HARVARD
SCHOOL OF PUBLIC HEALTH

DANA-FARBER
CANCER INSTITUTE

ChIP-Seq Data Analysis and Integration with Cistrome

Chongzhi Zang, PhD
Dana-Farber Cancer Institute
Harvard School of Public Health

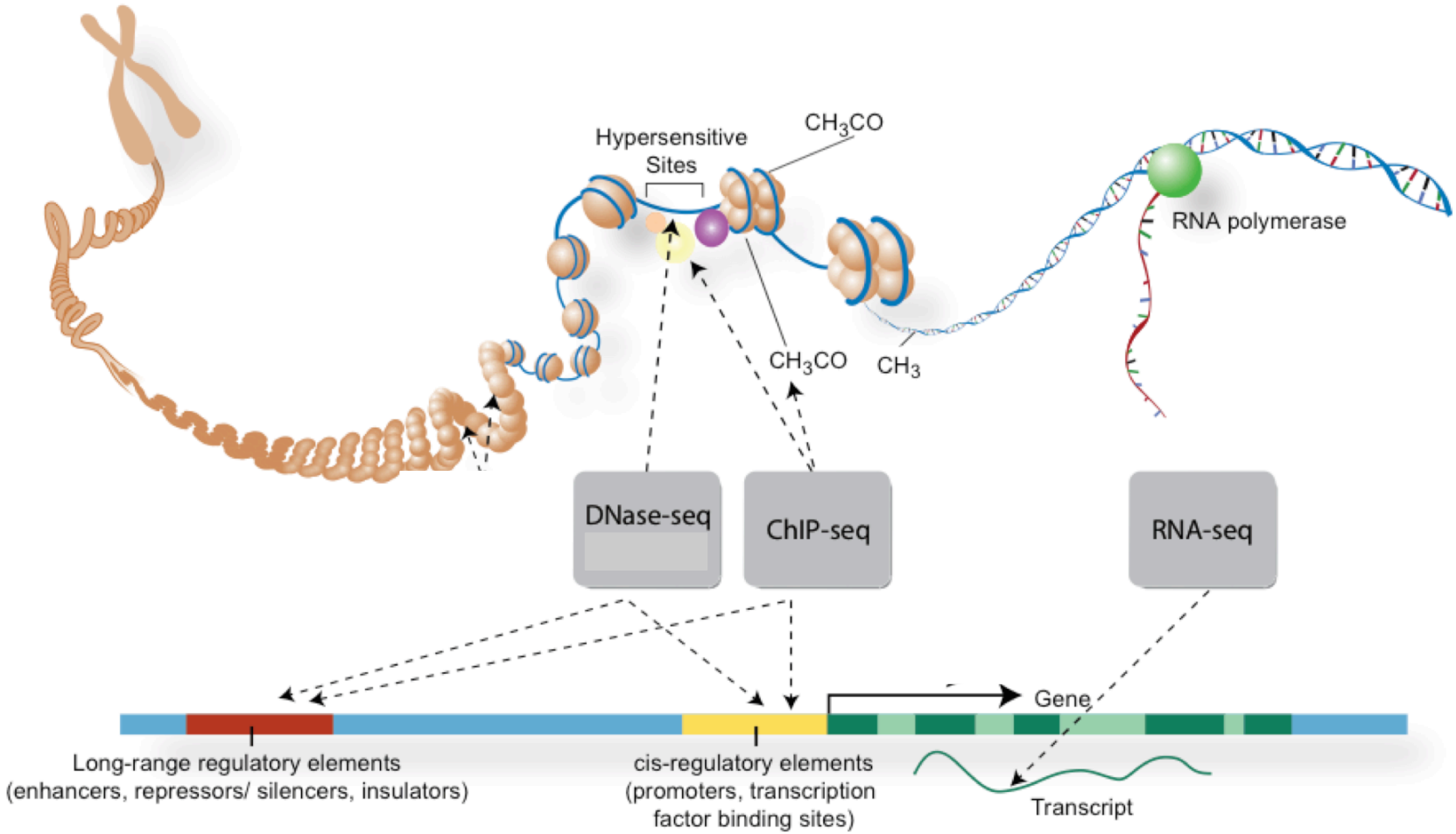
ChIP-Seq Data Analysis Workshop at NCI, NIH
November 19, 2014

Outline

- ChIP-Seq overview
- Cistrome Analysis Pipeline
  - Peak calling: MACS
  - ChIP-seq integrative analysis
  - BETA: Binding Expression Target Analysis
- Cistrome Dataset Browser

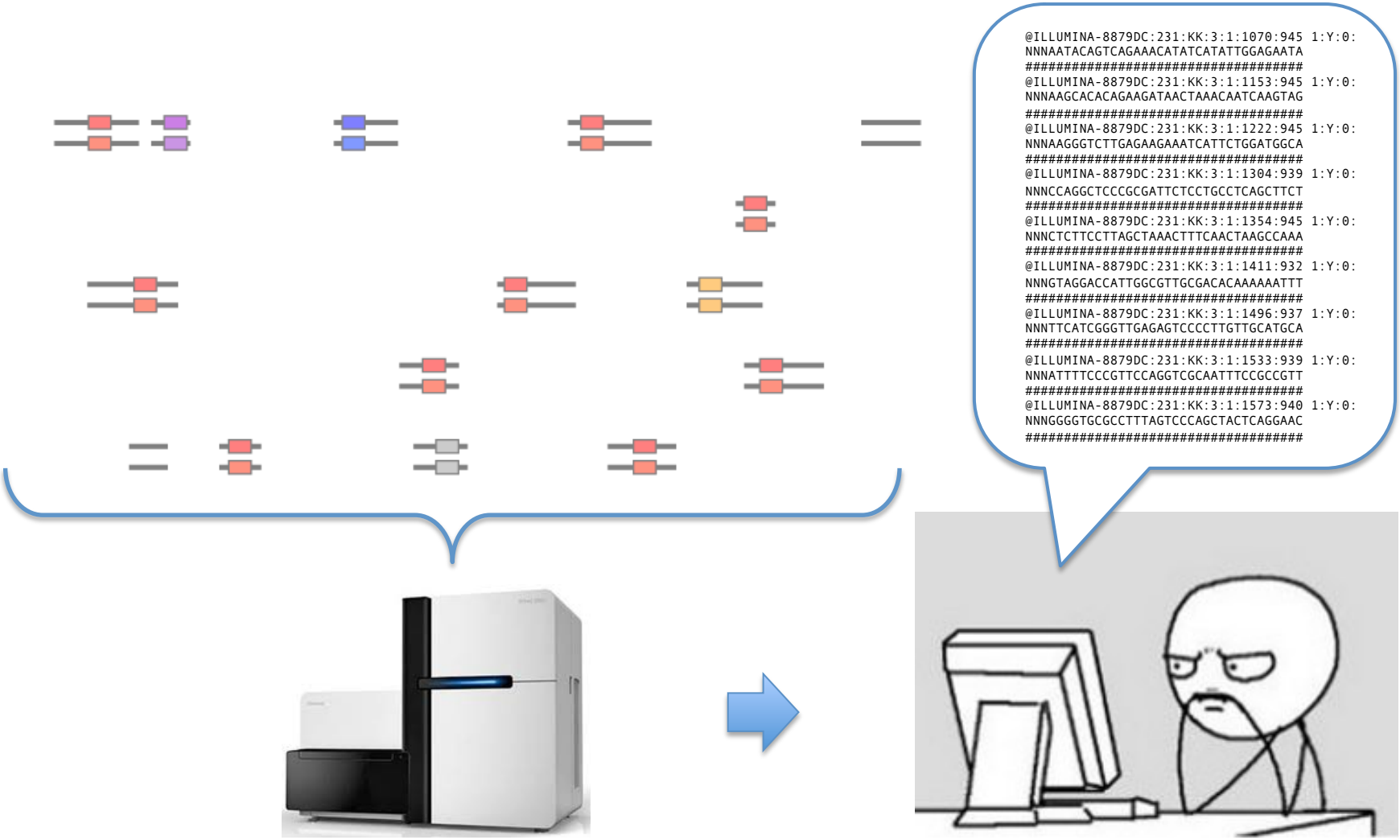
- Hands-on example with Cistrome Analysis

ChIP-Seq is used to study Cistrome, the in vivo genome-wide location of a transcription factor or a histone modification.



ENCODE Consortium

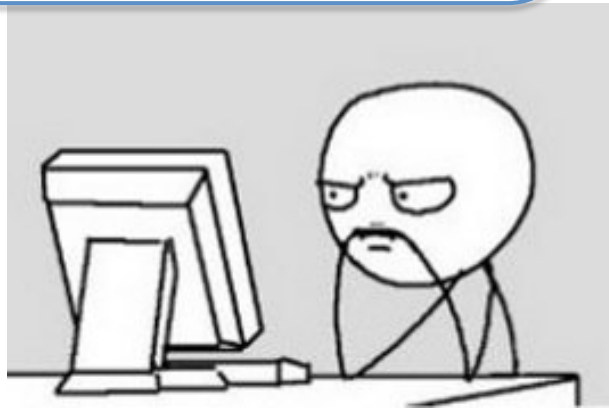
ChIP-Seq overview



ChIP-Seq data analysis overview

- Where in the genome do these sequence reads come from? - Sequence alignment and quality control
- What does the enrichment of sequences mean? - Peak calling: **MACS** or **SICER**
- What can we learn from these data? – Downstream analysis and integration

Cistrome can help!

