Nebula – A web-server for advanced ChIP-seq data analysis

Tutorial

by Valentina BOEVA

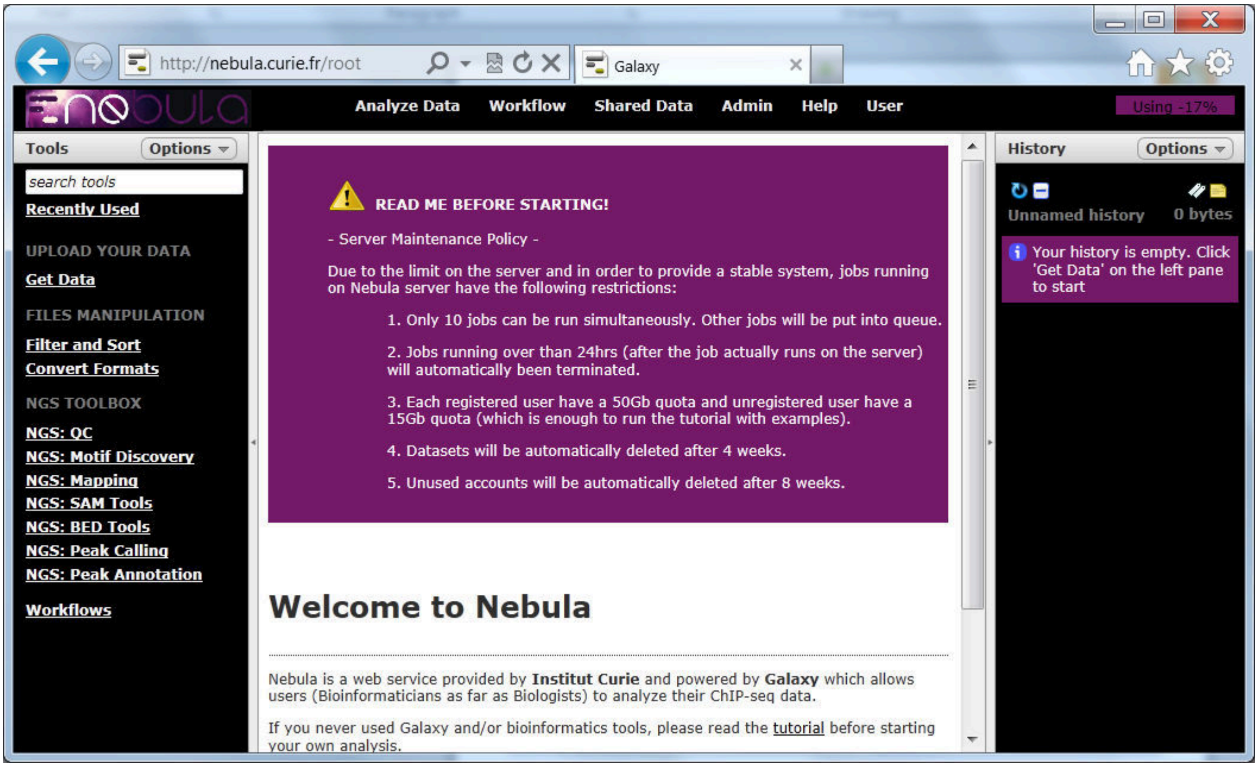
Content

PIPELINE FOR TRANSCRIPTION FACTORS

PIPELINE FOR HISTONE DATA

Our web server: http://nebula.curie.fr/

Our web service, Nebula, is based on the Galaxy open source framework.



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

READ ME BEFORE STARTING!

- Server Maintenance Policy -

Due to the limit on the server and in order to provide a stable system, jobs running on Nebula server have the following restrictions:

1. Only 10 jobs can be run simultaneously. Other jobs will be put into queue.
2. Jobs running over than 24hrs (after the job actually runs on the server) will automatically been terminated.
3. Each registered user have a 50Gb quota and unregistered user have a 15Gb quota (which is enough to run the tutorial with examples).
4. Datasets will be automatically deleted after 4 weeks.
5. Unused accounts will be automatically deleted after 8 weeks.

Welcome to Nebula

Nebula is a web service provided by **Institut Curie** and powered by **Galaxy** which allows users (Bioinformaticians as far as Biologists) to analyze their ChIP-seq data.

If you never used Galaxy and/or bioinformatics tools, please read the tutorial before starting your own analysis.

Tool box

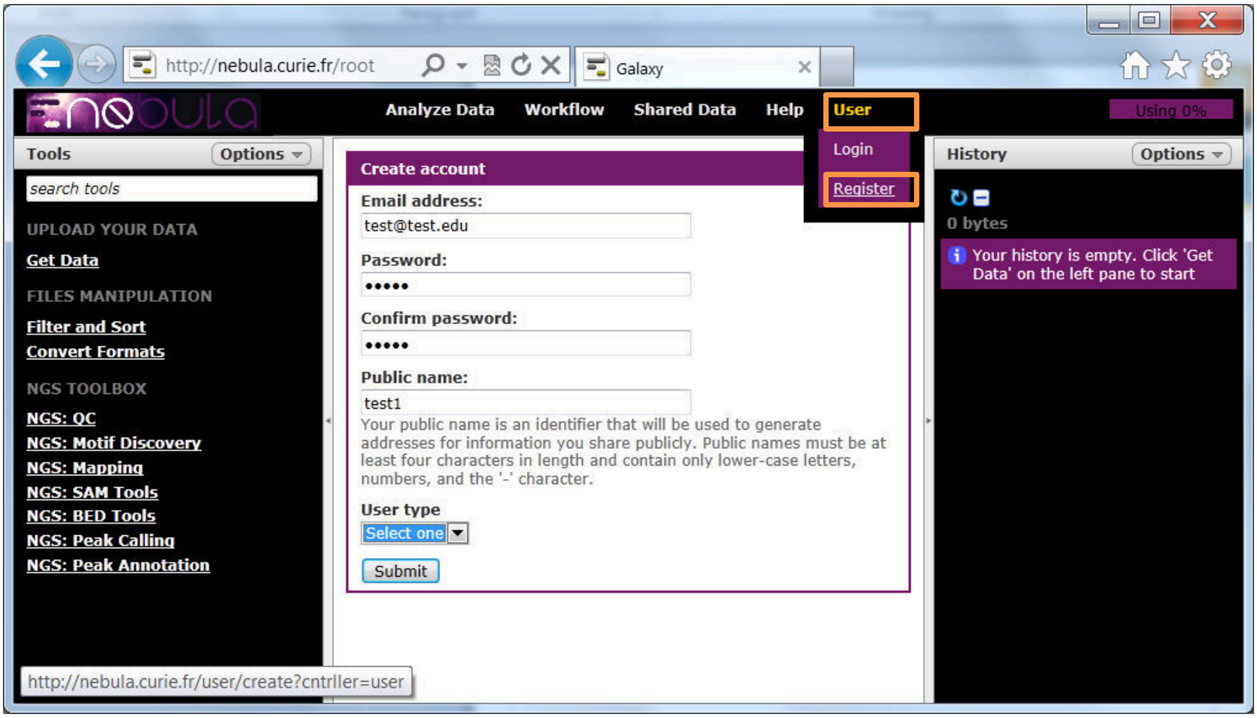
Work field

History

Main Galaxy server: [http://main.g2.bx.psu.edu/] **does not** include **all** our ChIP-seq analysis tools, but you can use it for other occasions.

Create your account

- Each registered user have a 50Gb quota and unregistered user have a 15Gb quota (which is enough to run the tutorial with examples).
- We would prefer you to register even if you don’t use your real email address.



Create account

Email address: test@test.edu

Password: •••••

Confirm password: •••••

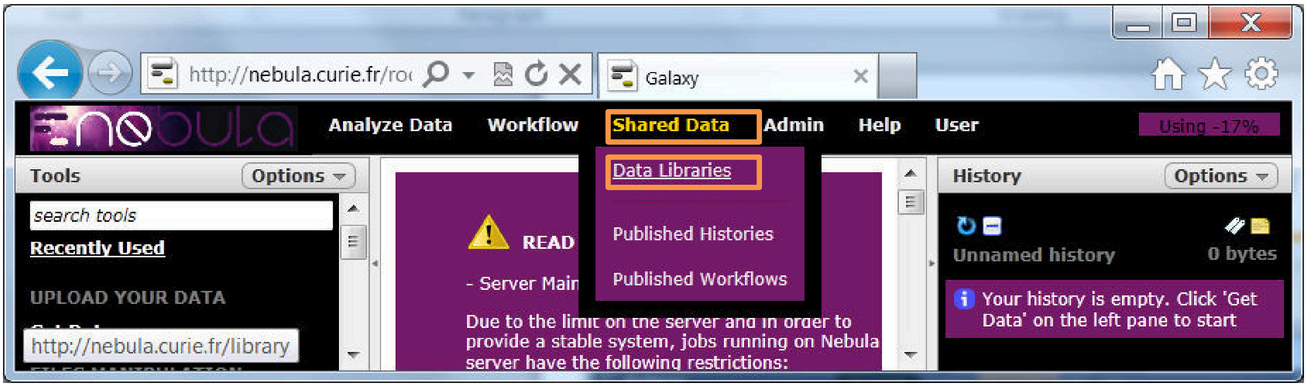
Public name: test1

Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least four characters in length and contain only lower-case letters, numbers, and the '-' character.

User type

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

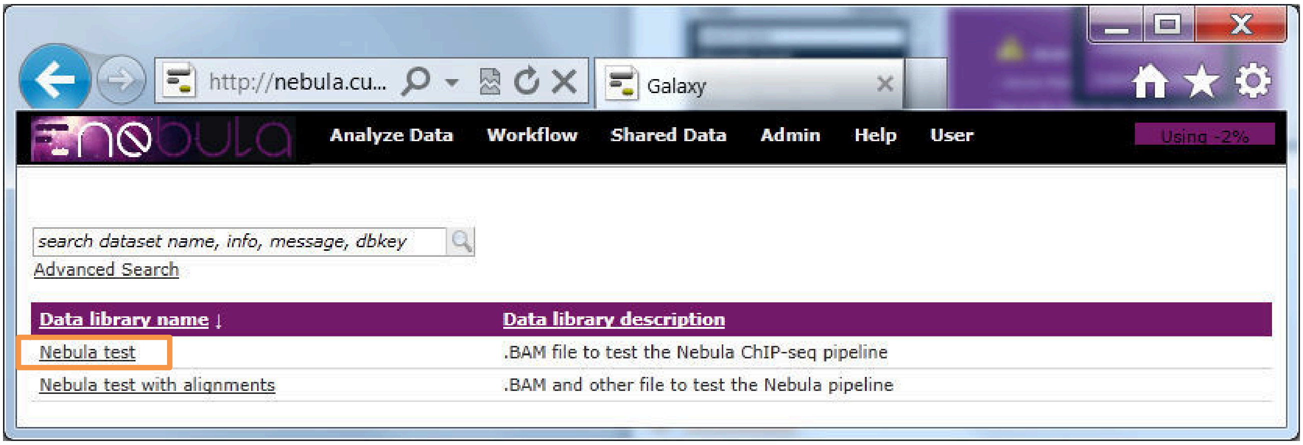
Download the test dataset to the history



- Server Main

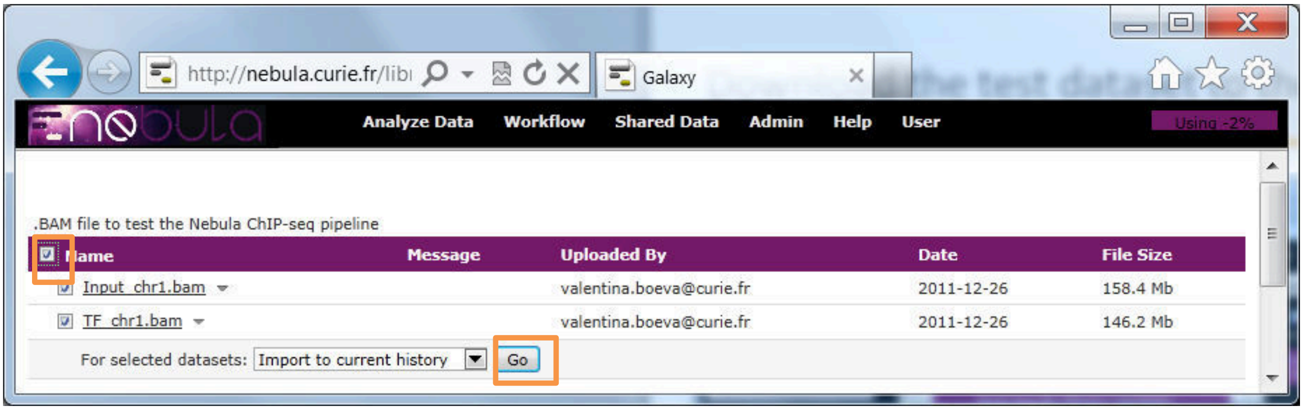
Due to the limit on the server and in order to provide a stable system, jobs running on Nebula server have the following restrictions:

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



Data library name ↓	Data library description
Nebula test	.BAM file to test the Nebula ChIP-seq pipeline
Nebula test with alignments	.BAM and other file to test the Nebula pipeline

- Select and import all datasets:



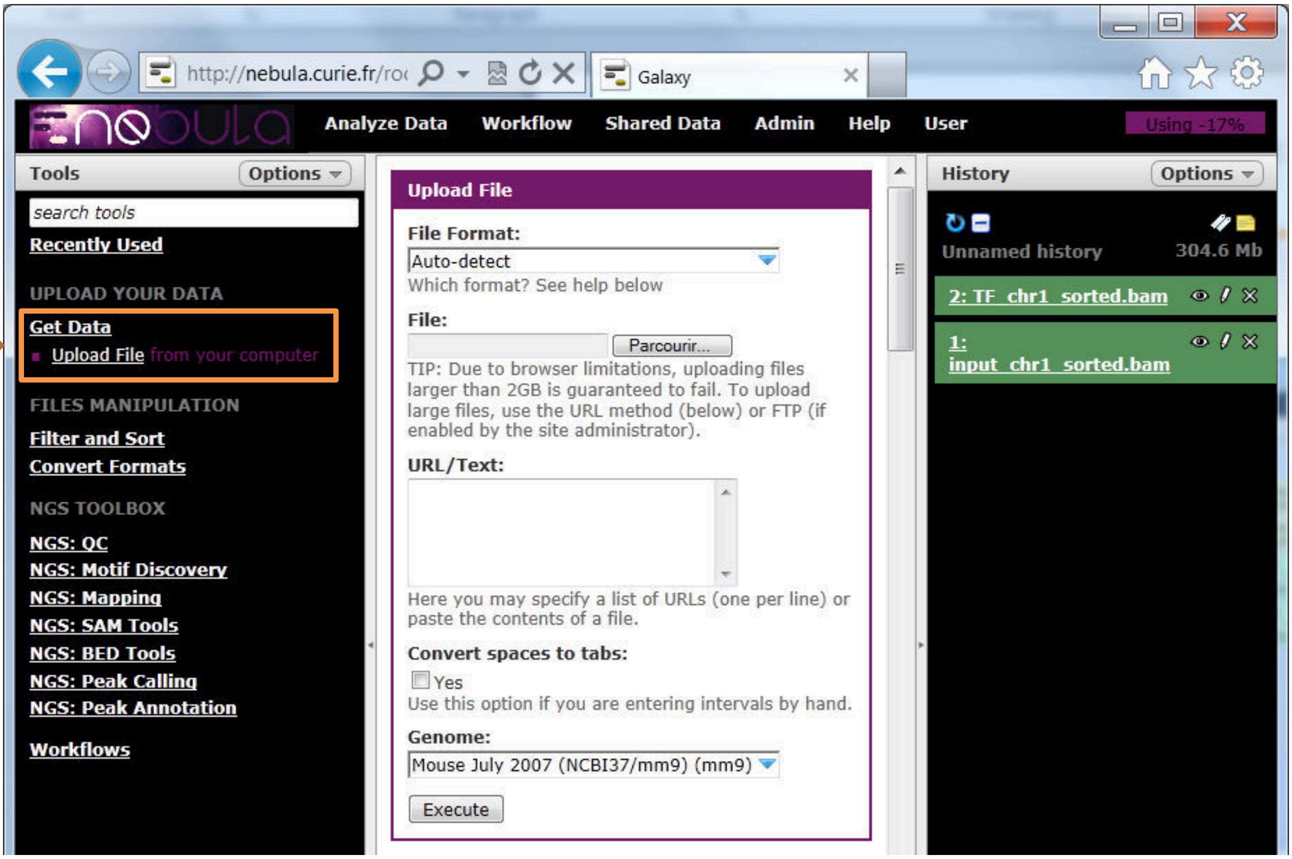
.BAM file to test the Nebula ChIP-seq pipeline

☑ Name	Message	Uploaded By	Date	File Size
☑ Input_chr1.bam		valentina.boeva@curie.fr	2011-12-26	158.4 Mb
☑ TF_chr1.bam		valentina.boeva@curie.fr	2011-12-26	146.2 Mb

For selected datasets: Import to current history Go

- Then go back by clicking “Analyze Data”

Alternative way to download your dataset to the history



Upload File

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, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

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

Convert spaces to tabs:

□ Yes

Use this option if you are entering intervals by hand.

Genome: Mouse July 2007 (NCBI37/mm9) (mm9)

This way you will use outside of this tutorial

- To upload files larger than 2GB, the user has to use the URL method through FTP/HTTP protocol. The user must have access to an open web server or ftp server where he should upload his data. If the user does not have access to any web or ftp server, he can install his own web server.

  The following servers are free and can be easily installed:

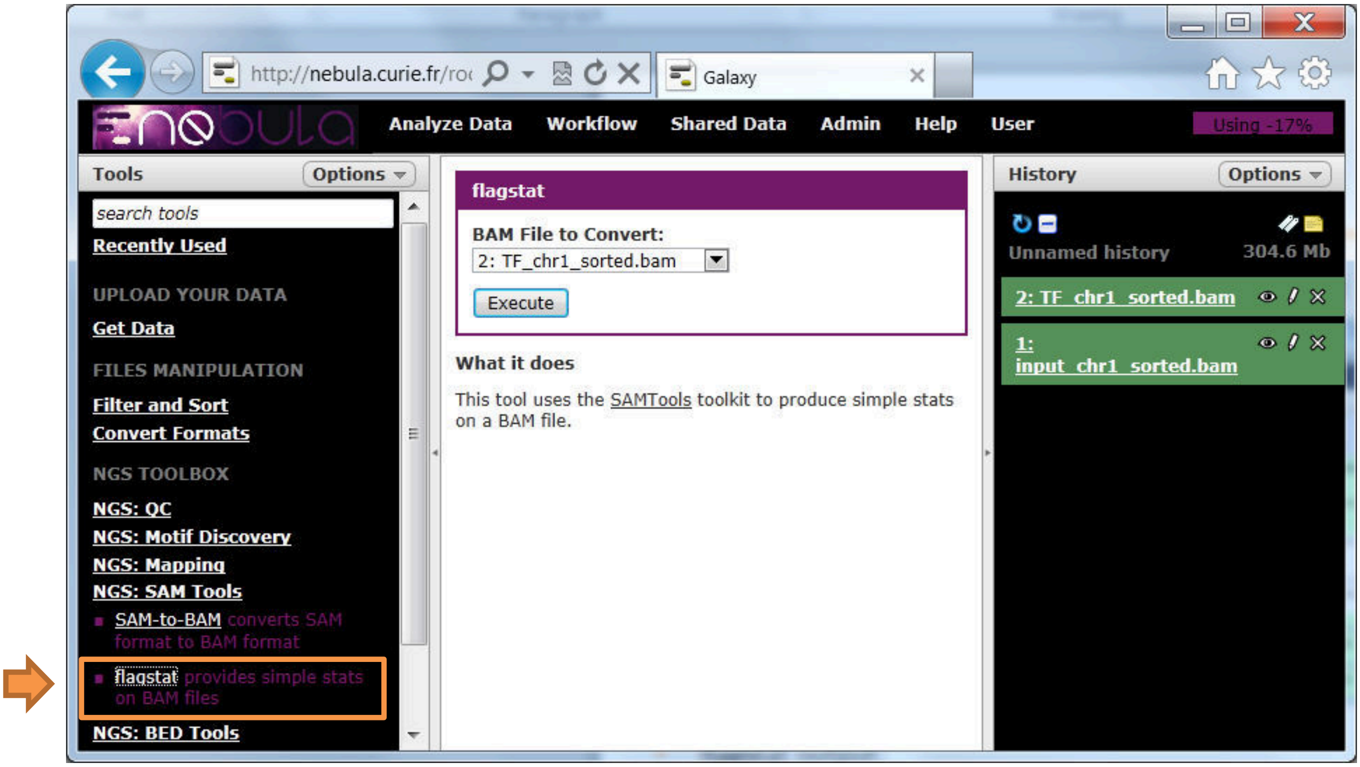
  Web servers:
  MAMP for Mac (http://www.mamp.info/en/index.html)
  WAMP for Windows (http://www.wampserver.com/)

  Ftp servers:
  FileZilla Server for Windows (http://filezilla-project.org/download.php?type=server)
  Pure-FTPd for Mac (http://www.pureftpd.org/project/pure-ftpd)

  Once the user has his own server installed, he can put his data on the server, copy the URL to the file (http://publicIP/path/to/file or ftp://user:passw@publicIP/path/to/file) and paste the URL into the URL Text box of the upload tool. After clicking on “execute”, the upload will start.
- **A more complete tutorial can be found at the main Galaxy server: https://main.g2.bx.psu.edu/ -> Live Quickies: Uploading Data using FTP, Galactic quickie #17**

Read statistics

- Run “flagstat” – to see how many reads were mapped



flagstat

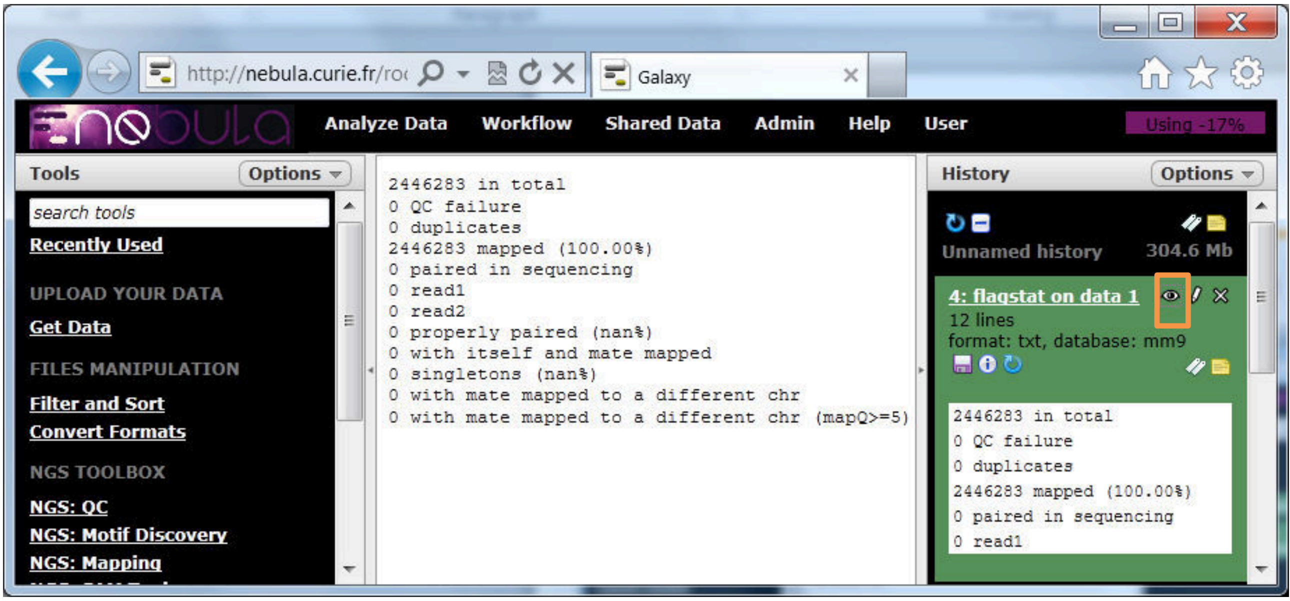
BAM File to Convert: 2: TF_chr1_sorted.bam

Execute

What it does

This tool uses the SAMTools toolkit to produce simple stats on a BAM file.

- flagstat output:



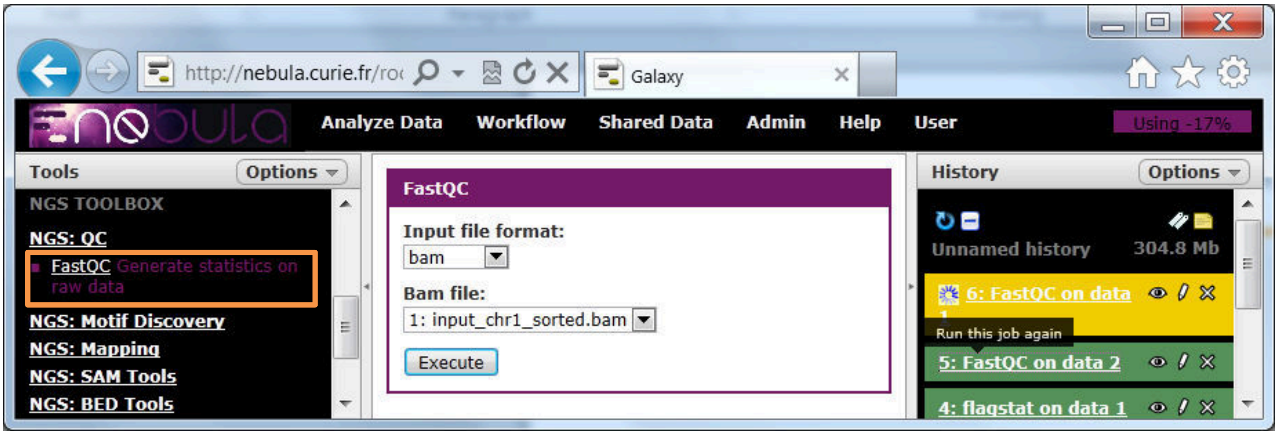
```
2446283 in total
0 QC failure
0 duplicates
2446283 mapped (100.00%)
0 paired in sequencing
0 read1
0 read2
0 properly paired (nan%)
0 with itself and mate mapped
0 singletons (nan%)
0 with mate mapped to a different chr
0 with mate mapped to a different chr (mapQ>=5)
```

4: flagstat on data 1
12 lines
format: txt, database: mm9

```
2446283 in total
0 QC failure
0 duplicates
2446283 mapped (100.00%)
0 paired in sequencing
0 read1
```

Check read quality before calling peaks

- Run FASTQC – to see statistics on read quality



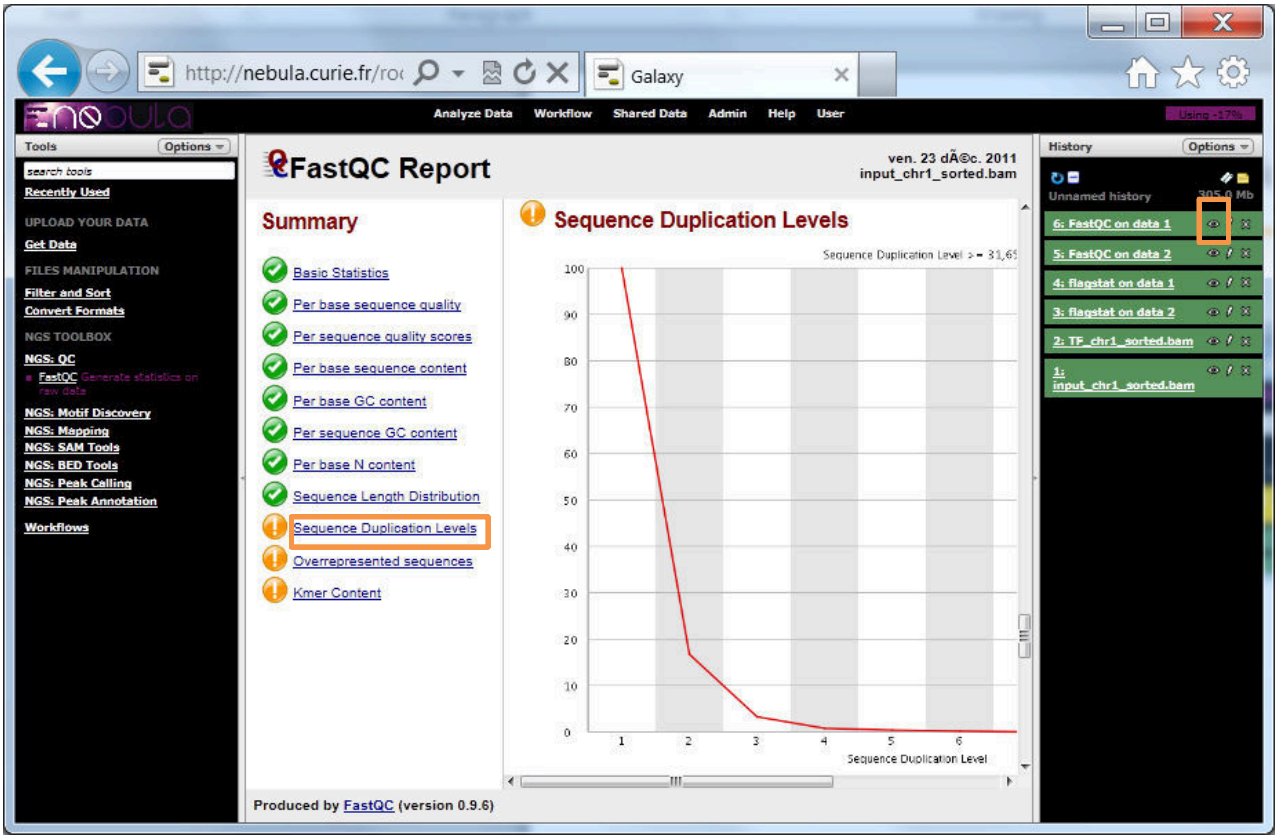
FastQC

Input file format: bam

Bam file: 1: input_chr1_sorted.bam

Execute

- Check FASTQC output:



FastQC Report

ven. 23 dÃ©c. 2011
input_chr1_sorted.bam

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per base GC content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ! Kmer Content

! Sequence Duplication Levels

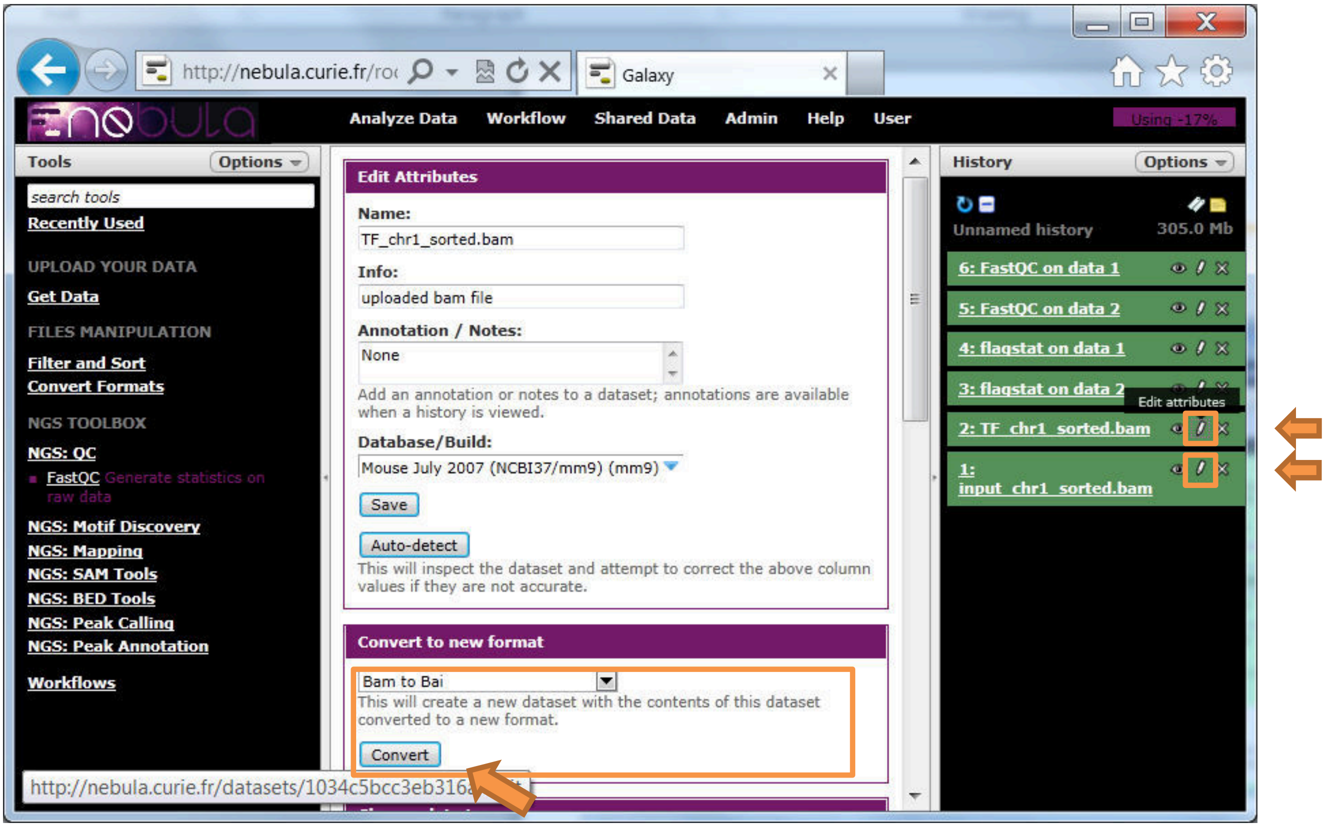
Produced by FastQC (version 0.9.6)

Check read quality before calling peaks

- Check how many reads you have in total by looking at the output of ‘flagstat’ (p. 7).
- How many reads were promised by the sequencing facilities? ☺
- I would say that 20 million mapped reads should be OK. In our example we have more then 2 million reads on chr1 (0.07 of the total mouse genome), this corresponds to about 30 million reads for the whole genome.
- Check the proportion of duplicate reads (‘FASTQC’, p. 8). High level of PCR duplicates means that you provided to little material for sequencing.
- Check whether you will have enough reads when you filter out duplicates. In our case we have about 30% of reads which are duplicates. Looking at the graph, we can say that filtering of duplicate reads will remove about 20% of reads. So it is still OK to continue our analysis and do peak calling. (MACS and FindPeaks will remove duplicate reads for you).

