NGS School

INSTITUTE of GENE BIOLOGY
RUSSIAN ACADEMY of SCIENCES

RUSSIAN ACADEMY OF SCIENCES
INSTITUTE FOR INFORMATION TRANSMISSION PROBLEMS
(Kharkevich Institute)

RTCB

Skoltech
Skolkovo Institute of Science and Technology

Single-cell Hi-C data analysis

Introductory lecture for NGS School 2017 workshop

Aleksandra Galitsyna

ARTICLE

2013

doi:10.1038/nature12593

Single-cell Hi-C reveals cell-to-cell variability in chromosome structure

Takashi Nagano[1]*, Yaniv Lubling[2]*, Tim J. Stevens[3]*, Stefan Schoenfelder[1], Eitan Yaffe[2], Wendy Dean[4], Ernest D. Laue[3], Amos Tanay[2] & Peter Fraser[1]

[1]Nuclear Dynamics Programme, The Babraham Institute, Cambridge CB22 3AT, UK. [2]Department of Computer Science and Applied Mathematics and Department of Biological Regulation, Weizmann Institute, Rehovot 76100, Israel. [3]Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK. [4]Epigenetics Programme, The Babraham Institute, Cambridge CB22 3AT, UK.

LETTER

2017

doi:10.1038/nature21711

Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition

Ilya M. Flyamer[1,2,3]*†, Johanna Gassler[1]*, Maxim Imakaev[4,5]*, Hugo B. Brandão[6], Sergey V. Ulianov[2,3], Nezar Abdennur[7], Sergey V. Razin[2,3], Leonid A. Mirny[4,5,6]§ & Kikuë Tachibana-Konwalski[1]§

[1]IMBA - Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna Biocenter (VBC), Dr Bohr-Gasse 3, 1030 Vienna, Austria. [2]Institute of Gene Biology, Russian Academy of Sciences, Moscow 119334, Russia. [3]Faculty of Biology, Lomonosov Moscow State University, Moscow 119234, Russia. [4]Institute for Medical Engineering and Science, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02139, USA. [5]Department of Physics, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02139, USA. [6]Harvard Program in Biophysics, Harvard University, Cambridge, Massachusetts 02138, USA. [7]Computational and Systems Biology Program, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02139 USA. †Present address: MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.

Single-cell Hi-C research in a large collaboration

- Kikuë Tachibana-Konwalski's lab from Institute of Molecular Biotechnology of the Austrian Academy of Sciences, VBC, Vienna, Austria
- Prof. Leonid Mirny's lab from MIT, Cambridge, Massachusetts, USA
- Prof. Sergey Razin's lab from Institute of Gene Biology of Russian Academy of Sciences (IGB RAS), Moscow, Russia
- Prof. Mikhail Gelfand's lab from Institute for Information Transmission Problems of Russian Academy of Sciences (IITP RAS), Moscow, and Skolkovo Institute of Science and Technology (Skoltech), Skolkovo, Russia

Outline

- Introduction
  - Eukaryotic chromatin structure
  - Hi-C and chromatin interaction map
  - Interaction map features: TADs, compartments, loops
  - Single-cell Hi-C
- From theory to practice: Hi-C data processing workflow
  - Reads mapping
  - Binning & filtering
  - Matrix balancing
  - TADs and compartments calling
  - Single-cell data analysis
- Workshop overview

1.1 Introduction: Eukaryotic chromatin structure

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

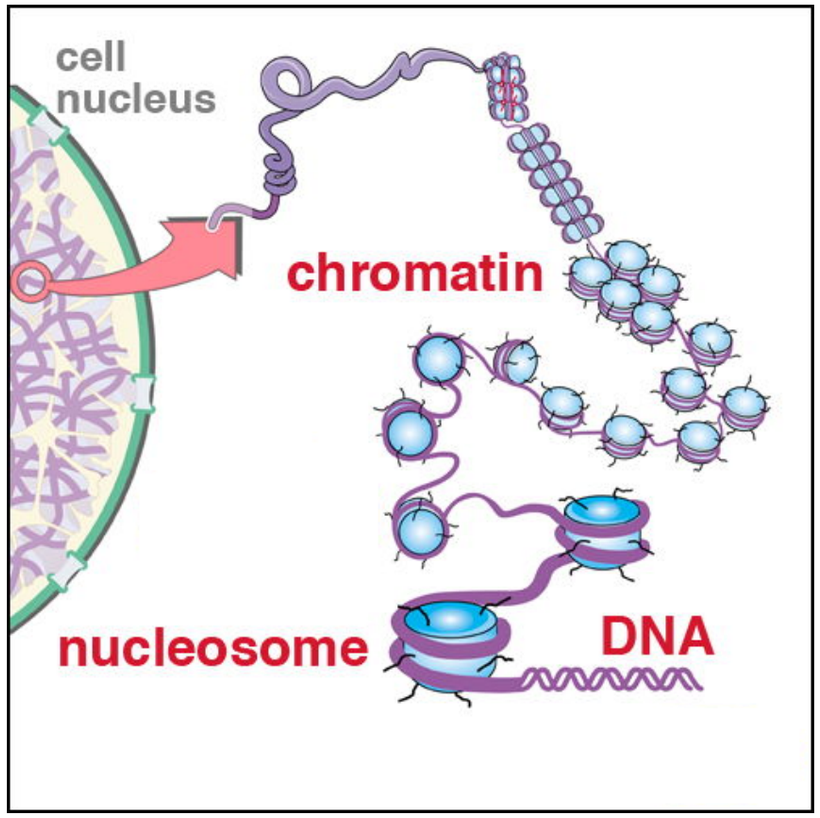
Histone modifications

Transcription factors binding

Non-coding RNAs

Nucleotide modifications

Binding to
the nucleolar envelope



Replication

Recombination

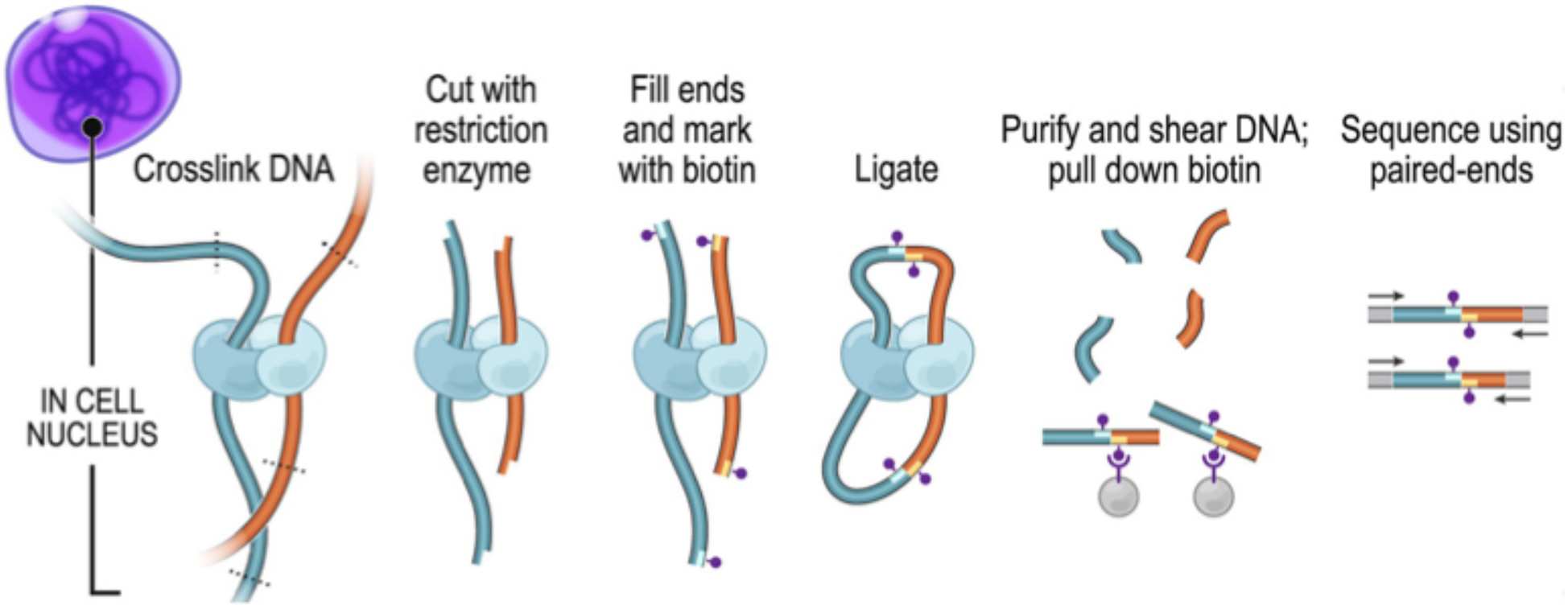
Regulation

Transcription

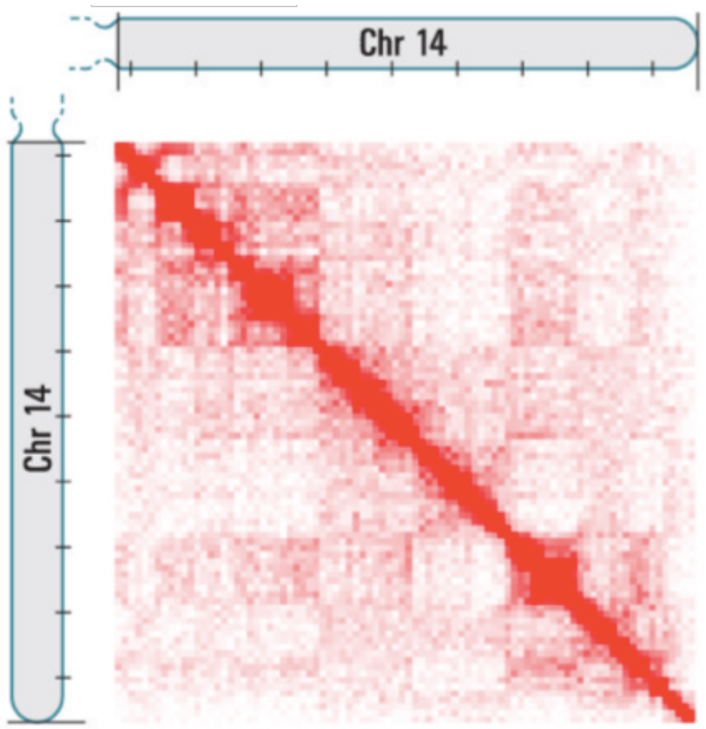
10^4-10^5 folding

1.2 Hi-C: high-throughput chromosomes conformation capture

Procedure:



Resulting interactions heatmap:



