

# **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



WID.WISC.EDU

# Long-range gene regulation by distal elements

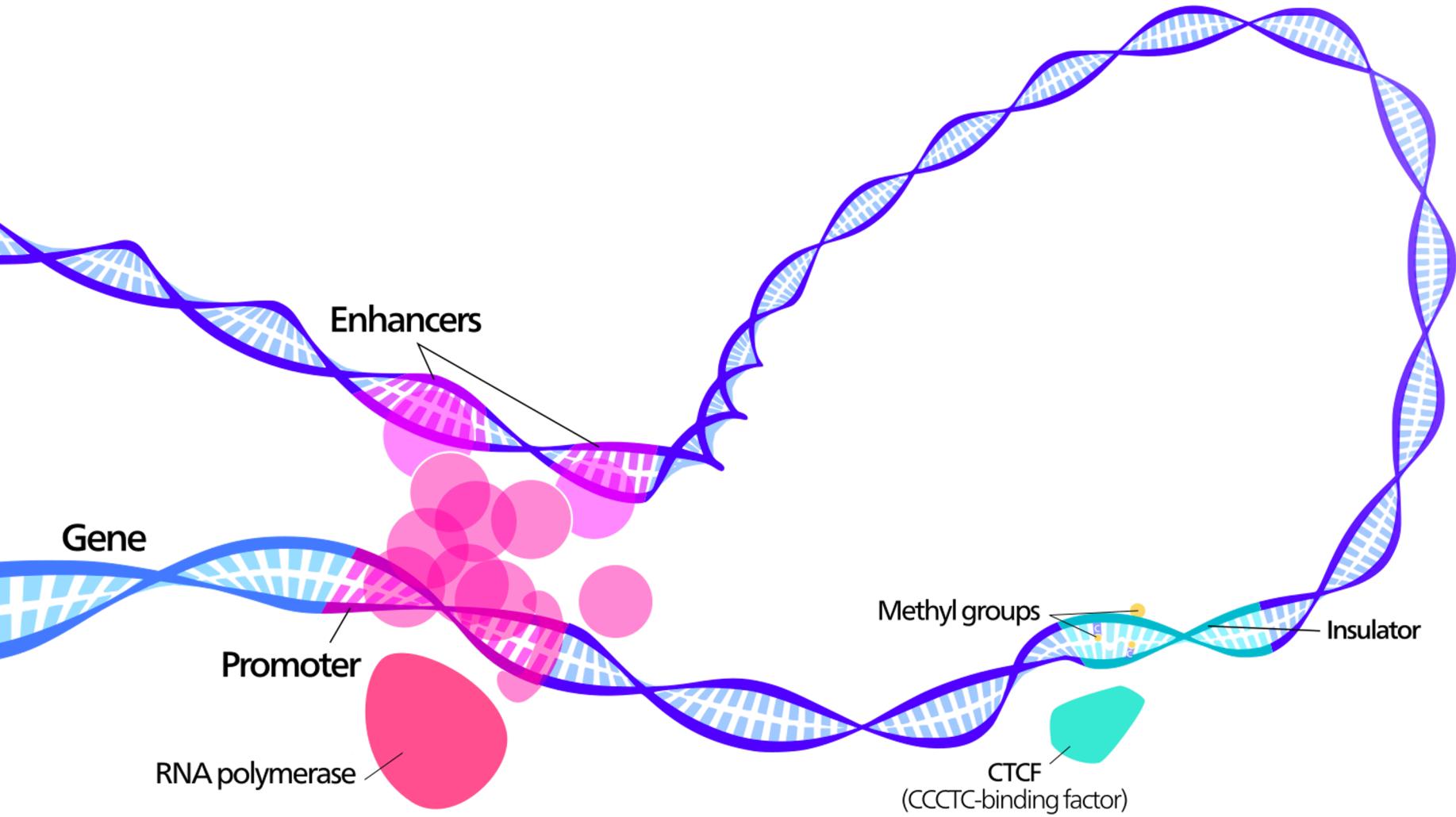
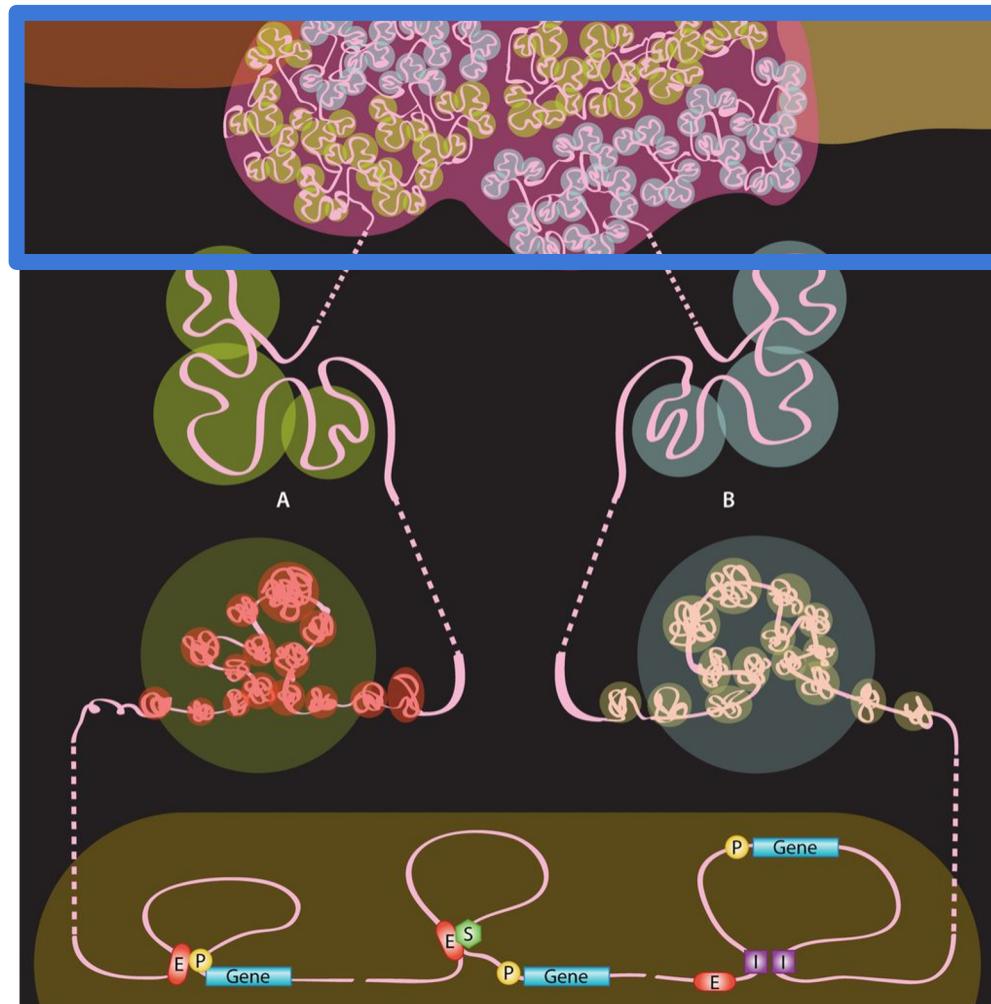