Visualize .SAM/.BAM files in UCSC

- First create an index (.bai) for .BAM files



Edit Attributes

Name: TF_chr1_sorted.bam

Info: uploaded bam file

Annotation / Notes: None

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build: Mouse July 2007 (NCBI37/mm9) (mm9)

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

Bam to Bai

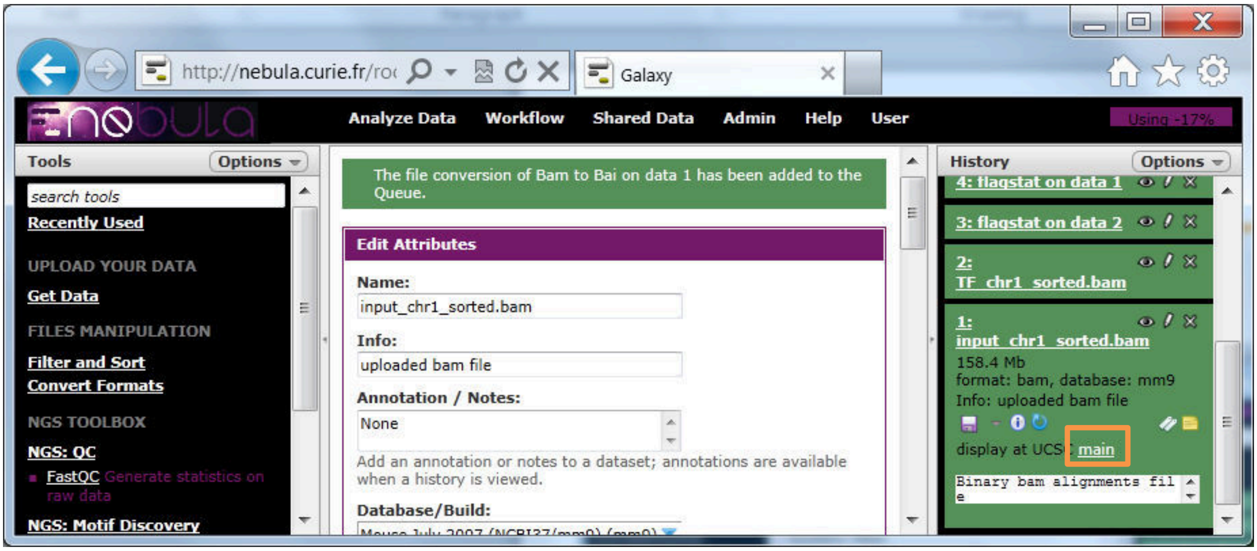
This will create a new dataset with the contents of this dataset converted to a new format.

Convert

- Do it both for **TF_chr1_sorted.bam** and **input_chr1_sorted.bam** !

Visualize .SAM/.BAM files in UCSC

- Click on ‘main’ of the initial .BAM file to visualize it in UCSC



The file conversion of Bam to Bai on data 1 has been added to the Queue.

Edit Attributes

Name: input_chr1_sorted.bam

Info: uploaded bam file

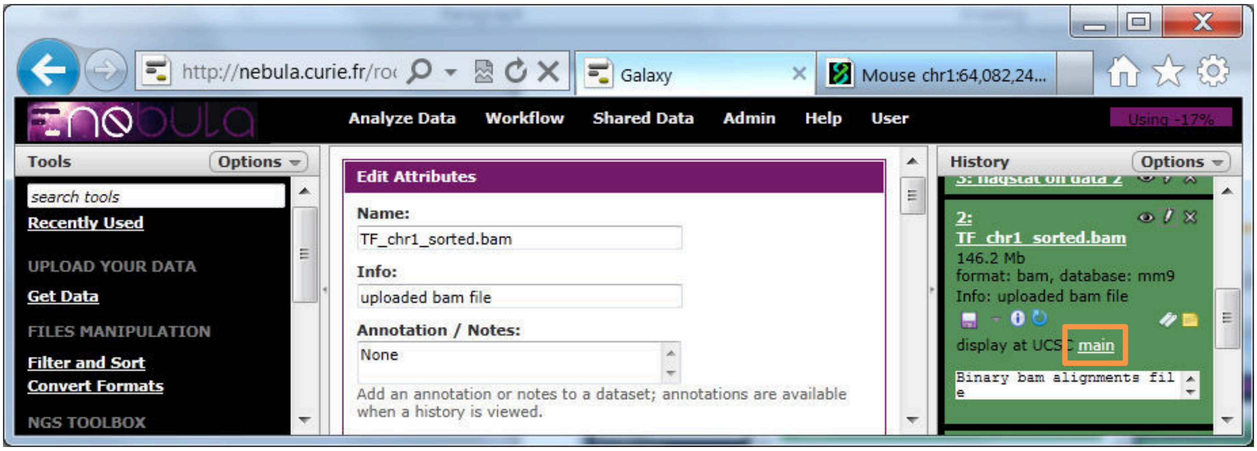
Annotation / Notes: None

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build: Mouse July 2007 (NCBI37/mm9) (mm9)

1: input_chr1_sorted.bam
158.4 Mb
format: bam, database: mm9
Info: uploaded bam file
display at UCSC main
Binary bam alignments file

- Do it twice: **TF_chr1_sorted.bam** and **input_chr1_sorted.bam** !
- Uploaded tracks will stay in your UCSC for several days. You can close and open the UCSC browser when you want and you won’t lose your tracks.



Edit Attributes

Name: TF_chr1_sorted.bam

Info: uploaded bam file

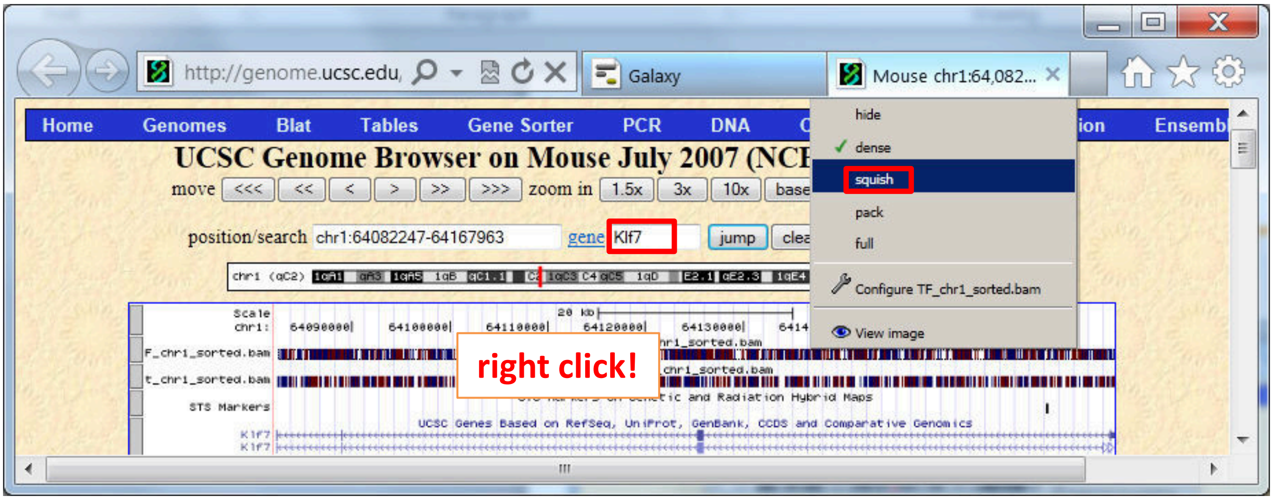
Annotation / Notes: None

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

2: TF_chr1_sorted.bam
146.2 Mb
format: bam, database: mm9
Info: uploaded bam file
display at UCSC main
Binary bam alignments file

Visualize .SAM/.BAM files in UCSC

- Go to the **Klf7** gene and change **view of the track**

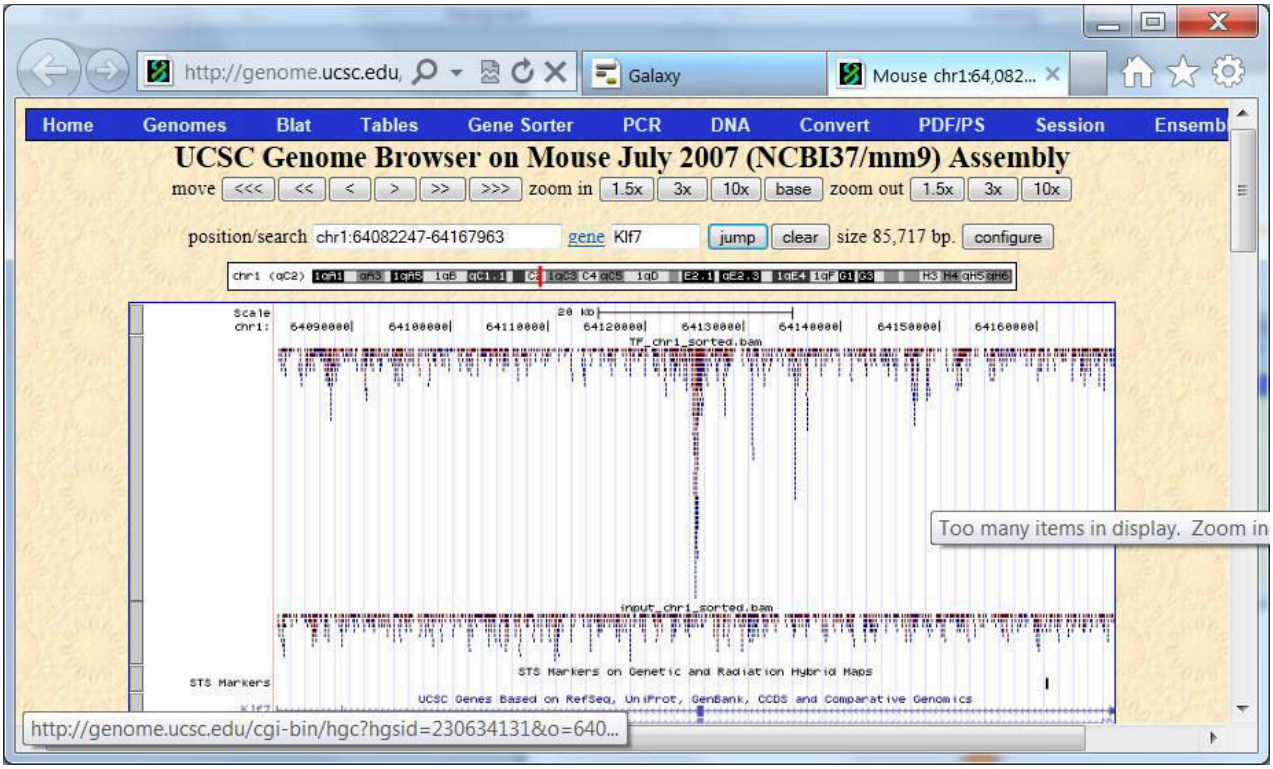


http://genome.ucsc.edu/

position/search chr1:64082247-64167963 gene Klf7

Galaxy
Mouse chr1:64,082...
Home
Genomes
Blat
Tables
Gene Sorter
PCR
DNA
Ensembl
UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base
jump
clear
hide
✓ dense
squish
pack
full
Configure TF_chr1_sorted.bam
View image
Scale
chr1:
64090000
64100000
64110000
64120000
64130000
64140000
20 kb
TF_chr1_sorted.bam
input_chr1_sorted.bam
STS Markers
STS Markers on Genetic and Radiation Hybrid Maps
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Klf7
Klf7

right click!



http://genome.ucsc.edu/

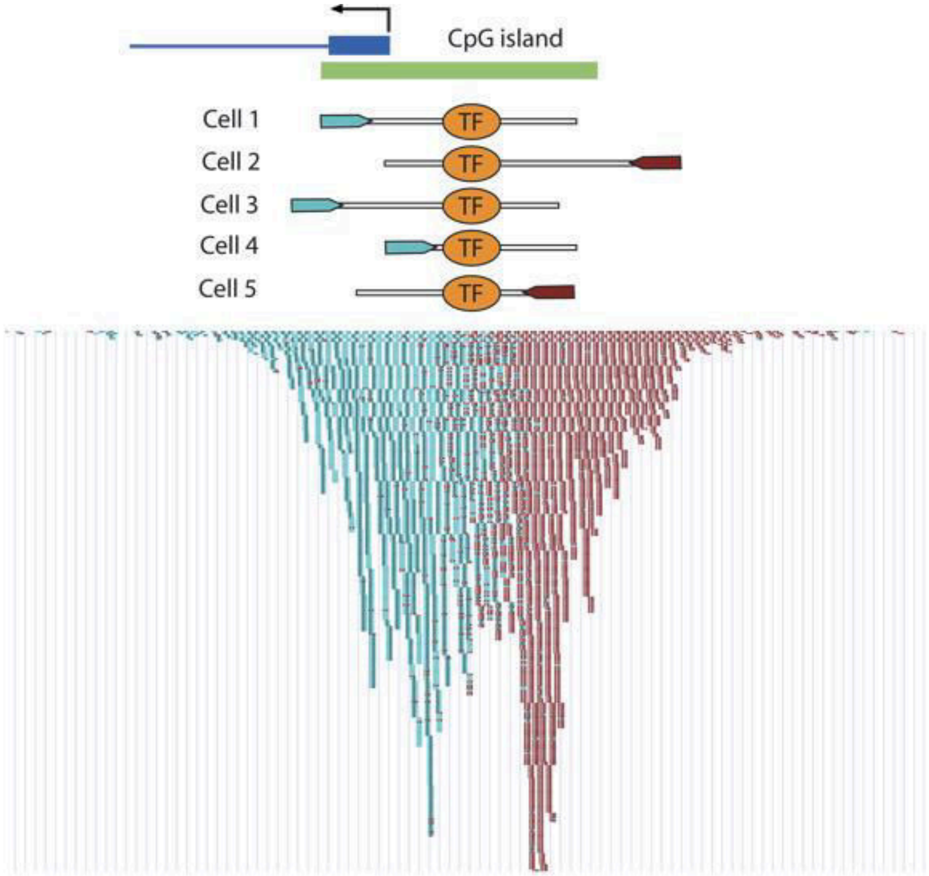
position/search chr1:64082247-64167963 gene Klf7 size 85,717 bp.

Too many items in display. Zoom in

http://genome.ucsc.edu/cgi-bin/hgc?hgsid=230634131&o=640...

Galaxy
Mouse chr1:64,082...
Home
Genomes
Blat
Tables
Gene Sorter
PCR
DNA
Convert
PDF/PS
Session
Ensembl
UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
jump
clear
configure
Scale
chr1:
64090000
64100000
64110000
64120000
64130000
64140000
64150000
64160000
20 kb
TF_chr1_sorted.bam
input_chr1_sorted.bam
STS Markers
STS Markers on Genetic and Radiation Hybrid Maps
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Klf7

How does a good binging site look like?

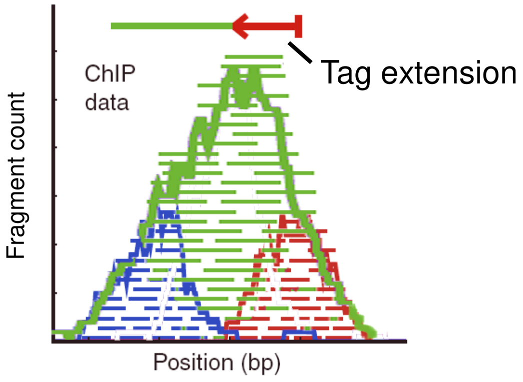


(from Valouev et al., Nat Methods 2008)

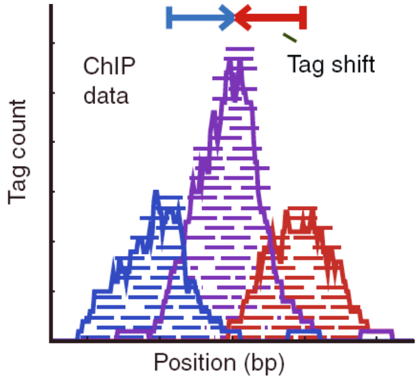
- In our case the separation of forward and revers reads is not as clear. This is because it is SOLiD reads and we performed double sonication (one before and one after immunoprecipitation)

There exist two main ways to extract the signal (construct peaks)

- Tools:
  - FindPeaks
  - QuEST
  - F-Seq
  - SICER
  - CisGenome
  - GLITR
  - PeakSeq
  - Spp
  - Useq
  - MACS
  - ERANGE
  - SiSSRs
- Main methods:



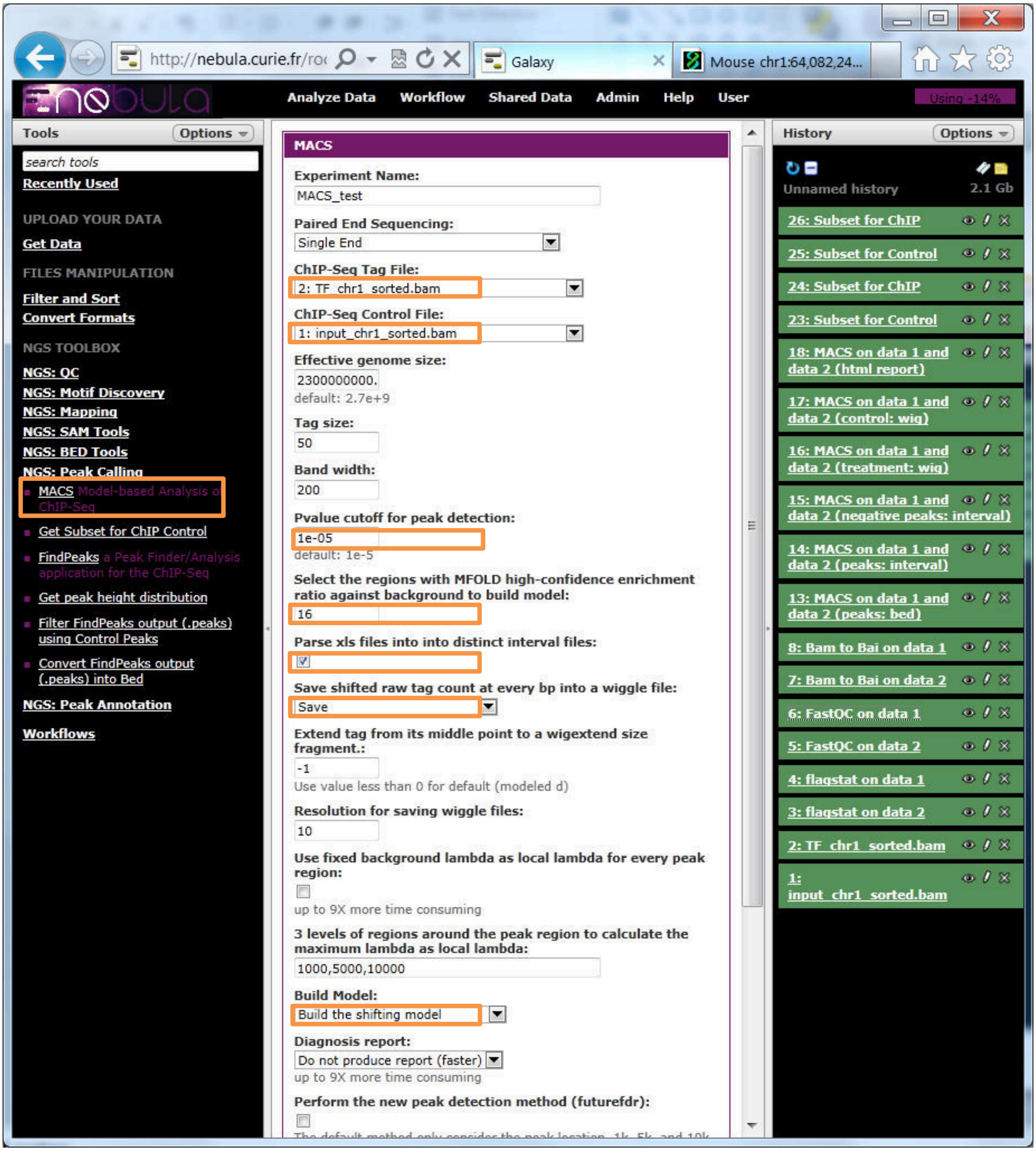
(FindPeaks)



Adopted from S. Pepke et al., 2009 Nat Methods

(MACS)

Run MACS (if you want to compare its output with the output of FindPeaks. **You can skip this step**.)



MACS

Experiment Name: MACS_test

Paired End Sequencing: Single End

ChIP-Seq Tag File: 2: TF_chr1_sorted.bam

ChIP-Seq Control File: 1: input_chr1_sorted.bam

Effective genome size: 2300000000.
default: 2.7e+9

Tag size: 50

Band width: 200

Pvalue cutoff for peak detection: 1e-05
default: 1e-5

Select the regions with MFOLD high-confidence enrichment ratio against background to build model: 16

Parse xls files into into distinct interval files: ☑

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

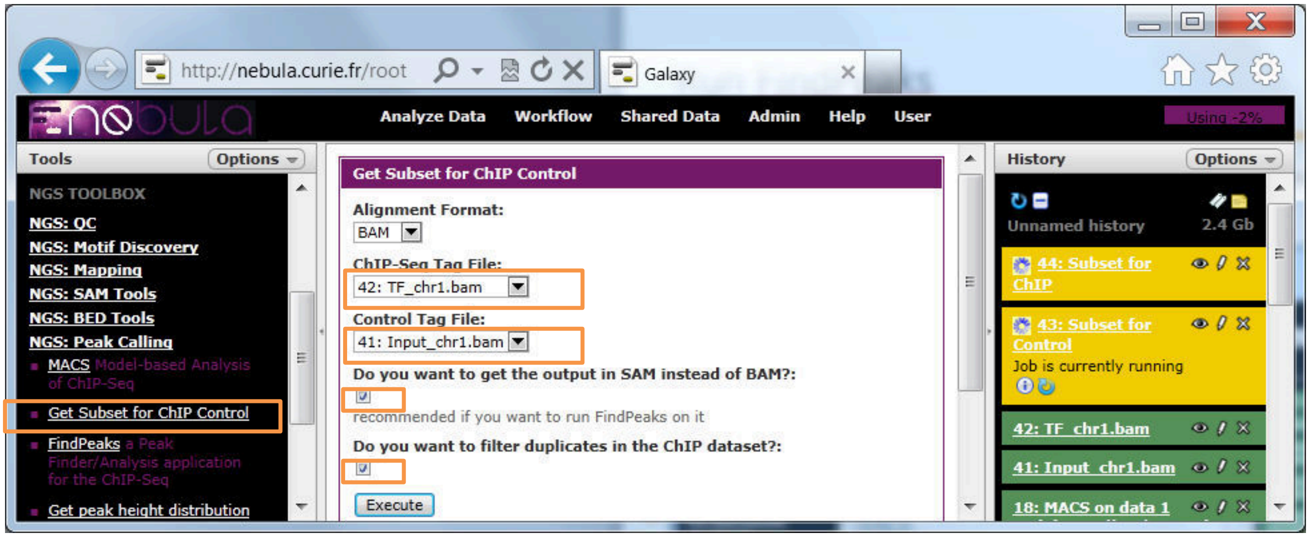
Diagnosis report: Do not produce report (faster)
up to 9X more time consuming

Perform the new peak detection method (futurefdr): □
The default method only consider the peak location, 1k, 5k, and 10k

- Band width: This value is only used while building the shifting model. Should be ≥ DNA fragment lengths.
- For transcription factors, it is important to check '**Parse xls files into into distinct interval files**' to get the locations of peak summits for peak annotation.

Run FindPeaks

- Create a subset of the control dataset if there are more reads for the control sample than for the ChIP sample



Get Subset for ChIP Control

Alignment Format: BAM

ChIP-Seq Tag File: 42: TF_chr1.bam

Control Tag File: 41: Input_chr1.bam

Do you want to get the output in SAM instead of BAM?: ☑

recommended if you want to run FindPeaks on it

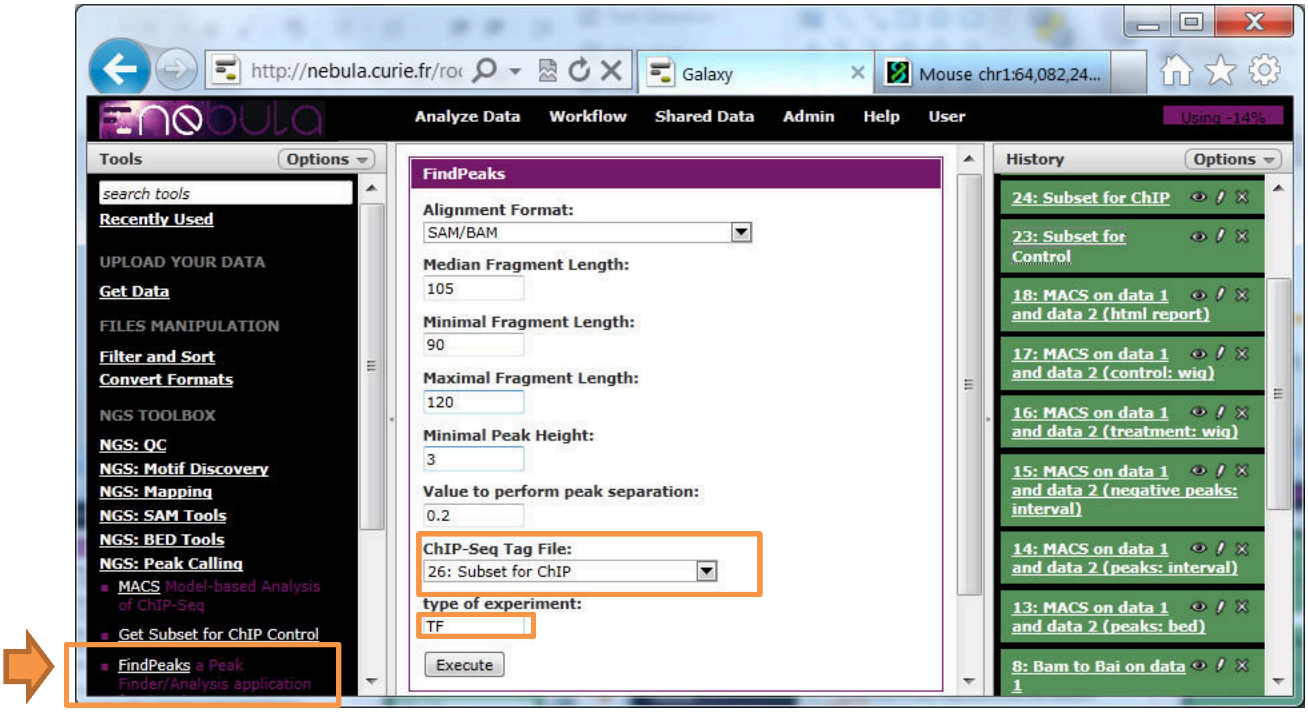
Do you want to filter duplicates in the ChIP dataset?: ☑

- This command will
  1. filter out duplicate reads from your ChIP and Control datasets,
  2. randomly select reads from the Control sample so that the total number of reads in both sample were equal.
  3. Transform .BAM into .SAM, because for some unknown reason FindPeaks does not like some .BAM …

- If you have the same number of reads in the ChIP and the control sample, you will be able to compare their outputs later on and filter out peaks detected in both datasets. Imagine, you have 10 times more reads in the control? – Then your real signal in the ChIP can appear weak…

Run FindPeaks

- Run FindPeaks on the TF and Input sample (twice!)



FindPeaks

Alignment Format: SAM/BAM

Median Fragment Length: 105

Minimal Fragment Length: 90

Maximal Fragment Length: 120

Minimal Peak Height: 3

Value to perform peak separation: 0.2

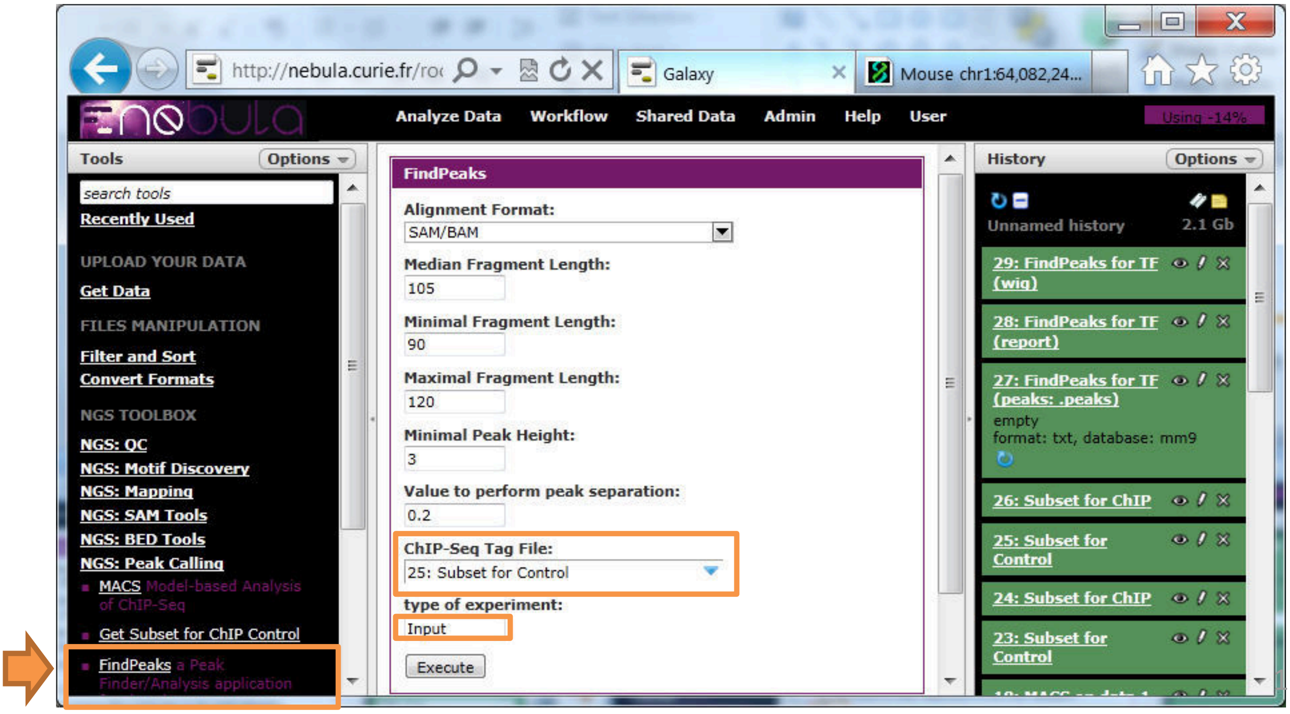
ChIP-Seq Tag File: 26: Subset for ChIP

type of experiment: TF

Execute

History:

- 24: Subset for ChIP
- 23: Subset for Control
- 18: MACS on data 1 and data 2 (html report)
- 17: MACS on data 1 and data 2 (control: wig)
- 16: MACS on data 1 and data 2 (treatment: wig)
- 15: MACS on data 1 and data 2 (negative peaks: interval)
- 14: MACS on data 1 and data 2 (peaks: interval)
- 13: MACS on data 1 and data 2 (peaks: bed)
- 8: Bam to Bai on data 1



FindPeaks

Alignment Format: SAM/BAM

Median Fragment Length: 105

Minimal Fragment Length: 90

Maximal Fragment Length: 120

Minimal Peak Height: 3

Value to perform peak separation: 0.2

ChIP-Seq Tag File: 25: Subset for Control

type of experiment: Input

Execute

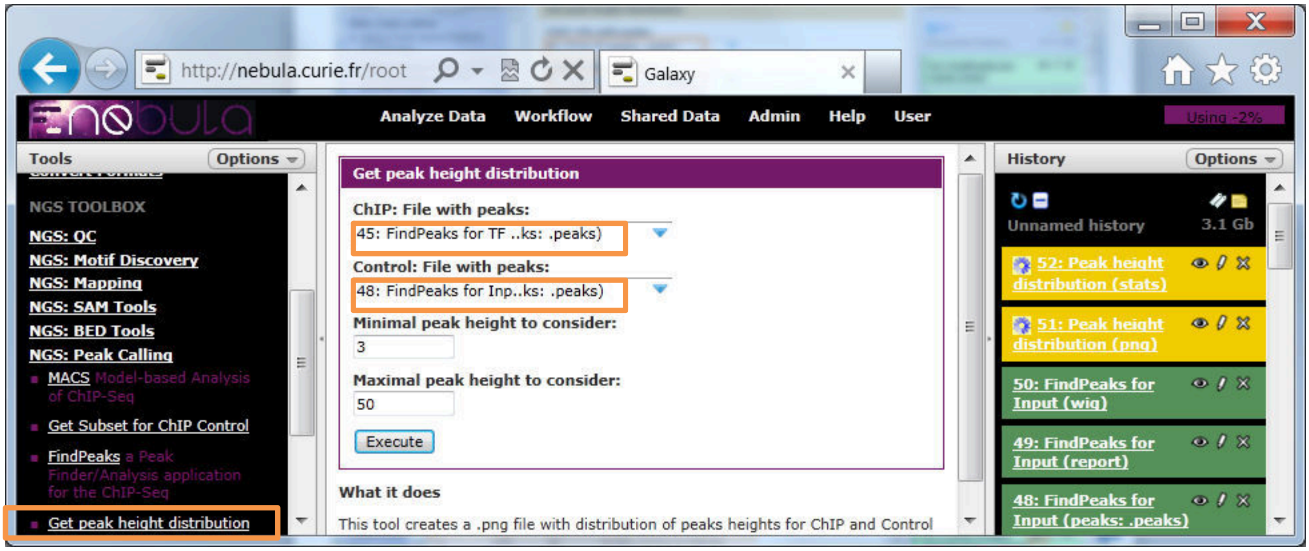
History:

Unnamed history 2.1 Gb

- 29: FindPeaks for TF (wig)
- 28: FindPeaks for TF (report)
- 27: FindPeaks for TF (peaks: .peaks)
  empty
  format: txt, database: mm9
- 26: Subset for ChIP
- 25: Subset for Control
- 24: Subset for ChIP
- 23: Subset for Control
- 18: MACS on data 1

Calculate peak height distribution – immunoprecipitation quality control

- You should enter FindPeaks output files **(.peaks)** for the **TF** and **Input**


http://nebula.curie.fr/root
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 2%
Tools
Options
NGS TOOLBOX
NGS: QC
NGS: Motif Discovery
NGS: Mapping
NGS: SAM Tools
NGS: BED Tools
NGS: Peak Calling
MACS Model-based Analysis of ChIP-Seq
Get Subset for ChIP Control
FindPeaks a Peak Finder/Analysis application for the ChIP-Seq
Get peak height distribution
History
Options
Unnamed history
3.1 Gb
52: Peak height distribution (stats)
51: Peak height distribution (png)
50: FindPeaks for Input (wig)
49: FindPeaks for Input (report)
48: FindPeaks for Input (peaks: .peaks)

Get peak height distribution

ChIP: File with peaks:
45: FindPeaks for TF ..ks: .peaks)

Control: File with peaks:
48: FindPeaks for Inp..ks: .peaks)

Minimal peak height to consider:
3

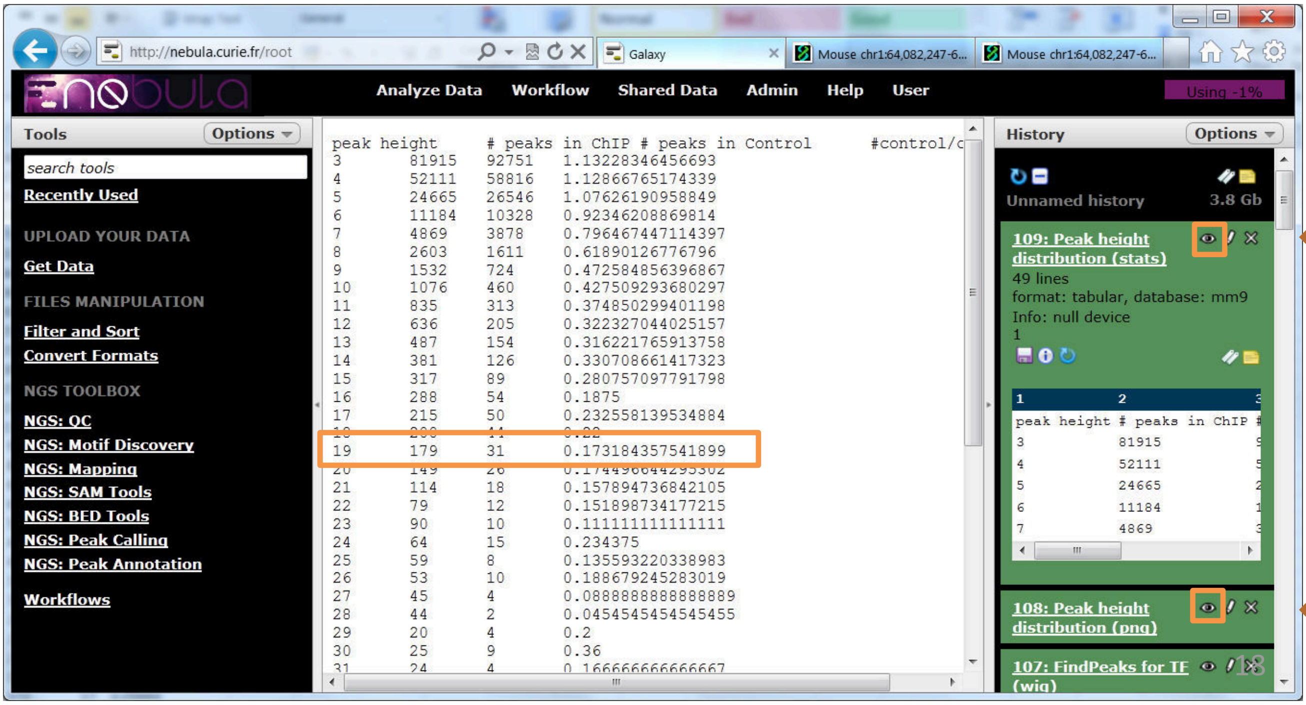
Maximal peak height to consider:
50

Execute

What it does

This tool creates a .png file with distribution of peaks heights for ChIP and Control

