KSBi-BIML 2021

Bioinformatics & Machine Learning (BIML)
Workshop for Life Scientists

생물정보학 & 머쉰러닝 워크샵(온라인)

3D Epigenome Data Analysis

정인경

SBi 한국생명정보학회 Korean Society for Bioinformatics
서울대학교 생물정보연구소 SNU BIOINFORMATICS INSTITUTE
Gene·2·us

Bioinformatics & Machine Learning for Life Scientists

BIML-2021

안녕하십니까?

한국생명정보학회의 동계 워크샵인 BIML-2021을 2월 15부터 2월 19일까지 개최합니다. 생명정보학 분야의 융합이론 보급과 실무역량 강화를 위해 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하였으며 올해로 7차를 맞이하게 되었습니다. 유례가 없는 코로나 대유행으로 인해 올해의 BIML 워크숍은 온라인으로 준비했습니다. 생생한 현장 강의에서만 느낄 수 있는 강의자와 수강생 사이의 상호교감을 가질 수 없다는 단점이 있지만, 온라인 강의의 여러 장점을 살려서 최근 생명정보학에서 주목받고 있는 거의 모든 분야를 망라한 강의를 준비했습니다. 또한 온라인 강의의 한계를 극복하기 위해서 실시간 Q&A 세션 또한 마련했습니다.

BIML 워크샵은 전통적으로 크게 생명정보학과 AI, 두 개의 분야로 구성되어오고 있으며 올해 역시 유사한 방식을 채택했습니다. AI 분야는 Probabilistic Modeling, Dimensionality Reduction, SVM 등과 같은 전통적인 Machine Learning부터 Deep Learning을 이용한 신약개발 및 유전체 연구까지 다양한 내용을 다루고 있습니다. 생명정보학 분야로는, Proteomics, Chemoinformatics, Single Cell Genomics, Cancer Genomics, Network Biology, 3D Epigenomics, RNA Biology, Microbiome 등 거의 모든 분야가 포함되어 있습니다. 연사들은 각 분야 최고의 전문가들이라 자부합니다.

이번 BIML-2021을 준비하기까지 너무나 많은 수고를 해주신 BIML-2021 운영위원회의 김태민 교수님, 류성호 교수님, 남진우 교수님, 백대현 교수님께 커다란 감사를 드립니다. 또한 재정적 도움을 주신, 김선 교수님 (AI-based Drug Discovery), 류성호 교수님, 남진우 교수님께 감사를 표시하고 싶습니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 강의자료를 만드는데 노력하셨을 뿐만 아니라 실시간 온라인 Q&A 세션까지 참여해 수고해 주시는 모든 연사분들께 깊이 감사드립니다.

2021년 2월

한국생명정보학회장 김동섭

강의개요

3D Epigenome Data Analysis

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. 이러한 염색질 3차구조는 게놈의 후성유전적 변화와 밀접한 연관이 있기에 최근 염색질 3차구조와 게놈의 후성유전적 변화를 통합 분석하려는 '3D epigenome' 연구가 급격하게 발전하고 있다.

본 강의에서는 염색질 3차구조를 중심으로 관련 이론, 실험 방법, 그리고 기본 데이터 분석을 실습과 함께 숙지하고자 한다. 간략하게 후성유전학 대한 강의 후, 염색질 3차구조에 대한 전반적인 소개와 관련 데이터 분석 방법을 소개하고 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:

- 후성유전학/염색질 3차구조 개요
- 염색질 3차구조 데이터 분석 방법
- 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: 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
Phone Number +82-42-350-7314

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

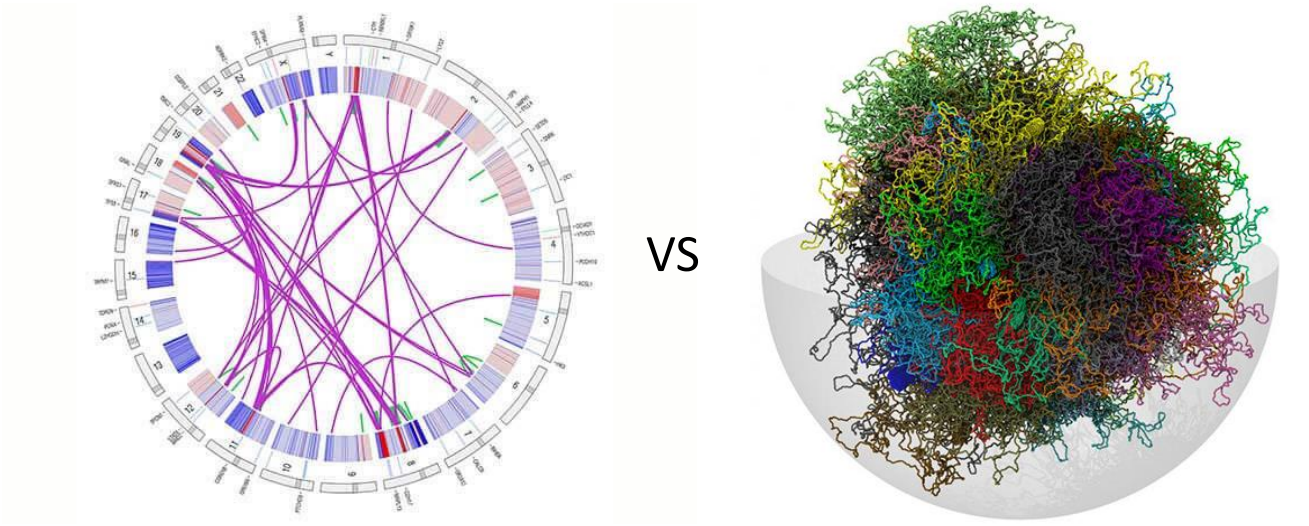
KSBi-BIML 2021

3D Epigenome Data Analysis – 염색질 3차구조 개요

정인경(KAIST)

Contents

Unsolved question: Mapping 1D Genome to 3D genome

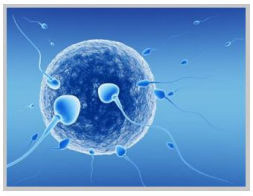


VS

Can we better understand genome through 3D genome?

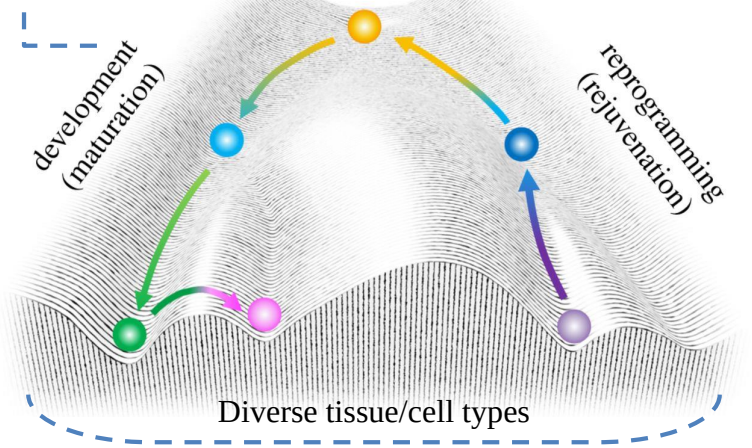
Epigenetic gene regulation determines cell fate

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

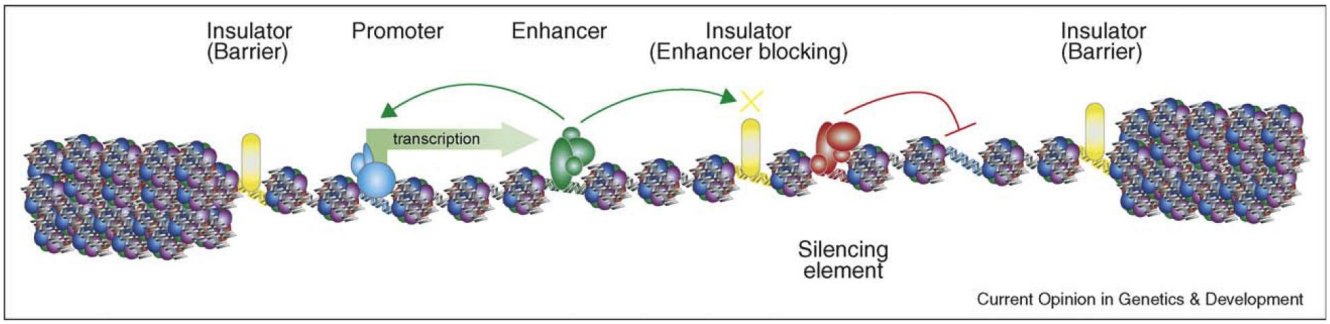
Conrad Waddington
(1905-1975)

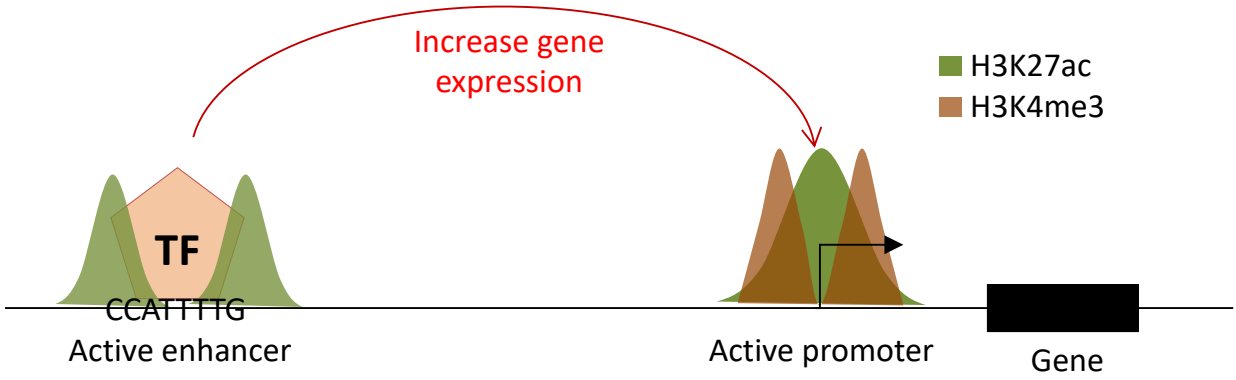


Leung, D. & Jung, I. (2015) Nature
Jung, I. (2019) Nature Genetics

Waddington's epigenetic landscape (Evolution, 1956)

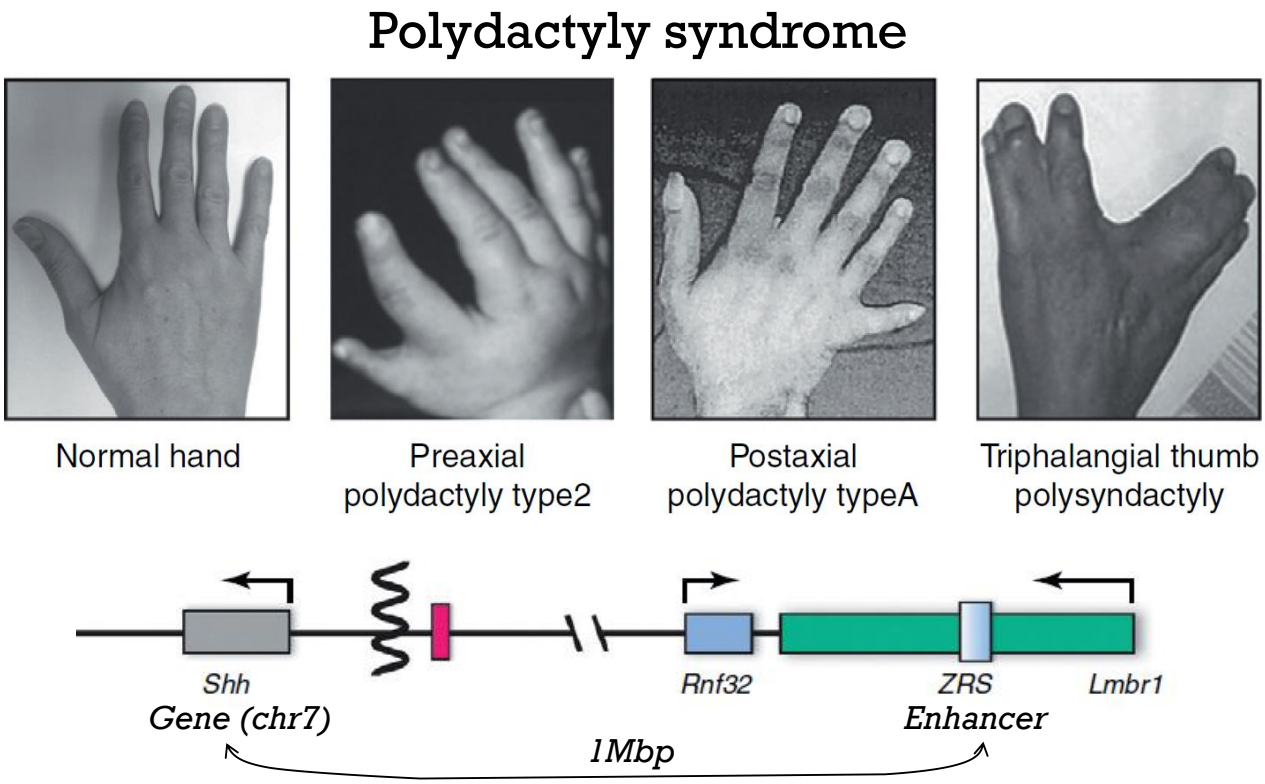
"Enhancer" is a major player in epigenetic gene regulation



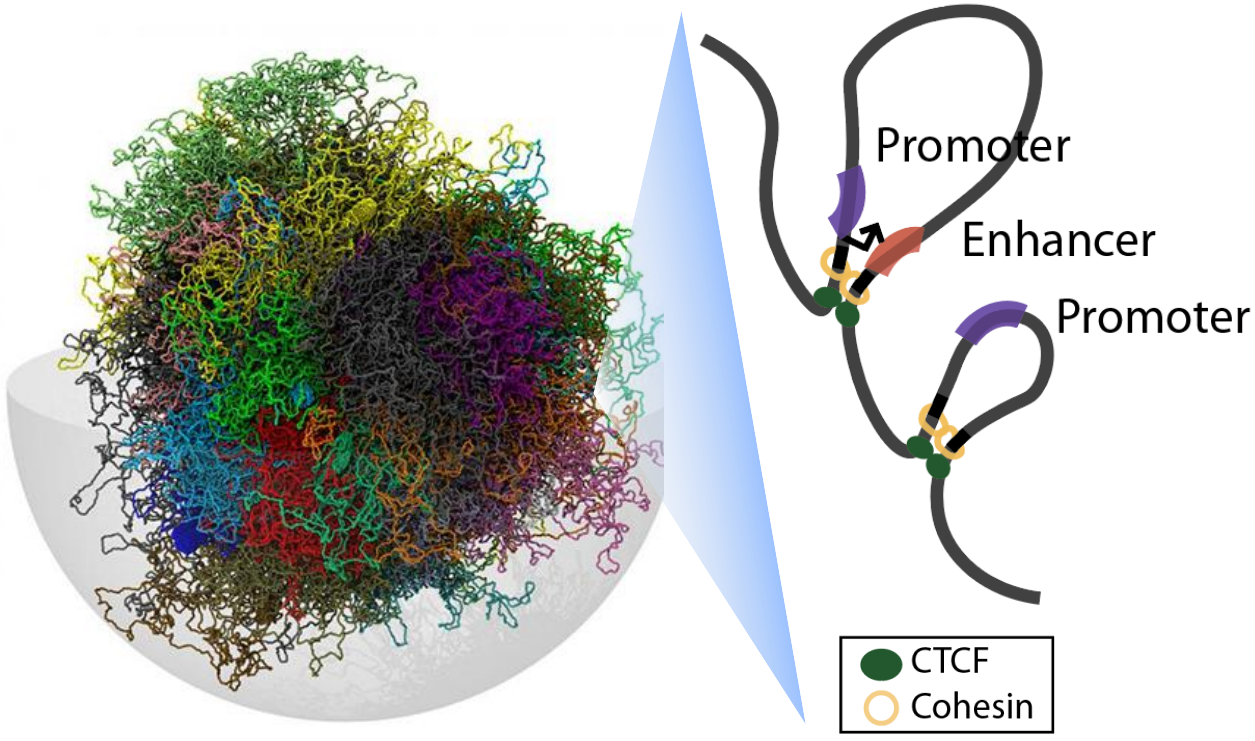


A epigenetic switch of gene expression

Enhancers can control distal target gene expression



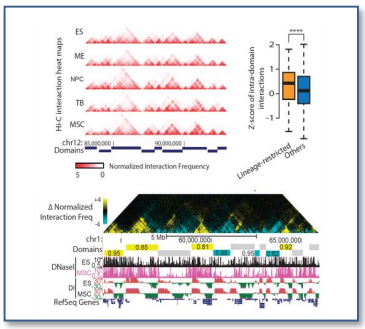
How do enhancers control distal target gene expression?



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

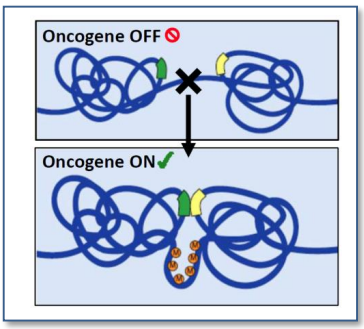
Genome functions are tightly coupled with 3D chromatin structure

Cellular differentiation



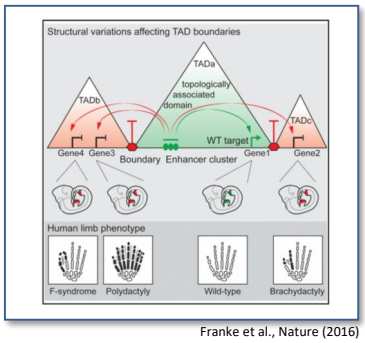
Dixon, JR*., **Jung, I***., et al., Nature (2015)

Oncogene activation



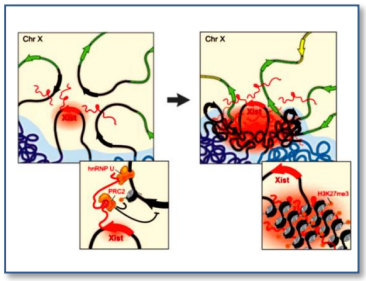
Hnisz et al., Science (2016)

Congenital disorder



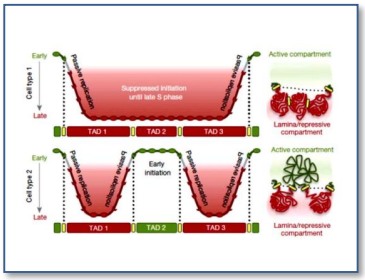
Franke et al., Nature (2016)

X-chromosome inactivation



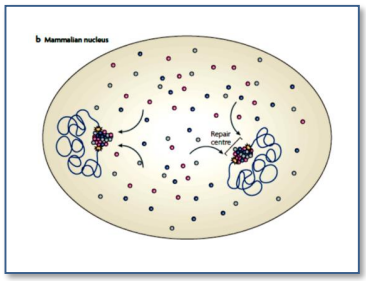
Engreitz et al., Science (2013)

DNA replication



Pope et al., Nature (2014)

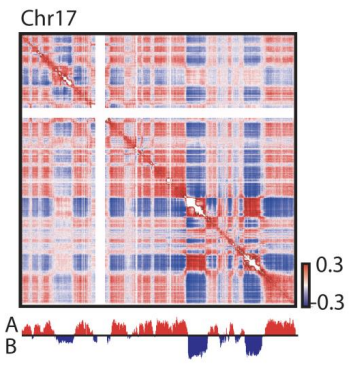
DNA repair



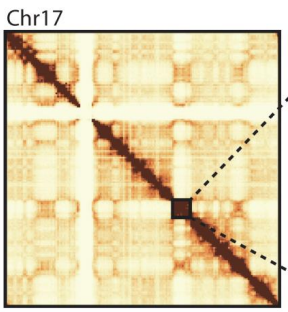
Misteli & Soutoglou, Mol Cell Biol (2009)

Multi-layered 3D genome organization

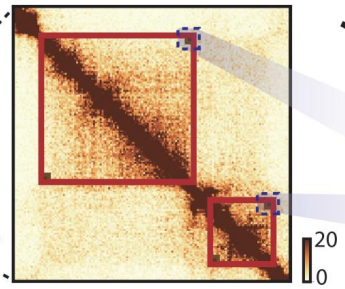
Compartment A/B



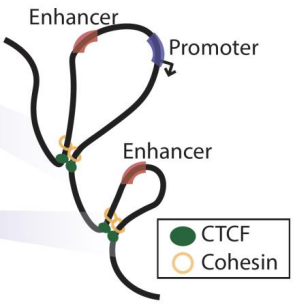
Individual chromosome



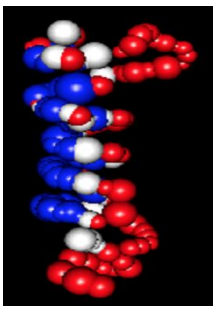
TAD (Topologically associating domains)



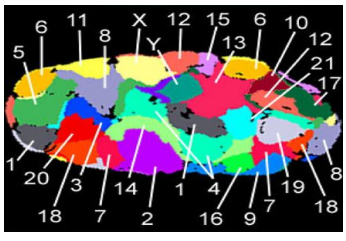
E-P interactions



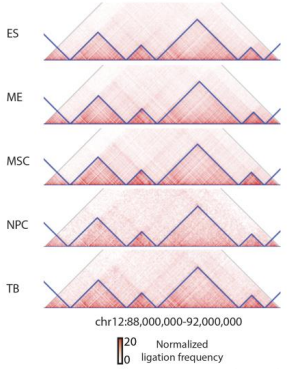
Kim et al., 2019



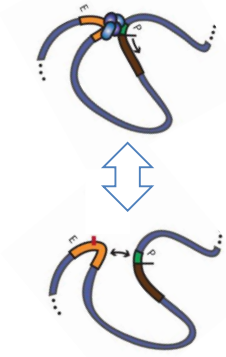
Hu et al., (2013)



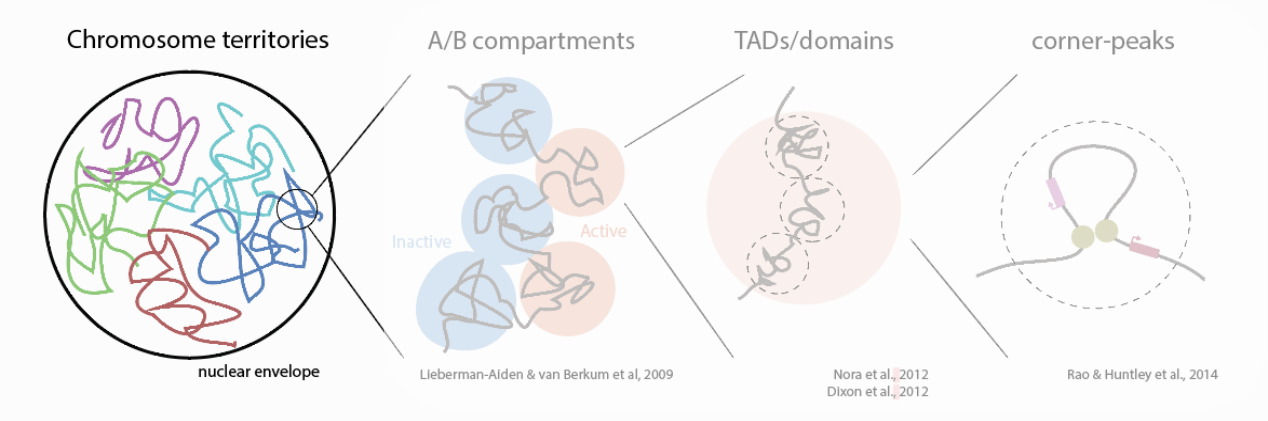
Bolzer et al., (2005)



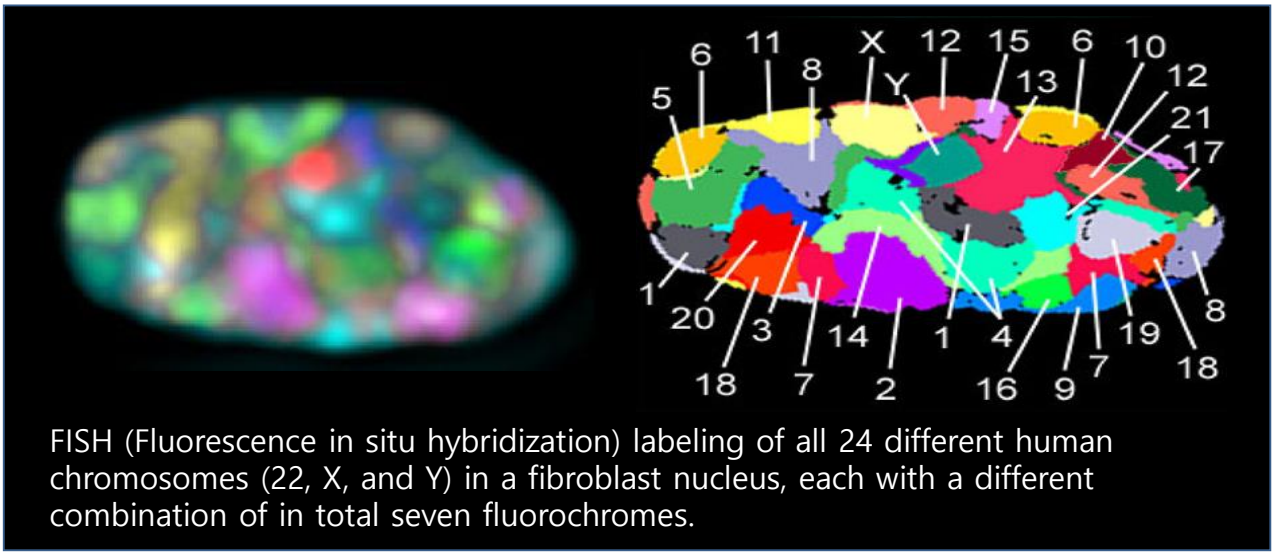
Dixon et al., (2015)



Jung et al., (2019)



A theory of chromosome territory



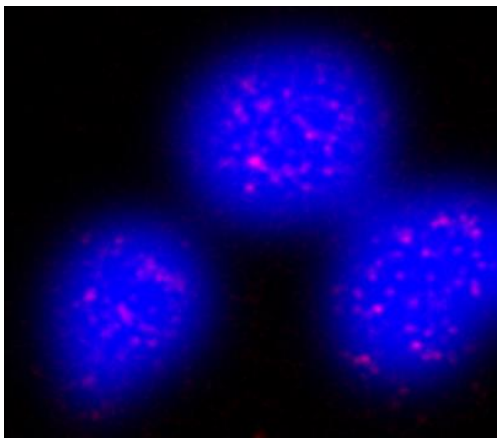
FISH (Fluorescence in situ hybridization) labeling of all 24 different human chromosomes (22, X, and Y) in a fibroblast nucleus, each with a different combination of in total seven fluorochromes.

Bolzer et al., (2005)

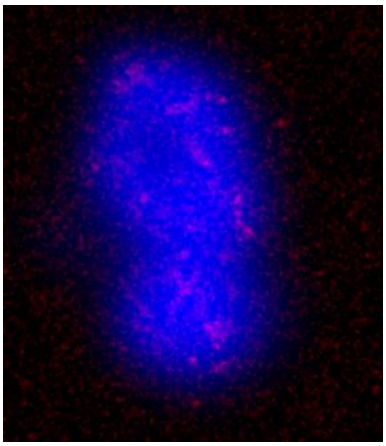
What is a driving force that directs macro-scale 3D genome?

Blue : Nucleus
Red : PRC2 hub

WT



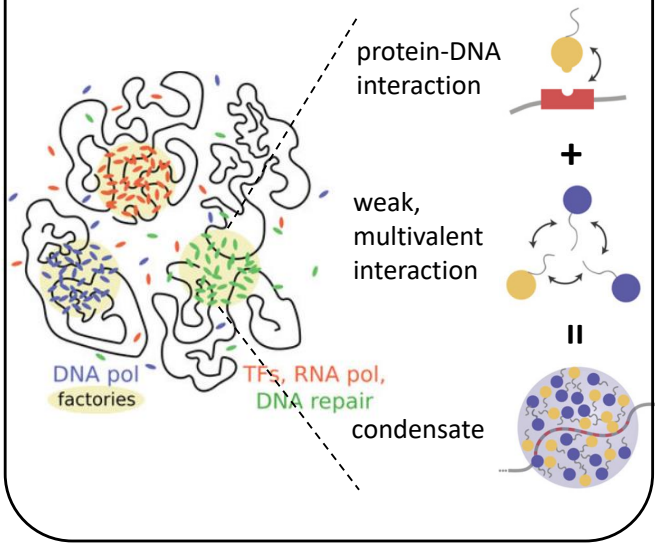
PRC2 mutant



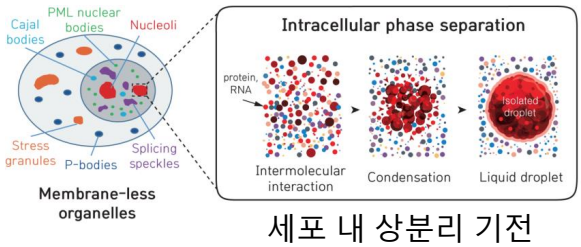
Chul-hwan Lee, SNU, unpublished

Phase separation, a new paradigm of 3D genome organization

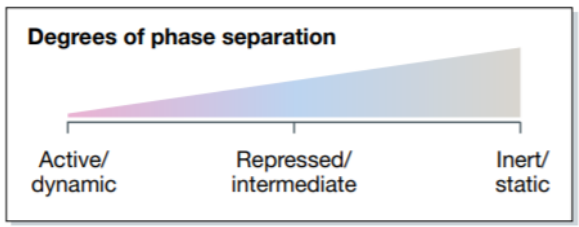
핵 내 다양한 응집체 (condensates)



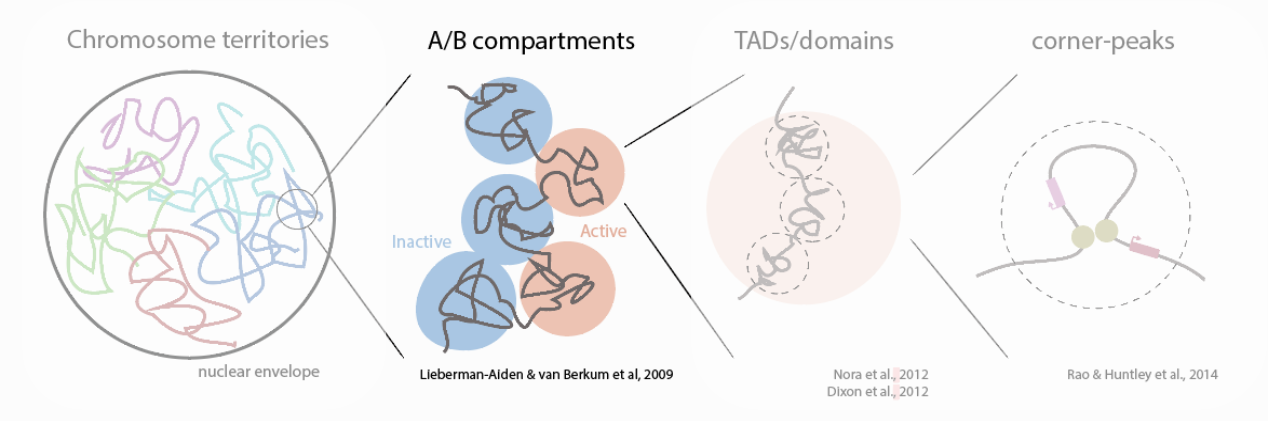
응집체와 염색질 3차구조



세포 내 상분리 기전



상분리 응집체와 염색질 3차 구조의 관계

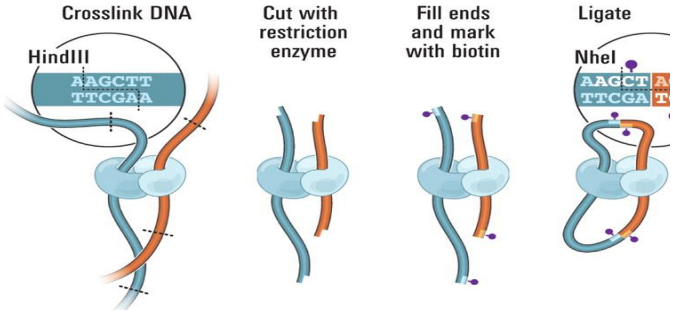


Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

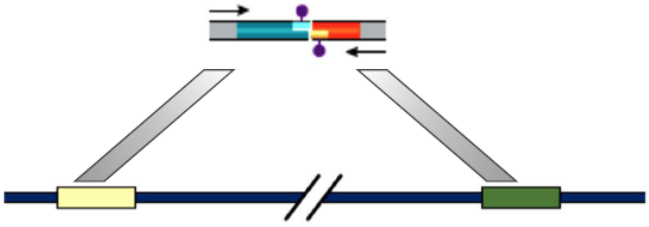
Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[5]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†

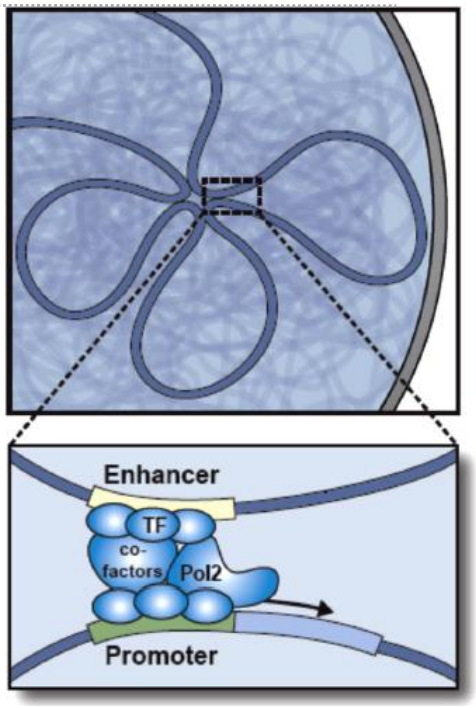
How can we investigate 3D genome organization?

Hi-C: **Hi**gh-throughput **c**hromatin conformation capture (3C)



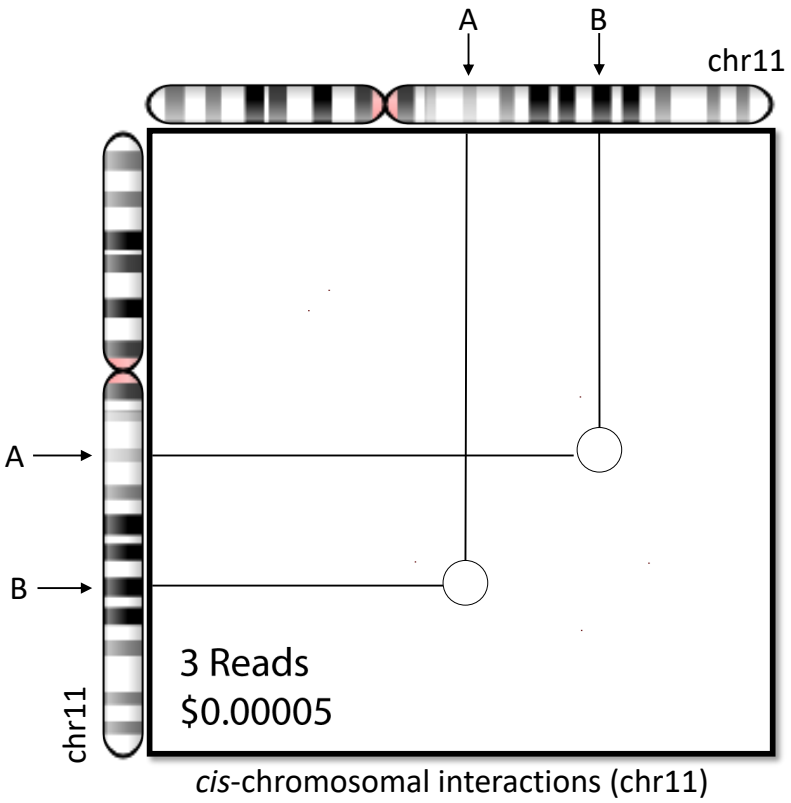
Cross-linking → Digestion → Re-l





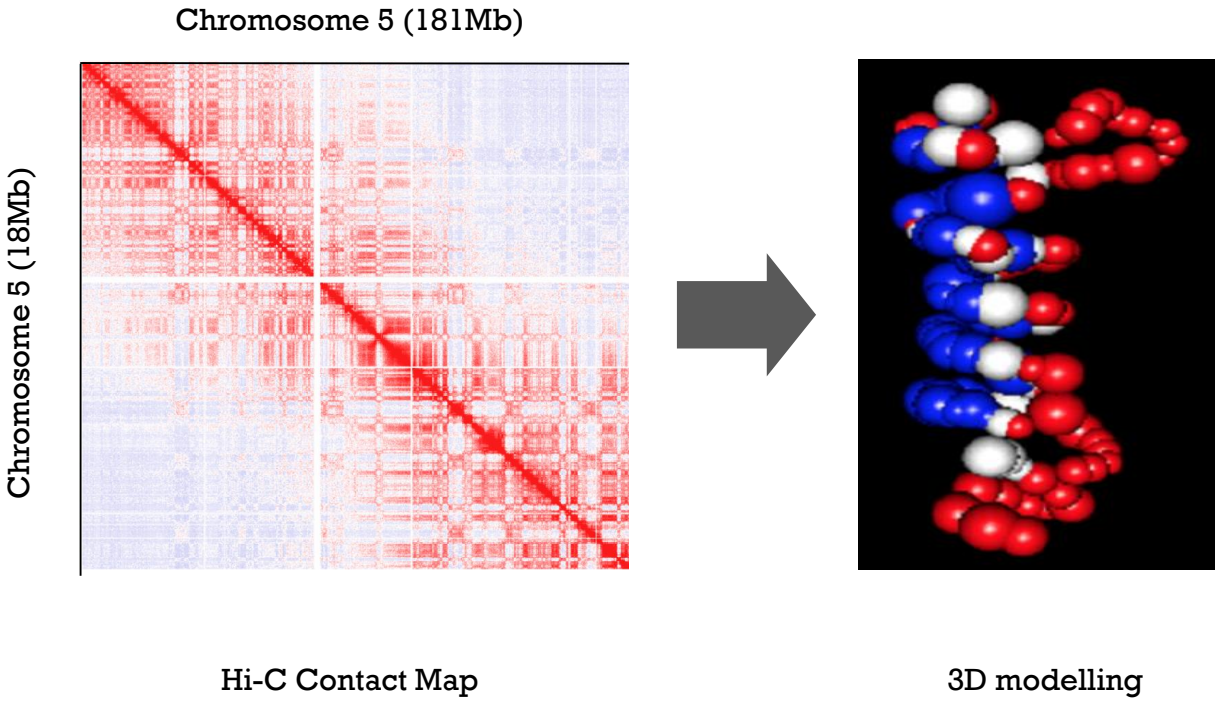
Lieberman et al., Science (2009)

Hi-C contact map to visualize 3D genome



cis-chromosomal interactions (chr11)

Modeling 3D chromatin structure from Hi-C contact map

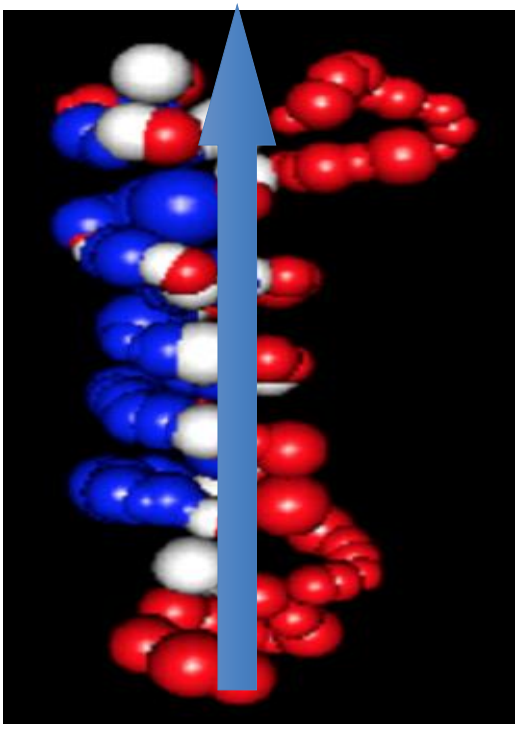


Hi-C Contact Map

3D modelling

What is a major structural component?