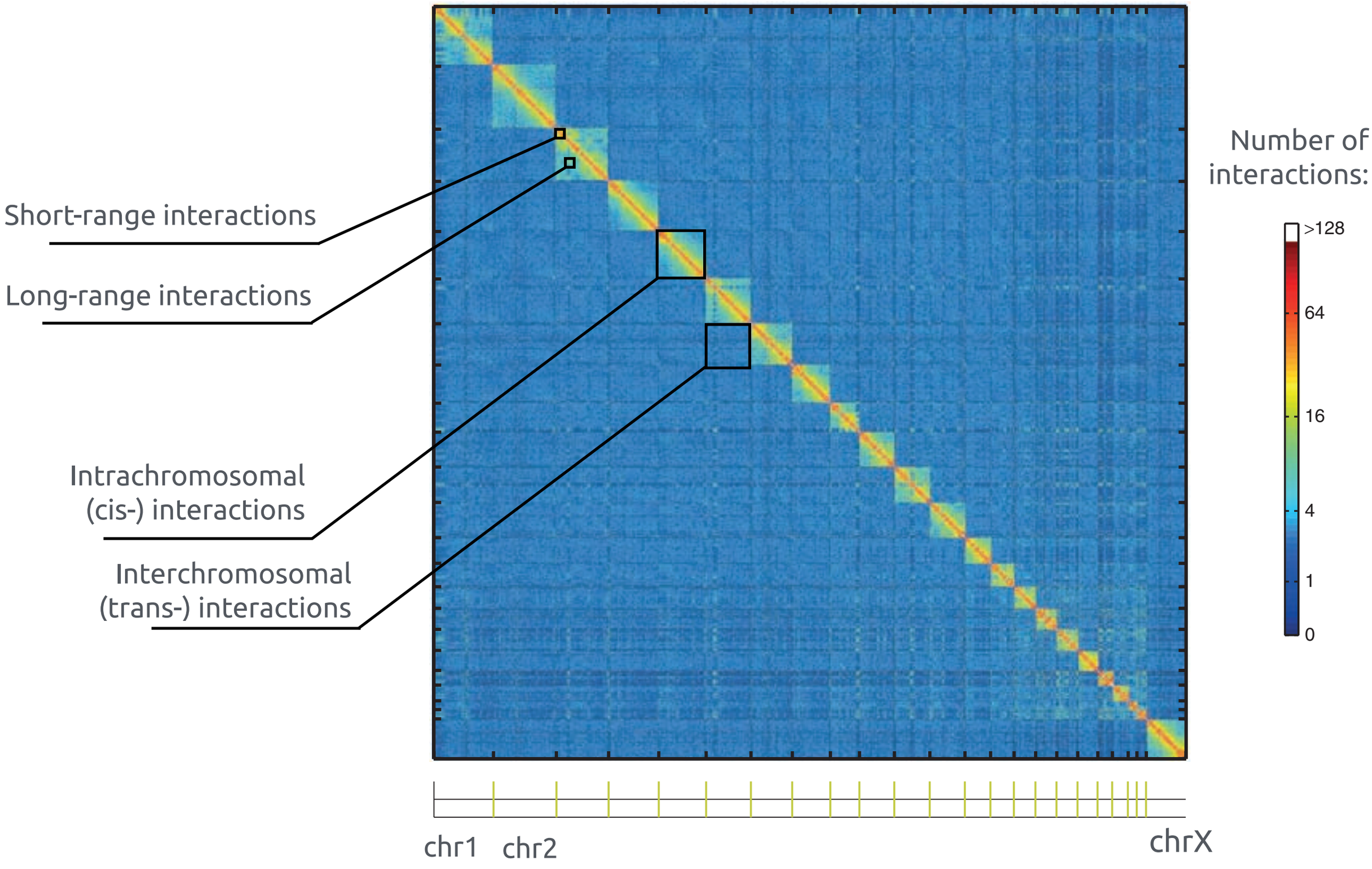
Lieberman-Aiden et al. *Science* 2009

1.2 Some of known conformation capture techniques

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

1.2 Chromatin interactions map

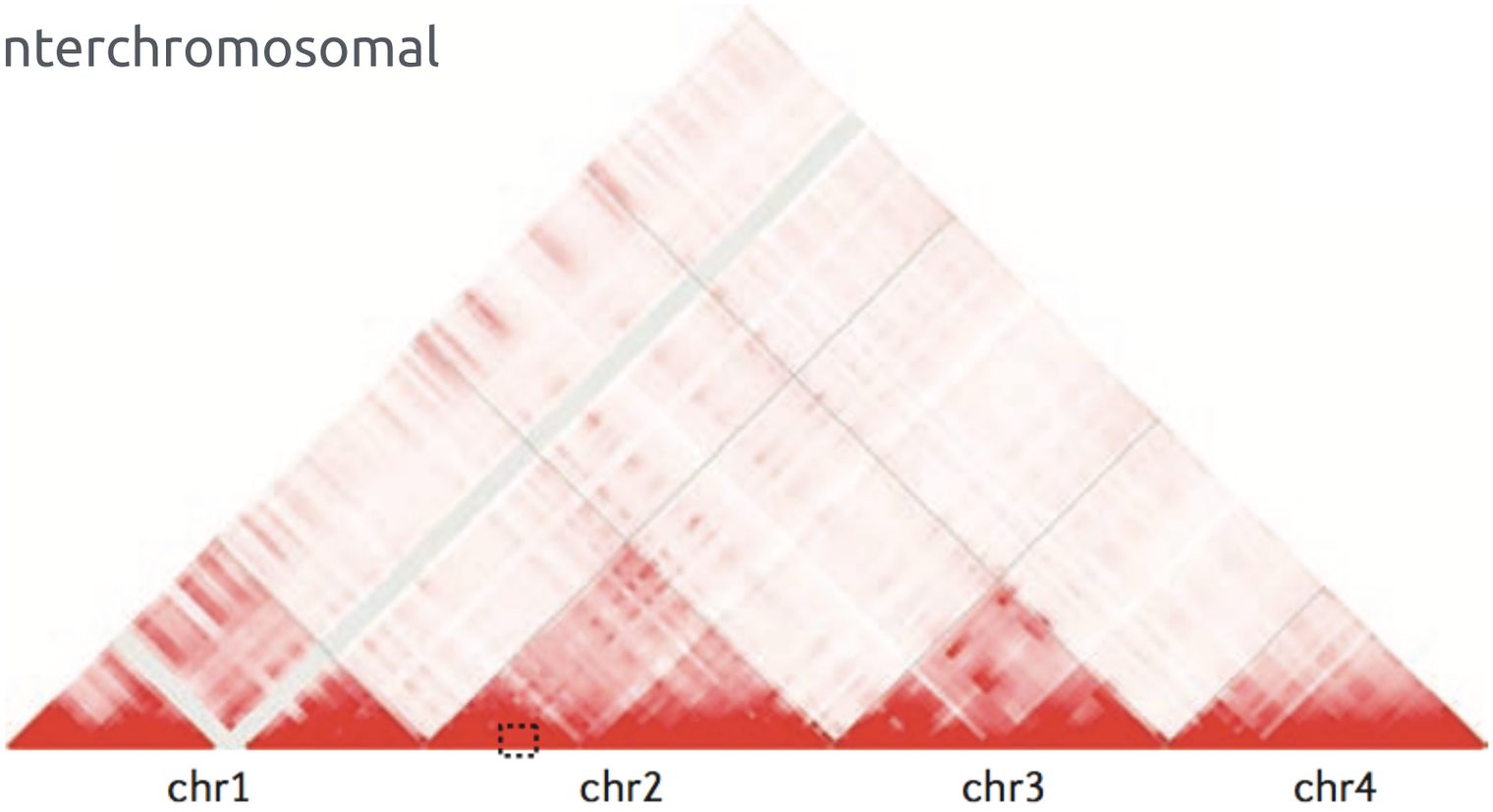


Adopted from Imakaev et al. *Nature Methods* 2012

1.3 Interaction map features: Chromosome territories

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

Interchromosomal

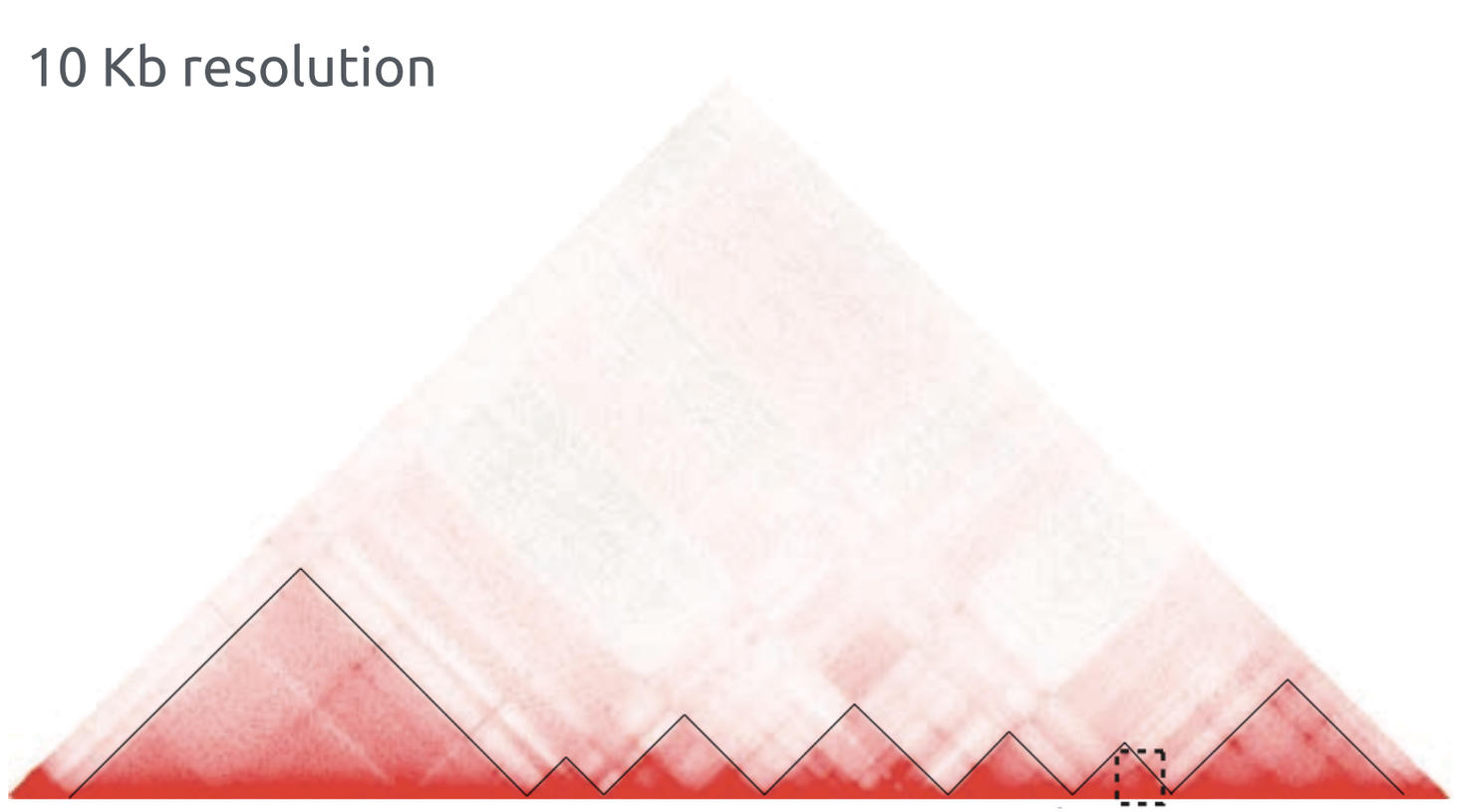


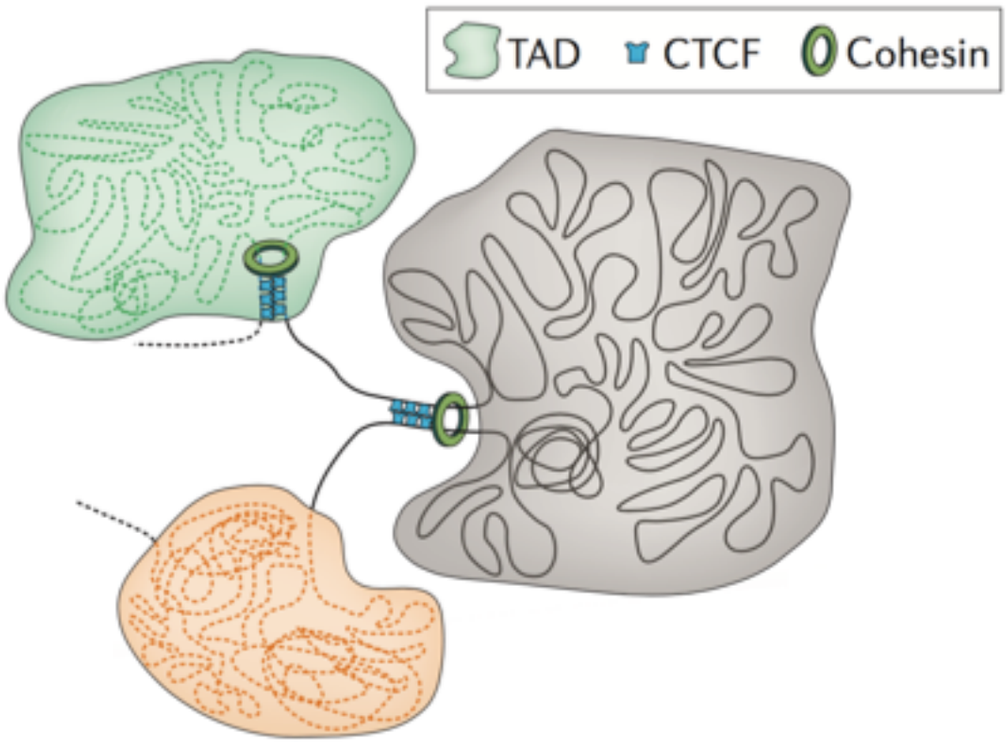
Map is rotated by 90°, upper triangle visualized

Bonev et al. *Nature Reviews* 2016

1.3 Topologically-associating domains (TADs)

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

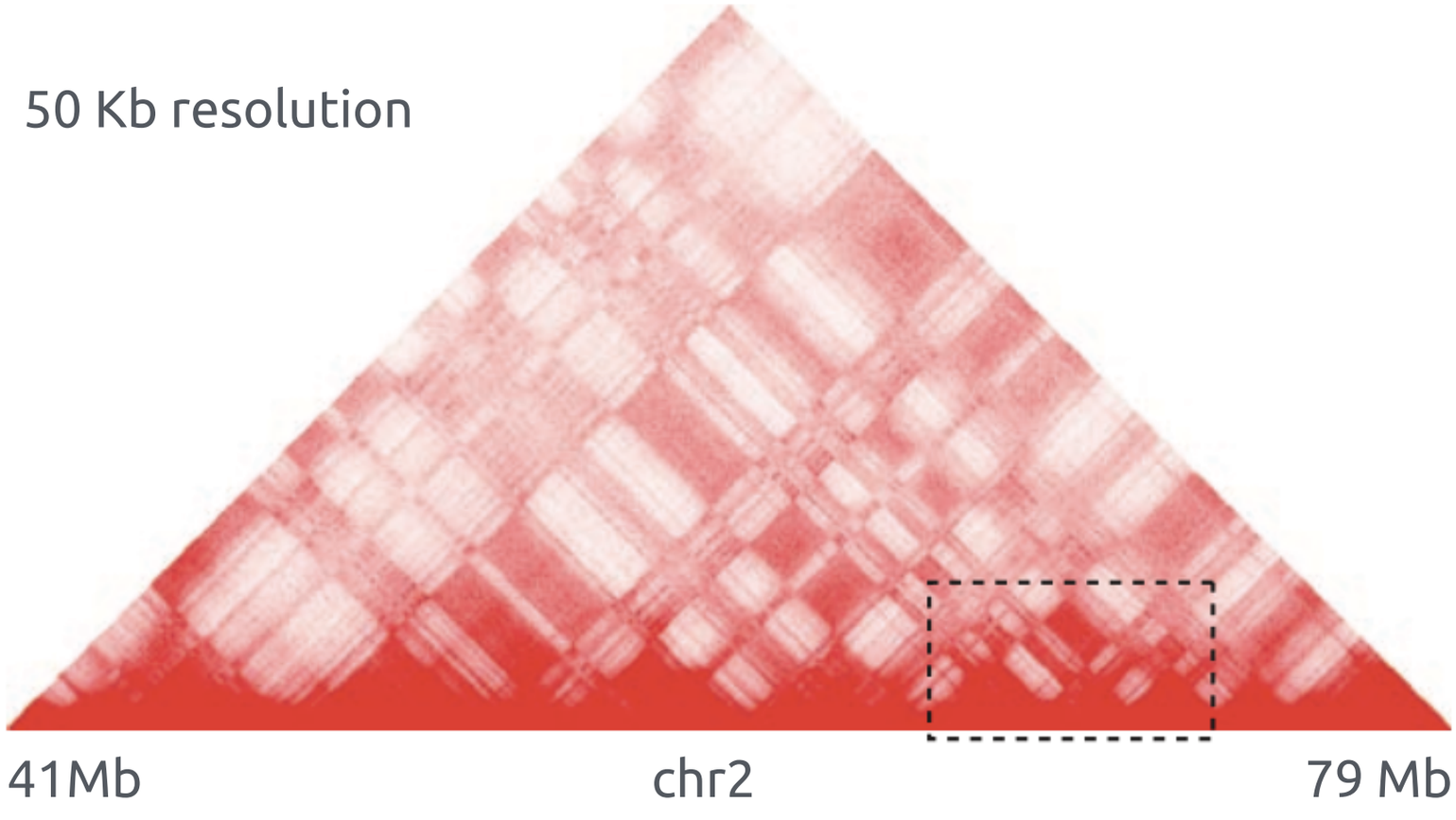


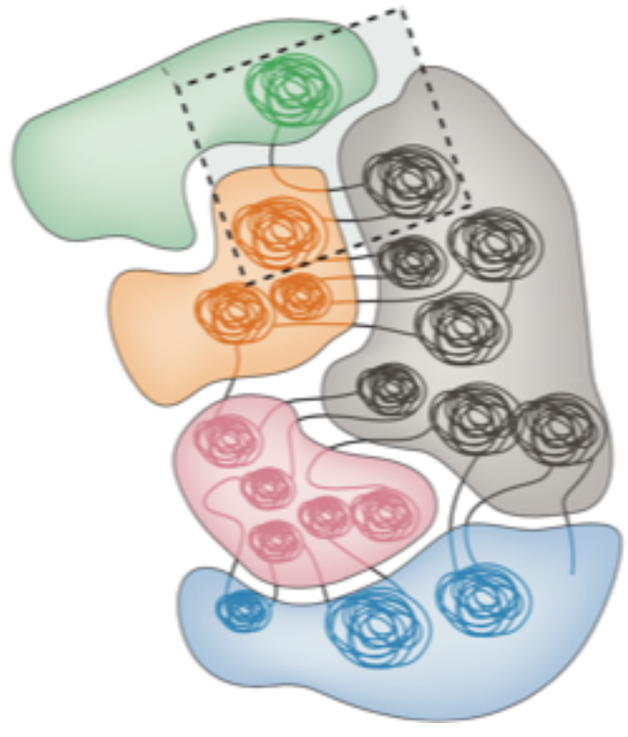


Bonev et al. *Nature Reviews* 2016

1.3 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.

