Encoding and Decoding 3D Genome Organization

Noam Kaplan

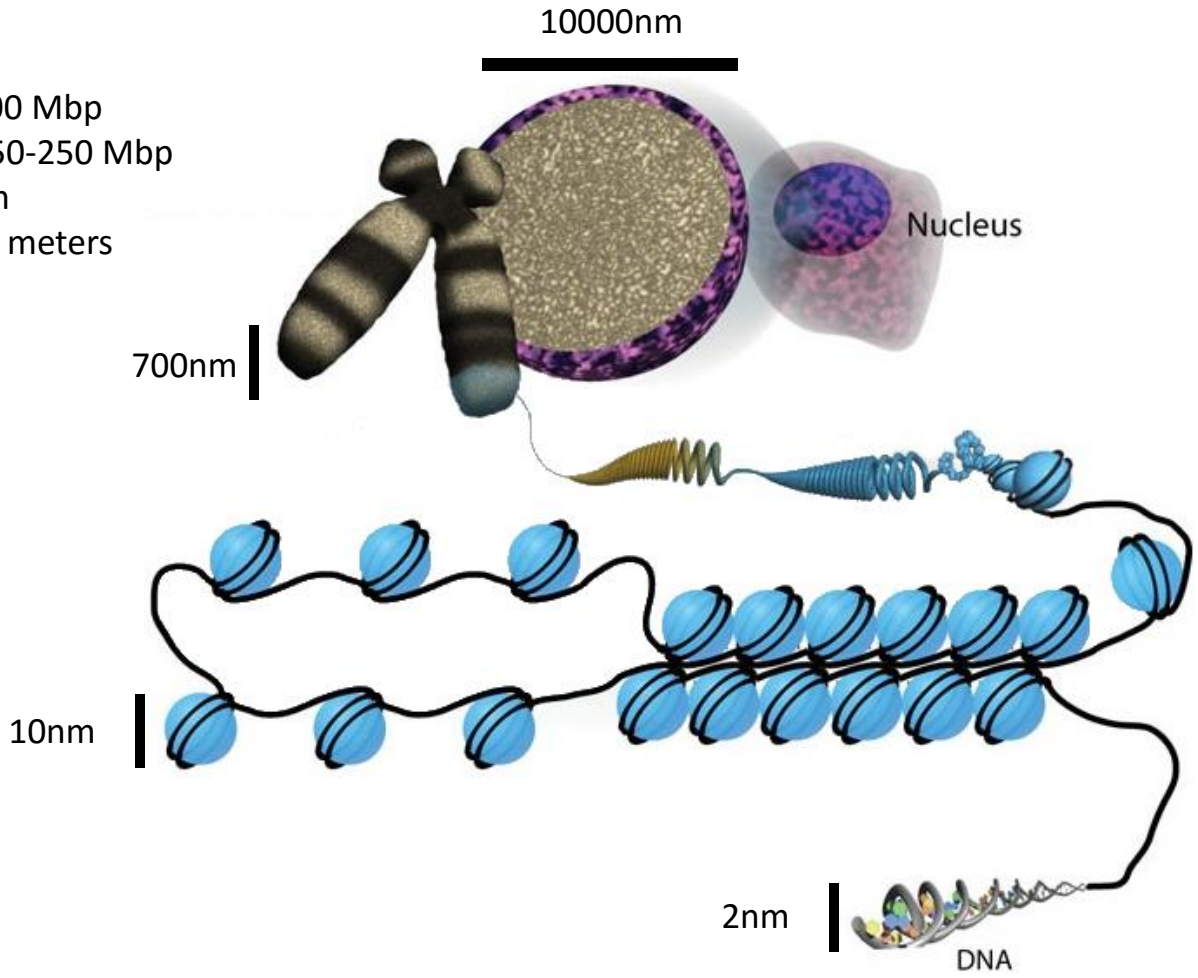
Dekker Lab

Program in Systems Biology

University of Massachusetts Medical School

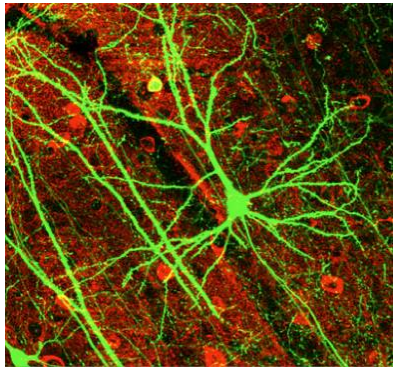
Chromatin organization

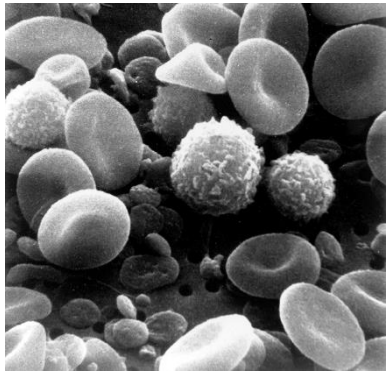
Human genome: 2x3000 Mbp
Human chromosome: 50-250 Mbp
Avg. bp length: 0.33 nm
Total genome length: 2 meters



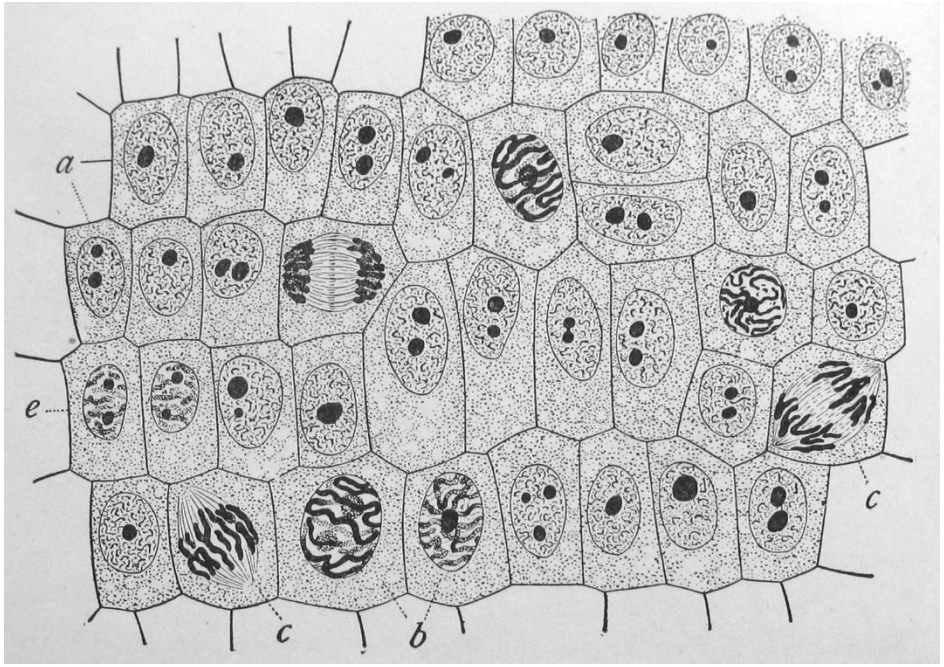
Chromatin structure and function

Gene regulation





Studying chromatin: From microscopy to genomics

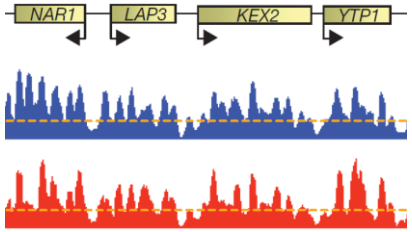


Wilson, *The Cell* 1900

Schermelleh et al., *Science* 2008

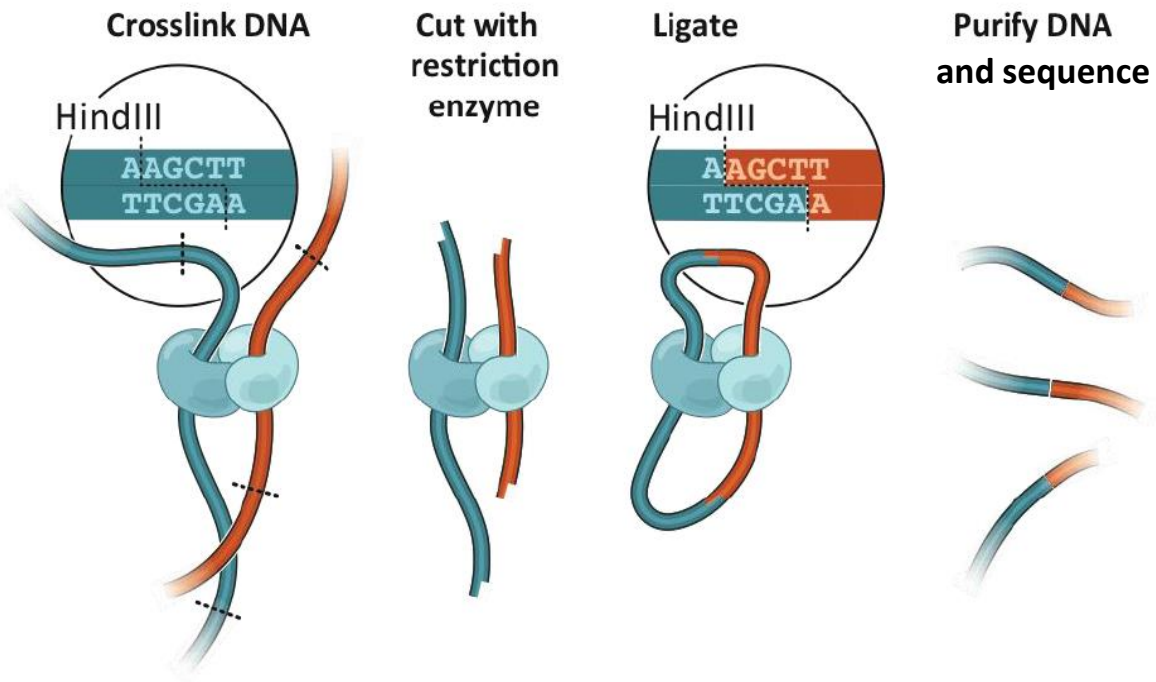
The human genome (2001)

Next Generation Sequencing



Genomic tracks

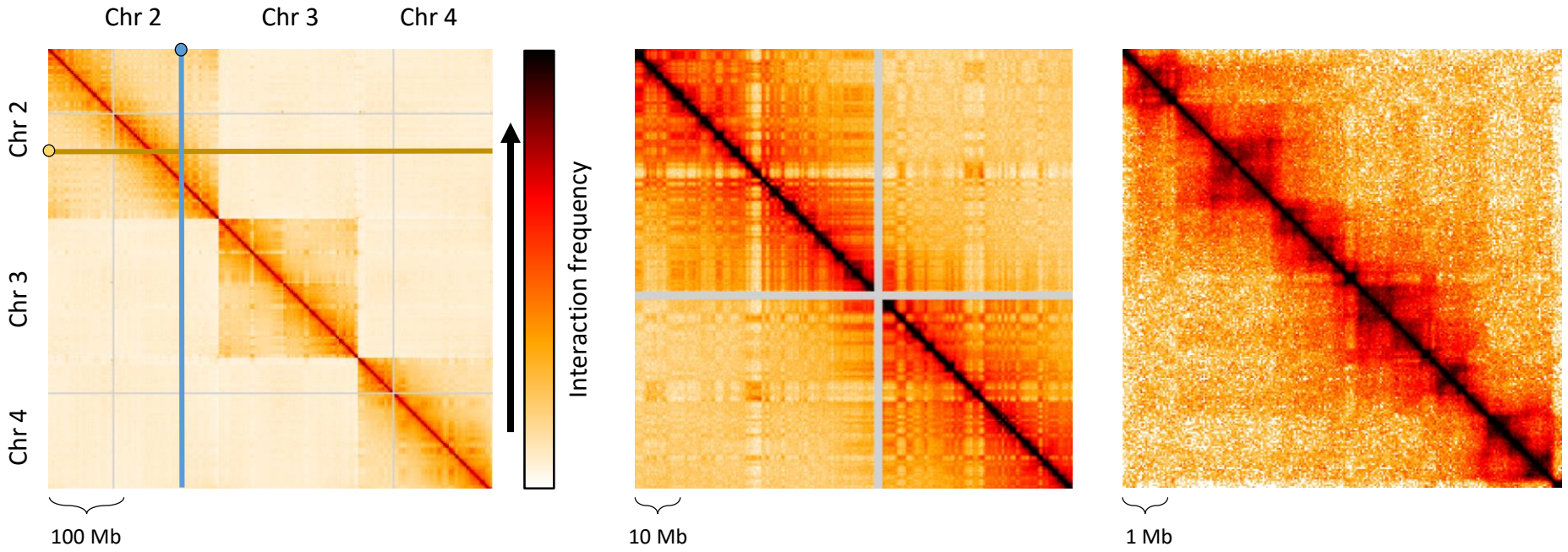
Chromosome Conformation Capture (3C)



One vs one: 3C-PCR
One vs all: 4C
Many vs many: 5C, capture Hi-C
All vs all: Hi-C, 3C-seq

3C: Dekker et al., *Science* 2002
Hi-C: Lieberman-Aiden et al., *Science* 2009

Interaction maps



Hi-C resolution: Some numbers

- **Human genome size:** 3*10^9 bp
- **Restriction enzyme site:** 6 bp (4 bp in some cases)
- **Restriction fragment length:** ~4 kbp "optimal resolution"
- **Number of restriction fragments:** ~7.5*10^5 bp
- **Size of interaction space:** 5.6*10^11 possible interactions

- **Next Generation Sequencing lane:** 10^8 usable reads, 2.5K USD
- **Number of cells per experiment**: 10^6-10^8
- **Number of interactions measured per cell:** 10^4-10^5 (?)

