

# The Epigenome Tools 2: ChIP-Seq and Data Analysis

Chongzhi Zang

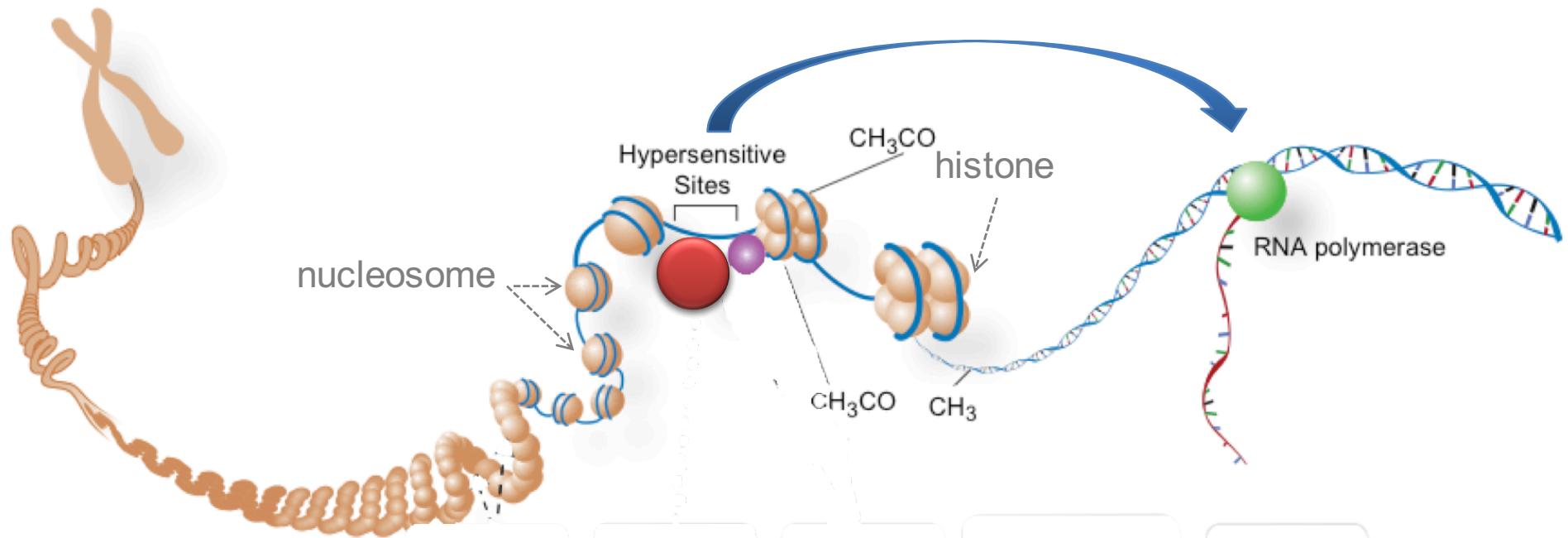
[zang@virginia.edu](mailto:zang@virginia.edu)  
<http://zanglab.com>

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

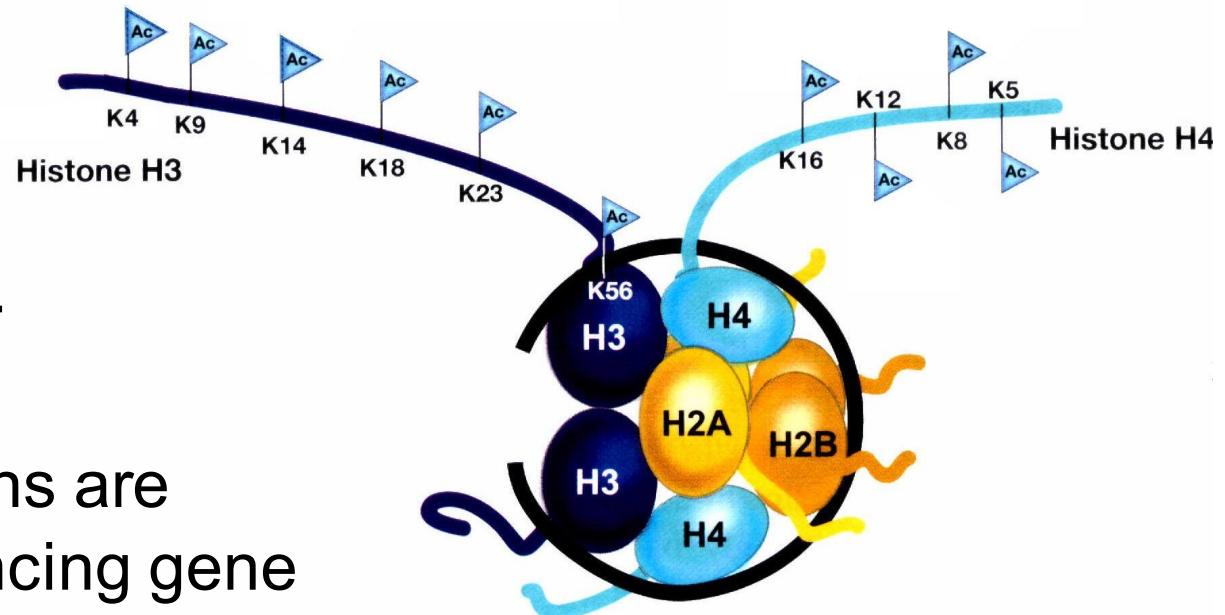
# Epigenomic marks

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

# Histone modifications

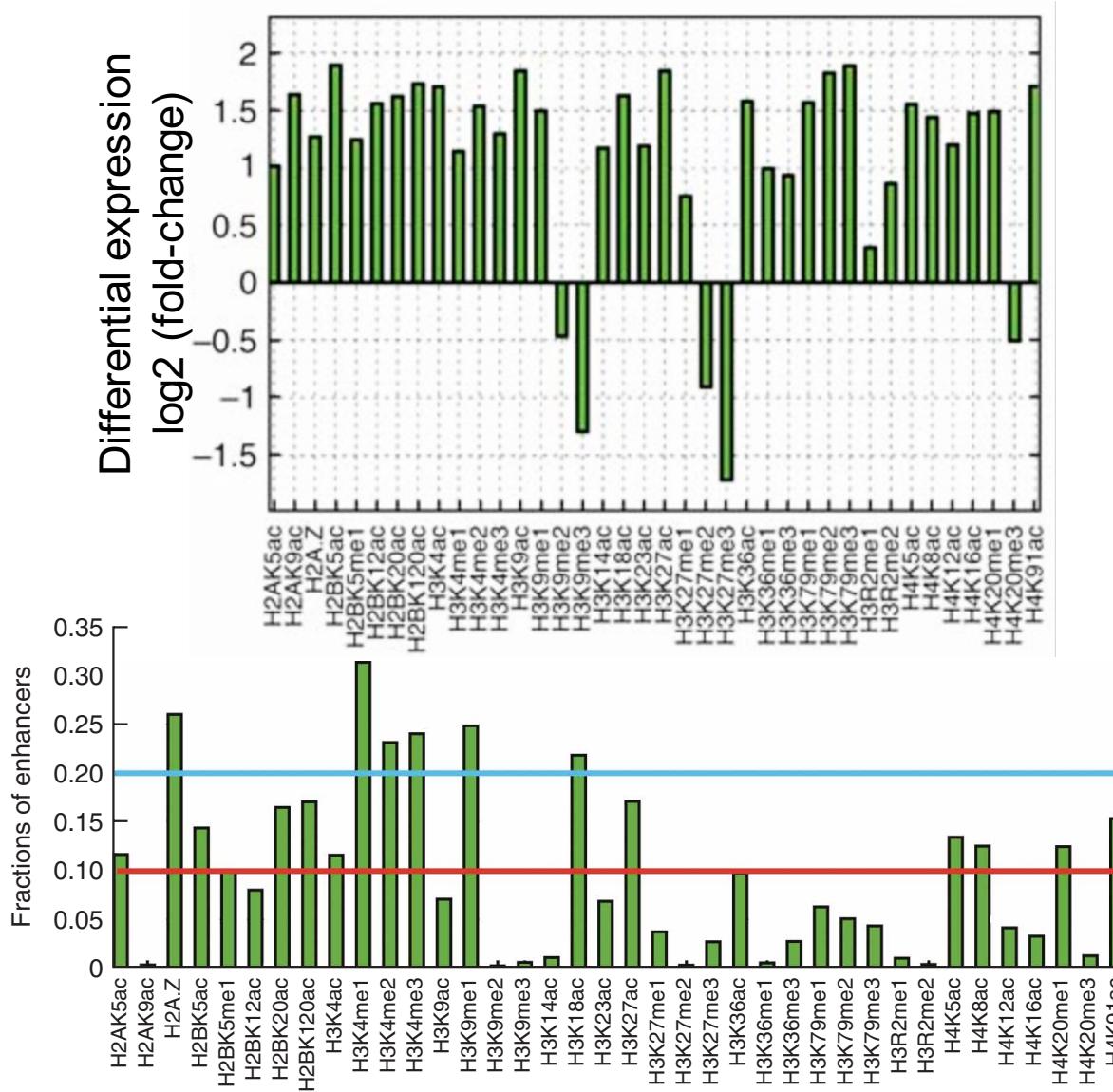
- Nucleosome Core Particles
- Core Histones: H2A, H2B, H3, H4
- Covalent modifications on histone tails include:
  - methylation (me),
  - acetylation (ac),
  - phosphorylation ...
- Histone variants
- Histone modifications are implicated in influencing gene expression.

