The epigenome: Transcription factors and histone modifications – How to measure them?

Chongzhi Zang

zang@virginia.edu West Complex 6131C
zanglab.org

Recent Advances in Public Health Genomics – Spring 2020
March 2, 2020

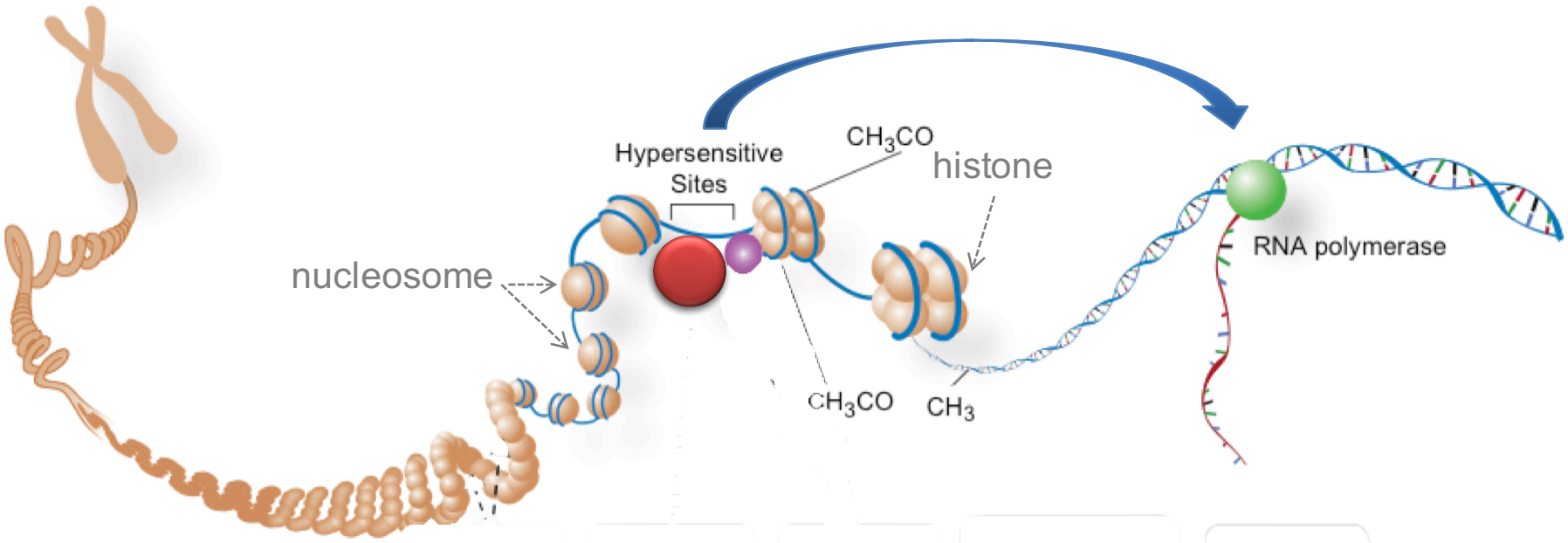
Outline

- Epigenome: an overview
- ChIP-seq technique
- ChIP-seq data analysis
- Future perspective

Outline

- **Epigenome: an overview**
- ChIP-seq technique
- ChIP-seq data analysis
- Future perspective

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)

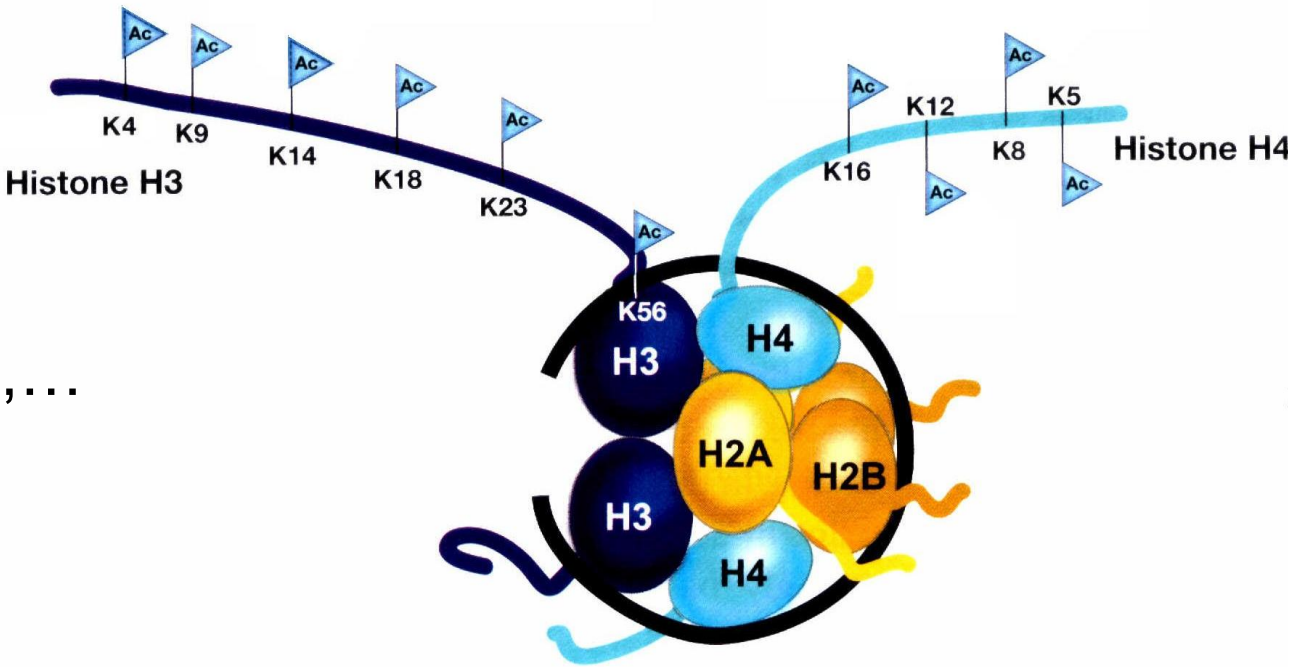
Marks of the epigenome

- DNA methylation
- Histone marks
  - Covalent modifications
  - Histone variants
- Transcription factors
- Chromatin regulators
  - Histone modifying enzymes: writers, readers, erasers
  - Chromatin remodeling complexes (e.g., SWI/SNF)

Histone modifications

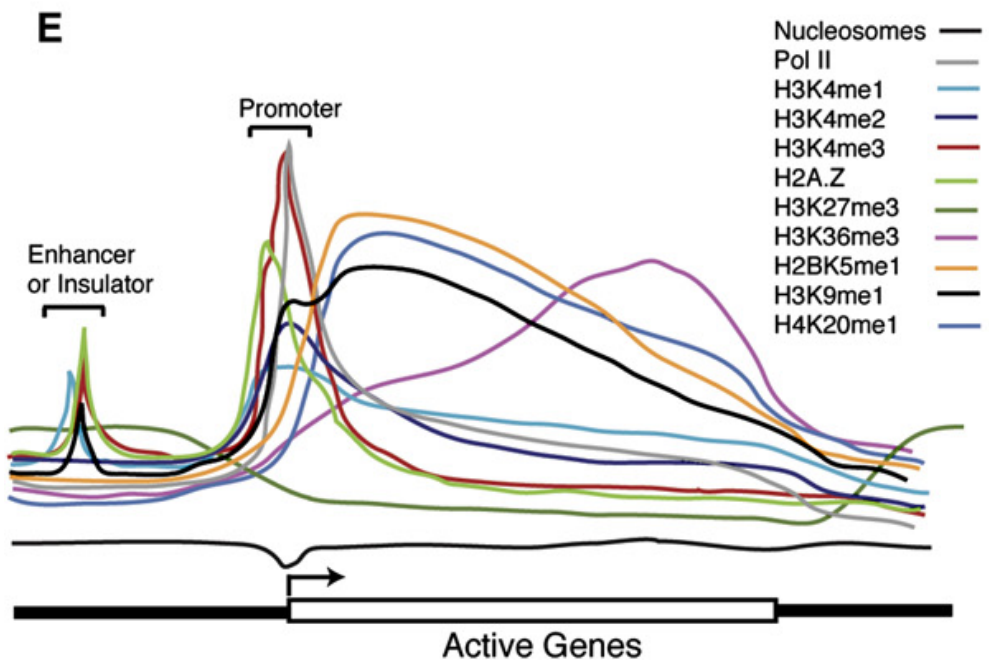
- Nucleosome Core Particles
- Core Histones: H2A, H2B, H3, H4
- Covalent modifications on histone tails include:
  methylation (me),
  acetylation (ac),
  ubiquitylation, …
- Histone variants: H2A.Z, H3.3,…
- Histone modifications are implicated in influencing gene expression.

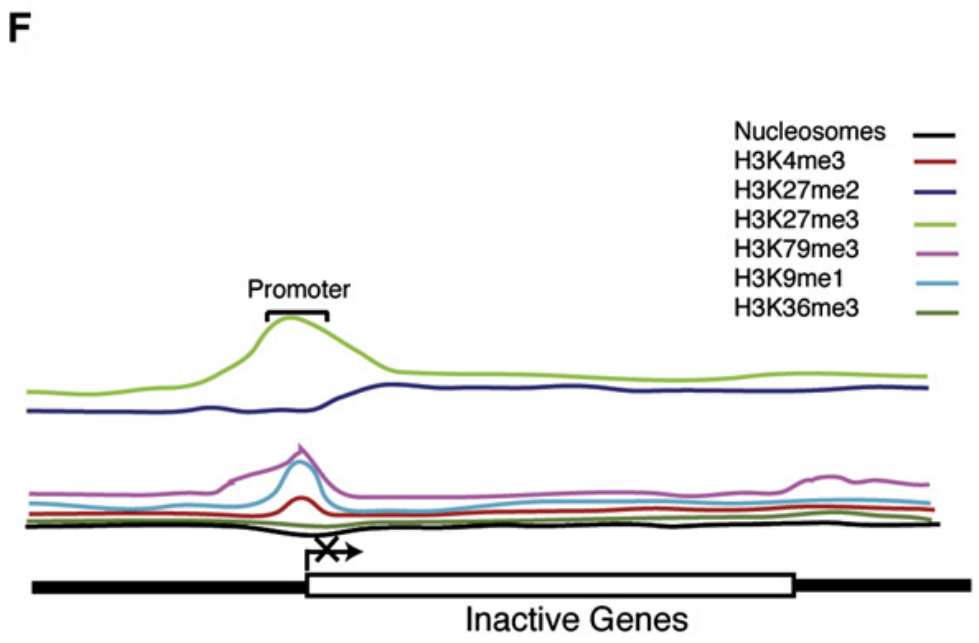
Notation:
H3K4me3



Allis C. *et al. Epigenetics* 2006

Histone modification patterns around genes

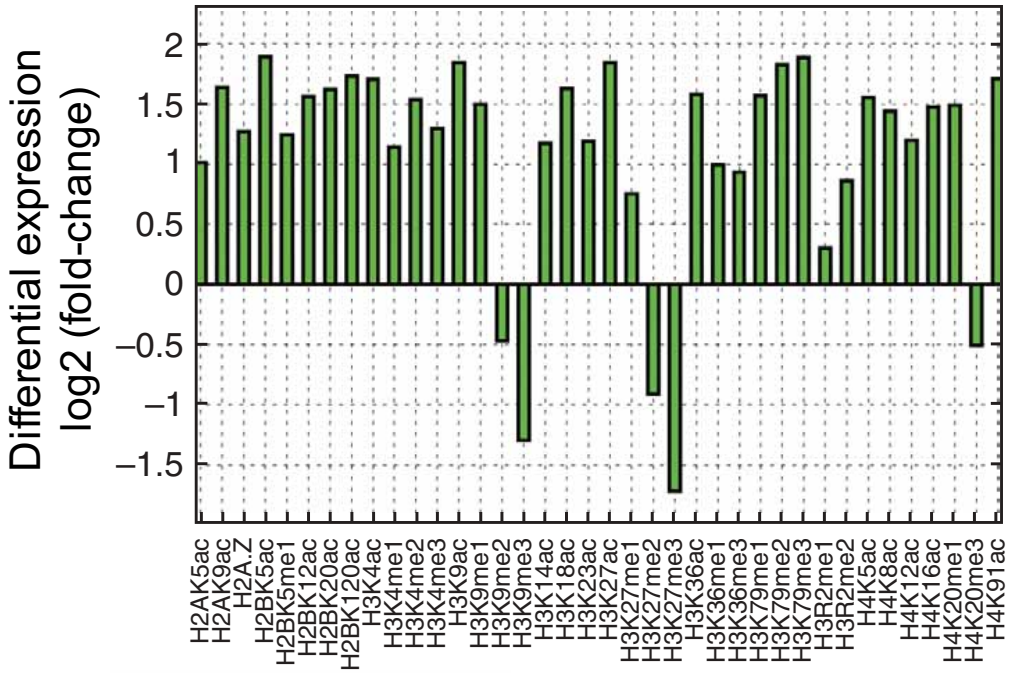




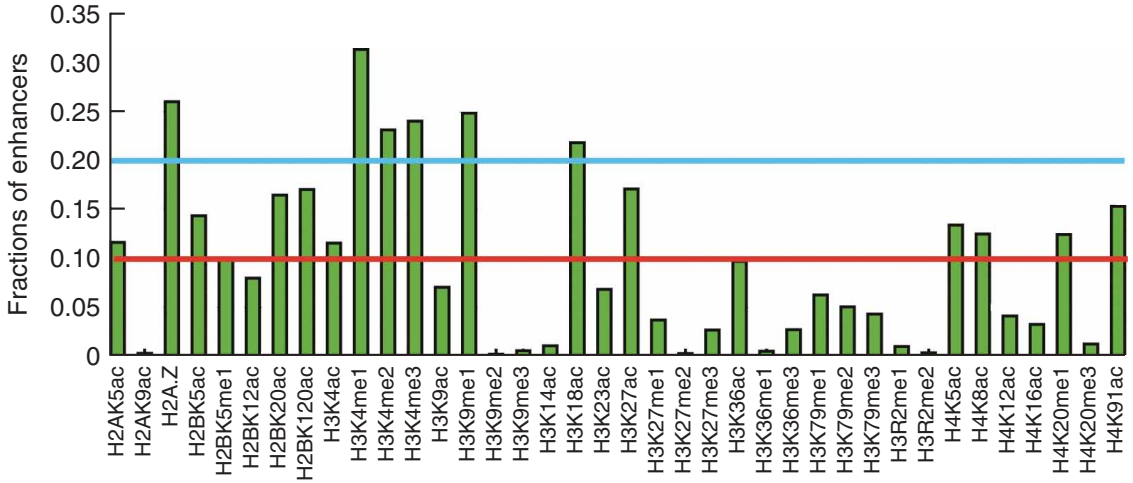
Barski A. *et al. Cell* 2007

Histone modifications associate with regulation of gene expression

Promoters



Putative enhancers



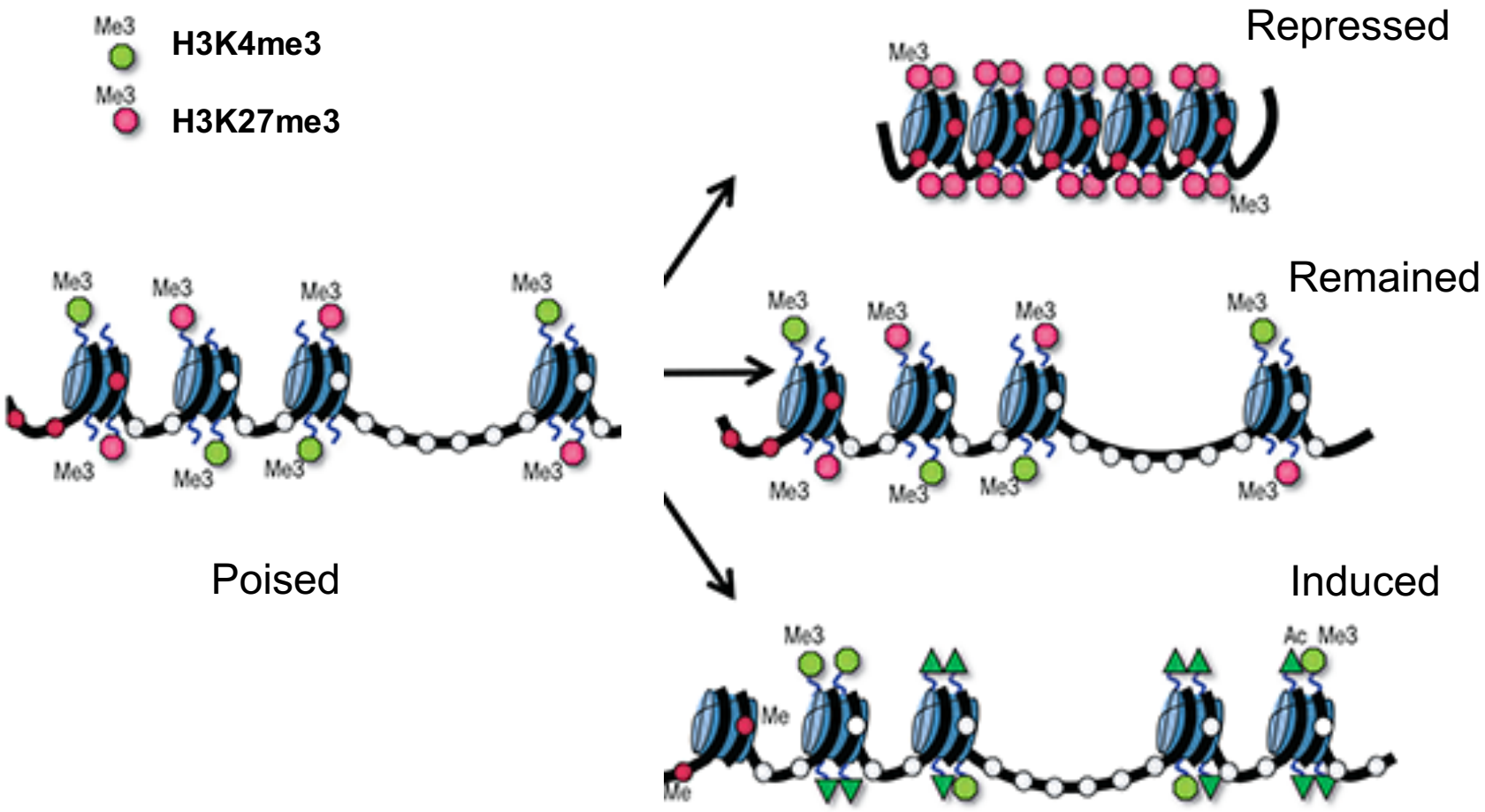
Wang, Zang et al. *Nat Genet* 2008

Functional annotation of common 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

