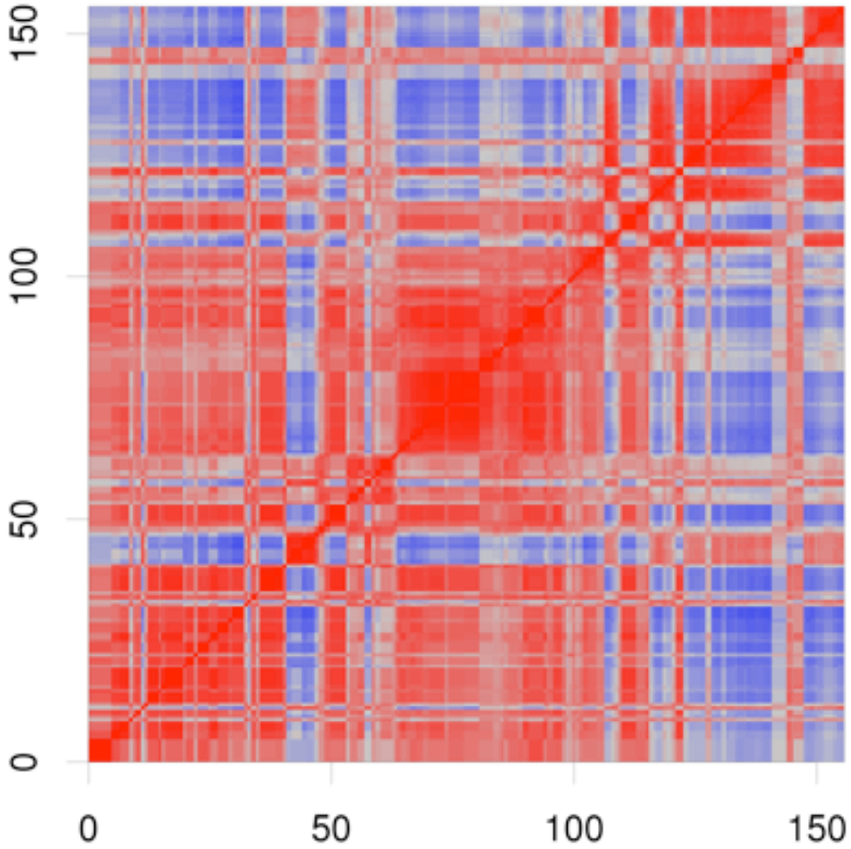
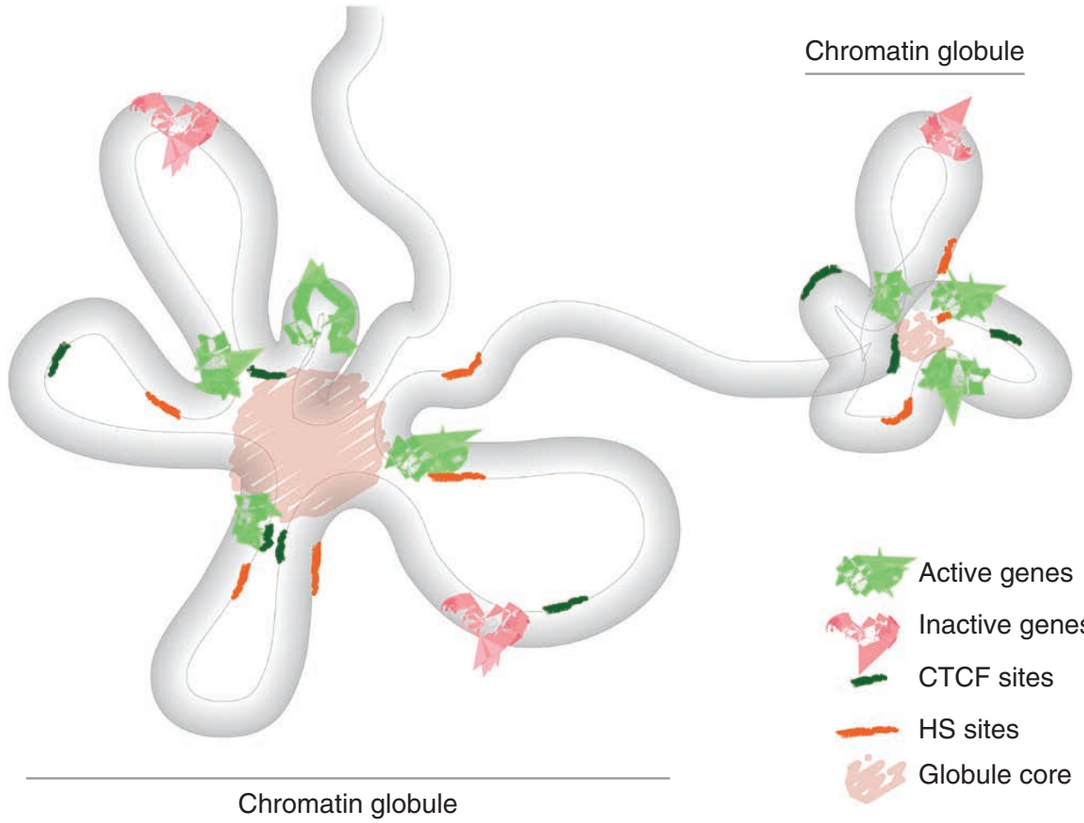
Hi-throughput Chromatin Conformation capture -- Hi-C





Bau 2010

CSAMA 2015, Brixen
18. 06. 2015.

Aleksandra Pekowska
aleksandra.pekowska@embl.de

EMBL 40 YEARS | 1974–2014

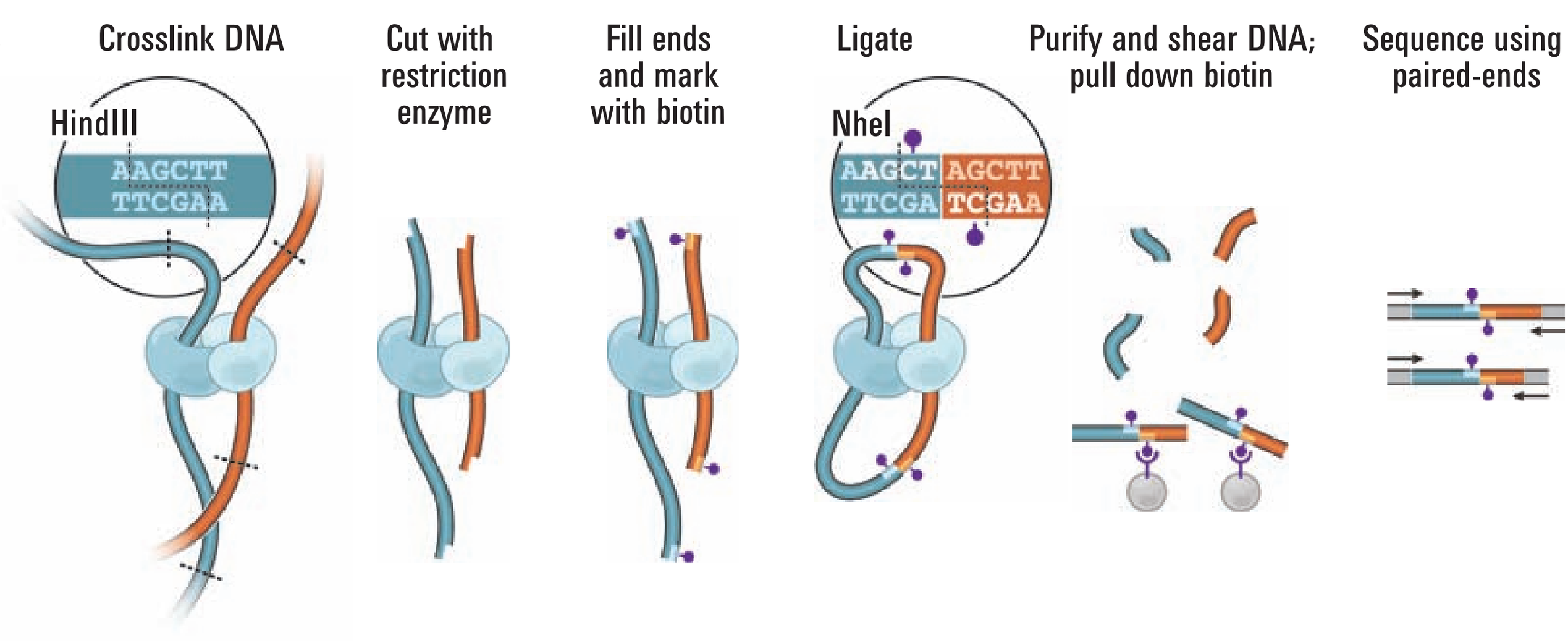
Outline of the lecture

Purpose: introduce basic steps and key considerations in Hi-C analysis

1. **The Hi-C/TCC method**
2. **What can we measure with Hi-C?**
3. **Study design**
4. **Hi-C analysis workflow:**
   a. Preprocessing
   b. Quality controls
   c. Normalization
   d. Chromosome-wide analysis
   e. Identification of local structures - TADs
   f. Identification of significant interactions

