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.)

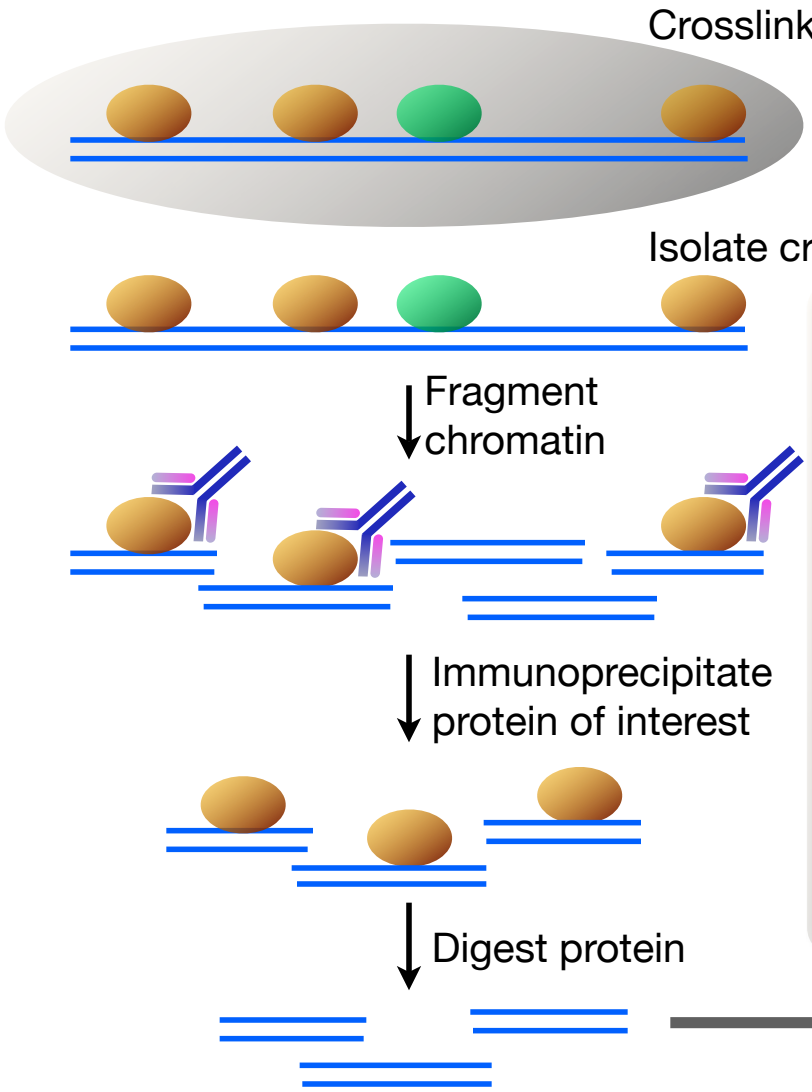
-Estimate peaks, identify DNA motif

-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:

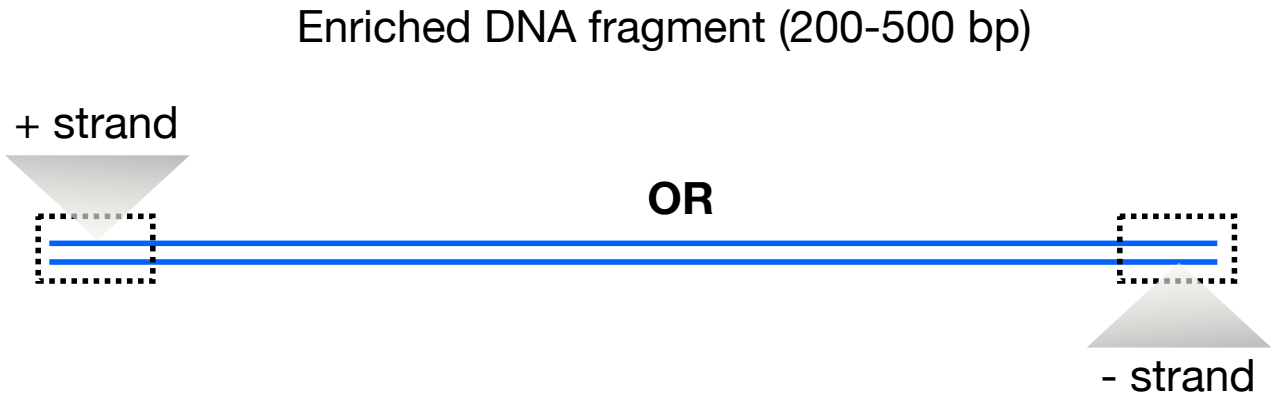
-This is an **enrichment** assay, not a purification

-Usually very small fraction of final DNA corresponds to actual signal

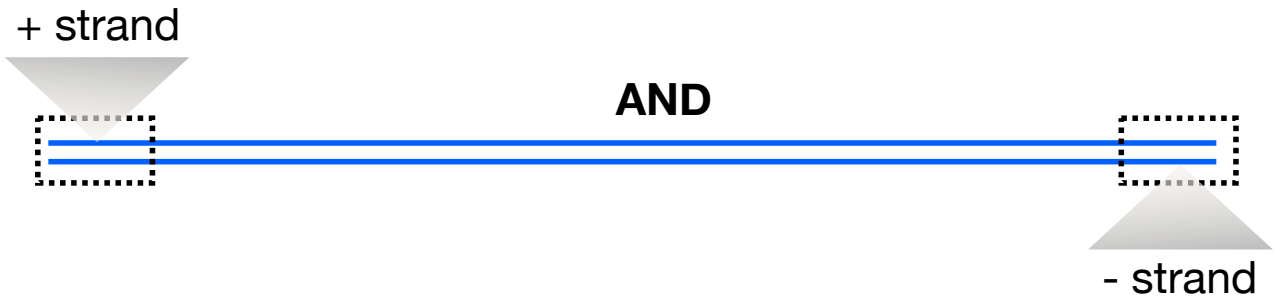
-How much of your final data corresponds to signal depends on **IP efficiency**, **abundance of protein** in a population of cells, **number of sites** in the genome bound by your protein

Sequencing can be done as single or paired end

Single end

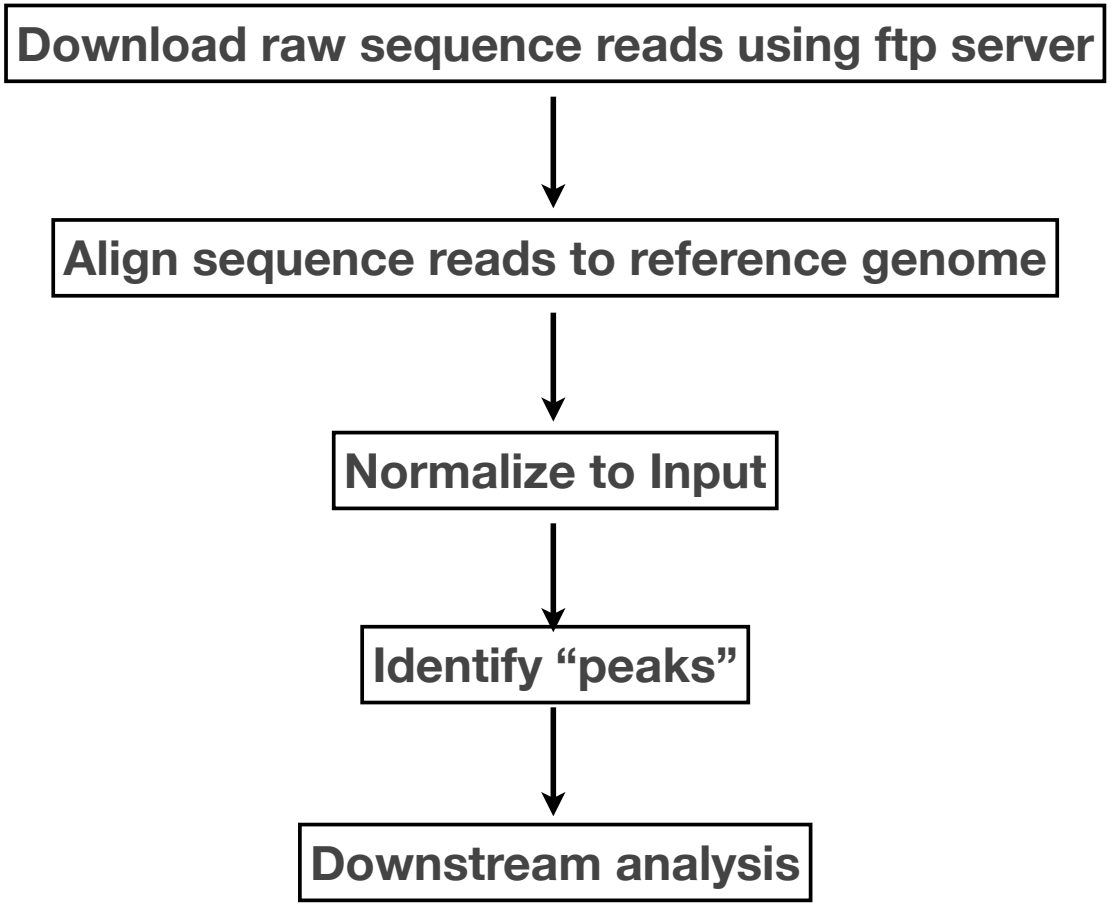


Paired end



For ChIP-Seq, single end sequencing is common

General workflow for analysis of ChIP-Seq data



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:

