

# Analysis of ChIP-seq data

## BIOC8145

Chongzhi Zang

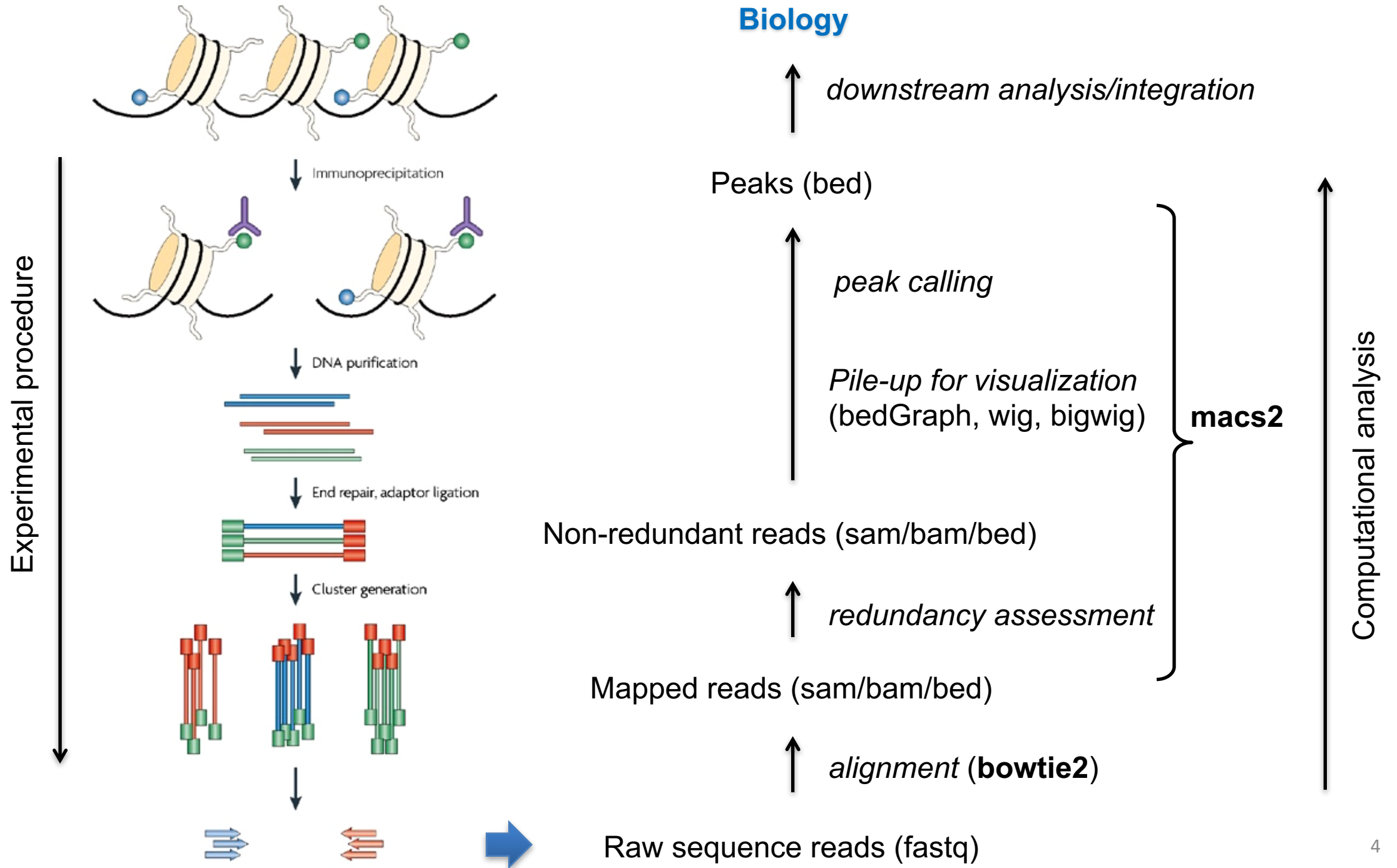
[zang@virginia.edu](mailto:zang@virginia.edu)  
[zanglab.org](http://zanglab.org)

