    71243089        71414910        NM_175210       0       -       71243296        71414582        0
chr1    80501072        80758553        NM_175291       0       -       80501706        80758448        0
chr1    88070778        88220002        NM_201644       0       +       88070829        88218424        0
chr1    89070461        89153793        NM_133816       0       +       89137685        89153351        0
chr1    88055410        88220002        NM_201641       0       +       88055481        88218424        0
chr1    84906704        84935083        NM_027921       0       -       84907277        84929551        0
chr1    86703803        87050097        NM_001172157    0       +       86744776        87049745        0
chr1    86703803        87050097        NM_153530       0       +       86744776        87049745        0
chr1    93309436        93342788        NM_080850       0       -       93310091        93337531        0
chr1    106934448       106957078       NM_027971       0       +       106946546       106956797       0
chr1    104768818       104995481       NM_011800       0       +       104934096       104994385       0
chr1    106938956       106957080       NM_001199213    0       +       106946546       106956797       0
chr1    109983736       110139001       NM_172853       0       +       109994179       110138355       0
chr1    115685136       116580674       NM_001077425    0       +       115685136       116580674       0
chr1    123332137       124045559       NM_199021       0       -       123334248       124045443       0
chr1    118389057       118609462       NM_029709       0       +       118419723       118606994       0
chr1    118389057       118609462       NM_001081276    0       +       118419723       118606994       0
chr1    118389057       118609462       NM_177548       0       +       118419723       118606994       0
chr1    131053703       131097543       NM_008551       0       -       131055091       131097525       0
chr1    127868772       127943876       NM_178690       0       +       127868811       127942610       0
chr1    132113546       132139685       NM_008795       0       -       132114935       132122435       0
chr1    129273303       130219278       NM_172485       0       +       129430830       130218180       0
chr1    140246256       140610261       NM_001081027    0       +       140246256       140609712       0
```

History — Options

14.7 Mb

data in browser

format: bed, d[ata]base: mm10

display at UCSC bitters.wi.mit.edu

```
1.Chrom 2.Start    3.End      4.Name
chr1    134199221  134235431  NM_0010
chr1    134199221  134235427  NM_0010
chr1    58713285   58733227   NM_0098
chr1    25067475   25829707   NM_1756
160945,328960,353082,363947,364951,
chr1    8362660    9299730    NM_0276
```

21: UCSC Main on Mouse: refGene (genome)

20: UCSC Main on Mouse: refGene (chr12:57795963-57815592)
0 region, 3 comments
format: bed, database: mm10

```
1.Chrom
# No results returned from query.
```

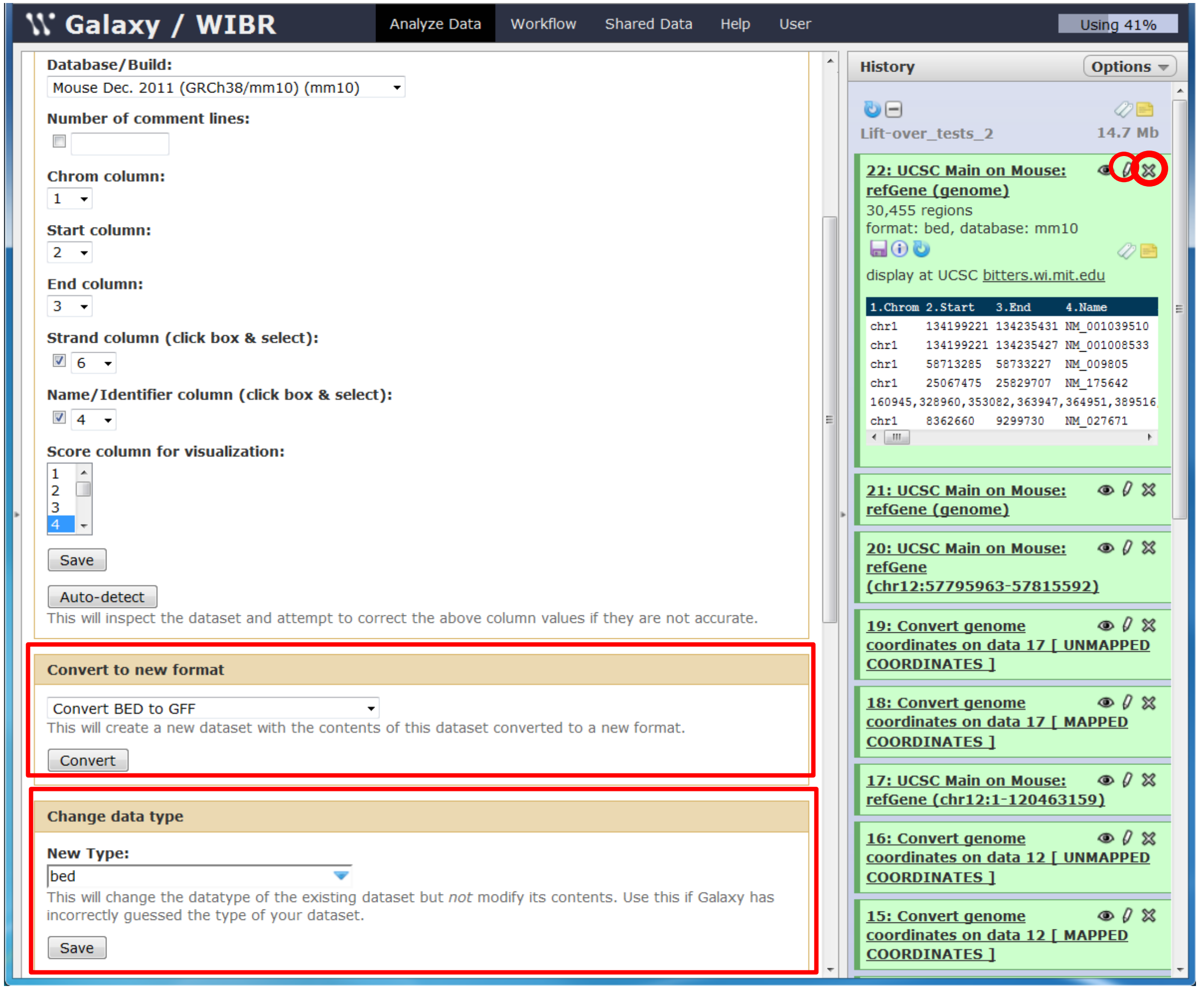
19: Convert genome coordinates on data 17 [UNMAPPED COORDINATES]
5 regions, 5 comments
format: bed, database: mm8

display at UCSC bitters.wi.mit.edu

View Data

Adapted from OpenHelix tutorial

Edit Attributes



Database/Build: Mouse Dec. 2011 (GRCh38/mm10) (mm10)

Number of comment lines: □

Chrom column: 1

Start column: 2

End column: 3

Strand column (click box & select): ☑ 6

Name/Identifier column (click box & select): ☑ 4

Score column for visualization: 1, 2, 3, 4

Save

Auto-detect

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

Convert to new format

Convert BED to GFF

This will create a new dataset with the contents of this dataset converted to a new format.

Convert

Change data type

New Type: bed

This will change the datatype of the existing dataset but *not* modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

Save

History

Lift-over_tests_2 14.7 Mb

22: UCSC Main on Mouse: refGene (genome)

30,455 regions

format: bed, database: mm10

display at UCSC bitters.wi.mit.edu

```
1.Chrom 2.Start   3.End     4.Name
chr1    134199221 134235431 NM_001039510
chr1    134199221 134235427 NM_001008533
chr1    58713285  58733227  NM_009805
chr1    25067475  25829707  NM_175642
160945,328960,353082,363947,364951,389516
chr1    8362660   9299730   NM_027671
```

21: UCSC Main on Mouse: refGene (genome)

20: UCSC Main on Mouse: refGene (chr12:57795963-57815592)

19: Convert genome coordinates on data 17 [UNMAPPED COORDINATES]

18: Convert genome coordinates on data 17 [MAPPED COORDINATES]

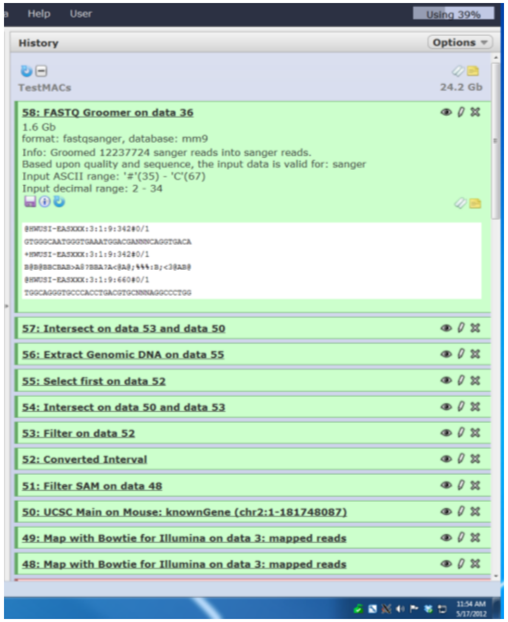
17: UCSC Main on Mouse: refGene (chr12:1-120463159)

16: Convert genome coordinates on data 12 [UNMAPPED COORDINATES]

15: Convert genome coordinates on data 12 [MAPPED COORDINATES]

History

- All steps are saved.
- Every time we do a new operation a new dataset is created. Data is not overwritten.
- Can share history with other Galaxy users.
- Can create workflow to repeat an analysis.



TestMACs 24.2 Gb

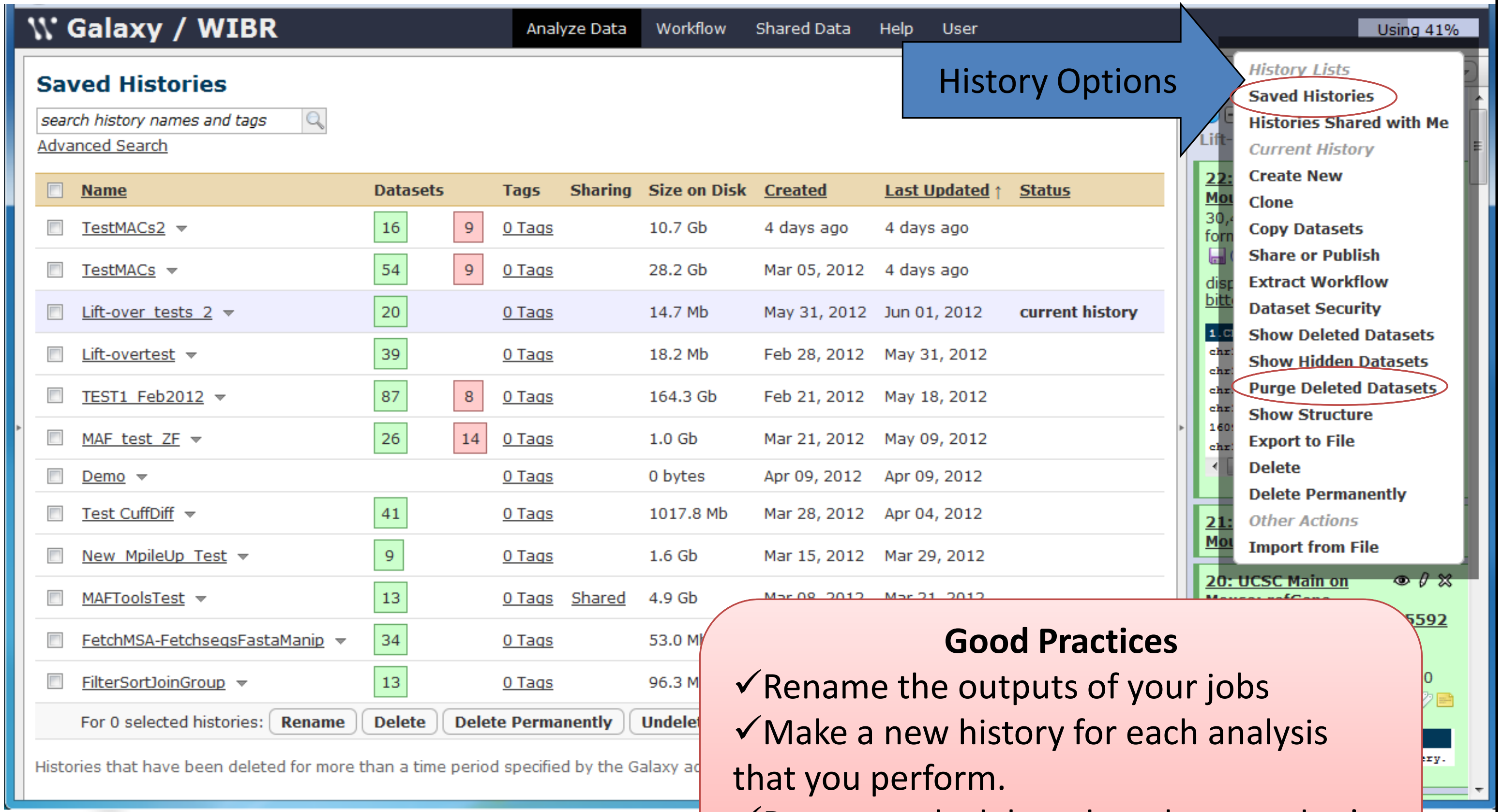
58: FASTQ Groomer on data 36

1.6 Gb
format: fastqsanger, database: mm9
Info: Groomed 12237724 sanger reads into sanger reads.
Based upon quality and sequence, the input data is valid for: sanger
Input ASCII range: '#'(35) - 'C'(67)
Input decimal range: 2 - 34