- You can the select the minimal peak height for further analyses using the calculated evaluation of false discovery rate:



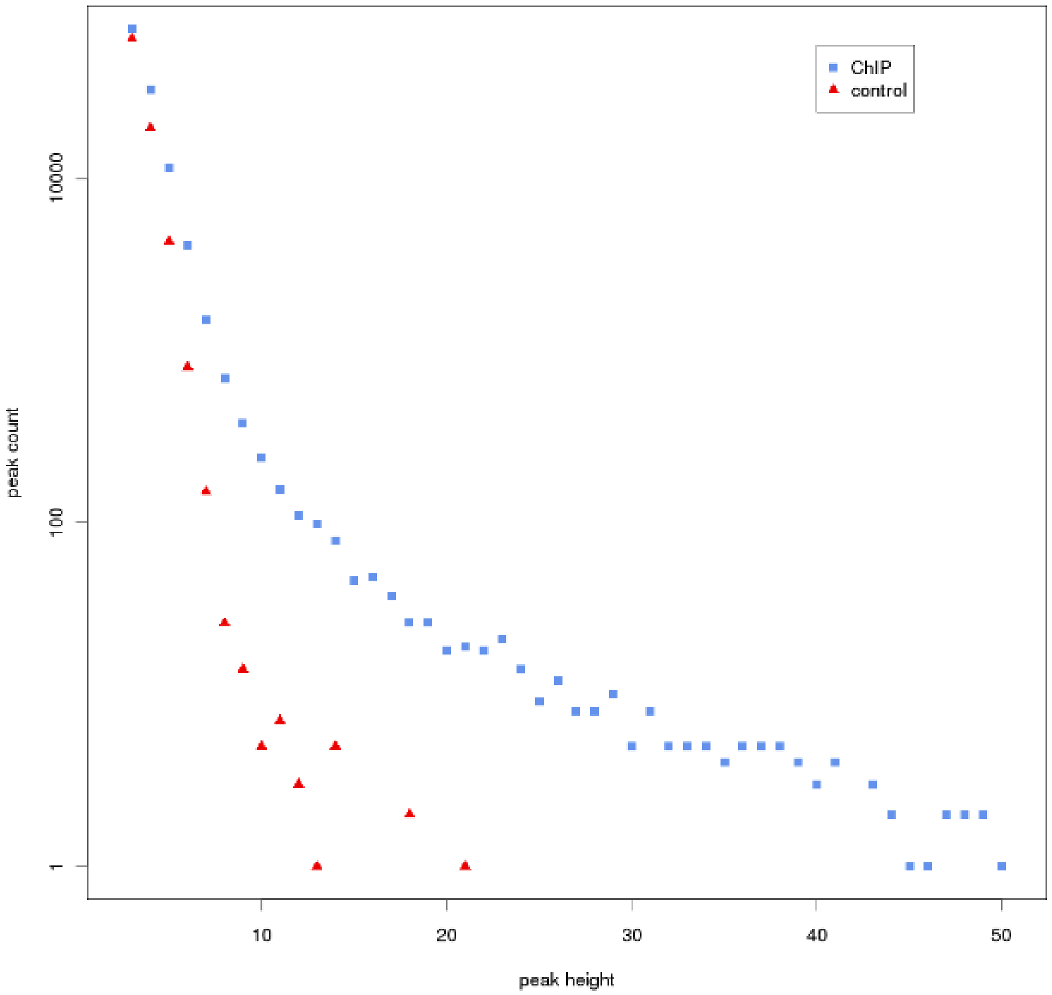
```
peak height     # peaks in ChIP # peaks in Control     #control/c
3         81915    92751    1.13228346456693
4         52111    58816    1.12866765174339
5         24665    26546    1.07626190958849
6         11184    10328    0.92346208869814
7         4869     3878     0.796467447114397
8         2603     1611     0.61890126776796
9         1532     724      0.472584856396867
10        1076     460      0.427509293680297
11        835      313      0.374850299401198
12        636      205      0.322327044025157
13        487      154      0.316221765913758
14        381      126      0.330708661417323
15        317      89       0.280757097791798
16        288      54       0.1875
17        215      50       0.232558139534884
18        200      44       0.22
19        179      31       0.173184357541899
20        149      26       0.174496644295302
21        114      18       0.157894736842105
22        79       12       0.151898734177215
23        90       10       0.111111111111111
24        64       15       0.234375
25        59       8        0.135593220338983
26        53       10       0.188679245283019
27        45       4        0.0888888888888889
28        44       2        0.0454545454545455
29        20       4        0.2
30        25       9        0.36
31        24       4        0.166666666666667
```

109: Peak height distribution (stats)
49 lines
format: tabular, database: mm9
Info: null device
1

```
1            2
peak height  # peaks in ChIP
3            81915
4            52111
5            24665
6            11184
7            4869
```

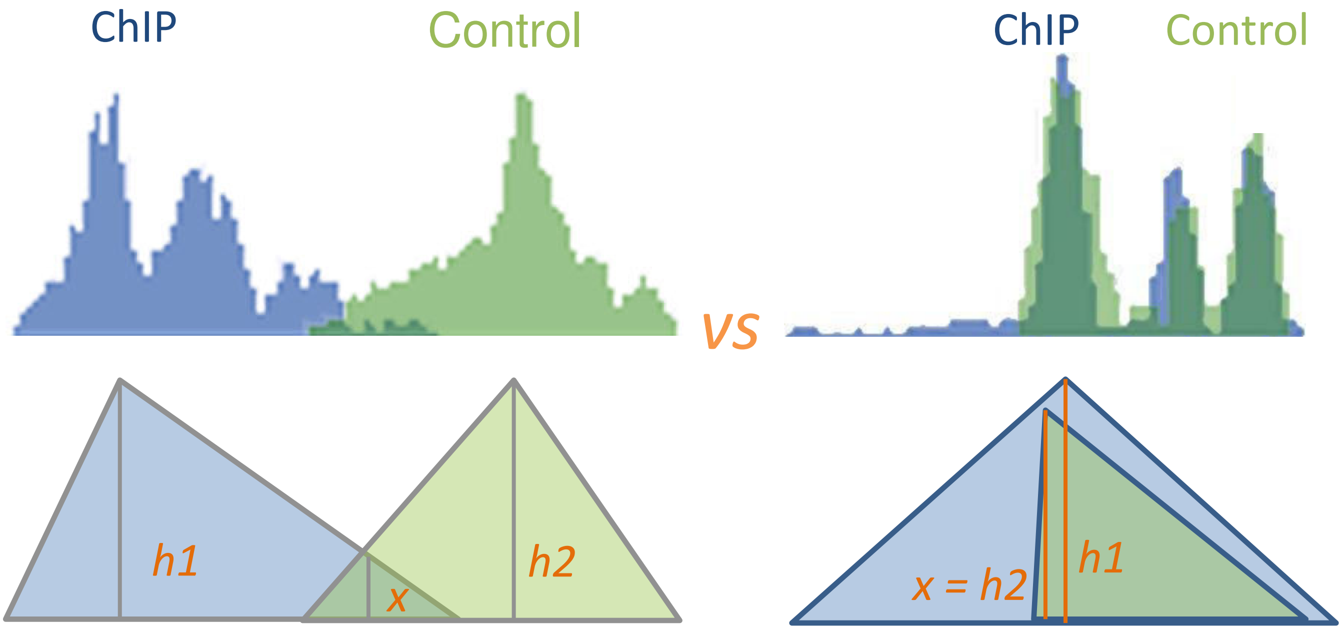
Calculate peak height distribution – immunoprecipitation quality control

- More high peaks in the ChIP sample – the better the immunoprecipitation was preformed

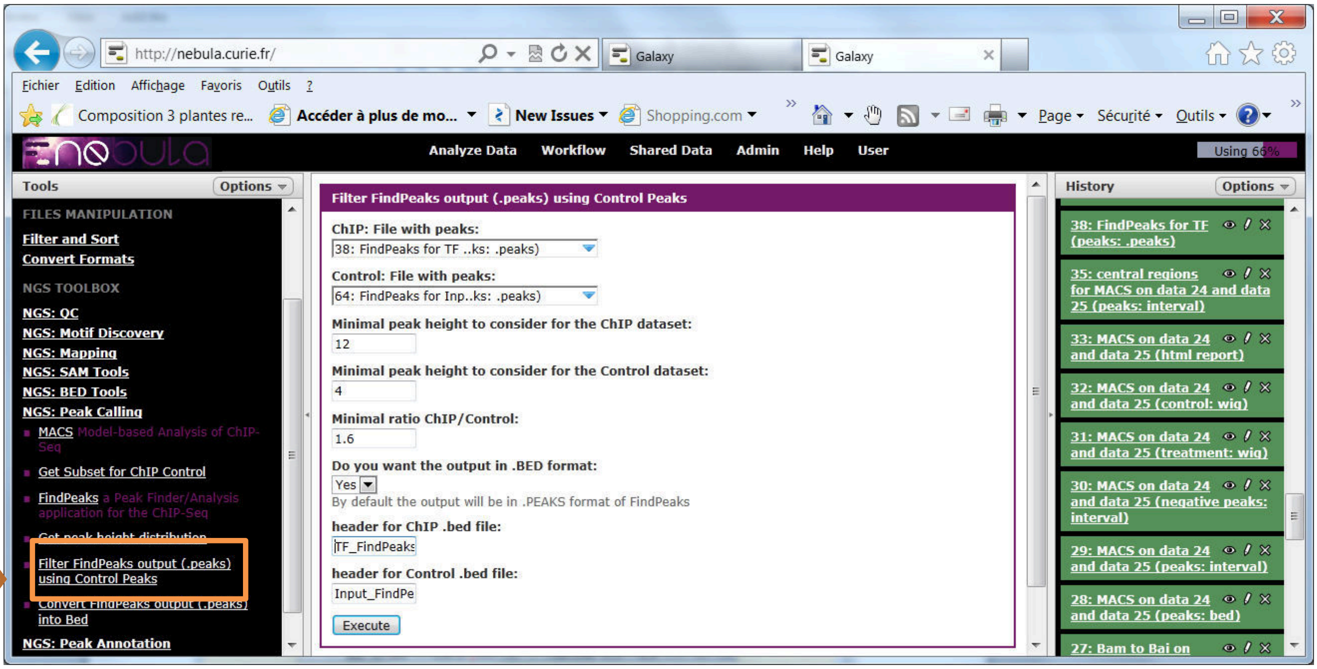


Filter FindPeaks' output using peaks from the control dataset

- The actual peak shapes is replaced by triangles (start, end, maximum and height). Then, the height (*x*) of maximal overlap is calculated. The ChIP peak is rejected if its height (*h1*) divided by *x* is less than or equal to a given threshold.



h1/*x* > 2? ➡ Keep the peak


http://nebula.curie.fr/
Galaxy
Galaxy
Fichier Edition Affichage Favoris Outils ?
Composition 3 plantes re...
Accéder à plus de mo...
New Issues
Shopping.com
Page
Sécurité
Outils
nebula
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 66%
Tools
Options
FILES MANIPULATION
Filter and Sort
Convert Formats
NGS TOOLBOX
NGS: QC
NGS: Motif Discovery
NGS: Mapping
NGS: SAM Tools
NGS: BED Tools
NGS: Peak Calling
MACS Model-based Analysis of ChIP-Seq
Get Subset for ChIP Control
FindPeaks a Peak Finder/Analysis application for the ChIP-Seq
Get peak height distribution
Filter FindPeaks output (.peaks) using Control Peaks
Convert FindPeaks output (.peaks) into Bed
NGS: Peak Annotation
History
Options
38: FindPeaks for TF (peaks: .peaks)
35: central regions for MACS on data 24 and data 25 (peaks: interval)
33: MACS on data 24 and data 25 (html report)
32: MACS on data 24 and data 25 (control: wig)
31: MACS on data 24 and data 25 (treatment: wig)
30: MACS on data 24 and data 25 (negative peaks: interval)
29: MACS on data 24 and data 25 (peaks: interval)
28: MACS on data 24 and data 25 (peaks: bed)
27: Bam to Bai on
Execute

Filter FindPeaks output (.peaks) using Control Peaks

ChIP: File with peaks:
38: FindPeaks for TF ..ks: .peaks)

Control: File with peaks:
64: FindPeaks for Inp..ks: .peaks)

Minimal peak height to consider for the ChIP dataset:
12

Minimal peak height to consider for the Control dataset:
4

Minimal ratio ChIP/Control:
1.6

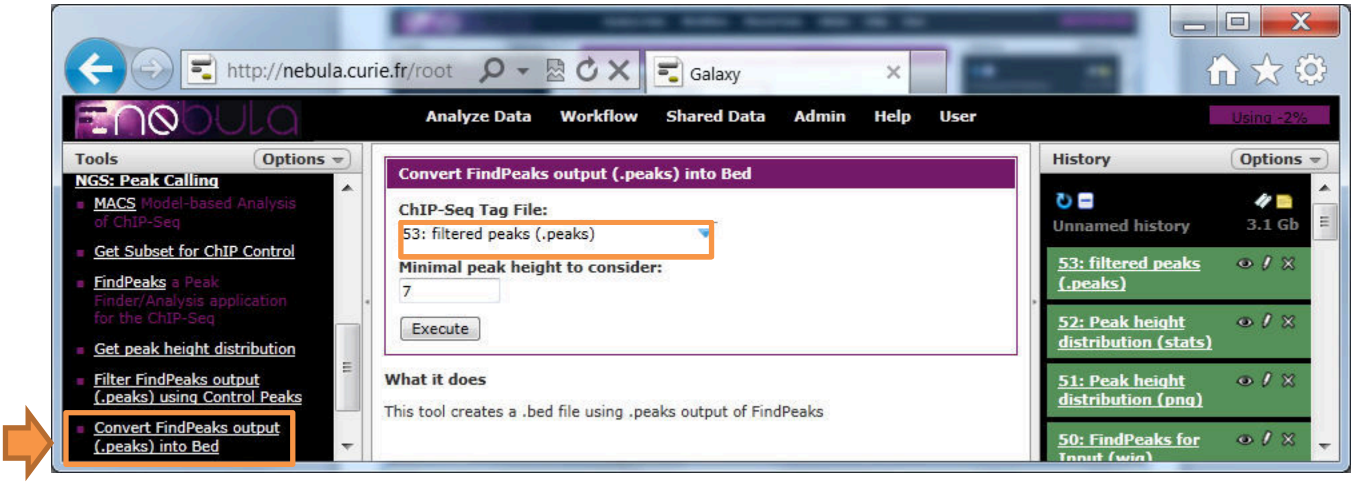
Do you want the output in .BED format:
Yes
By default the output will be in .PEAKS format of FindPeaks

header for ChIP .bed file:
TF_FindPeaks

header for Control .bed file:
Input_FindPe

Convert FindPeaks output (.peaks) into Bed if you did **not** select output in .BED at the previous step

- Convert “filtered” peaks into .BED (.BED is a standard format for genomic intervals):


http://nebula.curie.fr/root
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 2%
Tools
Options
NGS: Peak Calling
MACS Model-based Analysis of ChIP-Seq
Get Subset for ChIP Control
FindPeaks a Peak Finder/Analysis application for the ChIP-Seq
Get peak height distribution
Filter FindPeaks output (.peaks) using Control Peaks
Convert FindPeaks output (.peaks) into Bed
History
Options
Execute

Convert FindPeaks output (.peaks) into Bed

ChIP-Seq Tag File: 53: filtered peaks (.peaks)

Minimal peak height to consider: 7

What it does

This tool creates a .bed file using .peaks output of FindPeaks

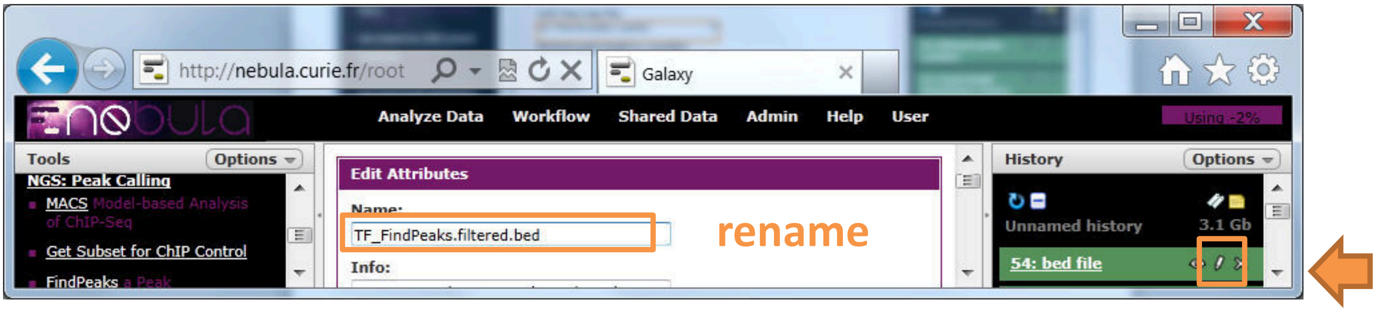
Unnamed history 3.1 Gb

53: filtered peaks (.peaks)

52: Peak height distribution (stats)

51: Peak height distribution (png)

50: FindPeaks for Input (wig)



Edit Attributes

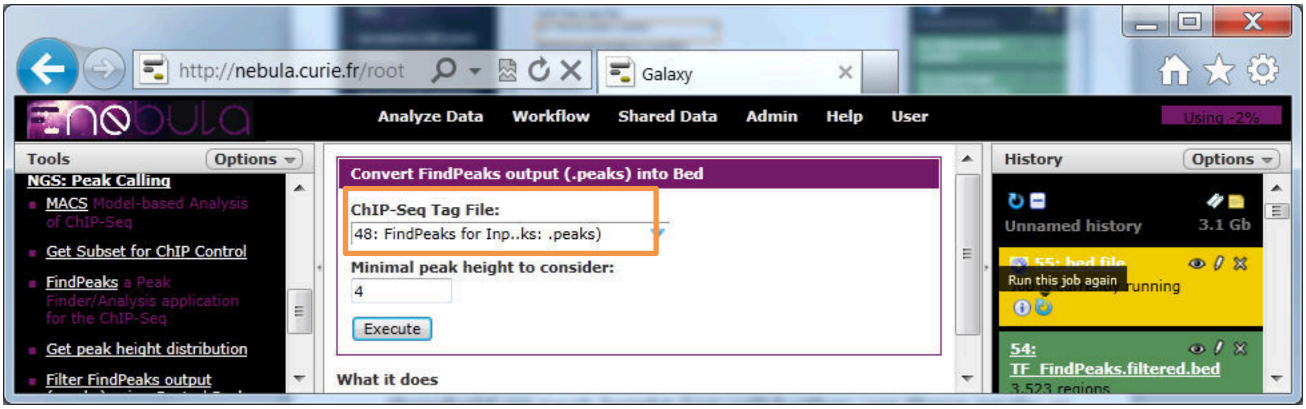
Name: TF_FindPeaks.filtered.bed

Info:

Unnamed history 3.1 Gb

54: bed file

- Convert the “control” peaks too. One should use a low threshold on peak height (we will further use these peaks as “random” control for peak location distribution):


http://nebula.curie.fr/root
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 2%
Tools
Options
NGS: Peak Calling
MACS Model-based Analysis of ChIP-Seq
Get Subset for ChIP Control
FindPeaks a Peak Finder/Analysis application for the ChIP-Seq
Get peak height distribution
Filter FindPeaks output
History
Options
Execute
Run this job again

Convert FindPeaks output (.peaks) into Bed

ChIP-Seq Tag File: 48: FindPeaks for Inp..ks: .peaks)

Minimal peak height to consider: 4

What it does

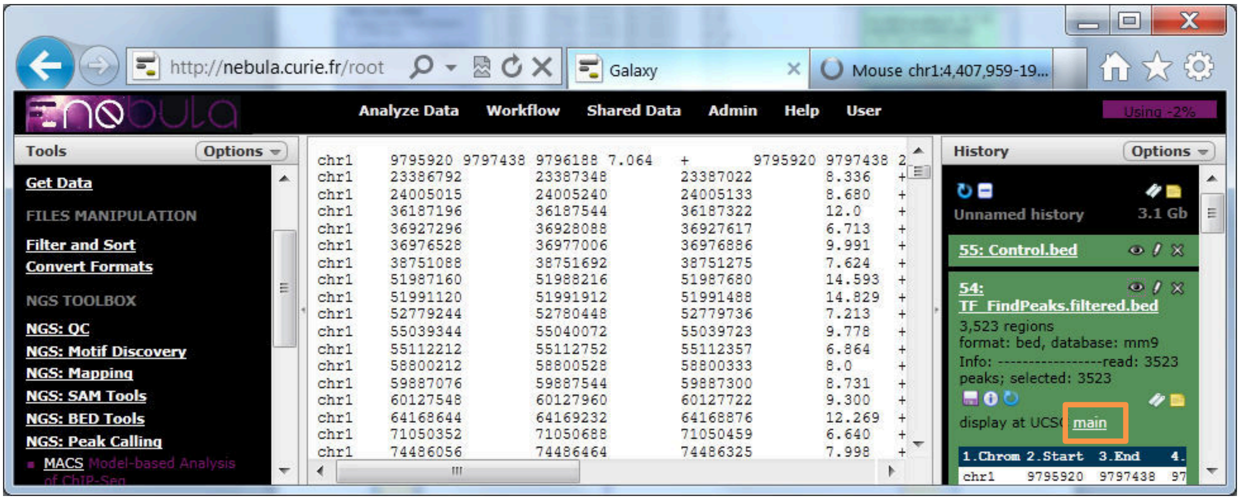
Unnamed history 3.1 Gb

55: bed file running

54: TF_FindPeaks.filtered.bed 3,523 regions

Visualize .bed in UCSC

- For .BED: visualize directly in UCSC



```
chr1     9795920 9797438 9796188 7.064   +      9795920 9797438 2
chr1     23386792        23387348        23387022        8.336   +
chr1     24005015        24005240        24005133        8.680   +
chr1     36187196        36187544        36187322        12.0    +
chr1     36927296        36928088        36927617        6.713   +
chr1     36976528        36977006        36976886        9.991   +
chr1     38751088        38751692        38751275        7.624   +
chr1     51987160        51988216        51987680        14.593  +
chr1     51991120        51991912        51991488        14.829  +
chr1     52779244        52780448        52779736        7.213   +
chr1     55039344        55040072        55039723        9.778   +
chr1     55112212        55112752        55112357        6.864   +
chr1     58800212        58800528        58800333        8.0     +
chr1     59887076        59887544        59887300        8.731   +
chr1     60127548        60127960        60127722        9.300   +
chr1     64168644        64169232        64168876        12.269  +
chr1     71050352        71050688        71050459        6.640   +
chr1     74486056        74486464        74486325        7.998   +
```

Unnamed history 3.1 Gb

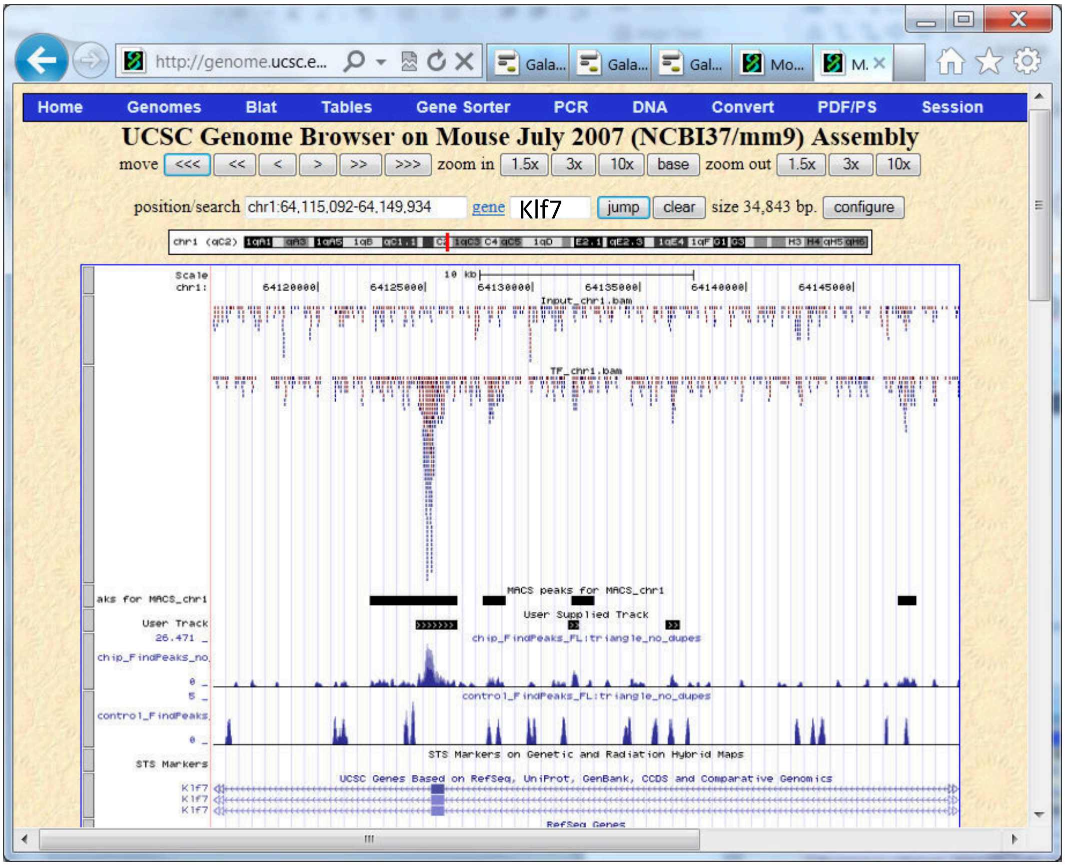
55: Control.bed

54: TF_FindPeaks.filtered.bed

3,523 regions
format: bed, database: mm9
Info: ----------------read: 3523 peaks; selected: 3523

display at UCSC main

```
1.Chrom 2.Start   3.End     4.
chr1     9795920  9797438   97
```

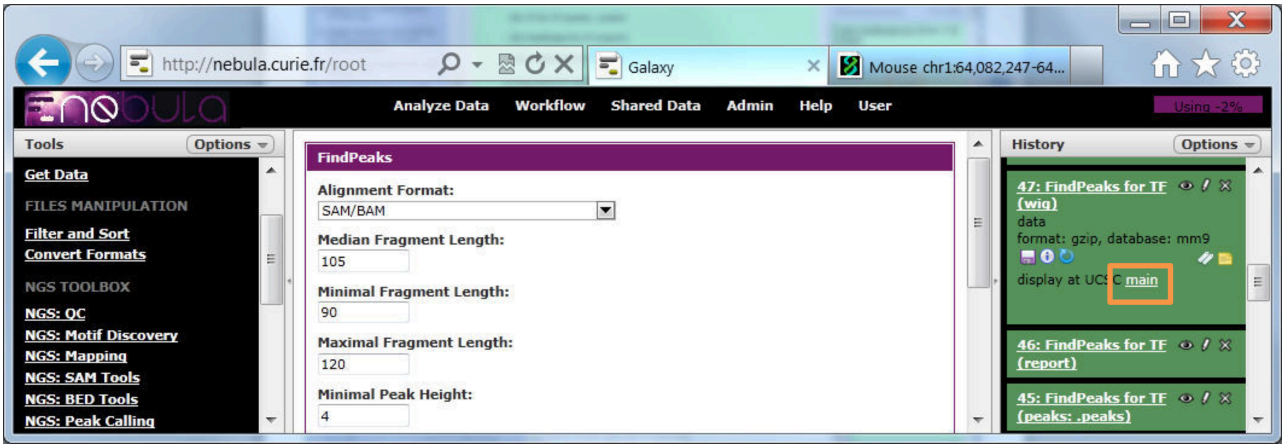


UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9) Assembly

position/search chr1:64,115,092-64,149,934 gene Klf7 size 34,843 bp.

Visualize .wig in UCSC

- For .wig.gz (output of FindPeaks): visualize directly in UCSC



FindPeaks

Alignment Format: SAM/BAM

Median Fragment Length: 105

Minimal Fragment Length: 90

Maximal Fragment Length: 120

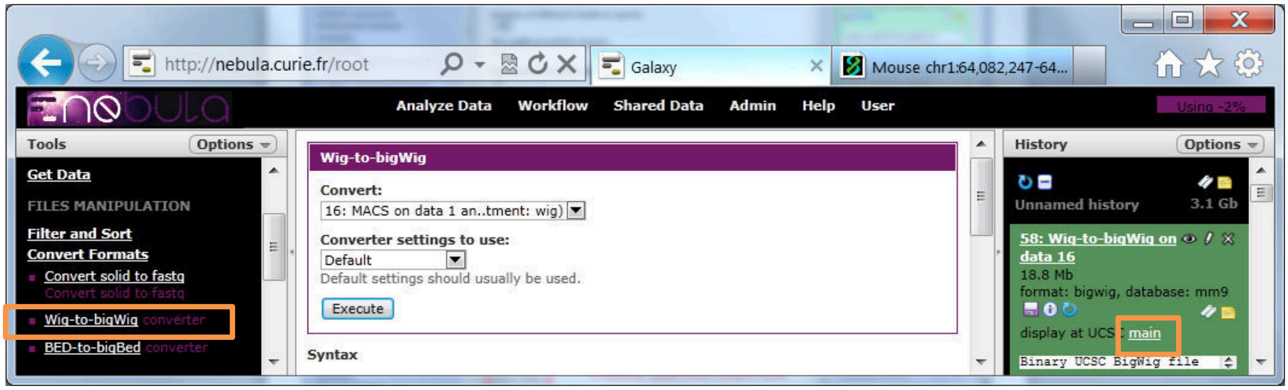
Minimal Peak Height: 4

47: FindPeaks for TF (wig)
data
format: gzip, database: mm9
display at UCSC main

46: FindPeaks for TF (report)

45: FindPeaks for TF (peaks: .peaks)

- For .Wig (output of MACS): you need to convert .wig to .bw (Big Wig) first and then you will visualyze the BigWig file:



Wig-to-bigWig

Convert: 16: MACS on data 1 an..tment: wig)

Converter settings to use: Default

Default settings should usually be used.