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

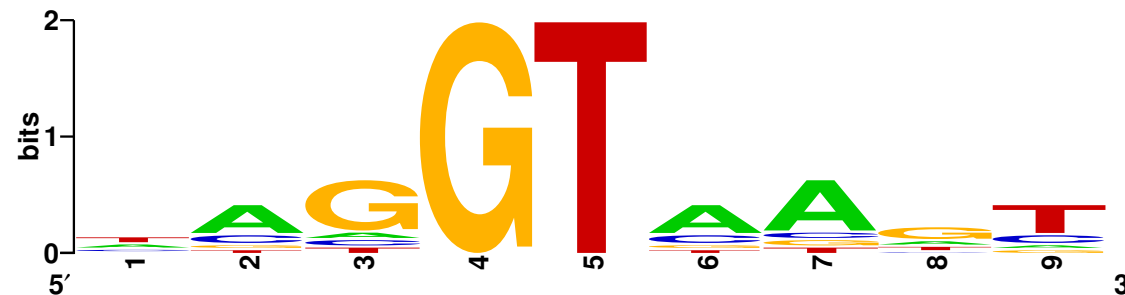
# Position weight matrix (PWM) representation of DNA sequence motifs

GAGGTAAAC  
TCCGTAAGT  
CAGGTTGGA  
ACAGTCAGT  
TAGGTCATT  
TAGGTACTG  
ATGGTAACT  
CAGGTATAC  
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 Suite 5.1.1

#### ▼ Motif Discovery

MEME  
DREME  
MEME-ChIP  
GLAM2  
MoMo

#### ► Motif Enrichment

#### ► Motif Scanning

#### ▼ Motif Comparison

Tomtom

#### ▼ Gene Regulation

T-Gene

#### ► Manual

#### ► Guides & Tutorials

#### ► Sample Outputs

#### ► File Format Reference

#### ► Databases

#### ► Download & Install

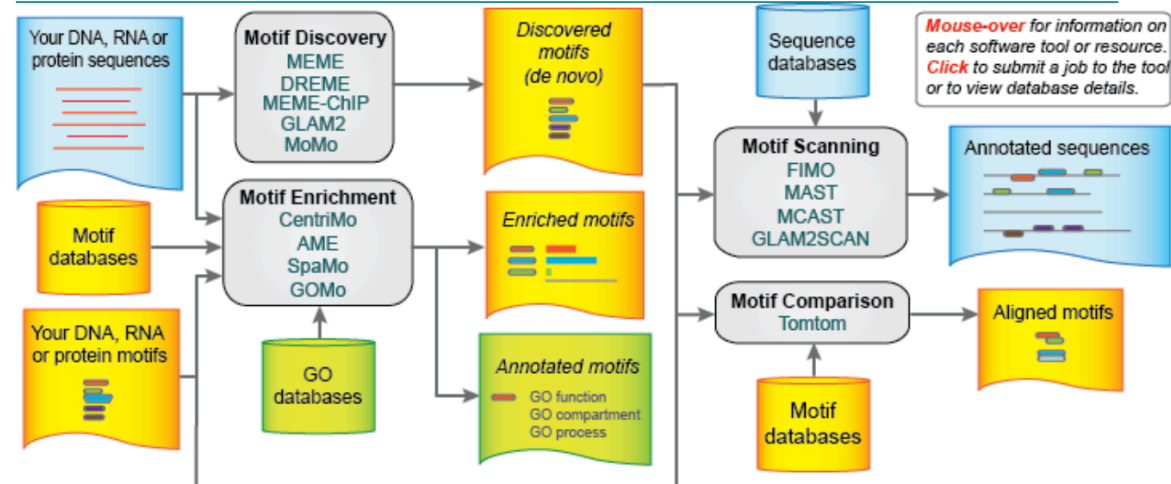
#### ► Help

#### ► Alternate Servers

#### ► Authors & Citing

#### ► Recent Jobs

↶ Previous version 5.1.0



**MEME**  
Multiple Em for Motif Elicitation

Improved

**CentriMo**  
Local Motif Enrichment Analysis

**FIMO**  
Find Individual Motif Occurrences

**DREME**  
Discriminative Regular Expression Motif Elicitation

**AME**  
Analysis of Motif Enrichment

Improved

**MAST**  
Motif Alignment & Search Tool

**MEME-ChIP**  
Motif Analysis of Large Nucleotide Datasets

Improved

**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

Improved

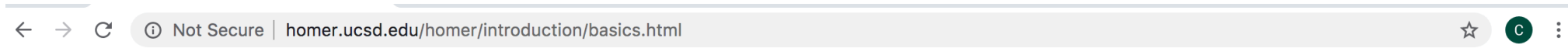
**Tomtom**  
Motif Comparison Tool

**GT-Scan**  
Identifying Unique Genomic Targets

**T-Gene**  
Predicting Target Genes

New

# HOMER ( [homer.ucsd.edu](http://homer.ucsd.edu) )



## 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.

### *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).

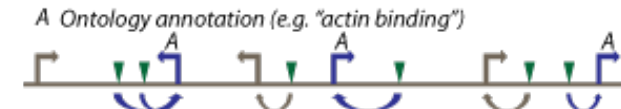
Example:  $\nabla$  SRF ChIP-Seq called peaks