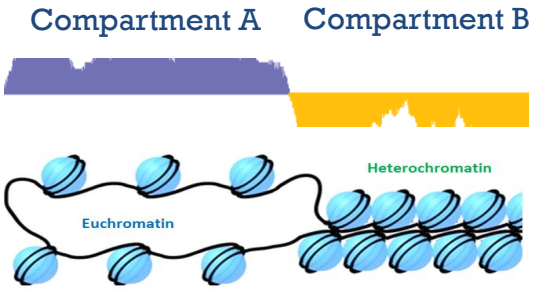
Compartment B

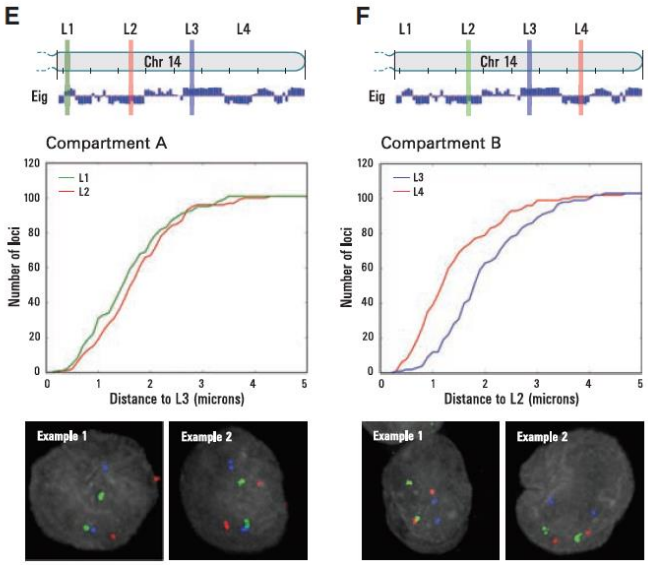


Compartment A

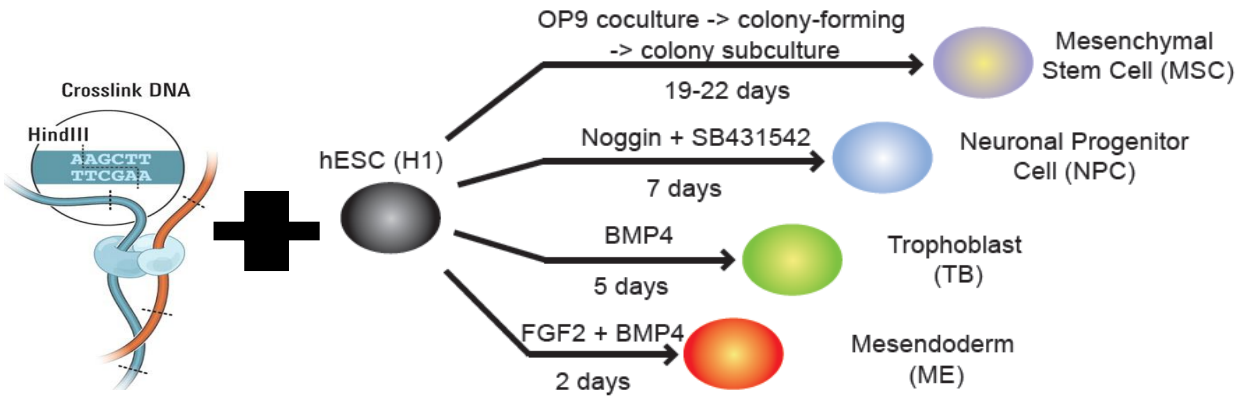
Two major compositions of chromatin structure: Compartment A/B

The loci in the same compartment showed spatial proximity





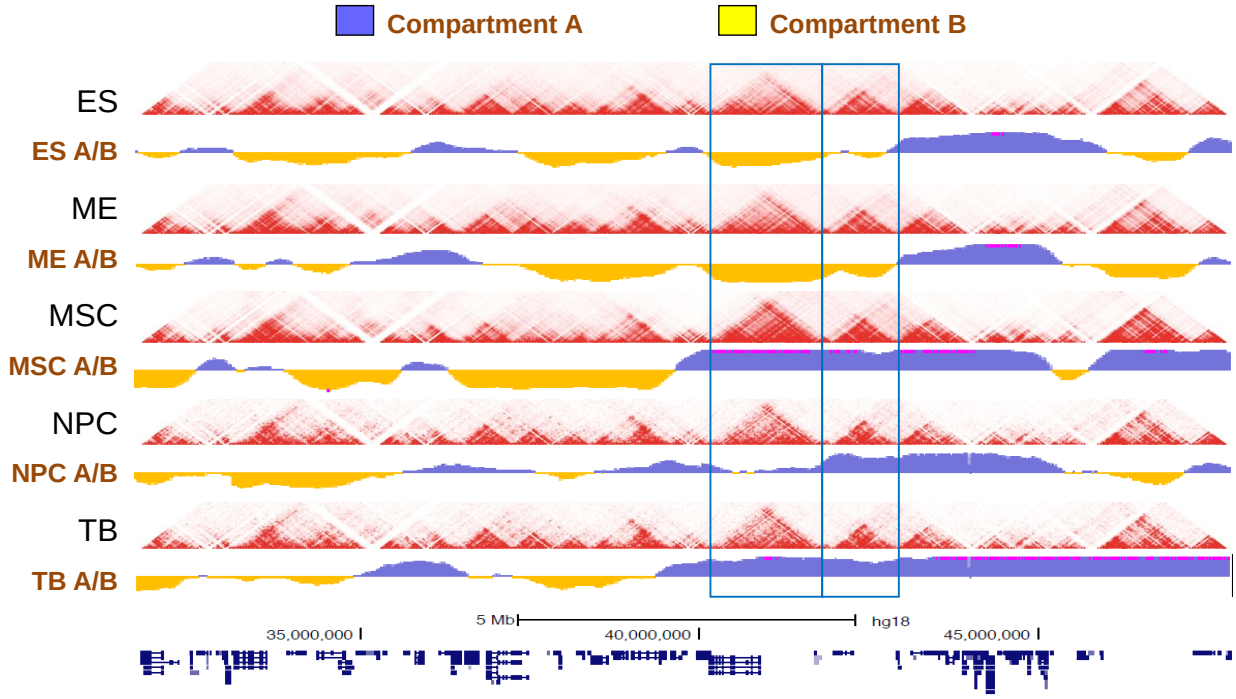
Compartment A/B dynamics during stem cell differentiation



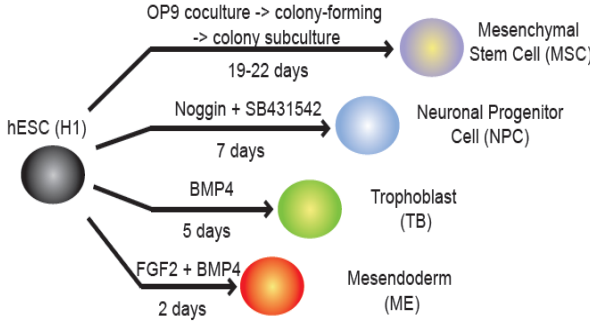
Perform Hi-C experiment
(3.85 billion unique read pairs, on average 770 million)

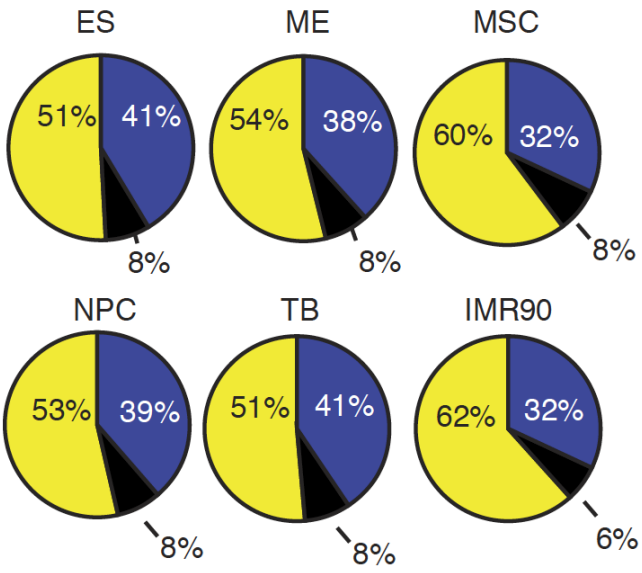
Dixon, JR., Jung, I., et al., Nature (2015)

Compartment A/B patterns are highly dynamic

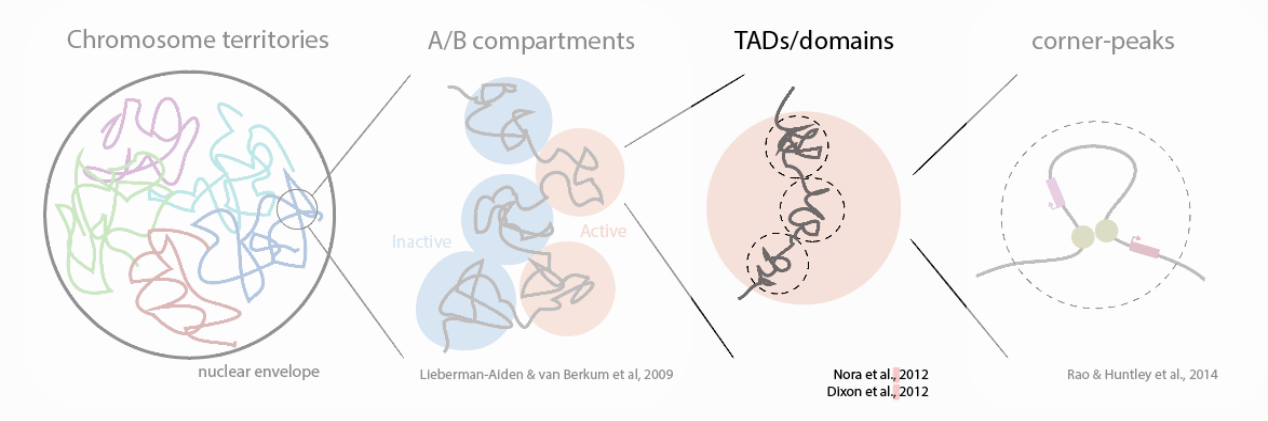


Fraction of compartment A/B in each cell-type

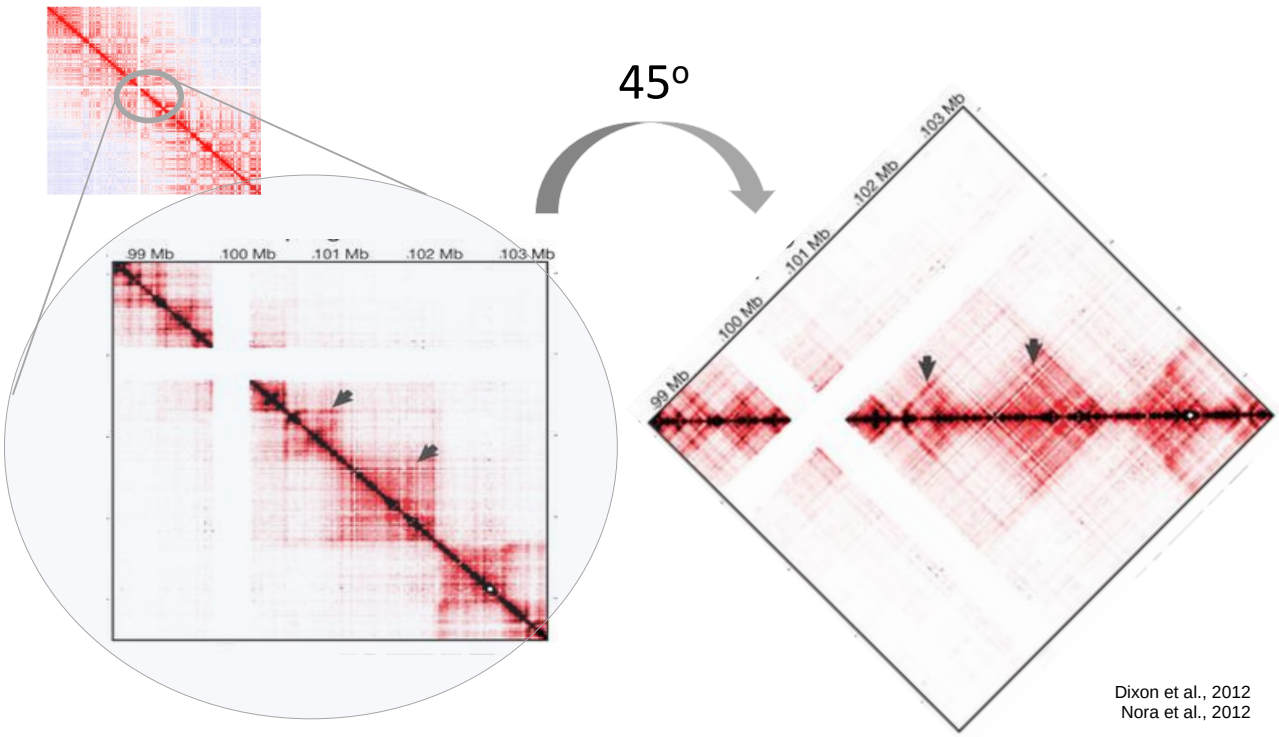




Fraction of genome marked as compartment A (blue)and B (yellow)

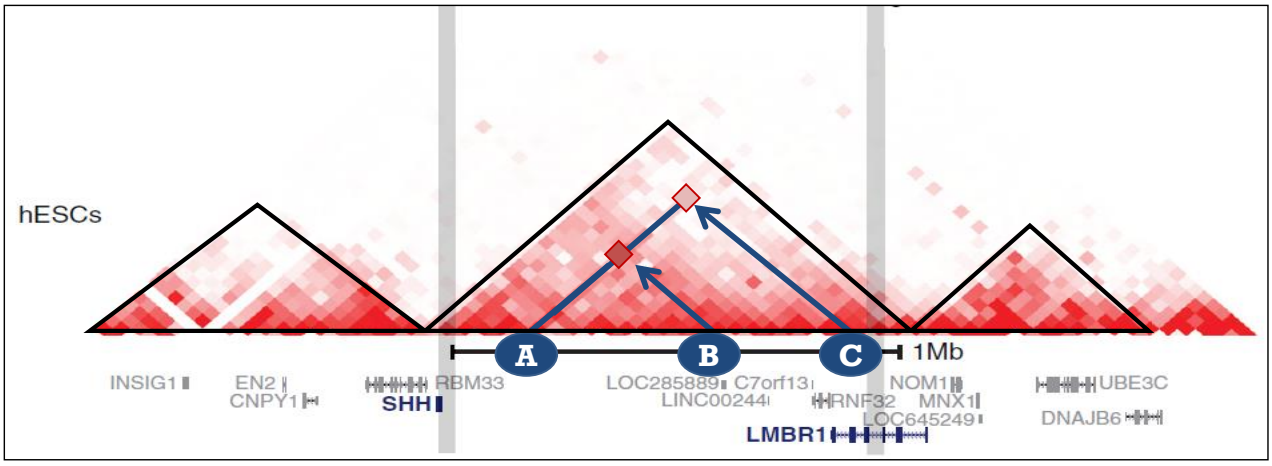


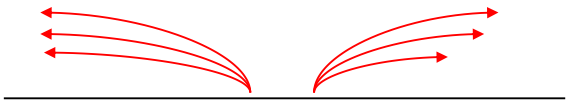
Topologically Associating Domains (TADs)



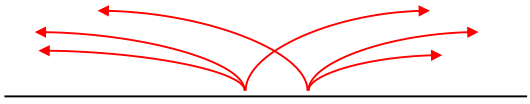
Dixon et al., 2012
Nora et al., 2012

Topologically Associating Domains (TADs)



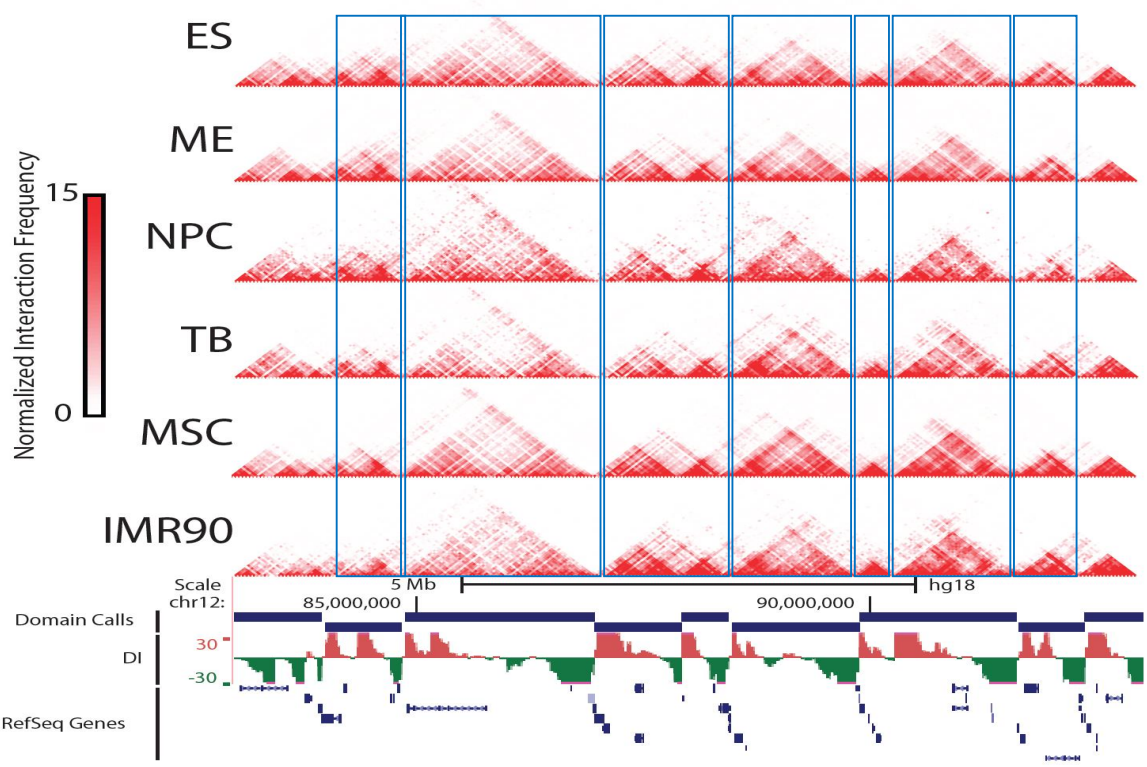


TAD boundary

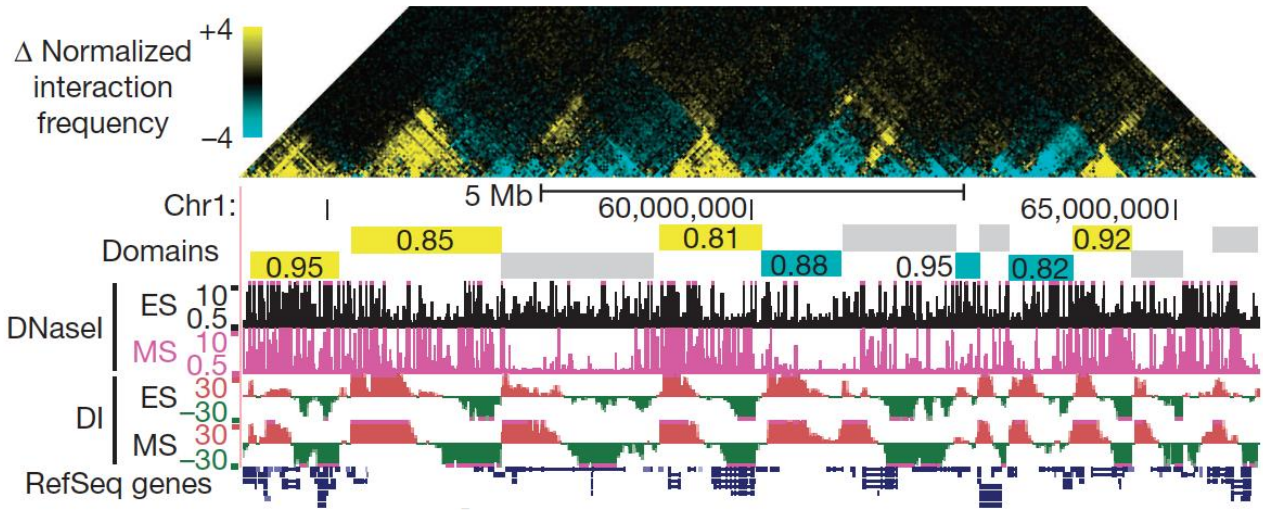


No boundary

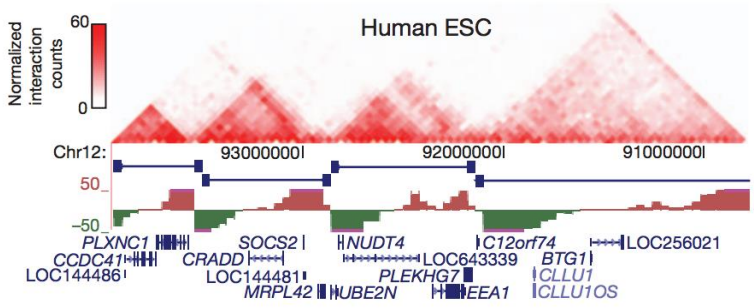
TAD boundaries are well maintained during differentiation

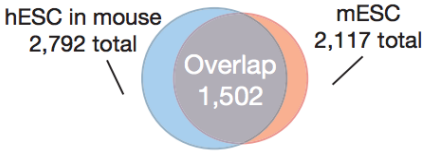


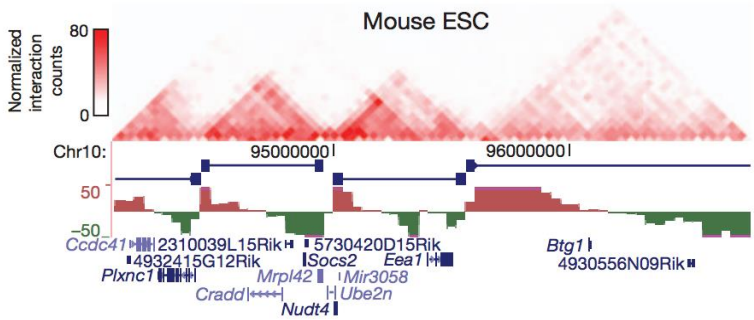
TAD-wise interaction changes during cellular differentiation

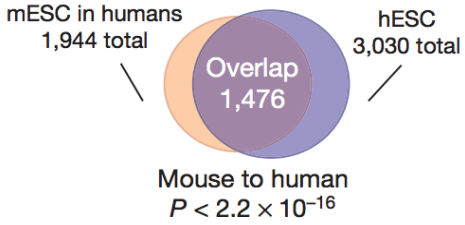


TAD boundaries are evolutionarily well conserved



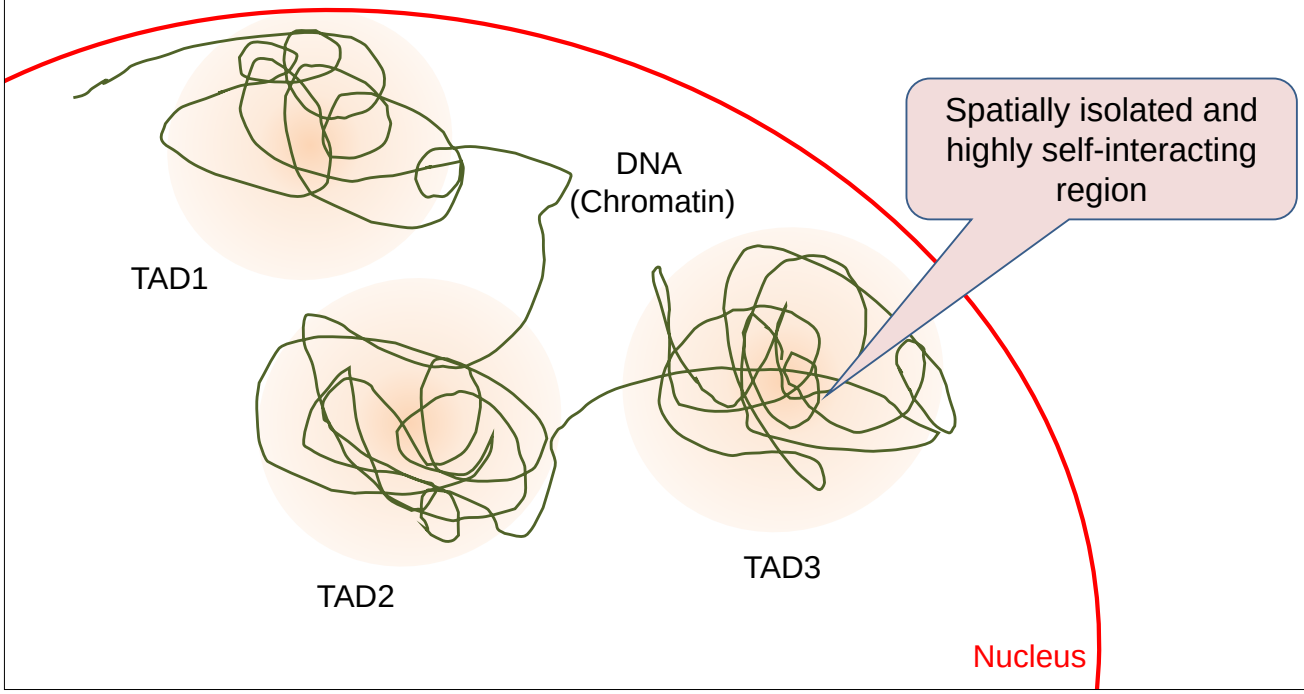




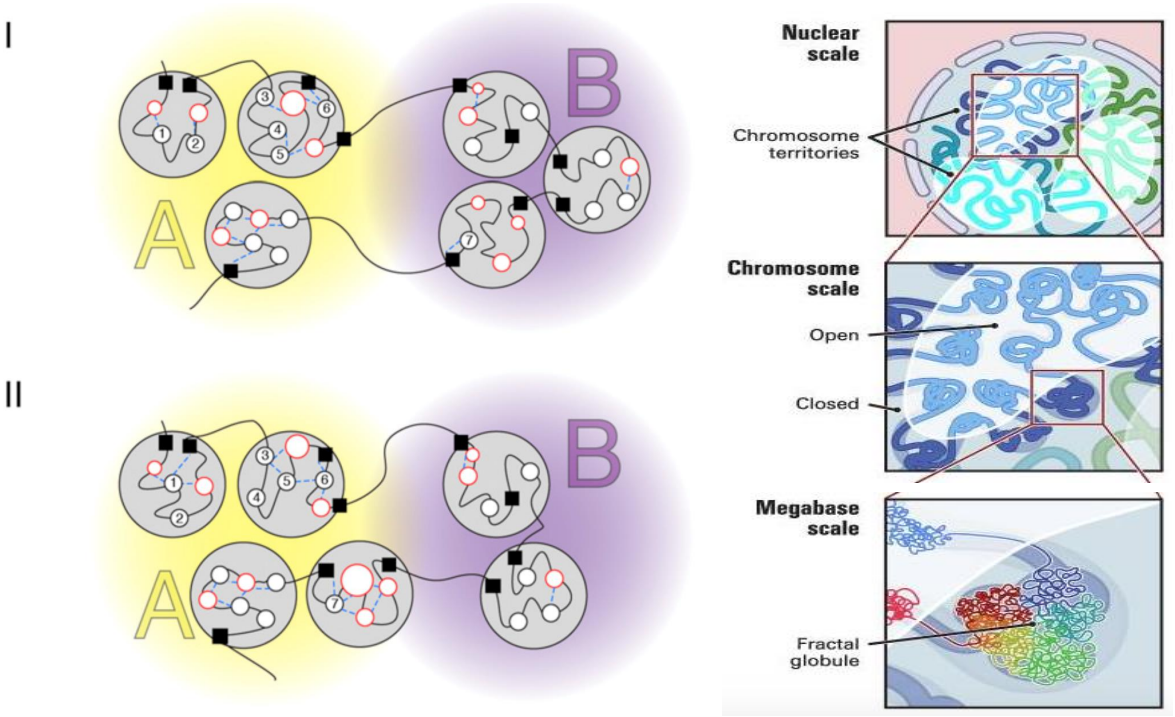


TAD is a basic unit of 3D chromatin structure

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

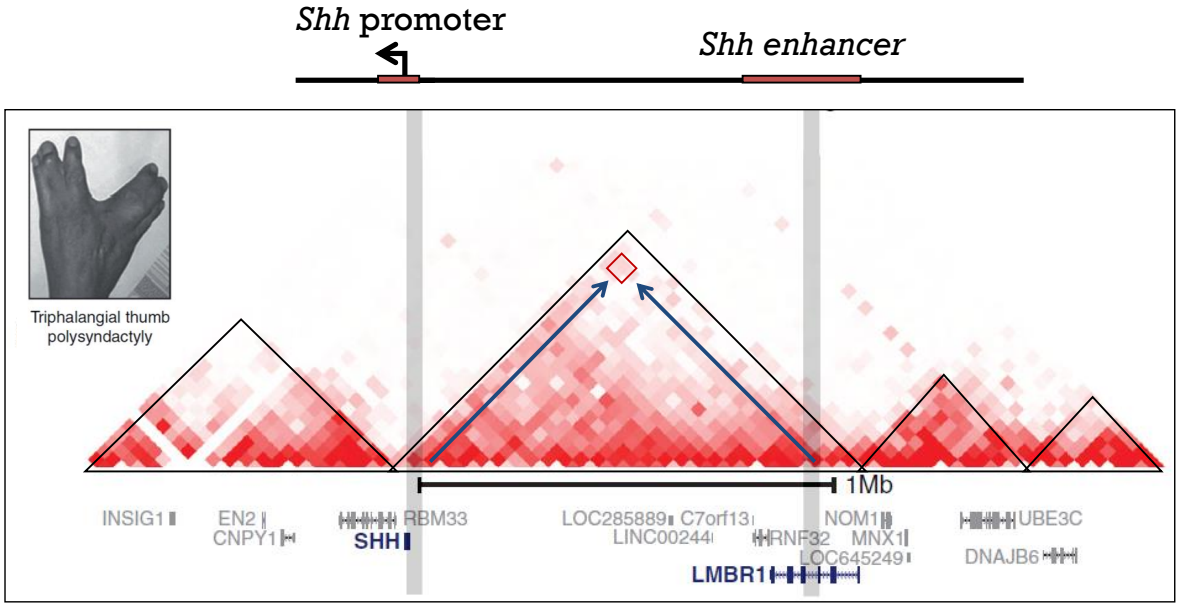


The relationship between TAD and Compartment A/B



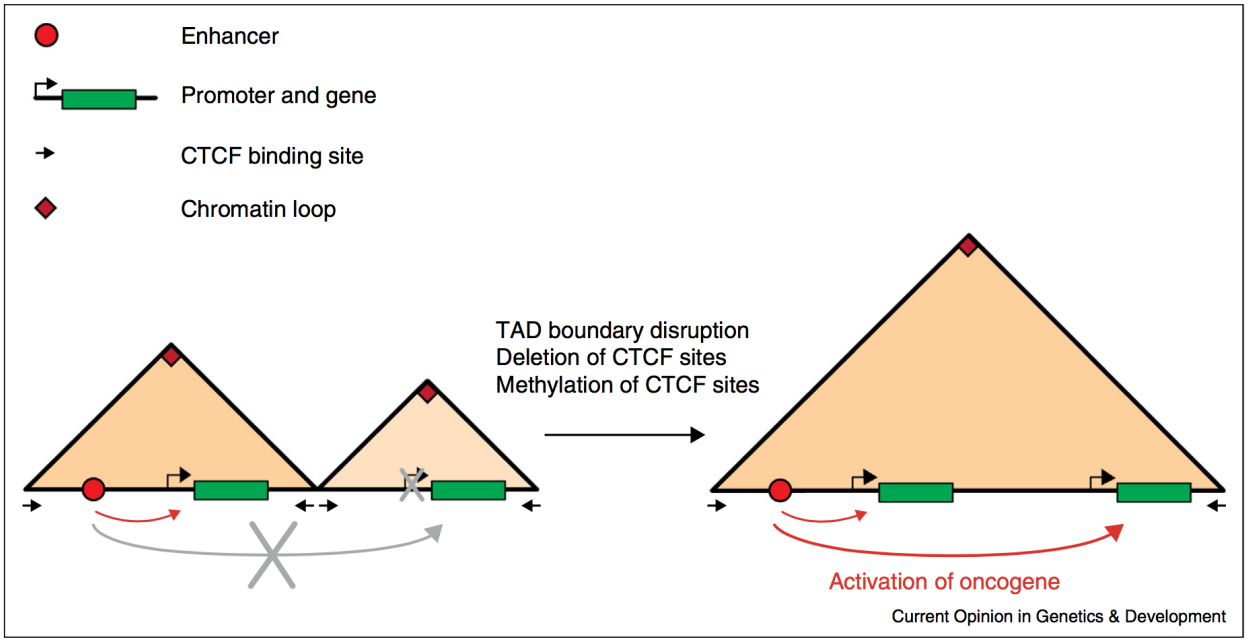
What is the functional role of TADs?

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



Activation of proto-oncogenes by disruption of TAD boundary

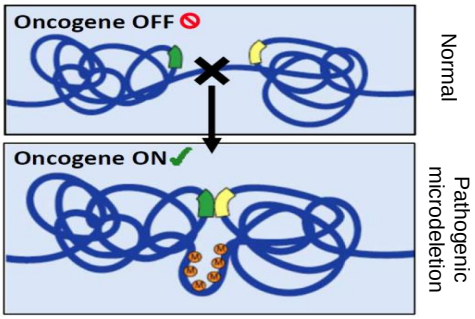
CANCER

Activation of proto-oncogenes by disruption of chromosome neighborhoods

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

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.

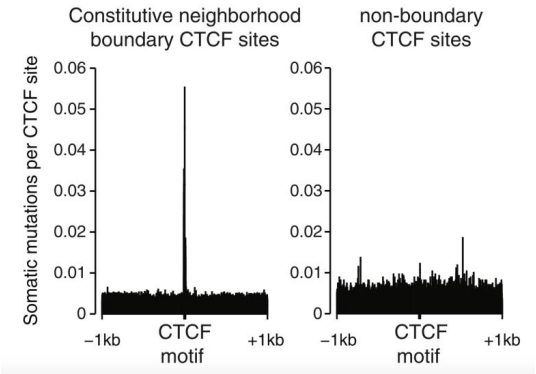
TAD boundary disruption induces oncogene activation



Hnisz et al., 2016 (Science)

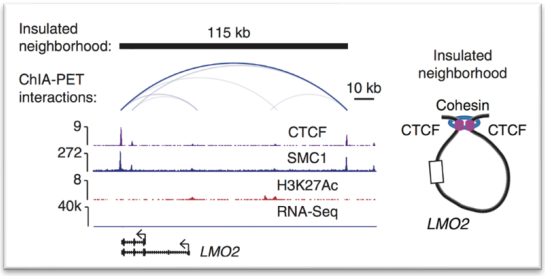
T-cell acute lymphoblastic leukemia (T-ALL)

Identification of recurrent deletion at TAD boundary

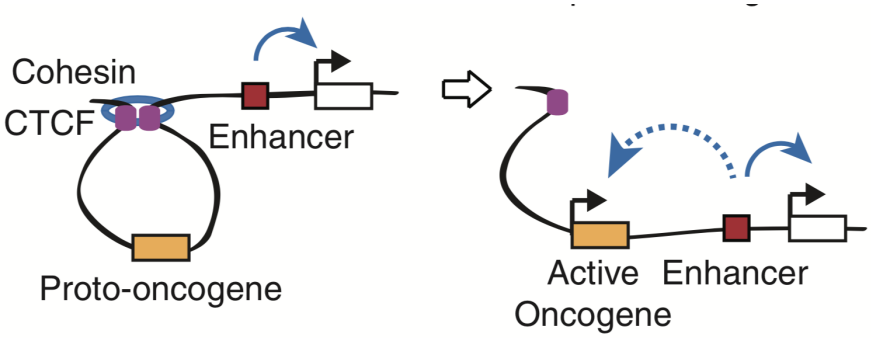


- Identified recurrent deletions in T-ALL genome
- 113 of 438 overlapped at least one boundary of TAD

TAD at the silent LMO2 locus

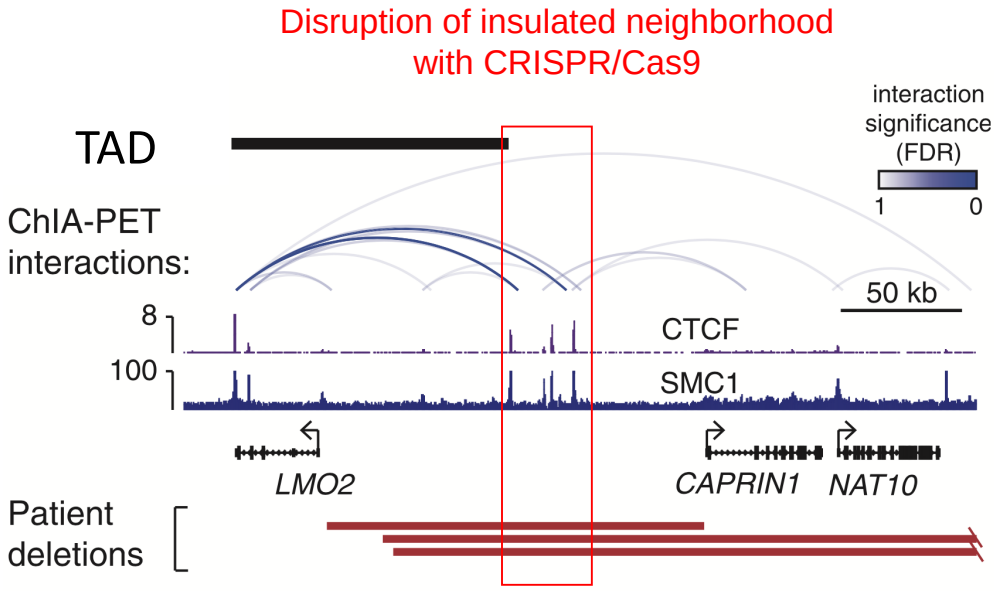


LMO2 (LIM Domain Only 2) is a protein coding gene, having a central and crucial role in hematopoietic development (T-ALL pathogenesis gene)

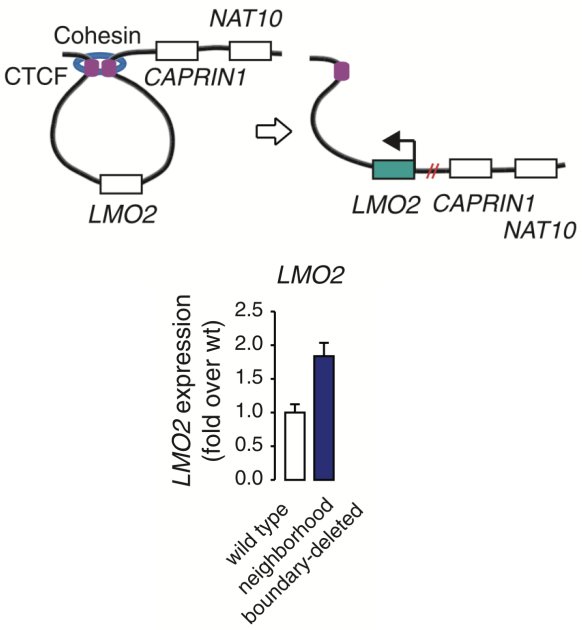


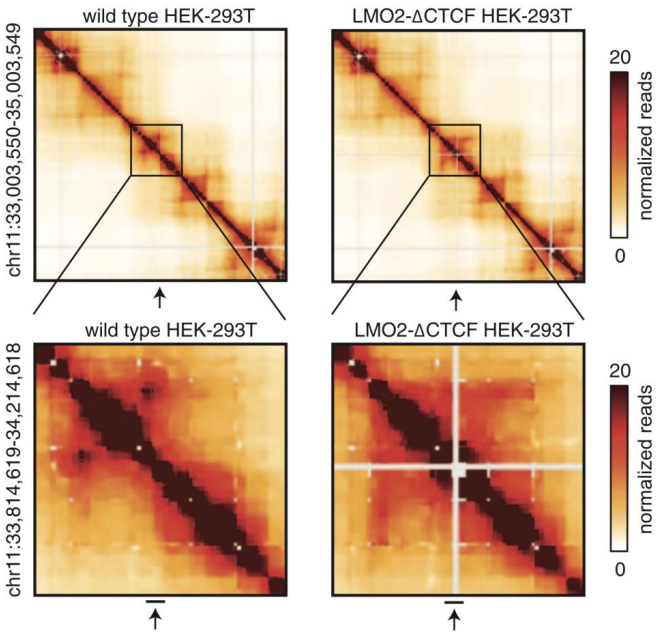
Can disruption of TAD boundary (TAD fusion) activate proto-oncogenes through enhancer-hijacking?