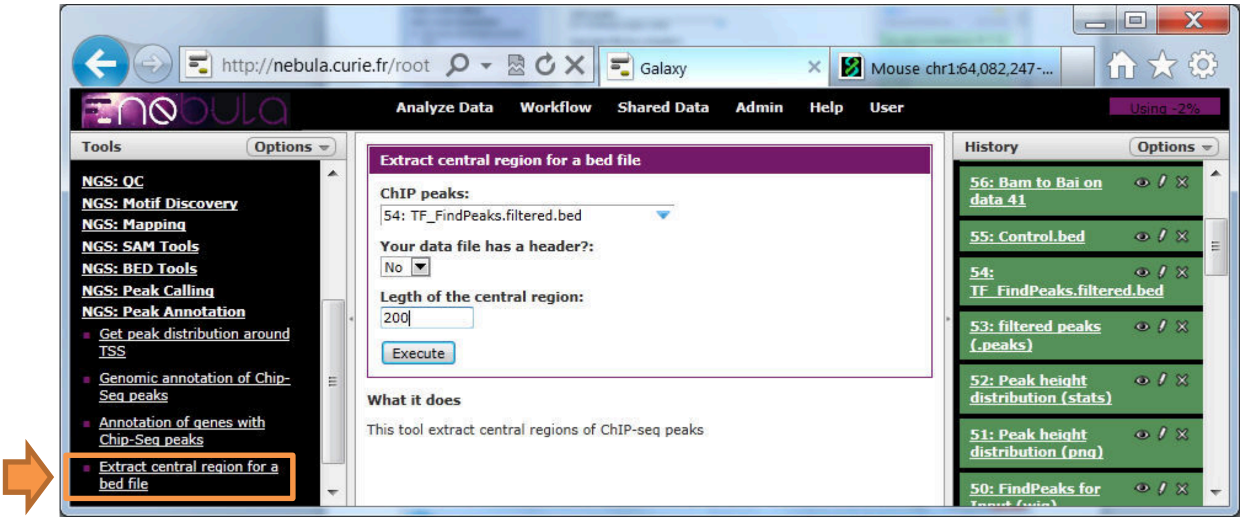
Syntax

Unnamed history 3.1 Gb

58: Wig-to-bigWig on data 16
18.8 Mb
format: bigwig, database: mm9
display at UCSC main
Binary UCSC BigWig file

Get .fasta sequences to find over-represented motifs

- Create .bed with coordinates of central regions of peaks (FindPeaks output: use .bed file)



http://nebula.curie.fr/root
Galaxy
Mouse chr1:64,082,247-...
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 2%
Tools
Options
NGS: QC
NGS: Motif Discovery
NGS: Mapping
NGS: SAM Tools
NGS: BED Tools
NGS: Peak Calling
NGS: Peak Annotation
Get peak distribution around TSS
Genomic annotation of Chip-Seq peaks
Annotation of genes with Chip-Seq peaks
Extract central region for a bed file
History
Options
56: Bam to Bai on data 41
55: Control.bed
54: TF_FindPeaks.filtered.bed
53: filtered peaks (.peaks)
52: Peak height distribution (stats)
51: Peak height distribution (png)
50: FindPeaks for Input (wig)

Extract central region for a bed file

ChIP peaks: 54: TF_FindPeaks.filtered.bed

Your data file has a header?: No

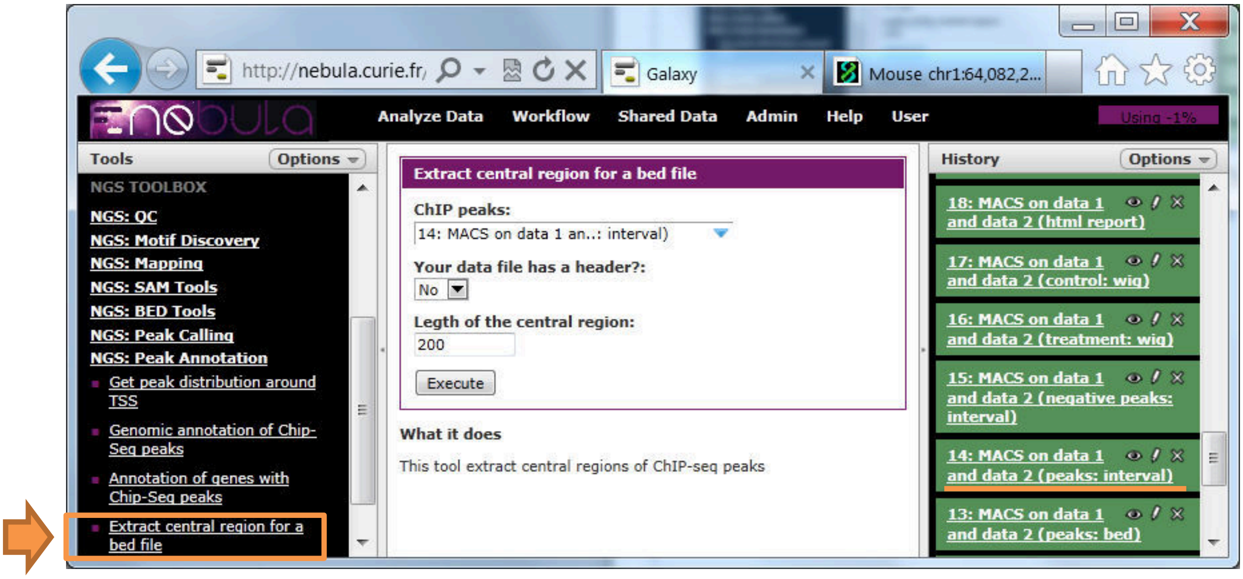
Legth of the central region: 200

Execute

What it does

This tool extract central regions of ChIP-seq peaks

- If you want to extract central regions for MACS use “peaks: interval” file instead of “peaks: bed”, since the former contains information about peak summits:



http://nebula.curie.fr/
Galaxy
Mouse chr1:64,082,2...
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 1%
Tools
Options
NGS TOOLBOX
NGS: QC
NGS: Motif Discovery
NGS: Mapping
NGS: SAM Tools
NGS: BED Tools
NGS: Peak Calling
NGS: Peak Annotation
Get peak distribution around TSS
Genomic annotation of Chip-Seq peaks
Annotation of genes with Chip-Seq peaks
Extract central region for a bed file
History
Options
18: MACS on data 1 and data 2 (html report)
17: MACS on data 1 and data 2 (control: wig)
16: MACS on data 1 and data 2 (treatment: wig)
15: MACS on data 1 and data 2 (negative peaks: interval)
14: MACS on data 1 and data 2 (peaks: interval)
13: MACS on data 1 and data 2 (peaks: bed)

Extract central region for a bed file

ChIP peaks: 14: MACS on data 1 an..: interval)

Your data file has a header?: No

Legth of the central region: 200

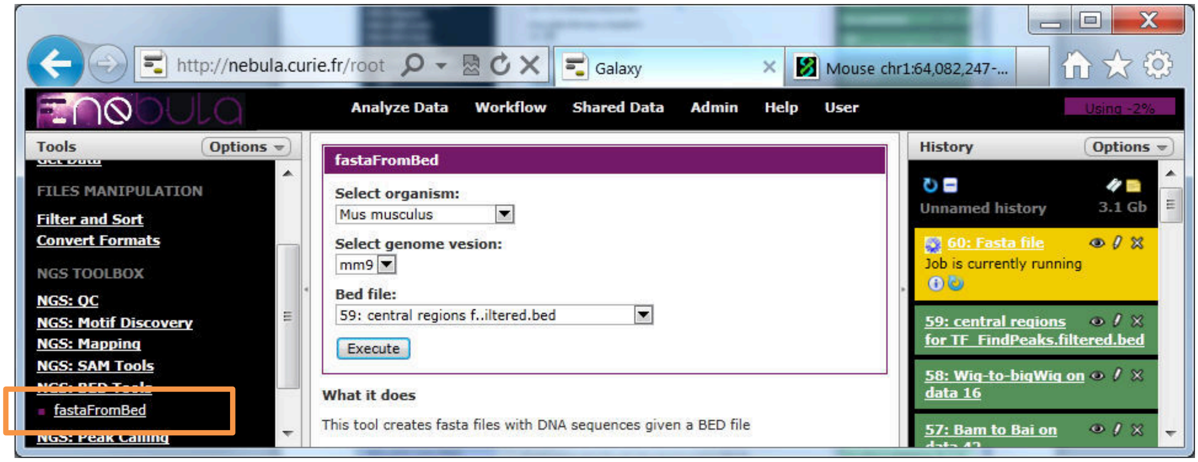
Execute

What it does

This tool extract central regions of ChIP-seq peaks

Get .fasta sequences to find over-represented motifs

- Extract .fasta



fastaFromBed

Select organism: Mus musculus

Select genome vesion: mm9

Bed file: 59: central regions f...iltered.bed

What it does

This tool creates fasta files with DNA sequences given a BED file

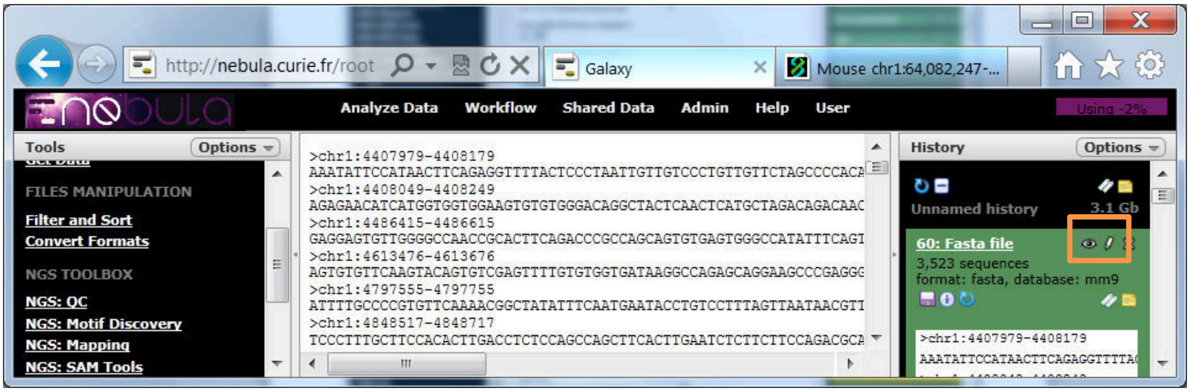
Unnamed history 3.1 Gb

60: Fasta file
Job is currently running

59: central regions for TF_FindPeaks.filtered.bed

58: Wig-to-bigWig on data 16

57: Bam to Bai on data 42



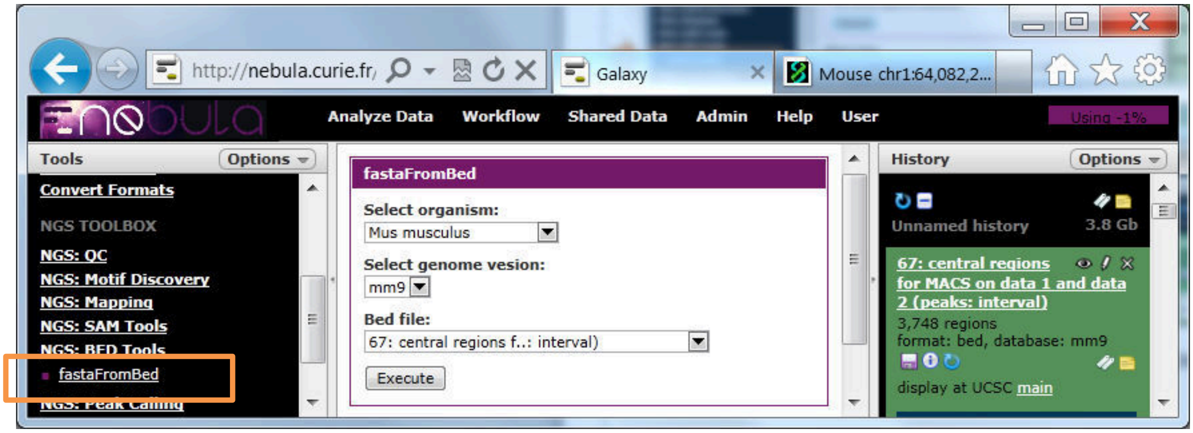
```
>chr1:4407979-4408179
AAATATTCCATAACTTCAGAGGTTTTACTCCCTAATTGTTGTCCCTGTTGTTCTAGCCCCACA
>chr1:4408049-4408249
AGAGAACATCATGGTGGTGGAAGTGTGTGGGACAGGCTACTCAACTCATGCTAGACAGACAAC
>chr1:4486415-4486615
GAGGAGTGTTGGGGCCAACCGCACTTCAGACCCGCCAGCAGTGTGAGTGGGCCATATTTCAGT
>chr1:4613476-4613676
AGTGTGTTCAAGTACAGTGTCGAGTTTTGTGTGGTGATAAGGCCAGAGCAGGAAGCCCGAGGG
>chr1:4797555-4797755
ATTTTGCCCCGTGTTCAAAACGGCTATATTTCAATGAATACCTGTCCTTTAGTTAATAACGTT
>chr1:4848517-4848717
TCCCTTTGCTTCCACACTTGACCTCTCCAGCCAGCTTCACTTGAATCTCTTCTTCCAGACGCA
```

Unnamed history 3.1 Gb

60: Fasta file
3,523 sequences
format: fasta, database: mm9

```
>chr1:4407979-4408179
AAATATTCCATAACTTCAGAGGTTTTA
```

- Extract .fasta for MACS central peak regions too if you used MACS peak calling:



fastaFromBed

Select organism: Mus musculus

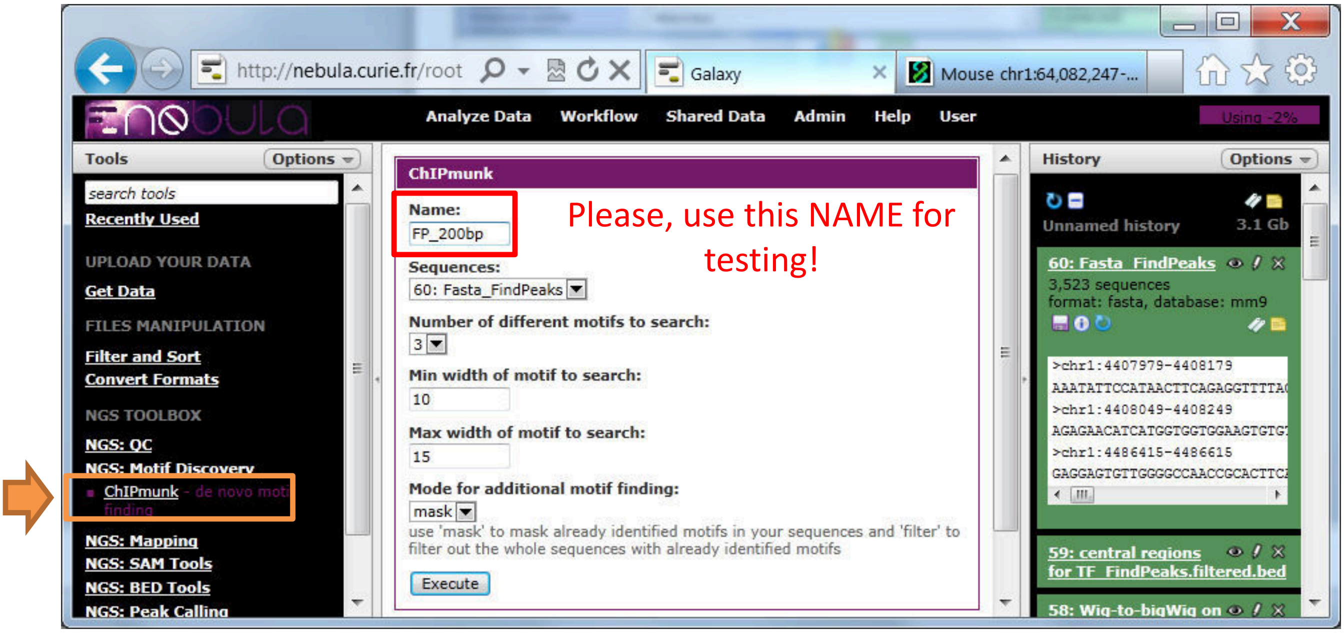
Select genome vesion: mm9

Bed file: 67: central regions f...: interval)

Unnamed history 3.8 Gb

67: central regions for MACS on data 1 and data 2 (peaks: interval)
3,748 regions
format: bed, database: mm9
display at UCSC main

Run motif finding on central regions of peaks



ChIPmunk

Name: FP_200bp

Please, use this NAME for testing!

Sequences: 60: Fasta_FindPeaks

Number of different motifs to search: 3

Min width of motif to search: 10

Max width of motif to search: 15

Mode for additional motif finding: mask

use 'mask' to mask already identified motifs in your sequences and 'filter' to filter out the whole sequences with already identified motifs

Unnamed history 3.1 Gb

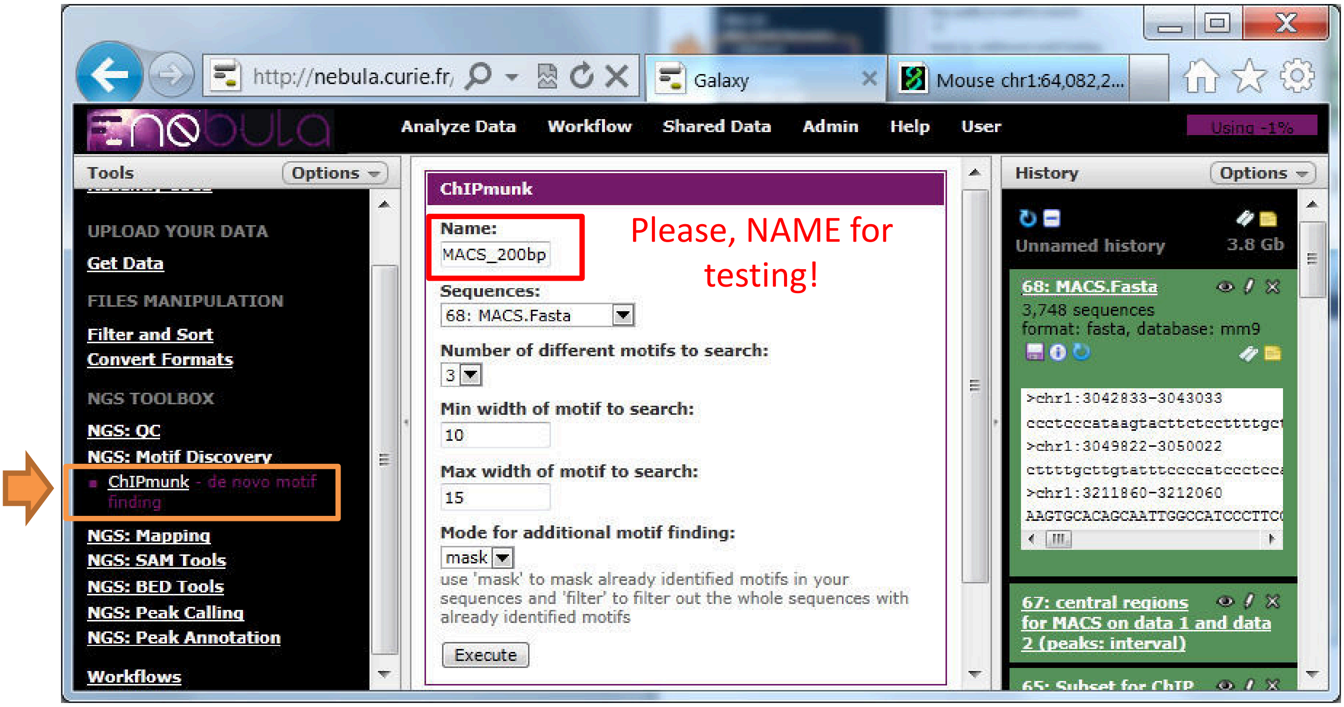
60: Fasta_FindPeaks
3,523 sequences
format: fasta, database: mm9

```
>chr1:4407979-4408179
AAATATTCCATAACTTCAGAGGTTTTA
>chr1:4408049-4408249
AGAGAACATCATGGTGGTGGAAGTGTG
>chr1:4486415-4486615
GAGGAGTGTTGGGGCCAACCGCACTTC
```

59: central regions for TF_FindPeaks.filtered.bed

58: Wig-to-bigWig on

- If you also created .fasta for the MACS peaks (p. 24), you can run motif finding on them too:



ChIPmunk

Name: MACS_200bp

Please, NAME for testing!

Sequences: 68: MACS.Fasta

Number of different motifs to search: 3

Min width of motif to search: 10

Max width of motif to search: 15

Mode for additional motif finding: mask

use 'mask' to mask already identified motifs in your sequences and 'filter' to filter out the whole sequences with already identified motifs

Unnamed history 3.8 Gb

68: MACS.Fasta
3,748 sequences
format: fasta, database: mm9

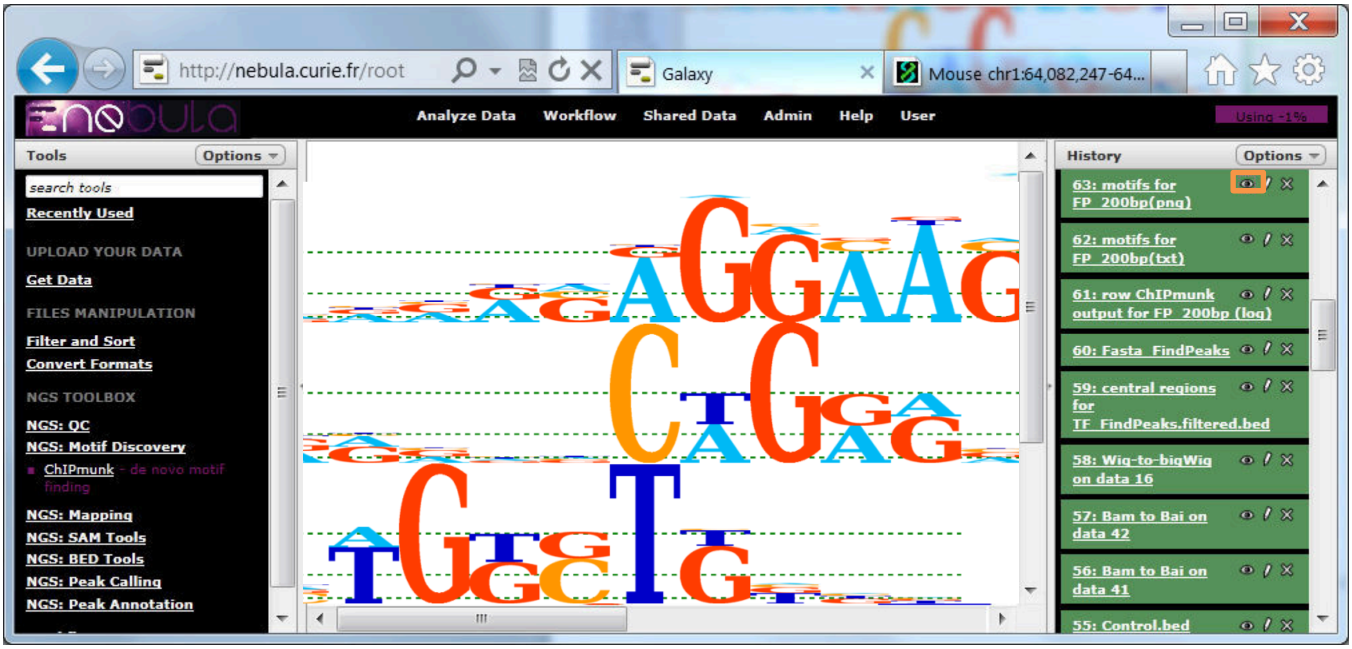
```
>chr1:3042833-3043033
ccctcccataagtacttctcctttttgc
>chr1:3049822-3050022
cttttgcttgtatttccccatccctcc
>chr1:3211860-3212060
AAGTGCACAGCAATTGGCCATCCCTTC
```

67: central regions for MACS on data 1 and data 2 (peaks: interval)

65: Subset for ChIP

Run motif finding on central regions of peaks

- Motifs found in peaks identified by FindPeaks (200bp central region, use UnZoom to see it better):



ChIPmunk has two modes to call multiple motifs

Mask sequences (“filter”)

Looking for several motifs of one TF

~~cgatcgagaCAGGAATGgct**agat**a~~

~~cacatgtacCAGGAATCcg**agat**at~~

~~acg**agat**cgCAGGAAAGgctacgat~~

~~cacat**agat**CCGGAATGcgatgcat~~

~~actgcgctgCAGGAATGagct**agat**~~

cac**agat**GGAAGGAAGGAAatgcat

agatcgcGGAAGGAAGGAActagca

CAGGAATG
CAGGAATC
CAGGAAAG
CCGGAATG
CAGGAATG

Motif 1

GGAAGGAAGGAA
GGAAGGAAGGAA

Motif 2

Mask motifs (“mask”)

Looking for motifs of co-factors

cgatcgaga~~CAGGAATG~~gct**agat**a

cacatgtac~~CAGGAATC~~cg**agat**at

acg**agat**cg~~CAGGAAAG~~gctacgat

cacat**agat**~~CCGGAATG~~cgatgcat

actgcgctg~~CAGGAATG~~agct**agat**

cac**agat**GGAAGGAAGGAAatgcat

agatcgcGGAAGGAAGGAActagca

CAGGAATG
CAGGAATC
CAGGAAAG
CCGGAATG
CAGGAATG

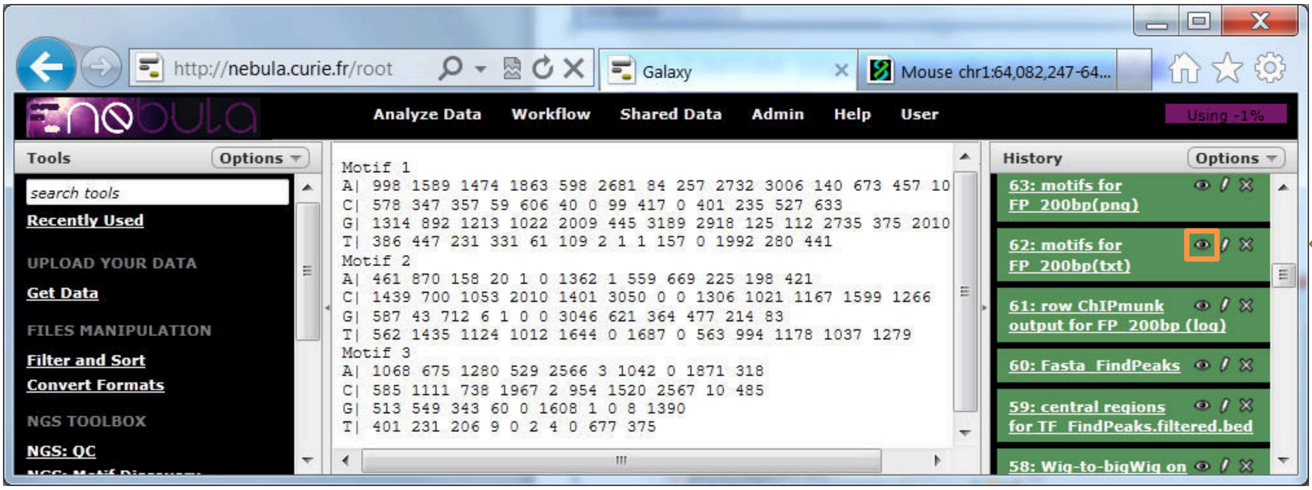
Motif 1

agat
agat
agat
agat
agat
agat
agat

Motif 2

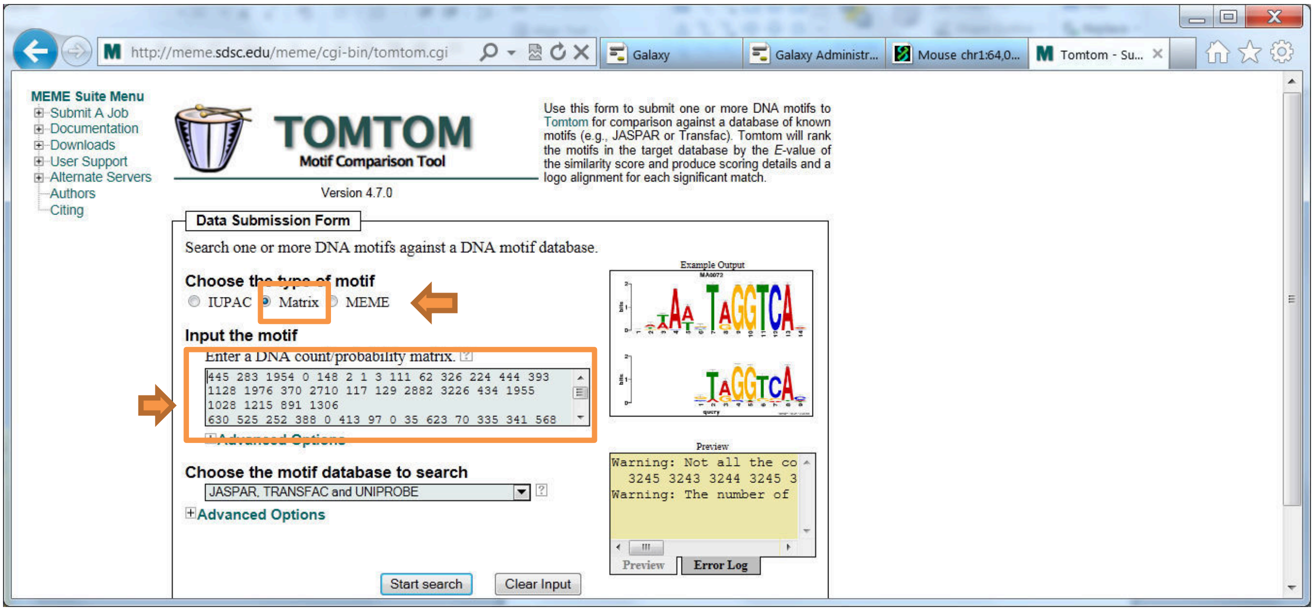
- Here we used the “mask” mode for motif finding (p. 25)

Which known transcription factors correspond to identified motifs?



```
Motif 1
A| 998 1589 1474 1863 598 2681 84 257 2732 3006 140 673 457 10
C| 578 347 357 59 606 40 0 99 417 0 401 235 527 633
G| 1314 892 1213 1022 2009 445 3189 2918 125 112 2735 375 2010
T| 386 447 231 331 61 109 2 1 1 157 0 1992 280 441
Motif 2
A| 461 870 158 20 1 0 1362 1 559 669 225 198 421
C| 1439 700 1053 2010 1401 3050 0 0 1306 1021 1167 1599 1266
G| 587 43 712 6 1 0 0 3046 621 364 477 214 83
T| 562 1435 1124 1012 1644 0 1687 0 563 994 1178 1037 1279
Motif 3
A| 1068 675 1280 529 2566 3 1042 0 1871 318
C| 585 1111 738 1967 2 954 1520 2567 10 485
G| 513 549 343 60 0 1608 1 0 8 1390
T| 401 231 206 9 0 2 4 0 677 375
```

- Run TOMTOM: http://meme.sdsc.edu/meme/cgi-bin/tomtom.cgi
- TOMTOM can be found by **google** using "**TOMTOM motif**"
- Select the type of motif "mask"
- Copy-paste your motif from Galaxy (remove motif name and nucleotide names in the beginning of each line)



Use this form to submit one or more DNA motifs to Tomtom for comparison against a database of known motifs (e.g., JASPAR or Transfac). Tomtom will rank the motifs in the target database by the E-value of the similarity score and produce scoring details and a logo alignment for each significant match.

Data Submission Form

Search one or more DNA motifs against a DNA motif database.

Choose the type of motif

○ IUPAC ● Matrix ○ MEME

Input the motif

Enter a DNA count/probability matrix.

```
445 283 1954 0 148 2 1 3 111 62 326 224 444 393
1128 1976 370 2710 117 129 2882 3226 434 1955
1028 1215 891 1306
630 525 252 388 0 413 97 0 35 623 70 335 341 568
```

Advanced Options

Choose the motif database to search

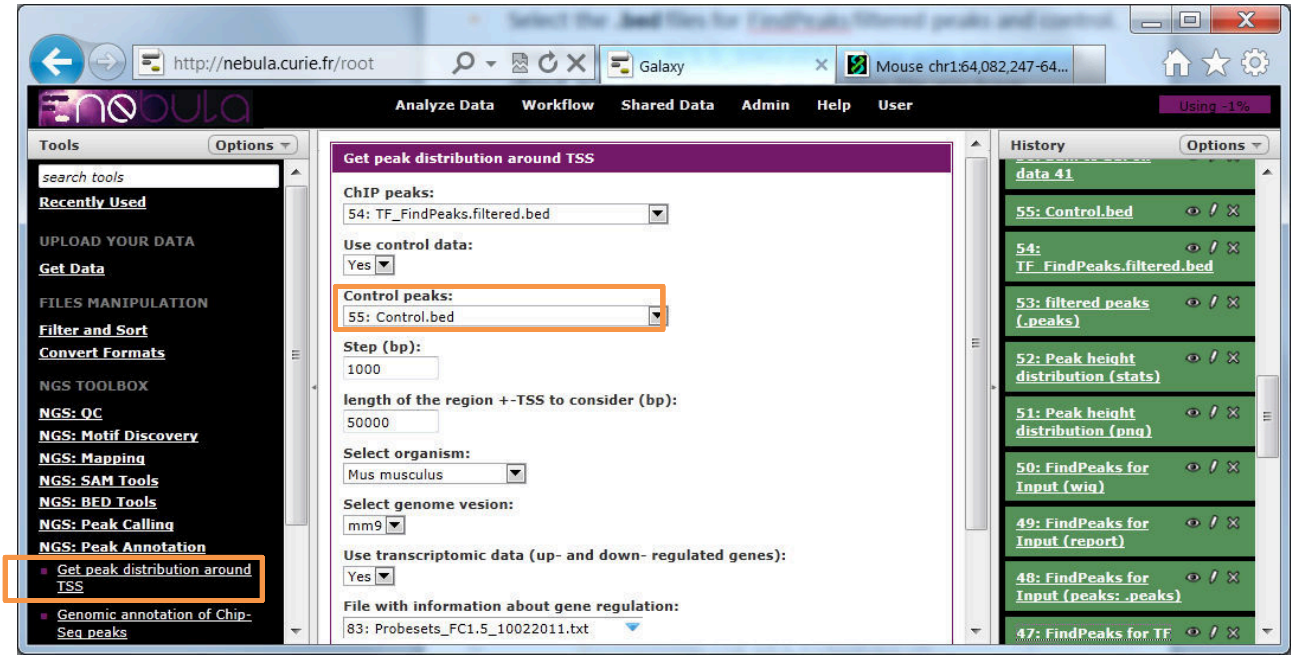
JASPAR, TRANSFAC and UNIPROBE

Advanced Options

```
Warning: Not all the co
  3245 3243 3244 3245 3
Warning: The number of
```

Calculate distribution of peak locations around gene TSS

- Select the **.bed** files for FindPeaks filtered peaks and control.
- Use ProbeSets_FC1.5_10022011.txt file with information about activated/repressed genes (this file you have uploaded to your history in the very beginning)



http://nebula.curie.fr/root

Galaxy

Mouse chr1:64,082,247-64...