```
@HWUSI-EASXXX:3:1:9:342#0/1
GTGGGCAATGGGTGAAATGGACGANNNCAGGTGACA
+HWUSI-EASXXX:3:1:9:342#0/1
B@B@BBCBAB>A87BBA?A<@A@;%%%:B;<3@AB@
@HWUSI-EASXXX:3:1:9:660#0/1
TGGCAGGGTGCCCACCTGACGTGCNNNAGGCCCTGG
```

57: Intersect on data 53 and data 50
56: Extract Genomic DNA on data 55
55: Select first on data 52
54: Intersect on data 50 and data 53
53: Filter on data 52
52: Converted Interval
51: Filter SAM on data 48
50: UCSC Main on Mouse: knownGene (chr2:1-181748087)
49: Map with Bowtie for Illumina on data 3: mapped reads
48: Map with Bowtie for Illumina on data 3: mapped reads

History



History Options

Saved Histories

search history names and tags

Advanced Search

☐	Name	Datasets		Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
☐	TestMACs2	16	9	0 Tags		10.7 Gb	4 days ago	4 days ago	
☐	TestMACs	54	9	0 Tags		28.2 Gb	Mar 05, 2012	4 days ago	
☐	Lift-over_tests_2	20		0 Tags		14.7 Mb	May 31, 2012	Jun 01, 2012	current history
☐	Lift-overtest	39		0 Tags		18.2 Mb	Feb 28, 2012	May 31, 2012	
☐	TEST1_Feb2012	87	8	0 Tags		164.3 Gb	Feb 21, 2012	May 18, 2012	
☐	MAF_test_ZF	26	14	0 Tags		1.0 Gb	Mar 21, 2012	May 09, 2012	
☐	Demo			0 Tags		0 bytes	Apr 09, 2012	Apr 09, 2012	
☐	Test CuffDiff	41		0 Tags		1017.8 Mb	Mar 28, 2012	Apr 04, 2012	
☐	New_MpileUp_Test	9		0 Tags		1.6 Gb	Mar 15, 2012	Mar 29, 2012	
☐	MAFToolsTest	13		0 Tags	Shared	4.9 Gb	Mar 08, 2012	Mar 21, 2012	
☐	FetchMSA-FetchseqsFastaManip	34		0 Tags		53.0 M			
☐	FilterSortJoinGroup	13		0 Tags		96.3 M			

For 0 selected histories: Rename Delete Delete Permanently Undelet

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

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

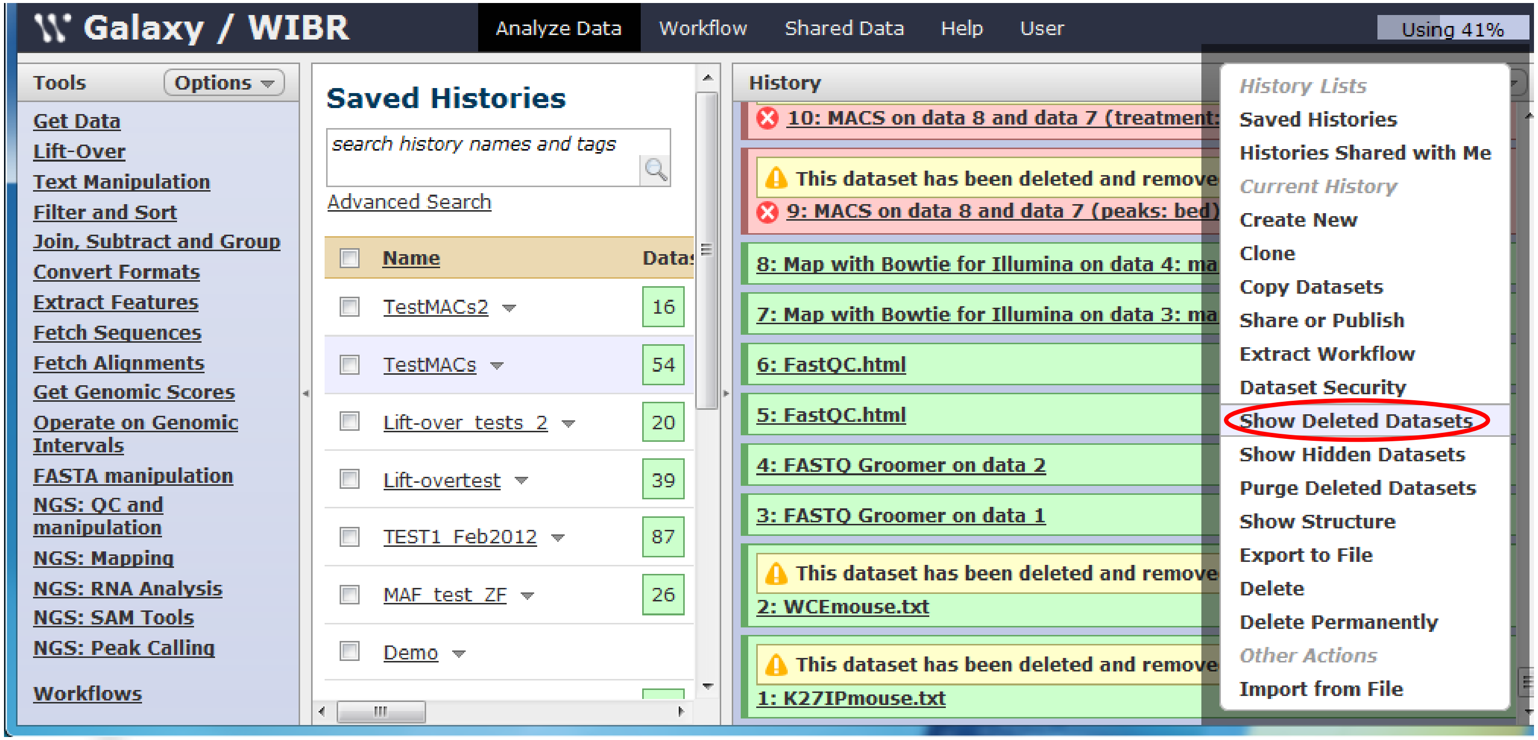
Good Practices

✓Rename the outputs of your jobs

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

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

History is not removed when datasets are removed



Saved Histories

search history names and tags

Advanced Search

□	Name	Datas
□	TestMACs2	16
□	TestMACs	54
□	Lift-over_tests_2	20
□	Lift-overtest	39
□	TEST1_Feb2012	87
□	MAF_test_ZF	26
□	Demo	

10: MACS on data 8 and data 7 (treatment:

This dataset has been deleted and remove

9: MACS on data 8 and data 7 (peaks: bed)

8: Map with Bowtie for Illumina on data 4: ma

7: Map with Bowtie for Illumina on data 3: ma

6: FastQC.html

5: FastQC.html

4: FASTQ Groomer on data 2

3: FASTQ Groomer on data 1

This dataset has been deleted and remove

2: WCEmouse.txt

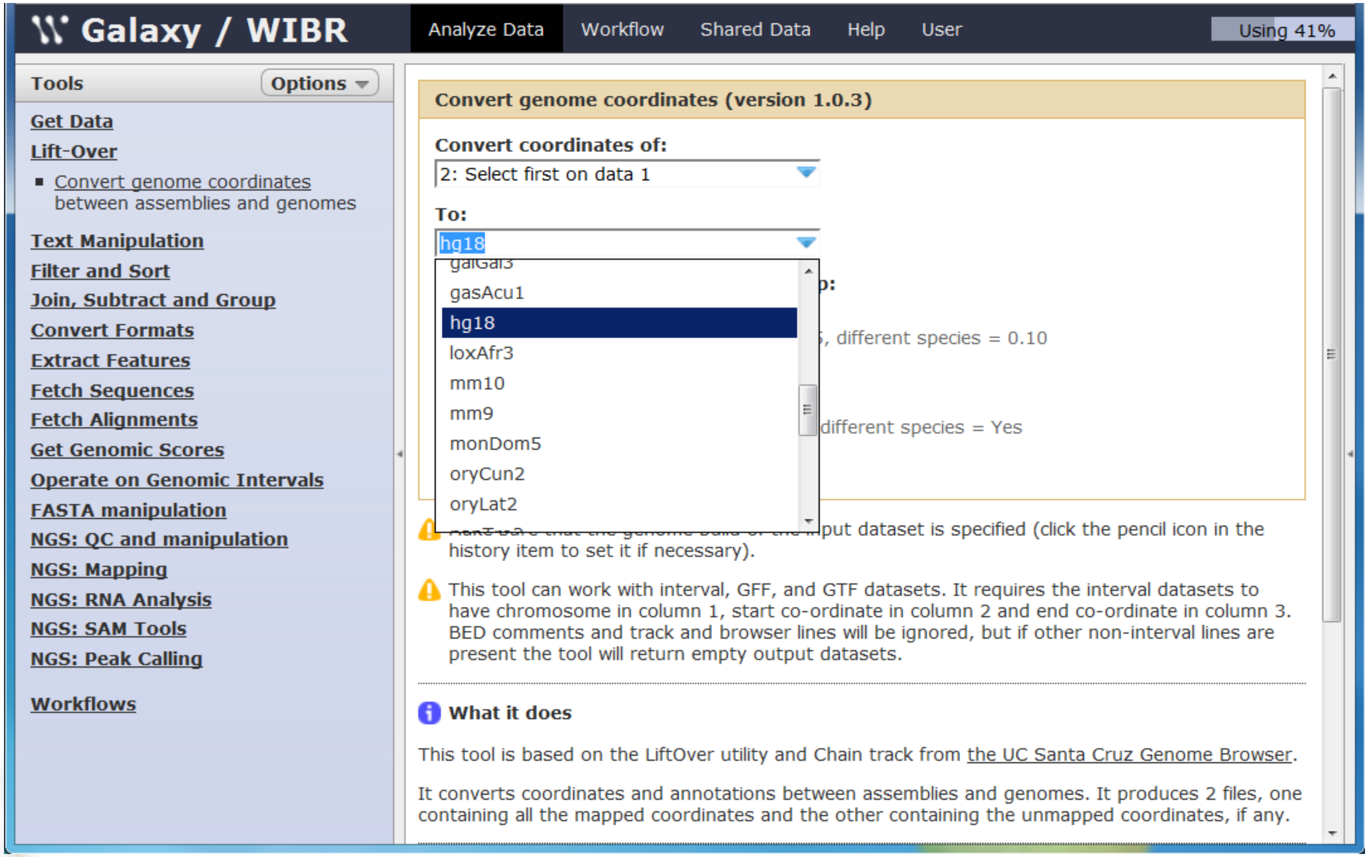
This dataset has been deleted and remove

1: K27IPmouse.txt

Talk Outline

- The Galaxy interface
- Getting data into Galaxy
- Overview of the tools
- The Next Generation Sequencing tool box
  - Preprocessing and quality control
  - Analysis of ChIP-seq
  - Analysis of RNA-seq
- Visualizing data on a genome browser and workflows available for analysis

Overview of the tools: Lift-Over



Convert genome coordinates (version 1.0.3)

Convert coordinates of:
2: Select first on data 1

To:
hg18

p:

, different species = 0.10

different species = Yes

put dataset is specified (click the pencil icon in the history item to set it if necessary).

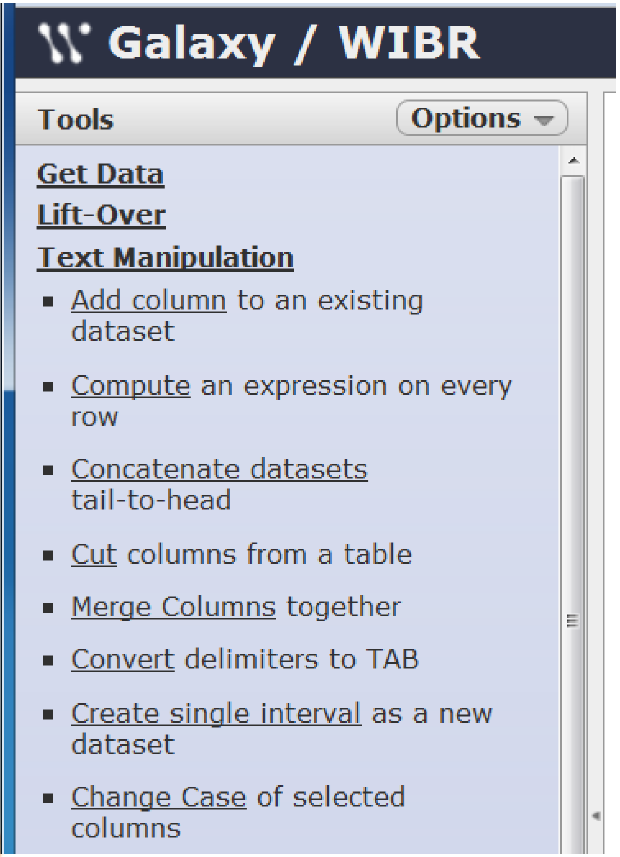
This tool can work with interval, GFF, and GTF datasets. It requires the interval datasets to have chromosome in column 1, start co-ordinate in column 2 and end co-ordinate in column 3. BED comments and track and browser lines will be ignored, but if other non-interval lines are present the tool will return empty output datasets.

What it does

This tool is based on the LiftOver utility and Chain track from the UC Santa Cruz Genome Browser.

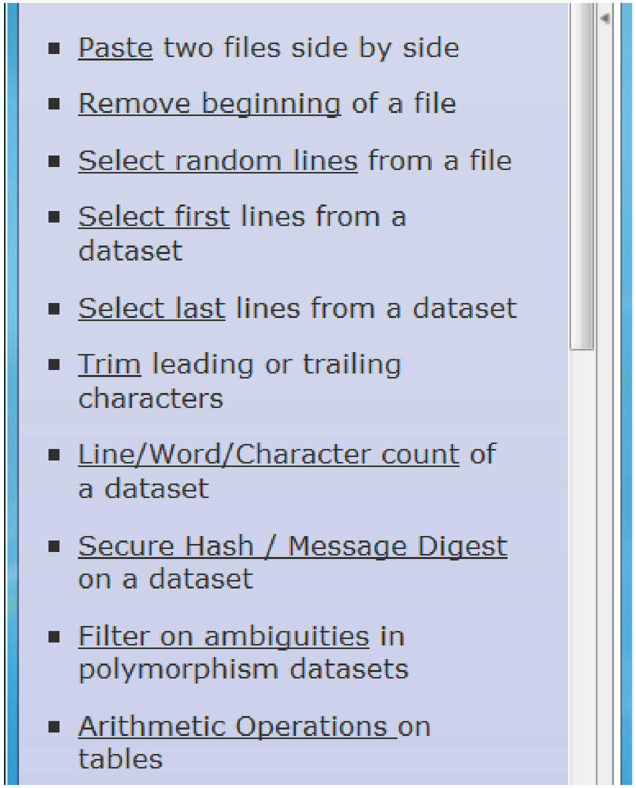
It converts coordinates and annotations between assemblies and genomes. It produces 2 files, one containing all the mapped coordinates and the other containing the unmapped coordinates, if any.

Text Manipulation



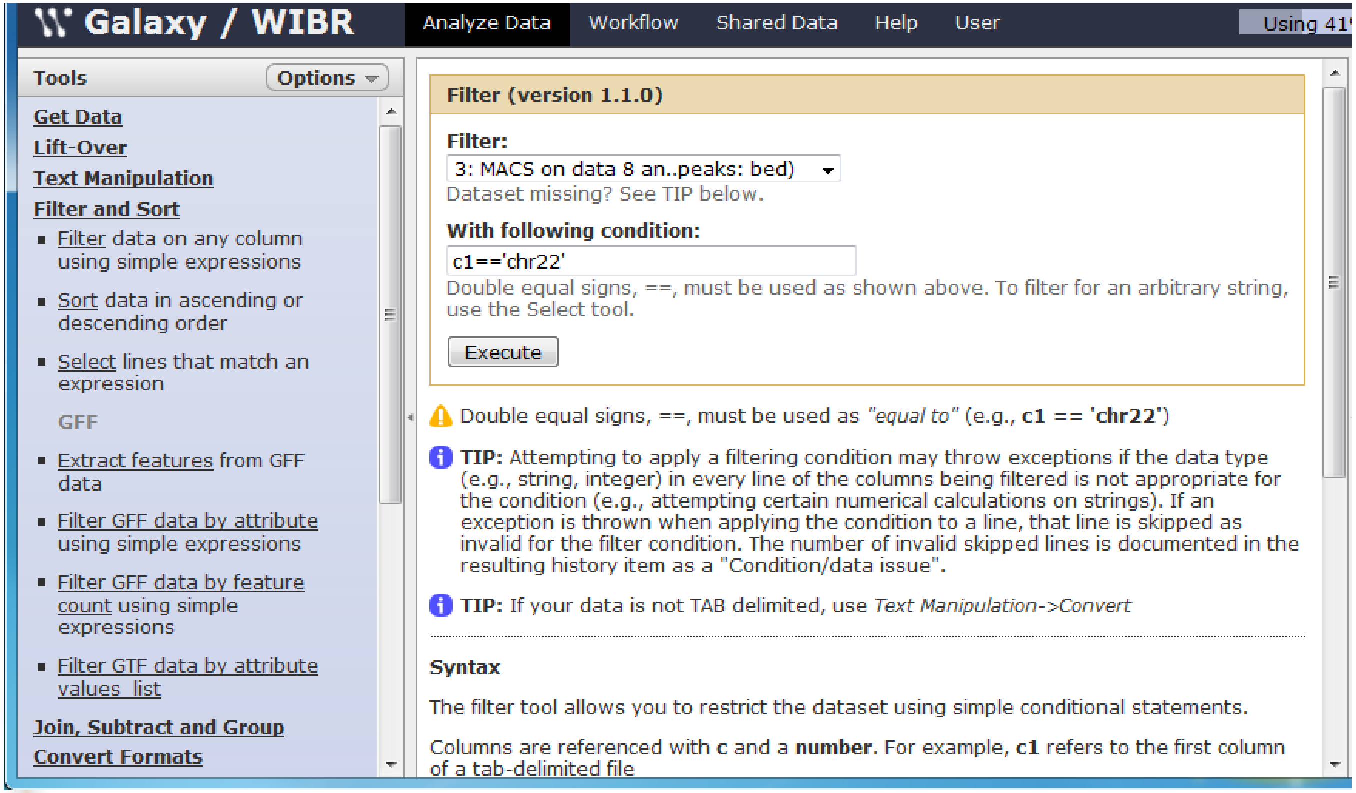
Get Data
Lift-Over
Text Manipulation

- 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

Filter and Sort:
Filter data on any column


Galaxy / WIBR
Analyze Data
Workflow
Shared Data
Help
User
Using 41
Tools
Options
Get Data
Lift-Over
Text Manipulation
Filter and Sort
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF data
Filter GFF data by attribute using simple expressions
Filter GFF data by feature count using simple expressions
Filter GTF data by attribute values_list
Join, Subtract and Group
Convert Formats
Execute

Filter (version 1.1.0)

Filter:
3: MACS on data 8 an..peaks: bed)
Dataset missing? See TIP below.

With following condition:
c1=='chr22'
Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

⚠ Double equal signs, ==, must be used as *"equal to"* (e.g., **c1 == 'chr22'**)

ℹ **TIP:** Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

ℹ **TIP:** If your data is not TAB delimited, use *Text Manipulation->Convert*

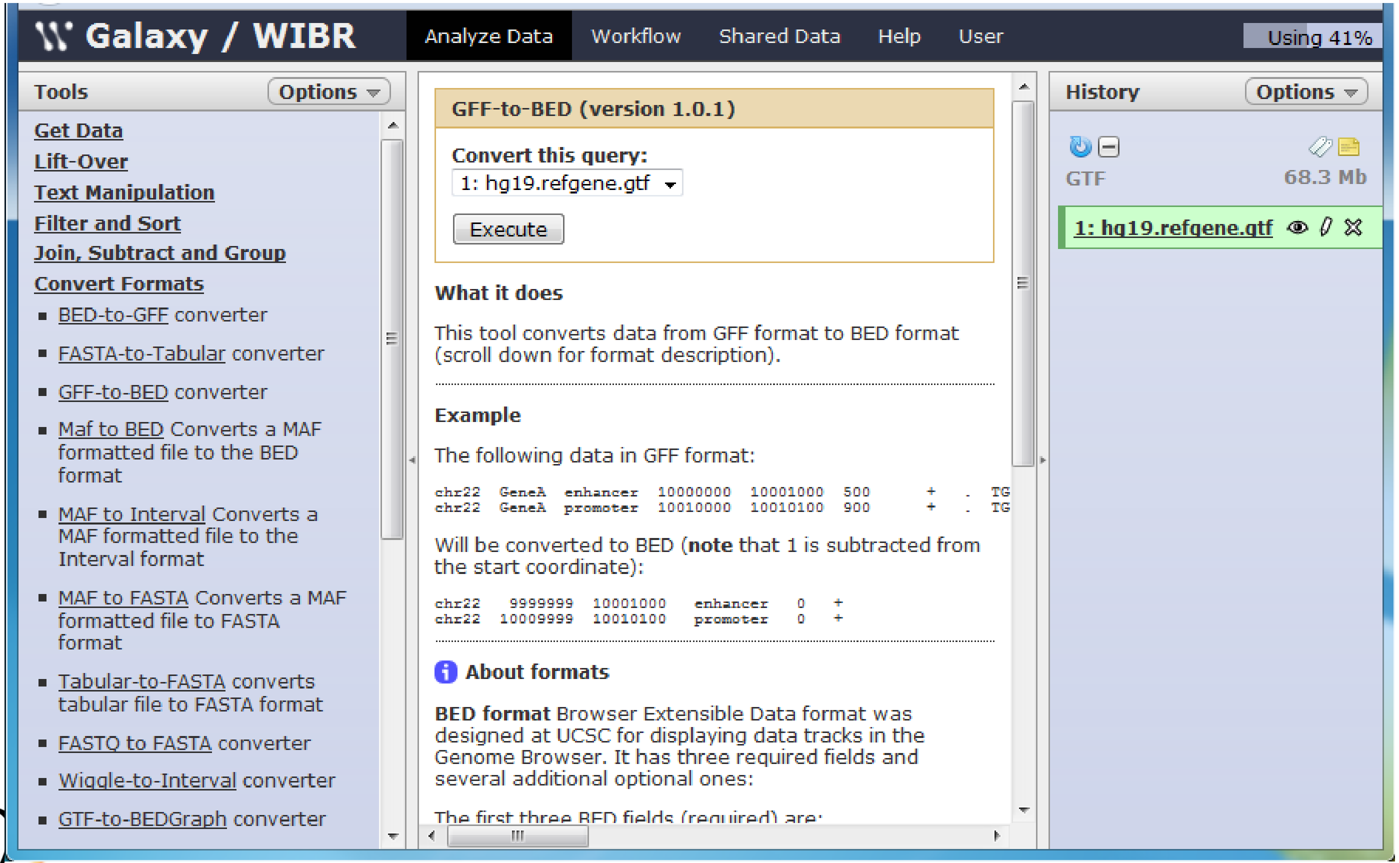
Syntax

The filter tool allows you to restrict the dataset using simple conditional statements.

Columns are referenced with **c** and a **number**. For example, **c1** refers to the first column of a tab-delimited file

BaRC Hot Topics *Galaxy*

Convert Formats:
GFF-to-BED



Get Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats

- BED-to-GFF converter
- FASTA-to-Tabular converter
- GFF-to-BED converter
- Maf to BED Converts a MAF formatted file to the BED format
- MAF to Interval Converts a MAF formatted file to the Interval format
- MAF to FASTA Converts a MAF formatted file to FASTA format
- Tabular-to-FASTA converts tabular file to FASTA format
- FASTQ to FASTA converter
- Wiggle-to-Interval converter
- GTF-to-BEDGraph converter

GFF-to-BED (version 1.0.1)

Convert this query: 1: hg19.refgene.gtf

Execute

What it does

This tool converts data from GFF format to BED format (scroll down for format description).

Example

The following data in GFF format:

