Data Analysis for Nuclear Compartmentalization

Jian Ma

Computational Biology Department

School of Computer Science

Carnegie Mellon University

July 21, 2017

3D Genome Data Processing, Analysis, and Visualization Tutorial

ISMB 2017, Prague

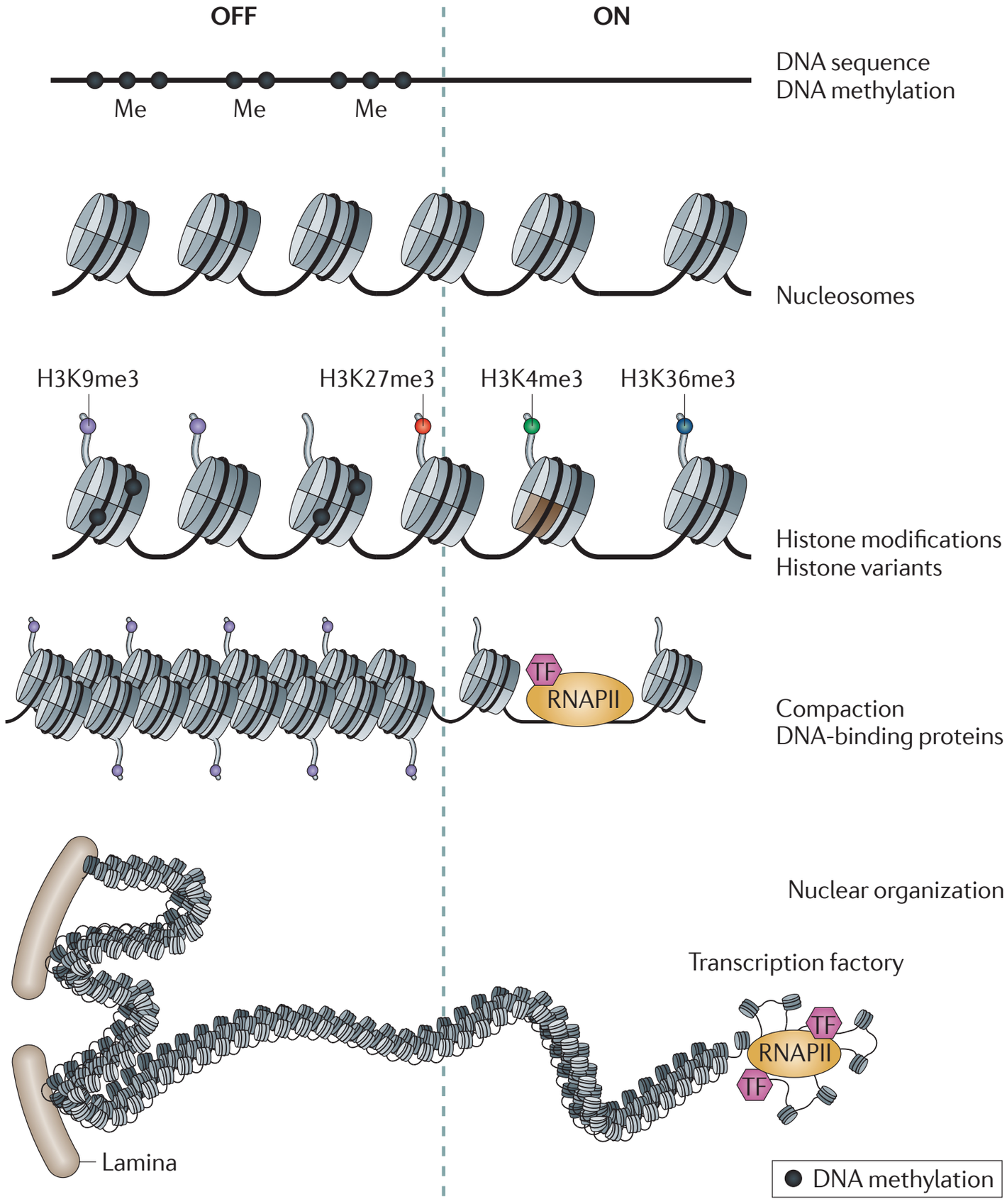
4D Nucleome

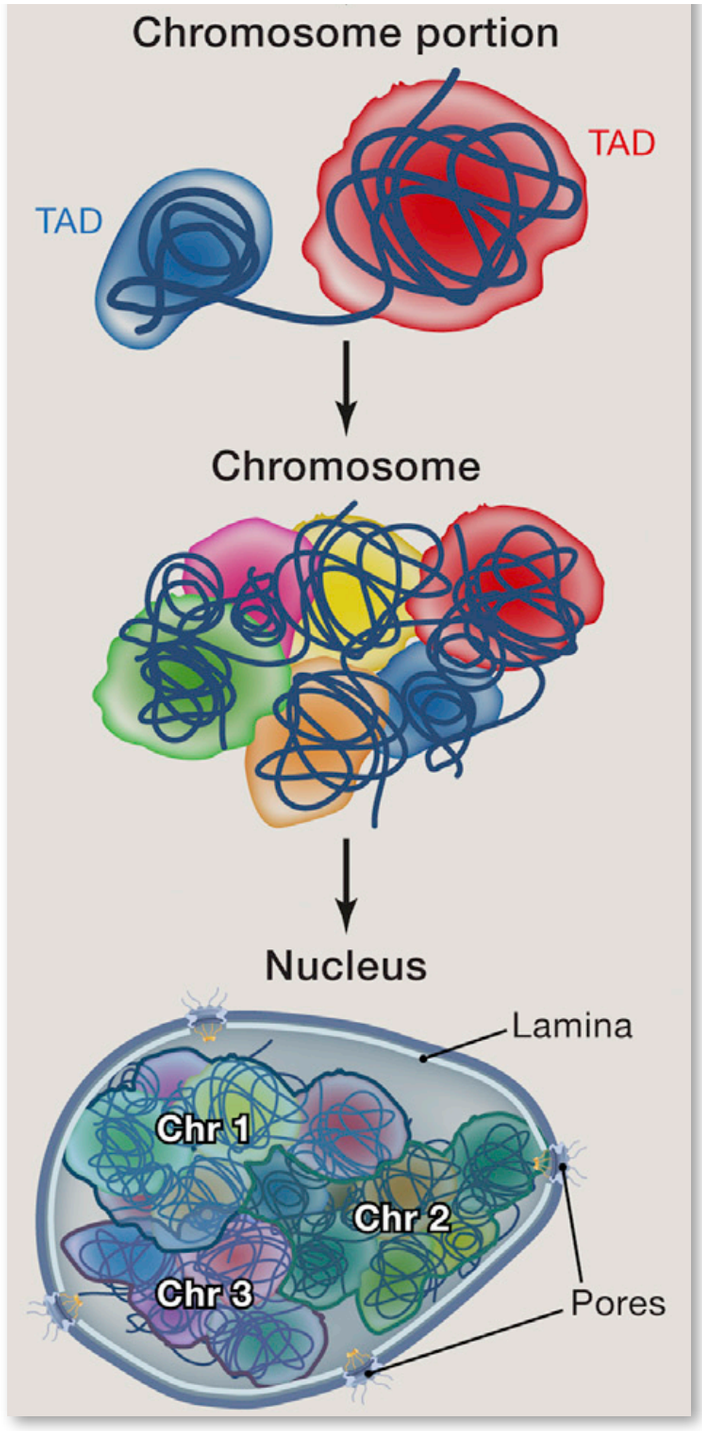
ISCB ISMB CELEBRATING 25 YEARS OF ISMB ECCB 2017 • PRAGUE JULY 21 TO JULY 25

Outline

- What is nuclear compartmentalization?
- Introduction to DamID data
- Introduction to Repli-seq data
- Demo on visualization of the data and integrated analysis

