ChIP Seq Analysis

Tommy Kaplan
Computational postdoc in Mike Eisen's lab
CGRL, Oct. 2011

Goals of this workshop

- So you decided to ChIP your protein. 8 months later you get back a pile of sequenced reads. Now what?

- Mapping reads back to genome
- Viewing data
- Finding bound regions
- Initial analysis

What this workshop won't do

- Web analysis tools won't bring you to publication-ready level.
- Programming or a computational collaborator are still needed:
  - More sophisticated enrichmentology
  - Motif analysis
  - Crossing data from time-points/conditions

But...

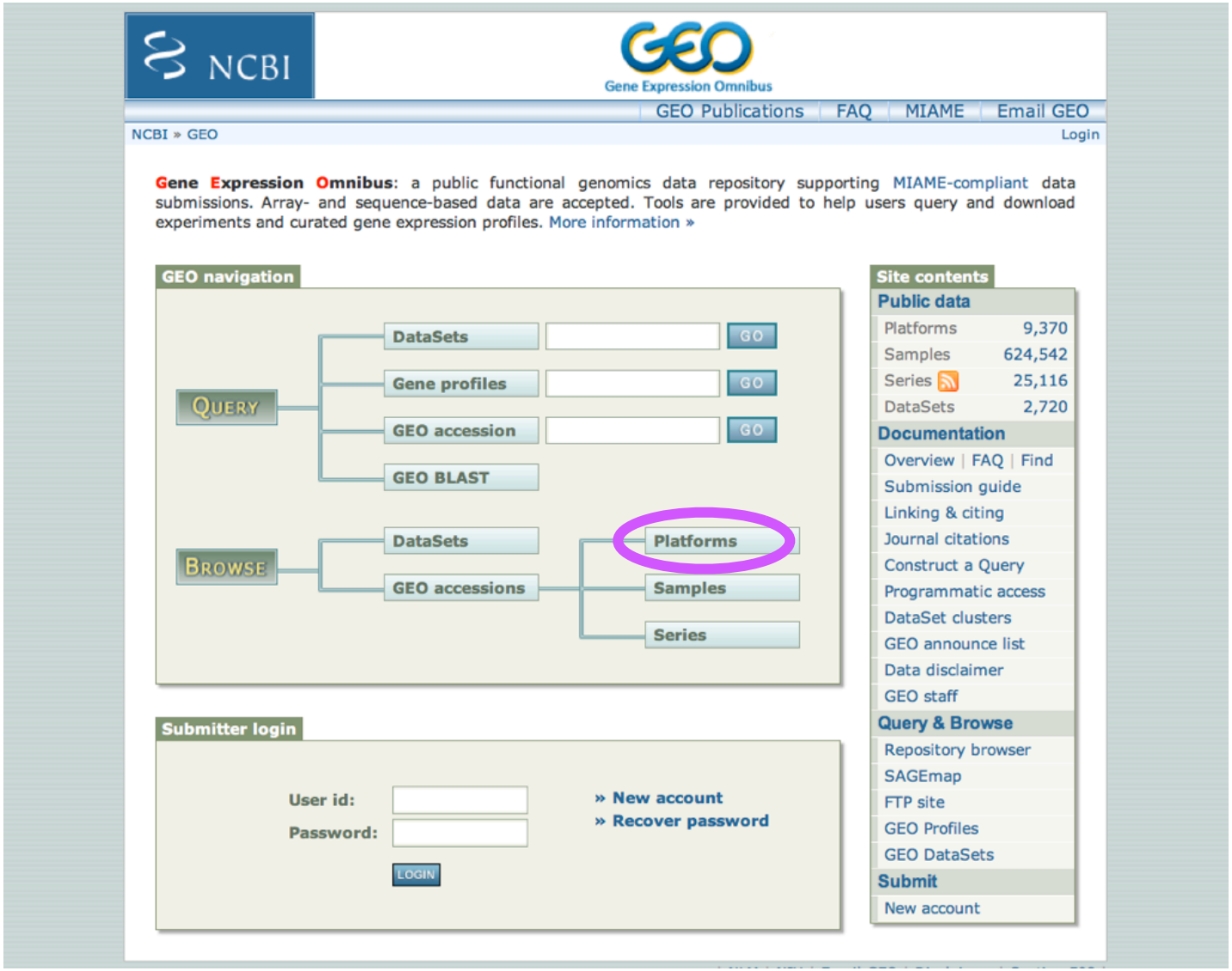
- We will overview tools that will help you find the main story by eyeballing the data

About me

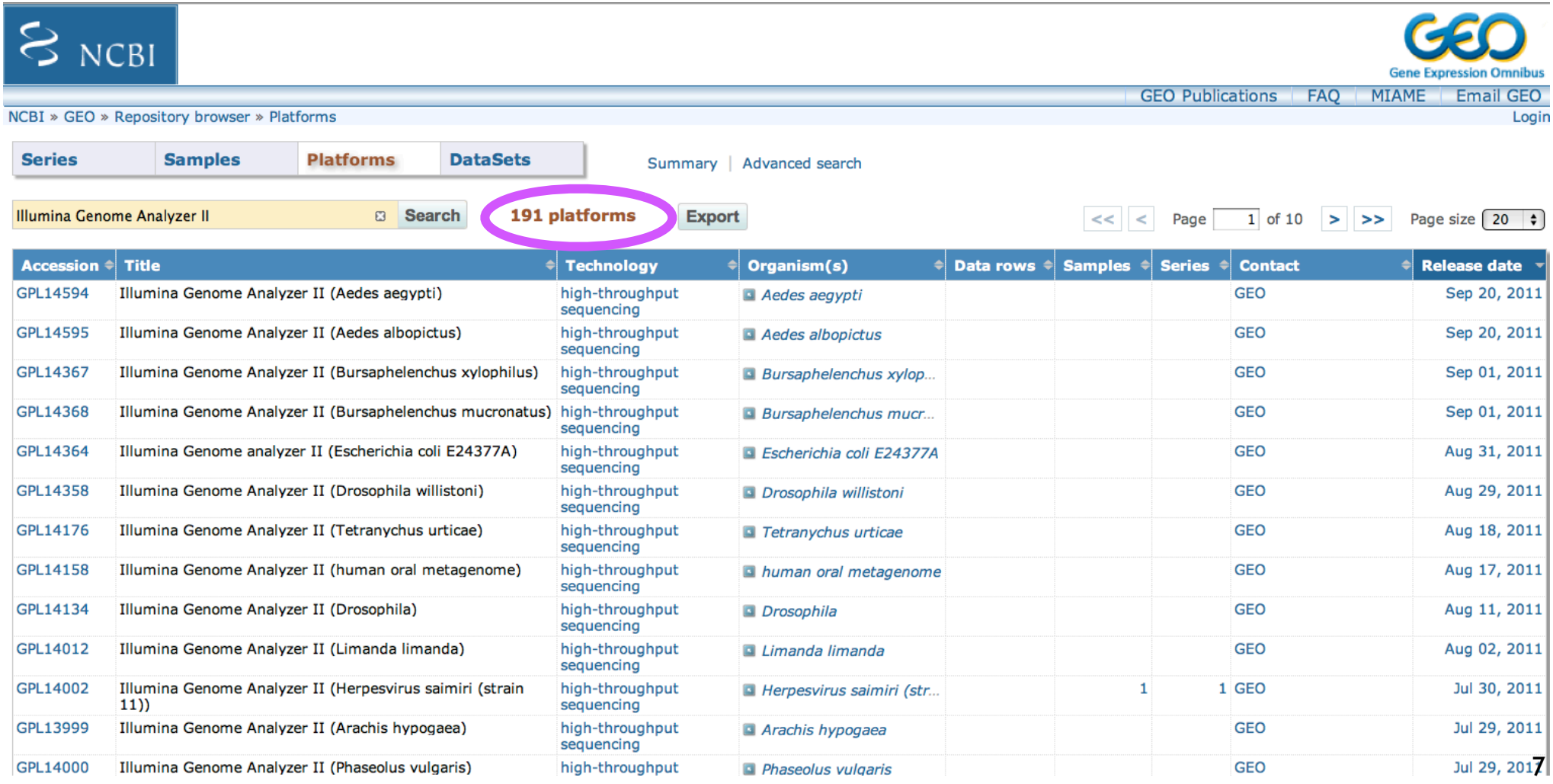
- PhD in Computer Science (and Computational Biology) from Hebrew University.
- Analyzed genomic data of histone modifications, nucleosome turnover, TF binding, gene expression, and DNA accessibility.

Obtaining ChIP-seq Data

- http://www.ncbi.nlm.nih.gov/projects/geo



Illumina Genome analyzer: hundreds of species



NCBI

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO » Repository browser » Platforms

Login

Series | Samples | Platforms | DataSets

Summary | Advanced search

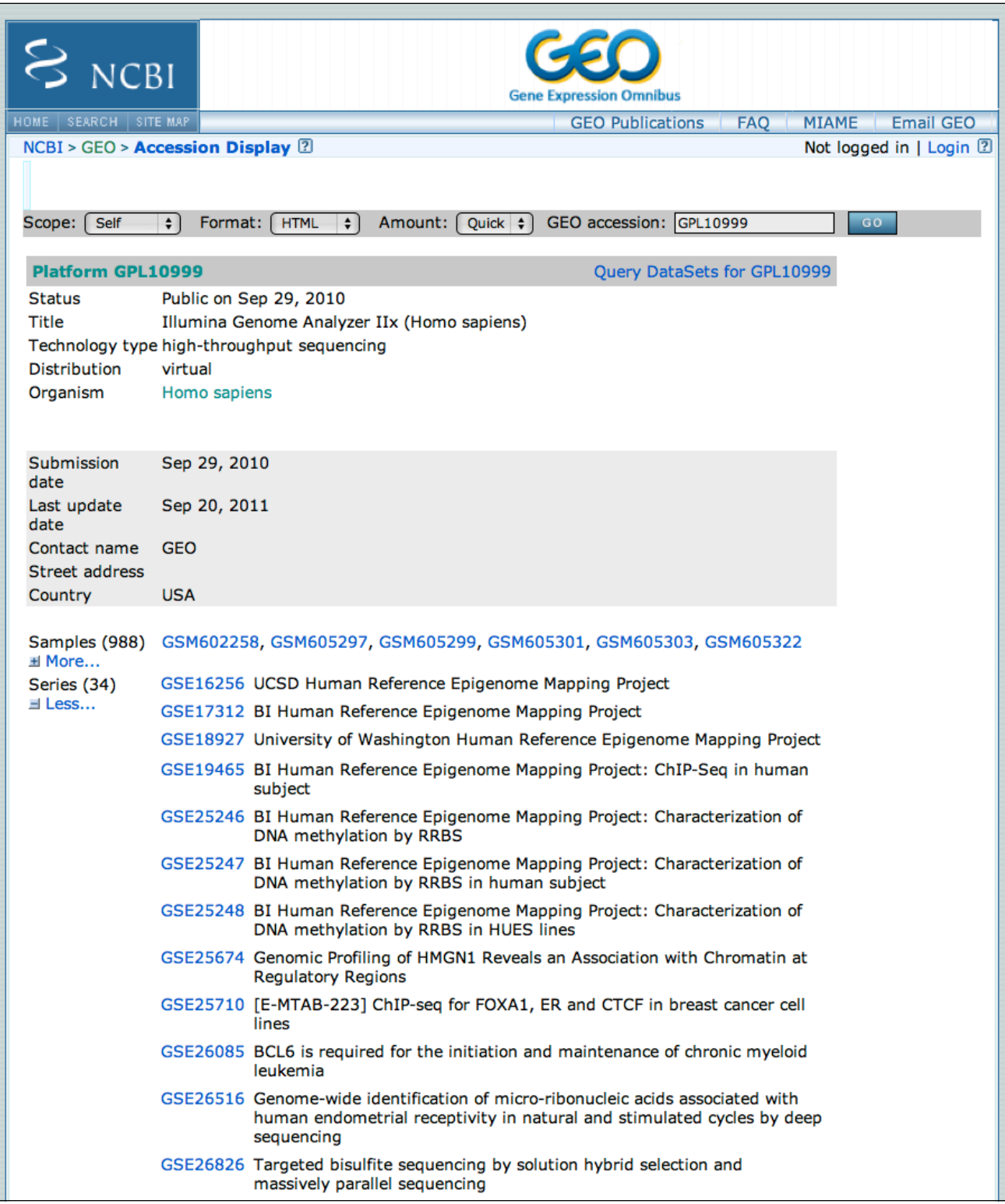
Illumina Genome Analyzer II Search 191 platforms Export

<< < Page 1 of 10 > >> Page size 20

Accession	Title	Technology	Organism(s)	Data rows	Samples	Series	Contact	Release date
GPL14594	Illumina Genome Analyzer II (Aedes aegypti)	high-throughput sequencing	*Aedes aegypti*				GEO	Sep 20, 2011
GPL14595	Illumina Genome Analyzer II (Aedes albopictus)	high-throughput sequencing	*Aedes albopictus*				GEO	Sep 20, 2011
GPL14367	Illumina Genome Analyzer II (Bursaphelenchus xylophilus)	high-throughput sequencing	*Bursaphelenchus xylop...*				GEO	Sep 01, 2011
GPL14368	Illumina Genome Analyzer II (Bursaphelenchus mucronatus)	high-throughput sequencing	*Bursaphelenchus mucr...*				GEO	Sep 01, 2011
GPL14364	Illumina Genome analyzer II (Escherichia coli E24377A)	high-throughput sequencing	*Escherichia coli E24377A*				GEO	Aug 31, 2011
GPL14358	Illumina Genome Analyzer II (Drosophila willistoni)	high-throughput sequencing	*Drosophila willistoni*				GEO	Aug 29, 2011
GPL14176	Illumina Genome Analyzer II (Tetranychus urticae)	high-throughput sequencing	*Tetranychus urticae*				GEO	Aug 18, 2011
GPL14158	Illumina Genome Analyzer II (human oral metagenome)	high-throughput sequencing	*human oral metagenome*				GEO	Aug 17, 2011
GPL14134	Illumina Genome Analyzer II (Drosophila)	high-throughput sequencing	*Drosophila*				GEO	Aug 11, 2011
GPL14012	Illumina Genome Analyzer II (Limanda limanda)	high-throughput sequencing	*Limanda limanda*				GEO	Aug 02, 2011
GPL14002	Illumina Genome Analyzer II (Herpesvirus saimiri (strain 11))	high-throughput sequencing	*Herpesvirus saimiri (str...*		1	1	GEO	Jul 30, 2011
GPL13999	Illumina Genome Analyzer II (Arachis hypogaea)	high-throughput sequencing	*Arachis hypogaea*				GEO	Jul 29, 2011
GPL14000	Illumina Genome Analyzer II (Phaseolus vulgaris)	high-throughput	*Phaseolus vulgaris*				GEO	Jul 29, 2011

>1000 human samples

Accession	Title	Technology	Organism(s)	Data rows	Samples	Series	Contact	Release date
Filter		high-throughput sequencing	*Homo sapiens*					
GPL14603	454 GS FLX Titanium (Homo sapiens)	high-throughput sequencing	*Homo sapiens*		16	2	GEO	Sep 21, 2011
GPL14583	Illumina Hiseq 2000 (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Sep 16, 2011
GPL14555	Illumina genome analyzer (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Sep 09, 2011
GPL13994	Illumina genome analyzer IIx (Homo sapiens)	high-throughput sequencing	*Homo sapiens*		67	1	GEO	Jul 29, 2011
GPL13978	Helicos Heliscope (Homo sapiens)	high-throughput sequencing	*Homo sapiens*		157	4	GEO	Jul 27, 2011
GPL13981	AB SOLiD System 4 (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Jul 27, 2011
GPL13873	AB SOLiD system 3.0 (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Jul 11, 2011
GPL13731	Illumina Genome Analyzer Iix (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Jun 16, 2011
GPL13484	454 Titanium (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	May 03, 2011
GPL13477	Illumina Genome Analyzer IIX (Homo sapiens)	high-throughput sequencing	*Homo sapiens*		8	2	GEO	May 02, 2011
GPL13393	AB SOLiD 4 System (Homo sapiens)	high-throughput sequencing	*Homo sapiens*		9	1	GEO	Apr 08, 2011
GPL13357	AB SOLiD System v3+ (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Mar 31, 2011
GPL13317	Heliscope (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Mar 23, 2011
GPL11436	AB SOLiD 3 Plus (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Jan 12, 2011
GPL11255	AB SOLiD System 3 (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Nov 30, 2010
GPL11154	Illumina HiSeq 2000 (Homo sapiens)	high-throughput sequencing	*Homo sapiens*		137	8	GEO	Nov 02, 2010
GPL10999	Illumina Genome Analyzer IIx (Homo sapiens)	high-throughput sequencing	*Homo sapiens*		988	34	GEO	Sep 29, 2010
GPL10400	454 GS (Homo sapiens)	high-throughput sequencing	*Homo sapiens*		4	3	GEO	May 06, 2010
GPL10329	Illumina Genome Analyzer (Homo sapiens; Mus musculus)	high-throughput sequencing	*Homo sapiens* *Mus musculus*		2	1	GEO	Apr 14, 2010
GPL10297	Illumina Genome analyzer II (Homo sapiens)	high-throughput sequencing	*Homo sapiens*				GEO	Apr 07, 2010



NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GPL10999 GO

Platform GPL10999 Query DataSets for GPL10999

Status	Public on Sep 29, 2010
Title	Illumina Genome Analyzer IIx (Homo sapiens)
Technology type	high-throughput sequencing
Distribution	virtual
Organism	Homo sapiens

Submission date	Sep 29, 2010
Last update date	Sep 20, 2011
Contact name	GEO
Street address	
Country	USA

Samples (988) GSM602258, GSM605297, GSM605299, GSM605301, GSM605303, GSM605322
More...

Series (34)
Less...

GSE16256 UCSD Human Reference Epigenome Mapping Project

GSE17312 BI Human Reference Epigenome Mapping Project

GSE18927 University of Washington Human Reference Epigenome Mapping Project

GSE19465 BI Human Reference Epigenome Mapping Project: ChIP-Seq in human subject

GSE25246 BI Human Reference Epigenome Mapping Project: Characterization of DNA methylation by RRBS

GSE25247 BI Human Reference Epigenome Mapping Project: Characterization of DNA methylation by RRBS in human subject

GSE25248 BI Human Reference Epigenome Mapping Project: Characterization of DNA methylation by RRBS in HUES lines

GSE25674 Genomic Profiling of HMGN1 Reveals an Association with Chromatin at Regulatory Regions

GSE25710 [E-MTAB-223] ChIP-seq for FOXA1, ER and CTCF in breast cancer cell lines

GSE26085 BCL6 is required for the initiation and maintenance of chronic myeloid leukemia

GSE26516 Genome-wide identification of micro-ribonucleic acids associated with human endometrial receptivity in natural and stimulated cycles by deep sequencing

GSE26826 Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing

ChIP-seq of FoxAI, ER & CTCF

NATURE GENETICS | ARTICLE

FOXA1 is a key determinant of estrogen receptor function and endocrine response

Antoni Hurtado, Kelly A Holmes, Caryn S Ross-Innes, Dominic Schmidt & Jason S Carroll

Affiliations | Contributions | Corresponding author

Nature Genetics **43**, 27–33 (2011) | doi:10.1038/ng.730
Received 24 August 2010 | Accepted 01 November 2010 | Published online 12 December 2010

Abstract

Abstract • Introduction • Results • Discussion • Methods • References • Acknowledgments • Author information • Supplementary information

Estrogen receptor-α (ER) is the key feature of most breast cancers and binding of ER to the genome correlates with expression of the Forkhead protein FOXA1 (also called HNF3α). Here we show that FOXA1 is a key determinant that can influence differential interactions between ER and chromatin. Almost all ER-chromatin interactions and gene expression changes depended on the presence of FOXA1 and FOXA1 influenced genome-wide chromatin accessibility. Furthermore, we found that CTCF was an upstream negative regulator of FOXA1-chromatin interactions. In estrogen-responsive breast cancer cells, the dependency on FOXA1 for tamoxifen-ER activity was absolute; in tamoxifen-resistant cells, ER binding was independent of ligand but depended on FOXA1. Expression of FOXA1 in non-breast cancer cells can alter ER binding and function. As such, FOXA1 is a major determinant of estrogen-ER activity and endocrine response in breast cancer cells.

