Having Fun in Hi-C Data Analysis

Fengling Chen
Supervisor: Michael Q. Zhang & Yang Chen

MOE Key Laboratory of Bioinformatics
Bioinformatics Division and Center for Synthetic and Systems Biology, BNRist
Department of Automation, Tsinghua University

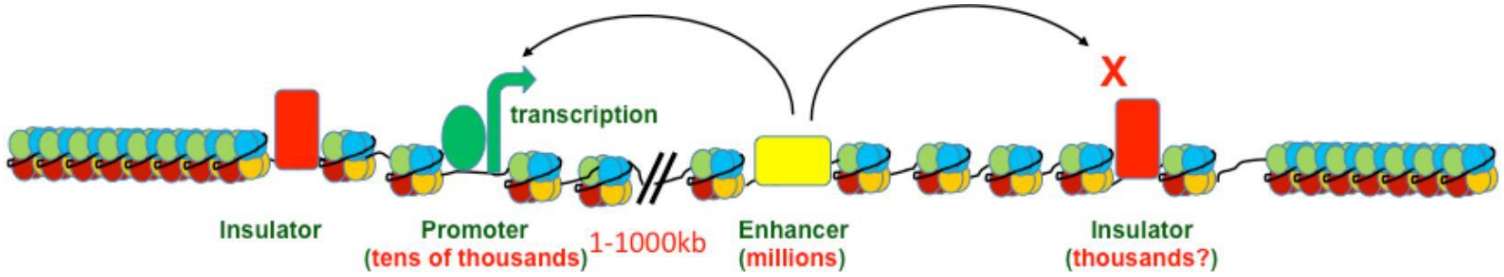
5th January 2019

Background

ROADMAP epigenomics PROJECT — Integrative analysis of 111 reference human epigenomes

ENCODE

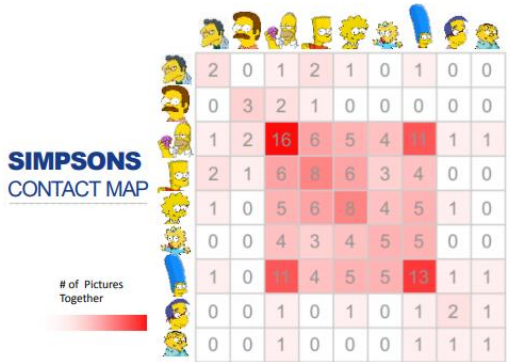
Human genome: 20,000 genes ; 100,000 promoters ;
500,000 enhancers; 5,000,000 regulatory elements

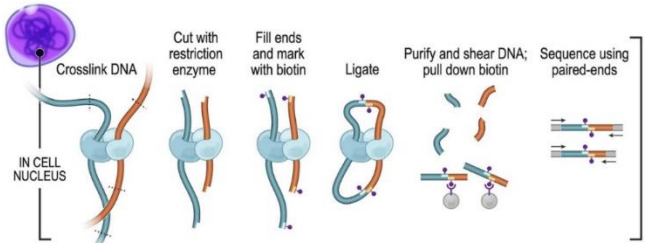


In development and disease，how these elements interact to regulate gene expression ?

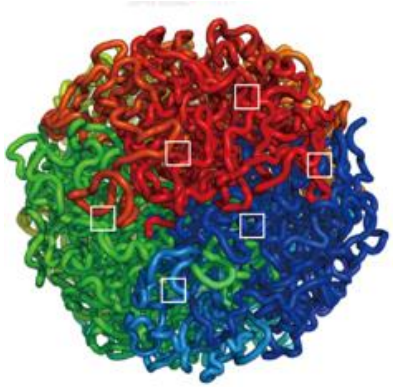
Roadmap Epigenome Consortium, Nature, 2015;
Organization and Regulation of Human Genome, Bing Ren, ENCODE 2016 Meeting, 2016

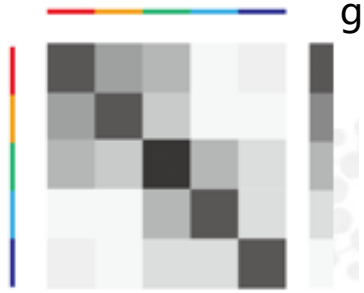
Background





Hi-C method





Contact matrix (pairwise)

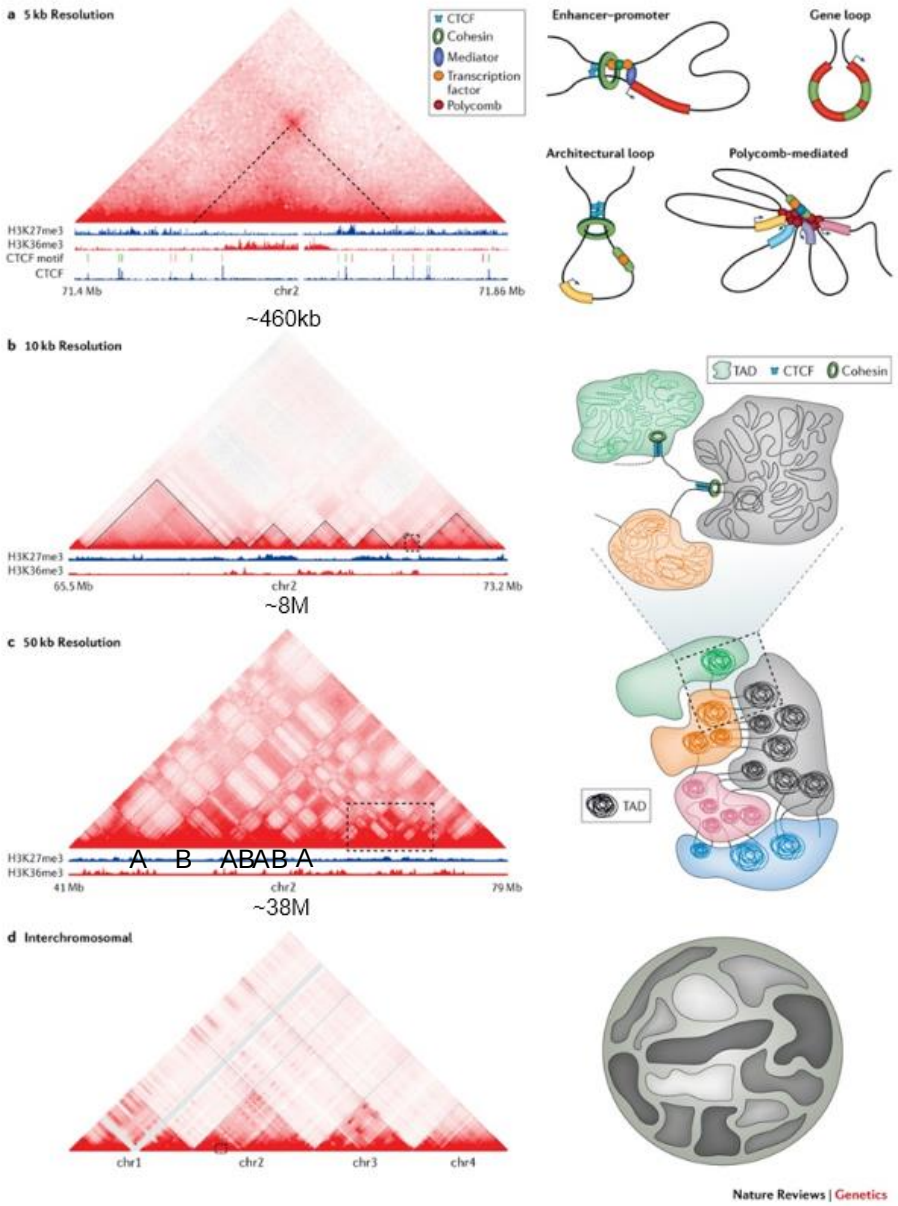
A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Loopinget al., Cell, 2014

