

# 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>

The screenshot displays the NCBI GEO website. At the top, the NCBI logo is on the left and the GEO logo (Gene Expression Omnibus) is on the right. Below the logos is a navigation bar with links: GEO Publications, FAQ, MIAME, and Email GEO. A 'Login' link is also present. A descriptive paragraph about the Gene Expression Omnibus follows. The main content area is divided into three sections: 'GEO navigation', 'Site contents', and 'Submitter login'. The 'GEO navigation' section has two main tabs: 'QUERY' and 'BROWSE'. Under 'QUERY', there are three search options: 'DataSets', 'Gene profiles', and 'GEO accession', each with a text input field and a 'GO' button. Under 'BROWSE', there are two main categories: 'DataSets' and 'GEO accessions'. 'DataSets' has a 'GO' button. 'GEO accessions' has three sub-categories: 'Platforms', 'Samples', and 'Series'. The 'Platforms' button is circled in purple. The 'Site contents' section on the right lists 'Public data' (Platforms: 9,370; Samples: 624,542; Series: 25,116; DataSets: 2,720) and 'Documentation' (Overview, FAQ, Find, Submission guide, Linking & citing, Journal citations, Construct a Query, Programmatic access, DataSet clusters, GEO announce list, Data disclaimer, GEO staff). Below 'Documentation' is a 'Query & Browse' section with links to Repository browser, SAGEmap, FTP site, GEO Profiles, and GEO DataSets. At the bottom is a 'Submitter login' section with fields for 'User id' and 'Password', and links for 'New account' and 'Recover password'. A 'LOGIN' button is at the bottom of the login section.

NCBI » GEO

**Gene Expression Omnibus:** a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. [More information »](#)

**GEO navigation**

**QUERY**

- DataSets  **GO**
- Gene profiles  **GO**
- GEO accession  **GO**
- GEO BLAST

**BROWSE**

- DataSets **GO**
- GEO accessions
  - Platforms**
  - Samples
  - Series

**Site contents**

**Public data**

Platforms	9,370
Samples	624,542
Series	25,116
DataSets	2,720

**Documentation**

- Overview | FAQ | Find
- Submission guide
- Linking & citing
- Journal citations
- Construct a Query
- Programmatic access
- DataSet clusters
- GEO announce list
- Data disclaimer
- GEO staff

**Query & Browse**

- Repository browser
- SAGEmap
- FTP site
- GEO Profiles
- GEO DataSets

**Submit**

- New account

**Submitter login**

User id:

Password:

**LOGIN**

- » New account
- » Recover password

# Illumina Genome analyzer: hundreds of species

NCBI

Gene Expression Omnibus

GEO Publications

FAQ

MIAME

Email GEO

Login

NCBI » GEO » Repository browser » Platforms

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 ( <i>Aedes aegypti</i> )	high-throughput sequencing	<i>Aedes aegypti</i>				GEO	Sep 20, 2011
GPL14595	Illumina Genome Analyzer II ( <i>Aedes albopictus</i> )	high-throughput sequencing	<i>Aedes albopictus</i>				GEO	Sep 20, 2011
GPL14367	Illumina Genome Analyzer II ( <i>Bursaphelenchus xylophilus</i> )	high-throughput sequencing	<i>Bursaphelenchus xylop...</i>				GEO	Sep 01, 2011
GPL14368	Illumina Genome Analyzer II ( <i>Bursaphelenchus mucronatus</i> )	high-throughput sequencing	<i>Bursaphelenchus mucr...</i>				GEO	Sep 01, 2011
GPL14364	Illumina Genome analyzer II ( <i>Escherichia coli</i> E24377A)	high-throughput sequencing	<i>Escherichia coli</i> E24377A				GEO	Aug 31, 2011
GPL14358	Illumina Genome Analyzer II ( <i>Drosophila willistoni</i> )	high-throughput sequencing	<i>Drosophila willistoni</i>				GEO	Aug 29, 2011
GPL14176	Illumina Genome Analyzer II ( <i>Tetranychus urticae</i> )	high-throughput sequencing	<i>Tetranychus urticae</i>				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 ( <i>Drosophila</i> )	high-throughput sequencing	<i>Drosophila</i>				GEO	Aug 11, 2011
GPL14012	Illumina Genome Analyzer II ( <i>Limanda limanda</i> )	high-throughput sequencing	<i>Limanda limanda</i>				GEO	Aug 02, 2011
GPL14002	Illumina Genome Analyzer II ( <i>Herpesvirus saimiri</i> (strain 11))	high-throughput sequencing	<i>Herpesvirus saimiri</i> (str...		1	1	GEO	Jul 30, 2011
GPL13999	Illumina Genome Analyzer II ( <i>Arachis hypogaea</i> )	high-throughput sequencing	<i>Arachis hypogaea</i>				GEO	Jul 29, 2011
GPL14000	Illumina Genome Analyzer II ( <i>Phaseolus vulgaris</i> )	high-throughput	<i>Phaseolus vulgaris</i>				GEO	Jul 29, 2011

# > 1 000 human samples

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



Scope:  Format:  Amount:  GEO accession:  
**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 HMG1 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 FoxA1, ER & CTCF



[nature.com](#) ▶ [journal home](#) ▶ [archive](#) ▶ [issue](#) ▶ [article](#) ▶ [full text](#)

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](#)

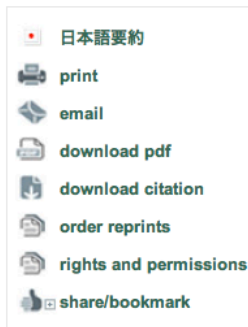
*Nature Genetics* **43**, 27–33 (2011) | doi:10.1038/ng.730

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- $\alpha$  (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 HNF3 $\alpha$ ). 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.



# DNase hyper-sensitivity



[nature.com](#) ▶ [journal home](#) ▶ [archive](#) ▶ [issue](#) ▶ [opinion and comment](#) ▶ [commentary](#) ▶ [full text](#)

[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.

Scope:  Format:  Amount:  GEO accession:

### 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) [GSE16256](#) UCSD Human Reference Epigenome Mapping Project

[Less...](#)

- [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
AAATGTTAATCTGAANAGCTGGAATCCAGTCTGGTGTGTTGTA
+FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
BACCBBCBCCBB:=!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		<a href="#">(ftp)</a>	SRA Experiment
GSM753974_UW.Breast_vHMEC.ChromatinAccessibility.RM035.DS18438.bed.gz	436.1 Mb	<a href="#">(ftp)</a> <a href="#">(http)</a>	BED
GSM753974_UW.Breast_vHMEC.ChromatinAccessibility.RM035.DS18438.wig.gz	168.2 Mb	<a href="#">(ftp)</a> <a href="#">(http)</a>	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		<a href="#">(ftp)</a>	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/sra/sra.cgi?cmd=show&f=software&m=software&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!

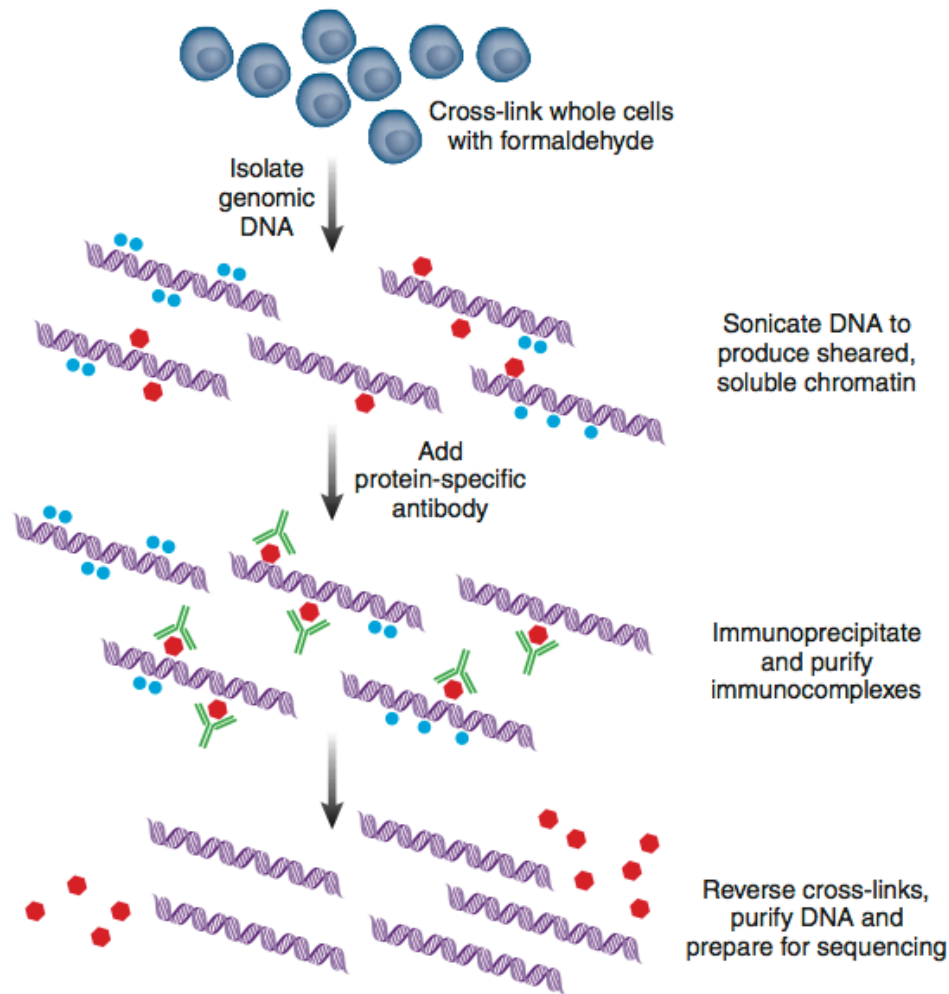
Quote

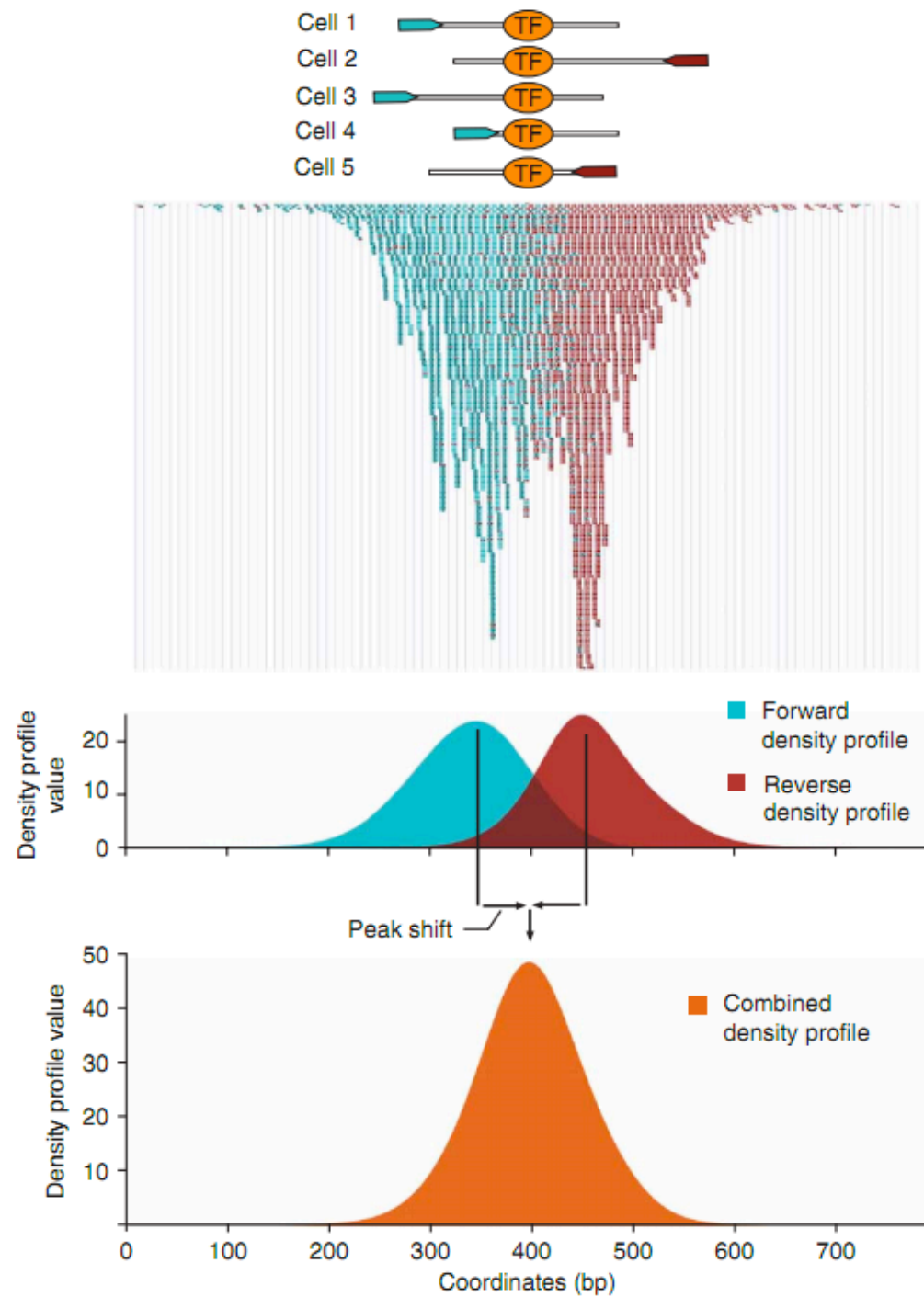
<http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software>

