The Epigenome Tools 2: ChIP-Seq and Data Analysis

Chongzhi Zang

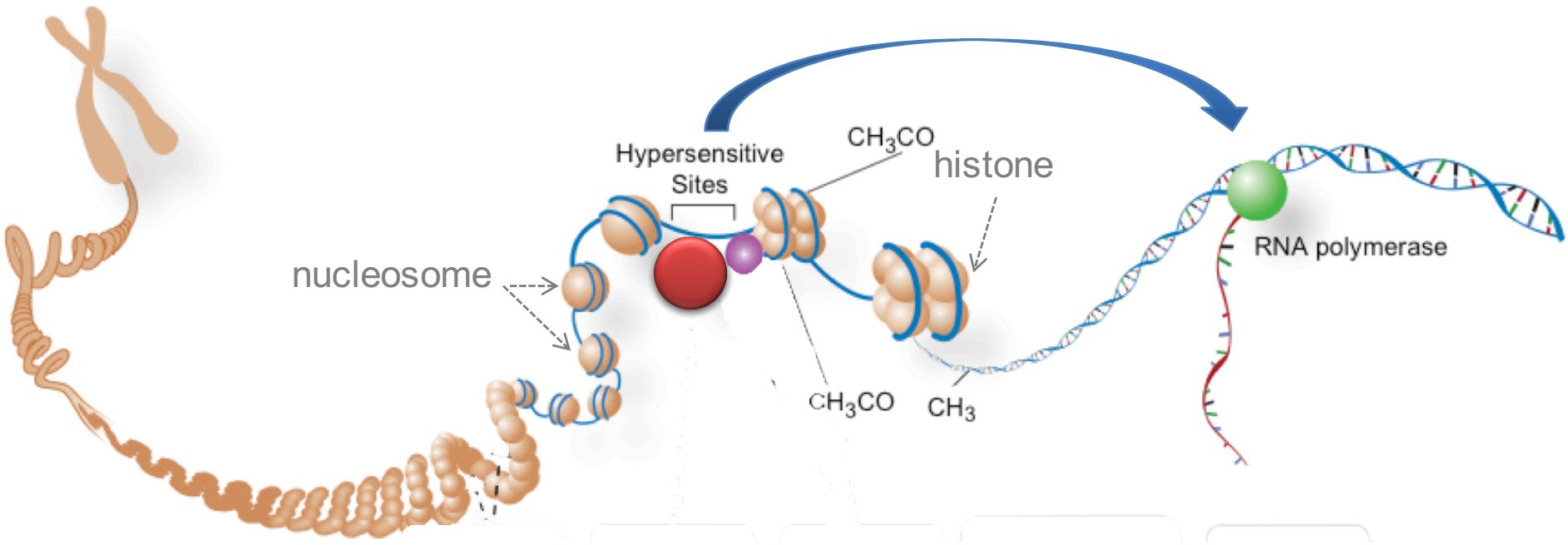
zang@virginia.edu
http://zanglab.com

PHS5705: Public Health Genomics
March 20, 2017

Outline

- Epigenome: basics review
- ChIP-seq overview
- ChIP-seq data analysis

Epigenome



The *epigenome* is a multitude of chemical compounds that can tell the *genome* what to do. The epigenome is made up of chemical compounds and proteins that can attach to DNA and direct such actions as turning genes on or off, controlling the production of proteins in particular cells. -- from genome.gov

Original figure from ENCODE, Darryl Leja (NHGRI), Ian Dunham (EBI)

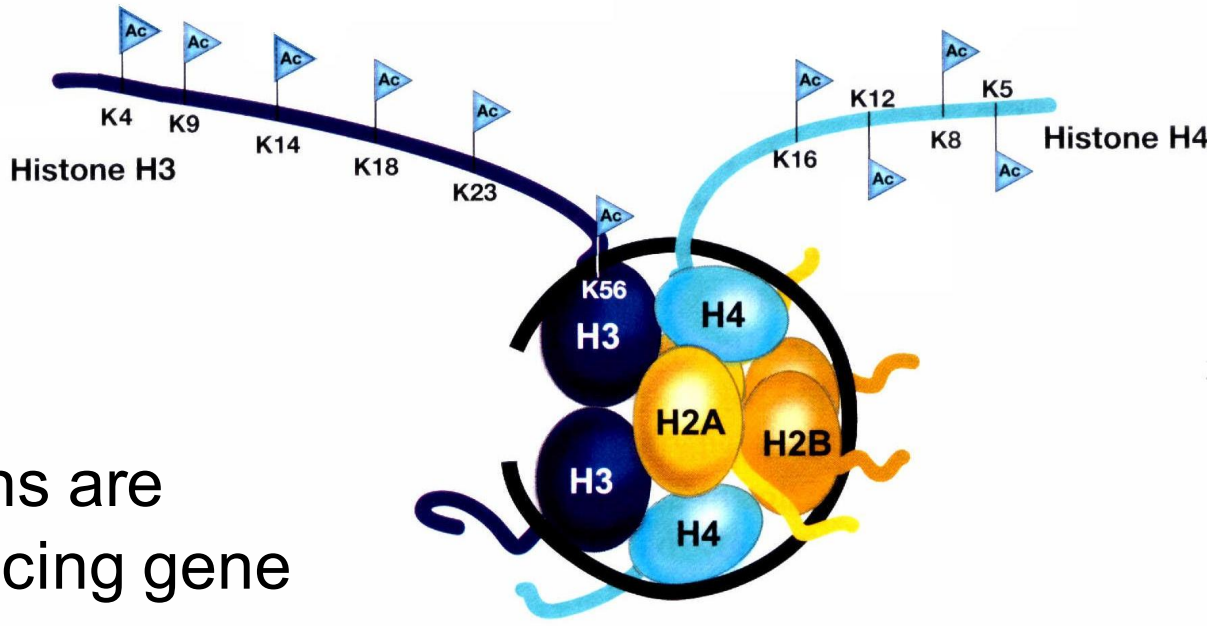
Epigenomic marks

- DNA methylation
- Histone marks
  - Covalent modifications
  - Histone variants
- Chromatin regulators
  - Histone modifying enzymes
  - Chromatin remodeling complexes
- * Transcription factors

Histone modifications

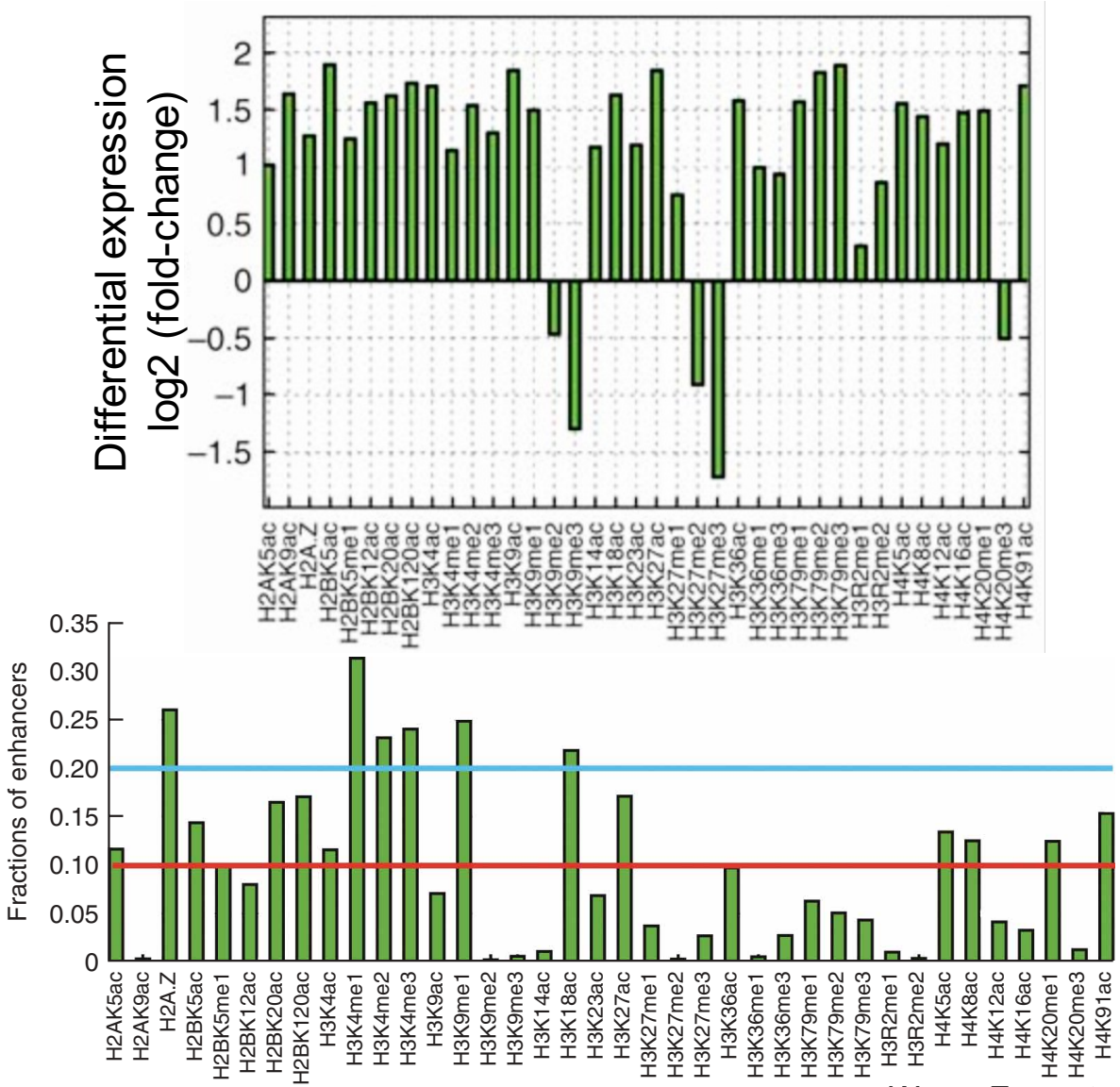
- Nucleosome Core Particles
- Core Histones: H2A, H2B, H3, H4
- Covalent modifications on histone tails include:
  - methylation (me),
  - acetylation (ac),
  - phosphorylation …
- Histone variants
- Histone modifications are implicated in influencing gene expression.

