

# KSBi-BIML 2023

Bioinformatics & Machine Learning(BIML)  
Workshop for Life Scientists, Data Scientists,  
and Bioinformaticians

생물정보학 & 머신러닝 워크샵 (온라인)

## 3D Epigenome in Gene Regulation

정인경 \_ KAIST



본 강의 자료는 한국생명정보학회가 주관하는 BIML 2023 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

# KSBi-BIML 2023

## Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists, Data Scientists, and Bioinformaticians

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크숍인 BIML-2023에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크숍은 2015년에 시작하여 올해로 9차를 맞이하게 되었습니다. 지난 2년간은 심각한 코로나 대유행으로 인해 아쉽게도 모든 강의가 온라인으로 진행되어 현장 강의에서만 가능한 강의자와 수강생 사이에 다양한 소통의 기회가 없음에 대한 아쉬움이 있었습니다. 다행히도 최근 사회적 거리두기 완화로 현장 강의를 가능해져 올해는 현장 강의를 재개함으로써 온라인과 현장 강의의 장점을 모두 갖춘 프로그램을 구성할 수 있게 되었습니다.

BIML 워크숍은 전통적으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되었습니다. 올해 AI 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 심층학습(Deep learning) 기법들에 대한 현장 강의를 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체 분석, 신약개발에 대한 이론과 실습 강의를 함께 제공할 예정입니다. 또한 싱글셀오믹스 분석과 메타유전체분석 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다. 이외에 다양한 생명정보학 분야에 대하여 30개 이상의 온라인 강좌가 개설되어 제공되며 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다. 특히 BIML은 각 분야 국내 최고 전문가들의 강의로 구성되어 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의를 될 것입니다.

이번 BIML-2023을 준비하기까지 너무나 많은 수고를 해주신 BIML-2023 운영위원회의 남진우, 우현구, 백대현, 정성원, 정인경, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 연사분께 깊은 감사를 드립니다.

2023년 2월

한국생명정보학회장 이 인 석

# 3D Epigenome in Gene Regulation

후성유전이란 유전자 조절의 핵심 기전으로 발생 및 분화 그리고 다양한 질환의 기전 연구에 있어 중요하다. 최근 후성유전적 유전자 조절 기전 연구는 염색질 3차구조 관점에서 이루어지고 있다. 염색질 3차구조란 핵 안에 3차원으로 배열된 게놈의 구조를 의미한다. 최근 연구 결과에 따르면 염색질 3차 구조는 무작위적 배열보다는 TAD (Topologically Associating Domain) 또는 Loop domain을 기본 단위로 여러 계층으로 구성되어 있으며, 이러한 구조적 제약에 의해 DNA 서열상 멀리 떨어진 인핸서, 프로모터 등 여러 전사 조절 인자들은 3차원 공간상에 인접할 수 있게 되어 전사 조절의 핵심 원리로 제시되고 있다.

본 강의에서는 후성유전체 및 염색질 3차구조를 중심으로 관련 이론 및 기본 데이터 분석을 실습과 함께 숙지하고자 한다. 간략하게 후성유전학에 대한 소개 이후, 염색질 3차구조에 대한 전반적인 소개를 하고, 실습시간에는 최근 본 연구팀이 개발한 3DIV 웹기반 염색질 3차구조 데이터 분석법과 covNorm기반 R을 활용한 Hi-C 데이터 기본 데이터 분석 방법을 익히려 한다.

강의는 다음의 내용을 포함한다:

- 후성유전체 및 ChIP-seq 개요
- 염색질 3차구조 개요
- 3DIV 기반 Hi-C 데이터 분석 실습
- covNorm 을 활용한 Hi-C 데이터 분석 실습

\* 교육생준비물:

노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

\* 강의 난이도: 중급

\* 강의: 정인경 교수 (한국과학기술원 생명과학과)

## Curriculum Vitae

**Speaker Name: Inkyung Jung, Ph.D.**



### ► Personal Info

Name Inkyung Jung  
Title Associate Professor  
Affiliation KAIST

### ► Contact Information

Address Department of biological sciences, KAIST  
Email [ijung@kaist.ac.kr](mailto:ijung@kaist.ac.kr)  
Phone Number 042-350-7314

---

### Research Interest

Epigenetic gene regulation, 3D chromatin structure

### Educational Experience

2006-2011 Ph.D. KAIST / Bio and Brain Engineering  
2002-2006 B.S. KAIST / Biosystems

### Professional Experience

2016-present Assistant Professor, Associate Professor, Department of Biological Sciences, KAIST  
2012-2016 Postdoctoral fellow, Ludwig Institute for Cancer Research  
2011-2012 Postdoctoral fellow, KAIST

### Selected Publications (5 maximum)

1. Kim K\*, Jang I\*, Kim M\*, Choi J, Kim MS, Lee B#, Jung I# (2020) 3DIV Update for 2021: a comprehensive resource of 3D genome and 3D cancer genome. *Nucleic Acids Res.* Jan 49(D1):38-46
2. Lee JS\*, Park S\*, Jeong HW\*, Ahn JY\*, Choi SJ, Lee H, Choi B, Nam SK, Kwon JS, Jeong SJ, Lee HK, Park SH, Park SH, Choi JY#, Kim SH#, Jung I#, Shin EC# (2020) Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Science Immunol.* Jul 4(49)
3. Jung I\*#, Schmitt A\*, Diao Y\*, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S, Chiang Z, Kim C, Masliah E, Barr CL, Li B, Kuan S, Kim D, Ren B#. (2019) A Compendium of Promoter-Centered Long-Range Chromatin Interactions in the Human Genome. *Nat Genet.* Oct 51(10):1442-1449
4. Ryu J\*, Kim H\*, Yang D, Lee AJ, Jung I. (2019) A new class of constitutively active super-enhancers is associated with fast recovery of 3D chromatin loops. *BMC Bioinformatics.* Mar 20:127
5. Yang D\*, Jang I\*, Choi J, Kim MS, Lee AJ, Kim H, Eom J, Kim D#, Jung I#, Lee B# (2018) 3DIV: A 3D-genome Interaction Viewer and database. *Nucleic Acids Res.* Jan 46(D1):52-67

# KSBi-BIML 2022

## 3D Epigenome in Gene Regulation (후성유전체 및 ChIP-seq 개요)

정인경(KAIST)

### Contents

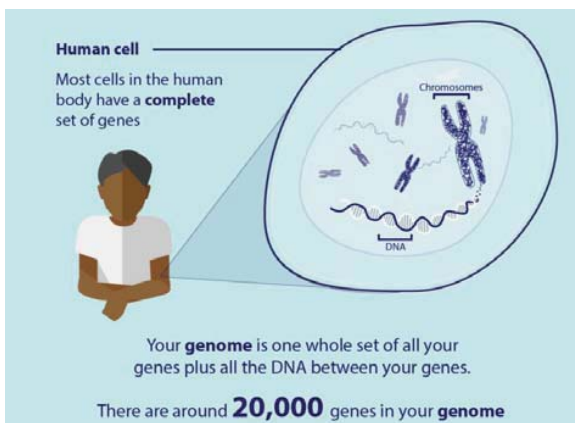
1. 후성유전체 및 ChIP-seq 개요
2. 염색질 3차구조 개요
3. 3DIV 기반 염색질 3차구조 및 유전자 조절 통합 분석 실습
4. 염색질 3차구조 데이터 분석 실습

**Proteins are essential molecules in cellular functions**

**Genes encode proteins**

**Then, how does DNA instruct genes to encode proteins?**

## What is the genome?



**Genome = Gene + Chromosome**

**Human Genome =** whole set of human genes + DNA between genes

Human Genome  
-3.2 X 10<sup>9</sup>bp (A/T/G/C)  
-22 paired chromosomes + sex chromosomes (X, Y)

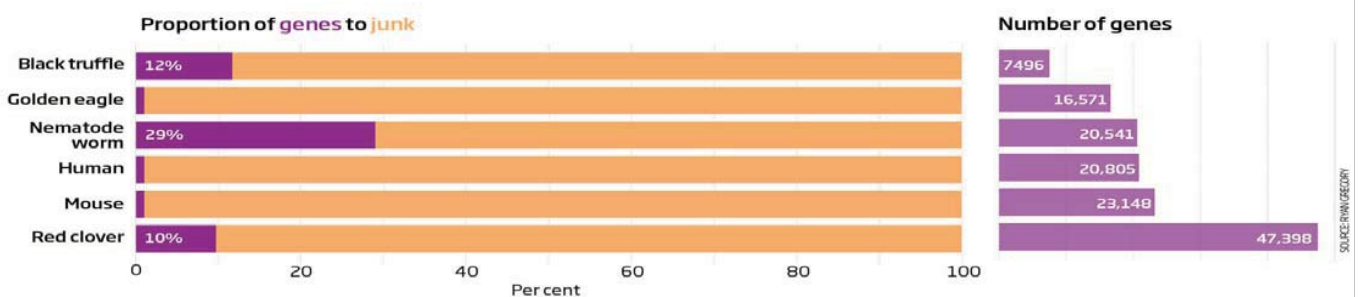
## How can we identify genes?

- **A priori method:** seek to recognize sequence patterns within expressed genes and the regions flanking them
- **A recent method:** Been there, seen that method:
  1. Recognize regions corresponding to previously known genes, from the similarity of their translated amino acid sequences to known proteins in another species
  2. Matching expressed sequence tags (ESTs)

“Need to know the genome sequence”

## Human Genome project (1990-2003)

Before sequencing human genome scientists estimated the number of genes in human genome as 1,000,000



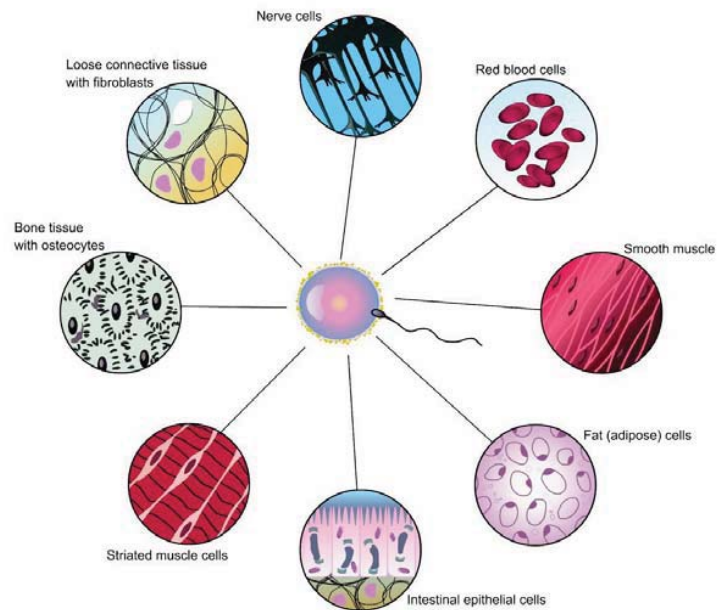
Protein-coding genes occupy a small fraction of the human genome  
– no more than about 2-3% -



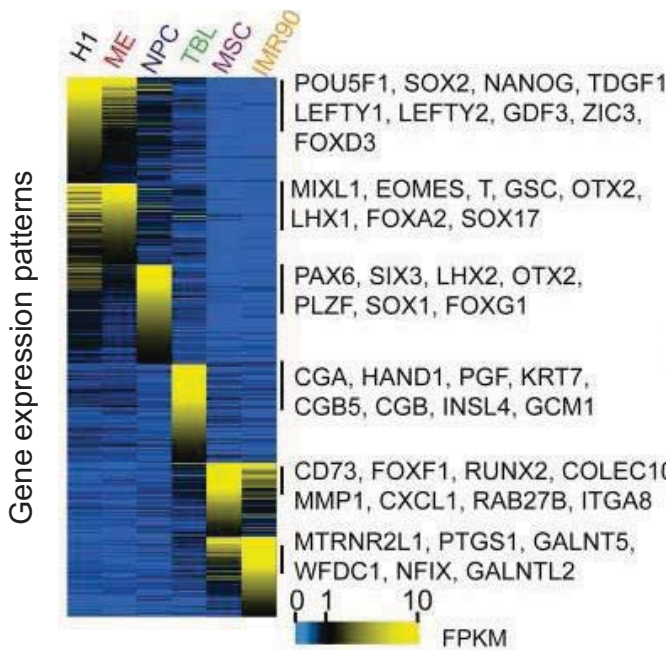
# One genome, but many different functions



30 trillion cells



# Cell-type specific transcriptome determines cell-type specific function



Cell type specific gene expression patterns can characterize cellular identity and define cell type specific biological functions

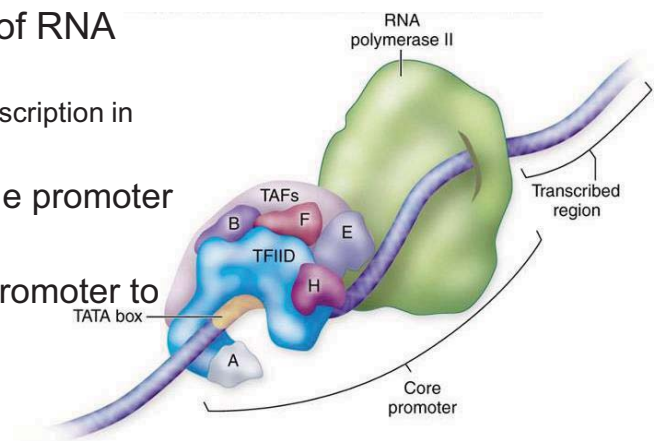
Gene structures are same for all cell types

Then, how are genes regulated cell-type specific manner?

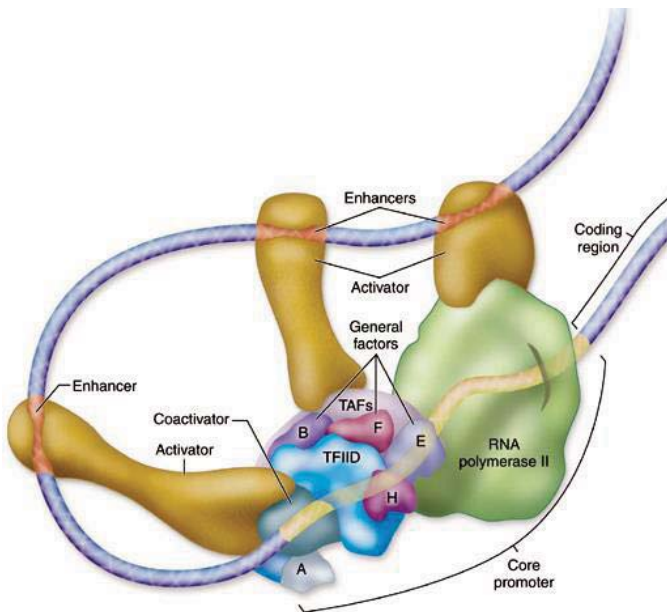
Can non-coding sequences facilitate cell-type specific transcriptome with limited number of genes?

## Promoter and epigenetic gene regulation

- Controlling the expression of eukaryotic genes requires **transcription factors**.
  - General transcription factors** are required for transcription initiation
    - required for proper binding of RNA polymerase to the DNA
  - Specific transcription factors** increase transcription in certain cells or in response to signals
- General transcription factors bind to the promoter region of the gene
- RNA polymerase II then binds to the promoter to begin transcription at the start site



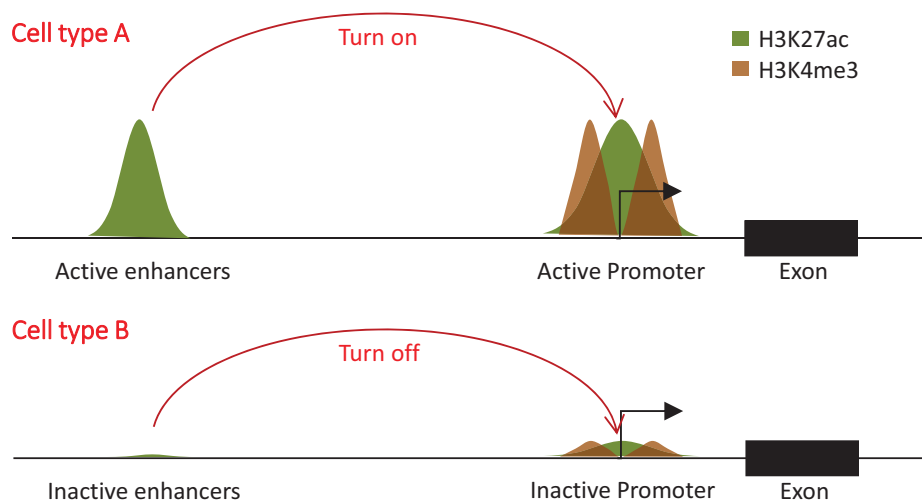
## Enhancers in gene regulation (or distal *cis*-regulatory elements)



- **Enhancers** are DNA sequences to which **specific transcription factors (activators)** bind to increase the rate of transcription.
- Enhancers are generally *cis*-acting, can be located up to 1 Mbp (1,000,000 bp) away from the gene, can be upstream or downstream from the start site, and either in the forward or backward direction
- There are hundreds of thousands of enhancers in the human genome

## Cell-type specific gene regulatory elements

Activities at *cis*-regulatory elements can determine gene expression



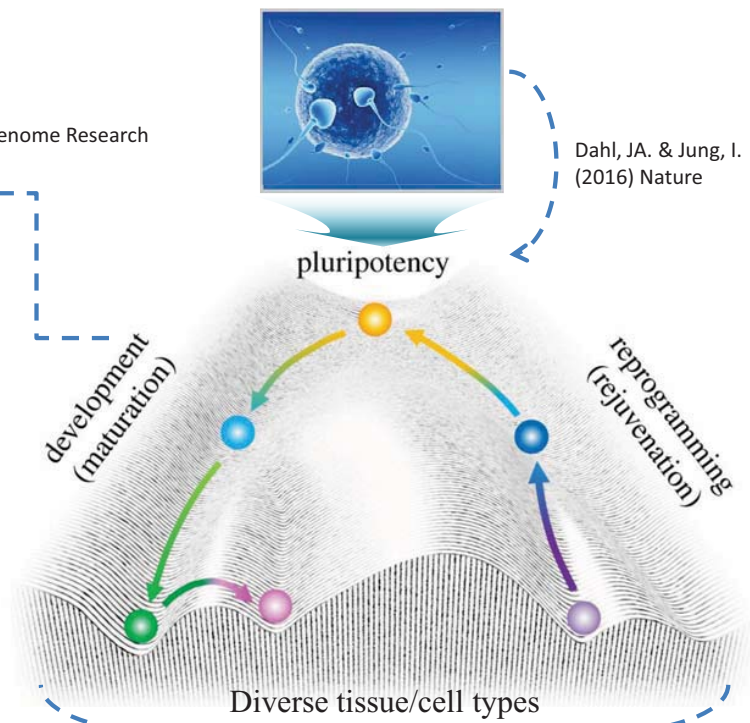
# Epigenetic gene regulation facilitates various cellular identity

Dixon, J. & Jung, I. (2015) Nature  
 Jung, I. Kim, SK, Kim, M. (2012) Genome Research  
 Kim, SK. & Jung, I. (2012) JBC

Dahl, JA. & Jung, I. (2016) Nature



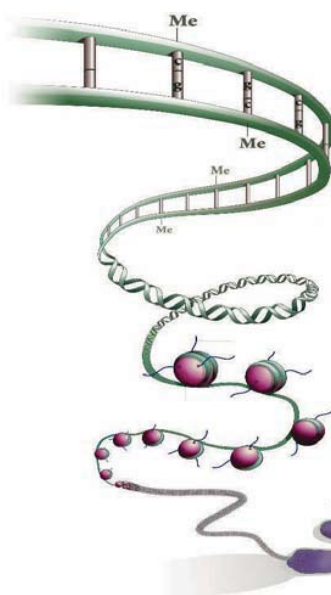
Conrad Waddington  
 (1905-1975)



Leung, D. & Jung, I. (2015) Nature  
 Jung, I. (2019) Nature Genetics

Waddington's epigenetic landscape (Evolution, 1956)

# Three Major Players in Shaping Epigenome

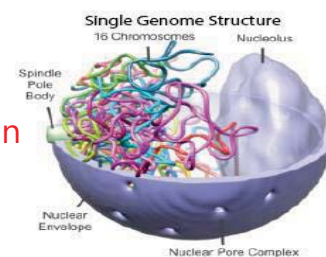


(1) DNA methylation

(6) 3D genome organization

(2) Chromatin modifications

- (3) Small RNAs
- (4) Nucleosome positioning
- (5) Chromatin Remodeling etc.



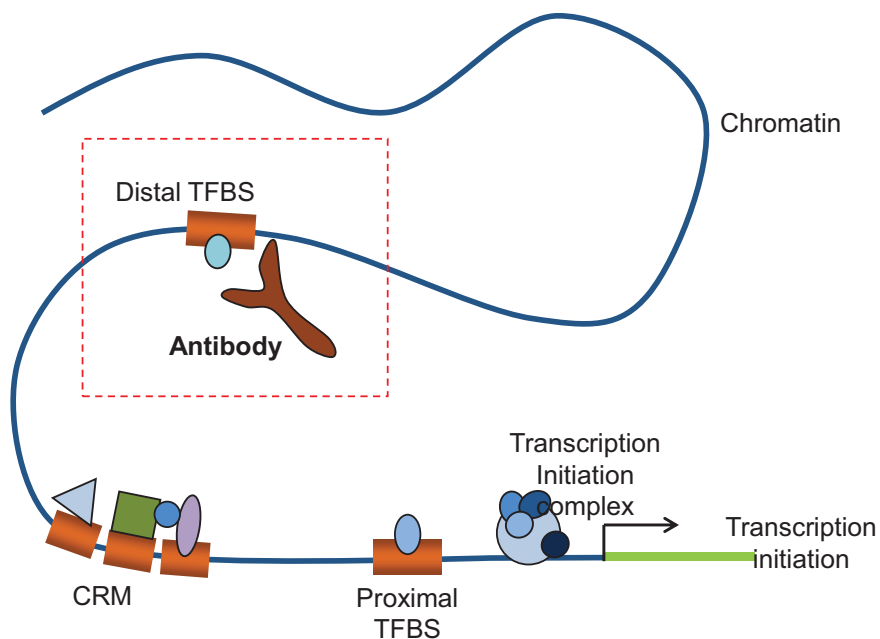
Nature 2006, 441, 143

Then, how can we characterize chromatin modifications?

## ChIP is a powerful tool to characterize chromatin modifications

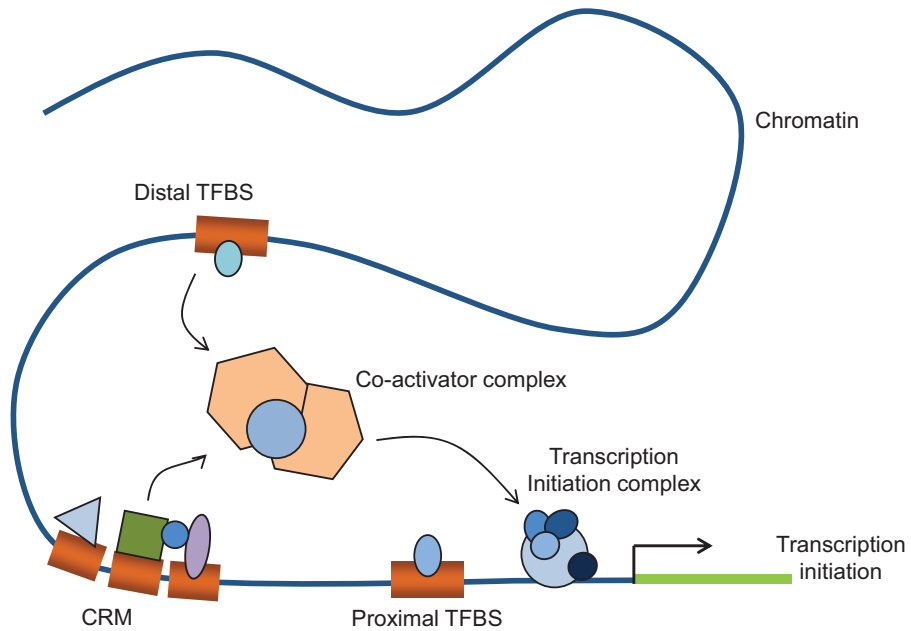
### ChIP (Chromatin immunoprecipitation)

:The method to use antibodies to pull down fragments of DNA that are bound by a protein (including histone)



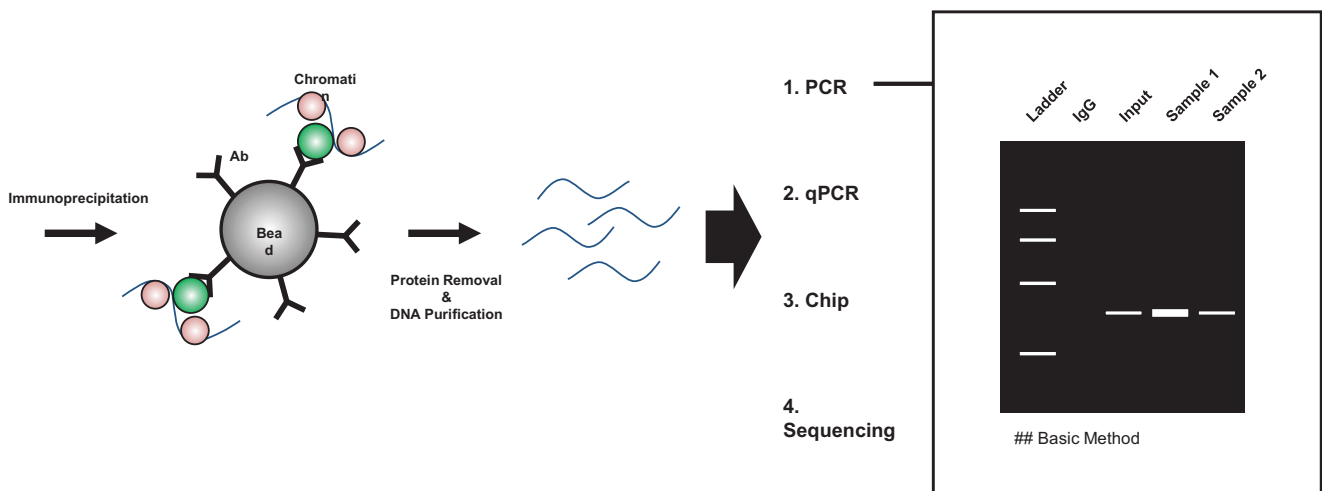
## Why do we need to perform ChIP?

- To understand **transcriptional regulation (or gene regulation)**
- Transcriptional regulation is largely controlled by protein-DNA interactions



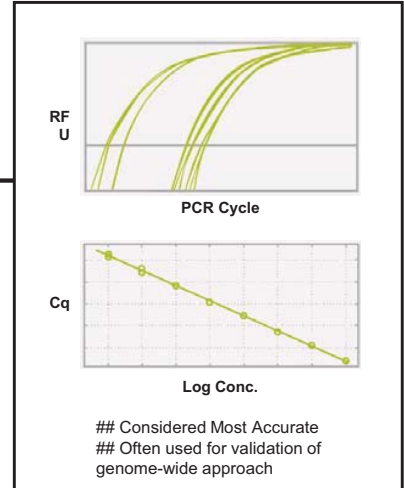
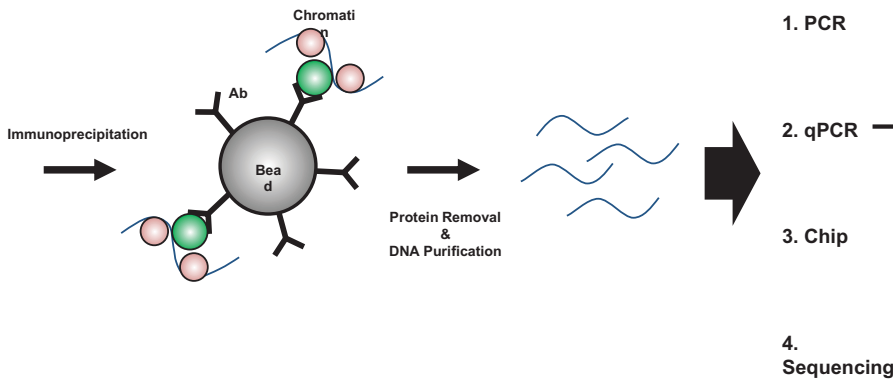
## What are the applications of ChIP?

### ChIP Applications (Chromatin immunoprecipitation)



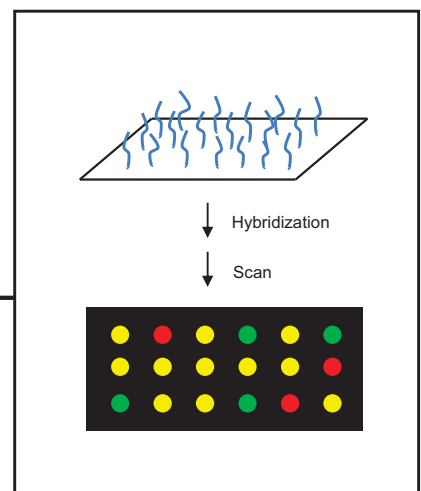
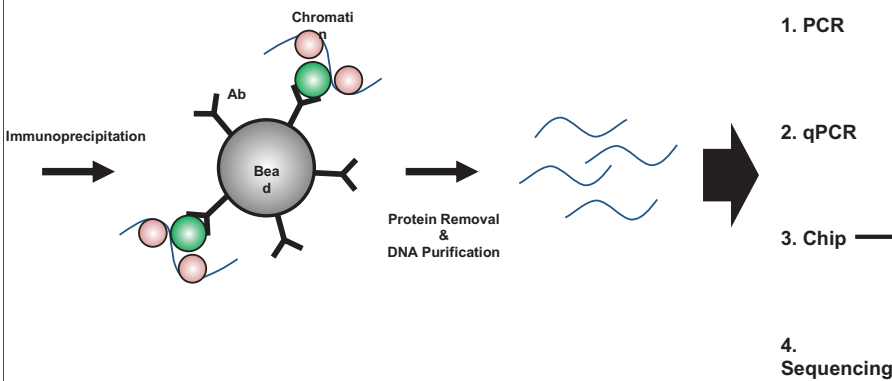
# What are the applications of ChIP?

## ChIP Applications (Chromatin immunoprecipitation)



# What are the applications of ChIP?


## ChIP Applications (Chromatin immunoprecipitation)



REPORTS

### Genome-Wide Location and Function of DNA Binding Proteins

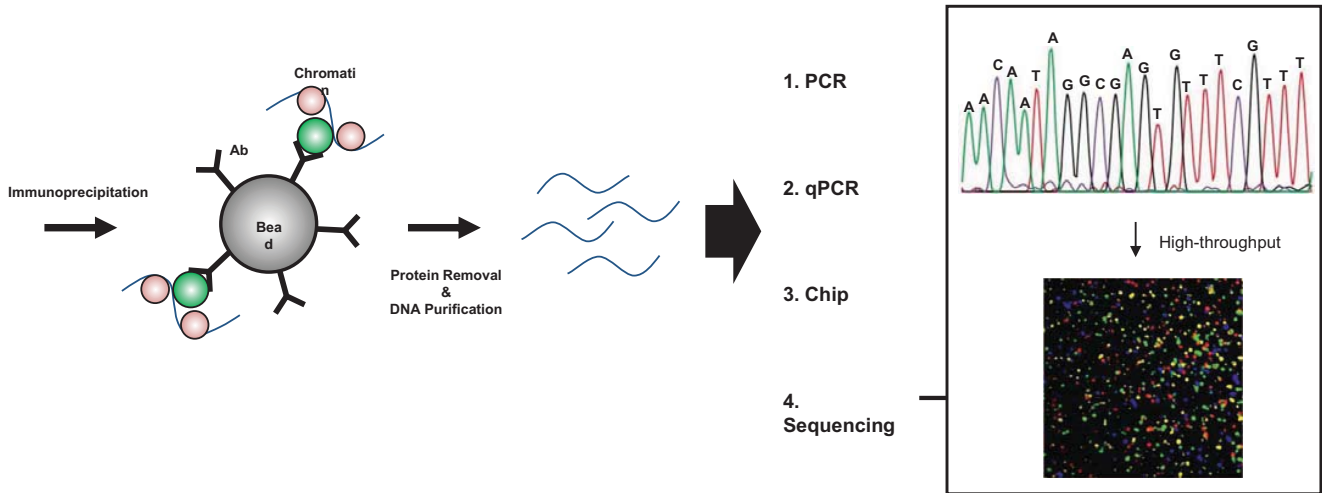
Bing Ren,<sup>1\*</sup> François Robert,<sup>1\*</sup> John J. Wyrick,<sup>1,2\*</sup> Oscar Aparicio,<sup>2,4</sup> Ezra G. Jennings,<sup>1,2</sup> Itamar Simon,<sup>1</sup> Julia Zeitlinger,<sup>1</sup> Jörg Schreiber,<sup>1</sup> Nancy Hannett,<sup>1</sup> Elenita Kanin,<sup>1</sup> Thomas L. Volkert,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



Dr. Bing Ren (Science, 2000)

# What are the applications of ChIP?

## ChIP Applications (Chromatin immunoprecipitation)



# Development of ChIP-seq

In 2007, there was a race to develop this apparently obviously technology. At least three groups worked on the development of genome-wide ChIP-seq assay.

ARTICLES

**Genome-wide maps of chromatin state in pluripotent and lineage-committed cells**

Tarjei S. Mikkelsen<sup>1,2</sup>, Manching Ku<sup>1,4</sup>, David B. Jaffe<sup>1</sup>, Biju Issac<sup>1,4</sup>, Erez Lieberman<sup>1,2</sup>, Georgia Giannoukos<sup>1</sup>, Pablo Alvarez<sup>1</sup>, William Brockman<sup>1</sup>, Tae-Kyung Kim<sup>1</sup>, Richard P. Koechel<sup>1,2</sup>, William Lee<sup>1</sup>, Eric Mendenhall<sup>1,4</sup>, Aisling O'Donovan<sup>1</sup>, Aviva Presser<sup>1</sup>, Carsten Ruse<sup>1</sup>, Xiaohui Xie<sup>1</sup>, Alexander Meissner<sup>1</sup>, Marius Wernig<sup>1</sup>, Rudolf Jaenisch<sup>1</sup>, Chad Nusbaum<sup>1</sup>, Eric S. Lander<sup>1,2\*</sup> & Bradley E. Bernstein<sup>1,4,6\*</sup>

Mikkelsen et al.  
Broad Institute  
Nature in August.

**High-Resolution Profiling of Histone Methylations in the Human Genome**

Artem Barski<sup>1,3</sup>, Suresh Cuddapah<sup>1,3</sup>, Kaiyong Cui<sup>1,3</sup>, Tae-Young Roh<sup>1,3</sup>, Dustin E. Schones<sup>1,3</sup>, Zhibin Wang<sup>1,3</sup>, Gang Wei<sup>1,3</sup>, Iouri Chepelev<sup>1</sup>, and Keji Zhao<sup>1,2</sup>

<sup>1</sup>Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA  
<sup>2</sup>Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA  
<sup>3</sup>These authors contributed equally to this work and are listed alphabetically.  
\*Correspondence: shuck@nih.gov  
DOI 10.1016/j.cell.2007.05.009

Barski et al.  
NHLBI, NIH  
Cell in May.

**Genome-Wide Mapping of in Vivo Protein-DNA Interactions**

David S. Johnson<sup>1\*</sup>, Ali Mortazavi<sup>2\*</sup>, Richard M. Myers<sup>1,†</sup>, Barbara Wold<sup>2,3,†</sup>

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 [ $\approx 50$  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  $\geq 0.96$ ] and statistical confidence ( $P < 10^{-8}$ ), properties that were important for inferring new candidate interactions. These include key transcription factors in the gene network that regulates pancreatic islet cell development.

Johnson et al.  
Stanford University  
Science in June.



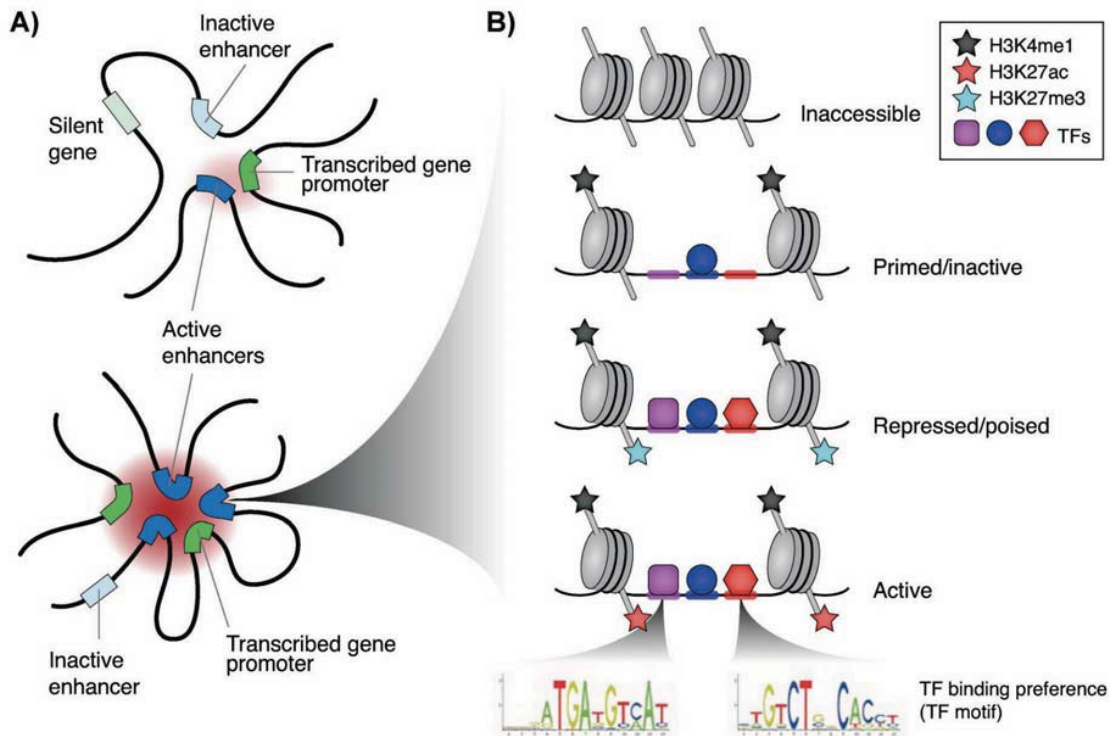
## Why do we need genome-wide CHIP?

- **Genome function on a global scale (epigenome)**
  - Cellular function / cell identity are determined by genome-wide manner
- **Finding transcription factor binding sites genome wide (TFBS)**
  - Transcription factors (TFs) are the determinants of context-specific transcription
- Chromatin/TFBS landscape of the **normal and diseased** cell.
- Understanding ES cell **growth/differentiation/reprogramming**.
- Made a huge impact on chromatin biology, epigenetics, transcription research, etc.

