

**Matrix factorization methods for examining 3D genome organization and
gene expression within and across species**

by
Da-Inn Lee

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Biomedical Data Science)

at the
UNIVERSITY OF WISCONSIN–MADISON
2024

Date of final oral examination: 12/03/2024

The dissertation is approved by the following members of the Final Oral Committee:

Sushmita Roy, Professor, Biostatistics and Medical Informatics

Colin Dewey, Professor, Biostatistics and Medical Informatics

Sunduz Keles, Professor, Biostatistics and Medical Informatics

Stephen Wright, Professor, Computer Sciences

© Copyright by Da-Inn Lee 2024

All Rights Reserved

Acknowledgments

I would like to thank my advisor, Sushmita, for, well, everything. She's the reason I was able to see myself doing research and the reason I didn't wash out of the program and the reason I made it this far. Now we just have to work on cloning her so she can do 8 people's worth of work instead of 4 (I am only half joking).

I also owe my general well-being to the former and current members of the Roy lab, where possibly the highest concentration of the most introverted people can be found (which means I feel right at home): Junha Shin, Deborah Chasman, Brittany Baur, Sara Knaack, Saptarshi Pyne, Suvojit Hazra, Chandrani Kumari, Alireza Fotuhi Siahpirani, Shilu Zhang, Matt Stone, Sunnie Grace McCalla, Khagani Eynullazada, Spencer Halberg-Spencer, Kristan Gimse, Marina Kotvanova, Prakriti Garg, Elias DeVoe, Yuda Liu, Swathisri Venkatesh, Harmon Bhasin, Jeremy Yang. Thank you for the scripts and programs you wrote, the data you have processed/the frustration you have saved me from, for checking in on me, supplying carbs, letting me borrow your chargers, helping me print posters, making me laugh, and listening to me.

I think any place is only as good as its people, and I definitely think our program is the best because of the faculty, administrators, and student leaders. I'd like to thank Dr. Colin Dewey for providing support and feedback on my research and for the time and effort he invested in revamping the bioinformatics courses of which I'm a direct beneficiary. I thank Dr. Sunduz Keles also for being part of my committee and always taking the time

during poster presentations to stop by, chat, and leave thoughtful and helpful feedback. I'm grateful to Dr. Stephen Wright for his feedback on my work, especially on the NMF regularization scheme back when I was still working out the math. I thank Dr. Irene Ong for letting me join her lab for a rotation, it was joy to work on a cool microbiome dataset. I'm grateful to Dr. Mark Craven, for providing an interesting and important rotation project on ML interpretation, and for being supportive in all CIBM-related activities including the NLM conferences. I'd like to thank Dr. Karl Broman for teaching us about "soft" yet important skills and ethics of being a scientist. I don't think it's an exaggeration to say I was able to get through the program thanks to our former and current coordinators, Beth Bierman and Shelley Maxtad - thank you for making sure I didn't slip through the crack and for everything you did/do for the students. I also had the honor and pleasure of being part of the Computation and Informatics in Biology and Medicine (CIBM) training program coordinated by Dr. Louise Pape - thank you for being the solid foundation of CIBM and always making me feel welcomed and safe. I'm very proud and impressed by the BDS student leaders including Spencer and Prakriti - thanks for all your hard work to create a community. It certainly takes a village to raise a PhD and I couldn't ask for a better village.

Speaking of a sense of community, I'd like to thank the UW-Madison Writing Center and the writing groups led by Chrissy Widmayer in 2021 and by Lisa Marvel Johnson in 2024. The writing group was the one of the few threads of human connection I had during the pandemic and also what got me through the last sprint of my dissertation.

I'd also like to thank our collaborators at the Hospital for Sick Children at Toronto - Michael Wilson, Liangxi (Dale) Wong, Mohamed Hawash, Huayun Hou - for generating and providing such cool and important cross-species data.

I'm grateful for the financial support from the National Institutes of Health (NIH) through the grant NIH NHGRI R01-HG010045-01 and by the Computation and Informatics

in Biology and Medicine (CIBM) training program (NLM 5T15LM007359), which allowed me to carry out a bulk of the work described here.

I owe my survival to my pandemic pod, Anna, Eric, and Avi. Thanks for feeding me and letting me be angry all the time. I look forward to founding our humble yet aggressive homestead out in the west.

Shoutout to Rachel, my eternal roommate - thanks for believing I can do whatever I set my mind to (even when I doubt it).

And my in-laws, Michele, Oscar, Elvia, Carlos, Gabby, thank you for being my family.

If it's possible to dedicate a dissertation to someone, I'd like to dedicate it to my mom, who taught me the importance of articulating myself clearly in Korean, English, in written, spoken, and visual form. Being born as her daughter was the luckiest break I had.

Last but not the least, my husband, Alejandro, and our cat, Raven. I'd like to thank my husband for being my rock and my roots and my shield and also just generally a good person. I thank Raven for keeping me company during my late-night writing and being my one-cat audience for practice talks. I sincerely hope getting a PhD means Raven will finally review my supplementary human status and consider promoting me to a co-primary.

Contents

Contents iv

List of Figures x

Abstract xiv

1 Introduction 1

1.1 3D genome organization and its dynamics across biological conditions and species 2

1.2 Single-cell transcriptomics and cross-species integration 4

1.3 Non-negative matrix factorization and its multi-task variants 6

1.4 Contributions and outline 8

2 GRiNCH: simultaneous smoothing and detection of topological units of genome organization from sparse chromatin contact count matrices with matrix factorization 11

2.1 Introduction 12

2.2 Results 15

2.2.1 GRiNCH, a non-negative matrix factorization-based method for analyzing high-throughput chromosome conformation capture datasets 15

2.2.2	GRiNCH TADs are high quality and stable to varying resolution and depth of input Hi-C data	17
2.2.3	GRiNCH TADs are enriched in architectural proteins and histone modification signals.	20
2.2.4	GRiNCH smoothing of low-depth datasets help recover structure and significant interactions.	21
2.2.5	GRiNCH application to chromosomal organization during development.	24
2.2.6	GRiNCH can be used for a variety of 3D conformation capture technologies	27
2.3	Methods	36
2.3.1	Graph-regularized Non-negative Matrix Factorization (NMF) and Clustering for Hi-C data (GRiNCH) framework	36
2.3.2	Datasets used in experiments and analysis	41
2.3.3	TAD-calling methods	43
2.3.4	TAD evaluation criteria	46
2.3.5	Identification of candidate genomic regions involved in 3D organization changes during mouse neural development	52
2.3.6	Identification of novel factor enrichment at GRiNCH TAD boundaries	52
2.3.7	Smoothing methods	53
2.3.8	Assessment of benefits from smoothing	54
2.4	Implementation and availability	55
2.5	Discussion	57
2.6	Conclusion	60
2.7	Acknowledgments	60

3	Examining dynamics of three-dimensional genome organization with multi-task matrix factorization	61
3.1	Introduction	62
3.2	Results	66
3.2.1	Tree-guided Integrated Factorization (TGIF) for examining dynamics in 3D genome organization	66
3.2.2	TGIF-DB identifies fewer false-positive differential boundaries in simulated and real Hi-C data.	67
3.2.3	TGIF-DC identifies compartment dynamics that are significantly enriched for differential regulatory signals.	71
3.2.4	TGIF-DC offers a unified framework to identify both compartment and subcompartment dynamics.	72
3.2.5	Changes in gene expression are associated with changes in boundaries during differentiation.	74
3.2.6	Persistent boundaries are enriched for SNPs from diverse disease phenotypes.	78
3.3	Methods	87
3.3.1	Tree-Guided Integrated Factorization (TGIF)	87
3.3.2	TGIF-DB for differential boundary identification	89
3.3.3	TGIF-DC for differential compartment and subcompartment identification	94
3.3.4	Estimating tree structure from input Hi-C matrices for unknown inter-dataset relationships	97
3.3.5	Datasets and preprocessing	98
3.3.6	Benchmarking methods for identifying differential domain boundaries	100
3.3.7	Comparison of TGIF-DC to existing compartment-calling methods .	107

3.3.8	Assessing differential gene expression near or within significantly differential boundaries and compartments	108
3.3.9	SNP enrichment within TGIF boundaries from cardiomyocyte differentiation data	110
3.4	Implementation and availability	110
3.5	Discussion	111
3.6	Acknowledgements	114
4	Analysis of evolutionarily conserved features in 3D genome organization with multi-task matrix factorization	116
4.1	Introduction	116
4.2	Methods	119
4.2.1	Single-window TGIF	119
4.2.2	Hi-C and CTCF ChIP-seq data from aortic endothelial cell of 5 mammalian species	119
4.2.3	Assessing across-species sequence similarity in the vicinity of human CTCF peaks	120
4.2.4	Application of TGIF to multi-species AEC data	121
4.2.5	Measuring structural similarity of Hi-C count matrices across species	122
4.3	Results	125
4.3.1	Sequence similarity between species rapidly decreases with distance from human CTCF peak.	125
4.3.2	Structural similarity across species in the neighborhood of human CTCF peak	125
4.3.3	Identification of conserved boundaries with swTGIF	126
4.4	Discussion	130

4.5	Acknowledgments	131
5	Multi-task matrix factorization leveraging gene orthology for cross-species integration of single-cell transcriptomics data	132
5.1	Introduction	132
5.2	Methods	135
5.2.1	Tree-guided integrated matrix factorization with branch-specific regularization (TIMBER)	135
5.2.2	From factors to clusters	141
5.2.3	From gene orthogroups to gene mapping matrices	142
5.2.4	Single cell RNAseq data from maize, sorghum, medicago	143
5.3	Preliminary results	144
5.4	Future direction	151
5.5	Acknowledgments	152
6	Concluding remarks	154
	Glossary	159
	Bibliography	163
A	GRiNCH supplementary materials	183
A.1	Supplementary figures	183
A.2	List of supplementary tables	204
A.3	Supplementary method	206
B	TGIF supplementary materials	207
B.1	Supplementary figures	207
B.2	List of supplementary tables	229

B.3 Supplementary methods231

List of Figures

2.1	Overview of GRiNCH.	16
2.2	Characterizing TADs with internal validation metrics, TAD size, and composition.	29
2.3	Evaluating the stability of different TAD-calling methods to datasets of different depths.	30
2.4	Summary of benchmarking TAD-calling methods.	31
2.5	Evaluating TAD-calling methods with enrichment of boundary elements and regulatory signals.	32
2.6	Evaluating the benefits of smoothing in GRiNCH.	33
2.7	GRiNCH applied to Hi-C datasets along developmental time courses.	34
2.8	Applying GRiNCH to datasets from different 3D genome conformation capture technologies.	35
3.1	Overview of TGIF	80
3.2	Benchmarking TGIF-DB	81
3.3	Benchmarking TGIF-DC on data from H1 and H1 differentiated to definitive endoderm	82
3.4	Characterizing compartments and subcompartments identified by TGIF-DC in mouse neural differentiation data.	83
3.5	Differential gene expression near or within differential structural features.	84