Interaction space is under-sampled

BUT interaction map is highly non-uniform

Solutions:

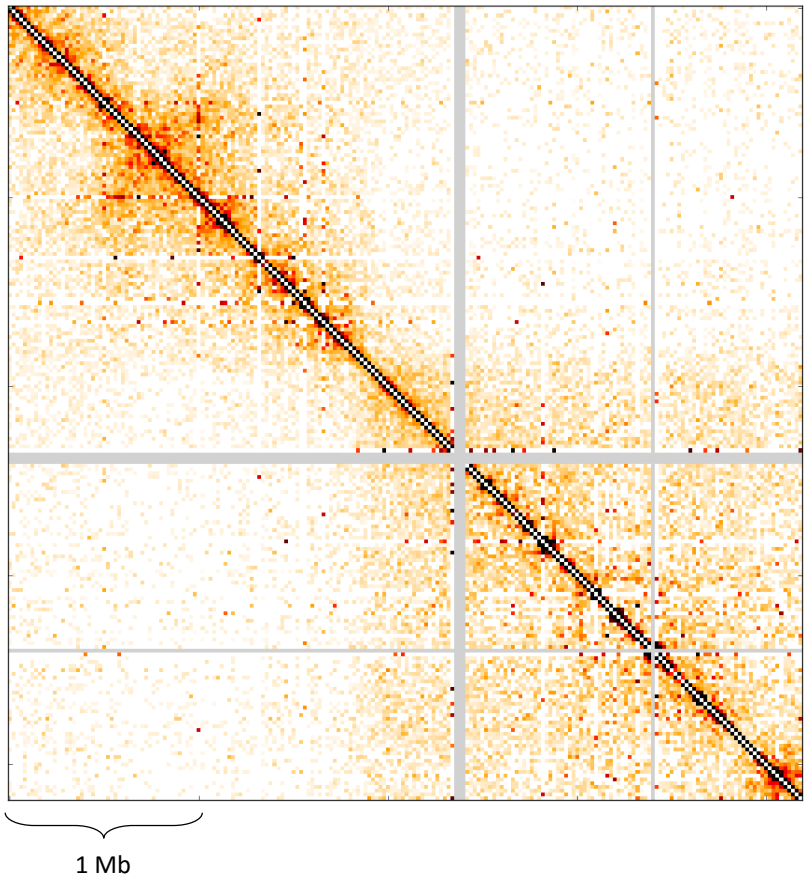
1. Reduce interaction space (binning; focused experiment)
2. Sequence more ($$$)
3. Acknowledge limitations in measurement of infrequent interactions

Practically:

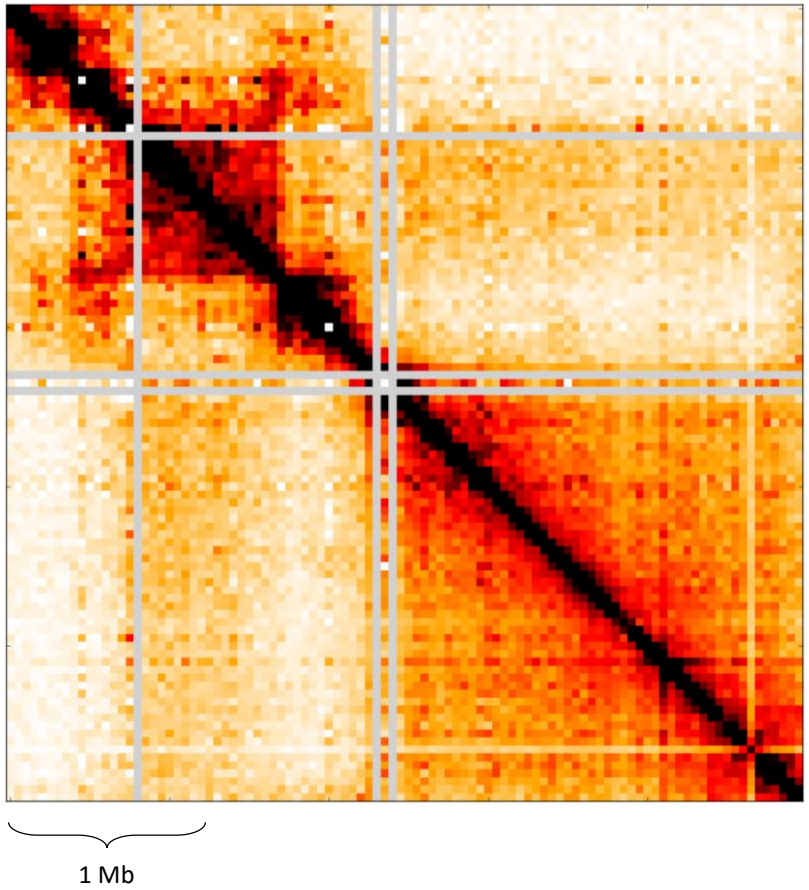
1. ~5-50kb resolution in human genome
2. Library complexity is not a major issue yet

Hi-C resolution

Fragment (~4kb) resolution



40 kb resolution



Bias normalization

- **Problem:** Locus-specific biases (e.g. sequencing, mapping)
- **Assumption:** Sum of reads from every row/col should be approx. equal

Imakaev et al., *Nature Methods* 2012

- Sinkhorn (1964): Given symmetric matrix A with positive elements, find a unique doubly stochastic matrix B, and diagonal matrix D with strictly positive elements, such that $A = DBD$.

Bias factors
(locus-specific)