Background

1. Hierarchical folding of chromatin
Chromatin territory
Compartmentalization of megabase-scale chromatin.
Topologically associating domains;
Chromatin loops

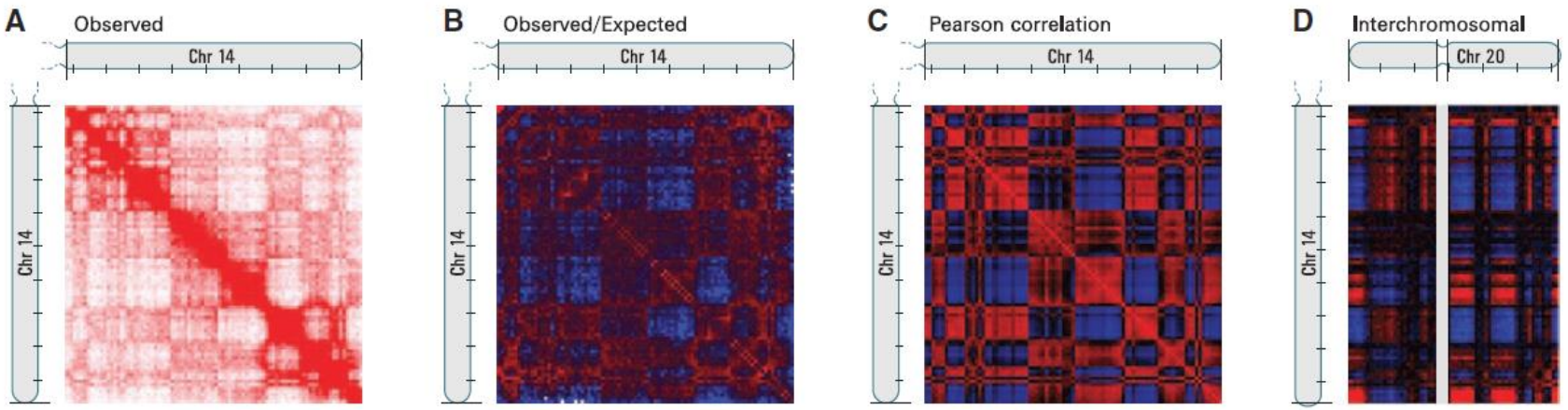
2. Mechanisms to organize chromatin in 3D
Architectural proteins: CTCF; cohesion; mediator
Non-coding RNAs
Histone modification
Phase separation

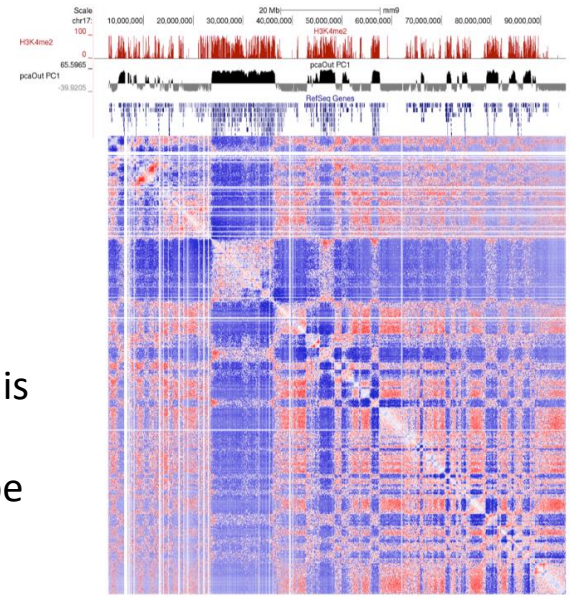
How to extract information on Hi-C maps ?



Organization and function of the 3D genome. Nature Reviews Genetics. 2016

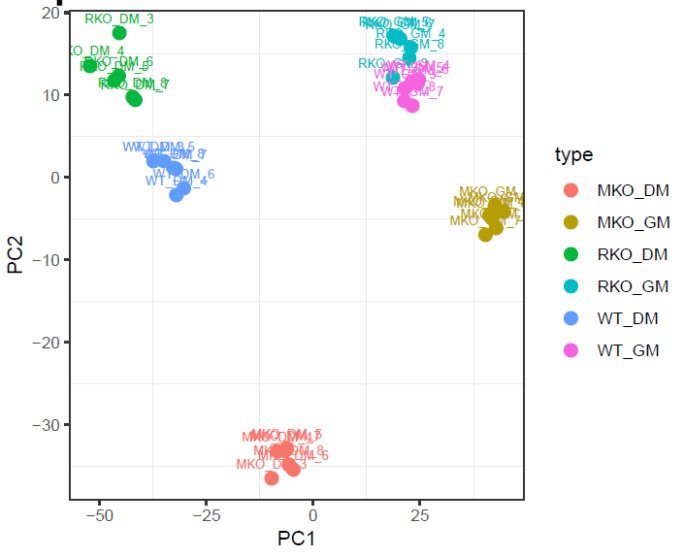
1. Compartment



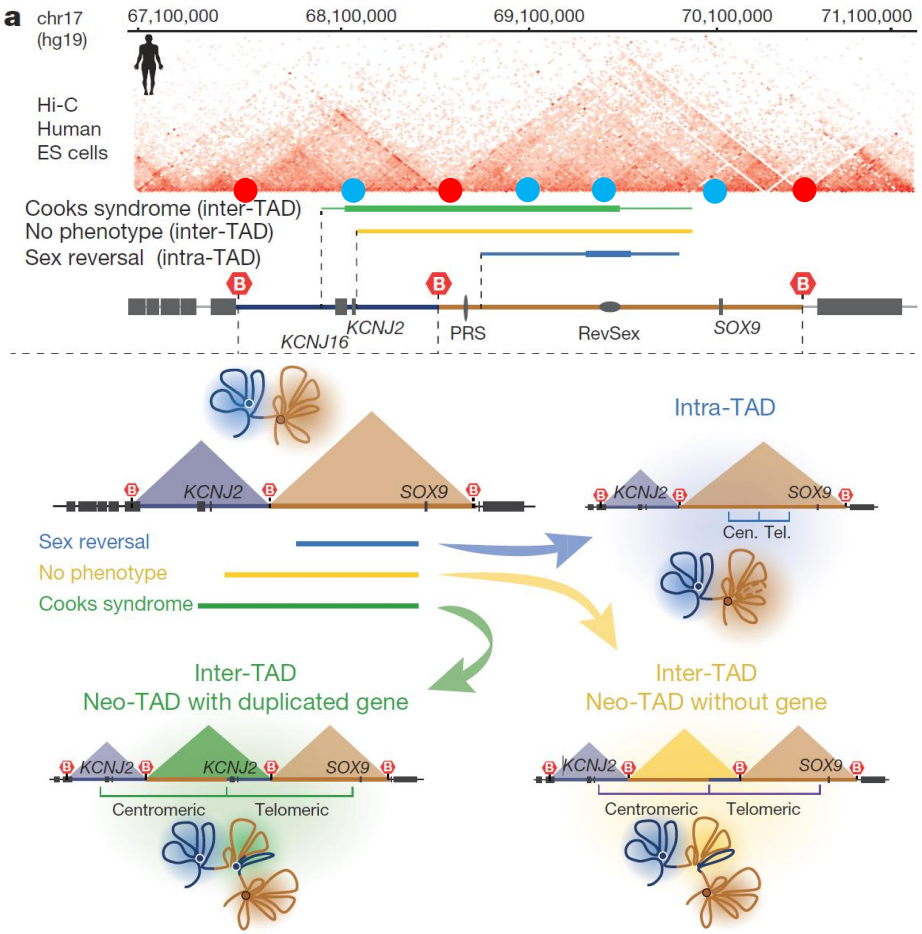


The basic idea behind **PCA** is to redefine the coordinate system such the data can be "described" with as few dimensions as possible.

PCA of PC1 values in muscle development



2. CDB



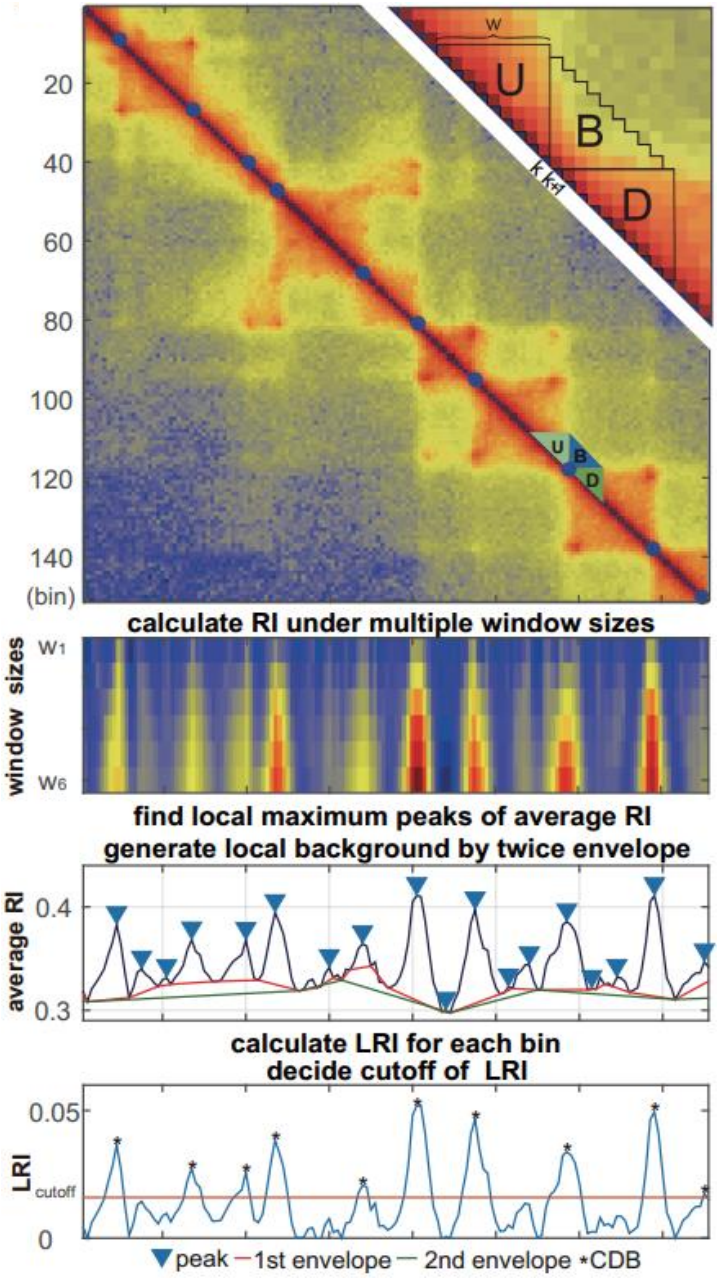
contact domain boundaries (CDBs) includes TAD boundaries and sub-TAD boundaries.

TADs: invariant/conserved; absolute insulated

sub-TADs : varied ; cell-type specific gene regulation; relatively insulated

Formation of new chromatin domains determines pathogenicity of genomic duplications. Franke et al. Nature. 2016

HiCDB Method



✓ Calculate relative insulation (RI) under different window size

$$RI(w,s) = \frac{U(w,s) + D(w,s) - B(w,s)}{U(w,s) + D(w,s) + B(w,s)}$$

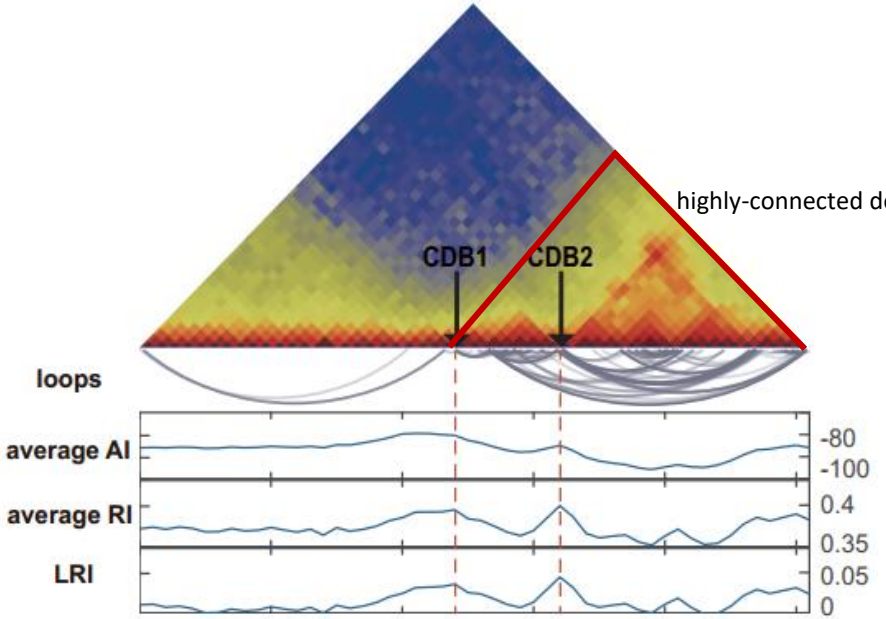
✓ Peak detection on average RI

$$\overline{RI}(s) = \frac{1}{w_n - w_1} \sum_{w=w_1}^{w_n} RI(w,s)$$

✓ Remove background

$$LRI(s) = \overline{RI}(s) - lower\_envelope(lower\_envelope(\overline{RI}(s)))$$