3.6	Human pluripotency-specific boundary elements.	85
3.7	SNP enrichment in persistent boundaries.	86
4.1	Single-window TGIF (swTGIF)	123
4.2	Calculating sequence overlap between uniform-sized bins	124
4.3	Sequence overlap between species around human CTCF peak	127
4.4	Structural similarity between species around human CTCF peak	128
4.5	Identification of conserved boundaries with swTGIF	129
5.1	Overview of TIMBER	136
5.2	Applying TIMBER on scRNAseq data from maize, sorghum, medicago	147
5.3	Effect of α on embeddings and clusters.	148
5.4	TIMBER cell cluster exploration	149
5.5	Marker gene enrichment within each TIMBER gene cluster ($k = 15, \alpha = 100$) . .	150
A.1	Selecting hyperparameters in GRiNCH.	184
A.2	Distribution of TAD lengths for different methods across resolutions.	185
A.3	Similarity of TADs across resolutions measured by mutual information	186
A.4	Similarity of TADs from different TAD-calling methods	188
A.5	Histone modification enrichment in TADs across different resolution.	189
A.6	Similarity of topology and significant interactions among Hi-C data using dif- ferent restriction enzymes and smoothed by GRiNCH.	190
A.7	Similarity of GRiNCH TADs from mouse neural development time-course Hi-C data	190
A.8	Interaction and regulatory profile near Arl6ip5 and Foxp1 during mouse neural development	191
A.9	Visual comparison of TADs around Zfp608 during mouse neural development	192

A.10 Visual comparison of TADs around Syap1 and Ap1s2 during mouse neural development	193
A.11 Visual comparison of TADs around Arl6ip5 and Foxp1 during mouse neural development	194
A.12 Similarity of GRiNCH TADs from pluripotency reprogramming time-course Hi-C data	196
A.13 Interaction and regulatory profile near Sox2 during mouse pluripotency reprogramming	197
A.14 Characterizing GRiNCH clusters of different size scales	198
A.15 Enrichment of regulatory signals in GRiNCH clusters of different size scales	199
A.16 GRiNCH TAD size distribution by regularization parameters	201
A.17 Memory consumption and runtime trend of GRiNCH algorithm	202
A.18 Comparison of NNDSVD initialization versus random initialization	203
B.1 Overview of TGIF-DB	208
B.2 Inferring a tree structure between input conditions based on the similarity of the input matrices. Here we use mouse chr19 intra-chromosomal matrices (25kb resolution) from neural differentiation data with 3 timepoints (ES, NPC, CN) as an example	209
B.3 Benchmarking TGIF-DB	210
B.4 Overview of datasets used in benchmarking and analysis	211
B.5 Characterizing sigDC in H1-endoderm differentiation	212
B.6 Characterizing TGIF-DC clusters and sigDC on mouse neural differentiation data	213
B.7 Heatmap visualization of pairwise difference in O/E counts	214
B.8 Post-hoc annotation of TGIF-DC clusters into A and B compartments in mouse neural differentiation data	215

B.9	Characterizing boundaries and sigDB in H1-endoderm differentiation	216
B.10	Characterizing boundaries and sigDB in mouse neural differentiation	217
B.11	Characterizing TGIF clusters and sigDC from applying TGIF-DC on cardiomyocyte differentiation data	218
B.12	Characterizing boundaries and sigDB in cardiomyocyte differentiation	219
B.13	Examples of differentially expression (DE) gene near significantly differential boundary (sigDB)	220
B.14	Persistent TGIF boundaries identified from human cardiomyocyte differentiation containing the cardiovascular disease associated GWAS SNP rs9349379	221
B.15	Hyperparameter α selection for TGIF-DB	222
B.16	Resource use by TGIF-DB	223
B.17	Hyperparameter α selection for TGIF-DC	224
B.18	Resource use by TGIF-DC	225
B.19	Post-hoc annotation of TGIF-DC clusters into A and B compartments in H1-endoderm dataset using accessibility	226
B.20	Post-hoc annotation of TGIF-DC clusters into A and B compartments in cardiomyocyte differentiation dataset using accessibility	227
B.21	TAD boundary perturbation procedure for Hi-C data simulation.	228

Abstract

Matrix factorization is a powerful and flexible dimensionality reduction method for high-dimensional and noisy genomic data. It finds factor matrices that preserve the low-dimensional structure of the input matrix. These factor embeddings can be used to co-cluster the row and the column entities, making matrix factorization a desirable approach to analyzing Hi-C data capturing 3D genome organization and scRNAseq data measuring gene expression at single-cell resolution. While Hi-C and scRNAseq technologies capture different aspects of gene regulation and cellular diversity, their datasets often suffer from noise and sparsity. The goal of studying 3D genome organization with Hi-C data is to identify structural units of 3D genome robust to noise and sparsity, and biologically significant differences in those units that are associated with dynamic processes such as development, disease progression, and evolution. Similarly, a key challenge in analyzing scRNAseq data is in overcoming batch effect and noise while preserving biological differences, a pronounced problem when analyzing data from multiple species to understand the evolution of cell types and gene expression patterns. To address these tasks, we developed three matrix-factorization methods, GRiNCH, TGIF, and TIMBER. GRiNCH factorizes a Hi-C matrix and uses neighborhood graph regularization to generate factors smooth in local neighborhoods. GRiNCH factors are used to recover structural units stable to noise in the input matrix and enriched in architectural proteins. TGIF uses multi-task learning to simultaneously factorize multiple Hi-C matrices. A tree is used to capture the

relationship among the input datasets, such as cell lineage and phylogeny. TGIF identifies differential structural units associated with changes in gene expression and chromatin states, as well as conserved structural elements across 5 mammalian species. TIMBER is also a tree-based multi-task factorization method, but it allows input matrices to have different dimensions. This in turn enables TIMBER to be applied to multi-species scRNAseq datasets with different number of genes. We show TIMBER is a promising approach to identifying homologous cell types across species. Together, GRiNCH, TGIF, and TIMBER provide a suite of computational tools for the discovery of shared and evolving patterns in 3D genome organization and gene expression.

Chapter 1

Introduction

Gene expression is governed by multiple layers of regulatory mechanisms: the interplay of architectural proteins, transcription factors, and various enzymes can organize the physical conformation of chromosomes inside the cell, bring regulatory sequence elements in contact with genes, and modify the chemical state of the chromatin to create an environment primed for gene expression (Schoenfelder and Fraser, 2019; Misteli and Finn, 2021; Hafner and Boettiger, 2023). Variations in these regulatory mechanisms lead to changes in gene expression underlying cellular diversity and heterogeneity observed during development, disease progression, and evolution (Stadhouders et al., 2018; Krijger and de Laat, 2016; Rowley et al., 2017). Advances in high-throughput sequencing technologies allow us to capture different aspects of such regulation and diversity. This dissertation primarily investigates 3D genome organization captured by Hi-C data, which quantifies pairwise interaction between genomic regions (Lieberman-Aiden et al., 2009; Rao et al., 2014; Kempfer and Pombo, 2020), and explores single-cell transcriptomics data which reveals heterogeneity of gene expression programs in different subpopulations of cells (Stuart and Satija, 2019; Adossa et al., 2021; Elmentaite et al., 2022; Heumos et al., 2023). As both genomic technologies generate high-dimensional and noisy datasets, we use non-negative matrix

factorization (NMF), a powerful and flexible dimension reduction method, to extract more interpretable embeddings and clusters from them (Lee and Seung, 1999, 2000; Stein-O’Brien et al., 2018). The flexibility of the NMF framework allows for regularization schemes and multi-task learning to simultaneously embed and cluster data from multiple cell types, biological conditions, and species, and enables the study of the dynamics and variations in the underlying biological process (Liu et al., 2013; Lee and Roy, 2024). This chapter further introduces the key goals in analyzing 3D genome and single-cell transcriptomics data, and provides an overview of NMF, its extensions, and its applications in these data domains.

1.1 3D genome organization and its dynamics across biological conditions and species

3D genome organization refers to how the DNA folds itself inside of the cell’s nucleus. A class of high-throughput sequencing technology called Hi-C enables the study of 3D genome organization by measuring the frequency of contact or interaction between pairs of genomic regions inside the 3D space of the nucleus (Lieberman-Aiden et al., 2009; Rao et al., 2014; Kempfer and Pombo, 2020). Analysis of Hi-C data has revealed the hierarchical organization of the 3D genome (Bonev and Cavalli, 2016; Rowley and Corces, 2018; Szabo et al., 2019): (1) chromosomal territories inside the nucleus preferentially occupied by specific chromosomes; (2) compartments segmenting the genome into active or repressive stretches millions of basepairs in size, each with distinct epigenetic and regulatory signatures; (3) topologically associating domains or TADs, the localized topological feature representing highly interacting neighboring regions of the genome; and (4) chromatin loops facilitating close contact between a pair of loci.

Detecting changes to the higher-order organization of the 3D genome is of particular

interest to understanding dynamic cellular processes underlying development, disease, and evolution. Rewiring of organizational units like TADs and compartments accompany specific stages of differentiation (Krijger et al., 2016; Hug and Vaquerizas, 2018; Stadhouders et al., 2018; Zheng and Xie, 2019; Boltsis et al., 2021). Causal structural variants behind various developmental diseases and cancer disrupt TAD boundaries and result in abnormal enhancer-promoter interactions (Lupiáñez et al., 2016; Hnisz et al., 2016; Norton and Phillips-Cremins, 2017; Chakraborty and Ay, 2019; Melo et al., 2020). TADs are considered the units of synteny across species, the conserved pockets of local interactions and regulations that travel together in the large-scale genome rearrangements during evolution (Rowley et al., 2017; McArthur and Capra, 2021; Liao et al., 2021; Álvarez González et al., 2022). An expanding catalog of publicly available Hi-C datasets covering developmental time courses, disease states, and multiples species provide an unprecedented opportunity to study the changes in these structural units and to link their dynamics with cellular function and phenotypic consequences (Dekker et al., 2017; Reiff et al., 2022).

Unfortunately, low depth and the sparsity of interaction counts in such datasets often hinder the efforts to identifying structural units like TADs in a reliable manner (Forcato et al., 2017; Zufferey et al., 2018; Xiong and Ma, 2019). While there are many TAD-calling methods available, they produce a widely different set of TADs depending on the depth, sparsity, resolution of the input data (Forcato et al., 2017; Zufferey et al., 2018; Lee and Roy, 2021). On the other hand, there remains a dearth of computational methods for identifying *changes* in TADs or compartments across biological contexts; here the key challenge is in distinguishing the underlying biological differences from noise and technical artifacts. Therefore, we need a suite of tools that can identify (1) reproducible TADs, and (2) significantly differential structural units and features like compartments and TAD boundaries.