DNaseI hyper-sensitivity

nature biotechnology

NATURE BIOTECHNOLOGY | OPINION AND COMMENT | COMMENTARY

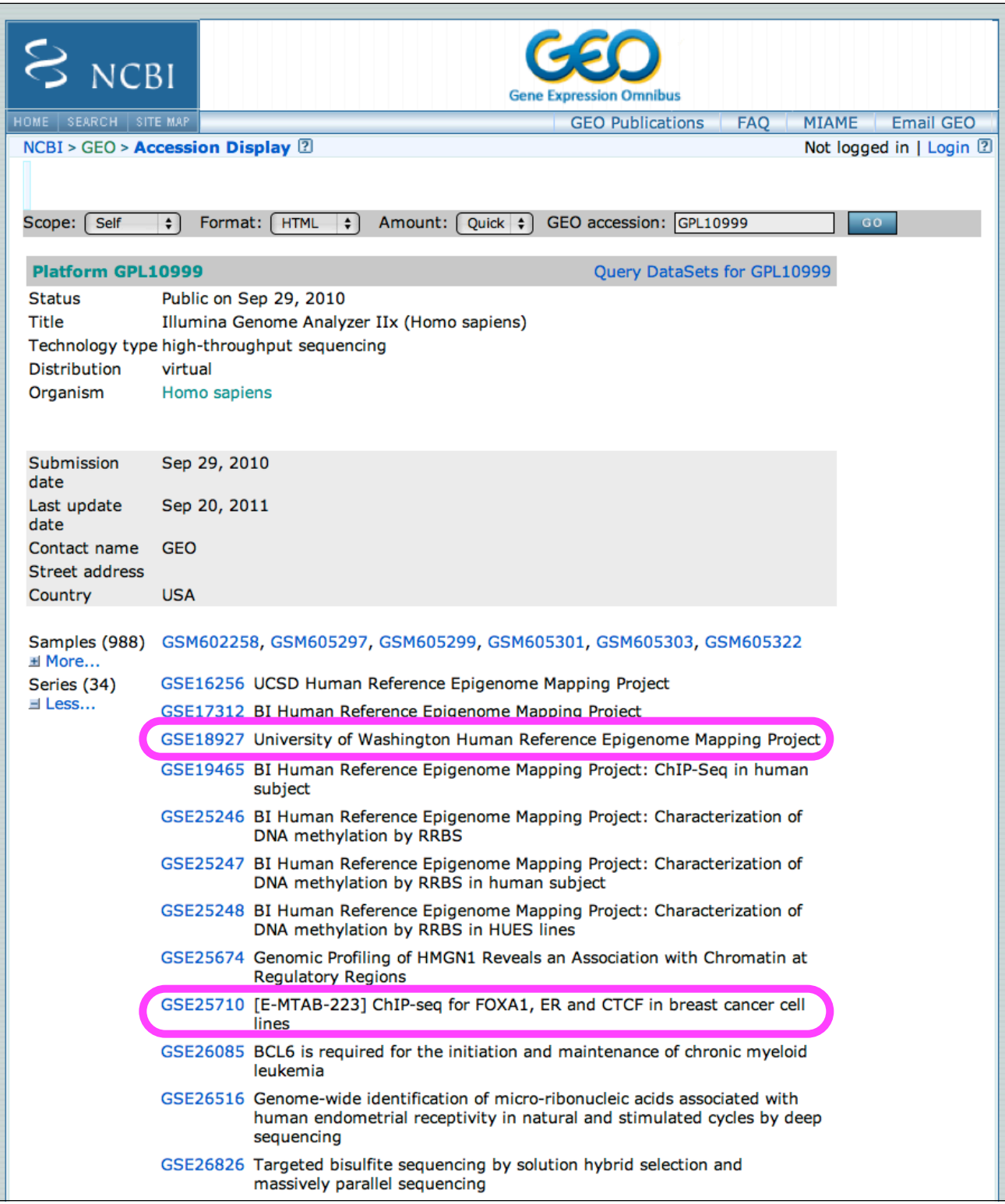
The NIH Roadmap Epigenomics Mapping Consortium

Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen & James A Thomson

Affiliations | **Corresponding author**

Nature Biotechnology **28**, 1045–1048 (2010) | doi:10.1038/nbt1010-1045
Published online 13 October 2010

The NIH Roadmap Epigenomics Mapping Consortium aims to produce a public resource of epigenomic maps for stem cells and primary *ex vivo* tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease.


NCBI
Gene Expression Omnibus
HOME SEARCH SITE MAP
GEO Publications FAQ MIAME Email GEO
NCBI > GEO > Accession Display
Not logged in | Login
Scope: Self Format: HTML Amount: Quick GEO accession: GPL10999 GO
Platform GPL10999
Query DataSets for GPL10999
Status Public on Sep 29, 2010
Title Illumina Genome Analyzer IIx (Homo sapiens)
Technology type high-throughput sequencing
Distribution virtual
Organism Homo sapiens
Submission date Sep 29, 2010
Last update date Sep 20, 2011
Contact name GEO
Street address
Country USA
Samples (988) GSM602258, GSM605297, GSM605299, GSM605301, GSM605303, GSM605322
More...
Series (34)
Less...
GSE16256 UCSD Human Reference Epigenome Mapping Project
GSE17312 BI Human Reference Epigenome Mapping Project
GSE18927 University of Washington Human Reference Epigenome Mapping Project
GSE19465 BI Human Reference Epigenome Mapping Project: ChIP-Seq in human subject
GSE25246 BI Human Reference Epigenome Mapping Project: Characterization of DNA methylation by RRBS
GSE25247 BI Human Reference Epigenome Mapping Project: Characterization of DNA methylation by RRBS in human subject
GSE25248 BI Human Reference Epigenome Mapping Project: Characterization of DNA methylation by RRBS in HUES lines
GSE25674 Genomic Profiling of HMGN1 Reveals an Association with Chromatin at Regulatory Regions
GSE25710 [E-MTAB-223] ChIP-seq for FOXA1, ER and CTCF in breast cancer cell lines
GSE26085 BCL6 is required for the initiation and maintenance of chronic myeloid leukemia
GSE26516 Genome-wide identification of micro-ribonucleic acids associated with human endometrial receptivity in natural and stimulated cycles by deep sequencing
GSE26826 Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing

Data deposited to GEO

- Raw reads (SRA or FASTQ format)

```
@FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
AAATGTTAATCTGAANAGCTGGAATCCAGTCTGGTGTTTGTA
+FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
BACCBBCBBCCBB:=!1=CAB7B@BBCCA<C?>A8B>@AAAC
@FoxA1.2 HWUSI-EAS582_229:6:1:1:569 length=42
CAGTATGGAGGTGAATAAACAGCAGATGGCCTGGAAGATACA
+FoxA1.2 HWUSI-EAS582_229:6:1:1:569 length=42
A?AB>CA@AB:833;>:2A@)@@?><6(?9@?;135B4>A??
```

- Mapped reads (BED)
- Read coverage along genome (WIG)
- Bound regions

Chromatin accessibility

Sample GSM753974 Query DataSets for GSM753974

Status	Public on Jul 06, 2011
Title	Chromatin accessibility assay of Breast_vHMEC; DS18438
Sample type	SRA
Source name	Breast, vHMEC, RM035; DS18438
Organism	Homo sapiens
Characteristics	sample alias: Breast vHMEC RM035 sample common name: Breast, vHMEC molecule: genomic DNA disease: None biomaterial_provider: Tlsty Lab, University of California, San Francisco biomaterial_type: Primary Cell Culture cell_type: Variant Human Mammary Epithelial Cells markers: NA culture_conditions: spilt 1:5 in MEGM medium (Lonza Inc.) donor_id: RM035 donor_age: 18 donor_health_status: Disease Free donor_sex: Female donor_ethnicity: Caucasian passage_if_expanded: 8 karyotype: 46,XX,1dmin parity: N/A experiment_type: Chromatin Accessibility extraction_protocol: Qiagen minElut dnase_protocol: Stamlab DNase Protocol, Sabo, P. J. et al. Nat Methods 3, 511-518 (2006)
Extracted molecule	genomic DNA
Extraction protocol	Library construction protocol: Single read - Illumina
Library strategy	DNase-Hypersensitivity
Library source	genomic
Library selection	DNAse
Instrument model	Illumina HiSeq 2000

Platform ID	GPL11154
Series (1)	GSE18927 University of Washington Human Reference Epigenome Mapping Project

Relations

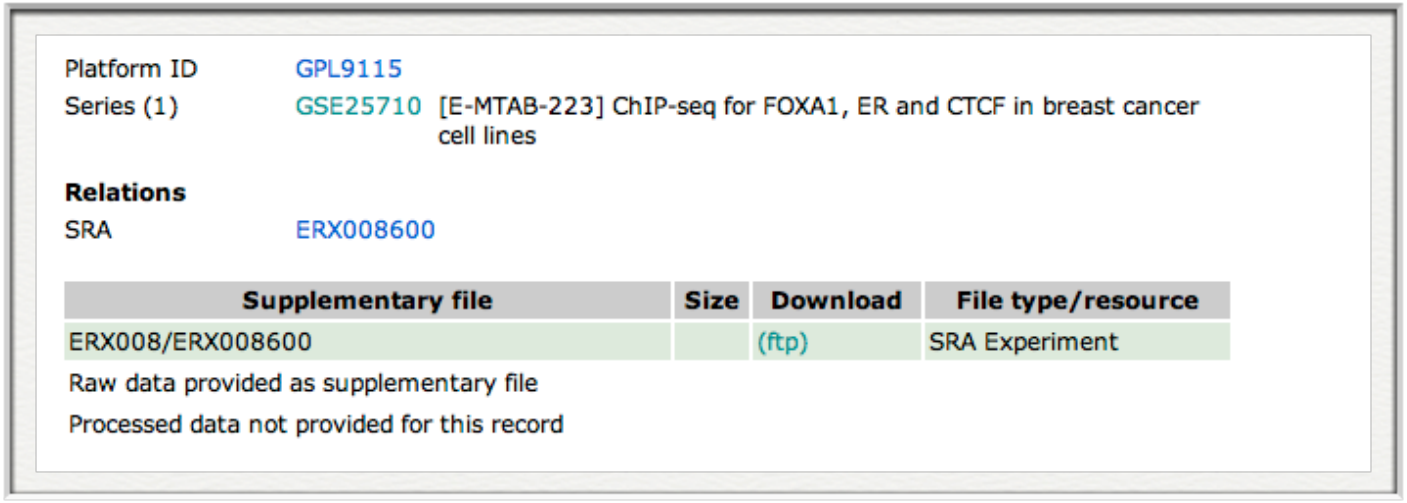
SRA	SRX081374

Supplementary file	Size	Download	File type/resource
SRX/SRX081/SRX081374		(ftp)	SRA Experiment
GSM753974_UW.Breast_vHMEC.ChromatinAccessibility.RM035.DS18438.bed.gz	436.1 Mb	(ftp)(http)	BED
GSM753974_UW.Breast_vHMEC.ChromatinAccessibility.RM035.DS18438.wig.gz	168.2 Mb	(ftp)(http)	WIG

Raw data provided as supplementary file

FoxA1, ER, CTCF

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM631469



Platform ID GPL9115

Series (1) GSE25710 [E-MTAB-223] ChIP-seq for FOXA1, ER and CTCF in breast cancer cell lines

Relations

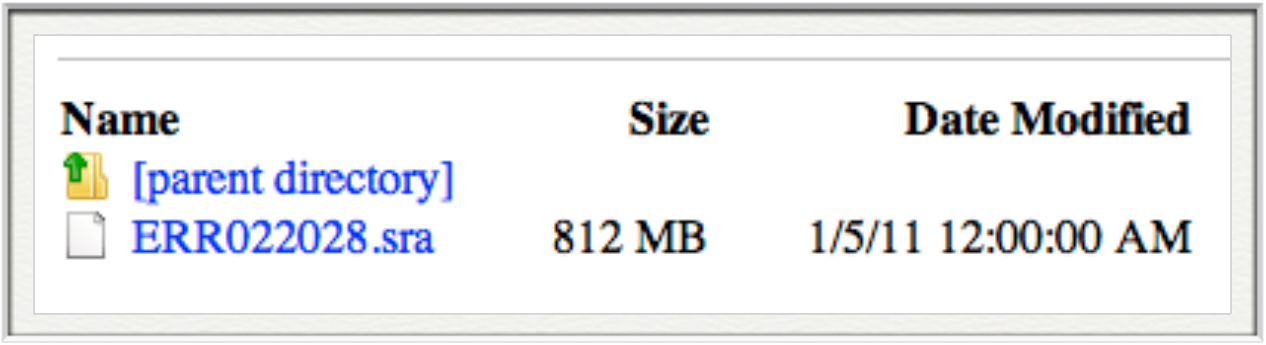
SRA ERX008600

Supplementary file	Size	Download	File type/resource
ERX008/ERX008600		(ftp)	SRA Experiment

Raw data provided as supplementary file

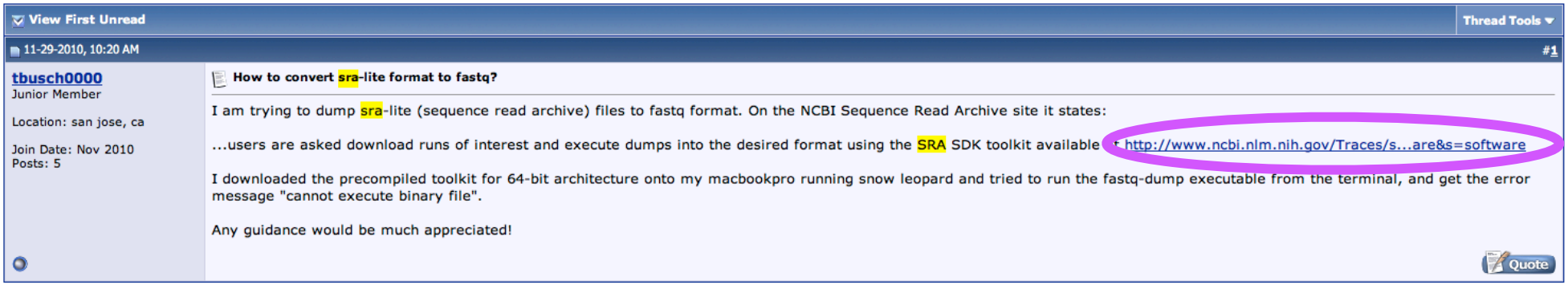
Processed data not provided for this record

ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByExp/sra/ERX%2FERX008%2FERX008600/ERR022028/



Name	Size	Date Modified
[parent directory]		
ERR022028.sra	812 MB	1/5/11 12:00:00 AM

Converting SRA to FASTQ


View First Unread

Thread Tools

11-29-2010, 10:20 AM

#1

tbusch0000
Junior Member

Location: san jose, ca

Join Date: Nov 2010
Posts: 5

How to convert sra-lite format to fastq?

I am trying to dump sra-lite (sequence read archive) files to fastq format. On the NCBI Sequence Read Archive site it states:

...users are asked download runs of interest and execute dumps into the desired format using the SRA SDK toolkit available at http://www.ncbi.nlm.nih.gov/Traces/s...are&s=software

I downloaded the precompiled toolkit for 64-bit architecture onto my macbookpro running snow leopard and tried to run the fastq-dump executable from the terminal, and get the error message "cannot execute binary file".

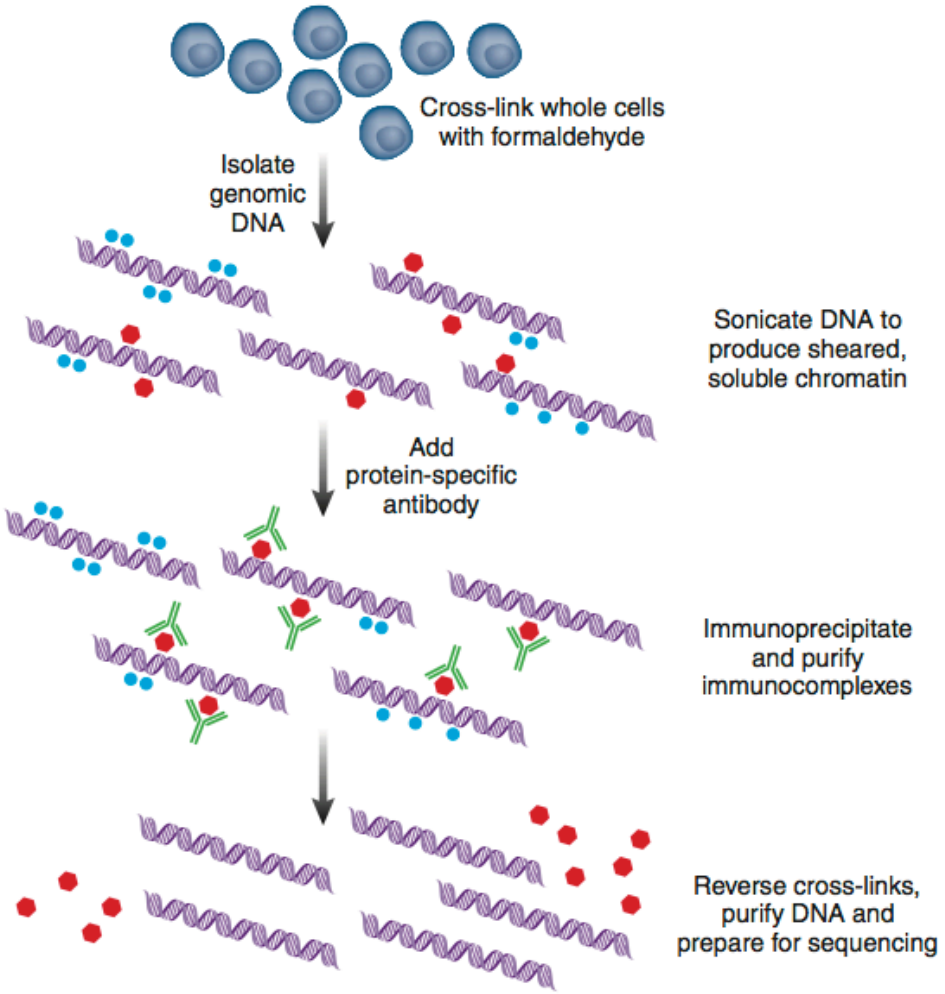
Any guidance would be much appreciated!

Quote

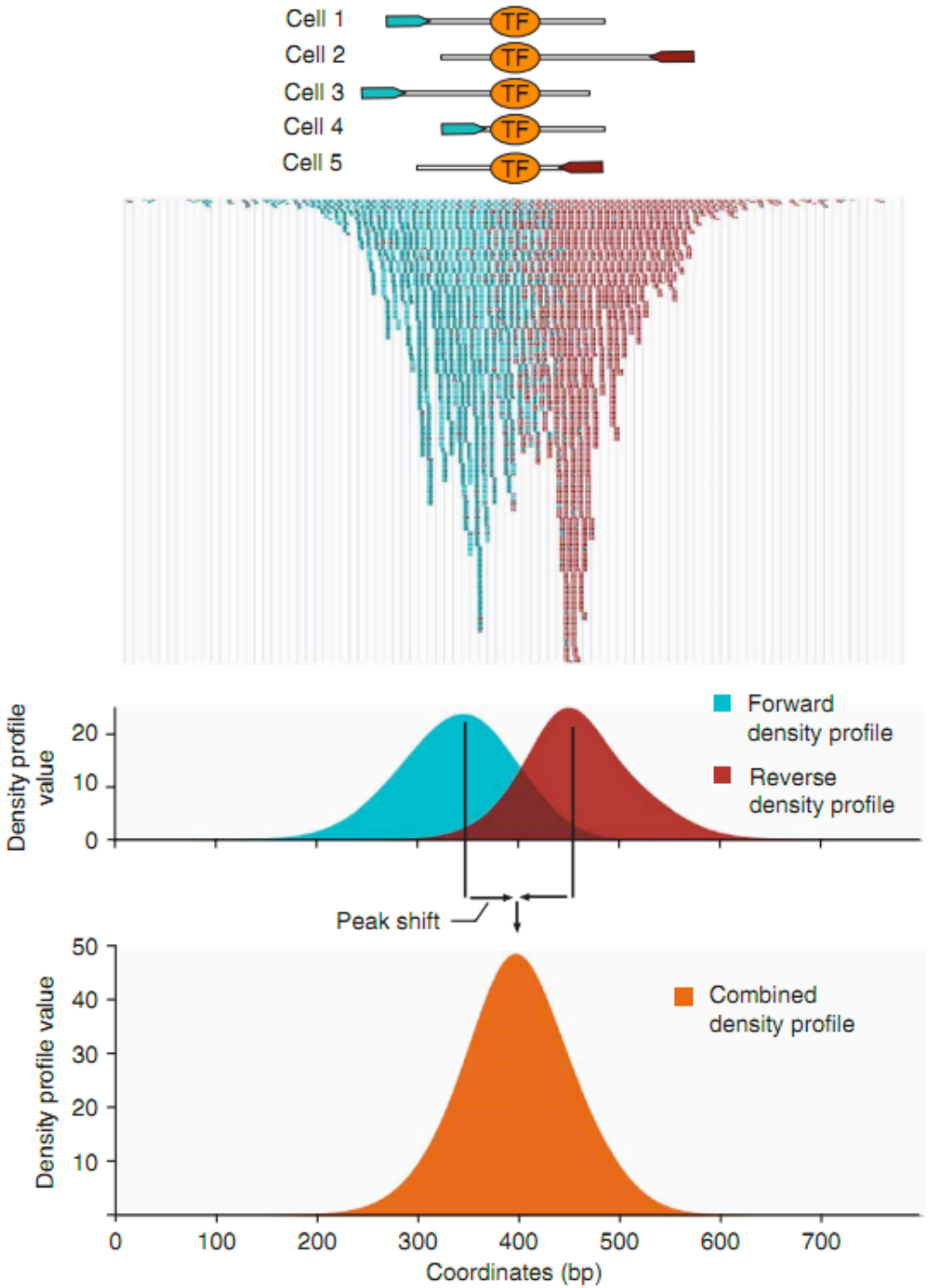
http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software

- tar zxvf sratoolkit.2.1.2-mac64.tgz
- fastq-dump FoxA1.sra

Chromatin Immunoprecipitation followed by sequencing



Mardis, 2007



Valouev et al, 2008

Basic questions

- Identify bound regions along genome
- Quantify binding occupancy
- Estimate peaks, identify DNA motif
- Where along the gene? Promoter, etc.
- Near which genes? Of specific function?
- Compare to other genomic data (time point, condition, cell line, other TF, etc)

Genomic mapping of sequenced data using BOWTIE

Software

Open Access

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Abstract

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Bowtie is open source http://bowtie.cbcb.umd.edu.