Notation:
H3K4me3



Allis C. et al. Epigenetics. 2006

Histone modifications associate with regulation of gene expression



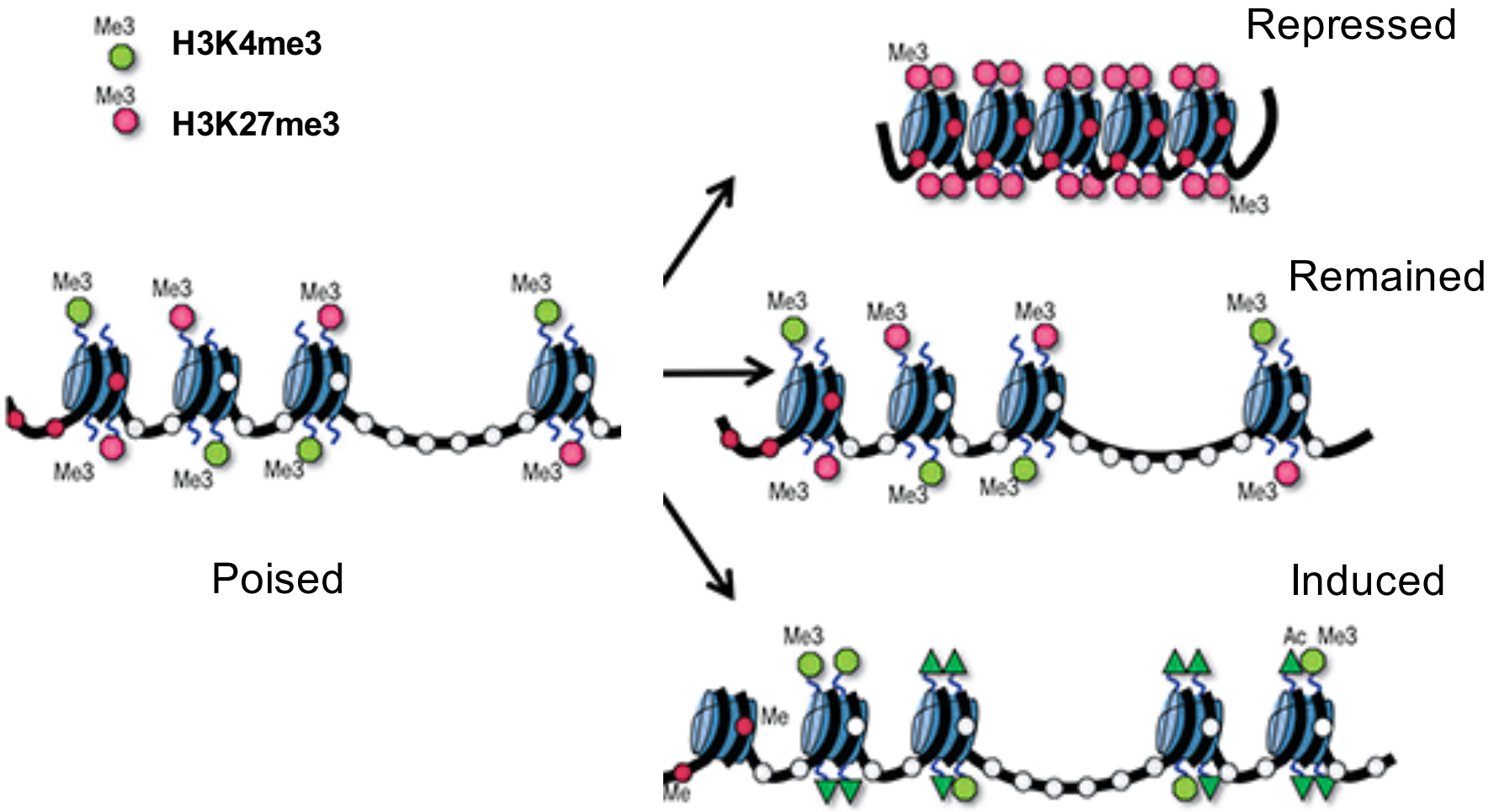
Wang, Zang et al. *Nat Genet* 2008

“Functions” of histone marks

Functional Annotation	Histone Marks
Promoters	H3K4me3
Bivalent/Poised Promoter	H3K4me3/H3K27me3
Transcribed Gene Body	H3K36me3
Enhancer (both active and poised)	H3K4me1
Poised Developmental Enhancer	H3K4me1/H3K27me3
Active Enhancer	H3K4me1/H3K27ac
Polycomb Repressed Regions	H3K27me3
Heterochromatin	H3K9me3

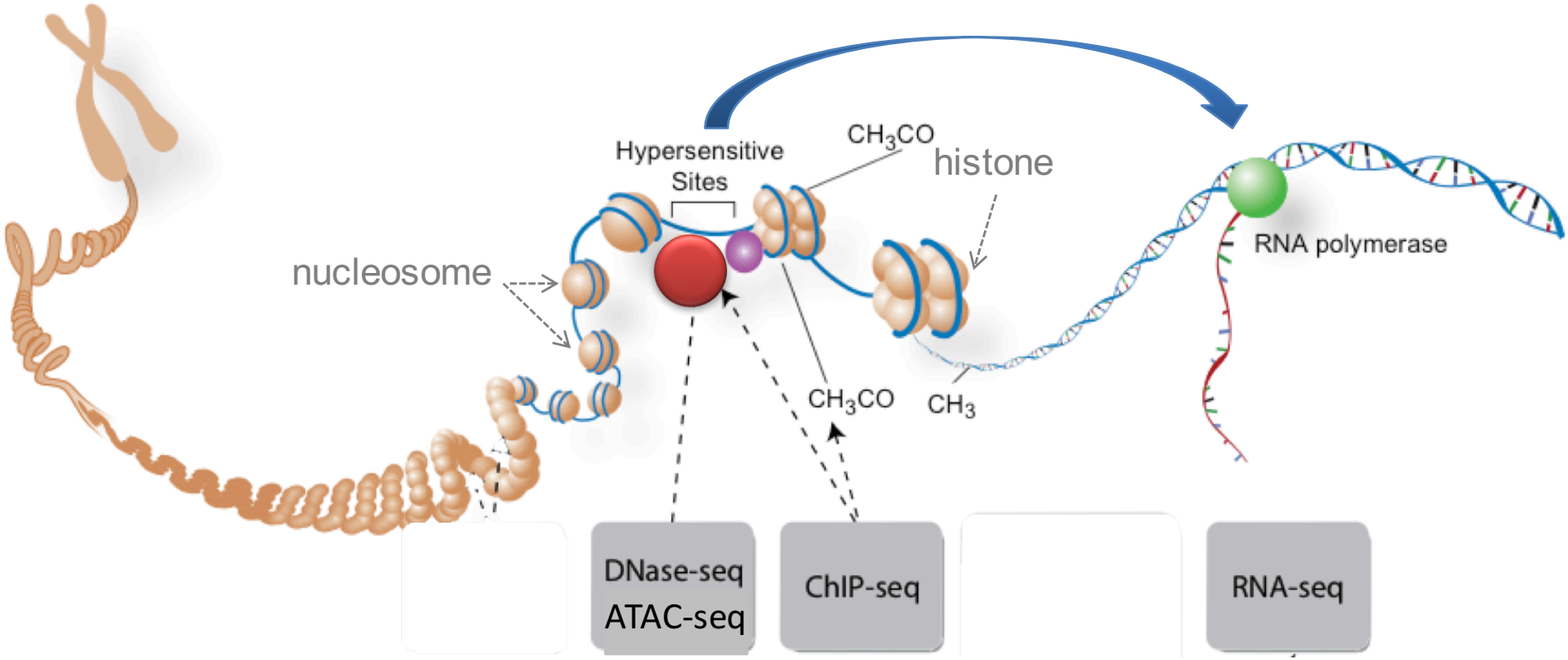
Rivera & Ren *Cell* 2013

H3K4me3/H3K27me3 Bivalent Domain



From: https://pubs.niaaa.nih.gov/publications/arcr351/77-85.htm

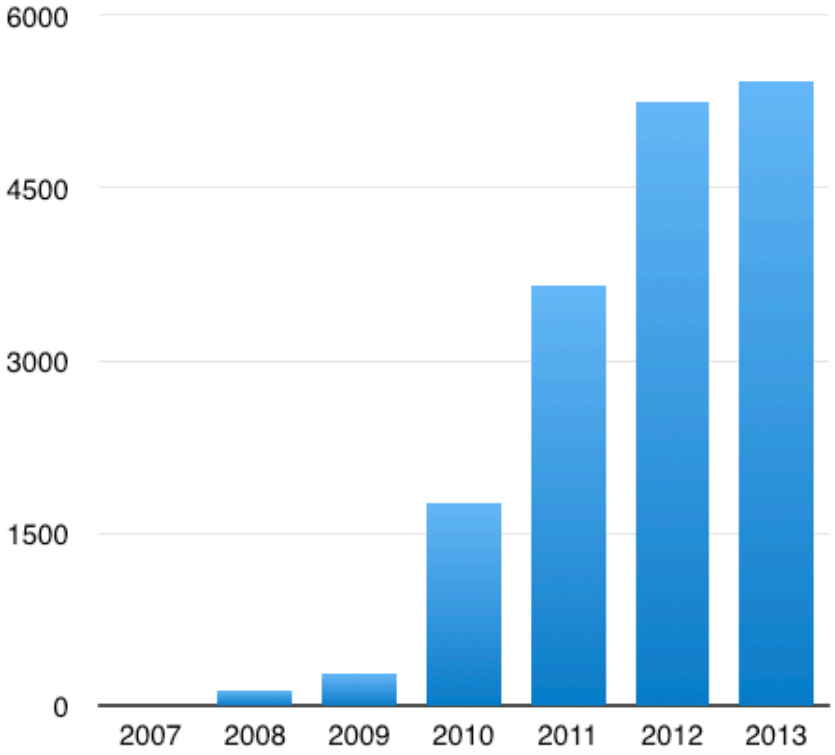
ChIP-seq: Profiling epigenomes with sequencing

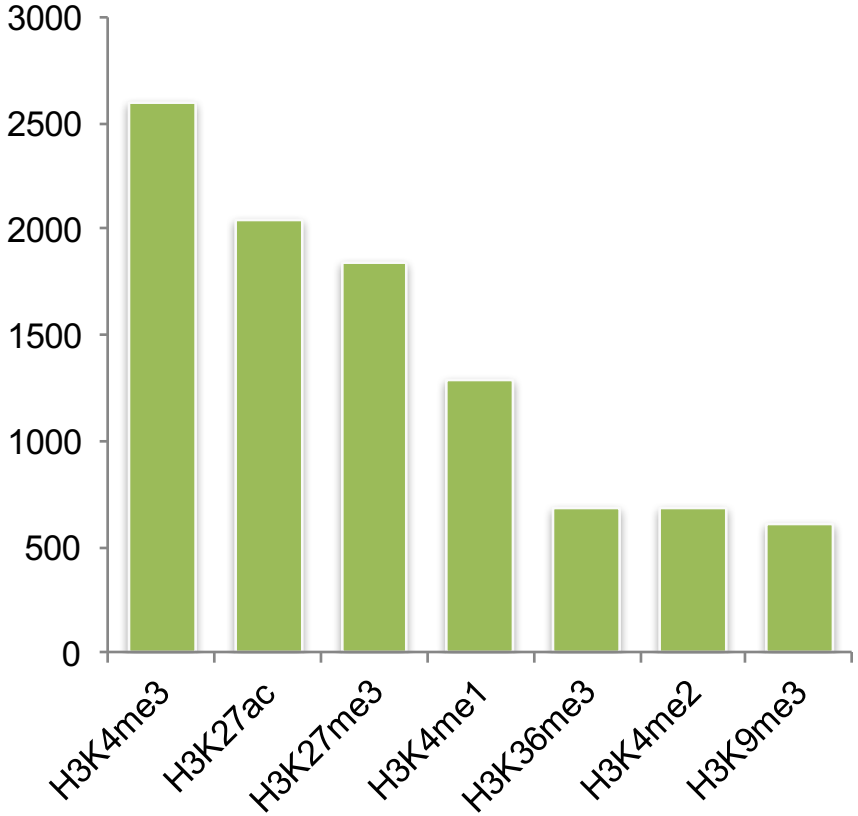


Original figure from ENCODE, Darryl Leja (NHGRI), Ian Dunham (EBI)