```
chr22  GeneA  enhancer  10000000  10001000  500     +   .  TG
chr22  GeneA  promoter  10010000  10010100  900     +   .  TG
```

Will be converted to BED (**note** that 1 is subtracted from the start coordinate):

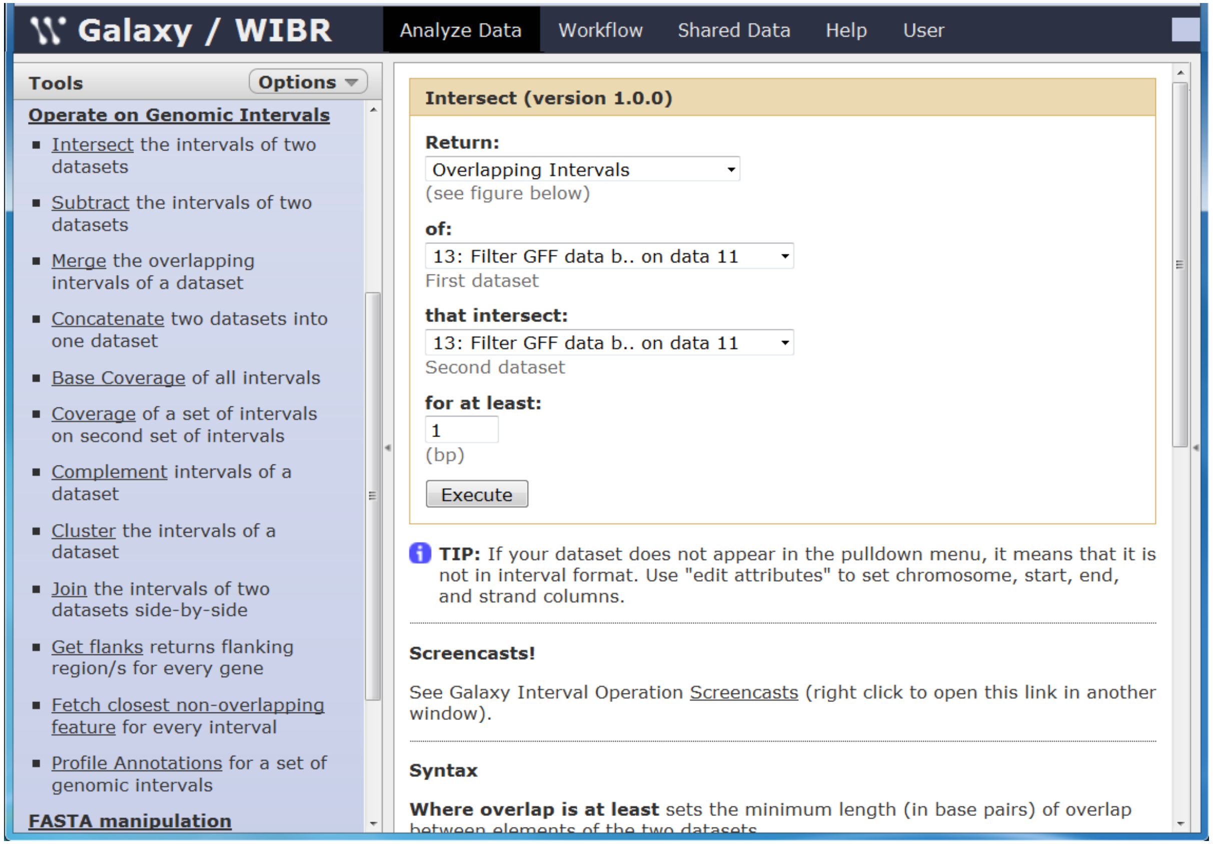
```
chr22   9999999  10001000   enhancer   0   +
chr22  10009999  10010100   promoter   0   +
```

About formats

BED format Browser Extensible Data format was designed at UCSC for displaying data tracks in the Genome Browser. It has three required fields and several additional optional ones:

The first three BED fields (required) are:

Operate on Genomic Intervals: Intersect the intervals of two datasets



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
- Profile Annotations for a set of genomic intervals

FASTA manipulation

Intersect (version 1.0.0)

Return:
Overlapping Intervals
(see figure below)

of:
13: Filter GFF data b.. on data 11
First dataset

that intersect:
13: Filter GFF data b.. on data 11
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.

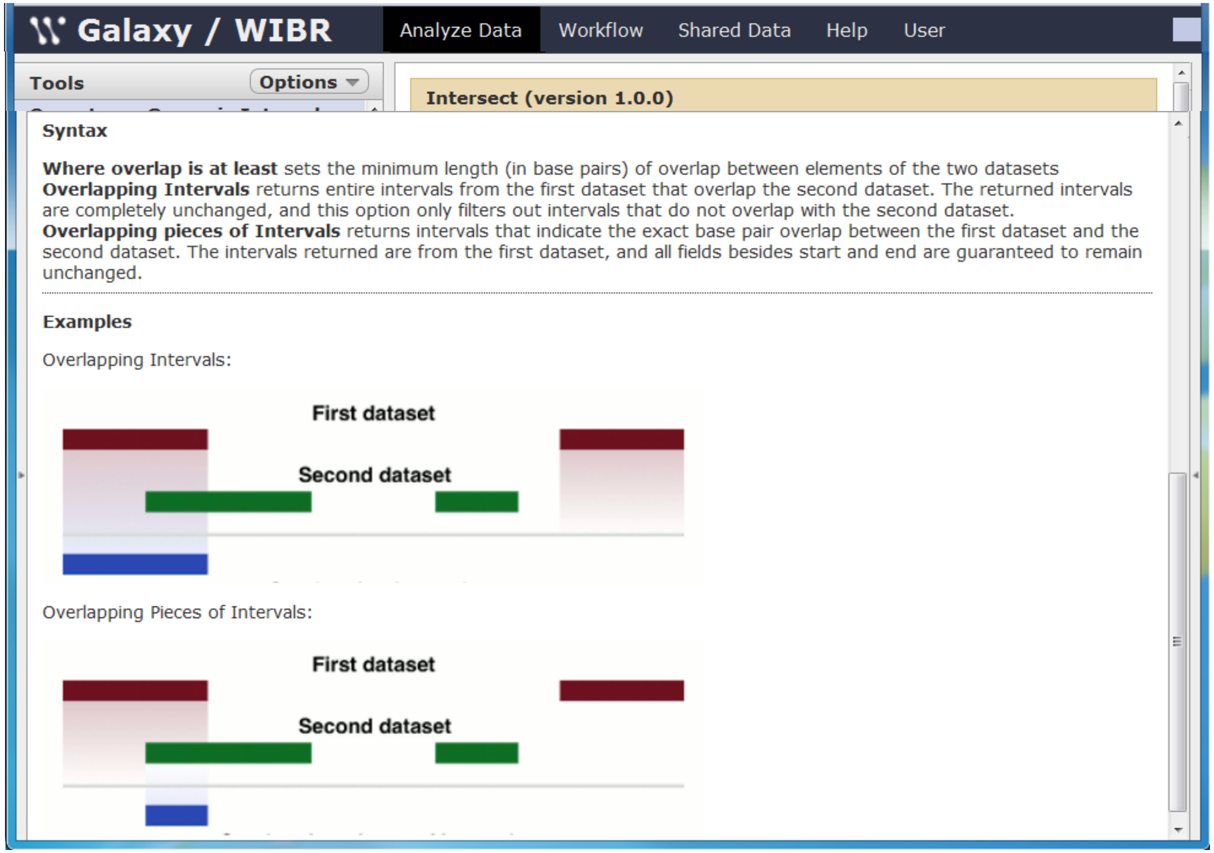
Screencasts!

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

Syntax

Where overlap is at least sets the minimum length (in base pairs) of overlap between elements of the two datasets

Operate on Genomic Intervals: Intersect the intervals of two datasets



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.

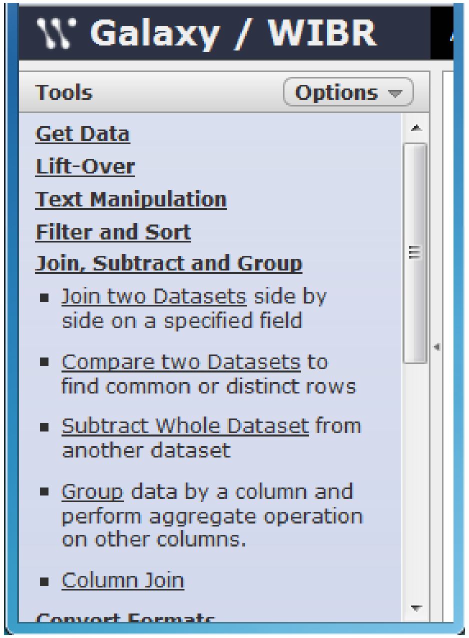
Examples

Overlapping Intervals:

Overlapping Pieces of Intervals:

Other tools

Join, Subtract and Group



Galaxy / WIBR

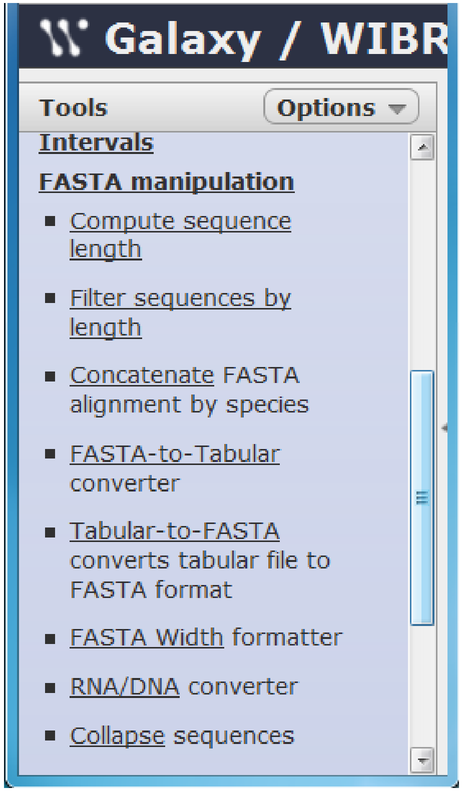
Tools Options

Get Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group

- Join two Datasets side by side on a specified field
- Compare two Datasets to find common or distinct rows
- Subtract Whole Dataset from another dataset
- Group data by a column and perform aggregate operation on other columns.
- Column Join

Convert Formats

FASTA manipulation



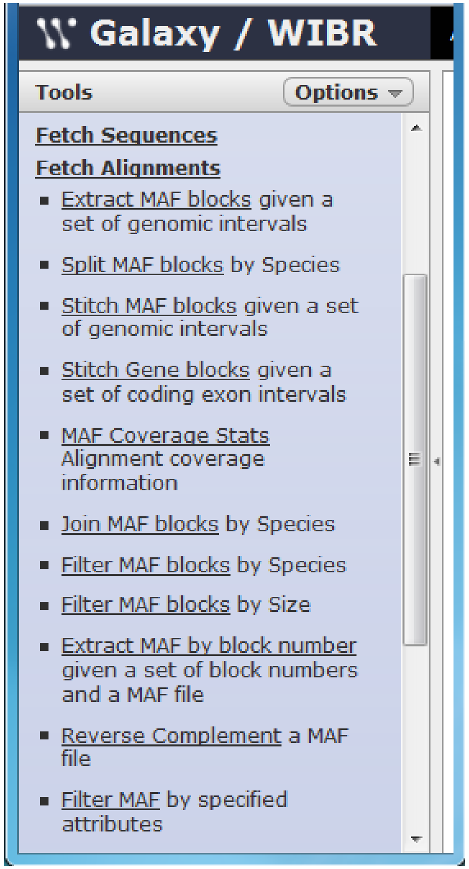
Galaxy / WIBR

Tools Options

Intervals
FASTA manipulation

- Compute sequence length
- Filter sequences by length
- Concatenate FASTA alignment by species
- FASTA-to-Tabular converter
- Tabular-to-FASTA converts tabular file to FASTA format
- FASTA Width formatter
- RNA/DNA converter
- Collapse sequences

Fetch Sequences and Fetch Alignments



Galaxy / WIBR

Tools Options

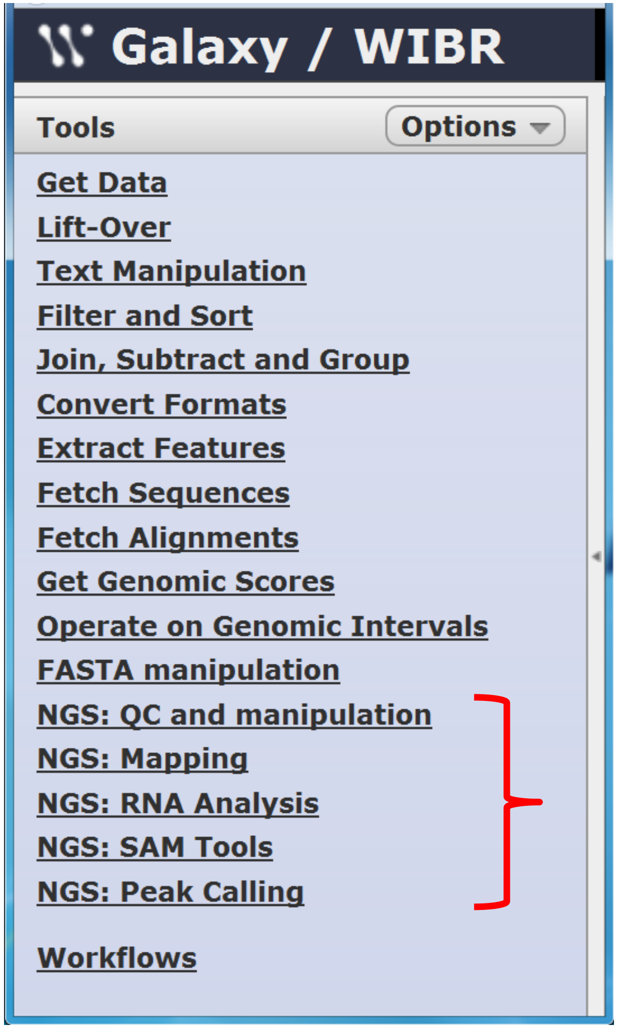
Fetch Sequences
Fetch Alignments

- Extract MAF blocks given a set of genomic intervals
- Split MAF blocks by Species
- Stitch MAF blocks given a set of genomic intervals
- Stitch Gene blocks given a set of coding exon intervals
- MAF Coverage Stats Alignment coverage information
- Join MAF blocks by Species
- Filter MAF blocks by Species
- Filter MAF blocks by Size
- Extract MAF by block number given a set of block numbers and a MAF file
- Reverse Complement a MAF file
- Filter MAF by specified attributes

Talk Outline

- The Galaxy interface
- Getting data into Galaxy
- Overview of the tools
- The Next Generation Sequencing tool box
  - Preprocessing and quality control
  - Analysis of ChIP-seq
  - Analysis of RNA-seq
- Visualizing data on a genome browser and workflows available for analysis

NGS Tools



Galaxy / WIBR

Tools

Options

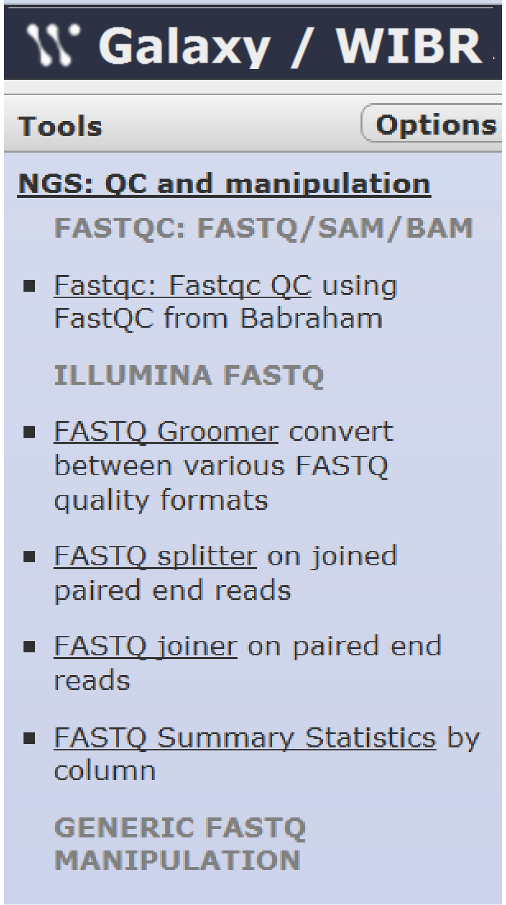
- Get Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: Peak Calling

Workflows

Next Generation Sequencing Tools

BaRC Hot Topics *Galaxy*

NGS: QC and manipulation



NGS: QC and manipulation

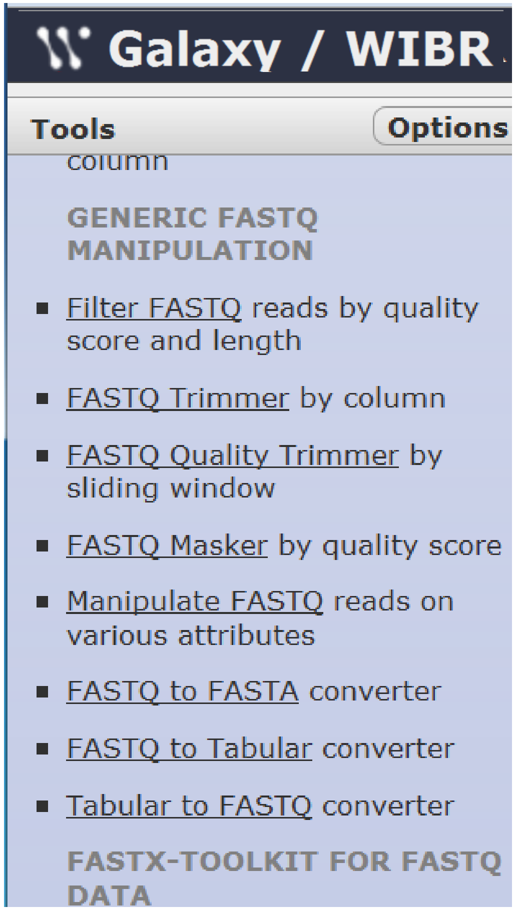
FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

ILLUMINA FASTQ

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

GENERIC FASTQ MANIPULATION

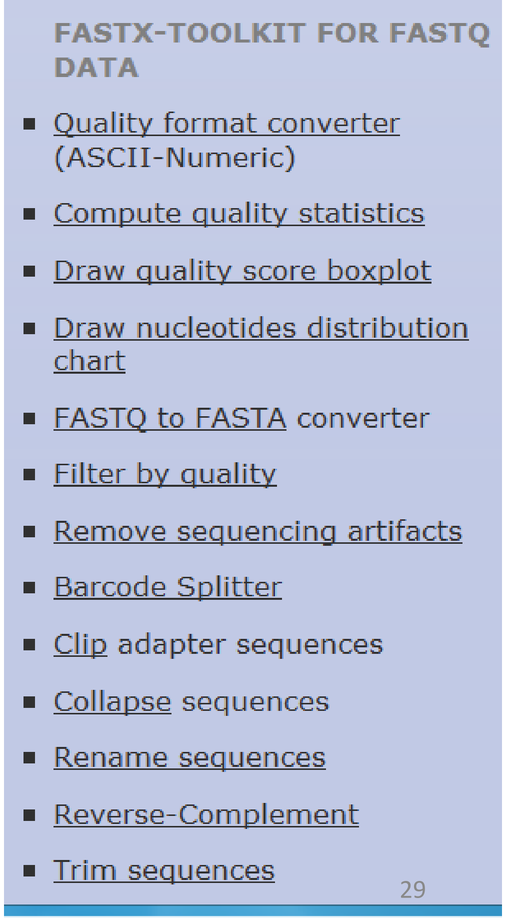


column

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

FASTX-TOOLKIT FOR FASTQ DATA

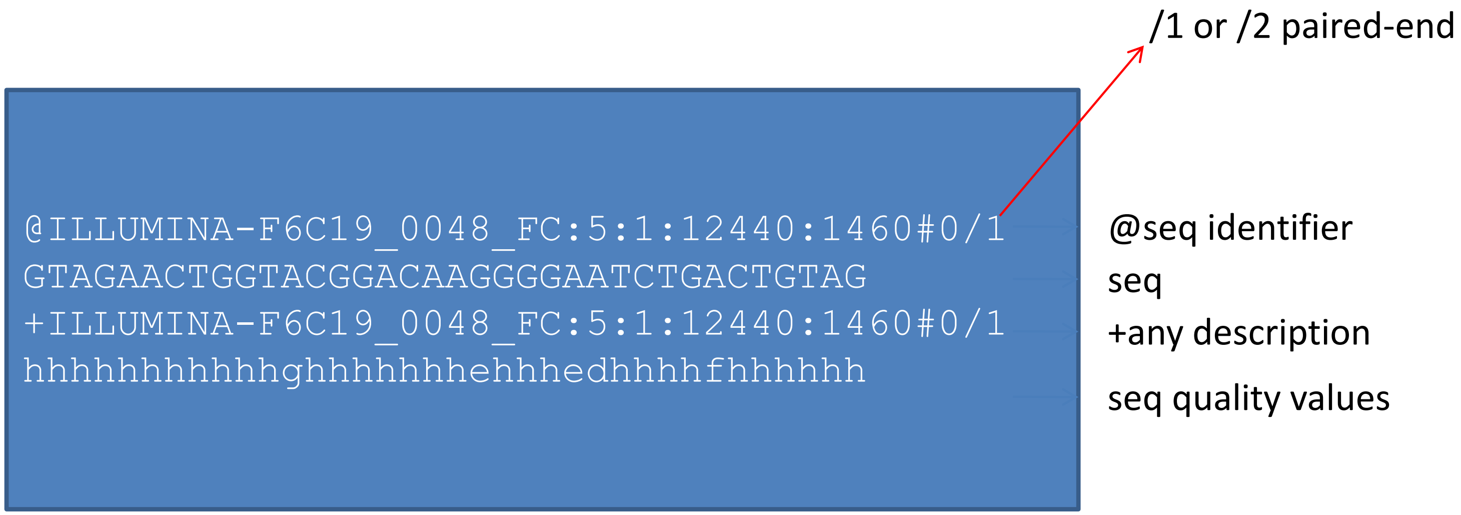


FASTX-TOOLKIT FOR FASTQ DATA

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Rename sequences
- Reverse-Complement
- Trim sequences

Illumina data format

- Fastq format:



/1 or /2 paired-end

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhghhhhhhhehhhedhhhhfhhhhhh
```

