

ChIP-Seq data analysis workshop

UC Berkeley, CGRL, March 2013

Chitra Kotwaliwale (chitra.kot@gmail.com)

What this workshop is meant to do

- You're a biologist. You did a ChIP-Seq experiment because you're studying a protein that binds to DNA. You want to know more about it.
- You just got your sequence data back. Now what?
- You will learn to do the following:
 - ❖ map reads back to genome
 - ❖ identify bound regions
 - ❖ view your data
 - ❖ do some initial analysis

What you will **NOT** learn from this workshop

- programming
- do the kind of sophisticated analysis that may be required for publication

What you can answer with ChIP-Seq

Used to investigate protein-DNA interactions including

- ❖ Transcription factors
- ❖ Polymerase
- ❖ Histone modifications
- ❖ Structural components (cohesins, condensins)

-Identify bound regions along genome

-Quantify binding occupancy (how much of the genome is bound by your protein etc.)

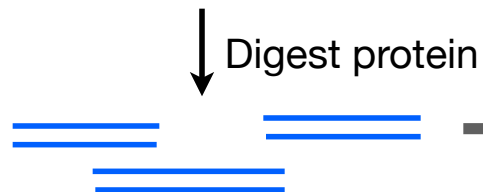
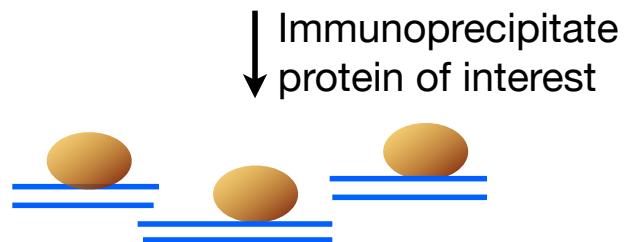
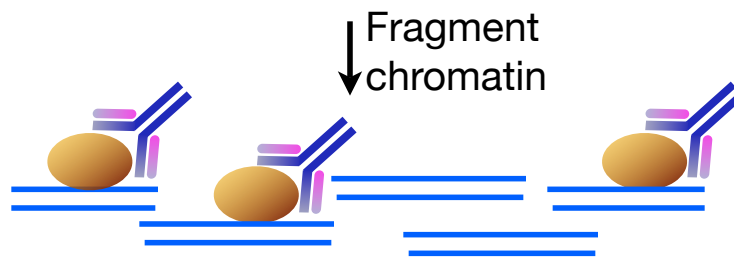
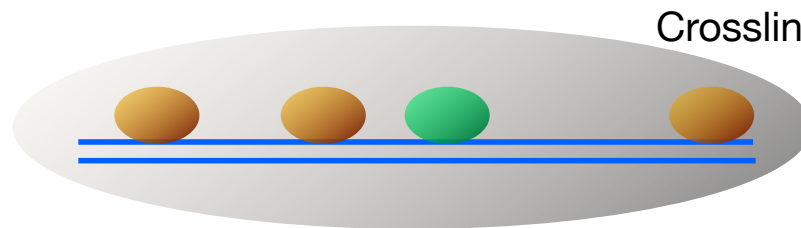
-Estimate peaks, identify DNA motif

-Where are binding sites in the genome? In genes? promoters? etc.

-Compare to other genomic data (other targets, time points, mutants, etc.)

Ultimately, the questions you need to answer with your data depends on your protein target

Overview of ChIP-Seq



Some words of caution:

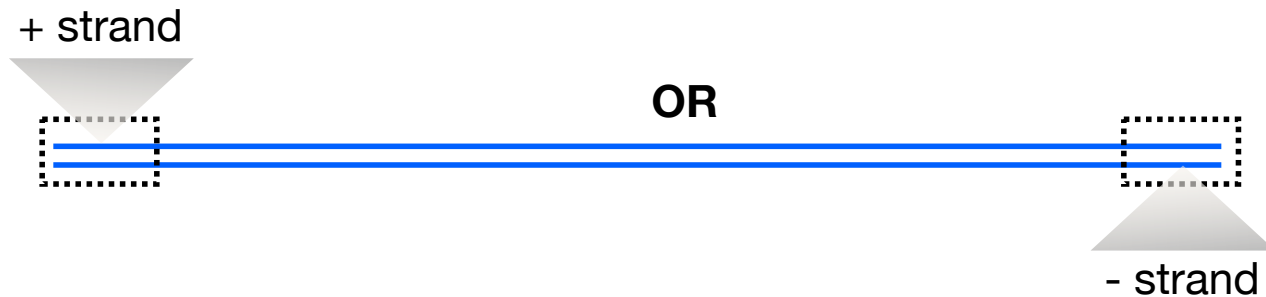
- This is an **enrichment** assay, not a purification
- Usually very small fraction of final DNA corresponds to actual signal
- How much of your final data corresponds to signal depends on **IP efficiency**, **abundance of protein** in a population of cells, **number of sites** in the genome bound by your protein

Make sequencing library

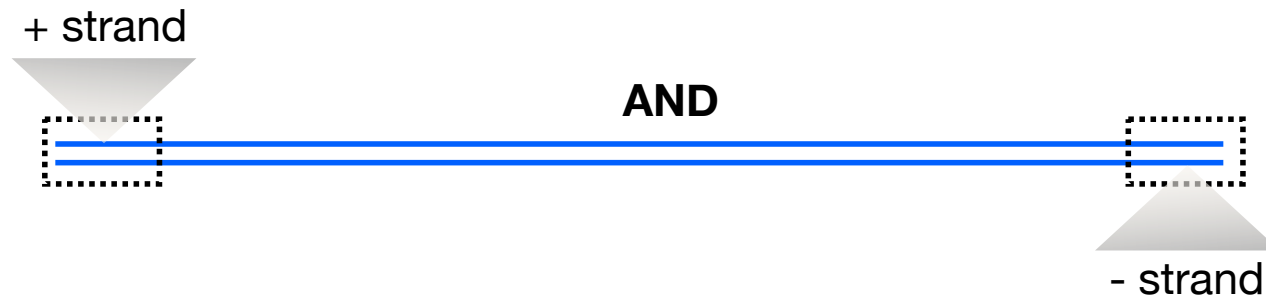
Sequencing can be done as single or paired end

Single end

Enriched DNA fragment (200-500 bp)



Paired end



For ChIP-Seq, single end sequencing is common

General workflow for analysis of ChIP-Seq data

Download raw sequence reads using ftp server



Align sequence reads to reference genome



Normalize to Input



Identify “peaks”



Downstream analysis

Mapping raw data to reference genome

The goal of mapping is to generate files with information about where reads align in the genome

Single end: chromosome, coordinate, strandedness

Paired end: chromosome, 5' position, 3' position

Sequence aligners:

Bowtie, MAQ, Eland etc.

(<http://seqanswers.com/forums/showthread.php?t=43>)

Bowtie: An ultrafast memory-efficient short read aligner

<http://bowtie-bio.sourceforge.net/index.shtml>
<http://bowtie-bio.sourceforge.net/manual.shtml>

Notable parameters:

- q Input file is fastq format. Also the default setting
- p Number of processors to use
- n Maximum number of mismatches permitted. This may be 0, 1, 2 or 3. Default is 2.
- m Suppress all alignments if more than <int> reportable alignments exist for it.
- S Print alignments in SAM format.