Chromatin organization hierarchies



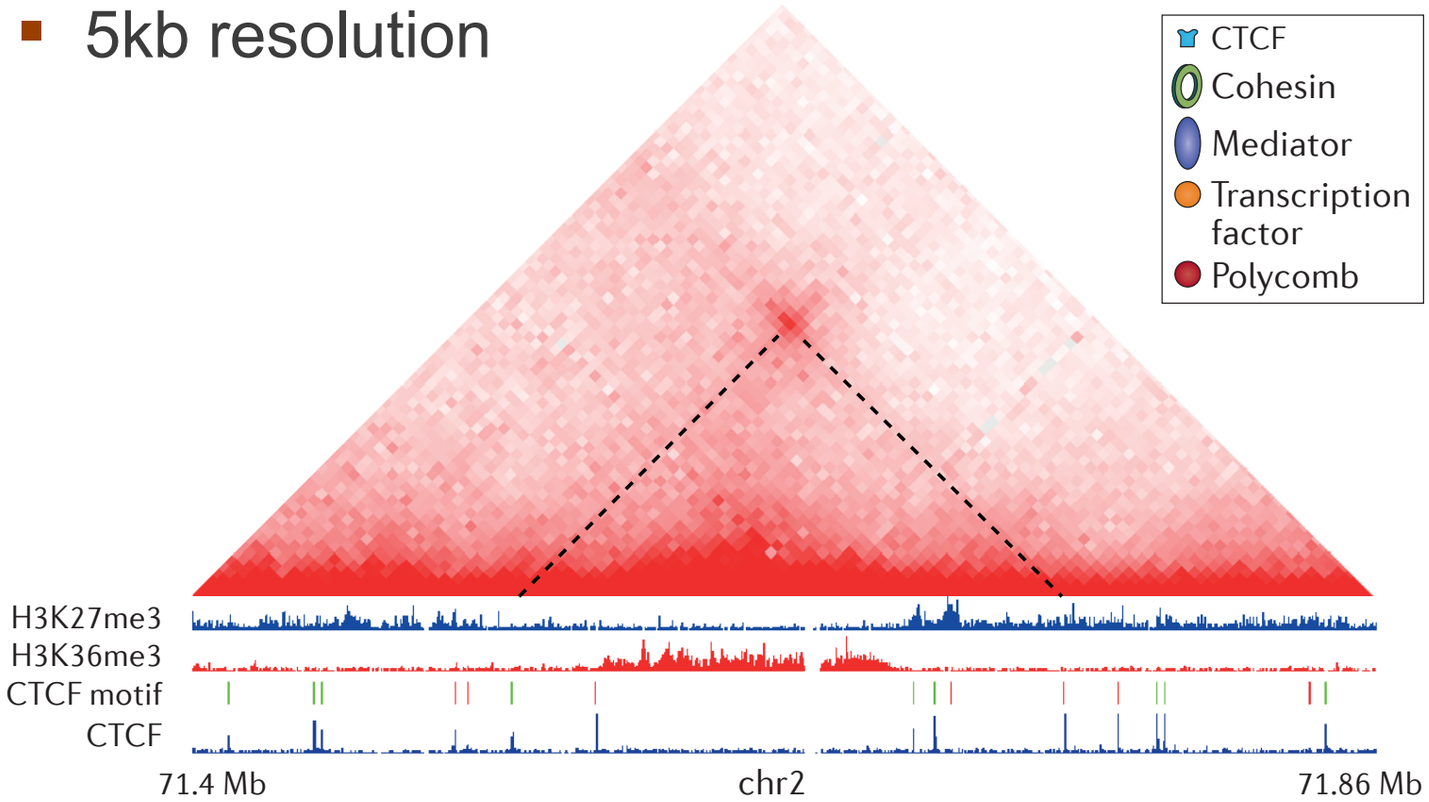


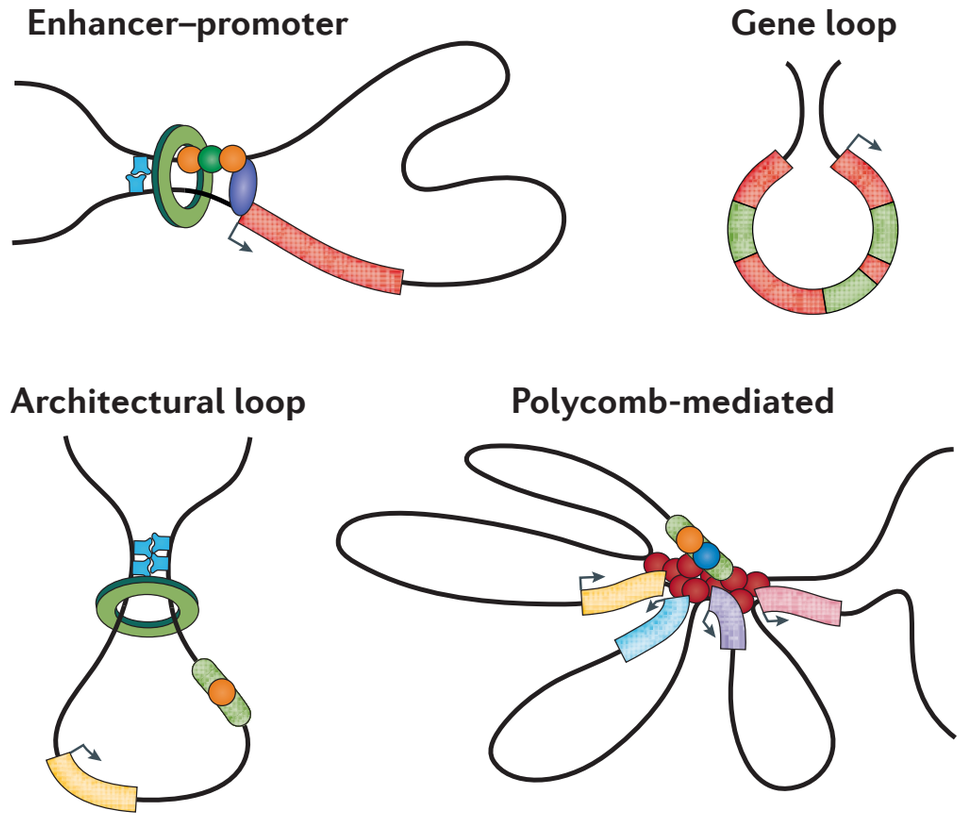
Zhou et al. *Nature Rev Gen* 2011

Sexton and Cavalli. *Cell* 2015

Chromatin interaction in different resolutions

- 5kb resolution



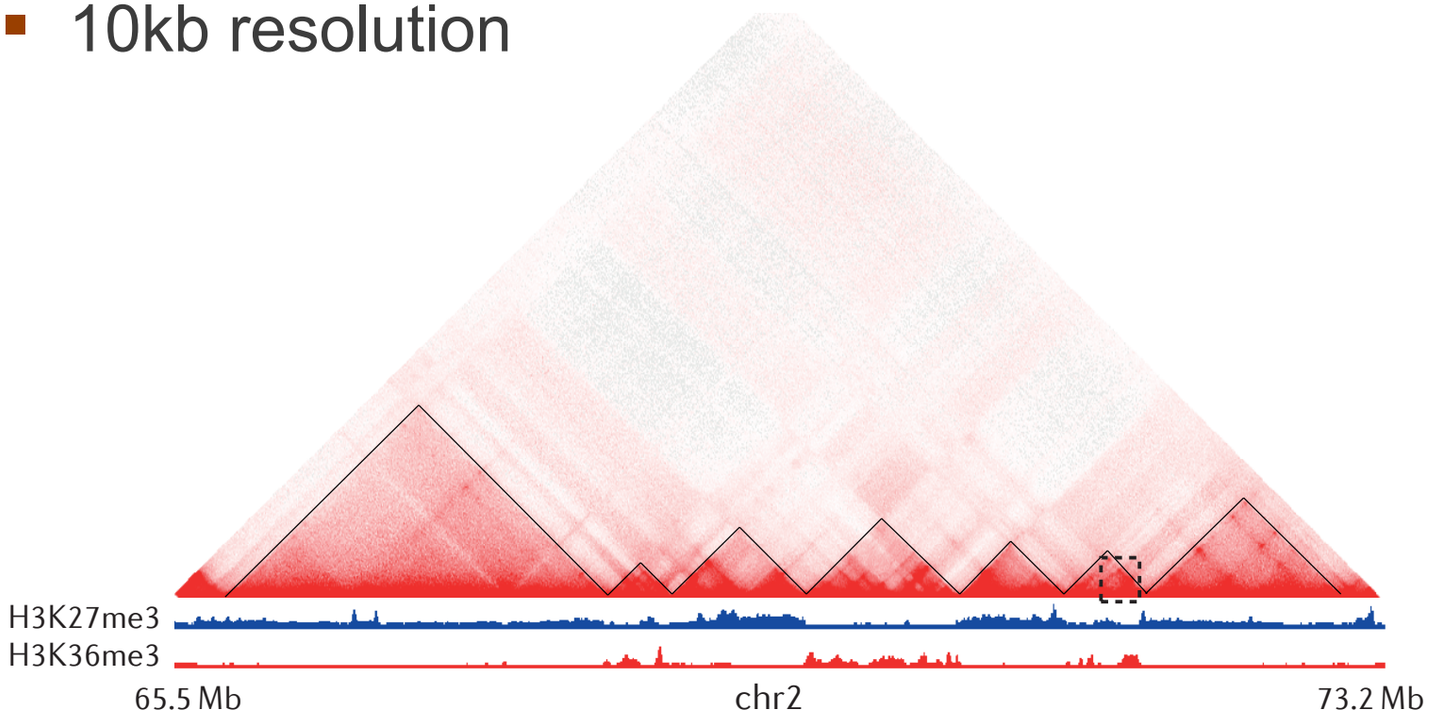


