

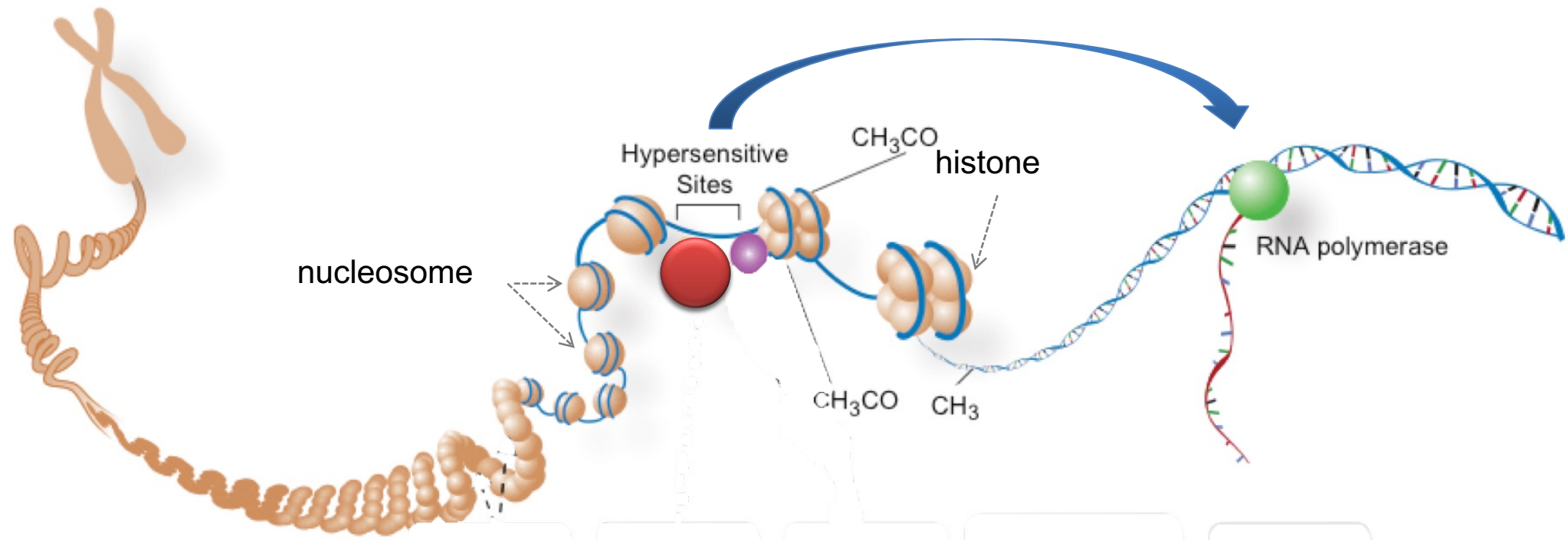
ChIP-seq, Epigenome Profiles, and Peak Detection

September 24, 2025

Outline

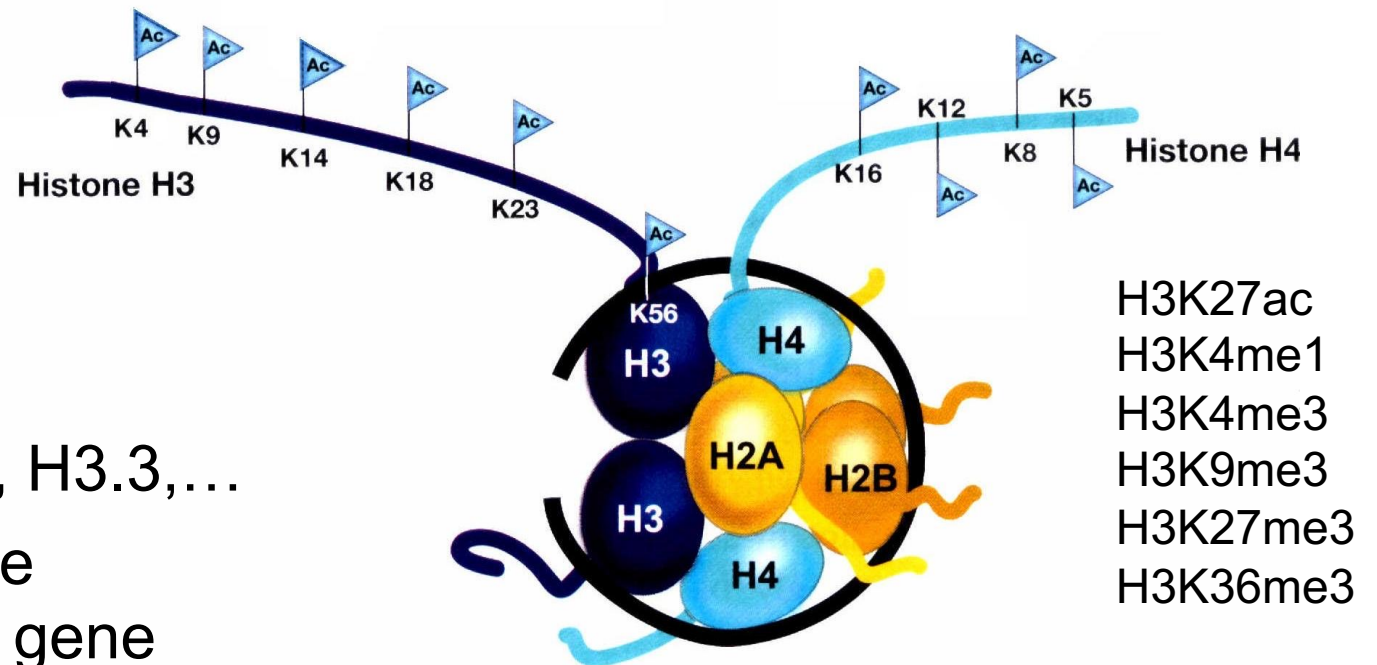
- Epigenome: chromatin and histone marks
- ChIP-seq technology and development
- ChIP-seq data analysis
 - Strategy
 - Peaks vs. islands
 - Peak signal detection (MACS)
 - Island signal detection (SICER)

Epigenome



Histone marks

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



Allis et al. *Epigenetics* 2006

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
Active Enhancer	H3K4me1/H3K27ac
Polycomb Repressed Regions	H3K27me3
Heterochromatin	H3K9me3

Modified from Rivera & Ren *Cell* 2013

The language of covalent histone modifications

Brian D. Strahl & C. David Allis

Department of Biochemistry and Molecular Genetics, University of Virginia Health Science Center, Charlottesville, Virginia 22908, USA

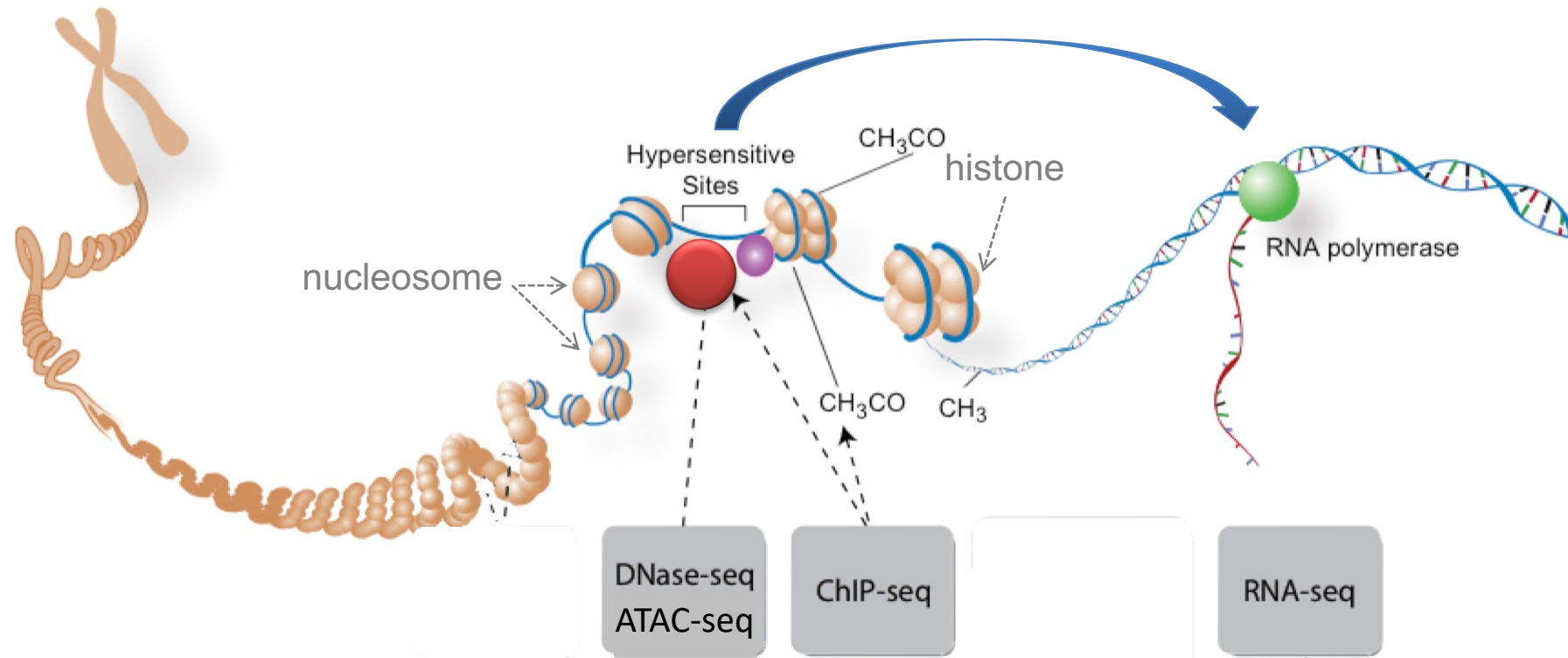
Histone proteins and the nucleosomes they form with DNA are the fundamental building blocks of eukaryotic chromatin. A diverse array of post-translational modifications that often occur on tail domains of these proteins has been well documented. Although the function of these highly conserved modifications has remained elusive, converging biochemical and genetic evidence suggests functions in several chromatin-based processes. We propose that distinct histone modifications, on one or more tails, act sequentially or in combination to form a 'histone code' that is, read by other proteins to bring about distinct downstream events.

How eukaryotic genomes are manipulated within a chromatin environment is a fundamental issue in biology. At the heart of chromatin structure are highly conserved histone proteins (H3, H4, H2A, H2B and H1) that function as building blocks to package eukaryotic DNA into repeating nucleosomal units that are folded into higher-order chromatin fibres^{1,2} (Fig. 1). Once thought of as static, non-participating structural elements, it is now clear that histones are integral and dynamic components of the machinery responsible for regulating gene transcription. The same is probably true for other DNA-templated processes such as replication, repair, recombination and chromosome segregation.

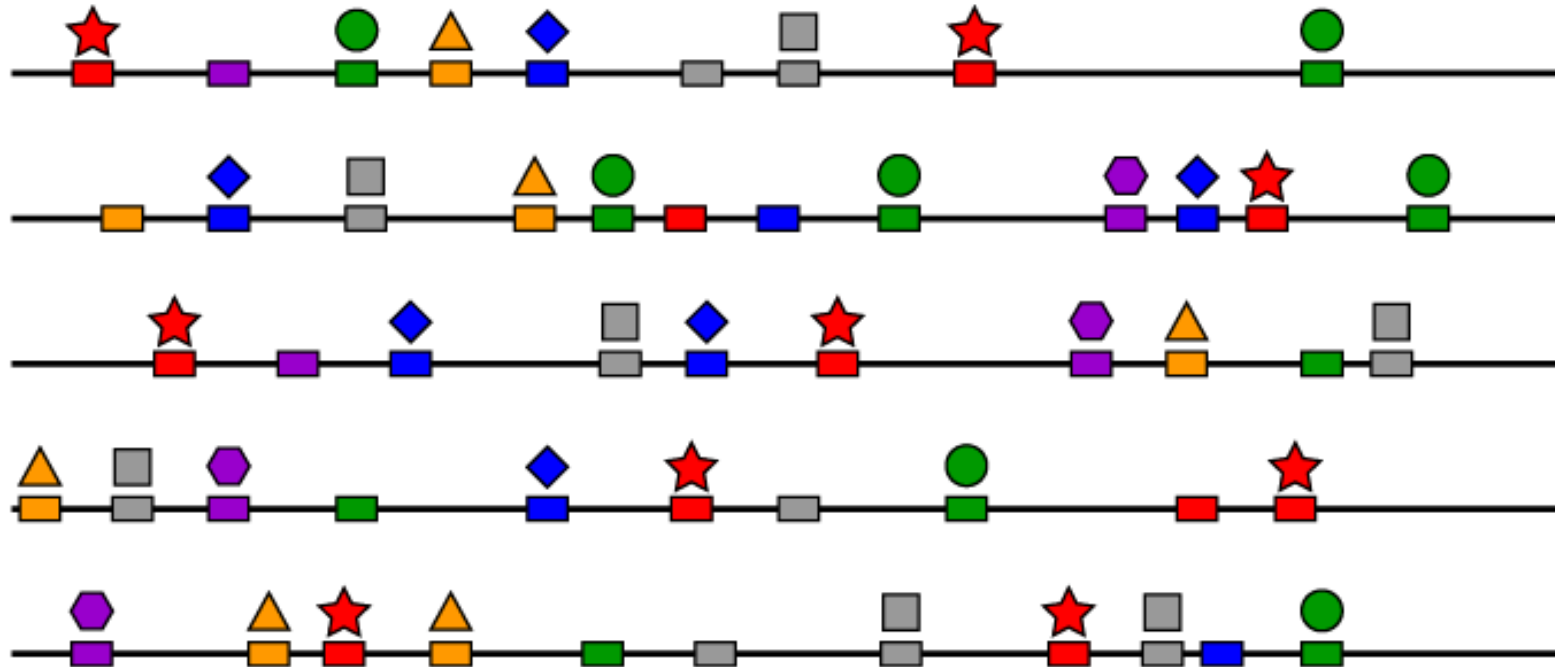
Transcription-linked acetylation, catalysed by the GCN5 family of HATs, shows a preference for lysine 14 of H3 *in vitro*¹⁷ although an expanded set of lysine residues is likely to be used *in vivo*^{18,19}. How is this acetylation site specificity in H3 brought about?

Solution and crystal structure data of various members of the GCN5 HAT family, including co-crystals of the enzyme with H3 tail peptides²⁰, have begun to yield important insights into the enzymatic mechanisms underlying the site specificity of these HATs^{21–25}. One important concept to emerge from these studies is that residues outside the preferred lysine 14 acetylation site in H3 are important for histone-binding specificity. For example, glycine 13 and proline

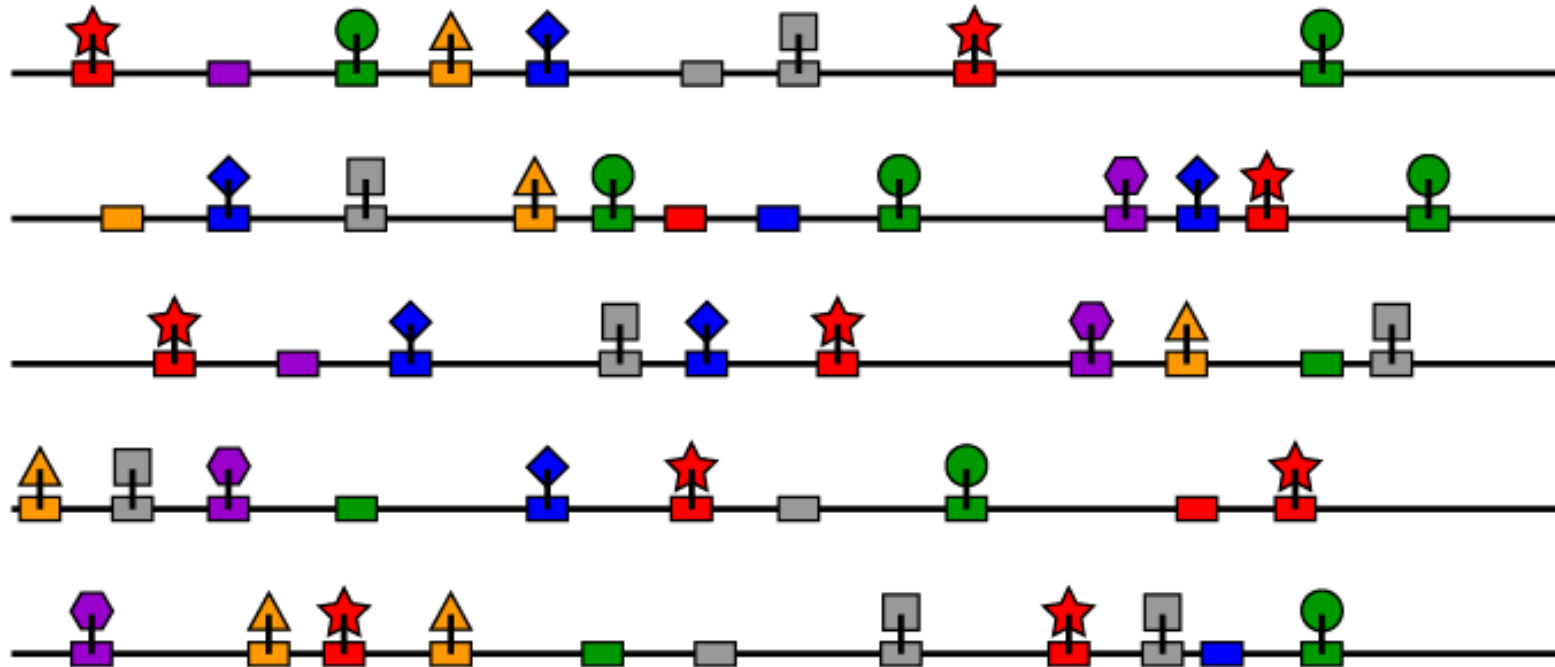
ChIP-seq: To determine the locations in the genome associating with a protein factor