Published ChIP-seq datasets are skyrocketing
We are entering the Big Data era

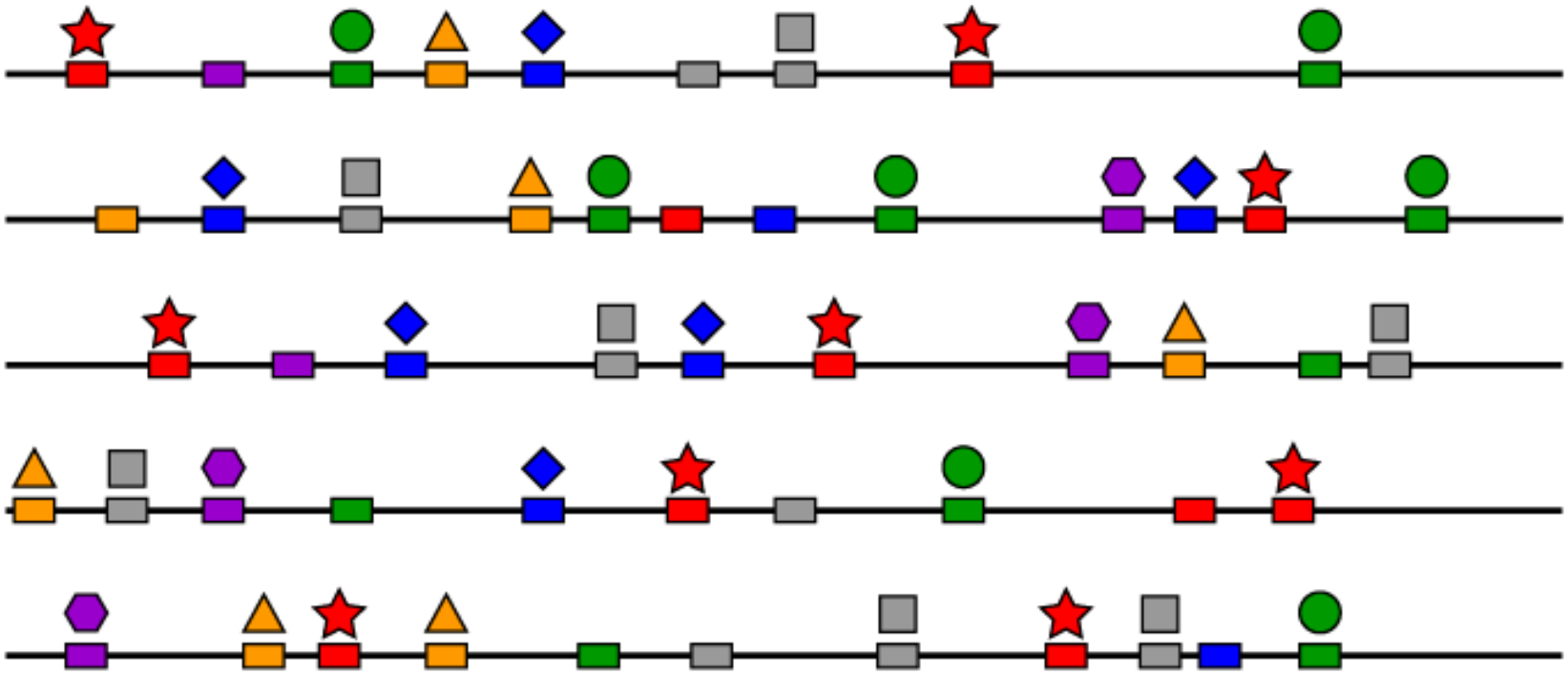
Number of ChIP-seq datasets on GEO



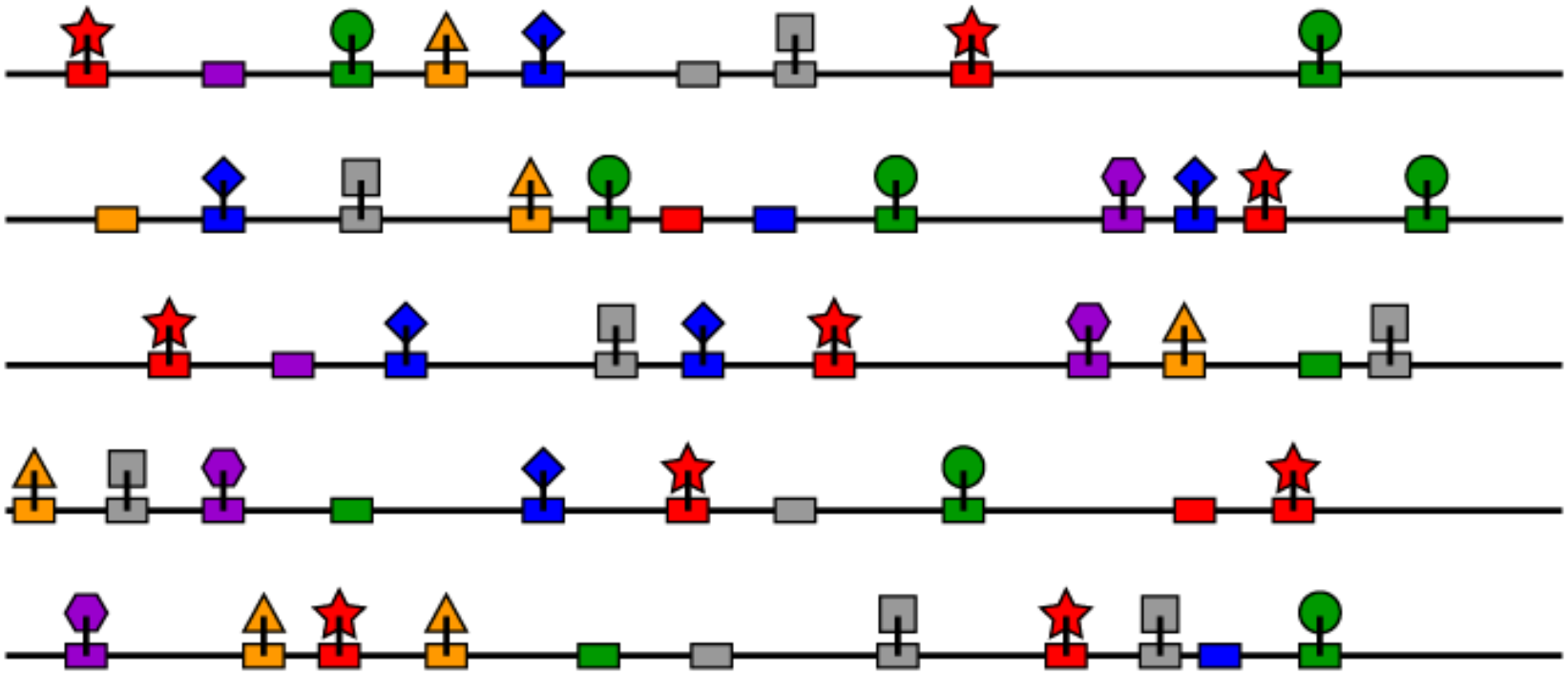


Mei et al. *Nucleic Acids Research* 2016

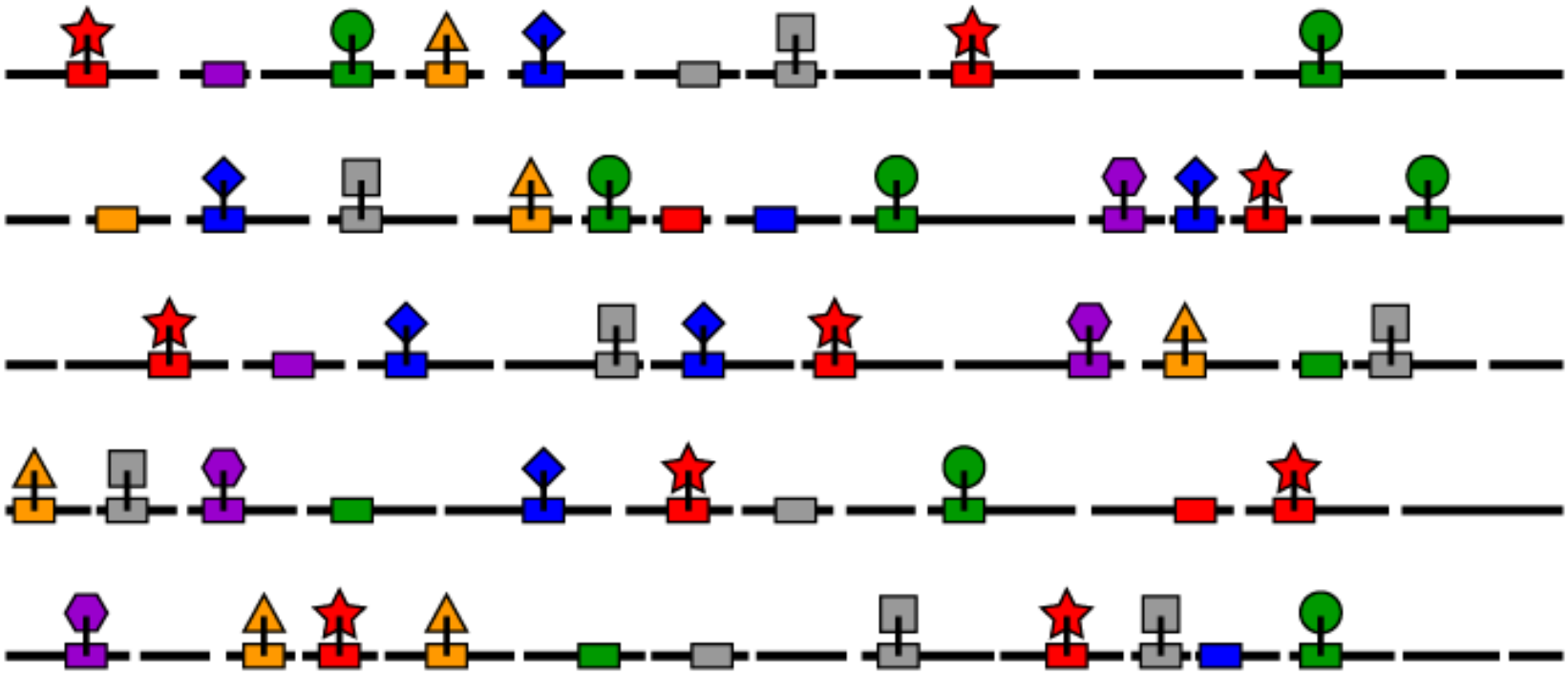
Chromatin ImmunoPrecipitation (ChIP)



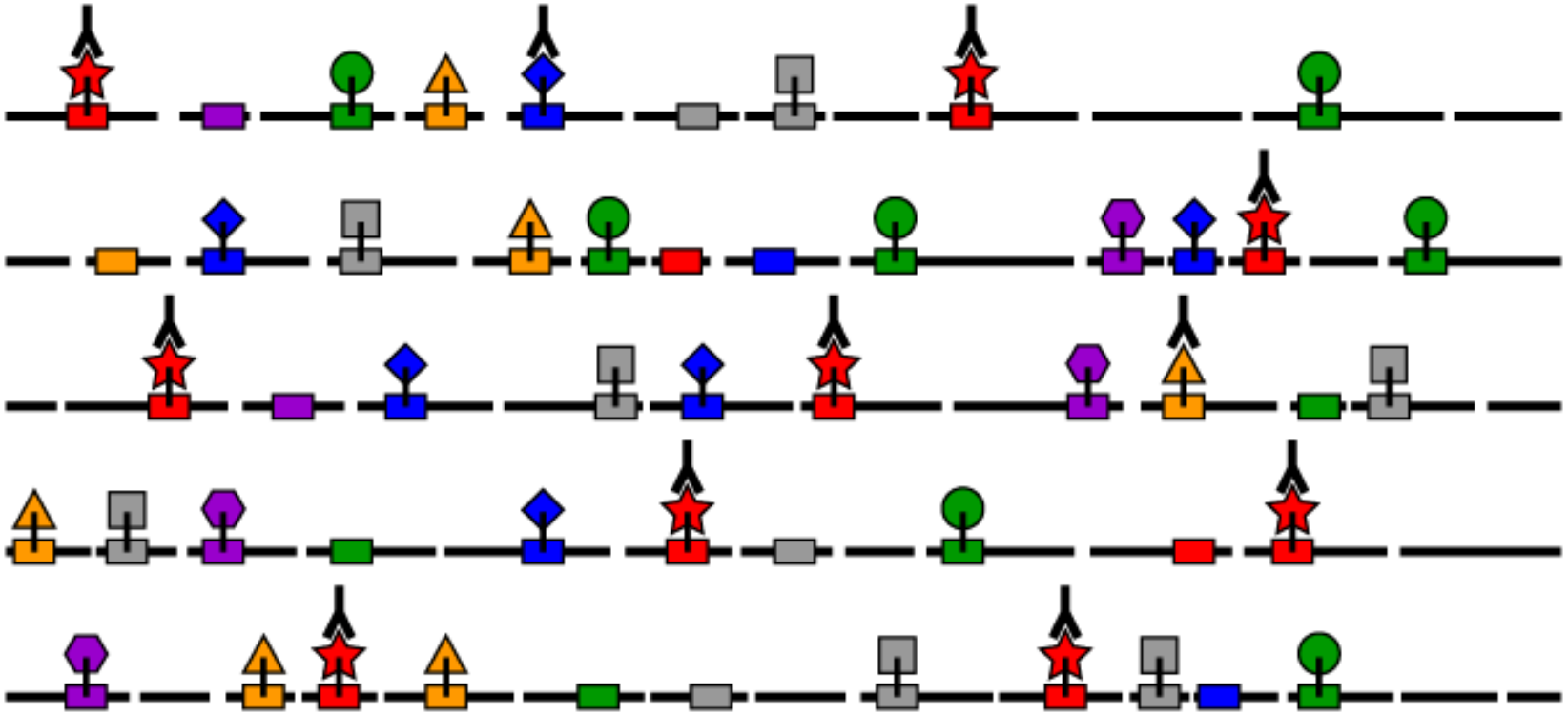
Protein-DNA crosslinking *in vivo* (for TF)