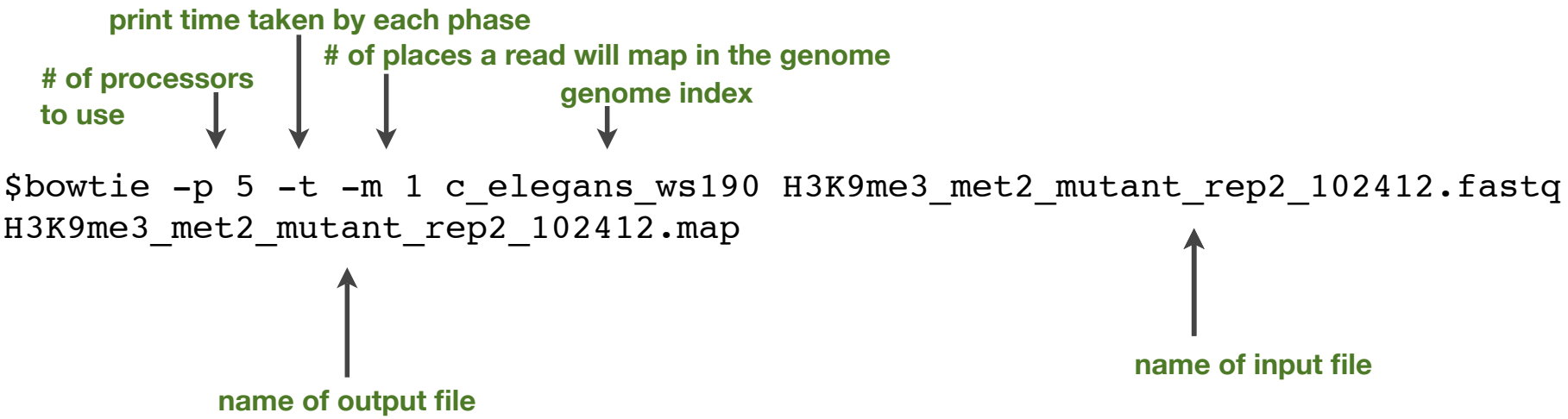
`-q` Input file is fastq format. Also the default setting
`-p` Number of processors to use
`-n` Maximum number of mismatches permitted. This may be 0, 1, 2 or 3. Default is 2.
`-m` Suppress all alignments if more than <int> reportable alignments exist for it.
`-S` Print alignments in SAM format.

Bowtie: An ultrafast memory-efficient short read aligner

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

Berkeley sequencing center provides data in fastq format ----> can be used as input for bowtie

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)

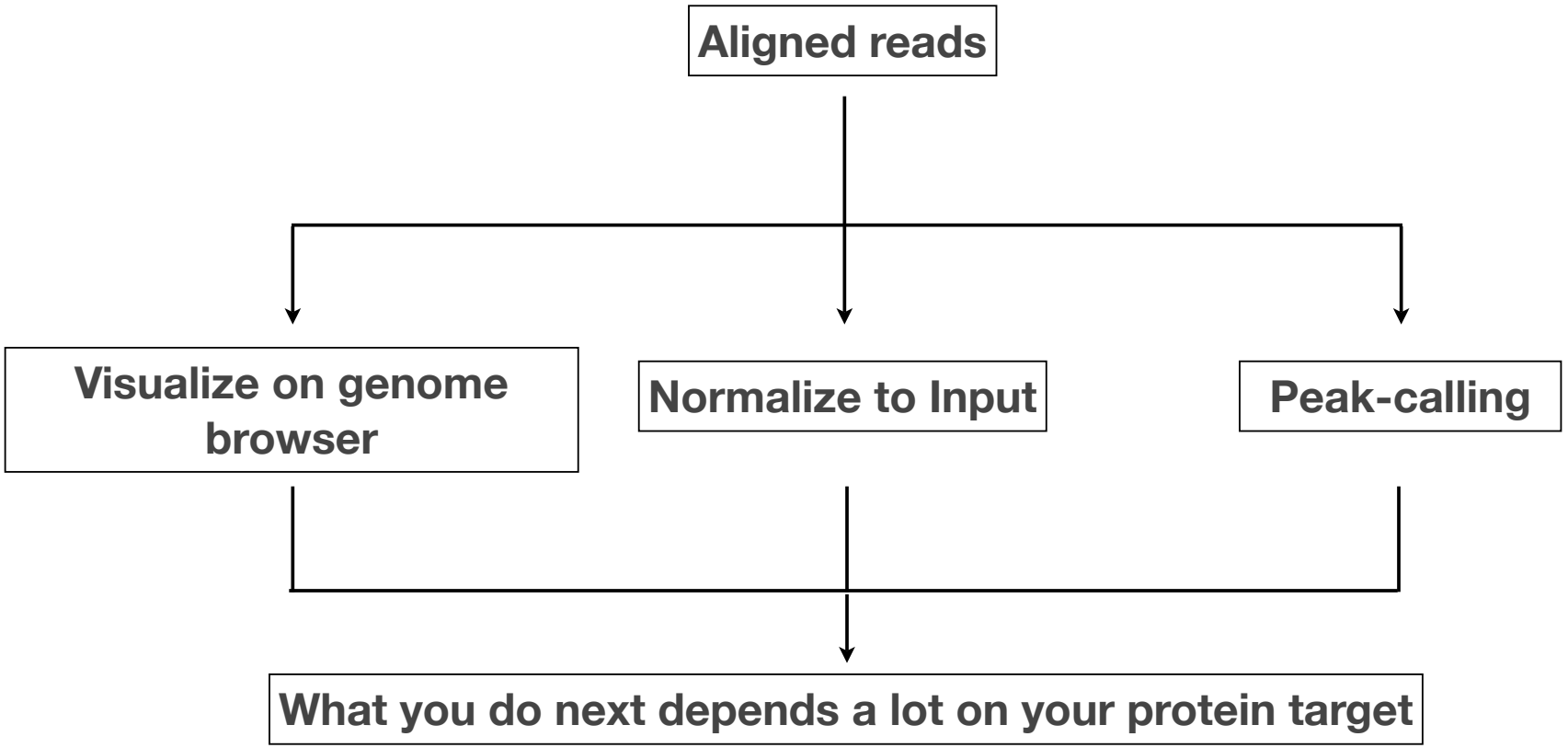
```
HS1:177:d0yyjacxx:6:1101:2688:49638 1:Y:0:ATCACG |+| IV| 16643508|
    CCATCTGAACCATGCGCGTCCAGACGCCTTCTCGGGCACCAAAAGAGCC|  :=+2<2<A=CA;
3<7<=?2A=)7*10):8=3999AA<''5=AAB==57>; |0 |0:T>C,26:A>C
```