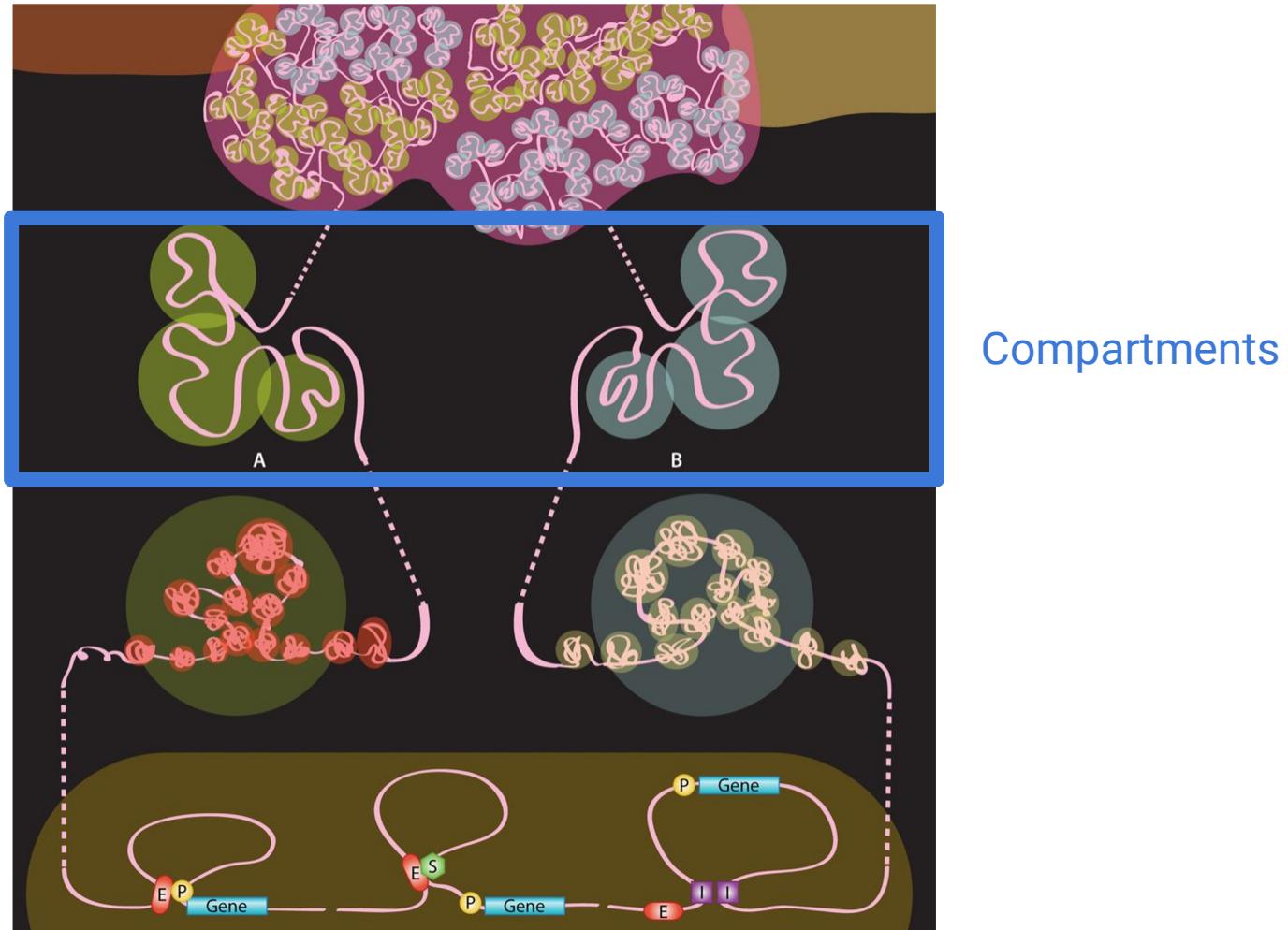
Illustration by Kelvin Ma

# Genome is organized into higher-order domains at multiple scales

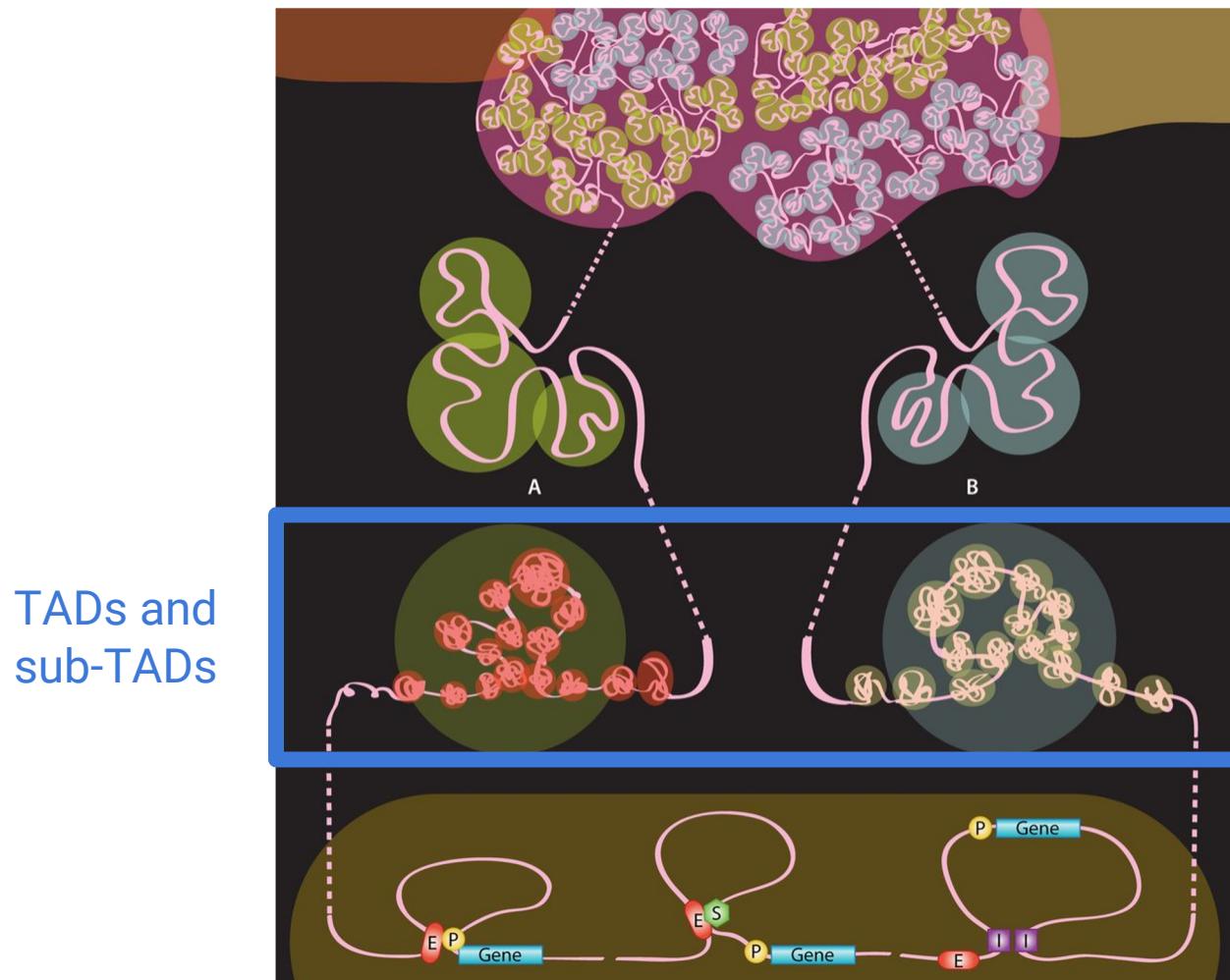
Chromosomal territories



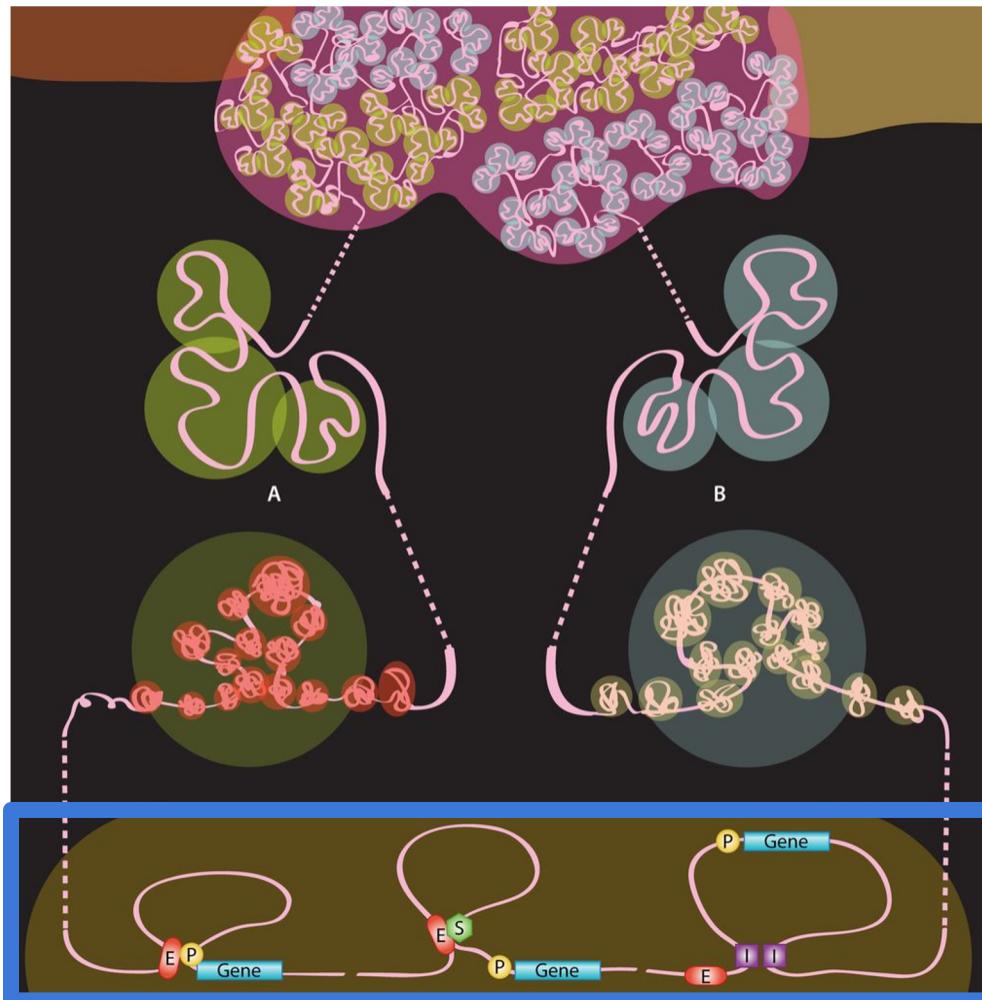
# Genome is organized into higher-order domains at multiple scales



# Genome is organized into higher-order domains at multiple scales



# Genome is organized into higher-order domains at multiple scales



# Existing methods for finding topological units of chromosomes

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



A 7x7 matrix representing user ratings for movies. The rows represent users and the columns represent movies. The matrix contains non-negative integer values ranging from 0 to 5.

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

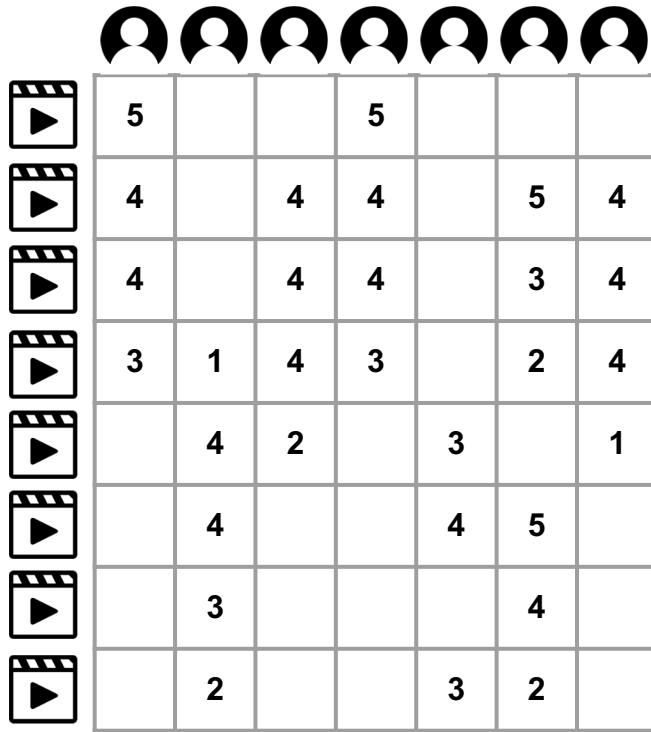


A 7x7 matrix representing a non-negative matrix. The matrix has 7 rows and 7 columns. The first row contains the values 5, empty, empty, empty, 5, empty, empty. The second row contains the values 4, empty, 4, 4, empty, 5, 4. The third row contains the values 4, empty, 4, 4, empty, 3, 4. The fourth row contains the values 3, 1, 4, 3, empty, 2, 4. The fifth row contains the values 4, 2, empty, 3, empty, empty, 1. The sixth row contains the values 4, empty, empty, empty, 4, 5, empty. The seventh row contains the values 3, empty, empty, empty, 4, empty, empty.

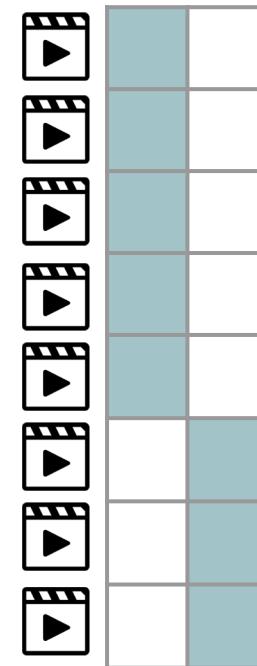
5				5		
4		4	4		5	4
4		4	4		3	4
3	1	4	3		2	4
4	2		3			1
4				4	5	
3				4		
2				3	2	

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

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

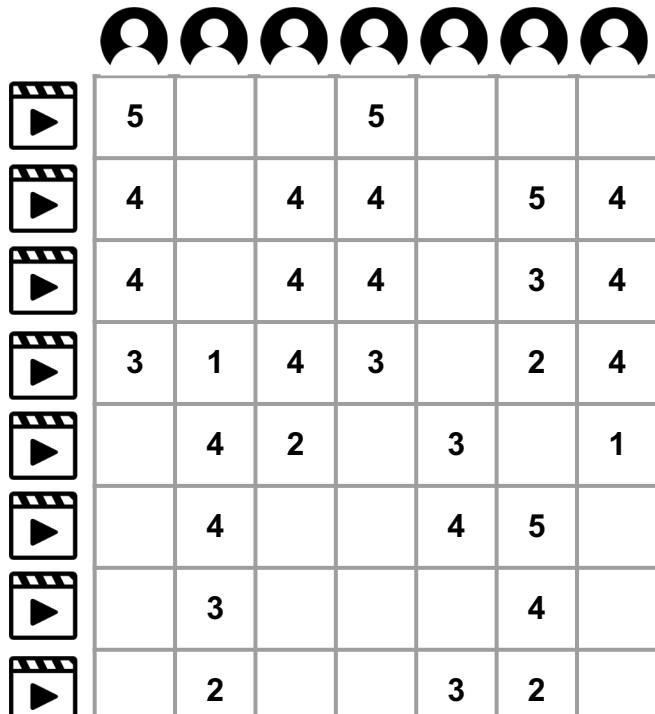


$$X = \mathbb{R}^{n \times m}$$

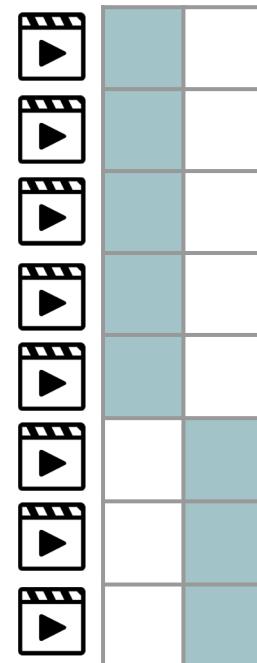


$$U = \mathbb{R}^{n \times k}$$

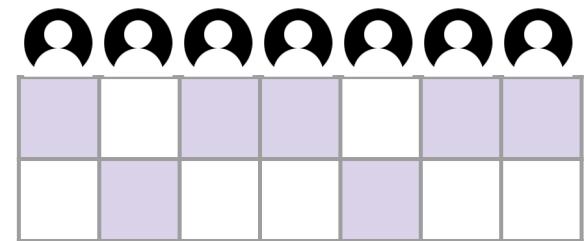
# Non-negative matrix factorization (NMF) reduces dimensions of data