**However, TF ChIP-seq requires a lot of optimization process**

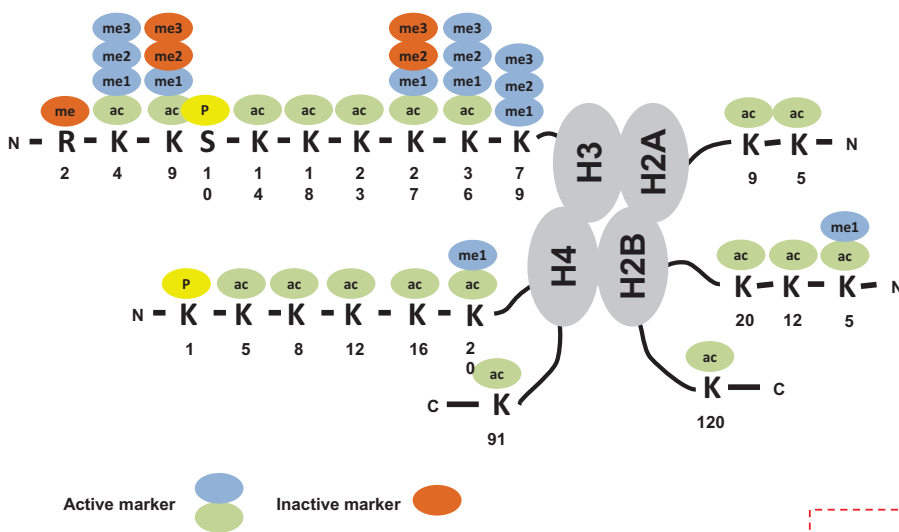
**Multiple TFs involve in directing cellular identify, but it is impossible to characterize all TF binding sites**

# How can we systematically identify regulatory elements genome-wide?



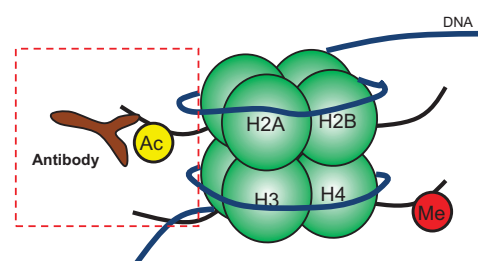
Andersson R. (2014)

# Histone modification can be an indicator of chromatin state



## Four major histone modification types

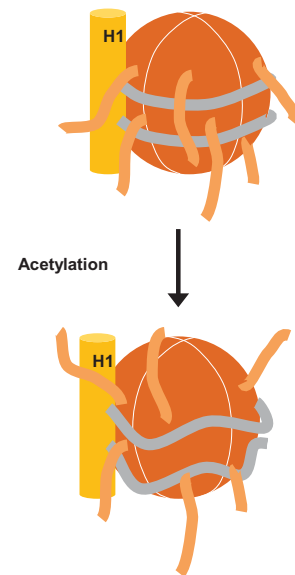
1. Acetylation : H3K27ac ...
2. Methylation : H3K4me3 ...
3. Phosphorylation: H3S10 ...
4. Ubiquitination : H2BUb ...



## Consequence effect of histone modification

### 1. Change in chromatin packing

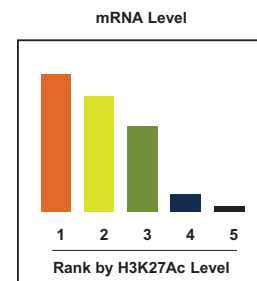
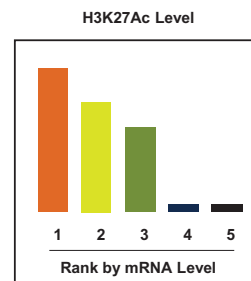
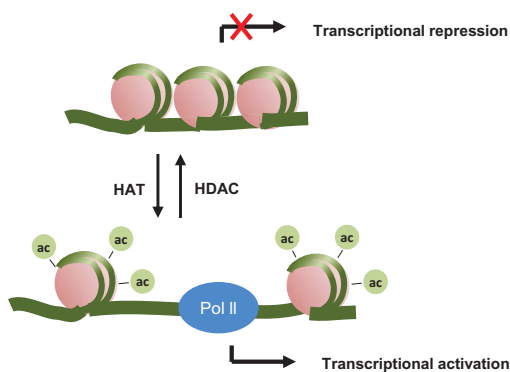
- The positive charge on the histone proteins are reduced through histone modification
- As a result, interactions between histone and negatively charged DNA are reduced and loosen chromatin packing



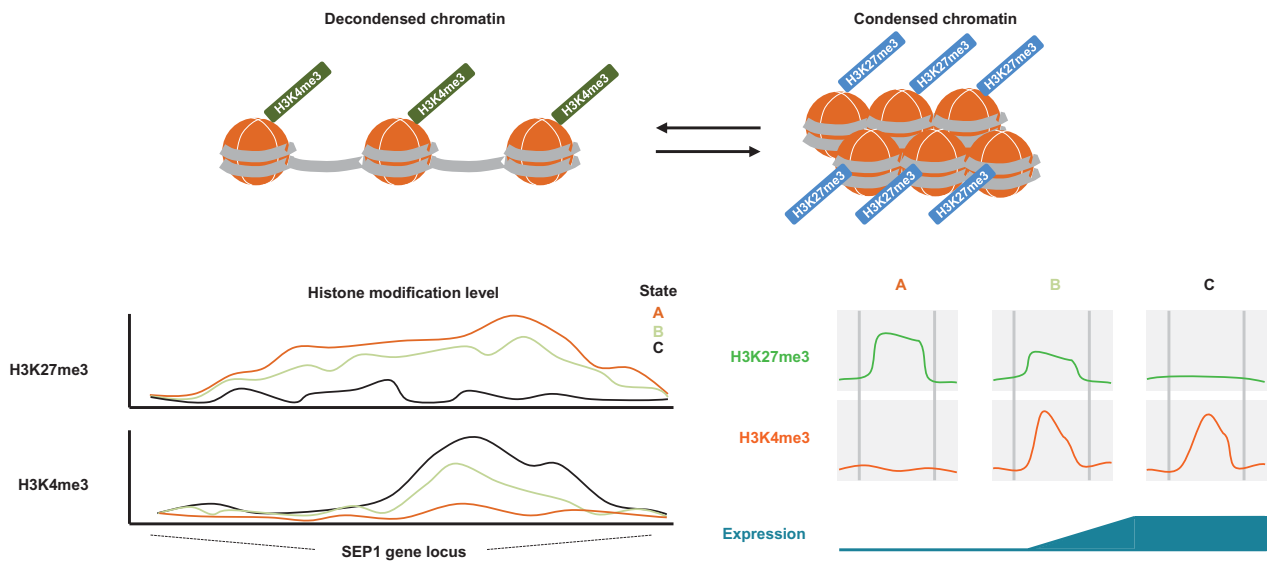
### 1. Site for recognition site for other protein

- Histone modifications can provide recognition sites for other regulatory proteins that alter chromatin compaction or induce other histone modification

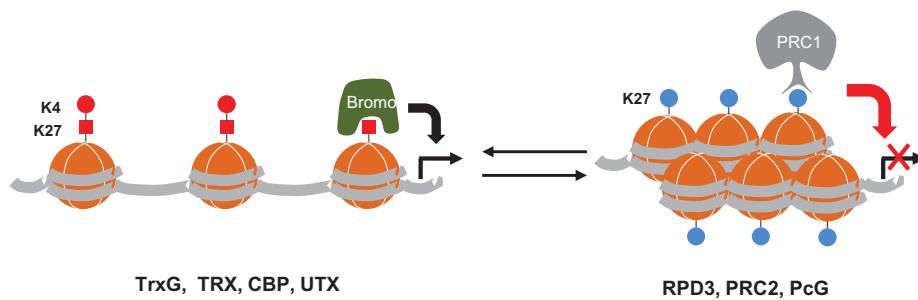
## Acetylation : H3K27ac



## Methylation : H3K4me3 vs H3K27me3

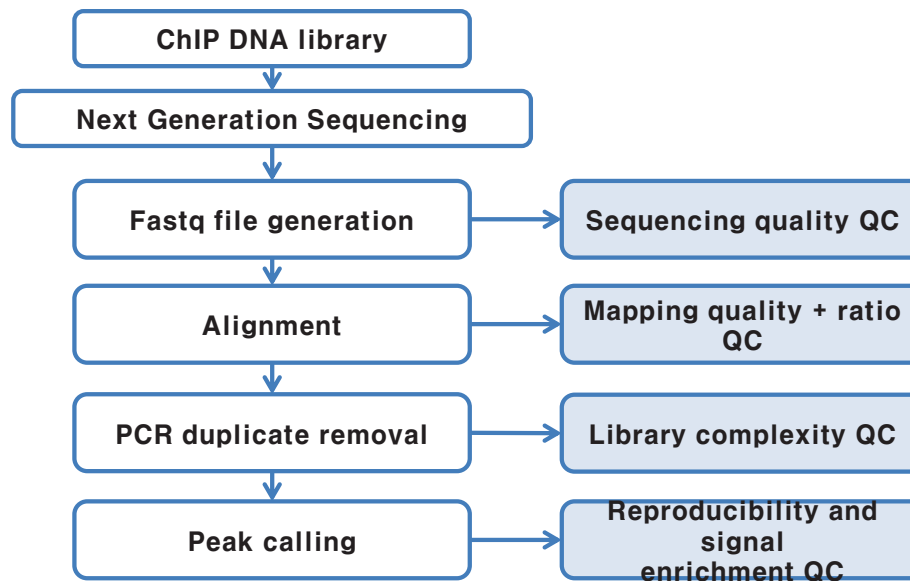


## Reader of acetylation and methylation



- PRC1 recognize  $K27me3$  and repress gene expression
- Bromodomain recognize  $K27ac$  and activate gene expression

## Workflow of ChIP-seq data preprocessing



## Quality Control of ChIP-seq – Alignment

### Resource

---

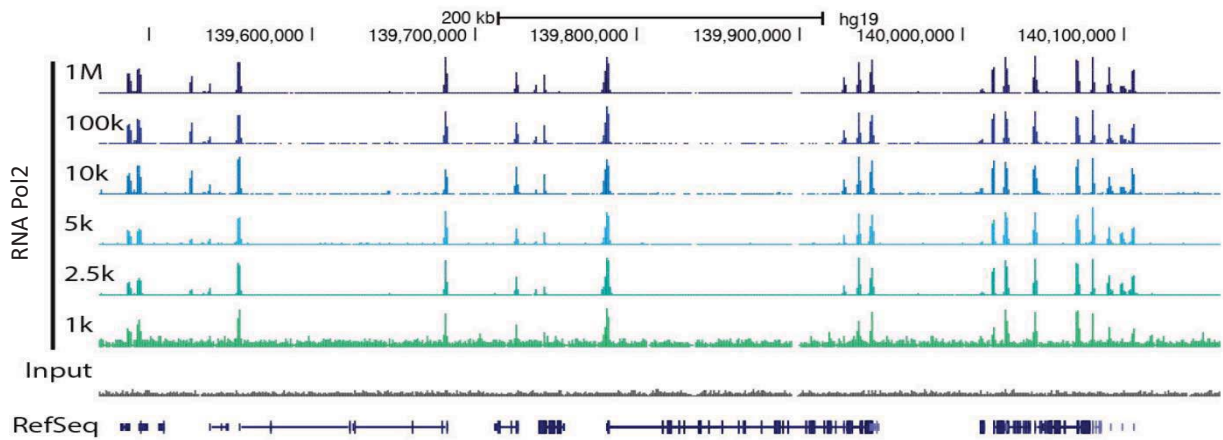
## ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Landt et al., Genome Research (2012)

### ChIP-seq quality guidelines after alignment and peak call

- Browser inspection for previously known sites
- Measuring global ChIP enrichment (FRiP)
- Cross-correlation analysis
- Consistency of replicates

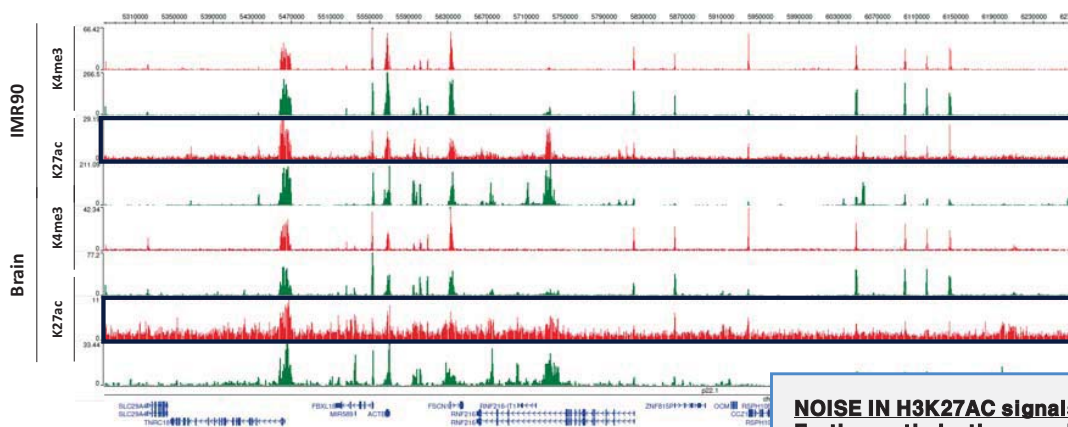
## Quality Control of ChIP-seq – Browser inspection



GenomeCoverageBed (bam file to bedgraph) / bedGraphToBigwig (bedgraph to bigwig)

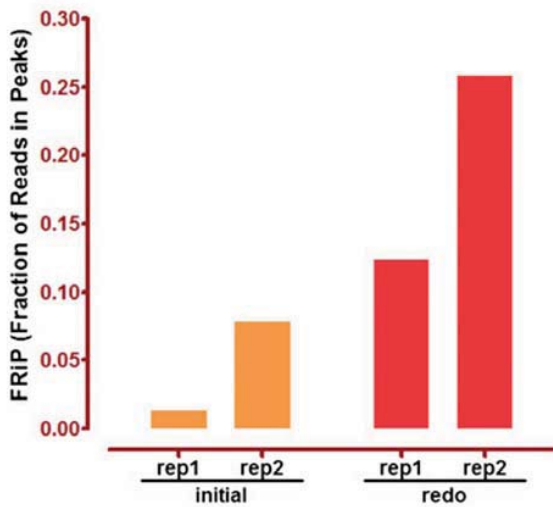
- RNA Pol2 is a mark for active promoter regions
- We should see RNA Pol2 peaks at known active gene promoter regions
- For example, if you use stem cells, oct4, nanog, and sox2 promoters are marked by Pol2 peaks

## Quality Control of ChIP-seq – Browser inspection



**NOISE IN H3K27AC signals  
Further optimization needed**

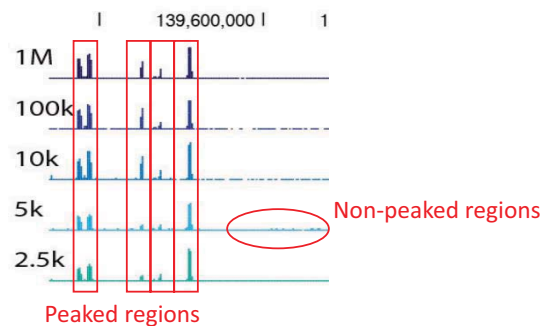
## Quality Control of ChIP-seq – Alignment: FRiP



K562 EGR1 ChIP-seq

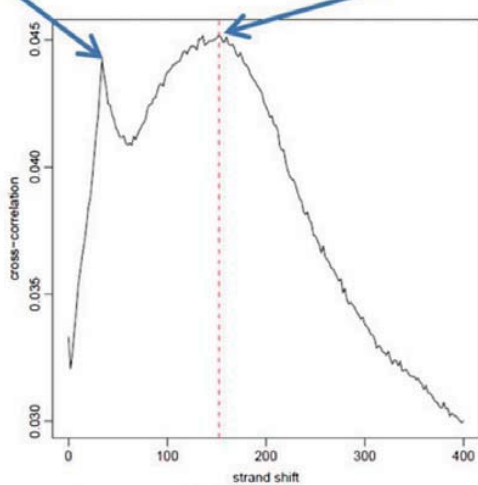
### FRiP: Fraction of reads in peaks

- A minority of reads in ChIP-seq experiments occur in significantly enriched genomic regions
- The remainder of the read represents background
- The fraction of reads falling within peak regions is a metric for the efficiency of the ChIP



## Quality Control of ChIP-seq – Alignment: Cross-correlation

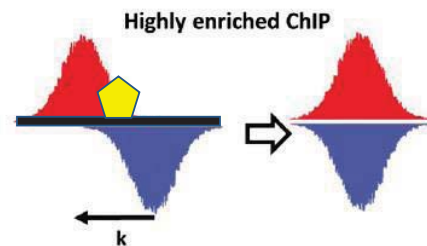
“phantom” peak      CHIP peak



Cross-correlation peaks

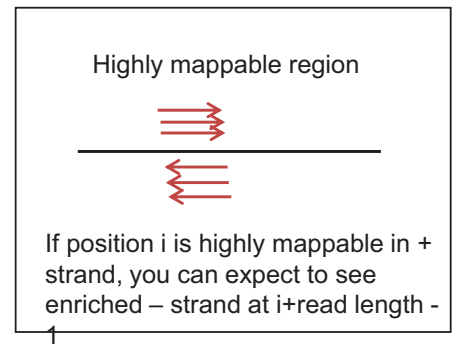
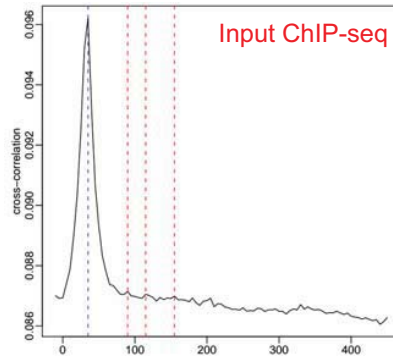
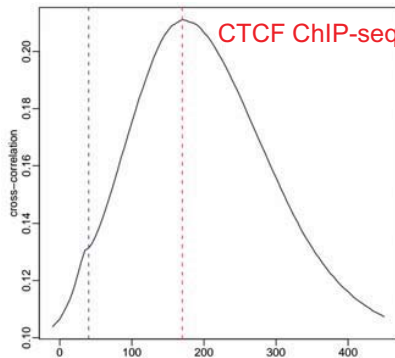
### Cross-correlation analysis

- A high-quality ChIP-seq experiment produces significant clustering of enriched DNA sequence tags at locations bound by the protein of interest
- Sequence tag density accumulates on forward and reverse strands centered around the binding site

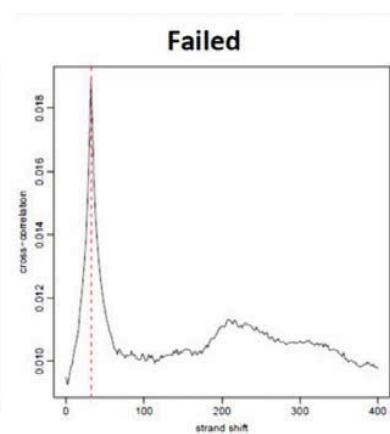
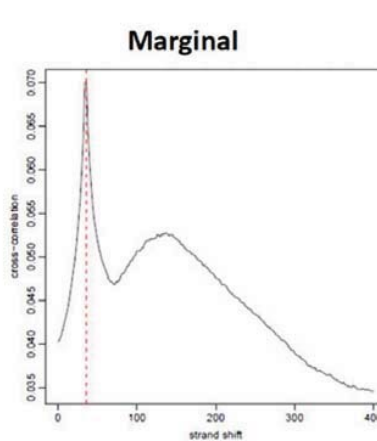
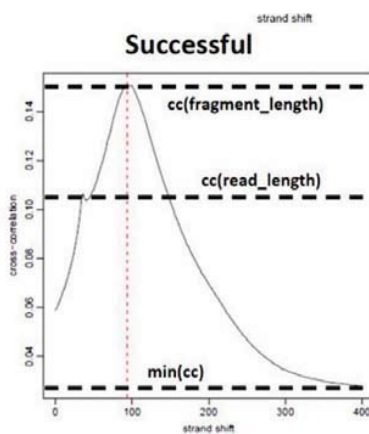


## Quality Control of ChIP-seq – Alignment: Cross-correlation

- Cross-correlation is a measure of similarity of two series as a function of the displacement of one relative to the other (a sliding dot product or sliding inner-product.)
- Cross-correlation: calculate Pearson correlation between genome-wide stranded tag densities (correlation between the Crick strand and the Watson strand after shifting Watson by k base pairs)
- Two peaks:
  1. A peak of enrichment corresponding to the predominant fragment length
  2. A peaks corresponding to the read length (phantom peak)



## Quality Control of ChIP-seq – Alignment: Cross-correlation



$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

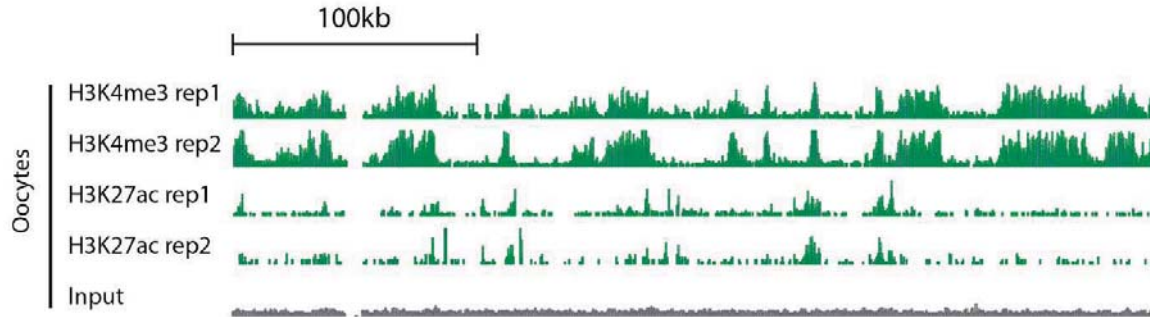
$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

Low quality ChIP-seq  
NSC value < 1.05  
RSC value < 0.8



# Quality Control of ChIP-seq - Reproducibility

Reproducibility is essential to reliable scientific discovery in high-throughput experiments

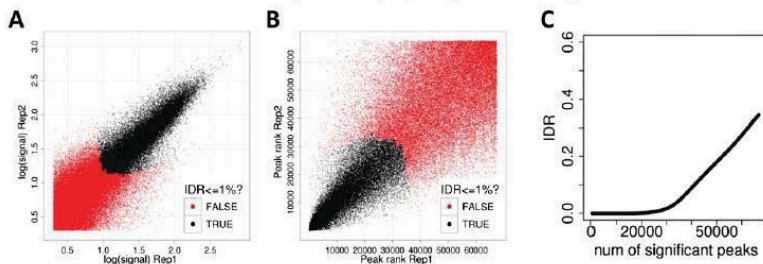


Highly reproducible broad H3K4me3 domains

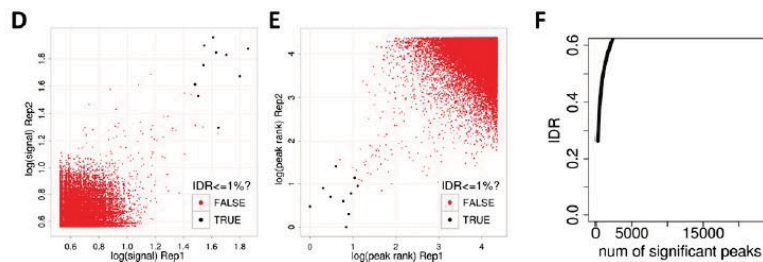
- For ChIP-seq, check the reproducibility of identified peaks

# Quality Control of ChIP-seq – Reproducibility (IDR)

## RAD21 Replicates (high reproducibility)



## SPT20 Replicates (low reproducibility)



- IDR: irreproducible discovery rate
- Rank identified peaks
- Check the consistency of highly ranked peaks
- Check whether consistent groups are ranked higher than the irreproducible group

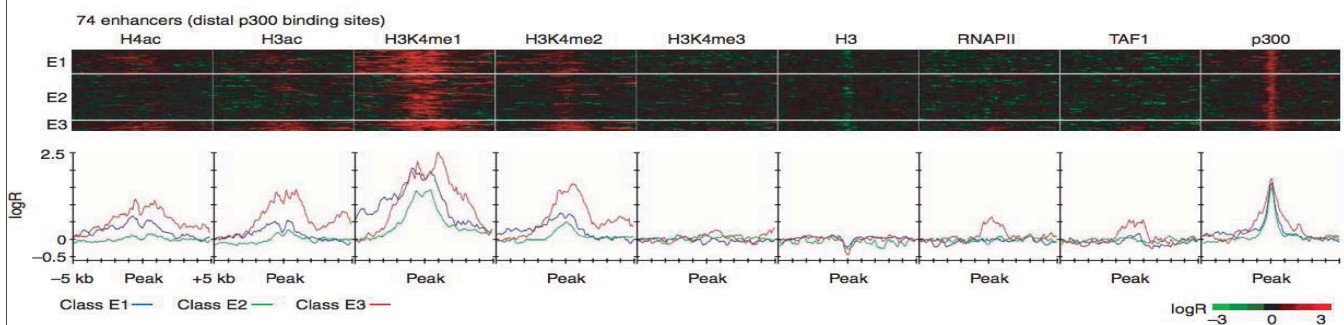
## Can we use histone modification signatures to define regulatory elements genome-wide instead of TF ChIP-seq?

### How can we identify cis-regulatory elements genome-wide?

#### Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome

Nathaniel D Heintzman<sup>1,2</sup>, Rhona K Stuart<sup>1</sup>, Gary Hon<sup>1,3</sup>, Yutao Fu<sup>4</sup>, Christina W Ching<sup>1</sup>, R David Hawkins<sup>1</sup>, Leah O Barrera<sup>1,3</sup>, Sara Van Calcar<sup>1</sup>, Chunxu Qu<sup>1</sup>, Keith A Ching<sup>1</sup>, Wei Wang<sup>5</sup>, Zhiping Weng<sup>4,6</sup>, Roland D Green<sup>7</sup>, Gregory E Crawford<sup>8</sup> & Bing Ren<sup>1,9</sup>

2007, Nature Genetics

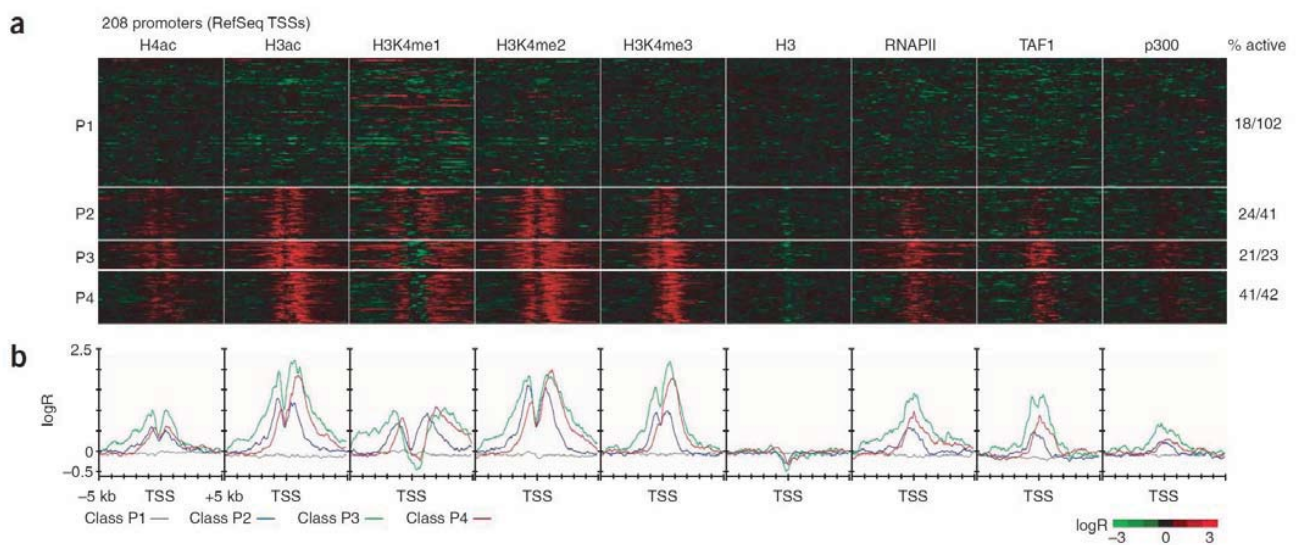


## How can we identify cis-regulatory elements genome-wide?

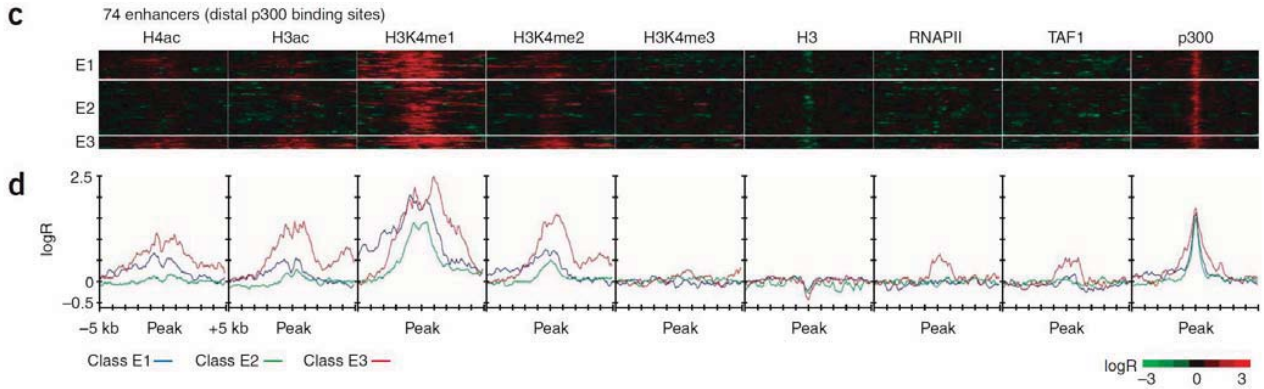
- Investigate the patterns of core histone H3 and five histone modifications: acetylated H3K9/14, acetylated H4K5/8/12/16 and mono-, di-, and trimethylated histone H3K4
- Examined binding of two components of the basal transcriptional machinery: RNAPII and TBP-associated factor 1 (TAF1) to identify active promoters
- Examined binding of transcriptional coactivator p300 to identify active enhancers
- ChIP-chip experiments for each marker in HeLa cells before and after treatment with interferon-gamma, as p300 is known to be involved in the cellular response to this cytokine at 38bp resolution

## Genome-wide identification of Promoter Signatures

Clustering of ChIP-chip profiles along 10kb regions surrounding promoters

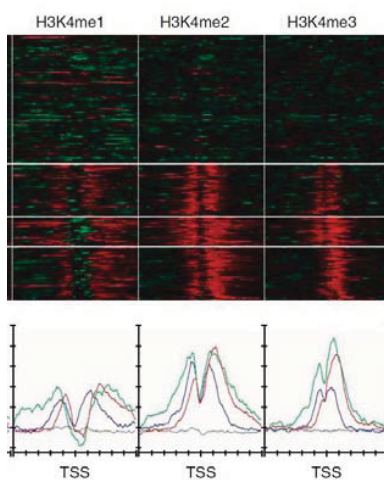


# Genome-wide identification of enhancer signatures

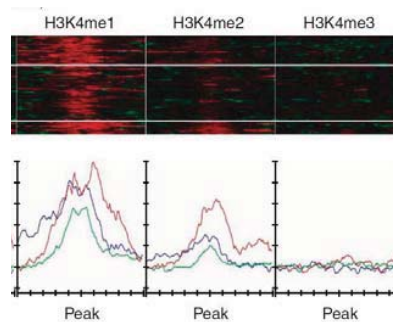


Strong H3K4me1 enrichment but depletion of H3K4me3

# Distinct chromatin signatures between promoters and enhancers



Can we predict promoters and enhancers via chromatin signatures?



Strong H3K4me1 enrichment at enhancers but bimodal distribution at promoters

## Genomic characteristic of predicted enhancer elements

Clustering of 10kb regions surrounding the distal p300 binding sites as putative enhancers regions (124 sites in untreated cells and 182 sites in treated cells)

The features of enhancers at p300 binding sites are..

1. Distribution of p300 sites was consistent with the widespread location of enhancers relative to their target genes (75% of p300 binding occurs > 2.5kb from TSS)
2. Significant number of overlap between p300 sites and DHS because enhancers have been known to show increased nuclease sensitivity (69.7%)
3. Most distal p300 sites were conserved across species (>60%)
4. Significant overlap between p300 sites and regulator modules predicted by TF binding motifs

**ENCODE/Roadmap Epigenomics provide  
reference epigenome maps in various cell-types and tissues**

# Systematic characterization of non-coding sequences



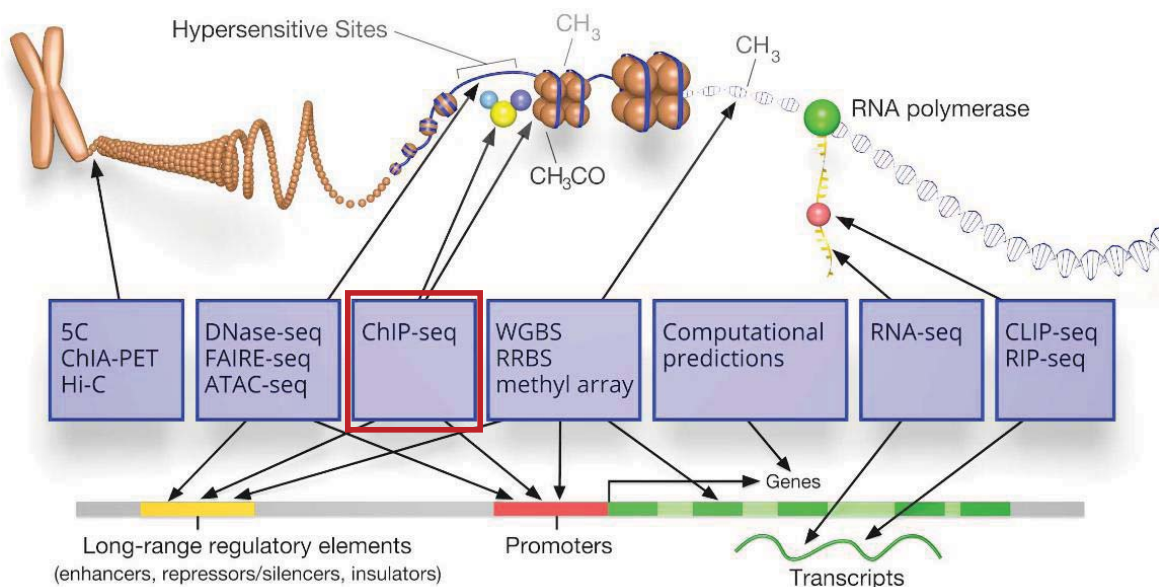
ENCODE consortium (2012)



Roadmap Epigenome consortium (2015)

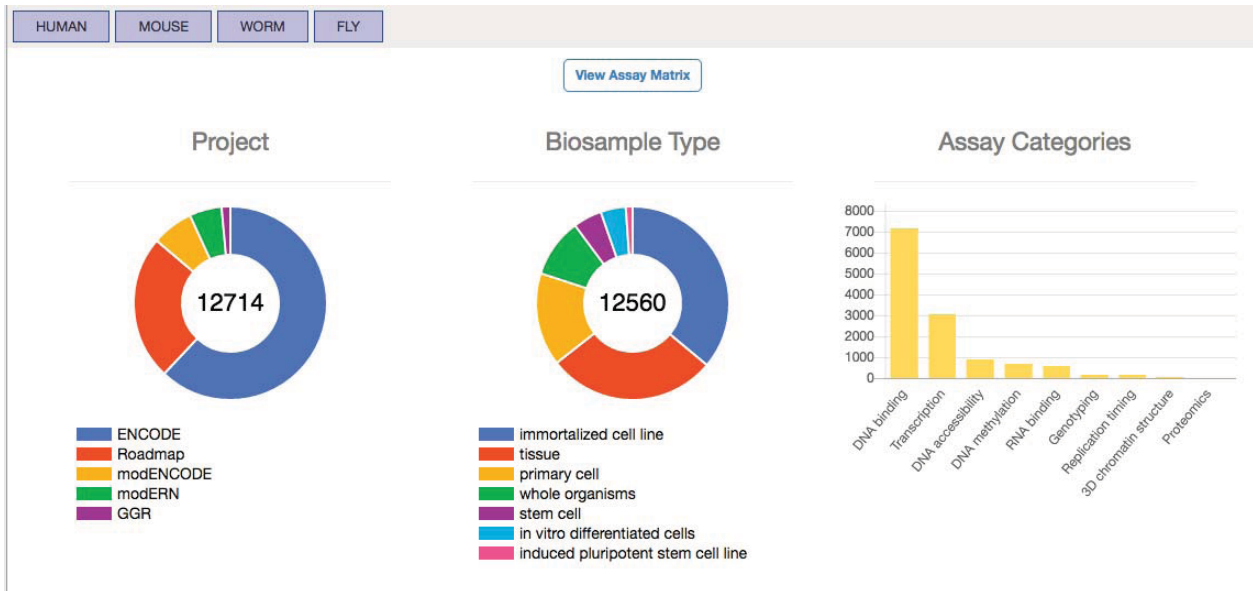
## ENCODE / Roadmap Epigenomics

- **Encyclopedia of DNA Elements (ENCODE)** : a public research project launched in 2003 (mostly cell lines)
- “aims to identify all functional elements in the human genome sequence.”
- **Roadmap Epigenomics**: Launched in 2008 (mostly primary human tissues)
- “aims to produce a public resource of epigenomic maps for stem cells and primary ex vivo tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease.”



# Data summary of ENCODE and Roadmap Epigenome

- 111 reference epigenomes
  - 1,821 histone modification datasets (ChIP-seq)
  - 360 DNase datasets
  - 277 DNA methylation datasets
  - 166 RNA-Seq datasets

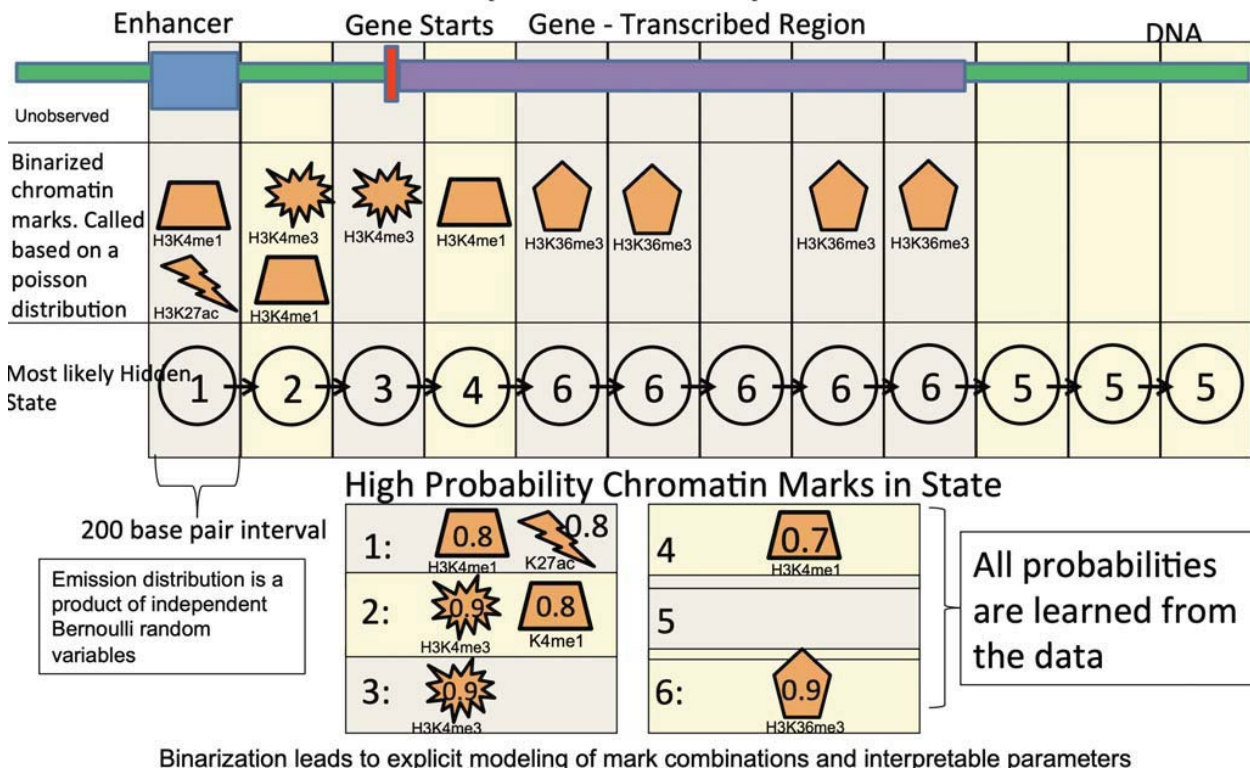


# WashU epigenome browser (<https://epigenomegateway.wustl.edu/>)

Annotation track for functional elements

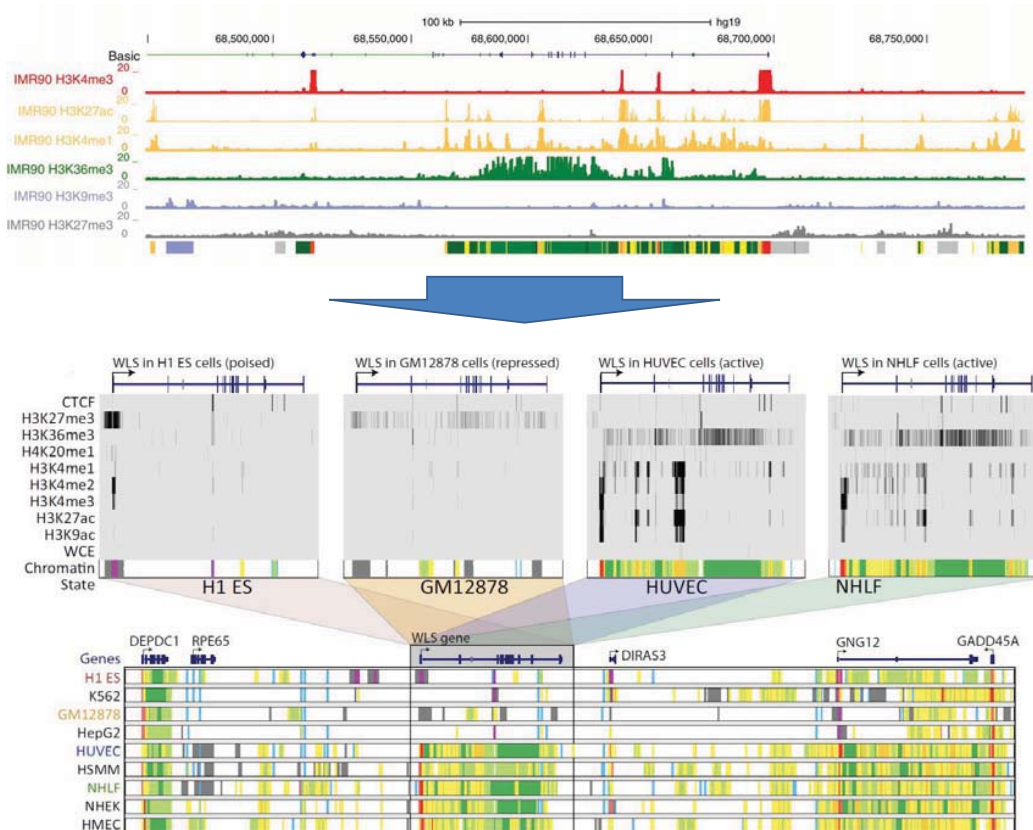


# A ChromHMM model to systematically annotate various chromatin state



Roadmap Epigenomics (2015)

# A ChromHMM model to systematically annotate various chromatin state

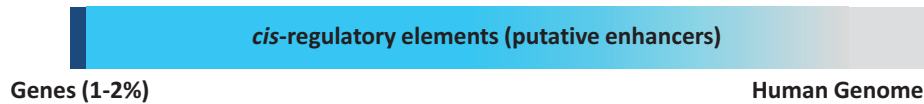


Nature (2011)



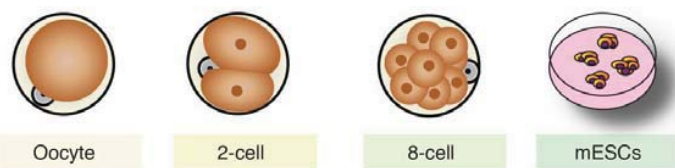
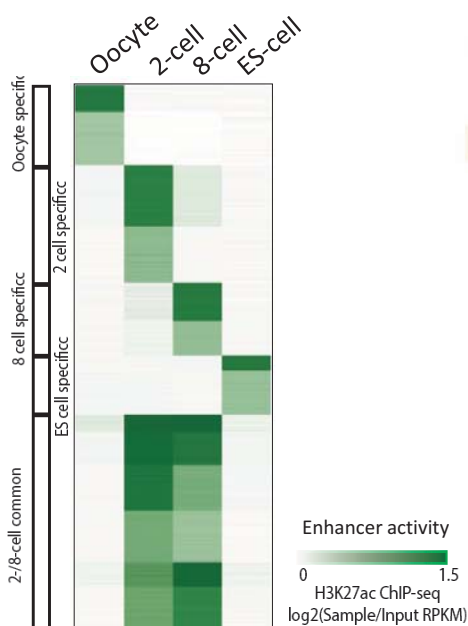
## What have we learned from ENCODE/Roadmap Epigenomics?

