Analysis of ChIP-seq data
BIOC8145

Chongzhi Zang

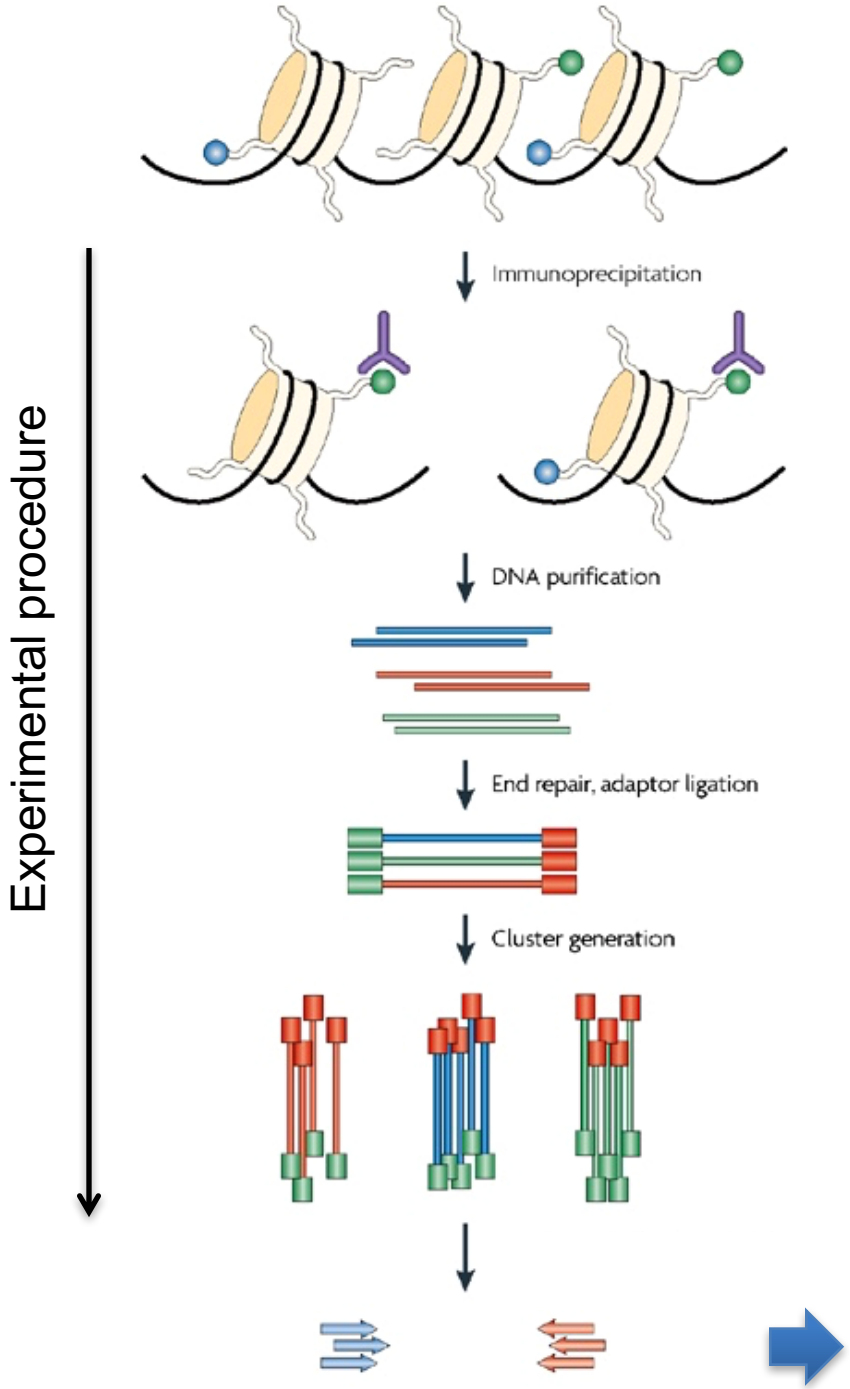
zang@virginia.edu
zanglab.org

BIOC8145 – Spring 2020
April 6-10, 2020

Outline

- Lecture 1
  - ChIP-seq technique introduction
  - ChIP-seq data analysis strategy
  - Read mapping (bowtie2)
  - Data formats
- Lecture 2
  - Peak calling (macs2)
  - Quality control
  - Data visualization (IGV)
- Lecture 3
  - Downstream analysis and integration
  - Online resources