<http://bowtie-bio.sourceforge.net/index.shtml>

Example bowtie command:

Be careful what version of the genome reads are mapped to!

Bowtie input and output

Input file (fastq)

```
@CCFFFFFFDDDHJIGIJJHIIIIII<@DDGGHJ2B*/B4=<FB=@@F##  
@HS3:245:C155KACXX:6:1101:3264:86731 1:N:0:ATCACG  
TCTCATCGAGTTTCTTCGATTTTCCTATGAGCTCCTGTTCCACTGCAATC
```

Output file (default bowtie output)

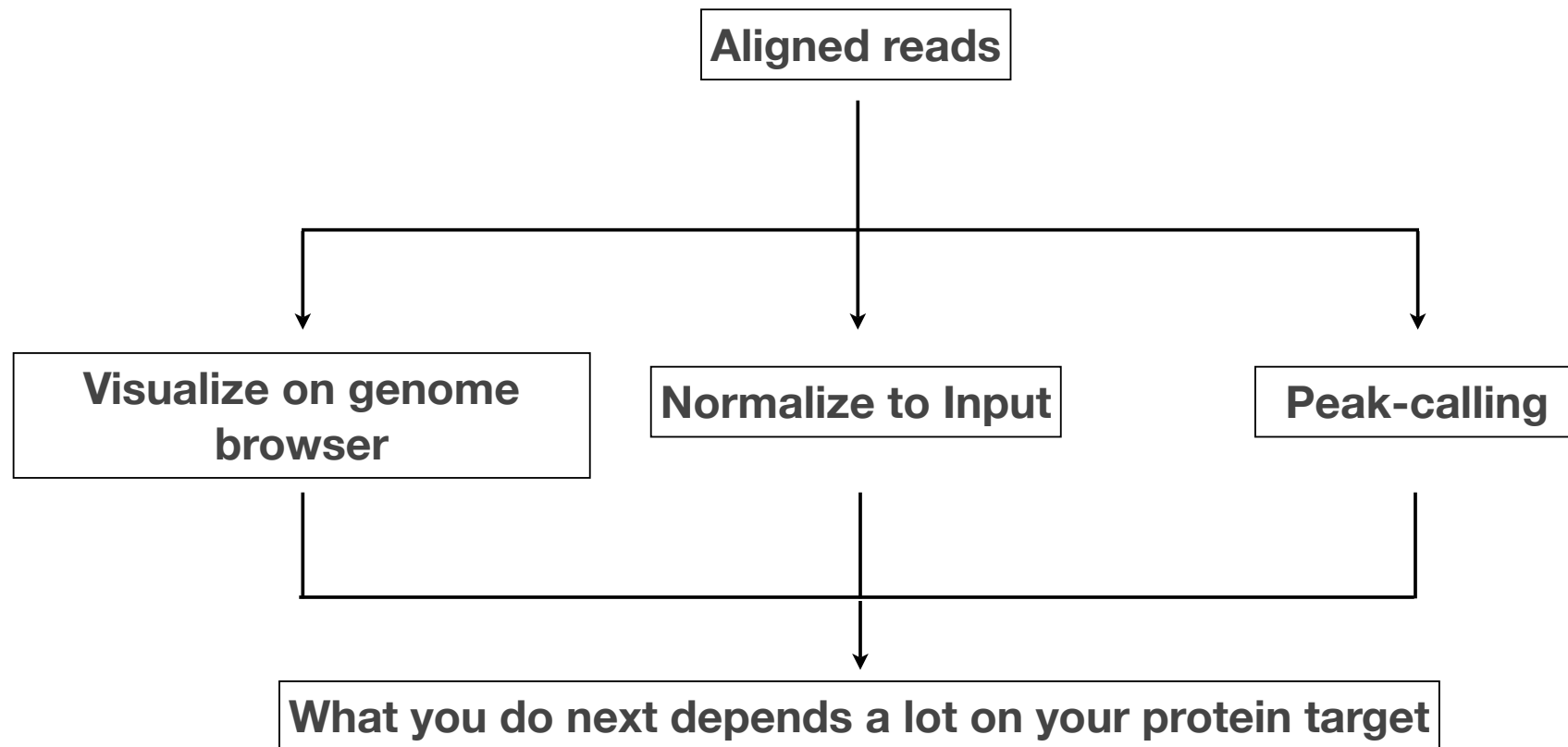
```
HS1:177:d0yyjacxx:6:1101:2688:49638 1:Y:0:ATCACG |+| IV| 16643508 |  
      CCATCTGAACCATGCGCGTCCAGACGCCCTTCTCGGGCACCAAAAGAGCC | :=+2<2<A=CA;  
3<7<=?2A=)7*10):8=3999AA<' '5=AAB==57>; |0|0:T>C,26:A>C
```

Name of read |strand |chr |position |sequence |read quality |number of other places read aligned to* |mismatch descriptors**

*This is *not* the number of other places the read aligns with the same number of mismatches.

**If there are no mismatches in the alignment, this field is empty. A single descriptor has the format offset:reference-base>read-base. The offset is expressed as a 0-based offset from the high-quality (5') end of the read.

How to proceed after reads are mapped



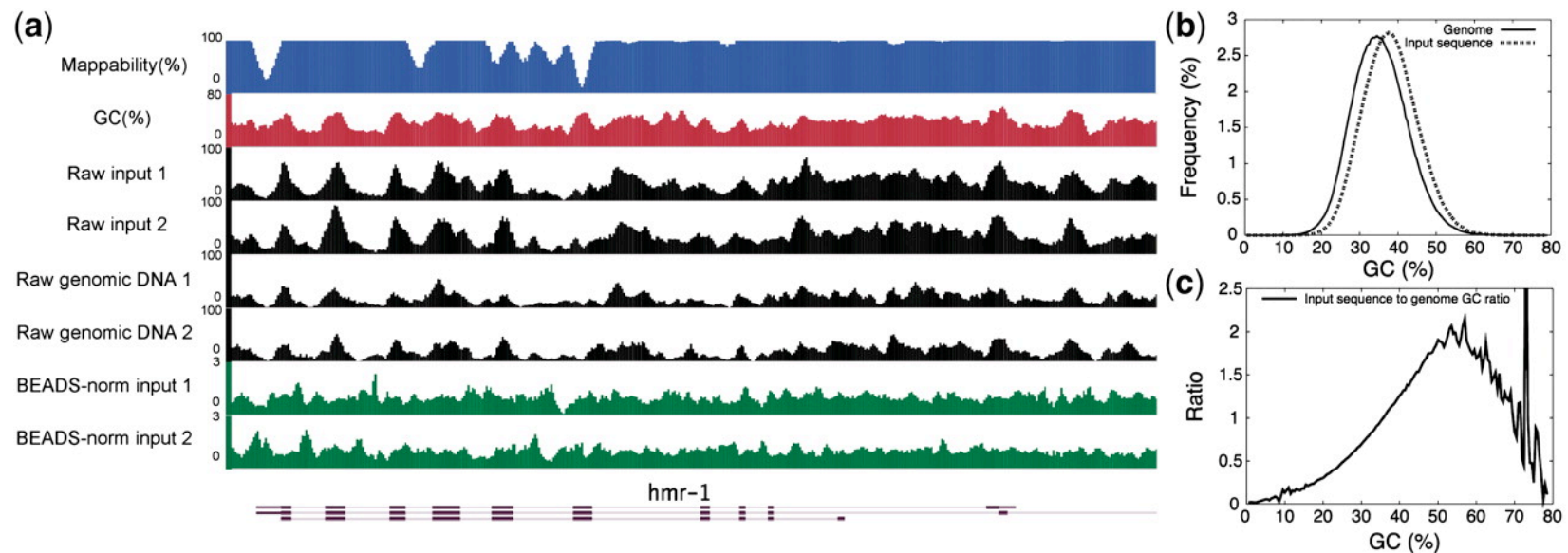
Nature of binding sites differs based on protein target



Certain tools perform better for certain targets (One size doesn't fit all!)