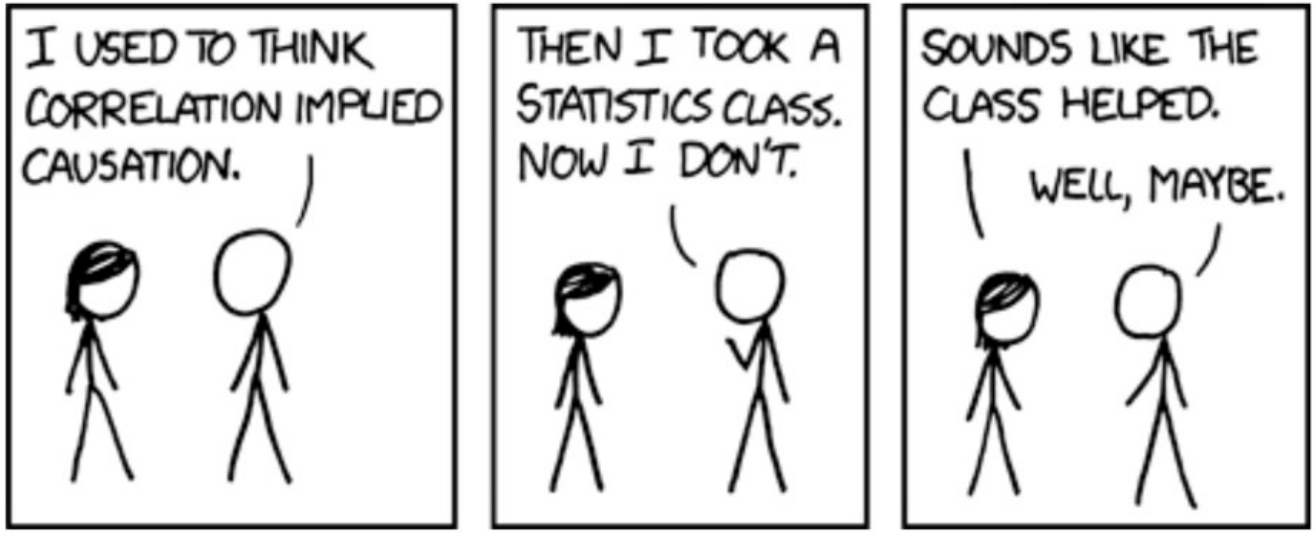
Rivera & Ren *Cell* 2013

H3K4me3/H3K27me3 Bivalent Domain



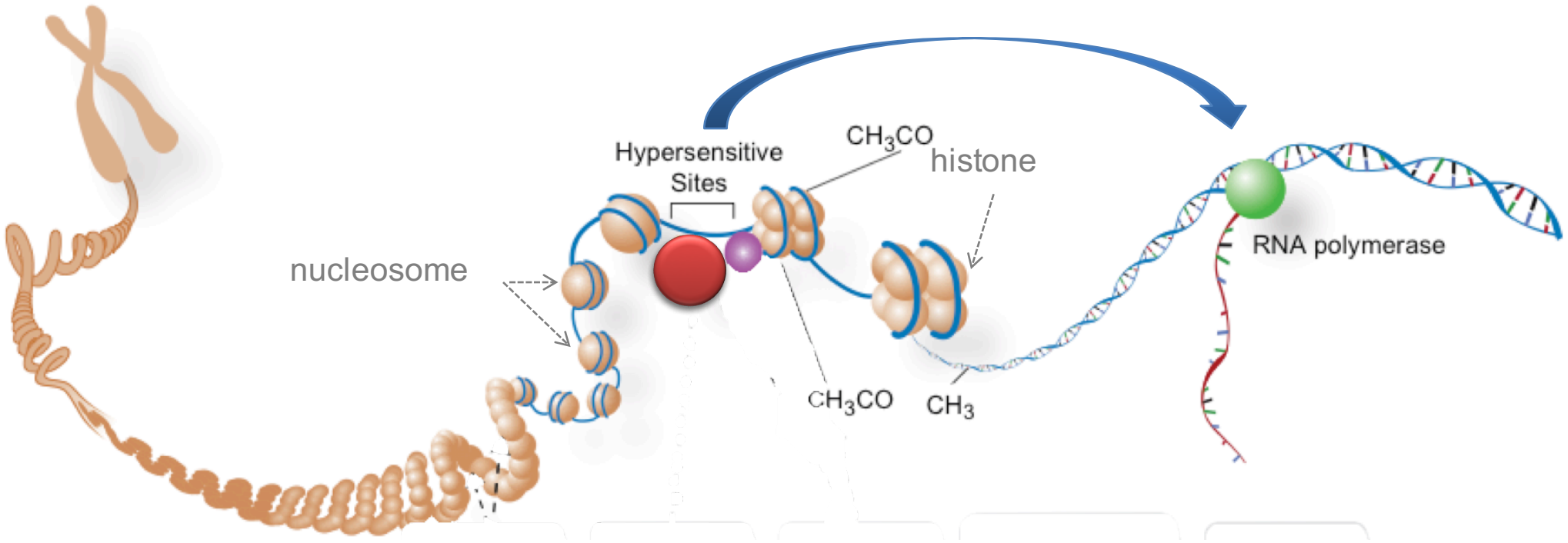
From: https://pubs.niaaa.nih.gov/publications/arcr351/77-85.htm

Correlation ≠ Causation

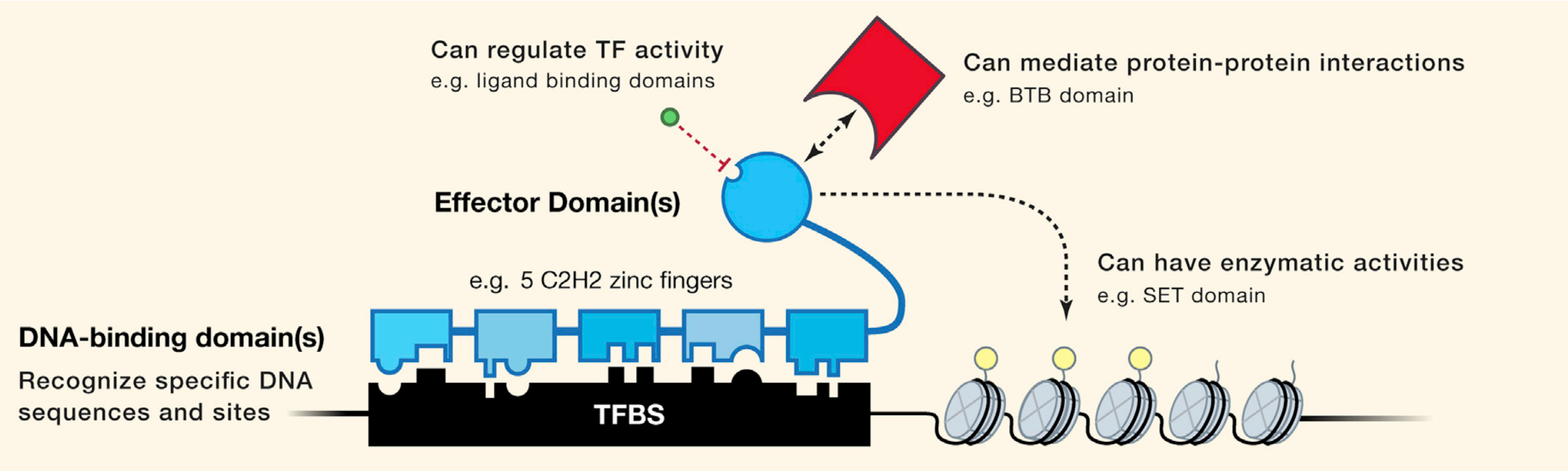


https://xkcd.com/552/

Transcription factors

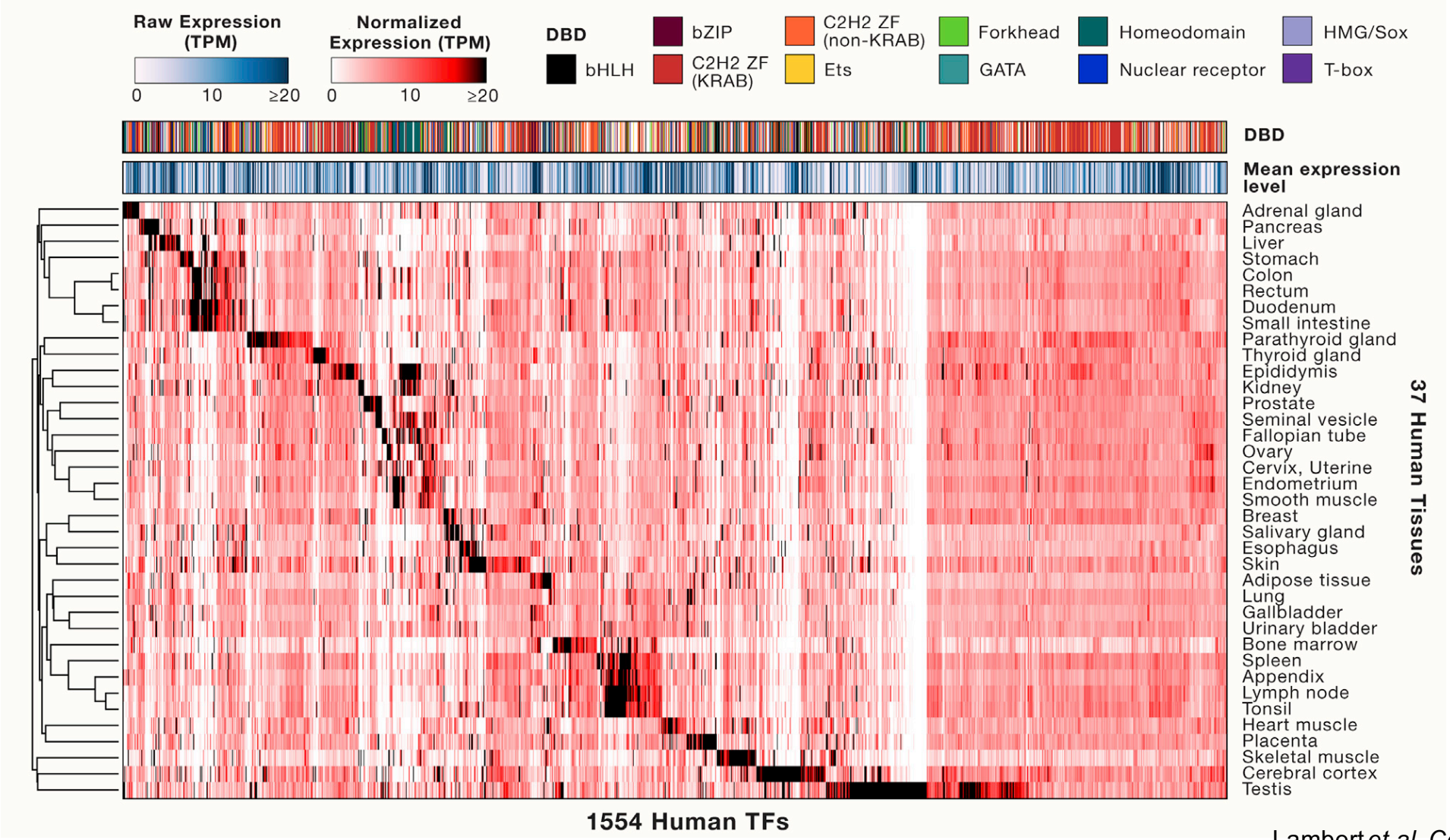


Transcription factors



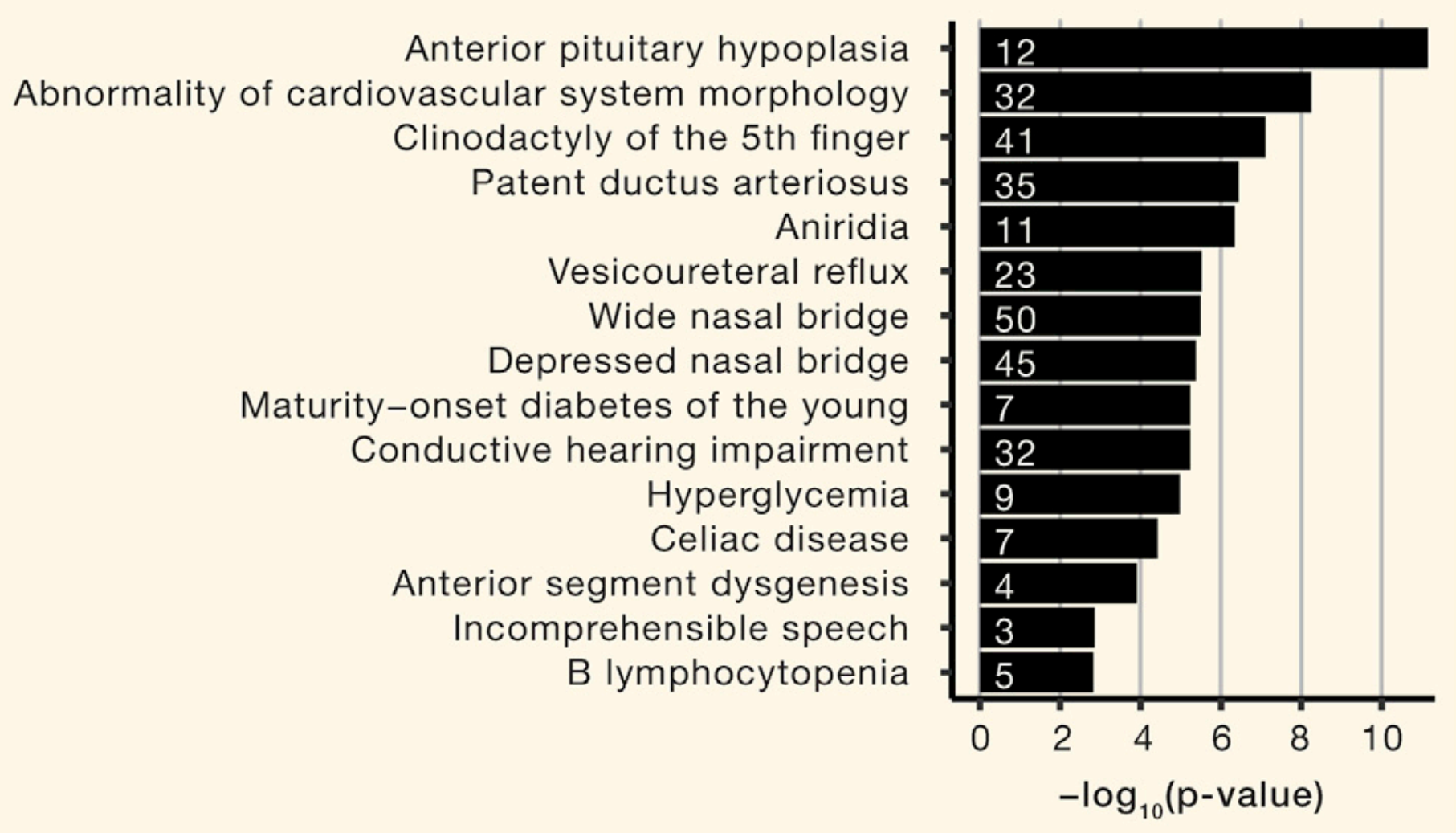
Lambert *et al. Cell* 2018

Many TFs exhibit tissue- and cell-type-specific expression patterns



Lambert *et al. Cell* 2018

TF gene set-associated disease phenotypes



Lambert *et al. Cell* 2018

Characterization of transcription factors

- Structure: Effector domain and DNA binding domain(s)
- Function:
  - Cell-type specific expression
  - Binding DNA sequence (motif)
  - Genome-wide binding sites
  - Target genes
  - Co-factors, etc.