Lecture 3: Downstream analysis, integration, and online resources



Biology

downstream analysis/integration

Peaks (bed)

peak calling

Pile-up for visualization
(bedGraph, wig, bigwig)

macs2

Non-redundant reads (sam/bam/bed)

redundancy assessment

Mapped reads (sam/bam/bed)

alignment (**bowtie2**)

Raw sequence reads (fastq)

Computational analysis

ChIP-seq: downstream analysis

1. DNA sequences at the peaks: motif discovery
2. Annotation of the peaks
3. Integration with other omics data/information for functional analyses

Position weight matrix (PWM) representation of DNA sequence motifs

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

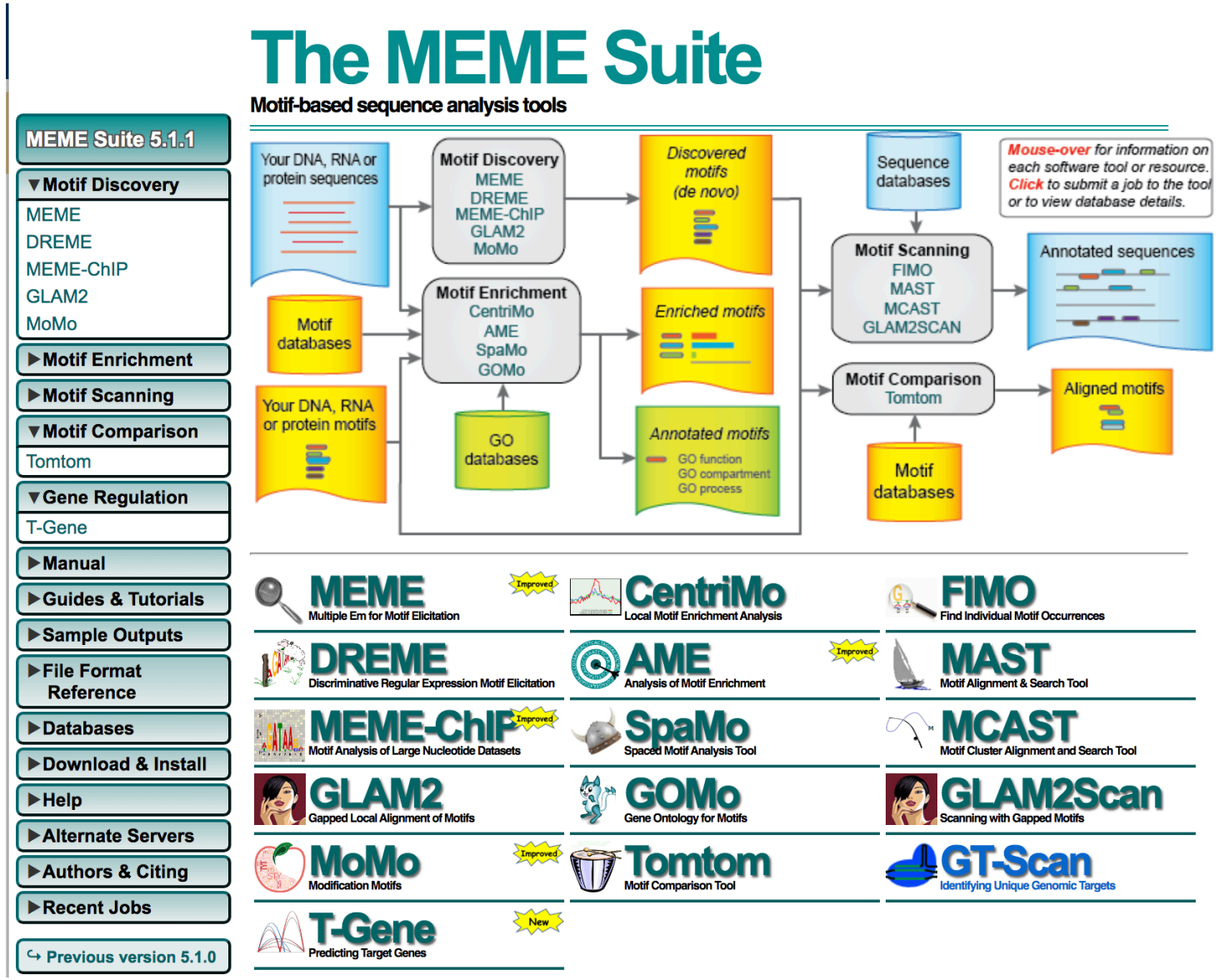
$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