@seq identifier

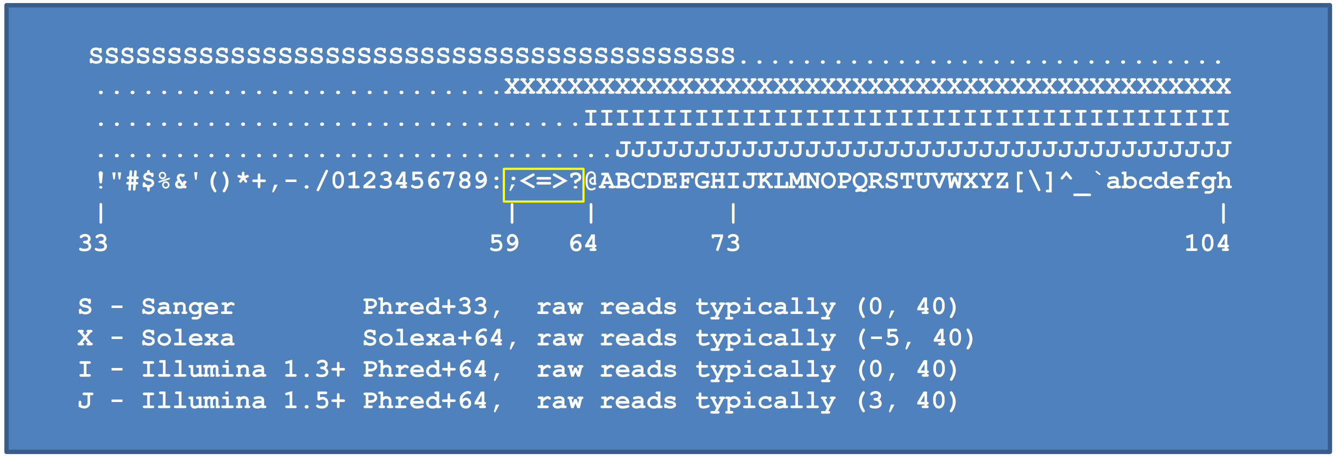
seq

+any description

seq quality values

Sequence quality values on different FASTQ formats

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



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS...............................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
 !"#$%&'()*+,-./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)
```

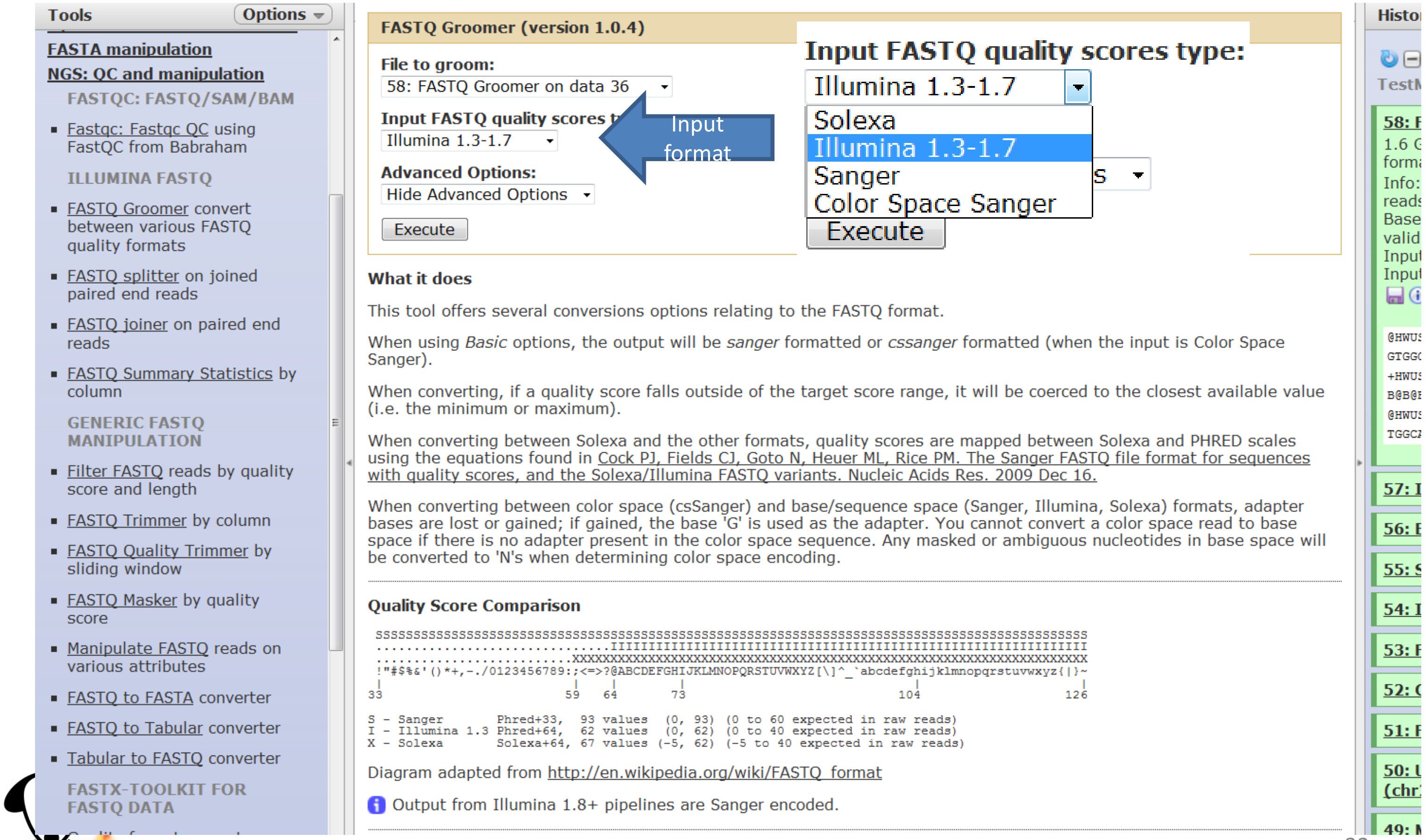
To discriminate between Solexa and Illumina 1.3+ check if your sequences' quality scores have any of the characters ;<=>?

BaRC Hot Topics *Galaxy*

FASTQ formats and FASTQ Groomer

ILLUMINA FASTQ

- FASTQ Groomer convert between various FASTQ quality formats



Tools
Options
FASTA manipulation
NGS: QC and manipulation
FASTQC: FASTQ/SAM/BAM
Fastqc: Fastqc QC using FastQC from Babraham
ILLUMINA FASTQ
FASTQ Groomer convert between various FASTQ quality formats
FASTQ splitter on joined paired end reads
FASTQ joiner on paired end reads
FASTQ Summary Statistics by column
GENERIC FASTQ MANIPULATION
Filter FASTQ reads by quality score and length
FASTQ Trimmer by column
FASTQ Quality Trimmer by sliding window
FASTQ Masker by quality score
Manipulate FASTQ reads on various attributes
FASTQ to FASTA converter
FASTQ to Tabular converter
Tabular to FASTQ converter
FASTX-TOOLKIT FOR FASTQ DATA
Histo
TestM
58: F
57: I
56: E
55: S
54: I
53: F
52: C
51: F
50: U (chr
49: M

FASTQ Groomer (version 1.0.4)

File to groom: 58: FASTQ Groomer on data 36

Input FASTQ quality scores t: Illumina 1.3-1.7

Input format

Advanced Options: Hide Advanced Options

Execute

Input FASTQ quality scores type: Illumina 1.3-1.7

Solexa
Illumina 1.3-1.7
Sanger
Color Space Sanger
Execute

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 sequence. Any masked or ambiguous nucleotides in base space will be converted to 'N's when determining color space encoding.

Quality Score Comparison

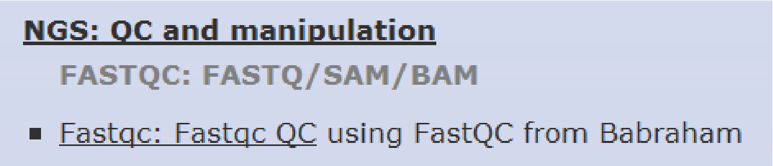
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3  Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa        Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

Diagram adapted from http://en.wikipedia.org/wiki/FASTQ_format

Output from Illumina 1.8+ pipelines are Sanger encoded.

BaRC Hot Topics Galaxy

NGS: Quality Control

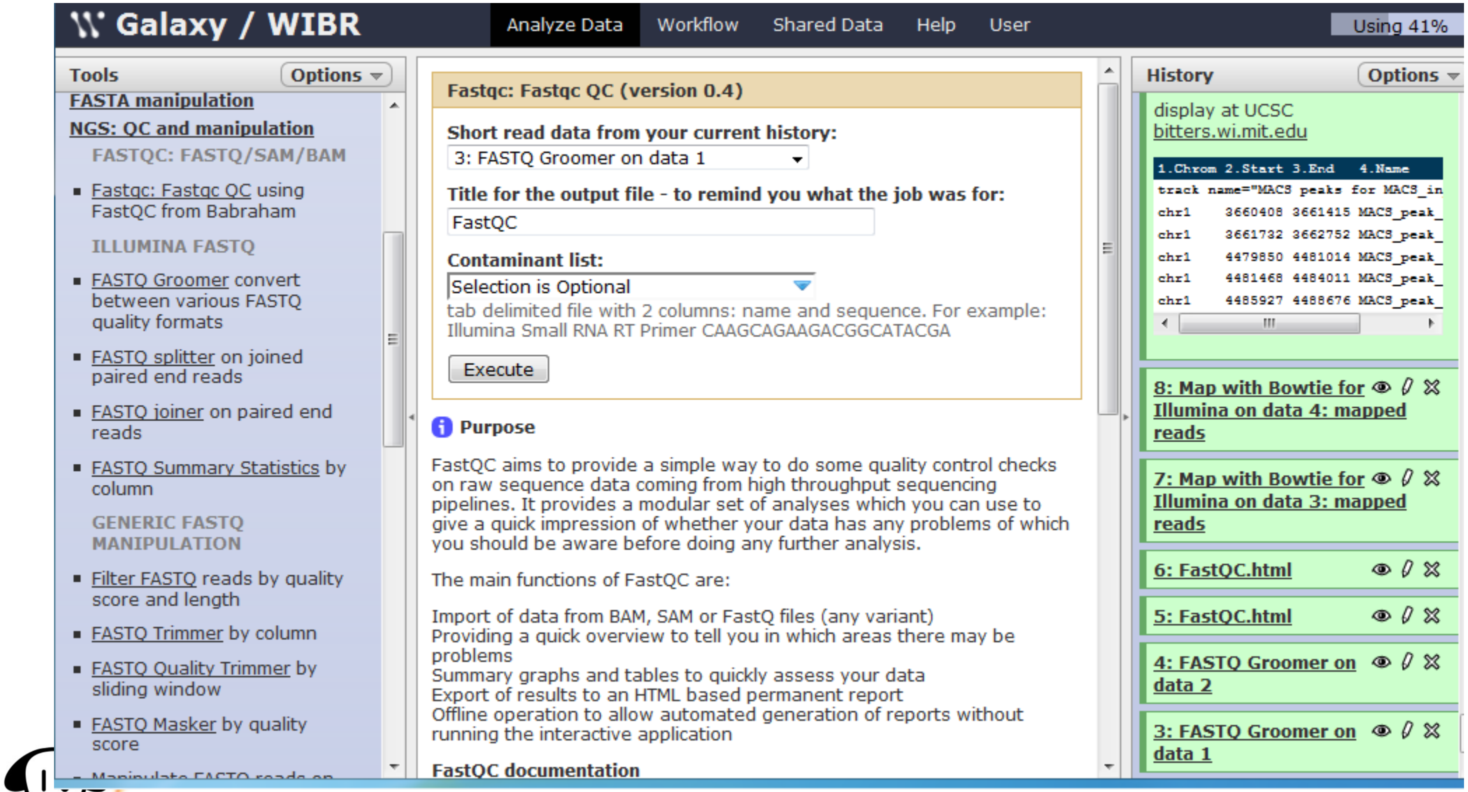


NGS: QC and manipulation

FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



FASTA manipulation

NGS: QC and manipulation

FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

ILLUMINA FASTQ

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- Manipulate FASTQ reads on

Fastqc: Fastqc QC (version 0.4)

Short read data from your current history:
3: FASTQ Groomer on data 1

Title for the output file - to remind you what the job was for:
FastQC

Contaminant list:
Selection is Optional
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Purpose

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

Import of data from BAM, SAM or FastQ files (any variant)
Providing a quick overview to tell you in which areas there may be problems
Summary graphs and tables to quickly assess your data
Export of results to an HTML based permanent report
Offline operation to allow automated generation of reports without running the interactive application

FastQC documentation

display at UCSC
bitters.wi.mit.edu

