UNIT 2 EPIGENOMICS

ChIP-seq

March 29, 2022

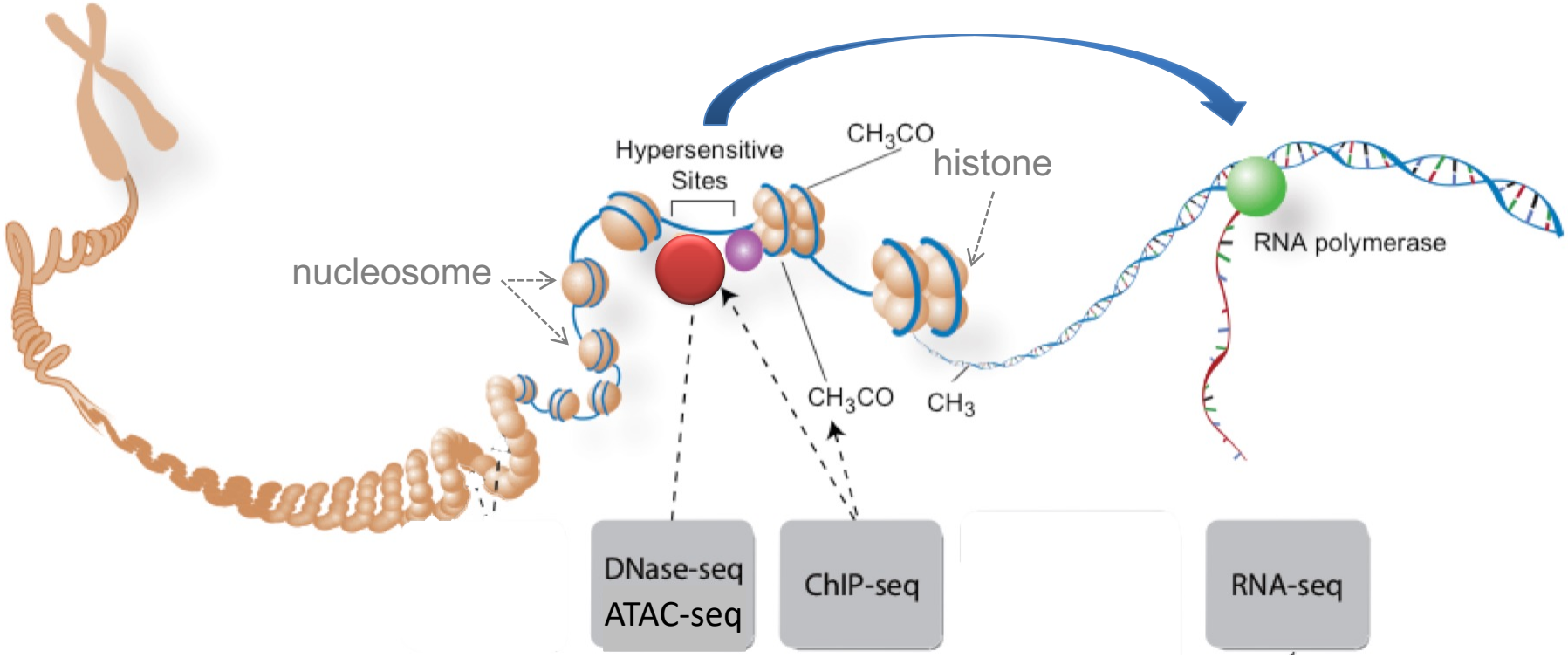
Outline

- ChIP-seq technology and development
- ChIP-seq data analysis
  - Strategy
  - Peak calling (MACS)
  - Peak calling (SICER)

Outline

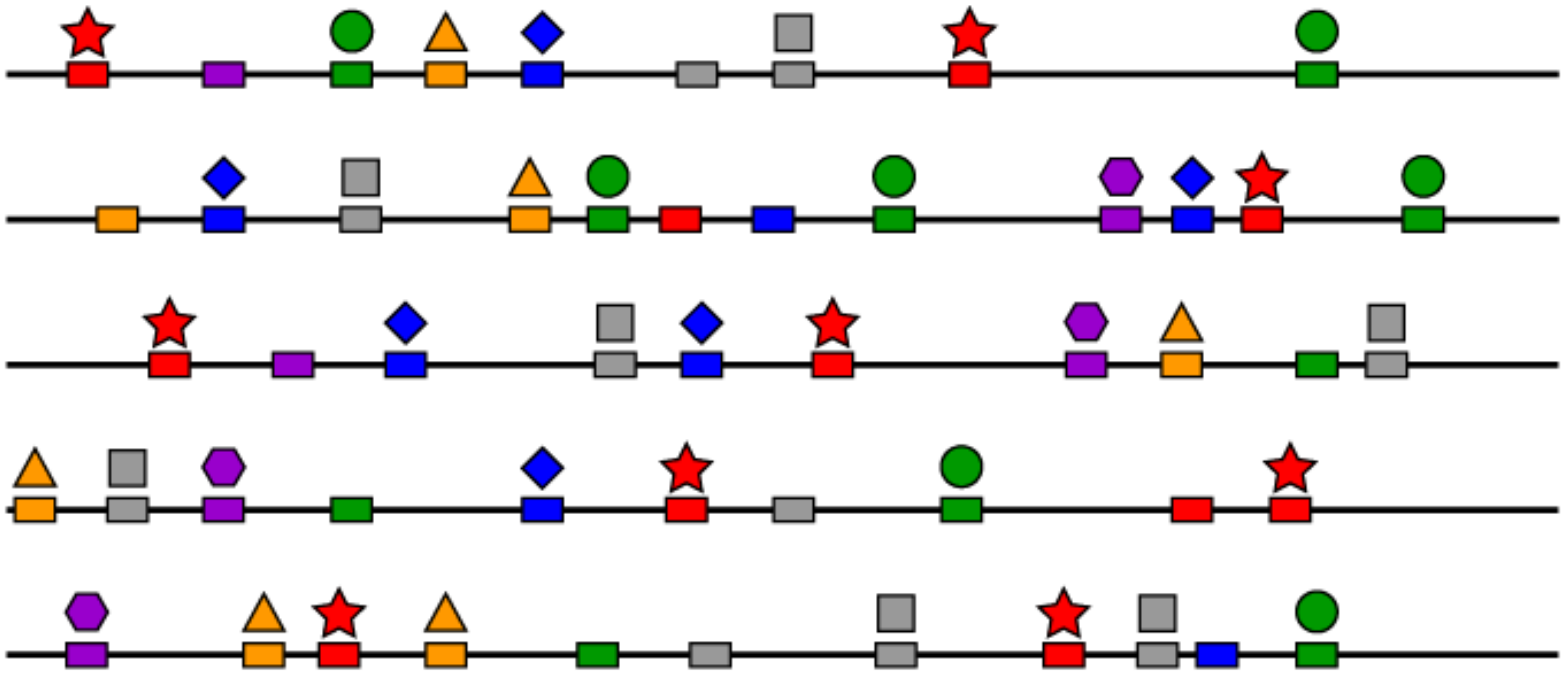
- **ChIP-seq technology and development**
- ChIP-seq data analysis
  - Strategy
  - Peak calling (MACS)
  - Peak calling (SICER)

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

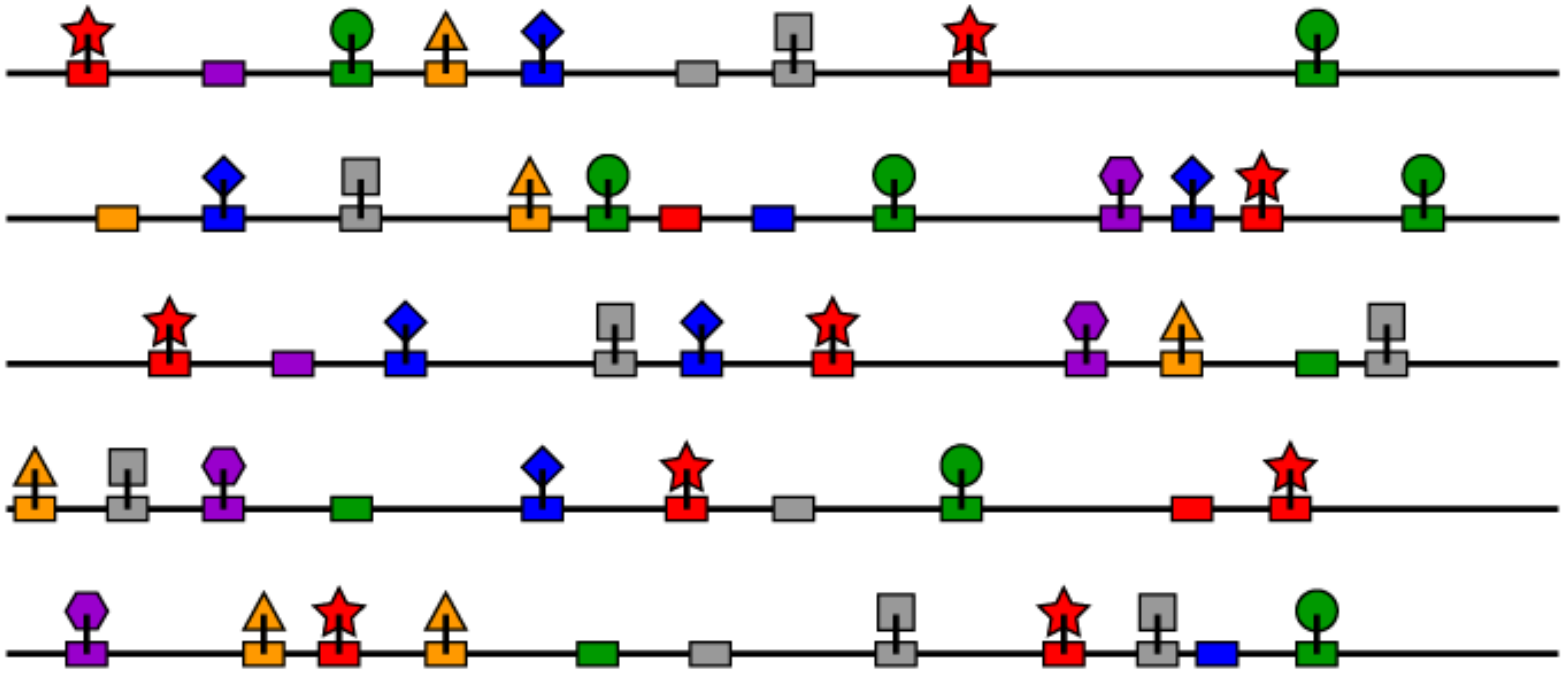


Original figure from ENCODE, Darryl Leja (NHGRI), Ian Dunham (EBI)

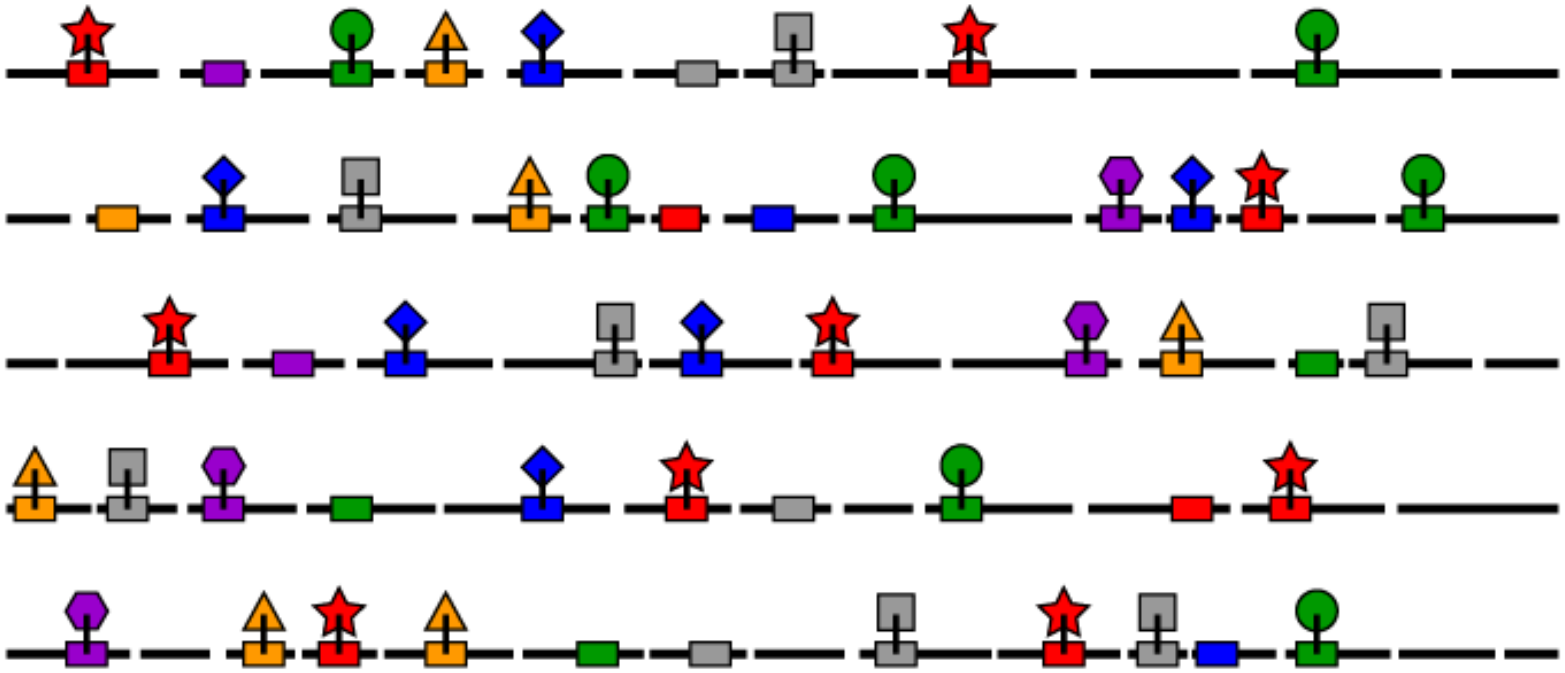
Chromatin ImmunoPrecipitation (ChIP)