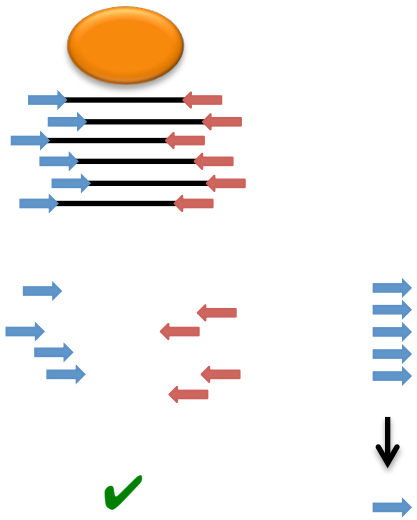


ChIP-Seq data analysis overview: basic processing

- alignment of each sequence read: **bowtie** or **BWA**

  - cannot map to the reference genome ✗
  - can map to multiple loci in the genome ✗
  - can map to a unique location in the genome ✔

- redundancy control: both **MACS** and **SICER** can do.

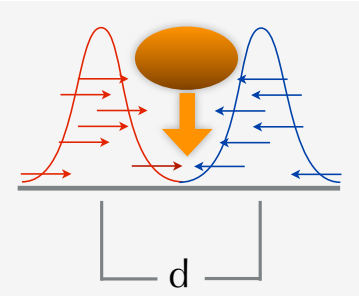


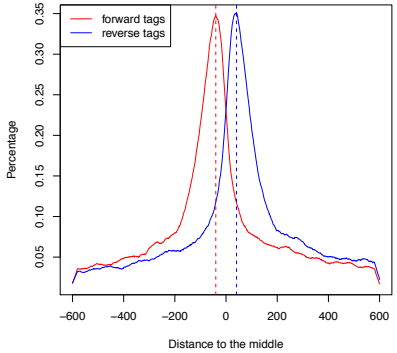
Langmead et al. 2009,
Zang et al. 2009

ChIP-Seq data analysis overview: peak calling

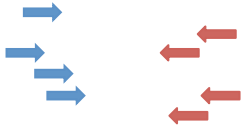
- DNA fragment size estimation

peak model

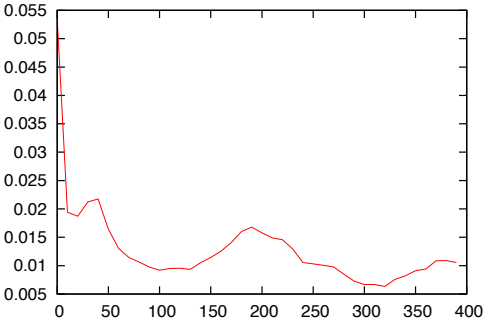




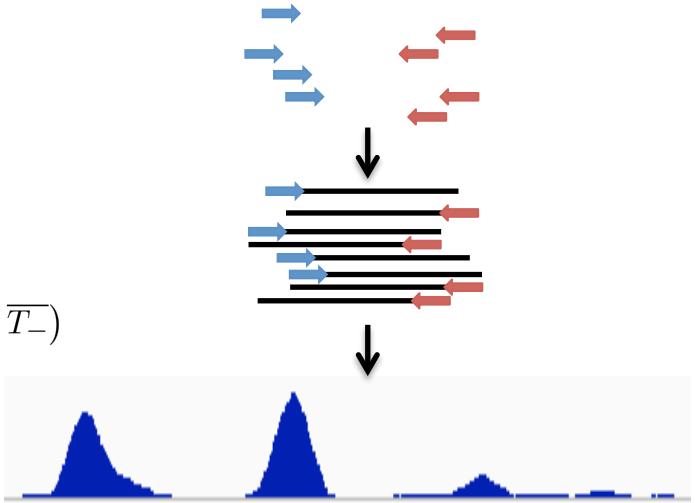
cross-correlation



$$C(r) = \frac{1}{X}\int_x \left(T_+(x) - \overline{T_+}\right)\left(T_-(x+r) - \overline{T_-}\right)$$



- pile-up profiling



- Data visualization:

  UCSC genome browser

  IGV

  WashU EpiGenome Browser

ChIP-Seq data analysis overview: peak calling

- **MACS** (Zhang, 2008)

  For sharp peaks (transcription factor binding, DNase HS)

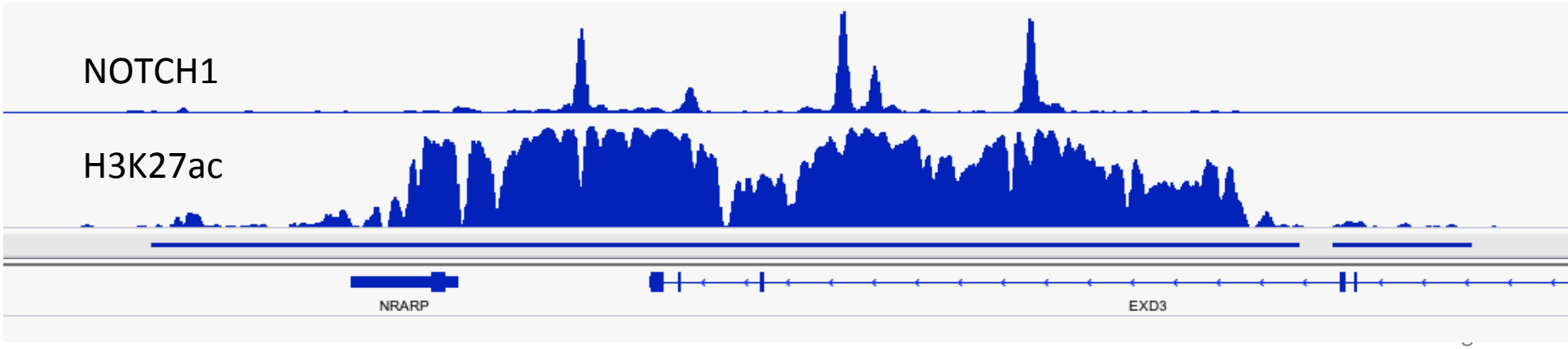
  Poisson model with dynamic local background

  $$\lambda_{local} = max(\lambda_d, \lambda_{1kb}, \lambda_{10kb})$$

- **SICER** (Zang, 2009)

  For broad peaks (histone modifications, "super-enhancers")

  Spatial clustering of localized weak signal and integrative Poisson model



Wang, Zang et al. 2014

cistrome.org/Cistrome/Cistrome_Project.html

Cistrome Project | Myles Brown Lab | Xiaole Shirley Liu Lab

Welcome to Cistrome

The cistrome refers to "the set of cis-acting targets of a trans-acting factor on a genome-wide scale, also known as the in vivo genome-wide location of transcription factor binding-sites or histone modifications". Here we build integrative analysis pipelines (Cistrome) to help experimental biologists, and conduct efficient data integration to better mine the hidden biological insights from publicly available high throughput data.

Learn more »

Cistrome Analysis Pipeline

An integrative and reproducible bioinformatics data analysis platform based on *Galaxy* open source framework. Besides standard Galaxy functions, Cistrome has 29 ChIP-chip- and ChIP-seq-specific tools in three major categories, from preliminary peak calling and correlation analyses to downstream genome feature association, gene expression analyses, and motif discovery.

Visit site »

CistromeMap Data Collection

A web server that provides a comprehensive knowledgebase of all of the publicly available ChIP-Seq and DNase-Seq data in mouse and human. We have manually curated metadata to ensure annotation consistency, and developed a user-friendly display matrix for quick navigation and retrieval of data for specific factors, cells and papers.

Visit site »

Nuclear Receptor Cistrome DB

A curated database of 88 nuclear receptor cistrome data sets and other associated high-throughput data sets including 121 collaborating factor cistromes, 94 epigenomes, and 319 transcriptomes. All the ChIP_chip/seq peak regions are annotated with enriched HRE and co-regulator motifs. A list of predicted hormone response genes from integration of nuclear receptor ChIP_chip/seq data and differential expression data is also readily available to the users.

Visit site »

Cistrome Chromatin Regulator

A knowledgebase on chromatin modifying enzymes and chromatin remodelers. All the chromatin regulators (CR) which possess ChIP-seq data are divided into four categories: reader, writer, eraser and remodeler. Then their basic information and their ChIP-seq data are collected and analysed.

Visit site »

CistromeFinder

CistromeFinder is an application for checking binding sites around a given gene. It has the most comprehensive collection of public ChIP/DNase-seq datasets in human and mouse (over 7,000 samples, including all of ENCODE, epigenome, and more published data from individual papers), which have all gone through a uniform QC and analysis pipeline. .

Visit site »

Cistrome Browser (Beta version)

A new portal to browser public ChIP-seq and DNase-seq datasets. It is intended to replace CistromeFinder and CistromeMap in the future.

Visit site »

CFCE CENTER for FUNCTIONAL CANCER EPIGENETICS | DANA-FARBER CANCER INSTITUTE | HARVARD SCHOOL OF PUBLIC HEALTH

Cistrome.org

Cistrome Analysis Pipeline

cistrome.org/ap/

Galaxy / Cistrome

Analyze Data | Workflow | Shared Data | Lab | Visualization | Help | User | Using 78.4 MB

Tools

search tools

CISTROME TOOLBOX

- Import Data
- 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
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis

NGS TOOLBOX BETA

- NGS: SAM Tools
- NGS: RNA Analysis
- Send Data

2014.5.28 Updates

Merge with latest galaxy code (2014.5.5 7a7985a007fb).