Good practice to sequence Input

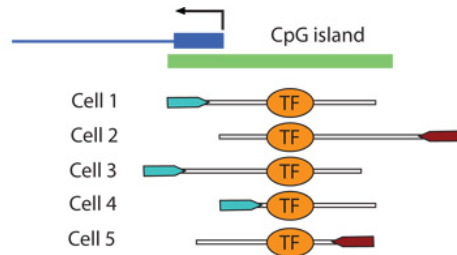
- ❖ sample to sample variability (introduced during extract preparation)
- ❖ sequencing biases (GC bias)
- ❖ variation in sequencing based on biological variation (for e.g. euchromatin generates more reads than heterochromatin)



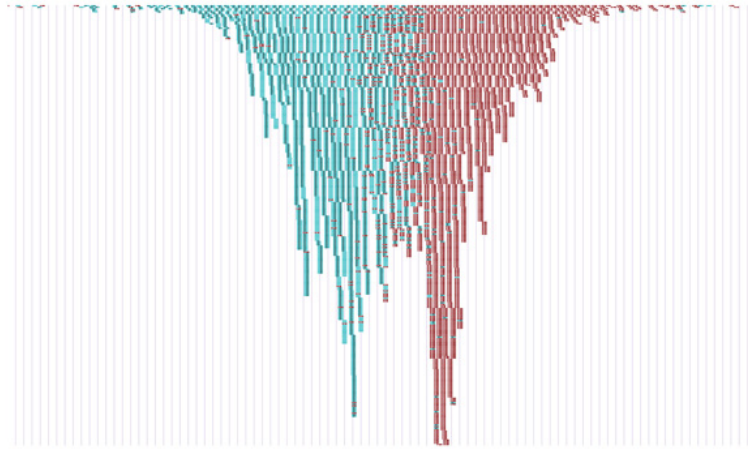
Identification of target binding sites

- ❖ Identify binding sites from aligned reads
- ❖ In principal, genomic intervals with lots of reads should indicate signal
- ❖ But regions with lots of reads could also be due to
 - Sequencing biases
 - Chromatin biases
 - PCR biases/artifacts
 - Biases/artifacts of unknown origin
- ❖ So need to separate signal from noise

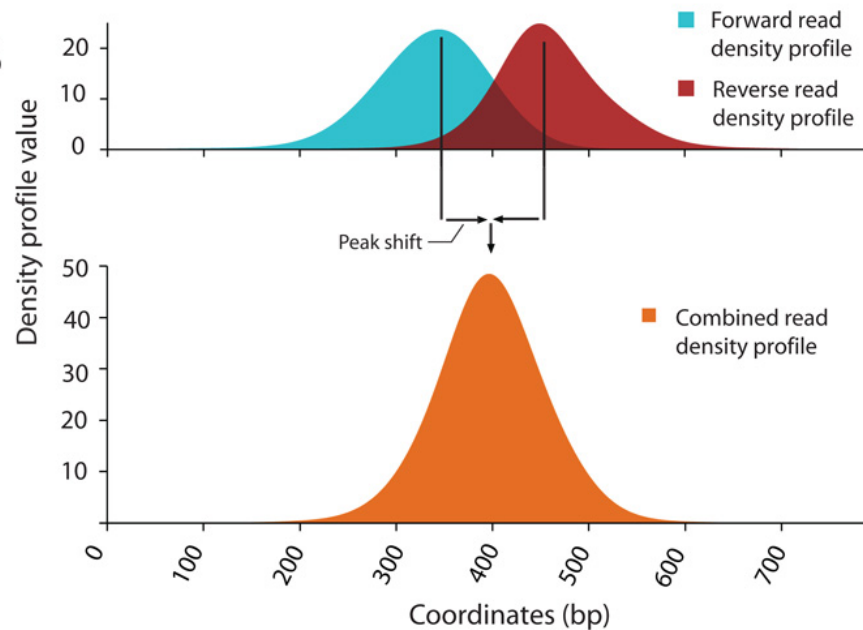
A



How are ChIP binding sites distinguished from noise?



B



Valouev et al., 2008

Peak-calling

Process of finding regions enriched due to events of interest and inferring the location of the event in those regions

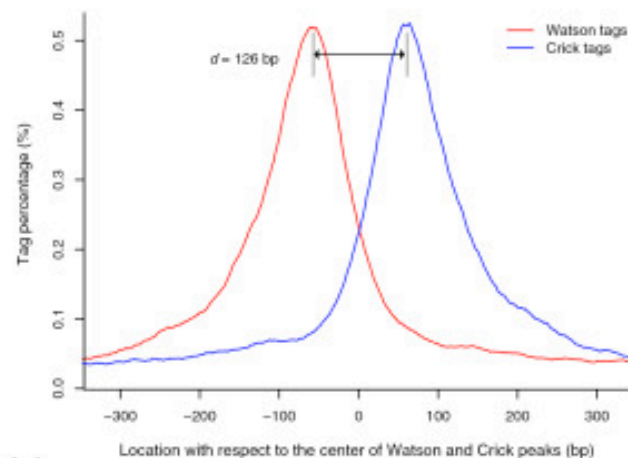
Tools for peak calling:

20+ packages out there: ERANGE, FindPeaks, **MACS**, QuEST, CisGenome, SISSRS, USeq, PeakSeq, **SPP**, ChIPSeqR, GLITR, ChIPDiff, T-PIC, BayesPeak, MOSAiCS, CCAT, CSAR, and others.

MACS

(<http://liulab.dfci.harvard.edu/MACS/>)

(a)



- empirically models the length of the sequenced ChIP fragments, which tends to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites
- uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence, allowing for more sensitive and robust prediction
- for experiments with a control, linearly scales the total control tag count to be the same as the total ChIP tag count
- removes duplicate tags that may arise as a result of over-amplification

MACS peak-calling & output files

MACS command

```
$macs2 -t ChIP -c Input -g ce -n test --nomodel --shiftsize 250 --nolambda
```

-t Treatment file (ChIP)
-c Control file (Input)
-g Genome size (can specify the number or name. ce=c. elegans, hs=homo sapiens etc.)
-n Name of Output file
--nomodel Do not build model of length of the sequenced ChIP fragments
--shiftsize The arbitrary shift size in bp. When nomodel is true, MACS will use this value as 1/2 of fragment size. DEFAULT: 100
--nolambda MACS will not consider the local bias at peak candidate regions
Many other parameters! Refer to MACS manual.