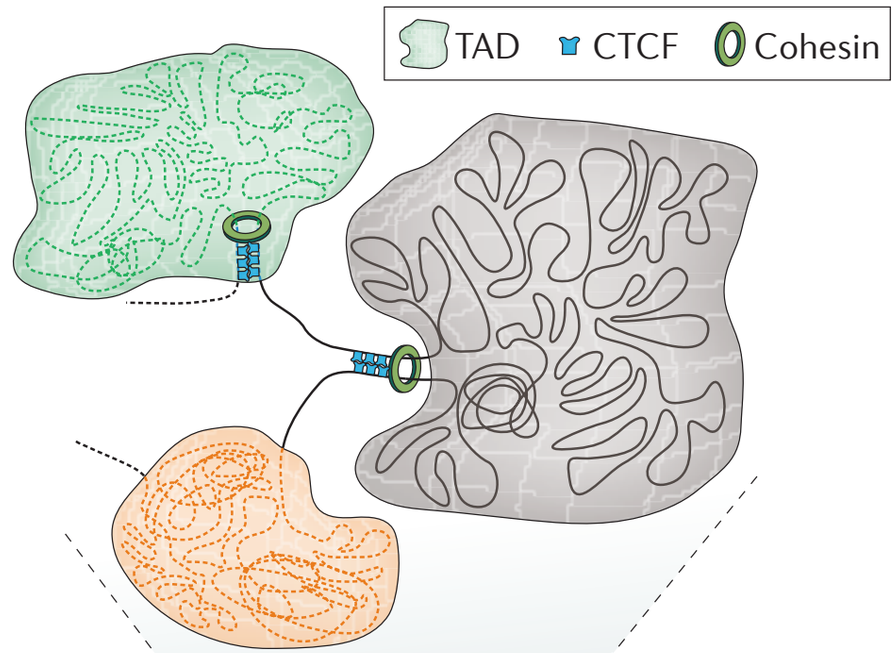
- Right — different types of **chromatin loop** within a domain
  - Enhancer-promoter loop, Polycomb-mediated loop, gene loop, or architectural loop via CTCF.
- Left — an example of an architectural loop observed in high-res Hi-C