Disruption of TAD boundary by CRISPR/Cas9

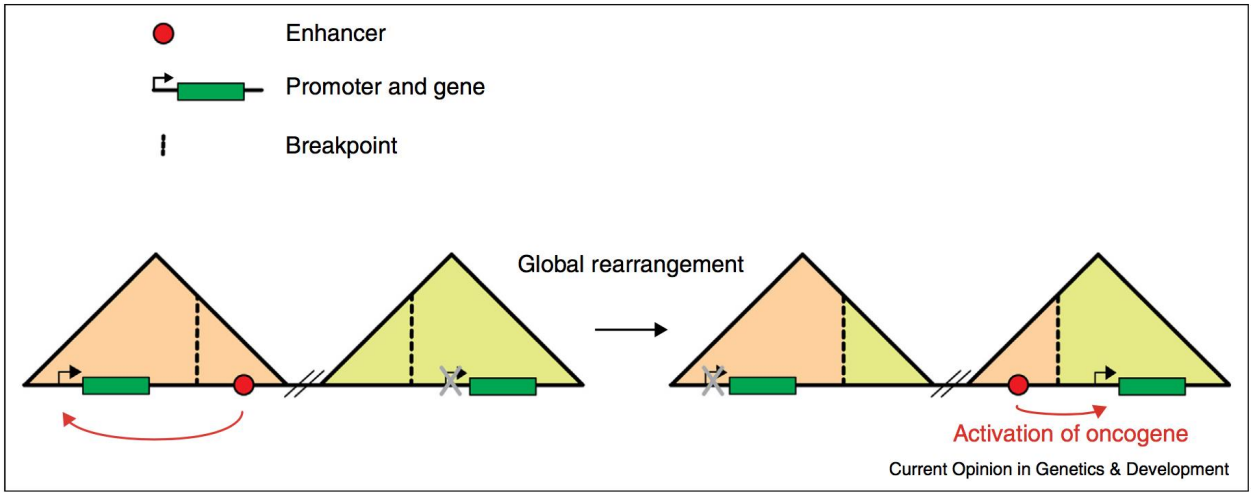


Disruption of TAD boundary activates LMO2

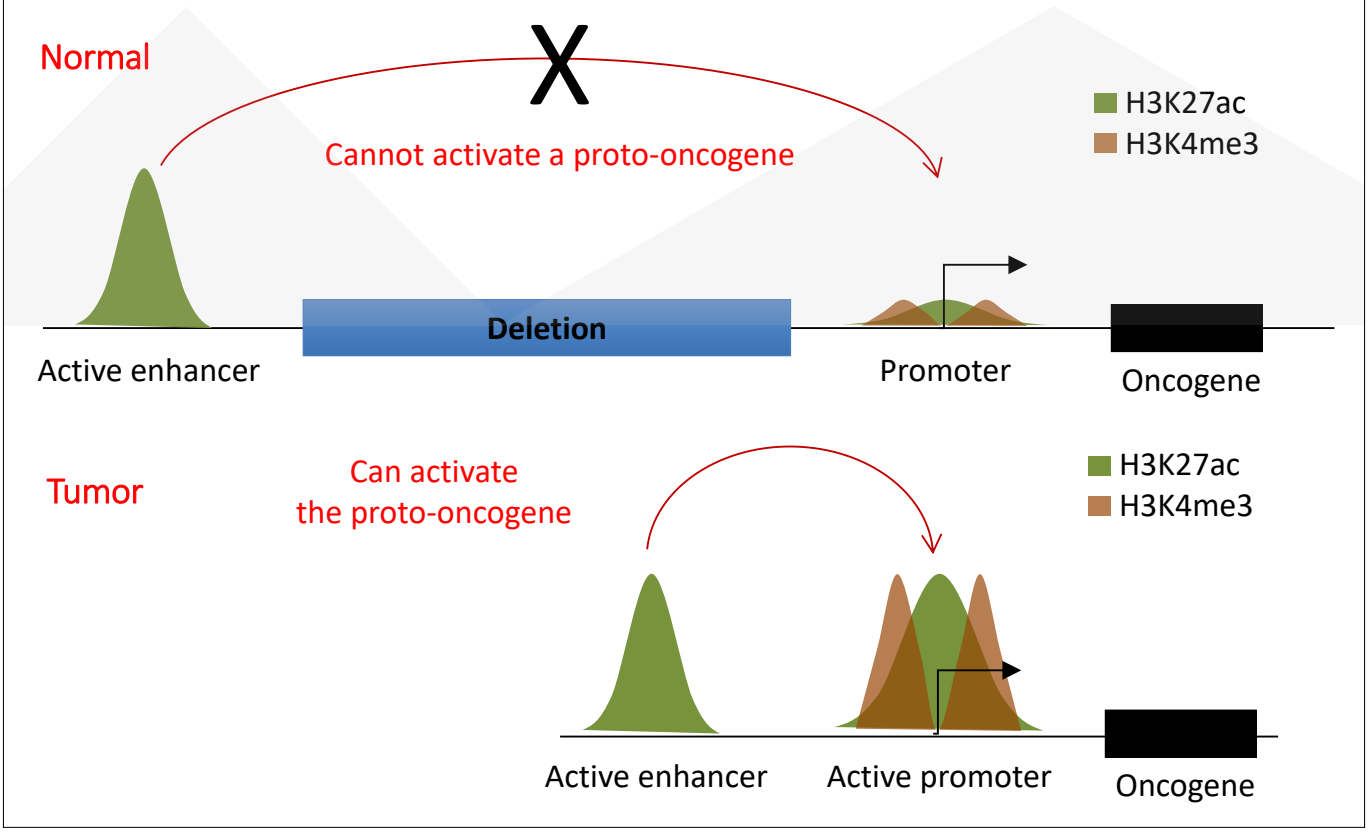




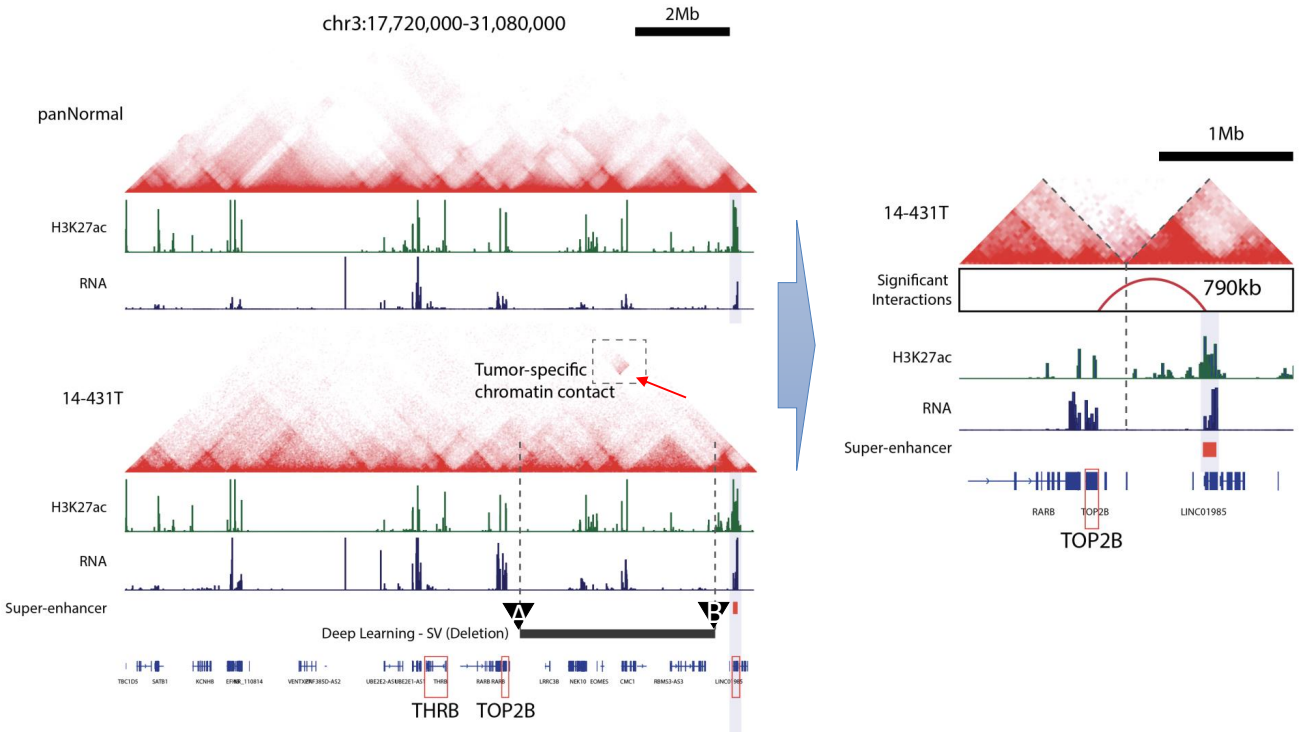
TAD disruption as oncogenic driver – Model 2



TAD rearrangments as oncogenic driver – Model 2

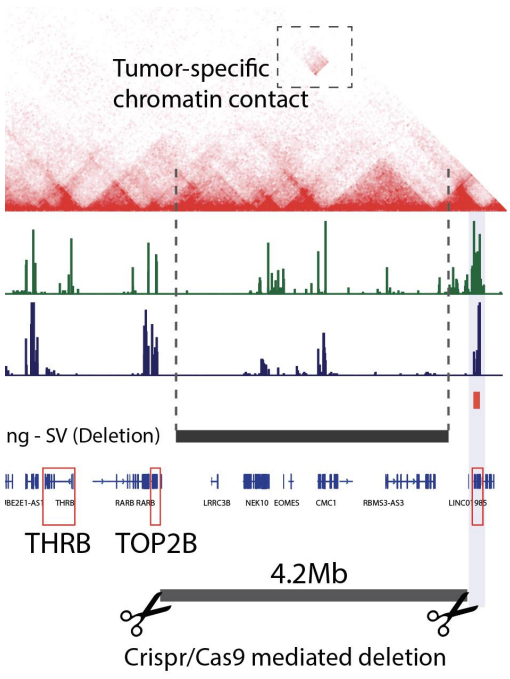


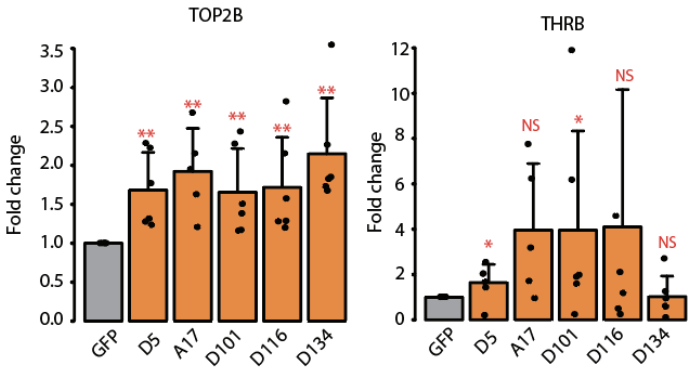
TAD fusion driven by a large-scale deletion activates a proto-oncogene

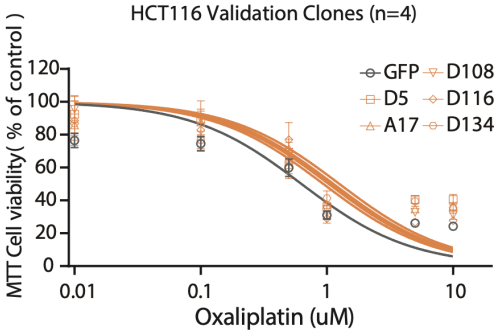


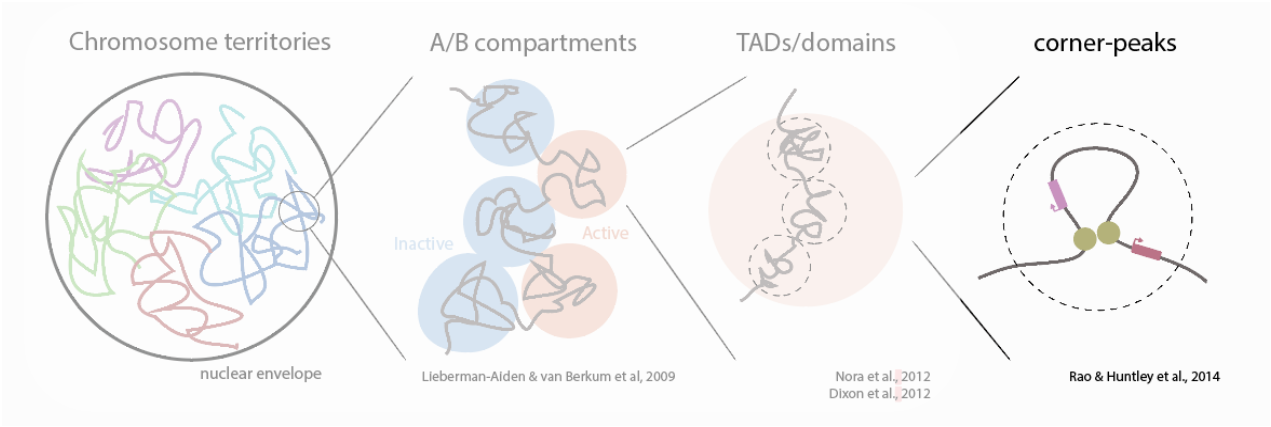
Unpublished data

CRISPR/Cas9 mediated site-specific chromosomal rearrangement at the TOP2B and distal super-enhancer into a single TAD

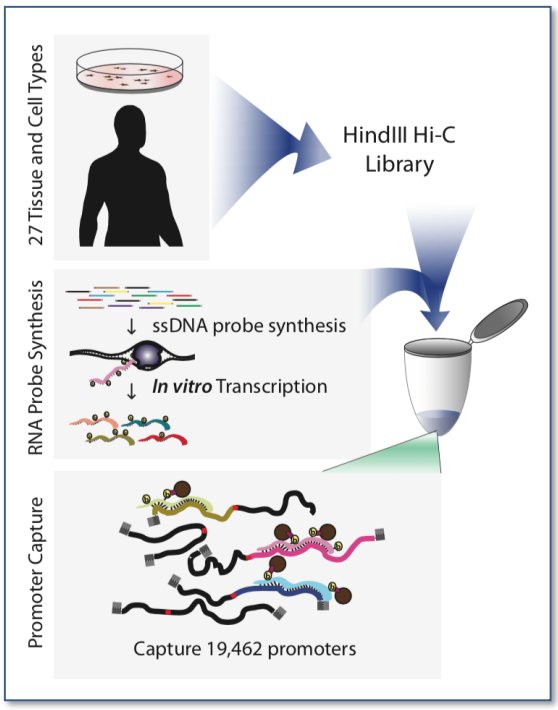


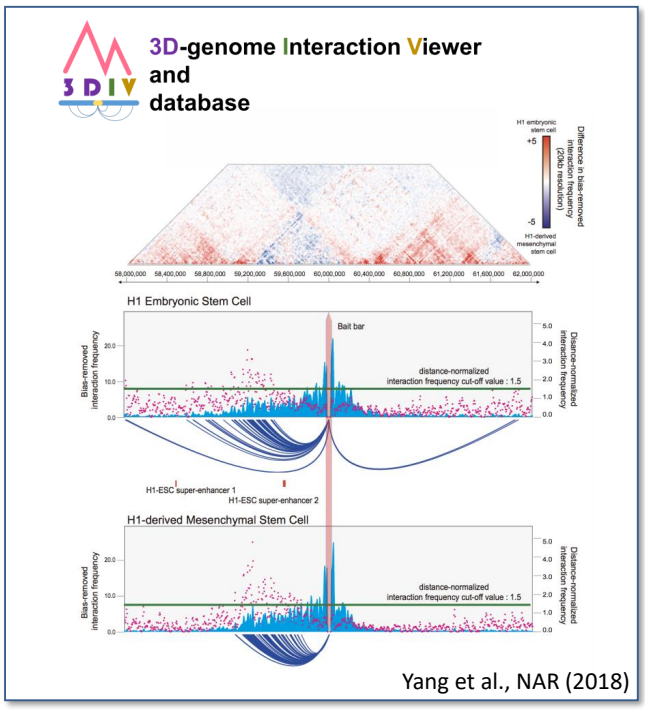






Promoter-capture Hi-C: Enhancer-promoter interaction maps



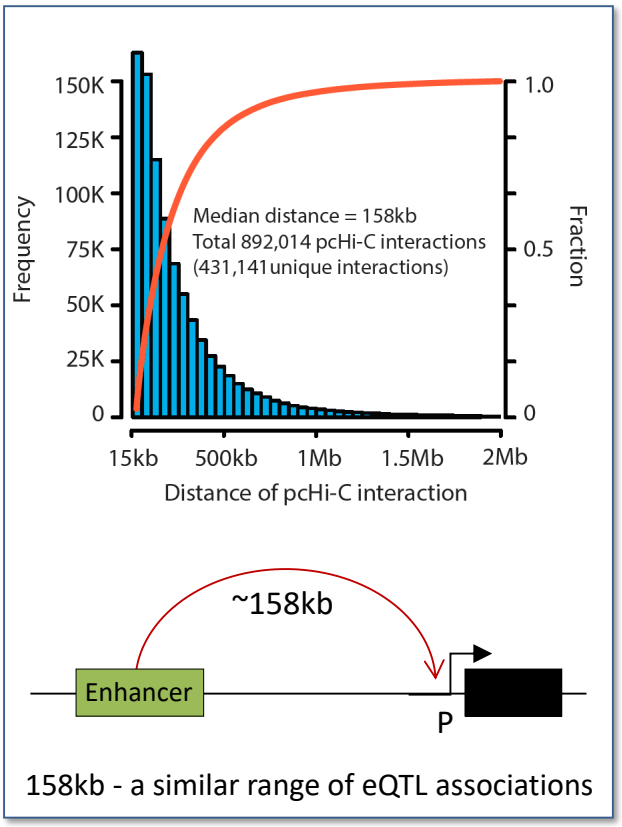


Yang et al., NAR (2018)

Jung et al., Nature Genetics (2019)

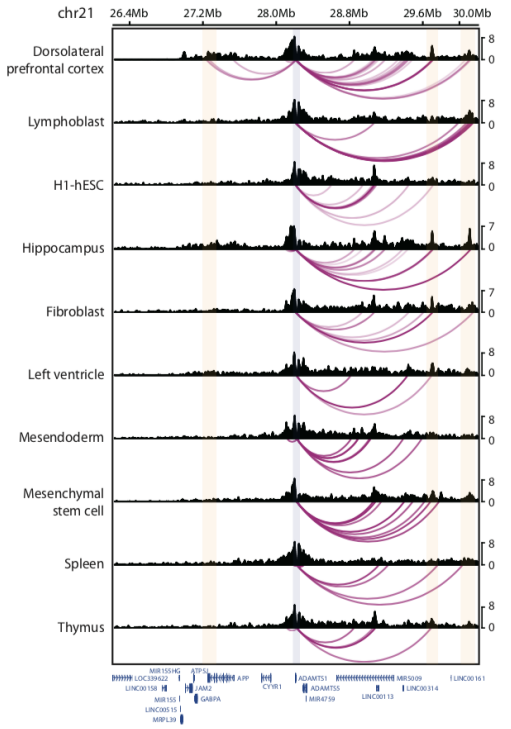
Basic principles of enhancer-promoter interactions

1. Enrichment of distal interactions



158kb - a similar range of eQTL associations

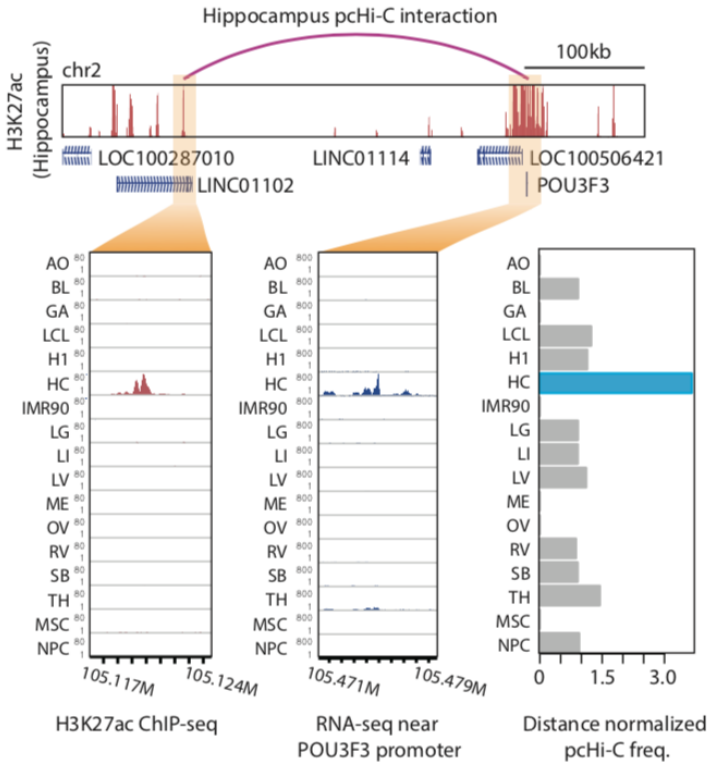
2. Interactions are tissue-specific

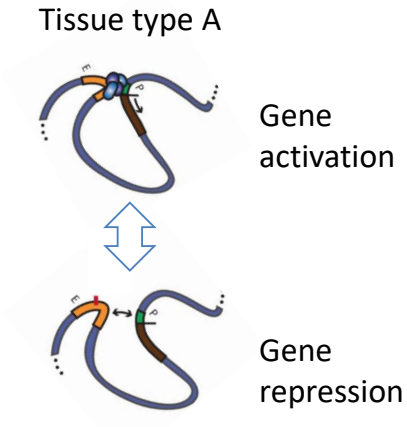


13~45% of pcHi-C interactions are unique

Basic principles of enhancer-promoter interactions

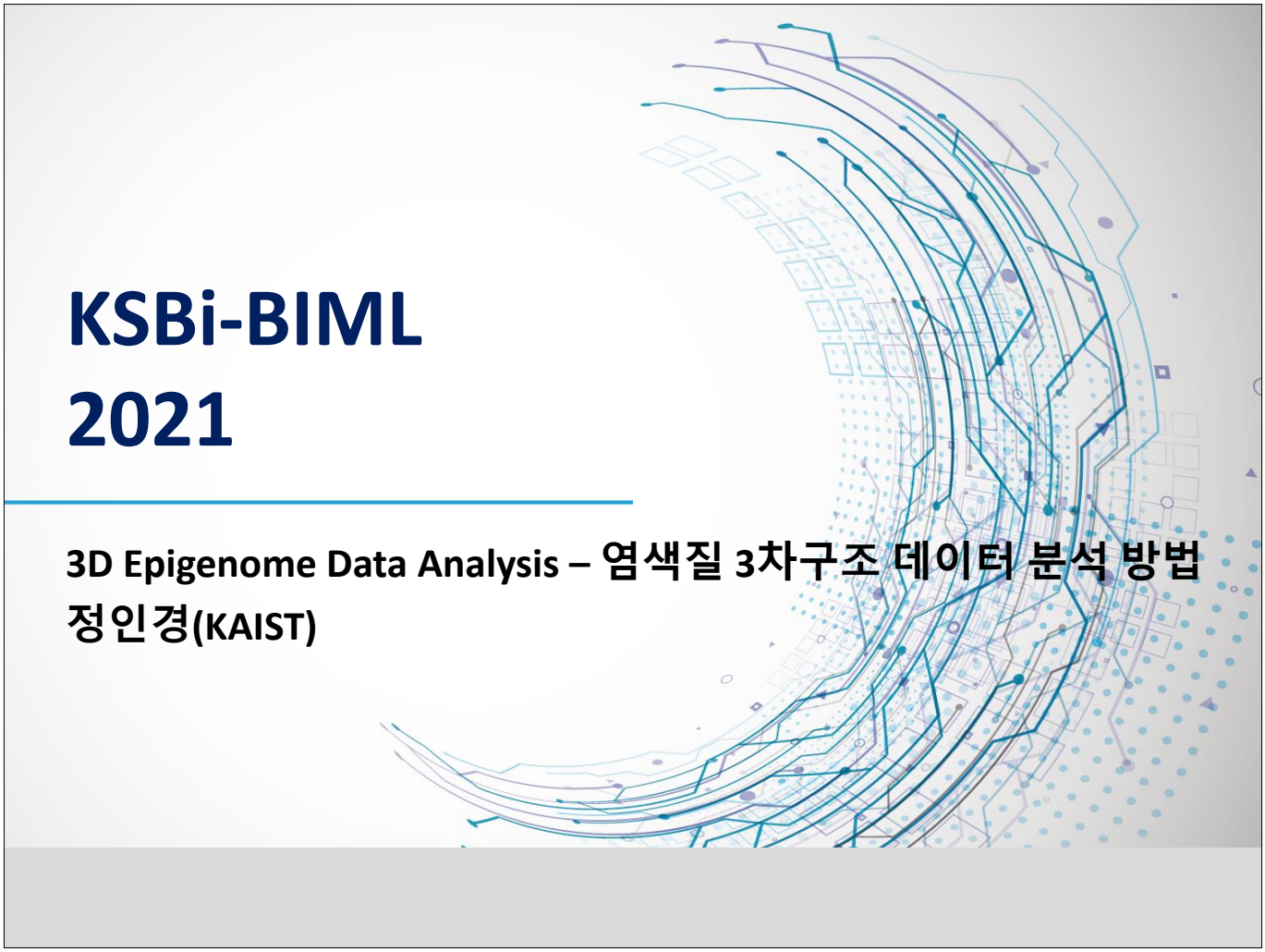
3. E-P interactions correlate with tissue-specific gene expression





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

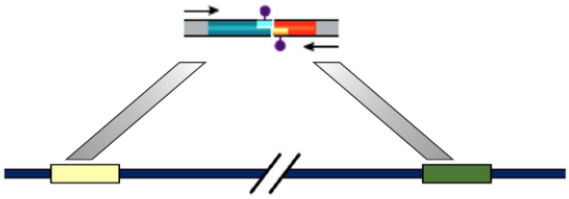


Contents

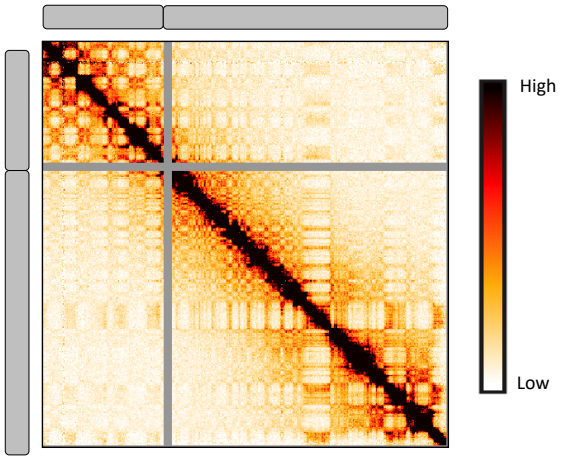
Normalization

Sequencing Hi-C Library and then What is the next step?

- Put the ends back together
  - Map to a reference genome
  - Determine ligation frequency

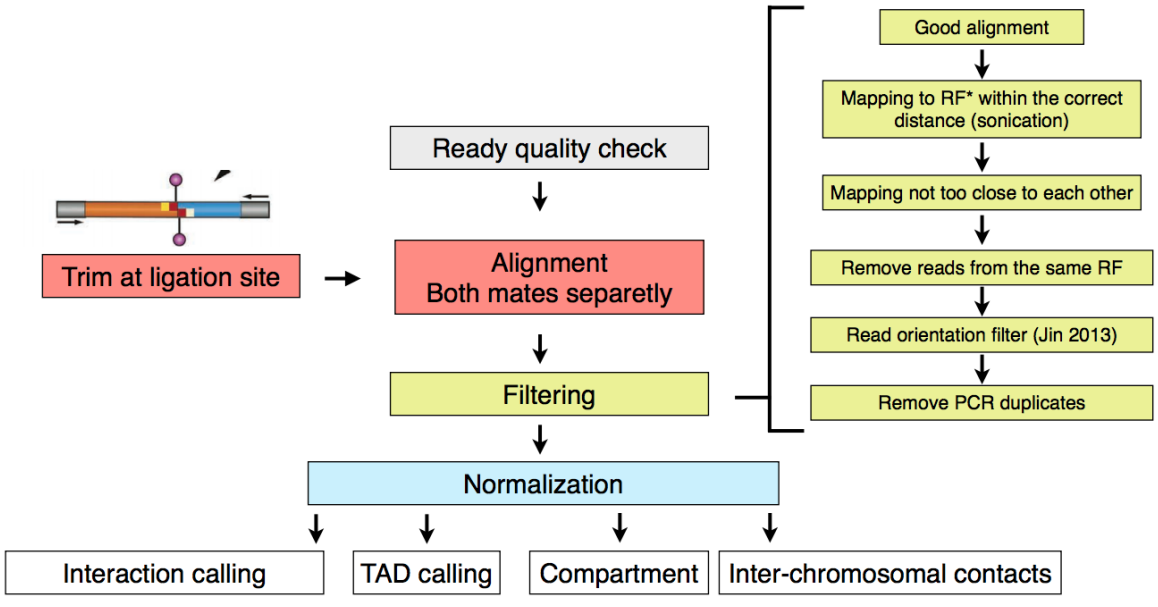


- Binning → Matrix
  - Sizes (resolution)

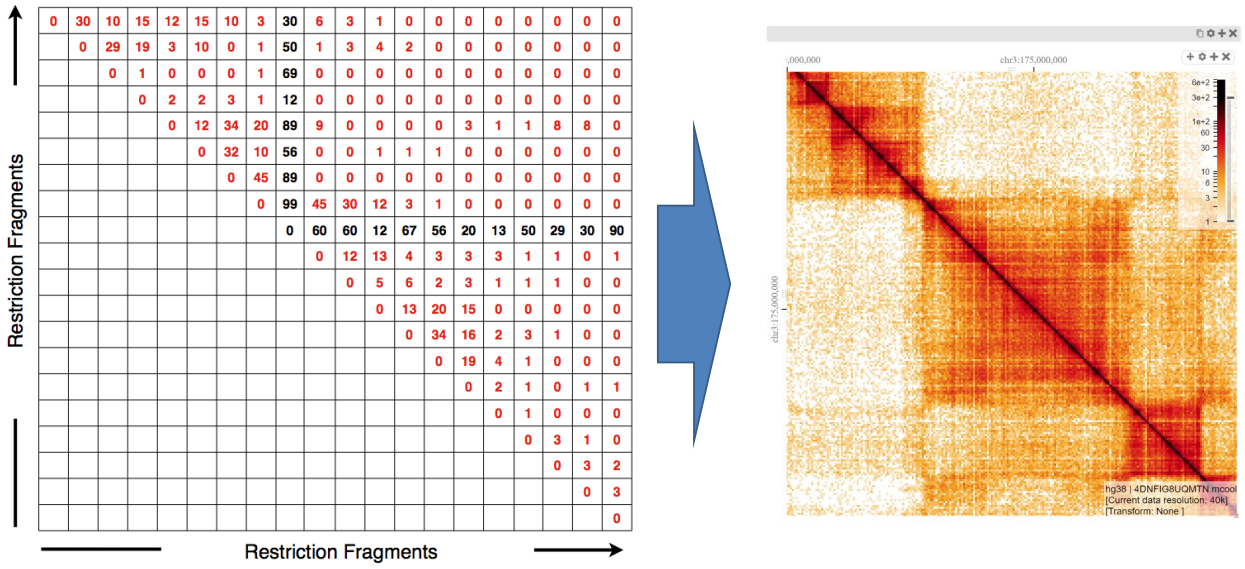


- Unless single cell, ensemble average (over 10^6 cells)
- Unless phased, averages over homologs.
- Unless synchronized, averages over the cell cycle.

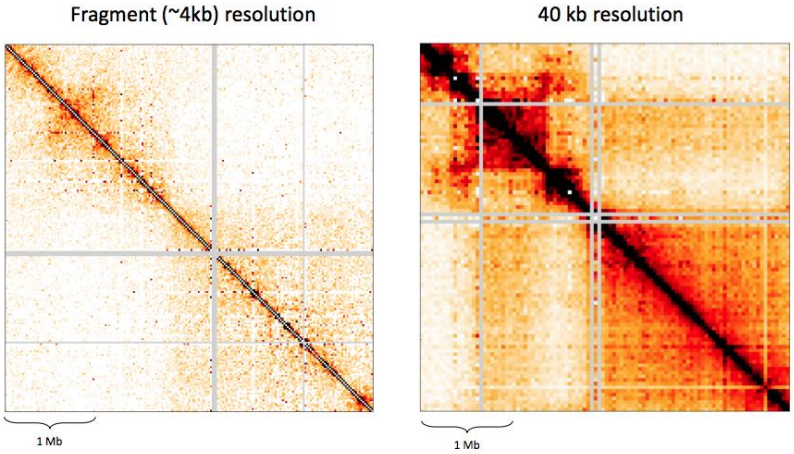
Analysis Workflow of Hi-C data



Ligation Frequency Matrix (Hi-C contact Map)



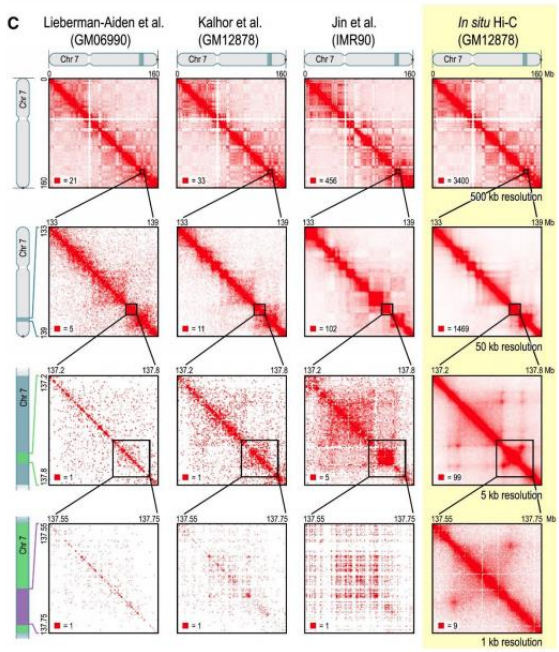
Resolution of Hi-C contact map



- For the same Hi-C result, but different visualization depending on the resolution (bin size)
  - At 40kb resolution, TAD looks clear
  - At 4kb resolution location interactions look clear
- Determining optimal resolution is critical to precisely interpret Hi-C result

No clear definition to determine Hi-C resolution so far since there is no clear resolution dependent properties. However one paper proposed that **bin size can be determined as at least 80 % of all possible bins having more than 1,000 contact**

Factors determining Hi-C Resolution



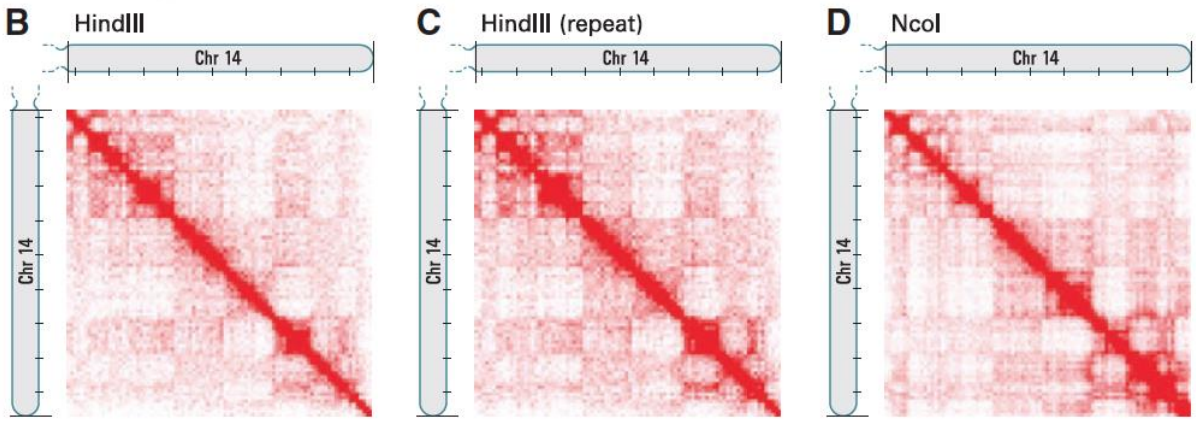
Some numbers in Hi-C resolution

- Human genome size: 3*10^9bp
- Restriction fragment length: ~4kb (6bp cutter)
- Number of fragments: 7.5*10^5
- Total interaction space: 5.6*10^11
- Number of cells per experiment: 10^6
- When we have 200M usable reads (2*10^8), 200 interactions were measured per cell
- Interaction space is under-sampled

How can we increase Hi-C resolution?

