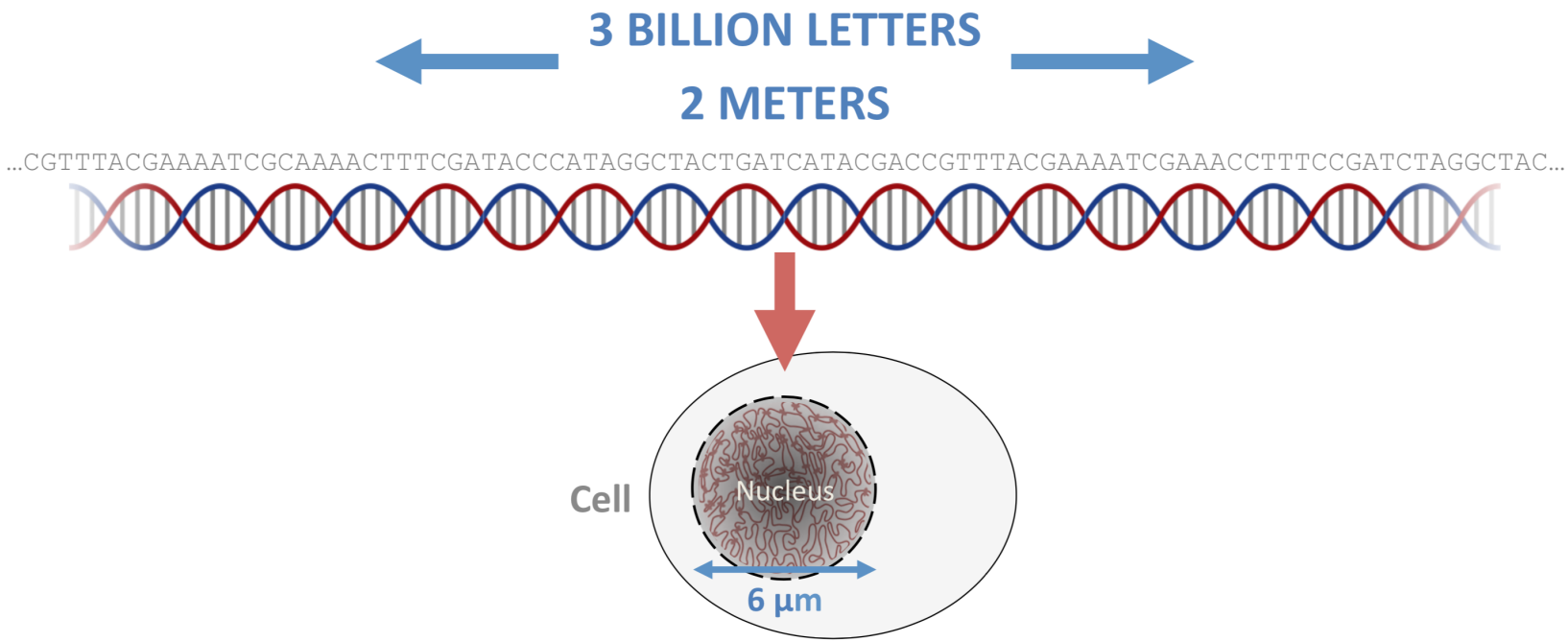
The 3-dimensional genome

Mikhail Dozmorov

Genome arithmetics

- Human genome is big - ~3.2 billion base pairs
- ~4 meters (~12ft) of diploid genome is packed into ~6-10um nucleus
- ~800 trips from Earth to Sun in ~30T cells from the human body

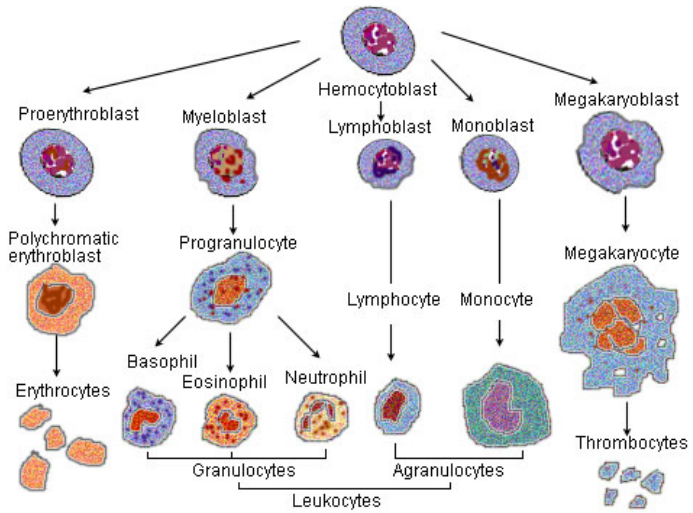


Human body has approximately 30 trillion human cells (excluding trillions of microbiome cells); Stretched haploid genome would be roughly 2 meters - each cell has 4 meters of DNA (1 m = 3.28 ft); 30 trillion * 4 meters = 120 trillion meters; Convert to miles: 120 trillion meters / 1609.34 = 7.45*10^{10}; *Convert to Earth-Sun distance: 7.45*10^{10} / 91.43*10^6 = 814.83

http://uswest.ensembl.org/Homo_sapiens/Location/Genome

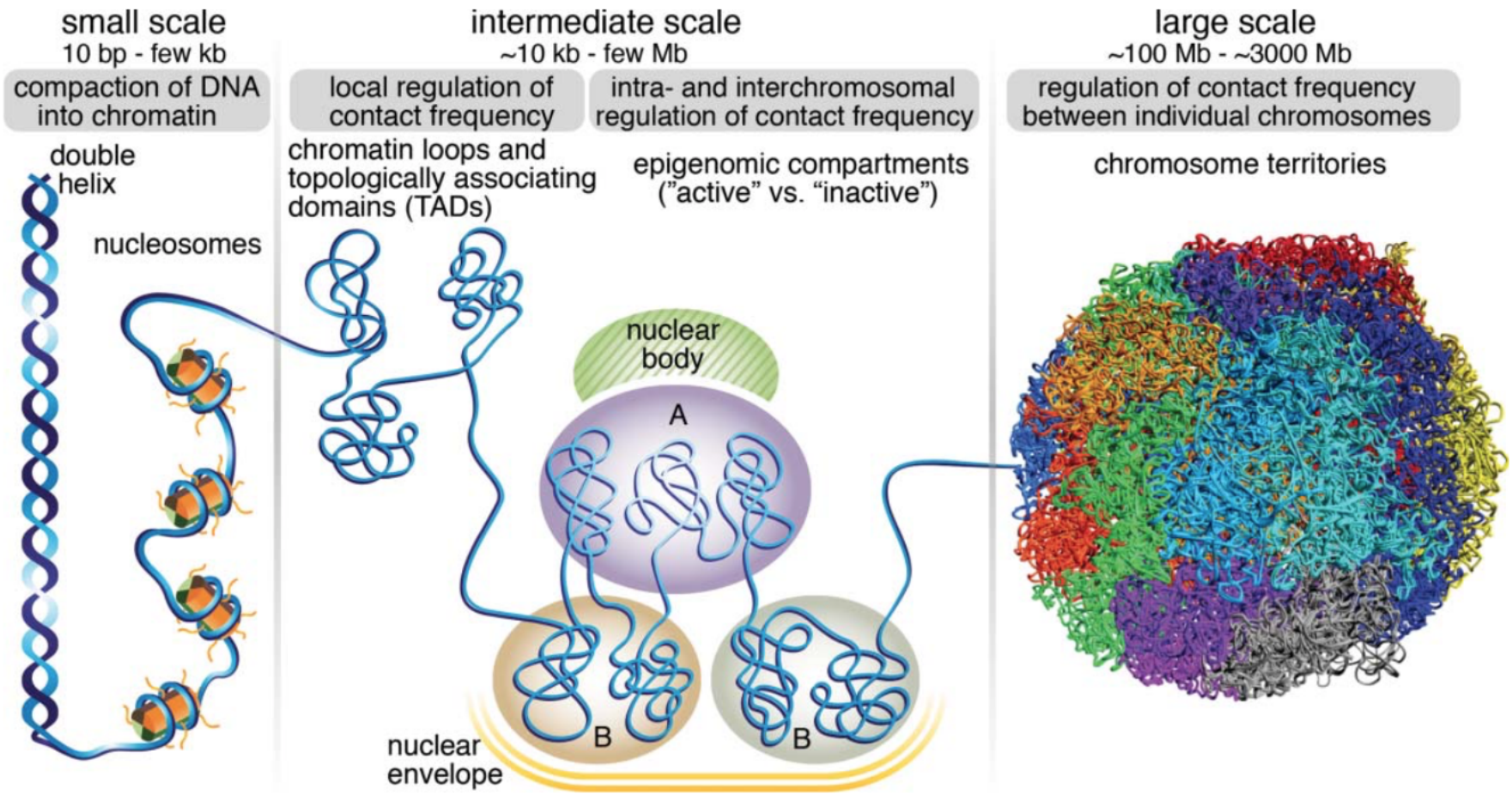
Why should we care?

- Every cell contains the same genetic material (DNA)
- Yet, the cells are very different from one another, because the genes are expressed differently and produce different protein quantities.
- These expression differences are probably linked, in part, to the three-dimensional (3D) organization of the DNA.



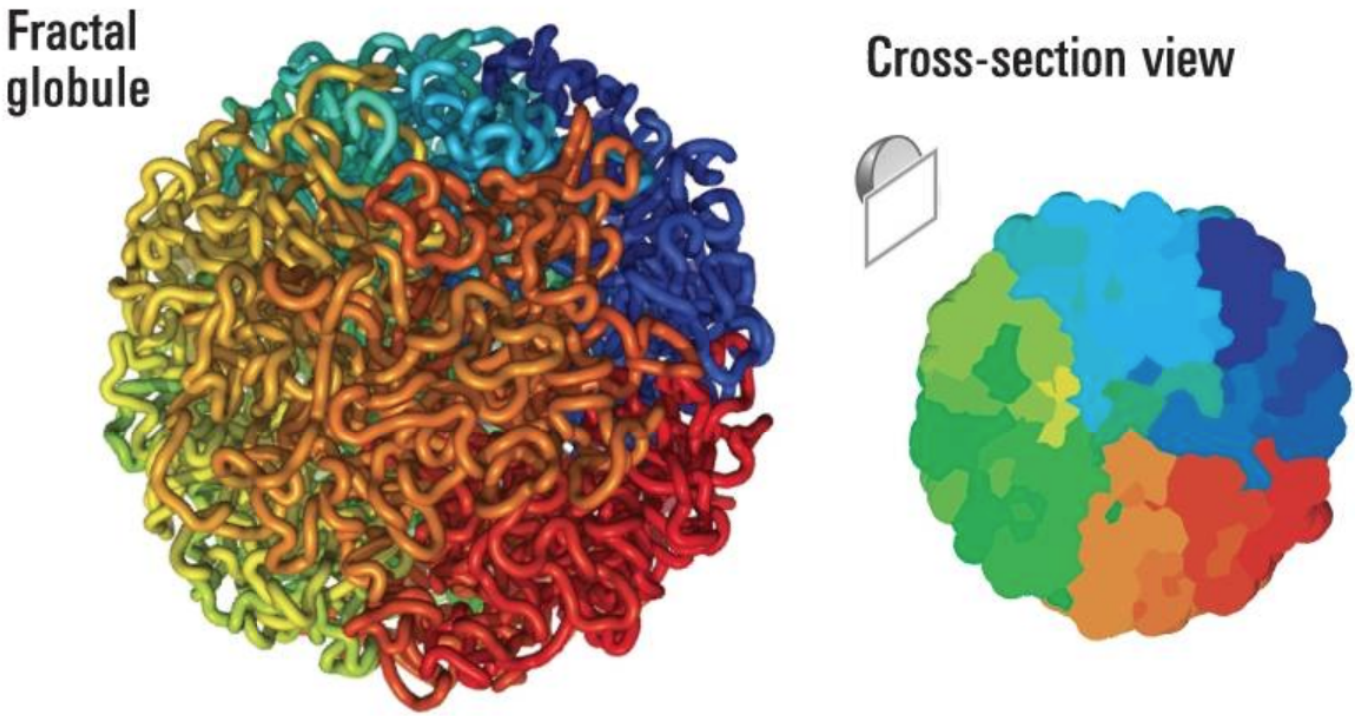
https://training.seer.cancer.gov/leukemia/anatomy/lineage.html

The human genome from a micro to macro scale



Hansen, Anders S., Claudia Cattoglio, Xavier Darzacq, and Robert Tjian. "Recent Evidence That TADs and Chromatin Loops Are Dynamic Structures." Nucleus (Austin, Tex.) 9, no. 1 (01 2018): 20–32. https://doi.org/10.1080/19491034.2017.1389365.

Genome in 3D



https://www.youtube.com/watch?v=0kJqhlyWVDA

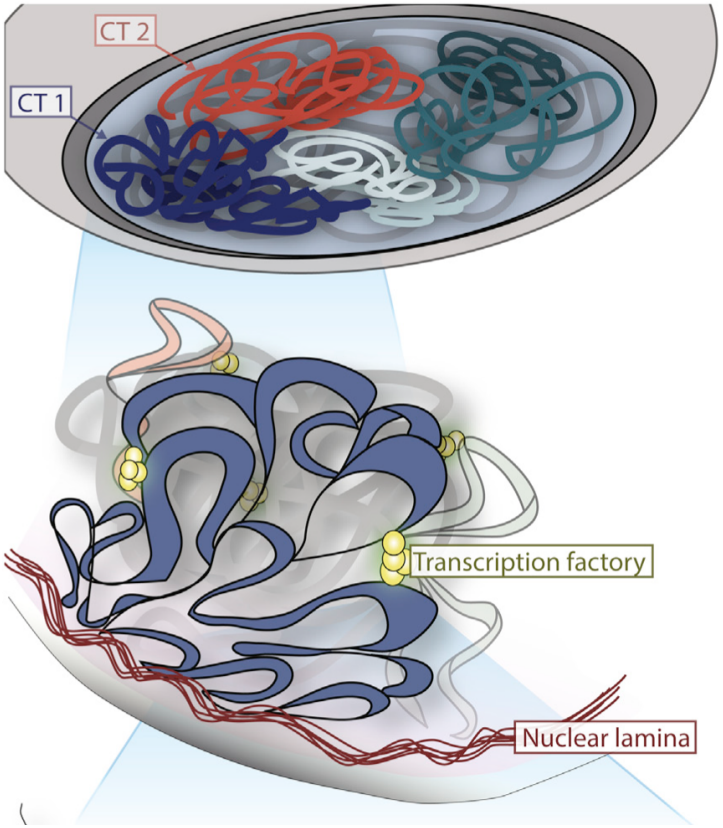
Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science, 2009. https://doi.org/10.1126/science.1181369.

Stevens, Tim J., David Lando, Srinjan Basu, Liam P. Atkinson, Yang Cao, Steven F. Lee, Martin Leeb, et al. "3D Structures of Individual Mammalian Genomes Studied by Single-Cell Hi-C". Nature, 2017. doi:10.1038/nature21429. Videos at http://www.nature.com/nature/journal/v544/n7648/full/nature21429.html#supplementary-information

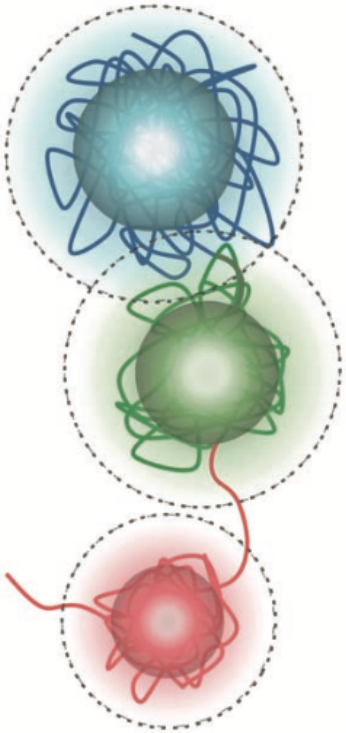
https://www.nature.com/news/dna-s-secret-weapon-against-knots-and-tangles-1.21838

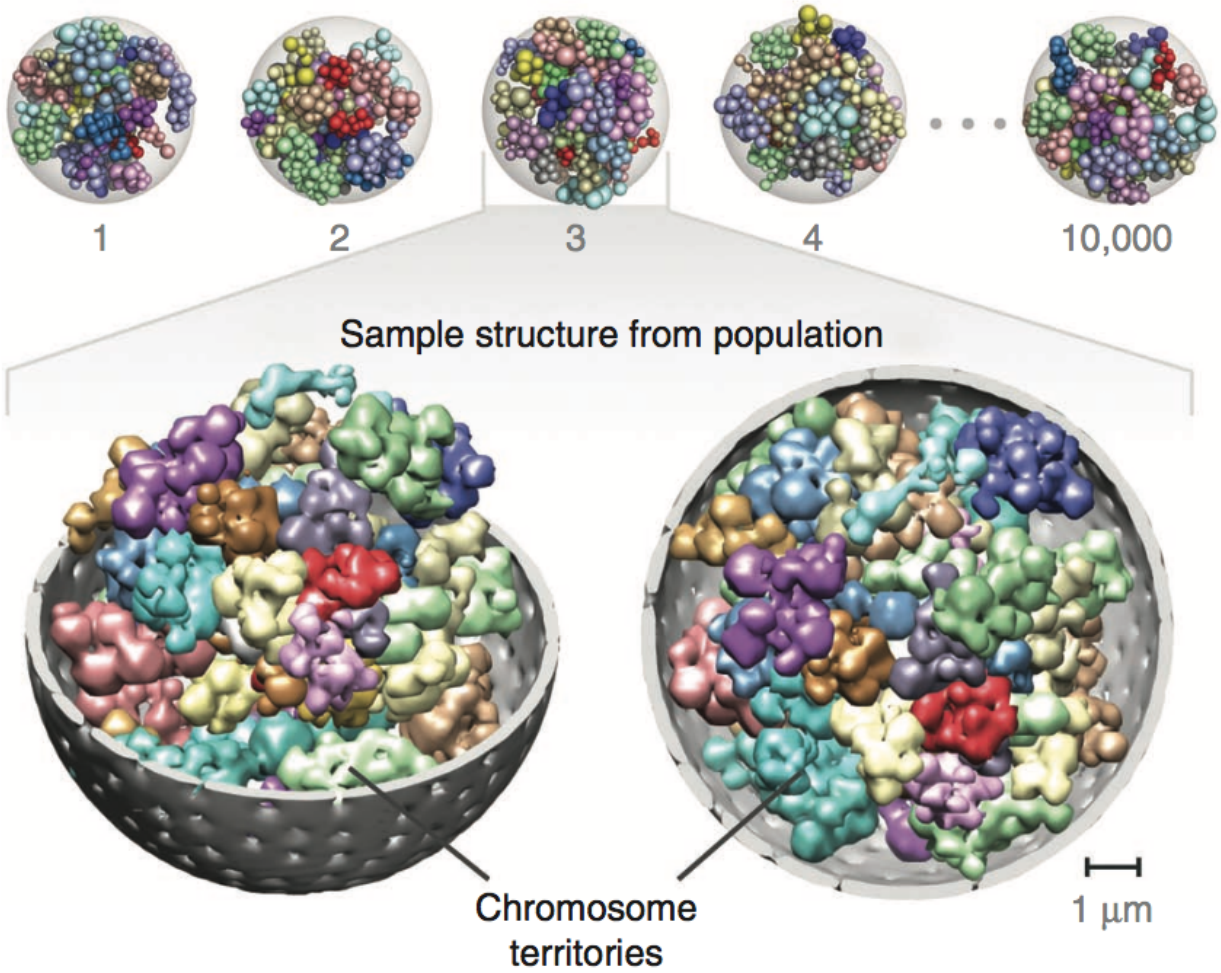
Levels of Genome Organization: Chromosome territories

- Chromosomes occupy distinct subregions of the nucleus known as chromosome territories (CTs).
- Transcriptionally inactive regions are enriched at the nuclear periphery, where they contact the nuclear lamina.
- Actively transcribed genes often colocalize at RNA polymerase II transcription factories.



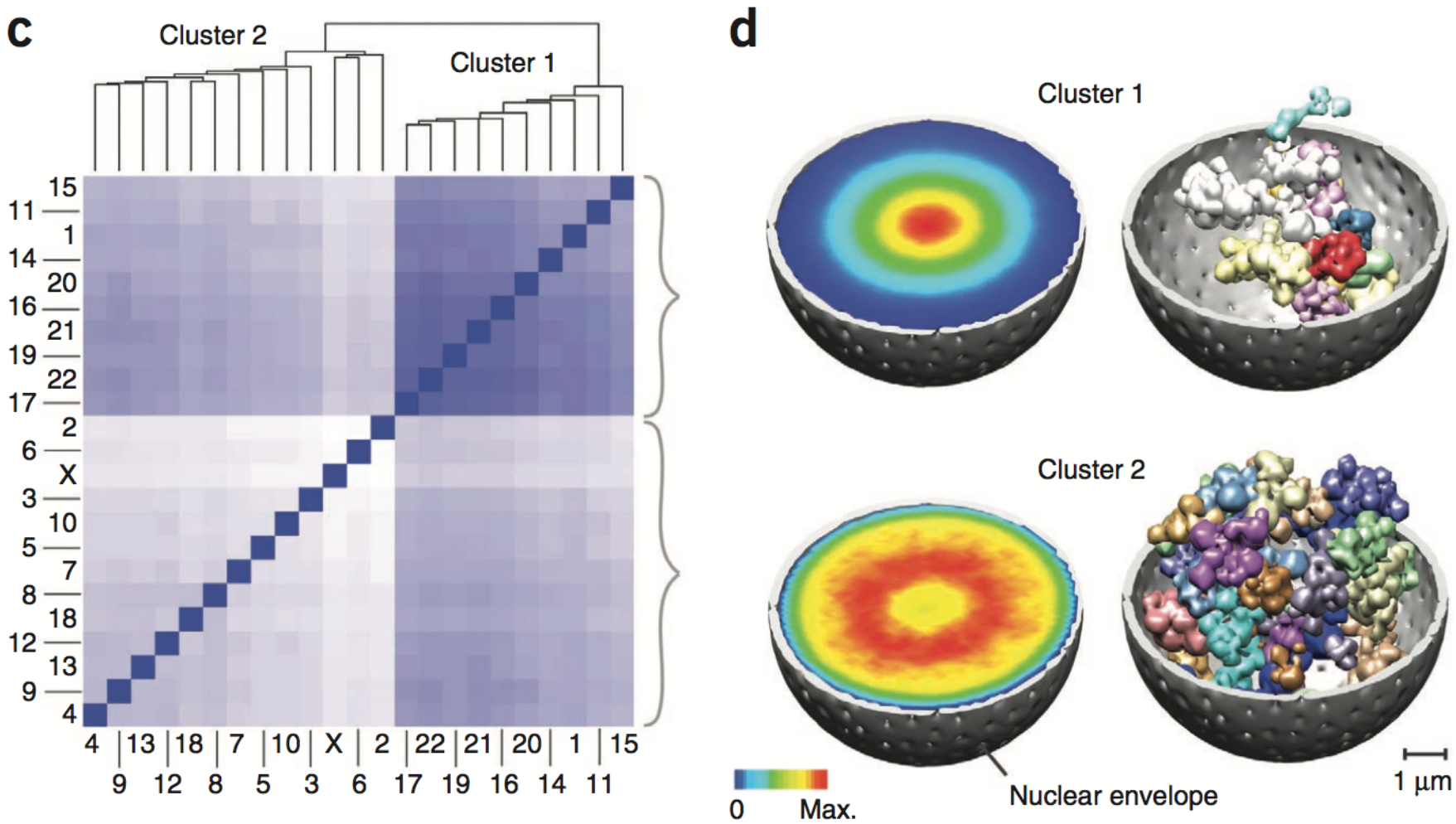
Chromosomal territories





Kalhor, Reza, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. "Genome Architectures Revealed by Tethered Chromosome Conformation Capture and Population-Based Modeling". Nature Biotechnology, 2011. https://doi.org/10.1038/nbt.2057.

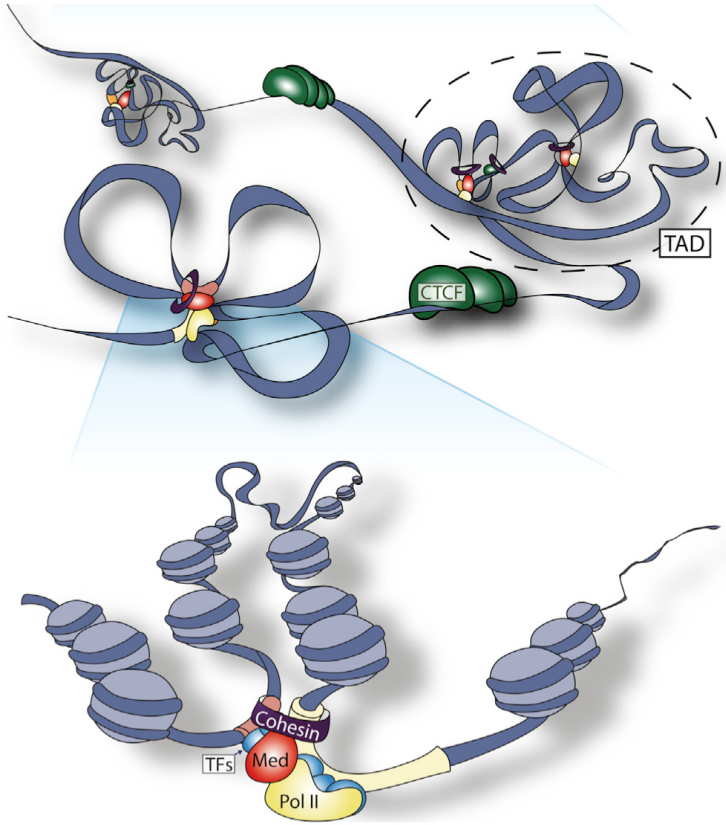
Gene-rich chromosomes are in the nuclear center, gene-poot - at the periphery



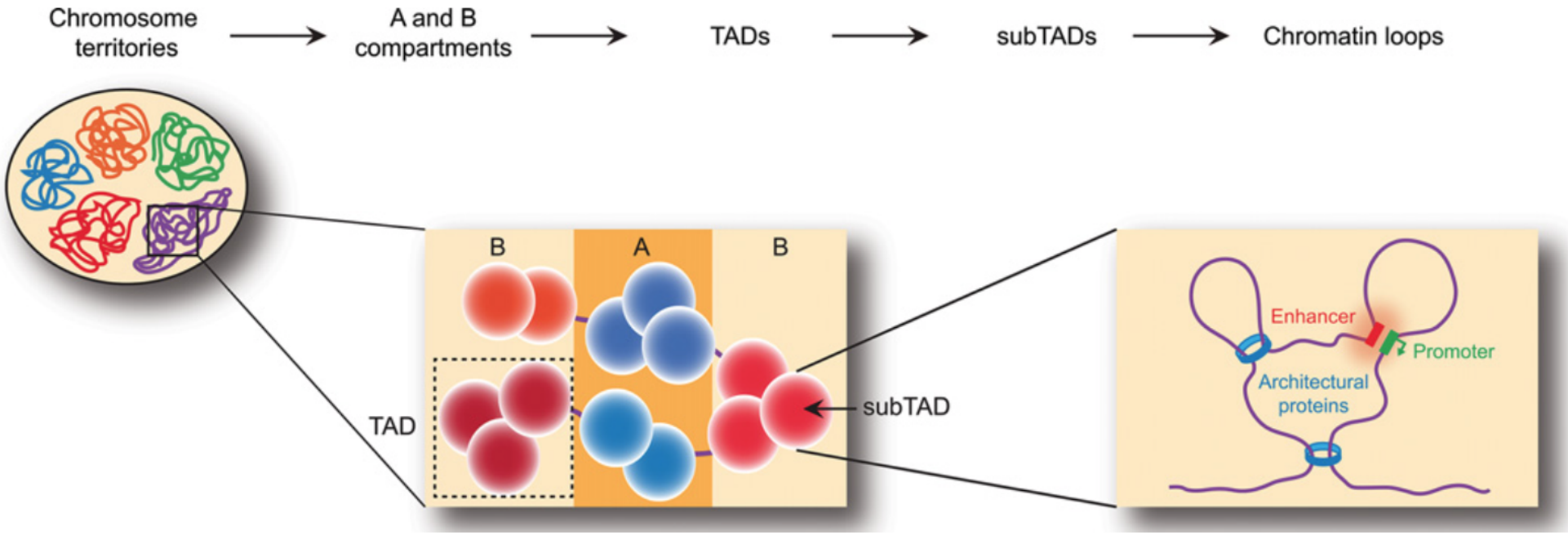
Kalhor, Reza, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome Architectures Revealed by Tethered Chromosome Conformation Capture and Population-Based Modeling. Nature Biotechnology, 2011.

Levels of Genome Organization

- The genome can be roughly partitioned into large A/B compartments corresponding to transcriptionally active/inactive chromatin, respectively.
- Topologically associating domains (TADs) are regions of frequent local interactions separated by boundaries across which interactions are less frequent.
- CTCF and cohesin binding sites are enriched at TAD boundaries.

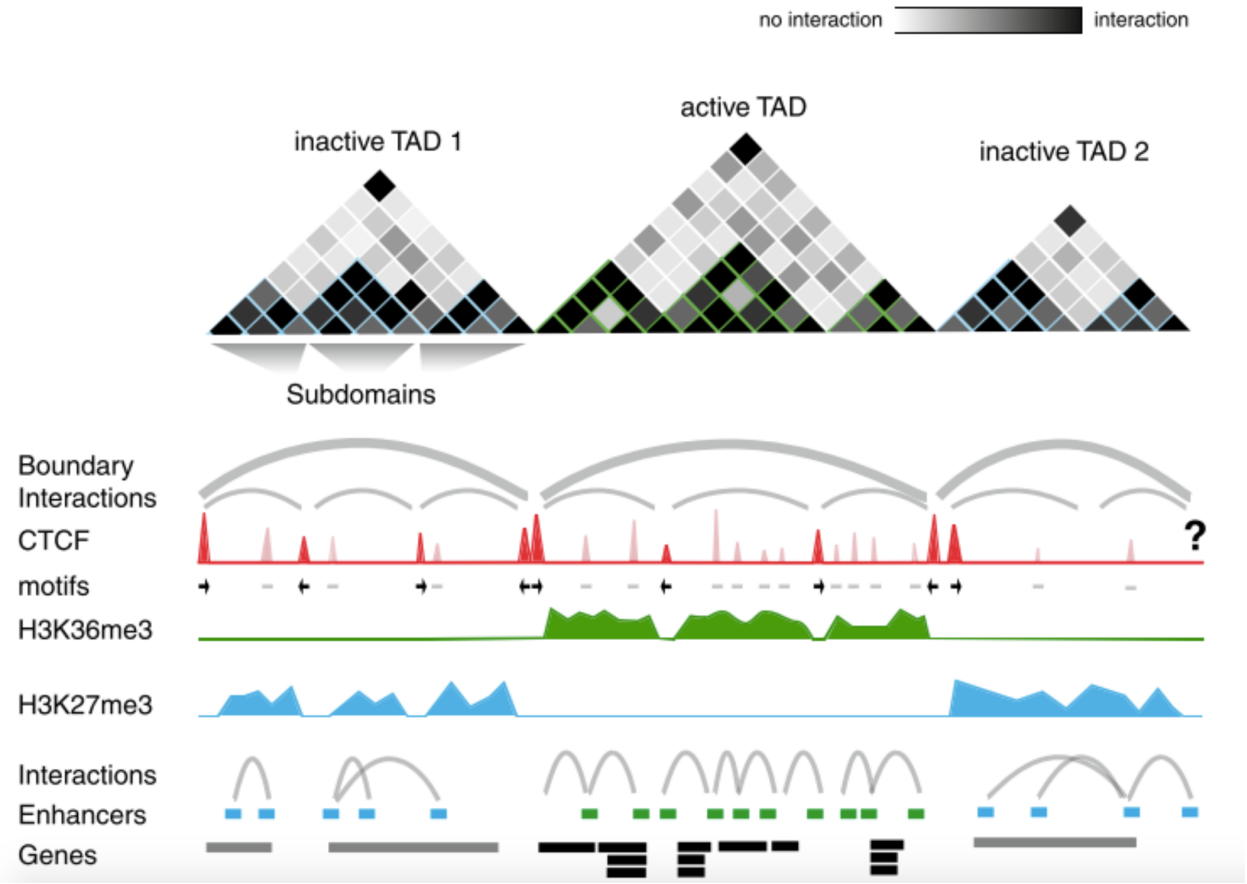


Hierarchical genome organization



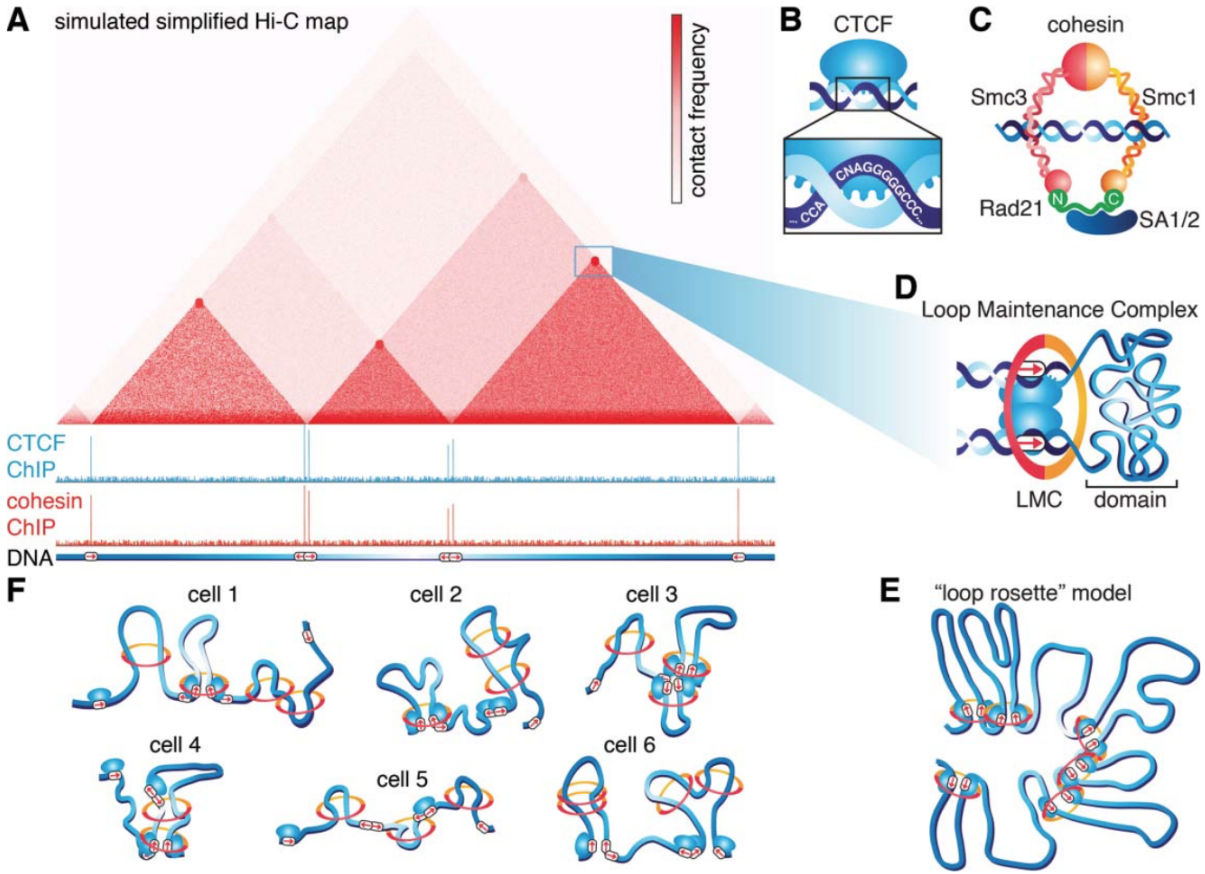
Schematic representation of the organization of the 3D genome into A (blue) and B (red) compartments and topologically associated domains (TADs), which are composed of several sub-TADs (depicted here as spheres), which in turn harbor several chromatin loops.

3D structure and the epigenetics



A combinatorial illustration showing TADs, CTCF bindings, corresponding epigenetic features, and long-range CTCF-mediated long-range interactions.

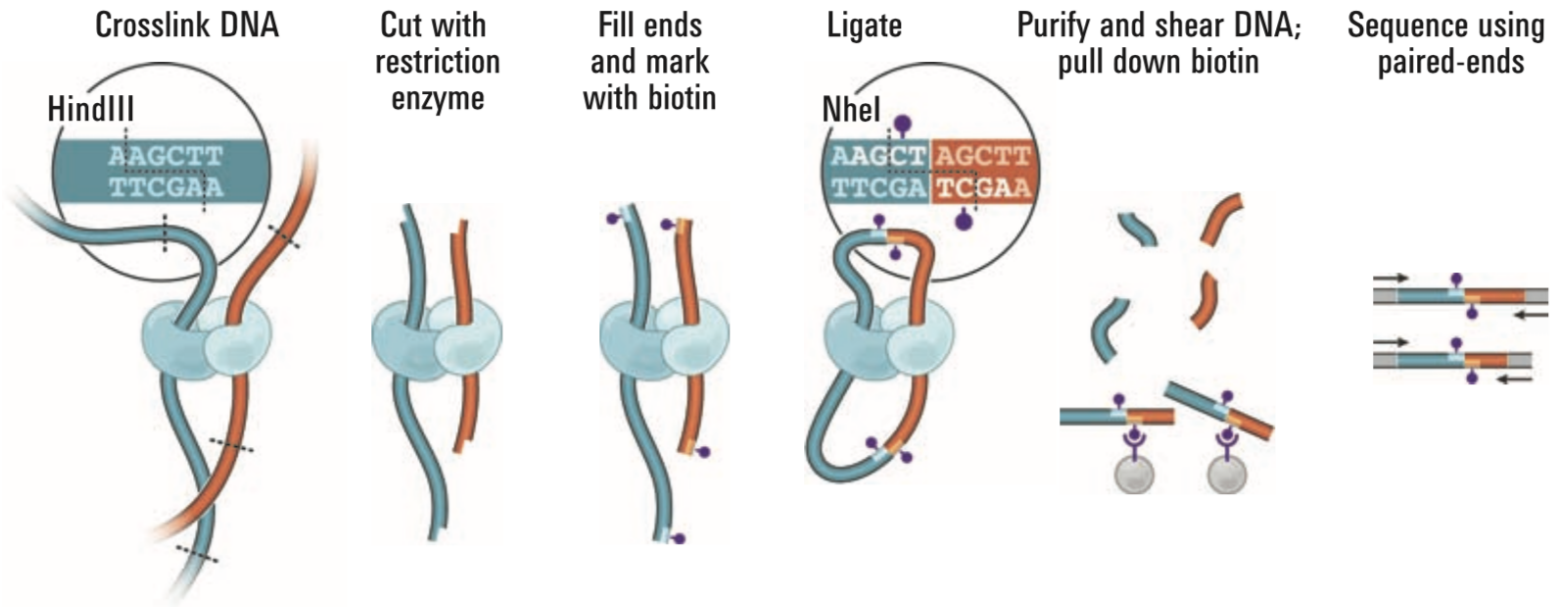
Loop extrusion mechanism



Fudenberg, Geoffrey, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A. Mirny. "Formation of Chromosomal Domains by Loop Extrusion." Cell Reports 15, no. 9 (May 31, 2016): 2038–49. https://doi.org/10.1016/j.celrep.2016.04.085

Chromosome conformation capture technologies

The core strategy in 3D genome mapping is nuclear proximity ligation (Cullen et al., 1993), which allows detection of distant genomic segments residing in close spatial proximity to one another, yet are linearly far away.

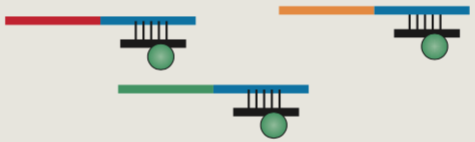
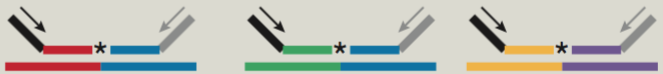


Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science, 2009. https://doi.org/10.1126/science.1181369.

Chromosome conformation capture technologies

- **3C** focuses on contacts between two selected DNA fragments.
- 3C-on chip (**4C**) and **Capture-C**-like approaches focus on all contacts made by one or more given DNA fragments of interest.
- 3C carbon copy (**5C**) analyses contacts among multiple selected DNA fragments.
- **Hi-C** provides an analysis of contacts among all genomic fragments. Requires deep sequencing (>20X RNA-seq, >600M reads)
- Chromatin interaction analysis by paired-end tag (**ChIA-PET**) combines 3C with chromatin immunoprecipitation (ChIP) to analyse the contacts of sequences bound by a protein of interest.

Chromosome conformation capture technologies

Technique	Purpose	Principle of detection
3C	One-to-one	qPCR 
4C	One-to-all	Inverse PCR from selected fragment plus sequencing 
Capture-C	One/many-to-all	Pull down of selected fragment plus sequencing 
5C	Many-to-many	Multiplex ligation-mediated amplification plus sequencing 
ChIA-PET	Many-to-many	ChIP plus sequencing 
Hi-C	All-to-all	Sequencing 

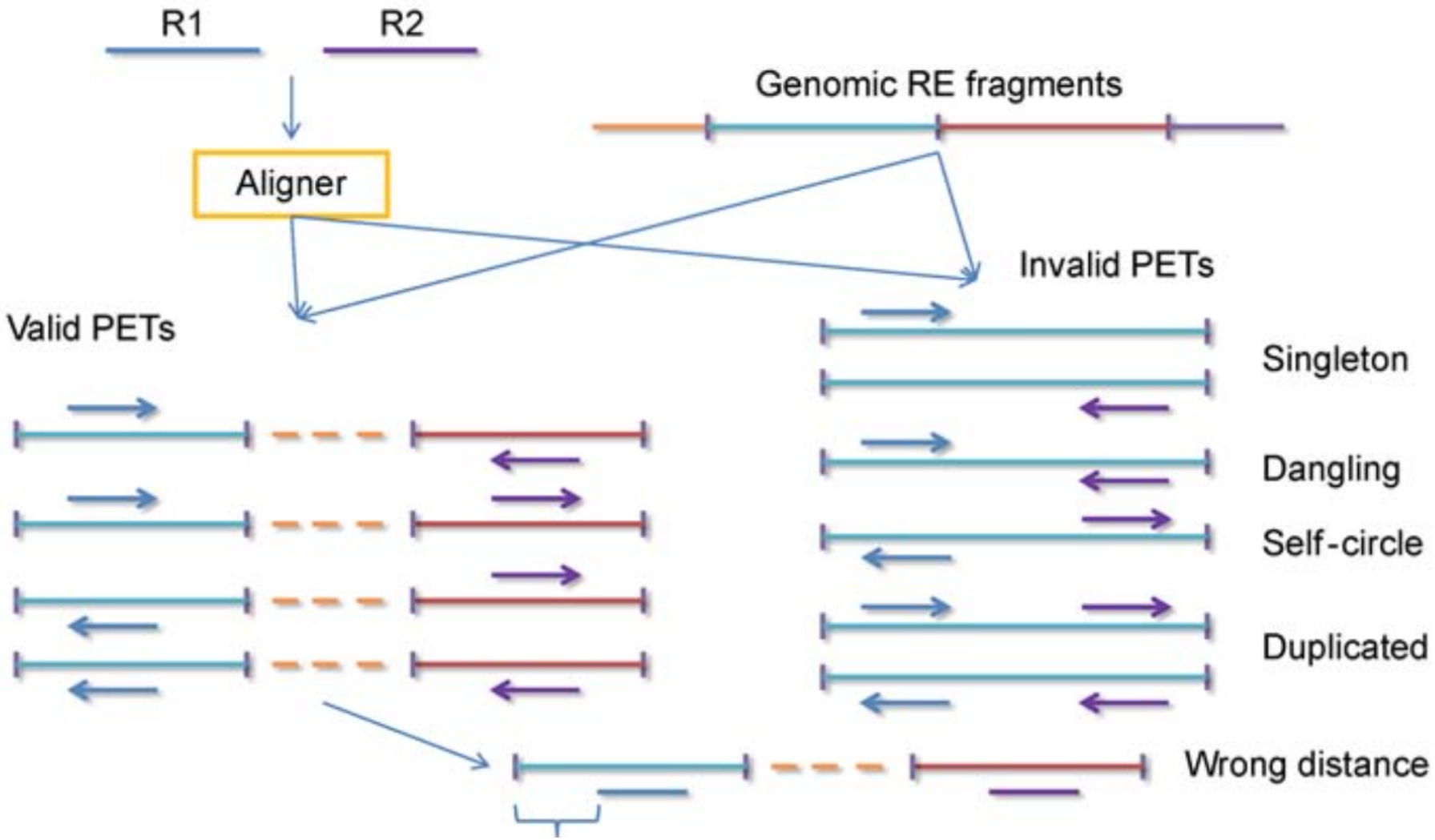
Cutting (aka restriction) enzymes

- Recognize specific short sequences in the genome
  - HindIII, recognizes a 6 bp motif AAGCTT
  - EcoRI, recognizes a 6 bp motif GAATTC
  - DpnII, recognizes a 4 bp motif GATC
- Using a 6-bp cutting restriction enzyme, there are $\sim 10^6$ restriction fragments, leading to an interaction space on the order of 10^{12} possible pairwise interactions.
- 4-cutters potentiate higher-resolution analyses of genome conformation by means of producing smaller restriction-fragment sizes, the total number of restriction fragments genome-wide is ~16-fold higher and the total number of possible pairwise contacts is 256-fold higher.

Hi-C sequencing products

1. Two different, noncontiguous restriction fragments, indicating a potential interaction between the two fragments.
2. Two ends map within the same restriction fragment, giving no information on chromatin topology.
3. Two ends map to adjacent restriction fragments that are contiguous in the genome, either due to incomplete restriction digestion or re-ligation of the two fragments. Also noninformative.
4. A concatemer of multiple restriction fragments.

Filtering mapped PETs

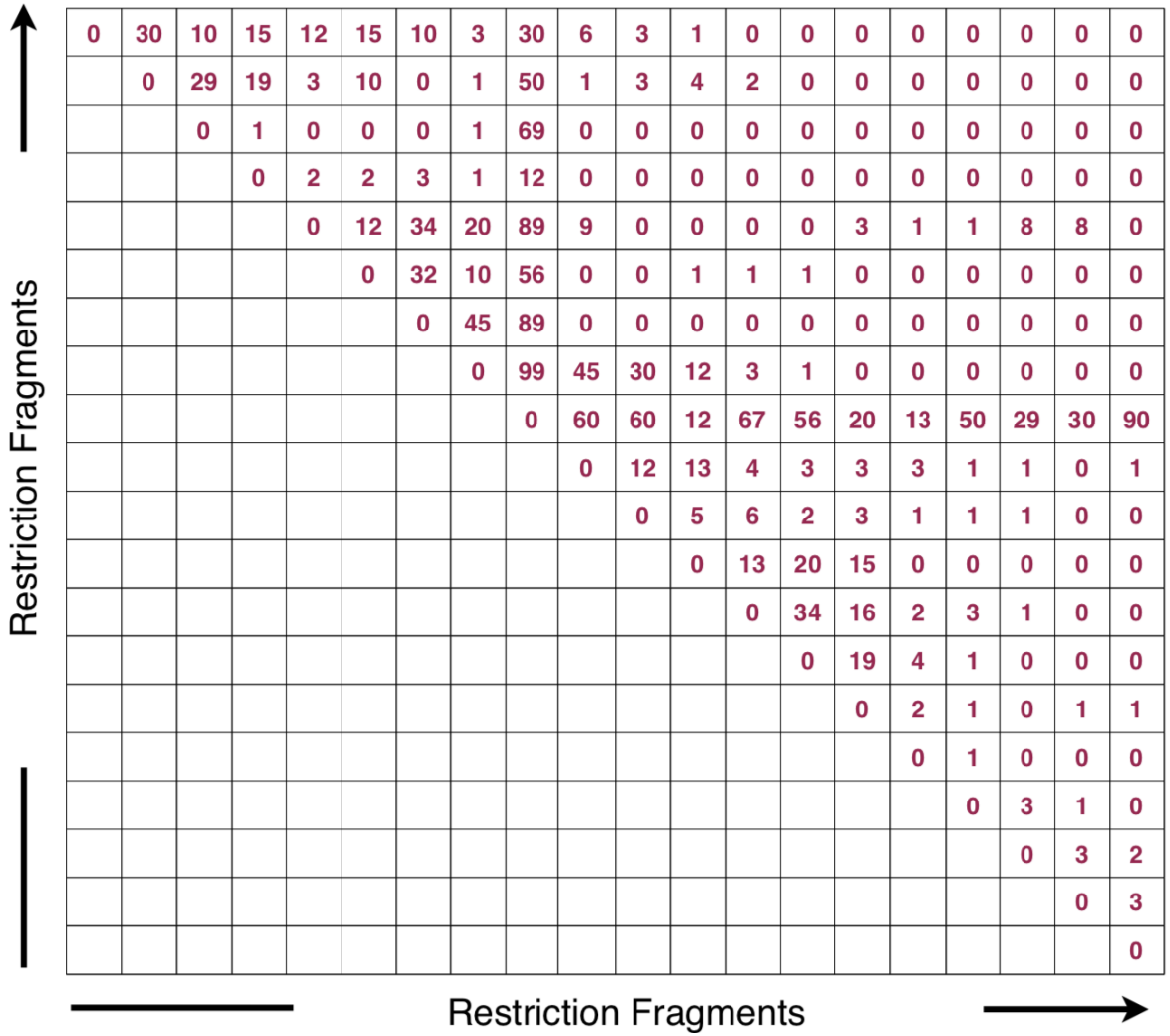


Han, Zhijun, and Gang Wei. "Computational Tools for Hi-C Data Analysis". Quantitative Biology, 2017. https://doi.org/10.1007/s40484-017-0113-6.

Hi-C data

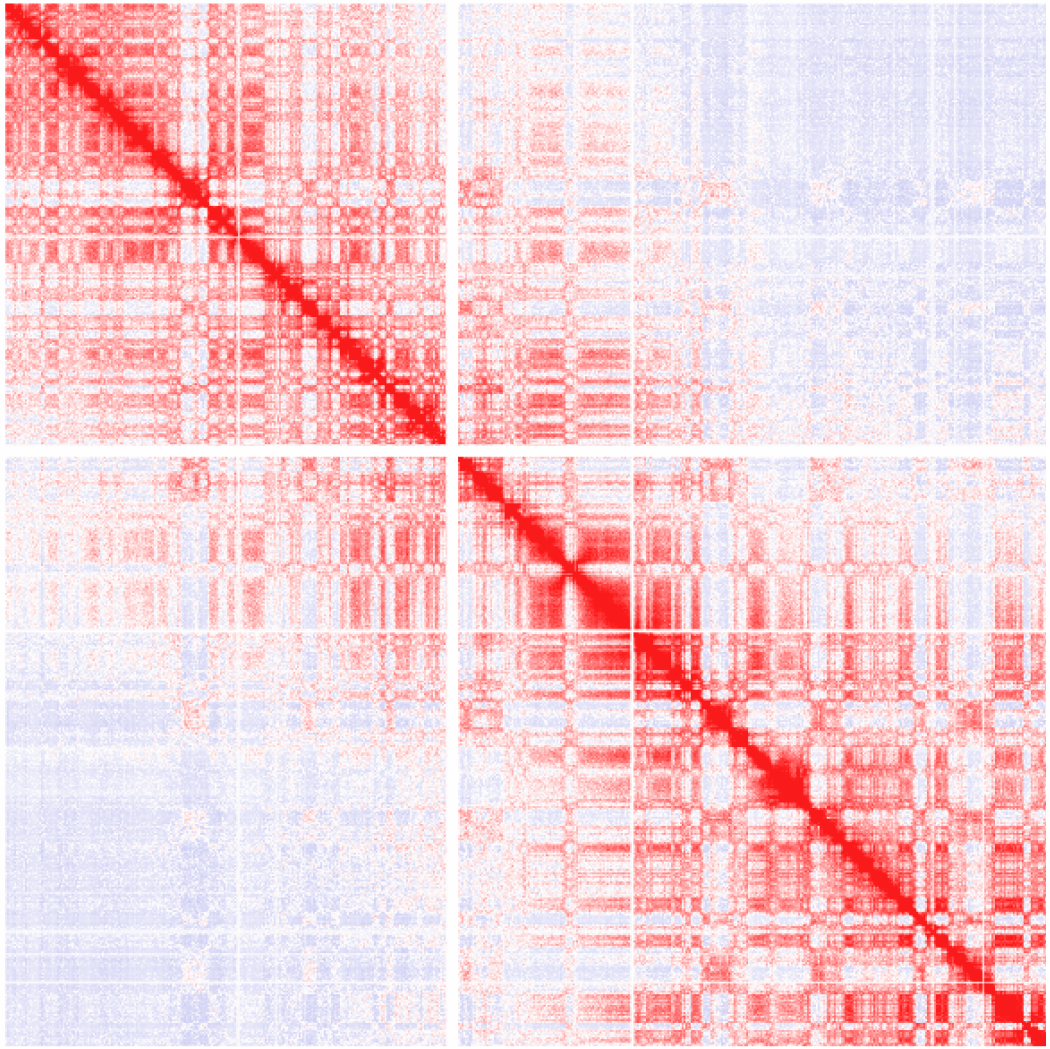
- Paired-end sequencing, each read in a pair is processed separately.
- Usually summarized the counts into a 2D matrix:
  - First cut genome into N equal sized bins (size depends on sequencing depth).
- Summarize the read counts into $N \times N$ matrix. The element (i, j) represents the number of pairs with one end from the ith window and the other end from the jth window.
- The counts represent the strength of interaction.
- Usually the numbers on diagonal are greater.

Hi-C interaction matrix

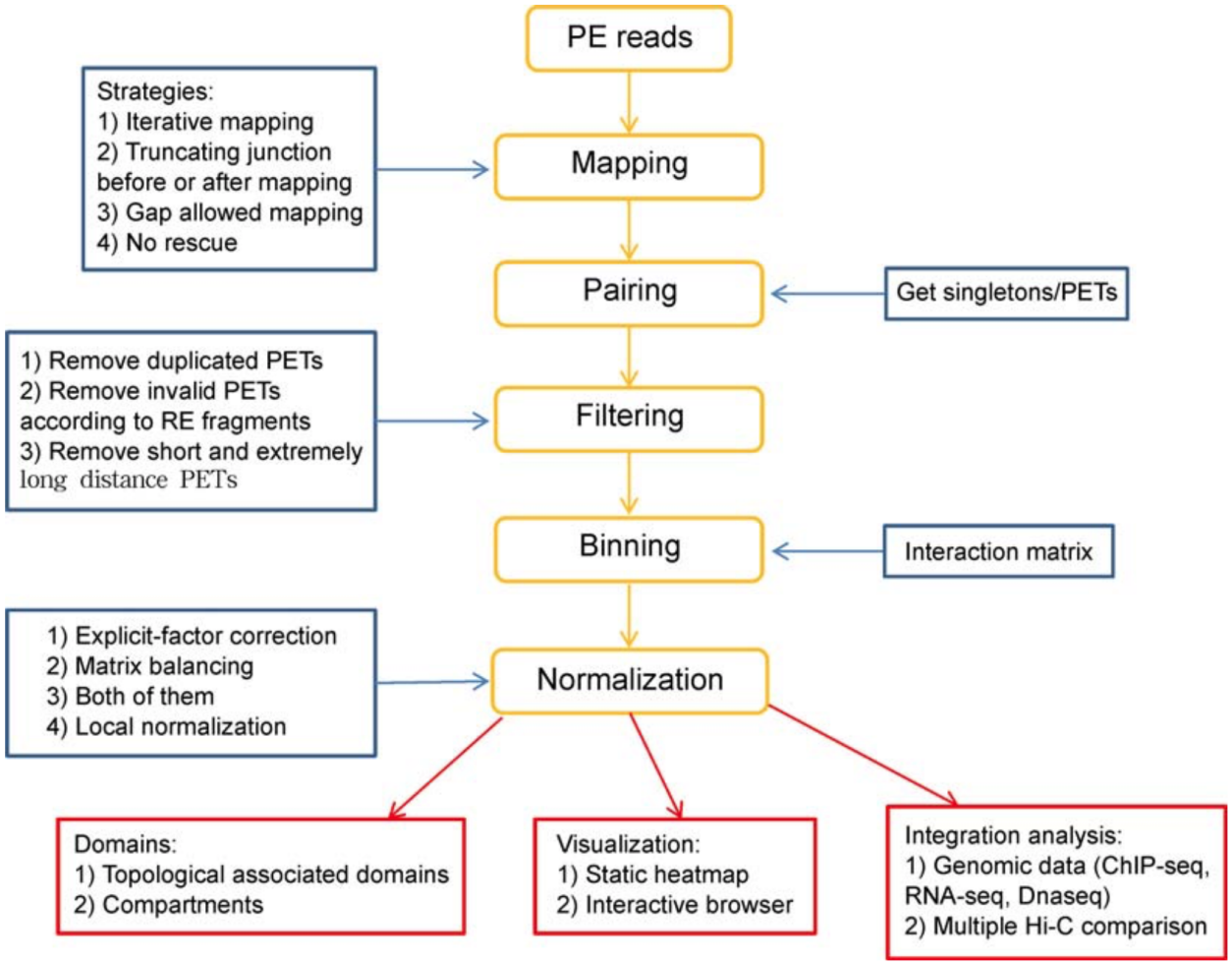


0	30	10	15	12	15	10	3	30	6	3	1	0	0	0	0	0	0	0	0
	0	29	19	3	10	0	1	50	1	3	4	2	0	0	0	0	0	0	0
		0	1	0	0	0	1	69	0	0	0	0	0	0	0	0	0	0	0
			0	2	2	3	1	12	0	0	0	0	0	0	0	0	0	0	0
				0	12	34	20	89	9	0	0	0	0	3	1	1	8	8	0
					0	32	10	56	0	0	1	1	1	0	0	0	0	0	0
						0	45	89	0	0	0	0	0	0	0	0	0	0	0
							0	99	45	30	12	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

Visualize Hi-C data in a heatmap

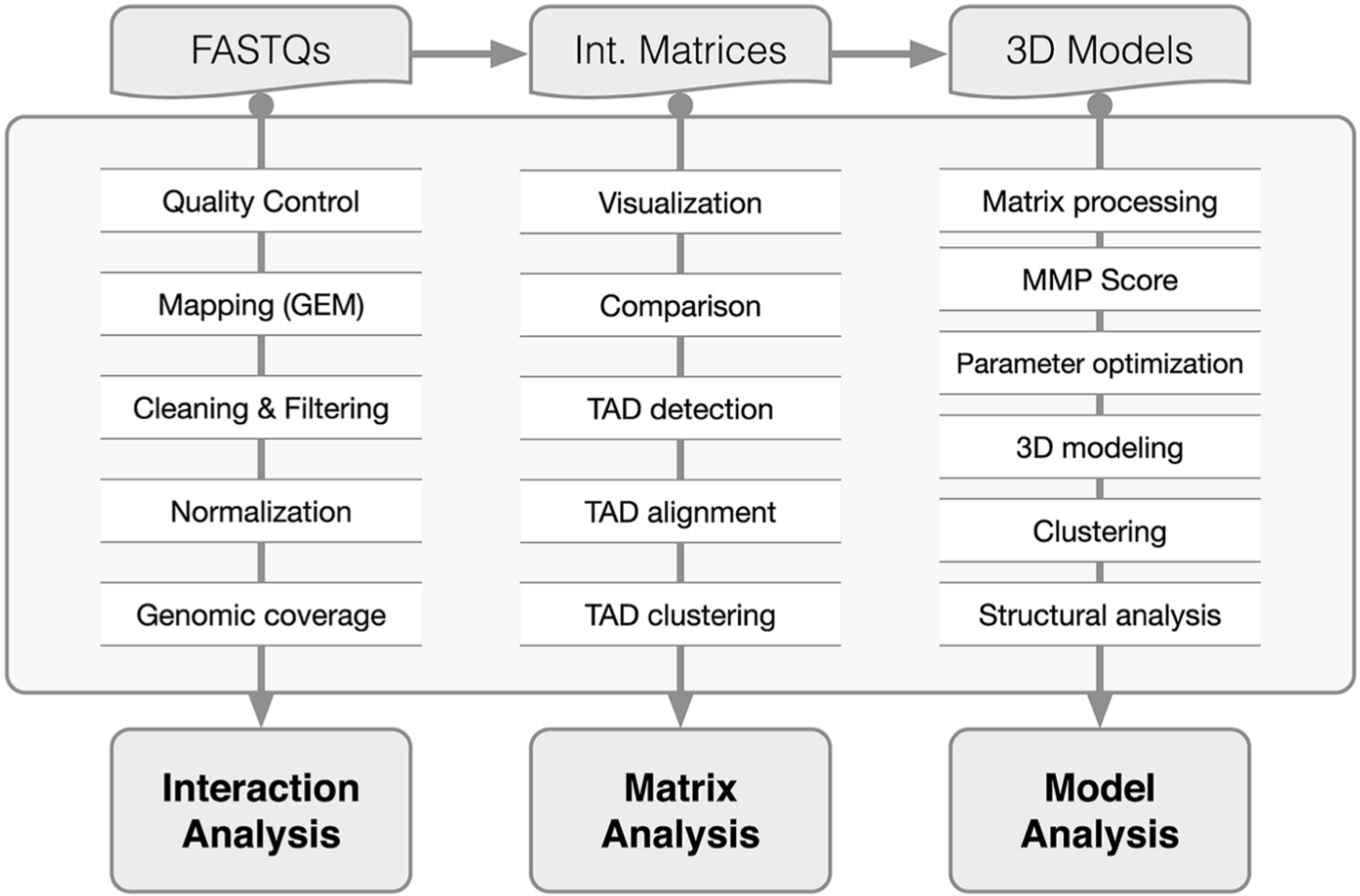


General Hi-C data processing workflow



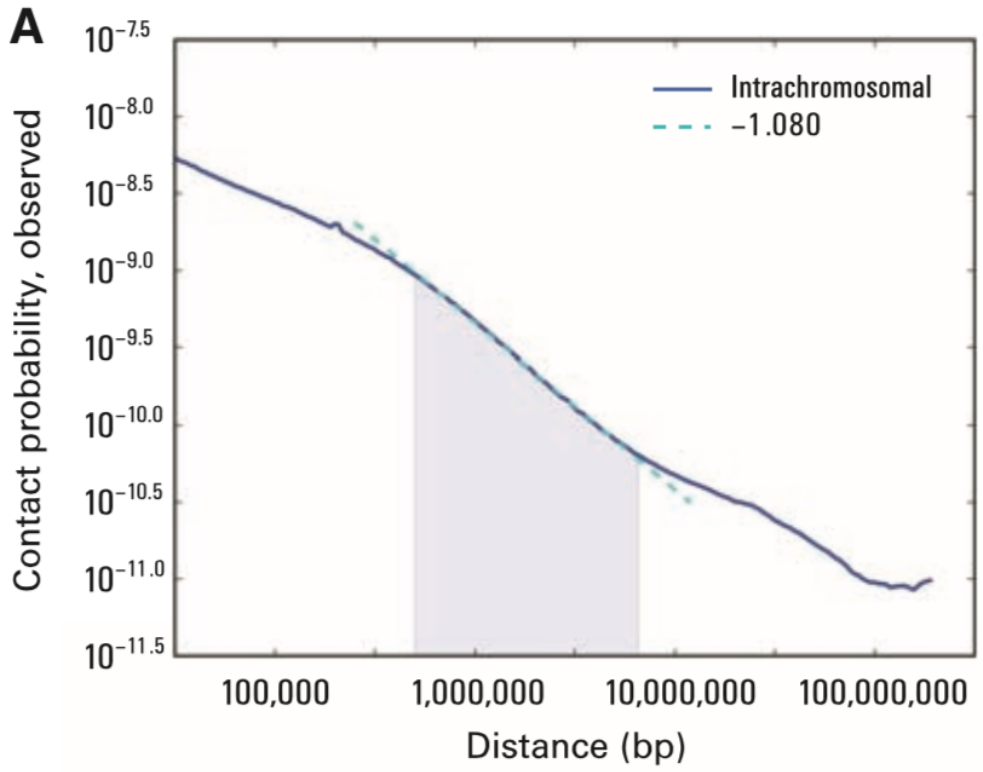
Han, Zhijun, and Gang Wei. "Computational Tools for Hi-C Data Analysis." Quantitative Biology, 2017. https://doi.org/10.1007/s40484-017-0113-6.

Hi-C data analysis



Serra, Francoois, Davide Bas, Mike Goodstadt, David Castillo, Guillaume J. Filion, and Marc A. Marti-Renom. "Automatic Analysis and 3D-Modelling of Hi-C Data Using TADbit Reveals Structural Features of the Fly Chromatin Colors". PLoS Computational Biology, 2017. https://doi.org/10.1371/journal.pcbi.1005665.

Distance-dependent decay of chromatin contacts



Contact probability as a function of genomic distance averaged across the genome (blue) shows a power law scaling between 500 kb and 7 Mb (shaded region) with a slope of ~ 1.08 (fit shown in cyan).

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science, 2009. https://doi.org/10.1126/science.1181369.

Hi-C data biases

- Raw Hi-C data have been observed to have both **technical** and **sequence-driven** biases (Yaffe and Tanay, 2011)
- Three predominant sources of sequence-driven bias identified so far
  1. Fragment length
  2. GC bias
  3. Mappability
- Removing biases (normalizing the data) usually improves correlation among replicates.

Normalization methods

- There are two general approaches to Hi-C bias correction: **explicit** and **implicit**.
  - Explicit bias models take into account factors such as mappability, GC content and fragment length.
  - Implicit approach, aka matrix balancing, iterative correction, is based on the assumption that since we are interrogating the entire interaction space in an unbiased manner, each fragment/bin should be observed approximately the same number of times in the experiment (interpreted as the sum of the genome-wide row/column in the interaction matrix).

Matrix balancing algorithms

The implicit or matrix-balancing approach does not require definition of predetermined factors that may introduce bias.

- Assumptions:
  - All fragments throughout the genome have equal visibility (i.e. equal propensity for detection via a proximity ligation assay).
  - The intrinsic fragment-specific biases can be represented as a single scalar value for each fragment that interacts multiplicatively with the intrinsic biases of its ligation partners.
- Idea: find a decomposition of the observed contact matrix into a vector of bias terms and a normalized contact map in which all rows have equal sums.

Knight-Ruiz (KR)

- Given a non-negative symmetric matrix D, the algorithm tries to find a vector x such that

$$diag(x)Dx = e$$

where $diag(x)$ is a diagonal matrix converted from x, and e represents a vector of all ones.

- Balancing a matrix in the 1-norm.

Knight, P. A., and D. Ruiz. "A Fast Algorithm for Matrix Balancing." IMA Journal of Numerical Analysis, July 1, 2013 https://doi.org/10.1093/imanum/drs019

Vanilla coverage (VC)

$$diag(x)Dx = e$$

- e_i is 1-norm of the ith row, e_j is 1-norm of the jth column

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science (New York, N.Y.) 326, no. 5950 (October 9, 2009): 289–93. https://doi.org/10.1126/science.1181369.

Sequential Component Normalization (SCN)

$$diag(x)Dx = e$$

- e_i is 2-norm of the ith row, e_j is 2-norm of the jth column

Cournac, Axel, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. "Normalization of a Chromosomal Contact Map." BMC Genomics 13 (2012): 436. https://doi.org/10.1186/1471-2164-13-436.

ICE - Iterative Correction and Eigenvalue decomposition, normalization of HiC data

- Don't explicitly model sources of bias. Only assume:

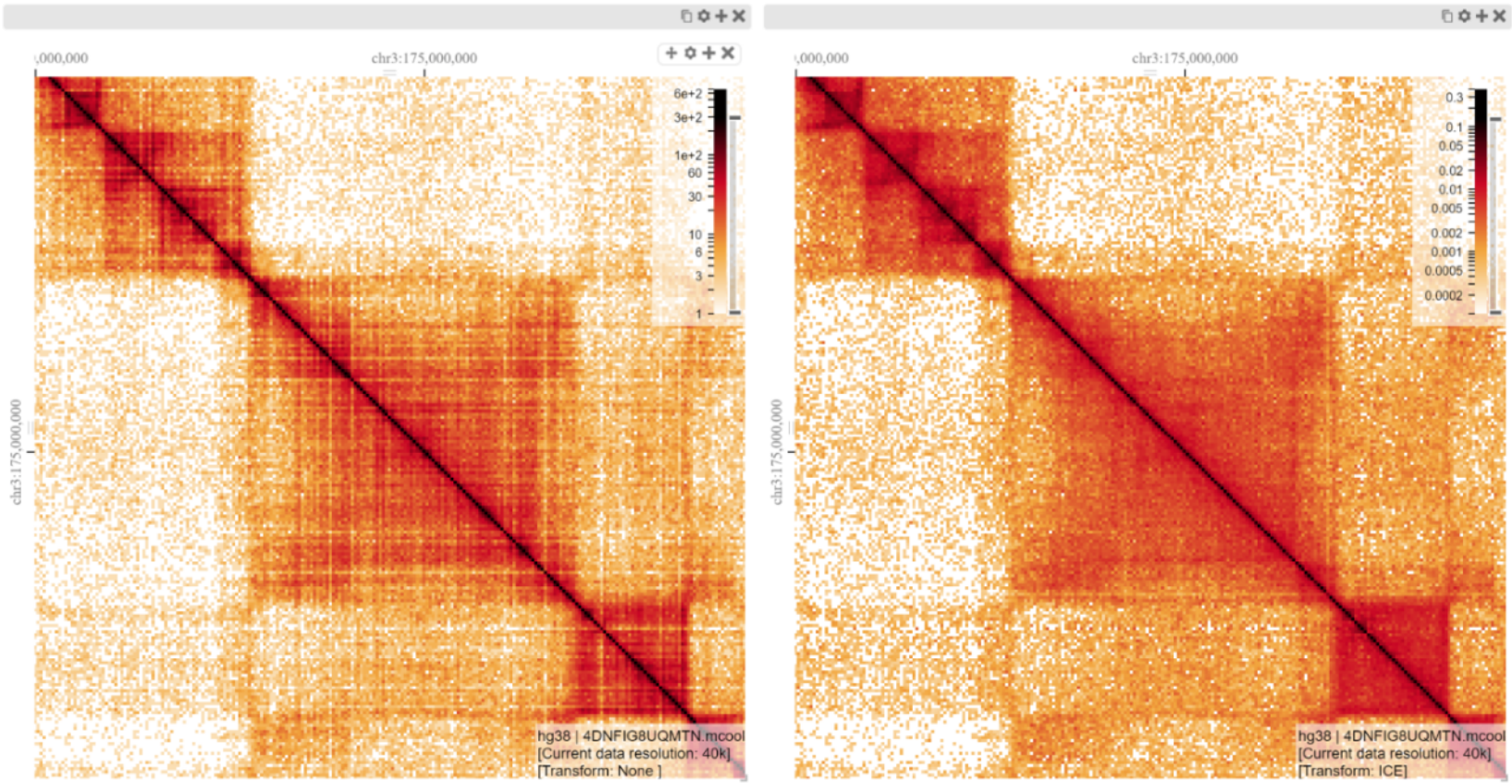
$$O_{i,j} = B_i * B_j * T_{i,j}$$

where B are bias vectors and $\sum_i T_{i,j} = 1$

- Two stopping criteria, the maximum iteration and error tolerance.
- Different names - Sinkhorn-Knapp, Knight-Ruiz.

Imakaev, Maxim, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. "Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization." Nature Methods 9, no. 10 (October 2012): 999–1003. https://doi.org/10.1038/nmeth.2148

Matrix balancing



Uncorrected

Corrected (balanced)

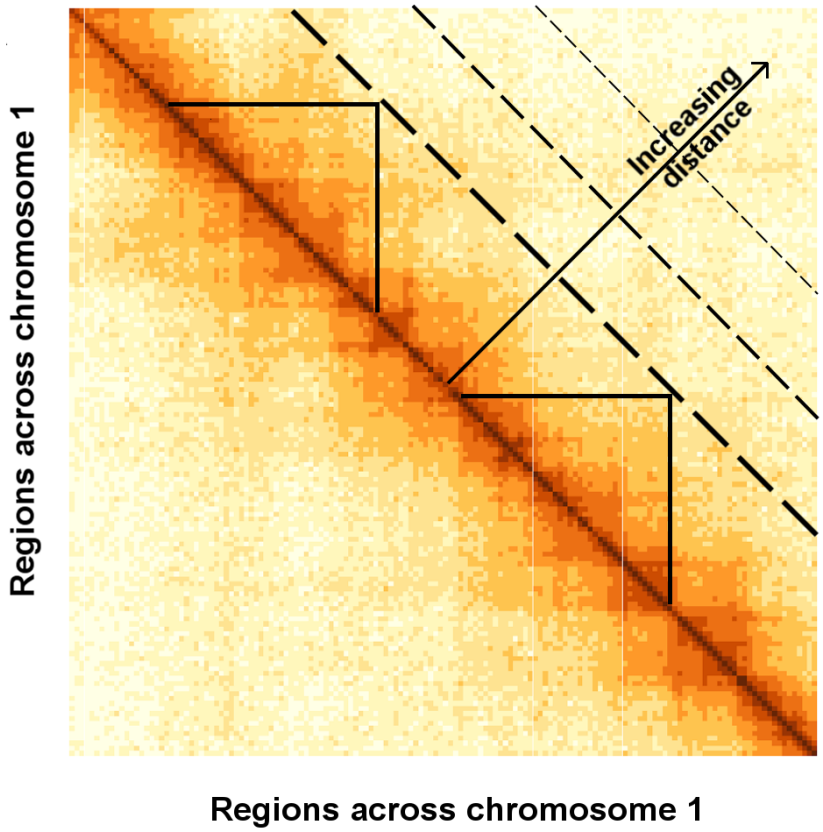
https://github.com/hms-dbmi/hic-data-analysis-bootcamp

Normalization of two or more Hi-C datasets

- Most normalization methods normalize individual Hi-C matrices.
- How to normalize two or more Hi-C matrices if we want to compare them?

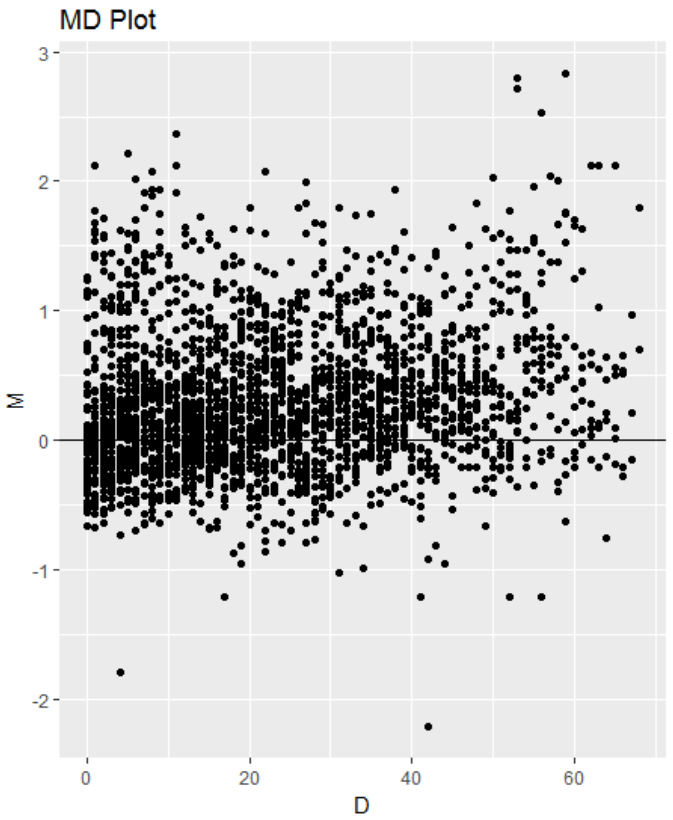
Distance-centric normalization of Hi-C Data

- Hi-C interaction matrix displays the linear genome on the diagonal (self-self interactions)
- Off-diagonal traces represent interaction frequencies between pairs of regions at a given distance
- Due to power-law decay, farther off-diagonal traces are very sparse, have low interaction frequencies



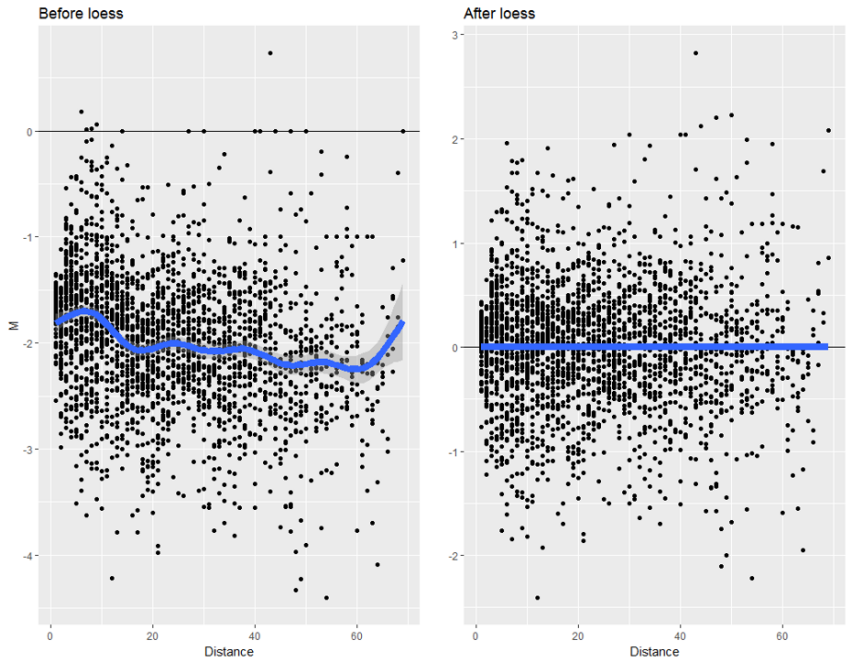
The MD Plot - plotting the differences between two Hi-C datasets at each distance

- The MD plot (Minus vs Distance plot)
- **M** = log2(IF2/IF1) on the Y-axis
- Genomic **D**istance on X-axis
- Similar to the MA plot (Bland-Altman plot)



Biases between two Hi-C datasets are shown as an offset from M=0

- We assume that the differences between two Hi-C datasets should be minimal across all distances (**M** should be ~ 0 across all **D**).
- Deviations from **M** ~ 0 are biases.
- Same principle is used in RNA-seq, ChIP-seq studies.



Joint Normalization of Hi-C Data

- Loess - Local Regression - fit based on local subsets of the data.
- Creates a smooth curve through the data
- Goal is to make the data symmetric around 0 on MD plot

Joint Normalization of Hi-C Data

- Loess technique adjusts the interaction frequencies (IFs) as follows:

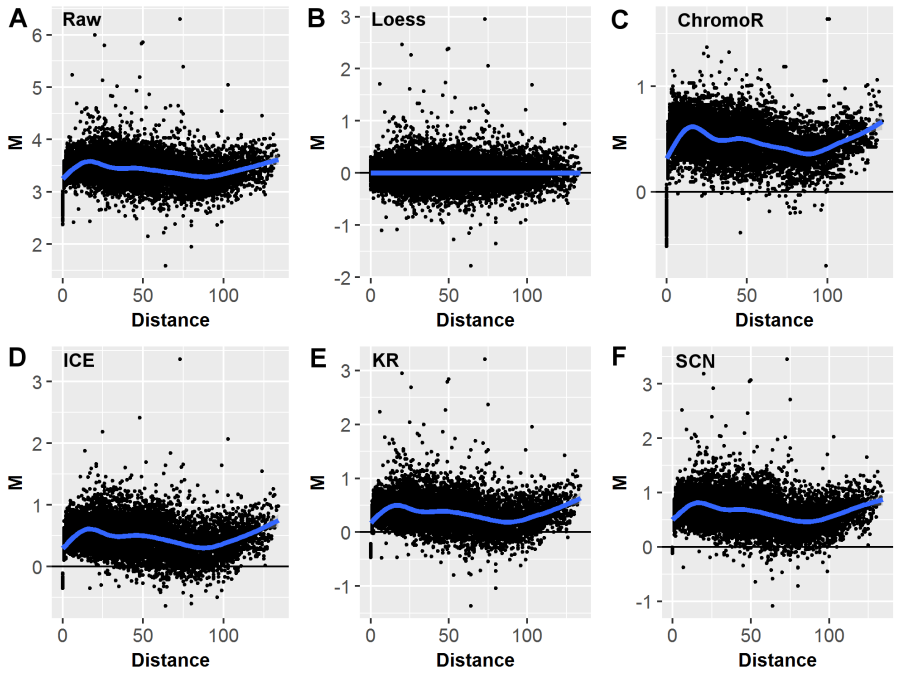
$$\begin{cases} log_2(\hat{IF}_{1D}) = log_2(IF_{1D}) + f(D)/2 \\ log_2(\hat{IF}_{2D}) = log_2(IF_{2D}) - f(D)/2 \end{cases}$$

- $f(D)$ is the predicted value from the loess regression at a distance D.
- The $log_2(\hat{IF})$ are anti-logged to obtain the normalized IF values.
- Average IF for the pair remains unchanged.

Stansfield, John C., Kellen G. Cresswell, Vladimir I. Vladimirov, and Mikhail G. Dozmorov. "HiCcompare: An R-Package for Joint Normalization and Comparison of HI-C Datasets." BMC Bioinformatics 19, no. 1 (December 2018). https://doi.org/10.1186/s12859-018-2288-x.

Joint Loess Normalization of Hi-C Data

- Differences between two datasets should be minimal (symmetric around M = 0, Y-axis)
- Perform loess regression on the MD plot to calculate $f(D)$ - the predicted interaction frequency IF value at distance D
- Adjust interaction frequencies



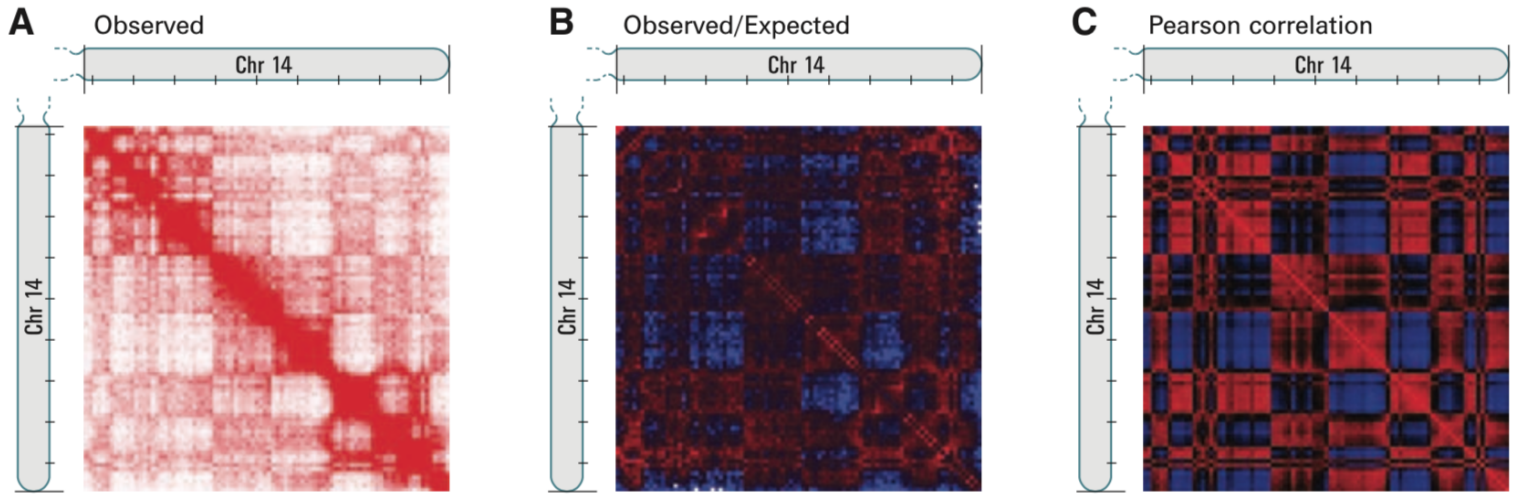
Benchmarking study: Lyu, Hongqiang, Erhu Liu, and Zhifang Wu. "Comparison of Normalization Methods for Hi-C Data" *BioTechniques*, October 7, 2019

A/B compartments detection

- Genomic regions at two distinct nuclear compartments, labeled compartment A and compartment B, display high contact frequency within the same compartment and low contact frequency between the compartments.
  - Compartment A roughly corresponds to the euchromatin and features higher gene density;
  - Compartment B corresponds to the heterochromatin and is largely made up of gene deserts. Also closely correlated with lamina-associated domains (LADs).
- Principal component analysis (PCA) on intra- or inter-chromosomal Hi-C contact maps can be applied to designate compartments A and B.
  - The sign of the first eigenvector, guided by density of genes/active epigenetic marks, determines the compartment label.

Partition the genome into A/B compartments

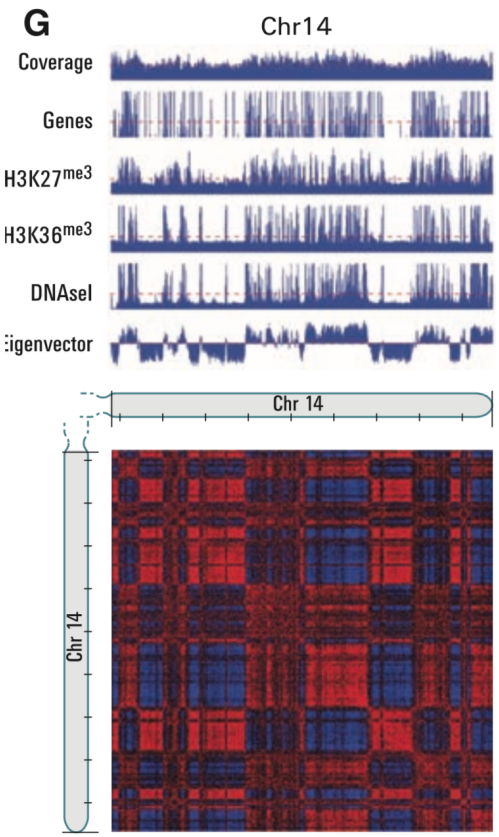
- **O/E matrix** - loci normalized to the average interaction frequency at that distance. Red/blue gradient highlights interactions more/less than expected at a given genomic distance.
- **Pearson matrix** - correlation between the intrachromosomal interaction profiles for every pair of loci.
- The plaid pattern indicates two compartments within the chromosome



Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science (New York, N.Y.) 326, no. 5950 (October 9, 2009): 289–93. https://doi.org/10.1126/science.1181369.

Partition the genome into A/B compartments

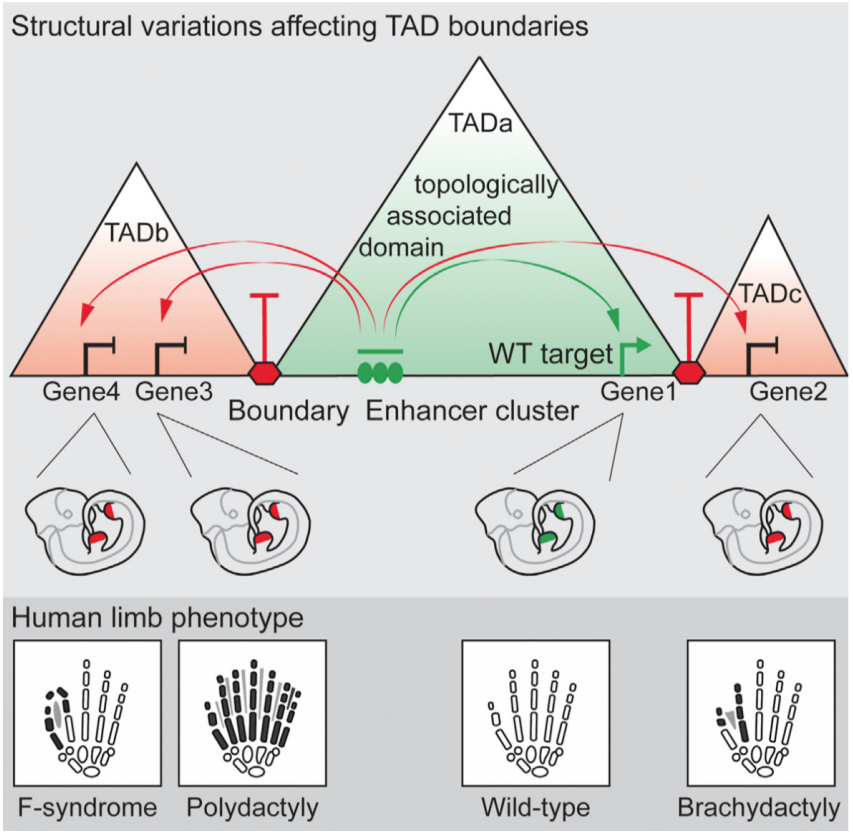
- Plaid patterns were consistently seen for all chromosomes.
- Genomic regions with the same sign of the first eigenvector (from PCA) had correlated contact profiles. The sign is used to assign A and B labels.
- Compartment A correlates strongly with the presence of genes, high gene expression, accessible chromatin marks.



Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science (New York, N.Y.) 326, no. 5950 (October 9, 2009): 289–93. https://doi.org/10.1126/science.1181369.

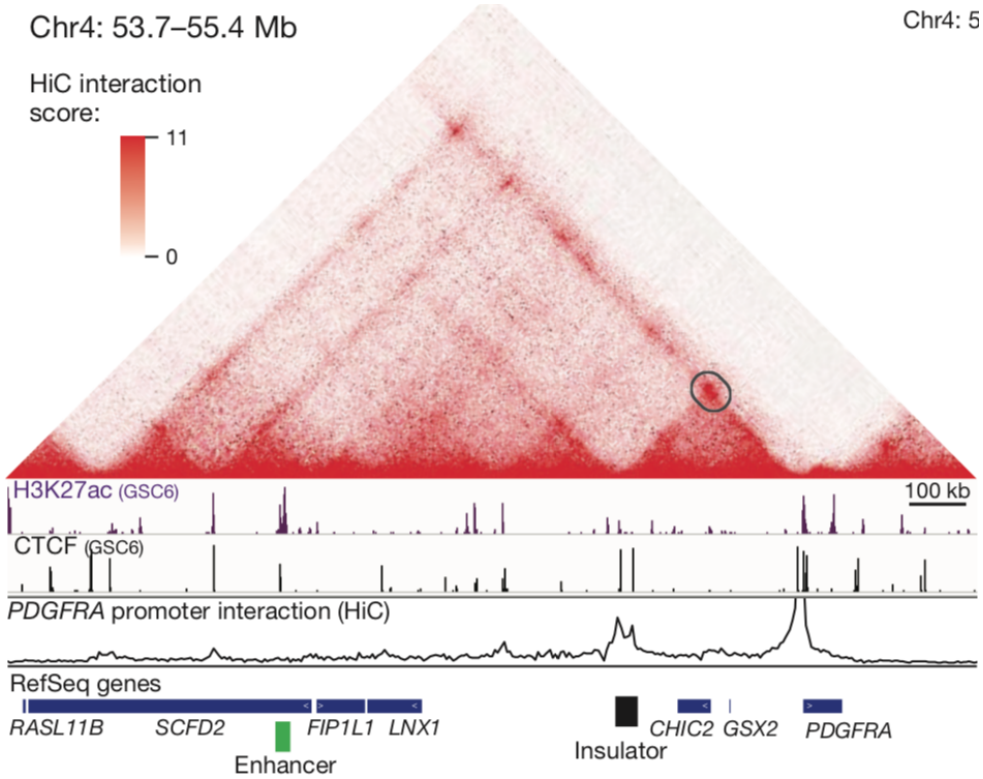
TADs control functional interactions

- Disruptions of TADs lead to de novo enhancer-promoter interactions and misexpression. - Misexpression occurs when CTCF-associated TAD boundary elements are disrupted. - Structural variations disrupting TAD structures can cause malformation syndromes. - Different phenotypes can result from one enhancer acting on different target genes.



Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, et al. "Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions." Cell, 2015. https://doi.org/10.1016/j.cell.2015.04.004.

TADs control functional interactions

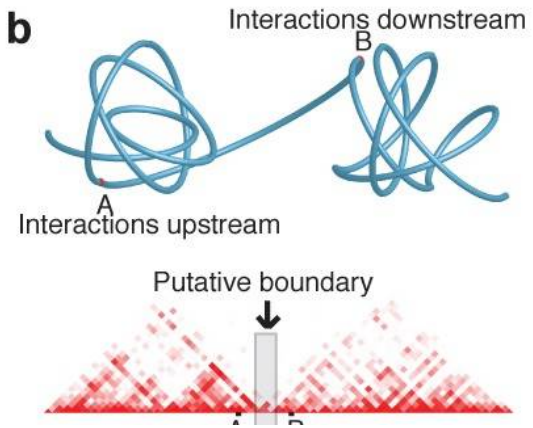


- Insulator loss allows PDGFRA to interact with a constitutive enhancer and become overexpressed.

Flavahan, William A., Yotam Drier, Brian B. Liau, Shawn M. Gillespie, Andrew S. Venteicher, Anat O. Stemmer-Rachamimov, Mario L. Suvà, and Bradley E. Bernstein. "Insulator Dysfunction and Oncogene Activation in IDH Mutant Gliomas." Nature, 2016. https://doi.org/10.1038/nature16490.

TAD boundaries detection

- Topologically associated domains (TADs) are defined as domains of increased self-interaction frequencies.
  - TAD boundaries are devoid of contacts crossing over them.
- TADs and A/B compartments are two different modes of genome organization



Schwarzer, Wibke, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A. Fonseca, et al. "Two Independent Modes of Chromatin Organization Revealed by Cohesin Removal." Nature 551, no. 7678 (02 2017): 51–56. https://doi.org/10.1038/nature24281

TAD boundary detection

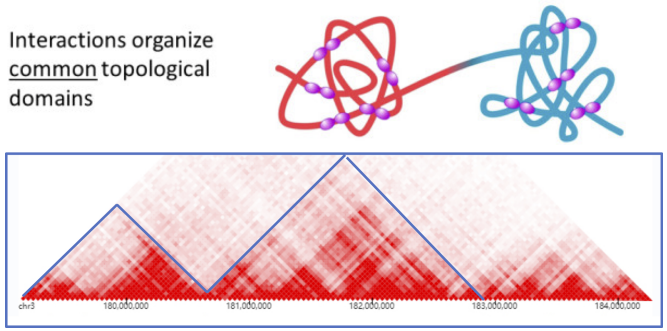
1) Methods that scan the genome for domain boundaries which are characterized by certain local genomic or topological features
   - **Directionality index (DI)** - quantifies the degree of up/downstream bias of Hi-C read pairs at given genomic loci
   - **Insulation score (IS)** - quantifies the level of interactions across given genomic loci
   - **Arrowhead score** - quantifies the likelihood that a spot will present as a corner of dark squares in the heatmap of a Hi-C contact matrix
   - Other methods: `IC-Finder`, `ClusterTAD`

J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, B. Ren, Topological domains in mammalian genomes identified by analysis of chromatin interactions, Nature 485 (7398) (2012) 376–380.

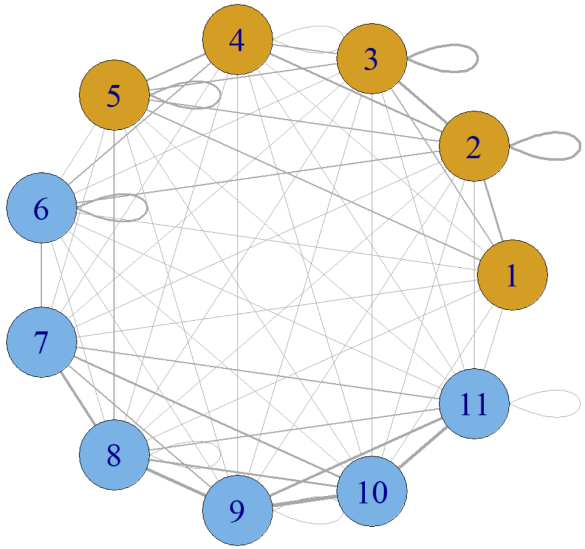
E. Crane, Q. Bian, R.P. McCord, B.R. Lajoie, B.S. Wheeler, E.J. Ralston, S. Uzawa, J. Dekker, B.J. Meyer, Condensin-driven remodelling of X chromosome topology during dosage compensation, Nature 523 (7559) (2015) 240–244.

S.S. Rao, M.H. Huntley, N.C. Durand, E.K. Stamenova, I.D. Bochkov, J.T. Robinson, A.L. Sanborn, I. Machol, A.D. Omer, E.S. Lander, E.L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell 159 (7) (2014) 1665–1680.

Spectral clustering for TAD boundary detection



- Strong interactions within TADs can be seen as color-dense triangles off the diagonal.
- Interacting regions can be represented as a graph of vertices V connected by edges E.
- Edges E are weighted by interaction frequencies.



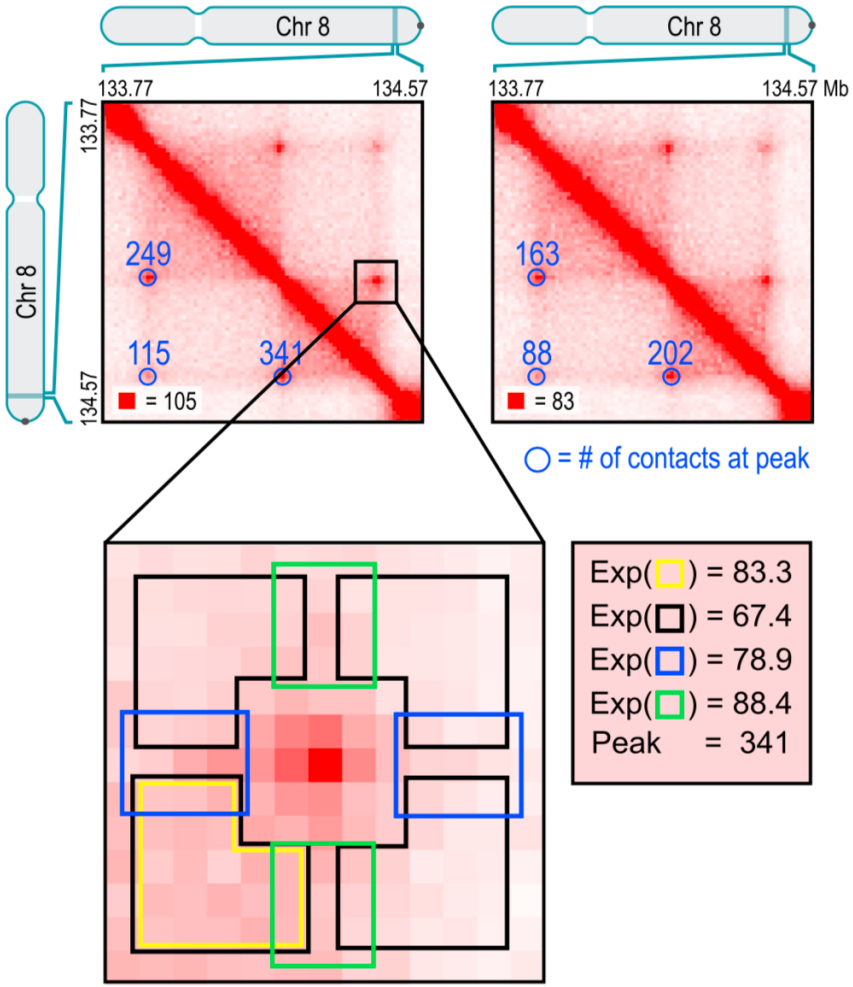
- Spectral clustering is designed to cluster graphs.
- Works by projecting the data into a lower dimensional space.
- Excels on noisy (Hi-C) and non-normally distributed data

https://bioconductor.org/packages/SpectralTAD/

Detecting chromatin loops

- Hi-C maps not only provide large scale information about A/B compartments and TADs but can also detect loops between specific loci.
- The challenge is to distinguish true contacts from noise and define a background model that suitably represents the background frequency of interaction between pairs of loci in the absence of specific interactions.
- Need to account for the distance-dependent signal decay.

Detecting chromatin loops



HiCCUPS

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

Detecting chromatin loops

- Idea: perform a parametric fit to best describe the data or bin all pairs of loci with the same genomic distance and then, compute a P-value comparing the observed count for a given contact as compared with all other possible interactions in that bin (Duan et al. 2010).
  - Background models can also take additional biases such as domain organization into account (HOMER, Fit-Hi-C)
- The HiCCUPS method identifies loops by seeking enrties substantially larger than surrounding entries (Rao et al. 2014).

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping". Cell, 2014. https://doi.org/10.1016/j.cell.2014.11.021.

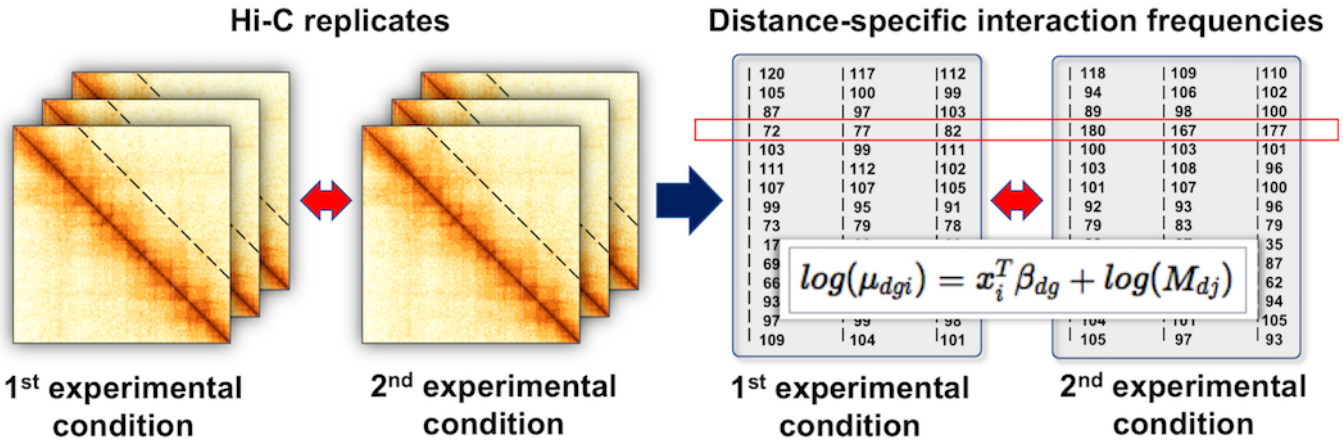
http://homer.salk.edu/homer/, https://noble.gs.washington.edu/proj/fit-hi-c/

Methods for detecting chromatin loops

- The `HiCCUPS` algorithm detects chromatin interaction "hotspots" within a Hi-C contact map and then can overlap hotspots between datasets.
- The `diffHic` method is an extension of `edgeR` as a Hi-C data processing pipeline.
- `chromoR` is an R package with functions for wavelet based normalization and a Poisson change point detection algorithm for difference detection.
- `FIND` is an R package that uses a spatial Poisson process to detect differences in high resolution Hi-C datasets by treating interactions as spatially dependent on neighboring interactions.

multiHiCcompare method for differential loop detection

- **Distance-centric analysis** – each off-diagonal data slice has unique statistical properties
- Split Hi-C data into *d* distance-centric matrices with *g* rows (indices for interacting pairs of regions) and *i* columns (samples)



https://bioconductor.org/packages/multiHiCcompare/

Benchmarking study: Zheng, Ye, Peigen Zhou, and Sündüz Keleş. "FreeHi-C Spike-in Simulations for Benchmarking Differential Chromatin Interaction Detection" *Methods*, July 12, 2020

Summary of Hi-C methods for TAD/loop identification

Table 1
Methods for TAD identification mentioned in this paper.

Category name	TAD caller name	References
Genomic or topological features	Domain Call	[13]
	Insulation Score (IS)	[38]
	HOMER	[35,36]
	Arrowhead	[17]
	IC-Finder	[39]
	ClusterTAD	[40]
Probabilistic model	GMAP	[41]
	PSYCHIC	[42]
	HiCseg	[43]
	TADbit	[44]
	TADtree	[45]
Dynamic programming	**Armatus**	[46]
	Matryoshka	[47]
Graph model	**MrTADFinder**	[48]
	3DnetMod	[49]
	deDoc	[30]
	Rajapakse	[50]

Table 2
Methods for chromatin loop identification mentioned in this paper.

Significant/differential loop	Loop caller names	References
Significant loop	**HiCCUPS**	[17]
	cLoops	[56]
	Fit-Hi-C	[52]
	Jin et al.	[53]
	CHICAGO	[54]
	HiC-DC	[55]
	PSYCHIC	[42]
	HMRF	[57]
	FastHiC	[58]
Differential loop	**HOMER**	[36]
	diffHiC	[63]
	FIND	[64]
	HiCcompare	[65]

straightforward algorithm and has been included in many Hi-C analysis packages, e.g., Juicer and HOMER [35,36]. Some alternative definitions

Li, Xiao, Ziyang An, and Zhihua Zhang. "Comparison of Computational Methods for 3D Genome Analysis at Single-Cell Hi-C Level." Methods, August 2019, S1046202319300891. https://doi.org/10.1016/j.ymeth.2019.08.005.

3D genome reconstruction

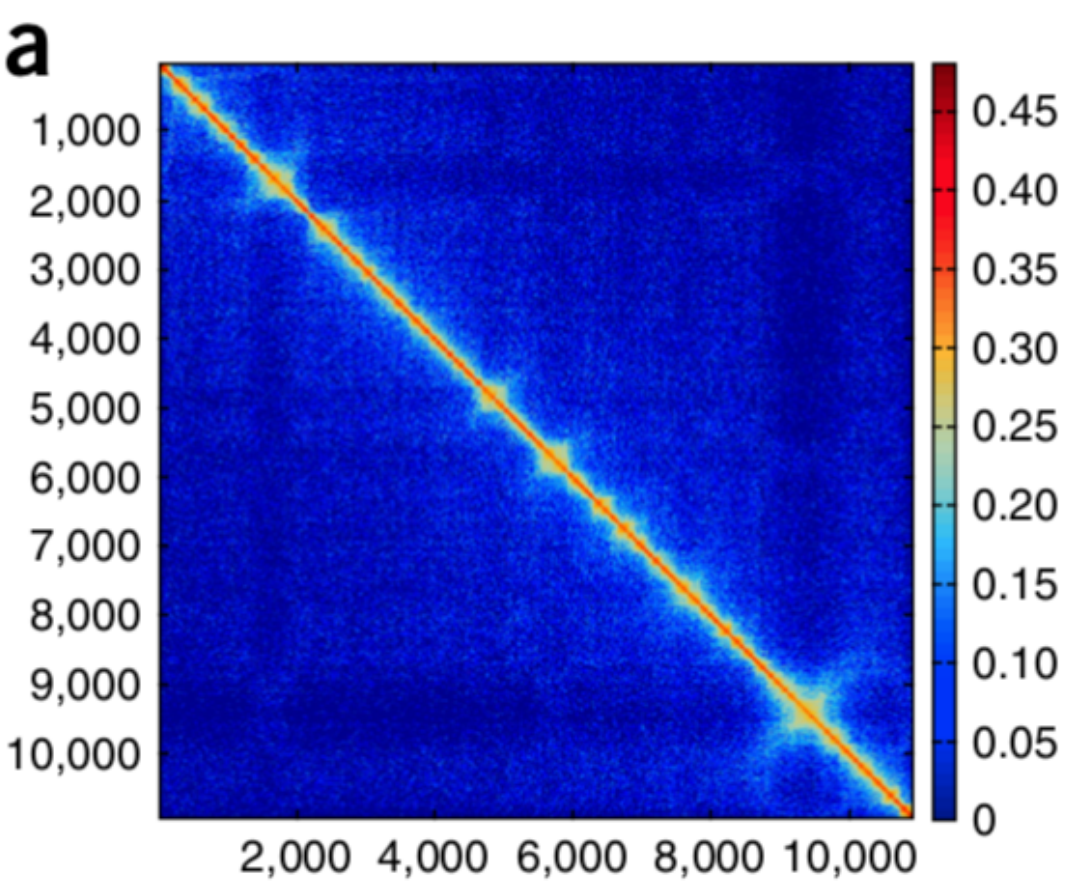
Given a map of self-contacts, how can you reconstruct the structures that produced it?

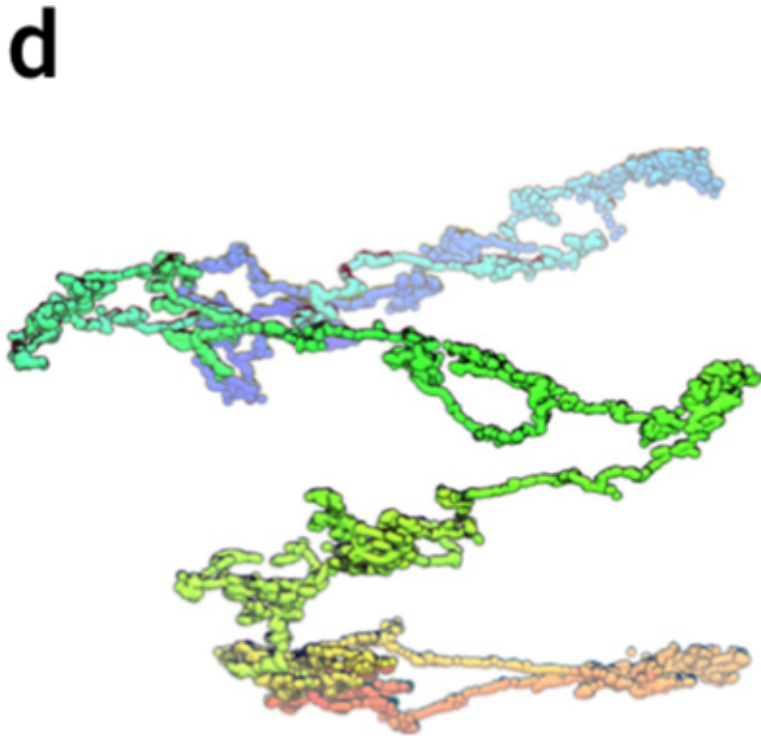
- Two different computational approaches for 3D modeling:
  - Simulations of polymer physics, an approach that has been referred to as direct, as it does not rely on indirect methods such as chromatin conformation capture.
  - The inverse or restraint-based approach, in which interaction probabilities derived from chromatin contact maps are used as restraints that are implemented in a scoring function, thereby basically constituting a computational optimization approach (Rosa and Zimmer 2014).

Dekker, Job, Marc A. Marti-Renom, and Leonid A. Mirny. "Exploring the Three-Dimensional Organization of Genomes: Interpreting Chromatin Interaction Data". Nature Reviews, 2013. https://doi.org/10.1038/nrg3454.

Serra, Francois, Marco Di Stefano, Yannick G. Spill, Yasmina Cuartero, Michael Goodstadt, Davide Bas, and Marc A. Marti-Renom. "Restraint-Based Three-Dimensional Modeling of Genomes and Genomic Domains". FEBS Letters, 2015. https://doi.org/10.1016/j.febslet.2015.05.012.

3D genome reconstruction (ShRec3D)



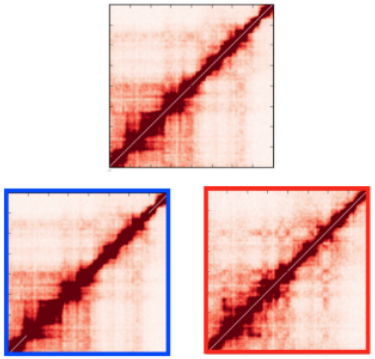


Lesne, Annick, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. "3D Genome Reconstruction from Chromosomal Contacts." Nature Methods 11, no. 11 (November 2014): 1141–43. https://doi.org/10.1038/nmeth.3104.

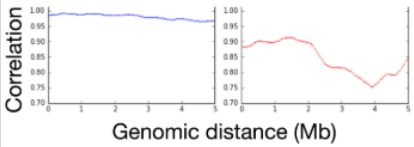
Reproducibility

HiCRep

Transformation: 2D mean filter



Comparison: weighted sum of correlation coefficients stratified by distance



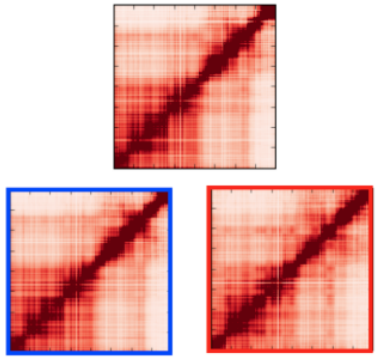
Reproducibility score:

$$\sum_d w_d \cdot \rho_d$$

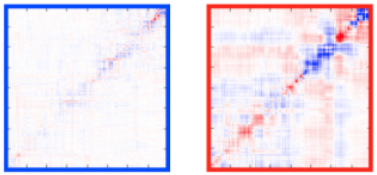
weight, correlation

GenomeDISCO

Transformation: smoothing using graph diffusion



Comparison: difference in smoothed contact maps

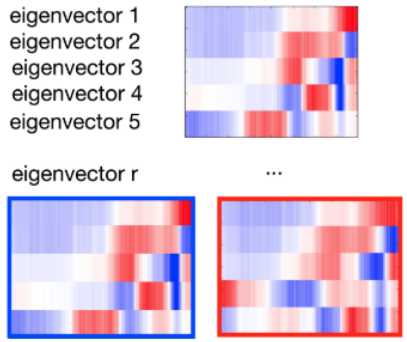


Reproducibility score:

$$\frac{\sum_i \sum_j |rw(A)_{ij} - rw(B)_{ij}|}{\#\ genomic\ bins}$$

HiC-Spector

Transformation: eigen-decomposition of Laplacian



Comparison: weighted difference of eigenvectors

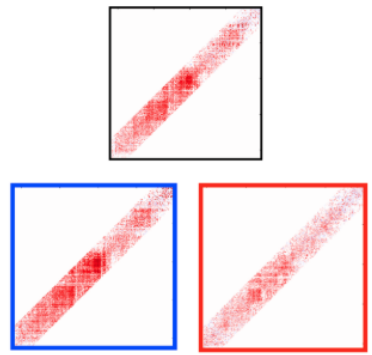
$$S_d(A,B) = \sum_{i=0}^{r-1} \|v_i^A - v_i^B\|$$

Reproducibility score:

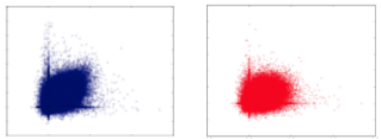
$$\left(1 - \frac{1}{r}\frac{S_d}{l}\right) \qquad l = \sqrt{2}$$