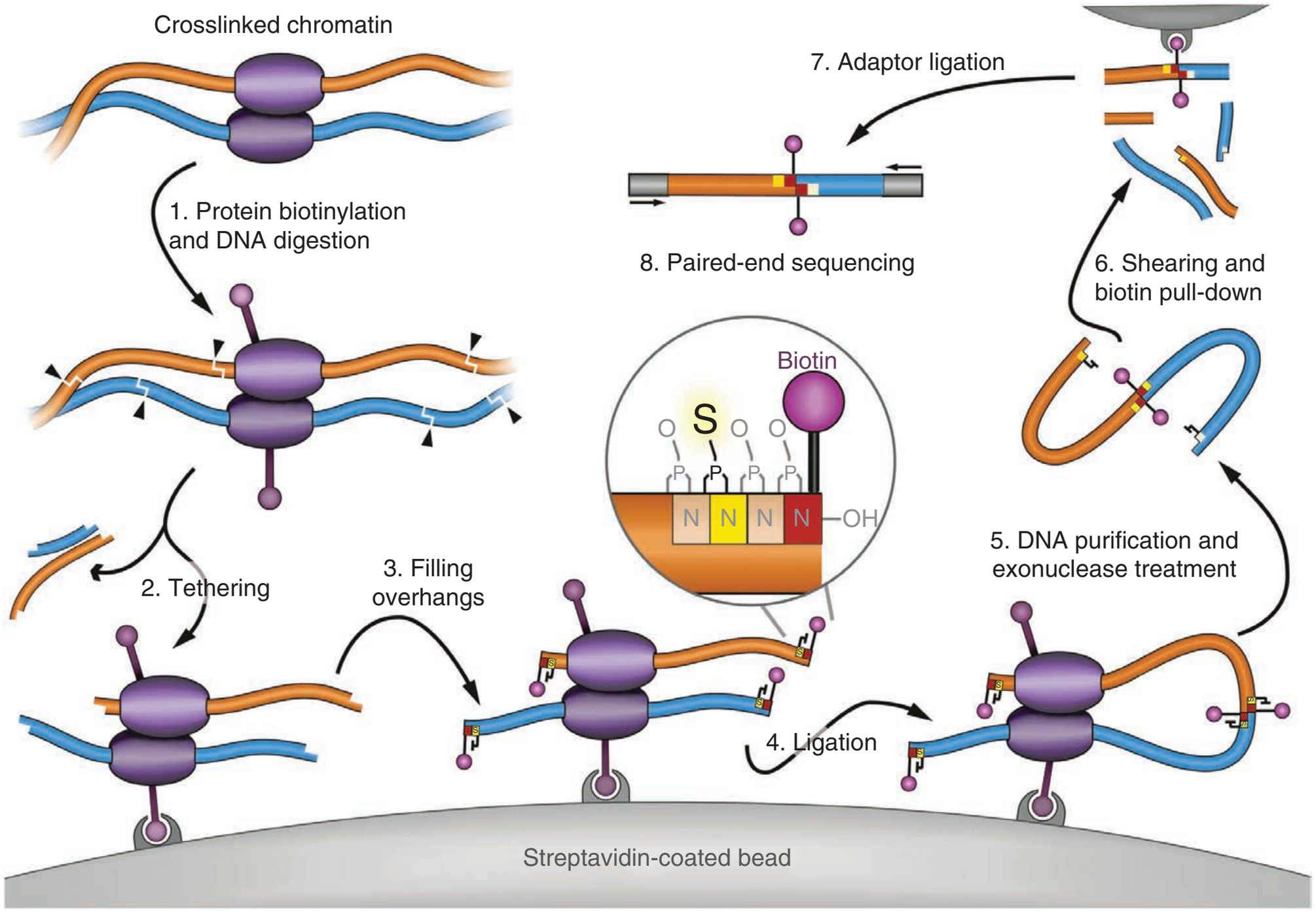
Hi-C and derivatives

A

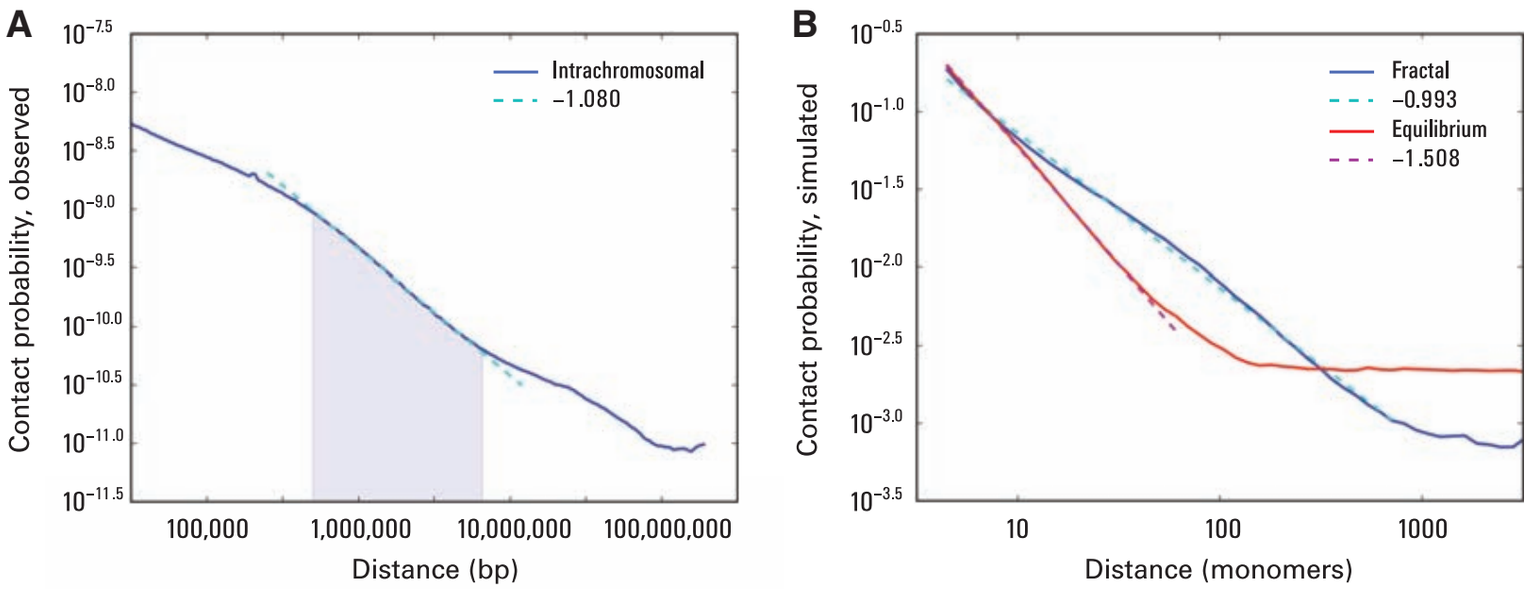


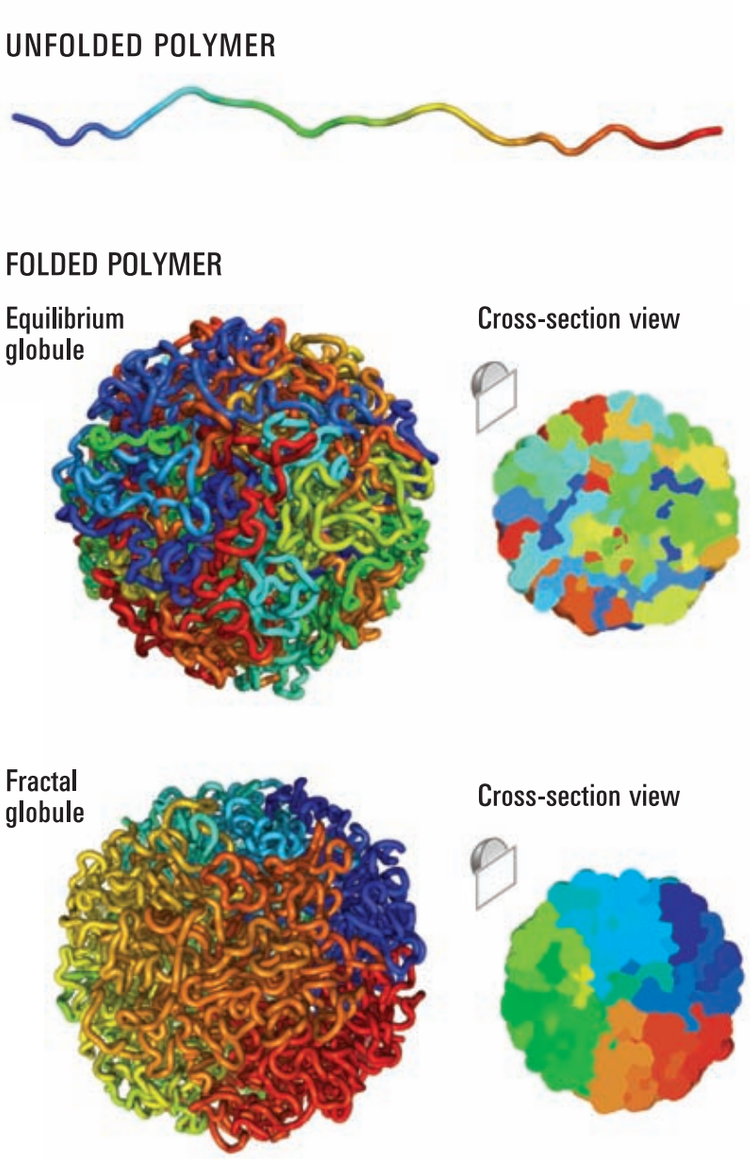
Lieberman-Aiden 2009

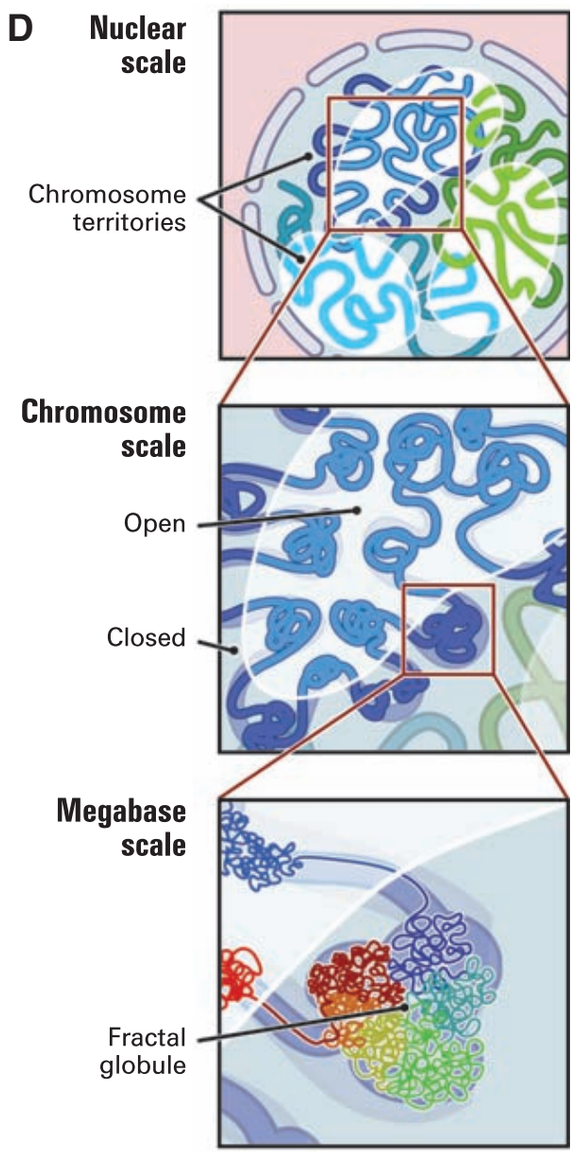
TCC - tethered chromatin conformation capture



Kalhor 2012







Hi-C methods - what can we learn from them?

Polymer biophysics

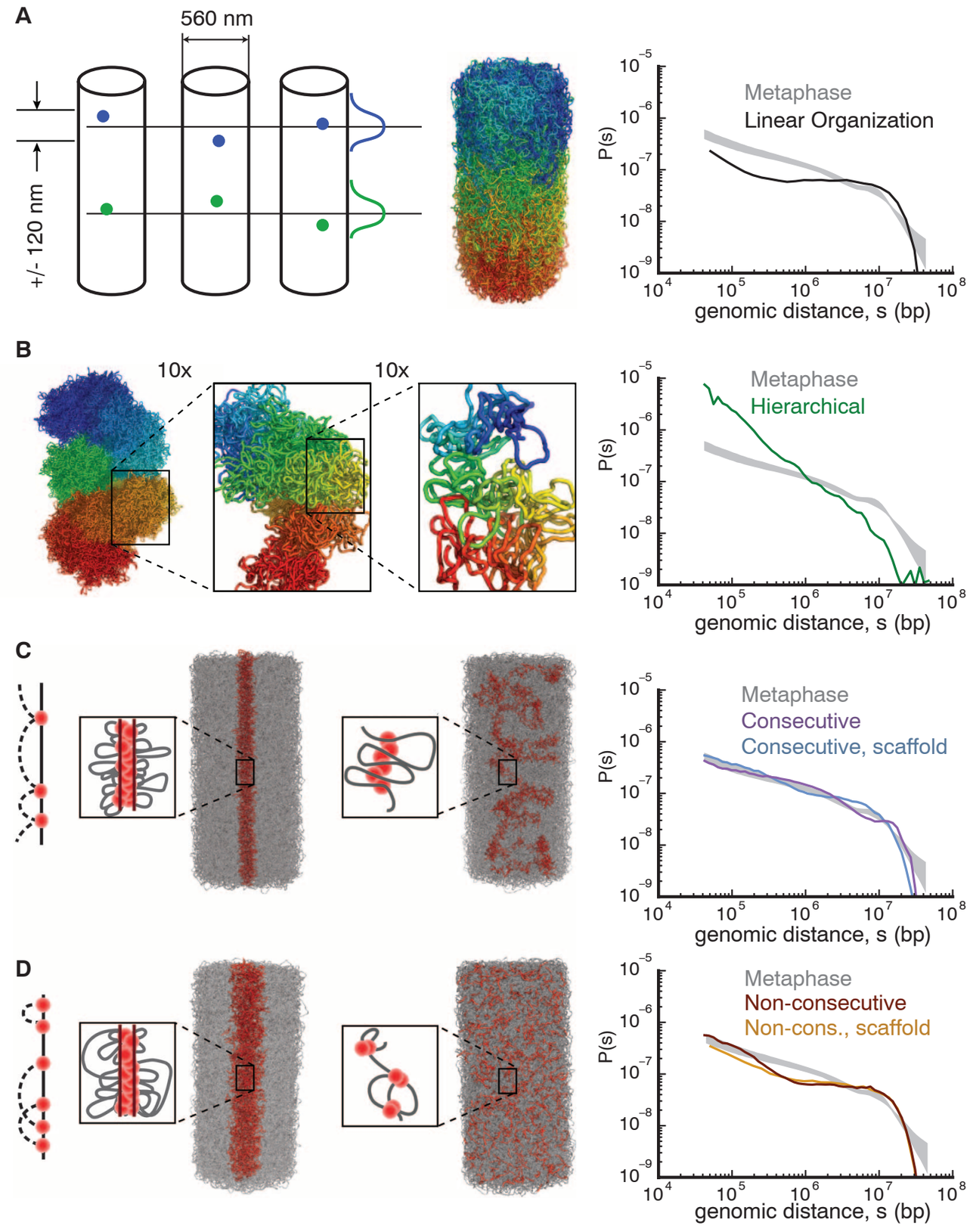
General chromatin structure in the interphase

Lieberman-Aiden 2009

Hi-C methods - what can we learn from them?