Bonev and Cavalli. *Nature Rev Gen* 2016

Chromatin interaction in different resolutions

- 10kb resolution



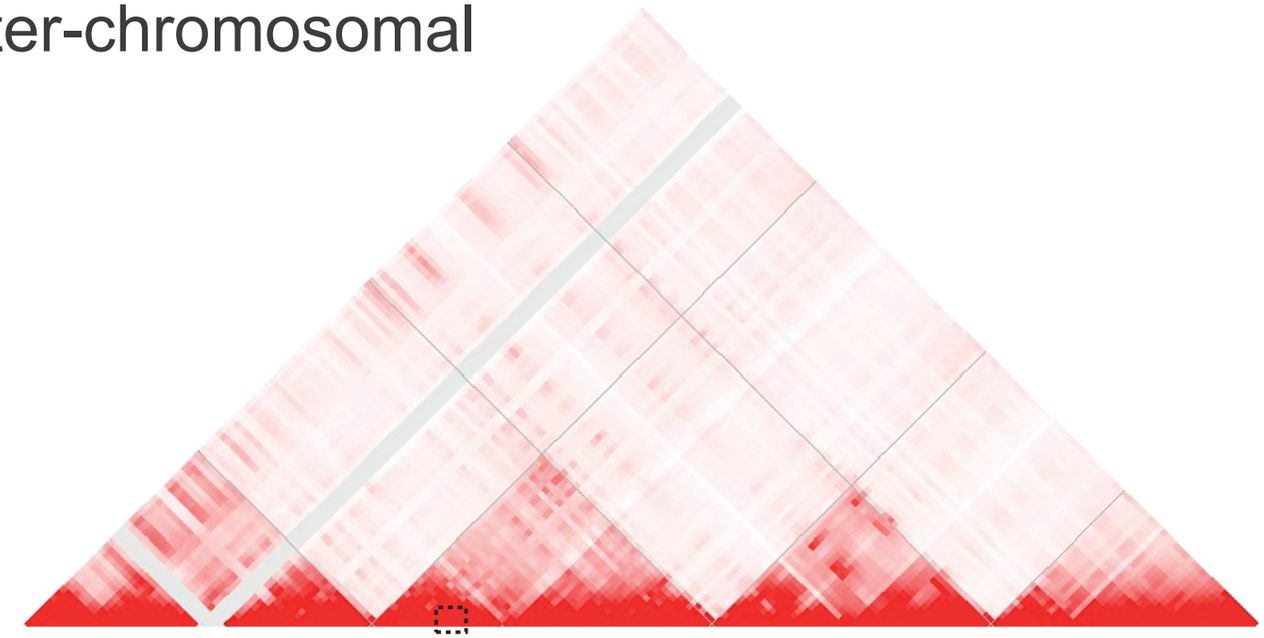


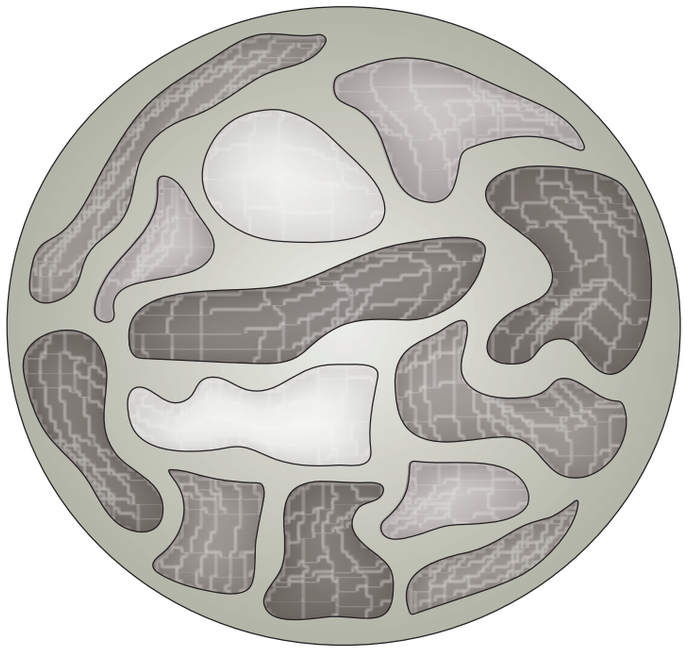
- Left — topologically associating domains (**TADs**) as seen in Hi-C maps
- Right — three different TADs

Bonev and Cavalli. *Nature Rev Gen* 2016

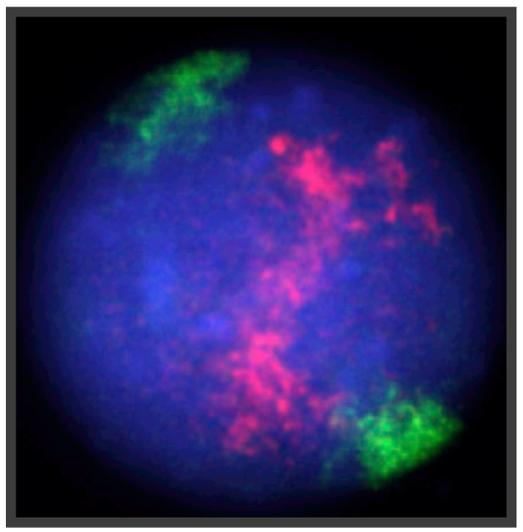
Chromatin interaction in different resolutions

- Inter-chromosomal





- Highest-level 3D interactions between individual chromosomes are less common.
  - Right — Chromosomes occupy different nuclear territories
  - General understanding is that gene-rich chromosomes are spatially located inside the nuclear interior, and gene-poor chromosomes are localized close to the nuclear periphery. However, the detailed nuclear organizations relative to different nuclear compartments remain mostly unclear



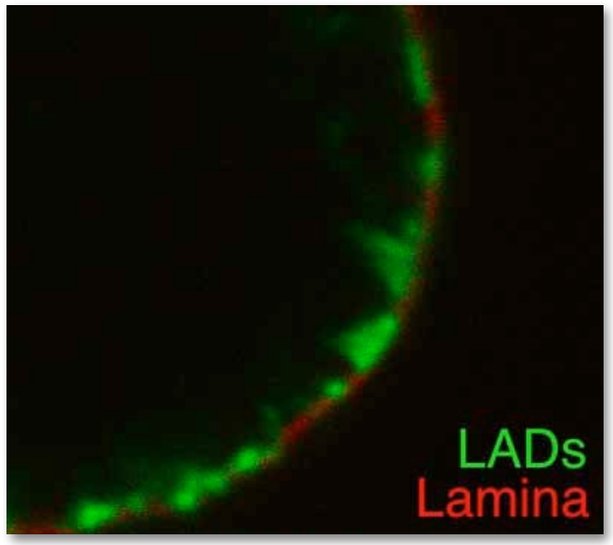
Human lymphoblastoid cell
gene-rich chr19 (red)
gene-poor chr18 (green)

Bonev and Cavalli. *Nature Rev Gen* 2016

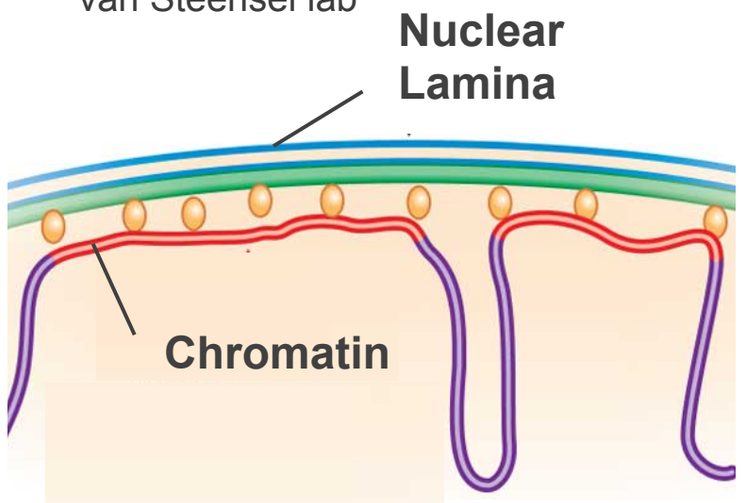
Nuclear compartments

- Examples include nuclear lamina, nuclear speckle, nucleolus ..