```
1.Chrom 2.Start 3.End   4.Name
track name="MACS peaks for MACS_in
chr1    3660408 3661415 MACS_peak_
chr1    3661732 3662752 MACS_peak_
chr1    4479850 4481014 MACS_peak_
chr1    4481468 4484011 MACS_peak_
chr1    4485927 4488676 MACS_peak_
```

8: Map with Bowtie for Illumina on data 4: mapped reads

7: Map with Bowtie for Illumina on data 3: mapped reads

6: FastQC.html

5: FastQC.html

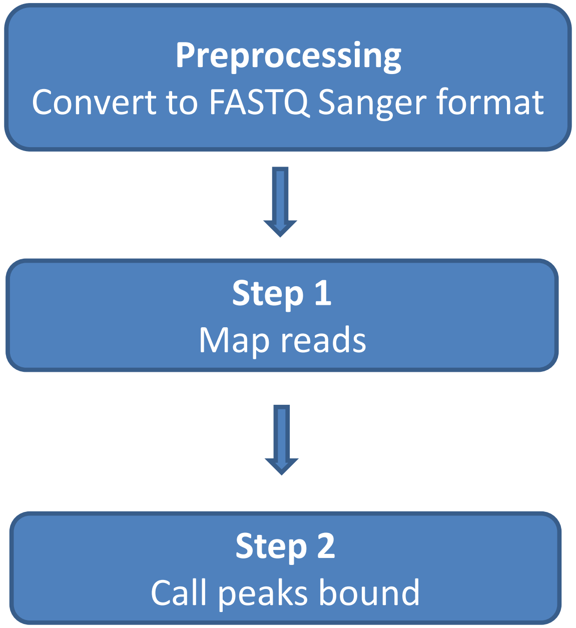
4: FASTQ Groomer on data 2

3: FASTQ Groomer on data 1

Talk Outline

- The Galaxy interface
- Getting data into Galaxy
- Overview of the tools
- The Next Generation Sequencing tool box
  - Preprocessing and quality control
  - Analysis of ChIP-seq
  - Analysis of RNA-seq
- Visualizing data on a genome browser and workflows available for analysis

Analysis of ChIP-seq experiments



Preprocessing
Convert to FASTQ Sanger format

FASTQ Groomer

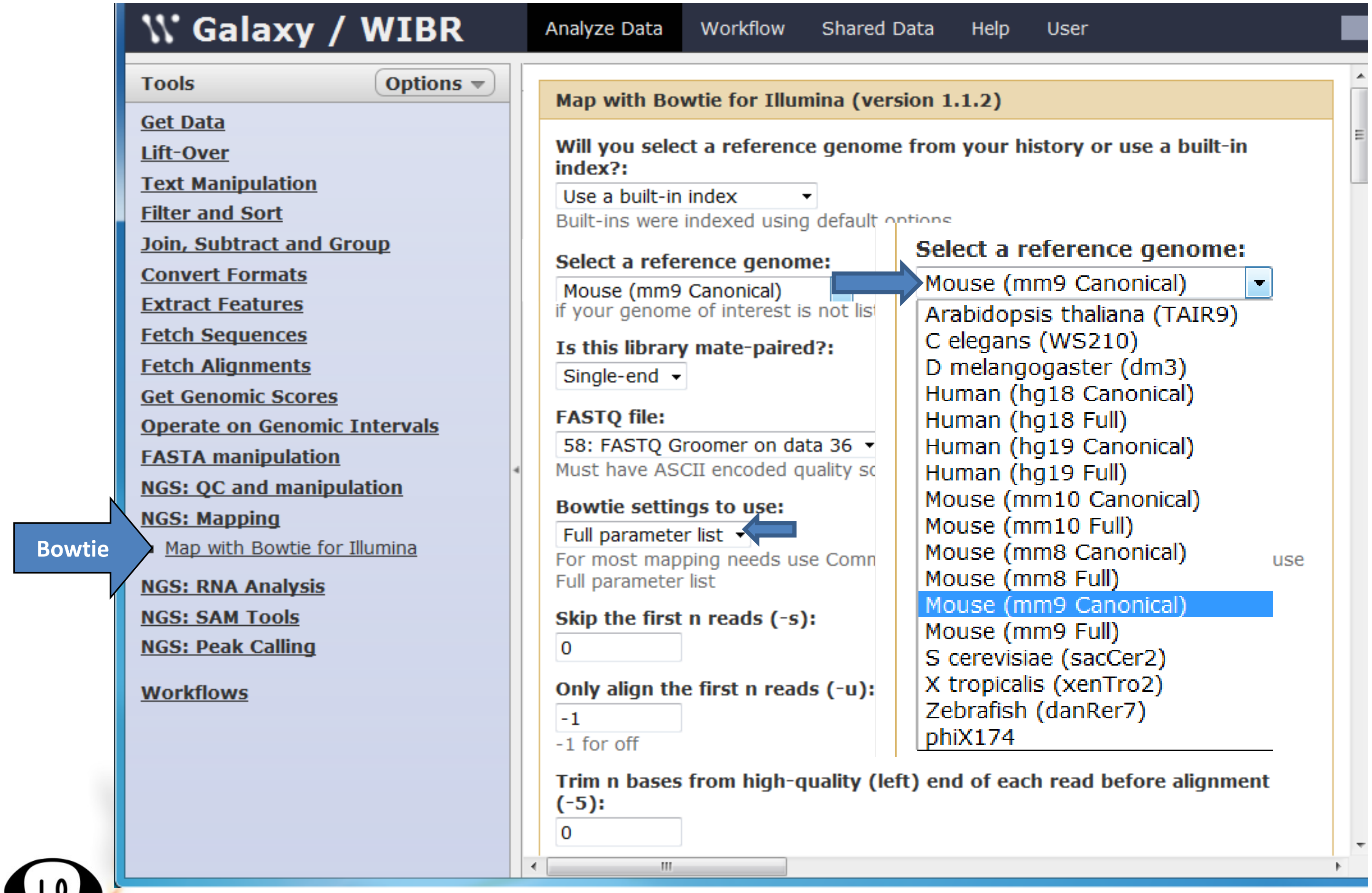
Step 1
Map reads

Bowtie

Step 2
Call peaks bound

MACS

Mapping Reads with **Bowtie**



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 lis

Is this library mate-paired?:
Single-end

FASTQ file:
58: FASTQ Groomer on data 36
Must have ASCII encoded quality sc

Bowtie settings to use:
Full parameter list
For most mapping needs use Comn
Full parameter list

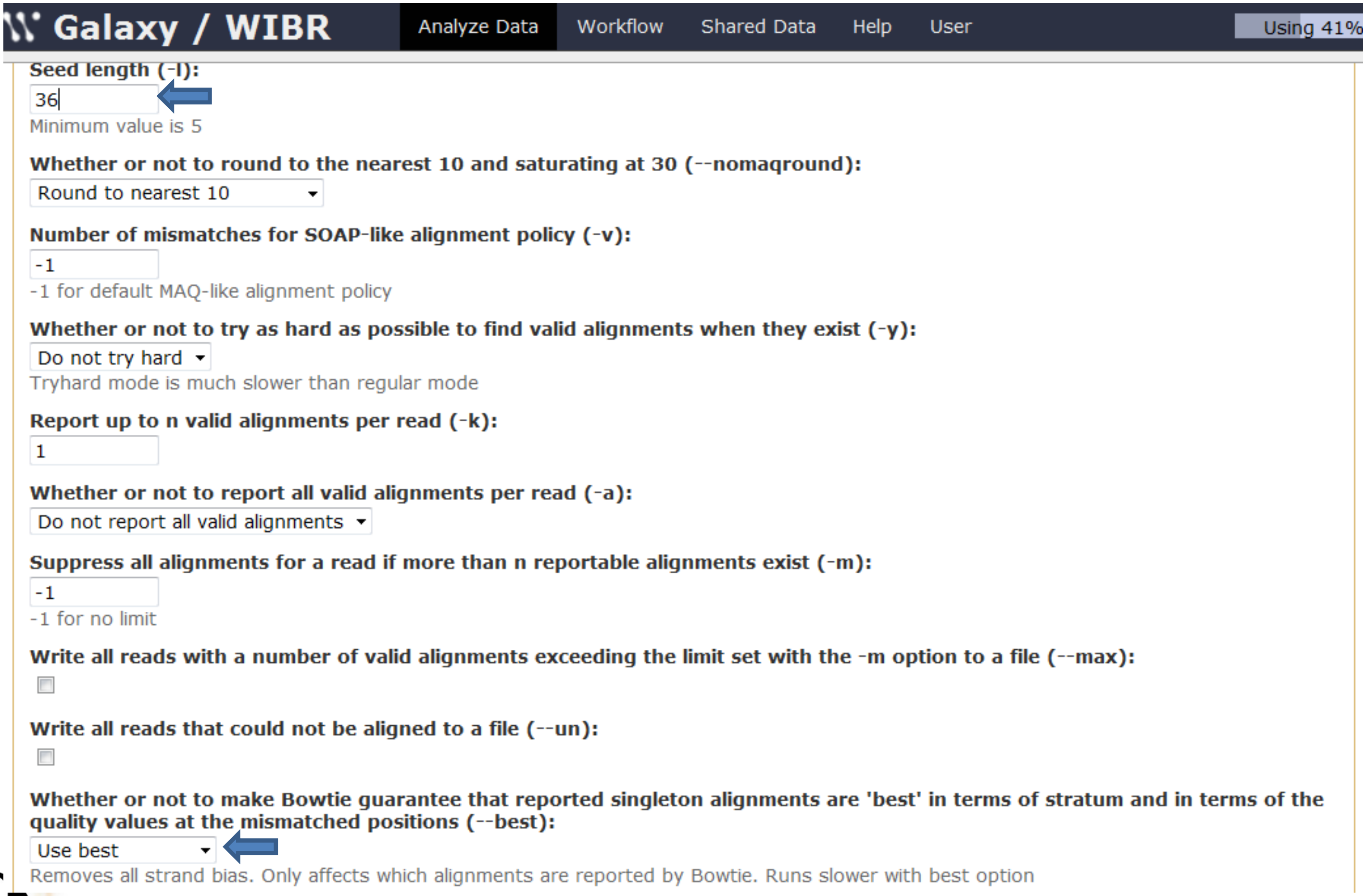
Skip the first n reads (-s):
0

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

Select a reference genome:
Mouse (mm9 Canonical)

Mapping Reads with **Bowtie**



Seed length (-l):

36

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):

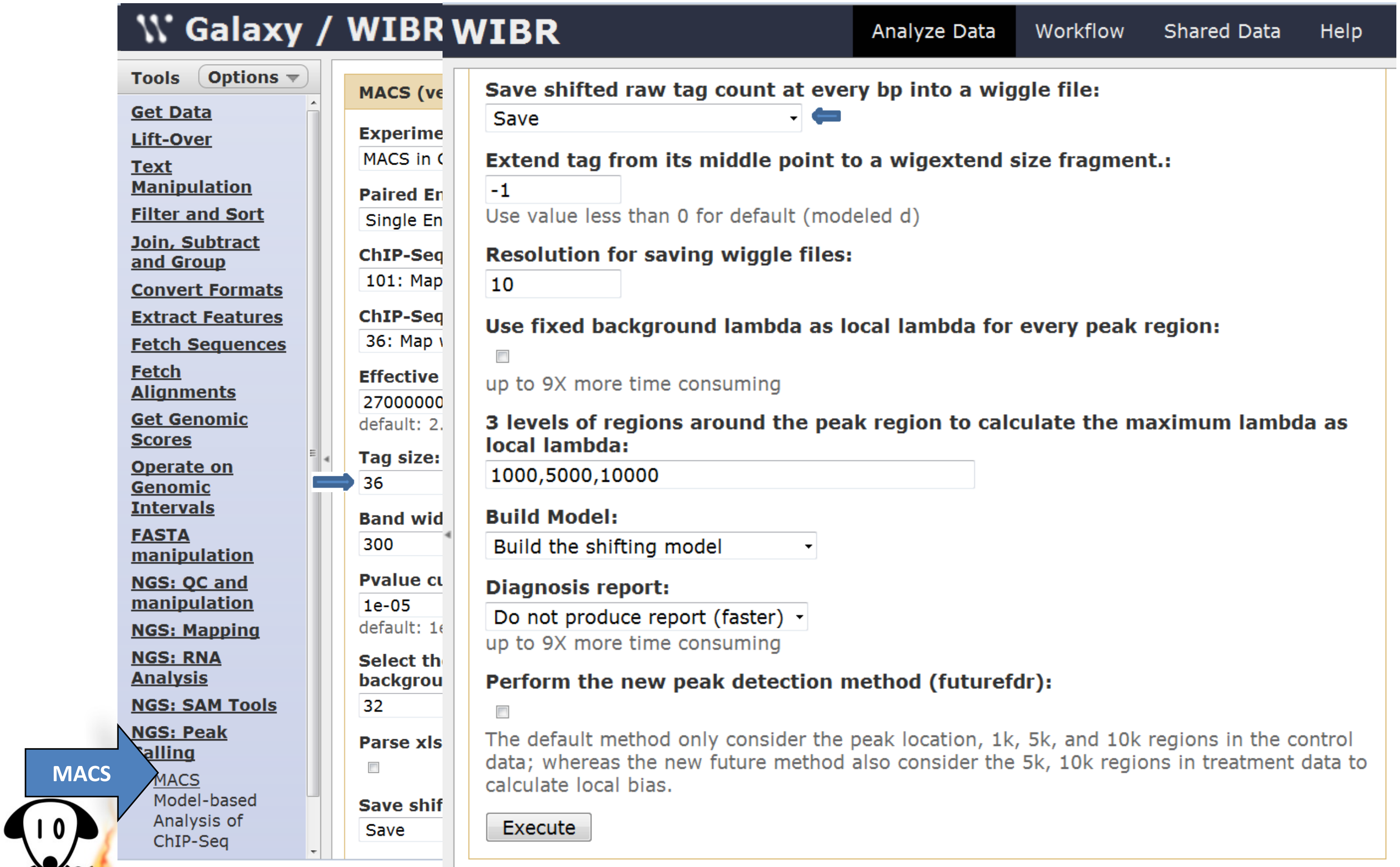
□

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

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



MACS (ve

Experime
MACS in C

Paired En
Single En

ChIP-Seq
101: Map

ChIP-Seq
36: Map w

Effective
27000000
default: 2.

Tag size:
36

Band wid
300

Pvalue cu
1e-05
default: 1e

Select th
backgrou
32

Parse xls
□

Save shif
Save

Save shifted raw tag count at every bp into a wiggle file:
Save

Extend tag from its middle point to a wigextend size fragment.:
-1
Use value less than 0 for default (modeled d)

Resolution for saving wiggle files:
10

Use fixed background lambda as local lambda for every peak region:
□
up to 9X more time consuming

3 levels of regions around the peak region to calculate the maximum lambda as local lambda:
1000,5000,10000

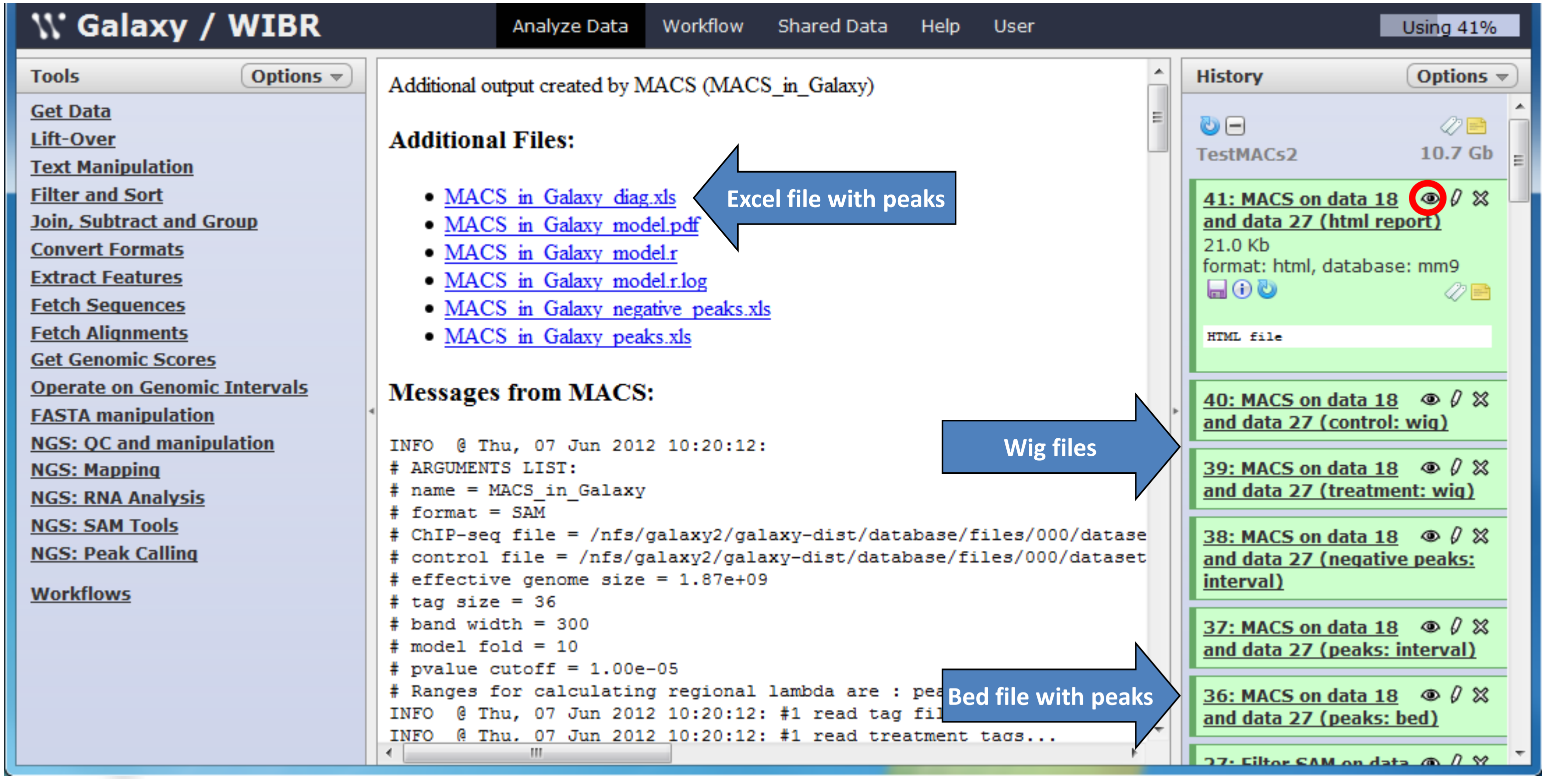
Build Model:
Build the shifting model

Diagnosis report:
Do not produce report (faster)
up to 9X more time consuming

Perform the new peak detection method (futurefdr):
□
The default method only consider the peak location, 1k, 5k, and 10k regions in the control data; whereas the new future method also consider the 5k, 10k regions in treatment data to calculate local bias.

Execute

MACS output



Additional output created by MACS (MACS_in_Galaxy)

Additional Files:

- MACS_in_Galaxy_diag.xls
- MACS_in_Galaxy_model.pdf
- MACS_in_Galaxy_model.r
- MACS_in_Galaxy_model.r.log
- MACS_in_Galaxy_negative_peaks.xls
- MACS_in_Galaxy_peaks.xls

Messages from MACS:

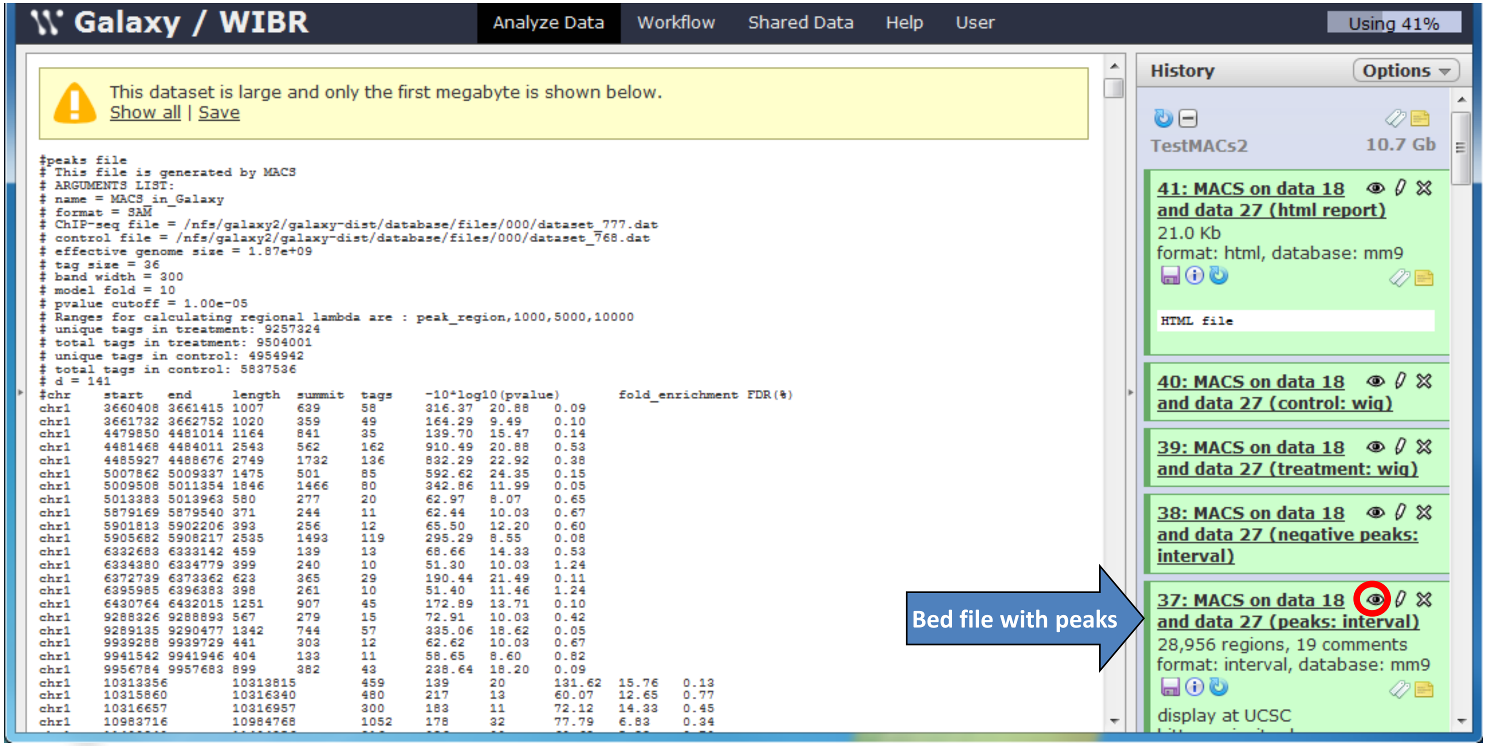
```
INFO  @ Thu, 07 Jun 2012 10:20:12:
# ARGUMENTS LIST:
# name = MACS_in_Galaxy
# format = SAM
# ChIP-seq file = /nfs/galaxy2/galaxy-dist/database/files/000/datase
# control file = /nfs/galaxy2/galaxy-dist/database/files/000/dataset
# effective genome size = 1.87e+09
# tag size = 36
# band width = 300
# model fold = 10
# pvalue cutoff = 1.00e-05
# Ranges for calculating regional lambda are : pea
INFO  @ Thu, 07 Jun 2012 10:20:12: #1 read tag fil
INFO  @ Thu, 07 Jun 2012 10:20:12: #1 read treatment tags...
```

Excel file with peaks

Wig files

Bed file with peaks

MACS output



This dataset is large and only the first megabyte is shown below.
Show all | Save