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

Narrow
e.g. transcription factors

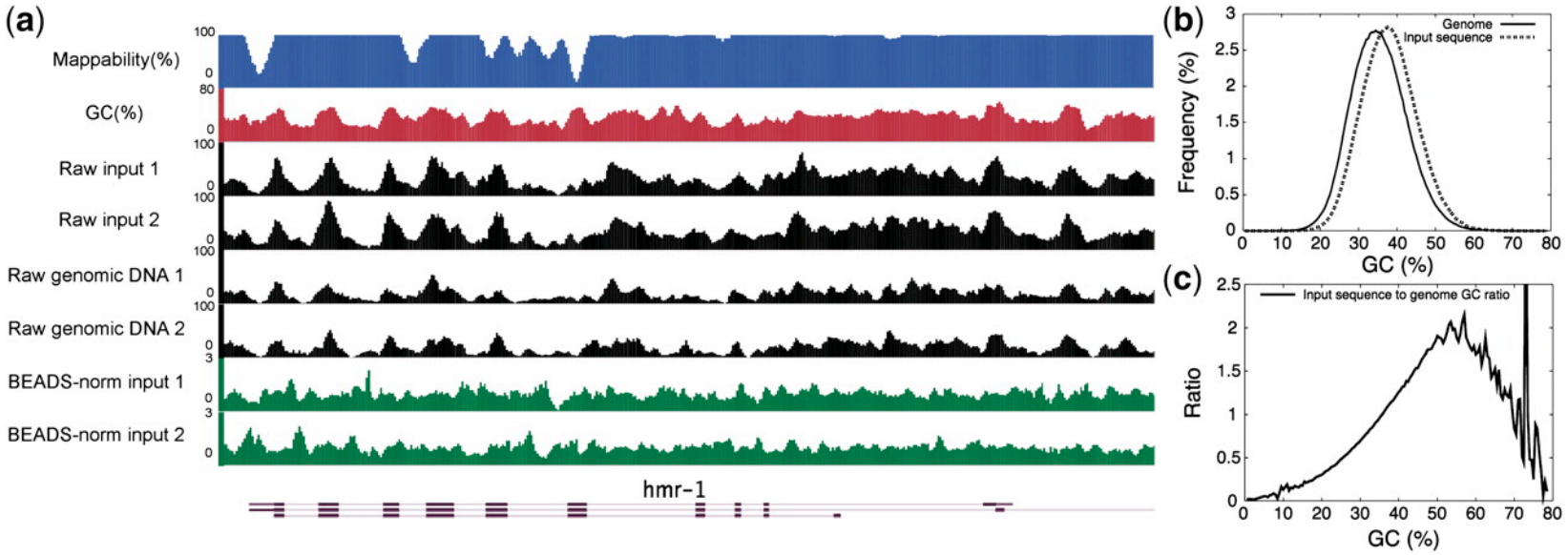
Mid-range
e.g. polymerase

Broad
e.g. histone modifications, structural proteins

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

Good practice to sequence Input

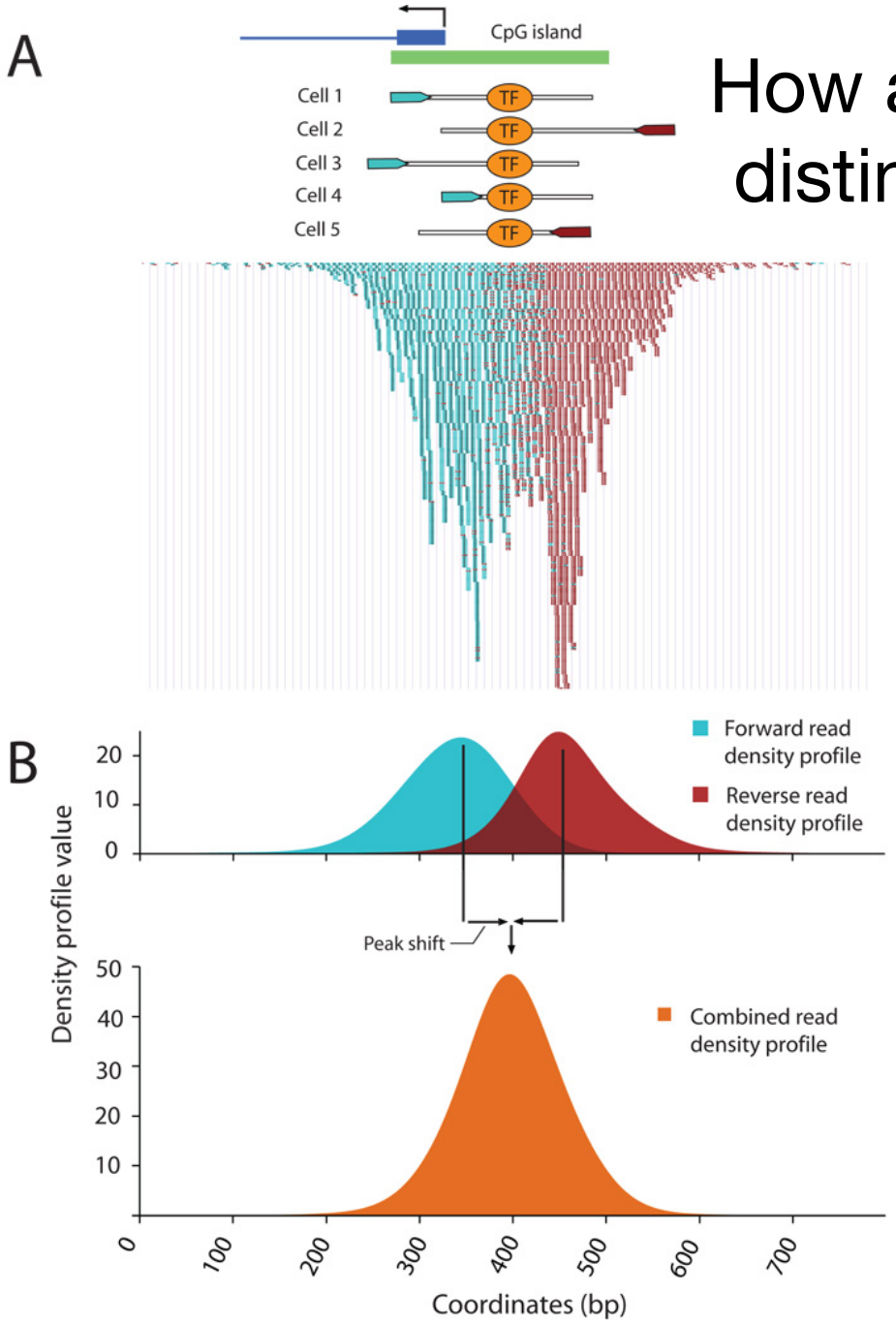
- sample to sample variability (introduced during extract preparation)
- sequencing biases (GC bias)
- variation in sequencing based on biological variation (for e.g. euchromatin generates more reads than heterochromatin)



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

How are ChIP binding sites distinguished from noise?



Valouev et al., 2008

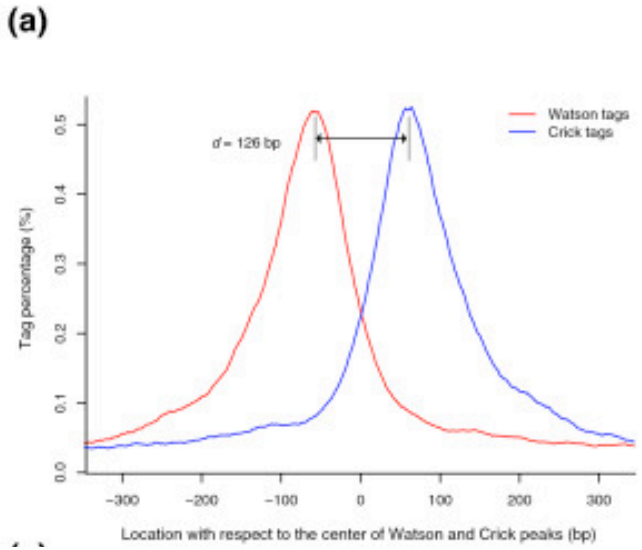
Peak-calling

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/)



- 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.

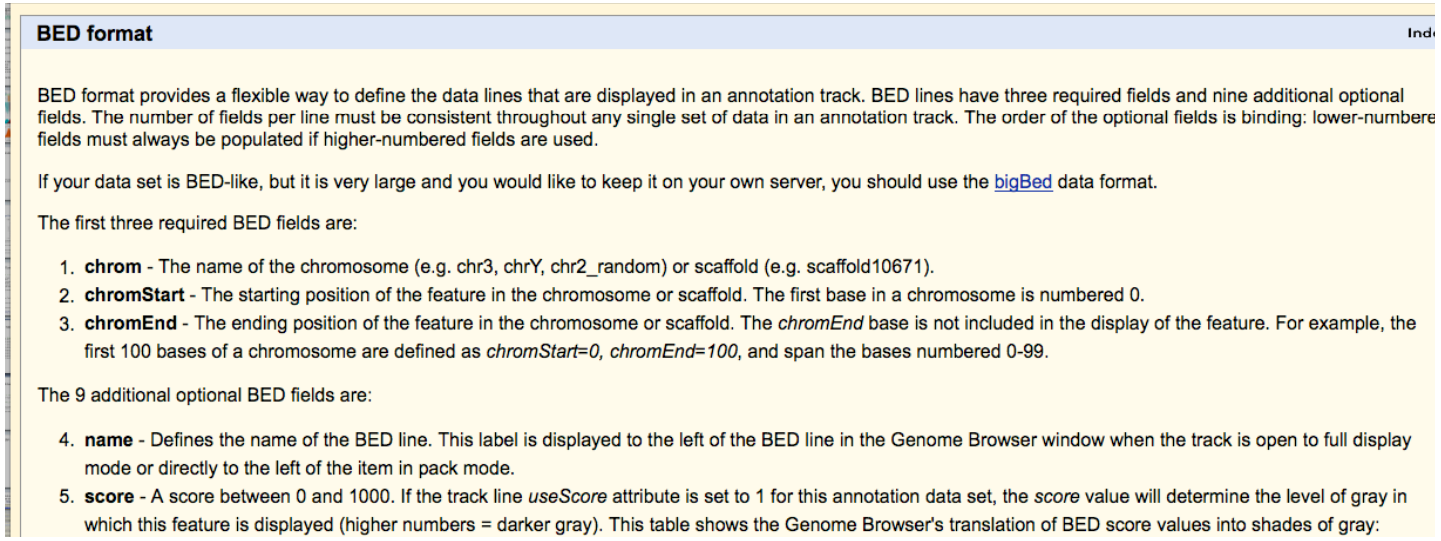
`test_summits.bed`
Peak summits in bed format. Can be loaded on genome browser

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

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-numbere 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