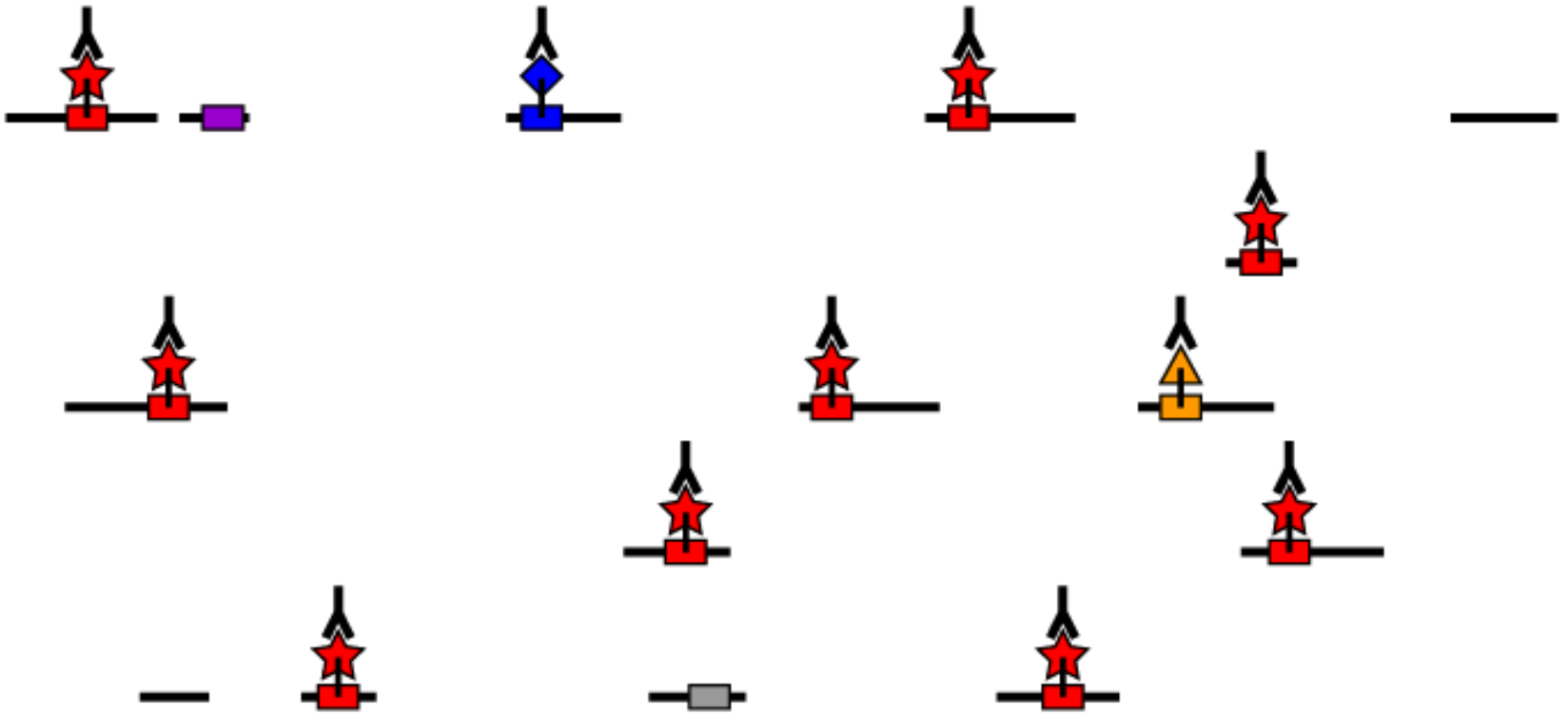
Chop the chromatin using sonication (TF) or micrococal nuclease (MNase) digestion (histone)