$$R_i = \log_2(4) - H_i$$

$$H_i = -\sum_b f_{b,i} \times \log_2 f_{b,i}$$

MEME (meme-suite.org)



HOMER (homer.ucsd.edu)

Not Secure | homer.ucsd.edu/homer/introduction/basics.html

HOMER

Software for motif discovery and ChIP-Seq analysis

Introduction to HOMER

The best way to learn about HOMER is to go through the tutorial pages. We've tried to spell out what happens in each step and explain the "why". A brief description of the Motif Finding component of HOMER is found below. Explanation of the sequencing analysis components of HOMER are integrated into the tutorials.

General Introduction to Motif Discovery with HOMER

HOMER is a collection of tools that are commonly needed for the analysis of gene expression profiling (microarray) and genome-wide location analysis experiments (ChIP-Seq or ChIP-Chip). There are also routines for other types of sequencing experiments, such as DNase-Seq or GRO-Seq.

Some of the things HOMER does NOT DO is find differentially expressed genes (although it has some routines to help with this), cluster gene expression profiles, or search for all the instances Transfac motifs in order to make you hopelessly confused!!! The idea was not to completely reinvent the wheel if possible.

Unfortunately, HOMER must be run as a command-line tool, and may be difficult to use if you are new to UNIX. While commands have been distilled to be as simple and user-friendly as possible, basic knowledge of the UNIX environment and file system is critical (but can probably be learned quickly after typing ﾟunix tutorialﾟ into google). I am proud to say that may of the people using HOMER are completely new to UNIX, so it is indeed possible. In addition, a spreadsheet program (i.e. EXCEL) is needed to graph and visualize some of the results produced by HOMER.

Below is a description of how motif analysis is executed with HOMER. Documentation describing the steps of analysis for Next-Gen Sequencing (or genomic position analysis) or Microarrays (gene-based analysis) are covered in separate sections.

***De Novo* Motif Discovery Strategy**

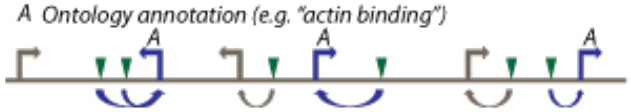
GREAT (great.stanford.edu)

GREAT predicts functions of *cis*-regulatory regions.

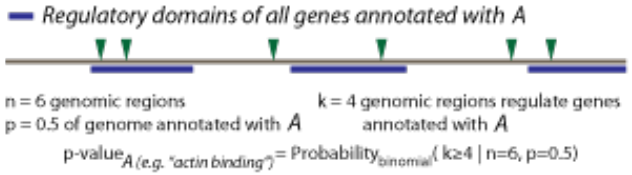
1. **Input:** A set of Genomic Regions (such as transcription factor binding events identified by ChIP-Seq).

2. GREAT associates both proximal and distal input Genomic Regions with their putative target genes.

3. GREAT uses gene Annotations from numerous ontologies to associate genomic regions with annotations.



4. GREAT calculates statistical Enrichments for associations between Genomic Regions and Annotations.



5. **Output:** Annotation terms that are significantly associated with the set of input Genomic Regions.

SRF peaks regulate genes involved in:

Ontology term	**p-value**
Actin cytoskeleton	10^{-9}
FOS gene family	10^{-8}
TRAIL signaling	10^{-7}

6. Users can create UCSC custom tracks from term-enriched subsets of Genomic Regions. Any track can be directly submitted to GREAT from the UCSC Table Browser.