- `tar zxvf sratoolkit.2.1.2-mac64.tgz`
- `fastq-dump FoxAI.sra`



# Chromatin Immunoprecipitation followed by sequencing





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
AAATGTTAATCTGAANAGCTGGAATCCAGTCTGGTGTGTTGTA
+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
TACAAACACCAGACTGGATTCCAGCTNTTCAGATTACA
CAAA@>B8A>?C<ACCB@B7BAC=1!=:BBCCBBCBBC
0 12:C>N
```

# Genomic landscape of binding

- Reformat BOWTIE output as BED file

```
awk -F'\t' '{OFS=" "; 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>

**UCSC Genome Bioinformatics**

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

**Genome Browser**

ENCODE

Neandertal

Blat

Table Browser

Gene Sorter

In Silico PCR

Genome Graphs

Galaxy

VisiGene

Proteome Browser

Utilities

Downloads

Release Log

Custom Tracks

Microbial Genomes

### 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.

#### 8 September 2011 - New Navigation and Display Features

We've added several new features to the Genome Browser that make it easier to quickly configure and navigate around in the browser's annotation tracks window.

**Automatic image resizing:** The first time the annotation track window is displayed, or after the Genome Browser has been reset, the size of the track window is now set by default to the width that best fits your Internet browser window. If you subsequently resize your browser window, you can automatically adjust the annotation track image size to the new width by clicking the *resize* button under the track image. The default width can still be manually overridden on the Track Configuration page.

**Scrolling left or right in the track window:** You can now scroll (pan) horizontally through the tracks image by clicking on the image, dragging the cursor to the left or right, then releasing the mouse button. The view may be scrolled by up to one image width.

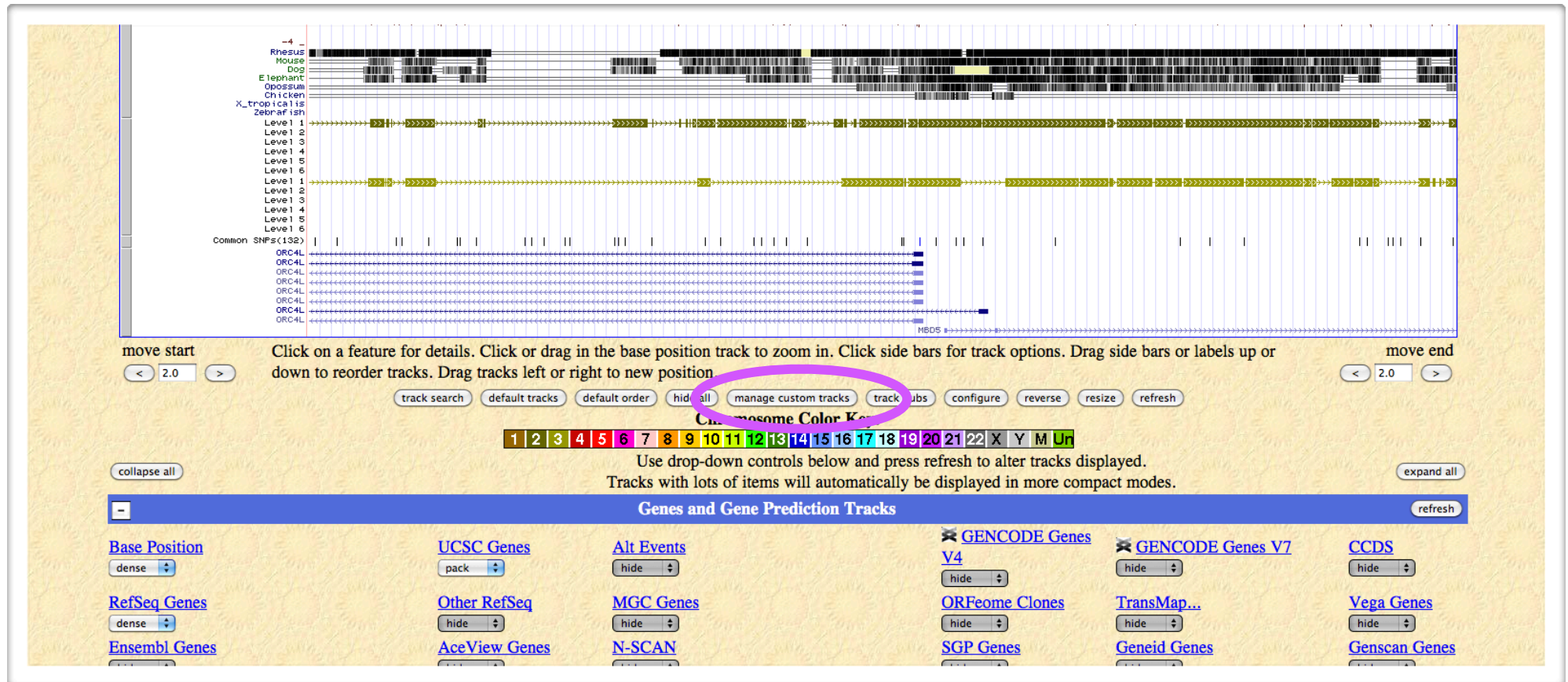
**Improved drag-zoom navigation:** The browser's "drag-and-zoom" feature lets you quickly zoom to a specific region of interest on the annotation tracks image. To define the region you wish to zoom to, depress the shift key, click-and-hold the mouse button on one edge of the desired zoom area (which can be anywhere in the tracks window), drag the mouse right or left to highlight the selection area, then release the mouse button. The annotation tracks image will automatically zoom to the new region. The Genome Browser still supports the earlier implementation of this feature, which restricted the click-drag to the Base Position track area of the image, but did not require the shift key to be pressed.

**Reordering groups of tracks:** You can now vertically reposition an entire group of associated tracks in the tracks image (such as all the displayed subtracks in a composite track) by clicking and holding the gray bar to the left of the tracks, dragging the group to the new position, then releasing the mouse button. To move a single track up or down, click and hold the mouse button on the side label, drag the highlighted track to the new position, then release the mouse button.

If you haven't yet tried the browser's right-click menu for quick access to frequently used track configuration features and functionality, read more [here](#).



# 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 |  
118388 |

# 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]

Name	Description	Type	Doc	Items	Pos	delete	update
<a href="#">User Track</a>	User Supplied Track	bed		5059	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">FoxA1 peaks</a>	FoxA1 peaks	bed		71450	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">ER peaks</a>	ER peaks	bed		21324	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">Breast_DNaseI.2 peaks</a>	Breast_DNaseI.2 peaks	bed		84121	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">Breast_DNaseI.1 peaks</a>	Breast_DNaseI.1 peaks	bed		85718	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">Breast_DNaseI.2</a>	Breast_DNaseI.2	bedGraph		4090778	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">Breast_DNaseI.1</a>	Breast_DNaseI.1	bedGraph		3620751	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">FoxA1</a>	FoxA1	bedGraph		5784344	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">ER</a>	ER	bedGraph		9449539	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">FoxA1.MACS</a>	FoxA1.MACS	bed		45758	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">ER.MACS</a>	ER.MACS	bed		5059	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">Breast_DNaseI.2.MACS</a>	Breast_DNaseI.2.MACS	bed		52329	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">Breast_DNaseI.1.MACS</a>	Breast_DNaseI.1.MACS	bed		51276	<a href="#">chr1:</a>	<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">FoxA1.bed_r1250_s25_coverage.wig</a>	FoxA1.bed_r1250_s25_coverage.wig	wiggle_0				<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">Breast_DNaseI.2.bed_r1250_s25_coverage.wig</a>	Breast_DNaseI.2.bed_r1250_s25_coverage.wig	wiggle_0				<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">Breast_DNaseI.1.bed_r1250_s25_coverage.wig</a>	Breast_DNaseI.1.bed_r1250_s25_coverage.wig	wiggle_0				<input type="checkbox"/>	<input type="checkbox"/>
<a href="#">ER.bed_r1250_s25_coverage.wig</a>	ER.bed_r1250_s25_coverage.wig	wiggle_0				<input type="checkbox"/>	<input type="checkbox"/>

check all / clear all + - + -

[add custom tracks](#)  
[go to genome browser](#)  
[go to table browser](#)