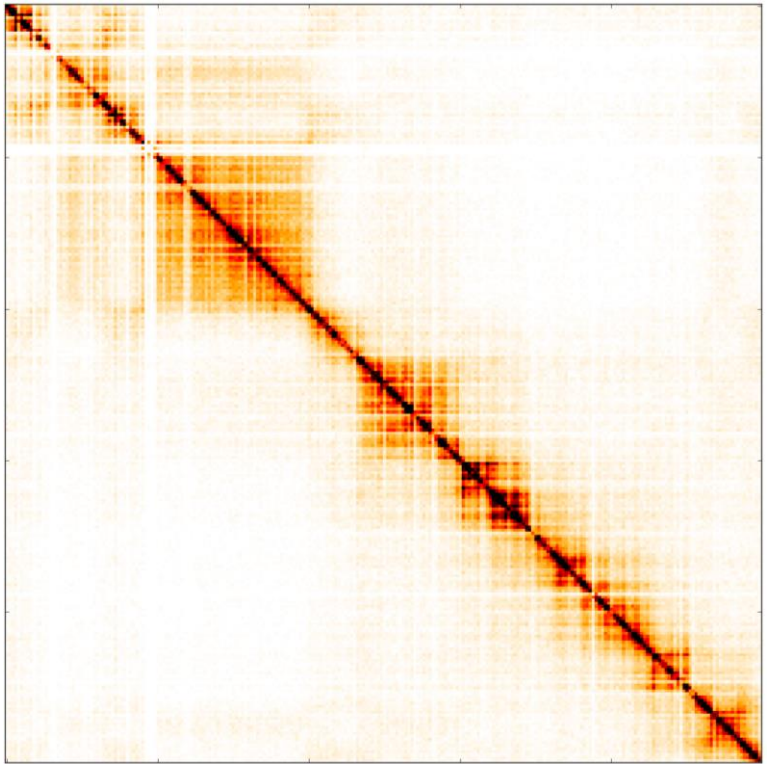
- So: $A_{i,j} = d_i d_j B_{i,j}$

Read count

Normalized
Read count

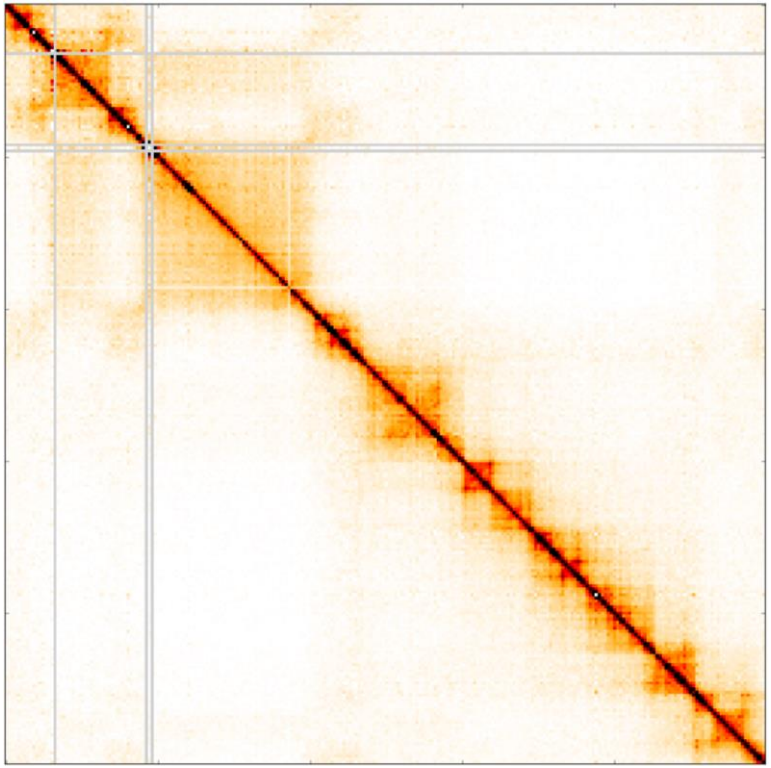
Normalization

Before



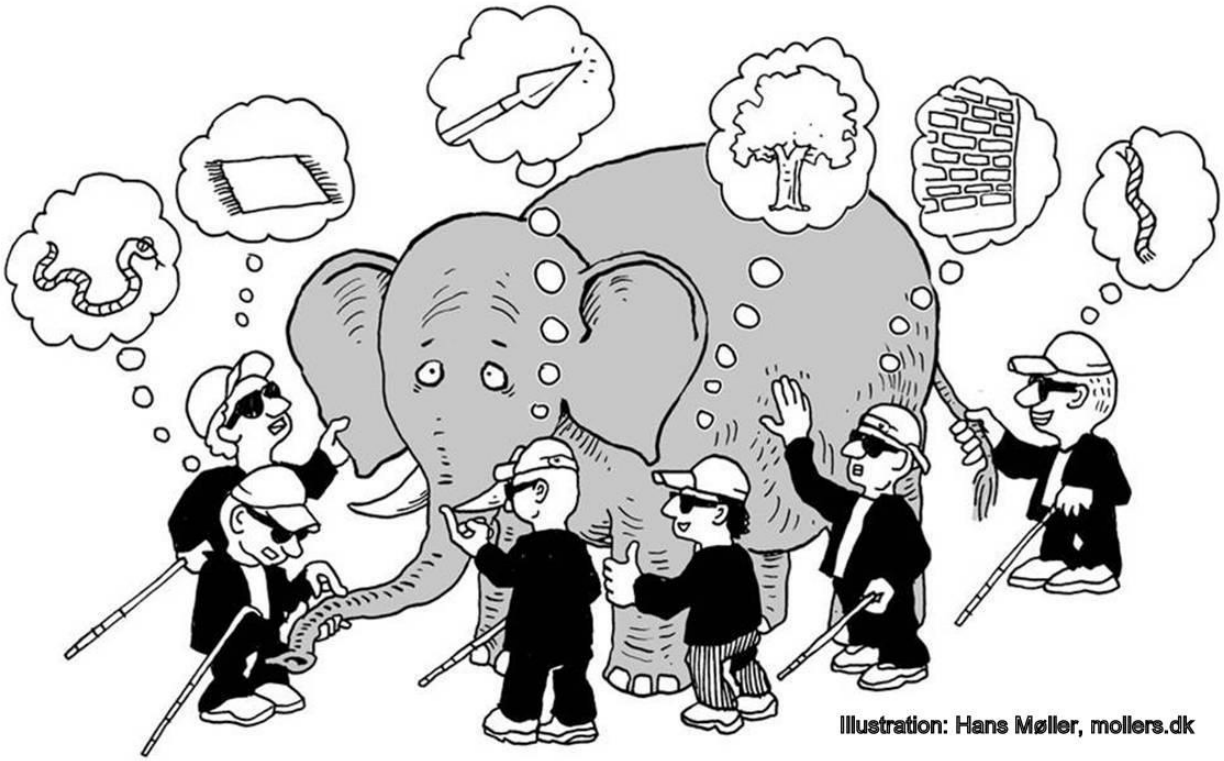
1 Mb

After



1 Mb

Hi-C interpretation

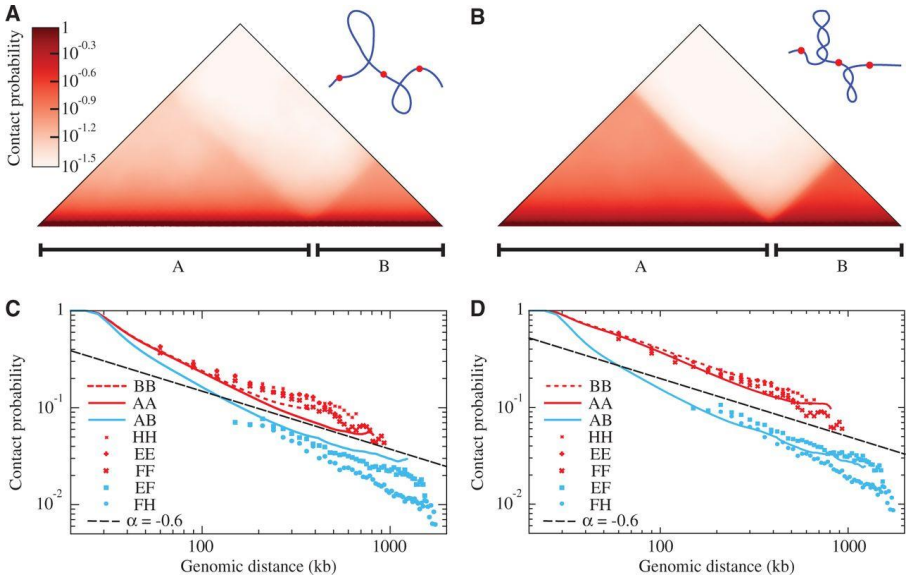


Approaches to Hi-C analysis

Physical	**Informatic**
Structure-based	Pattern-based
Specifies mechanism	Doesn’t specify mechanism
Physical assumptions required	No assumptions
Can be predictive	Typically not predictive
Rigorous	Not rigorous
Less useful for genomic analyses	Useful for genomic analyses

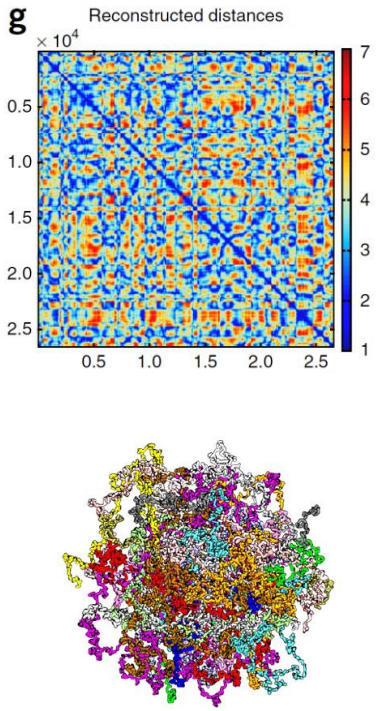
Structural approaches: 2 examples

"Model driven"



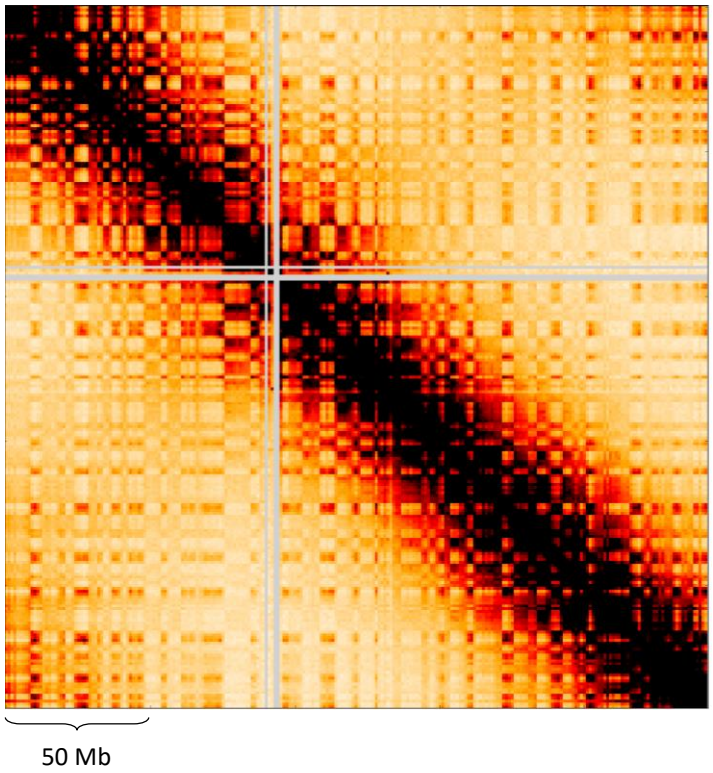
Benedetti et al., *Nucleic Acids Research* 2013

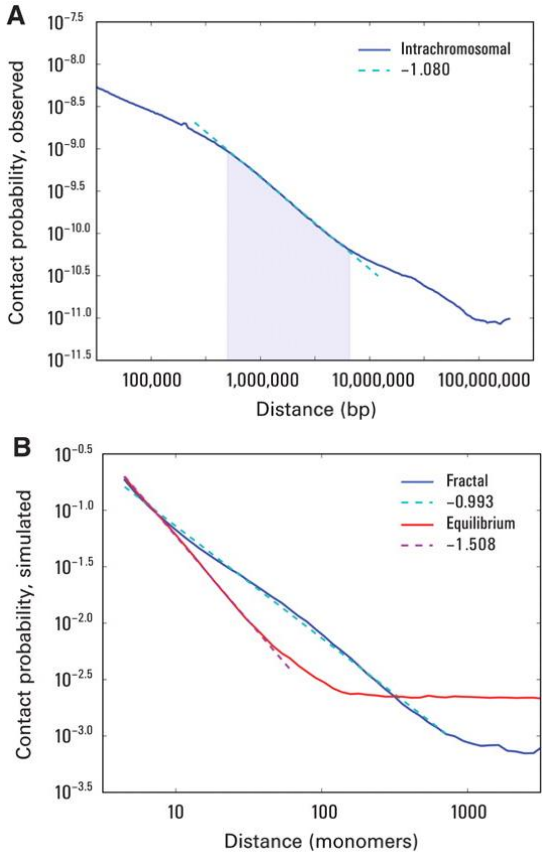
"Data driven"

