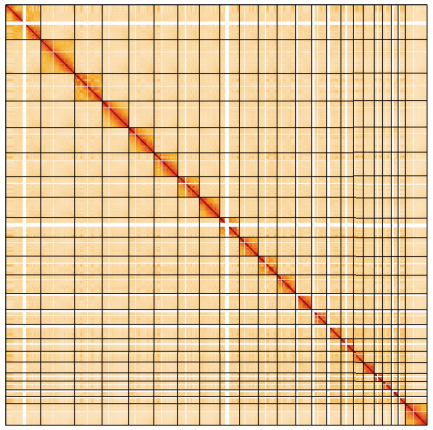
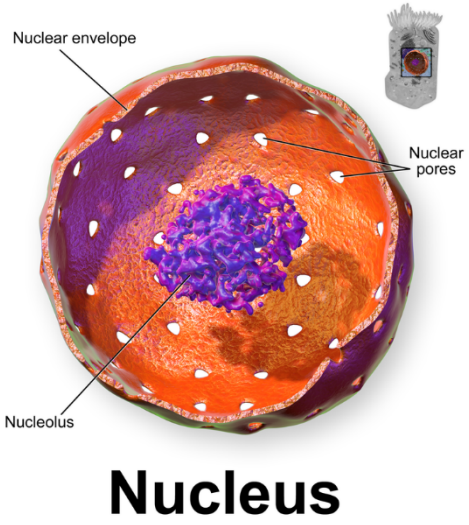
From contact matrix to biology

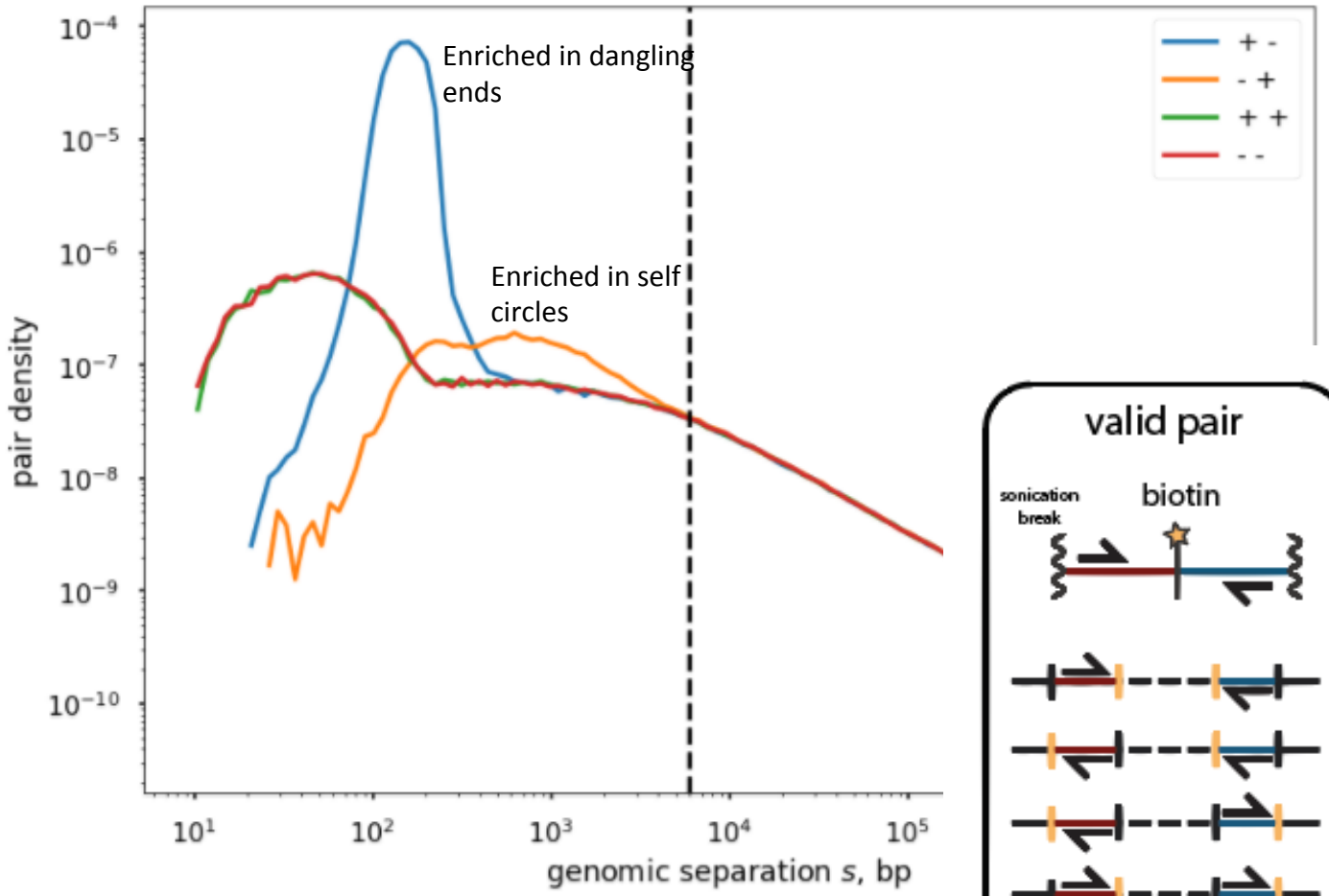
Nezar Abdennur

Mirny Lab, MIT



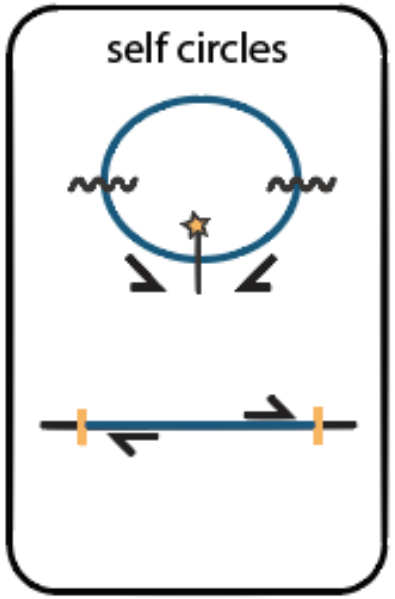


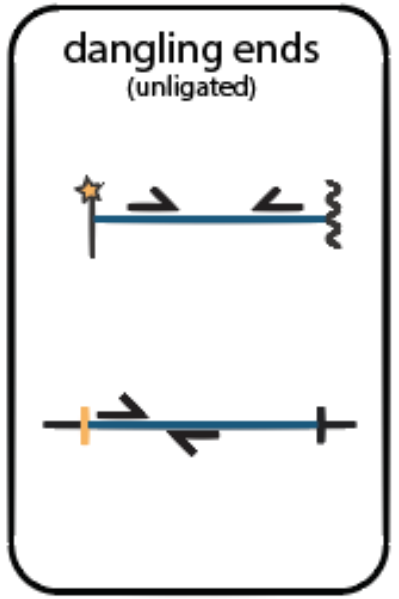
Genomic proximity and informative ligation products



align

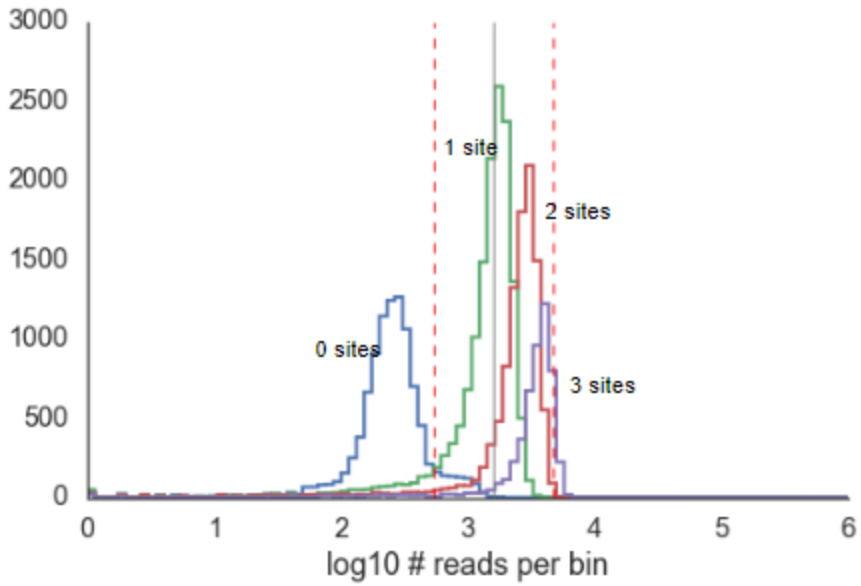
classify



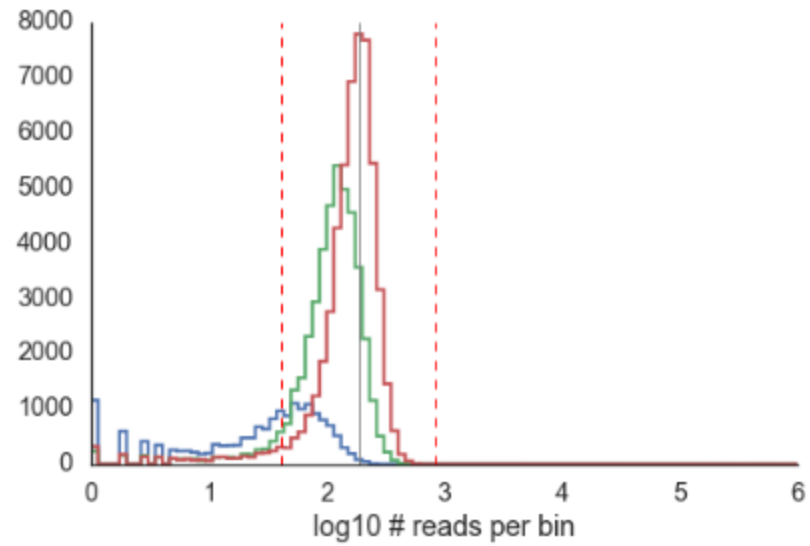


The modes in log-coverage correspond to the number of restriction sites in a bin

MADMax for 5kb HindIII split by n rsites



MADMax for 1kb MboI split by n rsites



align

classify

sort + dedup

filter + bin

balance

Cooler

build passing | docs latest | launch binder | chat on gitter

A cool place to store your Hi-C

align

classify

sort + dedup

filter + bin

balance

pairs

	read_id	chrom1	pos1	chrom2	pos2	strand1	strand2	pair_type
0	.	chr1	3000032	chr1	66691730	-	+	LL
1	.	chr1	3000035	chr1	4026372	-	+	LL
2	.	chr1	3000302	chr1	4430326	-	-	LL
3	.	chr1	3000356	chr1	3116276	-	-	LL
4	.	chr1	3000845	chr1	3001031	+	-	LL

aggregate

	chrom1	start1	end1	weight1	chrom2	start2	end2	weight2	count
100	chr1	10000	15000	NaN	chr10	102850000	102855000	1.503567	1
101	chr1	10000	15000	NaN	chr10	103105000	103110000	0.924921	1
102	chr1	10000	15000	NaN	chr10	134945000	134950000	1.480836	1
103	chr1	10000	15000	NaN	chr11	285000	290000	1.455694	1
104	chr1	10000	15000	NaN	chr11	420000	425000	1.107395	1
105	chr1	10000	15000	NaN	chr11	535000	540000	0.819330	1
106	chr1	10000	15000	NaN	chr11	715000	720000	0.911051	1
107	chr1	10000	15000	NaN	chr11	810000	815000	1.202326	1
108	chr1	10000	15000	NaN	chr11	935000	940000	0.889250	1
109	chr1	10000	15000	NaN	chr11	940000	945000	1.017561	1

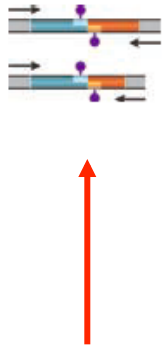
“join”
(annotate)

- Cooler
  - Sparse, binary, columnar, compressed storage schema based on HDF5 (.cool)
  - Python support library and CLI provides tools to create, aggregate, query, and normalize coolers
  - Provides sparse, parallel, out-of-core matrix balancing
  - Aims to integrate smoothly with the Python data ecosystem: numpy, pandas, dask, jupyter

Bias correction (normalization)

Digestion efficiency as a function of sequence composition (and DNA compaction)

Ligation efficiency as a function of fragment length



Sequencing efficiency as a function of sequence composition

- Can be applied to whole genome, cis maps, or other subsets

1. Explicit factor methods
   - Model bias due to GC content, fragment lengths, etc.
2. Matrix balancing (iterative correction)
   - Don't model explicit sources of bias. Only assumes factorizable biases.

