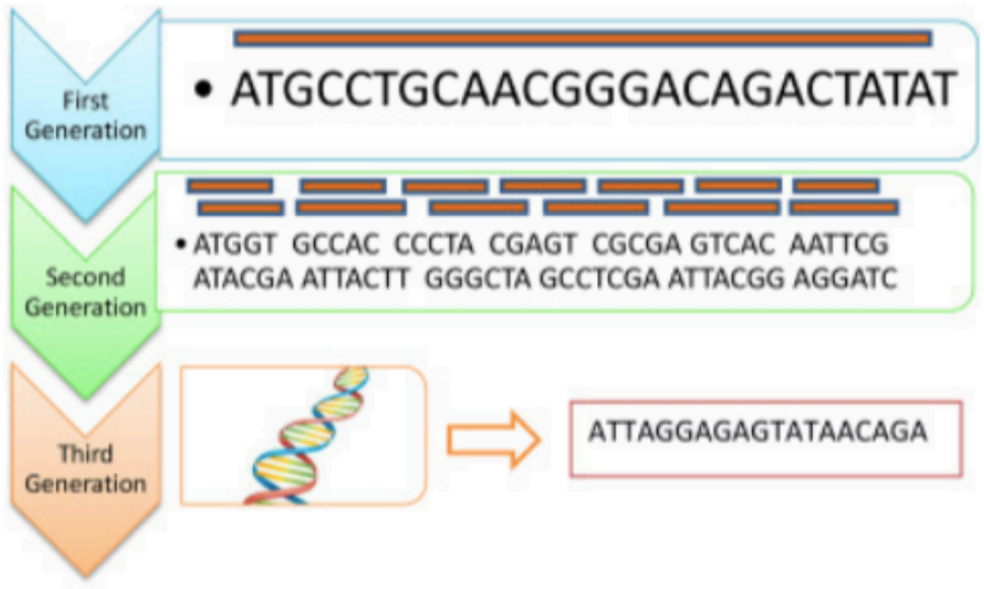
Bioinformatics Lab: EpiPractice1
NGS for chromatin structure analysis

Aleksandra Galitsyna
Aleksandra.Galitsyna@skoltech.ru

2 Apr 2019

Skoltech
Skolkovo Institute of Science and Technology

Introduction: Sequencing

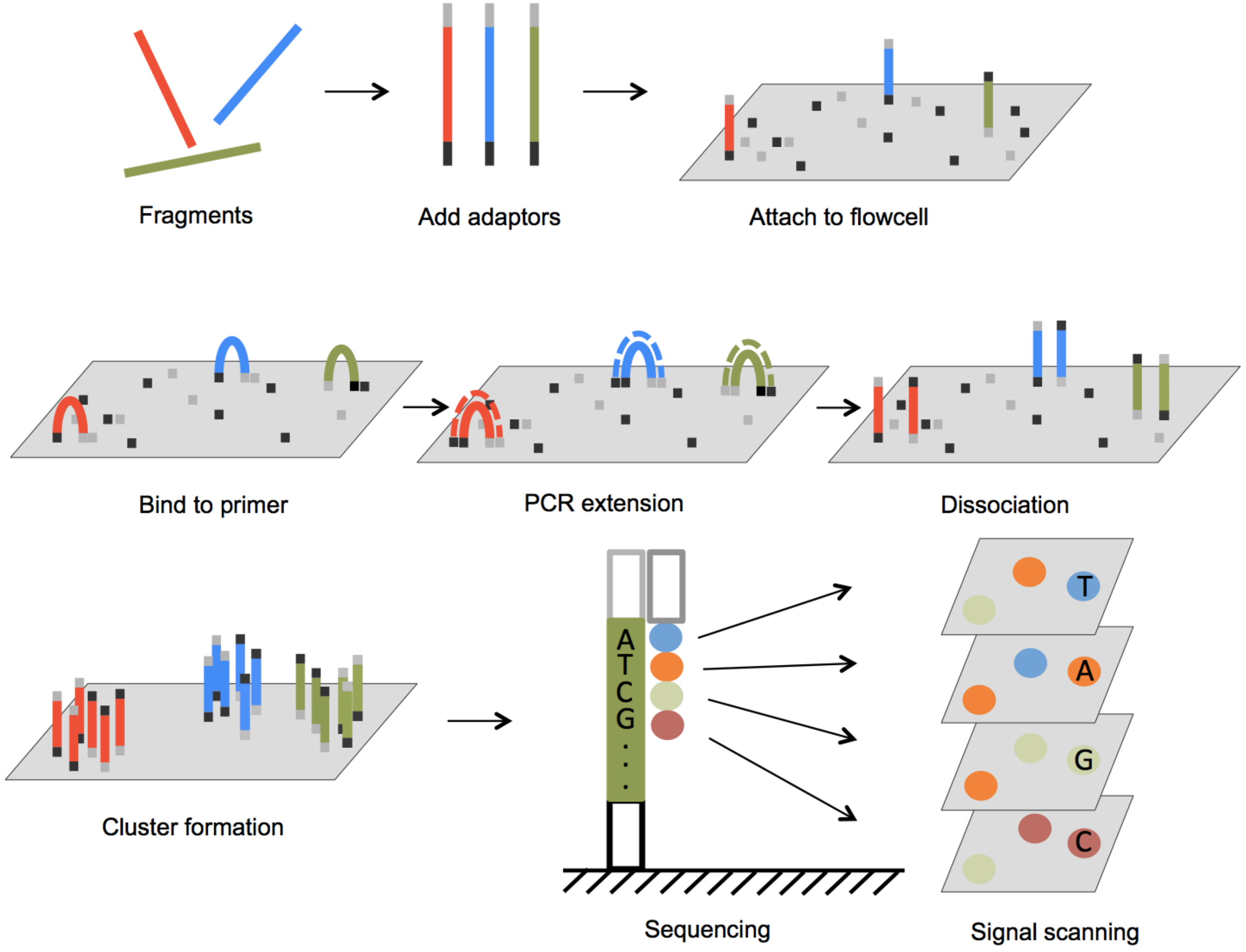

First Generation
• ATGCCTGCAACGGGACAGACTATAT
Second Generation
• ATGGT GCCAC CCCTA CGAGT CGCGA GTCAC AATTCG
ATACGA ATTACTT GGGCTA GCCTCGA ATTACGG AGGATC
Third Generation
ATTAGGAGAGTATAACAGA

Sanger

Illumina, Roche 454, SOLiD

Nanopore

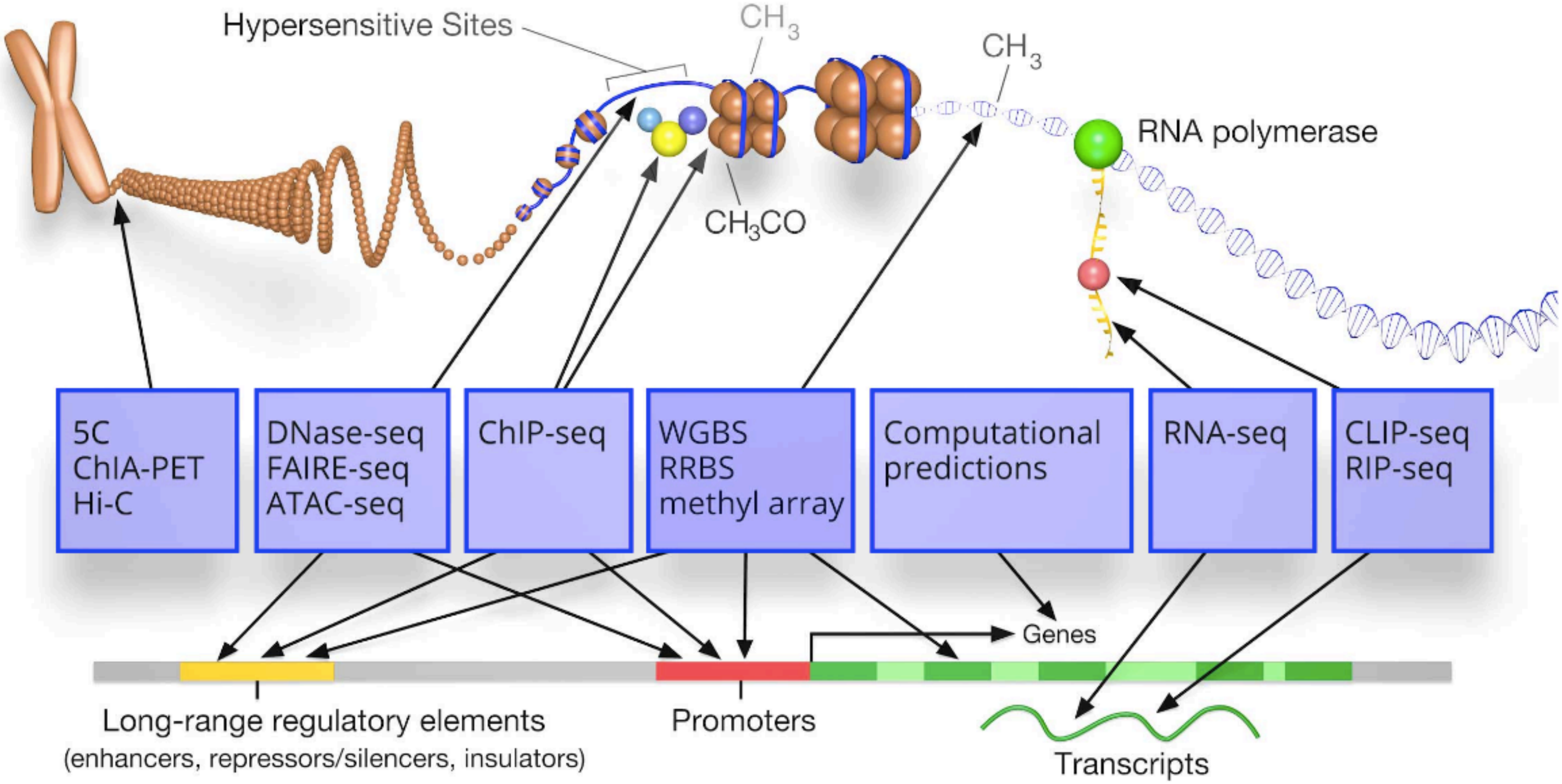
Illumina example:



"Computational analysis of next generation sequencing data and its applications in clinical oncology" Wadapukar&Vyas 2018

Databases for sequencing data

- GEO
- SRA
- ArrayExpress
- modENCODE
- ENCODE



5C
ChIA-PET
Hi-C

DNase-seq
FAIRE-seq
ATAC-seq

ChIP-seq

WGBS
RRBS
methyl array

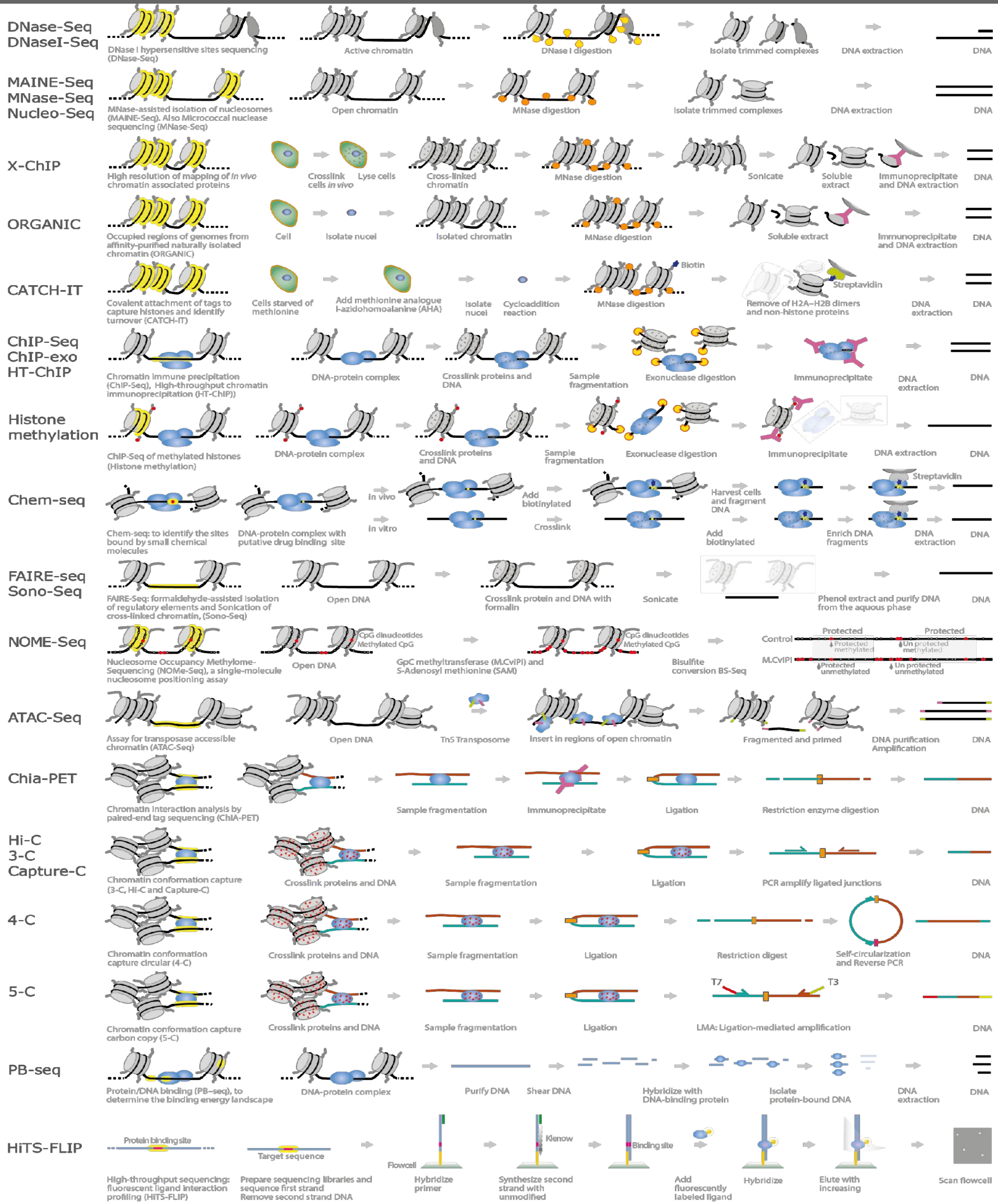
Computational predictions

RNA-seq

CLIP-seq
RIP-seq

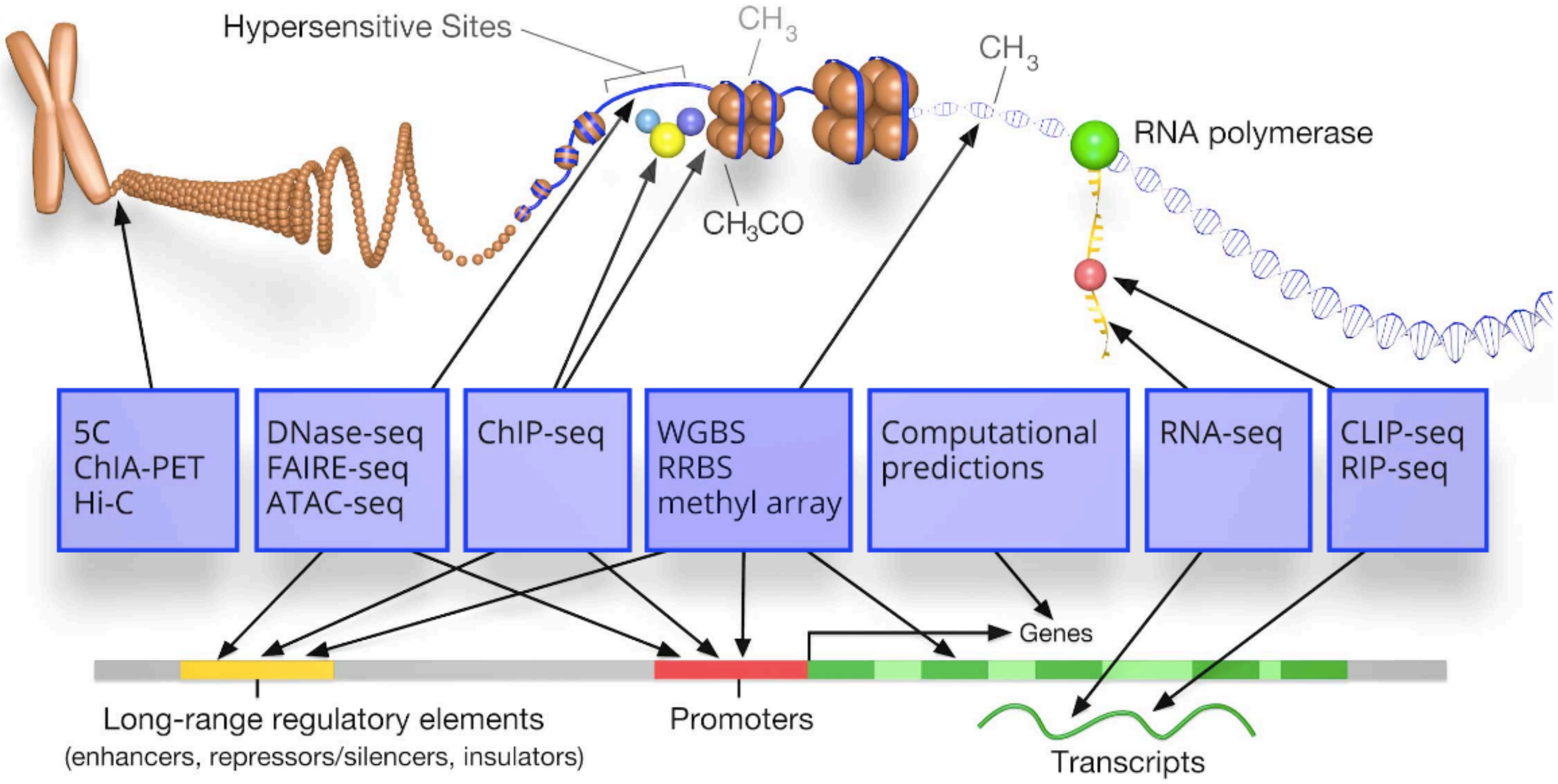
Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

NGS techniques diversity



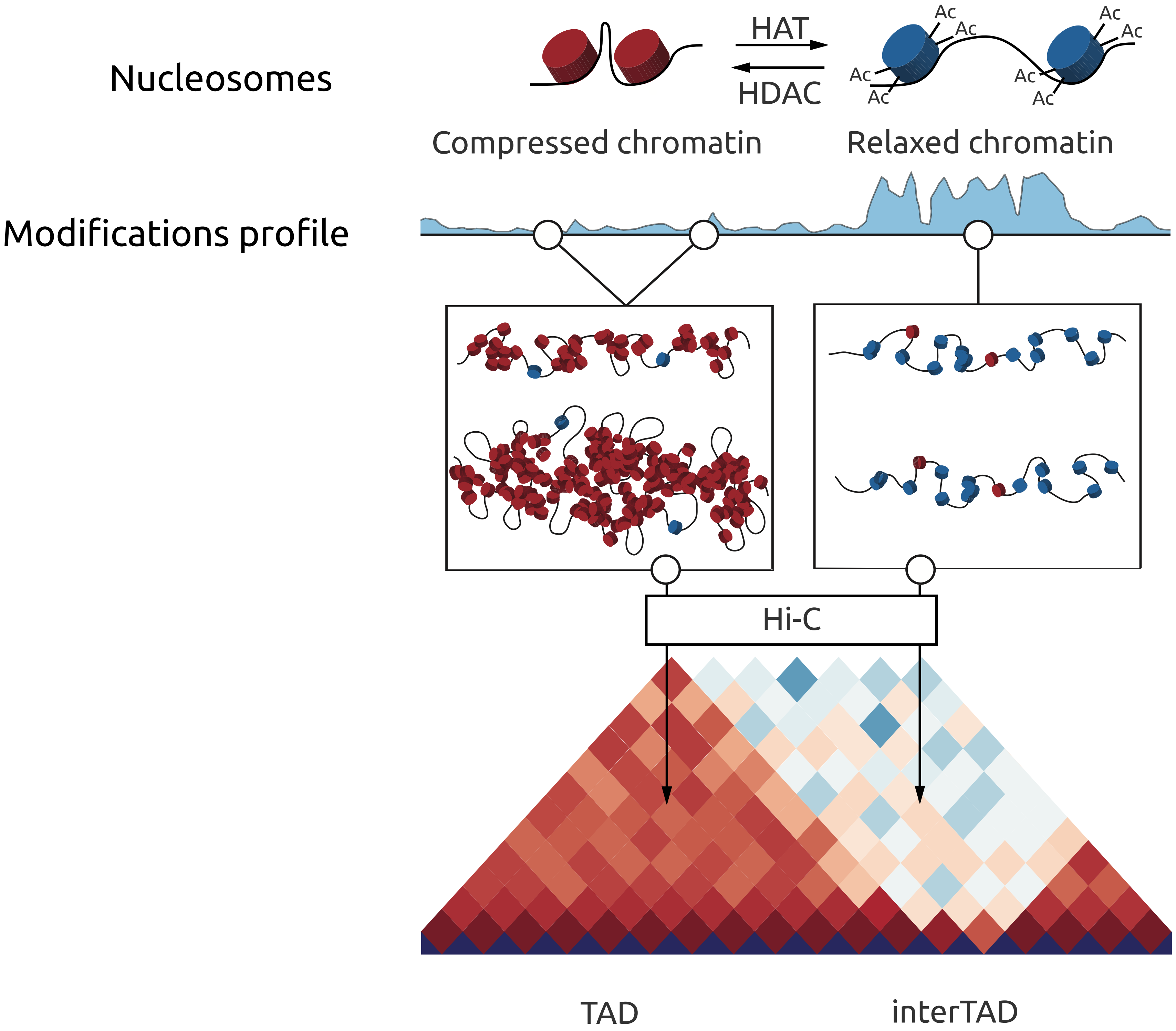
DNase-Seq
DNaseI-Seq
DNase I hypersensitive sites sequencing (DNase-Seq)
Active chromatin
DNase I digestion
Isolate trimmed complexes
DNA extraction
DNA
MAINE-Seq
MNase-Seq
Nucleo-Seq
MNase-assisted isolation of nucleosomes (MAINE-Seq). Also Micrococcal nuclease sequencing (MNase-Seq)
Open chromatin
MNase digestion
Isolate trimmed complexes
DNA extraction
DNA
X-ChIP
High resolution of mapping of in vivo chromatin associated proteins
Crosslink cells in vivo
Lyse cells
Cross-linked chromatin
MNase digestion
Sonicate
Soluble extract
Immunoprecipitate and DNA extraction
DNA
ORGANIC
Occupied regions of genomes from affinity-purified naturally isolated chromatin (ORGANIC)
Cell
Isolate nucei
Isolated chromatin
MNase digestion
Soluble extract
Immunoprecipitate and DNA extraction
DNA
CATCH-IT
Covalent attachment of tags to capture histones and identify turnover (CATCH-IT)
Cells starved of methionine
Add methionine analogue l-azidohomoalanine (AHA)
Isolate nucei
Cycloaddition reaction
MNase digestion
Biotin
Streptavidin
Remove of H2A–H2B dimers and non-histone proteins
DNA extraction
DNA
ChIP-Seq
ChIP-exo
HT-ChIP
Chromatin immune precipitation (ChIP-Seq), High-throughput chromatin immunoprecipitation (HT-ChIP))
DNA-protein complex
Crosslink proteins and DNA
Sample fragmentation
Exonuclease digestion
Immunoprecipitate
DNA extraction
DNA
Histone methylation
ChIP-Seq of methylated histones (Histone methylation)
DNA-protein complex
Crosslink proteins and DNA
Sample fragmentation
Exonuclease digestion
Immunoprecipitate
DNA extraction
DNA
Chem-seq
Chem-seq: to identify the sites bound by small chemical molecules
DNA-protein complex with putative drug binding site
in vivo
in vitro
Add biotinylated
Crosslink
Harvest cells and fragment DNA
Add biotinylated
Enrich DNA fragments
Streptavidin
DNA extraction
DNA
FAIRE-seq
Sono-Seq
FAIRE-Seq: formaldehyde-assisted Isolation of regulatory elements and Sonication of cross-linked chromatin, (Sono-Seq)
Open DNA
Crosslink protein and DNA with formalin
Sonicate
Phenol extract and purify DNA from the aquous phase
DNA
NOME-Seq
Nucleosome Occupancy Methylome-Sequencing (NOMe-Seq), a single-molecule nucleosome positioning assay
Open DNA
CpG dinucleotides
Methylated CpG
GpC methyltransferase (M.CviPI) and S-Adenosyl methionine (SAM)
CpG dinucleotides
Methylated CpG
Bisulfite conversion BS-Seq
Control
M.CviPI
Protected
Protected
Protected methylated
Un protected methylated
Protected unmethylated
Un protected unmethylated
ATAC-Seq
Assay for transposase accessible chromatin (ATAC-Seq)
Open DNA
Tn5 Transposome
Insert in regions of open chromatin
Fragmented and primed
DNA purification Amplification
DNA
Chia-PET
Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)
Sample fragmentation
Immunoprecipitate
Ligation
Restriction enzyme digestion
DNA
Hi-C
3-C
Capture-C
Chromatin conformation capture (3-C, Hi-C and Capture-C)
Crosslink proteins and DNA
Sample fragmentation
Ligation
PCR amplify ligated junctions
DNA
4-C
Chromatin conformation capture circular (4-C)
Crosslink proteins and DNA
Sample fragmentation
Ligation
Restriction digest
Self-circularization and Reverse PCR
DNA
5-C
Chromatin conformation capture carbon copy (5-C)
Crosslink proteins and DNA
Sample fragmentation
Ligation
T7
T3
LMA: Ligation-mediated amplification
DNA
PB-seq
Protein/DNA binding (PB–seq), to determine the binding energy landscape
DNA-protein complex
Purify DNA
Shear DNA
Hybridize with DNA-binding protein
Isolate protein-bound DNA
DNA extraction
DNA
HiTS-FLIP
Protein binding site
Target sequence
High-throughput sequencing: fluorescent ligand interaction profiling (HiTS-FLIP)
Prepare sequencing libraries and sequence first strand Remove second strand DNA
Flowcell
Hybridize primer
Klenow
Synthesize second strand with unmodified
Binding site
Add fluorescently labeled ligand
Hybridize
Elute with increasing
Scan flowcell

Types of epigenetics data



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Interconnected layers of epigenetics data



Nucleosomes

Modifications profile

Compressed chromatin

Relaxed chromatin

Hi-C

TAD

interTAD

How do we measure DNA-DNA interactions?