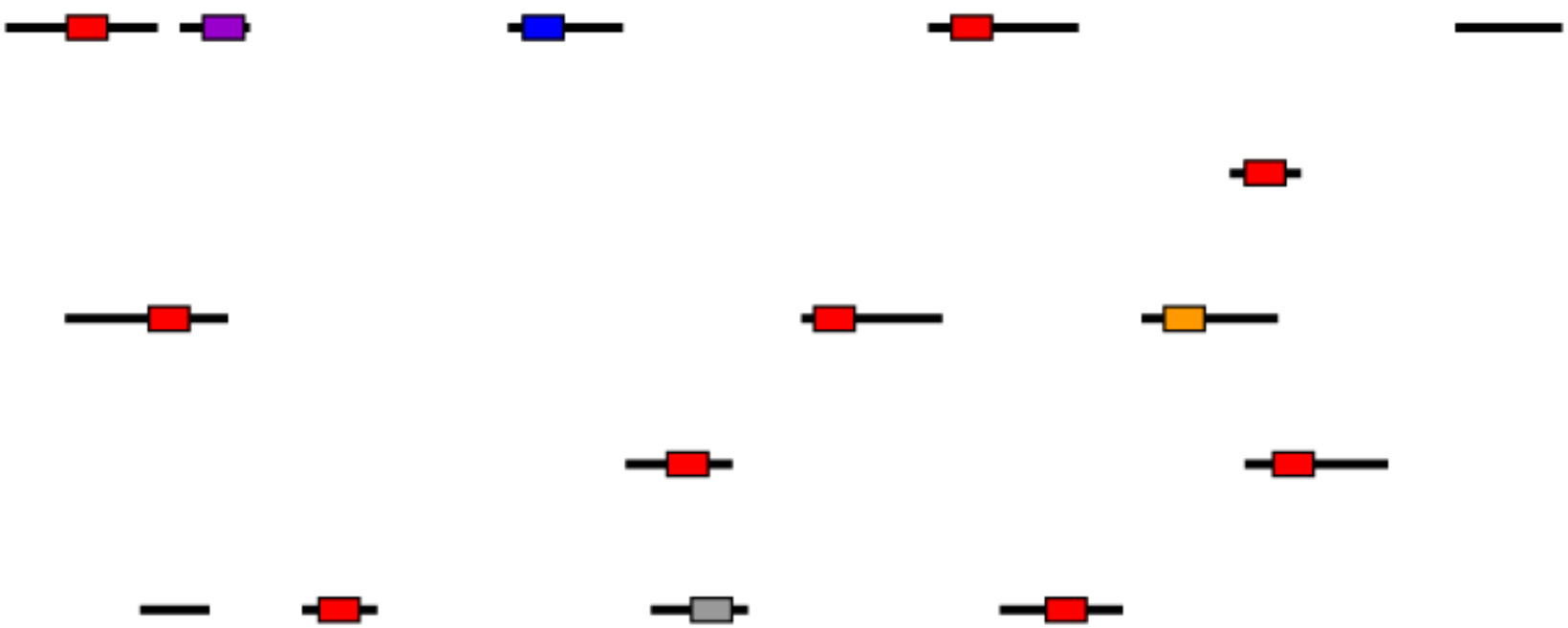
Specific factor-targeting antibody



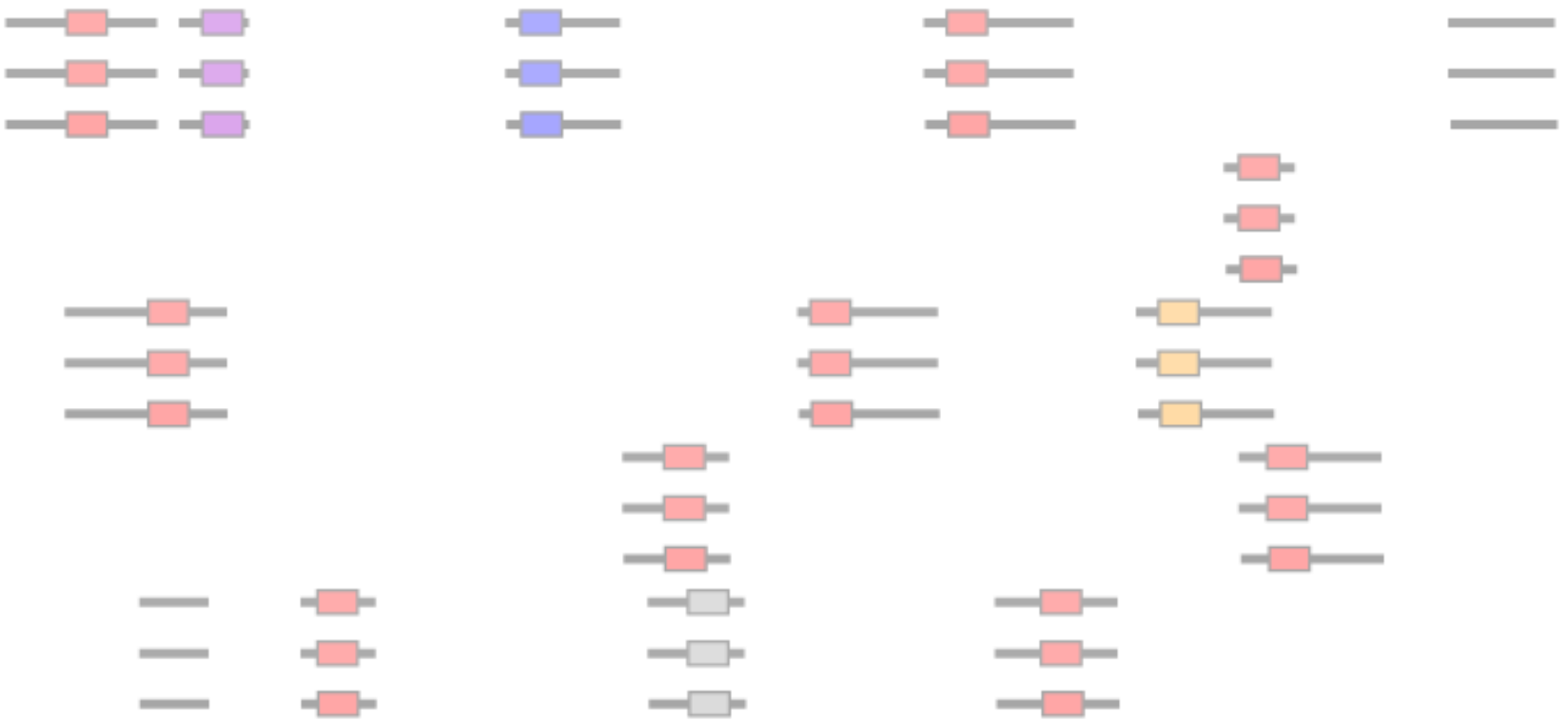
Immunoprecipitation



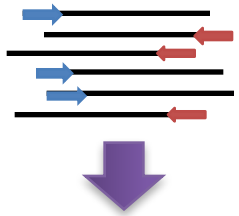
DNA purification



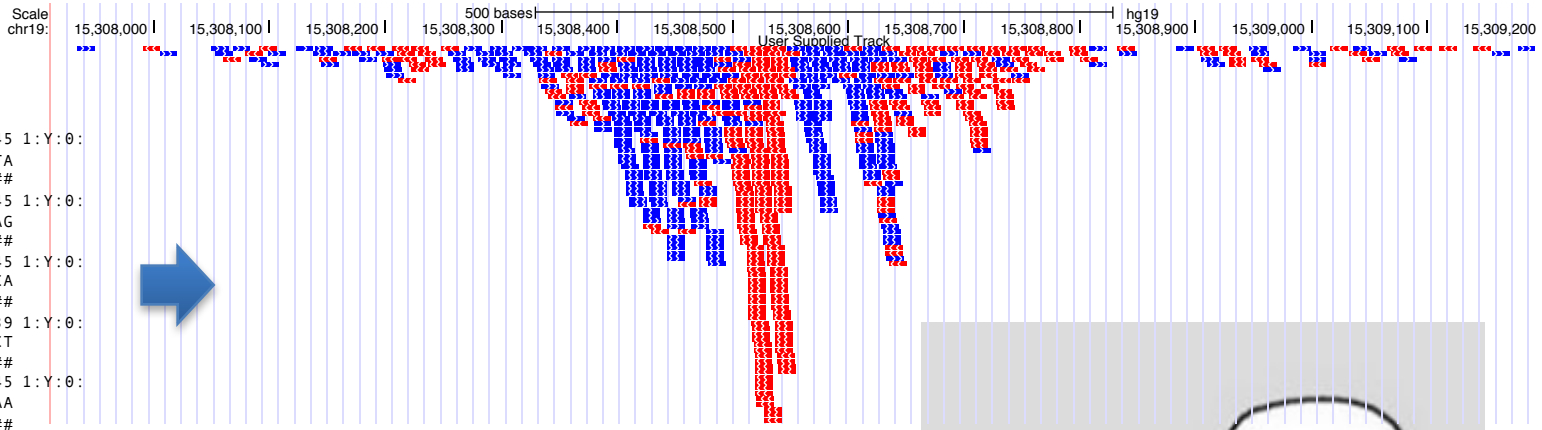
PCR amplification and sequencing

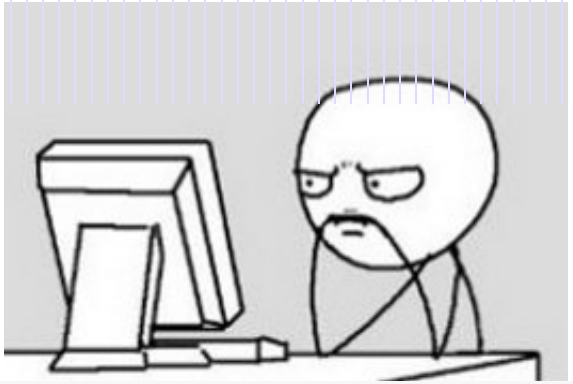


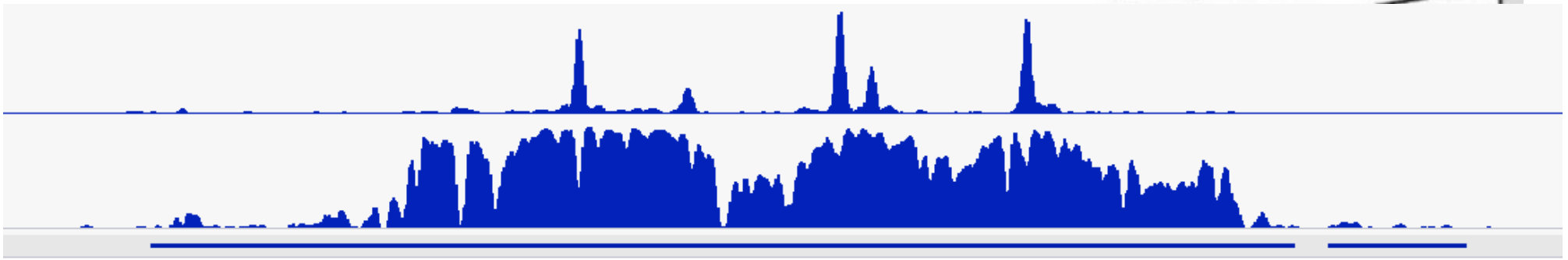
ChIP-seq data analysis overview



```
@ILLUMINA-8879DC:231:KK:3:1:1070:945 1:Y:0:
NNNAATACAGTCAGAAACATATCATATTGGAGAATA
####################################
@ILLUMINA-8879DC:231:KK:3:1:1153:945 1:Y:0:
NNNAAGCACACAGAAGATAACTAAACAATCAAGTAG
####################################
@ILLUMINA-8879DC:231:KK:3:1:1222:945 1:Y:0:
NNNAAGGGTCTTGAGAAGAAATCATTCTGGATGGCA
####################################
@ILLUMINA-8879DC:231:KK:3:1:1304:939 1:Y:0:
NNNCCAGGCTCCCGCGATTCTCCTGCCTCAGCTTCT
####################################
@ILLUMINA-8879DC:231:KK:3:1:1354:945 1:Y:0:
NNNCTCTTCCTTAGCTAAACTTTCAACTAAGCCAAA
####################################
@ILLUMINA-8879DC:231:KK:3:1:1411:932 1:Y:0:
NNNGTAGGACCATTGGCGTTGCGACACAAAAAATTT
####################################
@ILLUMINA-8879DC:231:KK:3:1:1496:937 1:Y:0:
NNNTTCATCGGGTTGAGAGTCCCCTTGTTGCATGCA
####################################
@ILLUMINA-8879DC:231:KK:3:1:1533:939 1:Y:0:
NNNATTTTCCCGTTCCAGGTCGCAATTTCCGCCGTT
####################################
@ILLUMINA-8879DC:231:KK:3:1:1573:940 1:Y:0:
NNNGGGGTGCGCCTTTAGTCCCAGCTACTCAGGAAC
####################################
```

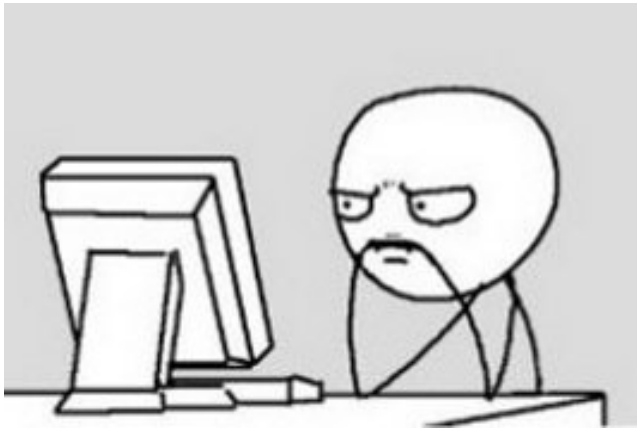






ChIP-seq data analysis overview

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

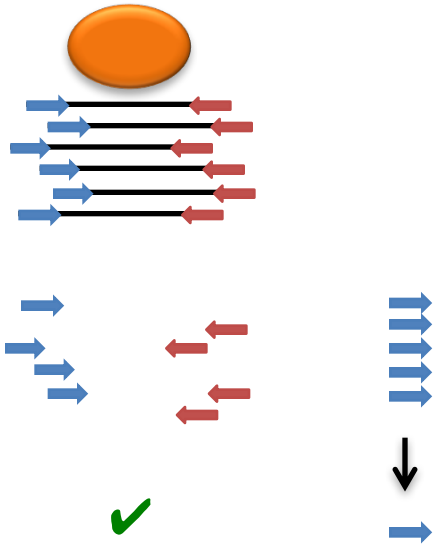


ChIP-seq data analysis: basic processing

- 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:

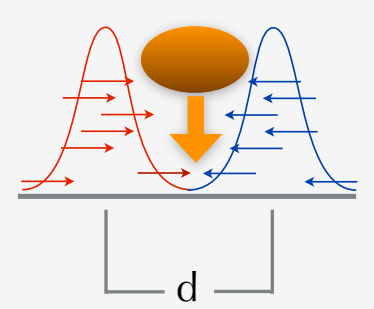


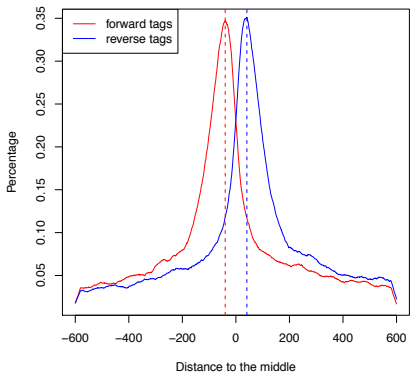
Langmead et al. 2009,
Zang et al. 2009

ChIP-seq data analysis: Peak calling

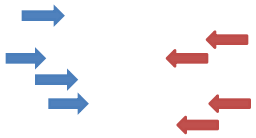
- DNA fragment size estimation

peak model

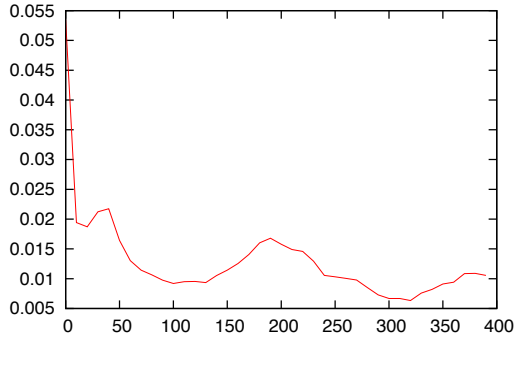




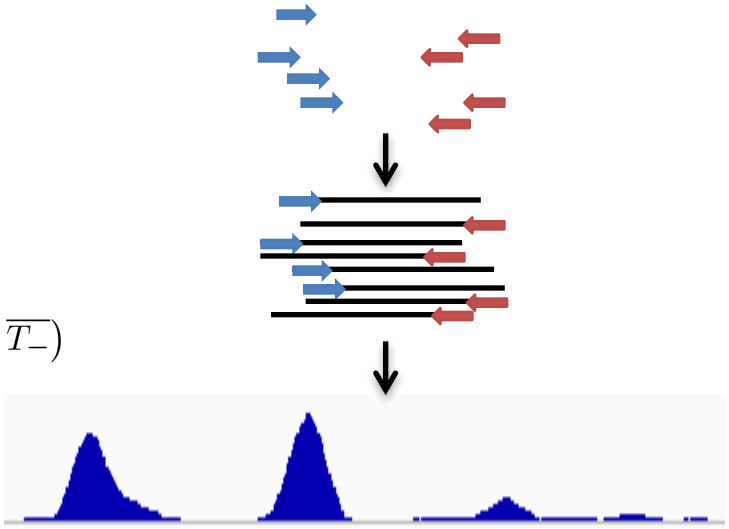
cross-correlation



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



- pile-up profiling



- Peak/signal detection

ChIP-seq data analysis: Peak calling

- **Sharp peaks**

  transcription factor binding, DNase, ATAC-seq

  MACS (Zhang, 2008)

  dynamic background

  Poisson model

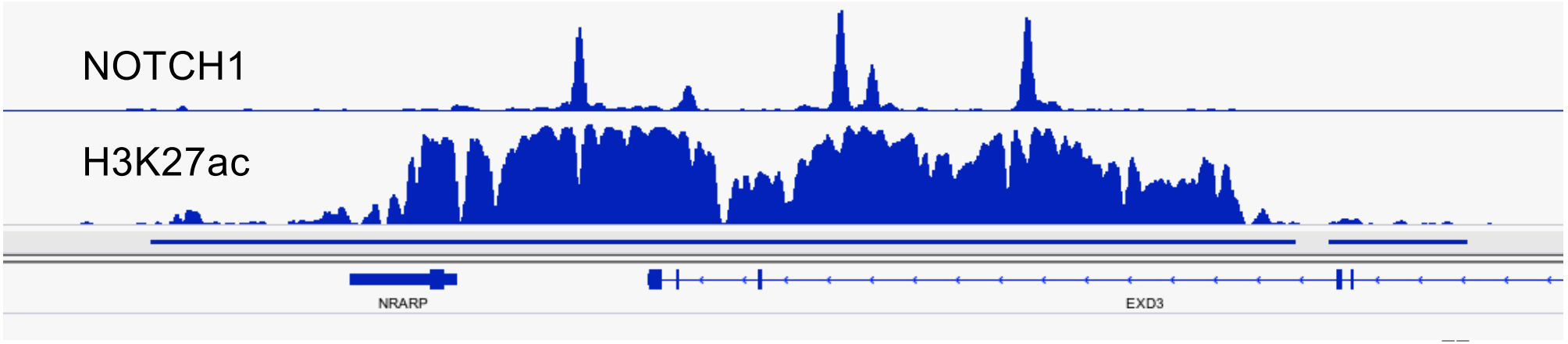
- **Broad peaks**

  Histone modifications, “super-enhancers”

  Diffuse

  SICER (Zang, 2009)

  Spatial clustering of localized weak signal and integrative Poisson model

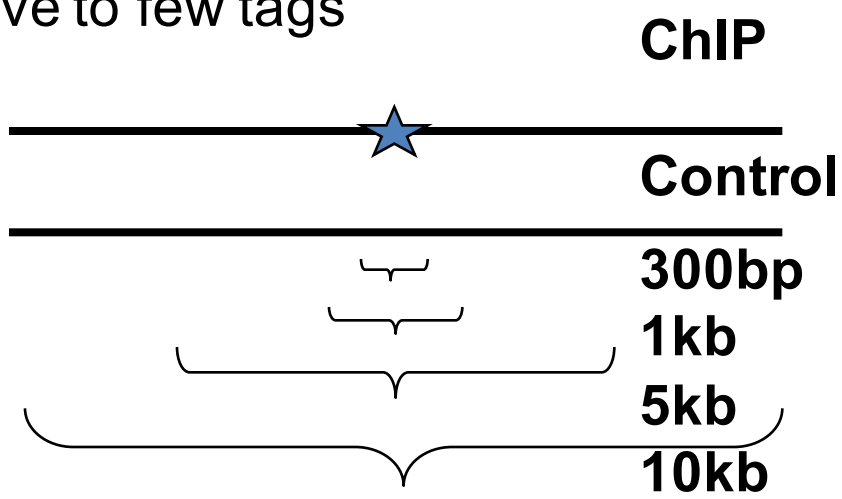


Wang, Zang et al. 2014

MACS

- **M**odel-based **A**nalysis for **C**hIP-**S**eq
- Tag distribution along the genome ~ Poisson distribution (λ_{BG} = total tag / genome size)
- ChIP-seq show local biases in the genome
  - Chromatin and sequencing bias
  - 200-300bp control windows have to few tags
  - But can look further

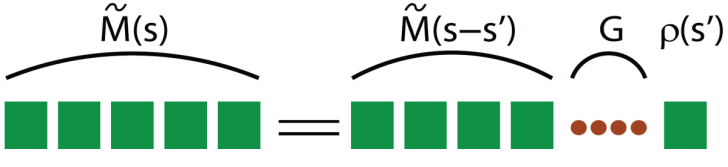
Dynamic λ_{local} =
max(λ_{BG}, [λ_{ctrl}, λ_{1k},] λ_{5k}, λ_{10k})



http://liulab.dfci.harvard.edu/MACS/
Zhang et al, *Genome Bio*, 2008

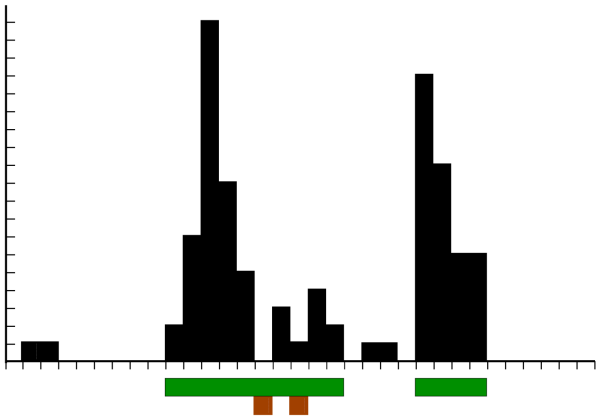
SICER

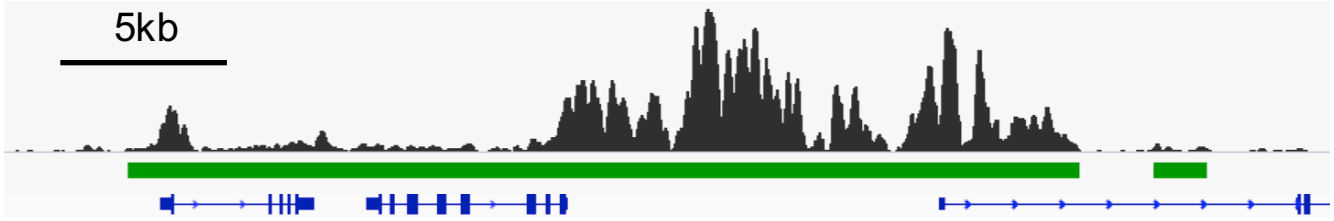
- **S**patial-clustering **I**dentification of **C**hIP-**E**nriched **R**egions

