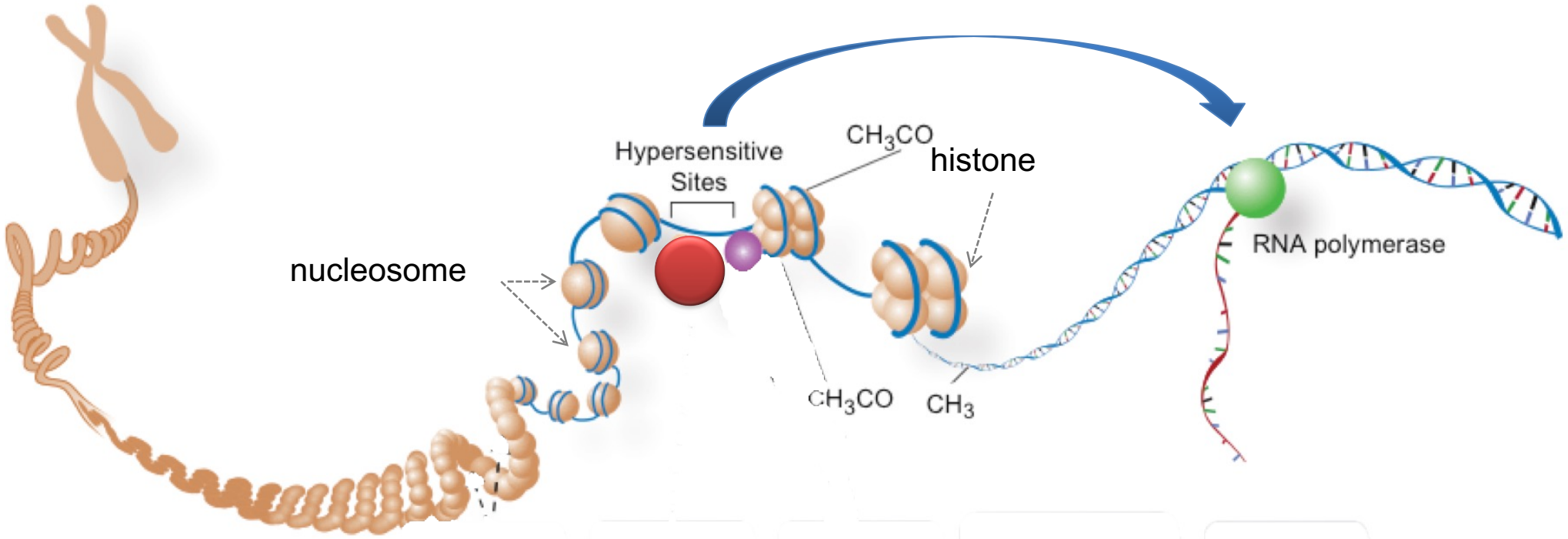
ChIP-seq, Hi-C, and Chromatin Domains

March 29, 2022

Outline

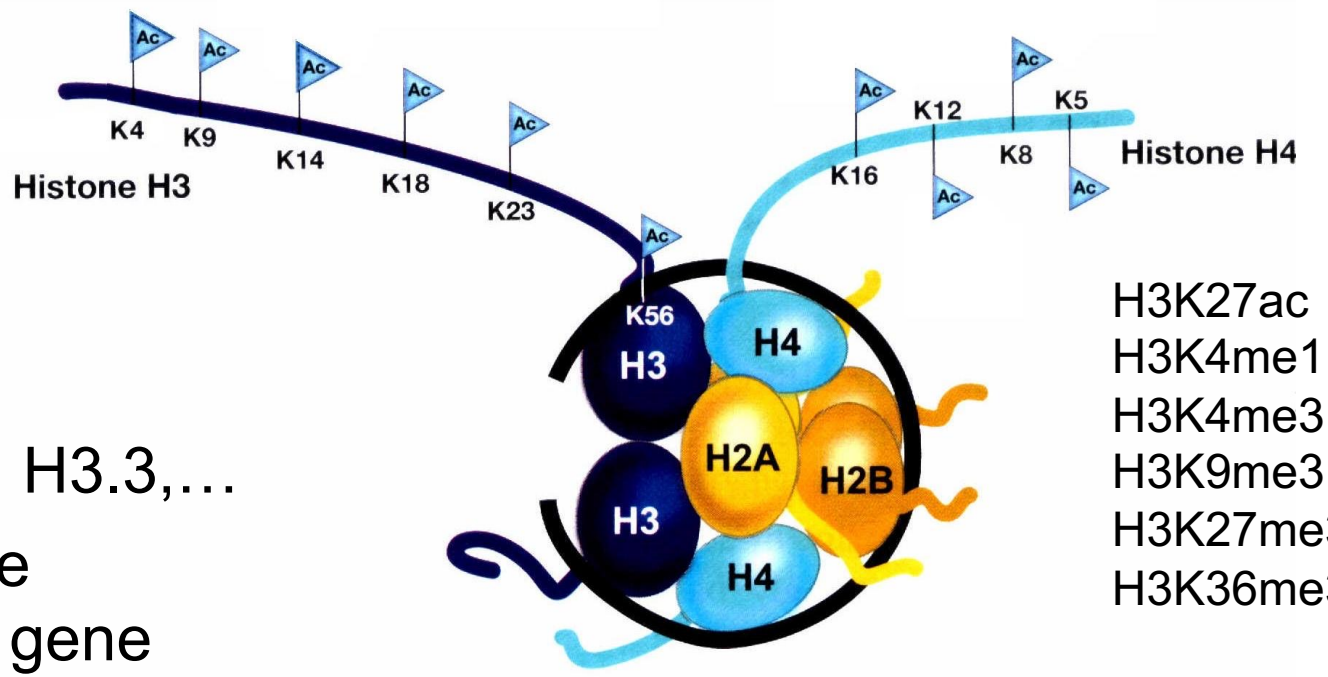
- Histone marks
- ChIP-seq data analysis continued
  - Peaks vs. islands
  - MACS vs. SICER
- Chromatin domains
- Hi-C
- Fractal structure of genome organization

Epigenome


nucleosome
Hypersensitive
Sites
CH_3CO
histone
RNA polymerase
CH_3CO
CH_3

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.



H3K27ac
H3K4me1
H3K4me3
H3K9me3
H3K27me3
H3K36me3

Allis C. *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

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 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

nature genetics

Combinatorial patterns of histone acetylations and methylations in the human genome

Zhibin Wang[1,5], Chongzhi Zang[2,5], Jeffrey A Rosenfeld[3–5], Dustin E Schones[1], Artem Barski[1], Suresh Cuddapah[1], Kairong Cui[1], Tae-Young Roh[1], Weiqun Peng[2], Michael Q Zhang[3] & Keji Zhao[1]

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 systems[3], 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 activation[5–10]. Various models, including the histone code, the signaling network and the charge neutralization model, have been proposed to account for the function of histone modifications[11–14]. 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 previously[15].

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 functionally[16]. 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 technique[15], 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-SAGE[18] with the separately examined patterns of H3K9ac and H3K14ac in

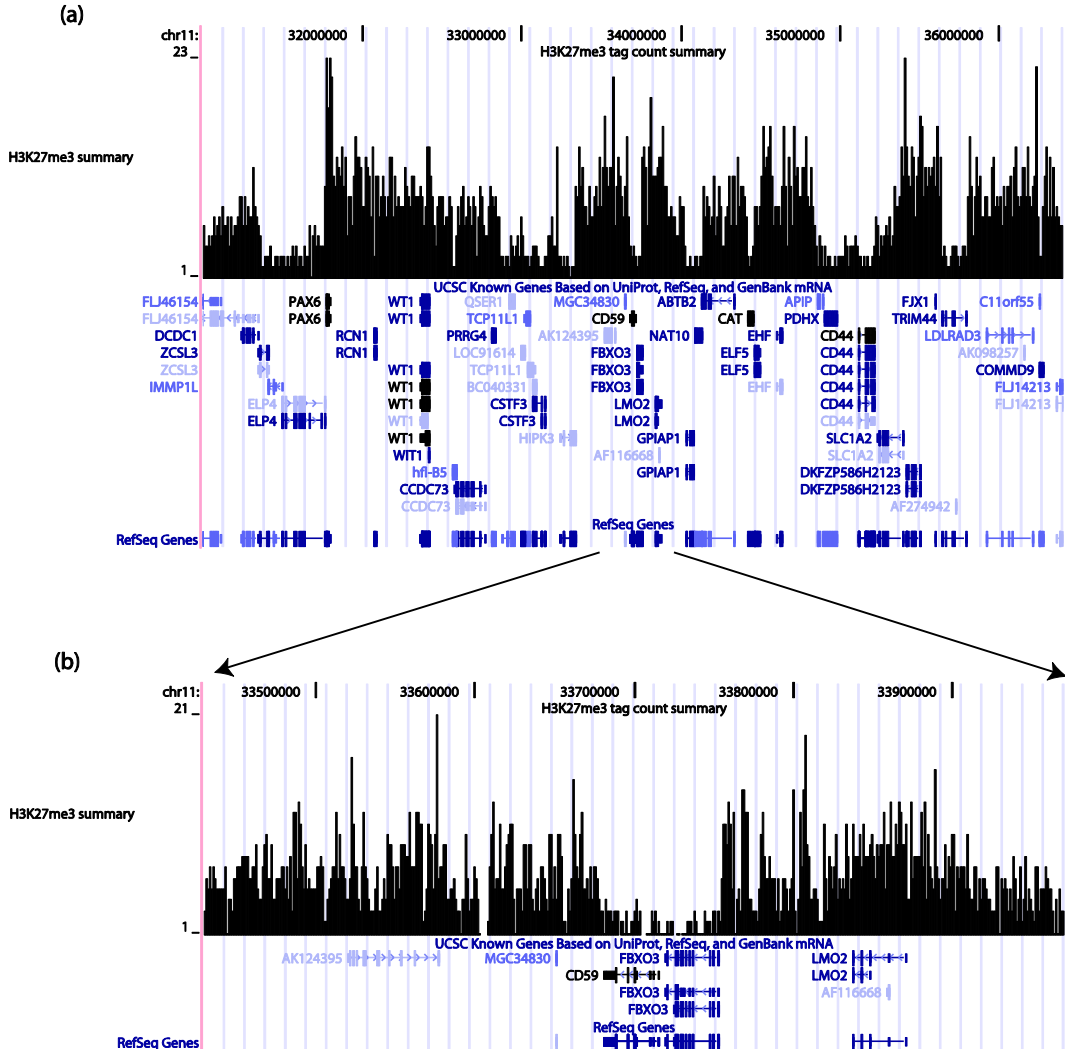
Transcription Factors vs. Histone Marks

	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)