Ulianov et al. 2016 Genome Biology

Outline: NGS for DNA-DNA interactions

- Introduction
    - Eukaryotic chromatin structure and methods to study it
    - Chromatin interaction map
    - Interaction map features: TADs, compartments, loops
- From theory to practice: Hi-C data processing workflow
    - Reads mapping
    - Binning & filtering
    - Matrix balancing
    - TADs and compartments calling
    - Variety of processing tools
- Some cases from chromatin study practice
- Seminar overview

1. Introduction

Chromatin spatial structure

Introduction: Eukaryotic chromatin structure

Chromatin factors → **Structure** → **Function**

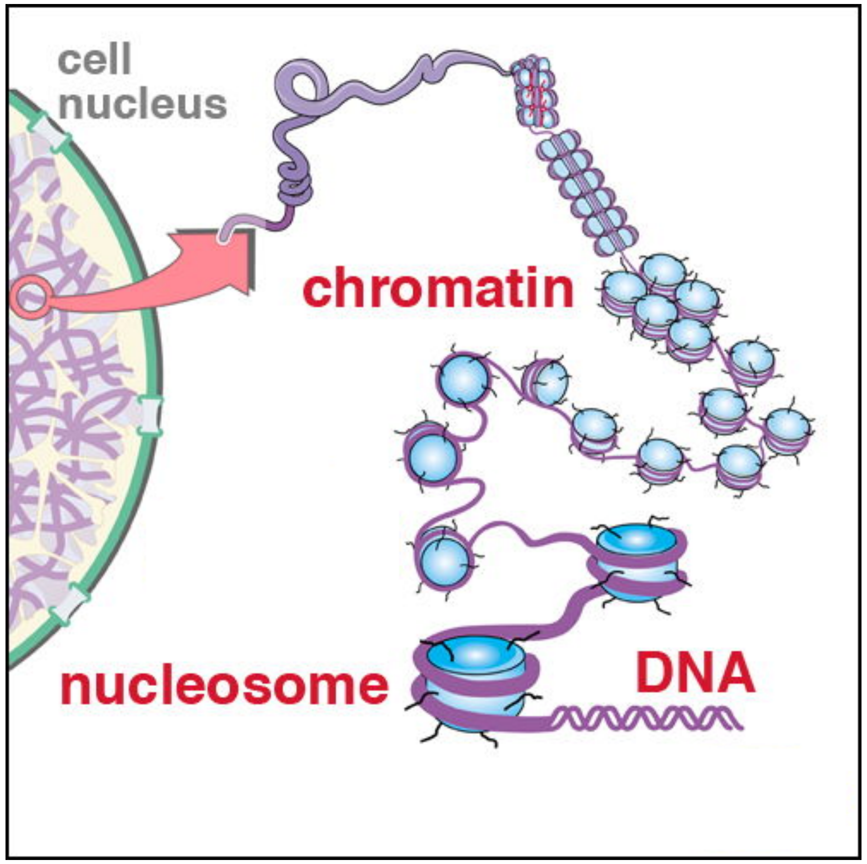
Histone modifications

Transcription factors binding

Non-coding RNAs

Nucleotide modifications

Binding to
the nucleolar envelope



10^4-10^5 folding

Replication

Recombination

Regulation

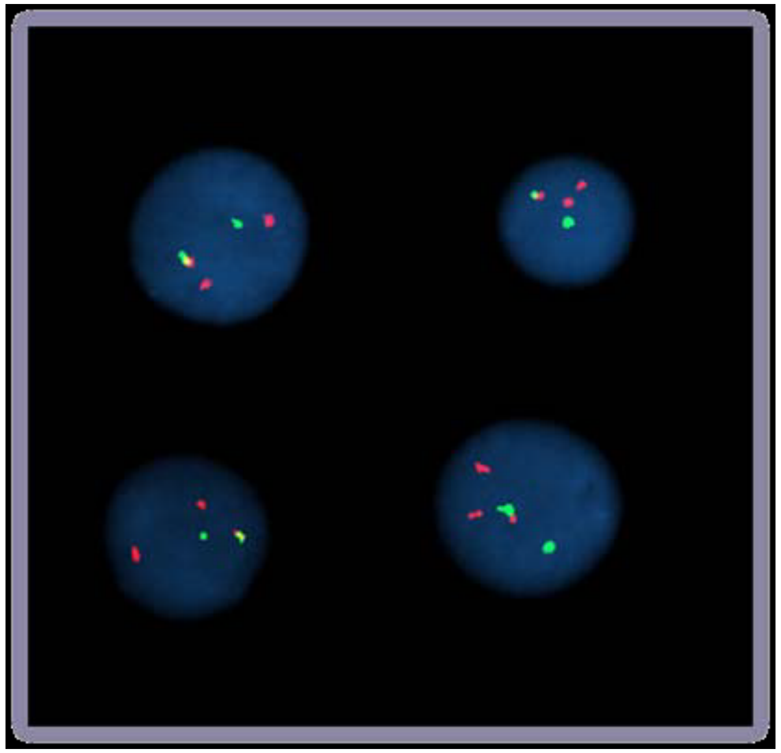
Transcription

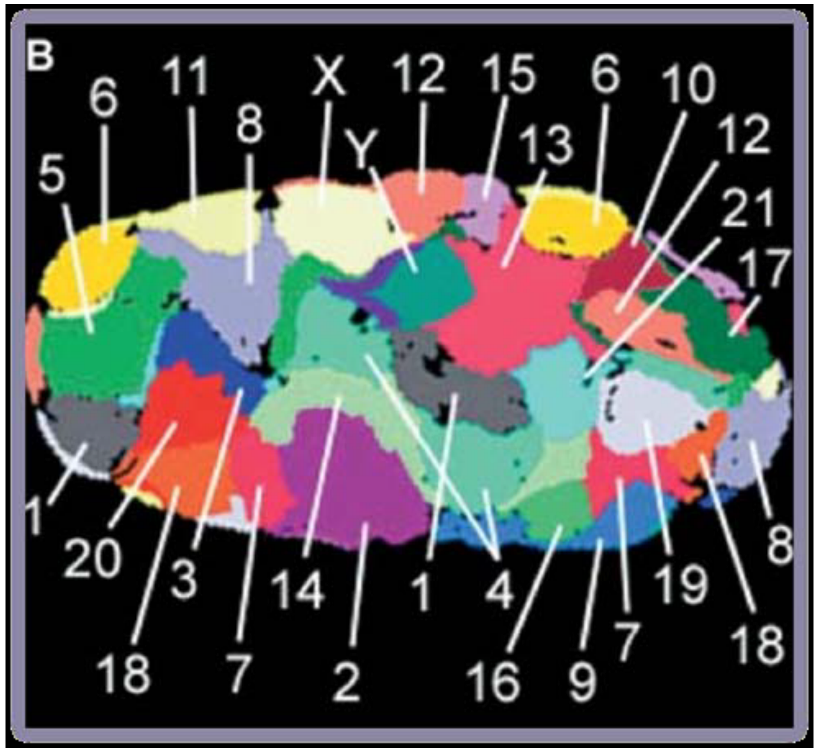
Some methods to probe chromatin structure

Bolzer et al., PLoS Biol. 2005

Some methods to probe chromatin structure

- Microscopy
- FISH (DNA fluorescence in situ hybridization)
- …

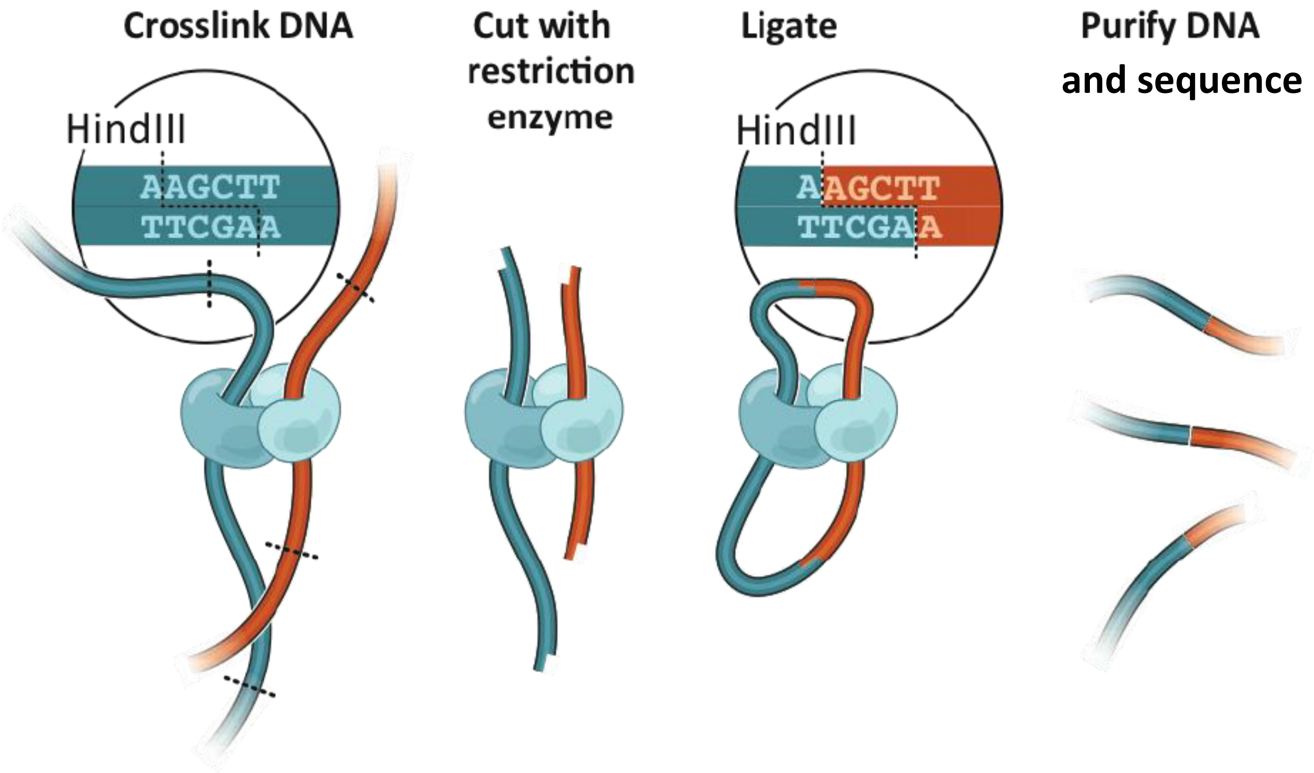




Bolzer et al., PLoS Biol. 2005

Some methods to probe chromatin structure

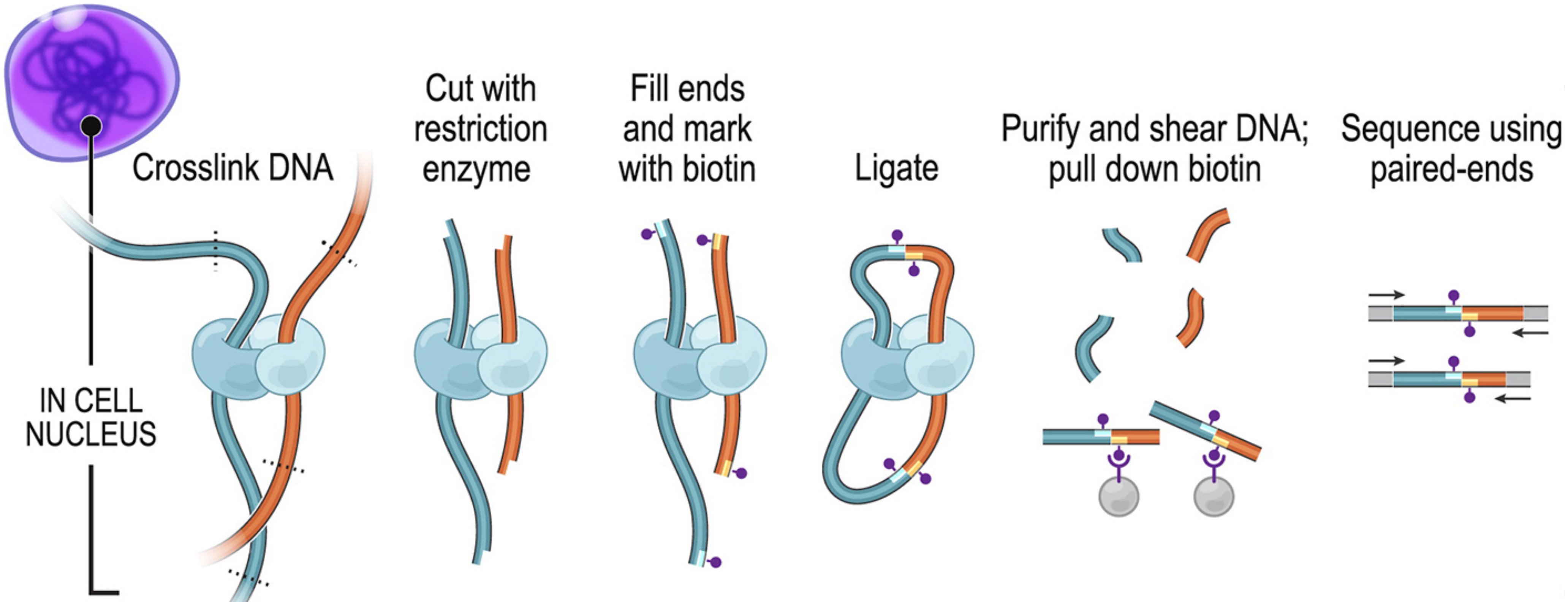
- Microscopy
- FISH (DNA fluorescence in situ hybridization)
- DamID (shows DNA fragments located at the periphery of the nucleus)
- 3C methods



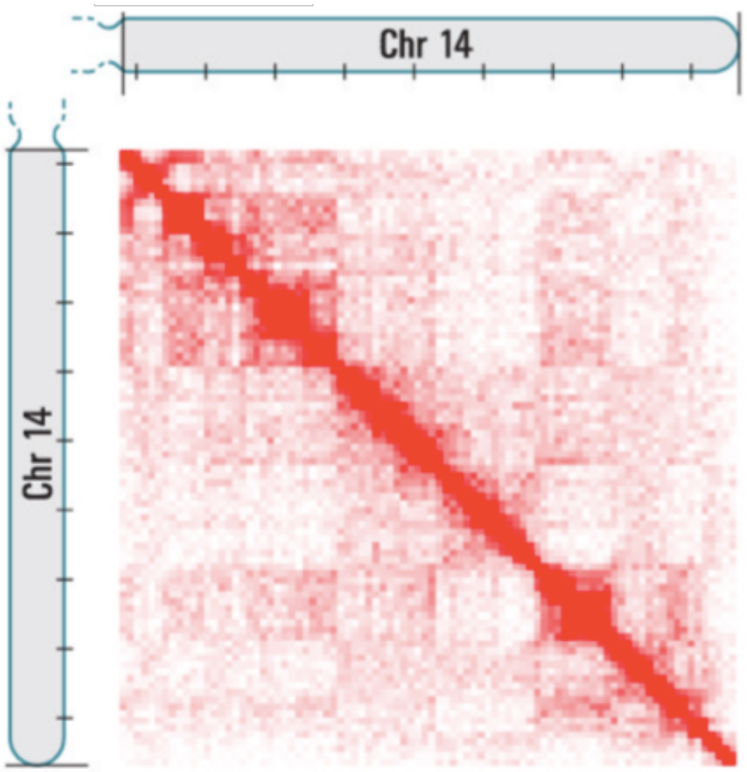
3C: Dekker et al., *Science* 2002

Hi-C: high-throughput chromosomes conformation capture

Procedure:


Crosslink DNA
IN CELL NUCLEUS
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Resulting interactions heatmap:



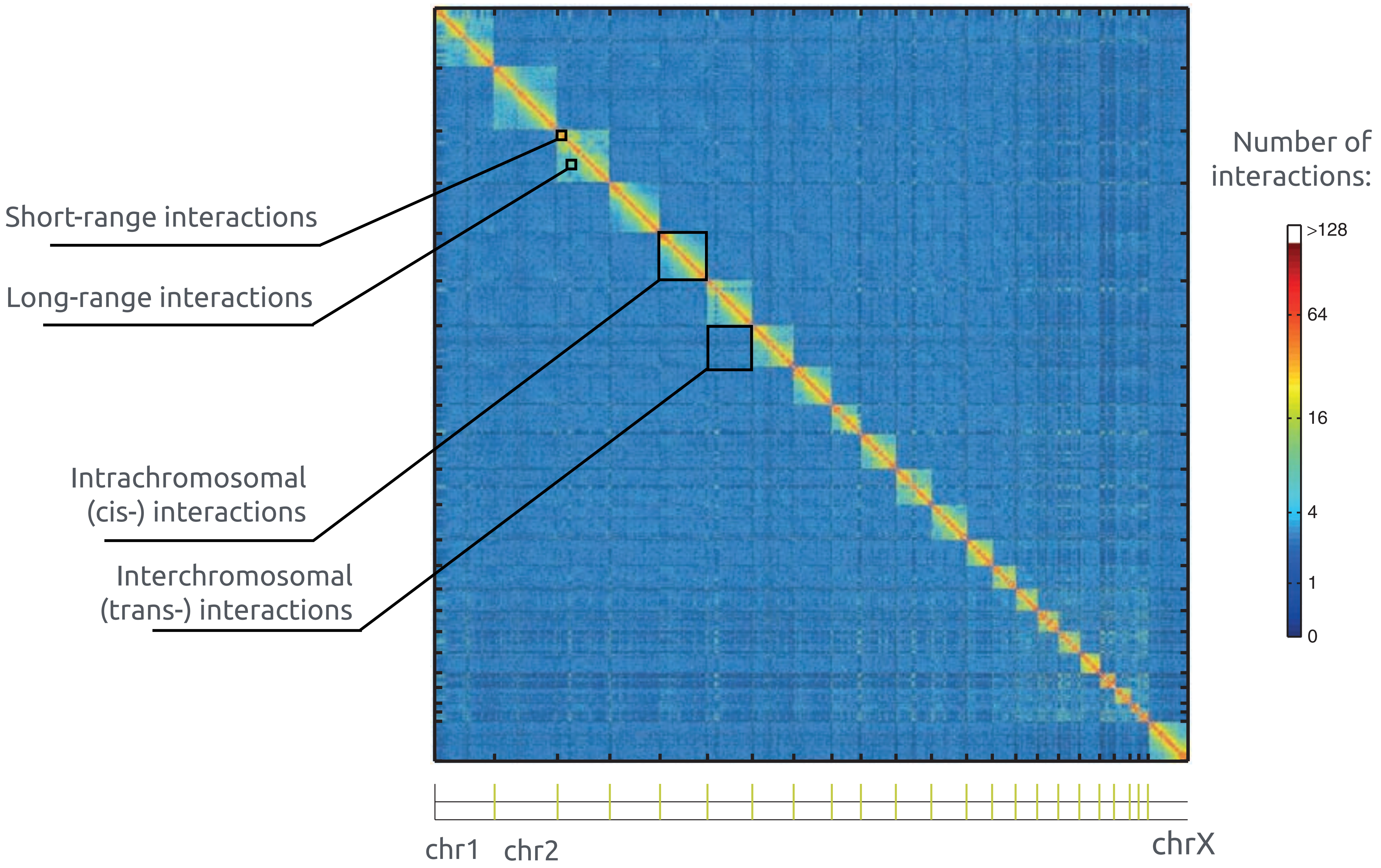
Lieberman-Aiden et al., Science 2009

The variety of 3C methods family

Type of probing	Assay abbreviation	Full assay name	Year
1 vs 1	3C	Chromosome conformation capture	2002
1 vs Many/All	Multiplexed 3C-seq	Multiplexed chromosome conformation capture sequencing	2011
	Open-ended 3C	Open-ended chromosome conformation capture	2006
	4C	Chromosome conformation capture-on-chip	2006
	ACT	Associated chromosome trap	2006
	e4C	Enhanced chromosome conformation capture-on-chip	2010
	3C-DSL	Chromosome conformation capture combined with DNA selection and ligation	2011
	4C-seq	Chromosome conformation capture-on-chip combined with high-throughput sequencing	2011
	4C	Circular chromosome conformation capture	2012
	TLA	Targeted locus amplification	2014
Many vs Many	5C	Chromosome conformation capture carbon copy	2006
	ChIA-PET	Chromatin interaction analysis paired-end tag sequencing	2009
Many vs All	Capture-3C	Chromosome conformation capture coupled with oligonucleotide capture technology	2014
	Capture-HiC	Hi-C coupled with oligonucleotide capture technology	2014
All vs All	GCC	Genome conformation capture	2009
	Hi-C	Genome-wide chromosome conformation capture	2009
	ELP	Genome-wide chromosome conformation capture with enrichment of ligation products	2010
	TCC	Tethered conformation capture	2012
	Single-cell Hi-C	Single-cell genome-wide chromosome conformation capture	2013
	In situ Hi-C	Genome-wide chromosome conformation capture with in situ ligation	2014
	DNase Hi-C	Genome-wide chromosome conformation capture with DNase I digestion	2015
	Micro-C	Genome-wide chromosome conformation capture with micrococcal nuclease digestion	2015
	GAM	Genome Architecture Mapping	2017

Adopted from Schmitt *Nature Reviews* 2016

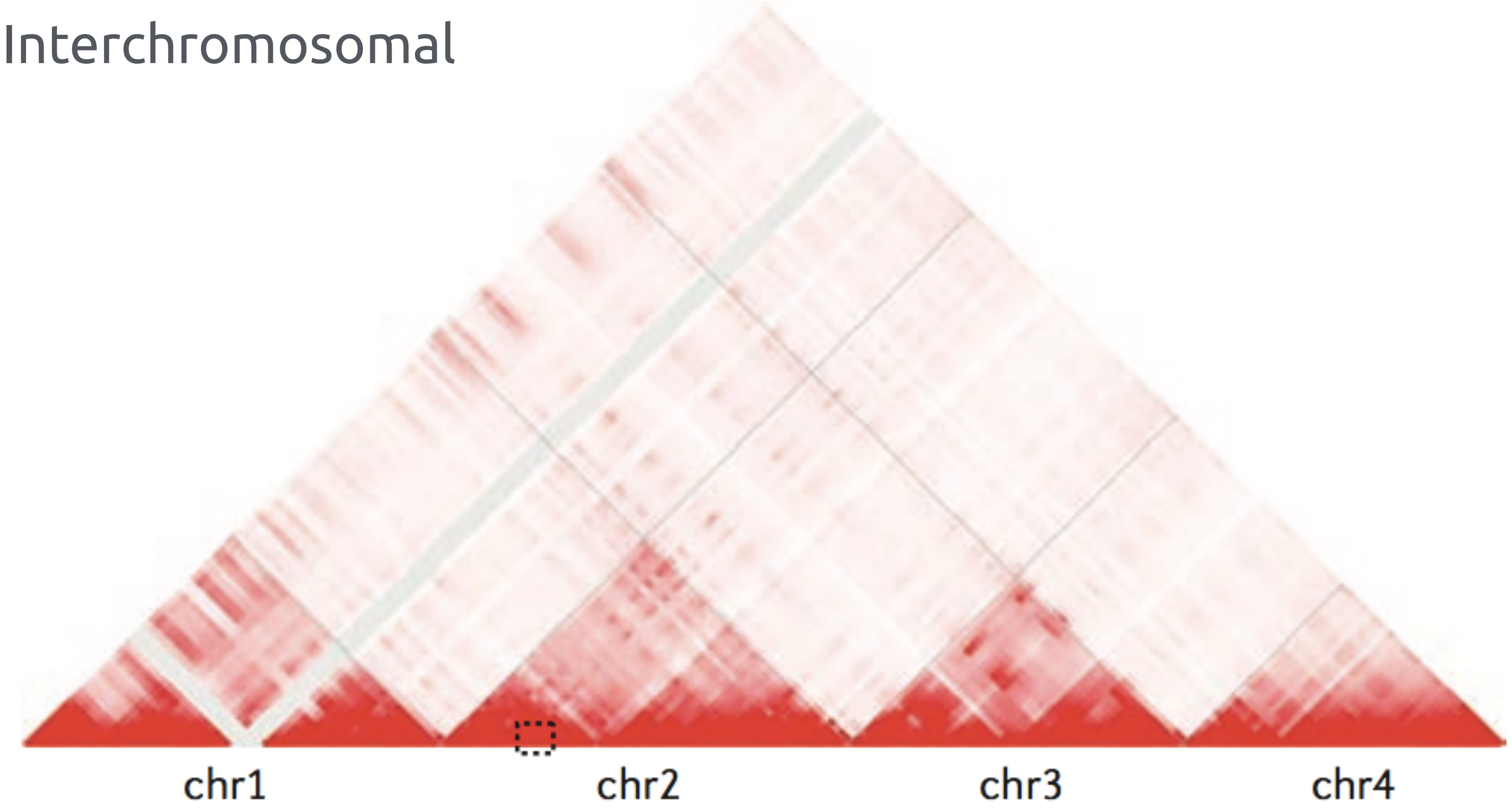
Chromatin interactions map



Adopted from Imakaev et al. *Nature Methods* 2012

Interaction map features: Chromosome territories

- At the highest-level of spatial organization, trans-interactions are rare.
- Individual chromosomes occupy distinct territories within the nucleus.