Cross-species analysis of Hi-C data in particular could provide insight about the rela-

tionship between sequence conservation and conservation of regulatory function. High degree of sequence conservation of an enhancer, for instance, does not guarantee nor is necessary for the conservation of its enhancer function in another species (Zemke et al., 2023; Snetkova et al., 2021). Therefore, to understand the evolution of regulatory elements, it is necessary to assess both the degree of sequence conservation as well as functional conservation across species. To carry about such analysis with Hi-C data first requires sequence alignment and mapping uniform-sized genomic regions or bins (typically 5000-50000 basepairs in size) across species, in order to find correspondence in 3D interactions among uniform-sized bins as measured by Hi-C (Yang et al., 2019; Eres et al., 2019; Zhang et al., 2019; Luo et al., 2021; Zemke et al., 2023; Keough et al., 2023). Once such mapping is completed and region-level homology identified, "homologous interactions" between pairs of homologous genomic regions is probed for the conservation or divergence of their regulatory relationship (e.g. an enhancer-promoter interaction; Yang et al., 2019; Eres et al., 2019). The regulatory role that homologous loci play in their respective species can also be assessed (e.g., the binding sites of a architectural protein to form TAD boundaries), along with the degree to which such role is species-specific or conserved (Zhang et al., 2019; Luo et al., 2021; Zemke et al., 2023; Keough et al., 2023). A systematic pipeline to carry out such analysis could be used to assess putative elements and proteins for their role in 3D genome organization across multiple species.

1.2 Single-cell transcriptomics and cross-species integration

Advances in single cell technologies have provided an unprecedented opportunity to analyze and understand complex systems by providing readouts of transcription, chromatin

accessibility, methylation, etc., at single-cell resolution (Stuart and Satija, 2019; Adossa et al., 2021; Elmentaite et al., 2022; Heumos et al., 2023). While bulk omics data can only provide the average measurement from an entire population of cells, single-cell omics can capture the diversity of transcriptomic and other epigenetic programs across a heterogeneous mix of cell types and states. Like Hi-C data, single-cell transcriptomics data (to be comprehensively referred to as scRNAseq from here on) is high-dimensional, typically with tens of thousands of genes and up to hundreds of thousands of cells. It also tends to suffer from sparsity and noise; a particular type of technical noise called batch effect is often observed in scRNAseq data as a large-scale shift in gene expression values due to experimental artifacts (Kiselev et al., 2019). Batch effect leads to spurious differences in signal from two biologically similar groups of cells (e.g., cells at a particular state of differentiation) from two different samples, experiments, or labs. Minimizing batch effect and integrating datasets across multiple samples, timepoints, investigators, and locations, while retaining signals from true biological differences is one of the key challenges in scRNAseq data analysis.

Batch correction or multi-condition integration methods are primarily used to integrate scRNAseq datasets collected from different species (Song et al., 2023). Understanding cell type homology and conservation or divergence of gene expression and regulatory programs across evolutionary timescale requires integrating scRNAseq data across species. Unsupervised cross-species integration can be challenging due to the fact that between-species difference can be larger than batch effect, while over-correction can diminish signals of cellular diversity within each species (Luecken et al., 2022; Song et al., 2023). Leveraging the phylogenetic relationship among the species that each dataset represents in order to inform the integration (and possibly mitigate over-integration) is a promising direction, although there is a scarcity in such computational approaches. SAMap, the sole method utilizing evolutionary information (by building a gene homology graph from BLAST alignment; Tarashansky et al., 2021) among a cross-species benchmarking study, was found

to excel in aligning homologous cell types across species, especially when the evolutionary distance among them is large (Song et al., 2023). There is much room to explore different ways of encoding phylogeny and homology for integration of multi-species scRNAseq data.

1.3 Non-negative matrix factorization and its multi-task variants

Non-negative matrix factorization (NMF) is a powerful dimension reduction method for noisy, high-dimensional data (Lee and Seung, 1999, 2000). Its objective is to recover two lower-dimensional factor matrices that can well approximate the original input matrix when multiplied together. The two factor matrices represent the embedding of the column entities and the row entities in the input, and can be used as latent features for clustering or further analysis. In a specialized case, the input matrix can be a representation of a network, where the rows and the columns are nodes, and the entry values the strength of connection between the nodes. Factorizing such matrix yields graph node embedding (Kuang et al., 2012; Jannesari et al., 2024; Hajiveisheh et al., 2024). The product of the factor matrices can also serve as a imputed or smoothed version of the input matrix (Cai et al., 2016; Nguyen et al., 2019). In the biomedical domain, NMF is often applied when the simultaneous embedding and clustering of a larger number of row entities (e.g. cells) and column entities (e.g. genes) is desirable (Stein-O’Brien et al., 2018; Brunet et al., 2004; Devarajan, 2008; Hamamoto et al., 2022).

NMF is a flexible framework that allows incorporation of further constraints or prior knowledge. One way to enforce a certain desirable quality in the output factors or the embedding is to impose regularization. For instance, if there is orthogonal information

available about the relationship between the row entities or the column entities that can be encoded as a network or a graph, the optimization objective can penalize the output factor loadings from deviating from the graph structure such that the final factor loadings or embeddings are more similar if the nodes they represent are more strongly connected in the graph (Cai et al., 2011). Such graph-based regularization scheme has been used to co-cluster cancer patients and oncogenes (Zhu et al., 2017), microRNAs with associated diseases (Xiao et al., 2018), and drug molecules to their targets (Zhang and Xie, 2022).

An extension to the regularization scheme facilitates multi-task learning. In cases where there are multiple related biological conditions, informing the embedding within each condition (or task) with those from closely related tasks can be beneficial for two key reasons: (1) the sparsity and the noise within the individual datasets can be better handled with information from other related datasets and (2) the resulting embedding in the shared latent space can make direct comparison across conditions easier.

Several multi-task NMF approaches have been used effectively in the scRNAseq data integration task. LIGER employs integrated NMF (iNMF) where the gene factor loadings for each input dataset consist of: a shared set of latent features across all inputs being integrated, and a “task-specific” set of latent features that captures the remaining variations in each input dataset (Welch et al., 2019; Liu et al., 2020). It has proven to be a very effective in removing batch effect across species, samples, and experimental techniques, but sometimes to a point of over-integrating (Luecken et al., 2022). UINMF is a variation of iNMF where each dataset keeps a subset of features (e.g., genes or accessibility peaks) to itself, and the subset does not contribute to learning the shared feature space (Kriebel and Welch, 2022). A benchmarking study found UINMF to be one of the most effective integration methods for multi-omic data with two or more modalities (Hu et al., 2024). Another approach, CoupledNMF is specifically designed to integrate single-cell transcriptomic (scRNAseq) and accessibility (scATACseq) data and incorporates both graph and multi-task regularization.

It forces the gene factor loading (for scRNAseq data) and the region-level factor load (for scATACseq data) to be smooth to a gene-region graph (Duren et al., 2018).

However, few attempts have been made to explicitly encode and exploit prior knowledge about the relationship among the input datasets (Hi-C or scRNAseq) in a multi-task NMF approach. Related Hi-C or scRNAseq datasets can be organized into a hierarchical relationship or be described as a member of a tree. For example, Hi-C can be measured from different cell types spawning from specific cell lineage during development; scRNAseq datasets can come from different perturbation experiments, each with its own nested set of other experimental conditions or designs; and of course, Hi-C and scRNAseq datasets from multiple species sharing ancestral species. Leveraging phylogenetic relationship in particular has been useful in identifying gene modules and inferring gene regulatory networks across multiple yeast, fungi, and plant species (Roy et al., 2013; Koch et al., 2017; Shin et al., 2021), identifying compartment structures simultaneously for human and mouse (Fotuhi Siahpirani et al., 2016), and finding groups of conserved or species-specific interactions among apes (Yang et al., 2019). Exploring tree-based regularization for multi-task learning is a promising approach to flexibly encode and leverage any known relationship among input datasets both for dynamic 3D genome analysis and for single-cell data integration.

1.4 Contributions and outline

The overarching goal of this work is to develop matrix factorization methods to understand context-specific and persistent patterns in 3D genome organization and gene expression across different biological conditions and species. To this end, we first developed Graph-Regularized NMF and Clustering for Hi-C (GRiNCH) to identify TAD-like clusters of genomic regions from Hi-C data in a manner stable and robust to various technical noises,

and to impute missing count values through matrix completion. We next expanded to a multi-task matrix factorization framework called Tree-Guided Integrated Factorization (TGIF), which can encode the relationship among multiple input Hi-C datasets as a tree, to simultaneously learn their embeddings and identify shared and differential structural features like TAD boundaries and compartments. We then applied a modified version of TGIF to analyze Hi-C data from multiple mammalian species and quantitatively validate the degree of conservation of a known boundary element. Finally, we extend the underlying NMF mechanism of TGIF to derive TIMBER, or Tree-guided Integrated Matrix Factorization with Branch-specific Regularization. TIMBER has the additional capacity to handle input datasets with different set or number of features, e.g., genes in single-cell gene expression matrices from multiple species. TIMBER was applied to single cell gene expression datasets from three different plant species with different nitrogen fixing strategies. The following is the outline of the subsequent chapters:

- In **Chapter 2**, we describe GRiNCH, a graph-regularized NMF approach for simultaneous smoothing and TAD identification in sparse and noisy Hi-C count matrices. GRiNCH outperforms seven existing TAD calling methods and 3 smoothing methods. It can be used on other 3D genome capture technologies (including SPRITE and HiCHIP) and to identify putative boundary factors that play context-specific roles.
- In **Chapter 3**, we present TGIF, a multi-task NMF framework for identifying differential structural units across multiple input Hi-C matrices. It streamlines differential analysis across multiple structural scales, i.e., boundary-, subcompartment-, and compartment-level changes. Application of TGIF to multiple developmental time-course datasets shows the association of significantly differential TAD boundaries and compartments to differential gene expression and changes in regulatory signals. Persistent TGIF boundaries are shown to be enriched in disease-associated sequence

variants.

