

Analysis of ChIP-seq data

BIOC8145

Chongzhi Zang

zang@virginia.edu

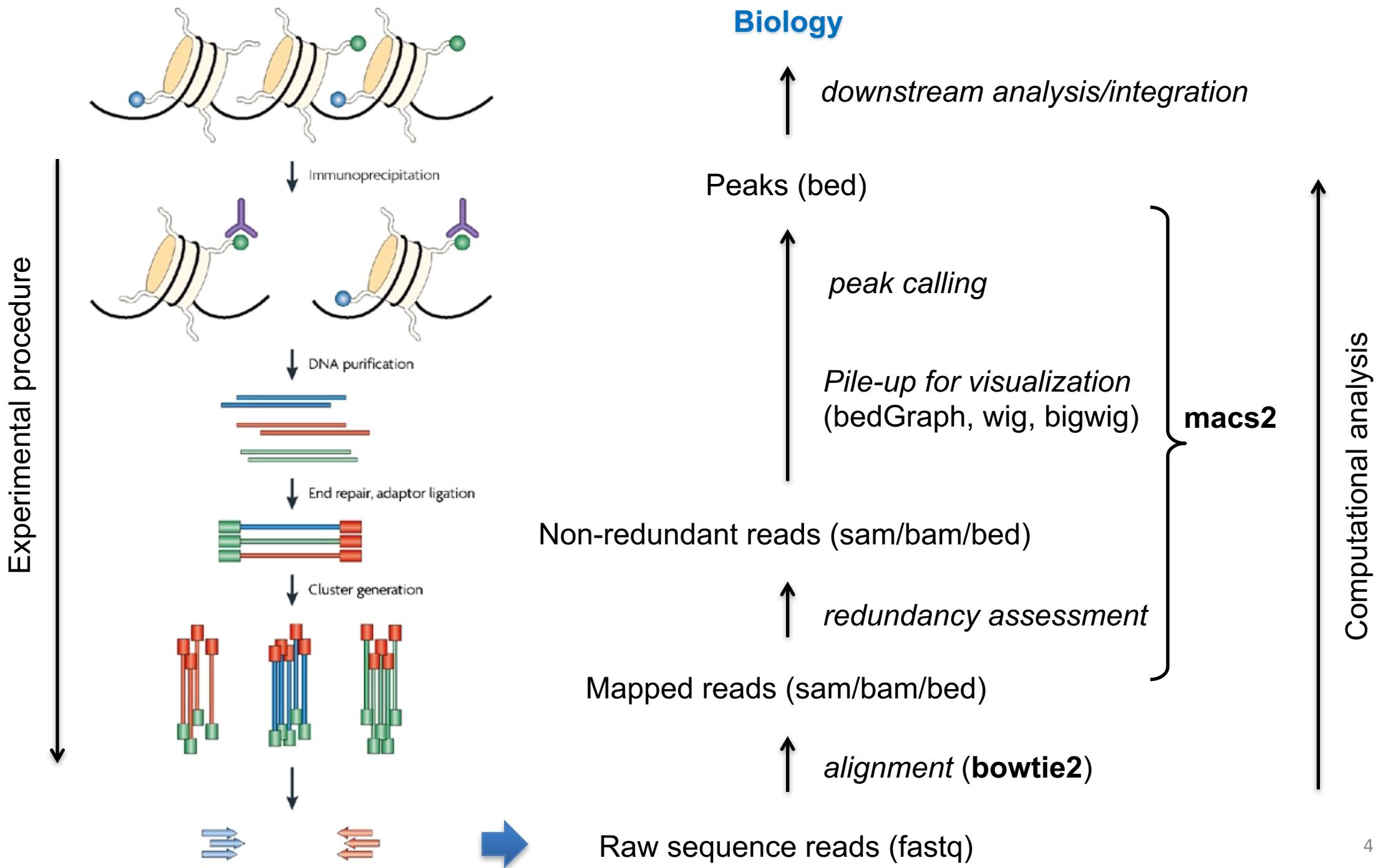
zanglab.org

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



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

TAGGTCTATT
TAGGTACTG
ATGGTAACT
CAGGTATAAC
TGTGTGAGT
AAGGTAAGT

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

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



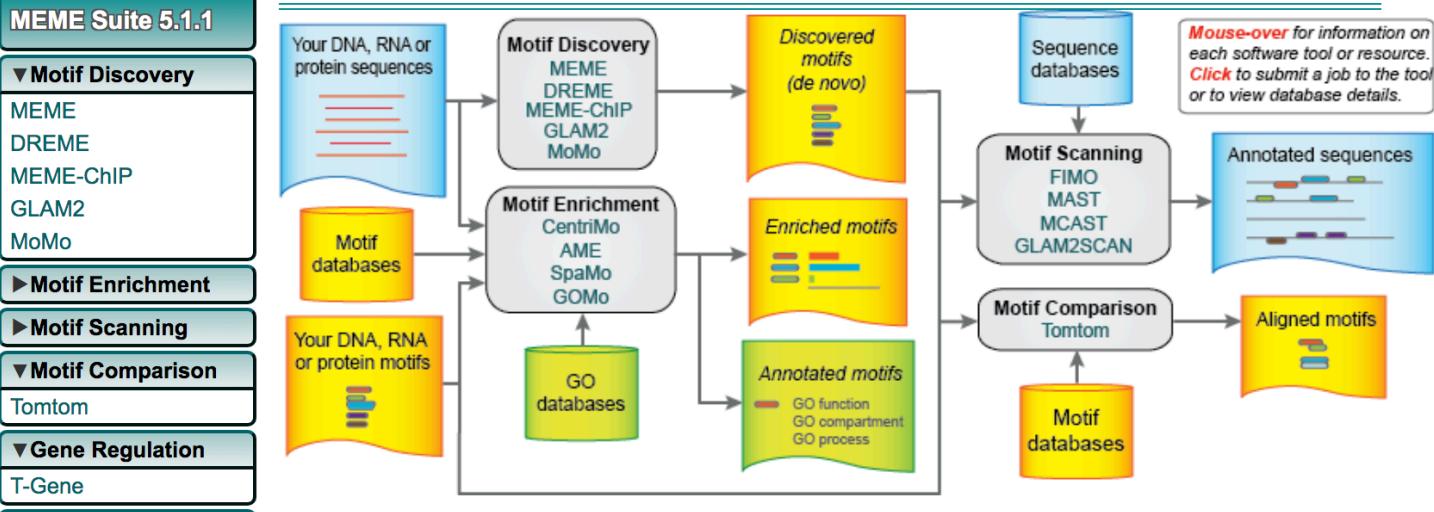
$$R_i = \log_2(4) - H_i$$

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

MEME (meme-suite.org)

The MEME Suite

Motif-based sequence analysis tools



MEME
Multiple Em for Motif Elicitation



CentriMo
Local Motif Enrichment Analysis



FIMO
Find Individual Motif Occurrences



DREME
Discriminative Regular Expression Motif Elicitation



AME
Analysis of Motif Enrichment



MAST
Motif Alignment & Search Tool



MEME-ChIP
Motif Analysis of Large Nucleotide Datasets



SpaMo
Spaced Motif Analysis Tool



MCast
Motif Cluster Alignment and Search Tool



GLAM2
Gapped Local Alignment of Motifs



GOMo
Gene Ontology for Motifs



GLAM2Scan
Scanning with Gapped Motifs



MoMo
Modification Motifs



Tomtom
Motif Comparison Tool



GT-Scan
Identifying Unique Genomic Targets



T-Gene
Predicting Target Genes



← Previous version 5.1.0

HOMER (homer.ucsd.edu)

← → C ⓘ 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 [unix tutorial](#) into google). I am proud to say that many 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.

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).