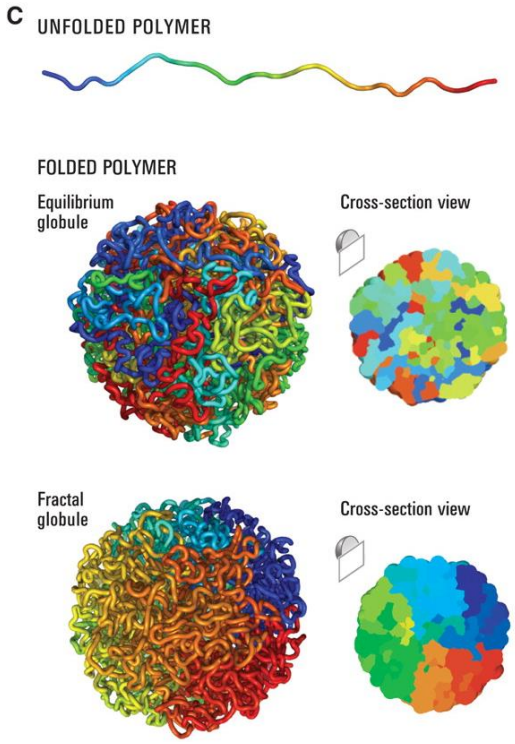


Lesne et al., *Nature Methods* 2013

Distance-dependent interaction

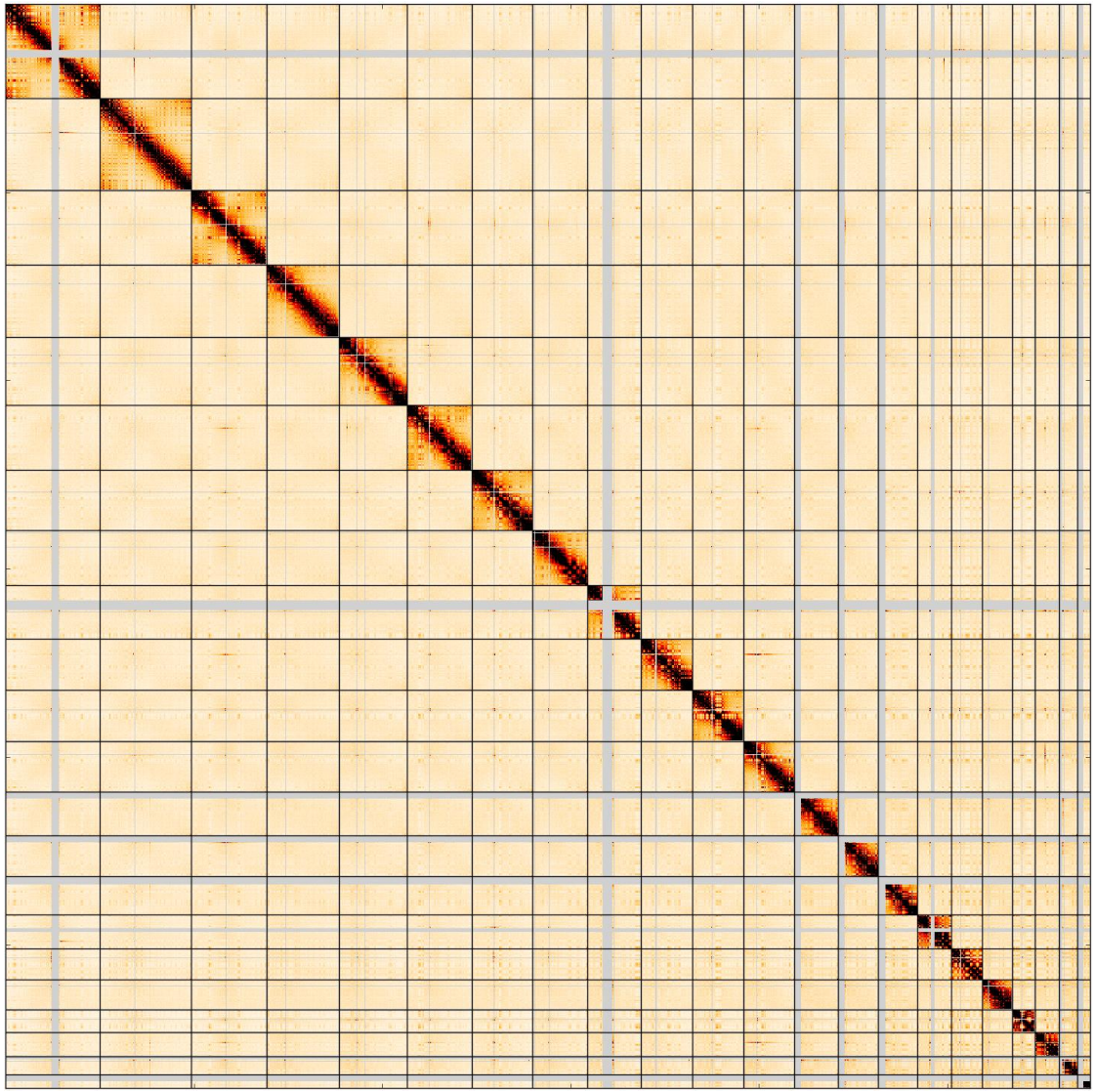


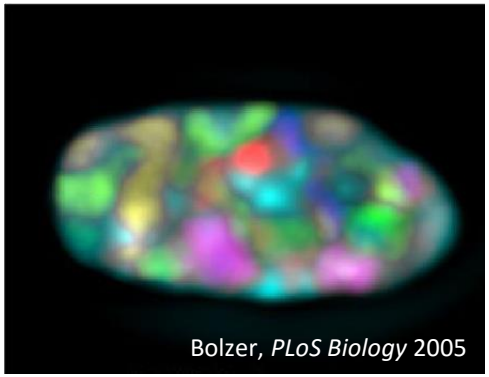


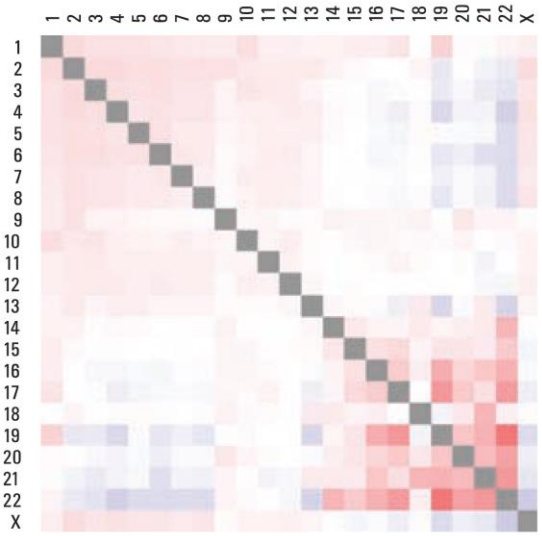


Lieberman-Aiden et al., *Science* 2009

Chromosome territories

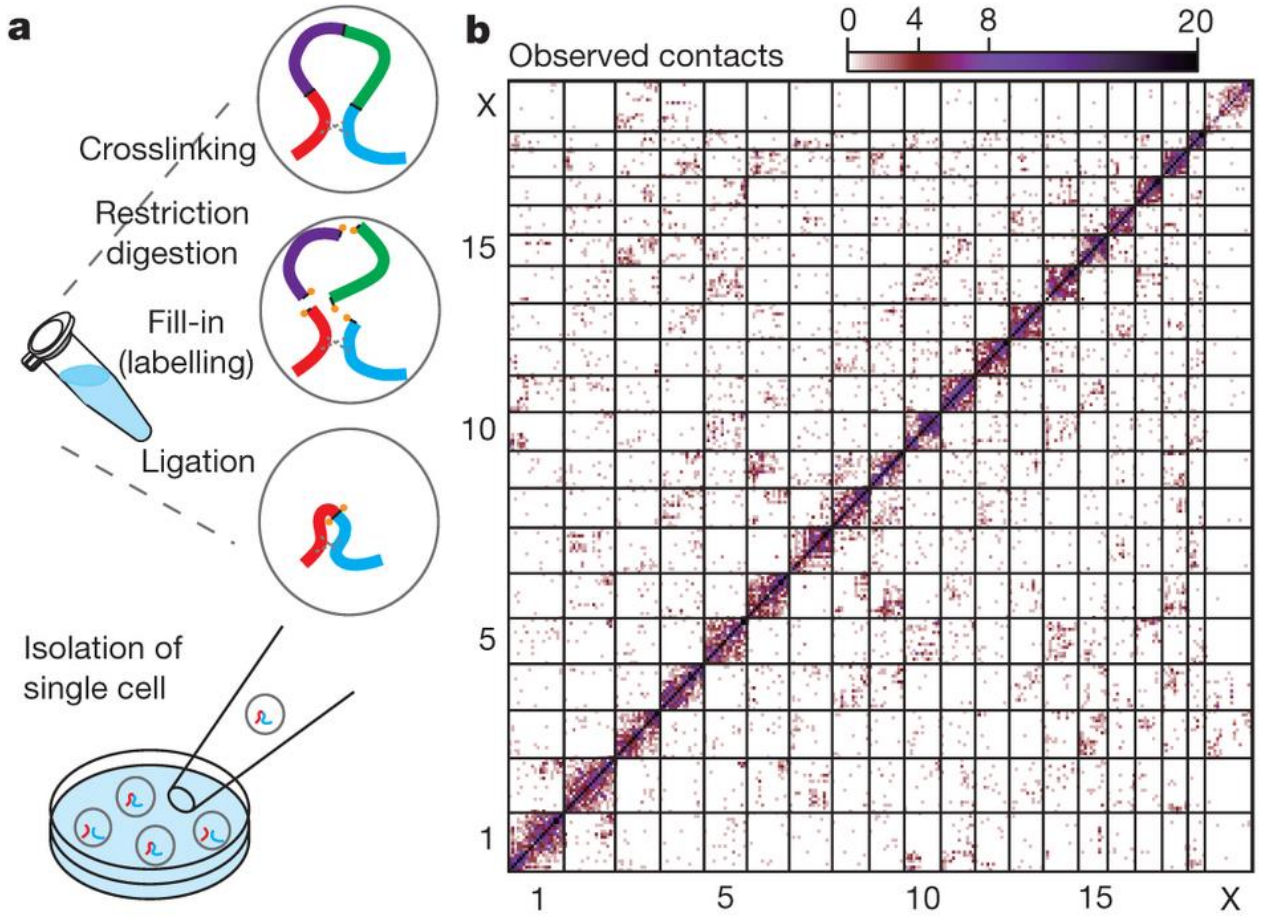






Lieberman-Aiden et al., *Science* 2009

Single cell Hi-C

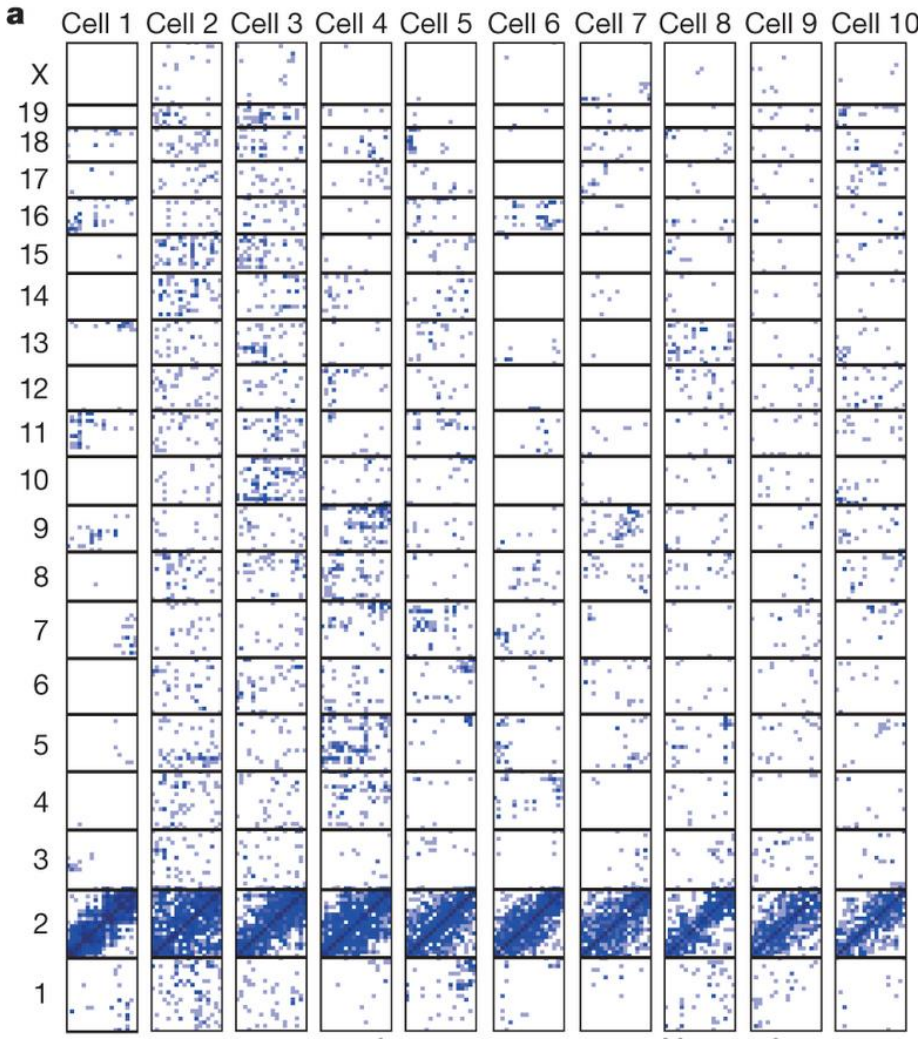


Nagano et al., *Nature* 2013

Caveats:

- Diploid
- ~10K-30K int. per cell

Single cell Hi-C

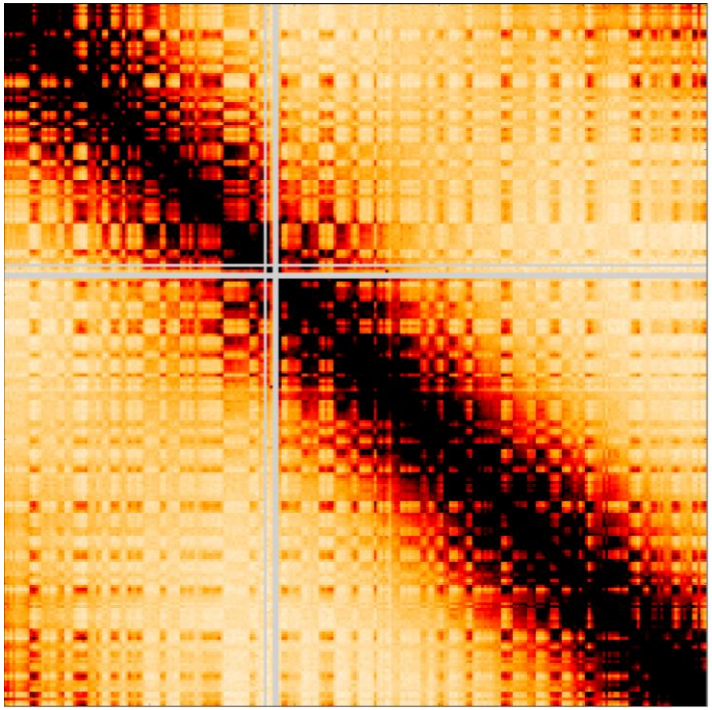


Nagano et al., *Nature* 2013

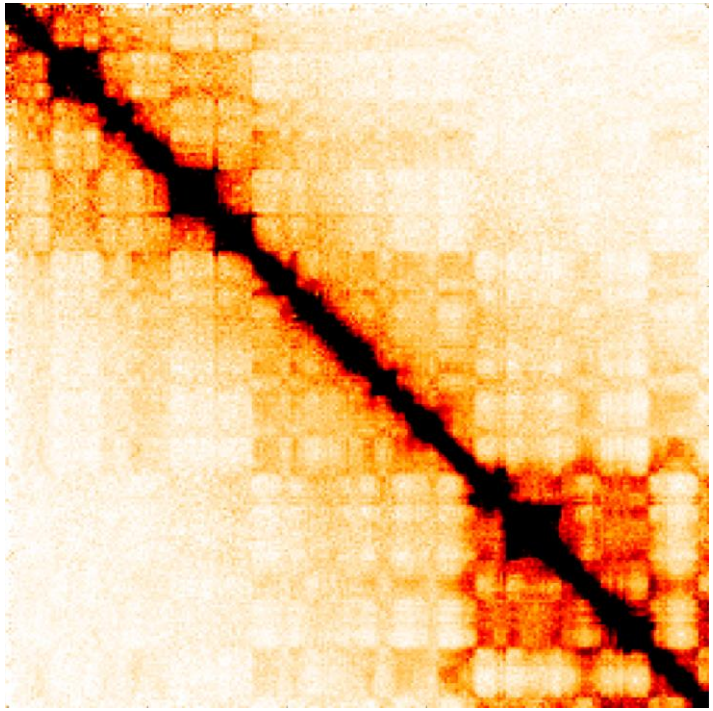
Coming soon (?):