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

Running BOWTIE

- bowtie -c -q -n 2 -m 1 -5 3 hg19 <.fastq> <.bw>
- 18,517,316 reads onto a 3e9 genome in 90 minutes (~3500 reads per second)
- Notable parameters:
  - -q = input in FASTQ format
  - -n 2 = max of 2 mismatches
  - -m 1 = only reads w/ unique match are reported
  - -5 x / -3 x = trim x bases from 5' or 3' end of reads

Running BOWTIE

- bowtie -c -q -n 2 -m 1 -5 3 hg19 <.fastq> <.bw>
- Input

```
@FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
AAATGTTAATCTGAANAGCTGGAATCCAGTCTGGTGTTTGTA
+FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
BACCBBCBBCCBB:=!1=CAB7B@BBCCA<C?>A8B>@AAAC
```

- Output

```
FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
- chr15 62798646
TACAAACACCAGACTGGATTCCAGCTNTTCAGATTAACA
CAAA@>B8A>?C<ACCBB@B7BAC=1!=:BBCCBBCBBC
0 12:C>N
```

Genomic landscape of binding

- Reformat BOWTIE output as BED file

  awk -F'\t' '{OFS="\t";print $3, $4, $4+length($5)-1, $2}' $f > $f:r.bed

  ```
  chr15 62798646 62798684 -
  ```

- Compute coverage

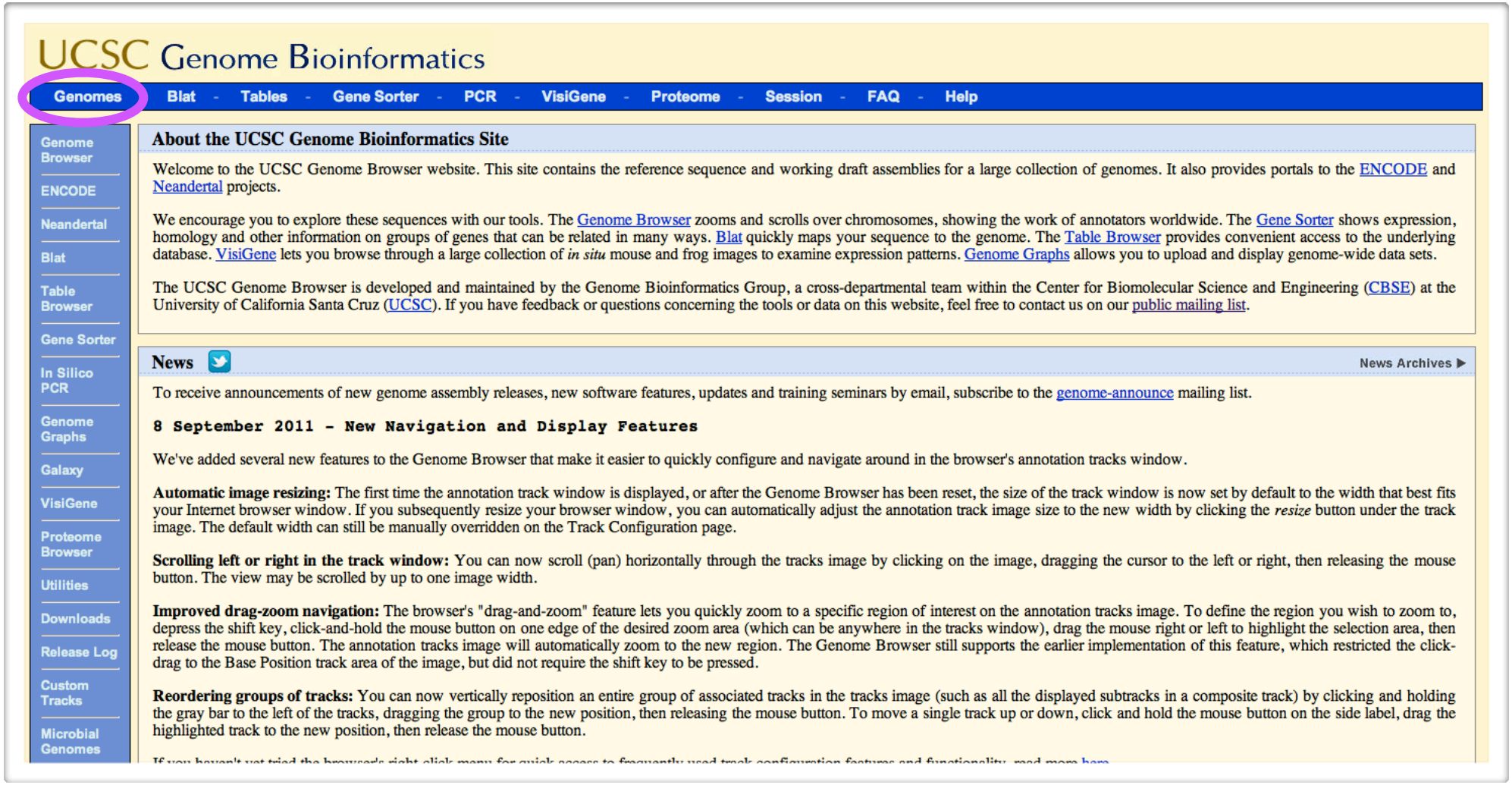
  ```
  create_coverage_VarStepWig.pl FoxA1.bed 250 25 1

  by Matt Blow, LBNL/JGI
  ```

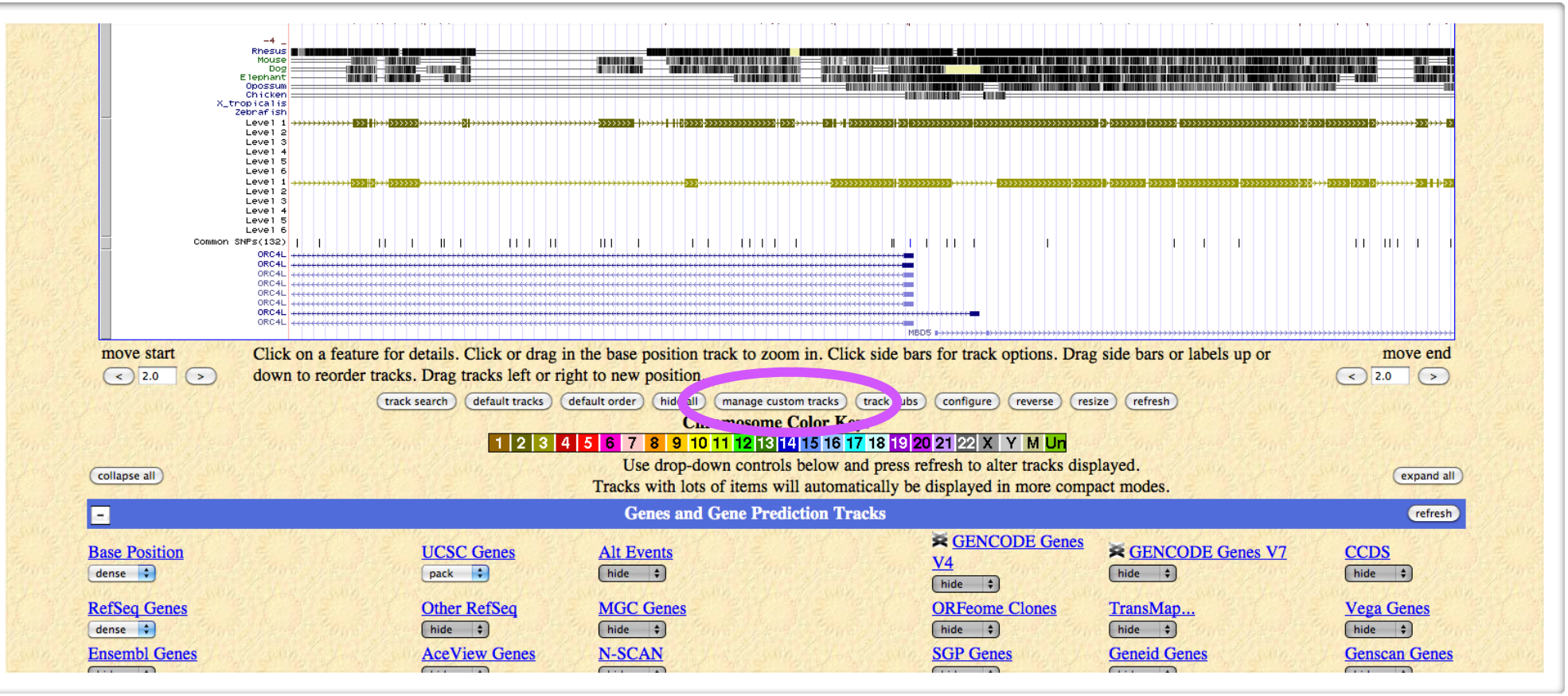
- Extends each read to 250 bp
- Calculate coverage in 25 bp windows

Visualization with UCSC browser

- http://genome.ucsc.edu



Visualize with UCSC browser

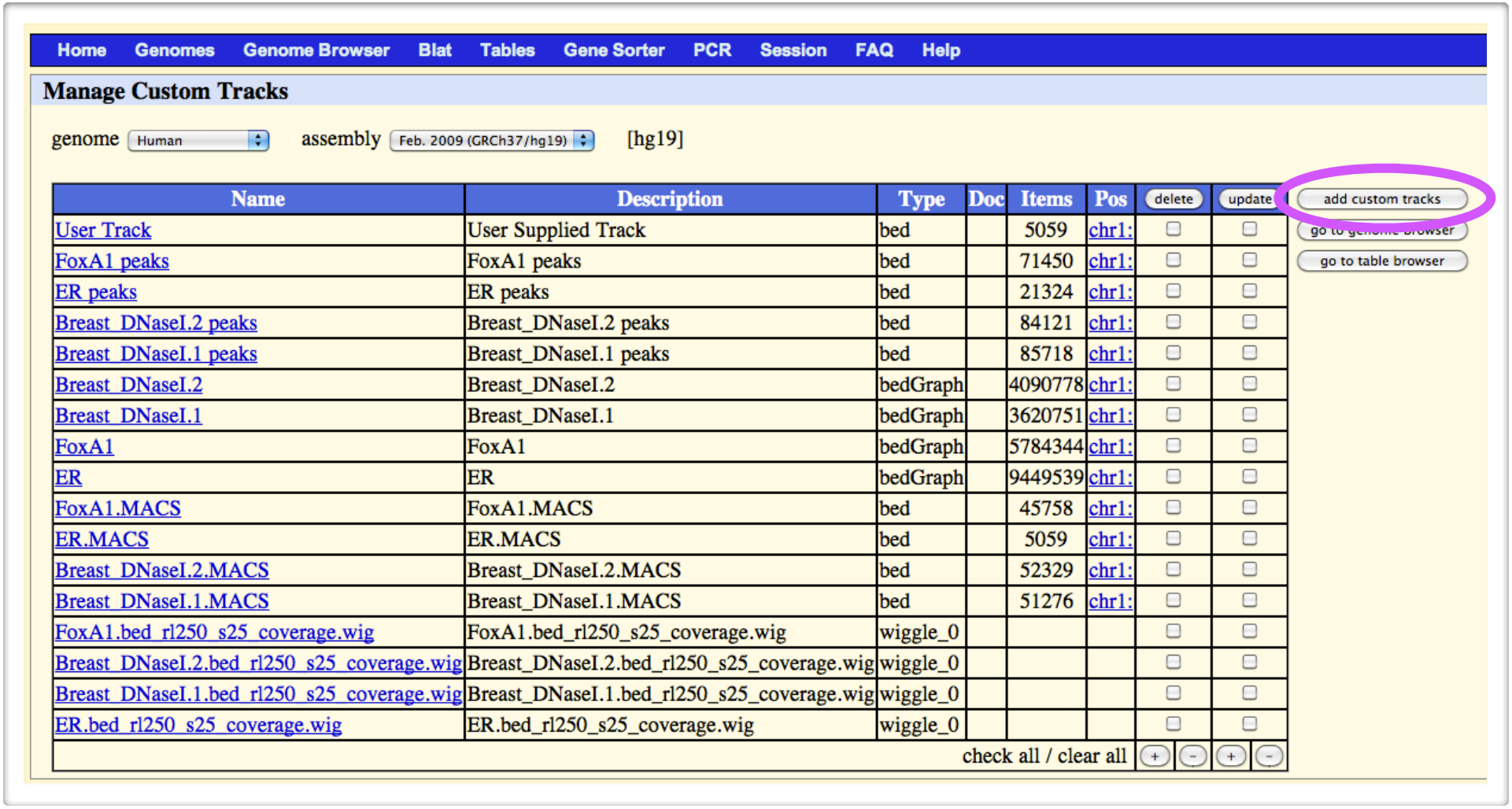


- Upload your own data as custom tracks

Visualize with UCSC browser

- Be sure to **bzip2** files before uploading
- BED
  - track type=bed name=... description=...
    chr1 867927 869317 peak_1 82.76
- WIG
  - track type=wiggle_0 name=... description=...
    chrom=chr1 span=25
    118363 1
    118388 1

Visualize with UCSC browser



Home | Genomes | Genome Browser | Blat | Tables | Gene Sorter | PCR | Session | FAQ | Help

Manage Custom Tracks

genome Human assembly Feb. 2009 (GRCh37/hg19) [hg19]

add custom tracks

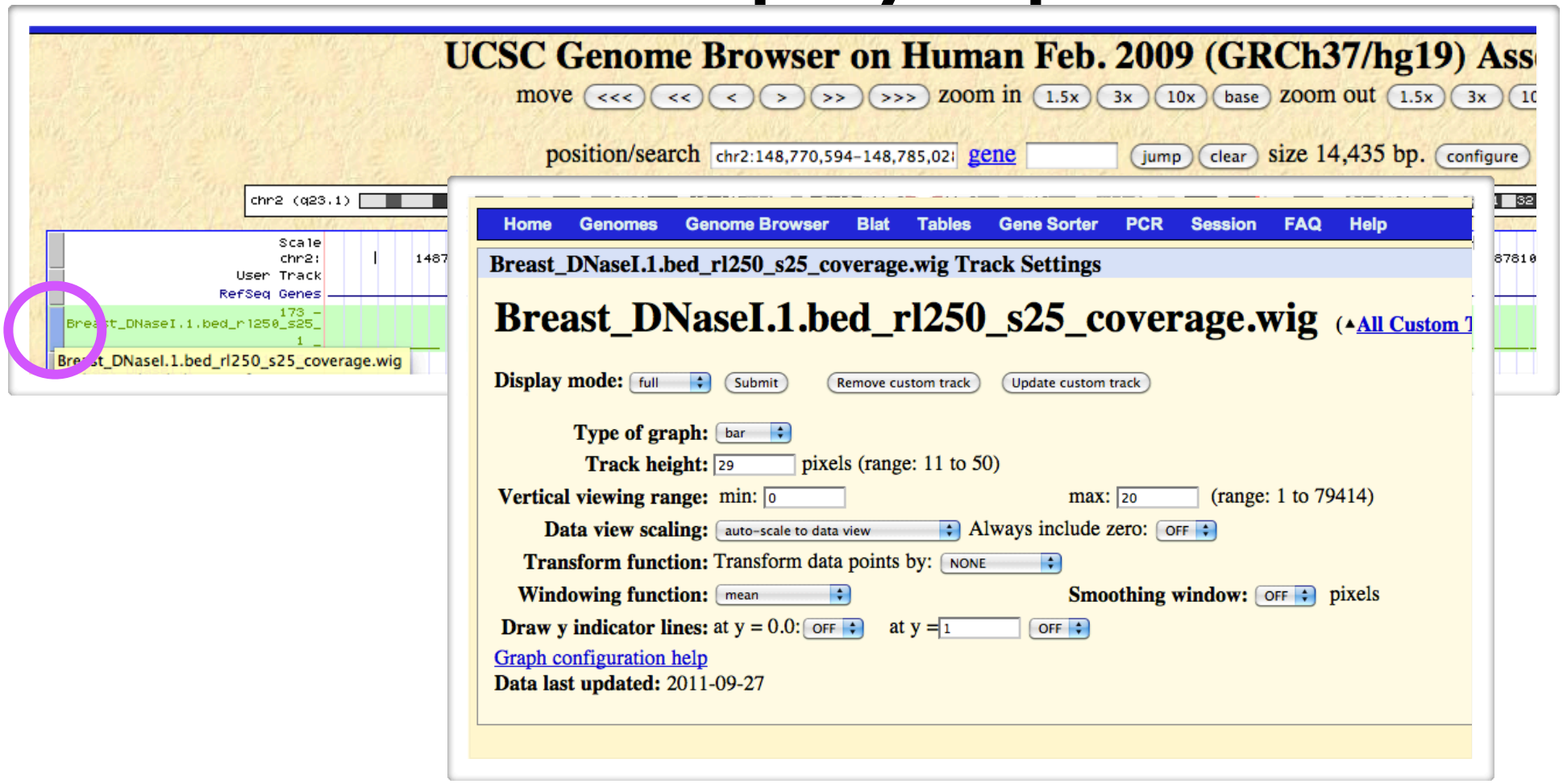
go to genome browser

go to table browser

Name	Description	Type	Doc	Items	Pos	delete	update
User Track	User Supplied Track	bed		5059	chr1:		
FoxA1 peaks	FoxA1 peaks	bed		71450	chr1:		
ER peaks	ER peaks	bed		21324	chr1:		
Breast_DNaseI.2 peaks	Breast_DNaseI.2 peaks	bed		84121	chr1:		
Breast_DNaseI.1 peaks	Breast_DNaseI.1 peaks	bed		85718	chr1:		
Breast_DNaseI.2	Breast_DNaseI.2	bedGraph		4090778	chr1:		
Breast_DNaseI.1	Breast_DNaseI.1	bedGraph		3620751	chr1:		
FoxA1	FoxA1	bedGraph		5784344	chr1:		
ER	ER	bedGraph		9449539	chr1:		
FoxA1.MACS	FoxA1.MACS	bed		45758	chr1:		
ER.MACS	ER.MACS	bed		5059	chr1:		
Breast_DNaseI.2.MACS	Breast_DNaseI.2.MACS	bed		52329	chr1:		
Breast_DNaseI.1.MACS	Breast_DNaseI.1.MACS	bed		51276	chr1:		
FoxA1.bed_rl250_s25_coverage.wig	FoxA1.bed_rl250_s25_coverage.wig	wiggle_0					
Breast_DNaseI.2.bed_rl250_s25_coverage.wig	Breast_DNaseI.2.bed_rl250_s25_coverage.wig	wiggle_0					
Breast_DNaseI.1.bed_rl250_s25_coverage.wig	Breast_DNaseI.1.bed_rl250_s25_coverage.wig	wiggle_0					
ER.bed_rl250_s25_coverage.wig	ER.bed_rl250_s25_coverage.wig	wiggle_0					
						check all / clear all	