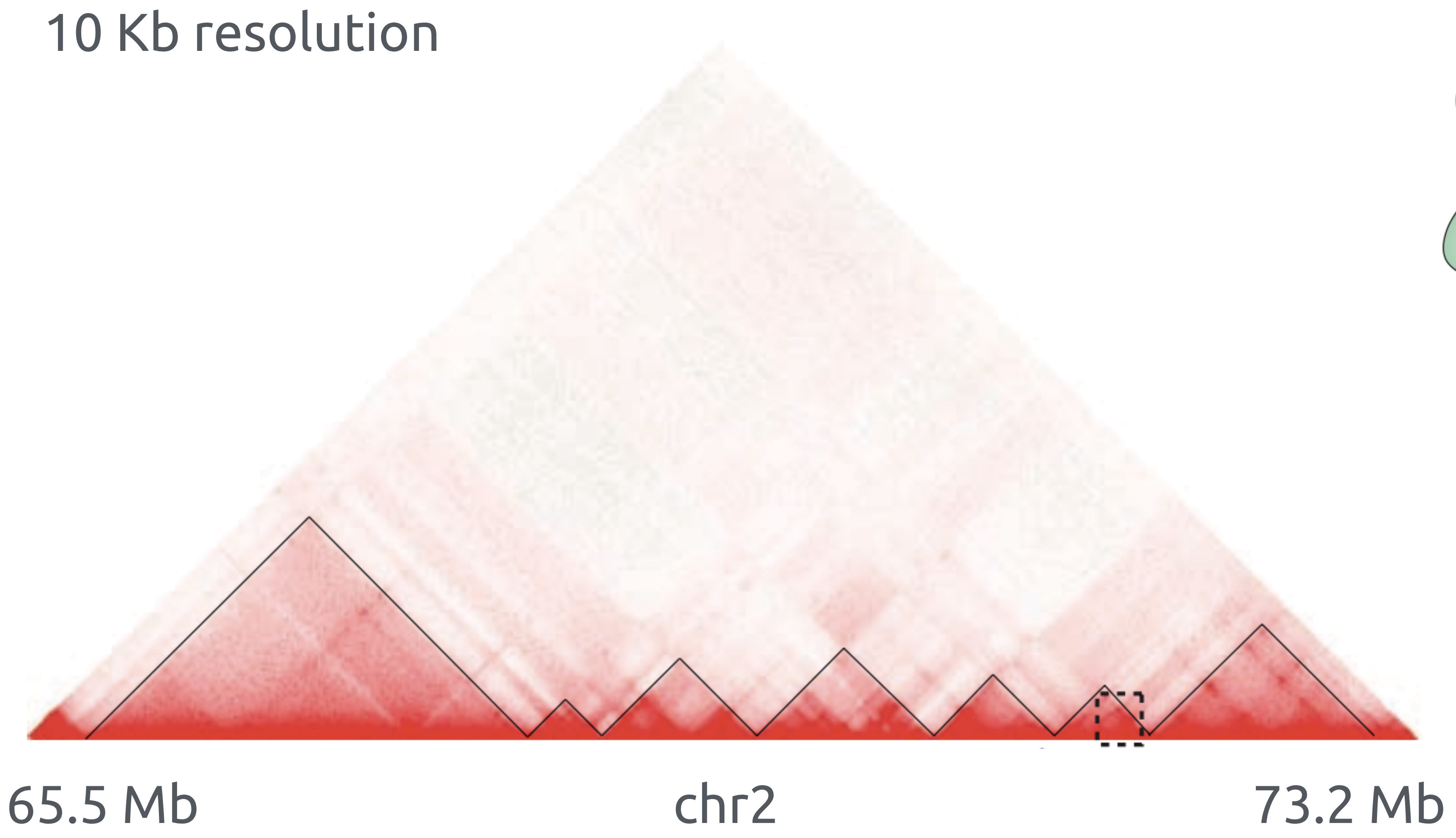


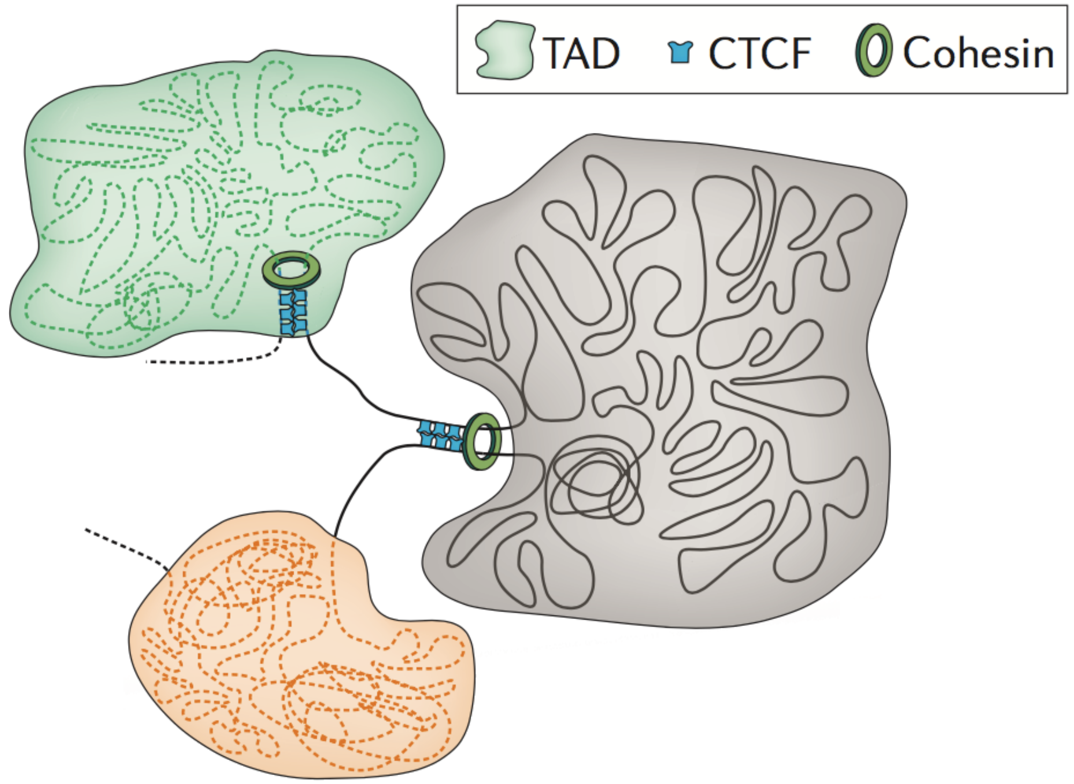
Map is rotated by 90°, upper triangle visualized

Bonev et al. *Nature Reviews* 2016

Topologically-associating domains (TADs)

- Chromosomes are further spatially segregated into sub-megabase scale domains, or TADs.

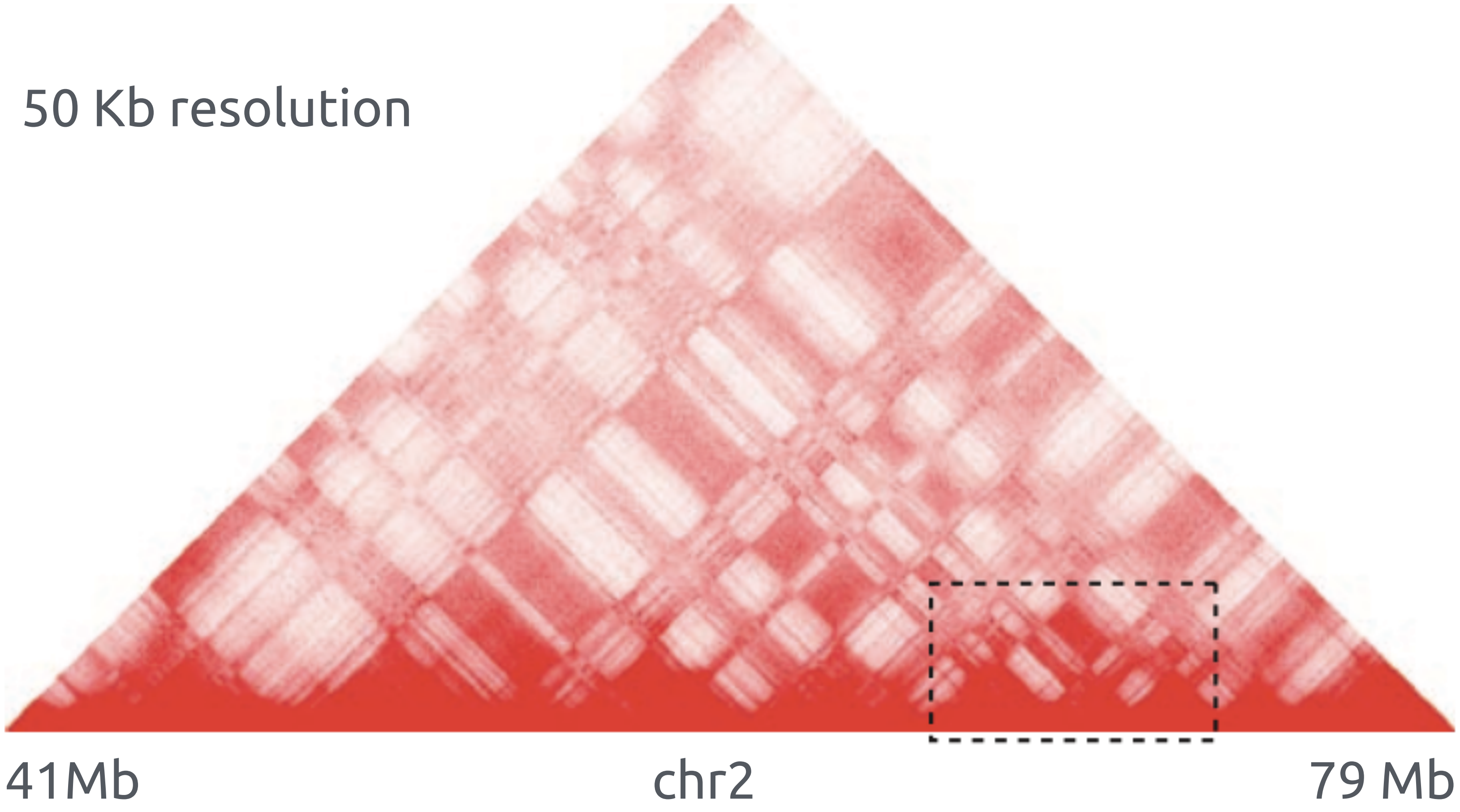


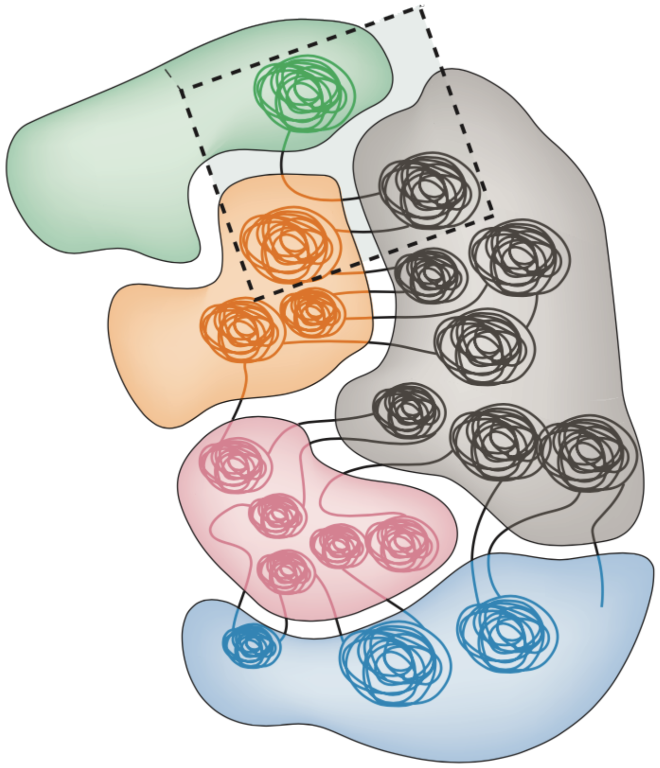


Bonev et al. *Nature Reviews* 2016

Chromatin compartments

- TADs have preferential long-range contacts with each other, forming two types of compartments, A and B (domains in compartment A interact mostly with other type A domains, and vice versa).
- Two major compartments can be further subdivided into six different subcompartments.

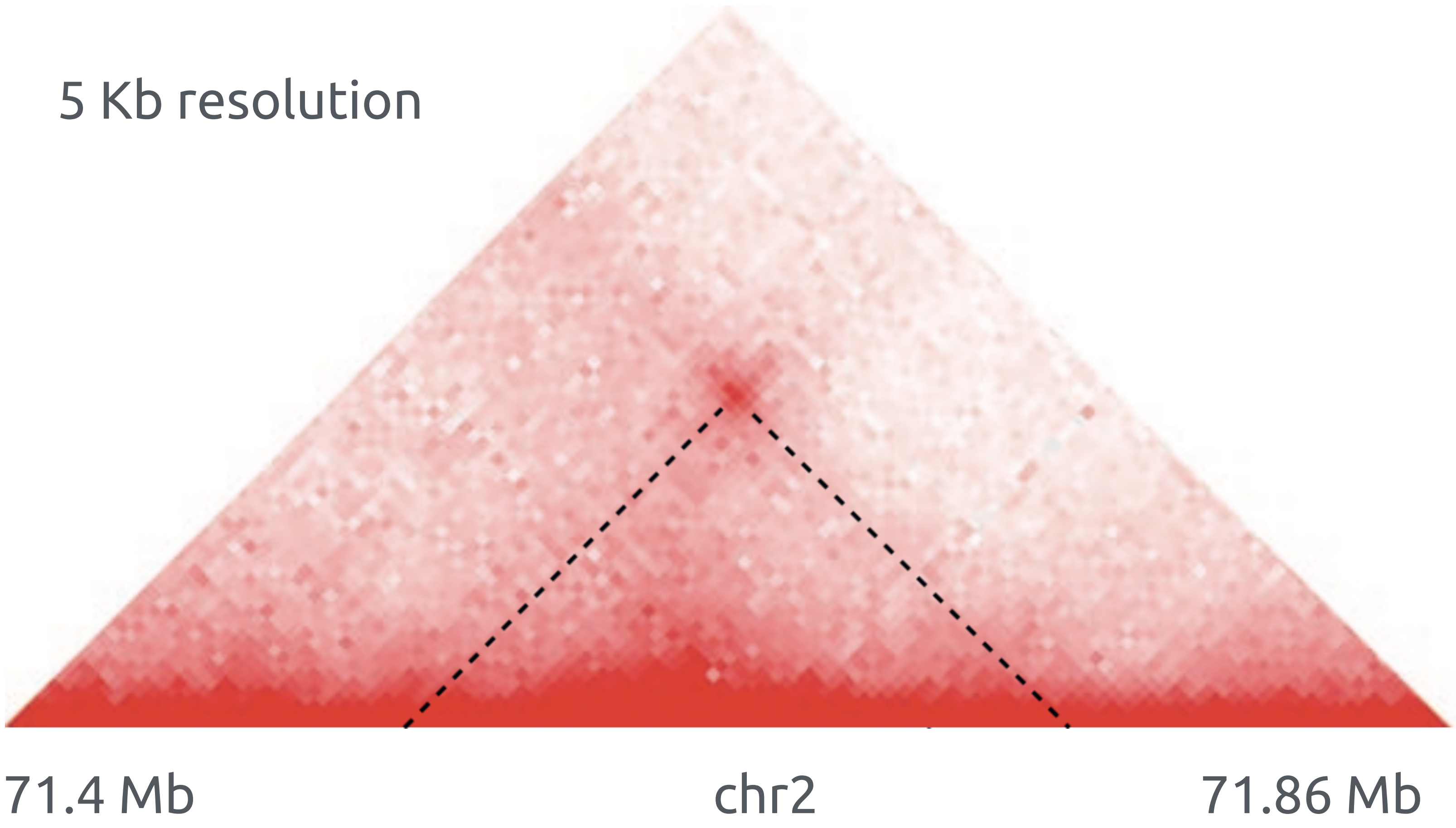


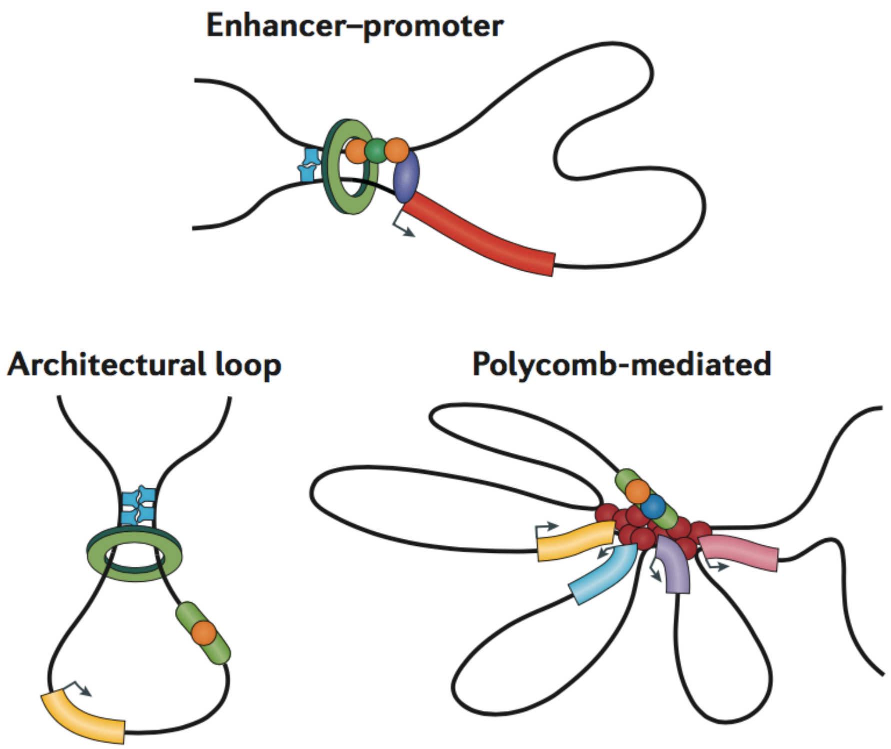


Bonev et al. *Nature Reviews* 2016

Chromatin loops

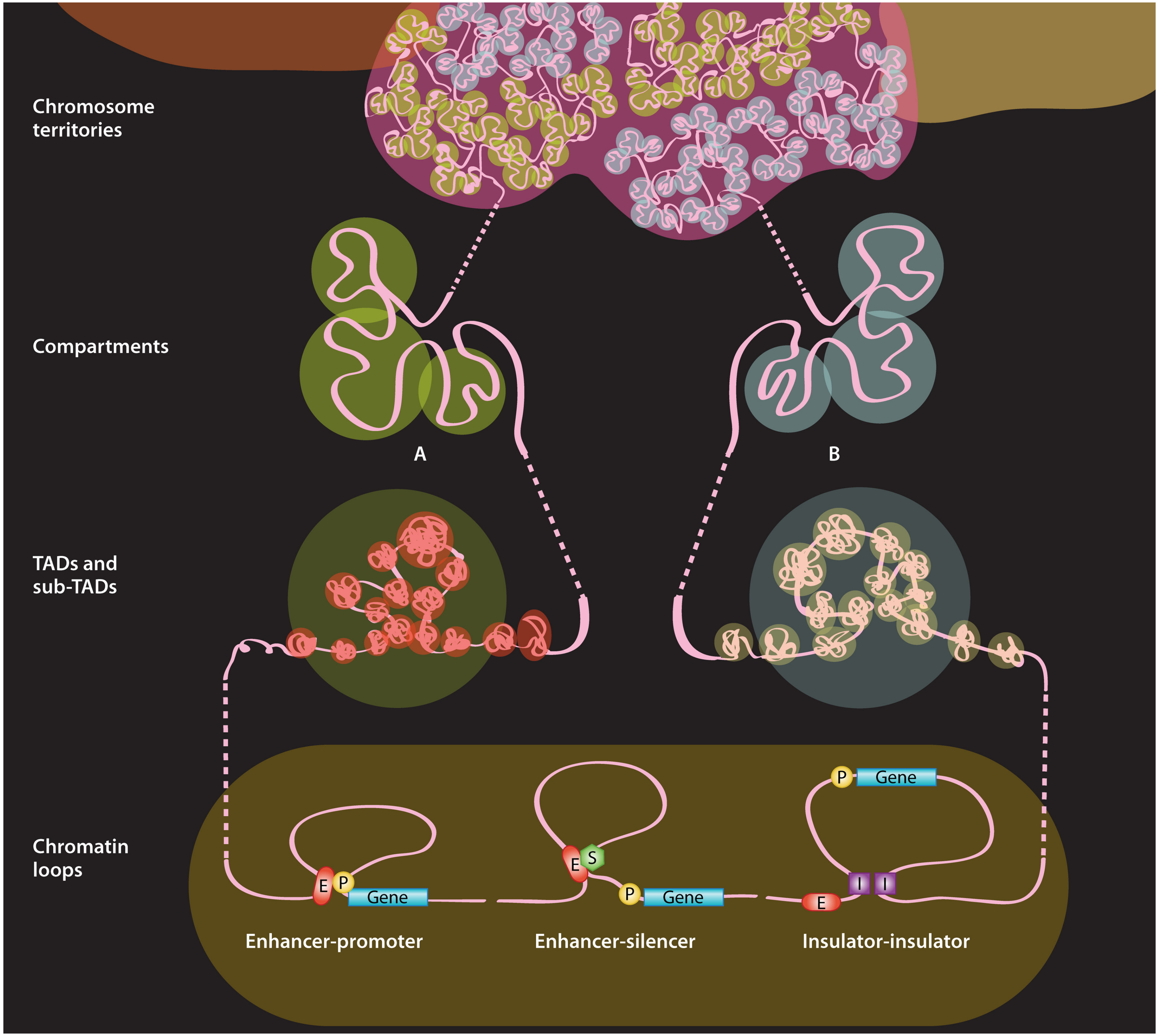
- Cis-regulatory elements of vertebrates, such as enhancers, are separated by relatively long distances and can be brought into close spatial proximity with its target through the formation of chromatin loops.
- There are also other cases of loops (e.g. between co-regulated genes, between Polycomb-repressed genes).