- Haploid cell line
- ~100K-300K int. per cell

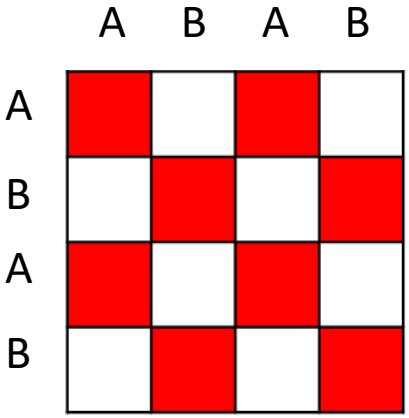
Genomic compartments



50 Mb

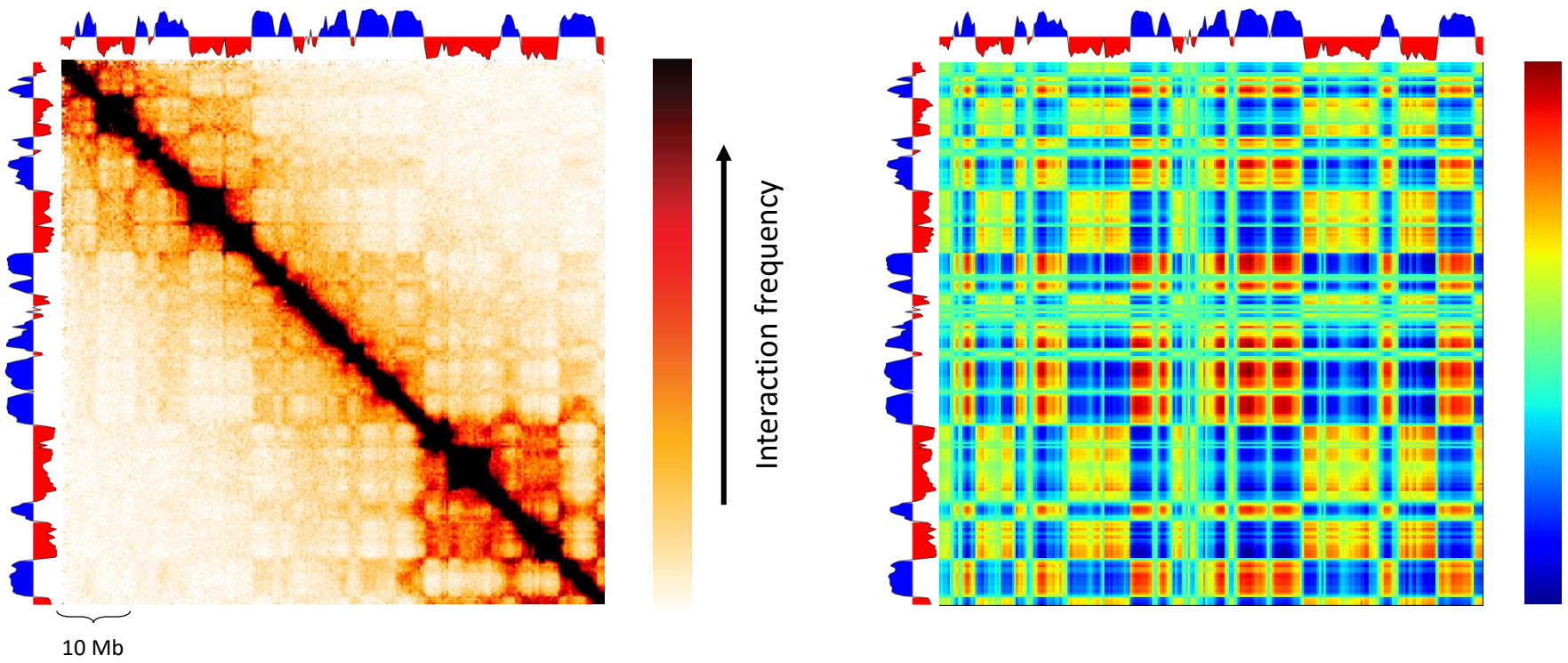


10 Mb

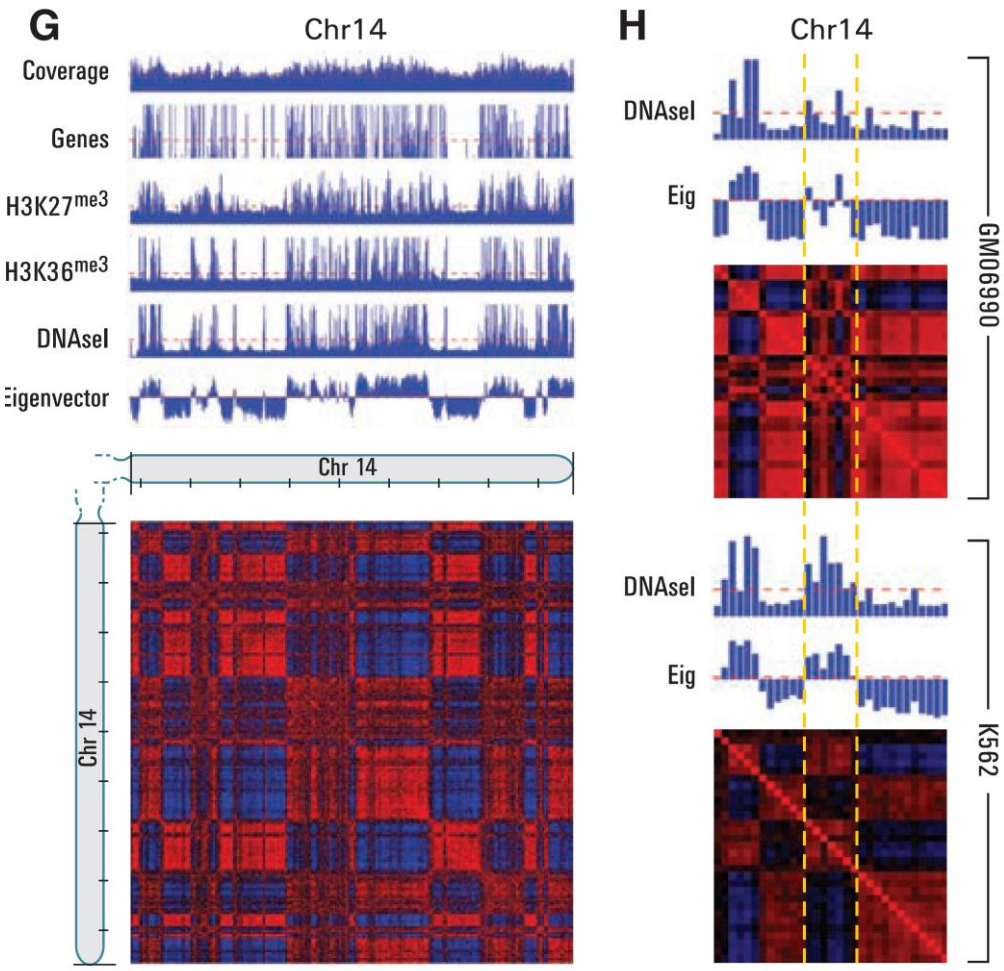


Finding genomic compartments

- Use the first eigenvector v of the PCA/SVD. Why?
- v is the vector that minimizes $\|svv^T - A\|_F$ (Eckart-Young 1936), or in other words $A_{i,j}$ will be "near" sv_iv_j.
- So:
  - if v_i and v_j have the same sign, their product will be positive (high interaction bins i,j)
  - if v_i and v_j have the opposite sign, their product will be negative (low interaction bins i,j)



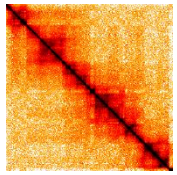
What are genomic compartments?

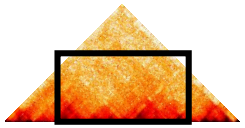


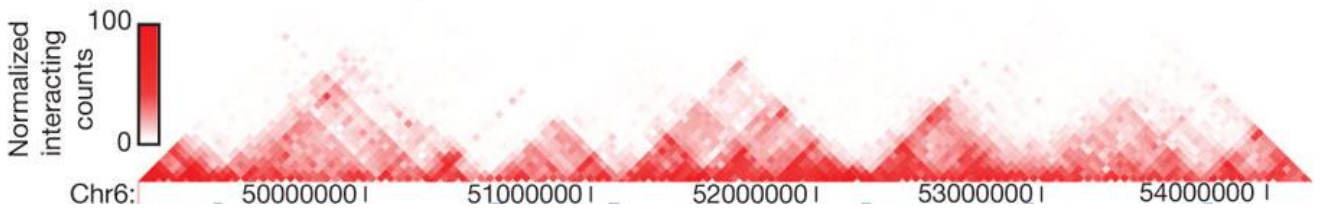
Lieberman-Aiden et al., *Science* 2009

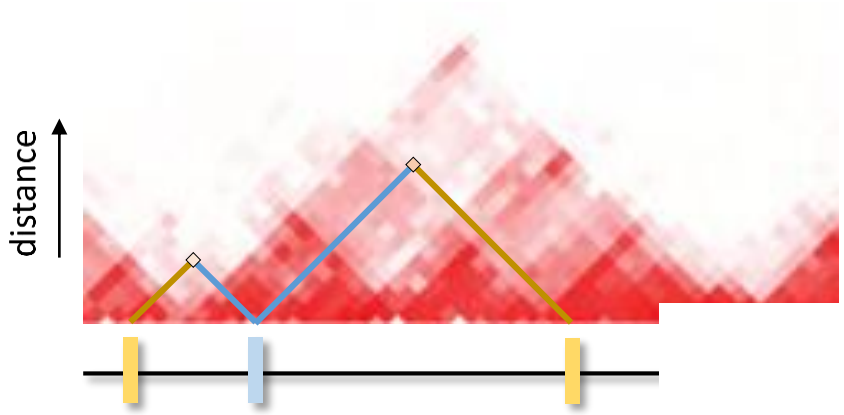
- Correlated with chromatin state
- Cell type-specific
- Structure unclear
- Probably variable on single-cell level

Topologically Associating Domains (TADs)



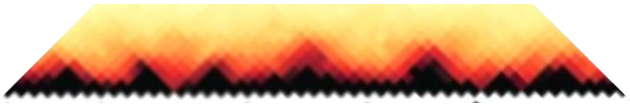


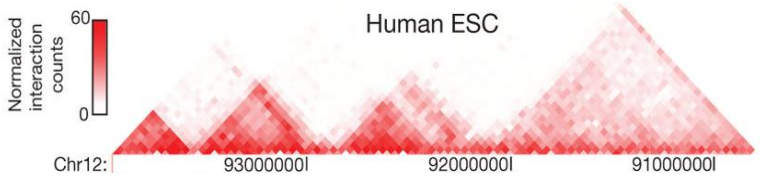




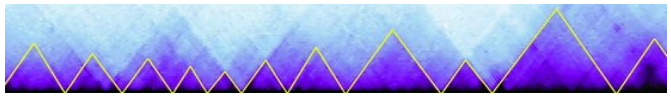
Dixon et al., *Nature* 2012
Nora et al., *Nature* 2012
Sexton et al., *Cell* 2012