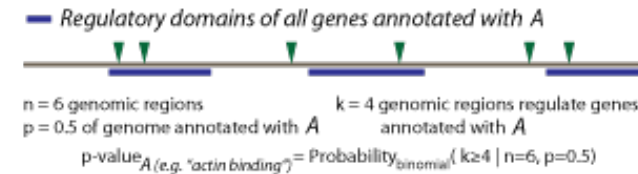
2. GREAT associates both proximal and distal input Genomic Regions with their putative target genes.



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



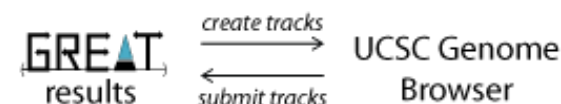
4. GREAT calculates statistical Enrichments for associations between Genomic Regions and Annotations.




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



Search:

HomeInstallHelpDevelopersAbout

Home » Bioconductor 3.10 » Software Packages » ChIPseeker

## ChIPseeker

platformsall

rank123 / 1823



posts2 / 0 / 1 / 0

in Bioc6 years

buildwarnings

updatedsince release


dependencies152

DOI: [10.18129/B9.bioc.ChIPseeker](https://doi.org/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, 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], 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](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

10

# **ChIP-seq: online resources**

# Galaxy: web-based analysis platform

- <https://usegalaxy.org/>

The screenshot shows the Galaxy web-based analysis platform homepage. The browser address bar displays <https://usegalaxy.org/>. The page features a dark blue header with navigation links: Analyze Data, Workflow, Visualize, Shared Data, Help, Login or Register, and a grid icon. A 'Using 0%' status bar is on the right. The left sidebar contains a 'Tools' section with a search bar and a list of tool categories: 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. The main content area includes a welcome message: '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.' Below this is a 'Galaxy Help' banner with the text 'Got Questions? Get Answers.' and the URL 'help.galaxyproject.org'. To the right is a 'Tweets' section by @galaxyproject, featuring a tweet about 'Single-Cell RNAseq Training 2020' from the Eartham Institute. The right sidebar shows a 'History' section with a search bar and a message: 'This history is empty. You can load your own data or get data from an external source'. The footer contains logos for PennState, Johns Hopkins University, Oregon Health & Science University, TACC, and CyVerse, along with text describing the Galaxy Team and the infrastructure provided by CyVerse and the National Science Foundation.

Galaxy

Analyze Data Workflow Visualize Shared Data Help Login or Register

Using 0%

Tools

search 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

Variant Calling

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

Tweets by @galaxyproject

Single-Cell RNAseq Training 2020

Date: 20 - 24 April 2020  
Register by: 02 March 2020  
Cost: £100 - full course  
£200 - bioinformatics days only (22-24 April)  
Venue: Earham Institute, Norwich, UK

Eartham Institute

Single-Cell RNAseq Training Course

The course will provide an introduction to Single Cell Genomics covering experimental design, cell sorting and processing, quality of sequence data, data

Embed View on Twitter

History

search datasets

Unnamed history

(empty)

This history is empty. You can load your own data or get data from an external source

PennState

JOHNS HOPKINS UNIVERSITY

OREGON HEALTH & SCIENCE UNIVERSITY

TACC

CYVERSE

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 displays the Cistrome Galaxy web interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Lab', 'Visualization', 'Help', and 'User'. The main content area is titled 'Upload File (version 1.1.4)'. It features a 'File Format' dropdown set to 'Auto-detect', a 'File (Please avoid Windows format text file)' section with a 'Choose File' button, and a 'URL/Text' section with a text area. Below these is a table for 'Files uploaded via ASPERA' which is currently empty. The 'Convert spaces to tabs' section has a 'Yes' checkbox. The 'Genome' dropdown is set to 'Human Dec. 2013 (GRCh38/hg38) (hg38)'. An 'Execute' button is at the bottom. The left sidebar contains a 'Tools' section with a search bar and a 'CISTROME TOOLBOX' with links for 'Import Data', 'Data Preprocessing', 'Gene Expression', 'Integrative Analysis', 'Liftover/Others', and 'GALAXY TOOLBOX'. The right sidebar shows a 'History' section with a list of 11 items, each with a name, size, and icons for viewing, editing, and deleting.