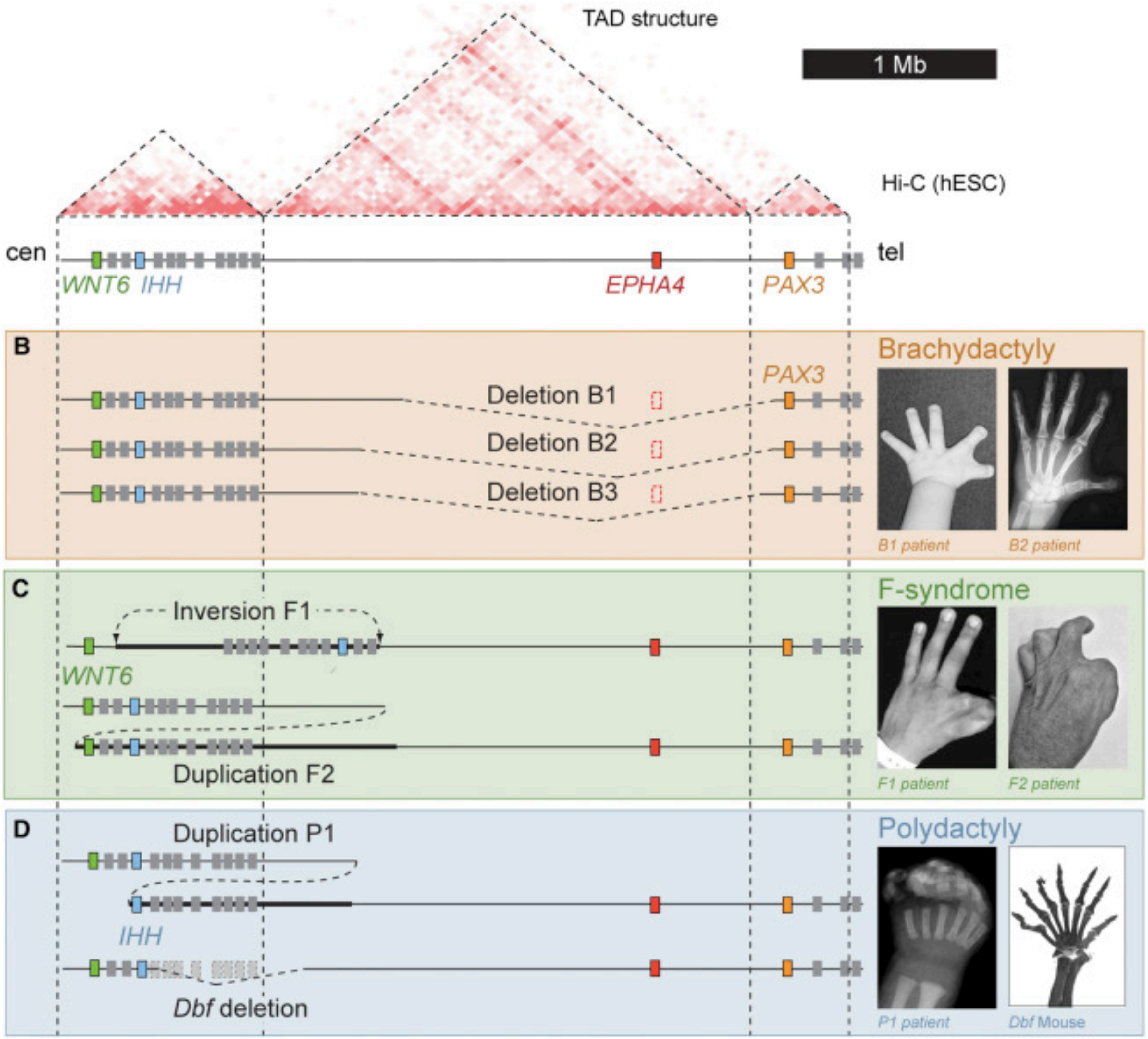
Bonev et al. *Nature Reviews* 2016

"The Zoo" of chromatin features



Fraser et al. *Microbiology and Molecular Biology Reviews* 2015

Why is it important?



Krijger, Nature, 2016

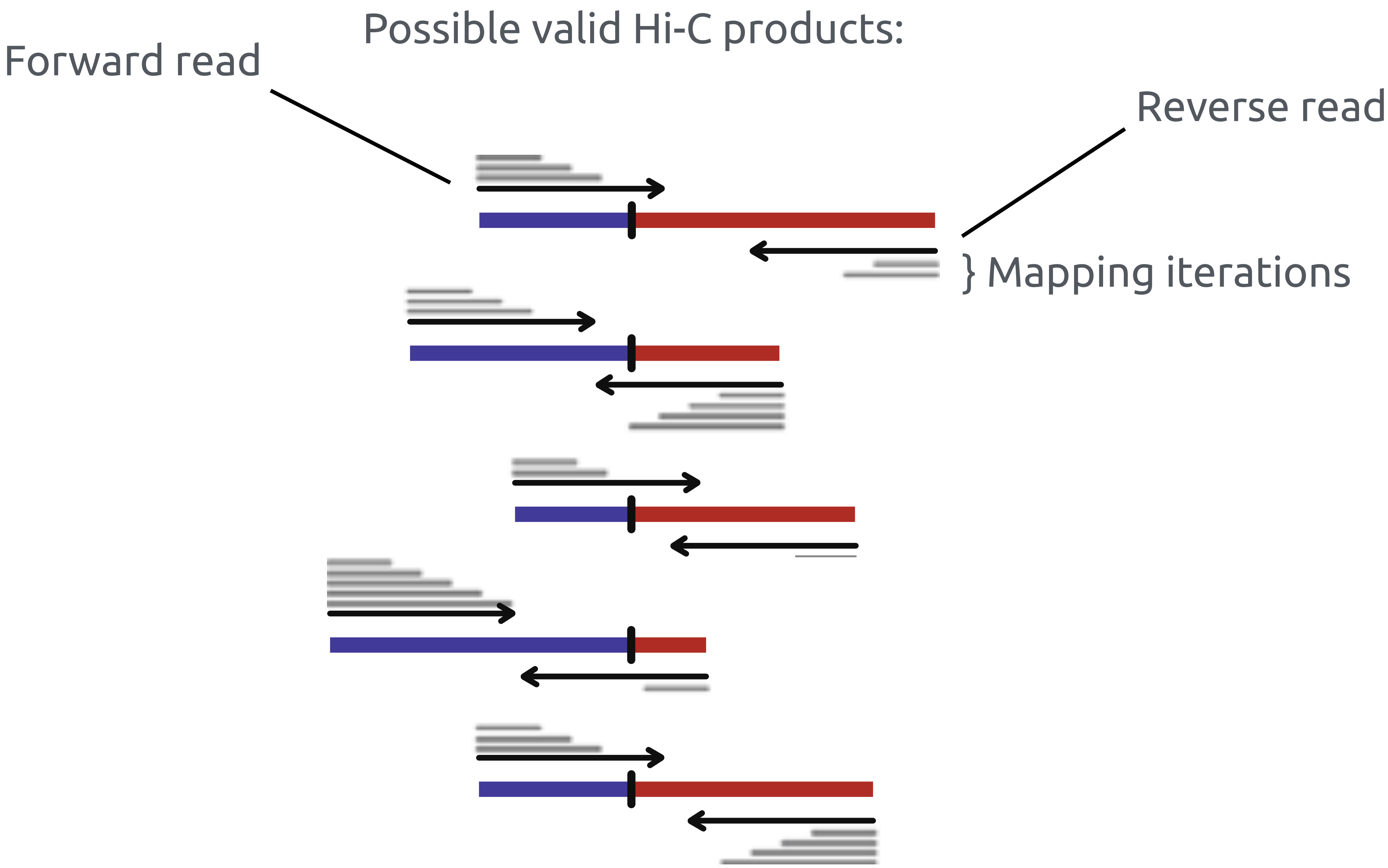
2. From theory to practice: Hi-C processing workflow

Hi-C processing workflow

1. Reads mapping: paired-end mode is not used, iterative mapping.
2. Filtering & binning
    - Fragment assignment: the mapped read is assigned according to its 5' mapped position, mapped read positions should fall close to a restriction site
    - Fragment filtering: multiple mapping, PCR duplicates, undigested restriction sites
    - Binning
    - Bin level filtering: remove 1% low signal rows/columns
3. Balancing: correction for technical biases
4. Features calling (TADs, compartments, loops, etc.)

Reads mapping

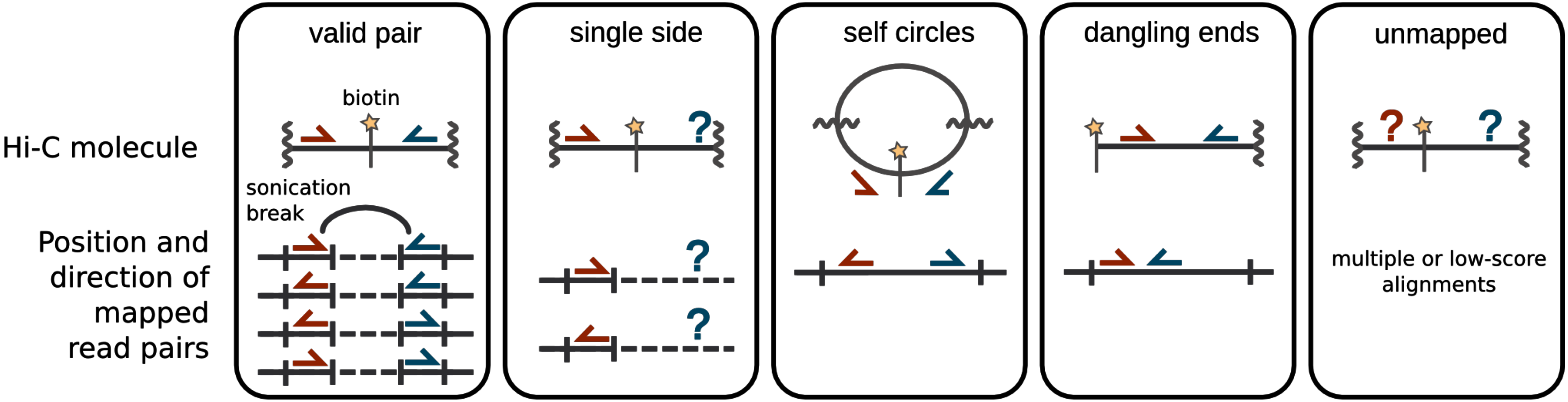
- Iterative or split reads mapping is required.



Adopted from Lajoie et al., The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods* 2015

Filtering at the level of fragments

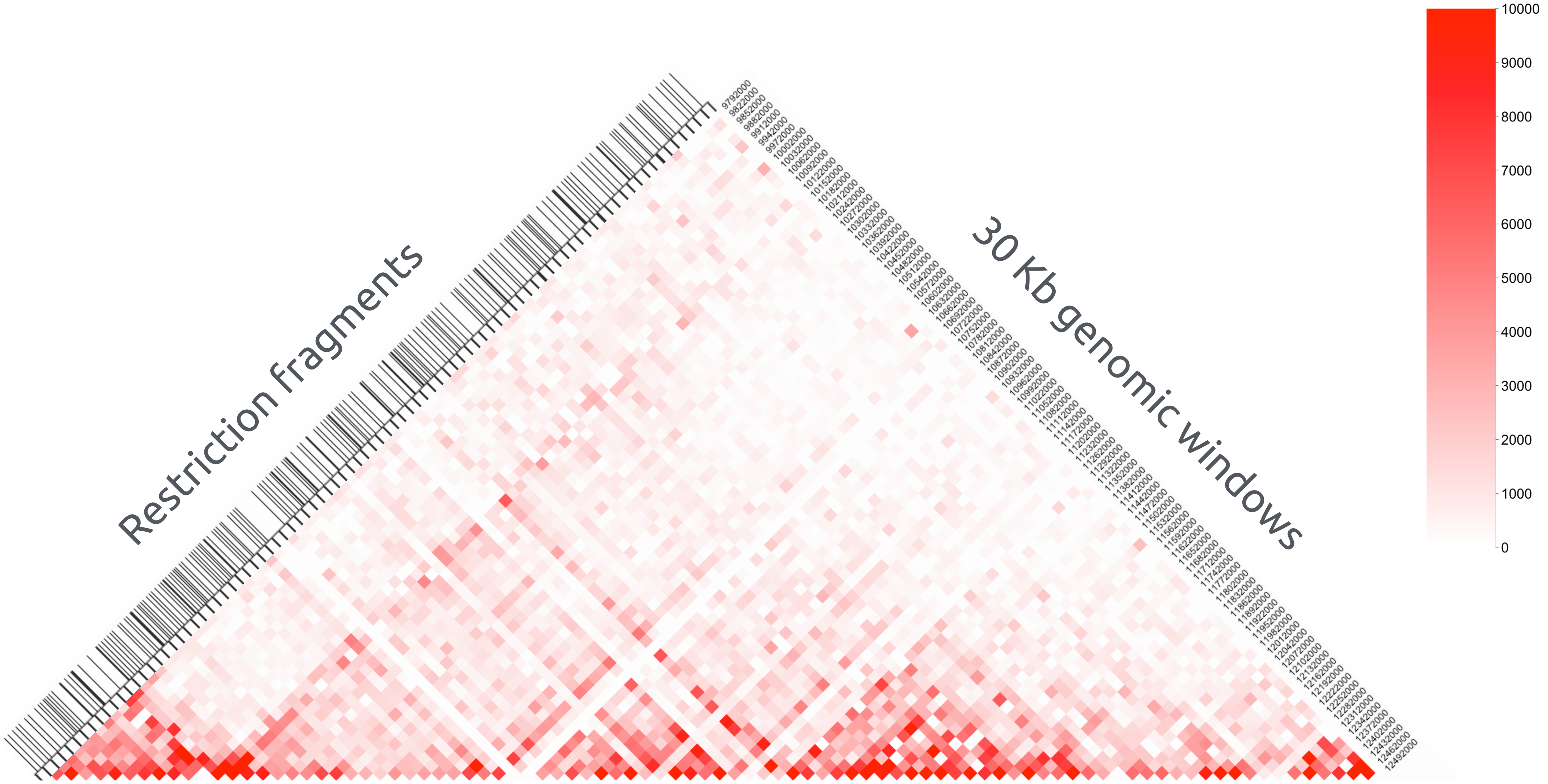
- Possible Hi-C mapping results:



Imakaev et al. *Nature Methods* 2012

Binning

- Hi-C restriction fragments are assigned to bins (sequential same size genomic windows) and aggregated by taking the sum:



Binned interactions heatmap

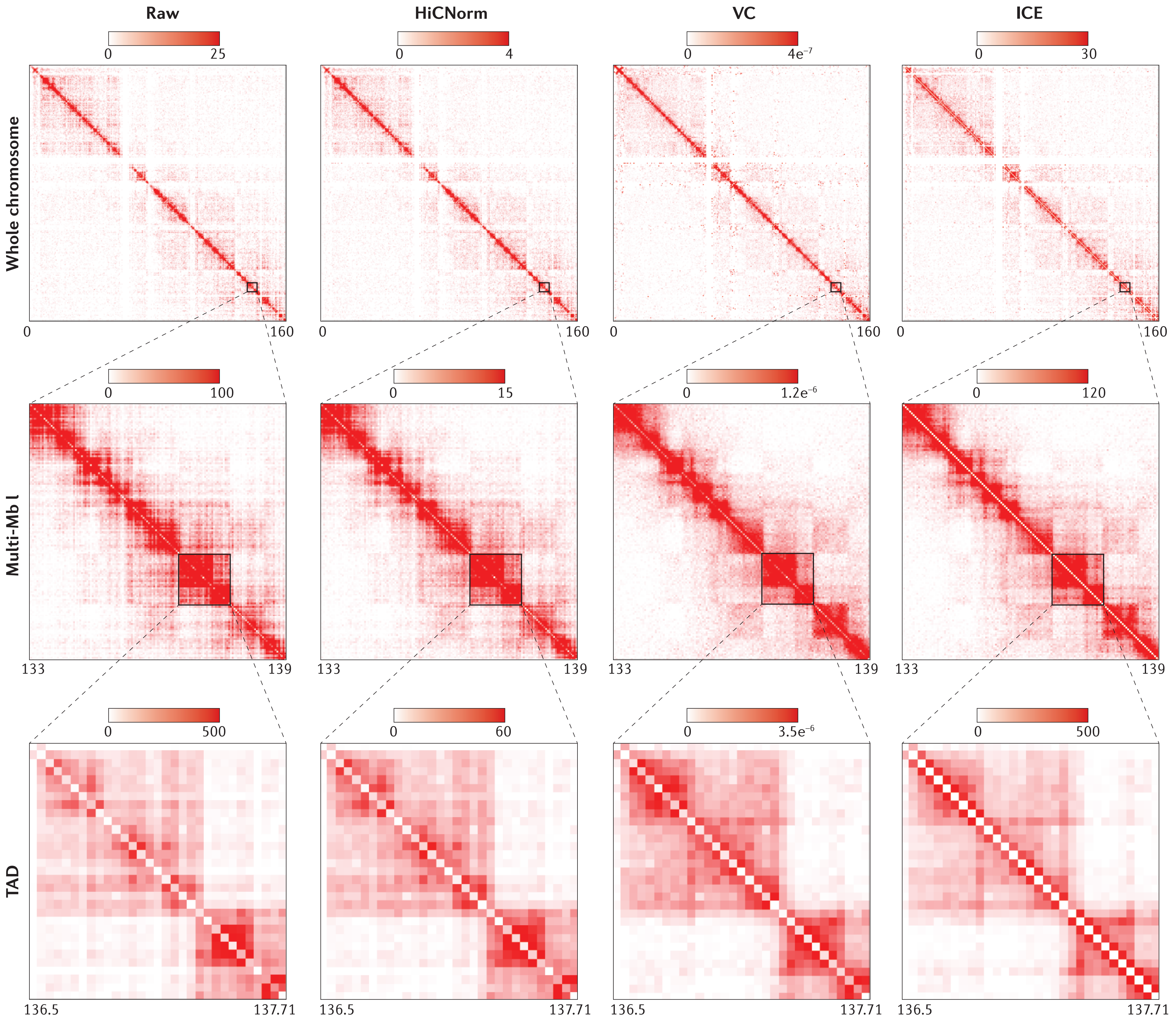
Matrix balancing

- Balancing is the procedure of correction of systematic technical bias in data.
- Major balancing methods and two general types of balancing:

Approach	**Type**	**Model assumption**	**Implementation**	**Computational speed**
Yaffe and Tanay	Explicit	Restriction enzyme fragment lengths, GC content and sequence mappability are three major systematic biases in Hi-C	Perl and R	Slow
HiCNorm			R	Fast
Iterative correction (ICE)	Implicit	All the bias is captured by the sequencing coverage of each bin, equal visibility	Python	Fast
Knight and Ruiz			JAVA	Fast
HiC-Pro			Python and R	Very fast

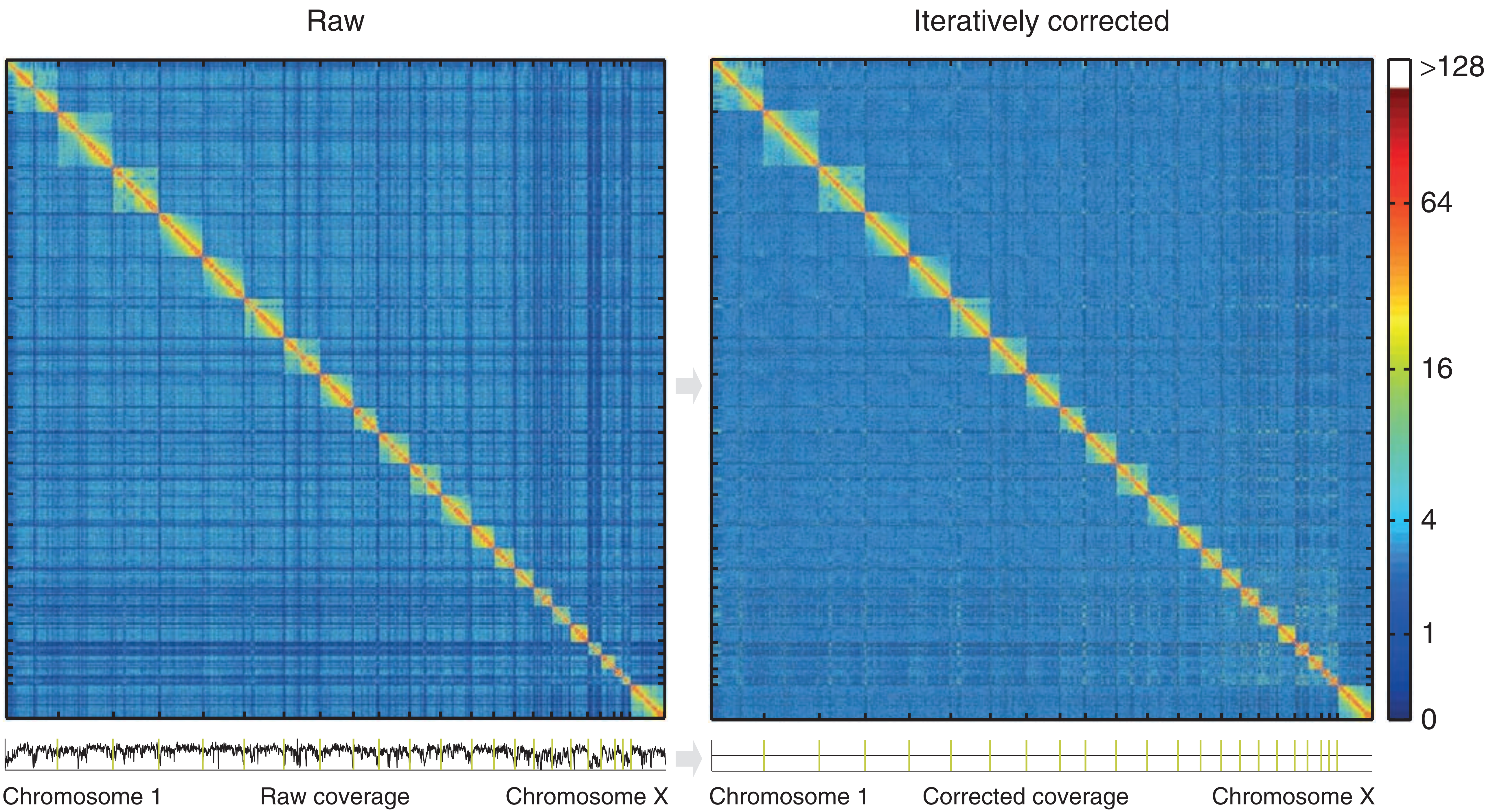
Adopted from Schmitt et al. *Nature Reviews* 2016

Matrix balancing



Schmitt et al. *Nature Reviews* 2016

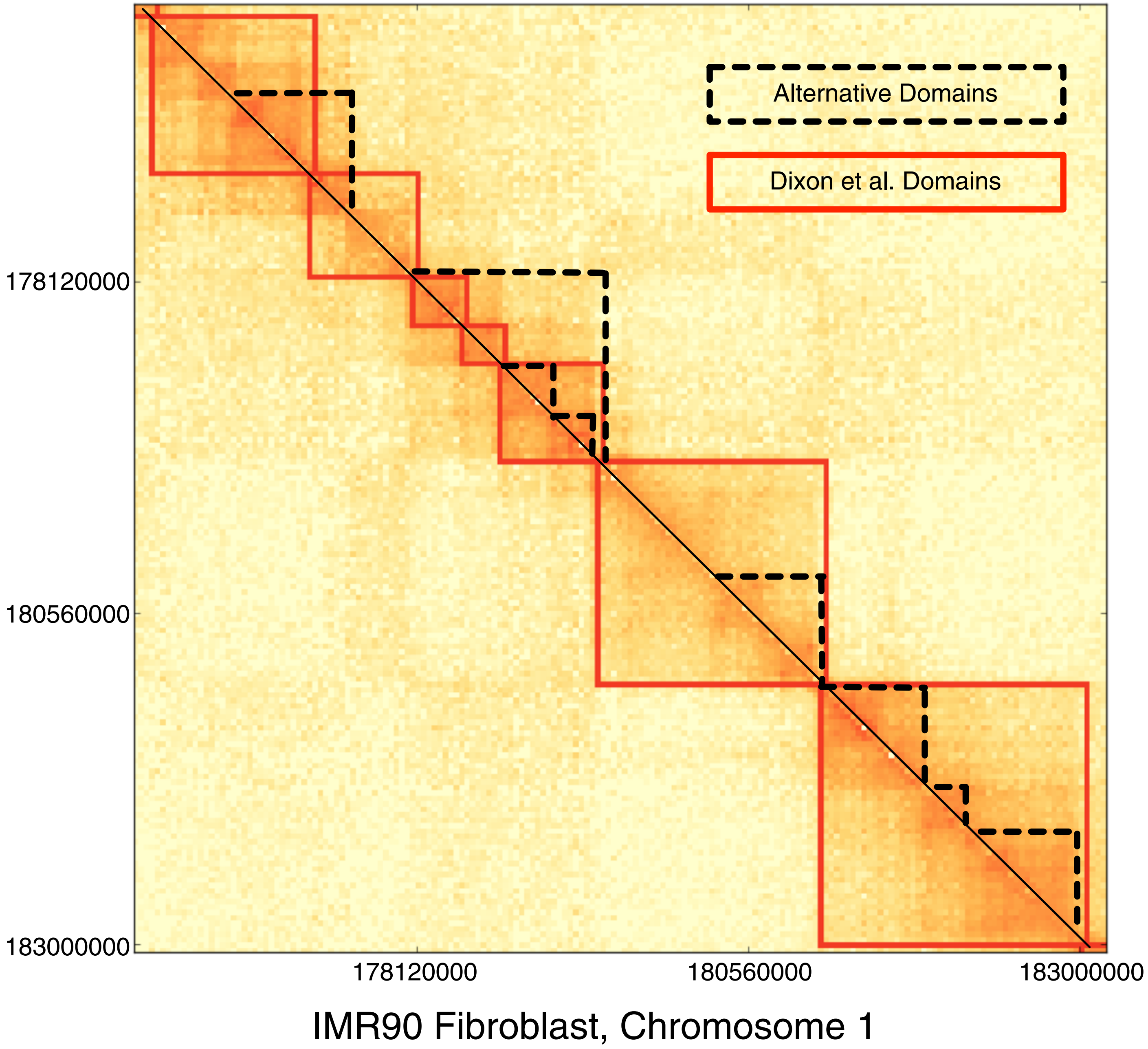
Iterative correction



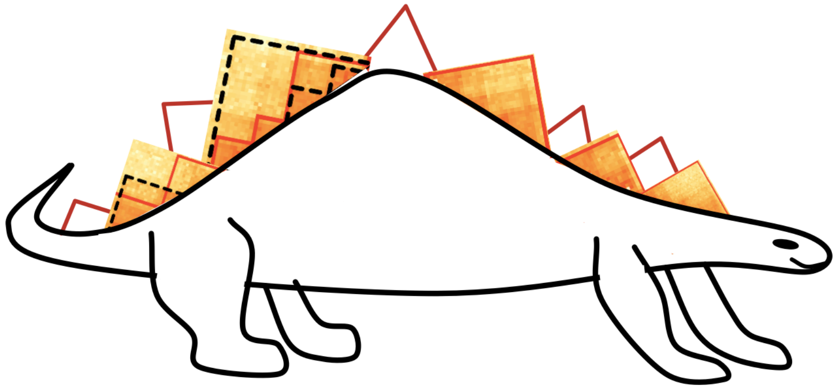
Imakaev et al. *Nature Methods* 2012

TADs calling

- TADs are hierarchical, there is no gold standard for TADs selection:



IMR90 Fibroblast, Chromosome 1



Armatus is a program predicting TADs using Hi-C contact matrices as an input.