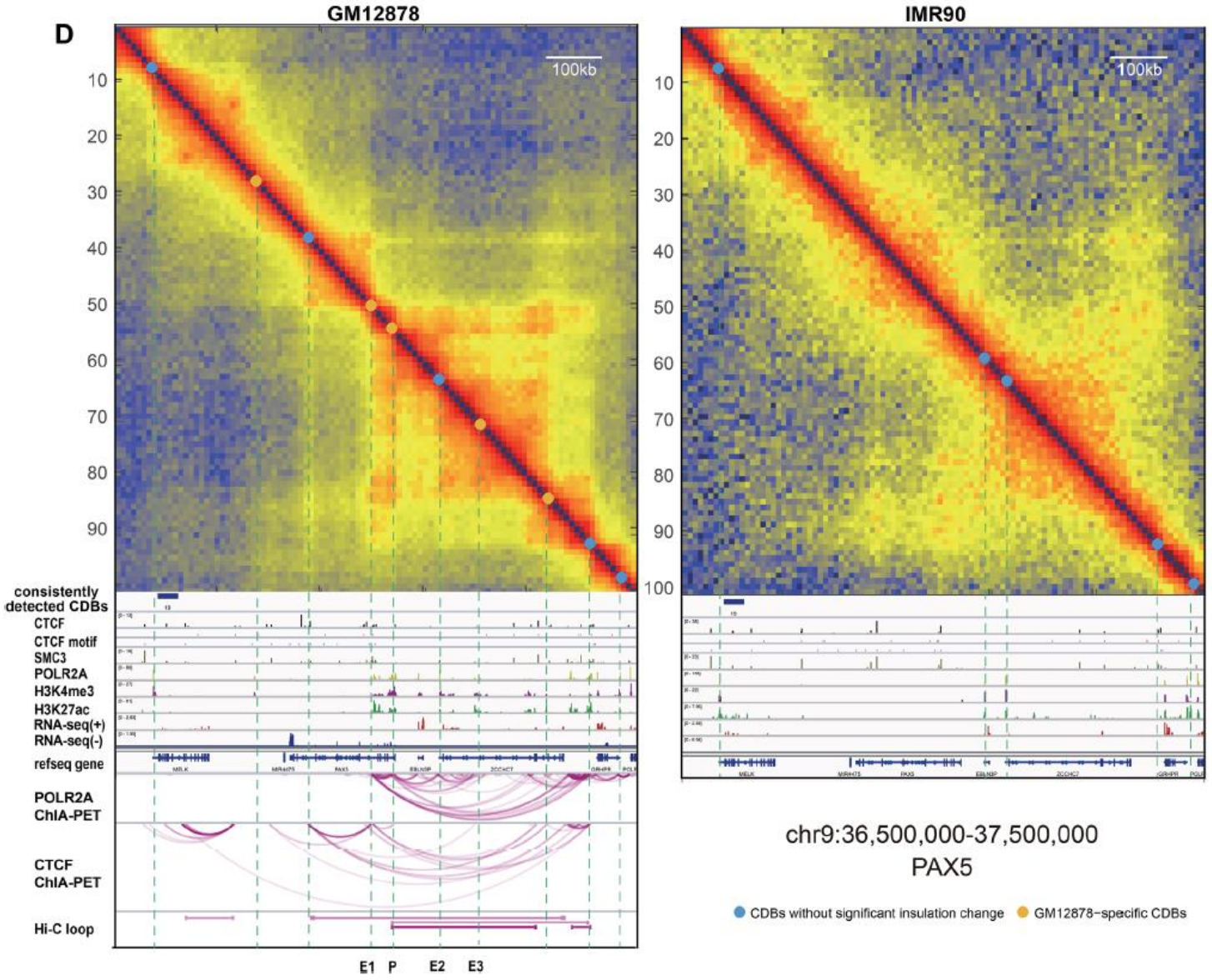
✓ Cut-off chosen (GSEA-like method)



aRI or LRI can be compared genome-wide

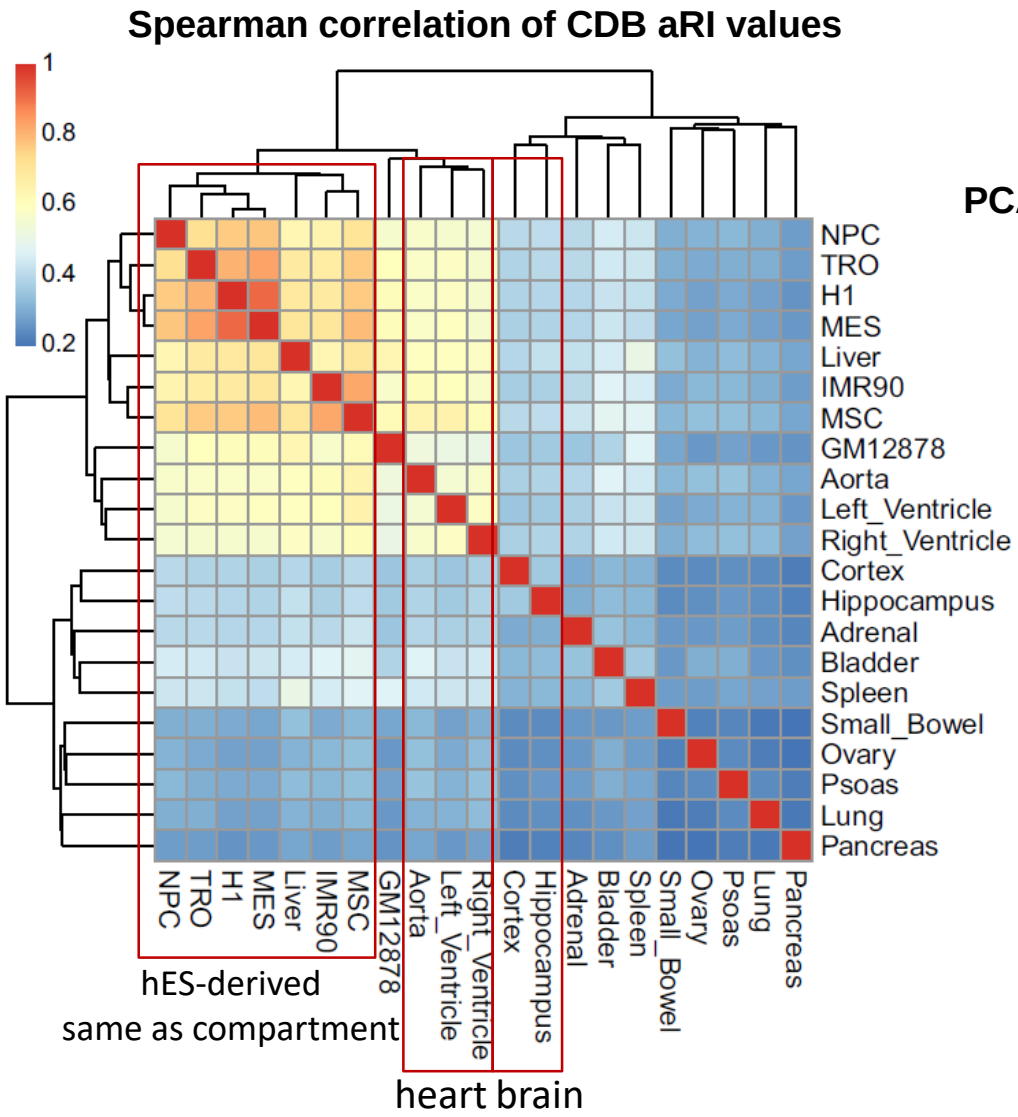
aRI or LRI help to find CDBs under highly-connected domains.