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](http://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 clustered regions

66: Heatmap R script

65: Heatmap image

64: Heatmap log

63: Heatmap k-means clustered regions

62: Heatmap R script

61: Heatmap image

60: Heatmap log

59: Heatmap k-means clustered regions

58: Heatmap R script

57: Heatmap image

56: Heatmap log

55: Heatmap k-means clustered regions



# ENCODE

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

Matrix - ENCODE

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

🔍 ☆ 🌐 C ⋮

Showing 16414 results

Download

Visualize

{:}

Assay type

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

Assay title

Q

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/>


Cistrome DB

Not Secure | cistrome.org/db/#/


☆

C

Cistrome Data BrowserHomeDocumentationAboutStatisticsBatch downloadToolKitCistrome-GOLiu 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

All

Homo sapiens

Mus musculus

### Biological Sources

All

1-cell pronuclei

1015c

10326

1064Sk

106A

<<

### Factors

All

AATF

ABCC9

ACSS2

ACTB

ADNP

### Results

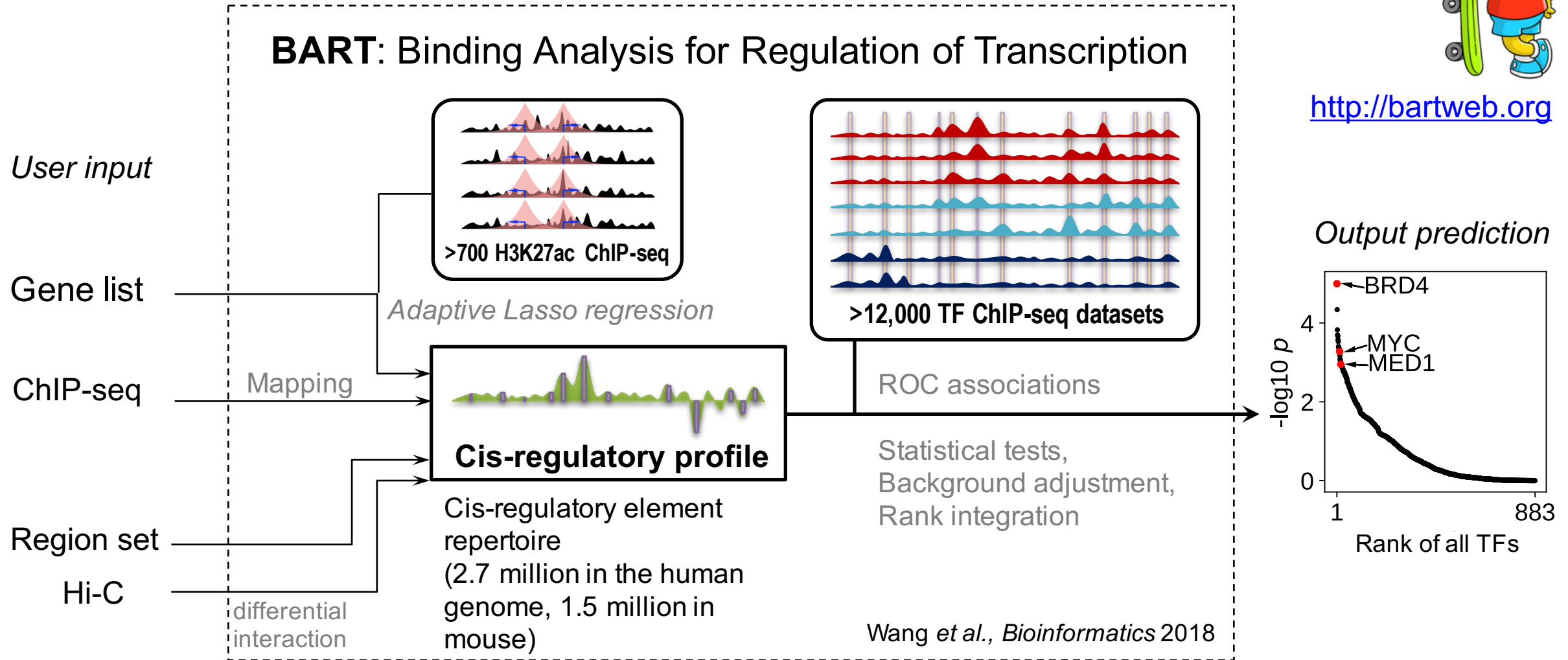
Batch	Species	Biological Source	Factor	Publication	Quality Control
<input type="checkbox"/>	Homo sapiens	HeLa; Epithelium; Cervix	BTAF1	Johannes F, et al. Bioinformatics 2010	<div><div></div><div></div><div></div><div></div><div></div><div></div></div>

