Genome 4D

Bio5488
Ting Wang
2.19/2018

https://www.nature.com/collections/rsxlmsyslk
Outline

• Motivation for studying 3D genome
• Method to investigate 3D genome organization
• 3D genome organization and human diseases
• 4DN project (https://www.4dnucleome.org)
Identify disease or trait-associated variants by genome-wide association studies (GWAS)
Majority of the GWAS hits are located in the non-coding regions

Hnisz et al, Cell, 2013
Majority of the GWAS hits are located in the cis-regulatory elements

Maurano et al. Science 2012
How non-coding mutations contribute to disease

- Gain of TF binding sites;
- Loss of TF binding sites;
- Enhancer hijacking

How non-coding mutations contribute to disease

How do we find the target genes for distal enhancers?

- Nearest genes by genomic distance;
- Correlation based on gene expression and enhancer activities;
- 3D space proximity;
- CRISPR/Cas9 to delete enhancers and observe their effect on gene expression;
Linking enhancers to target promoters

Gene 1

Enhancer

Gene 2

Bone Marrow
Cerebellum
Cortex
Heart
Kidney
Liver
Lung
Mef
ES
Spleen

Correlation coefficient

Correlation coefficient
Preaxial polydactyly

Enhancer

ZRS

Belgian 2

IV:2 nt

III:1 T/T

III:2 T/T

III:3 T/C

III:4 nt

II:1 nt

II:2 T/C

II:3 T/T

II:4 nt

I:1 nt

I:2 nt

B
3D genome and transcriptional regulation

Aixel, nature, 2009
Regulation at a distance

Obesity-associated variants within FTO form long-range functional connections with IRX3

Scott Smemo*, Juan J. Tenà*, Kyoung-Han Kim*, Eric R. Gamazon*, Flavia L. Credidio†, Débora R. Sobrelta†, Nora F. Wasserman†, Ju Hee Lee†, JoEun Son†, Nikil Alizadeh Vakili†, Hoon-Ki Sung†, Silvia Naranjo†, Raquel J. Cox†, Chi-Chung Hui†, Jose Luis Gomez-Skarmeta† & Marcelo Camargo†
Genome organization in mammals

Global view of the nucleus

Technologies used (and developed) to study genome folding

Can be divided into two broad categories:

1. **Imaging**
   1. Bright-field
   2. Fluorescence
   3. EM
   4. Fluorescence *in-situ* hybridization (FISH), etc.

2. **Genomics**
   1. DamID
   2. ChIA-PET
   3. GAM
   4. Chromosome conformation capture-derived, etc.

Adapted from [http://web.uvic.ca/ail/equipment.html](http://web.uvic.ca/ail/equipment.html)
Overview of chromatin structure and assays at three scales

Principles of nuclear organization revealed using microscopic techniques

From single genes to entire genomes: the search for a function of nuclear organization. Pueschel et al. Development (2016)
Fluorescence in situ hybridization

- Cytogenetic technique

- Uses fluorescent molecules to “paint” (regions of interest) chromosomes in cells often in Metaphase or Interphase

- Aids in analysis of chromosome structure, structural aberrations, ploidy determination, etc.
ADVANTAGES
• Rapid and sensitive
• Lots of cells can be analyzed
• No cell culture needed

DISADVANTAGES
• Low-throughput
• Limited number of commercial probes available
• Needs specialized camera and image capture system
Chromosome painting
Chromosome painting analysis of interphase nuclei reveals chromosome territories

Bolzer et al., PLoS Biol. 2005
Multiplexed super-resolution FISH and stochastic optical reconstruction microscopy (STORM)
a DNA-FISH
Chemical fixation, permeabilization, DNA denaturation

Hybridization
- Target region
- Probe

cryo-FISH
- 3D-FISH
 Imaging

b CRISPR-based live-cell imaging
Intact, living cell

sgRNA
dCas9
GFP

Image analysis
- Measure spatial distances
- Distance
  - A-B
  - A-C
- Non-interacting
- Interacting

Imaging
Capturing Chromosome Conformation

Job Dekker,¹* Karsten Rippe,² Martijn Dekker,³ Nancy Kleckner¹

15 FEBRUARY 2002 VOL 295 SCIENCE www.sciencemag.org
Advantages and drawbacks of “C”-technologies

### Table 1. Advantages and limits of 3C-derived methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Genomic Scale Investigated</th>
<th>Advantages</th>
<th>Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>3C-qPCR</td>
<td>~250 kilobases</td>
<td>Very high dynamic range (highly quantitative), easy data analysis</td>
<td>Very low throughput: limited to few viewpoints in a selected region</td>
</tr>
<tr>
<td>4C</td>
<td>Complete genome</td>
<td>Good sensitivity at large separation distances</td>
<td>Genome-wide contact map limited to a unique viewpoint (few viewpoints if multiplex sequencing is used)</td>
</tr>
<tr>
<td>5C</td>
<td>Few megabases</td>
<td>Good dynamic range, complete contact map (all possible viewpoints) of a specific locus</td>
<td>The contact map obtained is limited to a selected region</td>
</tr>
<tr>
<td>Hi-C</td>
<td>Complete genome</td>
<td>Very high throughput (complete contact map)</td>
<td>Poor dynamic range, complex data processing</td>
</tr>
</tbody>
</table>

