GRINCH: Discovering Structural Units of Chromosomes with Graph-Regularized Matrix Factorization

Da-Inn Lee[1] and Sushmita Roy[1,2]

[1]Department of Biostatistics and Medical Informatics - University of Wisconsin, Madison

[2]Wisconsin Institute of Discovery

School of Medicine and Public Health
UNIVERSITY OF WISCONSIN–MADISON

WISCONSIN INSTITUTE FOR DISCOVERY
WID.WISC.EDU

Introduction

- Long-range gene regulation is a critical factor in mammalian development and disease.
- The mechanism that brings together distal regulatory elements is governed by the three-dimensional organization of the genome.
- Chromosomes are organized into high-order domains at multiple scales: chromosomal territories, A/B compartments, topologically associating domains (TADs), chromatin loops, etc.
- Computational methods use data from high-throughput chromosomal conformation capture (Hi-C) assay to explore the organizational units of chromosomes.
- Recent studies comparing different TAD-finding methods found large variance in their replicability and stability, suggesting the need for more robust methods.
- Here we present GRINCH, a novel matrix-factorization-based approach for discovering topological units of chromosomes.
- GRINCH identifies clusters of regions that are stable to sparse input and are enriched for boundary factors like CTCF.

Non-negative matrix factorization (NMF) recovers underlying grouping structure in data

NMF, a key component of GRINCH, assumes there are latent groupings in the input data. For example, in the Netflix data challenge, you have a large user movie rating matrix. NMF decomposes the large matrix into two factors much smaller in dimension, revealing the underlying groupings of similar users and movies.

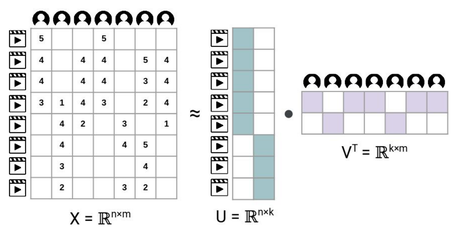

Fig 1. Using NMF, a large input matrix can be approximated by its two lower- dimensional factors that reveal the latent groupings of row or column entities.

The object of NMF is to minimize the difference between the original matrix X and its approximation from the dot product of the factors: $O = \|\mathrm{X} - \mathrm{UV}^T\|^2$

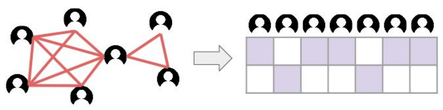

Fig 2. Prior knowledge in network format can be used in NMF by regularizing the factors with the graph Laplacian (L) which captures the network topology: $O = \|\mathrm{X} - \mathrm{UV}^T\|^2 + \lambda \mathrm{Tr}(\mathrm{V}^T\mathrm{LV}) + \lambda \mathrm{Tr}(\mathrm{U}^T\mathrm{LU})$

In many domains, there are known relationships between the column or row entities that could help recover the underlying groupings, for instance, the social network activities of Netflix users. A variation of NMF, graph-regularized NMF, can utilize this prior knowledge by regularizing or constraining each factor with the graph topology of the network whose information we wish to incorporate.

GRINCH: Graph-Regularized NMF and Clustering to analyze Hi-C data

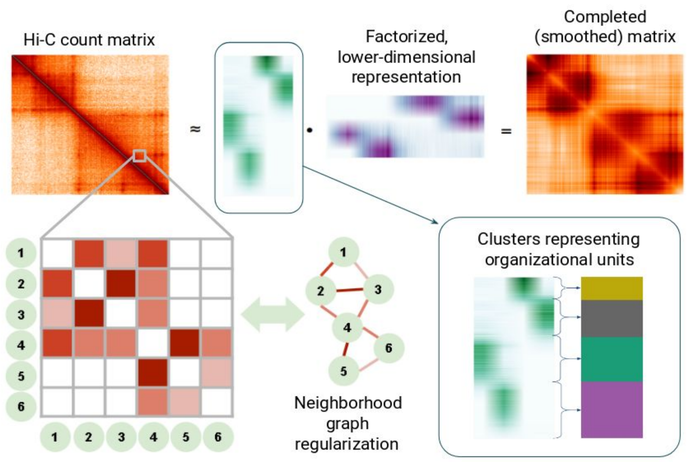

Fig 3. Overview of GRINCH

1. We apply NMF to a Hi-C matrix to recover latent groupings of genomic regions and to find the underlying structure captured by Hi-C data.
2. To exploit the distance dependency of interactions (i.e. nearby regions tend to interact more frequently), a neighborhood graph is used to influence the grouping of nearby regions through graph regularization.
3. From the neighborhood-regularized factors we recover the final clusters, representing organizational units of chromosomes.