- In **Chapter 4**, we apply a simplified version of TGIF to Hi-C matrices from 5 different mammalian species (human, rat, cow, pig, and dog) centered at human CTCF peaks and their mapped regions in other target species. We find that while sequence similarity between species rapidly decays away from the CTCF-peak-containing region, structural similarity based on the interaction patterns with neighboring regions are higher if a CTCF peak is also found in the target species. The simplified TGIF enables direct quantification and comparison of the presence of TAD boundaries across species.
- In **Chapter 5**, we describe TIMBER, another multi-task NMF framework which can specifically handle different sets of features (i.e. genes) across input scRNAseq datasets from multiple species. It encodes the relationship among the input datasets as a tree based on their phylogenetic relationship, and maps the different features across species using gene orthology information. We apply TIMBER to scRNAseq datasets from maize, sorghum, and medicago, with the goal of identifying sorghum- or species-specific mechanism behind nitrogen fixation from aerial roots.
- In **Chapter 6**, we summarize the key findings from each of the computational approaches and its application to 3D genome or single-cell gene expression data.

Chapter 2

GRiNCH: simultaneous smoothing and detection of topological units of genome organization from sparse chromatin contact count matrices with matrix factorization

High-throughput chromosome conformation capture assays, such as Hi-C, have shown that the genome is organized into organizational units such as topologically associating domains (TADs), which can impact gene regulatory processes. The sparsity of Hi-C matrices poses a challenge for reliable detection of these units. We present GRiNCH, a constrained matrix-factorization-based approach for simultaneous smoothing and discovery of TADs from sparse contact count matrices. GRiNCH shows superior performance against seven TAD-calling methods and three smoothing methods. GRiNCH is applicable to multiple platforms including SPRITE and HiChIP and can predict novel boundary factors with potential roles

in genome organization.

This work has been published in:

Lee DI and Roy S. 2021. GRiNCH: simultaneous smoothing and detection of topological units of genome organization from sparse chromatin contact count matrices with matrix factorization. *Genome Biology* **22**: 164

2.1 Introduction

The three-dimensional (3D) organization of the genome has emerged as an important layer of gene regulation in developmental processes, disease progression, and evolution (Bonev and Cavalli, 2016; Hug and Vaquerizas, 2018; Rowley et al., 2017; Krijger and de Laat, 2016; Szabo et al., 2019; Kempfer and Pombo, 2020). High-throughput chromosome conformation capture (3C) assays such as Hi-C (Lieberman-Aiden et al., 2009; Rowley and Corces, 2018), SPRITE (Quinodoz et al., 2018), and GAM (Kempfer and Pombo, 2020) provide a comprehensive view of 3D organization by measuring interactions among chromosomal regions on a genome-wide scale. High-throughput 3C data captured from diverse biological contexts and processes has led to an improved understanding of DNA packaging in the nucleus, the dynamics of 3D conformation across developmental stages (Zheng and Xie, 2019), and between normal and disease cellular states (Krijger and de Laat, 2016; Chakraborty and Ay, 2019). Analysis of such datasets has shown that chromosomal regions preferentially interact with one another, giving rise to higher-order structural units such as chromosomal territories, compartments, and topologically associating domains (TADs) which differ in the size of the structural unit and molecular features associated with the constituent regions. Although the relationship between TADs and changes in gene expression is debated (Kim et al., 2015; Ghavi-Helm et al., 2019; van Steensel and Furlong, 2019), these units have been shown to be conserved across species (Szabo et al., 2019; Eres

et al., 2019) and also associated with developmental (Stadhouders et al., 2018) and disease processes (Flavahan et al., 2016; Kleinjan and Lettice, 2008; Chakraborty and Ay, 2019; Valton and Dekker, 2016). Therefore, accurate identification of TADs is an important goal for linking 3D genome organization to cellular function.

Recently a large number of methods have been developed to identify TADs, utilizing different computational frameworks, such as dynamic programming (Filippova et al., 2014; Weinreb and Raphael, 2015), community and subgraph detection within networks (Filippova et al., 2014; Norton et al., 2018), Gaussian mixture modeling (Dixon et al., 2012; Yu et al., 2017), and signal processing approaches (Crane et al., 2015). However, comparison of TAD-finding methods (Forcato et al., 2017; Dali and Blanchette, 2017; Zufferey et al., 2018) have found large variability in the definition of TADs and high sensitivity to the resolution (size of the genomic region), sequencing depth, and sparsity of the input data. A lack of a clear definition for a TAD leads to difficulty in downstream interpretation of these structures (de Wit, 2019). To address the sparsity of datasets, different smoothing based approaches have been proposed (Yang et al., 2017; Ursu et al., 2018; Liu and Wang, 2019), however it is unclear whether and to what extent TAD identification or identification of significant loops can benefit from pre-smoothing the matrices.

Here, we present Graph Regularized Non-negative matrix factorization and Clustering for Hi-C (GRiNCH), a novel matrix-factorization-based method for the analysis of high-throughput 3C datasets. GRiNCH is based on non-negative matrix factorization (NMF), a powerful dimensionality reduction method used to recover interpretable low-dimensional structure from high-dimensional datasets (Lee and Seung, 2000; Wu et al., 2018; Soor et al., 2018). However, a standard application of NMF is not sufficient because of the strong distance dependence found in Hi-C data, that is, regions that are close to each other on the linear genome tend to have more interactions. We employ a graph regularized NMF approach, where the graph captures the distance dependence of contact counts such that

the learned lower-dimensional representation is smooth over the graph structure (Cai et al., 2011). Furthermore, by exploiting NMF's matrix completion property, which imputes missing entries of a matrix from the product of the low-dimensional factors, GRiNCH can smooth a sparse input matrix.

We perform a comprehensive comparison of GRiNCH and existing TAD-finding methods using a number of metrics: similarity of interaction profiles of regions belonging to the same TAD, stability to different resolutions and depth of input data, and enrichment of architectural proteins and histone modification known to facilitate or correlate with 3D genome organization. Despite the general trend of trade-off in performance among different criteria, e.g., a high performing method based on enrichment of architectural proteins is not as stable to resolution and depth, GRiNCH consistently ranks among the top across different measures. Furthermore, compared to existing smoothing approaches, GRiNCH-based smoothing of downsampled data leads to the recovery of TADs and significant interactions best in agreement with those from the original high-depth dataset. We apply GRiNCH to Hi-C data from two different developmental time courses; we successfully recapitulate previously identified topological changes around key genes, identify previously unknown topological changes around genes, and predict novel boundary factors that could interact with known architectural proteins to form topological domains. Taken together, GRiNCH is a robust and broadly applicable approach to discover structural units and smooth sparse high-throughput 3C datasets from diverse platforms including Hi-C, SPRITE and HiChIP.

2.2 Results

2.2.1 GRiNCH, a non-negative matrix factorization-based method for analyzing high-throughput chromosome conformation capture datasets

GRiNCH uses graph-regularized Non-negative Matrix Factorization (NMF) to identify topologically associating domains (TADs) from a high-dimensional 3C count matrix (**Figure 2.1, Methods**). GRiNCH has several properties that make it attractive for analyzing these count matrices: (1) matrix factorization methods including NMF have a “matrix completion” capability, which can be used to smooth noisy, sparse matrices, (2) the low-dimensional factors provide a clustering of the row and column entities that can be used to define chromosomal structural units, (3) the non-negativity constraint of the factors provide a parts-based representation of the data and is well suited for count datasets (such as Hi-C matrices), and (4) GRiNCH can be applied to any count matrix measuring chromosomal interactions between genomic loci such as Hi-C (Rao et al., 2014), SPRITE (Quinodoz et al., 2018), and HiChIP (Mumbach et al., 2016) datasets. Previously, NMF has been used for bias correction and dimensionality reduction of Hi-C data (Hu et al., 2016); however, this approach is applicable to only symmetric matrices while GRiNCH implementation can be easily extended to handle asymmetric matrices. Furthermore, smoothing properties of NMF has not been considered for Hi-C data.

For the ease of description, we will consider a Hi-C matrix as the input to GRiNCH. In GRiNCH, the count matrix is approximated by the product of two lower dimensional matrices, U and V , both with dimension $n \times k$, where n is the number of genomic regions in the given chromosome, and k is the rank of the lower-dimensional space. Because Hi-C matrices have a strong distance dependence, we use a constrained formulation of NMF,

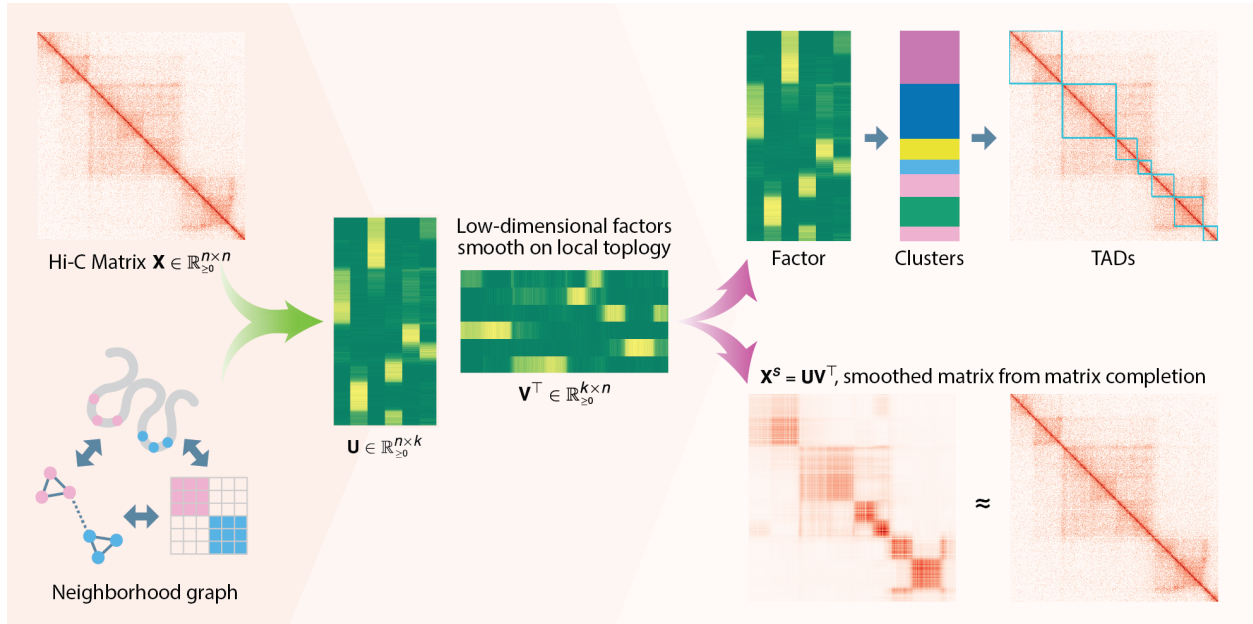


Figure 2.1: Overview of GRiNCH.. GRiNCH applies Non-negative Matrix Factorization (NMF) to a Hi-C or a similar high-throughput 3C matrix to find clusters of densely interacting genomic regions. NMF recovers low-dimensional factors U and V of the input matrix X that can be used to reconstruct the input matrix. As nearby genomic regions tend to interact more with each other, we regularize the factor matrices with a neighborhood graph to encourage neighboring regions to have a similar lower-dimensional representation, and subsequently belong to the same cluster. We cluster the regions by treating one of the factor matrices as a set of latent features and applying k-medoids clustering. The clusters represent topological units such as TADs. The factor matrices can be multiplied together to yield a smoothed version of the input matrix which is often sparse and noisy.

where the columns of the U and V matrices are favored to be smooth on a graph of genomic regions (**Figure 2.1**), such that regions that are connected in the graph have similar sets of values in the lower-dimensional space. The graph in turn captures the distance dependence using a local neighborhood, where two regions i and j have an edge between them if they are within a particular radius r of each other in linear distance along the chromosome. GRiNCH has three parameters, k , used for both the rank of the lower dimensional space and the number of TADs, r to control the size of the neighborhood, and λ to control the strength of graph regularization. After factorization, GRiNCH uses chain-constrained k-medoids clustering to define clusters of contiguous regions, which we consider as TADs.