Analyze Data Workflow Shared Data Admin Help User

Using -1%

Tools Options

search tools

Recently Used

UPLOAD YOUR DATA

Get Data

FILES MANIPULATION

Filter and Sort

Convert Formats

NGS TOOLBOX

NGS: QC

NGS: Motif Discovery

NGS: Mapping

NGS: SAM Tools

NGS: BED Tools

NGS: Peak Calling

NGS: Peak Annotation

- Get peak distribution around TSS
- Genomic annotation of Chip-Seq peaks

Get peak distribution around TSS

ChIP peaks: 54: TF_FindPeaks.filtered.bed

Use control data: Yes

Control peaks: 55: Control.bed

Step (bp): 1000

length of the region +-TSS to consider (bp): 50000

Select organism: Mus musculus

Select genome vesion: mm9

Use transcriptomic data (up- and down- regulated genes): Yes

File with information about gene regulation: 83: Probesets_FC1.5_10022011.txt

History Options

data 41

55: Control.bed

54: TF_FindPeaks.filtered.bed

53: filtered peaks (.peaks)

52: Peak height distribution (stats)

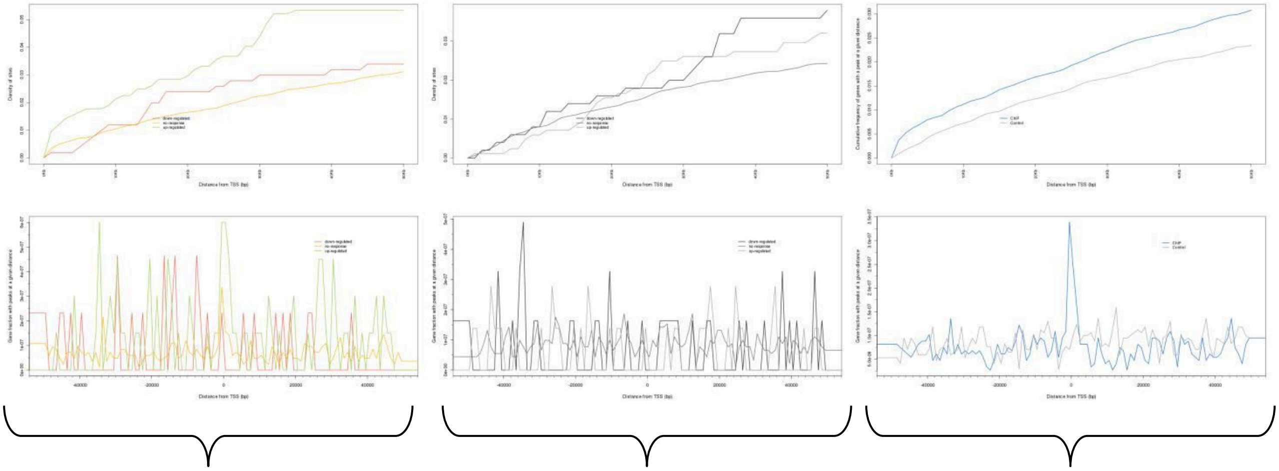
51: Peak height distribution (png)

50: FindPeaks for Input (wig)

49: FindPeaks for Input (report)

48: FindPeaks for Input (peaks: .peaks)

47: FindPeaks for TF



ChIP

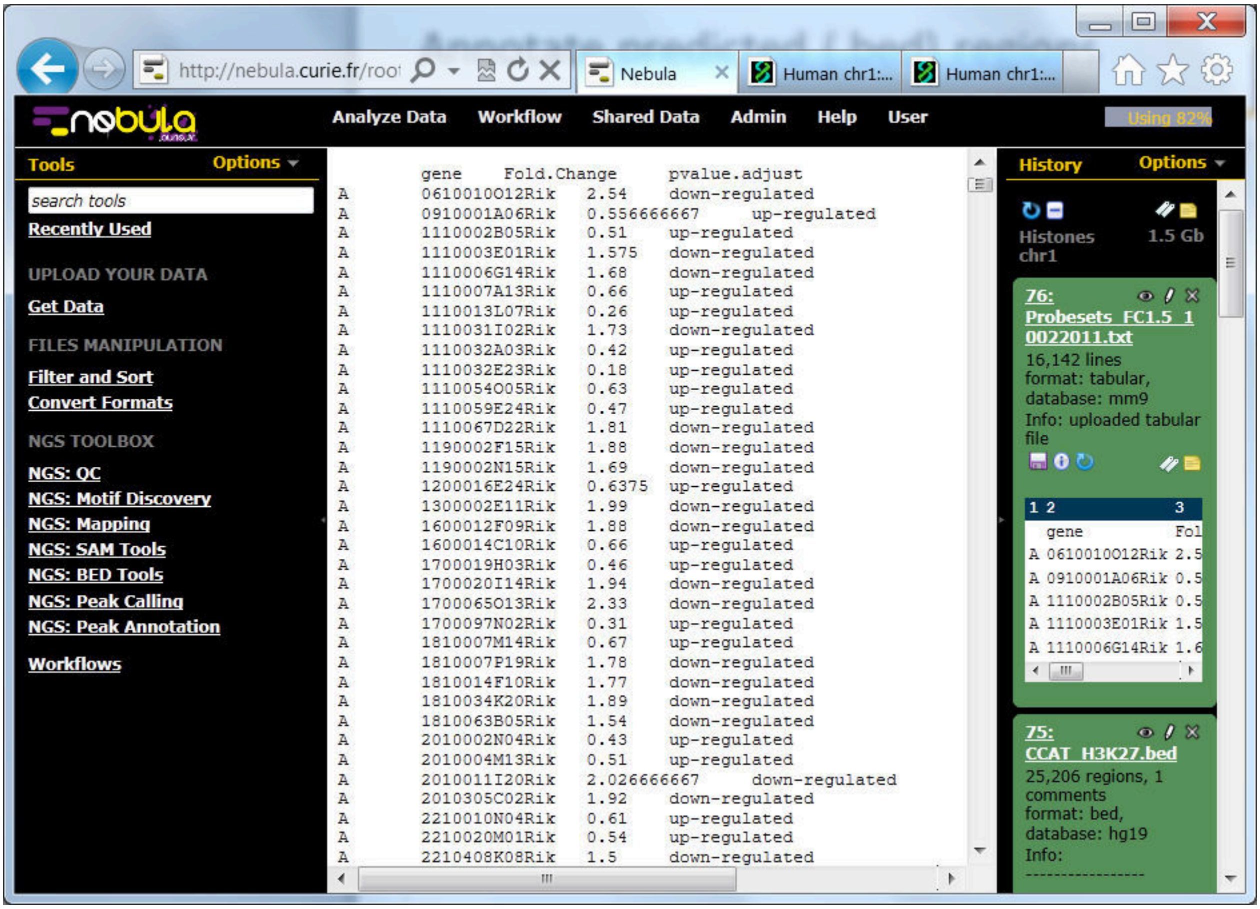
Control

ChIP vs Control

Format of gene expression/modulation file

- Tab-delimited:

1. Some field
2. Gene symbol
3. Some value (expression or fold change)
4. Gene feature: e.g., expressed, regulated, etc.



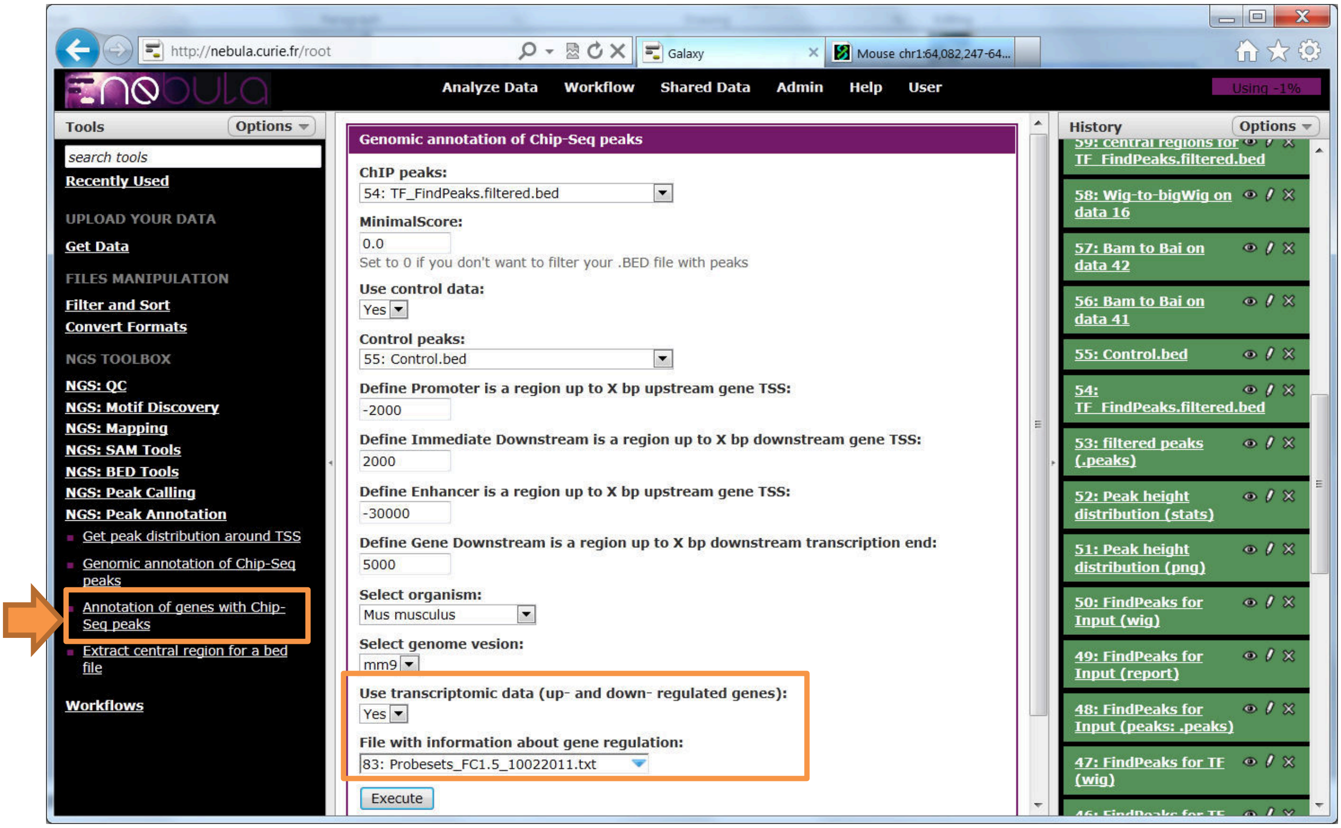
```
        gene         Fold.Change     pvalue.adjust
A       0610010O12Rik   2.54    down-regulated
A       0910001A06Rik   0.556666667     up-regulated
A       1110002B05Rik   0.51    up-regulated
A       1110003E01Rik   1.575   down-regulated
A       1110006G14Rik   1.68    down-regulated
A       1110007A13Rik   0.66    up-regulated
A       1110013L07Rik   0.26    up-regulated
A       1110031I02Rik   1.73    down-regulated
A       1110032A03Rik   0.42    up-regulated
A       1110032E23Rik   0.18    up-regulated
A       1110054O05Rik   0.63    up-regulated
A       1110059E24Rik   0.47    up-regulated
A       1110067D22Rik   1.81    down-regulated
A       1190002F15Rik   1.88    down-regulated
A       1190002N15Rik   1.69    down-regulated
A       1200016E24Rik   0.6375  up-regulated
A       1300002E11Rik   1.99    down-regulated
A       1600012F09Rik   1.88    down-regulated
A       1600014C10Rik   0.66    up-regulated
A       1700019H03Rik   0.46    up-regulated
A       1700020I14Rik   1.94    down-regulated
A       1700065O13Rik   2.33    down-regulated
A       1700097N02Rik   0.31    up-regulated
A       1810007M14Rik   0.67    up-regulated
A       1810007P19Rik   1.78    down-regulated
A       1810014F10Rik   1.77    down-regulated
A       1810034K20Rik   1.89    down-regulated
A       1810063B05Rik   1.54    down-regulated
A       2010002N04Rik   0.43    up-regulated
A       2010004M13Rik   0.51    up-regulated
A       2010011I20Rik   2.026666667     down-regulated
A       2010305C02Rik   1.92    down-regulated
A       2210010N04Rik   0.61    up-regulated
A       2210020M01Rik   0.54    up-regulated
A       2210408K08Rik   1.5     down-regulated
```

76: Probesets_FC1.5_10022011.txt
16,142 lines
format: tabular, database: mm9
Info: uploaded tabular file

```
1 2           3
  gene        Fol
A 0610010O12Rik 2.5
A 0910001A06Rik 0.5
A 1110002B05Rik 0.5
A 1110003E01Rik 1.5
A 1110006G14Rik 1.6
```

75: CCAT_H3K27.bed
25,206 regions, 1 comments
format: bed, database: hg19
Info:

Annotate peaks with genomic features


http://nebula.curie.fr/root
Galaxy
Mouse chr1:64,082,247-64...
nebula
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using -1%
Tools
Options
search tools
Recently Used
UPLOAD YOUR DATA
Get Data
FILES MANIPULATION
Filter and Sort
Convert Formats
NGS TOOLBOX
NGS: QC
NGS: Motif Discovery
NGS: Mapping
NGS: SAM Tools
NGS: BED Tools
NGS: Peak Calling
NGS: Peak Annotation
Get peak distribution around TSS
Genomic annotation of Chip-Seq peaks
Annotation of genes with Chip-Seq peaks
Extract central region for a bed file
Workflows
History
Options
59: central regions for TF_FindPeaks.filtered.bed
58: Wig-to-bigWig on data 16
57: Bam to Bai on data 42
56: Bam to Bai on data 41
55: Control.bed
54: TF_FindPeaks.filtered.bed
53: filtered peaks (.peaks)
52: Peak height distribution (stats)
51: Peak height distribution (png)
50: FindPeaks for Input (wig)
49: FindPeaks for Input (report)
48: FindPeaks for Input (peaks: .peaks)
47: FindPeaks for TF (wig)
46: FindPeaks for TF

Genomic annotation of Chip-Seq peaks

ChIP peaks: 54: TF_FindPeaks.filtered.bed

MinimalScore: 0.0

Set to 0 if you don't want to filter your .BED file with peaks

Use control data: Yes

Control peaks: 55: Control.bed

Define Promoter is a region up to X bp upstream gene TSS: -2000

Define Immediate Downstream is a region up to X bp downstream gene TSS: 2000

Define Enhancer is a region up to X bp upstream gene TSS: -30000

Define Gene Downstream is a region up to X bp downstream transcription end: 5000

Select organism: Mus musculus

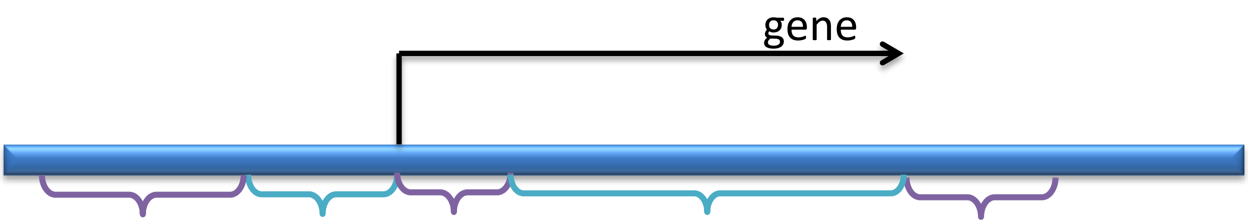
Select genome vesion: mm9

Use transcriptomic data (up- and down- regulated genes): Yes

File with information about gene regulation: 83: Probesets_FC1.5_10022011.txt

Execute

Annotate peaks with genomic features



gene

Promoter

Enhancer

Immediate downstream

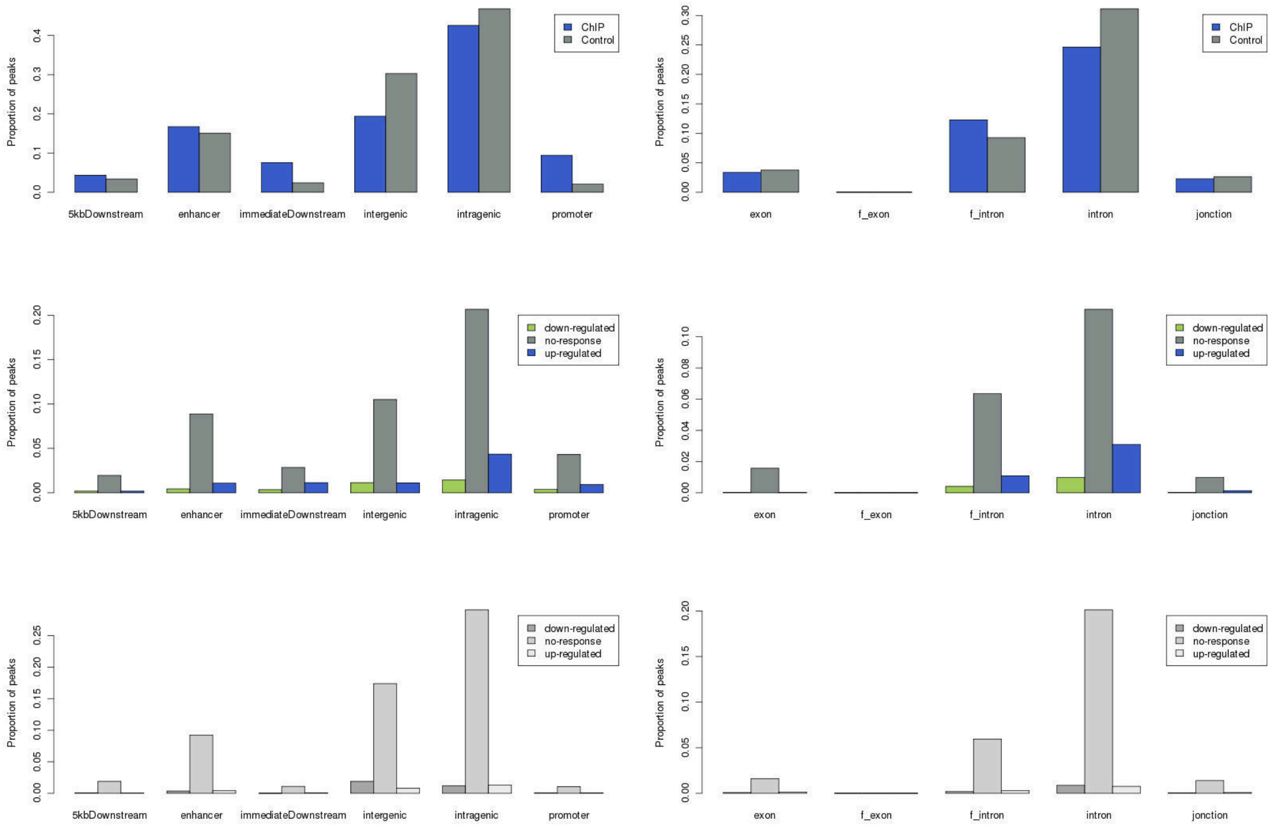
Intragenic

Gene downstream (default 5kb)

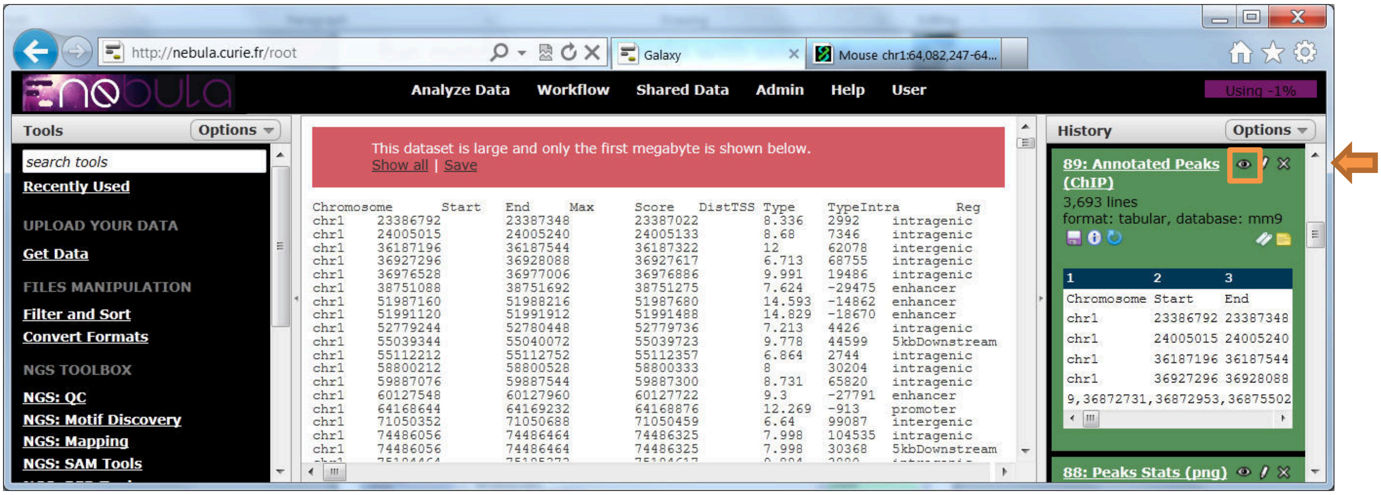
ChIP vs Control

Modulated vs NoResponse

Modulated vs NoResponse for Control



Run motif finding for peaks with selected genomic features



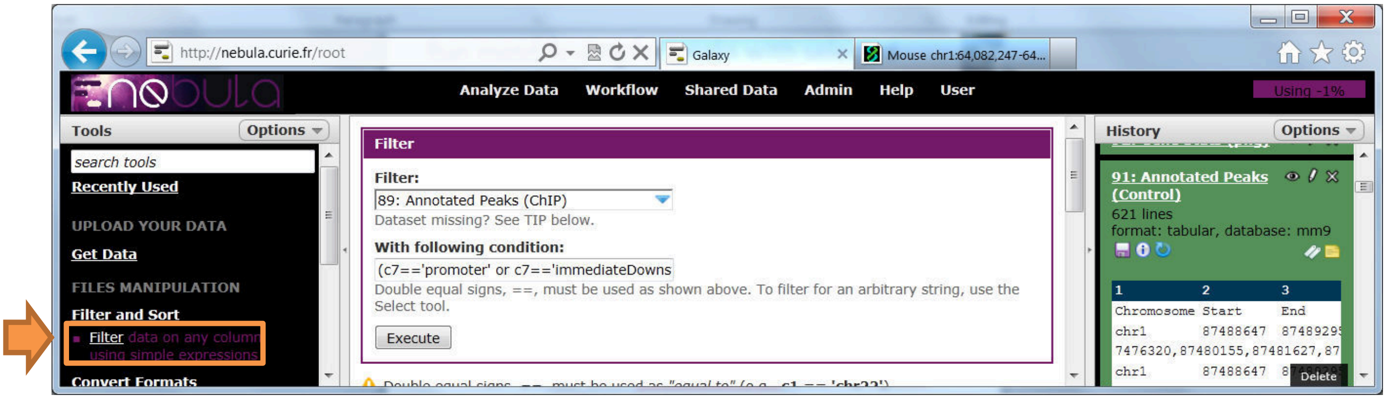
This dataset is large and only the first megabyte is shown below.
Show all | Save

```
Chromosome      Start     End      Max       Score    DistTSS Type     TypeIntra      Reg
chr1     23386792         23387348          23387022          8.336    2992     intragenic
chr1     24005015         24005240          24005133          8.68     7346     intragenic
chr1     36187196         36187544          36187322          12       62078    intergenic
chr1     36927296         36928088          36927617          6.713    68755    intragenic
chr1     36976528         36977006          36976886          9.991    19486    intragenic
chr1     38751088         38751692          38751275          7.624    -29475   enhancer
chr1     51987160         51988216          51987680          14.593   -14862   enhancer
chr1     51991120         51991912          51991488          14.829   -18670   enhancer
chr1     52779244         52780448          52779736          7.213    4426     intragenic
chr1     55039344         55040072          55039723          9.778    44599    5kbDownstream
chr1     55112212         55112752          55112357          6.864    2744     intragenic
chr1     58800212         58800528          58800333          8        30204    intragenic
chr1     59887076         59887544          59887300          8.731    65820    intragenic
chr1     60127548         60127960          60127722          9.3      -27791   enhancer
chr1     64168644         64169232          64168876          12.269   -913     promoter
chr1     71050352         71050688          71050459          6.64     99087    intergenic
chr1     74486056         74486464          74486325          7.998    104535   intragenic
chr1     74486056         74486464          74486325          7.998    30368    5kbDownstream
```

89: Annotated Peaks (ChIP)
3,693 lines
format: tabular, database: mm9

1	2	3
Chromosome	Start	End
chr1	23386792	23387348
chr1	24005015	24005240
chr1	36187196	36187544
chr1	36927296	36928088
9,36872731,36872953,36875502		

- **Use:** (c7==‘promoter’ or c7==‘immediateDownstream’) and c9==‘up-regulated’



Filter

Filter:
89: Annotated Peaks (ChIP)
Dataset missing? See TIP below.

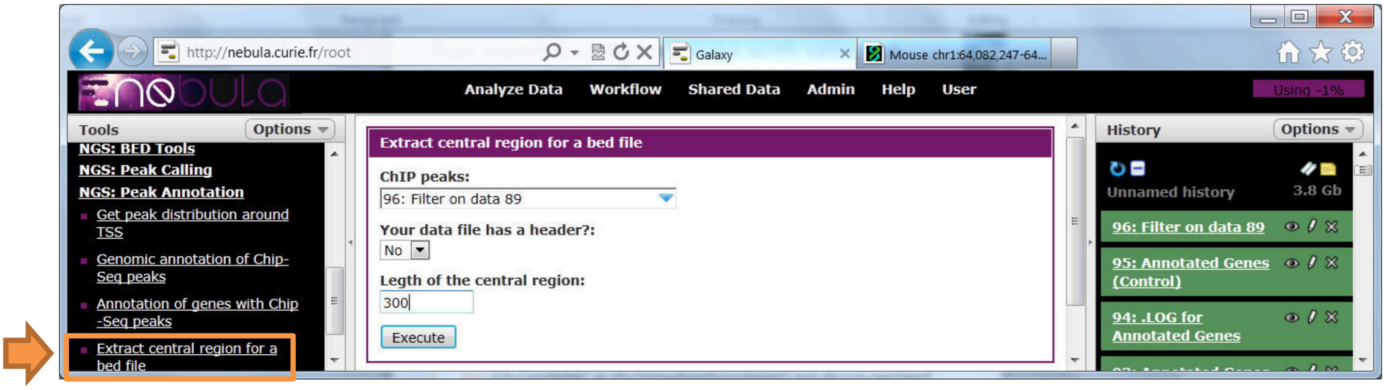
With following condition:
(c7=='promoter' or c7=='immediateDowns
Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Execute

91: Annotated Peaks (Control)
621 lines
format: tabular, database: mm9

1	2	3
Chromosome	Start	End
chr1	87488647	8748929
7476320,87480155,87481627,87		
chr1	87488647	8

- Select central regions of peaks:


http://nebula.curie.fr/root
Galaxy
Mouse chr1:64,082,247-64...
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using -1%
Tools
Options
NGS: BED Tools
NGS: Peak Calling
NGS: Peak Annotation
Get peak distribution around TSS
Genomic annotation of Chip-Seq peaks
Annotation of genes with Chip-Seq peaks
Extract central region for a bed file
History
Options
Unnamed history
3.8 Gb
96: Filter on data 89
95: Annotated Genes (Control)
94: .LOG for Annotated Genes

Extract central region for a bed file

ChIP peaks:
96: Filter on data 89

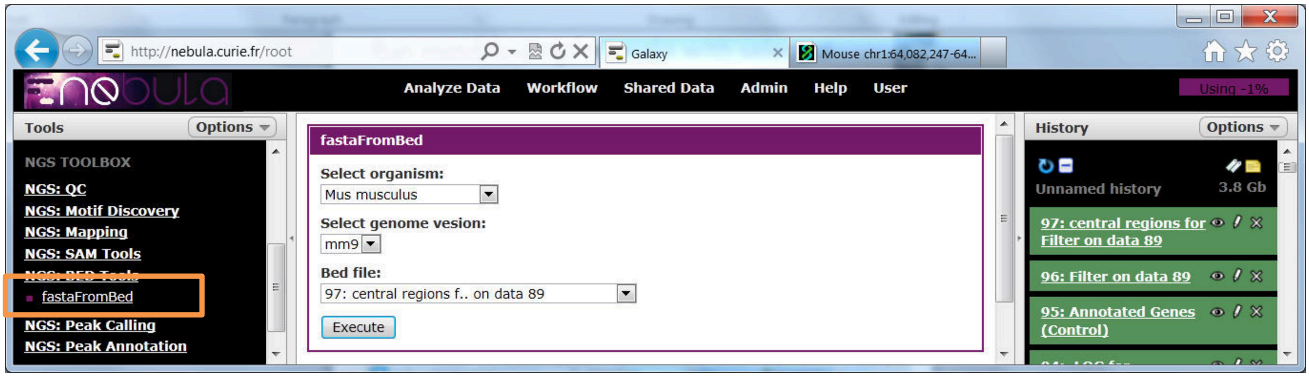
Your data file has a header?:
No

Legth of the central region:
300

Execute

Run motif finding for peaks with selected genomic features

- Get .fasta



http://nebula.curie.fr/root | Galaxy | Mouse chr1:64,082,247-64...

Analyze Data | Workflow | Shared Data | Admin | Help | User | Using -1%

Tools | Options

NGS TOOLBOX
- NGS: QC
- NGS: Motif Discovery
- NGS: Mapping
- NGS: SAM Tools
- NGS: BED Tools
  - fastaFromBed
- NGS: Peak Calling
- NGS: Peak Annotation

fastaFromBed

Select organism: Mus musculus

Select genome vesion: mm9

Bed file: 97: central regions f.. on data 89

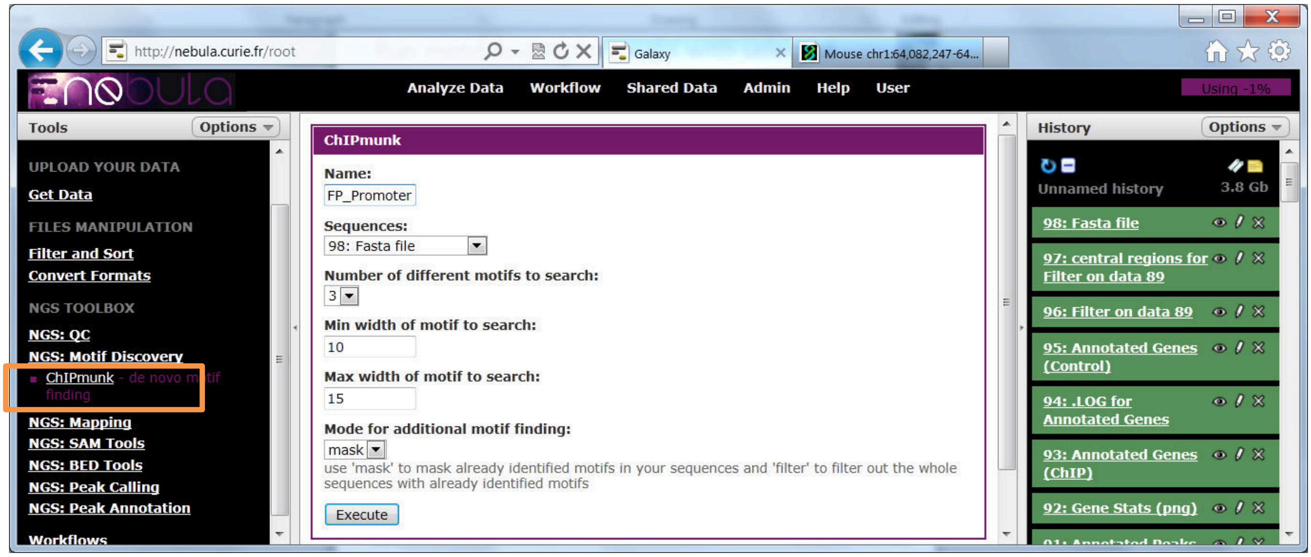
Execute

History | Options

Unnamed history 3.8 Gb

- 97: central regions for Filter on data 89
- 96: Filter on data 89
- 95: Annotated Genes (Control)
- 94: .LOG for...

- Run Motif Finding



http://nebula.curie.fr/root | Galaxy | Mouse chr1:64,082,247-64...

Analyze Data | Workflow | Shared Data | Admin | Help | User | Using -1%

Tools | Options

UPLOAD YOUR DATA
- Get Data

FILES MANIPULATION
- Filter and Sort
- Convert Formats

NGS TOOLBOX
- NGS: QC
- NGS: Motif Discovery
  - ChIPmunk - de novo motif finding
- NGS: Mapping
- NGS: SAM Tools
- NGS: BED Tools
- NGS: Peak Calling
- NGS: Peak Annotation

Workflows

ChIPmunk

Name: FP_Promoter

Sequences: 98: Fasta file

Number of different motifs to search: 3

Min width of motif to search: 10

Max width of motif to search: 15

Mode for additional motif finding: mask

use 'mask' to mask already identified motifs in your sequences and 'filter' to filter out the whole sequences with already identified motifs

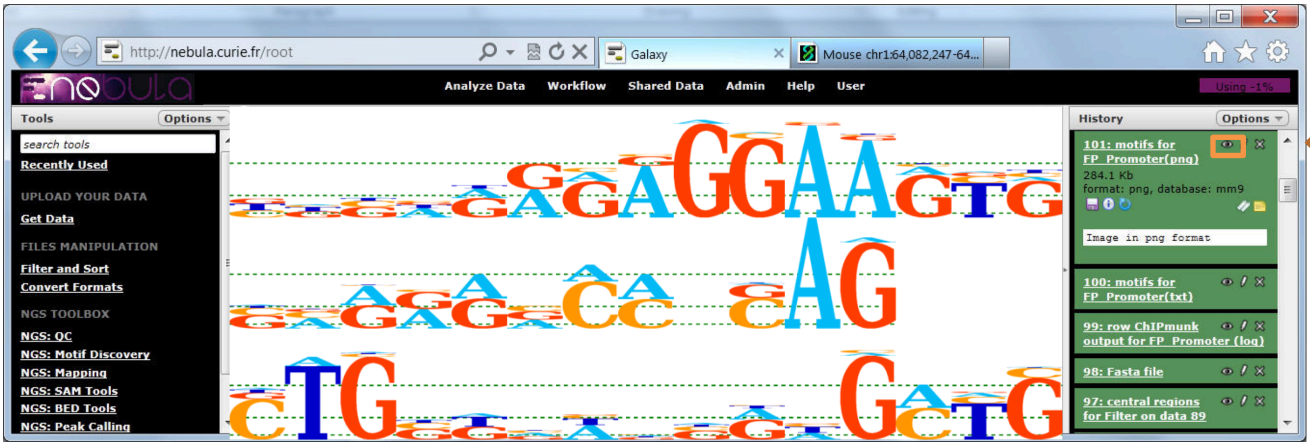
Execute

History | Options

Unnamed history 3.8 Gb

- 98: Fasta file
- 97: central regions for Filter on data 89
- 96: Filter on data 89
- 95: Annotated Genes (Control)
- 94: .LOG for Annotated Genes
- 93: Annotated Genes (ChIP)
- 92: Gene Stats (png)
- 91: Annotated Peaks

- Visualize motifs



http://nebula.curie.fr/root | Galaxy | Mouse chr1:64,082,247-64...