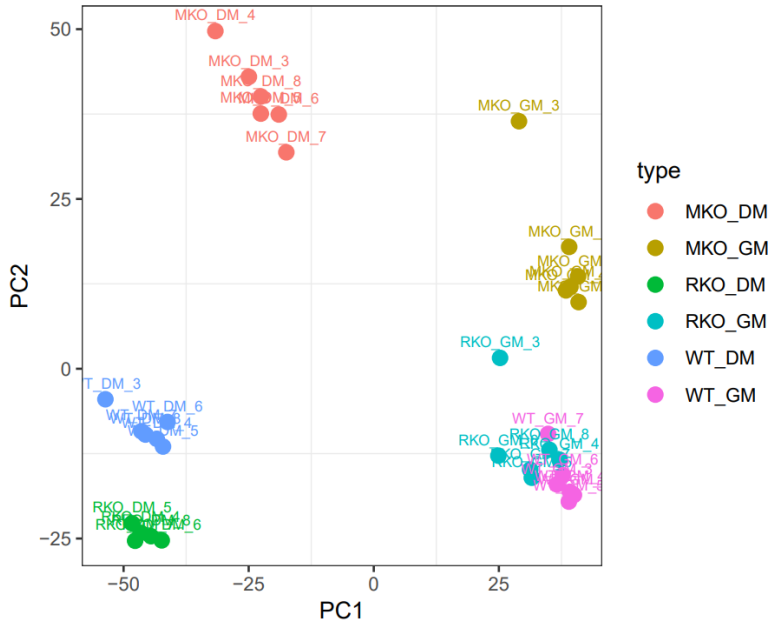
Differential CDB example



Cell-type-specific CDBs correlate with cell-type-specific histone modification and gene activation.

Clustering based on aRI values





aRI score is more sensitive to sequencing depth than compartment.

A compendium of chromatin contact maps reveals spatially active regions in the human genome[J]. Schmitt A D, et al. Cell reports, 2016.

3. Loop Calling

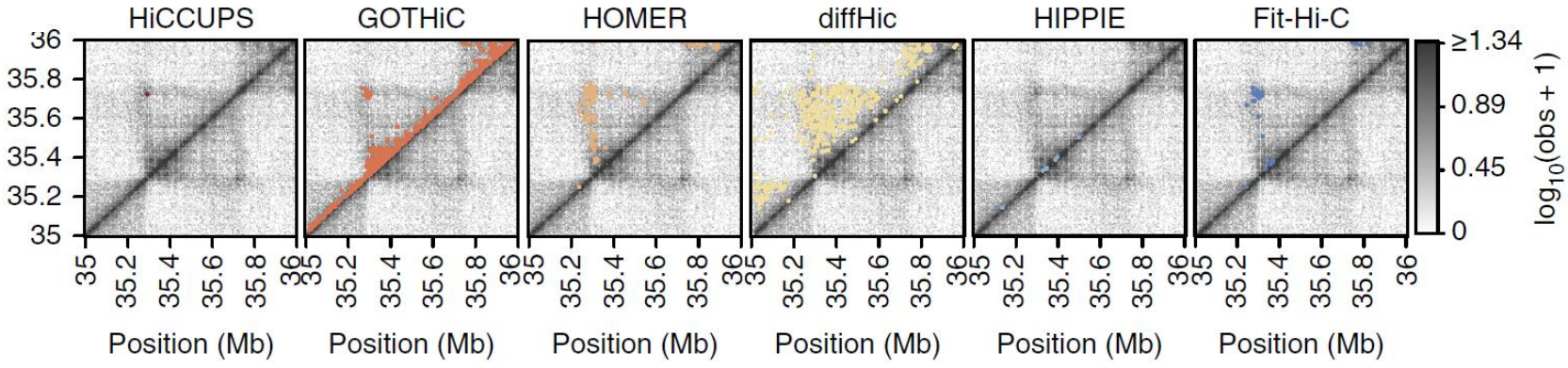
Table 3 Methods of identifying significant interactions

表 3 显著交互作用鉴别方法

算法或软件	模型	针对性	特点
HIPPIE[42]	负二项分布	显著交互作用	在酶切片段分辨率下，鉴别近距离交互作用
Fit-HiC[80]	二项分布	显著交互作用	对交互作用与距离的关系两次拟合
GOTHiC[81]	二项分布	显著交互作用	利用覆盖率估计期望的交互作用
HiC-DC[82]	零截尾负二项分布	显著交互作用	模型考虑了 Hi-C 数据的多零和高散度的特性
HOMER[43]	二项分布	显著或差异交互作用	多功能集成软件
HMRFBayes[83]& FastHiC[84]	负二项分布、隐 Markov 随机场	显著交互作用	考虑相邻交互作用的影响
HiCCUPS[31]	泊松过程	显著交互作用	去除 TAD 结构的影响
PSYCHIC[85]	对数正态分布	显著交互作用	基于 TAD 结构构建背景模型
CHiCAGO[86]	负二项方分布、泊松分布	显著交互作用	针对 Capture Hi-C 实验数据
Dynamic Interactions[13]	二项分布	差异交互作用	利用生物学重复
HiBrowse[87] & DiffHiC[88]	负二项分布	差异交互作用	借助 edgeR，利用生物学重复
FIND[89]	空间泊松过程	差异交互作用	考虑相邻交互作用的关联性

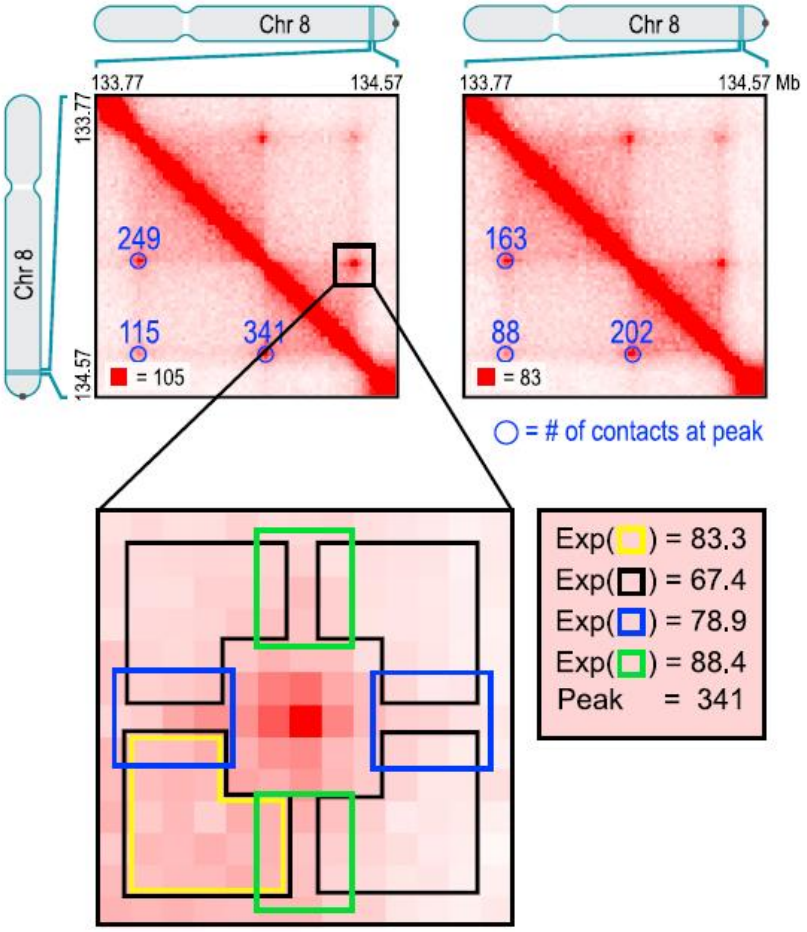
local background model

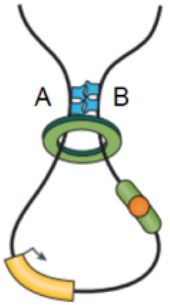
Global background model



3. Loop Calling

Local background model





Step1. build background model.

$$donut\ filter:\ E^{d*}_{i,j}=\frac{\sum_{a=i-w}^{i+w}\sum_{b=j-w}^{j+w}M^*_{a,b}-\sum_{a=i-p}^{i+p}\sum_{b=j-p}^{j+p}M^*_{a,b}-\sum_{a=i-w}^{i-p-1}M^*_{a,j}-\sum_{a=i+p+1}^{i+w}M^*_{a,j}-\sum_{b=j-w}^{j-p-1}M^*_{i,b}-\sum_{b=j+p+1}^{j+w}M^*_{i,b}}{\sum_{a=i-w}^{i+w}\sum_{b=j-w}^{j+w}E^*_{a,b}-\sum_{a=i-p}^{i+p}\sum_{b=j-p}^{j+p}E^*_{a,b}-\sum_{a=i-w}^{i-p-1}E^*_{a,j}-\sum_{a=i+p+1}^{i+w}E^*_{a,j}-\sum_{b=j-w}^{j-p-1}E^*_{i,b}-\sum_{b=j+p+1}^{j+w}E^*_{i,b}}\times E^*_{i,j}$$

$$lower\ left\ filter:\ E^{ll*}_{i,j}=\frac{\sum_{a=i+1}^{i+w}\sum_{b=j-w}^{j-1}M^*_{a,b}-\sum_{a=i+1}^{i+p}\sum_{b=j-p}^{j-1}M^*_{a,b}}{\sum_{a=i+1}^{i+w}\sum_{b=j-w}^{j-1}E^*_{a,b}-\sum_{a=i+1}^{i+p}\sum_{b=j-p}^{j-1}E^*_{a,b}}\times E^*_{i,j}$$

$$vertical\ filter:\ E^{v*}_{i,j}=\frac{\sum_{a=i-w}^{i-p-1}\sum_{b=j-1}^{j+1}M^*_{a,b}-\sum_{a=i+p+1}^{i+w}\sum_{b=j-1}^{j+1}M^*_{a,b}}{\sum_{a=i-w}^{i-p-1}\sum_{b=j-1}^{j+1}E^*_{a,b}-\sum_{a=i+p+1}^{i+w}\sum_{b=j-1}^{j+1}E^*_{a,b}}\times E^*_{i,j}$$

$$horizontal\ filter:\ E^{h*}_{i,j}=\frac{\sum_{b=j-w}^{j-p-1}\sum_{a=i-1}^{i+1}M^*_{a,b}-\sum_{b=j+p+1}^{j+w}\sum_{a=i-1}^{i+1}M^*_{a,b}}{\sum_{b=j-w}^{j-p-1}\sum_{a=i-1}^{i+1}E^*_{a,b}-\sum_{b=j+p+1}^{j+w}\sum_{a=i-1}^{i+1}E^*_{a,b}}\times E^*_{i,j}$$

Step2. calculate p-value ,fold and FDR. (FDR was distance related)

Step3. calculate enriched pixels.

Step4. merge pixels nearby.

a Poisson process with parameter λ= E_local×C'i ×C'j.

4. Differential Loop Detection

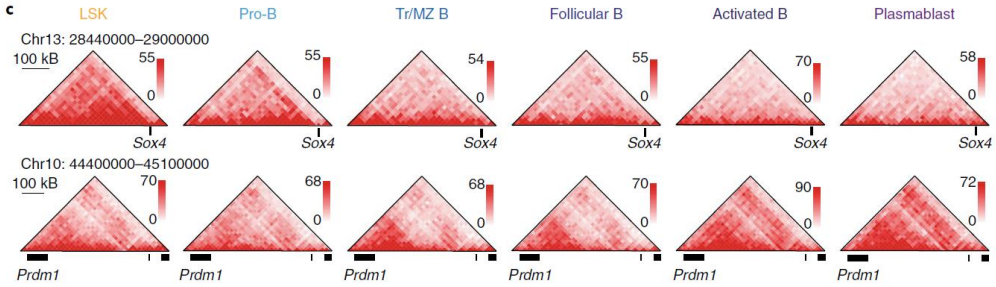
Differential interaction: test for each pixel (diffHiC, HiCcompare);
test for each pixel considering neighborhood (FIND)
Differential loop: test for each loop considering neighborhood

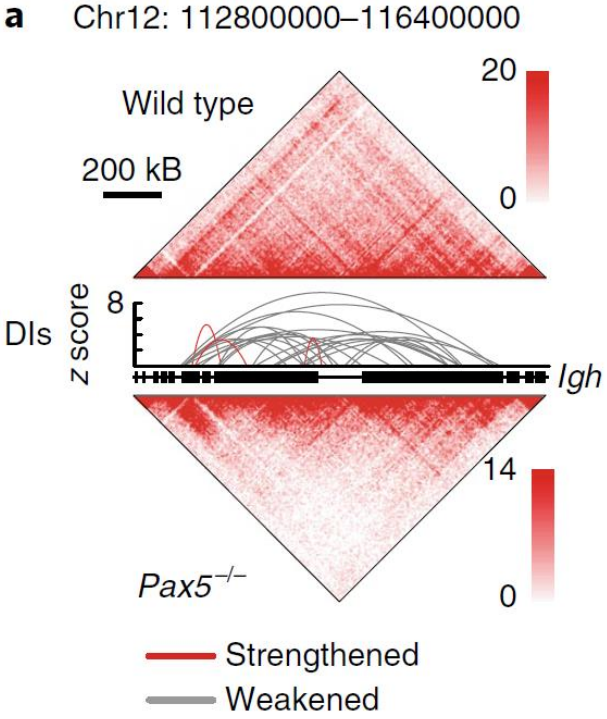
nature immunology
ARTICLES
https://doi.org/10.1038/s41590-018-0234-8

Transcription-factor-mediated supervision of global genome architecture maintains B cell identity

Timothy M. Johanson[1,2,4], Aaron T. L. Lun[1,2,4], Hannah D. Coughlan[1,2,4], Tania Tan[1,2], Gordon K. Smyth[1,3], Stephen L. Nutt[1,2,5*] and Rhys S. Allan[1,2,5*]

Recent studies have elucidated cell-lineage-specific three-dimensional genome organization; however, how such specific architecture is established or maintained is unclear. We hypothesized that lineage-defining transcription factors maintain cell identity via global control of genome organization. These factors bind many genomic sites outside of the genes that they directly regulate and thus are potentially implicated in three-dimensional genome organization. Using chromosome-conformation-capture techniques, we show that the transcription factor Paired box 5 (Pax5) is critical for the establishment and maintenance of the global lineage-specific architecture of B cells. Pax5 was found to supervise genome architecture throughout B cell differentiation, until the plasmablast stage, in which Pax5 is naturally silenced and B cell–specific genome structure is lost. Crucially, Pax5 did not rely on ongoing transcription to organize the genome. These results implicate sequence-specific DNA-binding proteins in global genome organization to establish and maintain lineage fidelity.