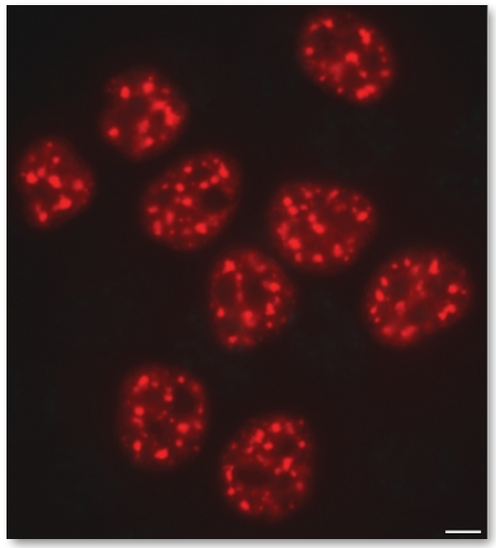
Nuclear lamina



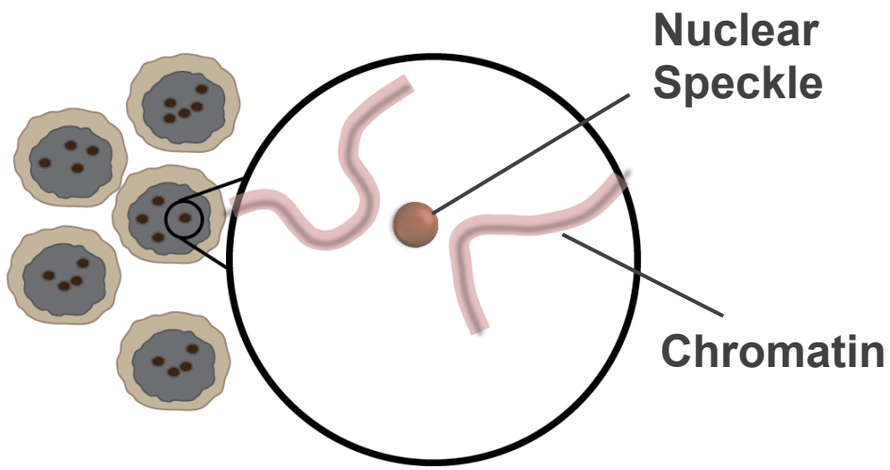
van Steensel lab



Nuclear speckle



Spector and Lamond
Cold Spring Harb Perspect Biol 2011

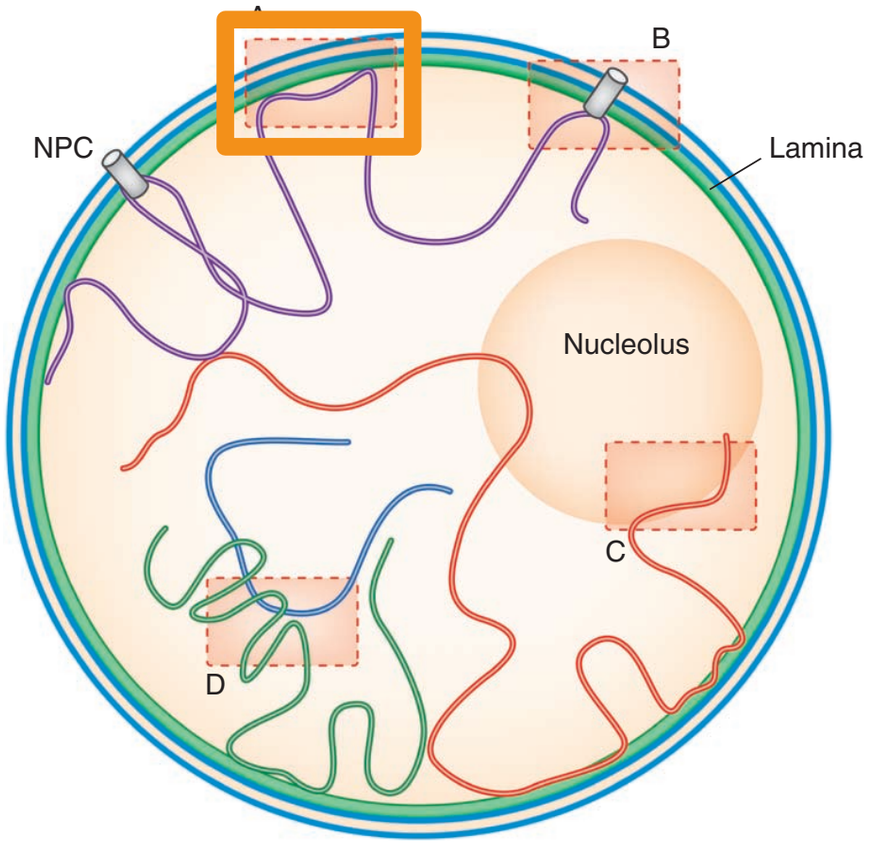


4D Nucleome

Andrew Belmont
Bas van Steensel
David Gilbert
Huimin Zhao
Jian Ma

- Spatial position of the chromosome relative to nuclear compartments has strong connection with gene regulation