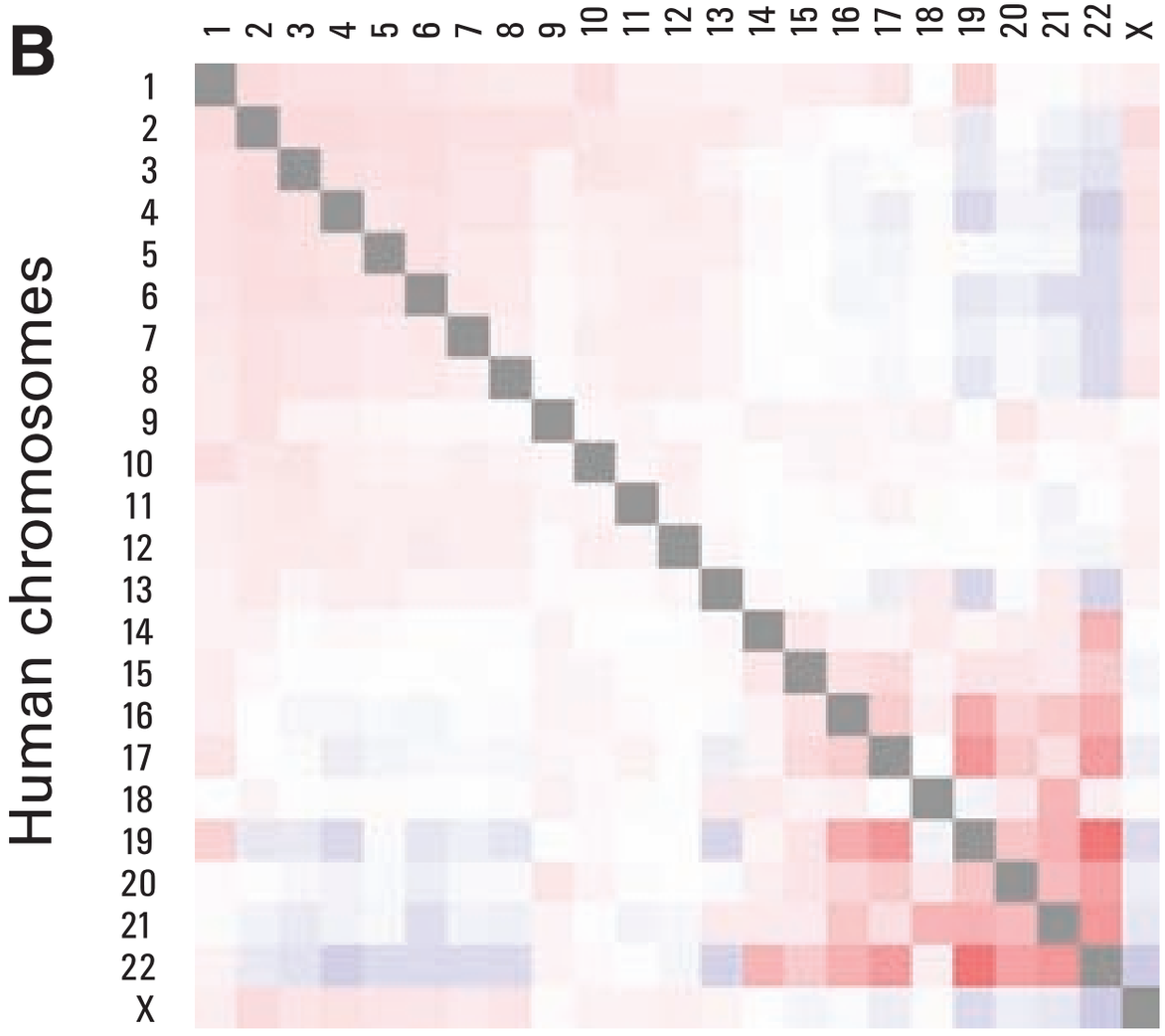
Polymer biophysics

The structure of metaphase chromosomes



Naumova 2013

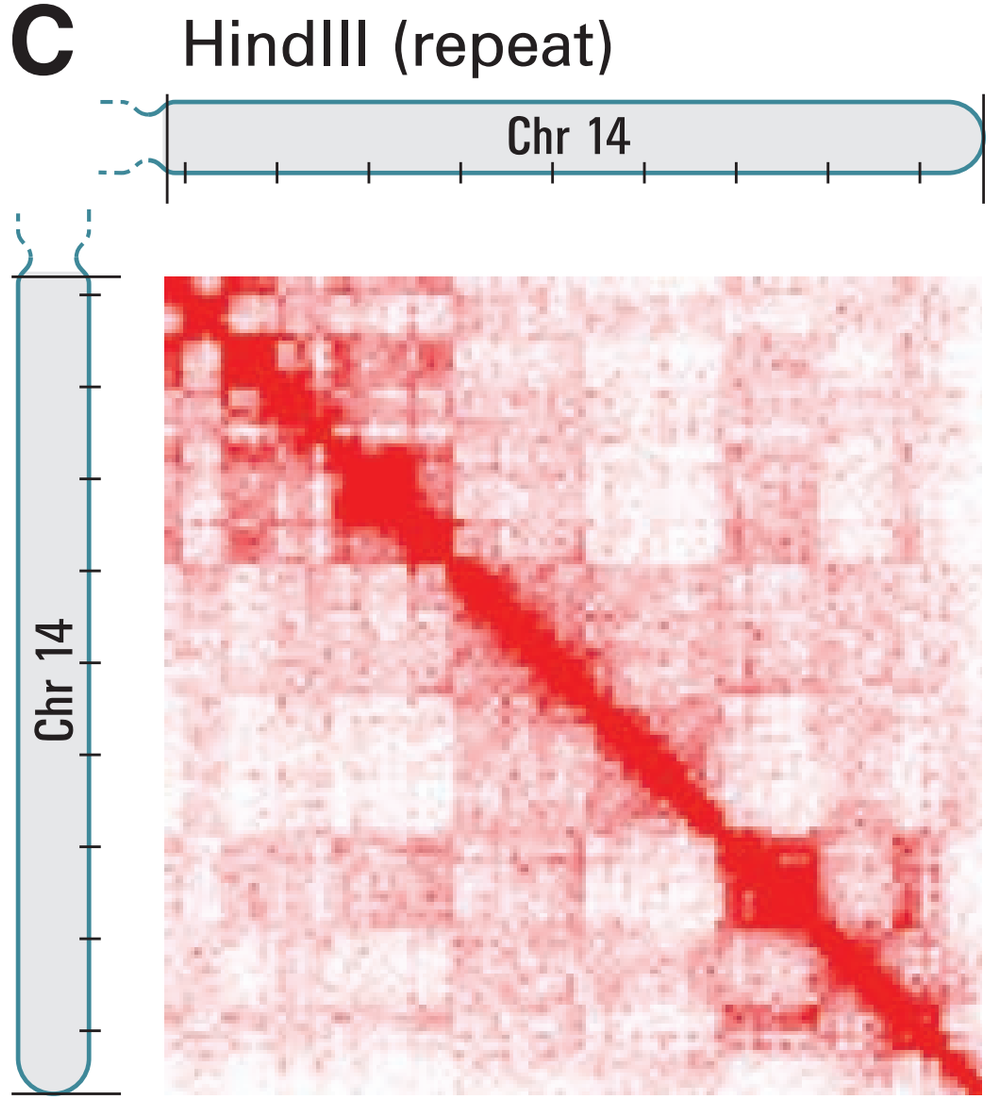
Hi-C methods - what can we learn from them?



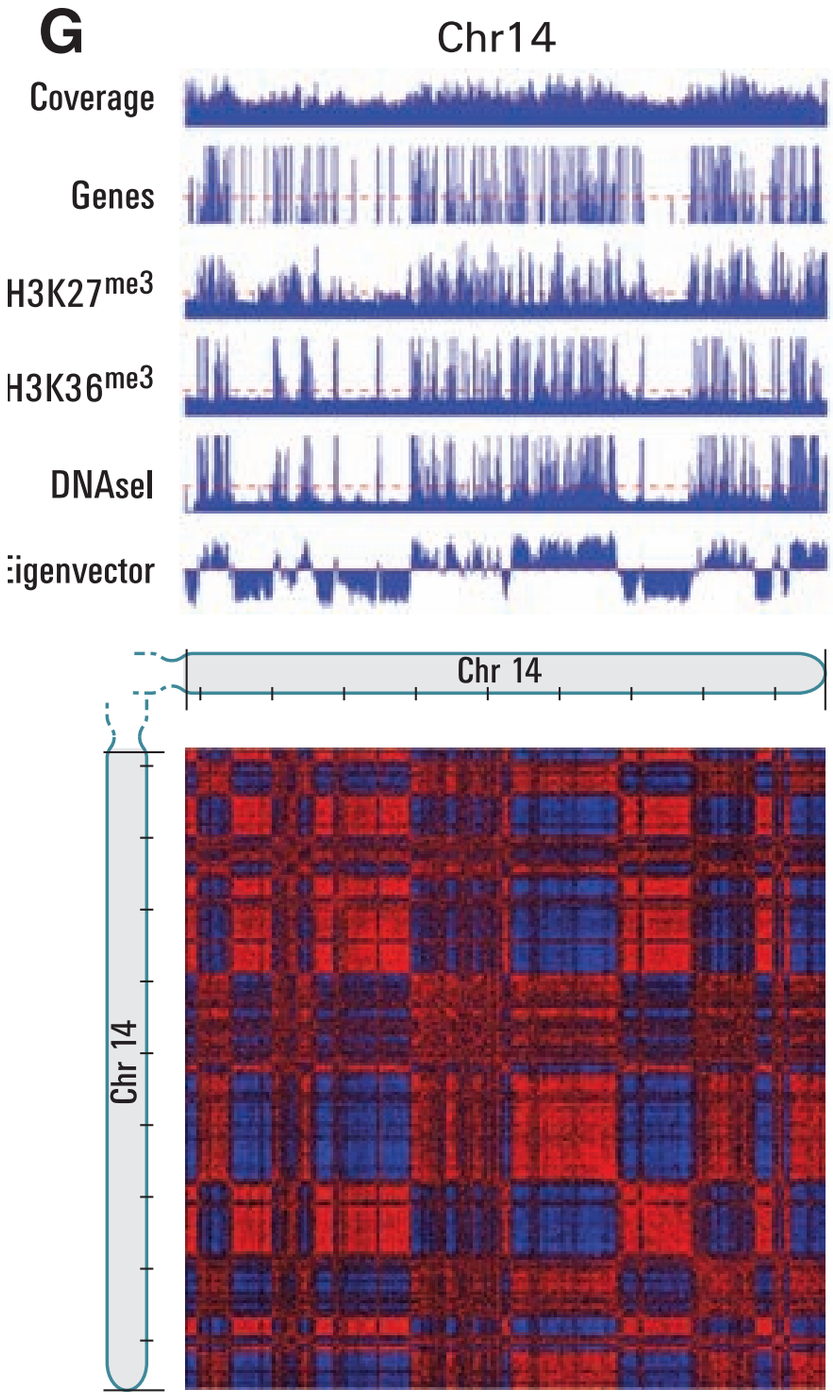
Lieberman-Aiden 2009

Cross-linking frequencies between sequences genome-wide

Often referred to as ‘interactions’

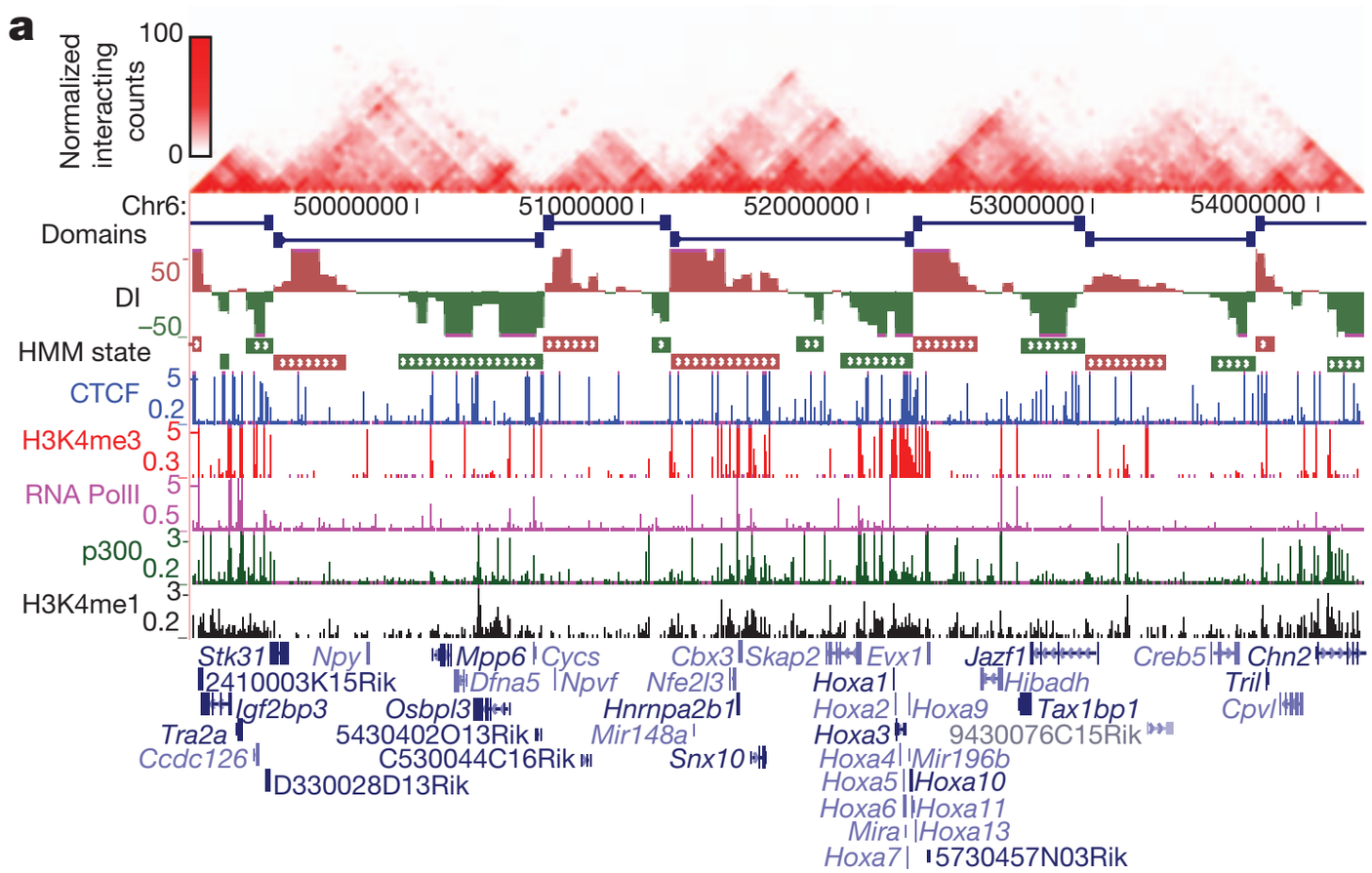


Hi-C methods - what can we learn from them?