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

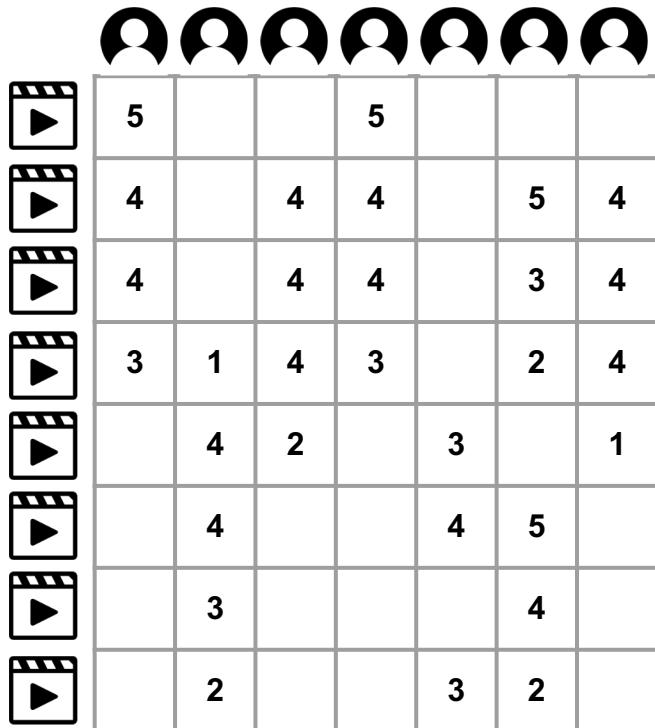


$$U = \mathbb{R}^{n \times k}$$

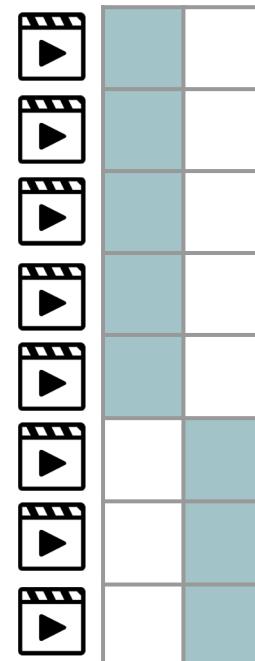


$$V^T = \mathbb{R}^{k \times m}$$

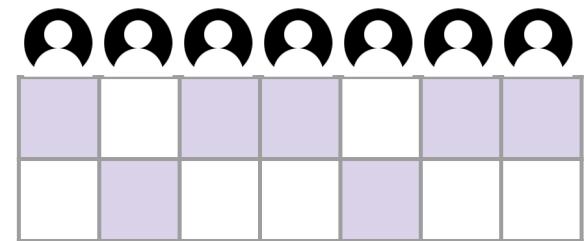
# Non-negative matrix factorization (NMF) reduces dimensions of data



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



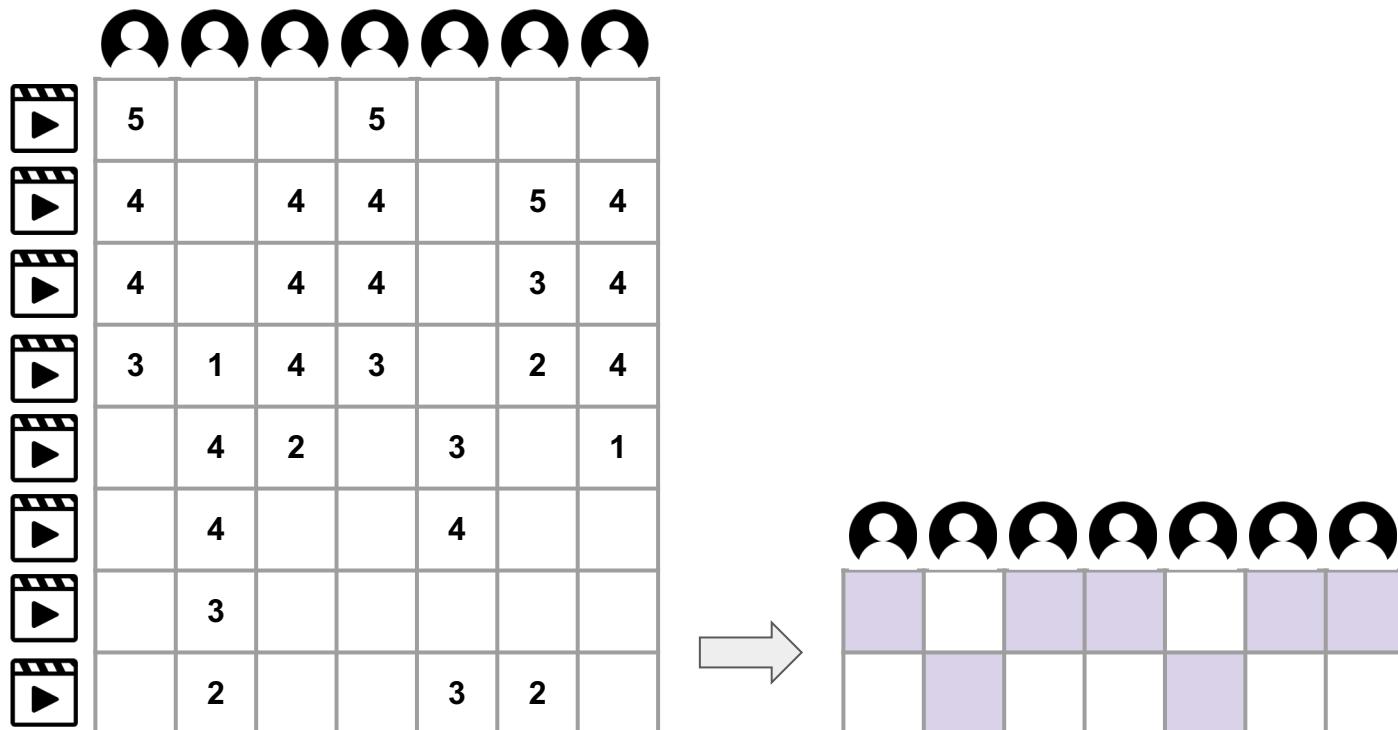
$$U = \mathbb{R}^{n \times k}$$



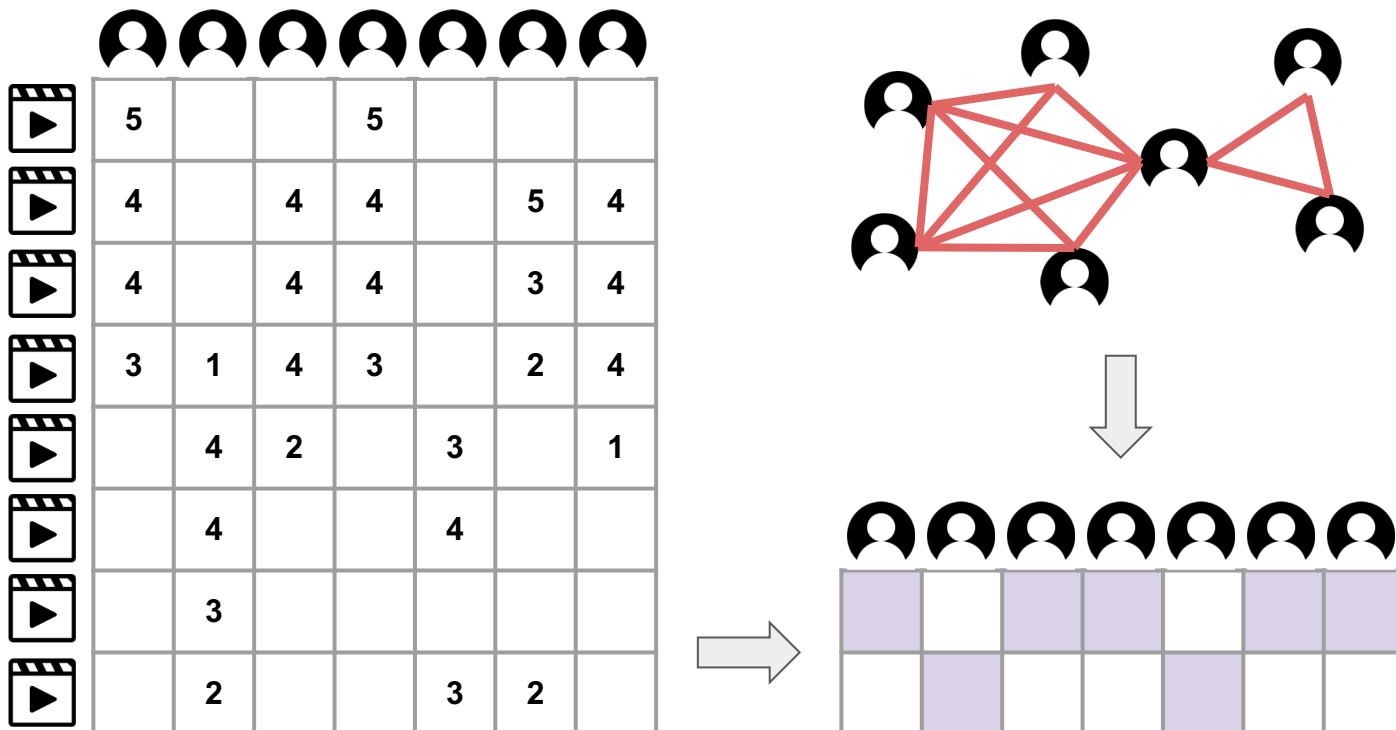
$$V^T = \mathbb{R}^{k \times m}$$

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

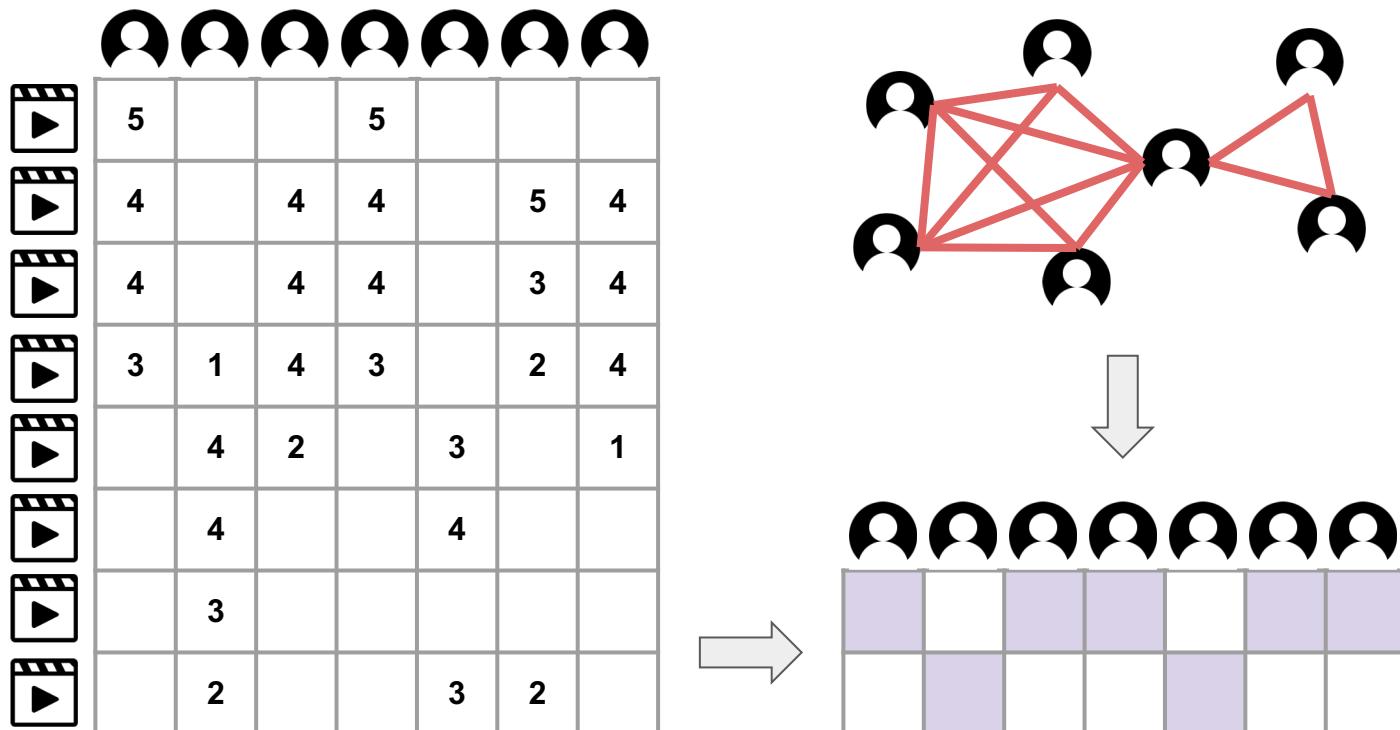
# Graph regularization incorporates prior knowledge in network form