QuASAR-Rep

Transformation: correlation matrix of distance-based contact enrichment



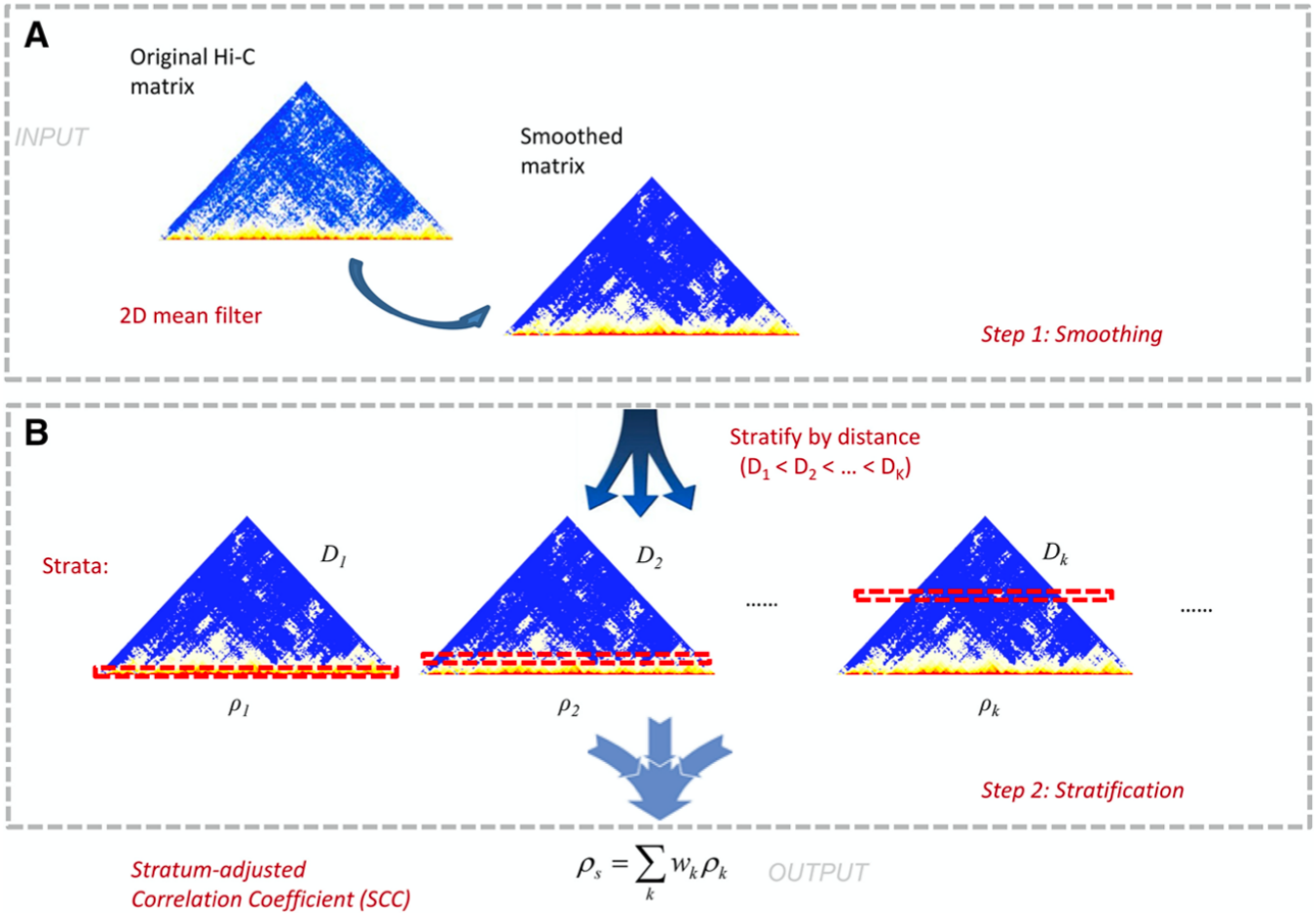
Comparison: compute correlation of values in the 2 transformed matrices



Reproducibility score:

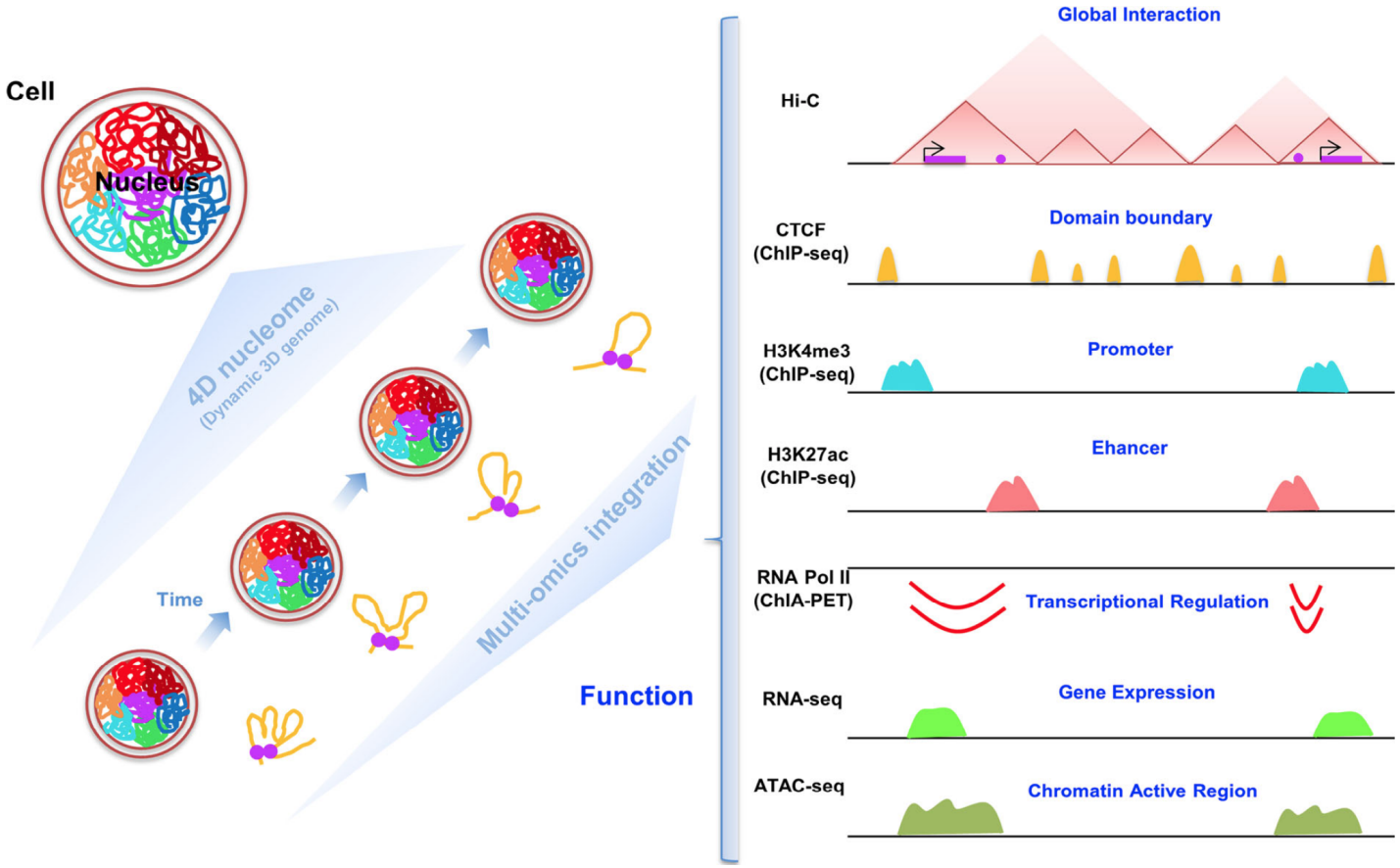
$$Pearson\ correlation$$
$$(\ quasar(A),\ quasar(B)\ )$$

HiCRep



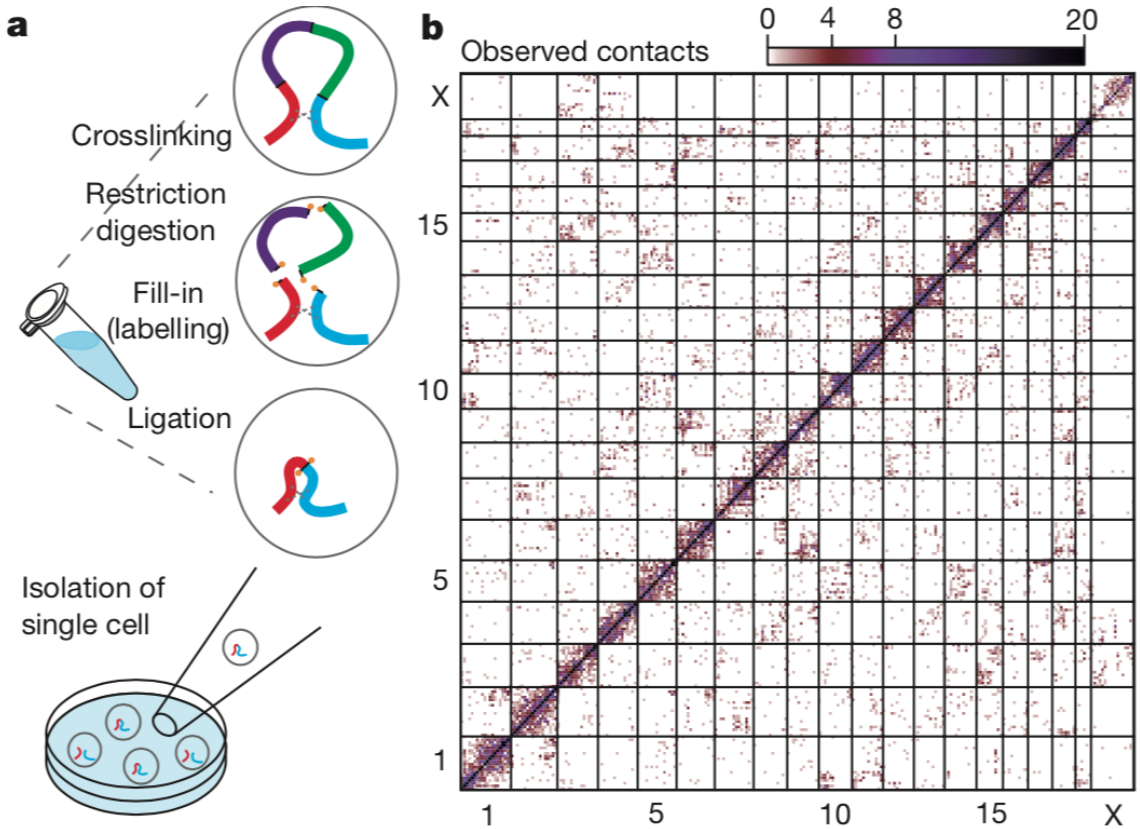
Yang, Tao, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, and Qunhua Li. "HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient." Genome Research, August 7, 2017, https://genome.cshlp.org/content/early/2017/10/06/gr.220640.117

Multi-omics integration



Kong, Siyuan, and Yubo Zhang. "Deciphering Hi-C: From 3D Genome to Function." Cell Biology and Toxicology, January 4, 2019. https://doi.org/10.1007/s10565-018-09456-2.

Single-cell Hi-C

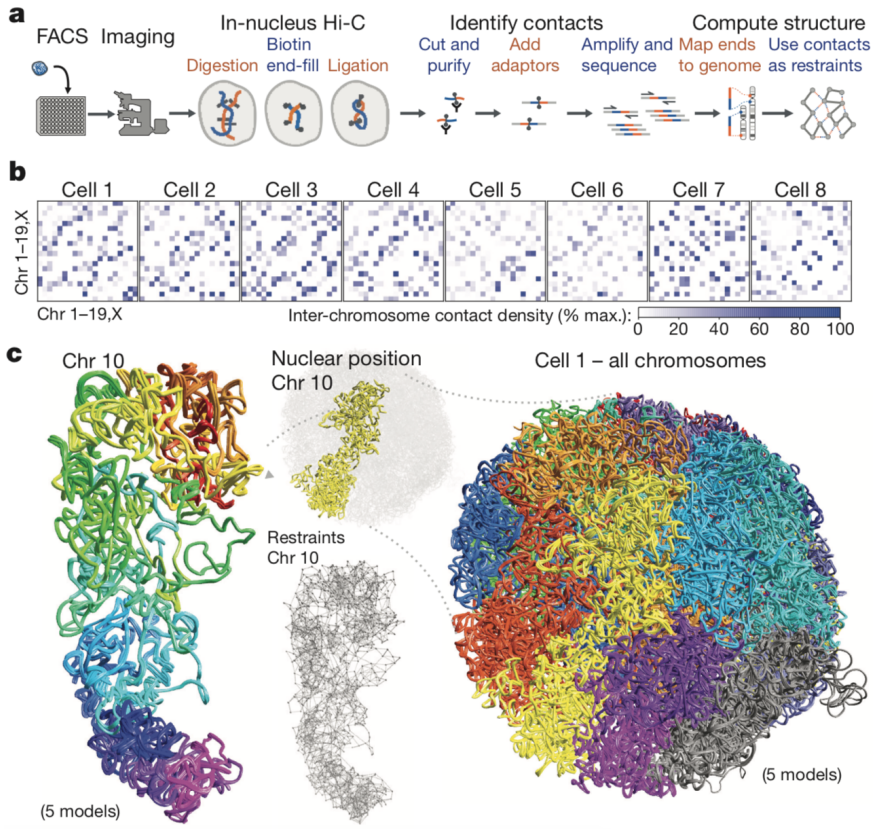


- Several types of restriction enzymes for more even cutting

Nagano,T.et al. "Single-cell Hi-C reveals cell-to-cell variability in chromosome structure". Nature, 2013.

Single-cell modeling

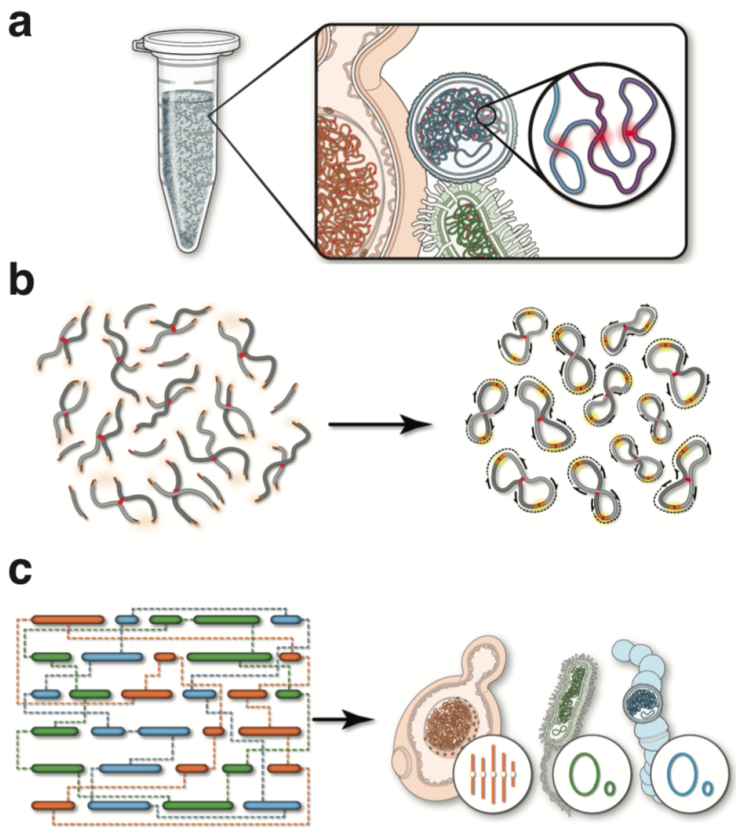
- With single-cell experiments, data is much more sparse, but corresponds to a unique structure.



Stevens, Tim J., David Lando, Srinjan Basu, Liam P. Atkinson, Yang Cao, Steven F. Lee, Martin Leeb, et al. "3D Structures of Individual Mammalian Genomes Studied by Single-Cell Hi-C." Nature, March 13, 2017. https://doi.org/10.1038/nature21429.

Other use of Hi-C: Genome reassembly

Reconstruction of microbial genomes



Press, Maximilian O., Andrew H. Wiser, Zev N. Kronenberg, Kyle W. Langford, Migun Shakya, Chien-Chi Lo, Kathryn A. Mueller, Shawn T. Sullivan, Patrick S. G. Chain, and Ivan Liachko. "Hi-C Deconvolution of a Human Gut Microbiome Yields High-Quality Draft Genomes and Reveals Plasmid-Genome Interactions.," October 5, 2017. https://doi.org/10.1101/198713.

Schematic of the ProxiMeta method for metagenomic deconvolution. a) First, a sample consisting of mixed organisms is

Many more methods, technologies, discoveries

https://github.com/mdozmorov/HiC_tools