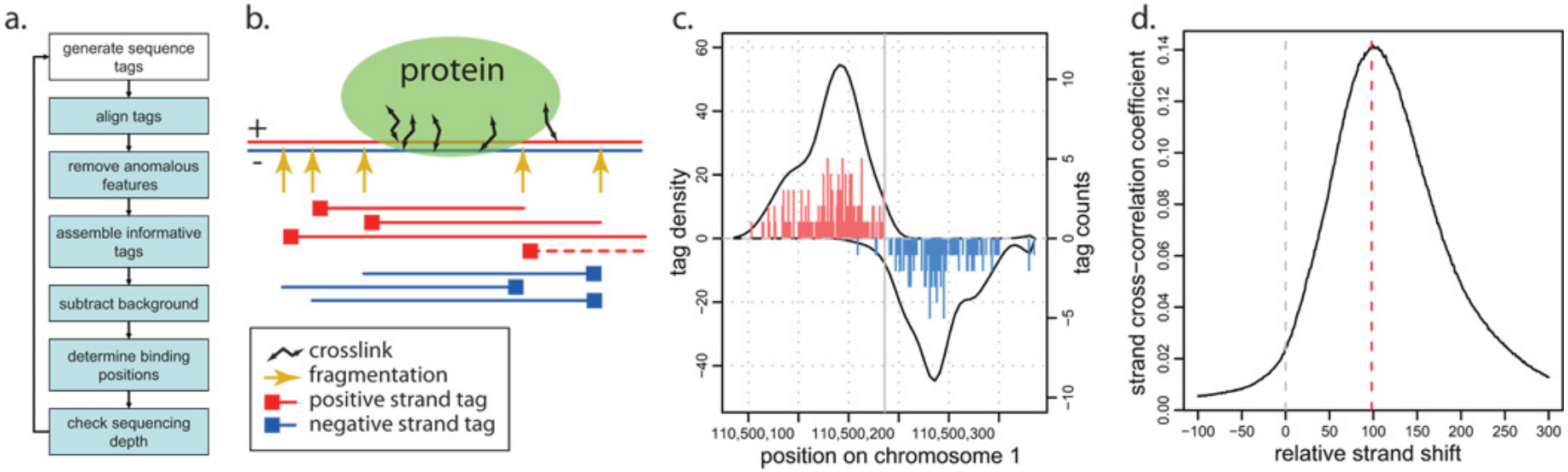
Example of fold-enrichment output file

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



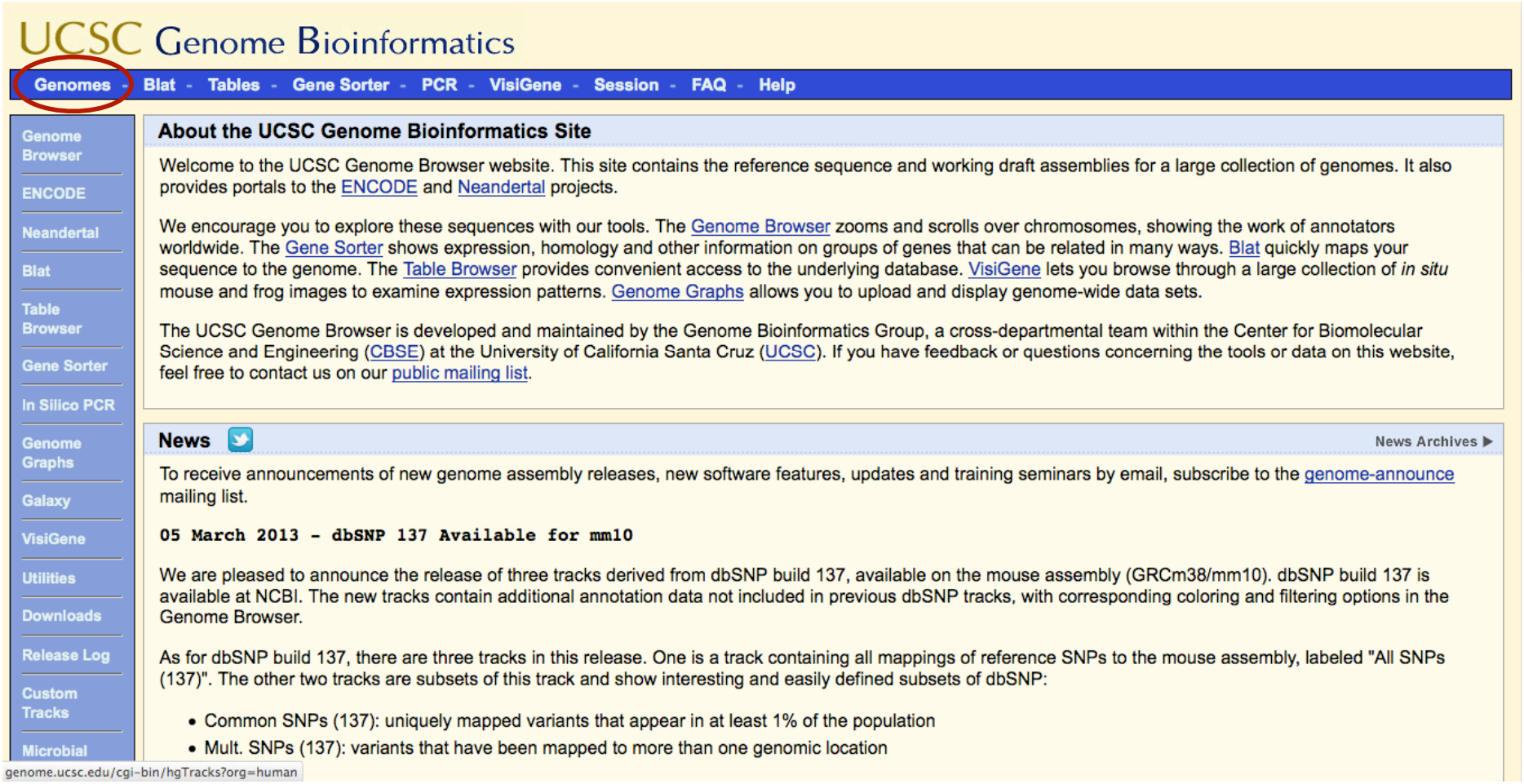
(Outputs a background-subtracted tag density file in addition to peak file after some smoothing)

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/



Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | About Us | Help

C. elegans (*Caenorhabditis elegans*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.

group	genome	assembly	position	search term
Nematode	C. elegans	May 2008 (WS190/ce6)	chrIII:4,263,097-4,350,441	enter position, gene symbol or search terms

submit

Click here to reset the browser user interface settings to their defaults.

track search | manage custom tracks | track hubs | configure tracks and display

Make sure appropriate genome version is selected!

C. elegans Genome Browser – ce6 assembly (sequences)

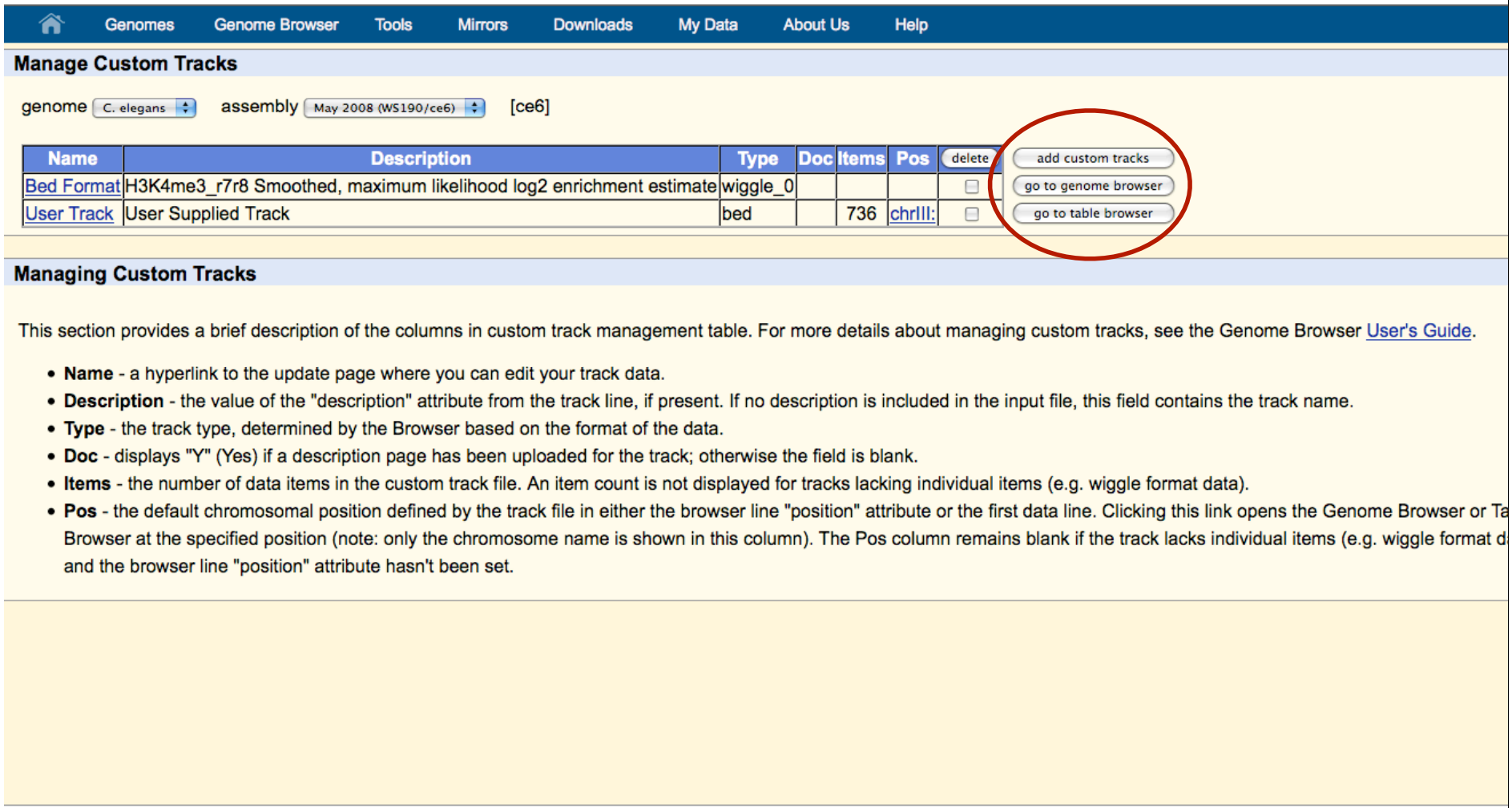
The May 2008 *Caenorhabditis elegans* assembly is based on sequence version WS190 deposited into WormBase as of 28 May 2008. The sequence was produced jointly by the Wellcome Trust Sanger Institute in Hinxton, England and the Genome Sequencing Center at Washington University in St. Louis (WUSTL) School of Medicine, MO, USA. For more information about this assembly, see WS190 in the NCBI Assembly database.