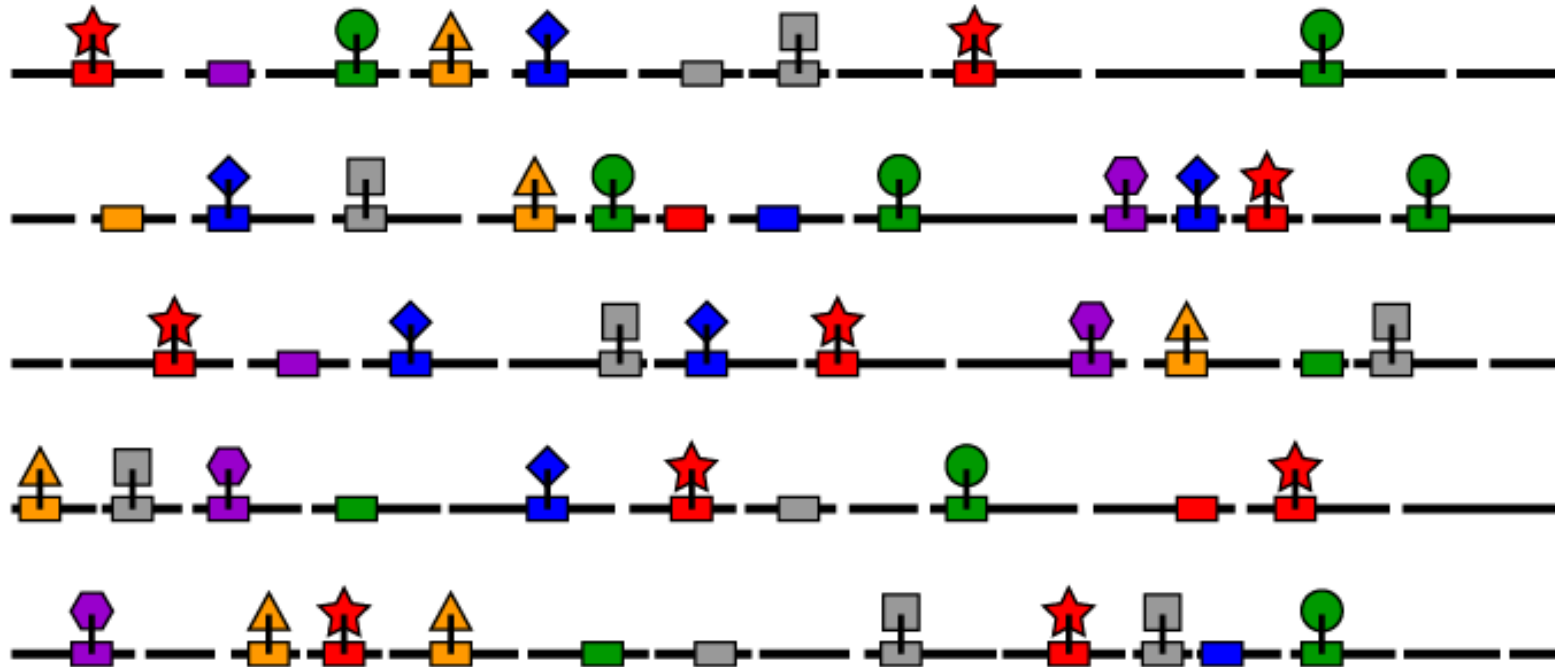
Chromatin ImmunoPrecipitation (ChIP)



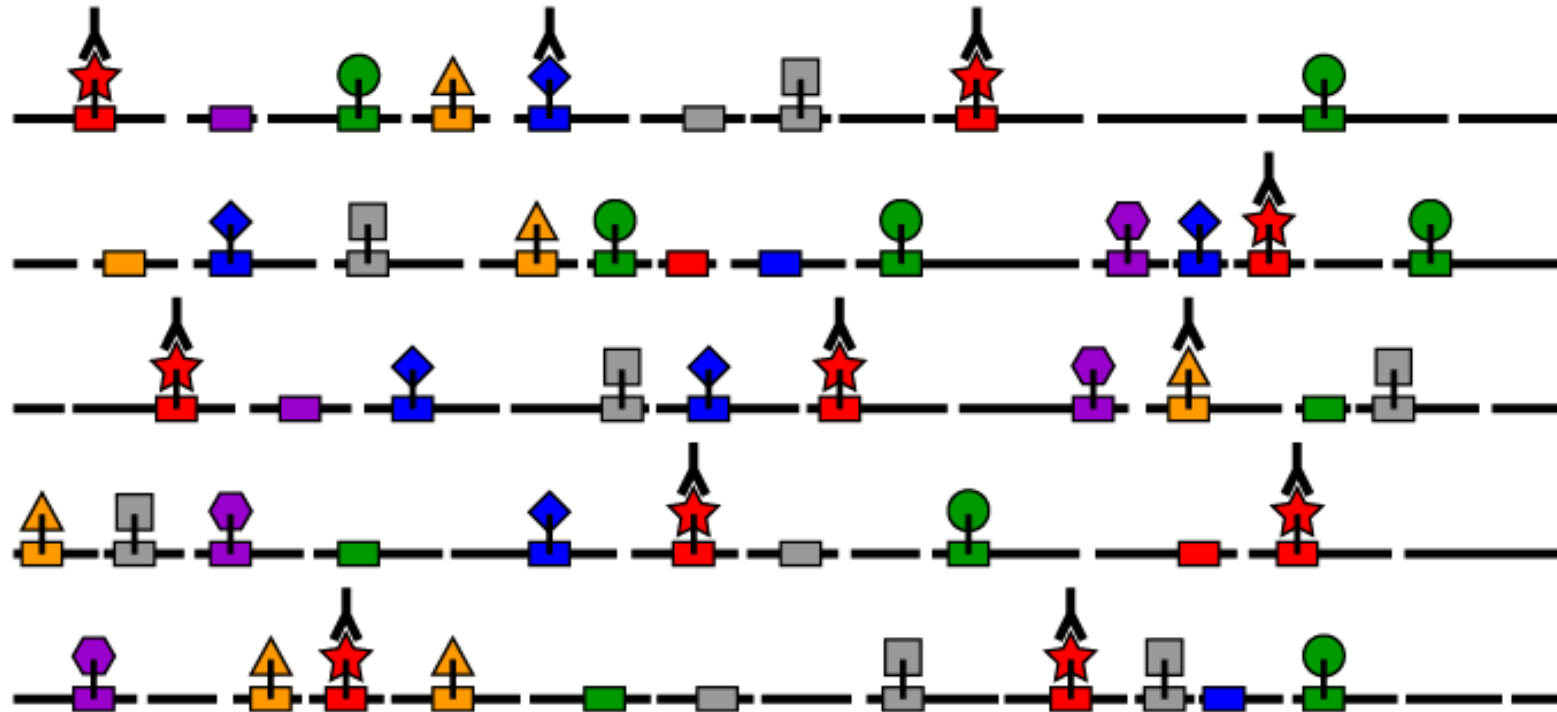
Protein-DNA crosslinking with formaldehyde (for TF)



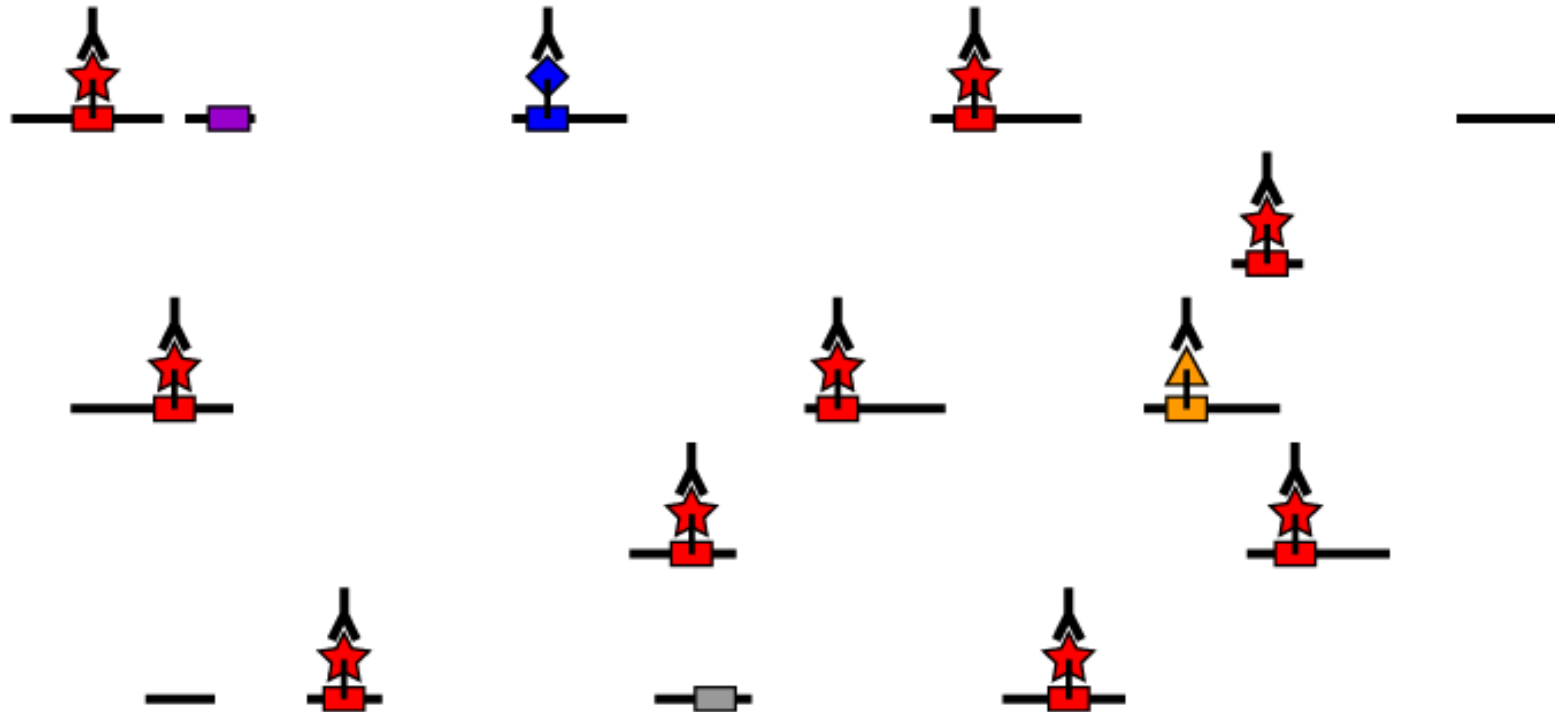
Chop the chromatin using sonication (TF) or micrococcal nuclease (MNase) digestion (histone)