```
#peaks file
# This file is generated by MACS
# ARGUMENTS LIST:
# name = MACS_in_Galaxy
# format = SAM
# ChIP-seq file = /nfs/galaxy2/galaxy-dist/database/files/000/dataset_777.dat
# control file = /nfs/galaxy2/galaxy-dist/database/files/000/dataset_768.dat
# effective genome size = 1.87e+09
# tag size = 36
# band width = 300
# model fold = 10
# pvalue cutoff = 1.00e-05
# Ranges for calculating regional lambda are : peak_region,1000,5000,10000
# unique tags in treatment: 9257324
# total tags in treatment: 9504001
# unique tags in control: 4954942
# total tags in control: 5837536
# d = 141
#chr    start   end     length  summit  tags    -10*log10(pvalue)       fold_enrichment FDR(%)
chr1    3660408 3661415 1007    639     58      316.37  20.88   0.09
chr1    3661732 3662752 1020    359     49      164.29  9.49    0.10
chr1    4479850 4481014 1164    841     35      139.70  15.47   0.14
chr1    4481468 4484011 2543    562     162     910.49  20.88   0.53
chr1    4485927 4488676 2749    1732    136     832.29  22.92   0.38
chr1    5007862 5009337 1475    501     85      592.62  24.35   0.15
chr1    5009508 5011354 1846    1466    80      342.86  11.99   0.05
chr1    5013383 5013963 580     277     20      62.97   8.07    0.65
chr1    5879169 5879540 371     244     11      62.44   10.03   0.67
chr1    5901813 5902206 393     256     12      65.50   12.20   0.60
chr1    5905682 5908217 2535    1493    119     295.29  8.55    0.08
chr1    6332683 6333142 459     139     13      68.66   14.33   0.53
chr1    6334380 6334779 399     240     10      51.30   10.03   1.24
chr1    6372739 6373362 623     365     29      190.44  21.49   0.11
chr1    6395985 6396383 398     261     10      51.40   11.46   1.24
chr1    6430764 6432015 1251    907     45      172.89  13.71   0.10
chr1    9288326 9288893 567     279     15      72.91   10.03   0.42
chr1    9289135 9290477 1342    744     57      335.06  18.62   0.05
chr1    9939288 9939729 441     303     12      62.62   10.03   0.67
chr1    9941542 9941946 404     133     11      58.65   8.60    0.82
chr1    9956784 9957683 899     382     43      238.64  18.20   0.09
chr1    10313356        10313815        459     139     20      131.62  15.76   0.13
chr1    10315860        10316340        480     217     13      60.07   12.65   0.77
chr1    10316657        10316957        300     183     11      72.12   14.33   0.45
chr1    10983716        10984768        1052    178     32      77.79   6.83    0.34
```

History Options

TestMACs2 10.7 Gb

41: MACS on data 18 and data 27 (html report)
21.0 Kb
format: html, database: mm9

HTML file

40: MACS on data 18 and data 27 (control: wig)

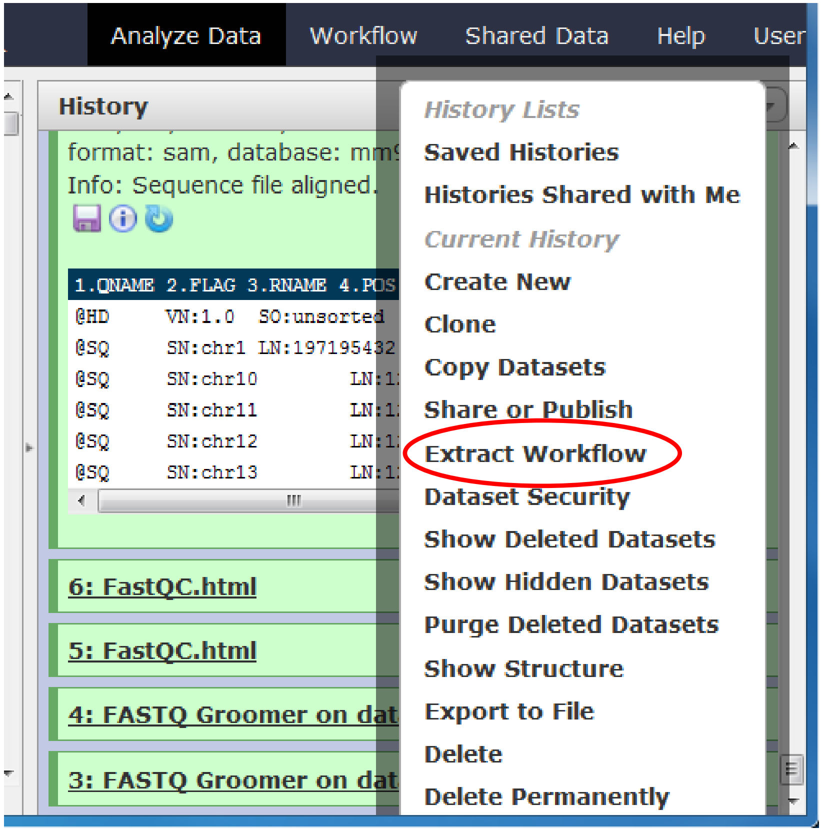
39: MACS on data 18 and data 27 (treatment: wig)

38: MACS on data 18 and data 27 (negative peaks: interval)

37: MACS on data 18 and data 27 (peaks: interval)
28,956 regions, 19 comments
format: interval, database: mm9
display at UCSC

Bed file with peaks

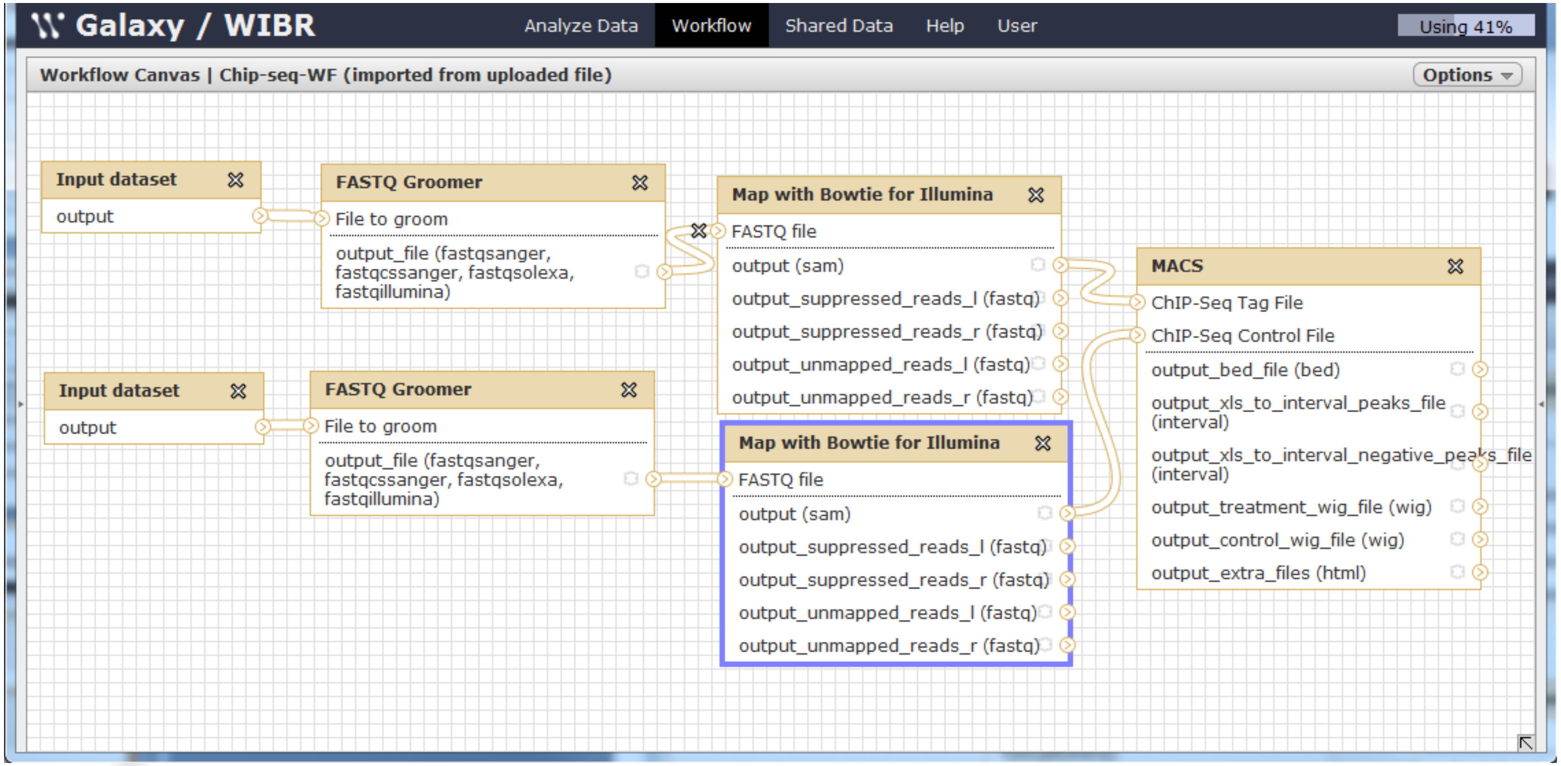
Creating Workflows



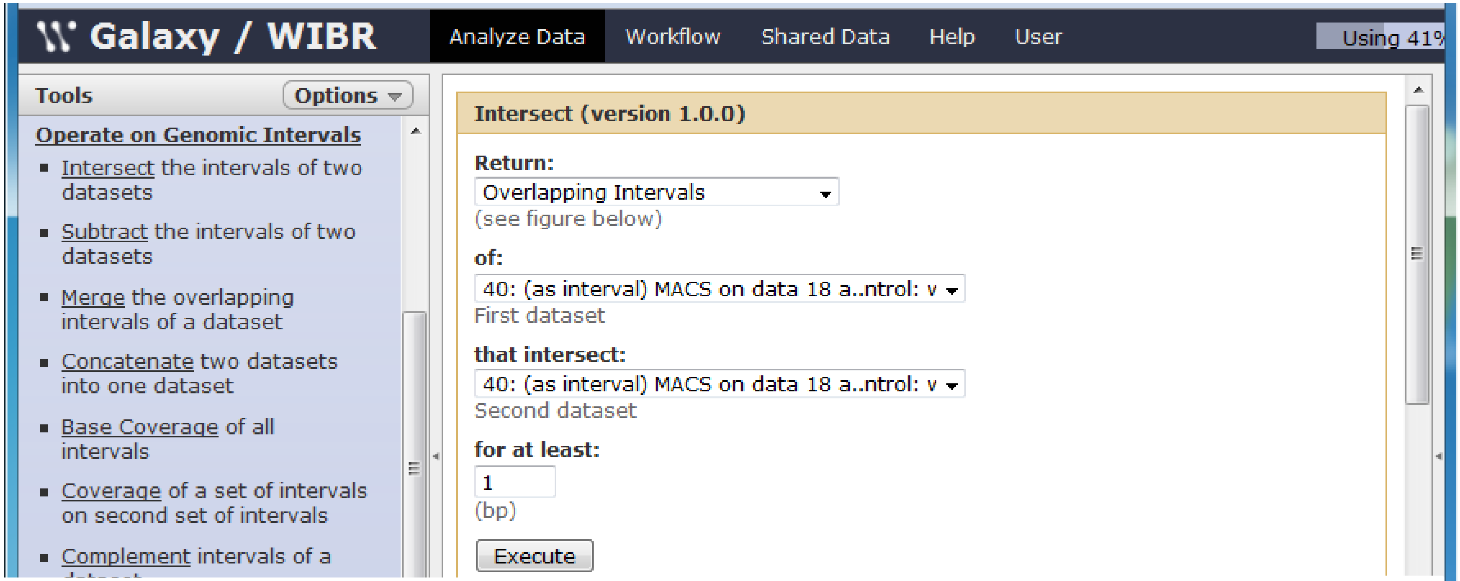
format: sam, database: mm9
Info: Sequence file aligned.

