

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 shows the GEO homepage with the NCBI logo at the top left and the GEO logo with "Gene Expression Omnibus" text at the top right. A navigation bar includes links for "GEO Publications", "FAQ", "MIAME", "Email GEO", and "Login". Below the navigation is a main content area with a "GEO navigation" section containing "QUERY" and "BROWSE" buttons. The "QUERY" section has four options: "DataSets", "Gene profiles", "GEO accession", and "GEO BLAST", each with a search input field and a "GO" button. The "BROWSE" section has two main categories: "DataSets" and "GEO accessions". "DataSets" leads to "Platforms", which is highlighted with a pink oval. "GEO accessions" leads to "Samples" and "Series". At the bottom is a "Submitter login" form with fields for "User id:" and "Password:", a "LOGIN" button, and links for "» New account" and "» Recover password". To the right is a "Site contents" sidebar with sections for "Public data" (listing Platforms: 9,370, Samples: 624,542, Series: 25,116, DataSets: 2,720), "Documentation" (links to Overview, FAQ, Find, Submission guide, Linking & citing, Journal citations, Construct a Query, Programmatic access, DataSet clusters, GEO announce list, Data disclaimer, GEO staff), and "Query & Browse" (links to Repository browser, SAGEmap, FTP site, GEO Profiles, GEO DataSets, Submit, New account).

GEO navigation

- QUERY**
 - DataSets
 - Gene profiles
 - GEO accession
 - GEO BLAST
- BROWSE**
 - DataSets
 - Platforms
 - GEO accessions
 - Samples
 - Series

Submitter login

User id: Password: [» New account](#) [» Recover password](#)

Platform

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](#)

Illumina Genome analyzer: hundreds of species

NCBI GEO Gene Expression Omnibus

NCBI » GEO » Repository browser » Platforms Login

Series Samples Platforms DataSets Summary Advanced search

Illumina Genome Analyzer II Search Export 191 platforms

<< < 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	Aedes aegypti				GEO	Sep 20, 2011
GPL14595	Illumina Genome Analyzer II (<i>Aedes albopictus</i>)	high-throughput sequencing	Aedes albopictus				GEO	Sep 20, 2011
GPL14367	Illumina Genome Analyzer II (<i>Bursaphelenchus xylophilus</i>)	high-throughput sequencing	Bursaphelenchus xylophilus				GEO	Sep 01, 2011
GPL14368	Illumina Genome Analyzer II (<i>Bursaphelenchus mucronatus</i>)	high-throughput sequencing	Bursaphelenchus mucronatus				GEO	Sep 01, 2011
GPL14364	Illumina Genome analyzer II (<i>Escherichia coli</i> E24377A)	high-throughput sequencing	Escherichia coli E24377A				GEO	Aug 31, 2011
GPL14358	Illumina Genome Analyzer II (<i>Drosophila willistoni</i>)	high-throughput sequencing	Drosophila willistoni				GEO	Aug 29, 2011
GPL14176	Illumina Genome Analyzer II (<i>Tetranychus urticae</i>)	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 (<i>Drosophila</i>)	high-throughput sequencing	Drosophila				GEO	Aug 11, 2011
GPL14012	Illumina Genome Analyzer II (<i>Limanda limanda</i>)	high-throughput sequencing	Limanda limanda				GEO	Aug 02, 2011
GPL14002	Illumina Genome Analyzer II (<i>Herpesvirus saimiri</i> (strain 11))	high-throughput sequencing	Herpesvirus saimiri (strain 11)	1	1	GEO	Jul 30, 2011	
GPL13999	Illumina Genome Analyzer II (<i>Arachis hypogaea</i>)	high-throughput sequencing	Arachis hypogaea				GEO	Jul 29, 2011
GPL14000	Illumina Genome Analyzer II (<i>Phaseolus vulgaris</i>)	high-throughput sequencing	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	<i>Homo sapiens</i>					
GPL14603	454 GS FLX Titanium (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>		16	2	GEO	Sep 21, 2011
GPL14583	Illumina Hiseq 2000 (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Sep 16, 2011
GPL14555	Illumina genome analyzer (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Sep 09, 2011
GPL13994	Illumina genome analyzer IIx (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>		67	1	GEO	Jul 29, 2011
GPL13978	Helicos Heliscope (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>		157	4	GEO	Jul 27, 2011
GPL13981	AB SOLiD System 4 (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Jul 27, 2011
GPL13873	AB SOLiD system 3.0 (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Jul 11, 2011
GPL13731	Illumina Genome Analyzer Iix (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Jun 16, 2011
GPL13484	454 Titanium (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	May 03, 2011
GPL13477	Illumina Genome Analyzer IIX (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>		8	2	GEO	May 02, 2011
GPL13393	AB SOLiD 4 System (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>		9	1	GEO	Apr 08, 2011
GPL13357	AB SOLiD System v3+ (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Mar 31, 2011
GPL13317	Heliscope (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Mar 23, 2011
GPL11436	AB SOLiD 3 Plus (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Jan 12, 2011
GPL11255	AB SOLiD System 3 (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Nov 30, 2010
GPL11154	Illumina HiSeq 2000 (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>		137	8	GEO	Nov 02, 2010
GPL10999	Illumina Genome Analyzer IIx (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>		988	34	GEO	Sep 29, 2010
GPL10400	454 GS (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>		4	3	GEO	May 06, 2010
GPL10329	Illumina Genome Analyzer (<i>Homo sapiens</i> ; <i>Mus musculus</i>)	high-throughput sequencing	<i>Homo sapiens</i> <i>Mus musculus</i>		2	1	GEO	Apr 14, 2010
GPL10297	Illumina Genome analyzer II (<i>Homo sapiens</i>)	high-throughput sequencing	<i>Homo sapiens</i>				GEO	Apr 07, 2010



HOME | SEARCH | SITE MAP

| GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > Accession Display

Not logged in | Login

Scope:

Format:

Amount:

GEO accession:

GPL10999

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

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- 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 HNF3 α). 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.

- [日本語要約](#)
- [print](#)
- [email](#)
- [download pdf](#)
- [download citation](#)
- [order reprints](#)
- [rights and permissions](#)
- [share/bookmark](#)

DNasel 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.



HOME | SEARCH | SITE MAP

GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > Accession Display

Not logged in | Login

Scope:

Format:

Amount:

GEO accession:

GPL10999

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
AAATGTTAACCTGAANAGCTGGATCCAGTCTGGTGTGTA
+FoxA1.1 HWUSI-EAS582_229:6:1:1:1235 length=42
BACCBBCBBCCBB:=!1=CAB7B@BBCCA<C?>A8B>@AAC
@FoxA1.2 HWUSI-EAS582_229:6:1:1:569 length=42
CAGTATGGAGGTGAATAAACAGCAGATGGCCTGGAAAGATACA
+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: split 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: Stalab 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			

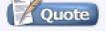
<ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByExp/sra/ERX%2FERX008%2FERX008600/ERR022028/>

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

Converting SRA to FASTQ

Screenshot of a forum post titled "How to convert sra-lite format to fastq?". The post is from user tbusch0000 on November 29, 2010, at 10:20 AM. The URL of the post is circled in pink.

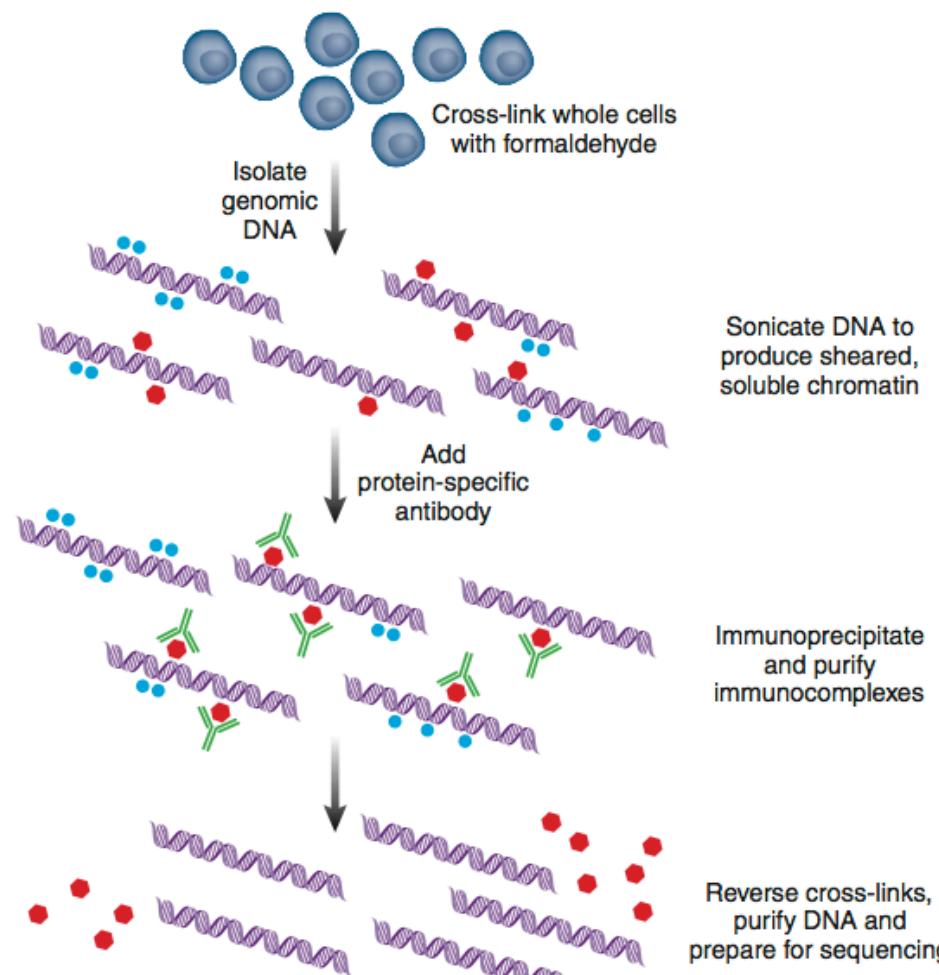
I am trying to dump sra-lite (sequence read archive) files to fastq format. On the NCBI Sequence Read Archive site it states:
...users are asked download runs of interest and execute dumps into the desired format using the SRA SDK toolkit available at <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software>
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!

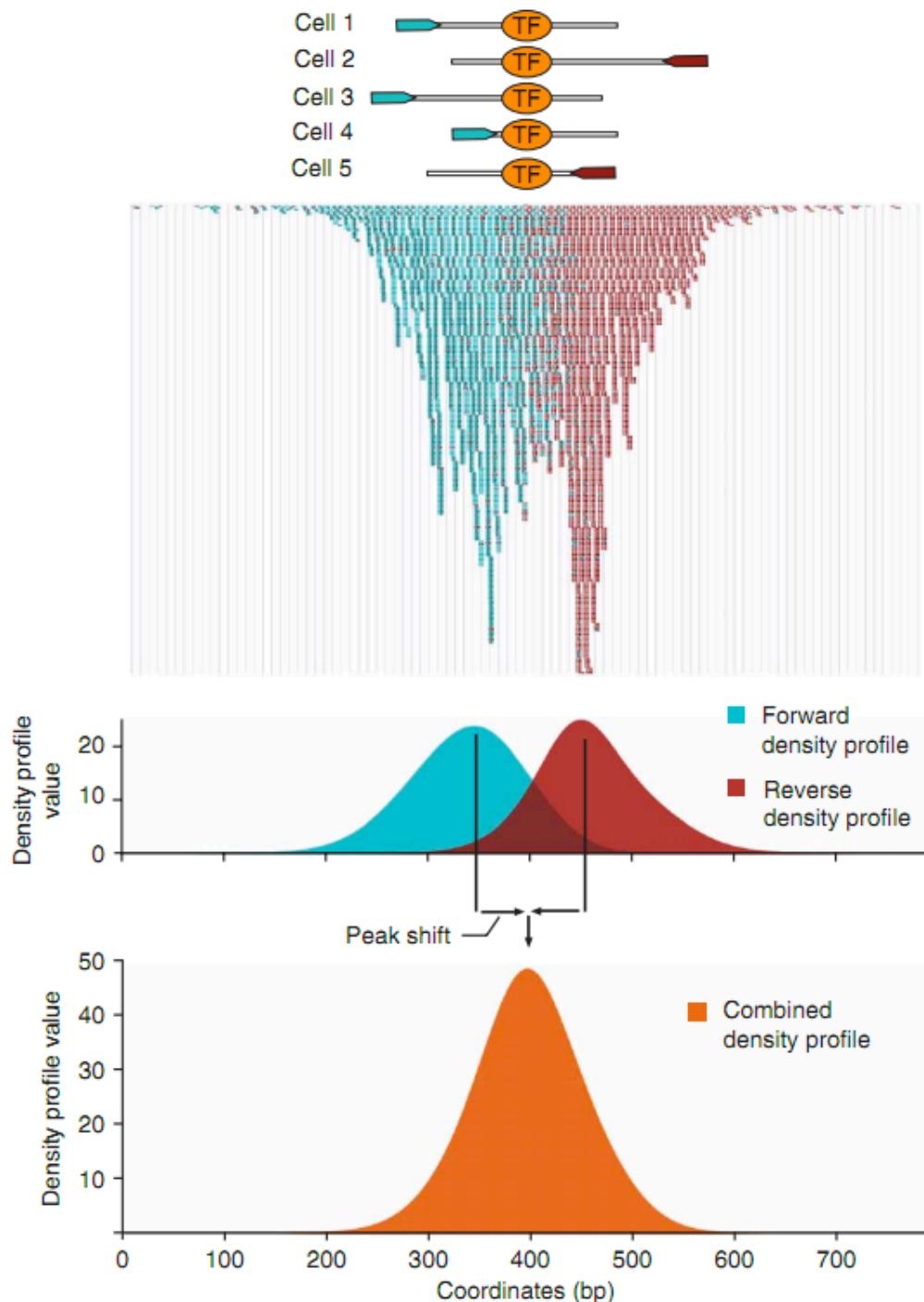
 Quote

<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 FoxA1.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.cbcn.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
AAATGTTAACCTGAANAGCTGGAATCCAGTCTGGTGTGTA
+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<ACCB@B7BAC=1!=:BBCCBBCB
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>

The screenshot shows the UCSC Genome Bioinformatics website. A purple circle highlights the 'Genomes' link in the top navigation bar. The main content area displays information about the site, news, and recent updates.

UCSC Genome Bioinformatics

Genomes (highlighted with a purple circle)

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

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

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.

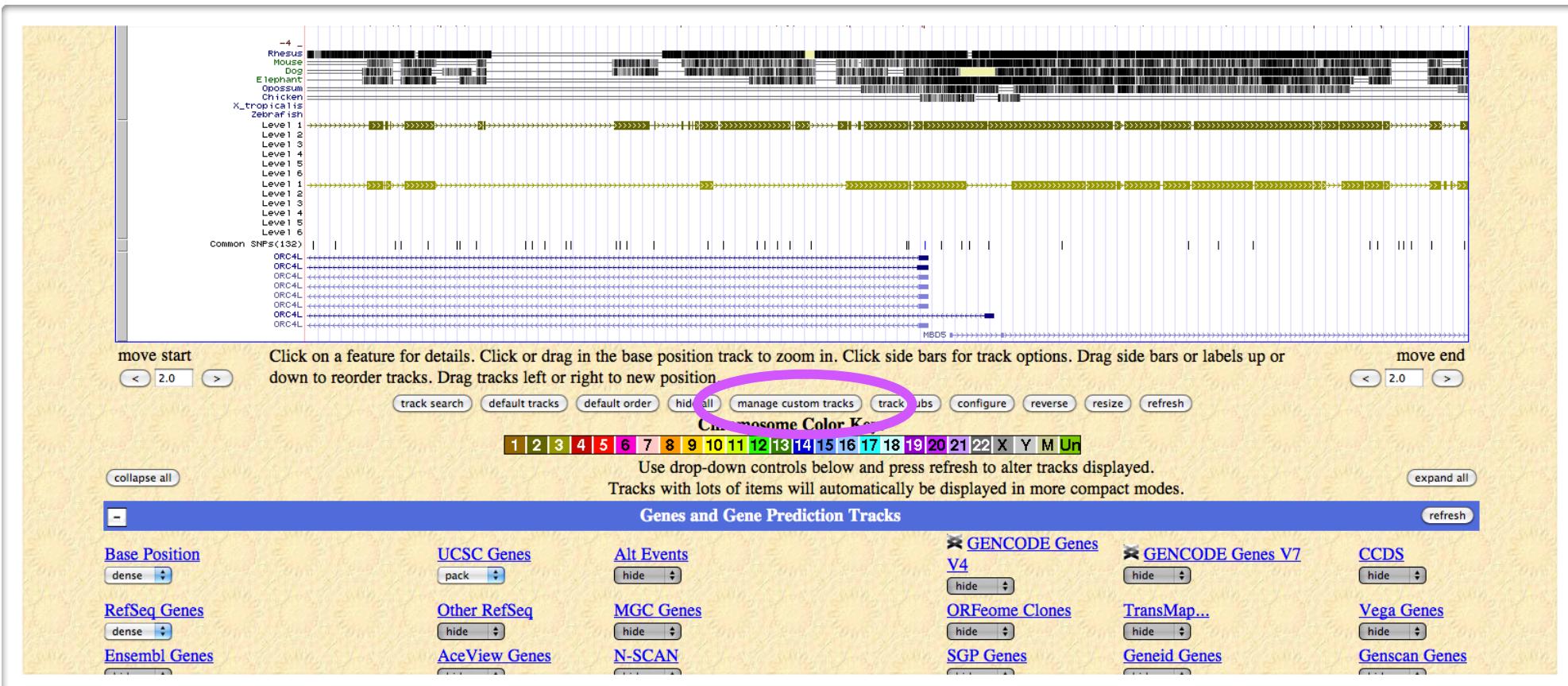
Scrolling left or right in the track window: You can now scroll (pan) horizontally through the tracks image by clicking on the image, dragging the cursor to the left or right, then releasing the mouse button. The view may be scrolled by up to one image width.

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_I 82.76`
- WIG
 - `track type=wiggle_0 name=... description=...`
`chrom=chr1 span=25`
`||8363 |`
`||8388 |`

Visualize with UCSC browser

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

Manage Custom Tracks

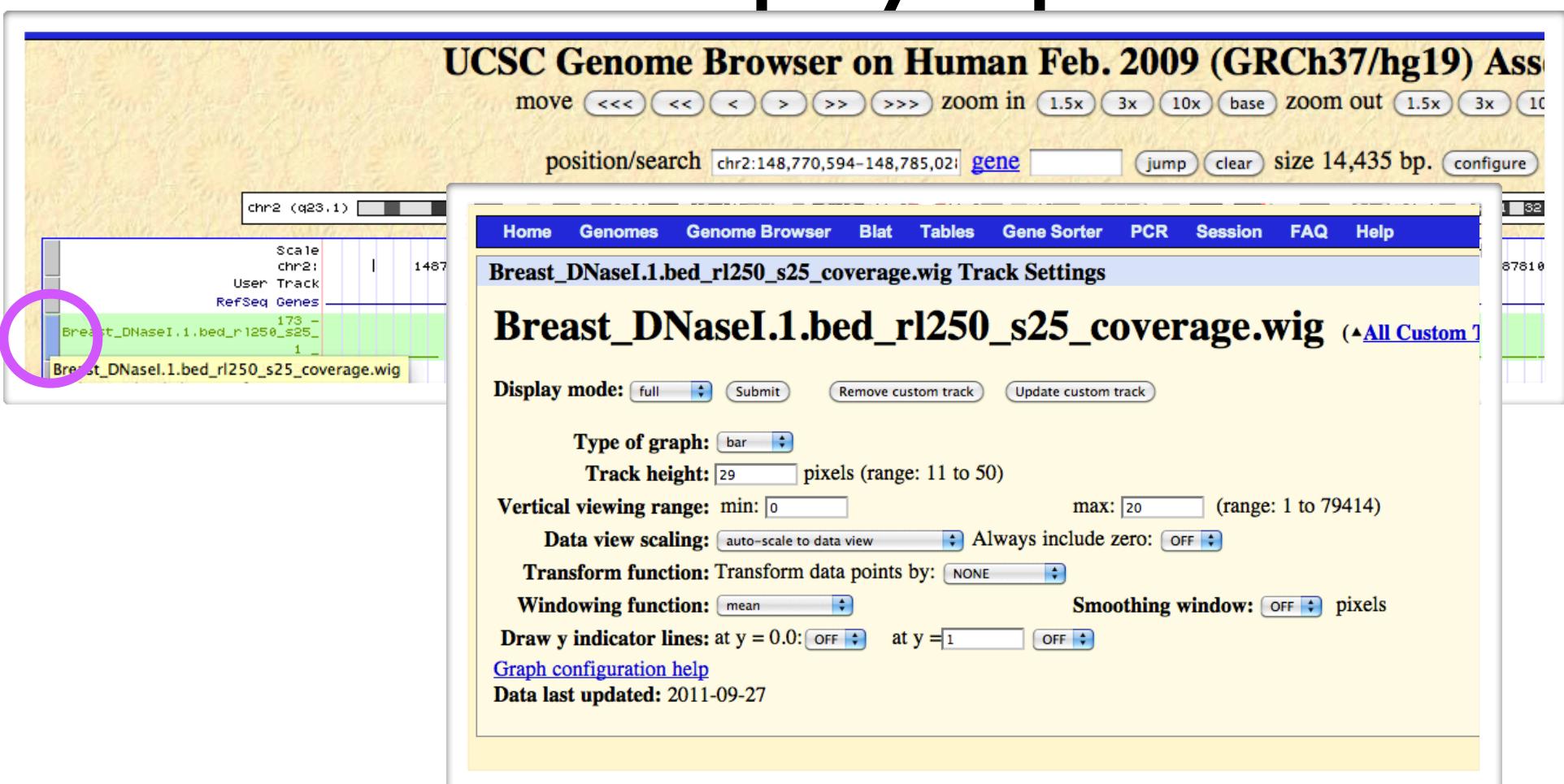
genome Human assembly Feb. 2009 (GRCh37/hg19) [hg19]

Name	Description	Type	Doc	Items	Pos	<input type="button" value="delete"/>	<input type="button" value="update"/>	<input type="button" value="add custom tracks"/>
User Track	User Supplied Track	bed		5059	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="button" value="add custom tracks"/>
FoxA1 peaks	FoxA1 peaks	bed		71450	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="button" value="go to genome browser"/>
ER peaks	ER peaks	bed		21324	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	<input type="button" value="go to table browser"/>
Breast_DNaseI.2_peaks	Breast_DNaseI.2 peaks	bed		84121	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
Breast_DNaseI.1_peaks	Breast_DNaseI.1 peaks	bed		85718	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
Breast_DNaseI.2	Breast_DNaseI.2	bedGraph		4090778	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
Breast_DNaseI.1	Breast_DNaseI.1	bedGraph		3620751	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
FoxA1	FoxA1	bedGraph		5784344	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
ER	ER	bedGraph		9449539	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
FoxA1.MACS	FoxA1.MACS	bed		45758	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
ER.MACS	ER.MACS	bed		5059	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
Breast_DNaseI.2.MACS	Breast_DNaseI.2.MACS	bed		52329	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
Breast_DNaseI.1.MACS	Breast_DNaseI.1.MACS	bed		51276	chr1:	<input type="checkbox"/>	<input type="checkbox"/>	
FoxA1.bed_r1250_s25_coverage.wig	FoxA1.bed_r1250_s25_coverage.wig	wiggle_0				<input type="checkbox"/>	<input type="checkbox"/>	
Breast_DNaseI.2.bed_r1250_s25_coverage.wig	Breast_DNaseI.2.bed_r1250_s25_coverage.wig	wiggle_0				<input type="checkbox"/>	<input type="checkbox"/>	
Breast_DNaseI.1.bed_r1250_s25_coverage.wig	Breast_DNaseI.1.bed_r1250_s25_coverage.wig	wiggle_0				<input type="checkbox"/>	<input type="checkbox"/>	
ER.bed_r1250_s25_coverage.wig	ER.bed_r1250_s25_coverage.wig	wiggle_0				<input type="checkbox"/>	<input type="checkbox"/>	

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

The screenshot shows the UCSC Genome Browser interface for the Human Feb. 2009 (GRCh37/hg19) Assembly. A purple circle highlights the "Session" button in the top navigation bar. Below the navigation bar, a search bar indicates a position of chr2:148,770,594-148,785,021. The main view displays genomic tracks for chromosomes 2 and 3, including RefSeq Genes, User Track, Breast_DNaseI.1, Breast_DNaseI.2, and FoxR1. On the right, a "Save Settings" dialog box is open. It contains fields for saving current settings as a named session ("name: hg19") and for saving to a local file ("file:"). It also includes sections for "Restore Settings" (using another user's session or a local file), and "Use settings from a URL".

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:

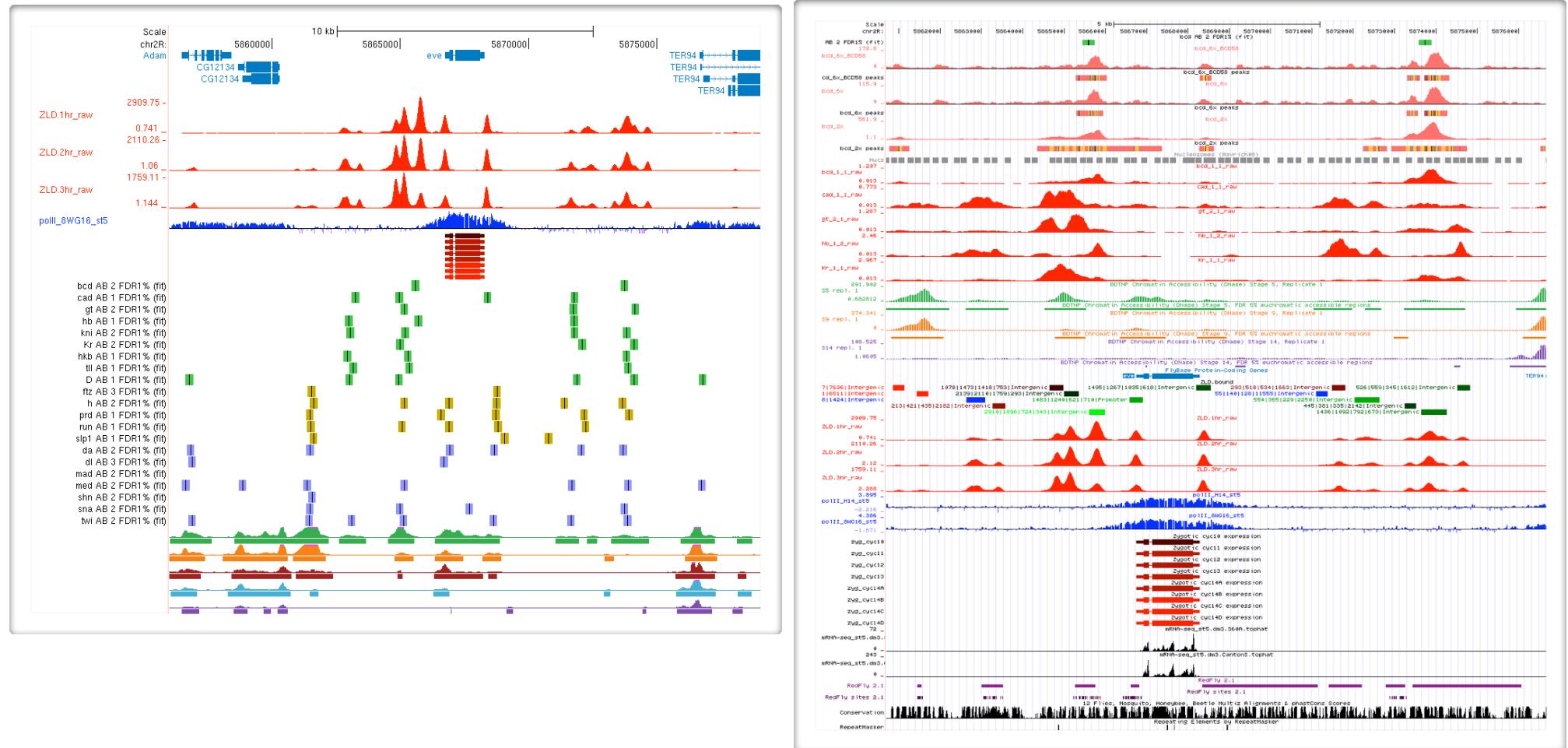
Use settings from a local file: No file chosen

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

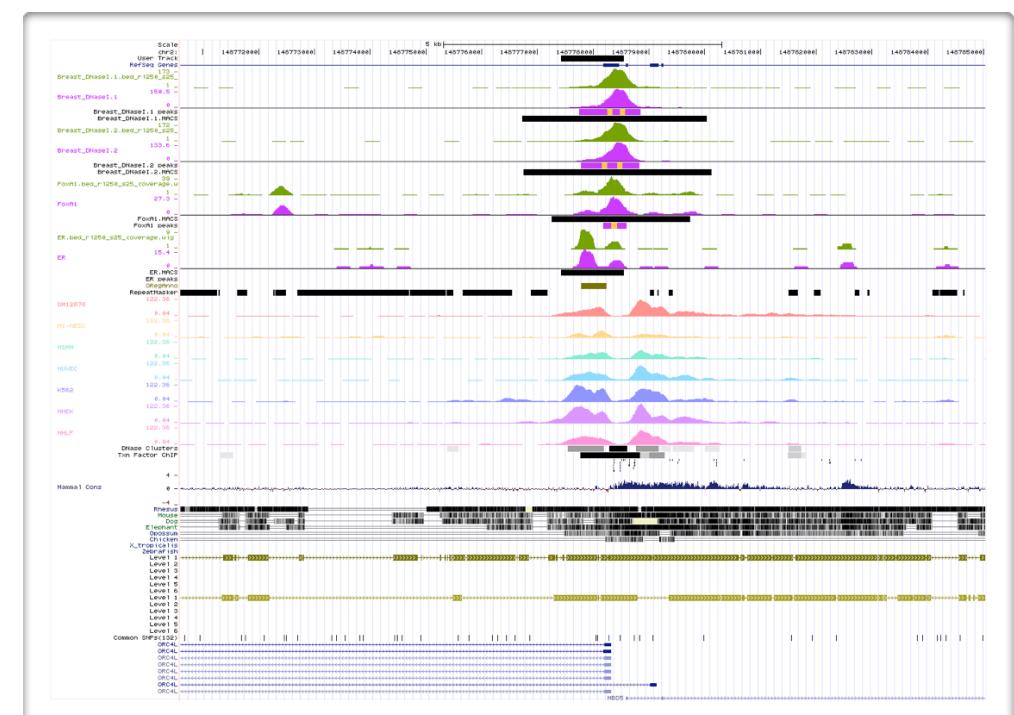
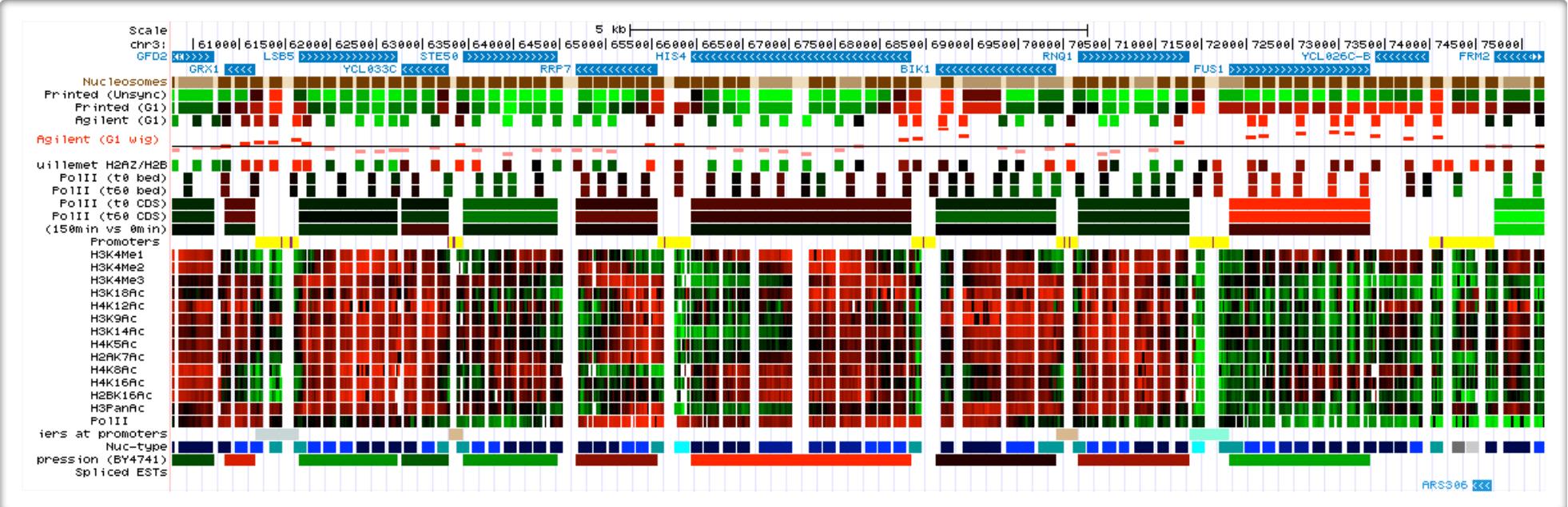
- Save as a session
- Share link with colleagues and friends
- Last 4 months since last access.
Access periodically, e.g. via www.followthatpage.com

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

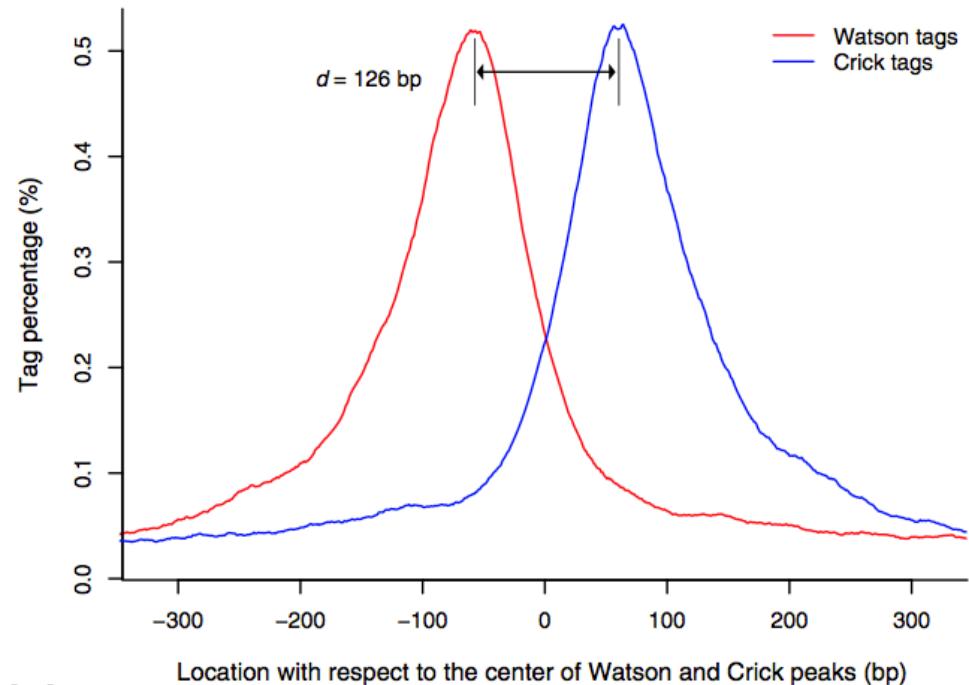
Genome Biology 2008, **9**:R137 (doi:10.1186/gb-2008-9-9-r137)

Abstract

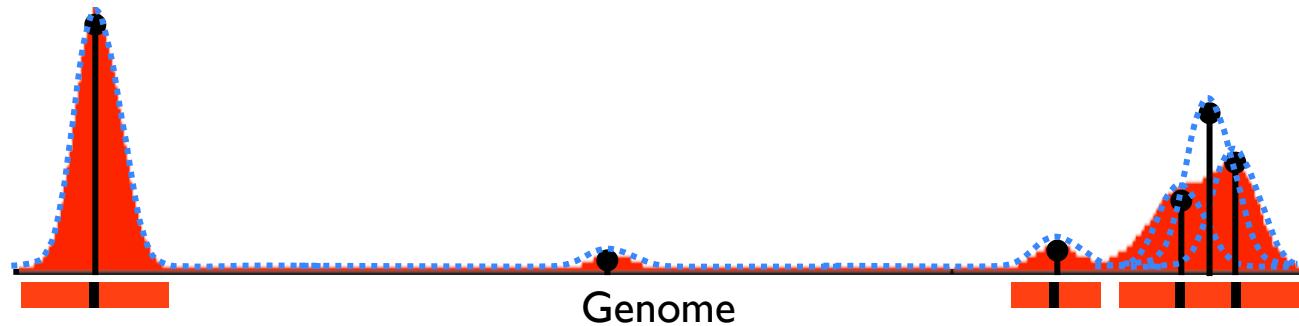
We present Model-based Analysis of ChIP-Seq data, MACS, which analyzes data generated by short read sequencers such as Solexa's Genome Analyzer. MACS empirically models the shift size of ChIP-Seq tags, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic 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
- `macsI4 -n FoxAI.MACS -t FoxAI.bed -g hs --off-auto --nomodel --shiftsize=125`

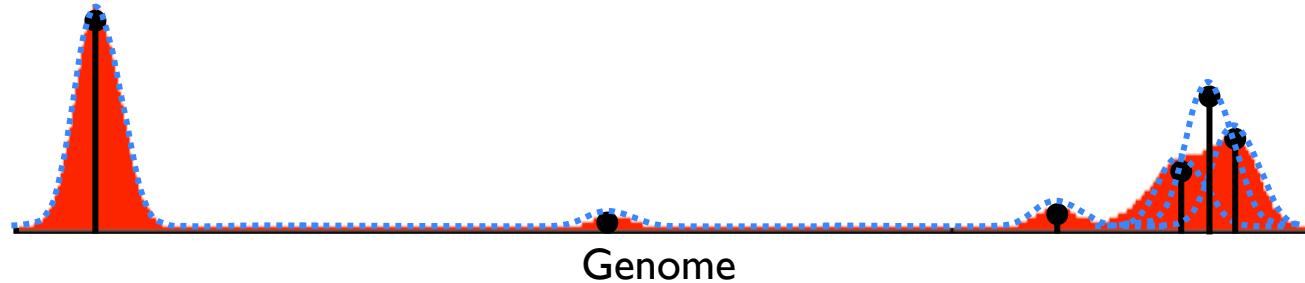


Grizzly Peak Fitting



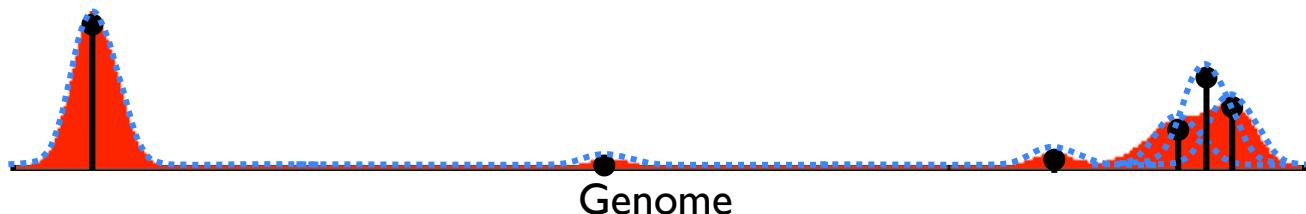
- Model-based, iterative, multi-peak peak finding algorithm
- MATLAB code, standalone executable*
- <http://rana.lbl.gov/software/grizzly>
- Capaldi, Kaplan et al, 2008, Harrison, Li, Kaplan et al, 2011

Grizzly Peak Fitting

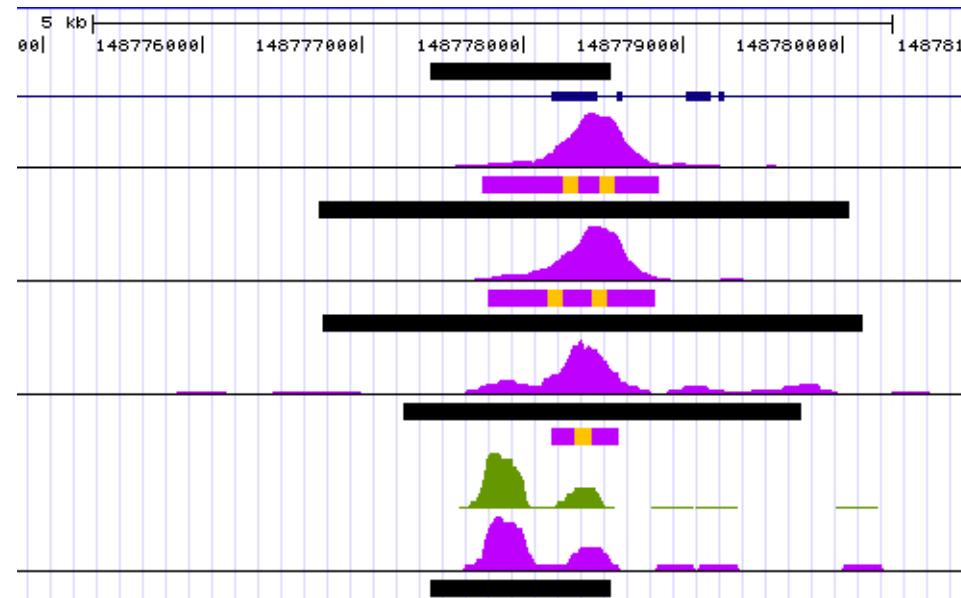


- Great in fitting the shape of a peak
- Given a motif, disentangle binding landscape into estimate occupancy at putative binding sites
- Bayesian confidence intervals for binding strengths
- Quite bad at stopping (separating real peaks from noise)
- pf0 FoxA1.bed 25 250 genome_hg19.mat 10

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

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. On the left, a sidebar titled "Tools" lists various genomic analysis tools: 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. Below the sidebar, the main content area features a large box titled "Galaxy 101" with the sub-titles "Start small" and "The very first tutorial you need". Below this, a section titled "Live Quickies" displays seven thumbnail cards, each representing a quick guide: "Mapping against custom genome" (Galaxy quickie # 10), "Illumina mapping: Single Ends" (Galaxy quickie # 11), "Illumina mapping: Paired Ends" (Galaxy quickie # 12), "Basic fastQ manipulation:" (Galaxy quickie # 13), "Advanced fastQ manipulation:" (Galaxy quickie # 14), "454 Mapping: Single End" (Galaxy quickie # 15), and "Uploading Data using FTP" (Galaxy quickie # 17). At the bottom of the page, there is a footer note about the Galaxy project's funding and a build identifier: "Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University." followed by "Galaxy build: \$Rev 6056:338ead4737ba\$".

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

Load called peaks

Galaxy

Analyze Data Workflow Shared Data Visualiz

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 such files, contact your 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 Galaxy server (see the "Log in" link above) and upload files via the "FTP" tab (you will need to provide your 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. Nar
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

40

Sort by height

Galaxy

Analyze Data Workflow Share

Tools Options ▾

[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 / X

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

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

Syntax

Select top 23 regions

Galaxy

Analyze Data Workflow Shared

Tools Options ▾

[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:

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

will produce:

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

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 shows the Galaxy web interface with the following components:

- Header:** Galaxy logo and Analyze tab.
- Left Panel (Tools):**
 - 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
 - Extract Genomic DNA using coordinates from assembled/unassembled genomes** (highlighted with a purple oval)
- Middle Panel (Tool Form):**
 - Extract Genomic DNA (version 2.2.2)**
 - Fetch sequences for intervals in:** 4: Top 23 peaks
 - Interpret features when possible:** Yes
 - Only meaningful for GFF, GTF datasets.**
 - Source for Genomic Data:** Locally cached
 - Output data type:** FASTA
 - Execute** button
- Right Panel (History):**
 - History tab, Unnamed history, 4.1 Mb
 - S: Seqs of top 23** dataset (23 sequences, format: fasta, database: hg19)
 - Download link: > Download_121482891_121485895_+
Sequence preview:
aatttotcagagaagtccaaatatatccacttgcatatc
tttgaacatgtccatcagaagatatgtcagctgtcg
atcattgcaaagaatttctgagaatgtttctgtctt
gttattttctttactacgataggctcaaaagggtcc
cagatttgcagaaggagtgttcaaacctgaactat

Look for motif with MEME

MEME Suite Menu
+ Submit A Job
+ Documentation
+ Downloads
+ User Support
+ Alternate Servers
+ Authors
+ Citing



Version 4.7.0

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

Data Submission Form

Required

Your e-mail address: Re-enter e-mail address:

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

Options

Description of your sequences:

MEME will find the optimum **number of sites** for each motif within the limits you specify here: Minimum sites (≥ 2) Maximum sites (≤ 300) Shuffle sequence letters

NEW Perform **discriminative motif discovery** – Enter the name of a file containing '**negative sequences**': No file chosen

Enter the name of a file containing a **background Markov model**: No file chosen

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

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 can check the status of your job at any time by clicking the **ACTIVATION** button.

ACTIVATION

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 \(format\)](#)
- [input sequences](#)

Click on any row to highlight sequence in all motifs.

Name	Strand	Start	p-value	Sites
hg19_chr17_1672500_1674182_+	-	75	7.57e-26	GGTTCAAGCC ATTCCTGCCCAGCTTACAGGCTGGGACTACAGGC GCCCACCAC
hg19_chr1_156440232_156443139_+	+	2373	1.67e-25	AGTTCAAGTG ATTCCTGCCCAGCTTACAGGCTGGGATTAACAGGC ACATGCCACT
hg19_chr17_73454188_73457495_+	-	822	2.58e-25	GGTTCAAGCC ATTCCTGCCCAGCTTACAGGCTGGGACTACAGGC ACCCGCCACA
hg19_chr9_110266406_110268805_+	+	1360	4.40e-25	GGTTCAAGCG ATTCCTGCCCAGCTTACAGGCTGGGACTACAGGC GCATGCTGCC
hg19_chr10_98622889_98625123_+	+	175	6.59e-25	GGTTCAAGTG ATTCCTGCCCAGCTTACAGGCTGGGATTAACAGGC ATGCCACCAC
hg19_chr17_61828830_61831978_+	-	2508	1.26e-23	GGTTCAAGCGA ATTCCTGCTTCAGCTTCTGAGTAGCTGGGATTACAGGC GTGACCCAC
hg19_chr9_128057870_128061033_+	+	2258	1.76e-23	GGTTCAAACA ATTCCTGCCCAGCTTCCCAGCTTAGCTAGGACTACAGGT GTACACCCAC
hg19_chr19_35495995_35498474_+	-	1143	5.42e-23	GGTTCAAGTG ATTCCTGCCCAGCTTCCCAGCTACAGGCTGGGATTACAGGC ACCCACACAC
hg19_chr20_55309000_55311968_+	+	589	6.77e-23	GGTTCAAGTG ATTCCTGCCCAGCTTCCCAGATACAGGCTGGGATTACAGGC GCCTGCCAC
hg19_chr1_40349932_40351562_+	+	1463	7.97e-23	GGTTCAAGCG ATTCCTGCCCAGCTTCTGAGTAGCTCGACTACAGGC ATTGGCACAC
hg19_chr5_40446525_40448658_+	-	23	1.20e-22	GGTTCACGTC ATTCCTGCCCAGCTTCCCAGGTAAGCTGGGCTACAGGC GCCCCCCCCA
hg19_chr19_54124466_54126761_+	-	1947	1.41e-22	GGTTCAAGCC ATTCCTGCTTCAGATACTGGGACTACAGGC GCCTGCCAC
hg19_chr3_193428681_193431475_+	+	2112	2.52e-22	GGCTCAAGCT ATTCCTGAGCTCAGCTTCCCAGCTACAGGCTGGGACTACAGGT GCATAAACAC
hg19_chr8_95540335_95542493_+	+	1636	3.97e-22	GGCTCAAGCGA ATTCCTGCCCAGCTCTGAGTAGCTGGGATTACAGGC GTGTAACAC
hg19_chr17_53509073_53512135_+	-	388	8.63e-22	GGCTCAAGTG CTTCCTGCCCAGCTTCCCAGTAGCTGGGACTACAGGC CTGGGCCAC
hg19_chr19_18140018_18141511_+	-	221	2.20e-19	GGTTCAAGCT ATTCCTCCCAGCTTCCCAGTAGCTGGGACCAACCCCC AGATAATTAC
hg19_chr13_37977646_37979671_+	+	1266	1.15e-18	GGTTCAAGCA AGTCTCTGCTCAGCATCCCATTAACTGGGTTAAAGGT ACCCATCAC

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

Messages

- [Processing Messages](#)
- [Error Messages](#)

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

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Tools', 'Options', 'Analyze Data', and 'Workflow'. On the left, a sidebar under 'Get Data' lists various sources like 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, and EpiGRAPH server. Below these are 'Send Data' and 'ENCODE Tools'. The main content area is titled 'Upload File (version 1.1.3)'. It has sections for 'File Format' (set to 'bed'), 'File' (with a 'Choose File' button pointing to 'ER.MACS_peaks.bed.bz2'), and 'URL/Text' (an empty text area). A note says 'TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed by site administrator'. Below this is a section for 'Files uploaded via FTP' which states 'Your FTP upload directory contains no files.' A note explains that the Galaxy server allows FTP uploads via address and password. At the bottom, there's an option 'Convert spaces to tabs' with a checked 'Yes' checkbox and a note about entering intervals by hand.

Galaxy

- Calculate base coverage for FoxA1

The screenshot shows the Galaxy web interface. The top navigation bar includes the Galaxy logo, a search bar, and tabs for "Tools", "Options", and "Analyze Data". The main content area is titled "Base Coverage (version 1.0.0)". A dropdown menu under "Compute coverage for:" shows "2: FoxA1 MACS peaks (sorted)". Below the dropdown is an "Execute" button. A tip message states: "TIP: If your dataset does not appear in the pulldown menu, click here to add columns." A descriptive text block says: "This operation counts the total bases covered by a set of intervals." A section titled "Screencasts!" provides links to video tutorials. On the left, a sidebar lists various tools under the "Tools" category, with "Base Coverage" highlighted and circled in pink. At the bottom right, a green box displays the results of the tool execution: "6: Base Coverage on data 2" (with icons for eye, copy, and close), "1 line", "format: txt, database: hg19", and the value "108992582".

Galaxy

Analyze Data

Tools Options Analyze Data

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, click here to add columns.

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

Screencasts!

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

6: Base Coverage on data 2

1 line

format: txt, database: hg19

108992582

Galaxy

● Intersect two databases

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 (highlighted with a pink circle)
- 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

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)

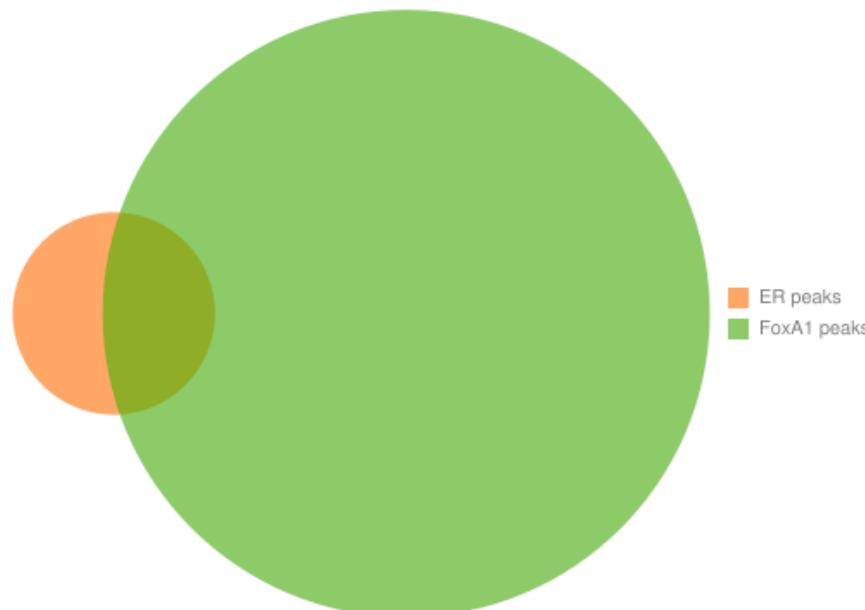
2,933 regions, 1 comments
format: bed, database: hg19
Info: Skipped 2 invalid lines of 1st dataset, 1st line #459:

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: Seqs 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 shows the Galaxy web interface with two main panels.

Left Panel: Upload File (version 1.1.3)

- Tools** menu is open, showing options like Get Data, Send Data, ENCODE Tools, Lift-Over, Text Manipulation, Convert Formats, FASTA manipulation, Filter and Sort, and Join, Subtract and Group.
- Get Data** section is selected, showing a list of servers:
 - Upload File from your computer (highlighted with a pink circle)
 - UCSC Eukaryote 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
- File Format:** wig (highlighted with a pink circle)
- File:** Choose File: ER.chr1.bed_r...erage.wig.b2z
- URL/Text:** Input field for URLs.
- Files uploaded via FTP:** File (Your FTP upload directory contains no files.)
- Convert spaces to tabs:** Yes
- Genome:** Human Feb. 2009 (GRCh37/hg19) (hg)
- Execute** button

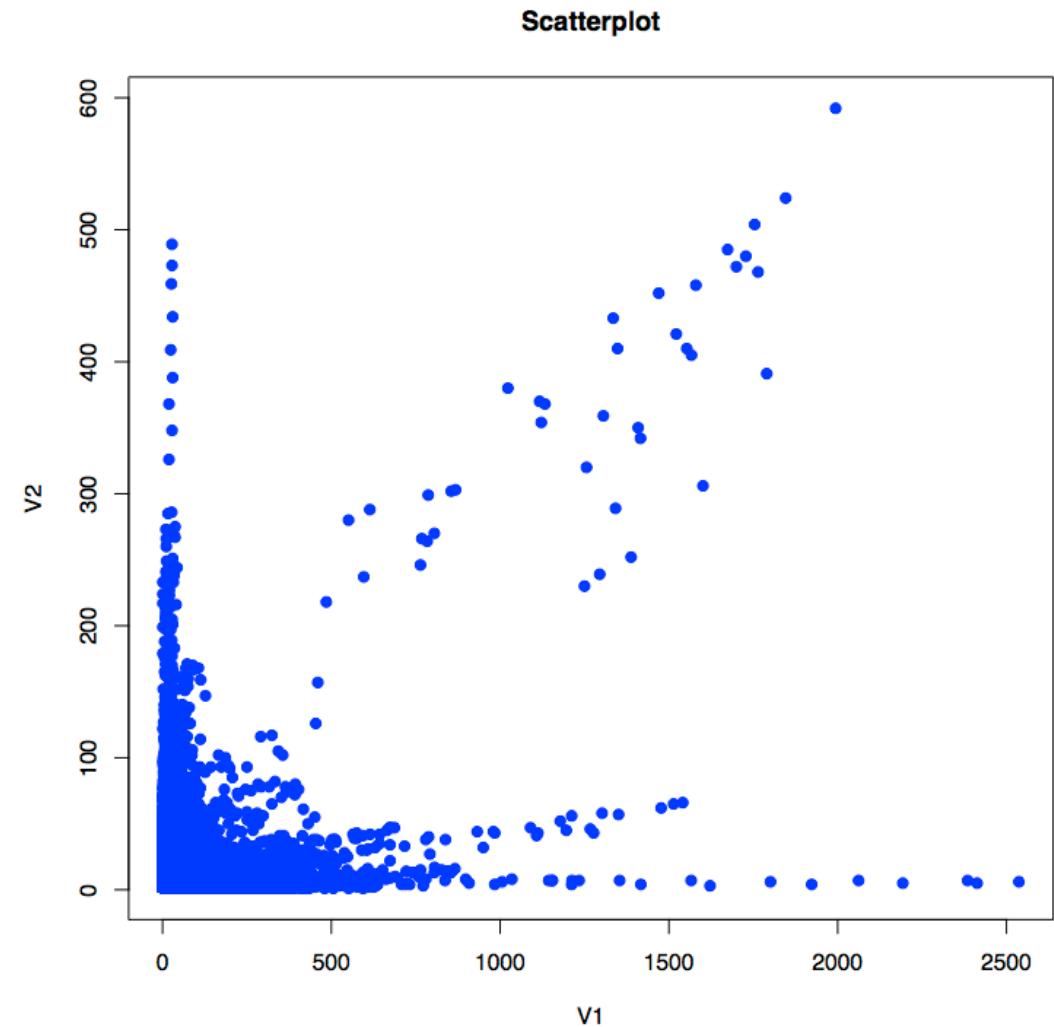
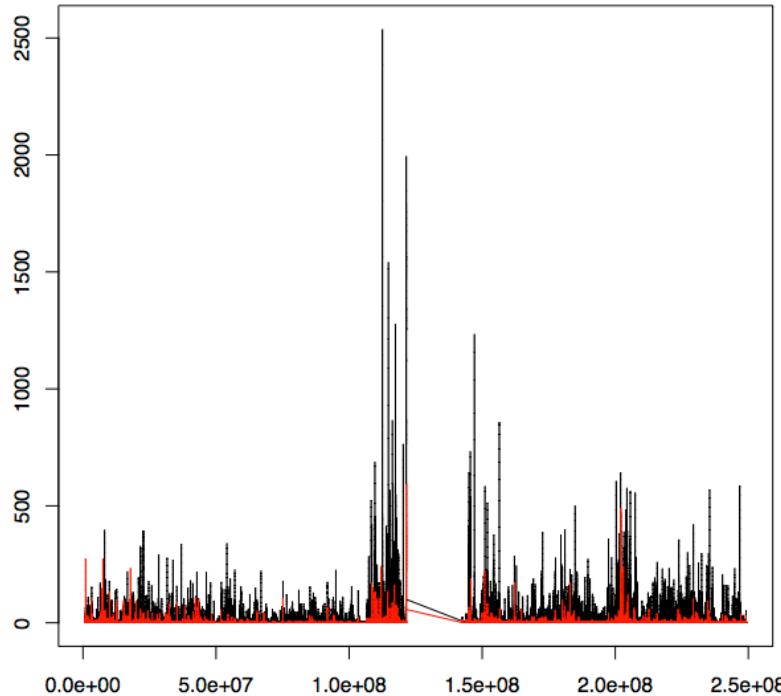
Right Panel: Wiggle-to-Interval (version 1.0.0)

- Tools** menu is open, showing options 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, Wiggle-to-Interval converter (highlighted with a pink circle), Aggregate datapoints such as phastCons, GERP, binCons, and others for a set of genomic intervals, Compute phastOdds score for each interval, Operate on Genomic Intervals, and Statistics.
- Convert:** 17: ER raw
- Execute** button
- Syntax**:
 - This tool converts wiggle data into interval type.
 - Wiggle format:** The .wig format is line-oriented. Wiggle data is preceded by a UCSC track definition line and followed by data lines.
 - BED format with no declaration line and four columns of data:**

```
chromA chromStartA chromEndA dataValueA
chromB chromStartB chromEndB dataValueB
```
 - variableStep** two column data; started by a declaration line and followed with chromosome position and data values.
 - fixedStep** single column data; started by a declaration line and followed with data values.

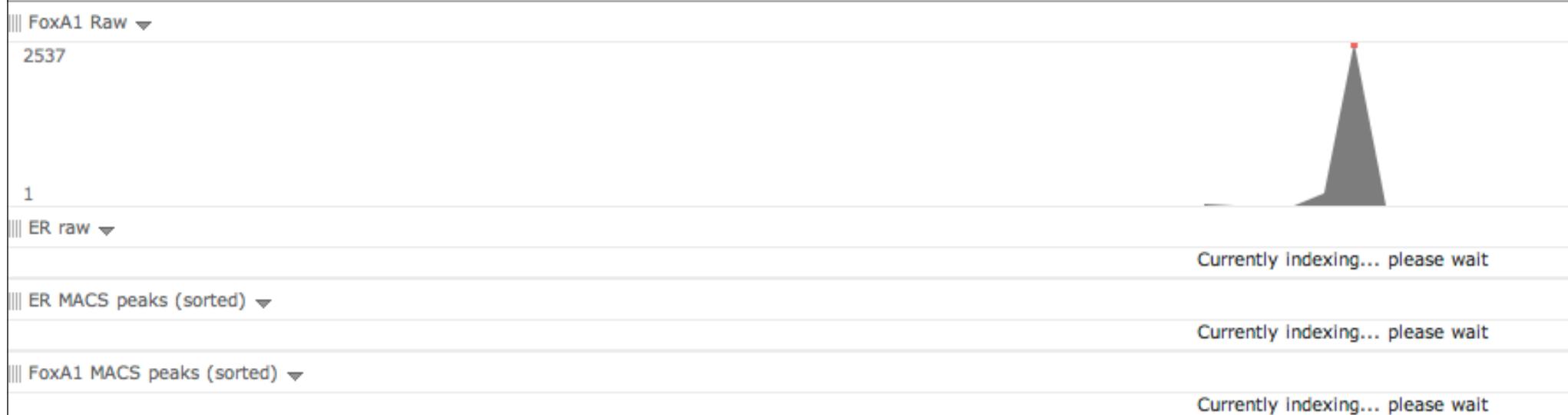
Galaxy

- 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

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](#)

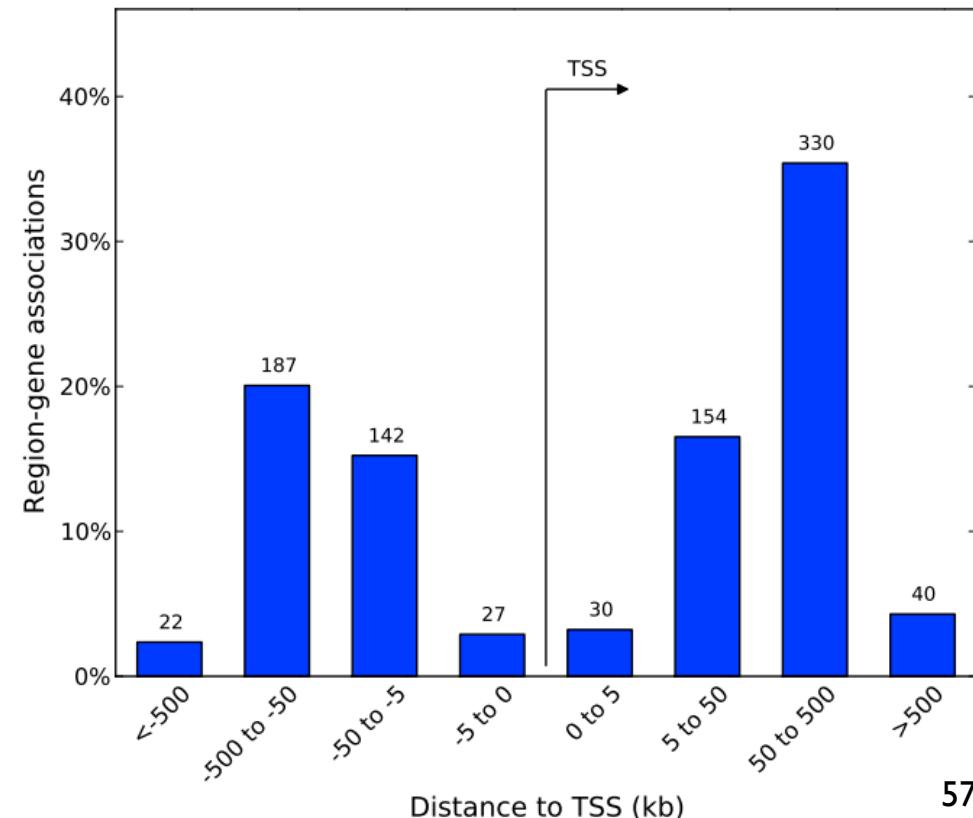
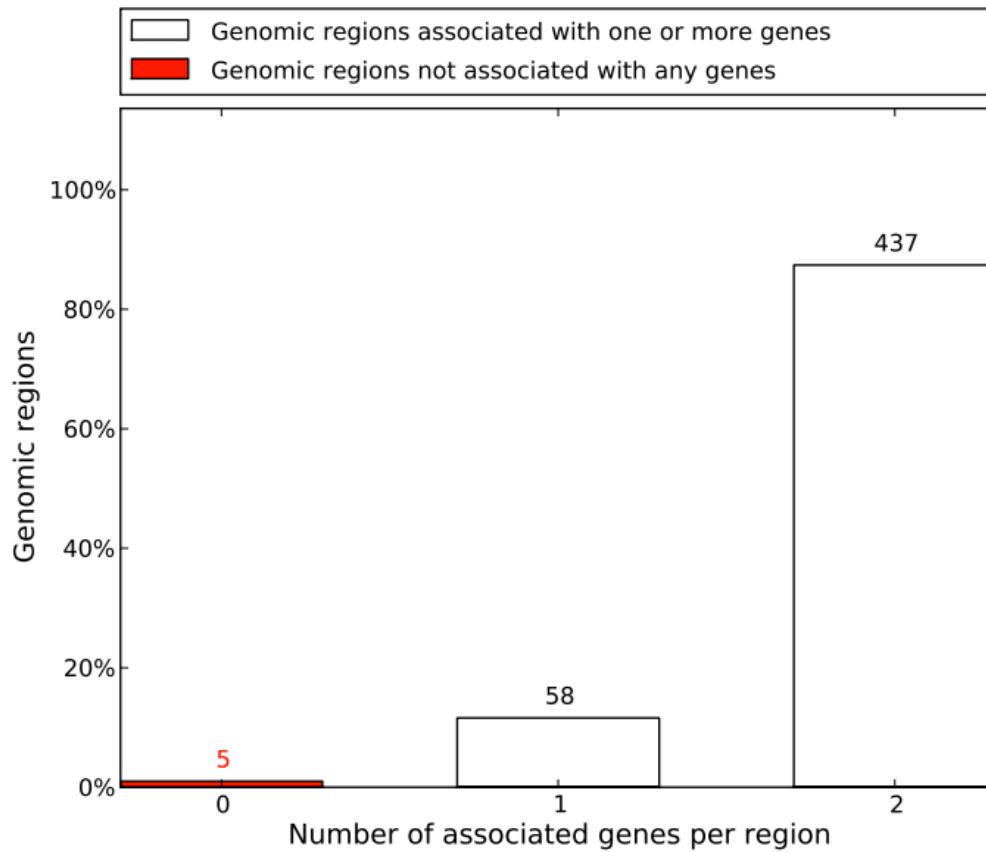
Genomic Regions Enrichment of Annotations Tool

<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)
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), RAD51L (+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)



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 r

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

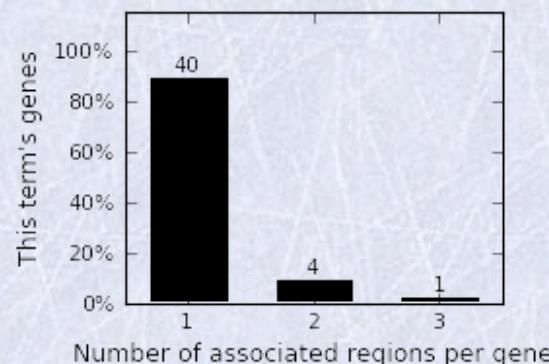
X GO Biological Process

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

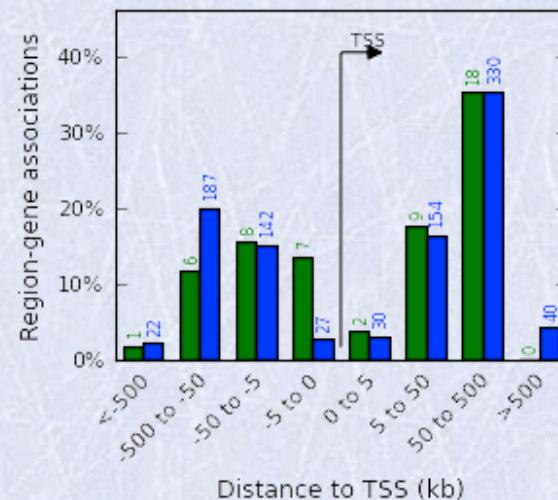
[Download as PDF.](#)



Binned by orientation and distance to TSS

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

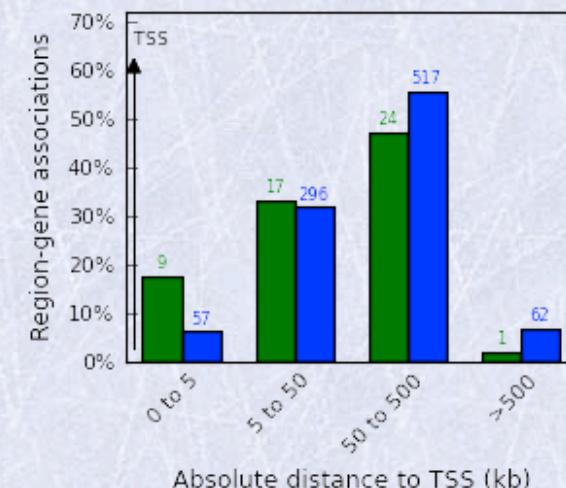
[Download as PDF.](#)



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

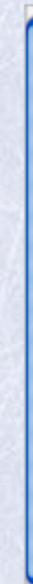
[Download table as text.](#)

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_2229	TIDBDP (+129,056)

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)
EFL5	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 FoxAI and ER (chr1)
2. Upload the two datasets to UCSC (save as a session)
3. Download output of BOWTIE for ER and FoxAI (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 DNase1 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