SeqPos updates 2014.4.30

1) The cistrome motif database updates.
2) Add an option "cutoff for hclust of the output", this option only affect how it display the result. Please refer to the result page to check the method.

Welcome to Galaxy/Cistrome!

Thanks to your support and test for our Cistrome AP site, which is based on Galaxy, a Metaserver for integrative analysis of genomic data. Please check our project site at bitbucket for more information.

Server Maintenance Policy

In order to provide a stable system, we schedule periodical server maintenance on **every Friday from 3:00am to 6:00am EST** During maintenance, we will:

1. Update Cistrome server codes if necessary.
2. Clean unused and unlinked data files older than 4 weeks.
3. Delete inactive user accounts older than 8 weeks.
4. Clean ASPERA directory for file upload, and reset ASPERA password. So please contact us again if necessary.

Due to the limit on the server, jobs running on Cistrome server have the following restrictions:

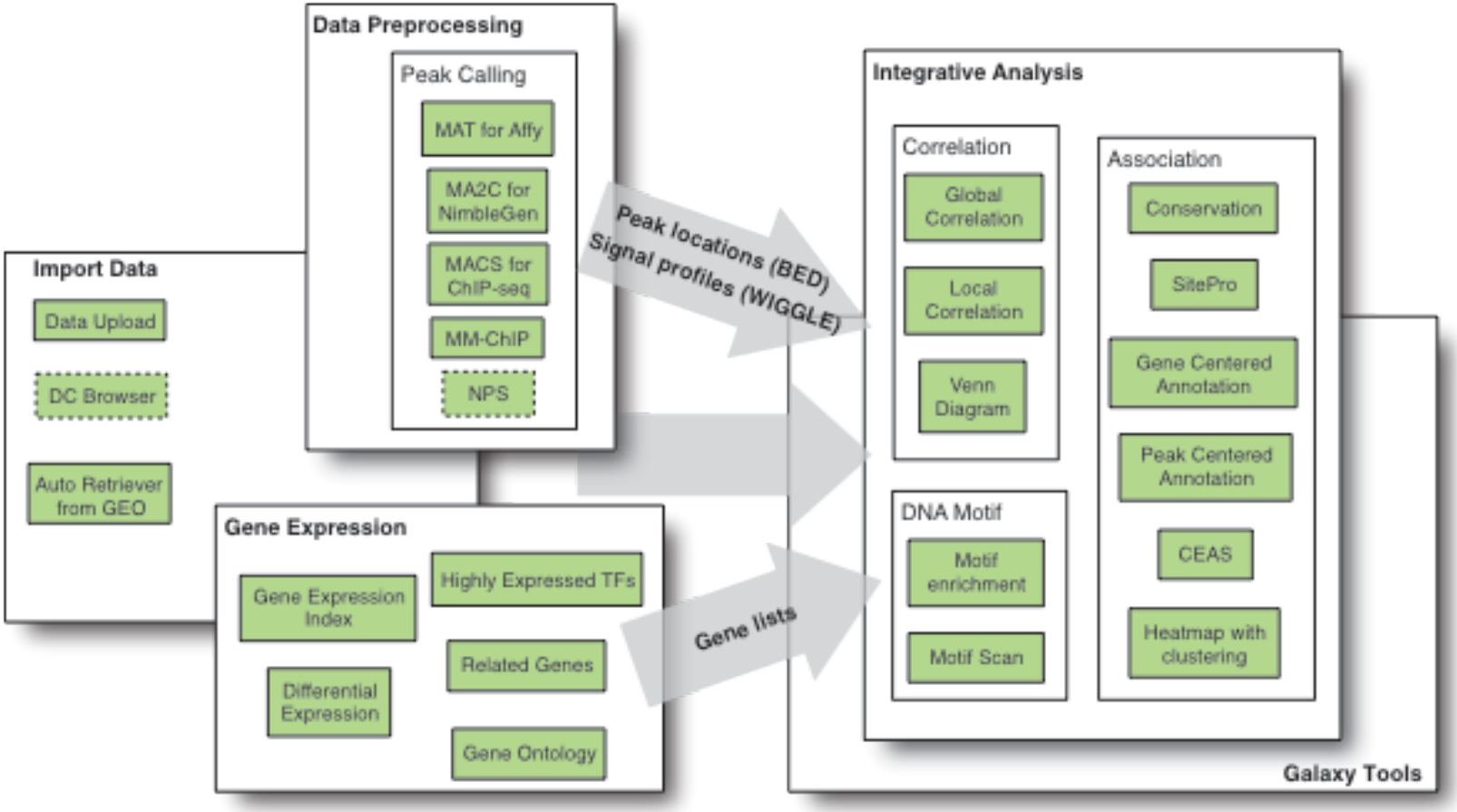
1. Only 20 jobs can be run simultaneously. Other jobs will be put into queue.
2. Any job generating over any single 10 Giga Bytes output will be stopped by Cistrome/Galaxy.
3. Cistrome will check every hour for jobs running over than 24hrs (after the job actually runs on the server), then terminate them. If your job is terminated in this way, the output will appear green and empty. We will improve this

History

test

78.4 MB

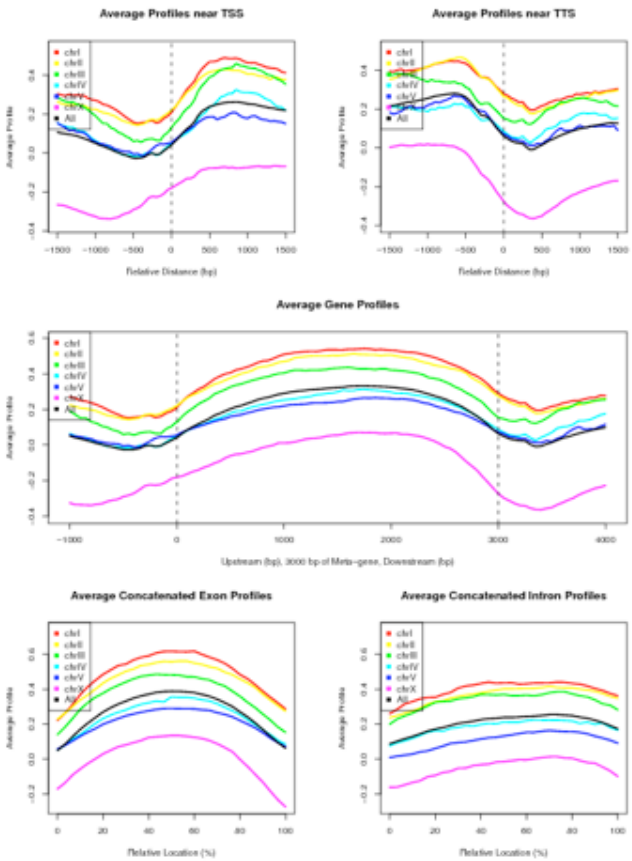
- 71: Log of BETA plus
- 70: Motifs up target regions versus down target regions
- 69: Motifs in down-target regions versus non-target regions
- 68: Motifs in down target regions
- 67: Motifs in up-target regions versus non-target regions
- 66: Motifs in up target regions
- 65: Motif analysis on target regions
- 64: Downtarget associated peaks
- 63: Uptarget associated peaks
- 62: BETA direct targets prediction on down regulated genes
- 61: BETA direct targets p

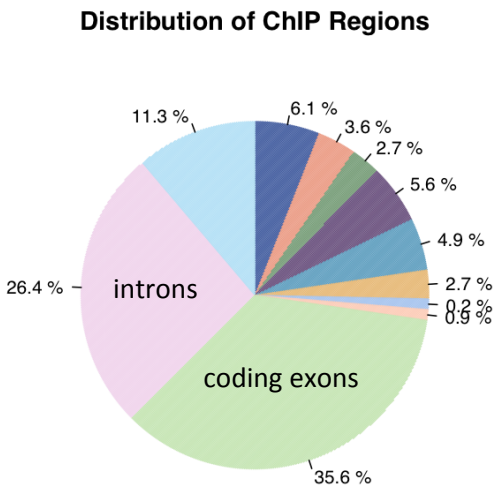


Overview of Cistrome AP project

Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y, et al. 2011. Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol 12: R83.