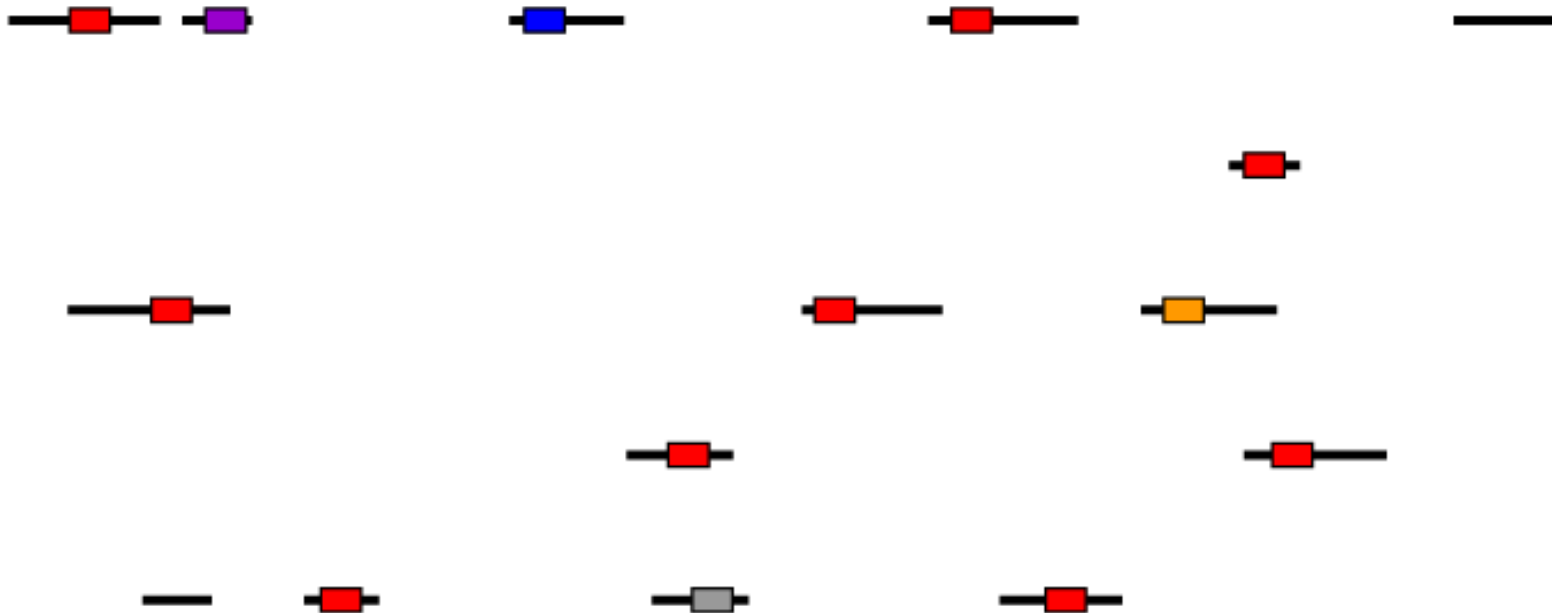
Specific factor-targeting antibody



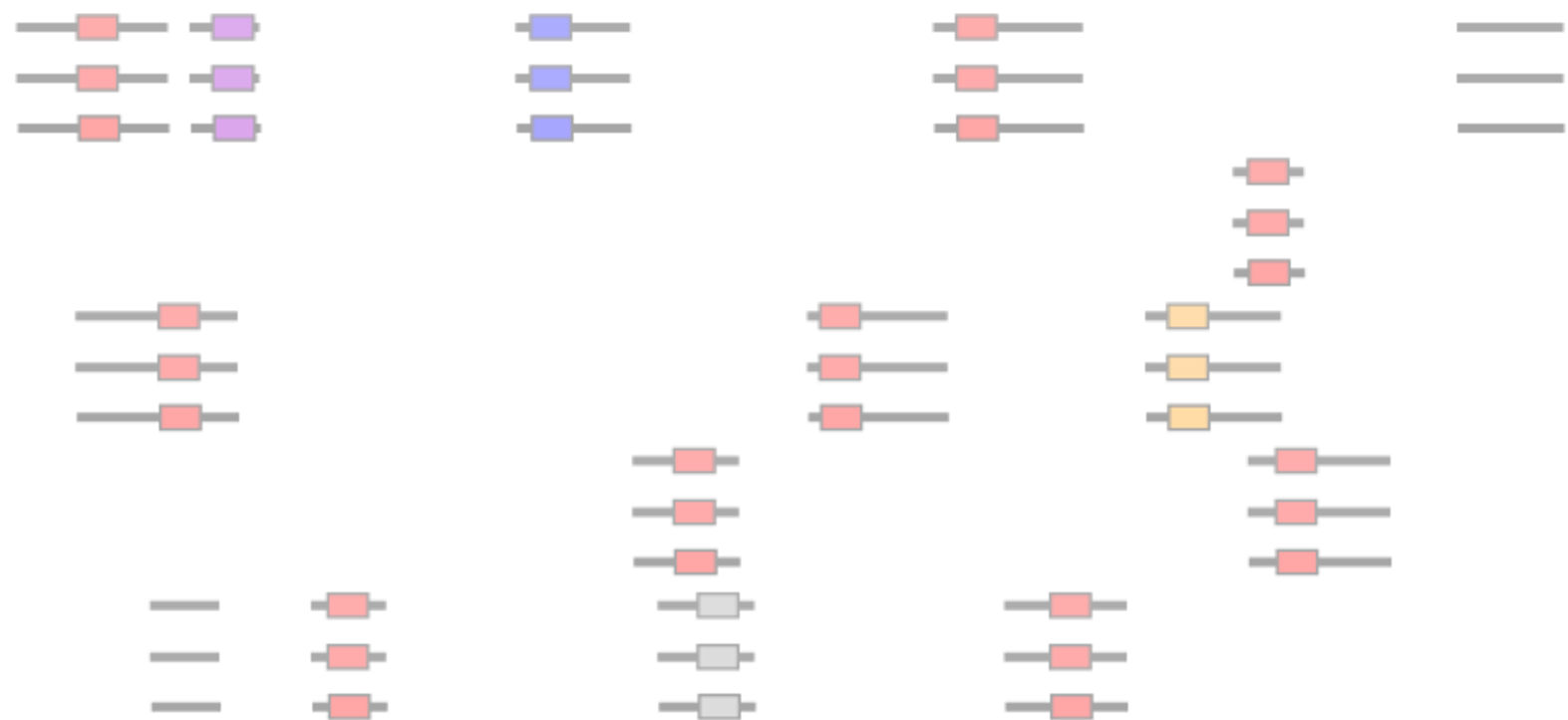
Immunoprecipitation



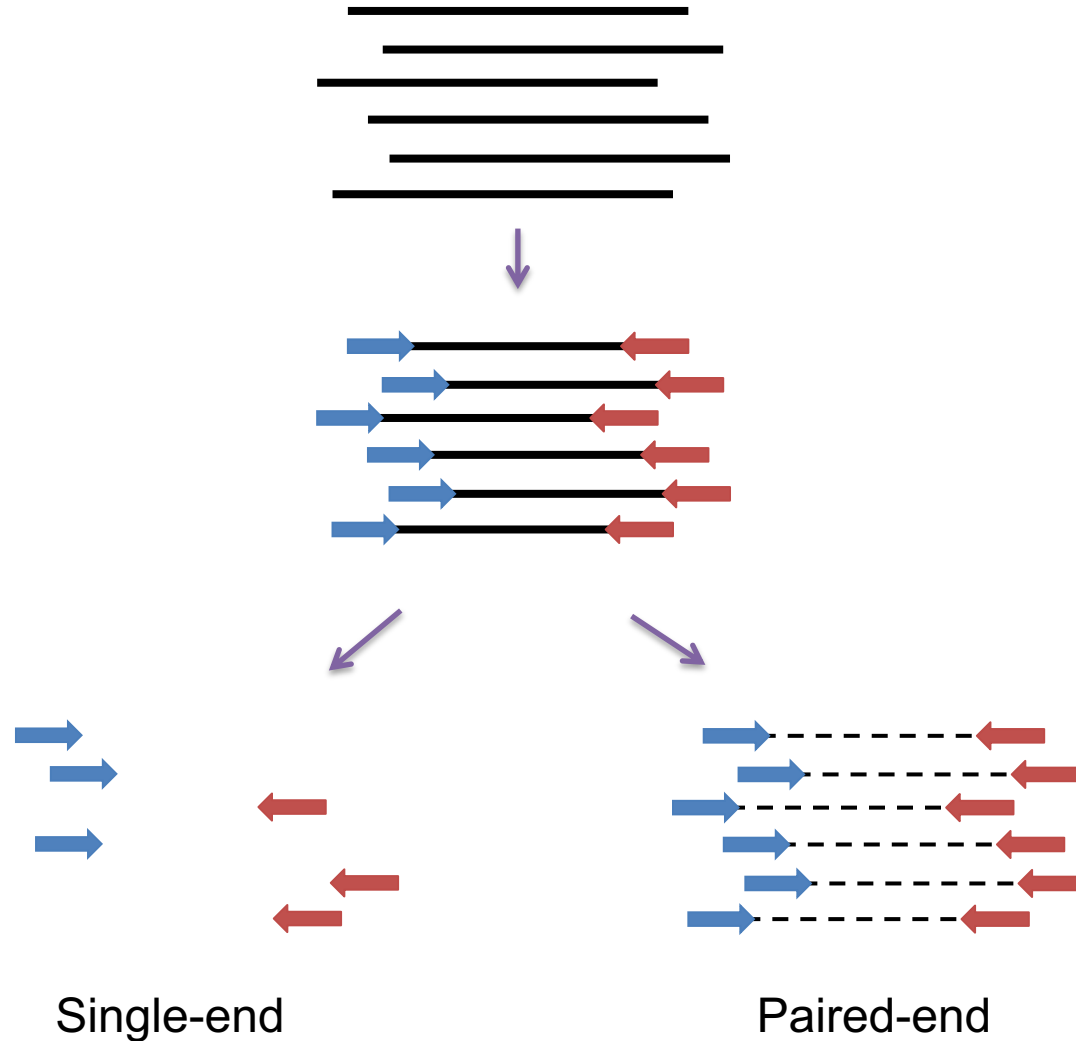
DNA purification

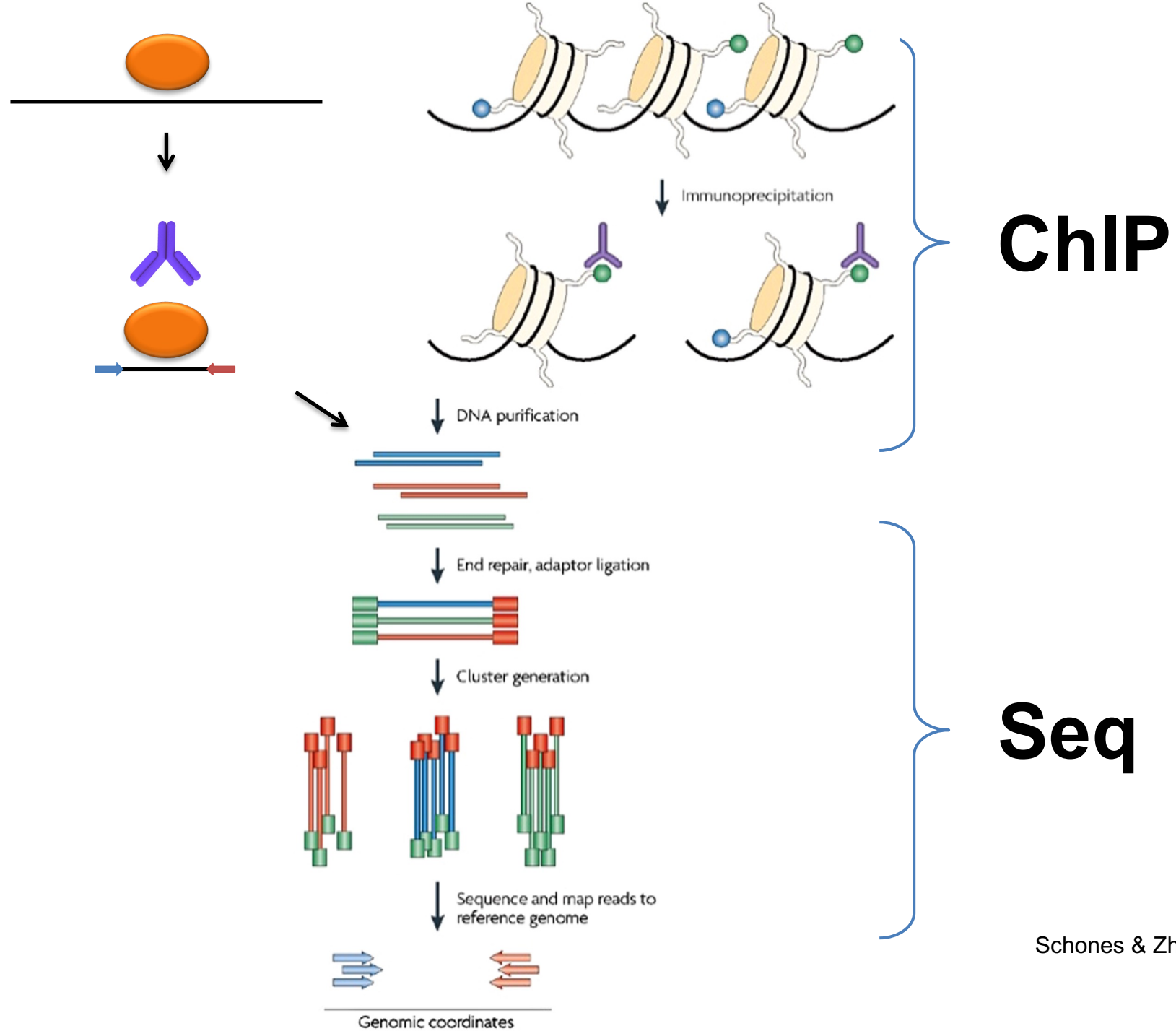


PCR amplification



High-throughput sequencing (Illumina)





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 *cl* 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 μ g 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 NaHCO₃/1% NaDodSO₄.

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

118

LIGATION MEDIATED - PCR OF IP-DNA

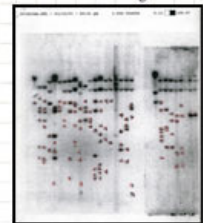
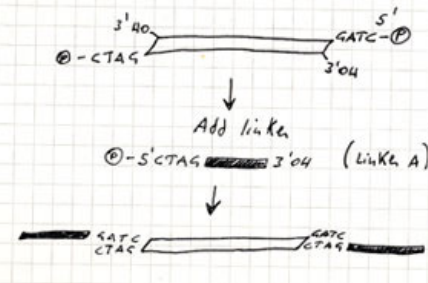
Hypothesis: DNA from IP of chromatin could be simply amplified by PCR using a linker.

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

2) ligate a linker to the IP DNA ~~maintaining~~ maintaining the restriction site

3) amplify the DNA with a linker antisense of the ligated one.

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



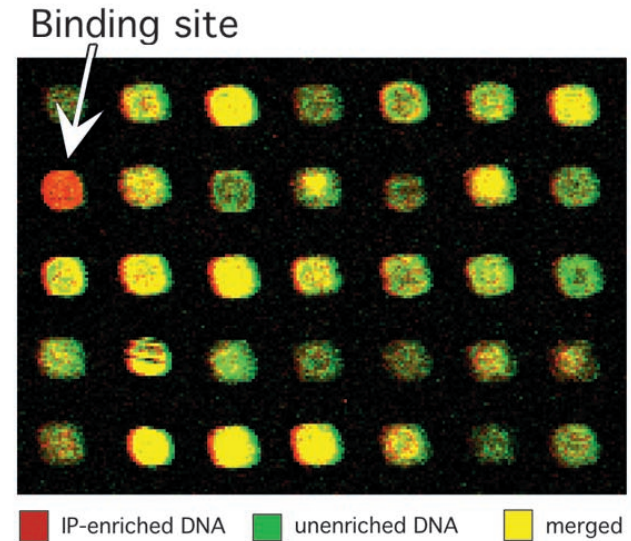
ChIP-chip (2000)

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.



Unbiased chromosomal coverage by tiling array

Cell, Vol. 116, 499–509, February 20, 2004, Copyright ©2004 by Cell Press

Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs

Simon Cawley,^{1,5} Stefan Bekiranov,^{1,5}
Huck H. Ng,^{2,3,4} Philipp Kapranov,¹
Edward A. Sekinger,² Dione Kampa,¹
Antonio Piccolboni,¹ Victor Sementchenko,¹
Jill Cheng,¹ Alan J. Williams,¹ Raymond Wheeler,¹
Brant Wong,¹ Jorg Drenkow,¹ Mark Yamanaka,¹
Sandeep Patel,¹ Shane Brubaker,¹ Hari Tammanna,¹
Gregg Helt,¹ Kevin Struhl,^{2,*}
and Thomas R. Gingeras^{1,*}

¹Affymetrix

3380 Central Expressway
Santa Clara, California 95051

²Department of Biological Chemistry
and Molecular Pharmacology
Harvard Medical School

Boston, Massachusetts 02115

³Department of Biological Sciences
National University of Singapore
Singapore 117543

⁴Genome Institute of Singapore
Singapore 138672

Summary

Using high-density oligonucleotide arrays representing essentially all nonrepetitive sequences on human chromosomes 21 and 22, we map the binding sites in vivo for three DNA binding transcription factors, Sp1, cMyc, and p53, in an unbiased manner. This mapping reveals an unexpectedly large number of transcription factor binding site (TFBS) regions, with a minimal estimate of 12,000 for Sp1, 25,000 for cMyc, and 1600 for p53 when extrapolated to the full genome. Only 22% of these TFBS regions are located at the 5' termini of protein-coding genes while 36% lie within or immediately 3' to well-characterized genes and are significantly correlated with noncoding RNAs. A significant number of these noncoding RNAs are regulated in response to retinoic acid, and overlapping pairs of protein-coding and noncoding RNAs are often coregulated. Thus, the human genome contains roughly comparable numbers of protein-coding and noncoding genes that are bound by common transcription factors and regulated by common environmental signals.

ChIP-seq (2007)

Resource

Cell

High-Resolution Profiling of Histone Methylation 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,*}

¹Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA

²Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

³These 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 mono-methylations 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 com-

putational 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

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

LETTERS

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

Istvan Albert¹, Travis N. Mavrich^{1,2}, Lynn P. Tomsho¹, Ji Qi¹, Sara J. Zanton^{1,2}, Stephan C. Schuster¹ & 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 expression¹. 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 complex². A nucleosome has two fundamental relationships with its DNA³. 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 sequence⁴⁻⁵. Indeed, a tendency of AA/TT dinucleotides to recur in 10-bp intervals and in counter-phase with GC dinucleotides generates a curved DNA structure that favours nucleosome formation⁵. 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 regions⁸⁻¹¹, 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?

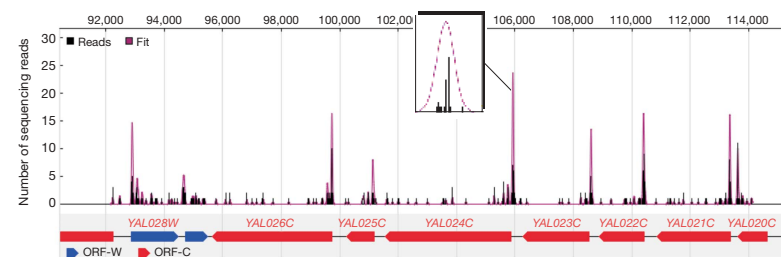


Figure 1 | Distribution of H2A.Z nucleosomal DNA at an arbitrary region of the yeast genome. Any region of the genome can be viewed in this way at <http://nucleosomes.sysbio.bx.psu.edu>. An enlarged view of a peak is shown in the inset, where each vertical bar corresponds to the number of

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.

¹Center for Comparative Genomics and Bioinformatics, ²Center 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	ChIP factors	# citations (9/23/25)
Translational and rotational settings of H2A.Z nucleosomes across the <i>Saccharomyces cerevisiae</i> genome	Albert...Pugh	<i>Nature</i>	10/20/2006	1/26/2007	3/29/2007	Yeast	H2A.Z	954
High-resolution profiling of histone methylations in the human genome	Barski, Cuddapah, Cui, Roh, Schones, Wang, Wei,..., Zhao	<i>Cell</i>	4/20/2007	5/3/2007	5/17/2007	Human CD4 ⁺ T cells	20 histone methylations, H2A.Z, PolII, CTCF	8494
Genome-wide mapping of in vivo protein-DNA interactions	Johnson, Mortazavi; Myers, Wold	<i>Science</i>	2/14/2007	4/26/2007	5/31/2007	Human Jurkat cell line	NRSF (REST)	3882
Genome-wide maps of chromatin state in pluripotent and lineage-committed cells	Mikkelsen,..., Lander, Bernstein	<i>Nature</i>	5/10/2007	6/13/2007	7/1/2007	Mouse ESC, NPC, MEF	4 histone methylations, PolII, H3	5122

First ChIP-seq for histone modifications

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 Chepelev², and Keji Zhao^{1,*}

¹Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA

²Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

³These 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 mono-methylations 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

nature
genetics

LETTERS

Combinatorial patterns of histone acetylations and methylations in the human genome

Zhibin Wang^{1,5}, Chongzhi Zang^{2,5}, Jeffrey A Rosenfeld³⁻⁵, Dustin E Schones¹, Artem Barski¹, Suresh Cuddapah¹, Kairong Cui¹, Tae-Young Roh¹, Weiqun Peng², Michael Q Zhang³ & Keji Zhao¹

Histones are characterized by numerous posttranslational modifications that influence gene transcription^{1,2}. However, because of the lack of global distribution data in higher eukaryotic systems³, the extent to which gene-specific combinatorial patterns of histone modifications exist remains to be determined. Here, we report the patterns derived from the analysis of 39 histone modifications in human CD4⁺ T cells. Our data indicate that a large number of patterns are associated with promoters and enhancers. In particular, we identify a common modification module consisting of 17 modifications detected at 3,286 promoters. These modifications tend to colocalize in the genome and correlate with each other at an individual nucleosome level. Genes associated with this module tend to have higher expression, and addition of more modifications to this module is associated with further increased expression. Our data suggest that these histone modifications may act cooperatively to prepare chromatin for transcriptional activation.