ChIPseeker: an R/Bioconductor package

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

Home » Bioconductor 3.10 » Software Packages » ChIPseeker

ChIPseeker

platforms all | rank 123 / 1823 | posts 2 / 0 / 1 / 0 | in Bioc 6 years
build warnings | updated since release | dependencies 152

DOI: 10.18129/B9.bioc.ChIPseeker

ChIPseeker for ChIP peak Annotation, Comparison, and Visualization

Bioconductor version: Release (3.10)

This package implements functions to retrieve the nearest genes around the peak, annotate genomic region of the peak, statstical methods for estimate the significance of overlap among ChIP peak data sets, and incorporate GEO database for user to compare the own dataset with those deposited in database. The comparison can be used to infer cooperative regulation and thus can be used to generate hypotheses. Several visualization functions are implemented to summarize the coverage of the peak experiment, average profile and heatmap of peaks binding to TSS regions, genomic annotation, distance to TSS, and overlap of peaks or genes.

Author: Guangchuang Yu [aut, cre], Yun Yan [ctb], Hervé Pagès [ctb], Michael Kluge [ctb], Thomas Schwarzl [ctb], Zhougeng Xu [ctb]

Maintainer: Guangchuang Yu <guangchuangyu at gmail.com>

Citation (from within R, enter `citation("ChIPseeker")`):

Yu G, Wang L, He Q (2015). "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization." *Bioinformatics*, **31**(14), 2382-2383. doi: 10.1093/bioinformatics/btv145.

Documentation »

Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

Support »

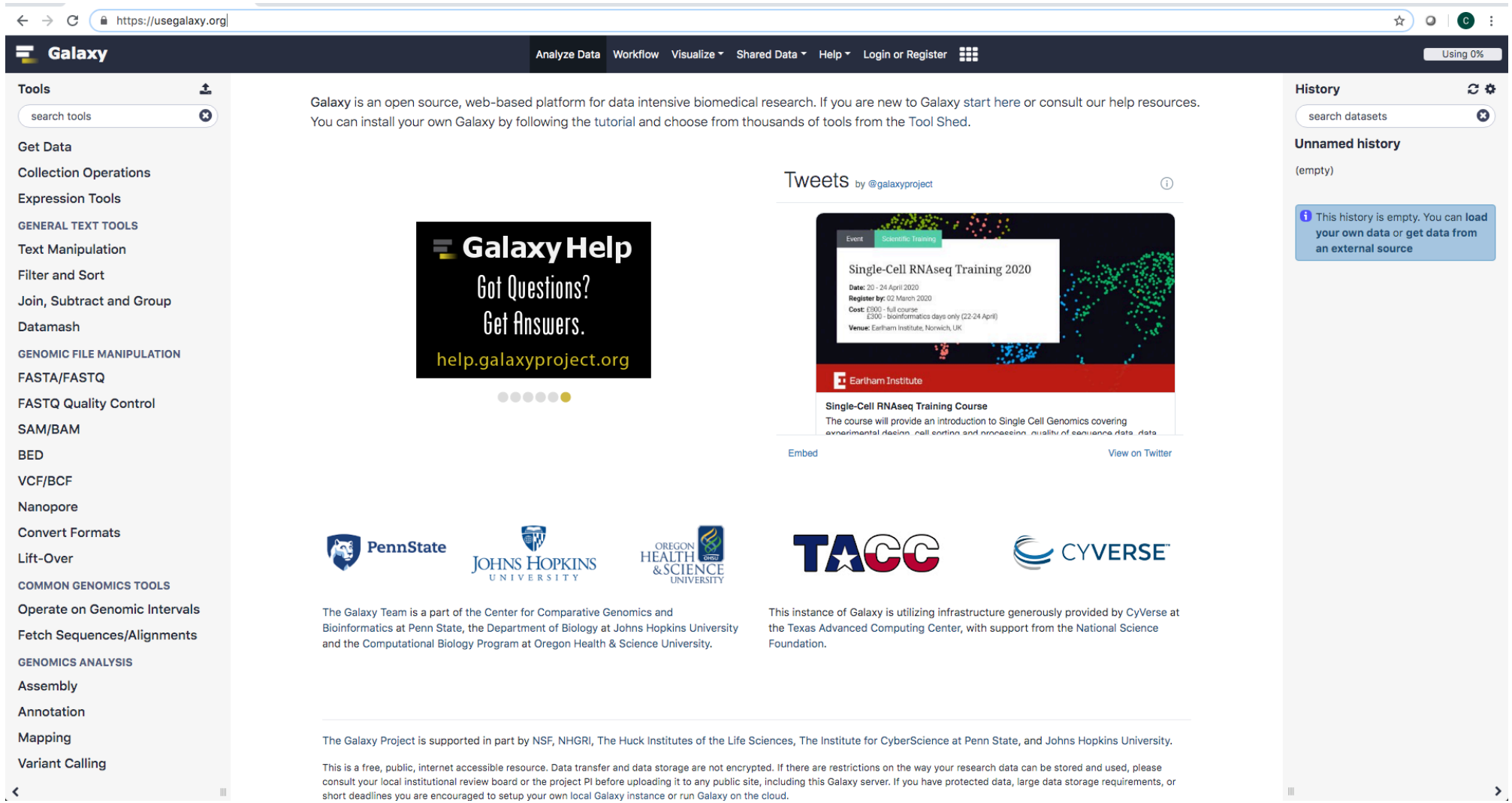
Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