Position weight matrix (PWM) representation of DNA sequence motifs

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

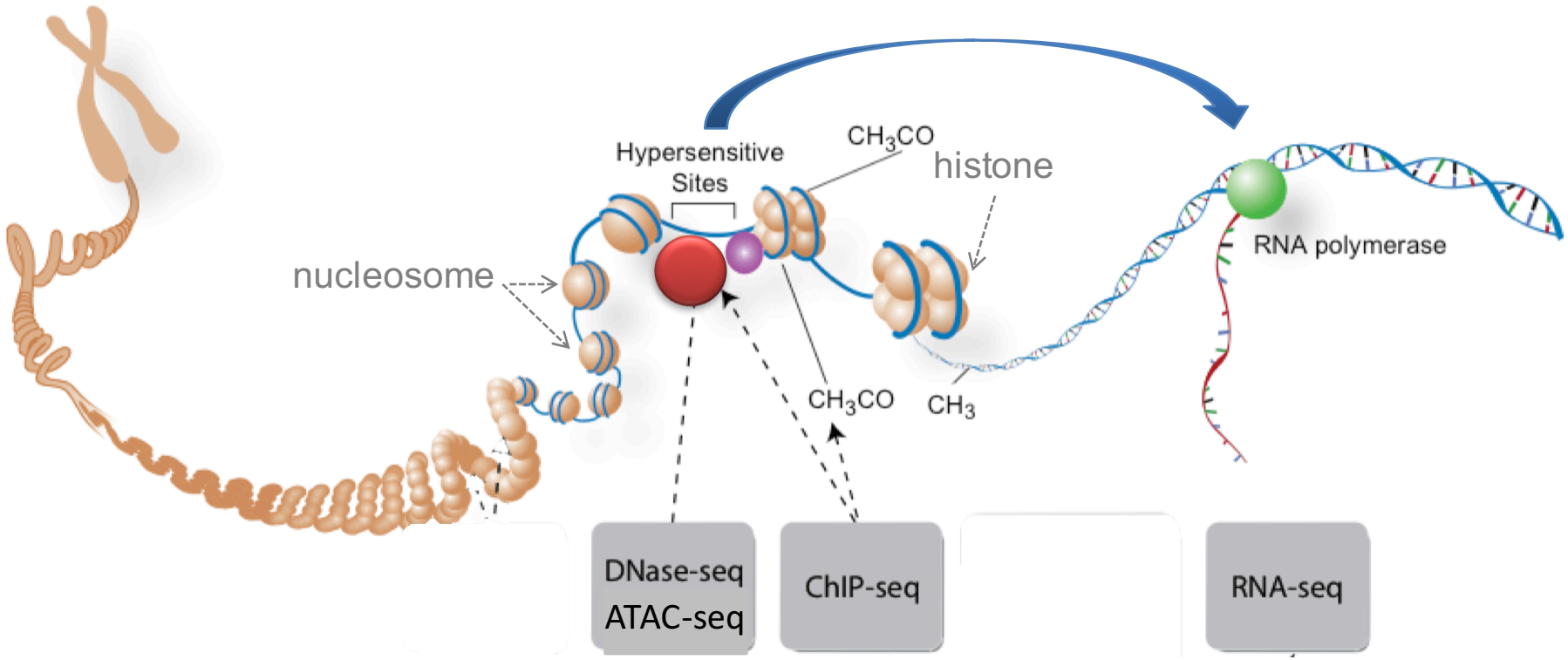
$$R_i = \log_2(4) - H_i$$

$$H_i = -\sum_b f_{b,i} \times \log_2 f_{b,i}$$

Outline

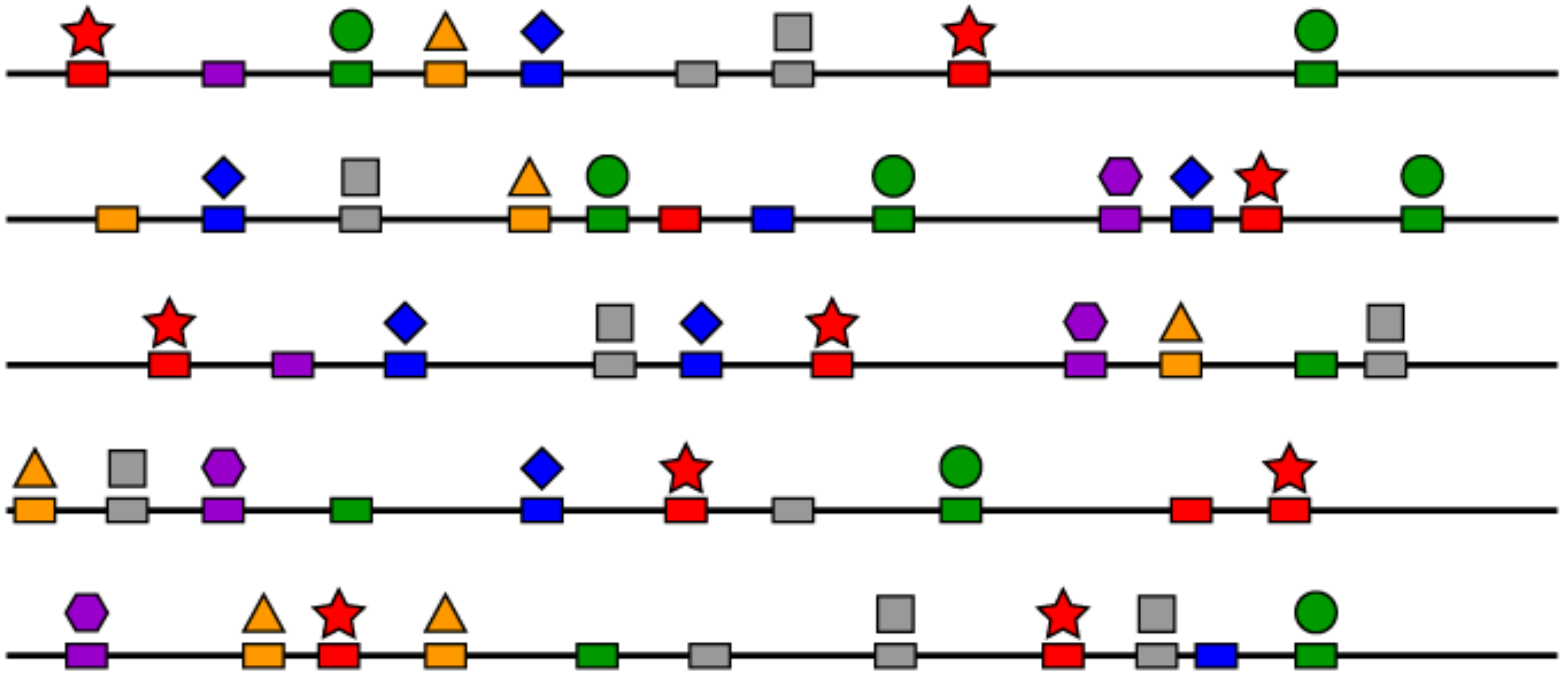
- Epigenome: an overview
- **ChIP-seq technique**
- ChIP-seq data analysis
- Future perspective

ChIP-seq: Profiling epigenomes with sequencing

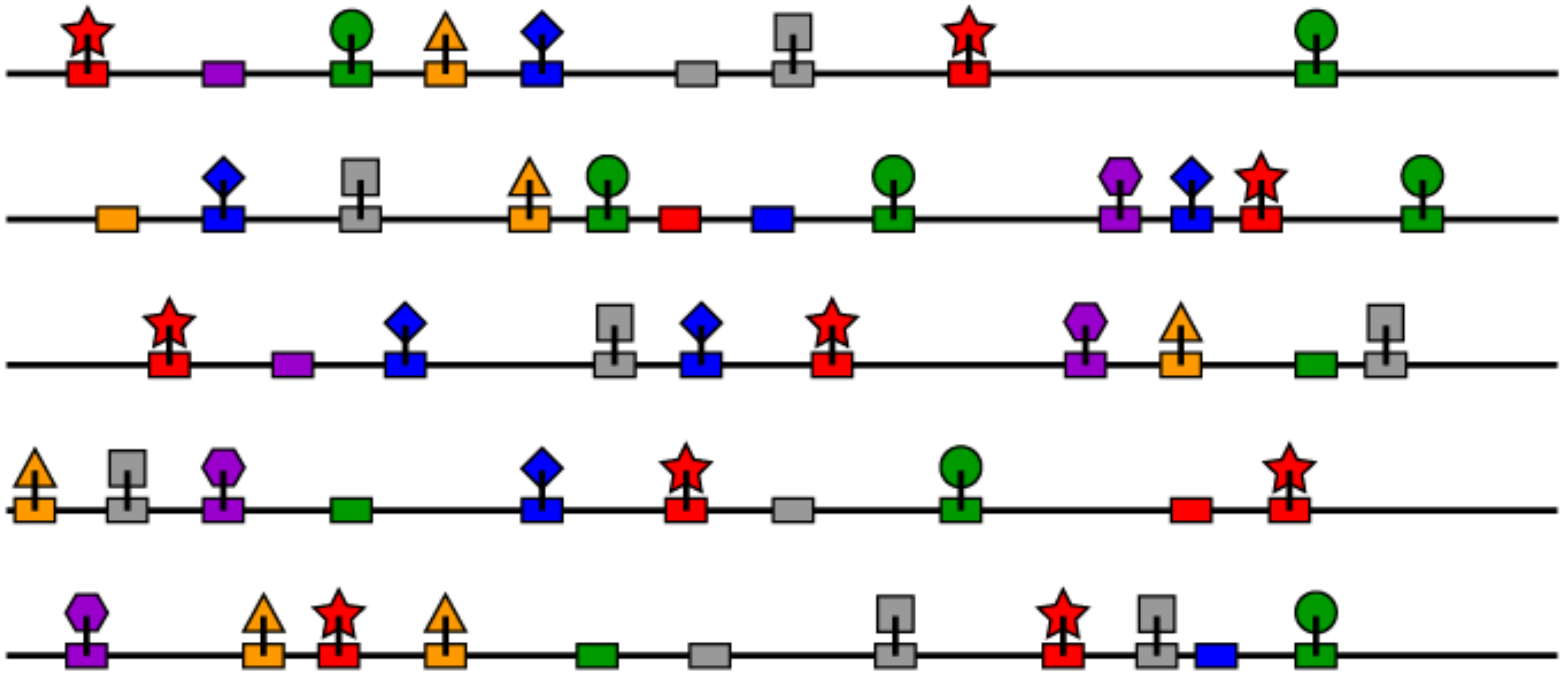


Original figure from ENCODE, Darryl Leja (NHGRI), Ian Dunham (EBI)

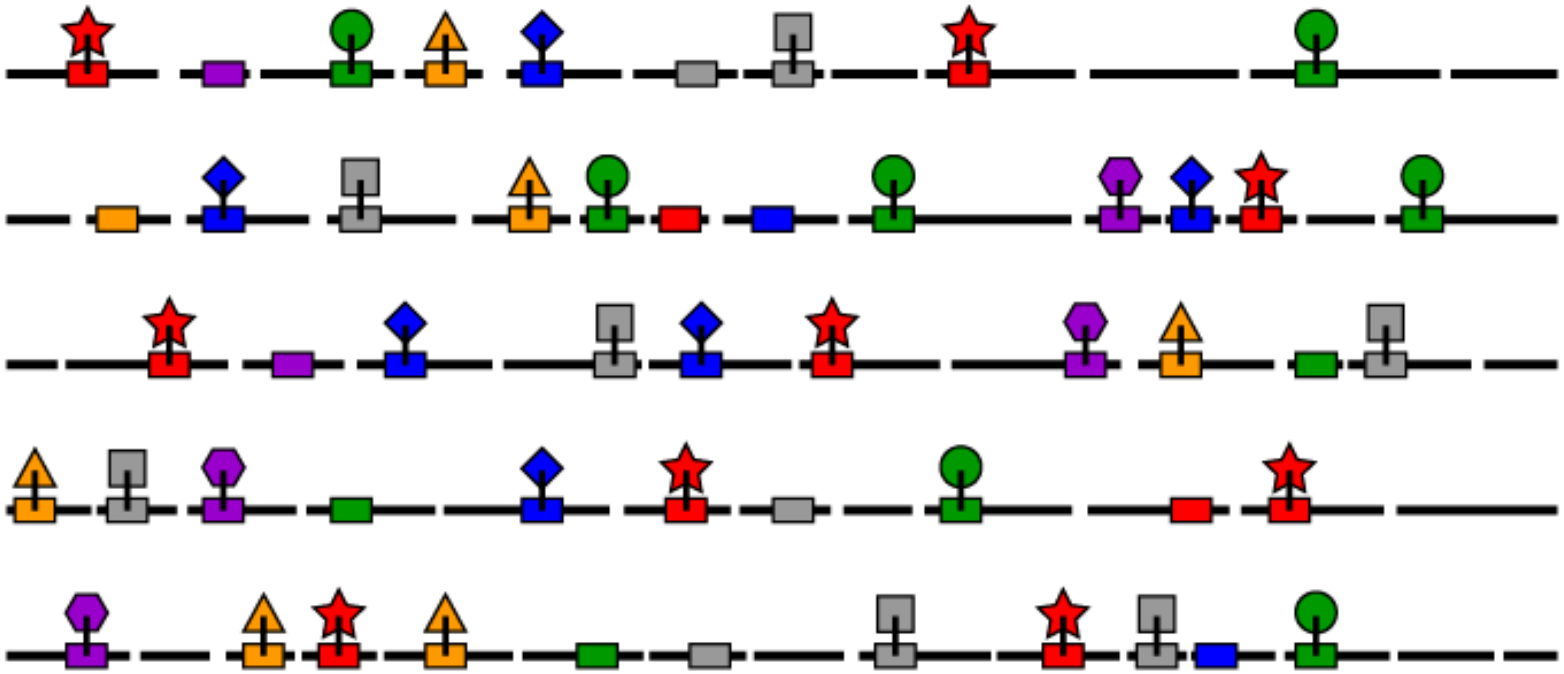
Chromatin ImmunoPrecipitation (ChIP)



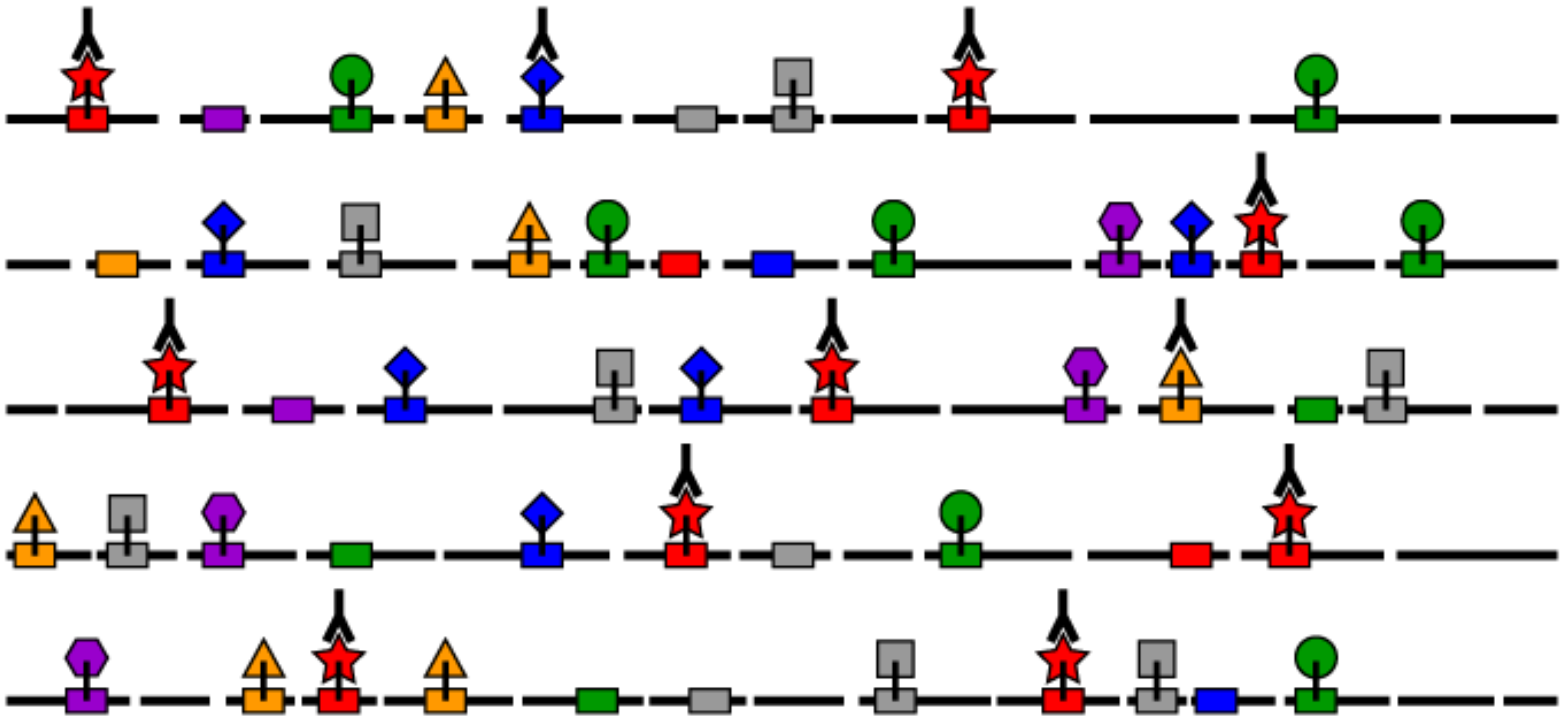
Protein-DNA crosslinking *in vivo* (for TF)