Histones are subject to numerous covalent modifications, including methylation and acetylation, that occur mainly at their N-terminal tails and that can affect transcription of genes^{1,2,4,5}. Extensive studies have established that histone acetylation is primarily associated with gene activation, whereas methylation, depending on its position and state, is associated with either repression or activation⁵⁻¹⁰. Various models, including the histone code, the signaling network and the charge neutralization model, have been proposed to account for the function of histone modifications¹¹⁻¹⁴. The histone code hypothesis

level (see Methods section for data deposition), and analyzed these together with the H2A.Z and 19 histone methylation maps we generated previously¹⁵.

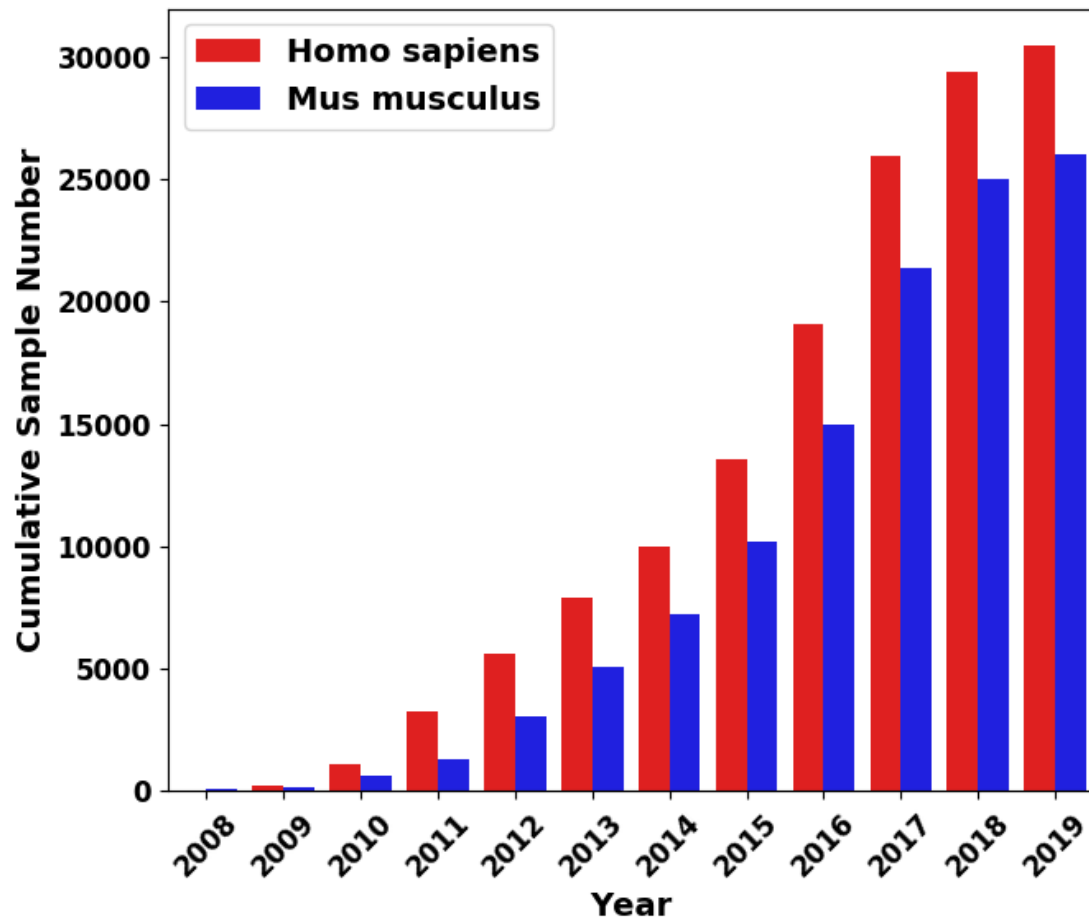
We first systematically evaluated the specificities of the acetylation antibodies used in this study (Supplementary Methods, Supplementary Table 1 and Supplementary Fig. 1 online). Competition assays using modified and unmodified peptides indicated that most antibodies showed specificity for the desired acetylation (Supplementary Fig. 1). The H4K5ac and H3K4ac antibodies demonstrated some crossreactivity toward H4K12ac and H3K9ac, respectively, in a condition with excess competitor peptides (Supplementary Fig. 1d,j), and the H4K91ac antibody did not work in protein blotting. Thus, the results for these modifications should be interpreted with caution. Of note, H2AK9ac has not been reported previously, and H3K4ac has only been identified by mass-spectrometry analysis and has not been previously characterized functionally¹⁶. Protein blotting indicated that these acetylations indeed exist in human CD4⁺ T cells (Supplementary Fig. 1j,o). We previously analyzed the genome-wide distribution of H2BK5me1 (ref. 15), and protein blotting data in this study indicated that this methylation exists in human cells and that the H2BK5me1 antibody is specific (Supplementary Fig. 1p).

Next, we determined the genomic distribution patterns of these histone acetylations using the ChIP-Seq technique¹⁵, which we previously confirmed yields H3K4me3 distribution patterns similar to those generated by the ChIP-SAGE (GMAT) strategy^{15,17}. To validate the histone acetylation data, we compared the genomic distribution patterns of the K9/K14-diacetylated histone H3 from ChIP-SAGE¹⁸ with the separately examined patterns of H3K9ac and H3K14ac in

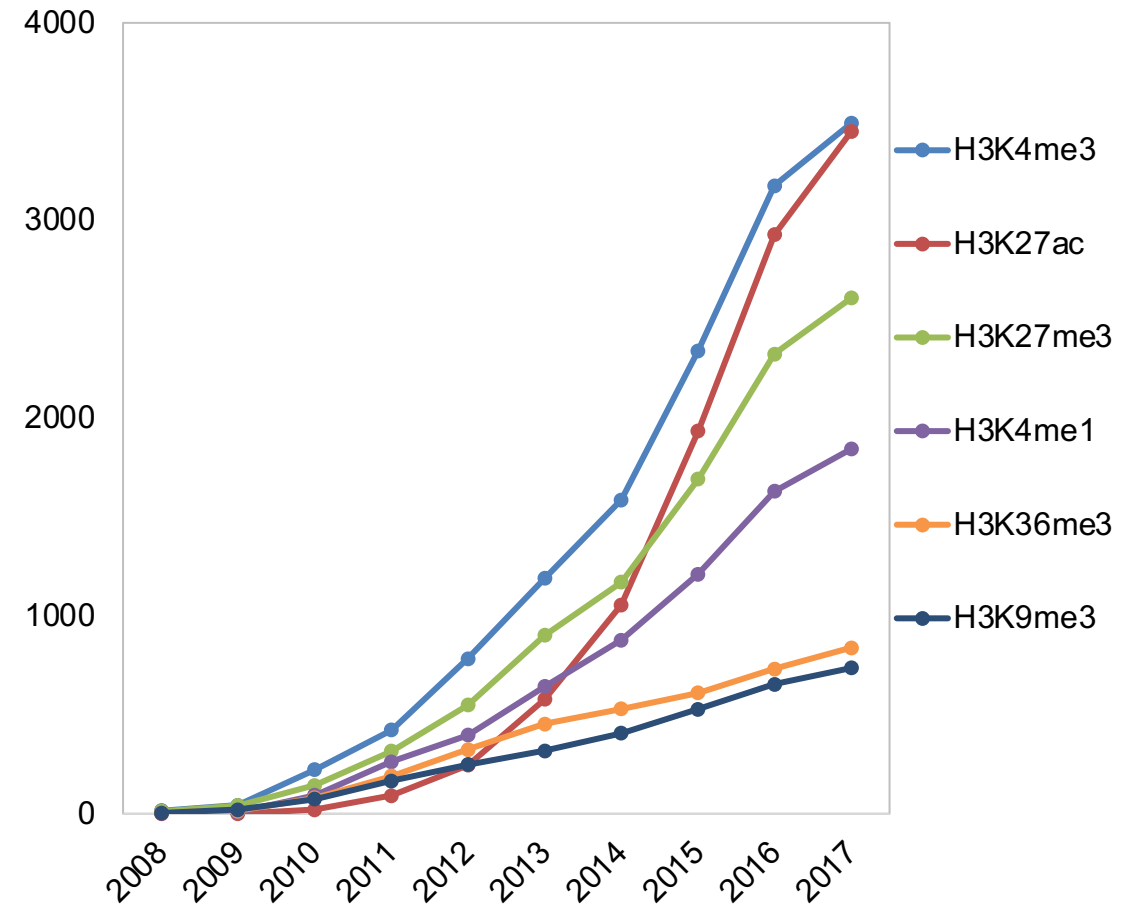
© 2008 Nature Publishing Group <http://www.nature.com/naturegenetics>

npg

ChIP-seq has become a predominant method for profiling chromatin epigenomes



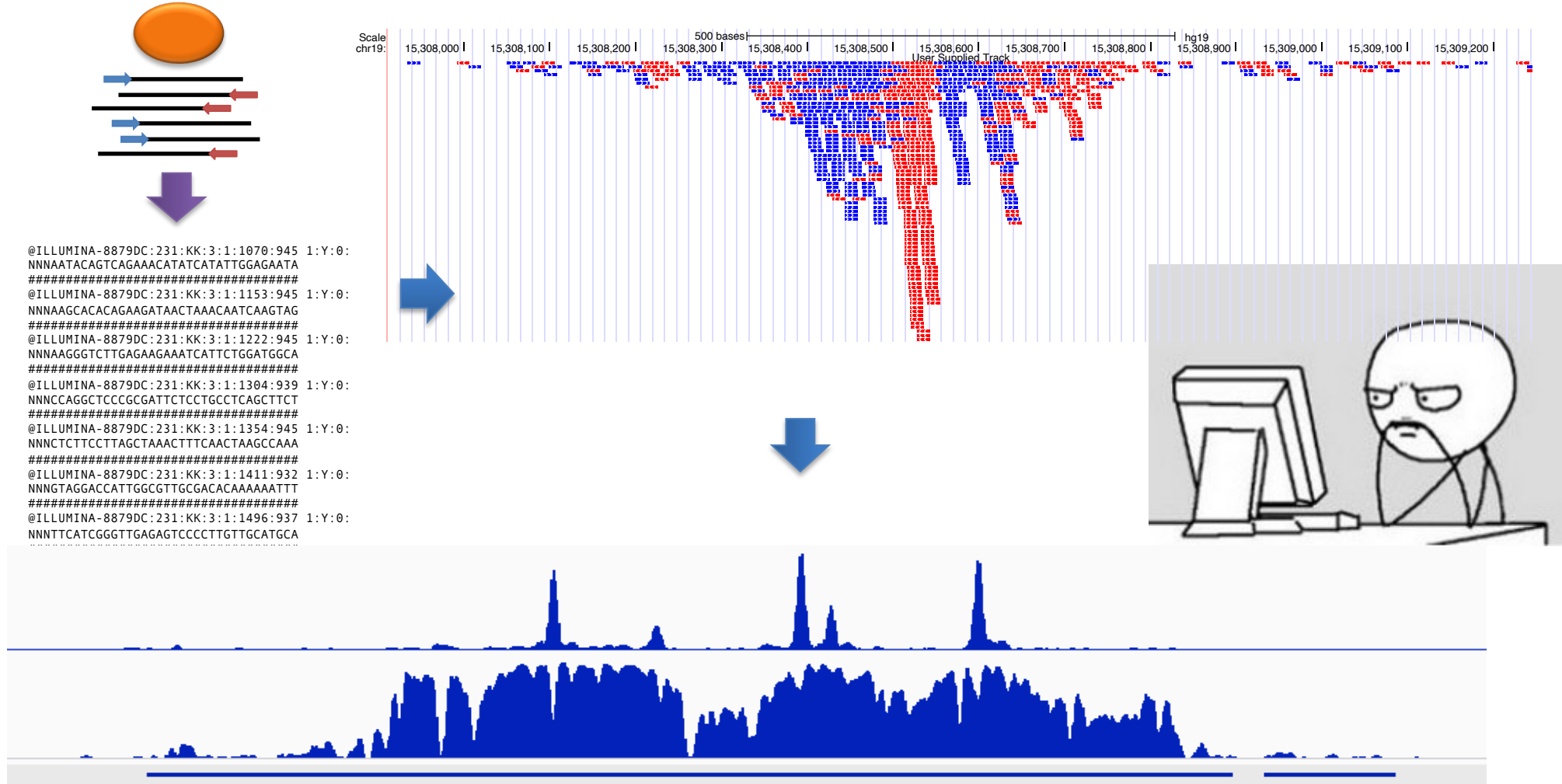
cistrome.org/db



Outline

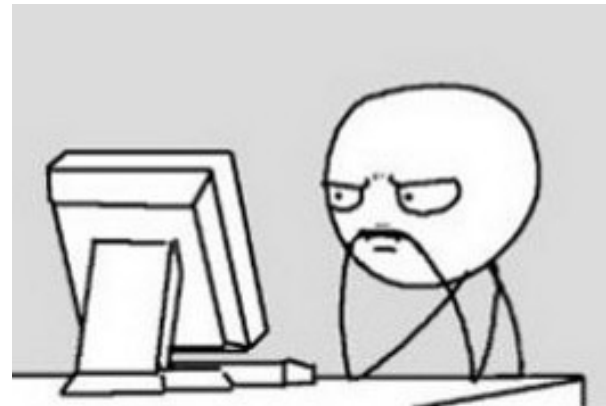
- Epigenome: chromatin and histone marks
- ChIP-seq technology and development
- **ChIP-seq data analysis**
 - Strategy
 - Peaks vs. islands
 - Peak signal detection (MACS)
 - Island signal detection (SICER)