GRINCH cluster boundaries are enriched in chromatin looping and remodeling elements

GRINCH cluster boundaries are significantly enriched in binding signals or accessible motif sites of proteins involved in chromatin looping (CTCF, RAD21, SMC3, YY1) and chromatin remodeling (MYC, SP1). The enrichment levels are comparable to other graph-based models like 3DNetMod.

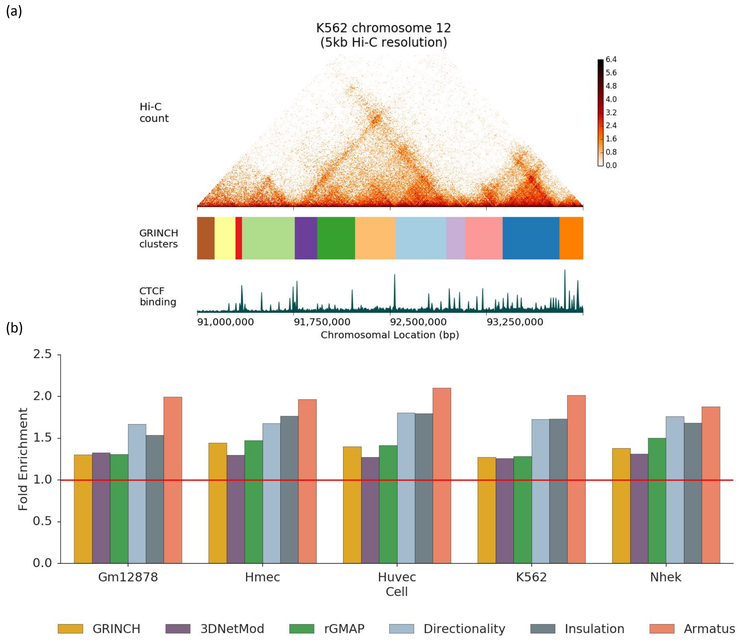

Fig 4. (a) GRINCH cluster boundaries align with ChIP-seq peaks of CTCF, a chromatin-loop-enforcing protein. (b) Levels above the red line are above-1-fold enrichment. GRINCH cluster boundaries have statistically significant levels of CTCF binding enrichment across 5 cell lines, comparable to other graph-based methods like 3DNetMod.

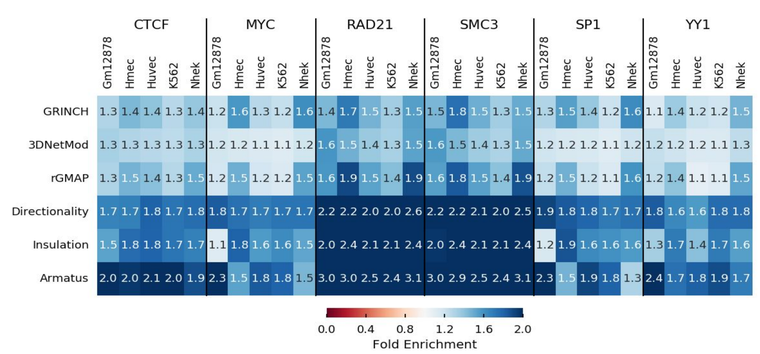

Fig 5. GRINCH cluster boundaries are significantly enriched in ChIP-seq binding peaks of CTCF and in accessible motif sites of MYC, RAD21, SMC3, SP1, and YY1 across 5 cell lines, levels comparable to other graph-based methods like 3DNetMod.

GRINCH is stable to sparse data

Since many Hi-C datasets are sparse and noisy, we evaluated the stability of each method by simulating sparser and lower-depth data. GRINCH is one of the most stable to both sparsity and low depth among methods compared.

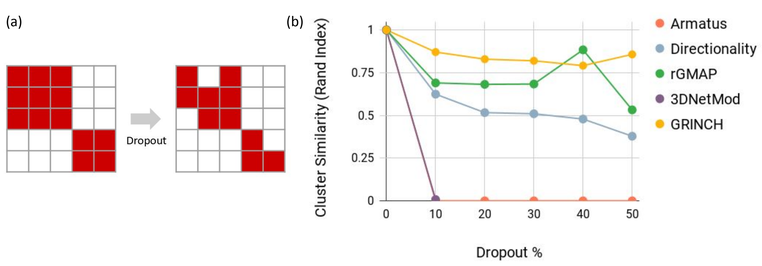

Fig 6. (a) Sparser data set was simulated by setting X% of original contact counts to 0 (dropout %). (b) GRINCH is the most stable to sparser datasets among methods compared.