- Reduced interaction space (targeted approaches)
- Increased sequencing depth
- Reduced fragment size

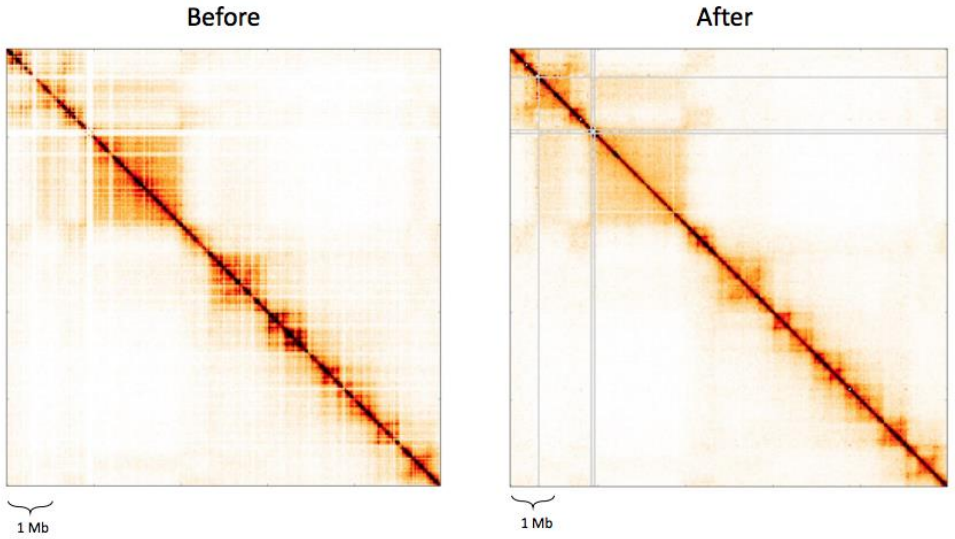
Restriction enzyme dependent Hi-C contact map



Red: Higher ligation frequency → Spatially proximal each other

White: Lower ligation frequency → Spatially distal each other

Bias Correction (Normalization)



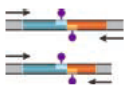
Uncorrected

Corrected (balanced)

Bias Correction (Normalization)

Digestion efficiency as a function of sequence composition (and DNA compaction)

Ligation efficiency as a function of fragment length



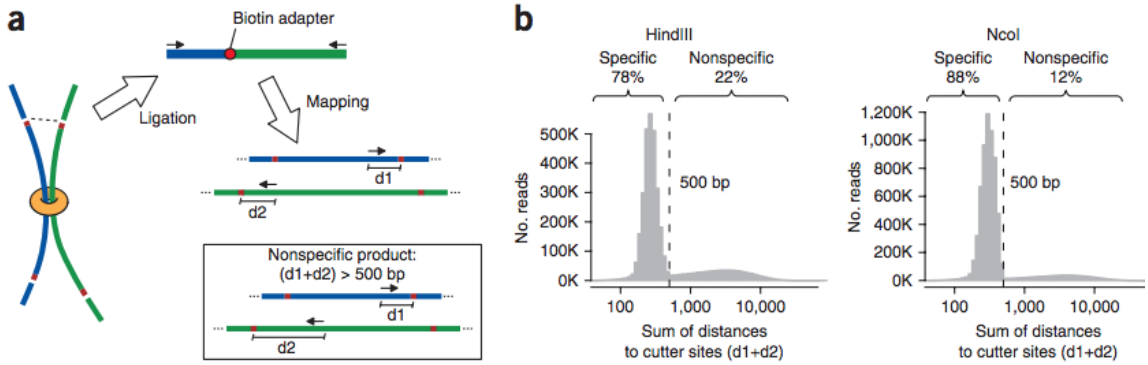
Sequencing efficiency as a function of sequence composition

1. Explicit factor methods (ex: HiCNorm)
   - Model bias due to GC content, fragment lengths, etc.
2. Coverage based methods (ex: ICE)
   - Don't model explicit sources of bias. Only assumes factorizable biases

Bias Correction - Explicit Factor Methods

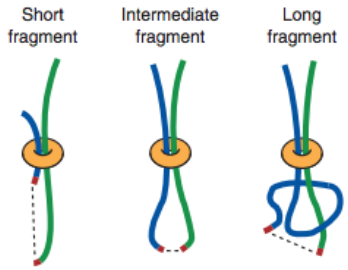
- Yaffe and Tanay or HiCNorm used explicit bias model based on 3 features that cause biases in Hi-C result
  - GC content, mappability, and effective length (or fragment length).
  - External information dependent
  - Effect of GC content on Hi-C library is enzyme-dependent.

Filtering for random ligation events

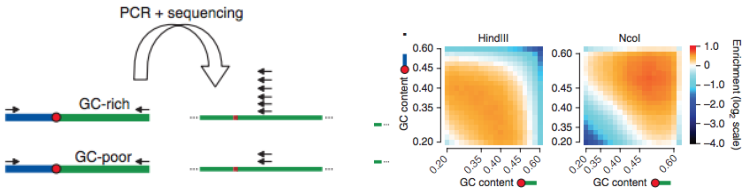


Bias Correction - Explicit Factor Methods

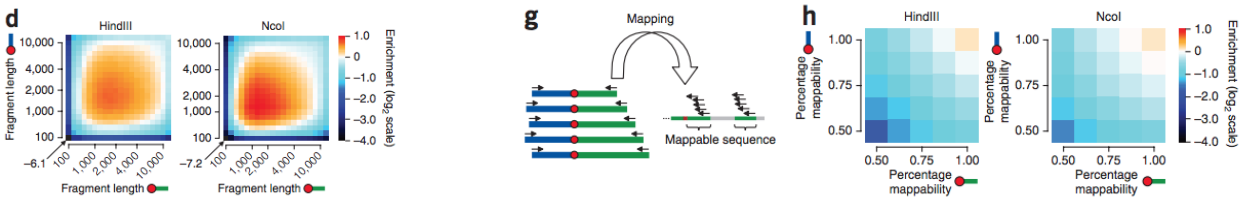
Fragment length



GC bias



Mappability bias



Bias Correction - Explicit Factor Methods

HiCNorm

Let $\boldsymbol{U^i} = \{u^i_{jk}\}_{1\le j,k\le n_i}$ represent the $n_i \times n_i$ Hi-C *cis* contact map for chromosome i, where n_i is the number of consecutive, disjoint 1 MB bins in chromosome i. Each entry u^i_{jk} represent the number of paired-end reads spanning two bins L^i_j and L^i_k. Let x^i_j, y^i_j and z^i_j represent the effective length feature, the GC content feature and the mappability feature at locus j for chromosome i, respectively. Similarly, let x^i_k, y^i_k and z^i_k represent the effective length feature, the GC content feature and the mappability feature at locus k for chromosome i, respectively. We assume that u^i_{jk} follows Poisson distribution with rate θ^i_{jk}:

$$\log(\theta^i_{jk}) = \beta^i_0 + \beta^i_{len}\log(x^i_j x^i_k) + \beta^i_{gcc}\log(y^i_j y^i_k) + \log(z^i_j z^i_k).$$

Here β^i_0 is the intercept term. β^i_{len} and β^i_{gcc} represent the effective length bias and the GC content bias, respectively. $\log(z^i_j z^i_k)$ is the Poisson offset term of the mappability bias. We fit this Poisson regression model, and let $\hat{\beta}^i_0$, $\hat{\beta}^i_{len}$ and $\hat{\beta}^i_{gcc}$ represent the corresponding parameter estimates. We further define the estimated Poisson rate $\hat{\theta}^i_{jk}$ as following:

$$\hat{\theta}^i_{jk} = exp\{\hat{\beta}^i_0 + \hat{\beta}^i_{len}\log(x^i_j x^i_k) + \hat{\beta}^i_{gcc}\log(y^i_j y^i_k) + \log(z^i_j z^i_k)\}.$$

The residual $e^i_{jk} = u^i_{jk}/\hat{\theta}^i_{jk}$ is the normalized *cis* interaction between two bins L^i_j and L^i_k.

Bias Correction – Coverage Based Methods

- Do not try to identify sources of biases but learn their effect from data

normalized ligation frequency between segments *i* and *j*

$$f_{ij} = \frac{c_{ij}\left(\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} c_{kl}\right)}{\left(\sum_{k=1}^{K} c_{ik}\right)\left(\sum_{k=1}^{K} c_{kj}\right)}$$

number of reads between segments *i* and *j*

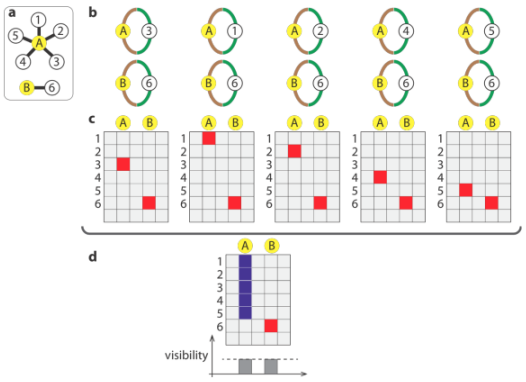
Total number of reads

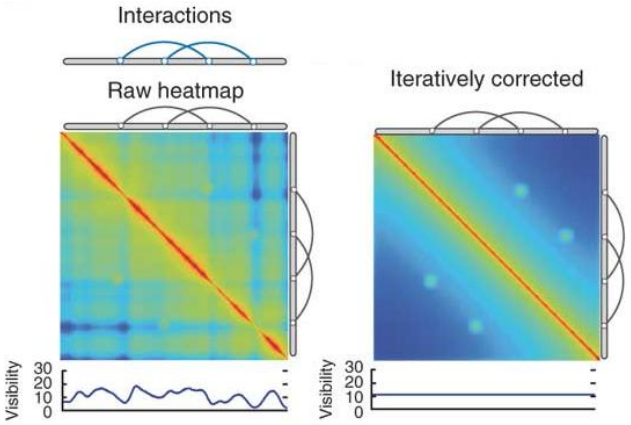
Total number of for segment *i*

Total number of for segment *j*

Bias Correction - Equalized Visibility

- ICE equalized visibility.
  - Each bin has equal coverage





Imakaev et al (2012)

Bias Correction - ICE

Biases

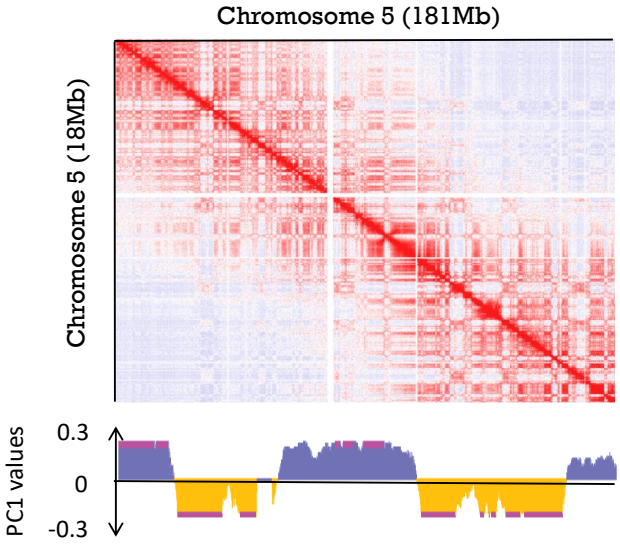
Observed $O_{ij} = B_i B_j T_{ij}$ True

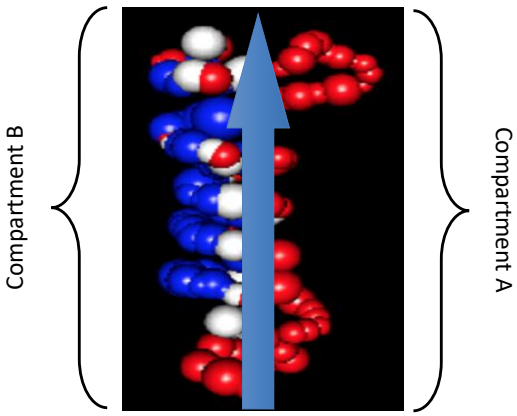
1. Start from Wij (=Oij) as the iterative process gradually changes this matrix to Tij
2. Calculate coverage of i as s_i=sum_j(Wij)
3. Additional biases delta Bi are calculated by renomralizitng s_i to have the unit mean as delta Bi = S_i / mean(S_i)
4. New Wij = Wij / (delta Bi * delta Bj)
5. Iterate step 2-4 until the variance of the additional biases becomes negligible

Issues:
- local signals can be removed
- interaction hubs such as transcriptional factory are not covered by ICE normalization

Compartment A/B

Two Major Compositions of Chromatin Structure: Compartment A/B

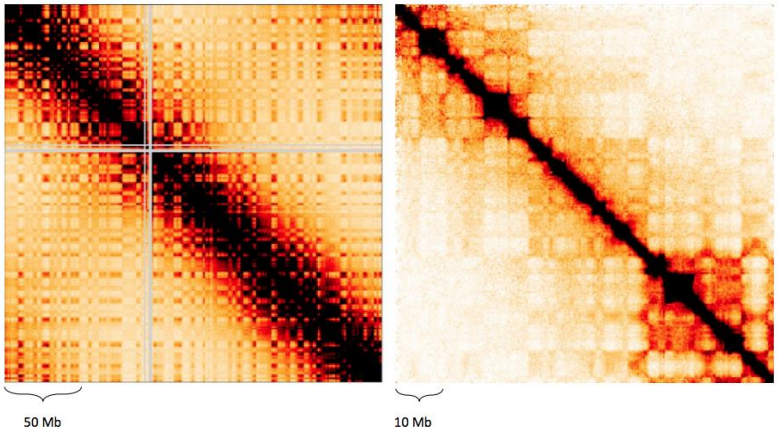


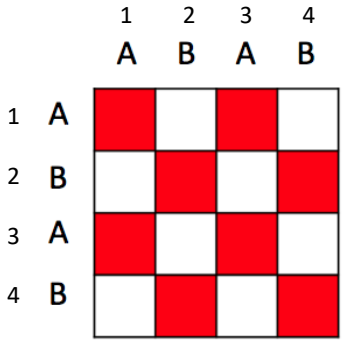


High gene density → Compartment A

Low gene density → Compartment B

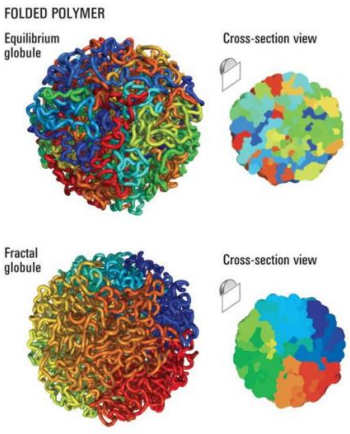
Compartment A/B patterns

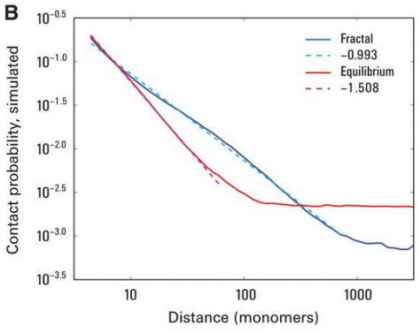


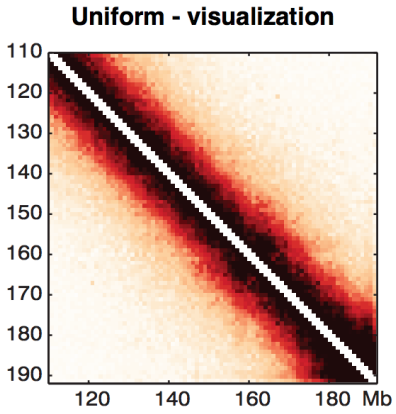


- Higher interaction between fragment 1 and 3 and between fragment 2 and 4
- Genome can be compartilized into two parts (compartment A and B)

Random collision (Distance Dependent Ligation Events)







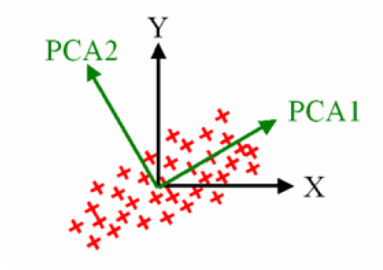
Random collision events

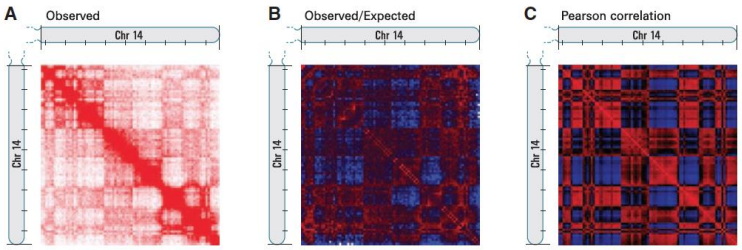
- Interaction occurring just because chromatin is a biopolymer and folds
- Random collisions can be estimated as expected interaction strength at a particular distance

PCA (Principle Component Analysis) to Hi-C contact matrix

Find component axes that maximize variance
(use the first eigenvector *v* of the PCA/SVD)

Apply PCA





v is the vector that minimizes $||svv^T-A||$, thus Aij will be near svivj

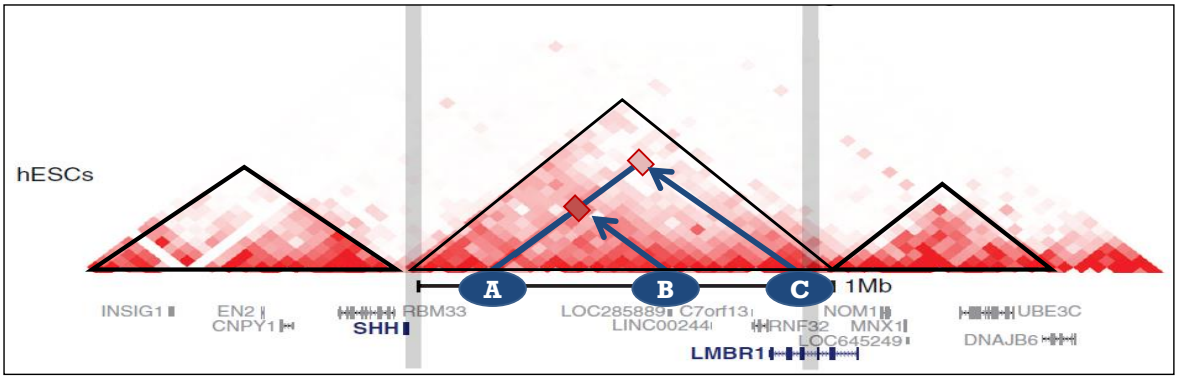
So, if vi and vj have the same sign their product will be positive (higher interaction bins i and j)

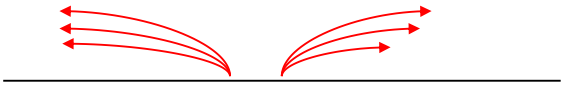
if vi and vj have the opposite sign their product will be negative (low interaction bins i and j)

Apply PCA to 3D chromatin structure to find a major structural component

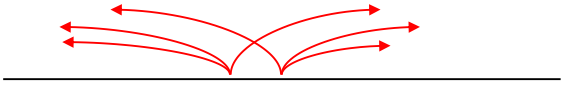
Topologically Associating Domains

Topologically Associating Domains (TADs)



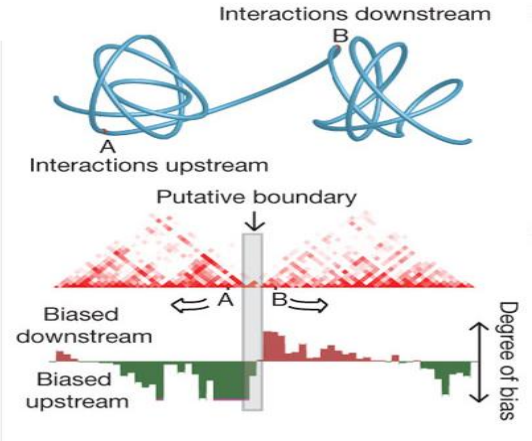


TAD boundary

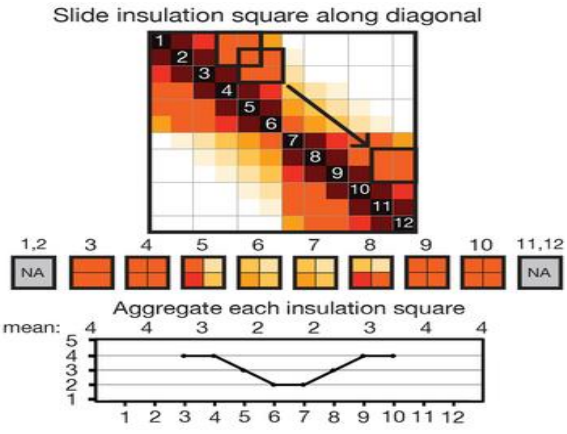


No boundary

Methods to Define TAD Boundaries



Directional Index Scores

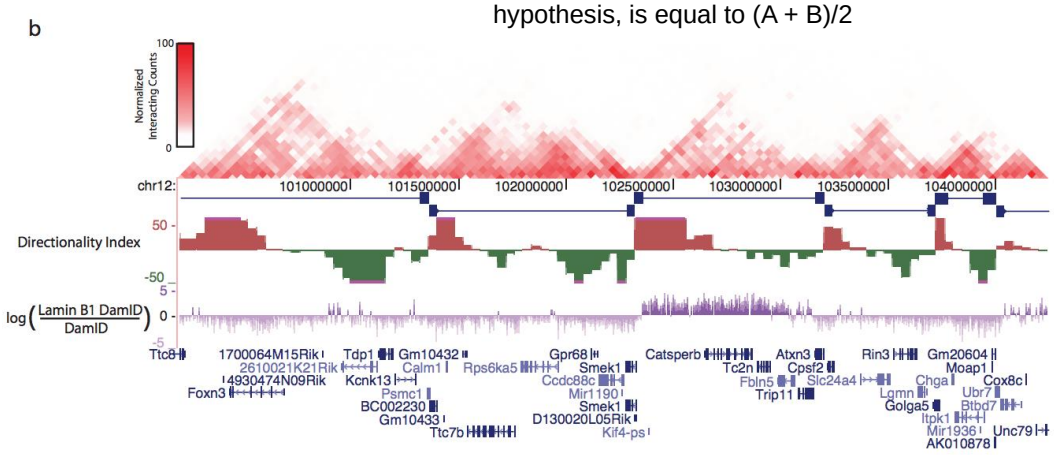


Insulation Scores

Methods to Define TAD Boundaries – Directional Index score

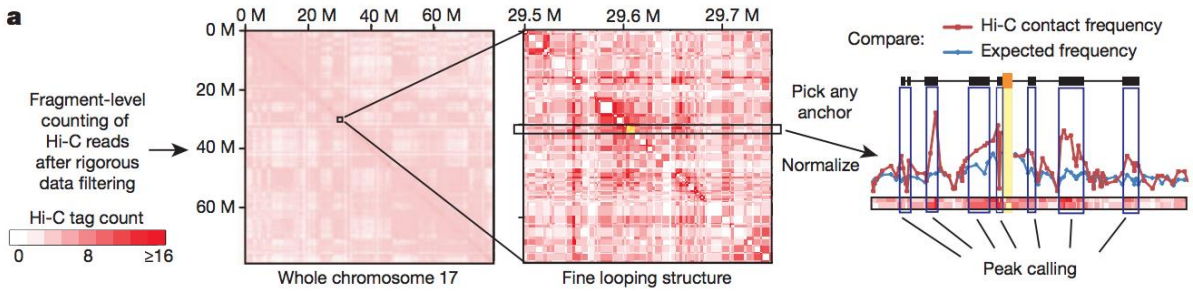
$$DI = \left(\frac{B - A}{|B - A|}\right)\left(\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}\right)$$

A: the number of reads that map from a given locus to the upstream 2Mb
B: the number of reads that map from the same locus to the downstream 2Mb
E: the expected number of reads under the null hypothesis, is equal to (A + B)/2



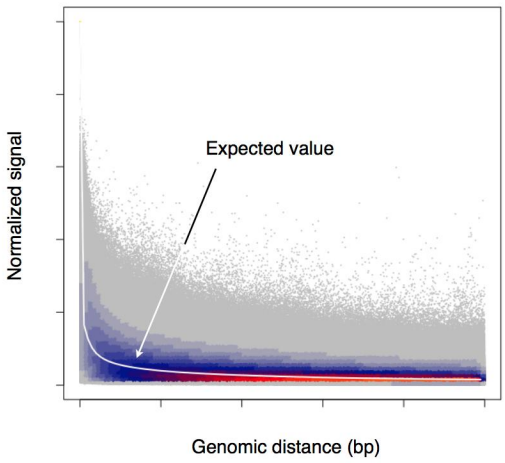
Long-range chromatin interactions

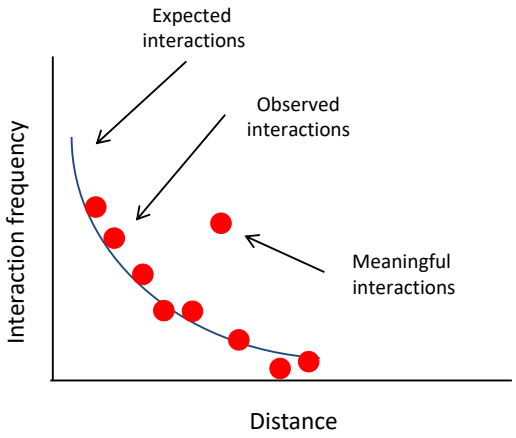
Isolation of Meaningful Interactions



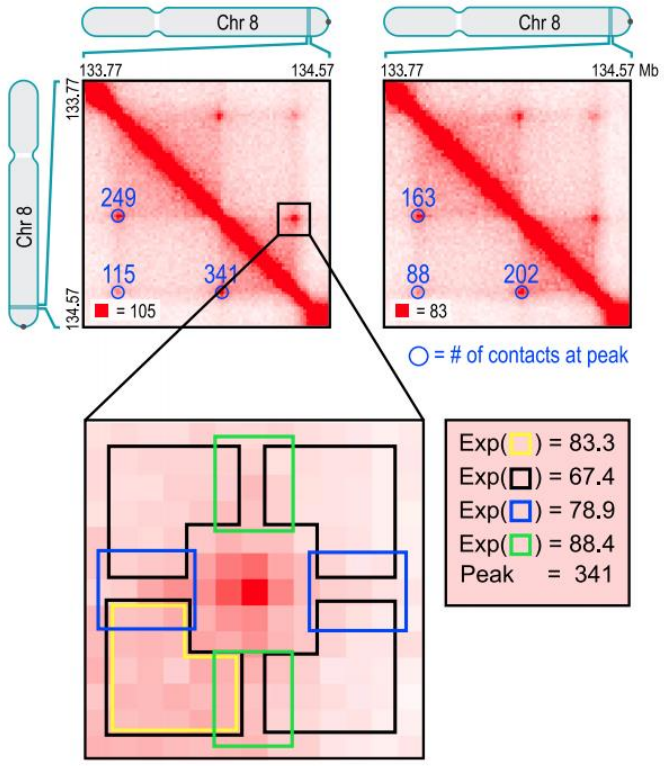
- Apply negative binomial distribution
- Test whether its strength is unexpectedly high given the biases, distances, and additional signal strength threshold

Random collision (Distance Dependent Ligation Events)





Isolation of Meaningful Interactions - HiCCUPS



- HiCCUPS
- Considering local background
- Pixels in the middle should have signal 50% higher than the surrounding

covNorm - Coverage based identification of long-range chromatin interactions

D. Removing biases based on coverage values of DNA fragments

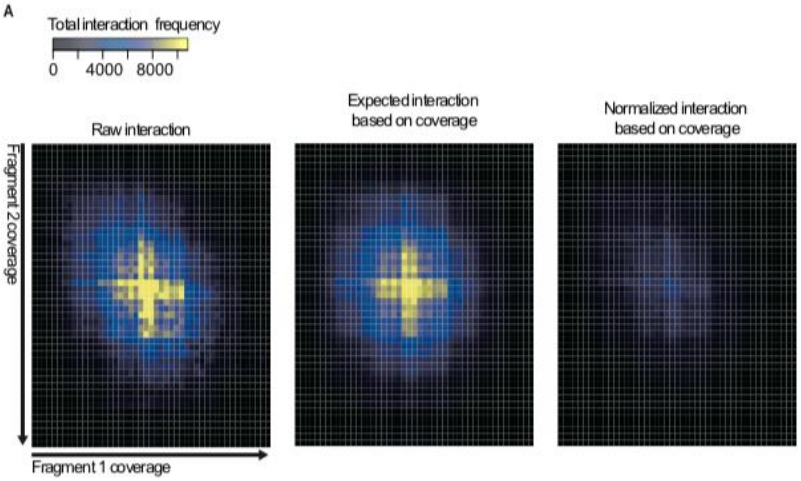
Let raw ligation frequency between two DNA fragments i and j as Y_{ij}, which is assumed to follow the negative binomial distribution ($Y_{ij} \sim NB$) with mean of μ and variance $\mu+\alpha\mu^2$ ($\alpha > 0$, over-dispersion parameter). The expected frequency of two DNA fragments based on the coverage C_i and C_j was obtained by fitting a generalized linear model:

$$\log(u_{ij}) = b_0 + b_1(C_i) + b_2(C_j) \quad (1)$$

The residual was defined as:

$$R_{ij} = Y_{ij} / \exp(\hat{b_0} + \hat{b_1}(C_i) + \hat{b_2}(C_j)) \quad (2)$$

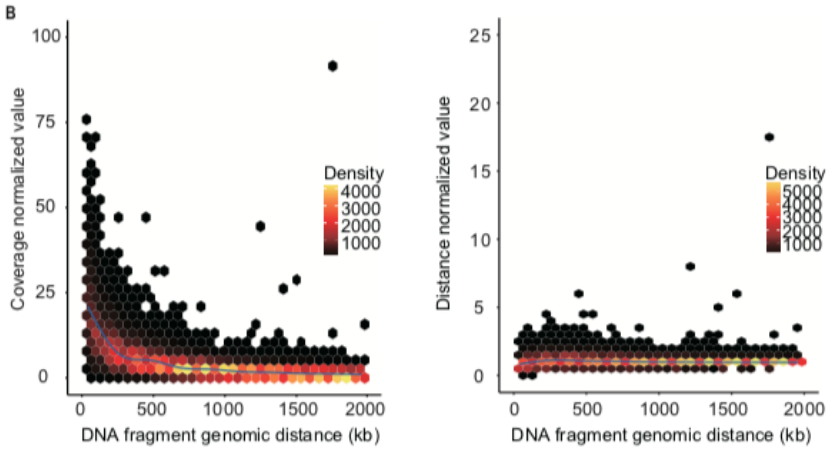
which indicates a bias-removed ligation frequency. was implemented by using R 'MASS::glm.nb' fur



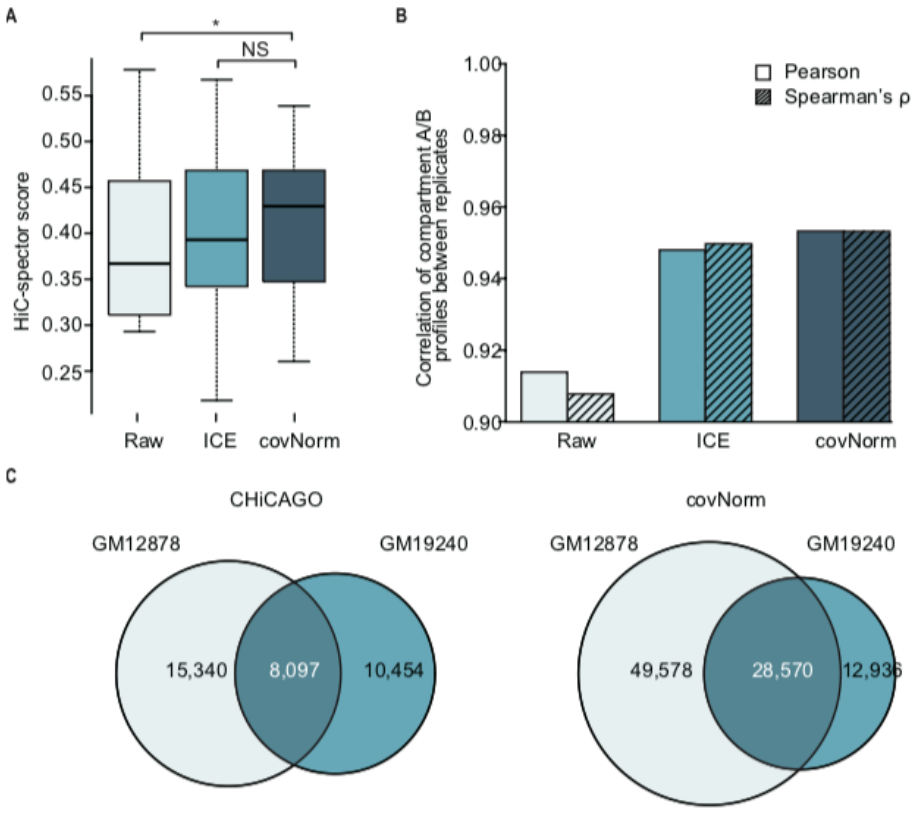
Distance normalization is critical to detect long-range chromatin interactions

E. Normalization against the genomic distance-dependent background signal

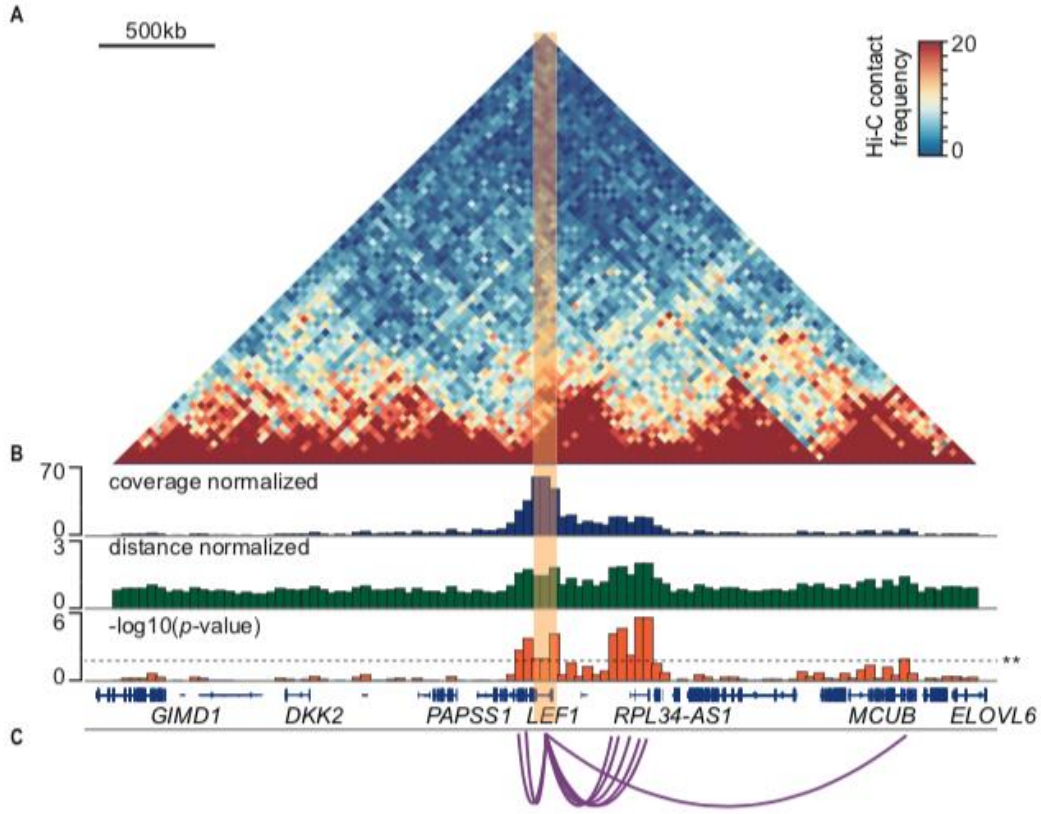
Hi-C contact probabilities decrease along with the genomic distance between DNA fragments due to the polymer nature of chromatin, thus another normalization step against genomic distance-dependent background signal is required to precisely identify biologically meaningful chromatin contacts such as enhancer-promoter interactions. To this end, given the residual R_{ij}, similar to the coverage normalization the model $\log(u_{ij}) = b_0 + b_1(D_{ij})$ was used to estimate E_d which is the expected ligation frequency at distance d, where the D_{ij} is the genomic distance between two DNA fragments. Distance-dependent background signal was removed by computing $(R_{ij} + avg(R_{ij})/(E_d + avg(R_{ij}))$.