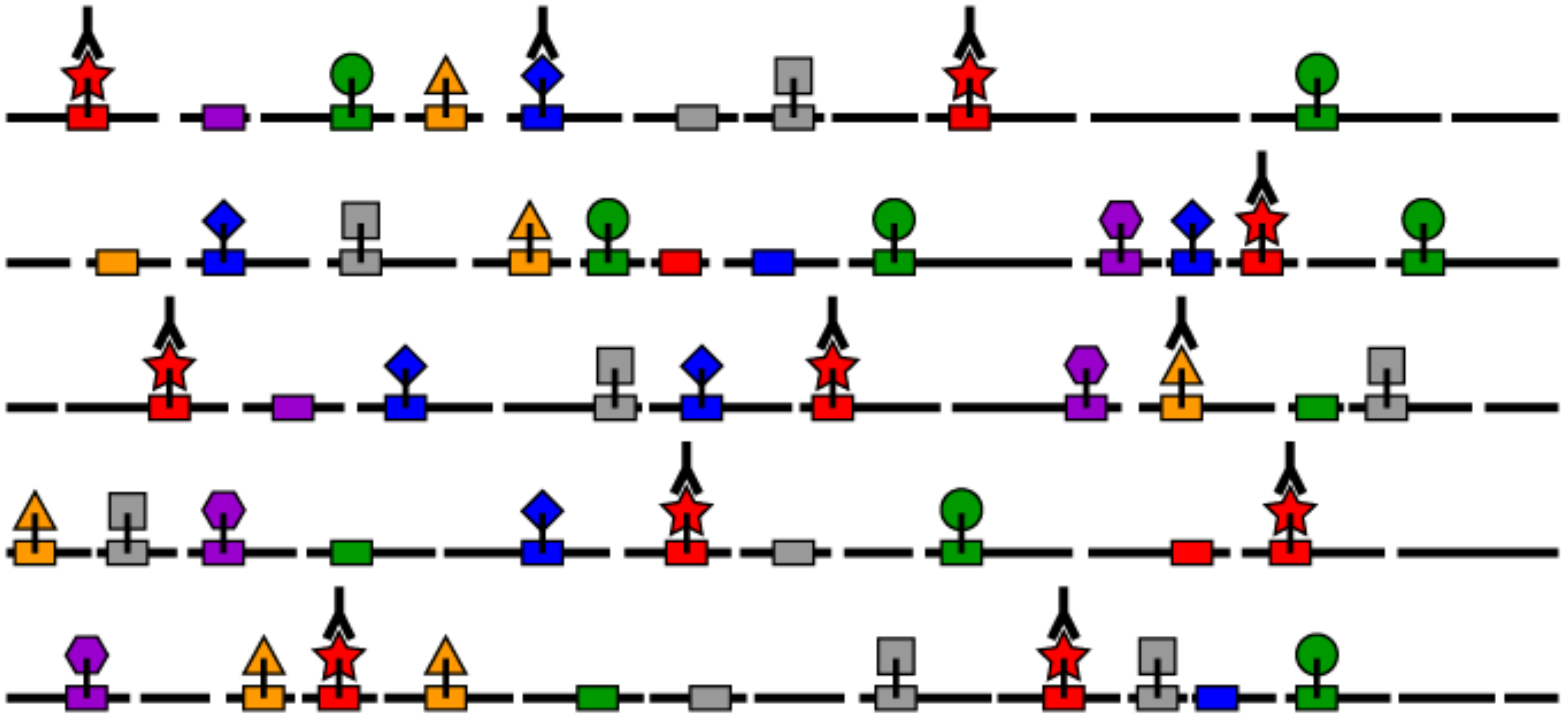
Protein-DNA crosslinking with formaldehyde (for TF)



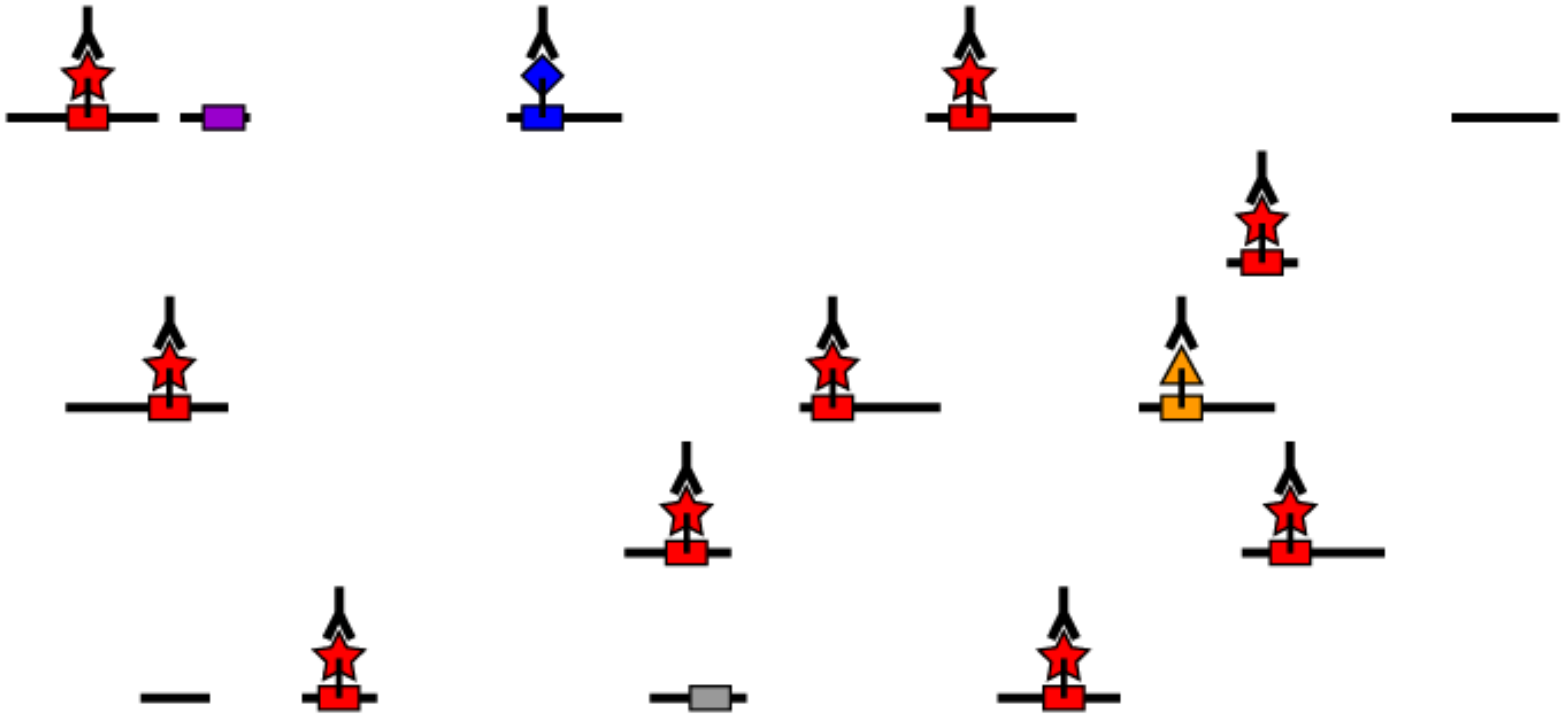
Chop the chromatin using sonication (TF) or micrococal nuclease (MNase) digestion (histone)