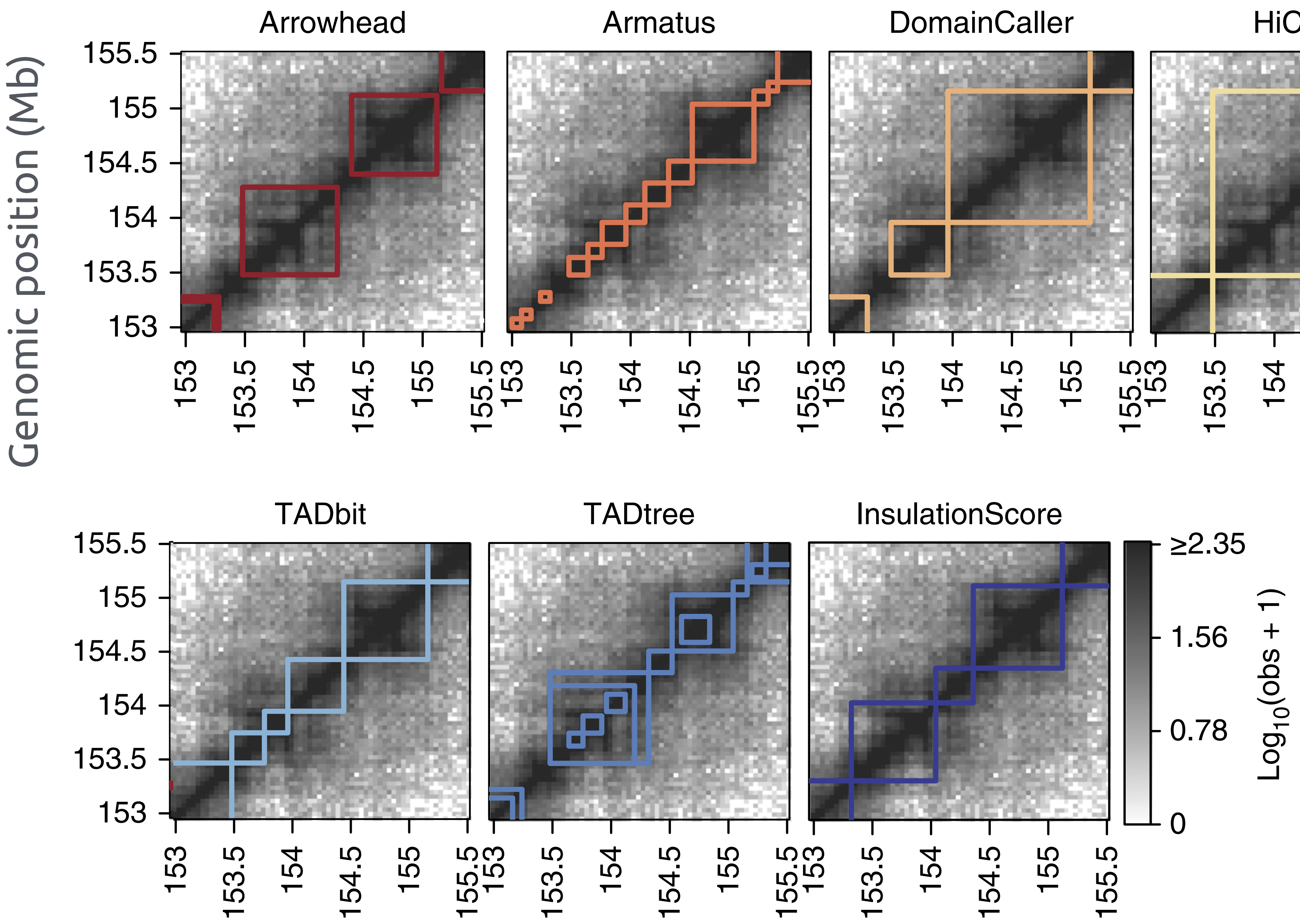
Armatus can produce several TAD annotations with different average TAD sizes.

Hierarchical structure of TADs: large TADs can be split into smaller ones.

Filippova et al. *Algorithms for Molecular Biology* 2014

TADs calling

- A recent comparison of multiple TADs calling tools:



Forcato et al. *Nature Methods* 2017

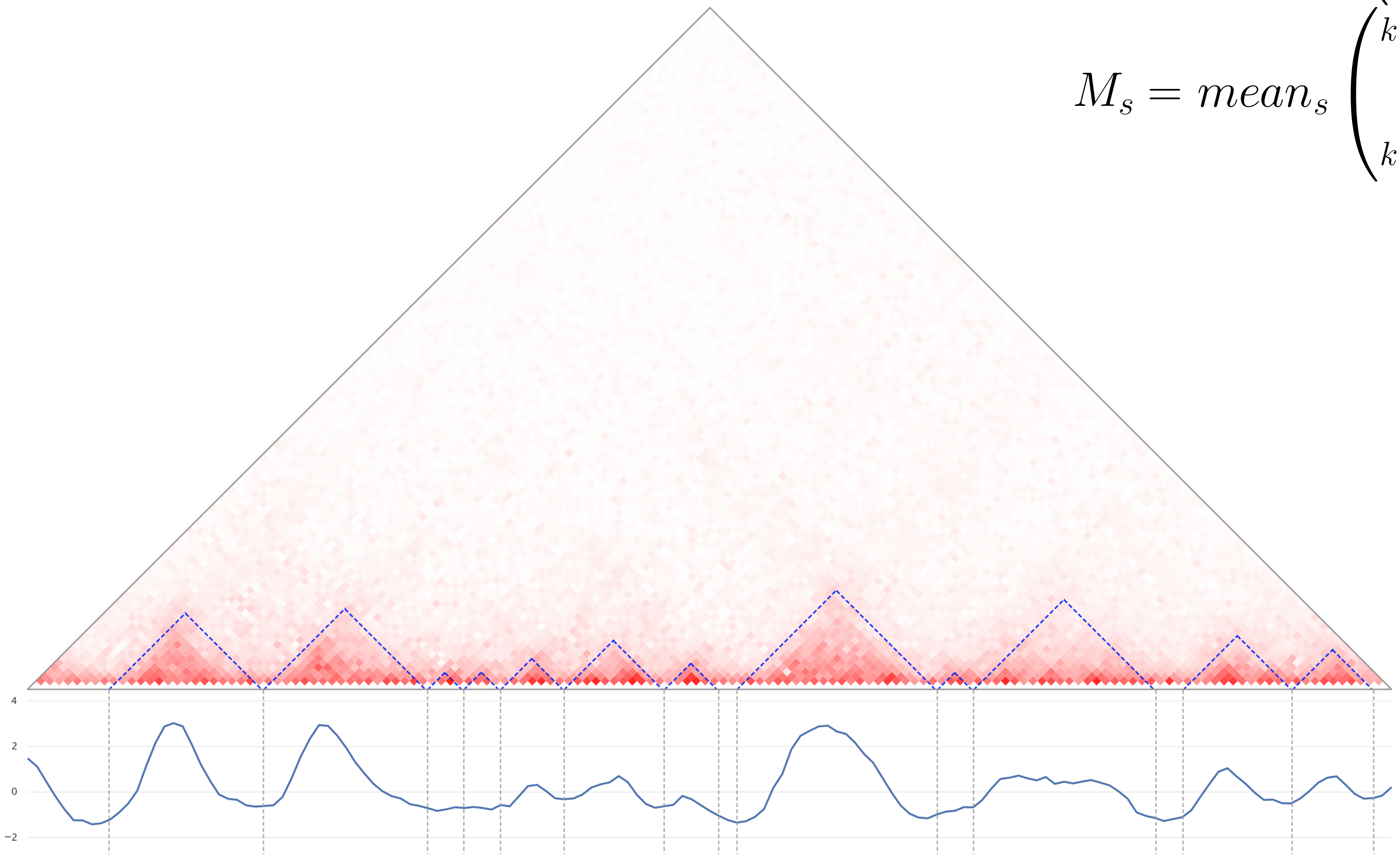
TADs calling

- Insulation score is intuitively easy way to calculate TAD boundaries:

1. Calculates insulation score (IS) for each bin:
2. Find local minima in IS profile

$$IS_j(s) = log_2 \left(\sum_{k=j-s/2}^{k=j+s/2} \frac{C_{k,k+s}}{M_s} \right)$$

$$M_s = mean_s \left(\sum_{k=j-s/2}^{k=j+s/2} C_{k,k+s} \right)$$

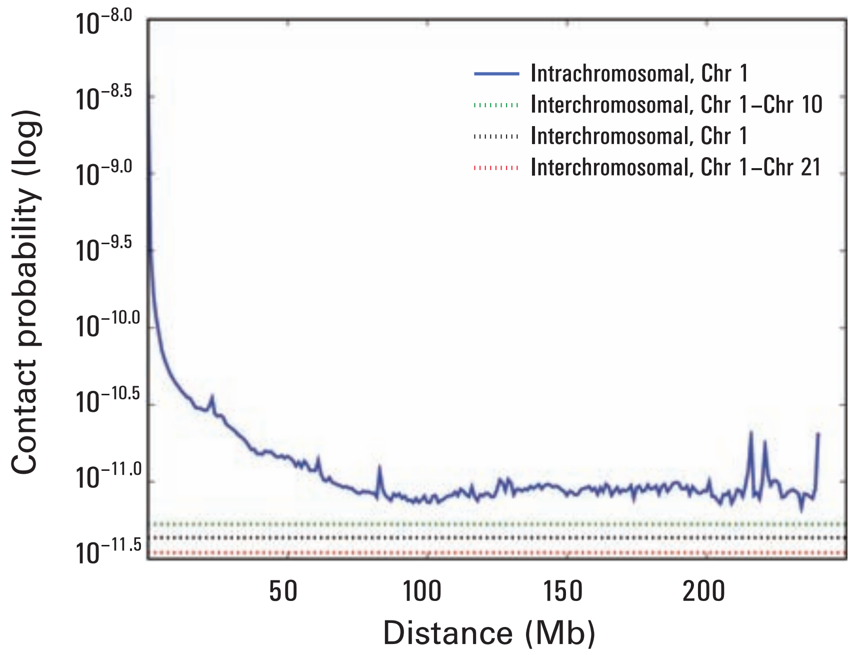


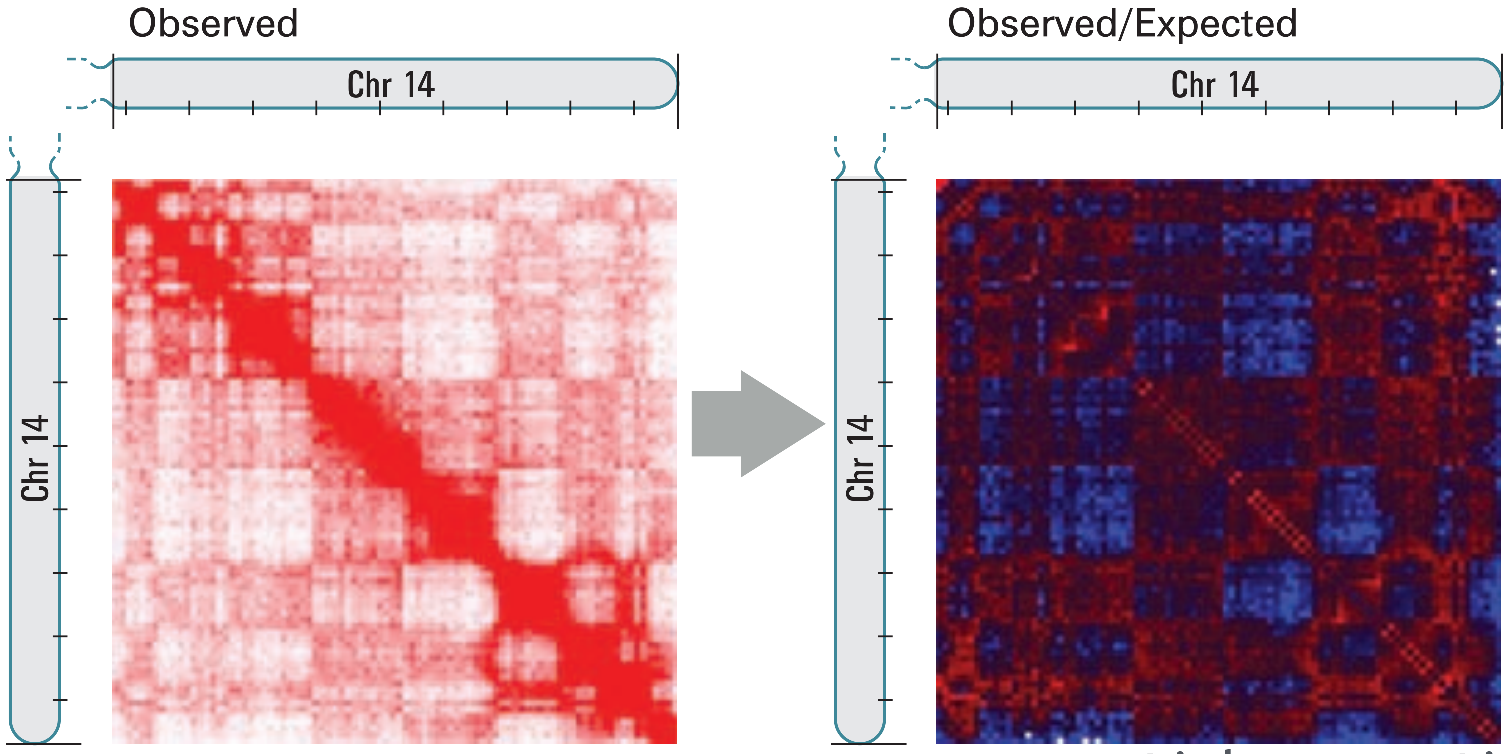
Based on Crane, 2015

Compartments calling

- Method from Lieberman-Aiden, 2009:

  1 Normalization of interaction matrix by expected interactions:



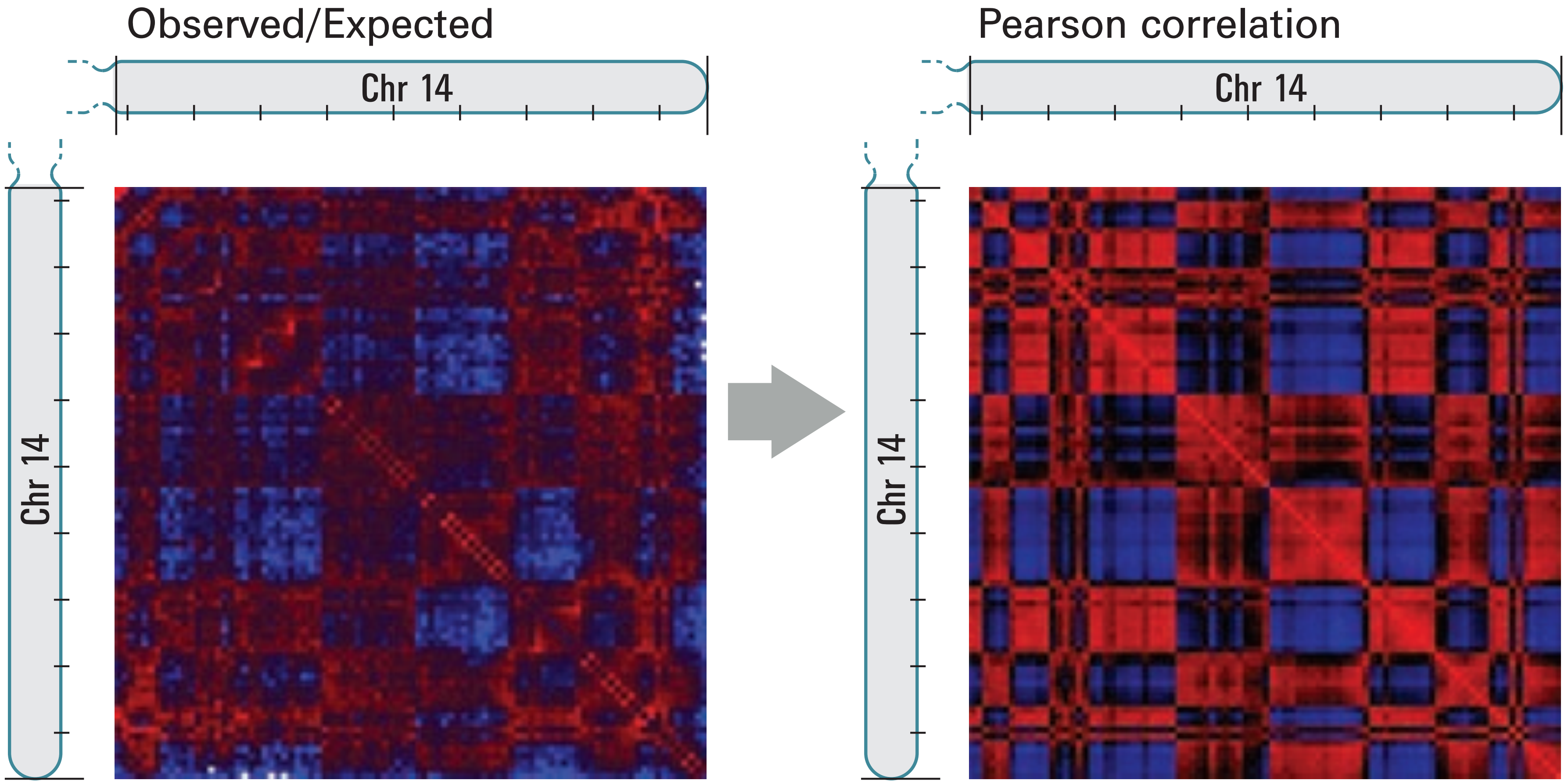


Lieberman-Aiden et al. *Nature* 2009

Compartments calling

- Method from 2009:

  ② Calculation of Pearson correlation

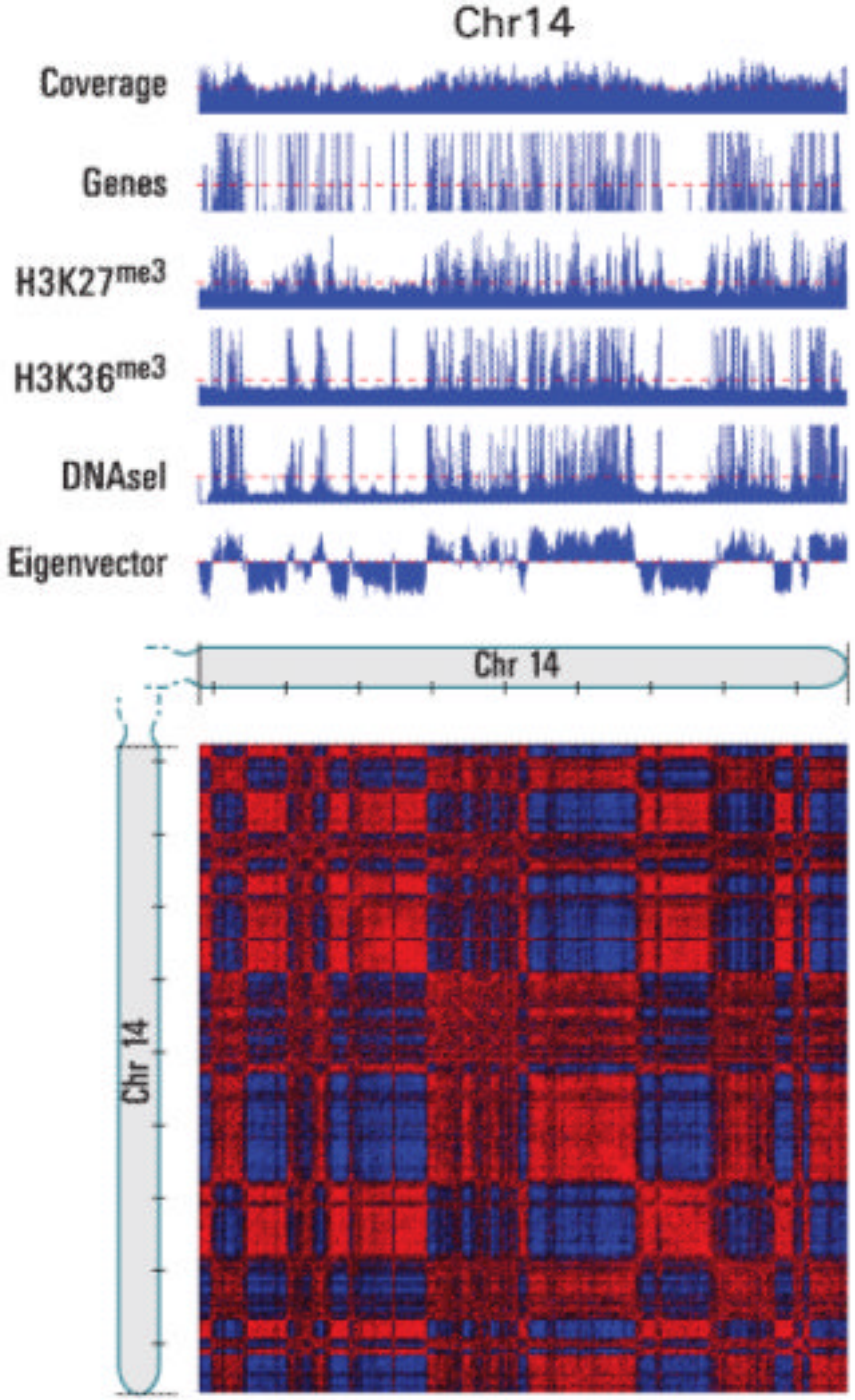


Lieberman-Aiden et al. *Nature* 2009

Compartments calling

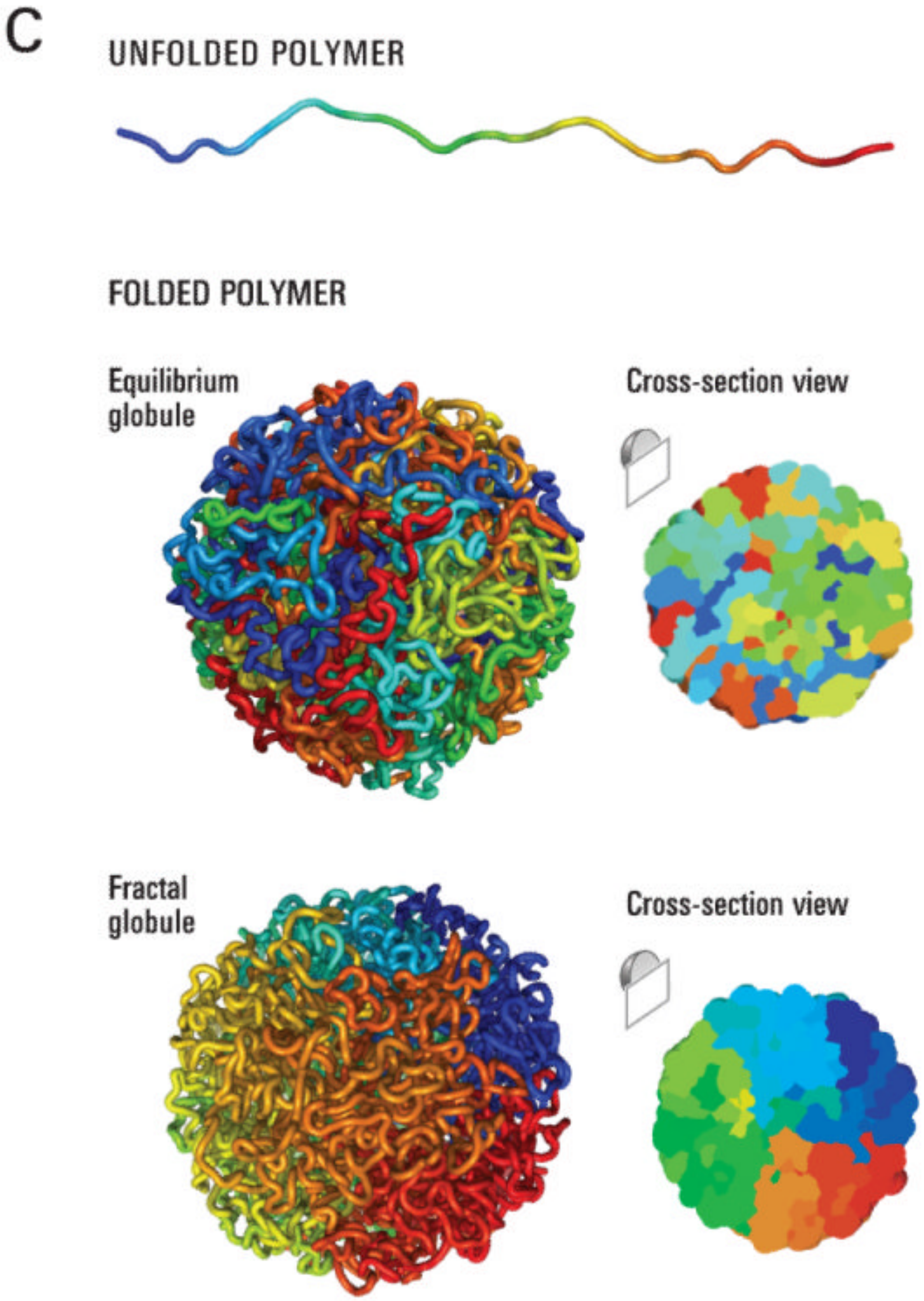
- Eigenvector decomposition:

  3 Eigenvector expansion (PCA, principal component analysis)



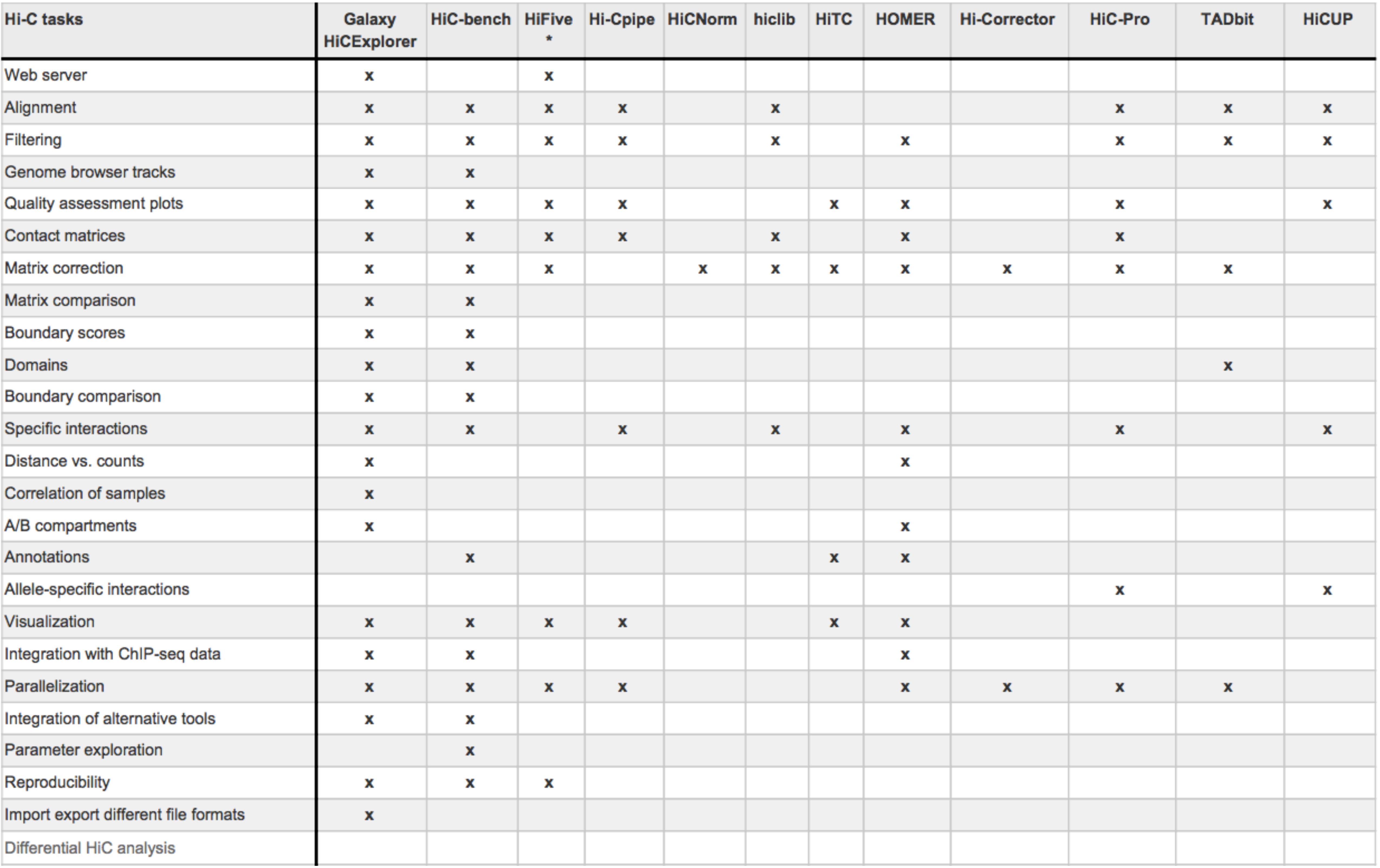
Lieberman-Aiden et al. *Nature* 2009

Chromatin modelling



Lieberman-Aiden et al. 2009

Variety of Hi-C processing tools



Hi-C tasks	Galaxy HiCExplorer	HiC-bench	HiFive *	Hi-Cpipe	HiCNorm	hiclib	HiTC	HOMER	Hi-Corrector	HiC-Pro	TADbit	HiCUP
Web server	x		x									
Alignment	x	x	x	x		x				x	x	x
Filtering	x	x	x	x		x		x		x	x	x
Genome browser tracks	x	x										
Quality assessment plots	x	x	x	x			x	x		x		x
Contact matrices	x	x	x	x		x		x		x		
Matrix correction	x	x	x		x	x	x	x	x	x	x	
Matrix comparison	x	x										
Boundary scores	x	x										
Domains	x	x									x	
Boundary comparison	x	x										
Specific interactions	x	x		x		x		x		x		x
Distance vs. counts	x							x				
Correlation of samples	x											
A/B compartments	x							x				
Annotations		x					x	x				
Allele-specific interactions										x		x
Visualization	x	x	x	x			x	x				
Integration with ChIP-seq data	x	x						x				
Parallelization	x	x	x	x				x	x	x	x	
Integration of alternative tools	x	x										
Parameter exploration		x										
Reproducibility	x	x	x									
Import export different file formats	x											
Differential HiC analysis												

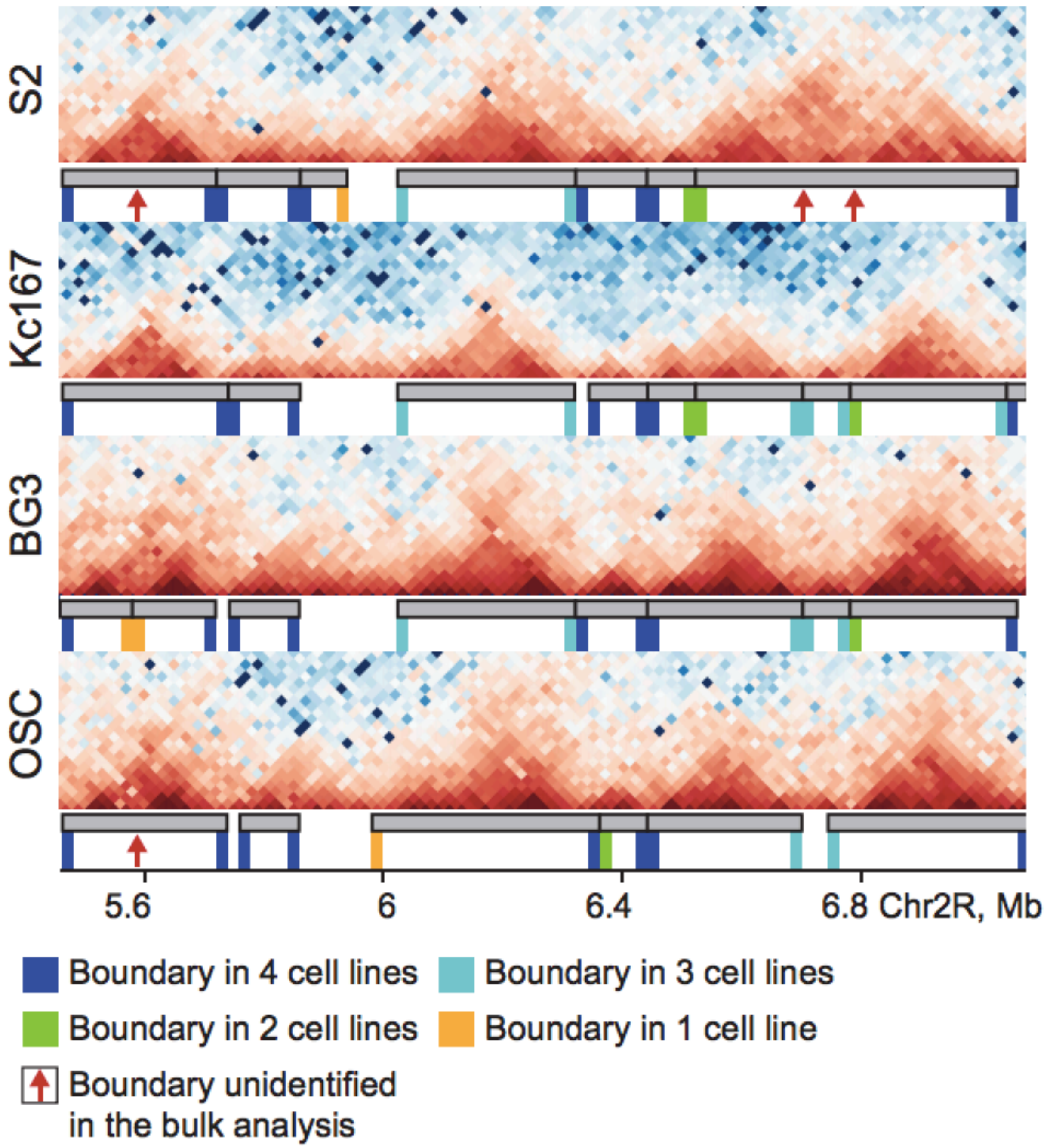
....

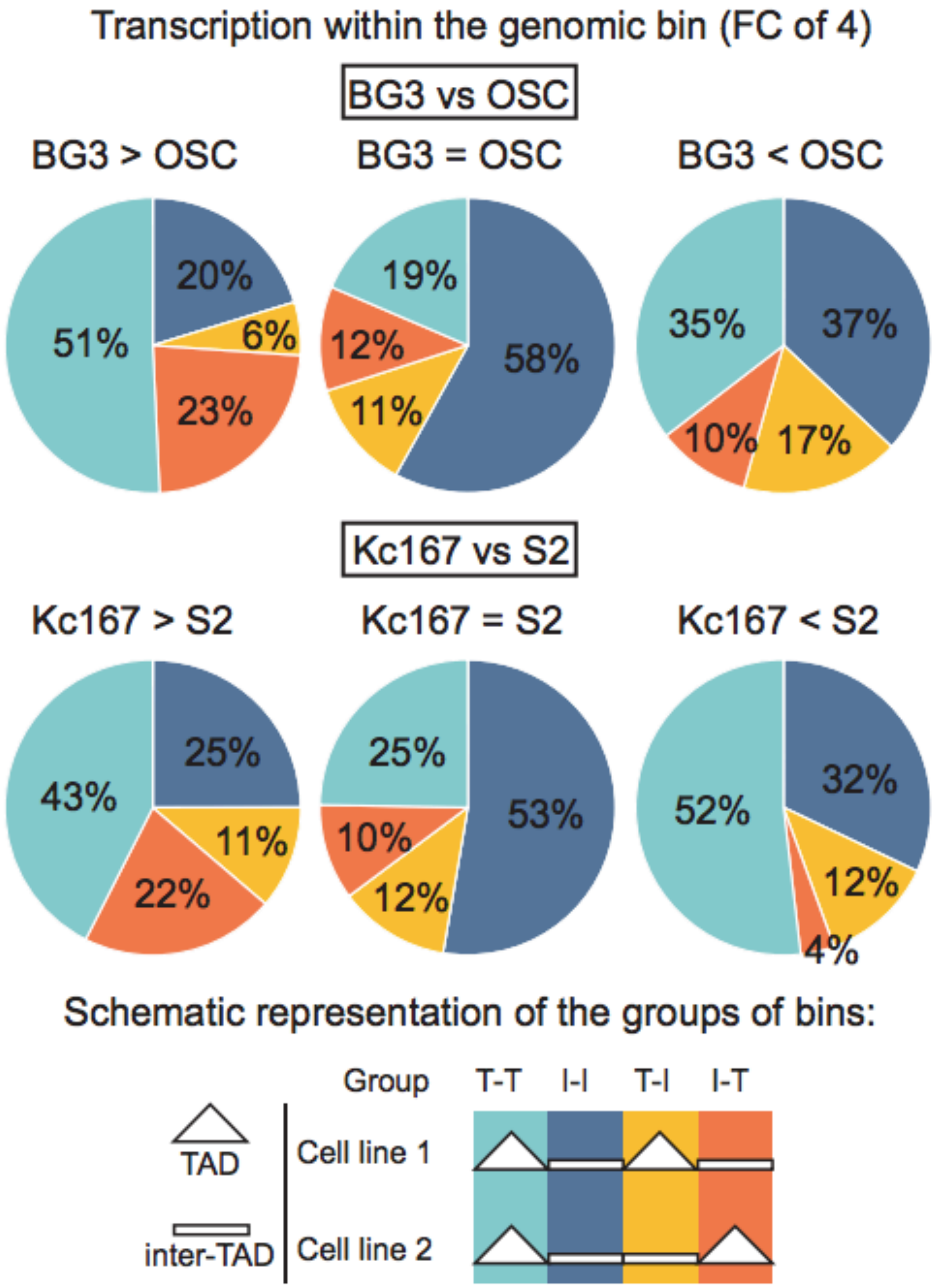
from Lazaris 2017 BMC Genomics, modified

3. Some cases from chromatin study practice

Comparison of cell lines

For example, we can compare chromatin properties in different cells and associate them with gene activity:

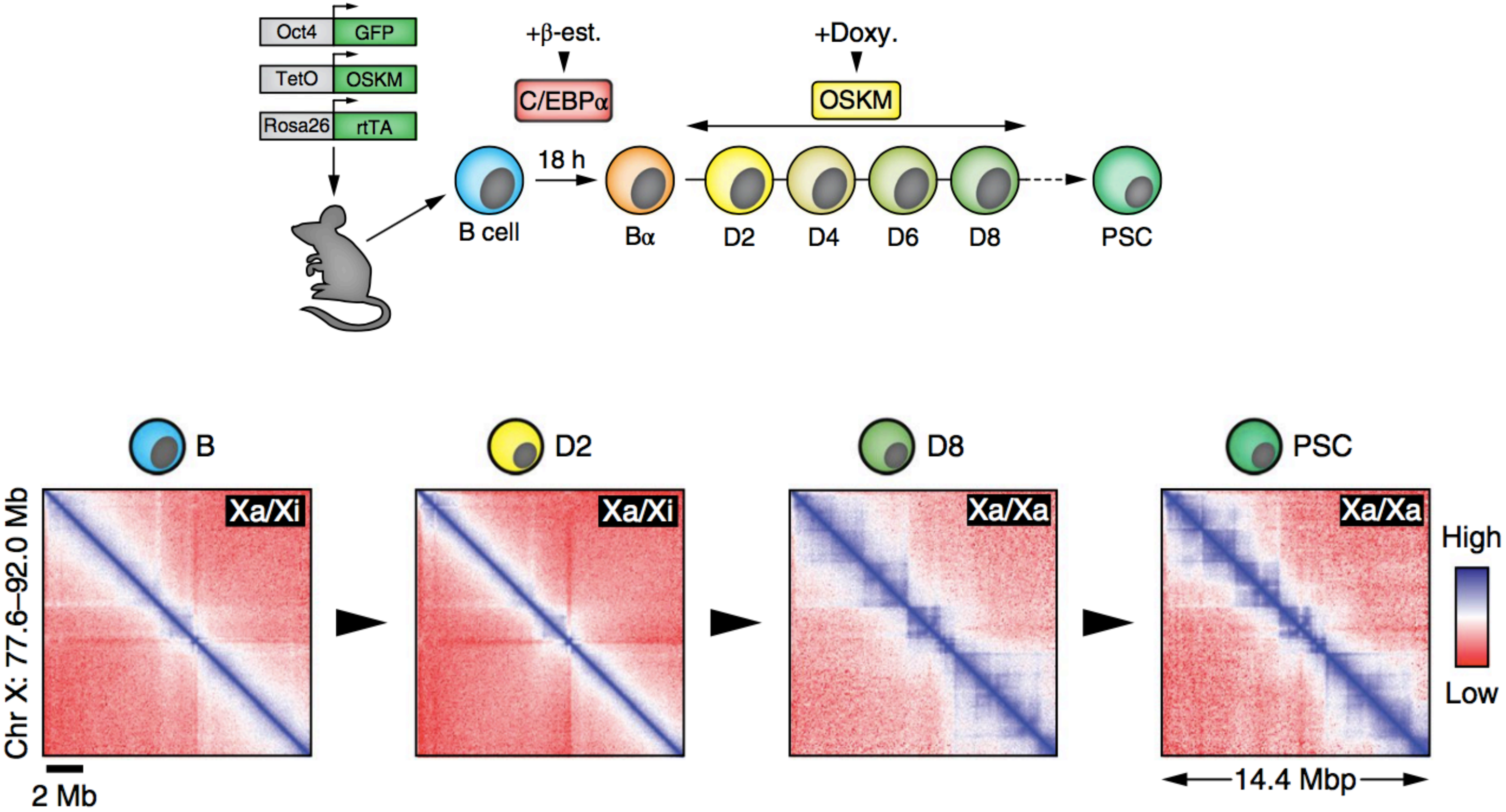




Ulyanov, 2016, Genome Research

"Time series"

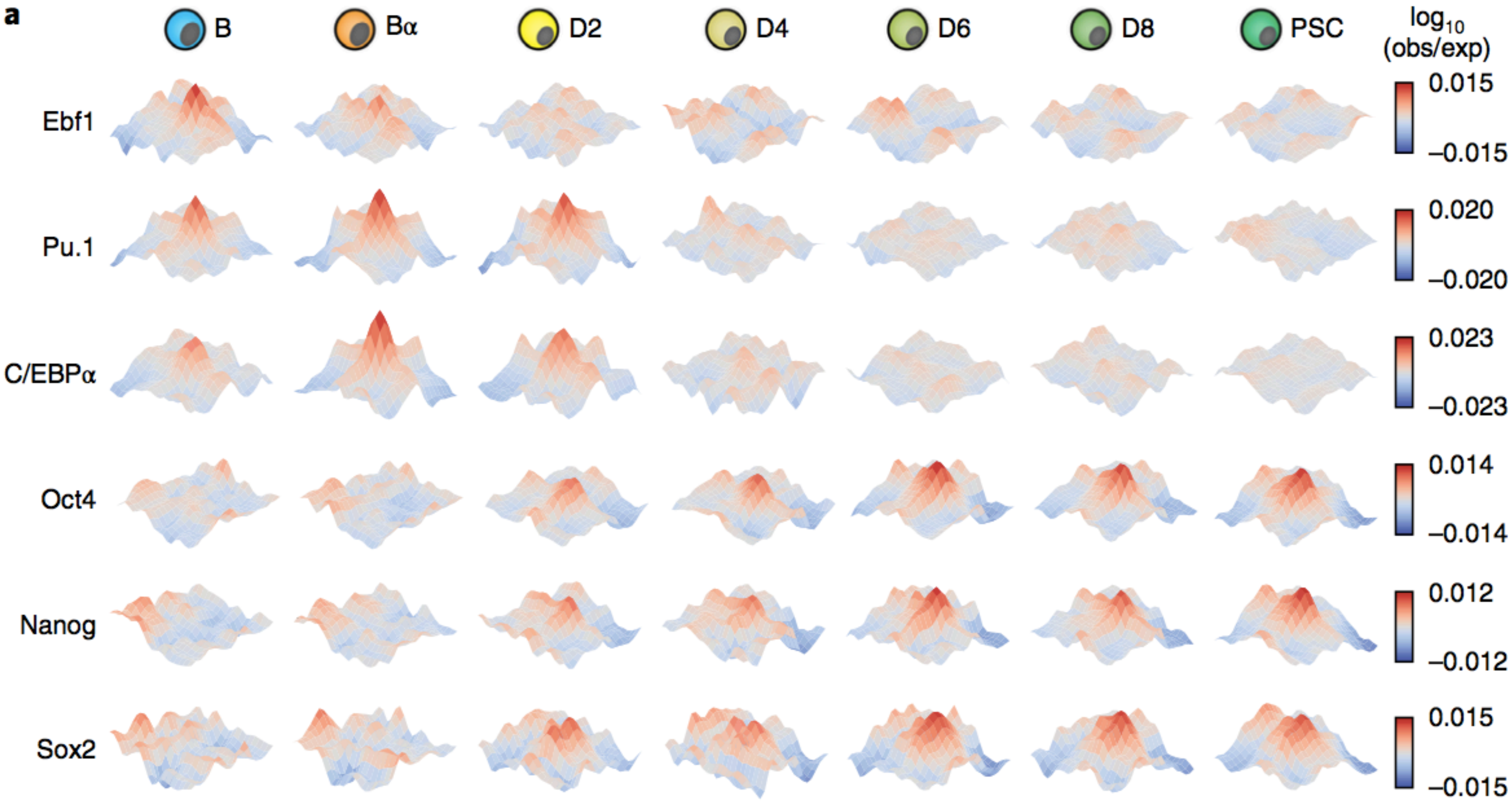
- For example, upon induction of pluripotency we can observe amazing topological transitions of chromatin:



Stadhouders, Nature Genetics, 2018

"Time series"