# Graph regularization incorporates prior knowledge in network form

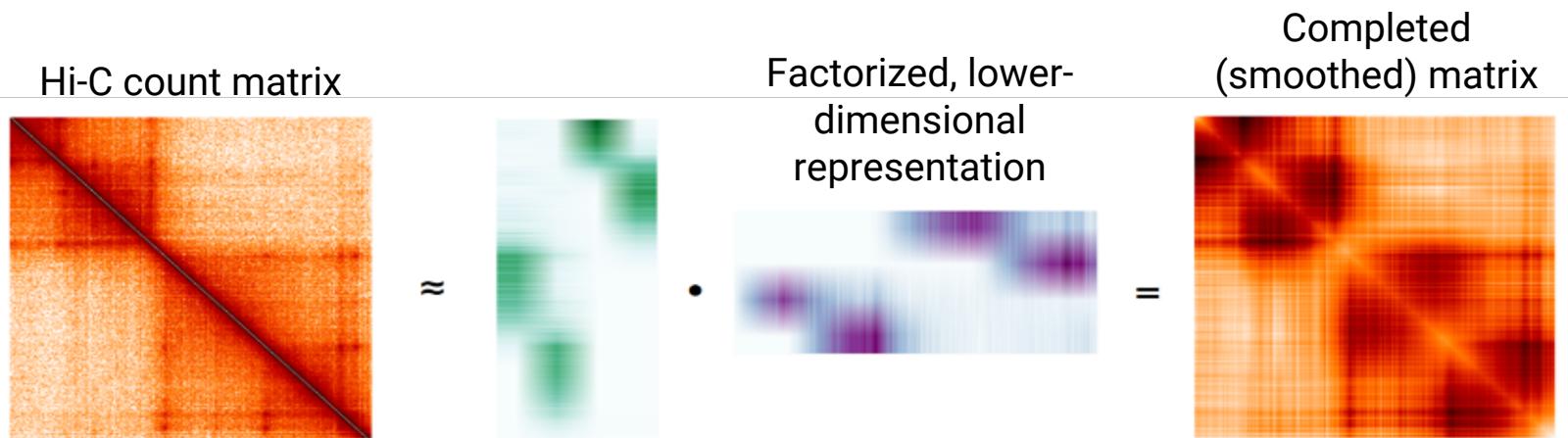


# Graph regularization incorporates prior knowledge in network form

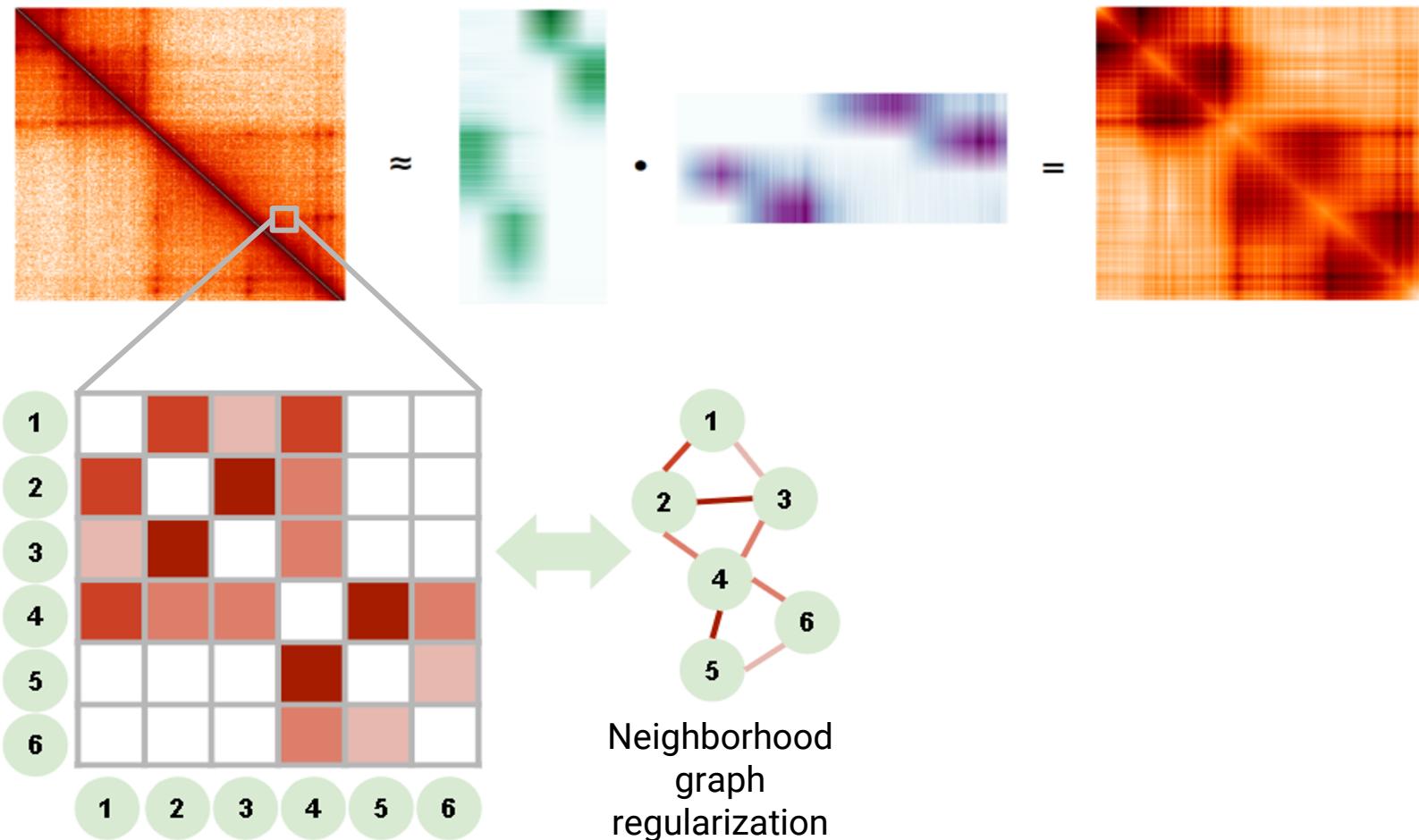


$$\text{Minimize } O = \|X - UV^T\|^2 + \lambda \text{Tr}(V^T L V) + \lambda \text{Tr}(U^T L U)$$

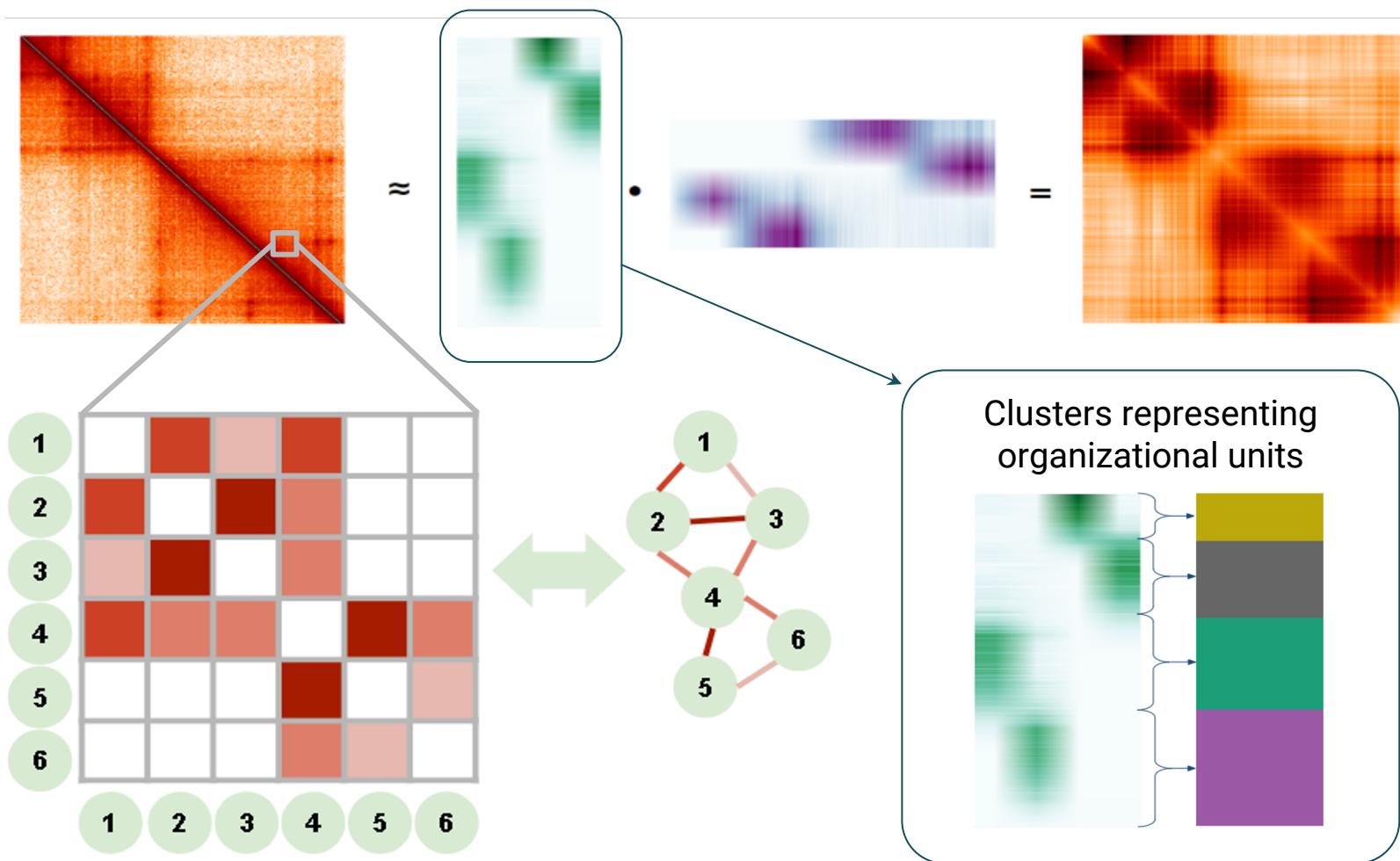
# GRINCH: graph-regularized NMF and clustering to analyze Hi-C data