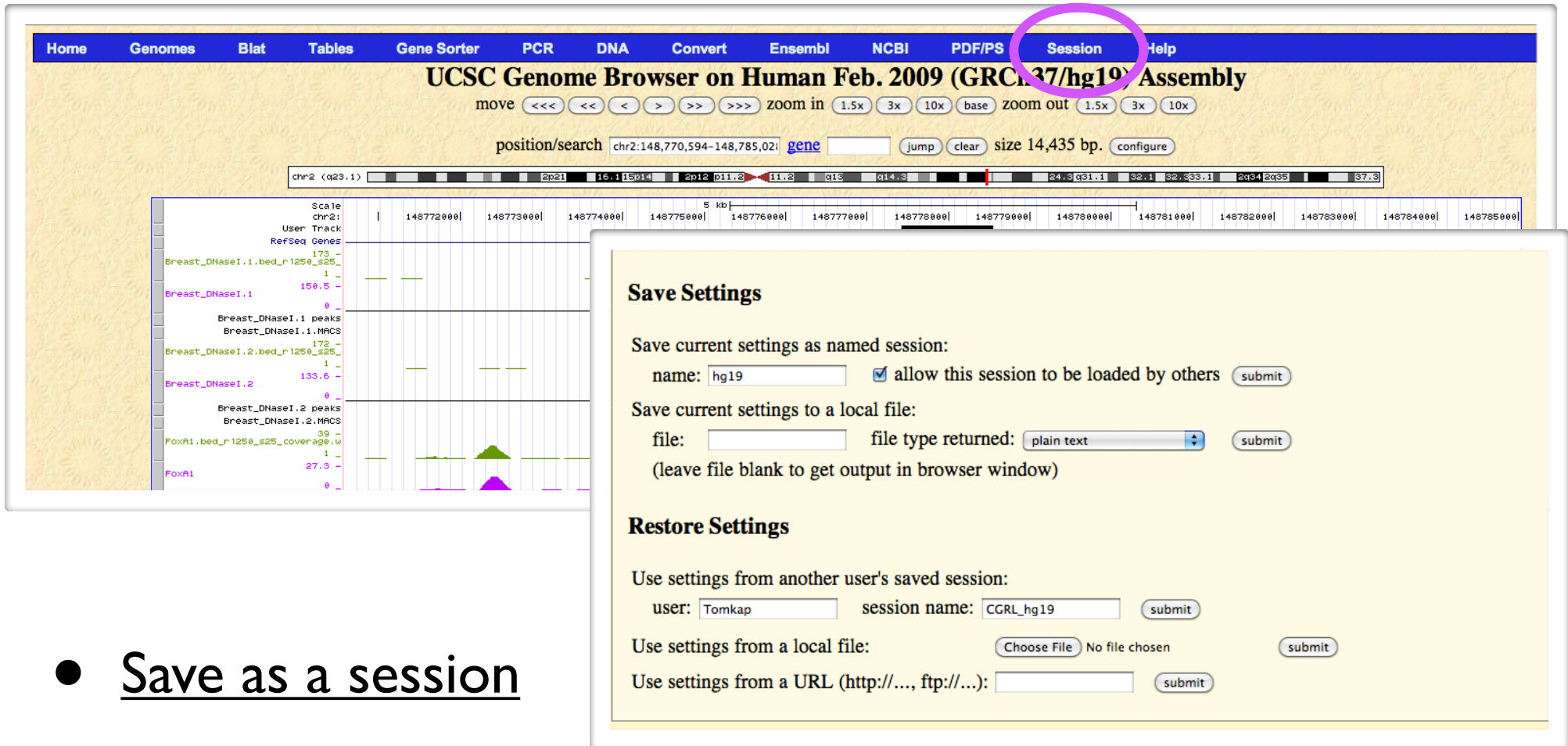
- Add as many tracks as needed

# Track display options

The screenshot shows the UCSC Genome Browser interface. At the top, it says "UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Ass". Below this are navigation controls: "move" with buttons for navigation, "zoom in" with buttons for zooming, and "zoom out" with buttons for zooming. A "position/search" field shows "chr2:148,770,594-148,785,021" and a "gene" field. A "size 14,435 bp." label and a "configure" button are also present. On the left, a track list shows "chr2 (q23.1)" and "Breast\_DNaseI.1.bed\_rl250\_s25\_coverage.wig". A purple circle highlights the "configure" button next to the track name. The main panel shows the "Breast\_DNaseI.1.bed\_rl250\_s25\_coverage.wig Track Settings" dialog. It includes a "Display mode" dropdown set to "full", a "Submit" button, and buttons for "Remove custom track" and "Update custom track". The "Type of graph" is set to "bar". The "Track height" is set to "29" pixels (range: 11 to 50). The "Vertical viewing range" has "min" set to "0" and "max" set to "20" (range: 1 to 79414). The "Data view scaling" is set to "auto-scale to data view" and "Always include zero" is set to "OFF". The "Transform function" is set to "NONE". The "Windowing function" is set to "mean" and the "Smoothing window" is set to "OFF" pixels. The "Draw y indicator lines" section has "at y = 0.0" set to "OFF" and "at y = 1" set to "OFF". A link for "Graph configuration help" and the text "Data last updated: 2011-09-27" are at the bottom of the dialog.

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

# Visualize with UCSC browser



The screenshot displays the UCSC Genome Browser interface. The top navigation bar includes links for Home, Genomes, Blat, Tables, Gene Sorter, PCR, DNA, Convert, Ensembl, NCBI, PDF/PS, Session, and Help. The 'Session' link is circled in purple. Below the navigation bar, the title reads 'UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly'. The main content area shows a genomic track for chr2:148,770,594-148,785,021. The track includes a scale bar, a user track, and several data tracks: Breast\_DNaseI.1, Breast\_DNaseI.1 peaks, Breast\_DNaseI.1 MACS, Breast\_DNaseI.2, Breast\_DNaseI.2 peaks, Breast\_DNaseI.2 MACS, FoxP1, and FoxP1 coverage. A 'Save Settings' dialog box is overlaid on the right, with the following content:

**Save Settings**

Save current settings as named session:

name:  ☒ allow this session to be loaded by others

Save current settings to a local file:

file:  file type returned:

(leave file blank to get output in browser window)

**Restore Settings**

Use settings from another user's saved session:

user:  session name:

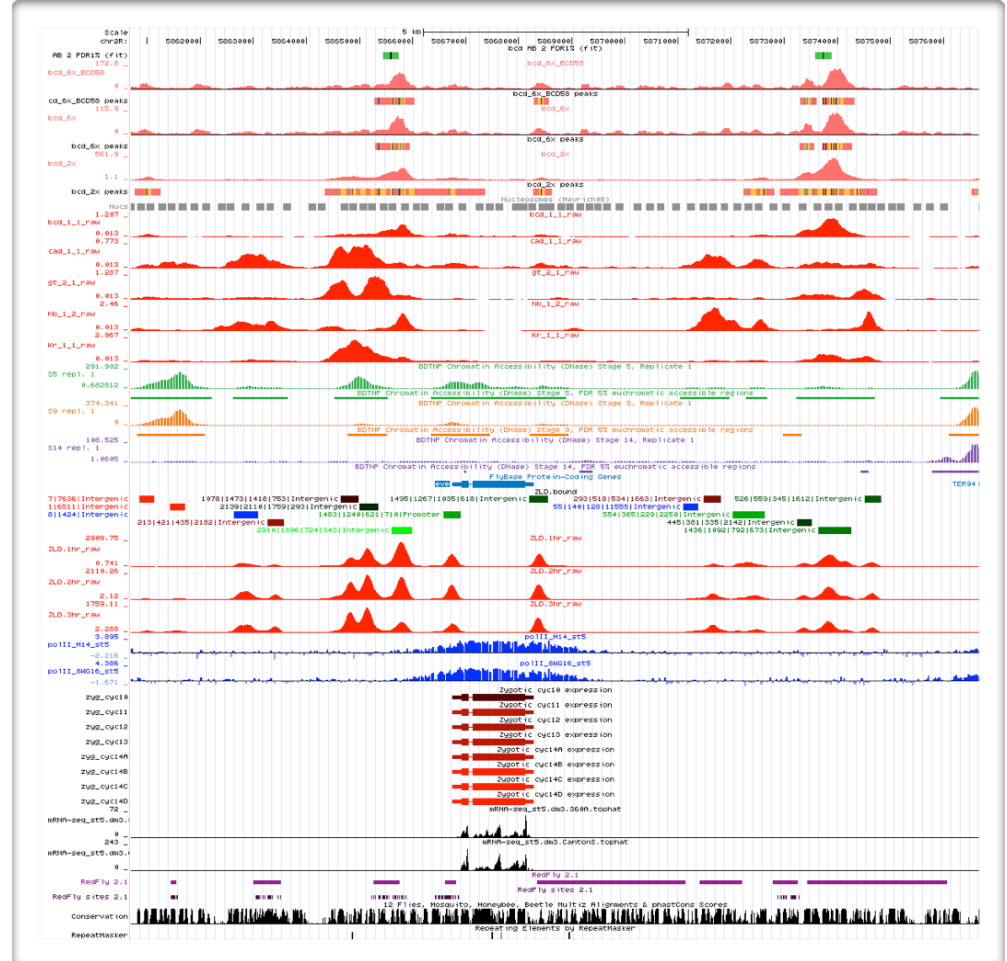
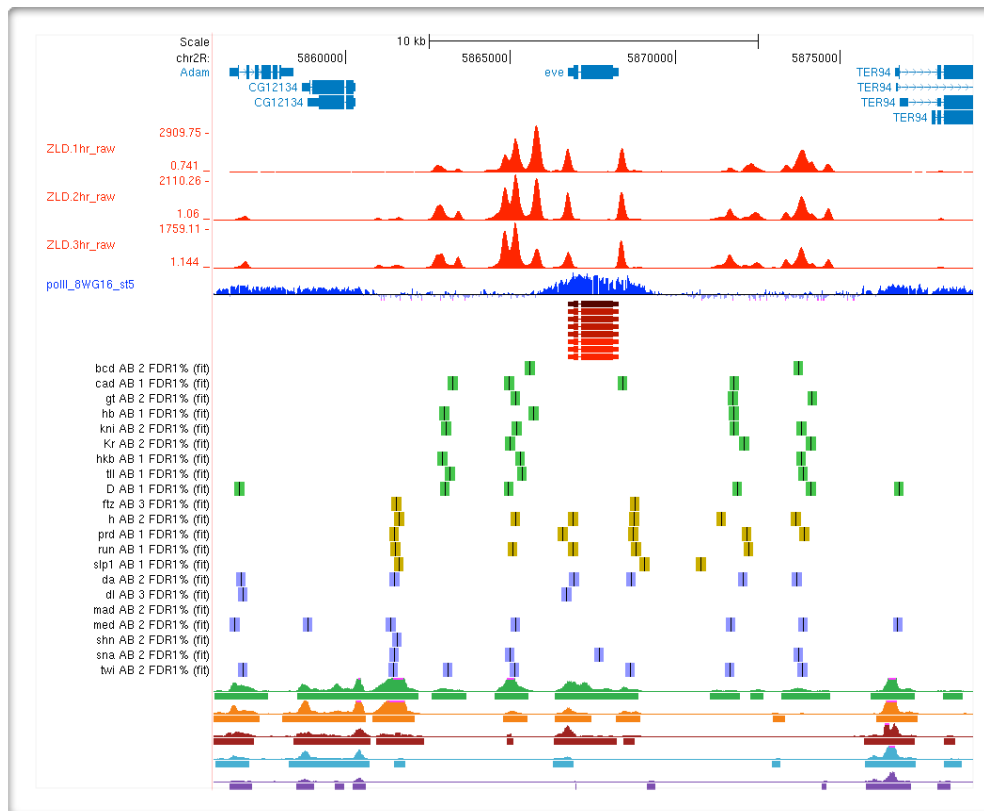
Use settings from a local file:  No file chosen

Use settings from a URL (http://..., ftp://...):

- Save as a session
- Share link with colleagues and friends
- Last 4 months since last access.  
Access periodically, e.g. via [www.followthatpage.com](http://www.followthatpage.com)