Schizosaccharomyces pombe (Mizuguchi et al., *Nature* 2014)



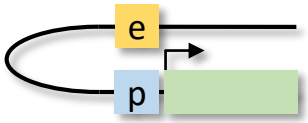


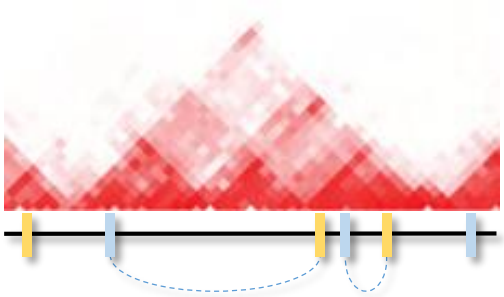
Caulobacter crescentus (Le et al., *Science* 2013)

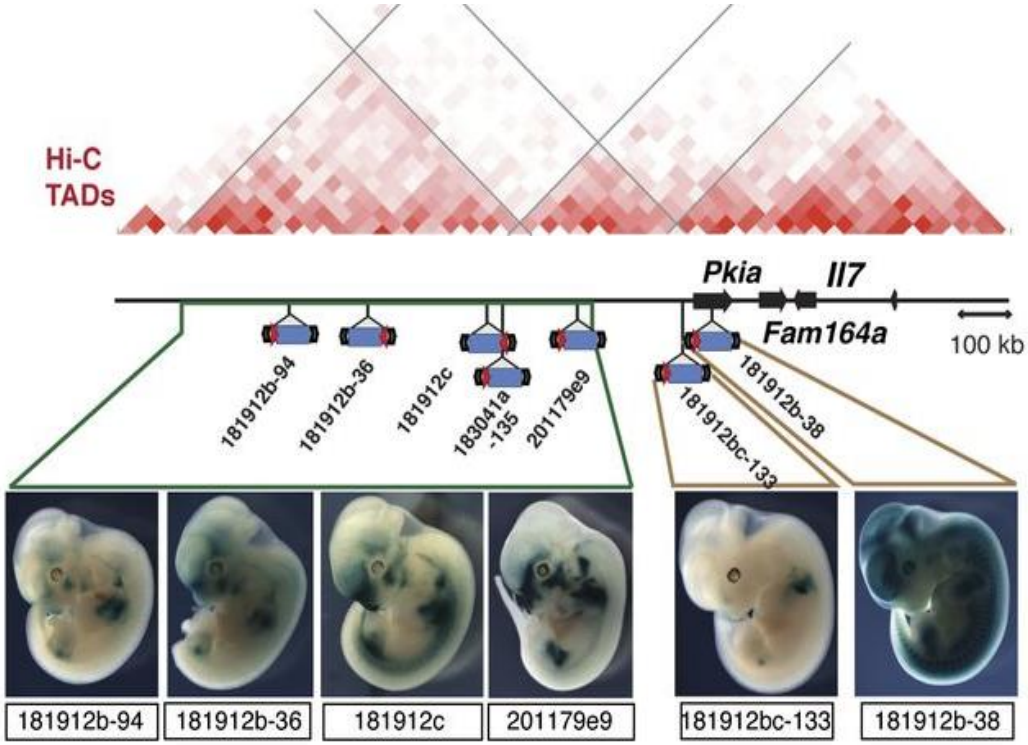


Dixon et al., *Nature* 2012

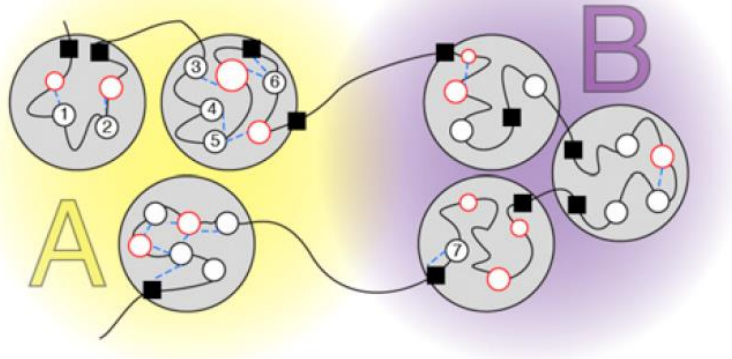
TADs in gene regulation

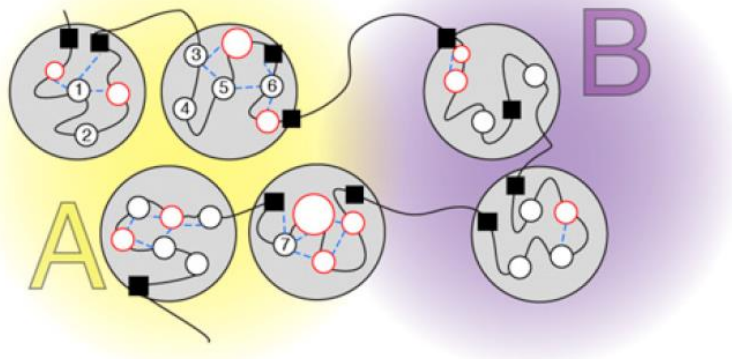






Symmons, *Genome Research* 2014



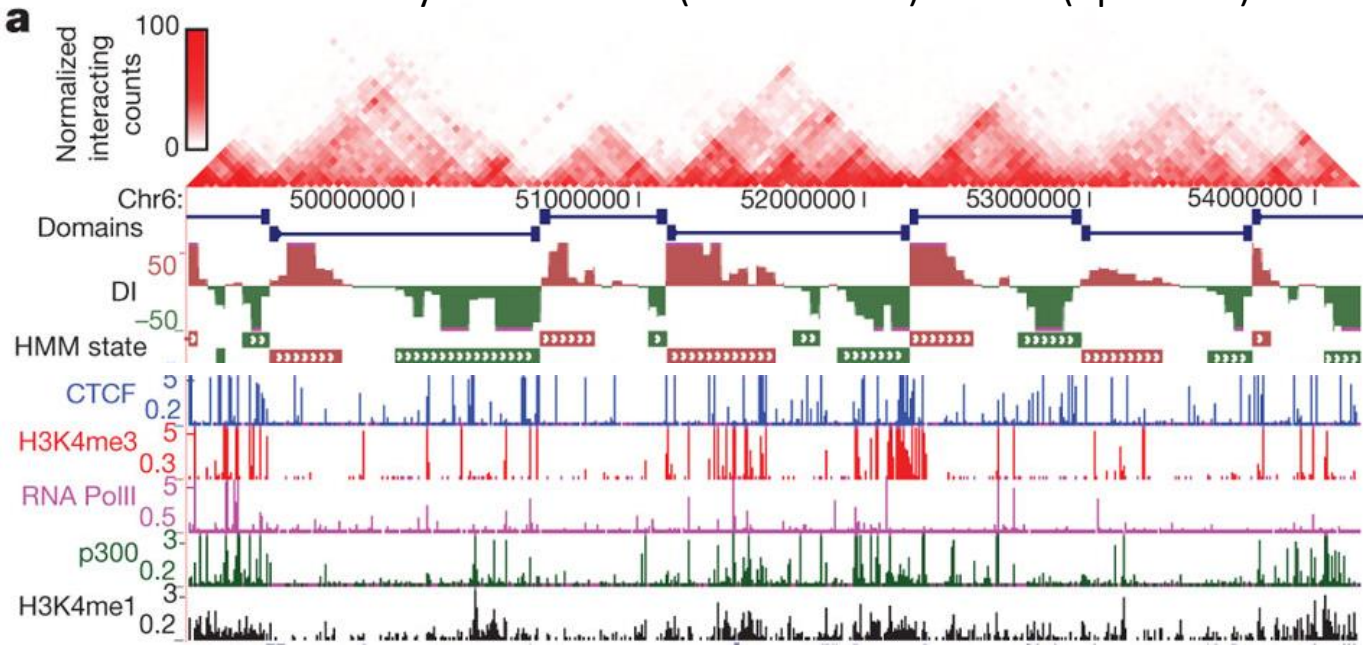


Gibcus et al., *Molecular Cell* 2015

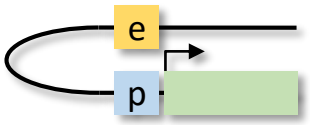
Finding TADs

Dixon et al., *Nature* 2012

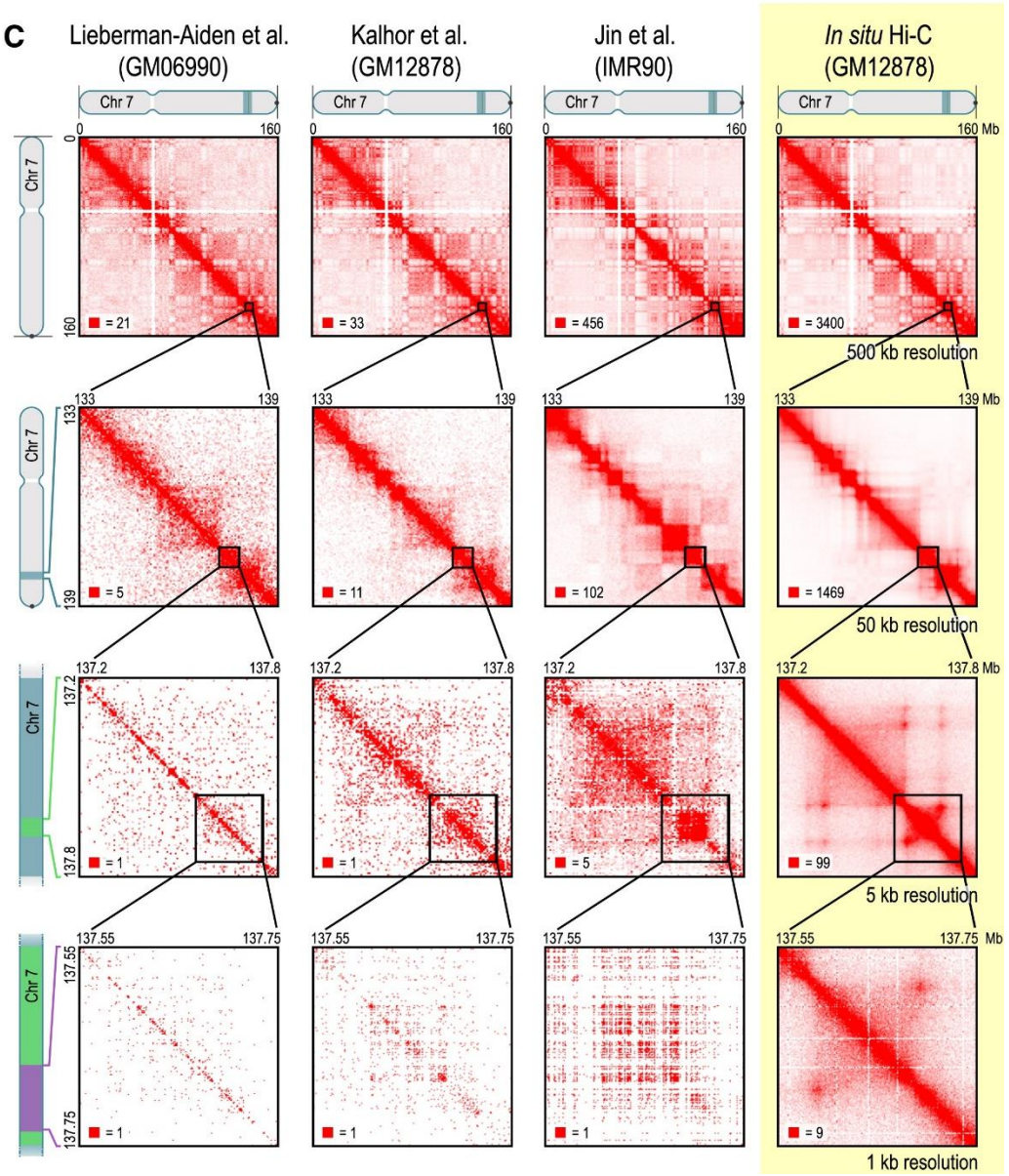
Directionality Index = mean(downstream) – mean(upstream)