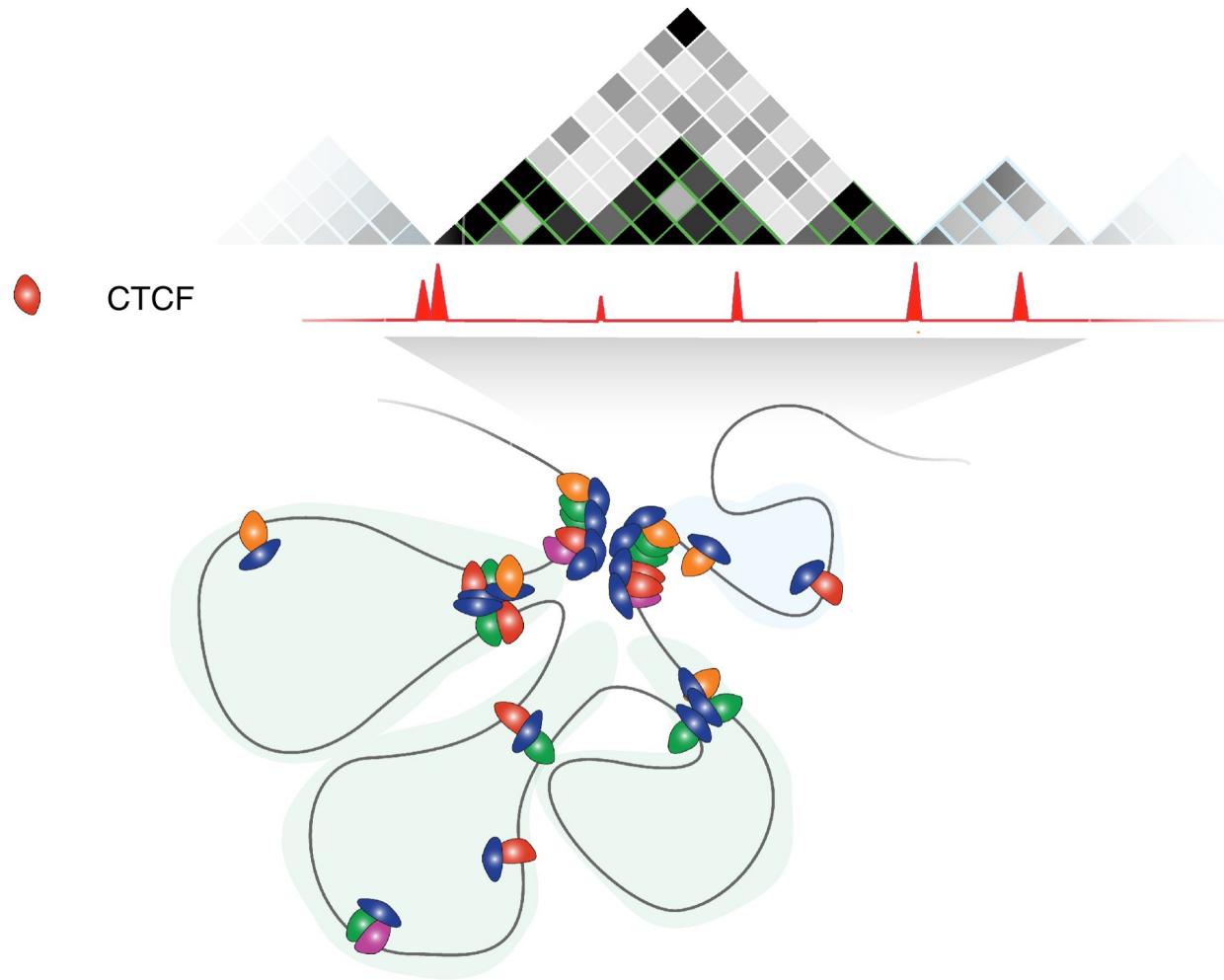
# GRINCH: graph-regularized NMF and clustering to analyze Hi-C data



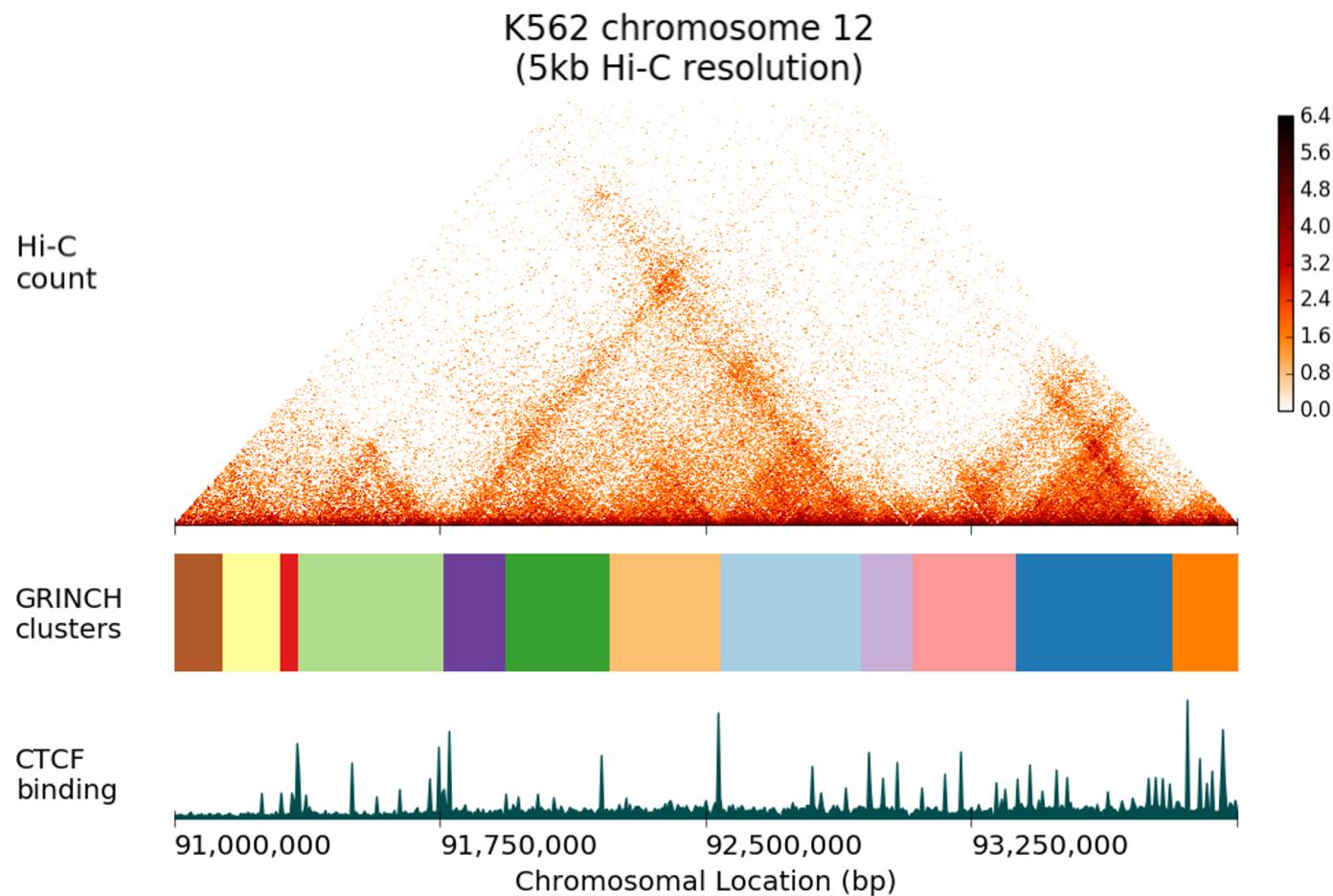
# GRINCH: graph-regularized NMF and clustering to analyze Hi-C data



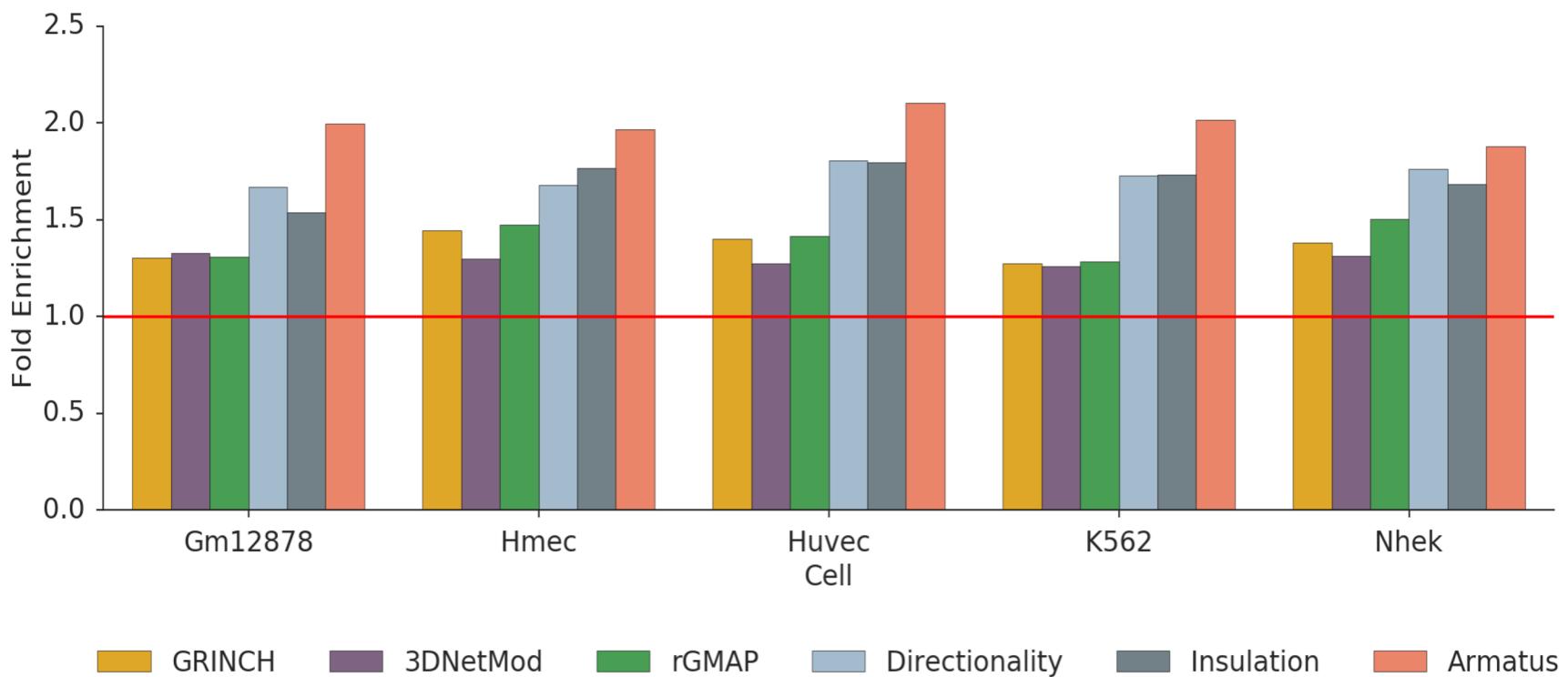
# CTCF binding is associated with TAD boundaries



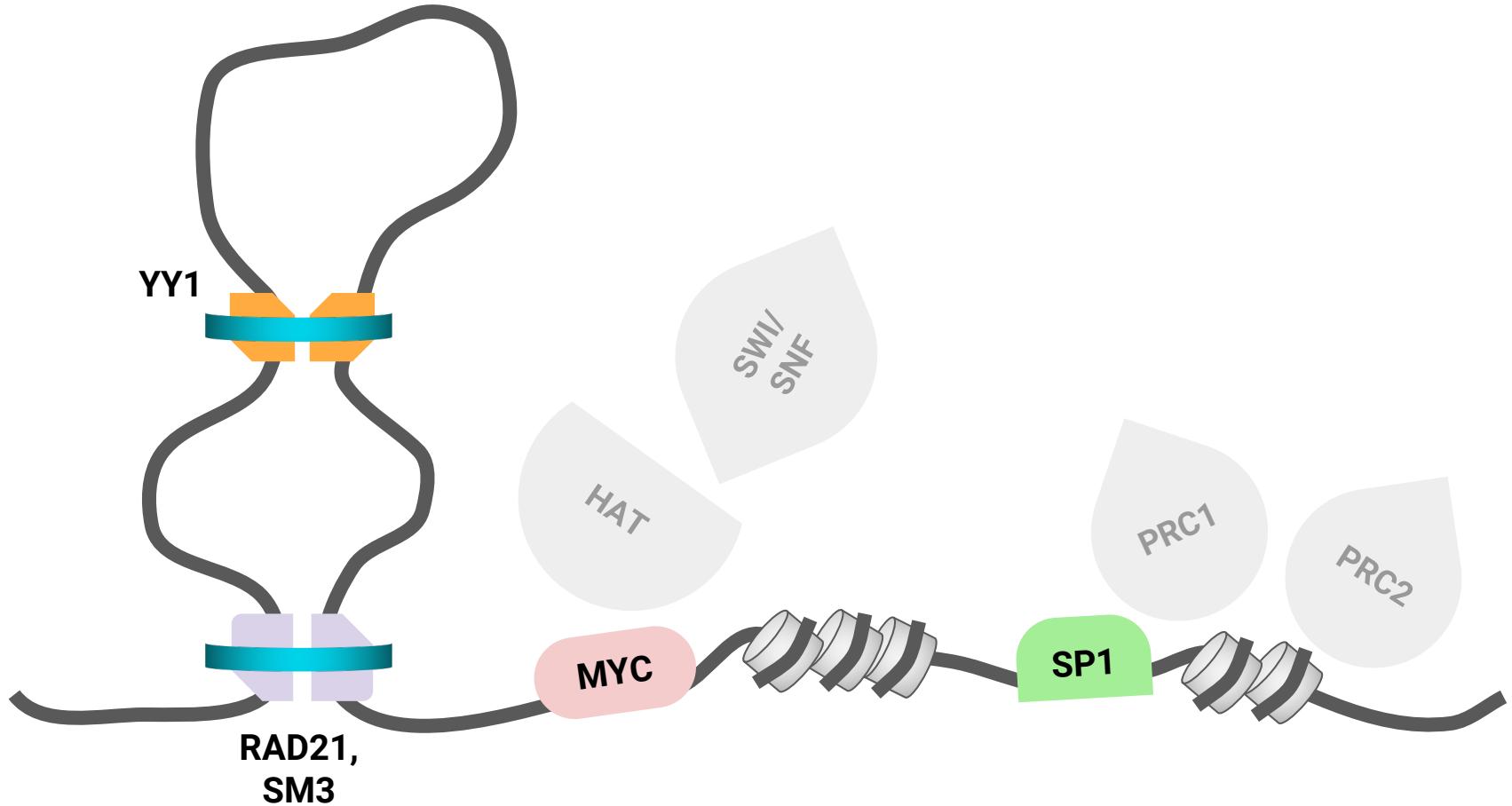
# GRINCH cluster boundaries are associated with CTCF signals