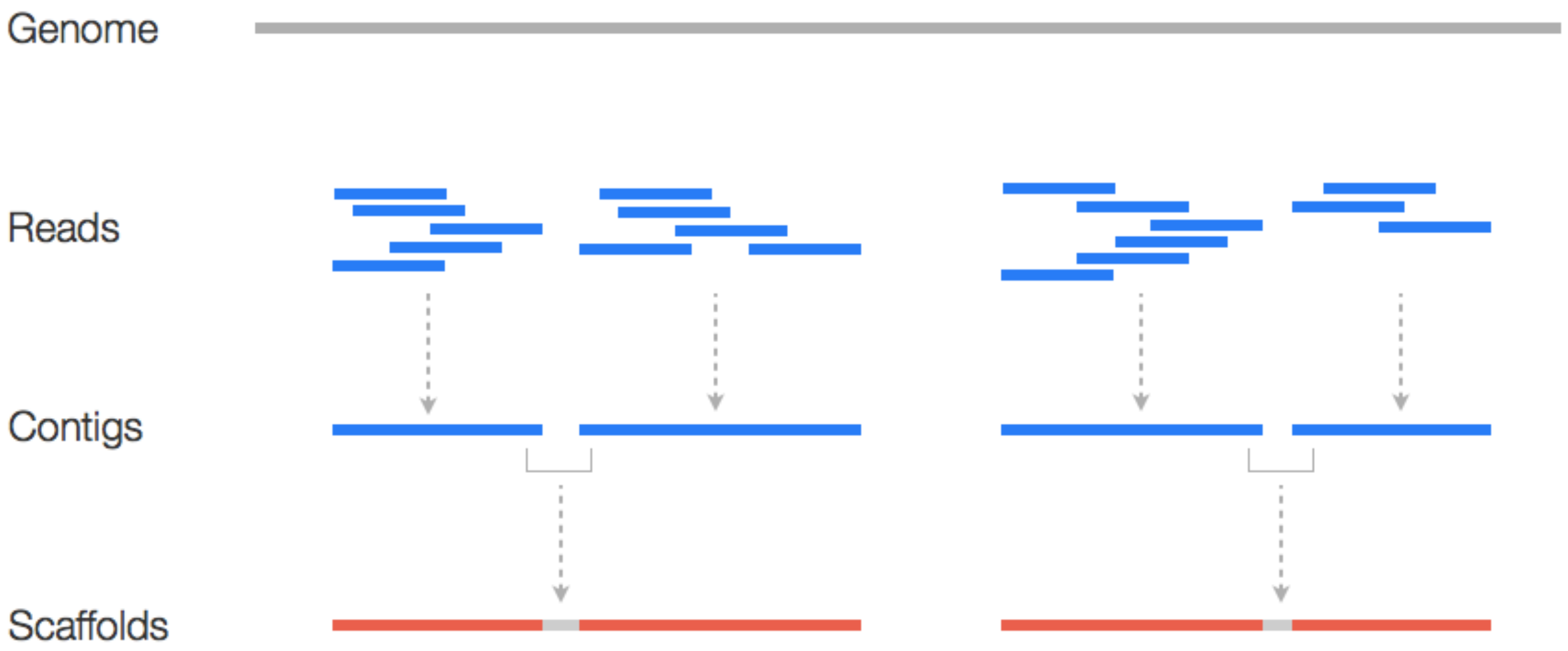
- Or we can observe how interactions of factors emerge:



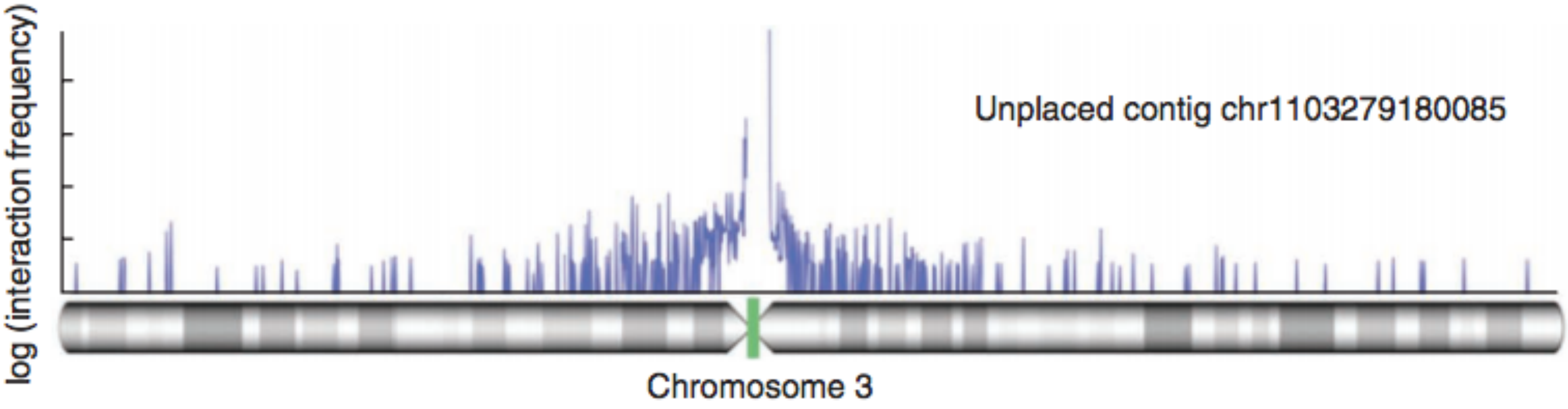
Stadhouders, Nature Genetics, 2018

Genome *de novo* assembly with Hi-C

- Conventional genome assembly:



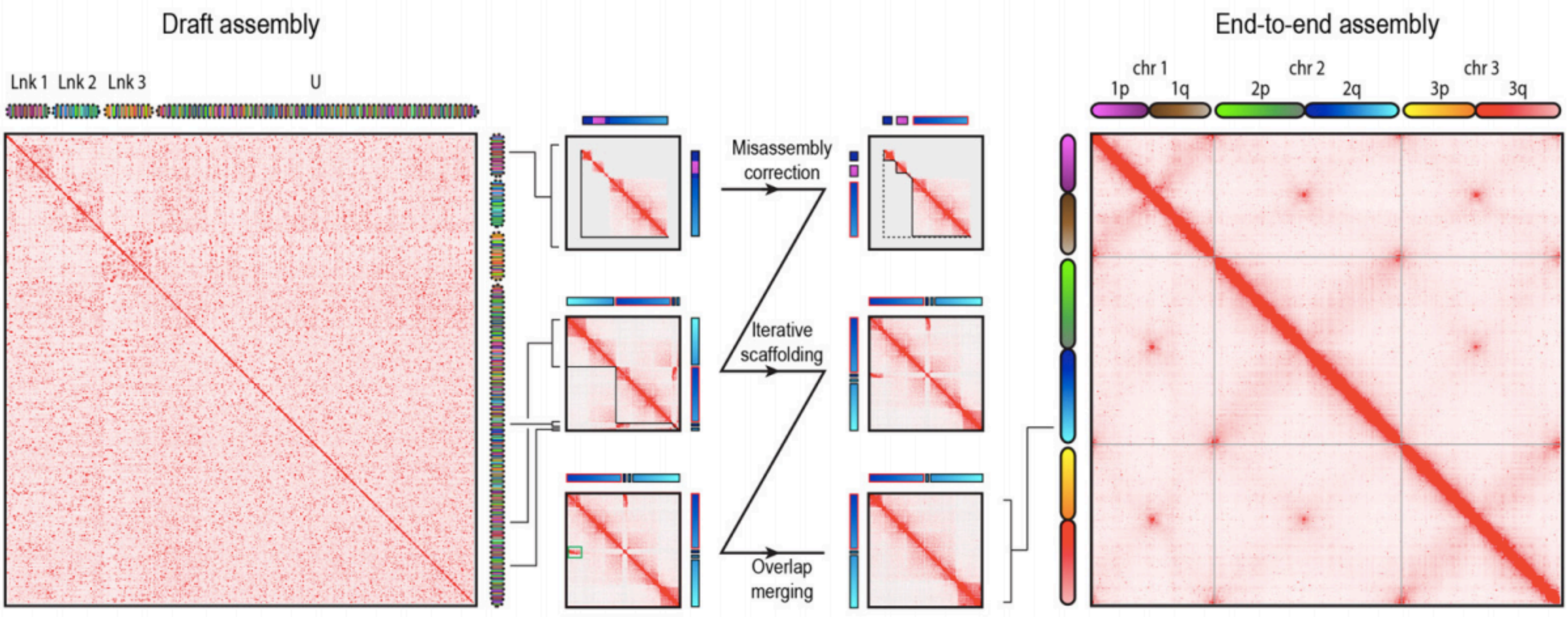
- Incorporation of Hi-C data:



Kaplan & Dekker, Nature Letters, 2013

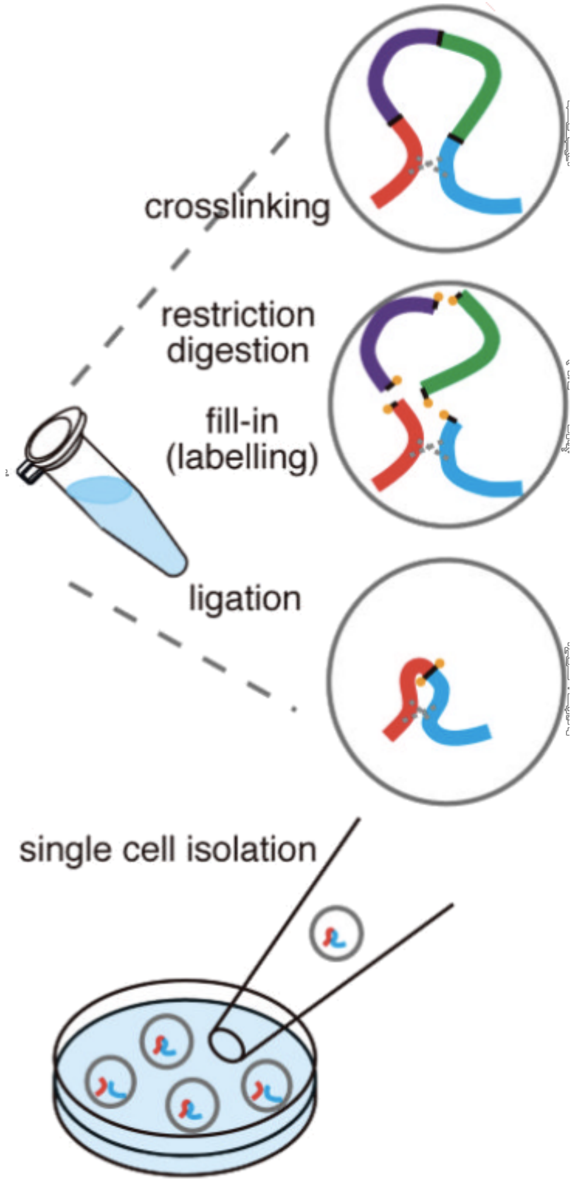
Genome *de novo* assembly with Hi-C

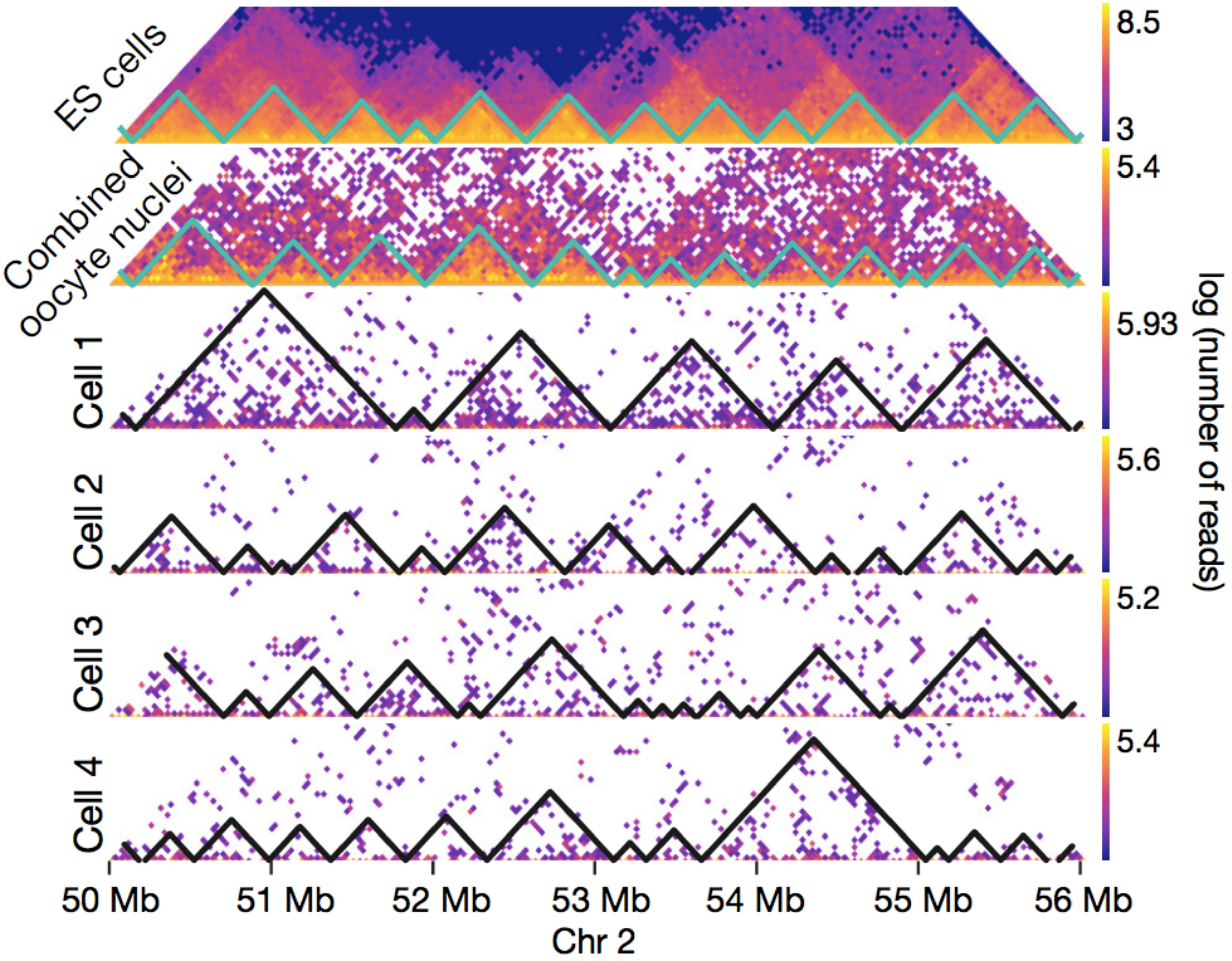
Aedes aegypti (the yellow fever mosquito)



Dudchenko, Science, 2017

Single-cell Hi-C

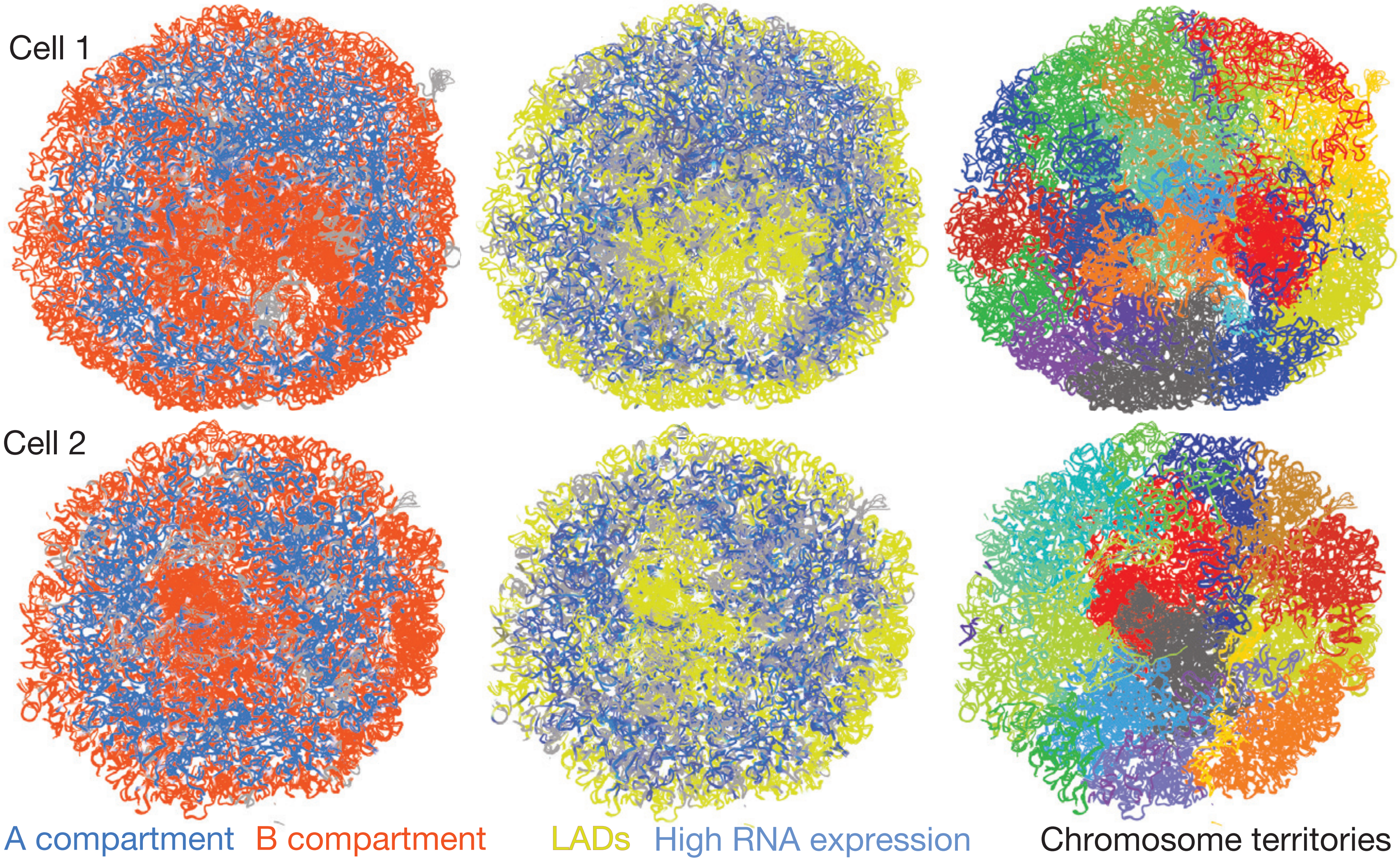




Flyamer et al. *Nature* 2017

Single-cell Hi-C

Data modelling based on single-cell can be very powerful, it can reproduce results visible previously by microscopy:



Stevens et al. *Nature* 2017

4. From theory to practice: EpiPractice1

Why you should not study low-level Hi-C data processing?

- Easy to use ready-made Hi-C maps and browsers Hi-C (GEO, ENCODE, 4dnucleome).
- Lots of methods for Hi-C data processing, no golden standard. Example of existing toolkits:

	Language	Year
Fit-Hi-C	Python	2014
GOTHiC	R	2015
HOMER	Perl, R	2010
HIPPIE	Python, Perl, R	2015
diffHic	R, Python	2015
HiCCUPS / Juicer	Java	2014, 2016
Juicer	Java	2016
TADbit	Python	2017
hiclib	Python	2012

- Constantly appearing new methods, written for one particular paper
- Sometimes it's easier to write your own processing pipeline…

Forcato et al. *Nature Methods* 2017

Some useful links

	4D Nucleome + Mirny lab	Lieberman-Aiden lab	Other labs
Hubs of recent updates	www.4dnucleome.org	www.aidenlab.org	
Data browsers	higlass.io data.4dnucleome.org	Juicebox	Promoter Yue lab browser
Data repos		Aiden lab datasets	ENCODE 3D-datasets
Online processing services			HiCExplorer on Galaxy
CLI/API processing tools	cooler docs cooler ipynb tutorial	Juicer	HiCExplorer

Practice outline

Two parts:

1. Hi-C data interpretation & Browsers comparison (10 pts)
    1. Yue lab Hi-C browser
    2. HiGlass
    3. Juicer
2. Hi-C data manipulation & Command line tools (10 pts)
    1. Setting up environment
    2. Hi-C data processing with CLI
        1. Data processing
        2. Data visualisation & TAD calling
        3. Data association

This is our seminar and home task (10 pts max per each part).
Send the reports in free form(doc or pdf with images) to Aleksandra.Galitsyna@skoltech.ru with subject:
" SK EpiPractice1 <Your name and surname>"
until 9th of April 23:00.

Each task has necessary sections (a, b, c, …) and additional (d*, …). You can get up to 20 points per this hometask. Optional tasks bring extra points that might compensate for incomplete or erroneous tasks

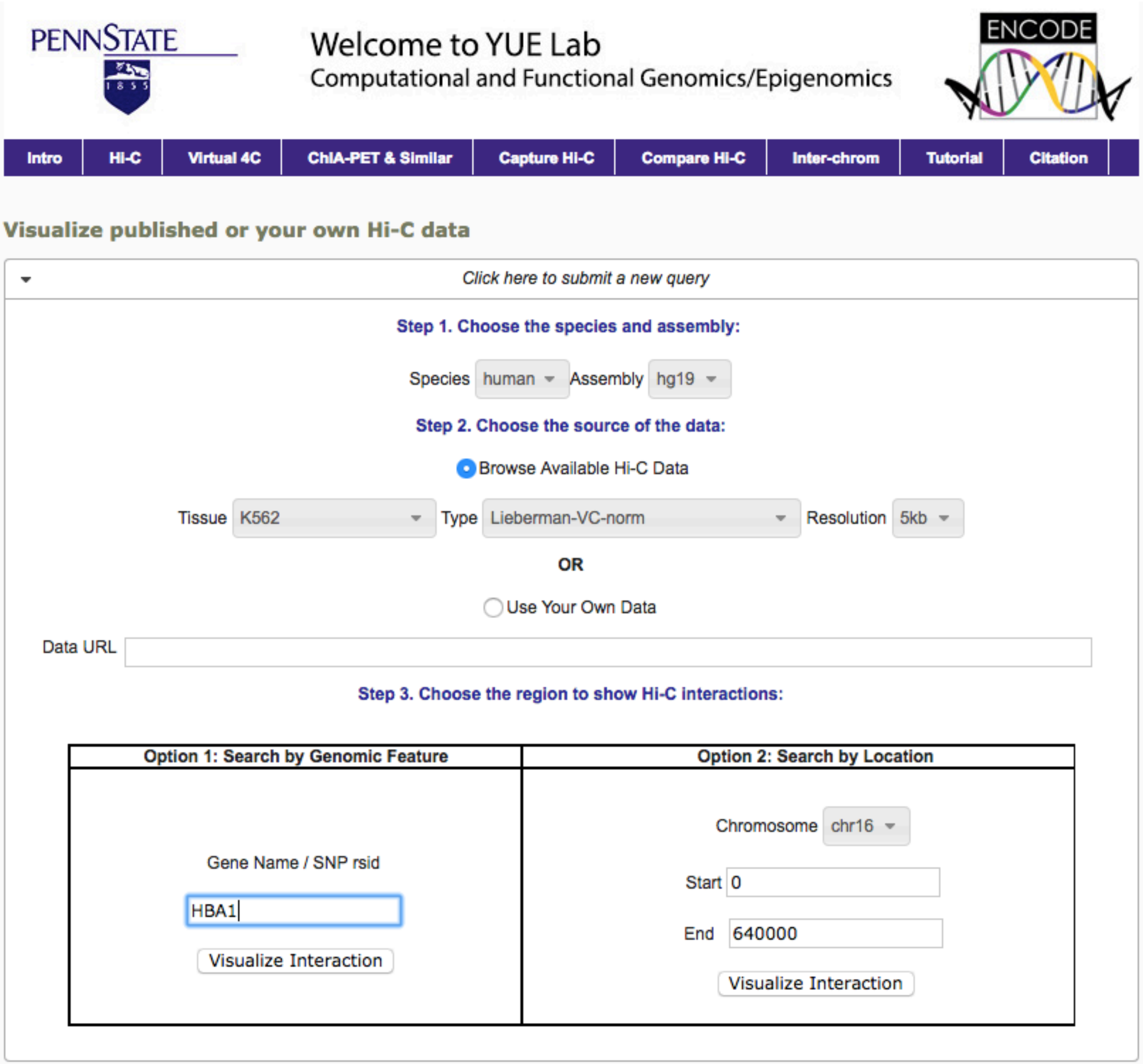
Task solutions for practices 1 and 2 will be presented at 10th of April, thus there is **no homework evaluation after 11:00 AM at 10th of April** (in case if you miss the deadline)

Task 1. Yue lab browser

- Go to "Promoter" browser from Yue lab (http://promoter.bx.psu.edu/hi-c/view.php)
- Select cell line K562 for genome assembly hg19, unbalanced Hi-C maps (raw), with 5 Kb resolution.
- Select the surroundings of HBA1 gene for view.
- Take a look a genes annotation. **Where the gene HBA1 is situated corresponding to TADs (in a TAD, at the boundary)? Report the screenshot and explanation.** (a)

Task 1. Yue lab browser

promoter.bx.psu.edu/hi-c/view.php



Welcome to YUE Lab

Computational and Functional Genomics/Epigenomics

Intro | Hi-C | Virtual 4C | ChIA-PET & Similar | Capture Hi-C | Compare Hi-C | Inter-chrom | Tutorial | Citation