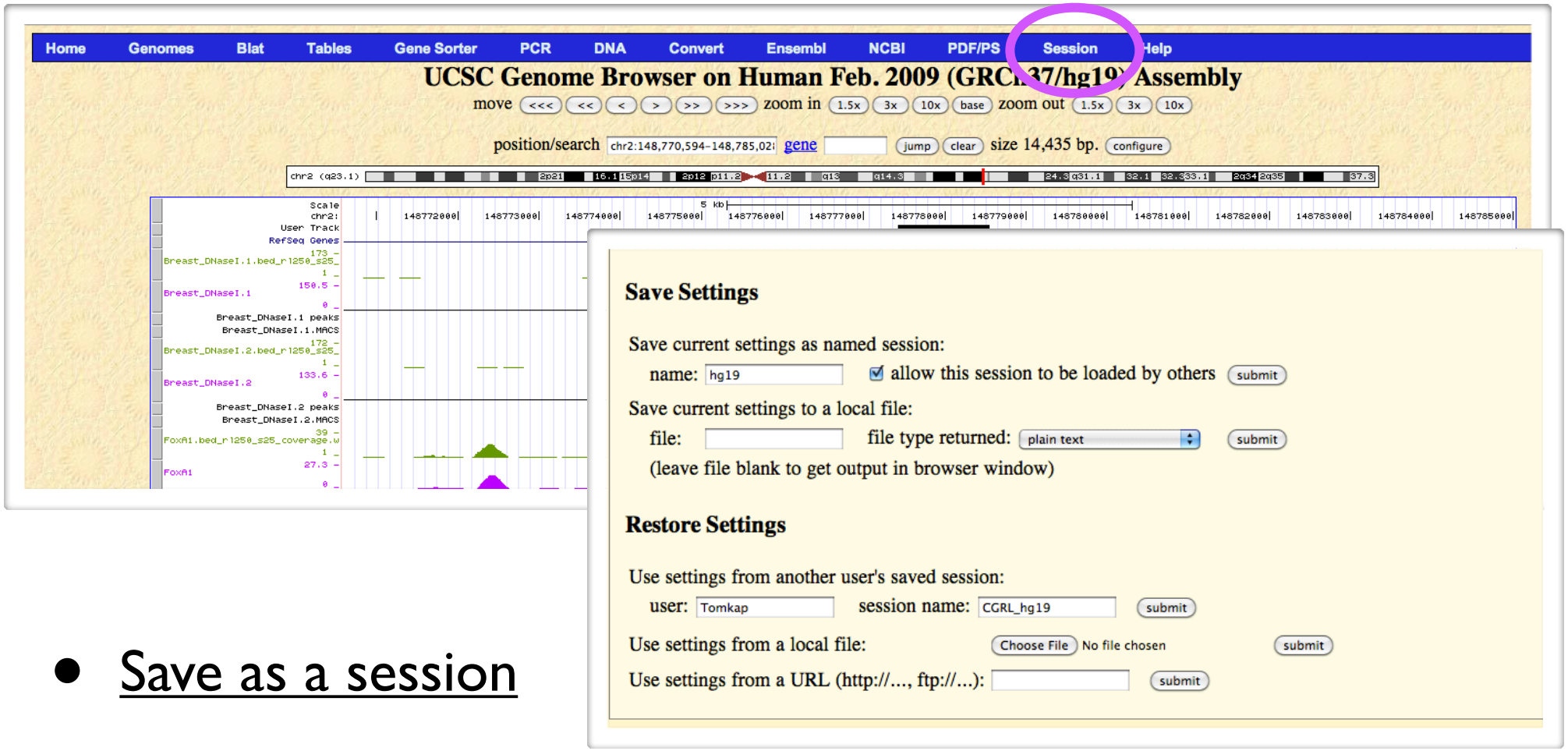
- Add as many tracks as needed

Track display options



- Line/filled area
- Height and range
- Smoothing and windowing
- Transformations
- Order of tracks

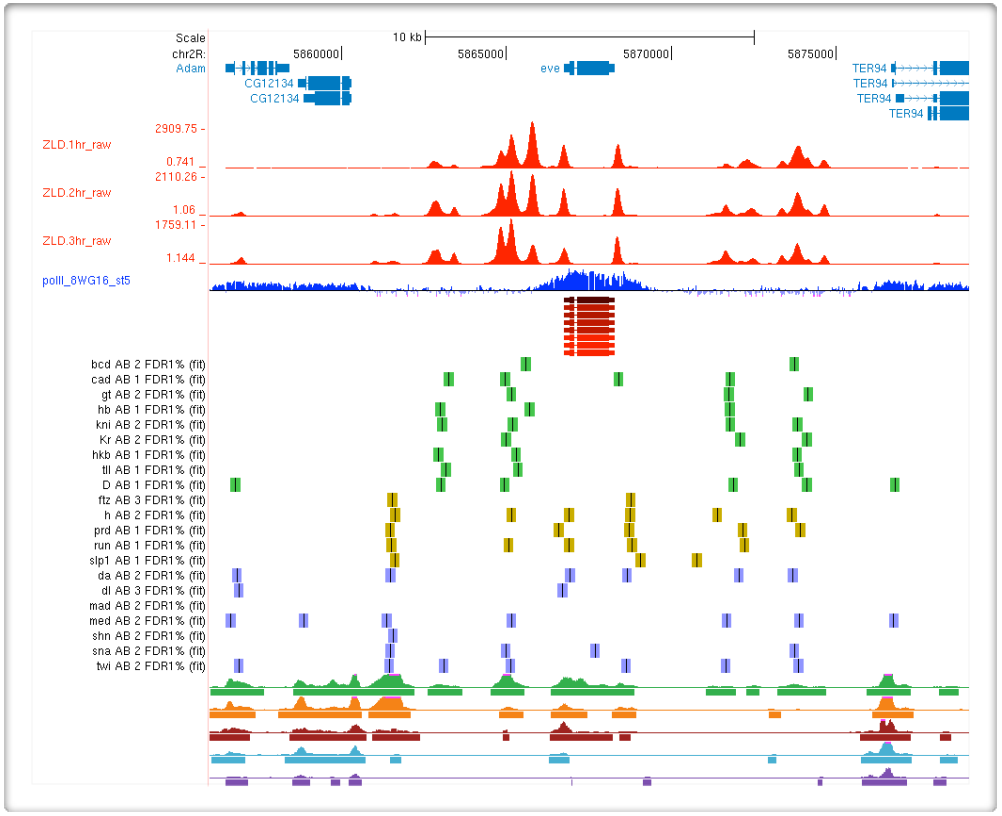
Visualize with UCSC browser

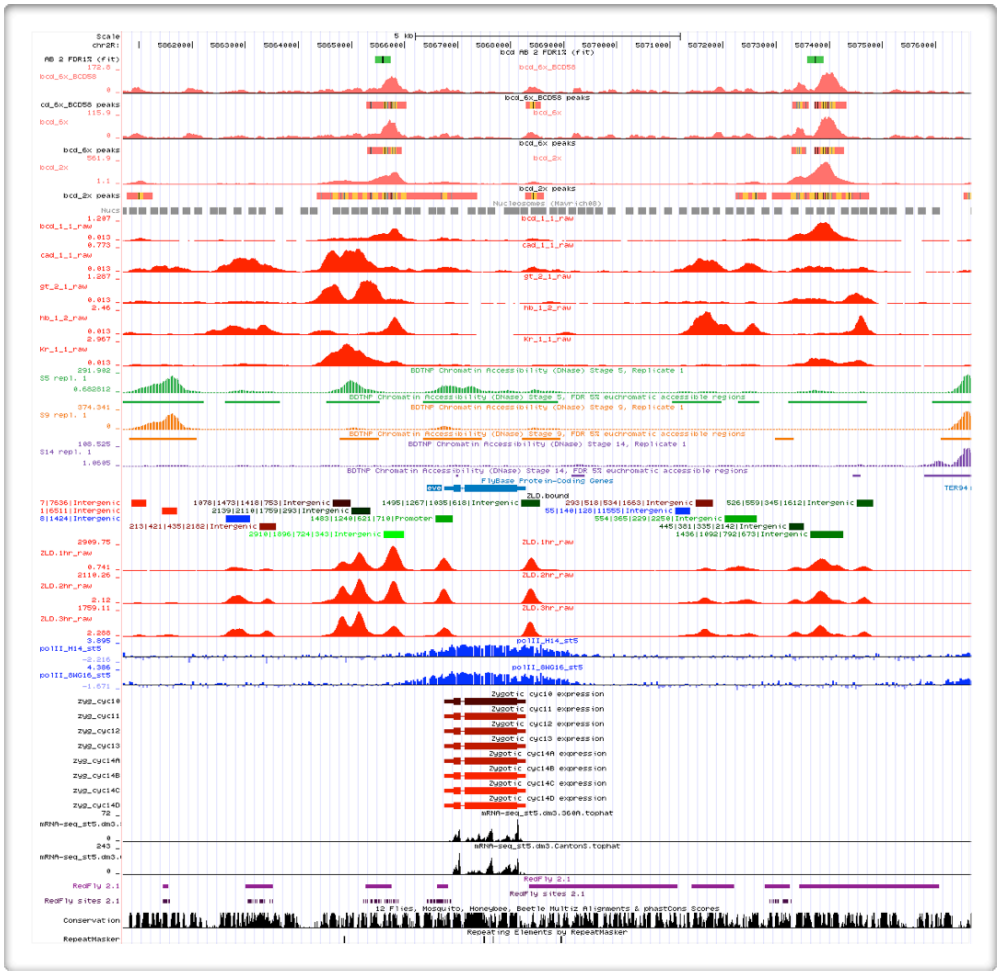


- Save as a session
- Share link with colleagues and friends
- Last 4 months since last access.
  Access periodically, e.g. via www.followthatpage.com

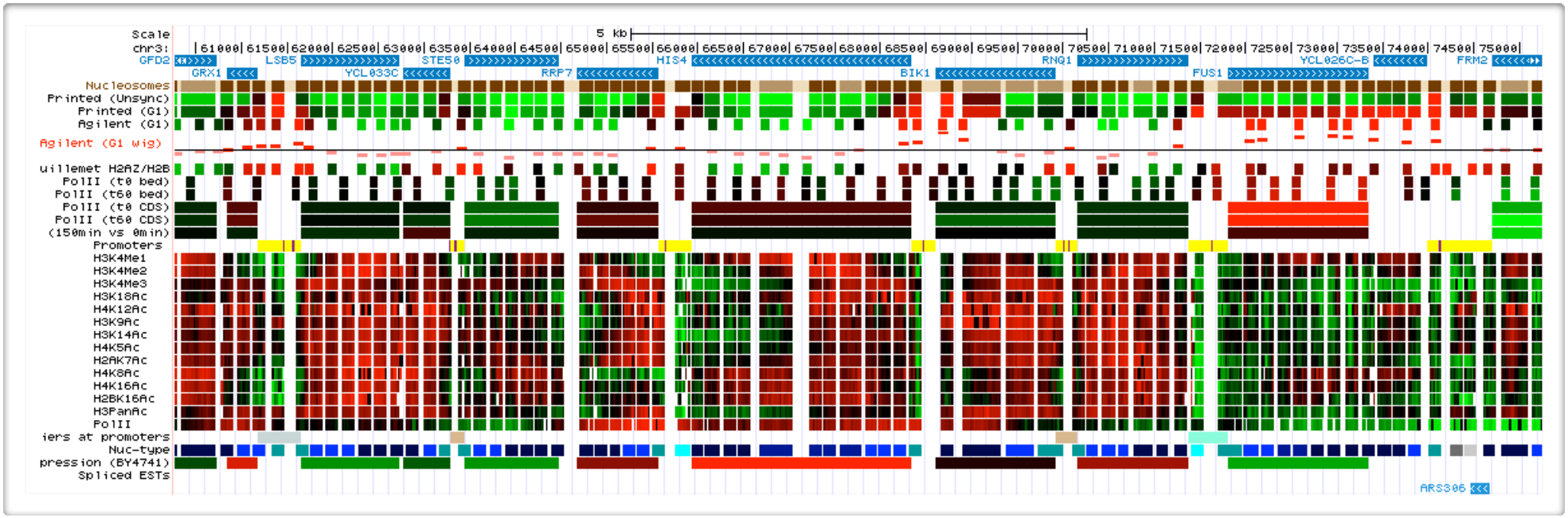
Some saved sessions

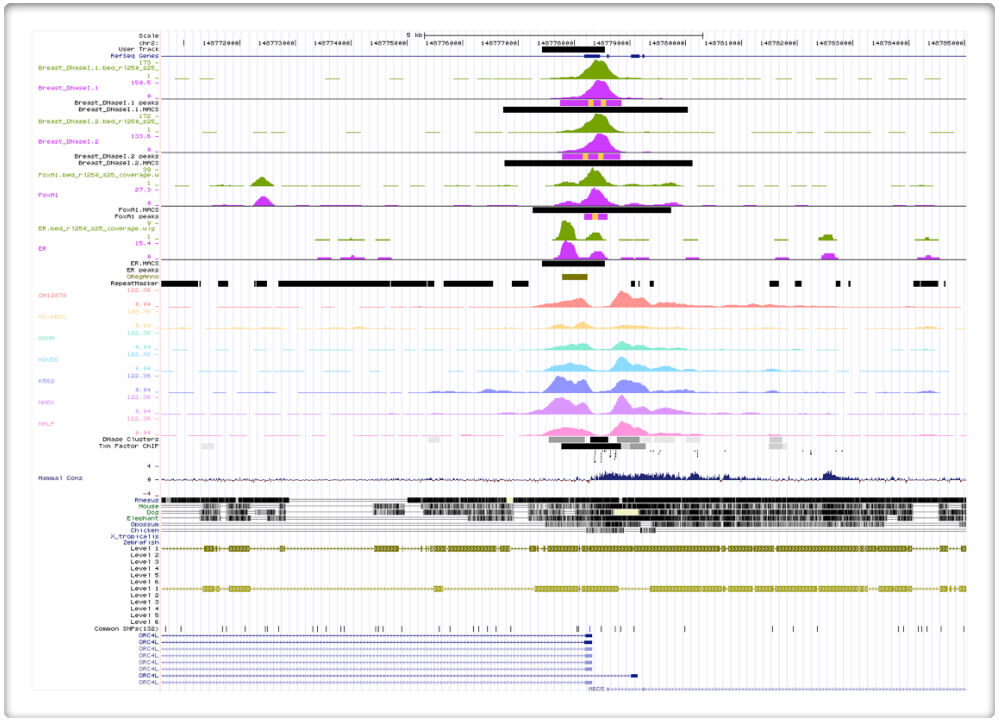
Binding and expression in Drosophila





Histone modifications and turnover rates in Yeast





CGRL example

Analyzing binding data

- Different data sets yield different questions

- Transcription factor ChIP
- Histone modifications
- DNaseI hypersensitivity

MACS

Open Access

Method

Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang[¤*], Tao Liu[¤*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nusbaum[¶], Richard M Myers[¥], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]

Published: 17 September 2008

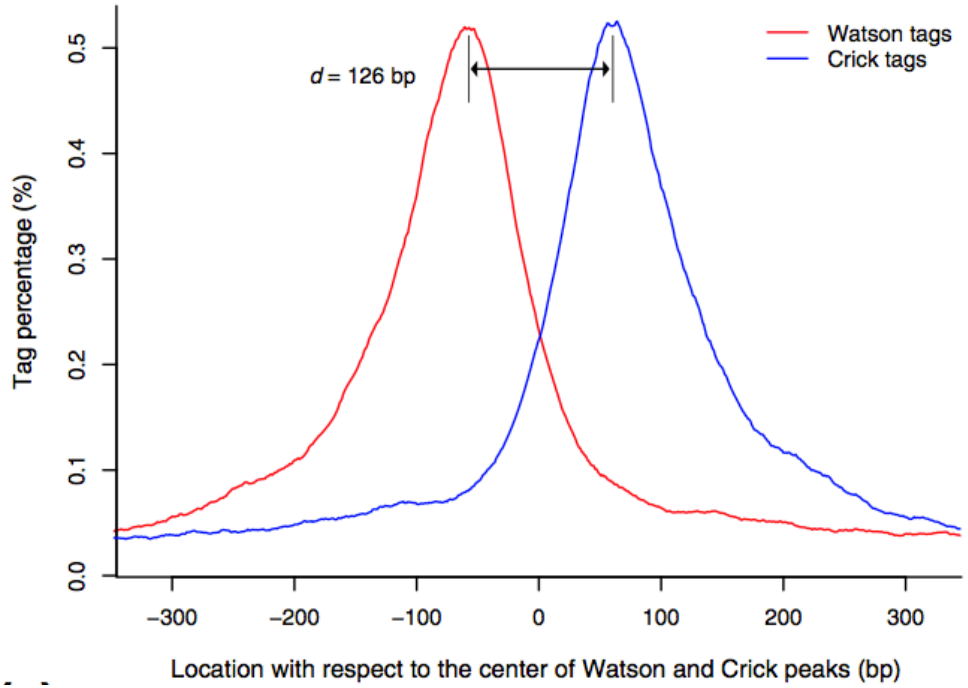
Genome ***Biology*** 2008, **9**:R137 (doi:10.1186/gb-2008-9-9-r137)

Abstract

We present Model-based Analysis of ChIP-Seq data, MACS, which analyzes data generated by short read sequencers such as Solexa's Genome Analyzer. MACS empirically models the shift size of ChIP-Seq tags, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome, allowing for more robust predictions. MACS compares favorably to existing ChIP-Seq peak-finding algorithms, and is freely available.

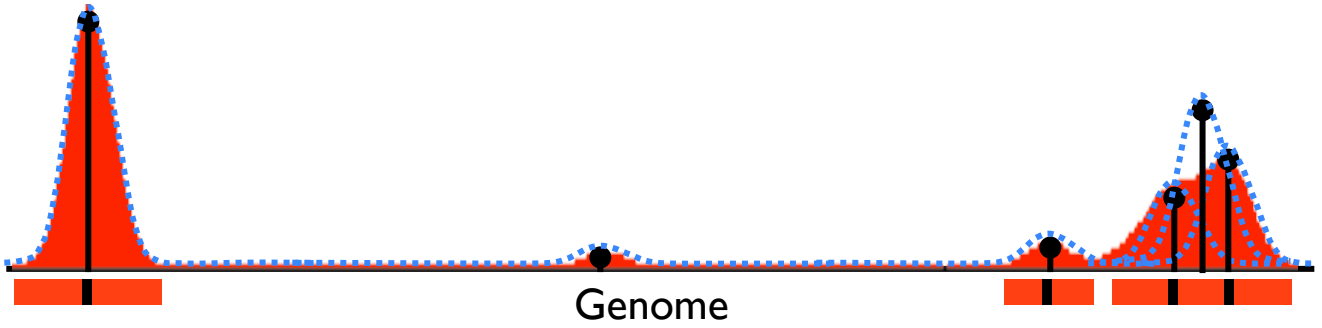
MACS

- Analyzes forward and reverse reads
- Identifies avg. length of DNA fragment
- Allows usage of mock IP for localized normalization



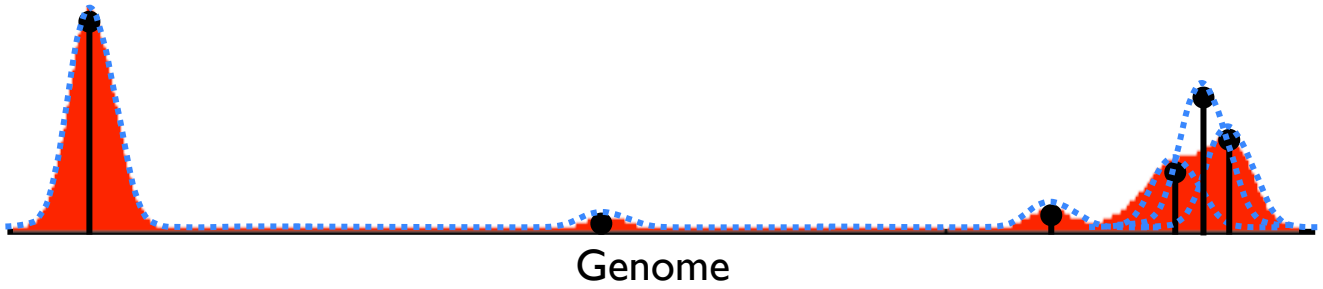
- macs14 -n FoxA1.MACS -t FoxA1.bed -g hs --off-auto --nomodel --shiftsize=125

Grizzly Peak Fitting



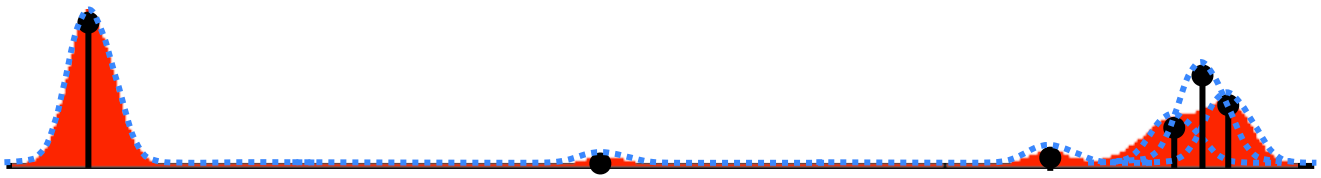
- Model-based, iterative, multi-peak peak finding algorithm
- MATLAB code, standalone executable*
- http://rana.lbl.gov/software/grizzly
- Capaldi, Kaplan et al, 2008, Harrison, Li, Kaplan et al, 2011