Analyze Data | Workflow | Shared Data | Admin | Help | User | Using -1%

Tools | Options

search tools

- Recently Used

UPLOAD YOUR DATA
- Get Data

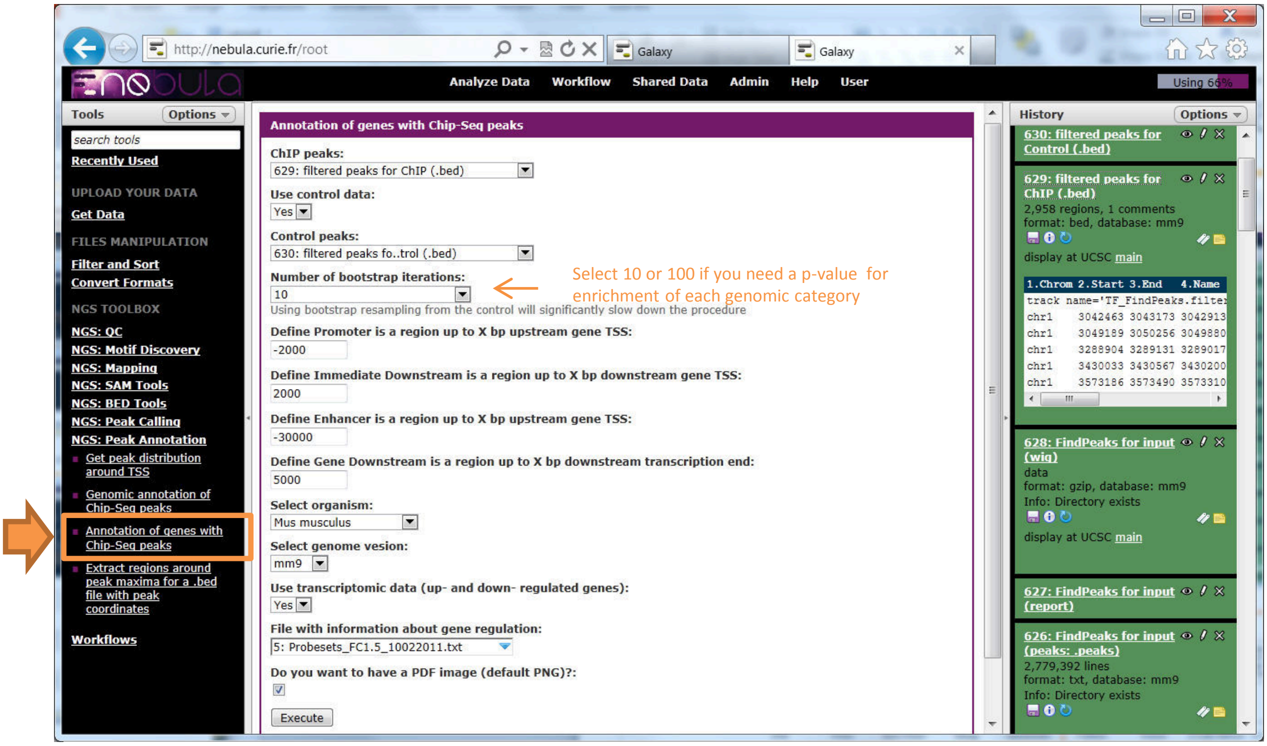
FILES MANIPULATION
- Filter and Sort
- Convert Formats

NGS TOOLBOX
- NGS: QC
- NGS: Motif Discovery
- NGS: Mapping
- NGS: SAM Tools
- NGS: BED Tools
- NGS: Peak Calling

History | Options

- 101: motifs for FP_Promoter(png)
  284.1 Kb
  format: png, database: mm9
  Image in png format
- 100: motifs for FP_Promoter(txt)
- 99: row ChIPmunk output for FP_Promoter (log)
- 98: Fasta file
- 97: central regions for Filter on data 89

Annotate genes with peak information



Annotation of genes with Chip-Seq peaks

ChIP peaks: 629: filtered peaks for ChIP (.bed)

Use control data: Yes

Control peaks: 630: filtered peaks fo..trol (.bed)

Number of bootstrap iterations: 10

Using bootstrap resampling from the control will significantly slow down the procedure

Select 10 or 100 if you need a p-value for enrichment of each genomic category

Define Promoter is a region up to X bp upstream gene TSS: -2000

Define Immediate Downstream is a region up to X bp downstream gene TSS: 2000

Define Enhancer is a region up to X bp upstream gene TSS: -30000

Define Gene Downstream is a region up to X bp downstream transcription end: 5000

Select organism: Mus musculus

Select genome vesion: mm9

Use transcriptomic data (up- and down- regulated genes): Yes

File with information about gene regulation: 5: Probesets_FC1.5_10022011.txt

Do you want to have a PDF image (default PNG)?: ☑

Execute

630: filtered peaks for Control (.bed)

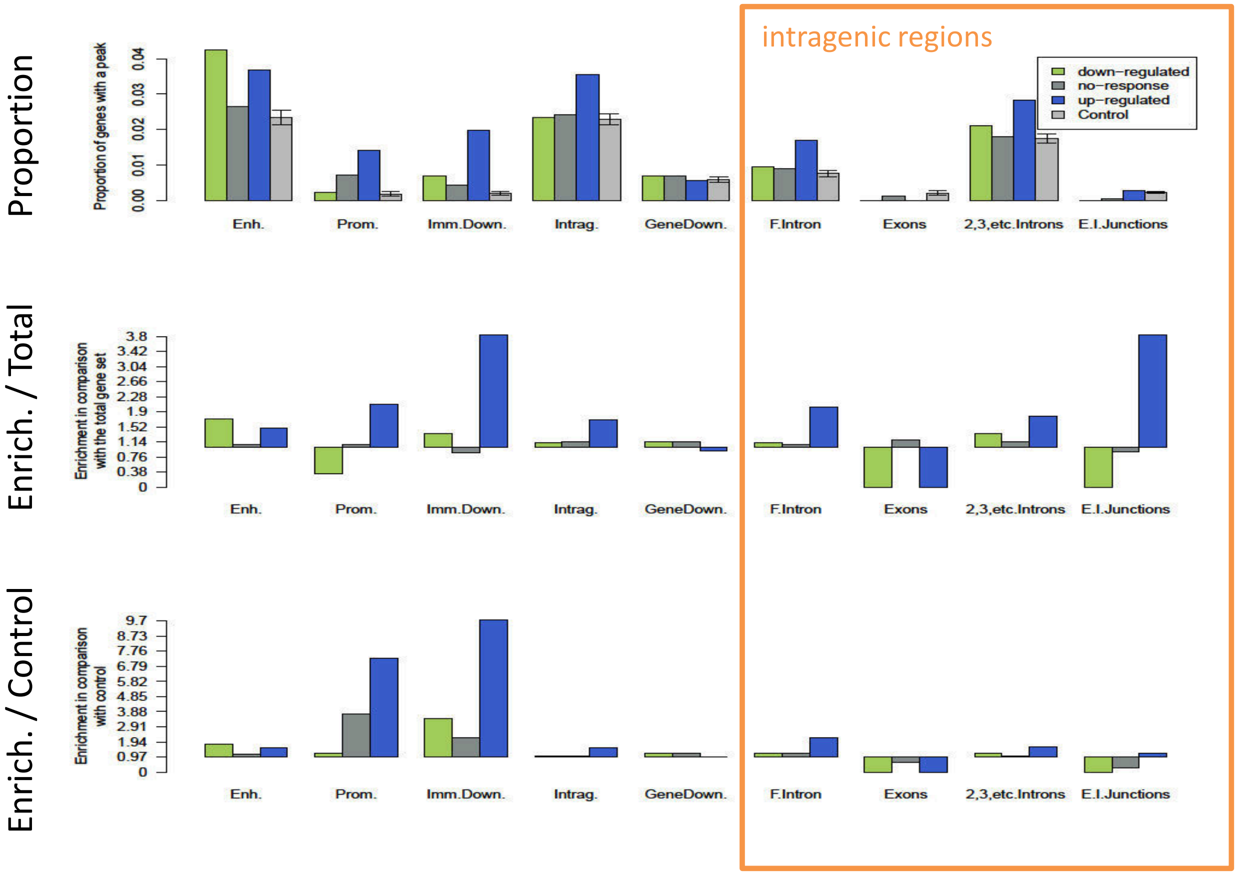
629: filtered peaks for ChIP (.bed)
2,958 regions, 1 comments
format: bed, database: mm9
display at UCSC main

```
1.Chrom 2.Start 3.End   4.Name
track name='TF_FindPeaks.filter
chr1    3042463 3043173 3042913
chr1    3049189 3050256 3049880
chr1    3288904 3289131 3289017
chr1    3430033 3430567 3430200
chr1    3573186 3573490 3573310
```

628: FindPeaks for input (wig)
data
format: gzip, database: mm9
Info: Directory exists
display at UCSC main

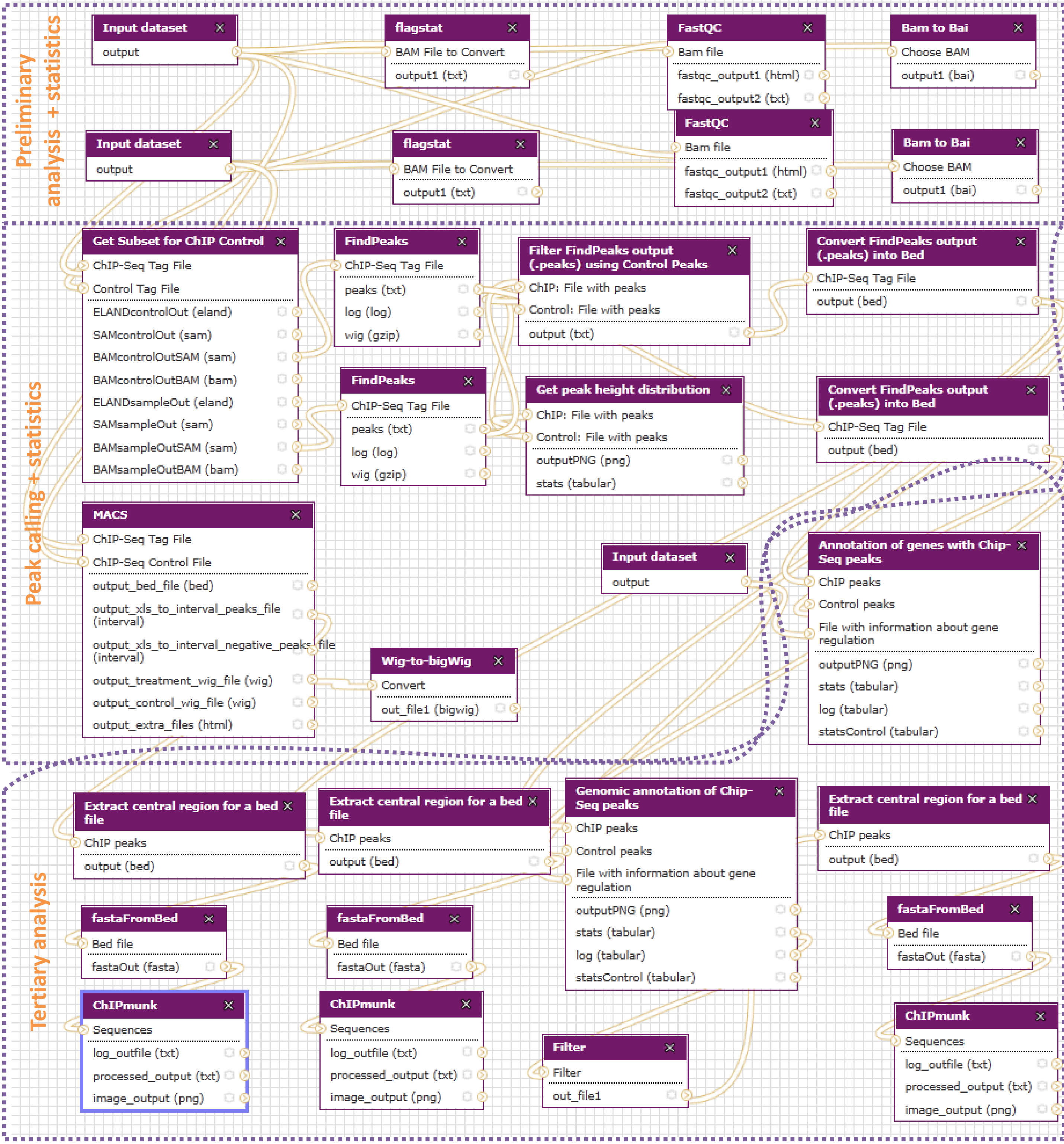
627: FindPeaks for input (report)

626: FindPeaks for input (peaks: .peaks)
2,779,392 lines
format: txt, database: mm9
Info: Directory exists



intragenic regions

Our workflow



Analysis of histone data

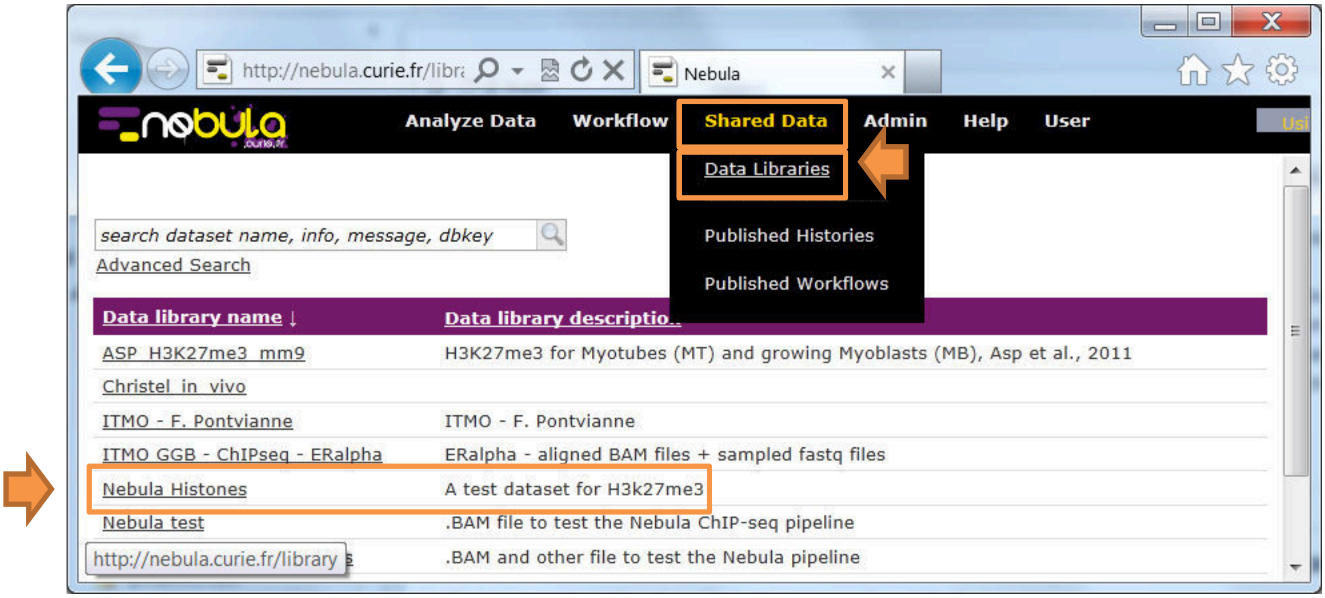
- Peak calling
- Peak visualization
- Peak statistics
- Gene annotation

Test Data: H3K27me3 (H.Ashoor, 2013) for a bladder cancer cell line. Data provided for chromosome 1 only.

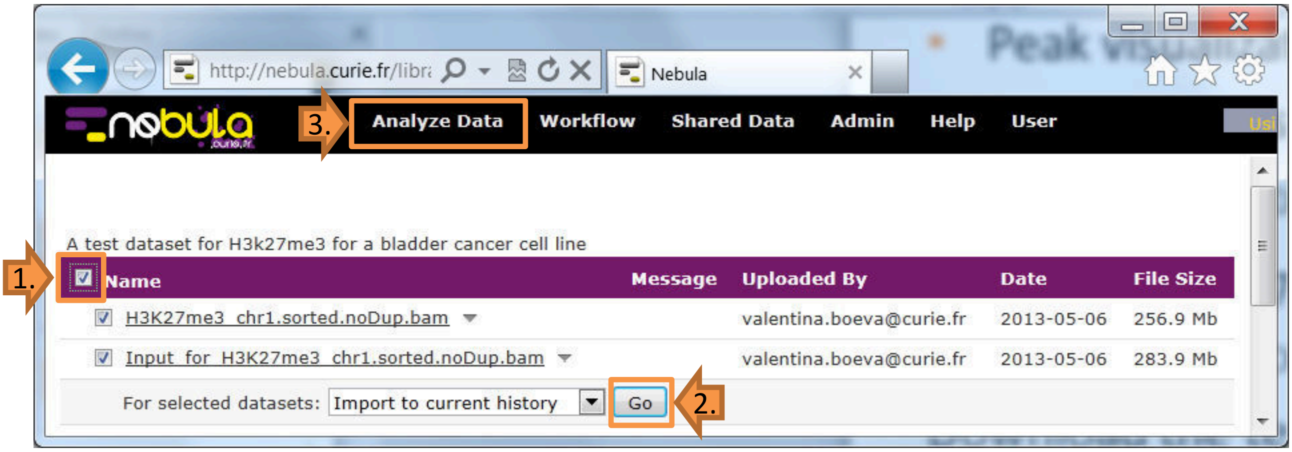
Upload data to the history

Test Data: H3K27me3 (H.Ashoor, 2013) for a bladder cancer cell line. ONLY chr1.

Download the test dataset:



Data library name ↓	Data library descriptio...
ASP_H3K27me3_mm9	H3K27me3 for Myotubes (MT) and growing Myoblasts (MB), Asp et al., 2011
Christel_in_vivo	
ITMO - F. Pontvianne	ITMO - F. Pontvianne
ITMO GGB - ChIPseq - ERalpha	ERalpha - aligned BAM files + sampled fastq files
Nebula Histones	A test dataset for H3k27me3
Nebula test	.BAM file to test the Nebula ChIP-seq pipeline
	.BAM and other file to test the Nebula pipeline



A test dataset for H3k27me3 for a bladder cancer cell line

☑ Name	Message	Uploaded By	Date	File Size
☑ H3K27me3_chr1.sorted.noDup.bam		valentina.boeva@curie.fr	2013-05-06	256.9 Mb
☑ Input_for_H3K27me3_chr1.sorted.noDup.bam		valentina.boeva@curie.fr	2013-05-06	283.9 Mb

For selected datasets: Import to current history [Go]

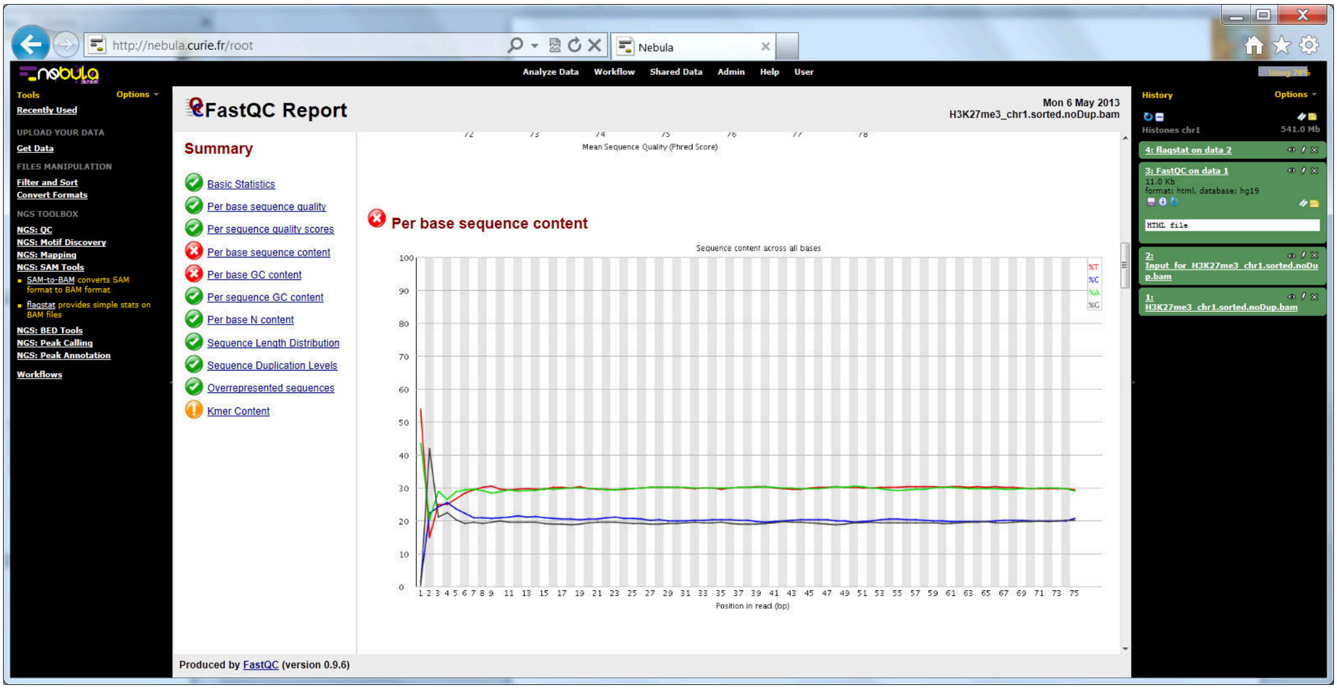
Check read number and sequencing quality pp. 7-9

Visualize .BAM files in UCSC pp. 10-13

Read statistics

- Use Samtools “flagstat” or “FastQC” to get statistics about reads

Check read number and sequencing quality	pp. 7-9
Visualize .BAM files in UCSC	pp. 10-13



FastQC Report

Mon 6 May 2013
H3K27me3_chr1.sorted.noDup.bam

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ! Kmer Content

✗ Per base sequence content

Produced by FastQC (version 0.9.6)

History

Histones chr1 541.0 Mb

4: flagstat on data 2

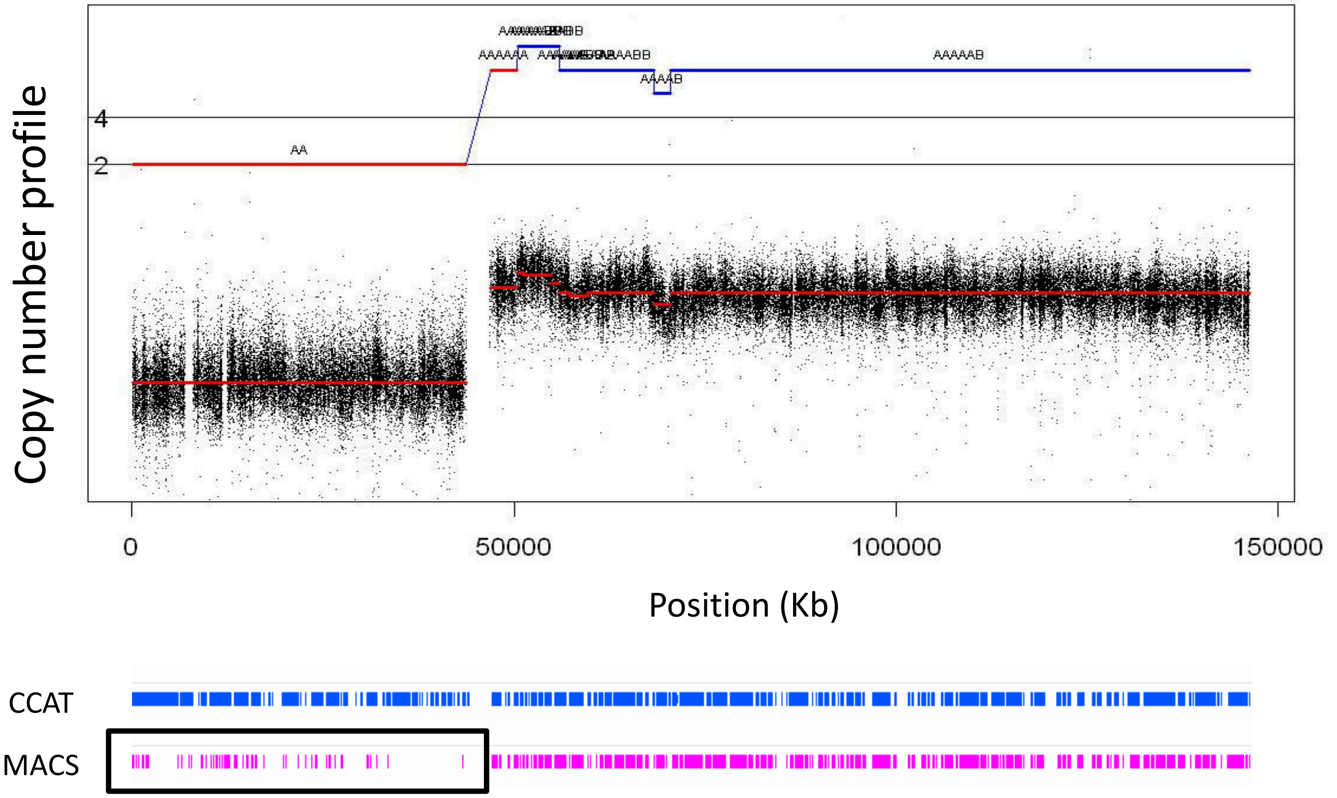
3: FastQC on data 1
11.0 Kb
format: html, database: hg19
HTML file

2: Input_for_H3K27me3_chr1.sorted.noDup.bam

1: H3K27me3_chr1.sorted.noDup.bam

Peak calling for histone data

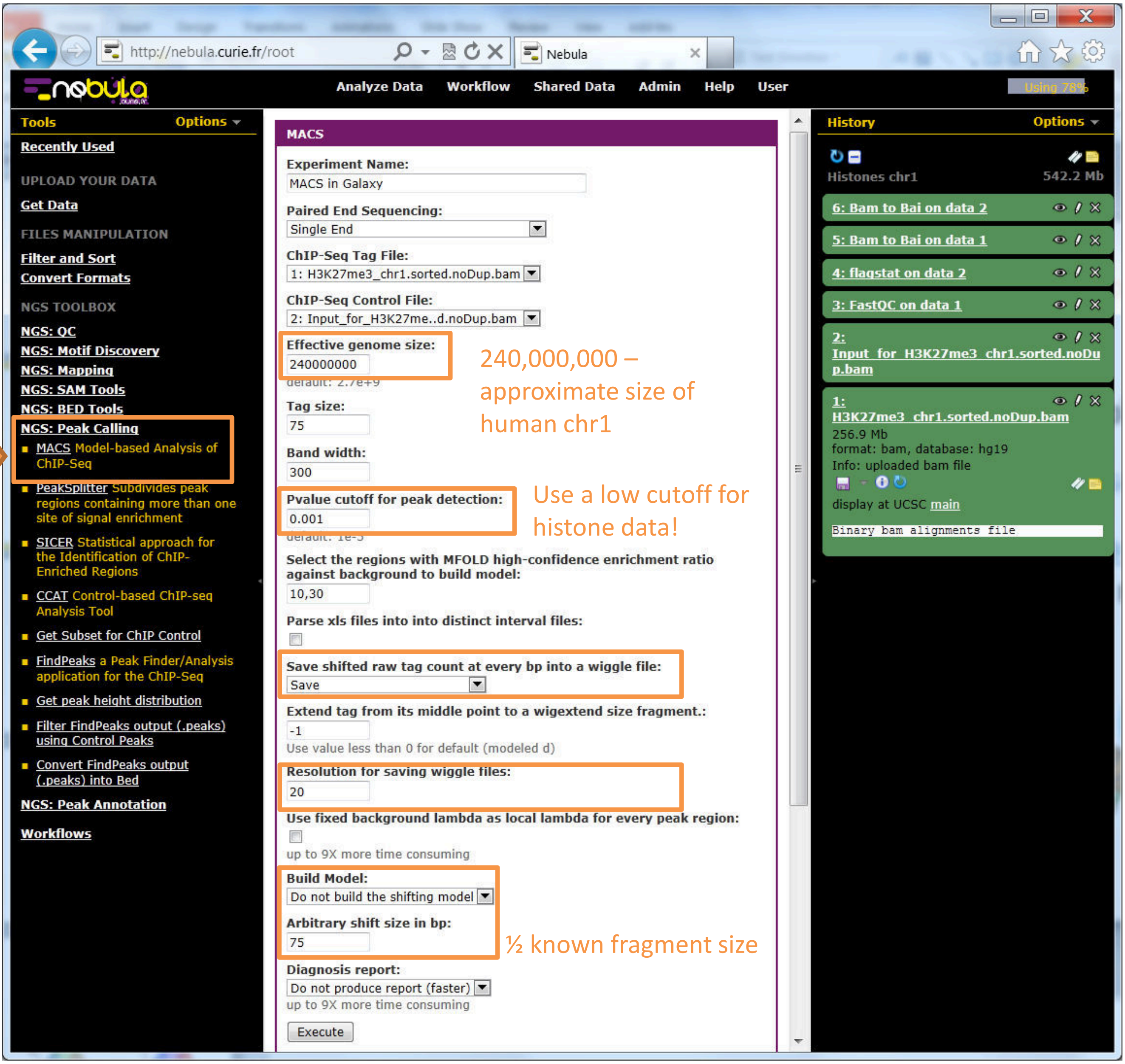
- To call peaks, use MACS or CCAT
- For cancer datasets, use CCAT since it does not show copy number bias:



low density of sites in predictions of MACS in the region of low copy number

- MACS generates .wig files

Peak calling with MACS for histone data



http://nebula.curie.fr/root
Nebula
nebula
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 78%
Tools
Options
Recently Used
UPLOAD YOUR DATA
Get Data
FILES MANIPULATION
Filter and Sort
Convert Formats
NGS TOOLBOX
NGS: QC
NGS: Motif Discovery
NGS: Mapping
NGS: SAM Tools
NGS: BED Tools
NGS: Peak Calling
MACS Model-based Analysis of ChIP-Seq
PeakSplitter Subdivides peak regions containing more than one site of signal enrichment
SICER Statistical approach for the Identification of ChIP-Enriched Regions
CCAT Control-based ChIP-seq Analysis Tool
Get Subset for ChIP Control
FindPeaks a Peak Finder/Analysis application for the ChIP-Seq
Get peak height distribution
Filter FindPeaks output (.peaks) using Control Peaks
Convert FindPeaks output (.peaks) into Bed
NGS: Peak Annotation
Workflows
History
Options
Histones chr1
542.2 Mb
6: Bam to Bai on data 2
5: Bam to Bai on data 1
4: flagstat on data 2
3: FastQC on data 1
2: Input_for_H3K27me3_chr1.sorted.noDup.bam
1: H3K27me3_chr1.sorted.noDup.bam
256.9 Mb
format: bam, database: hg19
Info: uploaded bam file
display at UCSC main
Binary bam alignments file

MACS

Experiment Name: MACS in Galaxy

Paired End Sequencing: Single End

ChIP-Seq Tag File: 1: H3K27me3_chr1.sorted.noDup.bam

ChIP-Seq Control File: 2: Input_for_H3K27me..d.noDup.bam

Effective genome size: 240000000
default: 2.7e+9

240,000,000 – approximate size of human chr1

Tag size: 75

Band width: 300

Pvalue cutoff for peak detection: 0.001
default: 1e-5

Use a low cutoff for histone data!

Select the regions with MFOLD high-confidence enrichment ratio against background to build model: 10,30

Parse xls files into into distinct interval files: □

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: 20

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

Build Model: Do not build the shifting model

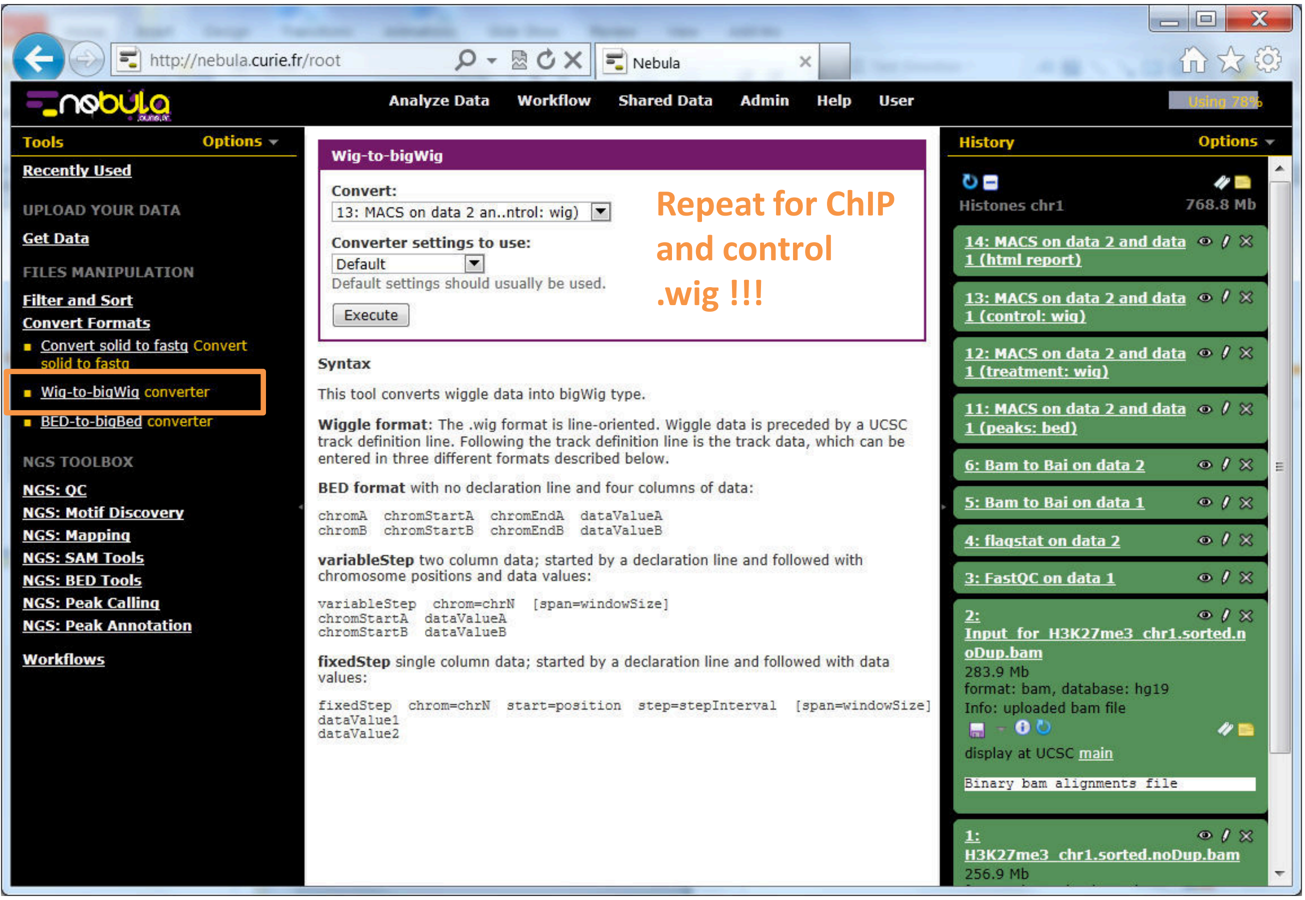
Arbitrary shift size in bp: 75

½ known fragment size

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

Execute

Visualize the .WIG



Wig-to-bigWig

Convert: 13: MACS on data 2 an..ntrol: wig)

Converter settings to use: Default

Default settings should usually be used.

Repeat for ChIP and control .wig !!!

Syntax

This tool converts wiggle data into bigWig type.

Wiggle format: The .wig format is line-oriented. Wiggle data is preceded by a UCSC track definition line. Following the track definition line is the track data, which can be entered in three different formats described below.