Quality control of identified long-range chromatin interactions

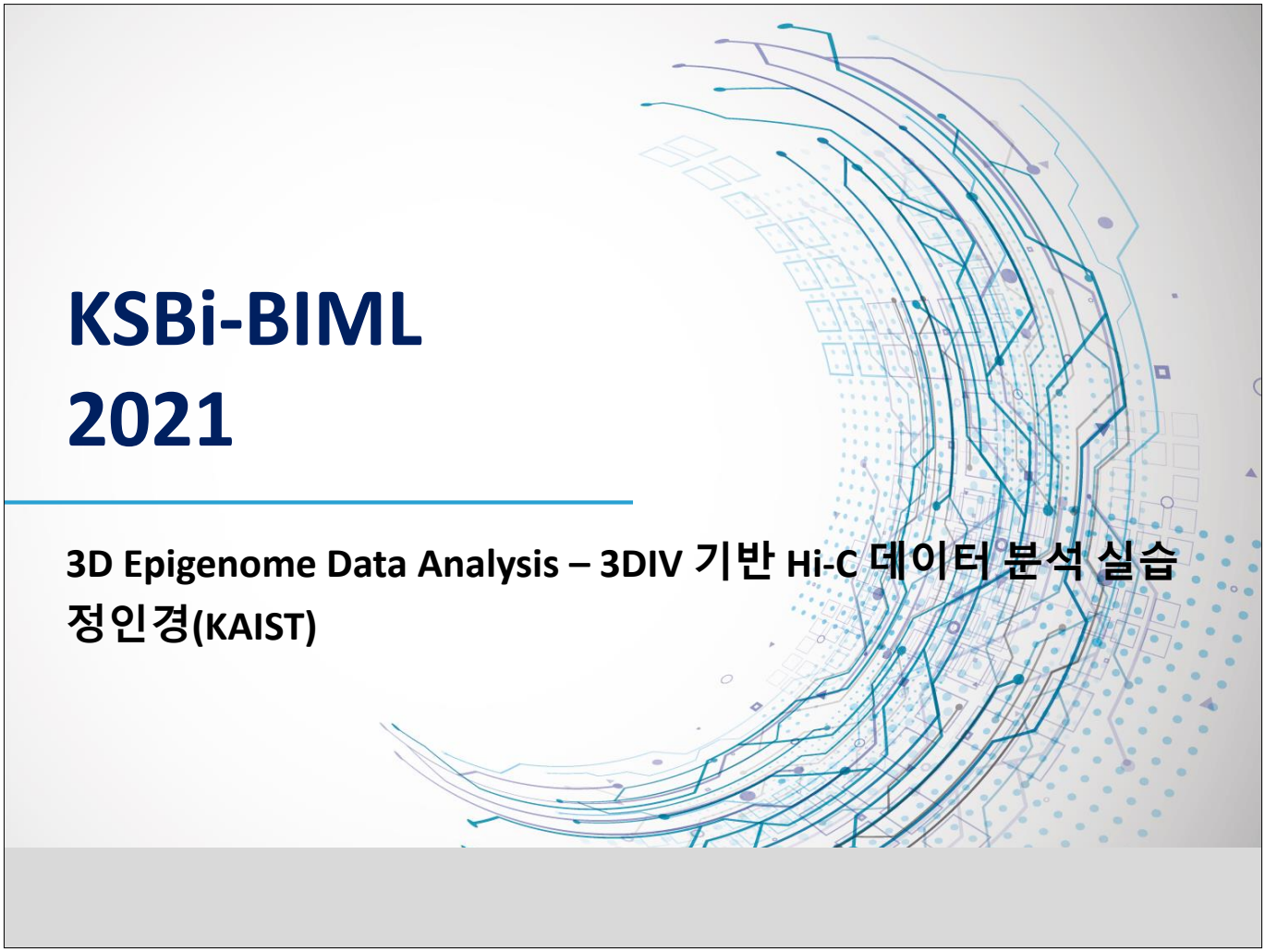


Visualization of coverage / distance normalized chromatin contacts



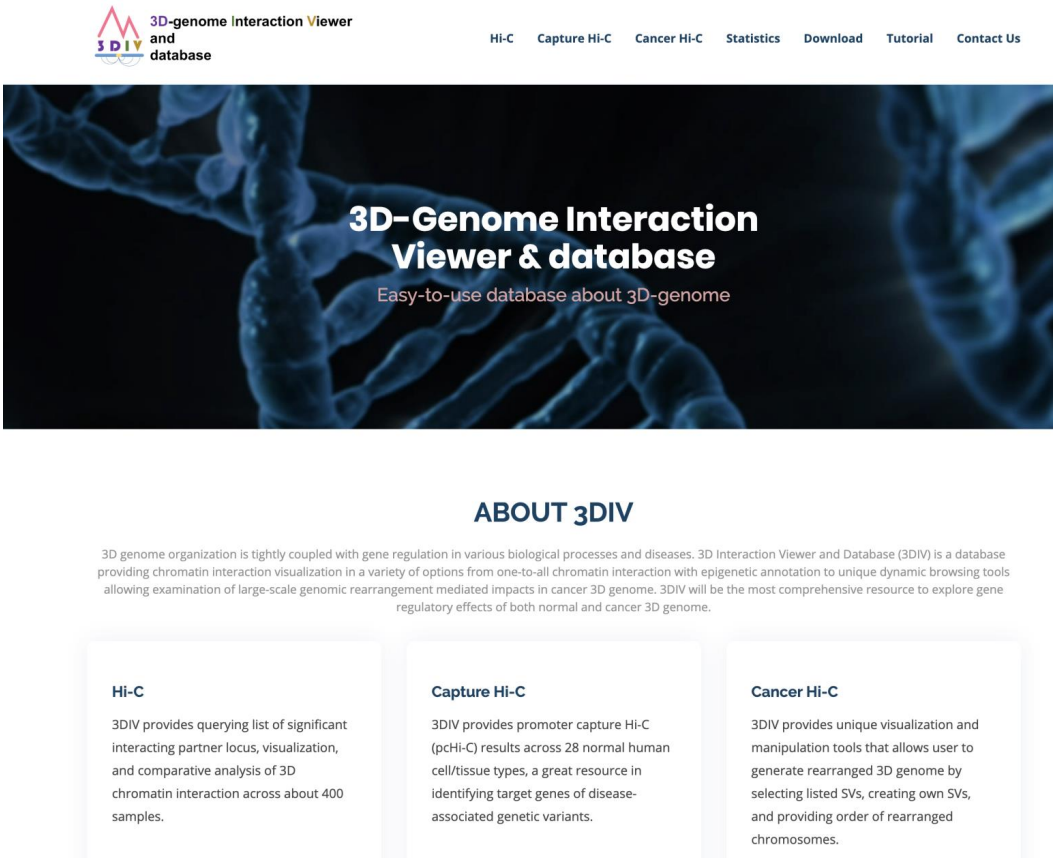
Summary

- Normalization step is required to remove experimental and innate biases in Hi-C data
- PCA analysis uncovers compartmentalized chromatin contact patterns
- Distance normalization is required to identify long-range chromatin contacts



Contents

3DIV: http://3div.kr



Hi-C data collection in 3DIV

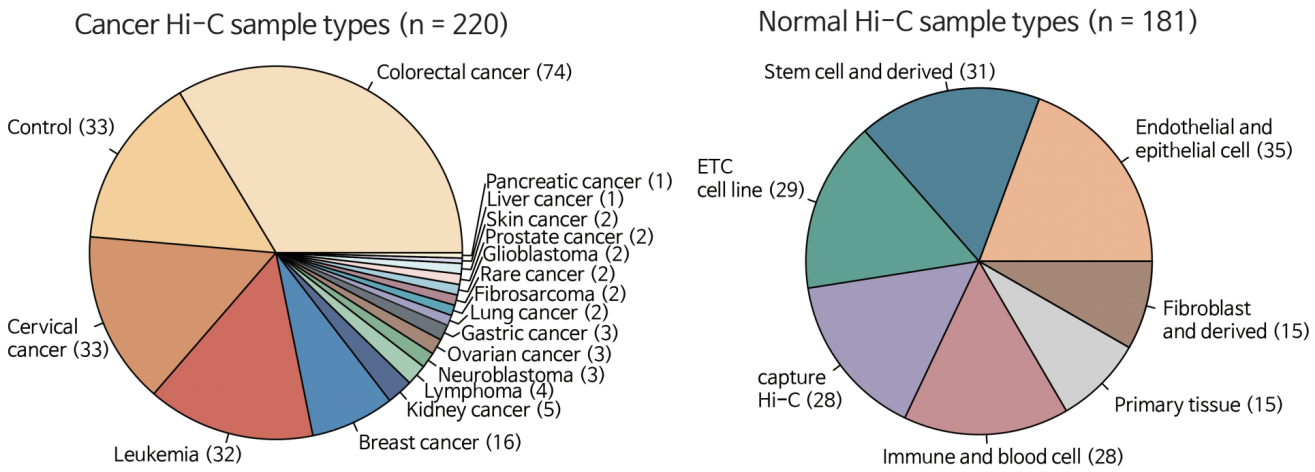


Table 1. Comparison of the updated 3DIV and other 3D genome databases as of October 2020

Software	Number of samples[a]	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[b]	✓					✓		
Nucleome Browser	138[c]	✓					✓		
WashU Epigenome Browser	36[c,d]	✓					✓		
HiView	2		✓	✓	✓	✓	✓		
HUGIn2	83	✓	✓	✓	✓	✓			
3D Genome Browser	113	✓	✓						
GITAR	20[c]	✓	✓						
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

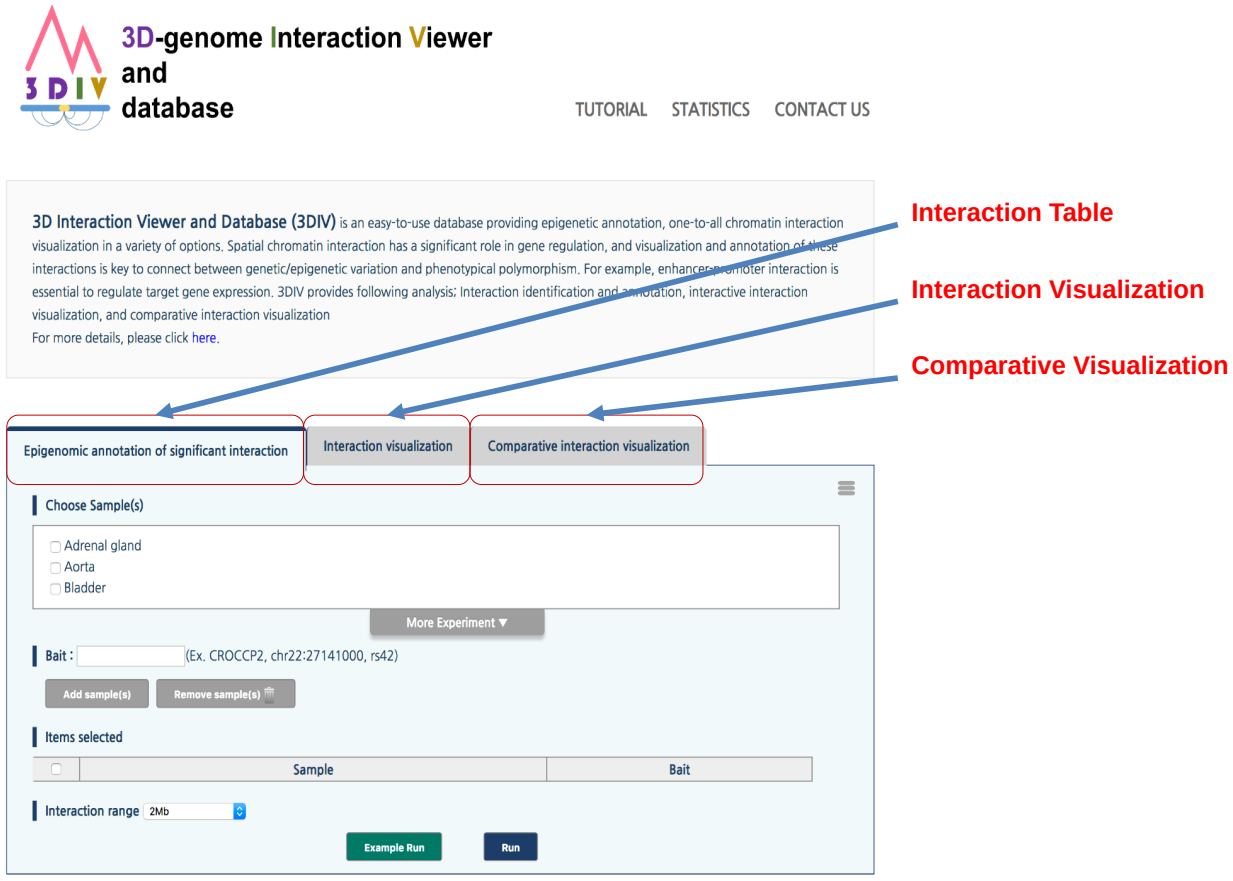
153 cell line/tissue Hi-C and 28 promoter capture Hi-C data

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



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

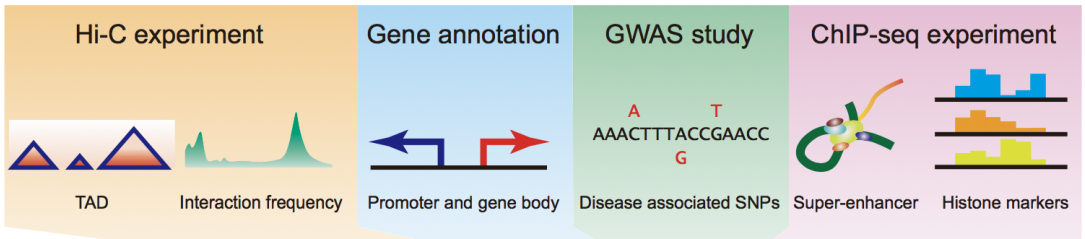
Interaction Visualization

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

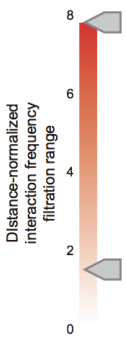
Comparative Visualization

- Comparative interaction frequency heatmap
- Synchronized interaction visualization

Module 1 : Interaction Table

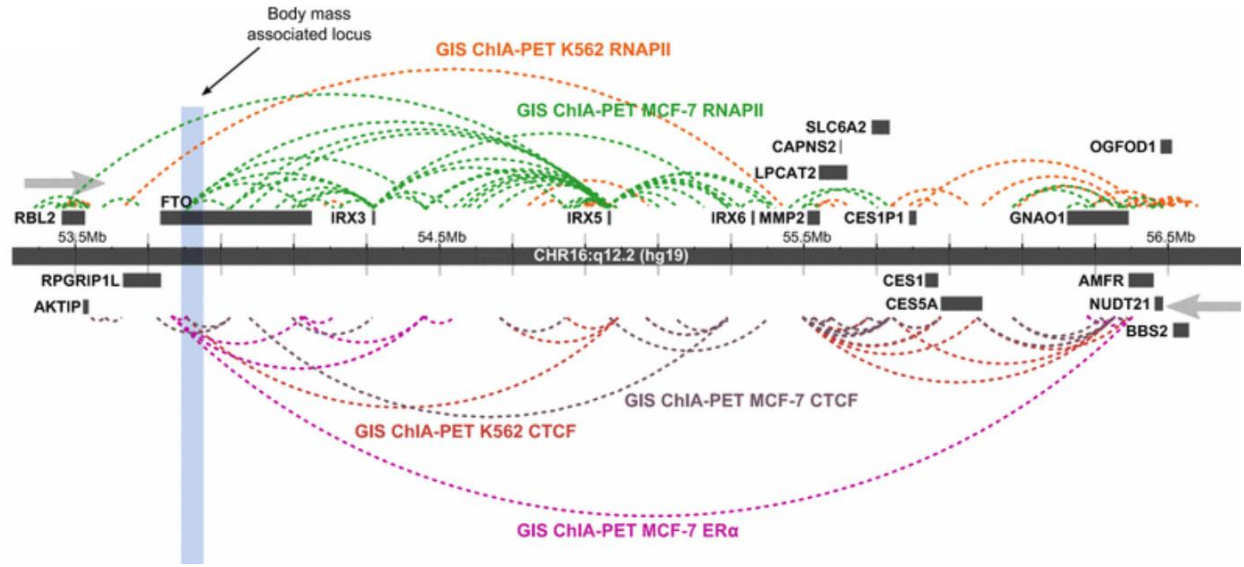


Sample	Locus (unit : kb)	Bias-removed interaction frequency	Distance-normalized interaction frequency	Gene Name	GWAS SNP ID	Enhancer or Super-enhancer	H3K27ac Fold change	H3K4me1 Fold change	H3K4me3 Fold change
Mesenchymal Stem Cell	chr16:54965-54970	3.16	7.07	IRX5			2.6	4.15	23.9
Mesenchymal Stem Cell	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	3.86
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
		6.22							1.33



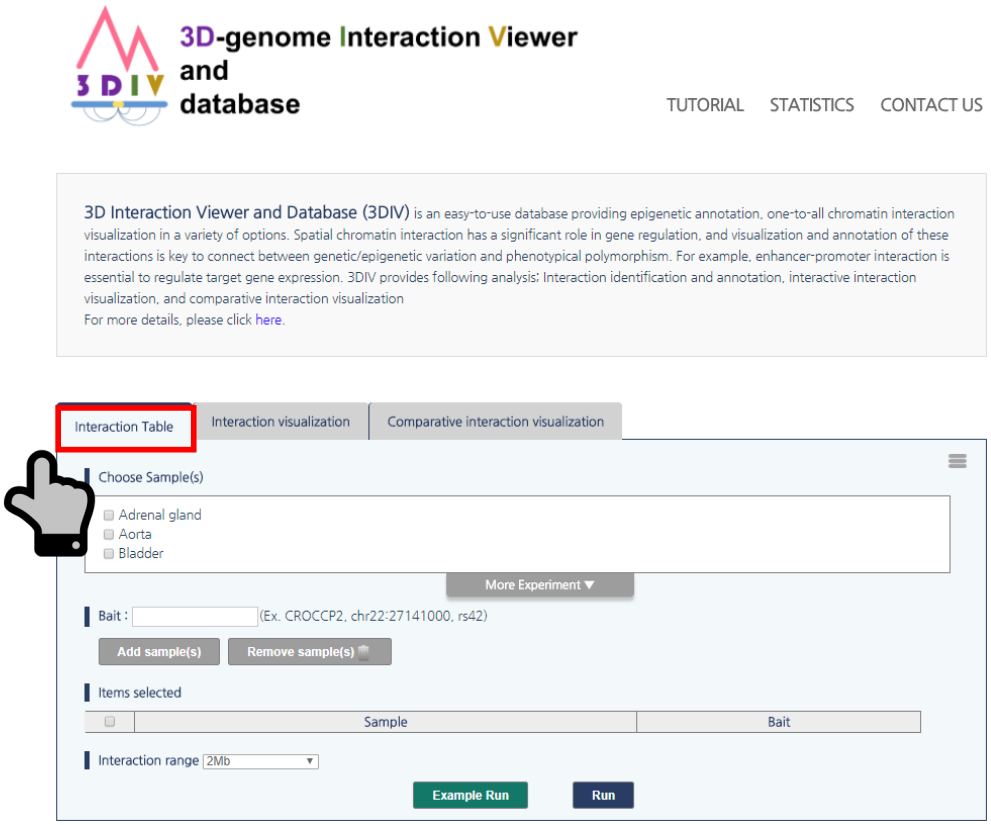
Example : Interaction profile of rs1421085

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

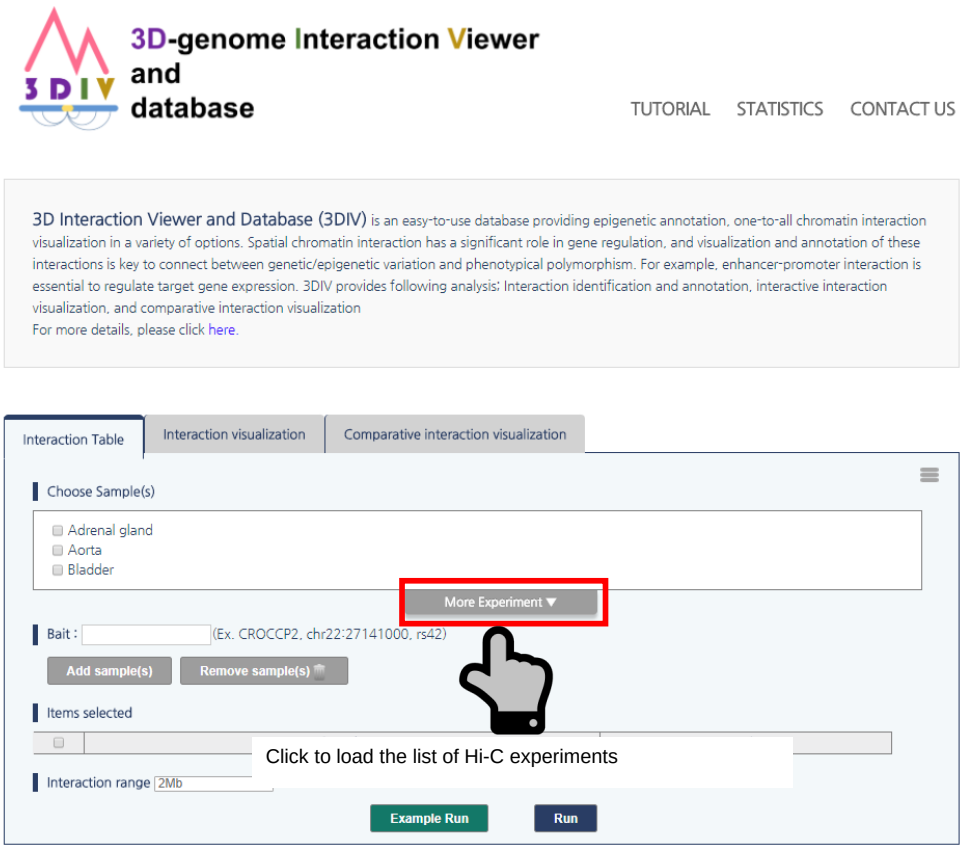


Rask-Andersen et al, Hum. Genet. (2015)

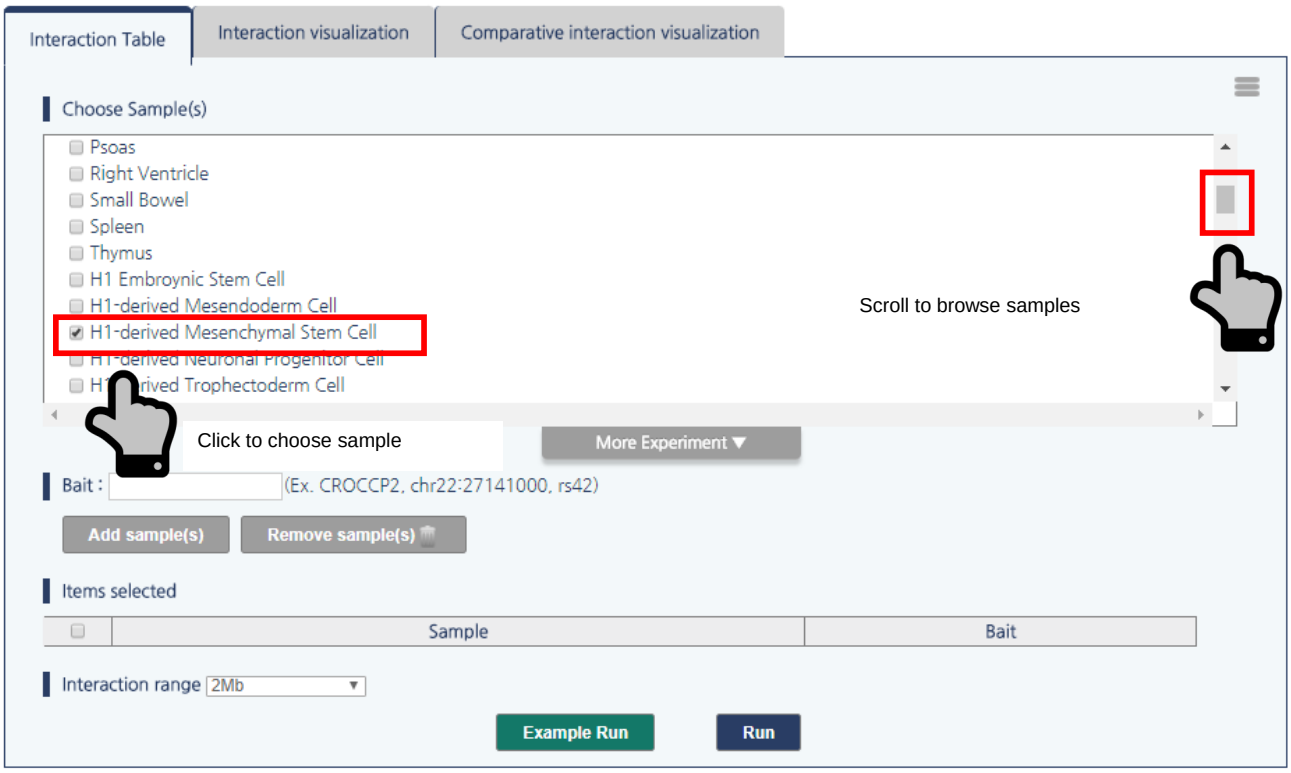
Step 1 : Open Interaction Table Module



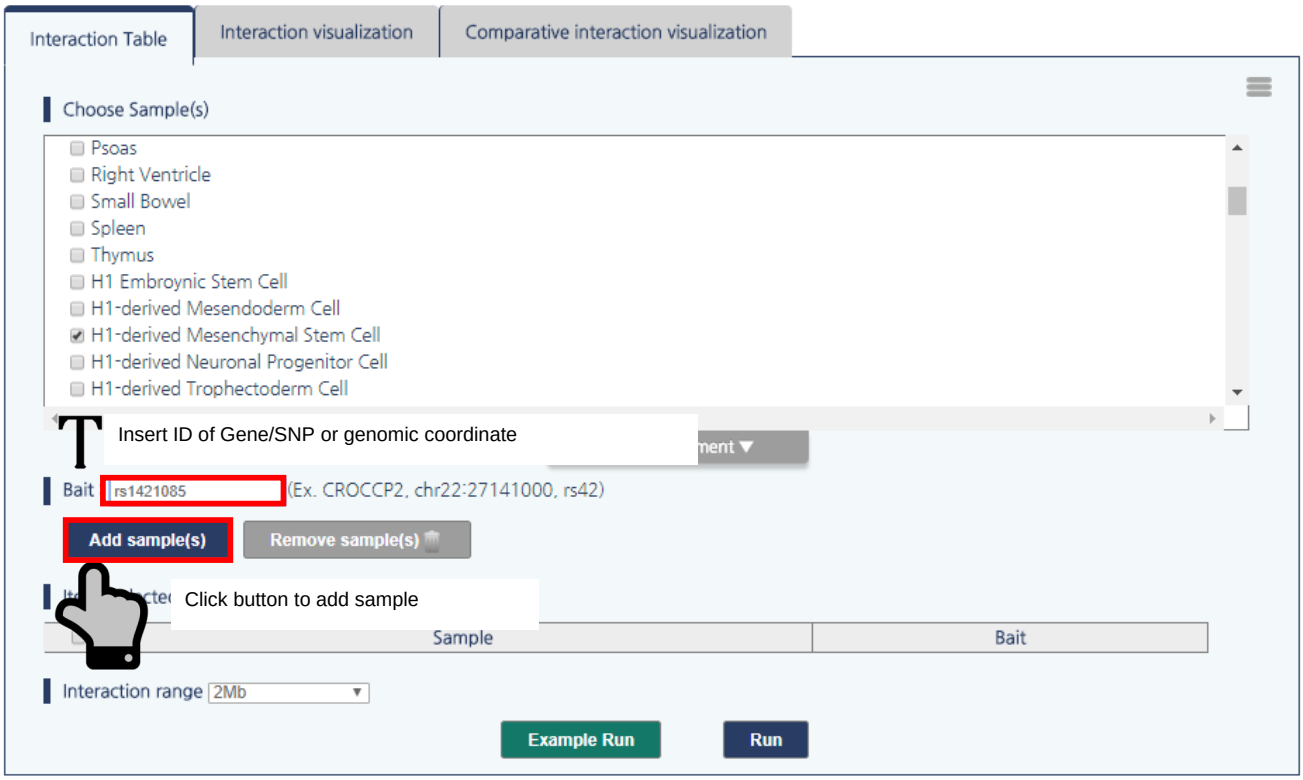
Step 2 : Choose a sample



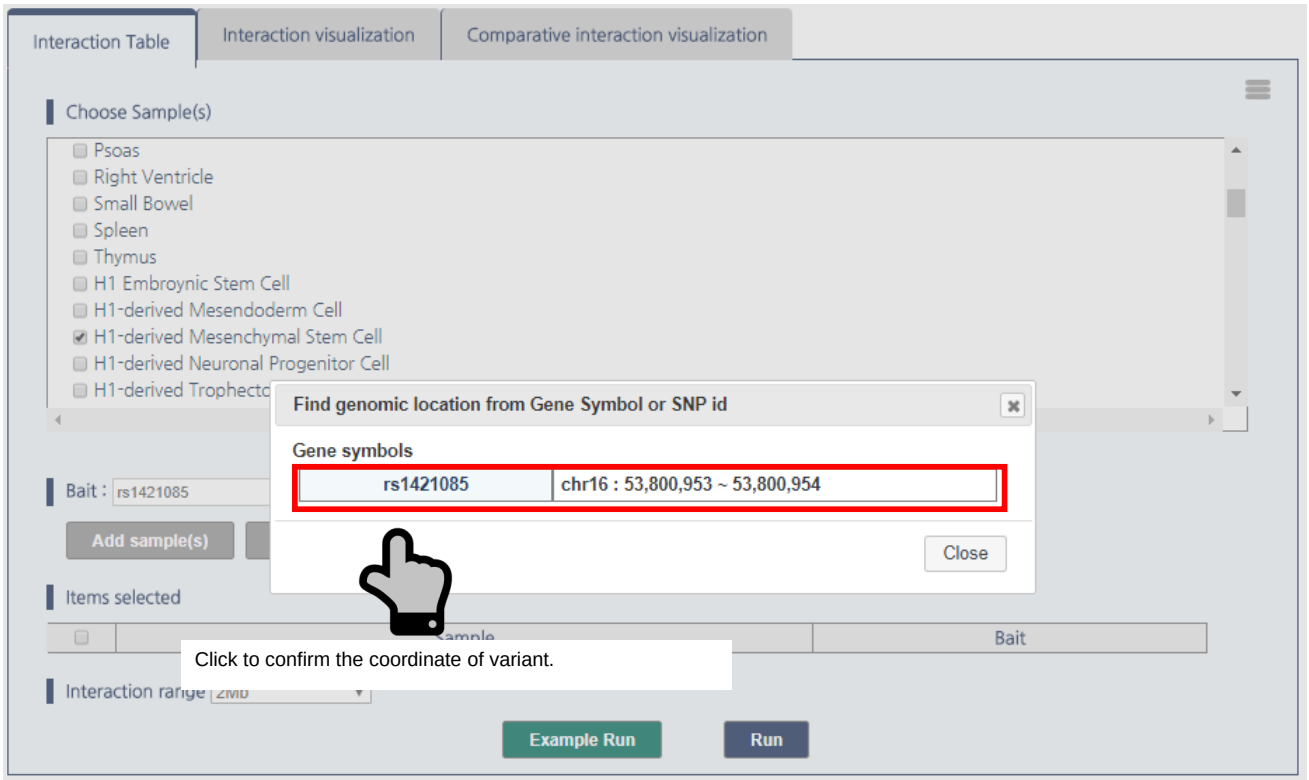
Step 2 : Choose a sample



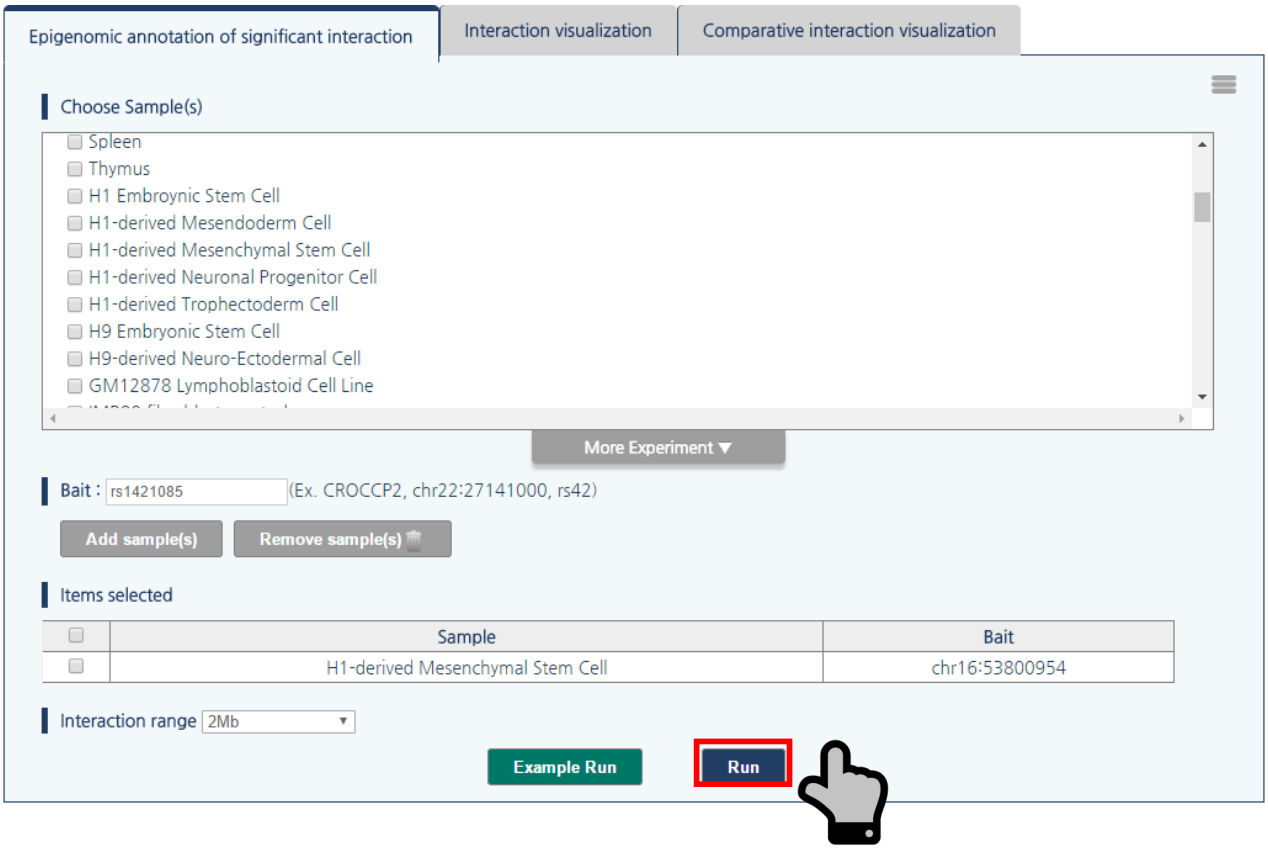
Step 3 : Choose a bait



Step 3 : Choose a bait

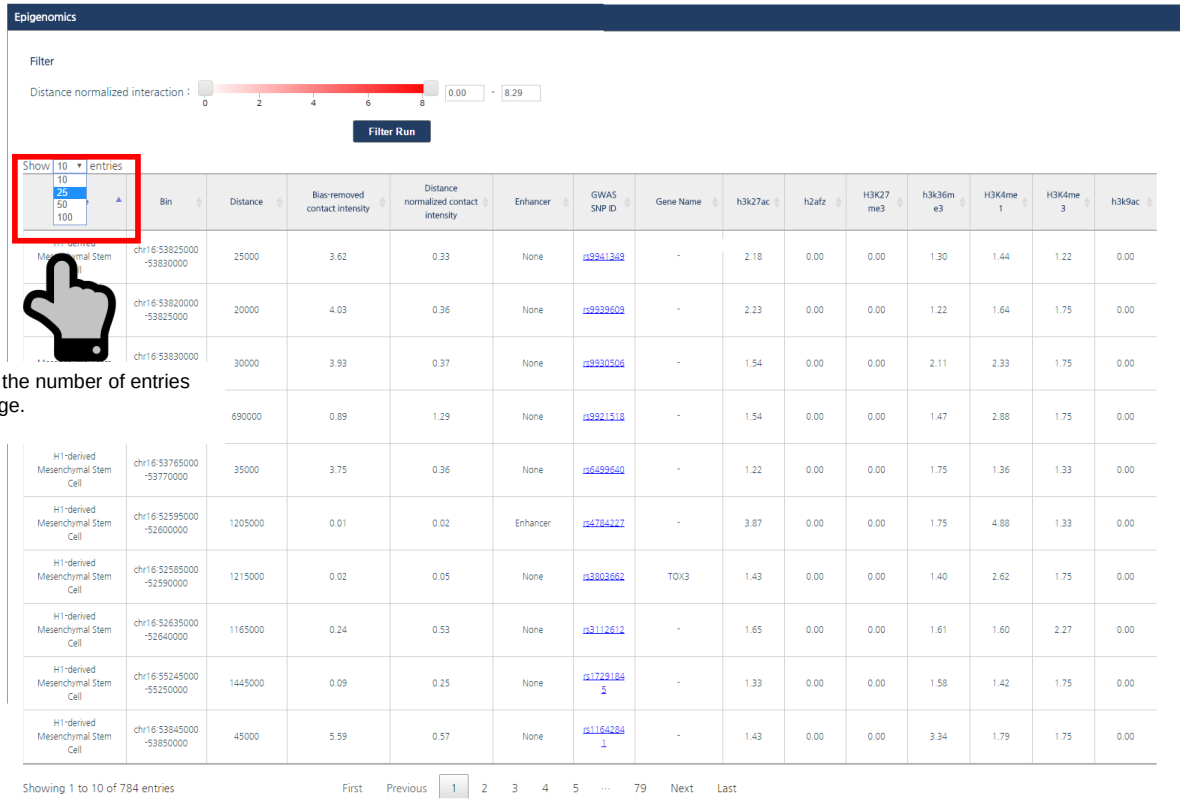


Step 4 : Run Module



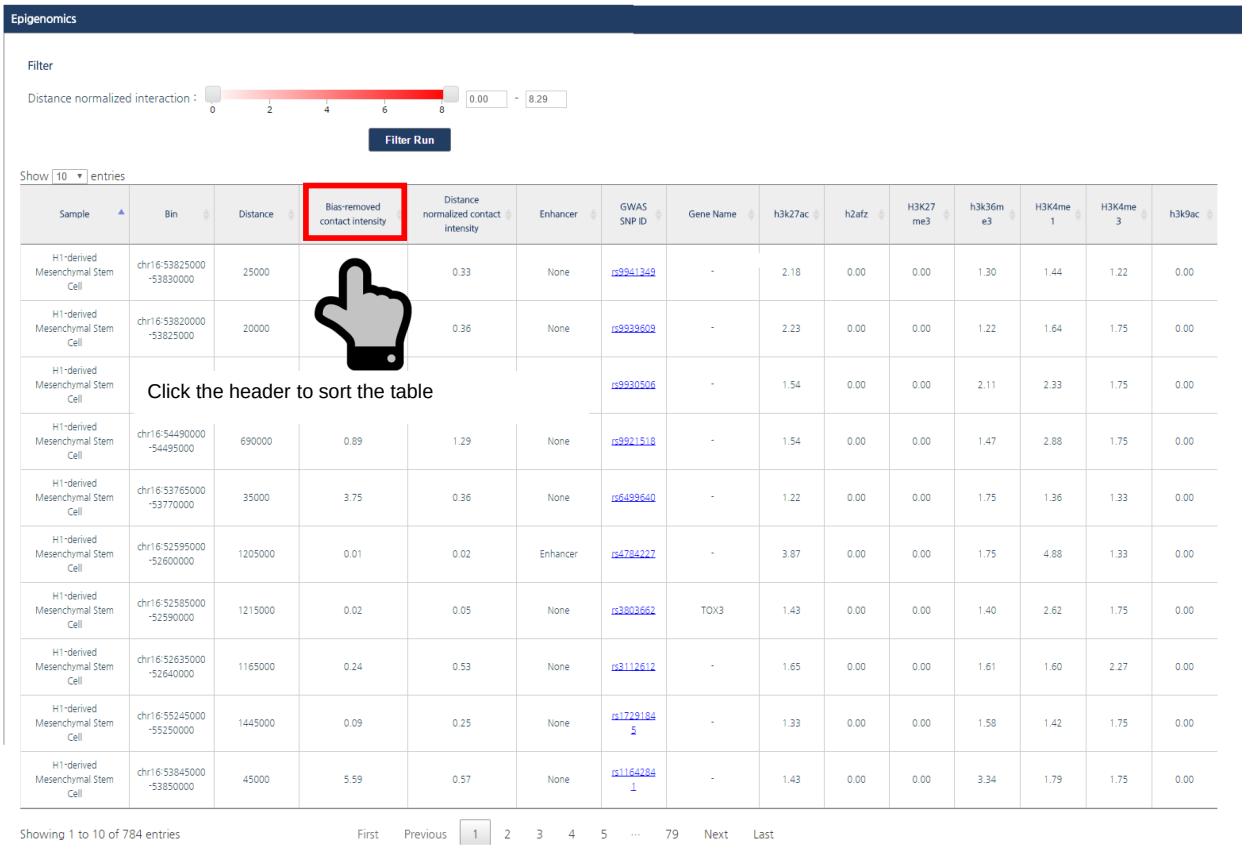
Step 5 : Browse the table

Step 5a : Adjust the table



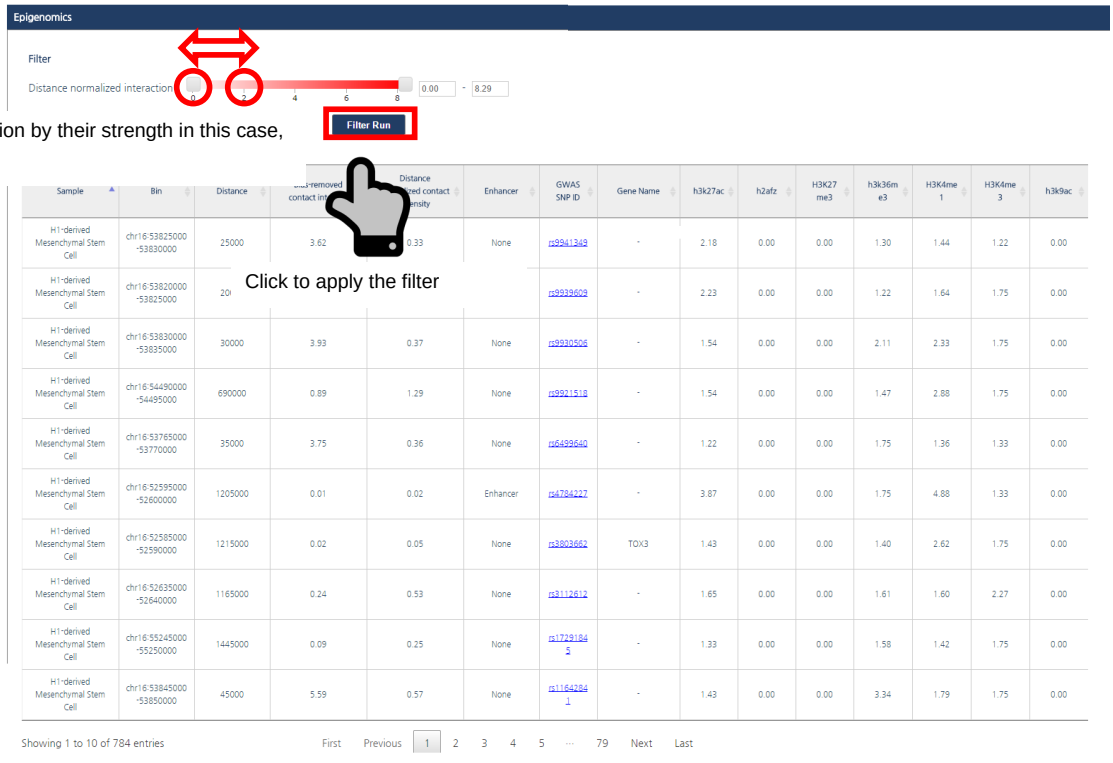
Adjust the number of entries per page.