Grizzly Peak Fitting

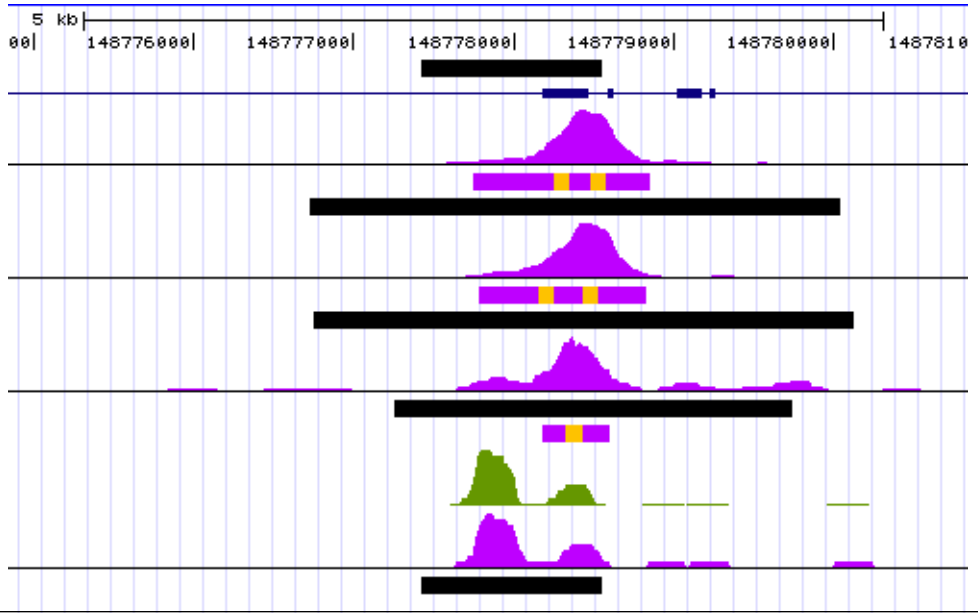


- Great in fitting the shape of a peak
- Given a motif, disentangle binding landscape into estimate occupancy at putative binding sites
- Bayesian confidence intervals for binding strengths
- Quite bad at stopping (separating real peaks from noise)
- pf0 FoxA1.bed 25 250 genome_hg19.mat 10

Intersecting Grizzly with MACS



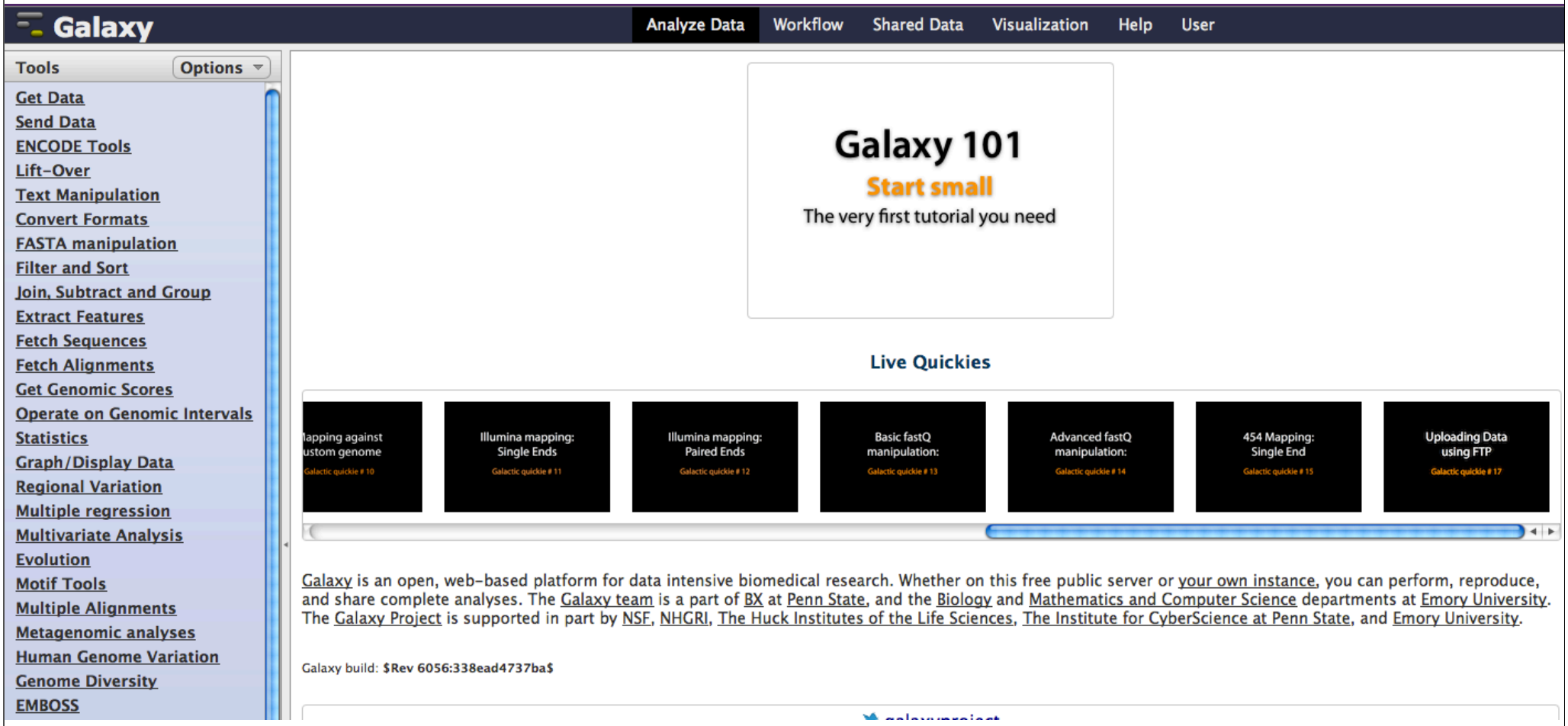
- Discard Grizzly peaks not overlapping a MACS peak
- Accurate prediction of binding position
- Robust calling of bound regions



What next?

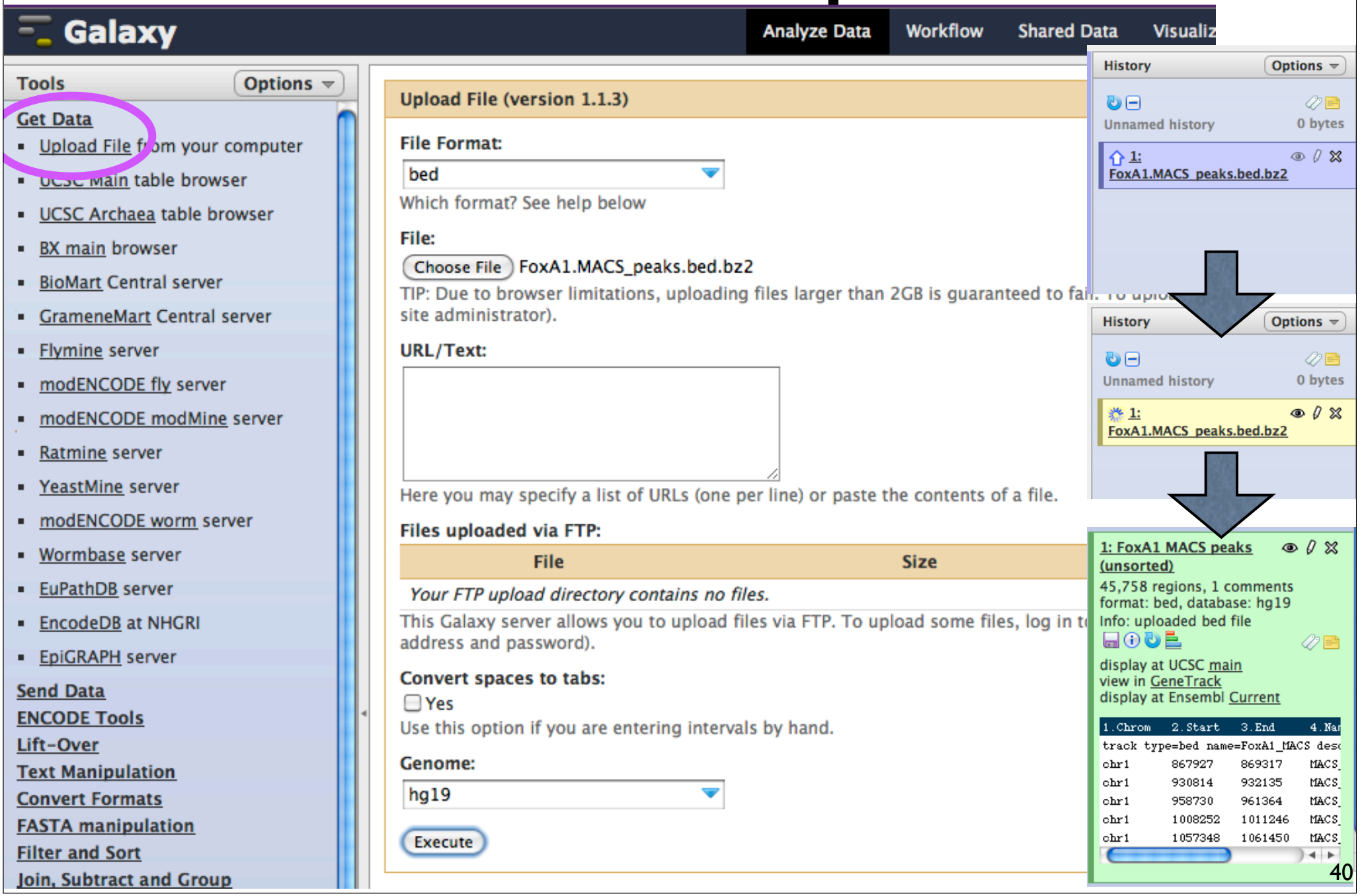
- Upload UCSC tracks of called peaks
- If it's your data, **eyeball the genome! Find the stories with your own eyes.**
  - Where are the peaks?
    - **DNA sequence motif?**
    - **Other correlated factors?**
    - **In regulatory regions?**
    - **Remotely from genes?**
    - **Which genes?**

Galaxy

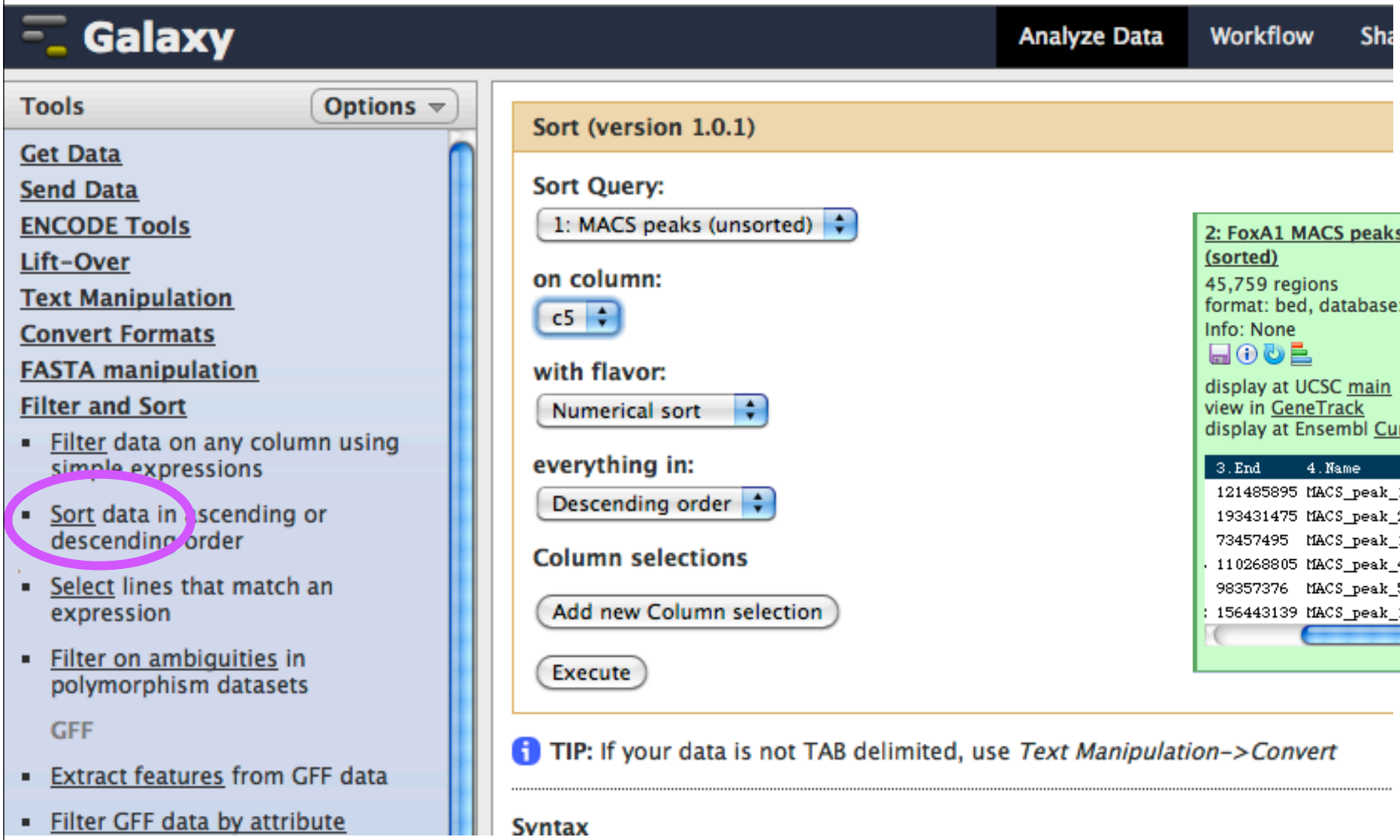


- Galaxy is an open, web-based platform for computational biomedical research.
- http://main.g2.bx.psu.edu

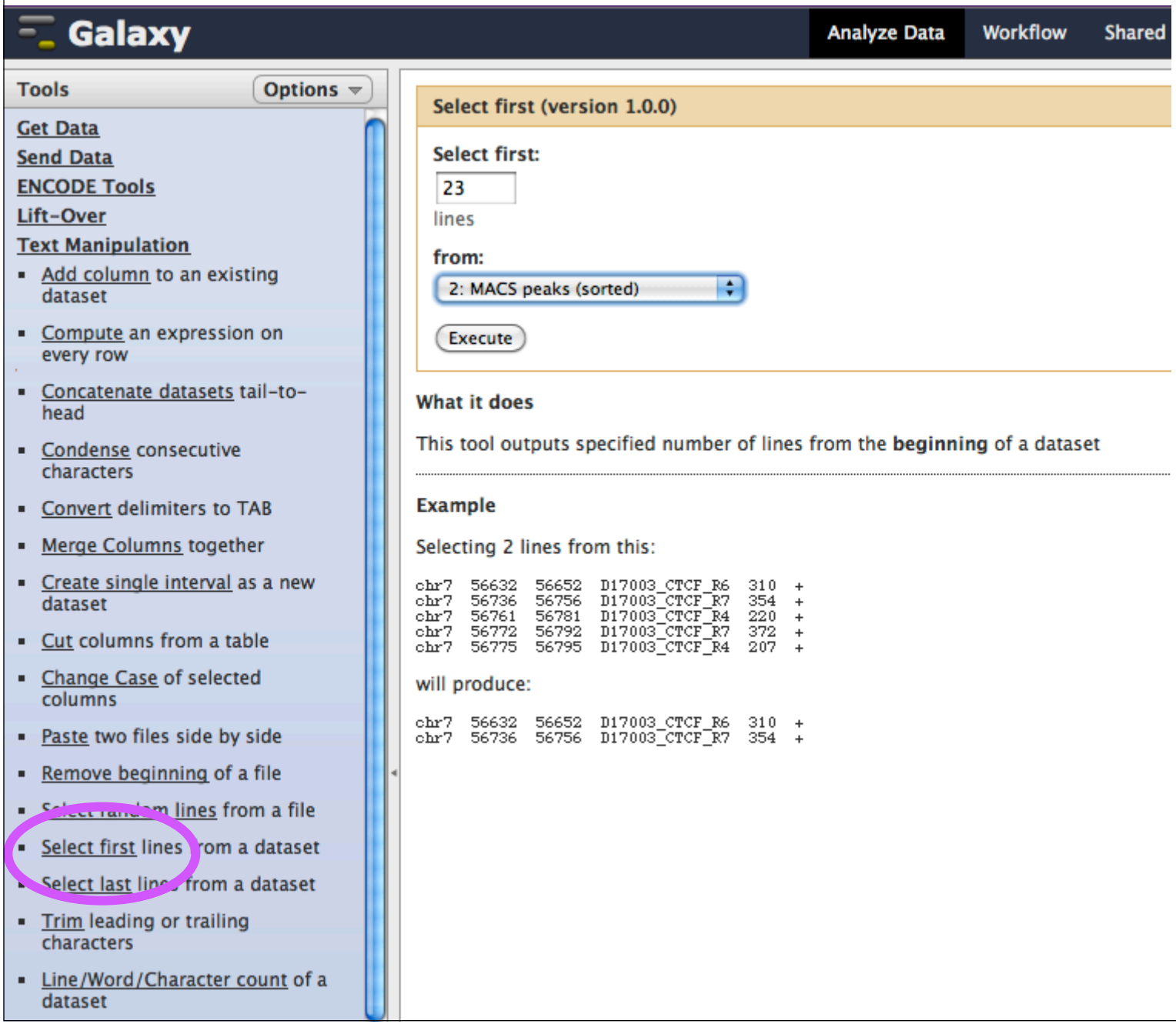
Load called peaks

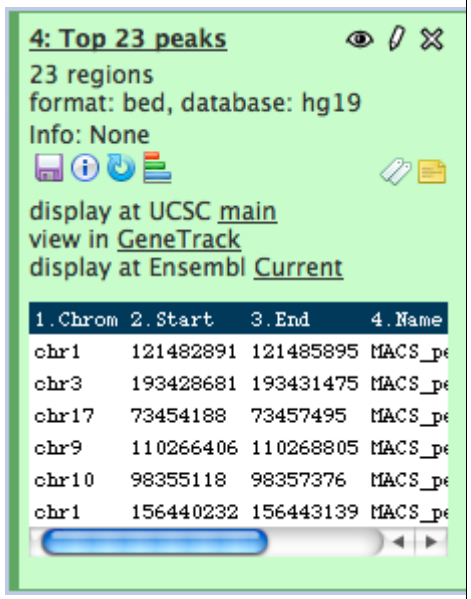


Sort by height

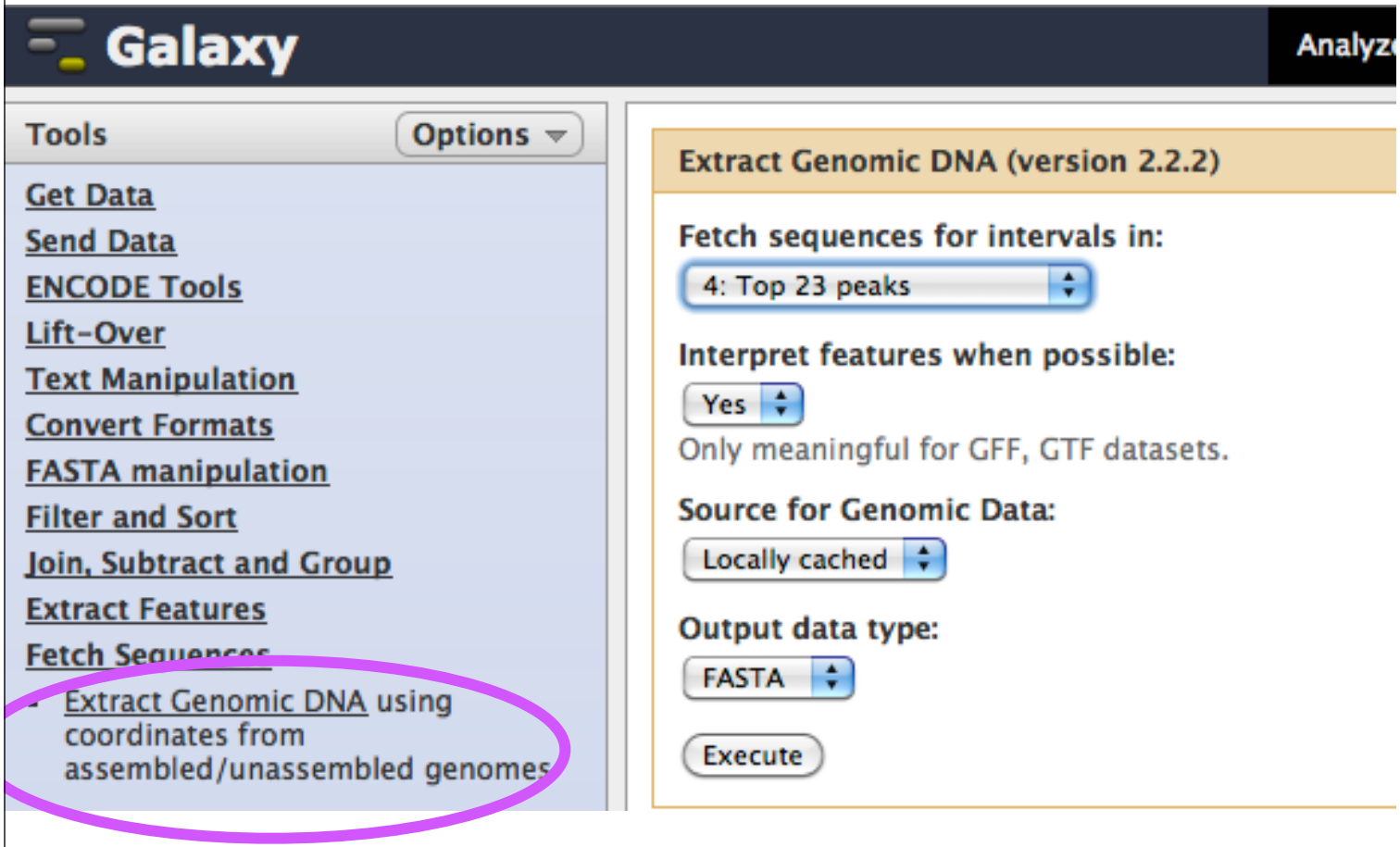


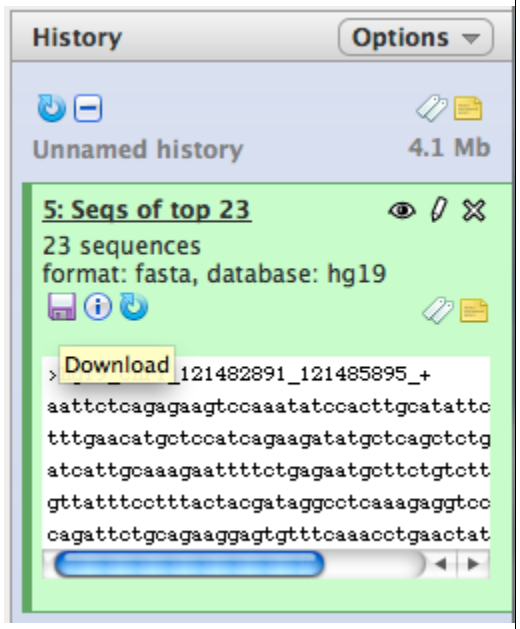
Select top 23 regions





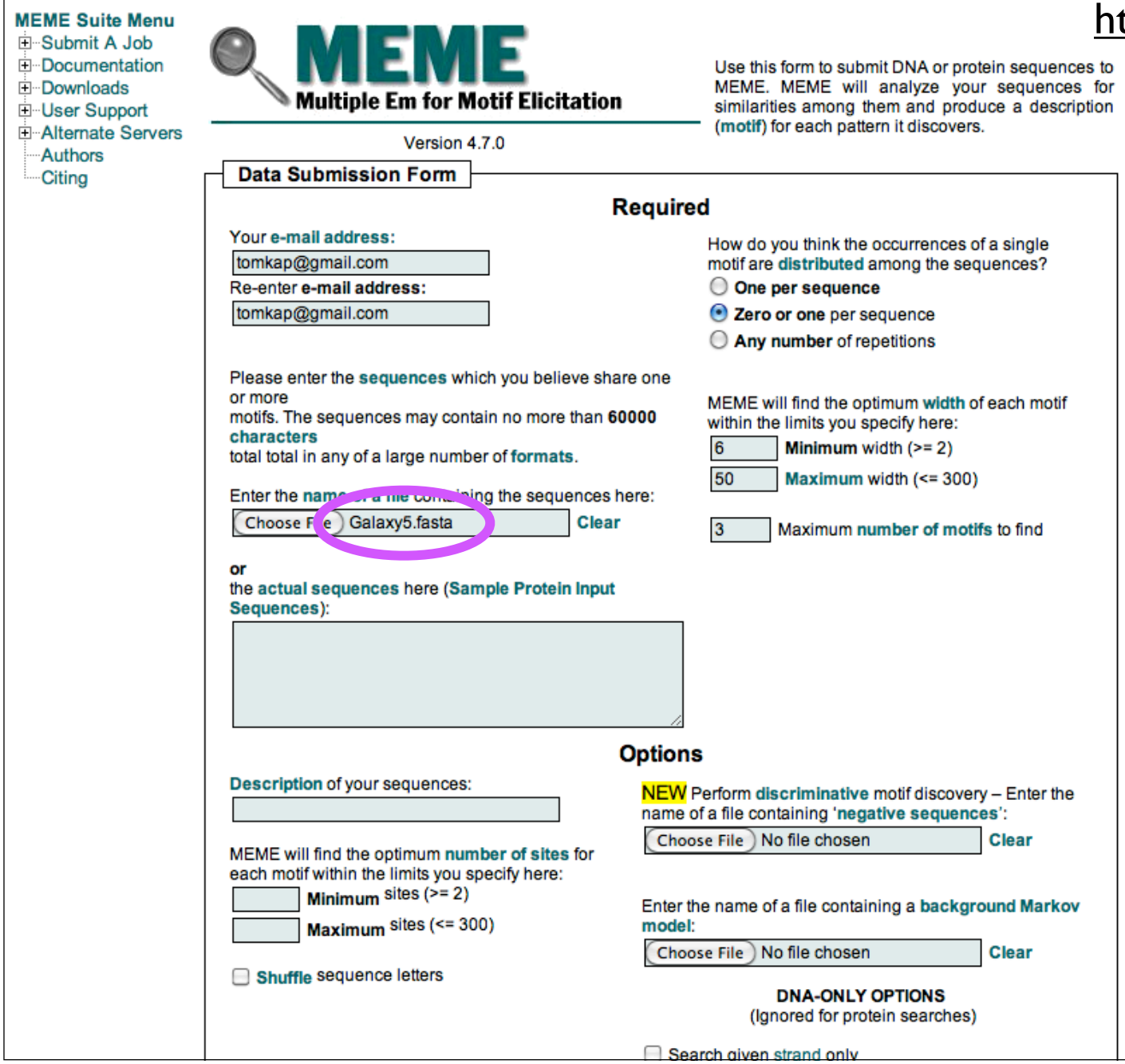
get genomic sequences





Look for motif with MEME

http://meme.sdsc.edu/meme/



* FASTA file from Galaxy had to be renamed for some reason

Look for motif with MEME

Your job id is: **app1318268648139**
You can view your job results at: http://meme.nbcr.net/meme4_7_0/cgi-bin/querystatus.cgi?jobid=app1318268648139&service=MEME
You can view server activity here.

- Sequence file: **Galaxy5.fasta**
- Distribution of motif occurrences: **Zero or one per sequence**
- Number of different motifs: **3**
- Minimum motif width: **6**
- Maximum motif width: **50**
- Statistics on your dataset:

type of sequence	**dna**
number of sequences	**23**
shortest sequence (residues)	**1493**
longest sequence (residues)	**3307**
average sequence length (residues)	**2540.8**
total dataset size (residues)	**58438**

You om.

ACTI

This pag

You can

The statu

View ser

ACT

This pa

You ca

The sta

View s

ACTIVE

This page will contain your job output when it is done.

You can bookmark it for later reference.

The status of your job will be checked again in 60 seconds.

View server activity.

A couple of hours later...

MEME Job app1318268648139

- meme sequences -sf Galaxy5.fasta -dna -mod zoops -nmotifs 3 -minw 6 -maxw 50 -time 7200 -maxsize 60000 -revcomp

Results

- MEME output as HTML
- MEME output as plain tex
- MEME output as XML
- MAST output as HTML
- MAST output as XML
- MAST output as text (forr
- input sequences

Messages

- Processing Messages
- Error Messages

MEME
Multiple Em for Motif Elicitation

For further information on how to interpret these results

If you use MEME in your research, please cite the followir
Timothy L. Bailey and Charles Elkan, "Fitting a mixture model b

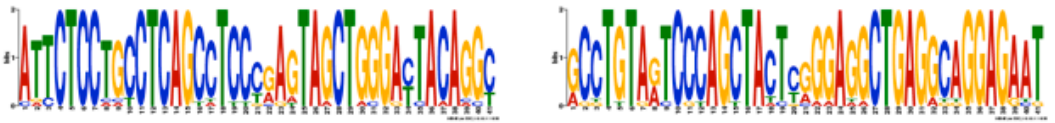
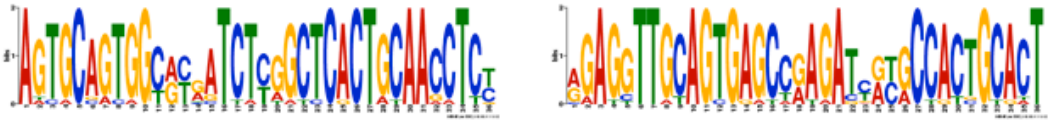
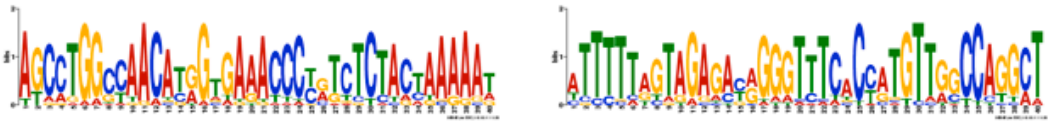
Click on any row to highlight sequence in all motifs.

Name	Strand	Start	p-value	Sites
hg19_chr17_1672500_1674182_+	-	75	7.57e-26	GGTTCACGCC ATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGC GCCCACCACC
hg19_chr1_156440232_156443139_+	+	2373	1.67e-25	AGTTCAAGTG ATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATTACAGGC ACATGCCACT
hg19_chr17_73454188_73457495_+	-	822	2.58e-25	GGTTCACGCC ATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGACTACAGGC ACCCGCCACA
hg19_chr9_110266406_110268805_+	+	1360	4.40e-25	GGTTCAAGCG ATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGT GCATGCTGCC
hg19_chr10_98622889_98625123_+	+	175	6.59e-25	GTTTCAAGTG ATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGATTACAGGC ATGCACCACC
hg19_chr17_61828830_61831978_+	-	2508	1.26e-23	GGTTCAAGCA ATTCTCCTGTCTCAGCCTCCTGAGTAGCTGGGATTACAGGC GTGCACCACC
hg19_chr9_128057870_128061033_+	+	2258	1.76e-23	GGTTCAAACA ATTCTCCTGCCTCAGCCTCCCGAGTAGCTAGGACTACAGGT GTACACCACC
hg19_chr19_35495955_35498474_+	-	1143	5.42e-23	GGTTCAAGTG ATTCTCCTGCCTCAGCCTCCGCAGTAGCTGGGATTACAGGC ACCCACCACC
hg19_chr20_55309000_55311968_+	+	589	6.77e-23	GGTTCAAGTG ATTCTCCTGCCTCAGCCTCCCGAATAGCTGGGATTACAAGC GCCTGCCACC
hg19_chr11_40349932_40351562_+	+	1463	7.97e-23	GGTTCAAGCG ATTCTCCTGCCTCAGCCTCTCGAGTAGCTGCGACTACAGGC ATTTGCCAAC
hg19_chr5_40446525_40448658_+	-	23	1.20e-22	GGTTCACGTC ATTCTCCTGCCTCAGCCTCCCAGGTAGCTGGGGCTACAGGC GCCCGCCCCA
hg19_chr19_54124486_54126761_+	-	1947	1.41e-22	GGTTCACGCC ATTCTCCTGCCTCAGCCTTCTGAATAGCTGGGACTACAGGC GCCTGCCACC
hg19_chr3_193428681_193431475_+	+	2112	2.52e-22	GGCTCAAGCT ATCCTCCAGCCTCAGCCTCCCAAGTAGCTGGGACTACAGGT GCATAAAACC
hg19_chr8_95540335_95542493_+	+	1636	3.97e-22	GGCTCAAGCA ATTCTCCTCCCTCAGTCTCCTGAGTAGCTGGGATTACAGGC GTGTAACACC
hg19_chr17_53509073_53512135_+	-	388	8.63e-22	GGCTCAAGTG CTCCTCCTGCCTCAGCTTCCCAAGTAGCTGGGACTACAGGC CTGGGCCACC
hg19_chr19_18140018_18141511_+	-	221	2.20e-19	GGTTCAAGCT AATCTCCCACCTCAGCCTCCCGAGTAGCTGGGACCACACCC AGATAATTAC
hg19_chr13_37977646_37979671_+	+	1266	1.15e-18	GGTTCAAGCA AGTCTCCTGTCTCAGCATCCCAATTAACTGGGATTAAAGGT ACCCATCACC

DISCOVERED MOTIFS | BLOCK DIAGRAMS OF MOTIFS | PROGRAM INFORMATION | EXPLANATION

DISCOVERED MOTIFS

Motif Overview

Motif 1	• 6.1e-206 • 17 sites	
Motif 2	• 6.3e-149 • 17 sites	
Motif 3	• 7.9e-133 • 21 sites	

Further Analysis

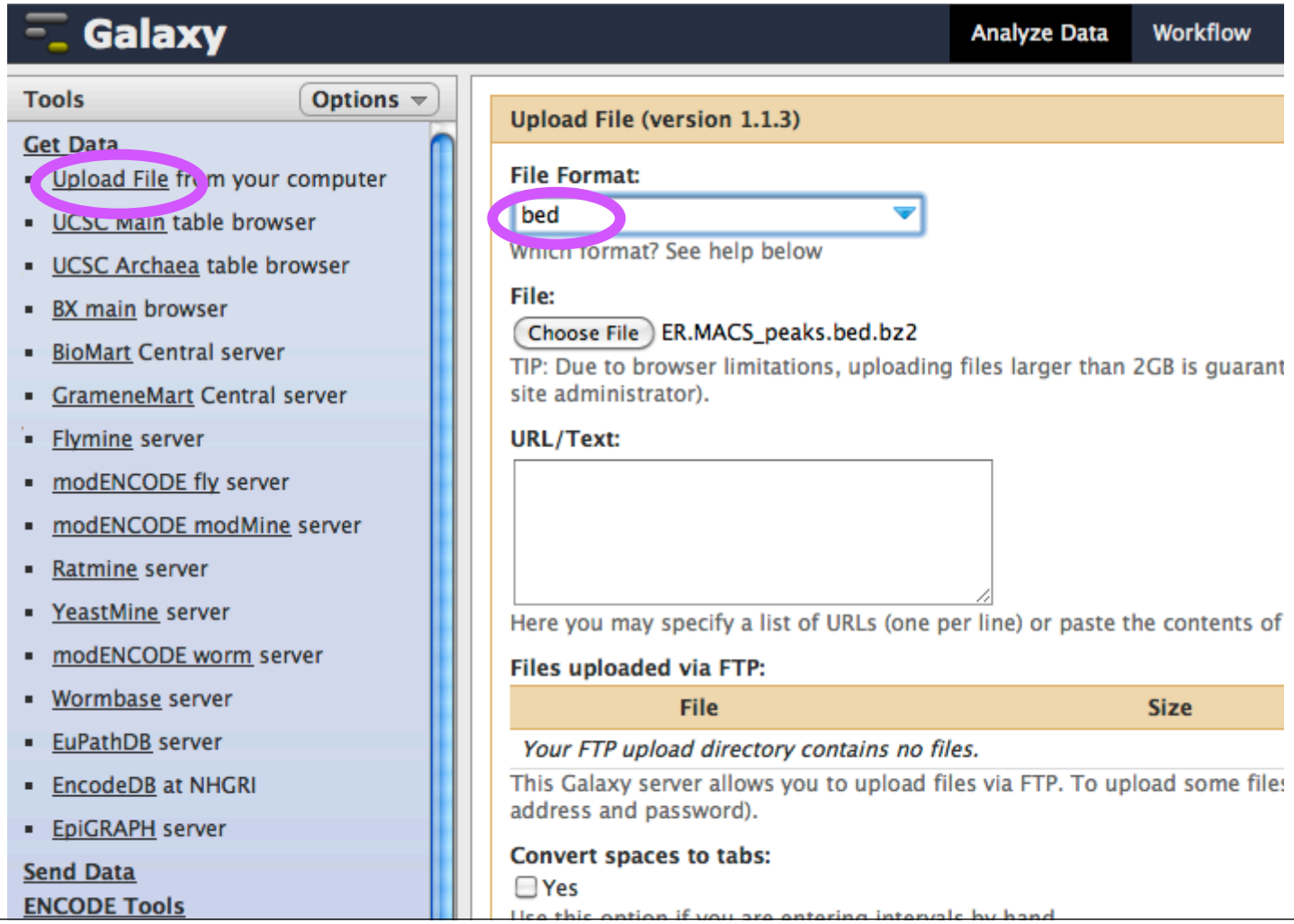
Submit all motifs to MAST FIMO GOMO BLOCKS Mouse-over buttons for more information.

Look for motif with MEME

- MEME is rather slow
- Try Weeder (Pavesi et al) for a much faster tool for enriched K-mers
- MEME allows up to 60,000 chars for web interface