Caenorhabditis elegans
(Photo courtesy of Mark Blaxter)

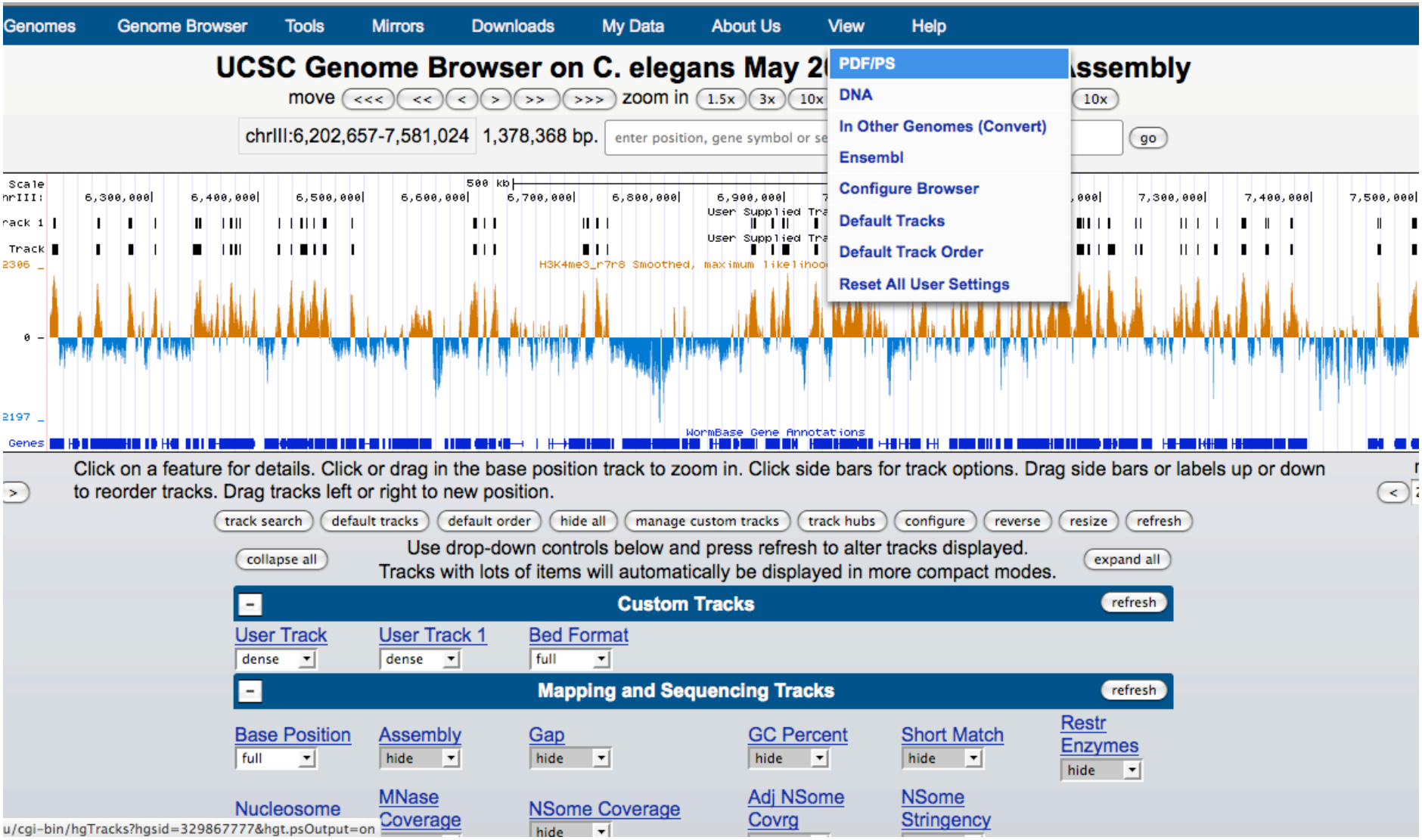
Sample position queries

A genome position can be specified by the accession number of a sequenced genomic region, an mRNA or EST, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the *C. elegans* genome. See the User's Guide for more information.

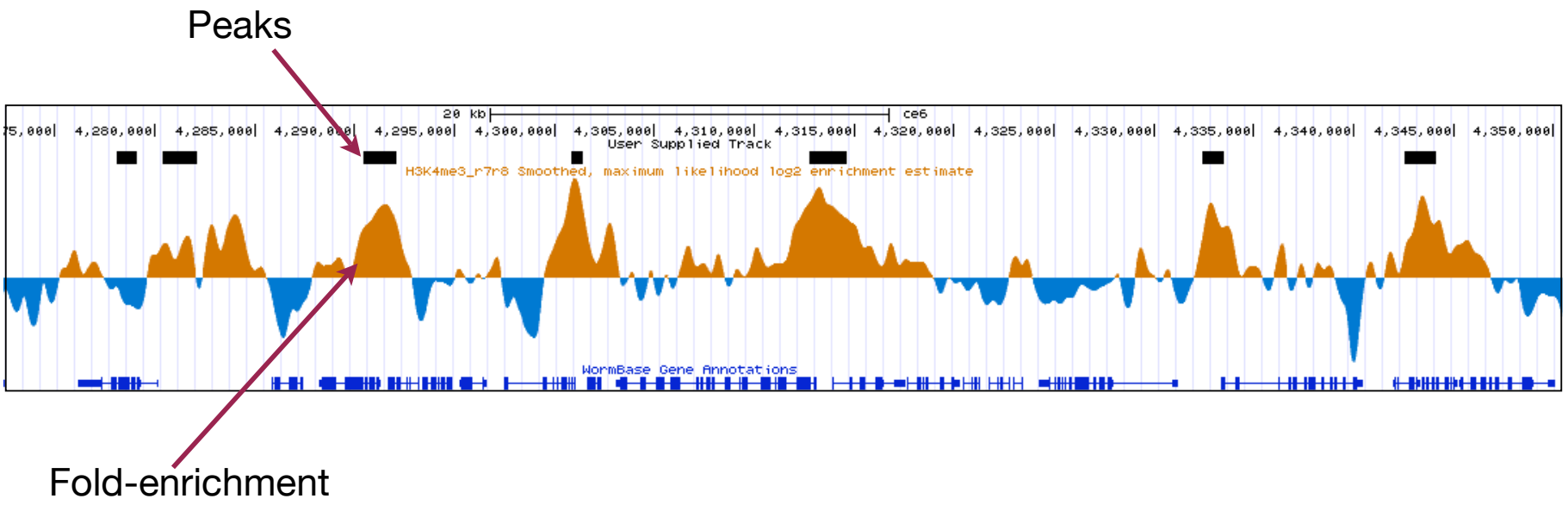
Request:	Genome Browser Response:
chrII	Displays all of chromosome II (valid chr entries include I, II, III, IV, V, M, X)
chrI:1-100000	Displays the first 100,000 bases of chrI
Z81037.1	Displays clone fragment Z81037.1
mec-8	Displays region of genome associated with RefSeq gene mec-8
F46A9.4	Displays the region associated with the Sanger predicted gene F46A9.4
NM_060107	Displays region of genome associated with RefSeq identifier NM_060107


Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
About Us
Help
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
User Track
User Supplied Track
bed
736
chrIII:
add custom tracks
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 Ta
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 d
and the browser line "position" attribute hasn't been set.

Track display and save options



Narrow peaks vs. broad regions



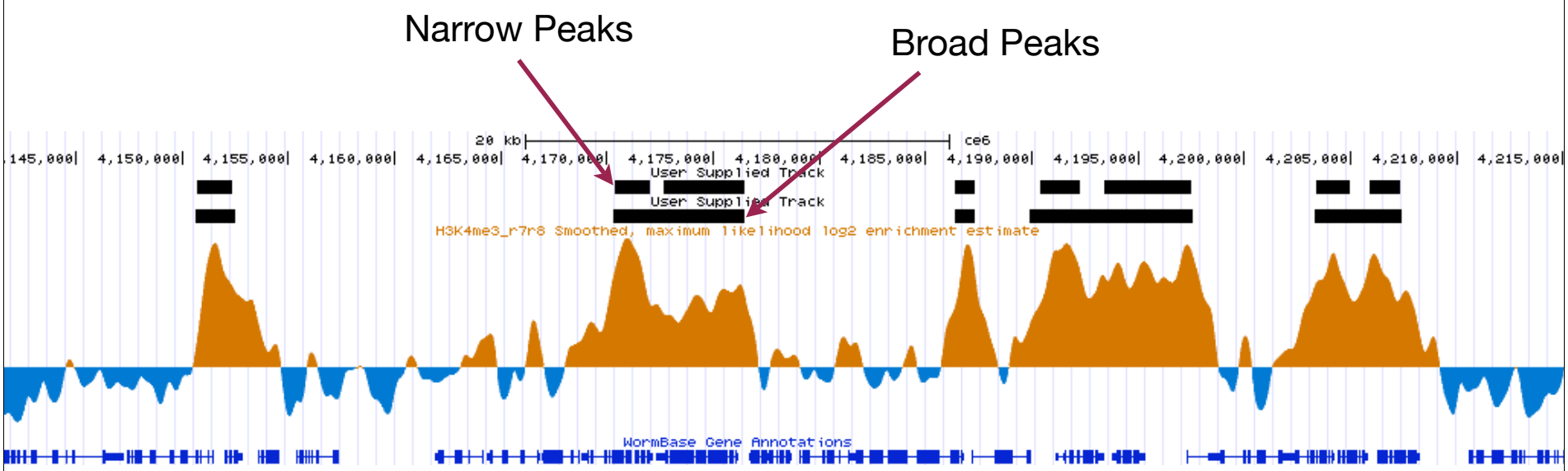
Narrow peaks vs. broad regions

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
20 kb
ce6
4,145,000
4,150,000
4,155,000
4,160,000
4,165,000
4,170,000
4,175,000
4,180,000
4,185,000
4,190,000
4,195,000
4,200,000
4,205,000
4,210,000
4,215,000
User Supplied Track
User Supplied Track
H3K4me3_r7r8 Smoothed, maximum likelihood log2 enrichment estimate
WormBase Gene Annotations