BED format with no declaration line and four columns of data:

```
chromA  chromStartA  chromEndA  dataValueA
chromB  chromStartB  chromEndB  dataValueB
```

variableStep two column data; started by a declaration line and followed with chromosome positions and data values:

```
variableStep  chrom=chrN  [span=windowSize]
chromStartA  dataValueA
chromStartB  dataValueB
```

fixedStep single column data; started by a declaration line and followed with data values:

```
fixedStep  chrom=chrN  start=position  step=stepInterval  [span=windowSize]
dataValue1
dataValue2
```

Histones chr1 768.8 Mb

14: MACS on data 2 and data 1 (html report)

13: MACS on data 2 and data 1 (control: wig)

12: MACS on data 2 and data 1 (treatment: wig)

11: MACS on data 2 and data 1 (peaks: bed)

6: Bam to Bai on data 2

5: Bam to Bai on data 1

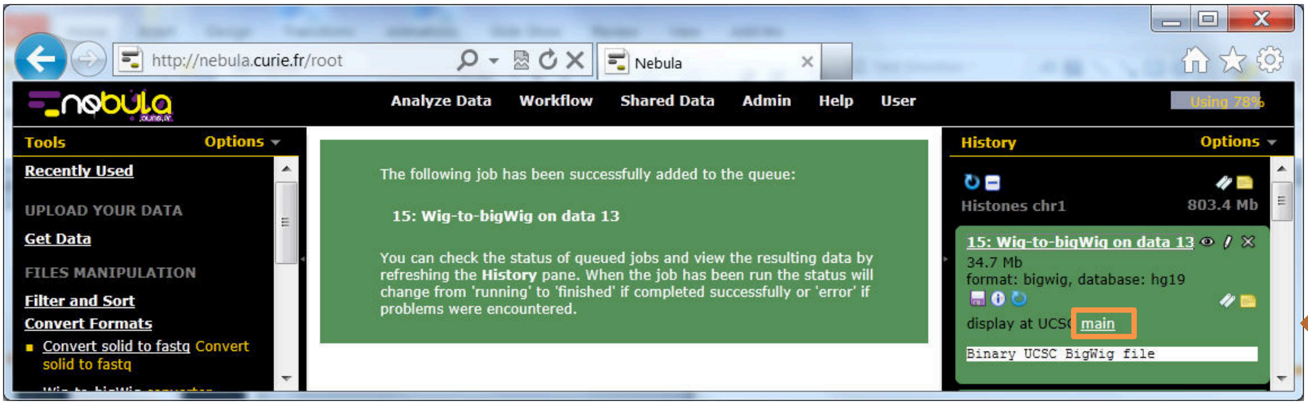
4: flagstat on data 2

3: FastQC on data 1

2: Input_for_H3K27me3_chr1.sorted.noDup.bam
283.9 Mb
format: bam, database: hg19
Info: uploaded bam file
display at UCSC main

```
Binary bam alignments file
```

1: H3K27me3_chr1.sorted.noDup.bam
256.9 Mb



The following job has been successfully added to the queue:

15: Wig-to-bigWig on data 13

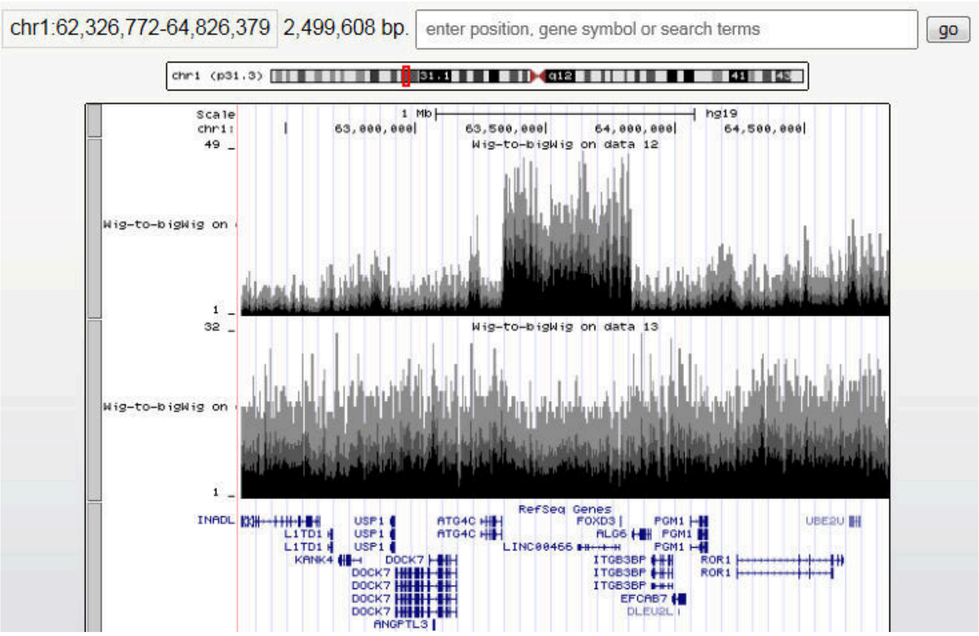
You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Histones chr1 803.4 Mb

15: Wig-to-bigWig on data 13
34.7 Mb
format: bigwig, database: hg19
display at UCSC main

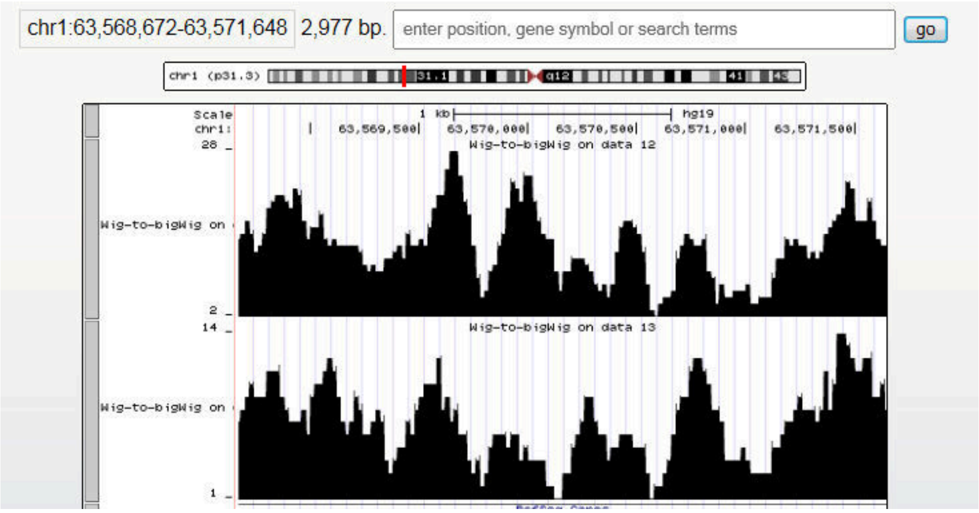
```
Binary UCSC BigWig file
```

Visualize the .WIG



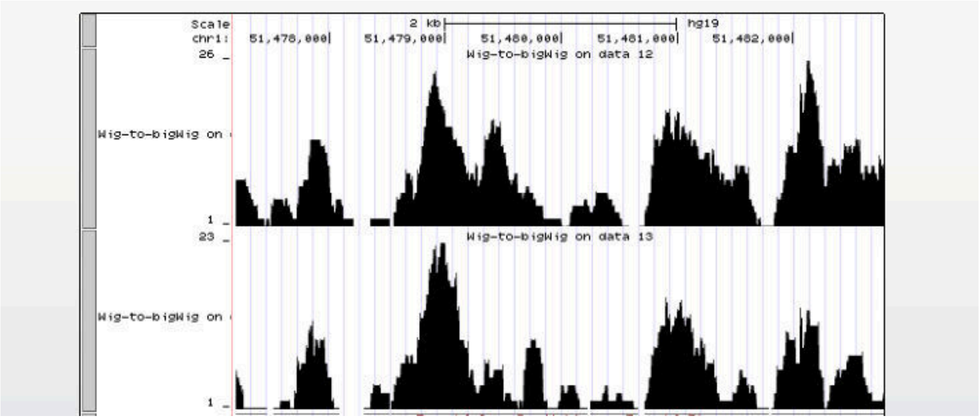
chr1:62,326,772-64,826,379 2,499,608 bp. enter position, gene symbol or search terms go

Unzoom



chr1:63,568,672-63,571,648 2,977 bp. enter position, gene symbol or search terms go

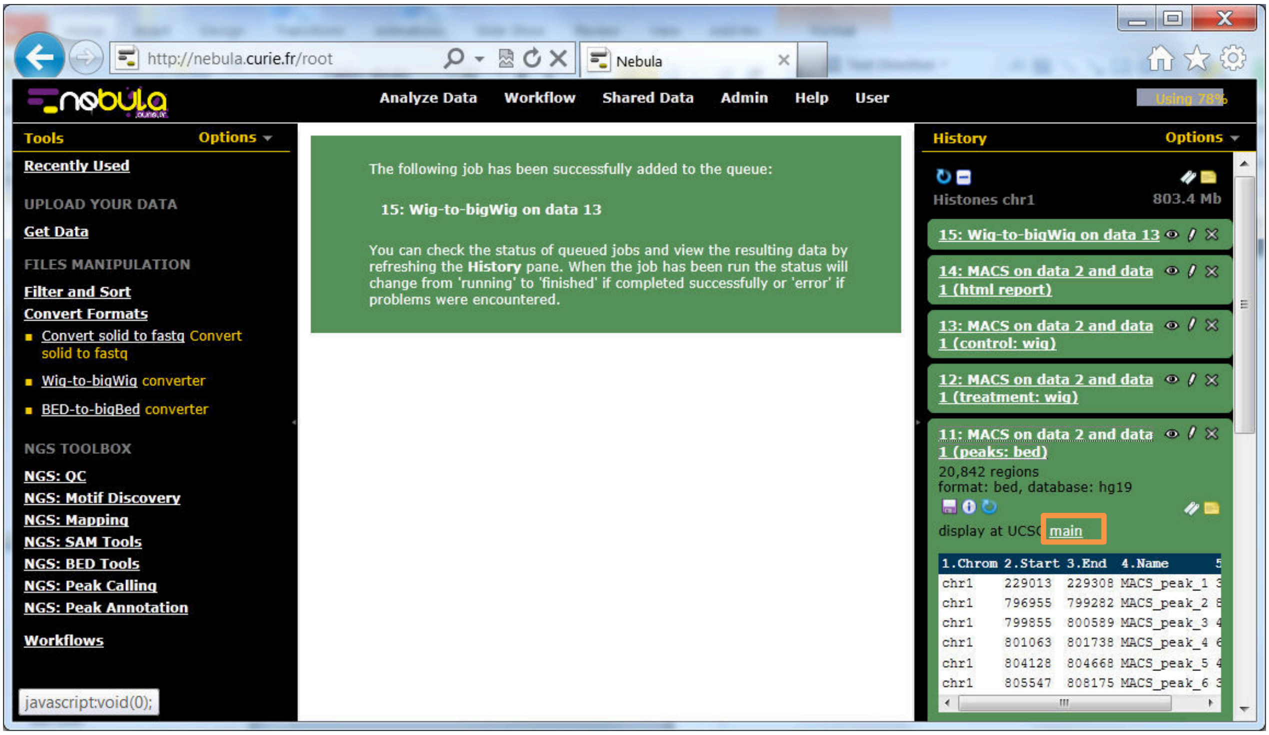
Big zoom 1



Big zoom 1

Visualize MACS peak as well as .WIG profiles

- Visualize the .BED file too:



Tools Options

Recently Used

UPLOAD YOUR DATA

Get Data

FILES MANIPULATION

Filter and Sort

Convert Formats

- Convert solid to fastq Convert solid to fastq
- Wig-to-bigWig converter
- BED-to-bigBed converter

NGS TOOLBOX

NGS: QC

NGS: Motif Discovery

NGS: Mapping

NGS: SAM Tools

NGS: BED Tools

NGS: Peak Calling

NGS: Peak Annotation

Workflows

javascript:void(0);

The following job has been successfully added to the queue:

15: Wig-to-bigWig on data 13

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.

History Options

Histones chr1 803.4 Mb

15: Wig-to-bigWig on data 13

14: MACS on data 2 and data 1 (html report)

13: MACS on data 2 and data 1 (control: wig)

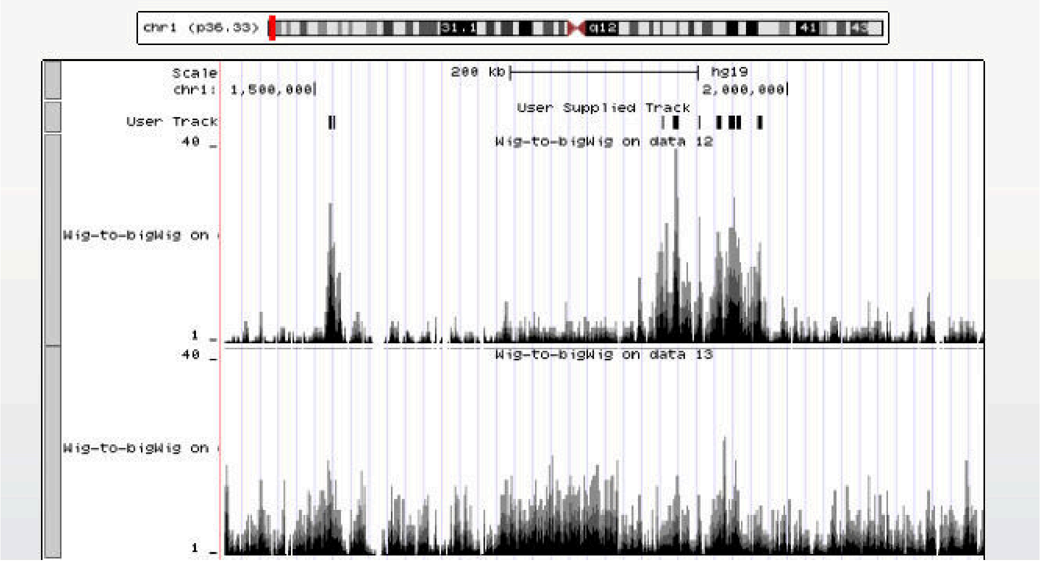
12: MACS on data 2 and data 1 (treatment: wig)

11: MACS on data 2 and data 1 (peaks: bed)

20,842 regions
format: bed, database: hg19

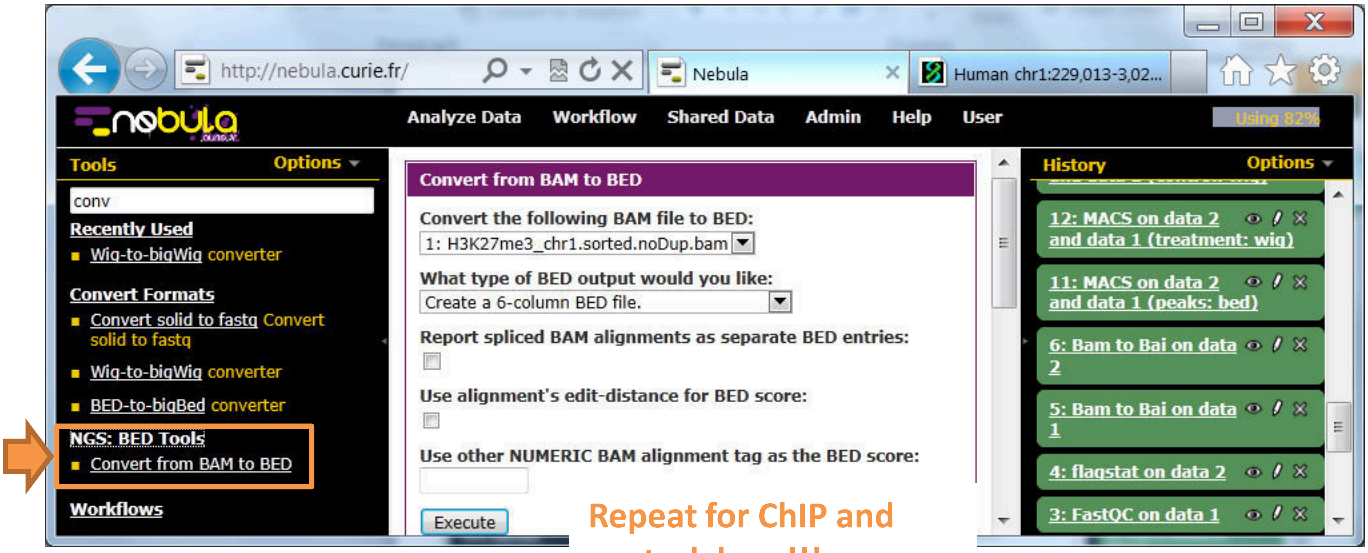
display at UCSC main

1.Chrom	2.Start	3.End	4.Name	5
chr1	229013	229308	MACS_peak_1	3
chr1	796955	799282	MACS_peak_2	8
chr1	799855	800589	MACS_peak_3	4
chr1	801063	801738	MACS_peak_4	6
chr1	804128	804668	MACS_peak_5	4
chr1	805547	808175	MACS_peak_6	3



Peak calling with CCAT for histone data

- Transform BAM to BED



Tools Options

conv

Recently Used
- Wig-to-bigWig converter

Convert Formats
- Convert solid to fastq Convert solid to fastq
- Wig-to-bigWig converter
- BED-to-bigBed converter

NGS: BED Tools
- Convert from BAM to BED

Workflows

Convert from BAM to BED

Convert the following BAM file to BED:
1: H3K27me3_chr1.sorted.noDup.bam

What type of BED output would you like:
Create a 6-column BED file.

Report spliced BAM alignments as separate BED entries:
□

Use alignment's edit-distance for BED score:
□

Use other NUMERIC BAM alignment tag as the BED score:

Execute

History Options

12: MACS on data 2 and data 1 (treatment: wig)

11: MACS on data 2 and data 1 (peaks: bed)

6: Bam to Bai on data 2

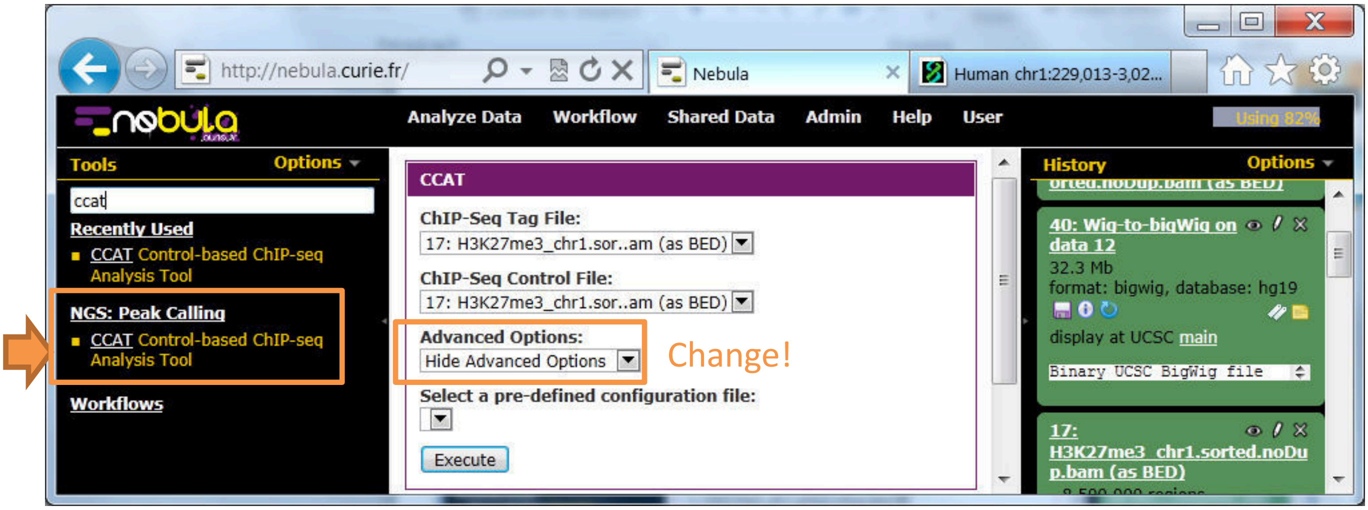
5: Bam to Bai on data 1

4: flagstat on data 2

3: FastQC on data 1

Repeat for ChIP and control .bam!!!

- Run CCAT



Tools Options

ccat

Recently Used
- CCAT Control-based ChIP-seq Analysis Tool

NGS: Peak Calling
- CCAT Control-based ChIP-seq Analysis Tool

Workflows

CCAT

ChIP-Seq Tag File:
17: H3K27me3_chr1.sor..am (as BED)

ChIP-Seq Control File:
17: H3K27me3_chr1.sor..am (as BED)

Advanced Options:
Hide Advanced Options

Select a pre-defined configuration file:

Execute

History Options

orted.noDup.bam (as BED)

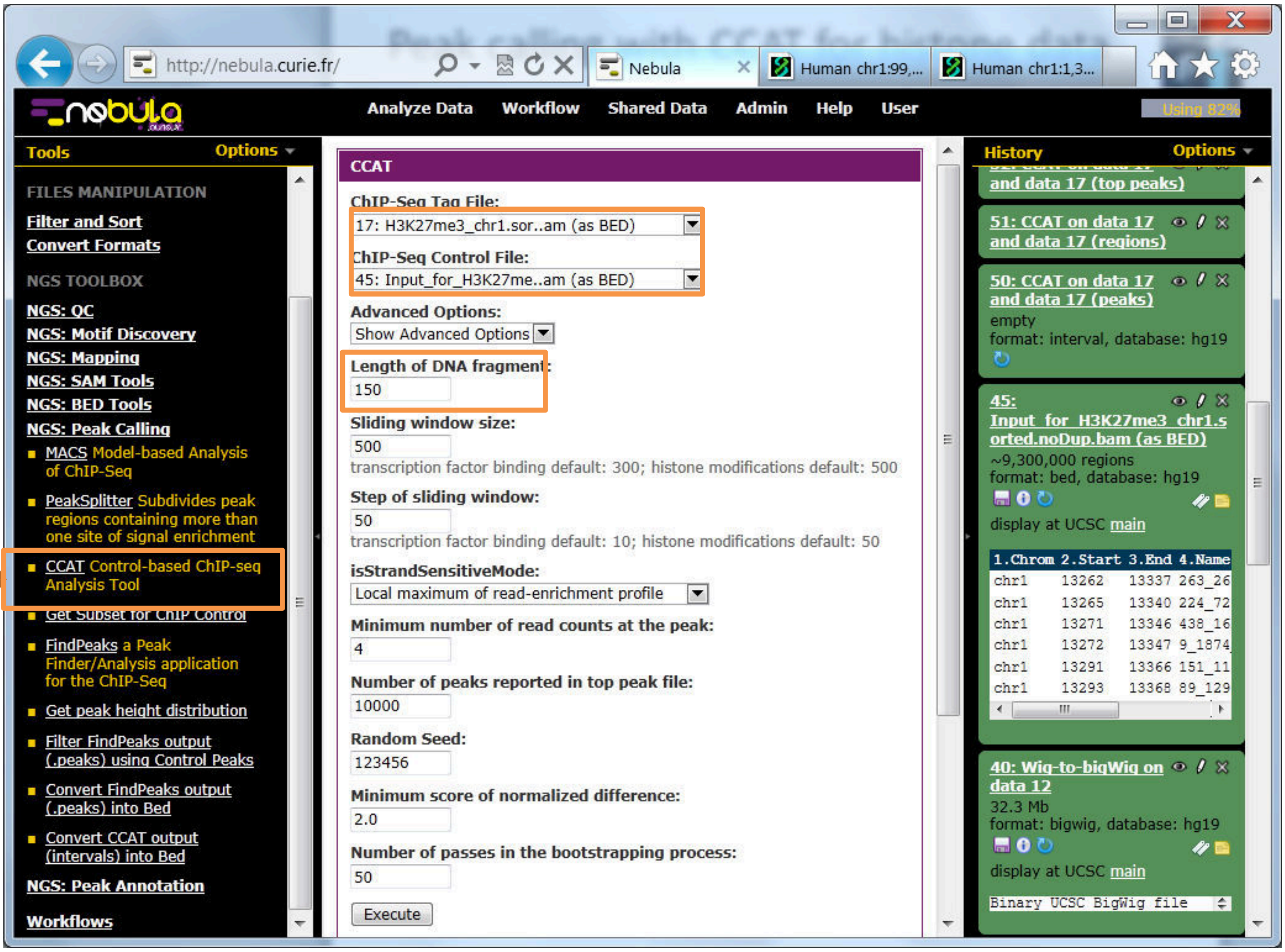
40: Wig-to-bigWig on data 12
32.3 Mb
format: bigwig, database: hg19
display at UCSC main
Binary UCSC BigWig file

17: H3K27me3_chr1.sorted.noDup.bam (as BED)

Change!

Peak calling with CCAT for histone data

- Run CCAT:


http://nebula.curie.fr/
Nebula
Human chr1:99,...
Human chr1:1,3...
nebula
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 82%
Tools
Options
FILES MANIPULATION
Filter and Sort
Convert Formats
NGS TOOLBOX
NGS: QC
NGS: Motif Discovery
NGS: Mapping
NGS: SAM Tools
NGS: BED Tools
NGS: Peak Calling
MACS Model-based Analysis of ChIP-Seq
PeakSplitter Subdivides peak regions containing more than one site of signal enrichment
CCAT Control-based ChIP-seq Analysis Tool
Get Subset for ChIP Control
FindPeaks a Peak Finder/Analysis application for the ChIP-Seq
Get peak height distribution
Filter FindPeaks output (.peaks) using Control Peaks
Convert FindPeaks output (.peaks) into Bed
Convert CCAT output (intervals) into Bed
NGS: Peak Annotation
Workflows
History
Options

CCAT

ChIP-Seq Tag File:
17: H3K27me3_chr1.sor..am (as BED)

ChIP-Seq Control File:
45: Input_for_H3K27me..am (as BED)

Advanced Options:
Show Advanced Options

Length of DNA fragment:
150

Sliding window size:
500
transcription factor binding default: 300; histone modifications default: 500

Step of sliding window:
50
transcription factor binding default: 10; histone modifications default: 50

isStrandSensitiveMode:
Local maximum of read-enrichment profile

Minimum number of read counts at the peak:
4

Number of peaks reported in top peak file:
10000

Random Seed:
123456

Minimum score of normalized difference:
2.0

Number of passes in the bootstrapping process:
50

Execute

and data 17 (top peaks)

51: CCAT on data 17 and data 17 (regions)

50: CCAT on data 17 and data 17 (peaks)
empty
format: interval, database: hg19

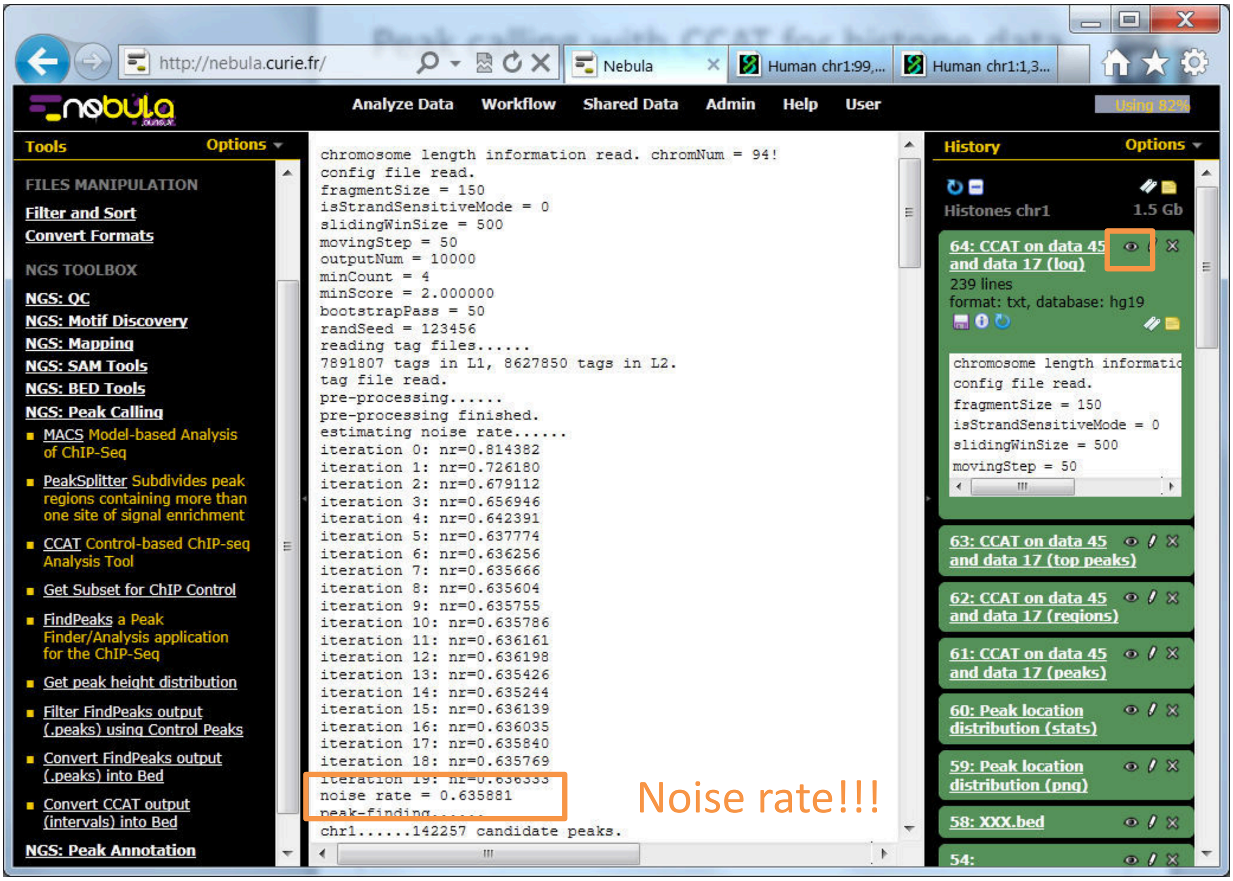
45: Input_for_H3K27me3_chr1.sorted.noDup.bam (as BED)
~9,300,000 regions
format: bed, database: hg19
display at UCSC main

1.Chrom	2.Start	3.End	4.Name
chr1	13262	13337	263_26
chr1	13265	13340	224_72
chr1	13271	13346	438_16
chr1	13272	13347	9_1874
chr1	13291	13366	151_11
chr1	13293	13368	89_129

40: Wig-to-bigWig on data 12
32.3 Mb
format: bigwig, database: hg19
display at UCSC main
Binary UCSC BigWig file

Peak calling with CCAT for histone data

- CCAT provides important information about noise in the ChIP-seq data

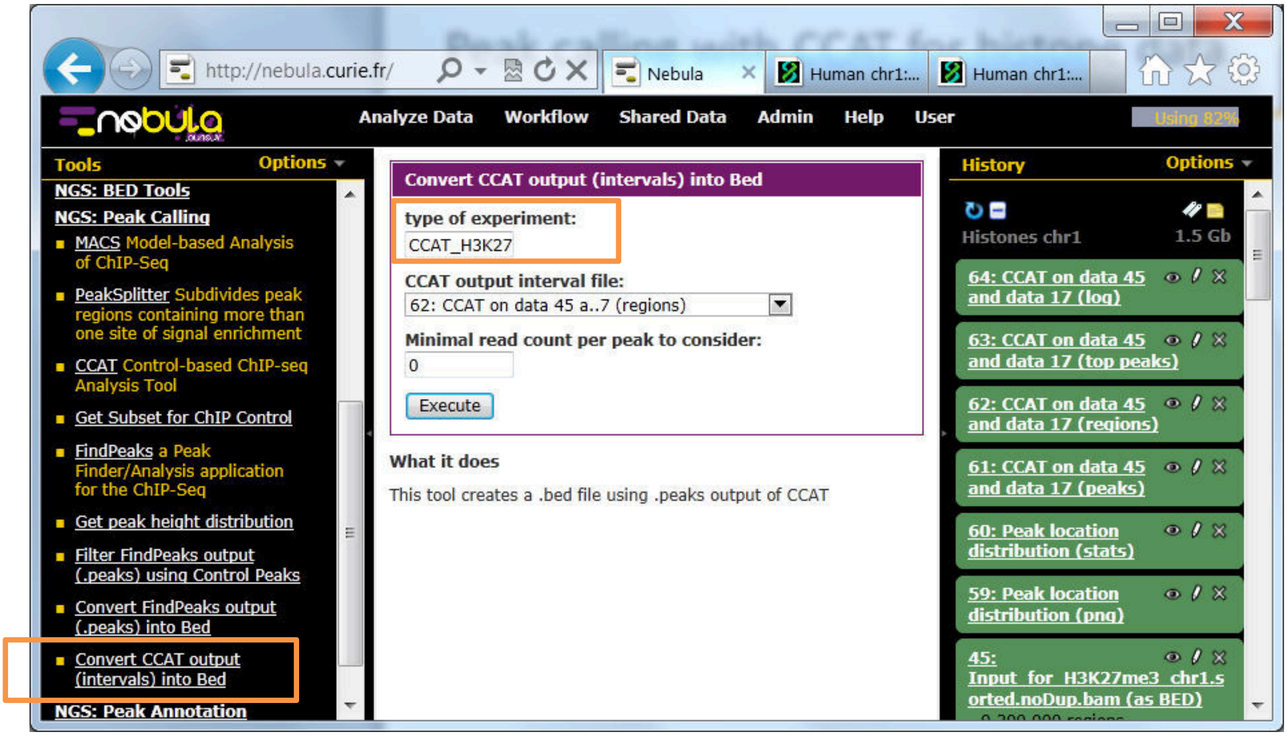

http://nebula.curie.fr/
Nebula
Human chr1:99,...
Human chr1:1,3...
nebula
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 82%
Tools
Options
FILES MANIPULATION
Filter and Sort
Convert Formats
NGS TOOLBOX
NGS: QC
NGS: Motif Discovery
NGS: Mapping
NGS: SAM Tools
NGS: BED Tools
NGS: Peak Calling
MACS Model-based Analysis of ChIP-Seq
PeakSplitter Subdivides peak regions containing more than one site of signal enrichment
CCAT Control-based ChIP-seq Analysis Tool
Get Subset for ChIP Control
FindPeaks a Peak Finder/Analysis application for the ChIP-Seq
Get peak height distribution
Filter FindPeaks output (.peaks) using Control Peaks
Convert FindPeaks output (.peaks) into Bed
Convert CCAT output (intervals) into Bed
NGS: Peak Annotation
History
Options
Histones chr1
1.5 Gb
64: CCAT on data 45 and data 17 (log)
239 lines
format: txt, database: hg19
chromosome length informatic
config file read.
fragmentSize = 150
isStrandSensitiveMode = 0
slidingWinSize = 500
movingStep = 50
63: CCAT on data 45 and data 17 (top peaks)
62: CCAT on data 45 and data 17 (regions)
61: CCAT on data 45 and data 17 (peaks)
60: Peak location distribution (stats)
59: Peak location distribution (png)
58: XXX.bed
54:

```
chromosome length information read. chromNum = 94!
config file read.
fragmentSize = 150
isStrandSensitiveMode = 0
slidingWinSize = 500
movingStep = 50
outputNum = 10000
minCount = 4
minScore = 2.000000
bootstrapPass = 50
randSeed = 123456
reading tag files......
7891807 tags in L1, 8627850 tags in L2.
tag file read.
pre-processing......
pre-processing finished.
estimating noise rate......
iteration 0: nr=0.814382
iteration 1: nr=0.726180
iteration 2: nr=0.679112
iteration 3: nr=0.656946
iteration 4: nr=0.642391
iteration 5: nr=0.637774
iteration 6: nr=0.636256
iteration 7: nr=0.635666
iteration 8: nr=0.635604
iteration 9: nr=0.635755
iteration 10: nr=0.635786
iteration 11: nr=0.636161
iteration 12: nr=0.636198
iteration 13: nr=0.635426
iteration 14: nr=0.635244
iteration 15: nr=0.636139
iteration 16: nr=0.636035
iteration 17: nr=0.635840
iteration 18: nr=0.635769
iteration 19: nr=0.636333
noise rate = 0.635881
peak-finding......
chr1......142257 candidate peaks.
```

Noise rate!!!