Notation:  
H3K4me3



Allis C. et al. Epigenetics. 2006

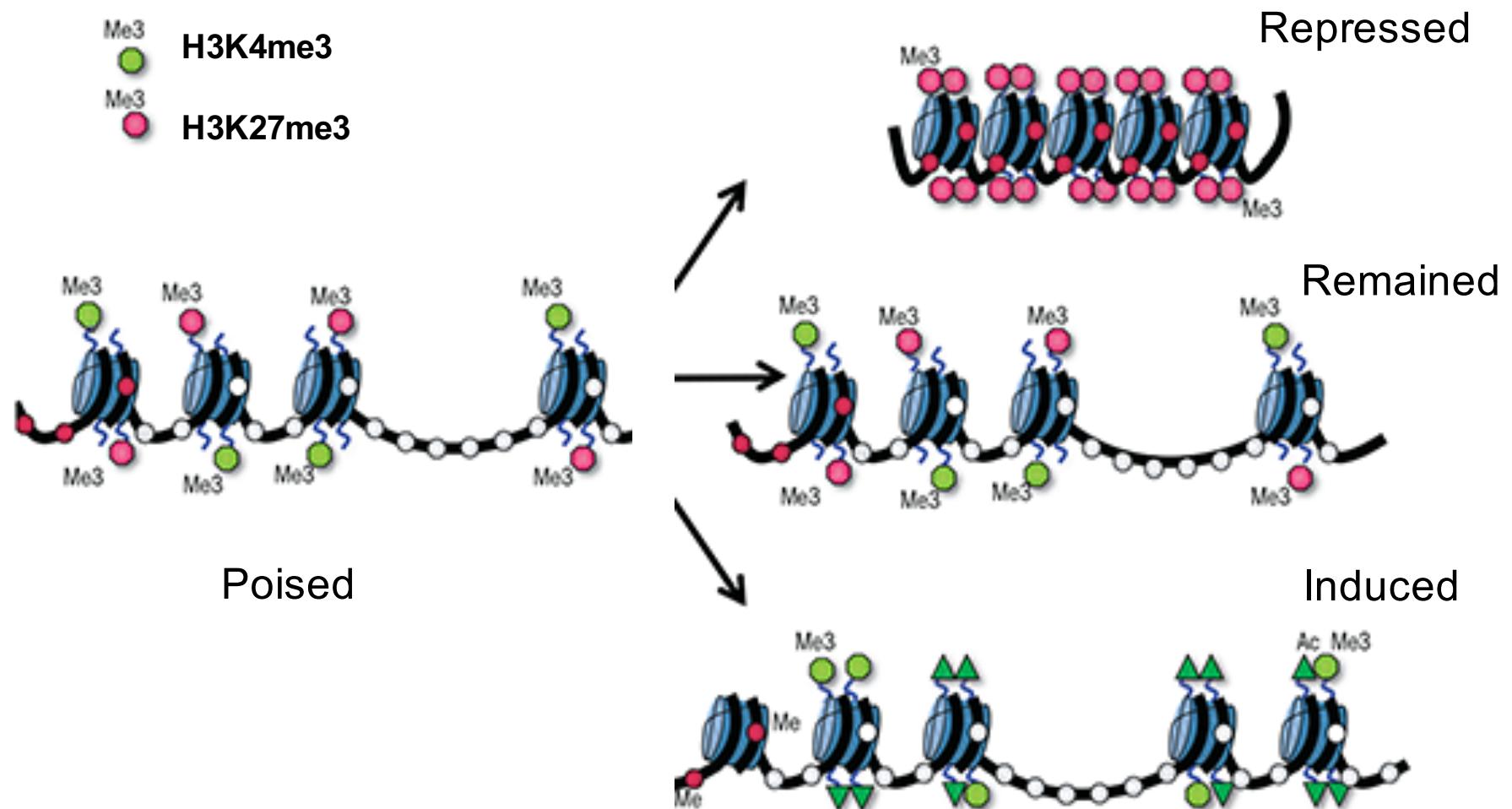
# Histone modifications associate with regulation of gene expression



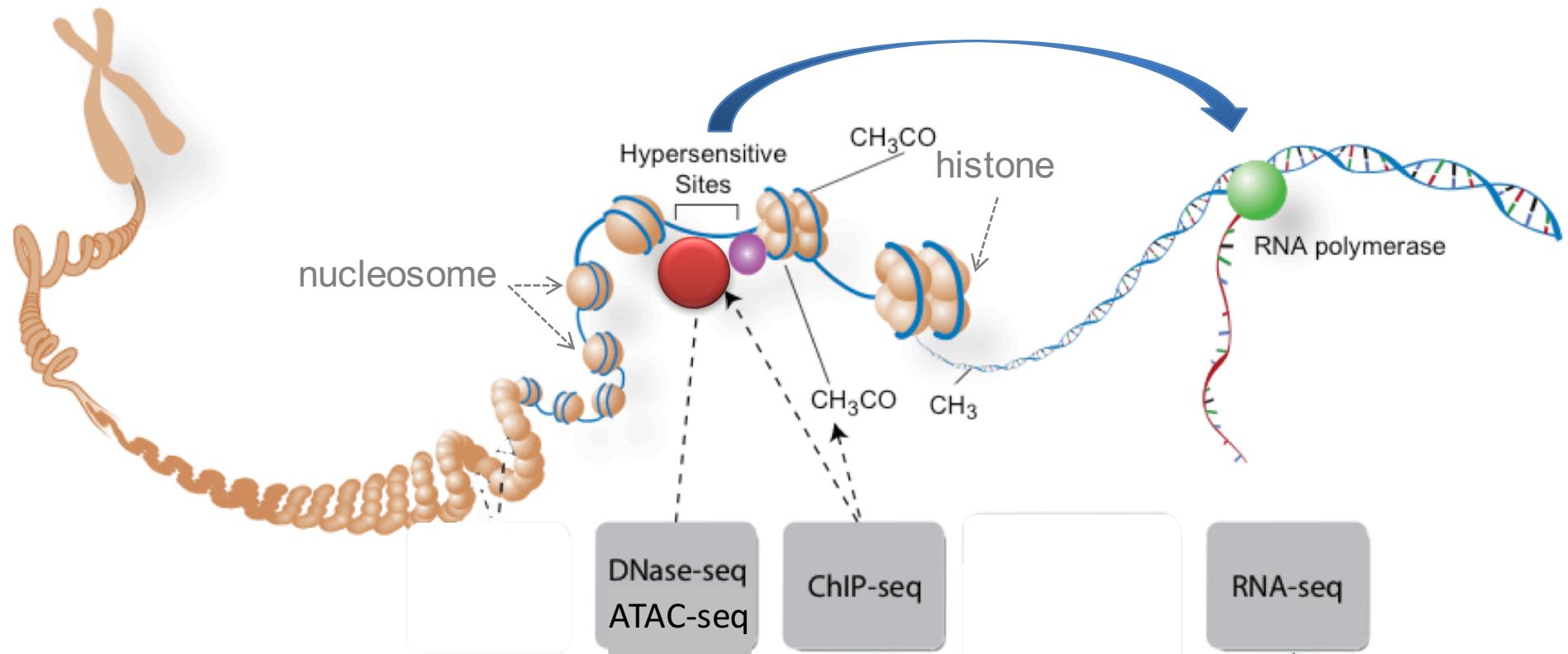
# “Functions” of histone marks

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

# H3K4me3/H3K27me3 Bivalent Domain

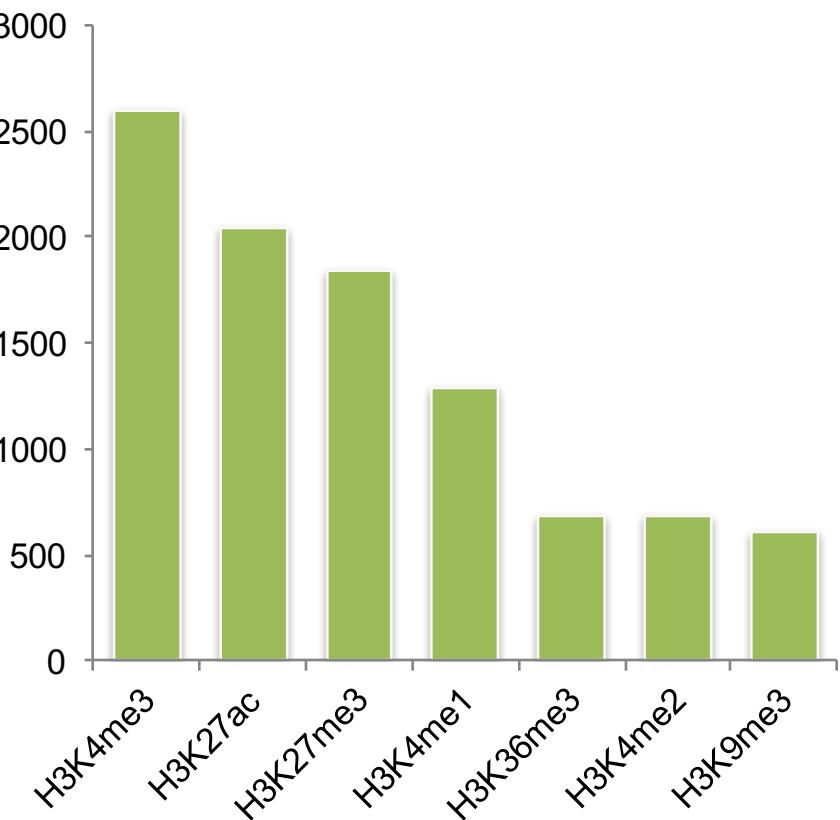
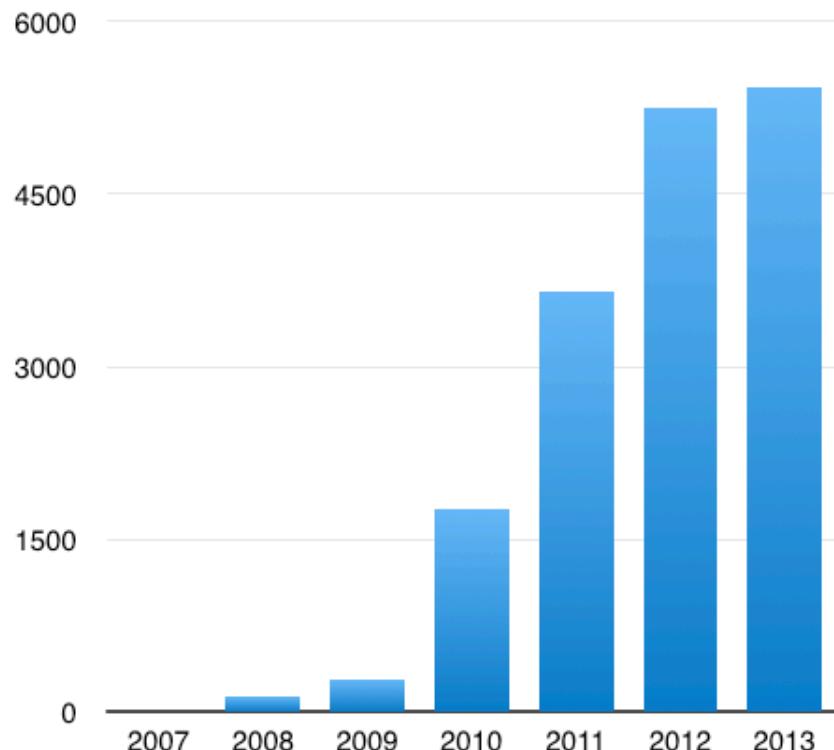


# ChIP-seq: Profiling epigenomes with sequencing

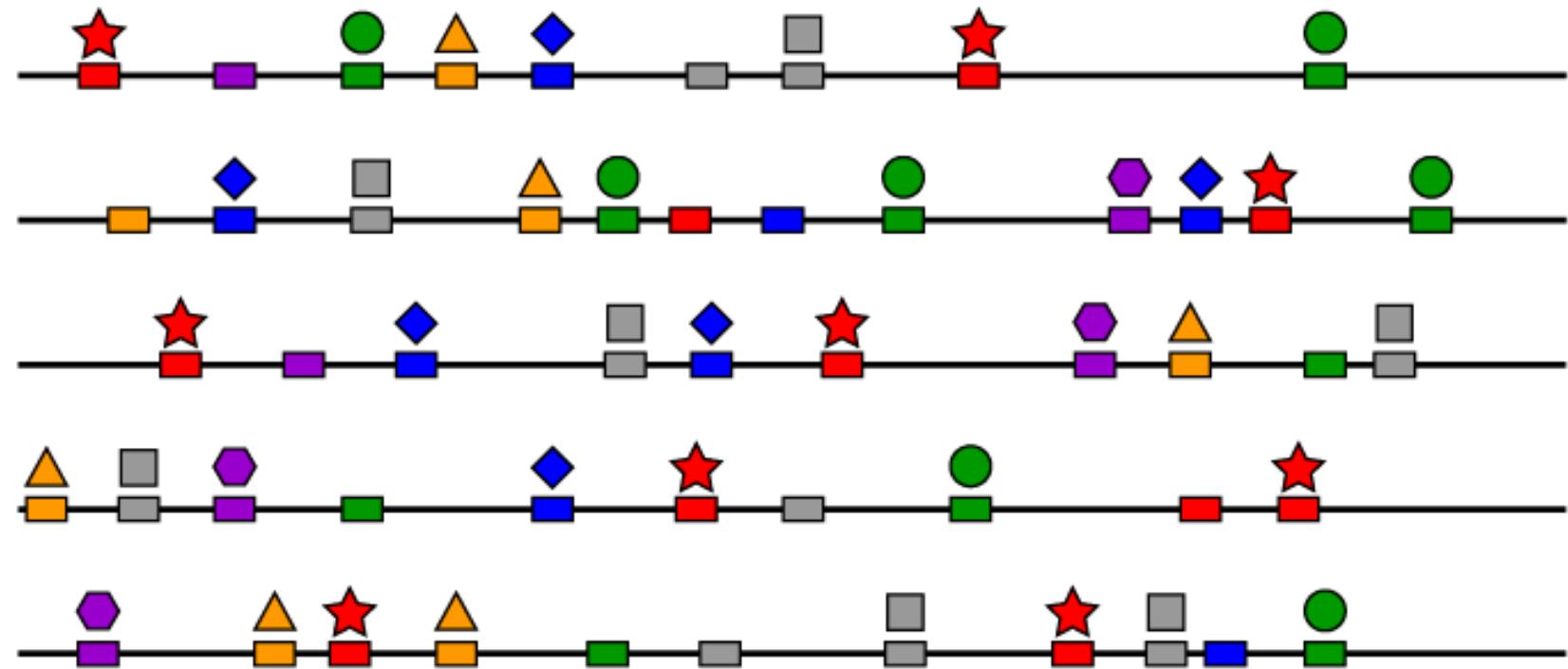


# Published ChIP-seq datasets are skyrocketing We are entering the Big Data era

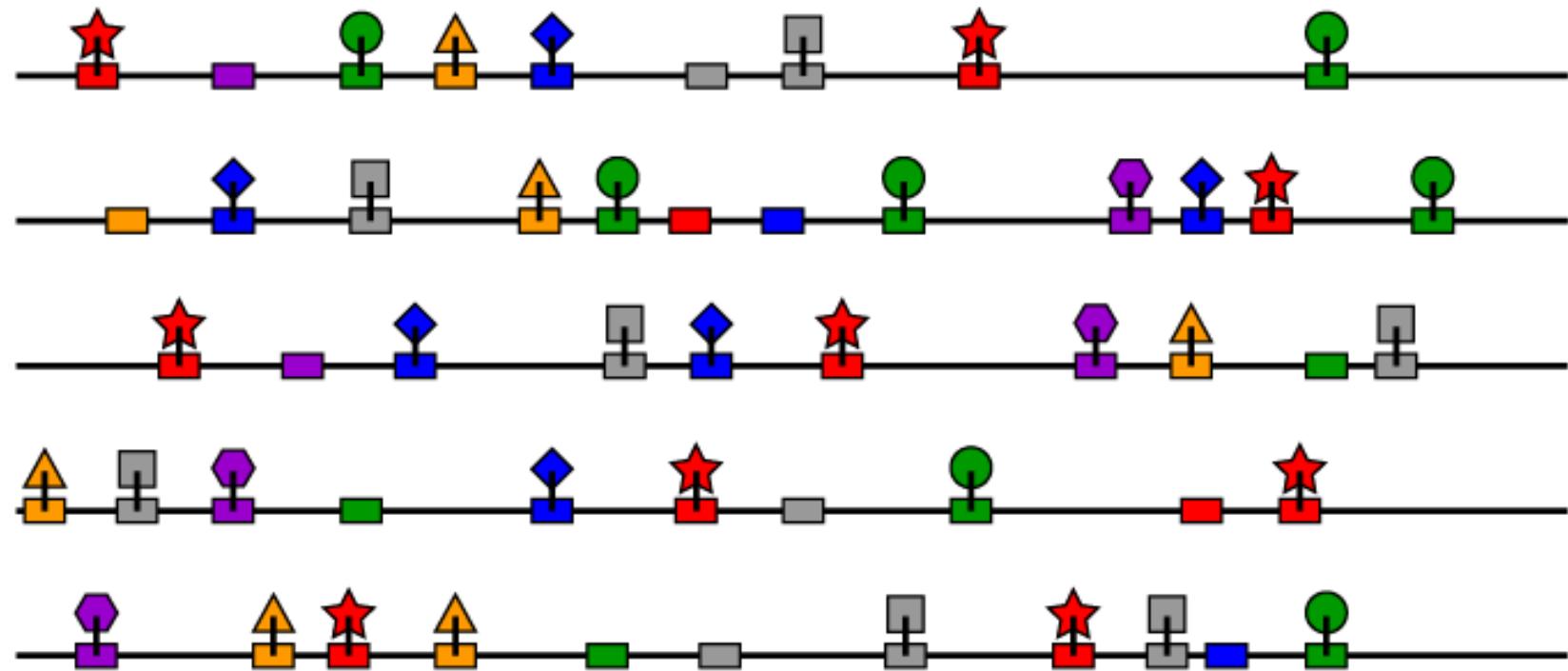
Number of ChIP-seq datasets on GEO



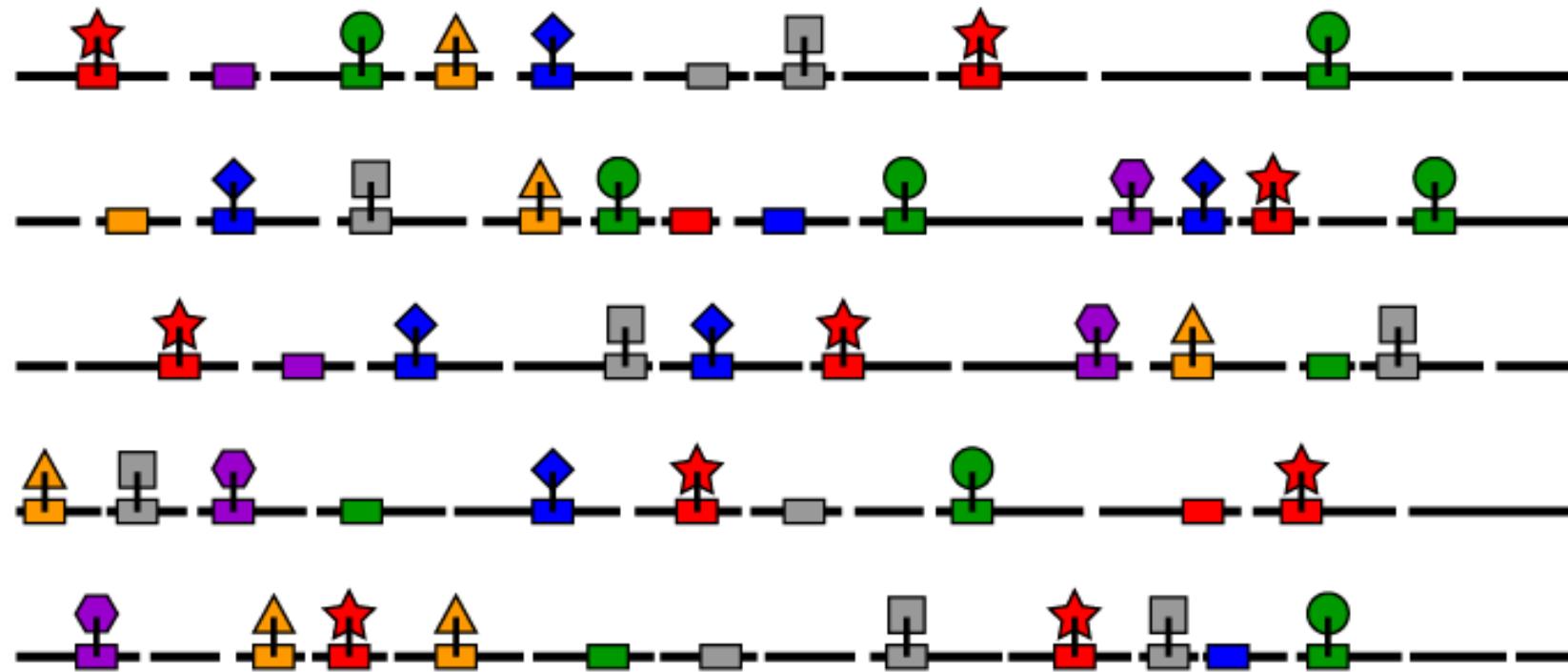
# Chromatin ImmunoPrecipitation (ChIP)



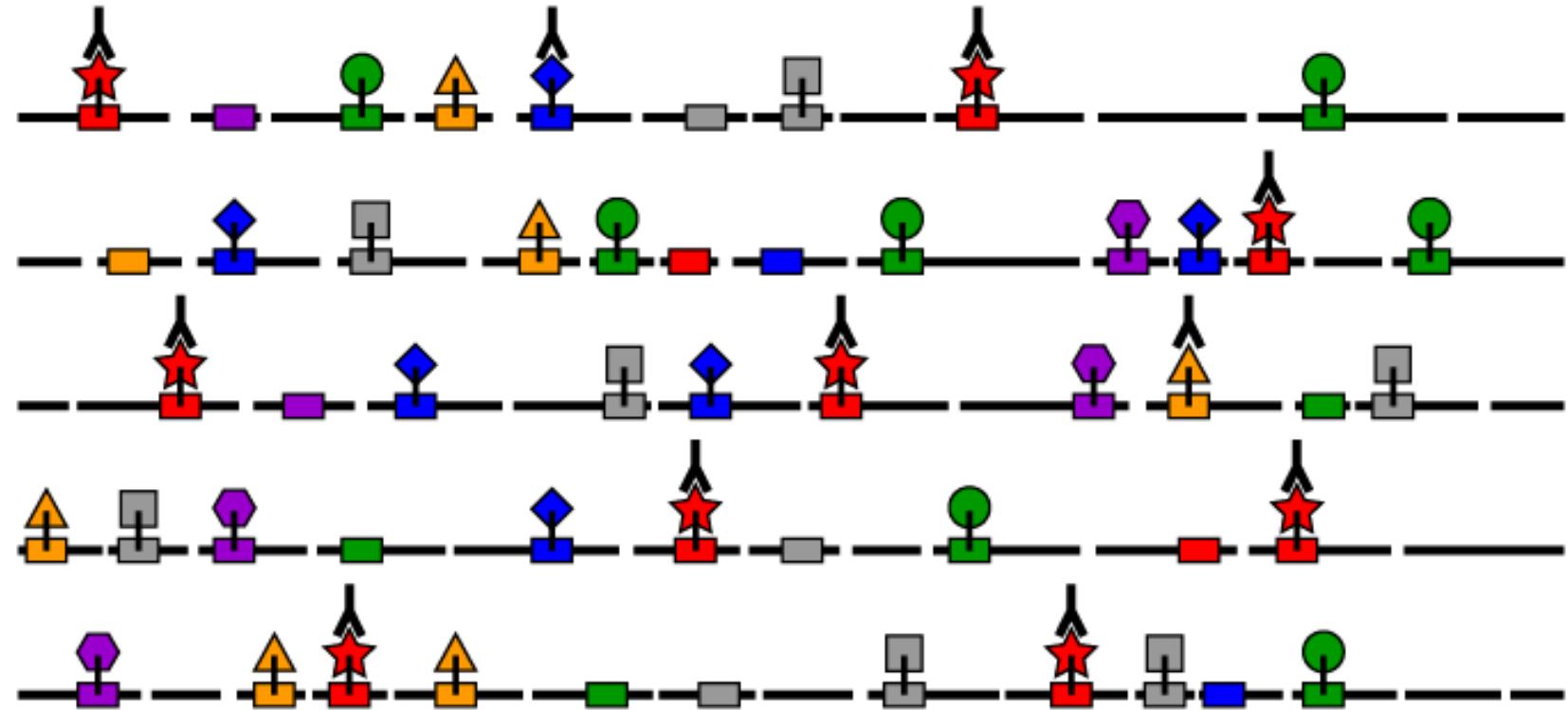
# Protein-DNA crosslinking *in vivo* (for TF)



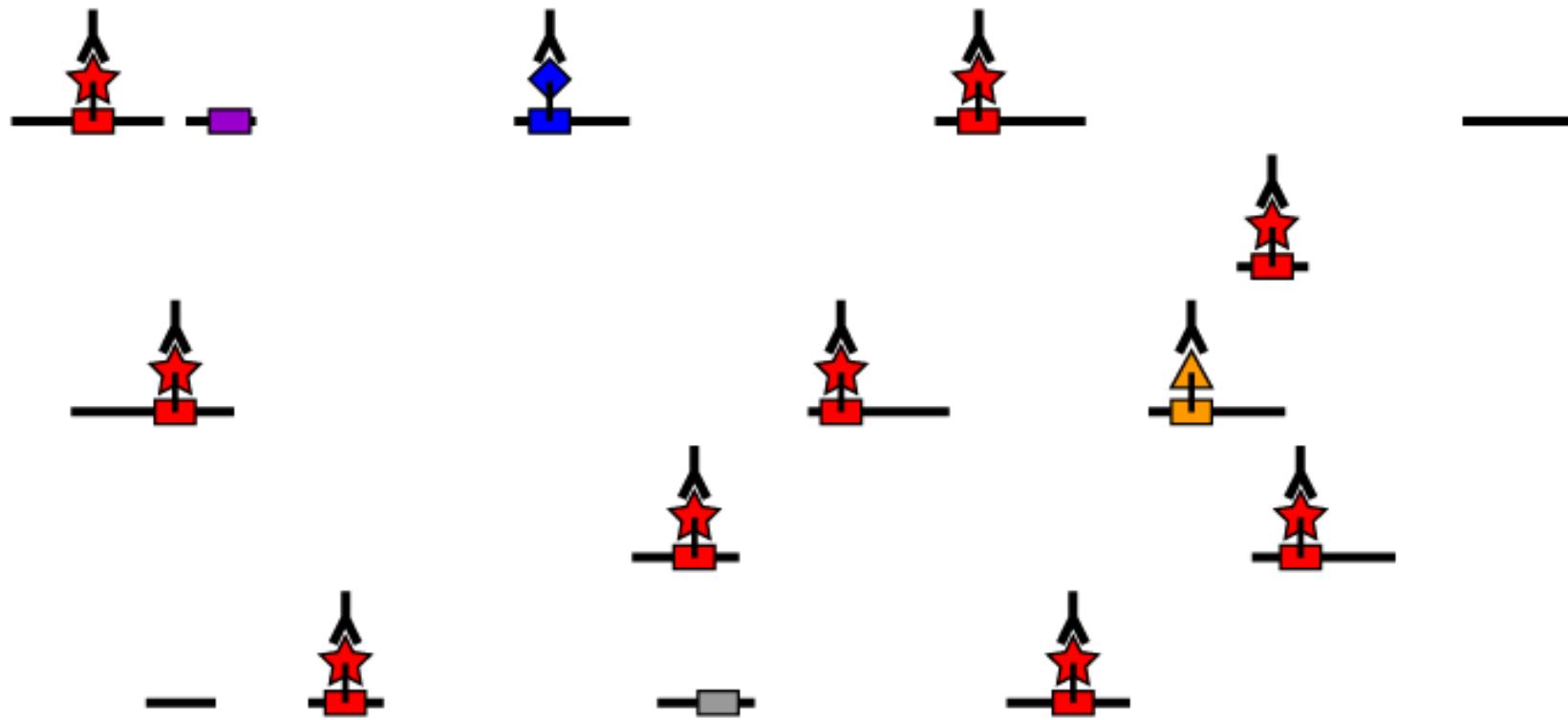
**Chop the chromatin using sonication (TF) or micrococcal nuclease (MNase) digestion (histone)**



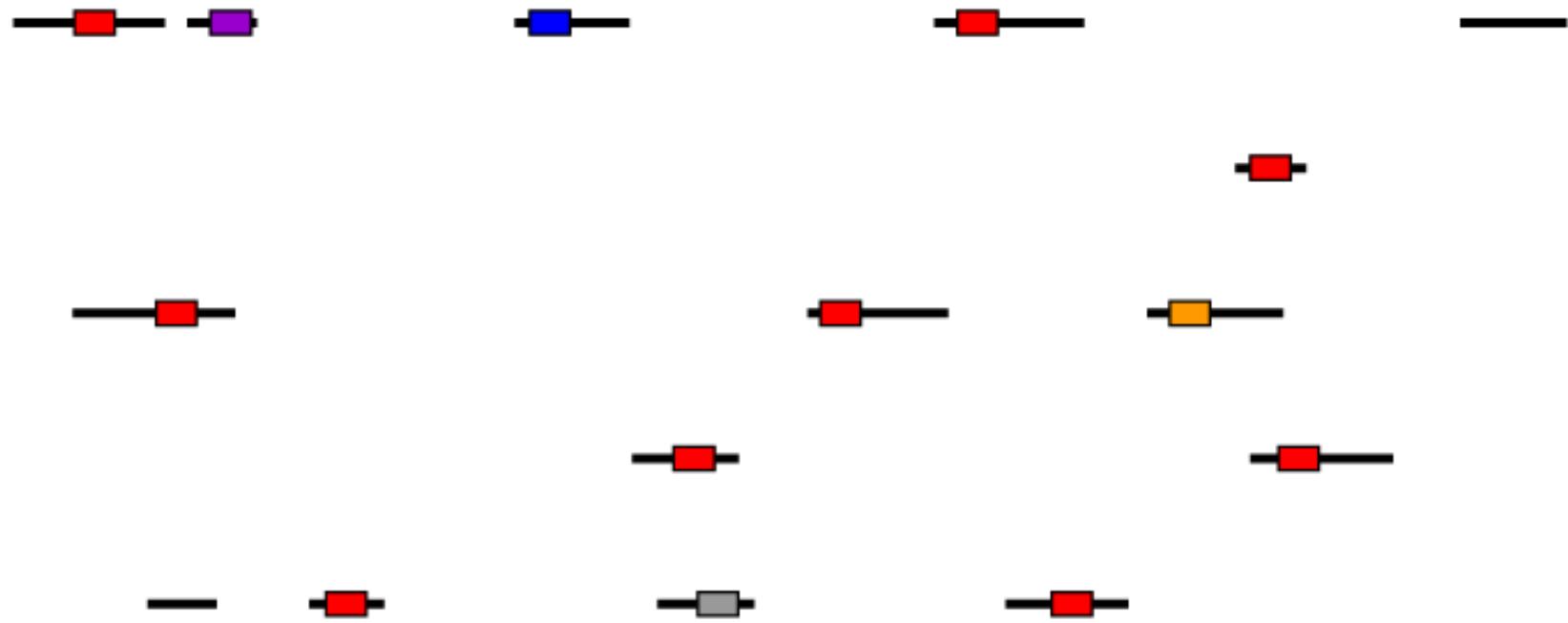
# Specific factor-targeting antibody



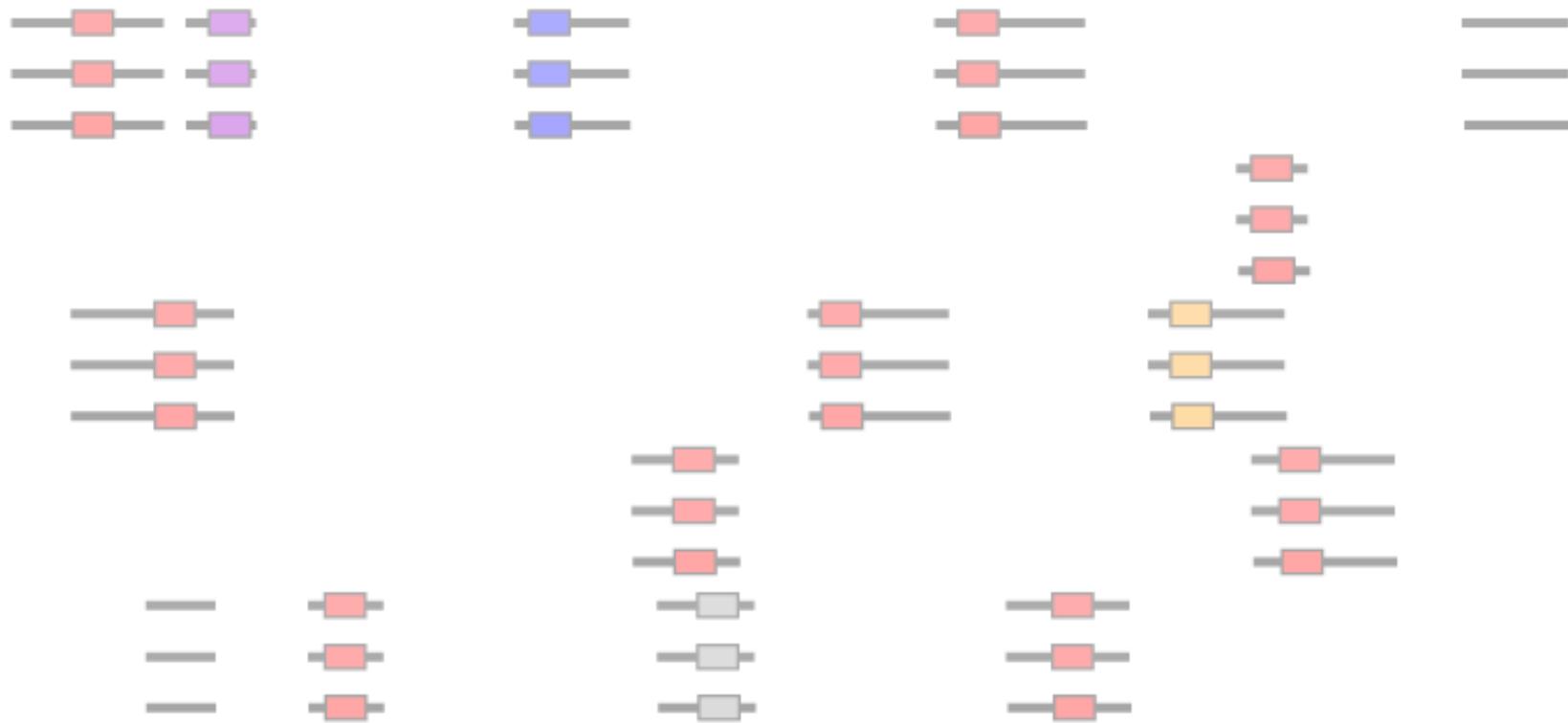
# Immunoprecipitation



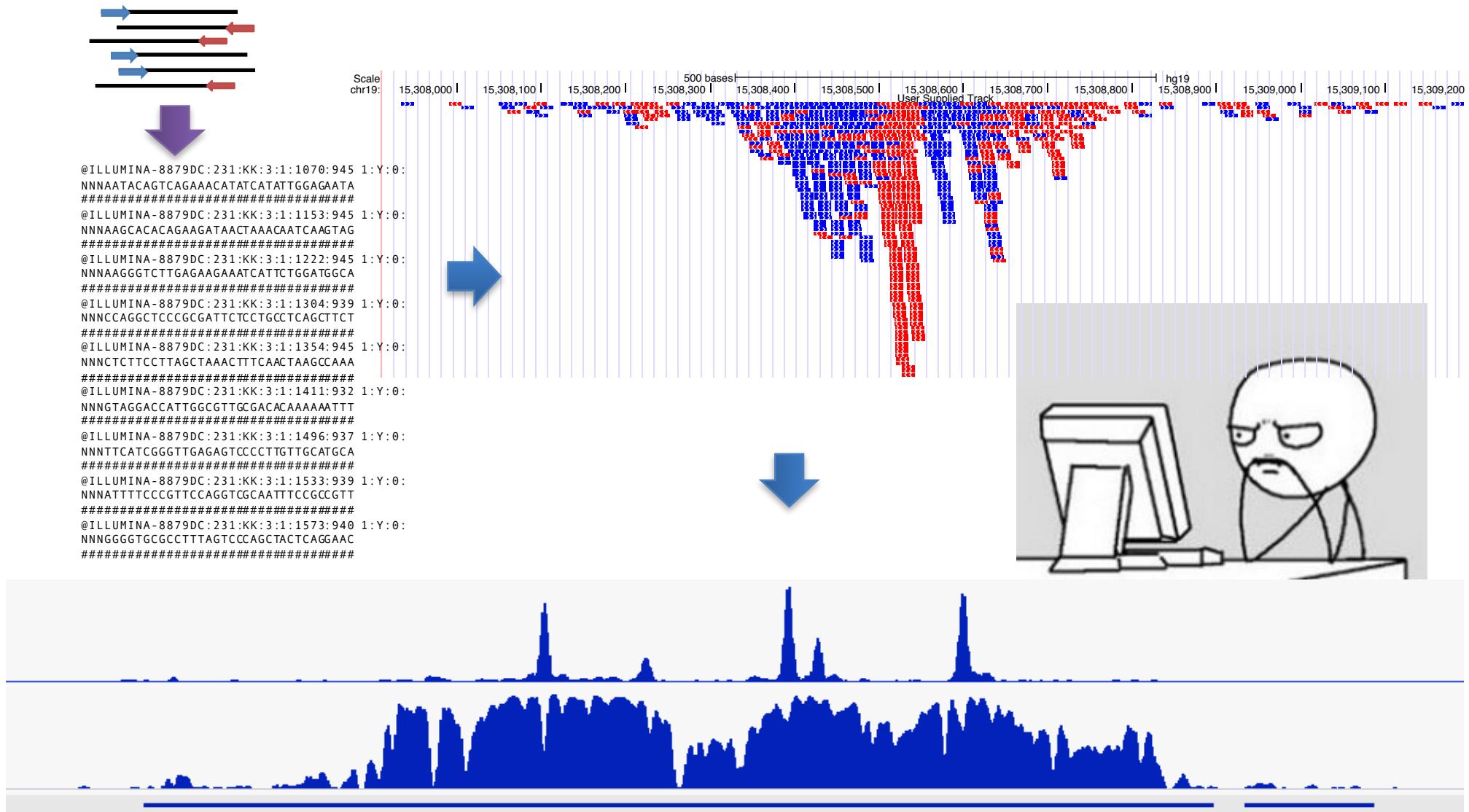
# DNA purification



# PCR amplification and sequencing

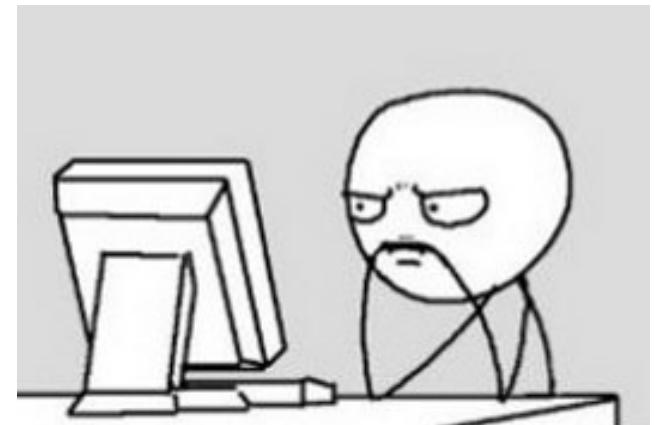


# ChIP-seq data analysis overview



# 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
- What can we learn from these data? – Downstream analysis and integration

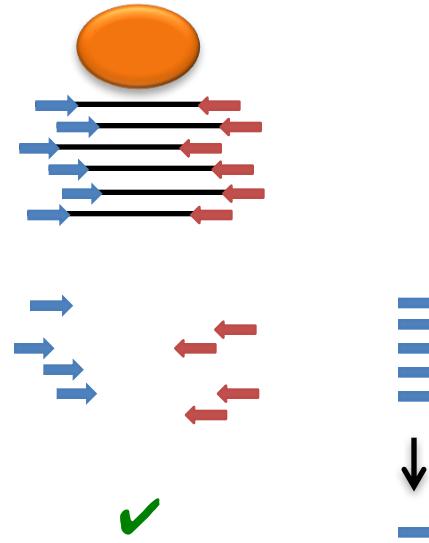


# ChIP-seq data analysis: basic processing

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

- {
  - cannot map to the reference genome X
  - can map to multiple loci in the genome X
  - can map to a unique location in the genome ✓

- redundancy control:

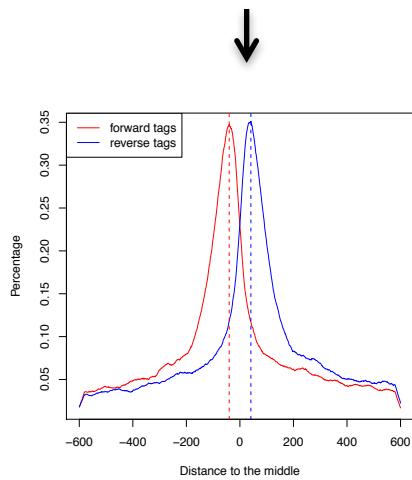
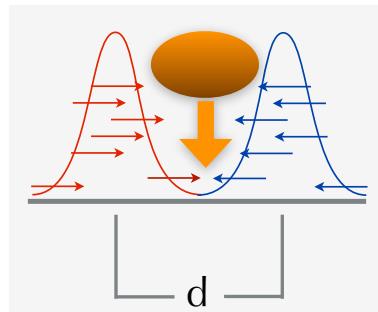


Langmead et al. 2009,  
Zang et al. 2009

# ChIP-seq data analysis: Peak calling

- DNA fragment size estimation
- pile-up profiling

peak model



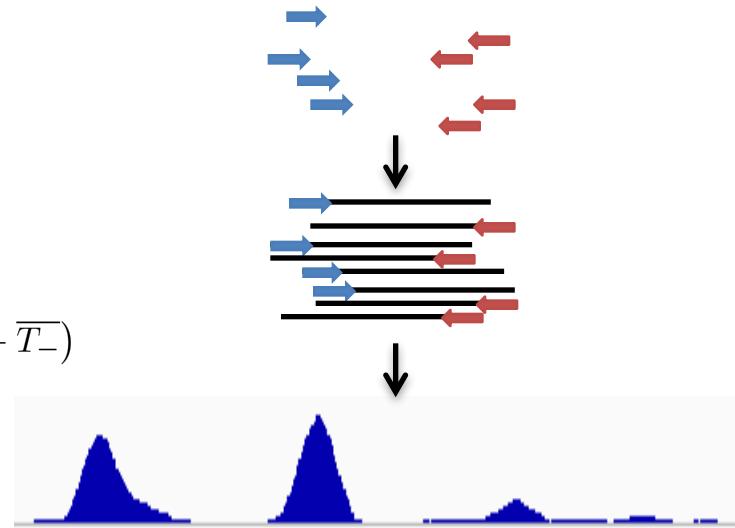
cross-correlation

$$C(r) = \frac{1}{X} \int_x (T_+(x) - \bar{T}_+) (T_-(x + r) - \bar{T}_-)$$

A diagram illustrating cross-correlation. It shows two sets of parallel horizontal lines. Blue arrows point to the right between the top set, and red arrows point to the left between the bottom set. A mathematical equation for cross-correlation is shown:  $C(r) = \frac{1}{X} \int_x (T_+(x) - \bar{T}_+) (T_-(x + r) - \bar{T}_-)$ .

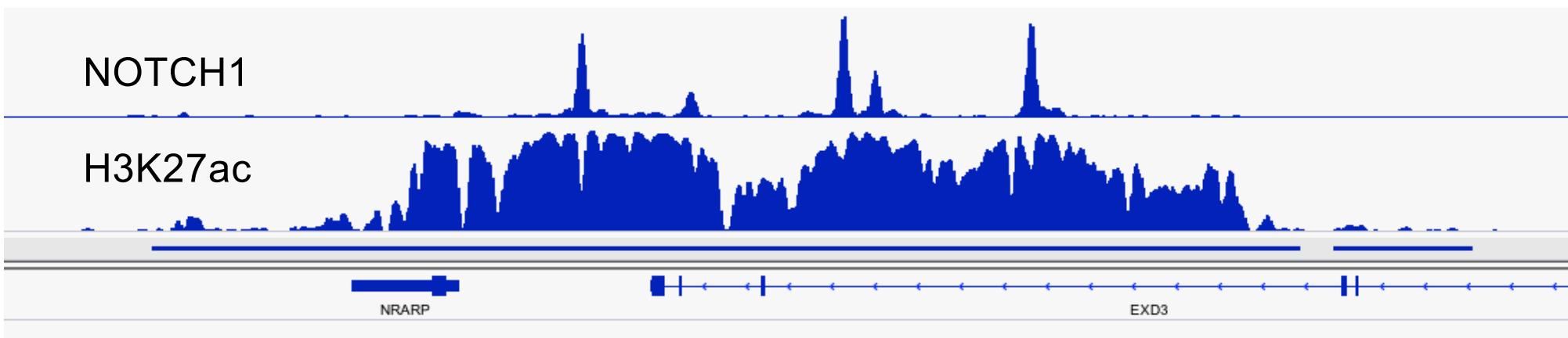


- Peak/signal detection



# ChIP-seq data analysis: Peak calling

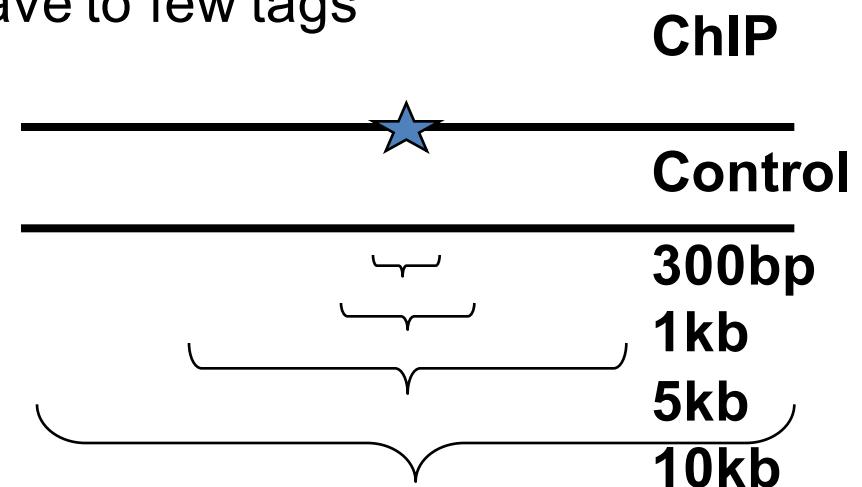
- **Sharp peaks**  
transcription factor binding,  
DNase, ATAC-seq  
  
**MACS** (Zhang, 2008)  
dynamic background  
Poisson model
- **Broad peaks**  
Histone modifications,  
“super-enhancers”  
Diffuse  
  
**SICER** (Zang, 2009)  
Spatial clustering of localized  
weak signal and integrative  
Poisson model



# MACS

- Model-based Analysis for ChIP-Seq
- Tag distribution along the genome ~ Poisson distribution ( $\lambda_{BG}$  = total tag / genome size)
- ChIP-seq show local biases in the genome
  - Chromatin and sequencing bias
  - 200-300bp control windows have few tags
  - But can look further

$$\text{Dynamic } \lambda_{local} = \max(\lambda_{BG}, [\lambda_{ctrl}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k}])$$



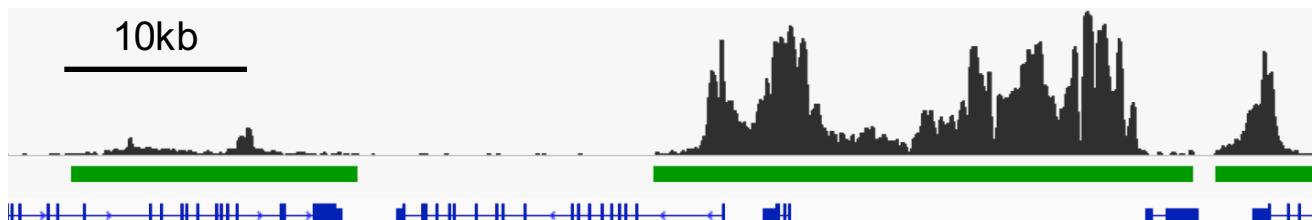
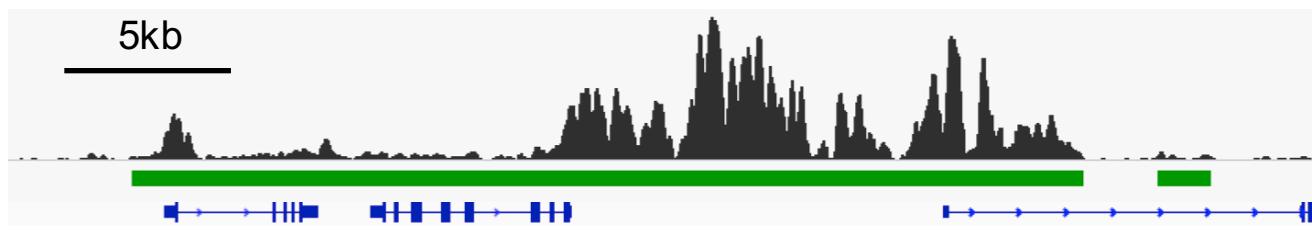
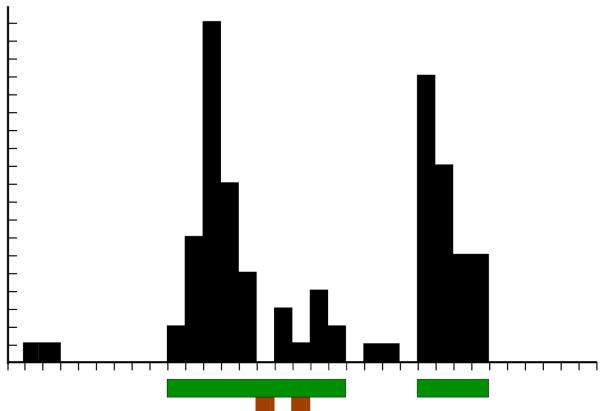
# SICER

- Spatial-clustering Identification of ChIP-Enriched Regions

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

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

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



omictools.com

# 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.

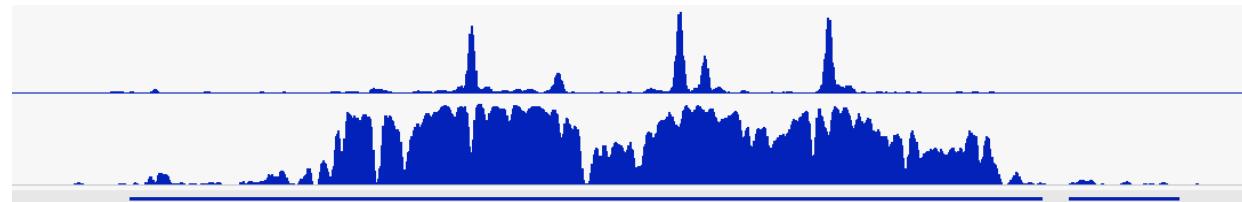
# ChIP-seq data analysis: Review

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

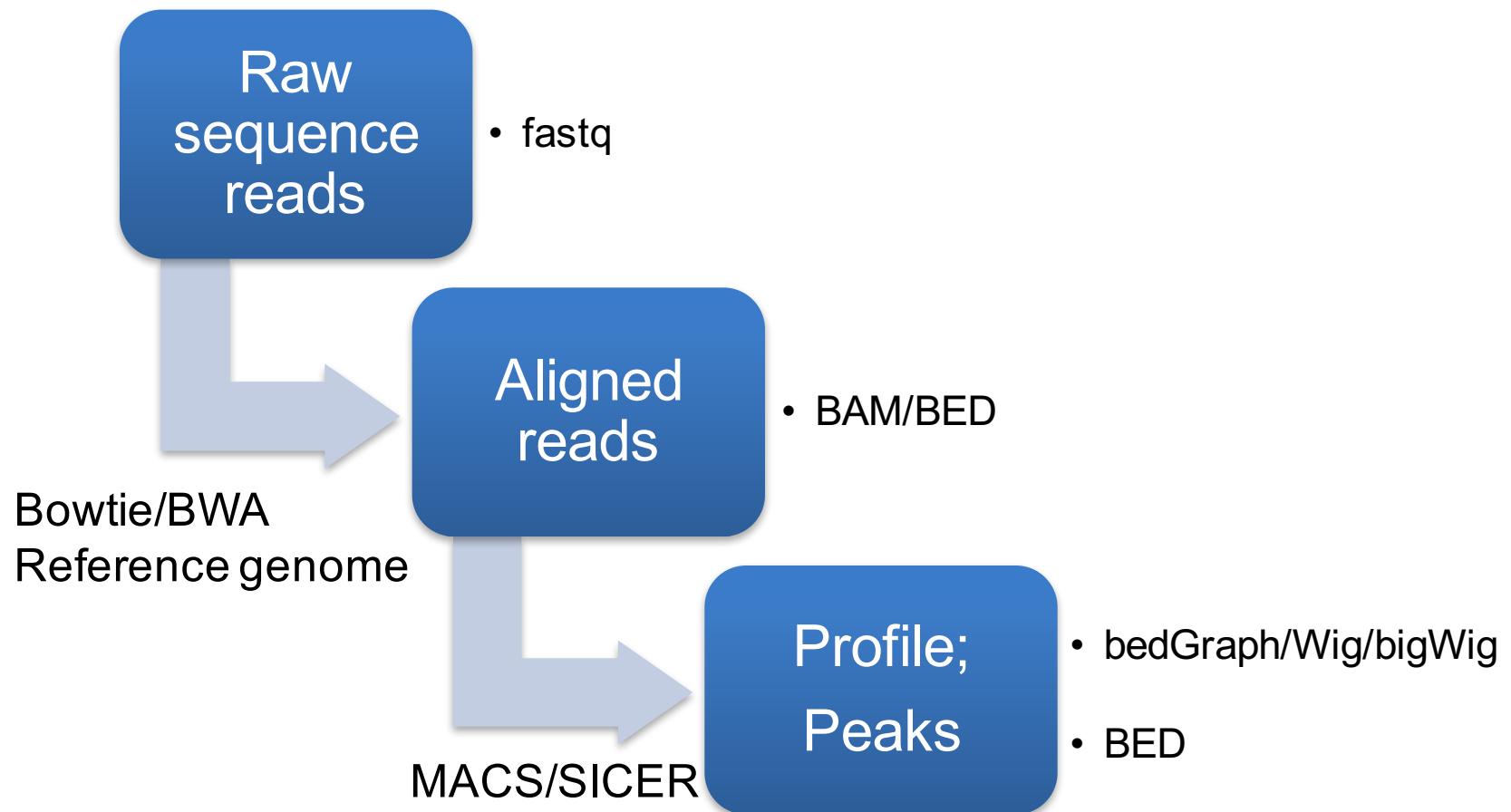
# Data formats

- 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



# Galaxy: web-interface analysis platform

- <https://usegalaxy.org/>

The screenshot shows the Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Help, User, and a grid icon. The main content area features a central text block about Galaxy, a logo for "IMMPORTGALAXY Flow Cytometry & CyTOF Analysis", and a "Tweets" section with two tweets from the Galaxy Project. The left sidebar contains a "Tools" section with a search bar and a "Get Data" section listing various servers like UCSC Main, EBI SRA, and BioMart. The right sidebar shows a "History" section with an empty history named "Unnamed history". Logos for Penn State, Johns Hopkins, Oregon Health & Science, TACC, and CYVERSE are visible at the bottom.

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](#).

**Tweets** by @galaxyproject

**Galaxy Project** @galaxyproject  
G-OnRamp Beta Testers Workshops: Create genome browsers for collaborative eukaryotic genome annotation  
[galaxyproject.org/news/2017-03-g...](http://galaxyproject.org/news/2017-03-g...)  
#usegalaxy

**Washington University in St.Louis**

**Galaxy Project** @galaxyproject  
Early career / student researcher working in

PENNSTATE JOHN HOPKINS OREGON HEALTH & SCIENCE TACC CYVERSE

# Run MACS on Cistrome, a Galaxy-based platform

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

The screenshot shows the Cistrome Galaxy-based platform interface. The top navigation bar includes tabs for "Galaxy / Cistrome", "Analyze Data", "Workflow", "Shared Data", "Lab", "Visualization", "Help", and "User". A "History" panel on the right lists 18 entries, all related to heatmap analysis (e.g., "68: Heatmap log", "67: Heatmap k-means classified regions"). The main content area displays the "Upload File (version 1.1.4)" tool. It has 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 about specifying URLs or pasting file contents), and "Files uploaded via ASPERA" (a table with no files listed). On the left, a "CISTROME TOOLBOX" sidebar lists various tools: Import Data (Upload File, CistromeFinder Import, CistromeCR Import, Expression CEL file packager, GenomeSpace import), Data Preprocessing (Gene Expression, Integrative Analysis, Liftover/Others), and GALAXY TOOLBOX (Get Data, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences). A "Tools" section at the top left contains a search bar and a "search tools" dropdown.

# Run SICER on Galaxy-based platforms

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

The screenshot shows a web browser window with the following details:

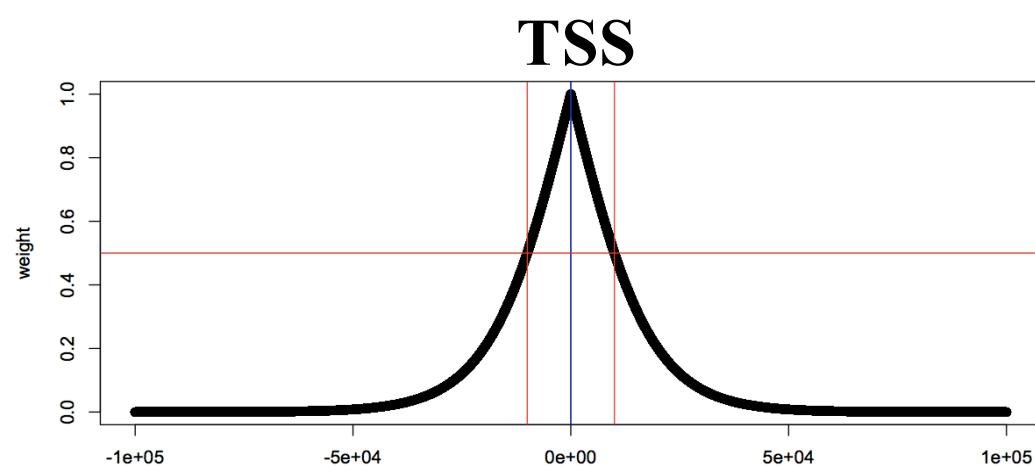
- Header:** Galaxy / CBiB. The title bar includes the Galaxy logo, a search bar with the URL "services.cbib.u-bordeaux.fr/galaxy/", and a user name "Chongzhi".
- Left Sidebar (Tools):** A list of bioinformatics tools categorized under "Metagenomic analyses", "FASTA manipulation", "EMBOSS", "NGS: QC and manipulation", "NCBI Blast plus", "NGS: Mapping", "NGS: RNA Analysis", "NGS: SAM Tools", "NGS: GATK Tools (beta)", "NGS: Peak Calling", "MACS", "CCAT", "NGS: Simulation", "Phenotype Association", "VCF Tools", "Assembly tools", "Prokaryotes Genome Annotation", "genome alignment", "Bed Manipulation", "FASTQ manipulation", and "Workflows".
- Middle Content:**
  - A logo for "cgfb BIOINFORMATIQUE" is displayed.
  - The text "Centre de Bioinformatique de Bordeaux" is centered below the logo.
  - A welcome message: "Welcome to Galaxy. This page will give you all the information you need to get you started. If you wish for further information or tutorials, please visit Galaxy's public instance: <http://usegalaxy.org>. If you encounter any difficulty, please e-mail us at: [admin.cbib@u-bordeaux2.fr](mailto:admin.cbib@u-bordeaux2.fr)".
  - The text "The CBiB Galaxy maintenance team" is at the bottom of the main content area.
- Right Sidebar (History):** Shows an "Unnamed history" section with a note: "This history is empty. You can load your own data or get data from an external source".
- Bottom Footer:** A dashed blue box highlights the "What's new in this release !!" section, which lists several software tools: Prokka, MUMmer, Bedtools, and MEME. Below this, a note states: "This is a free, public, internet accessible resource. This service is provided as an academic best effort in the hope that it will be useful, but WITHOUT ANY WARRANTY".

# ChIP-seq: Downstream analysis

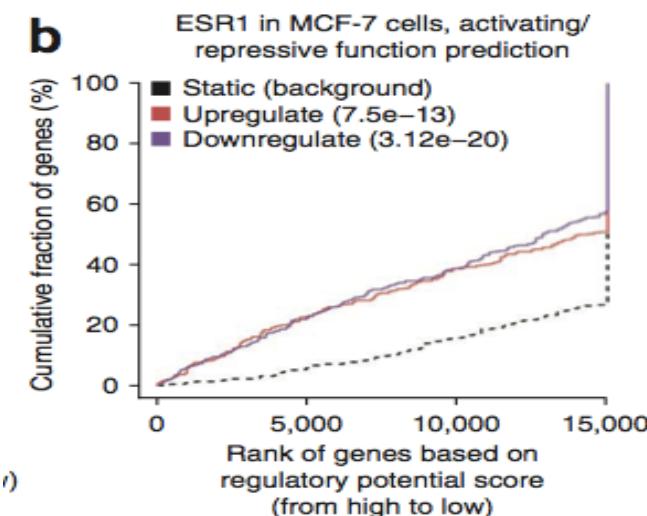
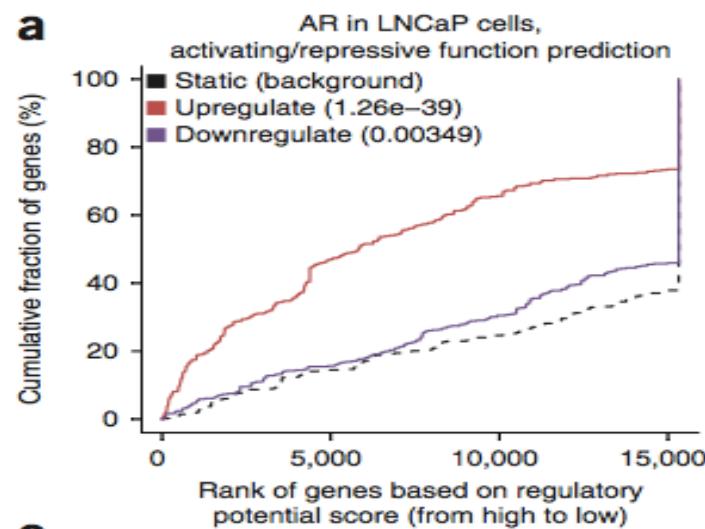
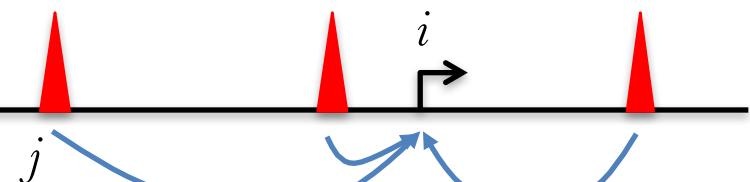
- Data visualization
  - UCSC genome browser: <http://genome.ucsc.edu/>
  - WashU epigenome browser:  
<http://epigenomegateway.wustl.edu/>
  - IGV: <http://software.broadinstitute.org/software/igv/>
- Meta analysis
  - CEAS: <http://liulab.dfci.harvard.edu/CEAS/>
- Integration with gene expression
  - 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

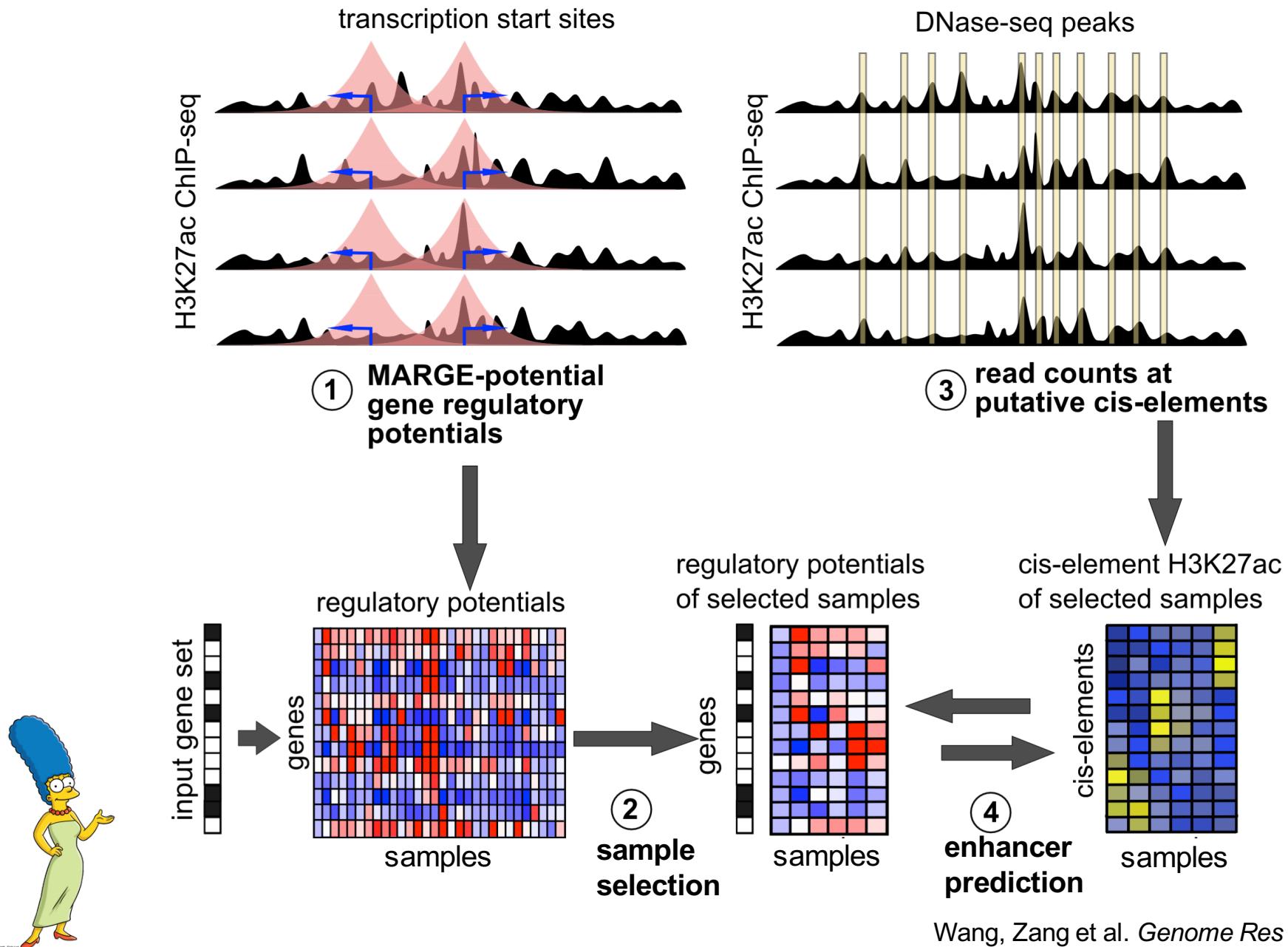
- Regulatory Potential

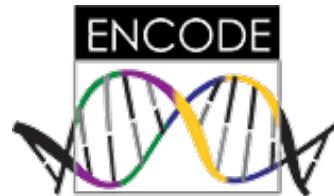


$$P(g_i) = \sum_{j \in S(i)} \exp\left(-\frac{\Delta_{ij}}{\lambda}\right)$$



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





# ENCODE

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

# Cistrome Data Browser

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

The screenshot shows the Cistrome Data Browser interface. At the top, there is a navigation bar with a back button, a refresh button, a search bar containing "cistrome.org/db/#/", and a star icon. Below the navigation bar is a large blue header with the text "Dataset Browser" and a logo consisting of three white circles connected by lines.

On the left side, there are three filter panels:

- Species**: A dropdown menu set to "All" with options "Homo sapiens" and "Mus musculus".
- Biological Sources**: A dropdown menu set to "All" with options "1015c", "10326", "1064Sk", "106A", and "10T1/2".
- Factors**: A dropdown menu set to "All" with options "ACTB", "ADNP", "ADNP2", "AEBP2", and "AFF1".

Below these filters is a search bar with a "Search" button and an "Options" dropdown.

The main area is titled "Results" and contains a table with the following columns:

Batch	Species	Biological Source	Factor	Publication	Status
<input type="checkbox"/>	Mus musculus	V6.5; Embryonic Stem Cell; Embryo	ATF7IP		completed
<input type="checkbox"/>	Homo sapiens	B Lymphocyte; Lymph Node	Dnase	Thurman RE, et al. Nature 2012	completed
<input type="checkbox"/>	Homo sapiens	MCF-7; Epithelium; Mammary Gland	ESR1	Welboren WJ, et al. EMBO J. 2009	completed
<input type="checkbox"/>	Homo sapiens	H9; Embryonic Stem Cell; Embryo	H3K23me2	Lister R, et al. Nature 2009	completed
<input type="checkbox"/>	Homo sapiens	Melanocyte; Foreskin	H3K27ac	Bernstein BE, et al. Nat. Biotechnol. 2010	completed
<input type="checkbox"/>	Mus musculus	B Lymphocyte; Bone Marrow	H3K27me3	Revilla-I-Domingo R, et al. EMBO J. 2012	completed
<input type="checkbox"/>	Mus musculus	Fibroblast; Embryo	H3K4me1	Koche RP, et al. Cell Stem Cell 2011	completed
<input type="checkbox"/>	Homo sapiens	H1; Embryonic Stem Cell; Embryo	H3K4me2	Lister R, et al. Nature 2009	36 completed
<input type="checkbox"/>	Mus musculus	Fibroblast; Embryo	H3K9ac	Feng TC, et al. J. Exp. Med. 2012	completed

# 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

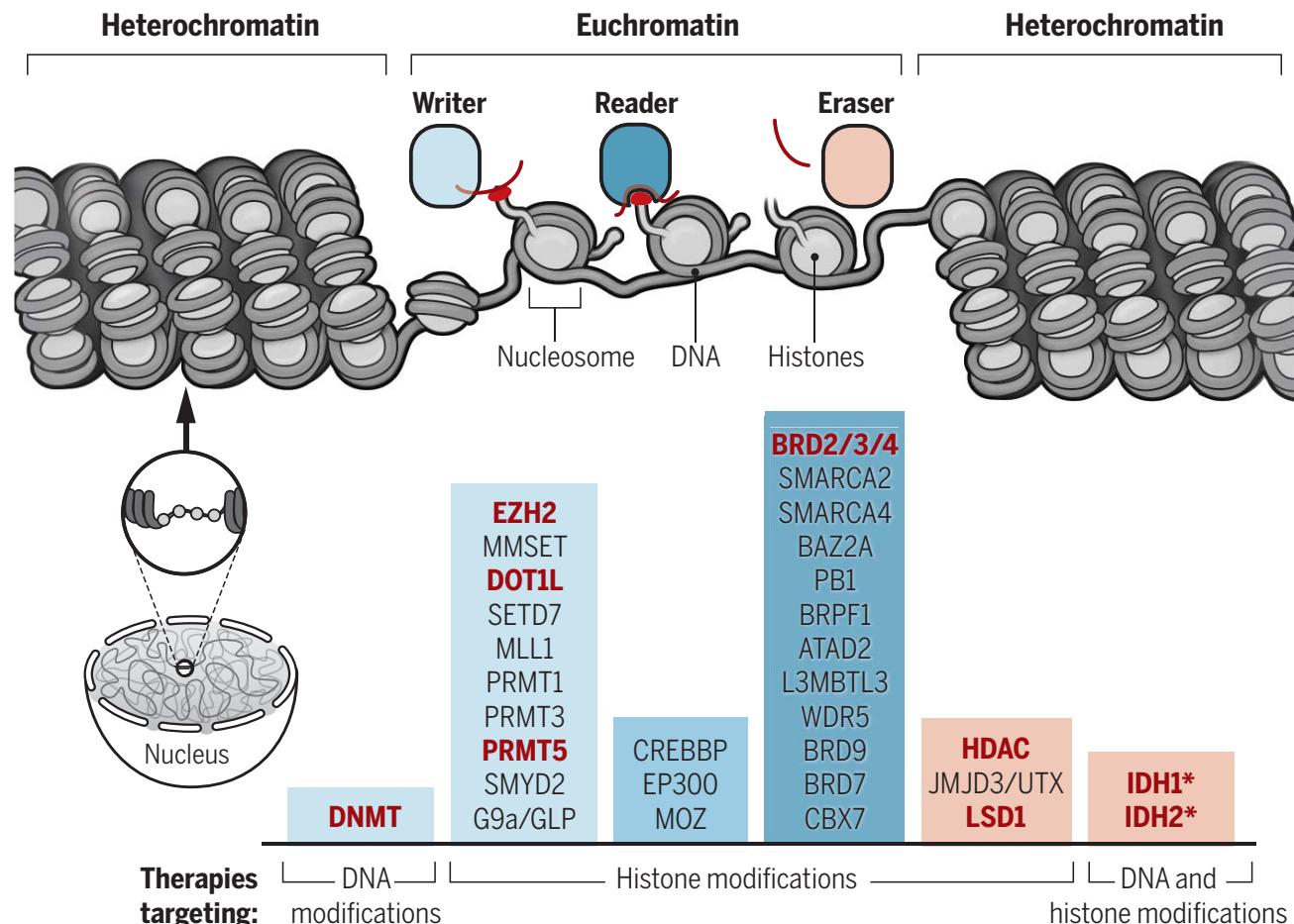
- ChIP-seq is used to profile epigenomes
- ChIP-seq data analysis
  - MACS for narrow peaks
  - SICER for broad peaks
- Online tools and resources

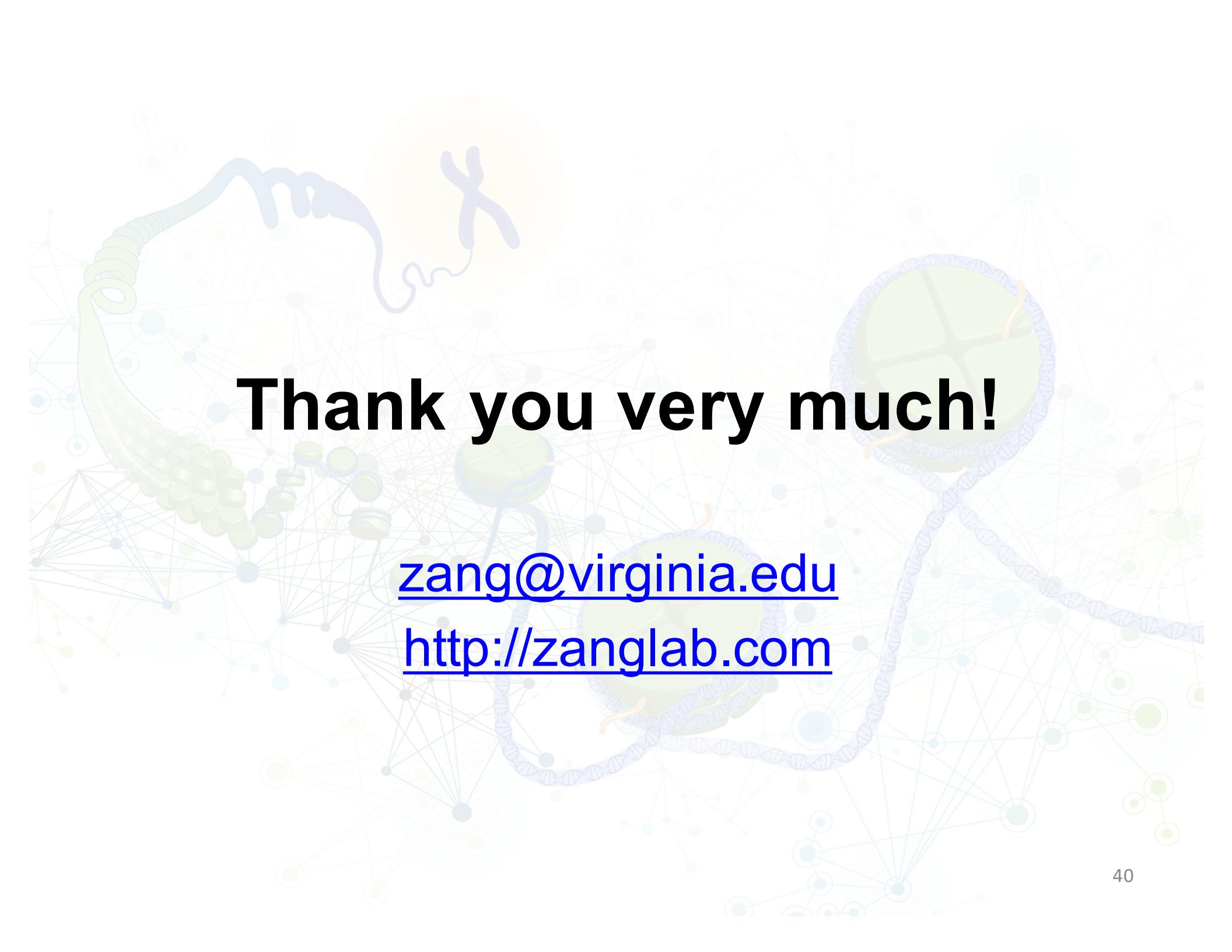
# Further Reading

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](mailto:zang@virginia.edu)

<http://zanglab.com>