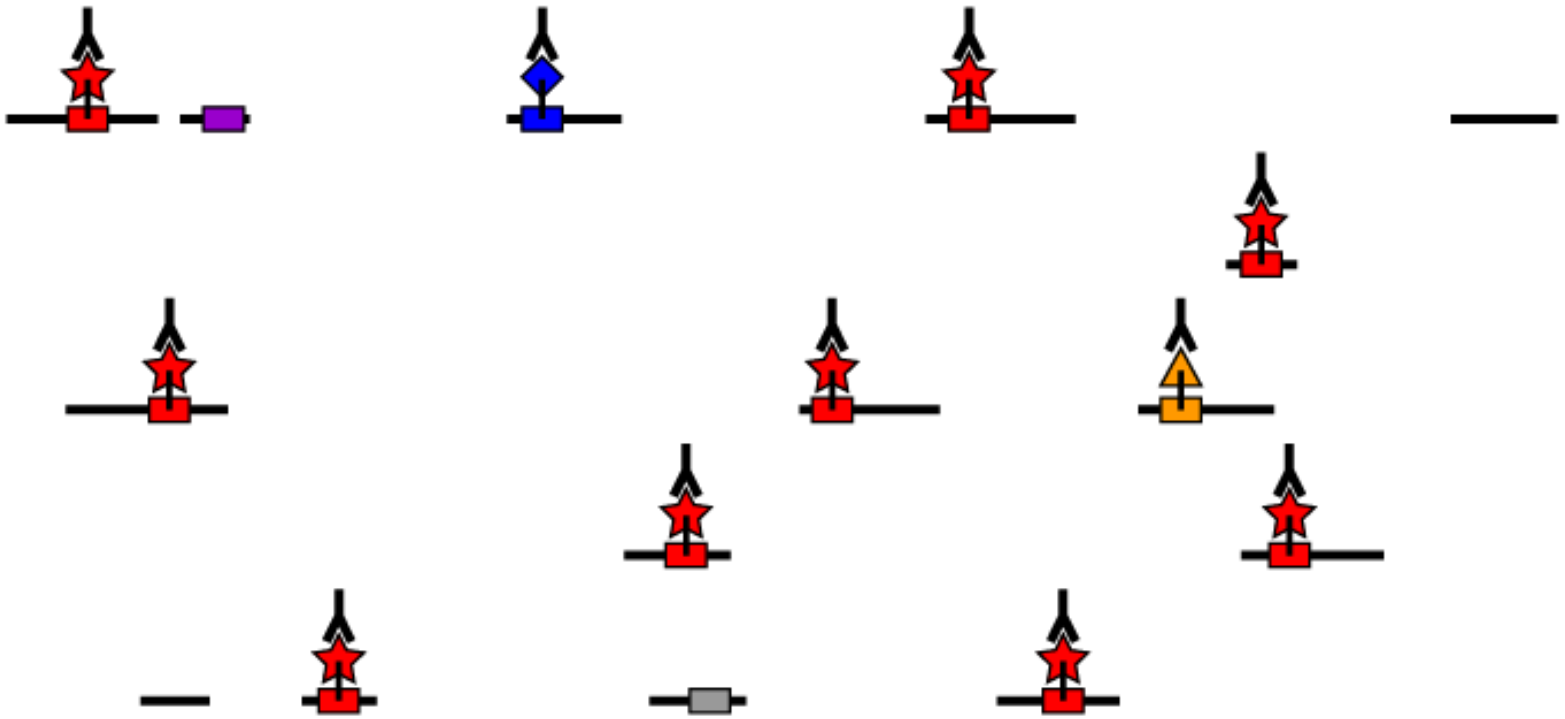
Chop the chromatin using sonication (TF) or micrococal nuclease (MNase) digestion (histone)



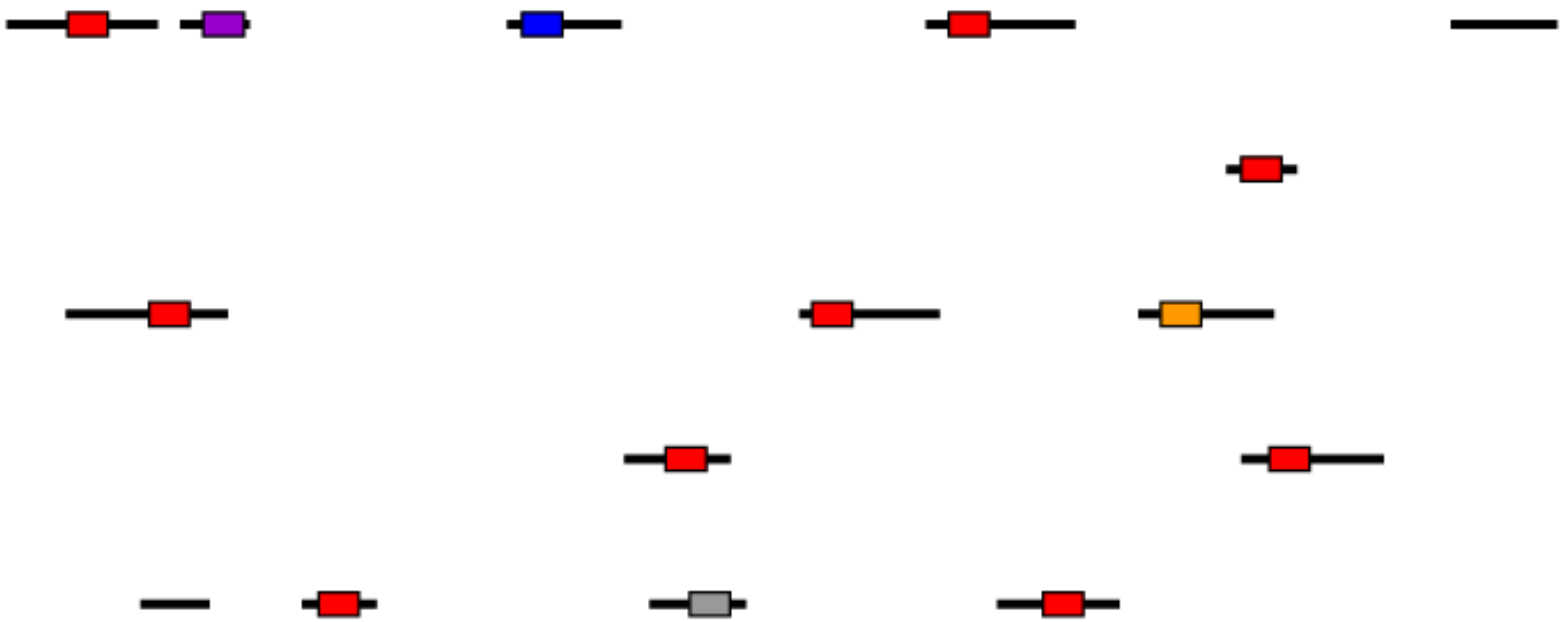
Specific factor-targeting antibody



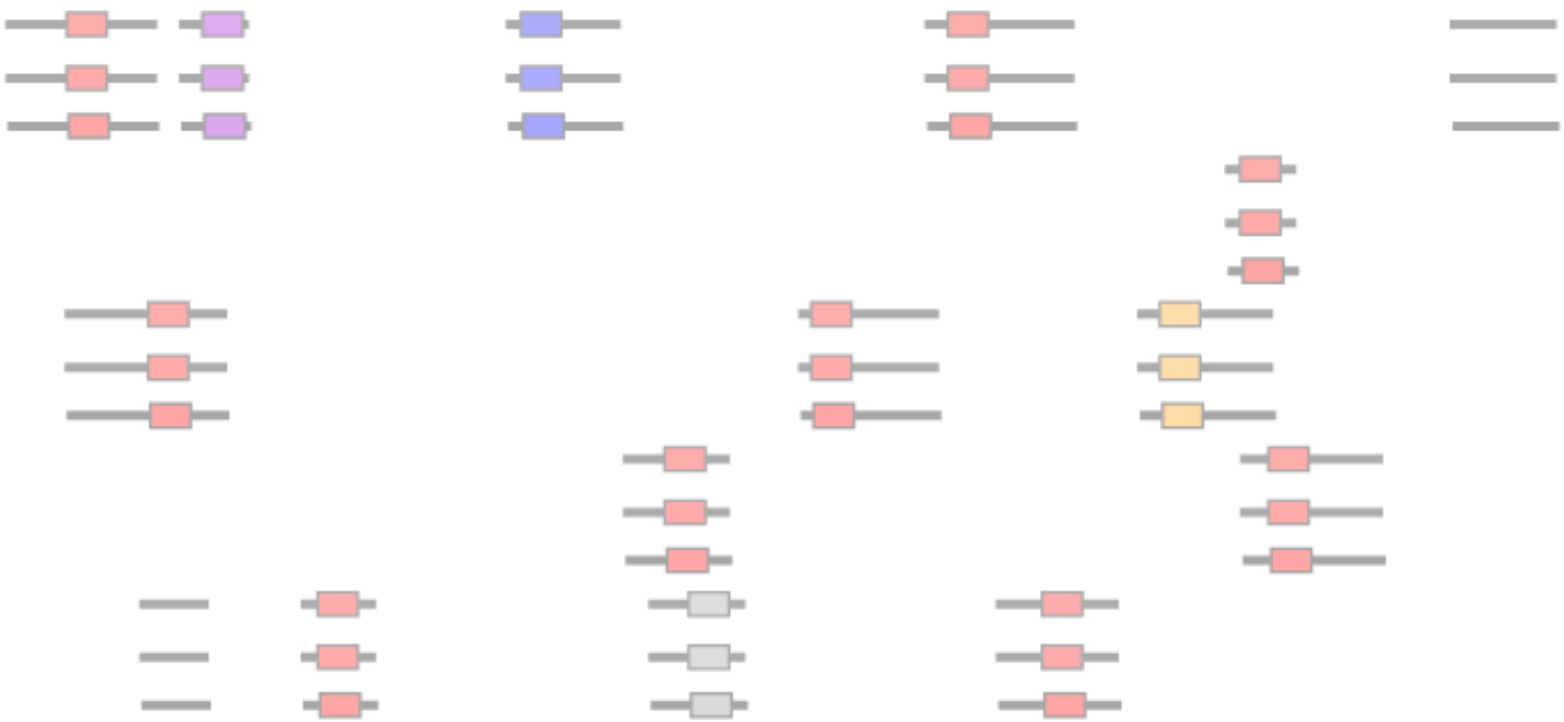
Immunoprecipitation

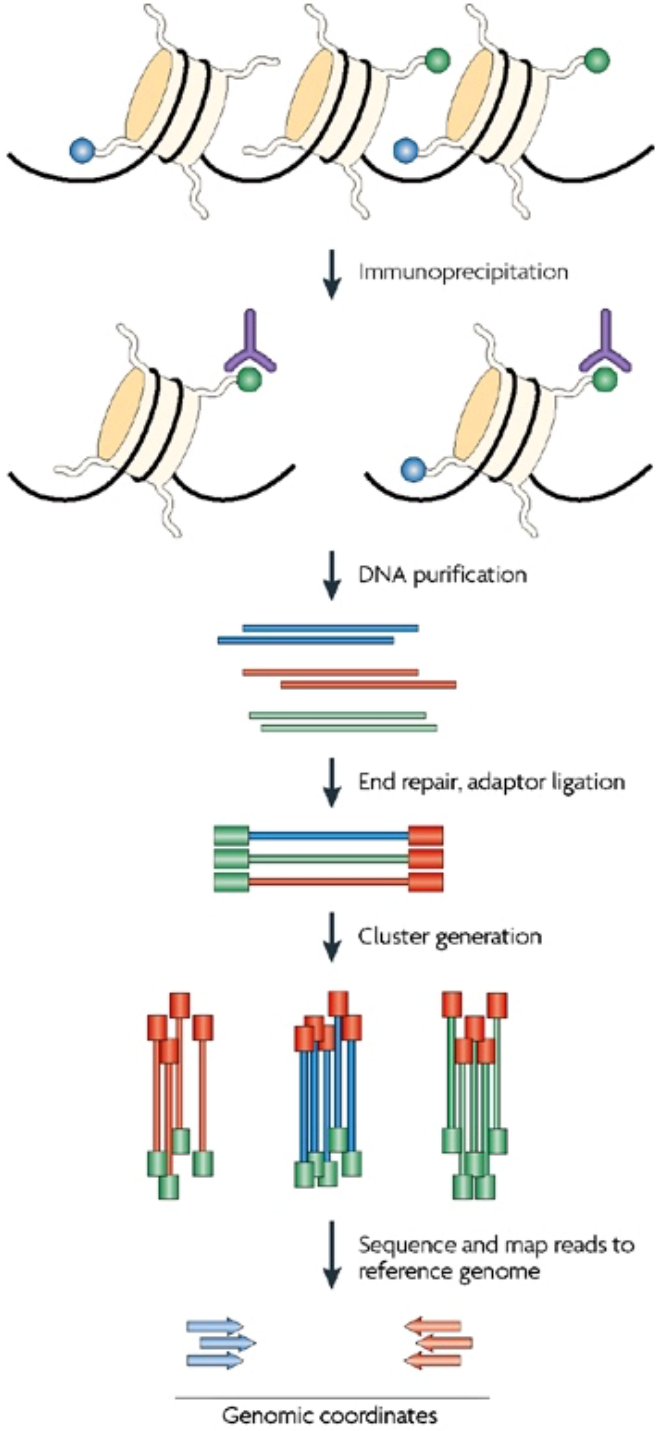


DNA purification



PCR amplification and sequencing





ChIP

Seq

Schones & Zhao. *Nat. Rev. Genet.* 2008

Some history: UV crosslinking (1984)

Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 4275–4279, July 1984
Biochemistry

Detecting protein–DNA interactions *in vivo*: Distribution of RNA polymerase on specific bacterial genes

(UV cross-linking/gene regulation/leucine operon/attenuation)

DAVID S. GILMOUR AND JOHN T. LIS

Section of Biochemistry, Molecular and Cell Biology, Cornell University, Ithaca, NY 14853

Communicated by Norman Davidson, March 23, 1984

ABSTRACT **We present an approach for determining the *in vivo* distribution of a protein on specific segments of chromosomal DNA. First, proteins are joined covalently to DNA by irradiating intact cells with UV light. Second, these cells are disrupted in detergent, and a specific protein is immunoprecipitated from the lysate. Third, the DNA that is covalently attached to the protein in the precipitate is purified and assayed by hybridization. To test this approach, we examine the cross-linking in *Escherichia coli* of RNA polymerase to a constitutively expressed, λ *cI* gene, and to the uninduced and isopropyl β-D-thiogalactoside (IPTG)-induced *lac* operon. As expected, the recovery of the constitutively expressed gene in the immunoprecipitate is dependent on the irradiation of cells and on the addition of RNA polymerase antiserum. The recovery of the *lac* operon DNA also requires transcriptional activation with IPTG prior to the cross-linking step. After these initial tests, we examine the distribution of RNA polymerase on the leucine operon of *Salmonella* in wild-type, attenuator mutant, and promoter mutant strains. Our *in vivo* data are in complete agreement with the predictions of the attenuation model of regulation. From these and other experiments, we discuss the resolution, sensitivity, and generality of these methods.**

RNA polymerase molecules can be associated with an actively transcribed gene, thereby enhancing the probability of generating a cross-link. Third, since regulatory mutations or chemical inducers can modulate the amount of RNA polymerase associated with a gene, the specificity of the interactions detected by our procedure can be rigorously tested. Moreover, the transcription level of some genes will remain unchanged, and these can serve as internal standards.

MATERIALS AND METHODS

Materials. *Escherichia coli* RNA polymerase had been purified as described (5). RNA polymerase antiserum was derived from a rabbit that was immunized as described (6) except 100 μg of purified RNA polymerase was used per injection. This antiserum immunoprecipitates the β and β′ subunits of both *E. coli* and *Salmonella* RNA polymerase. Protein A Sepharose (Pharmacia) was stored at 4°C in 150 mM NaCl/50 mM Tris·HCl, pH 8.0/1 mM EDTA, and was recycled after use by extensively washing with 50 mM $NaHCO_3$/1% $NaDodSO_4$.