Output files:

test_peaks.bed

Peak coordinates in bed format. Can be loaded on genome browser

test_peaks.xls

Information about peaks in an excel spreadsheet.

test_summits.bed

Peak summits in bed format. Can be loaded on genome browser

Information about different file formats can be found here:

<http://genome.ucsc.edu/FAQ/FAQformat.html>

MACS peaks bed file

(“head” is a quick way to print the first 10 lines of your file to terminal)

```
[ckotwali@poset chipseqWorkshop]$ head test_peaks.bed
```

```
III 41332 41878 MACS_peak_1 11.14
III 66533 67150 MACS_peak_2 6.45
III 85654 88386 MACS_peak_3 591.10
III 94608 95241 MACS_peak_4 10.27
III 131504 132477 MACS_peak_5 56.06
III 187281 189824 MACS_peak_6 338.76
III 234728 237213 MACS_peak_7 381.09
III 343360 344257 MACS_peak_8 34.59
III 347993 349954 MACS_peak_9 58.73
III 354984 356197 MACS_peak_10 31.63
```

BED format

Indi

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding: lower-numbered fields must always be populated if higher-numbered fields are used.

If your data set is BED-like, but it is very large and you would like to keep it on your own server, you should use the [bigBed](#) data format.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0*, *chromEnd=100*, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

Fold enrichment tracks without peak calling

```
$ macs2 callpeak -t ChIP -c Input -B --nomodel --shiftsize 250 -g ce -n test --bdg
```

`-B` generates pileup signal file of 'fragment pileup per million reads' in bedGraph format for ChIP and Input

```
$ macs2 bdgcmp -t test_treat_pileup.bdg -c test_control_lambda.bdg -o test_FE.bdg -m  
FE
```

`-m` FE means to calculate fold enrichment. Other options can be logLR for log likelihood, subtract for subtracting noise from treatment sample.

This gives a bedGraph file for fold-enrichment

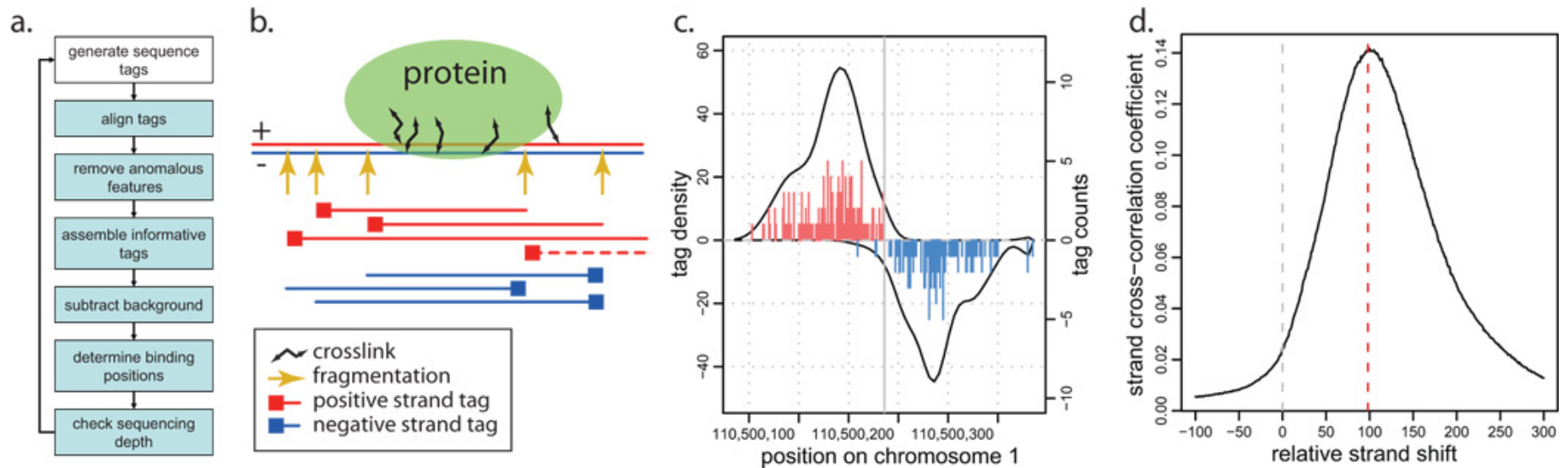
Example of fold-enrichment output file

Generates a fold-enrichment estimate for the entire genome, not just for regions identified as peaks

```
chrII 475236 475237 1.90872
chrII 475237 475238 1.95311
chrII 475238 475240 1.99750
chrII 475240 475243 1.98236
chrII 475243 475244 2.02642
chrII 475244 475248 2.01118
chrII 475248 475249 1.99617
chrII 475249 475250 1.98139
chrII 475250 475251 2.02446
chrII 475251 475252 2.00957
chrII 475252 475253 2.05233
chrII 475253 475255 2.03734
chrII 475255 475256 2.06472
chrII 475256 475257 2.04986
chrII 475257 475258 2.03522
chrII 475258 475261 2.04986
chrII 475261 475263 2.09170
```

SPP: A ChIP-seq peak calling algorithm, implemented as an R package (need to know a bit of R)

<http://compbio.med.harvard.edu/Supplements/ChIP-seq/tutorial.html>



(Outputs a background-subtracted tag density file in addition to peak file after some smoothing)

Example of background subtracted tag density output

chrI	17340	17349	63.161767709595
chrI	17350	17359	56.2122202206993
chrI	17360	17369	49.5706724963616
chrI	17370	17379	43.2426983605907
chrI	17380	17389	37.2081530953005
chrI	17390	17399	31.5863570003275
chrI	17400	17409	26.248253616307
chrI	17410	17419	21.3446957532495
chrI	17420	17429	16.5688547233586
chrI	17430	17439	12.2160383966229
chrI	17440	17449	7.90543164929034
chrI	17450	17459	3.83098053302168
chrI	17460	17469	0.185463404745037
chrI	17470	17479	-3.43966779154329
chrI	17480	17489	-6.86730541279676
chrI	17490	17499	-9.90233696573392
chrI	17500	17509	-12.6258154538576
chrI	17510	17519	-15.2672540209483
chrI	17520	17529	-17.7164576992632

Looking at data on genome browser (UCSC genome browser or Integrated genome browser)

<http://genome.ucsc.edu/>

UCSC Genome Bioinformatics

Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help

Genome Browser

ENCODE

Neandertal

Blat

Table Browser

Gene Sorter

In Silico PCR

Genome Graphs

Galaxy

VisiGene

Utilities

Downloads

Release Log

Custom Tracks

Microbial


About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the [ENCODE](#) and [Neandertal](#) projects.

We encourage you to explore these sequences with our tools. The [Genome Browser](#) zooms and scrolls over chromosomes, showing the work of annotators worldwide. The [Gene Sorter](#) shows expression, homology and other information on groups of genes that can be related in many ways. [Blat](#) quickly maps your sequence to the genome. The [Table Browser](#) provides convenient access to the underlying database. [VisiGene](#) lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns. [Genome Graphs](#) allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering ([CBSE](#)) at the University of California Santa Cruz ([UCSC](#)). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our [public mailing list](#).

News



News Archives ►

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the [genome-announce](#) mailing list.

05 March 2013 - dbSNP 137 Available for mm10

We are pleased to announce the release of three tracks derived from dbSNP build 137, available on the mouse assembly (GRCm38/mm10). dbSNP build 137 is available at NCBI. The new tracks contain additional annotation data not included in previous dbSNP tracks, with corresponding coloring and filtering options in the Genome Browser.