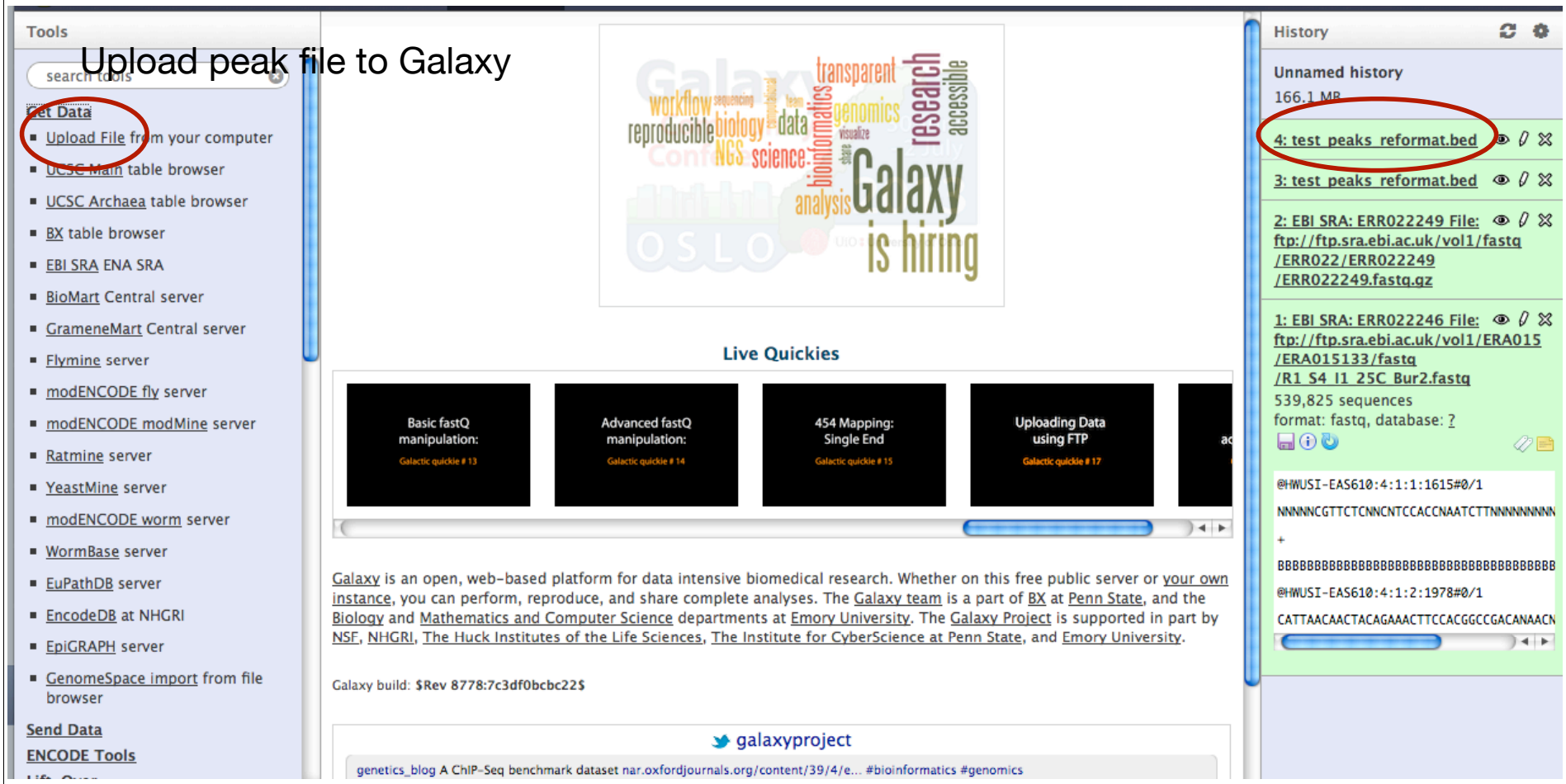
What next?

- 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?