All plasmid DNAs were maintained in *E. coli* HB101. Several of the plasmids are described elsewhere: pBGP120 (7), pKK3535 (8), pCV12 (9), and PUC13 (10). Plasmid pLRI was

Crosslinking + immunoprecipitation (1993)

Cell, Vol. 75, 1187–1198, December 17, 1993, Copyright © 1993 by Cell Press

Mapping Polycomb-Repressed Domains in the Bithorax Complex Using In Vivo Formaldehyde Cross-Linked Chromatin

Valerio Orlando and Renato Paro
Zentrum für Molekulare Biologie
Universität Heidelberg
Im Neuenheimer Feld 282
69120 Heidelberg
Federal Republic of Germany

Summary

The Polycomb group (Pc-G) proteins are responsible for keeping developmental regulators, like homeotic genes, stably and inheritably repressed during Drosophila development. Several similarities to a protein class involved in heterochromatin formation suggest that the Pc-G exerts its function at the higher order chromatin level. Here we have mapped the distribution of the Pc protein in the homeotic bithorax complex (BX-C) of Drosophila tissue culture cells. We have elaborated a method, based on the in vivo formaldehyde cross-linking technique, that allows a substantial enrichment for Pc-interacting sites by immunoprecipitation of the cross-linked chromatin with anti-Pc antibodies. We find that the Pc protein quantitatively covers large regulatory regions of repressed BX-C genes. Conversely, we find that the *Abdominal-B* gene is active in these cells and the region devoid of any bound Pc protein.

mined state, dispensing the cell from reproducing at every generation the complexity of a particular regulatory cascade.

The *Pc* gene is the prototype member of the Pc-G. As shown by polytene chromosome immunostainings, *Pc* encodes a nuclear protein associated with more than 100 loci in the genome, including the homeotic clusters of the Antennapedia (Antp) complex and bithorax complex (BX-C) (Zink and Paro, 1989). The Pc protein was not found to bind DNA sequence specifically in vitro, not even to sequences for which the protein is otherwise targeted in vivo, such as the *Antp* promoter (Zink and Paro, 1989). Other members of the Pc-G, like polyhomeotic and Posterior sex combs, have also been characterized, and although potential DNA-binding domains are present, these proteins, too, fail to bind DNA specifically in vitro (De Camillis et al., 1992; Rastelli et al., 1993). Thus, the ability of this class of proteins to bind specific genomic regions in vivo might involve the formation of higher order nucleoprotein complexes, a level of complexity not easily reproducible in vitro. Indeed, cytological and biochemical analysis showed that some Pc-G proteins share the same binding sites on polytene chromosomes and that they are part of a large multimeric complex (Franke et al., 1992; Rastelli et al., 1993).

An important feature of Pc is the presence of a highly conserved protein motif spanning over 48 amino acids at the amino-terminal end, called the chromodomain (Paro and Hogness, 1991). This protein domain is also found in the heterochromatin-associated protein HP1, encoded by

LIGATION MEDIATED - PCR OF IP - DNA

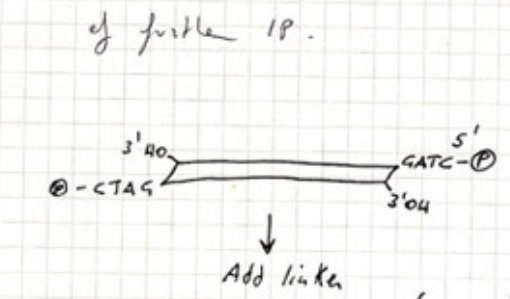
Hypothesis: DNA from IP of chromatin could be simply amplified by PCR using a linker.

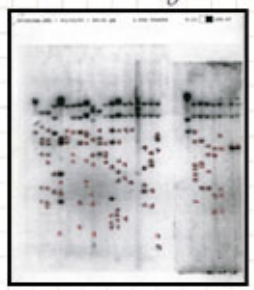
The idea is: 1) cut the IP DNA with a restriction enzyme (4 cutter) to have a more homogeneous population of DNA fragments in terms of length.

2) ligate an oligo to the IP DNA maintaining the restriction site

3) amplify the DNA with an oligo antisense of the ligated one.

4) End up with sufficient material virtually for several experiments without necessity of further IP.





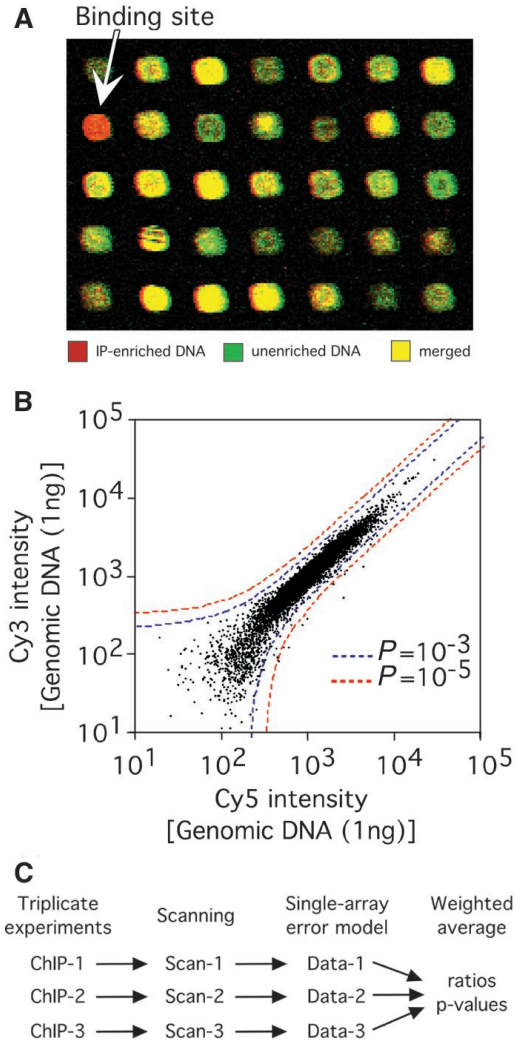
ChIP-chip (2000)

REPORTS

Genome-Wide Location and Function of DNA Binding Proteins

Bing Ren,[1*] François Robert,[1*] John J. Wyrick,[1,2*]
Oscar Aparicio,[2,4] Ezra G. Jennings,[1,2] Itamar Simon,[1]
Julia Zeitlinger,[1] Jörg Schreiber,[1] Nancy Hannett,[1]
Elenita Kanin,[1] Thomas L. Volkert,[1] Christopher J. Wilson,[5]
Stephen P. Bell,[2,3] Richard A. Young[1,2†]

Understanding how DNA binding proteins control global gene expression and chromosomal maintenance requires knowledge of the chromosomal locations at which these proteins function in vivo. We developed a microarray method that reveals the genome-wide location of DNA-bound proteins and used this method to monitor binding of gene-specific transcription activators in yeast. A combination of location and expression profiles was used to identify genes whose expression is directly controlled by Gal4 and Ste12 as cells respond to changes in carbon source and mating pheromone, respectively. The results identify pathways that are coordinately regulated by each of the two activators and reveal previously unknown functions for Gal4 and Ste12. Genome-wide location analysis will facilitate investigation of gene regulatory networks, gene function, and genome maintenance.



ChIP-seq (2007)

Resource

Cell

High-Resolution Profiling of Histone Methylations in the Human Genome

Artem Barski,[1,3] Suresh Cuddapah,[1,3] Kairong Cui,[1,3] Tae-Young Roh,[1,3] Dustin E. Schones,[1,3] Zhibin Wang,[1,3] Gang Wei,[1,3] Iouri Chepelev,[2] and Keji Zhao[1,*]
[1] Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA
[2] Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA
[3] These authors contributed equally to this work and are listed alphabetically.
*Correspondence: zhaok@nhlbi.nih.gov
DOI 10.1016/j.cell.2007.05.009

SUMMARY

Histone modifications are implicated in influencing gene expression. We have generated high-resolution maps for the genome-wide distribution of 20 histone lysine and arginine methylations as well as histone variant H2A.Z, RNA polymerase II, and the insulator binding protein CTCF across the human genome using the Solexa 1G sequencing technology. Typical patterns of histone methylations exhibited at promoters, insulators, enhancers, and transcribed regions are identified. The monomethylations of H3K27, H3K9, H4K20, H3K79, and H2BK5 are all linked to gene activation, whereas trimethylations of H3K27, H3K9, and H3K79 are linked to repression. H2A.Z associates with functional regulatory elements, and CTCF marks boundaries of histone methylation domains. Chromosome banding patterns are correlated with unique patterns of histone modifications. Chromosome breakpoints detected in T cell cancers frequently reside in chromatin regions associated with H3K4 methylations. Our data provide new insights into the function of histone methylation and chromatin organization in genome function.

biological processes. Among the various modifications, histone methylations at lysine and arginine residues are relatively stable and are therefore considered potential marks for carrying the epigenetic information that is stable through cell divisions. Indeed, enzymes that catalyze the methylation reaction have been implicated in playing critical roles in development and pathological processes.

Remarkable progress has been made during the past few years in the characterization of histone modifications on a genome-wide scale. The main driving force has been the development and improvement of the "ChIP-on-chip" technique by combining chromatin immunoprecipitation (ChIP) and DNA-microarray analysis (chip). With almost complete coverage of the yeast genome on DNA microarrays, its histone modification patterns have been extensively studied. The general picture emerging from these studies is that promoter regions of active genes have reduced nucleosome occupancy and elevated histone acetylation (Bernstein et al., 2002, 2004; Lee et al., 2004; Liu et al., 2005; Pokholok et al., 2005; Sekinger et al., 2005; Yuan et al., 2005). High levels of H3K4me1, H3K4me2, and H3K4me3 are detected surrounding transcription start sites (TSSs), whereas H3K36me3 peaks near the 3′ end of genes.

Significant progress has also been made in characterizing global levels of histone modifications in mammals. Several large-scale studies have revealed interesting insights into the complex relationship between gene expression and histone modifications. Generally, high levels of histone acetylation and H3K4 methylation are detected

Genome-Wide Mapping of in Vivo Protein-DNA Interactions

David S. Johnson,[1*] Ali Mortazavi,[2*] Richard M. Myers,[1†] Barbara Wold[2,3†]

In vivo protein-DNA interactions connect each transcription factor with its direct targets to form a gene network scaffold. To map these protein-DNA interactions comprehensively across entire mammalian genomes, we developed a large-scale chromatin immunoprecipitation assay (ChIPSeq) based on direct ultrahigh-throughput DNA sequencing. This sequence census method was then used to map in vivo binding of the neuron-restrictive silencer factor (NRSF; also known as REST, for repressor element–1 silencing transcription factor) to 1946 locations in the human genome. The data display sharp resolution of binding position [±50 base pairs (bp)], which facilitated our finding motifs and allowed us to identify noncanonical NRSF-binding motifs. These ChIPSeq data also have high sensitivity and specificity [ROC (receiver operator characteristic) area ≥ 0.96] and statistical confidence ($P < 10^{-4}$), properties that were important for inferring new candidate interactions. These include key transcription factors in the gene network that regulates pancreatic islet cell development.

Although much is known about transcription factor binding and action at specific genes, far less is known about the composition and function of entire factor-DNA interactomes, especially for organisms with large genomes. Now that human, mouse, and other large genomes have been sequenced, it is possible, in principle, to measure how any transcription factor is deployed across the entire genome for a given cell type and physiological condition. Such measurements are important for systems-level studies because they provide a global map of candidate gene network input connections. These direct physical interactions between transcription factors or cofactors and the chromosome can be detected by chromatin immunoprecipitation (ChIP) (*1*). In ChIP experiments, an immune reagent specific for a DNA binding factor is used to enrich target DNA sites to which the factor was bound in the living cell. The enriched DNA sites are then identified and quantified.

For the gigabase-size genomes of vertebrates, it has been difficult to make ChIP measurements that combine high accuracy, whole-genome completeness, and high binding-site resolution. These data-quality and depth issues dictate whether primary gene network structure can be inferred with reasonable certainty and comprehensiveness, and how effectively the data can be used to discover binding-site motifs by computational methods. For these purposes, statistical robustness, sampling depth across the genome, absolute signal and signal-to-noise ratio must be good enough to detect nearly all in vivo binding locations for a regulator with minimal inclusion of false-positives. A further challenge in genomes large or small is to map factor-binding sites with high positional resolution. In addition to making computational discovery of binding motifs feasible, this dictates the quality of regulatory site annotation relative to other gene anatomy landmarks, such as transcription start sites, enhancers, introns and exons, and conserved noncoding features (*2*). Finally, if high-quality protein-DNA interactome measurements can be performed routinely and at reasonable cost, it will open the way to detailed studies of interactome dynamics in response to specific signaling stimuli or genetic mutations. To address these issues, we turned to ultrahigh-throughput DNA sequencing to gain sampling power and applied size selection on immuno-enriched DNA to enhance positional resolution.

The ChIPSeq assay shown here differs from other large-scale ChIP methods such as ChIPArray, also called ChIPchip (*1*); ChIPSAGE (SACO) (*3*); or ChIPPet (*4*) in design, data produced, and cost. The design is simple (Fig. 1A) and, unlike SACO or ChIPPet, it involves no plasmid library construction. Unlike microarray assays, the vast majority of single-copy sites in the genome is accessible for ChIPSeq assay (*5*), rather than a subset selected to be array features. For example, to sample with similar completeness by an Affymetrix-style microarray design, a nucleotide-by-nucleotide sliding window design of roughly 1 billion features per array would be needed for the nonrepeat portion of the human genome. In addition, ChIPSeq counts sequences and so avoids constraints imposed by array hybridization chemistry, such as base composition constraints related to T_m, the temperature at which 50% of double-stranded DNA or DNA-RNA hybrids is denatured; cross-hybridization; and secondary structure interference. Finally, ChIPSeq is feasible for any sequenced genome, rather than being restricted to species for which whole-genome tiling arrays have been produced.

ChIPSeq illustrates the power of new sequencing platforms, such as those from Solexa/Illumina and 454, to perform sequence census counting assays. The generic task in these applications is to identify and quantify the molecular

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA, 94305–5120, USA. [2]Biology Division, California Institute of Technology, Pasadena, CA 91125, USA. [3]California Institute of Technology Beckman Institute, Pasadena, CA 91125, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: woldb@its.caltech.edu (B.W.); myers@shgc.stanford.edu (R.M.M.)

nature Vol 446|29 March 2007|doi:10.1038/nature05632

LETTERS

Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome

Istvan Albert[1], Travis N. Mavrich[1,2], Lynn P. Tomsho[1], Ji Qi[1], Sara J. Zanton[1,2], Stephan C. Schuster[1] & B. Franklin Pugh[1,2]

The nucleosome is the fundamental building block of eukaryotic chromosomes. Access to genetic information encoded in chromosomes is dependent on the position of nucleosomes along the DNA. Alternative locations just a few nucleotides apart can have profound effects on gene expression[1]. Yet the nucleosomal context in which chromosomal and gene regulatory elements reside remains ill-defined on a genomic scale. Here we sequence the DNA of 322,000 individual *Saccharomyces cerevisiae* nucleosomes, containing the histone variant H2A.Z, to provide a comprehensive map of H2A.Z nucleosomes in functionally important regions. With a median 4-base-pair resolution, we identify new and established signatures of nucleosome positioning. A single predominant rotational setting and multiple translational settings are evident. Chromosomal elements, ranging from telomeres to centromeres and transcriptional units, are found to possess characteristic nucleosomal architecture that may be important for their function. Promoter regulatory elements, including transcription factor binding sites and transcriptional start sites, show topological relationships with nucleosomes, such that transcription factor binding sites tend to be rotationally exposed on the nucleosome surface near its border. Transcriptional start sites tended to reside about one helical turn inside the nucleosome border. These findings reveal an intimate relationship between chromatin architecture and the underlying DNA sequence it regulates.

Chromatin is composed of repeating units of nucleosomes in which ~147 base pairs (bp) of DNA is wrapped ~1.7 times around the exterior of a histone protein complex[2]. A nucleosome has two fundamental relationships with its DNA[3]. A translational setting defines a nucleosomal midpoint relative to a given DNA locus. A rotational setting defines the orientation of DNA helix on the histone surface. Thus, DNA regulatory elements may reside in linker regions between nucleosomes or along the nucleosome surface, where they may face inward (potentially inaccessible) or outward (potentially accessible). Recent discoveries of nucleosome positioning sequences throughout the *S. cerevisiae* (yeast) genome suggest that nucleosome locations are partly defined by the underlying DNA sequence[4,5]. Indeed, a tendency of AA/TT dinucleotides to recur in 10-bp intervals and in counterphase with GC dinucleotides generates a curved DNA structure that favours nucleosome formation[3]. Genome-wide maps of nucleosome locations have been generated[6,7], but not at a resolution that would define translational and rotational settings. To acquire a better understanding of how genes are regulated by nucleosome positioning, we isolated and sequenced H2A.Z-containing nucleosomes from *S. cerevisiae*. Such nucleosomes are enriched at promoter regions[8–11], and thus maximum coverage of relevant regions can be achieved with fewer sequencing runs. With this high resolution map we sought to address the following questions: (1) what are the DNA signatures of nucleosome positioning *in vivo*? (2) How many translational and rotational settings do nucleosomes occupy? (3) Do chromosomal elements possess specific chromatin architecture? (4) What is the topological relationship between the location of promoter elements and the rotational and translational setting of nucleosomes?