Histone modification patterns are diffuse

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

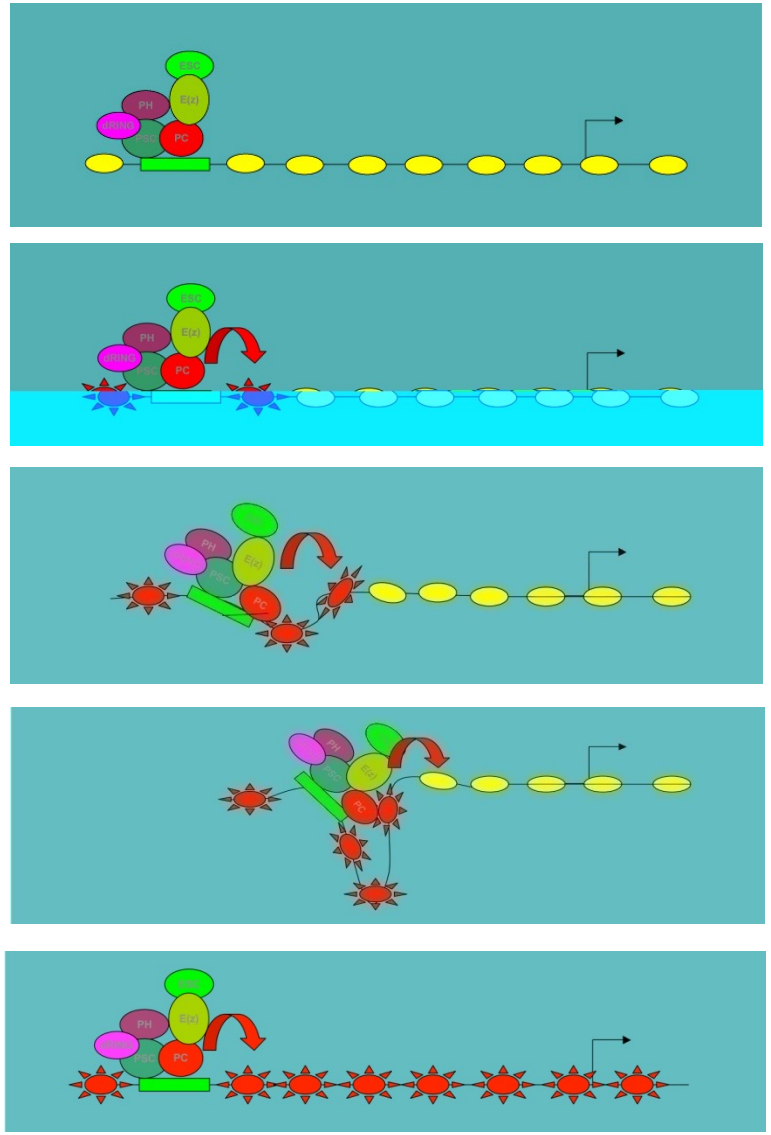
- *Why?*

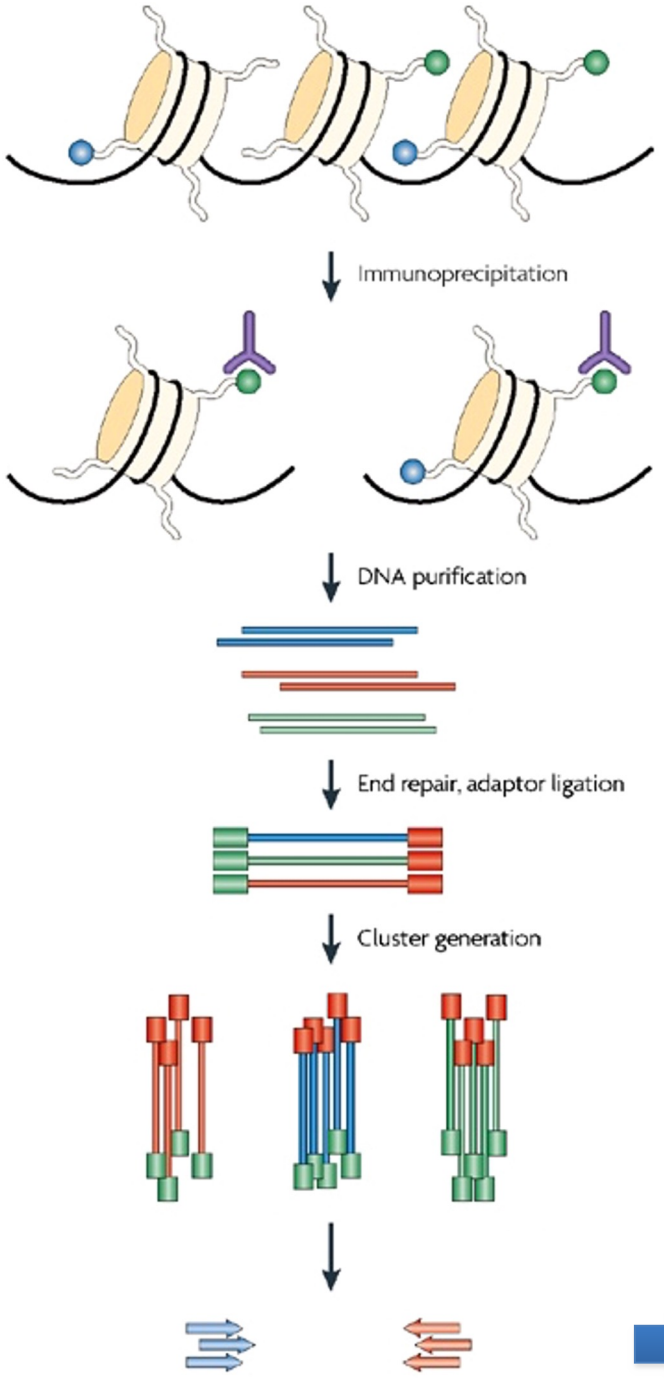


Histone modification tends to spread out

Domain formation model for repressive marks

- HP1
  H3K9me3
- PRC1/PRC2
  H3K27me3





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

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

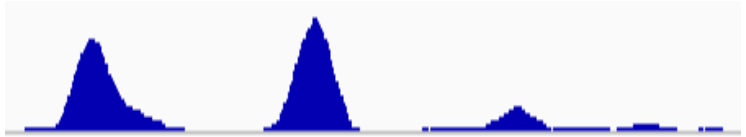
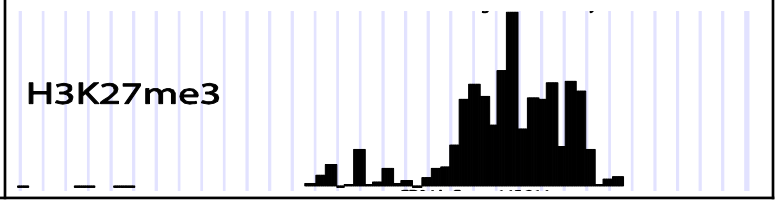
Covalent modifications of chromatin, including DNA methylation and histone modifications, play critical roles in gene regulation

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

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

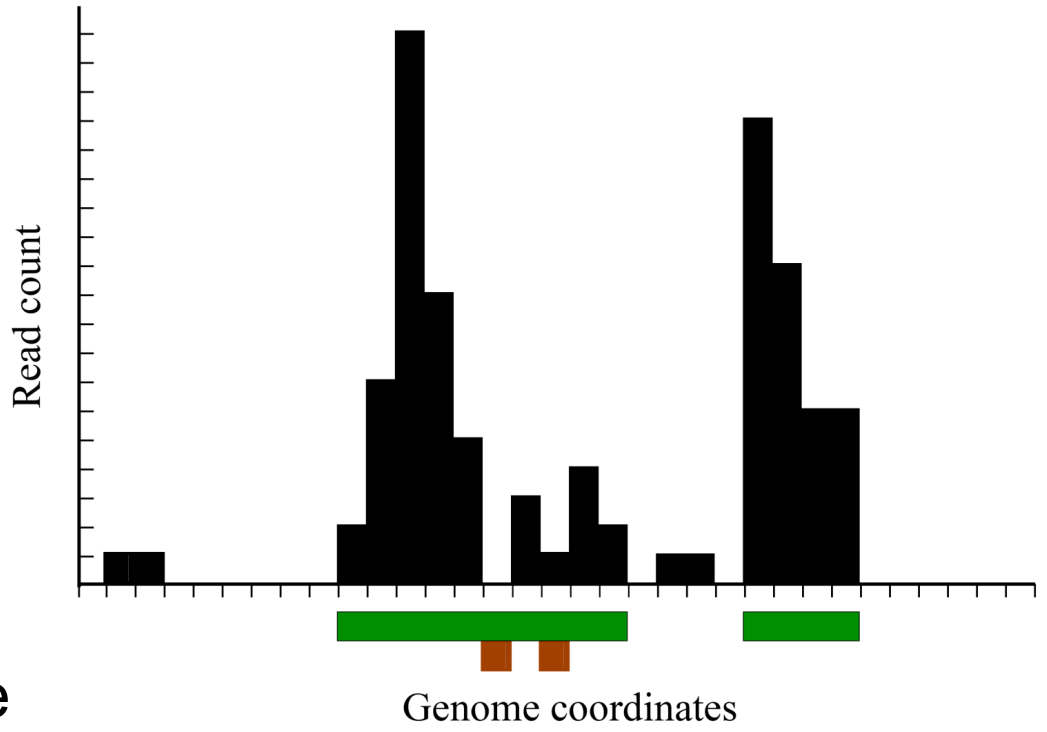
	MACS	SICER
Initial model	Poisson	Poisson
Signal scan	Sliding windows with bandwidth	Non-overlapping bin read count
Peak region identification	Merge windows	Merge windows allowing gaps
Peak scoring	Pile-up signal amount	Aggregate score on islands
Significance modeling	Poisson with dynamic λ	Asymptotic 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