**Diagram:**
- **3C:** one-to-one
- **4C:** one-to-all
- **5C:** many-to-many
- **Hi-C:** all-to-all

**References:**
Hi-C for genome-wide analysis of higher order chromatin structure

Mouse ES cells (from 433 Million Reads)

**in situ Hi-C**

In situ Hi-C maps DNA–DNA contacts occurring in intact nuclei, by proximity ligation.

While initial studies achieved only megabase resolution, the latest study with 15 billion contact reads, reaches kilobase resolution—which is a function of both the size of the restriction enzyme recognition sequence and the sequencing depth of the library.
HiChIP/PLAC-seq

Fang et al., 2016
Mumbach et al., 2016
Distribution of contact frequencies
Distribution of contact frequencies – clustering of smaller chromosomes
A and B Compartments

Lieberman-Aiden et al., 2009
Distribution of contact frequencies – cell to cell variability

Nagano et al. Nature 2013, October
Hi-C data reveals strong local chromatin interaction domains

\[ \text{DI}(x) = \frac{(B - A)}{\text{abs}(B - A)} \left( \frac{(A - E)^2}{E} + \frac{(B - E)^2}{E} \right) \]
The topological domains or TADs
The genome is composed of megabase sized topological domains

N = 2200
Dynamic chromatin organization during cell cycle

Naumova et al., Science 2013
Properties of the topological domains

- Topological domains are stable between different cell types.
- Topological domains are conserved between species

Dixon et al. Nature 2012
Overview of features revealed by Hi-C maps

Top: the long-range contact pattern of a locus (left) indicates its nuclear neighborhood (right).

Middle: squares of enhanced contact frequency along the diagonal (left) indicate the presence of small domains of condensed chromatin.

Bottom: peaks in the contact map (left) indicate the presence of loops (right). These loops tend to lie at domain boundaries and bind CTCF in a convergent orientation.
Overview of Hi-C analysis pipelines

These pipelines start from raw reads and produce raw and normalized contact maps for further interpretation. Colored boxes represent alternative ways to accomplish a given step in the pipeline. RE, restriction enzyme. At each step, commonly used file formats (‘.fq’, ‘.bam’, and ‘.txt’) are indicated.

A. The blue, pink and green boxes correspond to pre-truncation, iterative mapping and allowing split alignments, respectively.

B. Several filters are applied to individual reads

C. The blue and pink boxes correspond to strand filters and distance filters, respectively.

D. Three alternative methods for normalization

Fig. 3 Visualization of Hi-C data. An Epigenome Browser snapshot of a 4 Mb region of human chromosome 10. Top track shows Refseq genes. All other tracks display data from the human lymphoblastoid cell line GM12878. From top to bottom these tracks are: smoothed CTCF signal from ENCODE [130]; significant contact calls by Fit-Hi-C using 1 kb resolution Hi-C data (only the contacts >50 kb distance and $-\log(p\text{-value})<25$ are shown) [20]; arrowhead domain calls at 5 kb resolution [18]; Armatus multiscale domain calls for three different values of the domain-length scaling factor $\gamma$ [87]; DI HMM TAD calls at 50 kb resolution [15]; and the heatmap of 10 kb resolution normalized contact counts for GM12878 Hi-C data [18]. The color scale of the heatmap is truncated to the range 20 to 400, with higher contact counts corresponding to a darker color.
What you “C” might not be what you see

“C” technologies

✓ high throughput
✓ high resolution
✗ low ligation (detection) efficiency
✗ indirect cross-linking via nuclear structure
✗ not readily applicable to low cell number

Imaging

✗ low throughput
✗ limited resolution
✓ high detection efficiency
✓ direct visualization of proximity
✓ readily applicable to single cells

Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. Williamson et al. Genes and Dev (2014)
TADs and chromatin loops are two essential determinants of eukaryotic genome organization

**Topologically Associated Domains (TADs):** Contiguous stretches of chromosomal regions that exhibit enhanced physical interactions within them.