We probed the impact of the three parameters, k , r , and λ , on the resulting GRiNCH TADs (**Supp Figure A.1**). We determined that setting k to identify TADs of size $\sim 1\text{Mb}$, with a neighborhood size of $r = 250\text{kb}$ and a small amount of regularization ($\lambda = 1$), yields the best results. Notably, the regularization yields TADs with higher CTCF enrichment than vanilla matrix factorization without any regularization (i.e. $\lambda = 0$).

2.2.2 GRiNCH TADs are high quality and stable to varying resolution and depth of input Hi-C data

To assess the quality of GRiNCH TADs, we considered seven existing TAD identification methods (**Methods**) and applied them along with GRiNCH to Hi-C data of five different cell lines from Rao et al., 2014 for comparison. The quality of a TAD was measured with two internal validation metrics used for cluster evaluation, Davies-Bouldin index (DBI) and Delta Contact Count (DCC), both assessing the similarity of interaction profiles of regions within defined TADs. DBI of a cluster measures how well separated the given cluster is from other clusters; in our case, how distinct each TAD's interaction count profile is from other TADs (**Methods**); a lower value for DBI indicates a more distinct, better-separated cluster. DCC measures the difference between intra-TAD interaction counts and inter-TAD interaction counts, with higher difference associated with better TADs. For each TAD-finding algorithm, we measured the percentage of predicted TADs with significantly better DBI or DCC value compared to DBI or DCC values from randomly shuffled TADs within the same chromosome (**Methods**). When comparing DBI, TopDom, GRiNCH, and Directionality Index have the highest percentage of their TADs with significant DBI in majority of the cell lines (GM12878, HUVEC, K562); based on DCC, HiCseg, GRiNCH, and Directionality rank the highest across all cell lines (**Figure 2.2A**). Overall GRiNCH was among the top three methods for both internal validation metrics in TAD quality

evaluation.

Many TAD-calling methods are sensitive to the input data resolution (size of genomic region), with the resulting TAD lengths varying greatly as a function of resolution (Zufferey et al., 2018). A robust method is expected to yield TADs with consistent length distribution and composition when given the same user-specified parameter settings, regardless of the resolution. Therefore, we next assessed the ability of GRiNCH and the seven TAD calling methods for their ability to recover stable TADs across different resolutions, 10kb, 25kb, and 50kb. We first compared the overall length distribution across different resolutions (**Figure 2.2B; Supp Figure A.2**), and found that GRiNCH and Directionality Index are the most stable, with the exception of NHEK where Directionality Index learns longer TADs at 10k resolution (**Supp Figure A.2**). We next evaluated the overall similarity of TADs identified at different resolutions with metrics to quantify the similarity of pairs of clustering results: Rand Index and Mutual Information (**Methods**). Intuitively, Rand Index is a measure of cluster membership consistency; it measures whether two data points (in our case, two region bins) that belonged to the same cluster (TAD) in one clustering result also stayed together in the other result, and whether two data points that belonged to different clusters stayed separate. Rand Index ranges from 0 to 1, with 1 being perfect concordance. Mutual Information is an informational-theoretic metric measuring the dependency between two random variables, where each variable indicates a clustering result. A Mutual Information of 0 indicates complete disagreement and the higher the Mutual Information value the better the agreement between the corresponding clustering results. To enable comparison across resolutions with different number bins, we split the lower-resolution (10kb, 25kb, 50kb) bins to constituent bins of size 5kb, the size of the lowest common denominator. We assigned these 5kb bins the same cluster as the original lower-resolution bin (**Methods**). We find that for every pair of resolutions compared, e.g., TADs from 10kb vs. 50kb, TopDom, GRiNCH, and rGMAP rank in the top three for both

Rand Index (**Figure 2.2C**) and Mutual Information (**Supp Figure A.3**).

These results suggest that GRiNCH is robust to different resolutions, recovering consistent TADs across different resolutions.

TAD-calling methods can be sensitive to the sparsity of the Hi-C matrices due to low sequencing depth (Zufferey et al., 2018). To assess the robustness of each method to low-depth, sparse datasets with many zero entries, we first took the highest-depth dataset (GM12878, 4.9 billion mapped paired-end reads) and downsampled to the depth and sparsity level of lower-depth data from other cell lines (e.g. K562, the second “deepest” cell line with 932 million reads). We then compared the similarity of the TADs from the original high-depth data and those from the downsampled counterpart (**Figure 2.3A, Methods**), again using Rand Index and Mutual Information. Based on Rand Index, TopDom, HiCseg, and GRiNCH yield the most reproducible TADs across different depths, particularly at the lower depths of HMEC, HUVEC, and NHEK cell lines. Based on Mutual Information, TopDom is the most consistent followed by GRiNCH and HiCseg. Other methods were generally less consistent based on the Mutual Information metric.

A third hindrance in the interpretation of results from TAD finding methods is the disagreement on the TAD definitions (Zufferey et al., 2018; de Wit, 2019). Hence, we further evaluated whether different TAD-calling methods yielded relatively similar TADs, and which sets of methods yielded the most similar TADs to one another. Here again, we used Rand Index and Mutual Information as metrics to compare the sets of TADs from different methods. All pairwise comparisons of TAD-calling methods yielded high values of Rand Index (>0.8) and high Mutual Information (**Figure 2.3B,C**). Furthermore, GRiNCH and TopDom yield the most similar sets of TADs, followed by rGMAP across all cell lines. This pattern is fairly consistent even when analyzed for each cell line individually (**Supp Figure A.4**).

To summarize, our internal validation and stability analysis showed that the top per-

forming methods depends upon the evaluation criteria. However, GRiNCH is among the top performing methods for all the criteria we examined (**Figure 2.4**), producing TADs that are as good or better than existing methods and are stable to varying resolution and depth.

2.2.3 GRiNCH TADs are enriched in architectural proteins and histone modification signals.

We next characterized GRiNCH TADs as well as TADs from other methods for their ability to capture well-known one-dimensional signal enrichment patterns. In particular, one hallmark of TADs is the enrichment of architectural proteins such as CTCF and cohesin elements (RAD21, SMC3) on the boundaries of TADs (Chang et al., 2019; de Wit, 2019). We tested the TAD boundaries from each method for the enrichment of peaks of CTCF, RAD21, and SMC3 in the five Rao et al., 2014 cell lines with Hi-C data (**Figure 2.5A, Methods**). All methods identified boundaries enriched for peaks of these proteins; however, the methods varied in their relative performance across cell lines. GRiNCH TAD boundaries have comparable or better enrichment as the other top performing methods, namely, Directionality Index and Insulation Score in most cell lines, and HiCseg in K562 and HUVEC. All these methods including GRiNCH have significantly higher enrichment than 3DNetMod, rGMAP, Armatus across different cell lines. The lower performance of these three methods could be due to their focus on hierarchical topological domains.

As histone modifications have been shown to be associated with three-dimensional organization (Andrey et al., 2017), we next measured the proportion of TADs with significant levels of mean histone modification signals (**Figure 2.5B**) compared to randomly shuffled TADs (**Methods**). The histone modification signals include promoter- (H3K4me3, H3K4me2), elongation- (H3K79me2, H3K36me3), and enhancer-associated marks (H3K27ac),

and repressive chromatin marks (H3K27me3). A larger proportion of GRiNCH TADs, along with Armatus and HiCseg TADs, are consistently enriched for the activating histone marks such as H3K27ac, and the elongation marks, H3K36me3 and H3K79me2 across multiple cell lines and different resolutions (**Supp Figure A.5**). Interestingly, with the exception of GM12878, the enrichment of histone marks in the TADs from Insulation and Directionality index was much lower than the other methods suggesting these methods tend to find TADs defined by CTCF and might miss other types of TADs (Chang et al., 2019). These enrichment patterns show that when considering existing methods, there is a trade-off in the ability to recover TADs that are associated with CTCF and TADs that are associated with significant histone modifications. However, GRiNCH ranks among the top methods for both criteria (**Figure 2.4**) suggesting that GRiNCH TADs capture a diverse type of TADs.

2.2.4 GRiNCH smoothing of low-depth datasets help recover structure and significant interactions.

Our analysis so far compared different TAD finding methods for their ability to recover stable and biologically meaningful topological units. However, most Hi-C datasets are sparse, which can influence the TAD predictions significantly. Smoothing the input Hi-C matrix to impute missing values can enhance the visualization of topological units on the matrix (Yang et al., 2017; Ursu et al., 2018), improve the agreement among biological replicates (Yang et al., 2017), and assist in identifying loops and differential interactions (Rowley et al., 2020; Ardakany et al., 2019). Unlike existing TAD-calling methods, the matrix factorization framework of GRiNCH provides a natural matrix completion solution that can generate a smoothed version of the sparse input Hi-C matrix.

We first compared GRiNCH's smoothing functionality to common smoothing tech-

niques such as mean filter (Yang et al., 2017) and Gaussian filter (Ardakany et al., 2019), which have been used for Hi-C data pre-processing (Yang et al., 2017; Rowley et al., 2020; Ardakany et al., 2019). We additionally compared against a supervised learning method, HiCNN (Liu and Wang, 2019), which is based on a convolutional neural network and predicts high-resolution Hi-C data after training with high and low-depth data. We used three pre-trained models provided by HiCNN, trained on GM12878 data downsampled to 1/8, 1/16, and 1/25 depth. We used two metrics to assess the quality of smoothing: (a) recovery of TADs and (b) recovery of significant interaction after smoothing downsampled data (**Methods**). To perform these comparisons, we again used the downsampled GM12878 datasets with depths equal to each of the other four cell lines from Rao et al., 2014.

