GRINCH: a matrix factorization method to discover structural units of chromosomes

Da-Inn Lee and Sushmita Roy

Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Wisconsin Institute of Discovery

School of Medicine and Public Health
UNIVERSITY OF WISCONSIN–MADISON

WISCONSIN
INSTITUTE FOR DISCOVERY
WID.WISC.EDU

Long-range gene regulation by distal elements

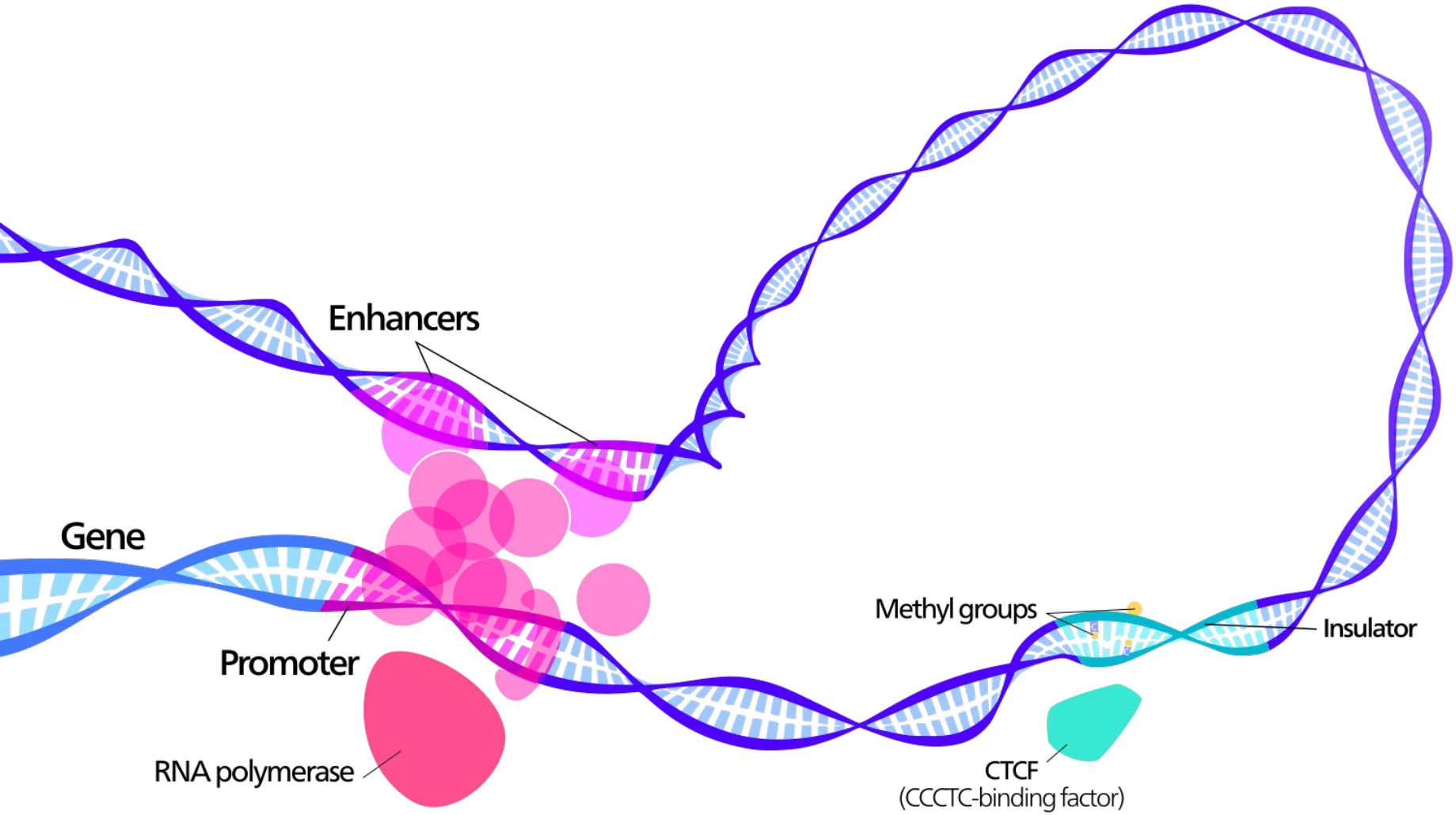
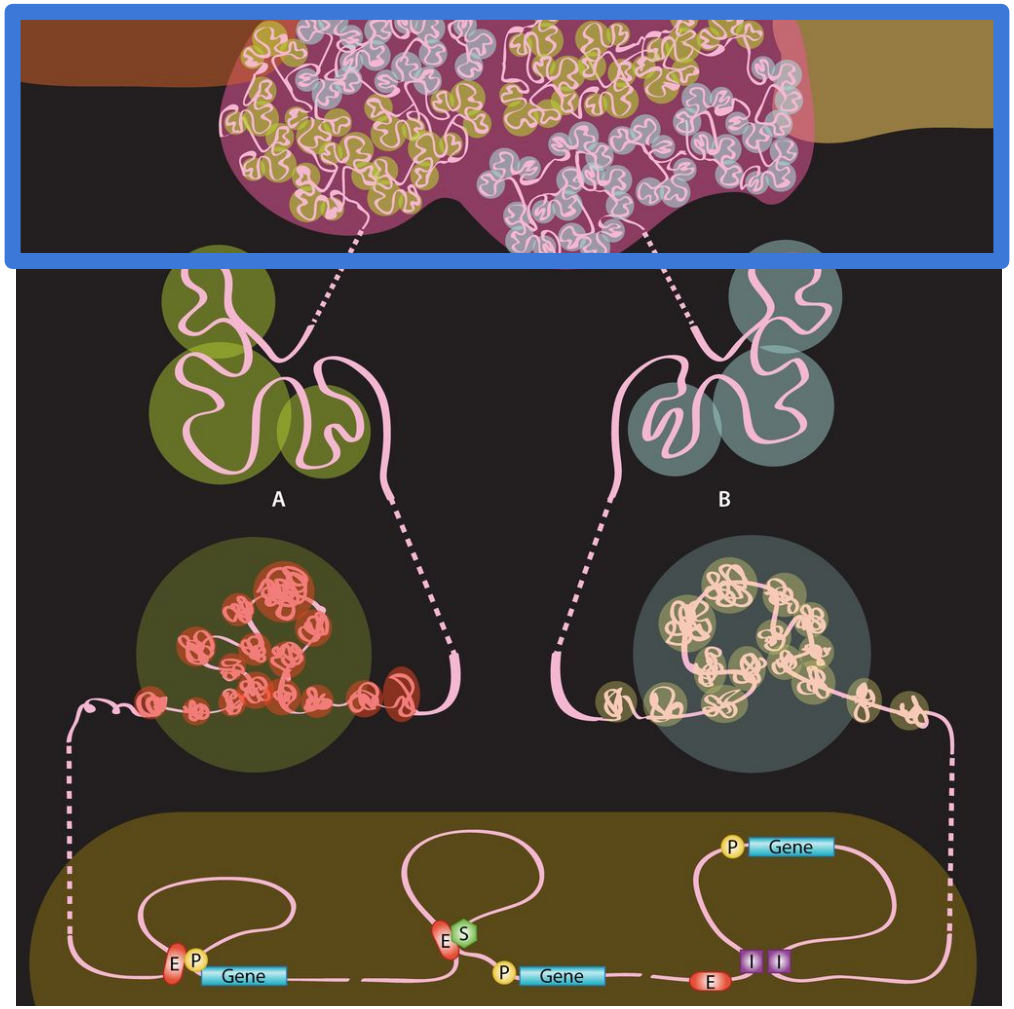


Illustration by Kelvin Ma

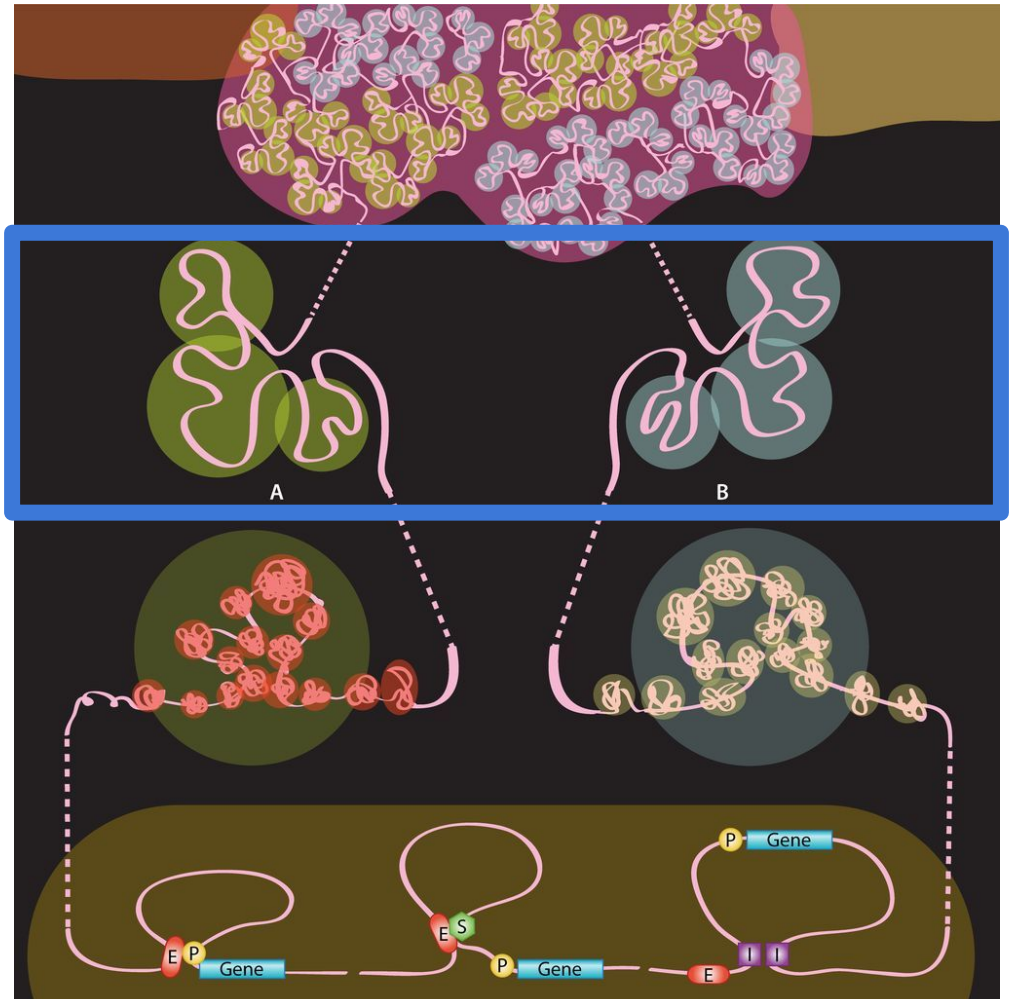
Genome is organized into higher-order domains at multiple scales