Visualize published or your own Hi-C data

Click here to submit a new query

Step 1. Choose the species and assembly:

Species: human; Assembly: hg19

Step 2. Choose the source of the data:

● Browse Available Hi-C Data

Tissue: K562; Type: Lieberman-VC-norm; Resolution: 5kb

OR

○ Use Your Own Data

Data URL:

Step 3. Choose the region to show Hi-C interactions:

Option 1: Search by Genomic Feature	Option 2: Search by Location
Gene Name / SNP rsid: HBA1 Visualize Interaction	Chromosome: chr16 Start: 0 End: 640000 Visualize Interaction

Task 1. Yue lab browser

- Load the dataset for the same cells (K562, Hi-C) with VC-correction. Has the map changed? **Report the difference, demonstrate the TAD close to HBA1 gene and send the picture.** (b)
- Take a look at DHS - DNase I hypersensitive sites, site of accessible chromatin. **Are there many DHS close to HBA1? Report and propose an explanation.** (c)
- HBA1 - is a globin gene involved in oxygen transport. K562 is a erythroleukemia-derived cell type. **Can you propose a biological explanation of observed HBA1 state? Would you expect the same effect in other cell lines? Provide your answer with proof.** (d*)

Task 2. Comparison of Hi-C in HiGlass

- Go to another Hi-C browser HiGlass: http://higlass.io/
- Go to "Two Linked Views" and adjust the view: human hg19 genome, position chr8:107,328,268-109,258,572 & chr8:107,461,673-109,232,887 [offset 0,0:0,0], comparison between GM12878 and K562.
- **Describe the difference between these two cell lines, loops, TADs or compartments.** (a)
- Change the heatmap properties. **Find the colouring pattern that makes both datasets look qualitatively the same. Send the screenshot.** (b)
- **Is the quality of the datasets the same? Describe the difference and possible qualitative effect of that.** (c)

Task 2. hi-C comparison in HiGlass

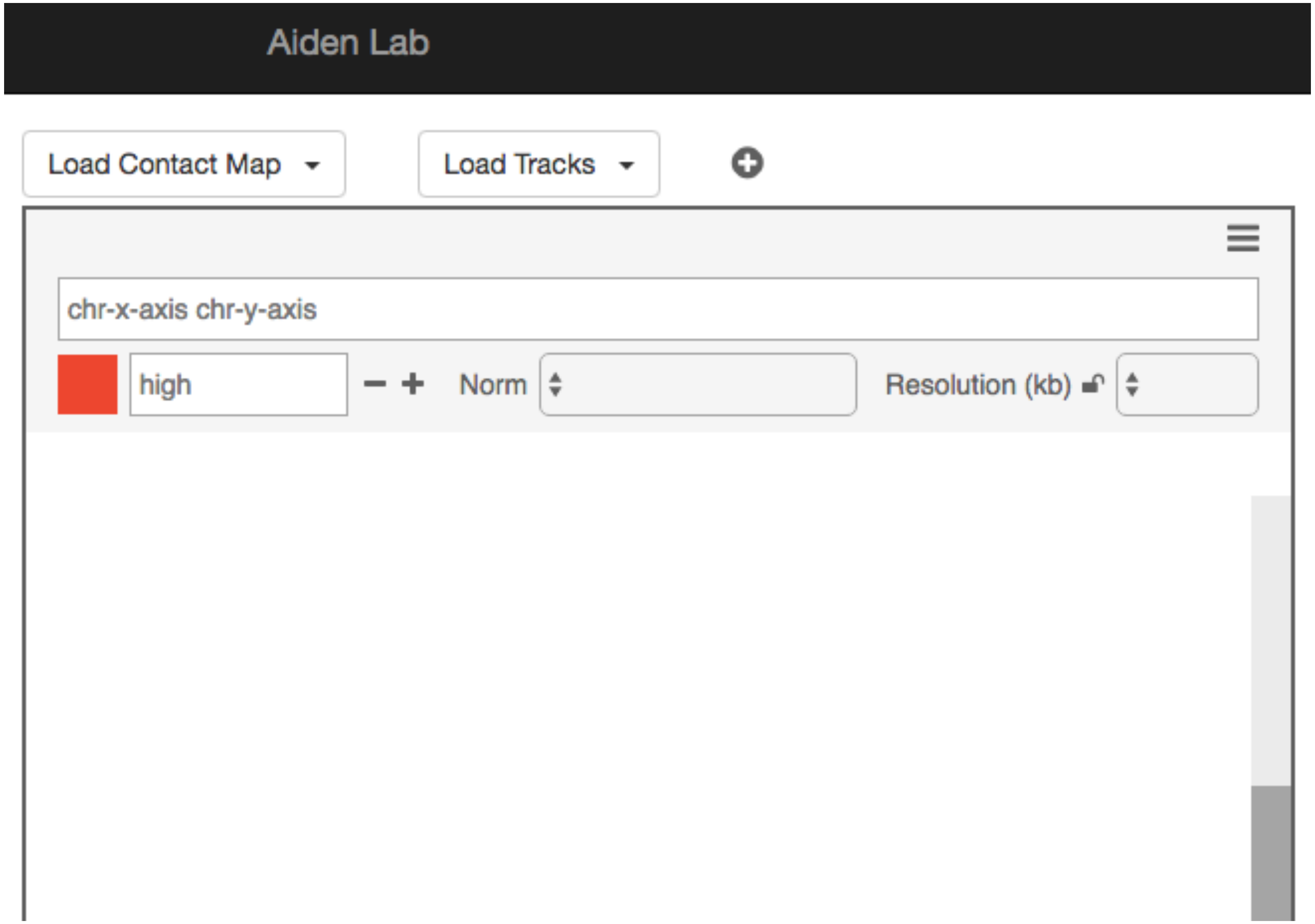
- Use the manual here: https://hms-dbmi.github.io/hic-data-analysis-bootcamp/#45 (slides 45-57) and adjust the view:
  hg19 genome
  central window Wutz2017.HeLa.Control_ProM_sync
  right window Wutz2017.HeLa.Control_G1_sync.

  Use the following options: Zoom limit 16 K, ICE

  This is the dataset for synchronised cells on mitosis and G1 phase. **Send the screenshot with description of differences. Propose biological explanation of the effect.** (d*)

Task 3. Loops, TADs and CTCF

- Go to online version of Juicebox from Aiden lab: http://www.aidenlab.org/juicebox



Task 3. Loops, TADs and CTCF

- Load Hi-C (!) for the same cells K562, se the resolution to 10 Kb, select the approptiate color scale (so that you can see TADs and loops). Select Balanced correction of the map. **Send a screenshot of some region with marked loops and TADs. Describe your observations.** (a)
- Load the TADs and loops annotation (Load tracks -> 2D Annotations -> combined domains, combined loops). **Is the annotation the same as you prodicted? Describe the difference between your expert judgement and software annotation.** (b)
- Load CTCF track for this cell line (Load tracks -> Genome Annotations -> CTCF). **Send the screenshot. Is CTCF associated with any Hi-C structures and why? Describe briefly in your report.** (c*)

Task 4. Command line tools

Currently there is no gold standard in raw Hi-C data processing.
Let's consider already prepared file with interactions heatmap in .cool format.

Processing steps:

1. Statistics retrieval
2. Changing resolution
3. Format conversions
4. Iterative correction
5. TADs calling
6. Data visualisation
7. TADs boundaries enrichment with ChIP-Seq data

Task 4. Command line tools

We will use the following tools:

- **cooler** for .cool manipulations (https://github.com/mirnylab/cooler)
- **HiCExplorer tools** for Hi-C data conversion, processing and visualisation (https://hicexplorer.readthedocs.io/en/documentation/content/list-of-tools.html)
- **deeptools** for data association with ChIP-Seq (http://deeptools.readthedocs.io/en/develop/content/list_of_tools.html)

Data formats:

- cooler is a sparse, compressed, binary persistent storage format for genomes interactions data
- h5 is some Hi-C data format used by HiCExplorer
- bed
- bigWig format for ChIP-Seq

0. Environment setup

- All the necessary packages are installed in anaconda environment at mg.uncb.iitp.ru server. Thus it's highly recommended to work there:

```
ssh -p9022 username@mg.uncb.iitp.ru
mkdir EpiPract1
cd EpiPract1
unset PYTHONPATH
export PATH="/mnt/local/bioinf_labs/home/galitsyna/anaconda3/bin:$PATH"
```

- Test for proper setup:

```
ls
pwd
conda list
deeptools --help
```

- Placement of all the datasets:
  `/mnt/local/bioinf_labs/home/galitsyna/DATA/EpiPract1`
- ChIp-Seq annotation files for different proteins and cell lines of *Drosophila*::
  `/mnt/local/bioinf_labs/home/galitsyna/DATA/EpiPract1/ANNOTATION/`
- Hi-C data files for different cell lines of *Drosophila*:
  `/mnt/local/bioinf_labs/home/galitsyna/DATA/EpiPract1/COOL/`

0. Exercise files variants

	cool file	ChIP-Seq file
Bella Bokan	BG3.10000.cool	BG3-Chriz.bigWig
Dilfuza Djamalova	BG3.10000.cool	BG3-CTCF.bigWig
Natalia Dranenko	BG3.10000.cool	BG3-H3K4me3.bigWig
Hilary Edema	BG3.10000.cool	BG3-JIL1.bigWig
Elizaveta Grigorashvili	BG3.10000.cool	BG3-RNAPolII.bigWig
Valeriia Kriukova	BG3.10000.cool	BG3-Su(Hw).bigWig
Ira Lisevich	BG3.10000.cool	BG3-WDS.bigWig
Anastasia Lubinets	Kc167.10000.cool	Kc167-Chriz.bigWig
Daniil Lukyanov	Kc167.10000.cool	Kc167-CTCF.bigWig
Valeriya Mikova	Kc167.10000.cool	Kc167-H3K4me3.bigWig
Anna Rybina	Kc167.10000.cool	Kc167-JIL1.bigWig
Marina Sarantseva	Kc167.10000.cool	Kc167-RNAPolII.bigWig
Natalia Trankova	Kc167.10000.cool	Kc167-Su(Hw).bigWig
Anastasiia Velikanova	Kc167.10000.cool	Kc167-WDS.bigWig
Artemy Zhigulev	S2.10000.cool	S2-Chriz.bigWig
Kulash Zhumadilova	S2.10000.cool	S2-CTCF.bigWig
Aleksandra Galitsyna	S2.10000.cool	S2-H3K4me3.bigWig

1. Statistics retrieval

Cooler contains multiple functions for cool manipulations, let's to find the number of contacts in file, e.g.:

```
cooler info OSC.10000.cool
```

Q. 1. What is the genome assembly, the resolution and number of contacts in your file?

Send me the answer in free text form as A.1

2. Changing resolution

We have data files with resolution 10000 bp (10 Kb), let's make it 20000 bp (20 Kb):

```
cooler coarsen -k 2 -o OSC.20000.cool OSC.10000.cool
```

This step has no answer, but it's required further

3. Format conversions

- cool is a very "young" format and some tools are not adjusted to process it. Thus file conversion is needed. hicExport tool from HiCExplorer can convert in between common Hi-C formats (https://hicexplorer.readthedocs.io/en/documentation/content/tools/hicExport.html).
- Let's convert cool to HiCExplorer format h5:

```
hicConvertFormat --matrices OSC.20000.cool --outFileName OSC.20000.h5 \
--inputFormat cool --outputFormat h5
```

You can now check file info with HiCExplorer (see chromosomes names, for example):

```
hicInfo -m OSC.20000.h5
```

This step has no answer, but it's required further

4. Iterative correction

- Now we need to normalise our dataset and correct for experimental biases with hicCorrectMatrix (https://hicexplorer.readthedocs.io/en/documentation/content/tools/hicCorrectMatrix.html):

```
hicCorrectMatrix correct --matrix OSC.20000.h5 \
--filterThreshold -10 10 -n 10 --out OSC.corr.20000.h5

hicPlotMatrix -m OSC.20000.h5 -o OSC.raw.mtx.png --log1p \
--clearMaskedBins --region chrX:10000000-12000000

hicPlotMatrix -m OSC.corr.20000.h5 -o OSC.corr.mtx.png --log1p \
--clearMaskedBins --region chrX:10000000-12000000
```

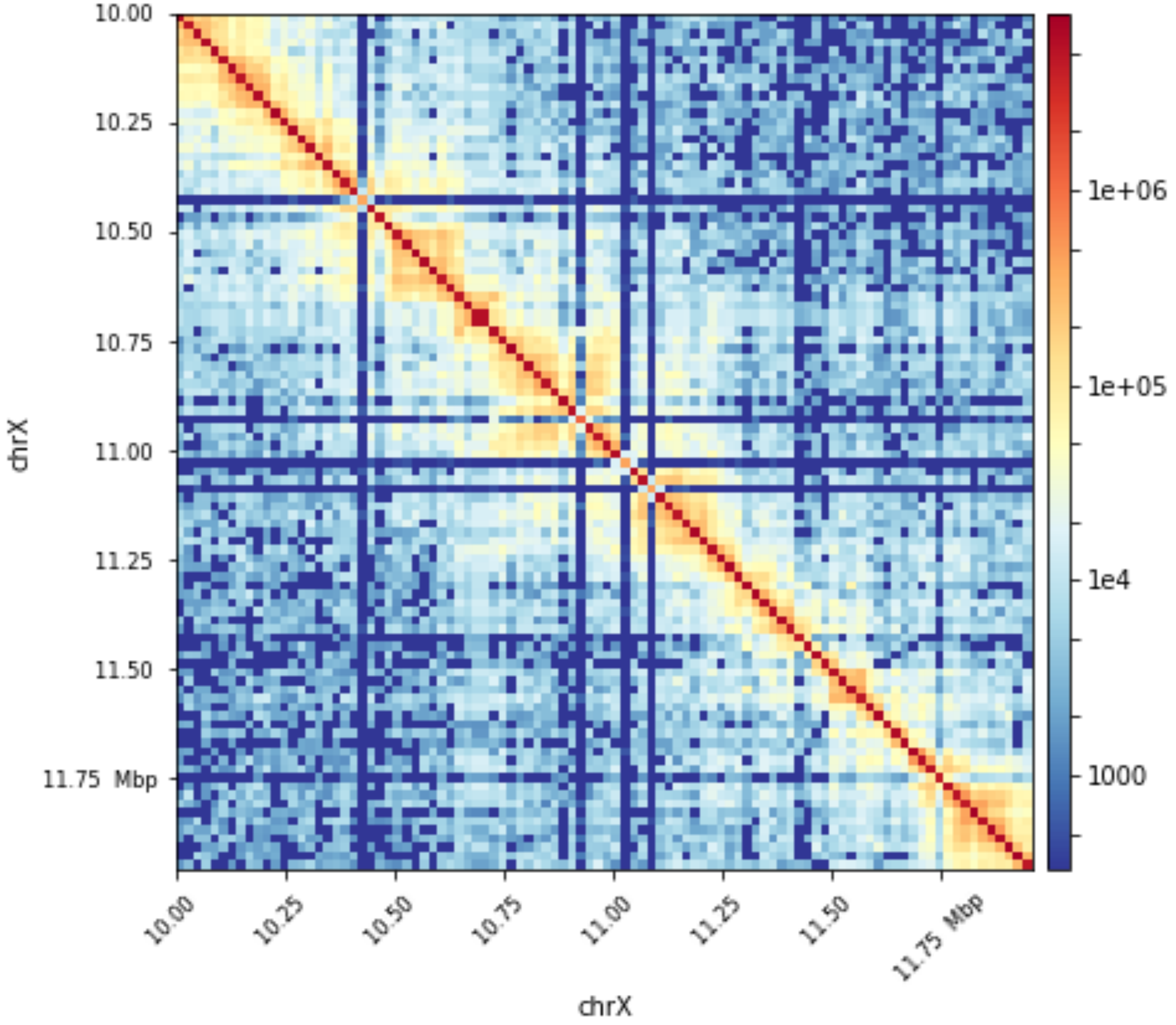
Q. 2. What is the difference between heatmaps? Send me both files and your brief observations.

Try to select the best parameters for the correction and visualisation. E.g. you can select any region of any size where you can see difference the most.

Send me the answer in free text form with 2 images as A.2

4. Iterative correction

- This is my result, but the visualisation might be improved, try different parameters for that:



5. TADs calling

- Let's try to call TADs in our dataset with TADs caller implemented in HiCExplorer (https://hicexplorer.readthedocs.io/en/documentation/content/tools/hicFindTADs.html). The concept is very similar to Insulation Score (IS).

```
hicFindTADs -m OSC.corr.20000.h5 --outPrefix OSC_TADs \
--minDepth 60000 --maxDepth 1000000 --step 20000 \
--thresholdComparisons 0.05 --delta 0.01 \
--correctForMultipleTesting fdr
```

- This command creates a list of files:

OSC_TADs_boundaries.bed
OSC_TADs_boundaries.gff
OSC_TADs_domains.bed
OSC_TADs_score.bedgraph
OSC_TADs_tad_score.bm
OSC_TADs_zscore_matrix.h5

This step has no answer, but it's required further

6. Data visualisation

- Let's plot interaction heatmap and TADs together:

```
hicPlotTADs --tracks tracks.ini --region chr2L:1000000-4000000 \
-o OSC.TADs.png
```

- As you can see, TADs visualisation with HiCExplorer required tracks.ini file with plot description. It seems to be quite complex, though it allows to adjust the very detail of your plot: http://hicexplorer.readthedocs.io/en/documentation/content/tools/hicPlotTADs.html?highlight=tracks.ini
- The minimal working version of tracks.ini file is placed at the next slide. It might be improved. Try to change parameters in the file and produce the better visualisation of TADs and heatmap features.

Q. 3. Does the TADs found by algorithm correspond to what you see? Send me the visualisation and your brief observations. Note that expected size of TADs in *Drosophila* is 120 Kb.

Send me the answer in free text form with 1 image as A.3

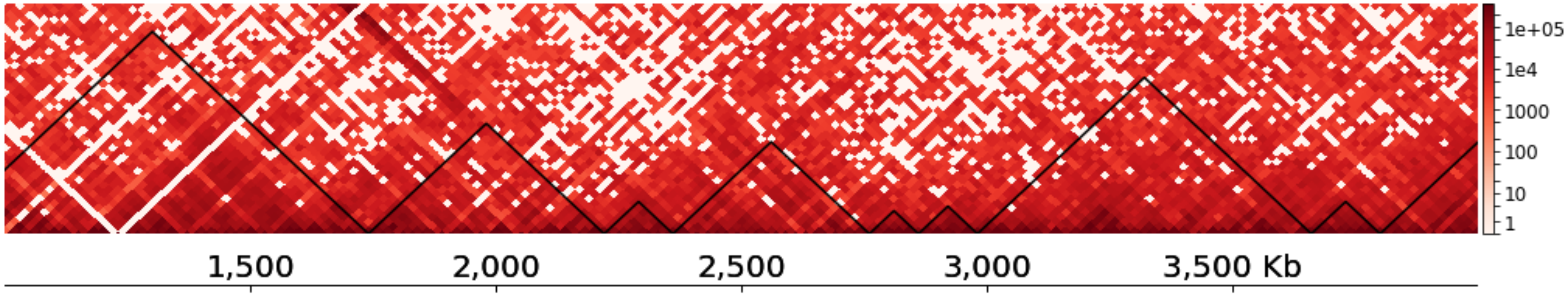
6. Data visualisation

```
cat tracks.ini

[hic]
file = OSC.corr.20000.h5
title = Hi-C
colormap = Reds
depth = 1000000
#min_value = 1
#max_value = 10000000
transform = log1p
boundaries_file = OSC_TADs_domains.bed
x labels = yes
type = interaction
file_type = hic_matrix
show_masked_bins = yes
scale factor = 1

[x-axis]
fontsize=20
where=top

[spacer]
width = 0.1
```



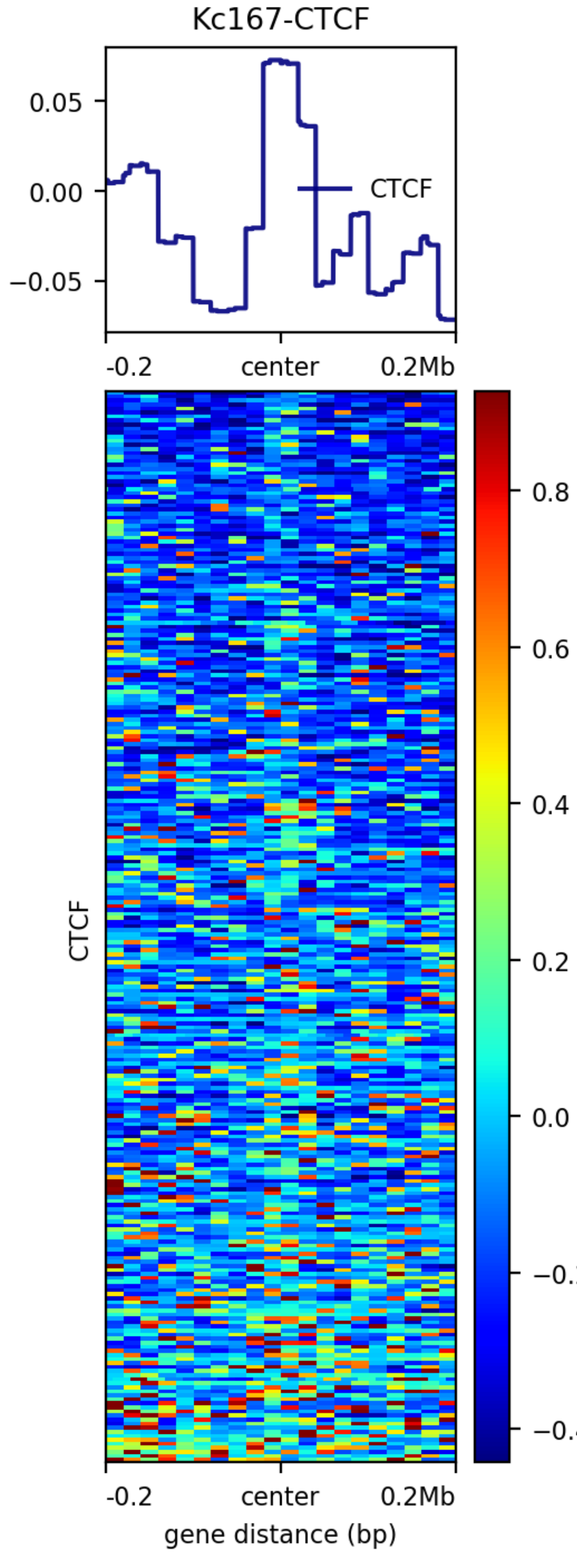
7. TADs boundaries enrichment with ChIP-Seq data

- Let's compare TADs boundaries with some ChIP-Seq profiles with deeptools (see http://deeptools.readthedocs.io/en/develop/content/example_gallery.html#dnase-accessibility-at-enhancers-in-murine-es-cells).

  Note that it might be time-consuming step! Try to select-a and -b smaller to make it faster or larger to make it more informative.

  ```
  computeMatrix reference-point -S S2-CTCF.bigWig \
  -R OSC_TADs_boundaries.bed --referencePoint center \
  -a 200000 -b 200000 -out matrix_enrichment.tab.gz

  plotHeatmap -m matrix_enrichment.tab.gz \
  -out enrichment.png --heatmapHeight 15 --colorMap jet \
  --sortRegions ascend  --regionsLabel 'CTCF'
  ```



- Q. 4. Is your factor enriched at TADs boundaries? Is there enough data to draw conclusions? Send me the visualisation and your brief observations.

Send me the answer in free text form with 1 image as A.4

8. Extra task *

- Collect the enrichment plots for the same factor, but for different cell types from your colleagues. Compare the results with yours. Is the abundance of factor the same at TAD boundaries?

  Add the results to your report and describe your observations for extra 2 points for this homework.

Expected exercise results of Task 4.

Report to Aleksandra.Galitsyna@skoltech.ru until 9th of April 23:00.
Subject: " SK EpiPractice1 <Your name and surname>"

Letter content (in free txt, word, pptx or whatever readable format):
Part 1. Work with Hi-C browsers. (10 pts)

Part 2. Description of your activity in command line highlighting:
Answer 1. Genome assembly, resolution and number of contacts in your file. (1 pt)
Answer 2. Two images (corrected and raw) of heatmaps for arbitrary genomic region with a brief description of differences. (2 pt)
Answer 3. One image with TADs plotted with interactions heatmap with a brief description. (3 pt)
Answer 4. One image with TADs boundaries enrichment with your factor with a brief description. (4 pt)
*Collecting results for different cell types from your colleagues and interpretation.
Note that each student has his own set of data files!
Extra points are added if you try to adjust commands parameters and send me the best final command (note that the final mark cannot exceed 20 pts).

In total: 20 pts