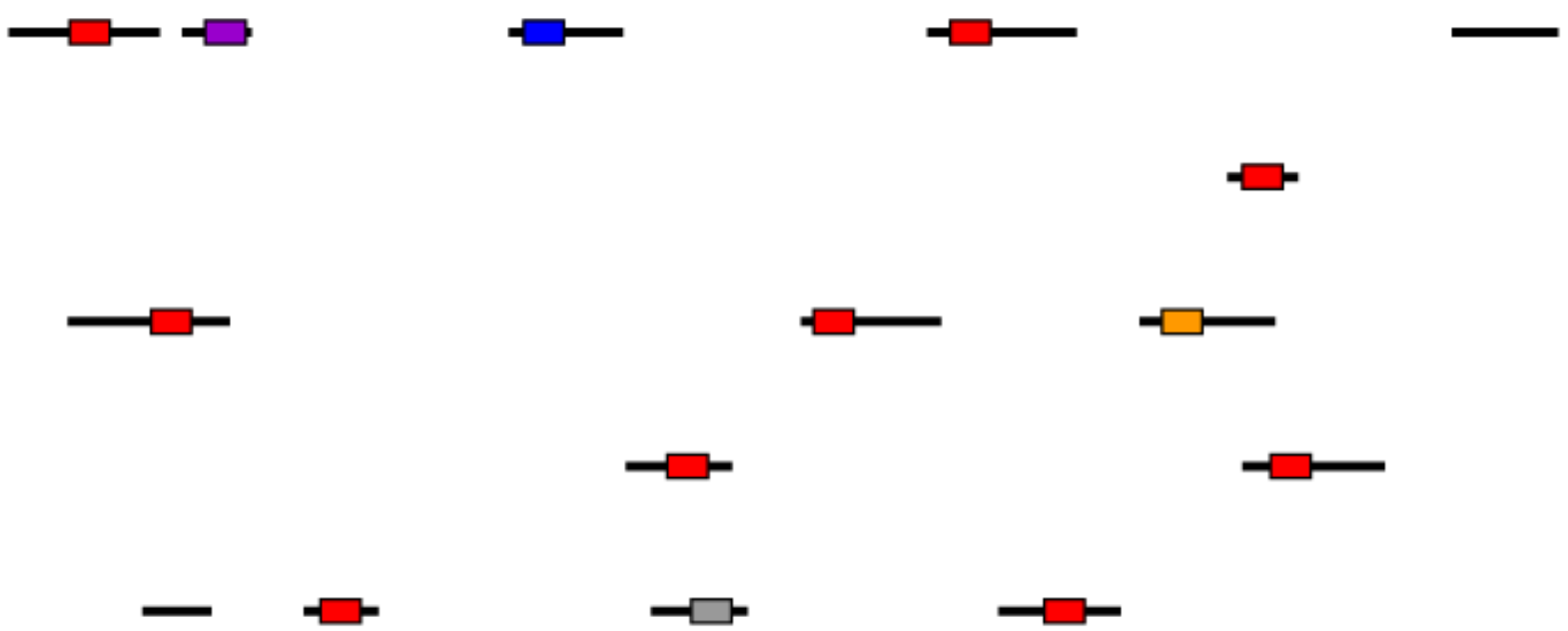
Specific factor-targeting antibody



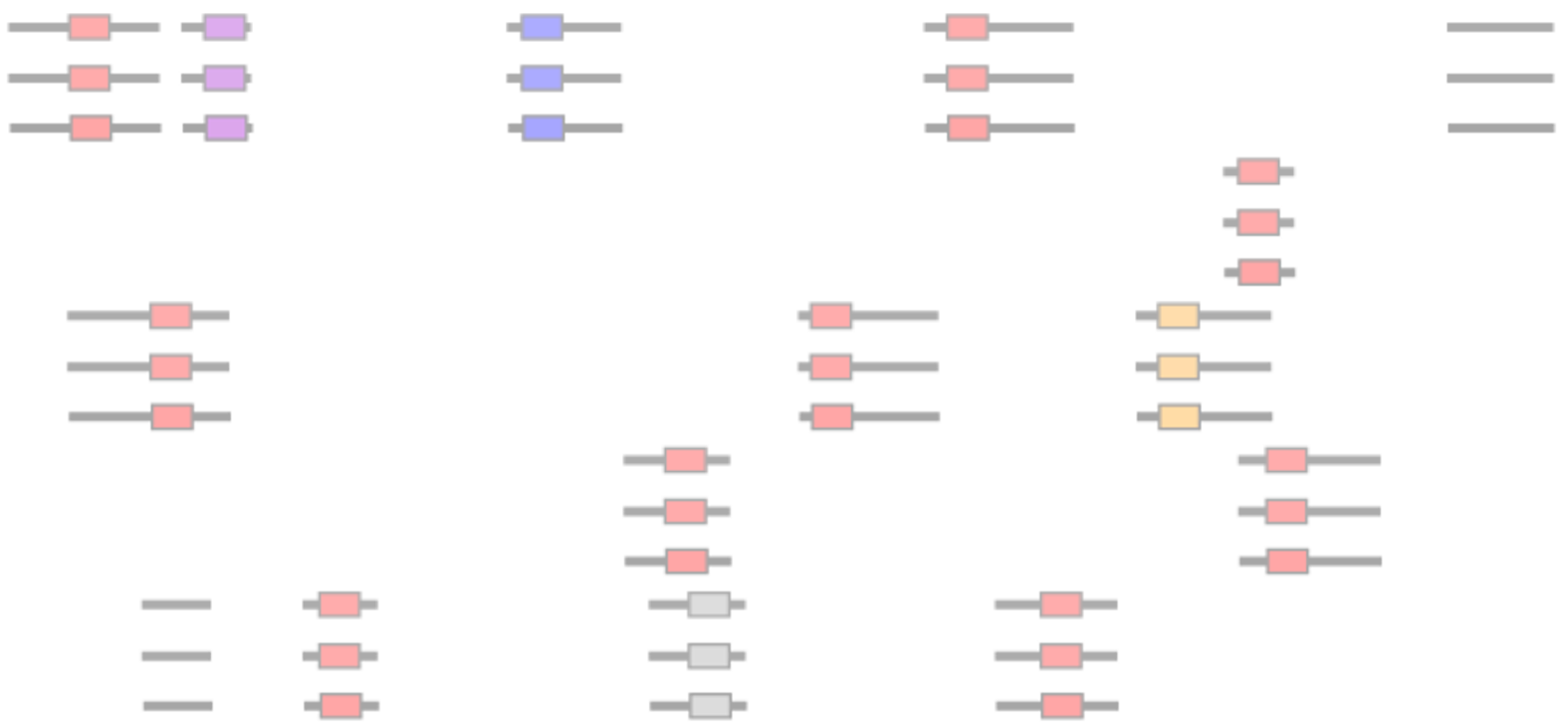
Immunoprecipitation



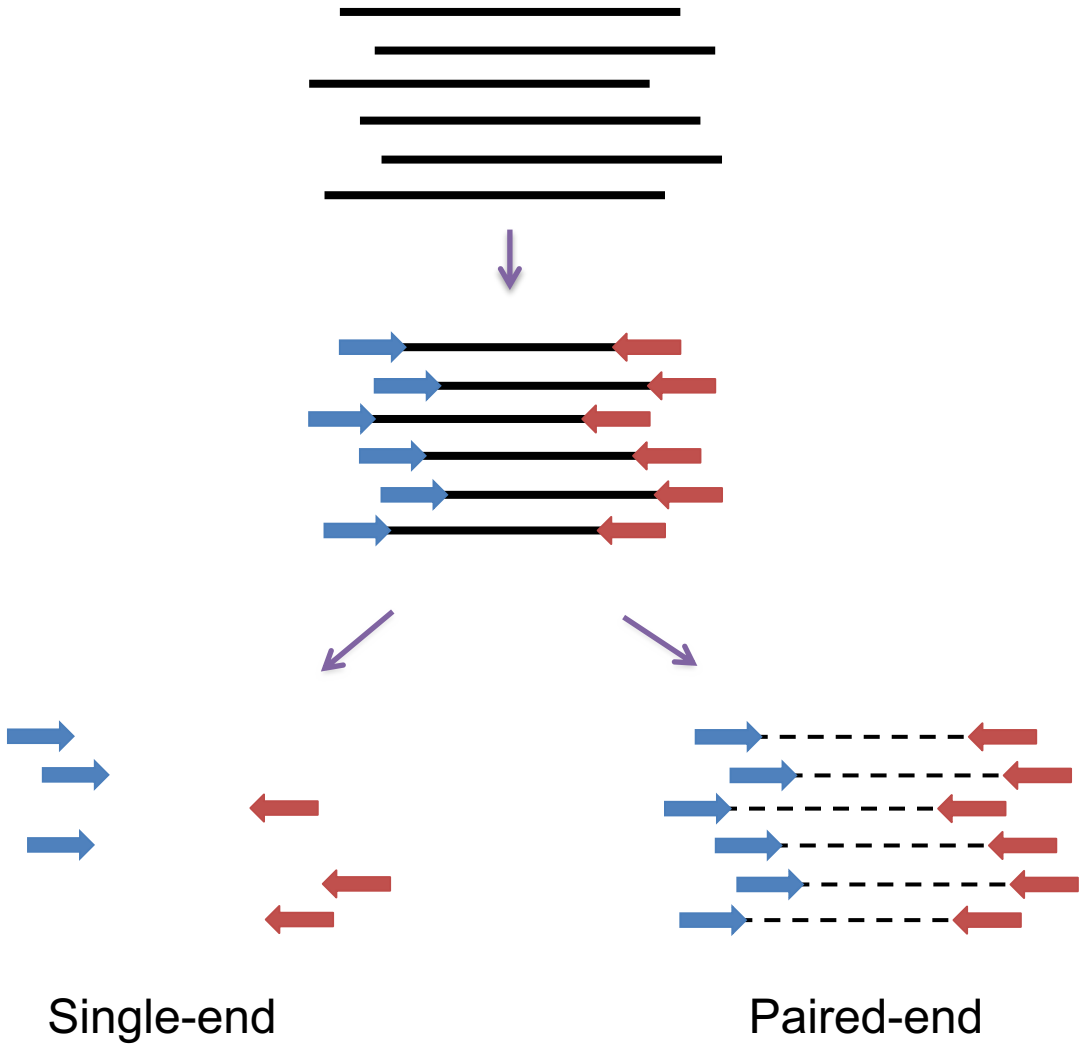
DNA purification

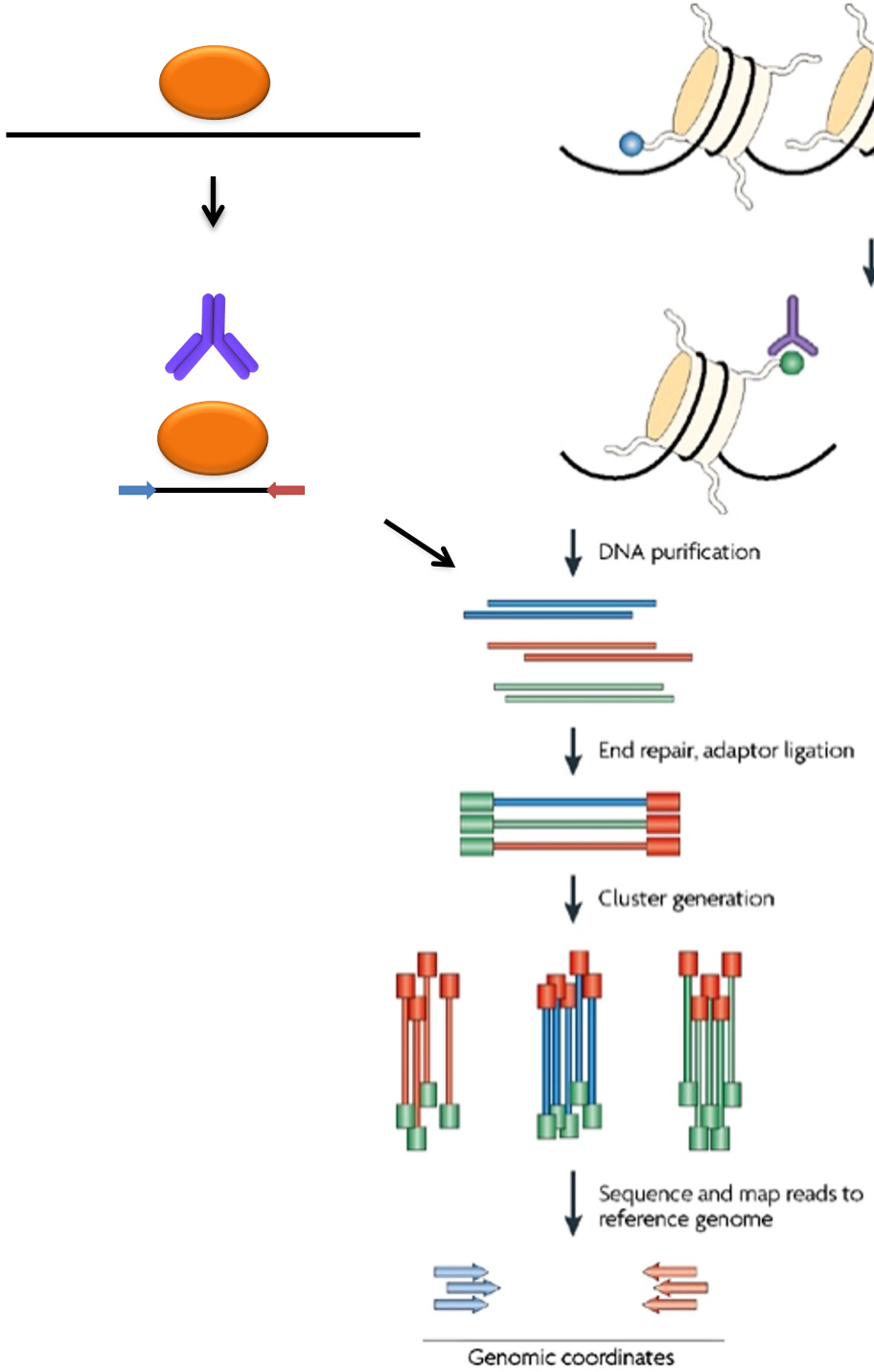


PCR amplification



High-throughput sequencing (Illumina)



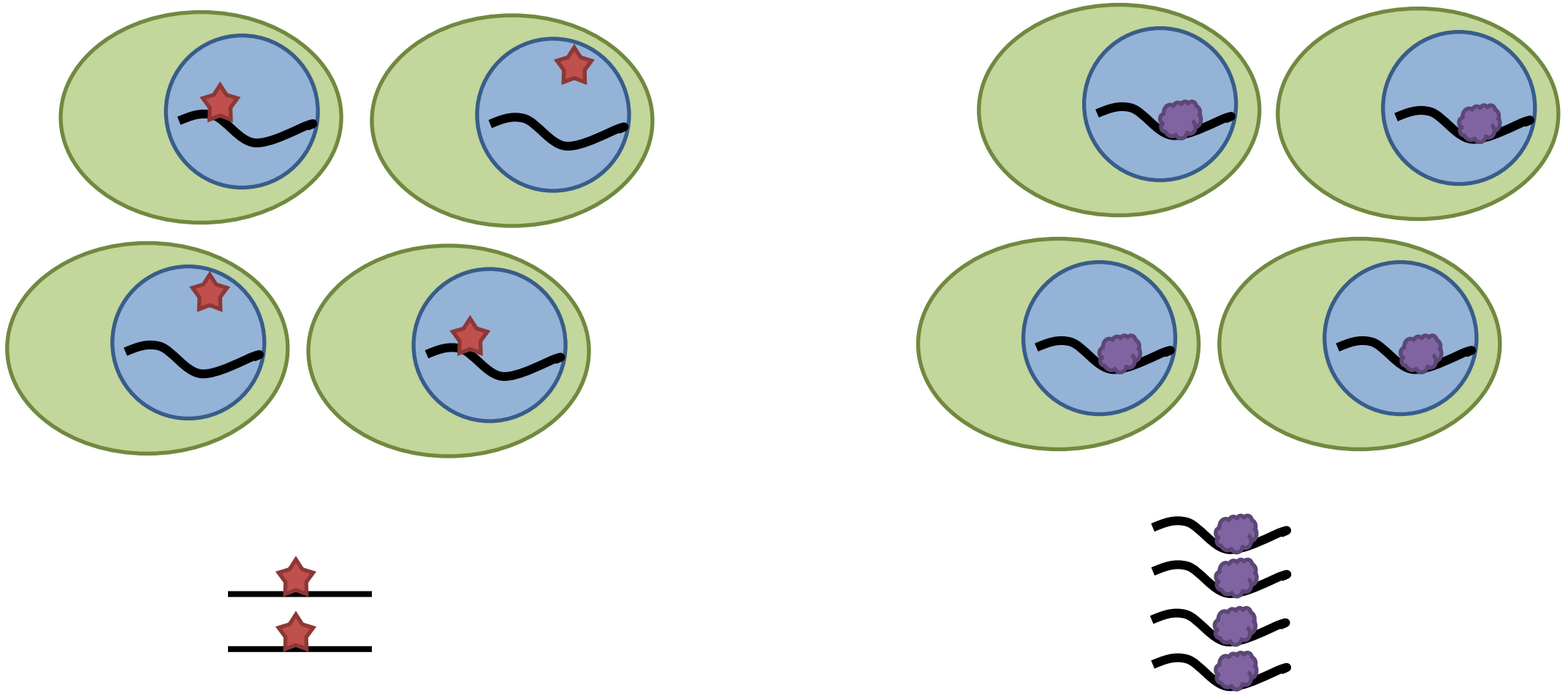


ChIP

Seq

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

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



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

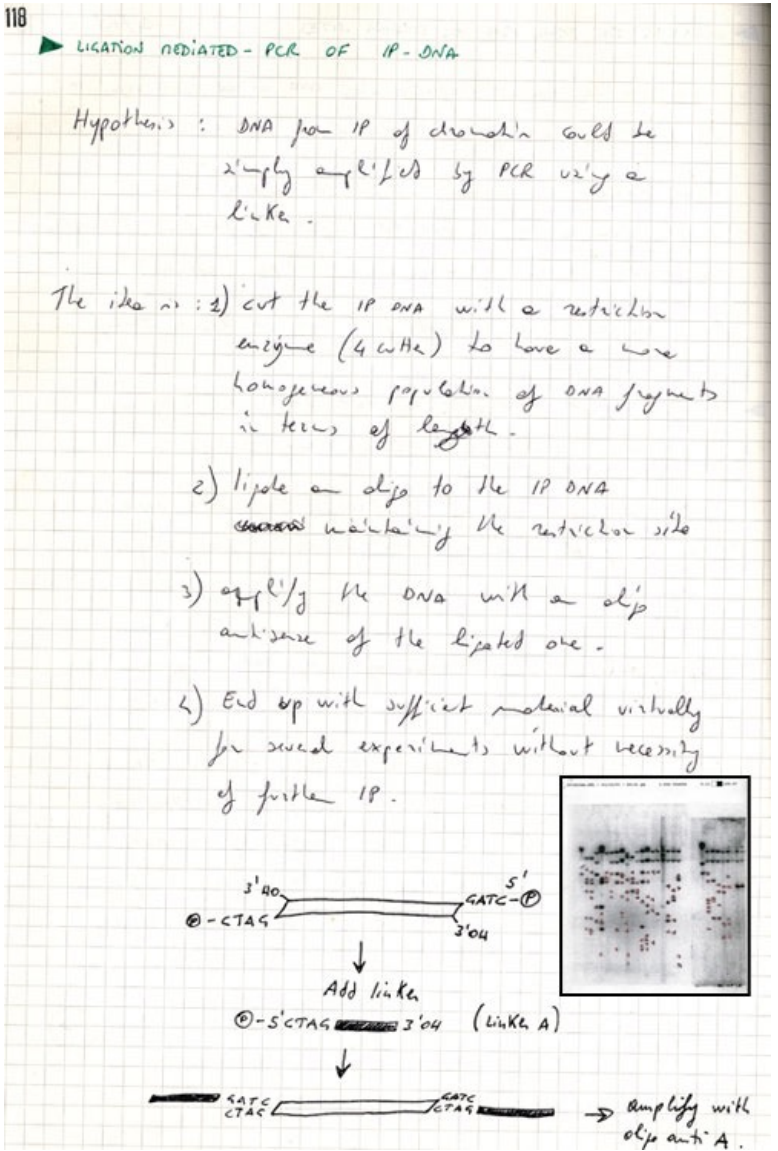
Summary

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



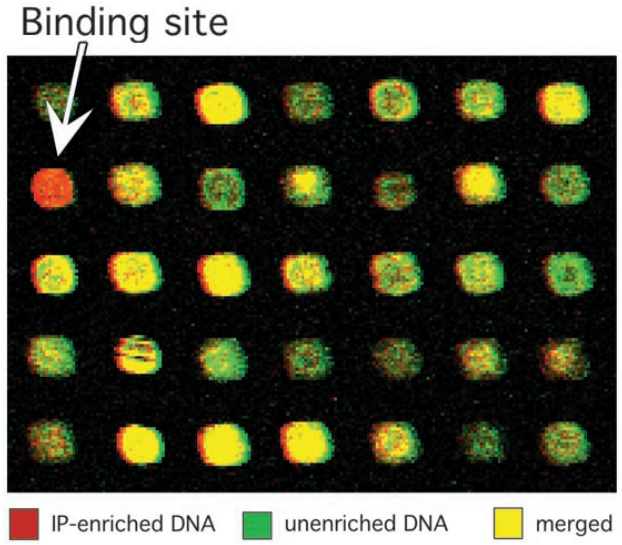
ChIP-chip (2000)

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 Tammana,[1] Gregg Helt,[1] Kevin Struhl,[2,*] and Thomas R. Gingeras[1,*]

[1]Affymetrix
3380 Central Expressway
Santa Clara, California 95051

[2]Deptartment of Biological Chemistry and Molecular Pharmacology
Harvard Medical School
Boston, Massachusetts 02115

[3]Department of Biological Sciences
National University of Singapore
Singapore 117543

[4]Genome 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 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.)

20

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 counter-phase 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?

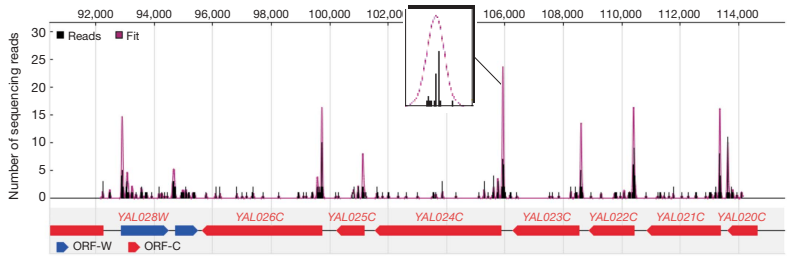


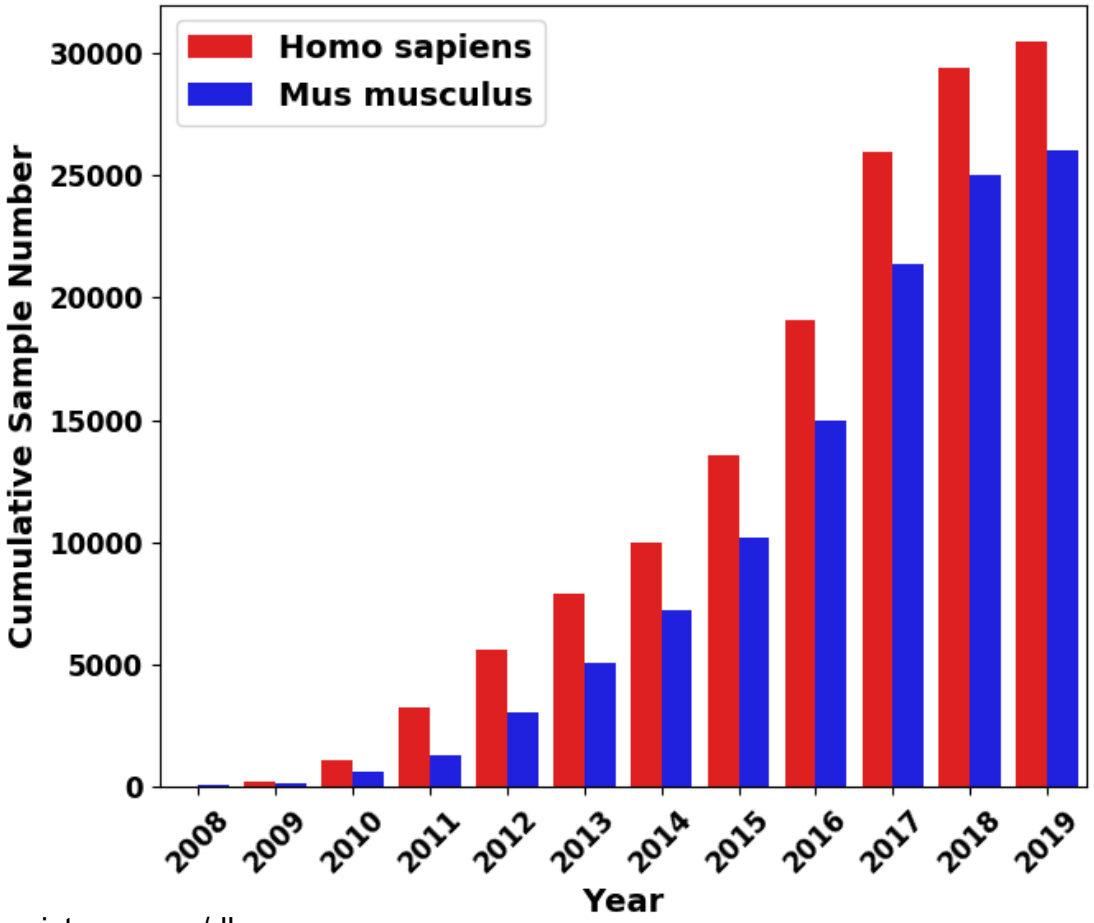
Figure 1 | Distribution of H2A.Z nucleosomal DNA at an arbitrary region of the yeast genome. Any region of the genome can be viewed in this way at http://nucleosomes.sysbio.bx.psu.edu. An enlarged view of a peak is shown in the inset, where each vertical bar corresponds to the number of sequencing reads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