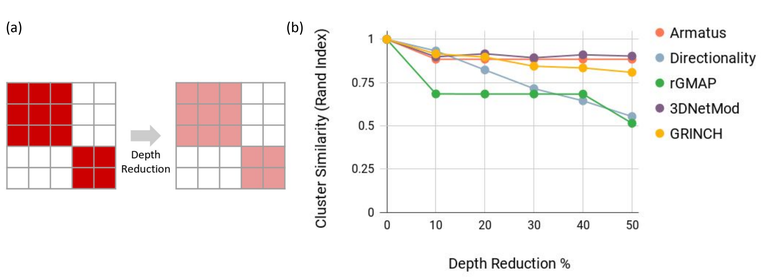

Fig 7. (a) Lower-depth data set was simulated by reducing the counts by X% across the original matrix (depth Reduction %). (b) GRINCH is the very stable to lower-depth data.

GRINCH captures large domains

GRINCH captures TADs of diverse lengths, in the mega-base-pair scale. It recovers TADs in the longer range than other methods.

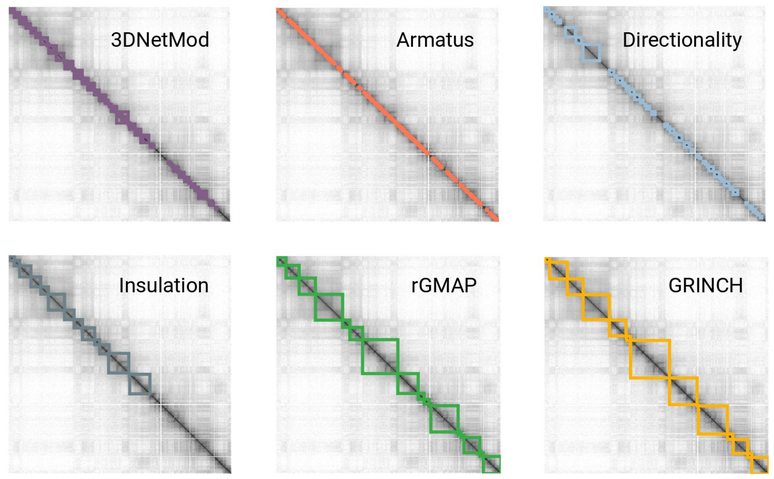

Fig 8. TADs from different methods for 115,850,000-141,100,000bp region of Gm12878 chromosome 9

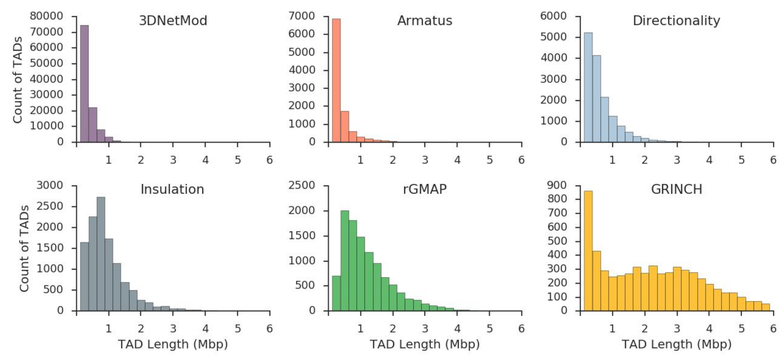

Fig 9. GRINCH recovers longer clusters than other methods, in the mega-base-pair scale.

GRINCH can smooth Hi-C data through matrix completion

Due to the sparse nature of Hi-C data, smoothing or denoising of Hi-C contact maps is an important preprocessing step in Hi-C analysis pipelines, allowing downstream replicability and concordance analysis (HiCRep, GenomeDISCO), joint normalization (HiCdiff), and resolution enhancement (HiCPlus). A useful byproduct of GRINCH is the smoothed Hi-C interaction matrix which can be obtained by multiplying back the NMF factors.

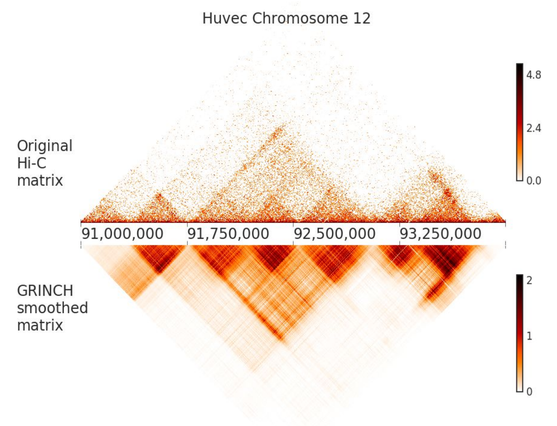

Fig 10. Smoothing of Hi-C matrices using GRINCH

Conclusion

- GRINCH is a method to discover structural units of the genome using non-negative matrix factorization with graph regularization.
- 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.

Acknowledgements

We acknowledge the support of NIH through NIH BD2K U54 AI117924 and NIH NIGMS 1R01GM117339.