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 $P(m,\lambda)$
  - $\lambda = w\mathrm{N}/L$ 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:

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

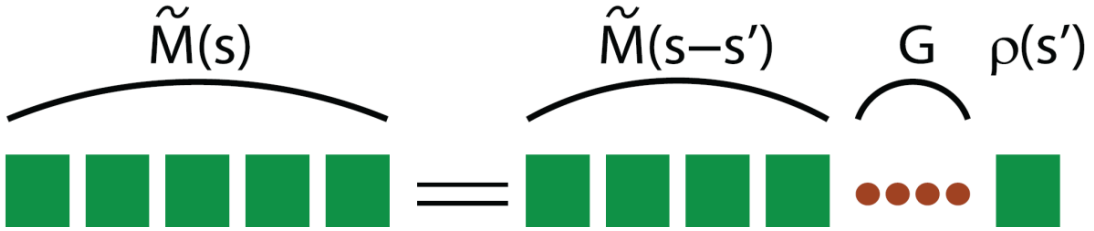
- Probability of a window being 'ineligible':

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

- Gap factor:

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

SICER: Island score statistics



- Recursion relation

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

- Probability of finding an island of score ***s***:

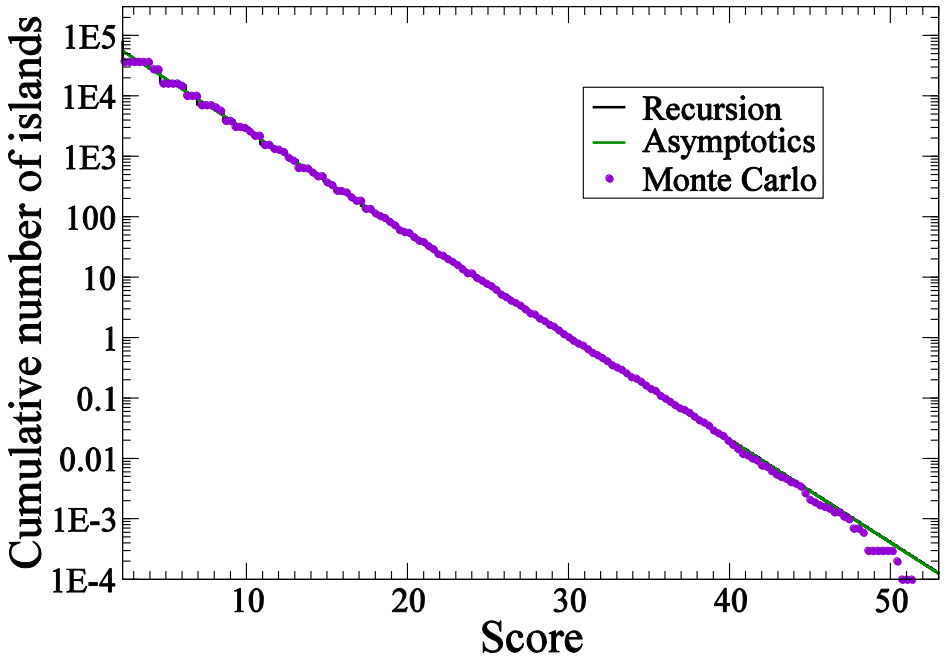
$$M(s) = t^{g+1} \tilde{M}(s) t^{g+1}$$

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

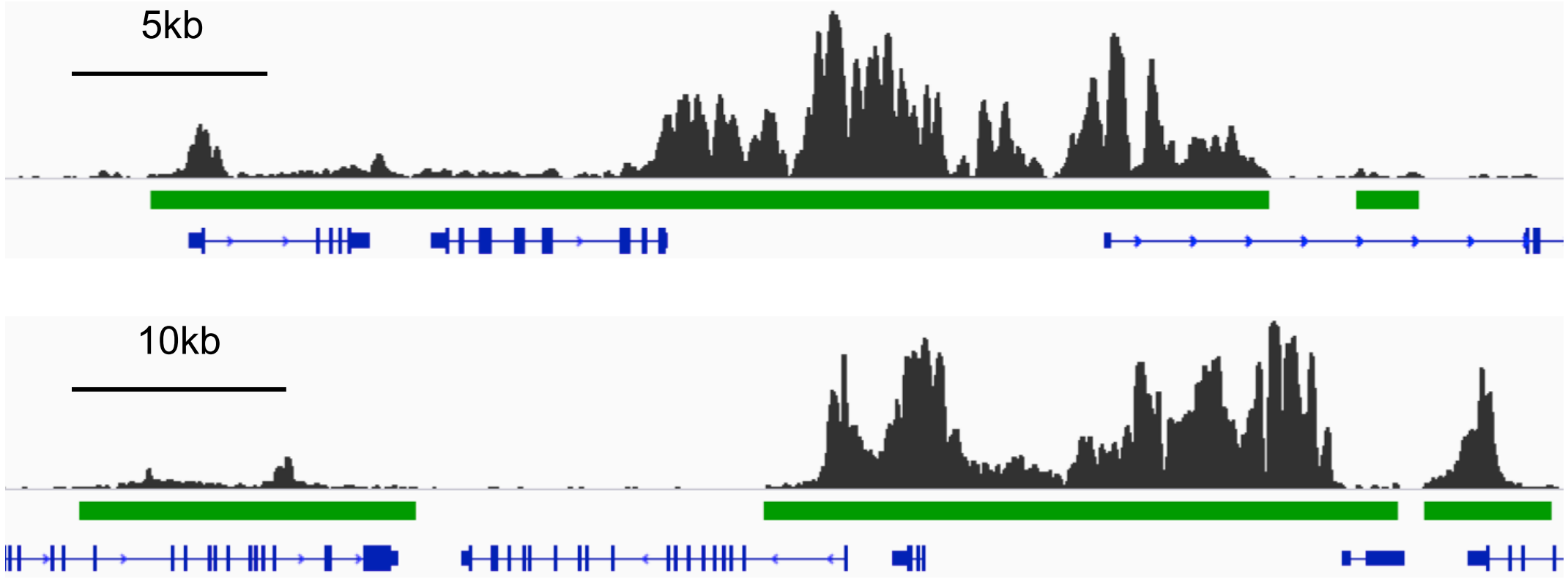
$$\sum_{s \geq s_T} LM(s) \leq e$$

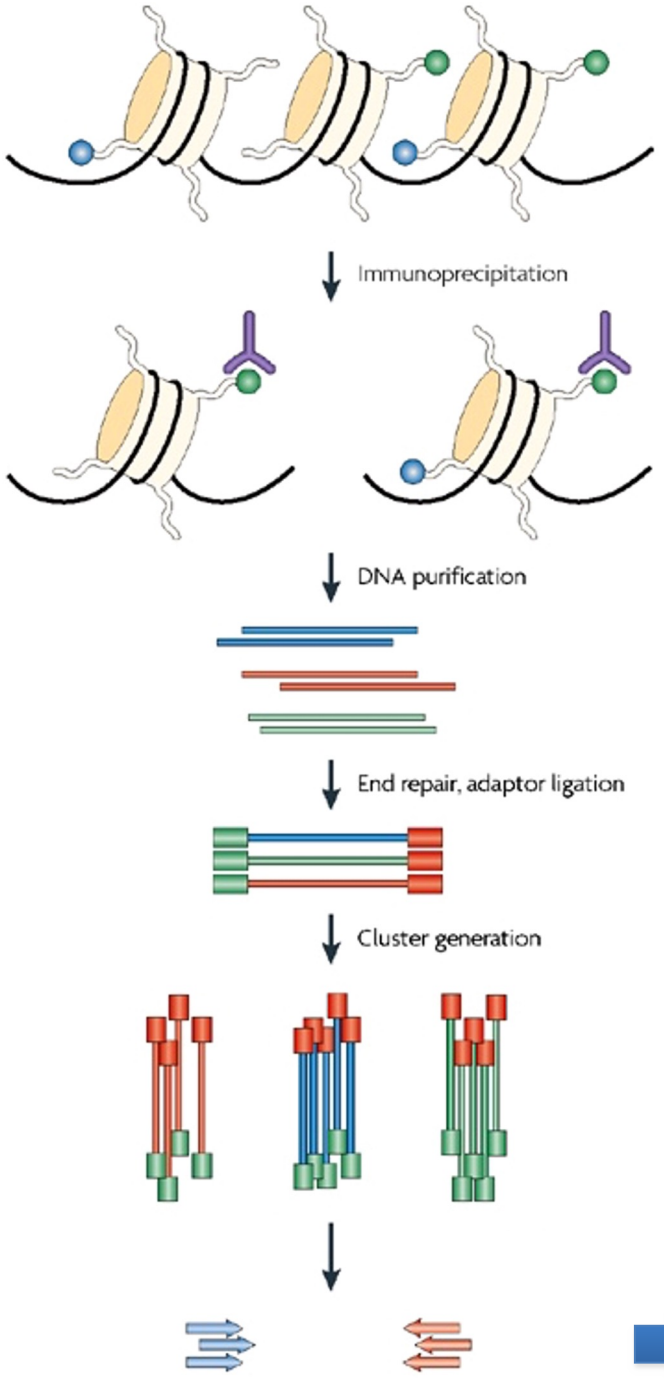
SICER: Significance determinations

- 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





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

Scales of histone mark islands and chromatin domains

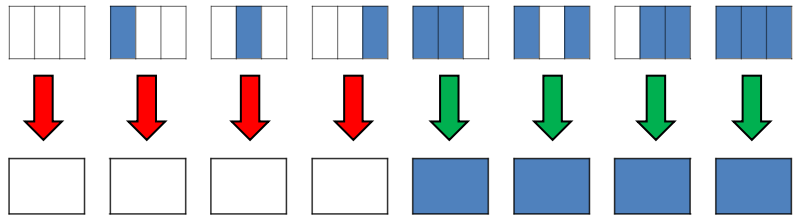
- Narrow: a few nucleosomes, 0.5kb ~ 5kb
  - H3K4me3, H2A.Z, etc.
- Broad: 5kb~100kb
  - Gene loci, chromatin domains, super-enhancers
  - H3K4me1, H3K27ac, H3K36me3, H3K27me3, etc.
- Very broad: >100kb
  - Large chromatin domains, chromatin compartments
  - H3K9me3, H3K27me3