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>



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



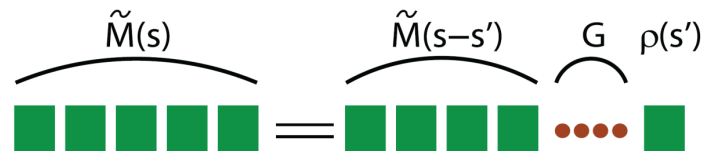
**Thank you very much!**

[zang@virginia.edu](mailto:zang@virginia.edu)

[zanglab.org](http://zanglab.org)

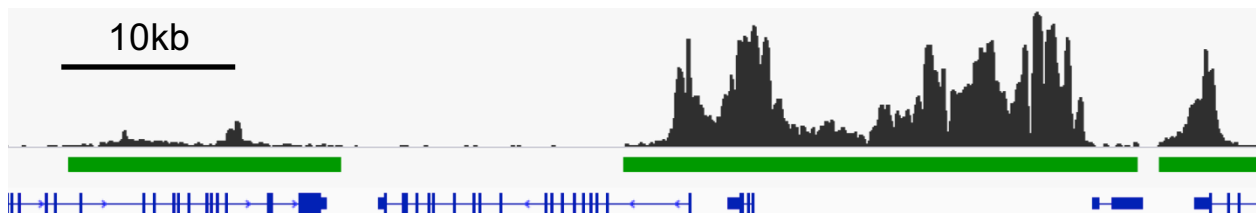
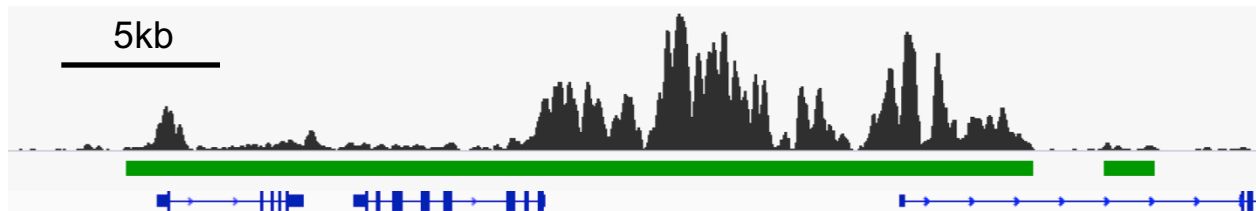
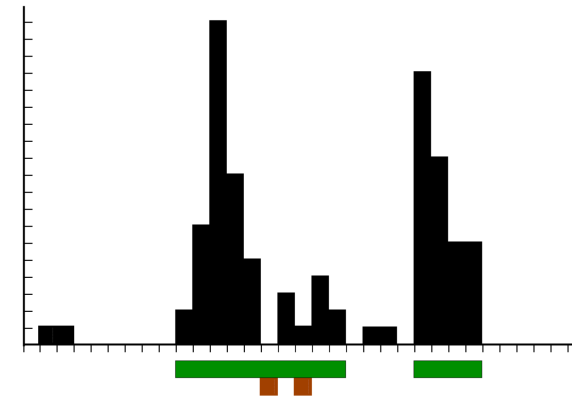
# Call broad peaks: SICER