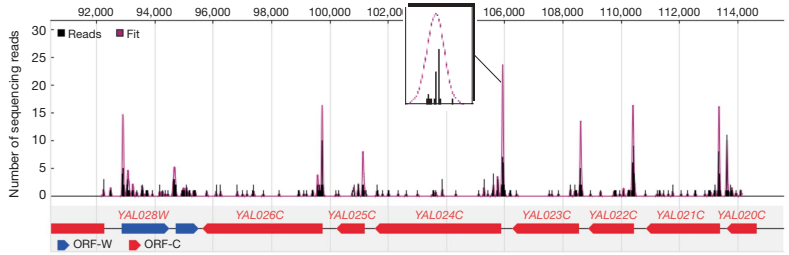


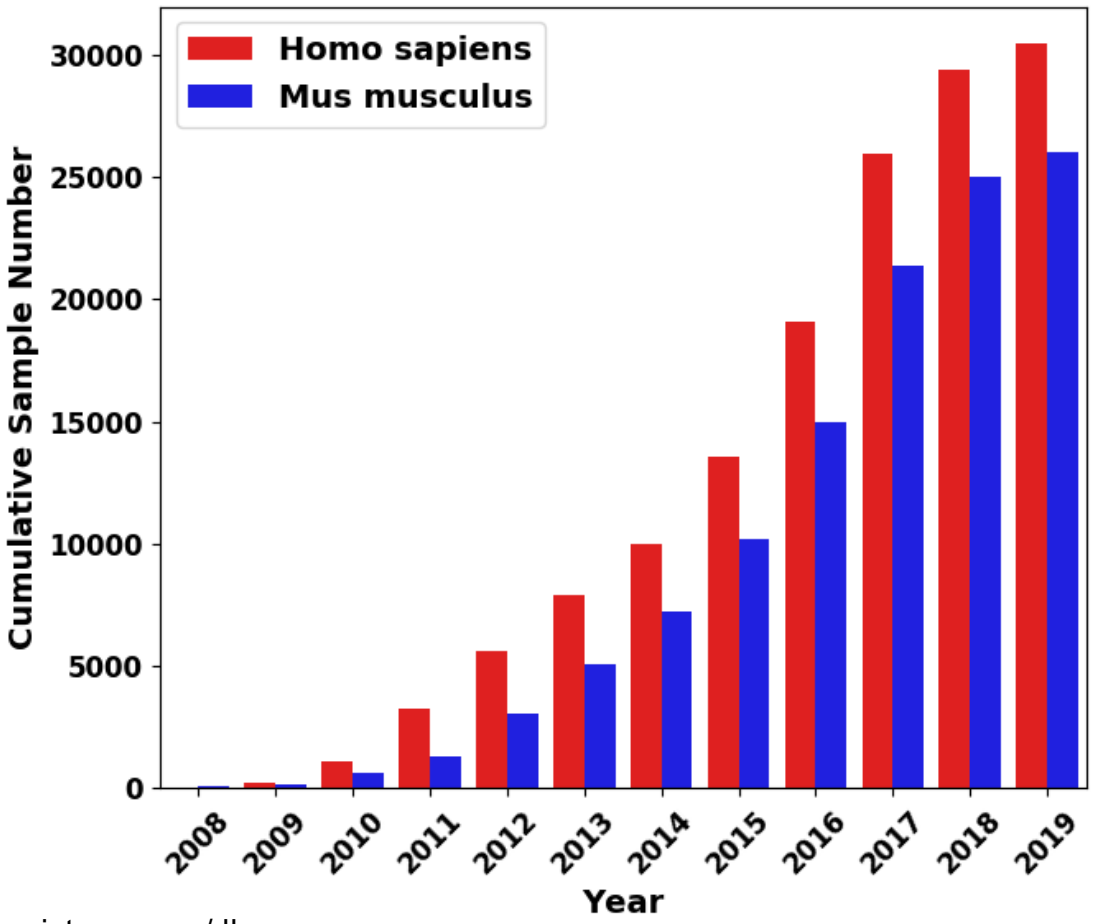
Figure 1 | Distribution of H2A.Z nucleosomal DNA at an arbitrary region of the yeast genome. Any region of the genome can be viewed in this way at http://nucleosomes.sysbio.bx.psu.edu. An enlarged view of a peak is shown in the inset, where each vertical bar corresponds to the number of sequencing reads located at individual chromosomal coordinates. The locations of ORFs are shown below the peaks. Additional browser shots are shown in Supplementary Fig. 1.

[1]Center for Comparative Genomics and Bioinformatics, [2]Center for Gene Regulation, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.

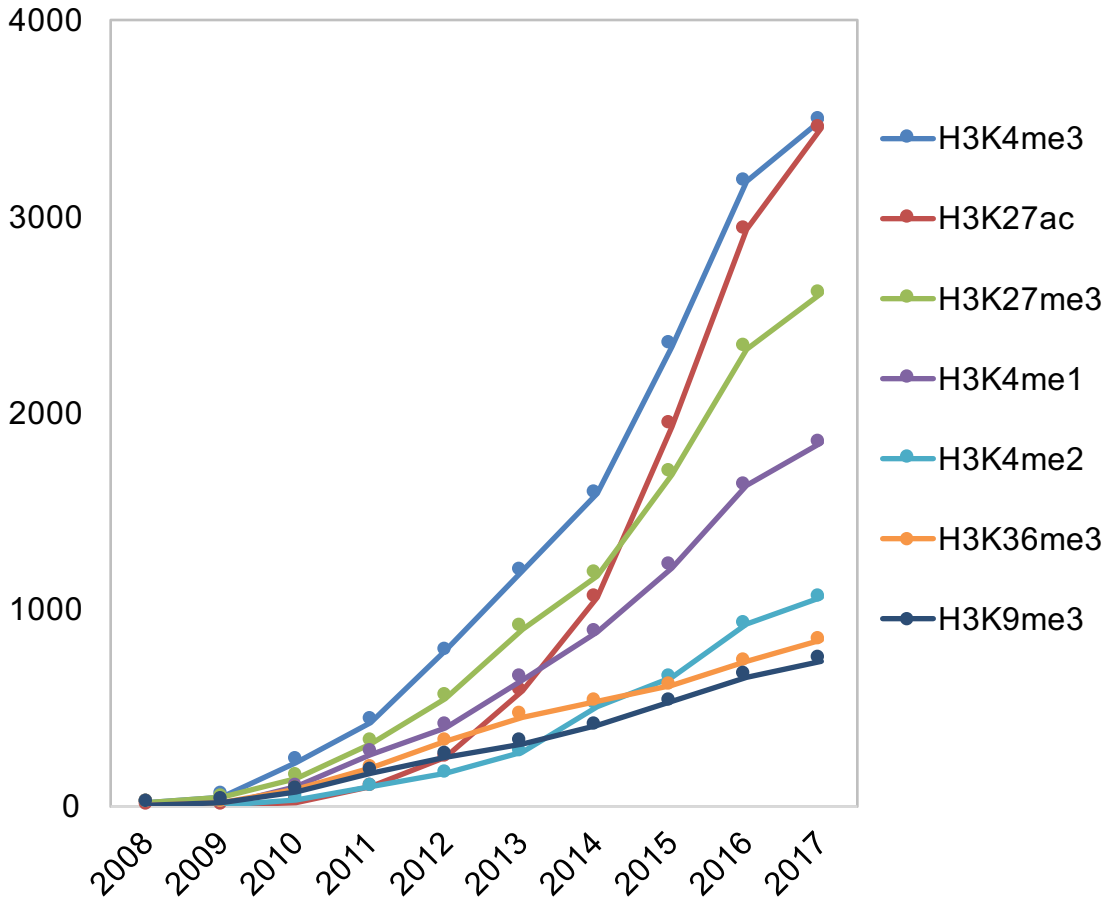
First ChIP-seq papers

Title	First/last authors	Journal	First submission date	Acceptance date	Publication date	Species/cell type	Target factors	# citations (2/28/20)
Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome	Albert…Pugh	*Nature*	10/20/2006	1/26/2007	3/29/2007	Yeast	H2A.Z	676
High-resolution profiling of histone methylations in the human genome	Barski, Cuddapah, Cui, Roh, Schones, Wang, Wei,…, Zhao	*Cell*	4/20/2007	5/3/2007	5/17/2007	Human $CD4^+$ T cells	20 histone methylations, H2A.Z, PolII, CTCF	6036
Genome-wide mapping of in vivo protein-DNA interactions	Johnson, Mortazavi; Myers, Wold	*Science*	2/14/2007	4/26/2007	5/31/2007	Human Jurkat cell line	NRSF (REST)	2633
Genome-wide maps of chromatin state in pluripotent and lineage-committed cells	Mikkelsen,…, Lander, Bernstein	*Nature*	5/10/2007	6/13/2007	7/1/2007	Mouse ESC, NPC, MEF	4 histone methylations, PolII, H3	3919

ChIP-seq has become a dominant method for profiling epigenomes

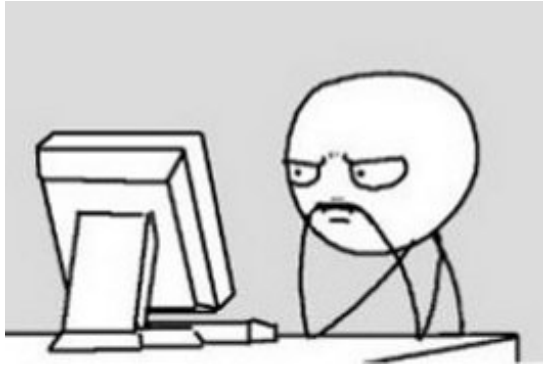


cistrome.org/db

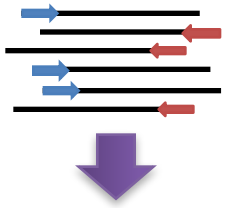


Outline

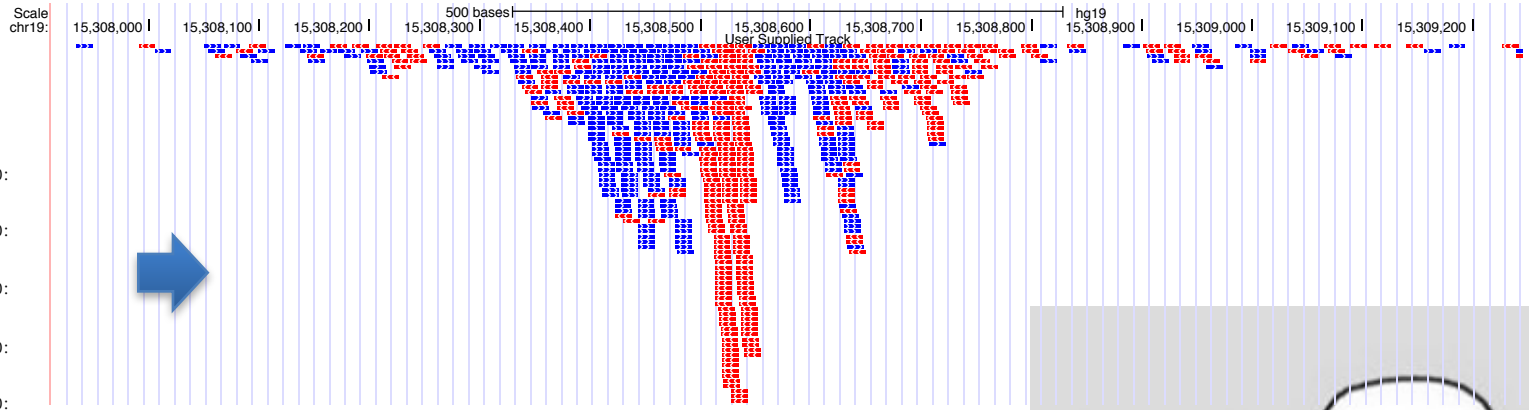
- Epigenome: an overview
- ChIP-seq technique
- **ChIP-seq data analysis**
- Future perspective

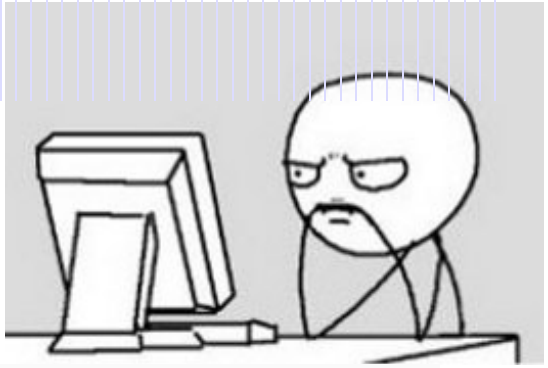


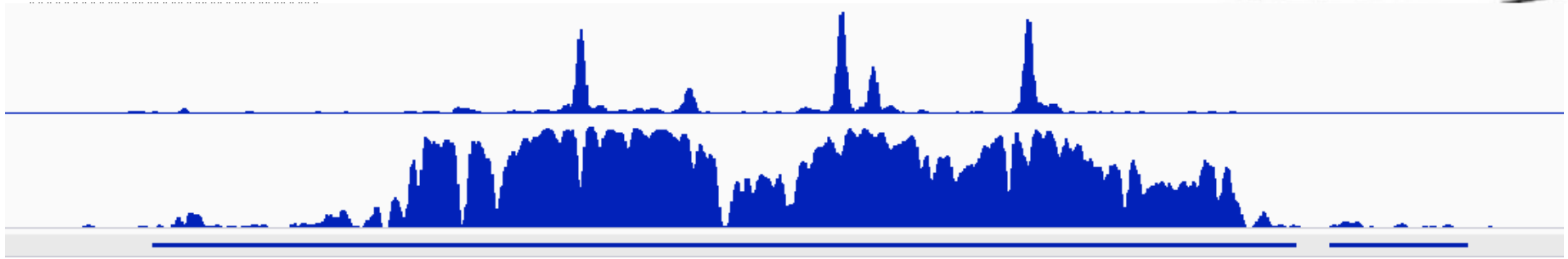
ChIP-seq data analysis overview



```
@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
```

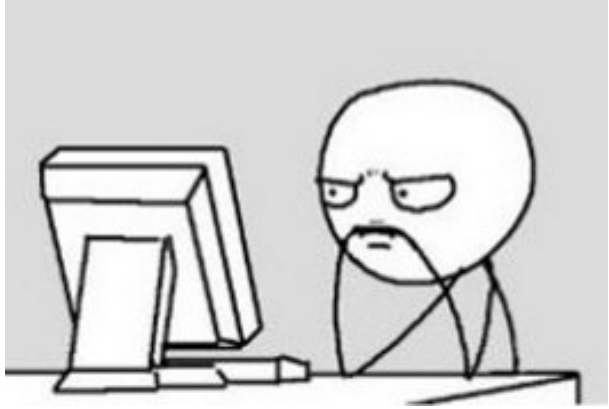






ChIP-seq data analysis overview

- 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

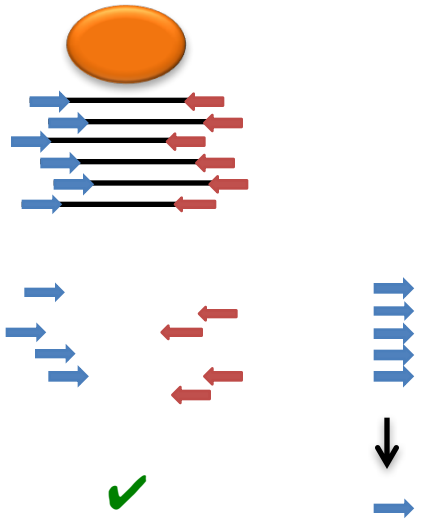


ChIP-seq data analysis: basic processing

- alignment of each sequence read: **bowtie**, **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:

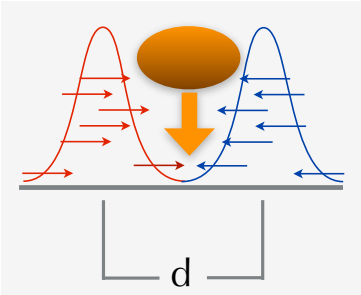


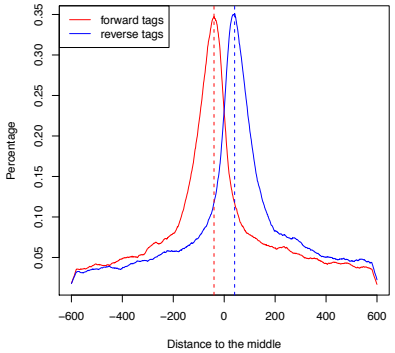
Langmead et al. 2009,
Zang et al. 2009

ChIP-seq data analysis: Peak calling

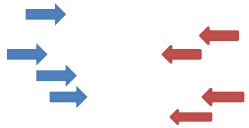
- DNA fragment size estimation

peak model

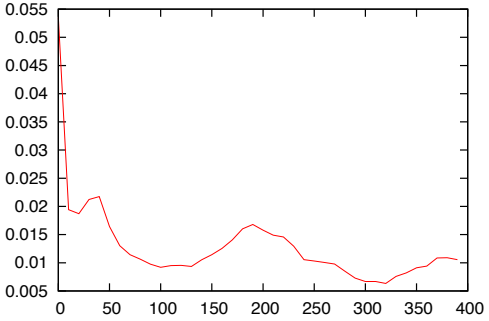




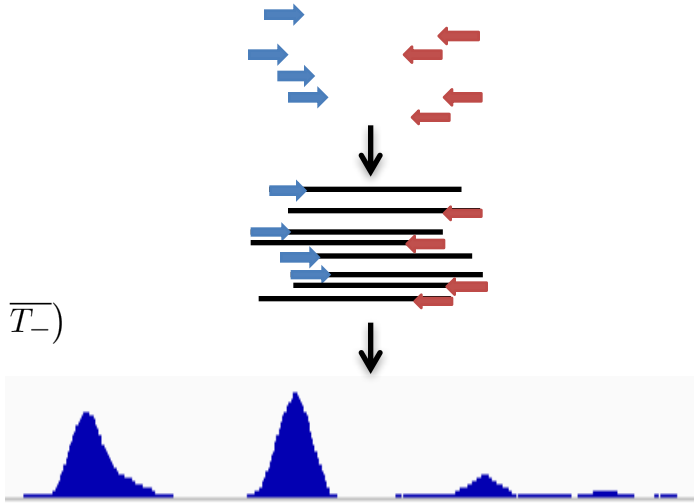
cross-correlation



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



- pile-up profiling



- Peak/signal detection

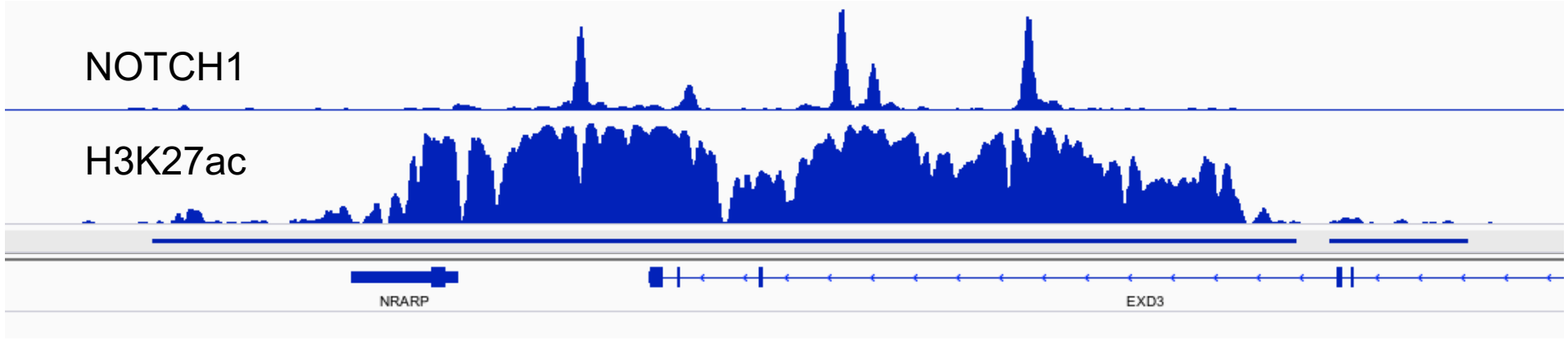
ChIP-seq data analysis: Peak calling

- **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

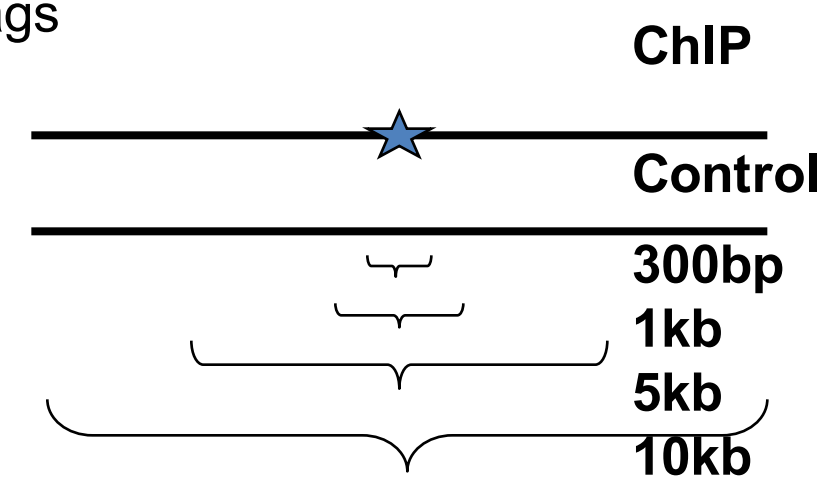


Wang, Zang et al. 2014

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

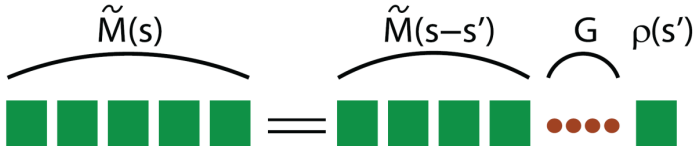
Dynamic λ_{local} =
max(λ_{BG}, [λ_{ctrl}, λ_{1k},] λ_{5k}, λ_{10k})



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

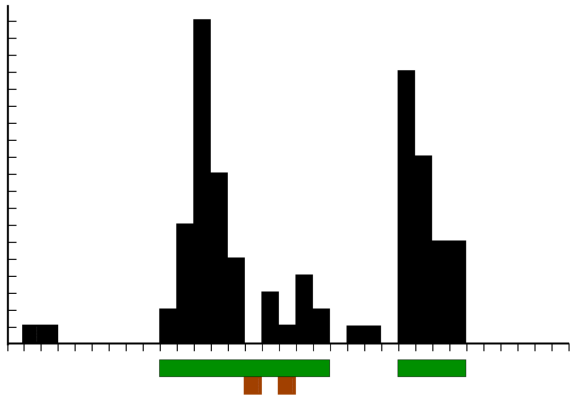
SICER

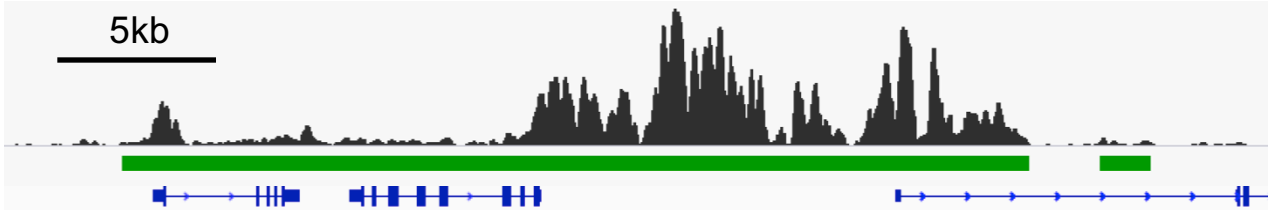
- **S**patial-clustering **I**dentification of **C**hIP-**E**nriched **R**egions

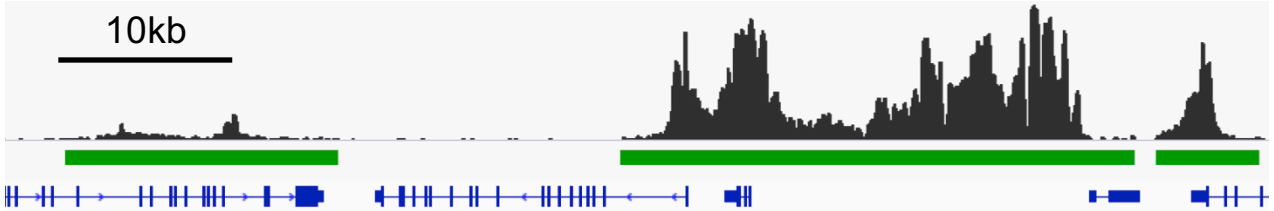


$$\tilde{M}(s) = G(\lambda, l_0, g) \int_{s_0}^{s} \mathrm{d}s' \tilde{M}(s - s')\rho(s')$$

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







Zang et al. *Bioinformatics* 2009

ChIP-seq peak calling: Parameters

Parameter	Remarks
Genome	Species and reference genome version, e.g. hg38, hg19, 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

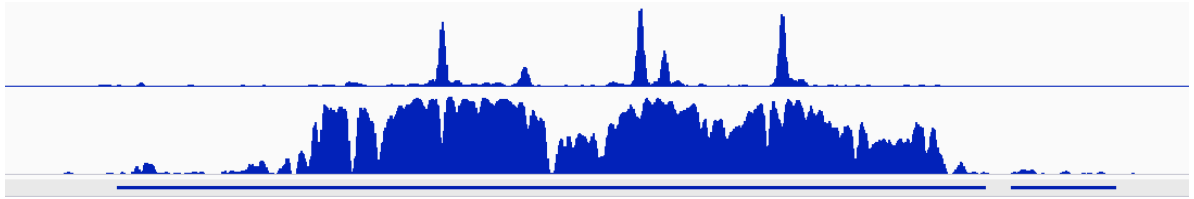
- fastq: raw sequences

- BED:

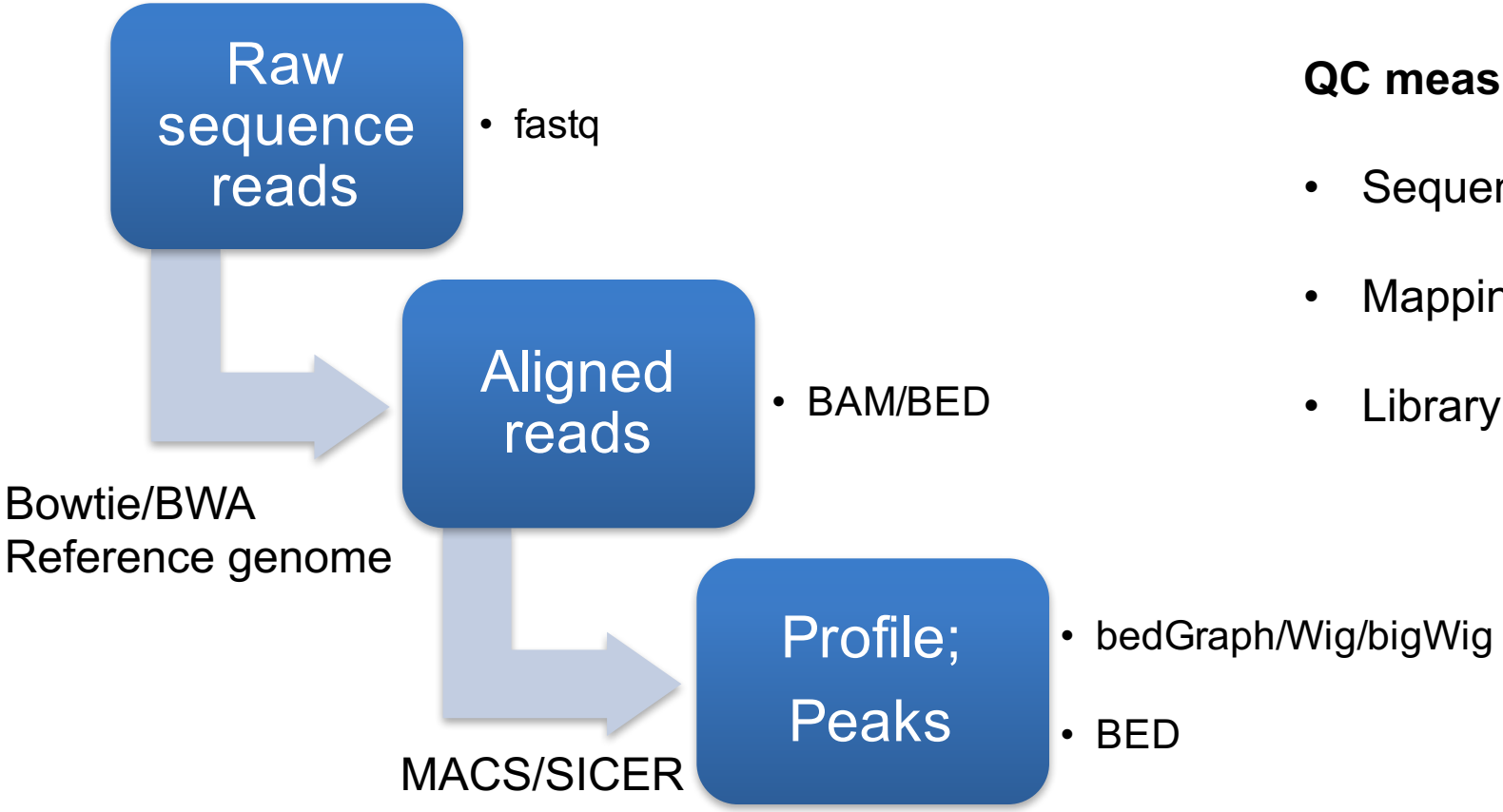
chr11	10344210	10344260	255	0	-
chr4	76649430	76649480	255	0	+
chr3	77858754	77858804	255	0	+
chr16	62688333	62688383	255	0	+
chr22	33031123	33031173	255	0	-

- SAM/BAM: aligned sequencing reads

- bedGraph, Wig, bigWig: pile-up profiles for browser visualization



Data flow



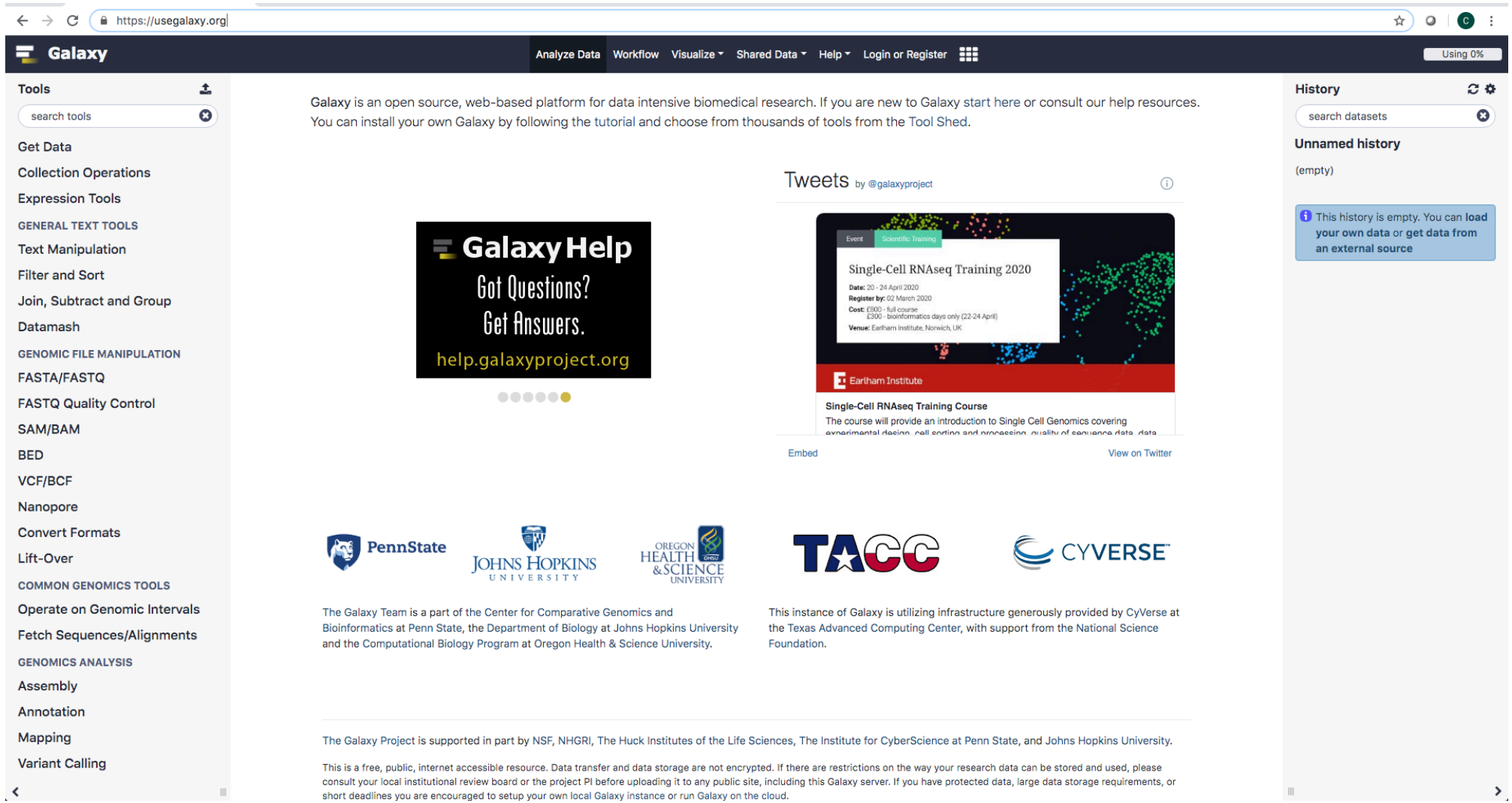
QC measures:

- Sequence quality (fastqc)
- Mapping quality (uniquely mapped ratio)
- Library complexity (PBC)

- ChIP enrichment
- Signal/Noise (FRIP score)
- Regulatory annotation

Galaxy: web-interface analysis platform

- https://usegalaxy.org/



Run MACS on Cistrome, a Galaxy-based platform

- http://cistrome.org/ap/

Galaxy / Cistrome / Cistrome

Chongzhi

cistrome.org/ap/root

Galaxy / Cistrome

Analyze Data | Workflow | Shared Data | Lab | Visualization | Help | User

Using 30.3 GB

Tools

search tools

CISTROME TOOLBOX

Import Data

Upload File from your computer

CistromeFinder Import from Cistrome Finder

CistromeCR Import from Cistrome Chromatin Regulator

Expression CEL file packager can download .cel files from GEO by given GSM IDs and prepare a cel.zip file for expression analysis.