1. 80.4% of the human genome participates in at least one biochemical chromatin associated events



2. Many important genetic variants are found at *cis*-regulatory elements
3. Enhancer elements are the major player in cell-type specific gene regulation

## Enhancer activities are cell-type specific

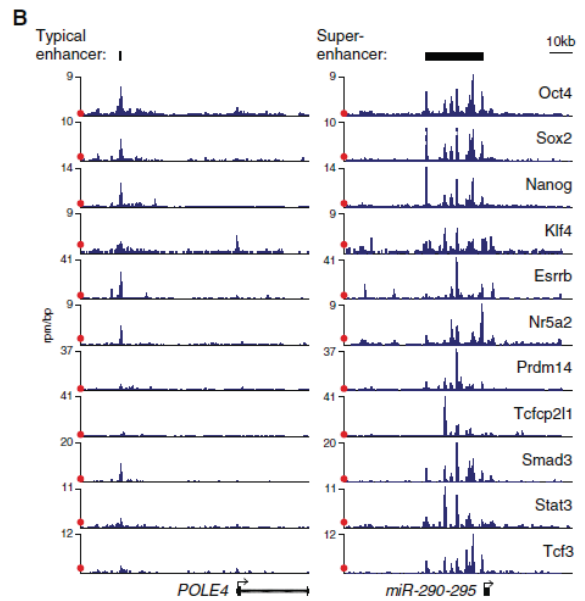
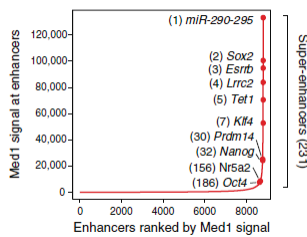


- 20,000~60,000 enhancers in each cell
- Enhancer activities are
  - Tissue/cell-type specific
  - Developmental stage specific
- Linked to tissue/cell-type specific gene expression

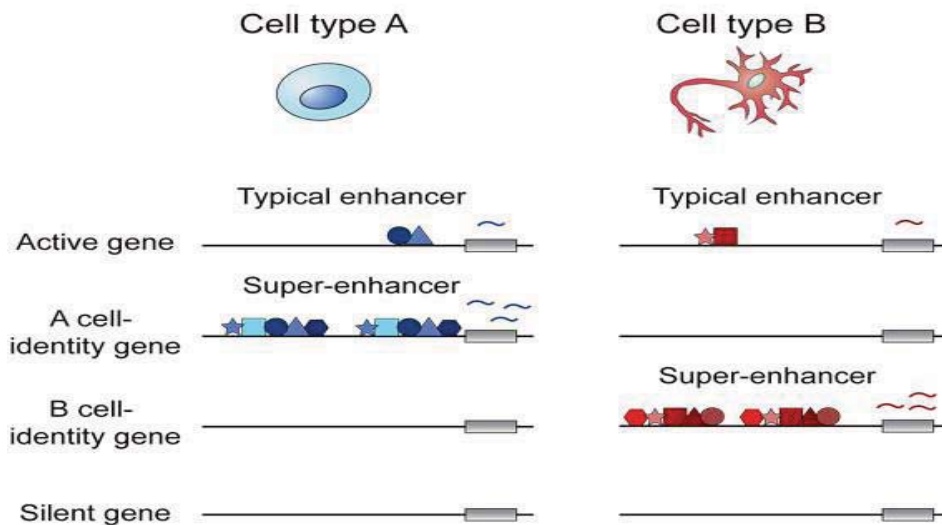
# Super-enhancer: a cluster of enhancers

## Super-Enhancers in the Control of Cell Identity and Disease

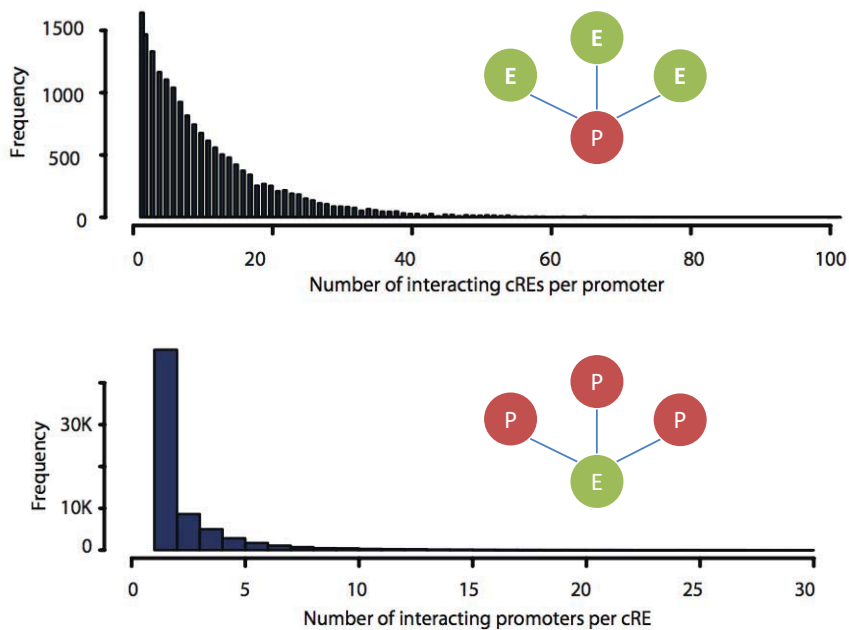
Denes Hnisz,<sup>1,2</sup> Brian J. Abraham,<sup>1,2</sup> Tong Ihn Lee,<sup>1,2</sup> Ashley Lau,<sup>1,2</sup> Violaine Saint-André,<sup>1</sup> Alla A. Sigova,<sup>1</sup> Heather A. Hoke,<sup>1,2</sup> and Richard A. Young<sup>1,2\*</sup>  
<sup>1</sup>Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA  
<sup>2</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA  
<sup>\*</sup>These authors contributed equally to this work  
<sup>\*</sup>Correspondence: young@wi.mit.edu  
<http://dx.doi.org/10.1016/j.cell.2013.09.053>



# Cell-type specific SE regulates cell-type specific gene expression



## One promoter can be controlled by multiple enhancers



## Summary

1. ChIP allows to investigate genome-wide location of DNA binding proteins and histone modifications
2. Histone ChIP-seq reveals that profiling H3K4me1 and H3K4me3 enable to define enhancer and promoter elements genome-wide
3. Enhancers are key sequences that control cell-type specific gene expression

# KSBi-BIML 2022

## 3D Epigenome in Gene Regulation (염색질 3차구조 개요)

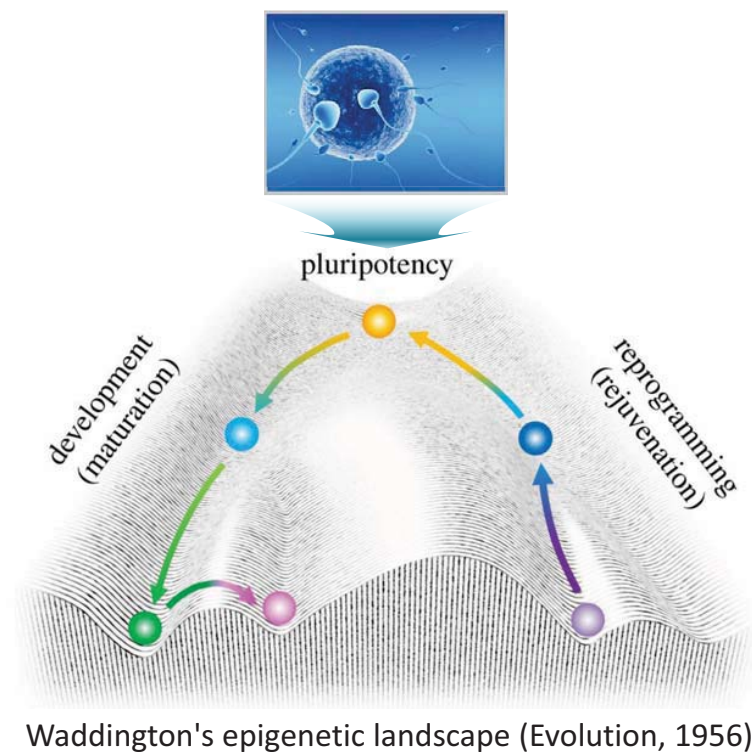
정인경(KAIST)

### Contents

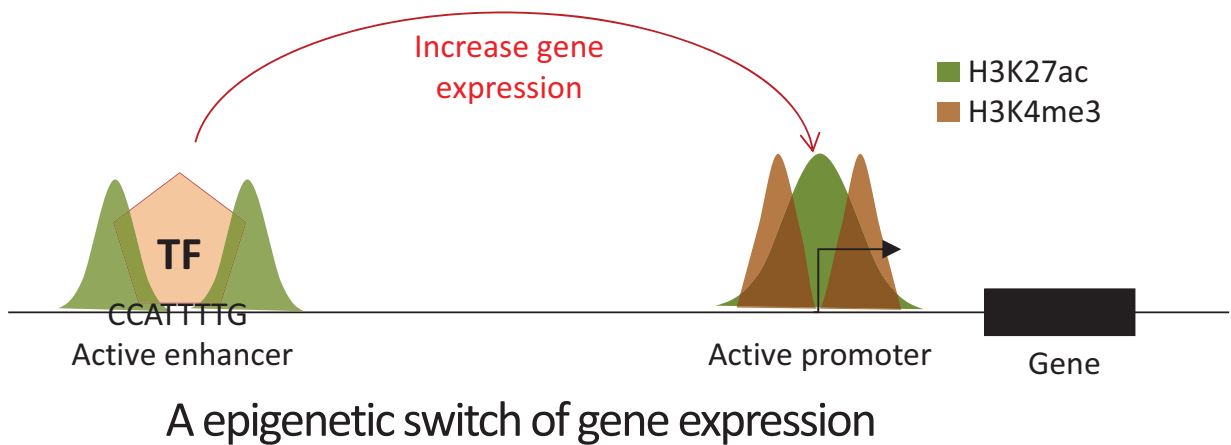
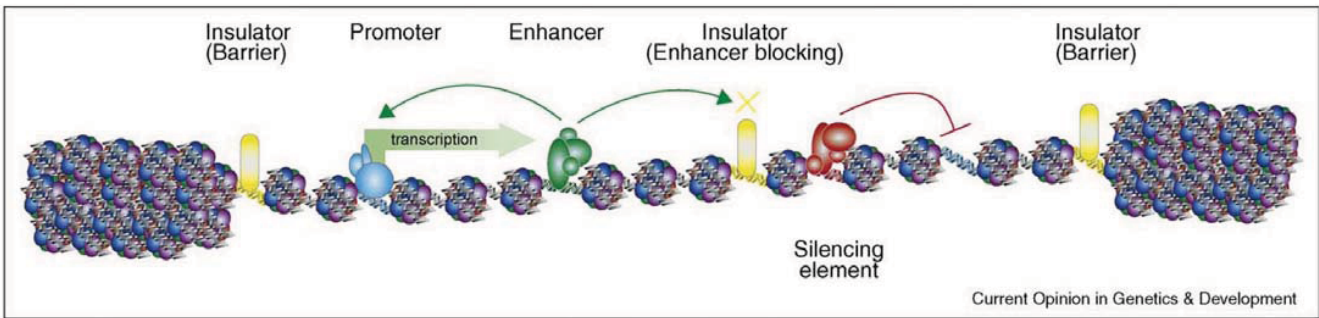
1. 후성유전체 및 CHIP-seq 개요
- 2. 염색질 3차구조 개요**
3. 3DIV 기반 염색질 3차구조 및 유전자 조절 통합 분석 실습
4. 염색질 3차구조 데이터 분석 실습

1. Introduction to 3D genome
2. Methods to explore 3D genome
3. Compartment A/B
4. Topologically associating domains
5. Long-range chromatin interactions

## Epigenetic gene regulation determines cell fate

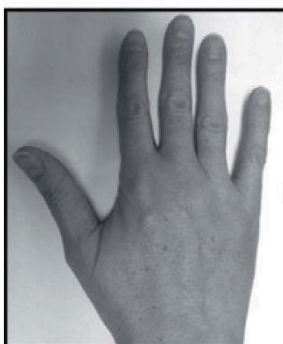


## “Enhancer” is a major player in epigenetic gene regulation



## Enhancers can control distal target gene expression

### Polydactyly syndrome



Normal hand



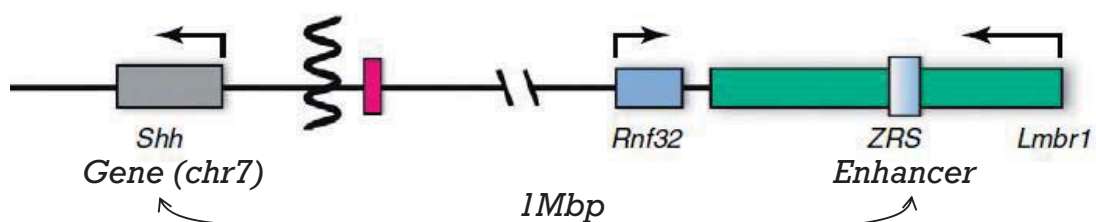
Preaxial polydactyly type2



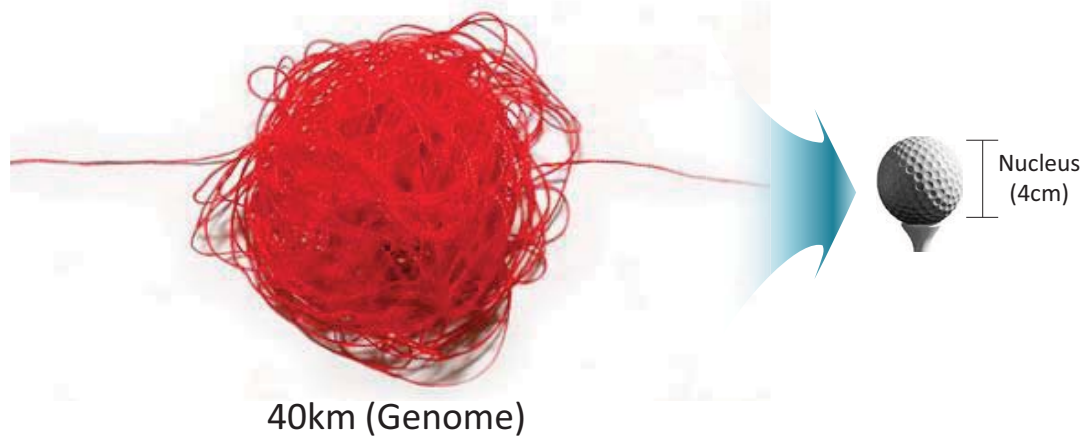
Postaxial polydactyly typeA



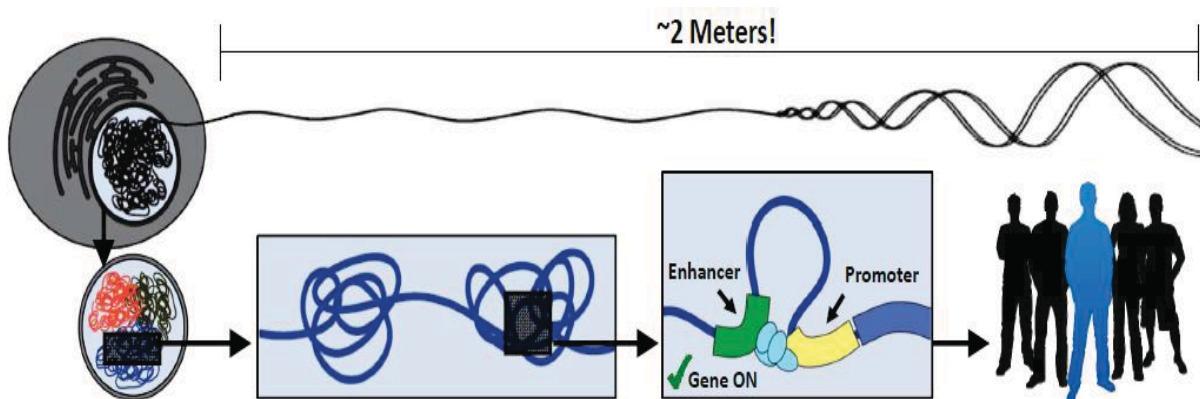
Triphalangial thumb polysyndactyly



## How does enhancer control distal gene expression?

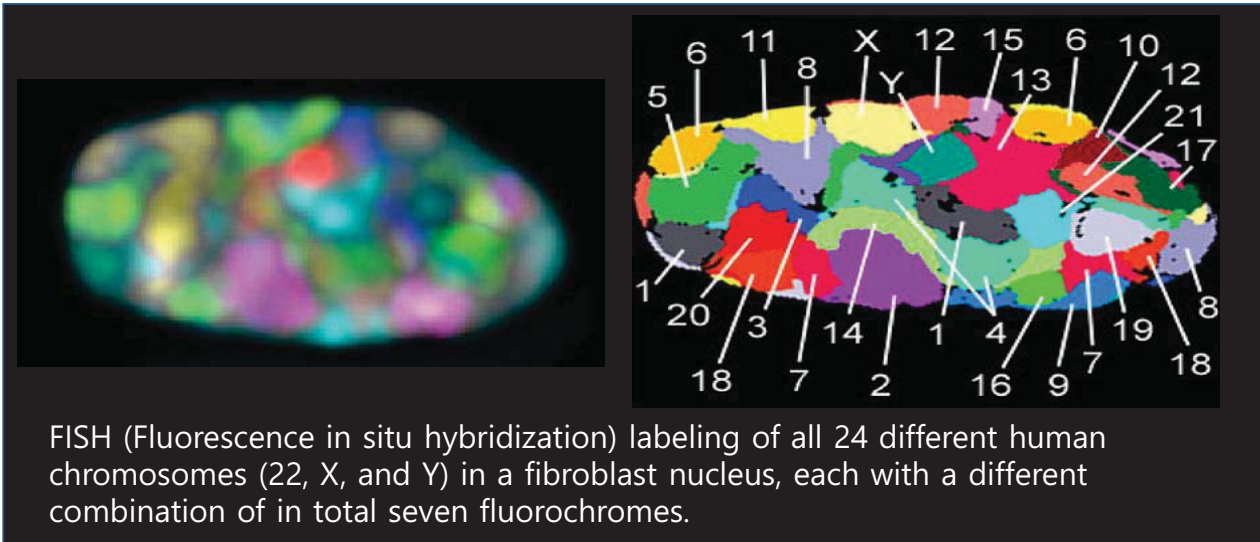


## Chromatin is not randomly folded into the nucleus



- Human DNA is well packaged
  - Length = near 2m
  - Average human cell nucleus : 6 micron  $\rightarrow$  1/300,000 compaction
- Chromatin is not randomly folded into the nucleus

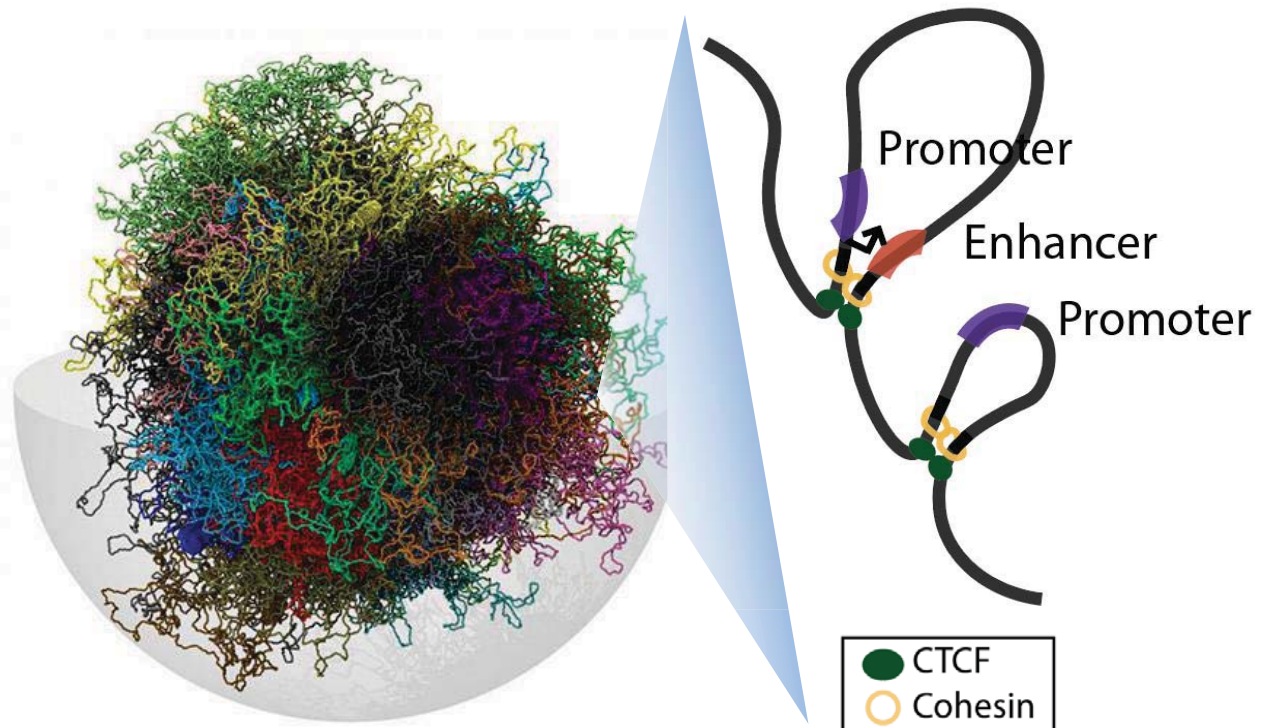
## A theory of chromosome territory



FISH (Fluorescence in situ hybridization) labeling of all 24 different human chromosomes (22, X, and Y) in a fibroblast nucleus, each with a different combination of in total seven fluorochromes.

[Bolzer et al., \(2005\)](#)

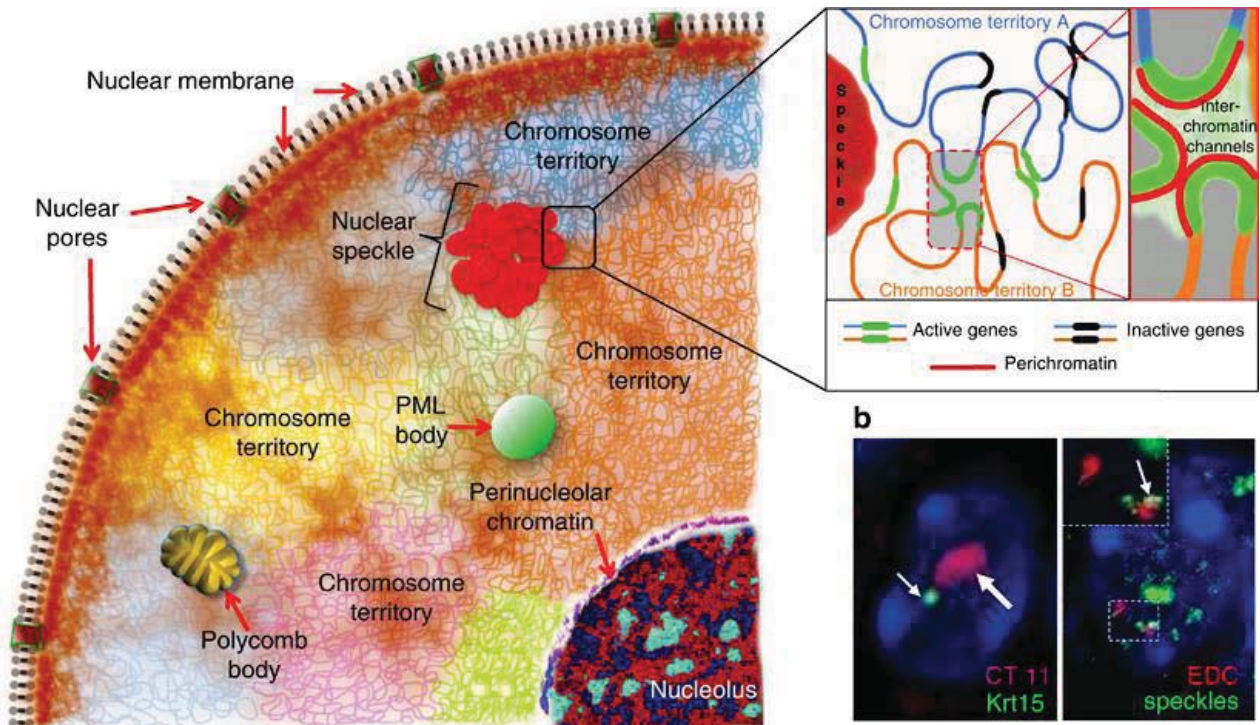
## 3D genome enables enhancers control distal gene expression



**3D genome: A spatial arrangement of the genome where distant DNA fragments can be juxtaposed in nuclear space**



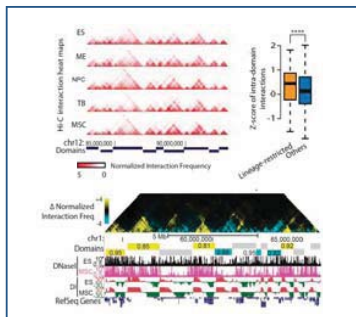
# Genome organization in 3D nuclear space



Botchkarev et al., 2012

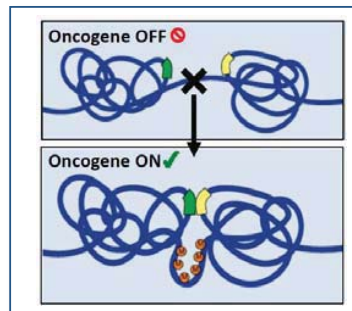
## Genome functions are tightly coupled with 3D chromatin structure

### Cellular differentiation



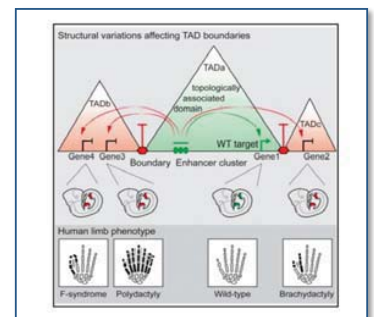
Dixon, JR\*, Jung, I\*, et al., Nature (2015)

### Oncogene activation



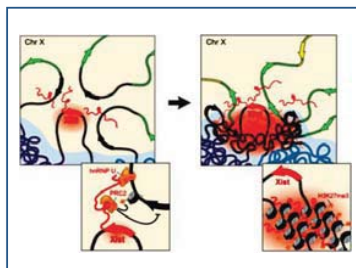
Hnisz et al., Science (2016)

### Congenital disorder



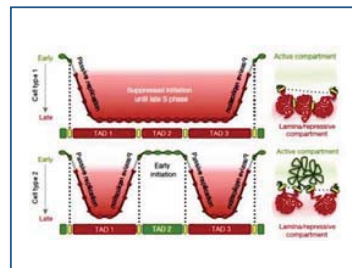
Franke et al., Nature (2016)

### X-chromosome inactivation



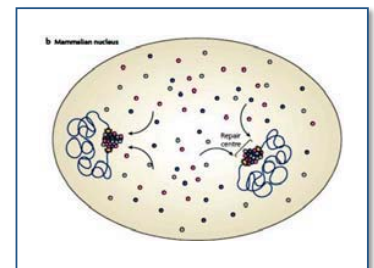
Engreitz et al., Science (2013)

### DNA replication



Pope et al., Nature (2014)

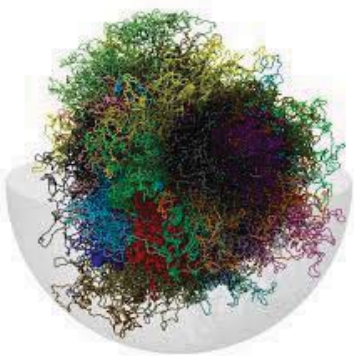
### DNA repair



Misteli & Soutoglou, Mol Cell Biol (2009)

1. Introduction to 3D genome
- 2. Methods to explore 3D genome**
3. Compartment A/B
4. Topologically associating domains
5. Long-range chromatin interactions

## Methods to detect 3D genome organization



### Imaging based methods:

1. Electron microscopy : labor intensive and not easily applicable to studies of specific loci
2. Light microscopy: Limited resolution (100~200 nm) to define chromosome conformation.
3. FISH (fluorescence in situ hybridization): Requires severe treatment that may affect chromosome organization

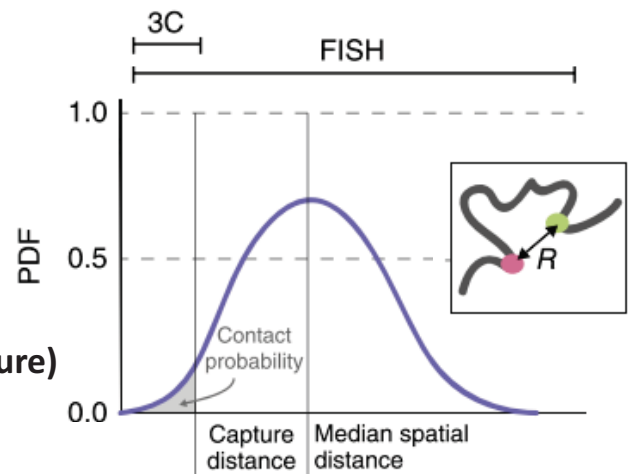
**Require alternative strategies to detect chromatin interactions at high-resolution genome-wide**

**3C**   **5C**   **4C-seq**   **ChIA-PET**   **Hi-C**   **Capture HiC**   **HiChIP**   **GAM**   **LAD**

# Imaging vs sequencing methods

## Imaging (FISH)

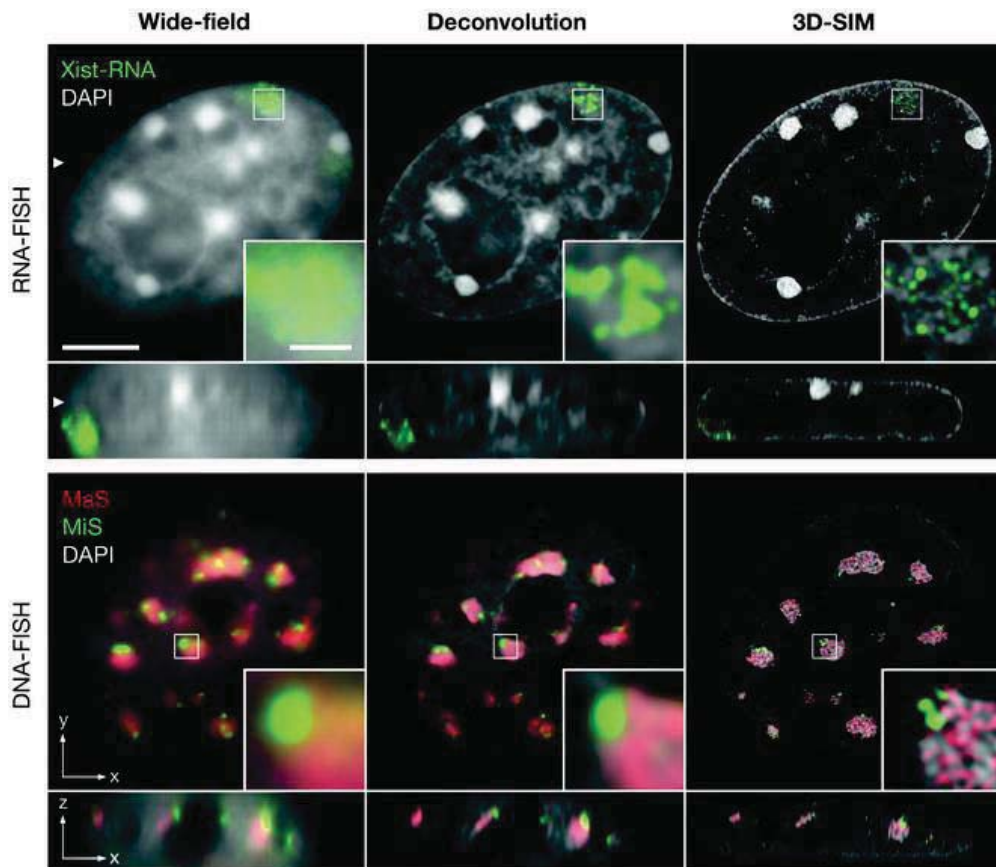
- In general: Single cell
- **Spatial distance**
  - Any distance outside probe “glare”



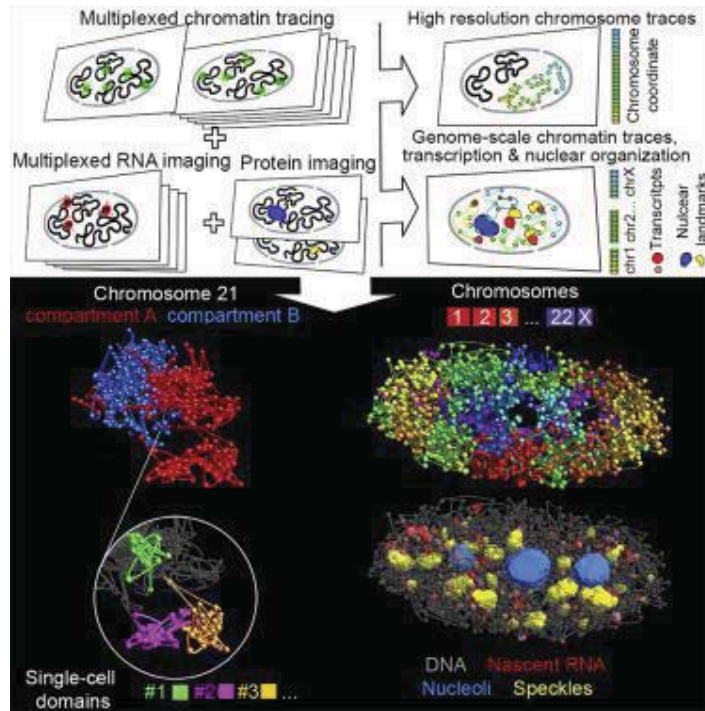
## Omics (Chromosome Conformation Capture)

- In general: population
- **Contact frequency**
  - Capture radius dependent
  - Long distances in close proximity

Belmont, *Curr.Opin.Cell Biol.* 2014  
 Giorgiotti, *Gen.Biol* 2016  
 Fudenberg and Imakaev, *Nat.Methods* 2017



[https://link.springer.com/protocol/10.1007/978-1-62703-137-0\\_4](https://link.springer.com/protocol/10.1007/978-1-62703-137-0_4)

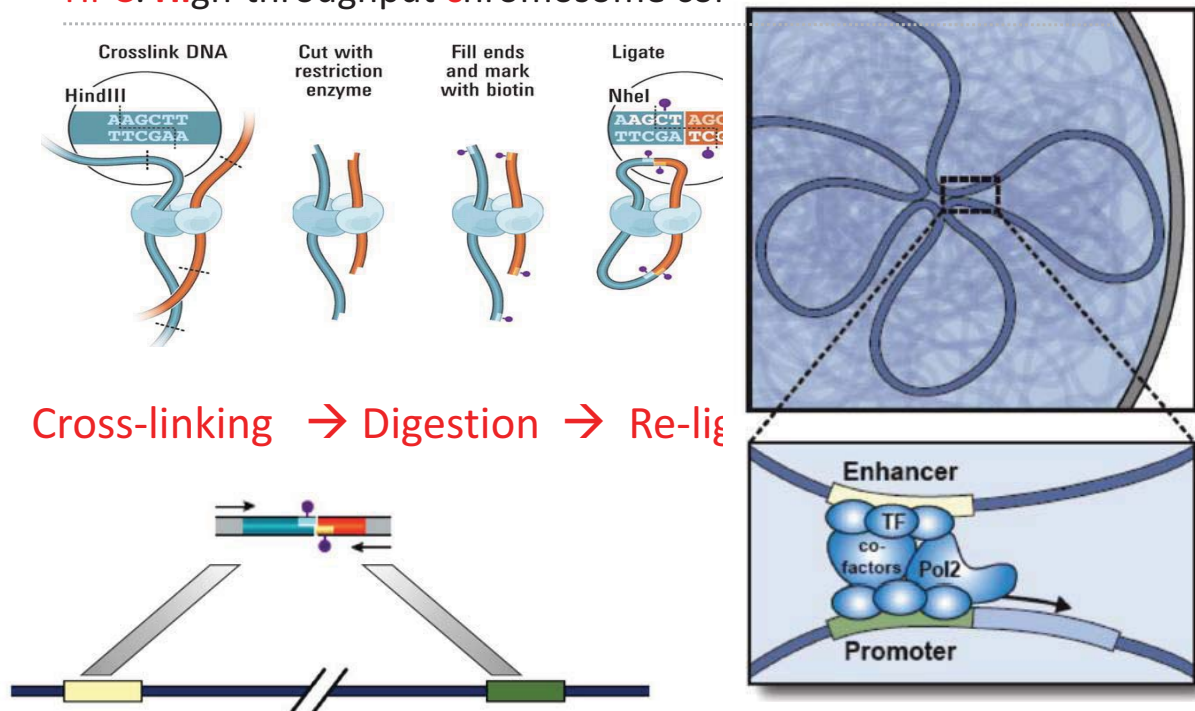


A multiplexed error-robust fluorescence *in situ* hybridization (MERFISH)

<https://www.sciencedirect.com/science/article/pii/S0092867420309405>

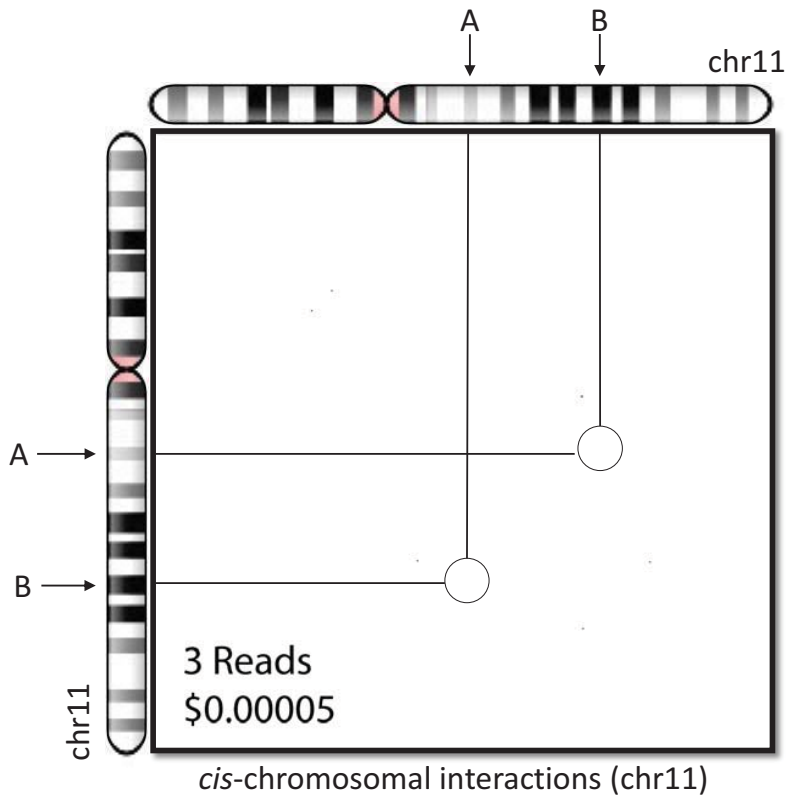
## How can we investigate 3D genome organization?

### Hi-C: High-throughput chromosome conformation capture (3C)

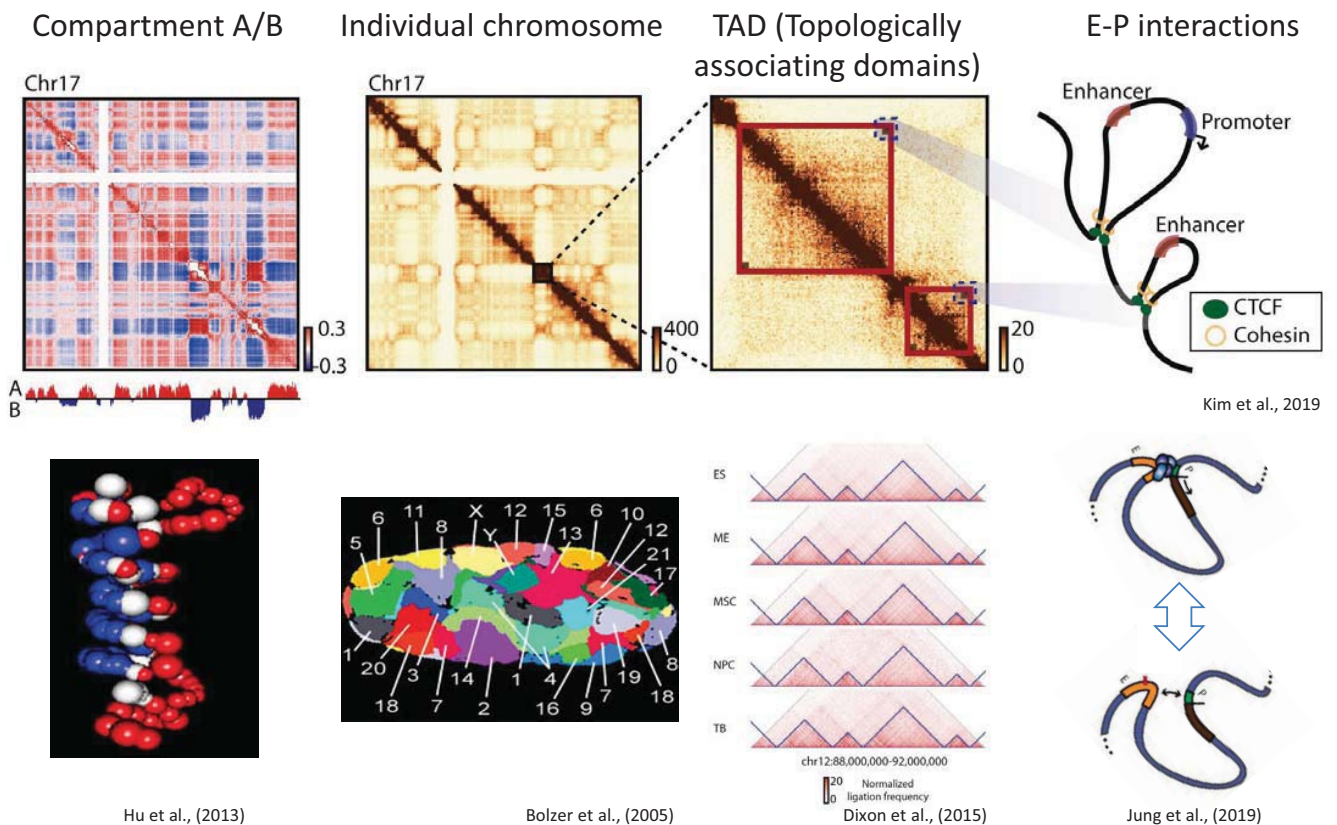


Lieberman et al., Science (2009)

# Hi-C contact map to visualize 3D genome

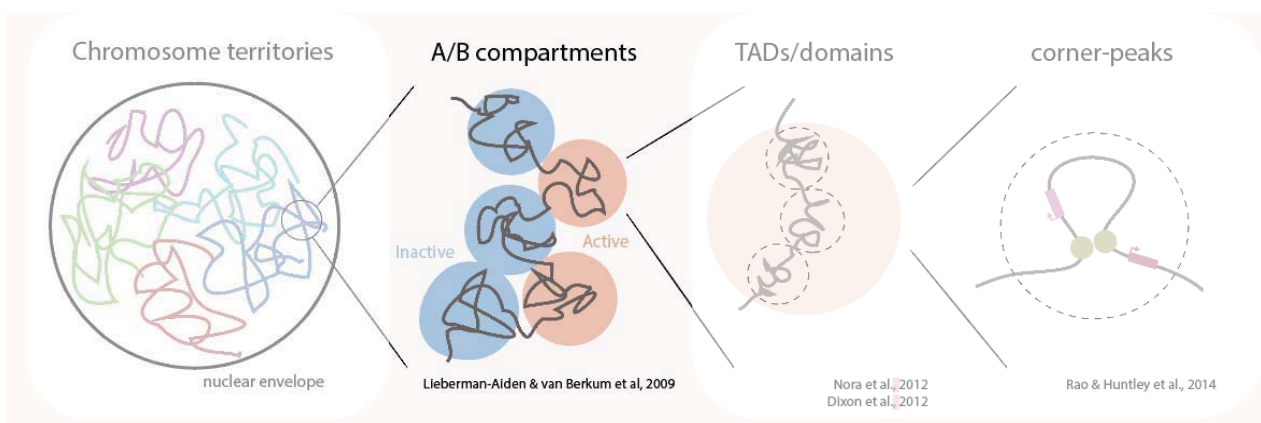


# Multi-layered 3D genome organization

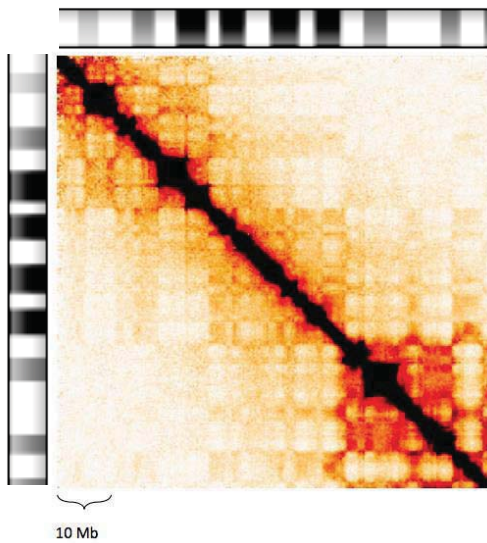


# Contents

1. Introduction to 3D genome
2. Methods to explore 3D genome
- 3. Compartment A/B**
4. Topologically associating domains
5. Long-range chromatin interactions



# Spatial compartmentalization of 3D genome

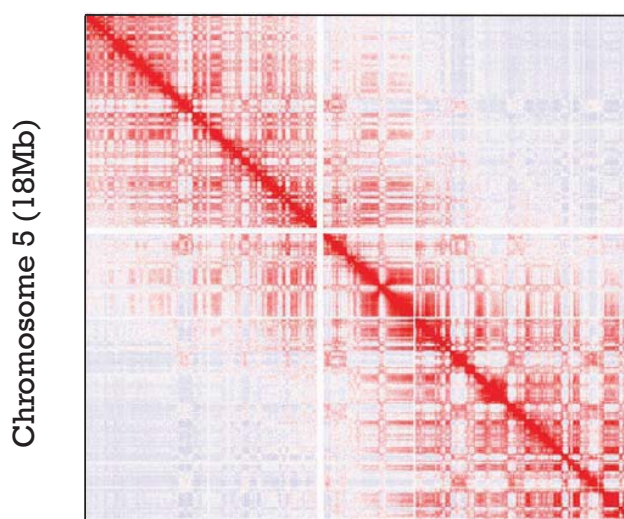


		1	2	3	4
		A	B	A	B
1	A				
2	B				
3	A				
4	B				

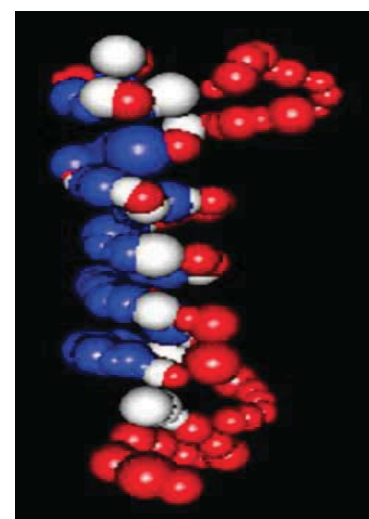
- What does a plaid pattern indicate for?
  - Higher interaction between fragment 1 and 3 and between fragment 2 and 4
- What is a biological meaning of the presence of a plaid pattern?
  - Genome can be compartmentalized into two parts (compartment A and B)

# Modeling 3D chromatin structure from Hi-C contact map

Chromosome 5 (181Mb)



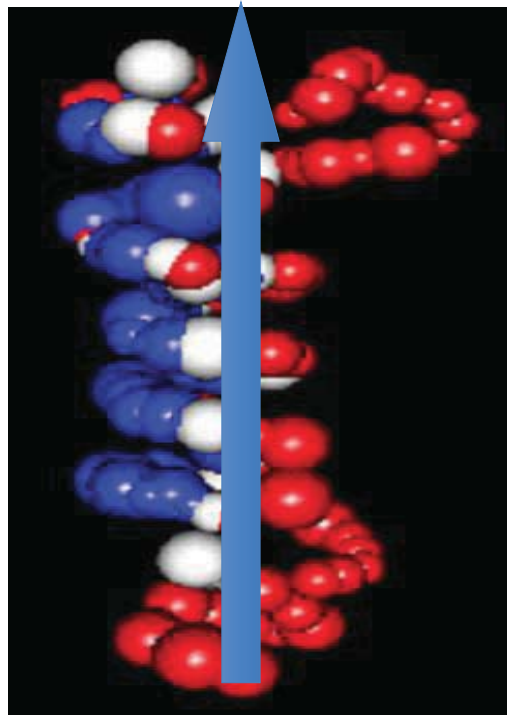
Hi-C Contact Map



3D modelling

# What is a major structural component?

Compartment B



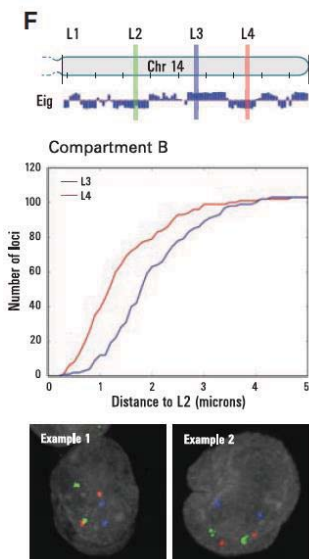
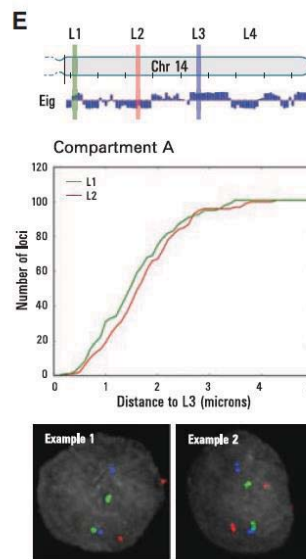
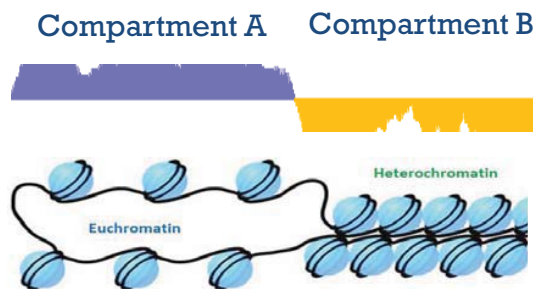
Compartment A

How can we systematically compartmentalize the 3D genome structure into two parts?

# Two major compositions of chromatin structure: Compartment A/B

How does compartment A/B affect spatial genome organization?

The loci in the same compartment showed spatial proximity

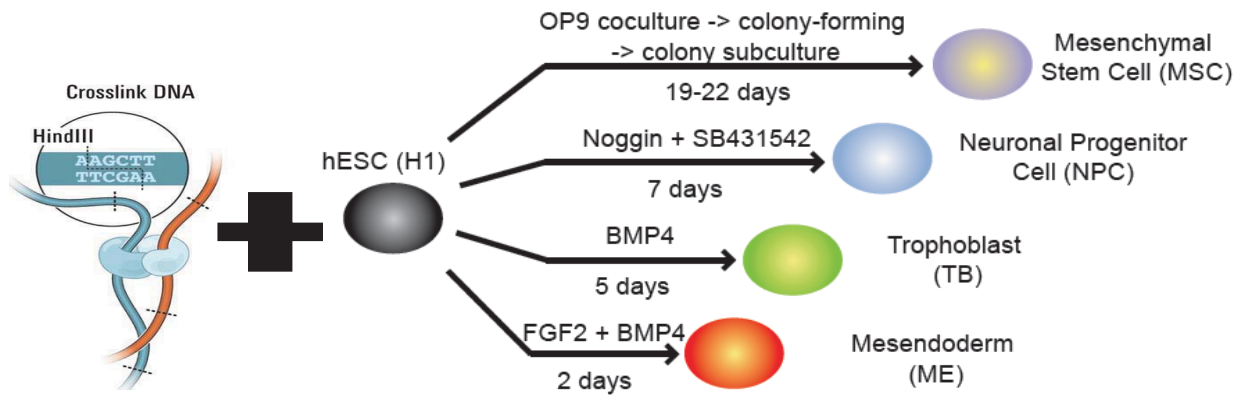




# Compartment A/B dynamics during stem cell differentiation

## Is compartment A/B cell-type specific?

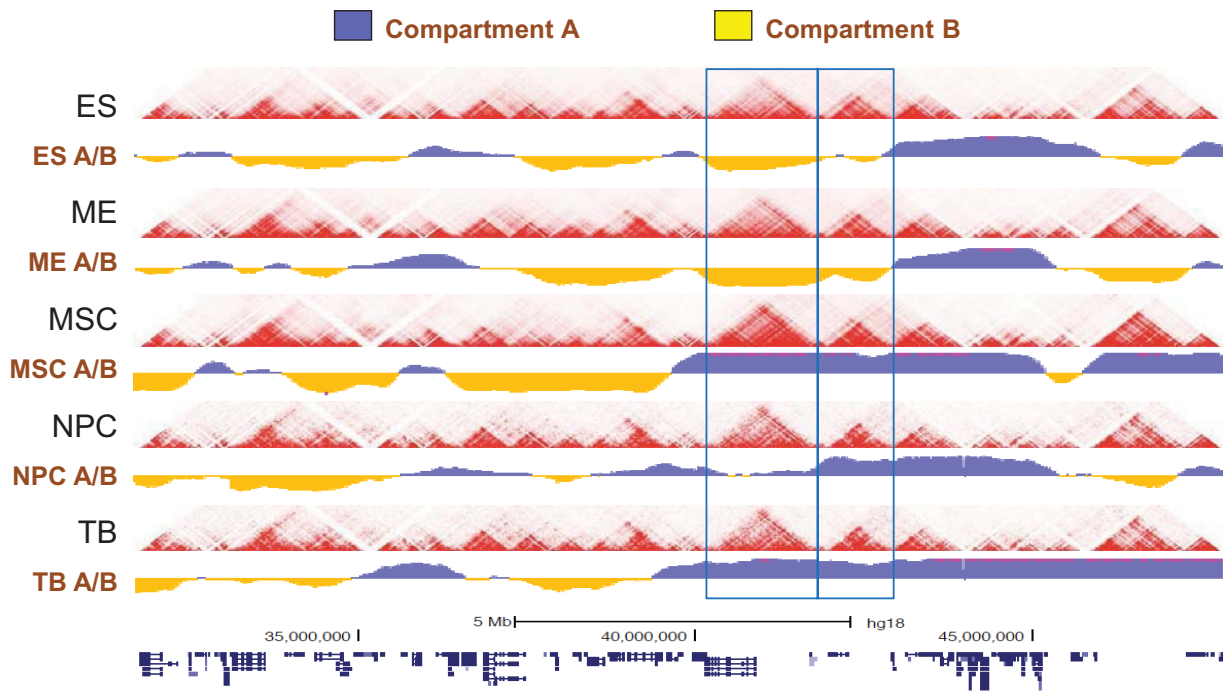
(How can we design a test experiment?)



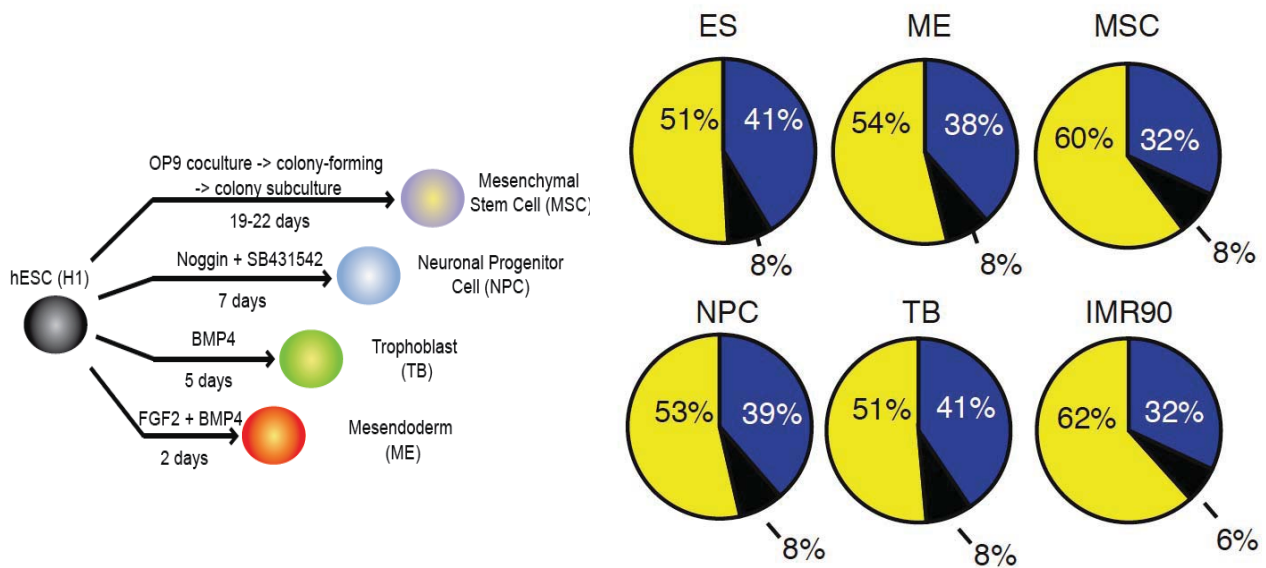
Perform Hi-C experiment

Dixon, JR., Jung, I., et al., Nature (2015)

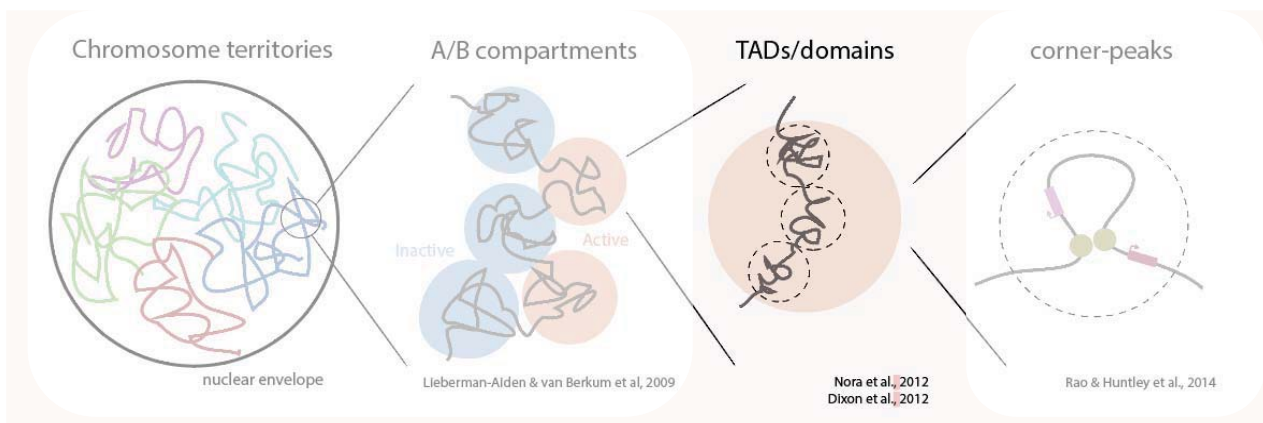
# Compartment A/B patterns are highly dynamic



# Fraction of compartment A/B in each cell-type

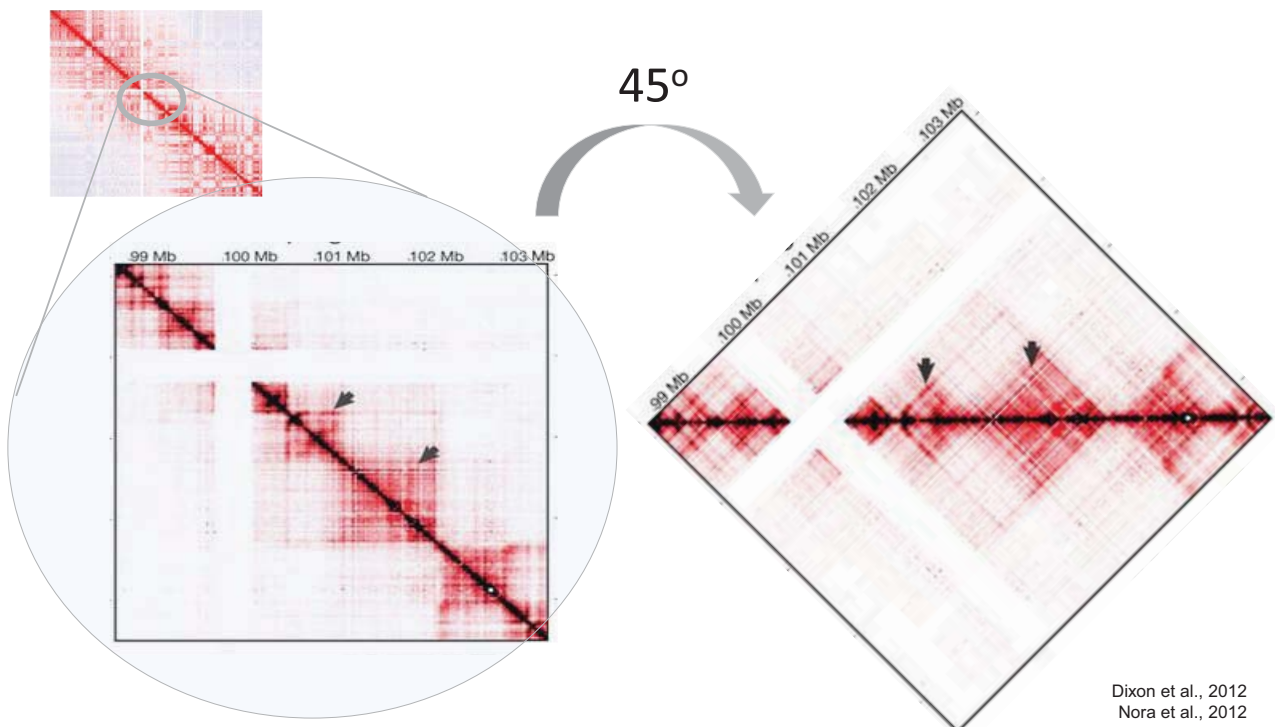


Fraction of genome marked as compartment A (blue) and B (yellow)

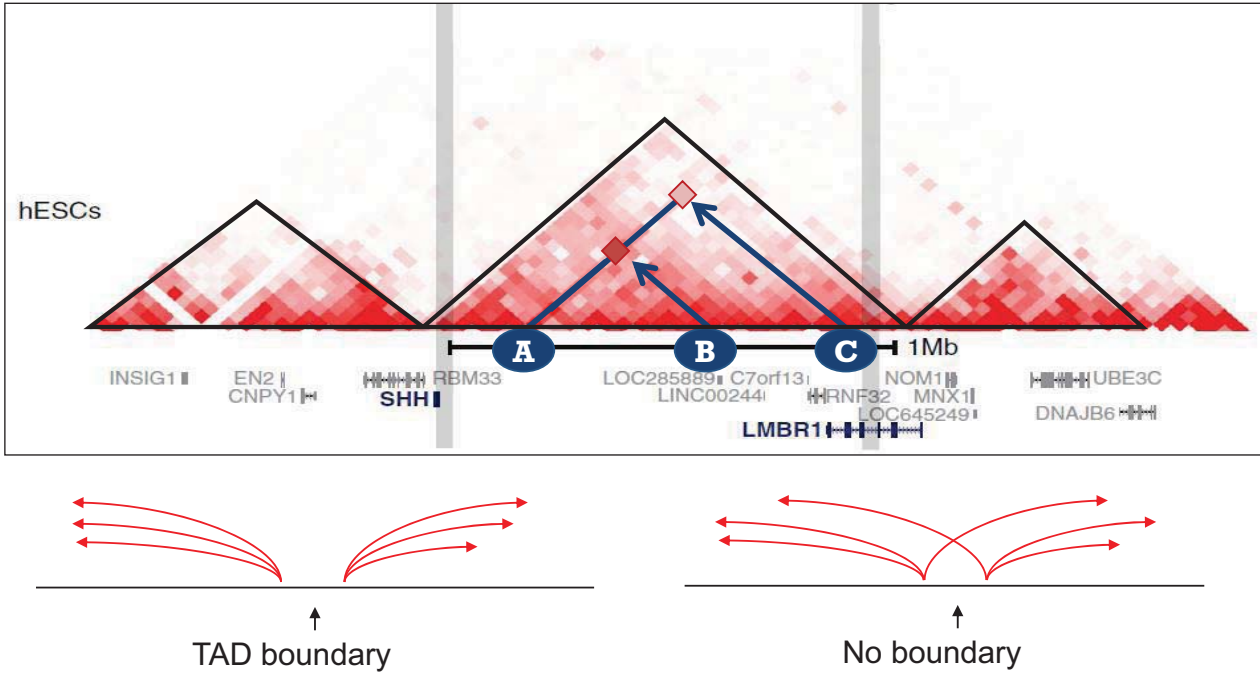


1. Introduction to 3D genome
2. Methods to explore 3D genome
3. Compartment A/B
- 4. Topologically associating domains**
5. Long-range chromatin interactions

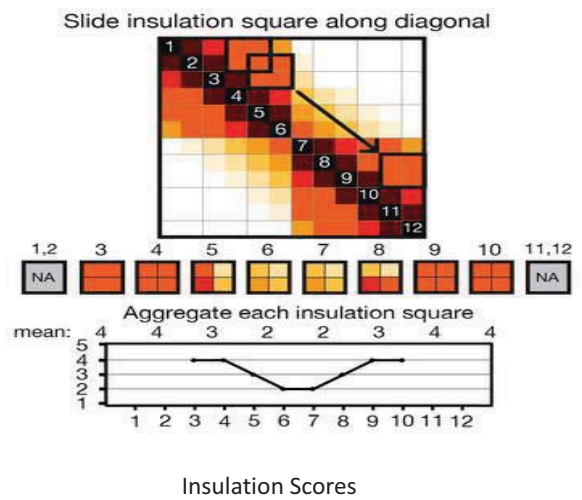
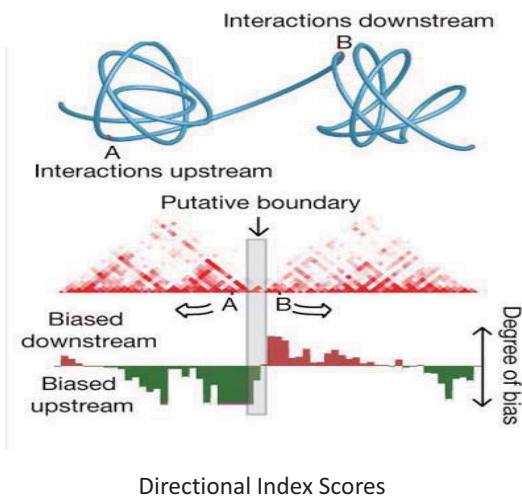
## Topologically Associating Domains (TADs)



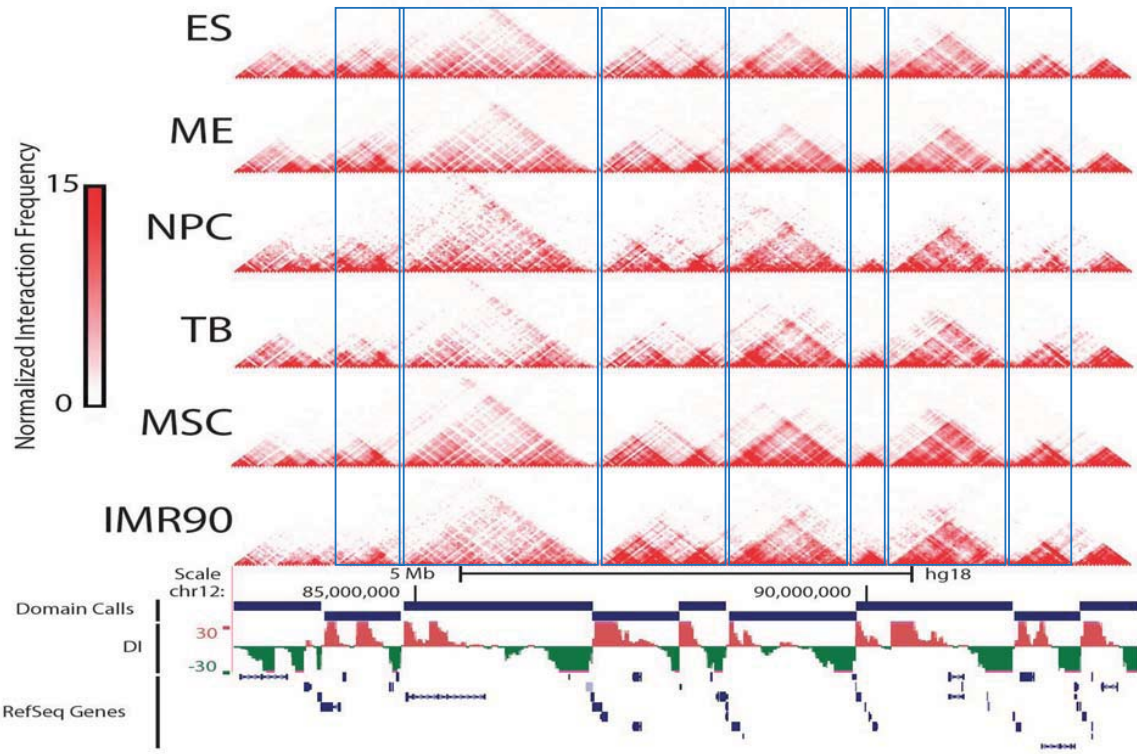
# Topologically Associating Domains (TADs)



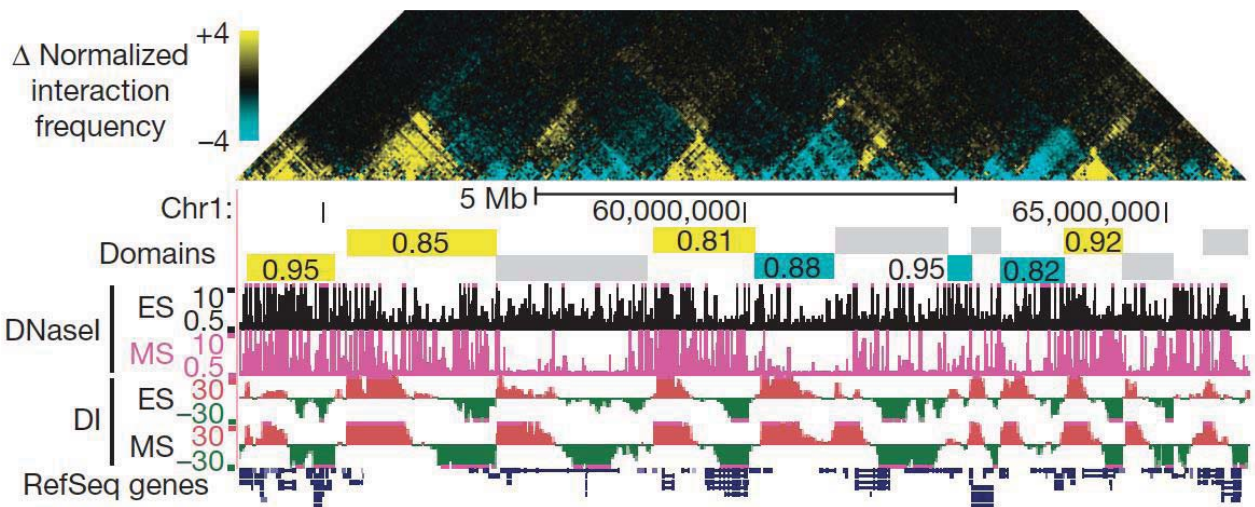
# Methods to define TAD boundaries



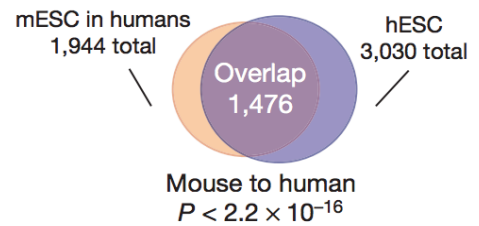
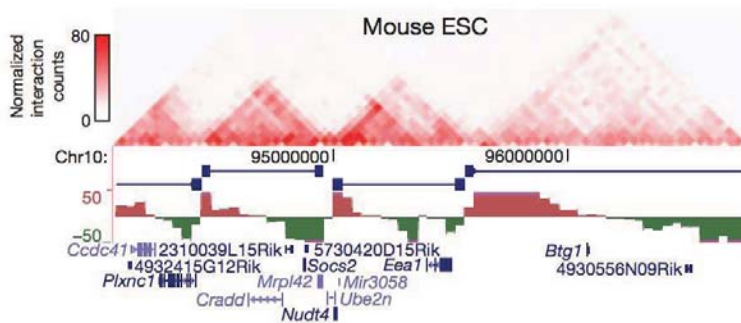
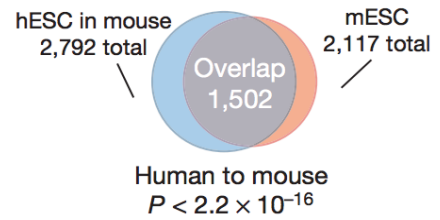
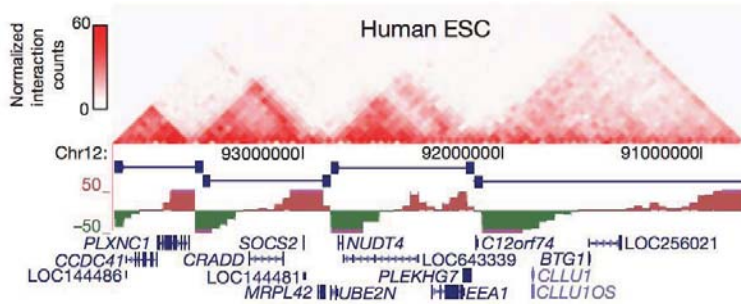
## TAD boundaries are well maintained during differentiation



## TAD-wise interaction changes during cellular differentiation

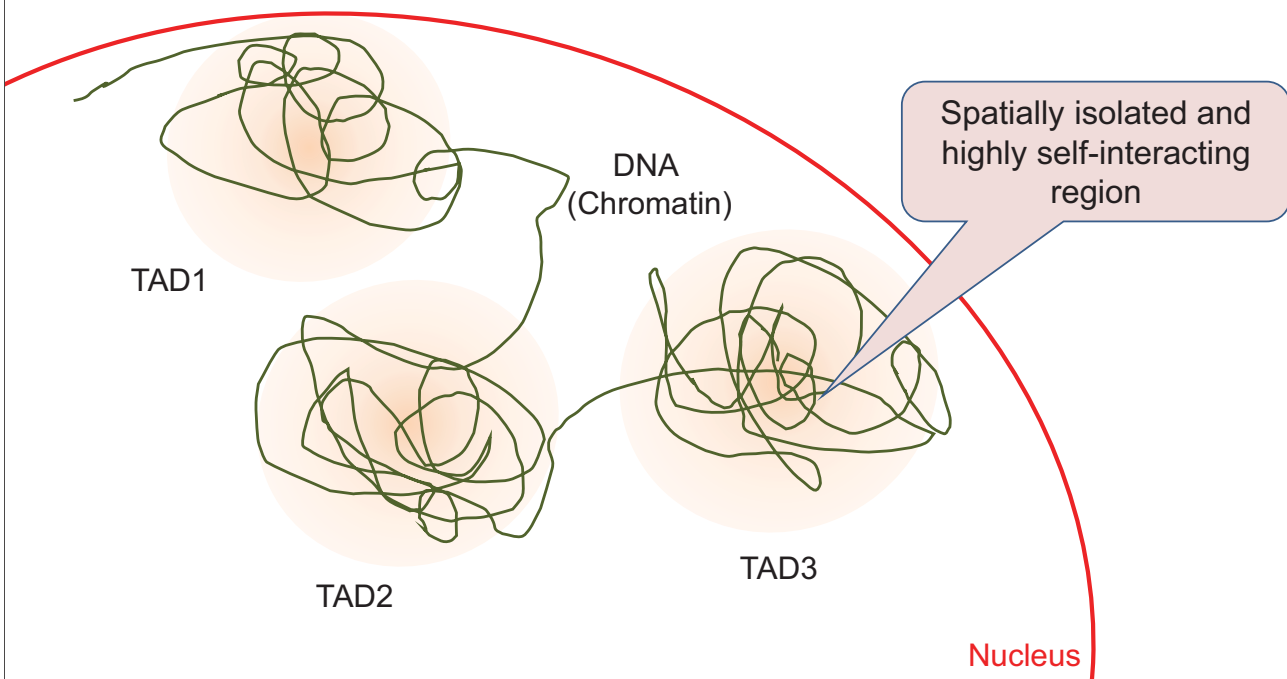


## TAD boundaries are evolutionarily well conserved



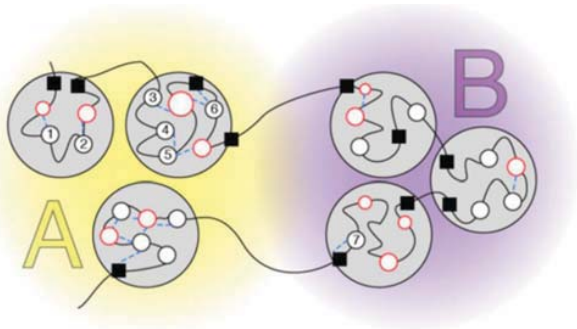
## TAD is a basic unit of 3D chromatin structure

1. The human genome is organized into 2000~3000 TADs
2. TAD boundaries are well maintained during cellular differentiation and evolution
3. However, within TAD interactions are dynamic in cell-type specific manner

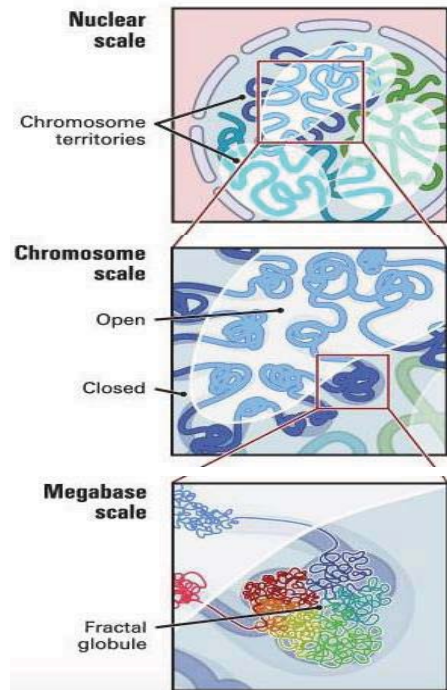
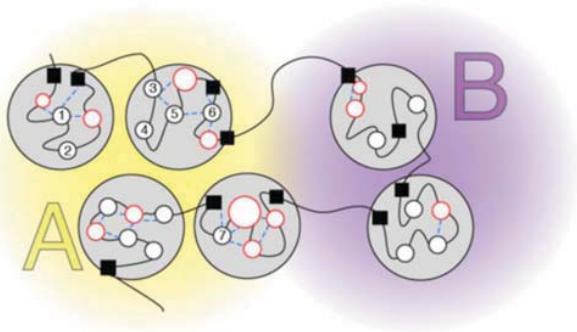


## What is a relationship between TAD and Compartment A/B?

I

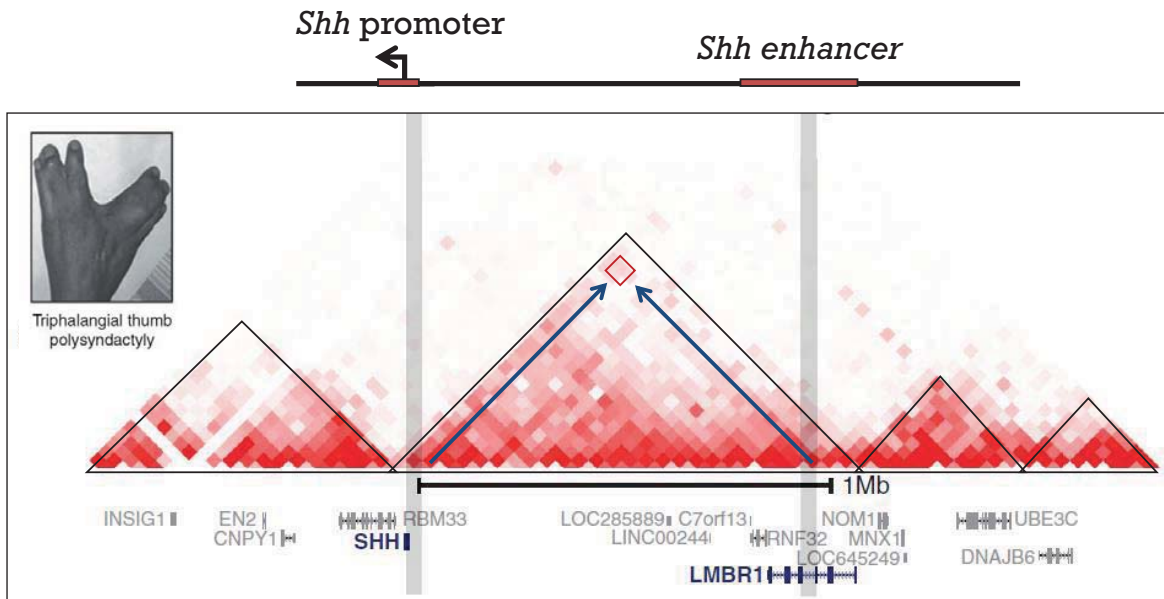


II



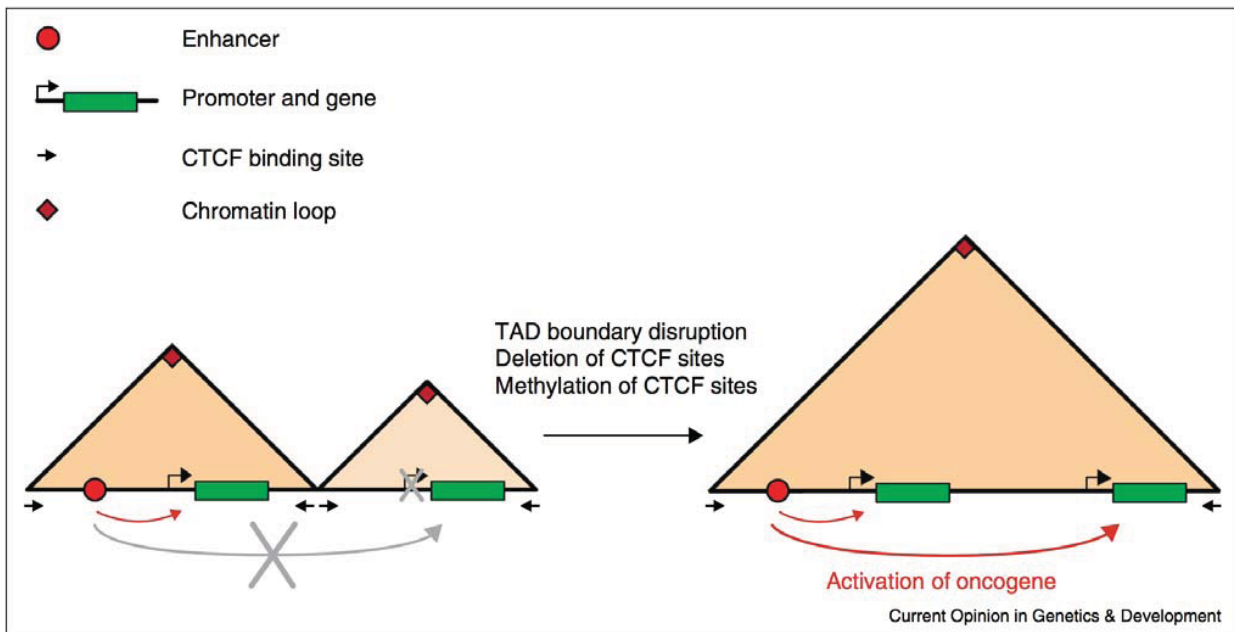
**What is a functional role of TADs?**

# TAD boundary restricts long-range enhancer controls



From Dixon et al, Nature (2012) and Smallwood et al, Current Opinion Cell Biology (2013)

# TAD boundary disruption as oncogenic driver – Model 1



Current Opinion in Genetics & Development



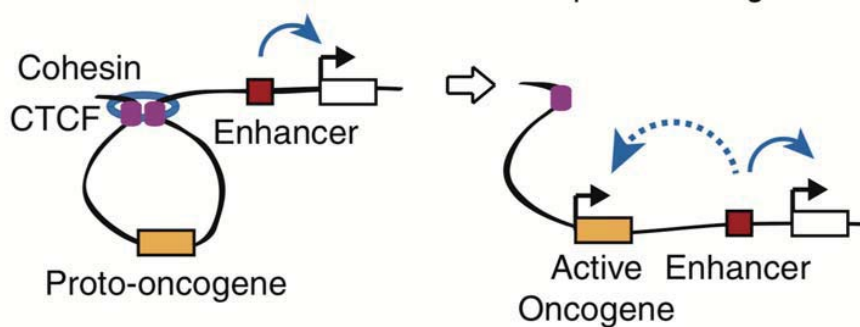
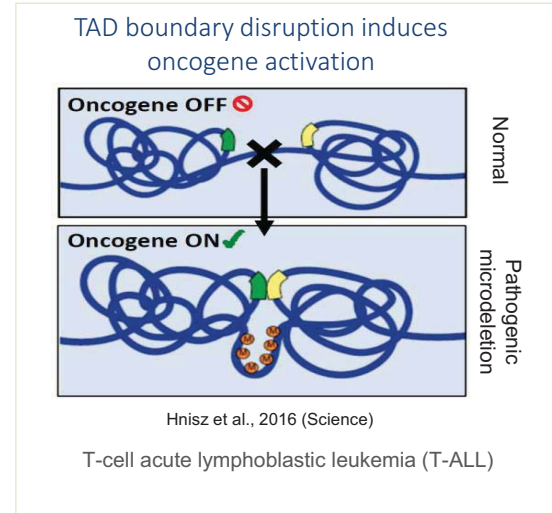
# Activation of proto-oncogenes by disruption of TAD boundary

CANCER

## Activation of proto-oncogenes by disruption of chromosome neighborhoods

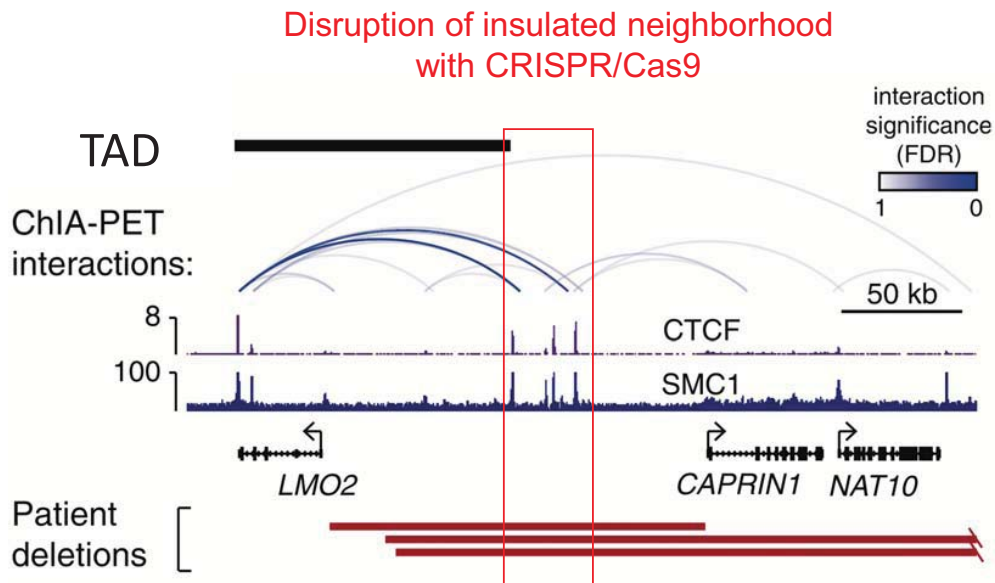
Denes Hnisz,<sup>1\*</sup> Abraham S. Weintraub,<sup>1,2\*</sup> Daniel S. Day,<sup>1</sup> Anne-Laure Valton,<sup>3</sup> Rasmus O. Bak,<sup>4</sup> Charles H. Li,<sup>1,2</sup> Johanna Goldmann,<sup>1</sup> Bryan R. Lajoie,<sup>3</sup> Zi Peng Fan,<sup>1,5</sup> Alla A. Sigova,<sup>1</sup> Jessica Reddy,<sup>1,2</sup> Diego Borges-Rivera,<sup>1,2</sup> Tong Ihn Lee,<sup>1</sup> Rudolf Jaenisch,<sup>1,2</sup> Matthew H. Porteus,<sup>4</sup> Job Dekker,<sup>3,6</sup> Richard A. Young<sup>1,2†</sup>

Oncogenes are activated through well-known chromosomal alterations such as gene fusion, translocation, and focal amplification. In light of recent evidence that the control of key genes depends on chromosome structures called insulated neighborhoods, we investigated whether proto-oncogenes occur within these structures and whether oncogene activation can occur via disruption of insulated neighborhood boundaries in cancer cells. We mapped insulated neighborhoods in T cell acute lymphoblastic leukemia (T-ALL) and found that tumor cell genomes contain recurrent microdeletions that eliminate the boundary sites of insulated neighborhoods containing prominent T-ALL proto-oncogenes. Perturbation of such boundaries in nonmalignant cells was sufficient to activate proto-oncogenes. Mutations affecting chromosome neighborhood boundaries were found in many types of cancer. Thus, oncogene activation can occur via genetic alterations that disrupt insulated neighborhoods in malignant cells.

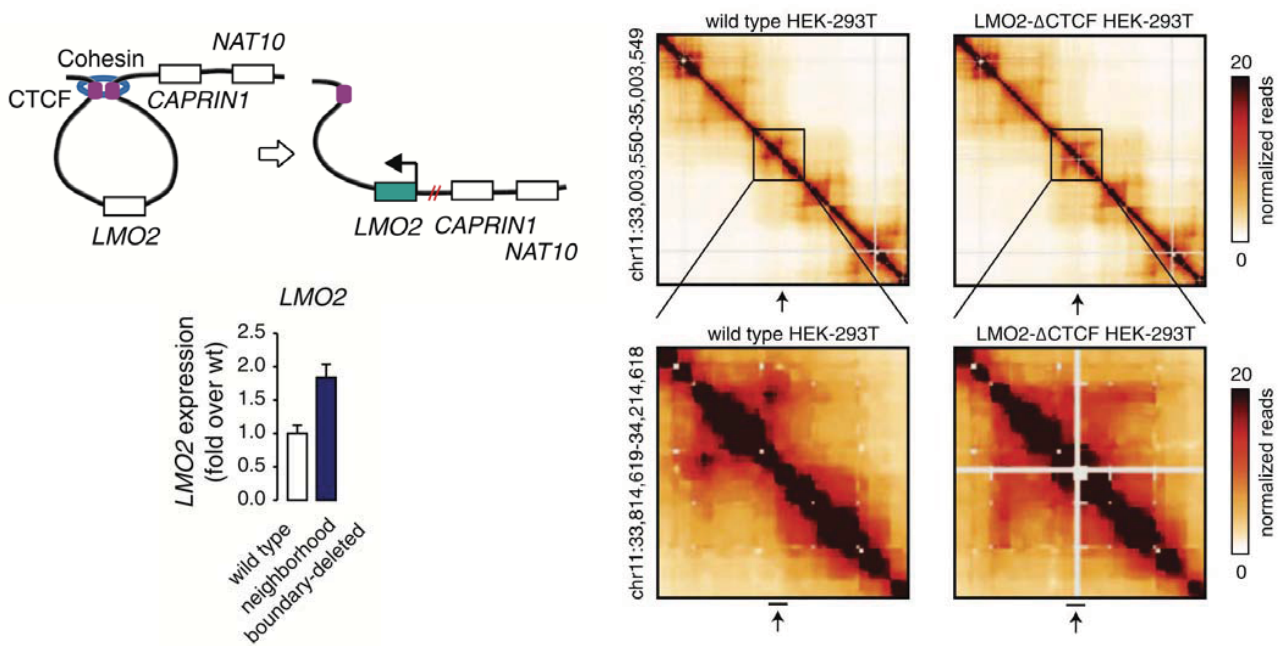


Can disruption of TAD boundary (TAD fusion) activate proto-oncogenes through **enhancer-hijacking**?

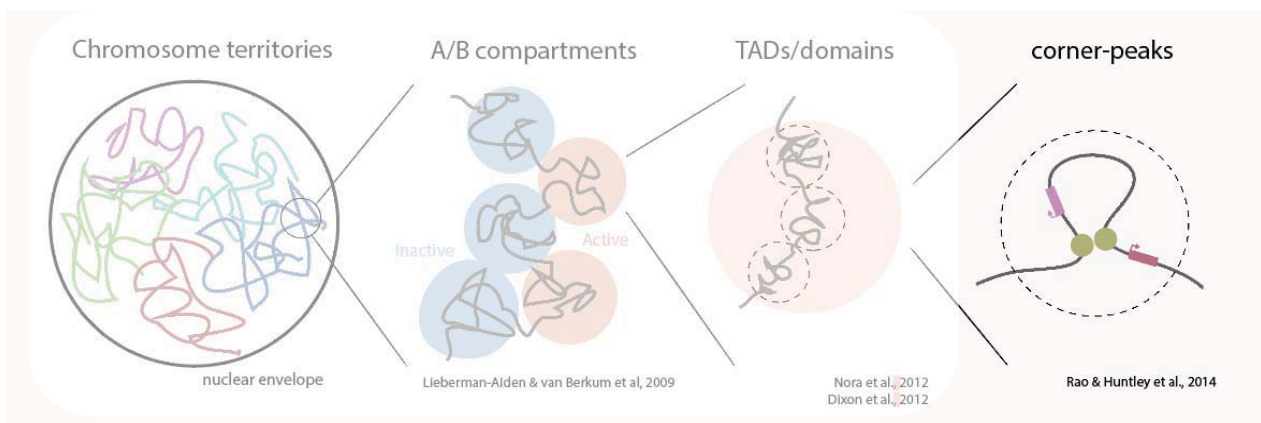
## Disruption of TAD boundary by CRISPR/Cas9



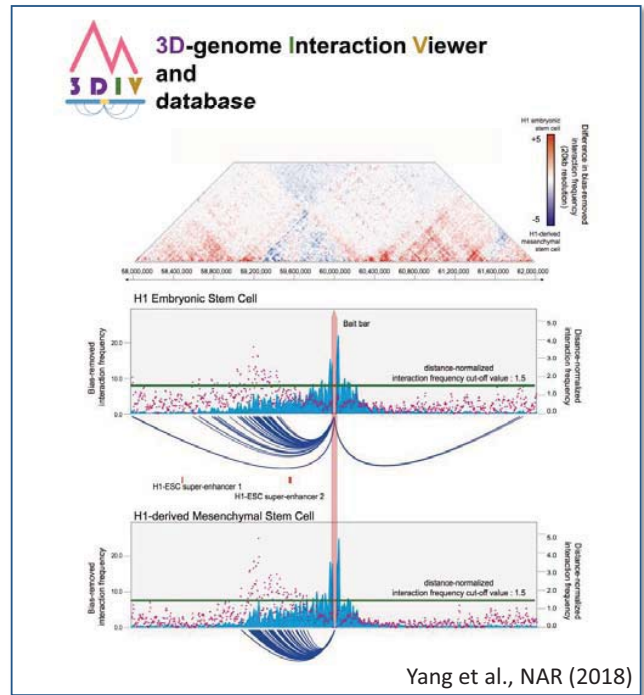
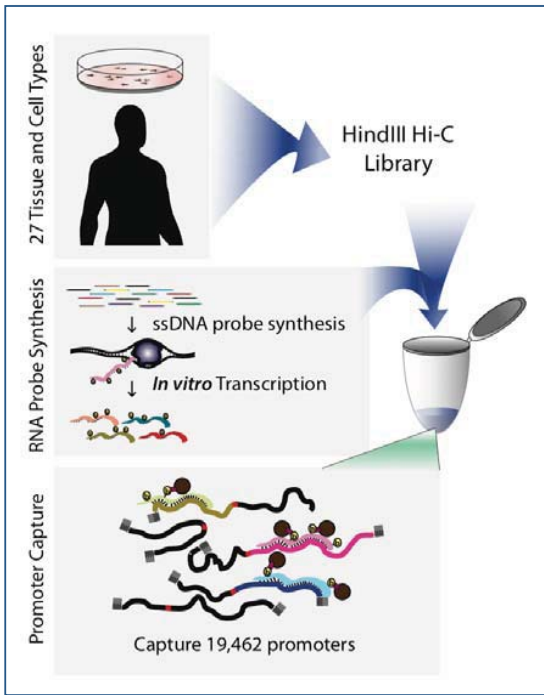
## Disruption of TAD boundary activates *LMO2*



1. Introduction to 3D genome
2. Methods to explore 3D genome
3. Compartment A/B
4. Topologically associating domains
- 5. Long-range chromatin interactions**



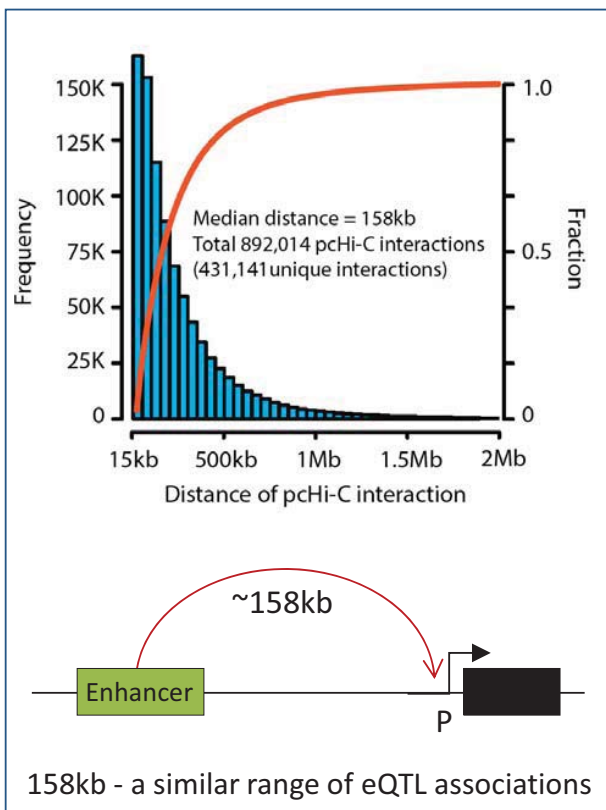
# Promoter-capture Hi-C: Enhancer-promoter interaction maps



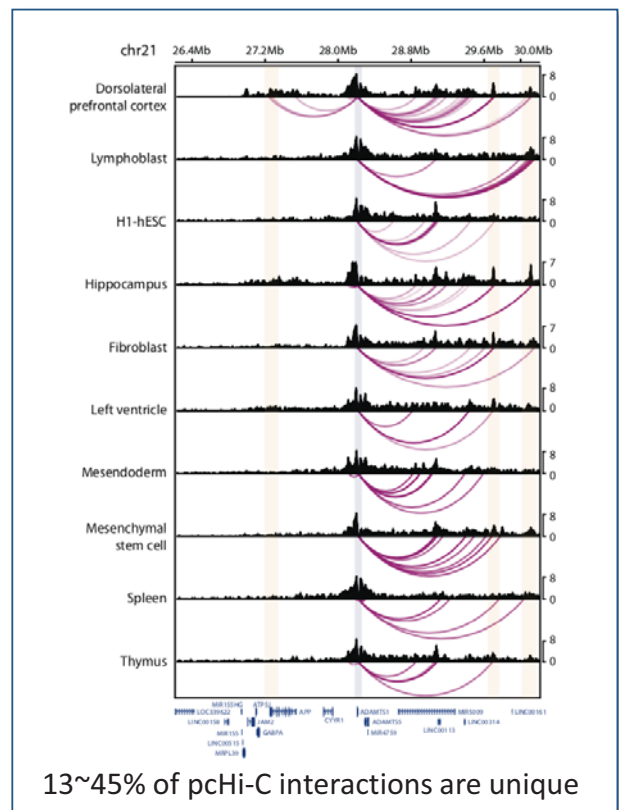
Jung et al., Nature Genetics (2019)

## Basic principles of enhancer-promoter interactions

### 1. Enrichment of distal interactions

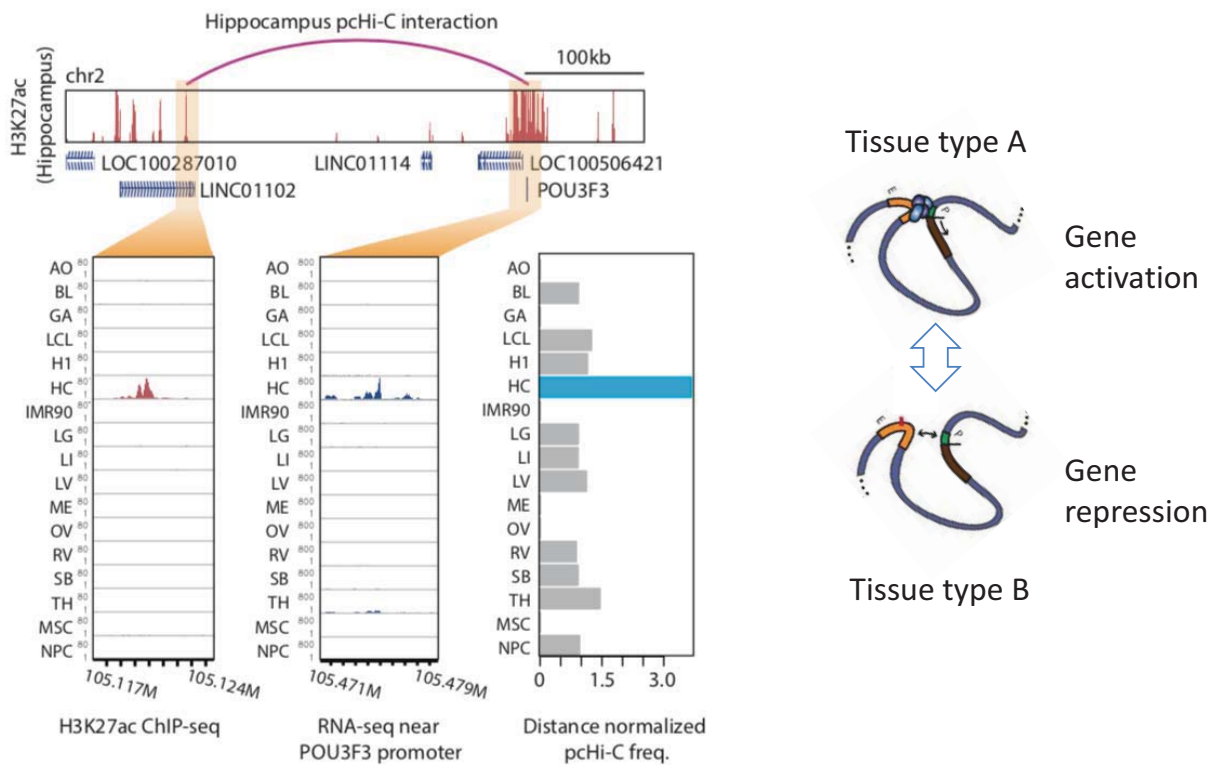


### 2. Interactions are tissue-specific



## Basic principles of enhancer-promoter interactions

### 3. E-P interactions correlate with tissue-specific gene expression



## Summary

- Genome is organized into multiple-layers
- TAD is a basic structural and functional unit of 3D chromatin structure
- Disruption TAD may potentiate disease-specific gene expression
- Long-range enhancer-promoter interactions are critical in cell/tissue-specific gene expression

# KSBi-BIML 2022

## 3D Epigenome in Gene Regulation (3DIV 기반 Hi-C 데이터 분석 실습)

정인경(KAIST)

### Contents

1. 후성유전체 및 ChIP-seq 개요
2. 염색질 3차구조 개요
- 3. 3DIV 기반 염색질 3차구조 및 유전자 조절 통합 분석 실습**
4. 염색질 3차구조 데이터 분석 실습



## ABOUT 3DIV

3D genome organization is tightly coupled with gene regulation in various biological processes and diseases. 3D Interaction Viewer and Database (3DIV) is a database providing chromatin interaction visualization in a variety of options from one-to-all chromatin interaction with epigenetic annotation to unique dynamic browsing tools allowing examination of large-scale genomic rearrangement mediated impacts in cancer 3D genome. 3DIV will be the most comprehensive resource to explore gene regulatory effects of both normal and cancer 3D genome.

### Hi-C

3DIV provides querying list of significant interacting partner locus, visualization, and comparative analysis of 3D chromatin interaction across about 400 samples.

### Capture Hi-C

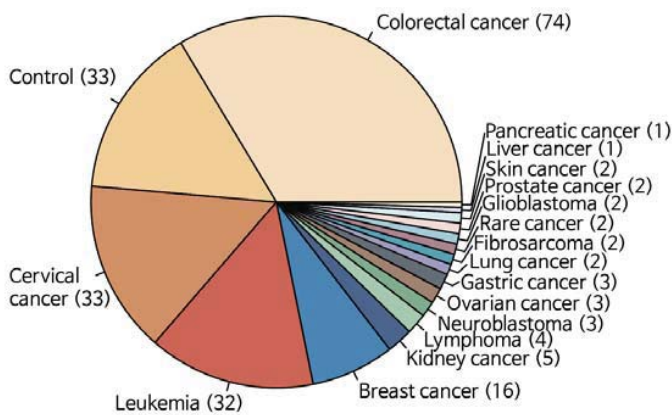
3DIV provides promoter capture Hi-C (pcHi-C) results across 28 normal human cell/tissue types, a great resource in identifying target genes of disease-associated genetic variants.

### Cancer Hi-C

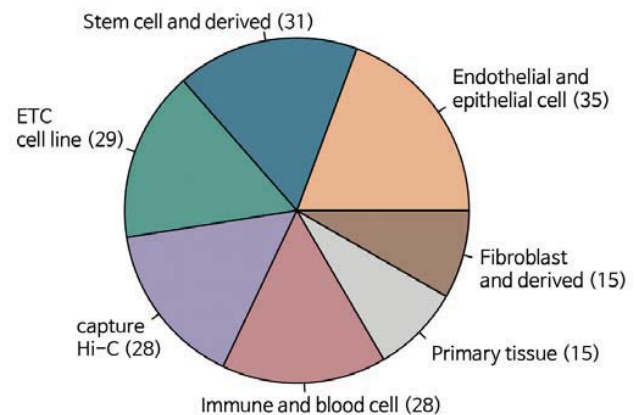
3DIV provides unique visualization and manipulation tools that allows user to generate rearranged 3D genome by selecting listed SVs, creating own SVs, and providing order of rearranged chromosomes.

## Hi-C data collection in 3DIV

Cancer Hi-C sample types (n = 220)



Normal Hi-C sample types (n = 181)



**Table 1.** Comparison of the updated 3DIV and other 3D genome databases as of October 2020

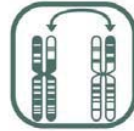
Software	Number of samples <sup>a</sup>	Hi-C contact map	TAD annotation	One-to-all interaction	Interaction table	Distance normalization	Interactive Hi-C contact map browsing	Live manipulation of genomic rearrangement	Structural variation annotation
3DIV 2021 Update	401	✓	✓	✓	✓	✓	✓	✓	✓
3DIV	80	✓	✓	✓	✓	✓	✓		
4D Nucleome	337 <sup>b</sup>	✓					✓		
Nucleome Browser	138 <sup>c</sup>	✓					✓		
WashU Epigenome Browser	36 <sup>c,d</sup>	✓					✓		
HiView	2		✓	✓	✓	✓	✓		
HUGIn <sup>2</sup>	83	✓	✓	✓	✓	✓			
3D Genome Browser	113	✓	✓						
GITAR	20 <sup>e</sup>	✓	✓						
Hi-C Data Browser	69	✓		✓					

## Unique functionalities of 3DIV

### Features of 3DIV



187 cancer/tumor tissue samples with 33 control samples



Pan-cancer SV data for corresponding cancer type



153 cell line/tissue Hi-C and 28 promoter capture Hi-C data



MySQL + Java Spring + HTML5 based webserver implementation



230 billion reads processed and normalized Hi-C contact maps



Interactive visualization function on web page

## Normal Hi-C Analysis



# Normal Hi-C Analysis



[Hi-C](#)
[Capture Hi-C](#)
[Cancer Hi-C](#)
[Statistics](#)
[Download](#)
[Tutorial](#)
[Contact Us](#)



## Normal Hi-C Analysis

### ABOUT 3DIV

3D genome organization is tightly coupled with gene regulation in various biological processes and diseases. 3D Interaction Viewer and Database (3DIV) is a database providing chromatin interaction visualization in a variety of options from one-to-all chromatin interaction with epigenetic annotation to unique dynamic browsing tools allowing examination of large-scale genomic rearrangement mediated impacts in cancer 3D genome. 3DIV will be the most comprehensive resource to explore gene regulatory effects of both normal and cancer 3D genome.

#### Hi-C

3DIV provides querying list of significant interacting partner locus, visualization, and comparative analysis of 3D chromatin interaction across about 400 samples.

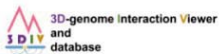
#### Capture Hi-C

3DIV provides promoter capture Hi-C (pcHi-C) results across 28 normal human cell/tissue types, a great resource in identifying target genes of disease-associated genetic variants.

#### Cancer Hi-C

3DIV provides unique visualization and manipulation tools that allows user to generate rearranged 3D genome by selecting listed SVs, creating own SVs, and providing order of rearranged chromosomes.

# Normal Hi-C Analysis



hg19
[Hi-C](#)
[Capture Hi-C](#)
[Cancer Hi-C](#)
[Statistics](#)
[Download](#)
[Tutorial](#)
[Contact Us](#)



Interaction Table

Interaction Visualization

Comparative Visualization

[Interaction table](#)
[Interaction visualization](#)
[Comparative interaction visualization](#)

> Choose sample(s)

Choose sample(s) by characteristics | Choose sample(s) by search | Choose sample(s)

> Type: Choose... |
 > Sample property: Choose... |
 > Condition: Choose... |
 > Sample: Choose...

> Input bait

Bait:  (Ex: CROCCP2, chr22:27141000, rs42)

> Interaction range

2Mb

> Selected region(s)

Sample	Bait
<input type="checkbox"/>	

# Functionalities of normal Hi-C in 3DIV

## Interaction Table

- Bias-removed/distance-normalized Interaction frequency
- Disease-associated GWAS SNPs
- Promoter/Enhancer/super-enhancer annotation
- Histone ChIP-seq signal

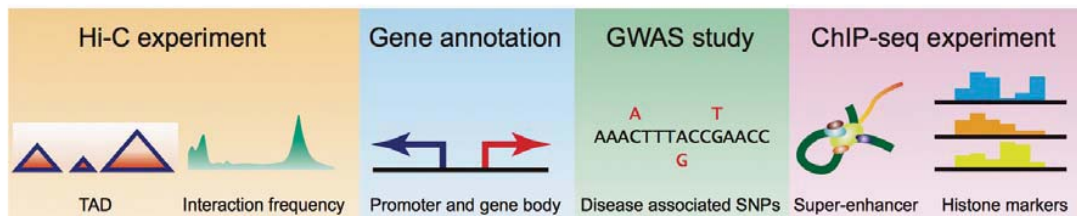
## Interaction Visualization

- Interaction frequency heatmap
- Topologically associating domains
- One-to-all interaction plot
- Arc-representation of significant interactions

## Comparative Visualization

- Comparative interaction frequency heatmap
- Synchronized interaction visualization

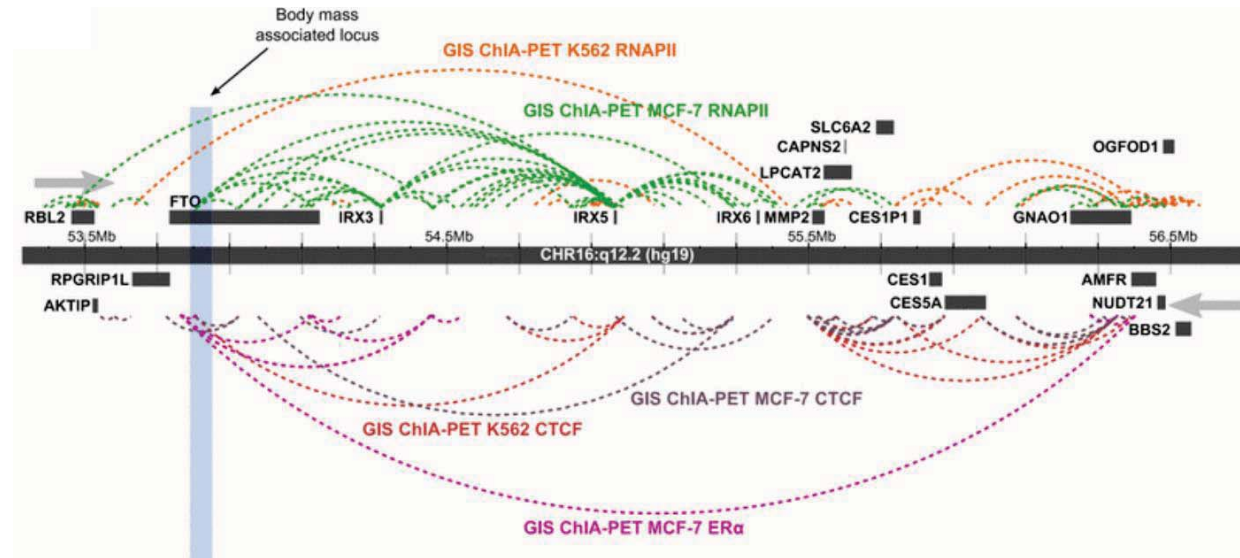
# Module 1 : Interaction Table



	Sample ▲▼	Locus (unit : kb) ▲▼	Bias-removed Interaction frequency ▲▼	Distance-normalized Interaction frequency ▼	Gene Name ▲▼	GWAS SNP ID ▲▼	Enhancer or Super-enhancer ▲▼	H3K27ac Fold change ▲▼	H3K4me1 Fold change ▲▼	H3K4me3 Fold change ▲▼
8 6 4 2 0	Mesenchymal Stem Cell	chr16:54965-54970	3.16	7.07	IRX5			2.6	4.15	23.9
	Mesenchymal Stem Cell	chr16:55505-55510	2.25	6.99			Enhancer	8.75	3.61	2.27
	Mesenchymal Stem Cell	chr16:55500-55505	2.01	6.24			Enhancer	6.41	4.15	3.66
	Mesenchymal Stem Cell	chr16:55540-55545	1.29	4.11	LPCAT2			1.54	5.61	11.77
	Mesenchymal Stem Cell	chr16:55510-55515	1.29	4.03	MMP2			2.07	5.61	23.5
	Mesenchymal Stem Cell	chr16:55535-55540	1.02	3.22			Enhancer	6.1	5.98	2.14
	Mesenchymal Stem Cell	chr16:55355-55360	1.06	3.01	IRX6			2.47	3.04	11.24
	Mesenchymal Stem Cell	chr16:54320-54325	2.56	2.88	IRX3			3.87	6.16	18.62
	Mesenchymal Stem Cell	chr16:55530-55535	0.86	2.73			Enhancer	6.63	5.39	2.8
	Mesenchymal Stem Cell	chr16:55515-55520	0.74	2.32	MMP2			1.34	3.73	1.75
	Mesenchymal Stem Cell	chr16:55310-55315	0.83	2.3			Enhancer	3.36	3.59	2.14
	Mesenchymal Stem Cell	chr16:52115-52120	0.73	2.24	LINC00919			1.33	1.24	1.22
	Mesenchymal Stem Cell	chr16:55705-55710	0.62	2.16	SLC6A2			1.12	1.79	2.27
	Mesenchymal Stem Cell	chr16:54375-54380	1.74	2.14			Enhancer	3.12	5.96	2.8
	Mesenchymal Stem Cell	chr16:55600-55605	0.62	2.05	CAPNS2			1.65	2.33	1.75
	Mesenchymal Stem Cell	chr16:55315-55320	0.74	2.05			Enhancer	6.31	4.15	1.75
	Mesenchymal Stem Cell	chr16:54490-54495	0.89	1.29		rs9921518		1.54	2.88	1.75

## Example : Interaction profile of rs1421085

rs1421085 : an obesity variant in FTO gene intron region.  
It is well characterized by significant interactions with IRX3 and IRX5 promoters.



Rask-Andersen et al, Hum. Genet. (2015)

## Step 1 : Open Interaction Table Module

3D-genome interaction Viewer  
and database

hg19 Hi-C Capture Hi-C Cancer Hi-C Statistics Download Tutorial Contact Us

Hi-C

Interaction table Interaction visualization Comparative interaction visualization

Choose sample(s)

Click "Interaction table"

Type: Choose... Sample property: Choose... Condition: Choose... Sample: Choose...

Input bait: Bait: (Ex: CROCCP2, chr22:27141000, rs42)

Interaction range: 2Mb

Add sample(s) Remove sample(s)

Selected region(s)

Sample	Bait
<input type="checkbox"/>	

Example Run Run

## Step 2 : Choose a sample

Interaction Table | Interaction visualization | Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics | Choose sample(s) by search | Choose sample(s)

Type: Stem cell and derived (3) | Sample property: H1 MSCs, Mesenchymal | Condition: No treatment (1) | Sample: H1\_Mesenchymal\_SCs

1) Choose samples with condition

Interaction table | Interaction visualization | Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics | Choose sample(s) by search | Choose sample(s)

Sample: Here is the placeholder  
H1 MSC  
H1 Mesenchymal Stem Cell  
H1 Mesendoderm Cell

Input bait | Interaction range

2) Choose samples with searching window

Interaction table | Interaction visualization | Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics | Choose sample(s) by search | Choose sample(s)

- fibroblast(CRL-2522) dexamethasone 40h
- fibroblast(CRL-2522) dexamethasone 48h
- fibroblast(CRL-2522) dexamethasone 56h
- GM23240 (primary skin fibroblasts)
- H1 Mesenchymal Stem Cell
- H1 Mesendoderm Cell
- H1 Neuronal Progenitor Cell
- H1 Trophoblast Cell
- H9 Human Embryonic Stem Cell Line, Heat shock condition
- H9 Human Embryonic Stem Cells
- H9 Human ESC-derived Neuroectodermal Cells
- TADP1 (near-hanoini rat line)

3) Choose samples from the list directly

## Step 3 : Choose a bait

Interaction table | Interaction visualization | Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics | Choose sample(s) by search | Choose sample(s)

Type: Stem cell and derived (3) | Sample property: H1 MSCs, Mesenchymal | Condition: No treatment (1) | Sample: H1\_Mesenchymal\_SCs

Input bait: Bait: rs1421085 (Ex: CROCCP2, chr22:27141000, rs42)

Interaction range: 2Mb

Add sample(s) | Remove sample(s)

> Selected region(s)

	Sample	Bait
<input type="checkbox"/>		

Example Run | Run

## Step 3 : Choose a bait

Interaction table | Interaction visualization | Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics | Choose sample(s) by search | Choose sample(s)

> Type  
Stem cell and derived

Find genomic location from Gene Symbol or SNP id

Gene symbols

rs1421085 | chr16 : 53,767,042 ~ 53,767,042

Close

> Input bait  
Click to confirm the coordinate of variant.  
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range  
2Mb

Add sample(s) Remove sample(s)

> Selected region(s)

<input type="checkbox"/>	Sample	Bait
<input type="checkbox"/>		

Example Run Run

## Step 4 : Run Module

Interaction table | Interaction visualization | Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics | Choose sample(s) by search | Choose sample(s)

> Type  
Stem cell and derived (3)

> Sample property  
H1 MSCs, Mesenchymal

> Condition  
No treatment (1)

> Sample  
H1\_Mesenchymal\_SCs

> Input bait  
Bait: rs1421085  
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range  
2Mb

Add sample(s) Remove sample(s)

> Selected region(s)

<input type="checkbox"/>	Sample	Bait
<input checked="" type="checkbox"/>	H1_Mesenchymal_SCs	chr16:53767042

Example Run Run

Click to run module

## Step 5 : Browse the table

### Epigenomics

#### Filter

Distance normalized Interaction frequency:  -

**Filter Run**

Show  entries

- 1) Bias-removed interaction frequency
- 2) Distance normalized interaction frequency
- 3) Annotation of Enhancer or Super-enhancer
- 4) Annotation of disease associated SNPs
- 5) Annotation of Promoter
- 6) CHIP-seq signals

Sample	Bin	Distance	Bias-removed Interaction frequency	Distance normalized Interaction frequency	Enhancer	GWAS SNP ID	Gene Name	H3K27ac	H3K27me3	H3K4me1	H3K4me3	H3K9me3	CTCF
H1 Mesenchymal Stem Cell	chr16:51815000-51820000	1950000	0.07	0.84	None	<a href="#">rs9935845</a>	-	1.76	1.85	1.64	1.92	1.82	0.00
H1 Mesenchymal Stem Cell	chr16:52510000-52515000	1255000	0.04	0.74	None	<a href="#">rs9933638</a>	-	2.01	2.06	2.28	1.47	1.41	0.00
H1 Mesenchymal Stem Cell	chr16:53495000-53500000	270000	1.14	0.82	None	<a href="#">rs9931702</a>	-	3.02	2.48	1.85	1.92	2.02	0.00
H1 Mesenchymal Stem Cell	chr16:53990000-53995000	225000	2.37	1.16	None	<a href="#">rs9924983</a>	-	2.01	2.27	2.71	2.38	1.82	0.00
H1 Mesenchymal Stem Cell	chr16:53800000-53805000	35000	5.78	0.61	None	<a href="#">rs9922619</a> ; <a href="#">rs9920506</a>	-	1.76	1.64	2.07	1.46	2.84	0.00
H1 Mesenchymal Stem Cell	chr16:54465000-54470000	700000	0.99	1.18	None	<a href="#">rs9921518</a>	-	3.52	2.06	3.78	1.92	1.82	0.00
H1 Mesenchymal Stem Cell	chr16:53015000-53020000	750000	0.10	0.67	None	<a href="#">rs9302592</a>	-	2.77	2.48	1.64	1.46	1.41	0.00
H1 Mesenchymal Stem Cell	chr16:53490000-53495000	275000	0.82	0.70	None	<a href="#">rs8057808</a>	AKTIP	2.01	1.85	1.85	2.38	4.68	0.00

## Step 5a : Adjust the table

### Epigenomics

Filter

Distance normalized Interaction frequency:  -

**Filter Run**

Show  entries

Sample	Bin	Distance	Bias-removed Interaction frequency	Distance normalized Interaction frequency	Enhancer	GWAS SNP ID	Gene Name	H3K27ac
H1 Mesenchymal Stem Cell	chr16:51815000-51820000	1950000	0.07	0.84	None	<a href="#">rs9935845</a>	-	1.76
H1 Mesenchymal Stem Cell	chr16:52510000-52515000	1255000	0.04	0.74	None	<a href="#">rs9933638</a>	-	2.01
H1 Mesenchymal Stem Cell	chr16:53495000-53500000	270000	1.14	0.82	None	<a href="#">rs9931702</a>	-	3.02
H1 Mesenchymal Stem Cell	chr16:53990000-53995000	225000	2.37	1.16	None	<a href="#">rs9924983</a>	-	2.01
H1 Mesenchymal Stem Cell	chr16:53800000-53805000	35000	5.78	0.61	None	<a href="#">rs9922619</a> ; <a href="#">rs9920506</a>	-	1.76
H1 Mesenchymal Stem Cell	chr16:54465000-54470000	700000	0.99	1.18	None	<a href="#">rs9921518</a>	-	3.52
H1 Mesenchymal Stem Cell	chr16:53015000-53020000	750000	0.10	0.67	None	<a href="#">rs9302592</a>	-	2.77
H1 Mesenchymal Stem Cell	chr16:53490000-53495000	275000	0.82	0.70	None	<a href="#">rs8057808</a>	AKTIP	2.01

Adjust the number of entries per page.

## Step 5b : Sort the interaction table

**Epigenomics**

**Filter**

Distance normalized Interaction frequency:  0.00 - 2.55

**Filter Run**

Show  entries

Sample	Bin	Distance	Bias-removed Interaction frequency	Distance normalized Interaction frequency	Enhancer	GWAS SNP ID	Gene Name	H3K27ac
H1 Mesenchymal Stem Cell	chr16:53805000-53810000	40000	5.85	0.69	None	<a href="#">rs72805613</a>	-	1.76
H1 Mesenchymal Stem Cell	chr16:53800000-53805000	35000	5.78	0.61	None	<a href="#">rs9922619</a> <a href="#">rs9930506</a>	-	1.76
H1 Mesenchymal Stem Cell	chr16:53845000-53850000	80000	5.78	1.17	None	NA	-	1.87
H1 Mesenchymal Stem Cell	chr16:53840000-53845000	75000	5.75	1.11	None	NA	-	2.77
H1 Mesenchymal Stem Cell	chr16:53850000-53855000	85000	5.73	1.22	None	NA	-	2.26
H1 Mesenchymal Stem Cell	chr16:53870000-53875000	105000	5.53	1.37	None	NA	-	2.77
H1 Mesenchymal Stem Cell	chr16:53835000-53840000	70000	5.49	1.01	None	NA	-	3.02
H1 Mesenchymal Stem Cell	chr16:53740000-53745000	25000	5.44	0.44	None	<a href="#">rs6499640</a>	-	1.76

Click the header to sort the table

## Step 5c : Filter interaction

**Epigenomics**

Drag to filter interaction by their strength in this case, 2.0 is the criteria.

**Filter**

Distance normalized Interaction frequency:  0.00 - 2.55

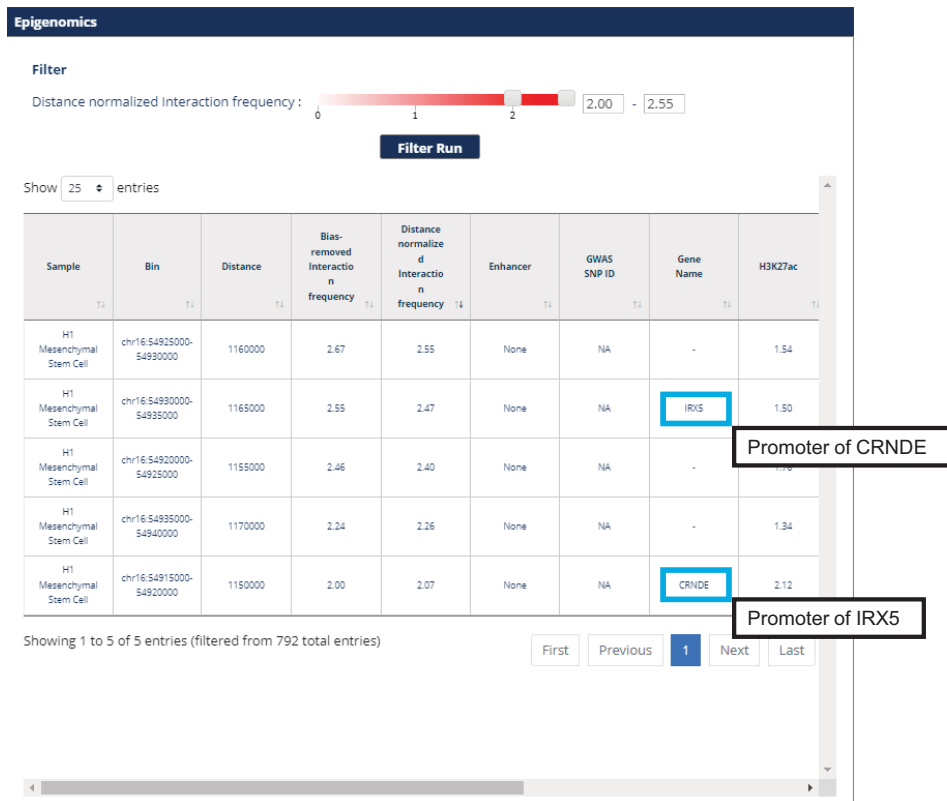
**Filter Run**

Show  entries

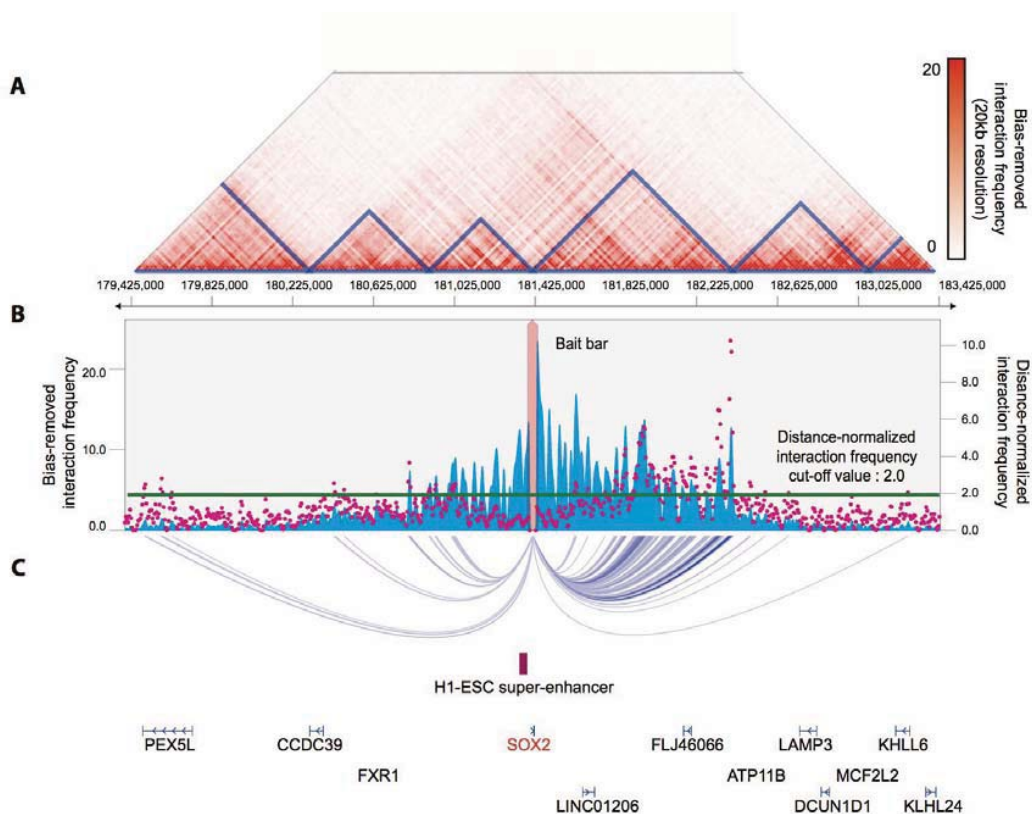
Click to apply the filter

Sample	Bin	Distance	Bias-removed Interaction frequency	Distance normalized Interaction frequency	Enhancer	GWAS SNP ID	Gene Name	H3K27ac
H1 Mesenchymal Stem Cell	chr16:53805000-53810000	40000	5.85	0.69	None	<a href="#">rs72805613</a>	-	1.76
H1 Mesenchymal Stem Cell	chr16:53800000-53805000	35000	5.78	0.61	None	<a href="#">rs9922619</a> <a href="#">rs9930506</a>	-	1.76
H1 Mesenchymal Stem Cell	chr16:53845000-53850000	80000	5.78	1.17	None	NA	-	1.87
H1 Mesenchymal Stem Cell	chr16:53840000-53845000	75000	5.75	1.11	None	NA	-	2.77
H1 Mesenchymal Stem Cell	chr16:53850000-53855000	85000	5.73	1.22	None	NA	-	2.26
H1 Mesenchymal Stem Cell	chr16:53870000-53875000	105000	5.53	1.37	None	NA	-	2.77
H1 Mesenchymal Stem Cell	chr16:53835000-53840000	70000	5.49	1.01	None	NA	-	3.02
H1 Mesenchymal Stem Cell	chr16:53740000-53745000	25000	5.44	0.44	None	<a href="#">rs6499640</a>	-	1.76

## Step 5 : Browse the table



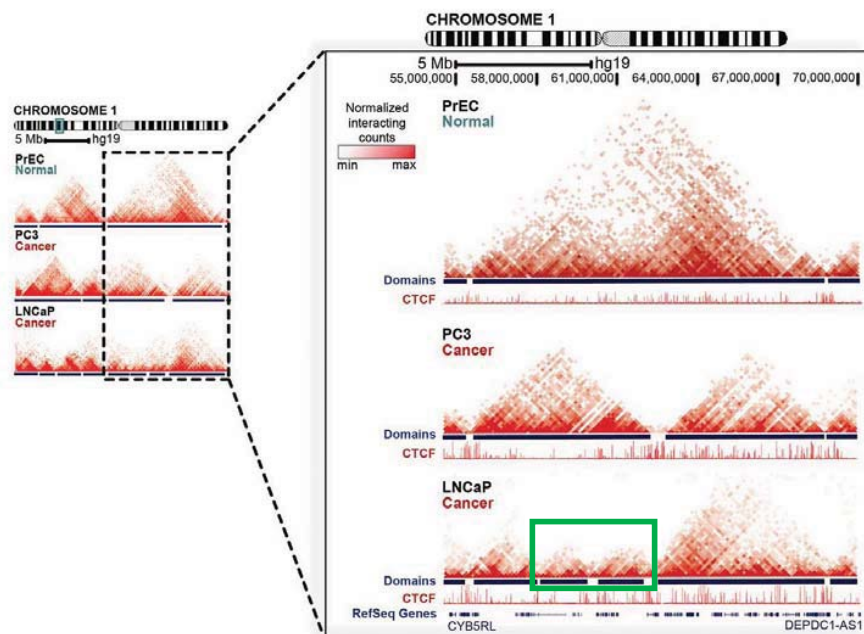
## Module 2 : Interaction Visualization





## Example : Interaction profile of SOX2

In cancer cells, the genomic structures are degraded into smaller sub-structures. In this session, we will reproduce this result with 3DIV.

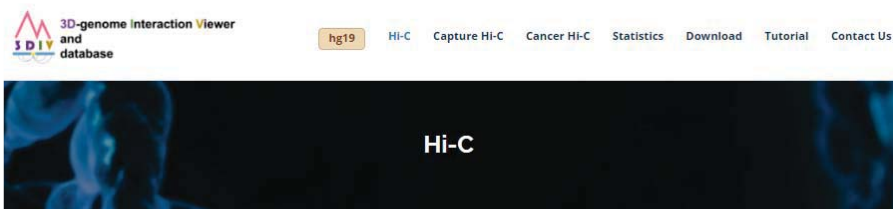


Taberlay et al, *Genome Res.* (2016)

## Step 1 : Open Interaction Visualization Module

The screenshot shows the 3DIV 3D-genome interaction viewer interface. The top navigation bar includes the 3DIV logo, the text "3D-genome interaction viewer and database", and a menu with options: hg19, Hi-C, Capture Hi-C, Cancer Hi-C, Statistics, Download, Tutorial, and Contact Us. Below the navigation bar is a banner image with the text "Hi-C". The main interface has three tabs: "Interaction table", "Interaction visualization" (highlighted with a red box and a hand icon), and "Comparative interaction visualization". A callout box with a hand icon points to the "Interaction visualization" tab with the text "Click 'Interaction visualization'". Below the tabs are several interactive panels: "Choose sample(s)" with sub-tabs for "by characteristics", "by search", and "sample(s)"; "Input bait" with a text field for "Bait:" (example: "CROCCP2, chr22:27141000, rs42"); "Interaction range" with a dropdown menu set to "2Mb"; "TAD" with a dropdown menu set to "DI (window size = 2Mb)"; "Add sample(s)" and "Remove sample(s)" buttons; and "Selected region(s)" with a table with columns for "Sample", "Bait", and "TAD". At the bottom are "Example Run" and "Run" buttons.

## Step 2 : Choose a sample



Interaction table | Interaction visualization | Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics | Choose sample(s) by search | **Choose sample(s)**

- A549 00h 100 nM dexamethasone
- A549 01h 100 nM dexamethasone
- A549 04h 100 nM dexamethasone
- A549 08h 100 nM dexamethasone
- A549 12h 100 nM dexamethasone
- ADAC418 (primary islet)
- Adrenal gland
- Aorta
- ASCs (Adipose-Derived Stem Cells), 0 day of differentiation induction
- ASCs (Adipose-Derived Stem Cells), 1 day after neuronal induction
- ASCs (Adipose-Derived Stem Cells), 1 day of differentiation induction
- ASCs (Adipose-Derived Stem Cells), 2 days before induction of differentiation
- ASCs (Adipose-Derived Stem Cells), 2 days after neuronal induction

**Click to load the list of Hi-C experiments**

> Input bait

Bait:   
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range

2Mb

> TAD

DI (window size = 2Mb)

Add sample(s) Remove sample(s)

## Step 2 : Choose a sample

Interaction table | Interaction visualization | Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics | Choose sample(s) by search | **Choose sample(s)**

- IMR90, in-situ Mbol
- K562, in-situ Mbol
- KBM7 cell line
- KBM7, in-situ Mbol
- Left Ventricle
- Liver
- LNCap prostate cancer cell line, BgIII**
- Lung
- MCF-10A
- MCF-10A BRG1 shRNA

**Click to choose sample**

> Input bait

Bait:   
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range

2Mb

> TAD

DI (window size = 2Mb)

Add sample(s) Remove sample(s)

### Step 3 : Choose a bait & TAD calling option

Interaction table   Interaction visualization   Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics   Choose sample(s) by search   Choose sample(s)

- IMR90, in-situ Mbol
- K562, in-situ Mbol
- KBM7 cell line
- KBM7, in-situ Mbol
- Left Ventricle
- Liver
- LNCap prostate cancer cell line, BgIII
- Lung
- MCF-10A
- MCF-10A BRG1 shRNA
- MCF-10A scramble shRNA
- MCF-7

> Input bait

Bait: chr1:60000000  
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range

2Mb

> TAD

DI (window size = 2Mb)

Add sample(s)   Remove sample(s)

> Selected region(s)

<input type="checkbox"/>	Sample	Bait	TAD
<input type="checkbox"/>			

Example Run   Run

Click button to adjust TAD calling option  
In this demo, DI-based caller with 2MB window is used

### Step 3 : Choose a Bait & TAD calling option

> Choose sample(s)

Choose sample(s) by characteristics   Choose sample(s) by search   Choose sample(s)

- IMR90, in-situ Mbol
- K562, in-situ Mbol
- KBM7 cell line
- KBM7, in-situ Mbol
- Left Ventricle
- Liver
- LNCap prostate cancer cell line, BgIII
- Lung
- MCF-10A
- MCF-10A BRG1 shRNA
- MCF-10A scramble shRNA
- MCF-7

> Input bait

Bait: chr1:60000000  
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range

2Mb

> TAD

DI (window size = 2Mb)

Add sample(s)   Remove sample(s)

> Selected region(s)

<input type="checkbox"/>	Sample	Bait	TAD
<input type="checkbox"/>	LNCap prostate cancer cell line, BgIII	chr1:60000000	DI (window size = 2Mb)

Example Run   Run

Click button to add sample

## Step 4 : Run Module

Interaction table   Interaction visualization   **Comparative interaction visualization**

> Choose sample(s)

Choose sample(s) by characteristics   Choose sample(s) by search   Choose sample(s)

- IMR90, in-situ Mbol
- K562, in-situ Mbol
- KBM7 cell line
- KBM7, in-situ Mbol
- Left Ventricle
- Liver
- LNCap prostate cancer cell line, BgIII
- Lung
- MCF-10A
- MCF-10A BRG1 shRNA
- MCF-10A scramble shRNA
- MCF-7

> Input bait

Bait : chr1:60000000  
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range

2Mb

> TAD

DI (window size = 2Mb)

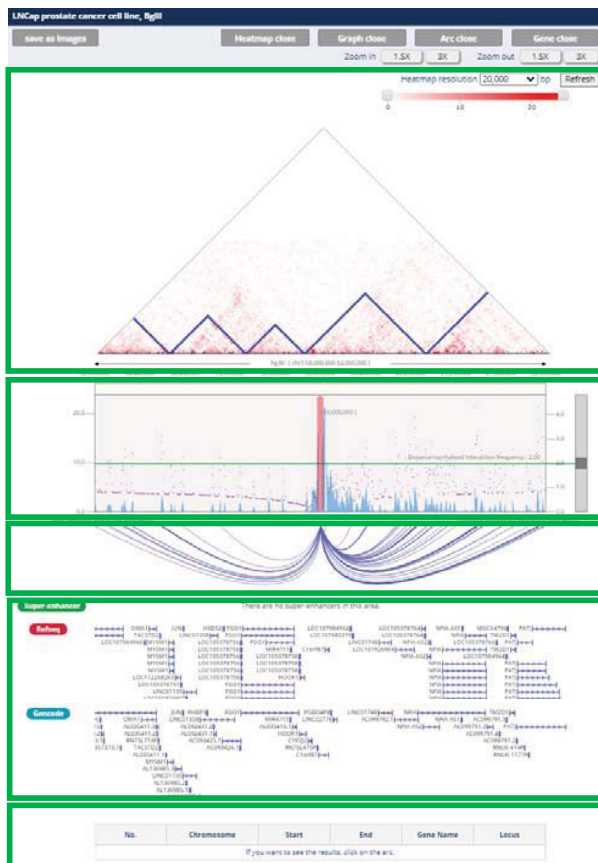
Add sample(s)   Remove sample(s)

> Selected region(s)

<input type="checkbox"/>	Sample	Bait	TAD
<input type="checkbox"/>	LNCap prostate cancer cell line, BgIII	chr1:60000000	DI (window size = 2Mb)

Example Run   Run

## Step 5 : Adjust the interaction visualization



Interaction frequency heatmap with topologically associating domains(TAD) annotation.

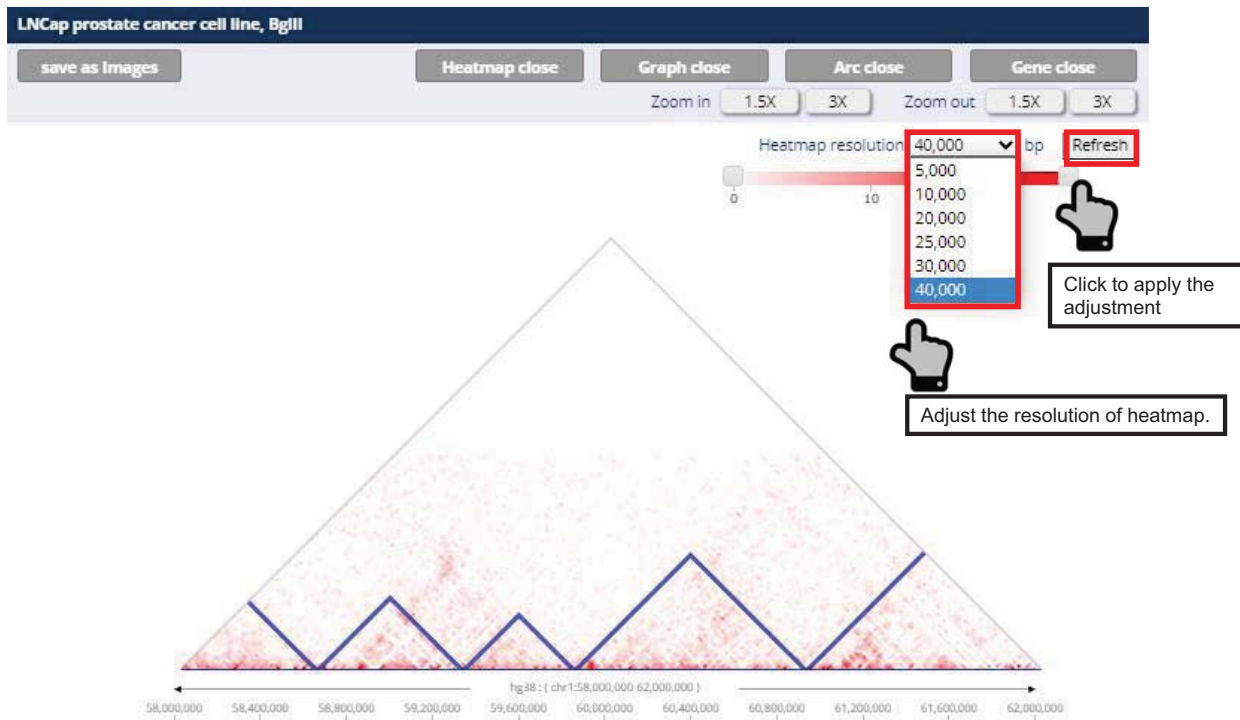
One-to-all interaction plot

Arc-representation of significant interactions

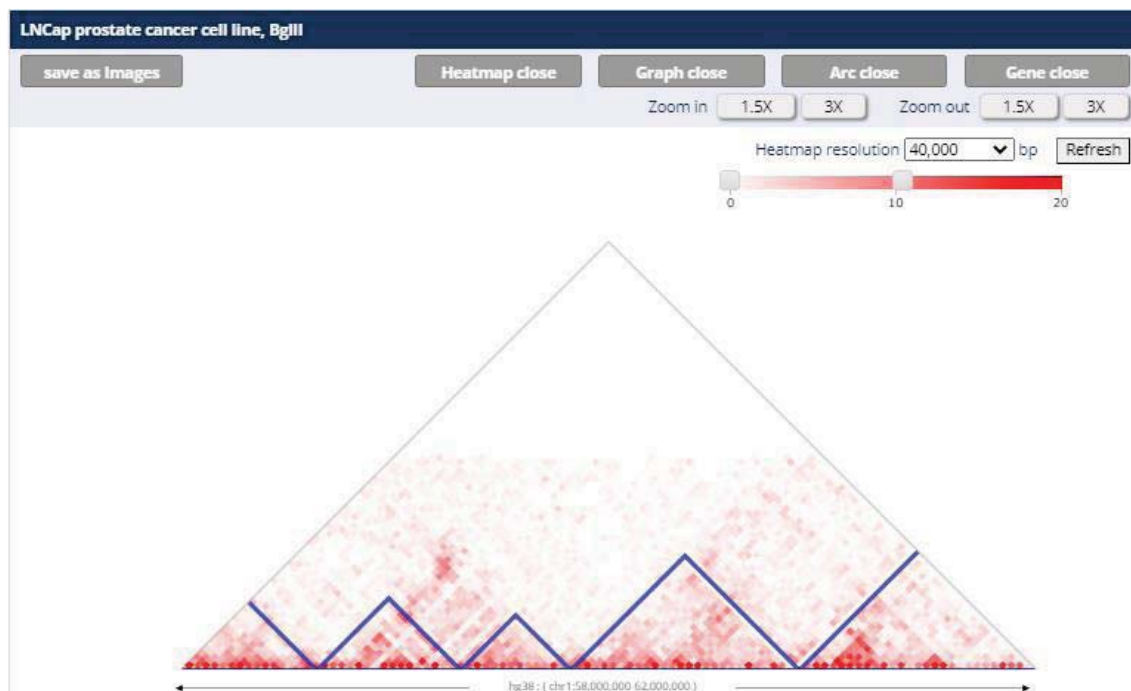
Gene annotations

Description of selected interaction

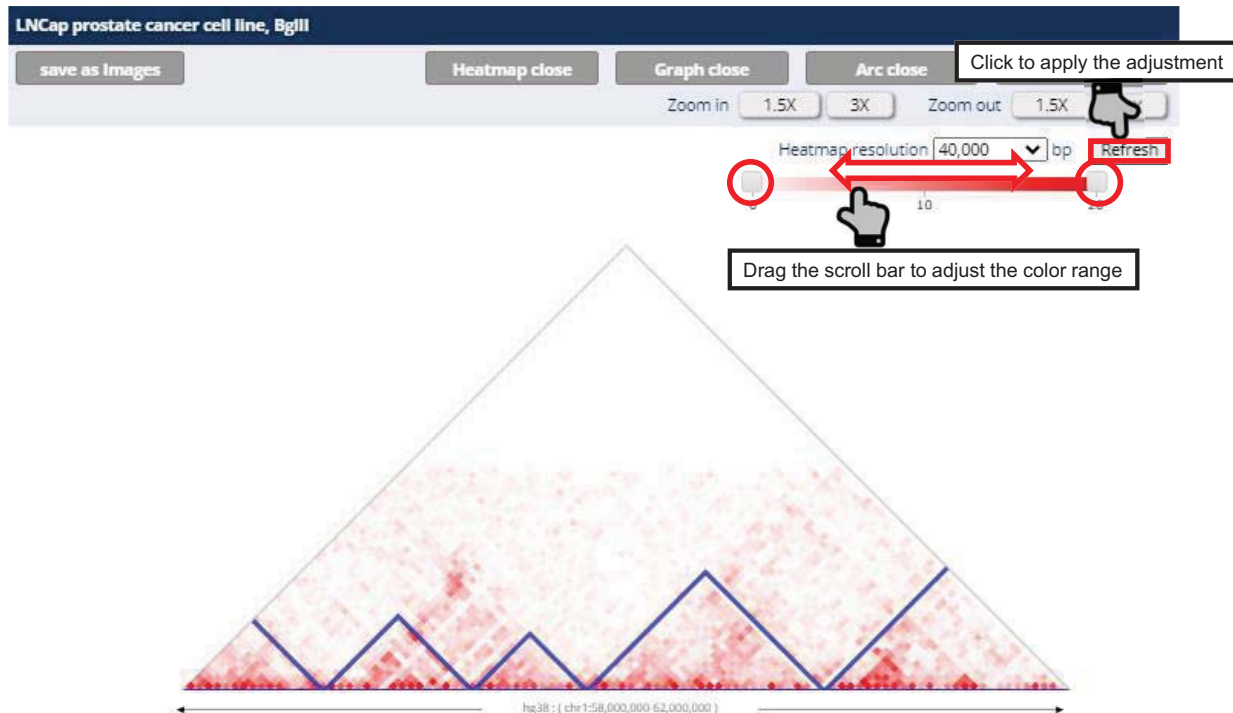
## Step 5a : Adjust the heatmap resolution



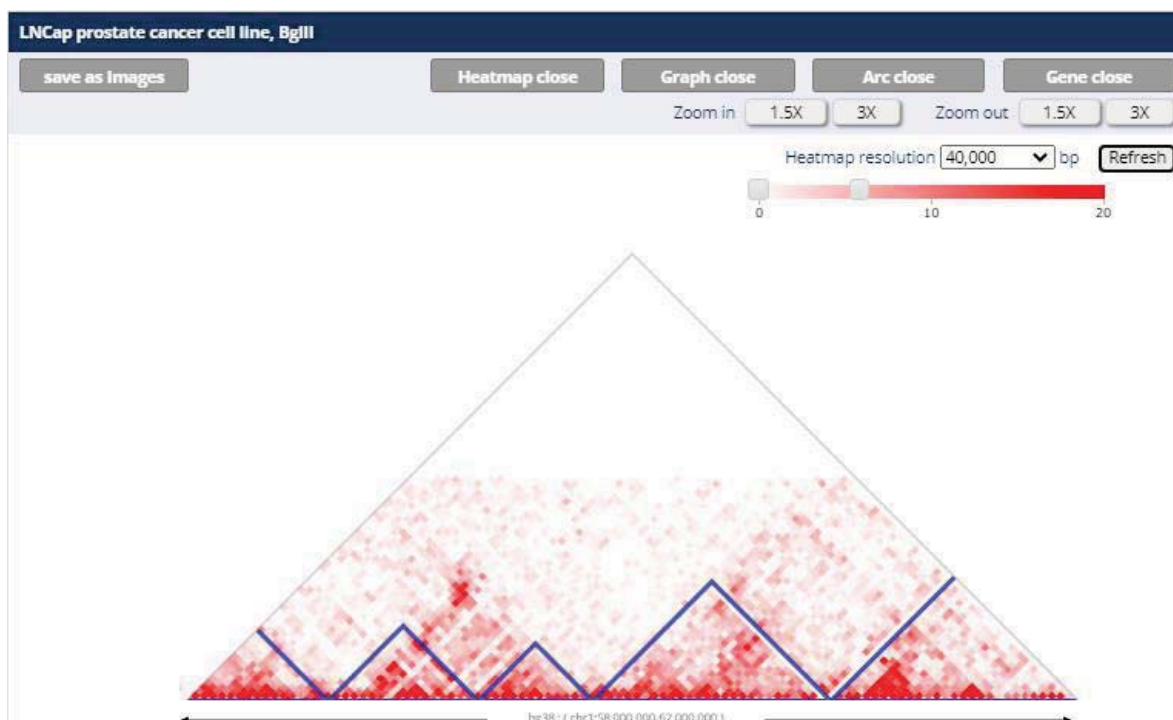
## Step 5a : Adjust the heatmap resolution



## Step 5b : Adjust the heatmap color range

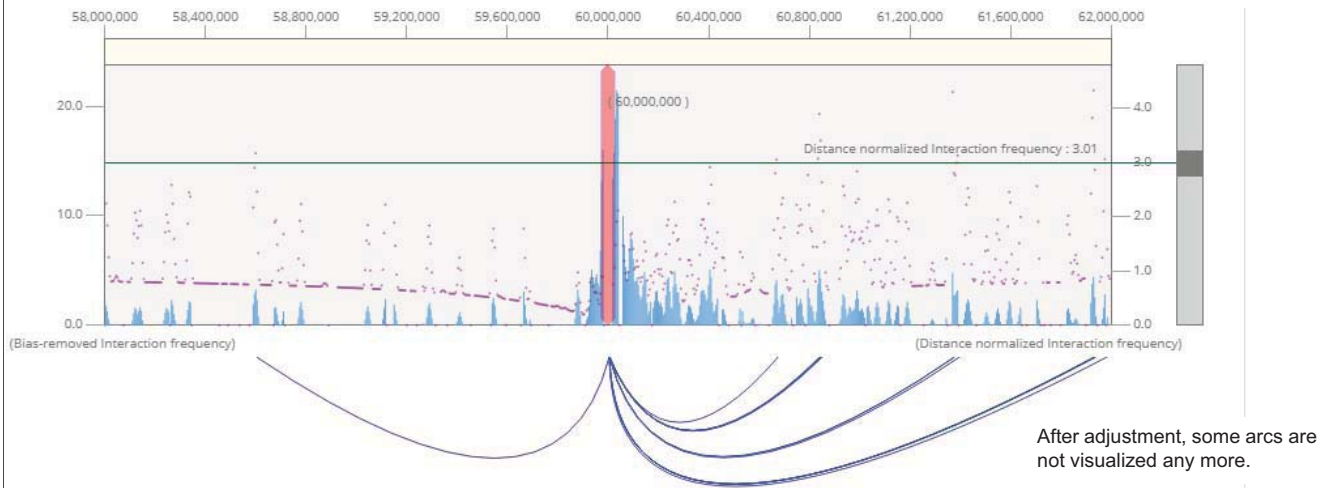


## Step 5b : Adjust the heatmap color range

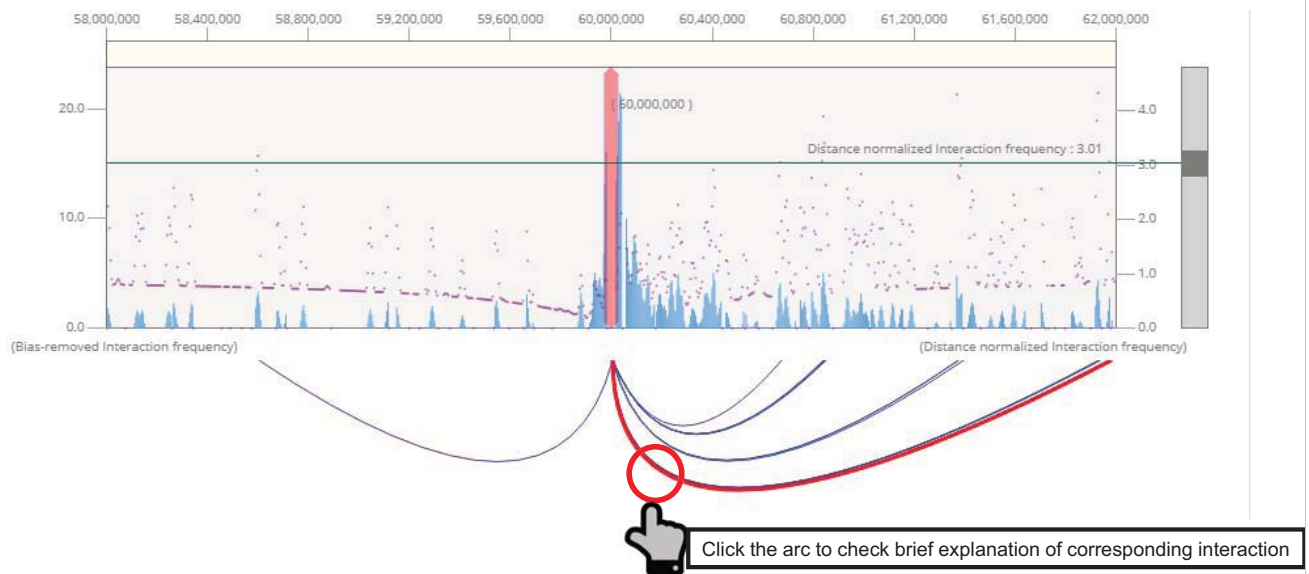




## Adjust the fold-change criteria



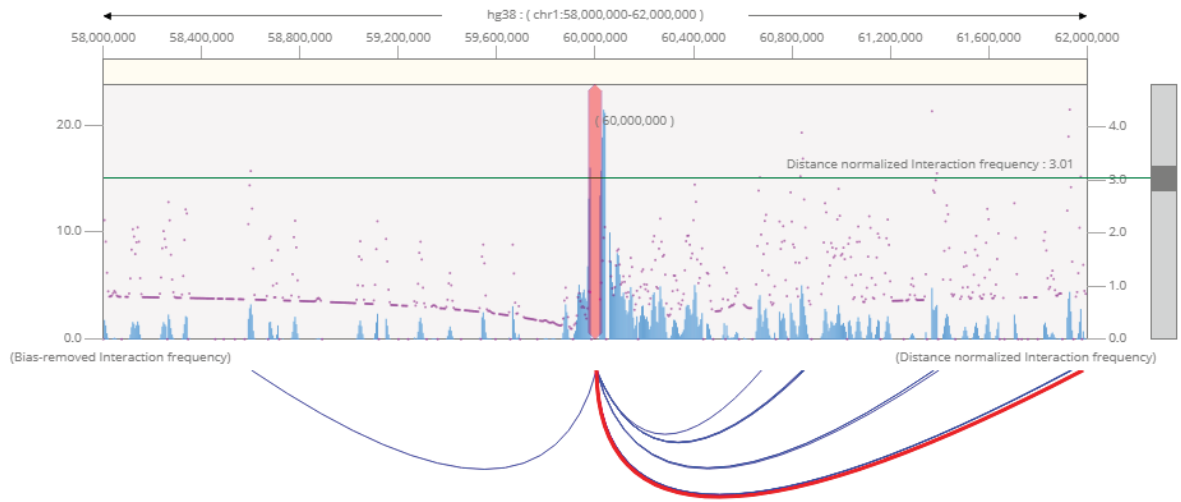
## Description of identified interactions



No.	Chromosome	Start	End	Gene Name	Locus
If you want to see the results, click on the arc.					