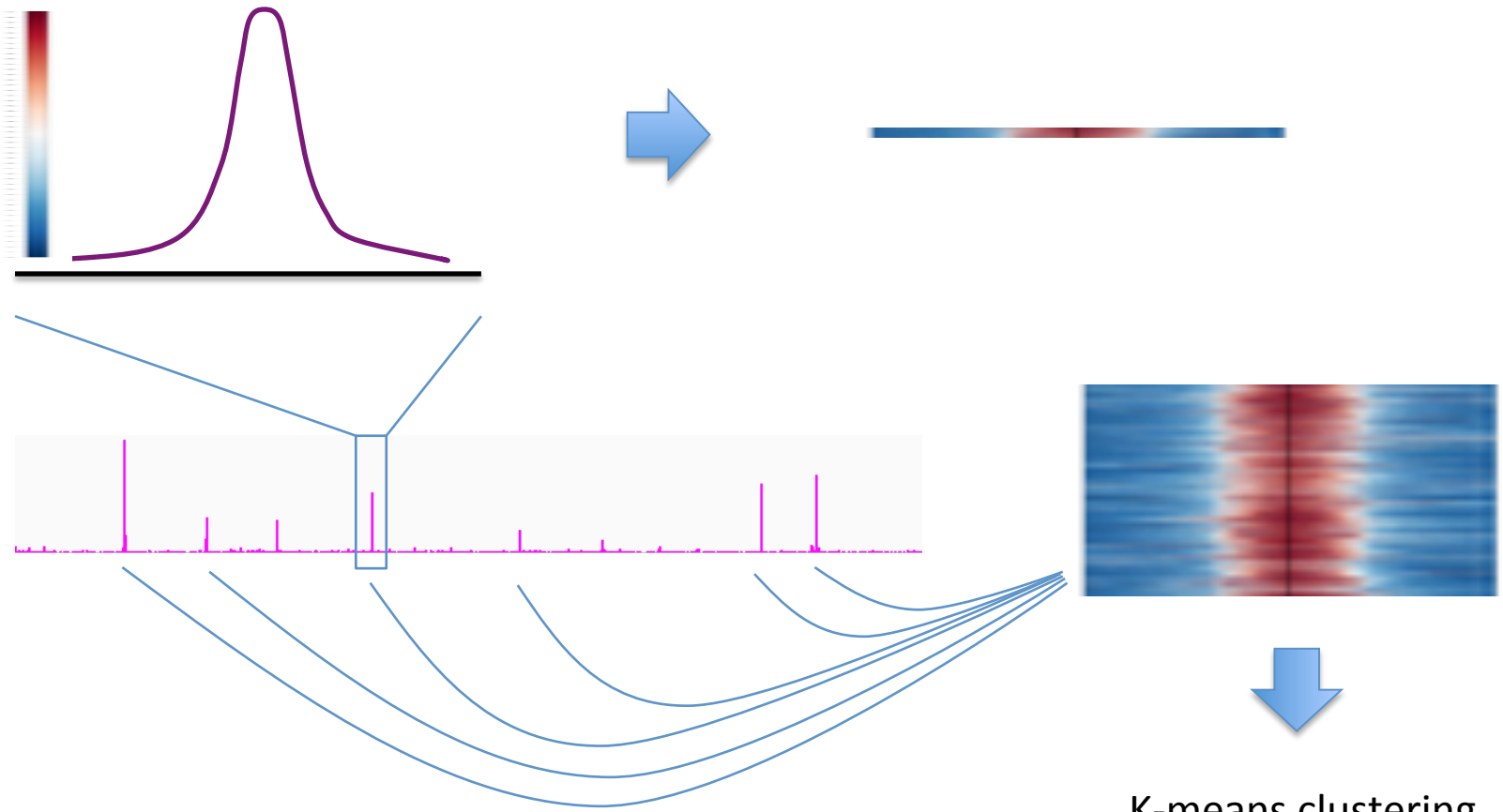
CEAS: annotation and visualization





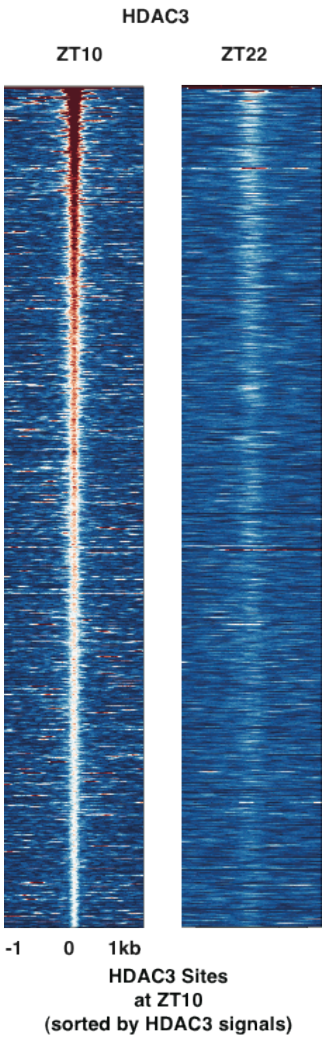
Data from:
Kolasinska-Zwierz P et al. Nat Genet 2009

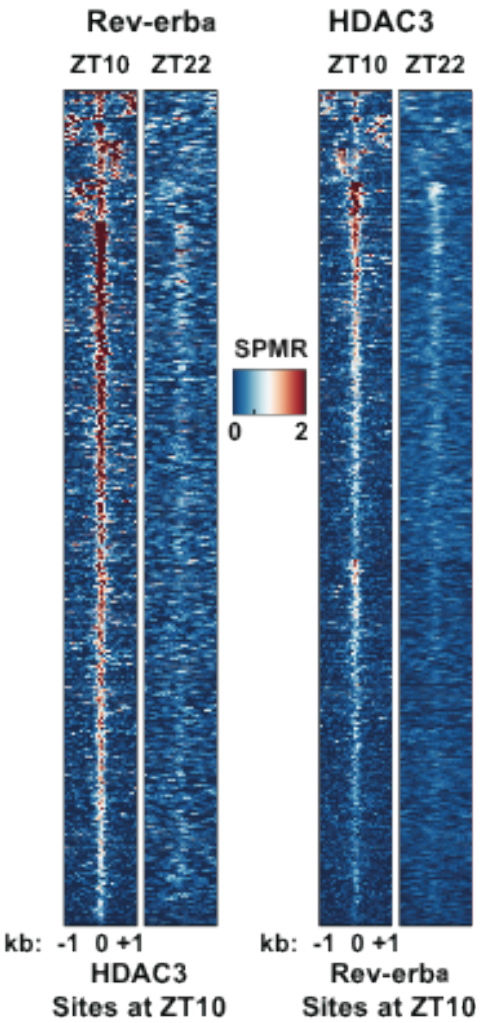
Heatmap and peak clustering

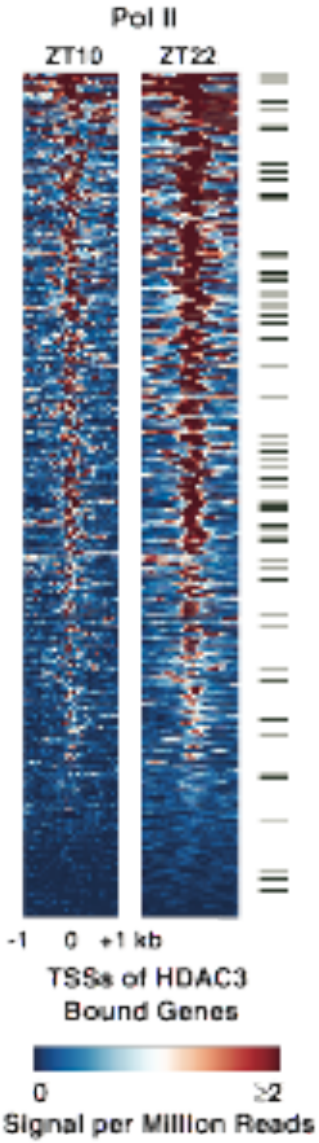


K-means clustering

Heatmap and peak clustering

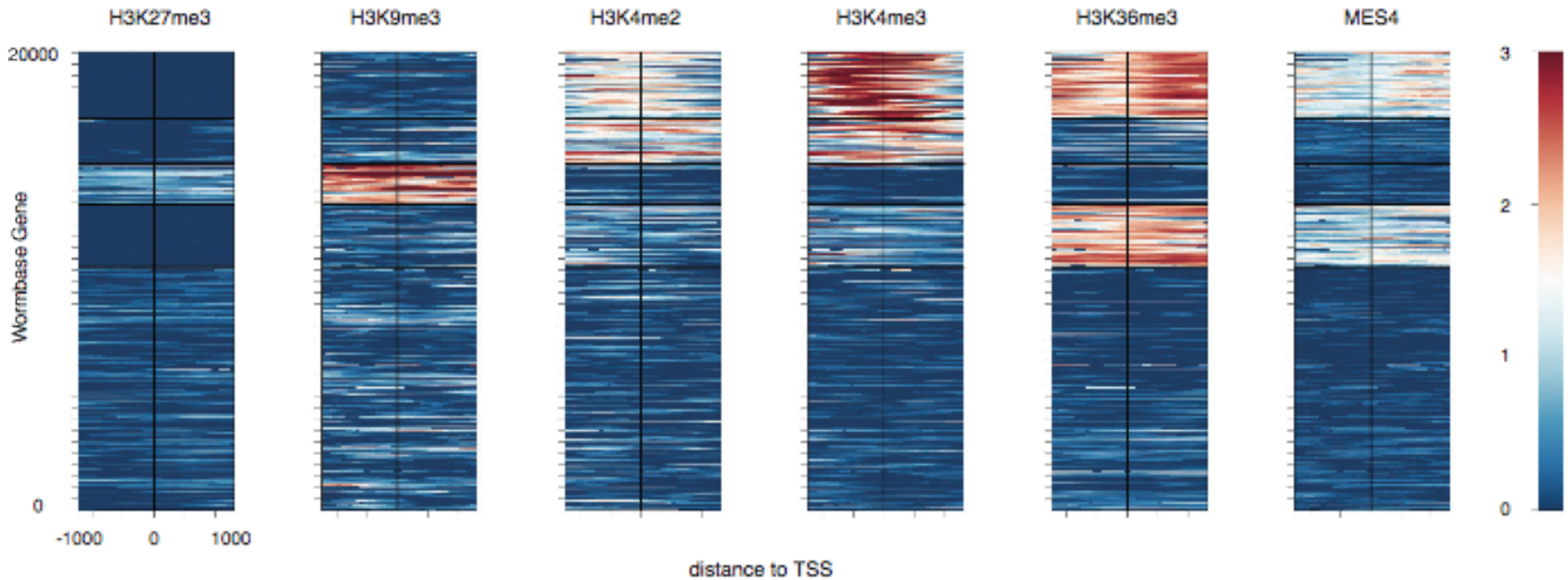






Feng D, et al. Science 2011

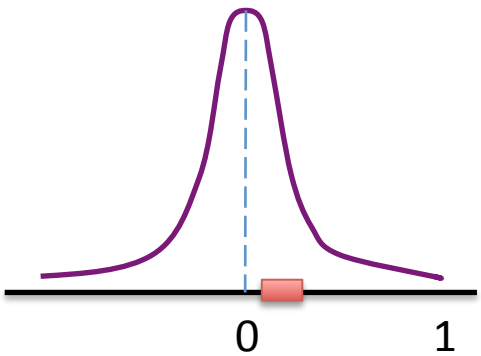
Heatmap and peak clustering



Data from:

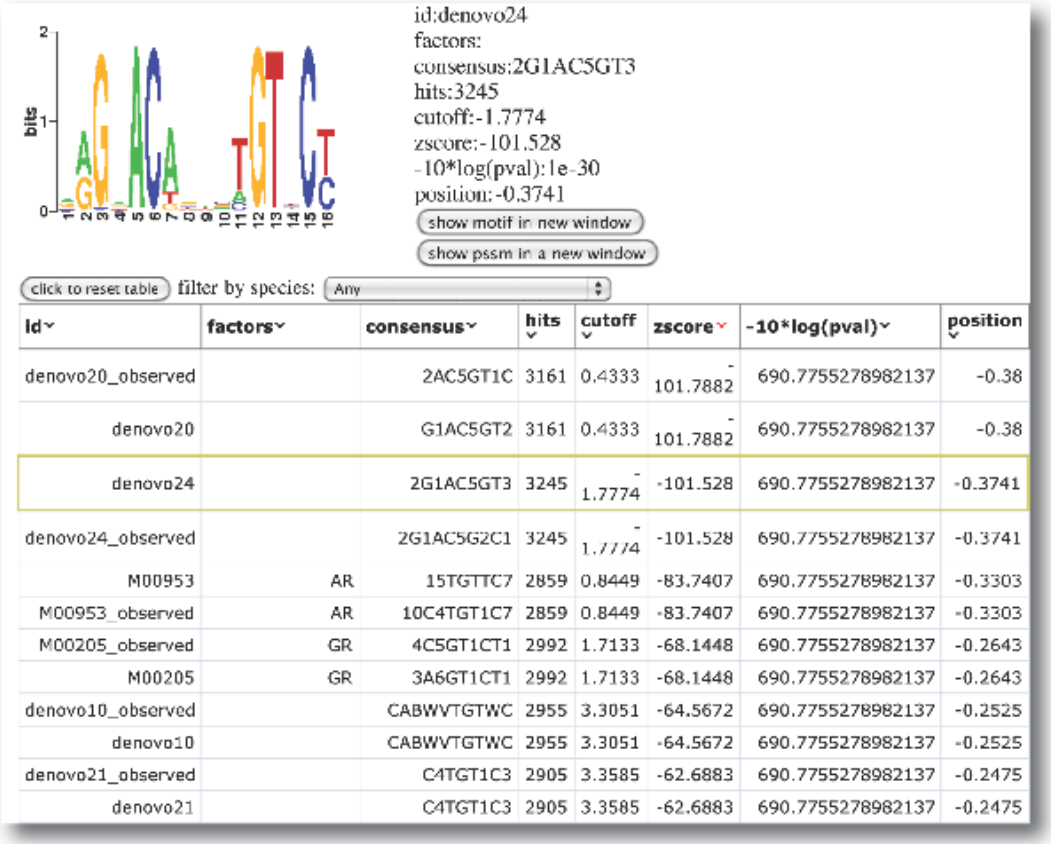
Gerstein MB, et al. Science 2010.

SeqPos motif analysis



$$Z = \sum_{i=1}^{N} \frac{x_i - 0.5}{\sqrt{N/12}}$$

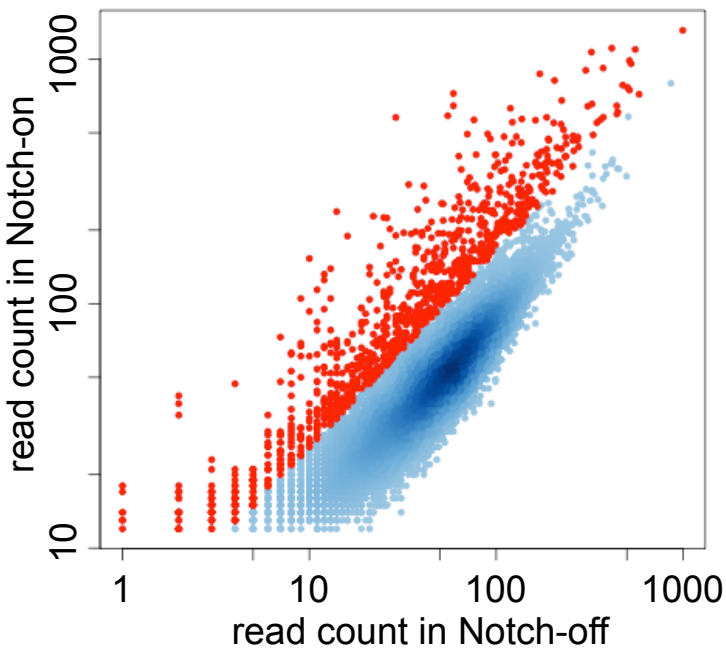
- Search the sequence motifs around peak center positions
- Search in known motif databases
- Perform de novo motif discovery based on MDscan algorithm
- Cluster similar motifs

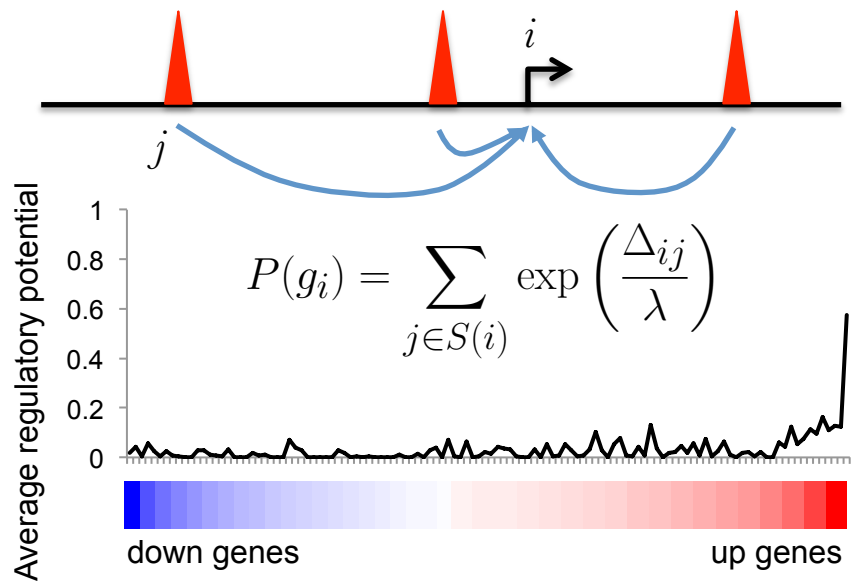


Id	factors	consensus	hits	cutoff	zscore	-10*log(pval)	position
denovo20_observed		2AC5GT1C	3161	0.4333	-101.7882	690.7755278982137	-0.38
denovo20		G1AC5GT2	3161	0.4333	-101.7882	690.7755278982137	-0.38
denovo24		2G1AC5GT3	3245	-1.7774	-101.528	690.7755278982137	-0.3741
denovo24_observed		2G1AC5G2C1	3245	-1.7774	-101.528	690.7755278982137	-0.3741
M00953	AR	15TGTTC7	2859	0.8449	-83.7407	690.7755278982137	-0.3303
M00953_observed	AR	10C4TGT1C7	2859	0.8449	-83.7407	690.7755278982137	-0.3303
M00205_observed	GR	4C5GT1CT1	2992	1.7133	-68.1448	690.7755278982137	-0.2643
M00205	GR	3A6GT1CT1	2992	1.7133	-68.1448	690.7755278982137	-0.2643
denovo10_observed		CABWVTGTWC	2955	3.3051	-64.5672	690.7755278982137	-0.2525
denovo10		CABWVTGTWC	2955	3.3051	-64.5672	690.7755278982137	-0.2525
denovo21_observed		C4TGT1C3	2905	3.3585	-62.6883	690.7755278982137	-0.2475
denovo21		C4TGT1C3	2905	3.3585	-62.6883	690.7755278982137	-0.2475

Integrative Analysis: BETA

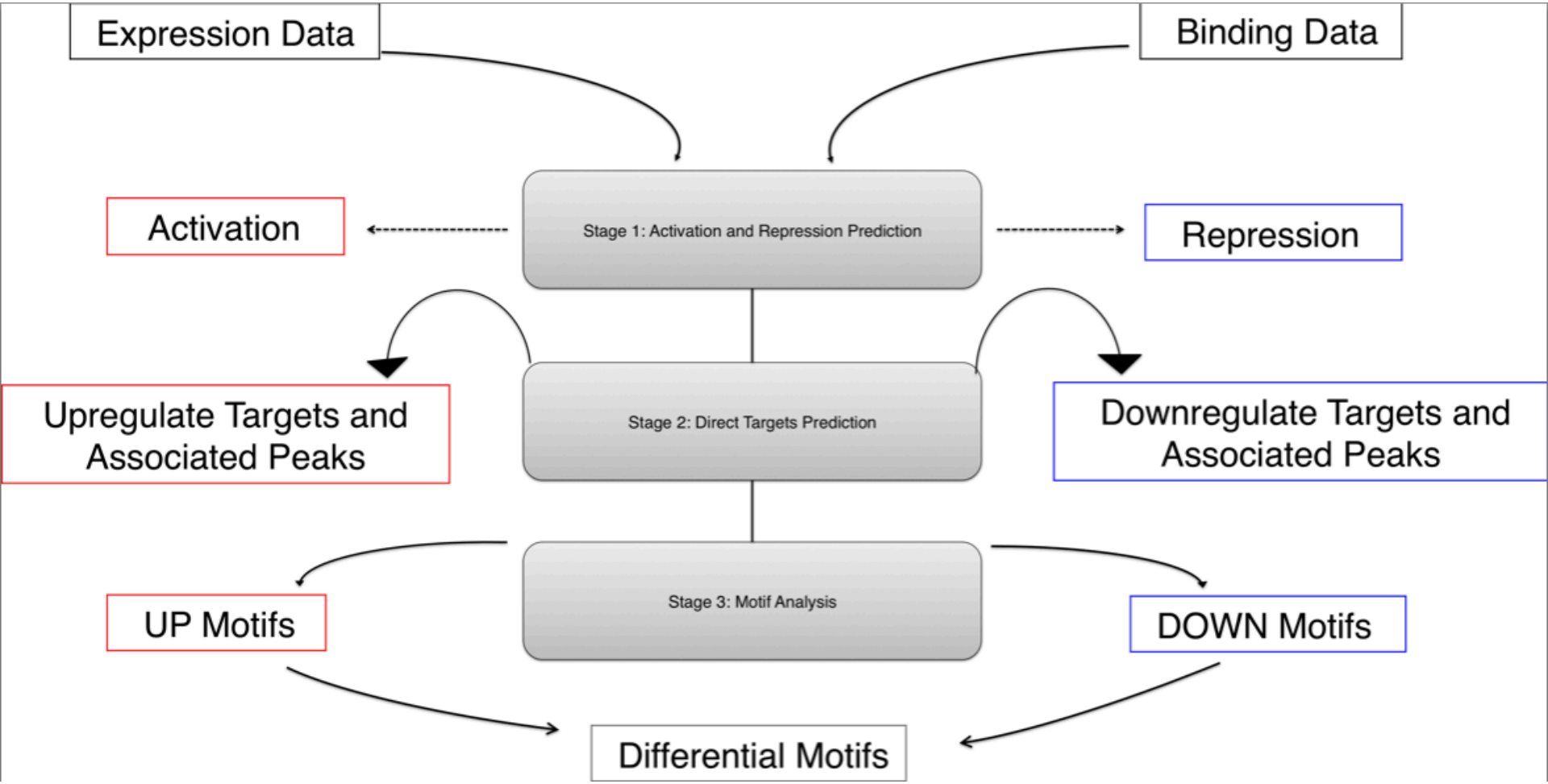
- Correlation with expression (Binding Expression Target Analysis, **BETA**, Wang et al. *Nat. Protoc.* 2013)



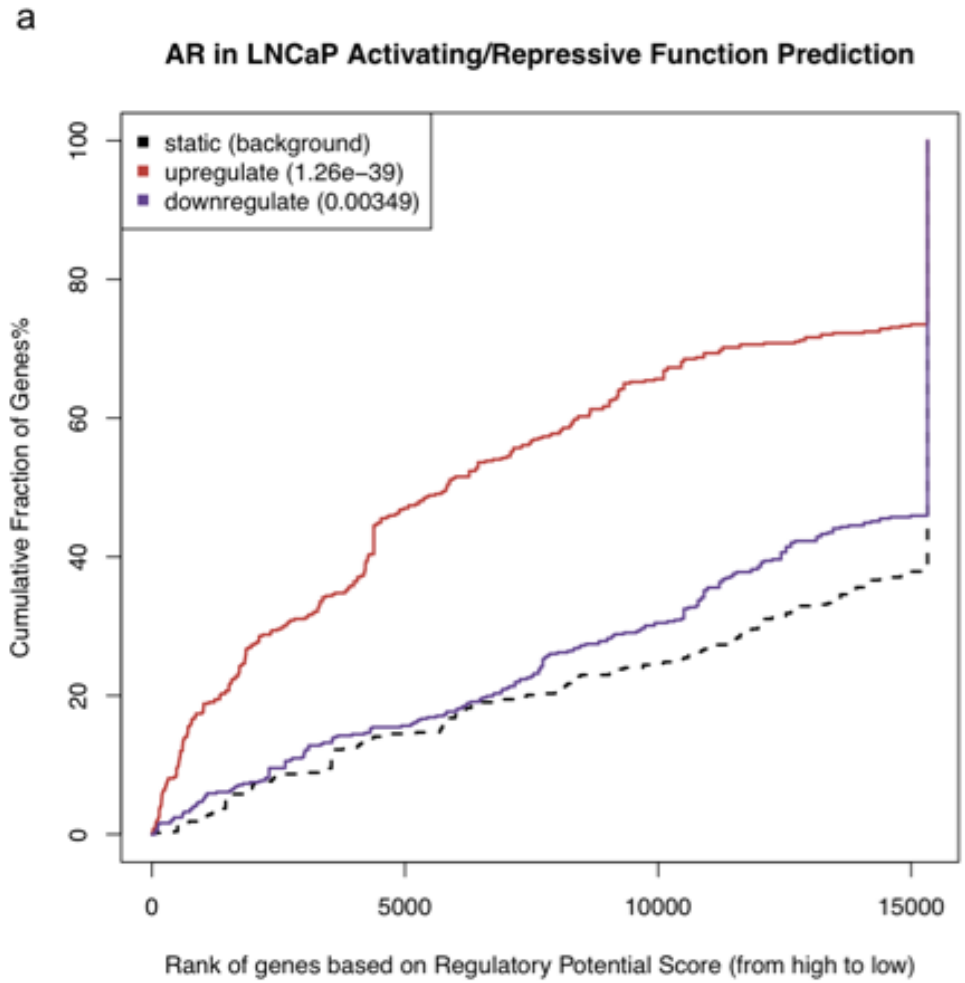


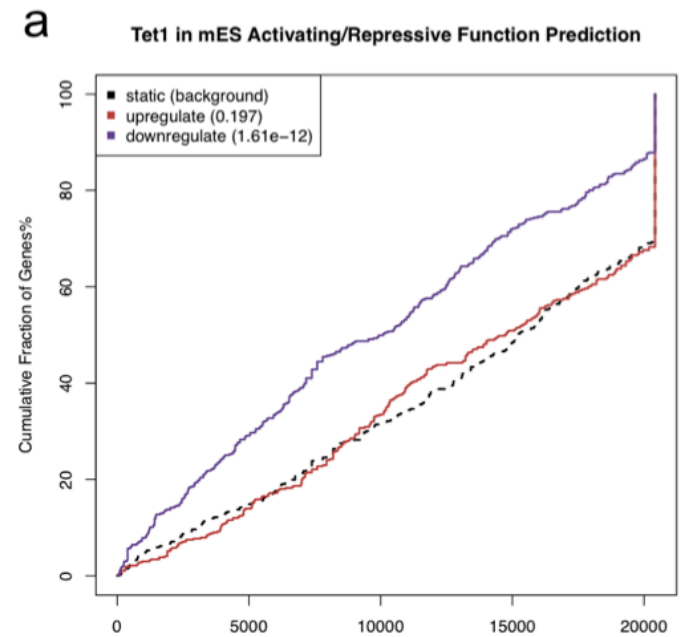
Wang, Zang et al. 2014

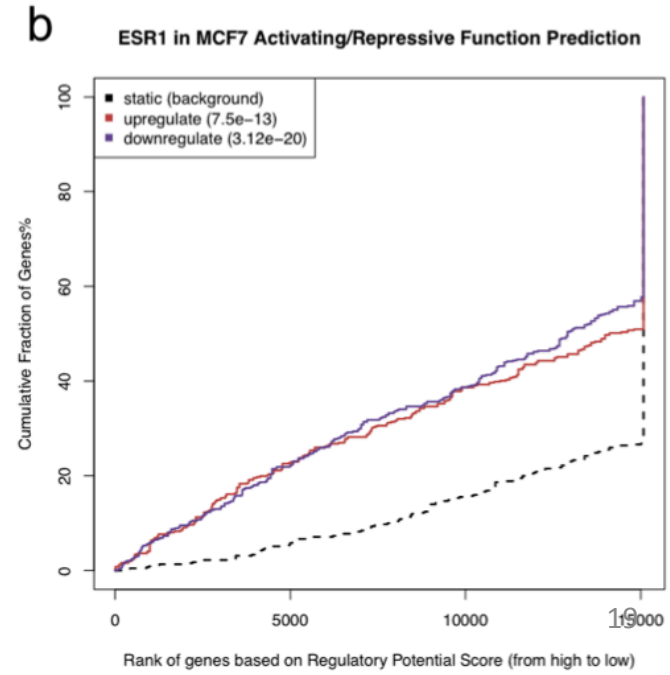
Workflow of BETA



Function prediction







Direct target prediction

$RPg = R_{gb}/n * R_{ge}/n$ (p-value)