Matrix balancing (iterative correction)

Alternatively, don't explicitly model sources of bias
Only assume:

$$O_{i,j} = B_i \cdot B_j \cdot T_{i,j}$$

biases are multiplicative

$$\sum_i T_{ij} = 1$$

total relative contact probabilities for each bin should sum to 1

Solution: Iterative correction algorithm

a.k.a.
Proportional fitting

a.k.a.
Matrix balancing

- Set $W_{ij}^{(0)} := O_{ij}, B^{(0)} := 1$
- Iterate until convergence:
  - Let $S_i = \sum_j W_{ij}^{(k)}$
  - Let $\Delta B_i = S_i / \bar{S}$
  - $W_{ij}^{(k+1)} = \frac{W_{ij}^{(k)}}{\Delta B_i \Delta B_j}$
  - $B_i^{(k+1)} = B_i^k \cdot \Delta B_i$

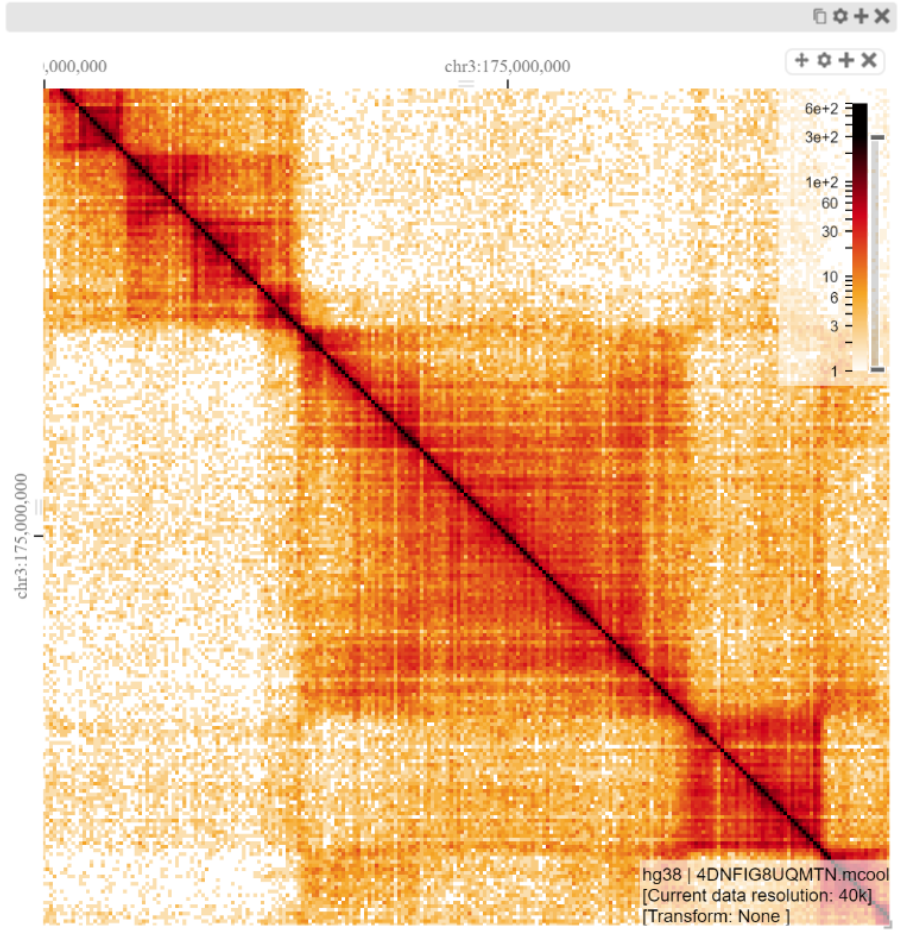
Imakaev et al., Nature Methods (2012)

Algorithms that solve the same problem go by various names:

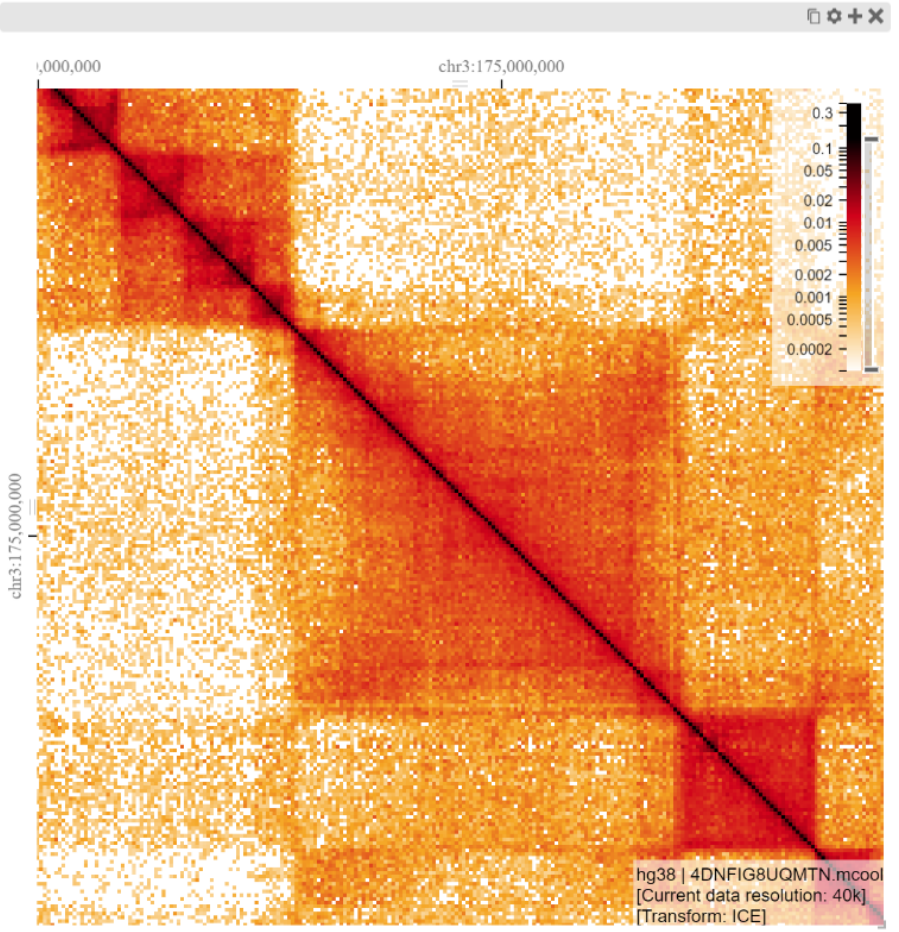
- Sinkhorn-Knapp
- Knight-Ruiz

Bias correction: Matrix Balancing

http://bit.ly/2uHd98l

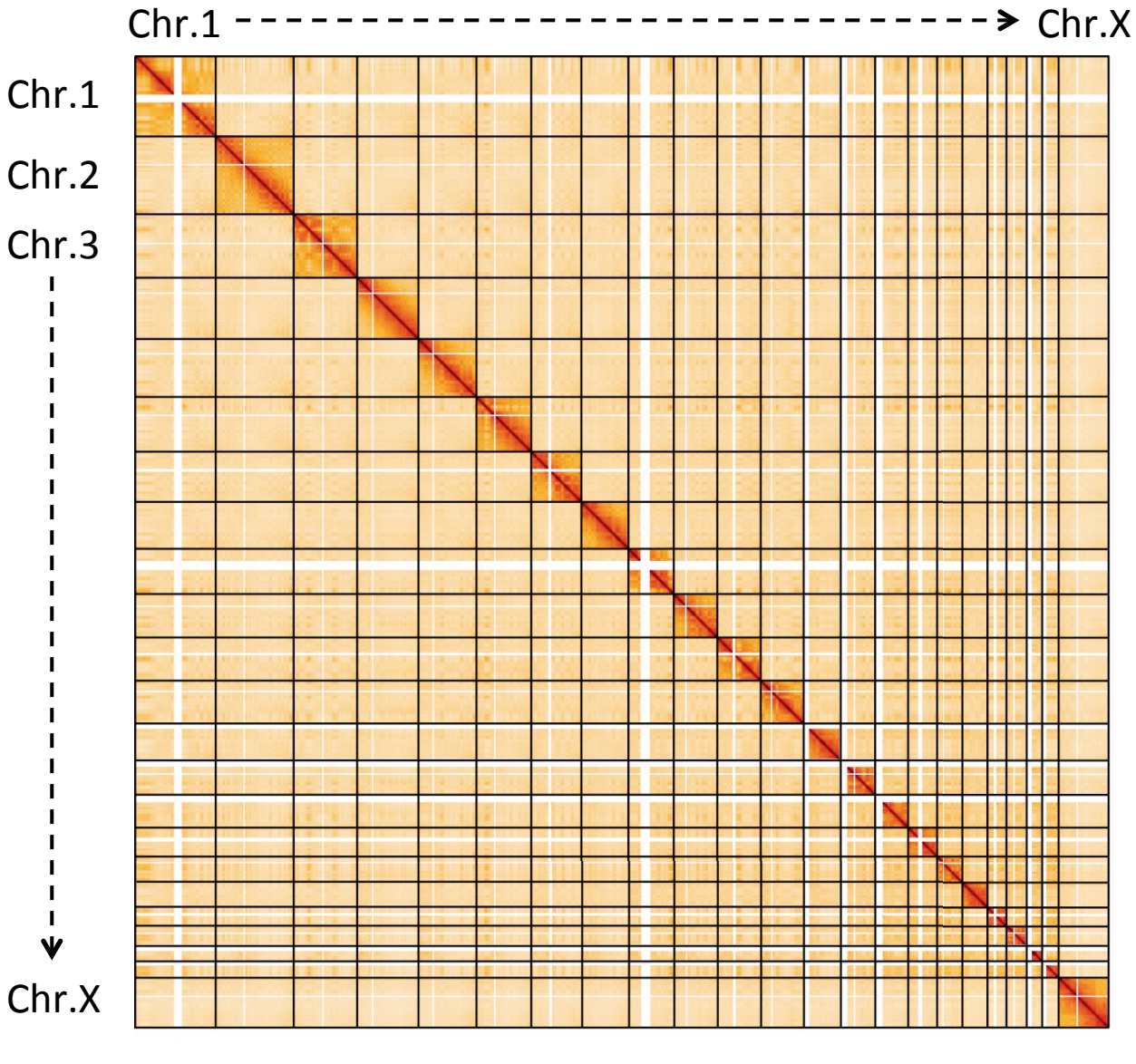


Uncorrected



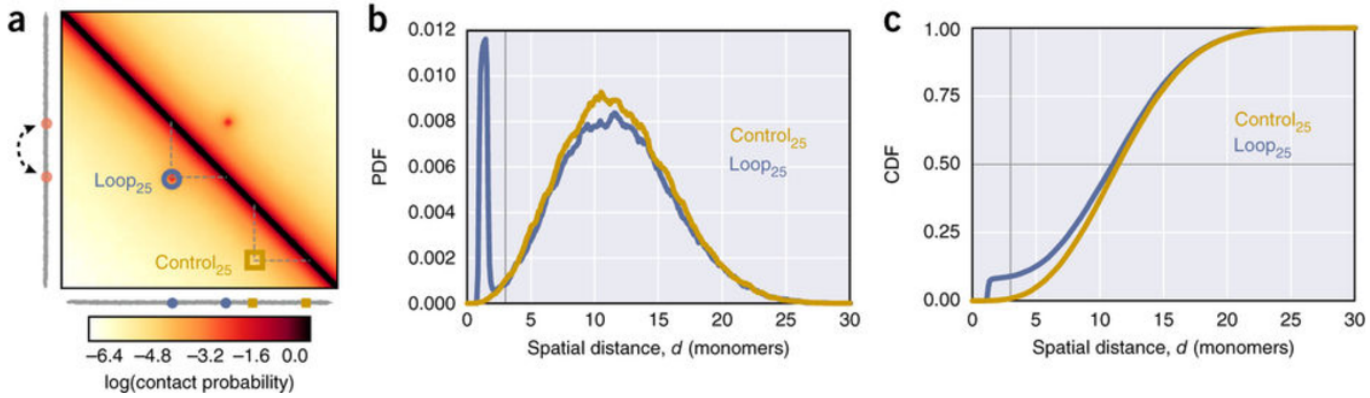
Corrected (balanced)

Hi-C Features



1. **P(s) curves**: decay of contact freq with genomic dist in *cis*
2. **Eigenvectors**: checkering pattern in *cis* and *trans*
3. **Insulation score**: Loci that deplete contact freq between neighbors
4. **"Pile-ups":** Aggregations of contact profiles centered at inferred features or at precise genomic landmarks

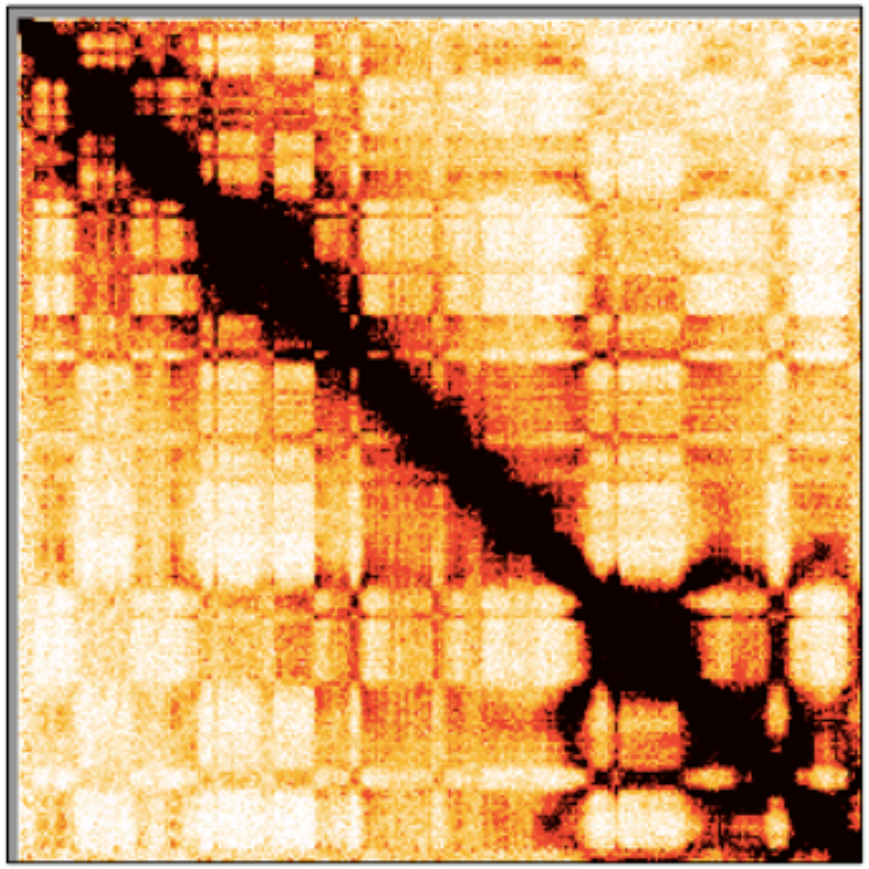
Caveat emptor



Fudenberg and Imakaev, 2017. *Nature*

- Contact probability != spatial distance
- Contact enrichment over background does not imply regulatory function
- TADs are not globules
- TADs are not static loops
- “Loops” are not static CTCF-CTCF loops
- “Loops” are not promoter-enhancer loops

.D. Bernal, and D. Crowfoot, X-ray photographs of crystalline pepsin, Nature, 133, 794–795 (1934).