Peak calling with CCAT for histone data

- Transform strange output format of CCAT (chr center start end reads_chip reads_control FC FDR) into .Bed:



http://nebula.curie.fr/
Nebula
Human chr1:...
Human chr1:...
nebula
Analyze Data
Workflow
Shared Data
Admin
Help
User
Using 82%
Tools
Options
NGS: BED Tools
NGS: Peak Calling
MACS Model-based Analysis of ChIP-Seq
PeakSplitter Subdivides peak regions containing more than one site of signal enrichment
CCAT Control-based ChIP-seq Analysis Tool
Get Subset for ChIP Control
FindPeaks a Peak Finder/Analysis application for the ChIP-Seq
Get peak height distribution
Filter FindPeaks output (.peaks) using Control Peaks
Convert FindPeaks output (.peaks) into Bed
Convert CCAT output (intervals) into Bed
NGS: Peak Annotation
History
Options
Histones chr1
1.5 Gb
64: CCAT on data 45 and data 17 (log)
63: CCAT on data 45 and data 17 (top peaks)
62: CCAT on data 45 and data 17 (regions)
61: CCAT on data 45 and data 17 (peaks)
60: Peak location distribution (stats)
59: Peak location distribution (png)
45: Input_for_H3K27me3_chr1.sorted.noDup.bam (as BED)
Execute

Convert CCAT output (intervals) into Bed

type of experiment: CCAT_H3K27

CCAT output interval file: 62: CCAT on data 45 a..7 (regions)

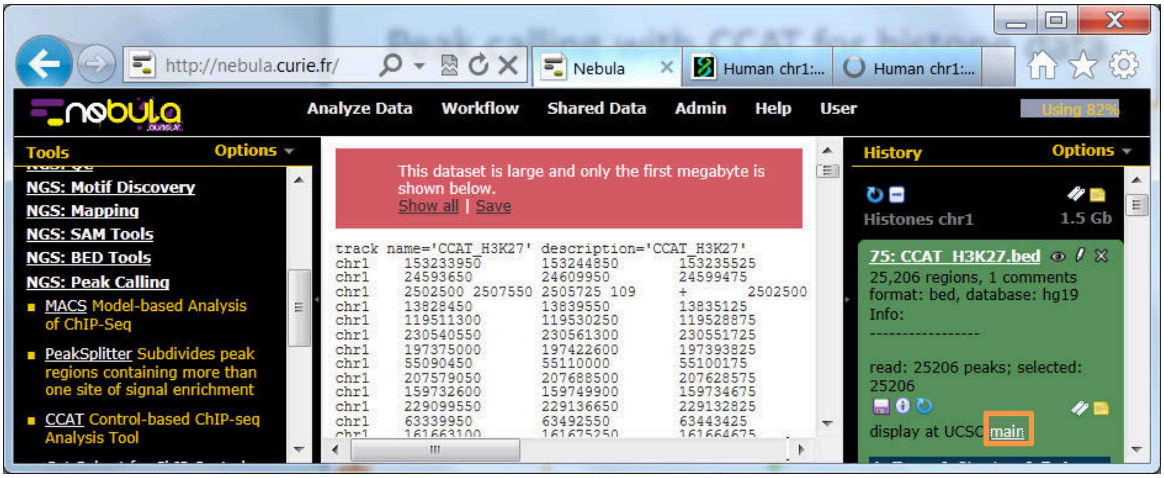
Minimal read count per peak to consider: 0

What it does

This tool creates a .bed file using .peaks output of CCAT

Peak calling with CCAT for histone data

- Visualize .Bed file:



NGS: Motif Discovery
NGS: Mapping
NGS: SAM Tools
NGS: BED Tools
NGS: Peak Calling

- MACS Model-based Analysis of ChIP-Seq
- PeakSplitter Subdivides peak regions containing more than one site of signal enrichment
- CCAT Control-based ChIP-seq Analysis Tool

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

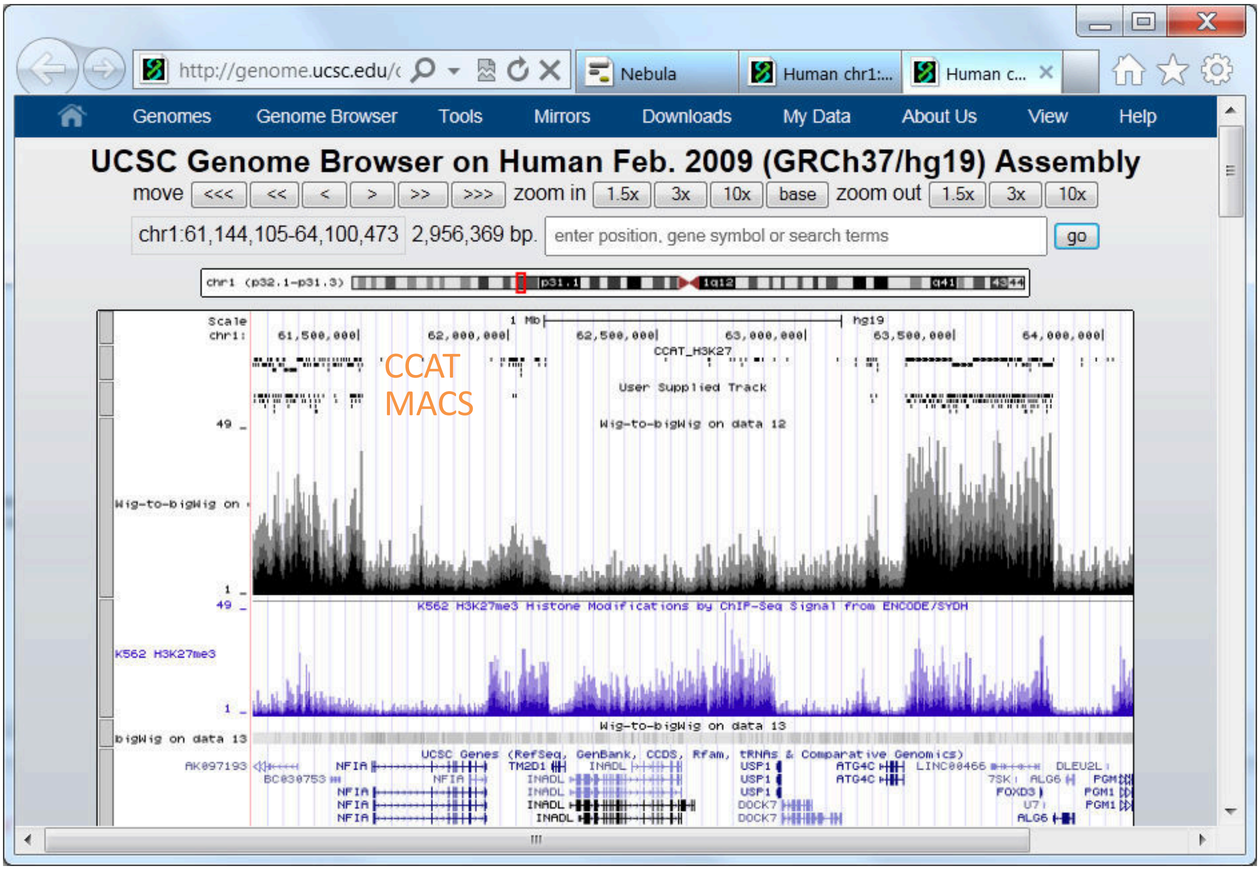
```
track name='CCAT_H3K27' description='CCAT_H3K27'
chr1    153233950       153244850       153235525
chr1    24593650        24609950        24599475
chr1    2502500 2507550 2505725 109     +       2502500
chr1    13828450        13839550        13835125
chr1    119511300       119530250       119528875
chr1    230540550       230561300       230551725
chr1    197375000       197422600       197393825
chr1    55090450        55110000        55100175
chr1    207579050       207688500       207628575
chr1    159732600       159749900       159734675
chr1    229099550       229136650       229132825
chr1    63339950        63492550        63443425
chr1    161663100       161675250       161664675
```

Histones chr1 1.5 Gb

75: CCAT_H3K27.bed
25,206 regions, 1 comments
format: bed, database: hg19
Info:
----

read: 25206 peaks; selected: 25206

display at UCSC main



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr1:61,144,105-64,100,473 2,956,369 bp. enter position, gene symbol or search terms

Annotate predicted histone marks

- Narrow marks:
  - H3K4me3
  - H3K4me1

The same kind of analysis as for transcription factors!

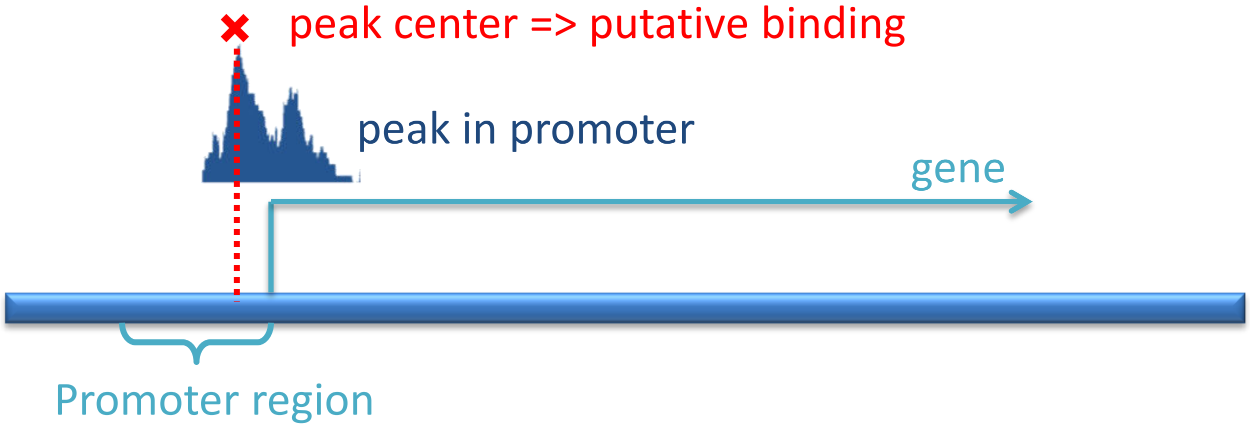
Where are they located respectively to the gene transcription starts?

- Large marks
  - HK36me3
  - H3K27me3
  - H3K9ac

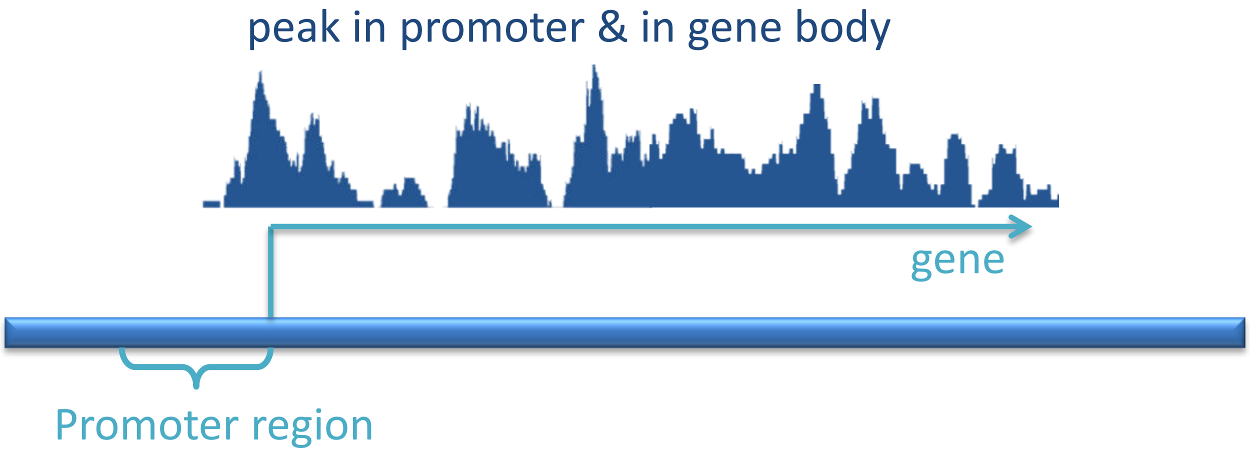
Are gene transcription starts covered by these marks? How much of the gene body is covered?

Annotate predicted histone marks

- Narrow marks:



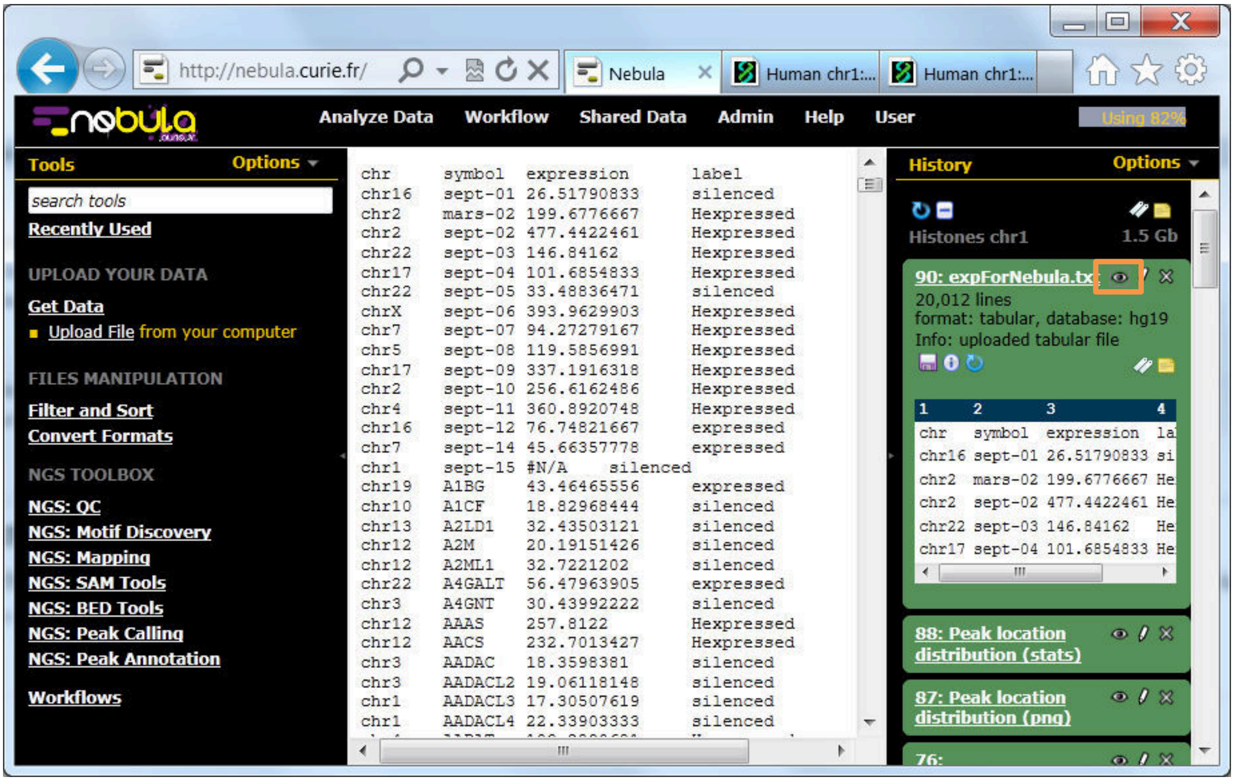
- Large marks



Format of gene expression/modulation file

- Tab-delimited:

1. Some field
2. Gene symbol
3. Some value (expression or fold change)
4. Gene feature: e.g., expressed, regulated, etc.

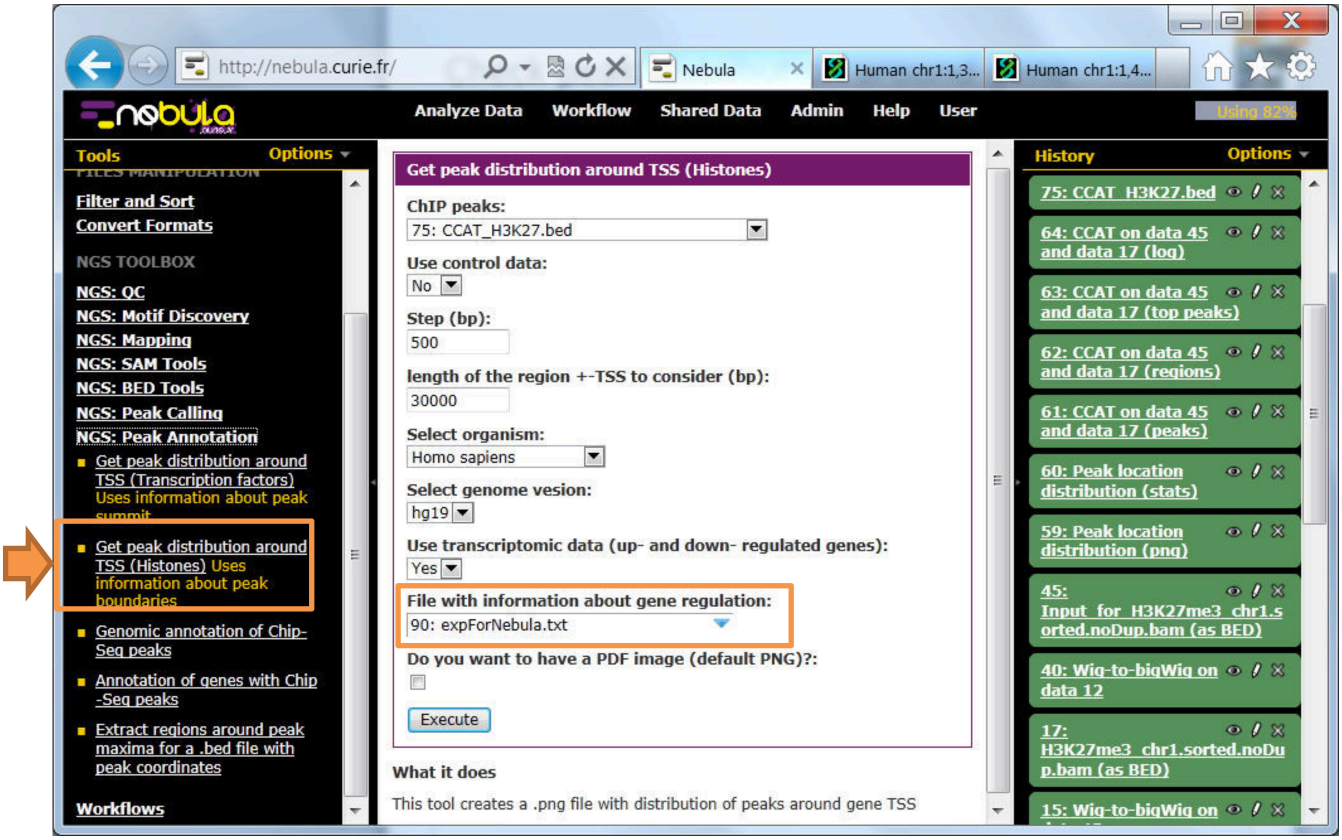


```
chr     symbol   expression      label
chr16   sept-01 26.51790833     silenced
chr2    mars-02 199.6776667     Hexpressed
chr2    sept-02 477.4422461     Hexpressed
chr22   sept-03 146.84162       Hexpressed
chr17   sept-04 101.6854833     Hexpressed
chr22   sept-05 33.48836471     silenced
chrX    sept-06 393.9629903     Hexpressed
chr7    sept-07 94.27279167     Hexpressed
chr5    sept-08 119.5856991     Hexpressed
chr17   sept-09 337.1916318     Hexpressed
chr2    sept-10 256.6162486     Hexpressed
chr4    sept-11 360.8920748     Hexpressed
chr16   sept-12 76.74821667     expressed
chr7    sept-14 45.66357778     expressed
chr1    sept-15 #N/A    silenced
chr19   A1BG    43.46465556     expressed
chr10   A1CF    18.82968444     silenced
chr13   A2LD1   32.43503121     silenced
chr12   A2M     20.19151426     silenced
chr12   A2ML1   32.7221202      silenced
chr22   A4GALT  56.47963905     expressed
chr3    A4GNT   30.43992222     silenced
chr12   AAAS    257.8122        Hexpressed
chr12   AACS    232.7013427     Hexpressed
chr3    AADAC   18.3598381      silenced
chr3    AADACL2 19.06118148     silenced
chr1    AADACL3 17.30507619     silenced
chr1    AADACL4 22.33903333     silenced
```

20,012 lines
format: tabular, database: hg19
Info: uploaded tabular file

```
1     2       3           4
chr   symbol  expression  la
chr16 sept-01 26.51790833 si
chr2  mars-02 199.6776667 He
chr2  sept-02 477.4422461 He
chr22 sept-03 146.84162   He
chr17 sept-04 101.6854833 He
```

Check peak distribution around gene transcription start sites (TSS)



Tools Options

FILES MANIPULATION

- Filter and Sort
- Convert Formats

NGS TOOLBOX

- NGS: QC
- NGS: Motif Discovery
- NGS: Mapping
- NGS: SAM Tools
- NGS: BED Tools
- NGS: Peak Calling
- NGS: Peak Annotation
  - Get peak distribution around TSS (Transcription factors) Uses information about peak summit
  - Get peak distribution around TSS (Histones) Uses information about peak boundaries
  - Genomic annotation of Chip-Seq peaks
  - Annotation of genes with Chip-Seq peaks
  - Extract regions around peak maxima for a .bed file with peak coordinates
- Workflows

Get peak distribution around TSS (Histones)

ChIP peaks: 75: CCAT_H3K27.bed

Use control data: No

Step (bp): 500

length of the region +-TSS to consider (bp): 30000

Select organism: Homo sapiens

Select genome vesion: hg19

Use transcriptomic data (up- and down- regulated genes): Yes

File with information about gene regulation: 90: expForNebula.txt

Do you want to have a PDF image (default PNG)?: □

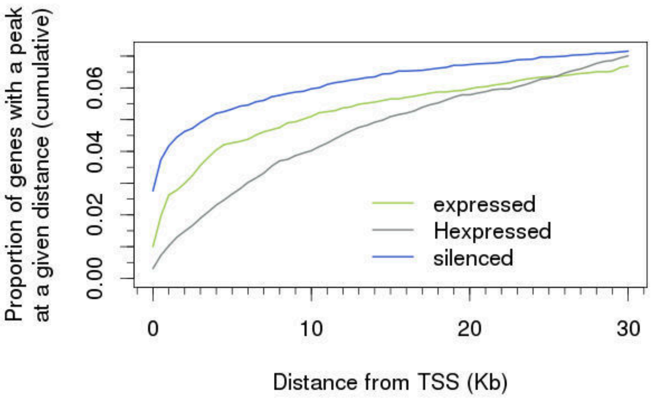
Execute

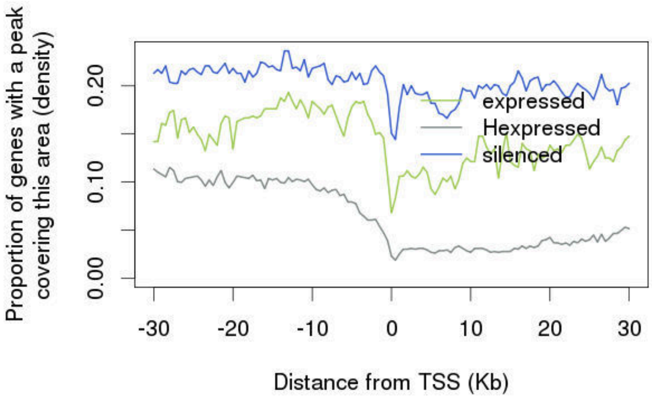
What it does

This tool creates a .png file with distribution of peaks around gene TSS

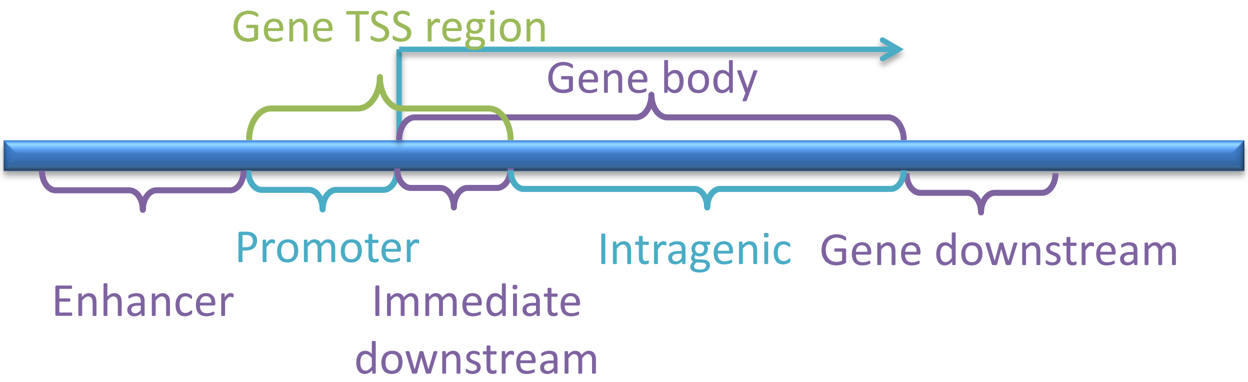
History Options

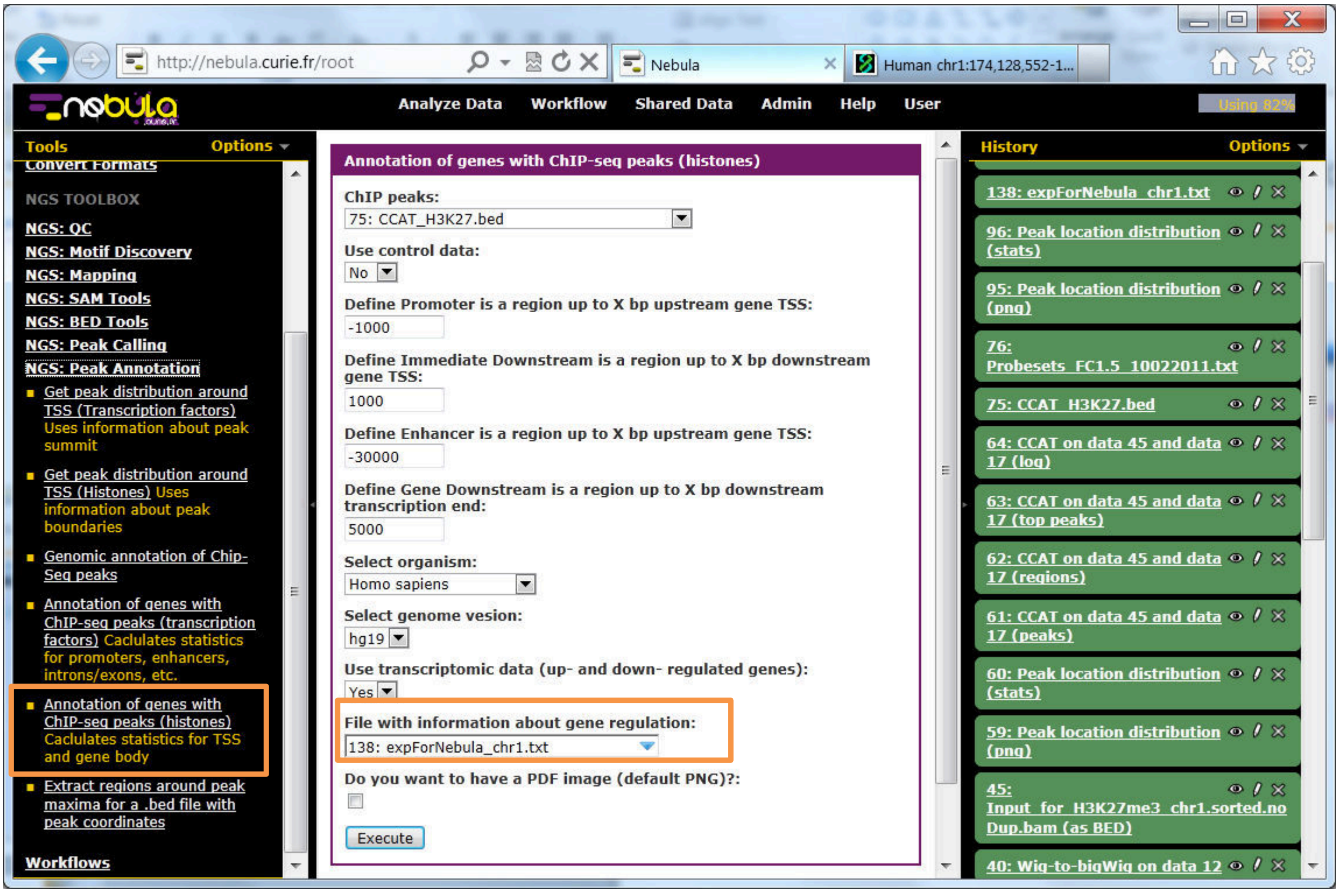
- 75: CCAT_H3K27.bed
- 64: CCAT on data 45 and data 17 (log)
- 63: CCAT on data 45 and data 17 (top peaks)
- 62: CCAT on data 45 and data 17 (regions)
- 61: CCAT on data 45 and data 17 (peaks)
- 60: Peak location distribution (stats)
- 59: Peak location distribution (png)
- 45: Input_for_H3K27me3_chr1.sorted.noDup.bam (as BED)
- 40: Wig-to-bigWig on data 12
- 17: H3K27me3_chr1.sorted.noDup.bam (as BED)
- 15: Wig-to-bigWig on





Check peak presence in genomic regions





Convert Formats

NGS TOOLBOX

- NGS: QC
- NGS: Motif Discovery
- NGS: Mapping
- NGS: SAM Tools
- NGS: BED Tools
- NGS: Peak Calling
- NGS: Peak Annotation
  - Get peak distribution around TSS (Transcription factors) Uses information about peak summit
  - Get peak distribution around TSS (Histones) Uses information about peak boundaries
  - Genomic annotation of Chip-Seq peaks
  - Annotation of genes with ChIP-seq peaks (transcription factors) Caclulates statistics for promoters, enhancers, introns/exons, etc.
  - Annotation of genes with ChIP-seq peaks (histones) Caclulates statistics for TSS and gene body
  - Extract regions around peak maxima for a .bed file with peak coordinates

Workflows

Annotation of genes with ChIP-seq peaks (histones)

ChIP peaks: 75: CCAT_H3K27.bed

Use control data: No

Define Promoter is a region up to X bp upstream gene TSS: -1000

Define Immediate Downstream is a region up to X bp downstream gene TSS: 1000

Define Enhancer is a region up to X bp upstream gene TSS: -30000

Define Gene Downstream is a region up to X bp downstream transcription end: 5000

Select organism: Homo sapiens

Select genome vesion: hg19

Use transcriptomic data (up- and down- regulated genes): Yes

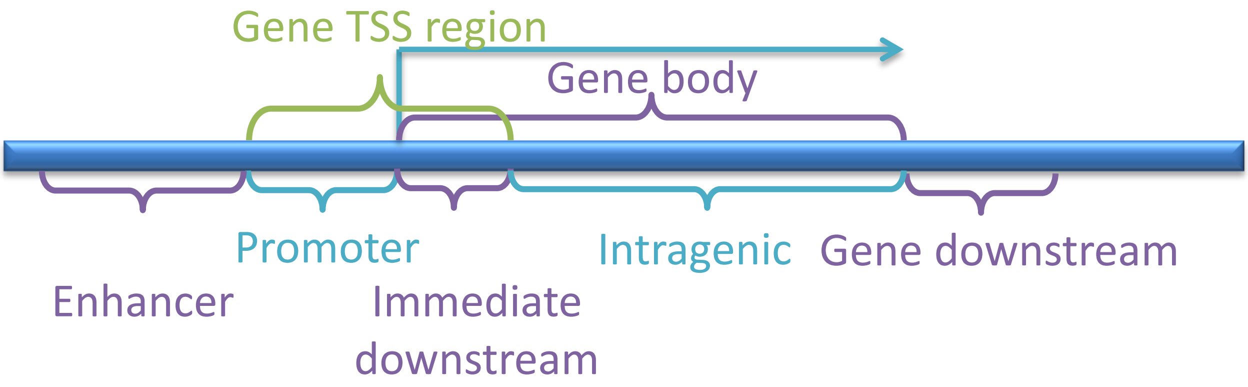
File with information about gene regulation: 138: expForNebula_chr1.txt

Do you want to have a PDF image (default PNG)?: □

Execute

- 138: expForNebula_chr1.txt
- 96: Peak location distribution (stats)
- 95: Peak location distribution (png)
- 76: Probesets_FC1.5_10022011.txt
- 75: CCAT_H3K27.bed
- 64: CCAT on data 45 and data 17 (log)
- 63: CCAT on data 45 and data 17 (top peaks)
- 62: CCAT on data 45 and data 17 (regions)
- 61: CCAT on data 45 and data 17 (peaks)
- 60: Peak location distribution (stats)
- 59: Peak location distribution (png)
- 45: Input_for_H3K27me3_chr1.sorted.noDup.bam (as BED)
- 40: Wig-to-bigWig on data 12

Check peak presence in genomic regions



Gene TSS region

Gene body

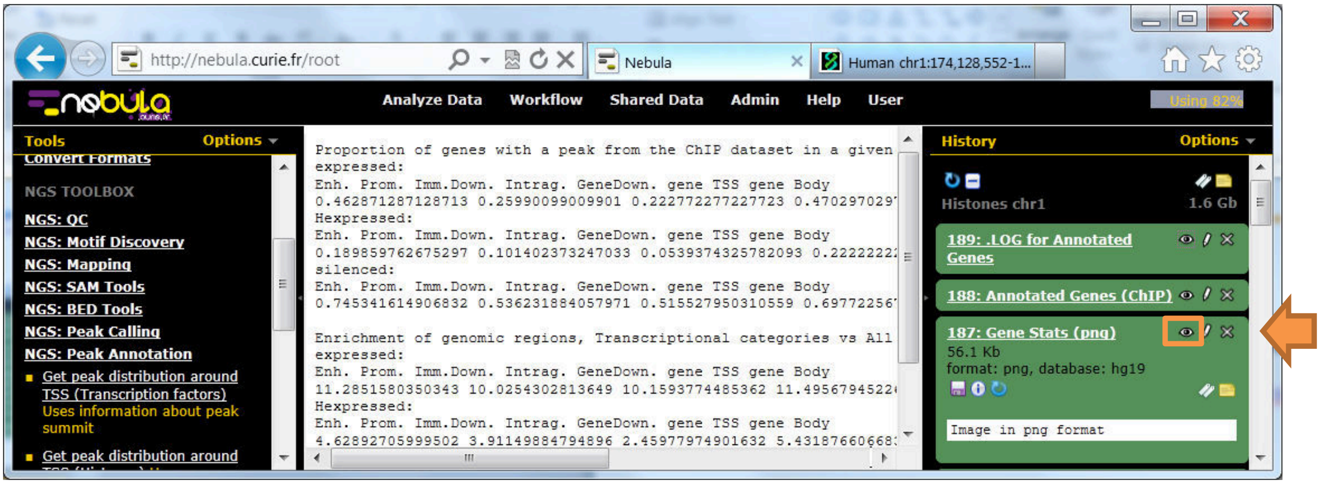
Enhancer

Promoter

Immediate downstream

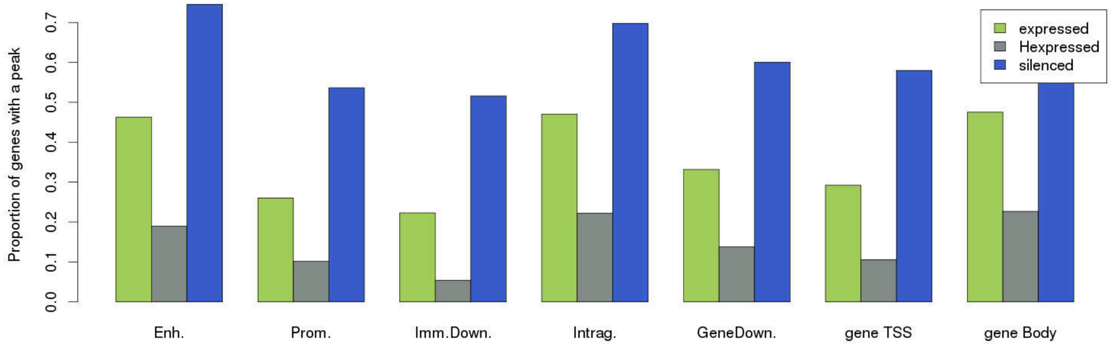
Intragenic

Gene downstream



```
Proportion of genes with a peak from the ChIP dataset in a given
expressed:
Enh. Prom. Imm.Down. Intrag. GeneDown. gene TSS gene Body
0.462871287128713 0.25990099009901 0.222772277227723 0.470297029
Hexpressed:
Enh. Prom. Imm.Down. Intrag. GeneDown. gene TSS gene Body
0.189859762675297 0.101402373247033 0.0539374325782093 0.2222222
silenced:
Enh. Prom. Imm.Down. Intrag. GeneDown. gene TSS gene Body
0.745341614906832 0.536231884057971 0.515527950310559 0.69772256

Enrichment of genomic regions, Transcriptional categories vs All
expressed:
Enh. Prom. Imm.Down. Intrag. GeneDown. gene TSS gene Body
11.2851580350343 10.0254302813649 10.1593774485362 11.4956794522
Hexpressed:
Enh. Prom. Imm.Down. Intrag. GeneDown. gene TSS gene Body
4.62892705999502 3.91149884794896 2.45977974901632 5.43187660668
```



Annexes

File formats:

- .BAI – index for a .BAM file (to visualize .BAM in UCSC)
- .BAM – aligned reads, binary .SAM
- .BED – genomic coordinates

```
chr1    23386792        23387348        23387022        8.336   +
chr1    24005015        24005240        24005133        8.680   +
chr1    36187196        36187544        36187322        12.0    +
```

- .BW (BIG WIG) – signal profile (to visualize .WIG in UCSC)
- .CSFASTA – read sequences in color code

```
>921_41_109_R17C7_F3
T21301022211321012213332130021213212202231322222222
```

- .FASTA – DNA sequences

```
>chr1:3525467-3526150
ACTGGGTAAATAGCAGGTAGCAATTTTATGCAGAGGTTGGAGCTCACTTGGAACACACTTCCACCTTTG
```

- .FASTQ – read sequences and qualities

```
@921_29_592_R17C7
T1002011.220200122120311200111132121311211322200220
+
&2*8411!6%'#,)##)$'#*&-5&&-&4-&%&,$&+%*$$+-,&'&4)#
```

- .SAM – aligned reads

```
119_171_1134_R17C1_     0       chr1    3000539 255     50M     *
286_1719_1498_R17C1_    0       chr1    3000539 255     50M     *
391_1794_580_R17C1_     0       chr1    3000539 255     50M     *
```

- .QUAL – read qualities

```
>921_41_132_R17C7_F3
8 13 3 5 6 4 4 4 10 6 2 13 5 8 6 5 6 11 7 3 5 13 15 5 5 2 5 10 6 15 12
```

- .WIG – signal profile (for visualization)

```
variableStep chrom=chr1 span=15
3000181 1
3000196 1.1
3000211 1.3
3000226 1.8
3000241 1.4
```