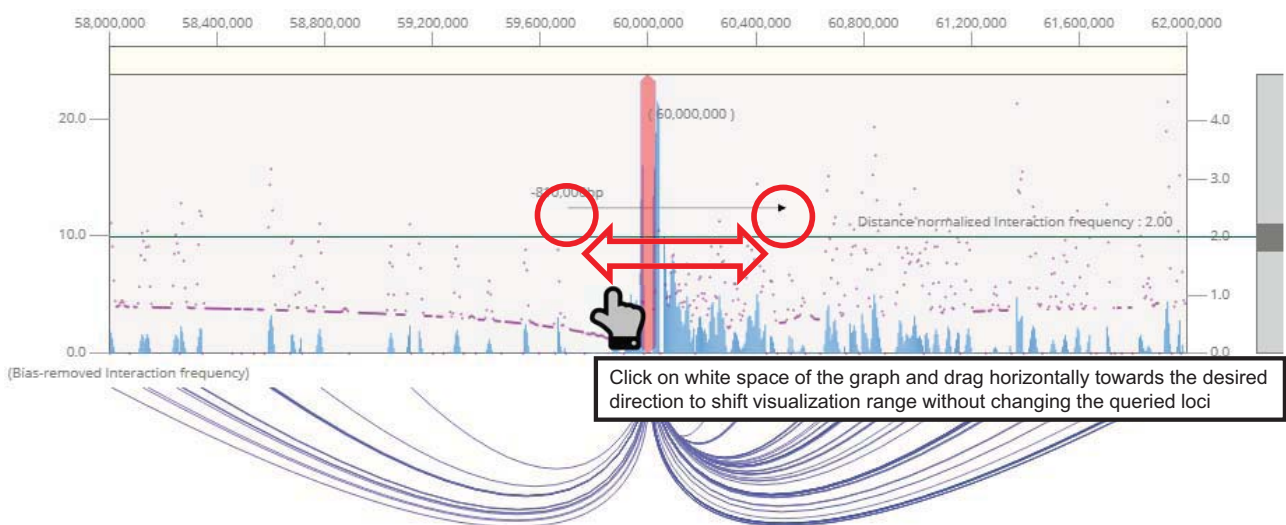


## Description of identified interactions

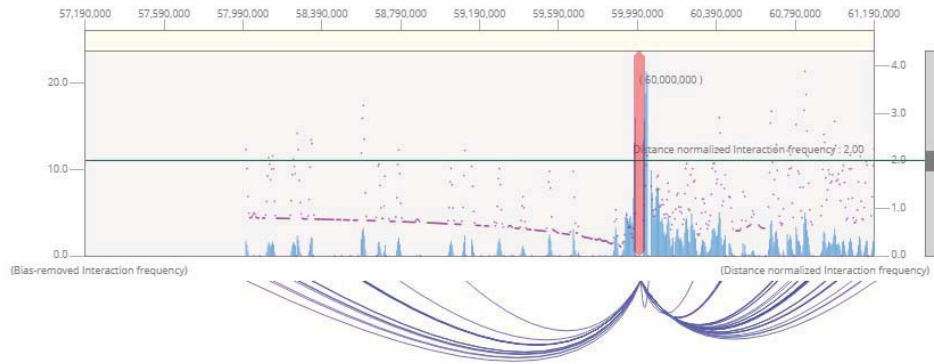


No.	Chromosome	Start	End	Gene Name	Locus
1	chr1	61975000	61980000		

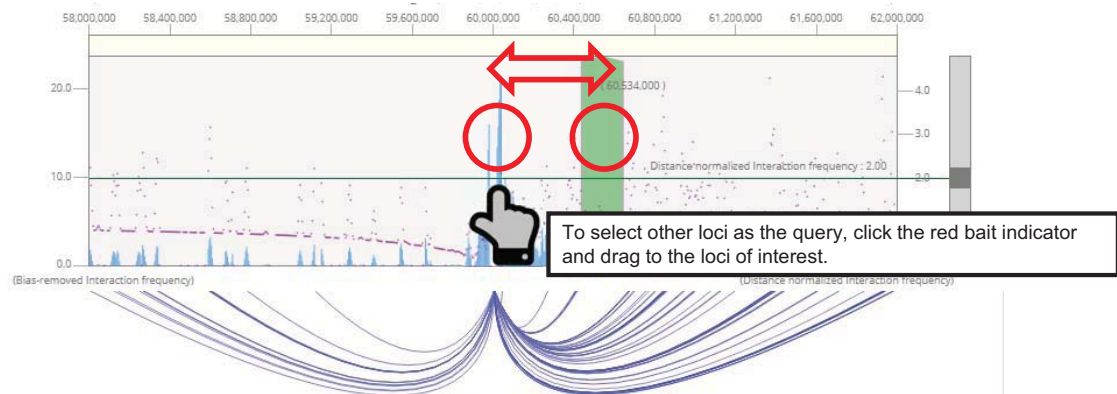
## Browse interaction frequency w/o change the bait



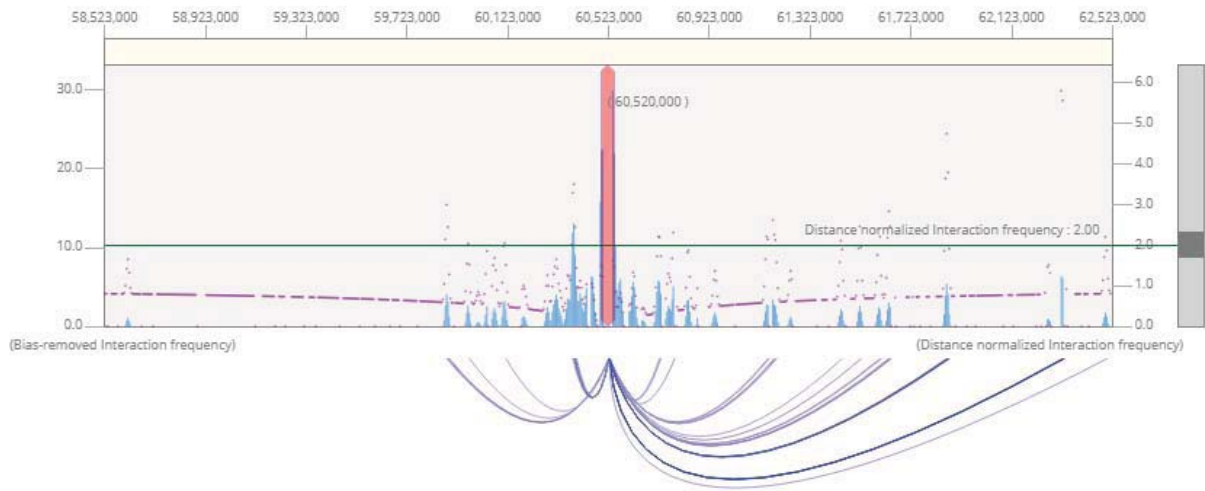
## Browse interaction frequency w/o change the bait



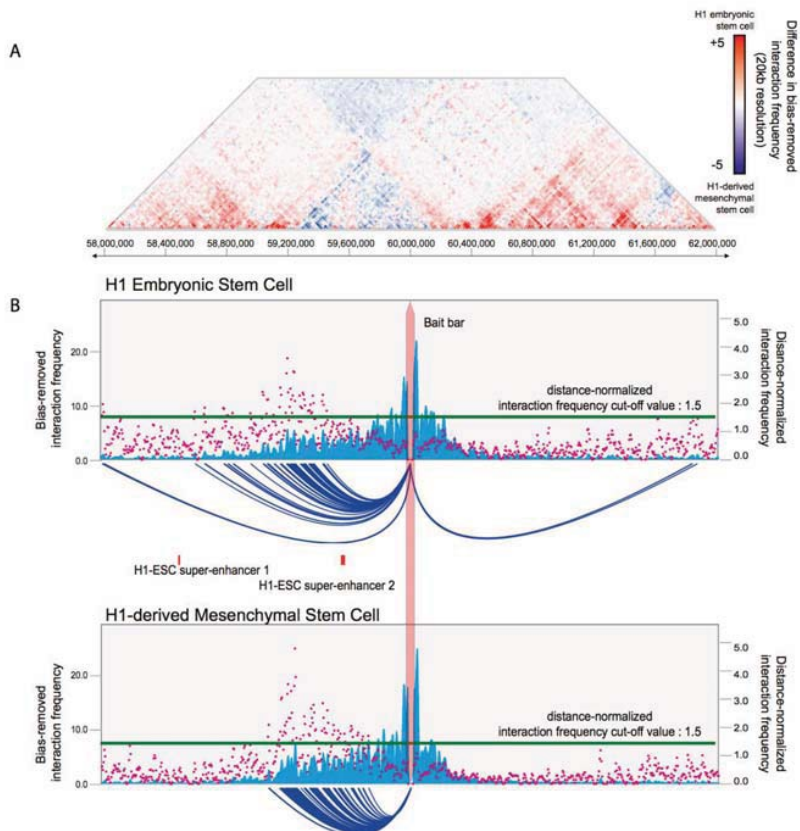
## Adjust bait without resubmission



## Adjust bait without resubmission

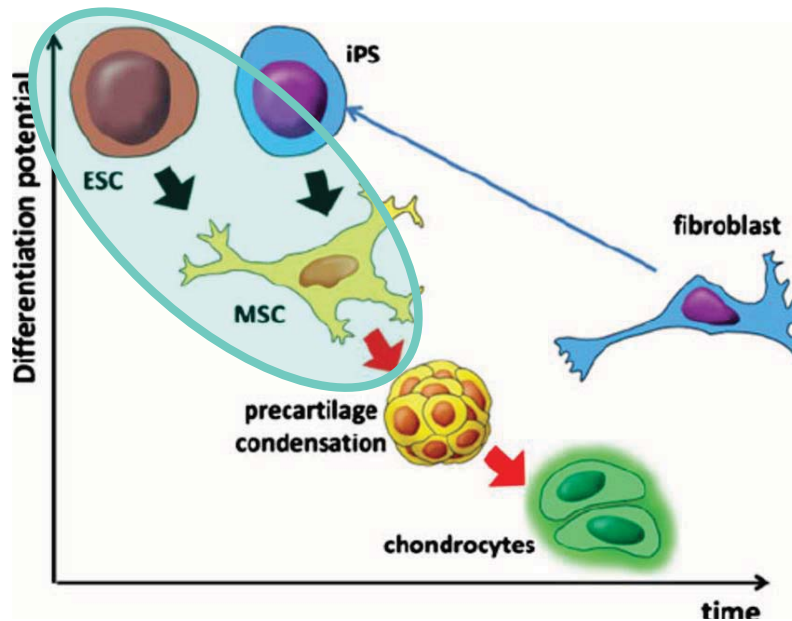


## Module 3 : Comparative Visualization



## Example : Interaction change during differentiation

During the differentiation, the interaction profile is dramatically changed. In this session, we will compare the interaction profile of ESC and MSC. ESC : Embryonic Stem Cell, MSC : Mesenchymal Stem Cell



Gadjanski et al, Stem Cell Rev. Rep. (2012)

## Step 1 : Open Comparative visualization Module

3D genome interaction viewer and database

hg19

Hi-C

Capture Hi-C

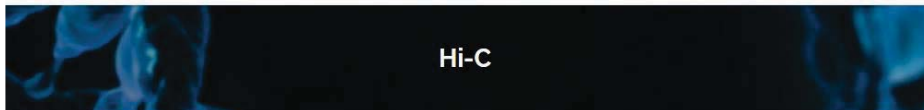
Cancer Hi-C

Statistics

Download

Tutorial

Contact Us



Interaction table   Interaction visualization   **Comparative interaction visualization**

> Choose sample(s)

Choose sample(s) by characteristics   Choose sample(s) by search   Choose sample(s)

> Type   > Sample property   > Condition   > Sample

Choose...   Choose...   Choose...   Choose...

> Input bait

Bait :  
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range

2Mb

Add sample(s)   Remove sample(s)

> Selected region(s)

	Sample	Bait
<input type="checkbox"/>		

Example Run   Run

## Step 2 : Choose a sample



Interaction table Interaction visualization Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics Choose sample(s) by search Choose sample(s)

- HAP1 (near-haploid cell line)
- HAP1 (near-haploid cell line), SSC Knock Out
- HAP1 (near-haploid cell line), WAPL and SSC Knock OUT
- HAP1 (near-haploid cell line), WAPL knock Out
- HEK293T (embryonic kidney cell line), transfected with dCas9-VPR targeting the exon CTCF binding site of Pcdha12
- HEK293T (embryonic kidney cell line), transfected with dCas9-VPR targeting the promoter CTCF binding site of Pcdha12
- Hippocampus
- HTBE (human tracheobronchial epithelial cells), infect active H5N1 influenza, infection time 6hour
- HTBE (human tracheobronchial epithelial cells), infect active H5N1 influenza, infection time 12hour
- HTBE (human tracheobronchial epithelial cells), infect active H5N1 influenza, infection time 18hour
- HTBE (human tracheobronchial epithelial cells), infect mock, infection time 6hour
- HTBE (human tracheobronchial epithelial cells), infect mock, infection time 12hour

> Input bait

Bait :  
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range

2Mb

Add sample(s) Remove sample(s)



Click to load the list of Hi-C experiments

## Step 2 : Choose a sample



Interaction table Interaction visualization Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics Choose sample(s) by search Choose sample(s)

- fibroblast(CRL-2522) dexamethasone 24h
- fibroblast(CRL-2522) dexamethasone 32h
- fibroblast(CRL-2522) dexamethasone 40h
- fibroblast(CRL-2522) dexamethasone 48h
- fibroblast(CRL-2522) dexamethasone 56h
- GM23248 (primary skin fibroblasts)
- H1 Embryonic Stem Cell
- H1 Mesenchymal Stem Cell
- H1 Mesendoderm Cell
- H1 Neuronal Progenitor Cell
- H1 Trophoctoderm Cell
- H9 human Embryonic Stem Cell Line, Heat shock condition
- H9 Human Embryonic Stem Cells

> Input bait

Bait :  
(Ex. CROCCP2, chr22:27141000, rs42)

> Interaction range

2Mb

Add sample(s) Remove sample(s)



Click to choose sample

## Step 3 : Choose a Bait

Interaction table   Interaction visualization   Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics   Choose sample(s) by search   Choose sample(s)

- fibroblast(CRL-2522) dexamethasone 24h
- fibroblast(CRL-2522) dexamethasone 32h
- fibroblast(CRL-2522) dexamethasone 40h
- fibroblast(CRL-2522) dexamethasone 48h
- fibroblast(CRL-2522) dexamethasone 56h
- GM23248 (primary skin fibroblasts)
- H1 Embryonic Stem Cell
- H1 Mesenchymal Stem Cell
- H1 Mesendoderm Cell
- H1 Neuronal Progenitor Cell
- H1 Trophectoderm Cell
- H9 human Embryonic Stem Cell Line, Heat shock condition
- H9 Human Embryonic Stem Cells

> Input bait   > Interaction range

Bait: chr1:60000000  
(Ex. CROCCP2, chr22:27141000)

Insert ID of Gene/SNP or genomic coordinate

Add sample(s)   Remove sample(s)

> Selected region(s)

Click button to add sample

<input type="checkbox"/>	Sample	Bait
<input type="checkbox"/>	H1 Embryonic Stem Cell	chr1:60000000
<input type="checkbox"/>	H1 Mesenchymal Stem Cell	chr1:60000000

Example Run   Run

## Step 4 : Run Module

Interaction table   Interaction visualization   Comparative interaction visualization

> Choose sample(s)

Choose sample(s) by characteristics   Choose sample(s) by search   Choose sample(s)

- fibroblast(CRL-2522) dexamethasone 24h
- fibroblast(CRL-2522) dexamethasone 32h
- fibroblast(CRL-2522) dexamethasone 40h
- fibroblast(CRL-2522) dexamethasone 48h
- fibroblast(CRL-2522) dexamethasone 56h
- GM23248 (primary skin fibroblasts)
- H1 Embryonic Stem Cell
- H1 Mesenchymal Stem Cell
- H1 Mesendoderm Cell
- H1 Neuronal Progenitor Cell
- H1 Trophectoderm Cell
- H9 human Embryonic Stem Cell Line, Heat shock condition
- H9 Human Embryonic Stem Cells

> Input bait   > Interaction range

Bait: chr1:60000000  
(Ex. CROCCP2, chr22:27141000, rs42)

2Mb

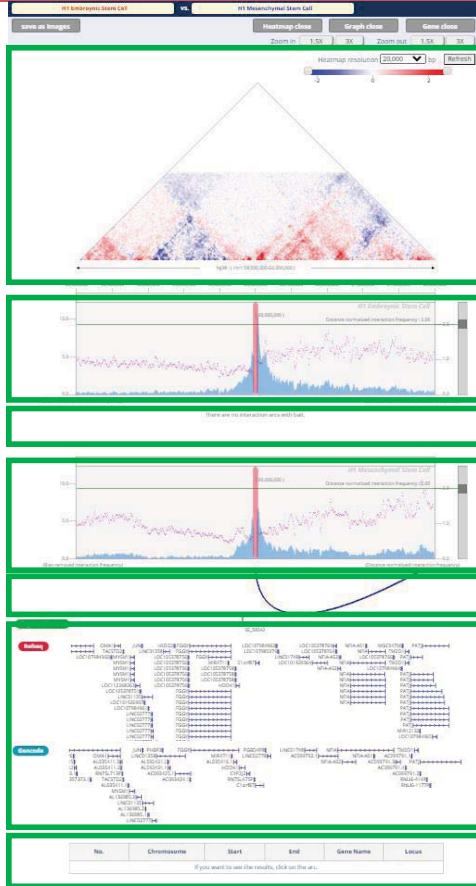
Add sample(s)   Remove sample(s)

> Selected region(s)

<input type="checkbox"/>	Sample	Bait
<input type="checkbox"/>	H1 Embryonic Stem Cell	chr1:60000000
<input type="checkbox"/>	H1 Mesenchymal Stem Cell	chr1:60000000

Example Run   Run

## Step 5 : Adjust comparative heatmap



Comparative heatmap of interaction frequency between 1<sup>st</sup> and 2<sup>nd</sup> samples.

Arc-representation of significant interactions in 1<sup>st</sup> sample

Arc-representation of significant interactions in 2<sup>nd</sup> sample

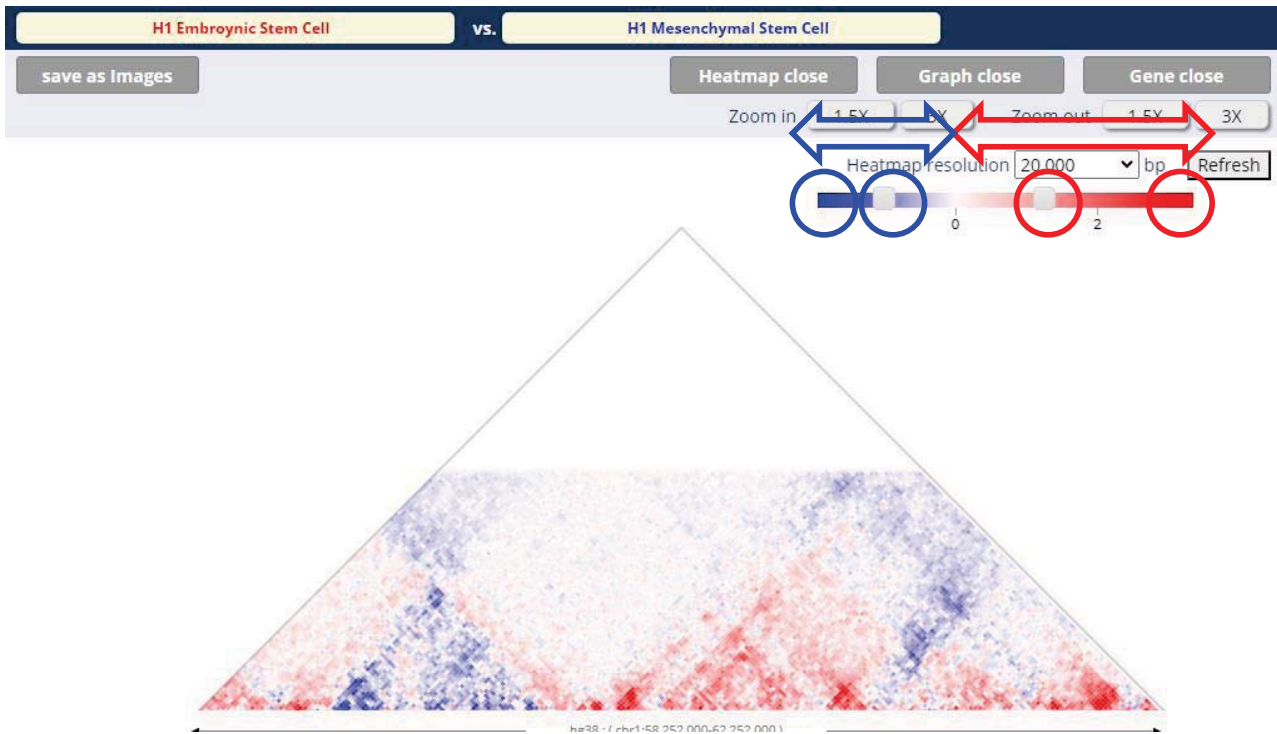
RefSeq Genes and super enhancer annotations

Description of selected interaction

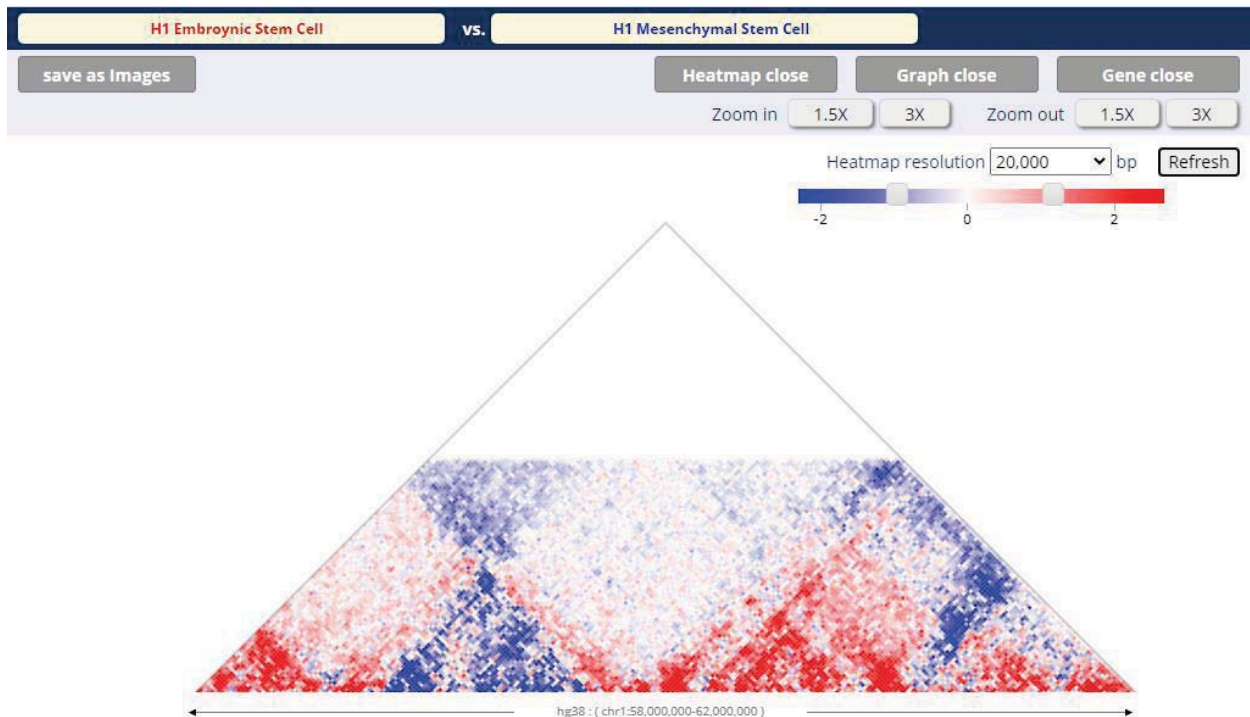
## Step 5a : Synchronized criteria change



## Step 5b : Adjust the heatmap color range

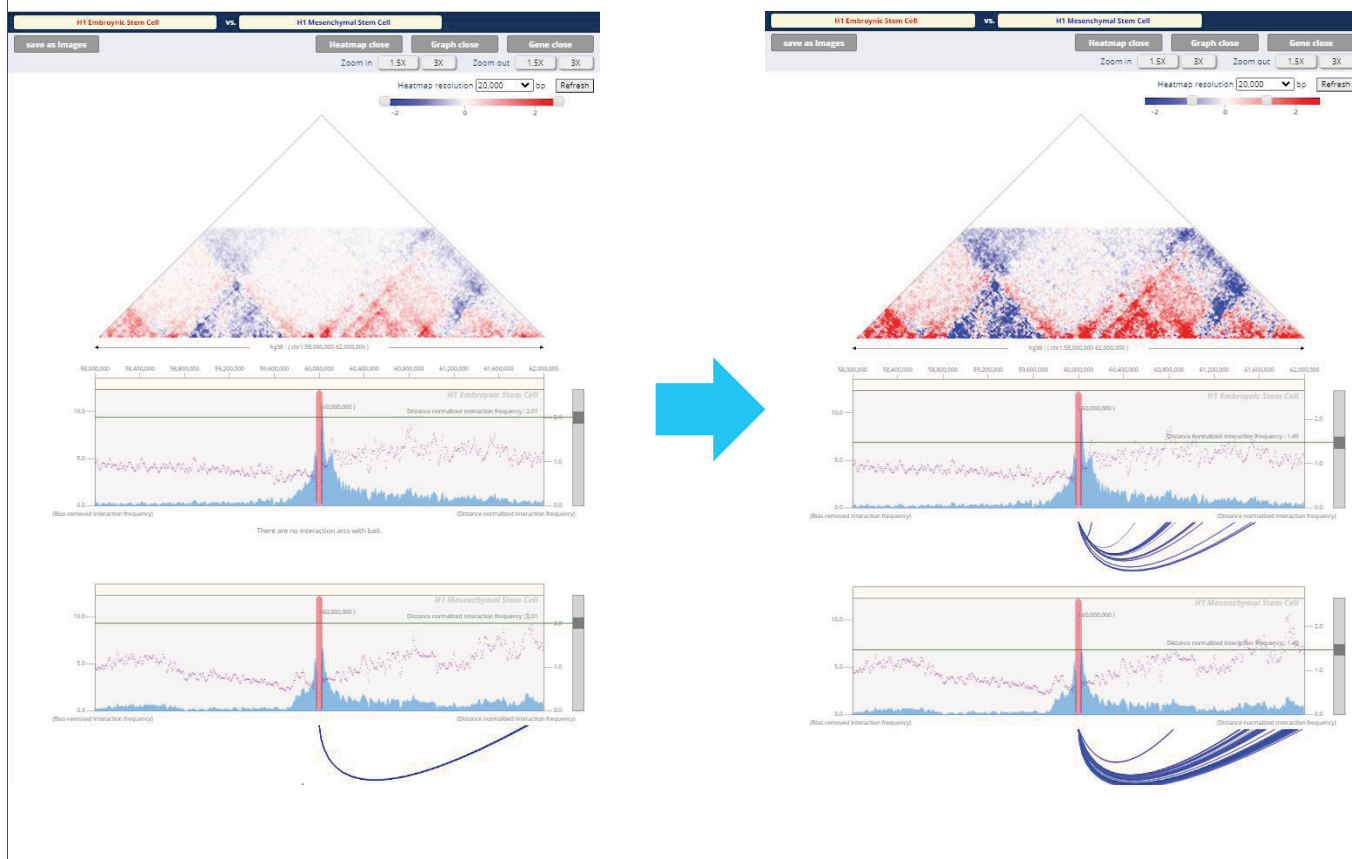


## Step 5b : Adjust the heatmap color range



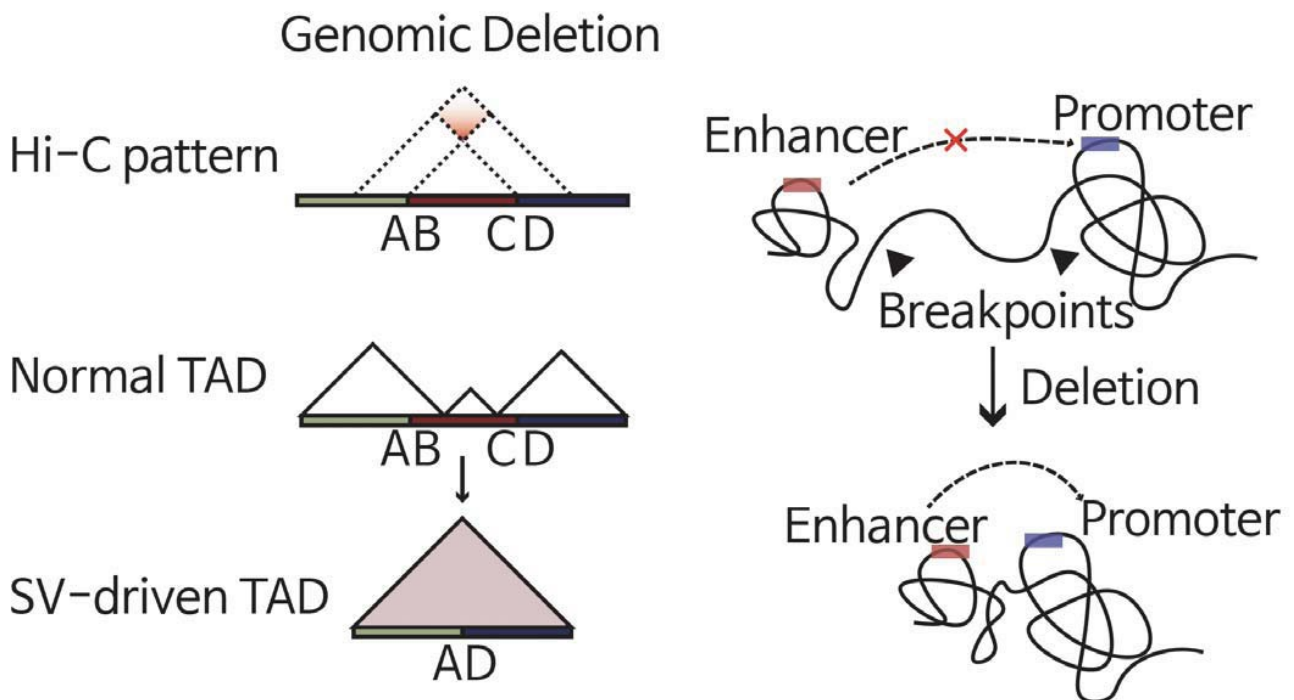


## Step 5 : Adjust comparative heatmap



## Cancer Hi-C Analysis

## The impact of large scale structural variations to cancer 3D genome



## Interactively visualize and simulate the impact of structural variations to cancer 3D genome

### Problem statement

1. Frequent genomic rearrangements in cancer alters 3D genome
2. Abberant gene expression based on rewired regulatory elements
3. Requires appropriate visualization tools and processed data

### Resolving issue

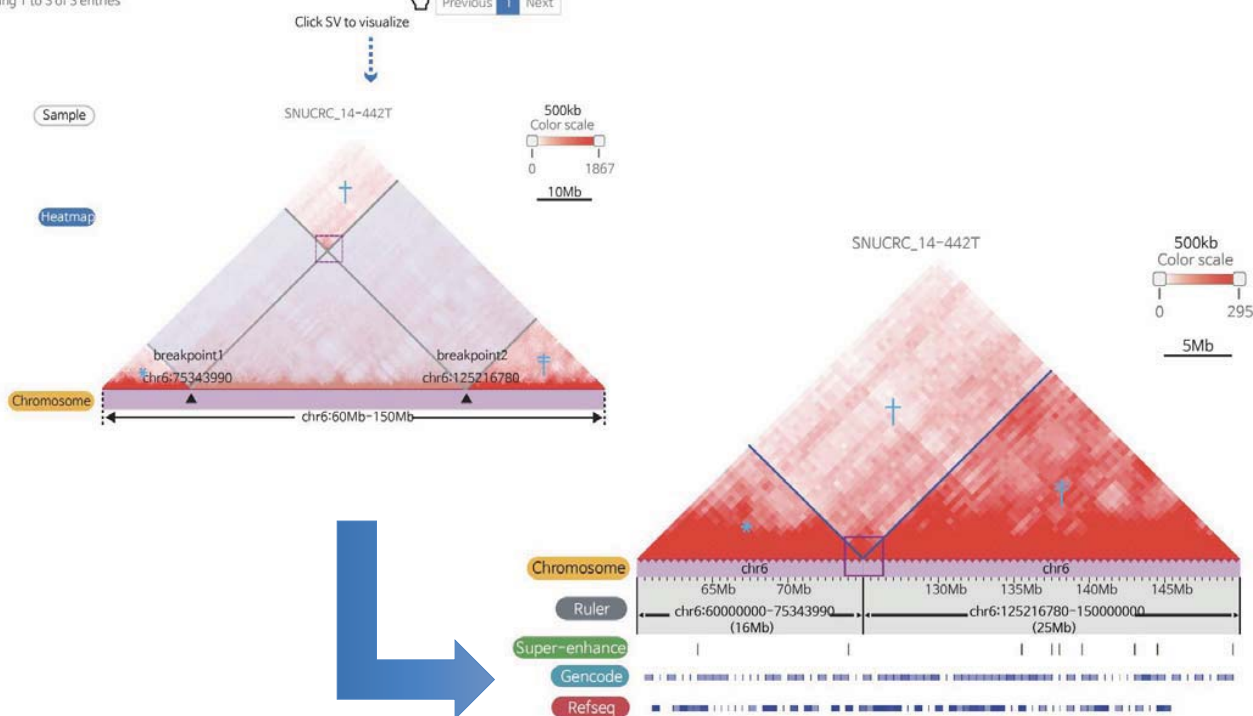
1. Collection of large cancer/normal Hi-C and pHi-C data
2. Visualization of cancer 3D genome
3. Hi-C contact map manipulation to examine impact of SVs

# Module I. Pre-called SV and 3D genome

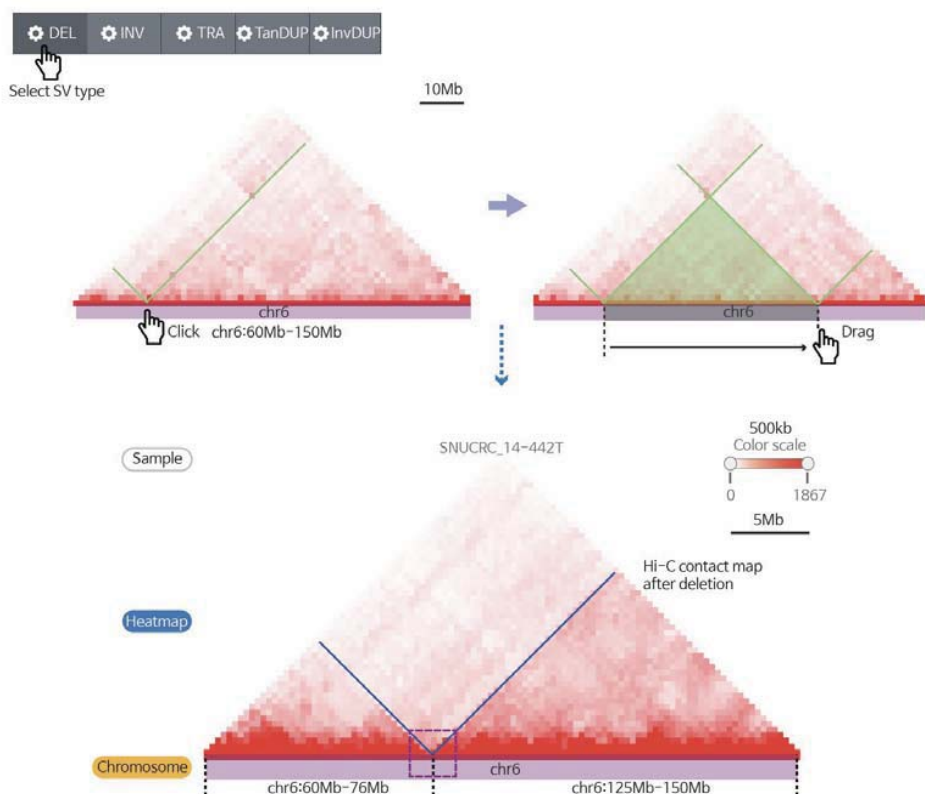
Sample	Chrom1	Breakpoint1	Chrom2	Breakpoint2	SV type	Orientation
14-442T	chr6	...	chr6	...	INV	3to3
14-442T	chr6	...	chr6	...	INV	5to5
14-442T	chr6	75343990	chr6	125216780	DEL	3to5