Direct target genes

Chroms	txStart	txEnd	refseqID	rank product	Strands	GeneSymbol
chr19	51376688	51383823	NM_001256080	2.186E-07	+	KLK2
chr19	51376688	51383823	NM_005551	2.186E-07	+	KLK2
chr19	51376688	51383823	NR_045762	2.186E-07	+	KLK2
chr19	51376688	51383823	NR_045763	2.186E-07	+	KLK2
chr19	51376688	51383823	NM_001002231	2.186E-07	+	KLK2
chr1	207191865	207206101	NM_023938	8.822E-07	-	C1orf116
chr1	207191865	207206101	NM_001083924	8.822E-07	-	C1orf116
chr21	42836477	42880085	NM_005656	1.03E-06	-	TMPRSS2
chr21	42836477	42879992	NM_001135099	1.04E-06	-	TMPRSS2

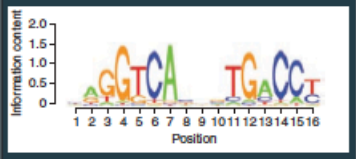
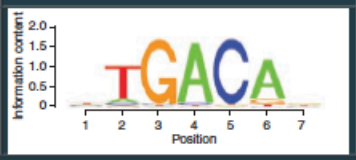
Target gene associated peaks

chrom	pStart	pEnd	Refseq	Symbol	Distance	Score
chr19	51354060	51354999	NM_001256080	KLK2	-22159	0.2500
chr19	51372841	51373704	NM_001256080	KLK2	-3416	0.5291
chr19	51392207	51393248	NM_001256080	KLK2	16039	0.3193
chr19	51354060	51354999	NM_005551	KLK2	-22159	0.2500
chr19	51372841	51373704	NM_005551	KLK2	-3416	0.5291
chr19	51392207	51393248	NM_005551	KLK2	16039	0.3193
chr19	51354060	51354999	NR_045762	KLK2	-22159	0.2500

Motif analysis on target regions

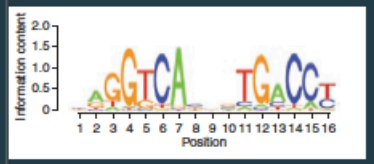
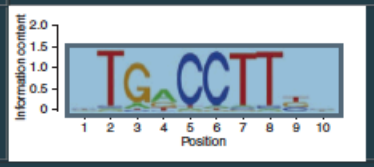
a

PART1: UP TARGET GENES

Symbol	DNA BindDom	Species	Pvalue (T Test)	T Score	Logo
ESR1	Hormone-nuclear Receptor Family	Homo sapiens	1.69e-58	17.54	
ESRRB	Hormone-nuclear Receptor Family				
NR4A2	Hormone-nuclear Receptor Family				
ESRRG	Hormone-nuclear Receptor Family				
ESRRA	Hormone-nuclear Receptor Family				
RARA	Hormone-nuclear Receptor Family				
PPARG	Hormone-nuclear Receptor Family				
NR2F1	Hormone-nuclear Receptor Family				
ESR2	Hormone-nuclear Receptor Family				
MEIS1	Homeodomain Family	Homo sapiens	1.21e-20	9.45	

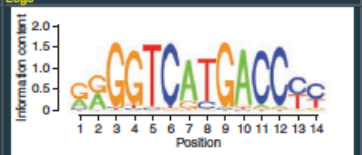
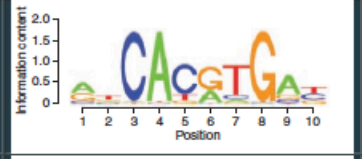
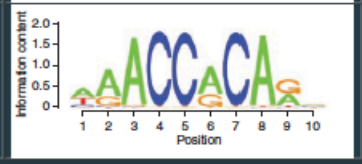
b

PART2: DOWN TARGET GENES

Symbol	DNA BindDom	Species	Pvalue (T Test)	T Score	Logo
ESR1	Hormone-nuclear Receptor Family	Homo sapiens	1.14e-133	27.12	
ESRRB	Hormone-nuclear Receptor Family				
RARA	Hormone-nuclear Receptor Family				
ESRRG	Hormone-nuclear Receptor Family				
ESRRA	Hormone-nuclear Receptor Family				
PPARG	Hormone-nuclear Receptor Family				
NR2F1	Hormone-nuclear Receptor Family				
ESR2	Hormone-nuclear Receptor Family				
NR4A1	Hormone-nuclear Receptor Family	Homo sapiens	3.22e-29	11.33	
RARG	Hormone-nuclear Receptor Family				

c

PART3: UP VS DOWN MOTIF SCAN

Symbol	DNA BindDom	Species	Pvalue (T Test)	T Score	Logo
RXRA	Hormone-nuclear Receptor Family	Homo sapiens	2.14e-06	-4.76	
TFEC	Helix-Loop-Helix Family	Homo sapiens	4.48e-06	-4.61	
RUNX1 FOXH1	Runt Domain Family Forkhead Domain Family	Homo sapiens	4.05e-04	3.55	

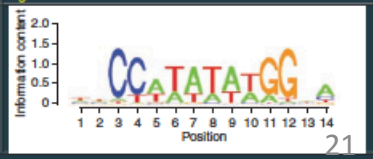
d

PART4: UP VS NON TARGET MOTIF

Symbol	DNA BindDom	Species	Pvalue (T Test)	T Score	Logo
GATA4	GATA Domain Family	Homo sapiens	1.96e-03	3.11	

e

PART5: DOWN VS NON TARGET MOTIF

Symbol	DNA BindDom	Species	Pvalue (T Test)	T Score	Logo
SRF	MADS Box Family	Homo sapiens	8.65e-08	5.37	

Details of Cistrome tools

- Import Data
- Data preprocessing
- Gene expression
- Integrative analysis
- BETA
- Liftover/Others
- Low level file operations

Import data

- Upload file
  - Upload through HTML for small files -- don't close webpage.
  - If you have uploaded files through Aspera, you can select them. Good for huge file >100MB
- Expression CEL file packager
  - It can download Affymetrix expression array CEL files directly from GEO, make treatment and control groups if necessary, then package a cel.zip file suitable for Cistrome expression analyses.

Data preprocessing

- MACS 1.4.1 for ChIP-seq peak calling
  - Input should be sequence alignment in ELAND, BED, SAM, or BAM formats
- MA2C for NimbleGen 2-color array
  - Need POS, NDF, and PairData files
- General peak caller
  - Take wig file from e.g. MACS/MA2C/MAT and recall peaks with other parameter setting
- MMChIP-chip and MMChIP-seq
  - Meta analysis to combine wiggle files from different platform or labs
  - ChIP-chip and ChIP-seq can't be combined

Gene expression

- Gene expression level
  - Input should be cel.zip (Affy) or xys.zip (other)
  - Zip file contains a pheno.txt -- check tool description on Cistrome site for detail
  - Can take output from Expression CEL file packager
- Differential expression
  - Take output from 'Gene expression' tool for either RefSeq, EntrezID, or gene symbols
- Highest expressed TF using GO term
  - Take output from 'Gene expression' tool
- Correlated gene or TFs given gene symbol and GO term
  - .eset file from 'Gene expression' tool
- GO: Gene Ontology
  - Take a list of EntrezIDs, run GO on BP, CC, MF, also send query to DAVID of the first 200 genes (due to limitation on DAVID)
  - If you only have refseq gene list or gene symbols, use 'convert gene ids' tool first
- Histogram or boxplot of expression levels
  - Take the gene expression level file, and a list of genes

Integrative analysis (correlation)

- Correlation of wiggle files in whole genome scale
  - take multiple wiggle files
  - Can draw heatmap with hierarchical clustering
- Correlation of wiggle files within special regions
  - Better to investigate the correlation at such as certain binding sites or DNAse regions or TSS regions
- Correlation of two wiggle files in the union region of two regional BED files
  - Better to check if two replicates are consistent
- Venn diagram
  - Show the overlap between up to 3 regional BED files
  - Better to show the co-localization of two or three TFs

Integrative analysis (association)

- CEAS: summarize the bias of cistrome
  - take regional BED file and optional wiggle file
  - multiple pages report including profiling on metagene body
- Sitepro: draw aggregation plot around given sites
  - Take multiple wig files or bed files
  - Can be used to show e.g. histone marks around TFBS
- Conservation plot at given sites
- Heatmap: the signal pattern around given regions
  - Can use multiple wiggle files or only use one of them to either do k-means clustering or sort, then reorder all sites.
  - Can output region in each cluster, to be combined with other tools

Integrative analysis (motif)

- SeqPos motif discovery and search
  - Perform both de novo motif discovery and known motif search in PBM, Y1H, Transfac, Jaspar, and our curated Cistrome motif collection
  - Consider the distance between the middle points of given sites and motif locations
- Screen motif tool
  - take a motif and given regions, scan the occurrences.