50kb diffHiC

4. Differential Loop Detection

DEseq2

Read count matrix K_{ij} (gene i, sample j)

Design matrix X_{jr} (sample j, condition r)

- $K_{ij} \sim NB(\mu_{ij}, \alpha_{ij})$
  - NB generalizes Poisson, mean: μ variance: $\mu + \alpha\mu^2$
  - α: dispersion parameter
- $\mu_{ij} = s_j q_{ij}$
  - $s_j = \underset{i:K_i^R \neq 0}{\text{median}} \frac{K_{ij}}{K_i^R}, K_i^R = \left(\sum_{j=1}^{m} K_{ij}\right)^{\frac{1}{m}}$
- $\log(q_{ij}) = \sum_r X_{jr}\beta_{ir}$
  - β: log fold change (LFC)

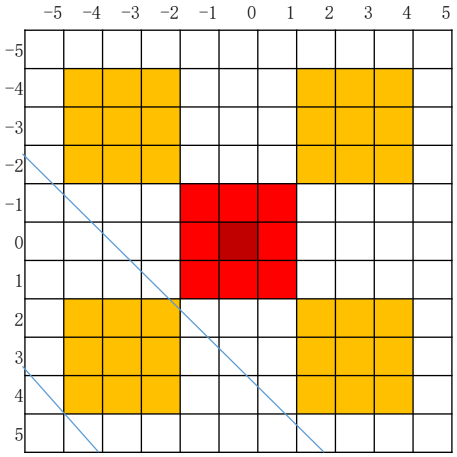
Null hypothesis: Beta=0 (gene is not differential as gene does not change when condition changes)

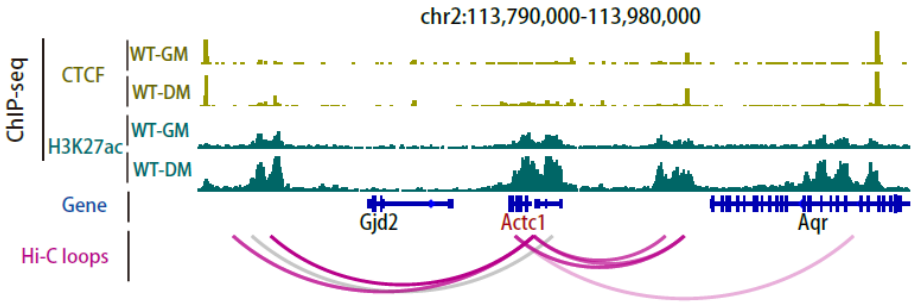
$$\frac{u_{ij} \times norm1 \times norm2}{norm\_expected} = q_{ij}$$

$$Sij = \frac{norm\_expected}{norm1 \times norm2}$$

local	sample	rep
D,P	GM,DM	1,2,3 (GM) 1,2,3 (DM)

design formula ~ rep + sample + rep:sample + local + sample:local

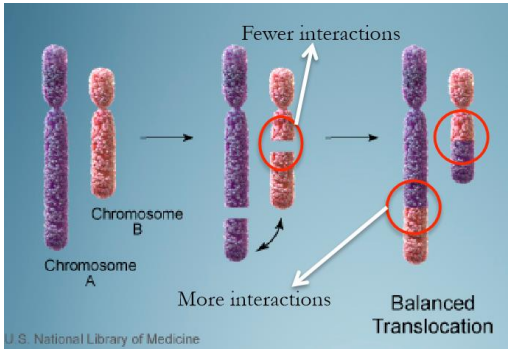


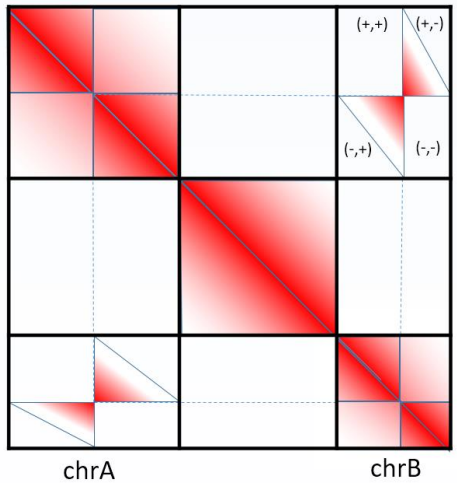


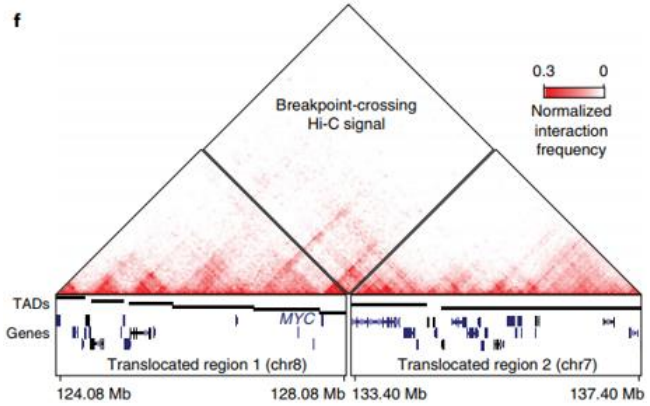
Phanstiel, Douglas H., et al. "Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development." Molecular cell 67.6 (2017): 1037-1048. (10 kb)

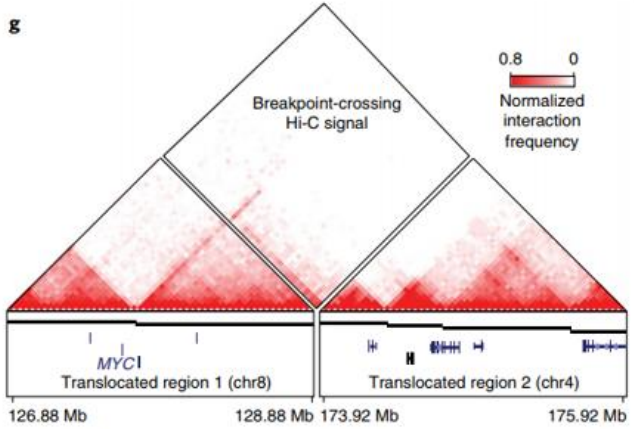
5. Structural Variation

✓20-100M reads (1-5X coverage);
✓complex / large-scale structural variation
✓breakpoint in repeat region