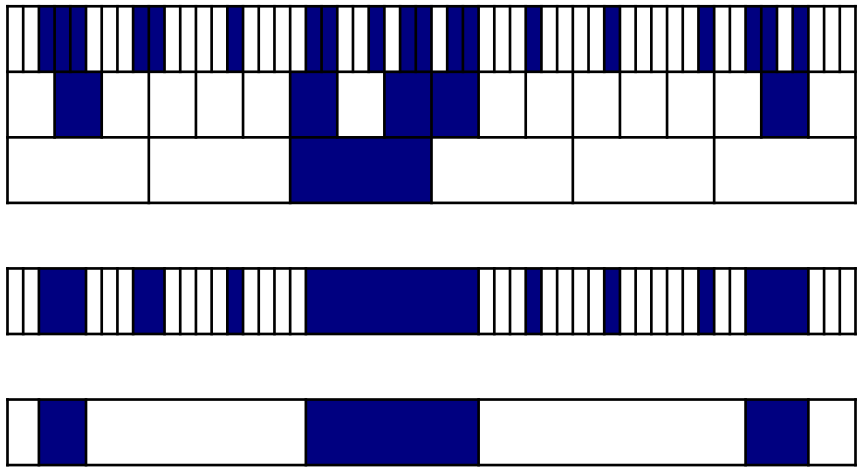
Other approaches for chromatin domains

- ChromHMM: Hidden Markov Models (Ernst & Kellis)
- Recognicer: Coarse-graining (Zang, et al. 2020)

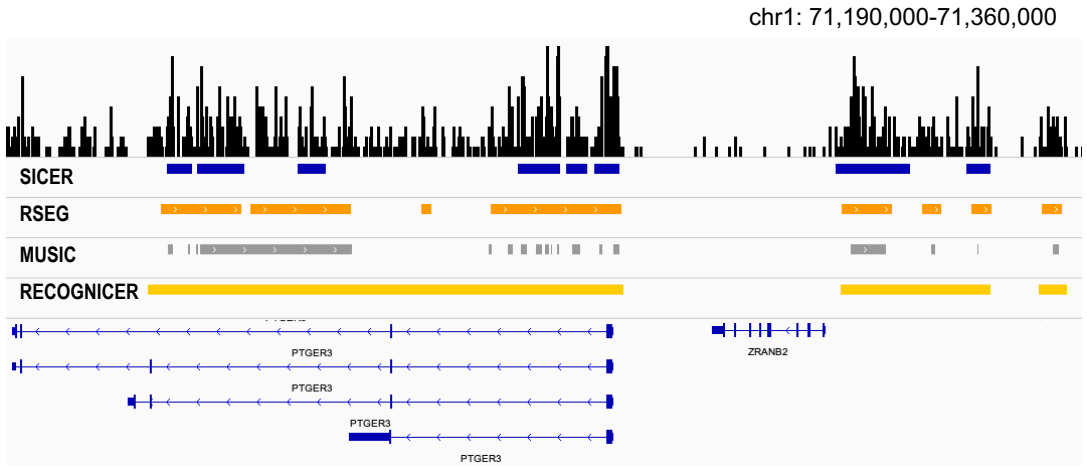
RECOGNICER: Coarse-graining

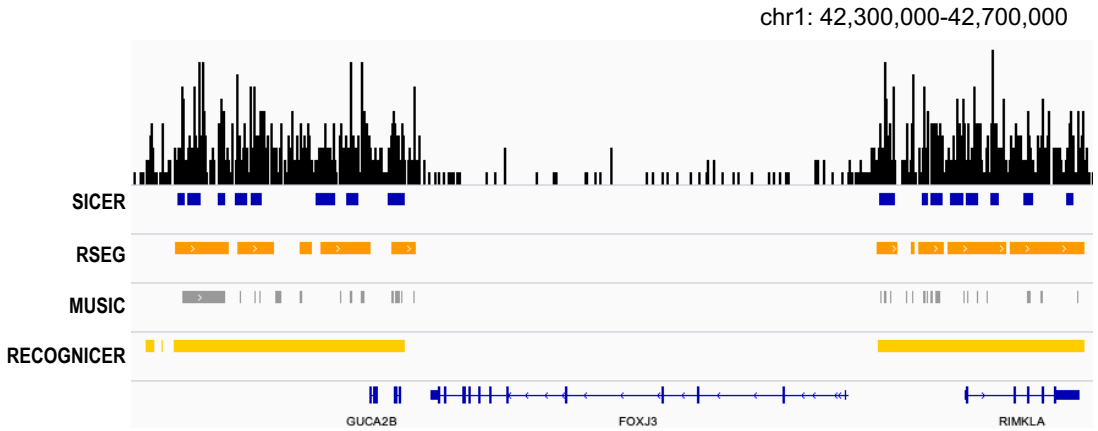
- Block transformation under a majority rule
- Approach:
  - Recursive block transformation
  - Trace back to identify candidate enriched regions
  - Significance determination
  - Scale-free