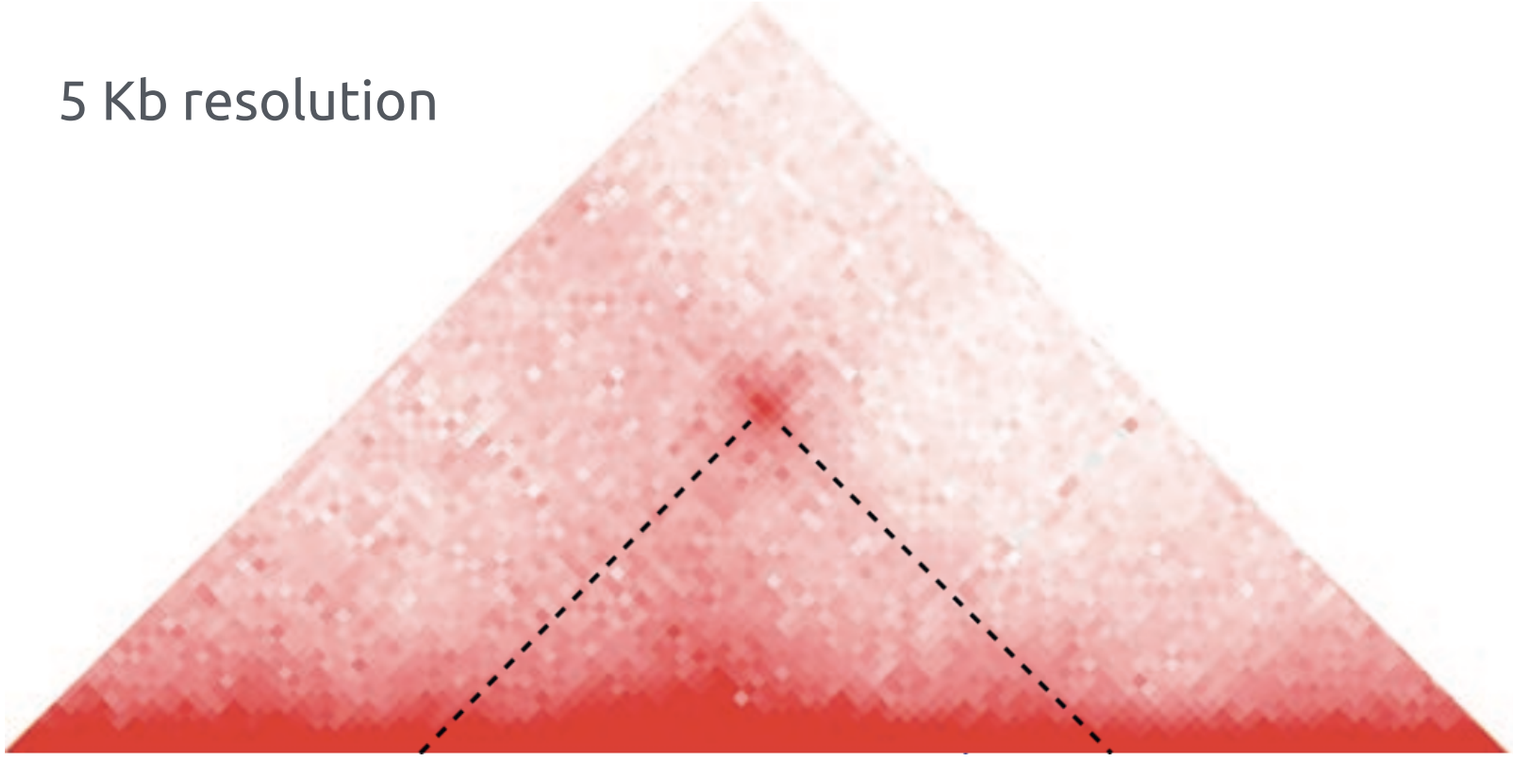


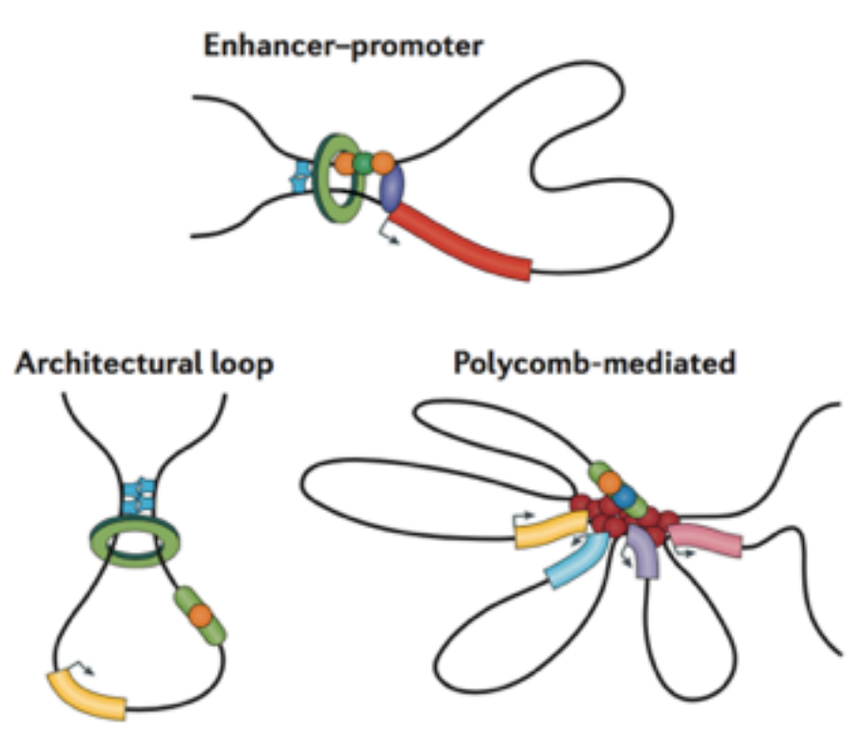


Bonev et al. *Nature Reviews* 2016

1.3 Chromatin loops

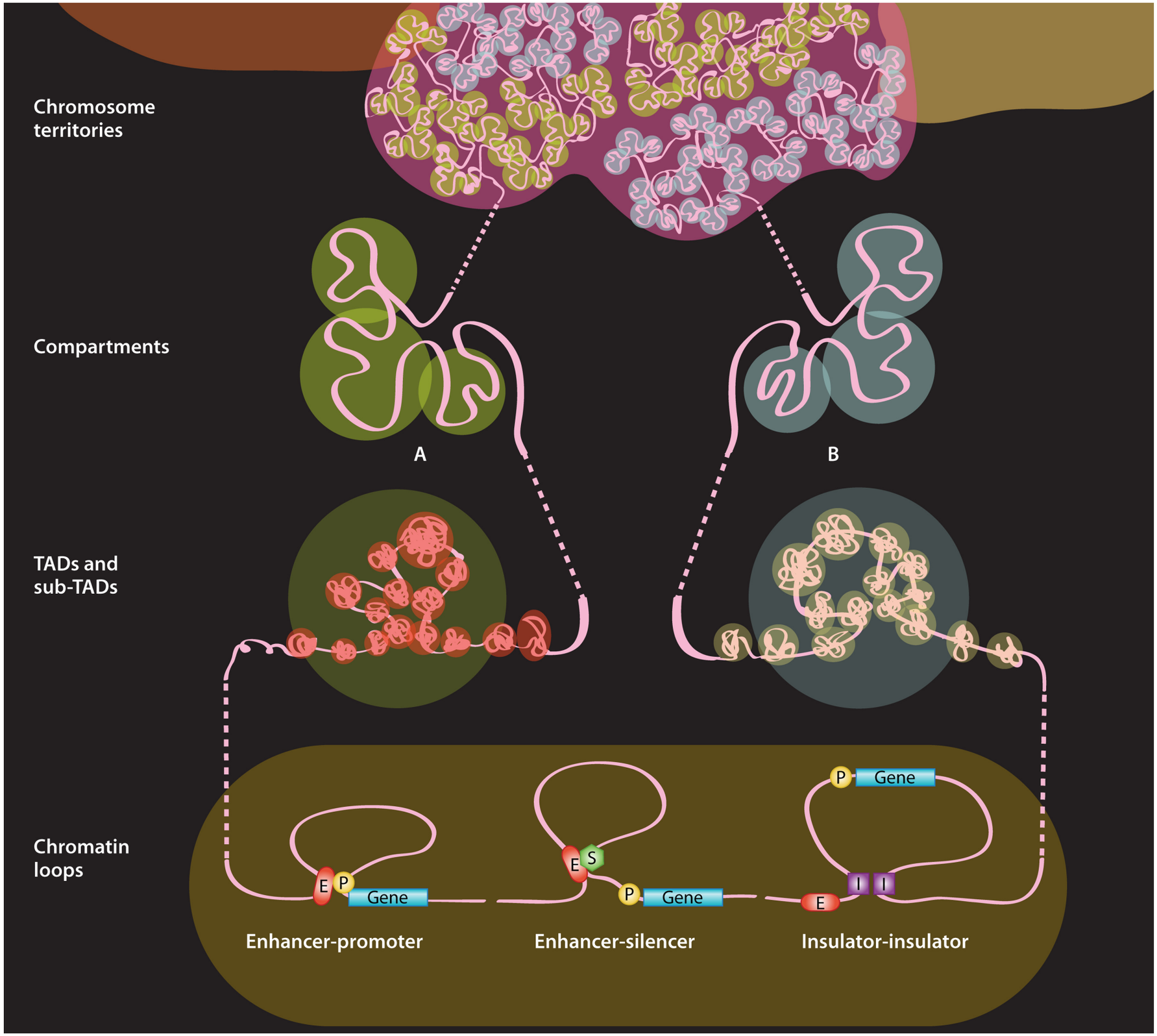
- Cis-regulatory elements of vertebrates, such as enhancers, are separated by relatively long distances and can be brought into close spatial proximity with its target through the formation of chromatin loops.
- There are also other cases of loops (e.g. between co-regulated genes, between Polycomb-repressed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