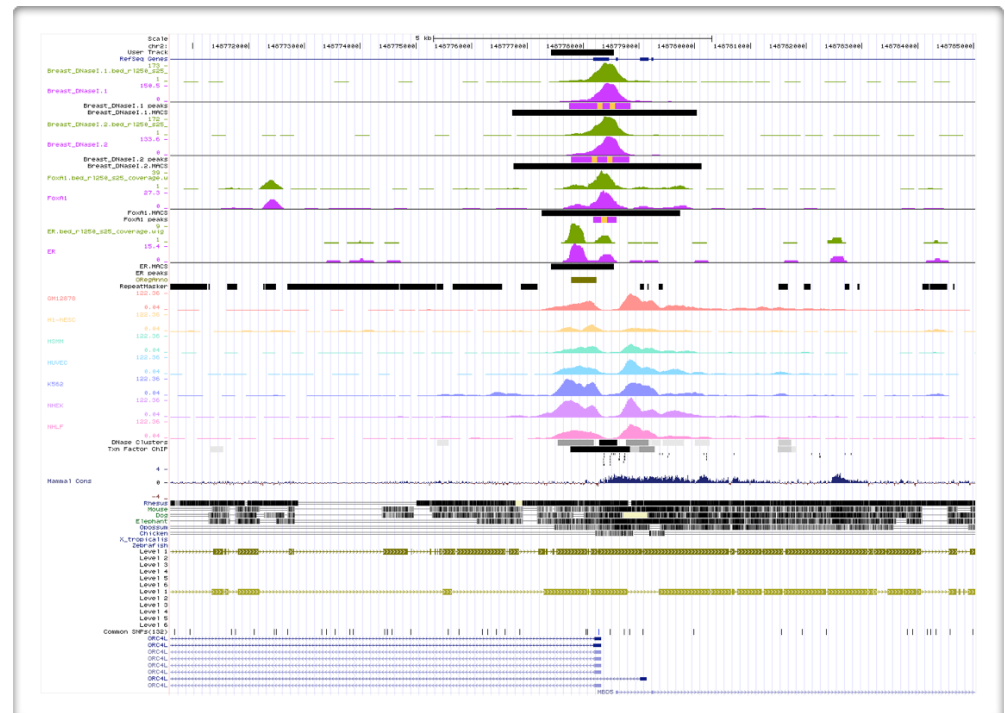
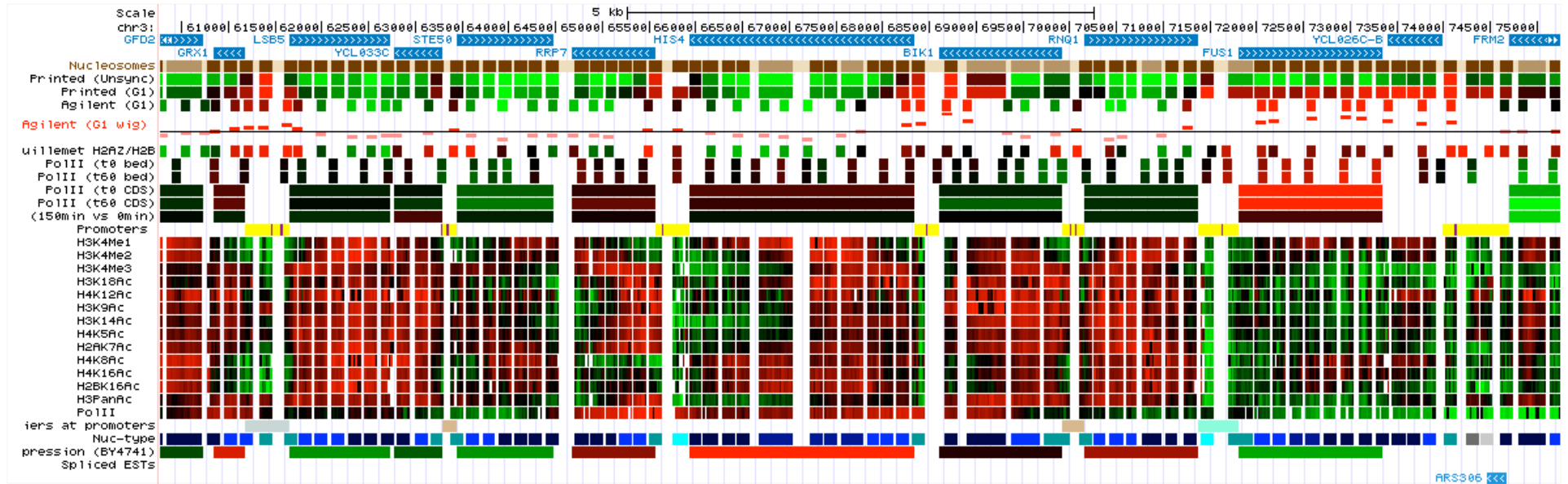
# Some saved sessions

## Binding and expression in Drosophila





# Histone modifications and turnover rates in Yeast



# CGRU example

# Analyzing binding data

- Different data sets yield different questions
- Transcription factor ChIP
- Histone modifications
- DNase hypersensitivity



# MACS

Open Access

Method

## Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang<sup>✉\*</sup>, Tao Liu<sup>✉\*</sup>, Clifford A Meyer<sup>\*</sup>, Jérôme Eeckhoute<sup>†</sup>,  
David S Johnson<sup>‡</sup>, Bradley E Bernstein<sup>§¶</sup>, Chad Nusbaum<sup>¶</sup>,  
Richard M Myers<sup>¥</sup>, Myles Brown<sup>†</sup>, Wei Li<sup>#</sup> and X Shirley Liu<sup>\*</sup>

Published: 17 September 2008

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

## Abstract

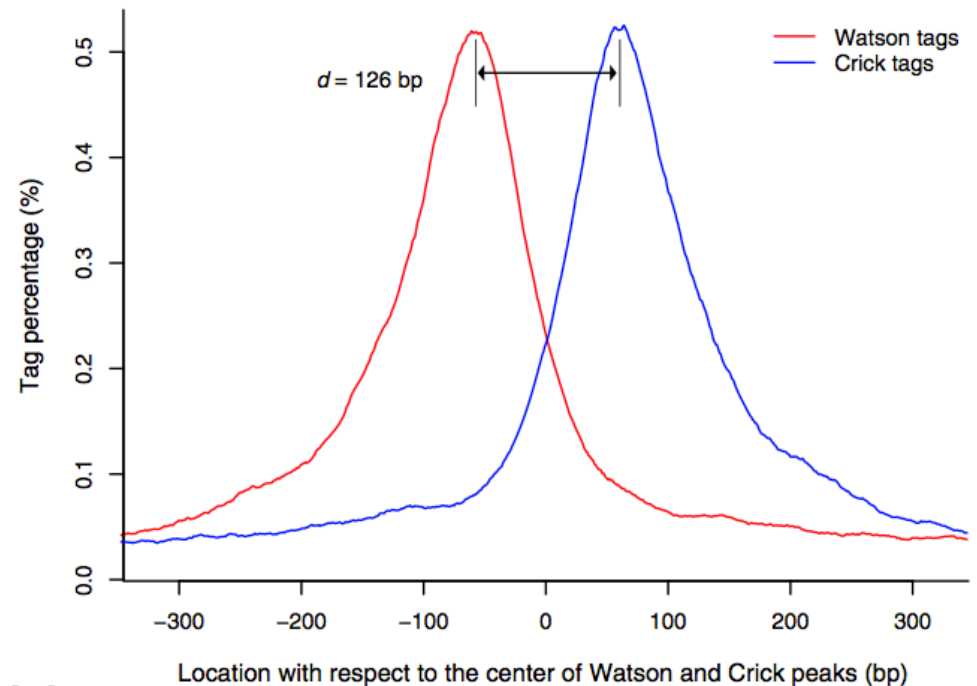
---

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.

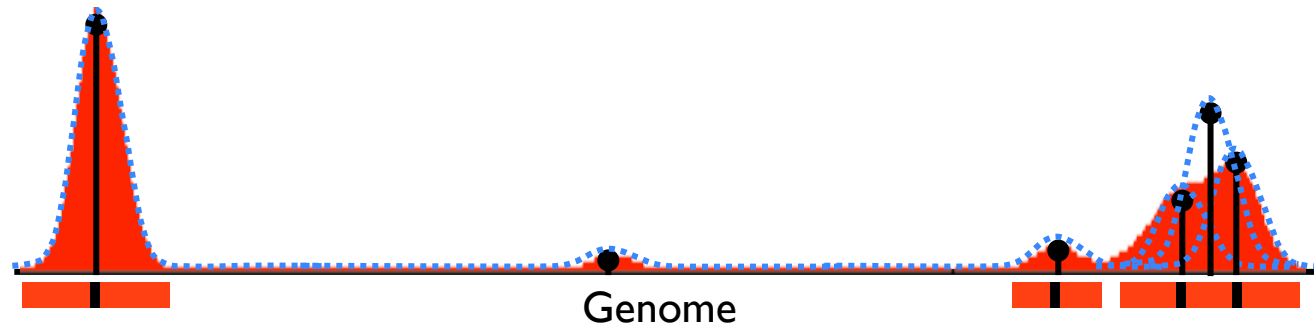
---

# MACS

- Analyzes forward and reverse reads
- Identifies avg. length of DNA fragment
- Allows usage of mock IP for localized normalization
- `macs14 -n FoxAI.MACS -t FoxAI.bed -g hs --off-auto --nomodel --shiftsize=125`

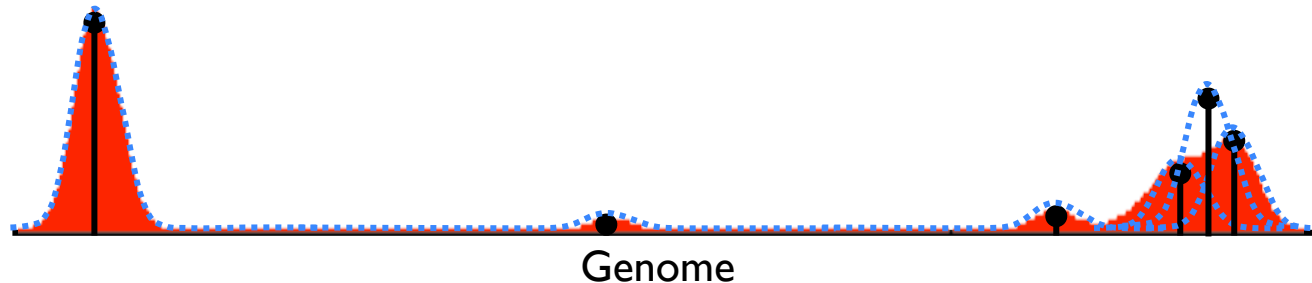


# Grizzly Peak Fitting



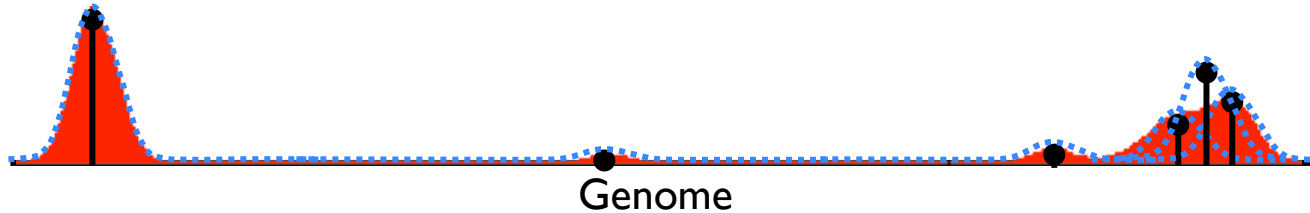
- Model-based, iterative, multi-peak peak finding algorithm
- MATLAB code, standalone executable\*
- <http://rana.lbl.gov/software/grizzly>
- Capaldi, Kaplan et al, 2008, Harrison, Li, Kaplan et al, 2011

# Grizzly Peak Fitting

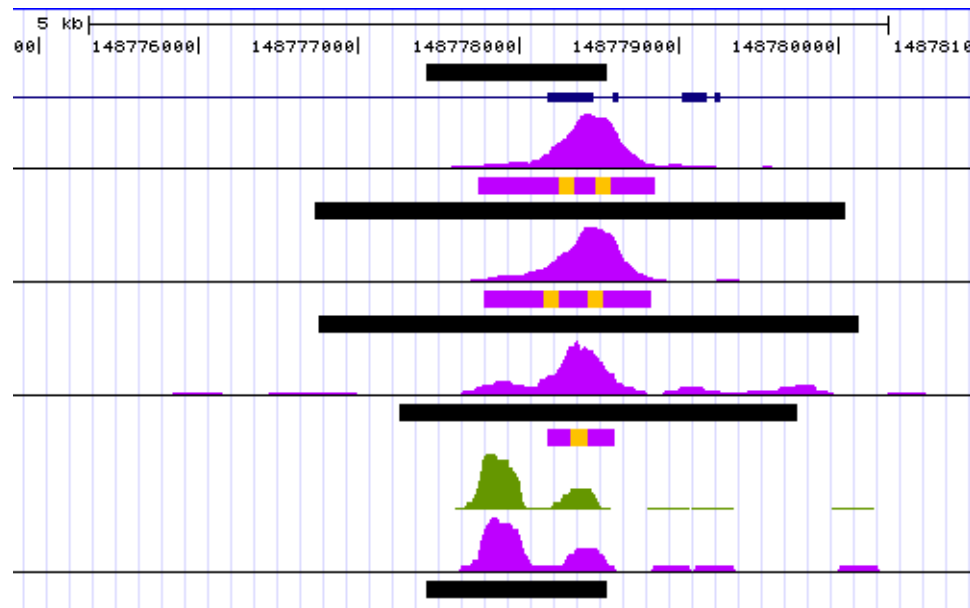


- Great in fitting the shape of a peak
- Given a motif, disentangle binding landscape into estimate occupancy at putative binding sites
- Bayesian confidence intervals for binding strengths
- Quite bad at stopping (separating real peaks from noise)
- pf0 FoxA1.bed 25 250 genome\_hgl9.mat 10