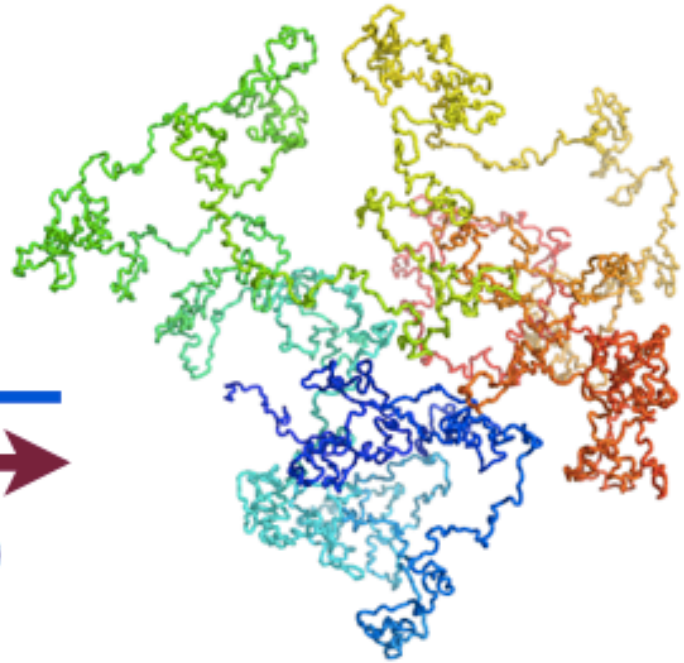


Protein contact map (NMR)

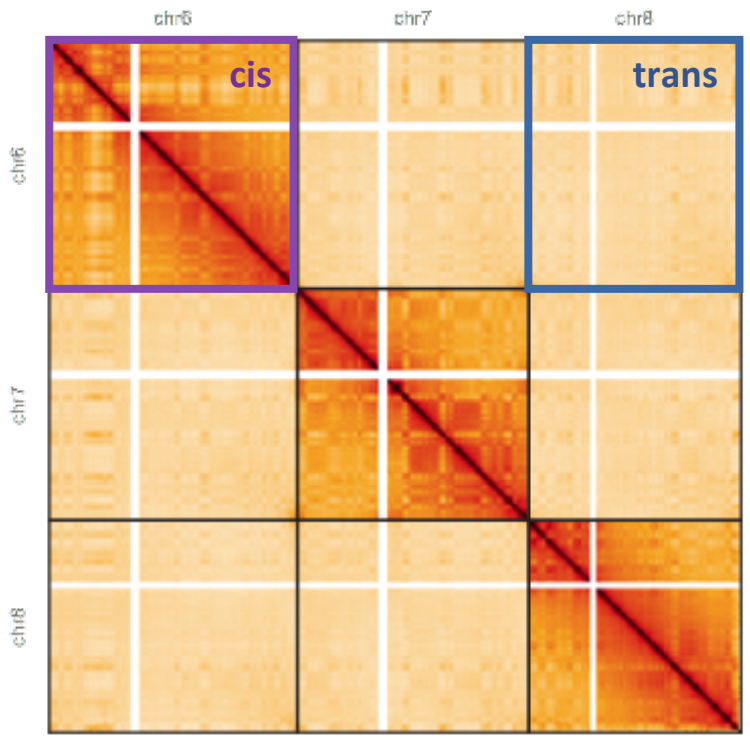
easy

HARD

ensemble

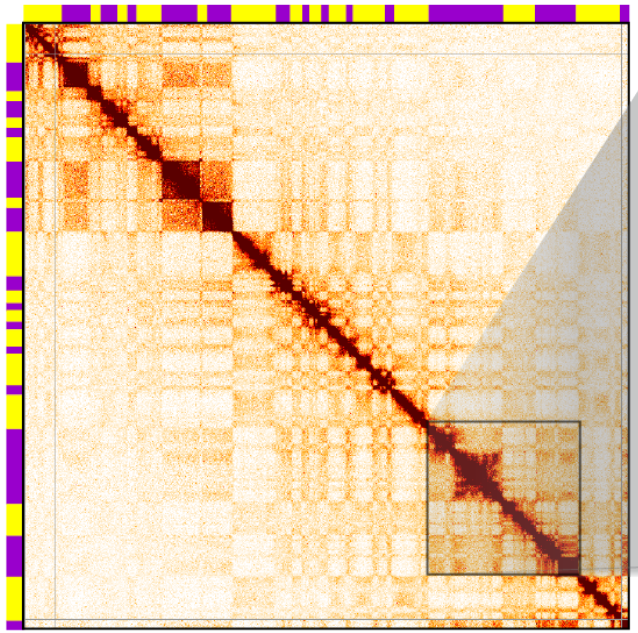


Scales of organization

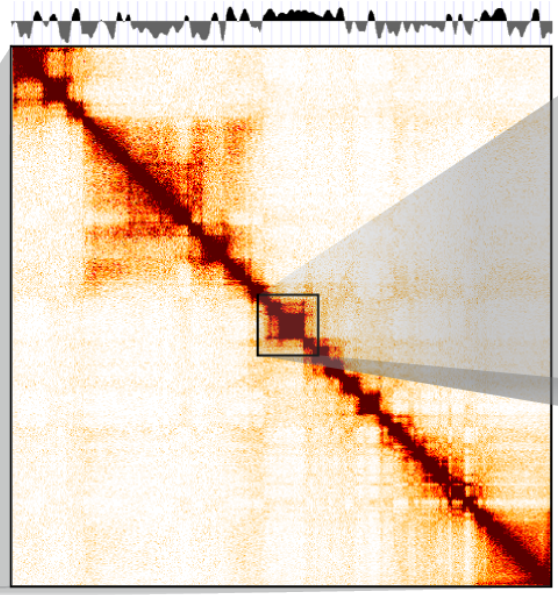


~~Territories / chromosomes~~

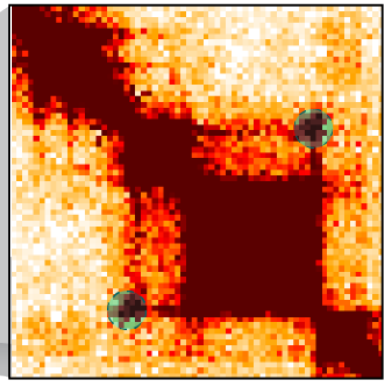
Chromosome 14, 106Mb



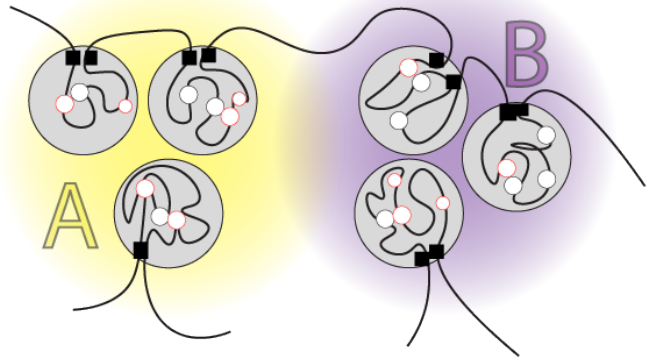
~~***Compartments 1-10 Mb***~~
100-500 Kb bins



~~***TADs 0.1-1Mb***~~
40-100 Kb bins



~~***Loops 0.01-0.1Mb***~~
10-40 Kb bins



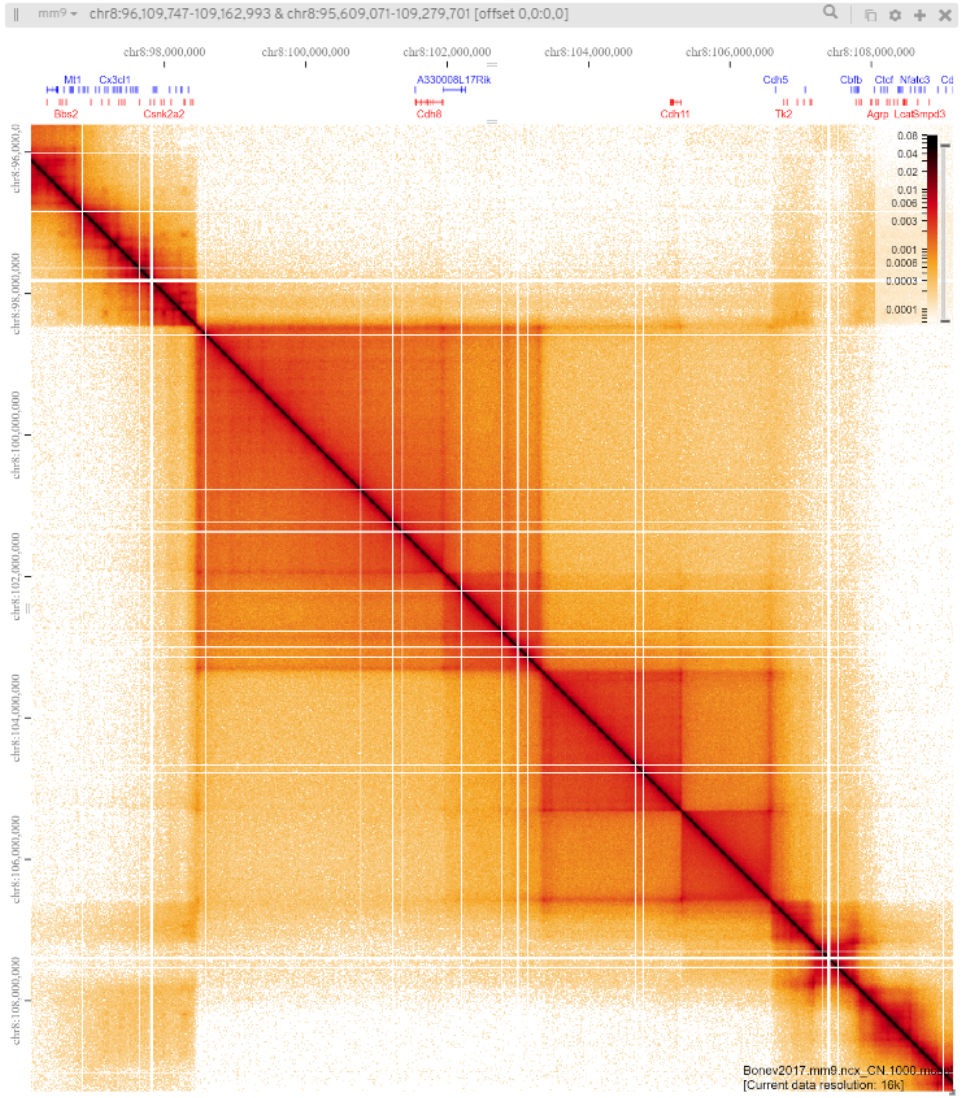
Patterns, quantifications & interpretations

- Contact decay profile with genomic distance
- Diagonal squares that checker in cis and in trans
- Diagonal squares that don’t checker
- Peaks
- Lines