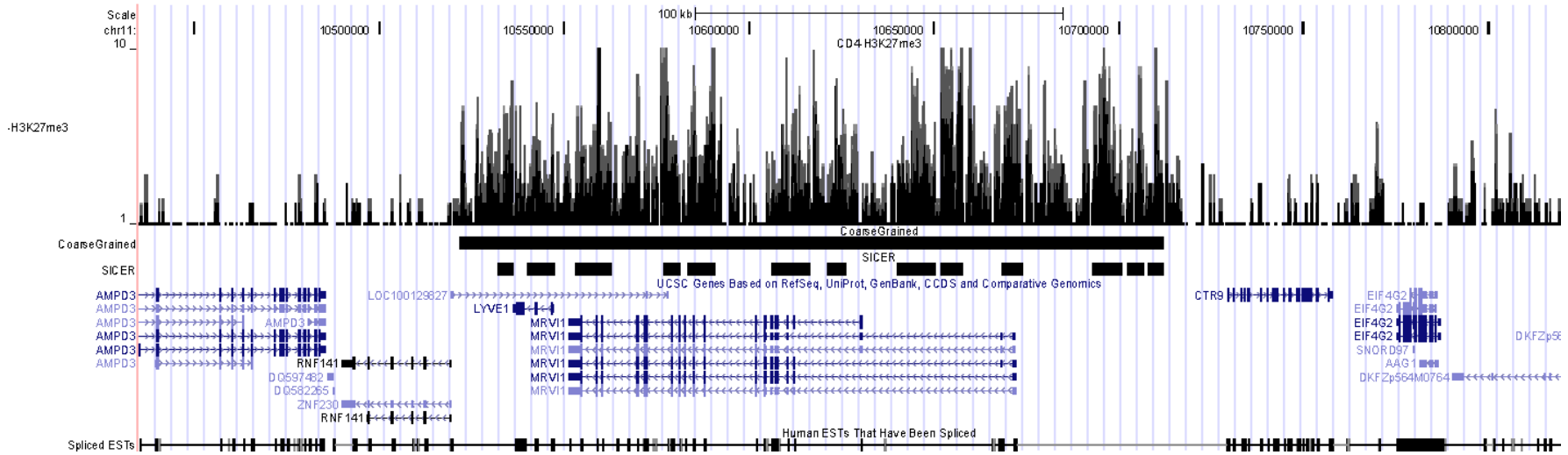


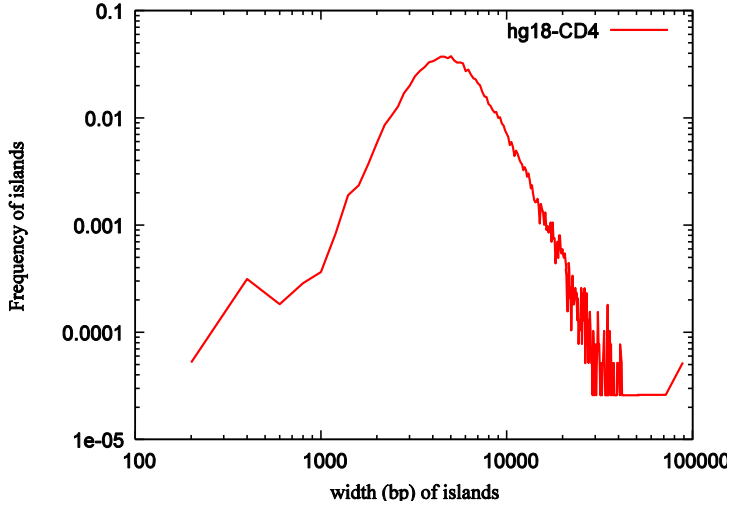
RECOGNICER: Coarse-graining



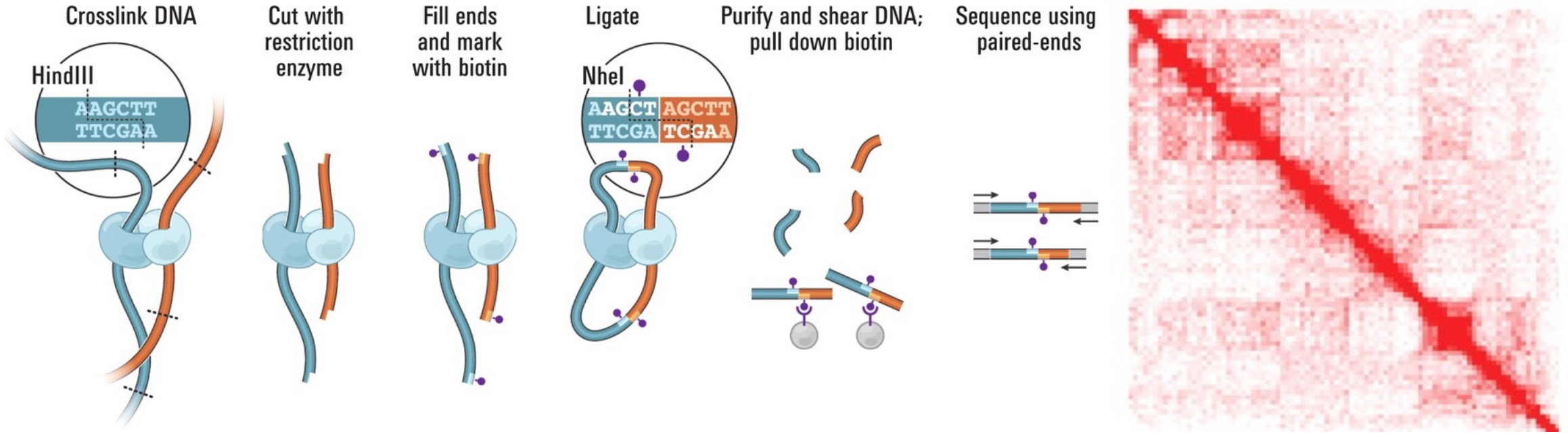


Scale-free property of chromatin domains

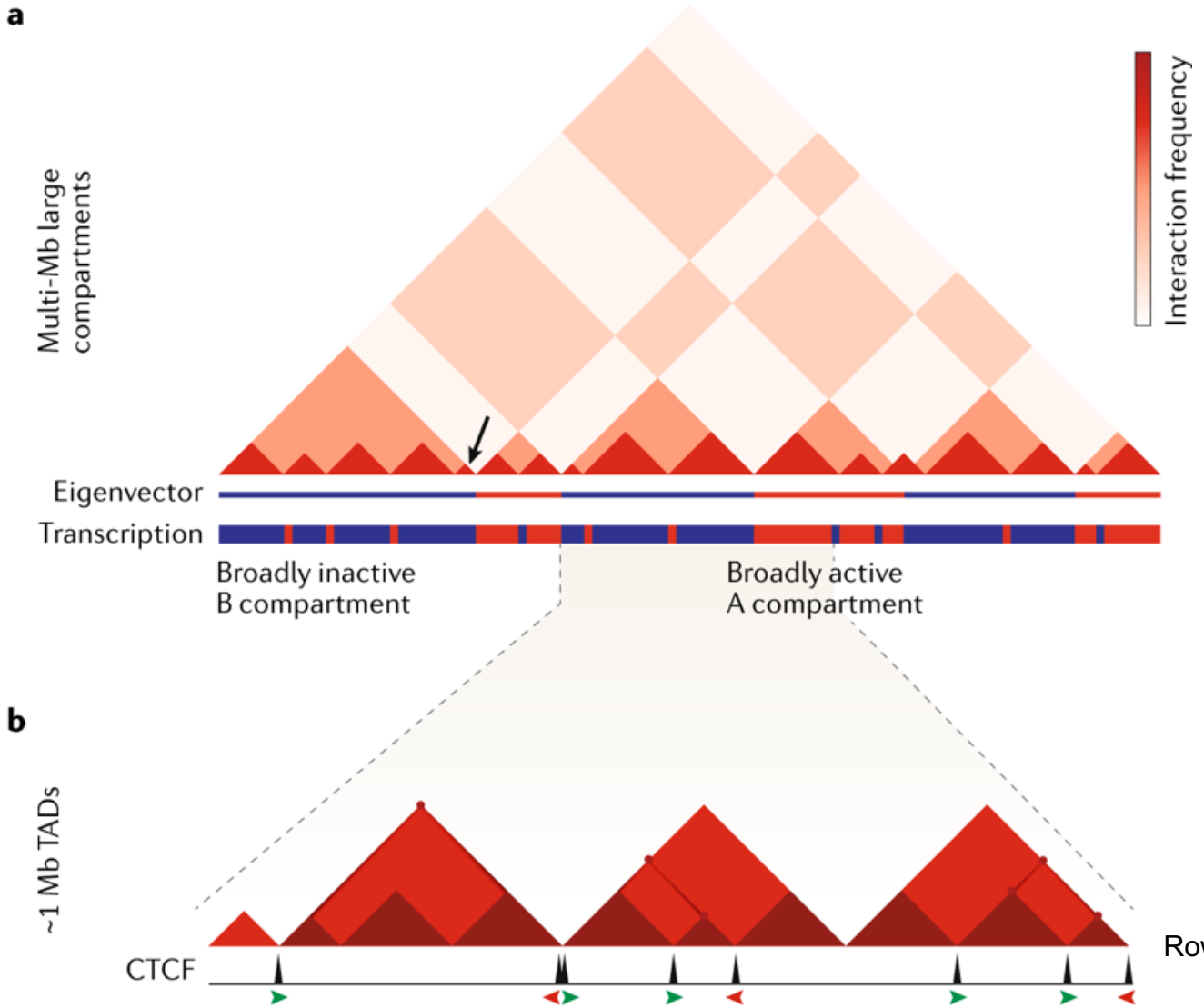




Hi-C


Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

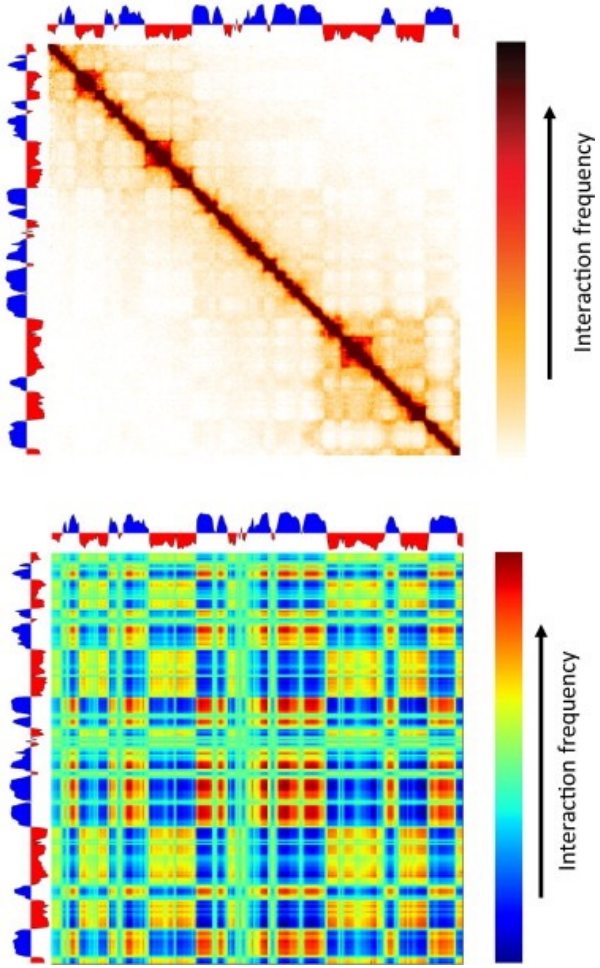
Scale-free property of Hi-C maps



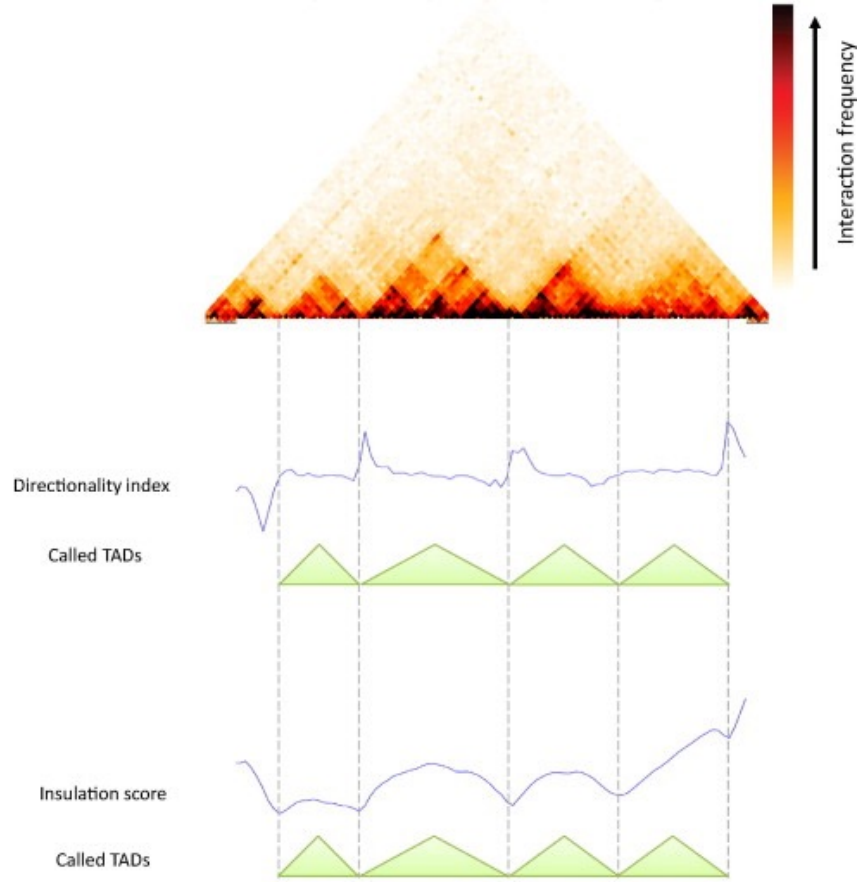
Rowley & Corces, *Nat Rev Genet* (2018)