The nuclear lamina as an anchoring platform for the genome



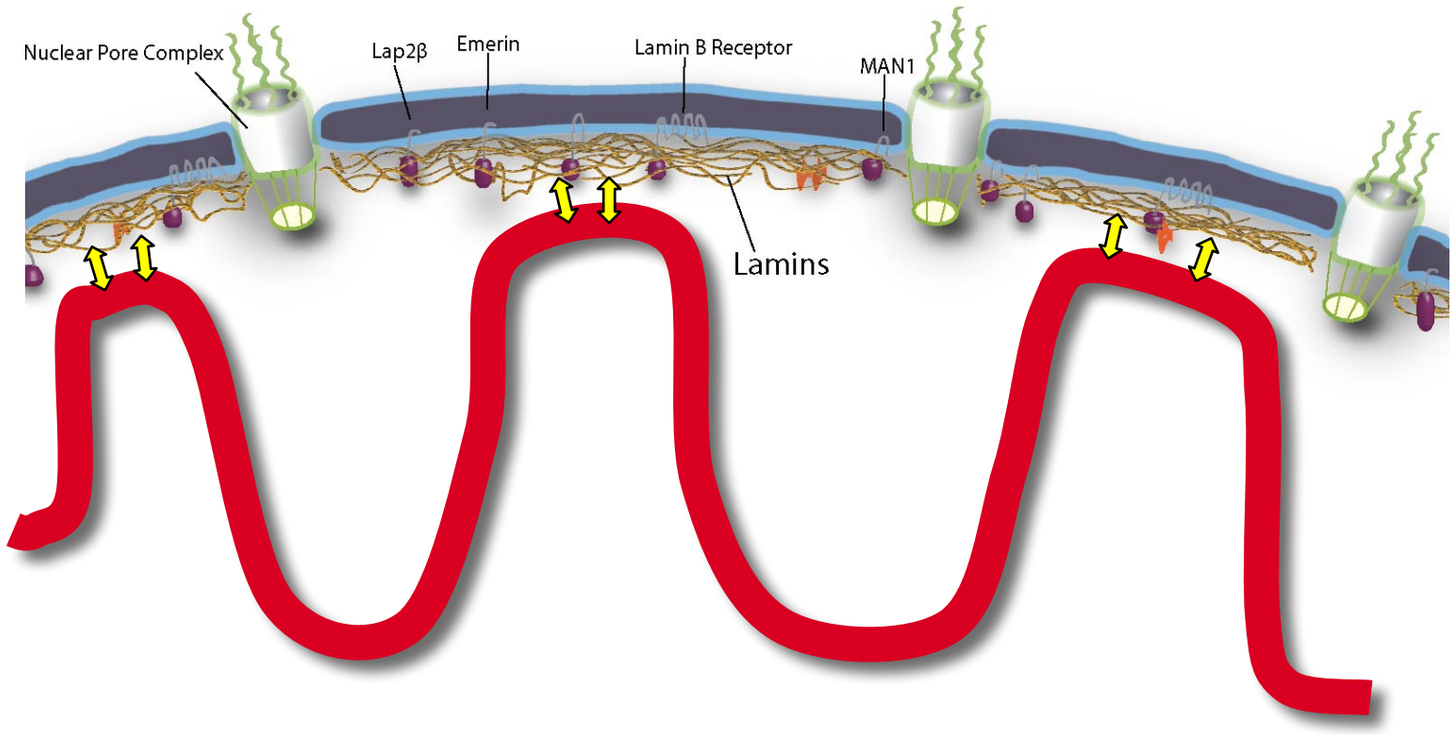
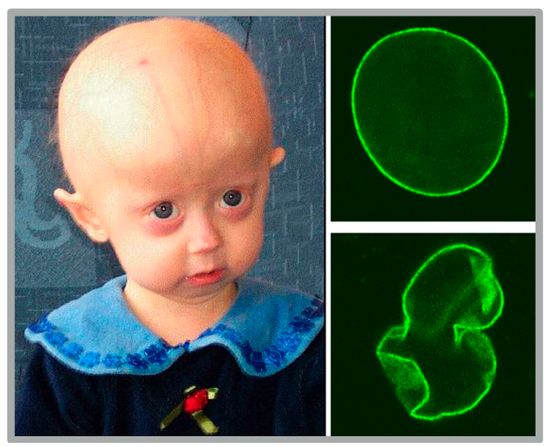