```
1.QNAME 2.FLAG 3.RNAME 4.POS
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr1 LN:197195432
@SQ     SN:chr10        LN:1
@SQ     SN:chr11        LN:1
@SQ     SN:chr12        LN:1
@SQ     SN:chr13        LN:1
```

Workflow for ChIP-seq analysis


Galaxy / WIBR
Analyze Data
Workflow
Shared Data
Help
User
Using 41%
Workflow Canvas | Chip-seq-WF (imported from uploaded file)
Options
Input dataset
output
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
Map with Bowtie for Illumina
FASTQ file
output (sam)
output_suppressed_reads_l (fastq)
output_suppressed_reads_r (fastq)
output_unmapped_reads_l (fastq)
output_unmapped_reads_r (fastq)
MACS
ChIP-Seq Tag File
ChIP-Seq Control File
output_bed_file (bed)
output_xls_to_interval_peaks_file (interval)
output_xls_to_interval_negative_peaks_file (interval)
output_treatment_wig_file (wig)
output_control_wig_file (wig)
output_extra_files (html)
Input dataset
output
FASTQ Groomer
File to groom
output_file (fastqsanger, fastqcssanger, fastqsolexa, fastqillumina)
Map with Bowtie for Illumina
FASTQ file
output (sam)
output_suppressed_reads_l (fastq)
output_suppressed_reads_r (fastq)
output_unmapped_reads_l (fastq)
output_unmapped_reads_r (fastq)

Example of downstream analysis: Intersect intervals of two datasets



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

Intersect (version 1.0.0)

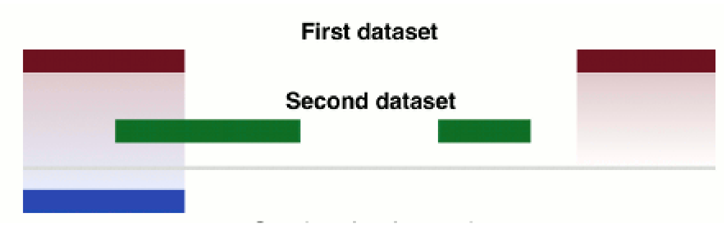
Return:
Overlapping Intervals
(see figure below)

of:
40: (as interval) MACS on data 18 a..ntrol: v
First dataset

that intersect:
40: (as interval) MACS on data 18 a..ntrol: v
Second dataset

for at least:
1
(bp)

Overlapping Intervals:



Genes, promoters, or regions of interest

ChIP-seq peaks

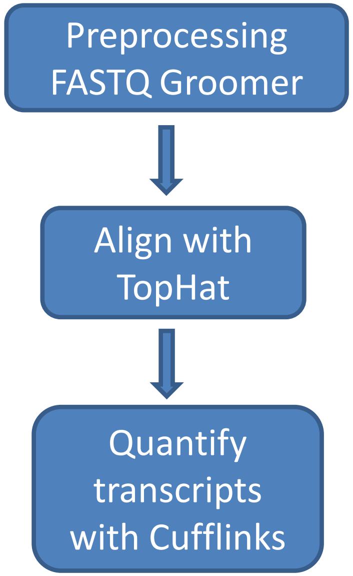
Genes overlapping with peaks

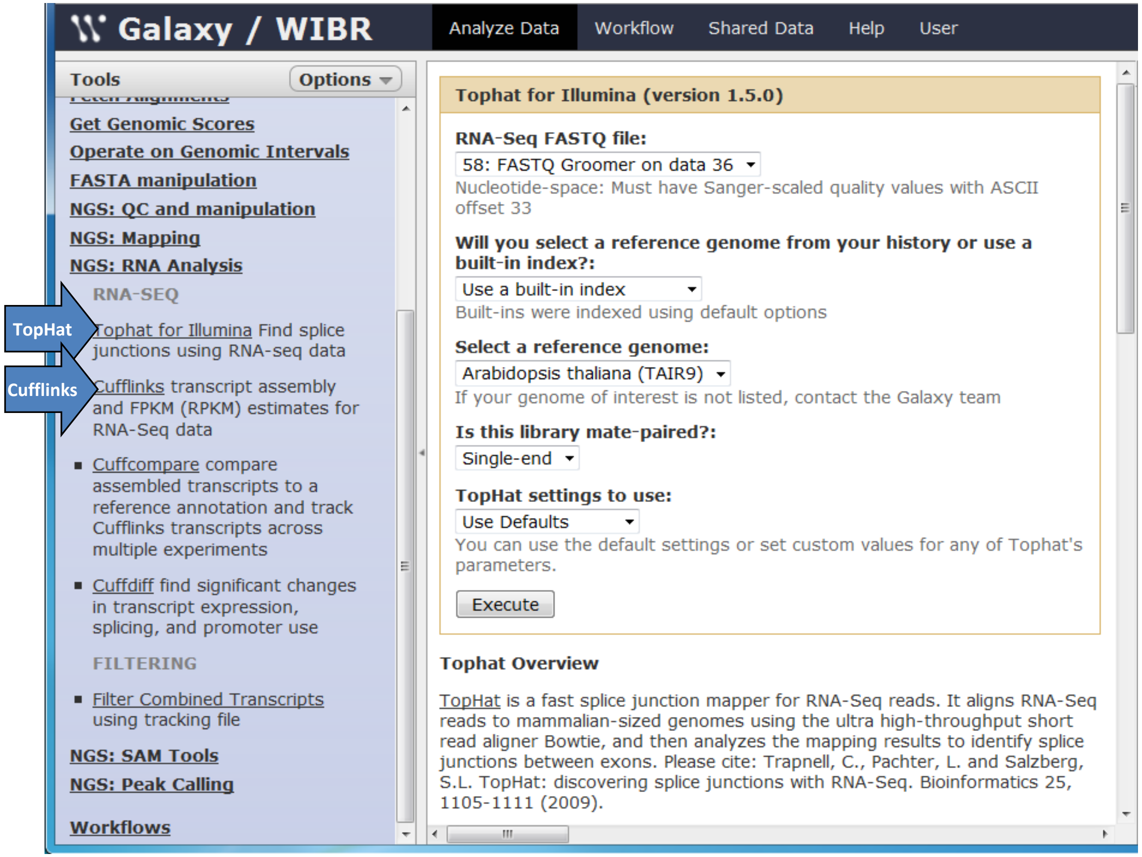
http://jura.wi.mit.edu/bio/education/hot_topics/galaxy/Galaxy.pdf

Talk Outline

- The Galaxy interface
- Getting data into Galaxy
- Overview of the tools
- The Next Generation Sequencing tool box
  - Preprocessing and quality control
  - Analysis of ChIP-seq
  - Analysis of RNA-seq
- Visualizing data on a genome browser and workflows available for analysis

Expression Profiling Workflow


Preprocessing FASTQ Groomer
Align with TopHat
Quantify transcripts with Cufflinks



Get Genomic Scores
Operate on Genomic Intervals
FASTA manipulation
NGS: QC and manipulation
NGS: Mapping
NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

NGS: SAM Tools
NGS: Peak Calling

Workflows

Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:
58: FASTQ Groomer on data 36
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

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:
Arabidopsis thaliana (TAIR9)
If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:
Single-end

TopHat settings to use:
Use Defaults
You can use the default settings or set custom values for any of Tophat's parameters.

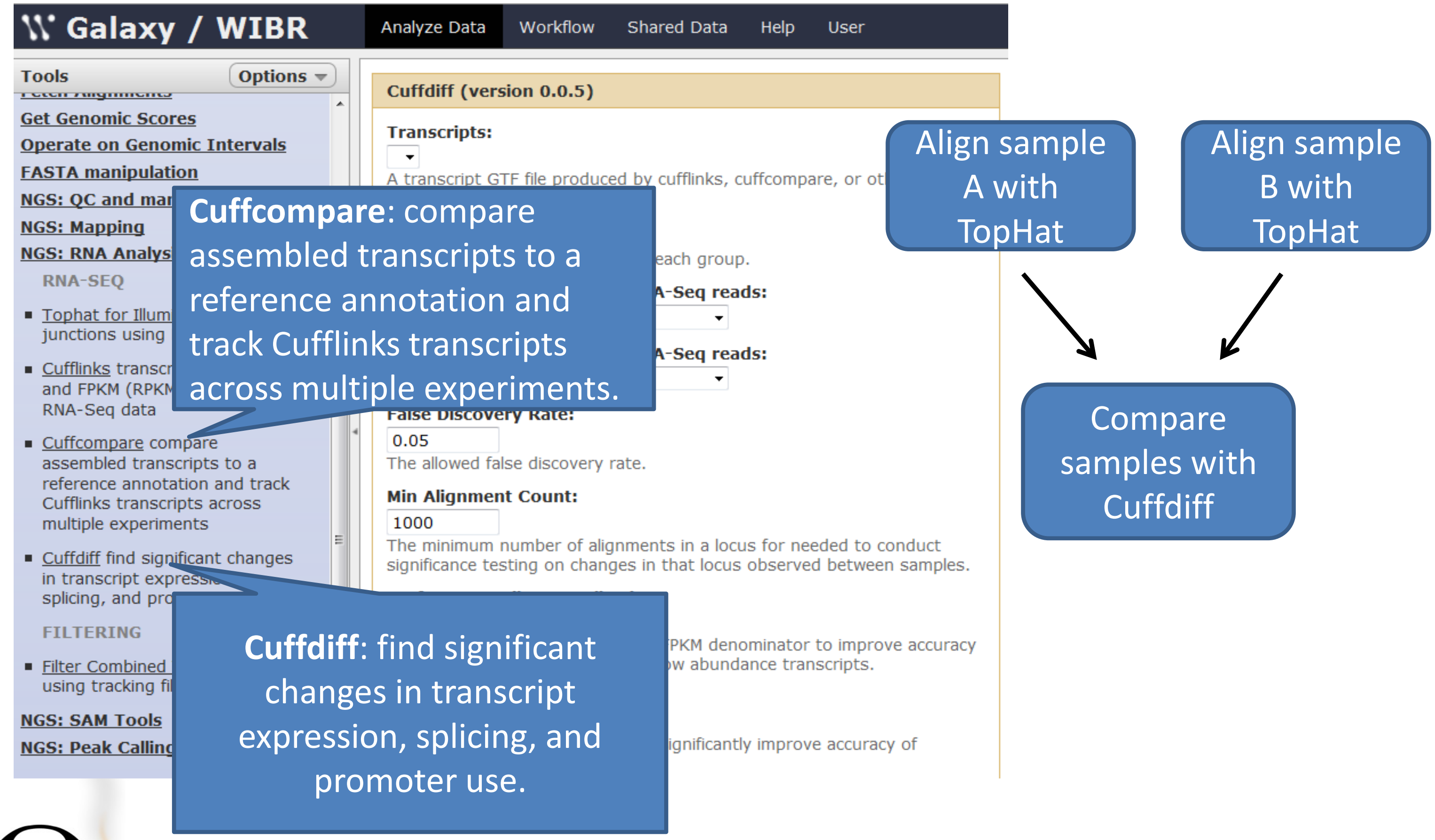
Execute

Tophat Overview

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Please cite: Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).

BaRC Hot Topics *Galaxy*

Other tools for expression profiling



Get Genomic Scores
Operate on Genomic Intervals
FASTA manipulation
NGS: QC and ma
NGS: Mapping
NGS: RNA Analys

RNA-SEQ

- Tophat for Illum junctions using
- Cufflinks transcr and FPKM (RPKM RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expressi splicing, and pro

FILTERING

- Filter Combined using tracking fi

NGS: SAM Tools
NGS: Peak Calling

Cuffdiff (version 0.0.5)

Transcripts:

A transcript GTF file produced by cufflinks, cuffcompare, or ot

each group.

A-Seq reads:

A-Seq reads:

False Discovery Rate:

0.05

The allowed false discovery rate.

Min Alignment Count:

1000

The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples.

FPKM denominator to improve accuracy w abundance transcripts.

ignificantly improve accuracy of

Cuffcompare: compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments.

Cuffdiff: find significant changes in transcript expression, splicing, and promoter use.

Align sample A with TopHat

Align sample B with TopHat

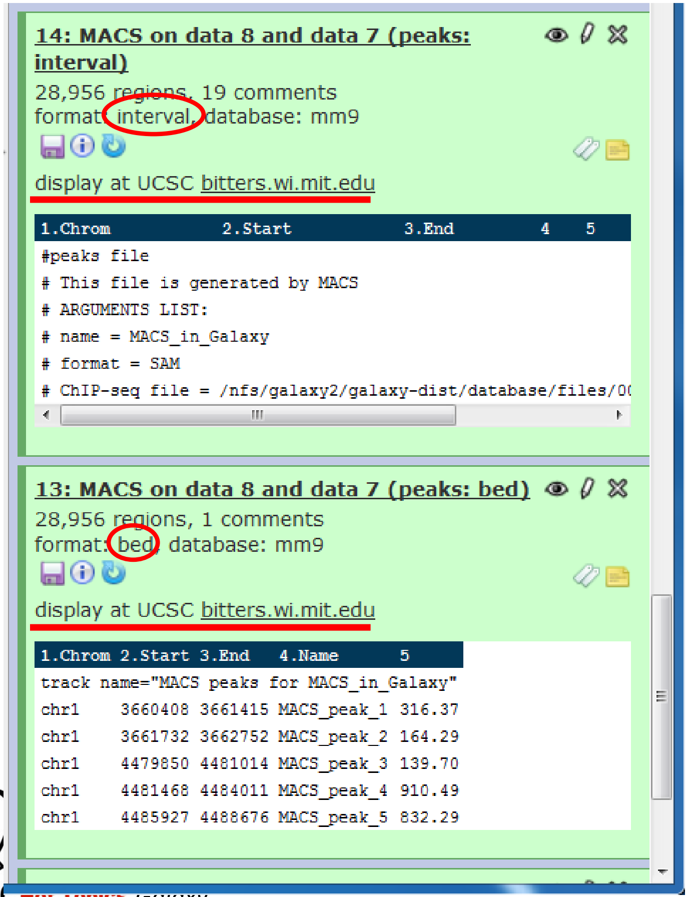
Compare samples with Cuffdiff

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc. 2012 Mar 1;7(3):562-78* . PMID: 22383036

Talk Outline

- The Galaxy interface
- Getting data into Galaxy
- Overview of the tools
- The Next Generation Sequencing tool box
  - Preprocessing and quality control
  - Analysis of ChIP-seq
  - Analysis of RNA-seq
- Visualizing data on a genome browser and workflows available for analysis

Visualizing data on UCSC



14: MACS on data 8 and data 7 (peaks: interval)

28,956 regions, 19 comments
format: interval, database: mm9

display at UCSC bitters.wi.mit.edu

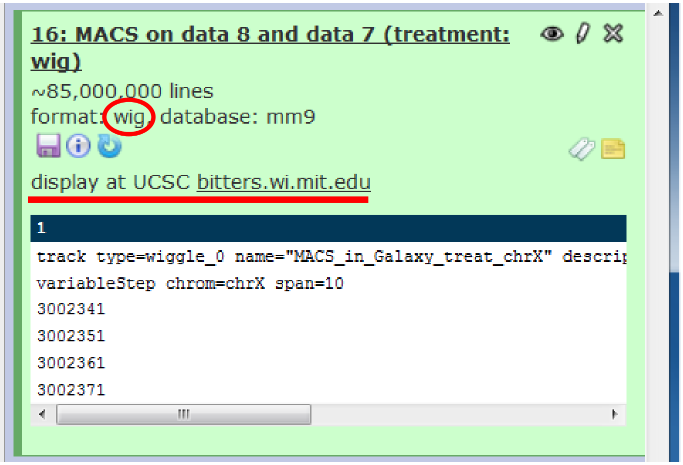
```
1.Chrom              2.Start              3.End          4     5
#peaks file
# This file is generated by MACS
# ARGUMENTS LIST:
# name = MACS_in_Galaxy
# format = SAM
# ChIP-seq file = /nfs/galaxy2/galaxy-dist/database/files/00
```

13: MACS on data 8 and data 7 (peaks: bed)

28,956 regions, 1 comments
format: bed, database: mm9

display at UCSC bitters.wi.mit.edu