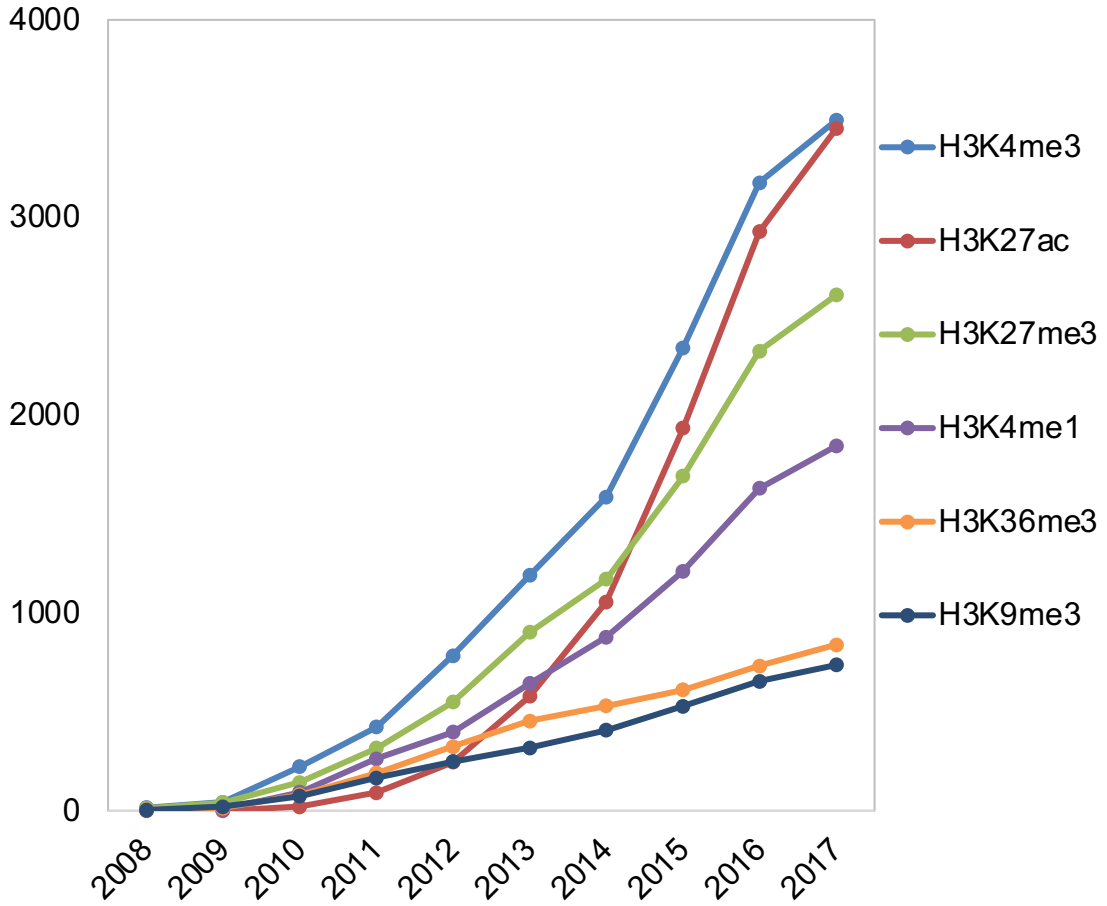
First ChIP-seq papers

Title	First/last authors	Journal	First submission date	Acceptance date	Publication date	Species/cell type	ChIP factors	# citations (3/28/22)
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	866
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	7096
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)	3100
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	4430

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



cistrome.org/db



Outline

- ChIP-seq technology and development
- **ChIP-seq data analysis**
  - Strategy
  - Peak calling (MACS)
  - Peak calling (SICER)

ChIP-seq data analysis

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

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 Zang[1], Dustin E. Schones[2], Chen Zeng[1], Kairong Cui[2], Keji Zhao[2] and Weiqun Peng[1,*]

[1]Department of Physics, The George Washington University, Washington, DC 20052 and [2]Laboratory 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

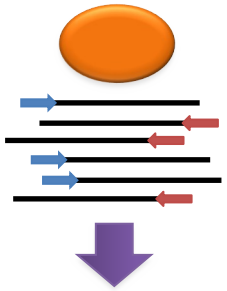
Covalent modifications of chromatin, including DNA methylation and histone modifications, play critical roles in gene regulation

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.

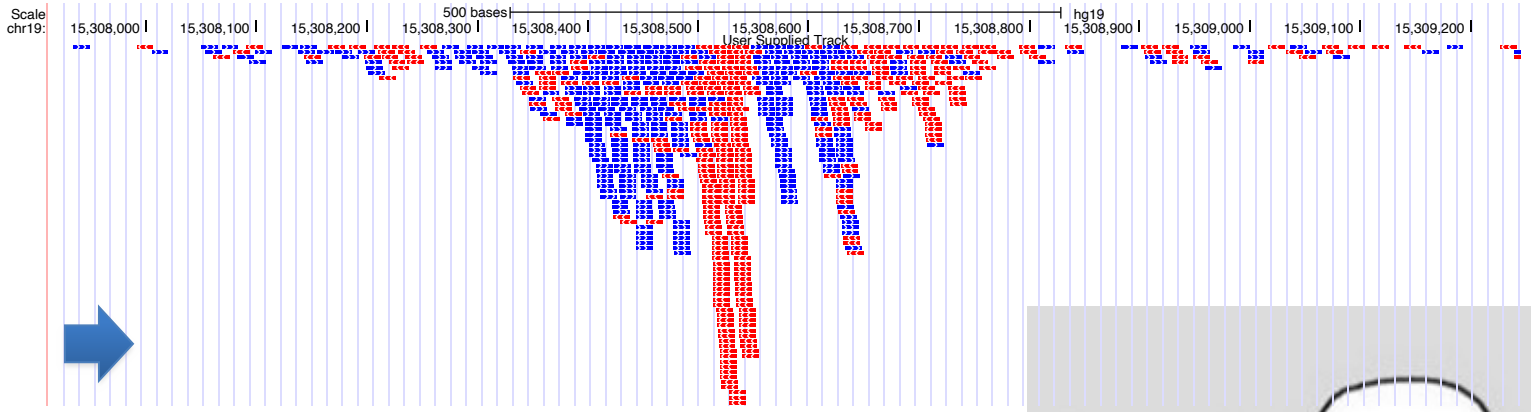
Binding of TFs is mainly governed by their sequence specificity and therefore is typically associated with very localized ChIP-Seq signals in the genome. A number of algorithms have been developed to find the exact locations of TF binding sites from ChIP-Seq data (Chen *et al.*, 2008; Fejes *et al.*, 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.

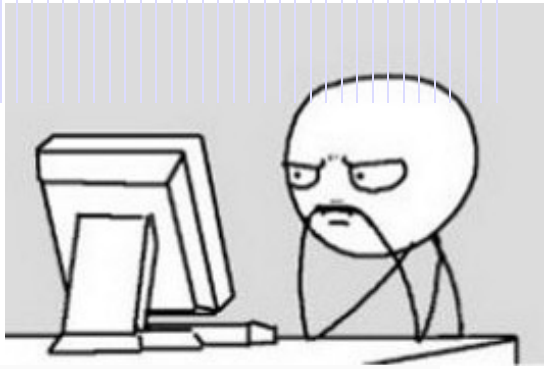
Many modification marks are known to form broad domains (Barski *et al.*, 2007; Wang *et al.*, 2008). This is believed to be helpful in stabilizing the chromatin state and propagating such states through cell division robustly (Bernstein *et al.*, 2007). A well-studied case is the trimethylation of histone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

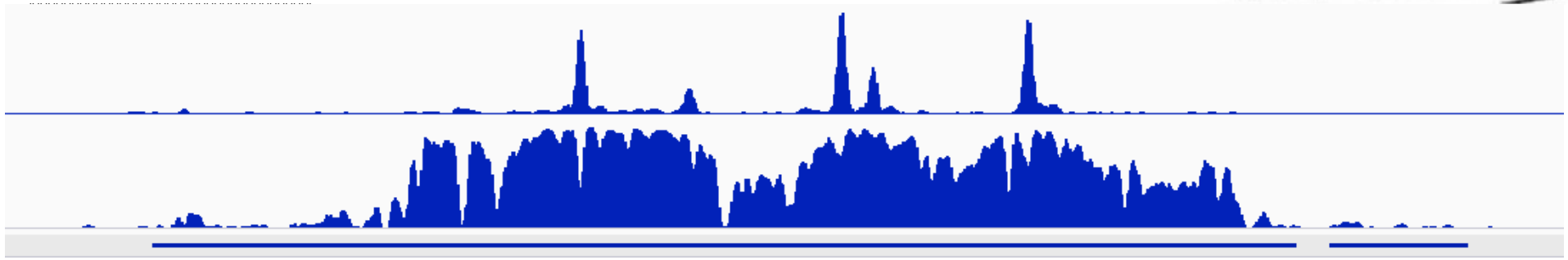
ChIP-seq data analysis overview



```
@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
```

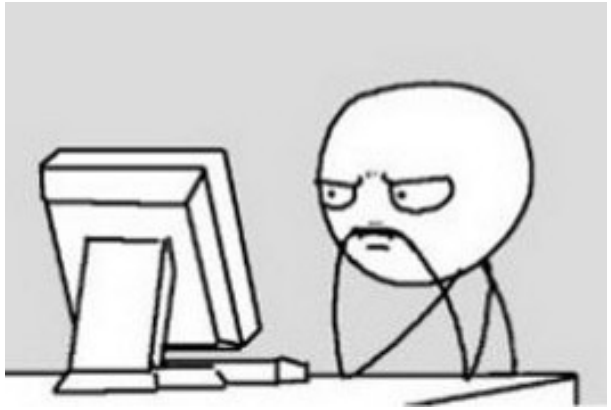


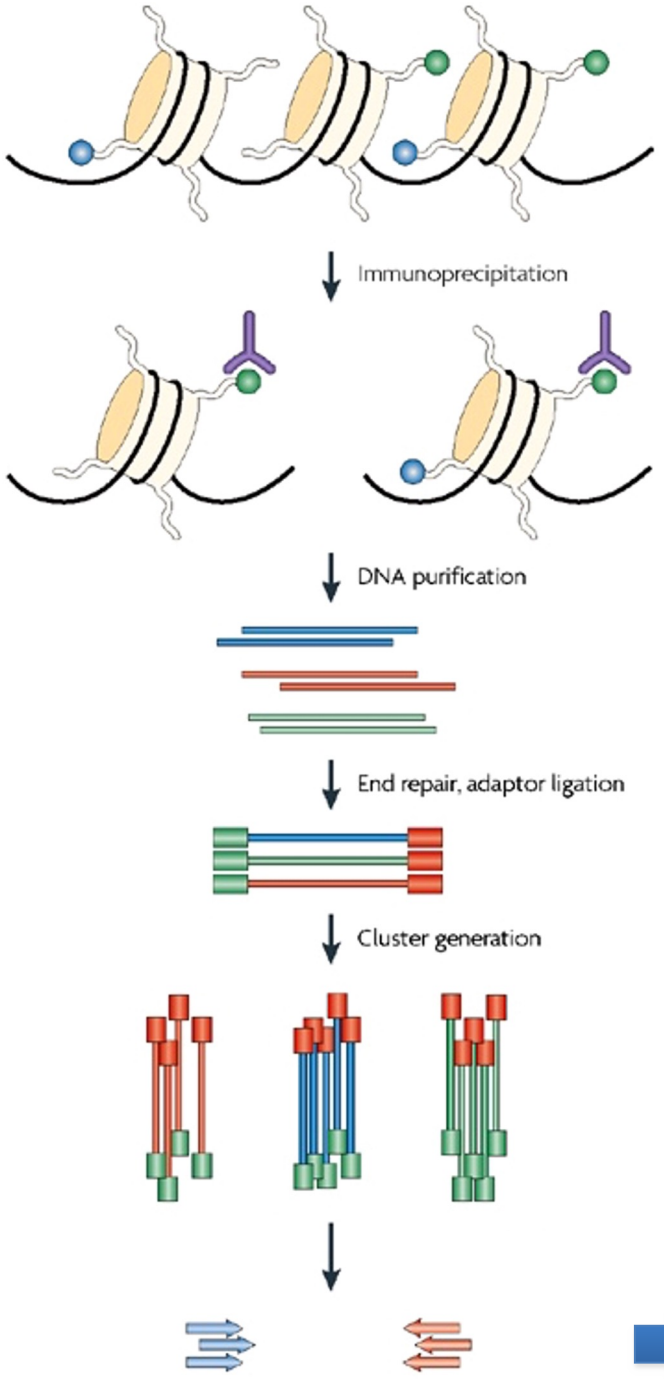




ChIP-seq data analysis goals

- 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





Experimental procedure

Biology

downstream analysis/integration

Peaks (bed)

peak calling

Pile-up for visualization (bedGraph, wig, bigwig)

MACS/ SICER

Non-redundant reads (sam/bam/bed)

redundancy assessment

Mapped reads (sam/bam/bed)

alignment (**bowtie2, BWA**)

Raw sequence reads (fastq)

Computational analysis

Data formats

- fastq data file:

```
@SRR3728822.1 97ZZTR1:422:C57C3ACXX:7:1101:1249:1883/1
NCAAGACCAGTGTCATGAAGCTTTCTCCCTCTTCTAGGAGTTTTACAGCTC
+
#11ADDFFFBFDHGGIJIJJJIEIIJJGHGJFIIHCEHHE?DFHGFHIIJE
@SRR3728822.2 97ZZTR1:422:C57C3ACXX:7:1101:1461:1902/1
NGATTCATAGCTGAATTCTACCAGATGTACAAAGAAGAGCTAGTACCATTA
+
#1=BDFFFHHHHHJJJJIJJJJJJJJJJJJJJIJJJJJGIIJHGIIJIJC
@SRR3728822.3 97ZZTR1:422:C57C3ACXX:7:1101:1479:1943/1
TGAAGATGGCTGAGAACTCCTTAACAGGCAAAATAGGTTTTGTTGGCCGGG
+
C@@FFFFFGHHGGJJJIJJJJJIJJJJIGJFGIJIJJFGHIIIJJIJJIIG
@SRR3728822.4 97ZZTR1:422:C57C3ACXX:7:1101:1287:1958/1
ATATGAACAAACCTTACCTCAGTGGATTCTCAGAACAACCTCTTGAGGTAT
+
CCCFFFFFGHGHHJJIJJJJJJHJJJIJJJIJGIJJJIGHGIJJJGIJEFE
@SRR3728822.5 97ZZTR1:422:C57C3ACXX:7:1101:1515:1796/1
NATTGTGTTTTAGTCTGAAATATCATTTCATGTGGAGAATTCCTTACTGTC
+
#1=DDDFDHHHDHHIIJIIIIJJIHHGHIIGIIIIGIIGIJJIJJIIJJFF
@SRR3728822.6 97ZZTR1:422:C57C3ACXX:7:1101:1585:1807/1
NATAGTTAAAACGGTCTTTCTTTTTGAGATGGAATTTTGCTCTTGTTGCCC
+
#4=DFFFFHHHHHJGIIJJJJJJJJJIIHIJJJIJJJJJJJJJJJJJJJJ
```