Figure provided by Bas van Steensel



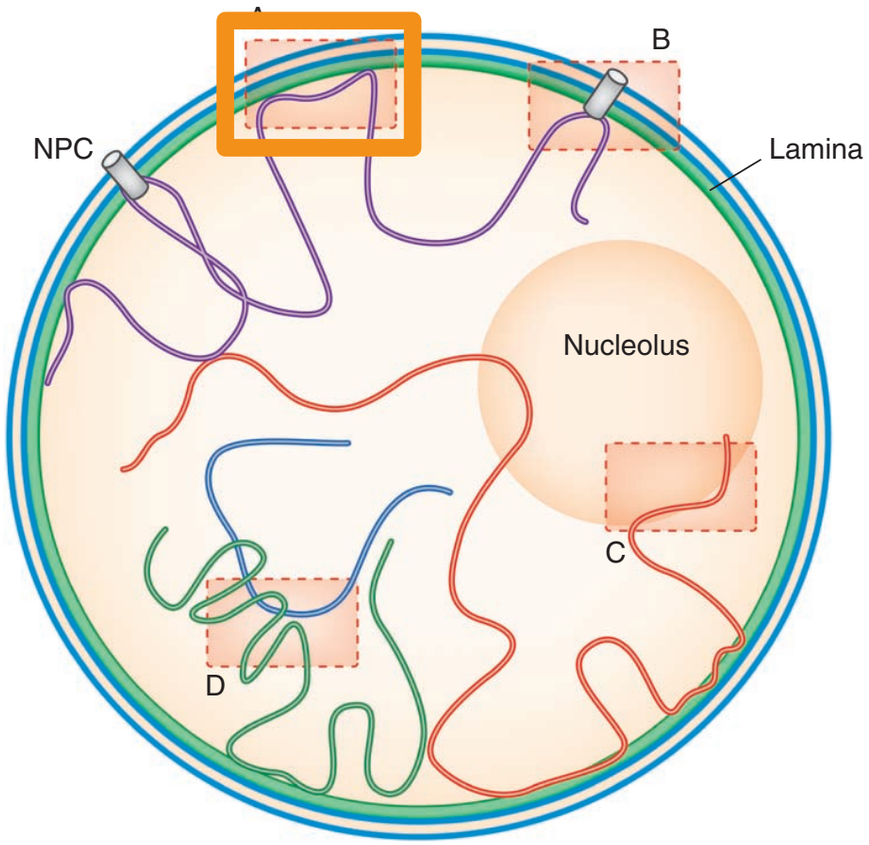
normal cell nucleus

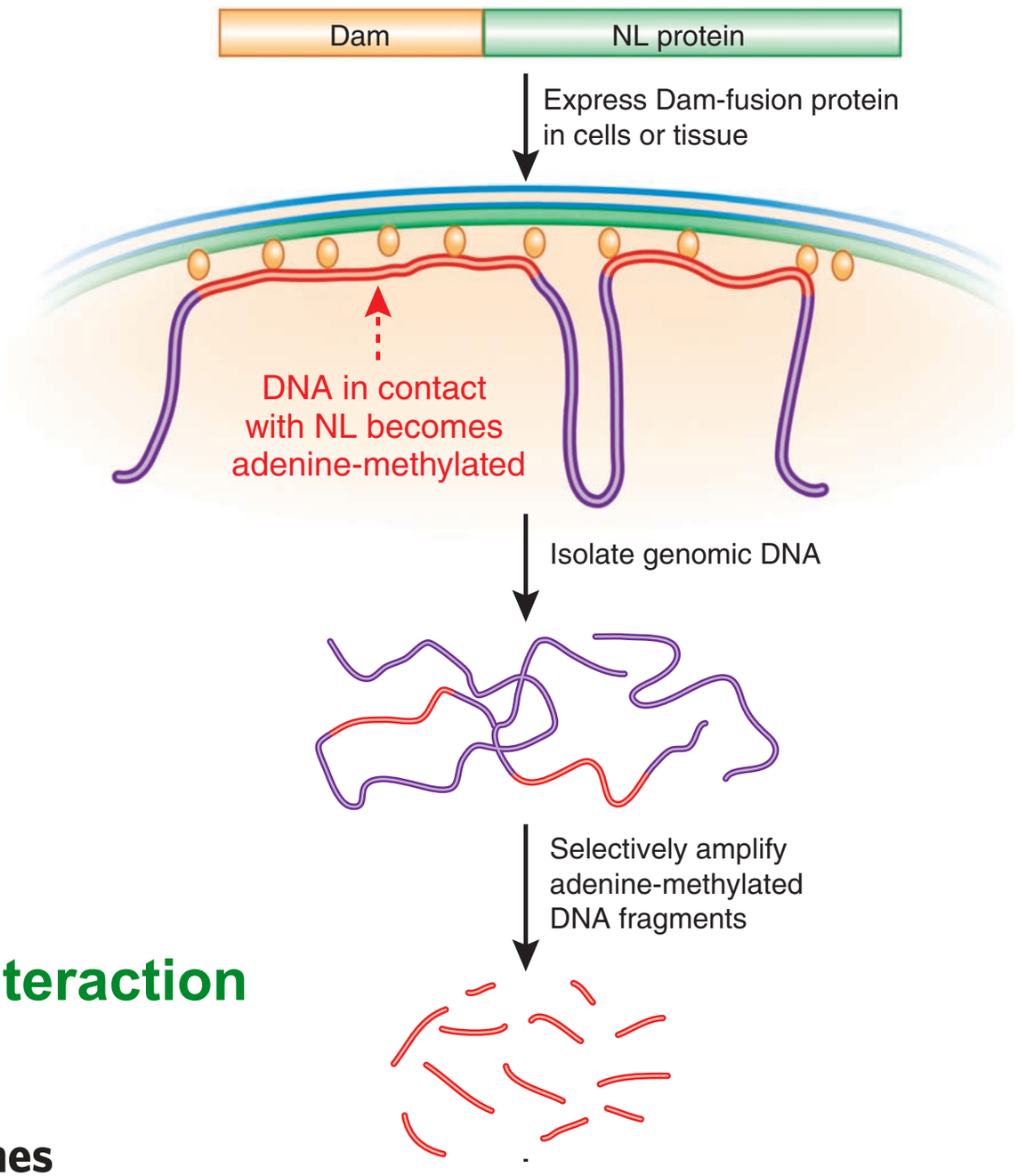
progeric cell nucleus

wikipedia

- Mutations of *LMNA* gene result in unstable nuclear envelope that progressively damages the nucleus, making cells more likely to die prematurely.
- Progeria patients typically die as teenager.

DamID — Contact Frequency Mapping





DamID to map nuclear lamina interaction

- Guelen et al. *Nature* 2008

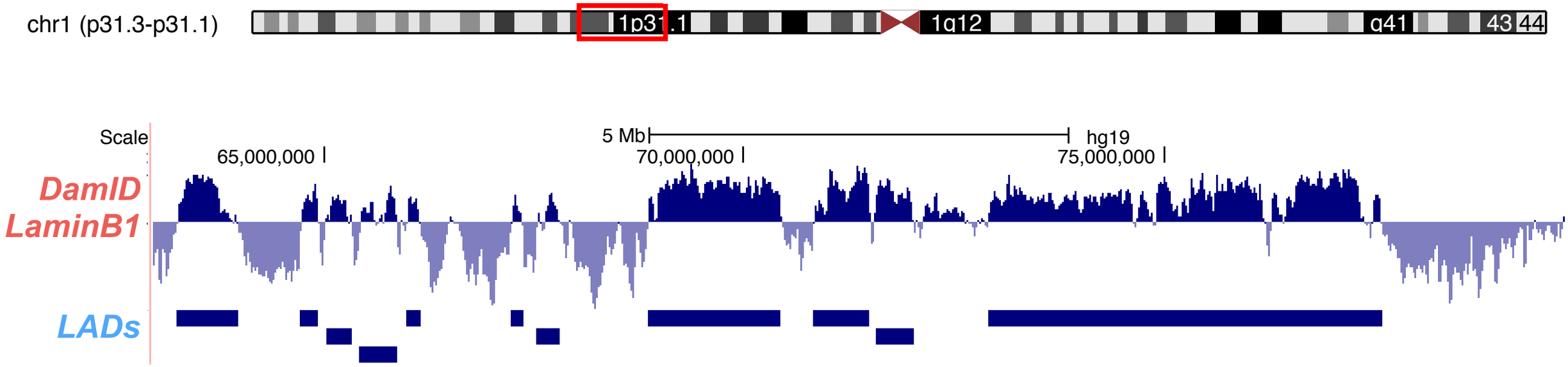
Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions

Lars Guelen[1], Ludo Pagie[1], Emilie Brasset[2], Wouter Meuleman[1,4], Marius B. Faza[1], Wendy Talhout[1], Bert H. Eussen[3], Annelies de Klein[3], Lodewyk Wessels[1,4], Wouter de Laat[2] & Bas van Steensel[1]

van Steensel and Dekker, *Nature Biot* 2010

DamID data processing

- Map raw sequencing reads to reference genome (e.g., Bowtie 2)
- Any reads that didn't map to GATC sites were discarded.
- Count the reads on Dam-LaminB and control, and use log2 ratio as DamID signal.



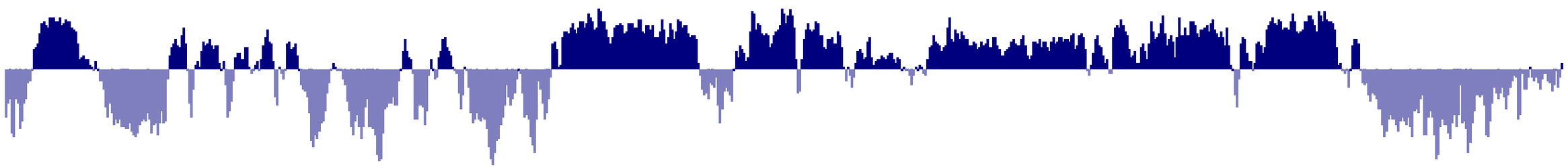
Lamina-associated domains (LADs)

Mouse and human cells:

- ~1,300 LADs
- size range: 0.1-10Mb (median 0.5Mb)
- total coverage: 35-40% of the genome
- thousands of genes in LADs: *most are inactive*
- differentiation: ~1,500 genes relocate

- How to identify LADs based on DamID data?