To assess TAD recovery from low-depth data, we identified TADs on the original high-depth GM12878 dataset and compared them to the TADs identified in the downsampled and smoothed data matrices using Rand Index and Mutual Information. Here, to avoid any bias in our interpretation, we used the Directionality Index method to call TADs.

We find that based on both Rand Index and Mutual Information, TADs recovered from GRiNCH-smoothed matrices are the most similar to the TADs from the high-depth dataset, performing better than mean filter and Gaussian filter for different parameter settings. Furthermore, GRiNCH outperforms HiCNN in all downsampled datasets across all three pre-trained HiCNN models (**Figure 2.6A**). The usefulness of GRiNCH is more apparent for lower-depth datasets (e.g. downsampled to NHEK depth).

To compare the smoothing methods on the recovery of significant interactions from low-depth data, we applied Fit-Hi-C on the original GM12878 dataset and on the downsampled and smoothed datasets to identify significant interactions ($q\text{-value} < 0.05$). Treating the significant interactions in the original high-depth dataset as the ground truth, we measured precision and recall as a function of the statistical significance of interactions from the smoothed datasets and computed the Area Under Precision-Recall curve (AUPR). The

higher the AUPR, the better the recovery of significant interactions after smoothing. As HiCNN predictions are limited to interactions less than 2Mb apart, we measured AUPR for interactions less than 2Mb and for all interactions separately (**Figure 2.6B**). When comparing interactions less than 2Mb apart, the HiCNN model trained with 1/8 depth of the original GM12878 dataset outperformed the other methods (mean filter, Gaussian filter, GRiNCH). This is not surprising as HiCNN was trained on the GM12878 cell line. HiCNN models trained on even lower depth (1/16, 1/25) data are at par or worse than GRiNCH for most datasets. Compared to mean filter and Gaussian filter, GRiNCH has a higher recovery of significant interactions on all the downsampled datasets with the exception of K562, where Gaussian filter outperformed both GRiNCH and HiCNN. When comparing all interactions including those further than 2Mb, GRiNCH has the highest AUPR compared to mean filter and Gaussian filter.

We additionally applied GRiNCH smoothing to Hi-C data collected from the same biological context but using different Hi-C protocols in order to evaluate whether it can help overcome artifacts introduced by the experimental protocol (e.g. the restriction enzyme used for digestion) and improve the concordance of TADs and significant interactions identified from these datasets. Using GRiNCH, we smoothed GM12878 25kb resolution datasets from three Hi-C protocols: *in situ* Hi-C using DpnII for digestion, *in situ* Hi-C using MboI, and a dilution Hi-C experiment using HindIII (**Methods**). To independently verify the smoothing capability of GRiNCH, we again used a different TAD-calling method (Directionality Index) to identify TADs on the original and the smoothed data. The similarity of TADs, measured by Rand Index and Mutual Information, was higher among GRiNCH-smoothed datasets than among the original datasets without smoothing (**Supp Figure A.6A,B**). We next used Fit-Hi-C (Ay et al., 2014) to identify significant interactions ($q\text{-value} < 0.05$) in the original and the smoothed data. We measured the overlap in the significant interactions identified from different datasets using Jaccard Index. We find

that GRiNCH-smoothed data shared a larger portion of significant interactions compared to the original unsmoothed data (**Supp Figure A.6C**). This demonstrates that GRiNCH smoothing is not sensitive to experimental artifacts such as restriction enzymes and can help improve the concordance between datasets from different platforms to detect shared topological units and significant interactions.

Overall, our experiments show that GRiNCH smoothing enables improved recovery of TAD structures and long-range interactions from lower-depth datasets, and helps recapitulate shared underlying biological signals beyond the experimental artifacts.

2.2.5 GRiNCH application to chromosomal organization during development.

To assess the value of GRiNCH in primary cells and to examine dynamics in chromosomal organization, we applied GRiNCH to two time-course Hi-C datasets profiling 3D genome organization during (a) mouse neural development (Bonev et al., 2017) and (b) pluripotency reprogramming in mouse (Stadhouder et al., 2018). Bonev et al., 2017 used high-resolution Hi-C experiments to measure 3D genome organization during neuronal differentiation from the embryonic stem cell state (mESC) to neural progenitor cells (NPC) and cortical neurons (CN). We applied GRiNCH on all chromosomes for all three cell types and compared them based on the overall similarity of TADs between the cell lines. Based on the two metrics of Mutual Information and Rand Index, the overall TAD similarity captured the temporal ordering of the cells, with mESC the most distinct and CN being closer to NPC (**Supp Figure A.7**). To assess whether GRiNCH can recover previously identified TAD dynamics, we next focused on a specific 4Mb region around the *Zfp608* gene, which was found by Bonev et al., 2017 as a neural-specific gene associated with a changing TAD boundary. In both NPC and CN, GRiNCH predicts a TAD near the *Zfp608*

gene, which is not present in the mESC state. Zfp608 was also associated with increased expression, and activating marks, H3K27ac and H3K4me3 at these time points, which is consistent with Zfp608 being a neural-specific gene (**Figure 2.7A**).

To identify novel genomic regions associated with changing 3D structure, we compared GRiNCH TADs across the time points (**Methods**) and identified 966 regions with dynamic 3D structure. Several of these regions are associated with neural-specific gene expression or implicated in neurological disorders. For example, we found TAD splits in the vicinity of Syap1 and Ap1s2 genes in the neural progenitor and cortical neuron cells, accompanied by corresponding increase in their gene expression (**Figure 2.7B**). Syap1-deficient mice have been shown to display motor and movement defects (R von Collenberg et al., 2019); Ap1s2 has been associated with intellectual disability, basal ganglia disease, and seizures accompanying Pettigrew syndrome (Cacciagli et al., 2014). Another example of dynamic 3D organization identified by GRiNCH was near the Arl6ip1 and Foxp1 genes (**Supp Figure A.8**). These genes are involved in glutamate neurotransmitter transport (Akiduki and Ikemoto, 2008) and neural differentiation (Braccioli et al., 2017), respectively. Visual inspection of results from other top-performing TAD-calling methods in the corresponding regions (**Supp Figure A.9, Supp Figure A.10, Supp Figure A.11**) did not capture these dynamic reconfigurations either because they did not predict any TADs or the TADs were too small. Overall this suggests that GRiNCH's ability to smooth and define TADs provides greater stability and sensitivity to detect these novel dynamic shifts in TAD structure between developmental stages.

We examined another time-course dataset which studied the 3D genome organization during reprogramming of mouse pre-B cells to pluripotent stem cells (PSC), with four intermediate time points (Day 2, 4, 6, and 8; see **Methods**). As in the neural developmental time course, we applied GRiNCH to all chromosomes from each time point and compared the overall 3D genome configuration over time. Here too we observed that time points

closer to each other generally had greater similarity in their TAD structure with replicates within the same time point displaying even greater similarity (**Supp Figure A.12**). We examined the interaction profile in the 1.3 Mb around the Sox2 gene, a known pluripotency gene (**Supp Figure A.13**). We see a gradual formation of a boundary around Sox2, which is also associated with concordant increase in expression, accessibility and the presence of H3K4me2, an active promoter mark.

While architectural proteins such as CTCF and cohesin play important roles in establishing TAD boundaries, it is currently unclear if there are additional DNA binding proteins that could, independently or in concert with CTCF, contribute towards establishing these boundaries, especially in a cell type-specific manner. Previous work to identify such regulatory proteins has focused on a single time point (Hong and Kim, 2017) or stage (Ramírez et al., 2018). As chromatin accessibility data was measured at each timepoint in the reprogramming dataset, we asked if we could identify additional regulatory proteins that could play a role in establishing TADs (**Methods**). Briefly, we tested the GRiNCH TAD boundaries from each mouse cell type, from pre-B cell to pluripotent cells, for enrichment of accessible motif instances of 746 transcription factors in the JASPAR 2020 core vertebrate motif database (Fornes et al., 2020). We ranked the TFs based on their significant enrichment in each cell type (**Figure 2.7C**, **Supp Table A.2**). The top-ranking TF across the cell types was CTCF, which is consistent with its role as an architectural protein in establishing TADs (**Figure 2.7C**). We also found other factors in the same zinc finger protein family as CTCF (Cassandri et al., 2017), such as ZBTB14, Plagl2/1, ZIC1/3/4/5, CTCFL, YY1/2 that were enriched across the cell types. YY1 and YY2, which are 65 and 56% identical in their DNA and protein sequence respectively in humans (Wu et al., 2017), are of interest as YY1 has been identified as an enforcer of long-range enhancer-promoter loops (Weintraub et al., 2017). Interestingly, we found several hematopoietic lineage factors, such as STAT3 and FOXP3, ranked highly in the pre-B cell TADs compared to other time points. STAT3

is needed for B cell development (Chou et al., 2006). FOXP3 is a master regulator of T cells (Lu et al., 2017), but could be involved in the suppression of B cells. We also found a number of HOX transcription factors, HOXA4, HOXA5, HOXB2, HOXB5, HOXB7, and the transcription factor MEIS3 to be ranked highly in the B cells. The HOX genes depend upon MEIS3 (Uribe and Bronner, 2015) to bind to their targets, supporting the simultaneous enrichment of these factors.

We repeated this analysis for the Rao et al., 2014 cell lines (**Supp Table A.3**). Here too we found CTCF and YY1/2 proteins highly enriched across cell lines. However, there was lesser degree of cell-line specificity for this dataset. Taken together, this analysis suggests that GRiNCH captures high-quality TADs, which can be used to define global and locus-specific similarities and differences in 3D genome organization between cell types. Furthermore, the GRiNCH boundary enrichment analysis identified novel transcription factors that could be followed up with downstream functional studies to examine their role in 3D genome organization.

2.2.6 GRiNCH can be used for a variety of 3D conformation capture technologies

Although Hi-C is still the most widely used technology to map 3D genome structure, recently several new methods have been developed to measure chromosomal contacts on a genome-wide scale (Kempfer and Pombo, 2020). To assess the applicability of GRiNCH to these technologies, we considered two complementary techniques to measure 3D genome organization: Split-Pool Recognition of Interactions by Tag Extension (SPRITE) (Quinodoz et al., 2018) and HiChIP (Mumbach et al., 2016). SPRITE measures multi-way chromatin interactions, and captures interactions across larger spatial distances than Hi-C. In HiChIP, long-range chromatin contacts are first established *in situ* in the nucleus before lysis; then

chromatin immunoprecipitation (ChIP) is performed with respect to a specific protein or histone mark, directly capturing interactions associated with a protein or histone mark of interest (Mumbach et al., 2016). A common property of both technologies is that they generate a contact count matrix, which is suitable for GRiNCH.