As for dbSNP build 137, there are three tracks in this release. One is a track containing all mappings of reference SNPs to the mouse assembly, labeled "All SNPs (137)". The other two tracks are subsets of this track and show interesting and easily defined subsets of dbSNP:

- Common SNPs (137): uniquely mapped variants that appear in at least 1% of the population
- Mult. SNPs (137): variants that have been mapped to more than one genomic location

genome.ucsc.edu/cgi-bin/hgTracks?org=human

Displays region of genome associated with RefSeq gene mec-8
Displays the region associated with the Sanger predicted gene F46A9.4
Displays region of genome associated with RefSeq identifier NM_060107
Displays region of genome associated with Sanger predicted gene F46A9.4

Manage Custom Tracks

genome C. elegans assembly May 2008 (WS190/ce6) [ce6]

Name	Description	Type	Doc	Items	Pos	delete	
Bed Format	H3K4me3_r7r8 Smoothed, maximum likelihood log2 enrichment estimate	wiggle_0				<input type="checkbox"/>	add custom tracks
User Track	User Supplied Track	bed		736	chrIII:	<input type="checkbox"/>	go to genome browser
							go to table browser

Managing Custom Tracks

This section provides a brief description of the columns in custom track management table. For more details about managing custom tracks, see the Genome Browser [User's Guide](#).

- **Name** - a hyperlink to the update page where you can edit your track data.
- **Description** - the value of the "description" attribute from the track line, if present. If no description is included in the input file, this field contains the track name.
- **Type** - the track type, determined by the Browser based on the format of the data.
- **Doc** - displays "Y" (Yes) if a description page has been uploaded for the track; otherwise the field is blank.
- **Items** - the number of data items in the custom track file. An item count is not displayed for tracks lacking individual items (e.g. wiggle format data).
- **Pos** - the default chromosomal position defined by the track file in either the browser line "position" attribute or the first data line. Clicking this link opens the Genome Browser or Table Browser at the specified position (note: only the chromosome name is shown in this column). The Pos column remains blank if the track lacks individual items (e.g. wiggle format data) and the browser line "position" attribute hasn't been set.

Track display and save options

The screenshot displays the UCSC Genome Browser interface for *C. elegans*. The main track shows a genomic region on chromosome III (chrIII:6,202,657-7,581,024) with a 1,378,368 bp scale. The track is divided into two main sections: a top section for H3K4me3_r7r8 Smoothed, maximum likelihood (orange) and a bottom section for WormBase Gene Annotations (blue). The interface includes a navigation bar at the top with links to Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, About Us, View, and Help. A search bar at the top center allows for moving and zooming in the genome. A dropdown menu is open, showing options for PDF/PS, DNA, In Other Genomes (Convert), Ensembl, Configure Browser, Default Tracks, Default Track Order, and Reset All User Settings. Below the tracks, there is a section for track search and default tracks, followed by a section for custom tracks and mapping and sequencing tracks. The bottom of the page shows the URL: `u/cgi-bin/hgTracks?hgsid=329867777&hgt.psOutput=on`.

UCSC Genome Browser on *C. elegans* May 2010 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x

chrIII:6,202,657-7,581,024 1,378,368 bp. enter position, gene symbol or search

Scale chrIII: 6,300,000 6,400,000 6,500,000 6,600,000 6,700,000 6,800,000 6,900,000 7,000,000 7,100,000 7,200,000 7,300,000 7,400,000 7,500,000

Track 1 User Supplied Tracks

Track 2 User Supplied Tracks

2396 H3K4me3_r7r8 Smoothed, maximum likelihood

0

2197 WormBase Gene Annotations

Genes

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

track search default tracks default order hide all manage custom tracks track hubs configure reverse resize refresh

collapse all Use drop-down controls below and press refresh to alter tracks displayed. expand all

Tracks with lots of items will automatically be displayed in more compact modes.

Custom Tracks refresh

User Track User Track 1 Bed Format

dense dense full

Mapping and Sequencing Tracks refresh

Base Position Assembly Gap GC Percent Short Match Restr Enzymes

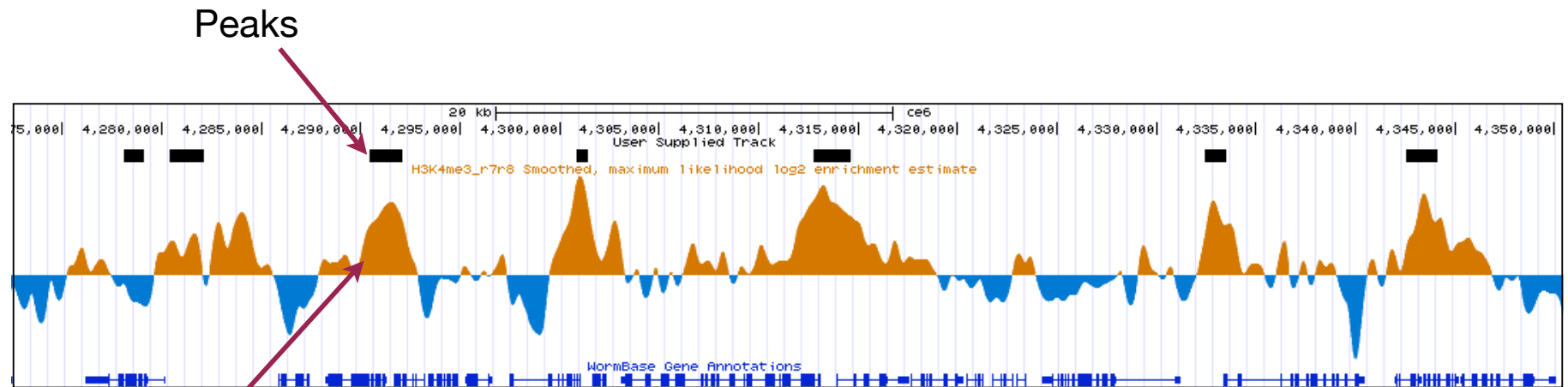
full hide hide hide hide hide

Nucleosome MNase Coverage NSome Coverage Adj NSome Covrg NSome Stringency

hide

u/cgi-bin/hgTracks?hgsid=329867777&hgt.psOutput=on

Narrow peaks vs. broad regions



Narrow peaks vs. broad regions

Point binding proteins --
Cross-correlation between
sense and anti-sense reads



Broadly binding proteins --
Weak cross-correlation between
sense and anti-sense reads.
Signal looks like noise