Hi-C analysis

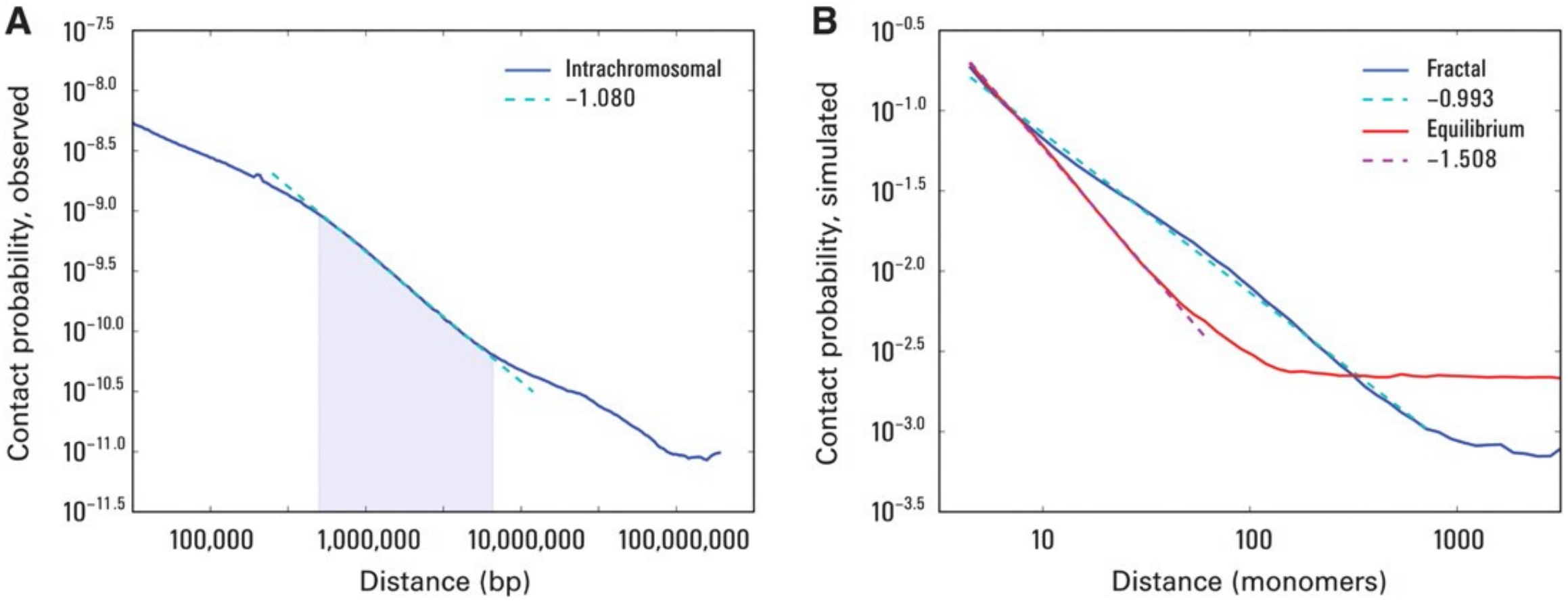
- Chromatin compartments



- Topologically Associating Domains (TADs)

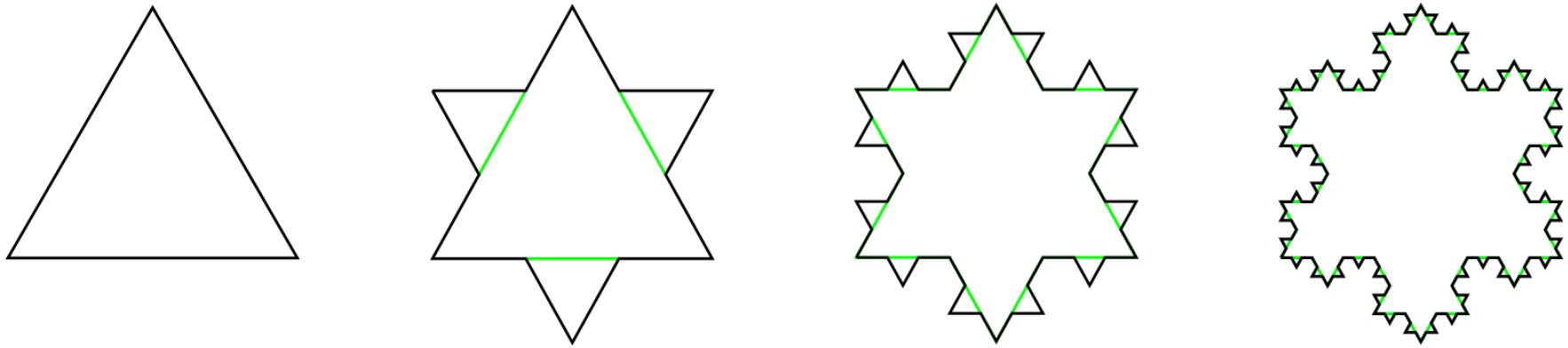


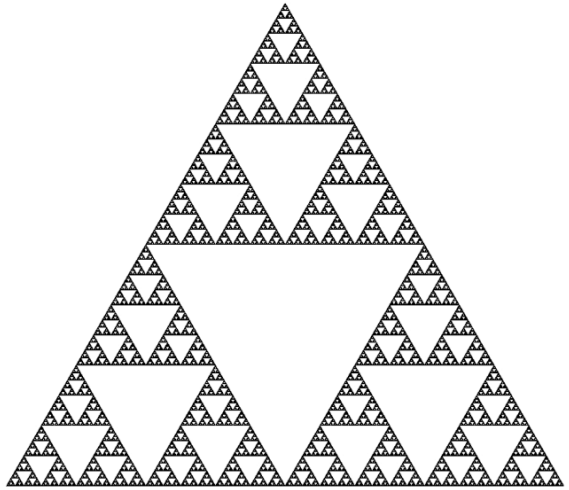
Hi-C: Power-law property of contact probability distribution