Step 5b : Sort the interaction table



Click the header to sort the table

Step 5c : Filter interaction

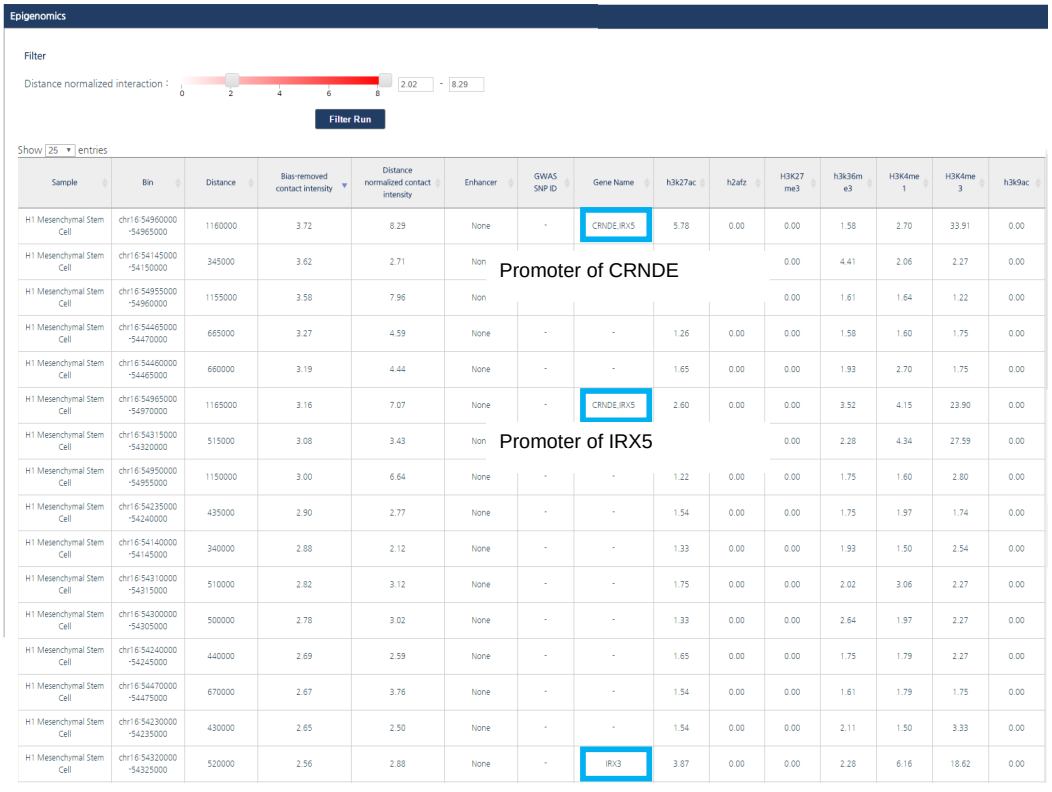


Drag to filter interaction by their strength in this case, 2.0 is the criteria.

Sample	Bin	Distance	Bias-removed contact intensity	Distance normalized contact intensity	Enhancer	GWAS SNP ID	Gene Name	h3k27ac	h2afz	H3K27me3	h3k36me3	H3K4me1	H3K4me3	h3k9ac
H1-derived Mesenchymal Stem Cell	chr16:53825000-53830000	25000	3.62	0.33	None	rs9941349	-	2.18	0.00	0.00	1.30	1.44	1.22	0.00
H1-derived Mesenchymal Stem Cell	chr16:53820000-53825000	20(				rs9939609	-	2.23	0.00	0.00	1.22	1.64	1.75	0.00
H1-derived Mesenchymal Stem Cell	chr16:53830000-53835000	30000	3.93	0.37	None	rs9930506	-	1.54	0.00	0.00	2.11	2.33	1.75	0.00
H1-derived Mesenchymal Stem Cell	chr16:54490000-54495000	690000	0.89	1.29	None	rs9921518	-	1.54	0.00	0.00	1.47	2.88	1.75	0.00
H1-derived Mesenchymal Stem Cell	chr16:53765000-53770000	35000	3.75	0.36	None	rs6499640	-	1.22	0.00	0.00	1.75	1.36	1.33	0.00
H1-derived Mesenchymal Stem Cell	chr16:52595000-52600000	1205000	0.01	0.02	Enhancer	rs4784227	-	3.87	0.00	0.00	1.75	4.88	1.33	0.00
H1-derived Mesenchymal Stem Cell	chr16:52585000-52590000	1215000	0.02	0.05	None	rs3803662	TOX3	1.43	0.00	0.00	1.40	2.62	1.75	0.00
H1-derived Mesenchymal Stem Cell	chr16:52635000-52640000	1165000	0.24	0.53	None	rs3112612	-	1.65	0.00	0.00	1.61	1.60	2.27	0.00
H1-derived Mesenchymal Stem Cell	chr16:55245000-55250000	1445000	0.09	0.25	None	rs17291845	-	1.33	0.00	0.00	1.58	1.42	1.75	0.00
H1-derived Mesenchymal Stem Cell	chr16:53845000-53850000	45000	5.59	0.57	None	rs11642841	-	1.43	0.00	0.00	3.34	1.79	1.75	0.00

Showing 1 to 10 of 784 entries First Previous 1 2 3 4 5 ... 79 Next Last

Step 5 : Browse the table



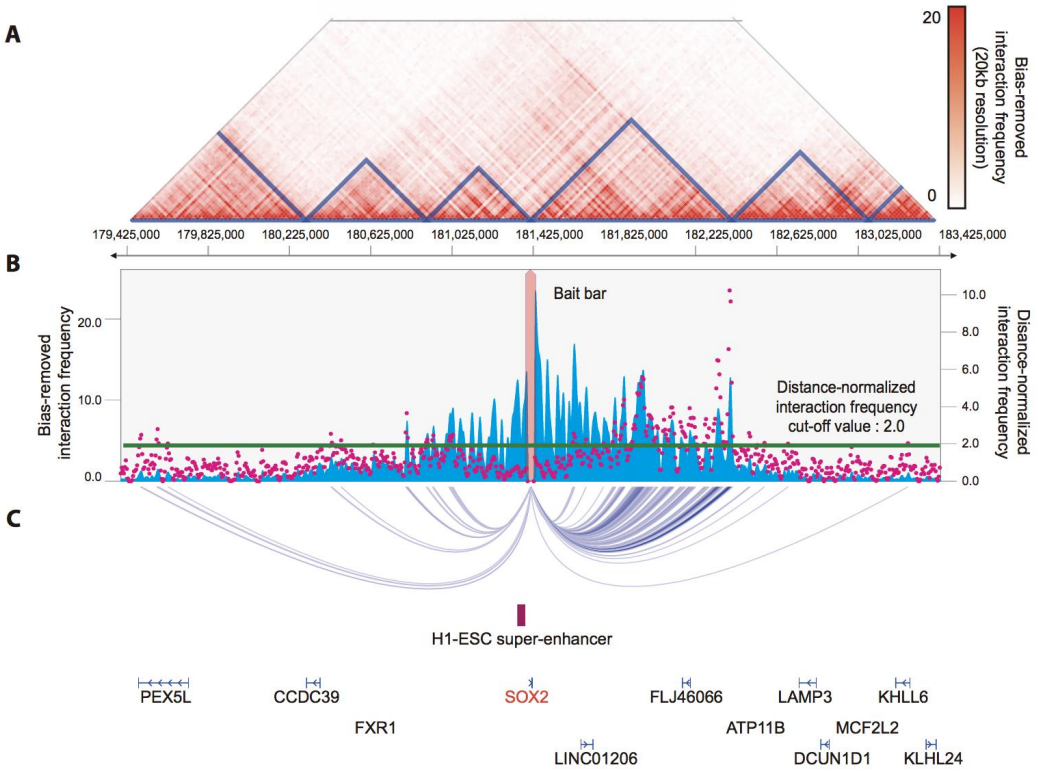
Sample	Bin	Distance	Bias-removed contact intensity	Distance normalized contact intensity	Enhancer	GWAS SNP ID	Gene Name	h3k27ac	h2afz	H3K27me3	h3k36me3	H3K4me1	H3K4me3	h3k9ac
H1 Mesenchymal Stem Cell	chr16:54960000-54965000	1160000	3.72	8.29	None	-	CRNDE,IRX5	5.78	0.00	0.00	1.58	2.70	33.91	0.00
H1 Mesenchymal Stem Cell	chr16:54145000-54150000	345000	3.62	2.71	Non					0.00	4.41	2.06	2.27	0.00
H1 Mesenchymal Stem Cell	chr16:54955000-54960000	1155000	3.58	7.96	Non					0.00	1.61	1.64	1.22	0.00
H1 Mesenchymal Stem Cell	chr16:54465000-54470000	665000	3.27	4.59	None	-	-	1.26	0.00	0.00	1.58	1.60	1.75	0.00
H1 Mesenchymal Stem Cell	chr16:54460000-54465000	660000	3.19	4.44	None	-	-	1.65	0.00	0.00	1.93	2.70	1.75	0.00
H1 Mesenchymal Stem Cell	chr16:54965000-54970000	1165000	3.16	7.07	None	-	CRNDE,IRX5	2.60	0.00	0.00	3.52	4.15	23.90	0.00
H1 Mesenchymal Stem Cell	chr16:54315000-54320000	515000	3.08	3.43	Non					0.00	2.28	4.34	27.59	0.00
H1 Mesenchymal Stem Cell	chr16:54950000-54955000	1150000	3.00	6.64	None	-	-	1.22	0.00	0.00	1.75	1.60	2.80	0.00
H1 Mesenchymal Stem Cell	chr16:54235000-54240000	435000	2.90	2.77	None	-	-	1.54	0.00	0.00	1.75	1.97	1.74	0.00
H1 Mesenchymal Stem Cell	chr16:54140000-54145000	340000	2.88	2.12	None	-	-	1.33	0.00	0.00	1.93	1.50	2.54	0.00
H1 Mesenchymal Stem Cell	chr16:54310000-54315000	510000	2.82	3.12	None	-	-	1.75	0.00	0.00	2.02	3.06	2.27	0.00
H1 Mesenchymal Stem Cell	chr16:54300000-54305000	500000	2.78	3.02	None	-	-	1.33	0.00	0.00	2.64	1.97	2.27	0.00
H1 Mesenchymal Stem Cell	chr16:54240000-54245000	440000	2.69	2.59	None	-	-	1.65	0.00	0.00	1.75	1.79	2.27	0.00
H1 Mesenchymal Stem Cell	chr16:54470000-54475000	670000	2.67	3.76	None	-	-	1.54	0.00	0.00	1.61	1.79	1.75	0.00
H1 Mesenchymal Stem Cell	chr16:54230000-54235000	430000	2.65	2.50	None	-	-	1.54	0.00	0.00	2.11	1.50	3.33	0.00
H1 Mesenchymal Stem Cell	chr16:54320000-54325000	520000	2.56	2.88	None	-	IRX3	3.87	0.00	0.00	2.28	6.16	18.62	0.00

Promoter of CRNDE

Promoter of IRX5

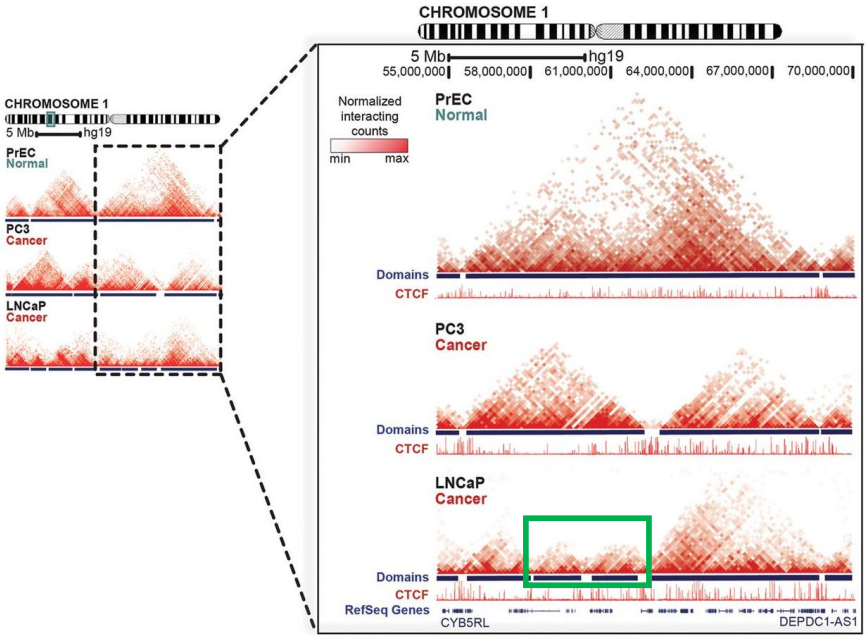
Promoter of IRX3

Module 2 : Interaction Visualization



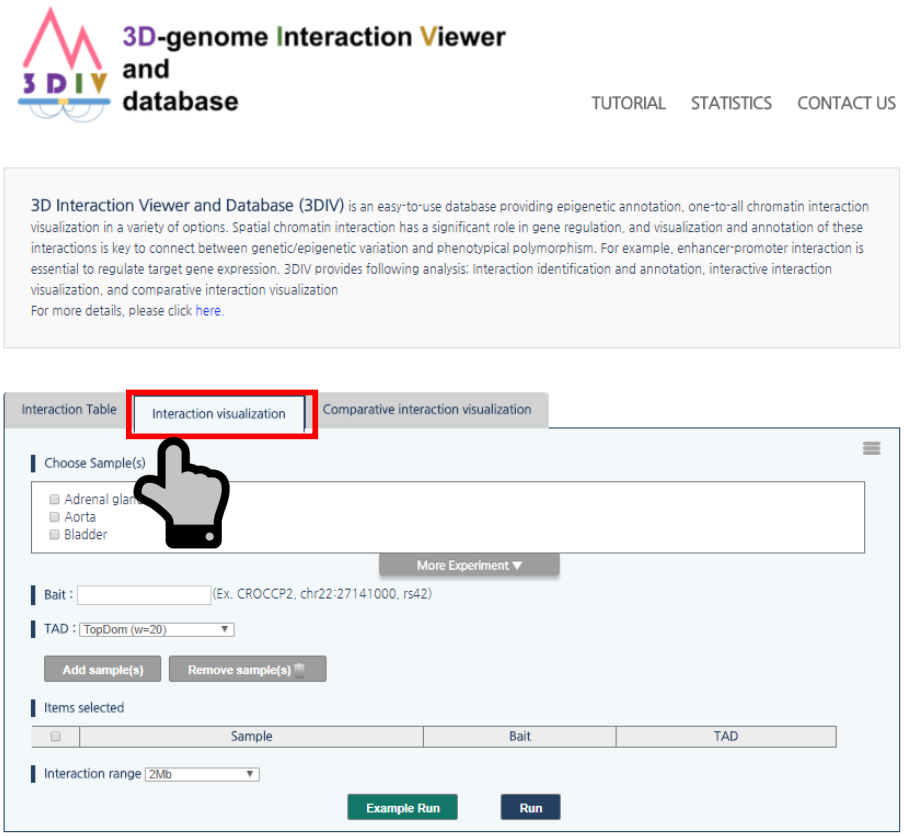
Example : Interaction profile of SOX2

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

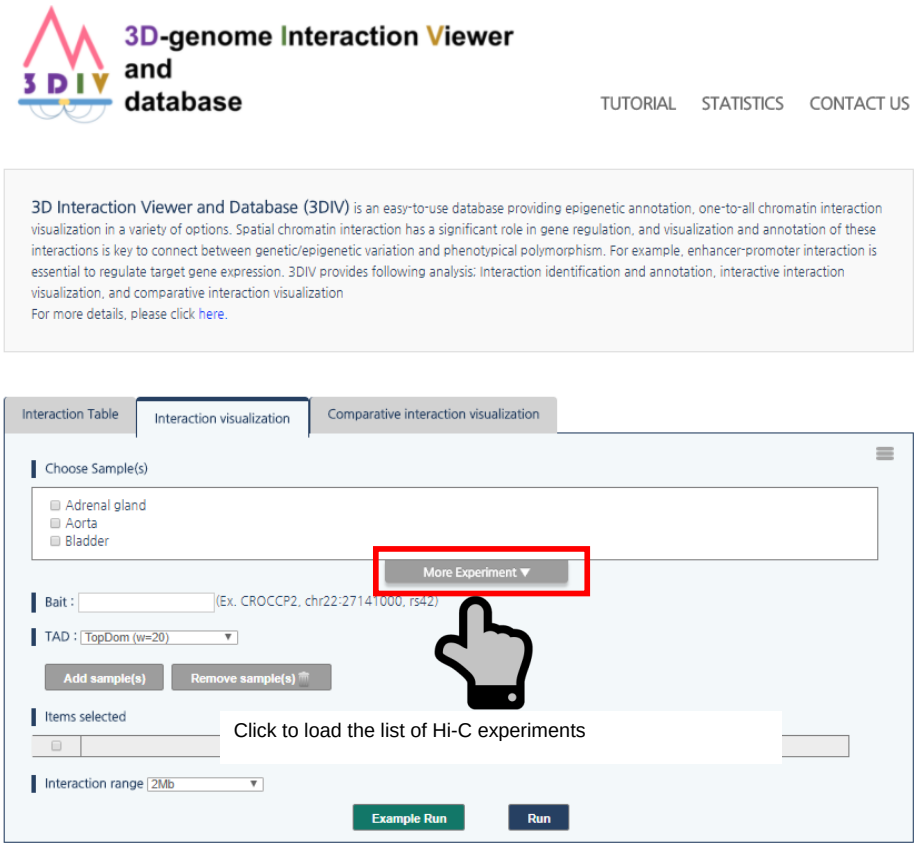


Taberlay et al, *Genome Res.* (2016)

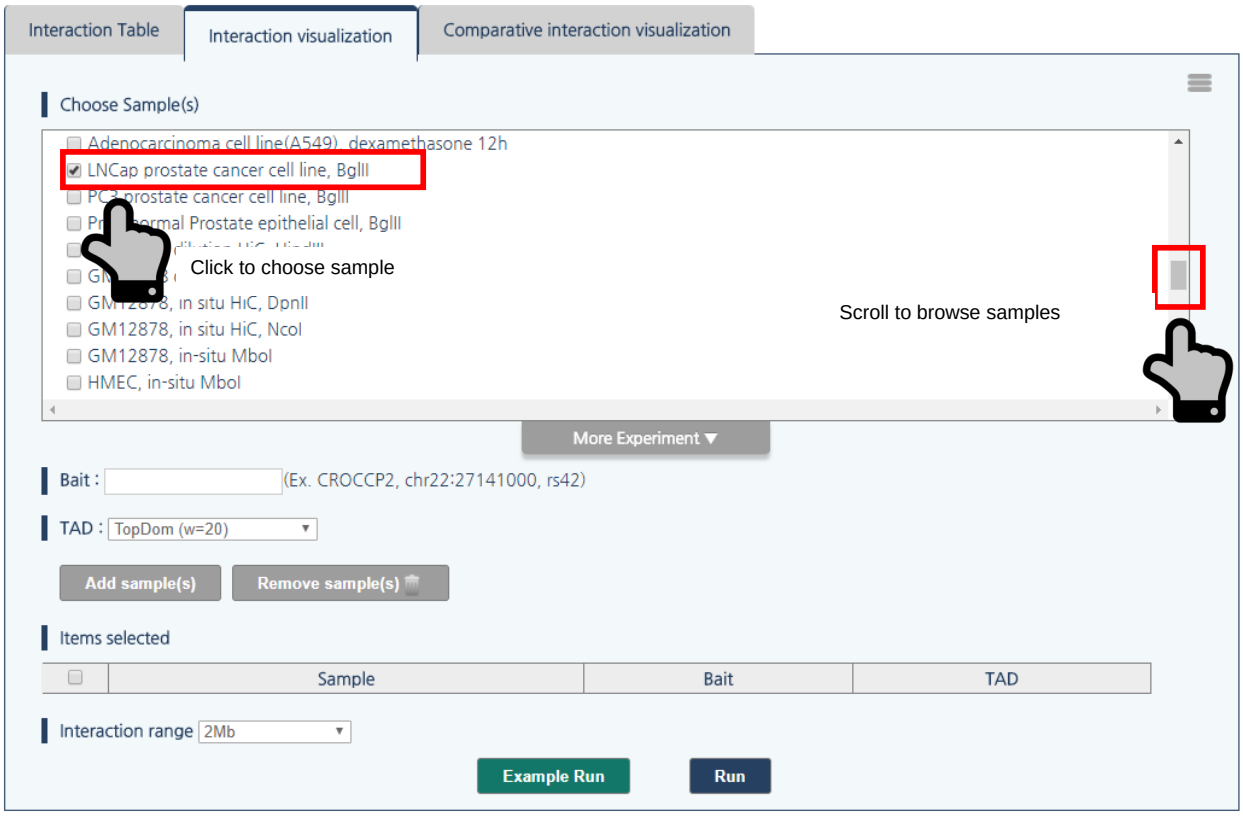
Step 1 : Open Interaction Visualization Module



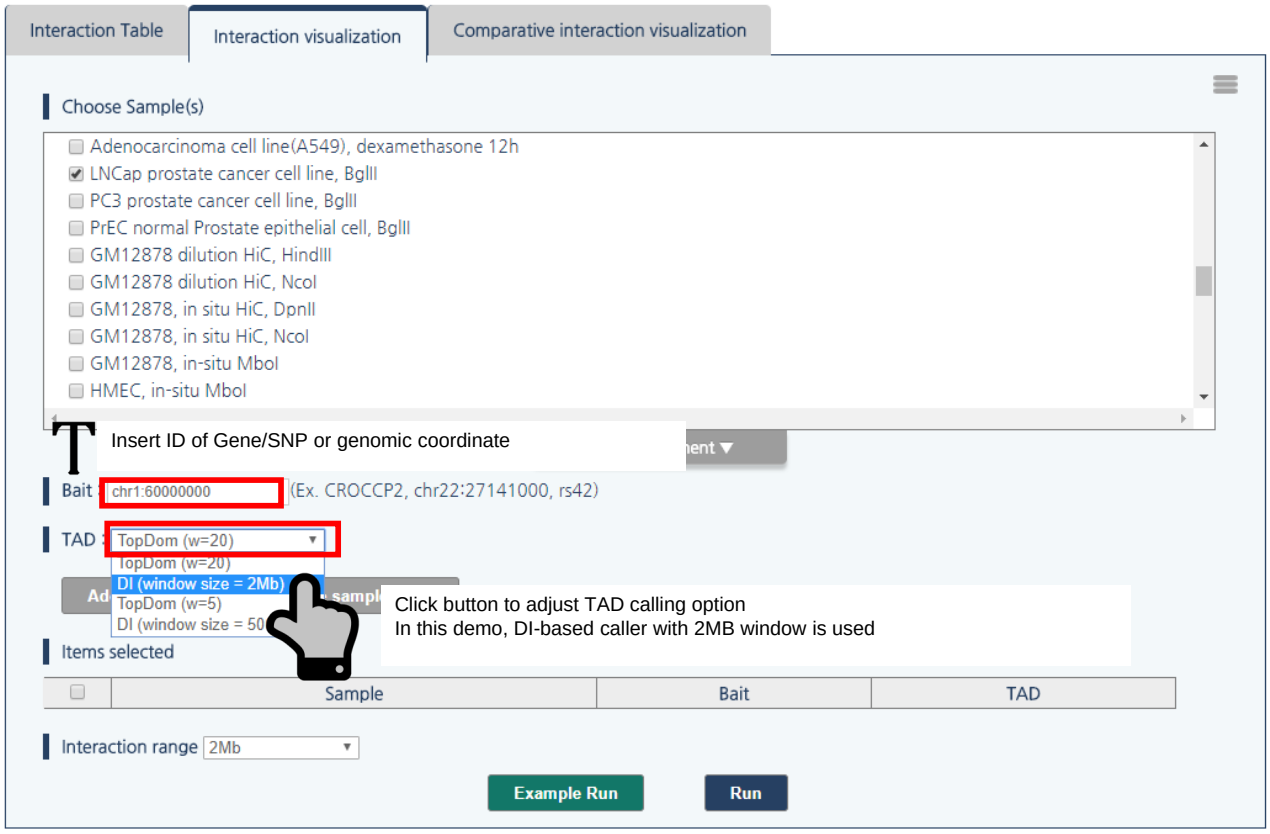
Step 2 : Choose a sample



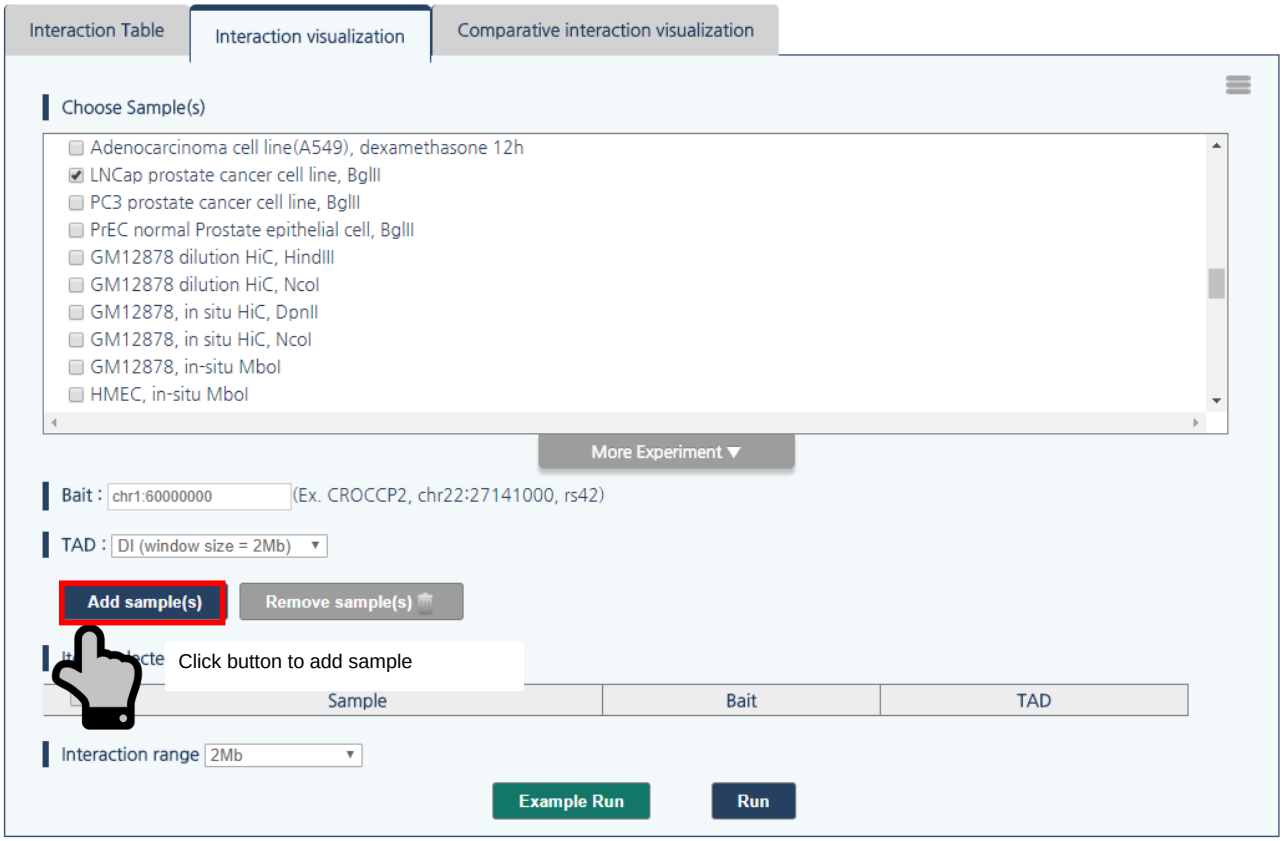
Step 2 : Choose a sample



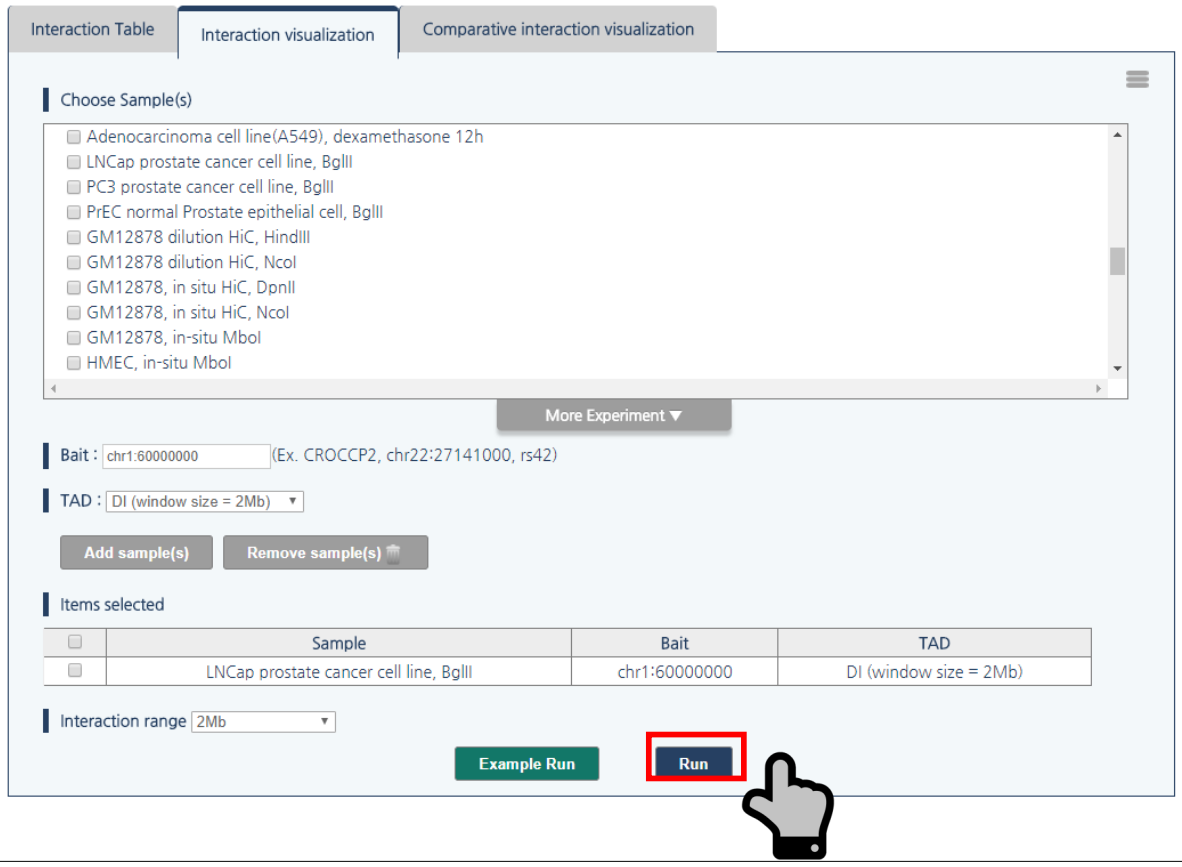
Step 3 : Choose a bait & TAD calling option



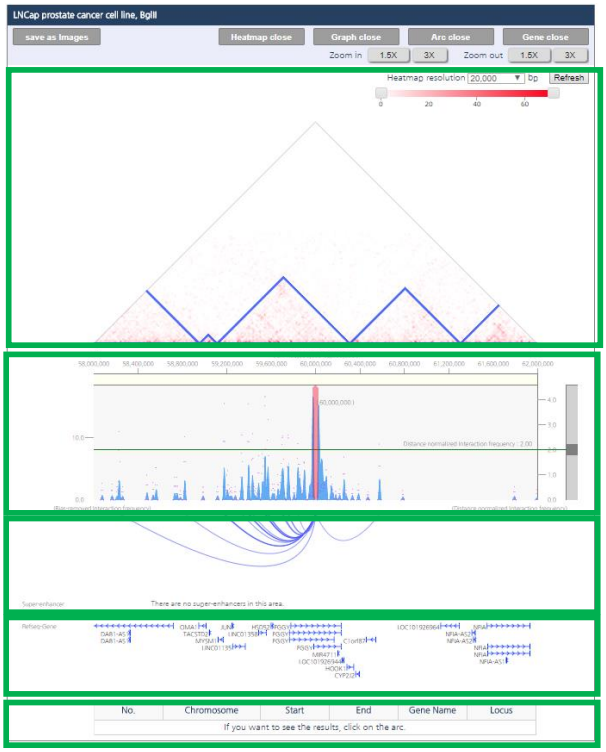
Step 3 : Choose a Bait & TAD calling option



Step 4 : Run Module



Step 5 : Adjust the interaction visualization



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

One-to-all interaction plot

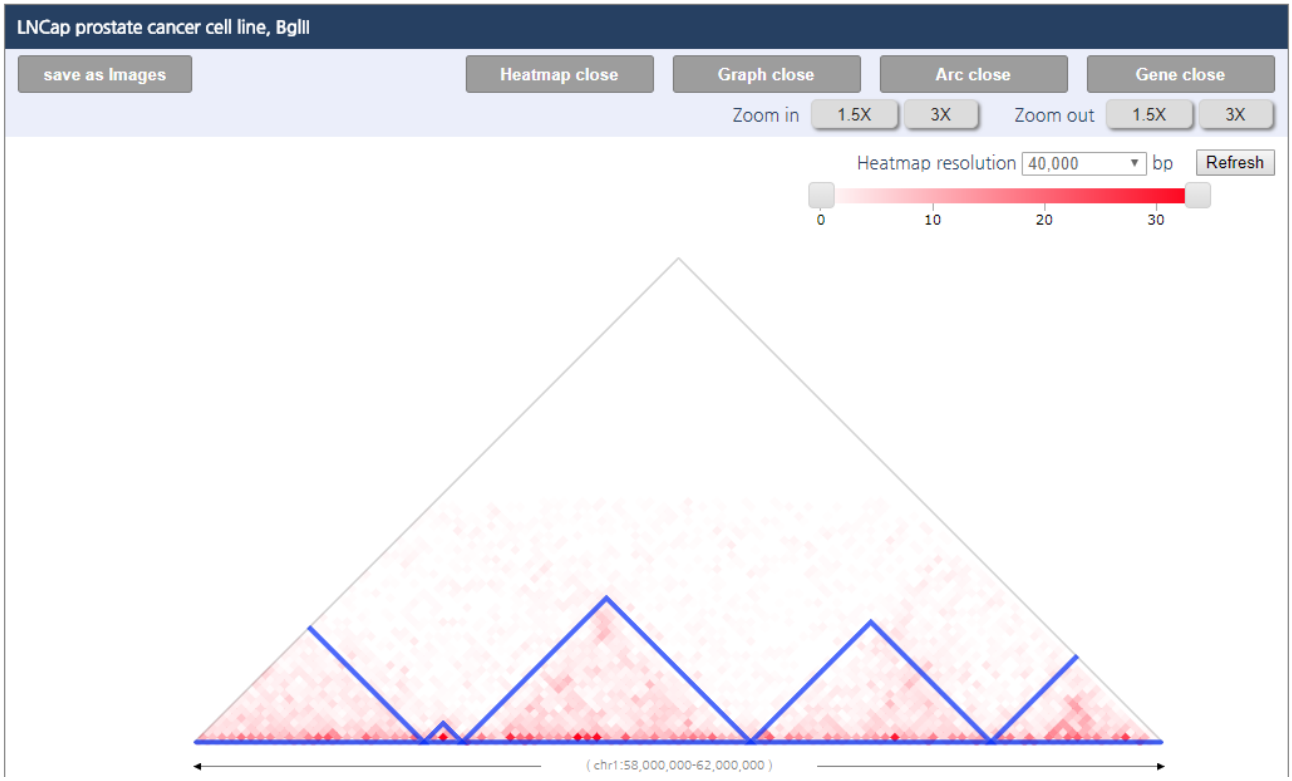
Arc-representation of significant interactions

RefSeq Gene annotations

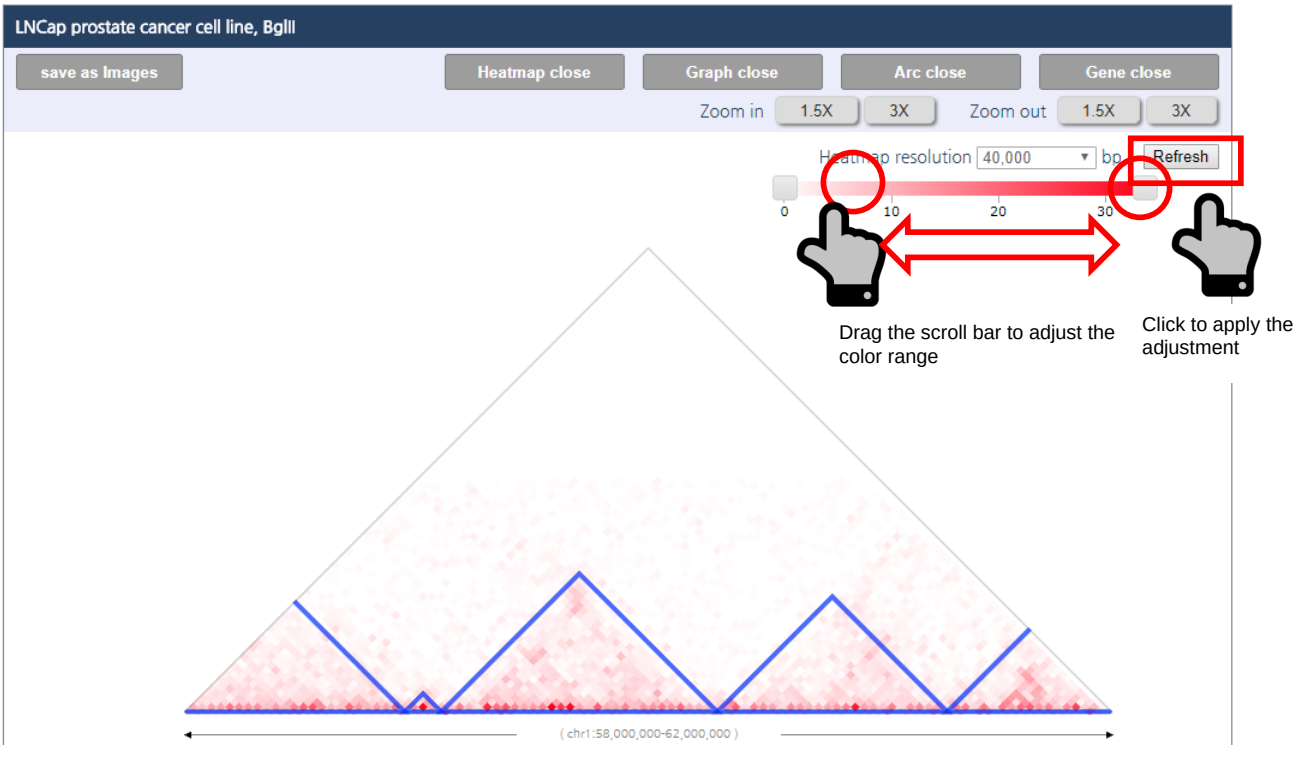
Description of selected interaction

Step 5a : Adjust the heatmap resolution

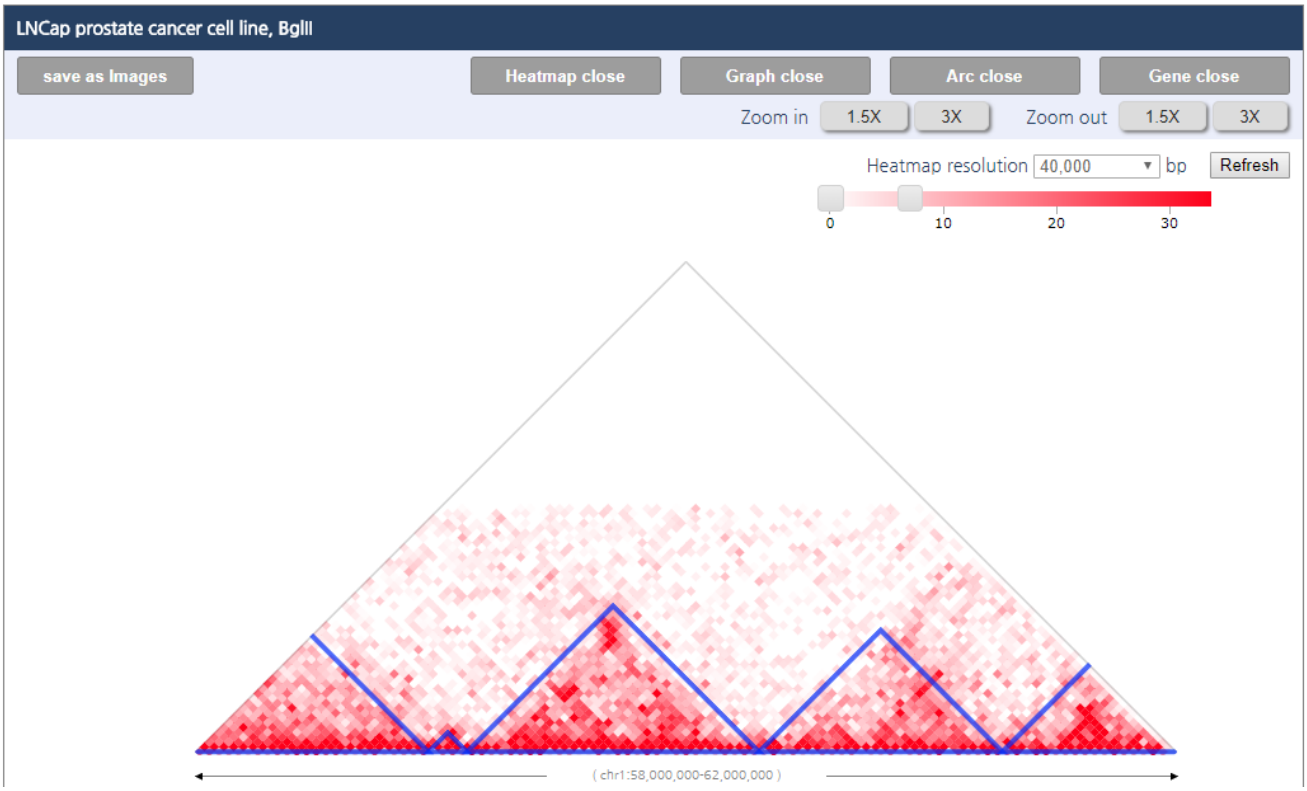
Step 5a : Adjust the heatmap resolution



Step 5b : Adjust the heatmap color range

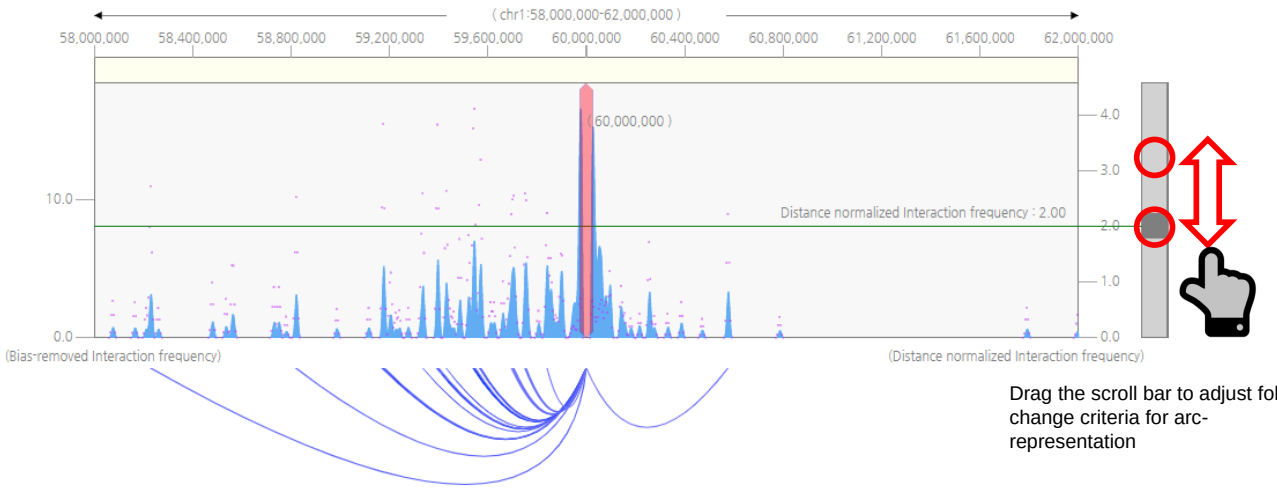


Step 5b : Adjust the heatmap color range



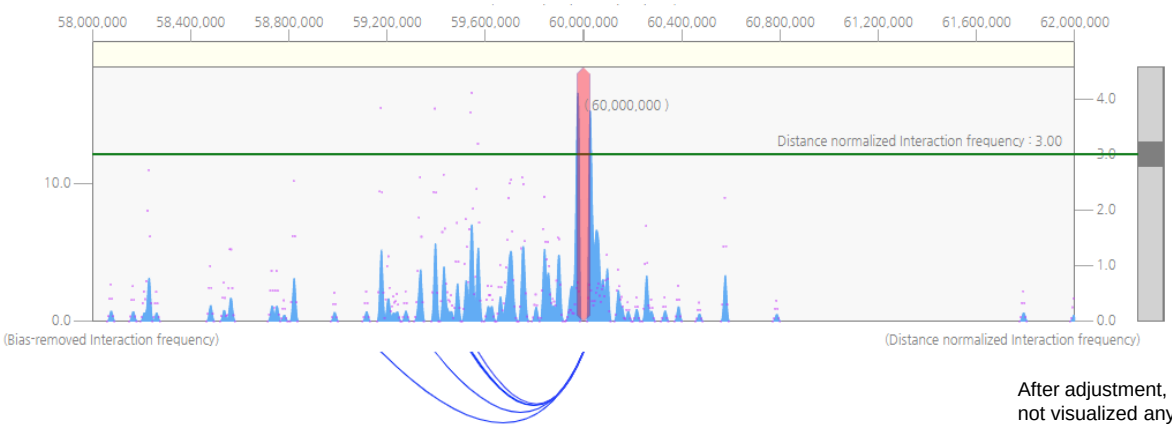
Step 5 : Adjust the interaction visualization