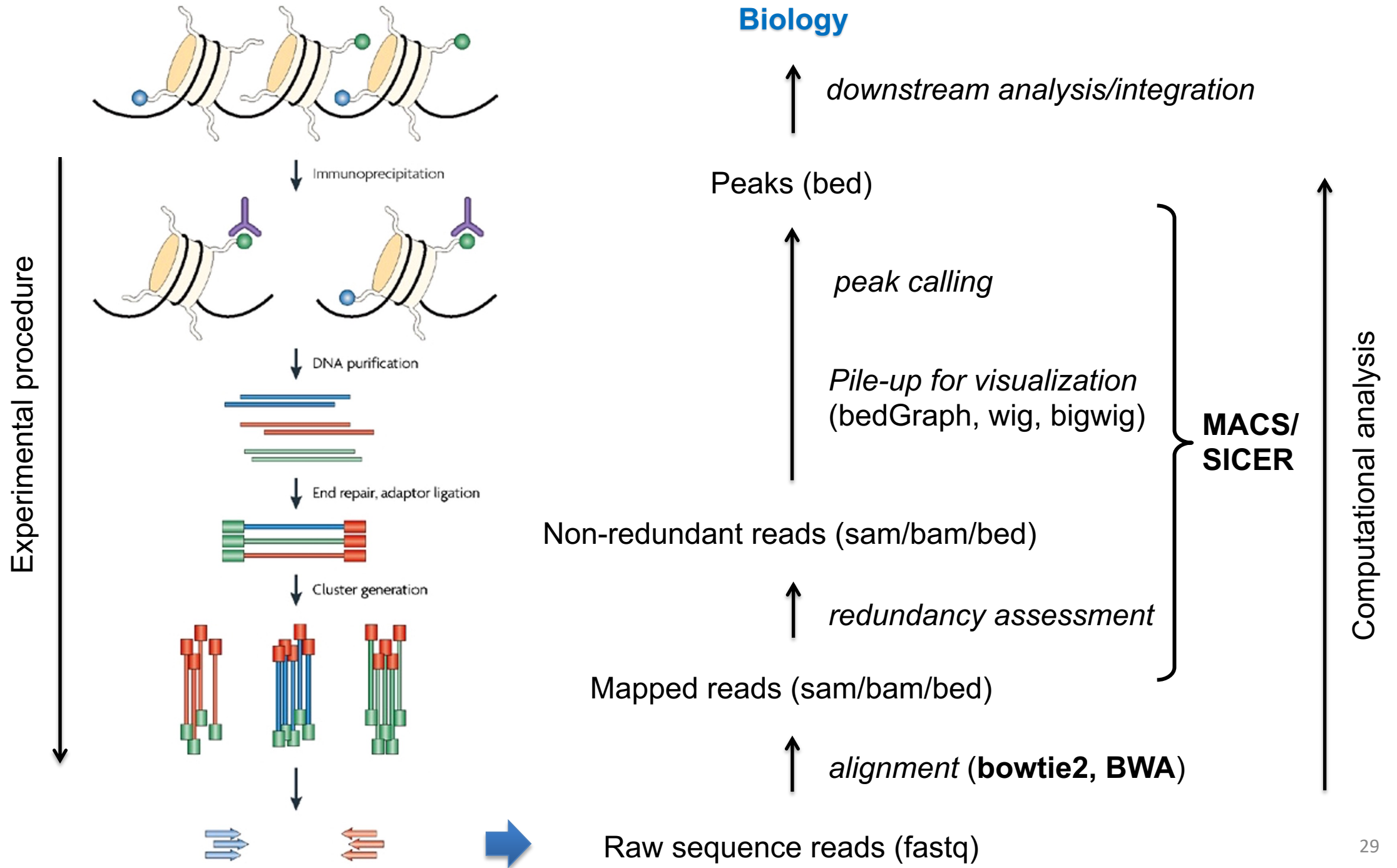
ChIP-seq data analysis overview



ChIP-seq data analysis goals

- 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



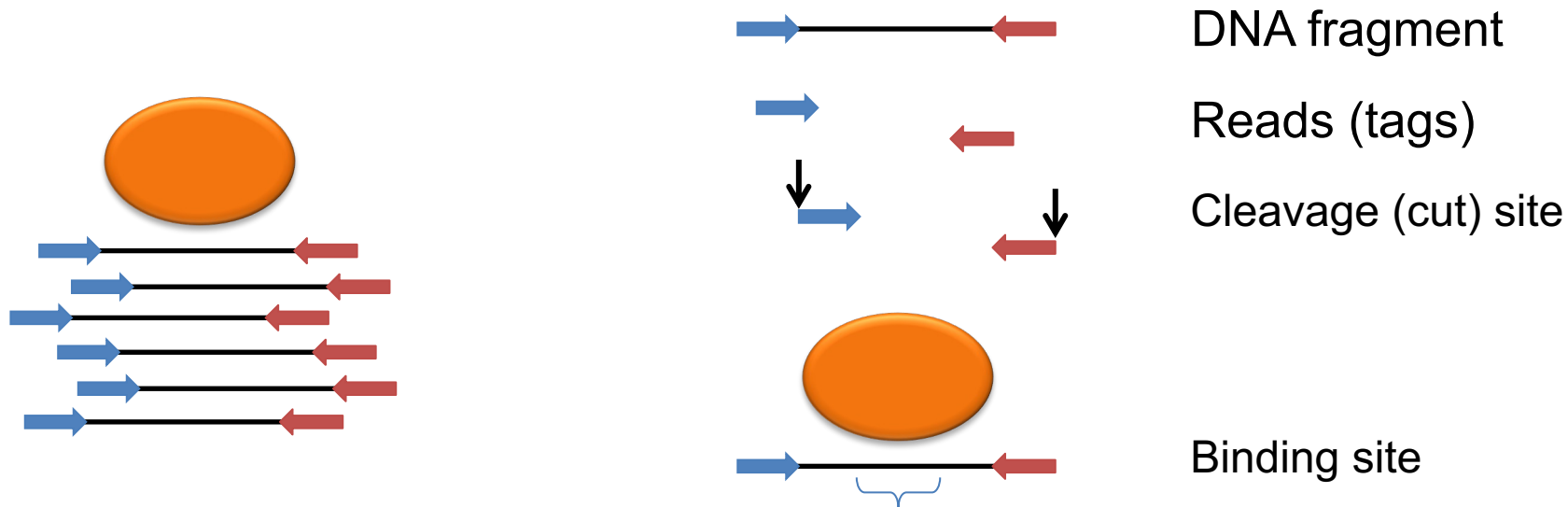


ChIP-seq read mapping

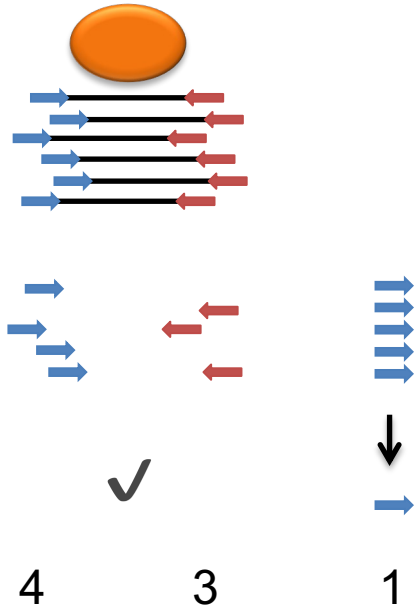
- alignment of each sequence read: **bowtie**, **BWA** (Burrows–Wheeler Algorithm)

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

- Concepts/Terminology:



Redundancy control



mapped reads: 12
 # non-redundant reads: 8
 # locations w/ reads: 8
 # locations w/ 1 read: 7

- Non-redundant rate:

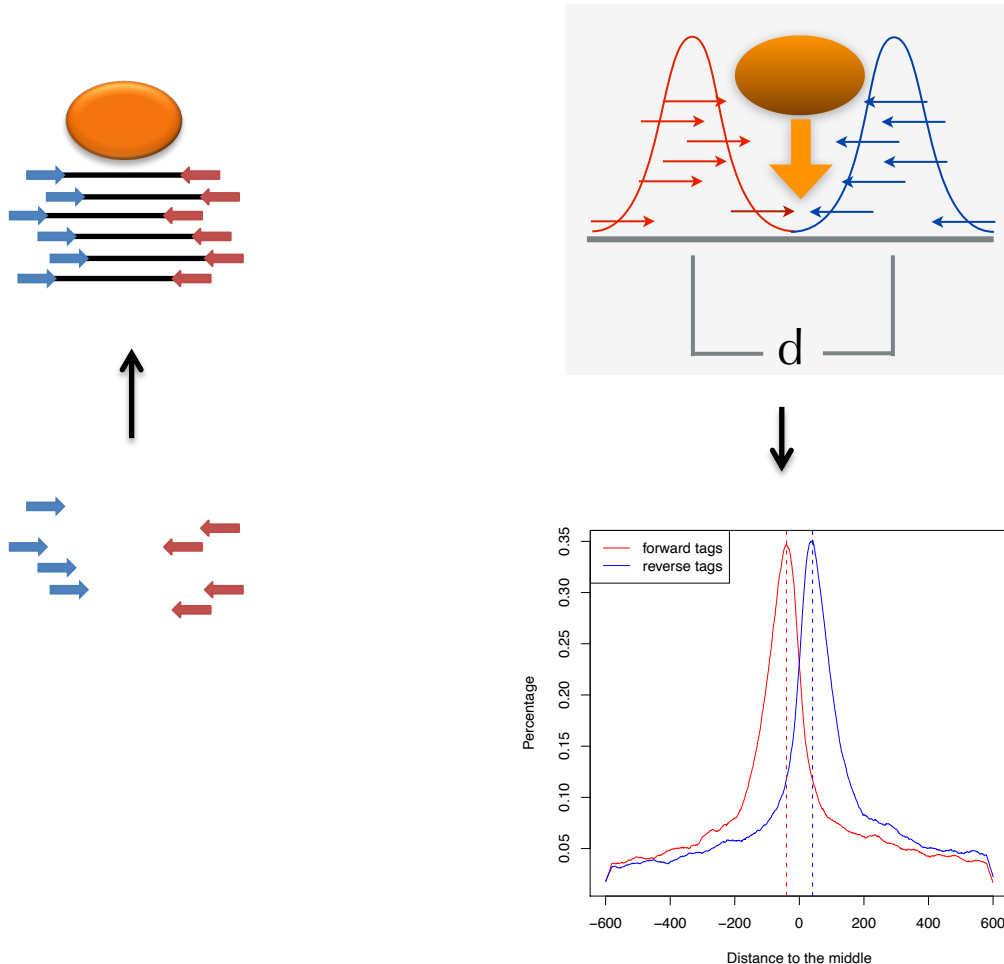
$$\frac{\text{\# non-redundant reads}}{\text{\# mapped reads}} = 8/12 = 66.7\%$$

- PBC (PCR Bottleneck Coefficient):

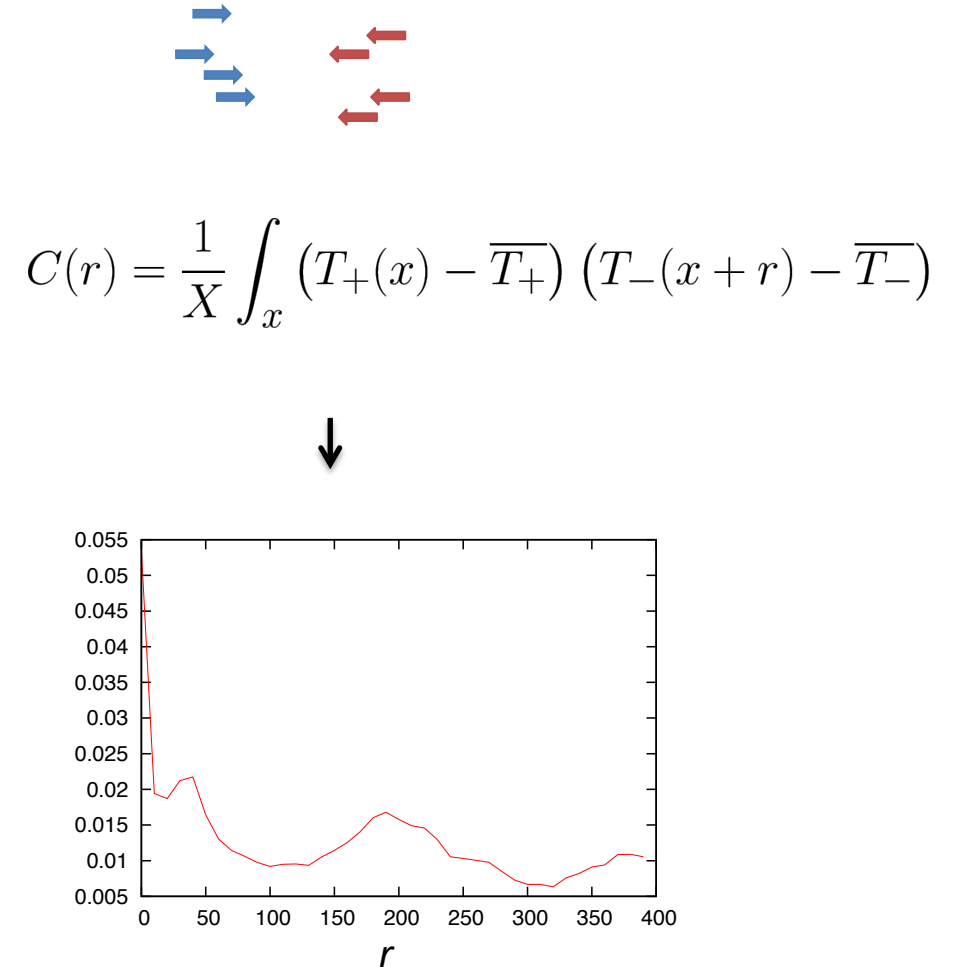
$$\frac{\text{\# locations w/ 1 read}}{\text{\# locations w/ reads}} = 7/8 = 87.5\%$$

DNA fragment size estimation

peak model (MACS)



cross-correlation (SICER)



Cross-correlation

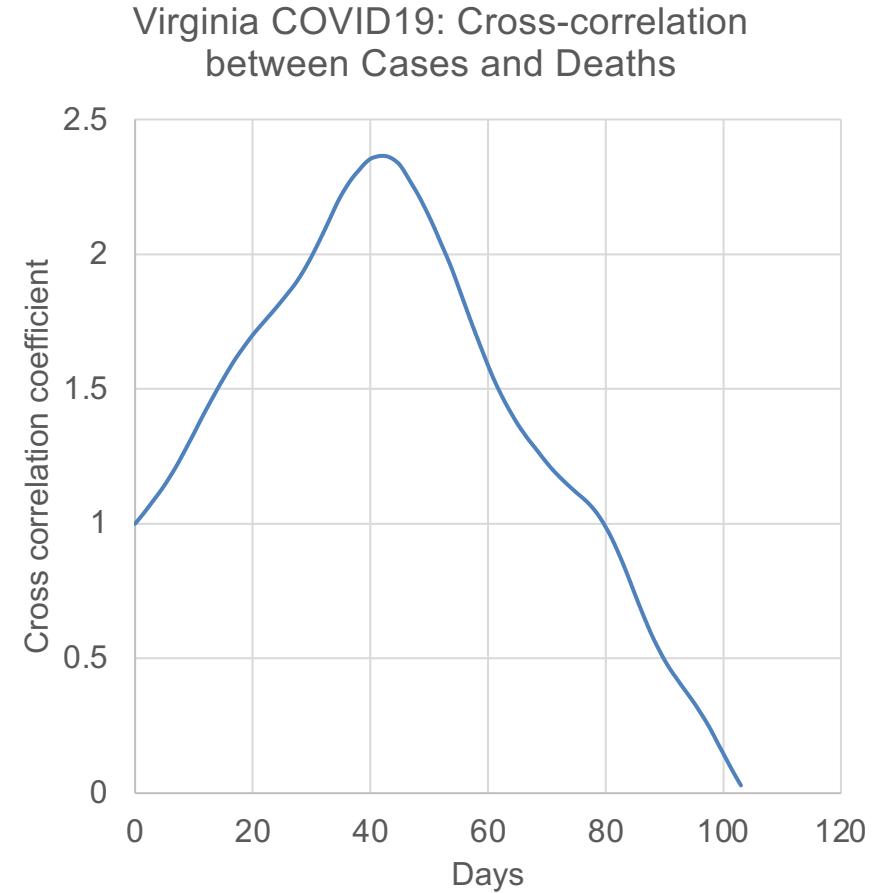
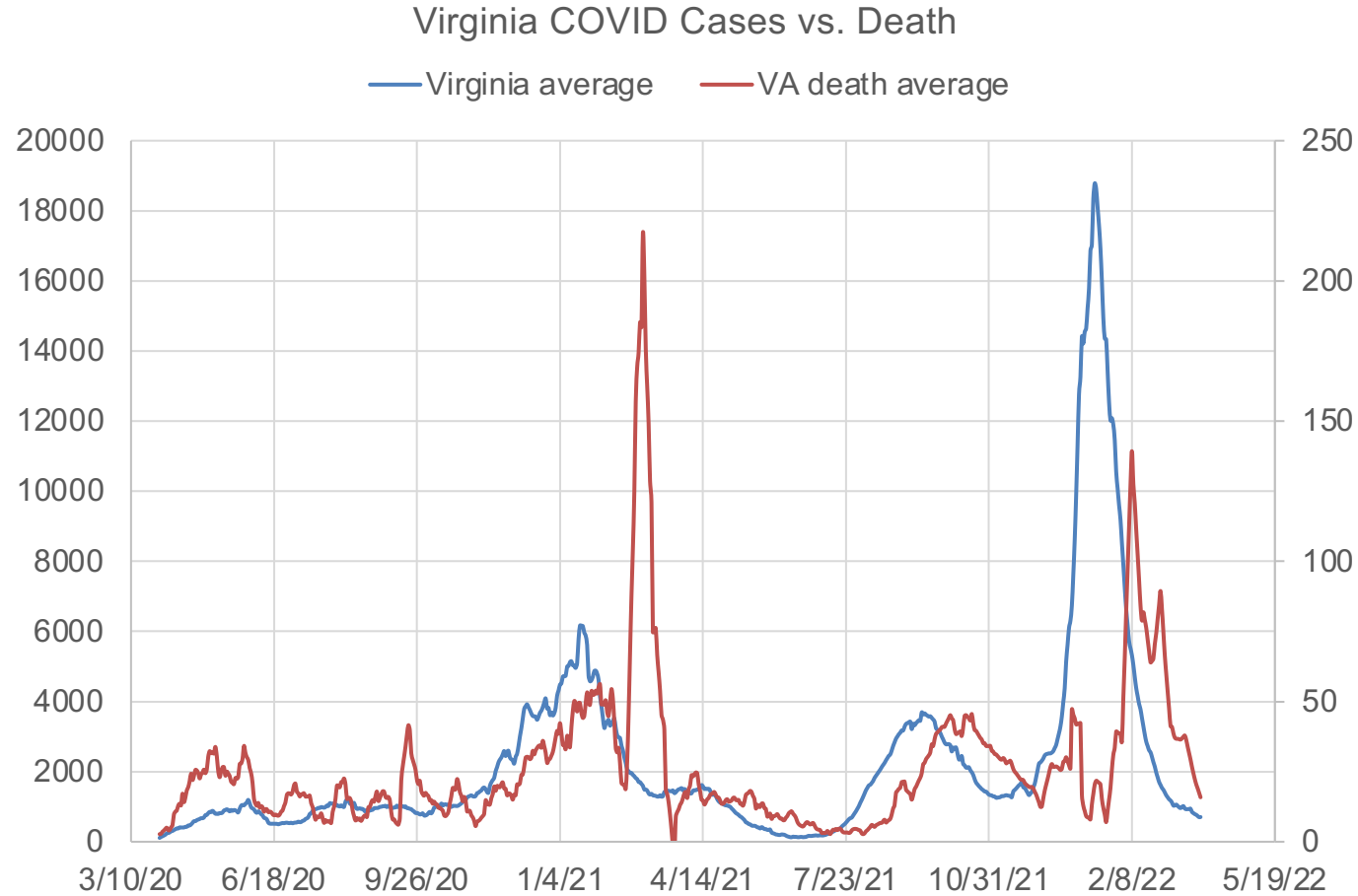
- Correlation between two strings with a displacement