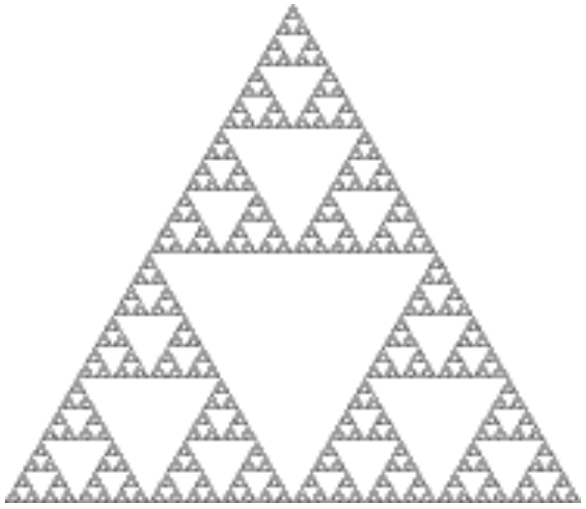


Lieberman-Aiden et al. Science 2009

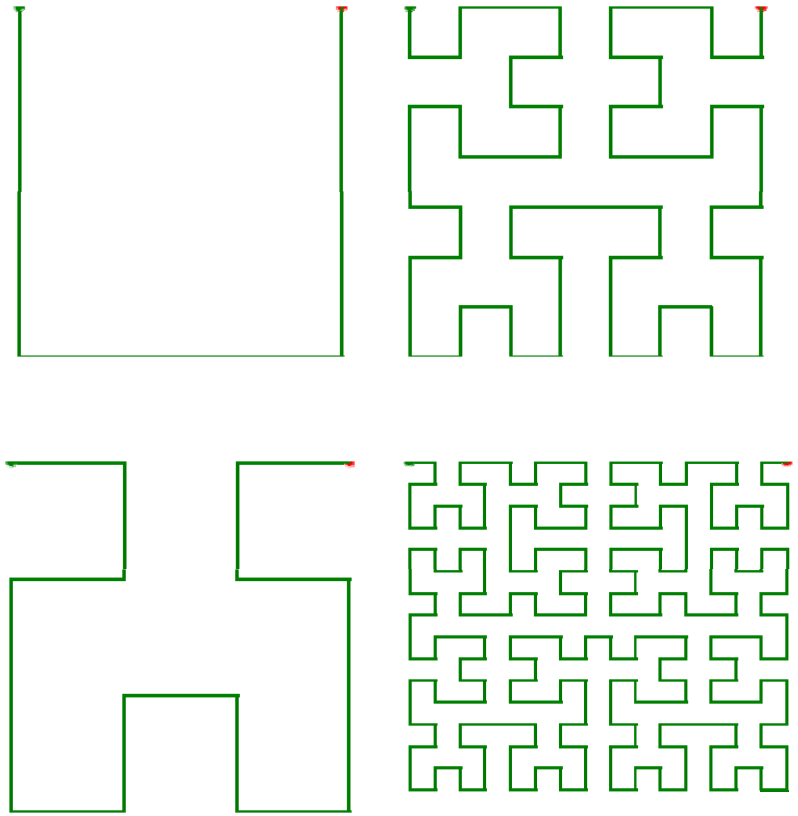
Fractal Structures

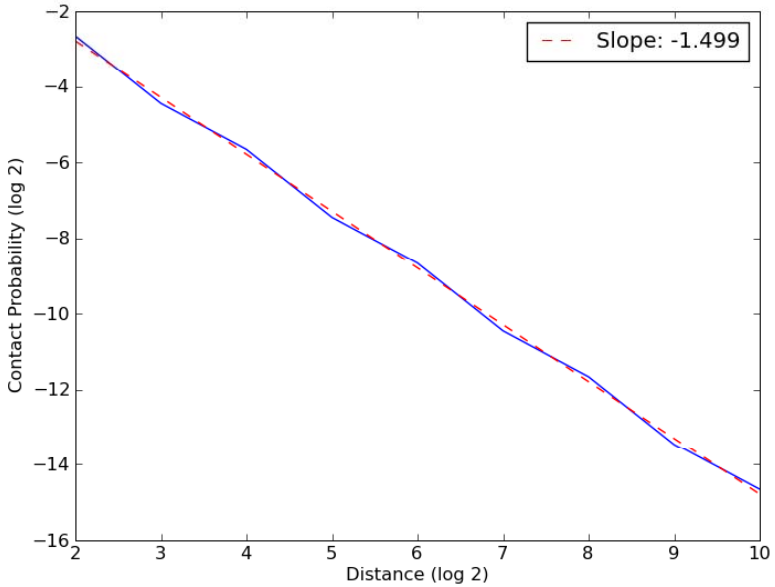






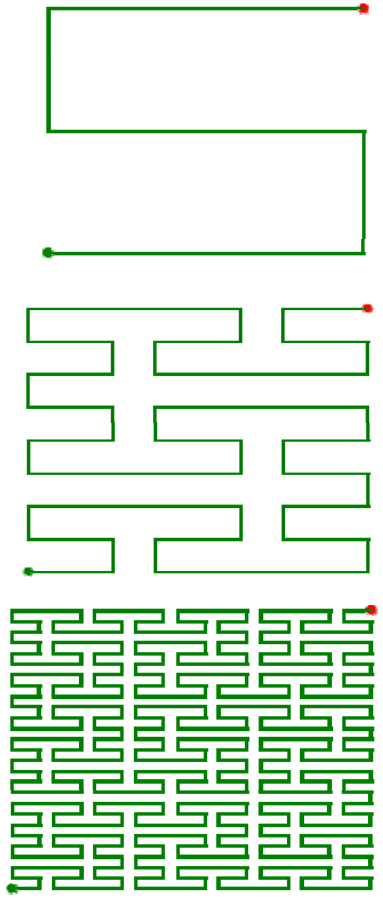
Hilbert Curve

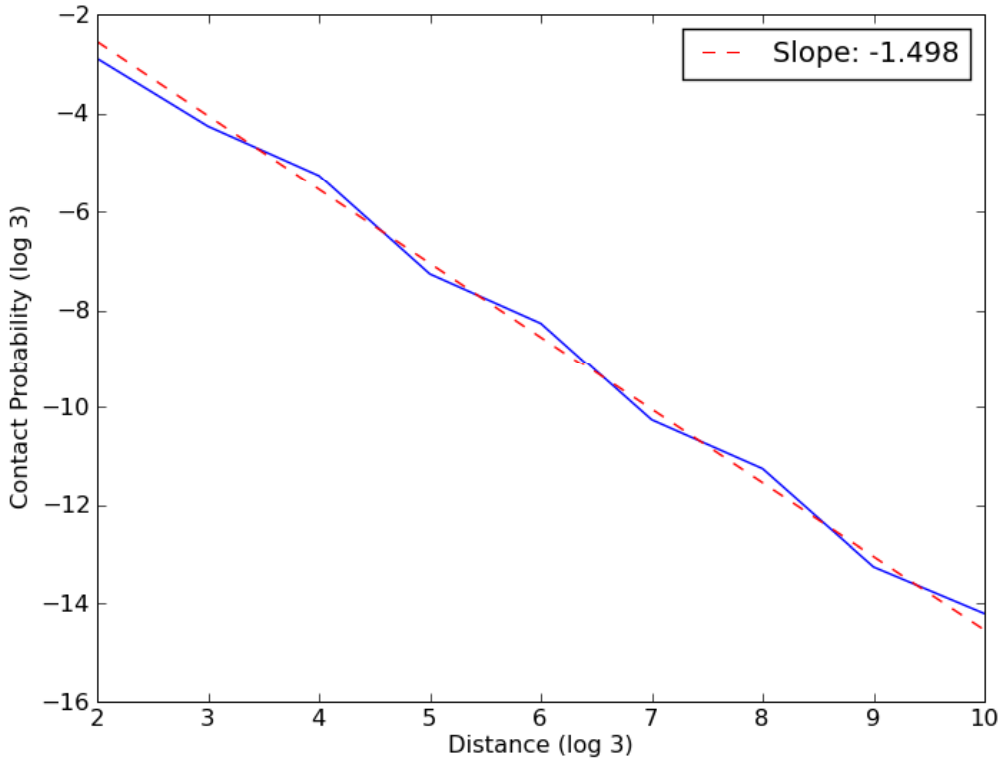




Lieberman-Aiden et al. Science 2009

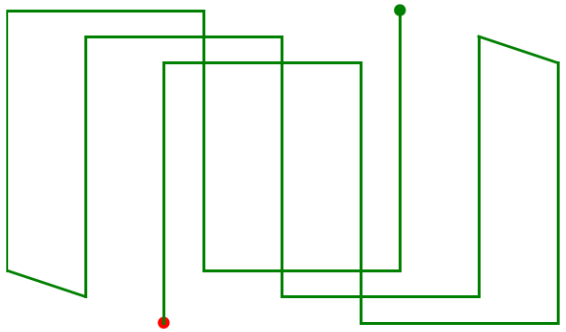
Peano Curve

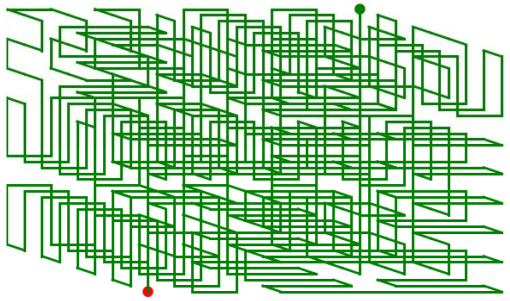


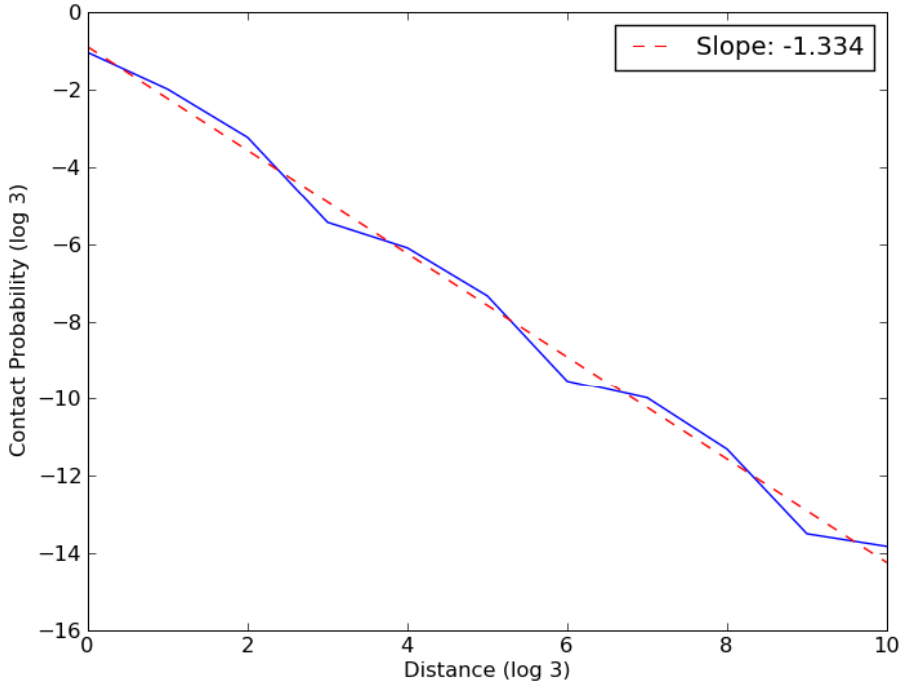


Lieberman-Aiden et al. Science 2009

3D Peano Curve





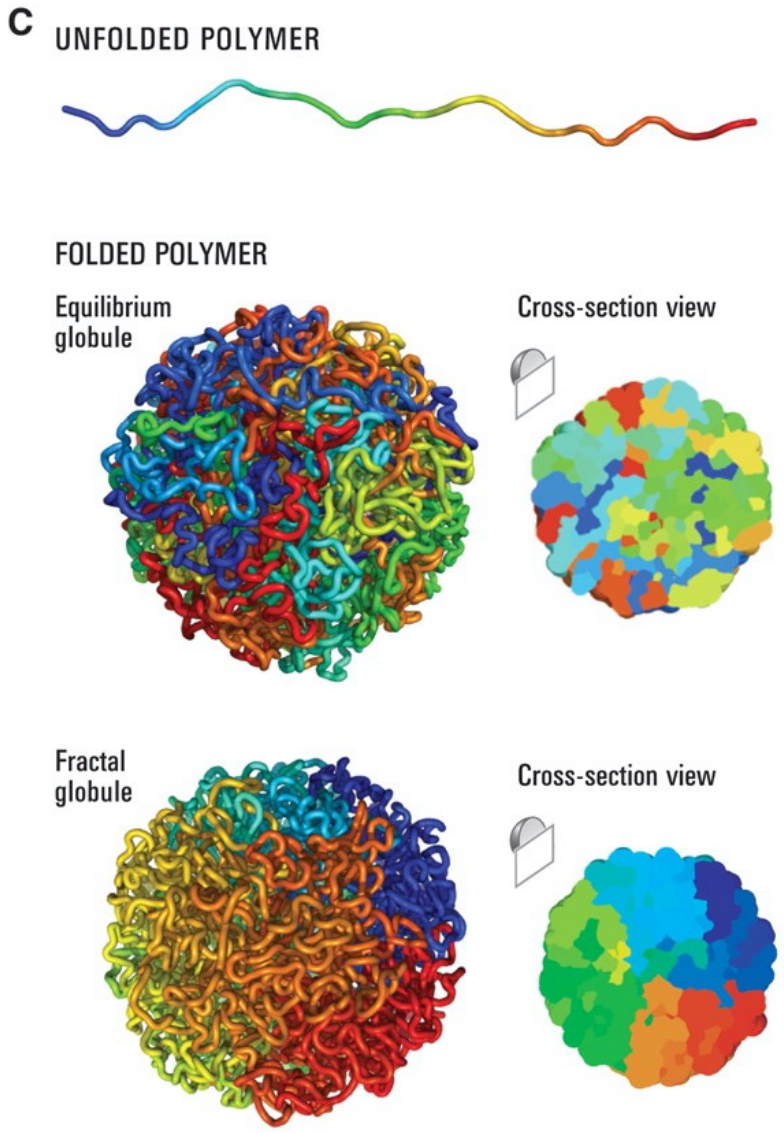


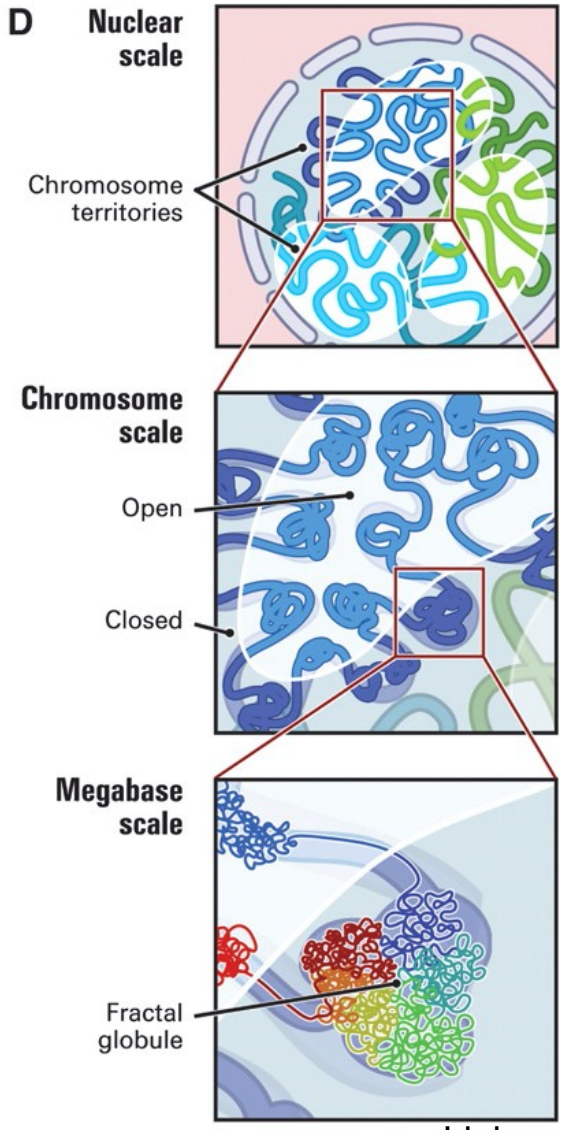
$P_{contact}(x) = kx^{\alpha}$, where α is given by