GenomeSpace import from file browser

Data Preprocessing

Gene Expression

Integrative Analysis

Liftover/Others

GALAXY TOOLBOX

Get Data

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Upload File (version 1.1.4)

File Format:

Auto-detect

Which format? If for expression data, choose cel.zip or xys.zip. See help below

File (Please avoid Windows format text file):

Choose File No file chosen

TIP1: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or ASPERA (please read the instruction). TIP2: If you want to upload expression data, please read the instruction and specify cel.zip or xys.zip for file format.

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via ASPERA:

File	Size	Date
Your ASPERA upload directory contains no files.		

This Galaxy server allows you to upload files via ASPERA. To upload some files, log in to the ASPERA server at **cistrome.dfci.harvard.edu** using your Cistrome credentials (email address and password).

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand.

Genome:

Human Dec. 2013 (GRCh38/hg38) (hg38)

Execute

History

Unnamed history

329.0 MB

68: Heatmap log

67: Heatmap k-means cl assified regions

66: Heatmapr R script

65: Heatmap image

64: Heatmap log

63: Heatmap k-means cl assified regions

62: Heatmapr R script

61: Heatmap image

60: Heatmap log

59: Heatmap k-means cl assified regions

58: Heatmapr R script

57: Heatmap image

56: Heatmap log

55: Heatmap k-means cl assified regions

Try SICER2

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

SICER2 Documentation

zanglab.github.io/SICER2/

SICER2 Documentation

Search docs

Docs » Quick Start

Edit on GitHub

SICER2

Redesigned and improved ChIP-seq broad peak calling tool SICER

build passing

GitHub Repo

Introduction

Chromatin immunoprecipitation combined with high-throughput sequencing (ChIP-seq) can be used to map binding sites of a protein of interest in the genome. Histone modifications usually occupy broad chromatin domains and result in diffuse patterns in ChIP-seq data that make it difficult to identify signal enrichment. SICER, a spatial clustering approach for the identification of ChIP-enriched regions, was developed for calling broad peaks from ChIP-seq data.

Usability of the original SICER software has been affected by increased throughputs of ChIP-seq experiments over the years. We now present SICER2 a more user-friendly version of SICER that has been redisgned and streamlined to handle large ChIP-seq data sets. This new Python package supports multiple job submissions on cluster systems and parallel processing on multicore architectures.

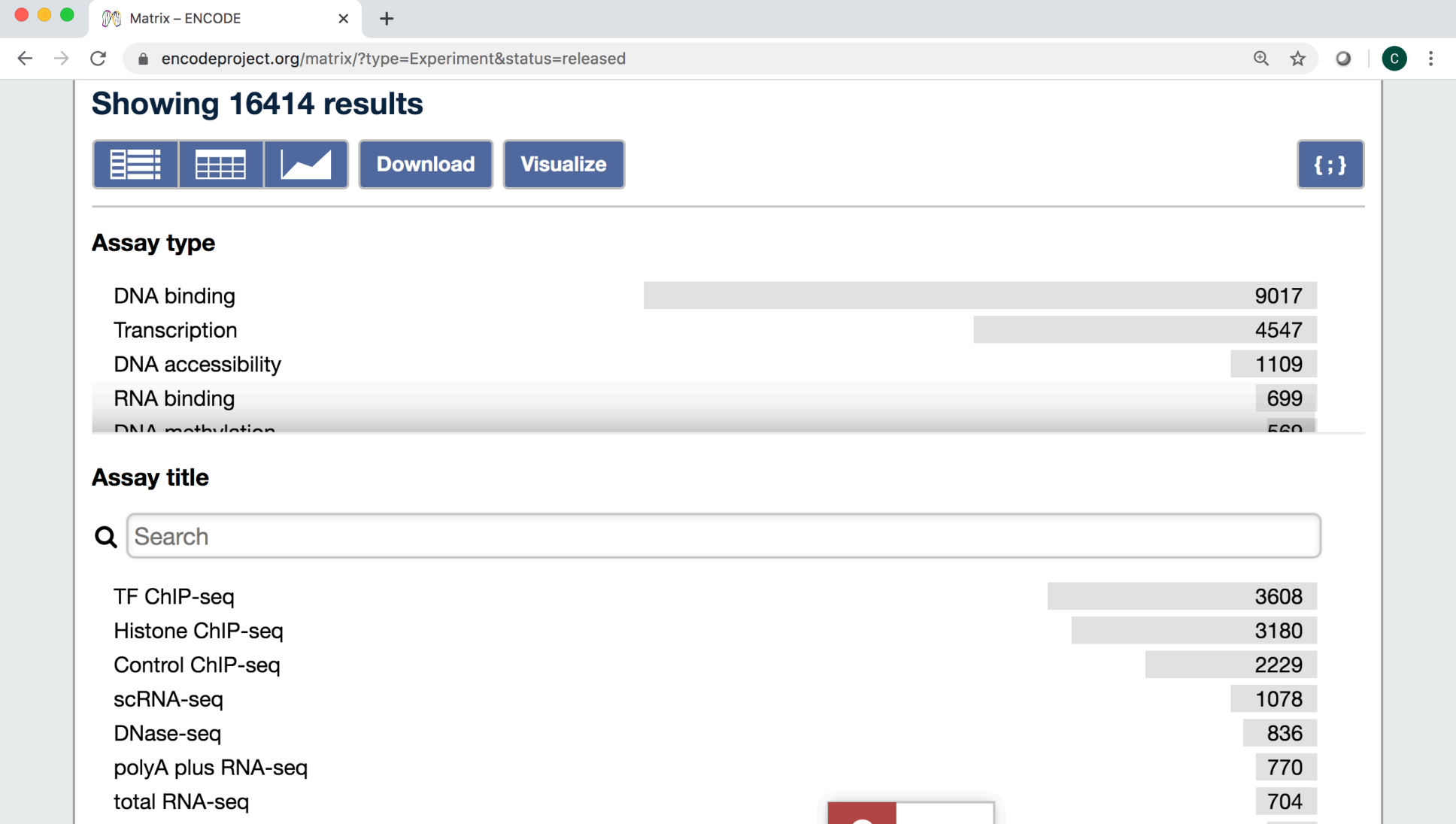
Jin Yong (Jeffrey) Yoo 2020

ChIP-seq: Downstream analyses

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

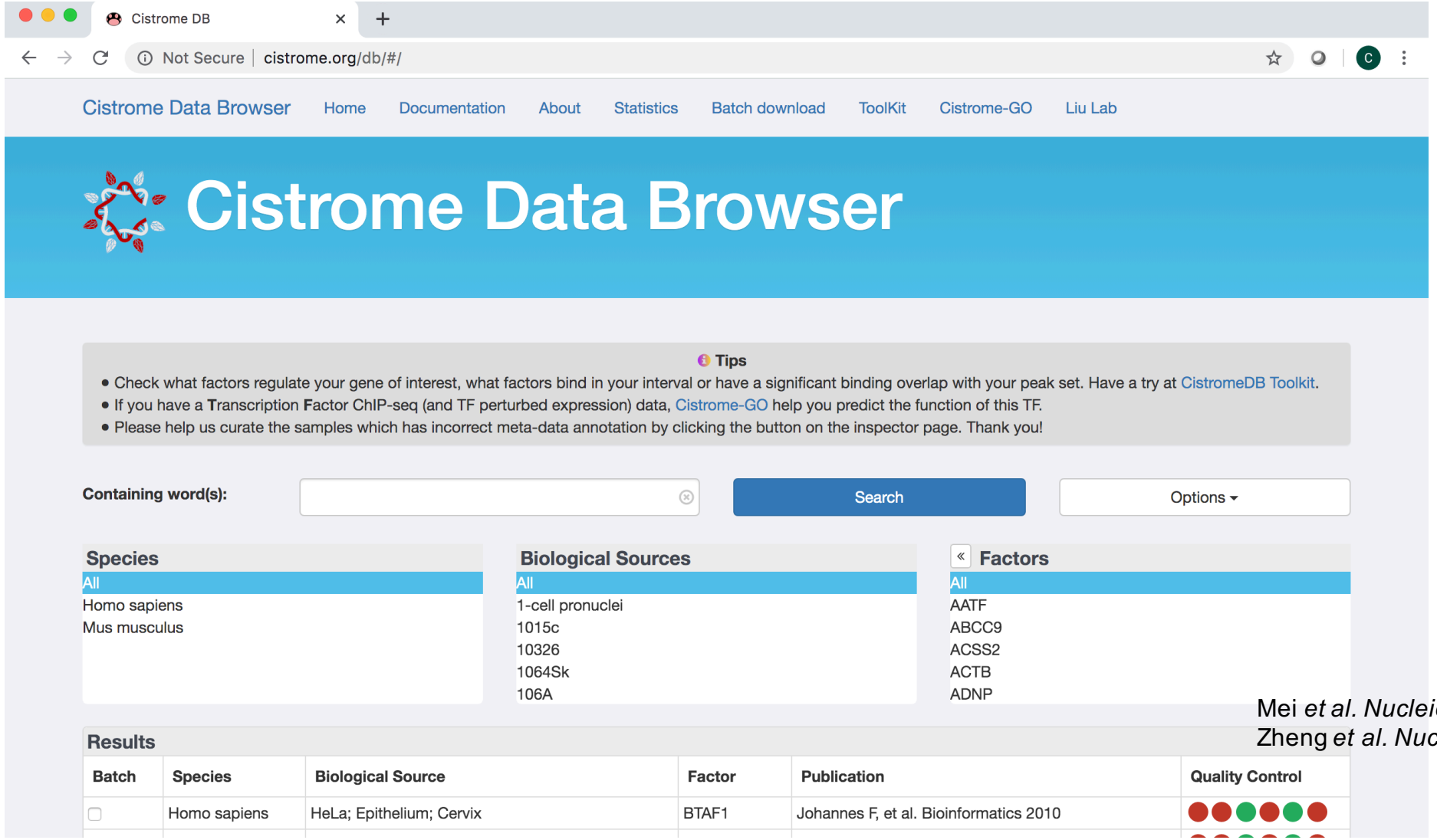
ENCODE

https://www.encodeproject.org/



Cistrome Data Browser

http://cistrome.org/db/



Mei *et al. Nucleic Acids Res.* 2017
Zheng *et al. Nucleic Acids Res.* 2018

BART: TF prediction using public ChIP-seq data

http://bartweb.org

BART: Binding Analysis for Regulation of Transcription

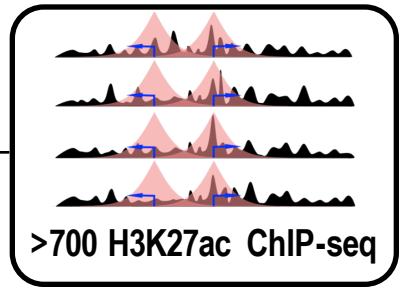
User input

Gene list

ChIP-seq

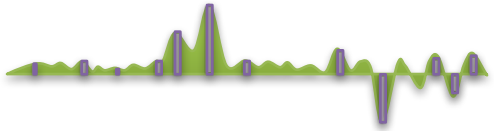
Region set

Hi-C



Adaptive Lasso regression

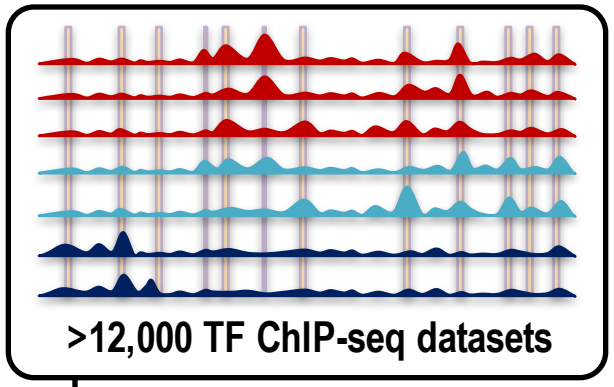
Mapping



Cis-regulatory profile

Cis-regulatory element repertoire (2.7 million in the human genome, 1.5 million in mouse)

differential interaction

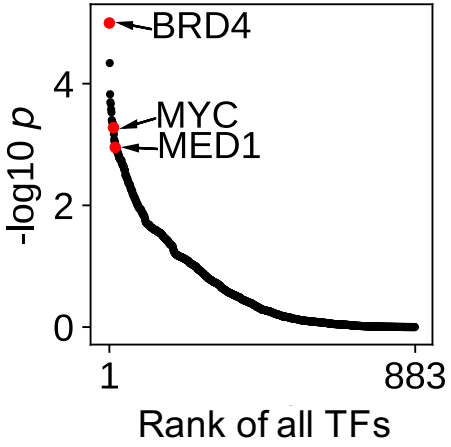


ROC associations

Statistical tests,
Background adjustment,
Rank integration

Wang *et al.*, *Bioinformatics* 2018

Output prediction



Ma *et al.*, under review 2020

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
4. Take advantage of public resources

Future Perspectives

- Limitations of ChIP-seq:
  - Dependent on antibody availability and quality
  - Semi-quantitative: does not detect global change
  - Needs many cells – difficult for clinical samples
  - Cellular heterogeneity
- Single-cell epigenomics
  - Single-cell ChIP-seq
  - Single-cell ATAC-seq
  - Joint assay of scRNA-seq and sc epigenome-seq
- Spatial genomics and epigenomics

ARTICLES

nature biotechnology

Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state

Assaf Rotem[1,2,7], Oren Ram[2–4,7], Noam Shoresh[2,7], Ralph A Sperling[1,6], Alon Goren[5], David A Weitz[1] & Bradley E Bernstein[2–4]

Chromatin profiling provides a versatile means to investigate functional genomic elements and their regulation. However, current methods yield ensemble profiles that are insensitive to cell-to-cell variation. Here we combine microfluidics, DNA barcoding and sequencing to collect chromatin data at single-cell resolution. We demonstrate the utility of the technology by assaying thousands of individual cells and using the data to deconvolute a mixture of ES cells, fibroblasts and hematopoietic progenitors into high-quality chromatin state maps for each cell type. The data from each single cell are sparse, comprising on the order of 1,000 unique reads. However, by assaying thousands of ES cells, we identify a spectrum of subpopulations defined by differences in chromatin signatures of pluripotency and differentiation priming. We corroborate these findings by comparison to orthogonal single-cell gene expression data. Our method for single-cell analysis reveals aspects of epigenetic heterogeneity not captured by transcriptional analysis alone.

2015

Lähnemann *et al. Genome Biology* (2020) 21:31
https://doi.org/10.1186/s13059-020-1926-6

Genome Biology

REVIEW Open Access

Eleven grand challenges in single-cell data science

David Lähnemann[1,2,3], Johannes Köster[1,4], Ewa Szczurek[5], Davis J. McCarthy[6,7], Stephanie C. Hicks[8], Mark D. Robinson[9], Catalina A. Vallejos[10,11], Kieran R. Campbell[12,13,14], Niko Beerenwinkel[15,16], Ahmed Mahfouz[17,18], Luca Pinello[19,20,21], Pavel Skums[22], Alexandros Stamatakis[23,24], Camille Stephan-Otto Attolini[25], Samuel Aparicio[13,26], Jasmijn Baaijens[27], Marleen Balvert[27,28], Buys de Barbanson[29,30,31], Antonio Cappuccio[32], Giacomo Corleone[33], Bas E. Dutilh[28,34], Maria Florescu[29,30,31], Victor Guryev[35], Rens Holmer[36], Katharina Jahn[15,16], Thamar Jessurun Lobo[35], Emma M. Keizer[37], Indu Khatri[38], Szymon M. Kielbasa[39], Jan O. Korbel[40], Alexey M. Kozlov[23], Tzu-Hao Kuo[3], Boudewijn P.F. Lelieveldt[41,42], Ion I. Mandoiu[43], John C. Marioni[44,45,46], Tobias Marschall[47,48], Felix Mölder[1,49], Amir Niknejad[50,51], Lukasz Raczkowski[5], Marcel Reinders[17,18], Jeroen de Ridder[29,30], Antoine-Emmanuel Saliba[52], Antonios Somarakis[42], Oliver Stegle[40,46,53], Fabian J. Theis[54], Huan Yang[55], Alex Zelikovsky[56,57], Alice C. McHardy[3], Benjamin J. Raphael[58], Sohrab P. Shah[59] and Alexander Schönhuth[27,28]*

Abstract

The recent boom in microfluidics and combinatorial indexing strategies, combined with low sequencing costs, has empowered single-cell sequencing technology. Thousands—or even millions—of cells analyzed in a single experiment amount to a data revolution in single-cell biology and pose unique data science problems. Here, we outline eleven challenges that will be central to bringing this emerging field of single-cell data science forward. For each challenge, we highlight motivating research questions, review prior work, and formulate open problems. This compendium is for established researchers, newcomers, and students alike, highlighting interesting and rewarding problems for the coming years.

Summay

- Transcription factors and histone modifications are two categories of functionally important marks of epigenomes.
- ChIP-seq is used to profile protein-DNA interaction information in the epigenomes
- ChIP-seq data analysis
  - MACS for narrow peaks
  - SICER for broad peaks
- Online tools and resources

Thank you very much!

zang@virginia.edu
zanglab.org