ChIP-seq: online resourses

Galaxy: web-based analysis platform

- https://usegalaxy.org/



Cistrome, a Galaxy-based platform for ChIP-seq analysis

- http://cistrome.org/ap/

Galaxy / Cistrome / Cistrome

Chongzhi

cistrome.org/ap/root

Galaxy / Cistrome

Analyze Data | Workflow | Shared Data | Lab | Visualization | Help | User

Using 30.3 GB

Tools

search tools

CISTROME TOOLBOX

Import Data

Upload File from your computer

CistromeFinder Import from Cistrome Finder

CistromeCR Import from Cistrome Chromatin Regulator

Expression CEL file packager can download .cel files from GEO by given GSM IDs and prepare a cel.zip file for expression analysis.

GenomeSpace import from file browser

Data Preprocessing

Gene Expression

Integrative Analysis

Liftover/Others

GALAXY TOOLBOX

Get Data

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Upload File (version 1.1.4)

File Format:

Auto-detect

Which format? If for expression data, choose cel.zip or xys.zip. See help below

File (Please avoid Windows format text file):

Choose File No file chosen

TIP1: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or ASPERA (please read the instruction). TIP2: If you want to upload expression data, please read the instruction and specify cel.zip or xys.zip for file format.

URL/Text:

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

Files uploaded via ASPERA:

File	Size	Date
Your ASPERA upload directory contains no files.		

This Galaxy server allows you to upload files via ASPERA. To upload some files, log in to the ASPERA server at **cistrome.dfci.harvard.edu** using your Cistrome credentials (email address and password).

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand.

Genome:

Human Dec. 2013 (GRCh38/hg38) (hg38)