We applied GRiNCH to GM12878 contact matrices measured with SPRITE (Quinodoz et al., 2018), cohesin HiChIP (Mumbach et al., 2016), and H3k27ac HiChIP (Mumbach et al., 2016). A visual comparison between these datasets for an 8Mb region of chr8 shows regions of good concordance between datasets (**Figure 2.8A-D**). We quantified the global similarity of GRiNCH TADs from the four different datasets, for all chromosomes with Rand Index and Mutual Information (**Figure 2.8E,F**). Interestingly, the GRiNCH TADs from Hi-C are the most similar to those from cohesin HiChIP and this similarity measure is higher than between the two HiChIP datasets. This is consistent with cohesin being a major determinant for the formation of loops detected in Hi-C datasets. The H3K27ac HiChIP data is as close to Hi-C as it is to cohesin HiChIP. Finally the most distinct set of TADs are identified by SPRITE, which is consistent with SPRITE capturing multi-way interactions and longer-distance interactions. Despite the differences in the specific TAD boundaries, overall the datasets look similar across different platforms (Rand Index >0.97). Taken together, this shows that GRiNCH is broadly apply to different experimental platforms for measuring genome-wide chromosome conformation.

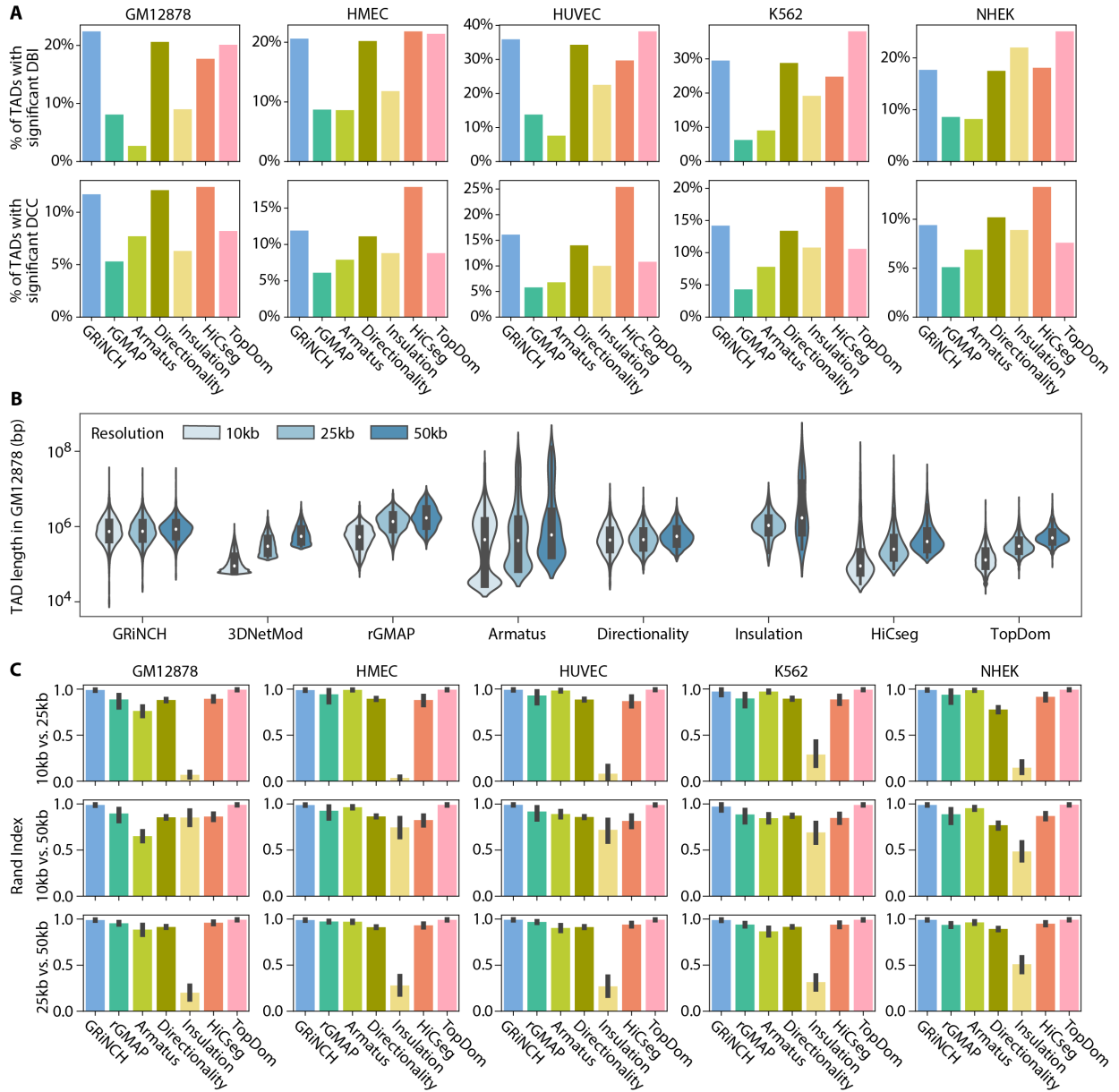


Figure 2.2: Characterizing TADs with internal validation metrics, TAD size, and composition.. **A.** Percentage of TADs with significant Davies-Bouldin Index (DBI) and Delta Contact Count (DCC) values. Shown are values for GRiNCH and six other methods. The higher the bar, the better a method. Note: 3DNetMod outputted overlapping TADs and was excluded from this analysis which involves TAD shuffling. **B.** The size distribution of TADs from GM12878. Y-axis is in log10 scale of base pairs. The white dot represents the median; the black box ranges from the 25th percentile to 75th. 10kb data from insulation is missing because it did not return any TADs when using the same hyperparameters as in 25kb and 50kb data. **C.** Similarity between TADs from higher- and lower-resolution data (e.g. 10kb vs. 25kb) measured by Rand Index. The higher the number, the higher the similarity. The error bar denotes the standard deviation from the mean across chromosomes. Note: 3DNetMod outputted overlapping TADs and was excluded from this analysis due to the requirement of unique cluster assignment for each region.

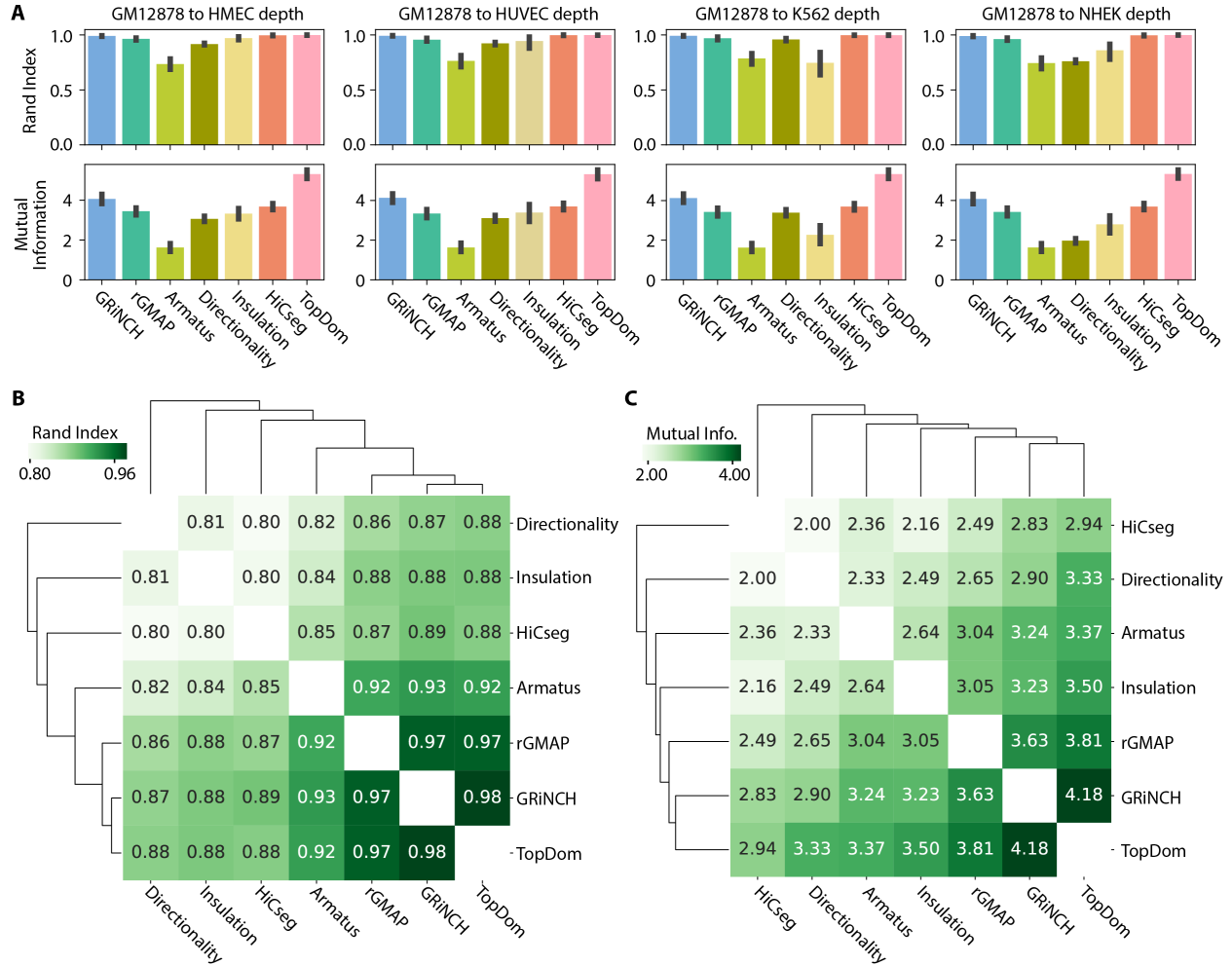


Figure 2.3: Evaluating the stability of different TAD-calling methods to datasets of different depths.. **A.** The mean similarity, across chromosomes, between TADs from high-depth GM12878 dataset and TADs from low-depth GM12878 datasets obtained by downsampling the GM12878 dataset to different depths observed in our five cell-line dataset. The similarity of the TADs is measured by Rand Index and Mutual Information. The error bar denotes the standard deviation from the mean. **B.** Similarity of TADs from pairs of TAD-calling methods (e.g. GRiNCH vs. TopDom), measured by Rand index. The higher the number, the higher the similarity. **C.** Similarity of TADs from pairs of TAD-calling methods measured by Mutual information. Note: 3DNetMod outputted overlapping TADs and was excluded from this analysis due to the requirement of unique cluster assignment for each region.