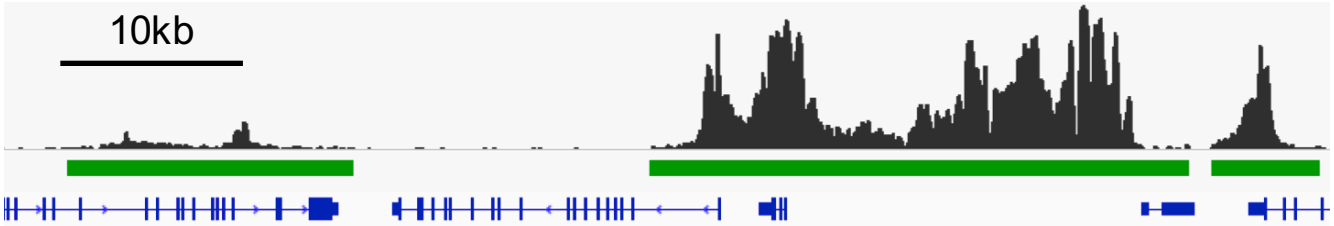


$$\tilde{M}(s) = G(\lambda, l_0, g) \int_{s_0}^{s} \mathrm{d}s' \tilde{M}(s - s')\rho(s')$$

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







omictools.com

Zang et al. *Bioinformatics* 2009

ChIP-seq peak calling: Parameters

Parameter	Remarks
Genome	Species and reference genome version, e.g. hg38, hg18, mm10, mm9
Effective genome rate	Fraction of the mappable genome, vary in species, read length, etc.
DNA fragment size	Estimated by default; can specify otherwise
Window size	Data resolution, usually nucleosome periodicity length, i.e. 200bp
Gap size	(for SICER only) Allowable gaps between eligible windows, usually 2 or 3 windows
P-value cut-off	Threshold for peak calling, from model
False discovery rate (FDR) cut-off	Threshold for peak calling, BH correction from p-value.

ChIP-seq data analysis: Review

1. Read mapping (sequence alignment)

2. Peak calling: ***MACS*** or ***SICER***
    1. QC
    2. DNA fragment size estimation (for Single-end)
    3. Pile-up profile generation
    4. Peak/signal detection

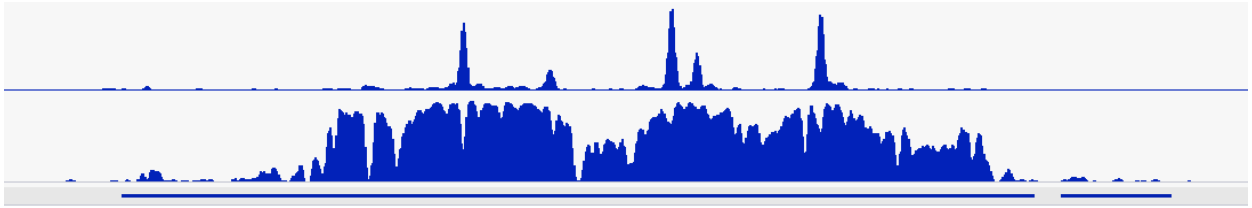
3. Downstream analysis/integration

Data formats

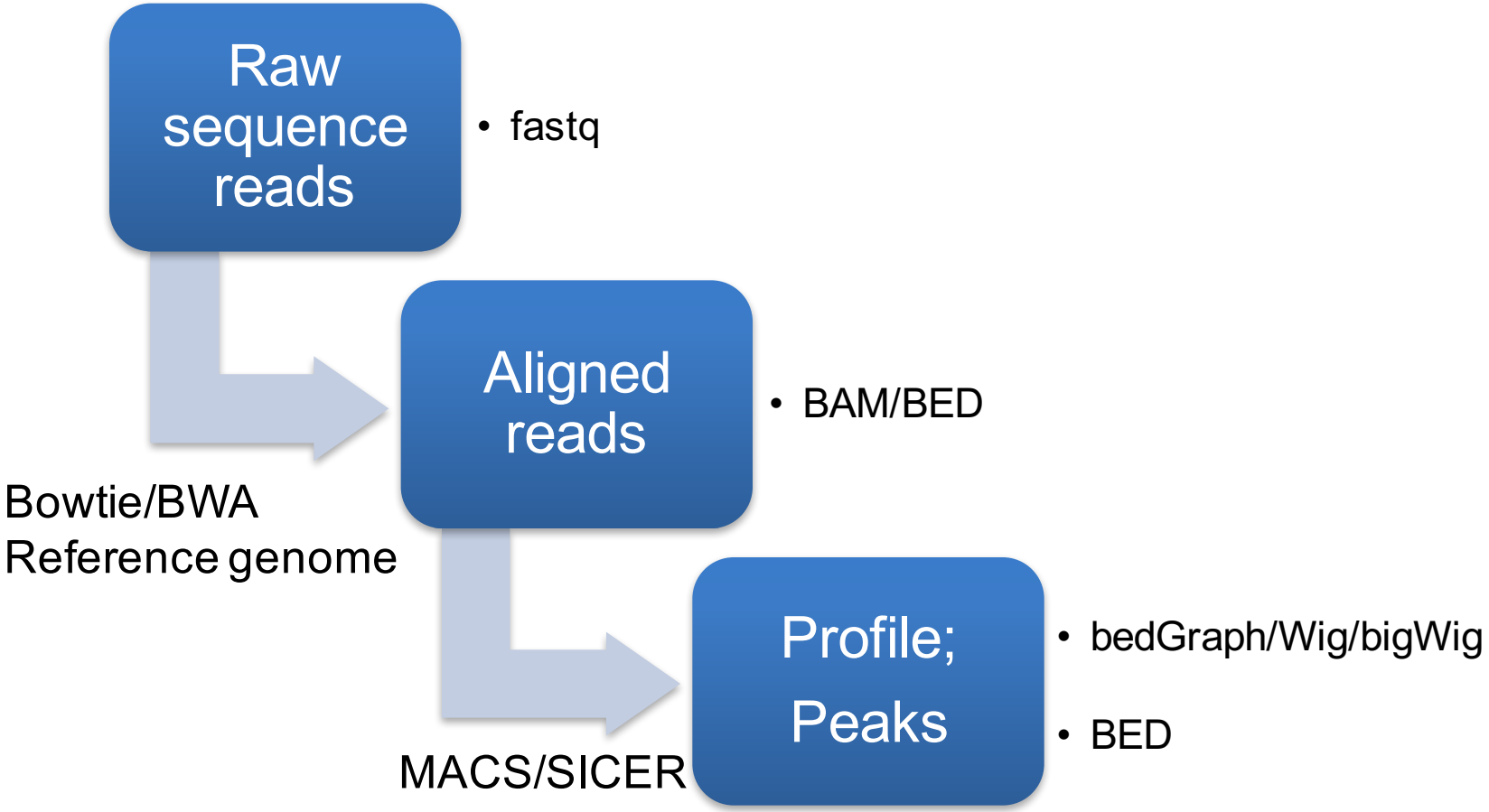
- fastq: raw sequences
- BED:

chr11	10344210 10344260 255	0	-
chr4	76649430 76649480 255	0	+
chr3	77858754 77858804 255	0	+
chr16	62688333 62688383 255	0	+
chr22	33031123 33031173 255	0	-

- SAM/BAM: aligned sequencing reads
- bedGraph, Wig, bigWig: pile-up profiles for browser visualization

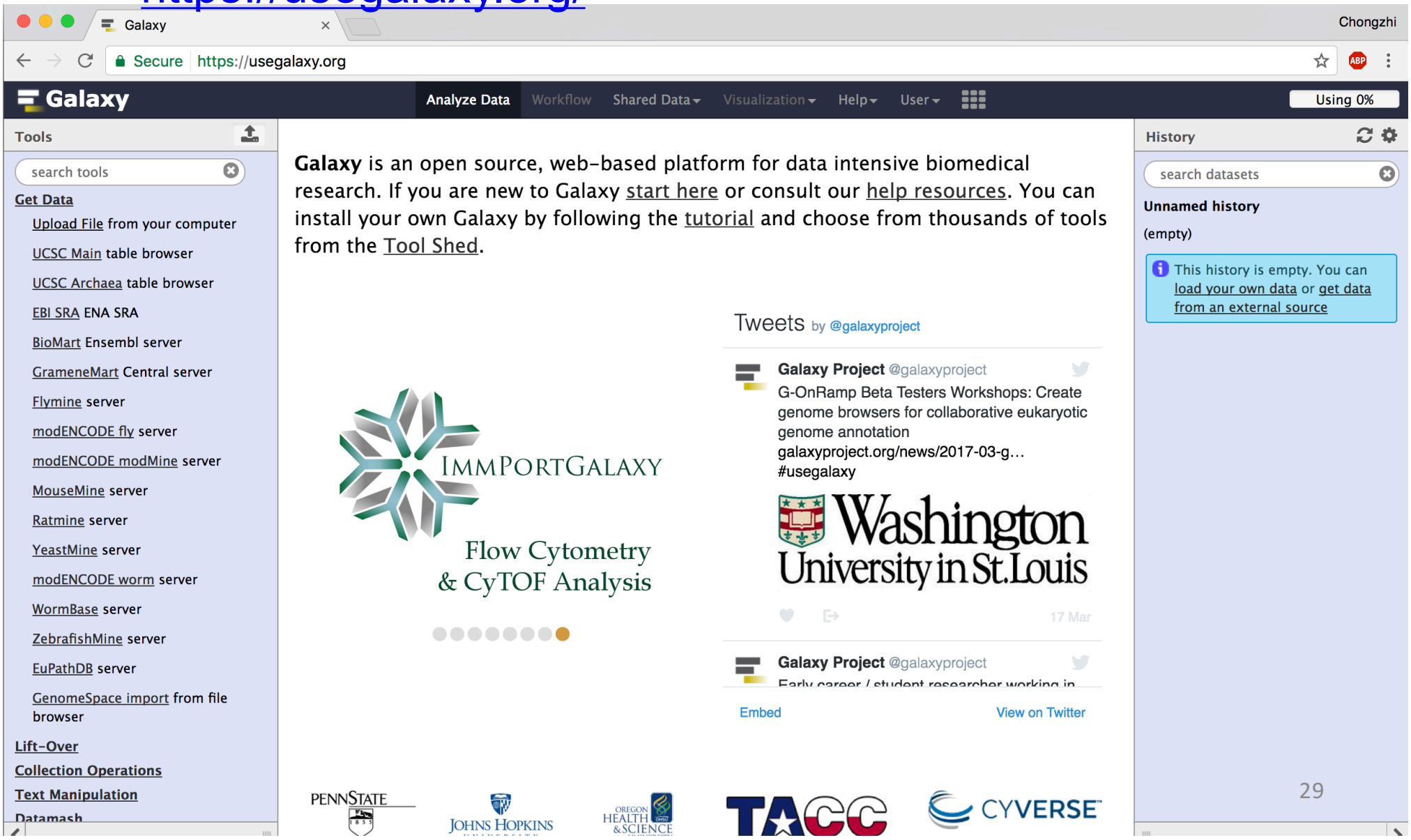


Data flow


Raw sequence reads
• fastq
Aligned reads
• BAM/BED
Bowtie/BWA
Reference genome
Profile; Peaks
• bedGraph/Wig/bigWig
• BED
MACS/SICER