Point interactions

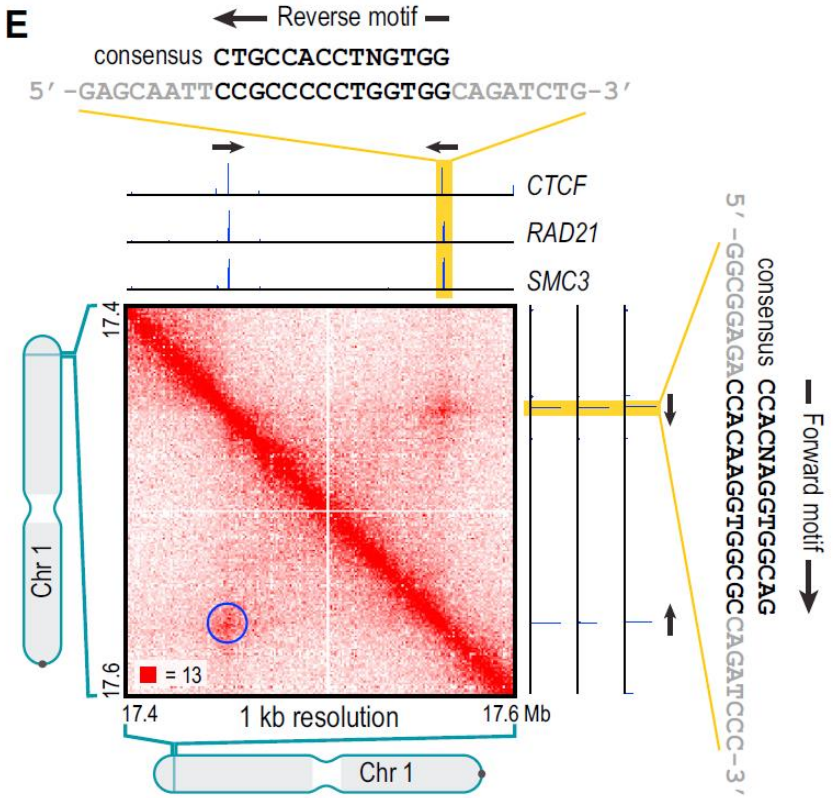


- Resolution problem
- Solutions:
  - Reduced interaction space (e.g. 5C)
  - More sequencing (& 4-cutter)



Rao et al., *Cell* 2015

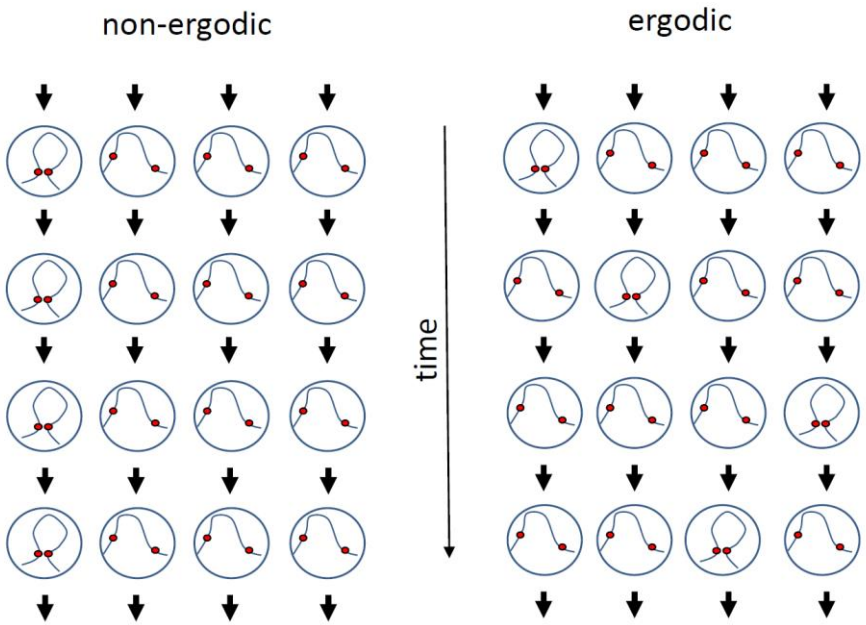
What drives point interactions?



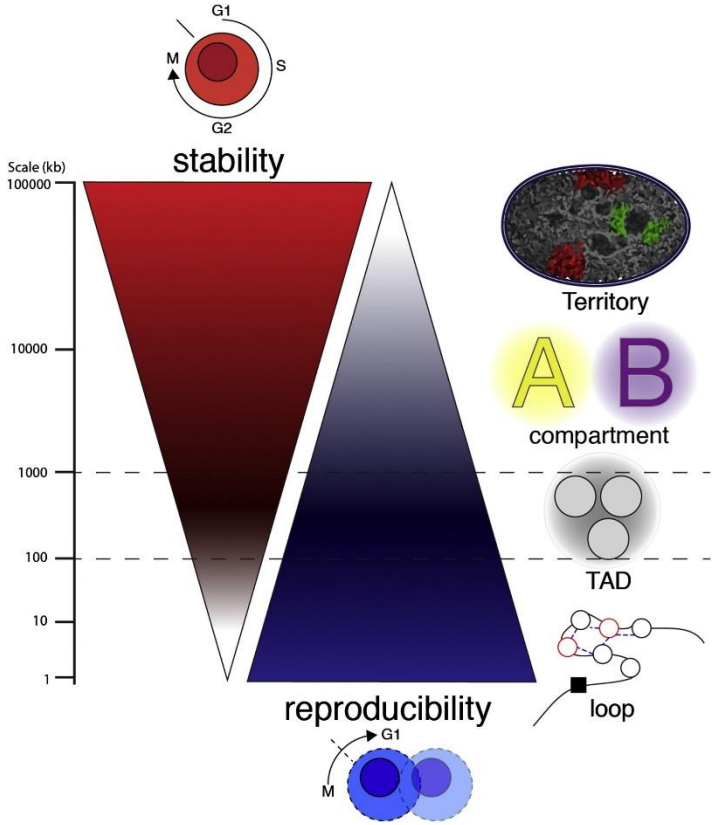
Rao et al., *Cell* 2015

- How to identify point interactions?
- Are these all the point interactions? (Rao et al.: 3K-8K; 30% P-E)
- Stability between cell types
- Relation to gene expression unclear