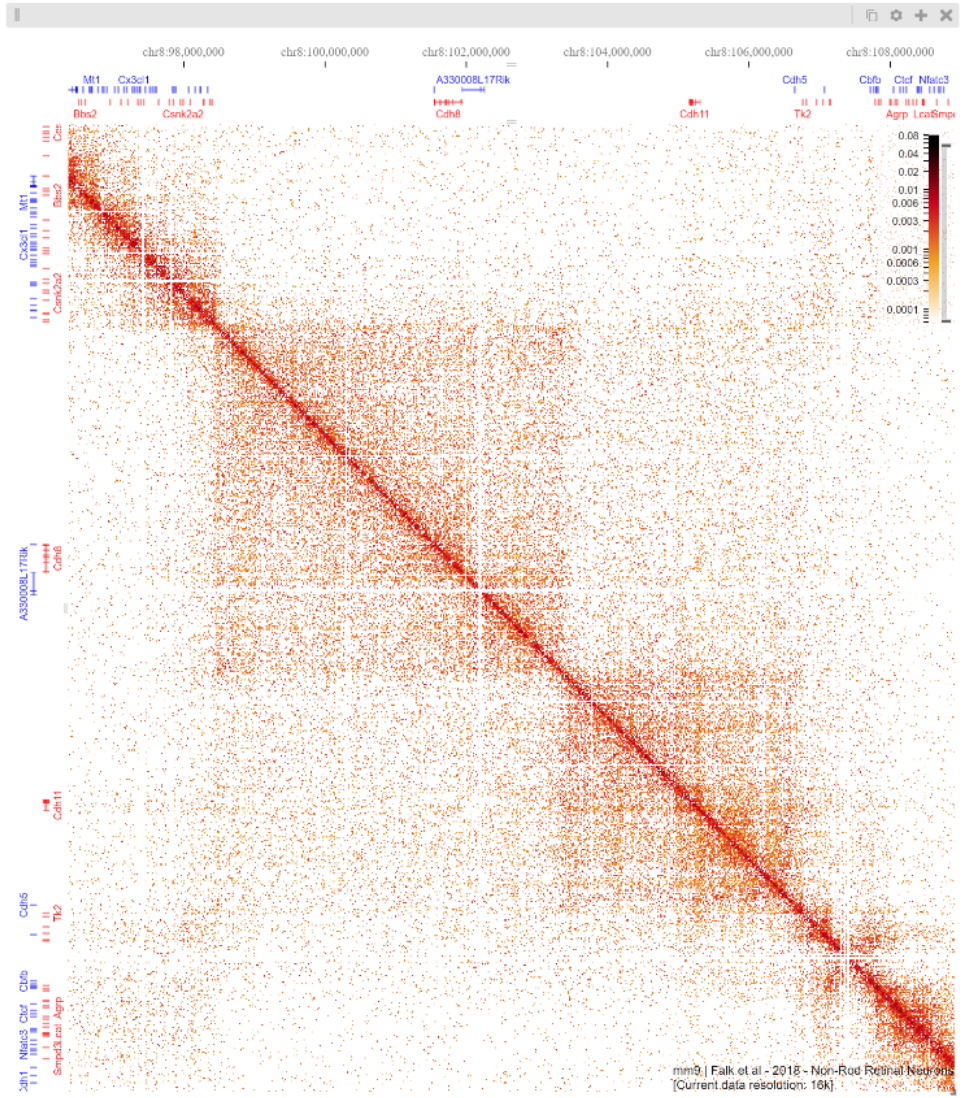
We are bad at naming things.

Mistaking map for territory:
https://en.wikipedia.org/wiki/Reification_(fallacy)

What the TAD?



~5B valid pairs



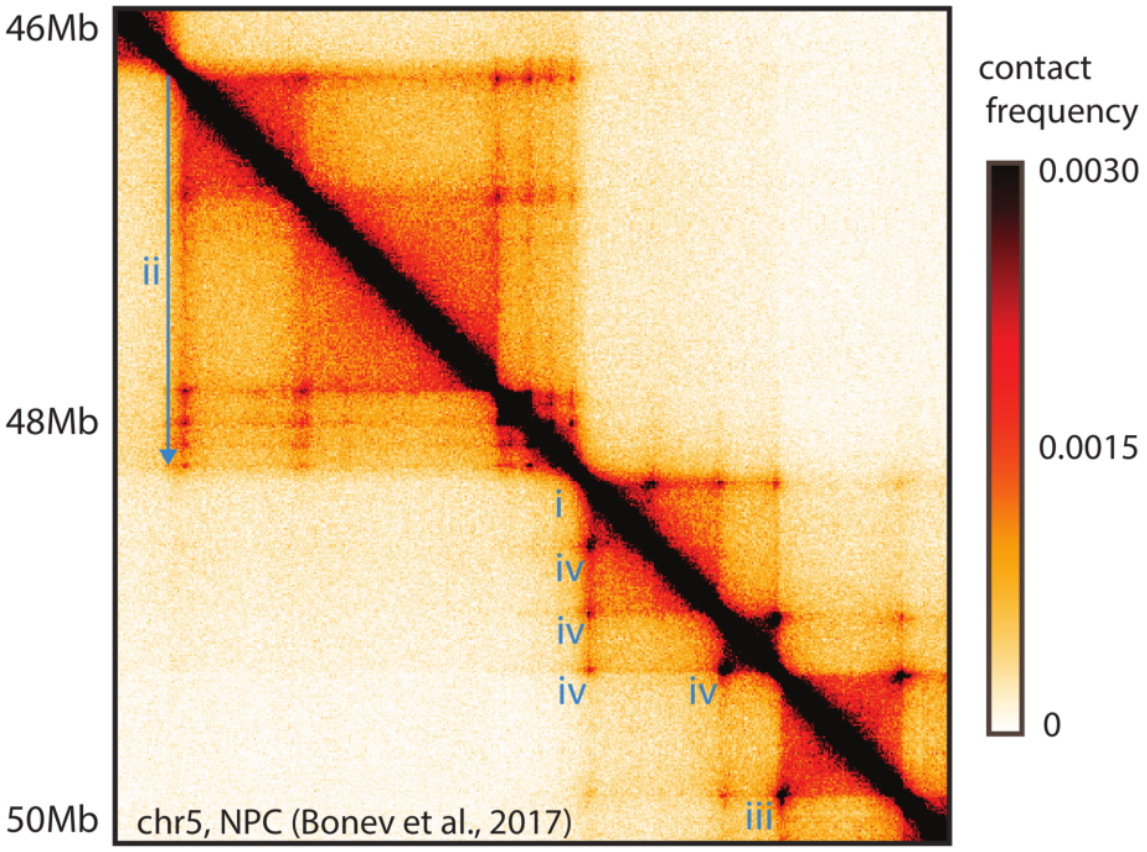
100M valid pairs

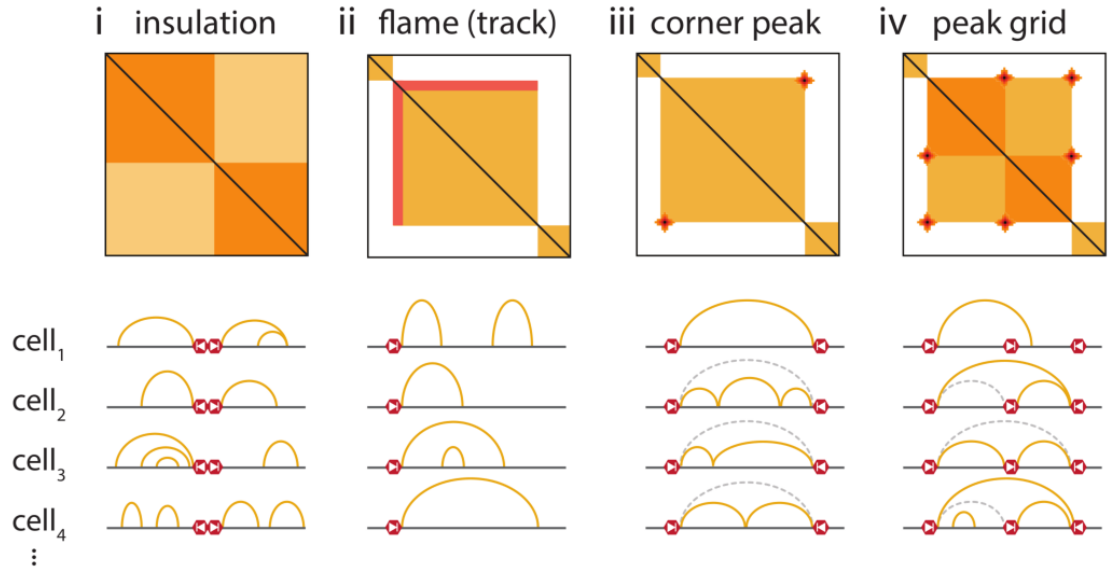
- Nora et al, 2012: We discover a series of discrete **200-kilobase to 1 Mb** topologically associating domains (TADs), … **differ from the higher-order organization** recently observed by Hi-C, **corresponding to much larger domains of open or closed chromatin, that come together** in the nucleus to form A and B types of compartments.
- Dixon et al, 2012: We identify large, **megabase-sized** local chromatin interaction domains, which we term topological domains, as a pervasive structural feature of the genome organization. .... Notably, a ***subset of the domain boundaries we identify appear to mark the transition between either LAD and non-LAD regions of the genome, the A and B compartments***, and ***early and late replicating chromatin***.
- In Drosophila (Sexton et al, 2012): physical domains

- Later: subTADs, contact domains, loop domains, metaTADs, etc.

http://mirnylab.mit.edu/projects/emerging-evidence-for-loop-extrusion/

TAD anatomy

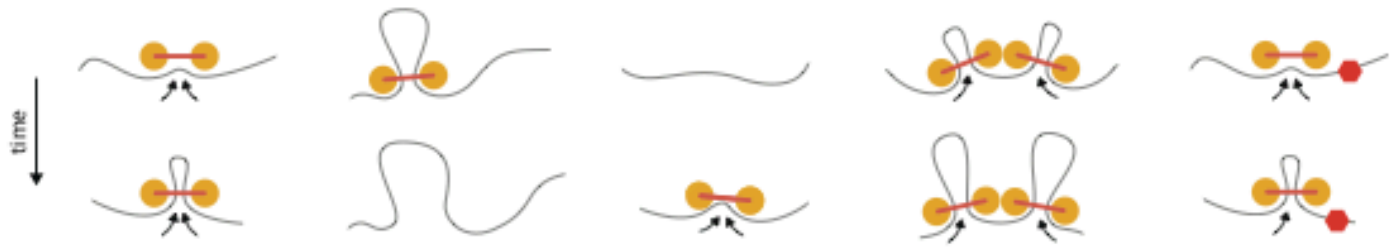




Fudenberg, Abdennur, et al, 2018. Cold Spring Harbor Symposia on Quantitative Biology

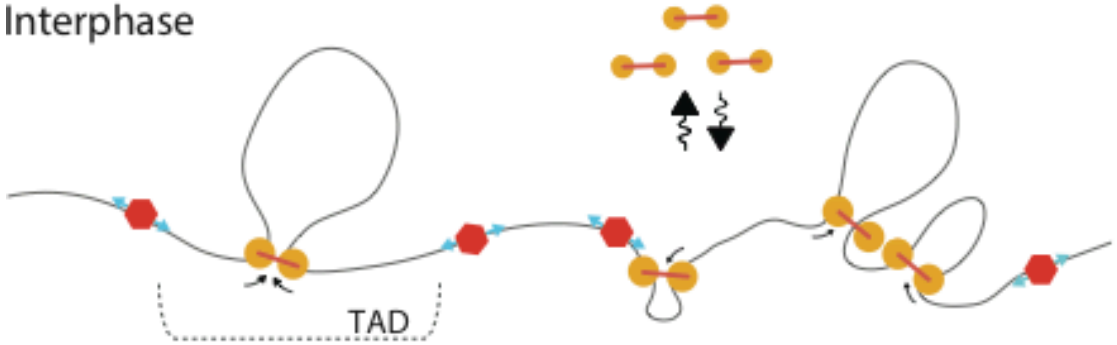
The loop extrusion model of TAD formation