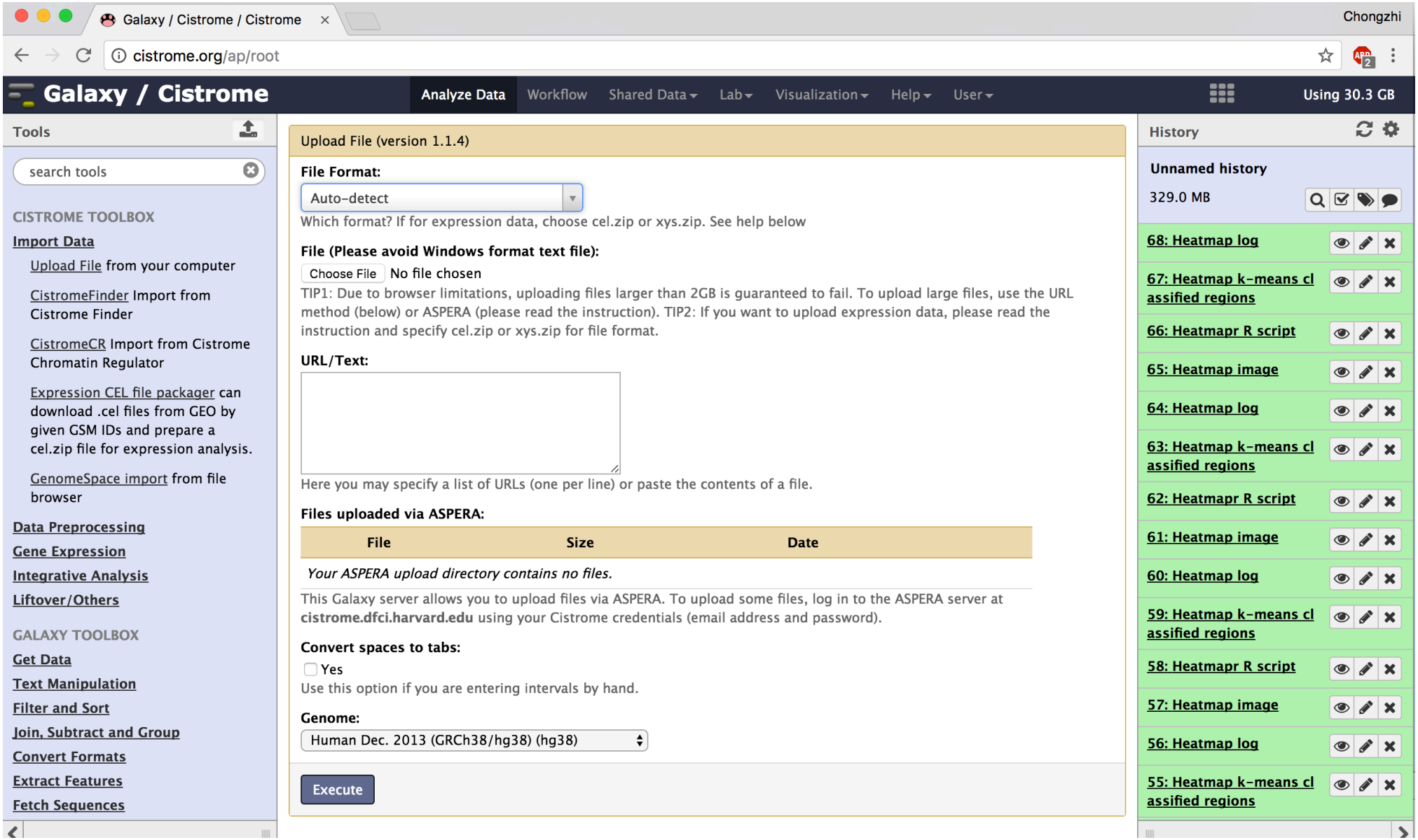
Galaxy: web-interface analysis platform

- https://usegalaxy.org/



Run MACS on Cistrome, a Galaxy-based platform

- http://cistrome.org/ap/



Run SICER on Galaxy-based platforms

- http://services.cbib.u-bordeaux.fr/galaxy/

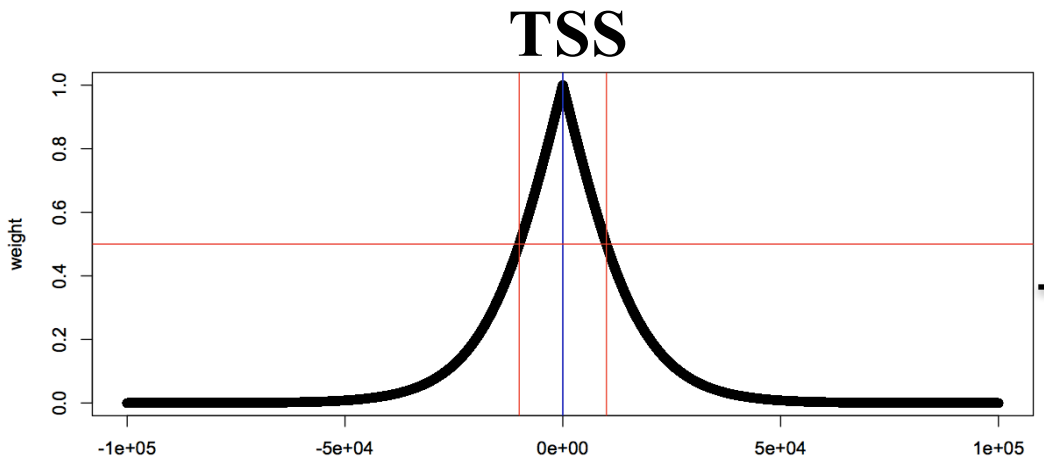
ChIP-seq: Downstream analysis

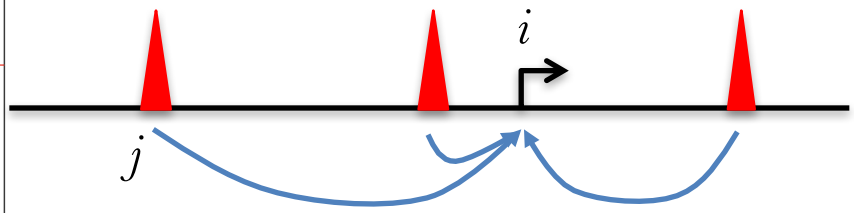
- Data visualization
  - UCSC genome browser: http://genome.ucsc.edu/
  - WashU epigenome browser: http://epigenomegateway.wustl.edu/
  - IGV: http://software.broadinstitute.org/software/igv/
- Meta analysis
  - CEAS: http://liulab.dfci.harvard.edu/CEAS/
- Integration with gene expression
  - BETA: http://cistrome.org/BETA/
  - MARGE: http://cistrome.org/MARGE/
- Integration with other epigenomic data
  - GREAT: http://great.stanford.edu
  - ENCODE SCREEN: http://screen.umassmed.edu/
  - MANCIE: https://cran.r-project.org/package=MANCIE
  - Cistrome DB: http://cistrome.org/db/

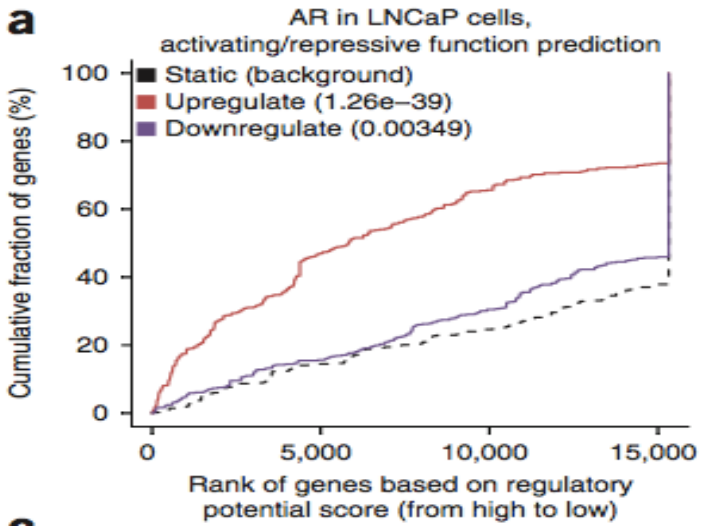
BETA: Binding Expression Target Analysis