$$R_{XY}(r) = \sum_t (X(t) - \overline{X}) (Y(t + r) - \overline{Y})$$

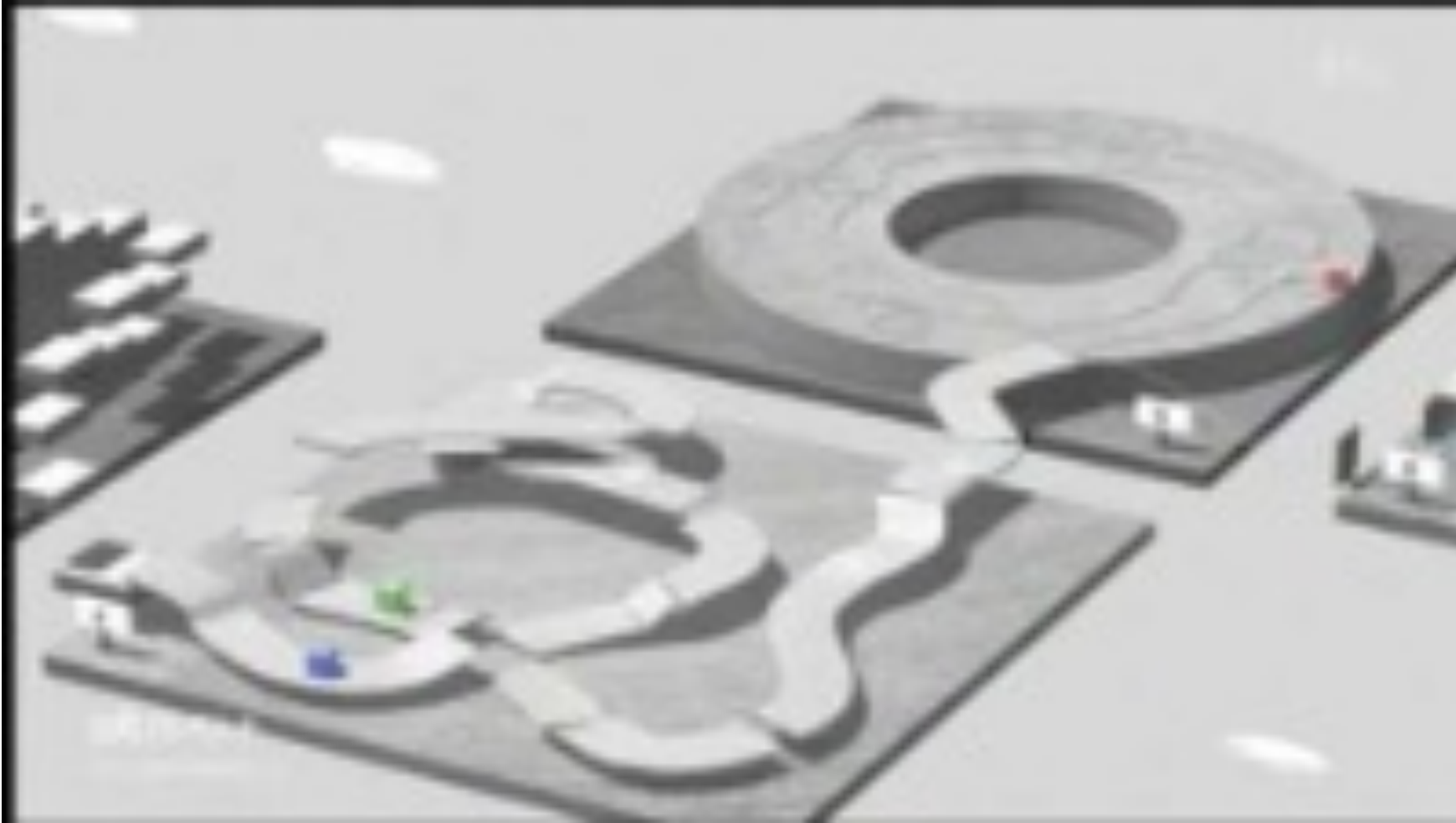
- Auto-correlation: Cross-correlation with itself

$$R_{XX}(r) = \sum_t (X(t) - \overline{X}) (X(t + r) - \overline{X})$$

Cross-correlation: example



Auto-correlation: example



<https://youtu.be/B7BFgCXCqGU>

Auto-correlation: example

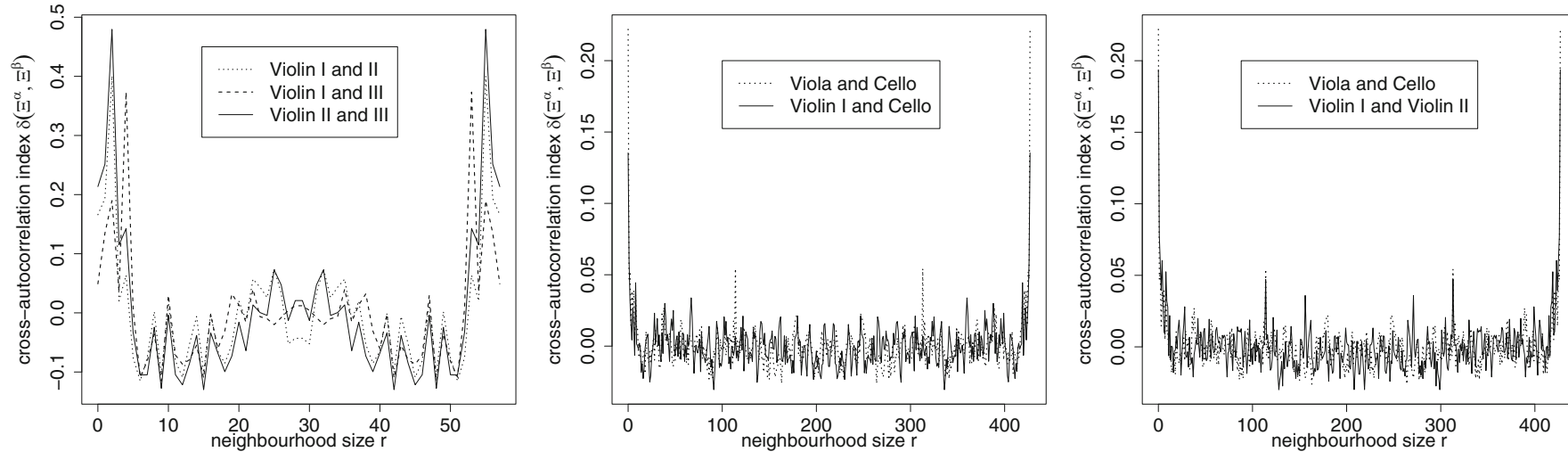
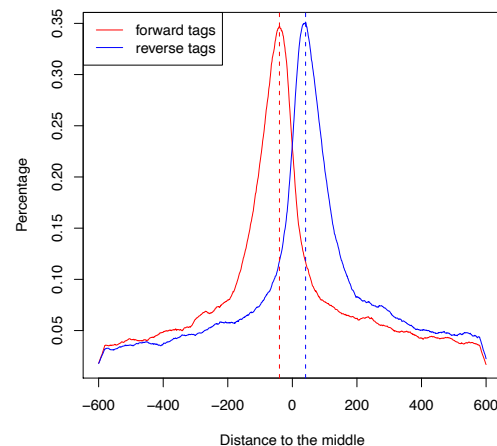
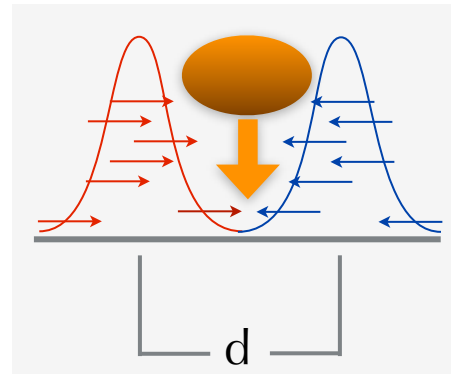
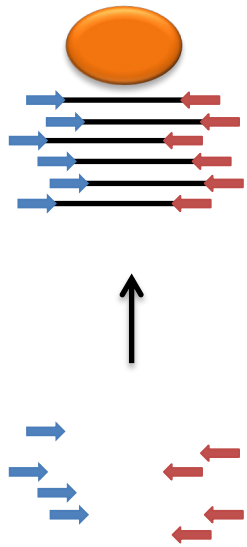


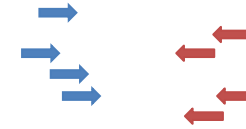
Fig. 6 Cross-autocorrelation index according to the lag r varying from 0 to n . *Left: Canon in D Major by Pachelbel with τ equal to a measure. Middle and right: first movement of the String Quartet No. 1 in F major, Op. 18 by Beethoven with τ equal to a measure*

DNA fragment size estimation

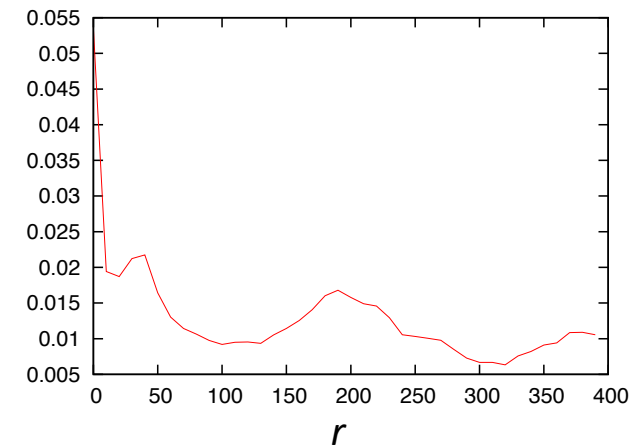
Peak model (MACS)
For TF



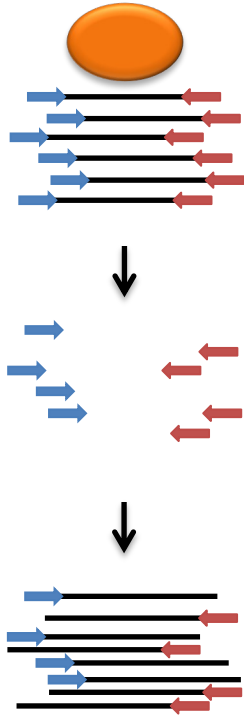
Cross-correlation (SICER)
for any ChIP-seq (input)



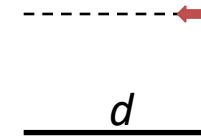
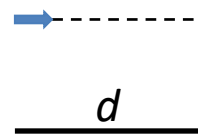
$$C(r) = \frac{1}{X} \int_x (T_+(x) - \overline{T_+}) (T_-(x+r) - \overline{T_-})$$



Retrieve DNA fragments

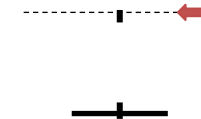
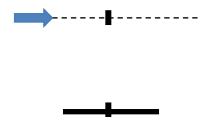


- Full length retrieval (MACS)



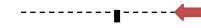
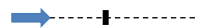
Extend d

- Partial retrieval (sharpen the signal)



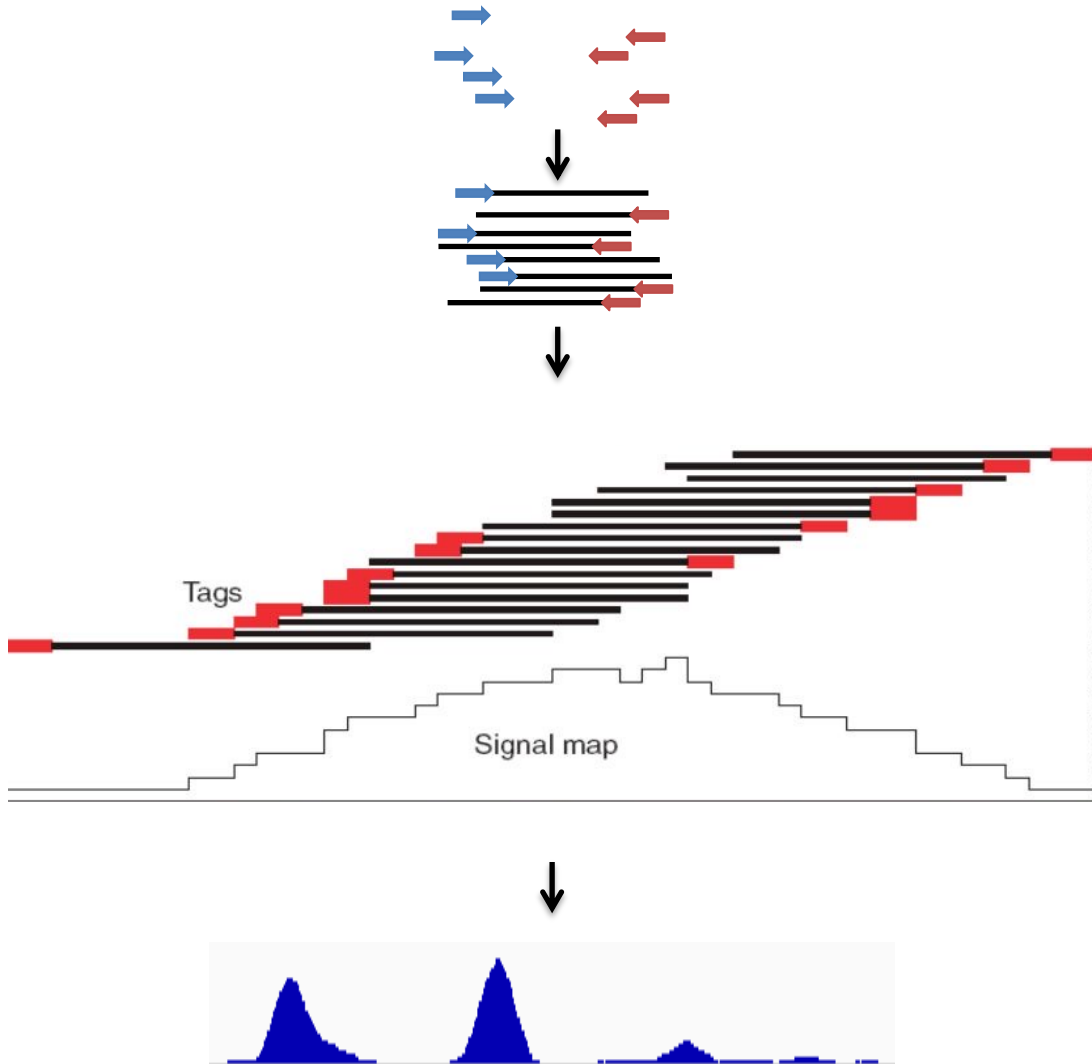
Shift $d/2$
Expand to $d/2$

- Point retrieval (SICER)



Shift $d/2$

Pile up: Signal map generation



- bedGraph:

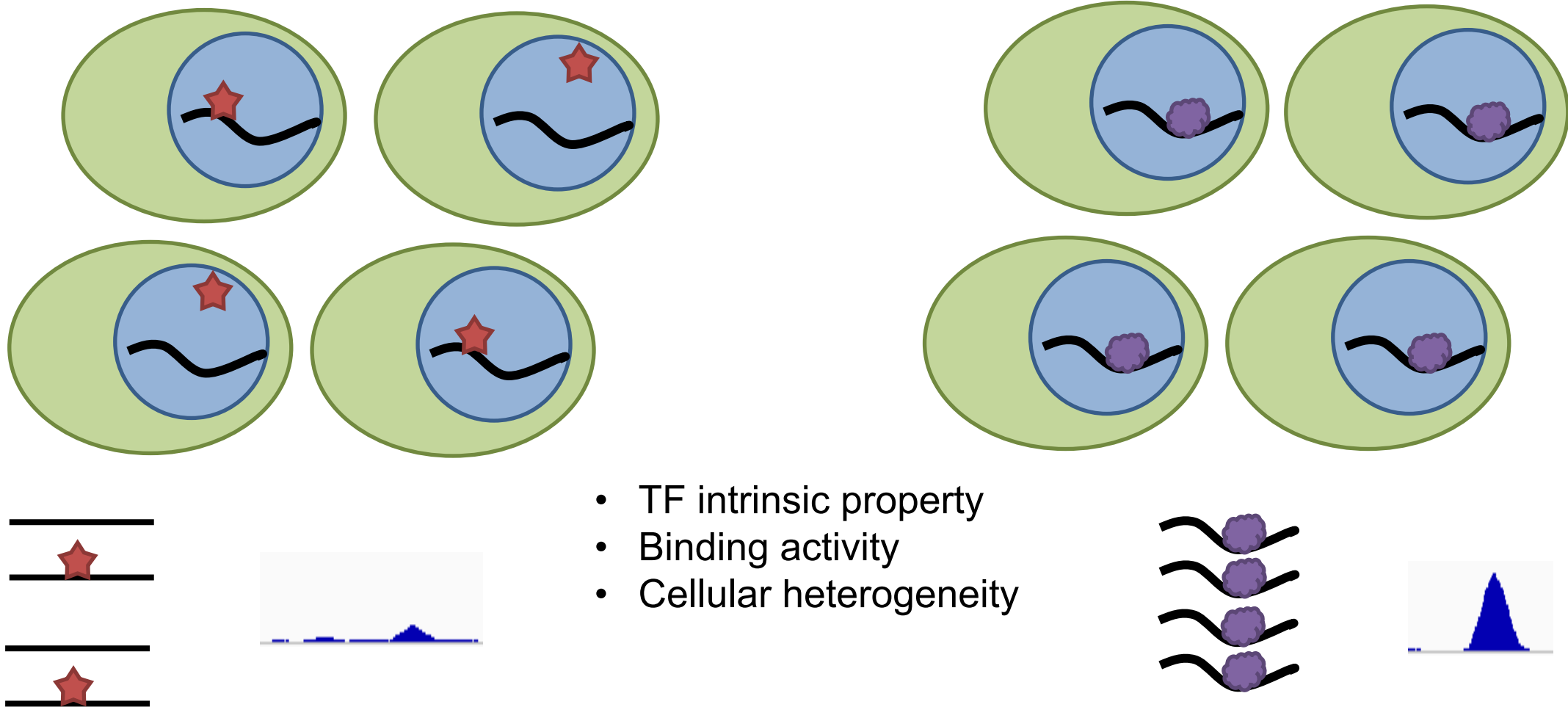
chr4	10344200	10344250	5
chr4	10344250	10344300	10
chr4	10344300	10344350	25
chr4	10344350	10344400	15
chr4	10344400	10344450	8