Loop Extrusion as a General Mechanism underlying Chromosome Organization



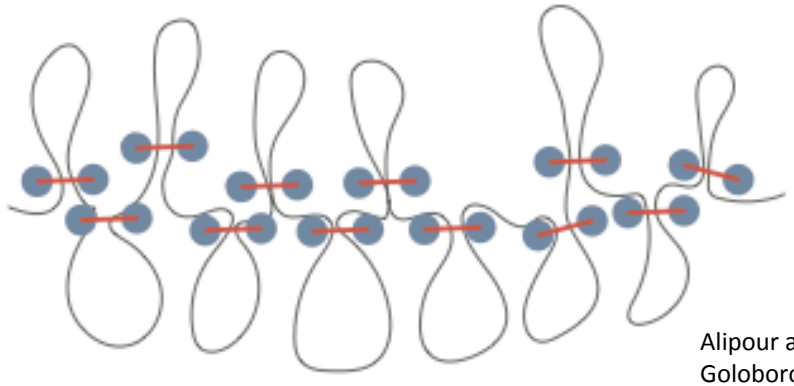
Riggs, 1990
Kimura, 1999
Nasmyth, 2001

Interphase

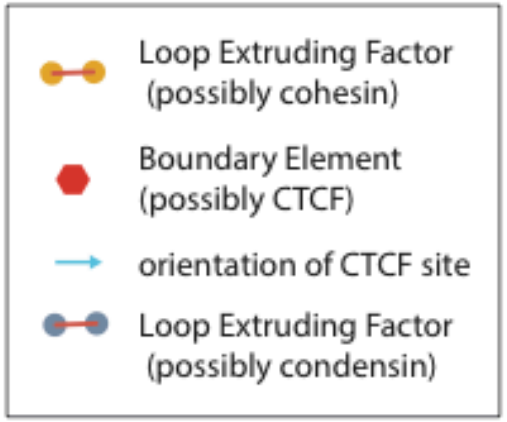


Fudenberg et al. 2016
Sanborn et al. 2015

Metaphase

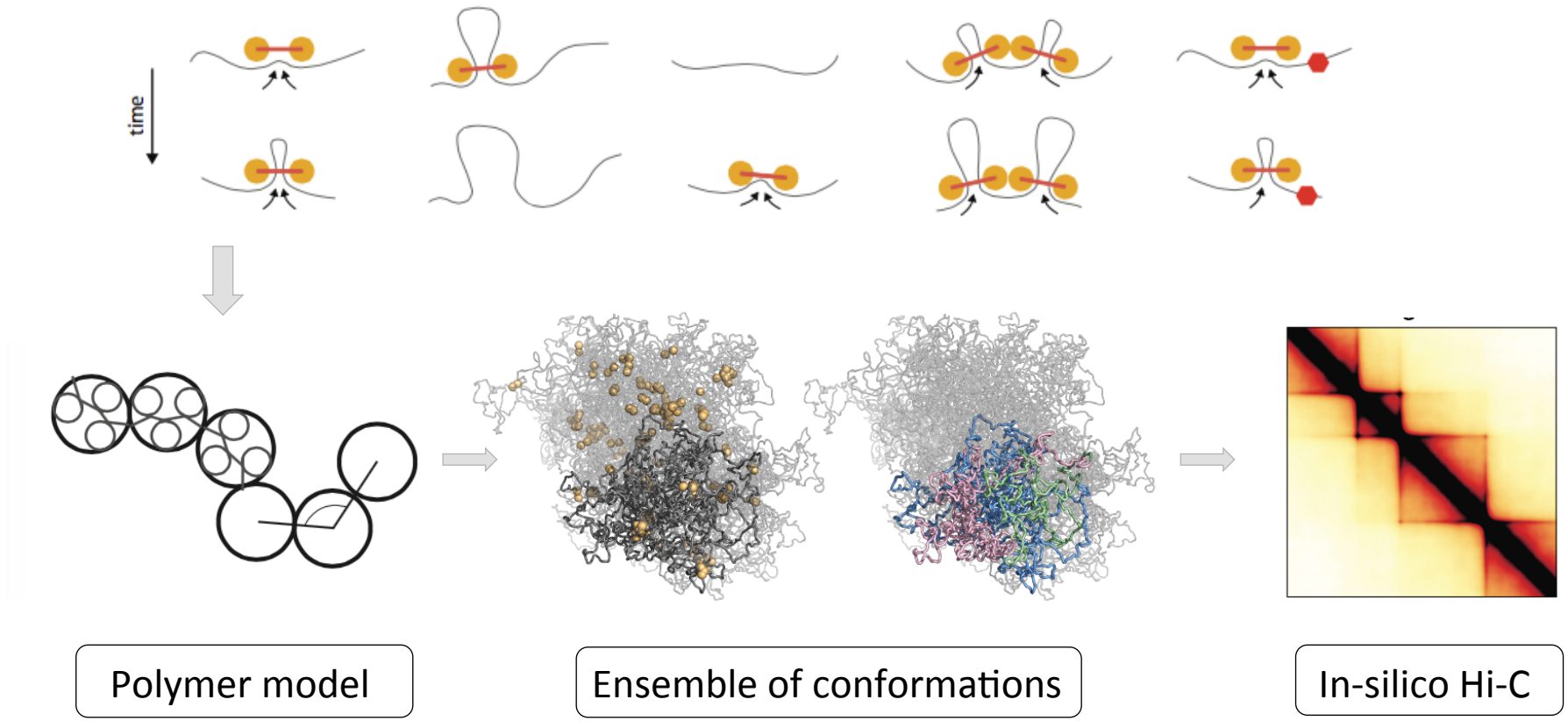


Alipour and Marko 2012
Goloborodko et al. 2016



1D extrusion model coupled to 3D polymer simulations

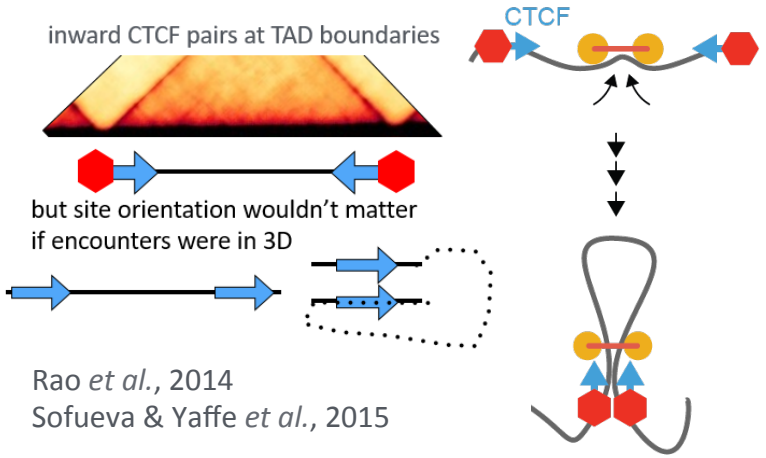
Update rules for 1D model



Polymer simulations in OpenMM, see: Eastman, 2013; Eastman, 2016
For review see Imakaev, GF, Mirny, 2015

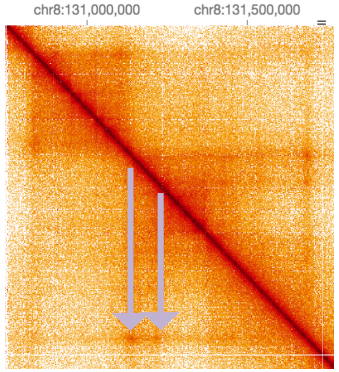
TAD model checklist

Loop extrusion can
orient over long distances

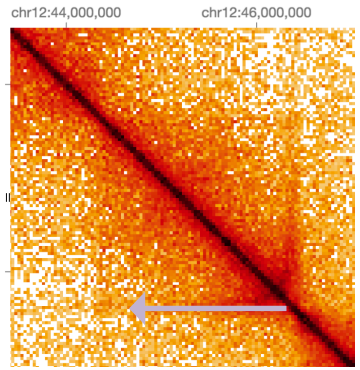


Rao *et al.*, 2014
Sofueva & Yaffe *et al.*, 2015

Flames (aka lines) are a pervasive feature of
Hi-C maps, consistent with loop extrusion



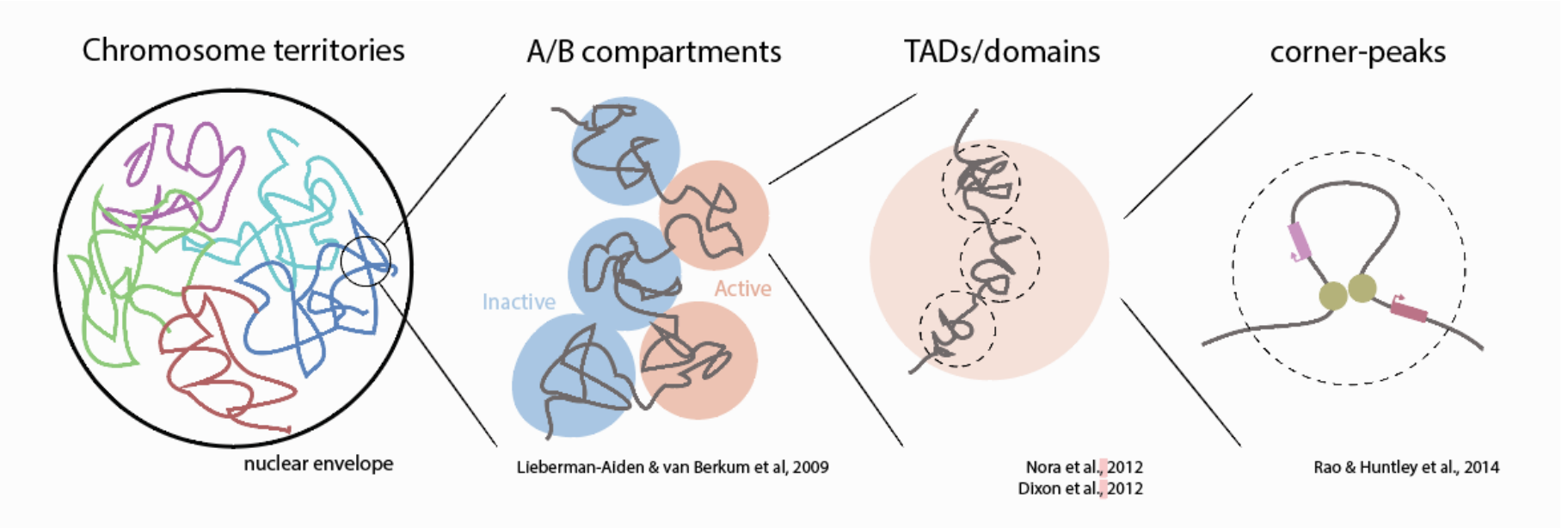
Human GM12878 from
Rao & Huntley, 2014



Mouse liver from
Schwarzer et al., 2016

Enriched interactions throughout TAD	✔
Does not form alternating compartments	✔
Boundary deletion → spreading	✔
Peaks only between proximal boundaries	✔
Explain inward CTCF at boundaries	✔
Accumulation LEFs inside of bound CTCFs	✔
Not simple stable loops	✔

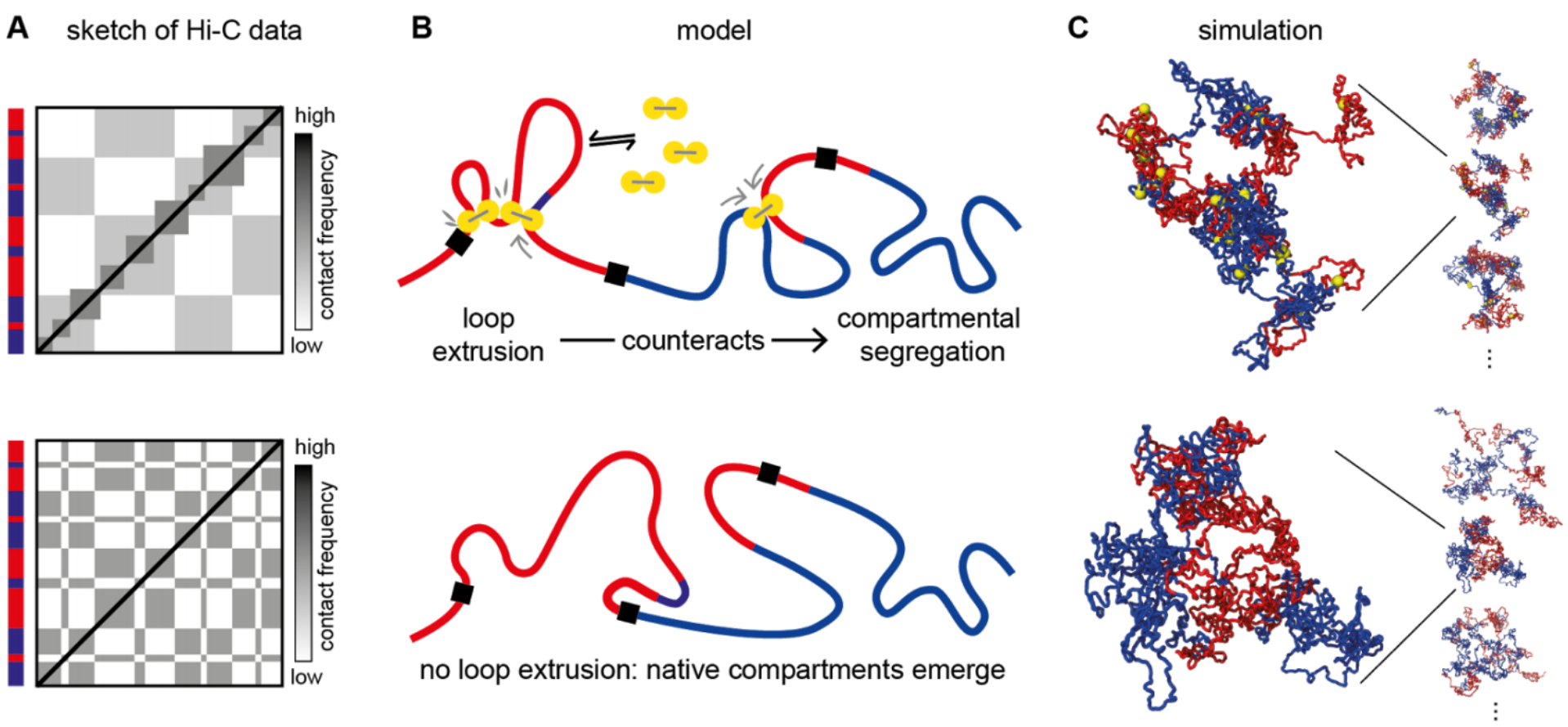
Interphase organizational hierarchy?