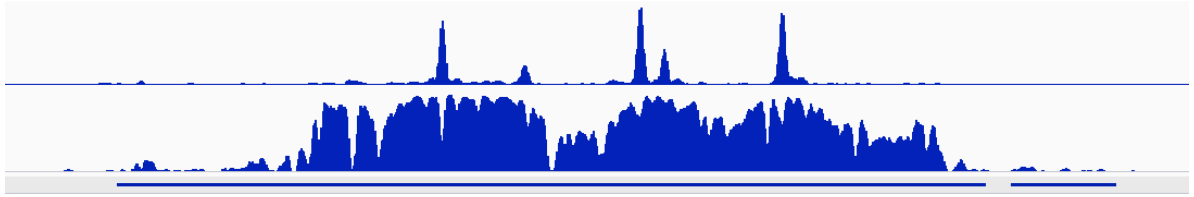
Data formats

- BED:

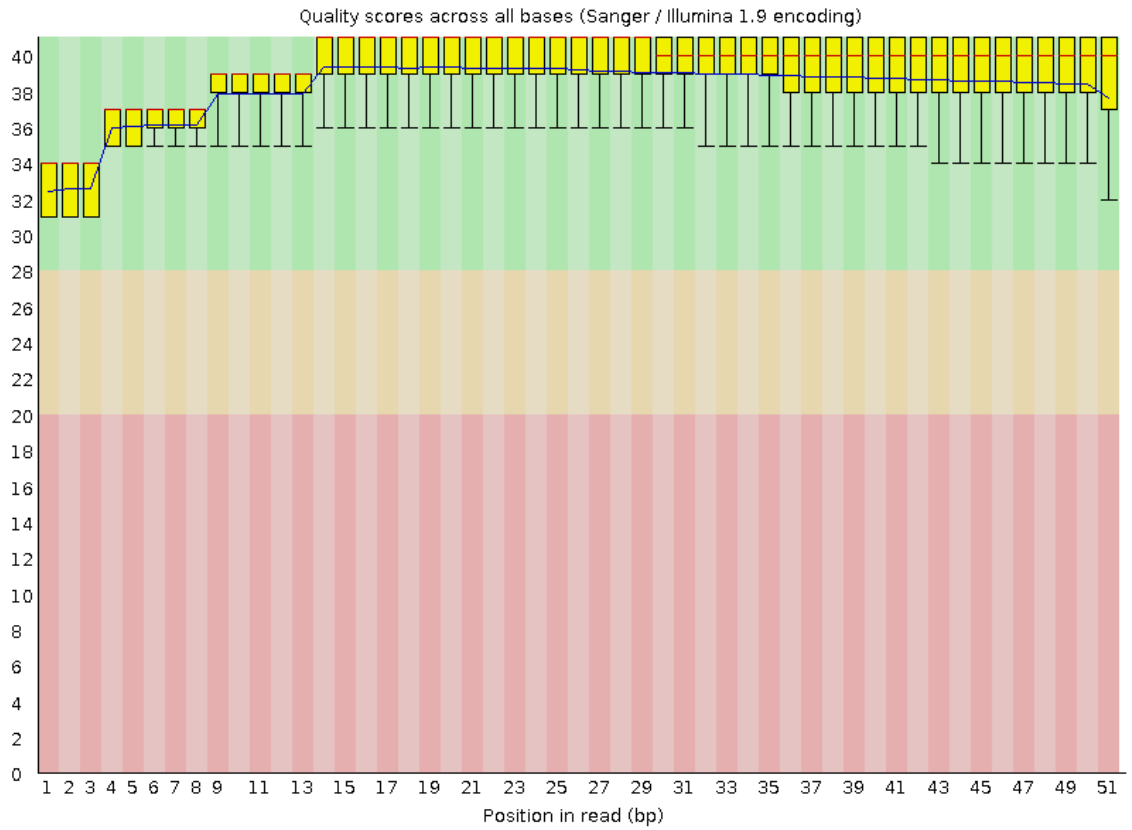
```
chr11    10344210    10344260    255    0    -
chr4     76649430    76649480    255    0    +
chr3     77858754    77858804    255    0    +
chr16    62688333    62688383    255    0    +
chr22    33031123    33031173    255    0    -
```

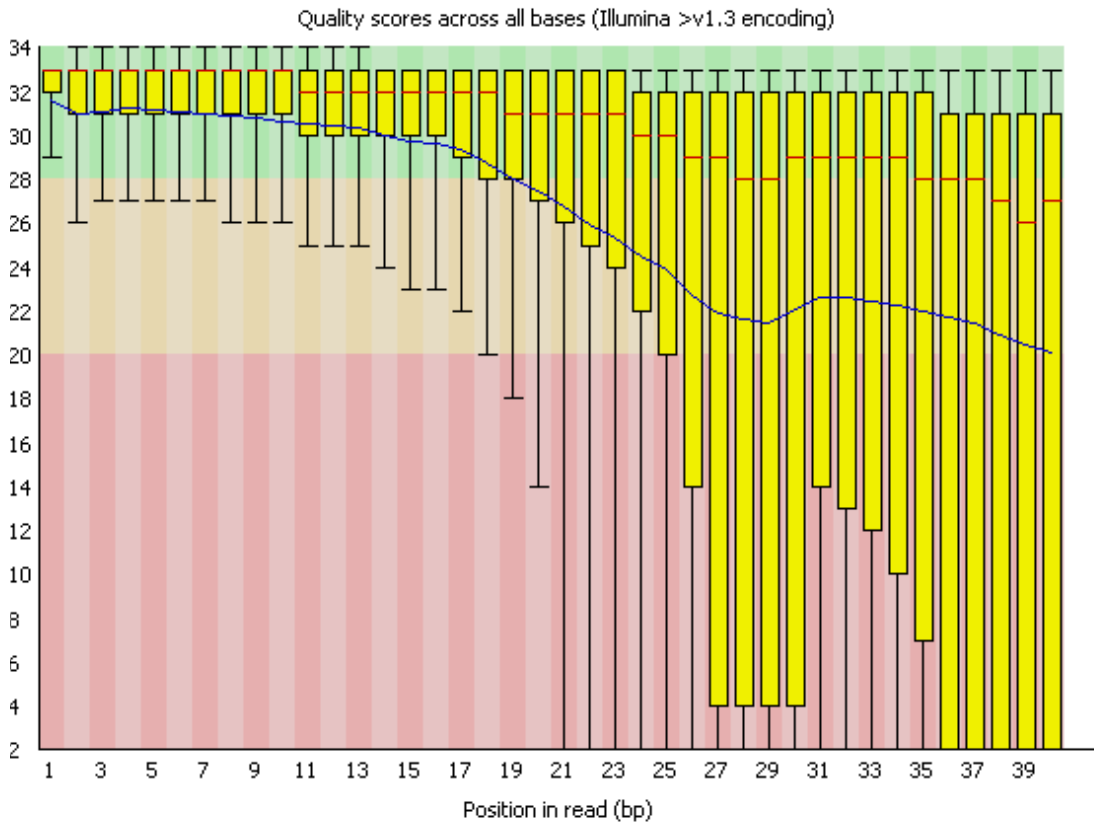
- SAM/BAM: aligned sequencing reads

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

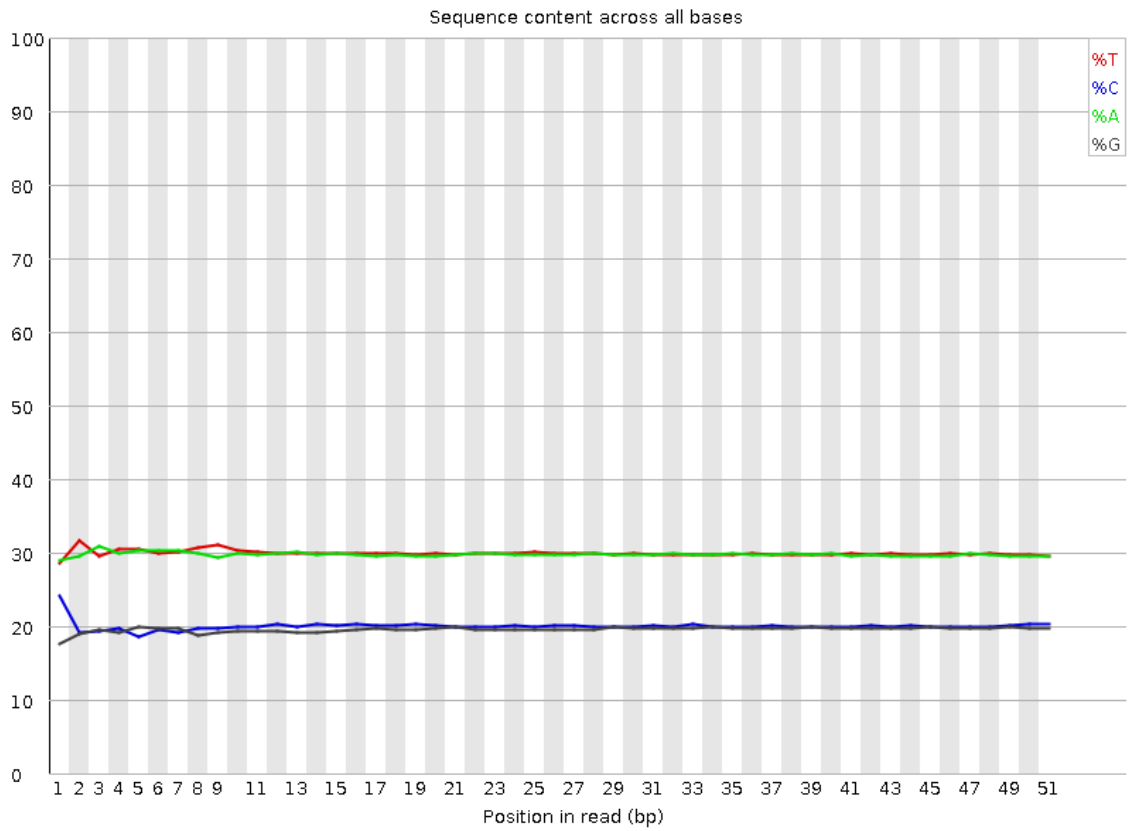


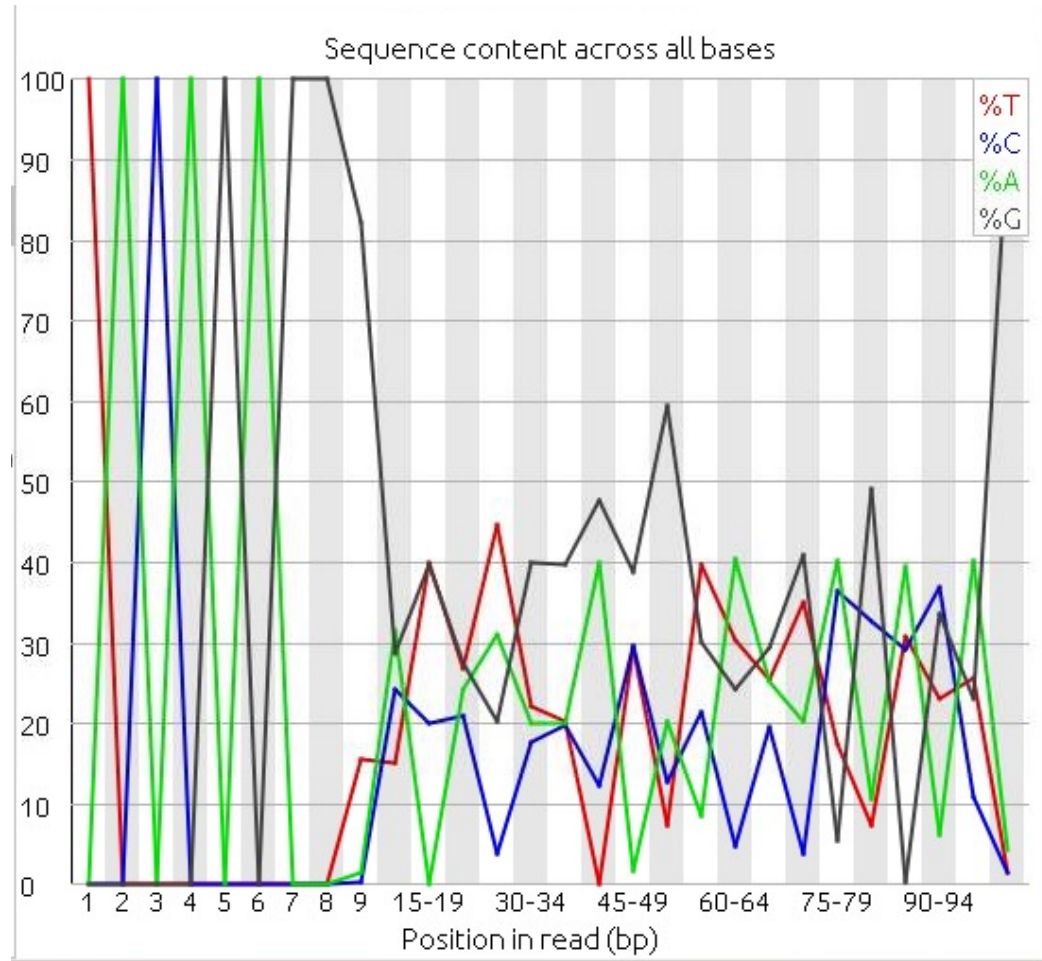
Sequencing quality assessment: fastqc





Sequencing quality assessment: fastqc



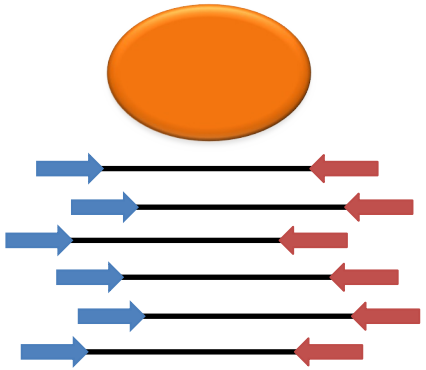


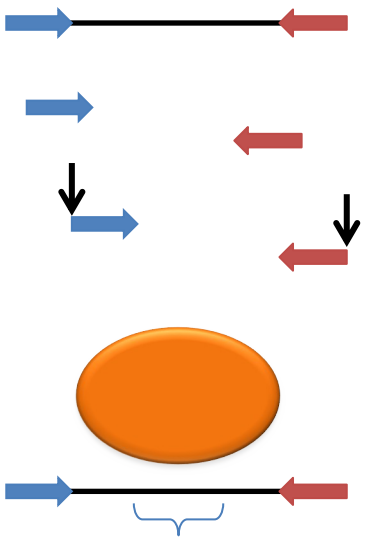
ChIP-seq read mapping

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

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

- Concepts/Terminology:





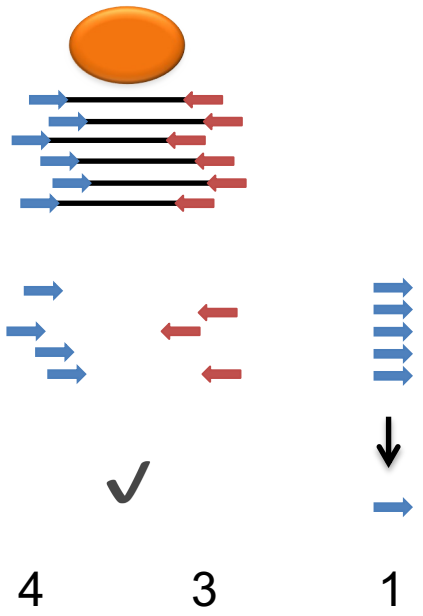
DNA fragment

Reads (tags)