Bonev et al. *Nature Reviews* 2016

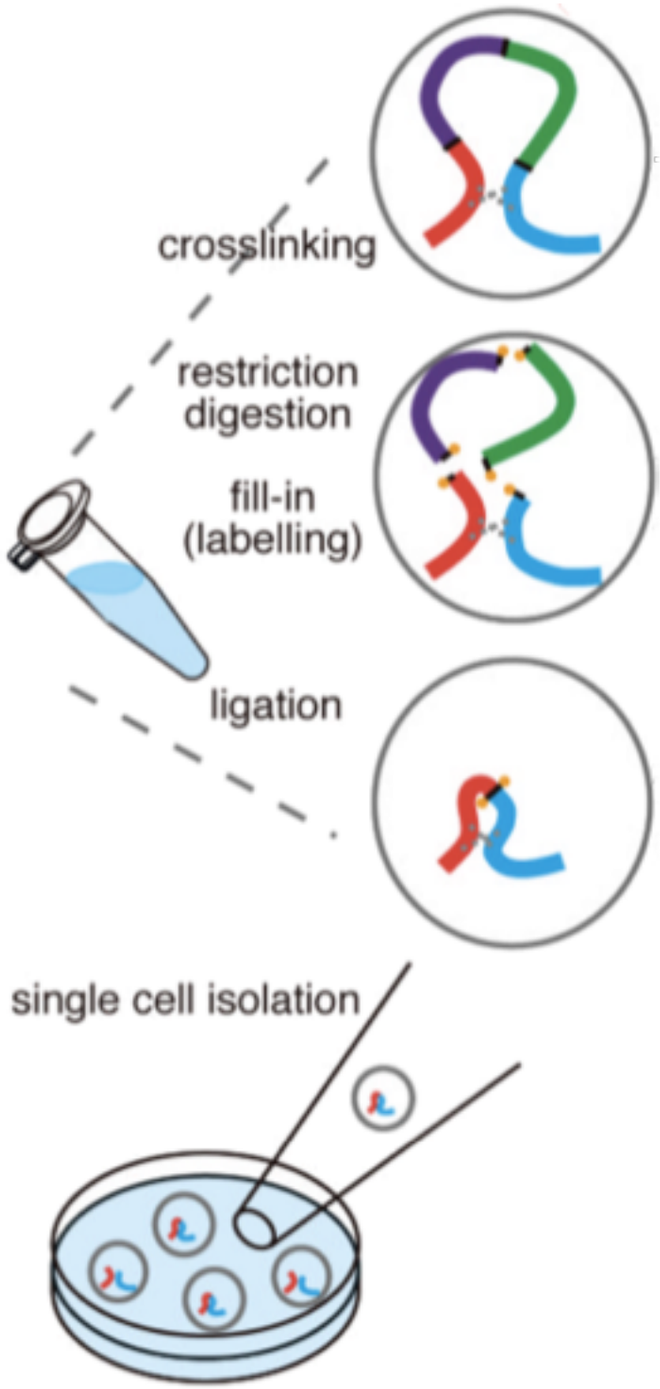
"The Zoo" of chromatin features



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

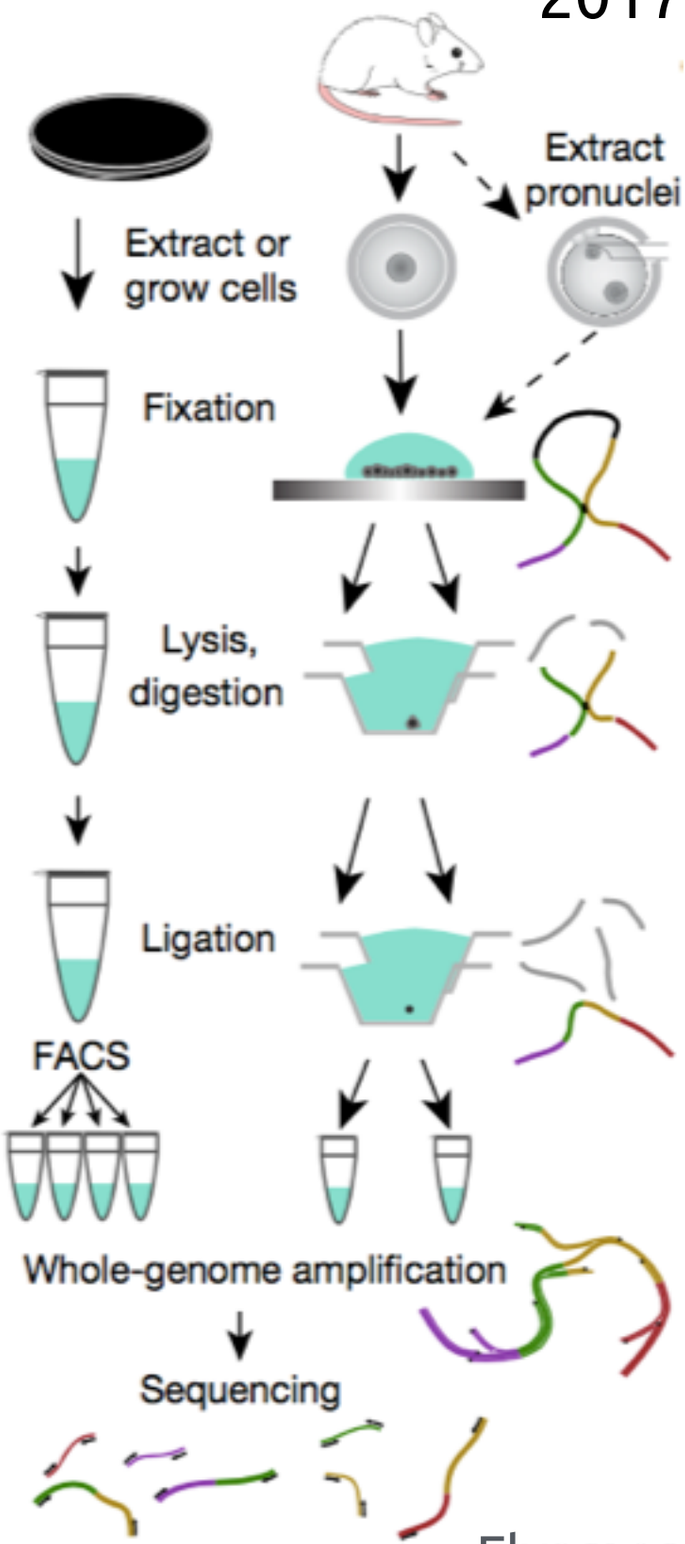
Single-cell Hi-C

2013 method:



Nagano et al. *Nature* 2013

2017 method:

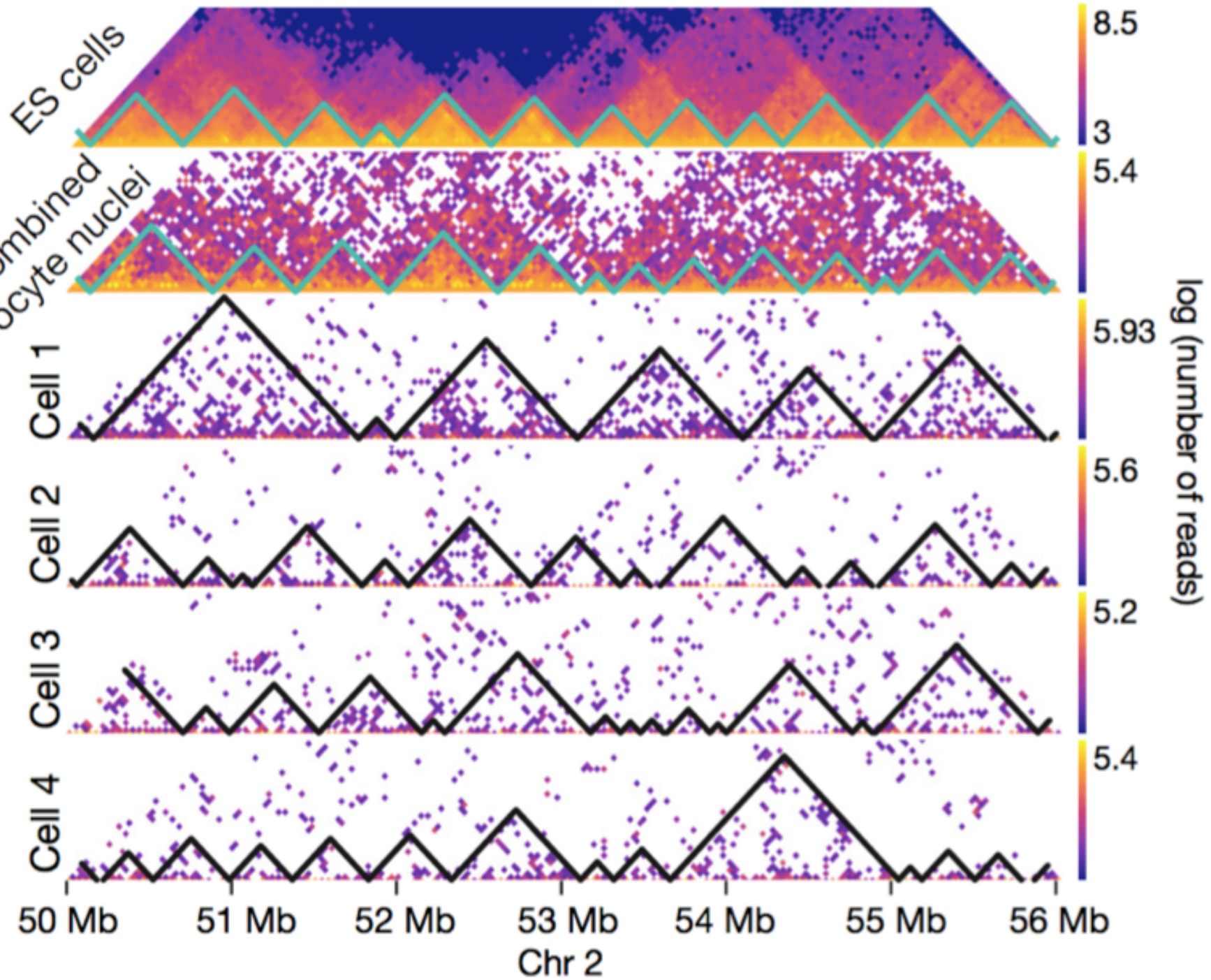


Flyamer et al. *Nature* 2017

1.4 Single-cell Hi-C

Ensemble (bulk) Hi-C:

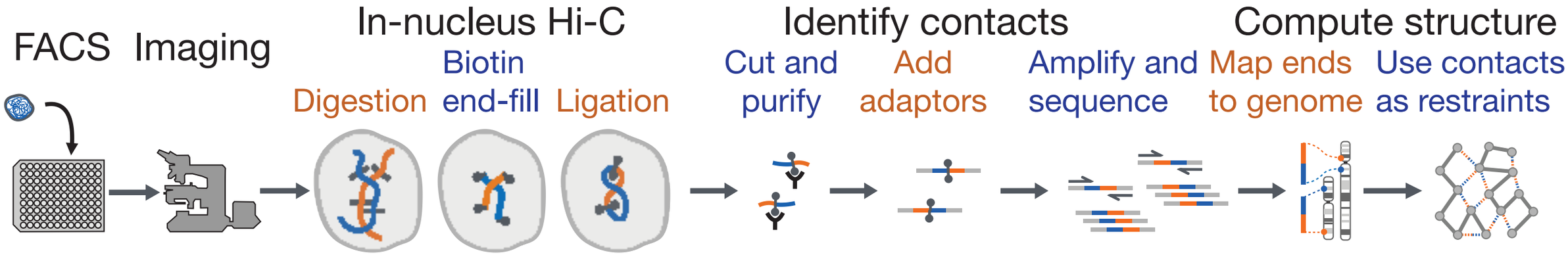
Single-cell Hi-C (2017): (multiple)



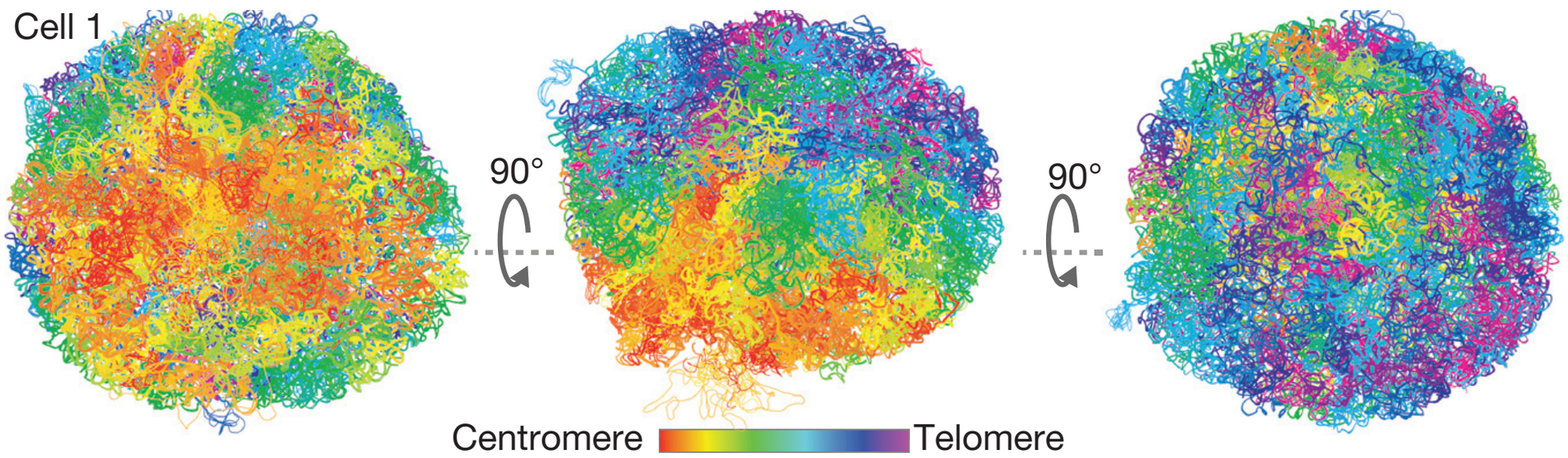
Flyamer et al. *Nature* 2017

1.4 Single-cell Hi-C

Another method from 2017:



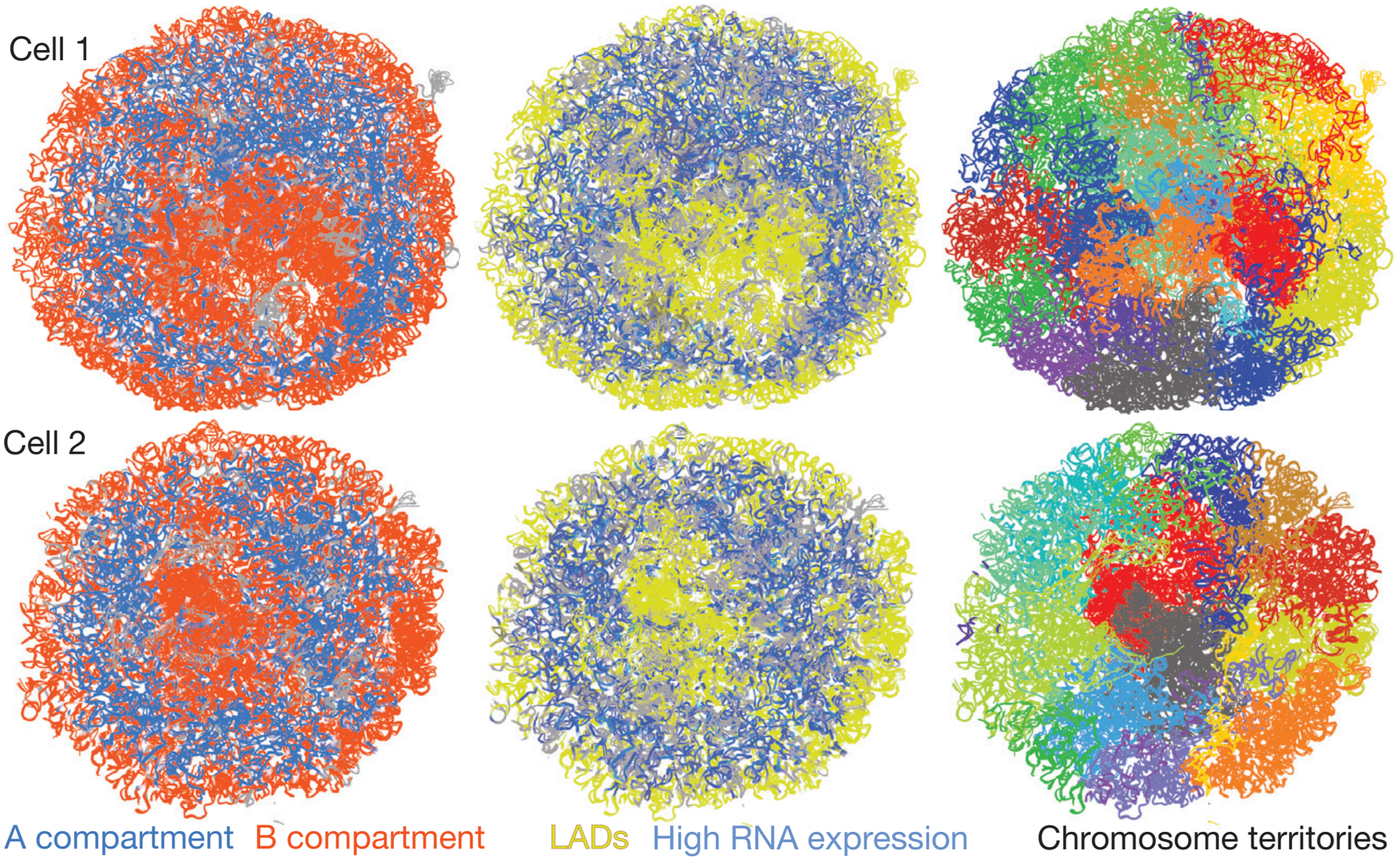
Structure modelling results:



Stevens et al. *Nature* 2017

Single-cell Hi-C

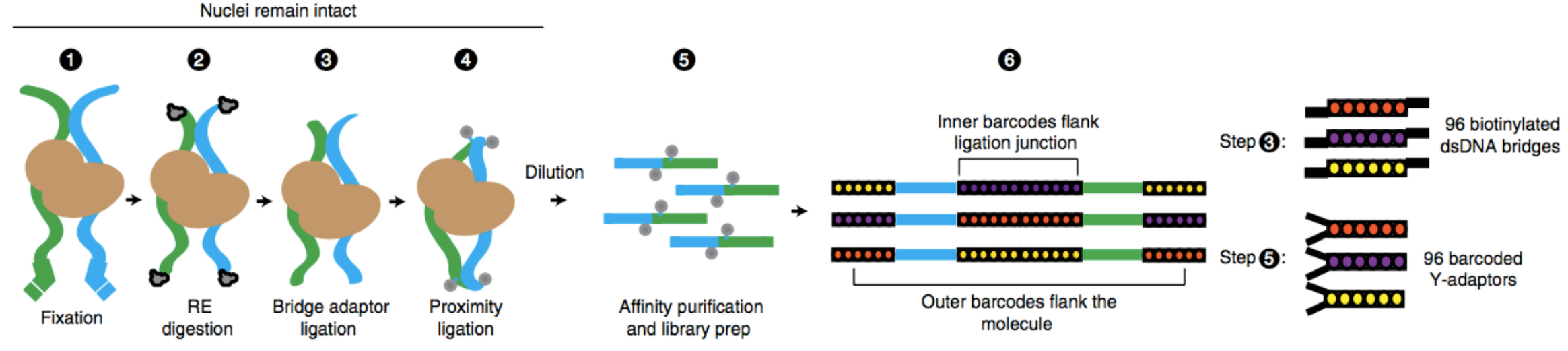
Data modelling based on single-cell can be very powerful:

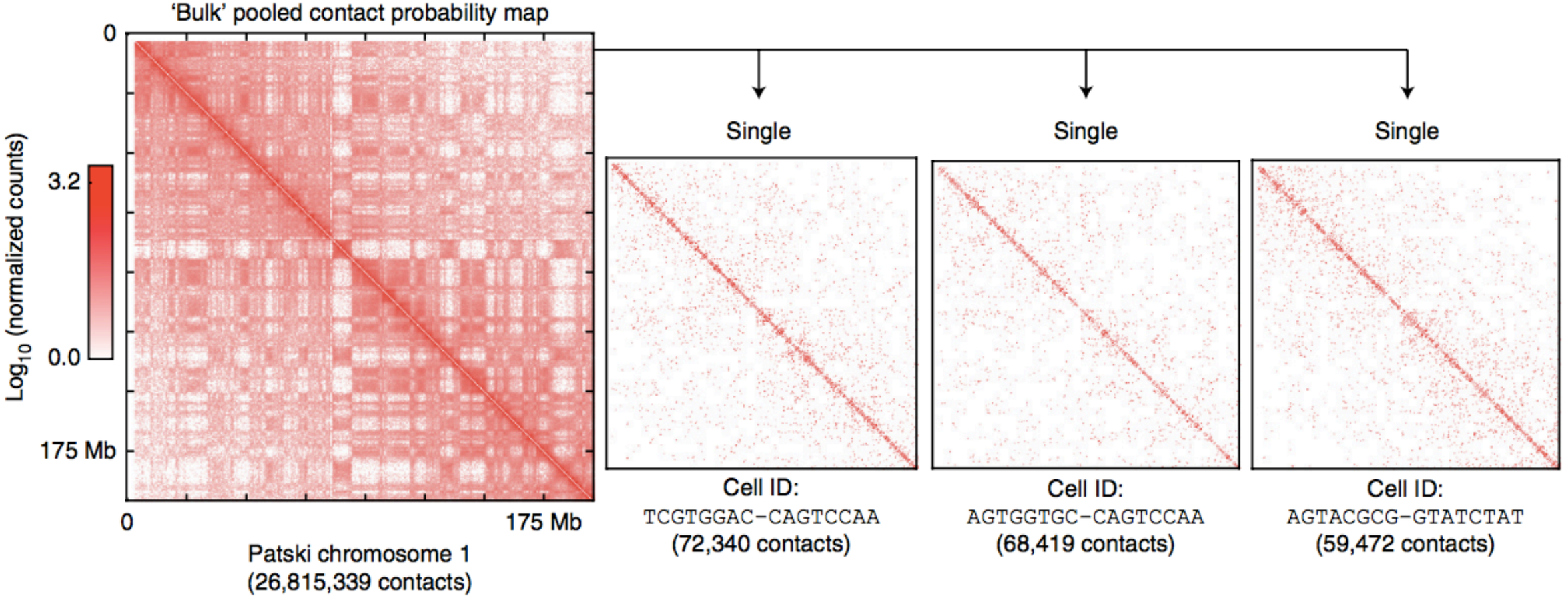


Stevens et al. *Nature* 2017

1.4 Single-cell Hi-C

One more method from 2017: single-cell combinatorial indexed Hi-C (sciHi-C)





Ramani et al. *Nature methods* 2017

How many contacts do we recover?

	Total number of reads per experiment	Number of cells	Number of contacts per cell	Recovery of the total possible ligation junction
Nagano 2013	5 - 15 mln	10 cells with > 1000 contacts	11,000-30,000	< 2%
Stevens, 2017	1.5 - 4.8 mln	8	37,000-122,000	1.2-4.1%
Flyamer, 2017	~15-83 mln	36 cells with >30,000 contacts 219 cells with > 1000 contacts	up to 1,906,000	~ 10%
Ramani, 2017	20-500 mln	10,696 cells with > 1,000 contacts	59,000-72,000	-

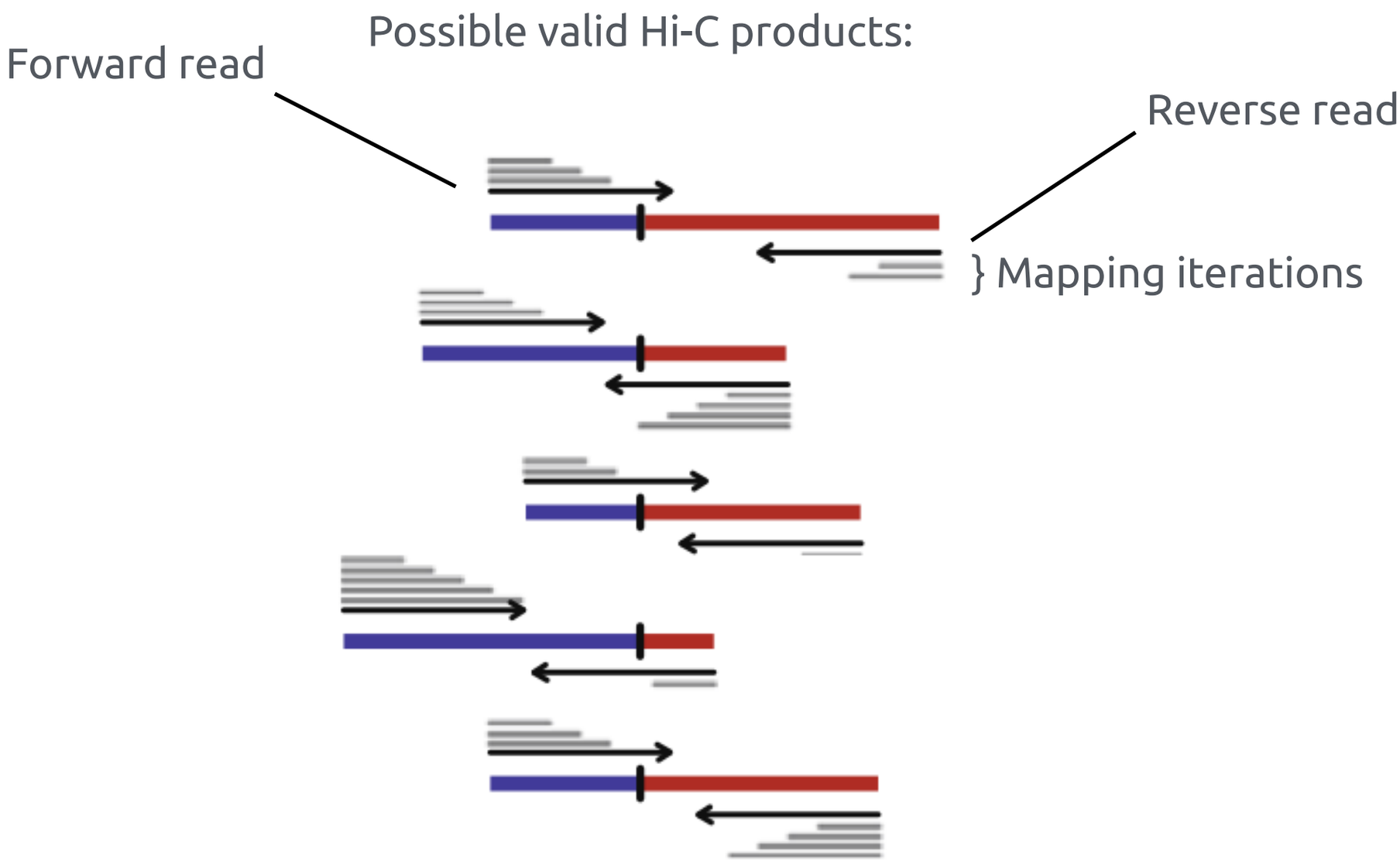
2. From theory to practice: Hi-C processing workflow

2. 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.)

2.1 Reads mapping

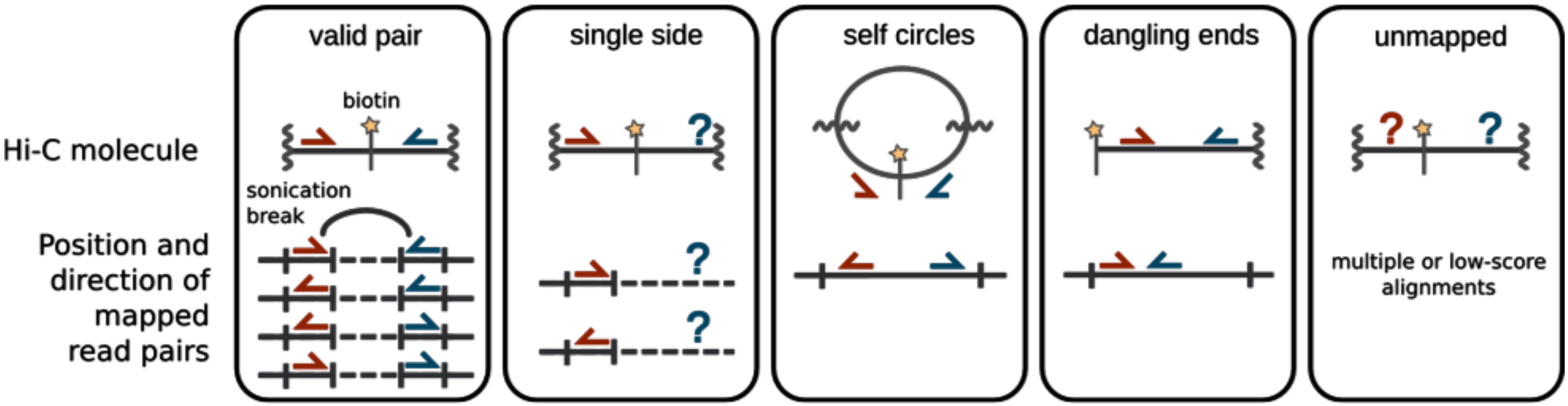
- Iterative or split reads mapping is required.