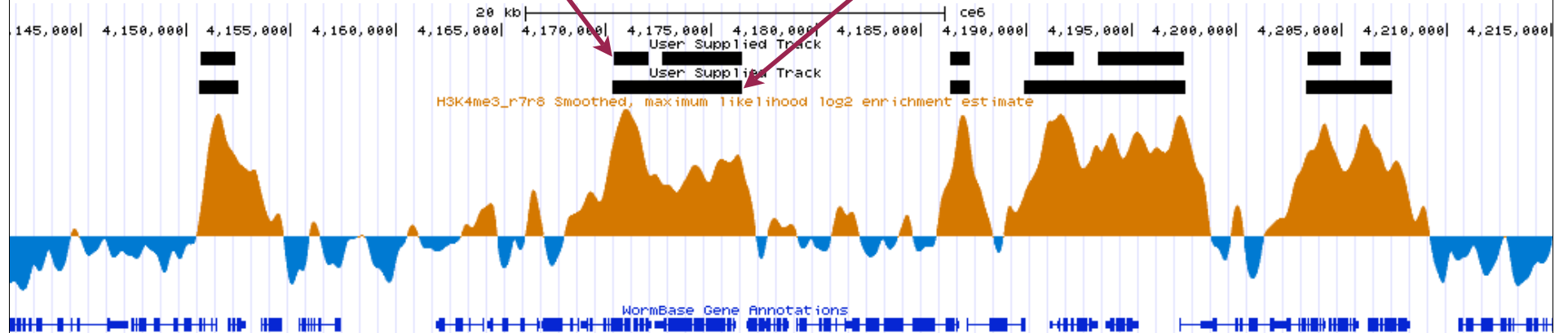


MACS has a **broadpeaks** option

```
$ macs2 -t ChIP -c Input --broad --nomodel --shiftsize 250 -g ce -n test
```

Narrow Peaks

Broad Peaks



What next?

- Downstream analysis strongly depends on what the protein target (looking at the genome browser often helps in deciding what to do next)
- Where are the peaks?
- In regulatory regions?
- In genes?
- Which genes?
- DNA sequence motif?
- What else is enriched in that region?

GALAXY

<https://main.g2.bx.psu.edu/>

Tools

search tools

Get Data

Upload File from your computer

UCSC Main table browser

UCSC Archaea table browser

BX table browser

EBI SRA ENA SRA

BioMart Central server

GrameneMart Central server

Flymine server

modENCODE fly server

modENCODE modMine server

Ratmine server

YeastMine server

modENCODE worm server

WormBase server

EuPathDB server

EncodeDB at NHGRI

EpiGRAPH server

GenomeSpace import from file browser

Send Data

ENCODE Tools

Galaxy is hiring

OSLO

workflows reproducible biology NGS science analysis

transparent data visualization

research accessible

Galaxy

Live Quickies

Basic fastQ manipulation:
Galactic quickie # 13

Advanced fastQ manipulation:
Galactic quickie # 14

454 Mapping: Single End
Galactic quickie # 15

Uploading Data using FTP
Galactic quickie # 17

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or [your own instance](#), you can perform, reproduce, and share complete analyses. The [Galaxy team](#) is a part of [BX at Penn State](#), and the [Biology](#) and [Mathematics and Computer Science](#) departments at [Emory University](#). The [Galaxy Project](#) is supported in part by [NSF](#), [NHGRI](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Emory University](#).

Galaxy build: \$Rev 8778:7c3df0bcb22\$

galaxyproject

genetics_blog A ChIP-Seq benchmark dataset nar.oxfordjournals.org/content/39/4/e... #bioinformatics #genomics

History

Unnamed history

166.1 MB

4: test peaks reformat.bed

3: test peaks reformat.bed

2: EBI SRA: ERR022249 File: [ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR022/ERR02249/ERR022249.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR022/ERR022249/ERR022249.fastq.gz)

1: EBI SRA: ERR022246 File: ftp://ftp.sra.ebi.ac.uk/vol1/ERA015/ERA015133/fastq/R1_S4_I1_25C_Bur2.fastq
539,825 sequences
format: fastq, database: ?

@HWI-ST-EAS610:4:1:1:1615#0/1

NNNNNGTTCTCNCNTCCACCAATCTTNNNNNNNN

+

BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB

@HWI-ST-EAS610:4:1:2:1978#0/1

CATTAACAACACAGAACTTCCACGCCGACANAACN

Sort by peak height

The screenshot shows the Galaxy web interface with the 'Sort' tool (version 1.0.1) selected. The tool configuration is as follows:

- Sort Dataset: 5: Sort on data 4
- on column: c5
- with flavor: Numerical sort
- everything in: Descending order
- Column selections: Add new Column selection

The 'Execute' button is visible. A red circle highlights the 'on column: c5' dropdown menu. Another red circle highlights the 'History' panel on the right, which shows the execution history of the tool. The history entry for '5: Sort on data 4' is highlighted, showing the output format and database information. Below the history entry, a table of sorted data is displayed, with columns for Chromosome, Start, End, and Peak Name. The table shows several rows of data, including chrIII, chrII, and chrI, with peak names like MACS_peak and test_peaks_reformat.bed. A red circle highlights the table header and the first few rows of data.

Tools

search tools

Get Data

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Convert Formats

FASTA 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
- Filter on ambiguities in polymorphism datasets

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

Extract Features

Fetch Sequences

Fetch Alignments

Sort (version 1.0.1)

Sort Dataset: 5: Sort on data 4

on column: c5

with flavor: Numerical sort

everything in: Descending order

Column selections: Add new Column selection

Execute

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

Syntax

This tool sorts the dataset on any number of columns in either ascending or descending order.

Numerical sort orders numbers by their magnitude, ignores all characters besides numbers, and evaluates a string of numbers to the value they signify.

Alphabetical sort is a phonebook type sort based on the conventional order of letters in an alphabet. Each nth letter is compared with the nth letter of other words in the list, starting at the first letter of each word and advancing to the second, third, fourth, and so on, until the order is established. Therefore, in an alphabetical sort, 2 comes after 100 (1 < 2).

Examples

The list of numbers 4,17,3,5 collates to 3,4,5,17 by numerical sorting, while it collates to 17,3,4,5 by alphabetical sorting.

Sorting the following:

Q d 7 II jhu 45

History

Unnamed history
166.1 MB

5: Sort on data 4

736 regions
format: interval, database: ce6
display at UCSC [main](#)
view in [GeneTrack](#)
display at Ensembl [Current](#)
display at GBrowse [WormBase](#)
[current](#)

1. Chrom	2. Start	3. End	4
chrIII	1416362	1421778	MACS_peak
chrIII	1200401	1204272	MACS_peak
chrIII	13461250	13467052	MACS_peak
chrIII	12904276	12906324	MACS_peak
chrIII	11919033	11922177	MACS_peak
chrIII	2784900	2788182	MACS_peak

4:
[test_peaks_reformat.bed](#)

3:
[test_peaks_reformat.bed](#)

2: EBI SRA: ERR022249
File: [ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR022/ERR022249/ERR022249.fastq.gz](#)

1: EBI SRA: ERR022246
File: [ftp://ftp.sra.ebi.ac.uk](#)

Select top 25 regions

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. The 'Tools' panel on the left lists various data manipulation tools, with 'Select first lines from a dataset' circled in red. The main workspace displays the 'Select first (version 1.0.0)' tool configuration. The 'Select first:' field is set to '25' and 'lines', and the 'from:' dropdown is set to '6: Select first on data 5'. The 'Execute' button is visible. Below the configuration, the 'What it does' section explains that the tool outputs a specified number of lines from the beginning of a dataset. The 'Example' section shows a table of genomic data and the resulting output after selecting the first 2 lines. The 'History' panel on the right shows a list of jobs, with the top job '6: Select first on data 5' circled in red. This job's details are expanded, showing the output format and links to view the data in UCSC, GeneTrack, Ensembl, and GBrowse. Below this, other jobs are listed, including '5: Sort on data 4' and '4: test peaks reformat.bed'.