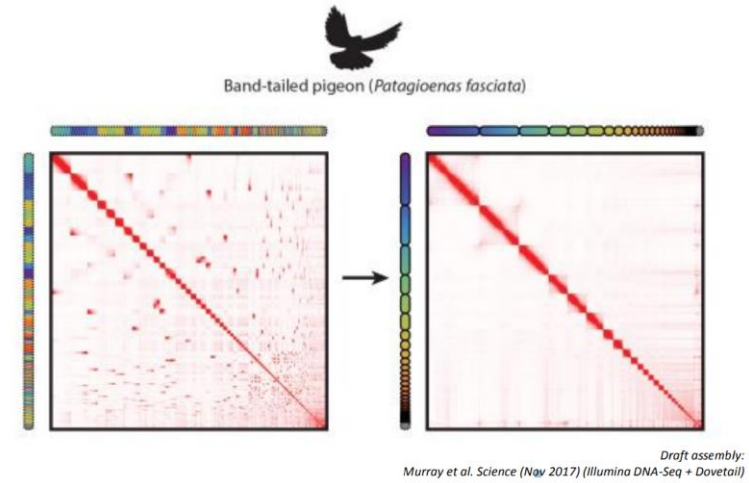


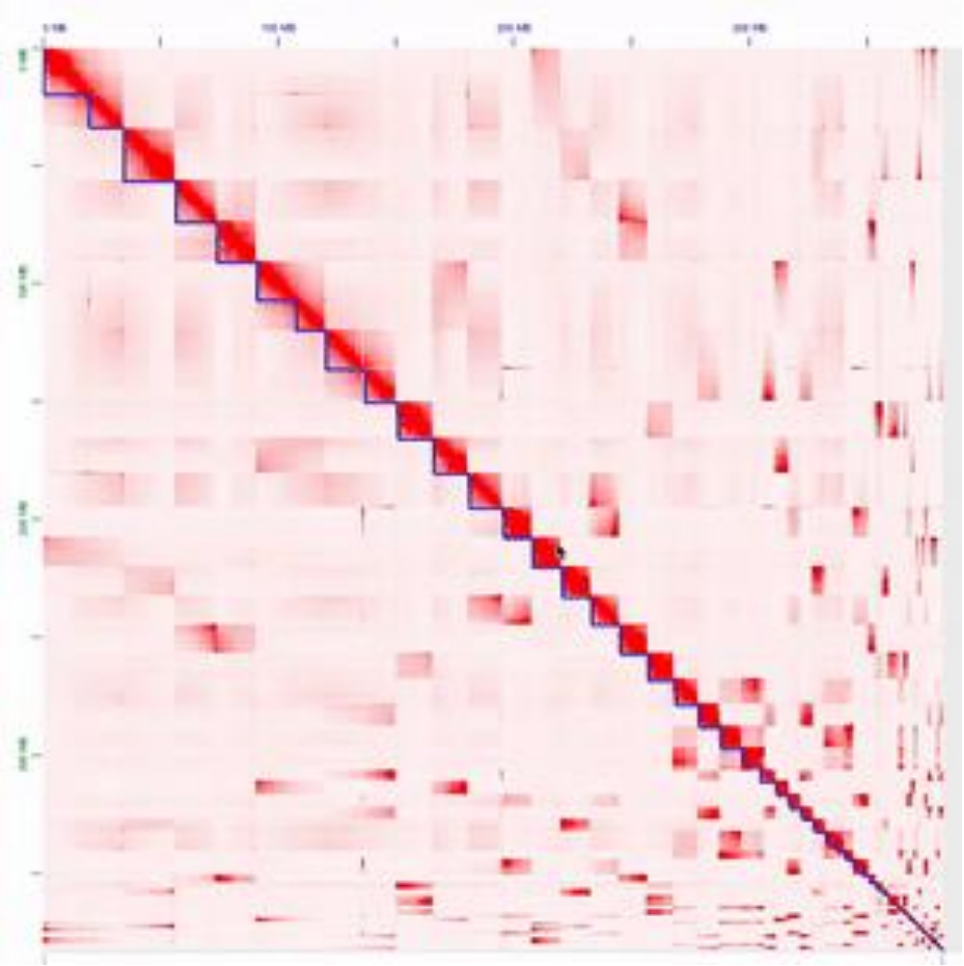


Features local maximal peak in distal regions → NMS
loss of interaction in breakpoint(possible)
right angle with direction → gradient/template matching

In development

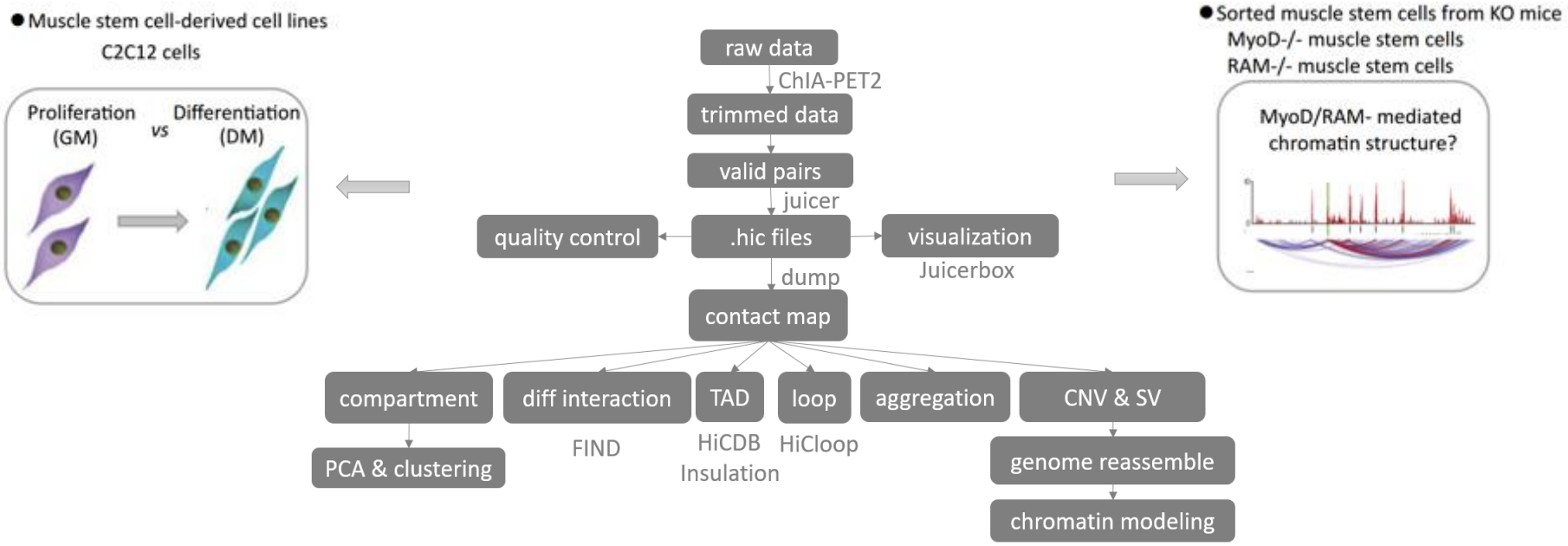
6. Genome Assembly





Take home message

Practice is the sole criterion for testing truth.



efficient; QC; visualization; multiple feature detection

Have fun in exploring biological data !

Michael Q. Zhang Lab @ THU

Thank you!

Fengling Chen 陈凤玲: Ph.D. candidate
Email: cfl15@mails.tsinghua.edu.cn
fchen@github