$\alpha_{smooth} = -(1+\frac{1}{d})$ and $\alpha_{interdigitated} = -1$

Lieberman-Aiden et al. Science 2009

Fractal Structure of Genome Organization



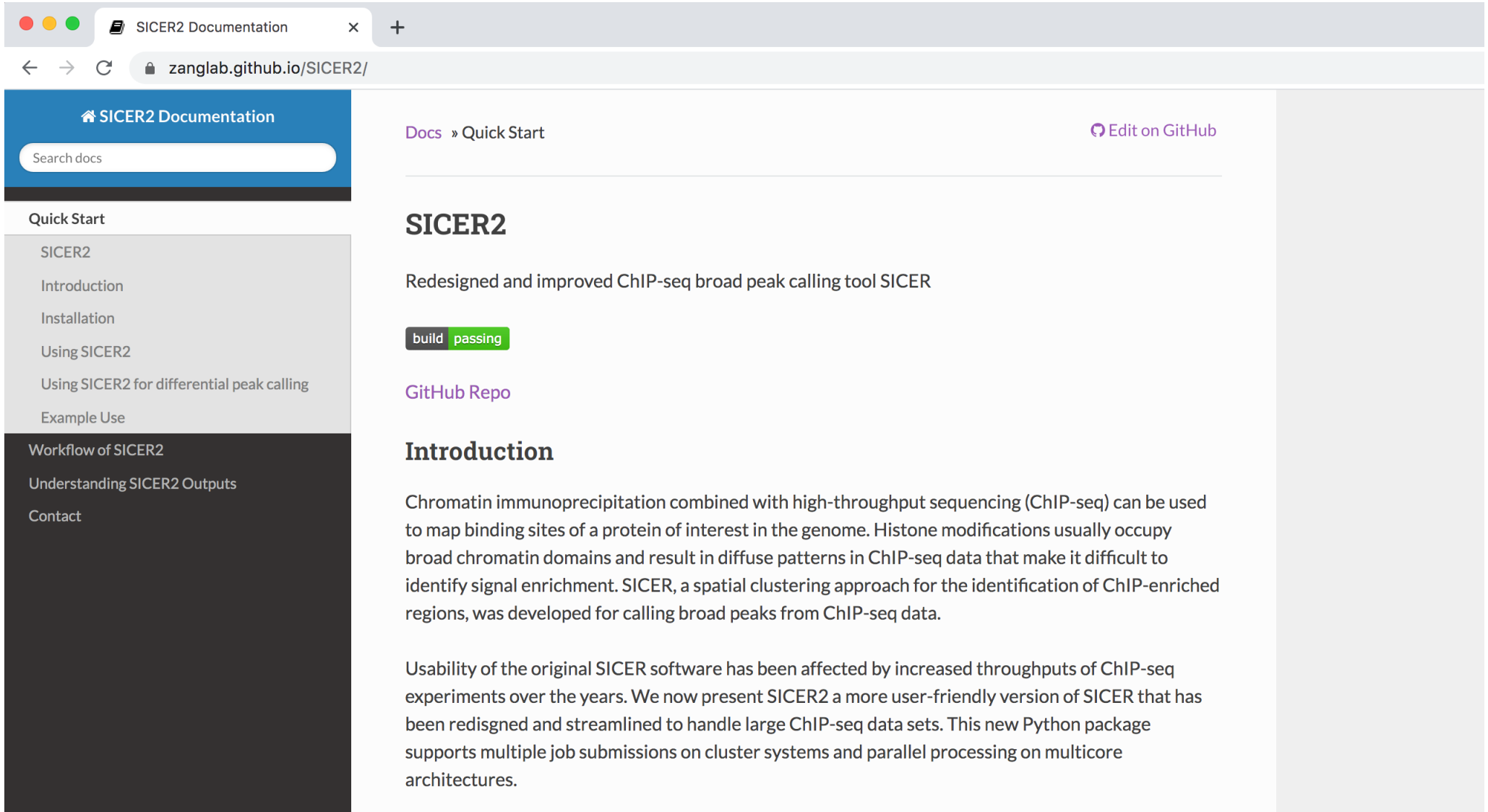


Lieberman-Aiden et al. Science 2009

SICER2

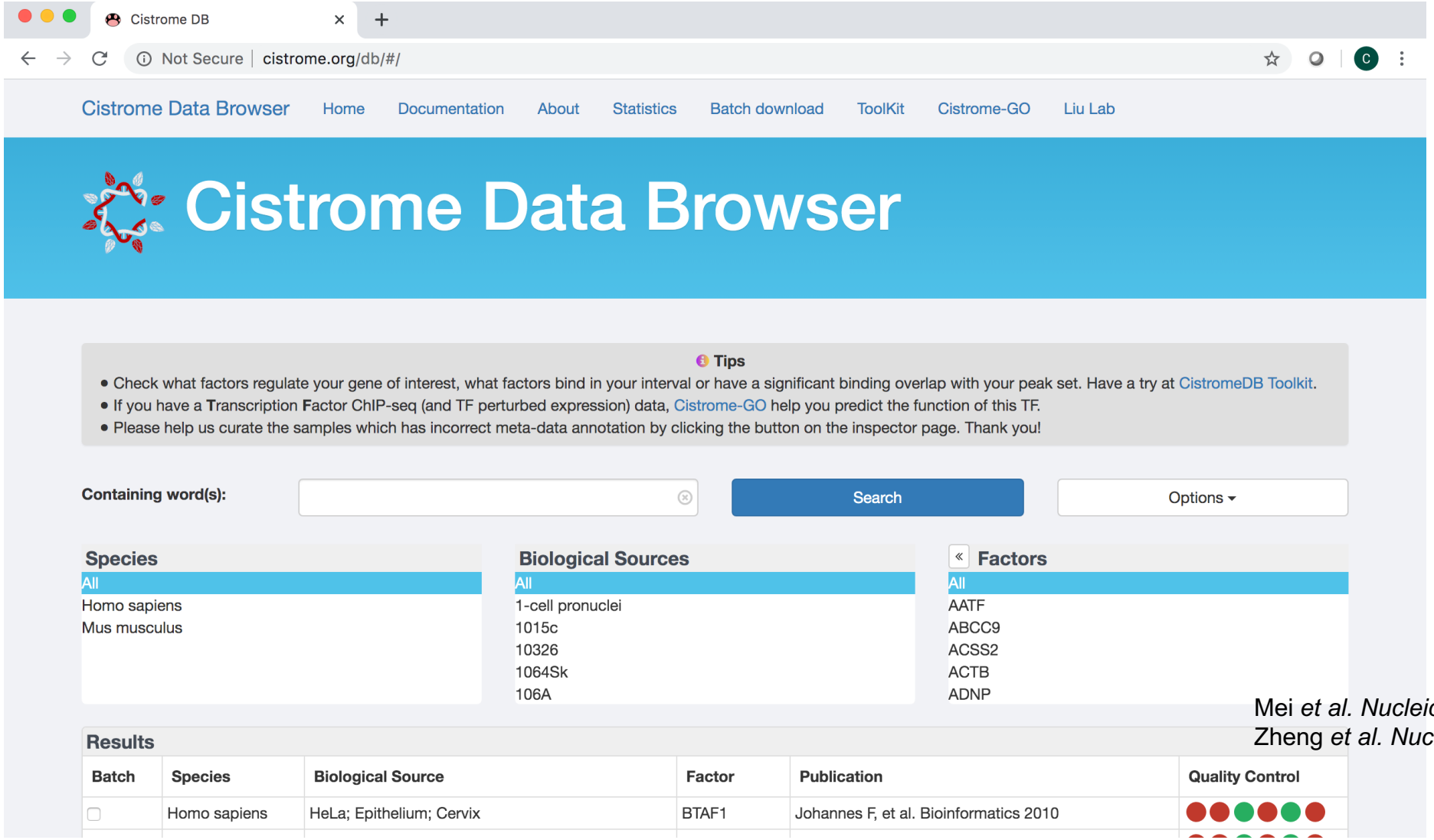
- https://zanglab.github.io/SICER2/

Jin Yong (Jeffrey) Yoo



Cistrome Data Browser

http://cistrome.org/db/



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

Summay

- Epigenome
- ChIP-seq
- ChIP-seq data analysis and signal detection
  - MACS for narrow peaks
  - SICER for broad domains
- Hi-C
- Domain structure of 3D genome organization