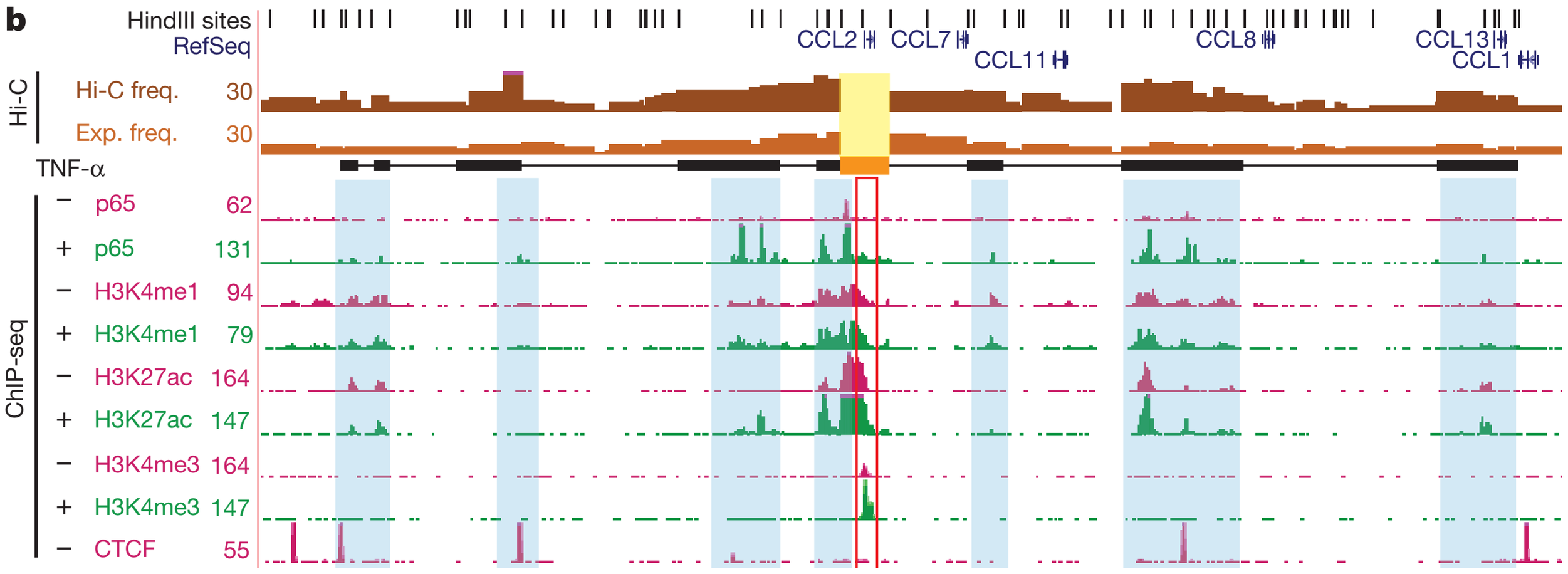
Lieberman-Aiden 2009

Chromosome compartmentalization



Dixon 2012

Hi-C methods - what can we learn from them?

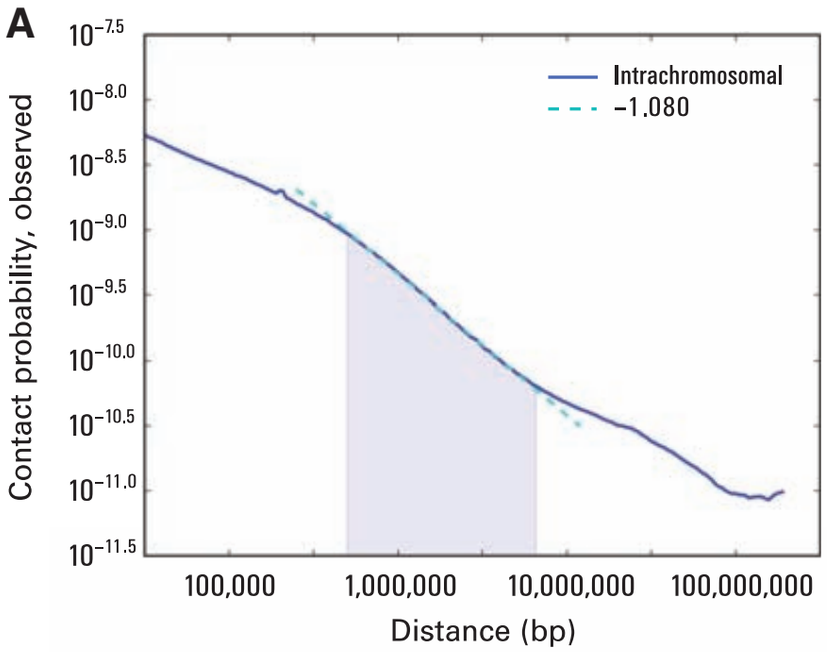


Jin 2013

Cis-regulatory element interactions

Study design

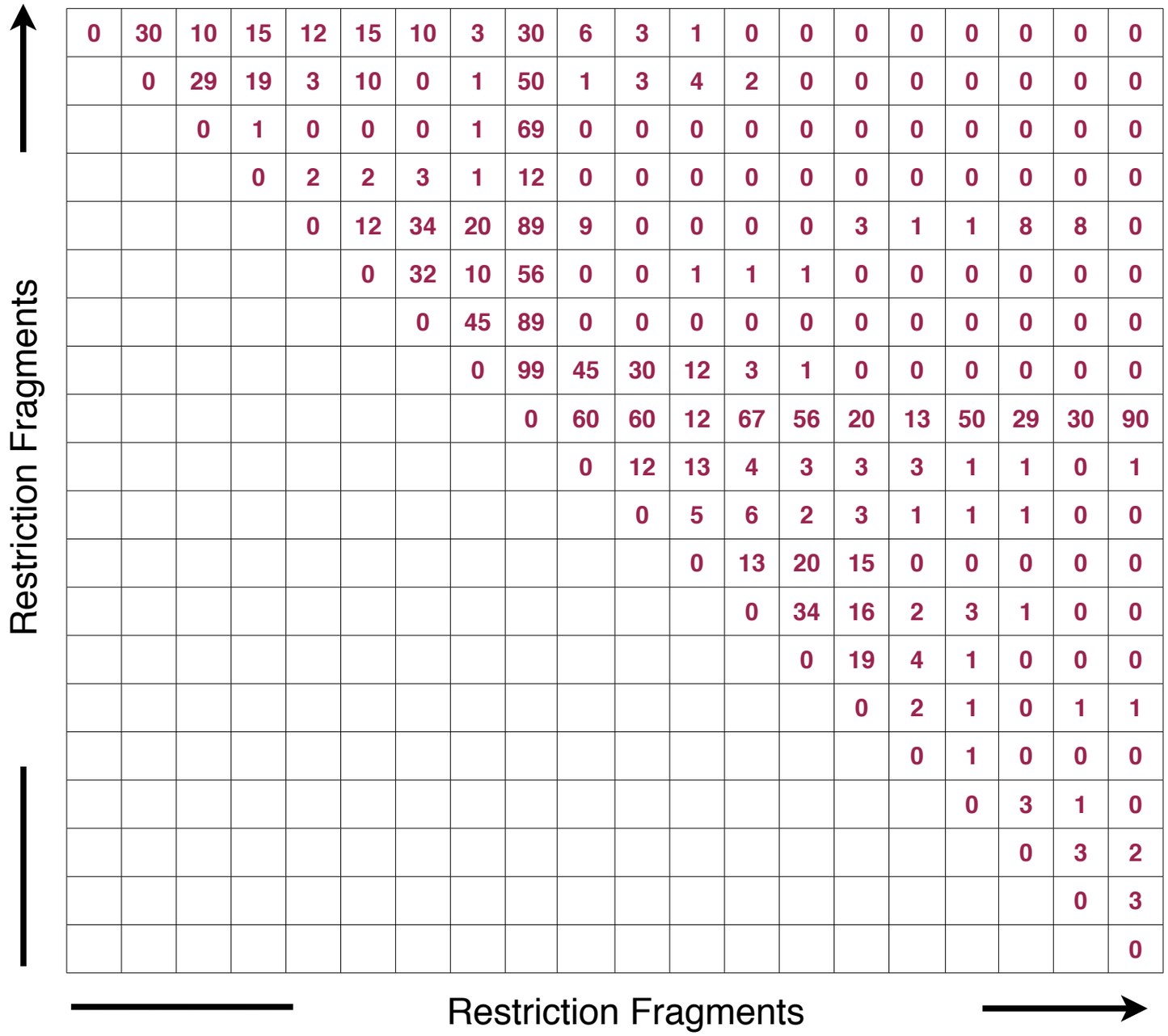
Lieberman-Aiden 2009



Signal declines very quickly with increasing genomic distance

Count noise...

Depending on the question we ask we would need appropriate sequencing depth



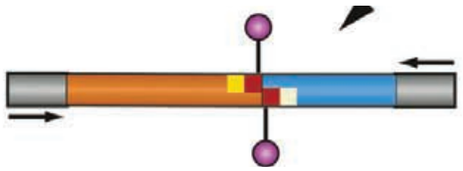
0	30	10	15	12	15	10	3	30	6	3	1	0	0	0	0	0	0	0	0
	0	29	19	3	10	0	1	50	1	3	4	2	0	0	0	0	0	0	0
		0	1	0	0	0	1	69	0	0	0	0	0	0	0	0	0	0	0
			0	2	2	3	1	12	0	0	0	0	0	0	0	0	0	0	0
				0	12	34	20	89	9	0	0	0	0	3	1	1	8	8	0
					0	32	10	56	0	0	1	1	1	0	0	0	0	0	0
						0	45	89	0	0	0	0	0	0	0	0	0	0	0
							0	99	45	30	12	3	1	0	0	0	0	0	0
								0	60	60	12	67	56	20	13	50	29	30	90
									0	12	13	4	3	3	3	1	1	0	1
										0	5	6	2	3	1	1	1	0	0
											0	13	20	15	0	0	0	0	0
												0	34	16	2	3	1	0	0
													0	19	4	1	0	0	0
														0	2	1	0	1	1
															0	1	0	0	0
																0	3	1	0
																	0	3	2
																		0	3
																			0

Study design - sequencing depth
‘personal observations’

1 Mb resolution, mammalian genome 1 lane of Hi-Seq per replicate should allow for comparative analysis of inter-chromosomal interactions (yield ~ 70M usable reads)

The same sequencing depth should allow for attempts in comparative analysis at 10 kb bin level (including ‘local’ interactions only - up to 1Mb)

Analysis workflow



Trim at ligation site

Ready quality check

Alignment
Both mates separetly

Filtering

- Good alignment
- Mapping to RF* within the correct distance (sonication)
- Mapping not too close to each other
- Remove reads from the same RF
- Read orientation filter (Jin 2013)
- Remove PCR duplicates

Normalization

Interaction calling

TAD calling

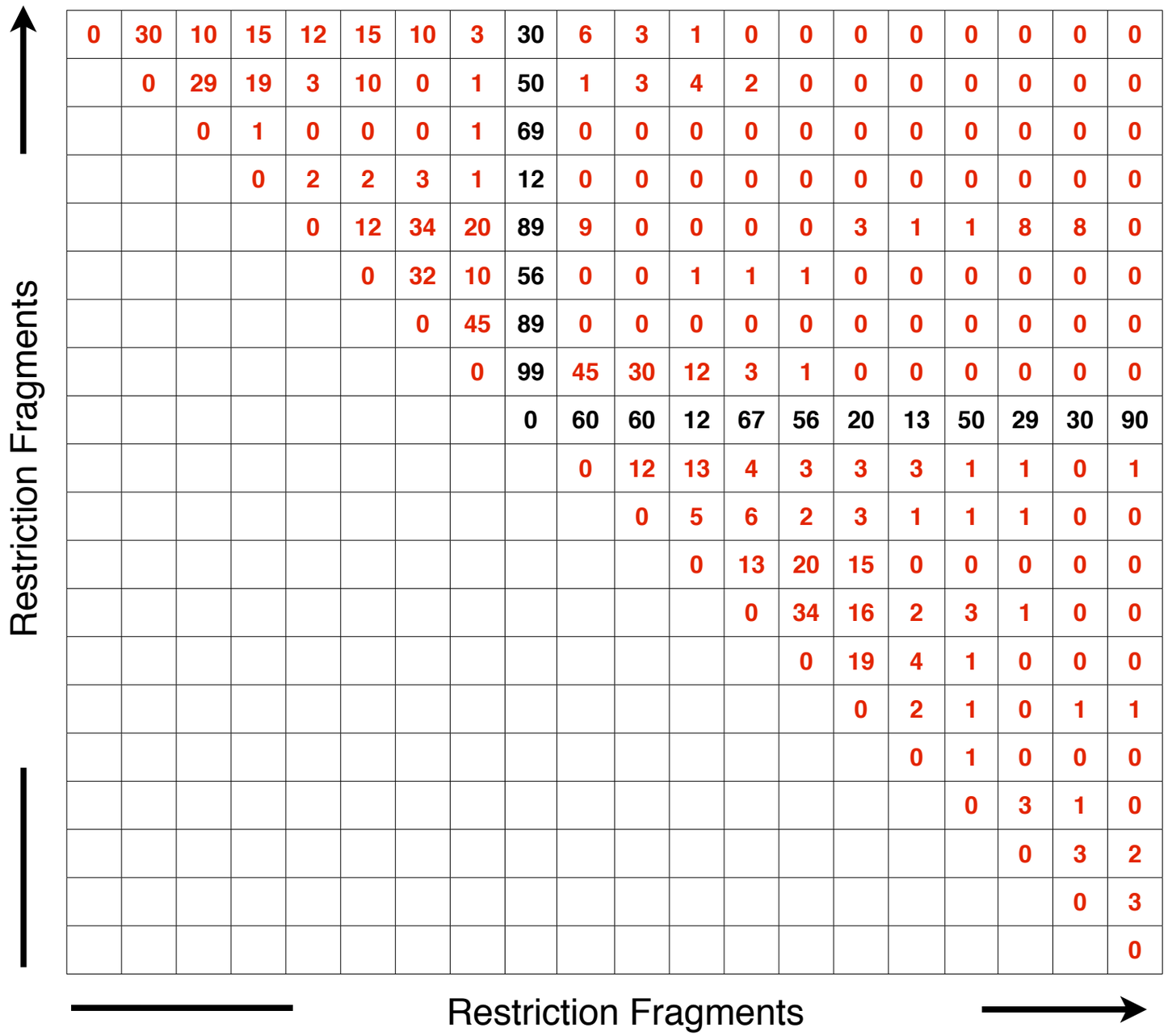
Compartment

Inter-chromosomal contacts

RF - restriction fragment

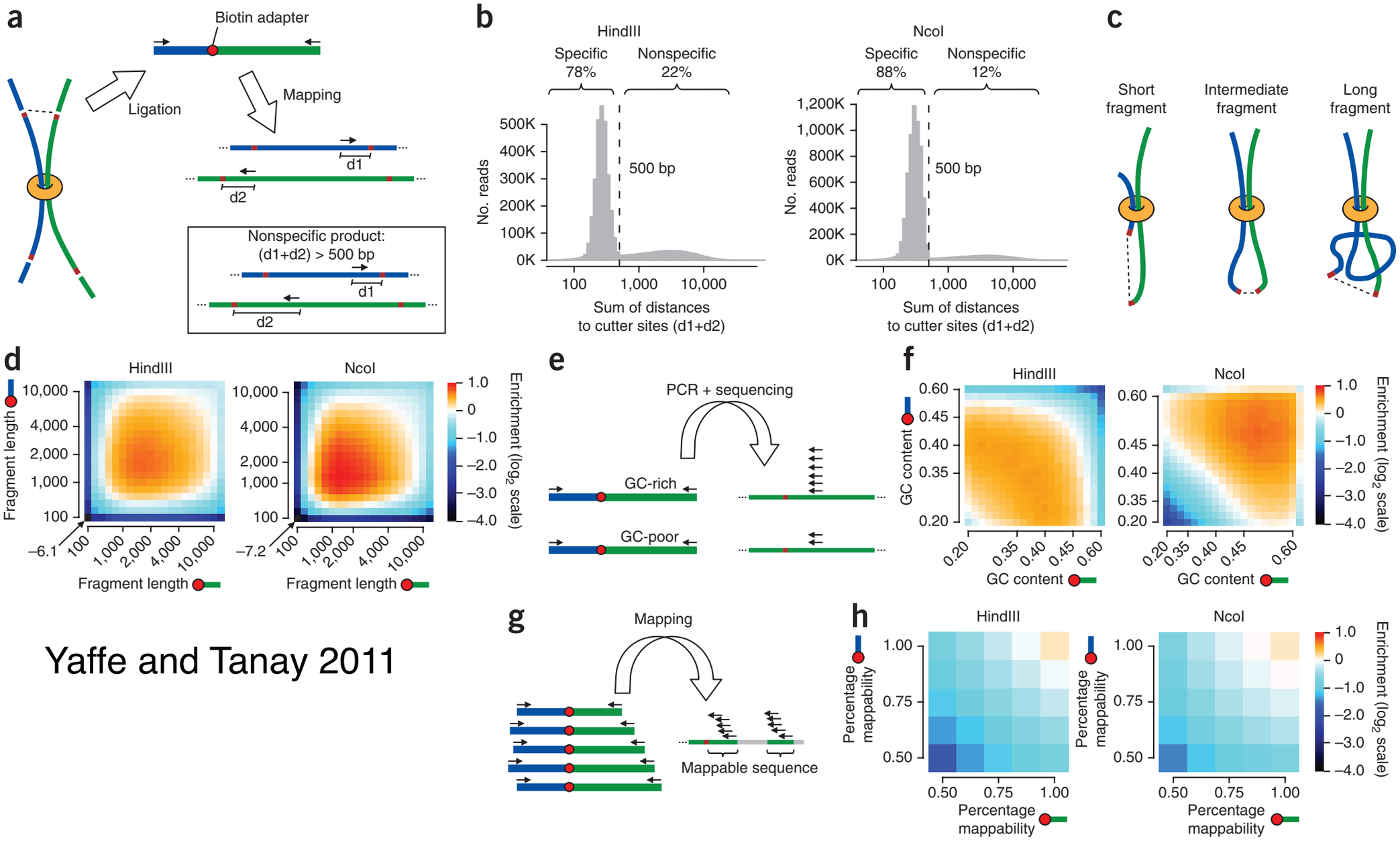
We can (should) perform normalization

Excellent news: we are genome-wide in this assay!



0	30	10	15	12	15	10	3	30	6	3	1	0	0	0	0	0	0	0	0
	0	29	19	3	10	0	1	50	1	3	4	2	0	0	0	0	0	0	0
		0	1	0	0	0	1	69	0	0	0	0	0	0	0	0	0	0	0
			0	2	2	3	1	12	0	0	0	0	0	0	0	0	0	0	0
				0	12	34	20	89	9	0	0	0	0	3	1	1	8	8	0
					0	32	10	56	0	0	1	1	1	0	0	0	0	0	0
						0	45	89	0	0	0	0	0	0	0	0	0	0	0
							0	99	45	30	12	3	1	0	0	0	0	0	0
								0	60	60	12	67	56	20	13	50	29	30	90
									0	12	13	4	3	3	3	1	1	0	1
										0	5	6	2	3	1	1	1	0	0
											0	13	20	15	0	0	0	0	0
												0	34	16	2	3	1	0	0
													0	19	4	1	0	0	0
														0	2	1	0	1	1
															0	1	0	0	0
																0	3	1	0
																	0	3	2
																		0	3
																			0

1st approach



Yaffe and Tanay 2011

1. Identify sources of biases: RF length, mapability and CG content
2. Normalize

2nd approach - first step

1. Do not try to identify sources of biases but learn their effect from data (coverage)
2. Normalize for the coverage

number of reads between segments *i* and *j*

$$f_{ij} = \frac{c_{ij}\left(\sum_{k=1}^{K\text{-}1}\sum_{l=k+1}^{K} c_{kl}\right)}{\left(\sum_{k=1}^{K} c_{ik}\right)\left(\sum_{k=1}^{K} c_{kj}\right)}$$

Total number of reads

normalized ligation frequency between segments *i* and *j*

Total number of for segment *i*

Total number of for segment *j*

Lieberman-Aiden 2009
Kalhor 2012

2nd approach ‘ICE’ - complete

1. Do not try to identify sources of biases but learn their effect from data (coverage)
2. Normalize for the coverage in an iterative fashion

Biases

Observed $O_{ij} = B_i B_j T_{ij}$ True

‘relative contact probabilities’ $\sum_{i=1,|i-j|>1}^{N} T_{ij} = 1$

* diagonal and 1st off-diagonal are removed
additional filtering required

Imakaev 2012

How does it work algorithmically?

“We start by creating a working copy of the matrix Oij, denoted **W*ij*** as the iterative process gradually changes this matrix to T*ij*.

We initialize the iterative procedure by setting each element of the vector of total biases B to 1. We begin each iteration by calculating the coverage

$$S_i = \Sigma_j W_{ij}$$

Next, additional biases ΔBi are calculated by renormalizing Si to have the unit mean

$$\Delta B_i = S_i/\text{mean}(S_i).$$

We then

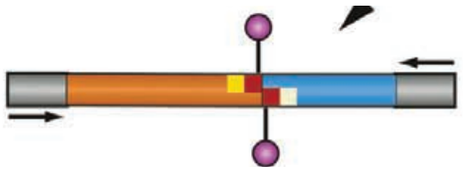
$$W_{ij} / \Delta B_i \Delta B_j \text{ for all } (i,j)$$

and update the total vector of biases by multiplying by the additional biases.

Iterations are repeated until the variance of the additional biases becomes negligible “

Imakaev 2012

Analysis workflow



Trim at ligation site

Ready quality check

Alignment
Both mates separetly

Filtering

Good alignment

Mapping to RF* within the correct distance

Mapping not too close to each other

Remove reads from the same RF

Read orientation filter (Jin 2013)

Remove PCR duplicates

Normalization

Interaction calling

TAD calling

Compartment

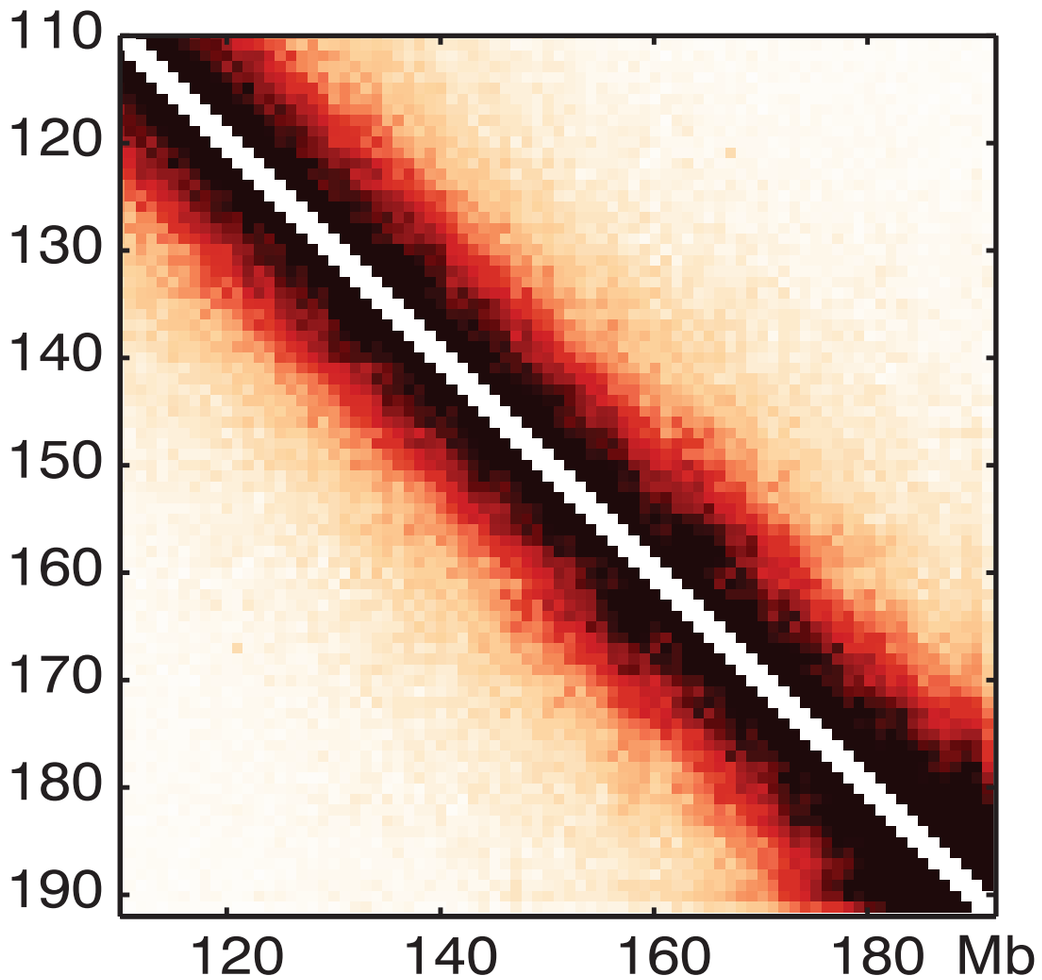
Inter-chromosomal contacts

RF - restriction fragment

Isolation of interactions

Random collisions

Uniform - visualization

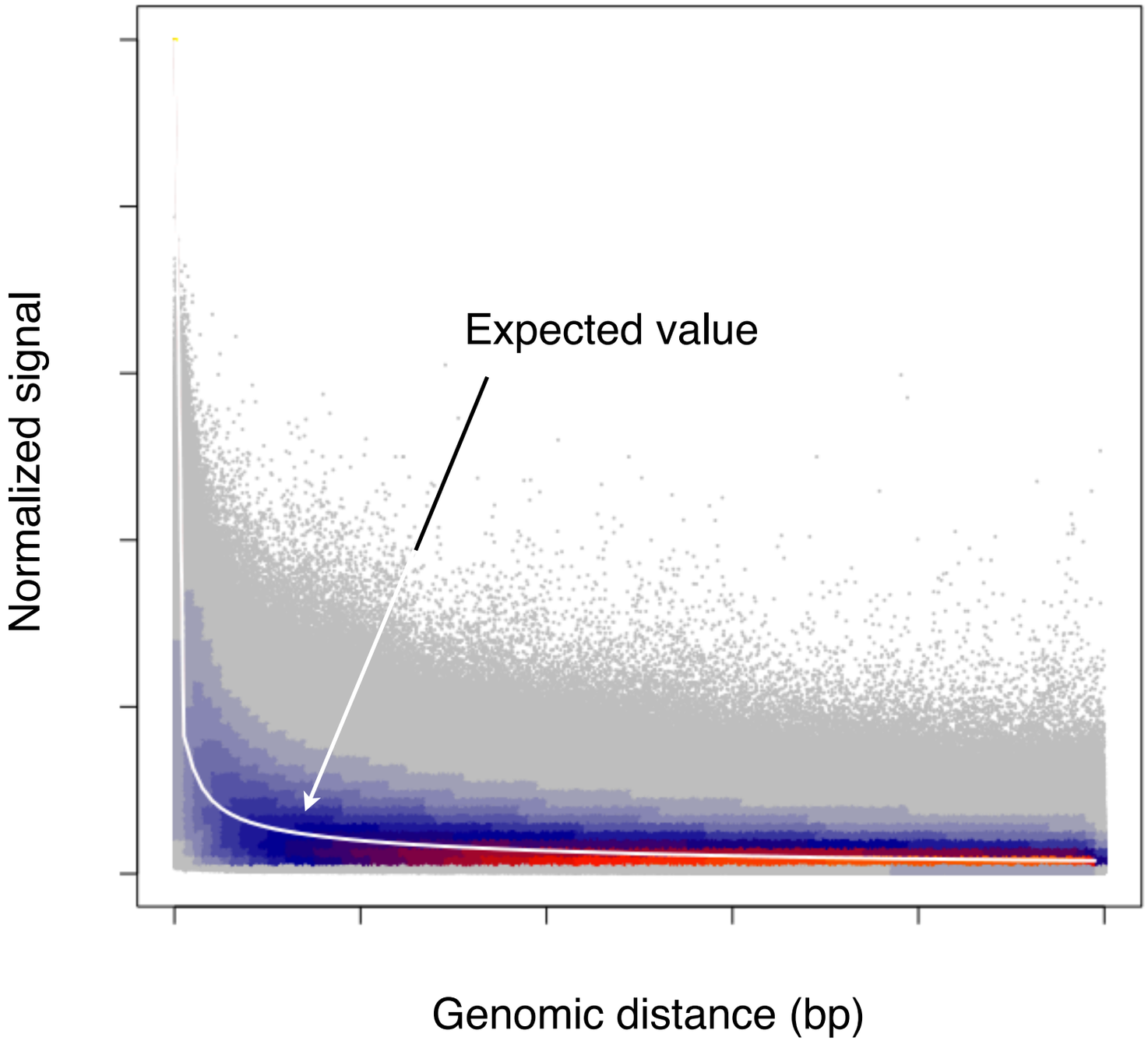


Naumova 2013

Interactions occurring just because chromatin is a biopolymer and folds

Isolation of interactions

Random collisions - expected interaction strength at a particular distance



Isolation of interactions

LETTER

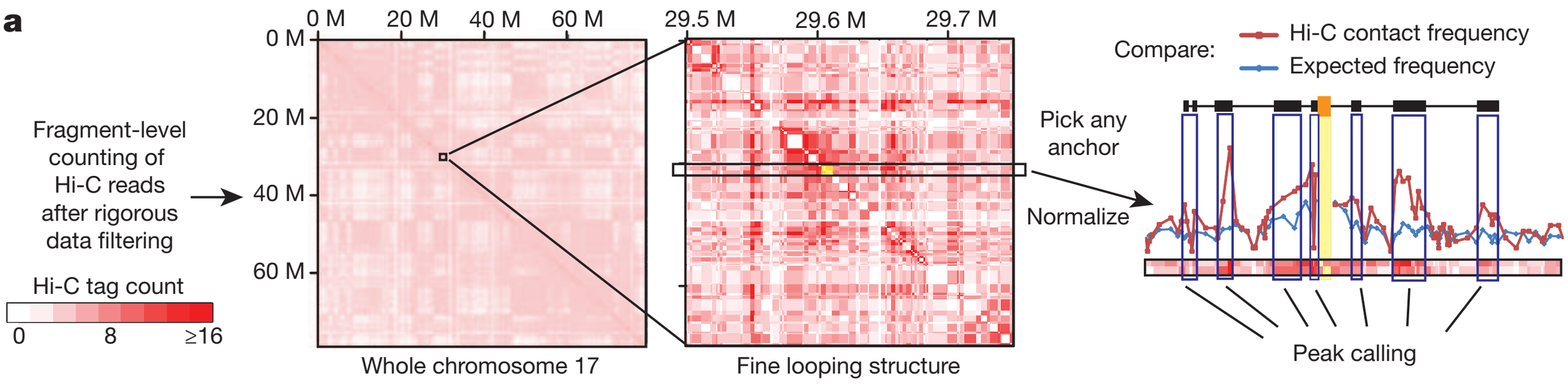
doi:10.1038/nature12644

A high-resolution map of the three-dimensional chromatin interactome in human cells

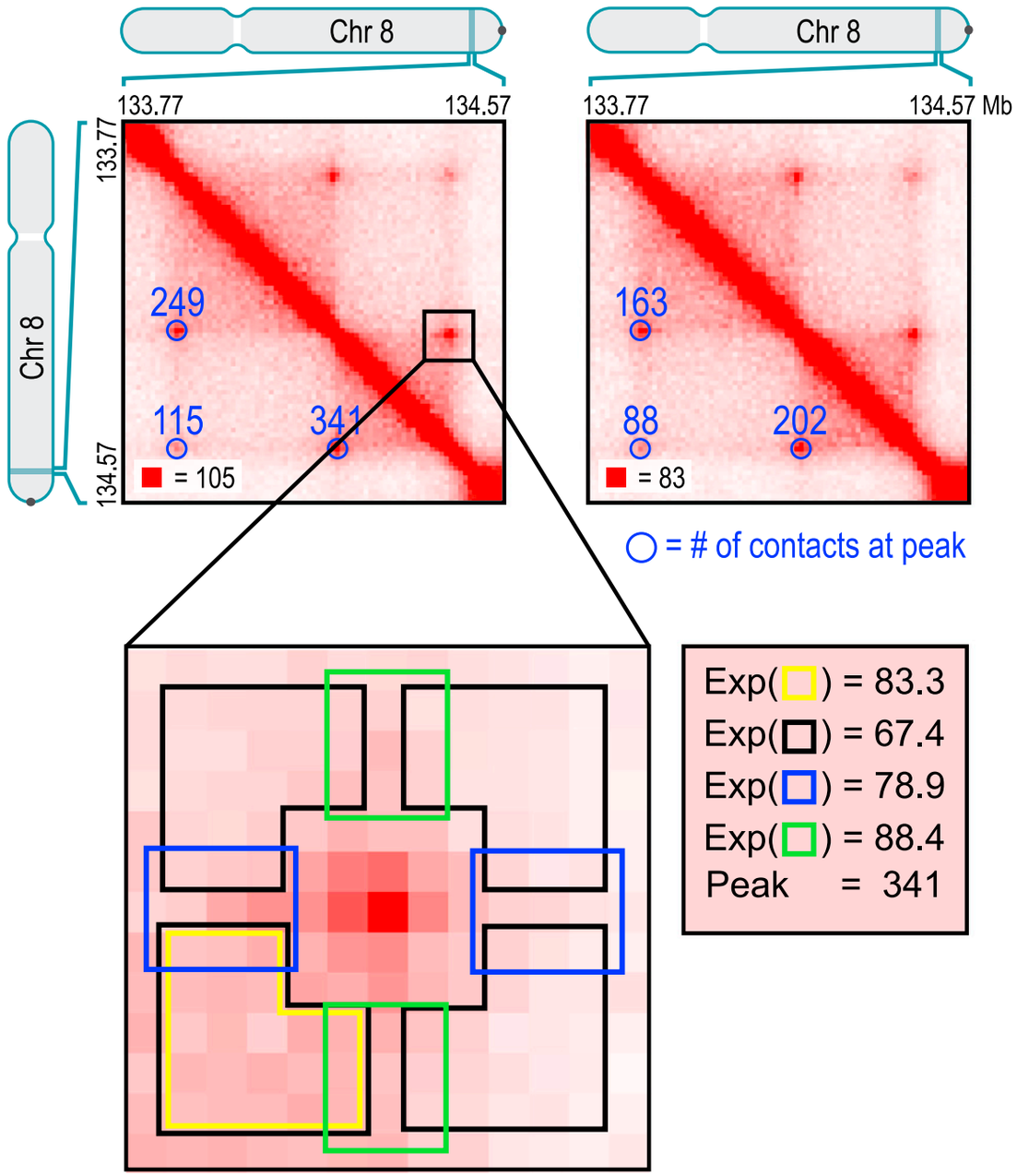
Fulai Jin[1]*, Yan Li[1]*, Jesse R. Dixon[1,2], Siddarth Selvaraj[1,3], Zhen Ye[1], Ah Young Lee[1], Chia-An Yen[1], Anthony D. Schmitt[1,4], Celso A. Espinoza[1] & Bing Ren[1,5]

Use negative binomial to asses for each interaction whether its strength is unexpectedly high given the:

- biases
- distance
- additional signal strength threshold



Isolation of interactions

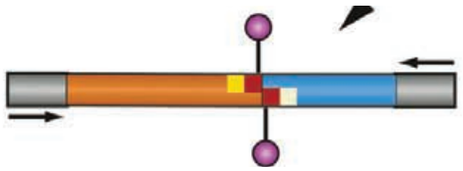


HiCCUPS

Pixels in the middle should have signal 50% higher than the surroundings.

Rao 2014

Analysis workflow



Trim at ligation site

Ready quality check

Alignment
Both mates separetly

Filtering

Good alignment

Mapping to RF* within the correct distance

Mapping not too close to each other

Remove reads from the same RF

Read orientation filter (Jin 2013)

Remove PCR duplicates

Normalization

Interaction calling

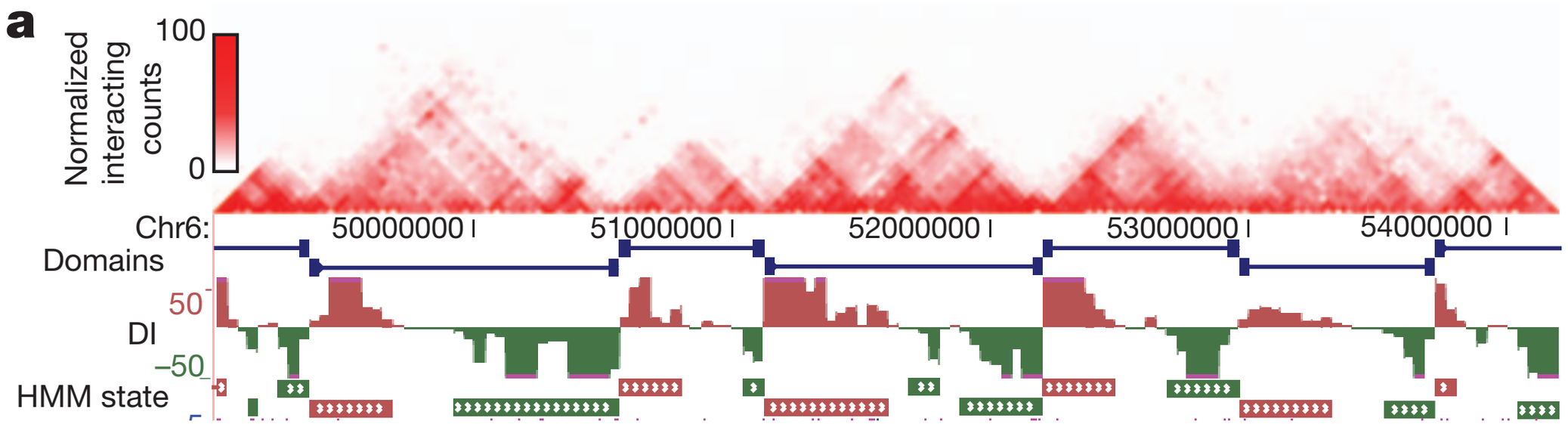
TAD calling

Compartment

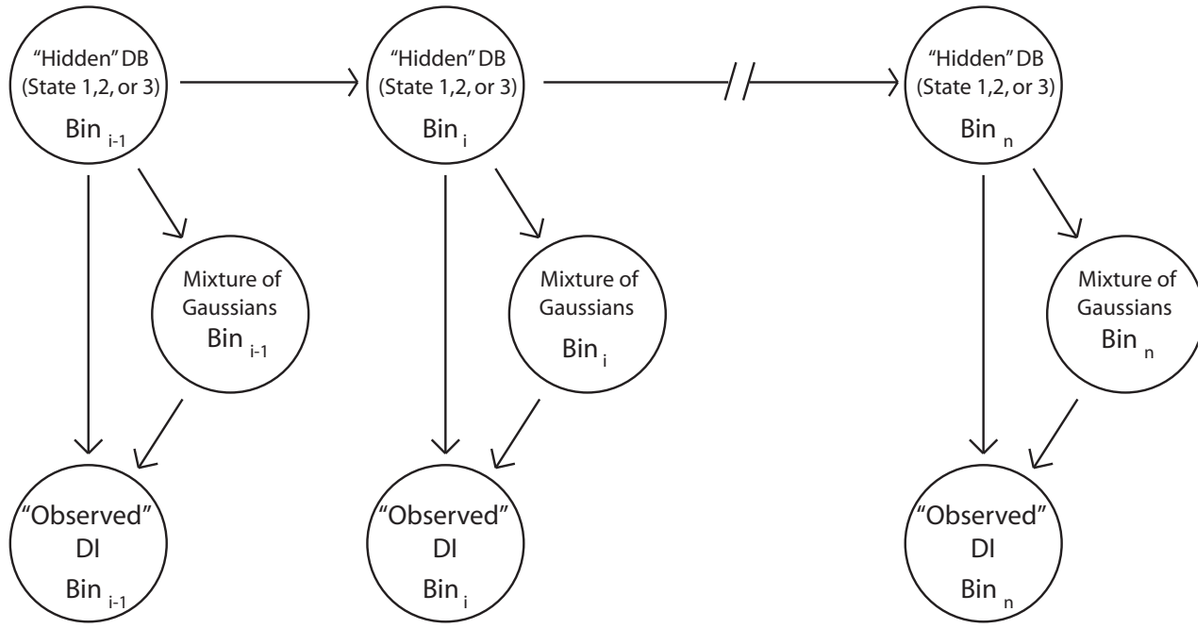
Inter-chromosomal contacts

RF - restriction fragment

Isolation of TADs - directionality index (DI)