Cleavage (cut) site

Binding site

Redundancy control



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

- Non-redundant rate:

$$\frac{\text{\# non-redundant reads}}{\text{\# mapped reads}}$$

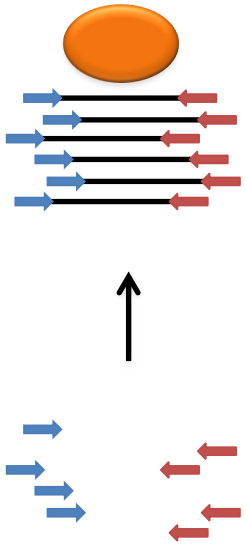
8/12 = 66.7%

- PBC (PCR Bottleneck Coefficient):

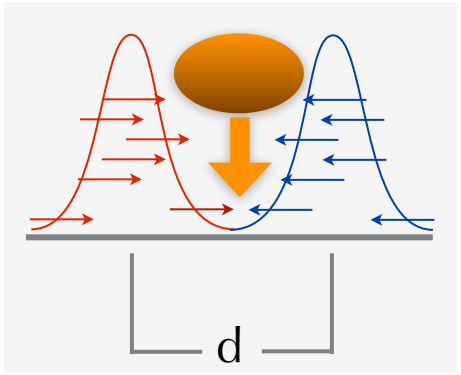
$$\frac{\text{\# locations w/ 1 read}}{\text{\# locations w/ reads}}$$

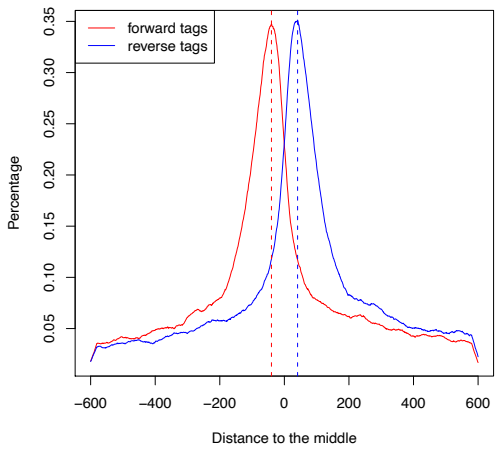
7/8 = 87.5%

DNA fragment size estimation

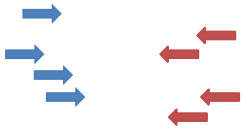


peak model (MACS)

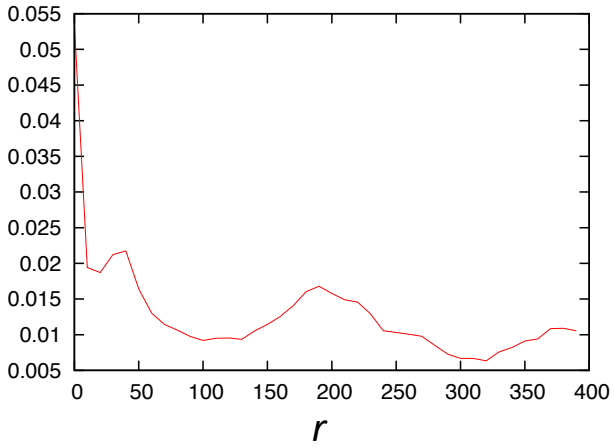




cross-correlation (SICER)



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



Cross-correlation

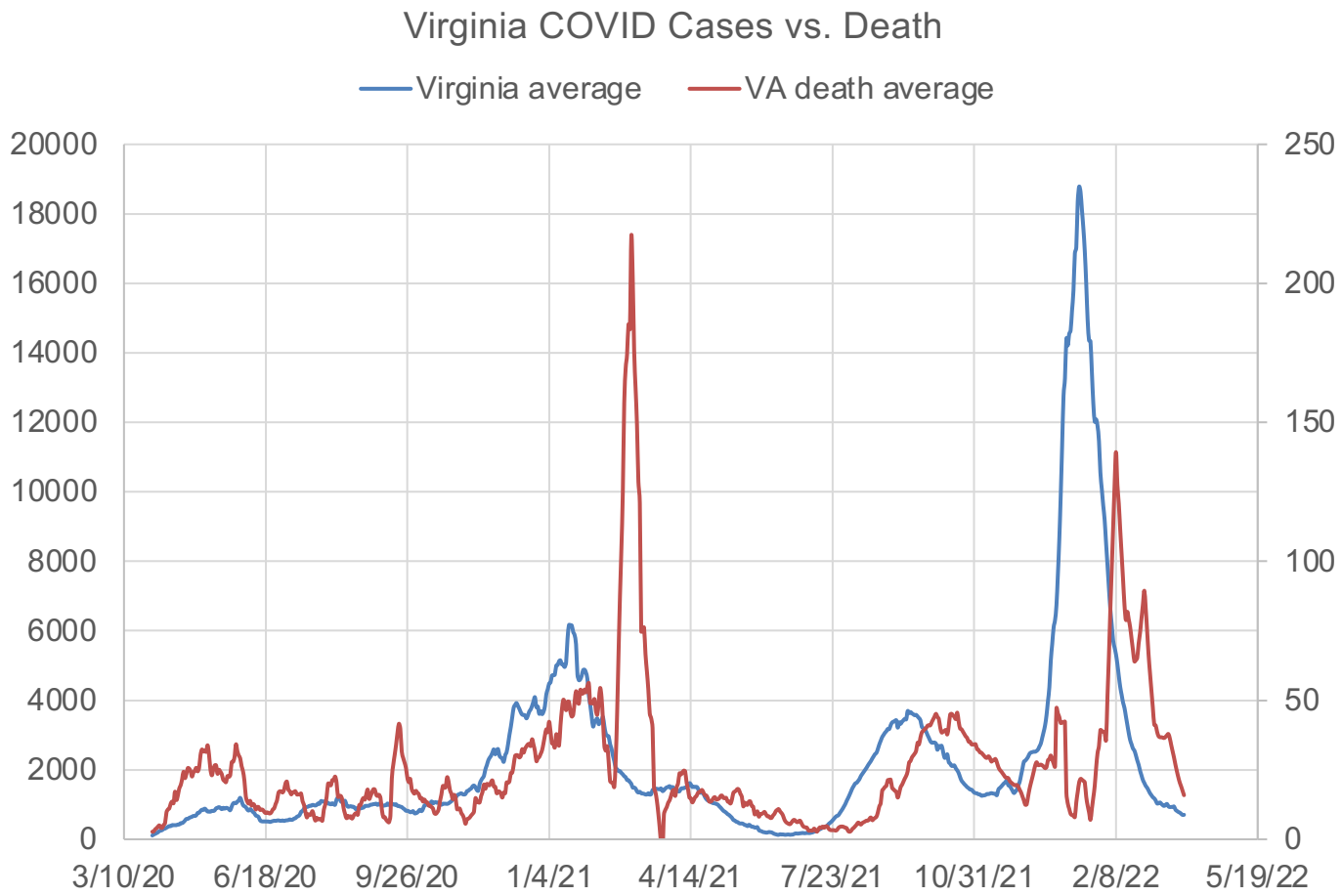
- Correlation between two strings with a displacement

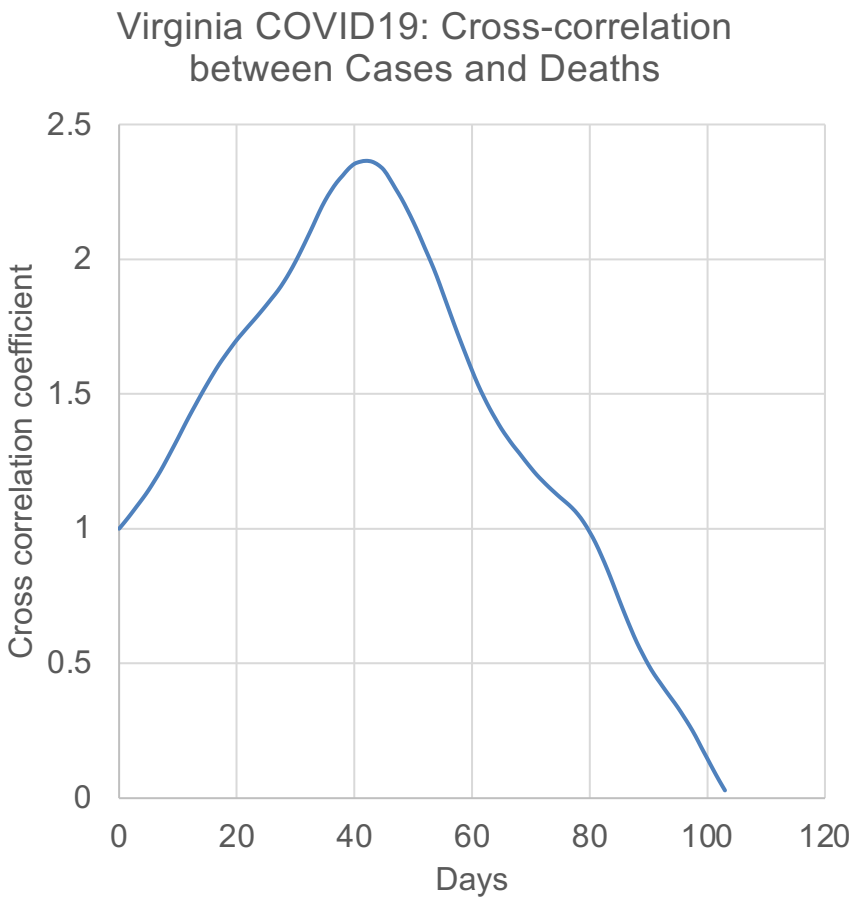
$$R_{XY}(r) = \sum_t \left(X(t) - \overline{X}\right)\left(Y(t+r) - \overline{Y}\right)$$

- Auto-correlation: Cross-correlation with itself

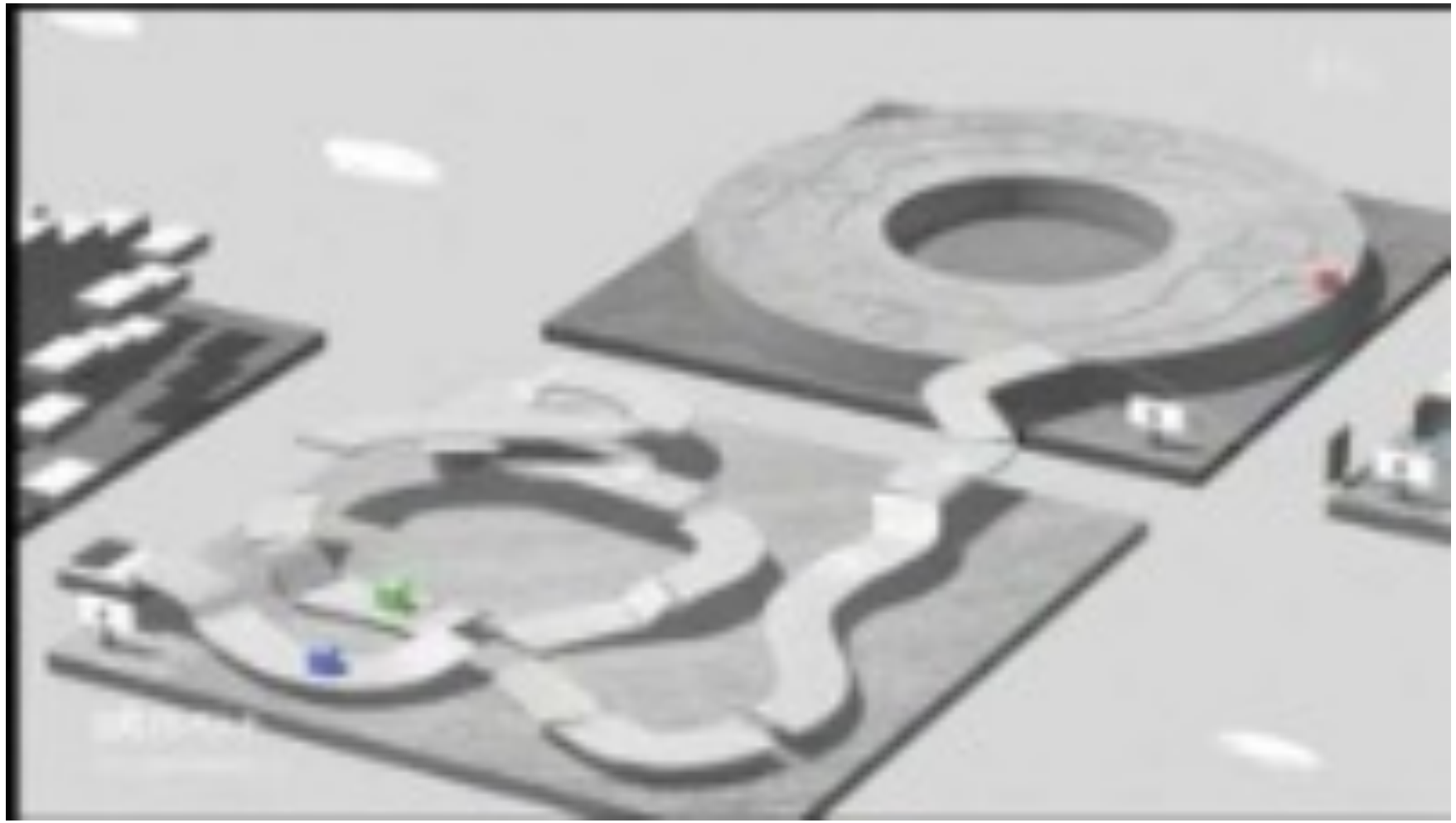
$$R_{XX}(r) = \sum_t \left(X(t) - \overline{X}\right)\left(X(t+r) - \overline{X}\right)$$

Cross-correlation: example





Auto-correlation: example



https://youtu.be/B7BFgCXCqGU

Auto-correlation: example

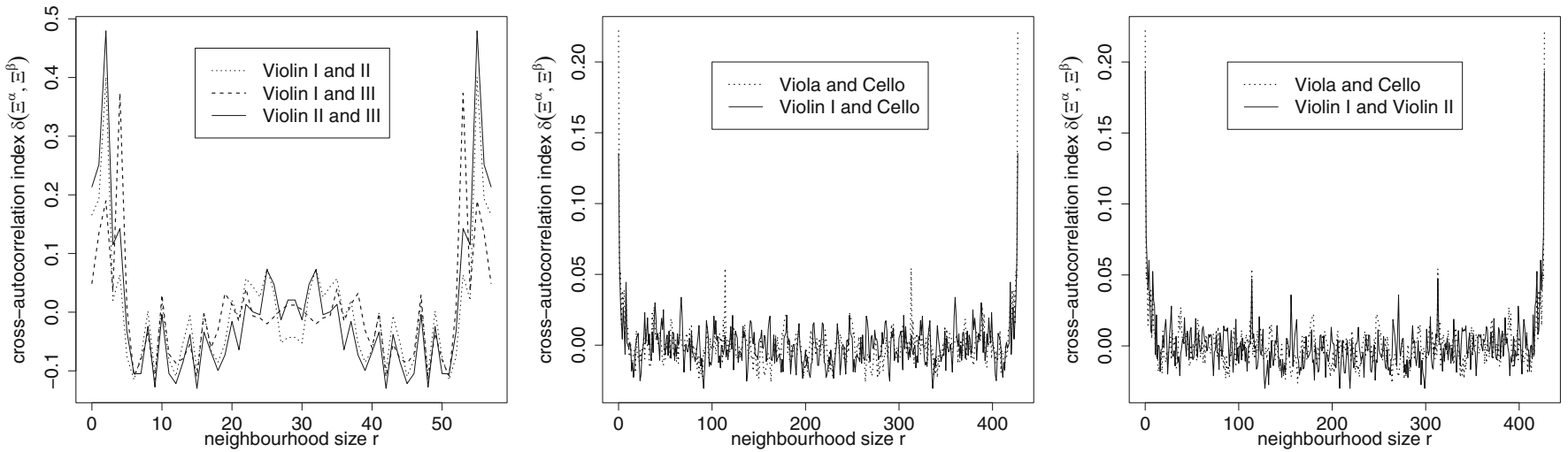
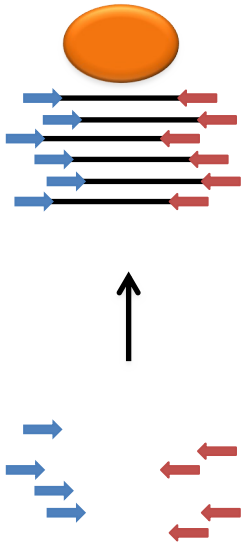