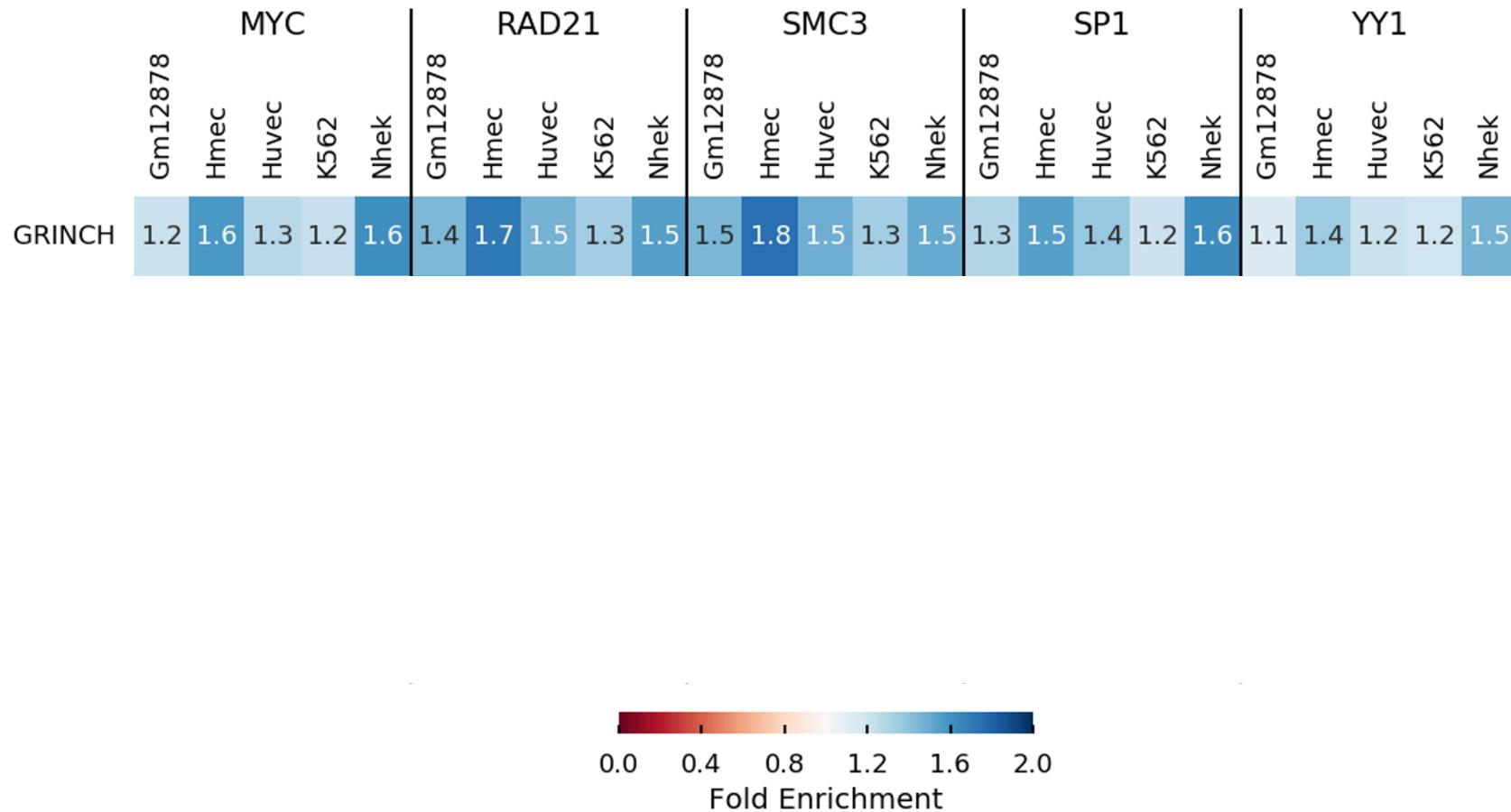
# GRINCH cluster boundaries are significantly enriched in CTCF binding



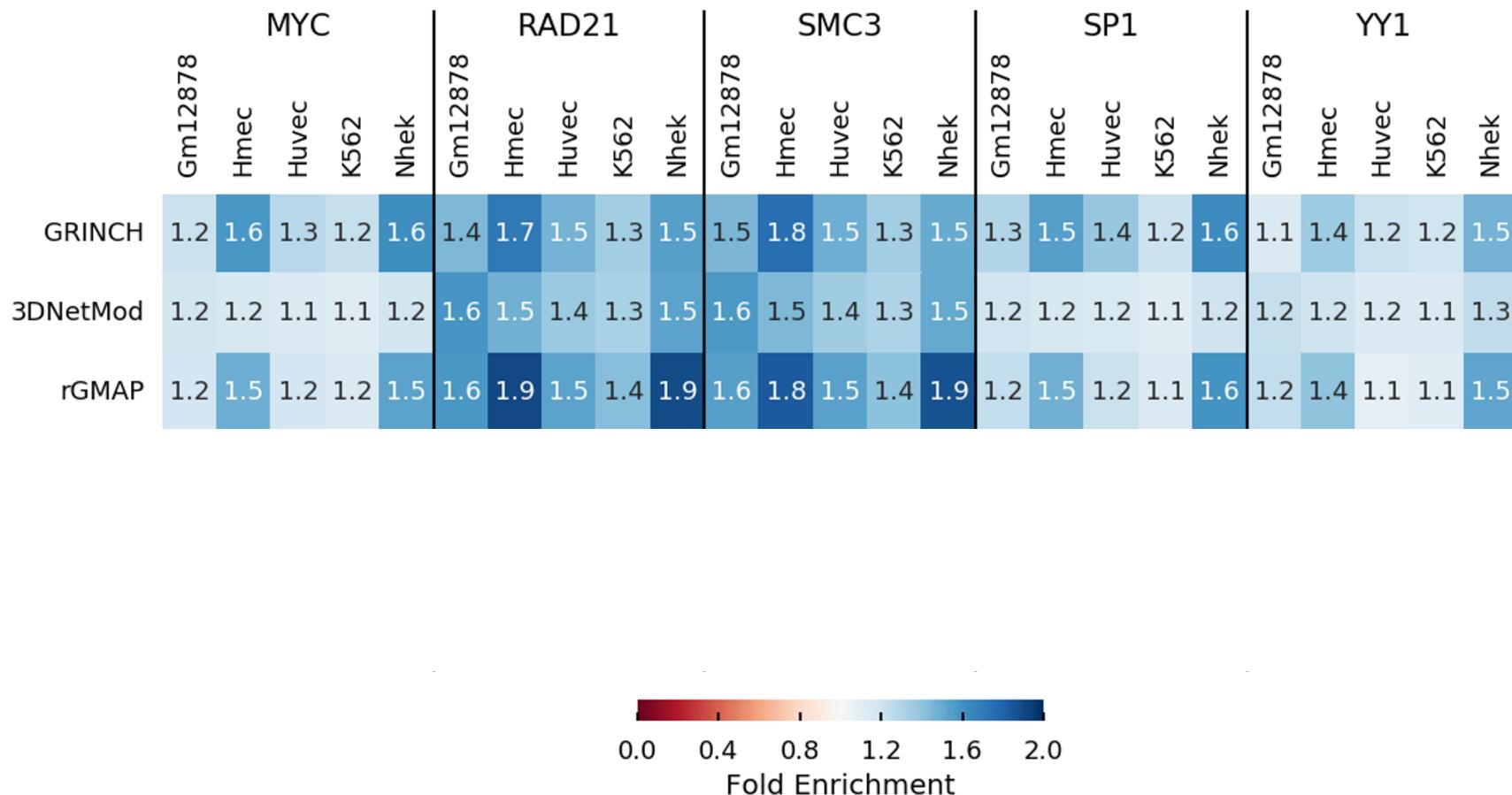
# Assessing cluster boundaries for architectural proteins



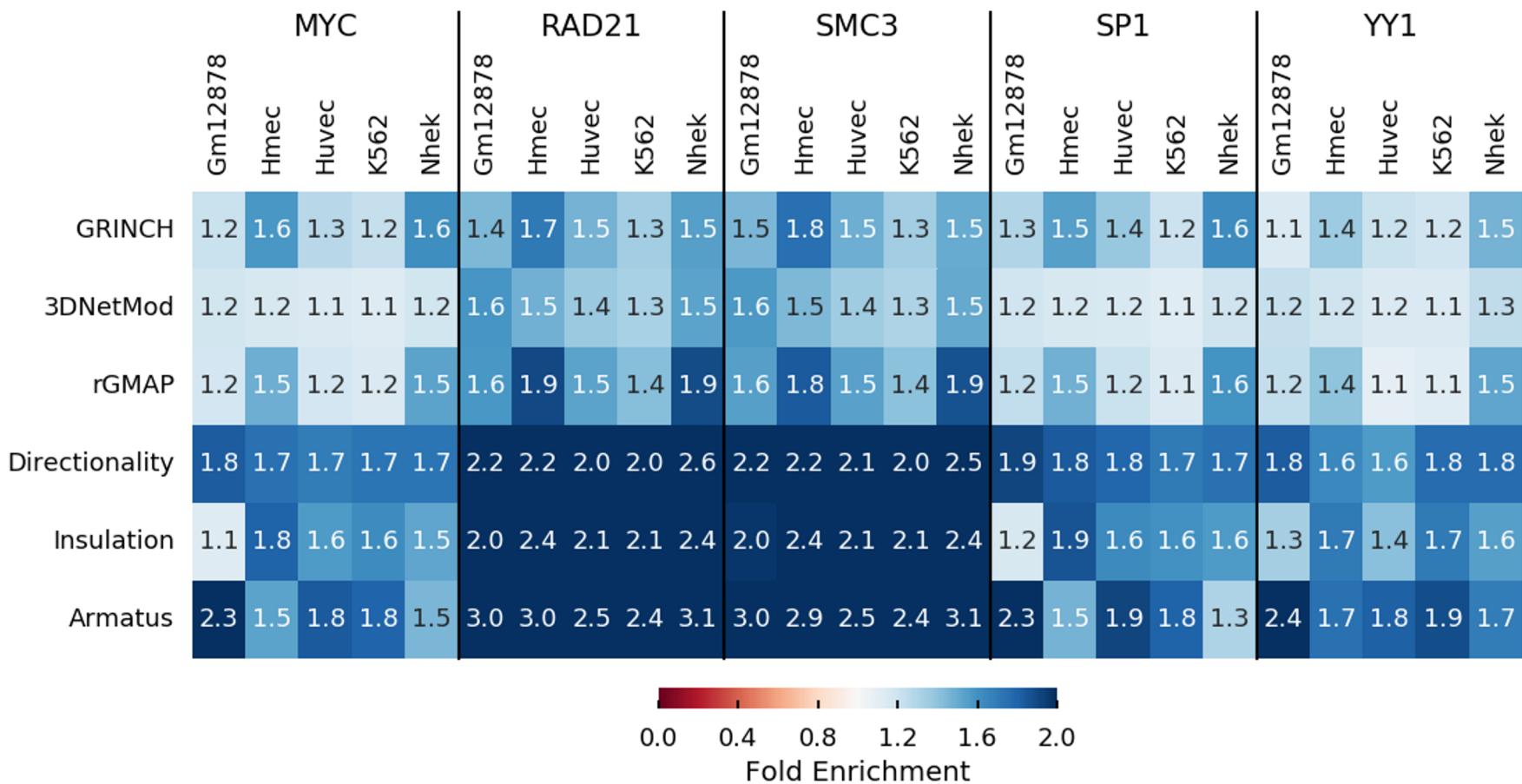
# GRINCH cluster boundaries are enriched for different architectural proteins