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

2.2 Filtering at the level of fragments

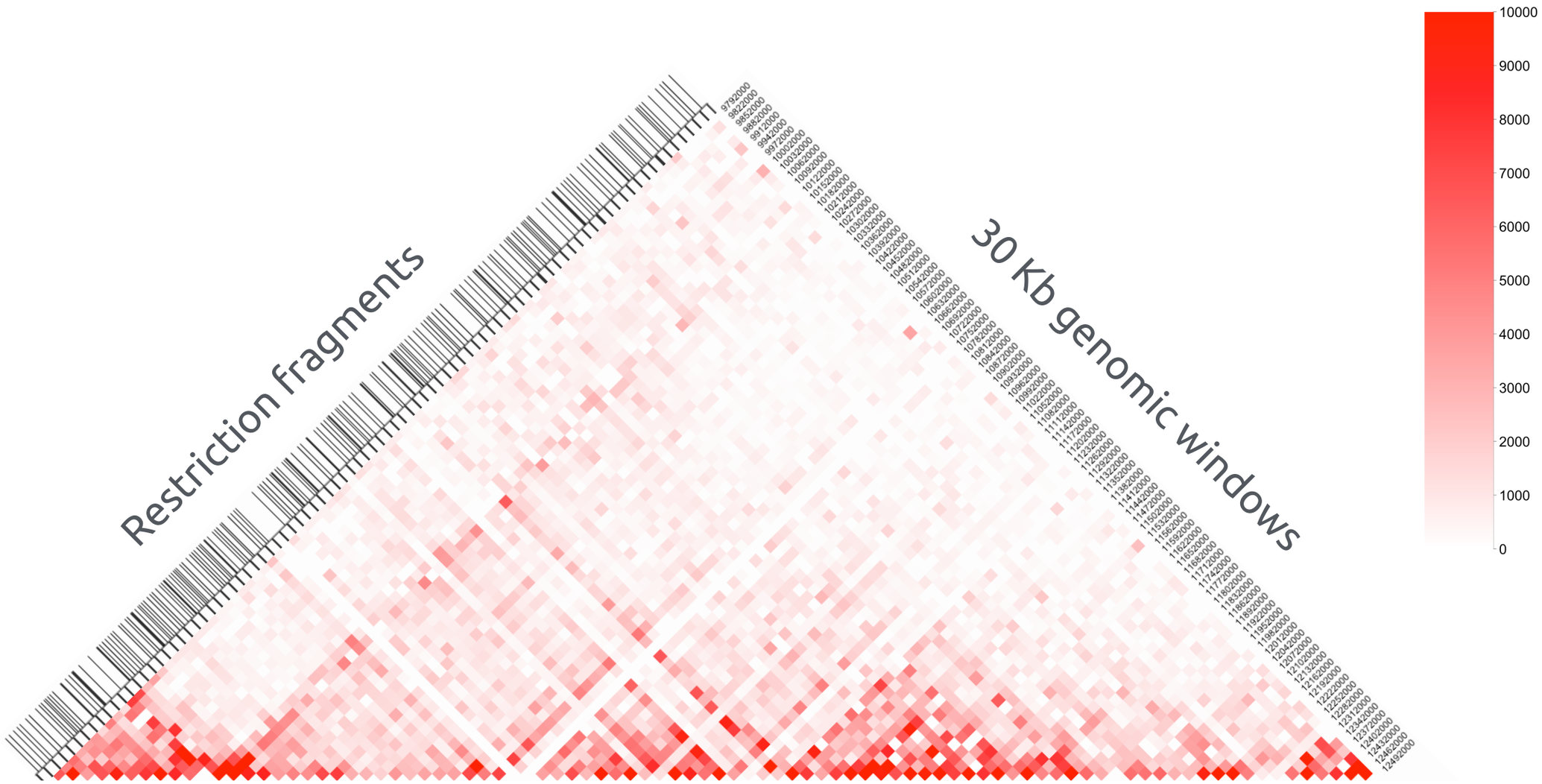
- Possible Hi-C mapping results:



Imakaev et al. *Nature Methods* 2012

2.2 Binning

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



Binned interactions heatmap

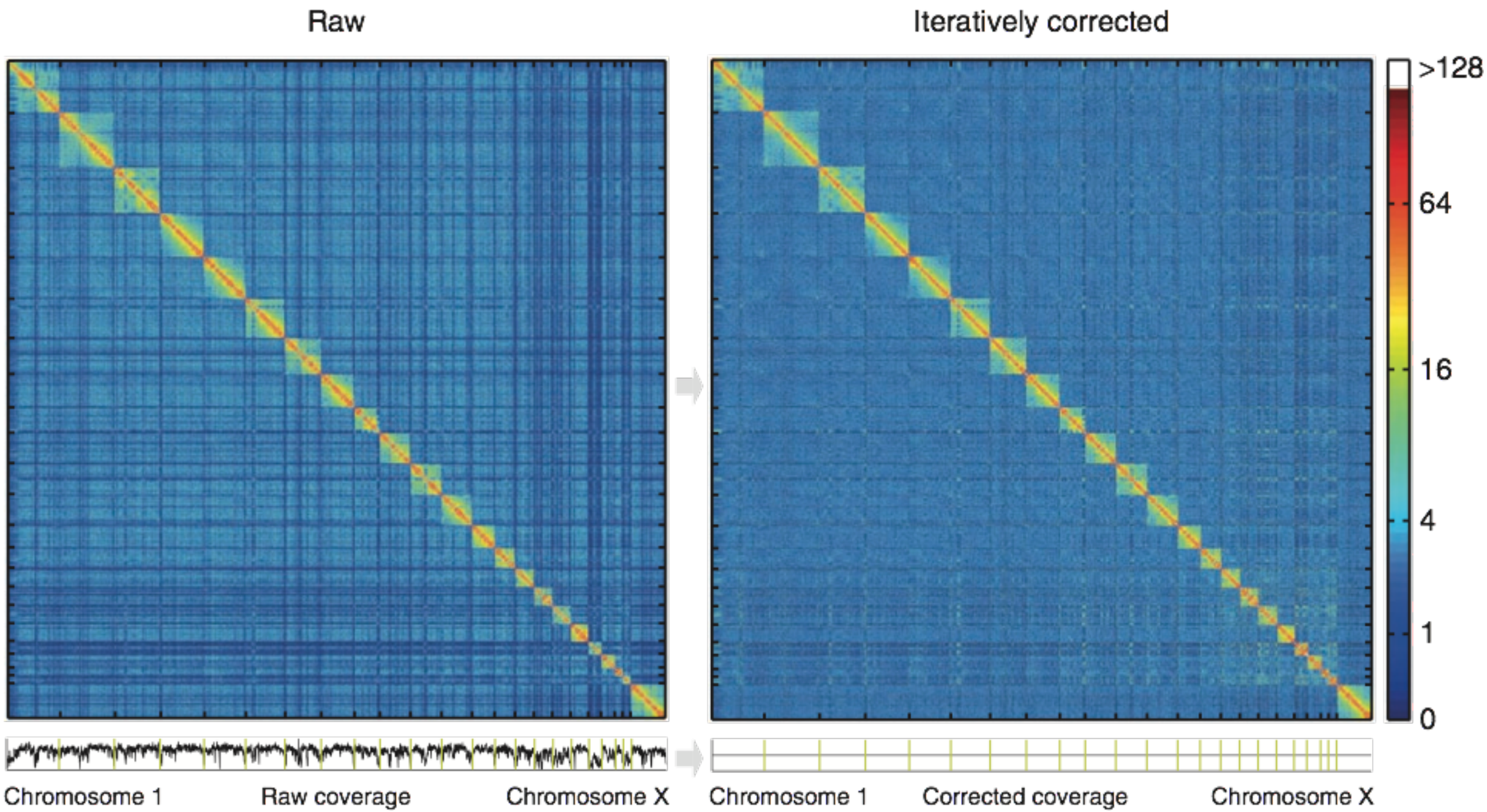
2.3 Matrix balancing

- 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

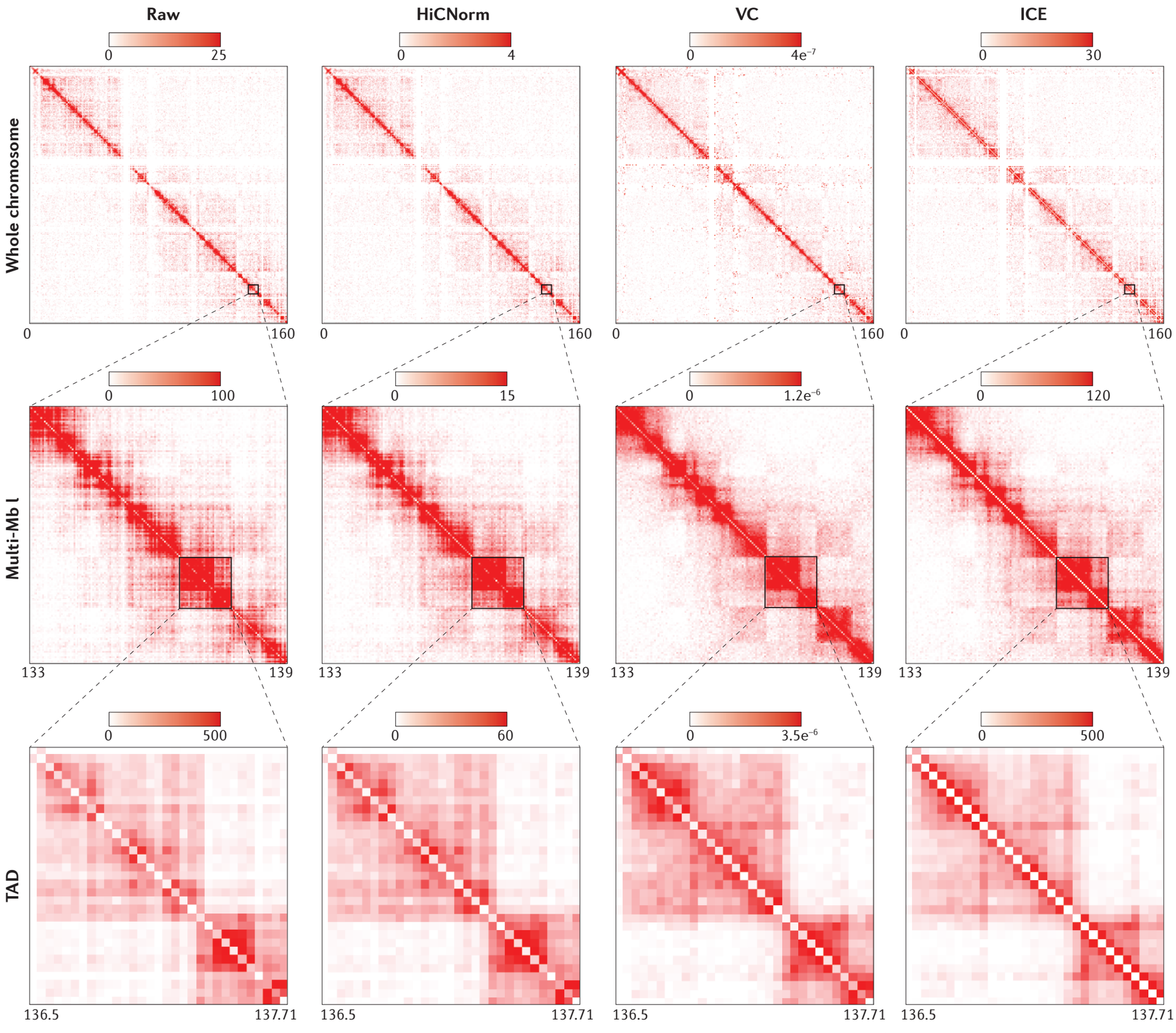
Adopted from Schmitt et al. *Nature Reviews* 2016