# 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

The screenshot shows the Galaxy web interface. At the top is a navigation bar with links: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. On the left is a sidebar with a 'Tools' section and an 'Options' dropdown. The 'Tools' section lists various categories: Get Data, Send Data, ENCODE Tools, Lift-Over, Text Manipulation, Convert Formats, FASTA manipulation, Filter and Sort, Join, Subtract and Group, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, Human Genome Variation, Genome Diversity, and EMBOSS. The main content area features a large box titled 'Galaxy 101' with the subtitle 'Start small' and the text 'The very first tutorial you need'. Below this is a section titled 'Live Quickies' which displays a horizontal scrollable list of seven quickie tutorials: Mapping against custom genome (Galactic quickie # 10), Illumina mapping: Single Ends (Galactic quickie # 11), Illumina mapping: Paired Ends (Galactic quickie # 12), Basic fastQ manipulation: (Galactic quickie # 13), Advanced fastQ manipulation: (Galactic quickie # 14), 454 Mapping: Single End (Galactic quickie # 15), and Uploading Data using FTP (Galactic quickie # 17). At the bottom of the interface, there is a paragraph of text about the Galaxy project and its support, followed by a 'Galaxy build' string: '\$Rev 6056:338ead4737ba\$'.

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

# Load called peaks

The screenshot shows the Galaxy web interface. On the left, the 'Tools' panel has 'Get Data' circled in purple. The main panel shows the 'Upload File (version 1.1.3)' workflow. The 'File Format' is set to 'bed'. The 'File' is 'FoxA1.MACS\_peaks.bed.bz2'. The 'Genome' is 'hg19'. The 'Execute' button is visible. On the right, the 'History' panel shows the upload history. A large blue arrow points from the 'History' panel to the 'Data View' panel, which displays the loaded data as a table.

**Galaxy** Analyze Data Workflow Shared Data Visualize

**Tools** Options

**Get Data**

- Upload File from your computer
- UCSC main table browser
- UCSC Archaea table browser
- BX main browser
- 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

**Send Data**

**ENCODE Tools**

**Lift-Over**

**Text Manipulation**

**Convert Formats**

**FASTA manipulation**

**Filter and Sort**

**Join, Subtract and Group**

**Upload File (version 1.1.3)**

**File Format:**

bed

Which format? See help below

**File:**

Choose File FoxA1.MACS\_peaks.bed.bz2

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload files larger than 2GB, use the FTP interface (see the site administrator).

**URL/Text:**

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

**Files uploaded via FTP:**

File	Size
Your FTP upload directory contains no files.	

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP site (see the site administrator for address and password).

**Convert spaces to tabs:**

☐ Yes

Use this option if you are entering intervals by hand.

**Genome:**

hg19

Execute

**History** Options

Unnamed history 0 bytes

1: FoxA1.MACS\_peaks.bed.bz2

**History** Options

Unnamed history 0 bytes

1: FoxA1.MACS\_peaks.bed.bz2

**1: FoxA1 MACS peaks (unsorted)**


45,758 regions, 1 comments  
format: bed, database: hg19  
Info: uploaded bed file

display at UCSC main  
view in GeneTrack  
display at Ensembl Current

1. Chrom	2. Start	3. End	4. Name
track type=bed name=FoxA1_MACS desc=			
chr1	867927	869317	MACS
chr1	930814	932135	MACS
chr1	958730	961364	MACS
chr1	1008252	1011246	MACS
chr1	1057348	1061450	MACS



# Sort by height

 **Galaxy**

Analyze DataWorkflowShare

ToolsOptions

[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

### Sort (version 1.0.1)

Sort Query:  
1: MACS peaks (unsorted)

on column:  
c5

with flavor:  
Numerical sort

everything in:  
Descending order

Column selections  
Add new Column selection

Execute

**2: FoxA1 MACS peaks (sorted)**

45,759 regions  
format: bed, database: hg19  
Info: None

display at UCSC [main](#)  
view in [GeneTrack](#)  
display at Ensembl [Current](#)

3. End	4. Name	5
121485895	MACS_peak_1714	3100.00
193431475	MACS_peak_28997	2287.62
73457495	MACS_peak_17610	1863.51
110268805	MACS_peak_43782	1841.39
98357376	MACS_peak_5375	1725.47
156443139	MACS_peak_1954	1715.69

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

Syntax

# Select top 23 regions

**Galaxy**

Analyze DataWorkflowShared

ToolsOptions

Get Data

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Add column to an existing dataset

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

Select first (version 1.0.0)

Select first:

23

lines

from:

2: MACS peaks (sorted)

Execute

What it does

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

Example

Selecting 2 lines from this:

chr75663256652D17003 CTCF\_R6310+

chr75673656756D17003 CTCF\_R7354+

chr75676156781D17003 CTCF\_R4220+

chr75677256792D17003 CTCF\_R7372+

chr75677556795D17003 CTCF\_R4207+

will produce:

chr75663256652D17003 CTCF\_R6310+

chr75673656756D17003 CTCF\_R7354+

4: Top 23 peaks

23 regions

format: bed, database: hg19

Info: None

display at UCSC [main](#)

view in [GeneTrack](#)

display at Ensembl [Current](#)

1. Chrom	2. Start	3. End	4. Name
chr1	121482891	121485895	MACS_pe
chr3	193428681	193431475	MACS_pe
chr17	73454188	73457495	MACS_pe
chr9	110266406	110268805	MACS_pe
chr10	98355118	98357376	MACS_pe
chr1	156440232	156443139	MACS_pe

# get genomic sequences

The screenshot displays the Galaxy web interface. The top navigation bar includes the Galaxy logo and an 'Analyze' button. On the left, a 'Tools' sidebar lists various categories: Get Data, Send Data, ENCODE Tools, Lift-Over, Text Manipulation, Convert Formats, FASTA manipulation, Filter and Sort, Join, Subtract and Group, Extract Features, and Fetch Sequences. The 'Fetch Sequences' category is expanded, and 'Extract Genomic DNA using coordinates from assembled/unassembled genomes' is highlighted with a purple oval.

The main workspace shows the 'Extract Genomic DNA (version 2.2.2)' tool configuration. The settings are as follows:

- Fetch sequences for intervals in:** 4: Top 23 peaks
- Interpret features when possible:** Yes
- Source for Genomic Data:** Locally cached
- Output data type:** FASTA
- Execute** button

On the right, the 'History' panel shows a list of jobs. The job '5: Seqs of top 23' is selected, showing 23 sequences in FASTA format from the hg19 database. A 'Download' button is visible next to the job name, and a preview of the FASTA sequence is shown below.

```
> 121482891_121485895_+
aatttcagagaagtccaaatocacttgcattat
tttgaacatgctccatcagaagatatgctcagct
atcattgcaaagaattttctgagaatgottctgt
gttatttcoctttactacgataggcootcaaagag
cagattctgcagaaggagtgtttcaaacotgaact
```

# Look for motif with MEME

<http://meme.sdsc.edu/meme/>

## MEME Suite Menu

- Submit A Job
- Documentation
- Downloads
- User Support
- Alternate Servers
- Authors
- Citing



Version 4.7.0

Use this form to submit DNA or protein sequences to MEME. MEME will analyze your sequences for similarities among them and produce a description (**motif**) for each pattern it discovers.

### Data Submission Form

#### Required

Your e-mail address:

tomkap@gmail.com

Re-enter e-mail address:

tomkap@gmail.com

Please enter the **sequences** which you believe share one or more motifs. The sequences may contain no more than **60000 characters** total in any of a large number of **formats**.

Enter the **name of a file** containing the sequences here:

Choose File Galaxy5.fasta

Clear

or  
the **actual sequences** here (**Sample Protein Input Sequences**):

How do you think the occurrences of a single motif are **distributed** among the sequences?

☐ One per sequence

☒ Zero or one per sequence

☐ Any number of repetitions

MEME will find the optimum **width** of each motif within the limits you specify here:

6 Minimum width ( $\geq 2$ )

50 Maximum width ( $\leq 300$ )

3 Maximum **number of motifs** to find

#### Options

**Description** of your sequences:

MEME will find the optimum **number of sites** for each motif within the limits you specify here:

Minimum sites ( $\geq 2$ )

Maximum sites ( $\leq 300$ )

☐ **Shuffle** sequence letters

**NEW** Perform **discriminative** motif discovery – Enter the name of a file containing '**negative sequences**':

Choose File No file chosen

Clear

Enter the name of a file containing a **background Markov model**:

Choose File No file chosen

Clear

**DNA-ONLY OPTIONS**  
(Ignored for protein searches)

☐ Search given **strand** only

\* 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](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	<b>dna</b>
number of sequences	<b>23</b>
shortest sequence (residues)	<b>1493</b>
longest sequence (residues)	<b>3307</b>
average sequence length (residues)	<b>2540.8</b>
total dataset size (residues)	<b>58438</b>

You can view your job results at: [http://meme.nbcr.net/meme4\\_7\\_0/cgi-bin/querystatus.cgi?jobid=app1318268648139&service=MEME](http://meme.nbcr.net/meme4_7_0/cgi-bin/querystatus.cgi?jobid=app1318268648139&service=MEME)

**ACTI** **ACT** **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 text](#)
- [MEME output as XML](#)
- [MAST output as HTML](#)
- [MAST output as XML](#)
- [MAST output as text \(for input sequences\)](#)

## 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 following:  
Timothy L. Bailey and Charles Elkan, "Fitting a mixture model b

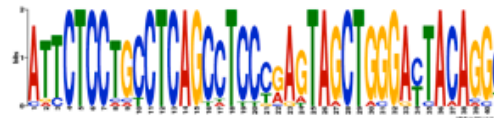
[DISCOVERED MOTIFS](#) | [BLOCK DIAGRAMS OF MOTIFS](#) | [PROGRAM INFORMATION](#) | [EXPLANATION](#)

## DISCOVERED MOTIFS

### Motif Overview

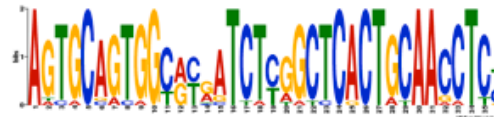
#### [Motif 1](#)

- 6.1e-206
- 17 sites



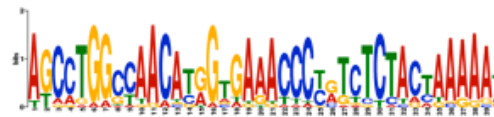
#### [Motif 2](#)

- 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.

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 ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGACTACAGGC GCCCACCACC
hg19_chr1_156440232_156443139_+	+	2373	1.67e-25	AGTTCAAGTG ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGATTACAGGC ACATGCCACT
hg19_chr17_73454188_73457495_+	-	822	2.58e-25	GGTTCACGCC ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGACTACAGGC ACCCGCCACA
hg19_chr9_110266406_110268805_+	+	1360	4.40e-25	GGTTCAAGCG ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGACTACAGGT GCATGTGCC
hg19_chr10_98622889_98625123_+	+	175	6.59e-25	GTTTCAAGTG ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGATTACAGGC ATGCACCACC
hg19_chr17_61828830_61831978_+	-	2508	1.26e-23	GTTTCAAGCA ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGATTACAGGC GTGCACCACC
hg19_chr9_128057870_128061033_+	+	2258	1.76e-23	GGTTCAACA ATTCTCCTGCCTCAGCCTCCGAGTAGCTAGGACTACAGGT GTACACCACC
hg19_chr19_35495955_35498474_+	-	1143	5.42e-23	GGTTCAAGTG ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGATTACAGGC ACCCACCACC
hg19_chr20_55309000_55311968_+	+	589	6.77e-23	GGTTCAAGTG ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGATTACAGGC GCCTGCCACC
hg19_chr11_40349932_40351562_+	+	1463	7.97e-23	GGTTCAAGCG ATTCTCCTGCCTCAGCCTCCGAGTAGCTGCGACTACAGGC ATTTGCCAAC
hg19_chr5_40446525_40448658_+	-	23	1.20e-22	GGTTCAAGTG ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGACTACAGGC GCCCGCCCA
hg19_chr19_54124486_54126761_+	-	1947	1.41e-22	GGTTCAAGCG ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGACTACAGGC GCCTGCCACC
hg19_chr3_193428681_193431475_+	+	2112	2.52e-22	GGCTCAAGCT ATCTCCAGCCTCAGCCTCCGAGTAGCTGGGACTACAGGT GCATAAACC
hg19_chr8_95540335_95542493_+	+	1636	3.97e-22	GGCTCAAGCA ATTCTCCTGCCTCAGCCTCCGAGTAGCTGGGATTACAGGC GTGTAAACC
hg19_chr17_53509073_53512135_+	+	388	8.63e-22	GGCTCAAGTG CTCCTCCTGCCTCAGCCTCCGAGTAGCTGGGACTACAGGC CTGGGCCACC
hg19_chr19_18140018_18141511_+	-	221	2.20e-19	GGTTCAAGCT AATCTCCACCTCAGCCTCCGAGTAGCTGGGACCACCC AGATAATTAC
hg19_chr13_37977646_37979671_+	+	1266	1.15e-18	GGTTCAAGCA AGTCTCCTGCTCAGCCTCCCAATTAACTGGGATTAAAGGT ACCCATCACC

# 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

The screenshot displays the Galaxy web interface. The top navigation bar includes the 'Galaxy' logo and tabs for 'Analyze Data' and 'Workflow'. On the left, a 'Tools' sidebar lists various data sources under the 'Get Data' section. The 'Upload File from your computer' tool is highlighted with a pink circle. The main panel shows the 'Upload File (version 1.1.3)' tool configuration. The 'File Format' dropdown menu is set to 'bed' and is also circled in pink. Below this, the 'File' section shows a 'Choose File' button and the filename 'ER.MACS\_peaks.bed.bz2'. A tip notes that files larger than 2GB cannot be uploaded via the browser. The 'URL/Text' section has a large text area for pasting content. Below that, a table titled 'Files uploaded via FTP:' shows no files. The 'Convert spaces to tabs' option is unchecked.

**Galaxy** Analyze Data Workflow

Tools Options ▾

**Get Data**

- Upload File from your computer
- UCSC main table browser
- UCSC Archaea table browser
- BX main browser
- 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

**Send Data**

**ENCODE Tools**

**Upload File (version 1.1.3)**

**File Format:**

bed

Which format? See help below

**File:**

Choose File ER.MACS\_peaks.bed.bz2

TIP: Due to browser limitations, uploading files larger than 2GB is guarant site administrator).

**URL/Text:**

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

**Files uploaded via FTP:**

File	Size
Your FTP upload directory contains no files.	

This Galaxy server allows you to upload files via FTP. To upload some file: address and password).

**Convert spaces to tabs:**

☐ Yes

Use this option if you are entering intervals by hand



# Galaxy

- Calculate base coverage for FoxA1

The screenshot shows the Galaxy web interface. On the left is a sidebar with a 'Tools' section containing various tool categories like 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Convert Formats', 'FASTA manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', and 'Operate on Genomic Intervals'. The 'Base Coverage' tool is highlighted in the 'Operate on Genomic Intervals' section. The main panel shows the 'Base Coverage (version 1.0.0)' tool interface. It has a dropdown menu set to '2: FoxA1 MACS peaks (sorted)' and an 'Execute' button. Below the tool interface, there is a tip: 'TIP: If your dataset does not appear in the pulldown menu, it may be in a different column.' and a description: 'This operation counts the total bases covered by a set of intervals.' There is also a 'Screencasts!' section with a link to 'See Galaxy Interval Operation Screencasts (right click to open in new window)'.

**Galaxy** Analyze Data

Tools Options

Get Data  
Send Data  
ENCODE Tools  
Lift-Over  
Text Manipulation  
Convert Formats  
FASTA manipulation  
Filter and Sort  
Join, Subtract and Group  
Extract Features  
Fetch Sequences  
Fetch Alignments  
Get Genomic Scores  
Operate on Genomic Intervals

- Intersect the intervals of two datasets
- Subtract the intervals of two datasets
- Merge the overlapping intervals of a dataset
- Concatenate two datasets into one dataset
- Base Coverage of all intervals**
- Coverage of a set of intervals on second set of intervals

**Base Coverage (version 1.0.0)**

Compute coverage for:  
2: FoxA1 MACS peaks (sorted)

Execute

**TIP:** If your dataset does not appear in the pulldown menu, it may be in a different column.

This operation counts the total bases covered by a set of intervals.

**Screencasts!**

See Galaxy Interval Operation [Screencasts](#) (right click to open in new window)

The screenshot shows a Galaxy job output window titled '6: Base Coverage on data 2'. It displays the following information: '1 line', 'format: txt, database: hg19', and a single line of output: '108992582'. The window has standard icons for viewing, editing, and deleting the output, as well as a download icon.

**6: Base Coverage on data 2**

1 line  
format: txt, database: hg19

108992582

# Galaxy

- Intersect two databases

The screenshot shows the Galaxy web interface. On the left is a sidebar with a 'Tools' menu. The 'Intersect' tool is highlighted under the 'Operate on Genomic Intervals' section. The main panel displays the 'Intersect (version 1.0.0)' tool configuration. The 'Return' dropdown is set to 'Overlapping pieces of Intervals'. The 'of' dropdown is set to '14: Intersect on data 8 and data 2'. The 'that intersect' dropdown is also set to '14: Intersect on data 8 and data 2'. The 'for at least' input is set to '1'. The 'Execute' button is visible. Below the configuration, there is a 'Screencasts!' section with a link to 'Galaxy Interval Operation Screencasts'. A 'Syntax' section explains the tool's parameters: 'Where overlap is at least' sets the minimum overlap length; 'Overlapping Intervals' returns entire intervals from the first dataset; 'Overlapping pieces of Intervals' returns the exact base pair overlap. An 'Example' section shows a diagram with a red bar (First query), a green bar (Intervals to intersect with (Second Query)), and the resulting overlapping intervals and pieces.

**Galaxy** Analyze Data Workflow Shared Data

Tools Options

Get Data  
Send Data  
ENCODE Tools  
Lift-Over  
Text Manipulation  
Convert Formats  
FASTA manipulation  
Filter and Sort  
Join, Subtract and Group  
Extract Features  
Fetch Sequences  
Fetch Alignments  
Get Genomic Scores  
Operate on Genomic Intervals  
**Intersect** the intervals of two datasets  
Subtract the intervals of two datasets  
Merge the overlapping intervals of a dataset  
Concatenate two datasets into one dataset  
Base Coverage of all intervals  
Coverage of a set of intervals on second set of intervals  
Complement intervals of a dataset  
Cluster the intervals of a dataset  
Join the intervals of two datasets side-by-side  
Get flanks returns flanking region/s for every gene  
Fetch closest non-overlapping feature for every interval  
Profile Annotations for a set of genomic intervals

**Intersect (version 1.0.0)**

Return:  
Overlapping pieces of Intervals  
(see figure below)

of:  
14: Intersect on data 8 and data 2  
First dataset

that intersect:  
14: Intersect on data 8 and data 2  
Second dataset

for at least:  
1  
(bp)

Execute

**TIP:** If your dataset does not appear in the pulldown menu, it means that it is not in Inter columns.

**Screencasts!**  
See Galaxy Interval Operation [Screencasts](#) (right click to open this link in another window).

**Syntax**  
Where overlap is at least sets the minimum length (in base pairs) of overlap between elements.  
Overlapping Intervals returns entire intervals from the first dataset that overlap the second only filters out intervals that do not overlap with the second dataset.  
Overlapping pieces of Intervals returns intervals that indicate the exact base pair overlap from the first dataset, and all fields besides start and end are guaranteed to remain unchanged.

**Example**

First query  
Intervals to intersect with (Second Query)  
Overlapping intervals  
Overlapping pieces of intervals

**14: Intersection of ER and FoxA1 (FoxA1 IDs)**

2,933 regions, 1 comments  
format: bed, database: hg19  
Info: Skipped 7 invalid lines of 1st dataset, 1st line #3011:

display at UCSC [main](#)  
view in [GeneTrack](#)  
display at Ensembl [Current](#)

1. Chrom	2. Start	3. End	4. Name
chr1	121482891	121485895	MACS_pe
chr3	193428783	193431035	MACS_pe
chr9	110266406	110268212	MACS_pe
chr19	35496340	35498474	MACS_pe
chr19	18140201	18141511	MACS_pe
chr10	7274362	7277079	MACS_pe

**13: Intersection of FoxA1 and ER (ER IDs)**

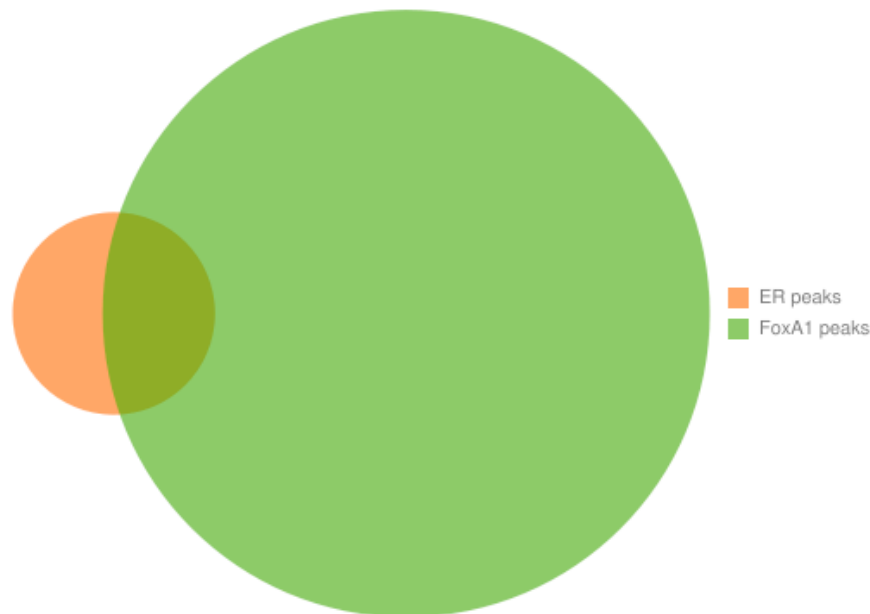
2,933 regions, 1 comments  
format: bed, database: hg19  
Info: Skipped 2 invalid lines of 1st dataset, 1st line #459:

display at UCSC [main](#)  
view in [GeneTrack](#)  
display at Ensembl [Current](#)