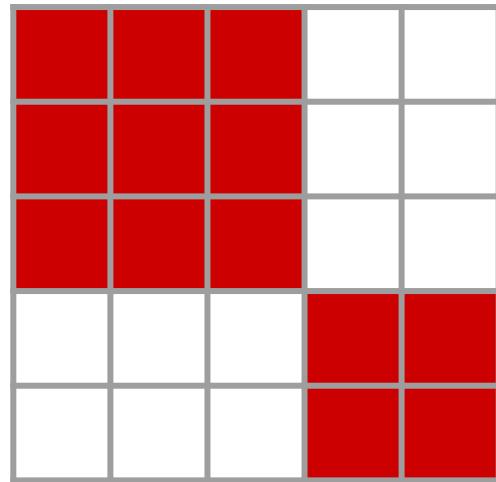
# GRINCH cluster boundaries are enriched for different architectural proteins



# GRINCH cluster boundaries are enriched for different architectural proteins

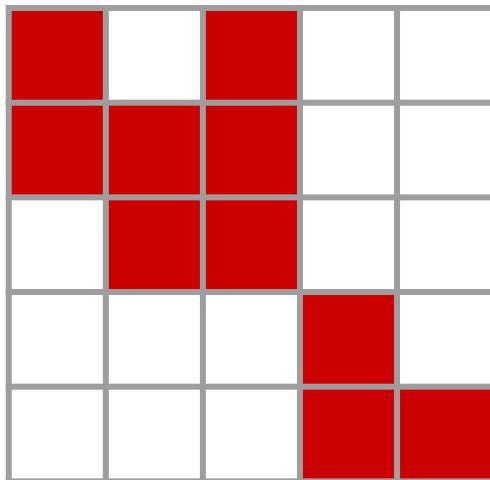
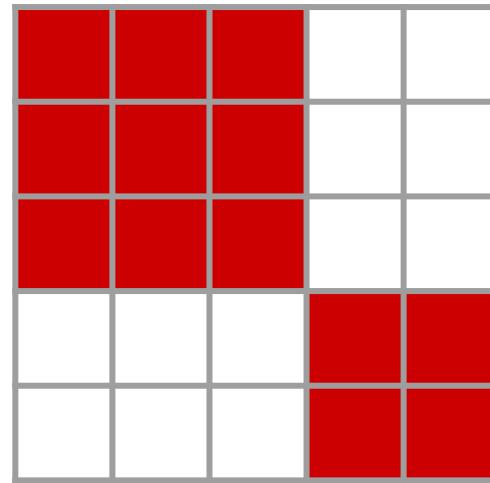
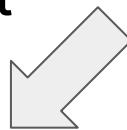


# Simulating sparsity to test stability



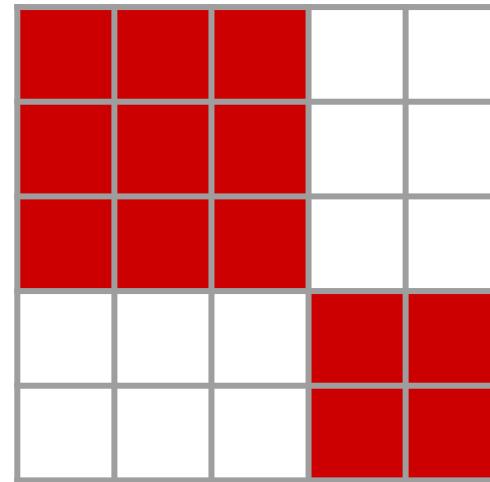
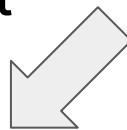
# Simulating sparsity to test stability

Dropout

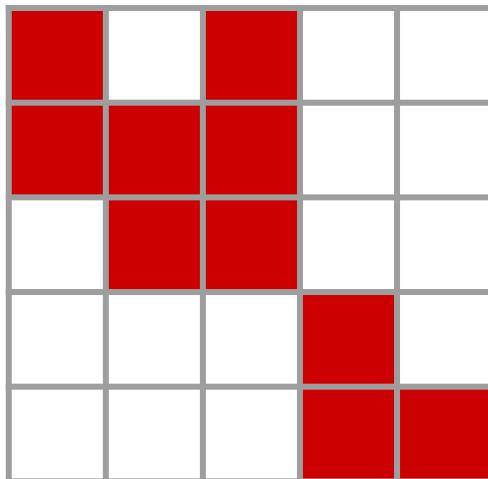
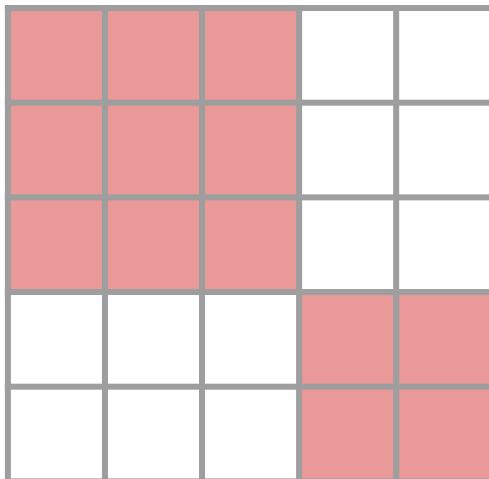


# Simulating sparsity to test stability

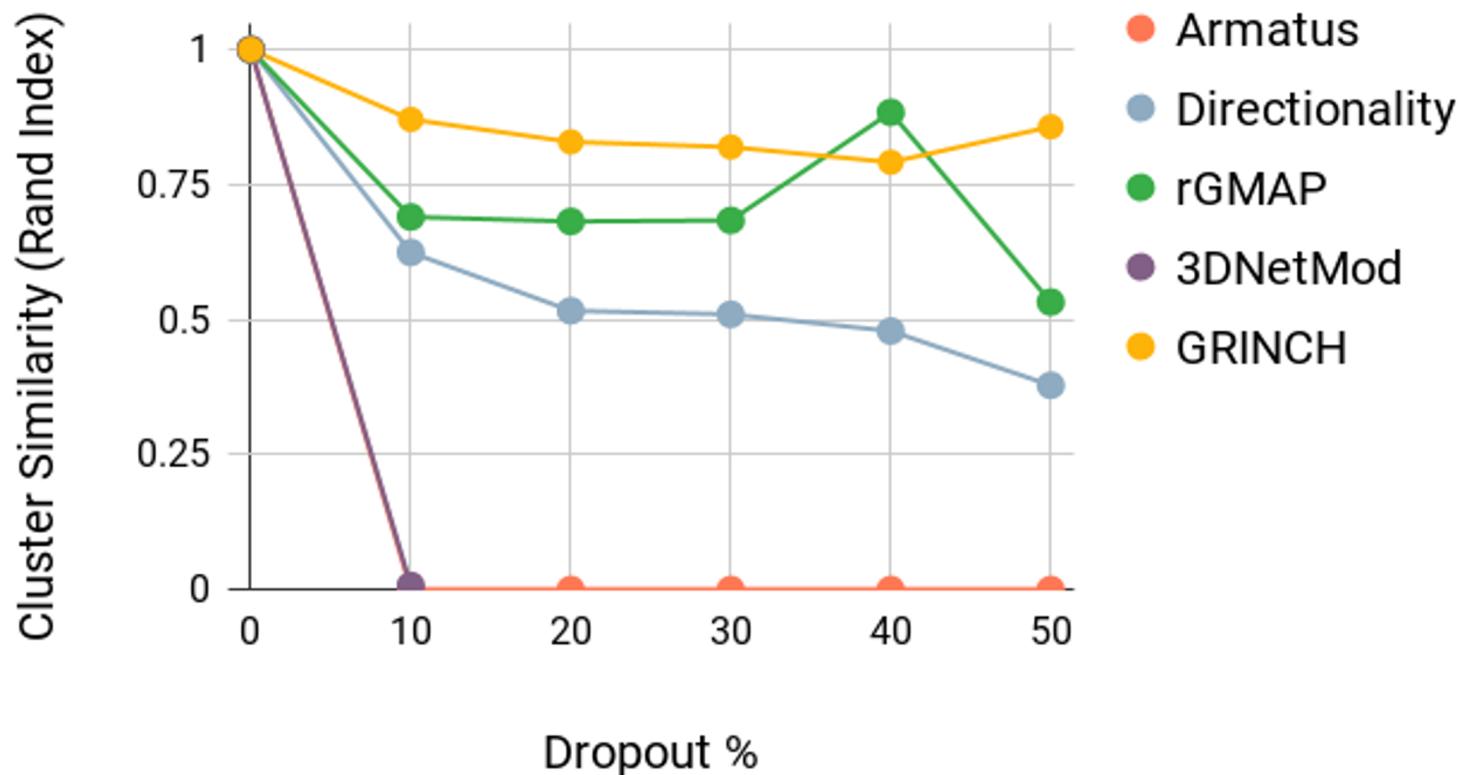
Dropout



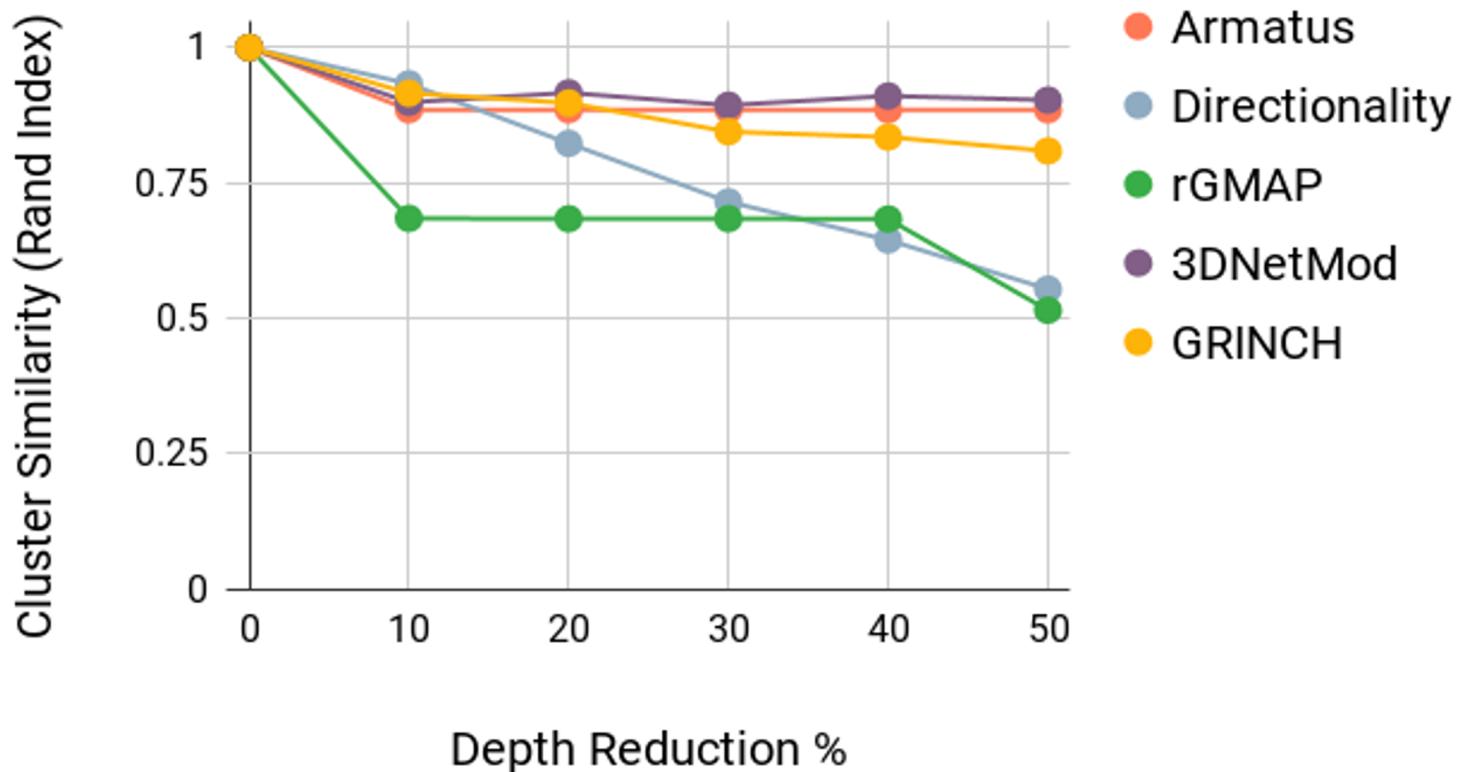
Depth  
Reduction



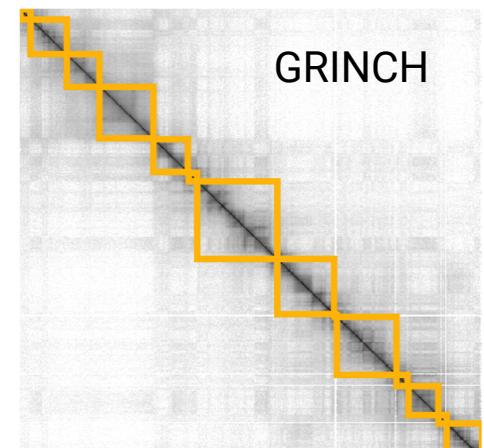
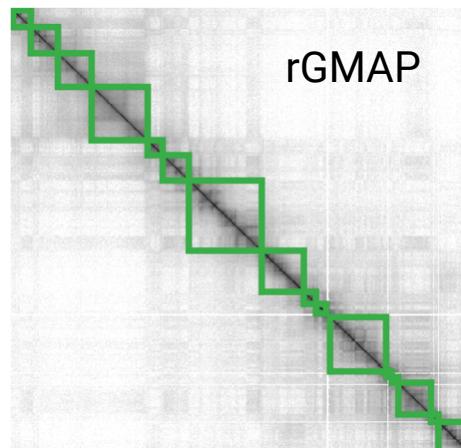
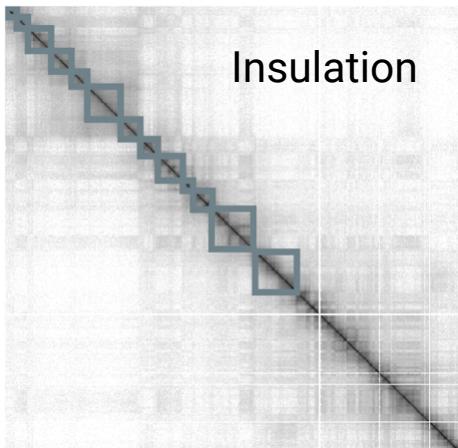
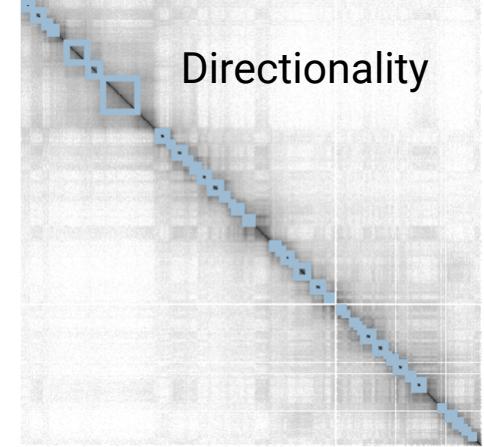
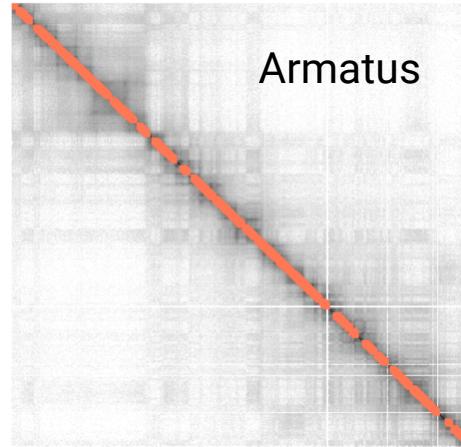
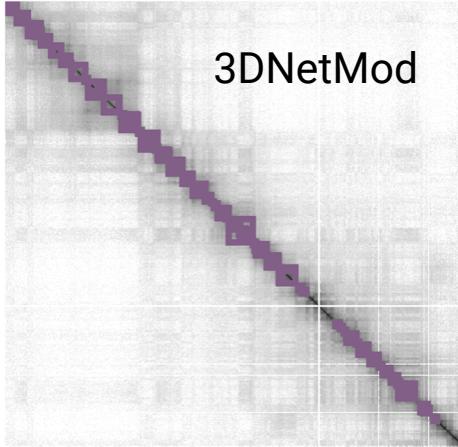
# GRINCH is the most stable method to dropout



# GRINCH is robust to lower-depth data

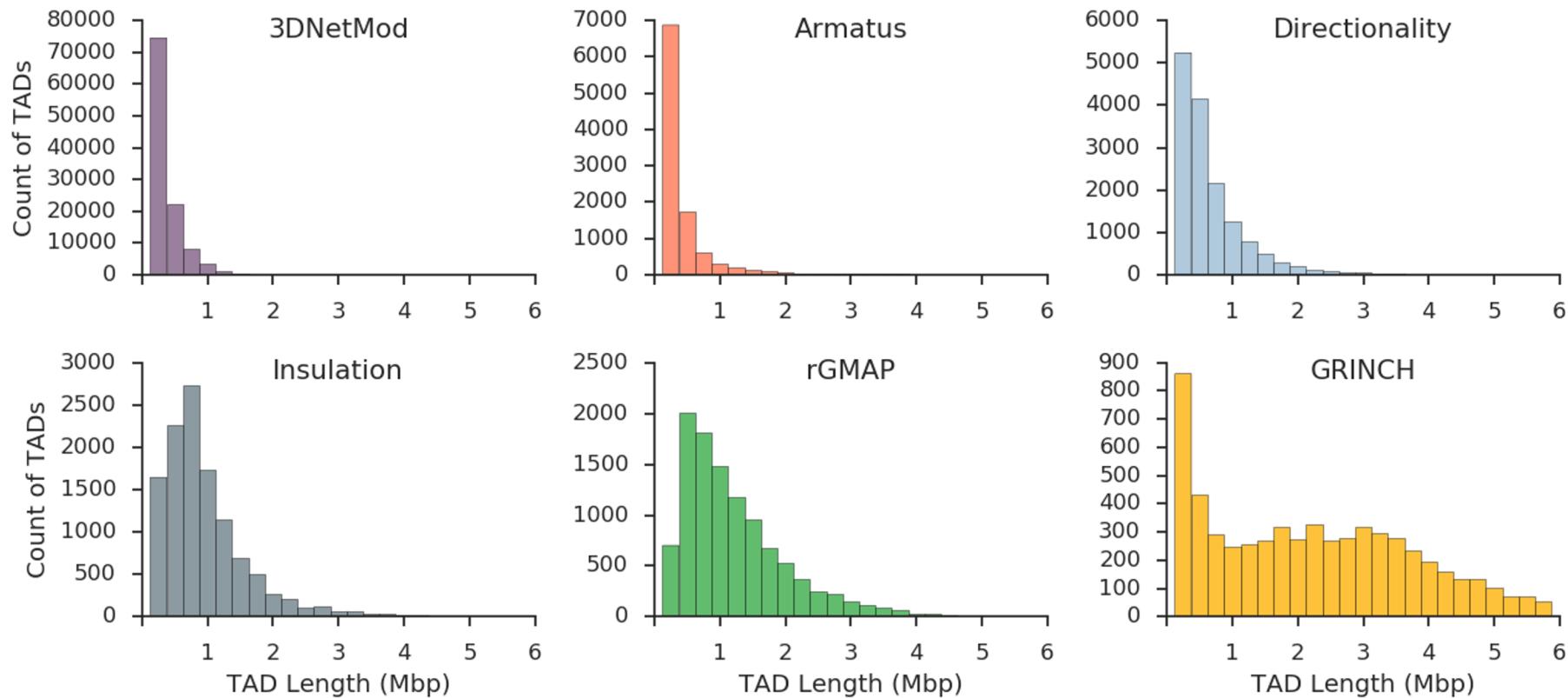


# GRINCH captures a wide range of domain sizes

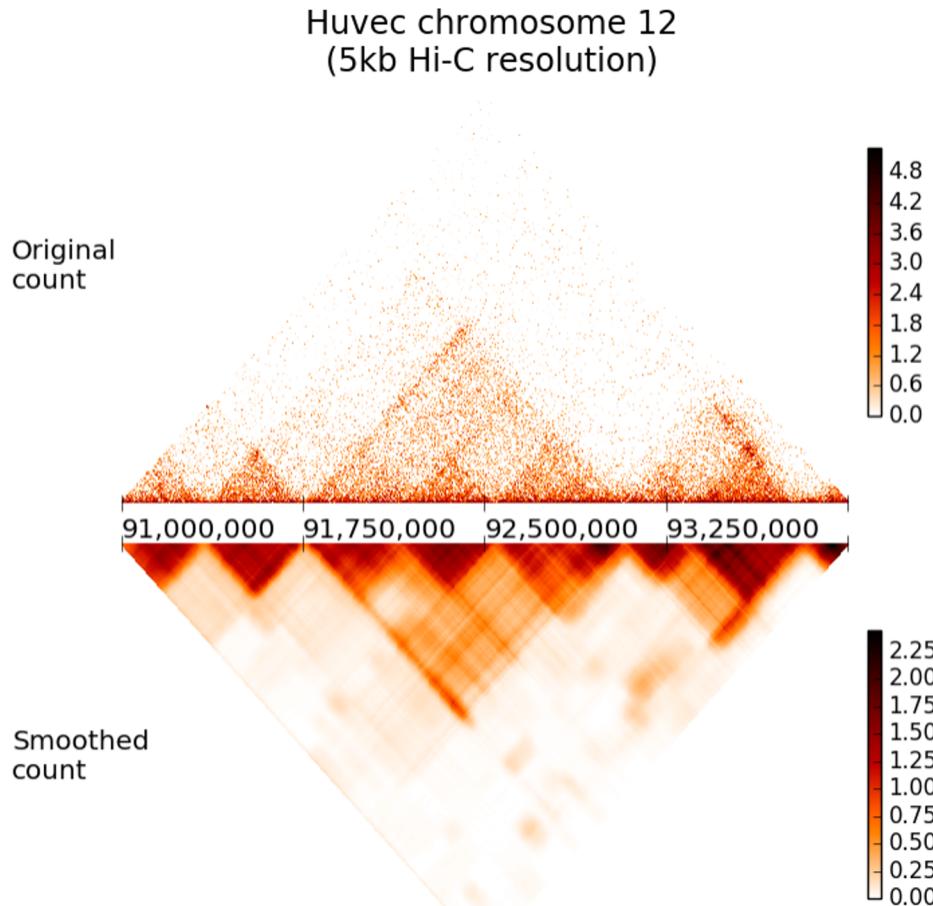


Gm12878 chr9 115,850,000-141,100,000bp

# 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)