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



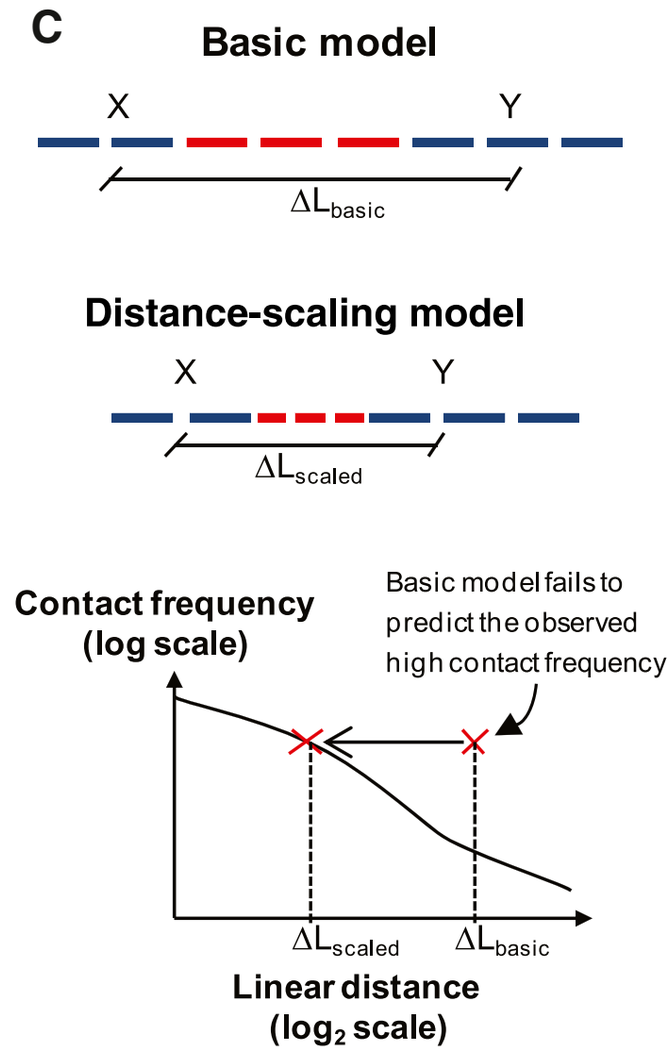
Dixon 2012

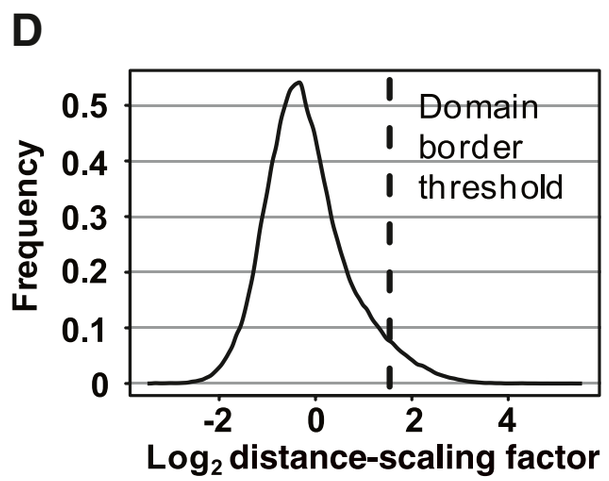
"Upstream Bias" - State 1
"Downstream Bias" - State 2
No Bias - State 3

Tom Sexton's and Amos Tanay's method

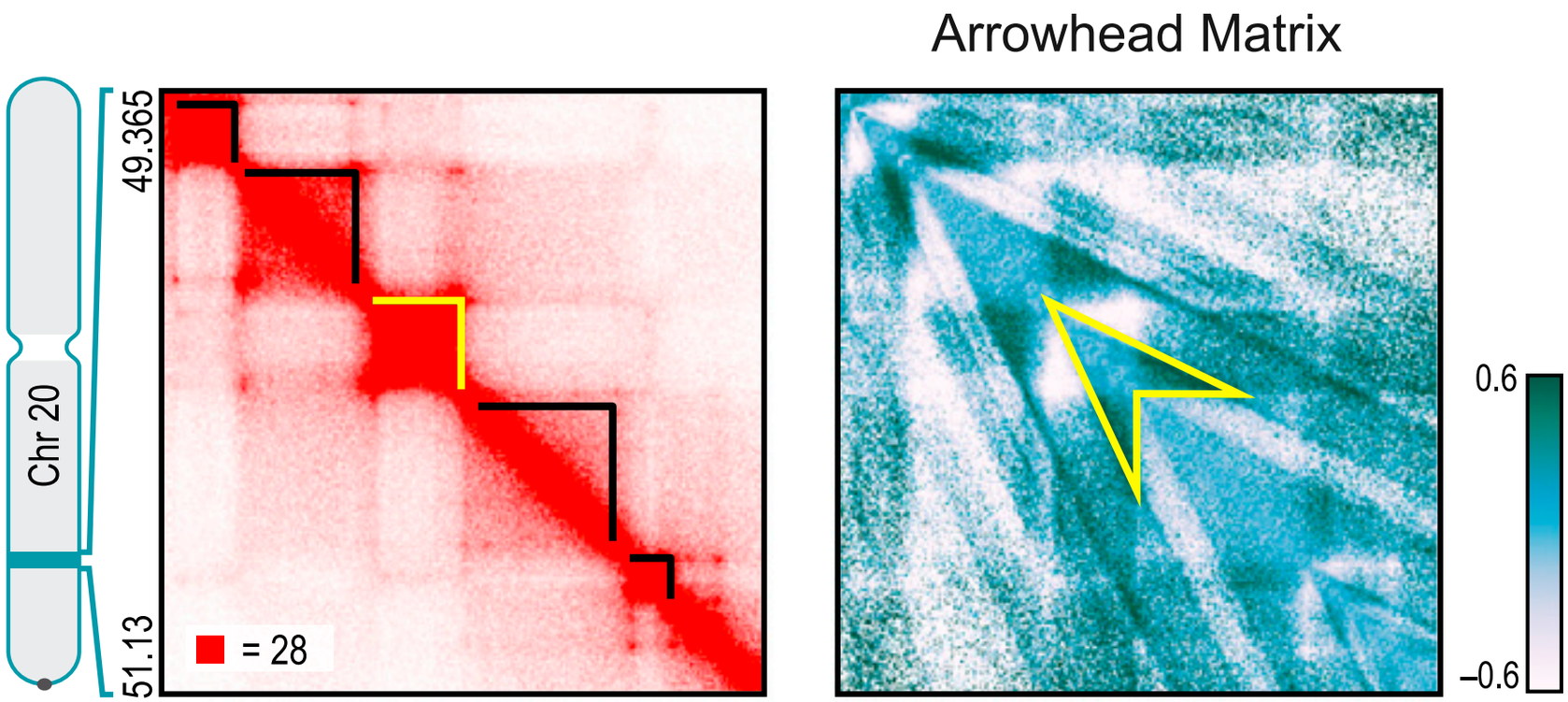
Particularly convenient for compact genomes

Here *D. melanogaster*





Isolation of TADs - other approaches

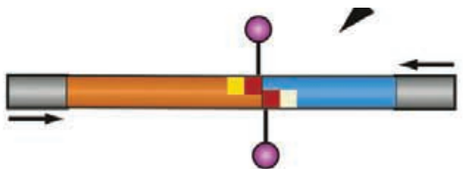


MATRIX -1x(Obs/Exp - 1)

'A "corner score" matrix, indicating each pixel's likelihood of lying at the corner of a domain, is efficiently calculated from the arrowhead matrix using dynamic programming.'

Rao 2014

Analysis workflow



Trim at ligation site

Ready quality check

Alignment
Both mates separetly

Filtering

Good alignment

Mapping to RF* within the correct distance

Mapping not too close to each other

Remove reads from the same RF

Read orientation filter (Jin 2013)