Execute

History

Unnamed history

329.0 MB

68: Heatmap log

67: Heatmap k-means cl assified regions

66: Heatmapr R script

65: Heatmap image

64: Heatmap log

63: Heatmap k-means cl assified regions

62: Heatmapr R script

61: Heatmap image

60: Heatmap log

59: Heatmap k-means cl assified regions

58: Heatmapr R script

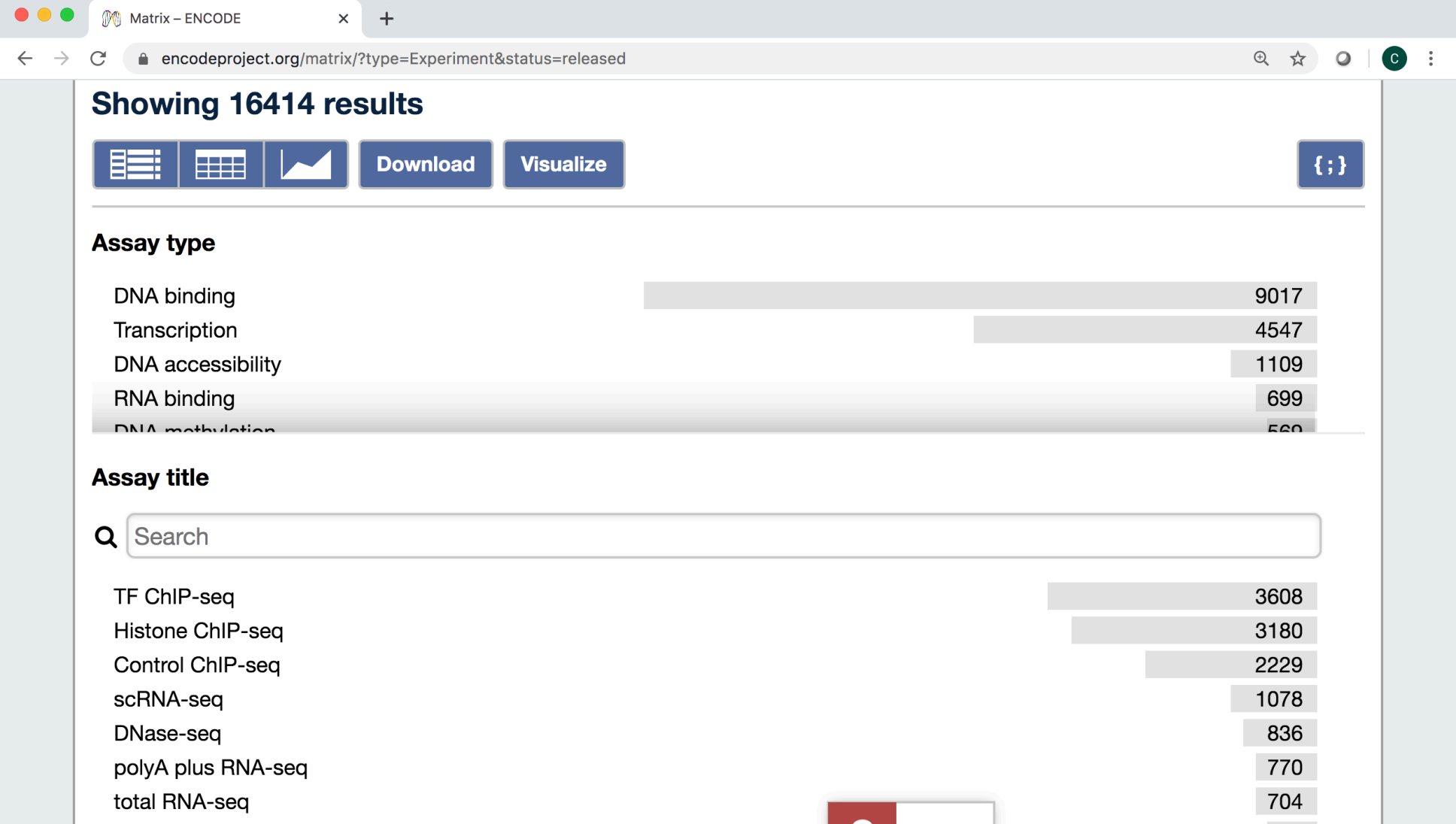
57: Heatmap image

56: Heatmap log

55: Heatmap k-means cl assified regions

ENCODE

https://www.encodeproject.org/



Cistrome Data Browser

http://cistrome.org/db/

Cistrome DB

Not Secure | cistrome.org/db/#/

Cistrome Data Browser | Home | Documentation | About | Statistics | Batch download | ToolKit | Cistrome-GO | Liu Lab

Cistrome Data Browser

Tips

- Check what factors regulate your gene of interest, what factors bind in your interval or have a significant binding overlap with your peak set. Have a try at CistromeDB Toolkit.
- If you have a Transcription Factor ChIP-seq (and TF perturbed expression) data, Cistrome-GO help you predict the function of this TF.
- Please help us curate the samples which has incorrect meta-data annotation by clicking the button on the inspector page. Thank you!