Chromosomal territories



Fraser et al. Microbiol. Mol. Biol. Rev. 2015

Genome is organized into higher-order domains at multiple scales

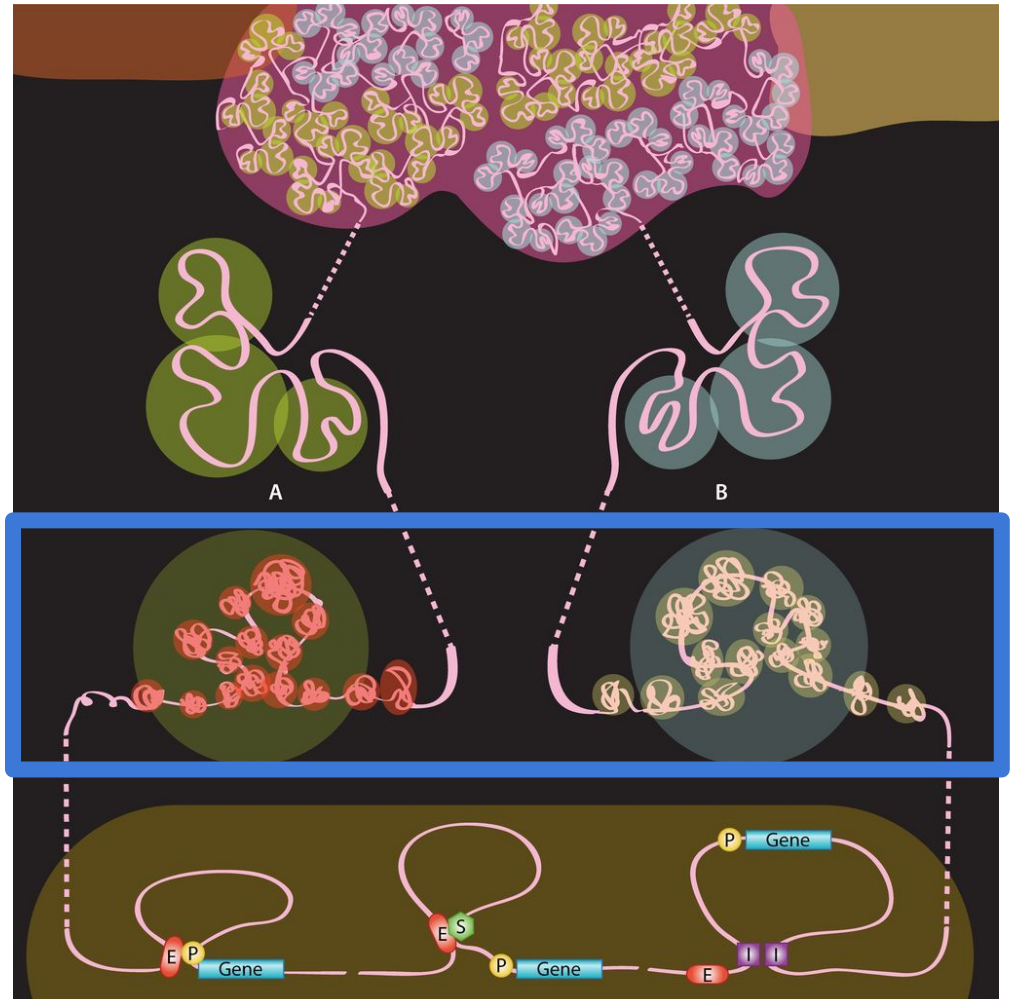


Compartments

Fraser et al. Microbiol. Mol. Biol. Rev. 2015

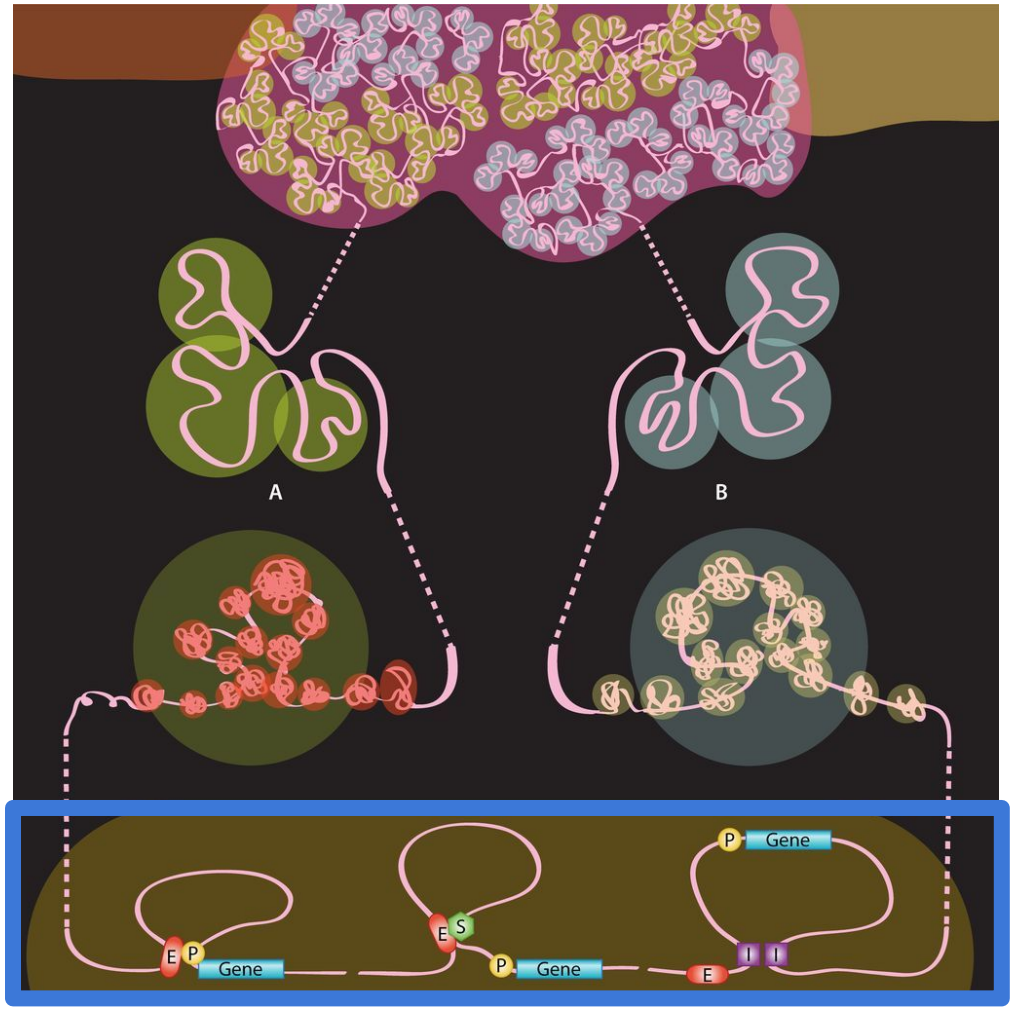
Genome is organized into higher-order domains at multiple scales

TADs and sub-TADs



Fraser et al. Microbiol. Mol. Biol. Rev. 2015

Genome is organized into higher-order domains at multiple scales



Chromatin loops

Fraser et al. Microbiol. Mol. Biol. Rev. 2015

Existing methods for finding topological units of chromosomes

Method	Algorithm	Objective
Directionality (Dixon et al. Nature 2012)	HMM	Find domains that maximize the difference between intra- and inter-domain interaction levels
Armatus (Filippova et al. Algorithm. Mol. Biol. 2014)	Dynamic programming	Find domains that maximize intra-domain sum of contact counts
Arrowhead (Rao et al. Cell 2014)	Dynamic programming	Find *boundaries* defining domains with observed counts significantly different from expected
Insulation Score (Crane et al. Nature 2015)	Aggregation, ratio calculation	Find domains with significantly higher ratio of observed counts to expected
3DNetMod (Norton et al. Nature 2018)	Network modularity maximization	Find communities within network with maximal modularity
rGMAP(Yu et al. Nature 2017)	Gaussian Mixture model	Find two components (intra- vs inter- domain interactions) and boundaries between the two

GRINCH: a method to discover topological units of chromosomes from Hi-C data

GRINCH: a method to discover topological units of chromosomes from Hi-C data

- Non-negative factorization (NMF)
- Graph regularization

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

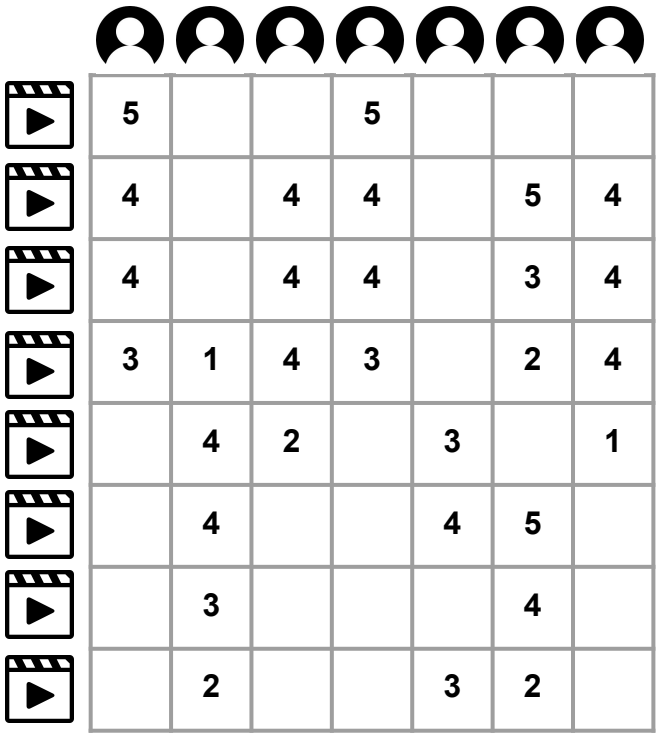
5			5			
4		4	4		5	4
4		4	4		3	4
3	1	4	3		2	4
	4	2		3		1
	4			4	5	
	3				4	
	2			3	2	

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

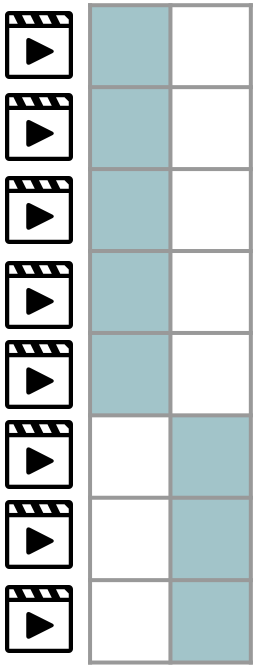
5			5			
4		4	4		5	4
4		4	4		3	4
3	1	4	3		2	4
	4	2		3		1
	4			4	5	
	3				4	
	2			3	2	

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

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

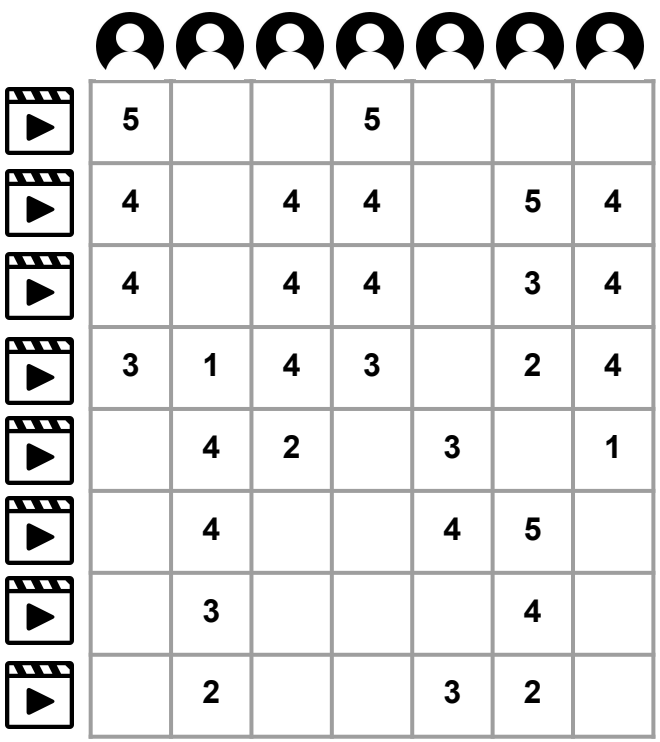


$X = \mathbb{R}^{n \times m}$

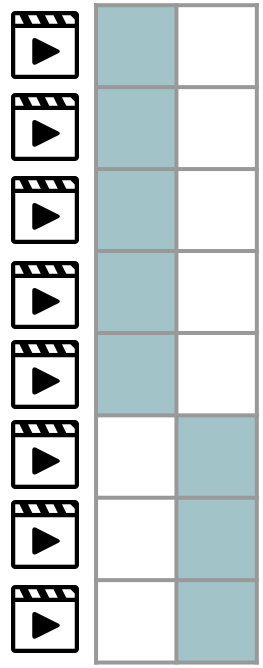


$U = \mathbb{R}^{n \times k}$

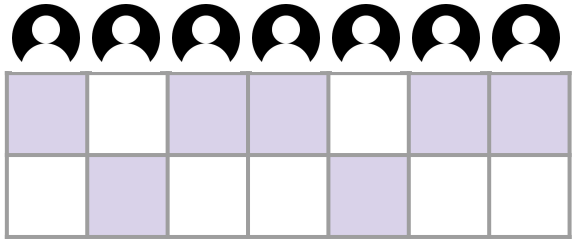
Non-negative matrix factorization (NMF) reduces dimensions of data



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

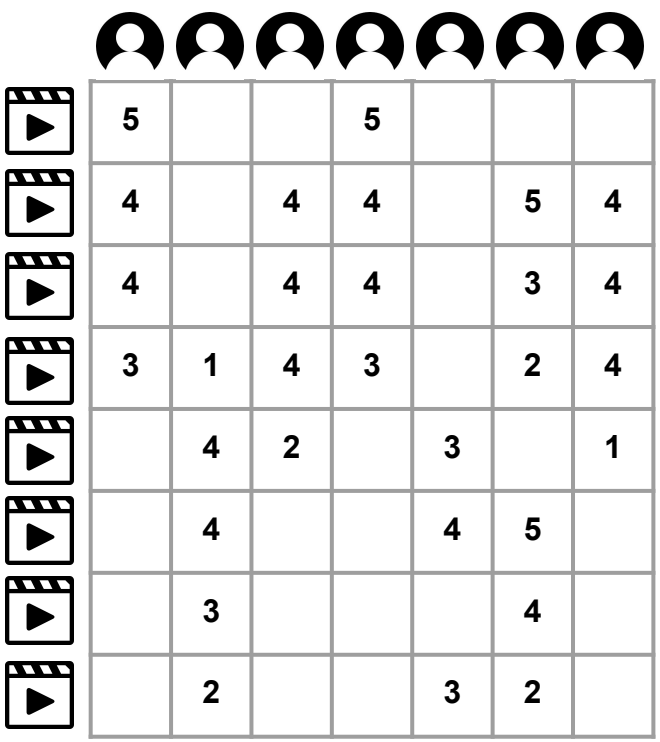


$U = \mathbb{R}^{n \times k}$

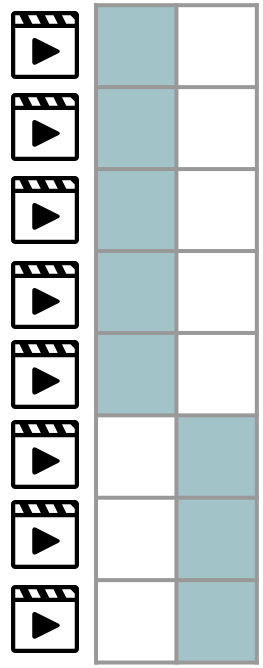


$V^T = \mathbb{R}^{k \times m}$

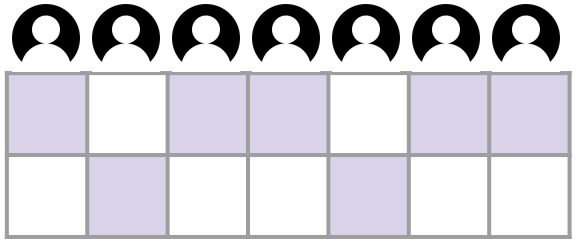
Non-negative matrix factorization (NMF) reduces dimensions of data



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



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

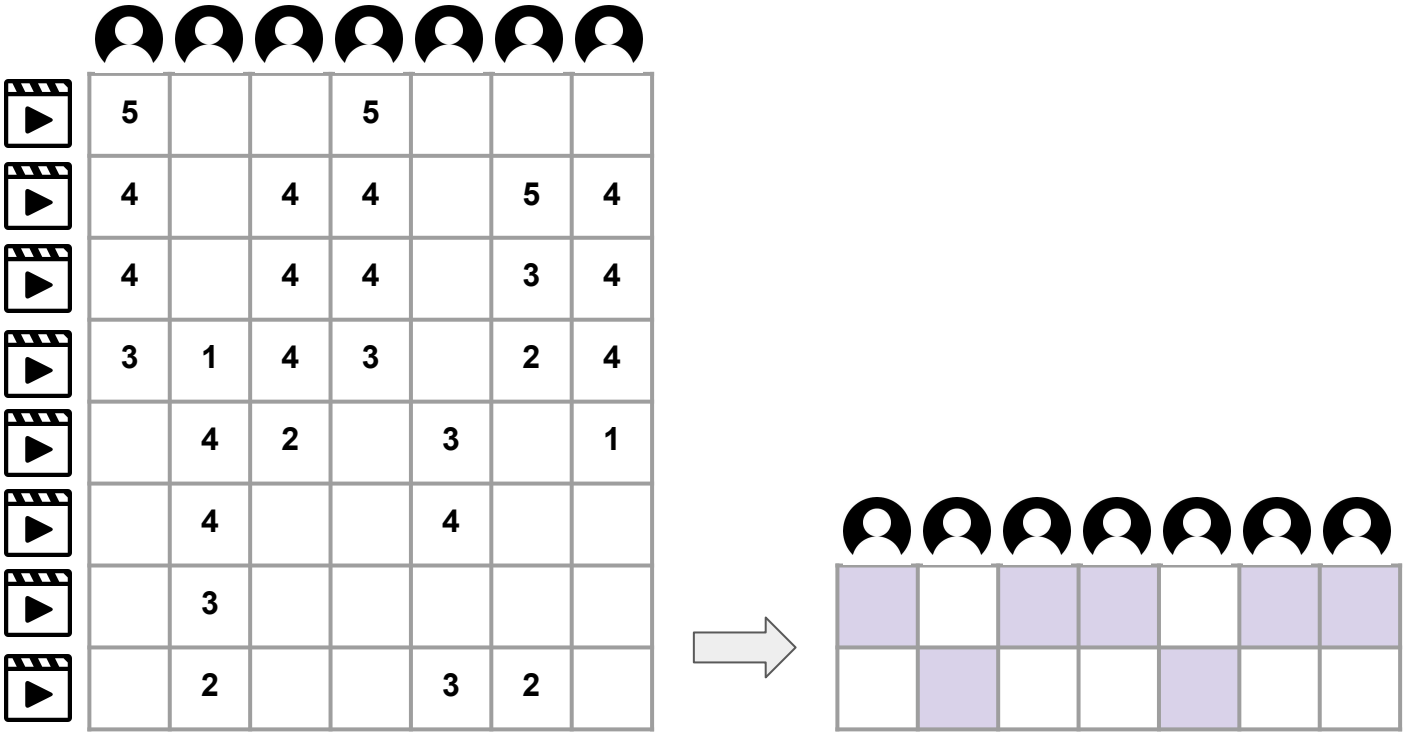


$V^T = \mathbb{R}^{k \times m}$

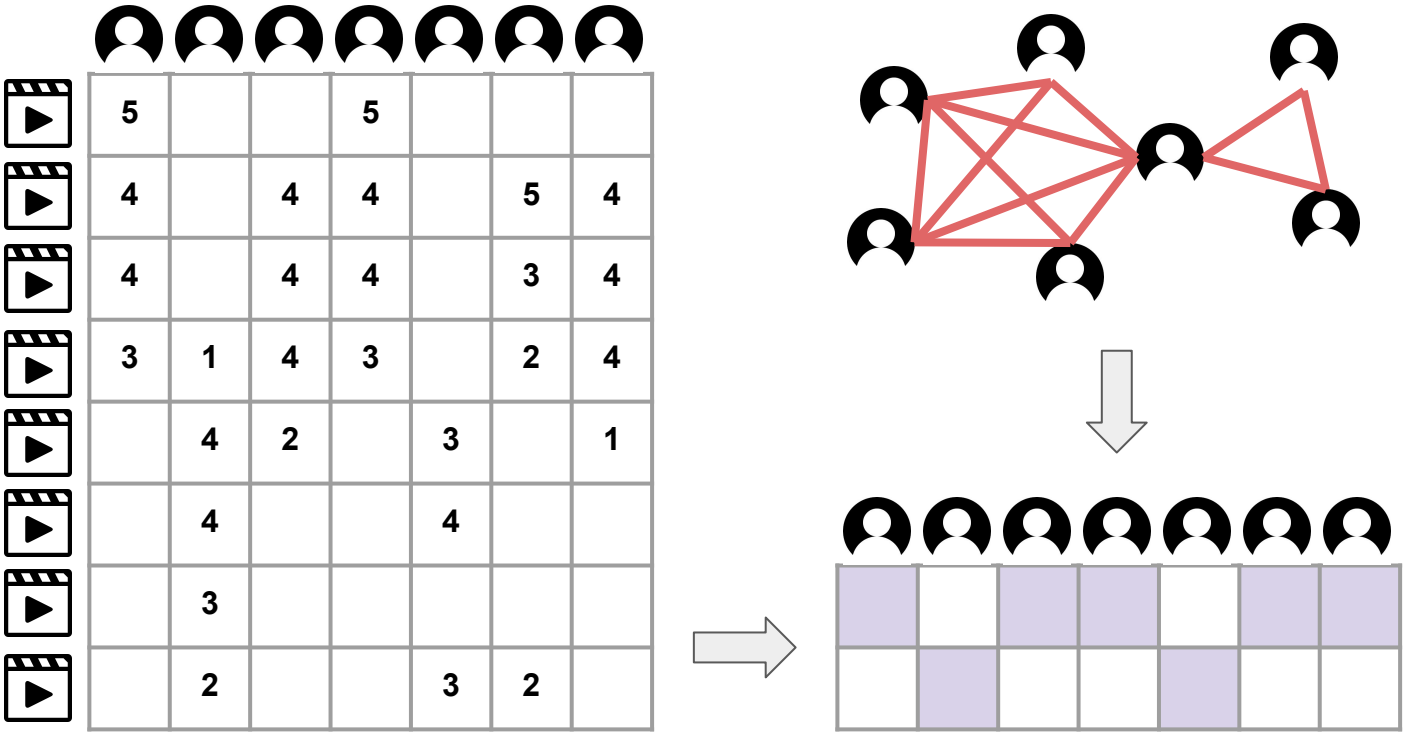
$$\text{Minimize} \quad O = \|X - UV^T\|^2$$

Lee and Seung Adv. Neur. In. 2001

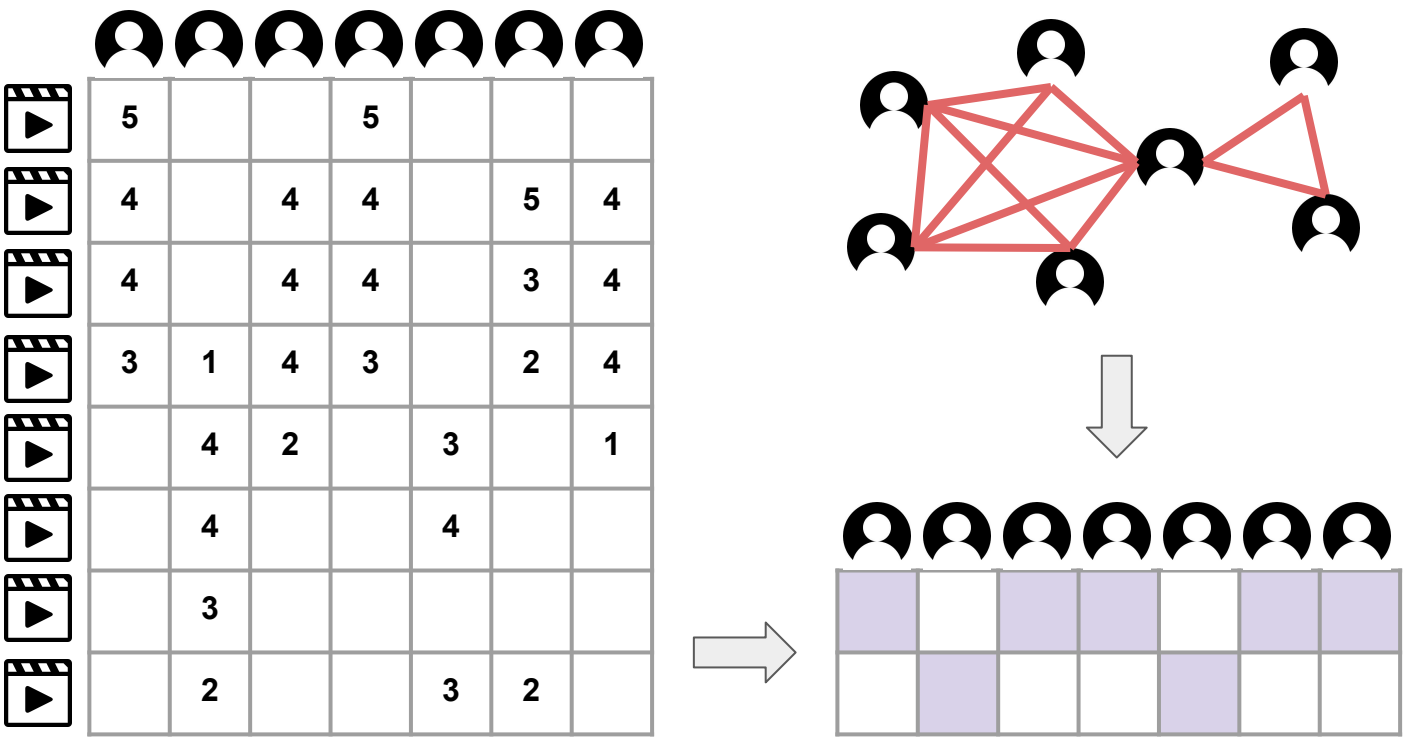
Graph regularization incorporates prior knowledge in network form



Graph regularization incorporates prior knowledge in network form



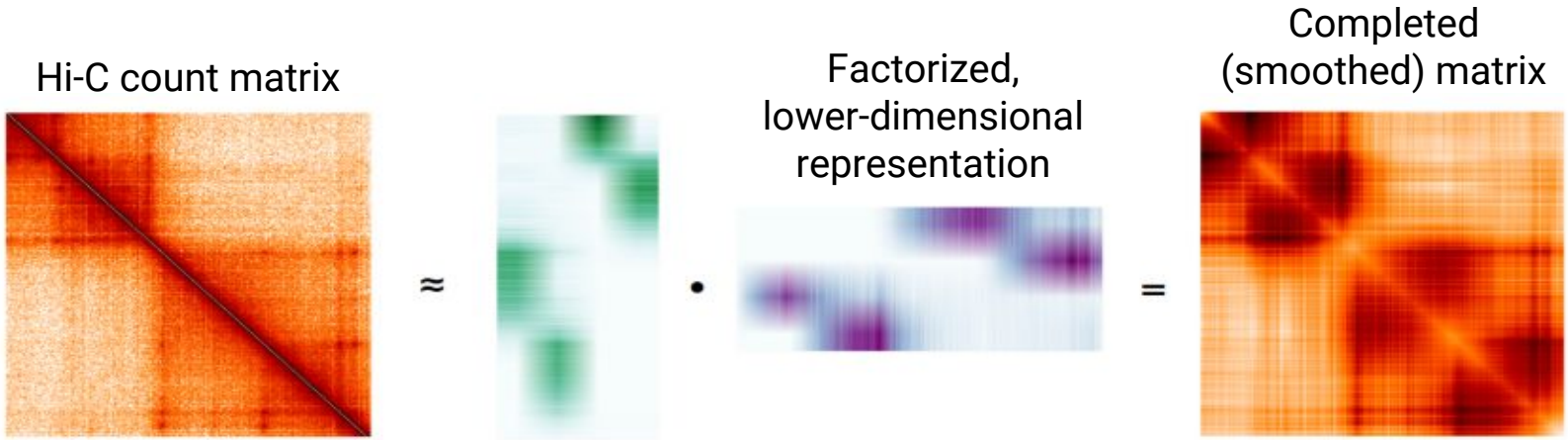
Graph regularization incorporates prior knowledge in network form



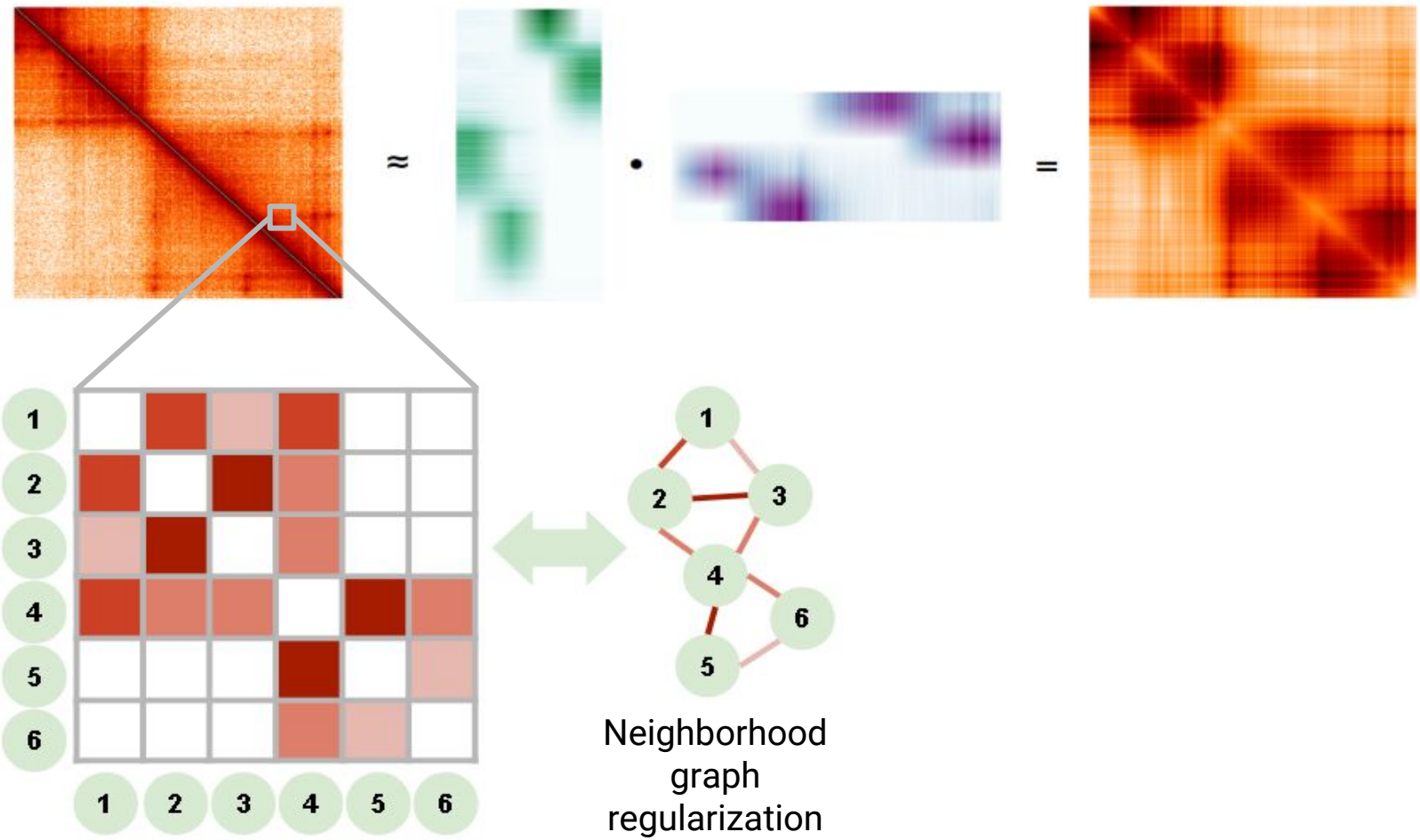
Minimize $O = \|\mathrm{X} - \mathrm{U}\mathrm{V}^T\|^2 + \lambda \mathrm{Tr}(\mathrm{V}^T\mathrm{L}\mathrm{V}) + \lambda \mathrm{Tr}(\mathrm{U}^T\mathrm{L}\mathrm{U})$

Cai et al. IEEE T. Pattern. Anal. 2011

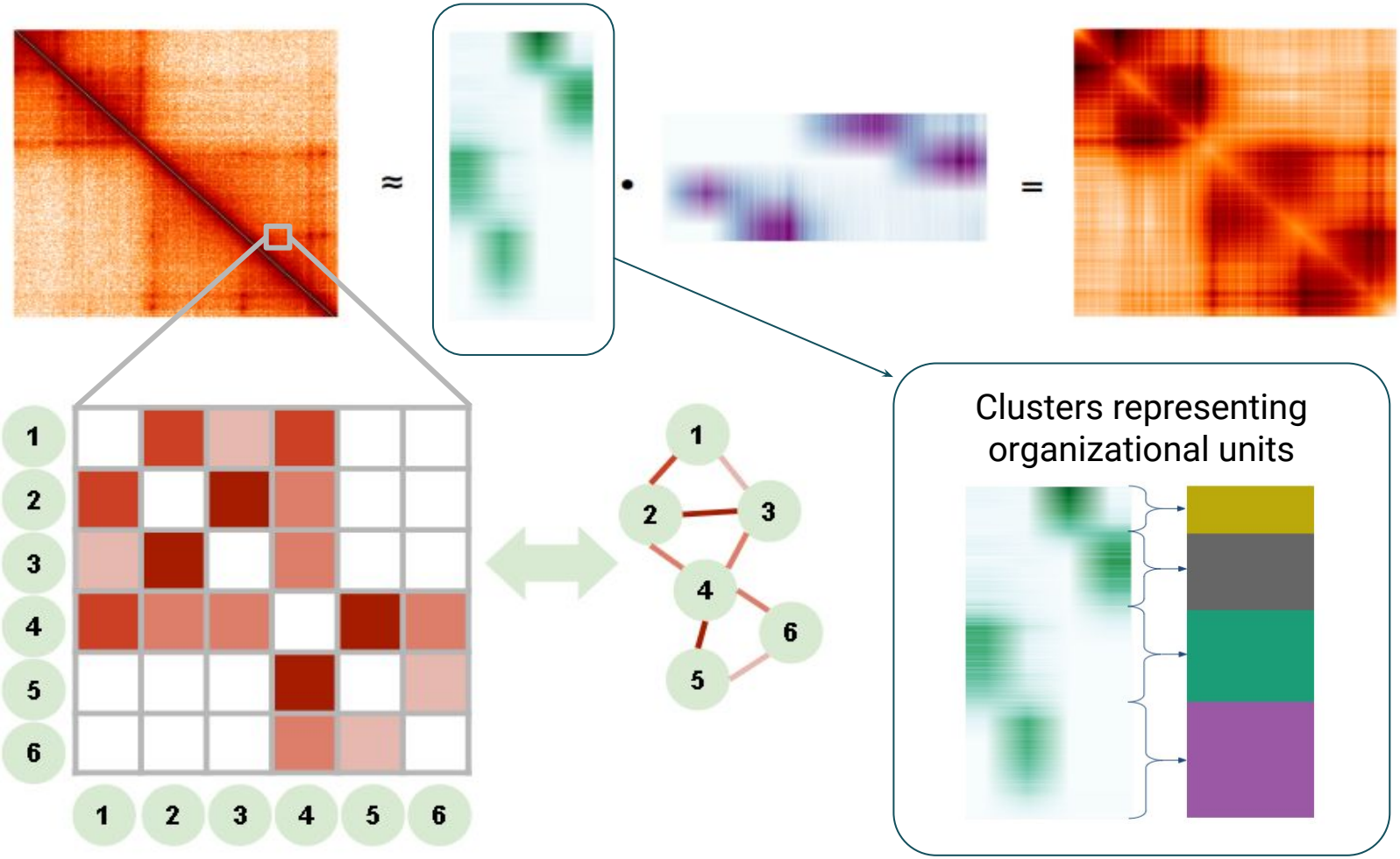
GRINCH: graph-regularized NMF and clustering to analyze Hi-C data



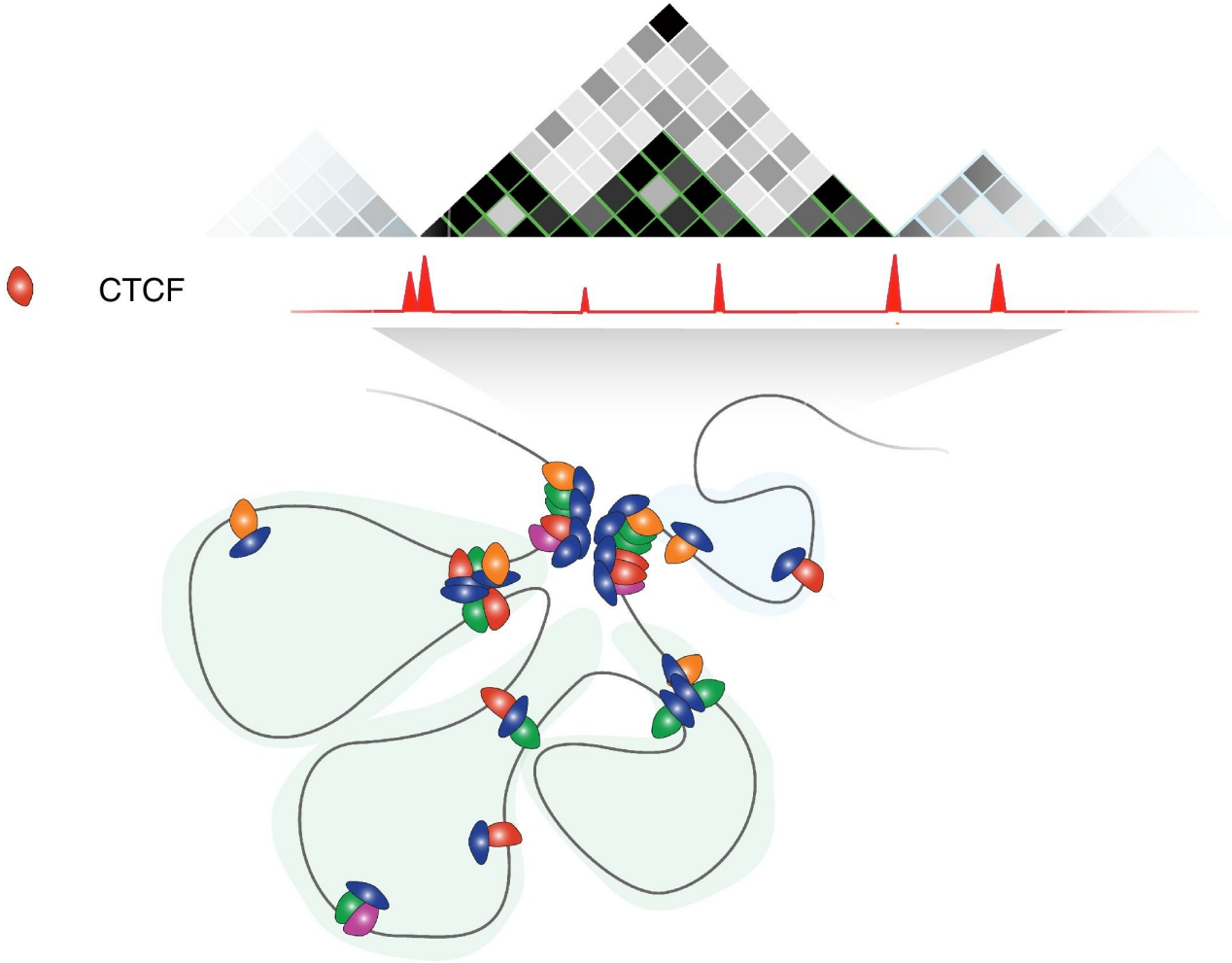
GRINCH: graph-regularized NMF and clustering to analyze Hi-C data



GRINCH: graph-regularized NMF and clustering to analyze Hi-C data

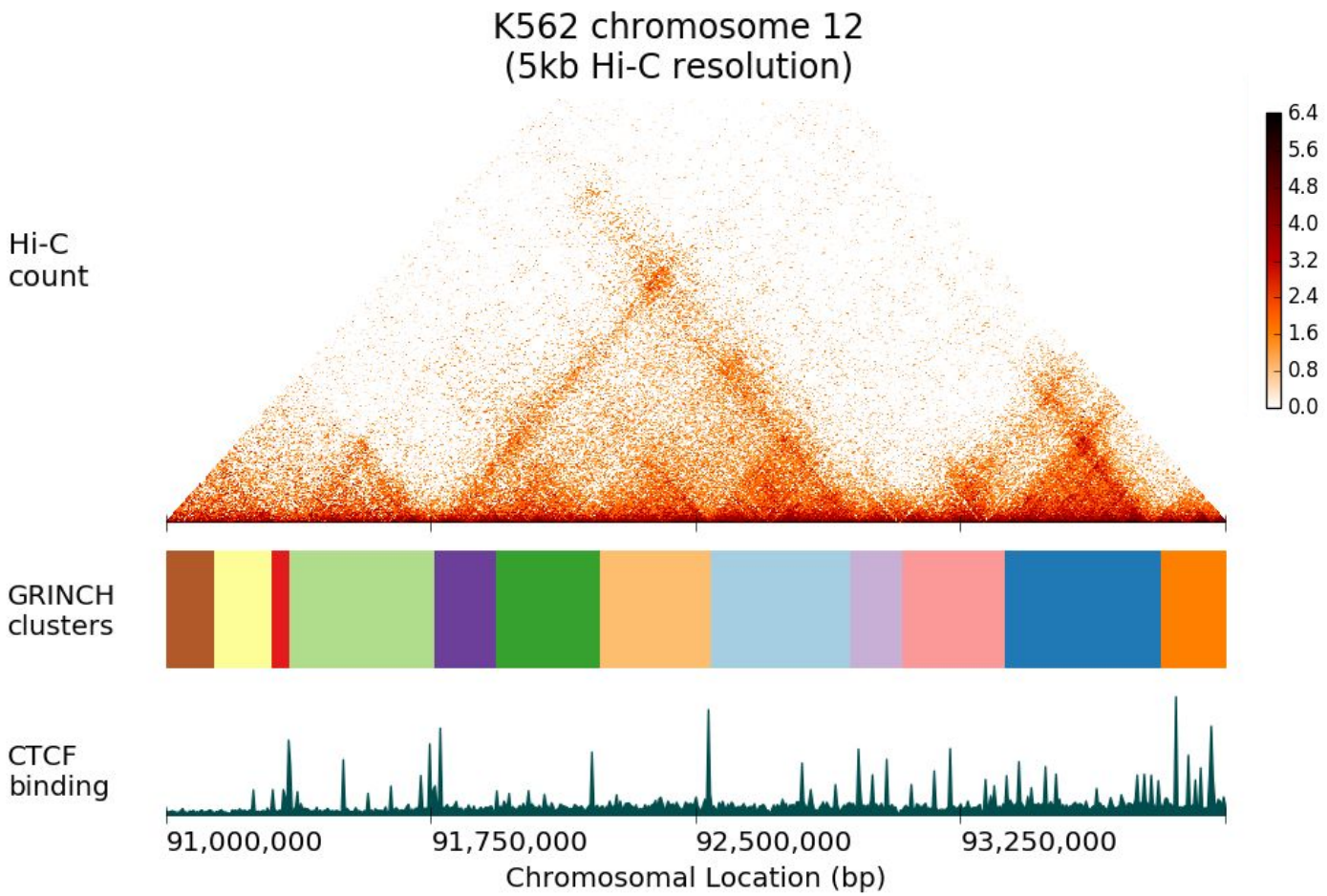


CTCF binding is associated with TAD boundaries

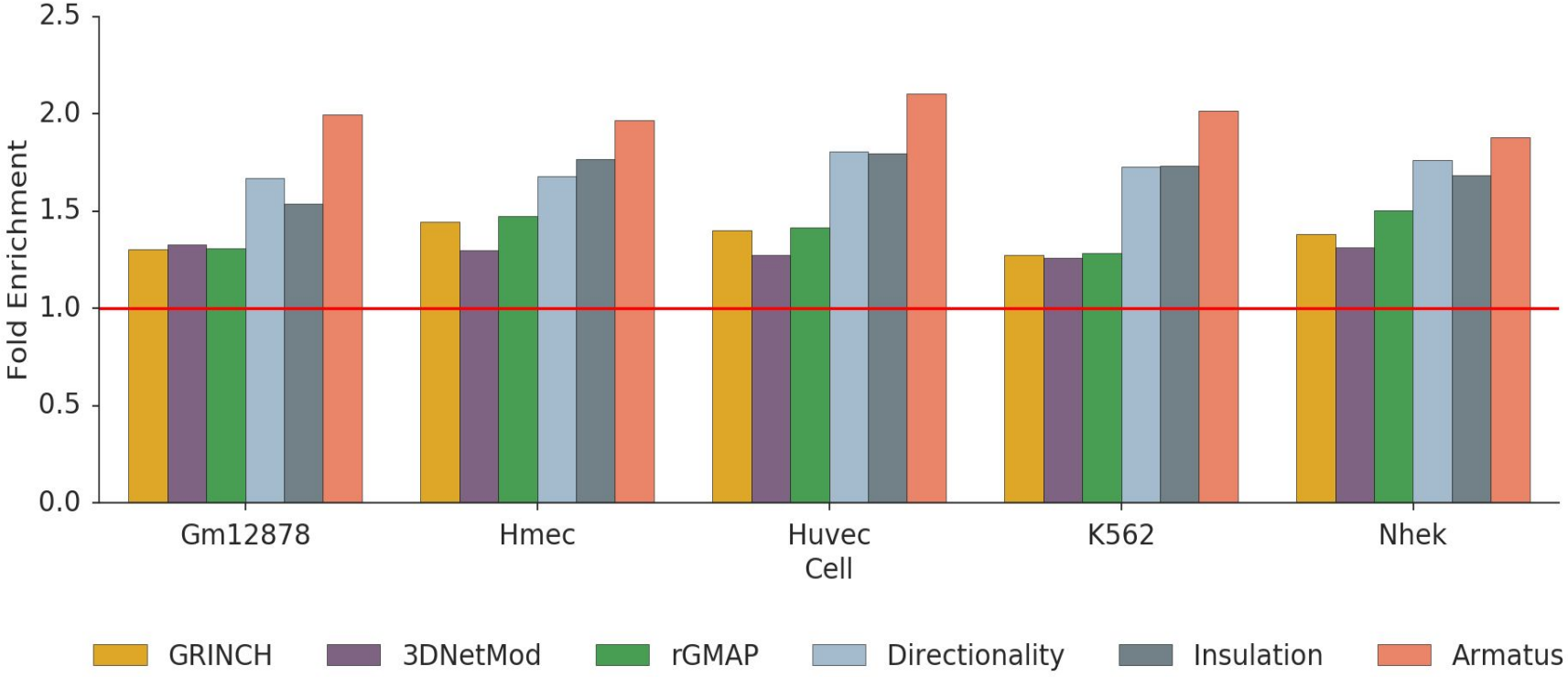


Ali et al. Curr. Opin. Genet. Dev. 2016

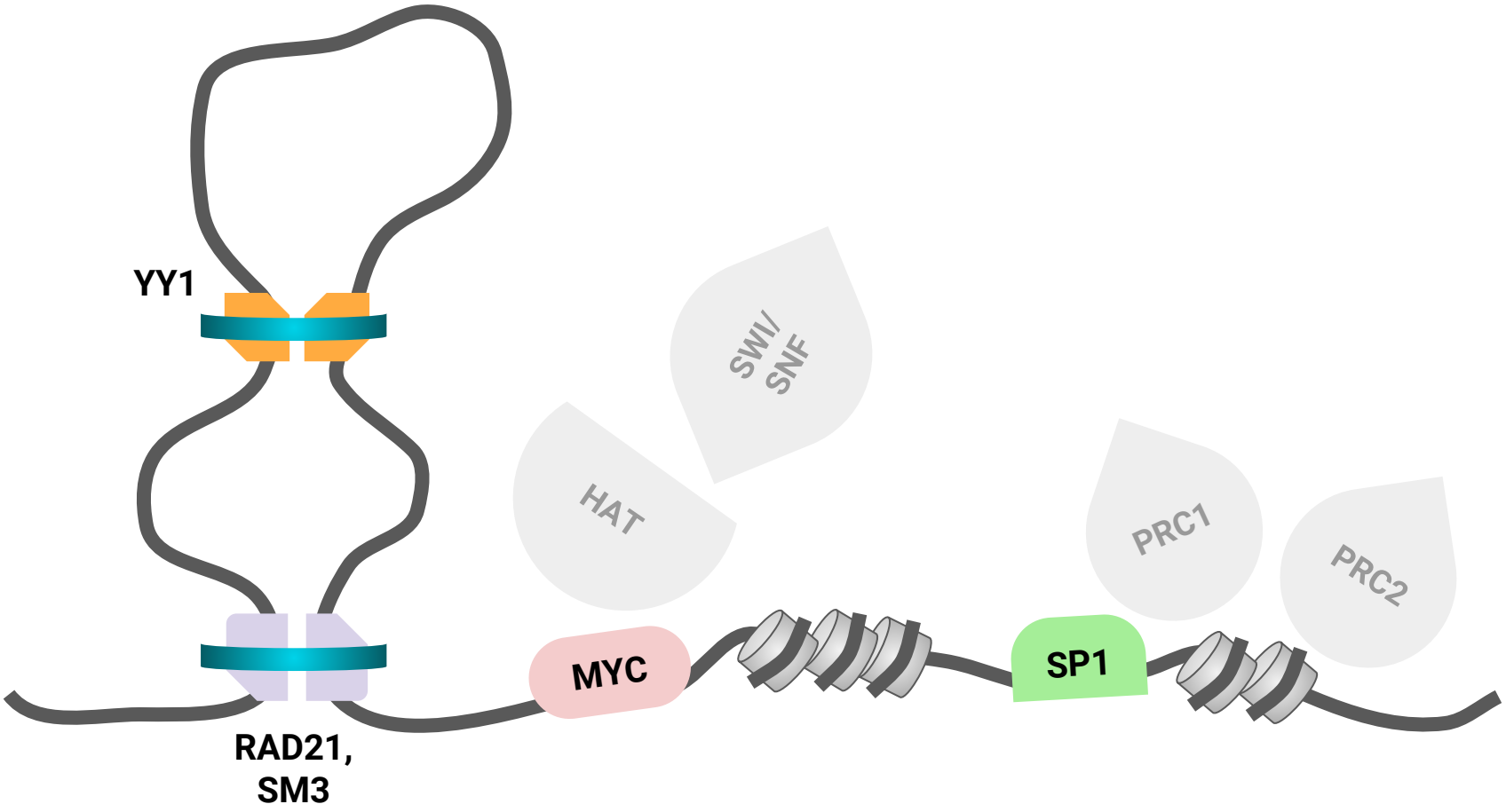
GRINCH cluster boundaries are associated with CTCF signals



GRINCH cluster boundaries are significantly enriched in CTCF binding



Assessing cluster boundaries for architectural proteins



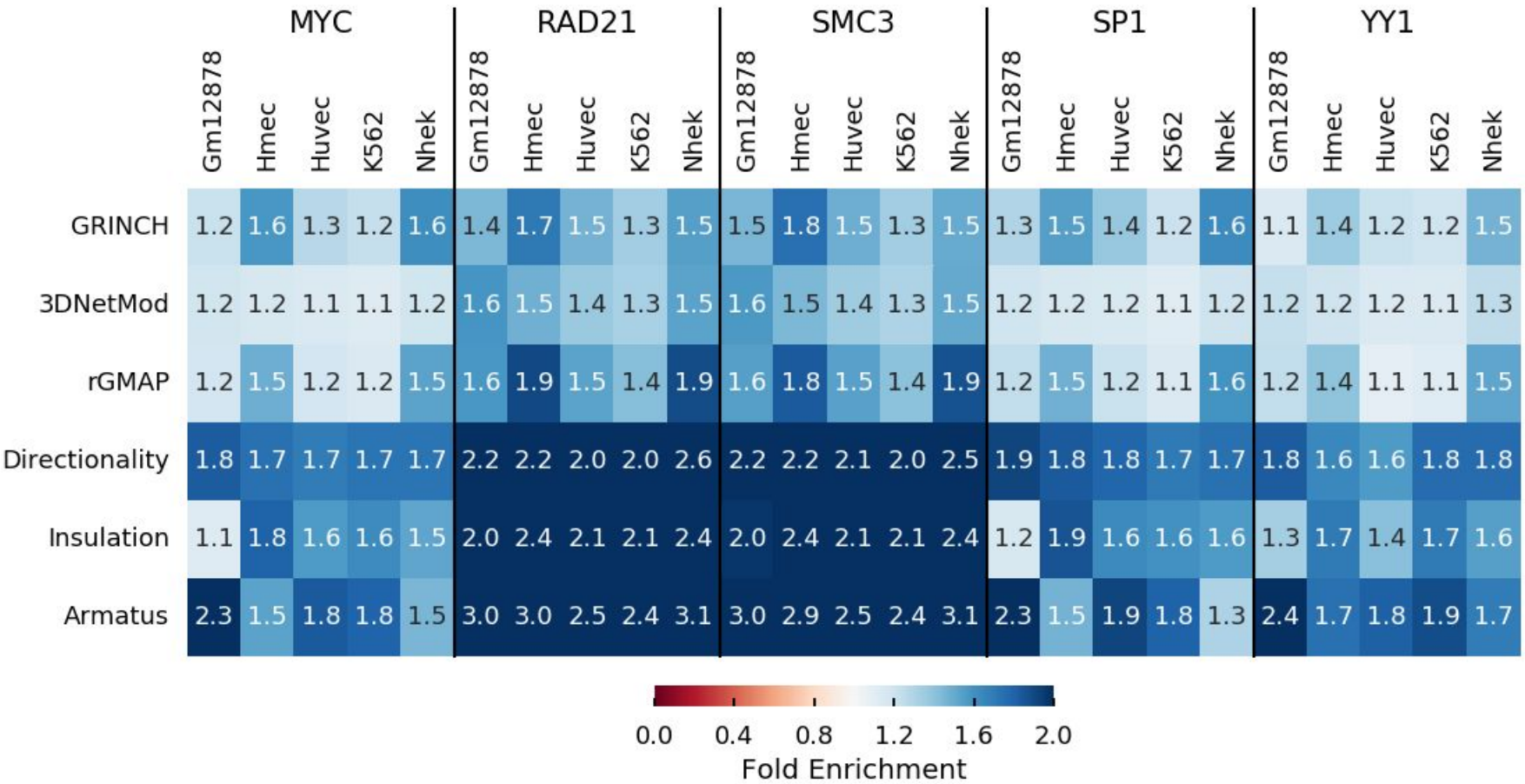
GRINCH cluster boundaries are enriched for different architectural proteins

	MYC					RAD21					SMC3					SP1					YY1				
	Gm12878	Hmec	Huvec	K562	Nhek	Gm12878	Hmec	Huvec	K562	Nhek	Gm12878	Hmec	Huvec	K562	Nhek	Gm12878	Hmec	Huvec	K562	Nhek	Gm12878	Hmec	Huvec	K562	Nhek
GRINCH	1.2	1.6	1.3	1.2	1.6	1.4	1.7	1.5	1.3	1.5	1.5	1.8	1.5	1.3	1.5	1.3	1.5	1.4	1.2	1.6	1.1	1.4	1.2	1.2	1.5

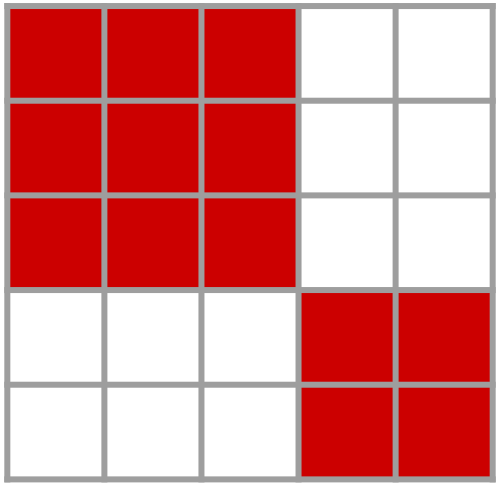
GRINCH cluster boundaries are enriched for different architectural proteins

	MYC					RAD21					SMC3					SP1					YY1				
	Gm12878	Hmec	Huvec	K562	Nhek	Gm12878	Hmec	Huvec	K562	Nhek	Gm12878	Hmec	Huvec	K562	Nhek	Gm12878	Hmec	Huvec	K562	Nhek	Gm12878	Hmec	Huvec	K562	Nhek
GRINCH	1.2	1.6	1.3	1.2	1.6	1.4	1.7	1.5	1.3	1.5	1.5	1.8	1.5	1.3	1.5	1.3	1.5	1.4	1.2	1.6	1.1	1.4	1.2	1.2	1.5
3DNetMod	1.2	1.2	1.1	1.1	1.2	1.6	1.5	1.4	1.3	1.5	1.6	1.5	1.4	1.3	1.5	1.2	1.2	1.2	1.1	1.2	1.2	1.2	1.2	1.1	1.3
rGMAP	1.2	1.5	1.2	1.2	1.5	1.6	1.9	1.5	1.4	1.9	1.6	1.8	1.5	1.4	1.9	1.2	1.5	1.2	1.1	1.6	1.2	1.4	1.1	1.1	1.5

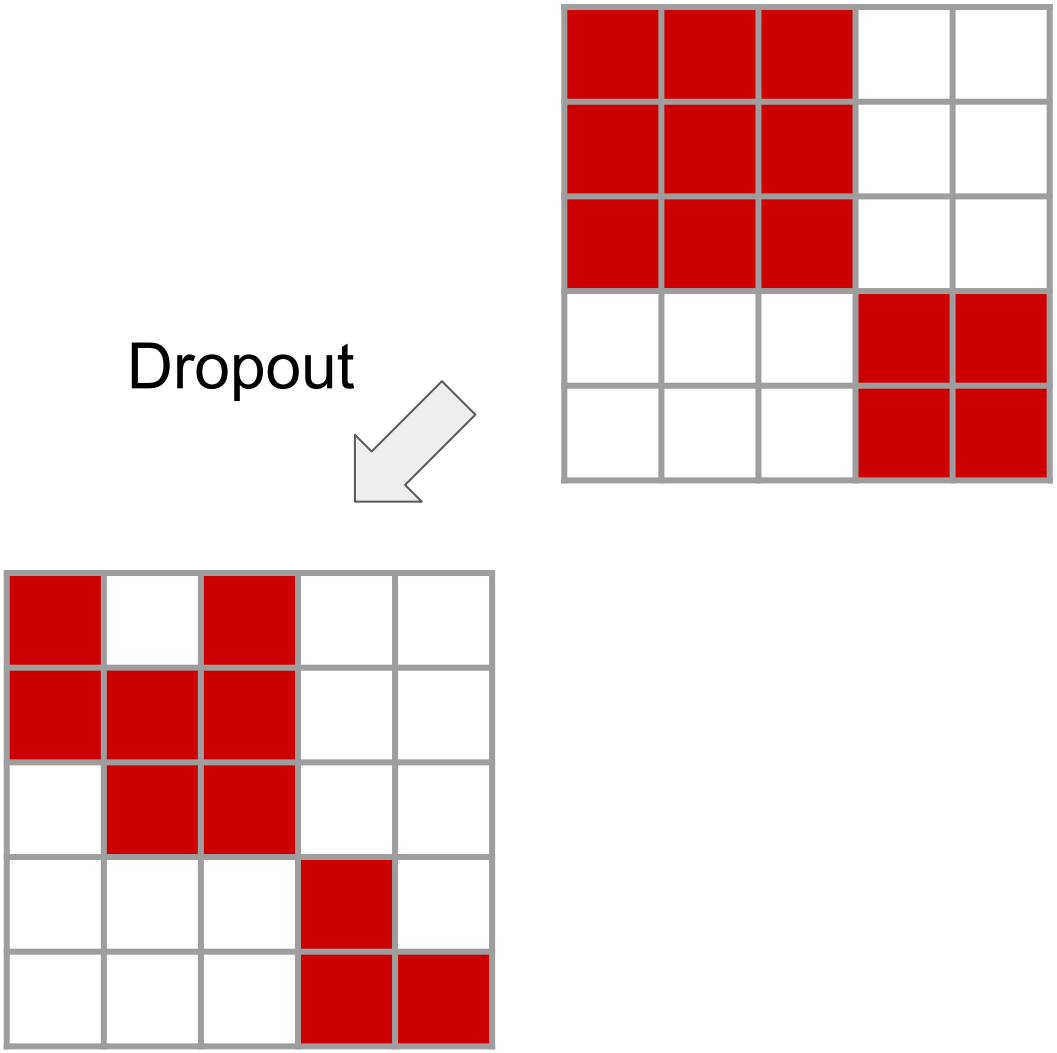
GRINCH cluster boundaries are enriched for different architectural proteins



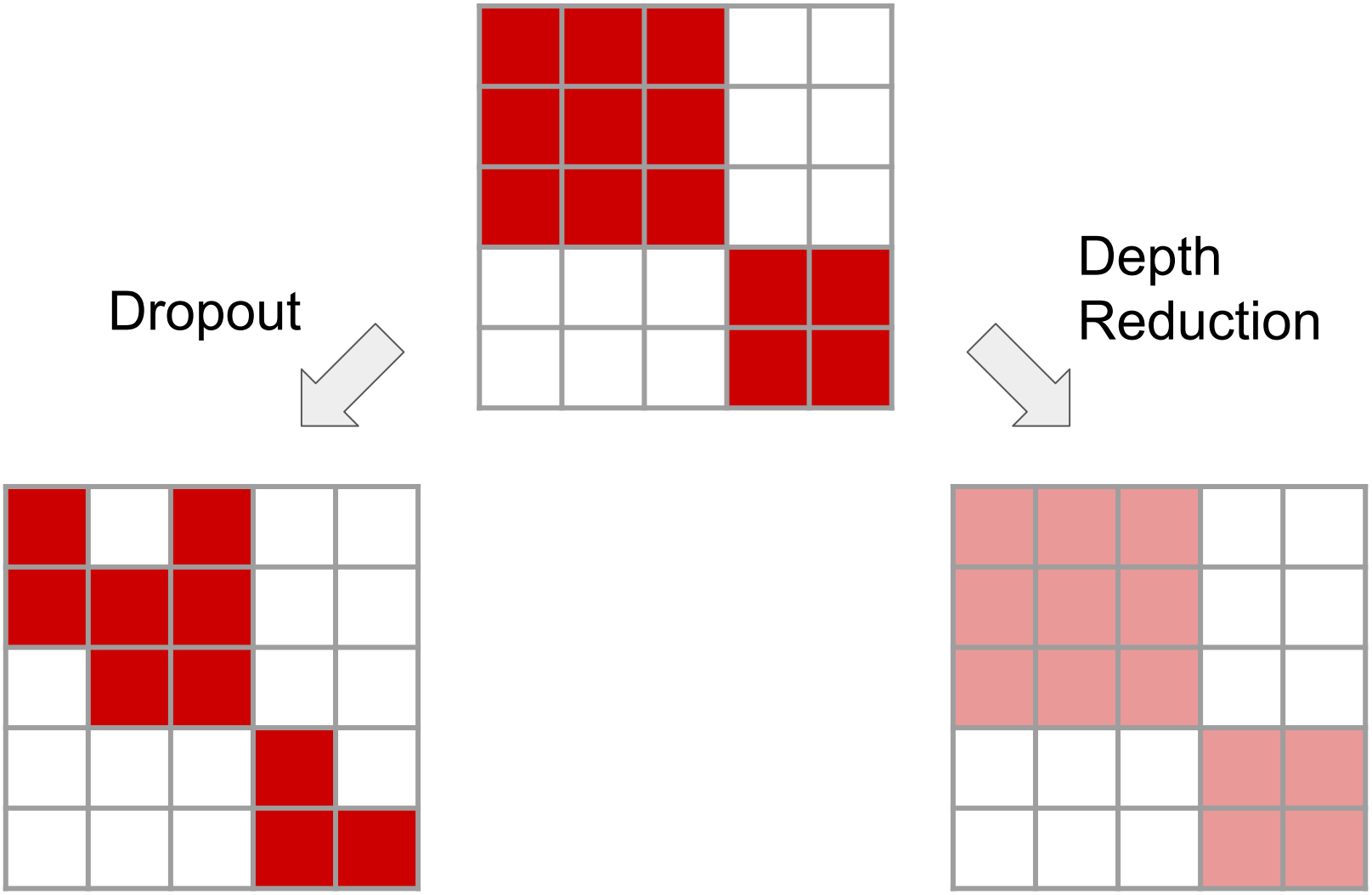
Simulating sparsity to test stability



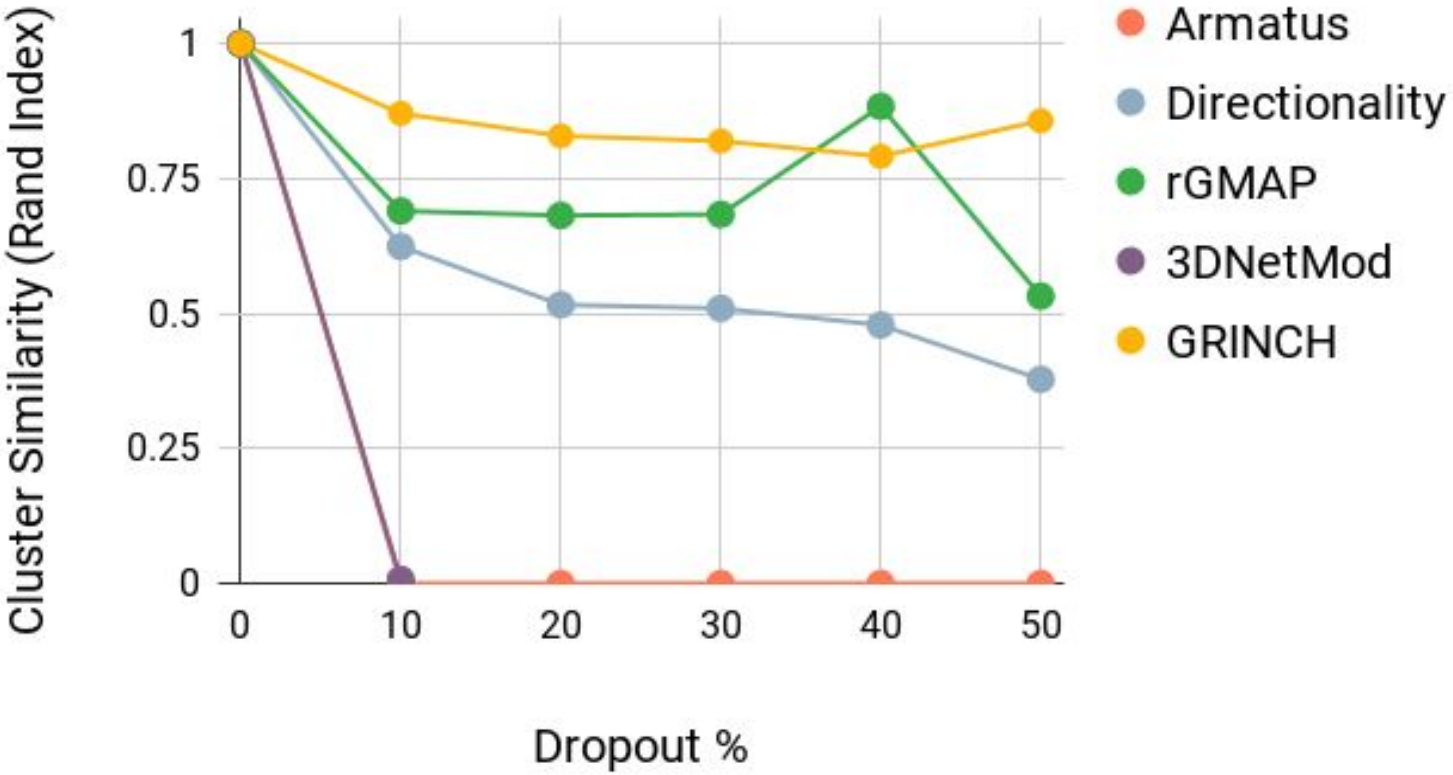
Simulating sparsity to test stability



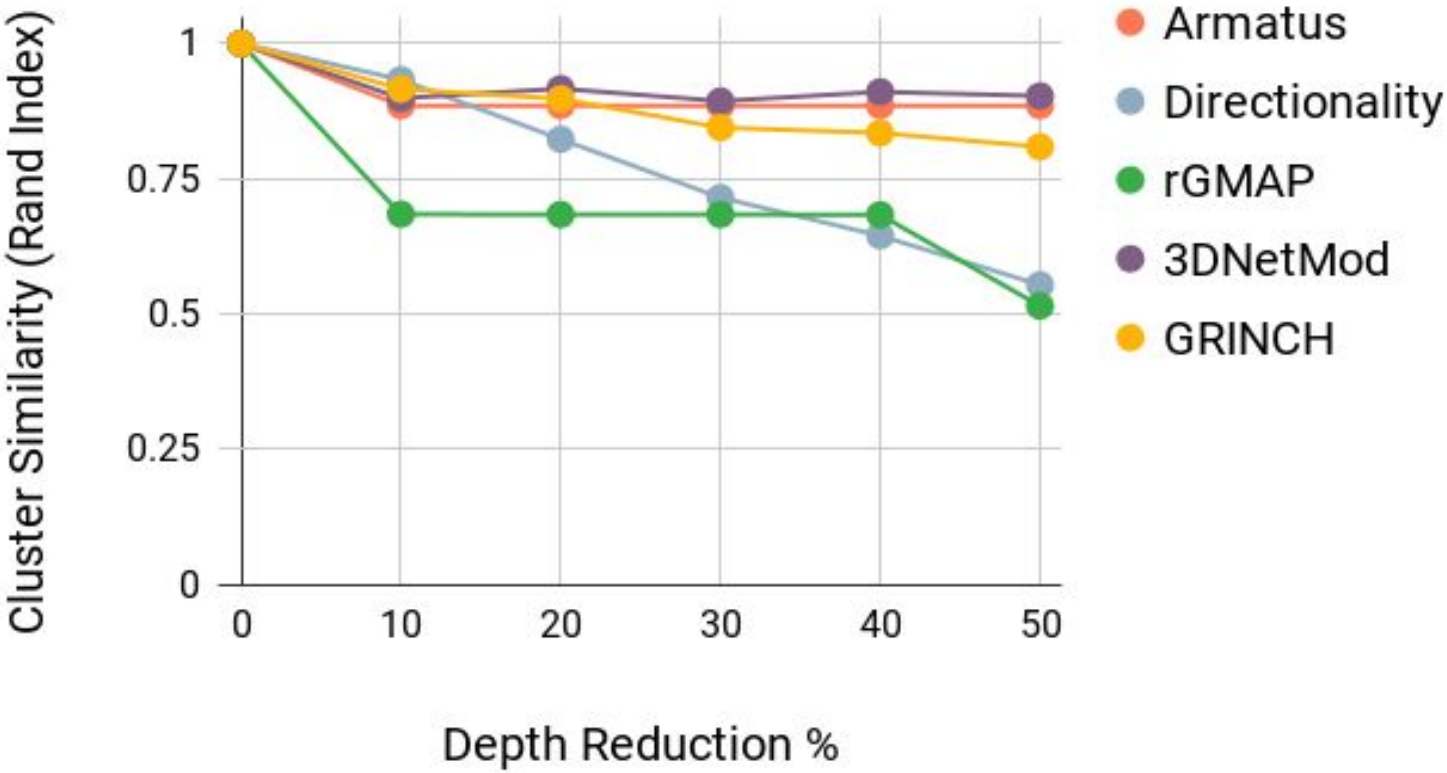
Simulating sparsity to test stability



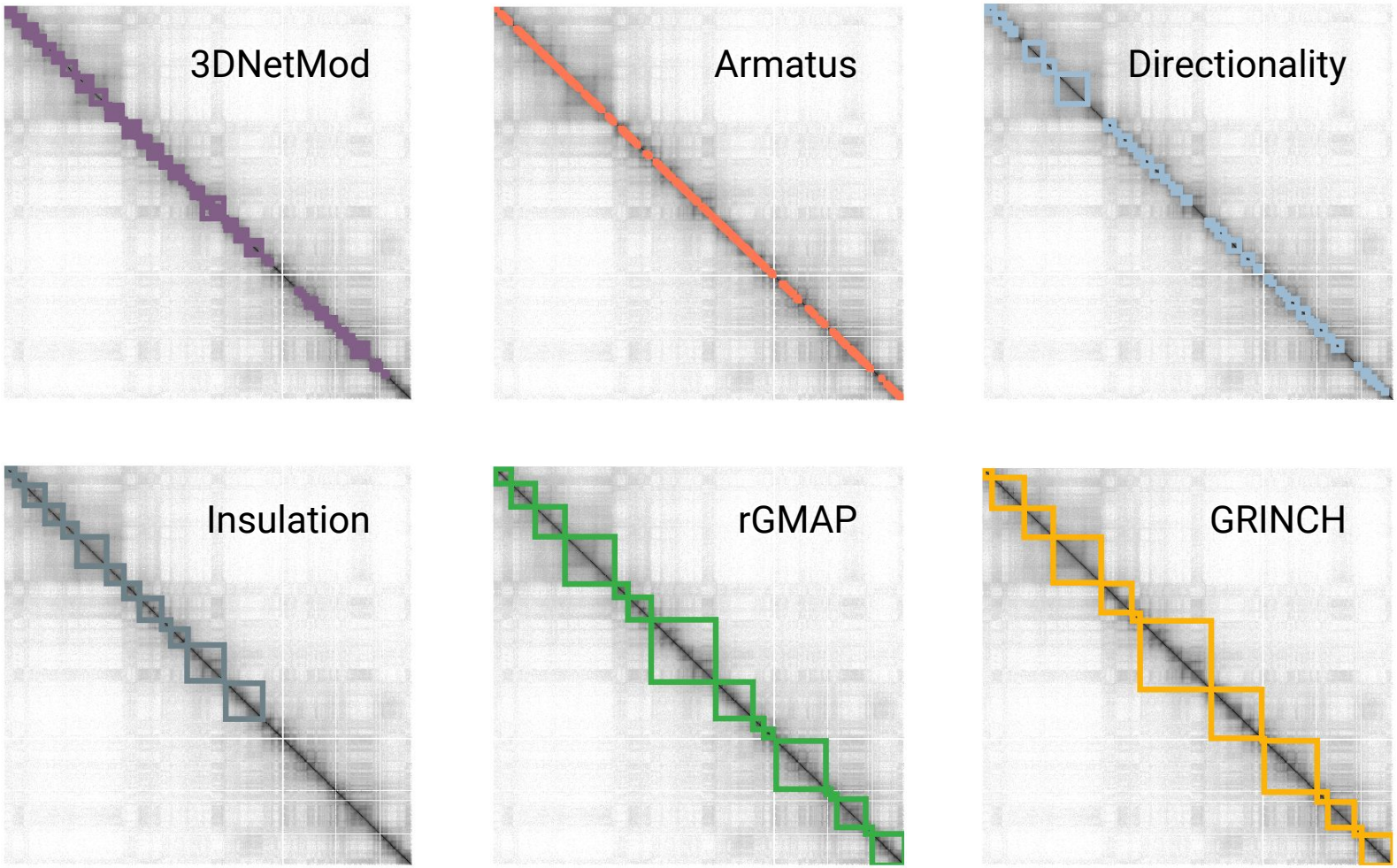
GRINCH is the most stable method to dropout



GRINCH is robust to lower-depth data

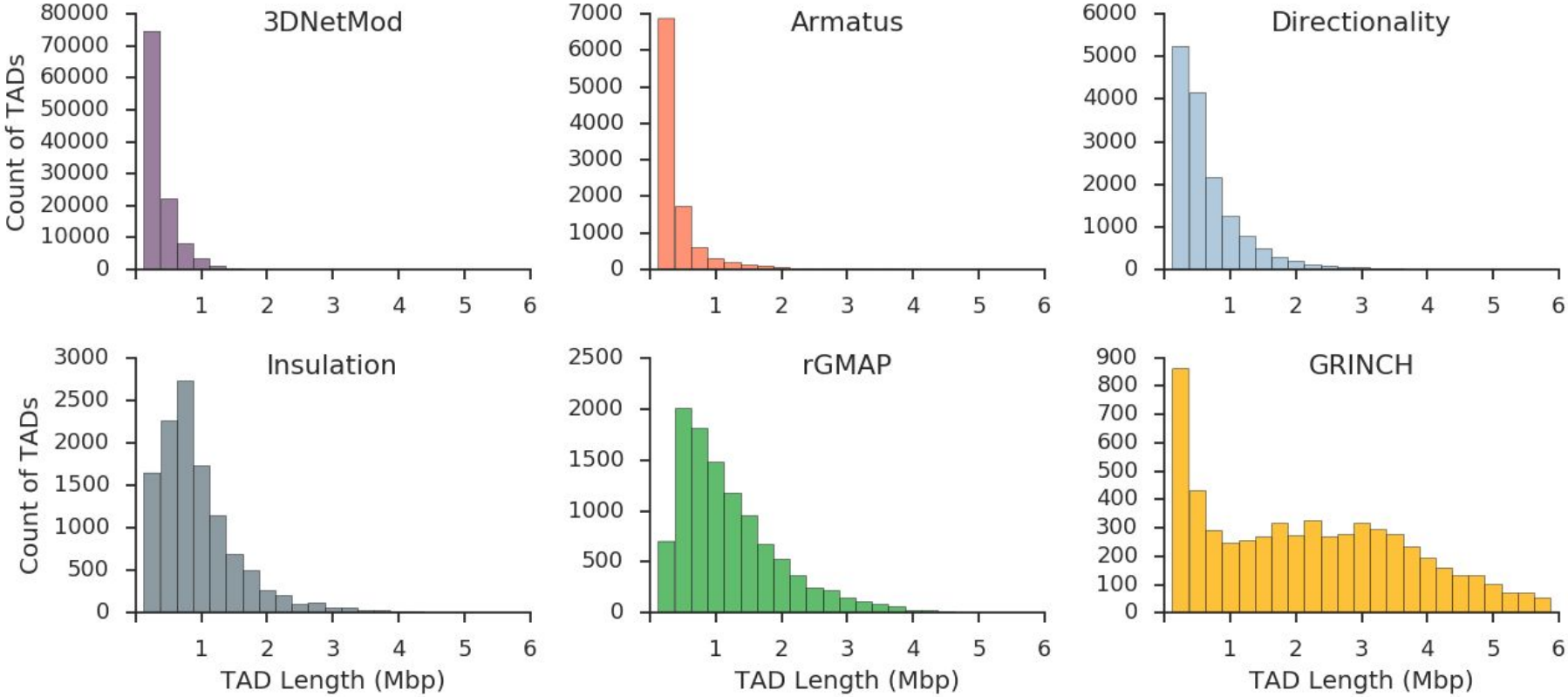


GRINCH captures a wide range of domain sizes

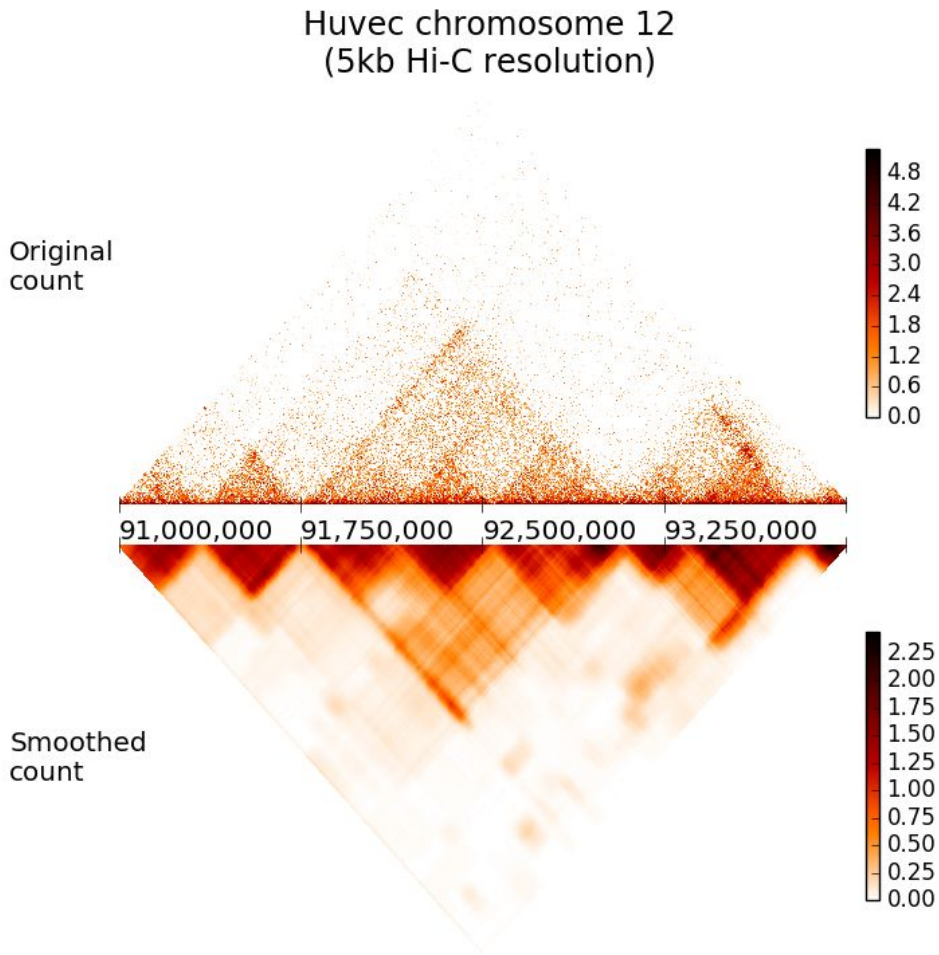


Gm12878 chr9 115,850,000-141,100,000bp

GRINCH captures a wide range of domain sizes



GRINCH can smooth Hi-C matrix through matrix completion



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 very stable to noisy datasets.
- GRINCH can find TADs of diverse lengths.
- GRINCH can smooth input Hi-C matrix.

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 very stable to noisy datasets.
- GRINCH can find TADs of diverse lengths.
- GRINCH can smooth input Hi-C matrix.

Poster A-73 on GRINCH

Acknowledgements

Members of Roy lab:

Sushmita Roy
Brittany Baur
Shilu Zhang
Deborah Chasman
Alireza Siahpirani
Sara Knaack
Jon Ide
Junha Shin
Sunnie Grace McCalla

Funding sources:

Center for Predictive Computational Phenotyping (NIH BD2K U54 AI117924)

NIH NIGMS 1R01GM117339

Quantitative Biology Initiative

Computing resources:

Center for High Throughput Computing (CHTC)