Organizational ~~hierarchy~~ processes

population average

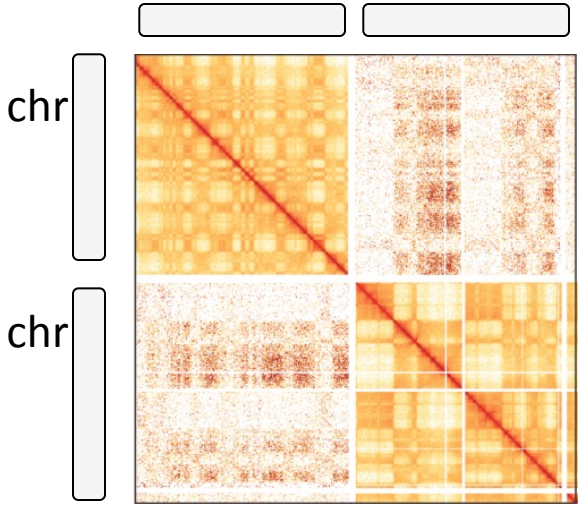
ensemble



Nuebler et al, 2017 (bioRxiv)

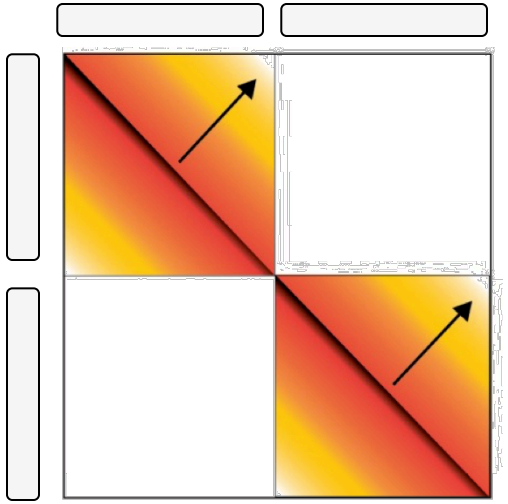
P(s): Polymer nature of the chromatin fiber

input



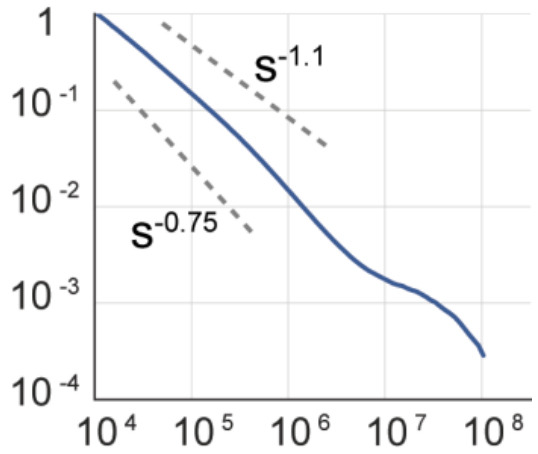
cis Hi-C maps

feature



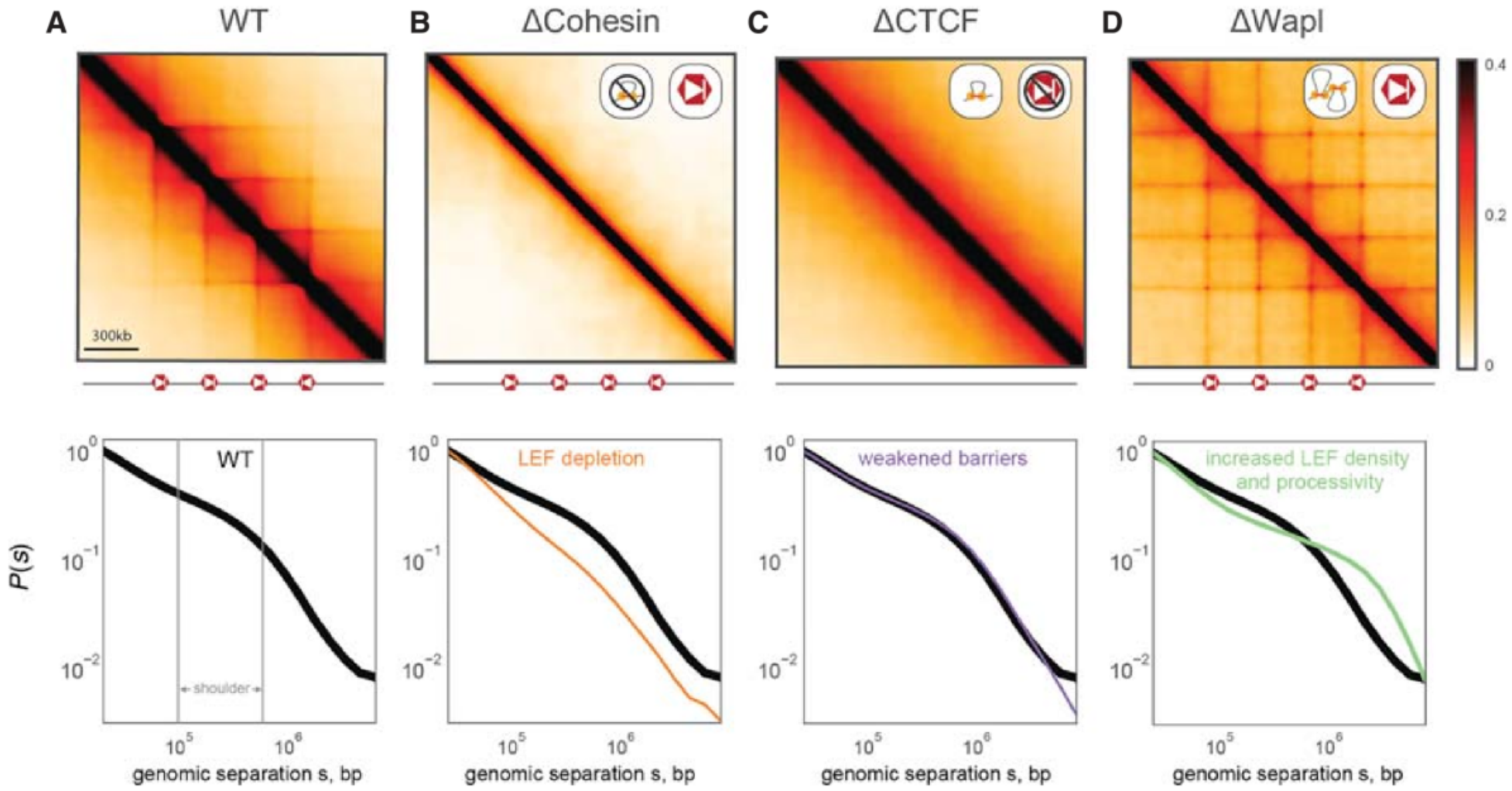
decay of intra-chrm contact frequency *P(s)* with genomic distance, s

quantification



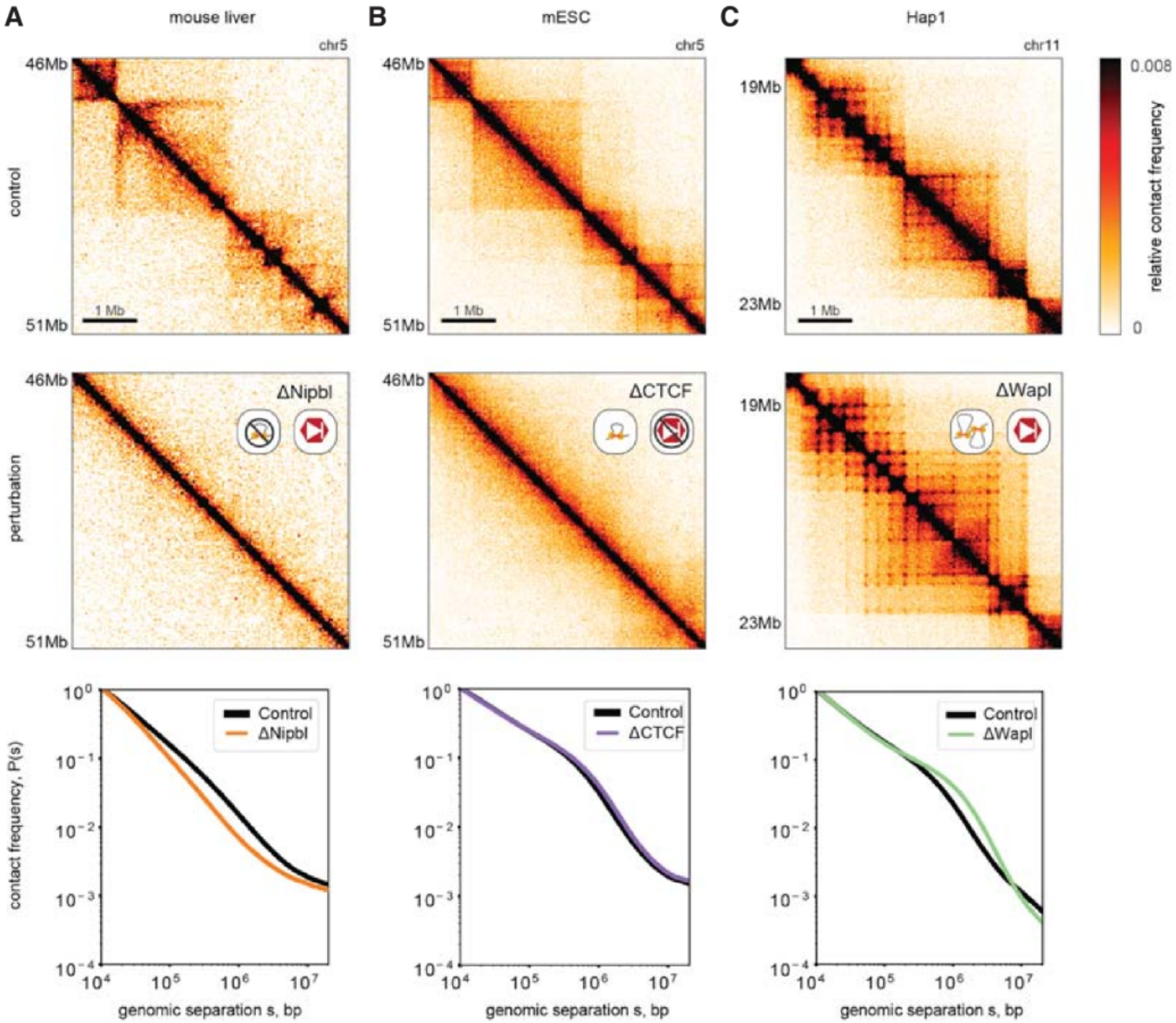
P(s) curves

The head and shoulders of loop extrusion



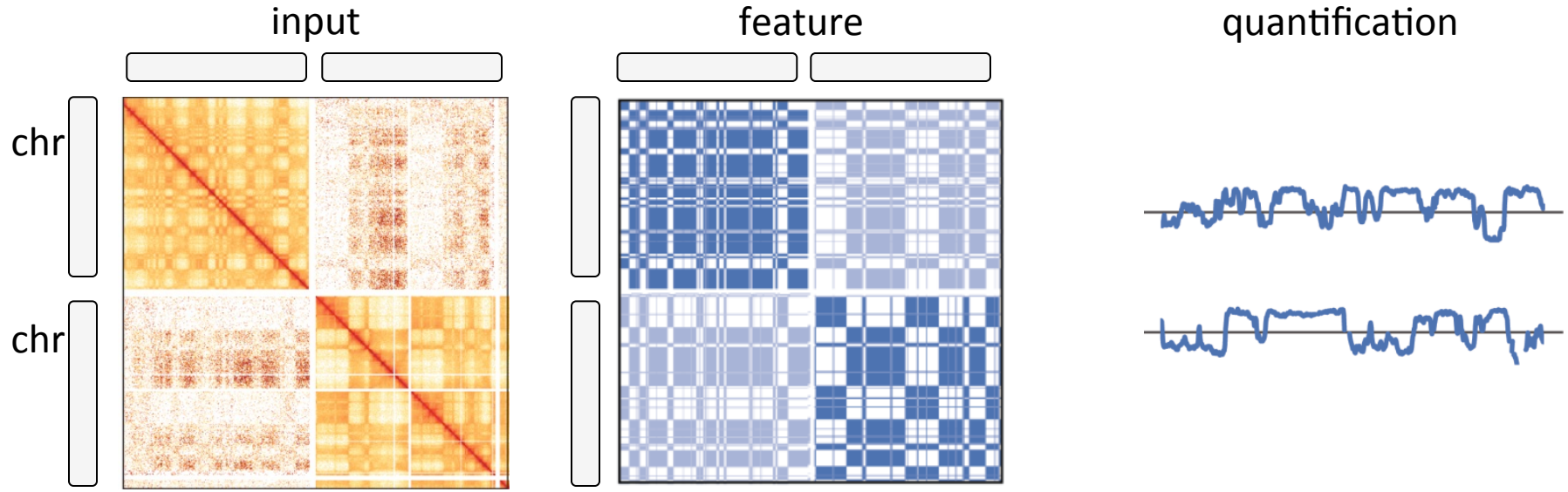
Fudenberg, Abdennur et al, 2018

Schwarzer et al, 2017 Nora et al, 2017 Haarhuis et al, 2017



Fudenberg, Abdennur et al, 2018

Compartmentalization (segregation)