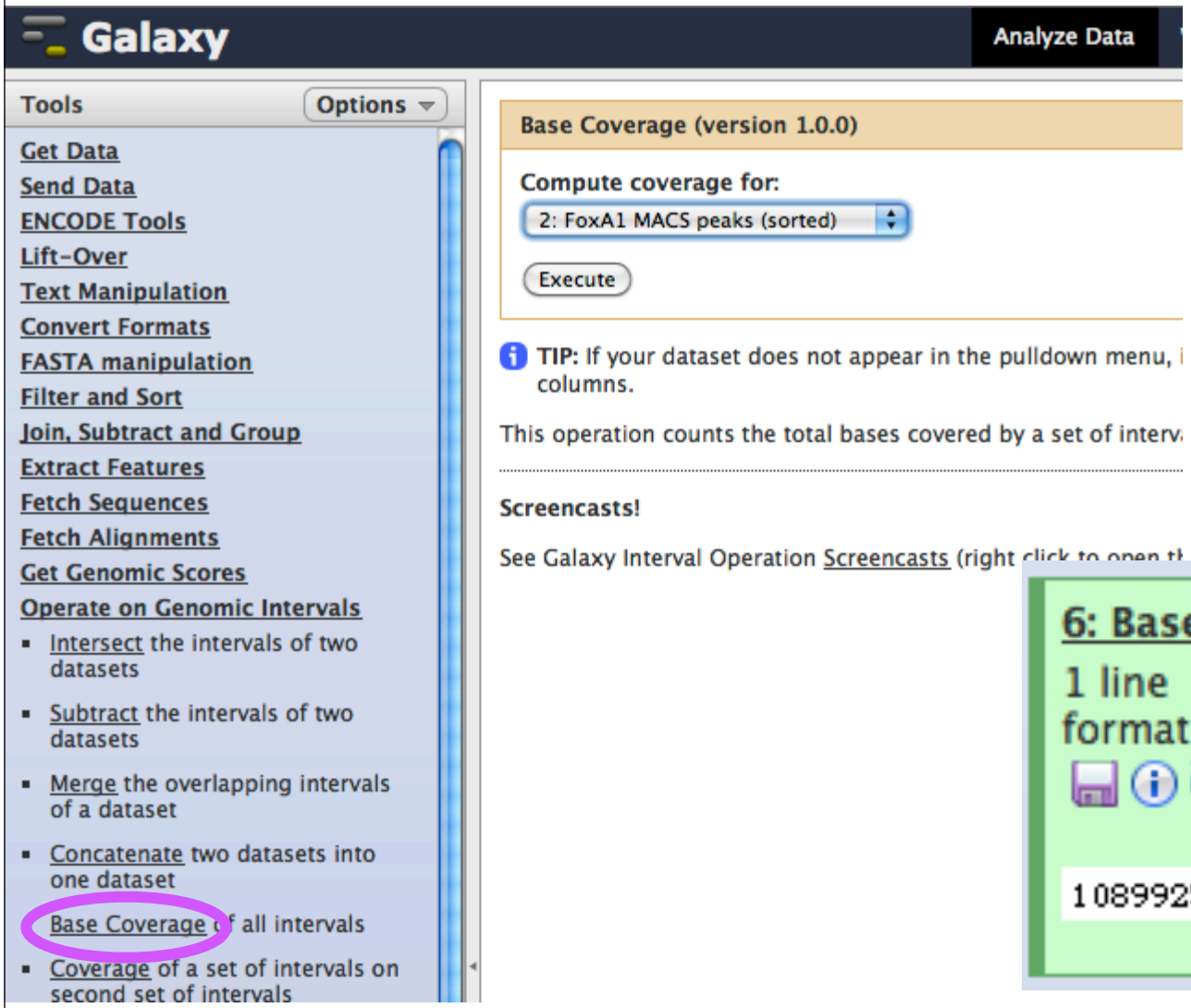
Galaxy

- Compare binding with other factors
- Add a second set of ChIP peaks



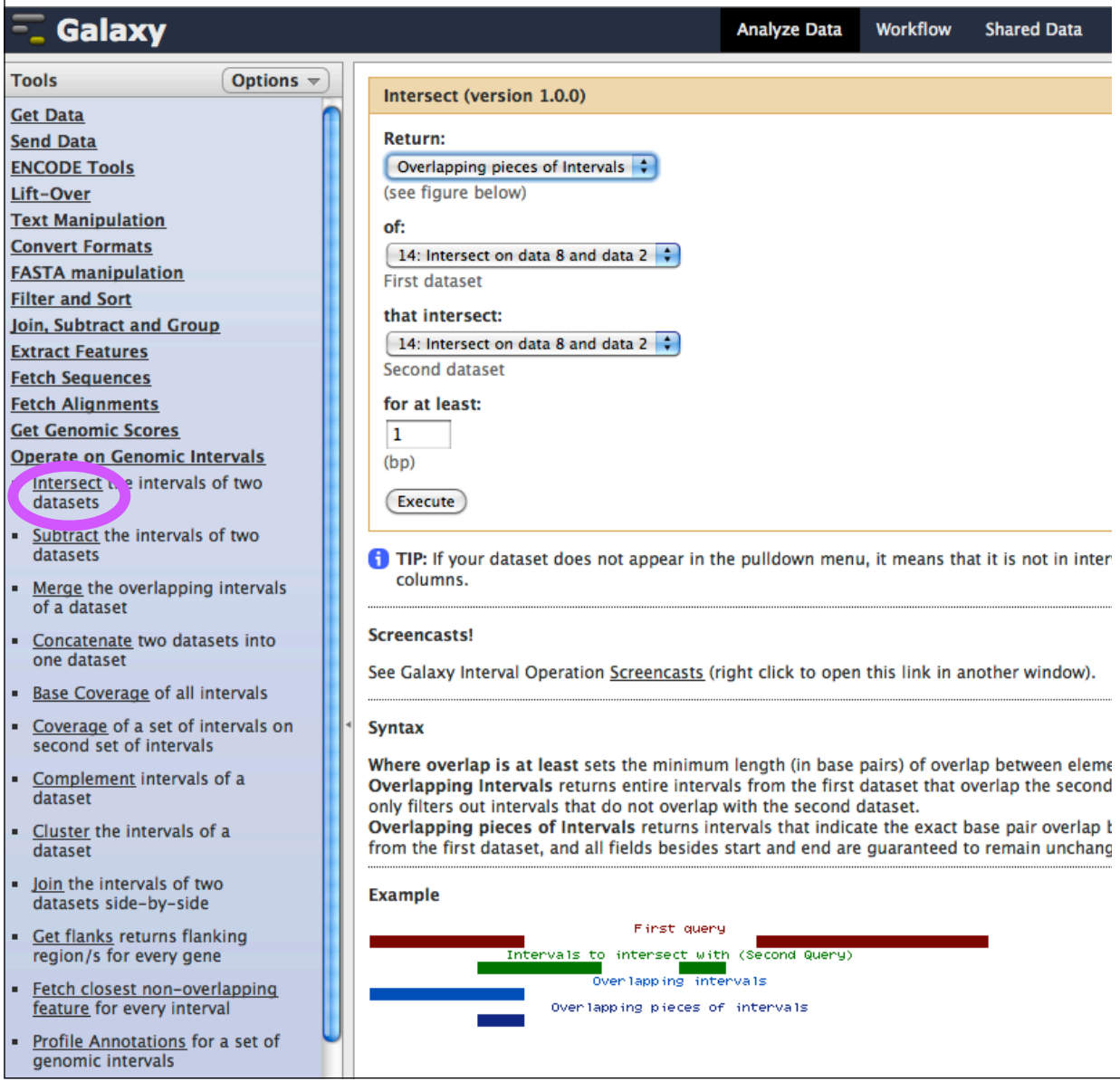
Galaxy

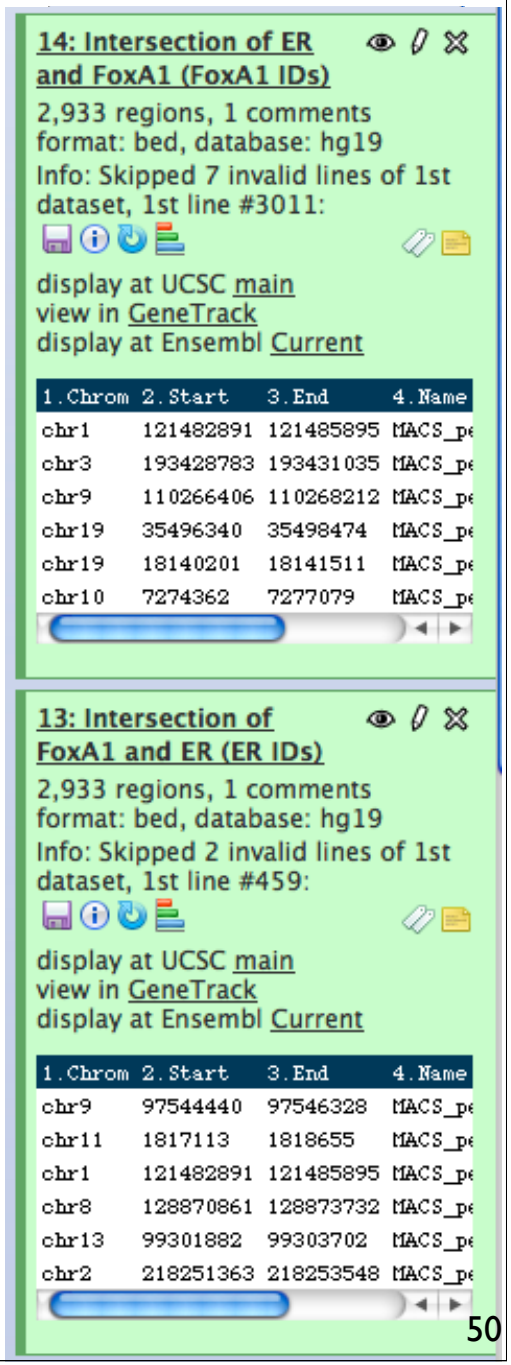
- Calculate base coverage for FoxA1



Galaxy

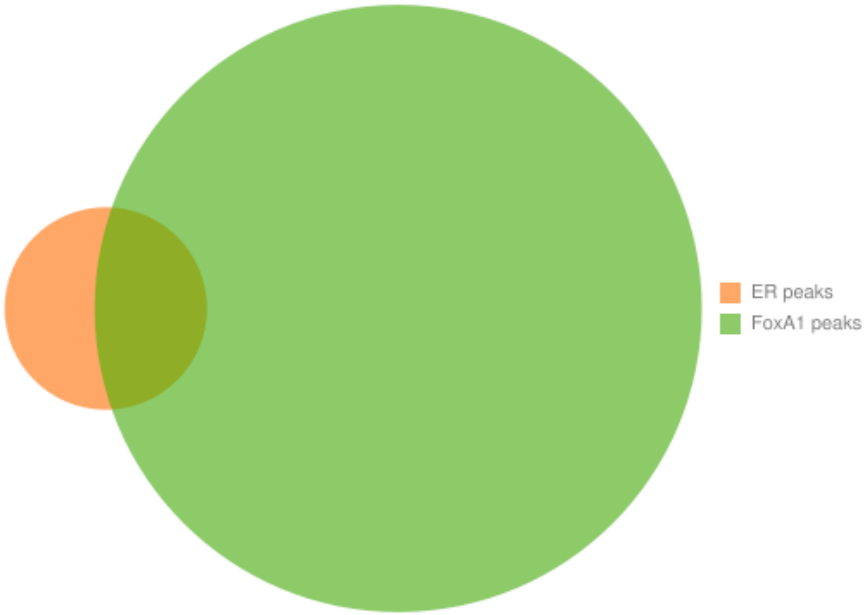
- Intersect two databases

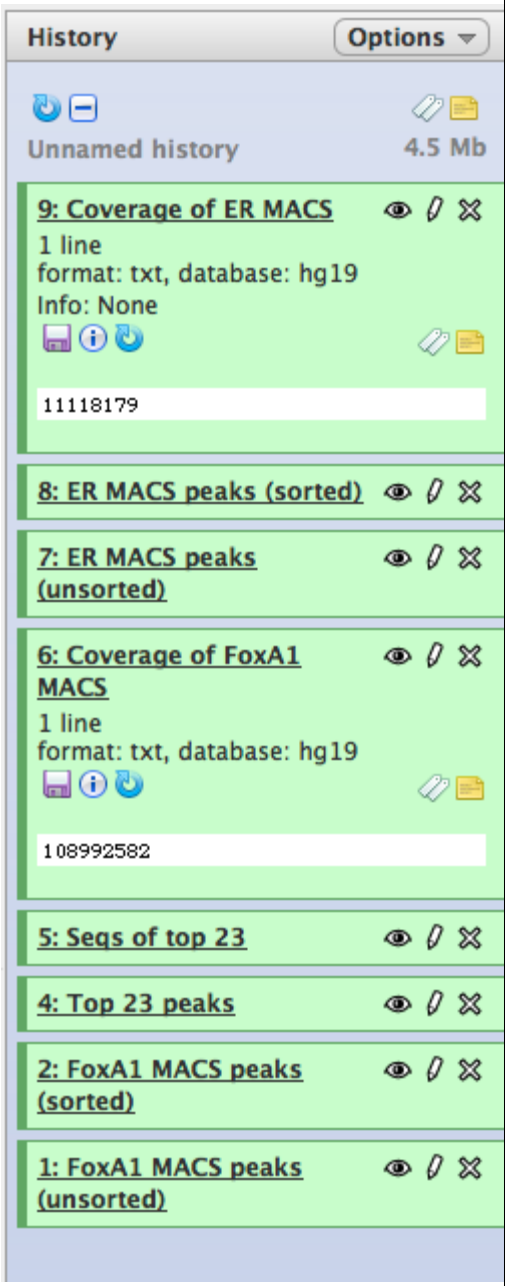




Galaxy

- Calculate base coverage for ER
- FoxA1 - 108,992,582 bp (3.6%)
- ER - 11,118,179 bp (0.37% of genome)
- ER & FoxA1 - 5,943,807 bp

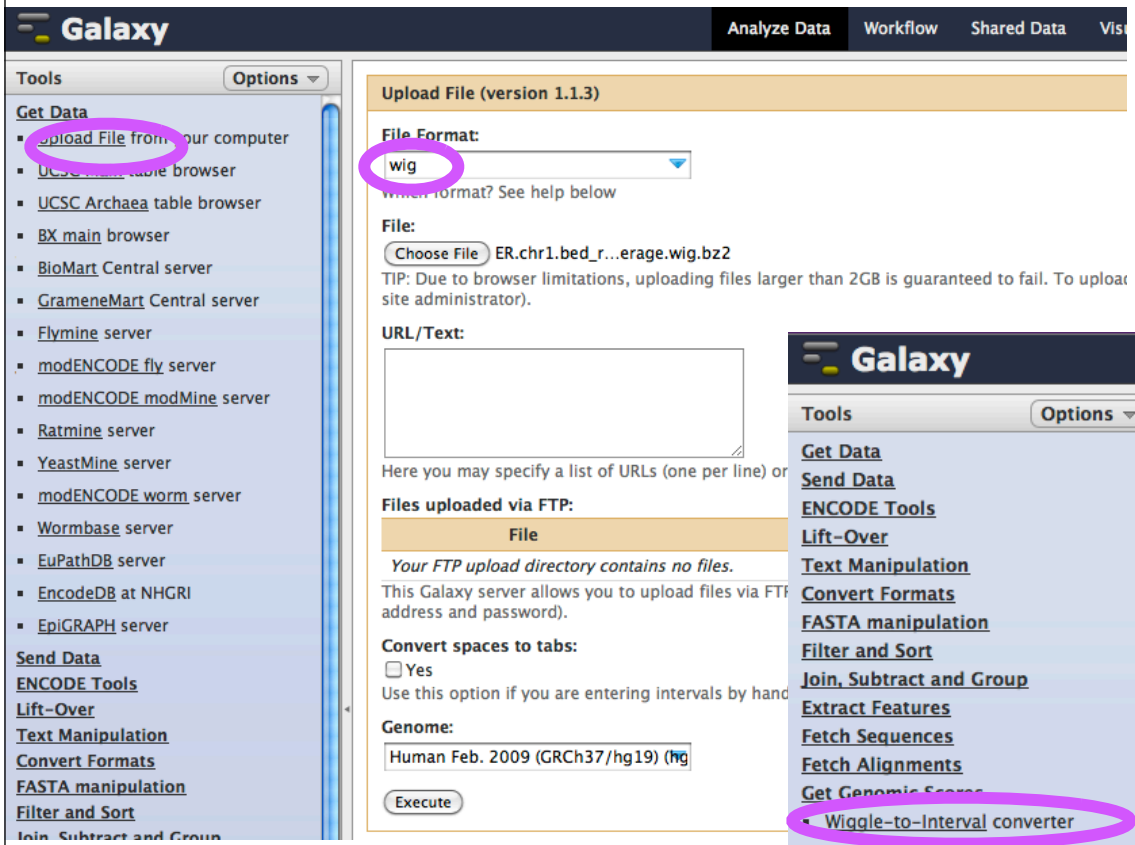


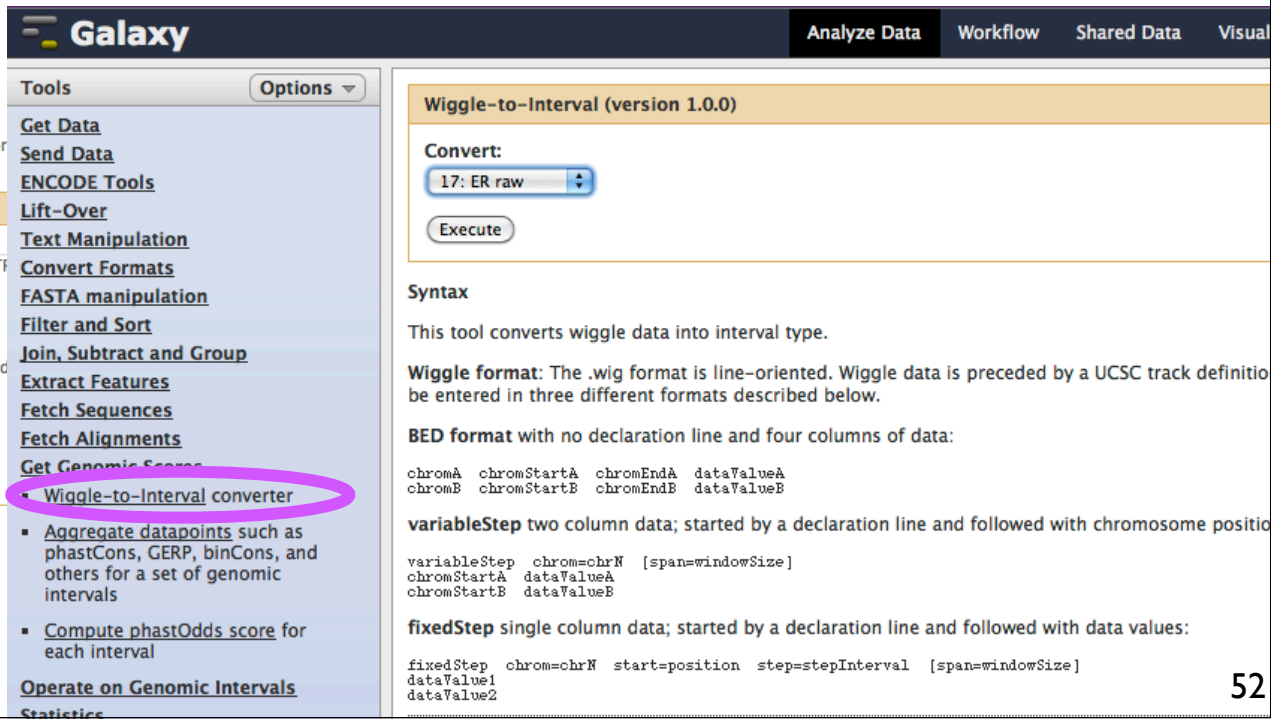


http://chart.apis.google.com/chart?cht=v&chd=t:11.118,108.99,0,5.9438&chs=500x500&chdl=ER%20peaks|FoxA1%20peaks

Galaxy

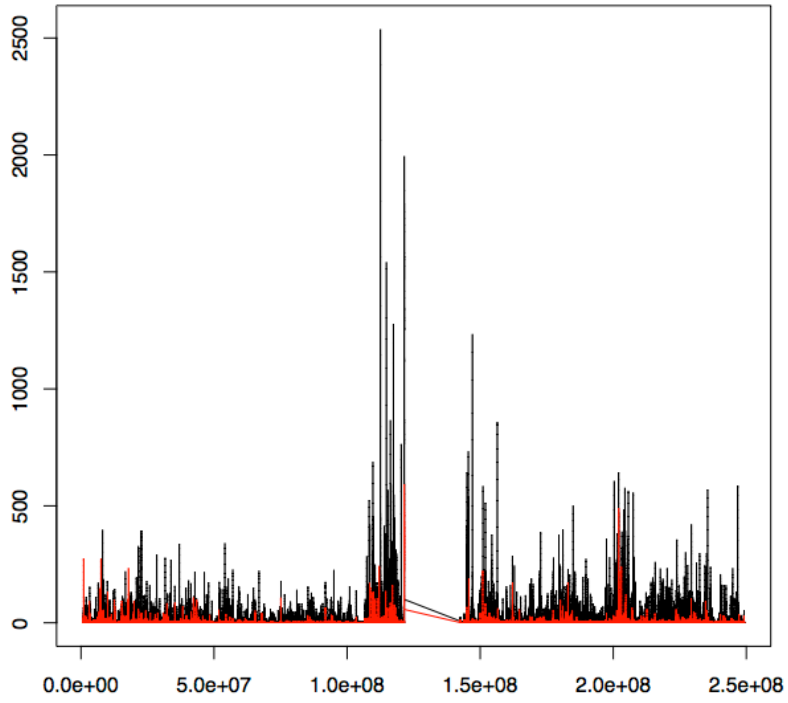
- Upload raw data, convert to intervals

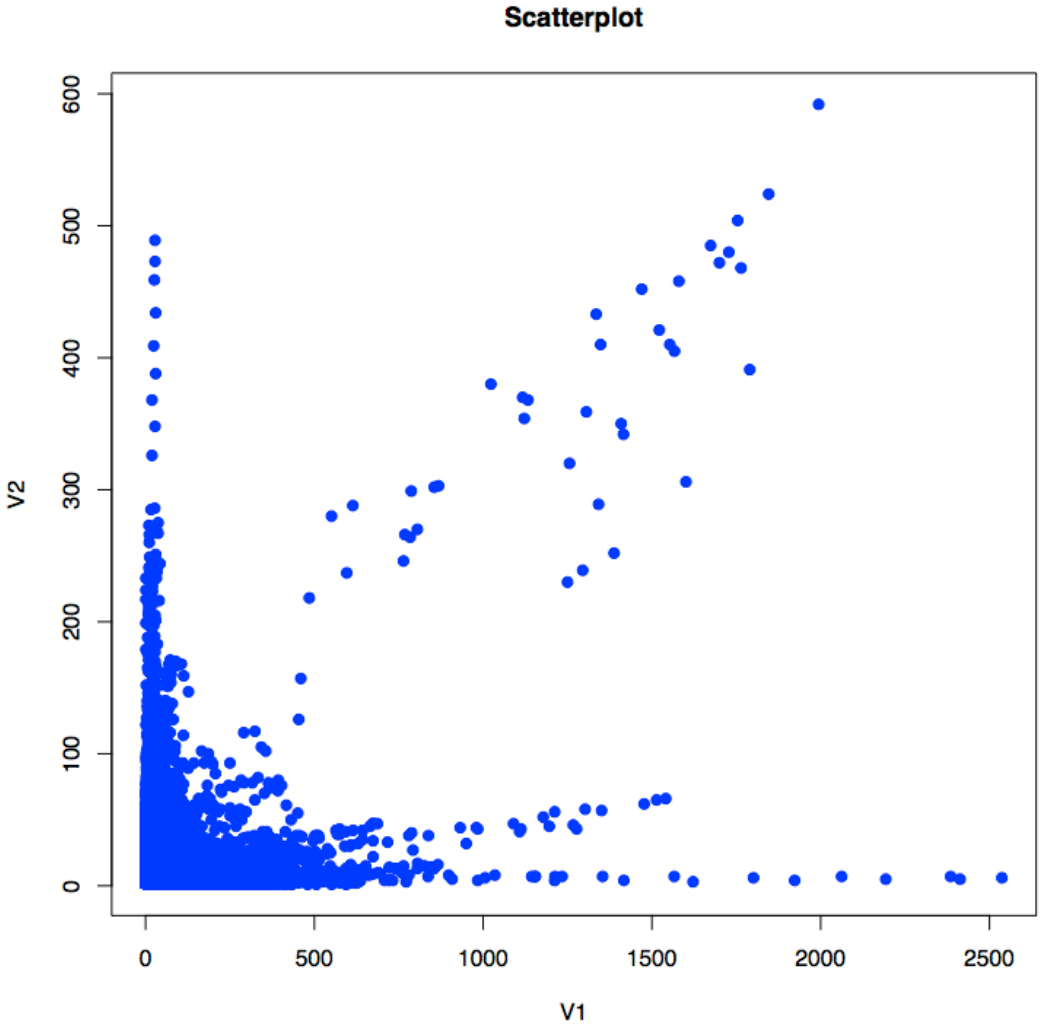




Galaxy

- Join two data sets
- Filter for intervals with values in both sets
- Scatterplot