1. Chrom	2. Start	3. End	4. Name
chr9	97544440	97546328	MACS_pe
chr11	1817113	1818655	MACS_pe
chr1	121482891	121485895	MACS_pe
chr8	128870861	128873732	MACS_pe
chr13	99301882	99303702	MACS_pe
chr2	218251363	218253548	MACS_pe

# 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



History Options

Unnamed history 4.5 Mb

9: Coverage of ER MACS  
1 line  
format: txt, database: hg19  
Info: None  
11118179

8: ER MACS peaks (sorted)

7: ER MACS peaks (unsorted)

6: Coverage of FoxA1 MACS  
1 line  
format: txt, database: hg19  
108992582

5: Segs of top 23

4: Top 23 peaks

2: FoxA1 MACS peaks (sorted)

1: FoxA1 MACS peaks (unsorted)

# Galaxy

- Upload raw data, convert to intervals

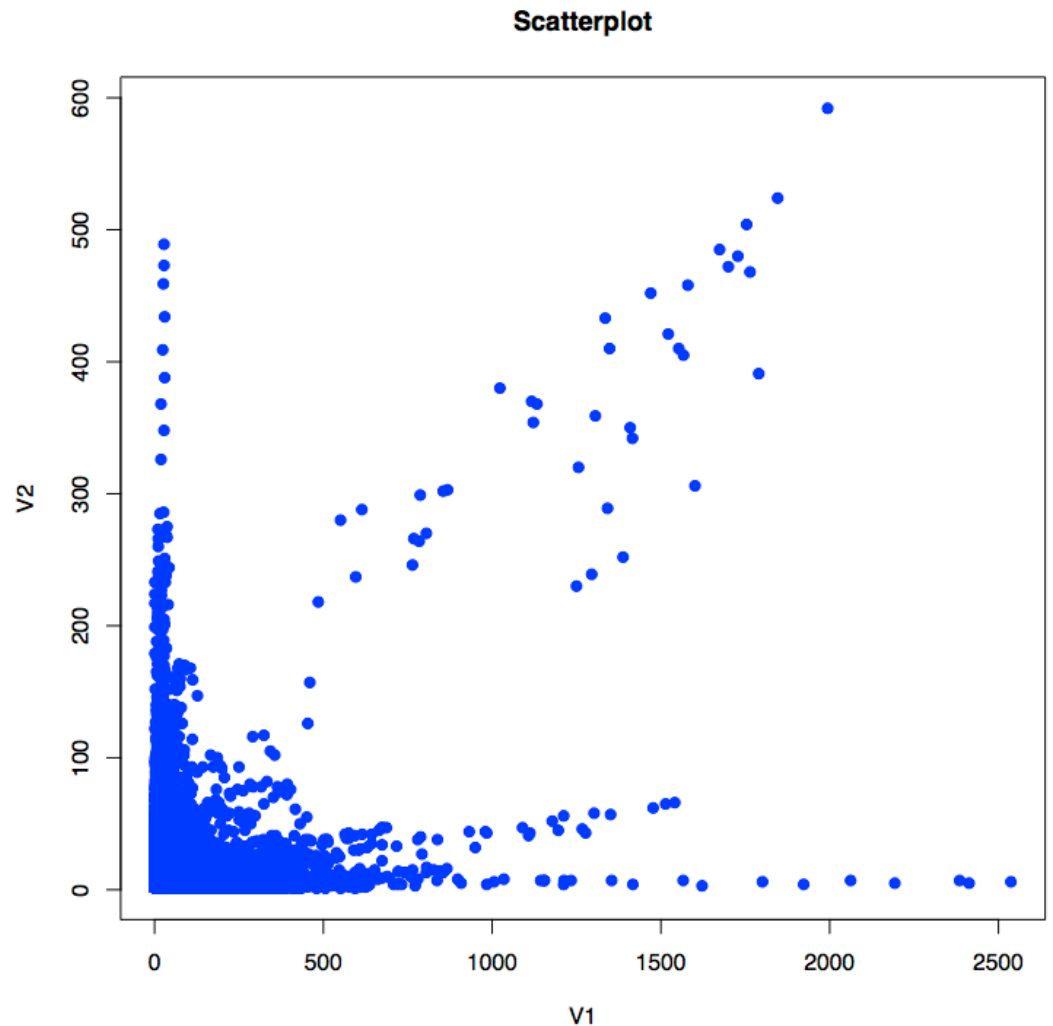
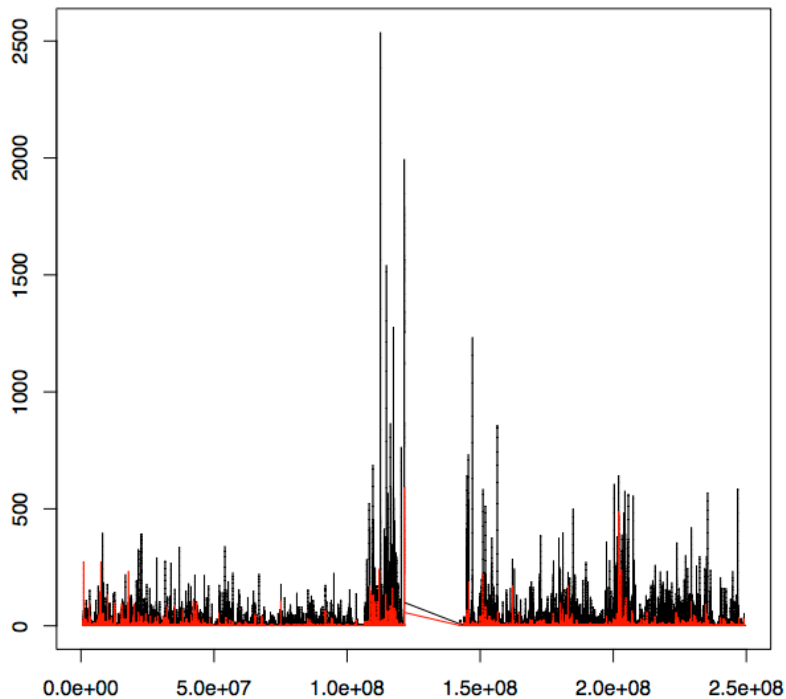
The screenshot displays the Galaxy web interface. On the left, the 'Tools' panel is visible, with 'Upload File from your computer' circled in pink. The main panel shows the 'Upload File (version 1.1.3)' tool. In this tool, 'File Format' is set to 'wig' (circled in pink), and the 'File' field contains 'ER.chr1.bed\_r...erage.wig.bz2'. Below this, the 'URL/Text' field is empty. The 'Files uploaded via FTP' section shows a message: 'Your FTP upload directory contains no files.' The 'Convert spaces to tabs' checkbox is unchecked. The 'Genome' dropdown is set to 'Human Feb. 2009 (GRCh37/hg19) (hg)'. The 'Execute' button is visible.

On the right, the 'Wiggle-to-Interval (version 1.0.0)' tool is shown. The 'Convert' dropdown is set to '17: ER raw' (circled in pink). The 'Execute' button is visible. Below the tool name, the 'Syntax' section explains that the tool converts wiggle data into interval type. It describes the 'Wiggle format' as line-oriented with a UCSC track definition. It also describes the 'BED format' with four columns of data and the 'variableStep' format with two columns of data. The 'fixedStep' format is also described as single column data.

The 'Tools' panel on the right side of the interface shows a list of tools, with 'Wiggle-to-Interval converter' circled in pink. Other tools listed include 'Get Data', 'Send Data', 'ENCODE Tools', 'Lift-Over', 'Text Manipulation', 'Convert Formats', 'FASTA manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Aggregate datapoints', 'Compute phastOdds score', and 'Operate on Genomic Intervals'.

# Galaxy

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



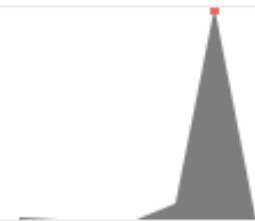
# Galaxy

- Very lame genome browser

||| FoxA1 Raw ▼

2537

1



||| ER raw ▼

Currently indexing... please wait

||| ER MACS peaks (sorted) ▼

Currently indexing... please wait

||| FoxA1 MACS peaks (sorted) ▼

Currently indexing... please wait

# 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

## NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

NGS: Indel Analysis

NGS: Peak Calling

- MACS Model-based Analysis of ChIP-Seq
- SICER Statistical approach for the Identification of ChIP-Enriched Regions
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

## *Genomic Regions Enrichment of Annotations Tool*

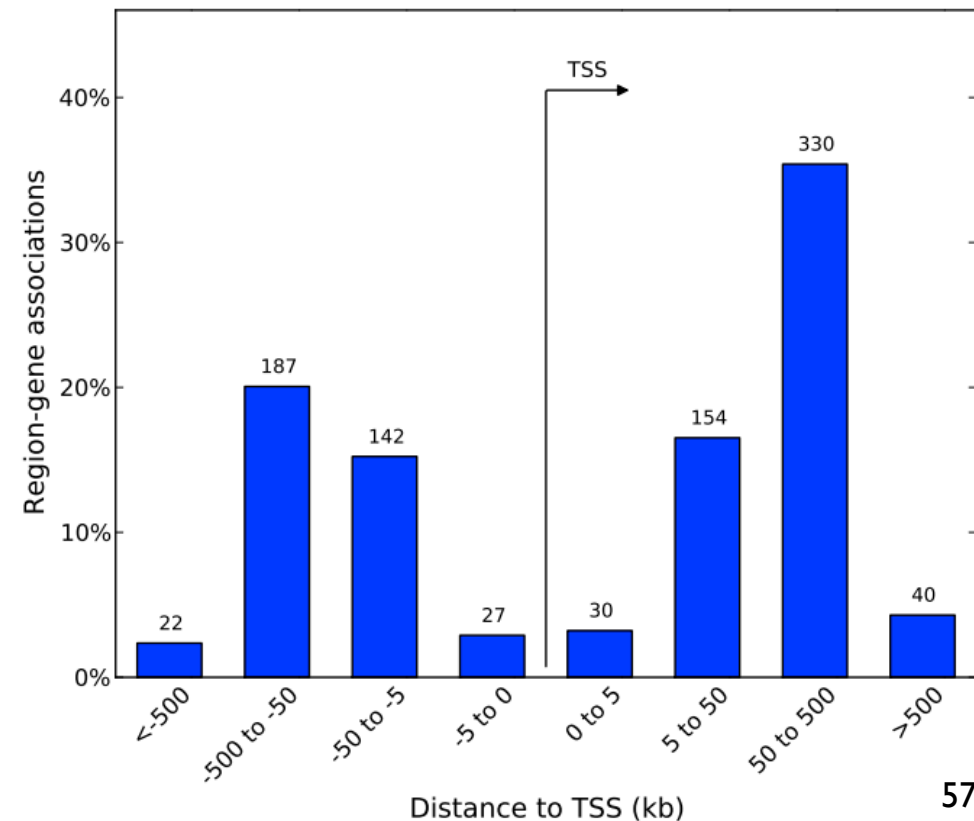
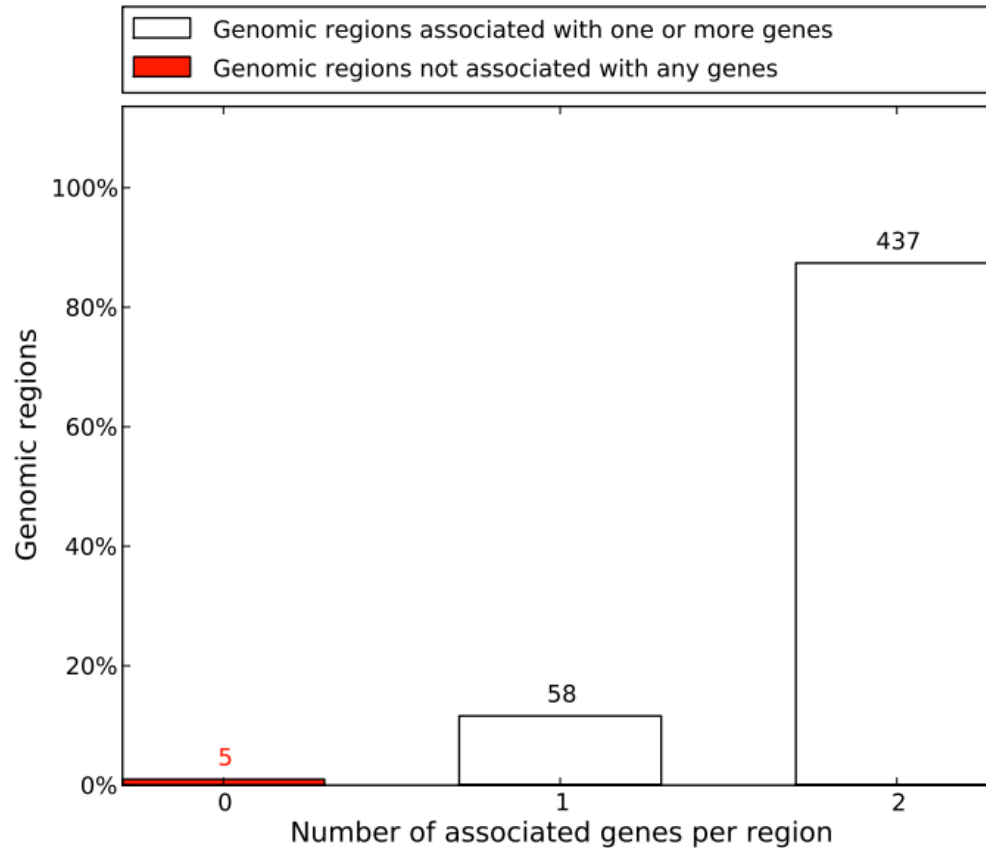
<http://great.stanford.edu>