Adjust the fold-change criteria



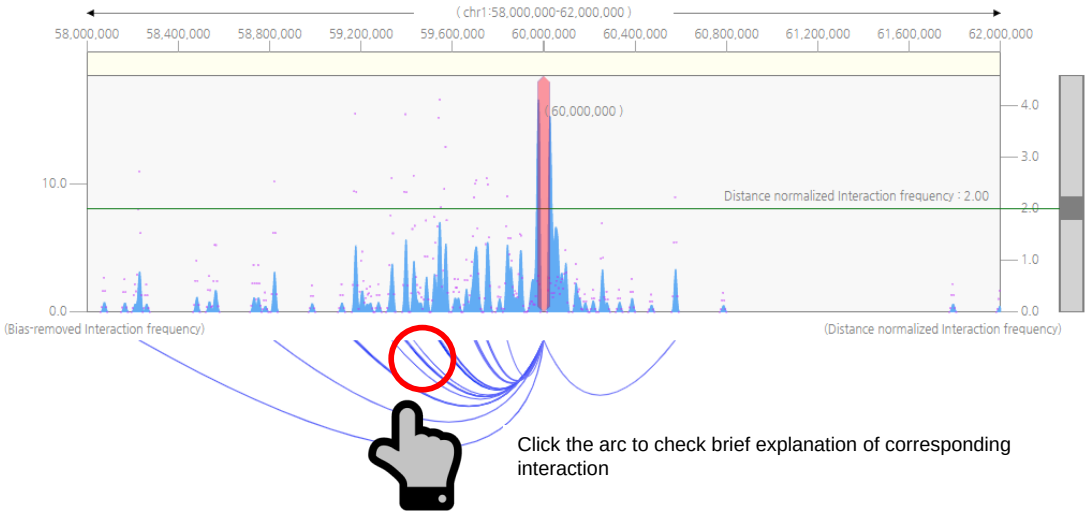
Drag the scroll bar to adjust fold-change criteria for arc-representation

Adjust the fold-change criteria



After adjustment, some arcs are not visualized any more.

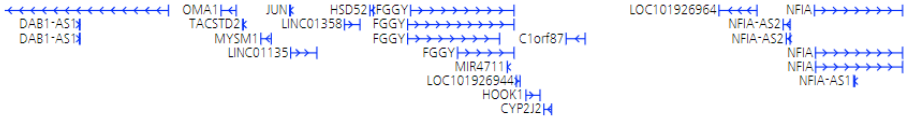
Description of identified interactions



Click the arc to check brief explanation of corresponding interaction

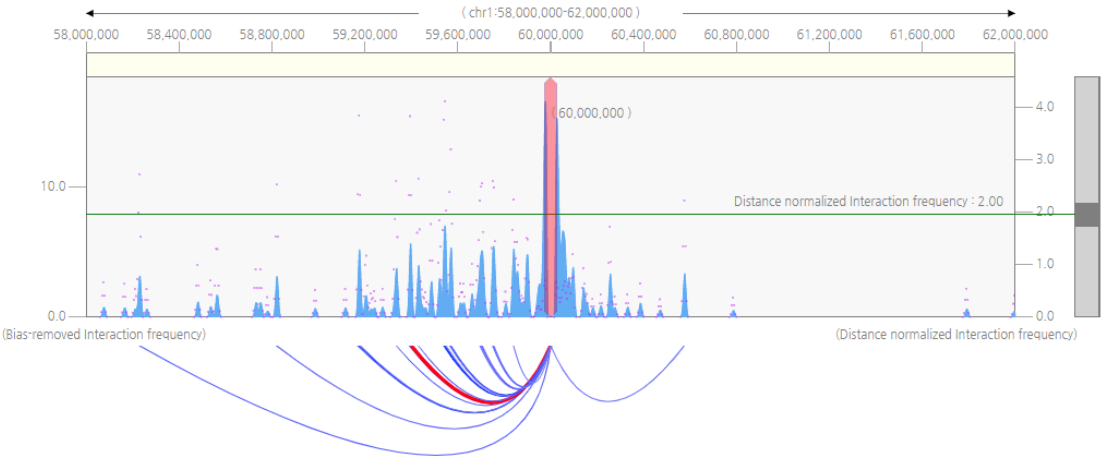
Super-enhancer There are no super-enhancers in this area.

Refseq-Gene



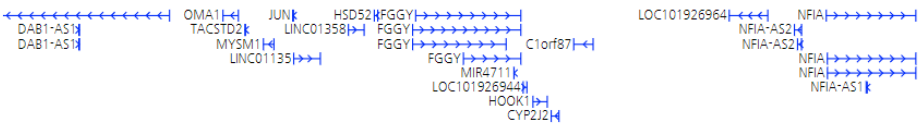
No.	Chromosome	Start	End	Gene Name	Locus
If you want to see the results, click on the arc.					

Description of identified interactions



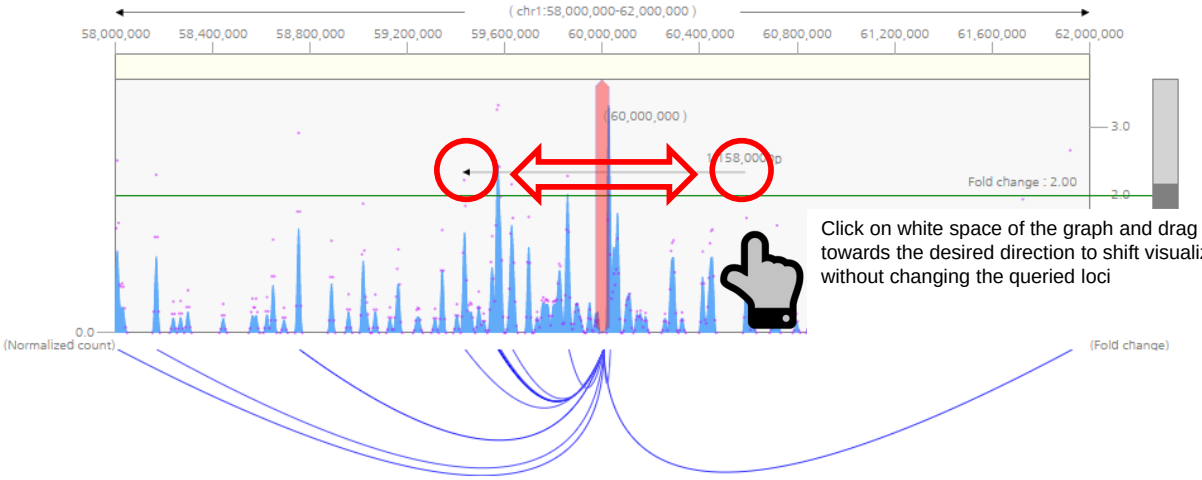
Super-enhancer There are no super-enhancers in this area.

Refseq-Gene



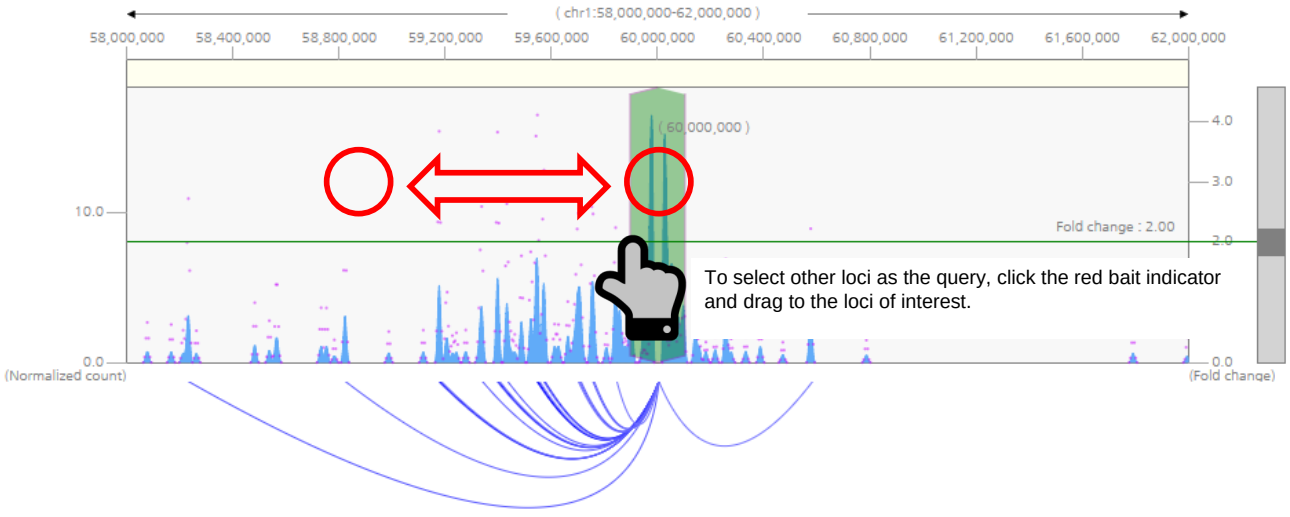
No.	Chromosome	Start	End	Gene Name	Locus
1	chr1	59400000	59405000		

Browse interaction frequency w/o change the bait



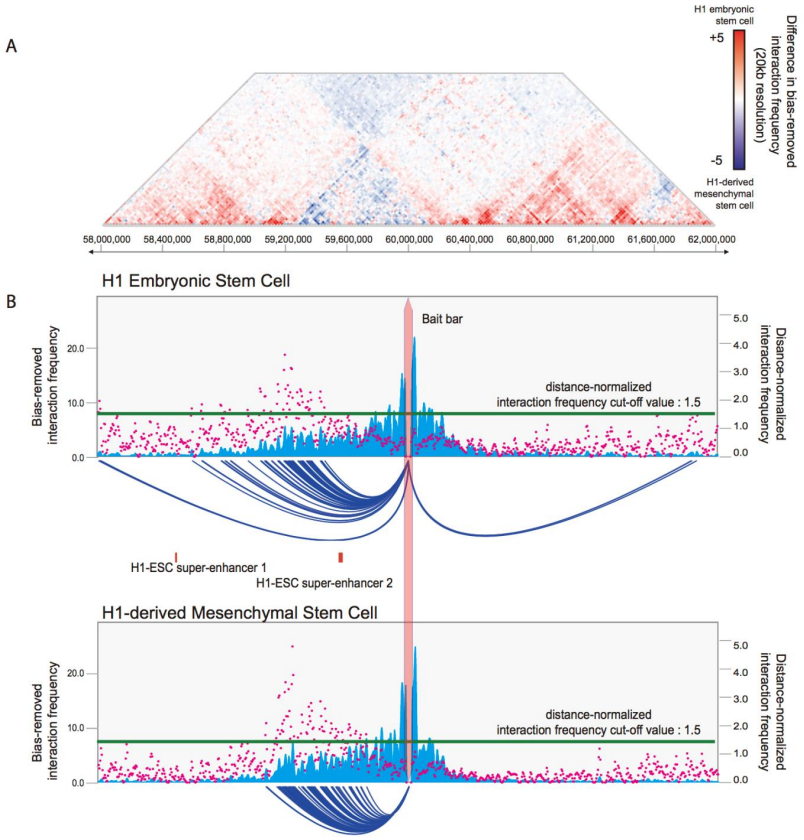
Click on white space of the graph and drag horizontally towards the desired direction to shift visualization range without changing the queried loci

Adjust bait without resubmission



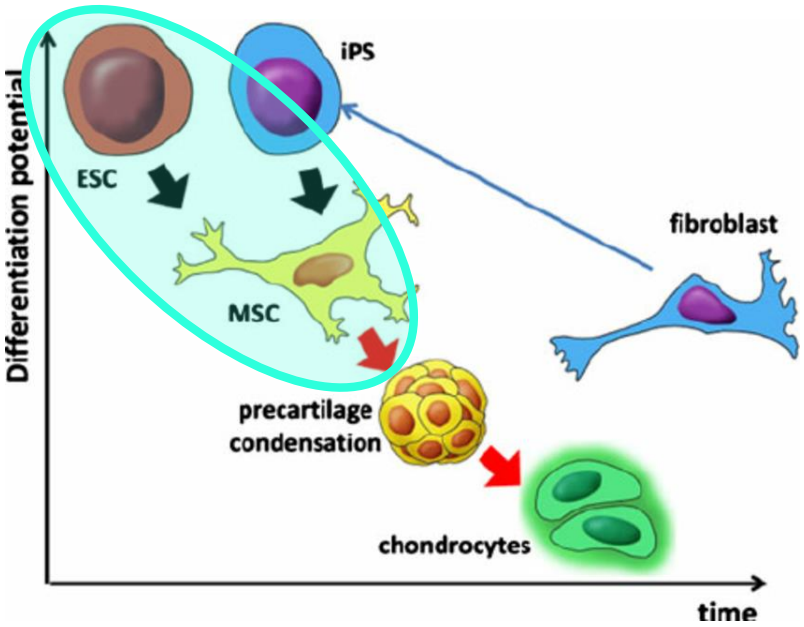
To select other loci as the query, click the red bait indicator and drag to the loci of interest.

Module 3 : Comparative Visualization



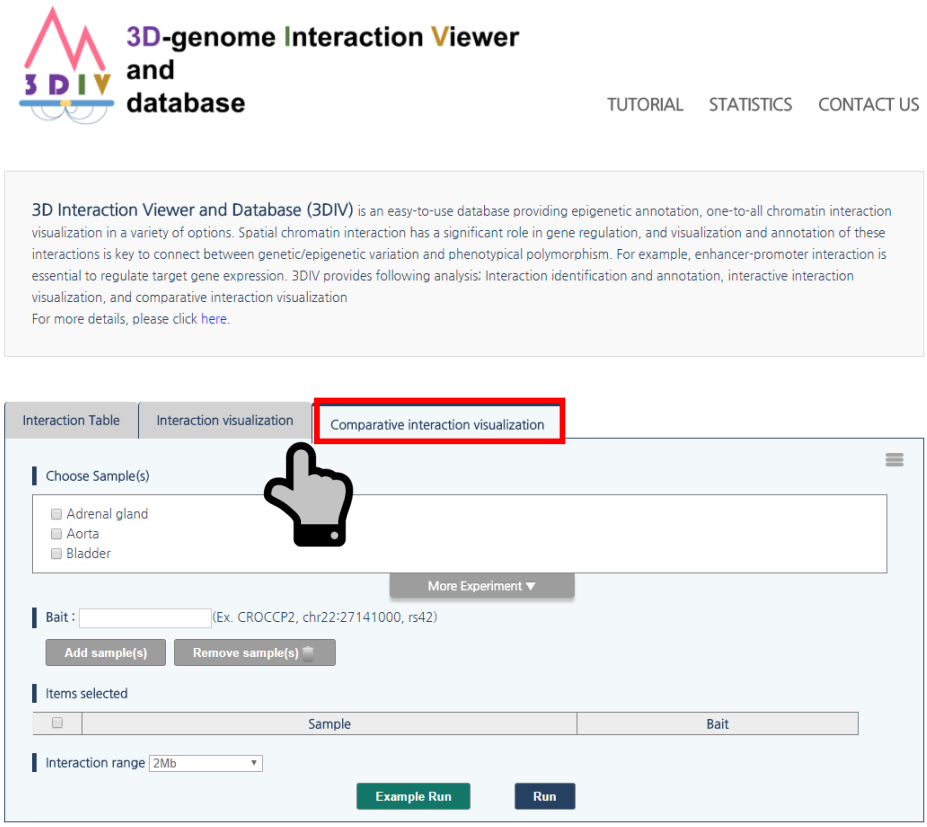
Example : Interaction change during differentiation

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

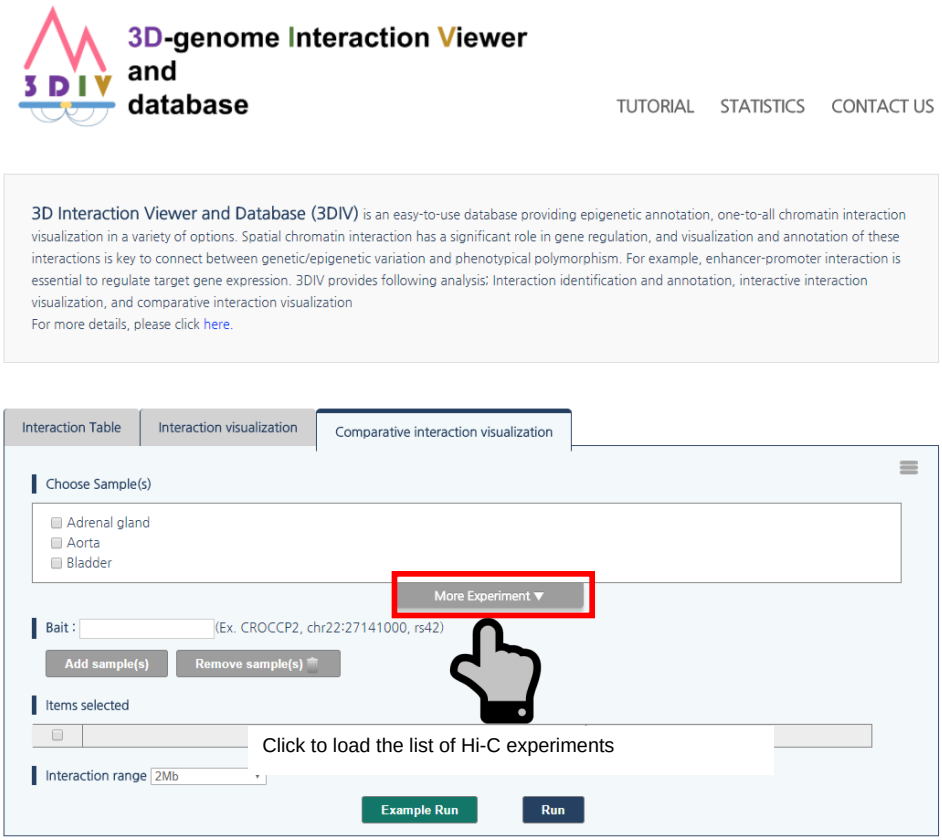


Gadjanski et al, Stem Cell Rev. Rep. (2012)

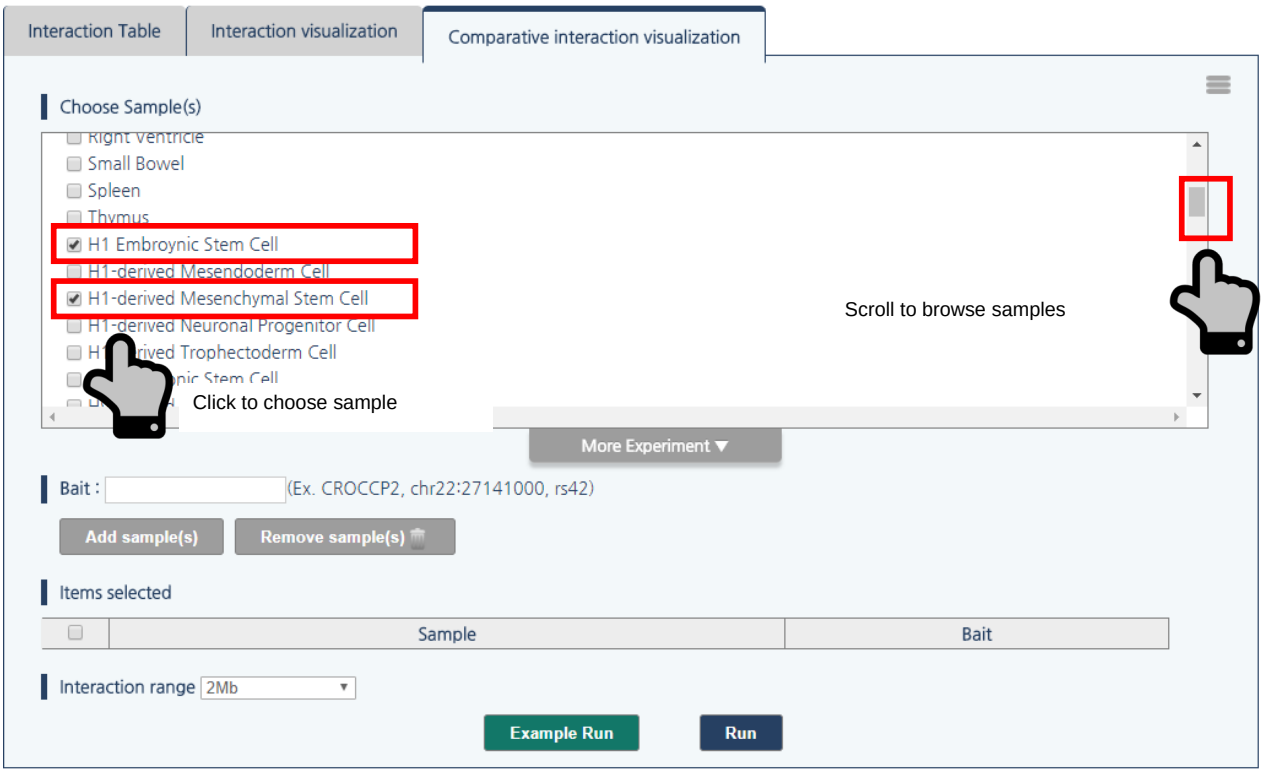
Step 1 : Open Comparative visualization Module



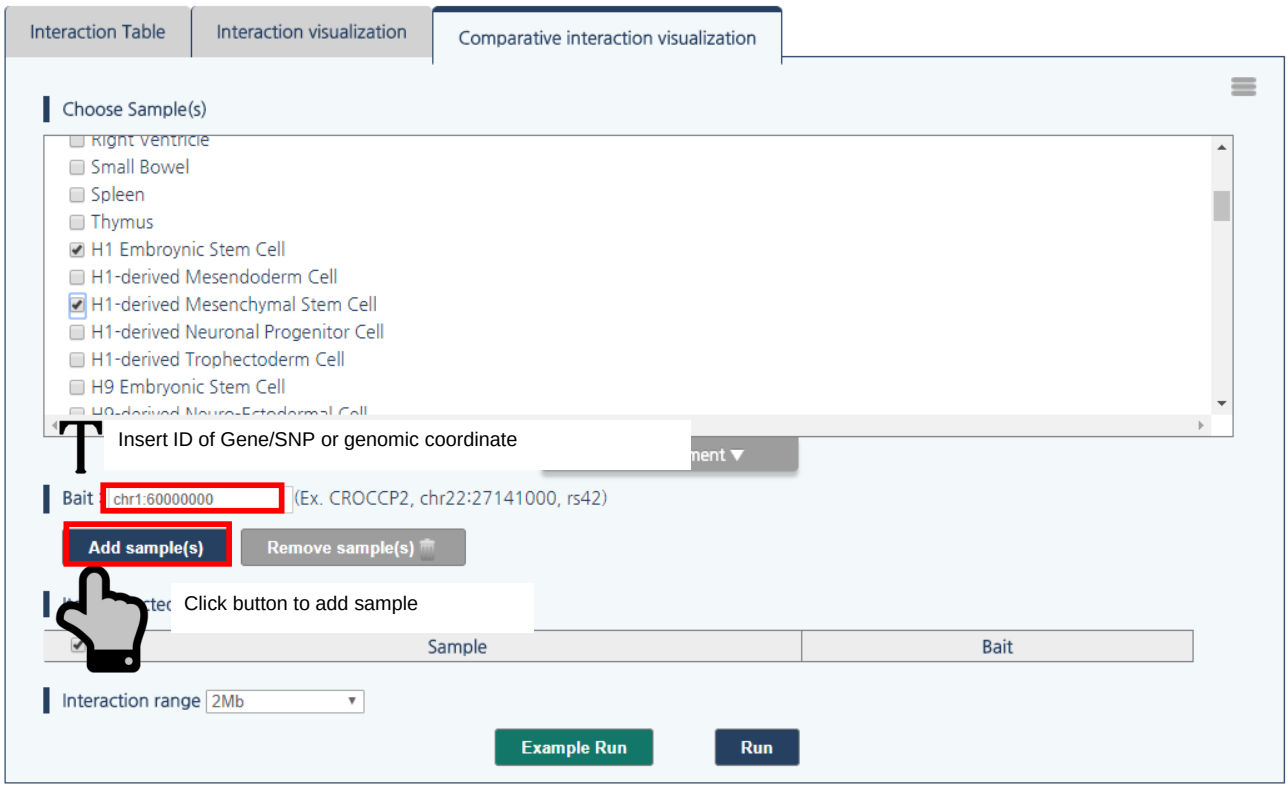
Step 2 : Choose a sample



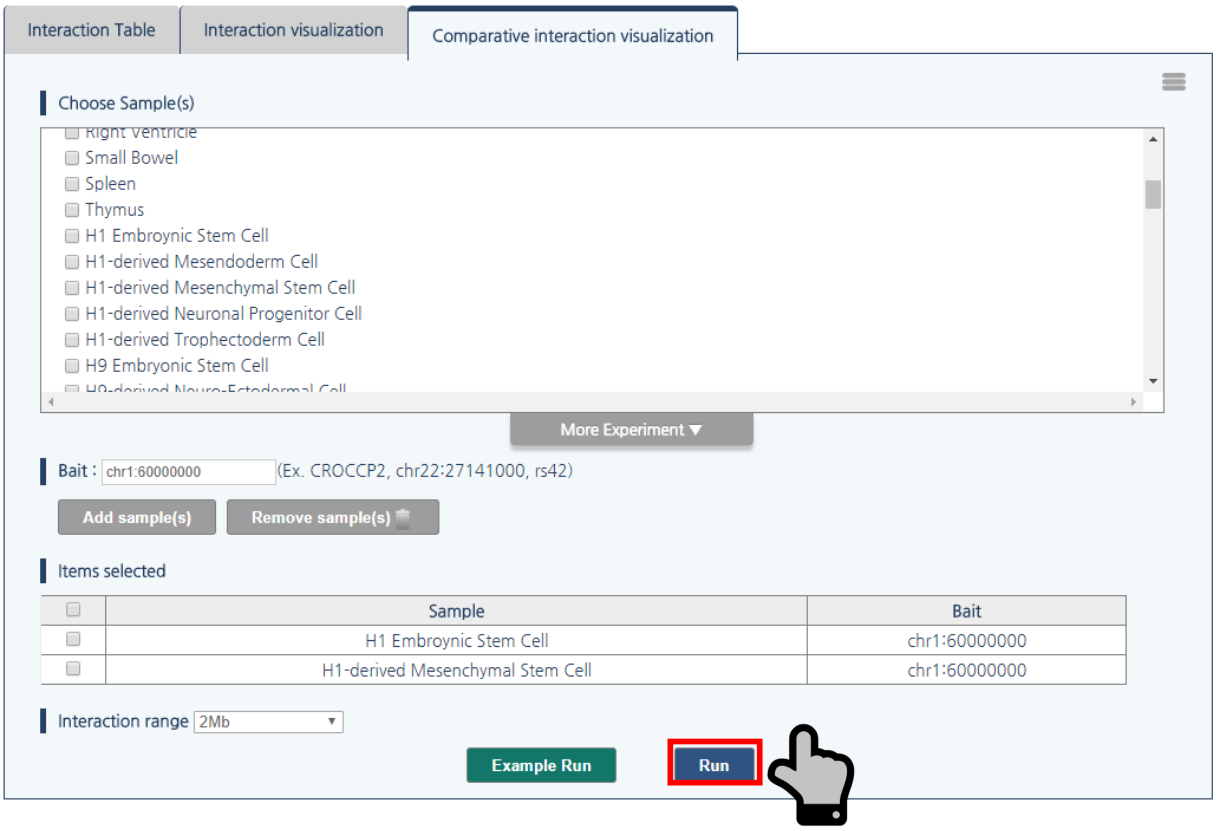
Step 2 : Choose a sample



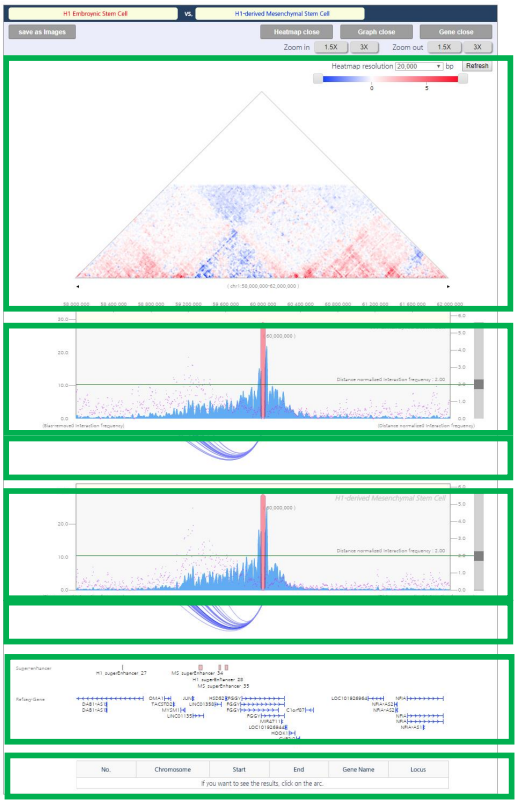
Step 3 : Choose a Bait



Step 4 : Run Module



Step 5 : Adjust comparative heatmap



Comparative heatmap of interaction frequency between 1st and 2nd samples.

Arc-representation of significant interactions in 1st sample

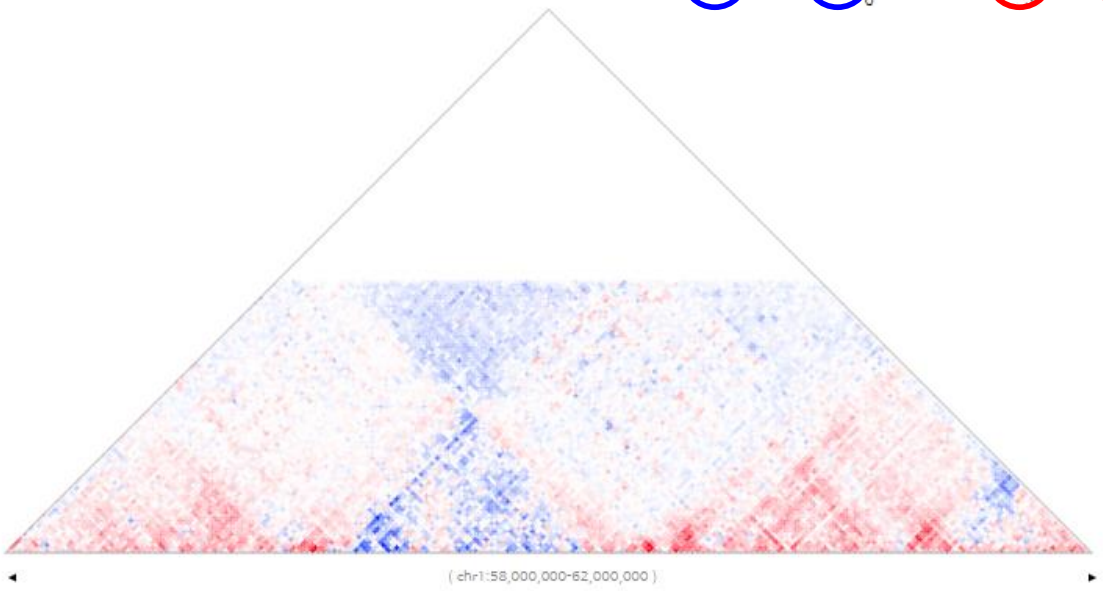
Arc-representation of significant interactions in 2nd sample

RefSeq Genes and super enhancer annotations

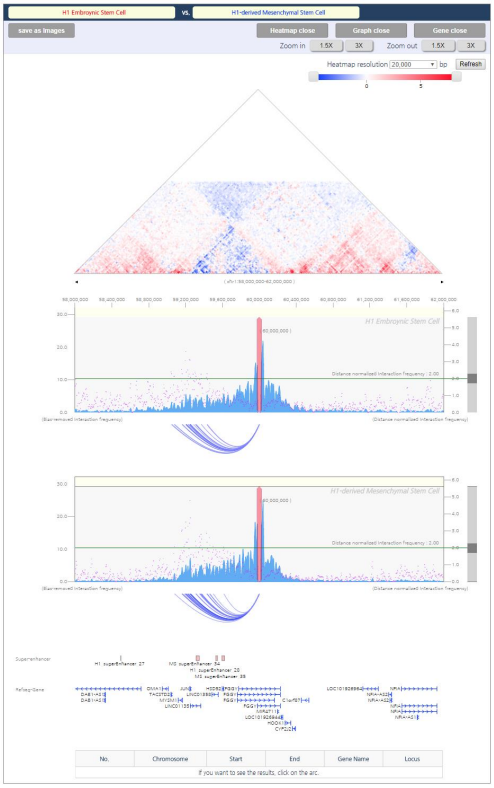
Description of selected interaction

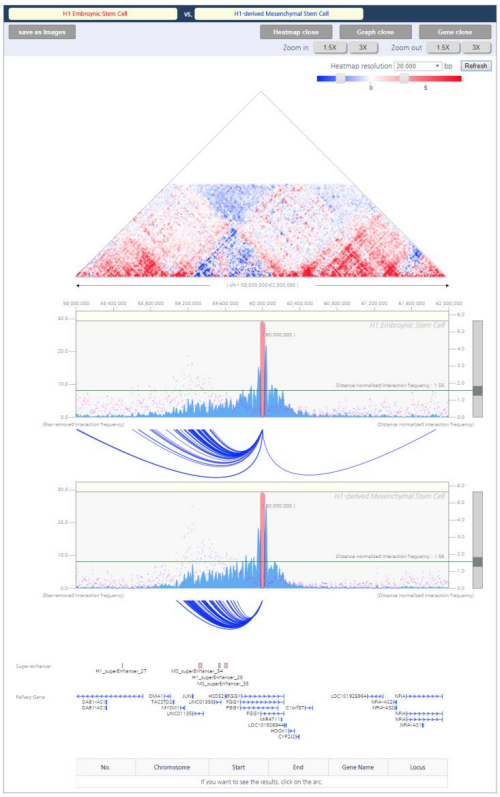
Step 5a : Synchronized criteria change

Step 5b : Adjust the heatmap color range



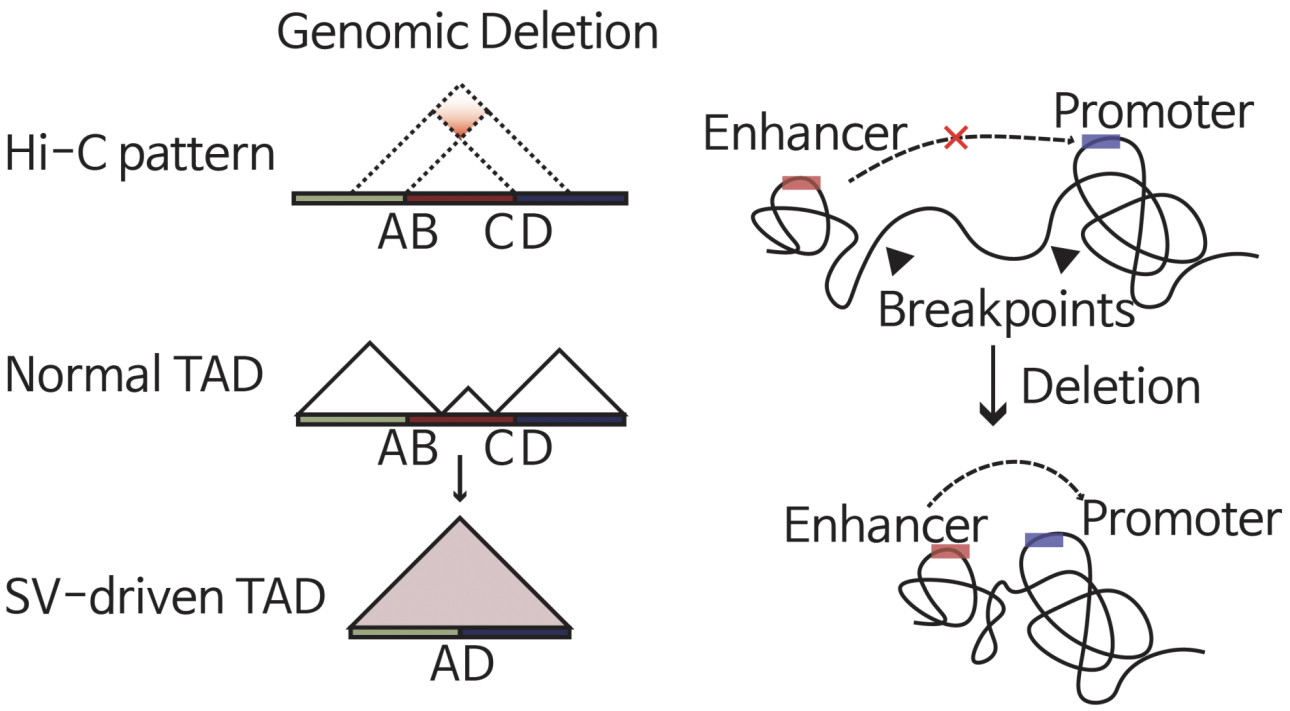
Step 5 : Adjust comparative heatmap





Cancer Hi-C Analysis

암 유전체의 대규모 구조 변이는 염색질 3차구조에 어떠한 변화를 가져올까?