2.3 Iterative correction



Imakaev et al. *Nature Methods* 2012

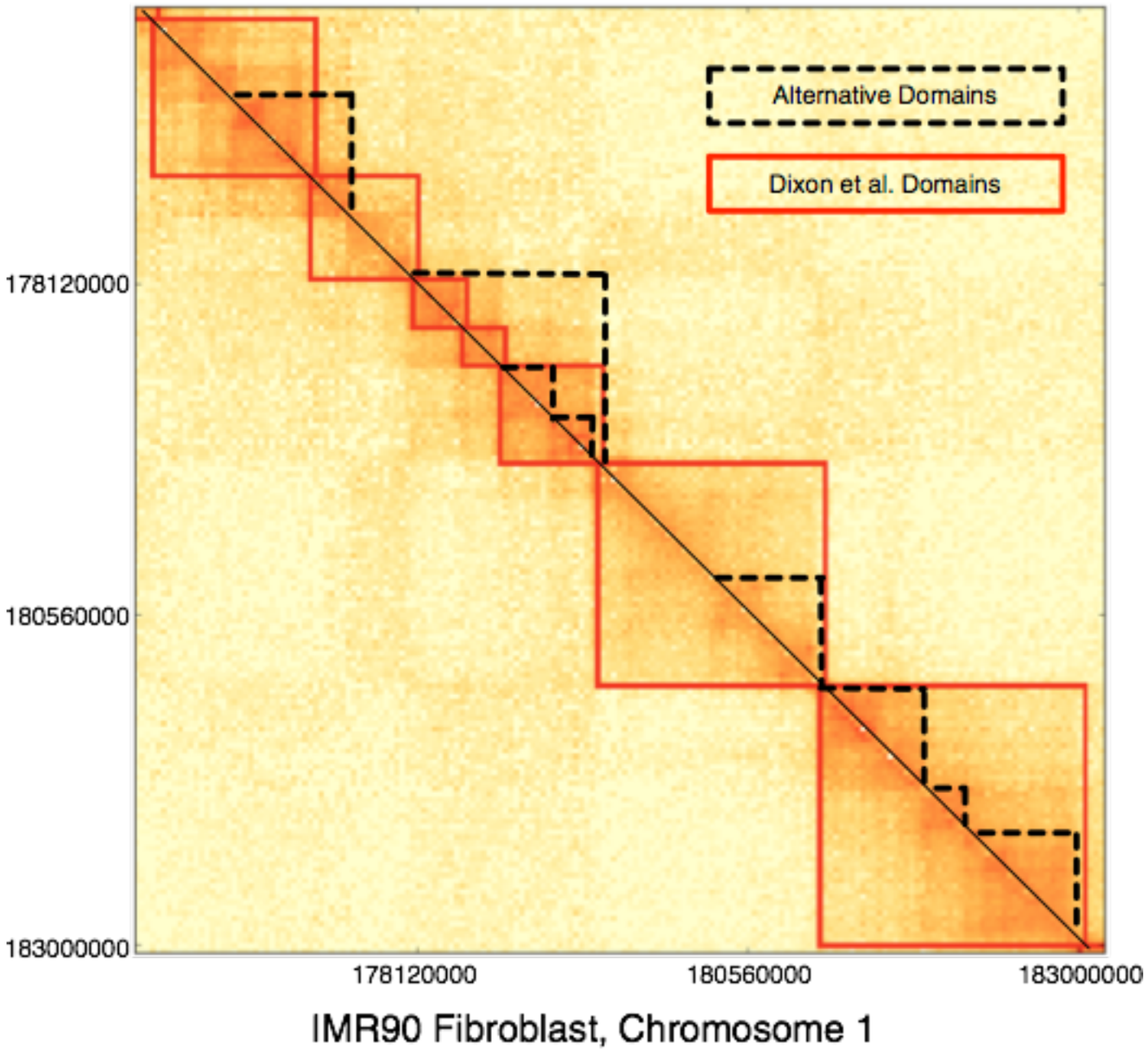
2.3 Matrix balancing

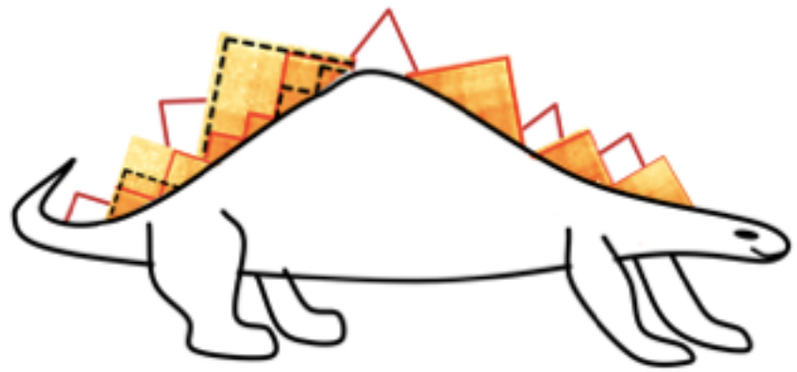


Schmitt et al. *Nature Reviews* 2016

2.4 TADs calling

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



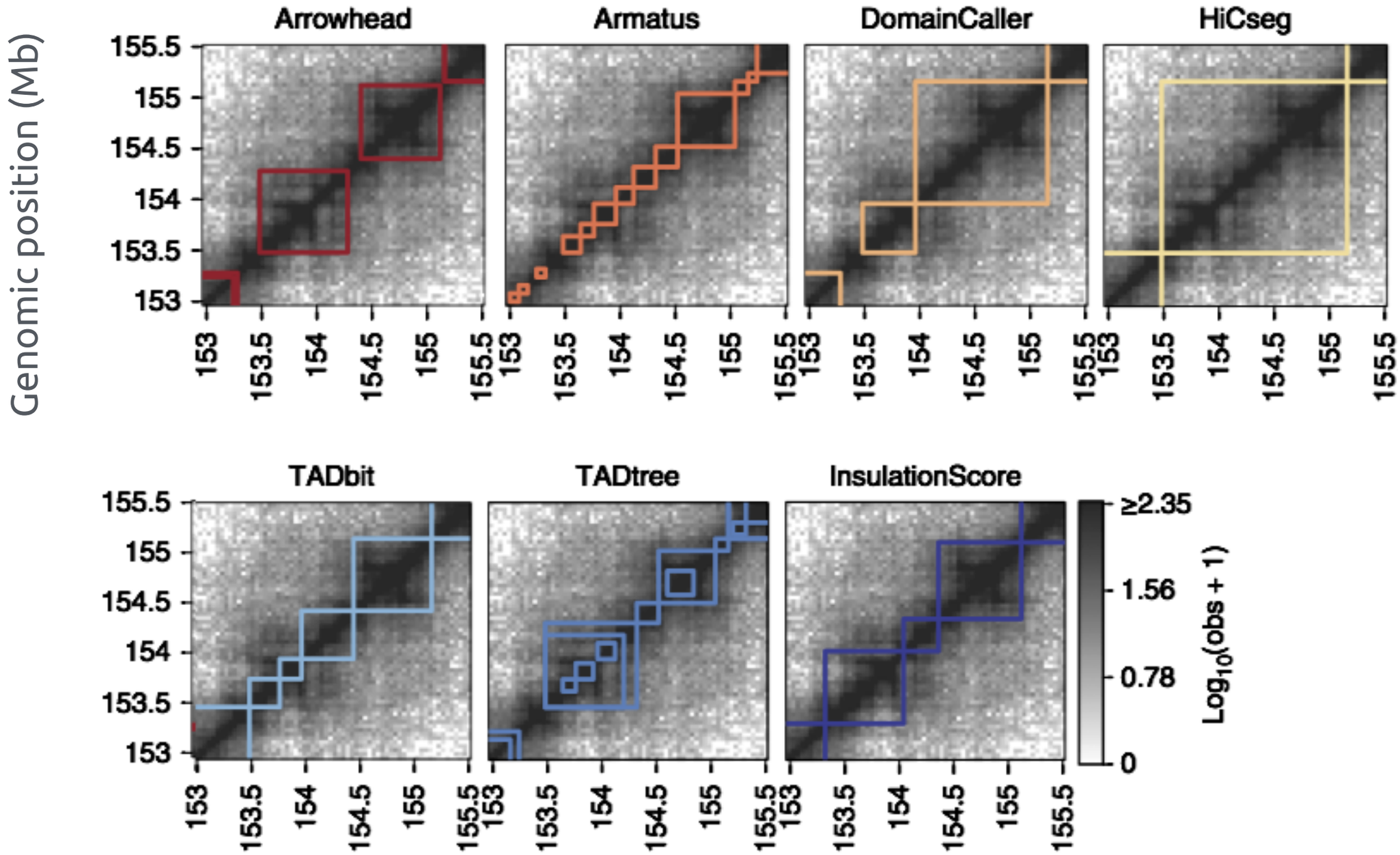


For example, Armatus algorithm is based on dynamic programming and has variable parameter, gamma.

Filippova et al. *Algorithms for Molecular Biology* 2014

2.4 TADs calling

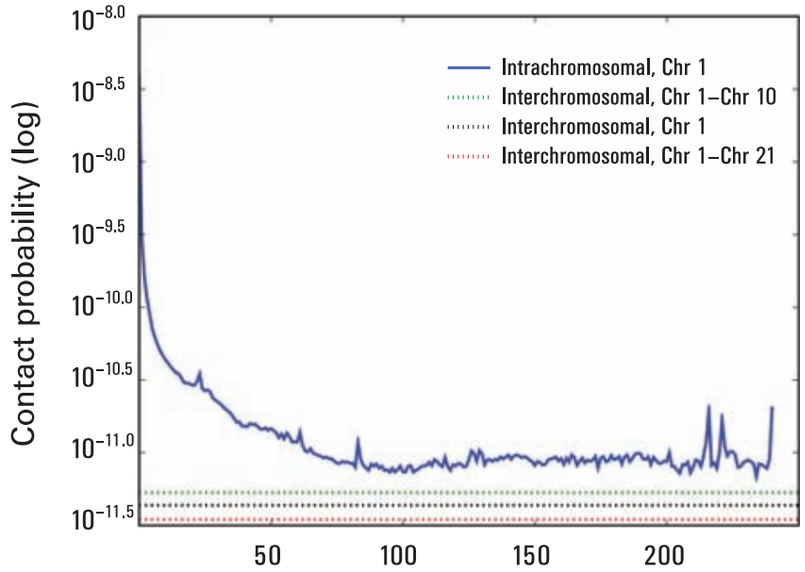
- A recent comparison of multiple TADs calling tools:

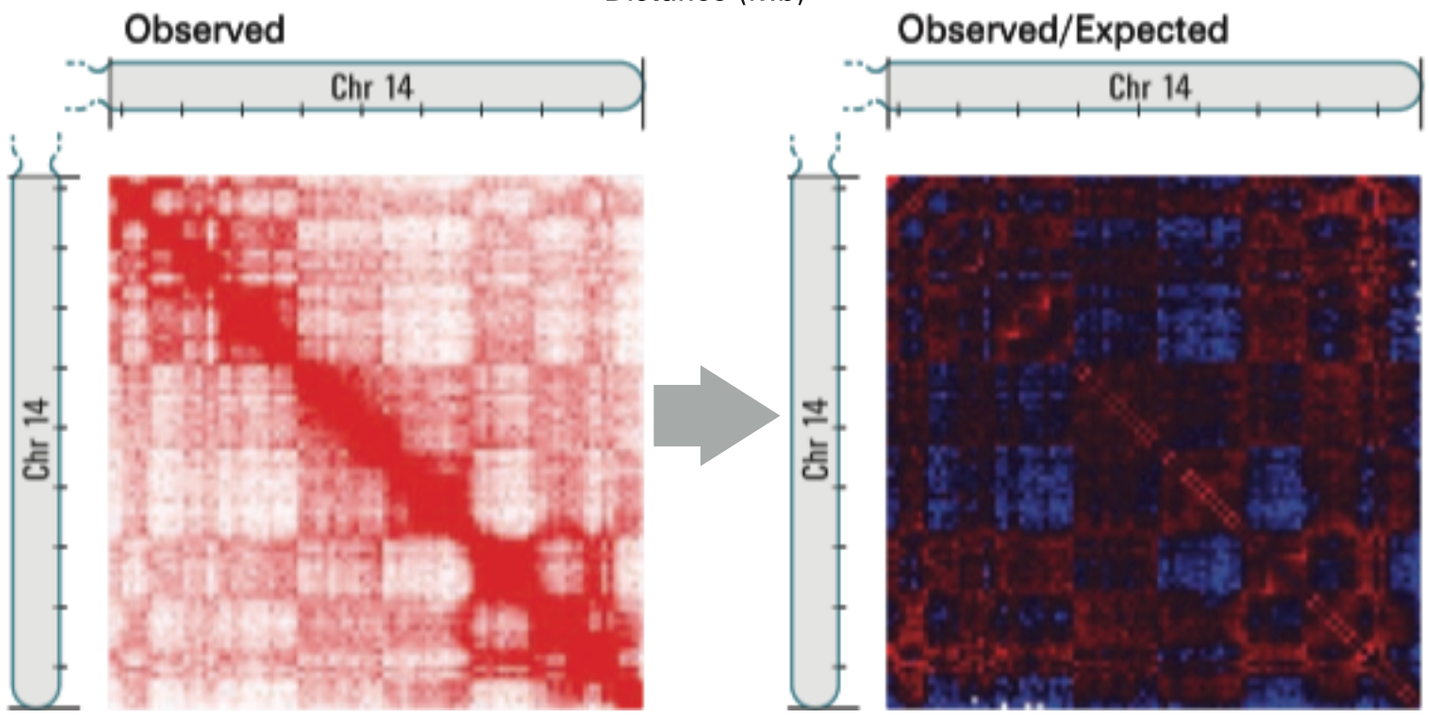


Forcato et al. *Nature Methods* 2017

2.4 Compartments calling

- Method from Lieberman-Aiden, 2009:
  1. Normalization of interaction matrix by expected interactions:



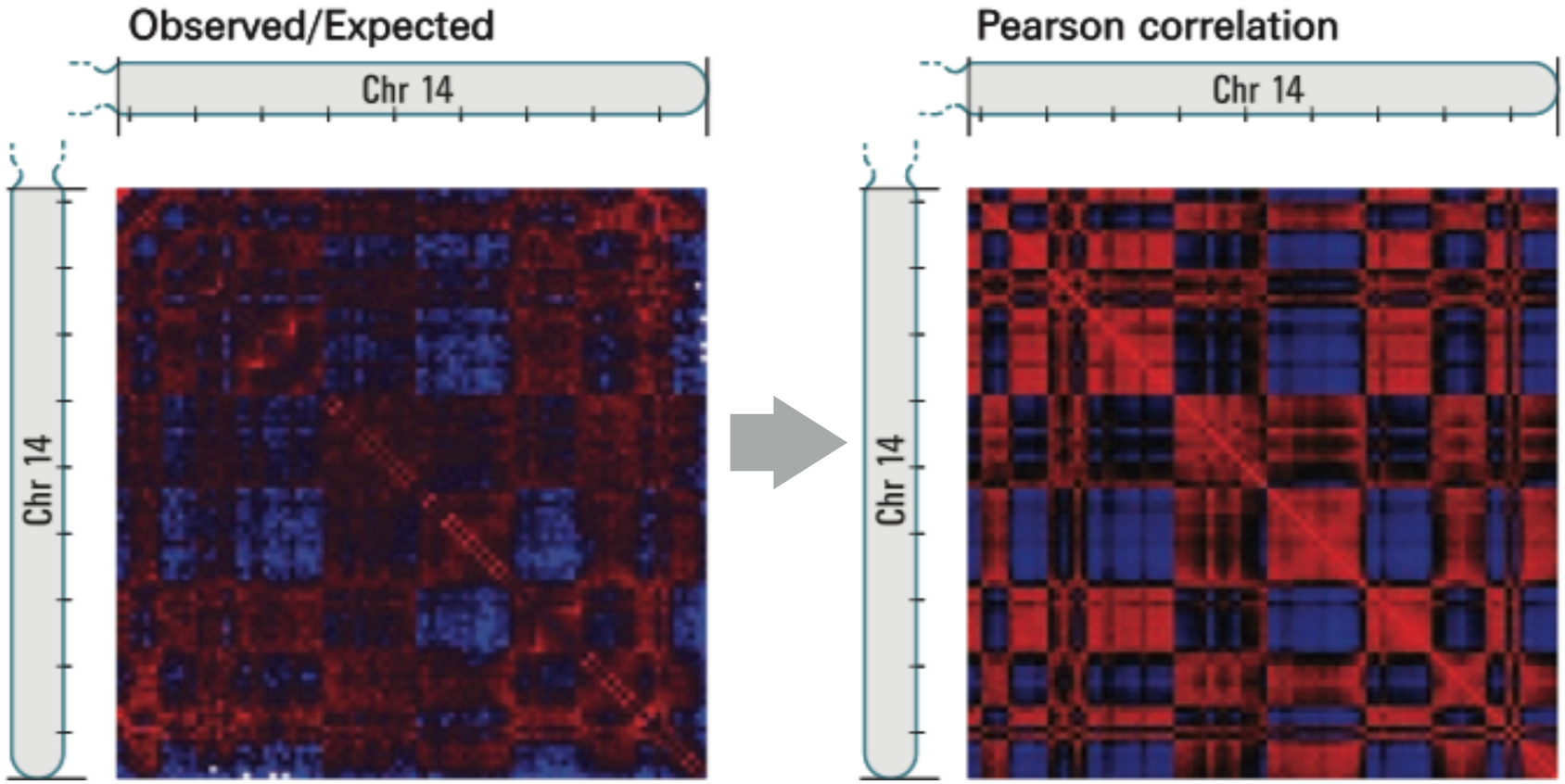


Lieberman-Aiden et al. *Nature* 2009

2.4 Compartments calling

- Method from 2009:

  (2) Calculation of Pearson correlation

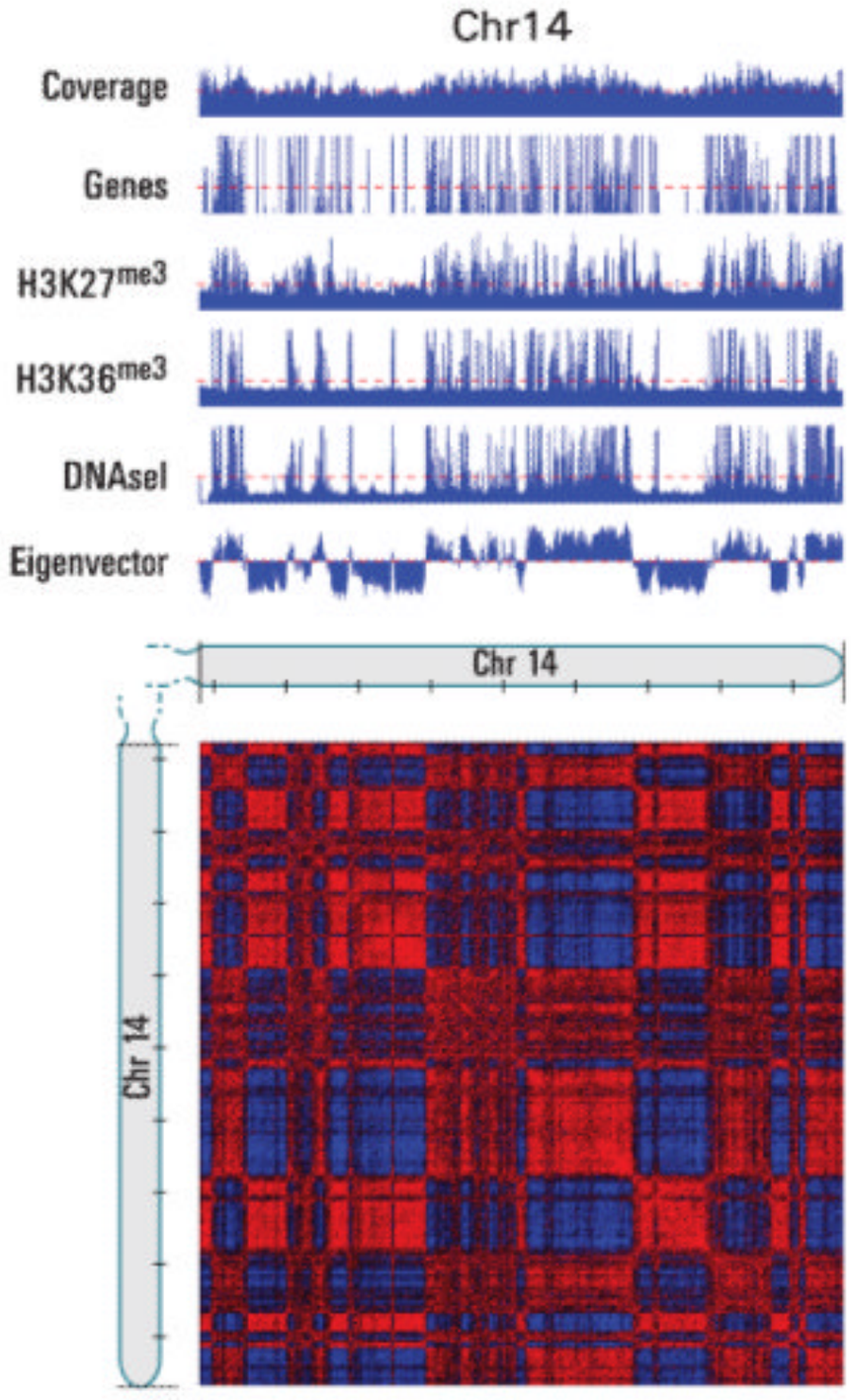


Lieberman-Aiden et al. *Nature* 2009

2.4 Compartments calling

- Eigenvector decomposition:

  (3) Eigenvector expansion (PCA, principal component analysis)



Lieberman-Aiden et al. *Nature* 2009

2.5 Some tools for basic Hi-C data processing

	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

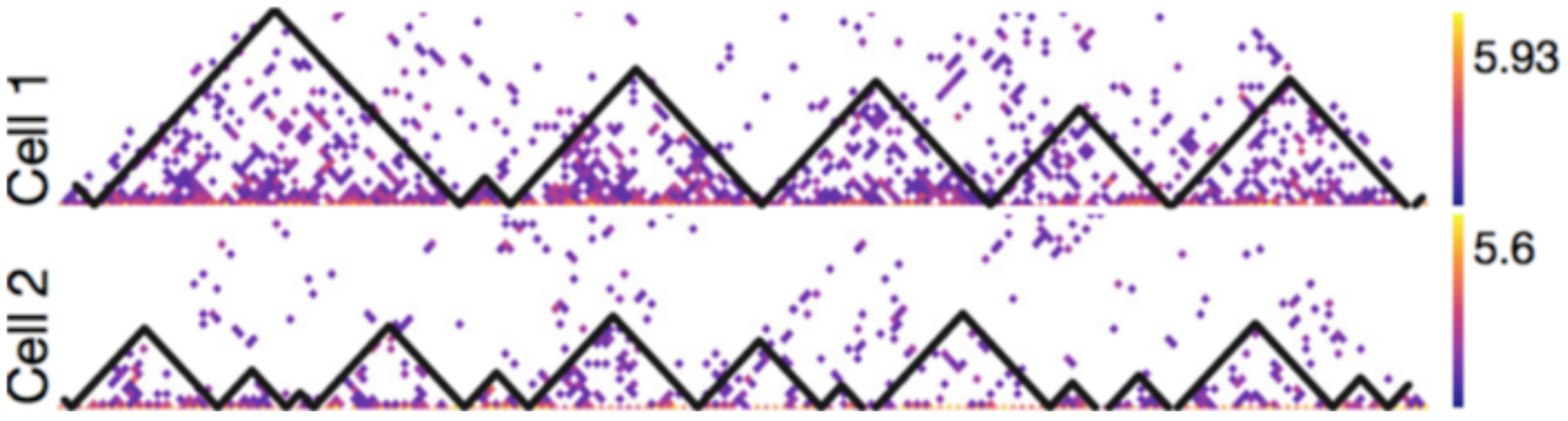
Forcato et al. *Nature Methods* 2017

2.5 Single-cell data analysis

- Generally the same processing workflow, except:
  - Stringent amplification duplicates filtering.

Example elimination of counting the same ligation junction many times (Flyamer et al. *Nature* 2017): if two reads map to the same strand, and each side of the read is within 500 bp of any side of the other read, only one copy of the read is retained.

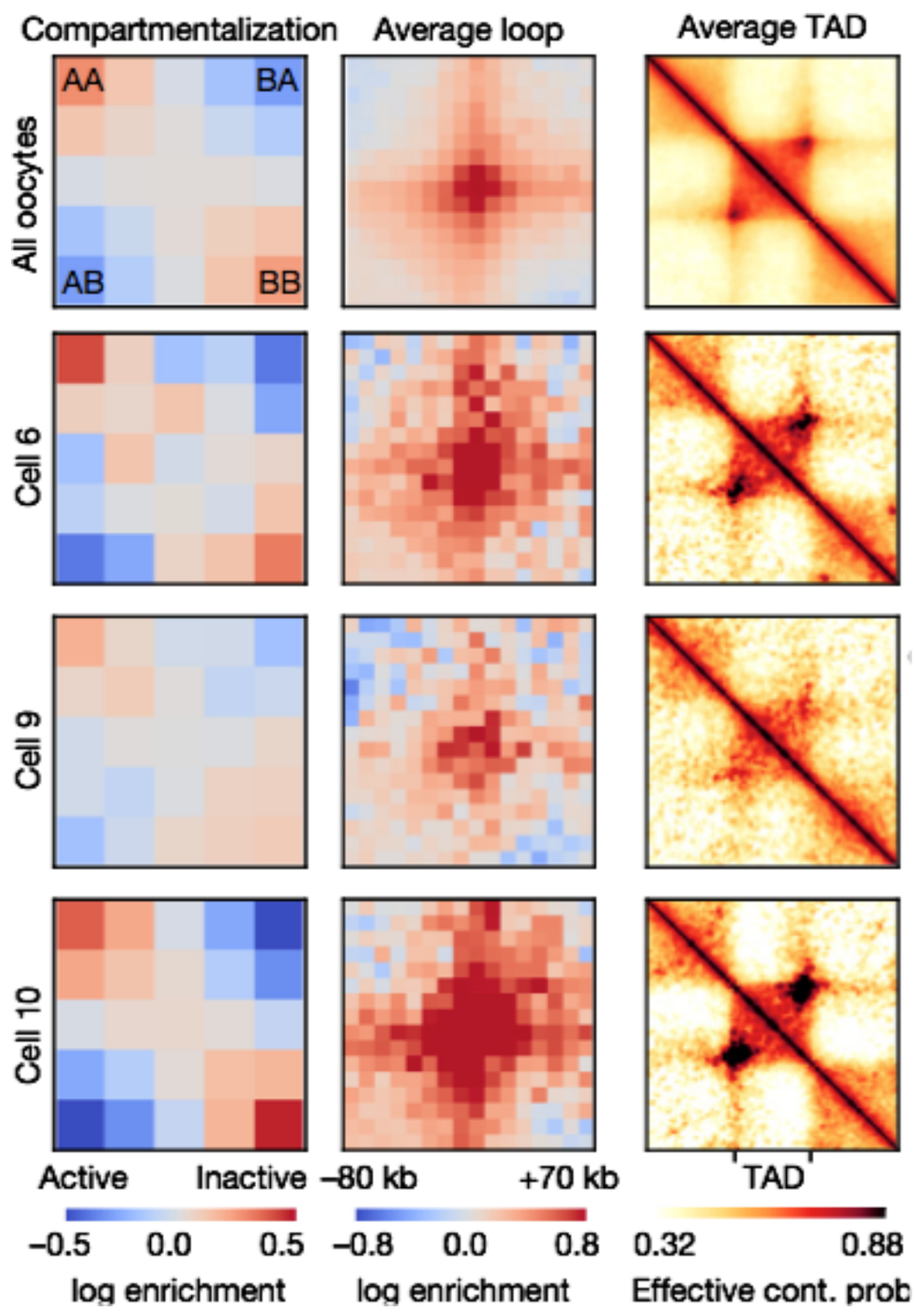
  - Iterative correction and normalization are not applicable due to data sparsity.



Flyamer et al. *Nature* 2017

Single-cell data analysis

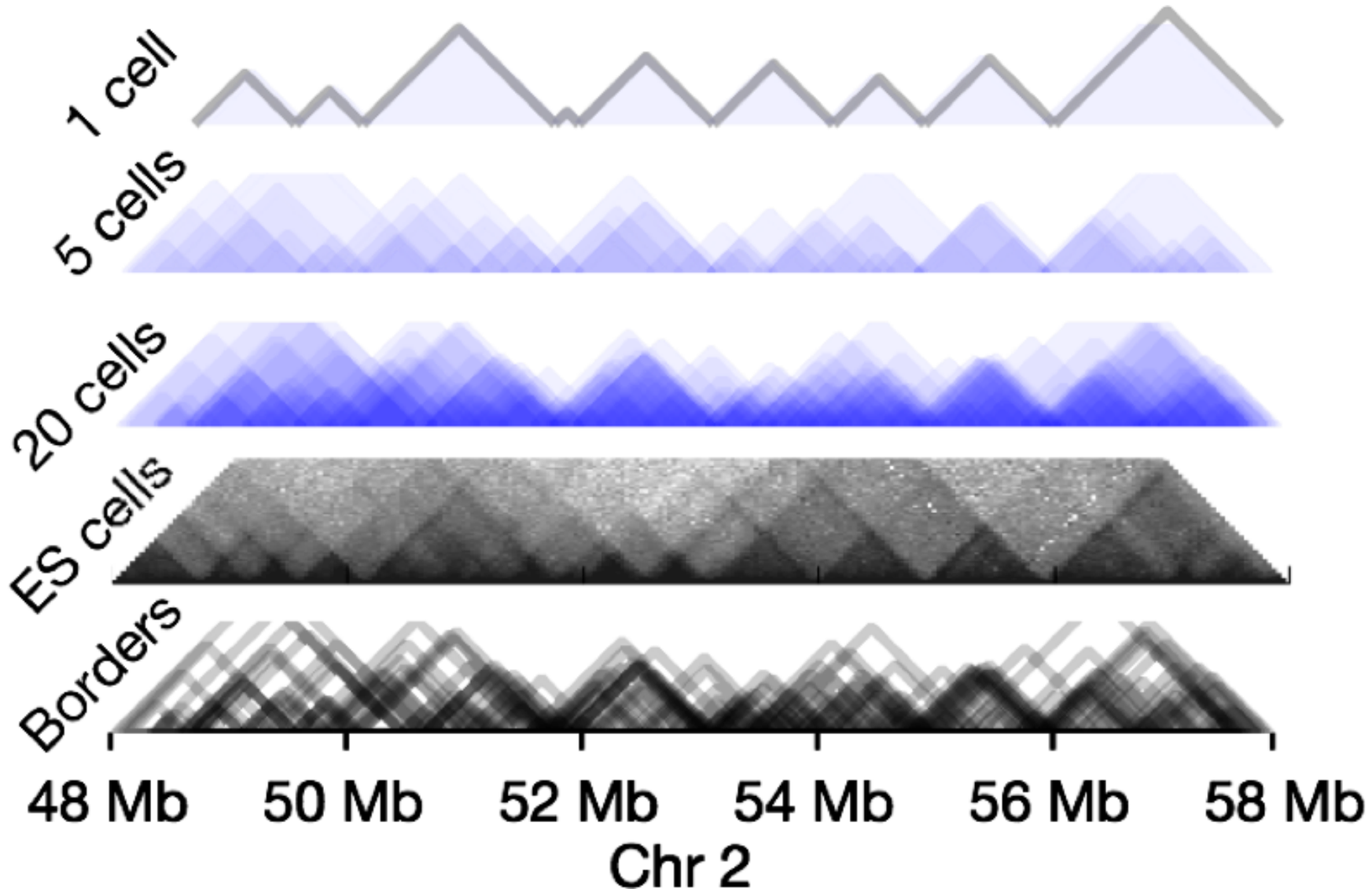
- Indirect detection of compartments, TADs and loops due to data sparsity:



Flyamer et al. *Nature* 2017

2.5 Single-cell data analysis

- Still, TAD-like structures ("contact domains") could be found directly:



Flyamer et al. *Nature* 2017

3. From theory to practice: workshop overview

Workshop overview

- Single-cell and bulk Hi-C raw datasets from Flyamer et al. *Nature* 2017 (GEO: GSE80006)
- Data processing with hiclib (one of the best Hi-C data practices since 2012):
  - Iterative mapping of reads with bowtie2
  - Data filtering
  - Binning
  - Data visualization
  - TADs calling
  - Comparison of single-cell and bulk Hi-C experiments
  - Compartments detection
  - …
- Powered by: