

Discovering structural units of chromosomal organization with matrix factorization and graph regularization

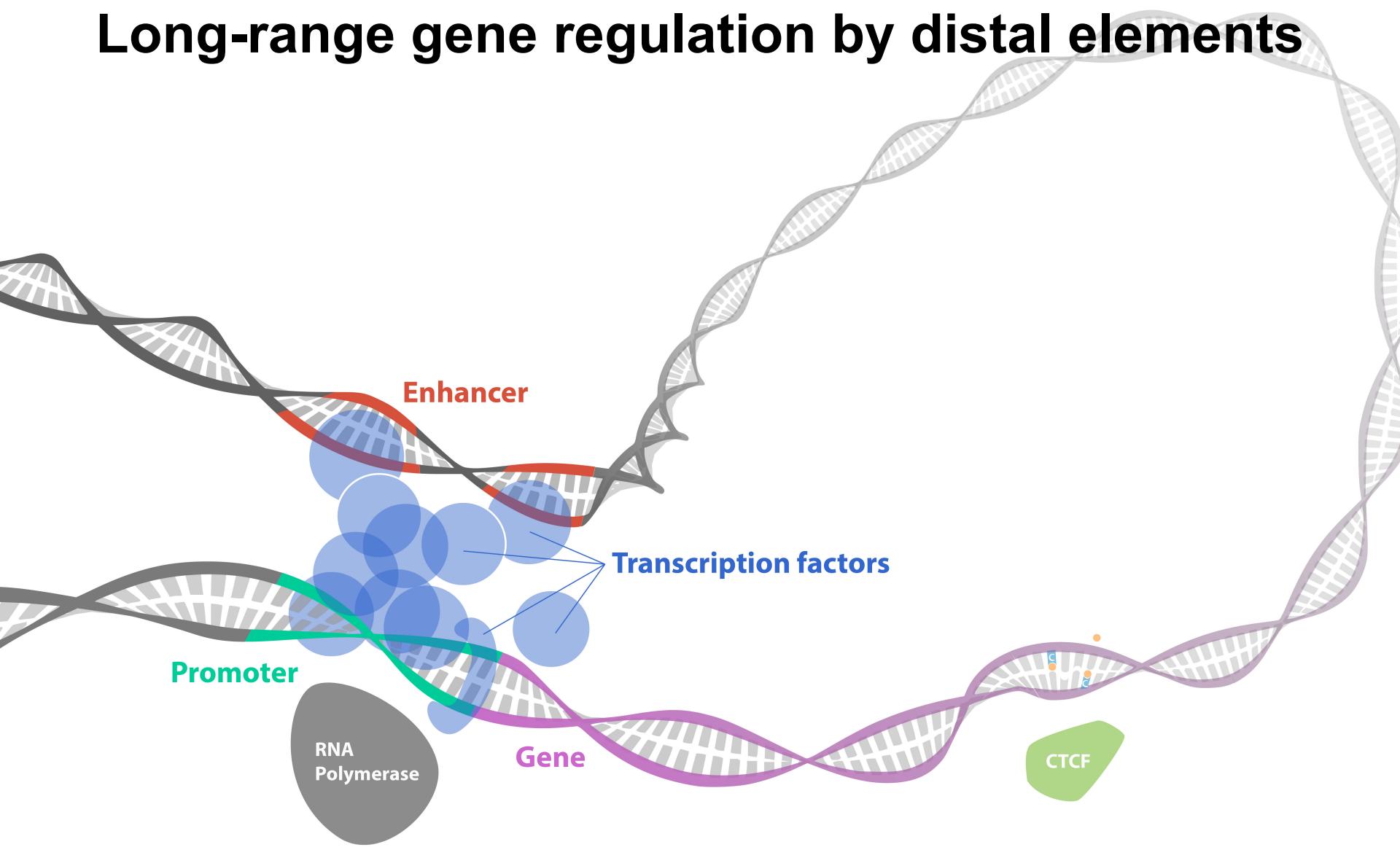
Da-Inn Lee¹ and Sushmita Roy^{1,2}

¹ Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

² Wisconsin Institute of Discovery

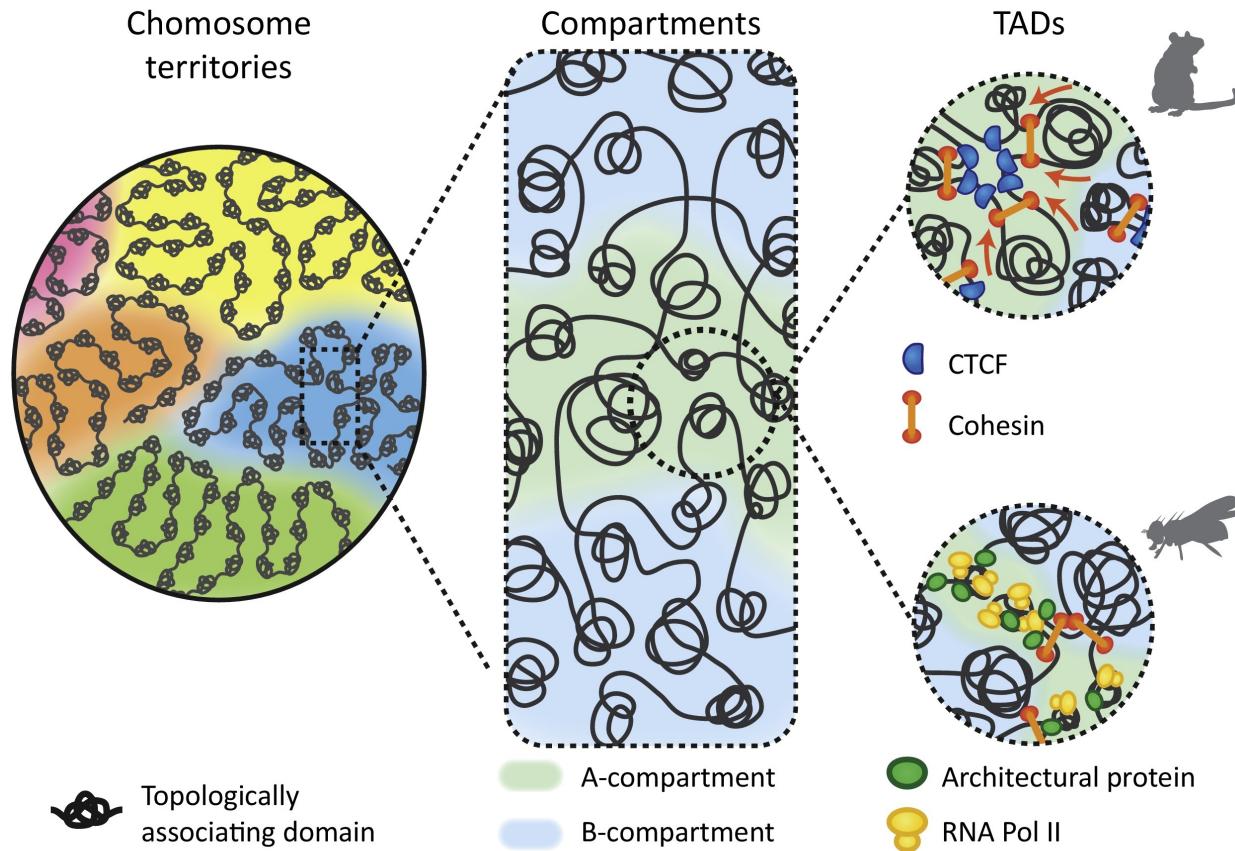


Long-range gene regulation by distal elements

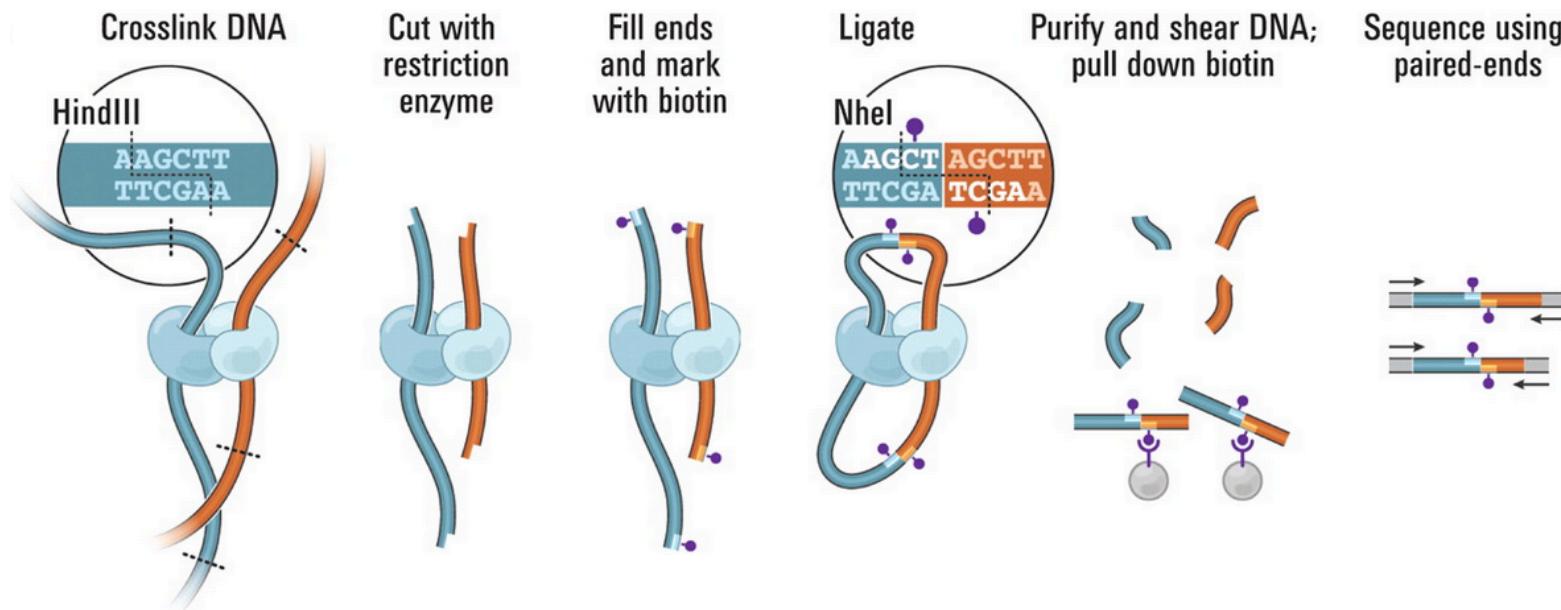


Adapted from illustration by Kelvin Ma

Genome is organized into higher-order domains at multiple scales

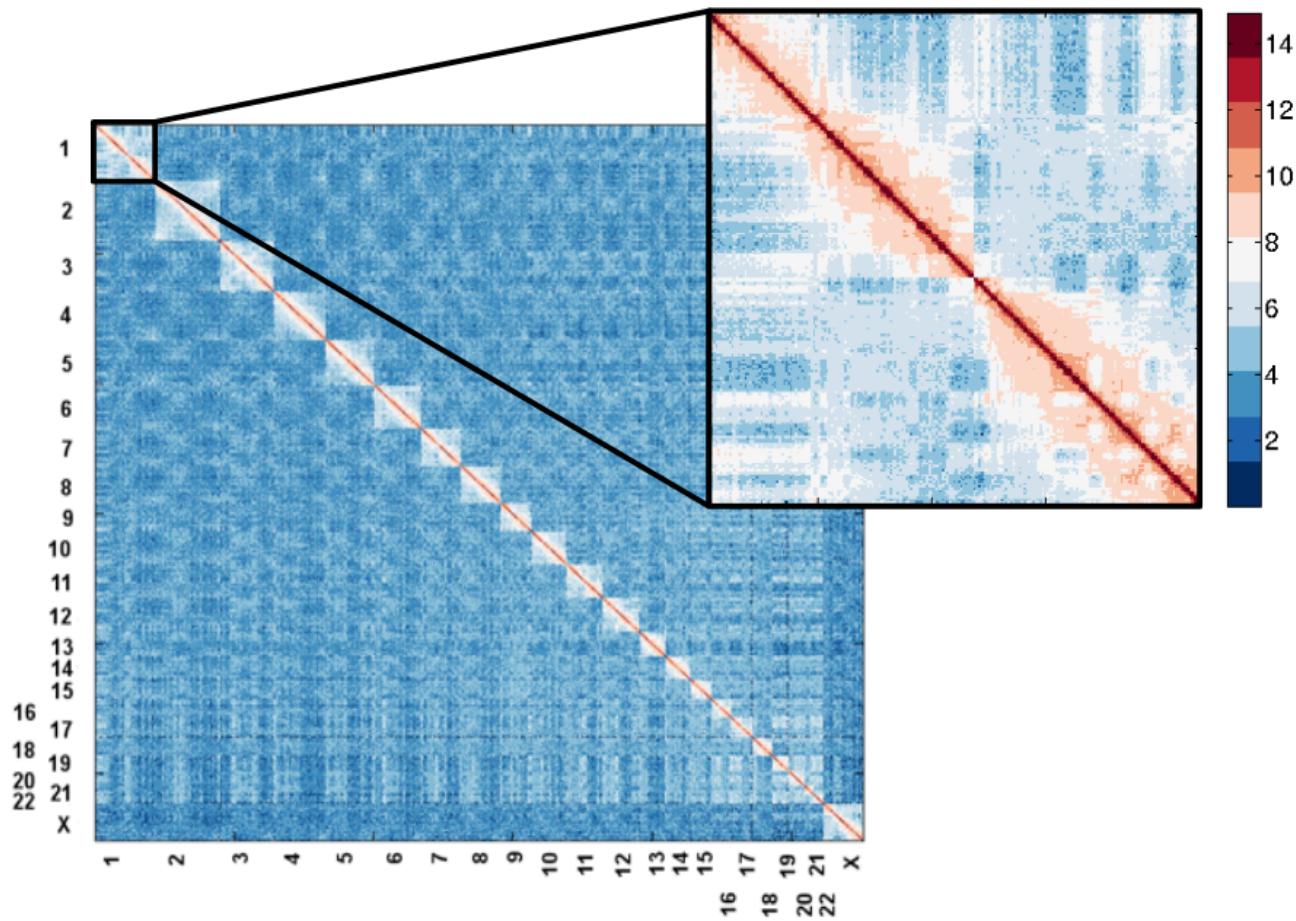


Hi-C enables study of 3D genome through pairwise interactions

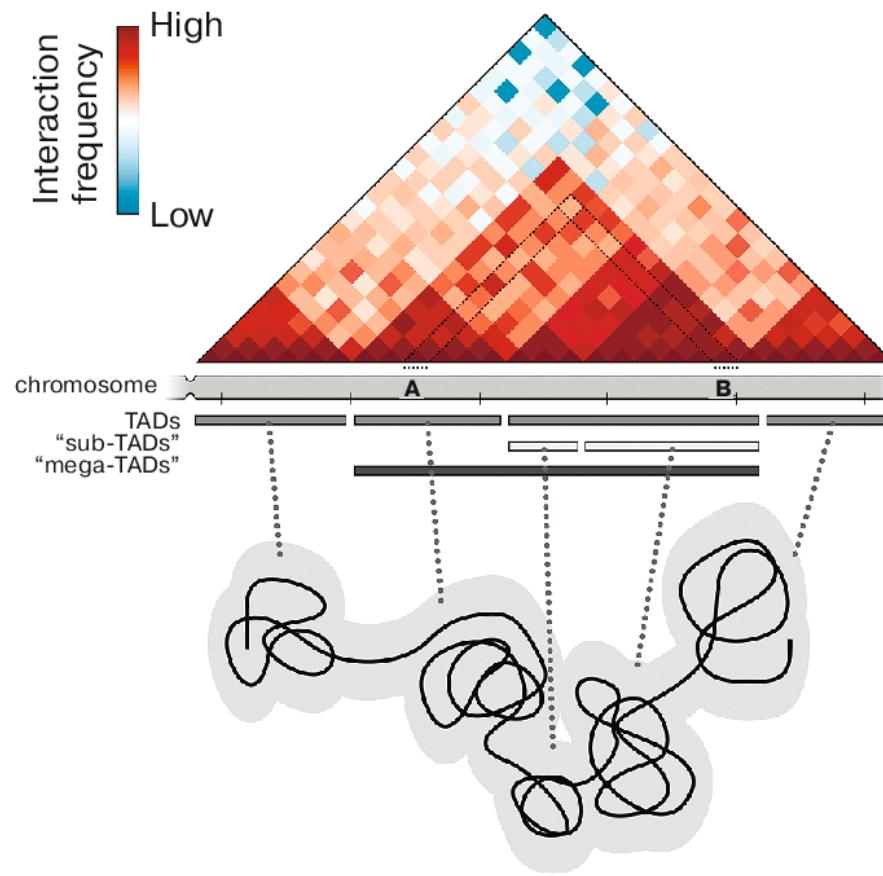


Lieberman-Aiden et al. Science. 2009

Hi-C enables study of 3D genome through pairwise interactions

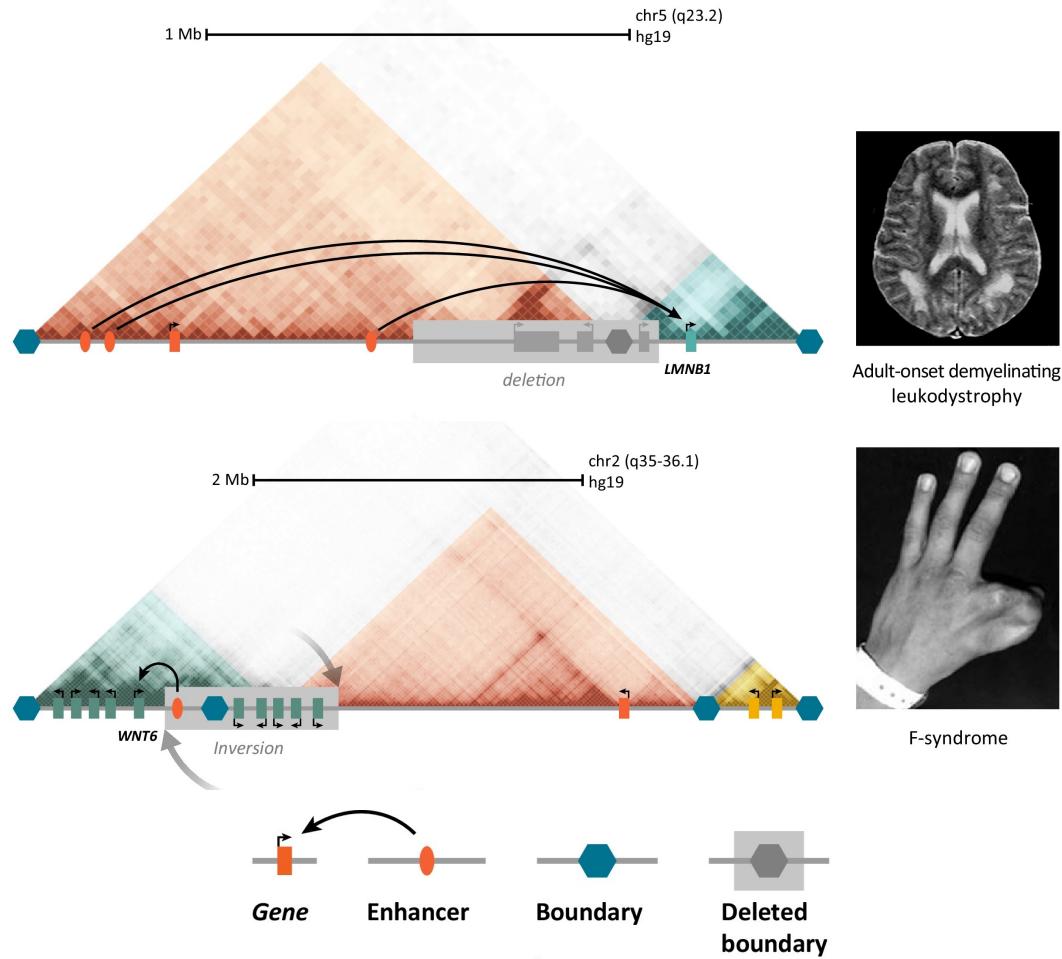


TADs are one of the dominant structural units observed in Hi-C matrices



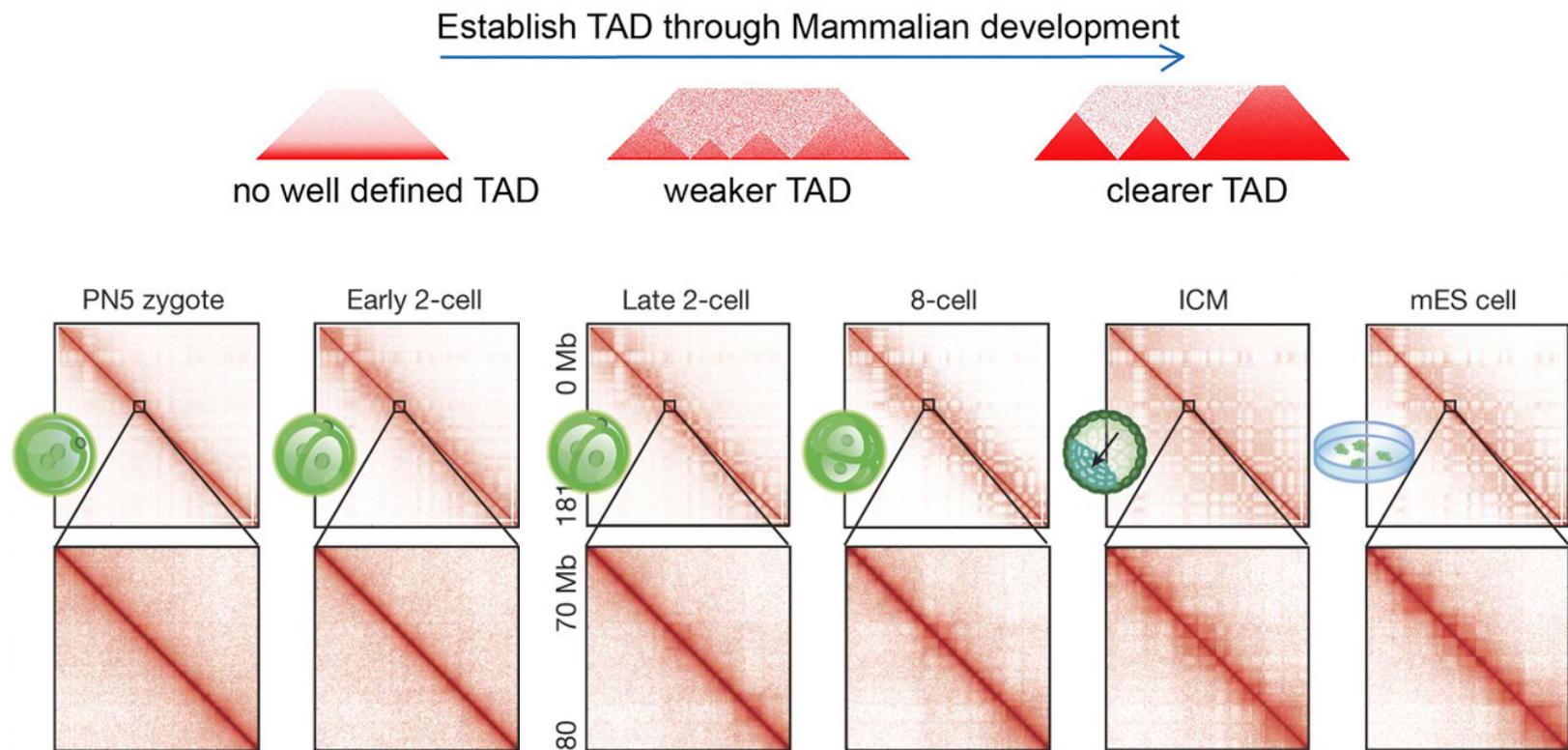
Razin and Gavrilov. Biochemistry (Moscow). 2018

Breaking TADs: alternations in chromatin domains can result in disease



Lupiáñez et al. Trends Genet. 2016

Structural reprogramming during early mammalian development



Key et al. Cell. 2017 and Du et al. Nature. 2017

Existing methods for finding topological units of chromosomes

Method	Algorithm	Objective
Directionality (Dixon et al. 2012)	HMM	Find a Gaussian mixture model, where a bin can take on one of three states: upstream biased, downstream biased, or not biased
Armatus (Filippova et al. 2014)	Dynamic programming	Find densest subgraph in a network where nodes = genomic regions, edge weights = interaction counts
HiCseg (Lévy-Leduc et al. 2014)	Dynamic programming	Find blocks/contours in Hi-C matrix whose count means form a maximum likelihood mixture model
Insulation Score (Crane et al. 2015)	Aggregation, ratio calculation	Find genomic regions with minimal insulation score, calculated as the mean of interaction counts centered on given region
TopDom (Shin et al. 2016)	Smooth curve fitting	Similar to insulation score
rGMAP (Yu et al. 2017)	Gaussian mixture model	Find a two-component Gaussian mixture model to group interactions to intra-domain and inter-domain contacts
3DNetMod (Norton et al. 2018)	Louvain-like algorithm	Partition a network of genomic regions to maximize modularity

Studies have shown existing methods' sensitivity to low depth and sparsity

nature | methods

Analysis | Published: 12 June 2017

Comparison of computational methods for Hi-C data analysis

Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari ✉ & Silvio Bicciato ✉

Nature Methods 14, 679–685 (2017) | Download Citation ↴

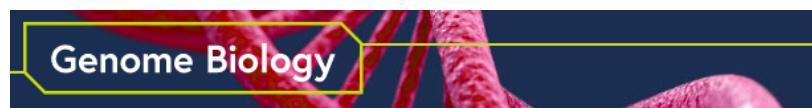
Nucleic Acids Research

A critical assessment of topologically associating domain prediction tools ⓘ

Rola Dali, Mathieu Blanchette ✉

Nucleic Acids Research, Volume 45, Issue 6, 7 April 2017, Pages 2994–3005,
<https://doi.org/10.1093/nar/gkx145>

Published: 02 March 2017 Article history ▾



Research | Open Access

Comparison of computational methods for the identification of topologically associating domains

Marie Zufferey †, Daniele Tavernari †, Elisa Oricchio and Giovanni Ciriello ✉ ⓘ

†Contributed equally

Genome Biology 2018 19:217

<https://doi.org/10.1186/s13059-018-1596-9> | © The Author(s). 2018

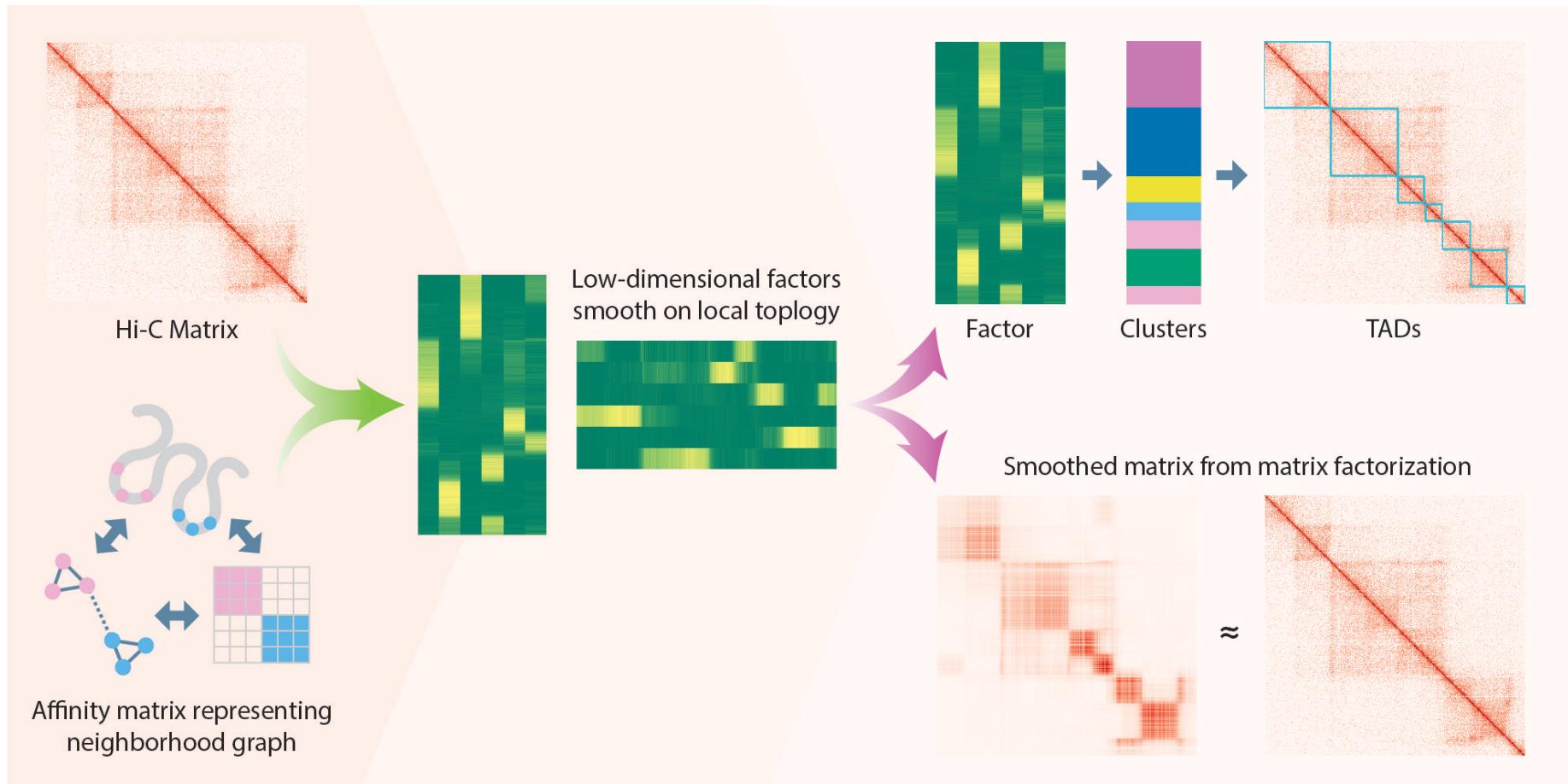
Received: 17 May 2018 | Accepted: 26 November 2018 | Published: 10 December 2018

GRiNCH discovers topological units of chromosomes from Hi-C data

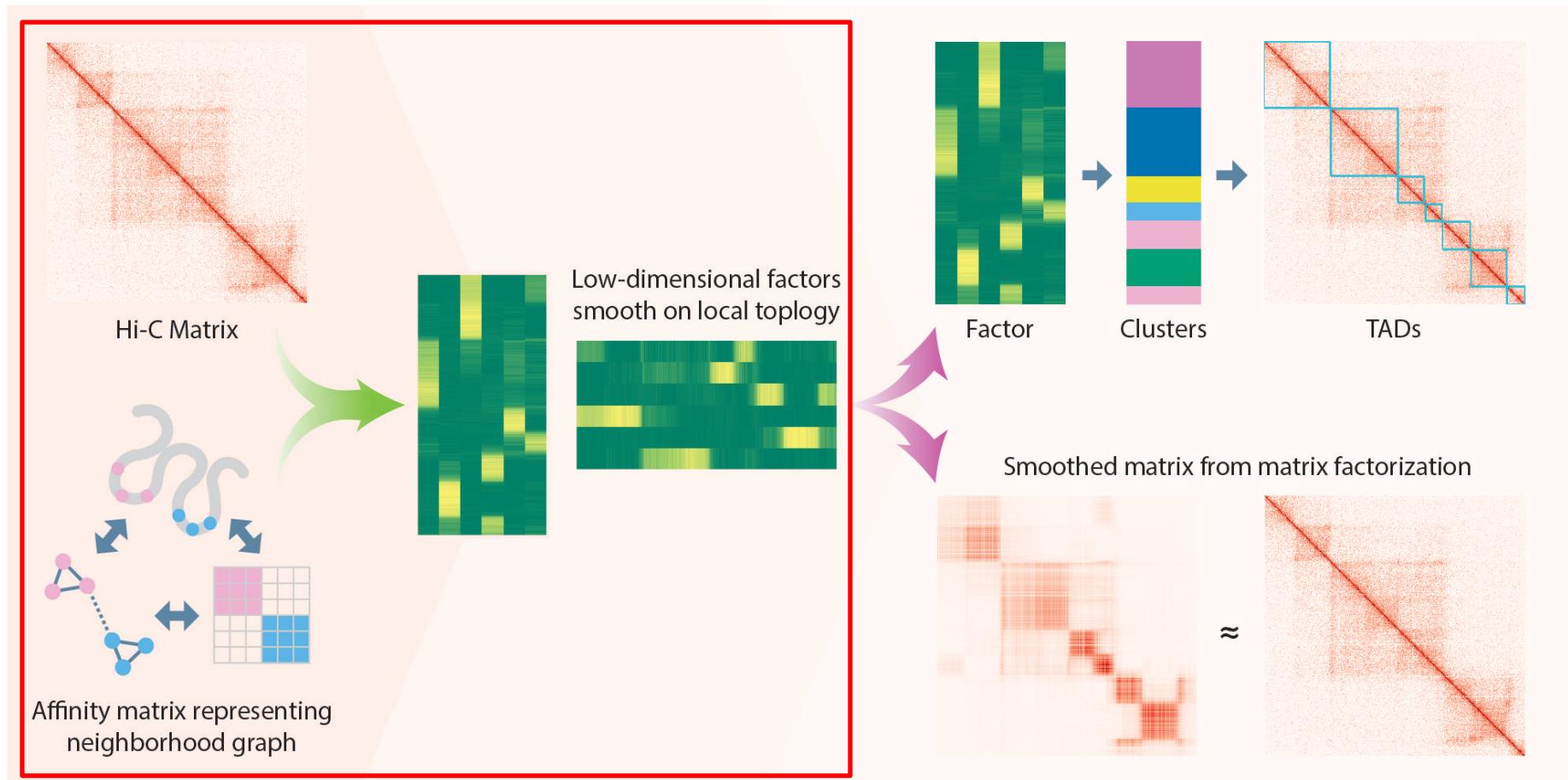


- ◆ Non-negative factorization (NMF)
- ◆ Graph regularization
- ◆ Chain-constrained k-medoids clustering

GRiNCH: Graph Regularized NMF and Clustering for Hi-C



GRiNCH: Graph Regularized NMF and Clustering for Hi-C



Non-negative matrix factorization (NMF) reduces dimensions of data

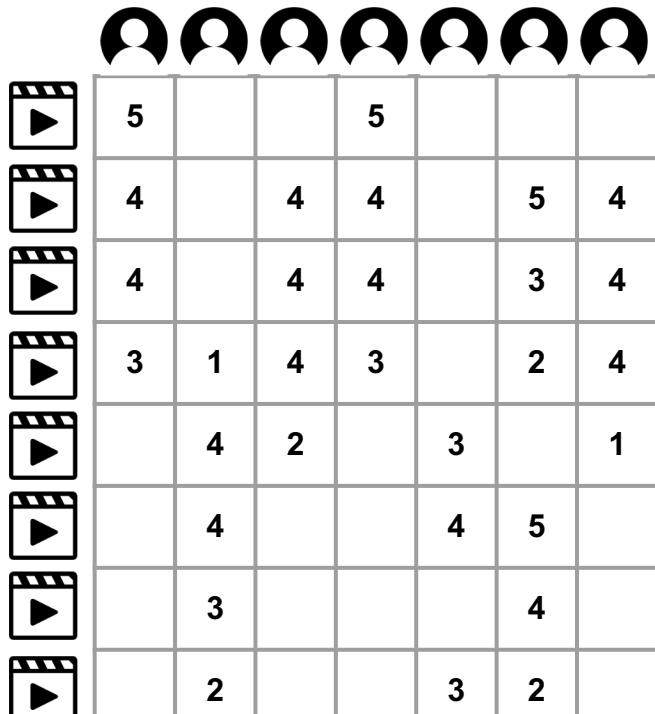


A 7x7 matrix representing a non-negative matrix. The columns represent users and the rows represent movies. The values in the matrix range from 0 to 5, indicating the rating given by each user to each movie.

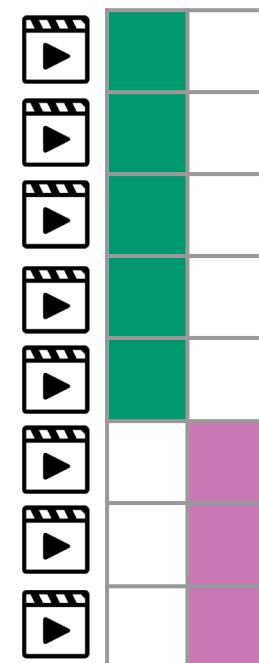
	5			5		
	4		4	4		5
	4		4	4		3
	3	1	4	3		2
		4	2		3	
		4			4	5
		3			4	
		2			3	2

$$X = \mathbb{R}^{n \times m}$$

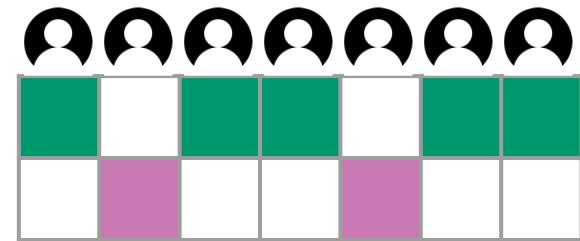
Non-negative matrix factorization (NMF) reduces dimensions of data



$$X = \mathbb{R}^{n \times m}$$



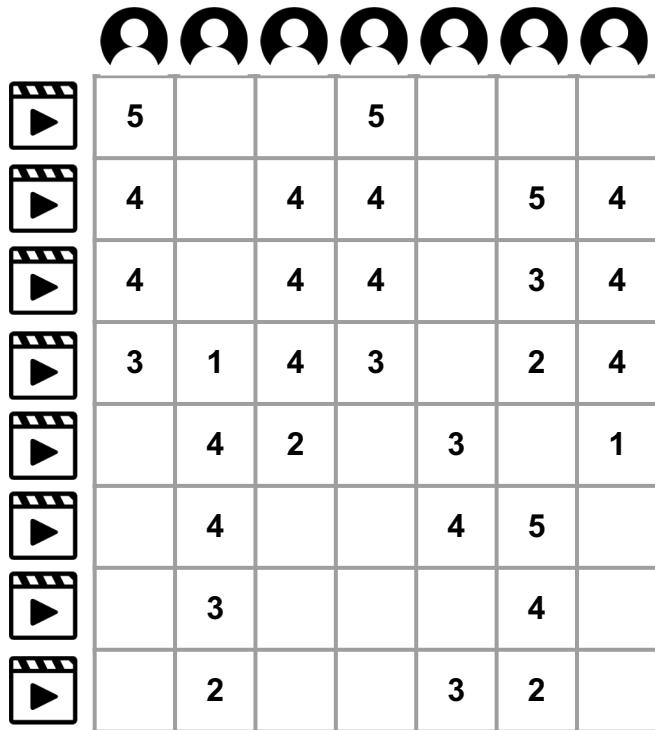
$$U = \mathbb{R}^{n \times k}$$



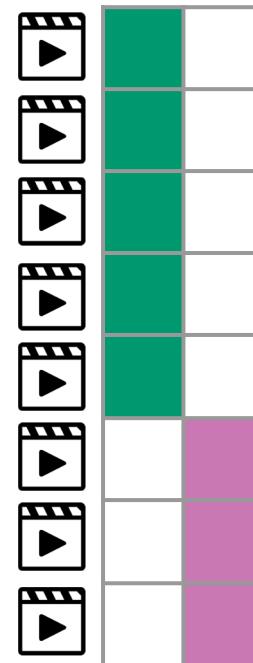
$$V^T = \mathbb{R}^{k \times m}$$

$$k \ll n, m$$

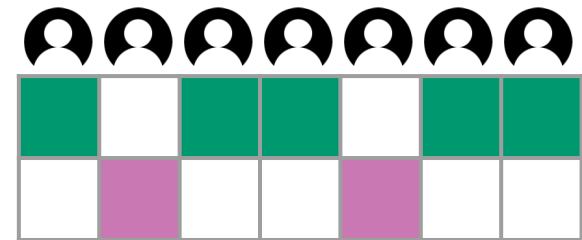
Non-negative matrix factorization (NMF) reduces dimensions of data



$$X = \mathbb{R}^{n \times m}$$



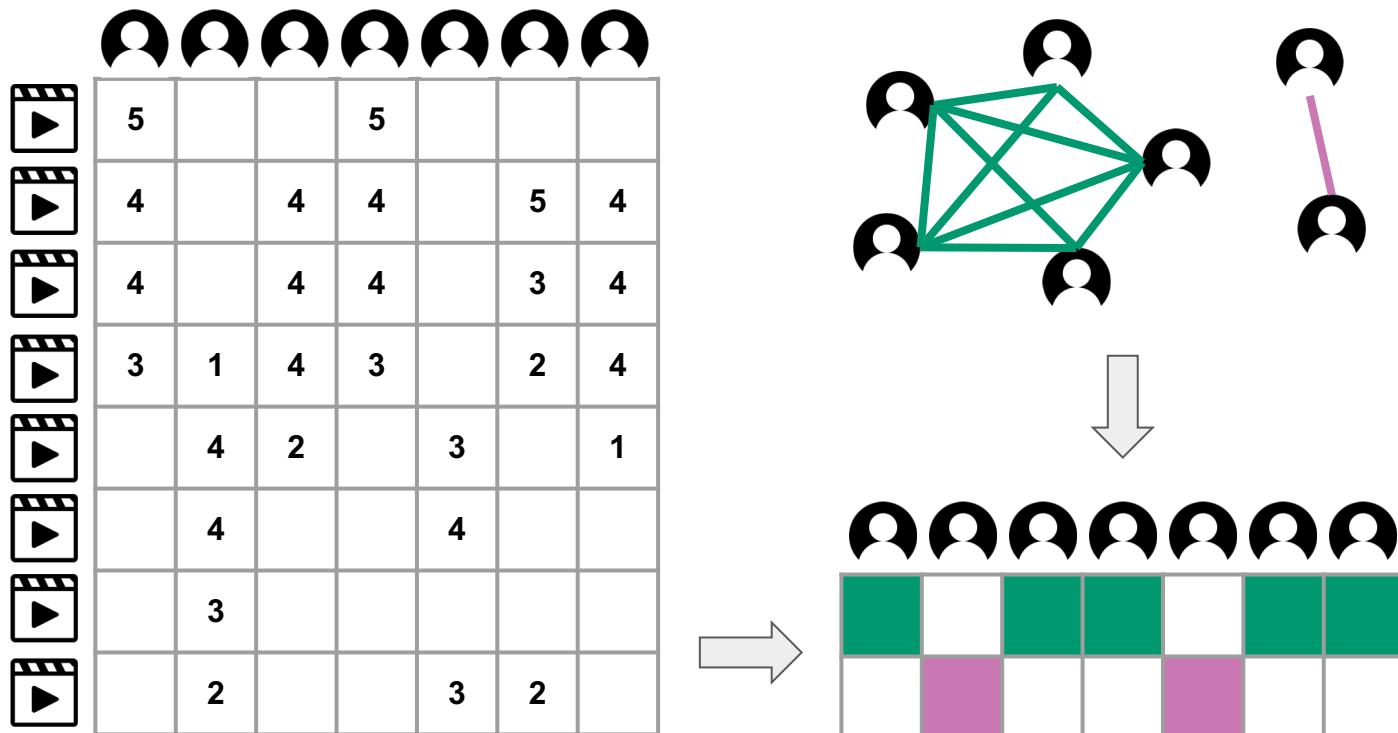
$$U = \mathbb{R}^{n \times k}$$



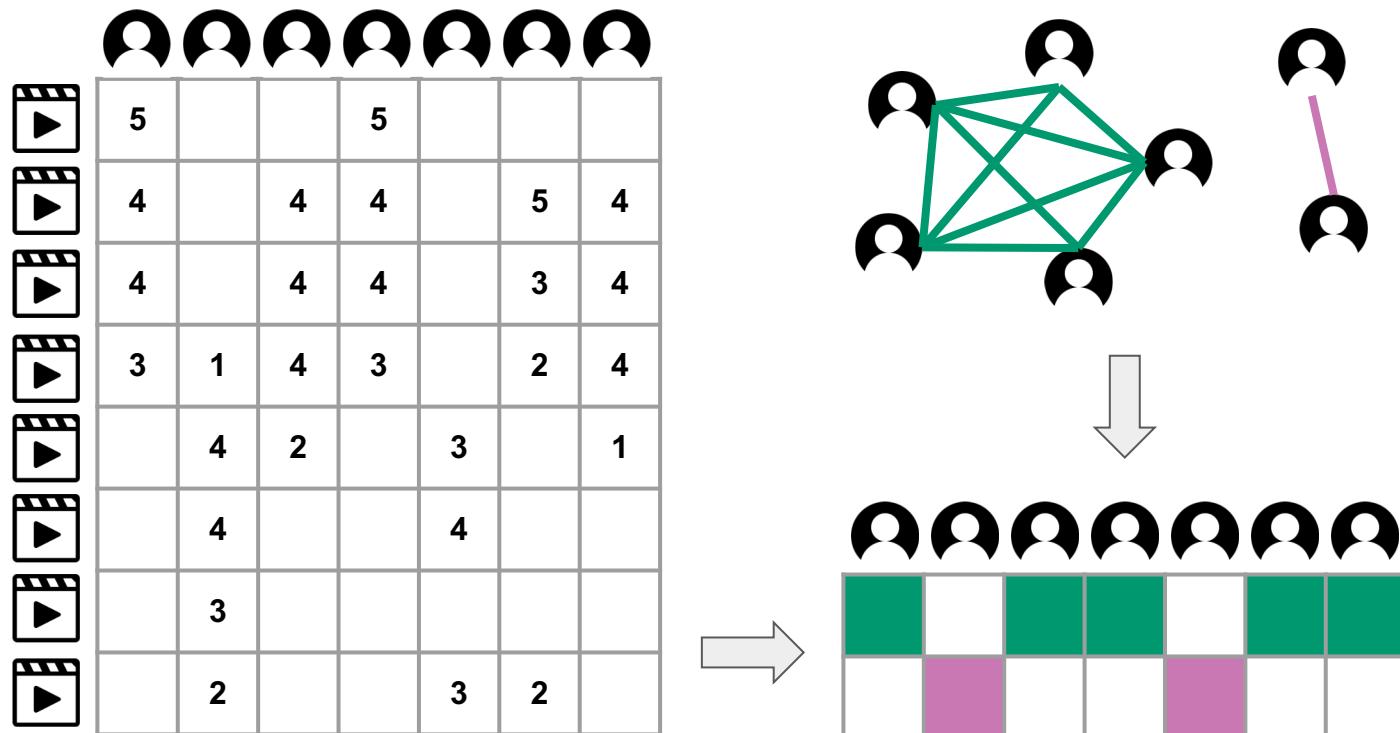
$$V^T = \mathbb{R}^{k \times m}$$

$$\text{Minimize } O = \|X - UV^T\|^2$$

Graph regularization incorporates prior knowledge in network form

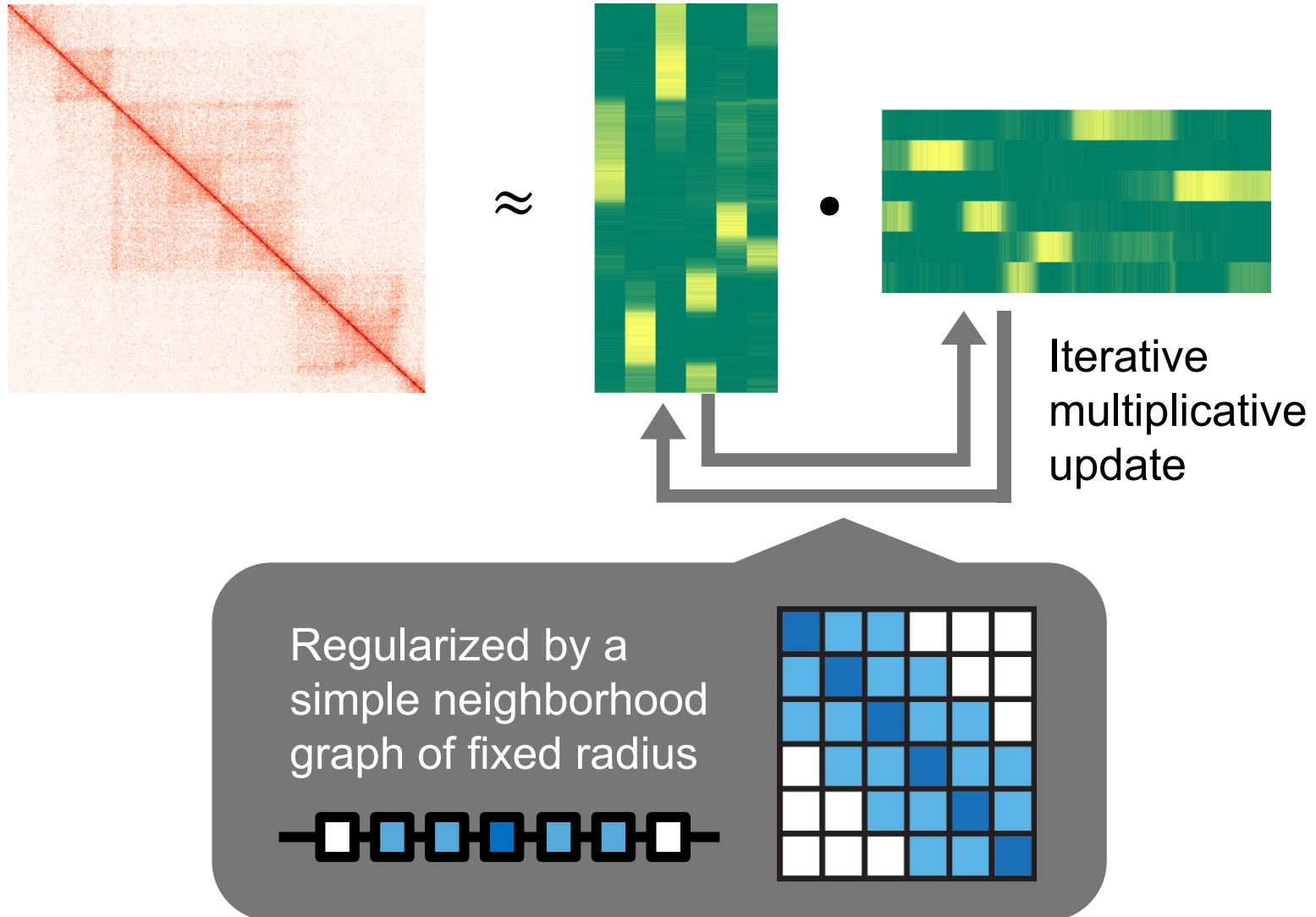


Graph regularization incorporates prior knowledge in network form

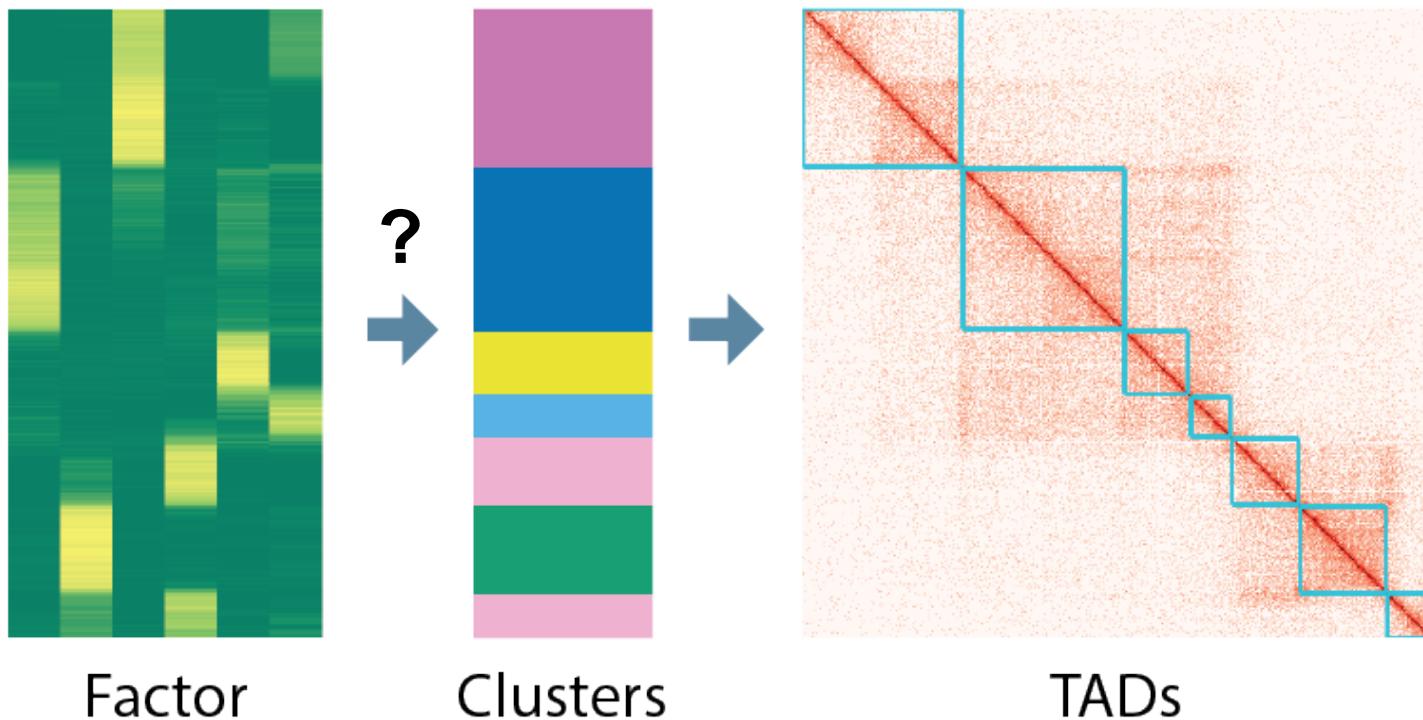


$$\text{Minimize } O = \|X - UV^T\|^2 + \lambda \text{Tr}(V^T L V) + \lambda \text{Tr}(U^T L U)$$

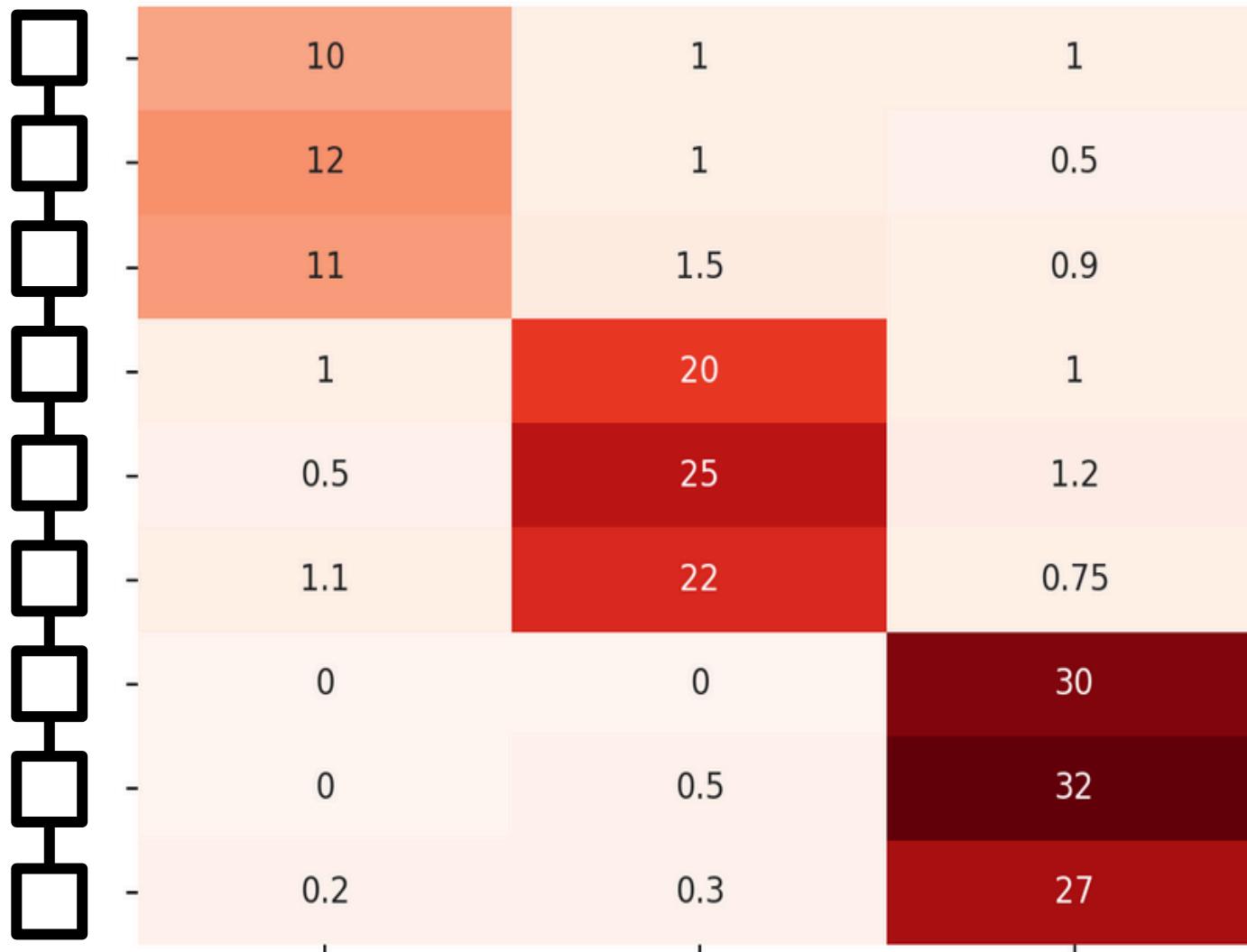
Graph regularized NMF on Hi-C data



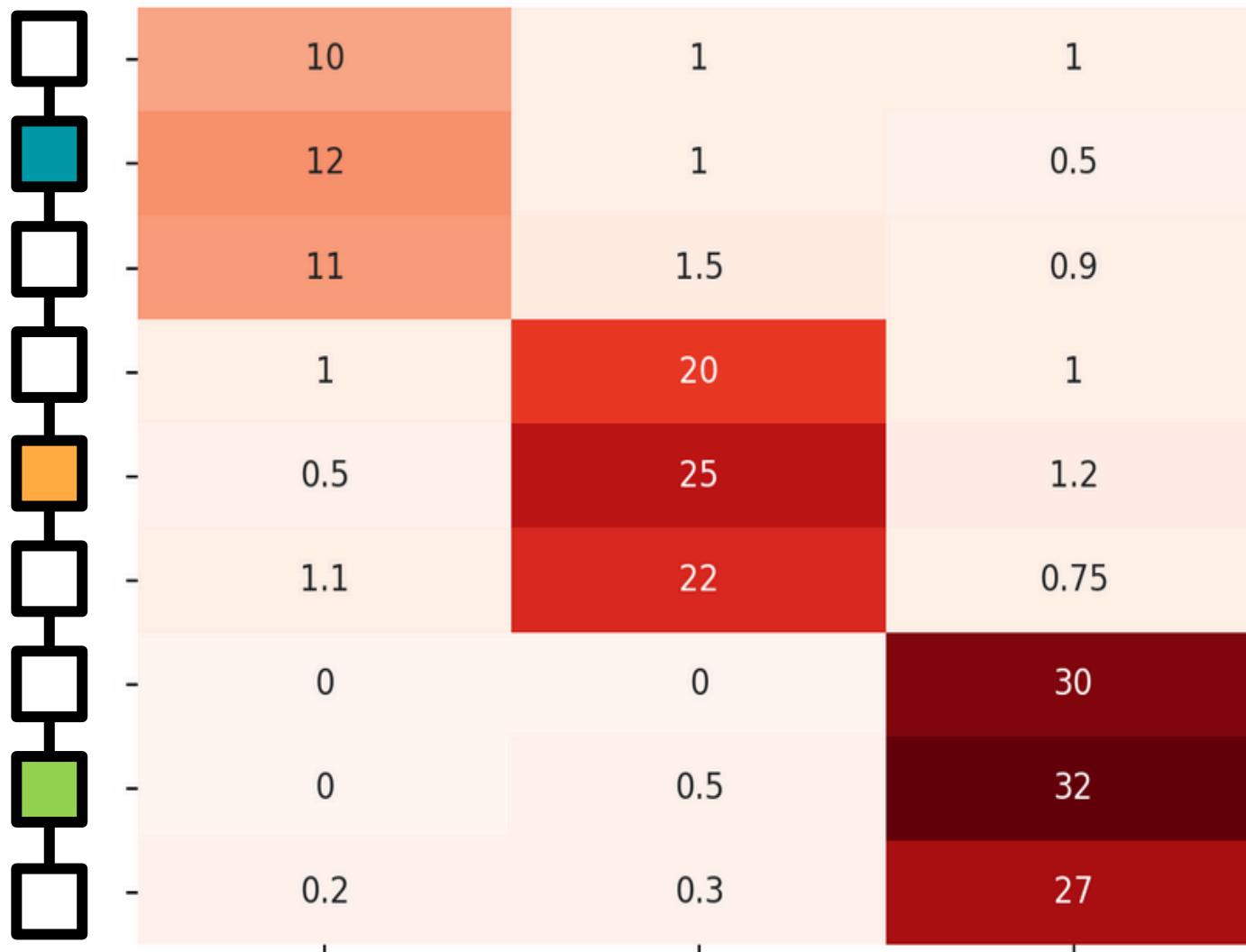
Lower-dimension factors to clusters



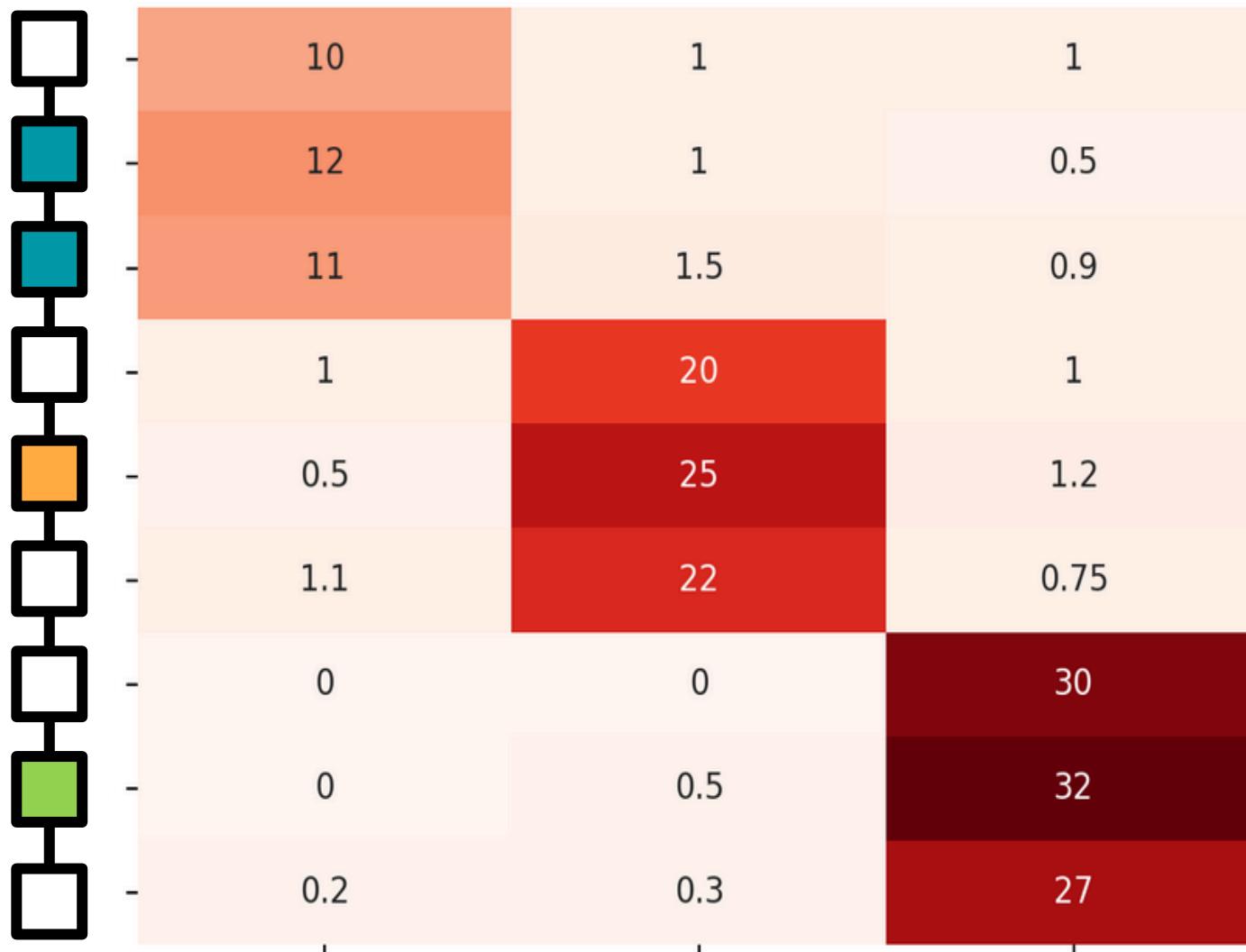
Chain-constrained k-medoids clustering



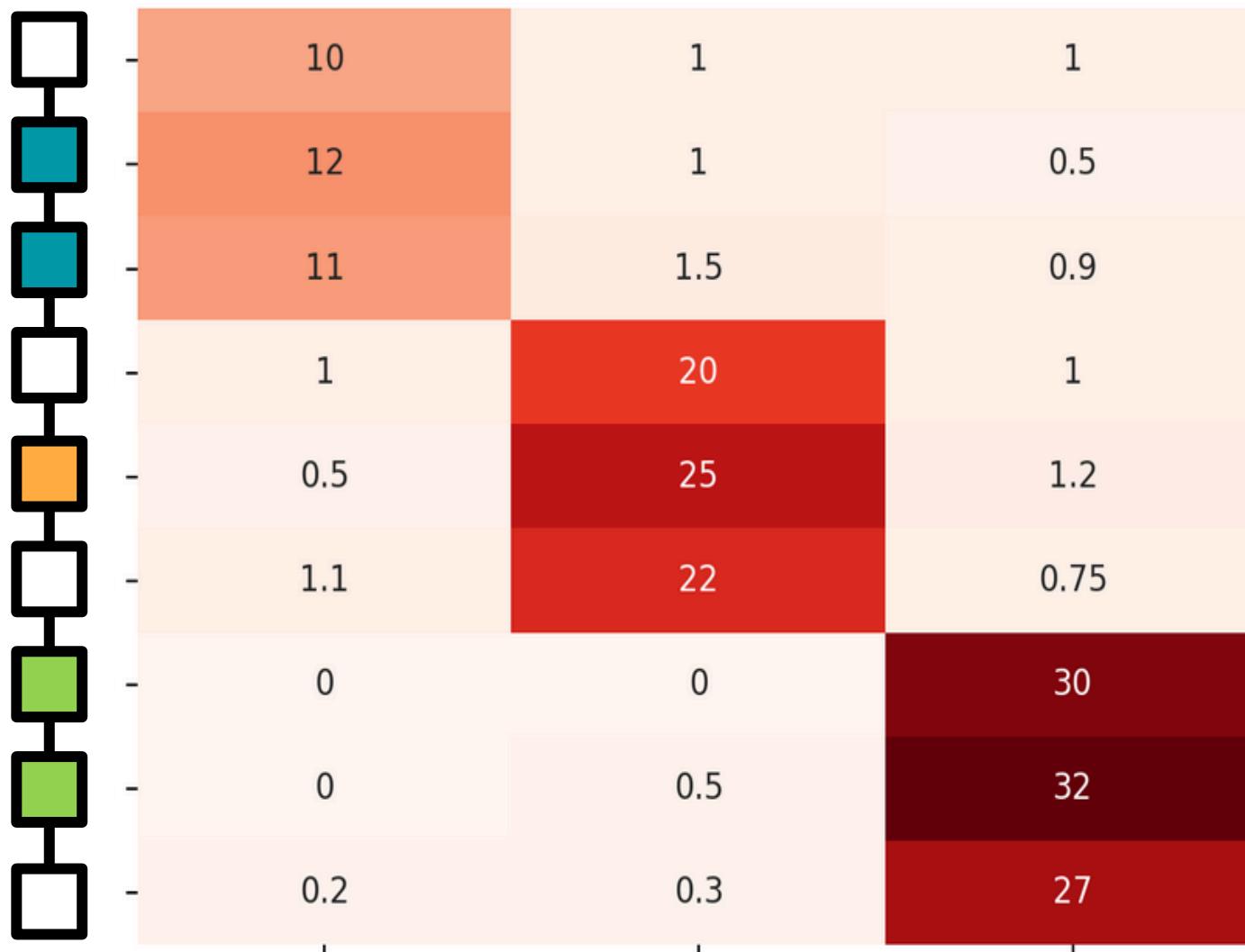
Chain-constrained k-medoids clustering



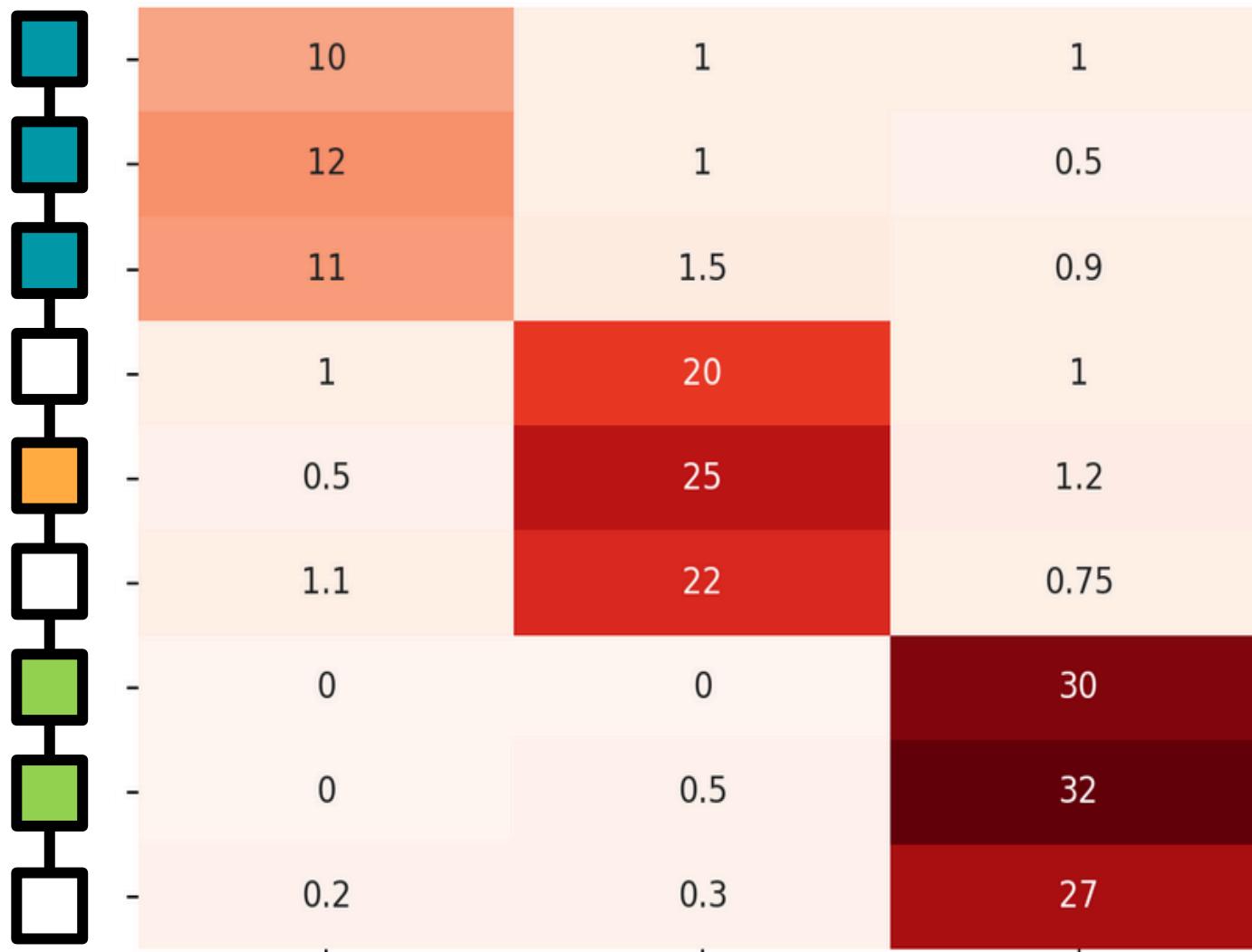
Chain-constrained k-medoids clustering



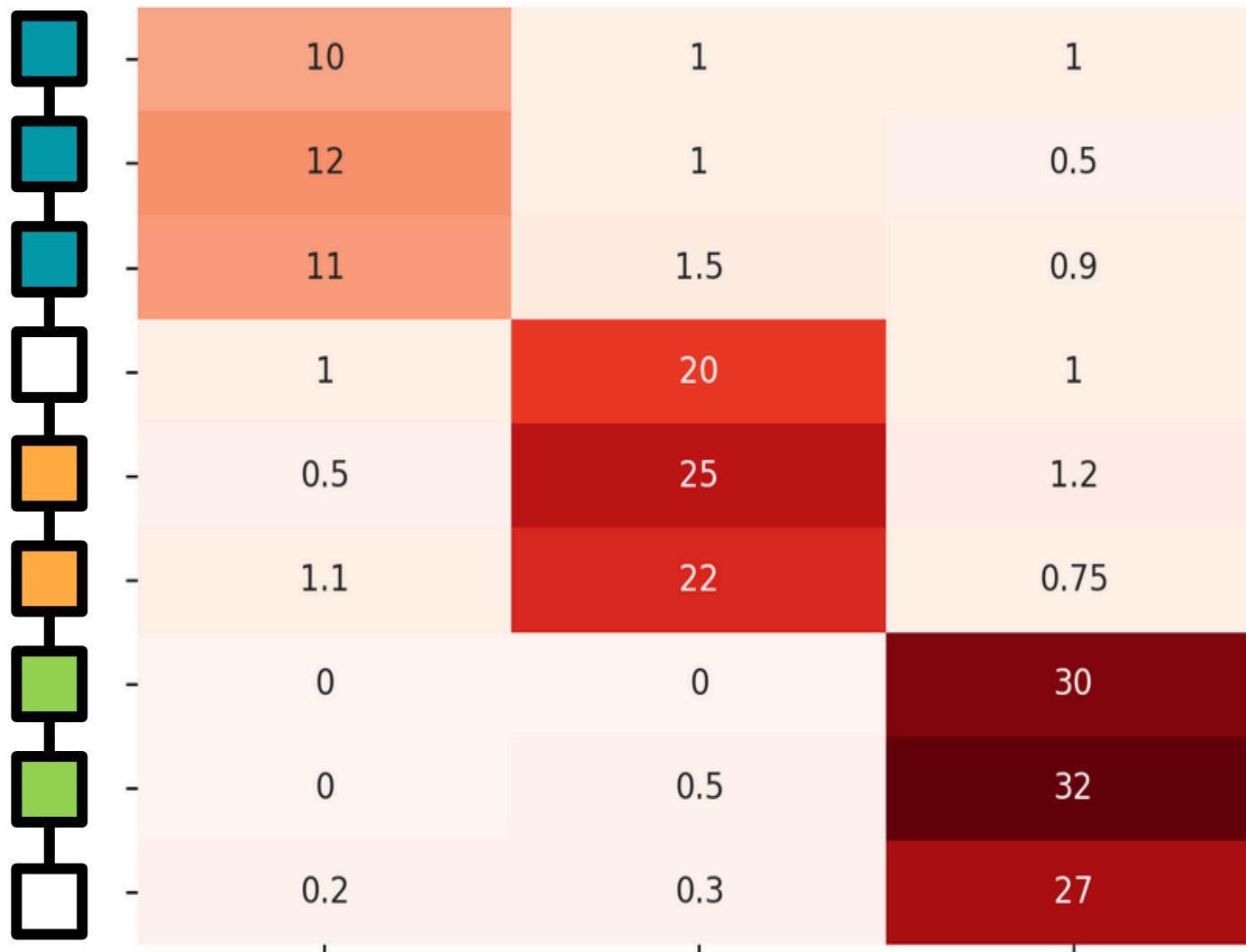
Chain-constrained k-medoids clustering



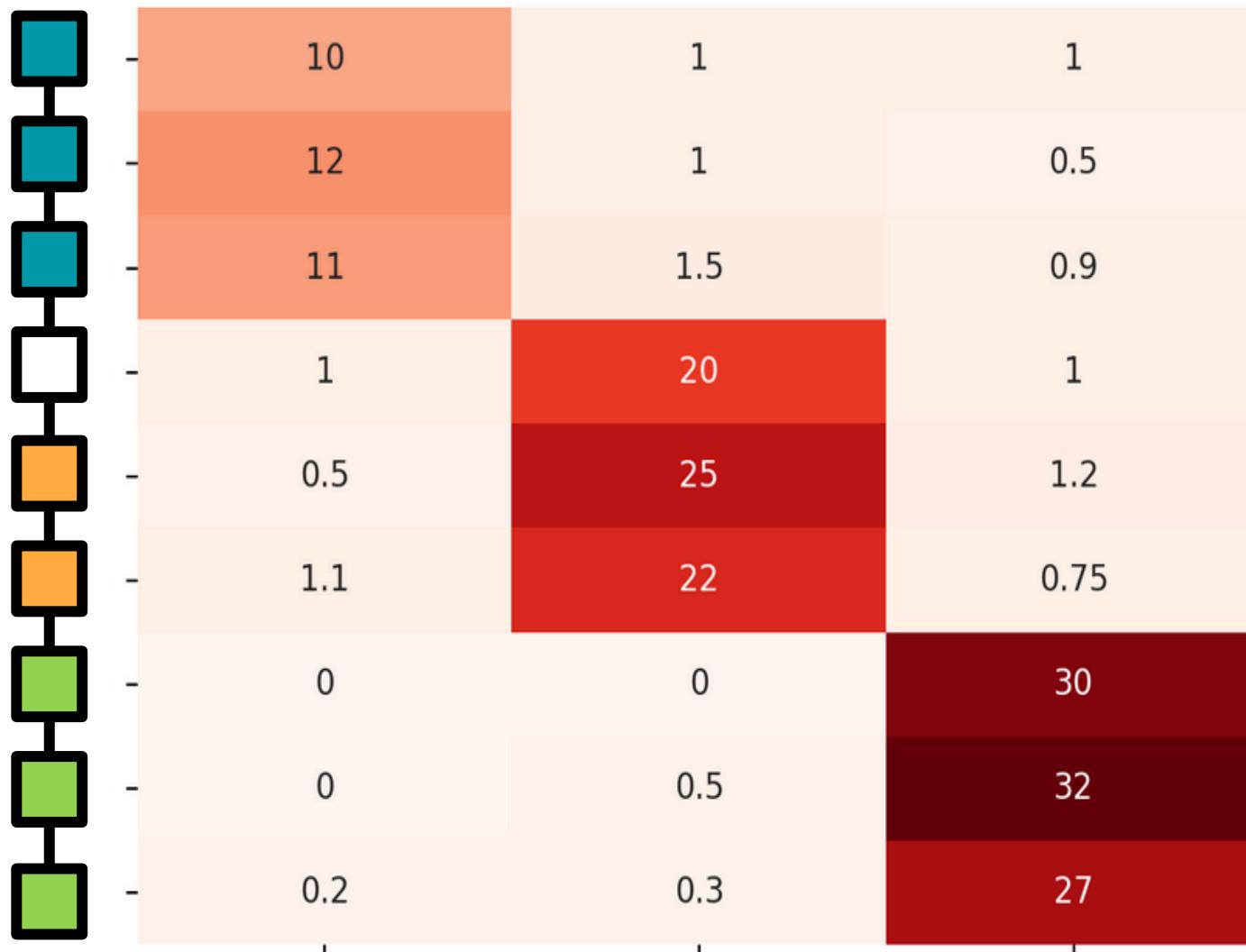
Chain-constrained k-medoids clustering



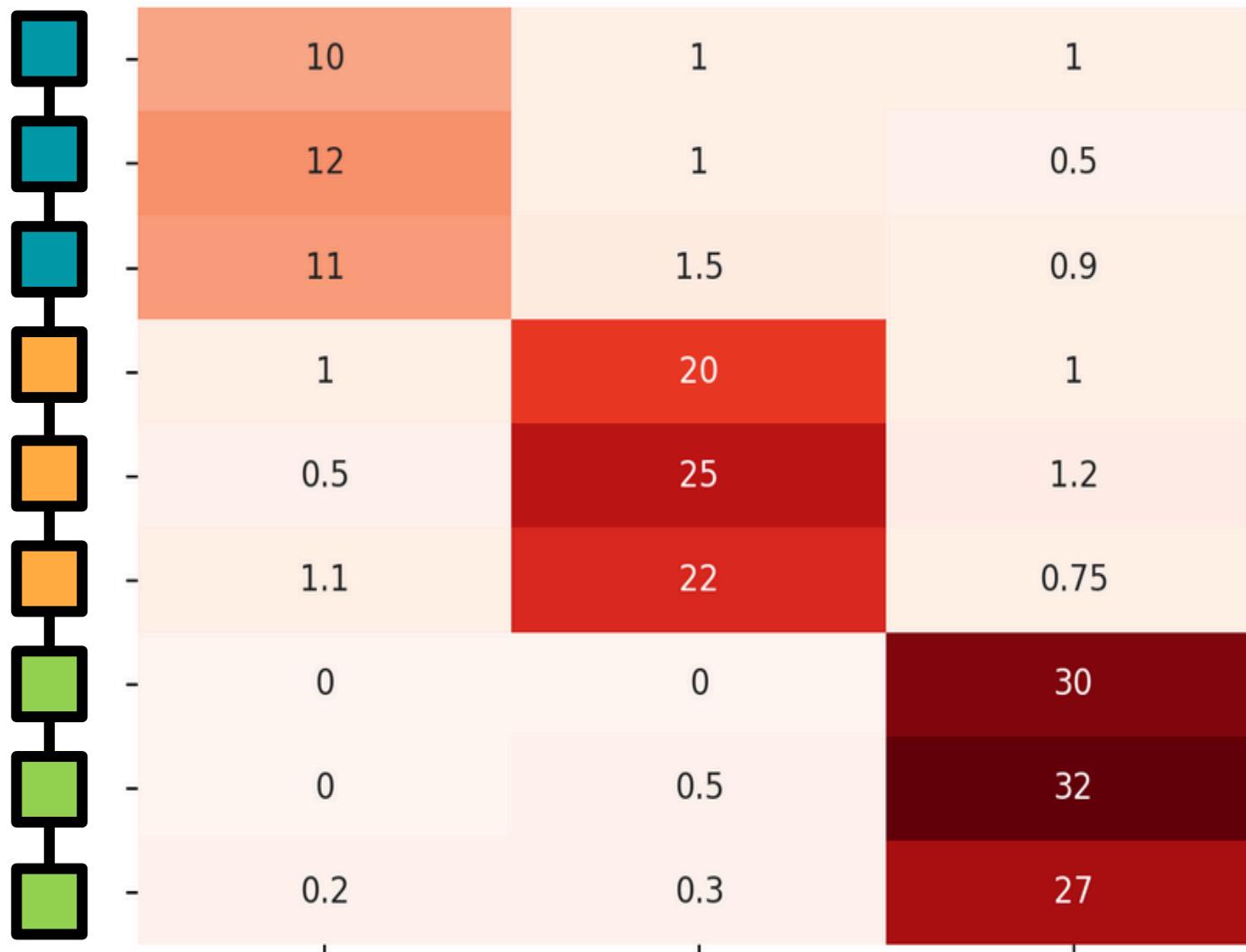
Chain-constrained k-medoids clustering



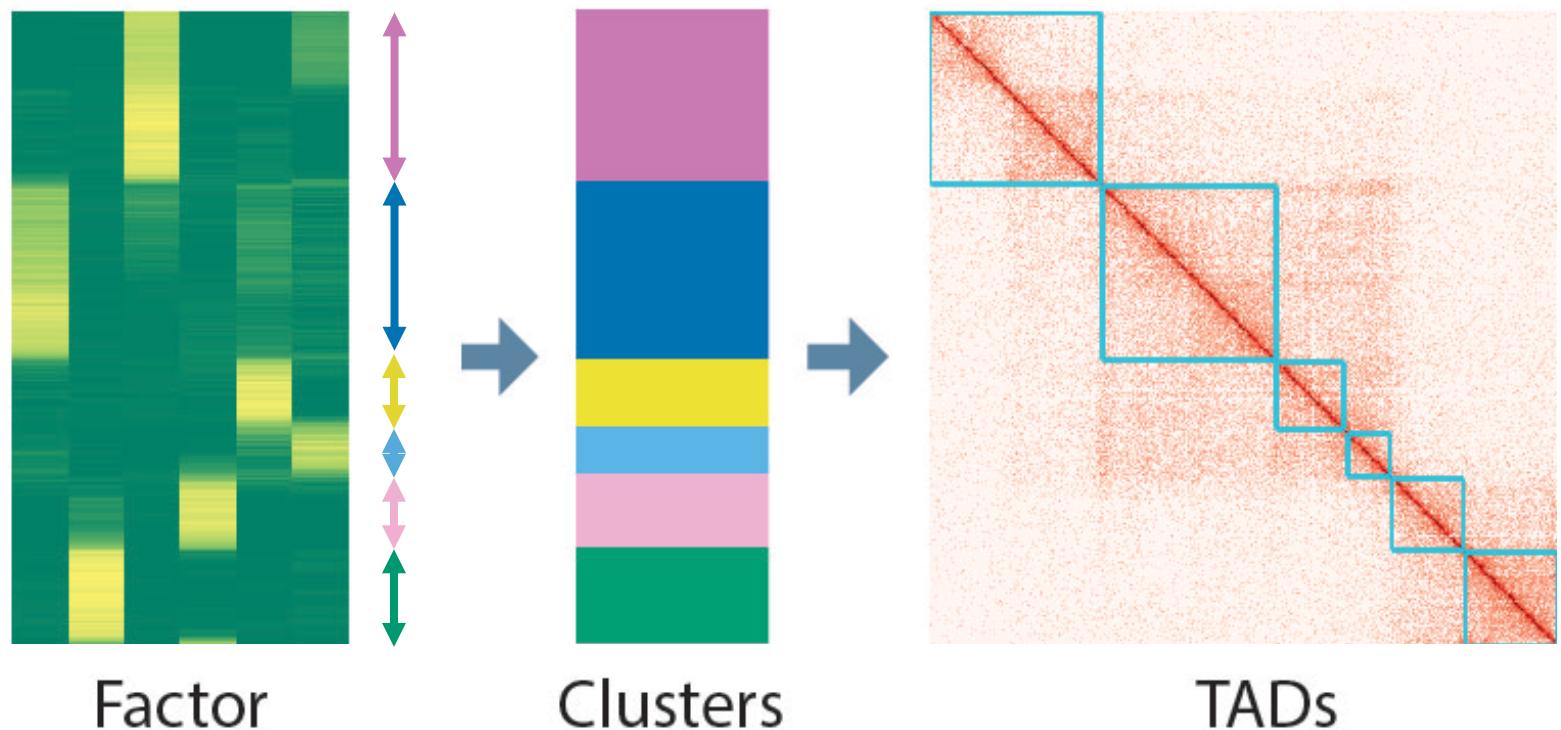
Chain-constrained k-medoids clustering



Chain-constrained k-medoids clustering



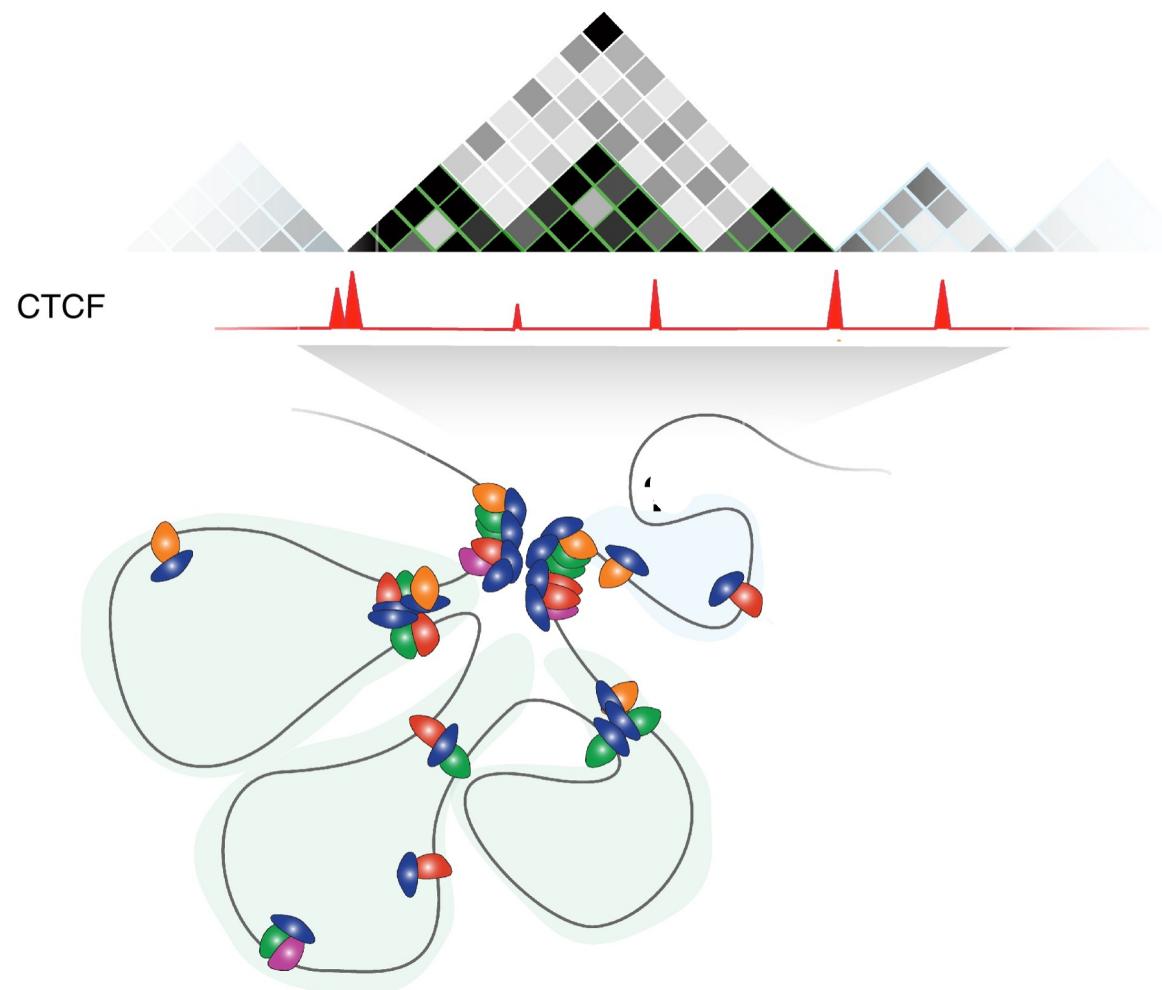
Chain-constrained k-medoids clustering on lower-dimensional factor



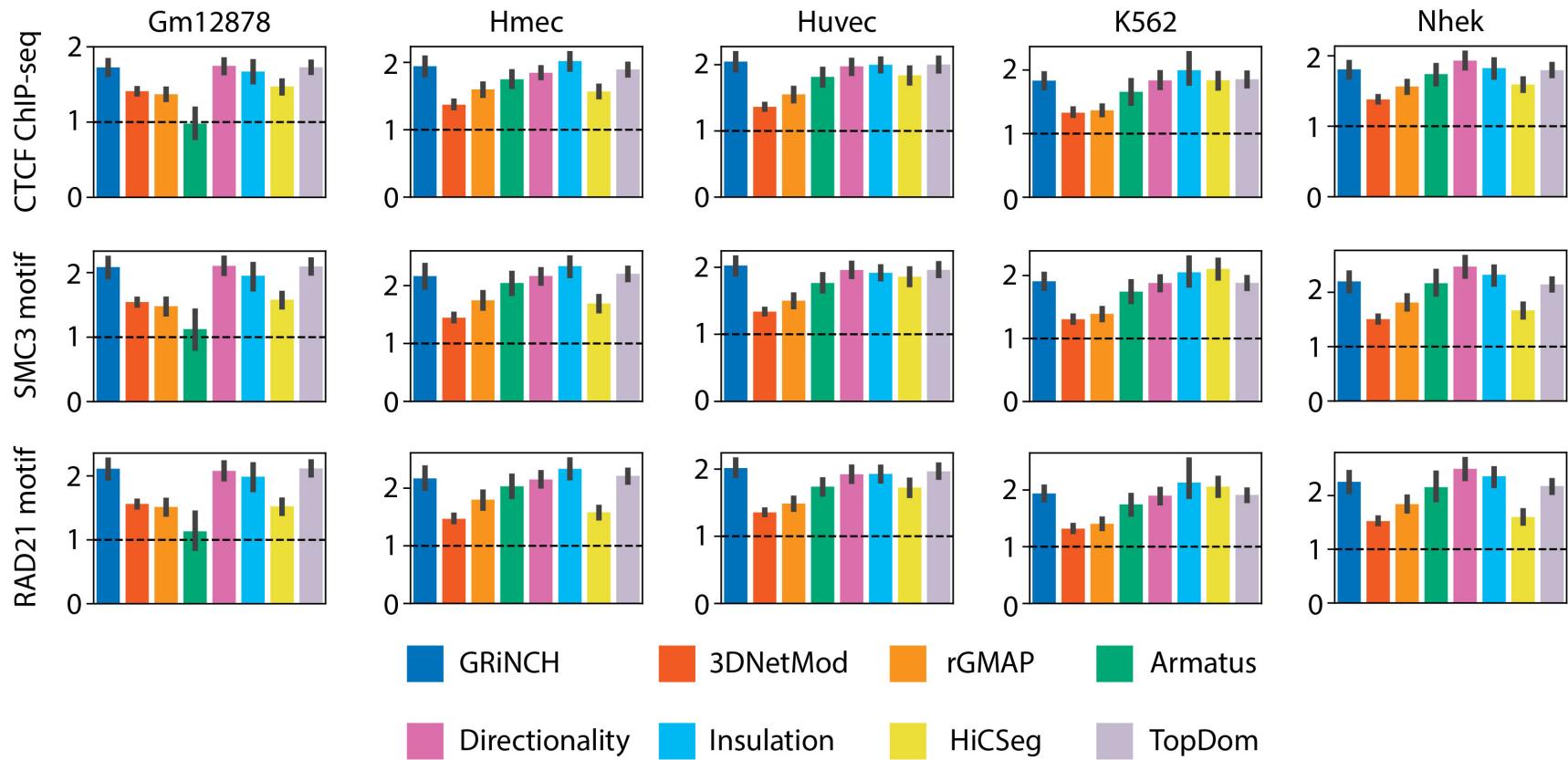
Evaluating effectiveness of GRiNCH

- ◆ CTCF and cohesion enrichment in TAD boundaries
- ◆ Stability of TADs to low-depth data
- ◆ Recovery of TADs from smoothed low-depth data

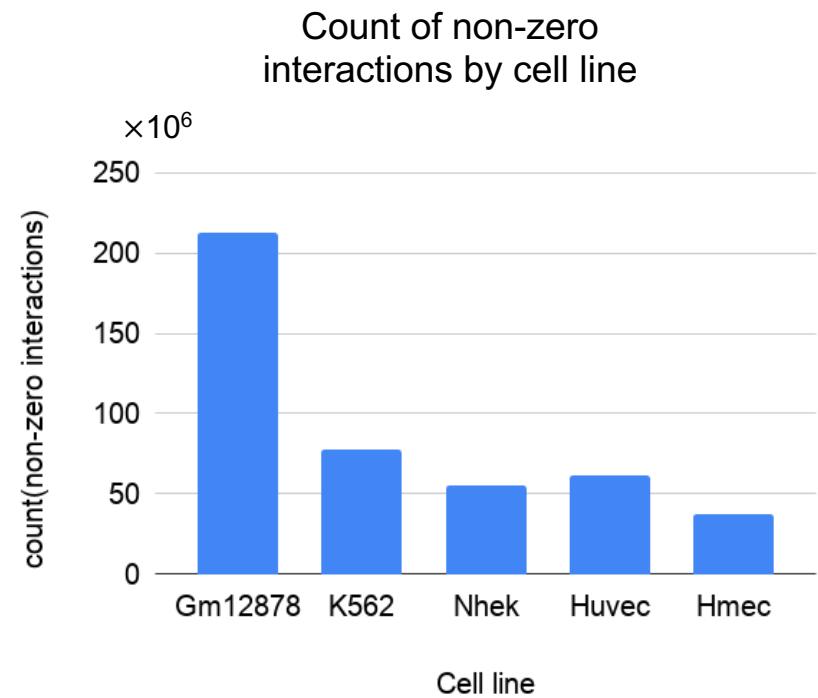
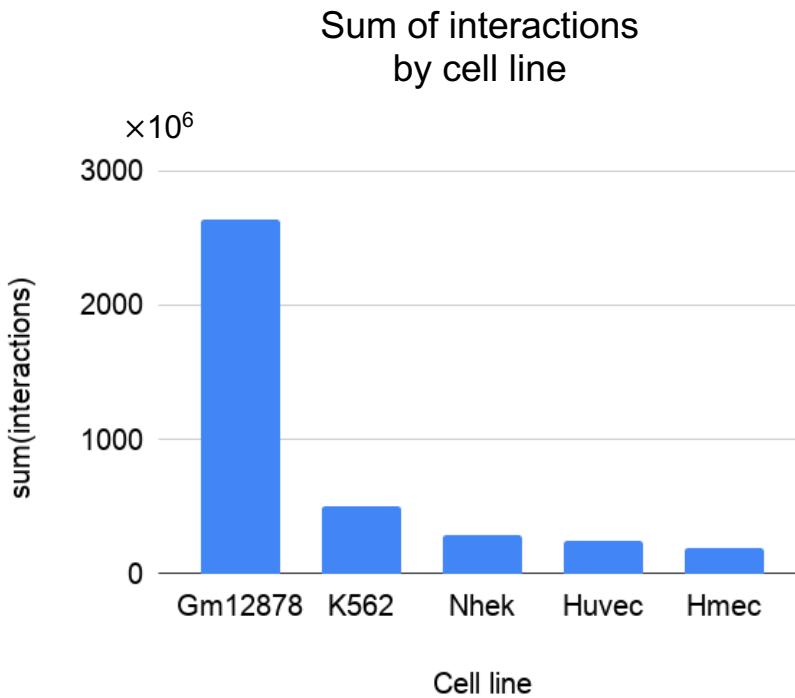
CTCF binding is associated with TAD boundaries



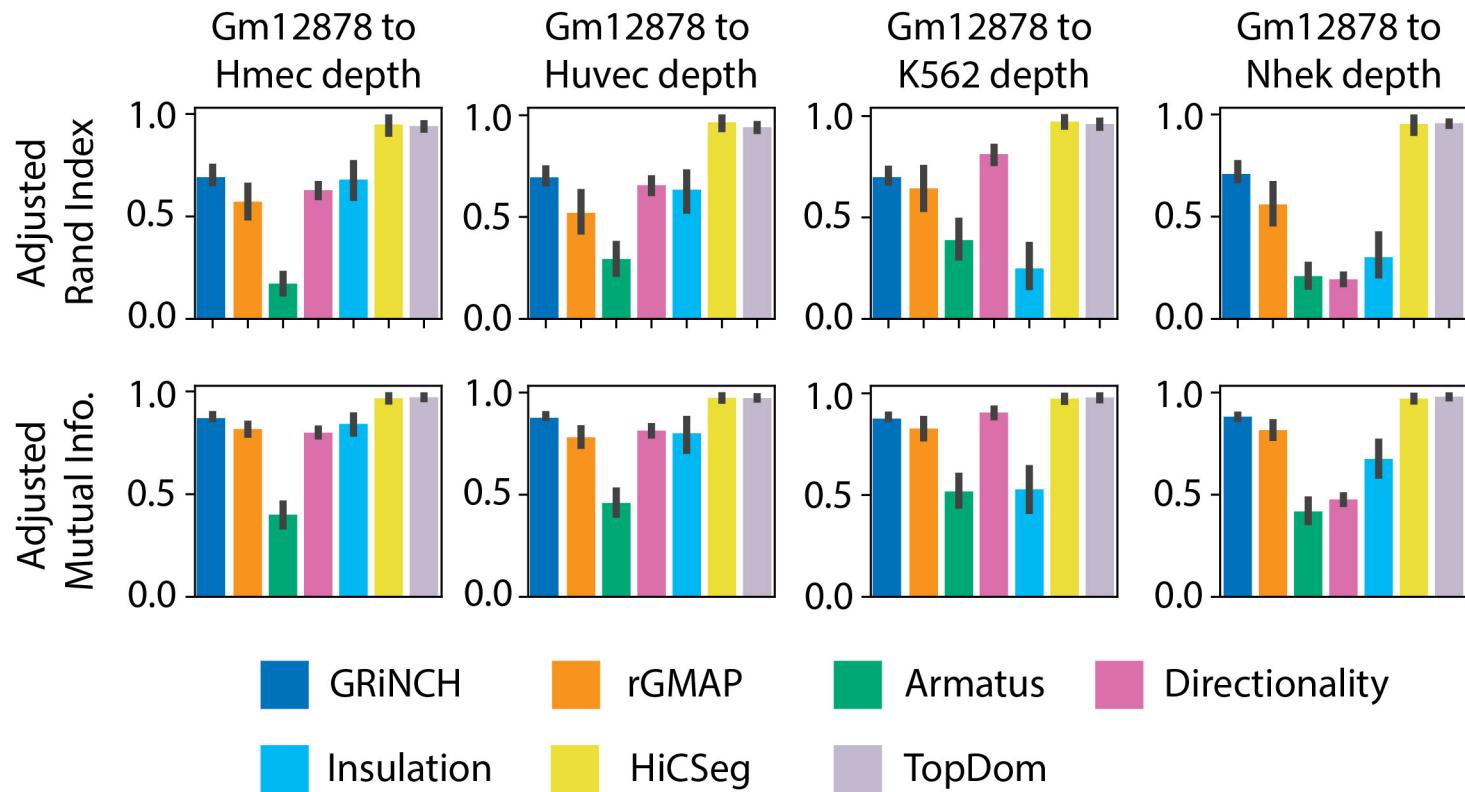
GRiNCH cluster boundaries are significantly enriched in CTCF and cohesin binding



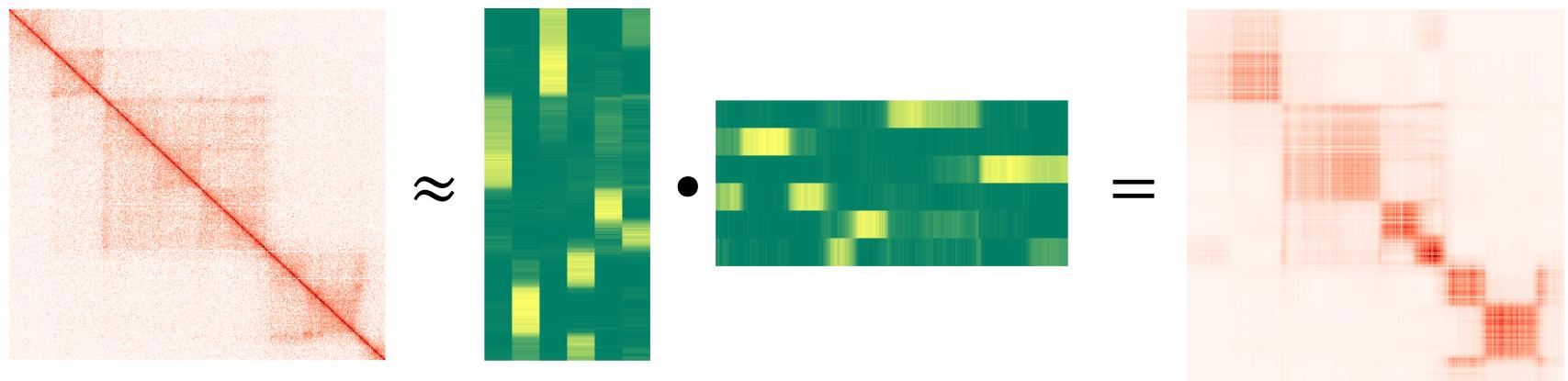
Many public high-resolution Hi-C datasets have low sequence depth



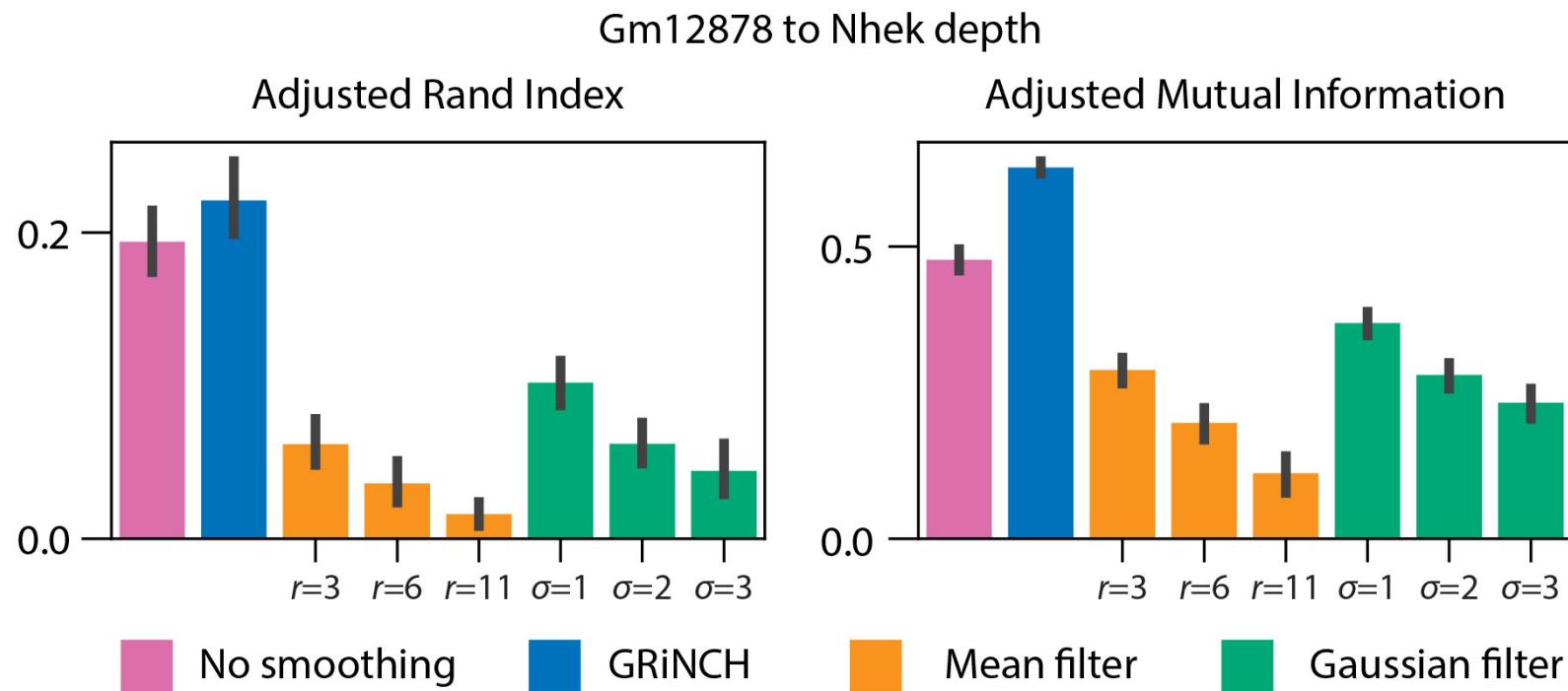
GRiNCH clusters are stable to low depth and sparsity



GRiNCH can smooth Hi-C matrix through matrix completion



GRiNCH smoothing can help recover structure in low-depth Hi-C data



Conclusion

- ◆ GRiNCH is an NMF-based method with graph regularization to find structural units of the genome.
- ◆ GRiNCH finds clusters with significant boundary element enrichment.
- ◆ GRiNCH is stable to low-depth, sparse datasets.
- ◆ GRiNCH can smooth input Hi-C matrix.

Acknowledgements

Members of Roy lab:

Sushmita Roy
Brittany Baur
Shilu Zhang
Alireza Siahpirani
Sara Knaack
Junha Shin
Sunnie Grace McCalla
Matt Stone

Funding sources:

Center for Predictive Computational Phenotyping
(NIH BD2K U54 AI117924)
NIH NIGRI R01-HG010045-01

Computing resources:

Center for High Throughput Computing (CHTC)