Tools

- Compute an expression on every row
- Concatenate datasets tail-to-head
- Condense consecutive characters
- Convert delimiters to TAB
- Merge Columns together
- Create single interval as a new dataset
- Cut columns from a table
- 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

Convert Formats

FASTA manipulation

Filter and Sort

Join, Subtract and Group

Select first (version 1.0.0)

Select first:
25
lines

from:
6: Select first on data 5

Execute

What it does

This tool outputs specified number of lines from the **beginning** of a dataset

Example

Selecting 2 lines from this:

chr7	56632	56652	D17003_CTCF_R6	310	+
chr7	56736	56756	D17003_CTCF_R7	354	+
chr7	56761	56781	D17003_CTCF_R4	220	+
chr7	56772	56792	D17003_CTCF_R7	372	+
chr7	56775	56795	D17003_CTCF_R4	207	+

will produce:

chr7	56632	56652	D17003_CTCF_R6	310	+
chr7	56736	56756	D17003_CTCF_R7	354	+

History

Unnamed history
166.1 MB

6: Select first on data 5
25 regions
format: interval, database: ce6
display at UCSC [main](#)
view in GeneTrack
display at Ensembl [Current](#)
display at GBrowse [WormBase current](#)

1. Chrom	2. Start	3. End	4
chrIII	1416362	1421778	MACS_peak
chrIII	1200401	1204272	MACS_peak
chrIII	13461250	13467052	MACS_peak
chrIII	12904276	12906324	MACS_peak
chrIII	11919033	11922177	MACS_peak
chrIII	2784900	2788182	MACS_peak

5: Sort on data 4

4: test peaks reformat.bed

3: test peaks reformat.bed

2: EBI SRA: ERR022249
File: <ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR022/ERR02249/ERR022249.fastq.gz>

Get genomic sequences

The screenshot displays the Galaxy web interface with a workflow titled "Extract Genomic DNA (version 2.2.2)".

Tools Panel (Left): A list of tool categories is shown, with "Fetch Sequences" circled in red. Under this category, the tool "Extract Genomic DNA using coordinates from assembled/unassembled genomes" is also circled in red.

Tool Configuration (Center):

- Fetch sequences for intervals in:** 6: Select first on data 5
- Interpret features when possible:** No
- Source for Genomic Data:** Locally cached
- Output data type:** FASTA
- Execute** button

Warnings (Center):

- ⚠ This tool requires interval or gff (special tabular formatted data). If your data is not TAB delimited, first use *Text Manipulation* → *Convert*.
- ⚠ Make sure that the genome build is specified for the dataset from which you are extracting sequences (click the pencil icon in the history item if it is not specified).
- ⚠ All of the following will cause a line from the input dataset to be skipped and a warning generated. The number of warnings and skipped lines is documented in the resulting history item.
 - Any lines that do not contain at least 3 columns, a chromosome and numerical start and end coordinates.
 - Sequences that fall outside of the range of a line's start and end coordinates.
 - Chromosome, start or end coordinates that are invalid for the specified build.
 - Any lines whose data columns are not separated by a TAB character (other white-space characters are invalid).

What it does (Center):

This tool uses coordinate, strand, and build information to fetch genomic DNAs in FASTA or interval format. If strand is not defined, the default value is "+".

History Panel (Right): A list of workflow steps is shown, with step 7 circled in red:

- 7: **Extract Genomic DNA on data 6**
 - 25 sequences
 - format: fasta, database: ce6
 - Preview of FASTA output:

```
>ce6_chrIII_1416362_1421778_+
GCTTTTCTGTAAATTTAGAAAAATCGGTGTTT
TATGGAAAAATCAGTTTTTCATGACATTCACATAAAA
TTGCGTTTTACATGGAAAAATTACTGAAAAATTGCAT
AAATGCTAAAAATTTGGTTTTTTTTTCAGTGAAAAAT
GGTTTTTCATGAAAAATGCTGAAAAATCGGTGTTT
```
- 6: Select first on data 5
- 5: Sort on data 4
- 4: test_peaks_reformat.bed
- 3: test_peaks_reformat.bed
- 2: EBI SRA: ERR022249 File: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR022/ERR022249/ERR022249.fastq.gz
- 1: EBI SRA: ERR022246 File: ftp://ftp.sra.ebi.ac.uk/vol1/ERA015/ERA015133/fastq/R1_S4_I1_25C_Bur2.fastq

Use Galaxy output as input to find motif

MEME Suite Menu

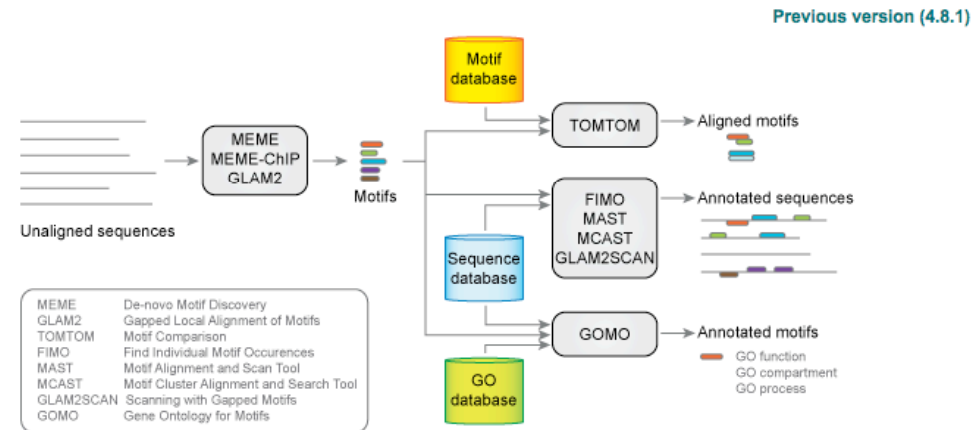
- Submit A Job
 - Discover New Motifs Using MEME
 - Discover Motifs in Large Data Sets Using MEME-ChIP
 - Discover New Gapped Motifs Using GLAM2
 - Compare Motifs Using Tomtom
 - Associate Motifs with GO-terms Using GOMO
- Documentation
- Downloads
- User Support
- Alternate Servers
- Authors
- Citing

The MEME Suite

Motif-based sequence analysis tools

<http://meme.nbcr.net/meme/>

On 19 March, from 5:00PM to 8:00PM PST, the SDSC datacenter will experience partial networking outages due to software upgrade to edge switches. All networking services will be unavailable intermittently while this maintenance is being performed.



The MEME Suite allows you to:

- discover motifs using [MEME](#), [DREME](#) (DNA only) or [GLAM2](#) on groups of related DNA or protein sequences,
- search sequence databases with motifs using [MAST](#), [FIMO](#), [MCAST](#) or [GLAM2SCAN](#),
- compare a motif to all motifs in a database of motifs,
- associate motifs with Gene Ontology terms via their putative target genes, and
- analyse motif enrichment using [SpaMo](#) or [CentriMo](#).

To submit a query, click on one of the logos below or select "Submit A Job" from the menu at the left.