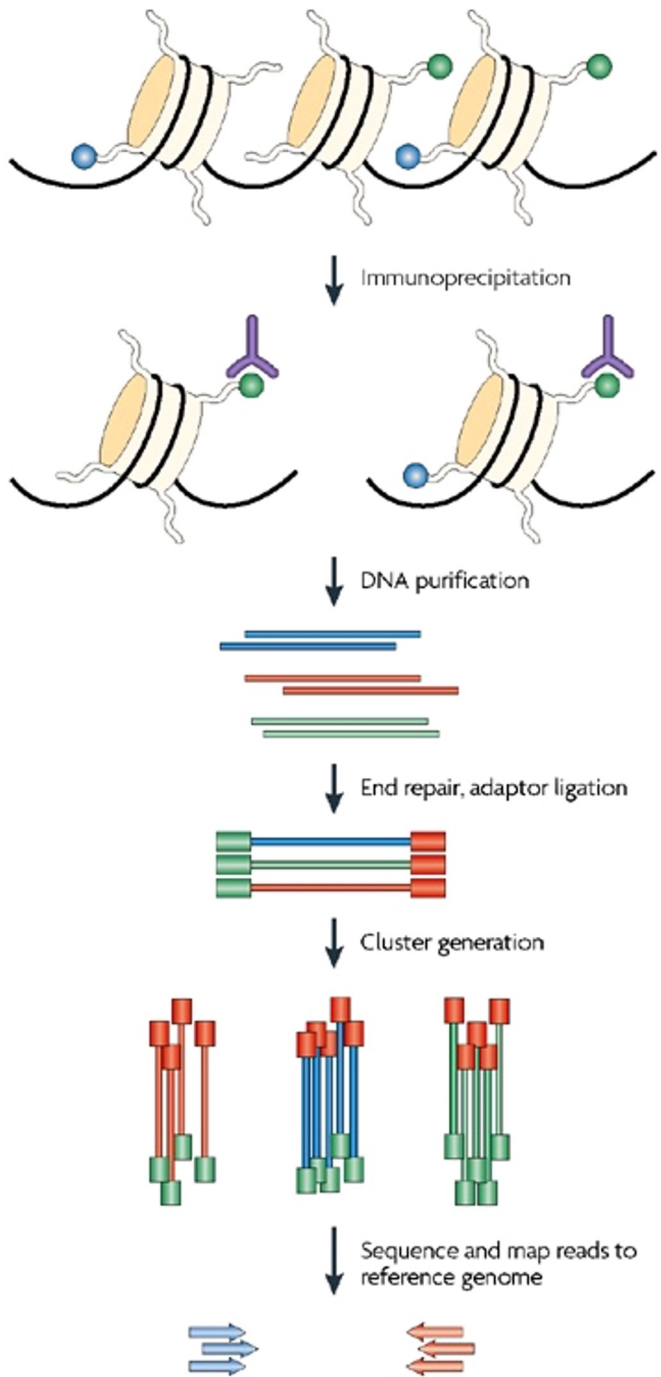
- wiggle:

```
track type=wiggle_0
variableStep chrom=chr4 span=50
10344200 5
10344250 10
10344300 25
10344350 15
10344400 8
```

- bigWig: indexed binary format

Biological interpretation: ChIP-seq captures a snapshot of binding patterns from a cell population



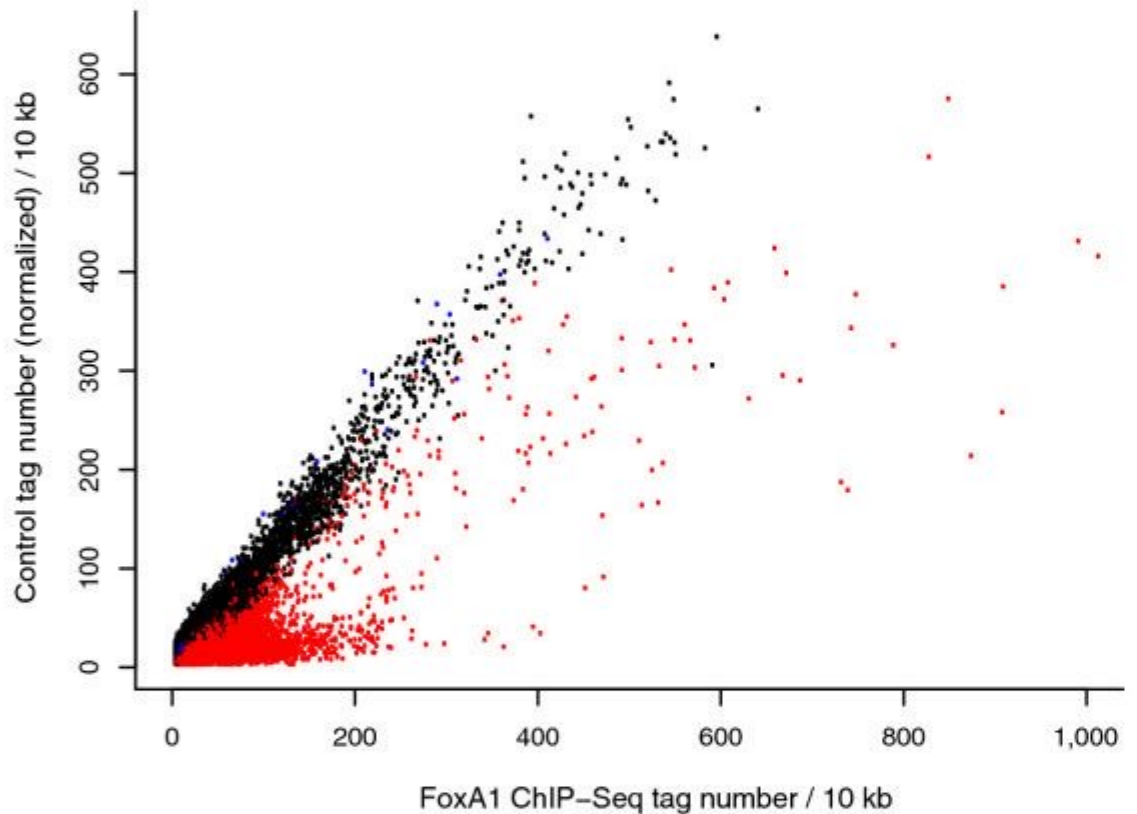
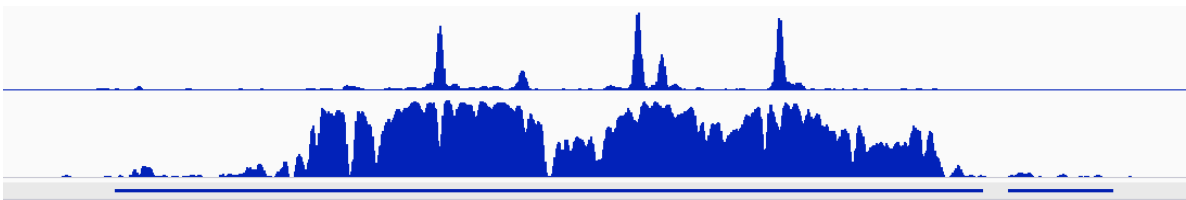


ChIP-seq: Study design

- Background Control: **Input** or **IgG**
 - Input chromatin: sonicated/digested chromatin without immunoprecipitation
 - IgG: “unspecific” immunoprecipitation
- Study Control:
 - Control condition: ChIP + input
 - Treated condition: ChIP + input

ChIP-seq: Signal detection (peak calling)

- Goal: Identify regions in the genome enriched for sequence reads:
 - Compared to genomic background
 - Compared to input control



ChIP-seq signal detection

Method

Open Access

Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang^{✉*}, Tao Liu^{✉*}, Clifford A Meyer^{*}, Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein^{§¶}, Chad Nusbaum[¶], Richard M Myers[‡], Myles Brown[†], Wei Li[#] and X Shirley Liu^{*}

Addresses: ^{*}Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, 44 Binney Street, Boston, MA 02115, USA. [†]Division of Molecular and Cellular Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute and Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA. [‡]Gene Security Network, Inc., 2686 Middlefield Road, Redwood City, CA 94063, USA. [§]Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital and Department of Pathology, Harvard Medical School, 13th Street, Charlestown, MA 02129, USA. [¶]Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA, 02142, USA. [‡]Department of Genetics, Stanford University Medical Center, Stanford, CA 94305, USA. [‡]Division of Biostatistics, Dan L Duncan Cancer Center, Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

✉ These authors contributed equally to this work.

Correspondence: Wei Li. Email: wl1@bcm.edu. X Shirley Liu. Email: xsliu@jimmy.harvard.edu

Published: 17 September 2008

Genome **Biology** 2008, **9**:R137 (doi:10.1186/gb-2008-9-9-r137)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2008/9/9/R137>

Received: 4 August 2008

Revised: 3 September 2008

Accepted: 17 September 2008

© 2008 Zhang *et al.*; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

We present Model-based Analysis of ChIP-Seq data, MACS, which analyzes data generated by short read sequencers such as Solexa's Genome Analyzer. MACS empirically models the shift size of ChIP-Seq tags, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome, allowing for more robust predictions. MACS compares favorably to existing ChIP-Seq peak-finding algorithms, and is freely available.

Background

The determination of the 'cistrome', the genome-wide set of *in vivo cis*-elements bound by *trans*-factors [1], is necessary

tional Sanger sequencing methods. Technologies such as Illumina's Solexa or Applied Biosystems' SOLiD™ have made ChIP-Seq a practical and potentially superior alternative to

BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 15 2009, pages 1952–1958
doi:10.1093/bioinformatics/btp340

Data and text mining

A clustering approach for identification of enriched domains from histone modification ChIP-Seq data

Chongzhi Zang¹, Dustin E. Schones², Chen Zeng¹, Kairong Cui², Keji Zhao² and Weiqun Peng^{1,*}

¹Department of Physics, The George Washington University, Washington, DC 20052 and ²Laboratory of Molecular Immunology, National Heart Lung and Blood Institute, NIH, Bethesda, MD 20892, USA

Received on March 3, 2009; revised on May 7, 2009; accepted on May 27, 2009

Advance Access publication June 8, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Chromatin states are the key to gene regulation and cell identity. Chromatin immunoprecipitation (ChIP) coupled with high-throughput sequencing (ChIP-Seq) is increasingly being used to map epigenetic states across genomes of diverse species. Chromatin modification profiles are frequently noisy and diffuse, spanning regions ranging from several nucleosomes to large domains of multiple genes. Much of the early work on the identification of ChIP-enriched regions for ChIP-Seq data has focused on identifying localized regions, such as transcription factor binding sites. Bioinformatic tools to identify diffuse domains of ChIP-enriched regions have been lacking.

Results: Based on the biological observation that histone modifications tend to cluster to form domains, we present a method that identifies spatial clusters of signals unlikely to appear by chance. This method pools together enrichment information from neighboring nucleosomes to increase sensitivity and specificity. By using genomic-scale analysis, as well as the examination of loci with validated epigenetic states, we demonstrate that this method outperforms existing methods in the identification of ChIP-enriched signals for histone modification profiles. We demonstrate the application of this unbiased method in important issues in ChIP-Seq data analysis, such as data normalization for quantitative comparison of levels of epigenetic modifications across cell types and growth conditions.

Availability: <http://home.gwu.edu/~wpeng/Software.htm>

Contact: wpeng@gwu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Covalent modifications of chromatin, including DNA methylation and histone modifications, play critical roles in gene regulation

high-throughput massively parallel sequencing technologies (Barski *et al.*, 2007; Mikkelsen *et al.*, 2007). ChIP-Seq combines chromatin immunoprecipitation (ChIP) with high-throughput sequencing to map genome-wide chromatin modification profiles and transcription factor (TF) binding sites. It is characterized by high resolution, a quantitative nature, cost effectiveness and no complication due to probe hybridization as encountered in ChIP-chip assays (Schones and Zhao, 2008). A large amount of data has recently been generated using the ChIP-Seq technique, and these datasets call for new analysis algorithms.

Binding of TFs is mainly governed by their sequence specificity and therefore is typically associated with very localized ChIP-Seq signals in the genome. A number of algorithms have been developed to find the exact locations of TF binding sites from ChIP-Seq data (Chen *et al.*, 2008; Fejes *et al.*, 2008; Ji *et al.*, 2008; Johnson *et al.*, 2007; Jothi *et al.*, 2008; Kharchenko *et al.*, 2008; Nix *et al.*, 2008; Rozowsky *et al.*, 2009; Valouev *et al.*, 2008; Zhang *et al.*, 2008a). In contrast, the signals for histone modifications, histone variants and histone-modifying enzymes are usually diffuse and lack of well-defined peaks, spanning from several nucleosomes to large domains encompassing multiple genes (Barski *et al.*, 2007; Pauler *et al.*, 2009; Wang *et al.*, 2008; Wen *et al.*, 2009) (see, e.g. Figure S1). The detection of diffuse signals often suffers from high noise level and lack of saturation in sequencing coverage. These generally weak signals render approaches seeking strong local enrichment, such as those peak-finding algorithms used in finding TF binding sites, inadequate.

Many modification marks are known to form broad domains (Barski *et al.*, 2007; Wang *et al.*, 2008). This is believed to be helpful in stabilizing the chromatin state and propagating such states through cell division robustly (Bernstein *et al.*, 2007). A well-studied case is the trimethylation of histone H3 lysine 9 (H3K9me3). H3K9me3 recruits HP1 via its chromodomain. HP1 in turn recruits H3K9 methyltransferase Suv39h, which modifies

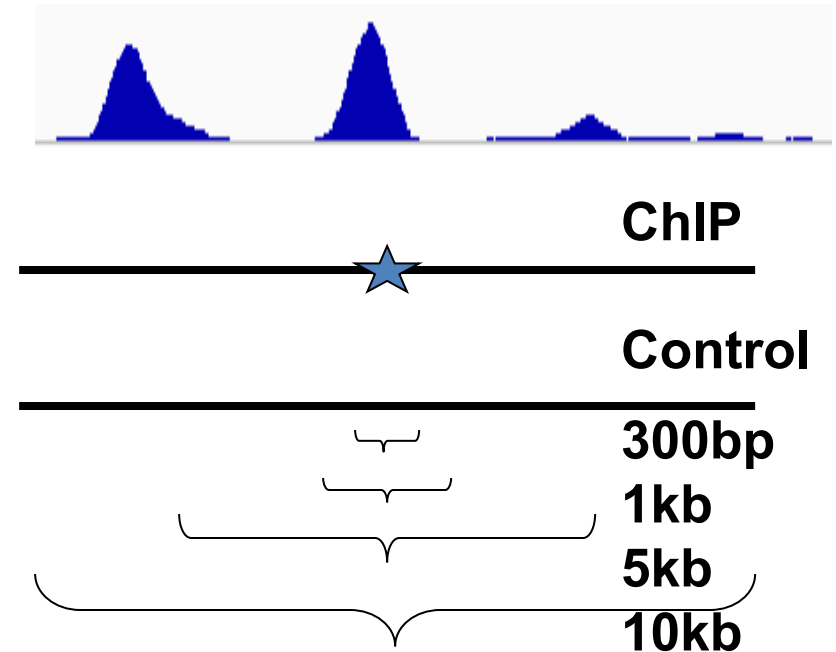
MACS: model

- **Model-based Analysis for ChIP-Seq**
- Read distribution along the genome ~ Poisson distribution
(λ_{BG} = total tag / genome size)
Negative binomial distribution (MACS2)

- ChIP-seq show local biases in the genome
 - Chromatin and sequencing bias
 - 200-300bp control windows have too few tags
 - But can look further