Showing 1 to 3 of 3 entries

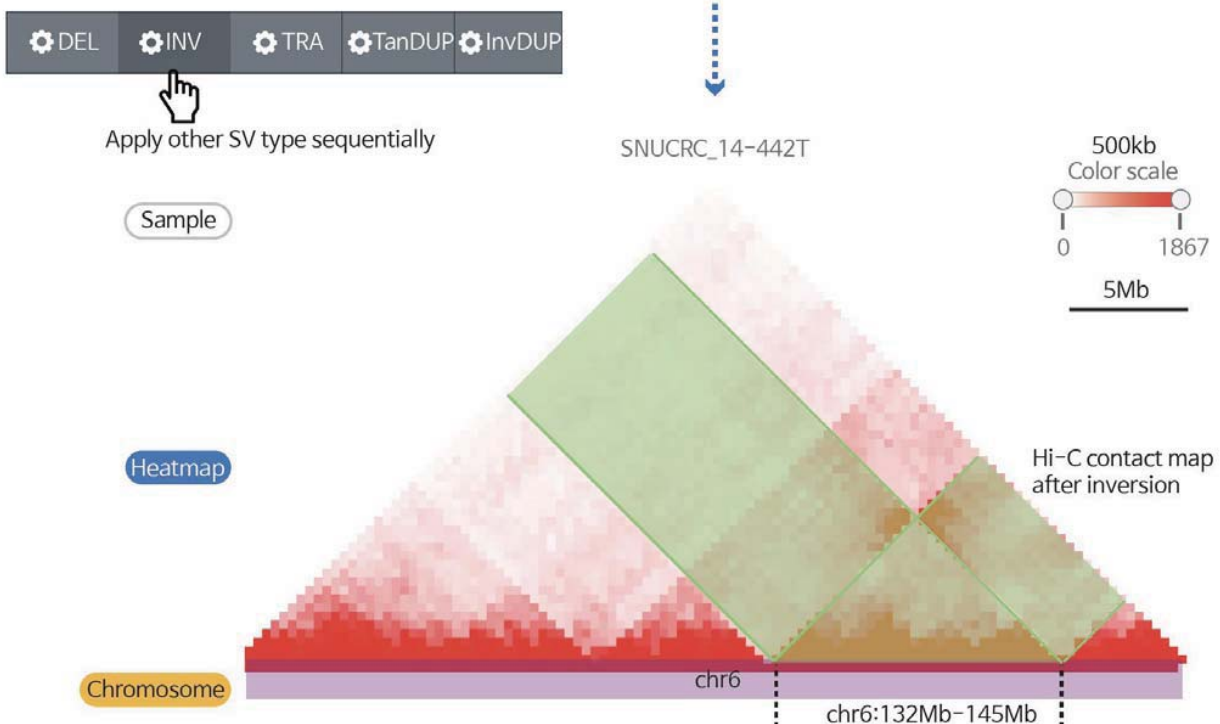
Previous 1 Next



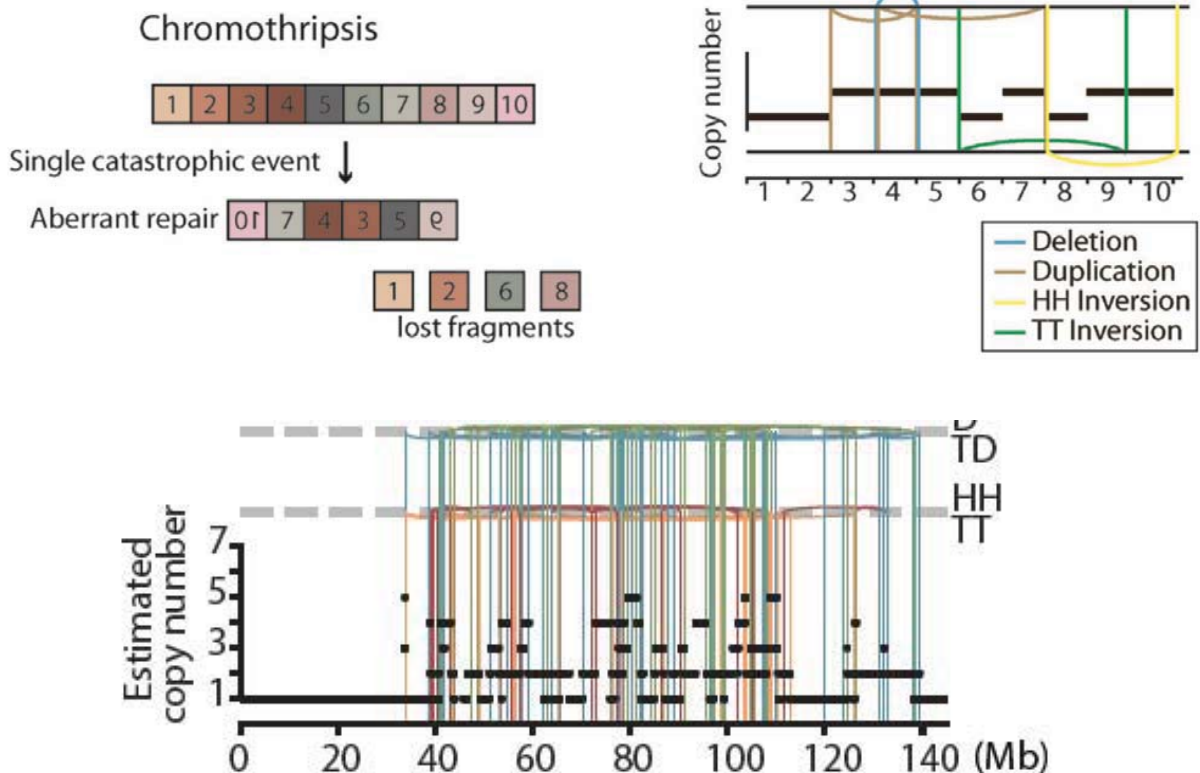
# Module II. Interactive 3D genome manipulation



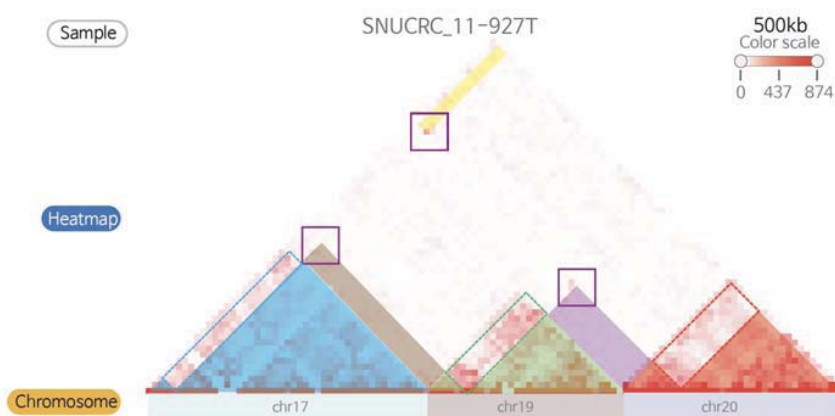
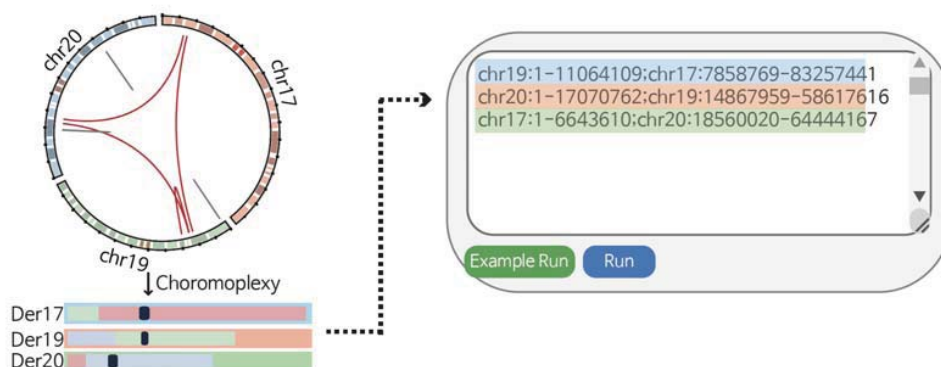
## Module II. Interactive 3D genome manipulation



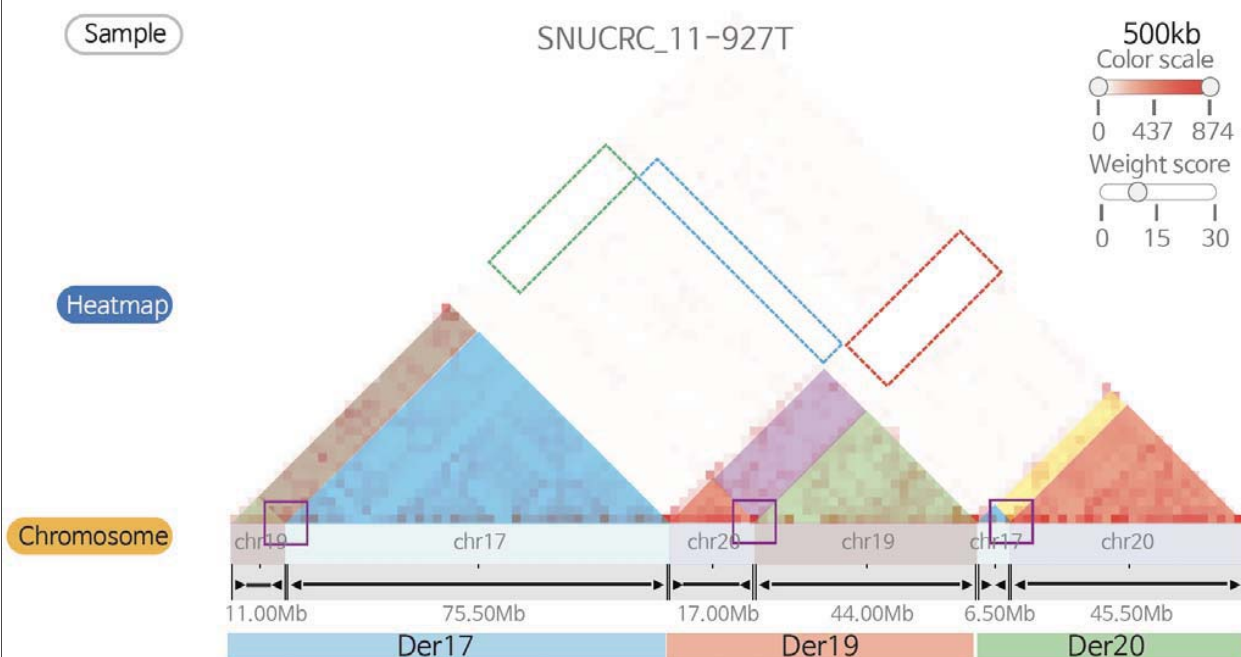
## Complex forms of large-scale structural variations



## Module III. Complex SV and 3D genome



## Module III. Complex SV and 3D genome



## Summary

- 3DIV provides the largest number of Hi-C samples
- 3DIV covers most of the required functionality in navigating the 3D cancer genome
- 3DIV is the most comprehensive resource to explore the gene regulatory effects of both the normal and cancer 3D genome

# KSBi-BIML 2022

covNorm R package tutorial  
covNorm을 활용한 Hi-C 분석 실습

## 1. Installation

- R (covNorm 구동) 과 Python (시각화 & 분석) 설치 필요
- 발표자료 내용은 Anaconda2 (5.2.0) 의 Python 2.7.15/3.8.10 과 R 3.4.3 환경에서 사용됨
- 정확히 R/Python version을 맞출 필요는 없으나 특정 버전에서는 패키지 의존성 문제 발생 가능 주의

67

## 1. Installation: R setup

- 방법 1: R 에서 의존성 패키지 직접 설치

```
install.packages(c("MASS", "propagate", "FAdist", "stringr", "splines")) # Imports
install.packages(c("reshape2", "gplots", "ggplot2", "corrplot"))      # Suggests
```

- 방법 2: conda 환경에서 설치 (Anaconda2 5.2.0)

```
conda create -n r-env r-essentials r-base
conda activate r-env
```

```
conda install -c anaconda libopenblas
conda install -c r r-gplots r-corrplot r-gmm r-mvtnorm
conda install -c conda-forge r-propagate r-fadist r-tmvtnorm
```

- R studio 설치 기능 사용 가능

68

## 1. Installation: package install

- 방법 1: Git clone 후 R CMD 로 build/INSTALL

```
git clone https://github.com/kaistcbfg/covNormRpkg.git
R CMD build covNormRpkg
R CMD INSTALL covNormRpkg_1.1.0.tar.gz
```

- 방법 2: R devtools 패키지로 R 콘솔에서 명령어 입력

```
devtools::install_github("kaistcbfg/covNormRpkg")
```

- 설치 이후 콘솔에서 `library(covNormRpkg)` 입력 시 에러 없이 해당 코드가 실행되어야 함

69

## 1. Installation: Python setup

- Python 패키지의 경우 anaconda 설치 시 numpy, matplotlib 등이 사전 설치되어 있음. miniconda의 경우 추가 설치 필요
- numpy, matplotlib, gzip, argparse, pickle 은 설치 필수. `import <package name>` 으로 설치 테스트
- 본 발표자료 시연 코드 사용 시 pysam과 hmmlearn 패키지 필요(추후 설명)
- pip 나 easy\_install, conda 등으로 설치 가능

70



## 2. Input data

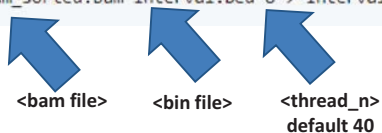
- BAM 파일 준비. 3DIV 파이프라인 사용.  
[https://github.com/kaistcbfg/3divv2/blob/master/hic\\_pipeline/Normal\\_proc\\_PBS.py](https://github.com/kaistcbfg/3divv2/blob/master/hic_pipeline/Normal_proc_PBS.py)
- BWA 로 fastq\_1, fastq\_2 파일 각각 mapping 후 chimeric read filtering.
- Filtered read 를 paired 로 merge 한 후 self ligation read 제거.
- Picard Markdup 으로 PCR duplicate 제거 후 samtools index.
- Inter-chromosomal (*trans*-) interaction을 분석하지 않을 경우 self ligation과 같이 *trans* read 제거

71

## 2. Input data

- coverage 계산
- Bedtools 의 coverageBed & sortBed로 계산. Bin 파일은 bed format으로 사전 생성  
coverageBed -counts -abam <bam file> -b <bin file> > <coverage file>  
sortBed -i <coverage file> > <sorted coverage file>
- coverageBed의 경우 bam 파일 용량이 클 경우 매우 느림.  
<https://github.com/gorliver/pyCoverage> pyCoverage 사용시 병렬 계산 가능.  
pysam 패키지 설치 필요.

```
pyCoverage.py bam_sorted.bam interval.bed 8 > interval.cov
```



<bam file>   <bin file>   <thread\_n>  
default 40

### 40kb bin file example

```
chr10 0 40000  
chr10 40000 80000  
chr10 80000 120000  
chr10 120000 160000  
chr10 160000 200000  
chr10 200000 240000  
chr10 240000 280000  
chr10 280000 320000  
chr10 320000 360000  
chr10 360000 400000  
chr10 400000 440000  
chr10 440000 480000  
chr10 480000 520000  
chr10 520000 560000  
chr10 560000 600000  
chr10 600000 640000  
chr10 640000 680000  
chr10 680000 720000  
chr10 720000 760000
```

72

## 2. Input data

- covNorm 입력을 위해서는 아래 table과 같이 formatting (gzipped) 필요.
- Row 1은 column name (table과 동일해야 함).
- frag1, frag2 는 interaction하는 두 bin.
- cov\_frag1, cov\_frag2 는 frag1과 frag2 bin의 coverage.
- freq는 두 bin의 interaction frequency (bam 파일의 read 수)
- dist는 두 bin의 genomic distance의 절대값.

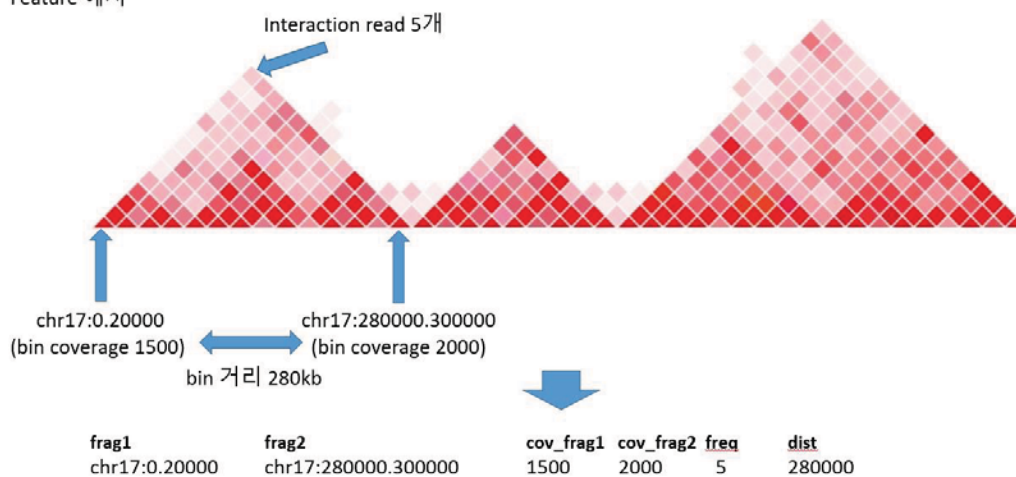
frag1	frag2	cov_frag1	cov_frag2	freq	dist
chr17.140000.160000	chr17.83160000.83180000	2296	2304	1.0	83020000
chr17.140000.160000	chr17.83180000.83200000	2296	2072	2.0	83040000
chr17.140000.160000	chr17.83200000.83220000	2296	778	2.0	83060000
...	...	...	...	...	...
chr17.160000.180000	chr17.200000.220000	2119	2253	12.0	40000
chr17.160000.180000	chr17.220000.240000	2119	1744	9.0	60000

<http://junglab.kaist.ac.kr/Dataset/GM19204.chr17.cis.feature.gz>

73

## 2. Input data

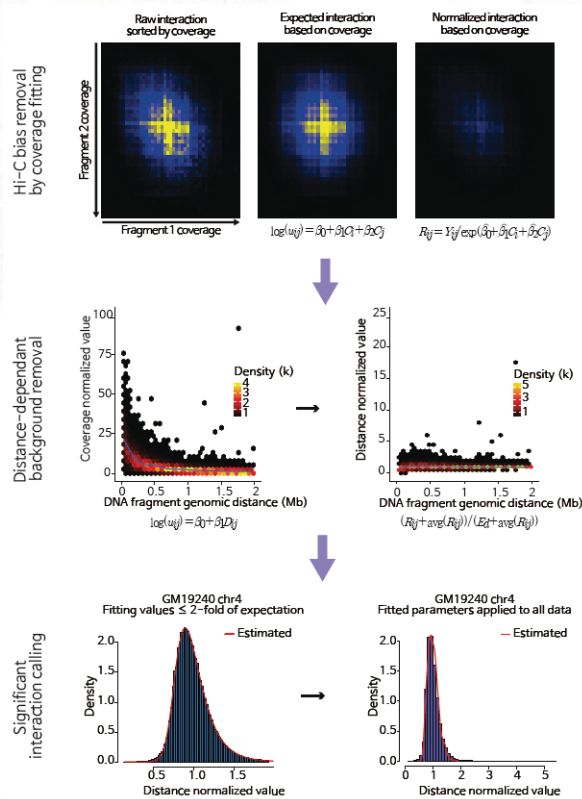
Feature 예시



Zero frequency bin (freq == 0)은 row에 추가 권장되지 않음

74

### 3. Normalization



75

### 3. Normalization

```
library('covNormRpkg')

args <- commandArgs(TRUE)
file_name=args[1]

raw_data <- read.table(gzfile(file_name),head=TRUE)
print("1: Data Loaded.")

raw_data_filter <- covNormRpkg::filterInputDF(raw_data)
print("2: Data Filtered.")

cov_result <- covNormRpkg::normCoverage(raw_data_filter)
cov_result$coeff_cov1
cov_result$coeff_cov2
cov_df <- cov_result$result_df
write.table(cov_df, file=gzfile("outFileName1"), row.names=FALSE, col.names=TRUE, sep="\t", quote=FALSE) #
print("3: Coverage normalized.")

covNormRpkg::checkFreqCovPCC(cov_df, outpdfname='QCplot_coverage_PCC.pdf')
covNormRpkg::plotCovNormRes( cov_df, outpdfname='QCplot_coverage_heatmap.pdf')
print("4: Plot coverage normalization results.")

dist_result <- covNormRpkg::normDistance(cov_df, max_dist=2000000)
dist_df <- dist_result$result_df
print("5: Distance normalized.")

covNormRpkg::checkFreqDistPCC(dist_df, outpdfname='QCplot_dist_PCC.pdf')
covNormRpkg::plotDistNormRes( dist_df, outpdfname='QCplot_dist_hexmap.pdf')
print("6: Plot distance normalization results.")

final_df <- covNormRpkg::contactPval(dist_df, 'fit.pdf')
print("7: Significant interactions called.")

#Uncomment 'saveEachChr' to split-save file for each chromosome.
#covNormRpkg::saveEachChr(final_df, "./outputFolder", "outputSampleName")
write.table(final_df, file=gzfile("outFileName2"), row.names=FALSE, col.names=TRUE, sep="\t", quote=FALSE)
```

Load data

Coverage normalization

Distance normalization

Significant interaction calling

Save result

76

### 3. Normalization

- coverage Normalization 우선 실행. 두 bin의 coverage를 GLM fitting 해 구한 expected value 대비 fold change 를 계산하여 cov\_result 변수에 저장.

```
cov_result <- covNormRpkg::normCoverage(raw_data_filter)
```

- Interaction frequency 기대값 (u) 이 Negative binomial 분포를 따르며 coverage에 대한 기대값을 fitting하여 parameter를 찾는 작업. 다음 수식과 같이 표현 가능.

$$\log(u_{ij}) = b_0 + b_1(C_i) + b_2(C_j)$$

- coverage가 너무 낮은 bin 의 경우 값을 지정해서 row 제거 가능. default 200. Self-ligation, 0 bin, trans-read 등도 사전 필터링 가능

```
raw_data_filter <- covNormRpkg::filterInputDF(raw_data)
```

- cov\_result의 result\_df 변수를 cov\_df 에 저장. (실제 normalization된 데이터). coeff\_cov1과 coeff\_cov2은 각 coverage의 fitting coefficient.

```
cov_result$coeff_cov1
cov_result$coeff_cov2
cov_df <- cov_result$result_df
```

77

### 3. Normalization

coverage Normalization 결과 해석: cov\_df의 경우 feature data.frame 뒤에 3개 컬럼 추가.

**rand:** 난수. coverage 1과 2의 shuffle에 사용.  
**exp\_value\_capture:** 두 bin의 coverage를 고려한 경우 interaction frequency의 기대값.  
**capture\_res:** residual. freq를 exp\_value\_capture로 나눠 준 값 (normalized frequency).

rand	exp_value_capture	capture_res
77	2.0625	2.9091
47	1.049	1.9066
40	1.6197	0.6174
38	1.7874	0.5595
...	...	...
84	0.7842	1.2752

$$R_{ij} = Y_{ij} / \exp(\hat{b}_0 + \hat{b}_1(C_i) + \hat{b}_2(C_j))$$

capture\_res
freq
exp\_value\_capture

78

### 3. Normalization

- coverage normalization 이후 distance-dependant background 제거를 위해 normalization 실행. coverage normalization과 유사하게 NB 분포를 따르는 기대값을 genomic distance에 대한 fitting으로 계산.

$$\log(u_{ij}) = b_0 + b_1(D_{ij})$$

- normDistance 함수에 cov\_df 입력. 일정 거리 이상에서는 Hi-C 신호가 0으로 수렴하기에 max\_dist 이내 (2Mb 사용) interaction에 대해서만 normalization 진행됨 (나머지 row는 drop). coverage\_normalization과 동일하게 dist\_result는 coeff와 result\_df로 구성됨.

```
dist_result <- covNormRpkg::normDistance(cov_df, max_dist=2000000)
dist_df <- dist_result$result_df
```

- dist\_df의 경우 cov\_df의 뒤에 column 2개 추가됨. **exp\_value\_dist**: 해당 **dist**의 기대값, **dist\_res**: normalized frequency.

exp_value_dist	dist_res
1.7021	1.4464
1.6201	1.1093
0.7307	0.9346
0.6794	0.9287
...	...
1.3979	0.9489

79

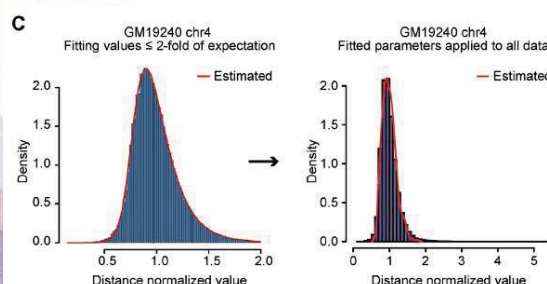
### 3. Normalization

- Distance normalization 이후 interaction의 significance와 FDR 계산 가능. 3-point Weibull distribution fitting을 통해 p-value와 FDR 계산.

- dist\_df를 입력 후 실행 시 data.frame으로 결과 출력. ('fit.pdf' 는 fitDistr 패키지에서 자동생성 되는 figure 이름 지정. 실제로는 사용하지 않음)

```
final_df <- covNormRpkg::contactPval(dist_df, 'fit.pdf')
```

- Outlier에 (기대값 대비 > 2-fold) 적절한 값을 assign 하기 위해 2-fold 이하의 값들로만 fitting하여 parameter를 얻은 후 전체 데이터를 해당 파라미터로 얻어진 3P-Weibull 분포에 넣어 p-value 계산. R의 p.adjust 함수로 FDR 계산. df 에 컬럼 추가됨



p_result_dist	FDR_dist_res
0.053347	0.709688
0.237543	0.945149
0.475677	0.948673
0.486306	0.948673
...	...
0.450677	0.948673

80

### 3. Normalization

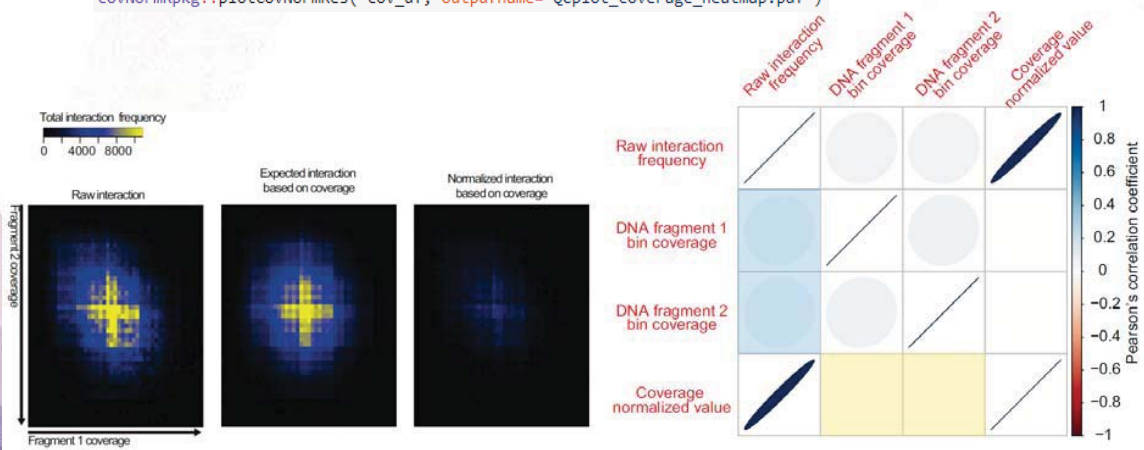
- Coverage normalization QC
- 각 bin당 coverage (cov\_frag1, cov\_frag2)에 dependent 한 결과가 나오면 안됨 (fitting 에러). coeff\_cov1과 coeff\_cov2를 확인하여 비슷한 값을 가지는지 체크
- Coverage가 높은 경우 여러 원인 (bias)에 의해 해당 bin의 interaction frequency가 높음. Normalization 후 coverage-interaction frequency간 correlation이 감소해야 함
- R에서 cor 함수로 correlation (Pearson's) 체크 가능. capture\_res가 freq 보다 PCC 값이 감소해야 정상

```
cor(df$cov_frag1, df$freq)
cor(df$cov_frag2, df$freq)
cor(df$cov_frag1, df$capture_res)
cor(df$cov_frag2, df$capture_res)
```

### 3. Normalization

- Coverage normalization QC
- covNorm R package에서는 coverage normalization 전후 결과 시각화 지원 (coverage sorted heatmap & corplot)

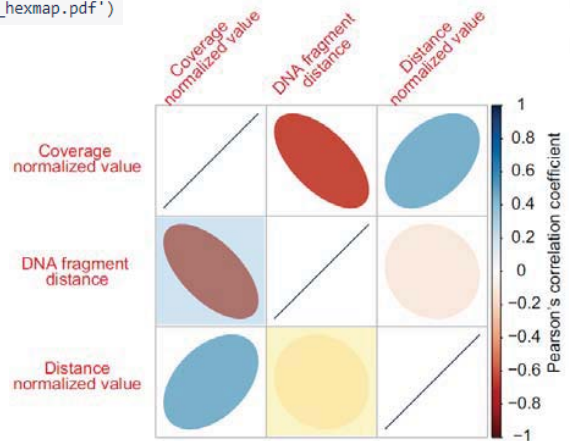
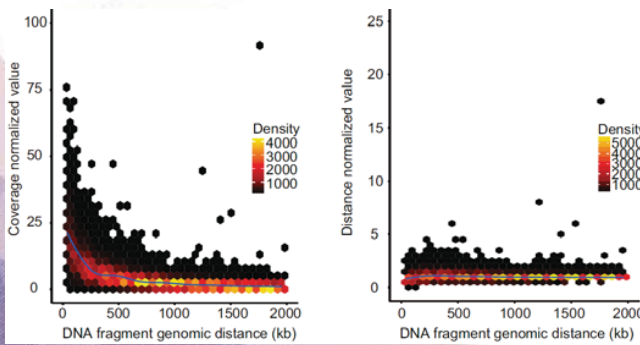
```
covNormRpkg::checkFreqCovPCC(cov_df, outpdfname='QCplot_coverage_PCC.pdf')
covNormRpkg::plotCovNormRes(cov_df, outpdfname='QCplot_coverage_heatmap.pdf')
```



### 3. Normalization

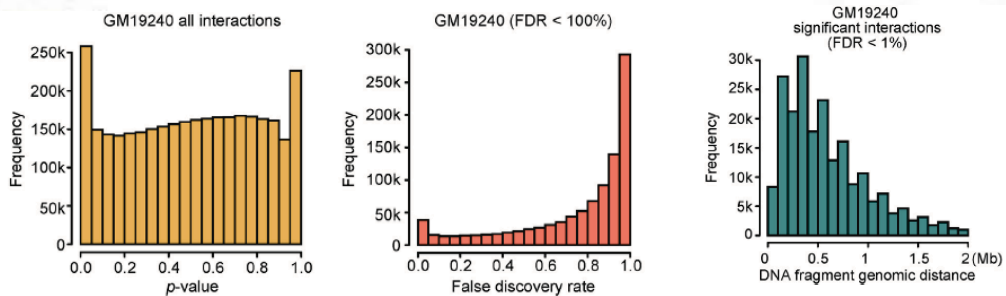
- Distance normalization QC
- Distance도 마찬가지로 normalization (dependent background removal) 후 correlation이 감소해야 함.
- R cor 함수로도 확인 가능하며 유사한 시각화 코드 제공 (distance-value plot & corplot)

```
covNormRpkg::checkFreqDistPCC(dist_df, outpdfname='QCplot_dist_PCC.pdf')
covNormRpkg::plotDistNormRes( dist_df, outpdfname='QCplot_dist_hexmap.pdf')
```



### 3. Normalization

- $p$ -value and FDR QC
- FDR 계산 후 값의 분포가 실제 분포 (exponential) 처럼 나오는지 확인. FDR < 1% 등의 threshold로 significant interaction 정의, 해당 interaction들의 genomic distance가 unbiased 되었는지 확인.



## 4. Visualization

- 앞서 실행한 coverage normalization 결과 df를 “GM19204.chr17.covnorm.gz” 로 저장
- 아래 코드 실행 (파일 이름/경로 적절히 교체) 시 pdf 포맷으로 오른쪽과 같은 figure 제공

```
import numpy as np
import matplotlib
matplotlib.use('Agg')
import matplotlib.pyplot as plt
import gzip

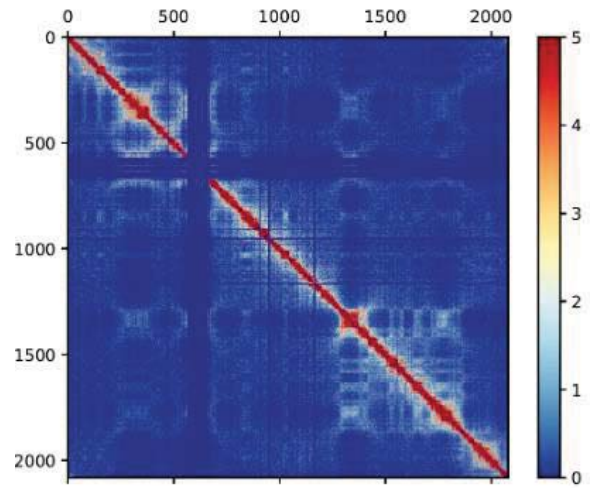
chromosome_size = 83257441 #hg38 chr17
resolution = 40000 # 40kb
bin_length = (chromosome_size/resolution) + 1

contact_map = np.zeros((bin_length, bin_length))

f = gzip.open('GM19240.chr17.covnorm.gz') # coverage normalization result file
f.readline() #header
for line in f:
    line = line.rstrip()
    linedata = line.split('\t')
    bin1 = int(linedata[0].split('.')[1])/resolution
    bin2 = int(linedata[1].split('.')[1])/resolution
    freq = float(linedata[8])

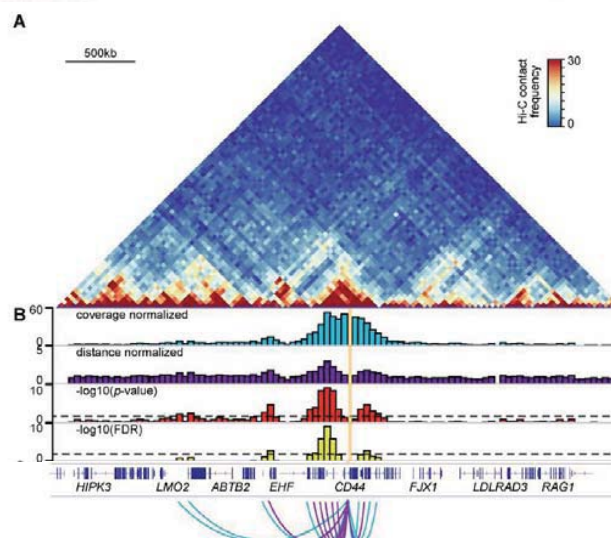
    contact_map[bin1][bin2] += freq
    contact_map[bin2][bin1] += freq
#
f.close()

fig = plt.figure(1)
ax = fig.add_subplot(111)
cax = ax.matshow(contact_map, cmap=plt.cm.RdYlBu_r, vmin=0, vmax=5)
fig.colorbar(cax)
plt.savefig("HiC_contact_map.pdf", dpi=1000)
```



## 4. Visualization

- Hi-C contact map, normalization 결과, genome track 생성 software 등 적절히 조합 시 아래와 같은 figure 생성 가능.





## 5. Further analysis

- Fitting기반 normalization의 경우 여러 장점이 있으나 data point가 과도하게 많을 경우 (very high depth) fitting 시간이 오래 걸린다는 단점 존재.
- “데이터가 충분할 경우” 일부 데이터를 “적절히” sampling 해서 fitting해도 원본 데이터의 분포 반영에 문제 없다고 가정 가능.
- covNorm의 경우 coverage/distance normalization 과정에서 subsample fitting 기능 제공. sample\_ratio 파라미터에 sample ratio 입력해서 사용 (default로 사용 안함).

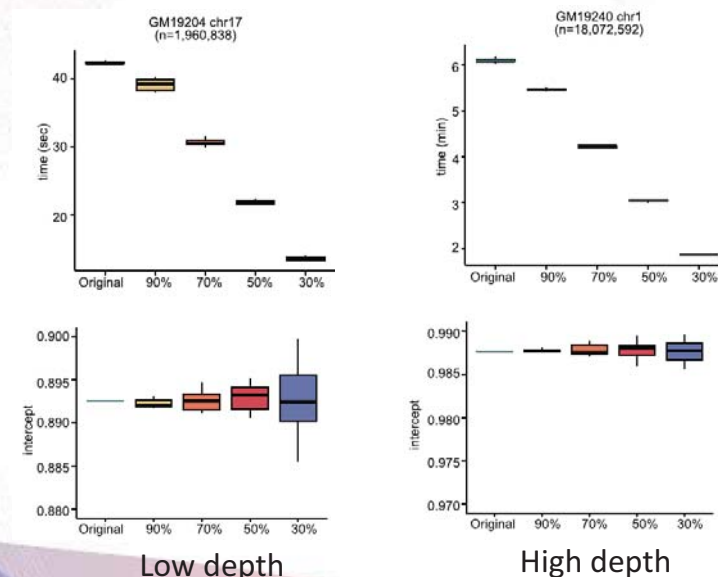
```
cov_result <- normCoverage(raw_data_filter, sample_ratio=0.5)
```

- Subsampling 된 데이터로 fitting parameter를 계산할 뿐 해당 parameter는 모든 data point에 적용됨 (데이터 손실 없음).

87

## 5. Further analysis

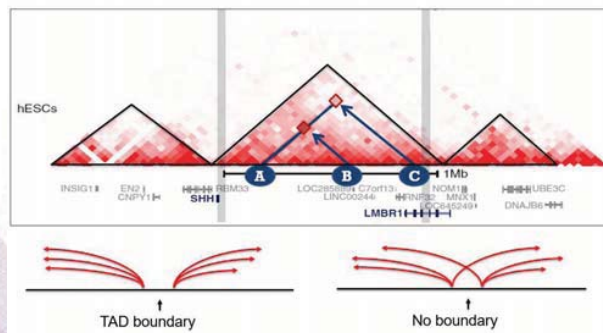
- Subsampling 비율에 비례하여 시간 감소.
- 낮은 depth에서 많이 subsampling 할수록 얻어진 parameter의 variance 증가.
- Benchmark 데이터에서는 50% 미만 subsampling부터 variance 크게 증가.
- Sample의 depth와 computing resource를 고려하여 조절 권장.



88

## 5. Further analysis

- DI score 계산
- 해당 bin의 up/downstream bias 측정
- 특정 bin에서 window 범위만큼 up (A) 과 down (B)에 map 된 read 비교
- Domain의 경계 지점에서 +/- 가 교차
- Dixon *et al.* 논문 (*Nature*, 2012)에서는 GMM & HMM으로 chromatin state를 추정해서 TAD 정의 (DI-HMM)



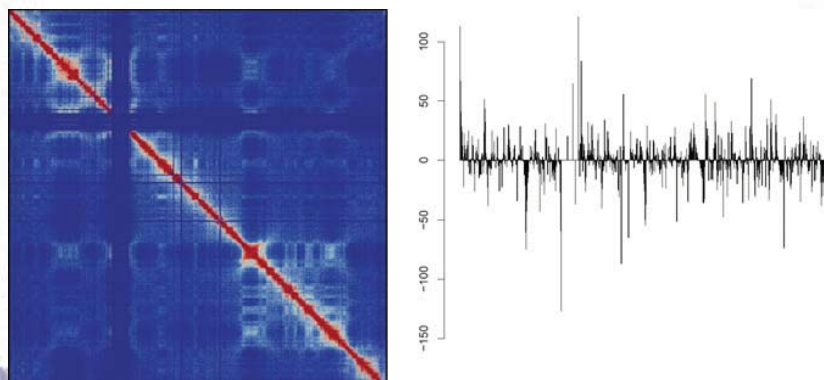
$$DI = \left( \frac{B-A}{|B-A|} \right) \left( \frac{(A-E)^2}{E} + \frac{(B-E)^2}{E} \right)$$

$$E = (A+B)/2$$

89

## 5. Further analysis

- DI score 계산
- pyDlcalc github에서 clone. (<https://github.com/kaistcbfg/pyDIscore>)
- covNorm output과 호환되어 구동 가능  
python pyDIscore.py --input-file <covnorm.gz> --chrname <chrname> --fai-file <fai file>
- 결과는 bed format으로 출력. chrname, bin\_start, bin\_end, DI\_score
- R 에서 DI\_score를 barplot으로 출력



90

## Summary

- coverage/distance 에 대한 normalization 및 QC 기능 제공.
- Fitting 기반의 유연한 설계로 Hi-C 외에도 capture Hi-C 등 variant 실험에 적용 가능.
- Data filter부터 significant interaction calling 까지 한 패키지로 수행 가능.
- Subsampling 기능을 이용한 scalability.
- DI score 등 연구실 개발/개발예정 SW들과 호환성.

Please check covNorm article for benchmark results.

Tutorial information available at Github



Computational and Structural Biotechnology  
Journal

Volume 19, 2021, Pages 3149-3159



covNorm: An R package for coverage based  
normalization of Hi-C and capture Hi-C data

Kyukwang Kim, Inkyung Jung  

<https://doi.org/10.1016/j.csbj.2021.05.041>

<https://github.com/kaistcbfg/covNormRpkg>