```
1.Chrom 2.Start 3.End    4.Name        5
track name="MACS peaks for MACS_in_Galaxy"
chr1     3660408 3661415 MACS_peak_1 316.37
chr1     3661732 3662752 MACS_peak_2 164.29
chr1     4479850 4481014 MACS_peak_3 139.70
chr1     4481468 4484011 MACS_peak_4 910.49
chr1     4485927 4488676 MACS_peak_5 832.29
```

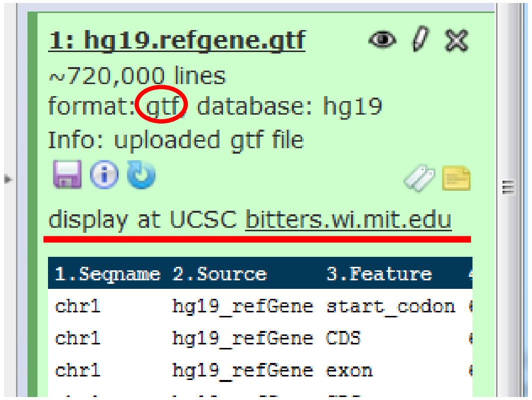


16: MACS on data 8 and data 7 (treatment: wig)

~85,000,000 lines
format: wig, database: mm9

display at UCSC bitters.wi.mit.edu

```
1
track type=wiggle_0 name="MACS_in_Galaxy_treat_chrX" descrip
variableStep chrom=chrX span=10
3002341
3002351
3002361
3002371
```



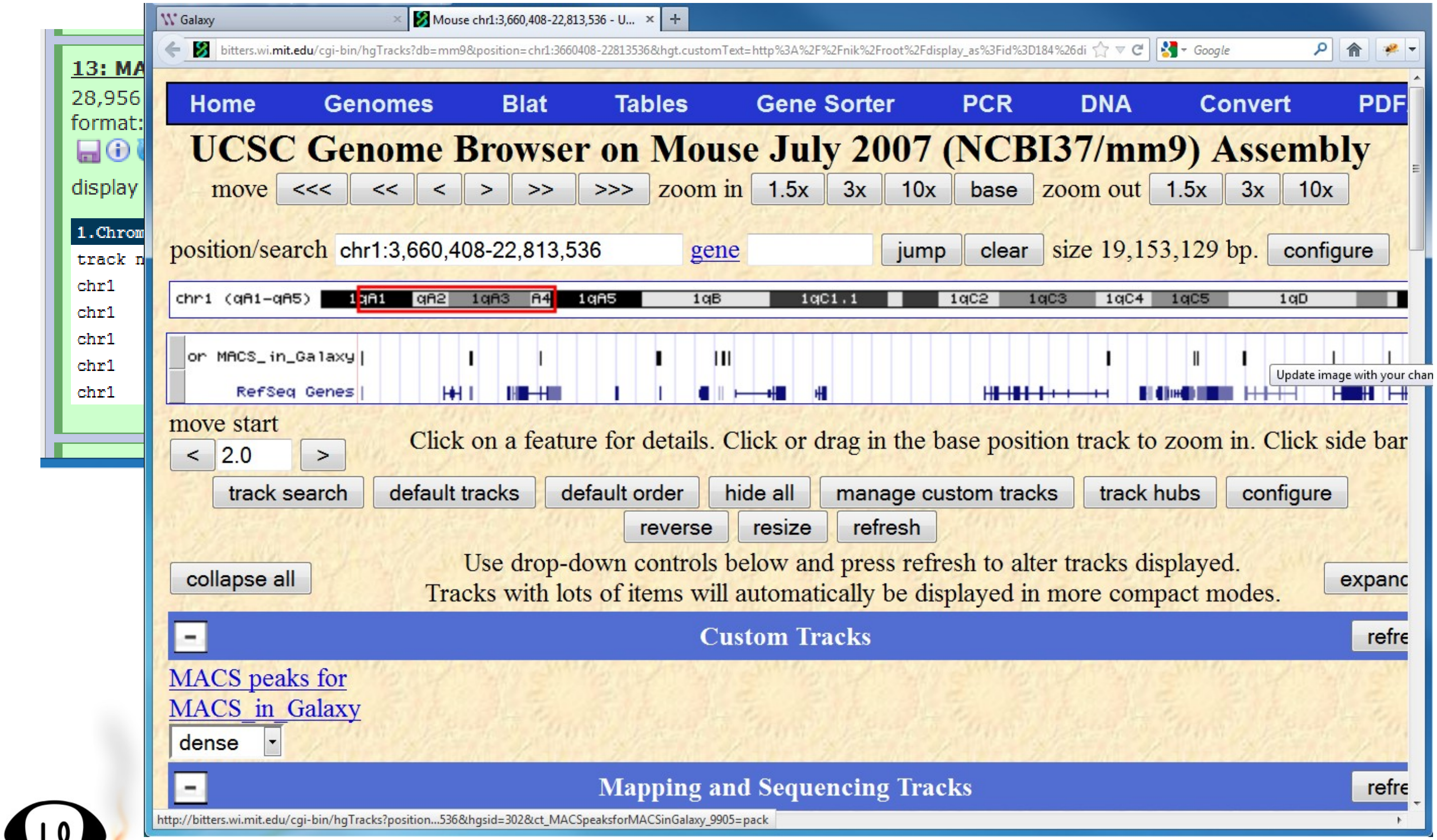
1: hg19.refgene.gtf

~720,000 lines
format: gtf, database: hg19
Info: uploaded gtf file

display at UCSC bitters.wi.mit.edu

```
1.Seqname  2.Source     3.Feature
chr1       hg19_refGene start_codon
chr1       hg19_refGene CDS
chr1       hg19_refGene exon
```

Visualizing data on UCSC



bitters.wi.mit.edu/cgi-bin/hgTracks?db=mm9&position=chr1:3660408-22813536&hgt.customText=http%3A%2F%2Fnik%2Froot%2Fdisplay_as%3Fid%3D184%26di

UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly

position/search chr1:3,660,408-22,813,536 size 19,153,129 bp.

move start 2.0

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bar

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Custom Tracks

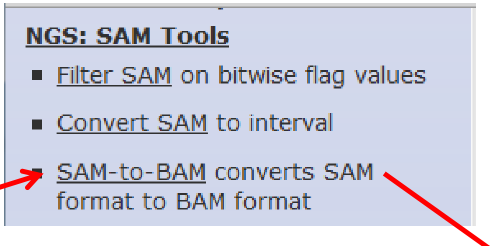
MACS peaks for MACS_in_Galaxy

dense

Mapping and Sequencing Tracks

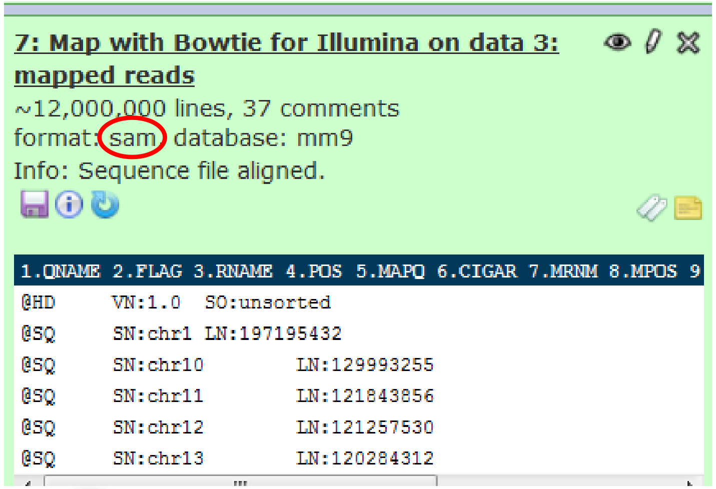
http://bitters.wi.mit.edu/cgi-bin/hgTracks?position...536&hgsid=302&ct_MACSpeaksforMACSinGalaxy_9905=pack

Visualizing data on UCSC: BAM files



NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format



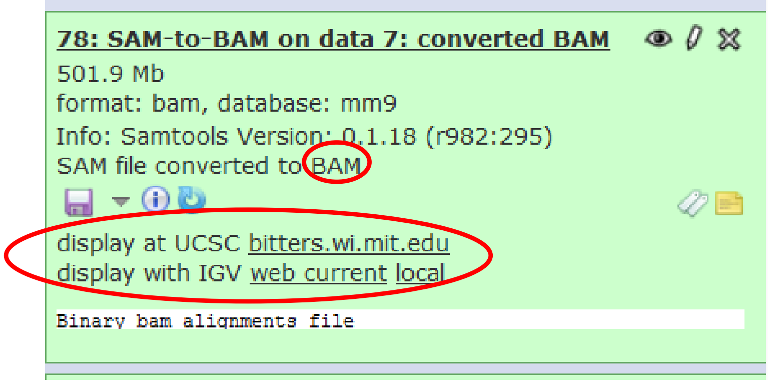
7: Map with Bowtie for Illumina on data 3: mapped reads

~12,000,000 lines, 37 comments

format: sam, database: mm9

Info: Sequence file aligned.

```
1.QNAME 2.FLAG 3.RNAME 4.POS 5.MAPQ 6.CIGAR 7.MRNM 8.MPOS 9
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr1 LN:197195432
@SQ     SN:chr10        LN:129993255
@SQ     SN:chr11        LN:121843856
@SQ     SN:chr12        LN:121257530
@SQ     SN:chr13        LN:120284312
```



78: SAM-to-BAM on data 7: converted BAM

501.9 Mb

format: bam, database: mm9

Info: Samtools Version: 0.1.18 (r982:295)

SAM file converted to BAM

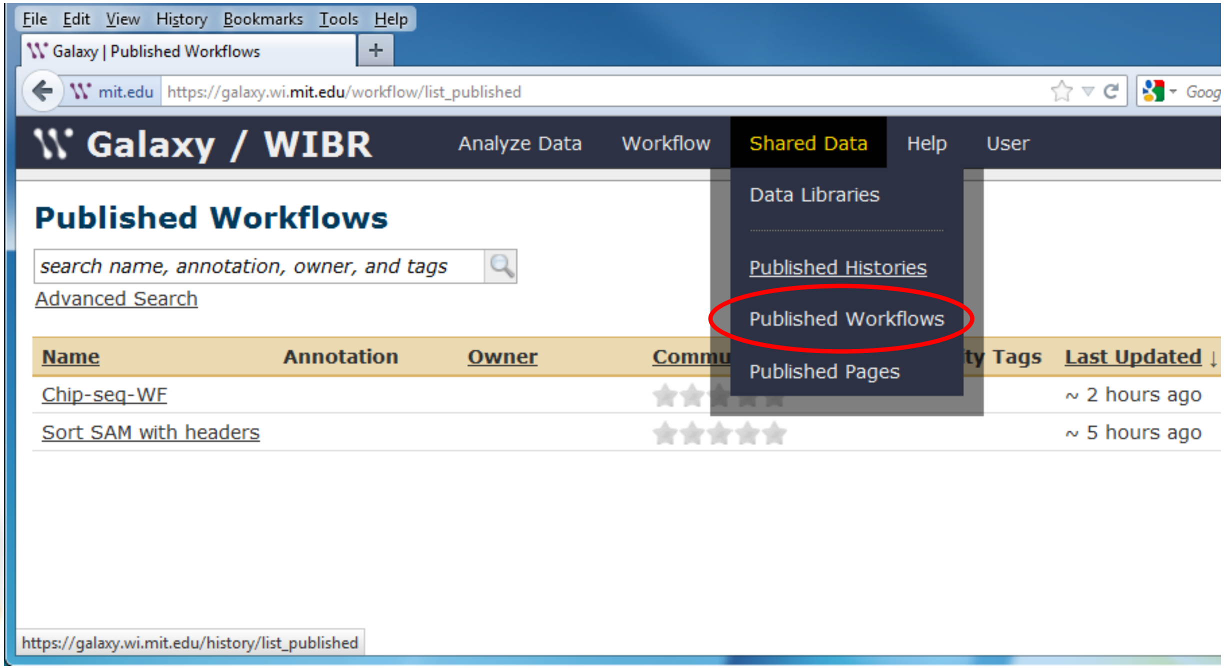
display at UCSC bitters.wi.mit.edu

display with IGV web current local

```
Binary bam alignments file
```

Workflows available inside Whitehead

1. Workflow for ChIP-seq analysis.
2. Workflow for sorting a SAM file.


File Edit View History Bookmarks Tools Help
Galaxy | Published Workflows
mit.edu https://galaxy.wi.mit.edu/workflow/list_published
Galaxy / WIBR
Analyze Data
Workflow
Shared Data
Help
User
Data Libraries
Published Histories
Published Workflows
Published Pages
search name, annotation, owner, and tags
Advanced Search
https://galaxy.wi.mit.edu/history/list_published

Published Workflows

Name	Annotation	Owner	Commu…ty Tags	Last Updated ↓
Chip-seq-WF				~ 2 hours ago
Sort SAM with headers				~ 5 hours ago

Documentation and Tutorials

- OpenHelix tutorials and exercises

  http://www.openhelix.com/cgi/tutorialInfo.cgi?id=82
- Galaxy tutorials

  http://galaxy.psu.edu/screencasts.html
- References

Galaxy developers: The Center for Comparative Genomics & Bioinformatics, Pennsylvania State University

Giardine, B., et al. Galaxy: a platform for interactive large-scale analysis. Genome Research (2005) 15:1451-1455

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

Blankenberg D., et al. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26(14):1783-5

Previous Hot Topics

- Previous Hot Topics in Galaxy
  http://jura.wi.mit.edu/bio/education/hot_topics/galaxy/Galaxy.pdf
  http://jura.wi.mit.edu/bio/education/hot_topics/GalaxyNGS/Galaxy_NGS.pdf
- Previous Hot Topics in NGS
  http://jura.wi.mit.edu/bio/education/hot_topics/shortRead_mapping/Mapping_HTseq.pdf
  http://jura.wi.mit.edu/bio/education/hot_topics/ChIPseq/ChIPSeq_HotTopics.pdf
  http://jura.wi.mit.edu/bio/education/hot_topics/RNAseq/RNA_Seq.pdf