- Web server for predicting the functions of cis-regulatory regions
- Let's try the top 500 bound regions of ER (convert height to integer)



# Genomic Regions Enrichment of Annotations Tool

- Associate each region to nearest genes



# Genomic Regions Enrichment of Annotations Tool

## Genomic region -> gene association table

[Download table as text.](#)

Region	Gene (distance to TSS)
<a href="#">MACS_peak_2412</a>	<a href="#">GREB1</a> (-35,170), <a href="#">E2F6</a> (-32,775)
<a href="#">MACS_peak_4803</a>	<a href="#">C9orf3</a> (+56,361), <a href="#">FANCC</a> (+534,636)
<a href="#">MACS_peak_729</a>	<a href="#">SYT8</a> (-37,790), <a href="#">HCCA2</a> (-32,383)
<a href="#">MACS_peak_225</a>	<a href="#">FCGR1B</a> (-548,445)
<a href="#">MACS_peak_4647</a>	<a href="#">MYC</a> (+123,952)
<a href="#">MACS_peak_1278</a>	<a href="#">STK24</a> (-73,396), <a href="#">SLC15A1</a> (+102,137)
<a href="#">MACS_peak_53</a>	<a href="#">RCC2</a> (-82,010), <a href="#">ARHGEF10L</a> (-19,263)
<a href="#">MACS_peak_2692</a>	<a href="#">TNF1</a> (-527,674), <a href="#">TNS1</a> (+556,340)
<a href="#">MACS_peak_3036</a>	<a href="#">ZNRF3</a> (-69,738), <a href="#">XBP1</a> (-13,592)
<a href="#">MACS_peak_2289</a>	<a href="#">ISYNA1</a> (-8,871), <a href="#">ELL</a> (+75,123)
<a href="#">MACS_peak_1945</a>	<a href="#">IGFBP4</a> (+4,959), <a href="#">TNS4</a> (+53,219)
<a href="#">MACS_peak_1143</a>	<a href="#">TMEM120B</a> (+36,403), <a href="#">RHOF</a> (+44,533)
<a href="#">MACS_peak_2202</a>	<a href="#">STK11</a> (-24,243), <a href="#">SBNO2</a> (-7,273)
<a href="#">MACS_peak_1844</a>	<a href="#">KLHDC4</a> (-24,301), <a href="#">SLC7A5</a> (+79,257)
<a href="#">MACS_peak_1048</a>	<a href="#">DYRK2</a> (-165,170), <a href="#">CAND1</a> (+214,281)
<a href="#">MACS_peak_4182</a>	<a href="#">KIAA0415</a> (-92,067), <a href="#">FOXK1</a> (+1,267)
<a href="#">MACS_peak_1947</a>	<a href="#">CCR7</a> (-11,276), <a href="#">SMARCE1</a> (+71,103)
<a href="#">MACS_peak_2977</a>	<a href="#">TMPRSS2</a> (-145,861), <a href="#">RIPK4</a> (+161,396)
<a href="#">MACS_peak_2194</a>	<a href="#">ATP9B</a> (-6,190), <a href="#">SALL3</a> (+82,932)
<a href="#">MACS_peak_3705</a>	<a href="#">IL6ST</a> (-434,434), <a href="#">MAP3K1</a> (-385,703)
<a href="#">MACS_peak_1395</a>	<a href="#">ZFP36L1</a> (+222,939), <a href="#">RAD51L1</a> (+750,337)
<a href="#">MACS_peak_2348</a>	<a href="#">CYP2B6</a> (-2,800)
<a href="#">MACS_peak_863</a>	<a href="#">NARS2</a> (-485,489), <a href="#">ODZ4</a> (+380,297)

## Gene -> genomic region association table

[Download table as text.](#)

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

# Genomic Regions Enrichment of Annotations Tool

- Then tests the associated genes for enrichment

**X** [GO Molecular Function](#)

Table controls:

Term Name
No results meet your chosen criteria.

**X** [Transcription Factor Targets](#)

Table controls:  Shown top results

Term Name	Binom Rank	Binom Raw P-Value
<a href="#">Targets of estrogen receptor alpha, identified by ChIP-DSL in MCF-7 cells</a>	1	3.5502e-9

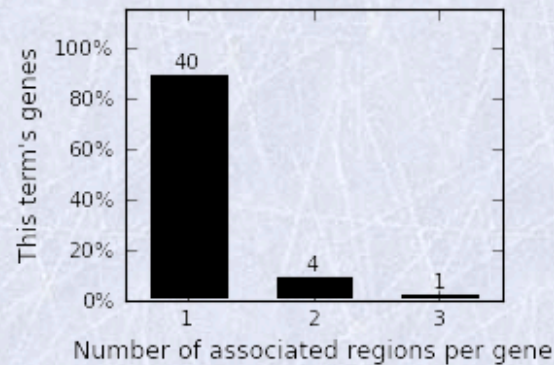
**X** [GO Biological Process](#)

Table controls:  Shown top results

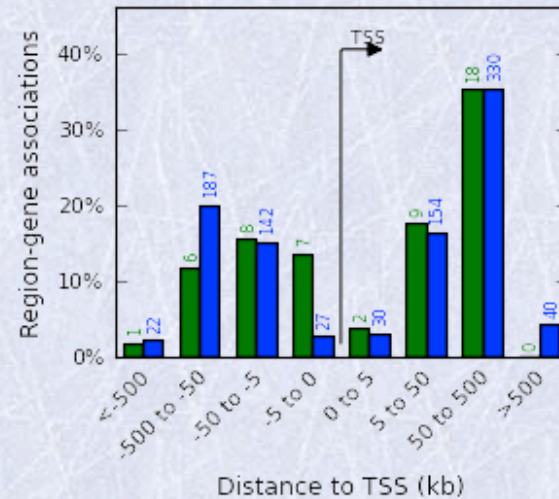
Term Name	Binom Rank	Binom Raw P-Value
<a href="#">gland development</a>	49	2.9902e-5
<a href="#">mammary gland development</a>	116	2.3782e-4



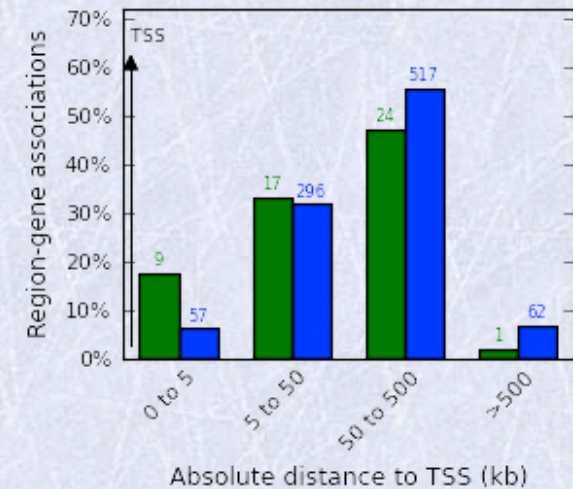
Number of associated regions per gene

[Download as PDF.](#)

Binned by orientation and distance to TSS

[Download as PDF.](#)

Binned by absolute distance to TSS

[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)
<a href="#">MACS_peak_2412</a>	<a href="#">GREB1</a> (-35,170)
<a href="#">MACS_peak_53</a>	<a href="#">ARHGEF10L</a> (-19,263)
<a href="#">MACS_peak_1048</a>	<a href="#">CAND1</a> (+214,281)
<a href="#">MACS_peak_2900</a>	<a href="#">TFAP2C</a> (+105,821)
<a href="#">MACS_peak_2415</a>	<a href="#">GREB1</a> (+5,985)
<a href="#">MACS_peak_4127</a>	<a href="#">ESR1</a> (-190,906)
<a href="#">MACS_peak_3899</a>	<a href="#">FOXC1</a> (+31,027)
<a href="#">MACS_peak_4082</a>	<a href="#">TPD52L1</a> (+1,254)
<a href="#">MACS_peak_3869</a>	<a href="#">STC2</a> (-206,271), <a href="#">BOD1</a> (+80,889)
<a href="#">MACS_peak_1715</a>	<a href="#">USP31</a> (+69,073)
<a href="#">MACS_peak_3062</a>	<a href="#">C1QTNF6</a> (-9,369)
<a href="#">MACS_peak_662</a>	<a href="#">CASP7</a> (+115)
<a href="#">MACS_peak_2338</a>	<a href="#">TIPARP</a> (+128,056)

#### This term's gene -> genomic region association table

[Download table as text.](#)

Gene	Region (distance to TSS)
<a href="#">ADAMTSL5</a>	<a href="#">MACS_peak_2207</a> (-450)
<a href="#">AMZ1</a>	<a href="#">MACS_peak_4178</a> (+9,213)
<a href="#">ARHGEF10L</a>	<a href="#">MACS_peak_53</a> (-19,263)
<a href="#">BCL3</a>	<a href="#">MACS_peak_2357</a> (-5,948)
<a href="#">BOD1</a>	<a href="#">MACS_peak_3869</a> (+80,889), <a href="#">MACS_peak_3868</a> (+143,681)
<a href="#">C1QTNF6</a>	<a href="#">MACS_peak_3062</a> (-9,369)
<a href="#">CALM1</a>	<a href="#">MACS_peak_1436</a> (+106,963)
<a href="#">CAND1</a>	<a href="#">MACS_peak_1048</a> (+214,281)
<a href="#">CASP7</a>	<a href="#">MACS_peak_662</a> (+115)
<a href="#">DCAKD</a>	<a href="#">MACS_peak_1964</a> (+8,243)
<a href="#">DICER1</a>	<a href="#">MACS_peak_1456</a> (+85,031)
<a href="#">FLS1</a>	<a href="#">MACS_peak_333</a> (+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 FoxAI and ER (chrI)
2. Upload the two datasets to UCSC (save as a session)
3. Download output of BOWTIE for ER and FoxAI (chrI) 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](mailto:tomkap@berkeley.edu)