Remove PCR duplicates

Normalization

Interaction calling

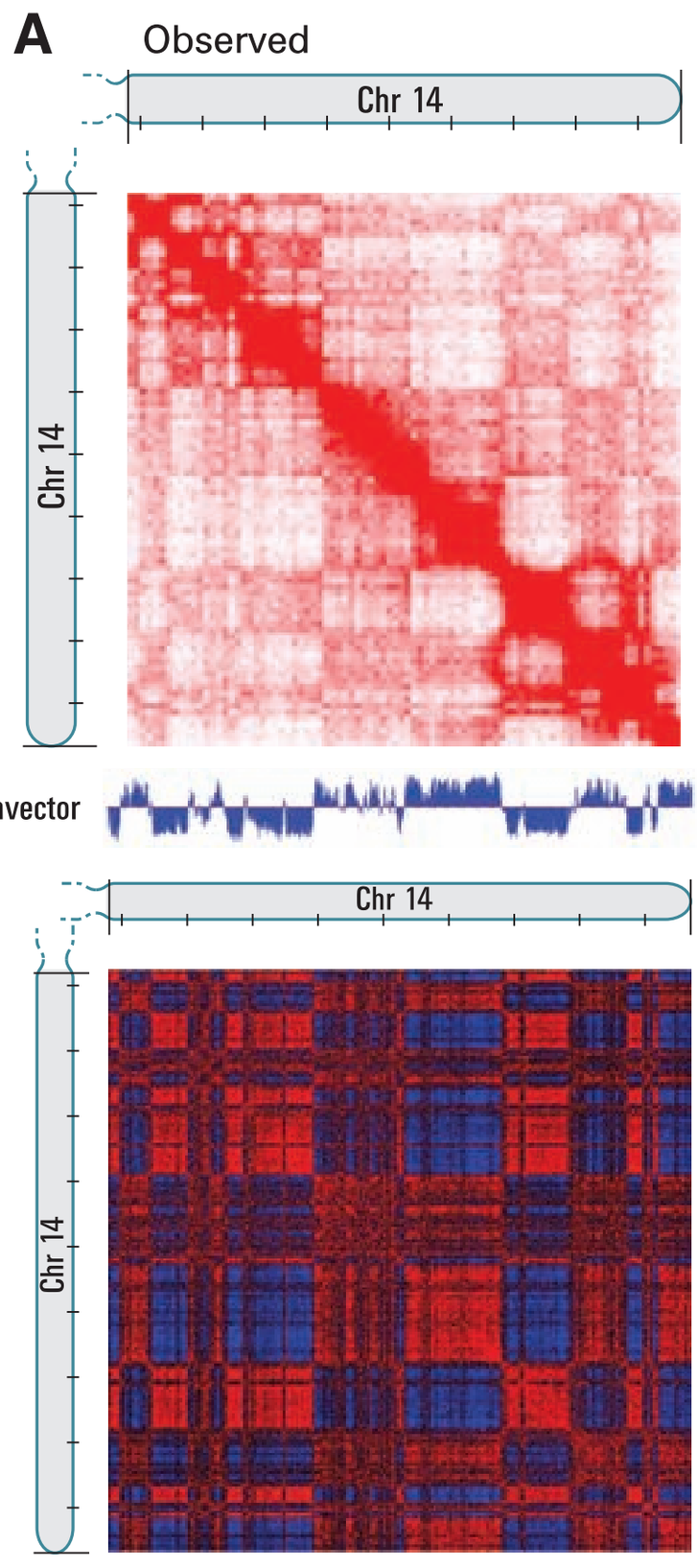
TAD calling

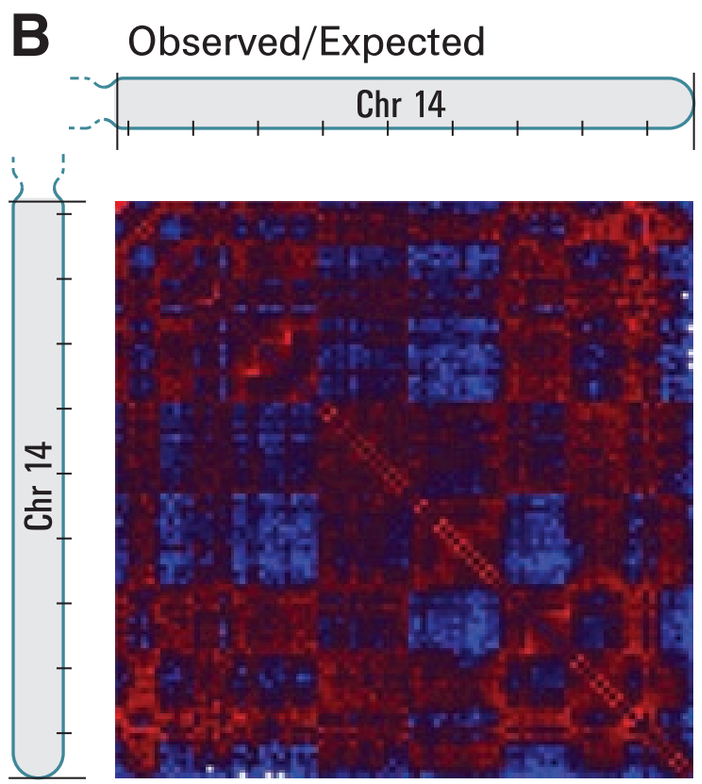
Compartment

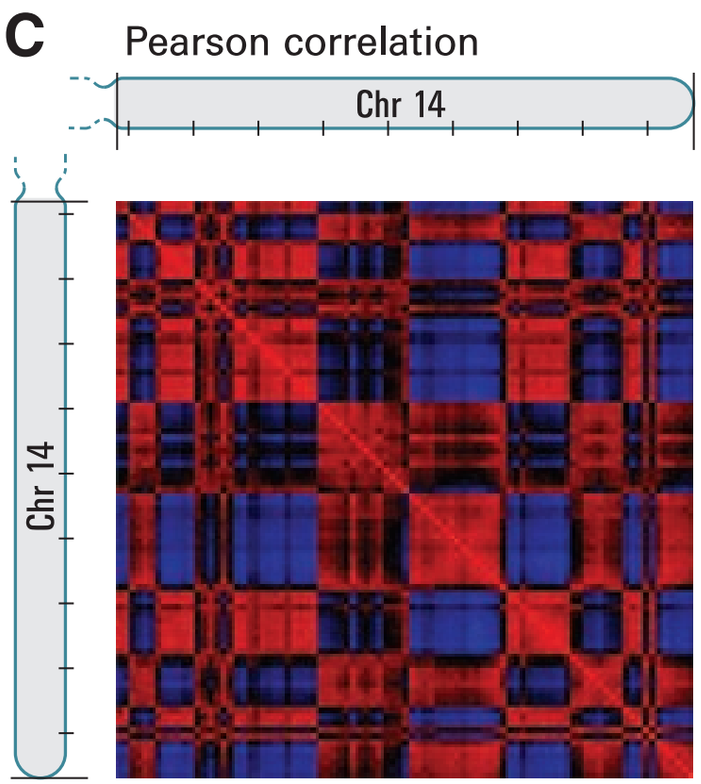
Inter-chromosomal contacts

RF - restriction fragment

The isolation of compartments







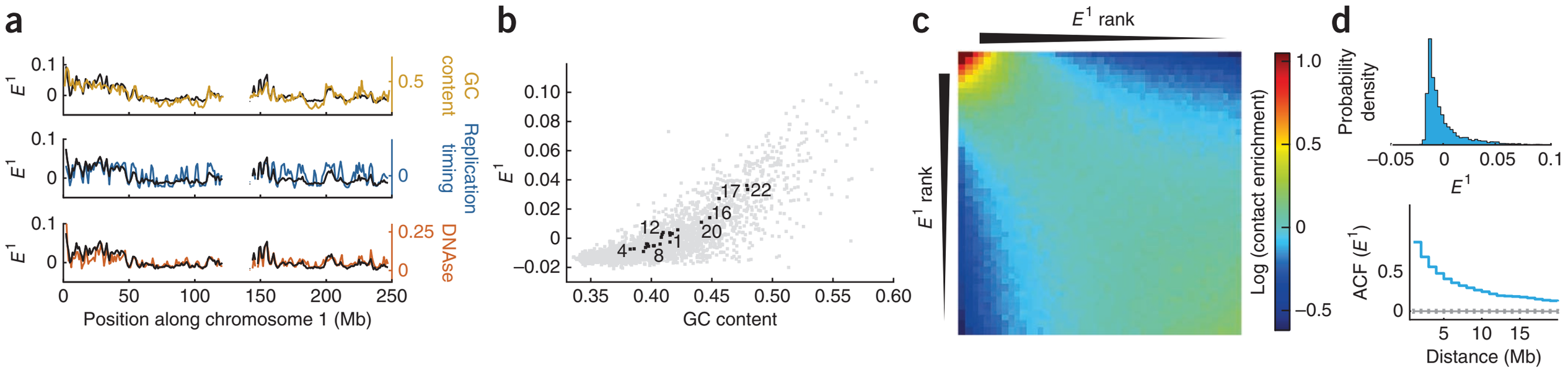
We apply eigen() function on the data

Visual inspection - which eigen vector corresponds best

sign: choice between A and B annotation is based on the overall expression

Lieberman-Aiden 2009

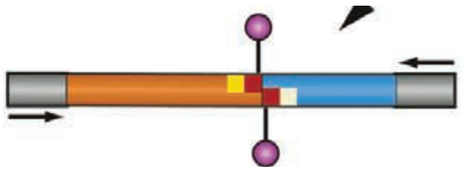
Inter-chromosomal interactions and compartments



Taking into account only inter-chromosomal contacts reveals a continuum in *E*

Imakaev 2012

Analysis workflow



Trim at ligation site

Ready quality check

Alignment
Both mates separetly

Filtering

Good alignment

Mapping to RF* within the correct distance

Mapping not too close to each other

Remove reads from the same RF

Read orientation filter (Jin 2013)

Remove PCR duplicates

Normalization

Interaction calling

TAD calling

Compartment

Inter-chromosomal contacts

RF - restriction fragment

Tools

HiTC - bioconductor package for Hi-C/5C data exploration, quality checks, binning, fitting, visualization

Our tool - Bioconductor package in preparation:
- binning/not
- normalization - ICE and other proportional fitting algorithms (convergence)
- TAD calling
- interaction calling
- compartment analysis
- visualization

Normalized
Interaction
strength

0.025
0.020
0.015
0.010
0.005
0.000

RNA-Seq
H3K4me1
H3K4me2
H3K4me3
H3K27me3
H3K27ac
PolII
CTCF
Smc1

I06Rik
Fam120b
Tbp
Chd1
Tcte3
Dll1
Pdcd2
Rgmb
25P20Rik
Psmb1
2210404J11Rik
Prdm9
Gm3417
Phf10
Gm3448
Gm3435

15.1 mb
15.3 mb
15.5 mb
15.7 mb
15.9 mb
15.2 mb
15.4 mb
15.6 mb
15.8 mb

Primer

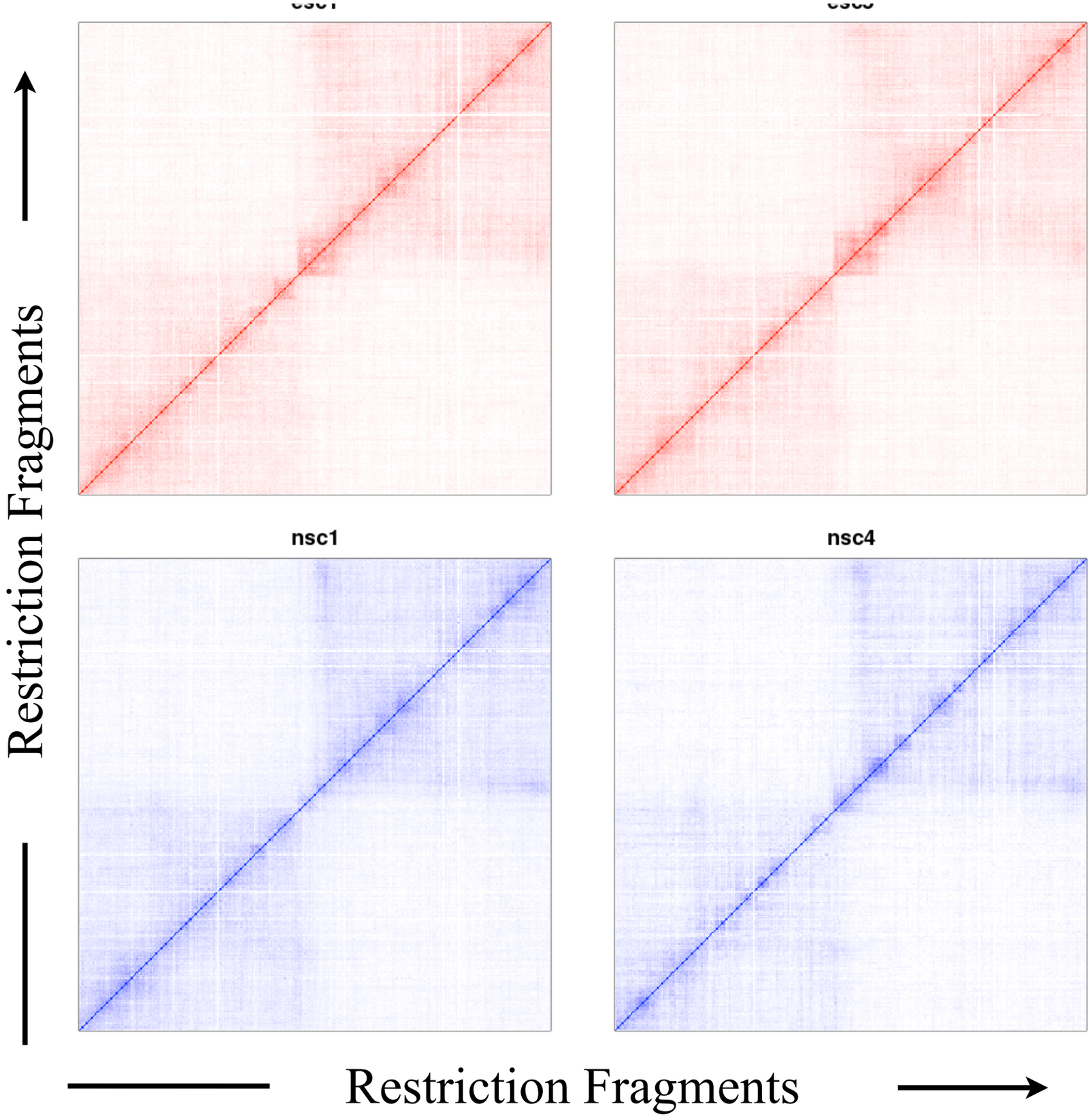
0	30	10	15	12	15	10	3	30	6	3	1	0	0	0	0	0	0	0	0
	0	29	19	3	10	0	1	50	1	3	4	2	0	0	0	0	0	0	0
		0	1	0	0	0	1	69	0	0	0	0	0	0	0	0	0	0	0
			0	2	2	3	1	12	0	0	0	0	0	0	0	0	0	0	0
				0	12	34	20	89	9	0	0	0	0	3	1	1	8	8	0
					0	32	10	56	0	0	1	1	1	0	0	0	0	0	0
						0	45	89	0	0	0	0	0	0	0	0	0	0	0
							0	99	45	30	12	3	1	0	0	0	0	0	0
								0	60	60	12	67	56	20	13	50	29	30	90
									0	12	13	4	3	3	3	1	1	0	1
										0	5	6	2	3	1	1	1	0	0
											0	13	20	15	0	0	0	0	0
												0	34	16	2	3	1	0	0
													0	19	4	1	0	0	0
														0	2	1	0	1	1
															0	1	0	0	0
																0	3	1	0
																	0	3	2
																		0	3
																			0

Restriction Fragments

Restriction Fragments

$$CS = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Cosine similarity

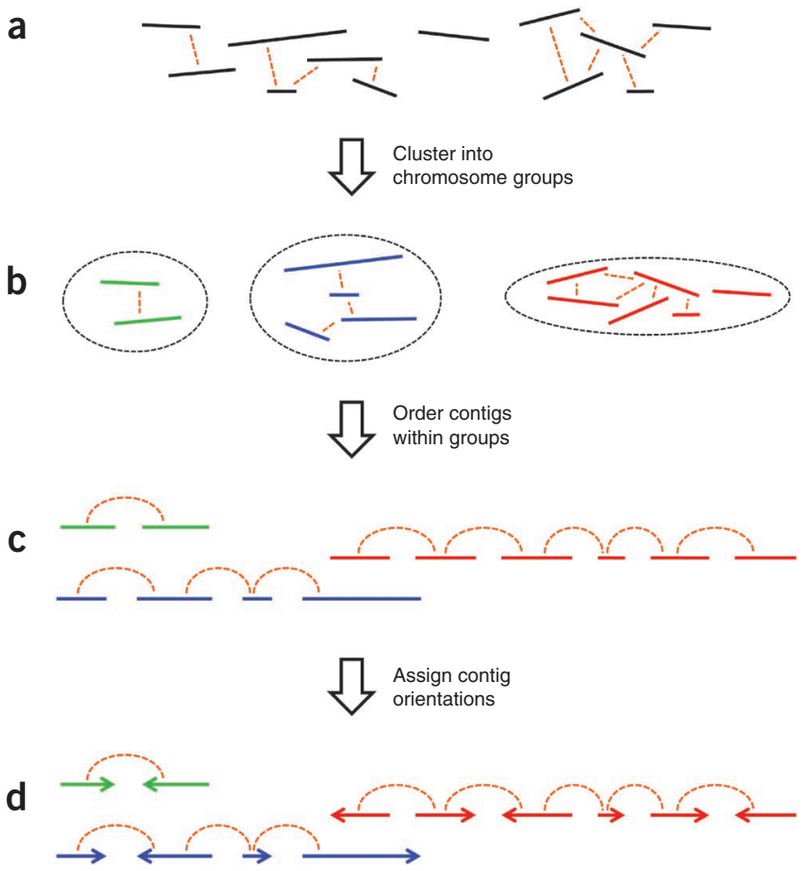


Acknowledgements

Wolfgang Huber
Bernd Klaus
Florian Hahne

Even more exciting use of Hi-C

Genome reassembly



LACHESIS

Burton 2013

Genome scaffolding

Kaplan 2013

nature COMMUNICATIONS

GRAAL

ARTICLE

Received 23 Sep 2014 | Accepted 29 Oct 2014 | Published 17 Dec 2014

DOI: 10.1038/ncomms6695 OPEN

High-quality genome (re)assembly using chromosomal contact data

Hervé Marie-Nelly[1,2,3,4,5,*], Martial Marbouty[1,2,*], Axel Cournac[1,2], Jean-François Flot[6], Gianni Liti[7], Dante Poggi Parodi[5,8], Sylvie Syan[9], Nancy Guillén[9], Antoine Margeot[8], Christophe Zimmer[3,4] & Romain Koszul[1,2]

References (I)

Ay F, Bailey TL, Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res 24: 999–1011.

Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom M a. 2011. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18**: 107–114.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**: 1119–25.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–80.

Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. 2012. HiCNorm: Removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**: 3131–3133.

Imakaev M, Fudenberg G, Mccord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza C a, Ren B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **14**: 290–294.

Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. 2011. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* **30**: 90–98.

Kaplan N, Dekker J. 2013. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol* **31**: 1143–7.

References (II)

Li W, Gong K, Li Q, Alber F, Zhou XJ. 2014. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics* **31**: 960–962.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-)* **326**: 289–93.

Marie-Nelly H, Marbouty M, Cournac A, Flot J-F, Liti G, Parodi DP, Syan S, Guillén N, Margeot A, Zimmer C, et al. 2014. High-quality genome (re)assembly using chromosomal contact data. *Nat Commun* **5**: 5695.

Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny L a, Dekker J. 2013. Organization of the mitotic chromosome. *Science* **342**: 948–53.

Rao SSP, Huntley MH, Durand NC, Stamenova EK. 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 1–16.

Selvaraj S, R Dixon J, Bansal V, Ren B. 2013. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **31**: 1111–8.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**: 458–72.

Yaffe E, Tanay A. 2011. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**: 1059–65.