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

- B-H adjustment to correct for FDR
 - p-value \rightarrow q-value



Zhang et al, *Genome Bio*, 2008

- Noisy
- Hard to see “peaks”
- Enriched regions are spread out
- Lack saturation

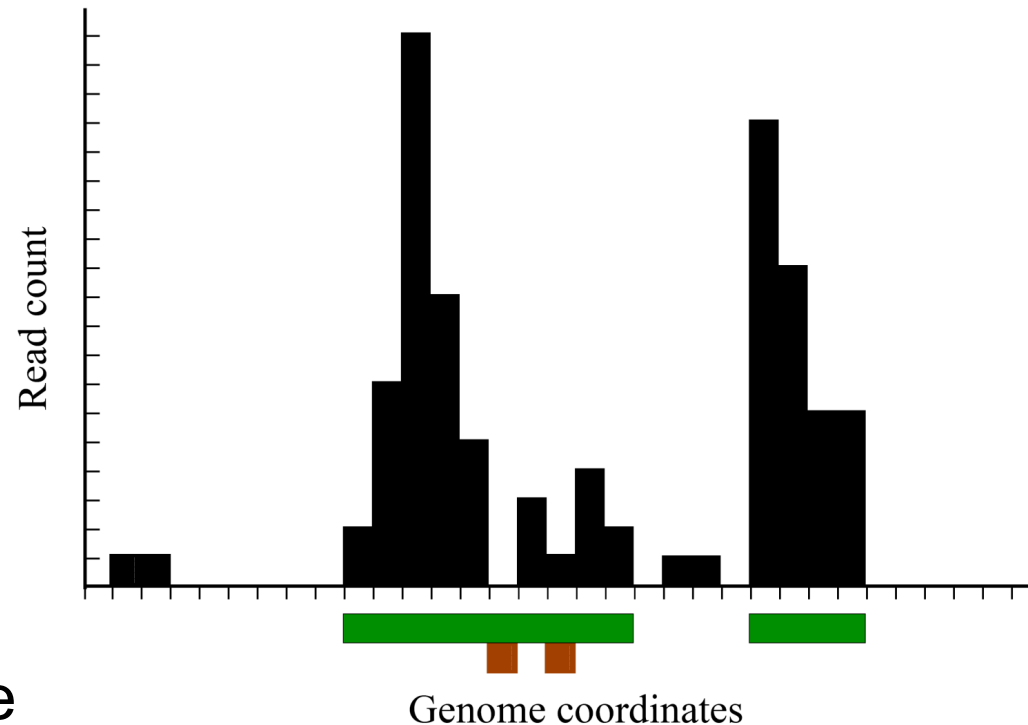


SICER: Definition of Island

- Eligible and ineligible windows

$$\sum_{l=l_0}^{\infty} P(l, \lambda) \leq p_0$$

- Eligible windows are separated by **gaps** of ineligible windows.
- Island**: cluster of eligible windows separated by gaps of size at most g windows.



Example islands for
 $l_0 = 2$ and $g = 2$

SICER: Scoring islands

- The scoring function is based on the probability of finding the observed tag count in a random background.
- For a window with ***m*** reads,
 - The probability of finding *m* reads is Poisson $P(m, \lambda)$
 - $\lambda = wN/L$ is the average number of reads in each window
- Scoring function for an eligible window:

$$S = -\ln P(m, \lambda)$$

- Key quantity: the score of an island
 - Aggregate score of all eligible windows in the island
 - It corresponds to the background probability of finding the observed pattern

SICER: Island score statistics

- Probability distribution of scores for a single window in a random background model:

$$\rho(s) = \sum_{l \geq l_0} \delta(s - s(l)) P(l, \lambda)$$

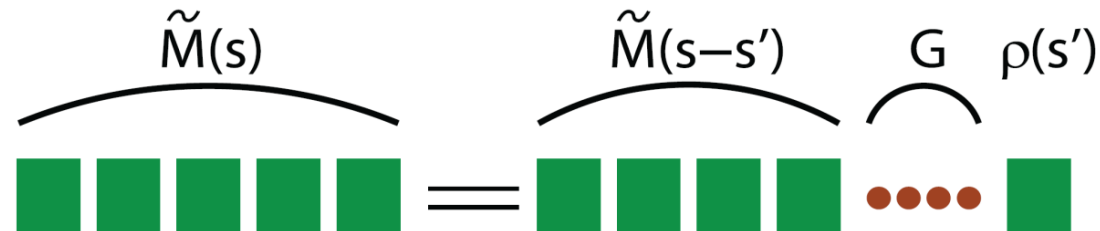
- Probability of a window being 'ineligible':

$$t = P(0, \lambda) + P(1, \lambda) + \dots + P(l_0 - 1, \lambda)$$

- Gap factor:

$$G = 1 + t + t^2 + \dots + t^g$$

SICER: Island score statistics



- Recursion relation

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

- Probability of finding an island of score \mathbf{s} :

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

SICER: Island score statistics

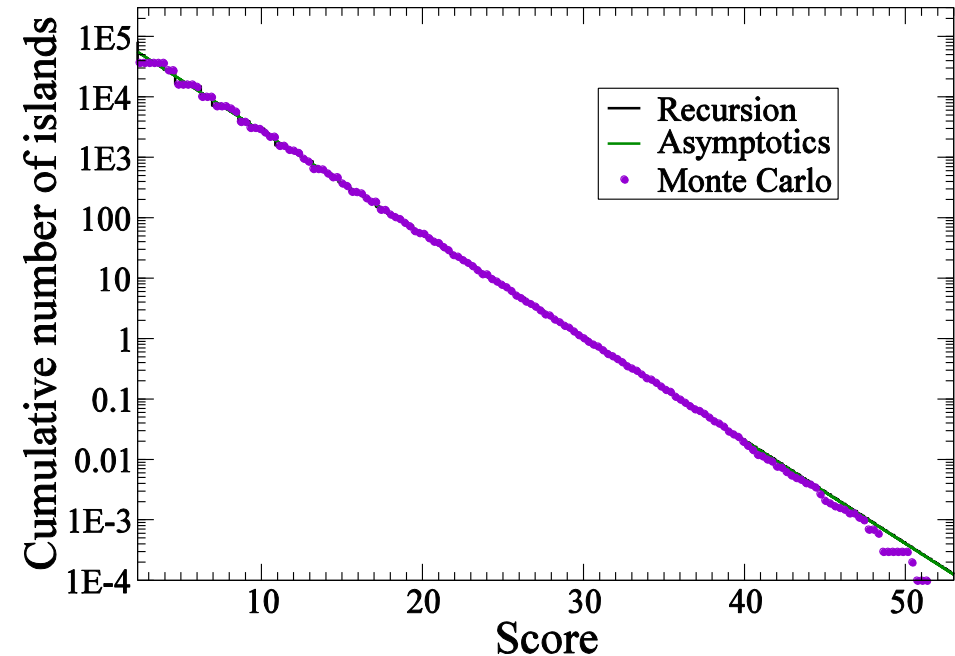
- Asymptotics of island score distribution in the background

$$\tilde{M}(s) = \alpha \exp(-\beta s)$$

$$G(\lambda, l_0, g) \sum_{l \geq l_0} P(l, \lambda)^{1-\beta} = 1$$

- Statistic: E -value
 - Expected number of islands with score above s_T in the background

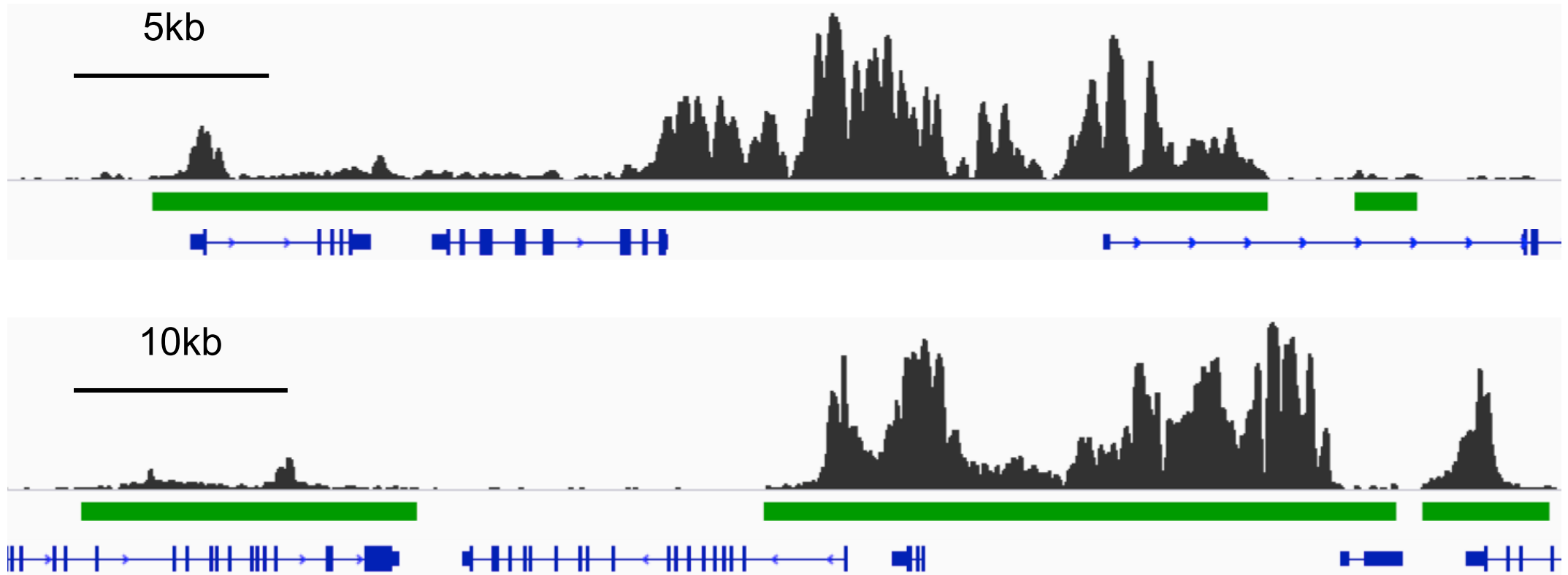
$$\sum_{s \geq s_T} LM(s) \leq e$$





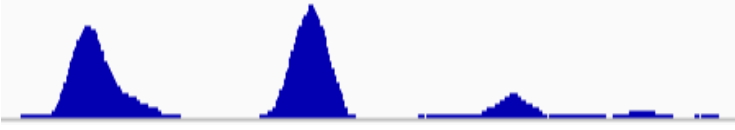
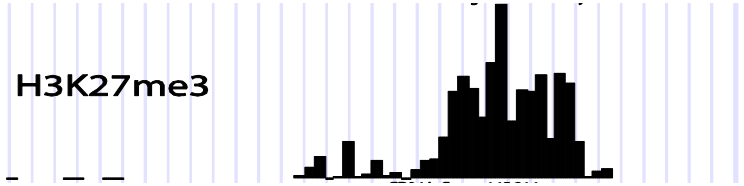
SICER: Significance determinations

- Significance determination with random background model:
 - E -value determines an island score threshold
- Significance determination with control sample
 - Identify candidate islands using random background
 - For each candidate island, compare sample with control
 - P -value $\sum_{n=n_s}^{\infty} P(n_s, cn_c)$
 - False Discovery Rate (FDR)

SICER result examples



Data preprocessing

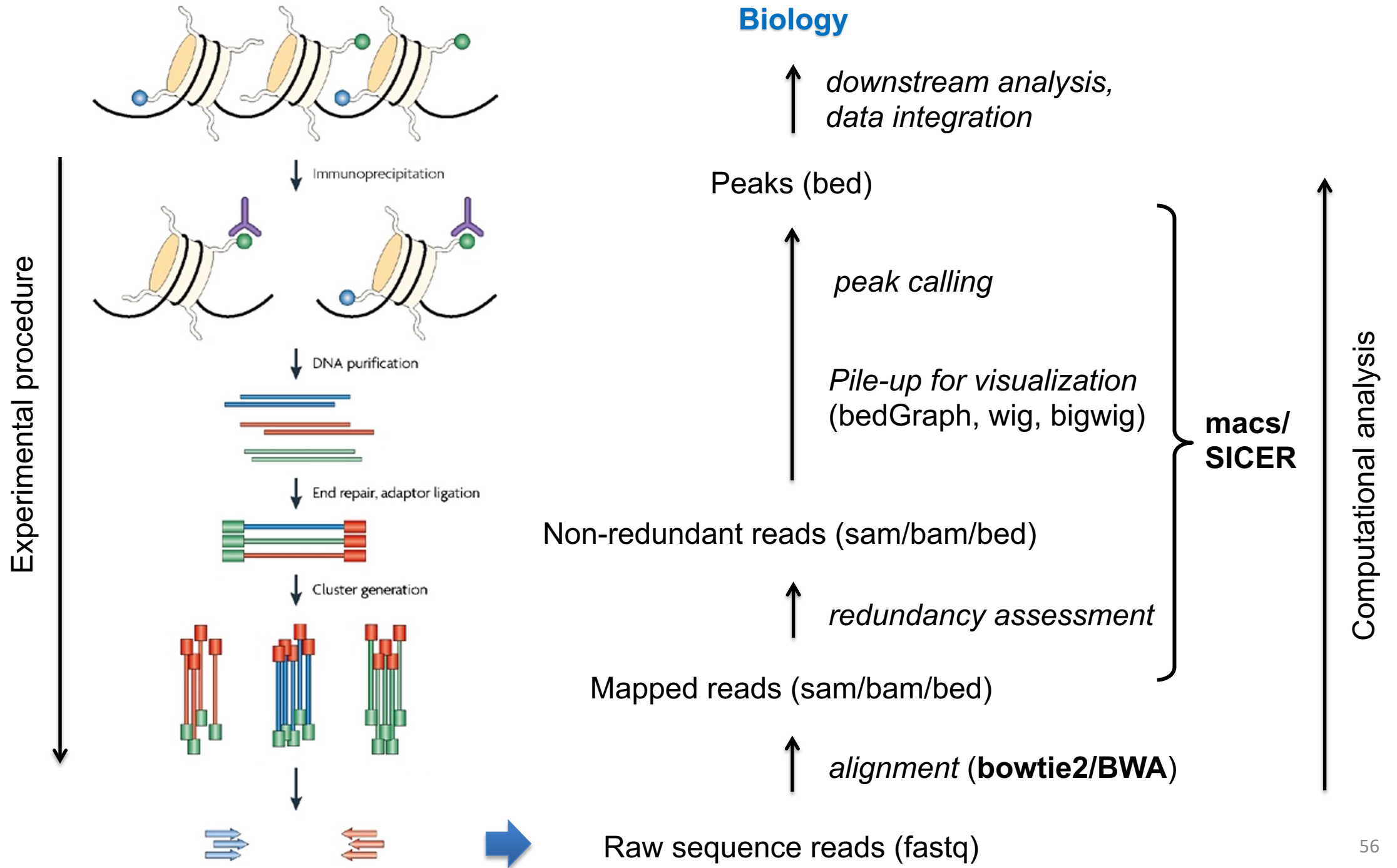
	MACS	SICER
DNA fragment size estimation	Peak model	Cross-correlation
DNA fragment retrieval	Full length (extend d) 	Point position (shift $d/2$) 
Signal profile generation	Fragment pile up	Read count in bins
		

Signal detection

	MACS	SICER
Initial model	Poisson	Poisson
Signal scan	Sliding windows with bandwidth	Non-overlapping bin read count
Peak region identification	Merge windows	Merge windows allowing gaps
Peak scoring	Pile-up signal amount	Aggregate score on islands
Significance modeling	Poisson with dynamic λ	Asymptotic estimation of island score statistics model, then compare with control
Additional information	Read count, Pile-up height, Summit position	Read count, peak score, E-value

Transcription Factors vs. Histone Marks

	DNA-binding proteins (Transcription factors)	Histone Marks (Histone modifications, histone variants, chromatin regulators*)
Cell type specificity	Both factor and profile	Profile
Signal width (“peak size”)	Narrow	Narrow or broad
Chromatin accessibility	High	High or low
DNA sequence motif	Yes	No
Resolution	Up to 1-10bp	Nucleosome (200bp)



HOW TO: DRAW A HORSE

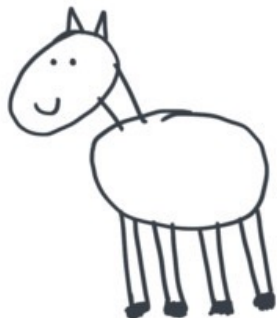
BY VAN OKTOP



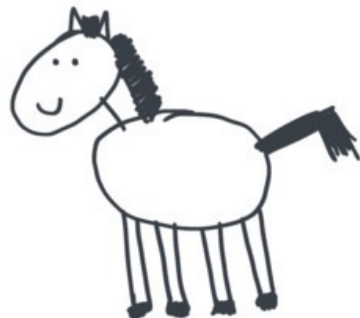
① DRAW 2 CIRCLES



② DRAW THE LEGS



③ DRAW THE FACE



④ DRAW THE HAIR

⑤
ADD
SMALL
DETAILS.