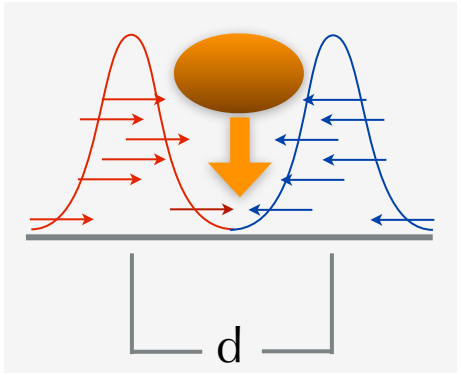
Fig. 6 Cross-autocorrelation index according to the lag r varying from 0 to n. *Left*: *Canon in D Major by Pachelbel* with τ equal to a measure. *Middle and right*: *first movement of the String Quartet No. 1 in F major, Op. 18 by Beethoven* with τ equal to a measure

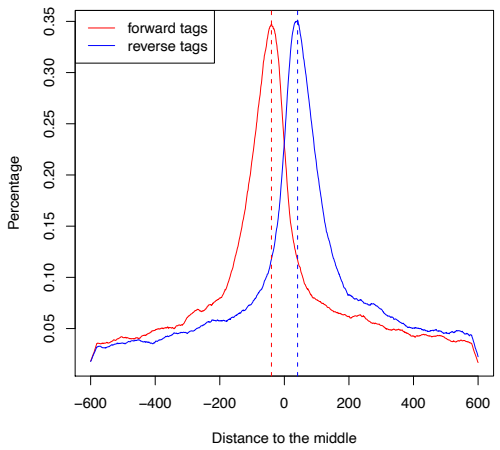
Cocco C & Bavaud F, 2015

DNA fragment size estimation

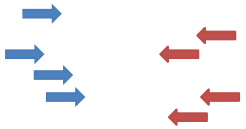


Peak model (MACS)
For TF

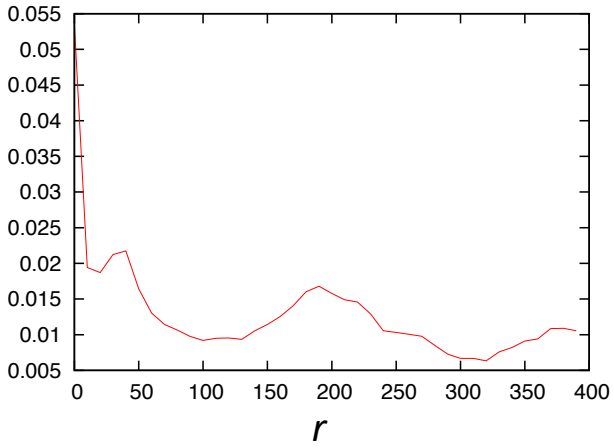




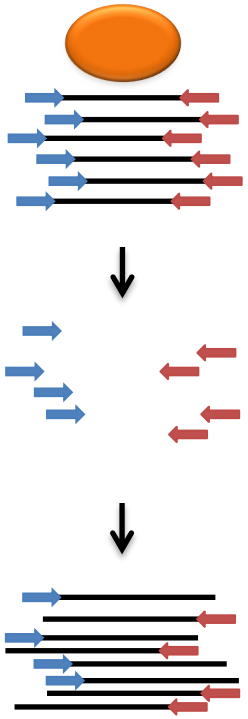
Cross-correlation (SICER)
for any ChIP-seq (input)



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



Retrieve DNA fragments



- Full length retrieval (MACS)

Extend *d*

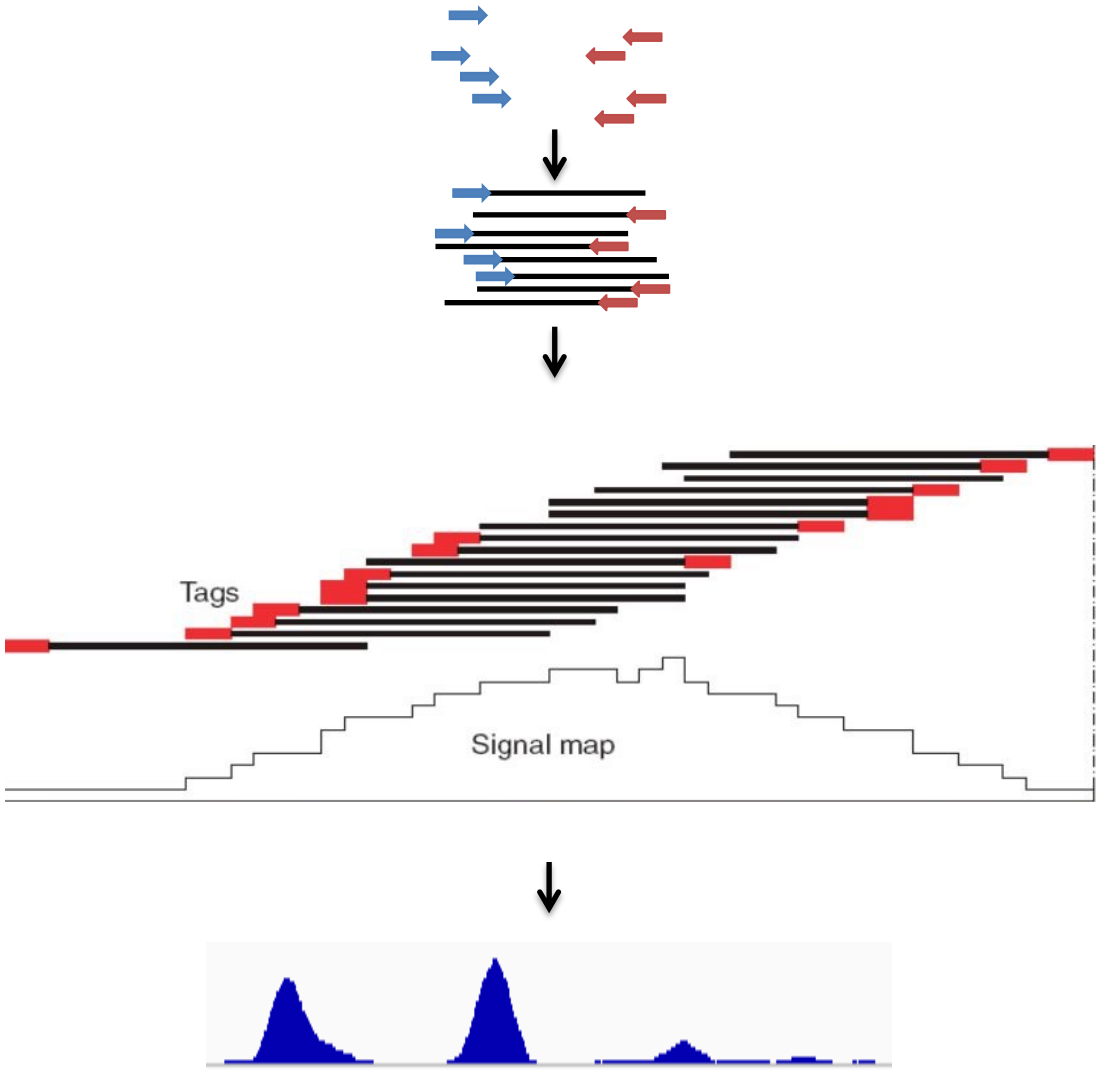
- Partial retrieval (sharpen the signal)

Shift *d/2*
Expand to *d/2*

- Point retrieval (SICER)

Shift *d/2*

Pile up: Signal map generation



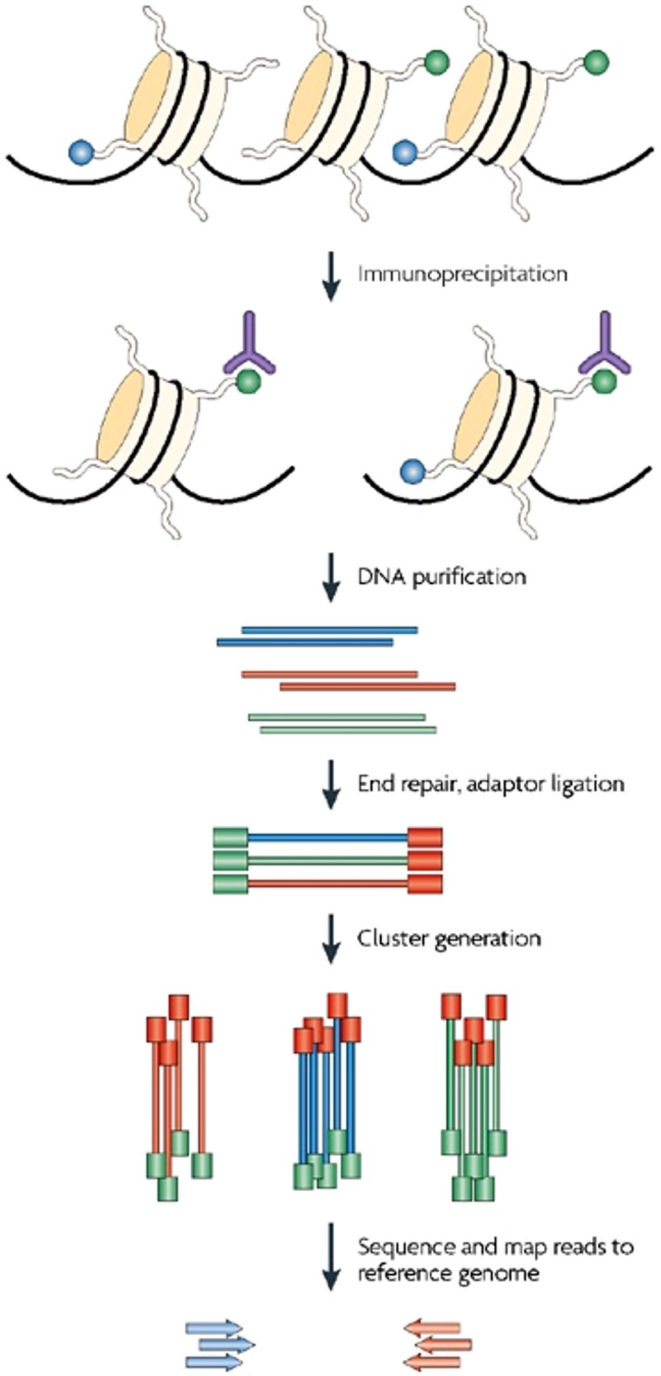
- bedGraph:

```
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



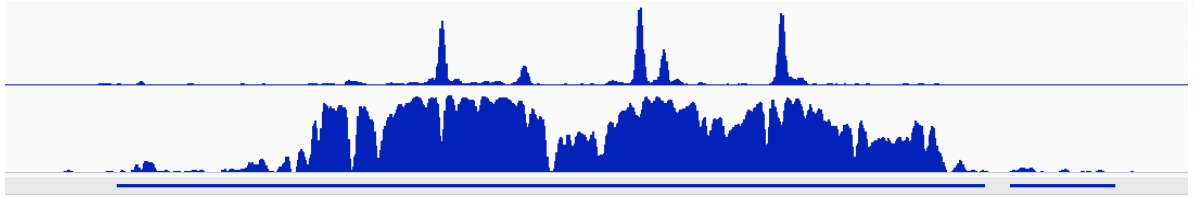
ChIP-seq: Study design

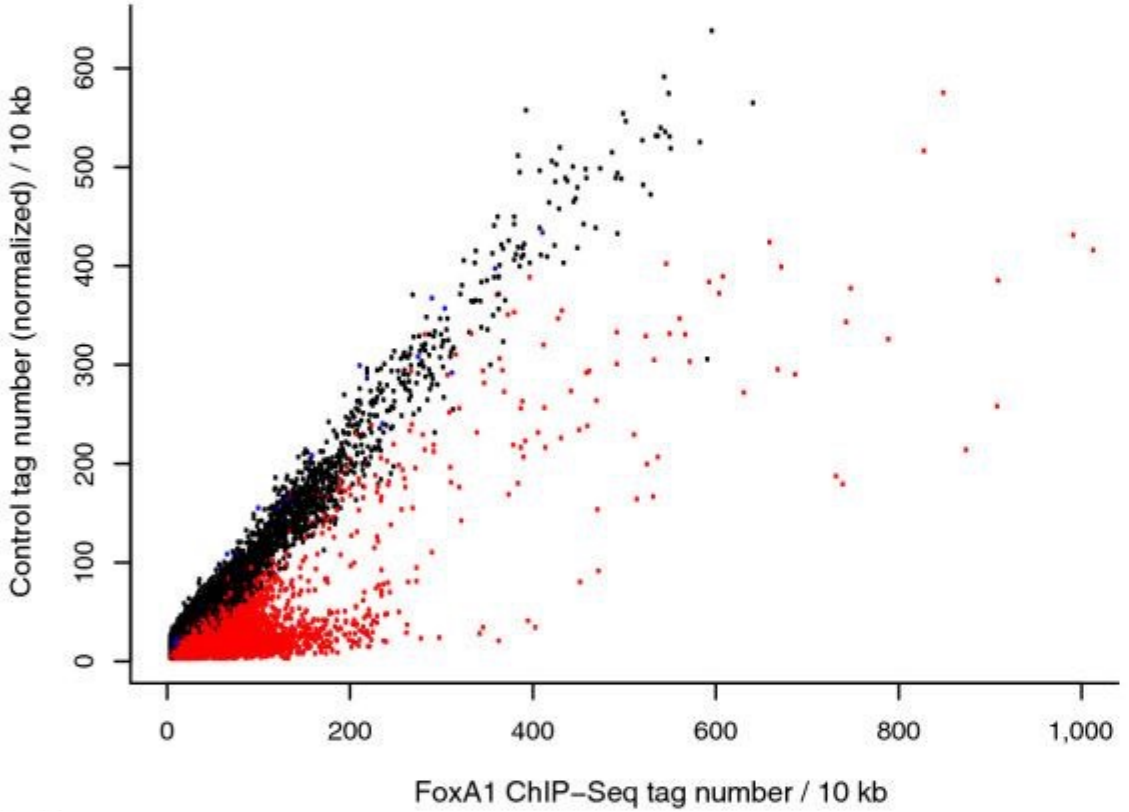
- Background Control: **Input** or **IgG**
  - Input chromatin: sonicated/digested chromatin without immunoprecipitation
  - IgG: “unspecific” immunoprecipitation

- Study Control:
  - Control exp sample: ChIP + input
  - Treated exp sample: ChIP + input

ChIP-seq: Peak calling

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





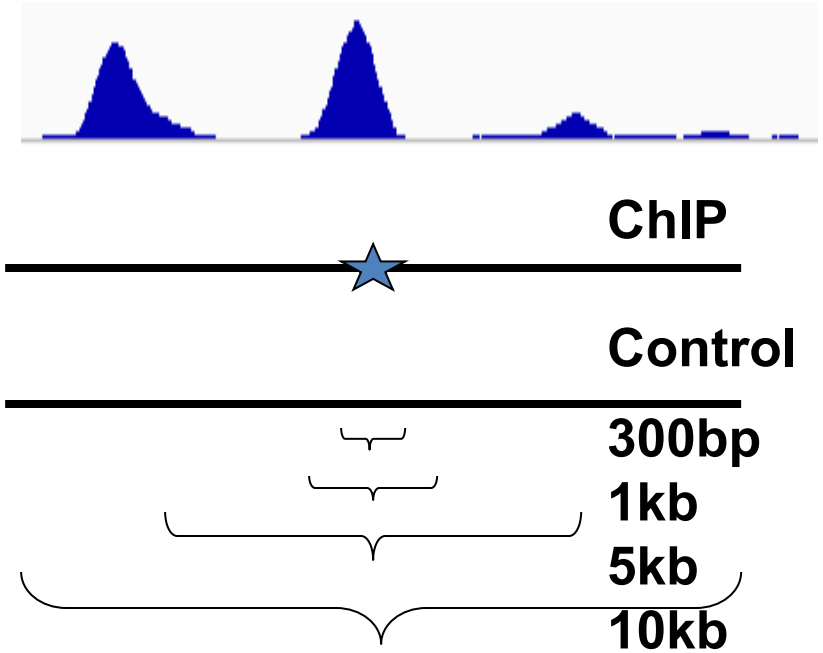
MACS: model

- **M**odel-based **A**nalysis for **C**hIP-**S**eq
- 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

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

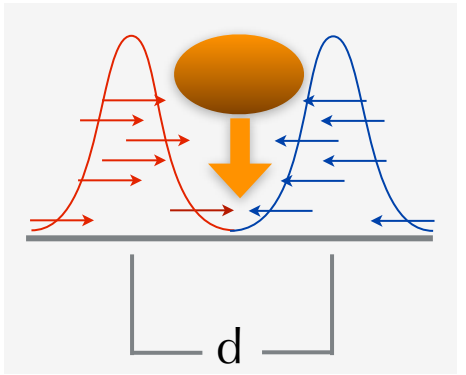
- B-H adjustment to correct for FDR
  - p-value → q-value

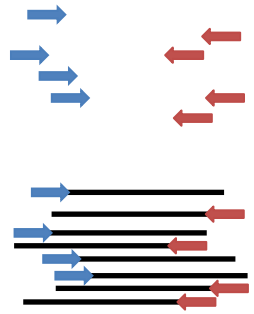


Zhang et al, *Genome Bio*, 2008

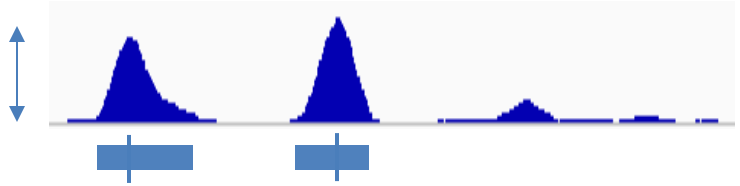
MACS: Output interpretation