- **S**patial-clustering **I**dentification of **ChIP-Enriched Regions**



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

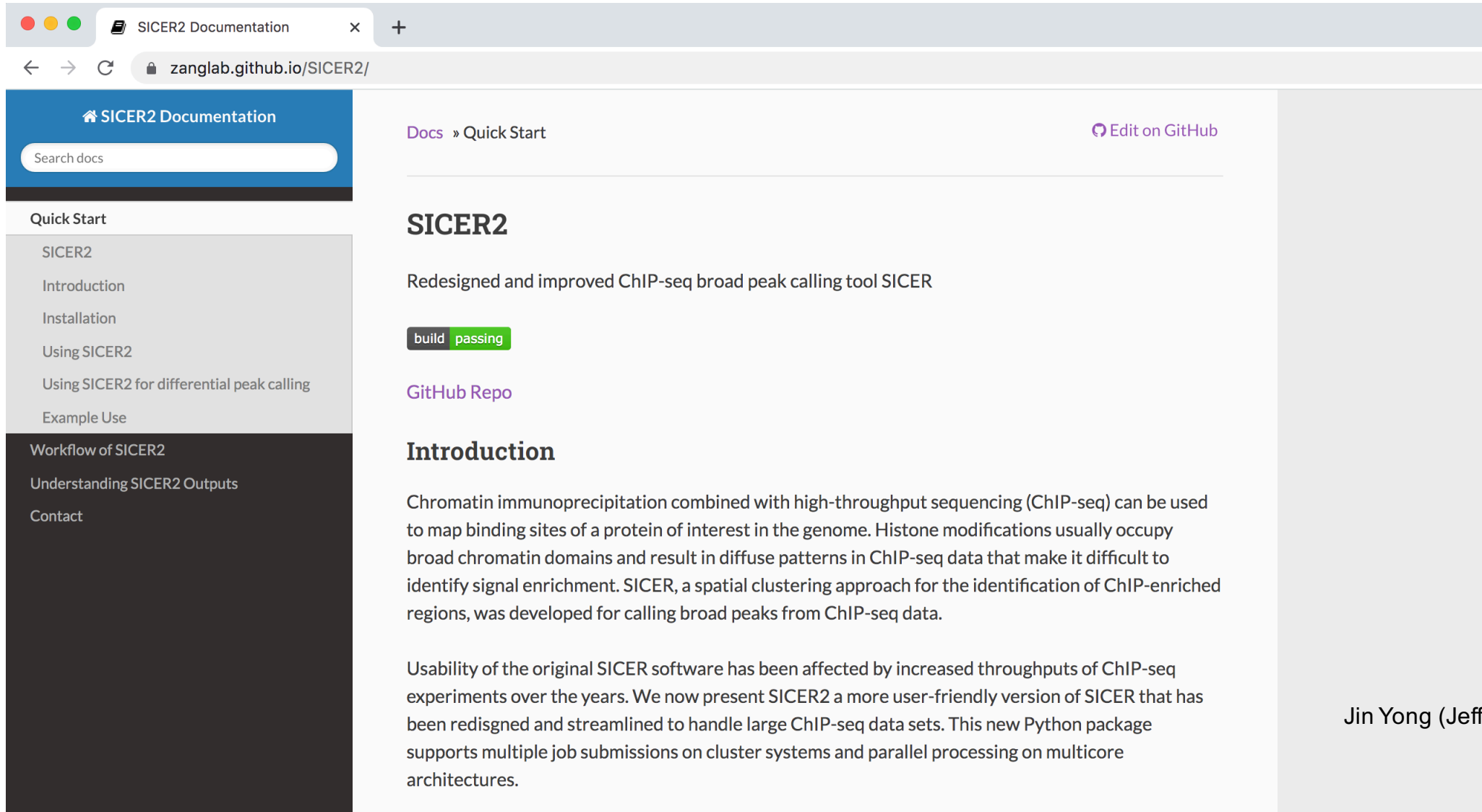
$$M(s) = t^{g+1} \tilde{M}(s) t^{g+1}$$



Zang et al. *Bioinformatics* 2009

# Try SICER2

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



The screenshot shows a web browser window with the address bar displaying `zanglab.github.io/SICER2/`. The page has a blue header with the text "SICER2 Documentation" and a search bar labeled "Search docs". A left sidebar contains a navigation menu with the following items: "Quick Start", "SICER2", "Introduction", "Installation", "Using SICER2", "Using SICER2 for differential peak calling", "Example Use", "Workflow of SICER2", "Understanding SICER2 Outputs", and "Contact". The main content area has a breadcrumb trail "Docs » Quick Start" and a link "Edit on GitHub". The title "SICER2" is prominently displayed, followed by the subtitle "Redesigned and improved ChIP-seq broad peak calling tool SICER". Below this is a status bar showing "build" in a grey box and "passing" in a green box. The section "Introduction" follows, containing two paragraphs of text. The first paragraph describes the tool's purpose in mapping binding sites. The second paragraph discusses the evolution from the original SICER software to the newer SICER2 version, highlighting its improved performance and user-friendliness.

SICER2 Documentation

Search docs

Quick Start

SICER2

Introduction

Installation

Using SICER2

Using SICER2 for differential peak calling

Example Use

Workflow of SICER2

Understanding SICER2 Outputs

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.