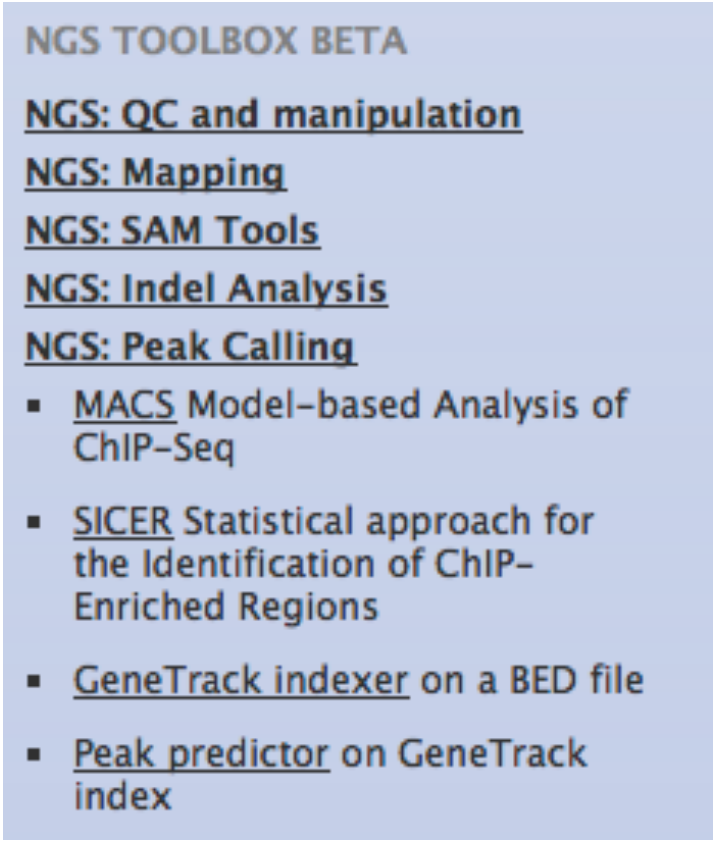


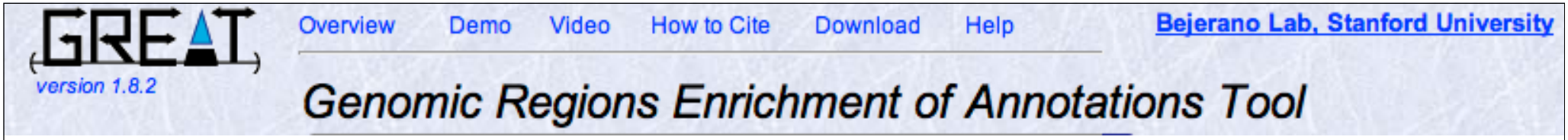
Galaxy

- Very lame genome browser

Galaxy

- In-house MACS caller and other options (beta version)
- Tons of other options.
- Takes some time to master, but offers a fast (limited) alternative to programming





http://great.stanford.edu

- Web server for predicting the functions of cis-regulatory regions

- Let's try the top 500 bound regions of ER (convert height to integer)

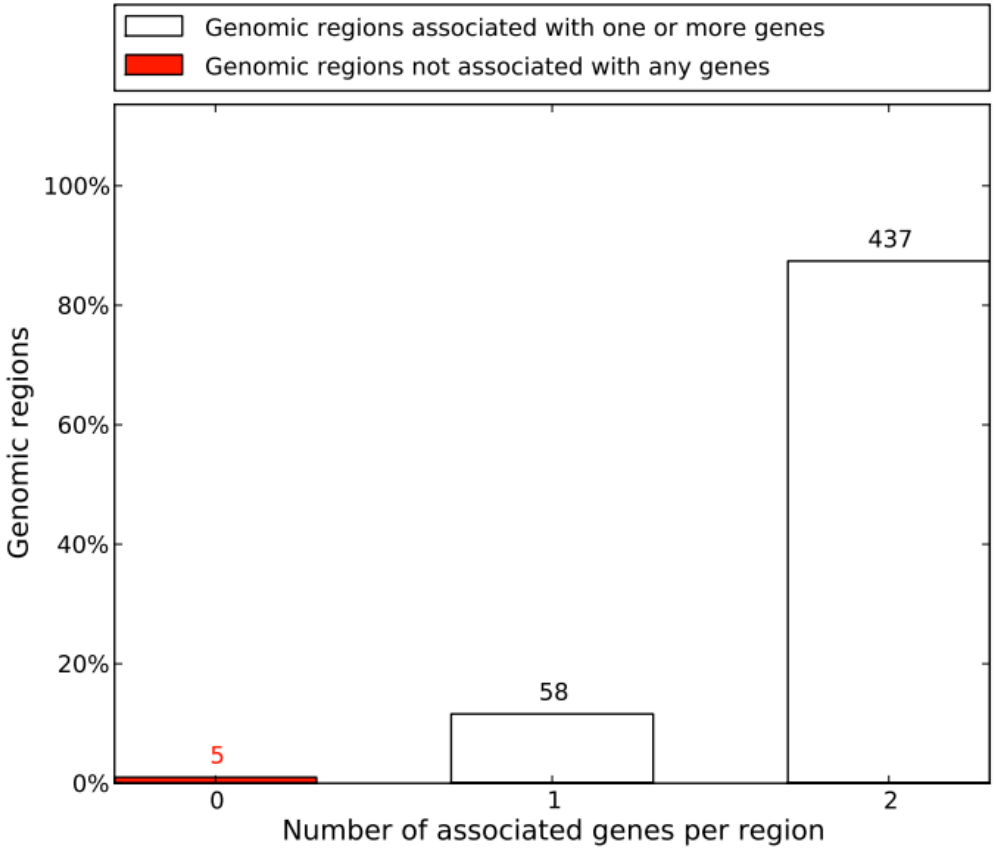
GREAT version 1.8.2

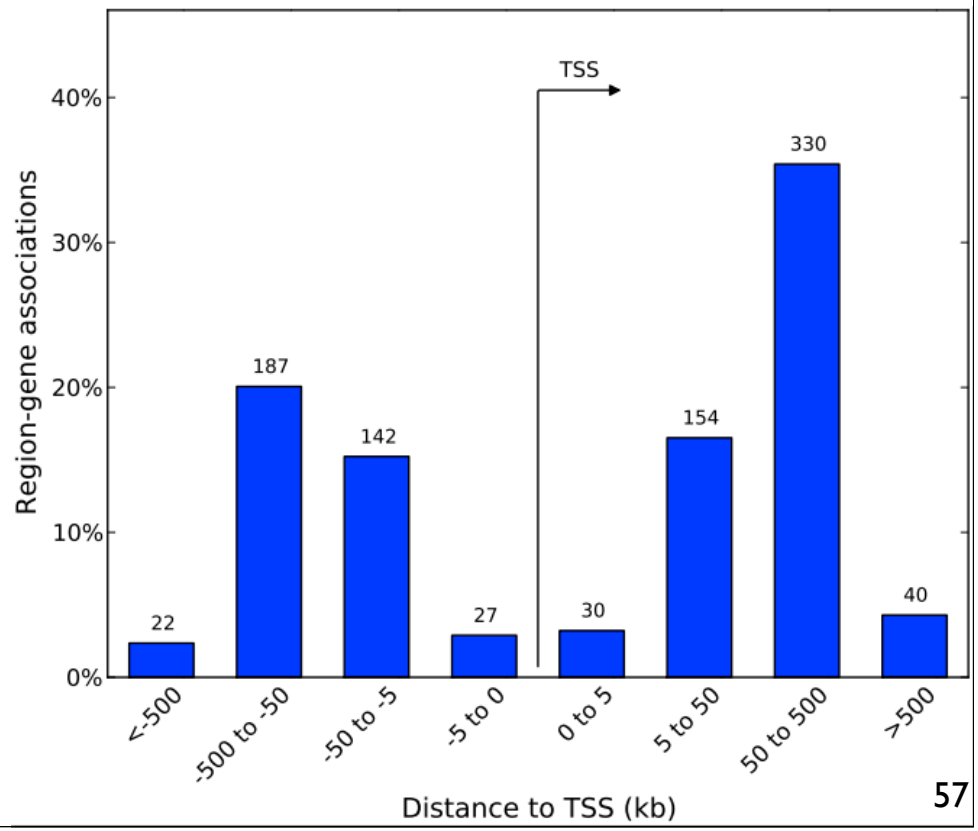
Overview Demo Video How to Cite Download Help

Bejerano Lab, Stanford University

Genomic Regions Enrichment of Annotations Tool

- Associate each region to nearest genes





GREAT version 1.8.2

Overview Demo Video How to Cite Download Help Bejerano Lab, Stanford University

Genomic Regions Enrichment of Annotations Tool

Genomic region -> gene association table

Download table as text.

Region	Gene (distance to TSS)
MACS_peak_2412	GREB1 (-35,170), E2F6 (-32,775)
MACS_peak_4803	C9orf3 (+56,361), FANCC (+534,636)
MACS_peak_729	SYT8 (-37,790), HCCA2 (-32,383)
MACS_peak_225	FCGR1B (-548,445)
MACS_peak_4647	MYC (+123,952)
MACS_peak_1278	STK24 (-73,396), SLC15A1 (+102,137)
MACS_peak_53	RCC2 (-82,010), ARHGEF10L (-19,263)
MACS_peak_2692	TNP1 (-527,674), TNS1 (+556,340)
MACS_peak_3036	ZNRF3 (-69,738), XBP1 (-13,592)
MACS_peak_2289	ISYNA1 (-8,871), ELL (+75,123)
MACS_peak_1945	IGFBP4 (+4,959), TNS4 (+53,219)
MACS_peak_1143	TMEM120B (+36,403), RHOF (+44,533)
MACS_peak_2202	STK11 (-24,243), SBNO2 (-7,273)
MACS_peak_1844	KLHDC4 (-24,301), SLC7A5 (+79,257)
MACS_peak_1048	DYRK2 (-165,170), CAND1 (+214,281)
MACS_peak_4182	KIAA0415 (-92,067), FOXK1 (+1,267)
MACS_peak_1947	CCR7 (-11,276), SMARCE1 (+71,103)
MACS_peak_2977	TMPRSS2 (-145,861), RIPK4 (+161,396)
MACS_peak_2194	ATP9B (-6,190), SALL3 (+82,932)
MACS_peak_3705	IL6ST (-434,434), MAP3K1 (-385,703)
MACS_peak_1395	ZFP36L1 (+222,939), RAD51L1 (+750,337)
MACS_peak_2348	CYP2B6 (-2,800)
MACS_peak_863	NARS2 (-485,489), ODZ4 (+380,297)

Gene -> genomic region association table

Download table as text.

Gene	Region (distance to TSS)
A4GALT	MACS_peak_3086 (-14,035)
ABAT	MACS_peak_1682 (+4,572)
ABHD2	MACS_peak_1606 (+27,679)
ACSL4	MACS_peak_5022 (-218,800)
ACTL7B	MACS_peak_4841 (+377,506)
ACTN4	MACS_peak_2336 (+66,208)
ACTR2	MACS_peak_2499 (+72,801)
ACTR3B	MACS_peak_4392 (-60,191)
ADA	MACS_peak_2853 (-41,825)
ADAMTS12	MACS_peak_3680 (+341,842)
ADAMTSL5	MACS_peak_2207 (-450)
ADCY9	MACS_peak_1679 (-36,649)
ADD2	MACS_peak_2513 (+151,622)
ADHFE1	MACS_peak_4513 (+80,520), MACS_peak_4514 (+90,242)
ADORA3	MACS_peak_195 (-9,976)
ADRA2C	MACS_peak_3444 (+60,732)
ADRBK1	MACS_peak_811 (-17,343)
AFF3	MACS_peak_2550 (+468,876)
AGRN	MACS_peak_2 (+53,567)
AHCYL1	MACS_peak_189 (-33,761)
AK3L1	MACS_peak_138 (-164,345)
ALG8	MACS_peak_858 (-16,945)
AMZ1	MACS_peak_4178 (+9,213)
ANLN	MACS_peak_4222 (-116,421)
ANO1	MACS_peak_833 (-216,777)

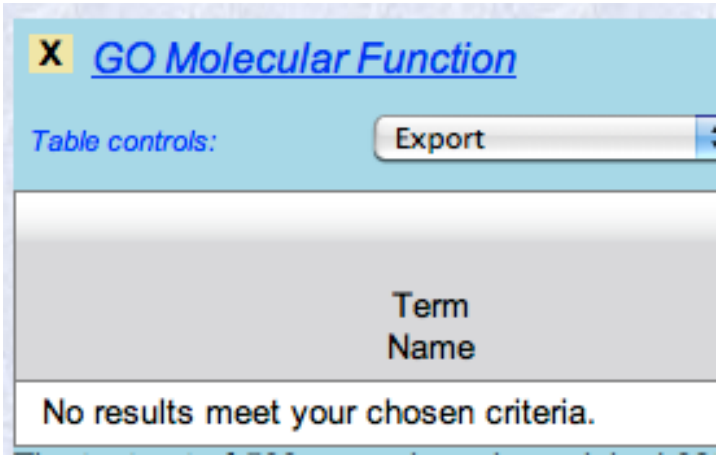
GREAT version 1.8.2

Overview Demo Video How to Cite Download Help

Bejerano Lab, Stanford University

Genomic Regions Enrichment of Annotations Tool

- Then tests the associated genes for enrichment



X GO Molecular Function

Table controls: Export

Term Name
No results meet your chosen criteria.

X Transcription Factor Targets

Table controls: Export Shown top rc

Term Name	Binom Rank	Binom Raw P-Value
Targets of estrogen receptor alpha, identified by ChIP-DSL in MCF-7 cells	1	3.5502e-9

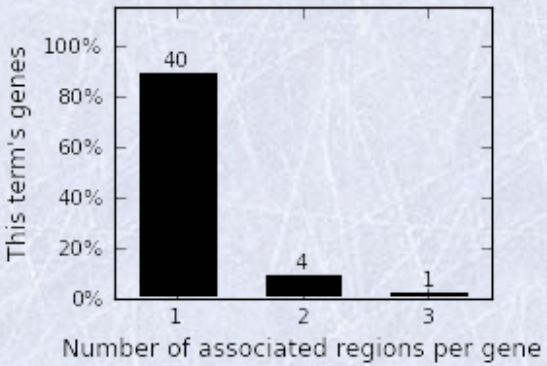
X GO Biological Process

Table controls: Export Shown top r

Term Name	Binom Rank	Binom Raw P-Value
gland development	49	2.9902e-5
mammary gland development	116	2.3782e-4

Number of associated regions per gene

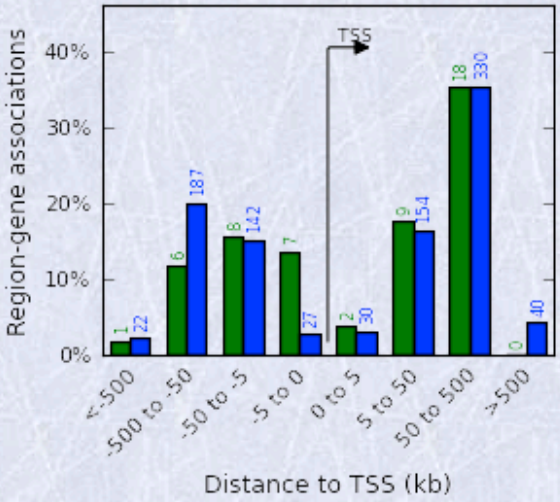
Download as PDF.



Binned by orientation and distance to TSS

This term's region-gene associations
Set-wide region-gene associations

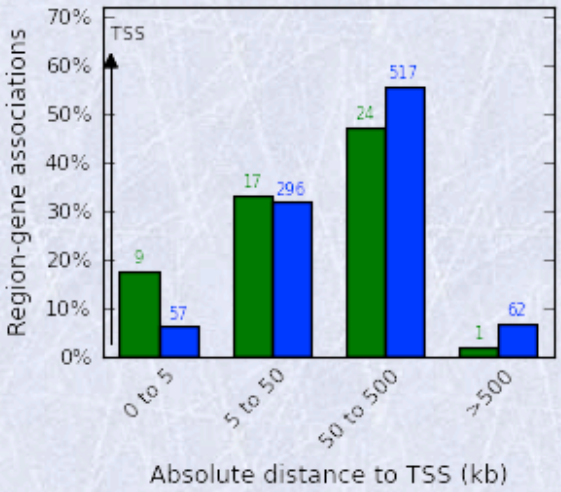
Download as PDF.



Binned by absolute distance to TSS

This term's region-gene associations
Set-wide region-gene associations

Download as PDF.



back to top

This term's genomic region-gene association tables

What do these tables show?

This term's genomic region -> gene association table

Download table as text.

Region	Gene (distance to TSS)
MACS_peak_2412	GREB1 (-35,170)
MACS_peak_53	ARHGEF10L (-19,263)
MACS_peak_1048	CAND1 (+214,281)
MACS_peak_2900	TFAP2C (+105,821)
MACS_peak_2415	GREB1 (+5,985)
MACS_peak_4127	ESR1 (-190,906)
MACS_peak_3899	FOXC1 (+31,027)
MACS_peak_4082	TPD52L1 (+1,254)
MACS_peak_3869	STC2 (-206,271), BOD1 (+80,889)
MACS_peak_1715	USP31 (+69,073)
MACS_peak_3062	C1QTNF6 (-9,369)
MACS_peak_662	CASP7 (+115)
MACS_peak_3238	TIRAP (+138,056)

This term's gene -> genomic region association table

Download table as text.

Gene	Region (distance to TSS)
ADAMTSL5	MACS_peak_2207 (-450)
AMZ1	MACS_peak_4178 (+9,213)
ARHGEF10L	MACS_peak_53 (-19,263)
BCL3	MACS_peak_2357 (-5,948)
BOD1	MACS_peak_3869 (+80,889), MACS_peak_3868 (+143,681)
C1QTNF6	MACS_peak_3062 (-9,369)
CALM1	MACS_peak_1436 (+106,963)
CAND1	MACS_peak_1048 (+214,281)
CASP7	MACS_peak_662 (+115)
DCAKD	MACS_peak_1964 (+8,243)
DICER1	MACS_peak_1456 (+85,031)
ELF3	MACS_peak_333 (+47,442),

QuEST

- Be sure to check QuEST - UNIX-based software for analysis of ChIP-Seq data
- By Anton Valouev, a postdoc in Arend Sidow's lab at Stanford
- http://www.stanford.edu/~valouev/QuEST/QuEST.html

Some practice...

1. Download the read coverage landscape for FoxA1 and ER (chr1)
2. Upload the two datasets to UCSC (save as a session)
3. Download output of BOWTIE for ER and FoxA1 (chr1) and run MACS, upload peaks to UCSC browser
4. Compare the peaks to the landscape. Should the parameters be changed? Rerun if needed.

cont./

Some practice...

5. Upload MACS-called peaks to Galaxy
6. Get top 1000 regions for FoxA1, ER
7. Compute intersection (and Venn diagram) - compare to the intersection of the entire sets

8. Download Breast Cancer DNaseI accessibility data (already mapped using BOWTIE)
9. Run MACS and find bound regions
10. Do ER and FoxA1 peaks tend to occur in highly accessible DNA regions?

Thanks!

Tommy Kaplan
tomkap@berkeley.edu