Containing word(s): Search Options

Species	Biological Sources	Factors
All	All	All
Homo sapiens	1-cell pronuclei	AATF
Mus musculus	1015c	ABCC9
	10326	ACSS2
	1064Sk	ACTB
	106A	ADNP

Results

Batch	Species	Biological Source	Factor	Publication	Quality Control
☐	Homo sapiens	HeLa; Epithelium; Cervix	BTAF1	Johannes F, et al. Bioinformatics 2010	

Mei *et al. Nucleic Acids Res.* 2017
Zheng *et al. Nucleic Acids Res.* 2018

BART: TF prediction using public ChIP-seq data

http://bartweb.org

BART: Binding Analysis for Regulation of Transcription

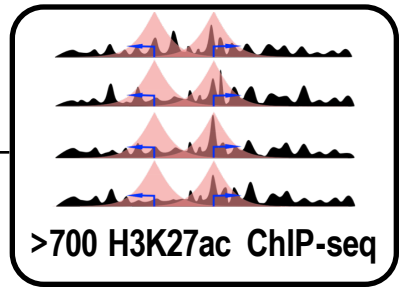
User input

Gene list

ChIP-seq

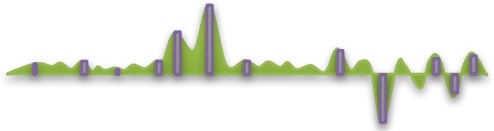
Region set

Hi-C



Adaptive Lasso regression

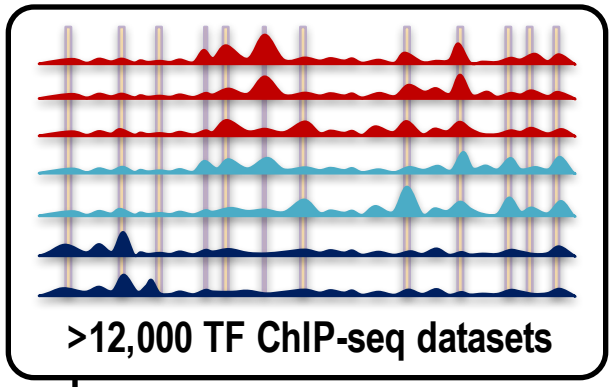
Mapping



Cis-regulatory profile

Cis-regulatory element repertoire (2.7 million in the human genome, 1.5 million in mouse)

differential interaction

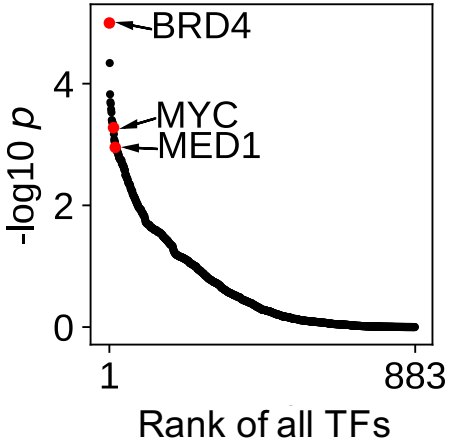


ROC associations

Statistical tests,
Background adjustment,
Rank integration

Wang *et al.*, *Bioinformatics* 2018

Output prediction



Ma *et al.*, under review 2020

Limitations of ChIP-seq

- Dependent on antibody availability and quality
- Semi-quantitative: does not detect global change
- Needs many cells – difficult for clinical samples
- Cellular heterogeneity

Take-home message

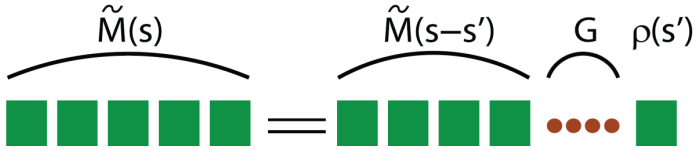
- Why am I learning these if I am not a bioinformatician?
  - Help improve experimental design
  - Quality control
  - Better interpret the experimental data
  - Take advantage of existing tools and data resources

Thank you very much!

zang@virginia.edu
zanglab.org

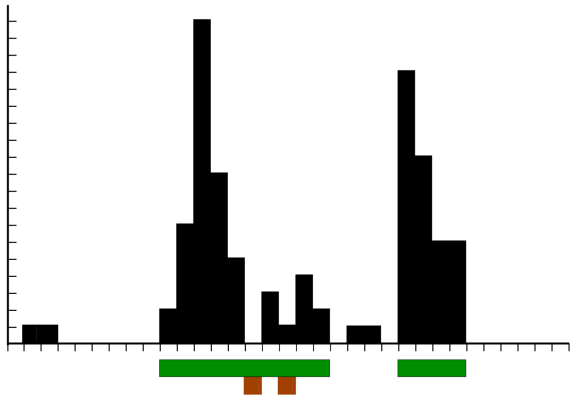
Call broad peaks: SICER

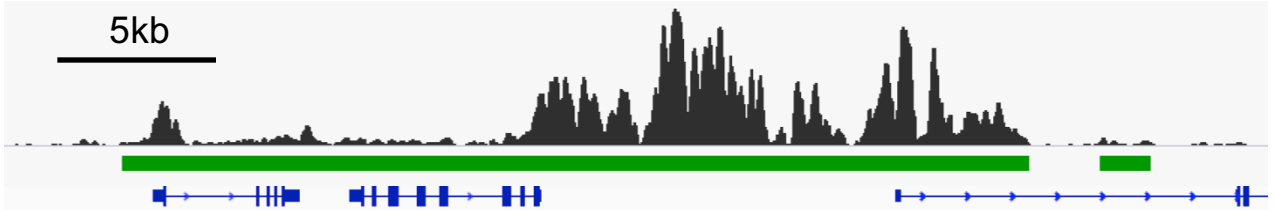
- **S**patial-clustering **I**dentification of **C**hIP-**E**nriched **R**egions

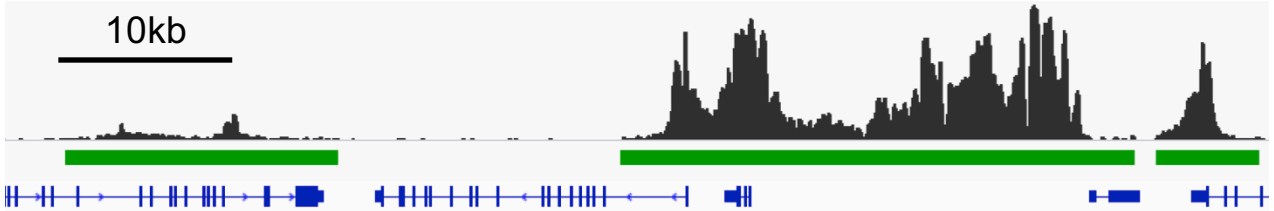


$$\tilde{M}(s) = G(\lambda, l_0, g) \int_{s_0}^{s} \mathrm{d}s' \tilde{M}(s-s')\rho(s')$$

$$M(s) = t^{g+1}\tilde{M}(s)t^{g+1}$$







Zang et al. *Bioinformatics* 2009

Try SICER2

- https://zanglab.github.io/SICER2/

SICER2 Documentation

zanglab.github.io/SICER2/

SICER2 Documentation

Search docs

Docs » Quick Start

Edit on GitHub

SICER2

Redesigned and improved ChIP-seq broad peak calling tool SICER

build passing

GitHub Repo

Introduction

Chromatin immunoprecipitation combined with high-throughput sequencing (ChIP-seq) can be used to map binding sites of a protein of interest in the genome. Histone modifications usually occupy broad chromatin domains and result in diffuse patterns in ChIP-seq data that make it difficult to identify signal enrichment. SICER, a spatial clustering approach for the identification of ChIP-enriched regions, was developed for calling broad peaks from ChIP-seq data.

Usability of the original SICER software has been affected by increased throughputs of ChIP-seq experiments over the years. We now present SICER2 a more user-friendly version of SICER that has been redisgned and streamlined to handle large ChIP-seq data sets. This new Python package supports multiple job submissions on cluster systems and parallel processing on multicore architectures.

Jin Yong (Jeffrey) Yoo 2020