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

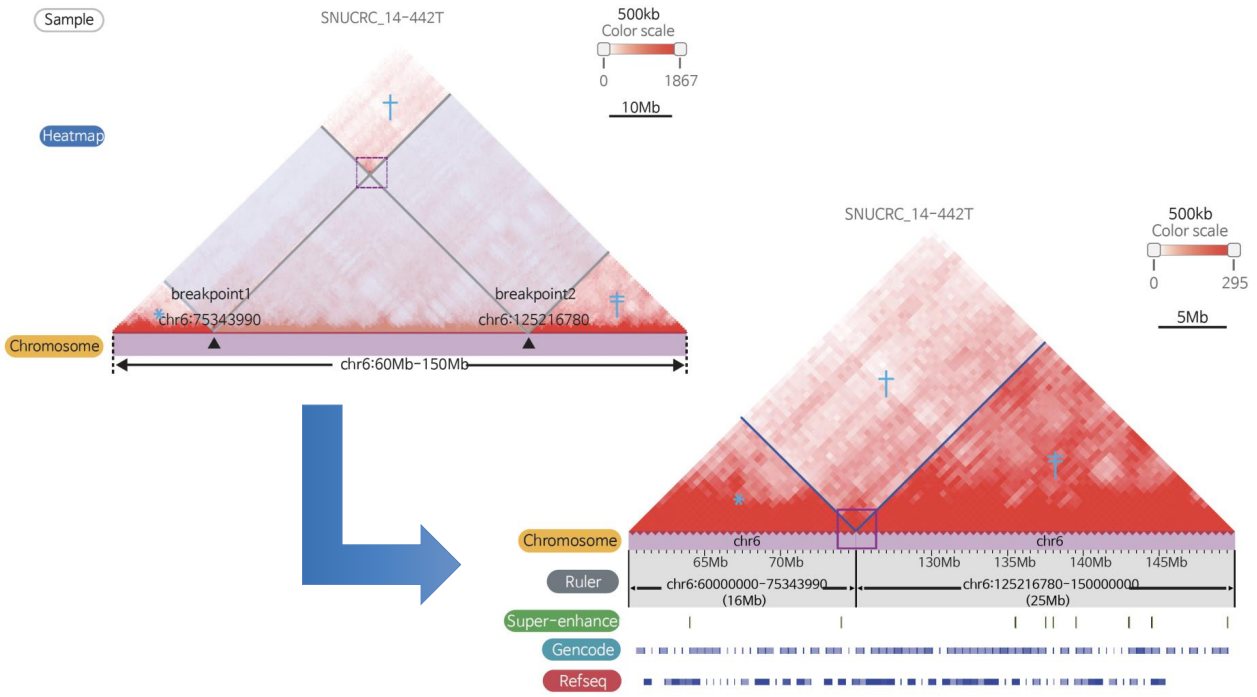
Resolving issue
1. Collection of large cancer/normal Hi-C and pcHi-C data
2. Visualization of cancer 3D genome
3. Hi-C contact map manipulation to examine impact of SVs

Module I.
Pre-called SV and 3D genome

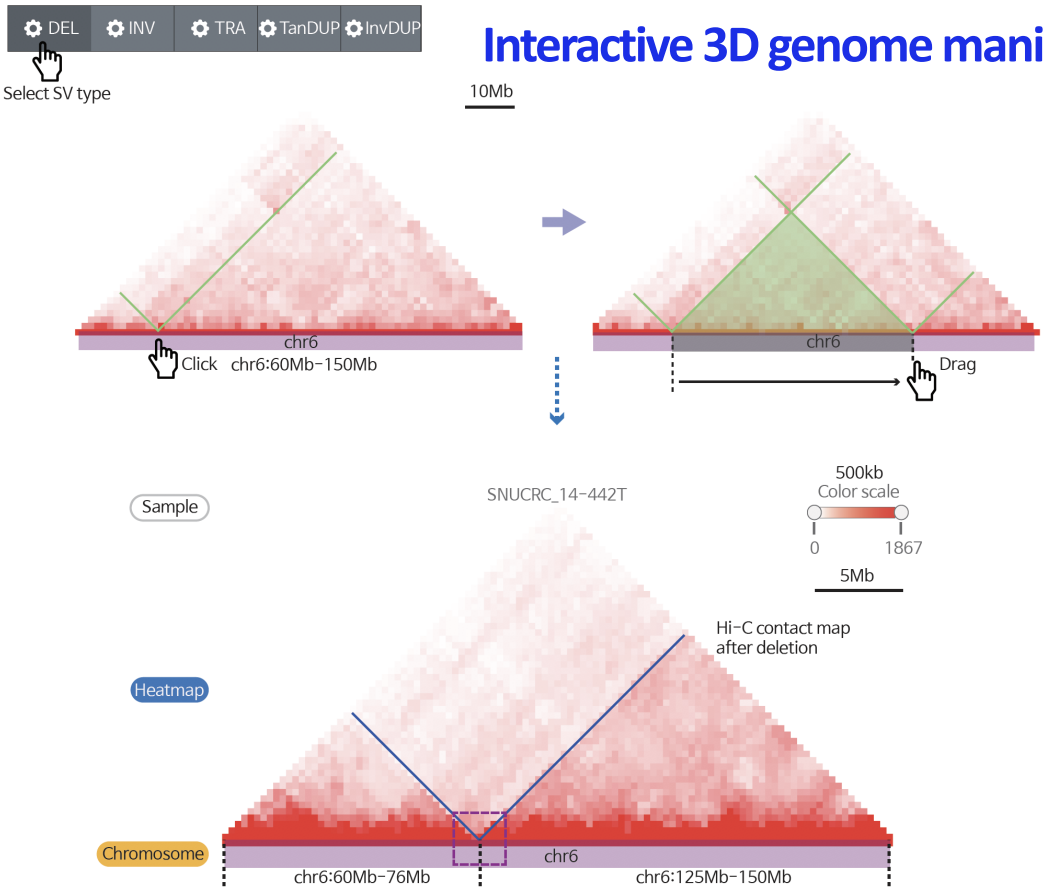
Sample	Chrom1	Breakpoint1	Chrom2	Breakpoint2	SV type	Orientation
14-442T	chr6	...	chr6	...	INV	3to3
14-442T	chr6	...	chr6	...	INV	5to5
14-442T	chr6	75343990	chr6	125216780	DEL	3to5

Showing 1 to 3 of 3 entries Previous 1 Next

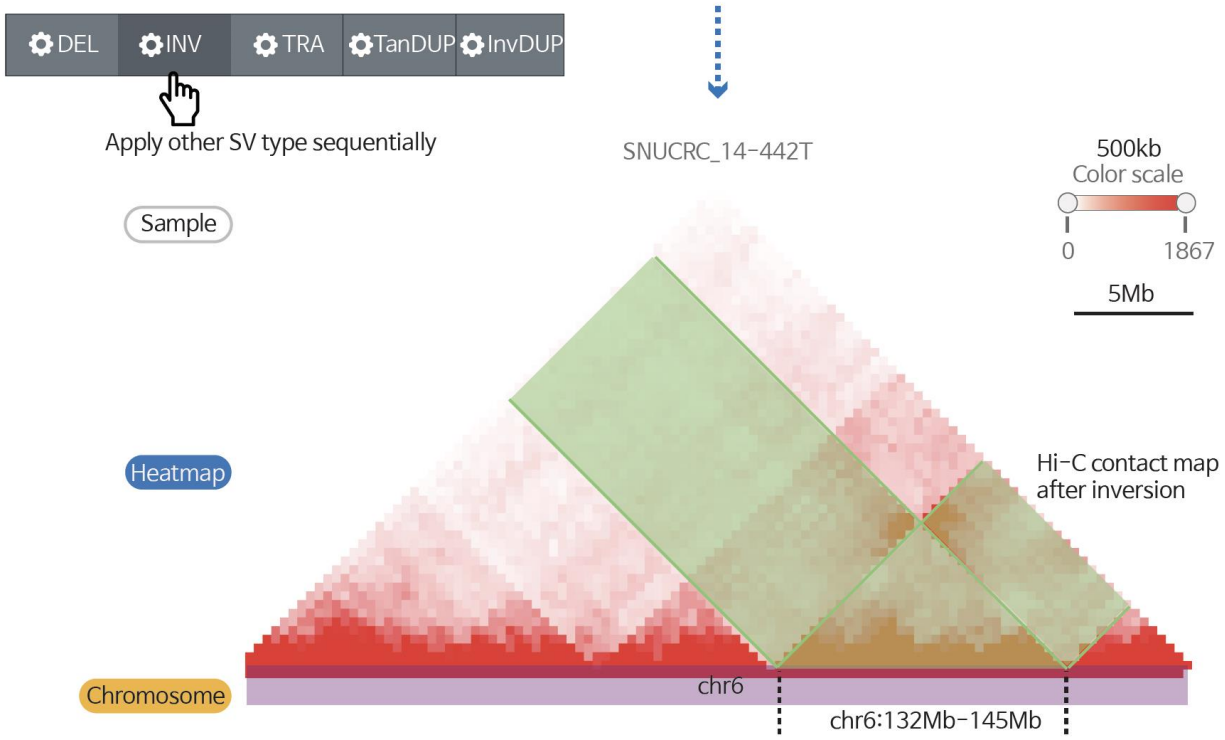
Click SV to visualize



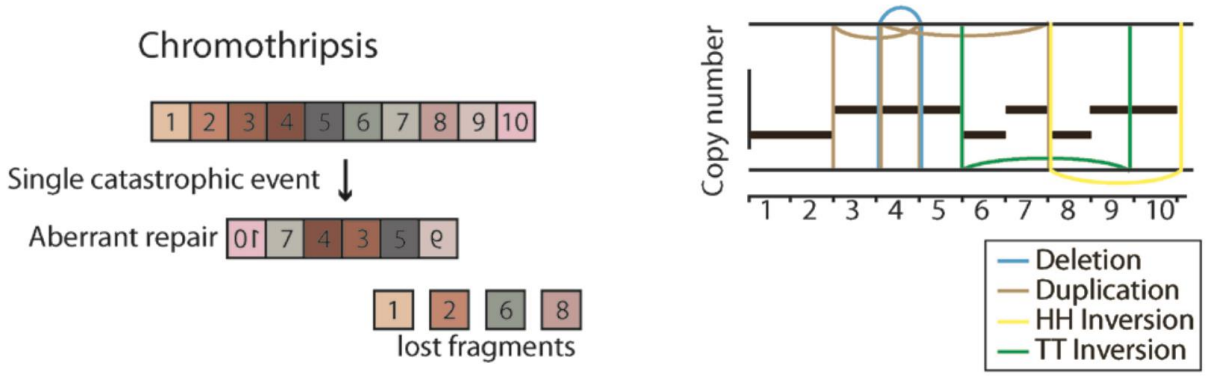
Module II.
Interactive 3D genome manipulation

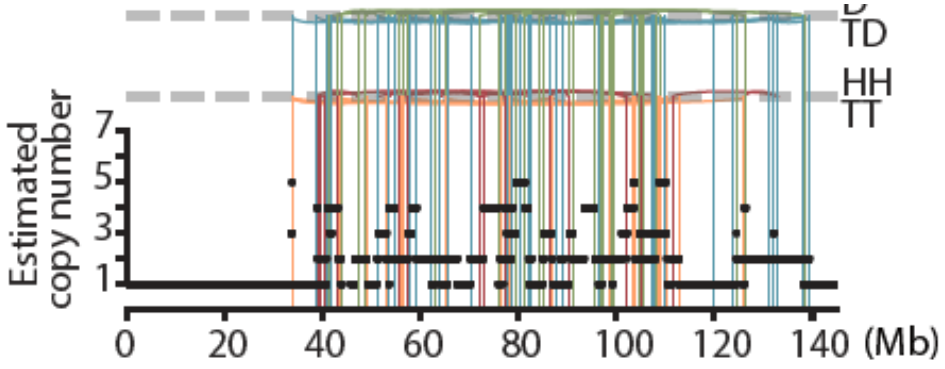


Module II.
Interactive 3D genome manipulation

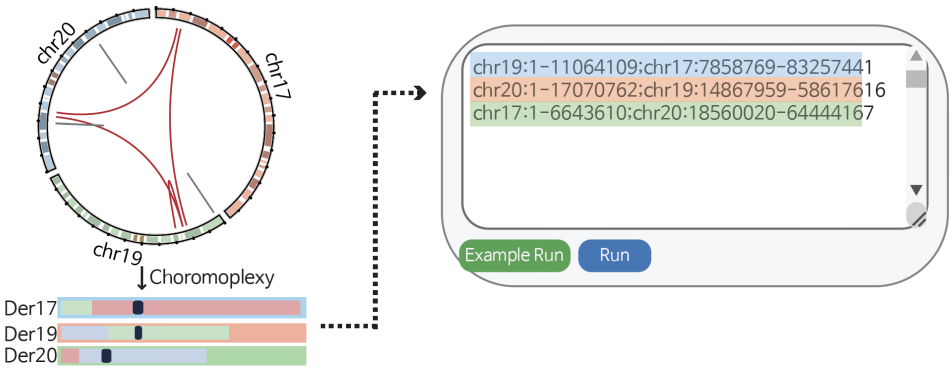


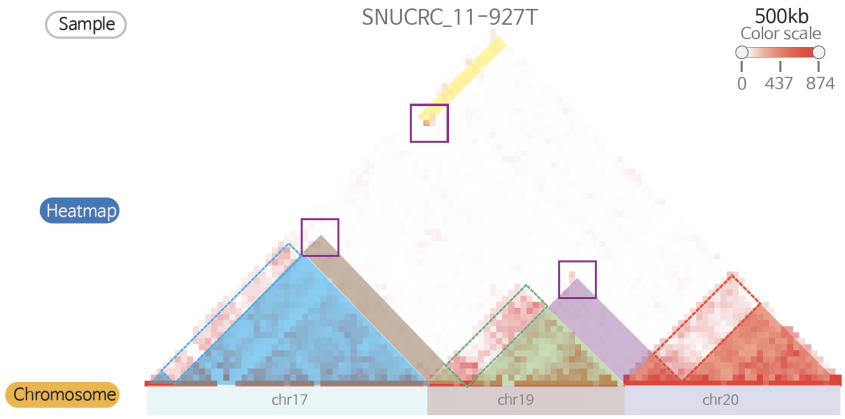
Complex forms of large-scale structural variations



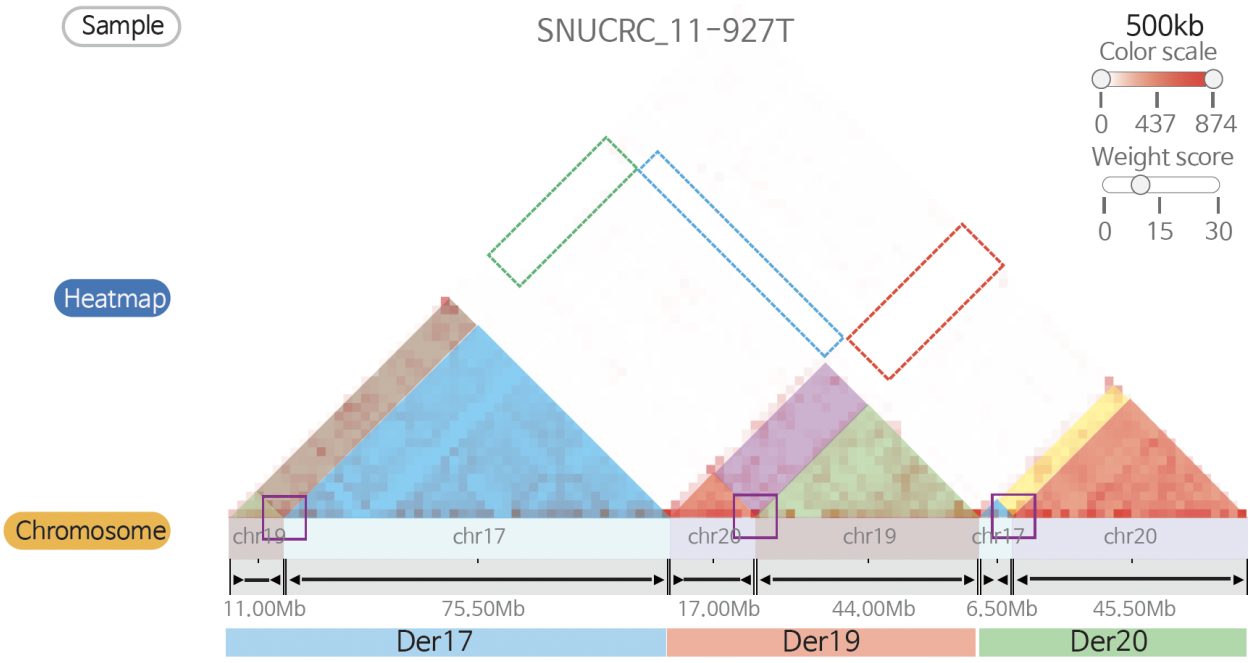


Module III. Complex SV and 3D genome



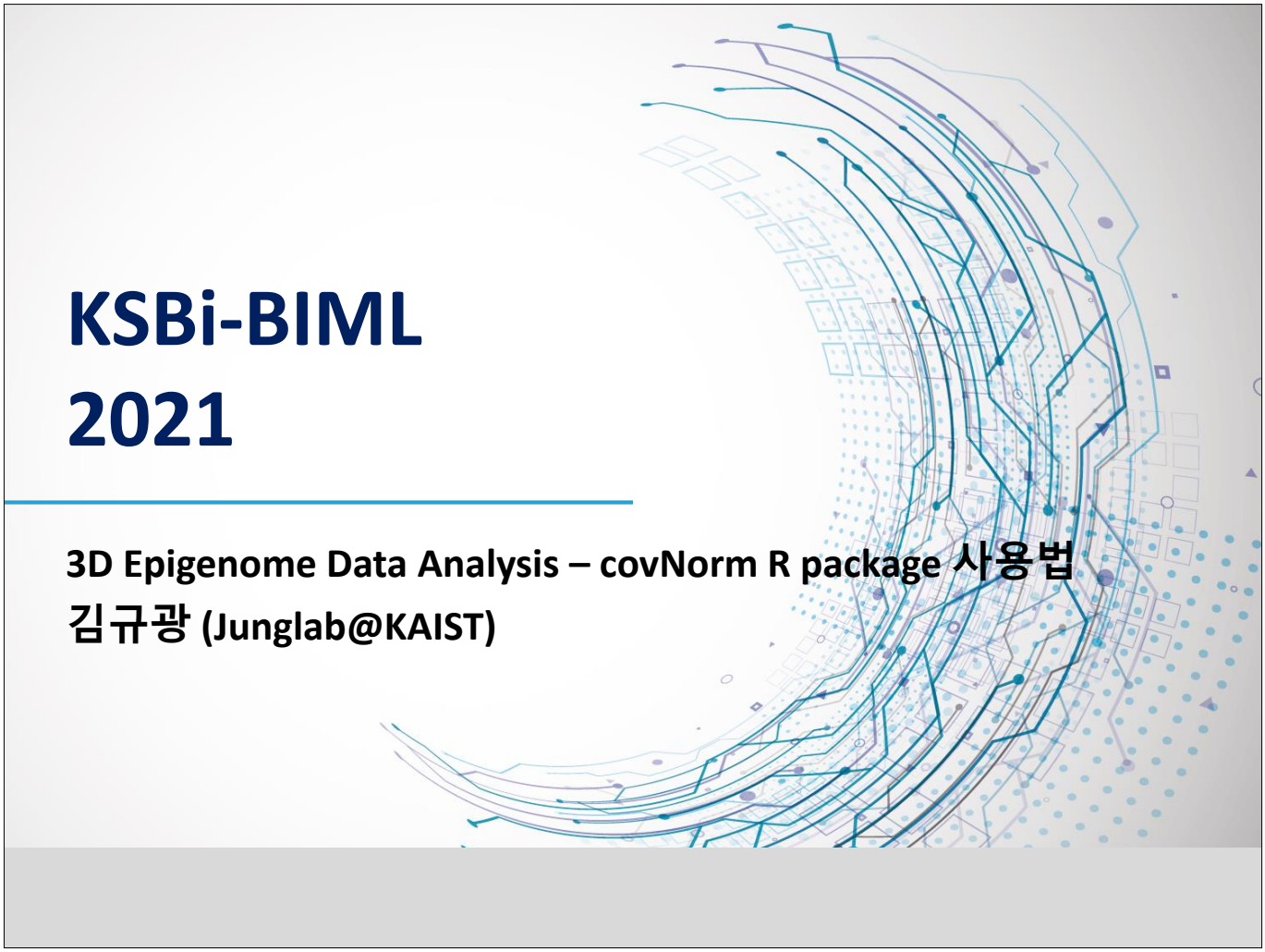


Module III. Complex SV and 3D genome



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



Contents

1. Package installation: Prerequisite

- R (covNorm 패키지 구동) 과 Python (Hi-C contact map 시각화) 필요

- Anaconda2 (5.2.0) 의 Python 2.7.15 와 R 3.4.3 으로 시연. R의 경우 anaconda 설치 후 conda 명령어를 사용하여 설치

1. Package installation: R package

- 설치 방법 1
- covNorm의 필요 R package를 직접 설치. Suggests package의 경우 covNorm 구동에 필수는 아니나 QC용 plot 생성 등 부가 기능에 필요함.

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

- GitHub에서 covNorm 소스 다운로드 후 R CMD로 빌드-설치. INSTALL 시 버전의 경우 추후 covNorm 버전 변경에 따라 변동 가능. 현재는 1.0.0 버전.

```
git clone https://github.com/kaistcbfg/covNormRpkg.git
R CMD build covNormRpkg
R CMD INSTALL covNormRpkg_1.0.0.tar.gz
```

- 설치 후 R에서 명령어 library(covNormRpkg) 실행으로 정상 설치 확인. 에러 메시지 없이 해당 코드가 실행되어야 함

5

1. Package installation: R package

- 설치 방법 2
- R devtools package가 있는 경우 Github 주소 입력으로 설치 가능. devtools 버전에 따른 패키지 의존성 자동 설치 주의.

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

- 설치 후 R에서 명령어 library(covNormRpkg) 실행으로 정상 설치 확인. 에러 메시지 없이 해당 코드가 실행되어야 함

6

1. Package installation: Python package

- Numpy (array 자료구조) 와 Matplotlib (시각화), gzip (압축) 패키지 필요.
- Anaconda로 Python 설치 시 해당 패키지들 기본 제공. miniconda 나 다른 python이 설치된 경우 해당 패키지들 설치. (pip install, conda install 사용)
- python 쉘에서 import numpy, import matplotlib, import gzip 실행으로 설치 테스트

2. Input feature file

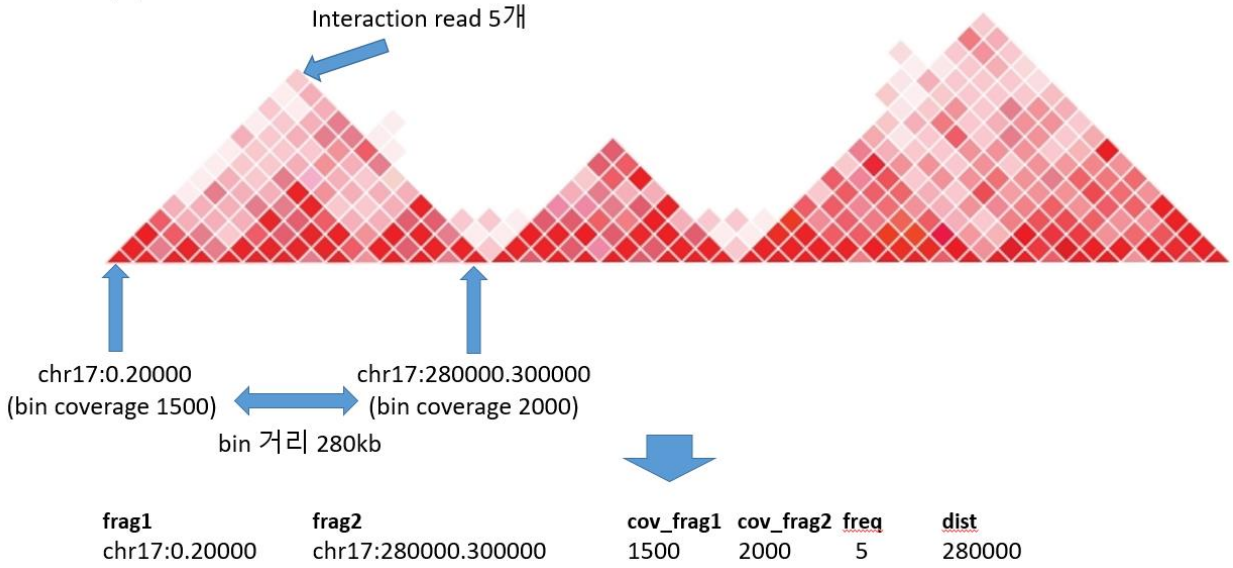
- covNorm 입력을 위해서는 아래 table과 같은 format의 파일 (gzipped) 필요.
- Row 1은 column name (table과 동일해야 함)
- frag1, frag2 는 interaction하는 두 bin
- cov_frag1, cov_frag2 는 frag1과 frag2 bin의 coverage
- freq는 두 bin의 interaction frequency (bam 파일의 read 수)
- dist는 두 bin의 genomic distance의 절대값

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

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

2. Input feature file

Feature 예시



frag1	frag2	cov_frag1	cov_frag2	freq	dist
chr17:0.20000	chr17:280000.300000	1500	2000	5	280000

2. Input feature file

- 각 bin의 coverage는 bedtools의 coverageBed 를 사용하여 계산. https://bedtools.readthedocs.io/en/latest/content/tools/coverage.html
- 큰 bam 파일 모든 bin의 coverage 계산은 시간이 오래 필요함. Tutorial 용으로 GM19204 cell line의 chromosome 17 intra-chromosomal Hi-C interaction feature파일 제공 (40kb bin). http://junglab.kaist.ac.kr/Dataset/GM19204.chr17.cis.feature.gz 리눅스의 경우 wget으로 바로 다운로드 가능
- Resolution이나 depth가 높은 경우 다수의 interaction이 기록되어 텍스트 파일 크기 증가. 압축 효율이 좋으므로 gzip 포맷으로 입출력 사용 권장.
- 리눅스의 경우 zcat 과 less 사용 시 텍스트 파일처럼 gzip 파일 열람 가능.

3. Normalization

- 예시 feature 파일을 다운받은 디렉토리에서 R 실행 후 해당 명령어 입력. 디렉토리가 다를 경우 file_name 변수에 feature 파일 경로 입력

```
library('covNormRpkg')

file_name = "GM19204.chr17.cis.feature.gz"
raw_data <- read.table(gzfile(file_name),header=TRUE)
head(raw_data)
```

- read.table 함수를 사용해서 raw_data 변수에 feature 파일을 data.frame 형식으로 loading. head() 함수를 사용하여 정상적으로 loading 되었는지 확인

```
> head(raw_data)
                  frag1                  frag2 cov_frag1 cov_frag2 freq
1 chr17.82760000.82800000 chr17.77680000.77720000       519      1273    1
2 chr17.12240000.12280000 chr17.53000000.53040000      1078       998    1
3 chr17.13200000.13240000 chr17.61160000.61200000      1090      1390    2
4 chr17.66920000.66960000 chr17.71040000.71080000      1295      1428    1
5 chr17.47320000.47360000 chr17.45400000.45440000      1226       529    1
6 chr17.35480000.35520000 chr17.35280000.35320000      1413      1677   48
      dist
1  5080000
2 40760000
3 47960000
4  4120000
5  1920000
6   200000
>
```

11

3. Normalization

- covNorm의 경우 (1) intra-chromsome interaction간의 normalization, (2) zero contact bin 없음, (3) self ligation read filtered 를 전제로 구동.

- 만약 사전 제거되지 않은 경우 covNormRpkg::filterInputDF 함수로 data.frame에서 해당 데이터 제거 가능

- 함수 실행 시 default 옵션으로 **frag1**, **frag2**의 chromosome이 다른 row, **freq**가 0인 row, **dist** 가 15kb 이내 인 row 제거.

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

12

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)
```

- coverage가 너무 낮은 bin 의 경우 값을 지정해서 row 제거 가능. default 200.

- 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
```

- 결과 저장 (outfilename 에 적절한 이름 사용)

```
write.table(cov_df, file=gzfile("outFileName1"), row.names=FALSE, col.names=TRUE, sep="\t", quote=FALSE)
```

13

3. Normalization

- coverage Normalization 결과 해석: cov_df의 경우 feature data.frame 뒤에 3개 컬럼 추가

- **rand**의 경우 난수. coverage 1과 2의 shuffle에 사용. **exp_value_capture**의 경우 두 bin의 coverage를 고려한 경우 interaction frequency의 기대값. **capture_res**의 경우 residual. **freq**를 **exp_value_capture**로 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

14

3. Normalization

- coverage Normalization 결과 QC
- (1) 각 bin당 coverage (cov_frag1, cov_frag2)에 dependent 한 결과가 나오면 안됨 (fitting 에러). coeff_cov1과 coeff_cov2를 확인하여 비슷한 값을 가지는지 체크.
- (2) 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. (df 대신 cov_df 사용)

```
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)
```

- coverage Normalize 후 PCC 감소한 경우 정상적으로 normalized 됨.

15

3. Normalization

- coverage Normalization 결과 QC
- covNorm R package에서는 normalization 전후 결과 시각화 지원

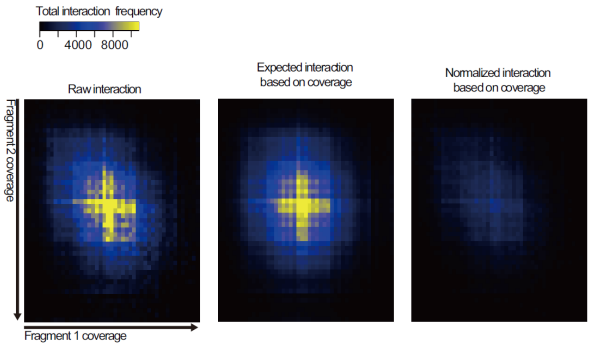
```
covNormRpkg::checkFreqCovPCC(cov_df, outpdfname='QCplot_coverage_PCC.pdf')
covNormRpkg::plotCovNormRes( cov_df, outpdfname='QCplot_coverage_heatmap.pdf')
```

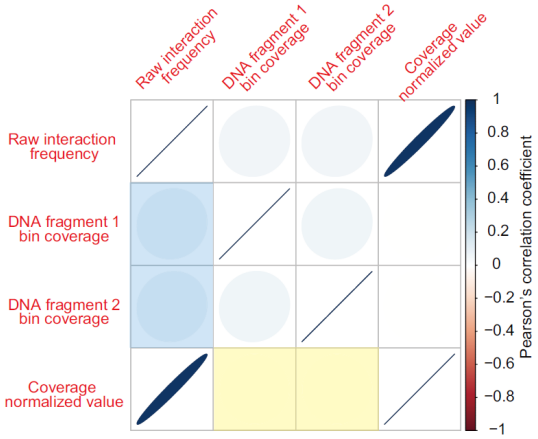
- 해당 함수 실행 시 coverage 대비 pdf 포맷으로 결과 시각화 이미지 생성. 사전 소개된 figure와 동일한 figure 생성 가능

16

3. Normalization

- coverage Normalization 결과 QC
- coverage normalization QC figure





3. Normalization

- coverage normalization 이후 distance Normalization 실행.
- normDistance 함수에 cov_df 입력. max_dist 이내 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
```

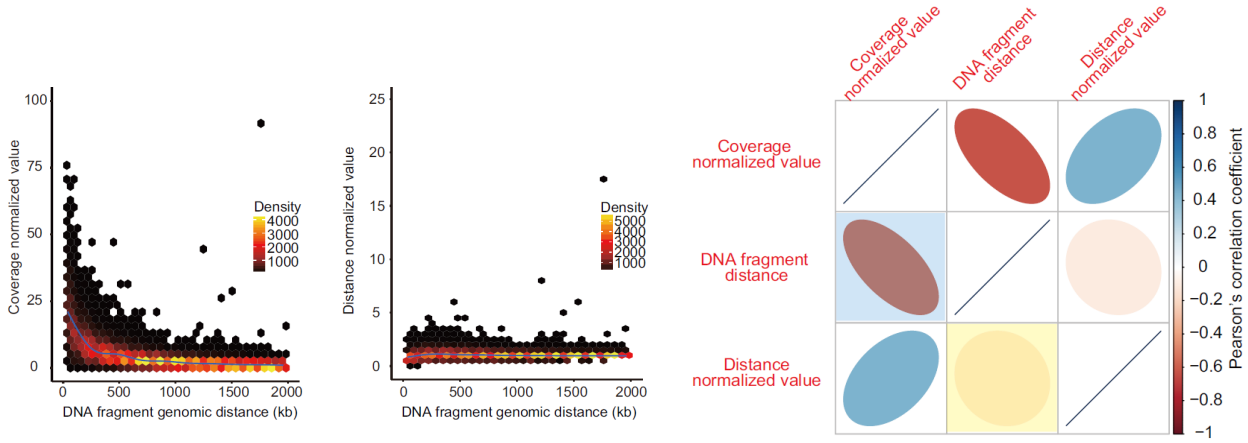
- dist_df의 경우 cov_df의 뒤에 column 2개 추가됨.
- **exp_value_dist**: 해당 **dist**의 기대값, **dist_res**: normalized frequency

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

3. Normalization

- distance Normalization 의 QC.
- distance 대비 correlation의 '감소' (negative correlation이므로 절대값의 감소 체크) 로 QC. coverage QC와 유사한 시각화 기능 제공

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



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

```
final_df <- covNormRpkg::contactPval(dist_df)
```

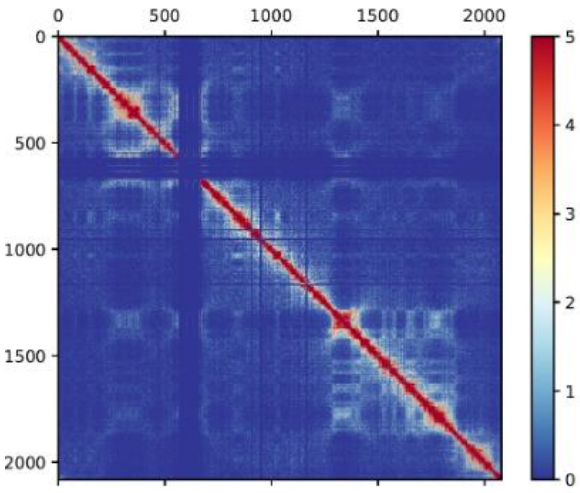
- final_df의 경우 dist_df 뒤에 p-value와 FDR 컬럼 추가됨.

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

4. Visualization

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

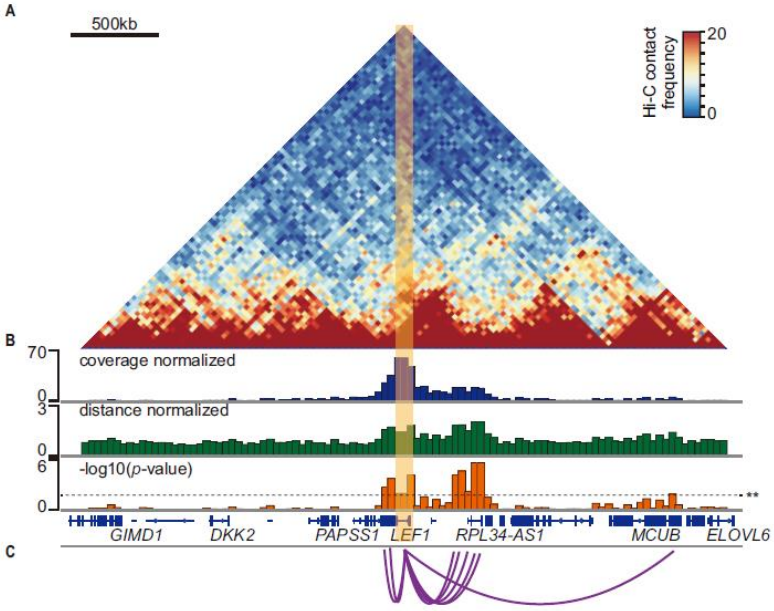
```
import numpy as np
import matplotlib
matplotlib.use('Agg')
import matplotlib.pyplot as plt
import gzip

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)
```



21

4. Visualization

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



22

5. Custom parameter

- Normalization, 시각화에 사용되는 함수들의 경우 다양한 parameter 를 입력 가능.

- Experiment type (pcHi-C 등) 나 sequencing depth 등 다양한 요인에 의해 filtering의 적정 threshold, visualization range 등이 달라질 수 있음.

- Github에서 함수 선언 확인 시 입력 parameter/default 옵션 확인 가능. 해당 값 변경 가능

```
normCoverage <- function(df, do_shuffle=TRUE, cov1_thresh=200, cov2_thresh=200, max_covnorm_value=50, sample_ratio=-1)
```

- 사용 예) pcHi-C 데이터에서 Promoter-other interaction 분석 시 coverage shuffle 사용 안함/차등 coverage filter threshold (200, 50) 적용

```
cov_result <- covNormRpkg::normCoverage(raw_data_filter, do_shuffle=FALSE, cov1_thresh=200, cov2_thresh=50)
#in pcHi-C PO-interaction normalization, lower coverage threshold for 'other' interaction
```

23

5. Custom parameter

- 사용 예 2) downsampling option 사용. depth가 높아 파일 사이즈가 너무 큰 경우 random downsampling 하여 fitting 가능. **적정 수치** downsampling 시 fitting 결과가 전체 데이터 사용시와 크게 다르지 않으나 시간은 줄어들음.

- sample ratio 옵션에 0~1.0 사이 값을 입력해서 사용. 전체의 50%를 가지고 fitting 진행 시 해당 파라미터로 0.5 설정.

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

- distance normalization에도 downsample 기능 있음.

24

KSBi-BIML 2021

covNorm R package 사용법. The end