CISTROME analysis pipeline

<http://cistrome.org/ap/>

Same interface as galaxy

The screenshot displays the Galaxy / Cistrome web interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Lab', 'Visualization', 'Help', and 'User'. The main content area is divided into three panels:

- Tools Panel (Left):** Lists various tools. The tool 'CEAS: Enrichment on chromosome and annotation' is highlighted with a red circle. Other tools include 'Multiple wiggle files correlation in given regions', 'Two wiggle file correlation in union regions', 'Venn Diagram', 'SitePro: Aggregation plot tool for signal profiling', 'GCA: Gene centered annotation', and 'peak2gene: Peak Center Annotation'.
- Configuration Panel (Center):** Shows the settings for the 'wig / bigwig file' tool. The 'BED file(maximum 100000 lines):' is set to '649: (as bed) test_peaks_reformat.bed'. Other settings include 'Span: 3000', 'Profiling resolution: 50', 'Promoter/downstream lower-interval: 1000', 'Promoter/downstream middle-interval: 2000', 'Promoter/downstream upper-interval: 3000', 'Bi-Promoter lower range: 2500', 'Bi-Promoter upper range: 5000', 'Relative distance: 3000', 'Image Type: PNG format', and 'Specify gene list in the signal profiling: No'. An 'Execute' button is at the bottom.
- History Panel (Right):** Shows a list of previous jobs. The top job is 'RAD-51 for Paper' (5.1 GB). Below it are several jobs related to '649: test_peaks_reformat.bed', including '648: MACS diagnosis report on RAD51 Rep2 input.bed', '647: MACS job log on RAD51 Rep2 input.bed', '646: MACS wiggle on RAD51 Rep2 input.bed', '645: MACS xls on RAD51 Rep2 input.bed', '644: MACS summits on RAD51 Rep2 input.bed', '643: MACS peaks on RAD51 Rep2 input.bed', '642: MACS diagnosis report on RAD51 Rep2 IP.bed', '641: MACS job log on RAD51 Rep2 IP.bed', and '640: MACS wiggle on RAD51 Rep2 IP.bed' (~10,000,000 lines, format: wig, database: ce4).

Can save output files....

Galaxy / Cistrome Analyze Data Workflow Shared Data Lab Visualization Help User

tools

Multiple wiggle files correlation in given regions Calculate the correlation coefficient on the genome scale using one bed file and multiple wiggle / bigwig files

Two wiggle file correlation in union regions Calculate the correlation coefficient of two wiggle / bigwig files in the union regions from two bed files

Venn Diagram Given 2 or 3 intervals, generate a venn diagram of their intersections

ASSOCIATION STUDY

CEAS: Enrichment on chromosome and annotation Annotate the given intervals and scores with genome features such as gene body

SitePro: Aggregation plot tool for signal profiling Draw the score profile near a given interval

GCA: Gene centered annotation Find the nearest interval in the given intervals set for every annotated coding gene

peak2gene: Peak Center Annotation Input a peak file, and it will search each peak on

Genome

Feature	Percentage
Promoter (<=1000 bp)	15.4 %
Promoter (1000-2000 bp)	6.6 %
Promoter (2000-3000 bp)	3.9 %
Downstream (<=1000 bp)	9.3 %
Downstream (1000-2000 bp)	3.9 %
Downstream (2000-3000 bp)	2.1 %
5'UTR	0.7 %
3'UTR	1.6 %
Coding exon	22.2 %
Intron	26.3 %
Distal intergenic	8.0 %

ChIP

Feature	Percentage
Promoter (<=1000 bp)	35.0 %
Promoter (1000-2000 bp)	3.8 %
Promoter (2000-3000 bp)	1.0 %
Downstream (<=1000 bp)	5.8 %
Downstream (1000-2000 bp)	2.2 %
Downstream (2000-3000 bp)	1.3 %
5'UTR	3.0 %
3'UTR	1.5 %
Coding exon	27.3 %
Intron	17.7 %
Distal intergenic	1.4 %

History

RAD-51 for Pap

651: ceas job lo

650: CEAS: Enric
chromosome an
data 649

649:
test peaks refo

648: MACS diag
report on RAD5

647: MACS job l
RAD51 Rep2 in

646: MACS wigg
RAD51 Rep2 in

645: MACS xls o
RAD51 Rep2 in

644: MACS sum
RAD51 Rep2 in

643: MACS peak
RAD51 Rep2 in

642: MACS diag
report on RAD5

641: MACS job l
RAD51 Rep2 IP

Find: tag density Next Previous Highlight all Match case

Quality control

Before embarking on a long analysis journey, make sure you're convinced your experiment worked

- Do replicates correlate well? Correlation, peak overlap etc.
- Do genomic annotations look similar between replicates
- Do controls if possible
- If anything is known about your protein, do your ChIPs agree?

Assessing reproducibility between ChIP replicates

<https://sites.google.com/site/anshulkundaje/projects/idr>

[Projects](#) >

IDR: Reproducibility and automatic thresholding of ChIP-seq data

Contents

- 1 Mailing list
- 2 Summary
- 3 Intuitive explanation of IDR and IDR plots
- 4 CODE for IDR analysis
 - 4.1 IDR CODE README
- 5 IDR pipeline run-through
 - 5.1 CALL PEAKS ON INDIVIDUAL REPLICATES
 - 5.2 CALL PEAKS ON POOLED REPLICATES
 - 5.3 FOR SELF-CONSISTENCY ANALYSIS CALL PEAKS ON PSEUDOREPLICATES OF INDIVIDUAL REPLICATES
 - 5.4 CREATE PSEUDOREPLICATES OF POOLED DATA AND CALL PEAKS
 - 5.5 INPUT TO IDR ANALYSIS
 - 5.6 IDR ANALYSIS ON ORIGINAL REPLICATES
 - 5.7 IDR ANALYSIS ON SELF-PSEUDOREPLICATES
 - 5.8 IDR ANALYSIS ON POOLED-PSEUDOREPLICATES
 - 5.9 GETTING THRESHOLDS TO TRUNCATE PEAK LISTS
 - 5.10 FLAGGING REPLICATES FOR LOW CONSISTENCY
 - 5.11 FINAL SET OF PEAK CALLS


Last Updated: Nov 26, 2012

Mailing list

Please join the IDR mailing list <https://groups.google.com/group/idr-discuss> for FAQs, discussions and updates on software.

Summary

<http://cgrlucb.wikispaces.com/ChIPSeqSpring2013>

ucb·ucsc·ucsf

Wiki Home

Recent Changes

Pages and Files

Members

Manage Wiki

Spring 2013 Workshops

Fall 2012 Workshops

Spring 2012 Workshops

Fall 2011 Workshops

edit navigation

★ ChIPSeqSpring2013

Edit

0

0

13


...

ChIP-Seq NGS Data Analysis


Instructor: Chitra Kotwaliwale (chitra.kot@gmail.com)

March 18, 2013


Tools:

MACS 

A widely-used, fast, robust ChIP-seq peak-finding algorithm.


SPP 


A ChIP-seq peak calling algorithm, implemented as an R package.


GALAXY 


An open, web-based platform with some useful tools for ChIP-Seq data analysis.


Relevant papers:

[This paper discusses GC and other biases in high-throughput sequencing](#) 

[Paper by Liu and colleagues describing MACS](#) 

[Detailed "protocol" for how to install and use MACS](#) 

[ChIP-Seq guidles outlined by ENCODE and modENCODE](#) 

[Paper by Kharchenko and colleagues describing spp](#) 

Thanks!