cis and *trans*
Hi-C maps

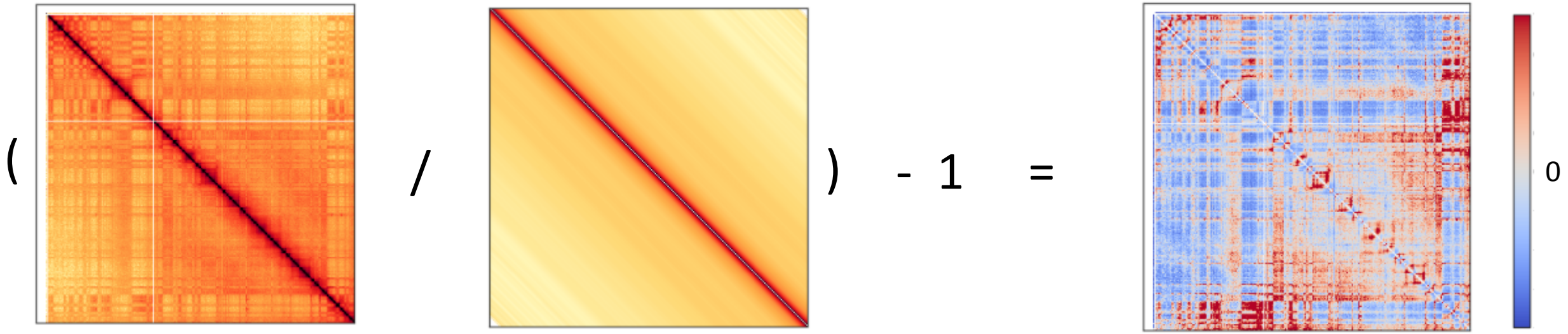
checkered pattern
in *cis* and *trans*

A/B compartment
eigenvector

*Other approaches: clustering

Eigendecomposition

$O\downarrow ij$ $S\downarrow |i-j|$



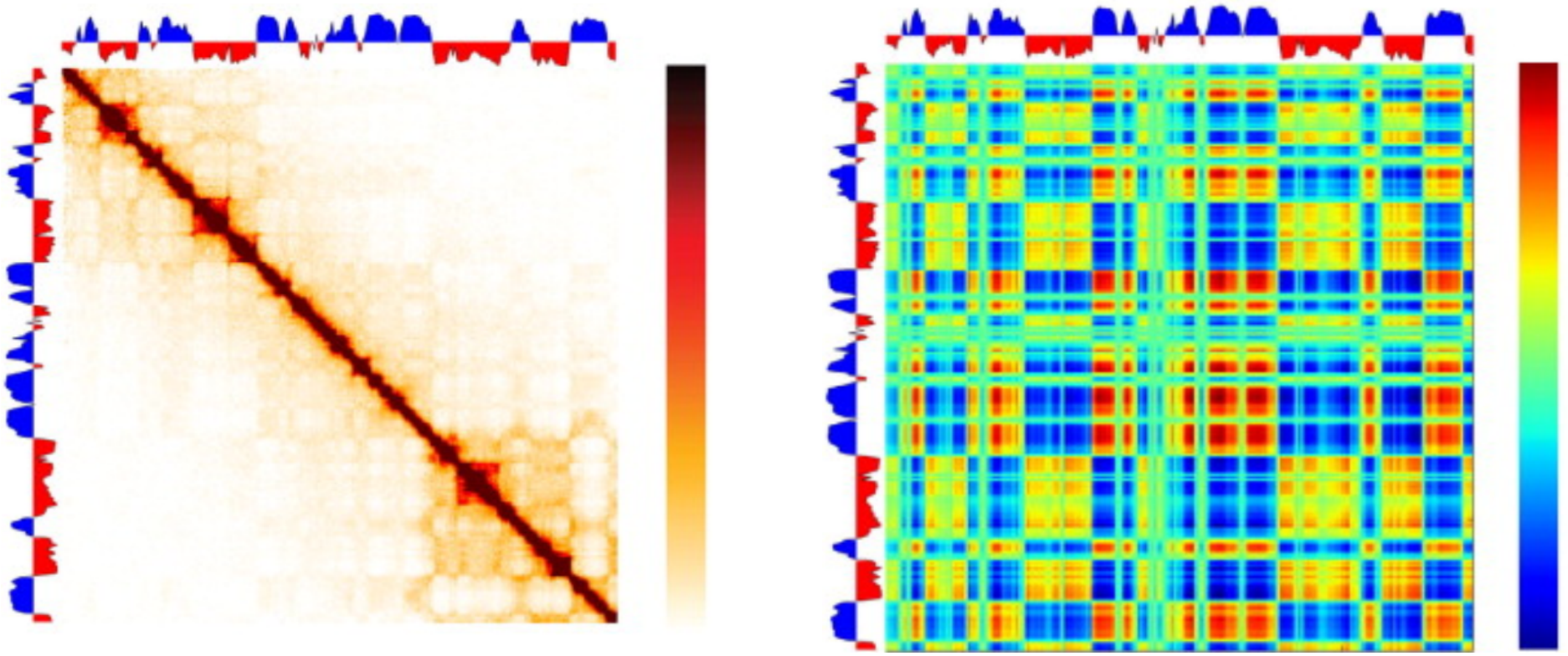
Def. of eigenvector decomposition says:

$A = Q\Lambda Q^{-1}$, where columns of Q are unit eigenvectors e_1, e_2, e_3,... ($\lambda_1 > \lambda_2 > \lambda_3 \ldots$)

$A = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \ldots$

$A \approx \lambda_1 e_1 e_1^T$ rank-1 approximation

Eigendecomposition

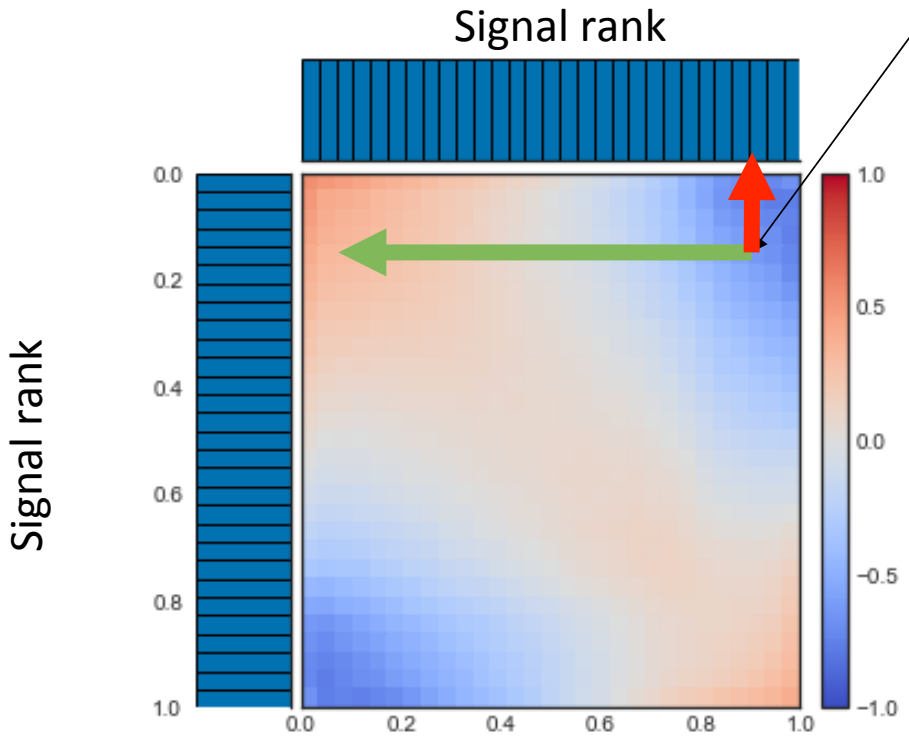


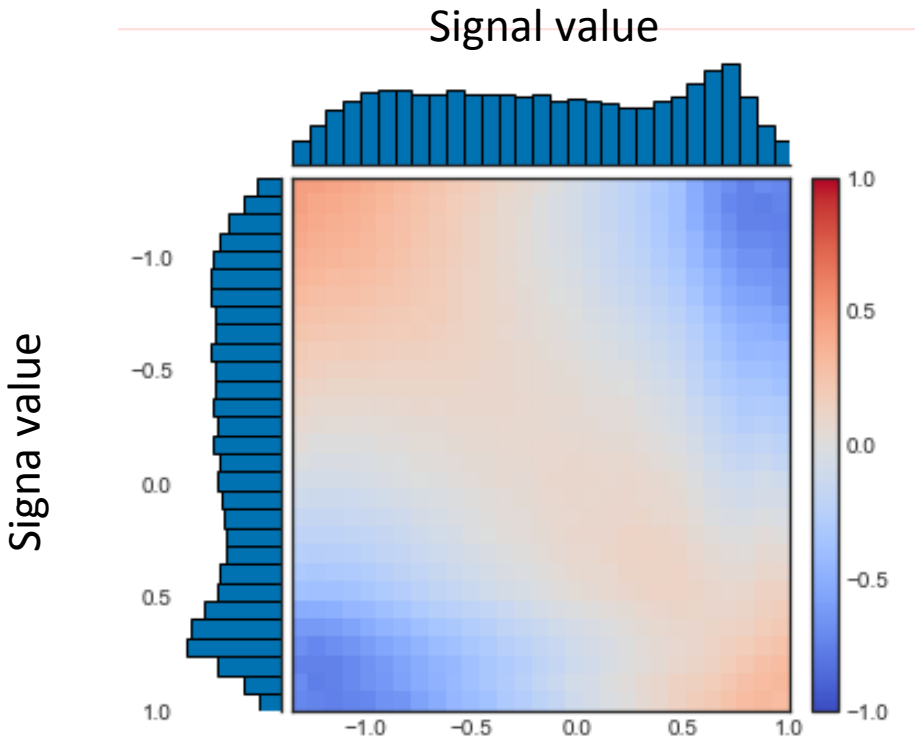
Lajoie et al, 2015

E1 captures the checkerboard pattern!

The “saddle” plot

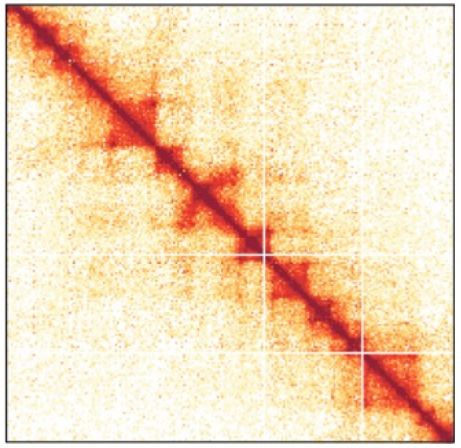
Average OBS/EXP contact frequency for all pixels whose bin1 EIG signal falls in the signal quantile in green, and bin2 EIG signal falls in the signal quantile in red





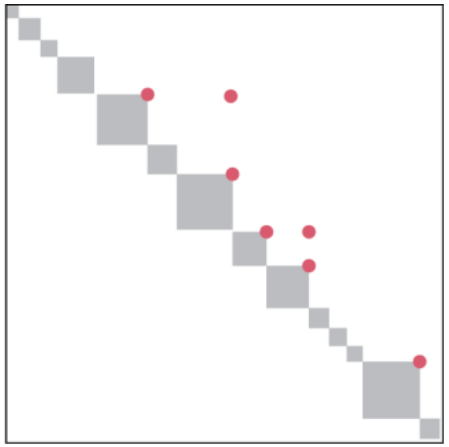
Local features

input



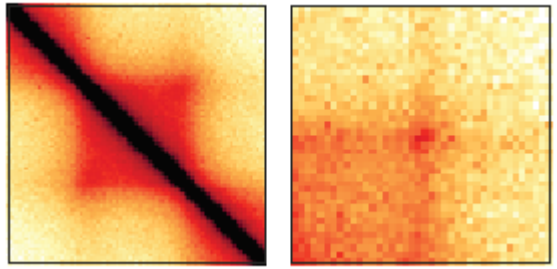
cis, short distance
(< few Mb)

feature



Squares and peaks of
contact enrichment

quantification

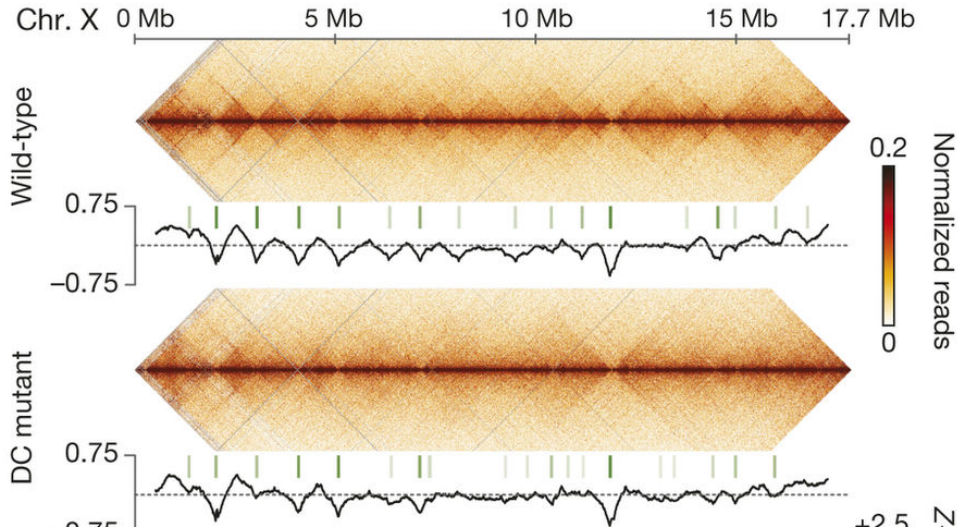


Various detection
algorithms
Metrics of insulation,
directionality, etc.

1-D score tracks

- Insulation score

  Sliding diamond

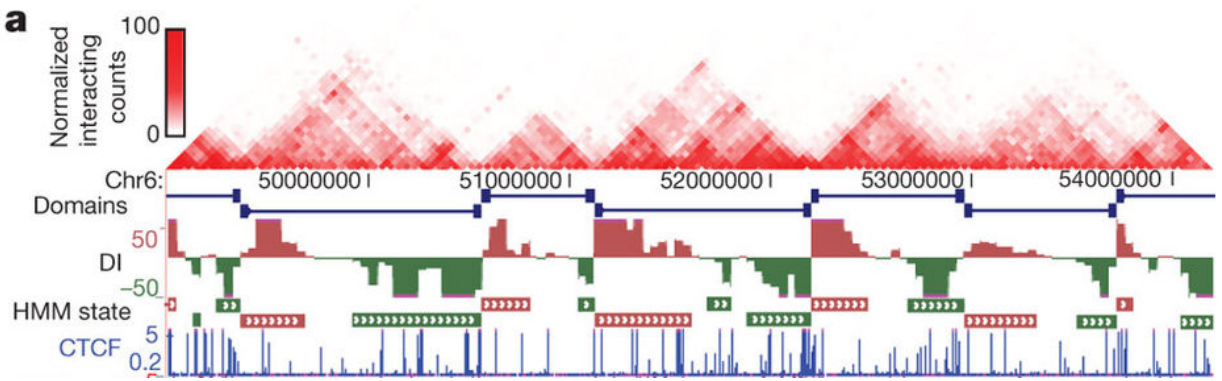


Crane et al, 2013

- Directionality Index

  Upstream vs downstream

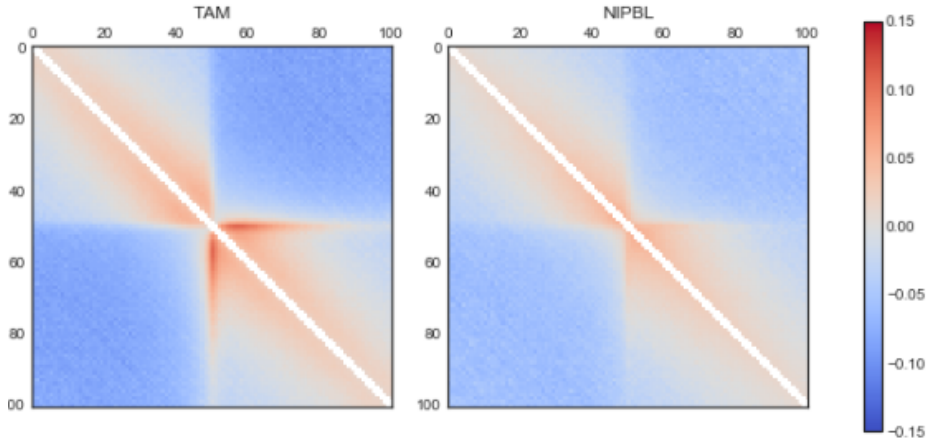
$$DI = \left(\frac{B - A}{|B - A|}\right)\left(\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}\right)$$

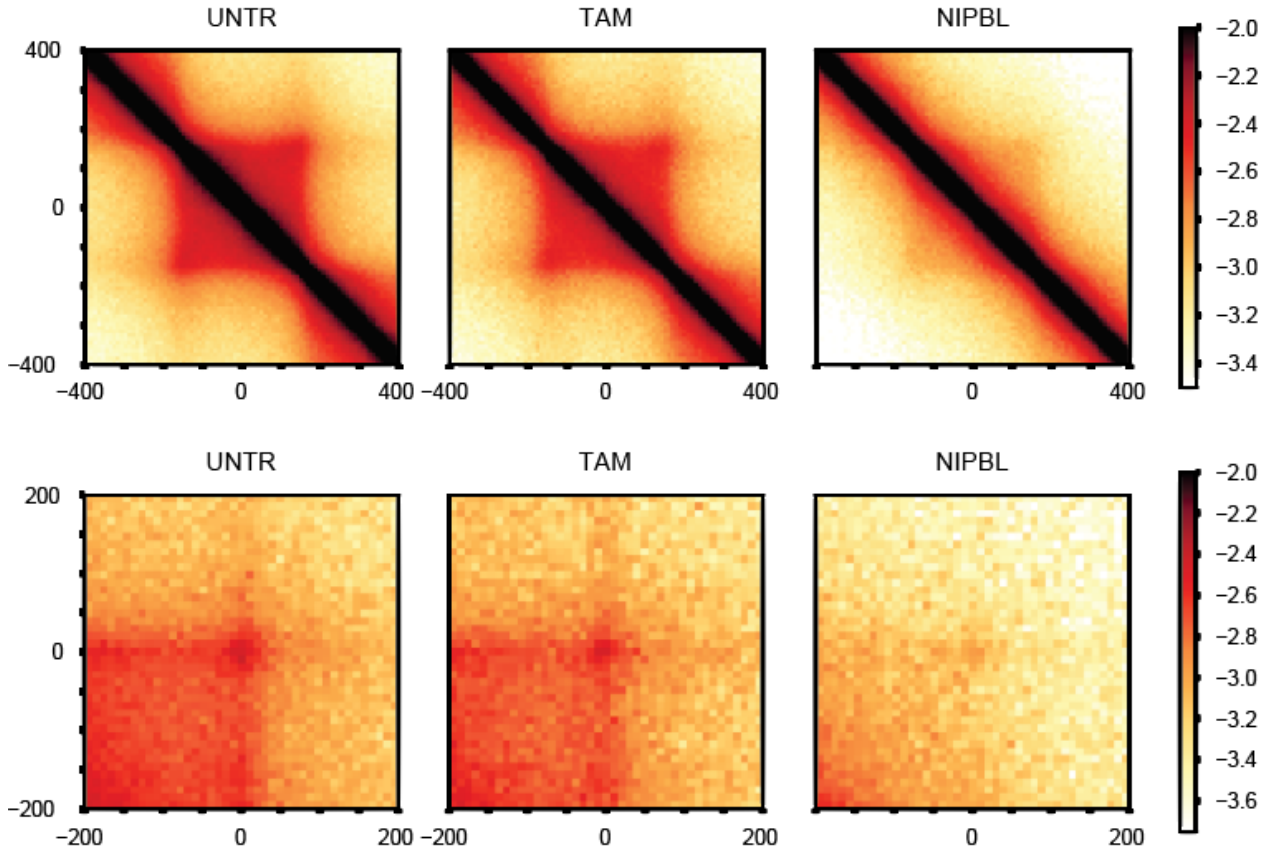


Dixon et al, 2012

The mighty pileup

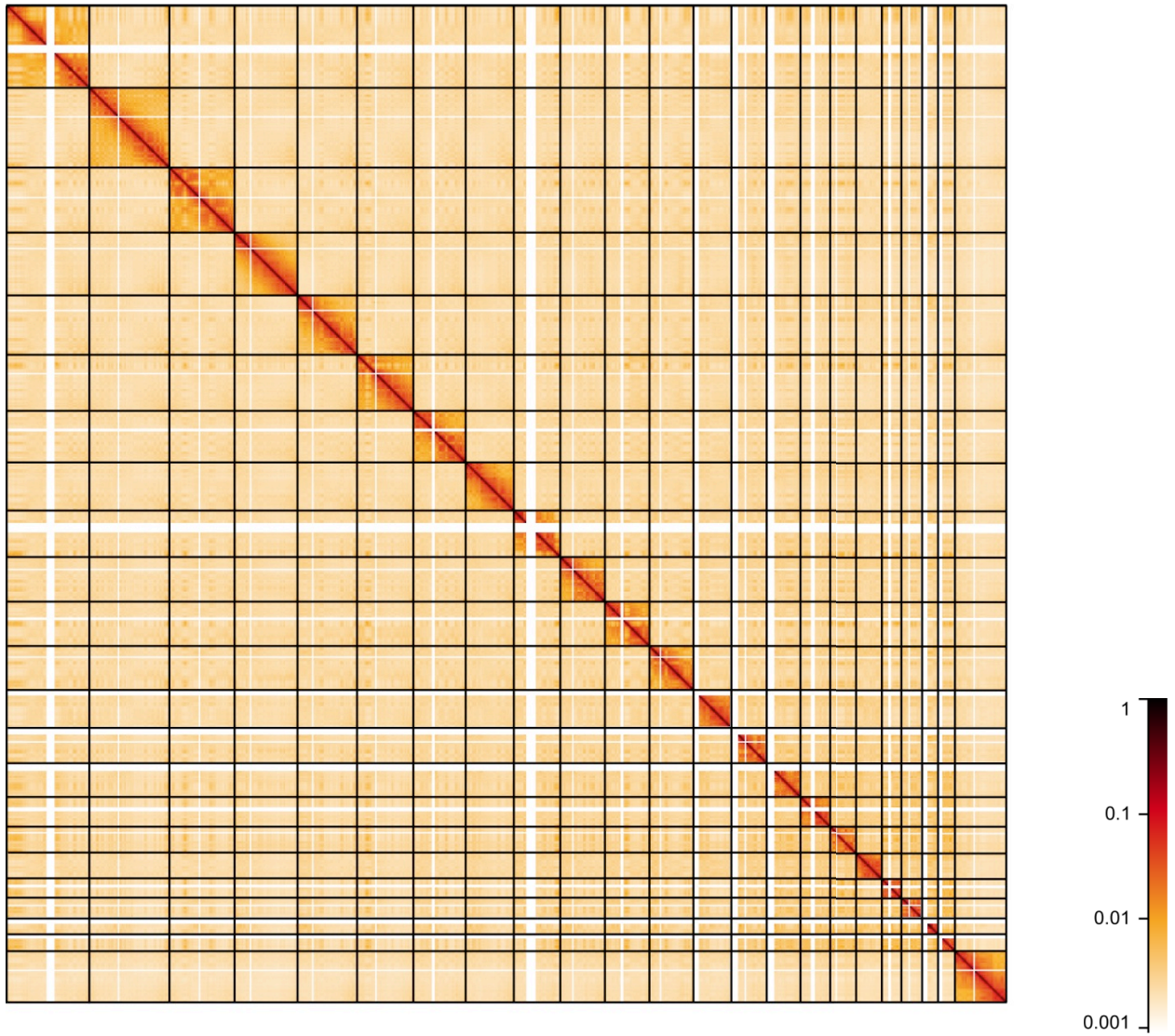
* Enhance your resolution!
* Beware of “average footprints”



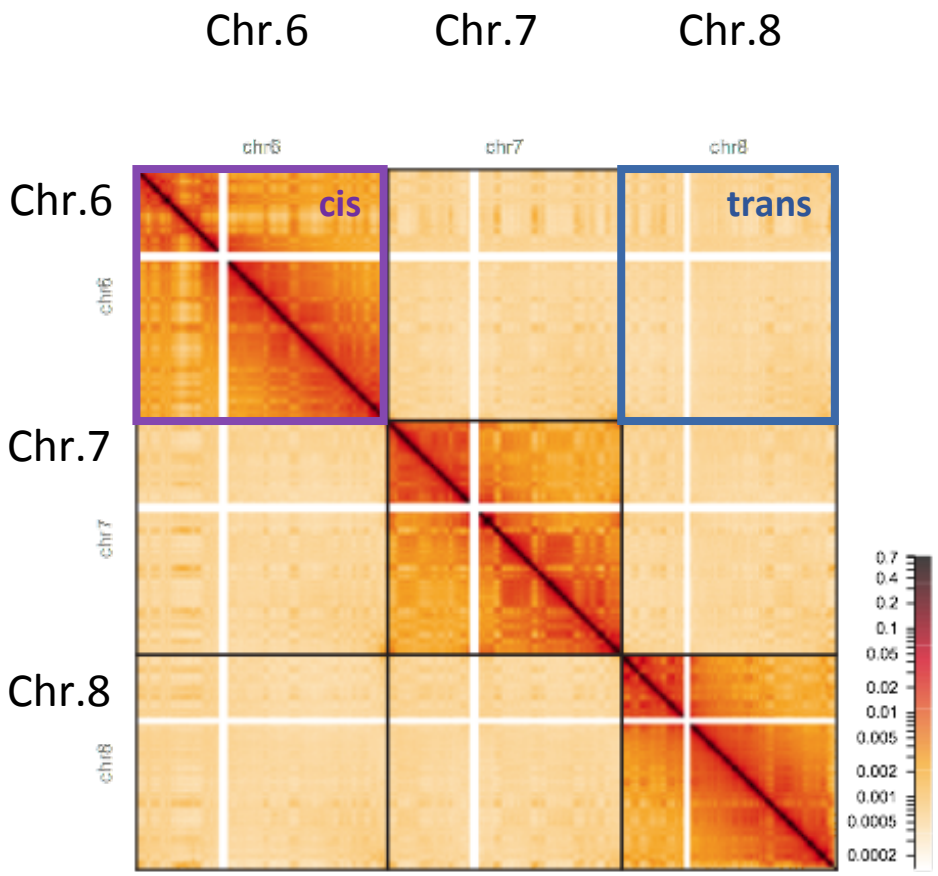


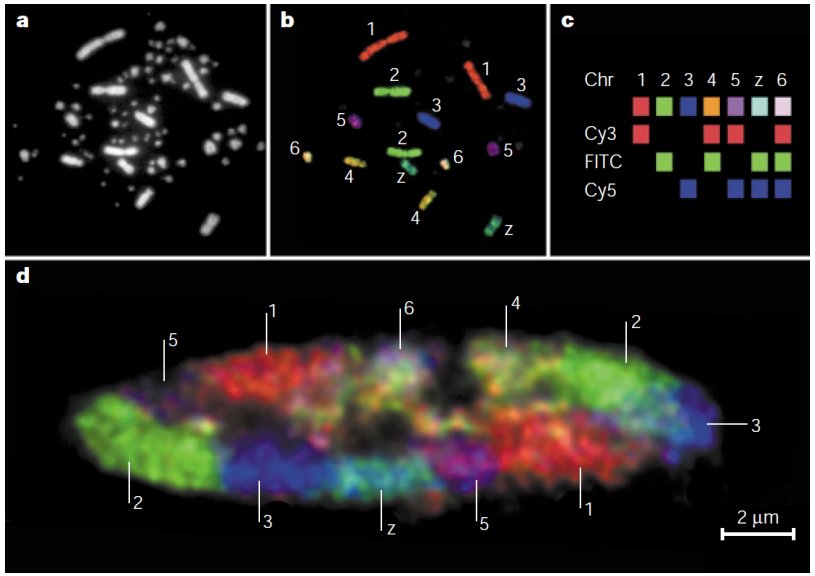
Schwarzer et al, 2017. Nature

http://bit.ly/2hdhkAn



Territoriality





Cremer & Cremer, Nature Review Genetics (2001)