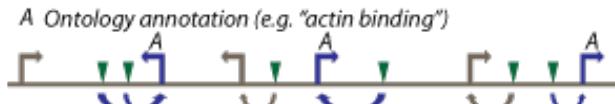
Example: SRF ChIP-Seq called peaks



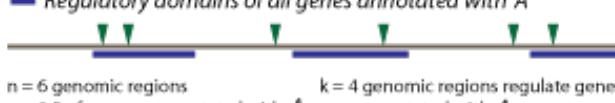
2. GREAT associates both proximal and distal input Genomic Regions with their putative target genes.



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



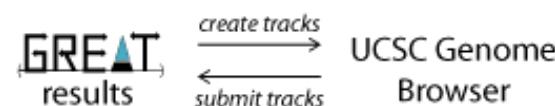
4. GREAT calculates statistical Enrichments for associations between Genomic Regions and Annotations.



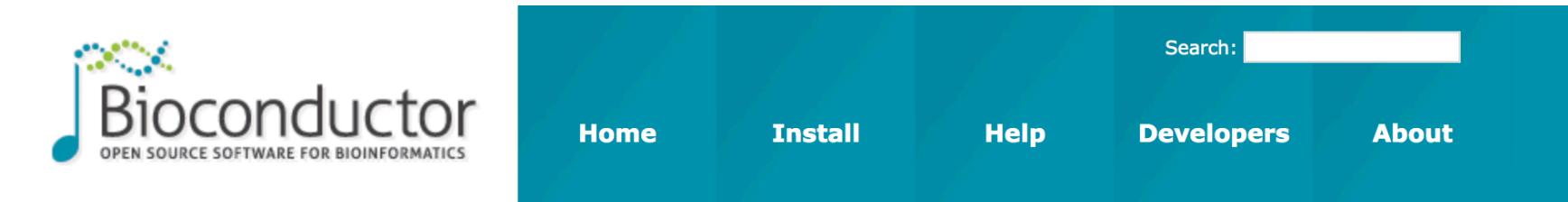
5. **Output:** Annotation terms that are significantly associated with the set of input Genomic Regions.

	Ontology term	p-value
SRF peaks regulate genes involved in:	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



The screenshot shows the ChIPseeker package page on the Bioconductor website. The header features the Bioconductor logo and navigation links for Home, Install, Help, Developers, and About. A search bar is also present. The main content area displays the package's details, including its version (3.10), platforms (all), rank (123 / 1823), posts (2 / 0 / 1 / 0), and Bioconductor history (6 years). It also shows build warnings, updated since release, and dependencies (152). Social sharing icons for Facebook and Twitter are included. Below this, a section titled "ChIPseeker for ChIP peak Annotation, Comparison, and Visualization" provides a brief description of the package's functionality. The Bioconductor version is listed as Release (3.10). The package implements functions for retrieving nearest genes, annotating genomic regions, estimating significance of overlap, and incorporating the GEO database for comparison. Several visualization functions are provided for summarizing peak coverage, average profiles, heatmaps, genomic annotation, distance to TSS, and peak overlap. Author information includes Guangchuang Yu [aut, cre] (ORCID iD), Yun Yan [ctb], Hervé Pagès [ctb], Michael Kluge [ctb], Thomas Schwarzl [ctb], and Zhoueng Xu [ctb]. The maintainer is Guangchuang Yu <guangchuangyu at gmail.com>. Citation information is provided for the paper: 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](https://doi.org/10.1093/bioinformatics/btv145).

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](https://doi.org/10.18129/B9.bioc.ChIPseeker) [f](#) [t](#)

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, statistical 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], Zhoueng 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](https://doi.org/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 resources

Galaxy: web-based analysis platform

- <https://usegalaxy.org/>

The screenshot shows the Galaxy homepage at https://usegalaxy.org/. The page features a dark header with the Galaxy logo, navigation links for Analyze Data, Workflow, Visualize, Shared Data, Help, Login or Register, and a search bar. A progress bar indicates "Using 0%". On the left, a sidebar lists various tool categories: Tools, Get Data, Collection Operations, Expression Tools, GENERAL TEXT TOOLS, Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash, GENOMIC FILE MANIPULATION, FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED, VCF/BCF, Nanopore, Convert Formats, Lift-Over, COMMON GENOMICS TOOLS, Operate on Genomic Intervals, Fetch Sequences/Alignments, GENOMICS ANALYSIS, Assembly, Annotation, Mapping, and Variant Calling. A central banner for "Galaxy Help" encourages users to "Got Questions? Get Answers." with a link to help.galaxyproject.org. Below the banner, logos for Penn State, Johns Hopkins University, and Oregon Health & Science University are displayed. A section about the Galaxy Team's affiliation with these institutions follows. To the right, a "Tweets" feed from @galaxyproject shows a tweet about a Single-Cell RNAseq Training course at Earlham Institute. A "History" panel on the far right shows an "Unnamed history" entry with a message about loading data. At the bottom, a footer notes Galaxy's support by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University, and provides a disclaimer about data storage and encryption.

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Galaxy Help
Got Questions?
Get Answers.
help.galaxyproject.org

PennState JOHNS HOPKINS OREGON HEALTH & SCIENCE UNIVERSITY

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, the Department of Biology at Johns Hopkins University and the Computational Biology Program at Oregon Health & Science University.

This instance of Galaxy is utilizing infrastructure generously provided by CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

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 Johns Hopkins University.

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.

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

- <http://cistrome.org/ap/>

The screenshot shows the Cistrome Galaxy-based platform interface. The top navigation bar includes tabs for Analyze Data, Workflow, Shared Data, Lab, Visualization, Help, and User, along with a search bar and user information. The main content area is titled "Upload File (version 1.1.4)". It contains sections for "File Format" (set to "Auto-detect"), "File (Please avoid Windows format text file)" (with a "Choose File" button and "No file chosen" message), "URL/Text" (a large text input field with placeholder text "Here you may specify a list of URLs (one per line) or paste the contents of a file."), and "Files uploaded via ASPERA" (a table with columns "File", "Size", and "Date", showing the message "Your ASPERA upload directory contains no files."). On the left, the "CISTROME TOOLBOX" sidebar lists various import and data processing tools. On the right, the "History" panel shows a list of recent analyses, each with a preview icon, edit icon, and delete icon.

Galaxy / Cistrome / Cistrome x 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 classified regions

66: Heatmapr R script

65: Heatmap image

64: Heatmap log

63: Heatmap k-means classified regions

62: Heatmapr R script

61: Heatmap image

60: Heatmap log

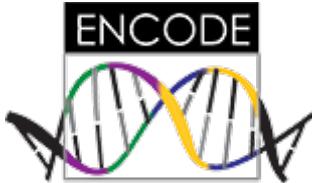
59: Heatmap k-means classified regions

58: Heatmapr R script

57: Heatmap image

56: Heatmap log

55: Heatmap k-means classified regions



ENCODE

<https://www.encodeproject.org/>

Matrix – ENCODE

encodeproject.org/matrix/?type=Experiment&status=released

Showing 16414 results

Assay type

DNA binding	9017
Transcription	4547
DNA accessibility	1109
RNA binding	699
DNA methylation	560

Assay title

Search

TF ChIP-seq	3608
Histone ChIP-seq	3180
Control ChIP-seq	2229
scRNA-seq	1078
DNase-seq	836
polyA plus RNA-seq	770
total RNA-seq	704

Cistrome Data Browser

<http://cistrome.org/db/>

The screenshot shows the Cistrome Data Browser homepage. At the top, there is a navigation bar with links to "Cistrome Data Browser", "Home", "Documentation", "About", "Statistics", "Batch download", "ToolKit", "Cistrome-GO", and "Liu Lab". Below the navigation bar is a large blue header with the "Cistrome Data Browser" logo and text. A "Tips" section contains the following bullet points:

- 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!

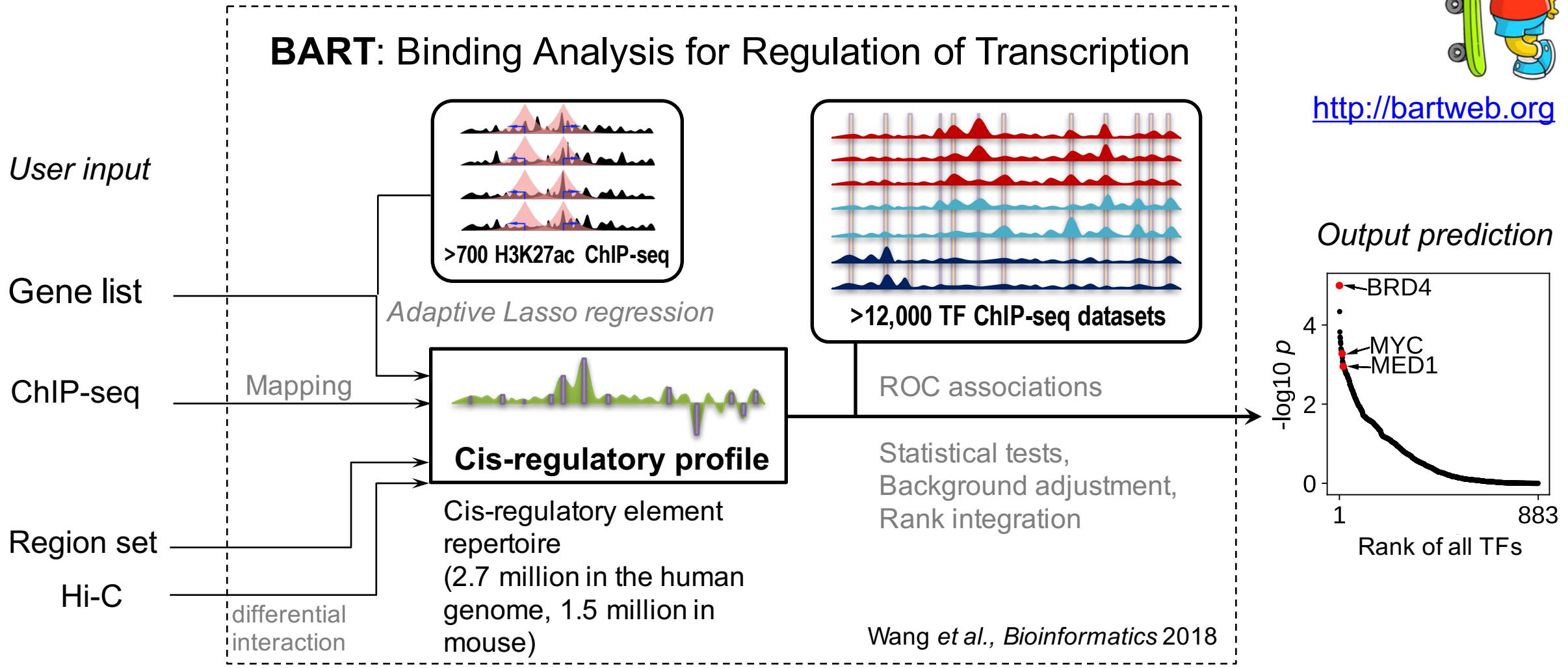
Below the tips section are three filter panels: "Containing word(s)" (with a search input and "Search" button), "Species" (with "All" selected, "Homo sapiens", and "Mus musculus" options), "Biological Sources" (with "All" selected, "1-cell pronuclei", "1015c", "10326", "1064Sk", and "106A" options), and "Factors" (with "All" selected, "AATF", "ABCC9", "ACSS2", "ACTB", and "ADNP" options). At the bottom, there is a "Results" table with columns: Batch, Species, Biological Source, Factor, Publication, and Quality Control. One row is shown: "Batch" (checkbox), "Species" (Homo sapiens), "Biological Source" (HeLa; Epithelium; Cervix), "Factor" (BTAF1), "Publication" (Johannes F, et al. Bioinformatics 2010), and "Quality Control" (red and green dots). To the right of the table, there are two citations: "Mei et al. Nucleic Acids Res. 2017" and "Zheng et al. Nucleic Acids Res. 2018".

Results					
Batch	Species	Biological Source	Factor	Publication	Quality Control
<input type="checkbox"/>	Homo sapiens	HeLa; Epithelium; Cervix	BTAF1	Johannes F, et al. Bioinformatics 2010	

BART: TF prediction using public ChIP-seq data



<http://bartweb.org>

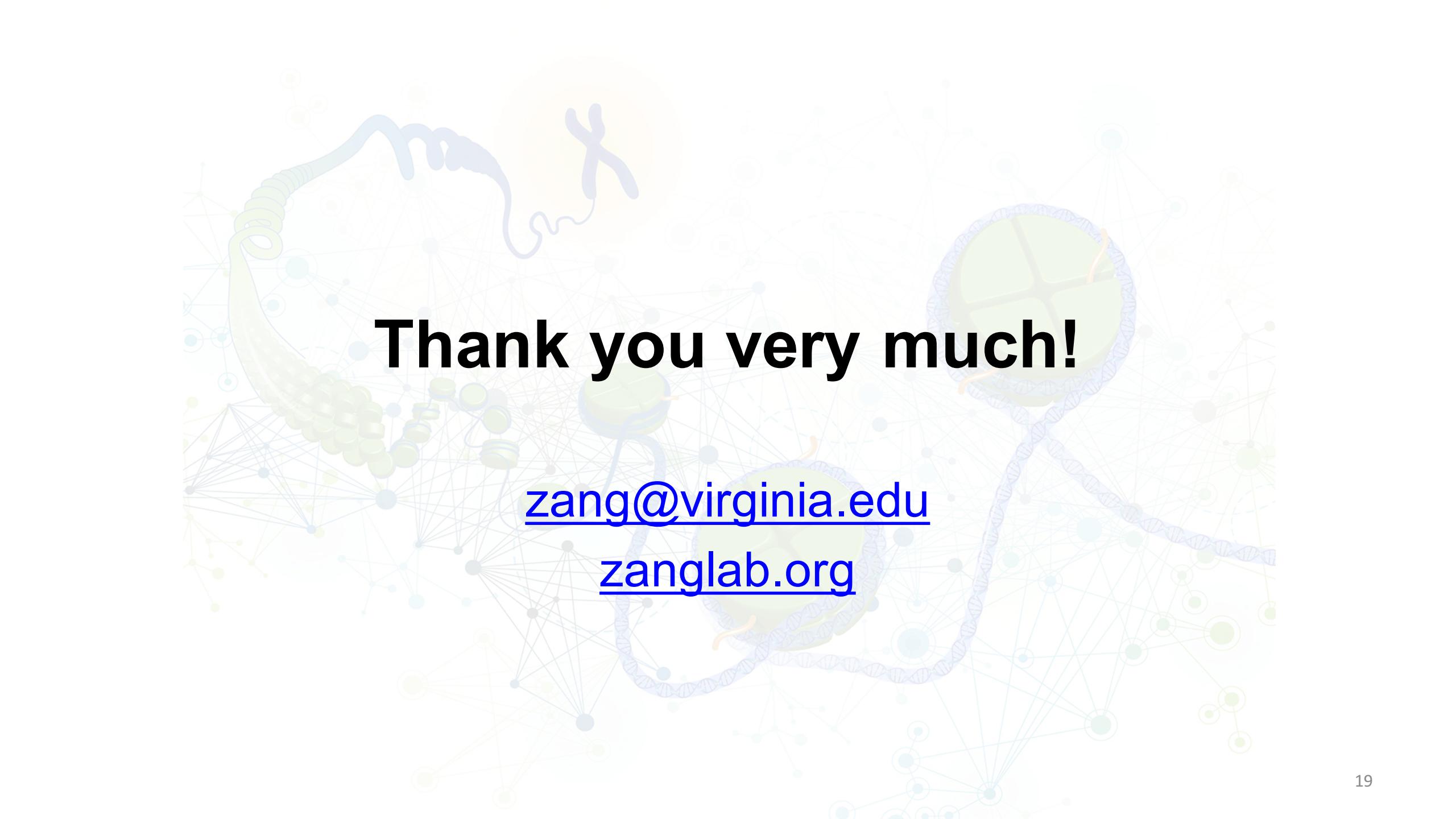


Limitations of ChIP-seq

- 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



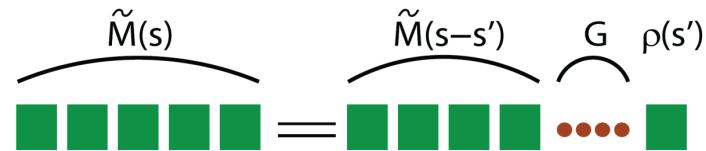
Thank you very much!

zang@virginia.edu

zanglab.org

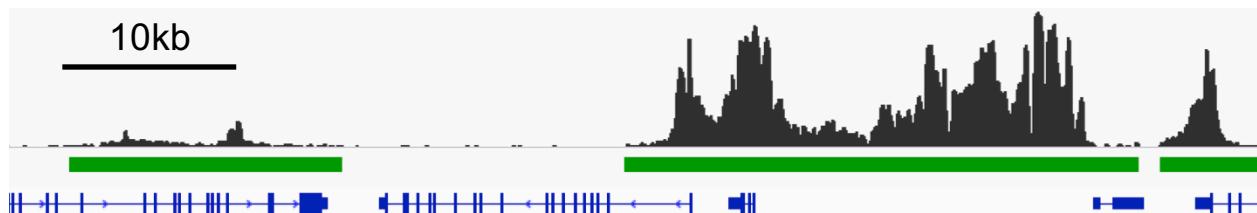
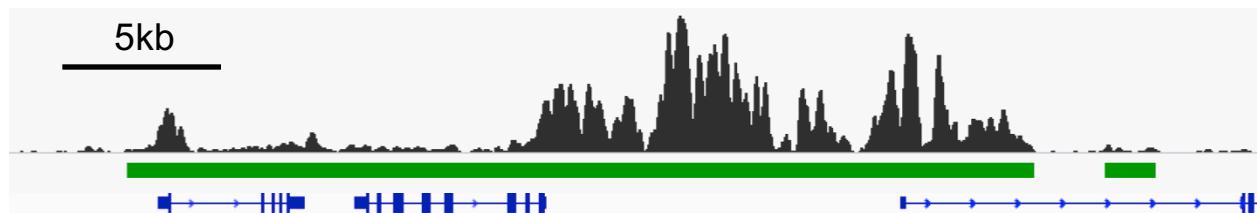
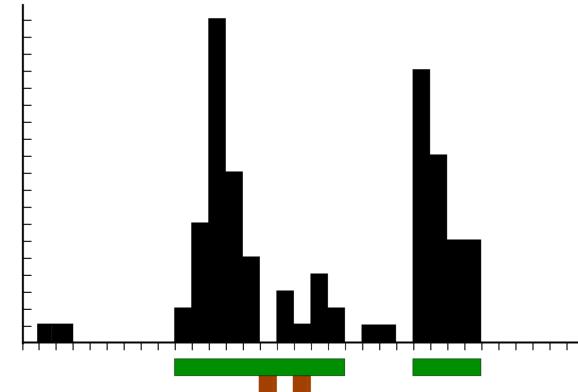
Call broad peaks: SICER

- Spatial-clustering Identification of ChIP-Enriched Regions

$$\tilde{M}(s) = \tilde{M}(s-s') \cdot G \cdot \rho(s')$$


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

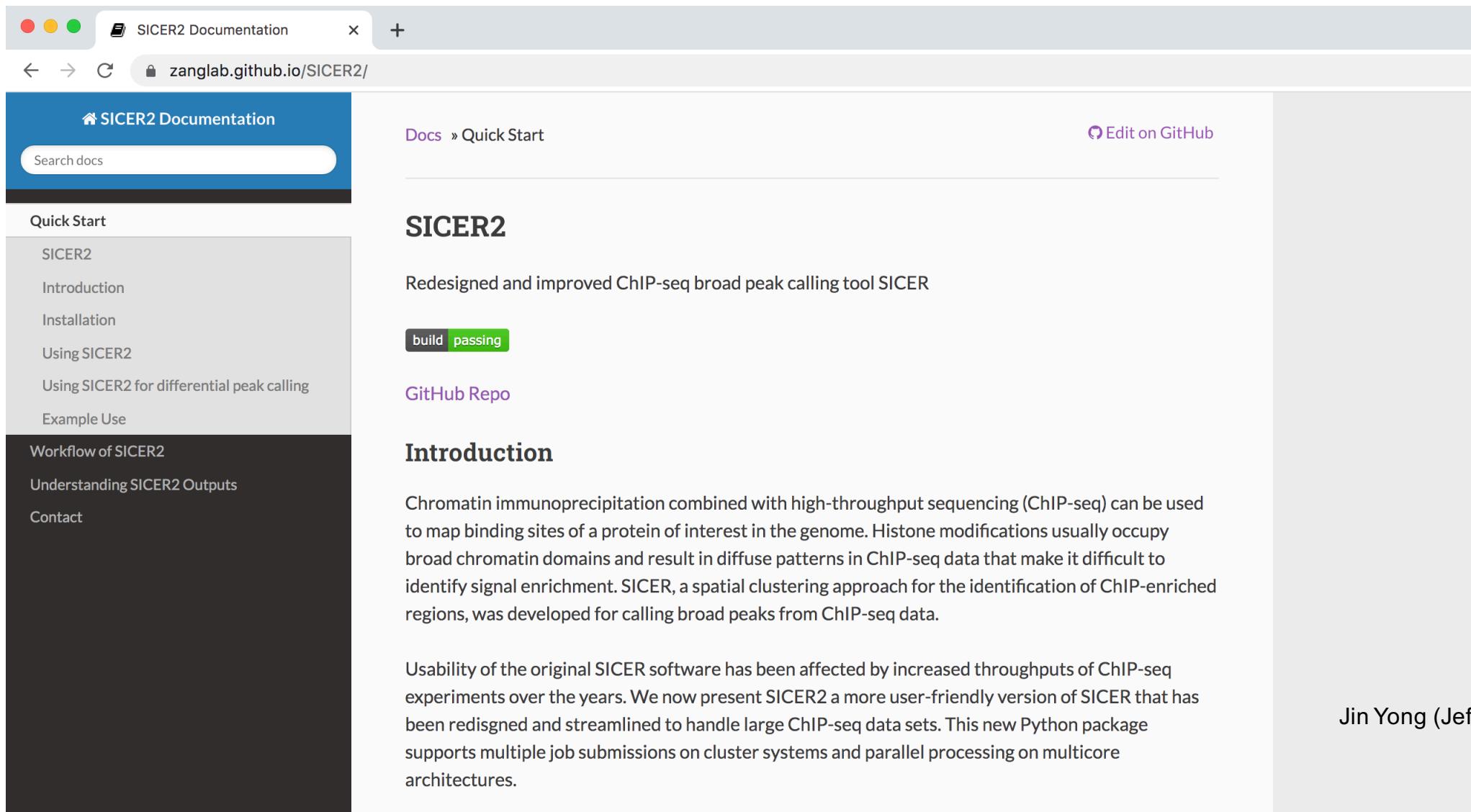
$$M(s) = t^{g+1} \tilde{M}(s) t^{g+1}$$



Zang et al. Bioinformatics 2009

Try SICER2

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



The screenshot shows a web browser window displaying the SICER2 Documentation at zanglab.github.io/SICER2/. The page has a blue header bar with the title "SICER2 Documentation". Below the header is a search bar labeled "Search docs". The main content area features a large heading "SICER2" and a sub-headline "Redesigned and improved ChIP-seq broad peak calling tool SICER". A green button labeled "build passing" indicates a successful build status. There is also a link to the "GitHub Repo". On the left side, there is a sidebar with a dark background containing a navigation menu. The menu items under "Quick Start" are: SICER2, Introduction, Installation, Using SICER2, Using SICER2 for differential peak calling, and Example Use. Under "Workflow of SICER2", the items are: Understanding SICER2 Outputs and Contact.

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 redesigned 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.