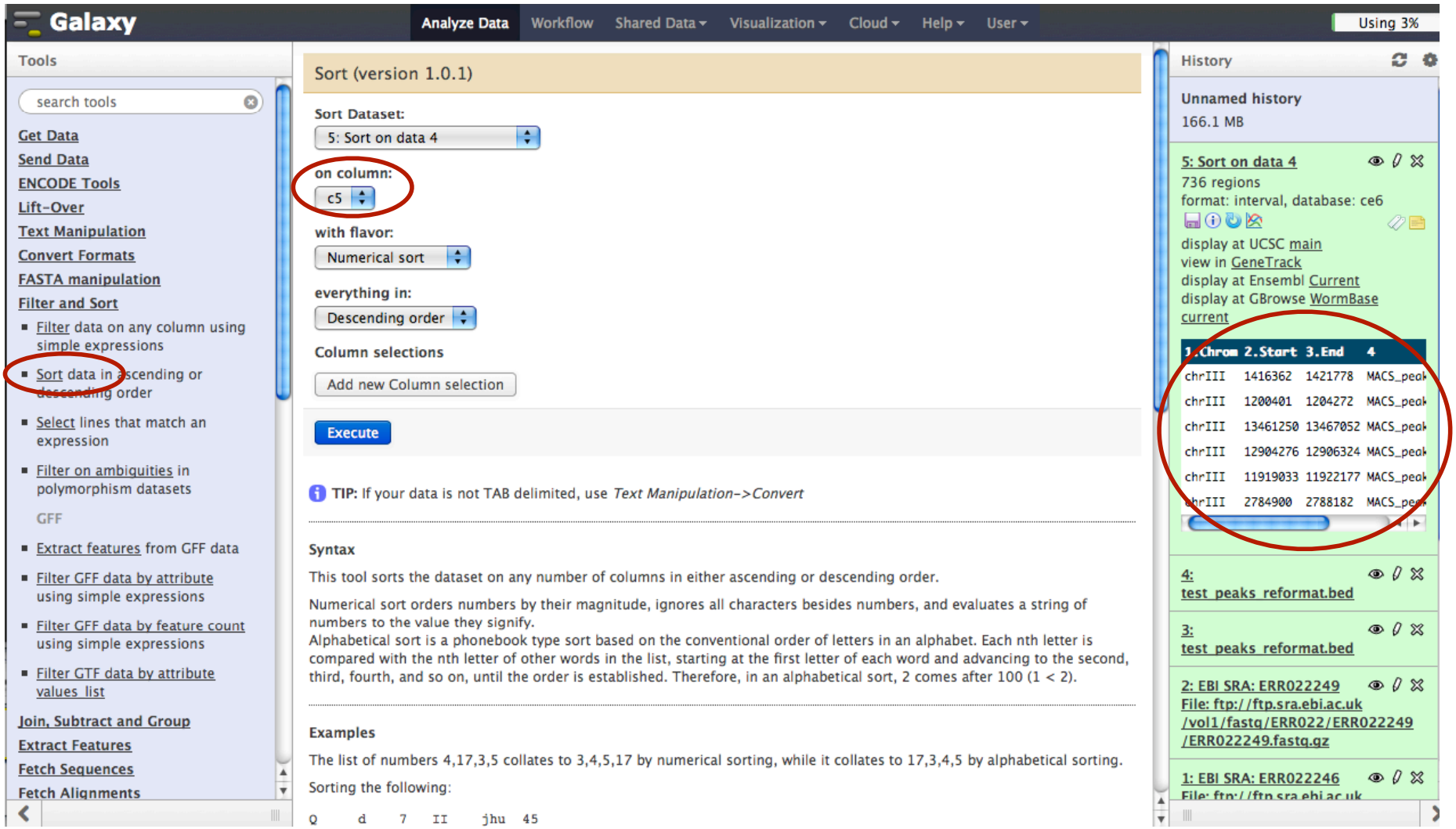
GALAXY

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

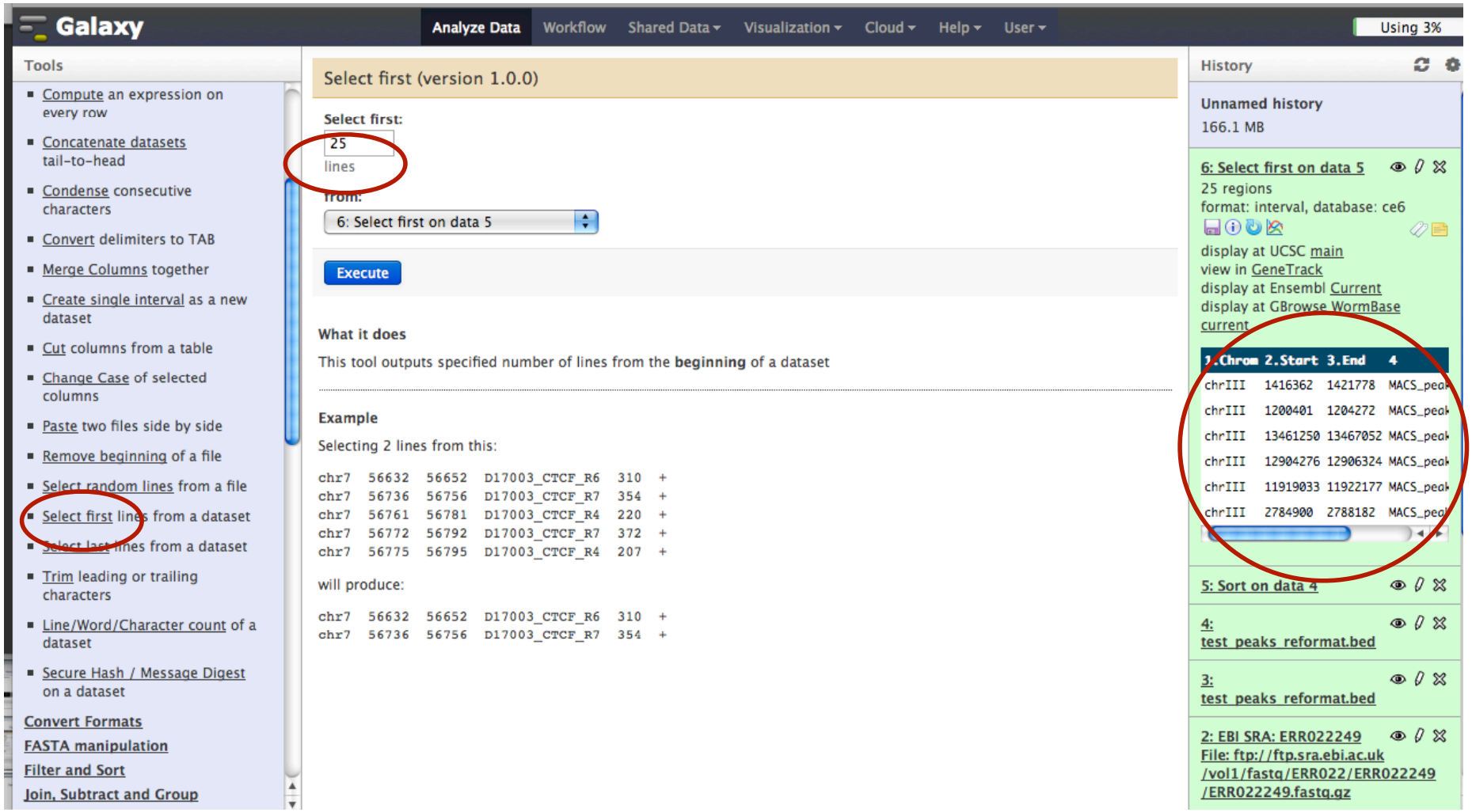
Upload peak file to Galaxy



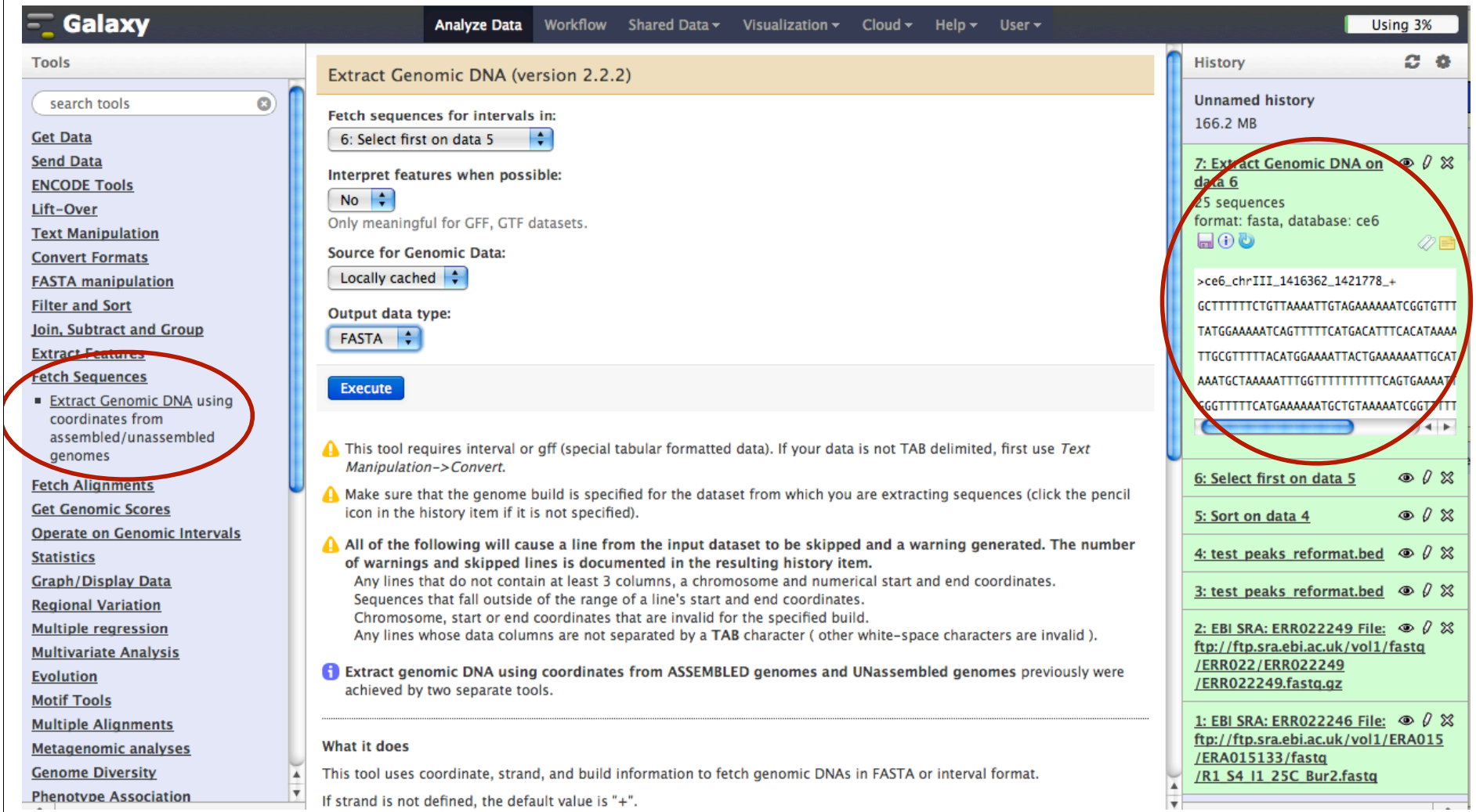
Sort by peak height



Select top 25 regions



Get genomic sequences



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

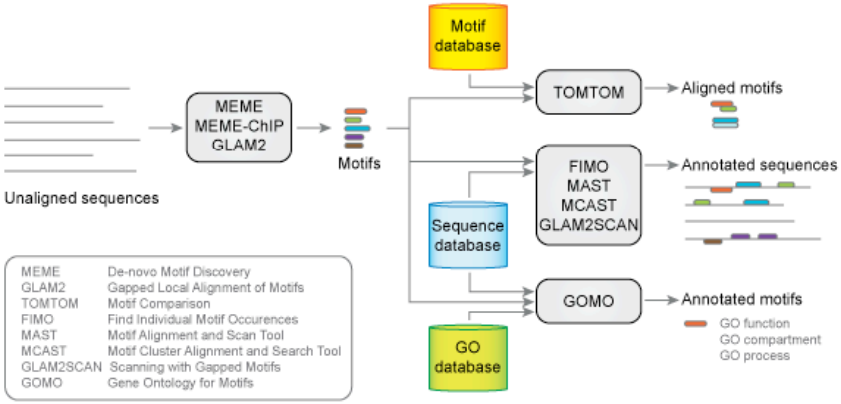
The MEME Suite

http://meme.nbcr.net/meme/

Motif-based sequence analysis tools

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.

Previous version (4.8.1)



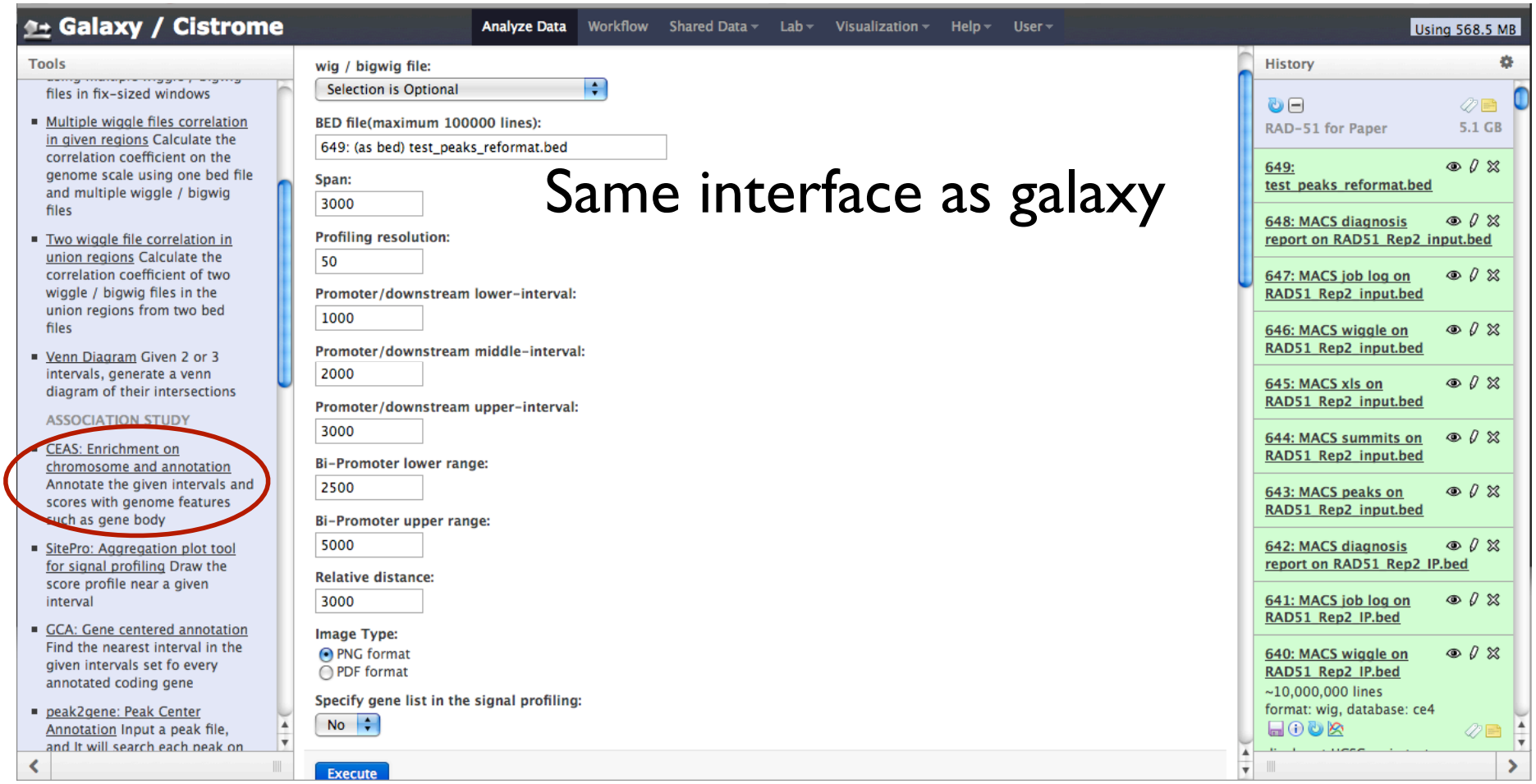
The MEME Suite allows you to:

- 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.

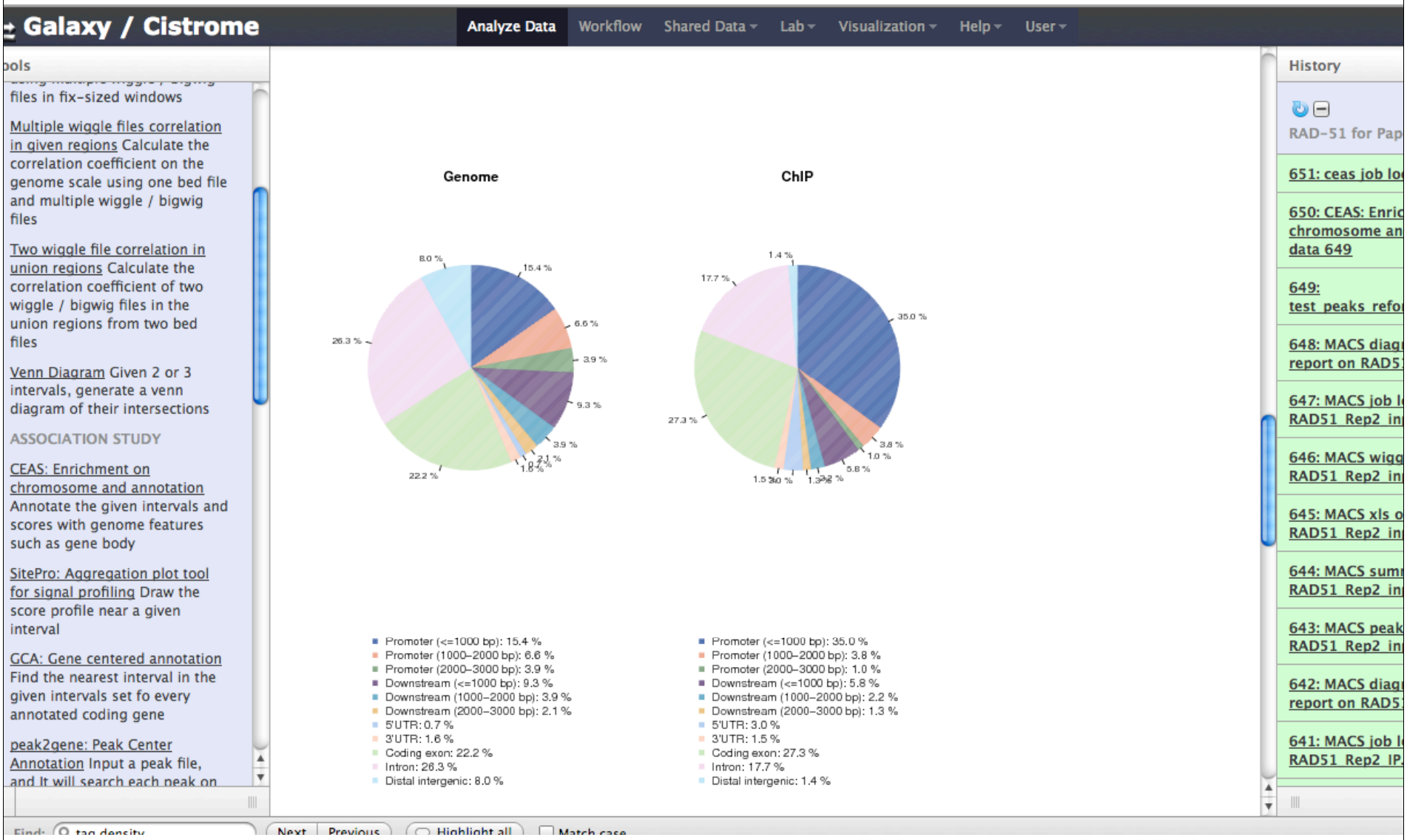
CISTROME analysis pipeline

http://cistrome.org/ap/



Same interface as galaxy

Can save output files....



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

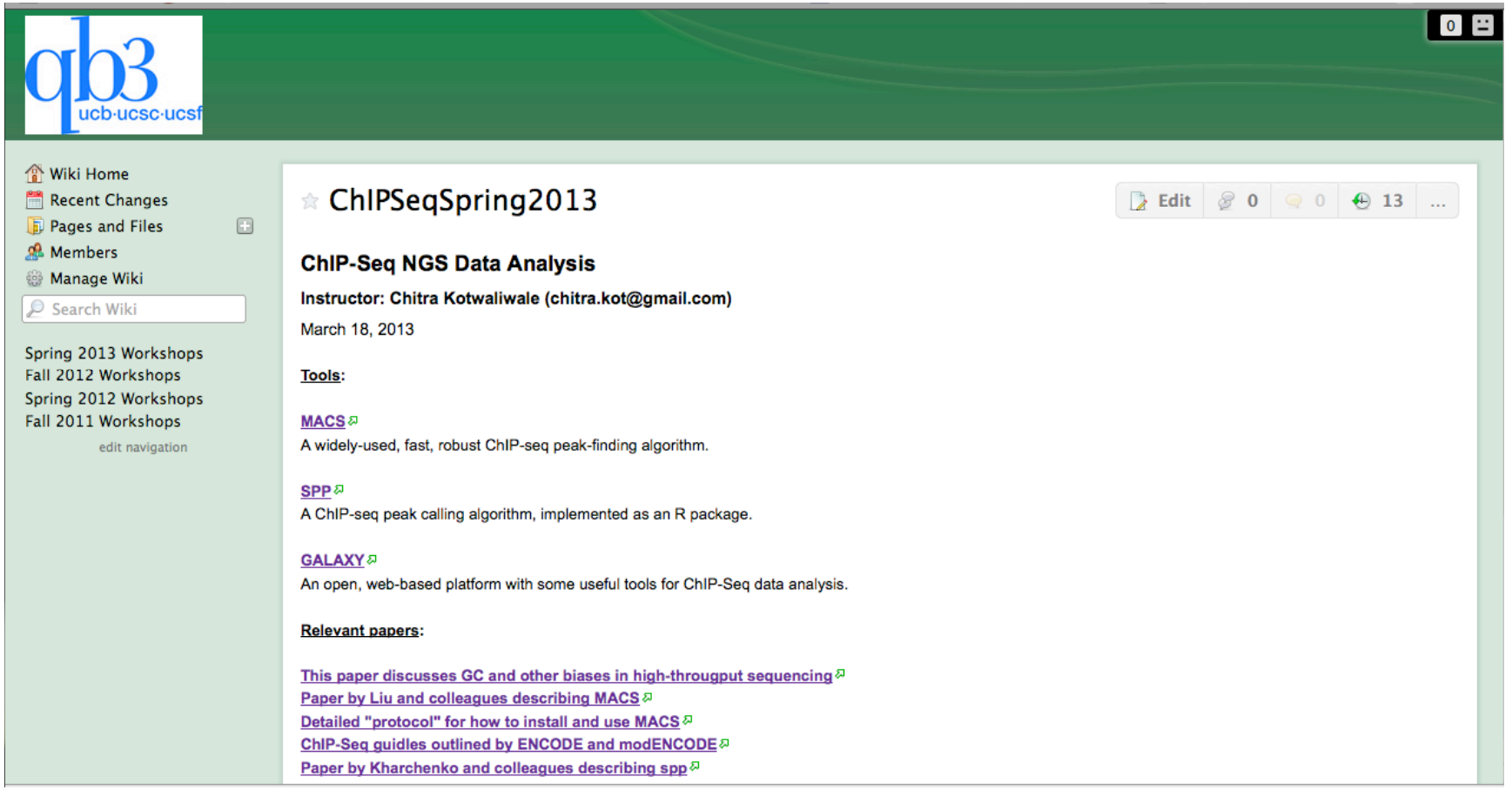
Last Updated: Nov 26, 2012

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



Thanks!