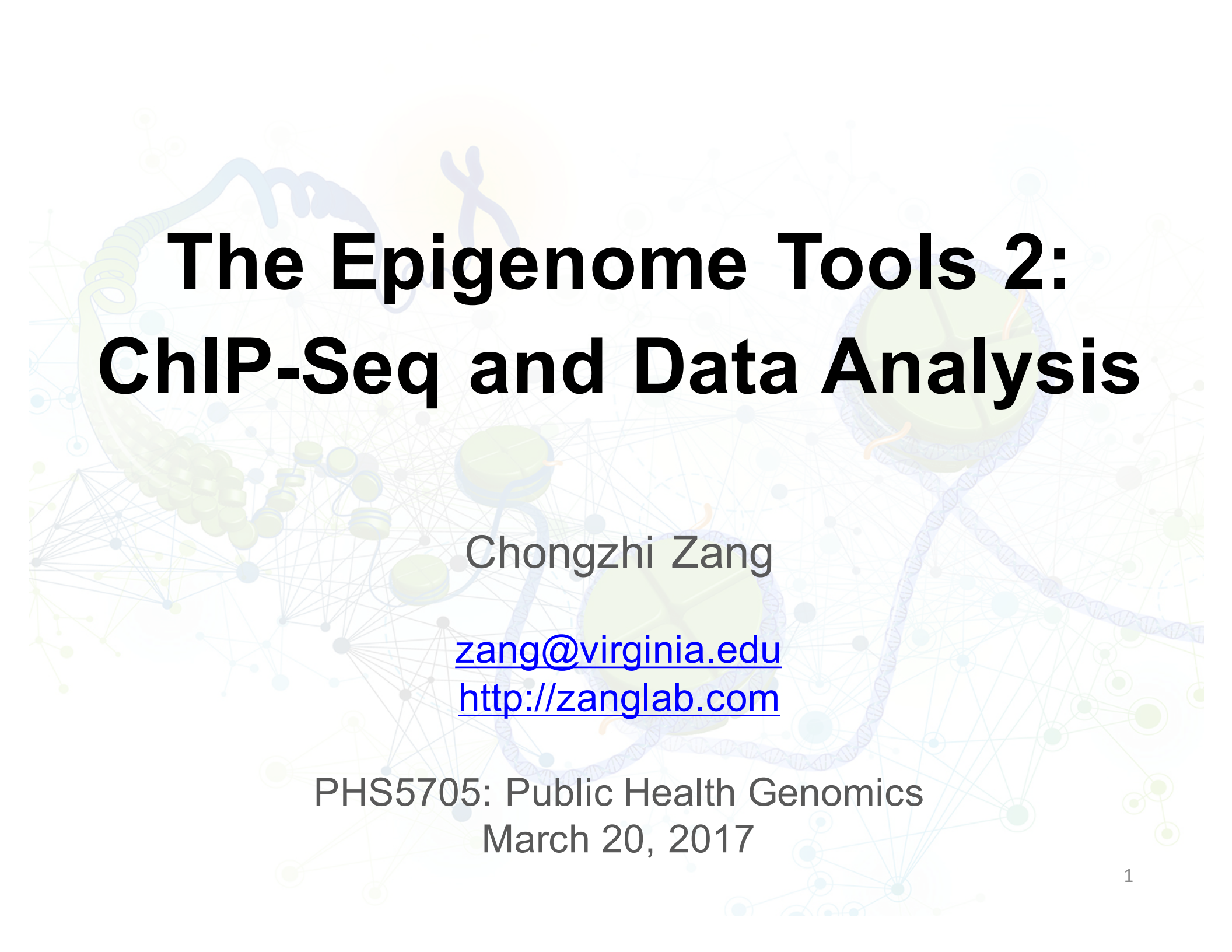




The Epigenome Tools 2: ChIP-Seq and Data Analysis

Chongzhi Zang

zang@virginia.edu

<http://zanglab.com>

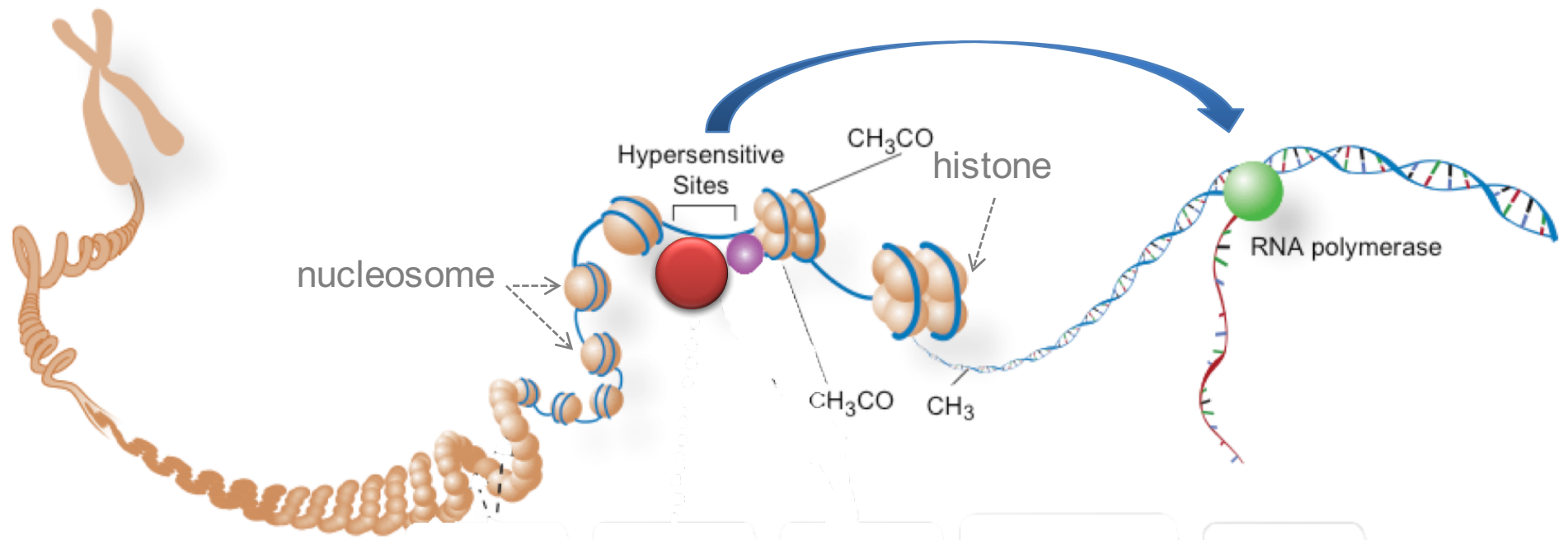
PHS5705: Public Health Genomics

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

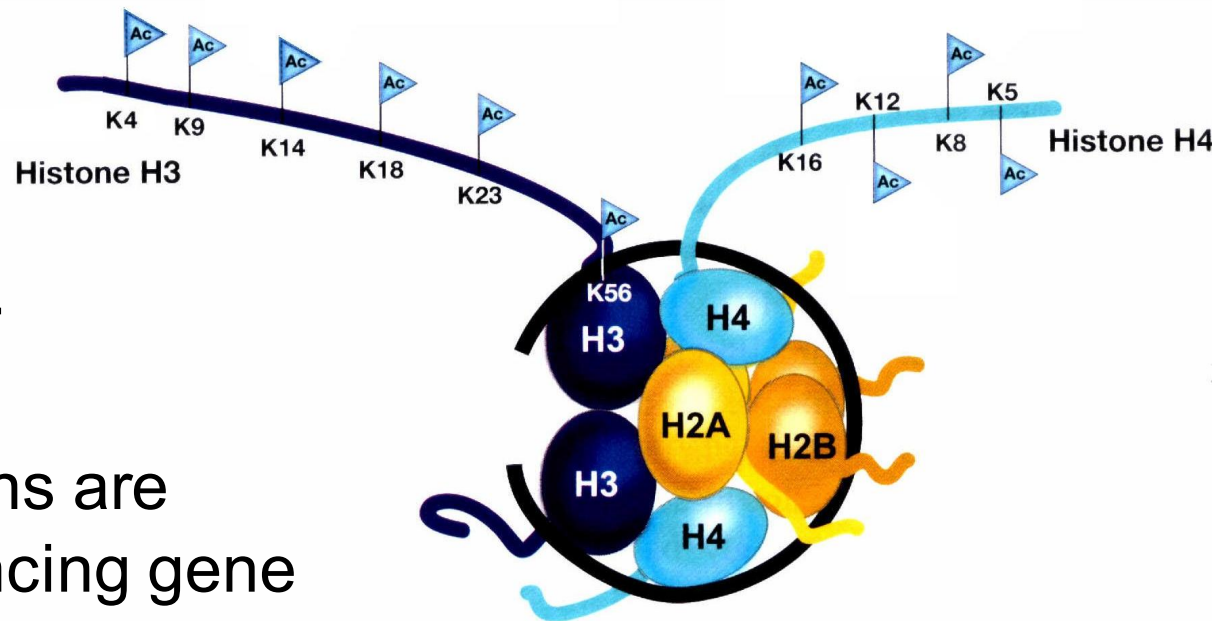
Epigenomic marks

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

Histone modifications

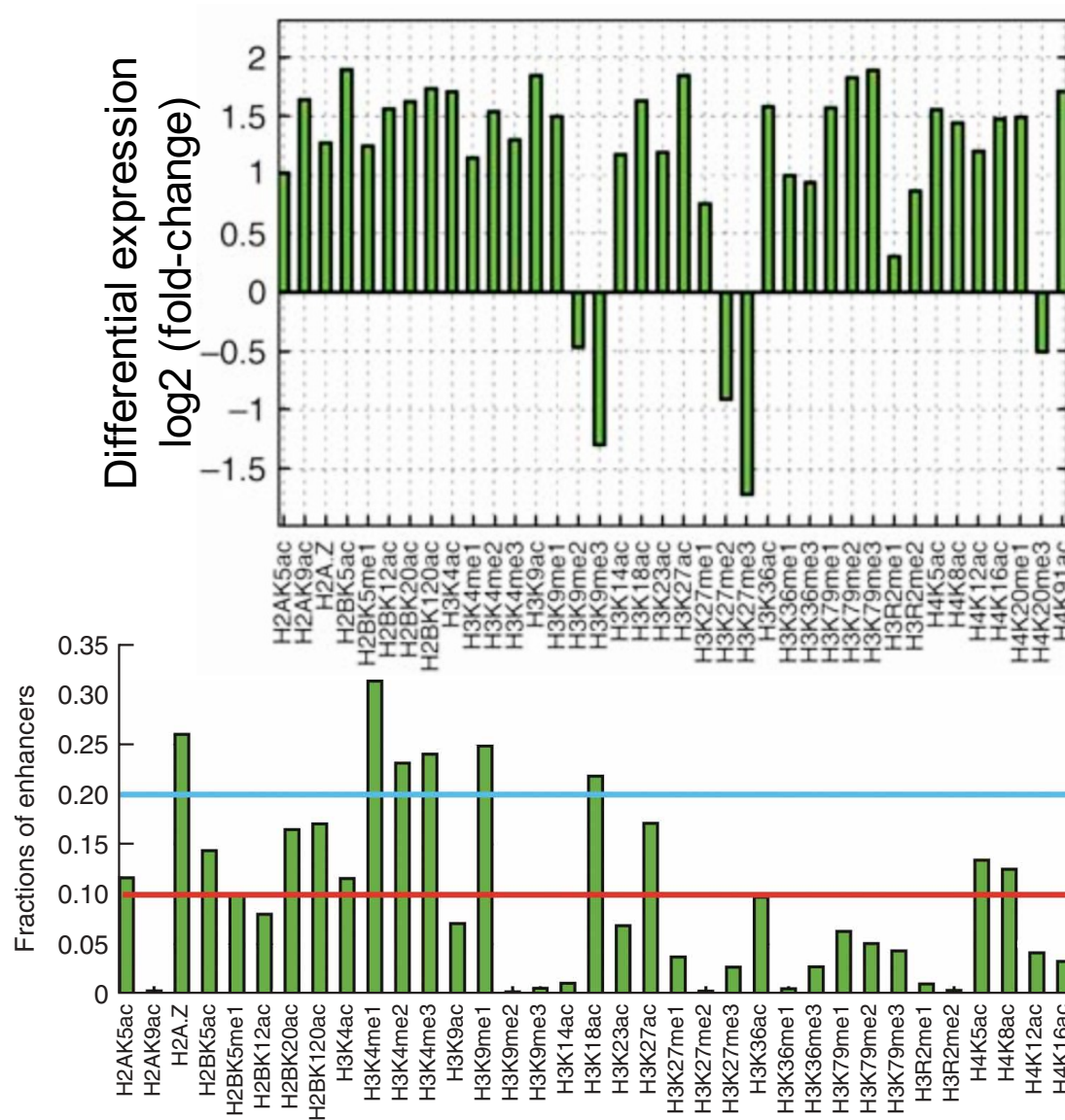
- Nucleosome Core Particles
- Core Histones: H2A, H2B, H3, H4
- Covalent modifications on histone tails include:
 - methylation (me),
 - acetylation (ac),
 - phosphorylation ...
- Histone variants
- Histone modifications are implicated in influencing gene expression.

Notation:
H3K4me3



Allis C. et al. Epigenetics. 2006

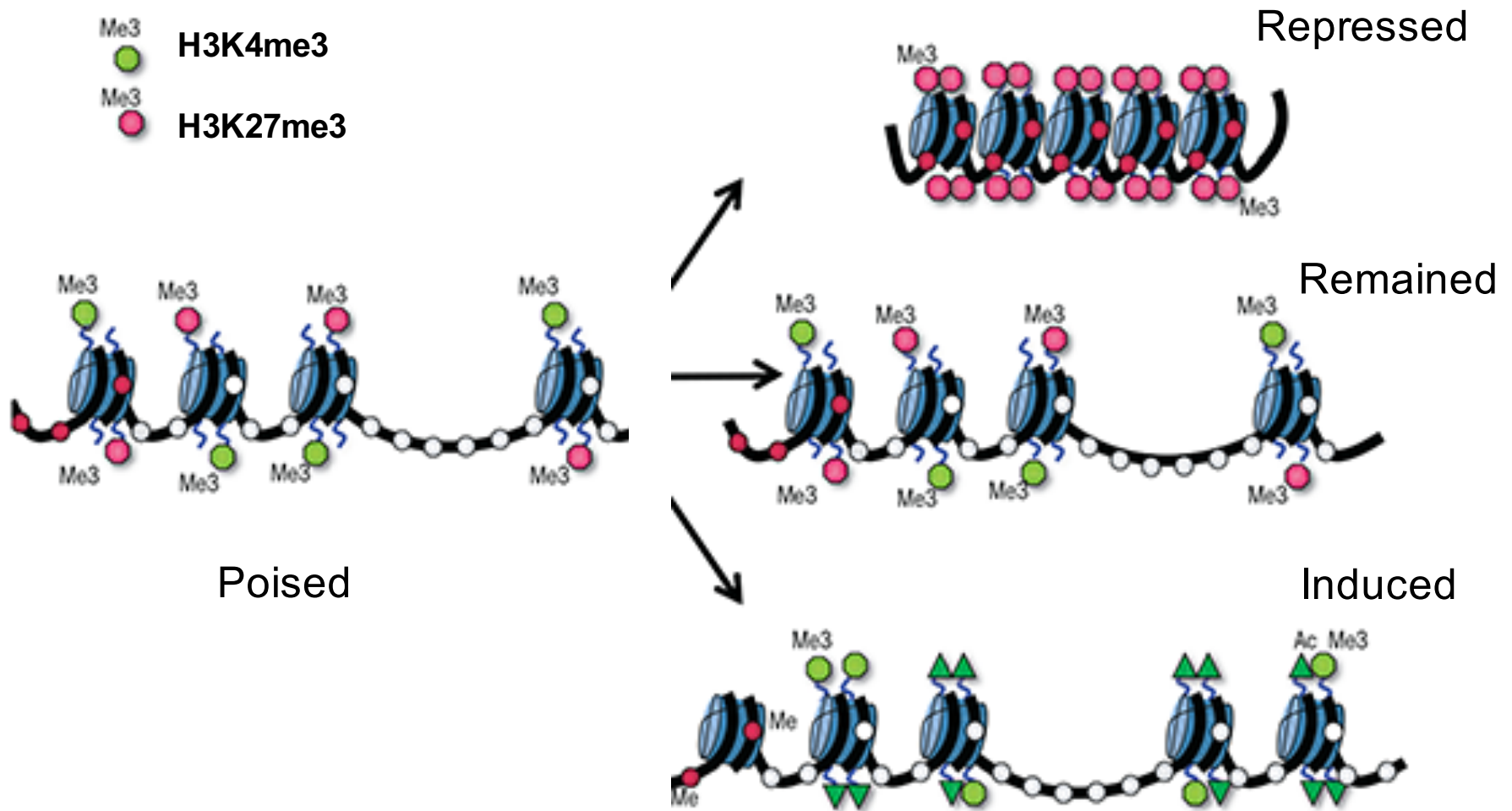
Histone modifications associate with regulation of gene expression



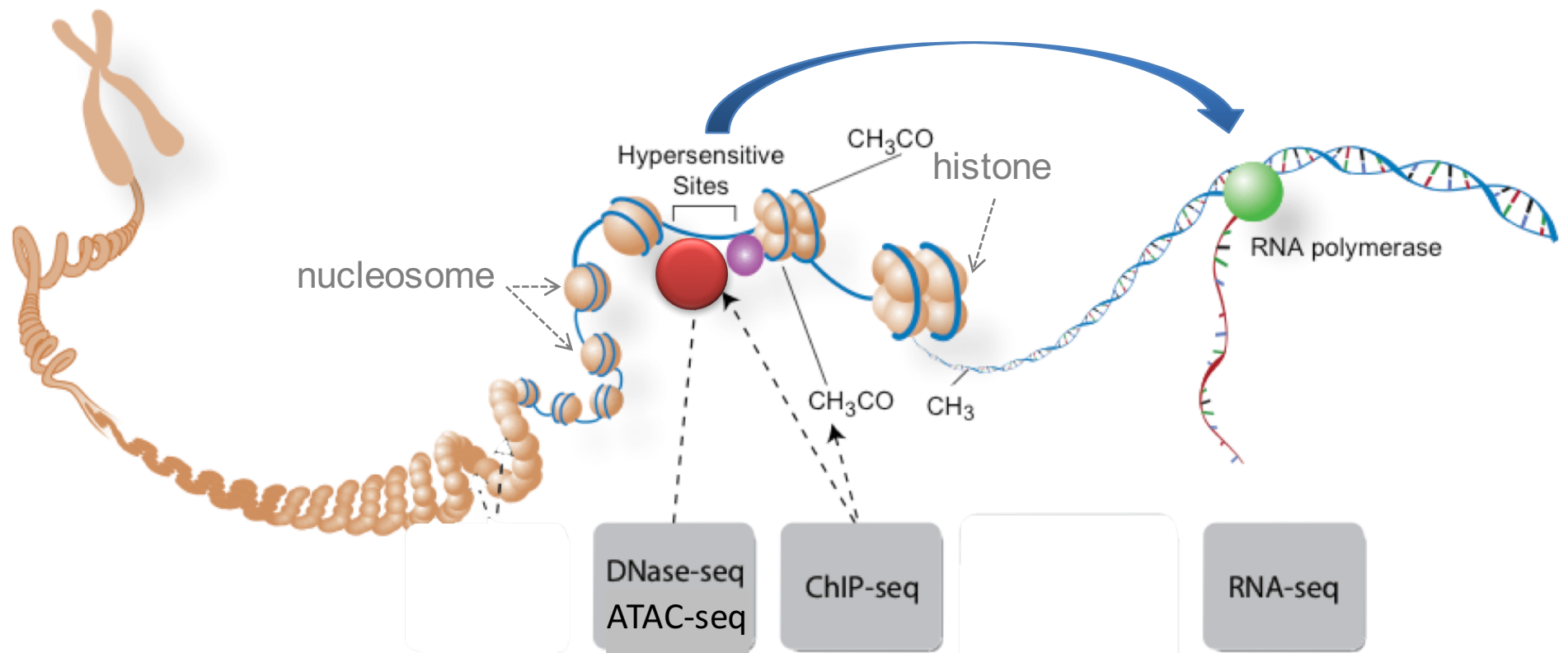
“Functions” of histone marks

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



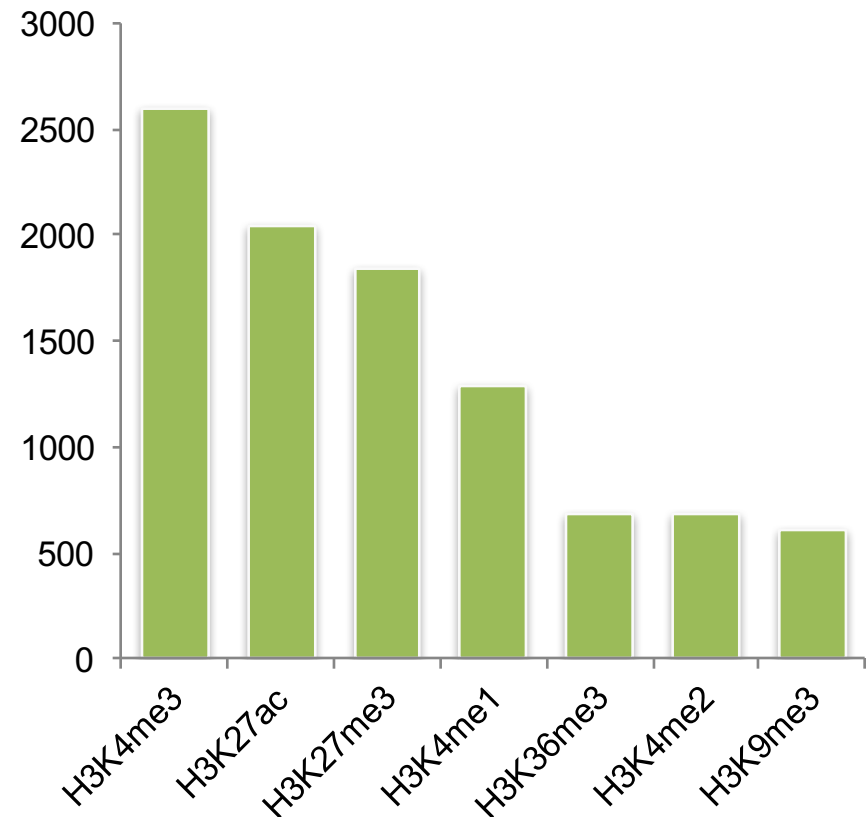
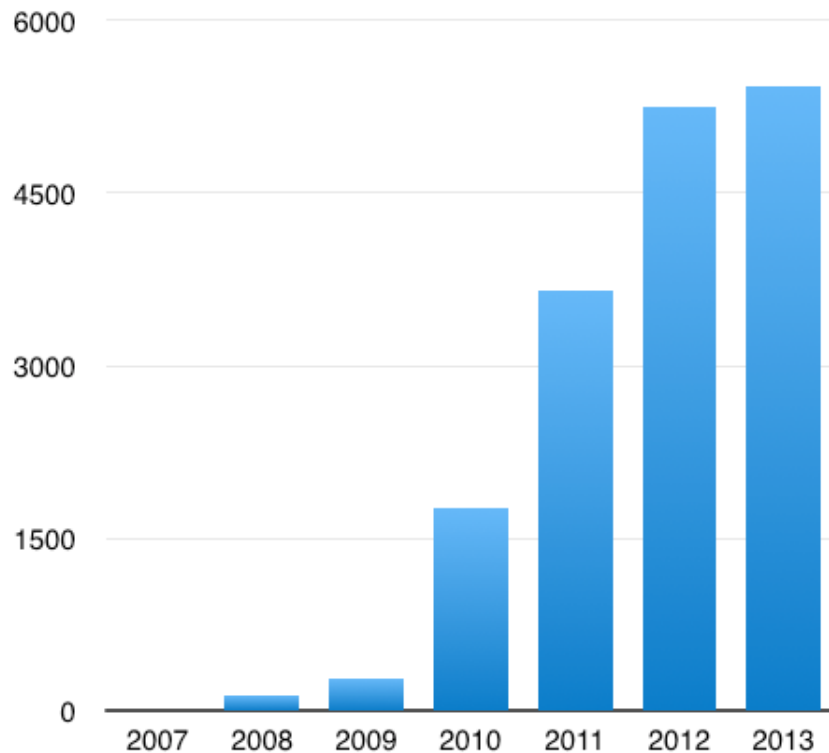
ChIP-seq: Profiling epigenomes with sequencing



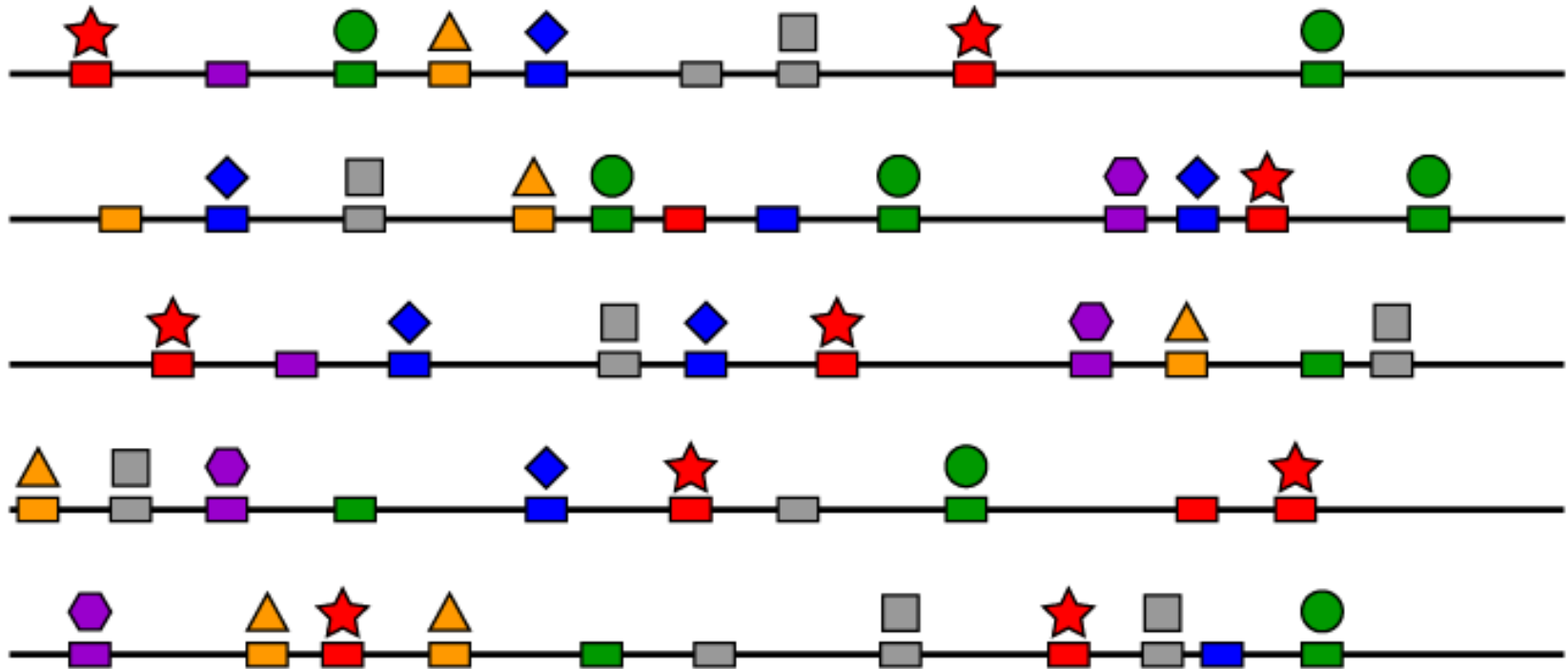
Published ChIP-seq datasets are skyrocketing

We are entering the Big Data era

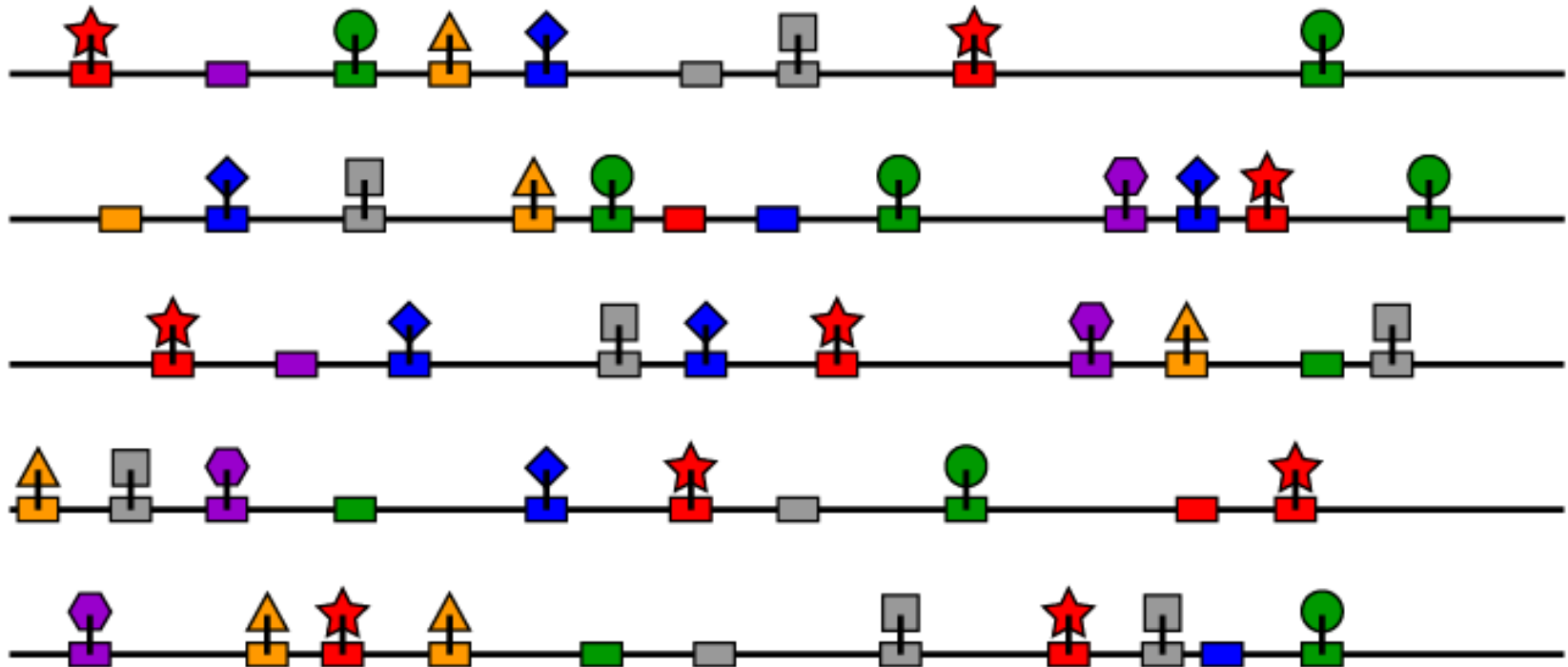
Number of ChIP-seq datasets on GEO



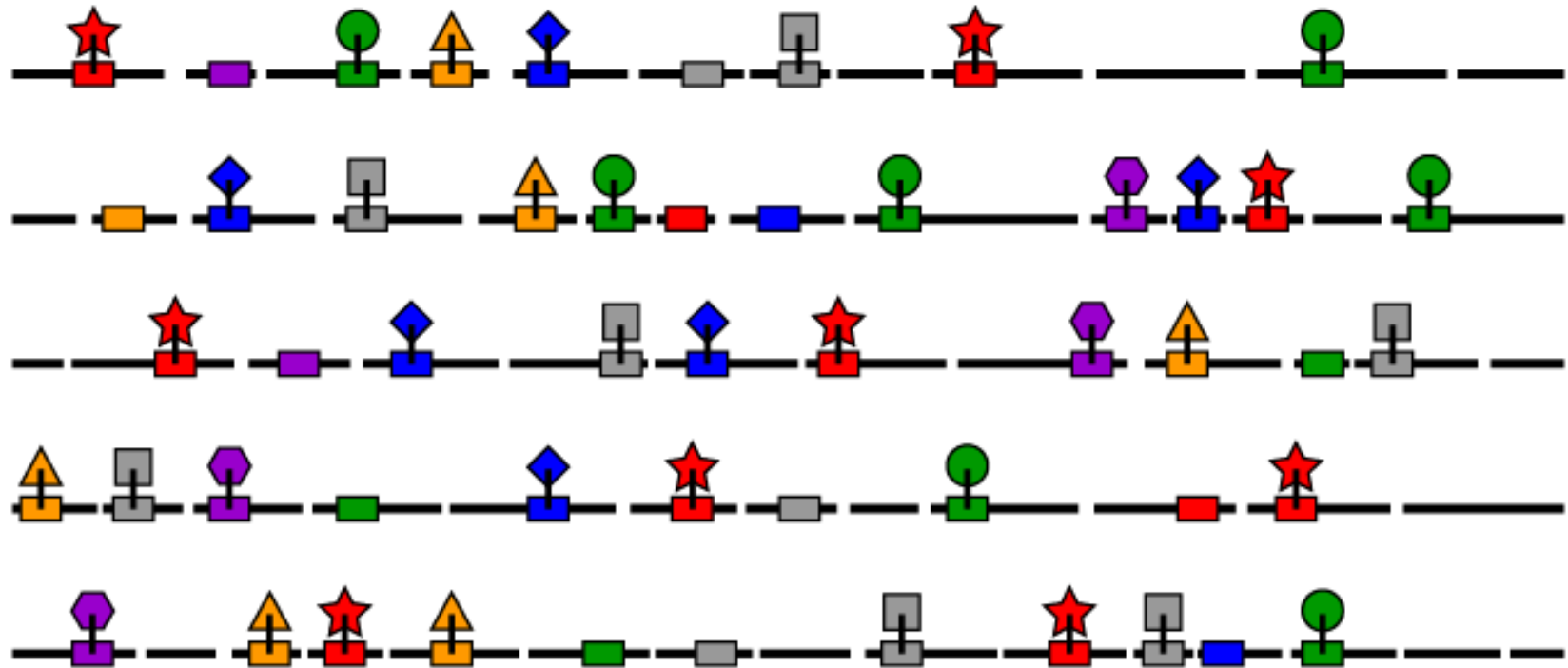
Chromatin ImmunoPrecipitation (ChIP)



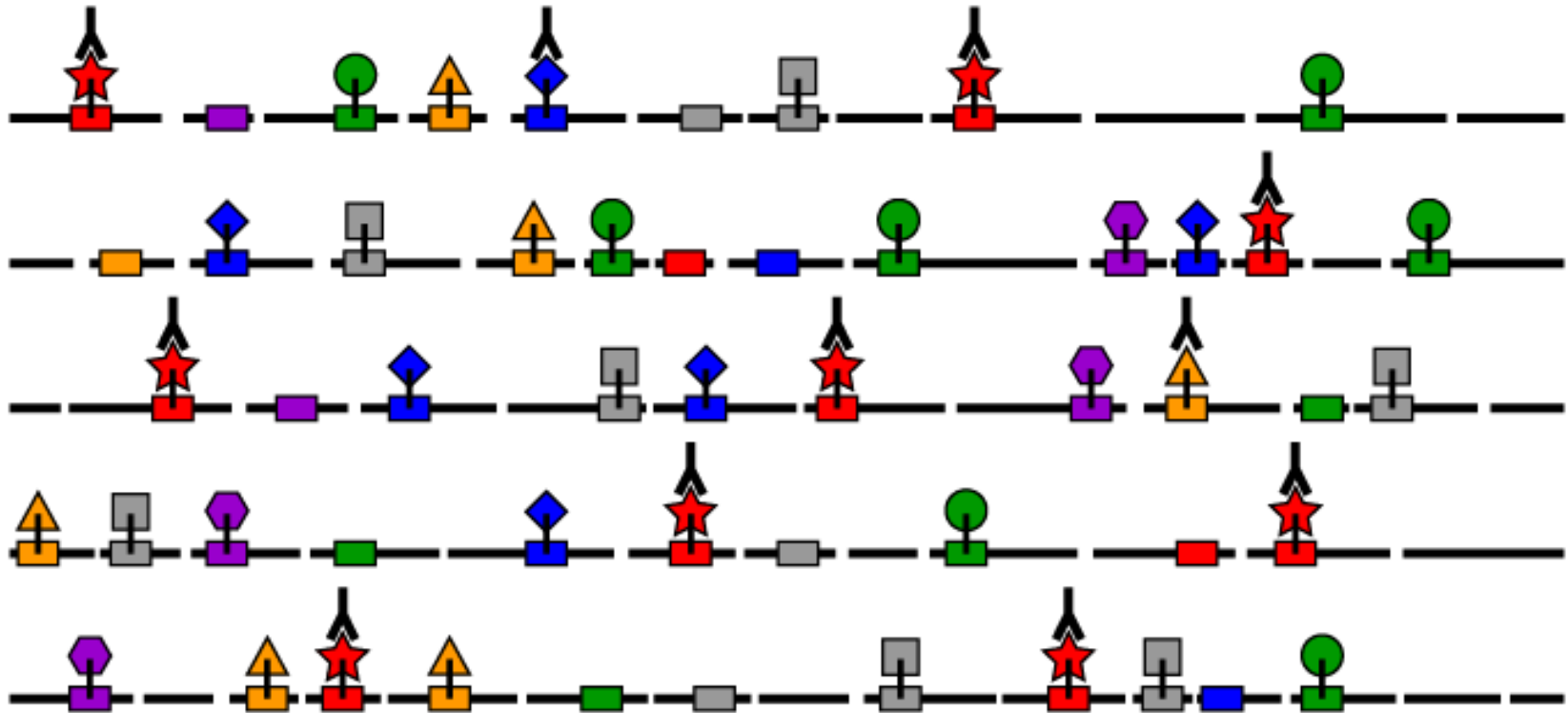
Protein-DNA crosslinking *in vivo* (for TF)



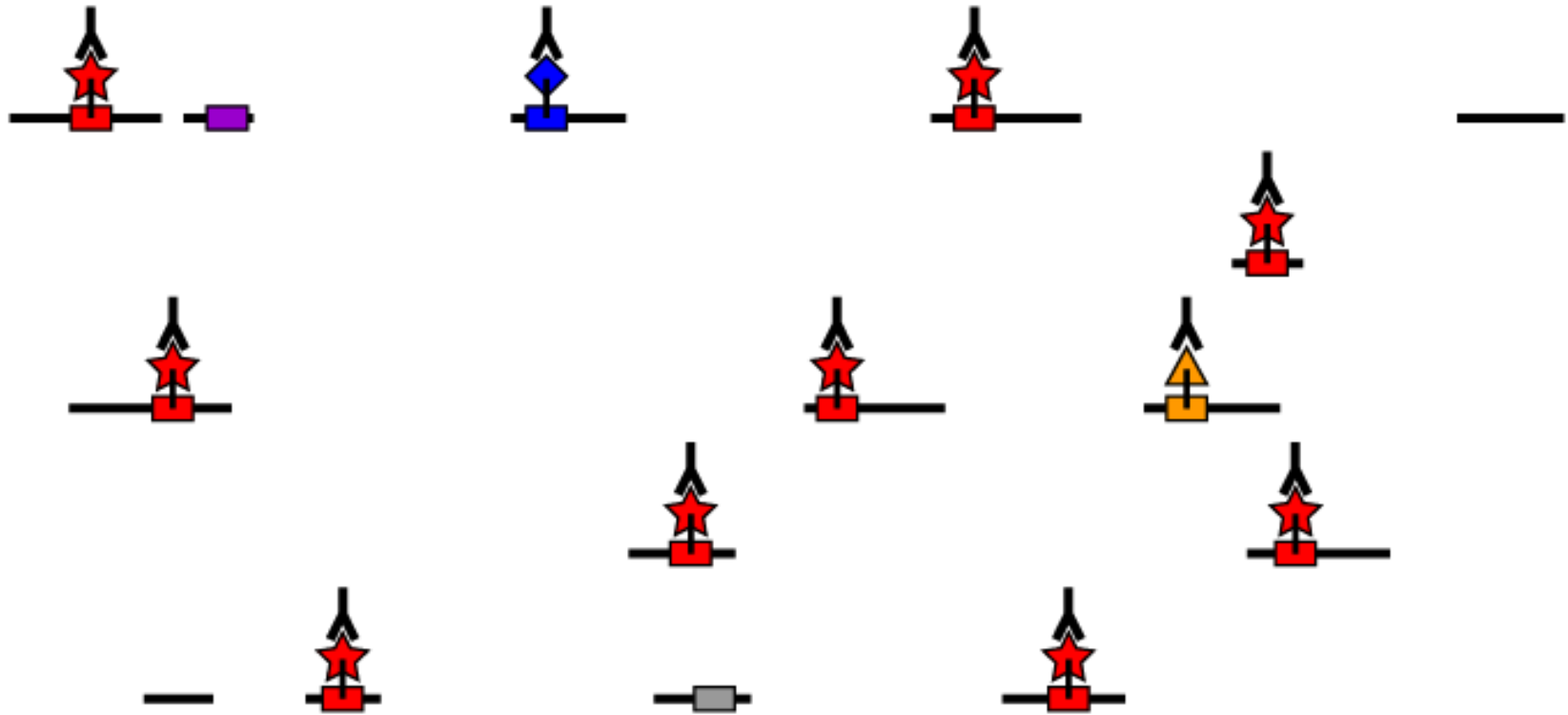
Chop the chromatin using sonication (TF) or micrococcal nuclease (MNase) digestion (histone)



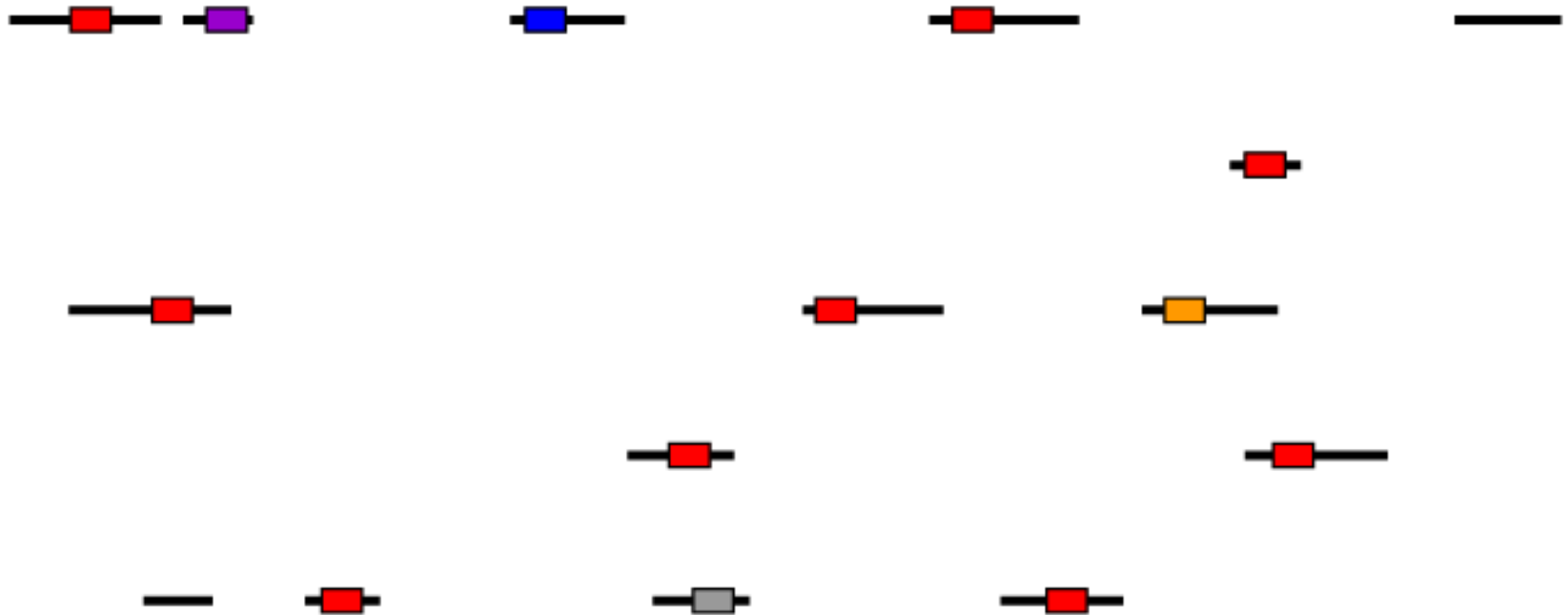
Specific factor-targeting antibody



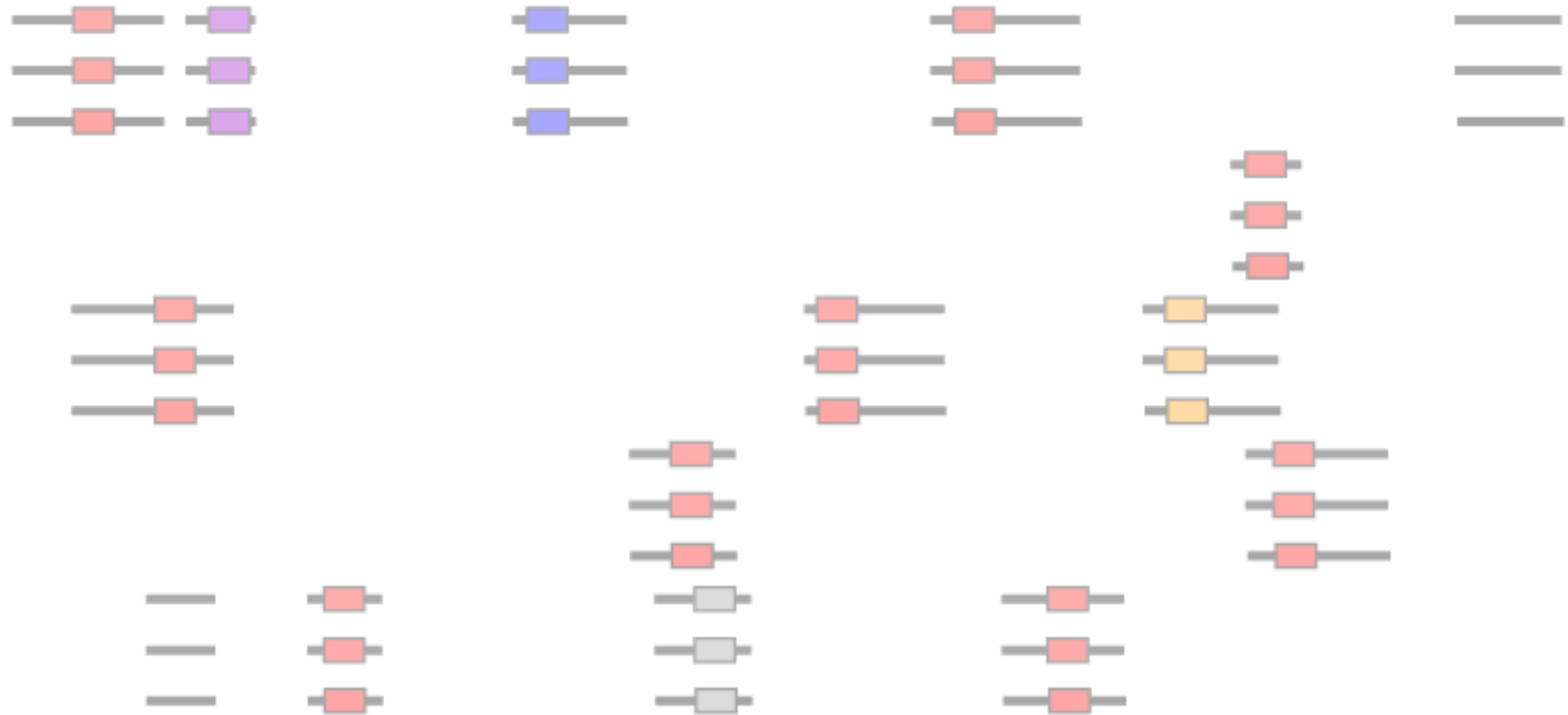
Immunoprecipitation



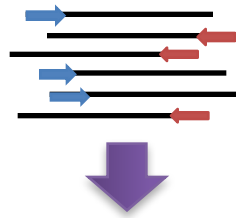
DNA purification



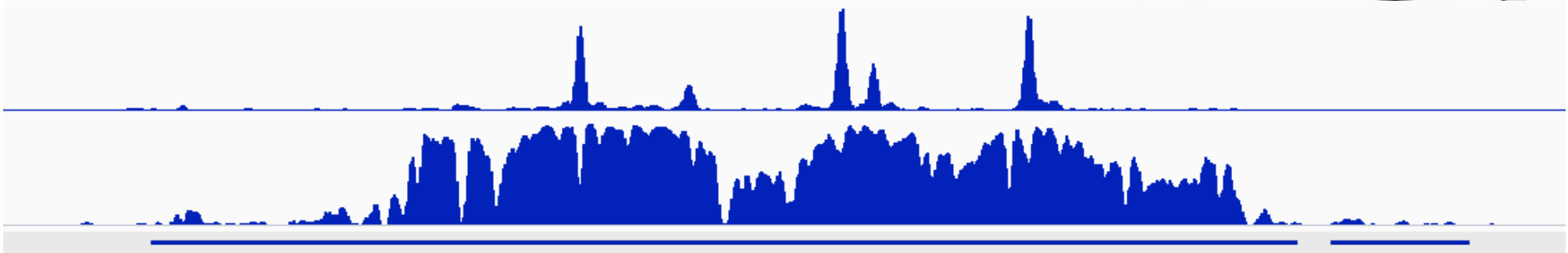
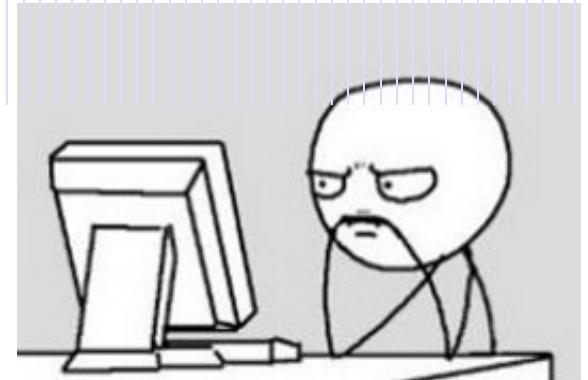
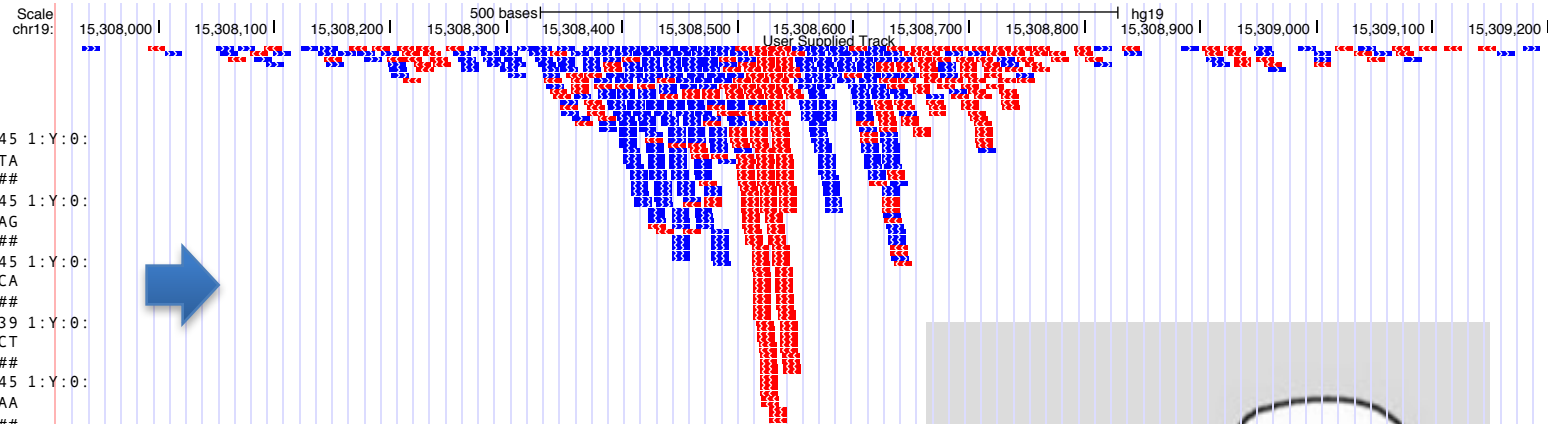
PCR amplification and sequencing



ChIP-seq data analysis overview

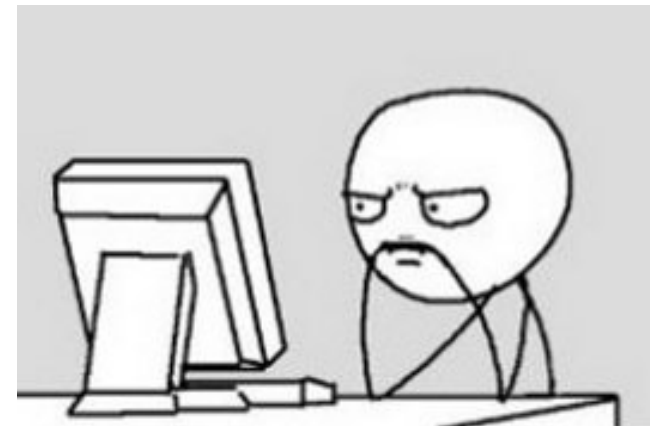


```
@ILLUMINA-8879DC:231:KK:3:1:1070:945 1:Y:0:
NNNAATACAGTCAGAAACATATCATATTGGAGAATA
#####
@ILLUMINA-8879DC:231:KK:3:1:1153:945 1:Y:0:
NNNAAGCACACAGAAGATAACTAAACAATCAAGTAG
#####
@ILLUMINA-8879DC:231:KK:3:1:1222:945 1:Y:0:
NNNAAGGGCTCTTGAGAAGAAATCATTCTGGATGGCA
#####
@ILLUMINA-8879DC:231:KK:3:1:1304:939 1:Y:0:
NNNCCAGGCTCCCGCATTCTCCTGCCTCAGCTTCT
#####
@ILLUMINA-8879DC:231:KK:3:1:1354:945 1:Y:0:
NNNCTCTTCTTAGCTAAACTTTCACTAAGCCAAA
#####
@ILLUMINA-8879DC:231:KK:3:1:1411:932 1:Y:0:
NNNGTAGGACCATTGGCGTTGCGACACAAAAATTT
#####
@ILLUMINA-8879DC:231:KK:3:1:1496:937 1:Y:0:
NNNTTCATCGGGTTGAGAGTCCCTTGTTCATGCA
#####
@ILLUMINA-8879DC:231:KK:3:1:1533:939 1:Y:0:
NNNATTTTCCCGTTCCAGGTCGAATTTCCGCGTT
#####
@ILLUMINA-8879DC:231:KK:3:1:1573:940 1:Y:0:
NNNGGGTGCGCCTTTAGTCCCAGCTACTCAGGAAC
#####
```



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

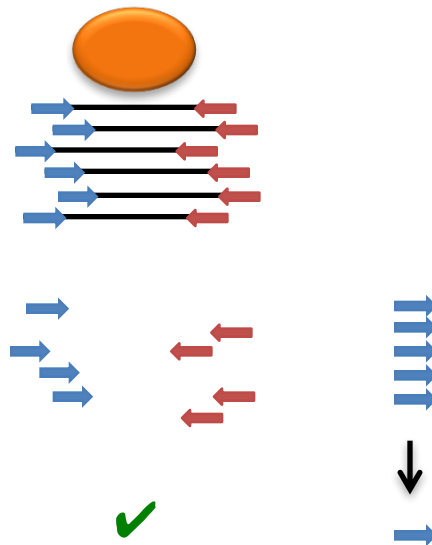


ChIP-seq data analysis: basic processing

- 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:

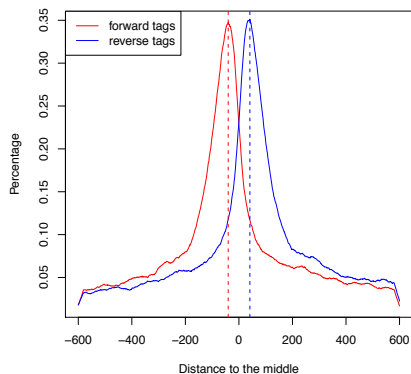
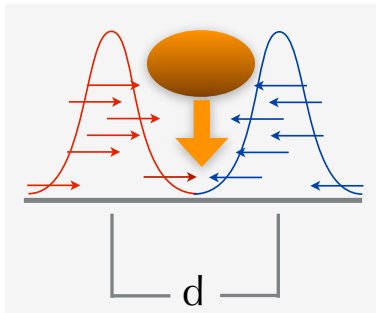


Langmead et al. 2009,
Zang et al. 2009

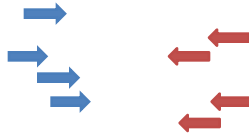
ChIP-seq data analysis: Peak calling

- DNA fragment size estimation
- pile-up profiling

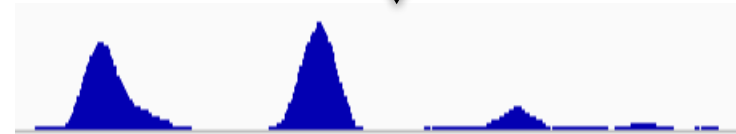
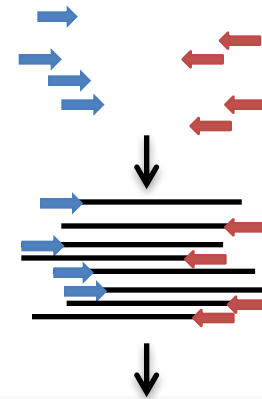
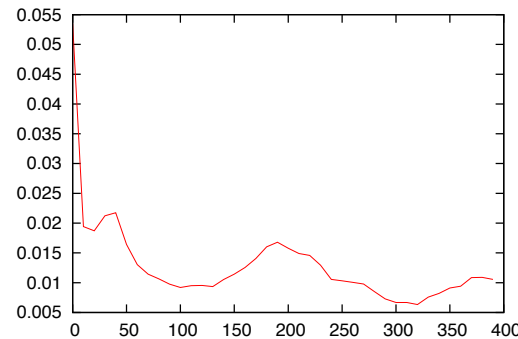
peak model



cross-correlation



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



- Peak/signal detection

ChIP-seq data analysis: Peak calling

- **Sharp peaks**

transcription factor binding,
DNase, ATAC-seq

MACS (Zhang, 2008)

dynamic background

Poisson model

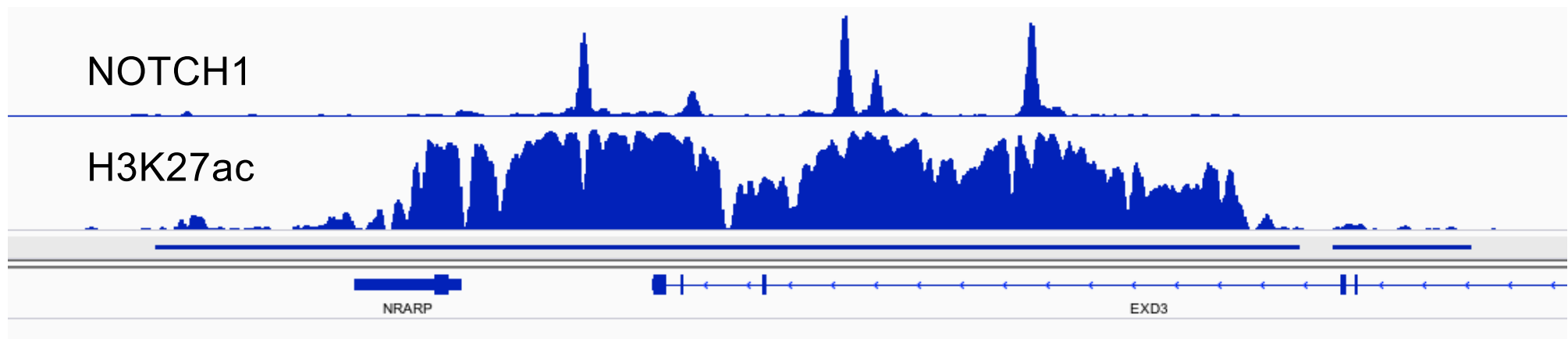
- **Broad peaks**

Histone modifications,
“super-enhancers”

Diffuse

SICER (Zang, 2009)

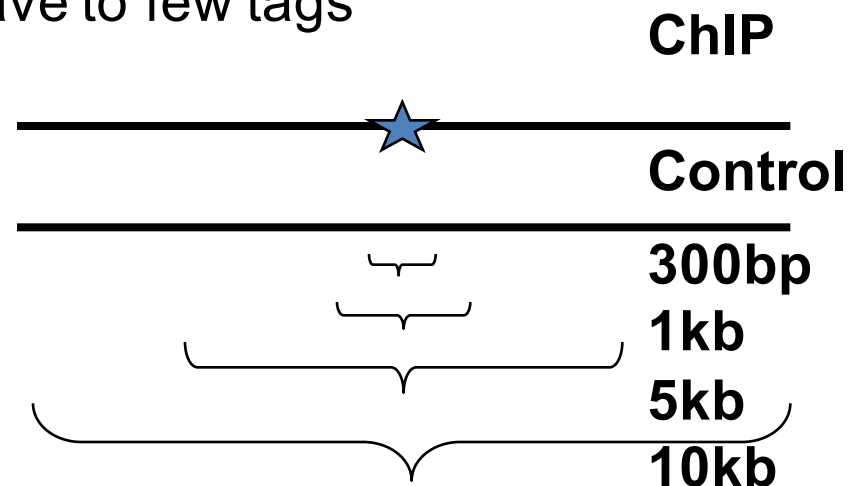
Spatial clustering of localized
weak signal and integrative
Poisson model



MACS

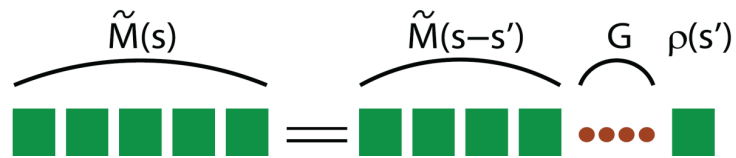
- **Model-based Analysis for ChIP-Seq**
- Tag distribution along the genome $\sim$ Poisson distribution (λ_{BG} = total tag / genome size)
- ChIP-seq show local biases in the genome
 - Chromatin and sequencing bias
 - 200-300bp control windows have too few tags
 - But can look further

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



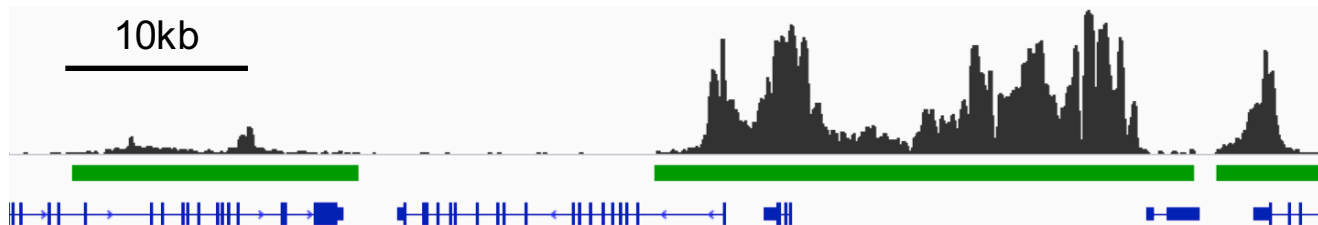
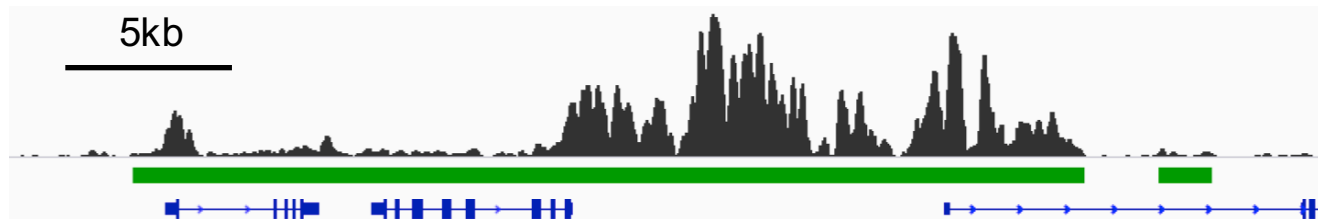
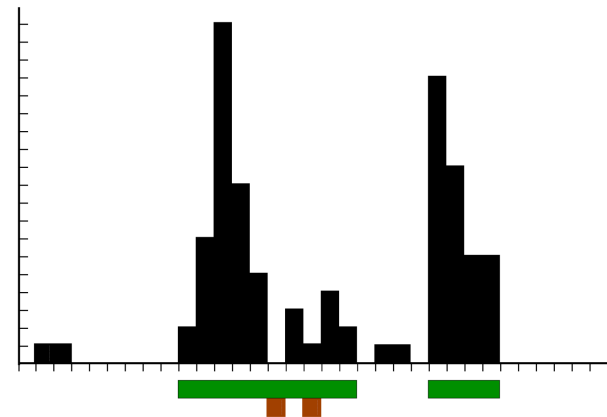
SICER

- **S**patial-clustering **I**dentification of **Ch**IP-**E**nriched **R**egions



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

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



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

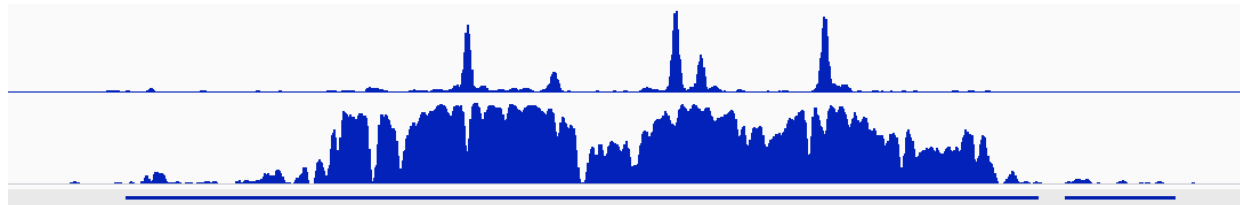
Data formats

- fastq: raw sequences

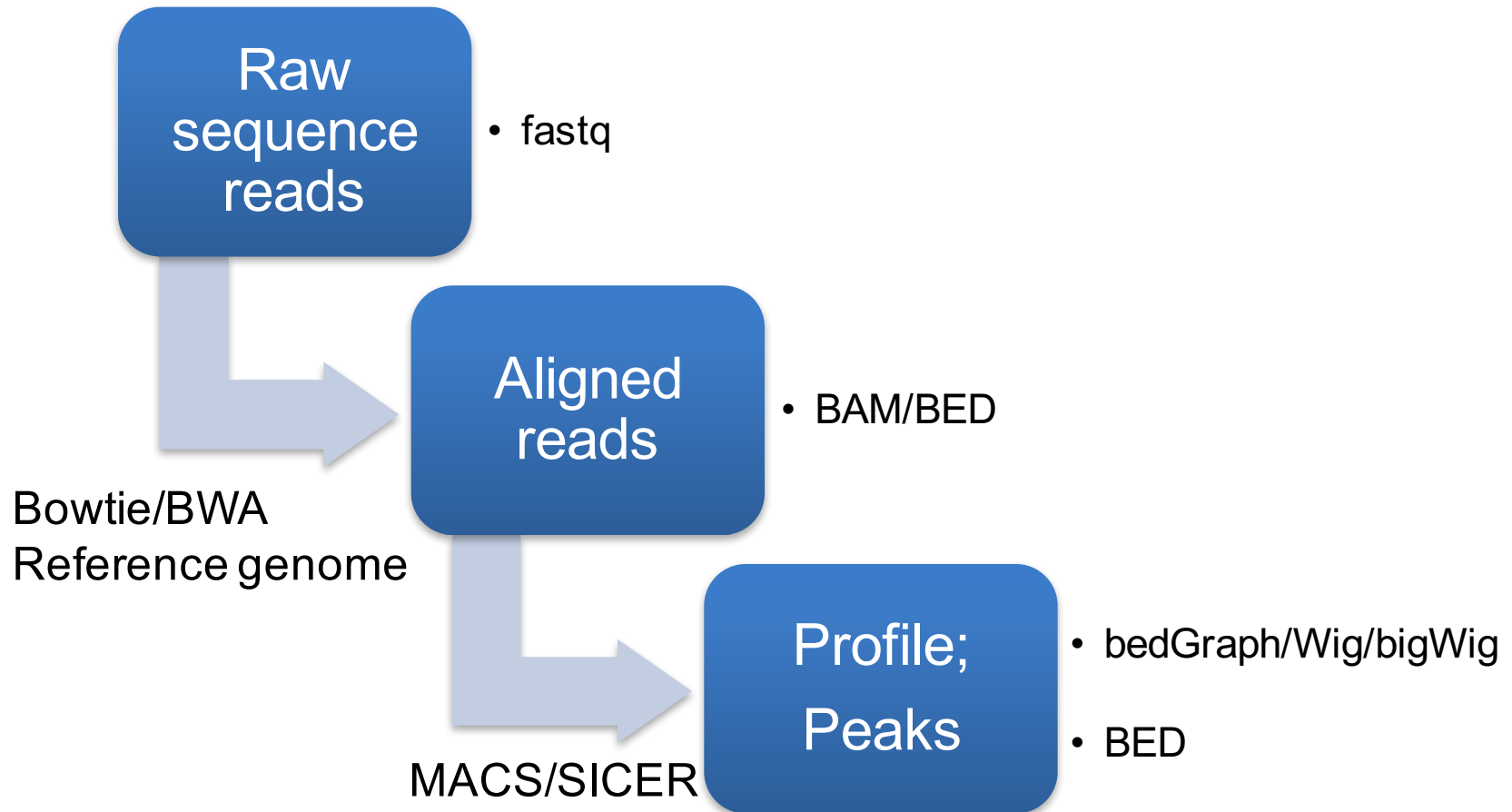
- BED:

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 displays the Galaxy web interface in a browser window. The browser's address bar shows the URL <https://usegalaxy.org/>. The interface features a dark blue header with navigation links: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. A 'Using 0%' indicator is visible in the top right corner.

Tools Panel (Left): Contains a search bar and a list of tools under the 'Get Data' section, including 'Upload File from your computer', 'UCSC Main table browser', 'UCSC Archaea table browser', 'EBI SRA ENA SRA', 'BioMart Ensembl server', 'GrameneMart Central server', 'Flymine server', 'modENCODE fly server', 'modENCODE modMine server', 'MouseMine server', 'Ratmine server', 'YeastMine server', 'modENCODE worm server', 'WormBase server', 'ZebrafishMine server', and 'EuPathDB server'. Other sections include 'Lift-Over', 'Collection Operations', 'Text Manipulation', and 'Datamash'.

Main Content Area (Center): Features a large text block stating: '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 text is a logo for 'IMMPORTGALAXY Flow Cytometry & CyTOF Analysis' and a row of seven colored dots.

History Panel (Right): Displays a search bar for datasets and a message: 'Unnamed history (empty)'. A blue box contains the text: 'This history is empty. You can [load your own data](#) or [get data from an external source](#)'.

Tweets (Bottom Right): Shows two tweets from the Galaxy Project (@galaxyproject). The first tweet mentions 'G-OnRamp Beta Testers Workshops' and includes a link to 'galaxyproject.org/news/2017-03-g...'. The second tweet mentions 'Early career / student researcher working in...'. Both tweets include 'Embed' and 'View on Twitter' links.

Logos (Bottom): A row of logos for partner institutions: PENN STATE, JOHNS HOPKINS, OREGON HEALTH & SCIENCE, TACC, and CYVERSE.

Run MACS on Cistrome, a Galaxy-based platform

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

The screenshot displays the Cistrome Galaxy web interface. The top navigation bar includes 'Galaxy / Cistrome' and a user profile 'Chongzhi'. The main header shows 'Galaxy / Cistrome' and a navigation menu with 'Analyze Data', 'Workflow', 'Shared Data', 'Lab', 'Visualization', 'Help', and 'User'. The right side of the header indicates 'Using 30.3 GB'.

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', and 'Liftover / Others'. Below these are 'GALAXY TOOLBOX' links for 'Get Data', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', and 'Fetch Sequences'.

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 'No file chosen' message, and a 'URL/Text:' section with a text area. Below these is a table for 'Files uploaded via ASPERA:' which is currently empty, showing a message: 'Your ASPERA upload directory contains no files.' At the bottom of the main area is an 'Execute' button.

The right sidebar shows a 'History' section with a list of recent jobs. The jobs are listed in descending order of size, starting with '68: Heatmap log' (329.0 MB) and ending with '55: Heatmap k-means clustered regions'.

File	Size	Date
Your ASPERA upload directory contains no files.		

Job Name	Size	Actions
68: Heatmap log	329.0 MB	View, Edit, Delete
67: Heatmap k-means clustered regions		View, Edit, Delete
66: Heatmap R script		View, Edit, Delete
65: Heatmap image		View, Edit, Delete
64: Heatmap log		View, Edit, Delete
63: Heatmap k-means clustered regions		View, Edit, Delete
62: Heatmap R script		View, Edit, Delete
61: Heatmap image		View, Edit, Delete
60: Heatmap log		View, Edit, Delete
59: Heatmap k-means clustered regions		View, Edit, Delete
58: Heatmap R script		View, Edit, Delete
57: Heatmap image		View, Edit, Delete
56: Heatmap log		View, Edit, Delete
55: Heatmap k-means clustered regions		View, Edit, Delete

Run SICER on Galaxy-based platforms

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

The screenshot shows a web browser window with the URL <http://services.cbib.u-bordeaux.fr/galaxy/>. The page features a dark navigation bar with the 'Galaxy / CBIb' logo and various menu items like 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. On the left, a sidebar lists tool categories such as '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', 'SICER', 'MACS', 'CCAT', 'NGS: Simulation', 'Phenotype Association', 'VCF Tools', 'Assembly tools', 'Prokaryotes Genome Annotation', 'genome alignment', 'Bed Manipulation', 'FASTQ manipulation', and 'Workflows'. The main content area displays the 'cgfb BIOINFORMATIQUE' logo and the title 'Centre de Bioinformatique de Bordeaux'. Below this, a welcome message states: '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'. The maintenance team is identified as 'The CBIb Galaxy maintenance team'. A dashed box highlights a section titled 'What's new in this release !!' which lists updates: 'Prokka', a software tool to rapidly annotate bacterial, archaeal and viral genomes; 'MUMmer', a system for rapidly aligning entire genomes; 'Bedtools', utilities for a wide-range of genomics analysis tasks; and 'MEME', which discovers novel motifs. A 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 bottom right corner of the page displays the number '31'.

Galaxy / CBIb Analyze Data Workflow Shared Data Visualization Help User Using 0 bytes

Tools

- 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
- SICER Statistical approach for the Identification of ChIP-Enriched Regions
- MACS Model-based Analysis of ChIP-Seq
- CCAT Control-based ChIP-seq Analysis Tool
- NGS: Simulation
- Phenotype Association
- VCF Tools
- Assembly tools
- Prokaryotes Genome Annotation
- genome alignment
- Bed Manipulation
- FASTQ manipulation
- Workflows
 - All workflows

cgfb BIOINFORMATIQUE

Centre de Bioinformatique de Bordeaux

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

The **CBIb** Galaxy maintenance team

What's new in this release !!

Prokka, a software tool to rapidly annotate bacterial, archaeal and viral genomes, and produce output files that require only minor tweaking to submit to GenBank/ENA/DDBJ.

MUMmer, a system for rapidly aligning entire genomes, whether in complete or draft form.

Bedtools utilities are a swiss-army knife of tools for a wide-range of genomics analysis tasks. The most widely-used tools enable genome arithmetic: that is, set theory on the genome. For example, bedtools allows one to intersect, merge, count, complement, and shuffle genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF. While each individual tool is designed to do a relatively simple task (e.g., intersect two interval files), quite sophisticated analyses can be conducted by combining multiple bedtools operations on the UNIX command line.

MEME discovers novel, ungapped motifs (recurring, fixed-length patterns) in your sequences (sample output from sequences). MEME splits variable-length patterns into two or more separate motifs. See this Manual for more information.

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

History

search datasets

Unnamed history

0 b

This history is empty. You can [load your own data](#) or [get data from an external source](#)

31

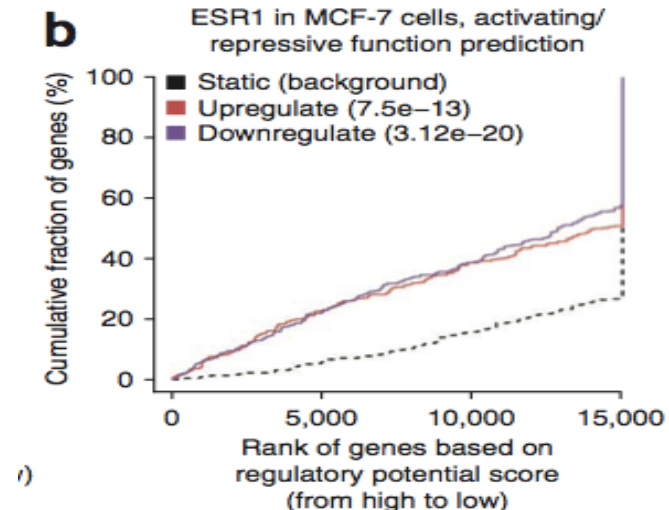
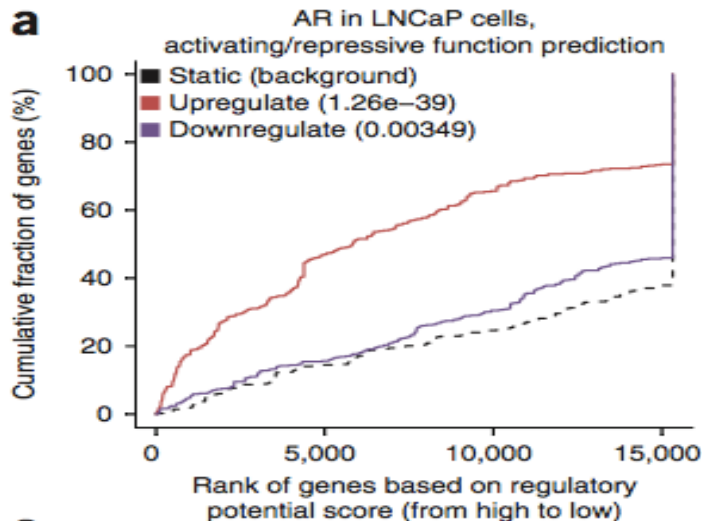
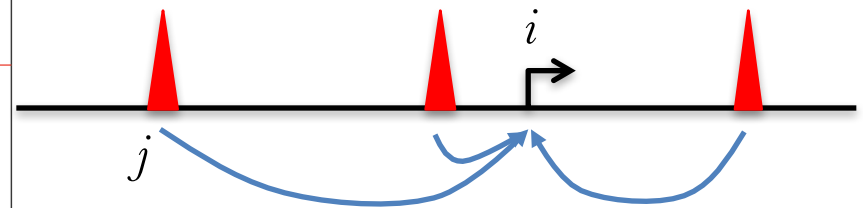
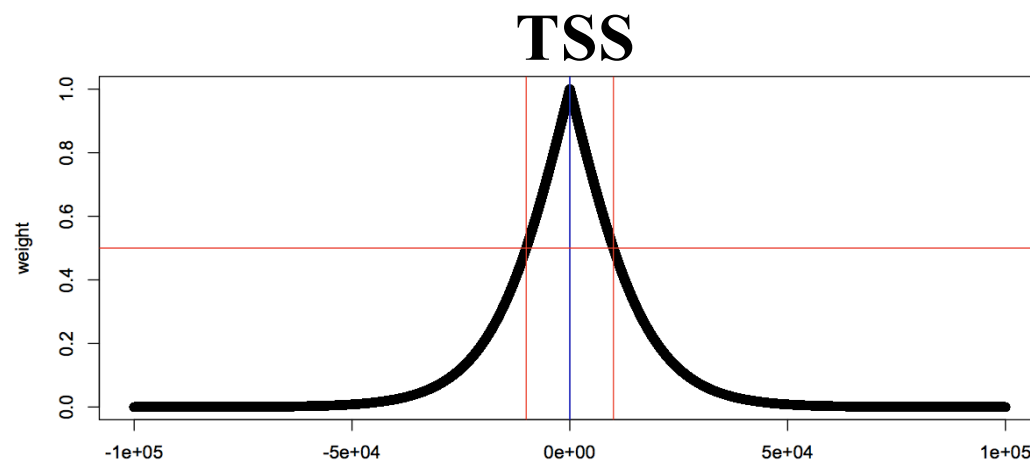
ChIP-seq: Downstream analysis

- 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/>

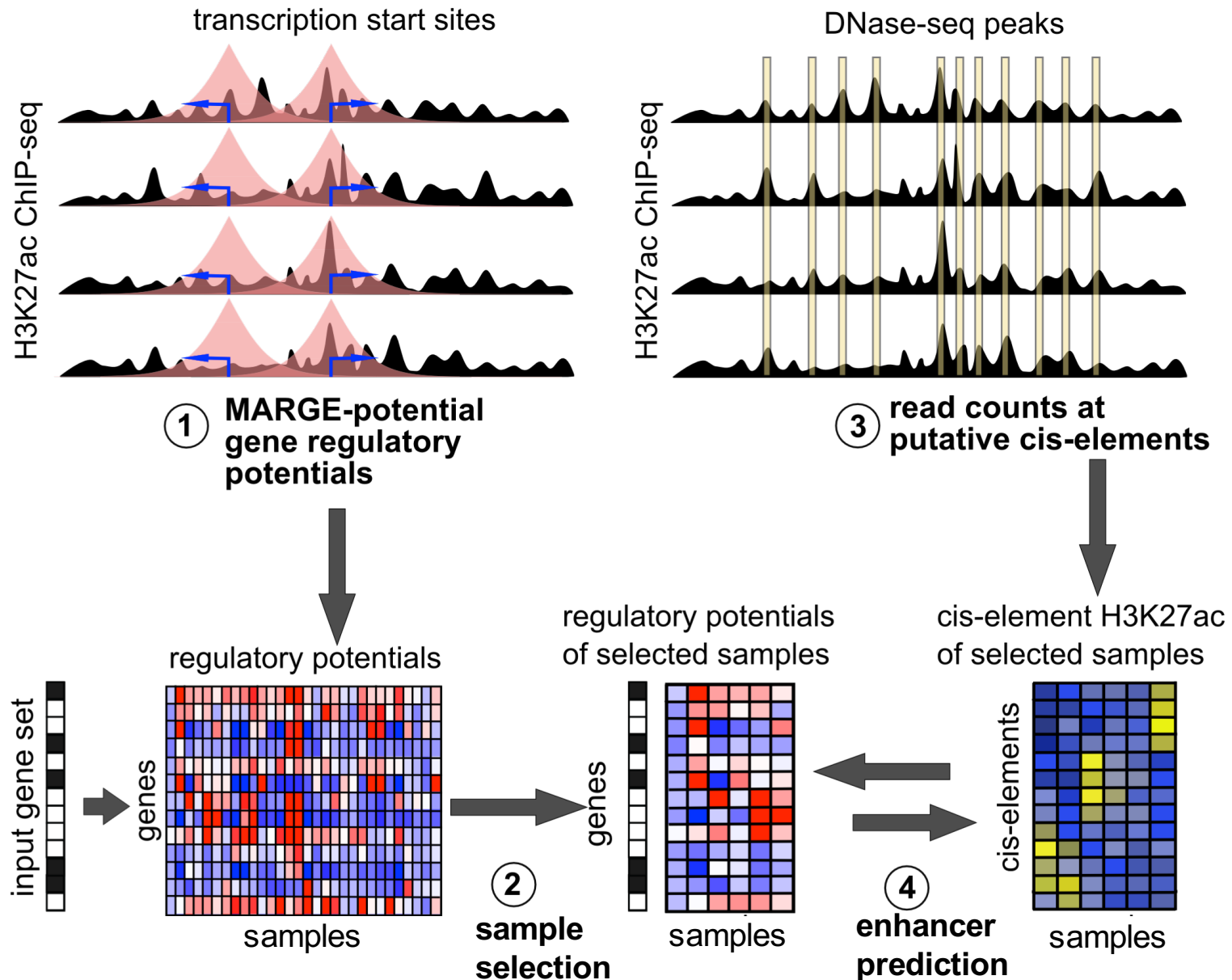
BETA: Binding Expression Target Analysis

- 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

Data

Encyclopedia

Materials & Methods

Help

Search...



[+ See more...](#)

Biosample type

[+ See more...](#)

Organ

[+ See more...](#)

Project

ENCODE	8378
Roadmap	2848
modENCODE	883
modERN	773
GGR	331

Genome assembly (visualization)

hg19	5248
GRCh38	4460
mm10	1335
dm3	602
dm6	408

13213 results



Clear Filters ✕

BIOSAMPLE

23
more

immortalized cell line

[illegible]

tissue

[illegible]

primary cell

[illegible]

whole organisms

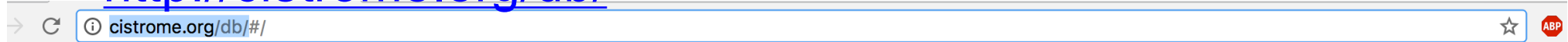
multi-cellular organism	1146	73	50	15
carcass		12	4	4

in vitro differentiated cells

mesenchymal stem cell	61	1	4			2	1						
dendritic cell	11			25									
neuronal stem cell	35	1	4			1	2						
chondrocyte	32					3							

Cistrome Data Browser

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



Dataset Browser

Containing word(s):

Search

Options ▾

Species

All
Homo sapiens
Mus musculus

Biological Sources

All
1015c
10326
1064Sk
106A
10T1/2

Factors

All
ACTB
ADNP
ADNP2
AEBP2
AFF1

Results

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

36
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

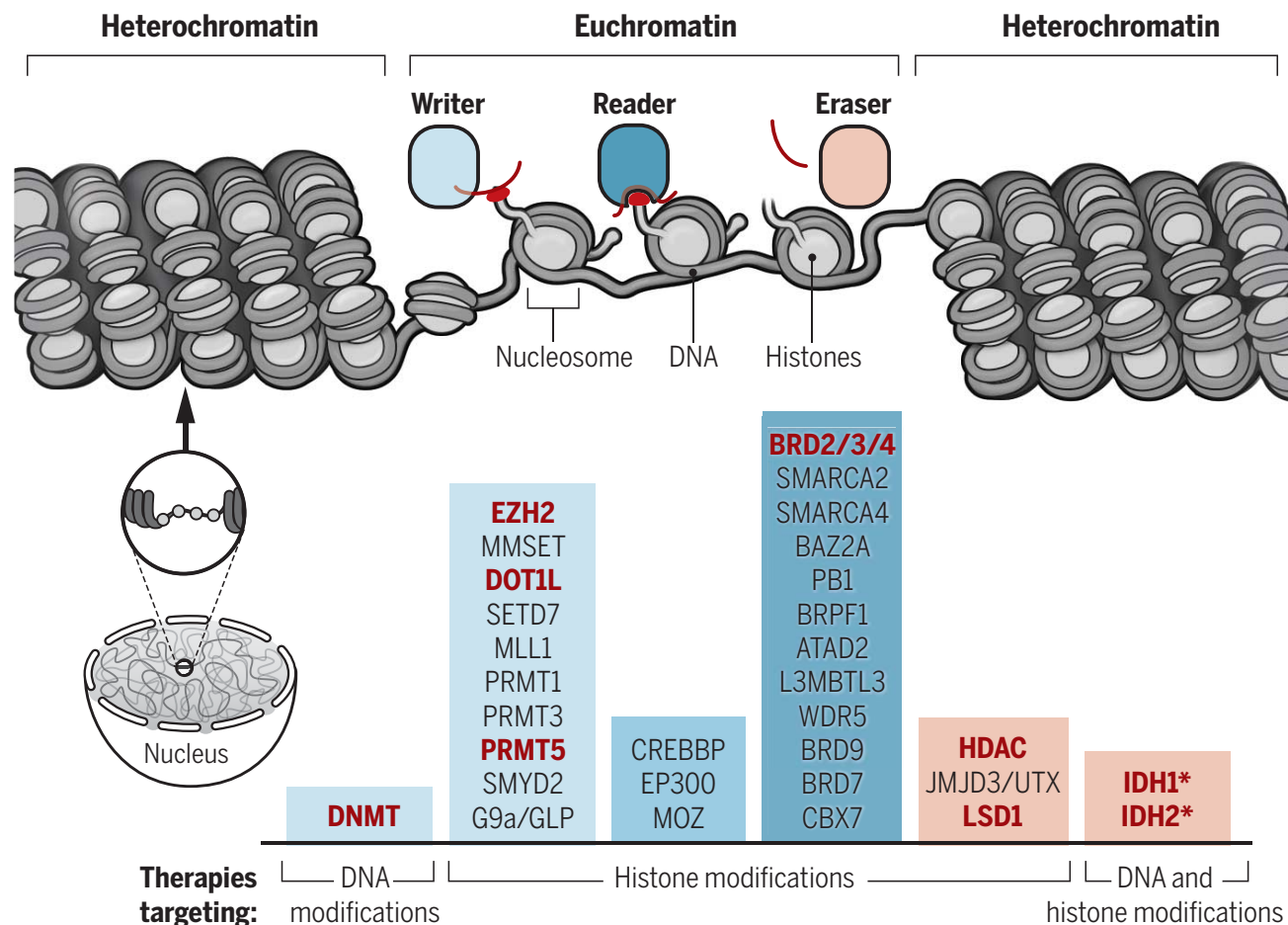
- ChIP-seq is used to profile epigenomes
- ChIP-seq data analysis
 - MACS for narrow peaks
 - SICER for broad peaks
- Online tools and resources

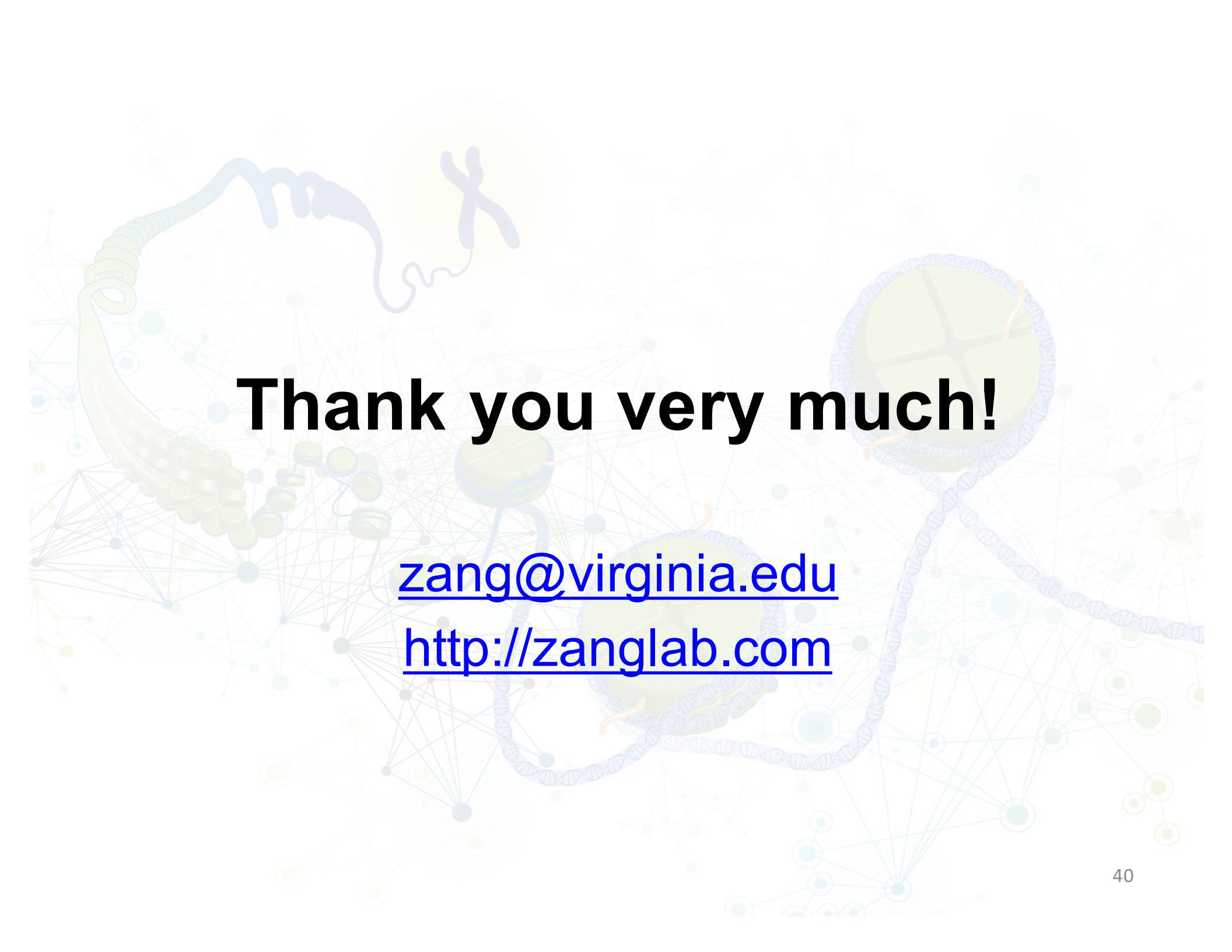
Further Reading

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>



The background features a complex network diagram with numerous nodes and connecting lines. Overlaid on this are several molecular models: a blue ribbon structure on the left, a yellow X-shaped molecule at the top center, and two green spherical structures with internal patterns on the right. A blue double helix structure is also visible at the bottom right.

Thank you very much!

zang@virginia.edu

<http://zanglab.com>