**Chromatin Loops:** Intra-TAD interactions that bring distal regulatory elements in proximity of their target genes by means of chromatin looping that are held together in part by CTCF.
Test: draw the loop structures
Loop extrusion model

- [https://www.youtube.com/watch?v=Tn5qqEqWqW8](https://www.youtube.com/watch?v=Tn5qqEqWqW8)
One sided extrusion
Disruption of chromatin organization by structural variation at a genetic locus containing genes and enhancers relevant to mammalian limb formation leads to pathological rewiring of genetic regulatory interactions resulting in three related human genetic disorders.
Insulator loss allows PDGFRA to interact with a constitutive enhancer in IDH gliomas.

(A) HiC map for IMR90 cells. Contact domain structure shown for a 1.7-Mb region containing PDGFRA. Convergent CTCF sites anchor a loop that separates PDGFRA and FIP1L1 (black circle). Interaction trace (below) depicts HiC signals between the PDGFRA promoter and all other positions in the region. Genes, FIP1L1 enhancer (per H3K27ac) and insulator (per HiC and CTCF binding) are indicated.

(B) The right CTCF peak in the insulator contains a CpG methylation sensitive CTCF-motif.

(C, D) ChIP-qPCR data shows that CTCF occupancy over the boundary is reduced in IDH mutant (red) gliomas and models, relative to wild type (black).

(E) Methylation levels of the CpG in the CTCF motif plotted as percentage of alleles protected from conversion measured in gliomaspheres by bisulfite sequencing.

(F) Methylation levels of the CpG in the CTCF motif plotted as relative protection from methylation-sensitive restriction measured in gliomaspheres.

Boundary methylation and CTCF occupancy affect PDGFRA expression and proliferation

(A) Schematic depicts chromatin loops and boundaries in the PDGFRA locus. In IDH wild-type cells (left), intact boundary insulates oncogene. Disruption of the boundary by removing the CTCF motif should activate the oncogene. In IDH mutant cells (right), hypermethylation blocks CTCF, compromising the boundary and allowing enhancer to activate the oncogene. Demethylation should restore CTCF-mediated insulation. meCpG, methylated CpG

(B) Plot compares CpG methylation in the CTCF motif in IDH wild-type gliomaspheres (black), IDH1 mutant gliomaspheres (red), and IDH1 mutant gliomaspheres treated with 5μM 5-aza for 8 days (purple).

(C) Plot compares CTCF occupancy over the boundary. DMSO, dimethylsulfoxide; WCE, whole-cell extract.

(D) Plot compares PDGFRA expression. Demethylation restores PDGFRA insulation in IDH1 mutant gliomaspheres.

Boundary methylation and CTCF occupancy affect PDGFRA expression and proliferation

(E) CTCF binding shown for the FIP1L1/PDGFRA region. Expanded view shows CTCF motif in the insulator targeted for CRISPR-based deletion. sgRNA and PAM direct Cas9 nuclease to the motif.

(G) Sequencing of target site reveals the indicated deletions. CTCF motif disrupted on ~25% of alleles (compare to <0.01% in control).

(I) qPCR reveals increased PDGFRA expression in insulator CRISPR cells.

(J) PDGFRA and FIP1L1, which are normally confined to separate loop domains rarely interact.

(K) But can become closely associated in IDH-mutant tumors.
Formation of new TADs as a result of translocation
Formation of new TADs as a result of translocation
4D Nucleome Project

a) Mapping – molecular genomic mapping of contacts, imaging of contacts and dynamics

b) Model building – by constraint modelling and polymer simulation

- Cell
- Nuclear membrane
- TADs
- Chromosome
- Chromosomal regions
- Regions brought into topological contact by polymer rearrangement
- ΔT

C) Functional validation – by genetic and biophysical perturbation experiments

- Binding sequences
- CRISPR complex
- Loop-related proteins
- Loop formation

- Loop-related protein
4D Nucleome Project

De novo modelling

Hypotheses, mechanisms, prior knowledge

Model

Comparison to data

Revise model/parameters

Data-driven modelling

Self-consistent model of ensemble

More predictions: dynamics, mechanism, effects of mutations
Summary

• Hi-C analysis reveals that the mammalian genome is spatially compartmentalized, and consists of mega-base sized topological domains (also known as TADs).

• Topological domains have been independently observed in flies (Sexton et al. Cell 2012; Hou et al, Mol Cell 2012) and with different approaches (5C, Nora et al., Nature 2012)

• Topological domains are stable across cell types and largely preserved during evolution, suggesting that they are a basic property of the chromosome architecture.

• Partitioning of the genome into topological domains would naturally restrict the enhancers to selective promoters

• Long-range looping interactions between enhancers and promoters correlate with higher transcriptional responsiveness of promoters.

• Cell specific enhancer/promoter interactions are formed in each cell type, some time prior to activation of the genes, and are not significantly altered by transient signaling induction

• Pre-existing, lineage specific chromatin looping interactions between enhancers and promoters predict transcriptional responses to extracellular signaling, suggesting that chromatin conformation is another layer of transcriptional control