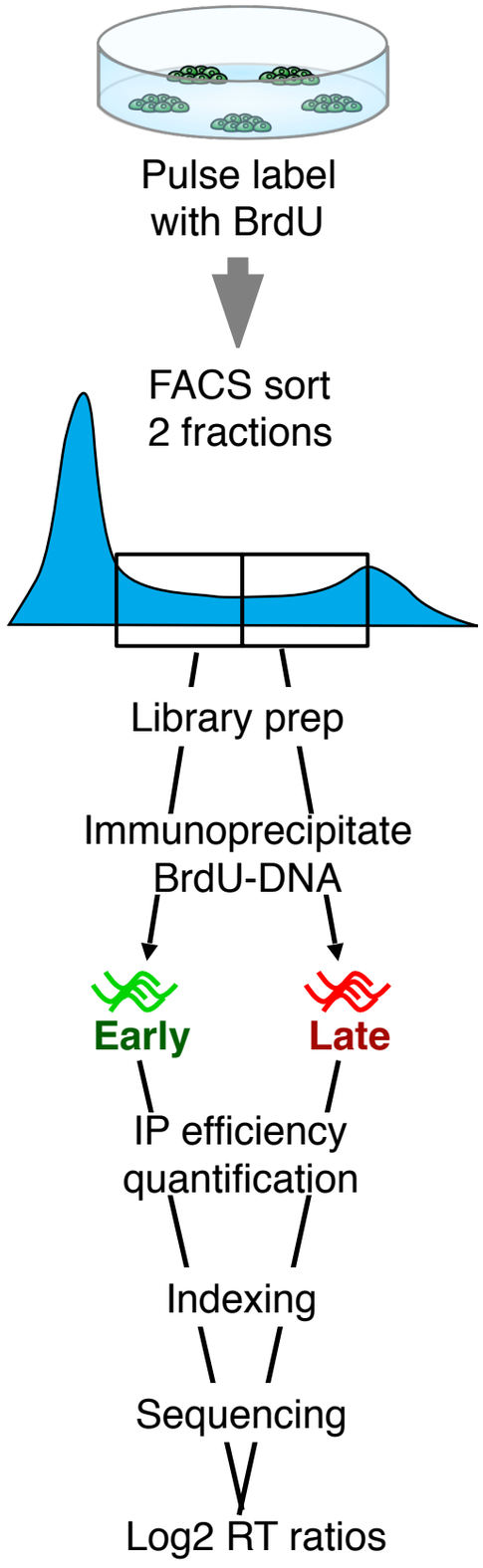
Identifying LADs



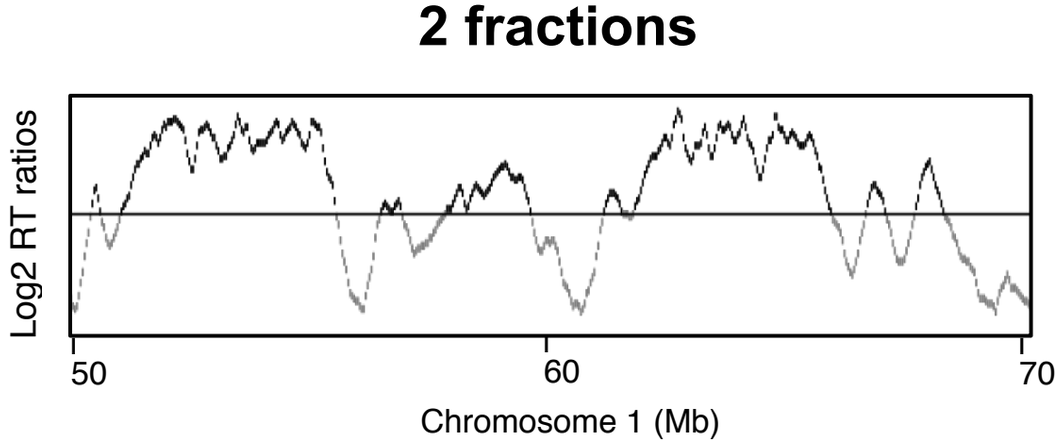
- We can use a HMM to identify LADs
- **Hidden states: LADs** and inter-LADs (**iLADs**)
- **Observations:** signals as shown above
  - In Meuleman et al. *Genome Res* 2013, the Student's *t*-distribution is used as the emission probability distribution (mean and variance are different between the two states but the degree of freedom is the same)

- How to learn the parameters in this HMM?
- How to identify LADs and iLADs?

Repli-seq — Mapping Replication Timing



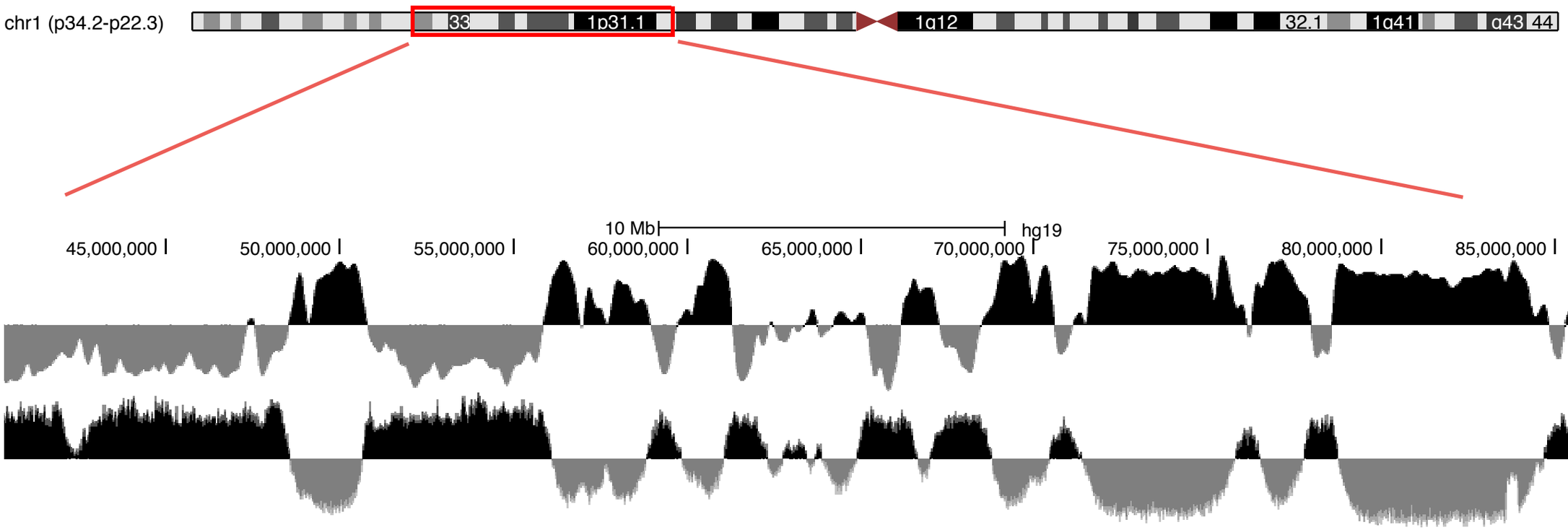
- Map raw reads from each fraction (e.g., E/L) to the reference genome.
- The raw repli-seq signals within each fraction are determined by the number of mapped reads within a window (e.g., 60kb).
- Use log2 ratio as E/L replication timing signals.



David Gilbert lab

(Gilbert *PNAS* 1986; Gilbert & Cohen *Cell* 1987; Hiratani et al. *PLoS Biology* 2008)

Correlation between DamID LamB and Repli-seq



Demo