Precompiled workflows

Name	Description
General ChIP-seq	A generic ChIP-seq pipeline for Next Generation Sequencing platform data of single replicate
ChIP-seq with two replicates	Calculate correlation of two ChIP-seq replicates
Generate differential gene list	Take the differential expression result and generate the up/ down-regulated genes, which can be used in CEAS.
From Heatmap clustering to Gene names	Take the Heatmap clustering results on gene TSSs, then separate the first 5 clusters with distinct patterns, which can be followed by GO analysis
BAM to BED	Convert BAM format file to BED while filtering out unmapped reads
Randomly select reads in BAM	Randomly sample BAM file to given number of reads in BED format
Find regions with two different motifs	Scan given regions of two different motifs, find the regions with two non-overlapping different motifs

Cistrome Dataset Browser

cistrome.org/Cistrome/Cistrome_Project.html

Cistrome Project | Myles Brown Lab | Xiaole Shirley Liu Lab

Welcome to Cistrome

The cistrome refers to "the set of cis-acting targets of a trans-acting factor on a genome-wide scale, also known as the in vivo genome-wide location of transcription factor binding-sites or histone modifications". Here we build integrative analysis pipelines (Cistrome) to help experimental biologists, and conduct efficient data integration to better mine the hidden biological insights from publicly available high throughput data.

Learn more »

Cistrome Analysis Pipeline

An integrative and reproducible bioinformatics data analysis platform based on *Galaxy* open source framework. Besides standard Galaxy functions, Cistrome has 29 ChIP-chip- and ChIP-seq-specific tools in three major categories, from preliminary peak calling and correlation analyses to downstream genome feature association, gene expression analyses, and motif discovery.

Visit site »

CistromeMap Data Collection

A web server that provides a comprehensive knowledgebase of all of the publicly available ChIP-Seq and DNase-Seq data in mouse and human. We have manually curated metadata to ensure annotation consistency, and developed a user-friendly display matrix for quick navigation and retrieval of data for specific factors, cells and papers.

Visit site »

Nuclear Receptor Cistrome DB

A curated database of 88 nuclear receptor cistrome data sets and other associated high-throughput data sets including 121 collaborating factor cistromes, 94 epigenomes, and 319 transcriptomes. All the ChIP_chip/seq peak regions are annotated with enriched HRE and co-regulator motifs. A list of predicted hormone response genes from integration of nuclear receptor ChIP_chip/seq data and differential expression data is also readily available to the users.

Visit site »

Cistrome Chromatin Regulator

A knowledgebase on chromatin modifying enzymes and chromatin remodelers. All the chromatin regulators (CR) which possess ChIP-seq data are divided into four categories: reader, writer, eraser and remodeler. Then their basic information and their ChIP-seq data are collected and analysed.

Visit site »

CistromeFinder

CistromeFinder is an application for checking binding sites around a given gene. It has the most comprehensive collection of public ChIP/DNase-seq datasets in human and mouse (over 7,000 samples, including all of ENCODE, epigenome, and more published data from individual papers), which have all gone through a uniform QC and analysis pipeline. .

Visit site »

Cistrome Browser (Beta version)

A new portal to browser public ChIP-seq and DNase-seq datasets. It is intended to replace CistromeFinder and CistromeMap in the future.

Visit site »

CFCE CENTER for FUNCTIONAL CANCER EPIGENETICS | DANA-FARBER CANCER INSTITUTE | HARVARD SCHOOL OF PUBLIC HEALTH

Cistrome.org

cistrome.org/dc

Cistrome Browser Home About LOGIN

Dataset Browser

Containing word(s): Search Options

Species
All
Homo sapiens
Mus musculus

Biological Sources
All
106A
10T1/2
22RV1
266.6
2TS22C

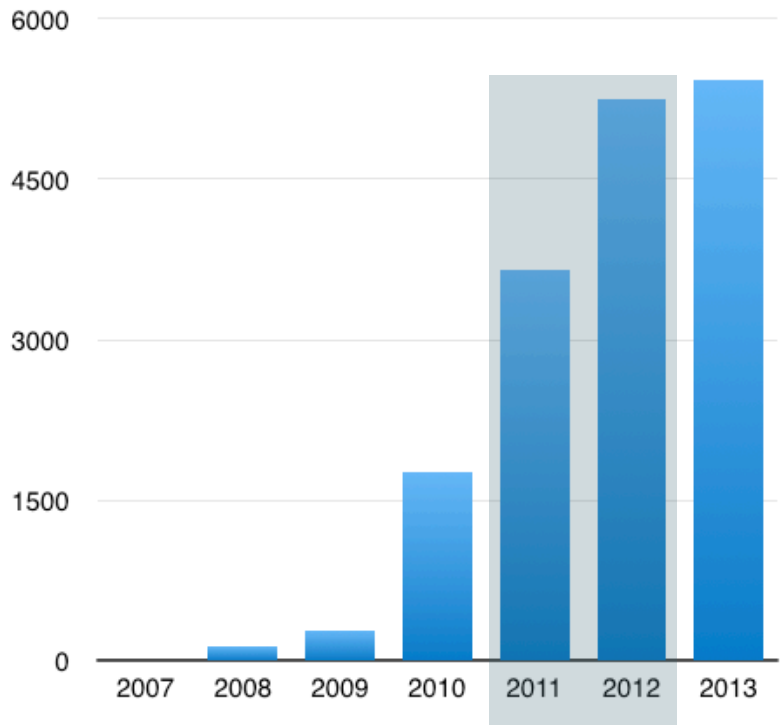
Factors
All
5HMC
7SK
ADNP
ADNP2
AEBP2

Results

Species	Biological Source	Factor	Publication	Status
Mus musculus	HPC-7; Hematopoietic Progenitor; Blood	LMO2	Knezevic K, et al. Mol. Cell. Biol. 2011	complete
Mus musculus	HPC-7; Hematopoietic Progenitor; Blood	FLI1	Knezevic K, et al. Mol. Cell. Biol. 2011	complete
Mus musculus	MEL86; Erythroleukemia Cell; C57BL/6; Blood	GATA1	Yu M, et al. Mol. Cell 2009	complete
Mus musculus	E11.5; Forebrain	EP300	Blow MJ, et al. Nat. Genet. 2010	complete
Mus musculus	E11.5; CD-1; Embryonic Midbrain	EP300	Blow MJ, et al. Nat. Genet. 2010	complete
Mus musculus	E11.5; CD-1; Embryonic Limb	EP300	Blow MJ, et al. Nat. Genet. 2010	complete

Cistrome Dataset Browser

12,937 ChIP-seq datasets have been collected



Numbers of ChIP-seq samples on GEO

Features of ChIP-seq datasets

- Species
- Biological source: tissue, cell type, disease, condition, etc.
- Factor
- Publication
- Quality Control:
  - Sequence quality: Raw sequence median quality score and raw read GC content
  - Mapping quality: Uniquely mapped ratio
  - Library complexity (PBC): PCR bottleneck coefficient
  - ChIP enrichment: sufficient number of peaks with good enrichment
  - Signal to noise ratio (FRiP): Fraction of reads in peaks
  - Evolutionary conservation: Phastcons score around the peak summits
  - Regulatory regions: DHS overlapped ratio in top 5000 peaks
  - Motif: enrichment of corresponding motifs in peaks

Features of ChIP-seq samples

- Visualize:
  - WashU Browser
  - UCSC Browser
- Download:
  - Peaks (BED)
  - BigWig
  - Putative target genes
- Similar datasets

cistrome.org/dc

Inspector

Title:	**Treatments** • LNCaP_AR_Vehicle **Controls** • LNCaP_Input_AR-stimulated
Species:	Homo sapiens
Citation:	Choudhary V, et al. Novel role of androgens in mitochondrial fission and apoptosis. Mol. Cancer Res. 2011 PMID: 21724752
Factor:	AR
Biological Source:	**Cell Line:** LNCaP **Cell Type:** Epithelial **Tissue:** Prostate **Disease:** Prostate Carcinoma

Quality Control

Visualize

WashU Browser

UCSC Browser

Download

BED Peaks

BIGWIG File

Putative Targets

Tools

QC reports | Get top putative targets | Check a putative target | Find similar datasets

QC	2722_AR_treat_rep1
Raw sequence median quality score	30
% Reads uniquely mapped	64.4%
PCR bottlenect coefficient (PBC)	99.3%
Number of merged Total/Fold 10/Fold 20 peaks	261 / 77 / 9

Cistrome Dataset Browser

Summary

- Cistrome Analysis Pipeline
  - Peak calling
  - Integrative analysis
  - BETA
- Cistrome Dataset Browser
  - Browse and reuse published ChIP-seq data

Acknowledgments

Xiaole Shirley Liu
Tao Liu
Len Taing
Hanfei Sun
Su Wang
Yong Zhang
Clifford Meyer
Hyunjin Shin
Jian Ma
Chenfei Wang
Qiu Wu
Qian Qin
Shenglin Mei
Bo Qin

Myles Brown
Keji Zhao
Weiqun Peng
Henry Long
Ramesh A. Shivdasani
Jon C. Aster

All Cistrome users

NIH

LEUKEMIA & LYMPHOMA SOCIETY®
fighting blood cancers