Chromatin dynamics



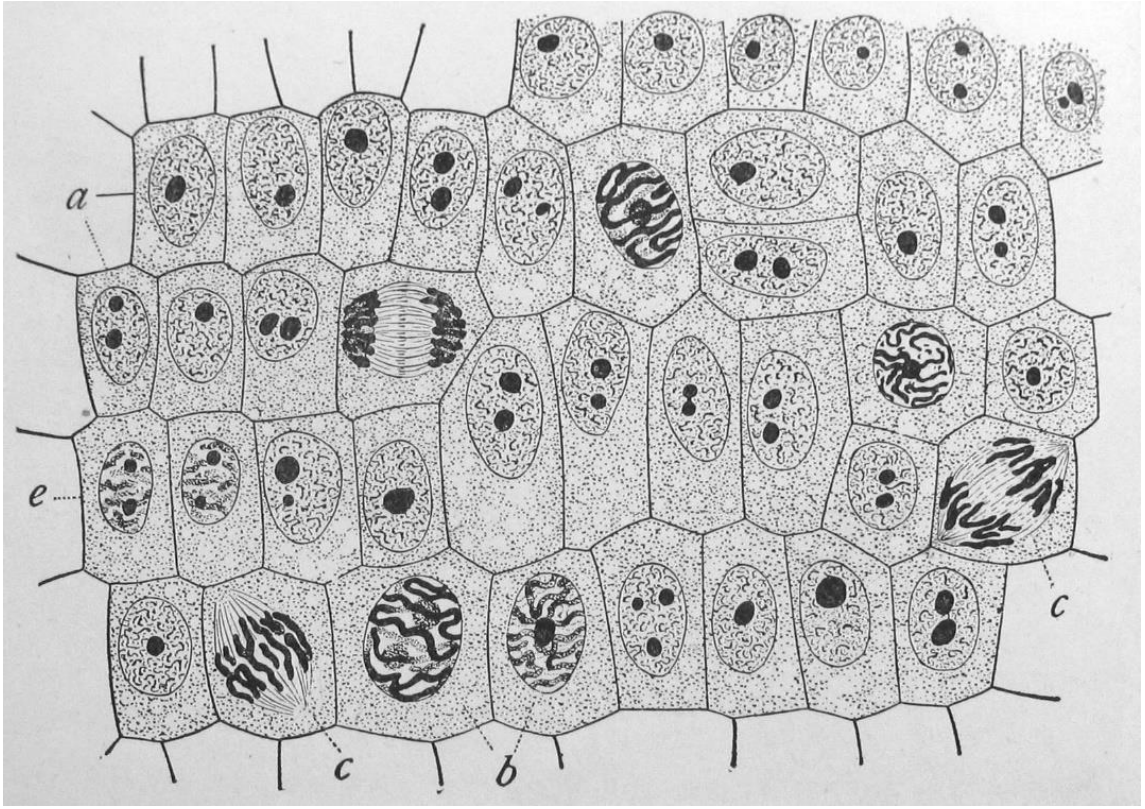
Lajoie et al., *Methods* 2015



Gibcus et al., *Molecular Cell* 2015

Generally Hi-C does not provide dynamics

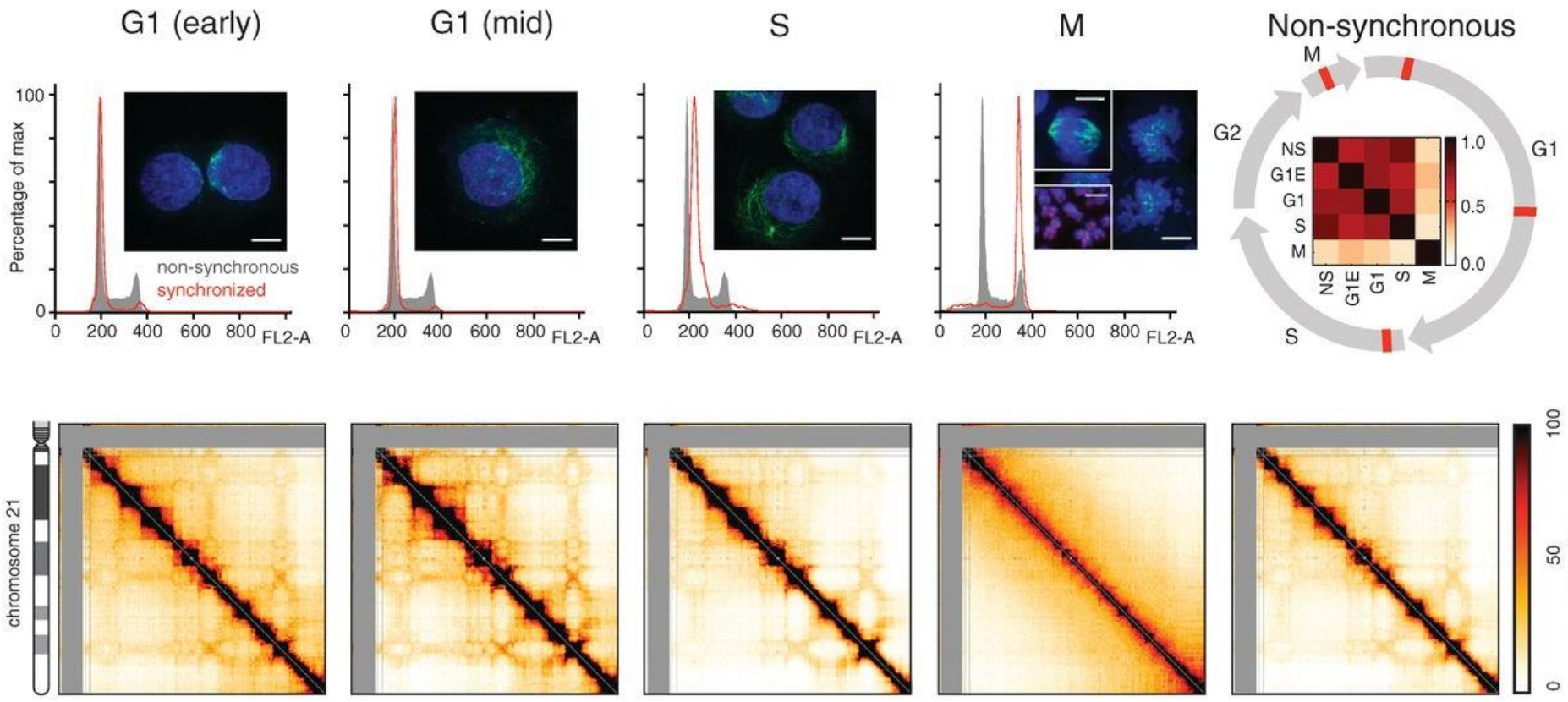
Back to the cell cycle



Wilson, *The Cell* 1900

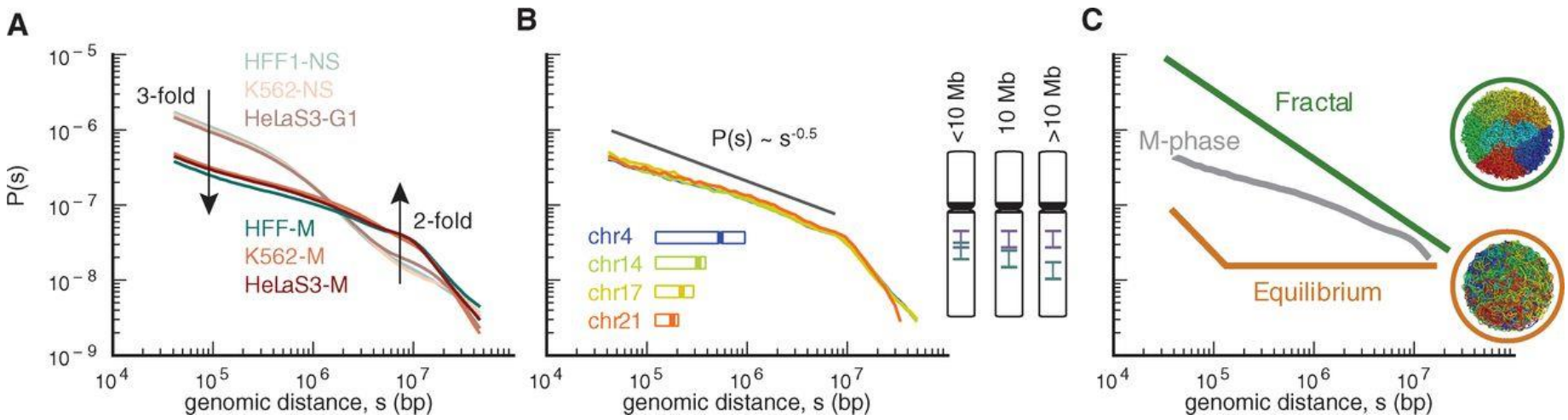
Cell cycle Hi-C

Naumova et al., *Science* 2013

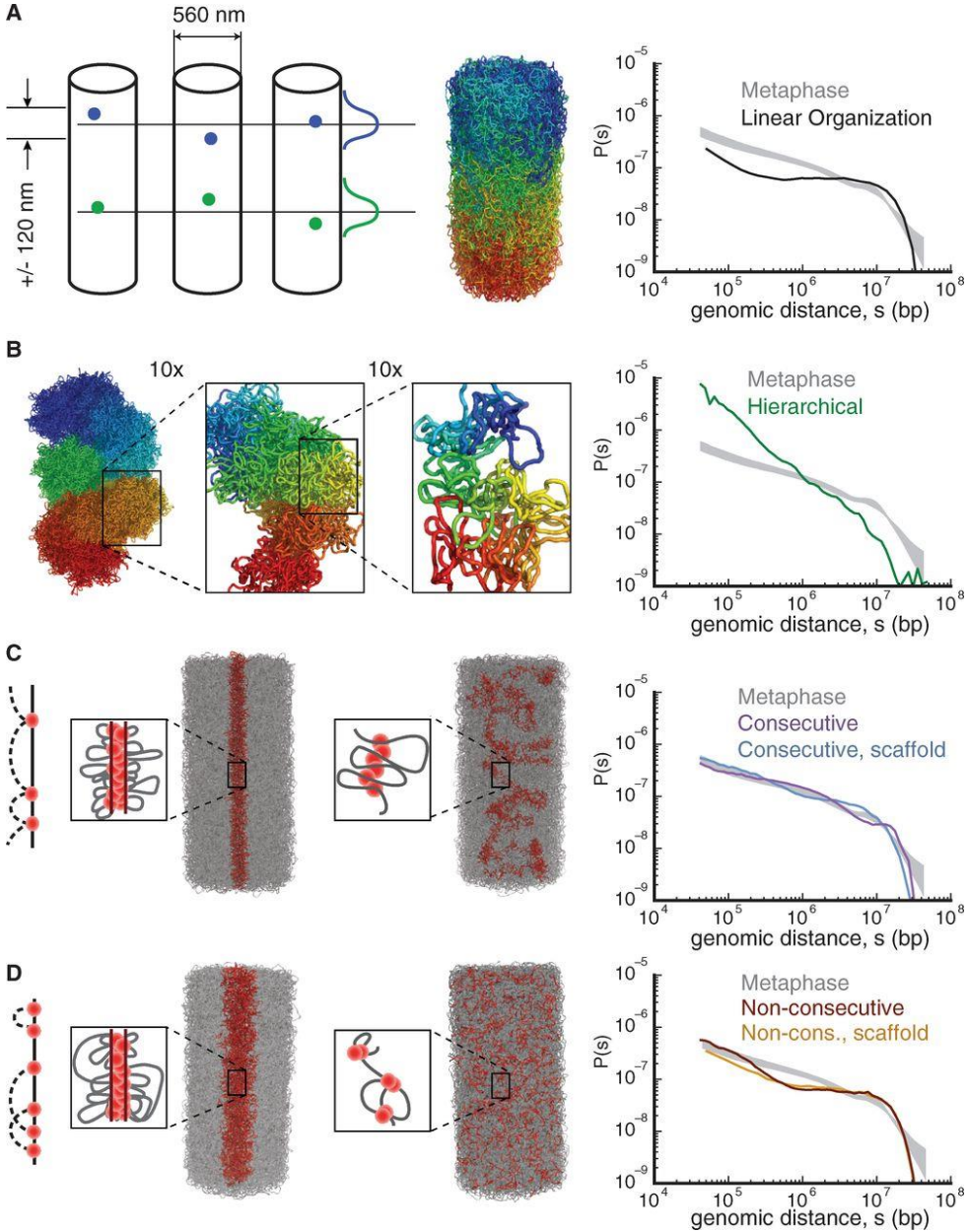


Modelling metaphase chromosomes

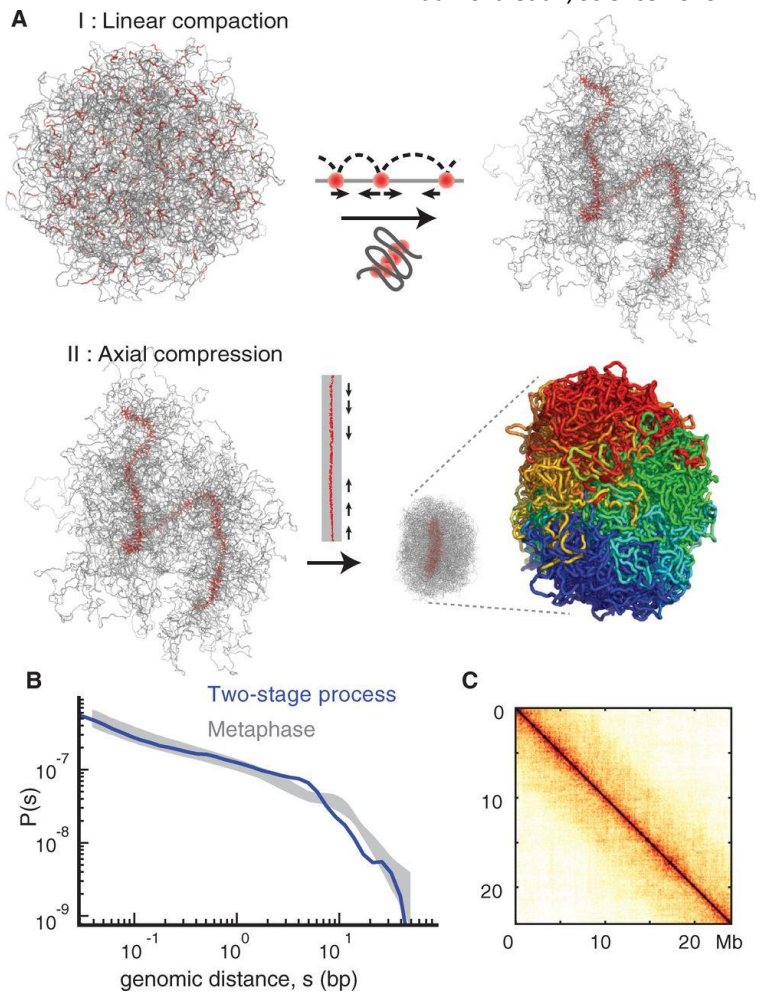
Naumova et al., *Science* 2013



Modelling metaphase chromosomes



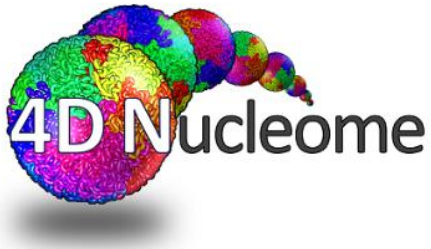
Naumova et al., *Science* 2013



How is structure re-established?

Questions and challenges

- What are the patterns/structures we observe?
- How and when are structures established?
- How do the structures relate to each other?
- How are structures related to function? Causality?
- How is functional robustness achieved for highly variable interactions?
- What are the dynamics? How are functional movements achieved?



Acknowledgements

Prof. Job Dekker

Houda Belaghzal

Johan Gibcus

Bryan R. Lajoie

Sunny Liu

Rachel P. McCord

Marlies Oomen

Anne-Laure Valton

Liyan Yang

Ye Zhan

Some references

Original 3C paper:
Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. Science (New York, N.Y.), 295(5558), 1306–11.
http://doi.org/10.1126/science.1067799

Original Hi-C paper (also relevant for chromosomes territories, genomic compartments and fractal globule):
Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science, 326(5950), 289–293.
http://doi.org/10.1126/science.1181369

Single-cell Hi-C:
Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., … Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature, 502(7469), 59–64.
http://doi.org/10.1038/nature12593

First genome-wide measurement of TAD structures ("topologically associating domains"):
Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398), 376–380.
http://doi.org/10.1038/nature11082

High resolution Hi-C (where point interactions are observed):
Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7), 1665–80.
http://doi.org/10.1016/j.cell.2014.11.021

Cell cycle Hi-C (genomic structures disappear when chromosomes are condensed):
Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A., & Dekker, J. (2013). Organization of the Mitotic Chromosome. Science (New York, N.Y.), 342, 948–53.
http://doi.org/10.1126/science.1236083

Some references

Biological review of Hi-C/genome structure:
Gibcus, J. H., & Dekker, J. (2013). The Hierarchy of the 3D Genome. Molecular Cell.
http://doi.org/10.1016/j.molcel.2013.02.011

Review on structural modeling of Hi-C:
Imakaev, M. V., Fudenberg, G., & Mirny, L. A. (2015). Modeling chromosomes: Beyond pretty pictures. FEBS Letters, 589(20), 3031–3036.
http://doi.org/10.1016/j.febslet.2015.09.004

Overview of Hi-C data processing and analysis (descriptive):
Lajoie, B. R., Dekker, J., & Kaplan, N. (2014). The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. Methods, 72, 65–75
http://doi.org/10.1016/j.ymeth.2014.10.031

Solving problems in 1D genome assembly by using Hi-C data:
Kaplan, N., & Dekker, J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. Nature Biotechnology, 31, 1143–1147.
http://doi.org/10.1038/nbt.2768

Effect of DNA sequence on genomic nucleosome organization:
Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., … Segal, E. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature, 458(7236), 362–6.
http://doi.org/10.1038/nature07667