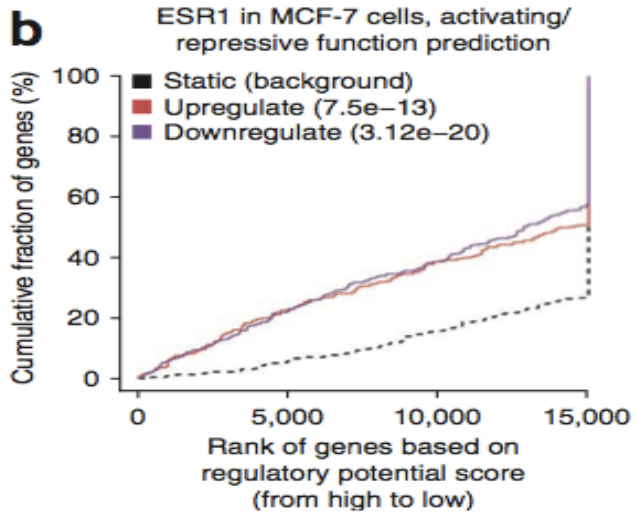
- Regulatory Potential

$$P(g_i) = \sum_{j \in S(i)} \exp\left(-\frac{\Delta_{ij}}{\lambda}\right)$$

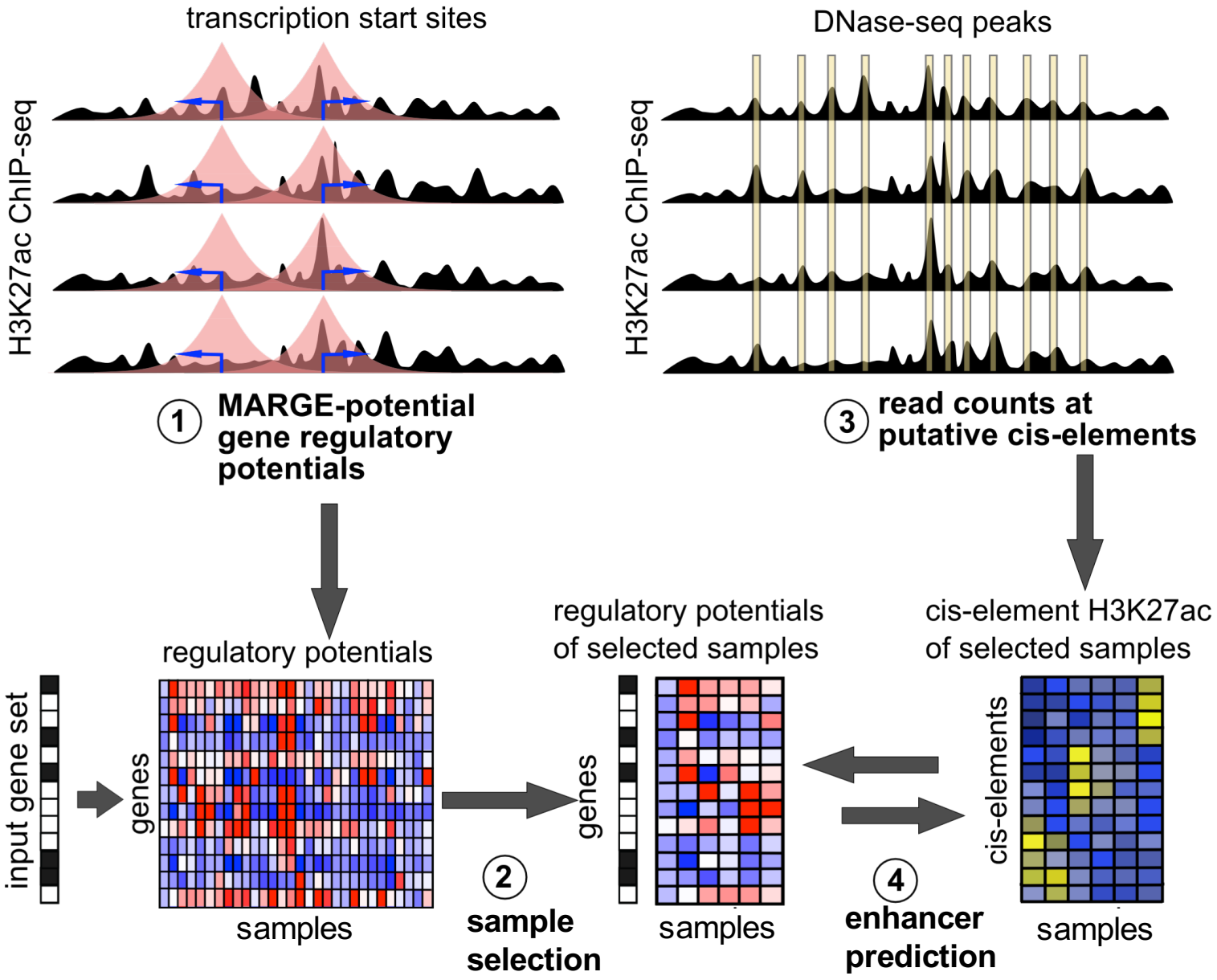








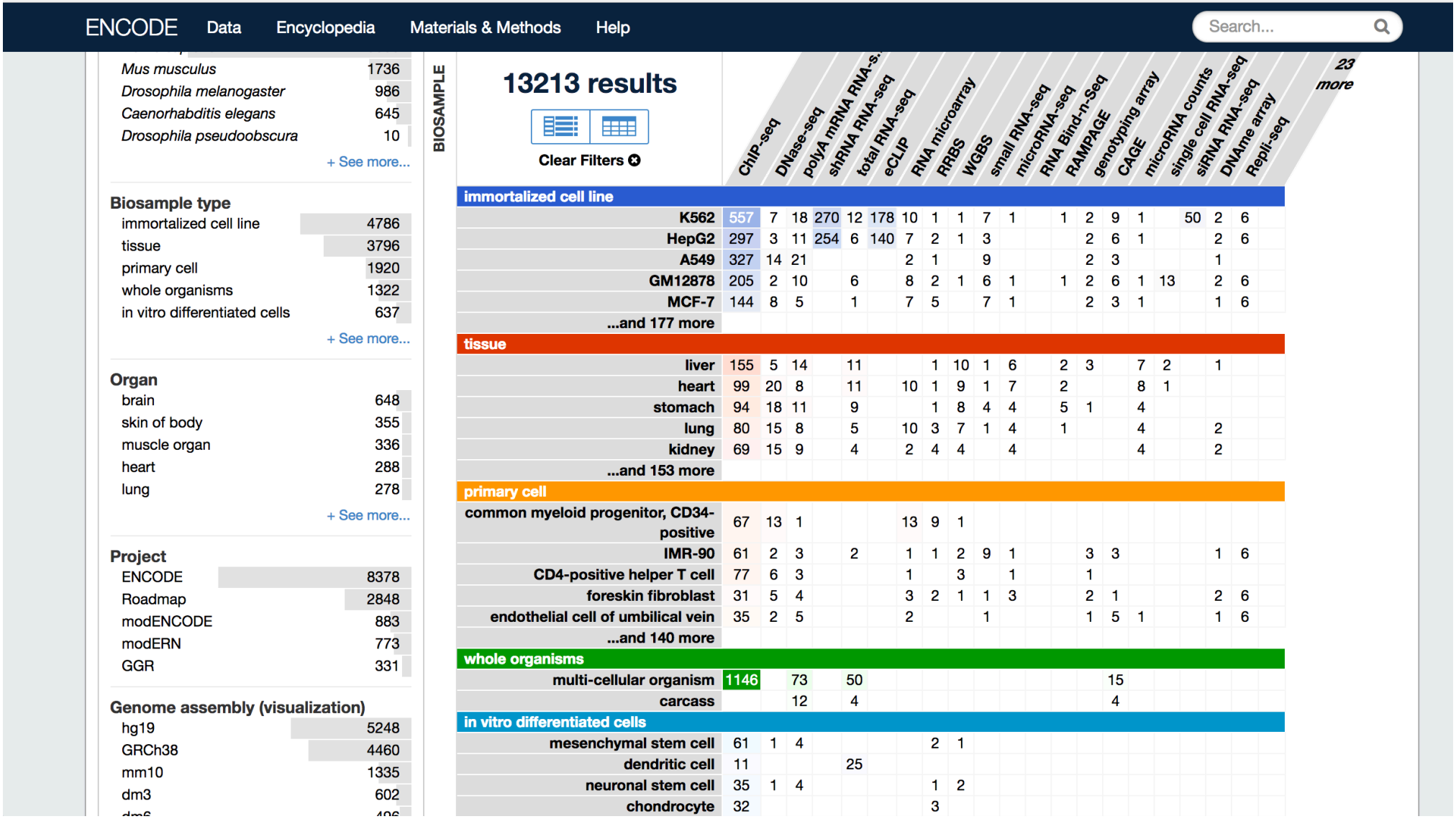
MARGE: A big data driven, integrative regression and semi-supervised approach for predicting functional enhancers



Wang, Zang et al. *Genome Res 2016*

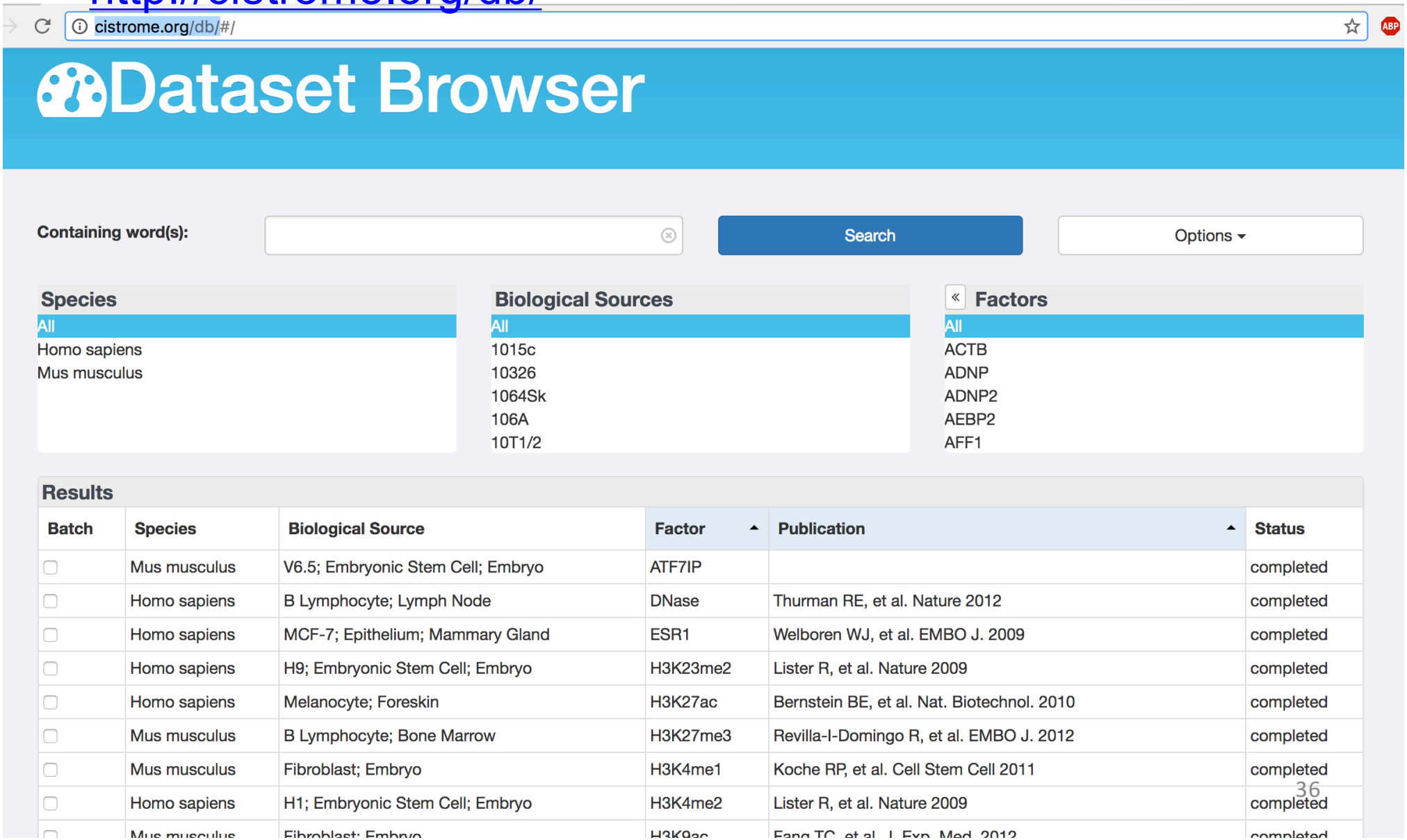
ENCODE

https://www.encodeproject.org/



Cistrome Data Browser

http://cistrome.org/db/



ChIP-seq data analysis: Review

1. Read mapping (sequence alignment)

2. Peak calling: ***MACS*** or ***SICER***
    1. QC
    2. DNA fragment size estimation (for Single-end)
    3. Pile-up profile generation
    4. Peak/signal detection

3. Downstream analysis/integration

Summary

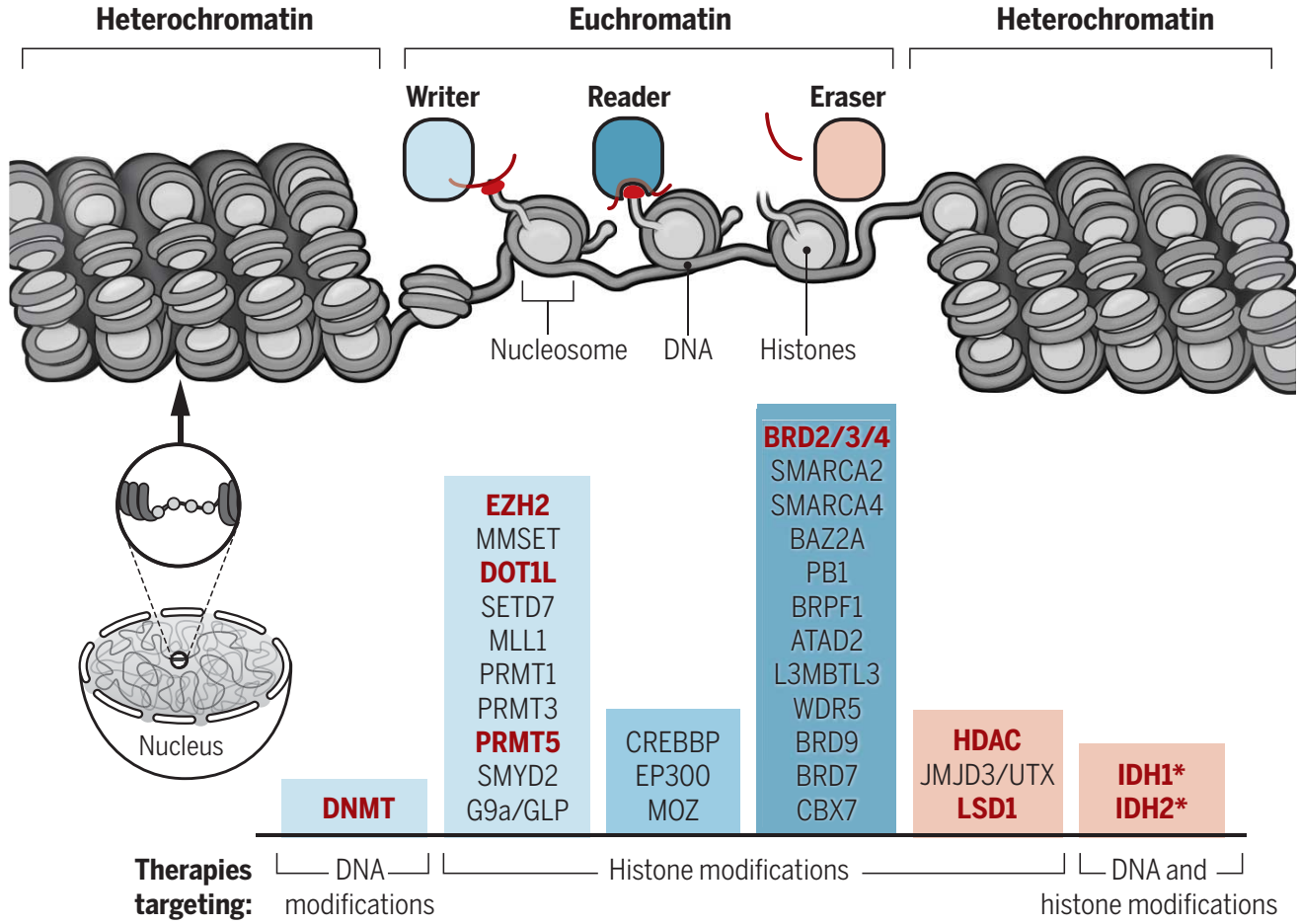
- ChIP-seq is used to profile epigenomes
- ChIP-seq data analysis
  - MACS for narrow peaks
  - SICER for broad peaks
- Online tools and resources

Further Reading

The cancer epigenome: Concepts, challenges, and therapeutic opportunities

Science 17 Mar 2017: Vol. 355, Issue 6330, pp.1147-1152

http://science.sciencemag.org/content/355/6330/1147



Thank you very much!

zang@virginia.edu
http://zanglab.com