```
# tag size is determined as 51 bps
# total tags in treatment: 19442622
# tags after filtering in treatment: 17218335
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.11
# d = 141
# alternative fragment length(s) may be 141 bps
```



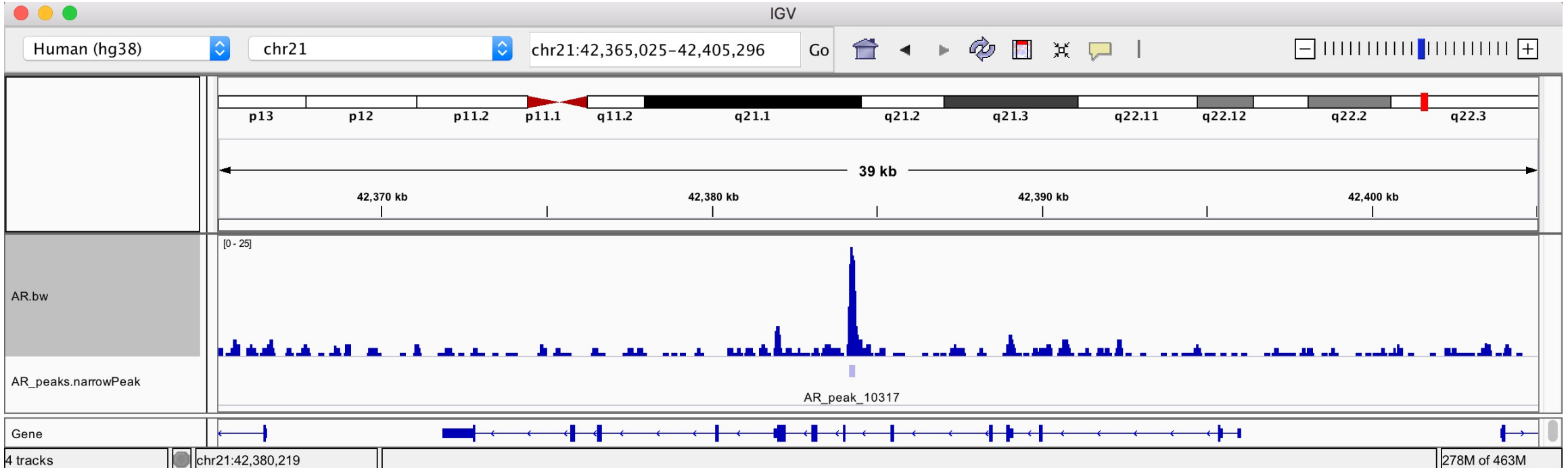


chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
chr1	2603	2989	387	2870	18.00	6.68596	3.52825	3.66748	AR_peak_1
chr1	138179	138371	193	138281	18.00	14.90779	7.93021	11.47829	AR_peak_2
chr1	36515	36714	200	36609	16.00	12.59143	7.05394	9.25447	AR_peak_3
chr1	201091	201231	141	201114	10.00	7.58293	5.23859	4.50002	AR_peak_4
chr1	69373	69558	186	69452	18.00	9.61904	4.93737	6.41821	AR_peak_5



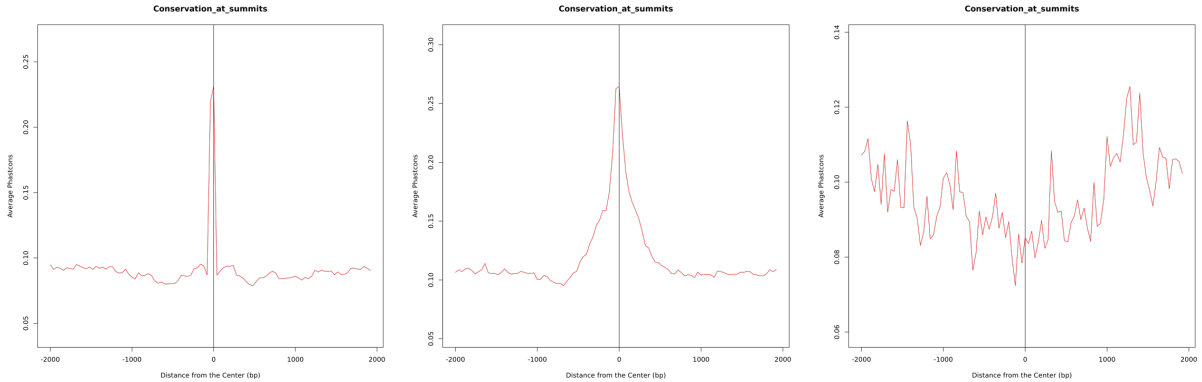
Data Visualization

- bedGraph to bigWig
- macs2 output data
- IGV

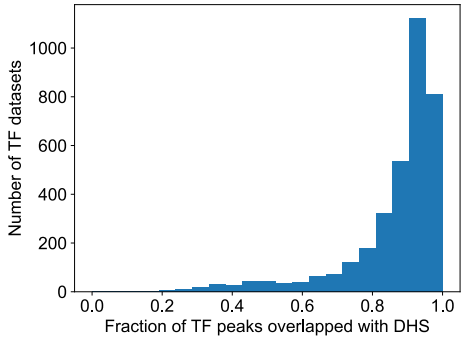


Quality Control

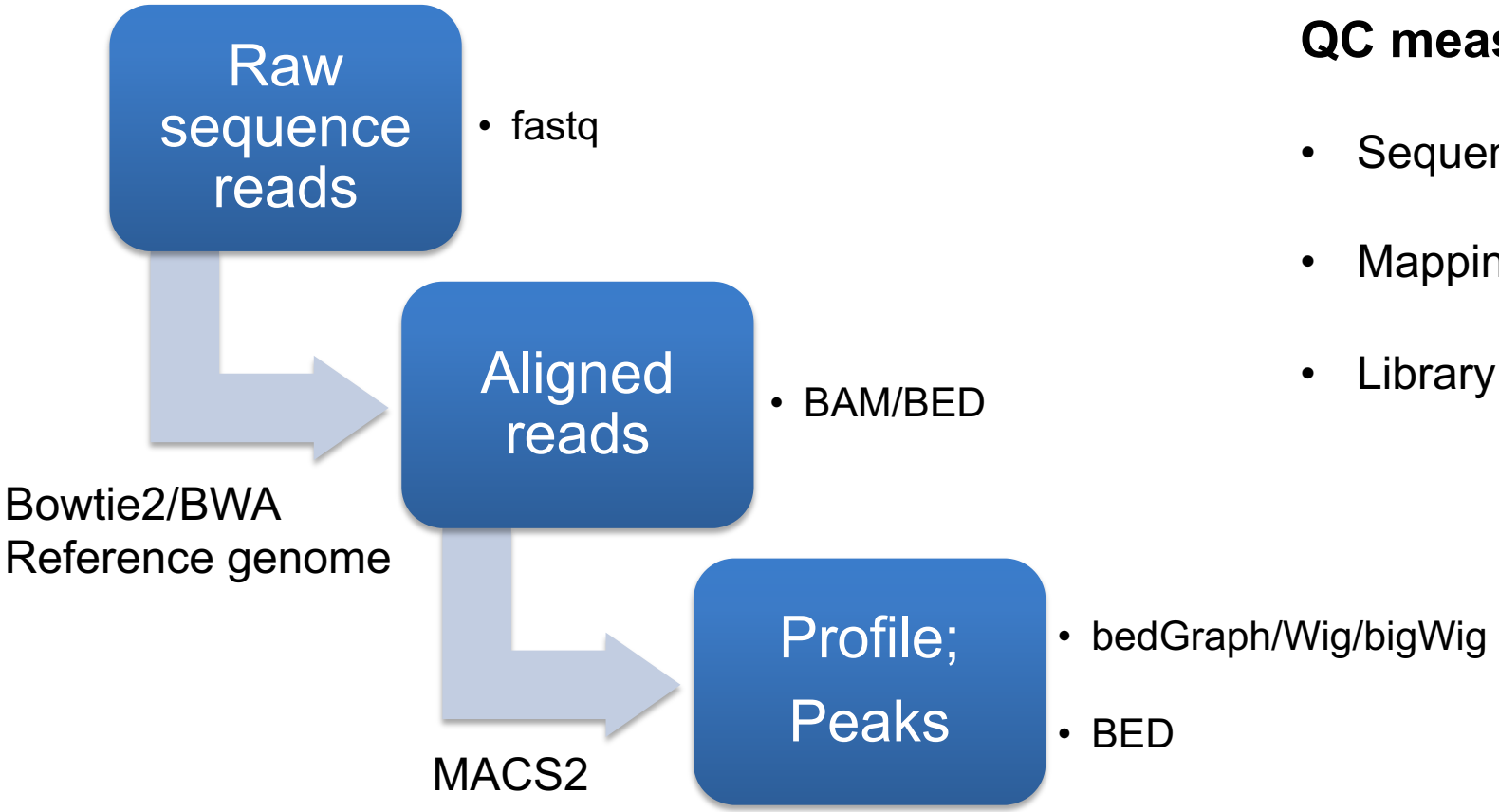
- FRiP (Fraction of Reads in Peaks) score
  - 1-10% for TF is normal
- Number of peaks
  - Number of peaks with high fold-enrichment, e.g, 5, 10, …
  - 2000
- Sequence conservation



- Fraction of peaks within regulatory regions
  - 80%

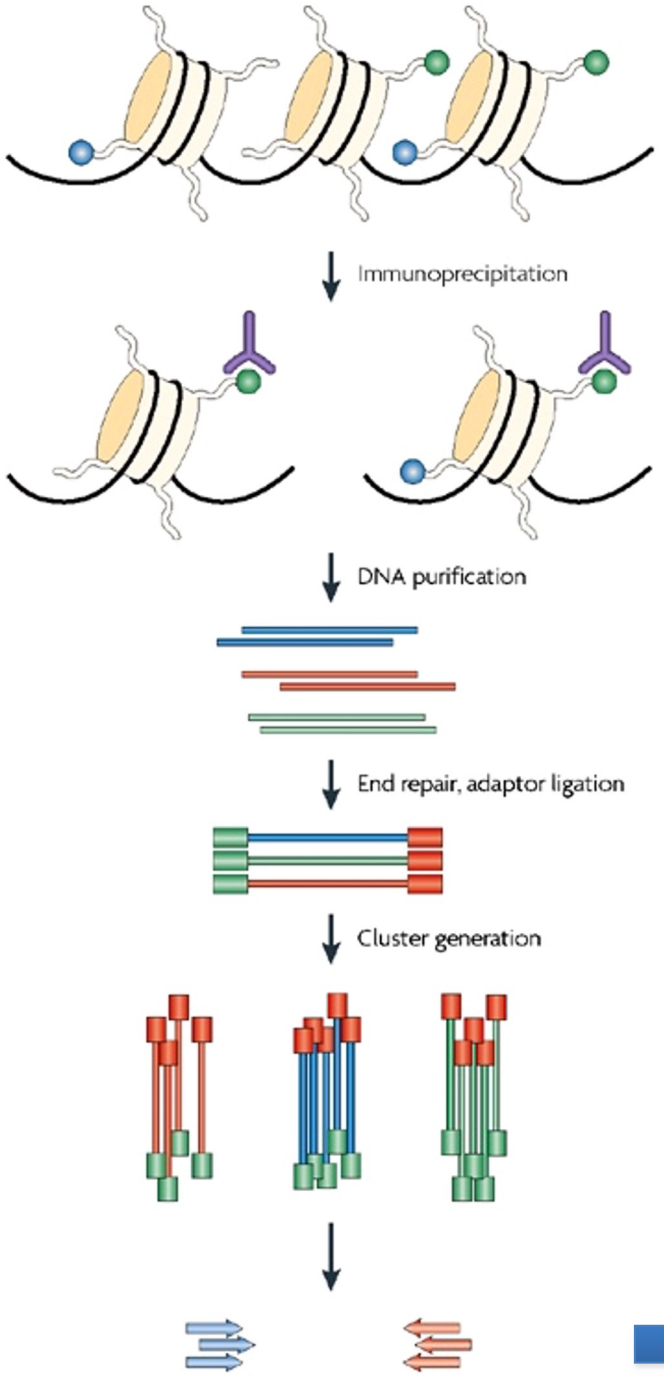


Data flow and QC summary



QC measures

- Sequence quality (fastqc)
- Mapping quality (uniquely mapped ratio)
- Library complexity (PBC)
- Fold enrichment, peaks
- Signal/Noise (FRIP score)
- Regulatory annotation



Experimental procedure

Biology

downstream analysis, data integration

Peaks (bed)

peak calling

Pile-up for visualization (bedGraph, wig, bigwig)

Non-redundant reads (sam/bam/bed)

redundancy assessment

Mapped reads (sam/bam/bed)

alignment (**bowtie2/BWA**)

Raw sequence reads (fastq)

macs/ SICER

Computational analysis