	Validation		Resolution			Depth		Consistency	Enrichment	
	DBI	DCC	Size	RI	MI	RI	MI		CTCF	Histone
GRiNCH										
3DNetMod										
rGMAP										
Armatus										
Directionality										
Insulation										
HiCseg										
TopDom										

Figure 2.4: Summary of benchmarking TAD-calling methods.. Shown are different criteria of evaluation. A medal denotes whether the given TAD-calling method is among the top 3 methods for a particular criteria (gold/yellow: 1st place; silver/grey: 2nd place; bronze/brown: 3rd place). **Validation:** internal validation metrics for measuring the cohesiveness of predicted TADs. DBI: percentage of TADs with significant Davies-Bouldin Index (**Supp Table A.1A**); DCC: percentage of TADs with significant Delta Contact Counts (**Supp Table A.1B**). **Resolution:** measuring stability of TADs to changing input data resolution (e.g. 10kb, 25kb, 50kb). Size: stability of median TAD size to Hi-C resolution (**Supp Table A.1C**); RI, MI: similarity of TADs from high- and low-resolution data, measured by Rand Index (RI, **Supp Table A.1D**) and Mutual Information (MI, **Supp Table A.1E**). **Depth:** measuring stability of TADs to the depth and sparsity of input data. RI, MI: similarity of TADs from high-depth and low-depth data, measured by Rand Index (RI, **Supp Table A.1F**) and Mutual Information (MI, **Supp Table A.1G**). **Consistency:** a group of methods yielding TADs with highest similarity, with gold for the pair of methods with highest similarity according to hierarchical clustering. **Enrichment:** measuring enrichment of regulatory signals. CTCF: fold enrichment of CTCF and cohesin elements in TAD boundaries (**Supp Table A.1H**); Histone: proportion of TADs with significant mean histone signal (**Supp Table A.1I**).

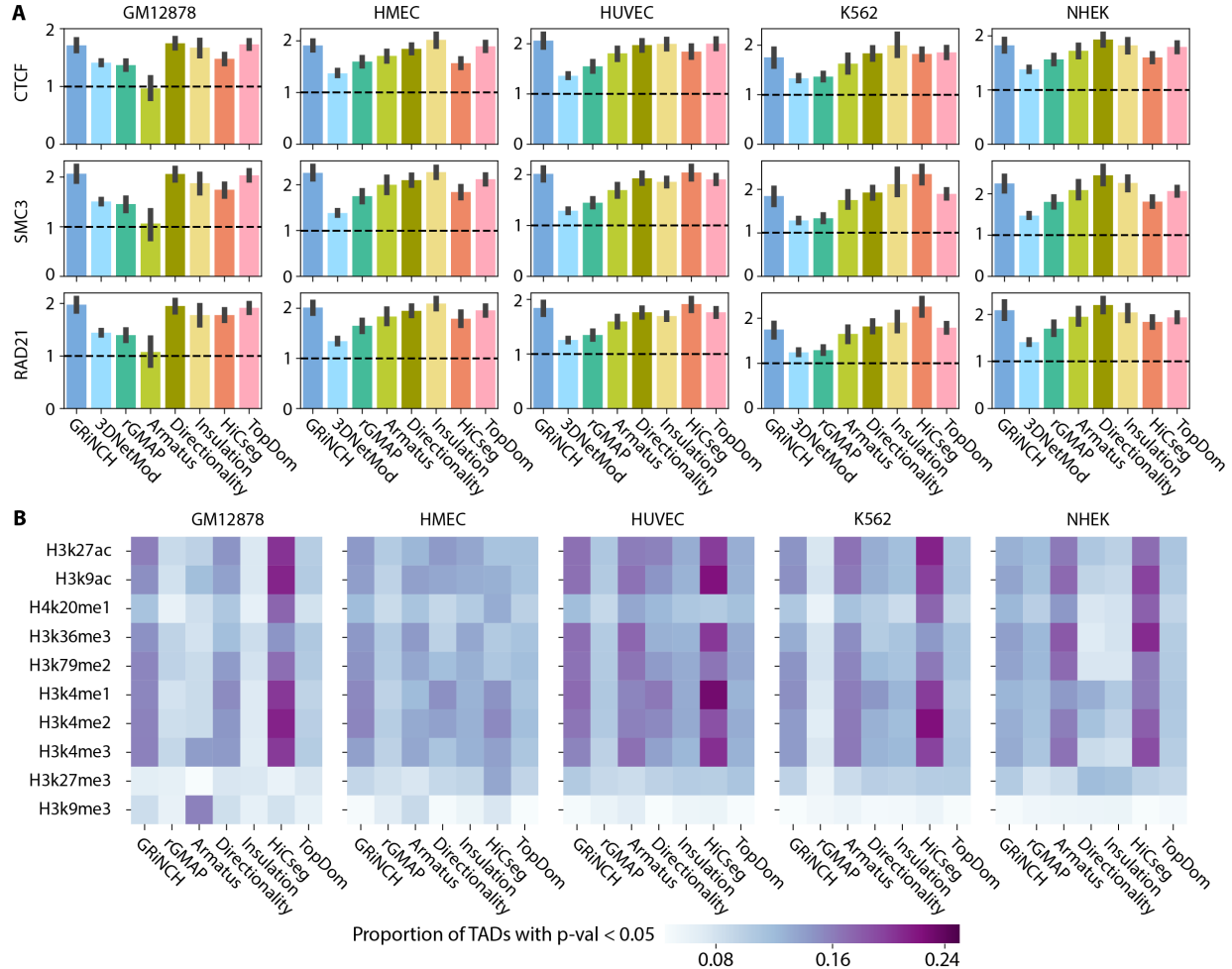


Figure 2.5: Evaluating TAD-calling methods with enrichment of boundary elements and regulatory signals.. **A.** Fold enrichment of binding signals of architectural protein in TAD boundaries. Shown are the mean fold enrichment of CTCF ChIP-seq peaks and accessible motif instances of cohesin proteins, RAD21 and SMC3, estimated across multiple chromosomes. The error bar denotes the standard deviation from the mean. **B.** Proportion of TADs with significant mean histone modification signal (i.e. empirical p-value < 0.05). The darker the entry the higher the proportion of TADs with significant histone enrichment. The average ChIP-seq signal for each histone modification mark was taken from within each TAD; the p-value of each TAD is derived from an empirical null distribution of mean signals in randomly shuffled TADs. Note: 3DNetMod outputted overlapping TADs and was excluded from this analysis as it involves TAD randomization/shuffling.

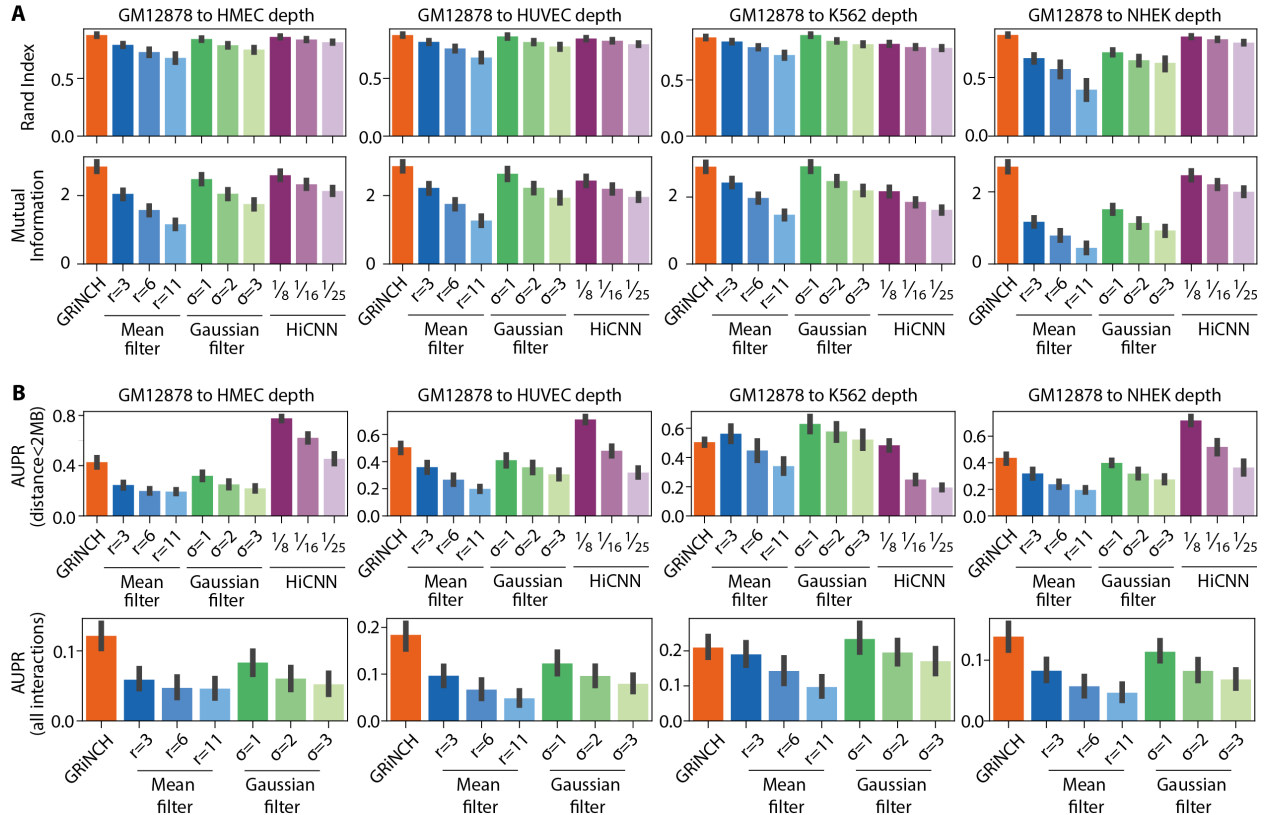


Figure 2.6: Evaluating the benefits of smoothing in GRiNCH.. Recovery of topology and significant interactions from downsampled then smoothed data. **A.** Rand Index and Mutual Information were used to measure the similarity between TADs from high-depth GM12878 dataset and TADs from downsampled datasets smoothed by different methods (GRiNCH, Mean Filter, Gaussian Filter, HiCNN). Directionality was used as a TAD-calling method independent of any of the smoothing methods, i.e., GRiNCH. The mean is computed across chromosomes and the error bar denotes deviation from the mean. **B.** Area Under Precision-Recall curve (AUPR) was used to measure the recovery of significant interactions called by Fit-Hi-C. Precision and recall were measured for significant interactions from downsampled and smoothed datasets against the “ground truth” defined by the significant interactions from the high-depth GM12878 dataset. Since the pretrained HiCNN models imputes interactions up to 2MB apart, the AUPR for interactions < 2MB apart and for all interactions are shown here.

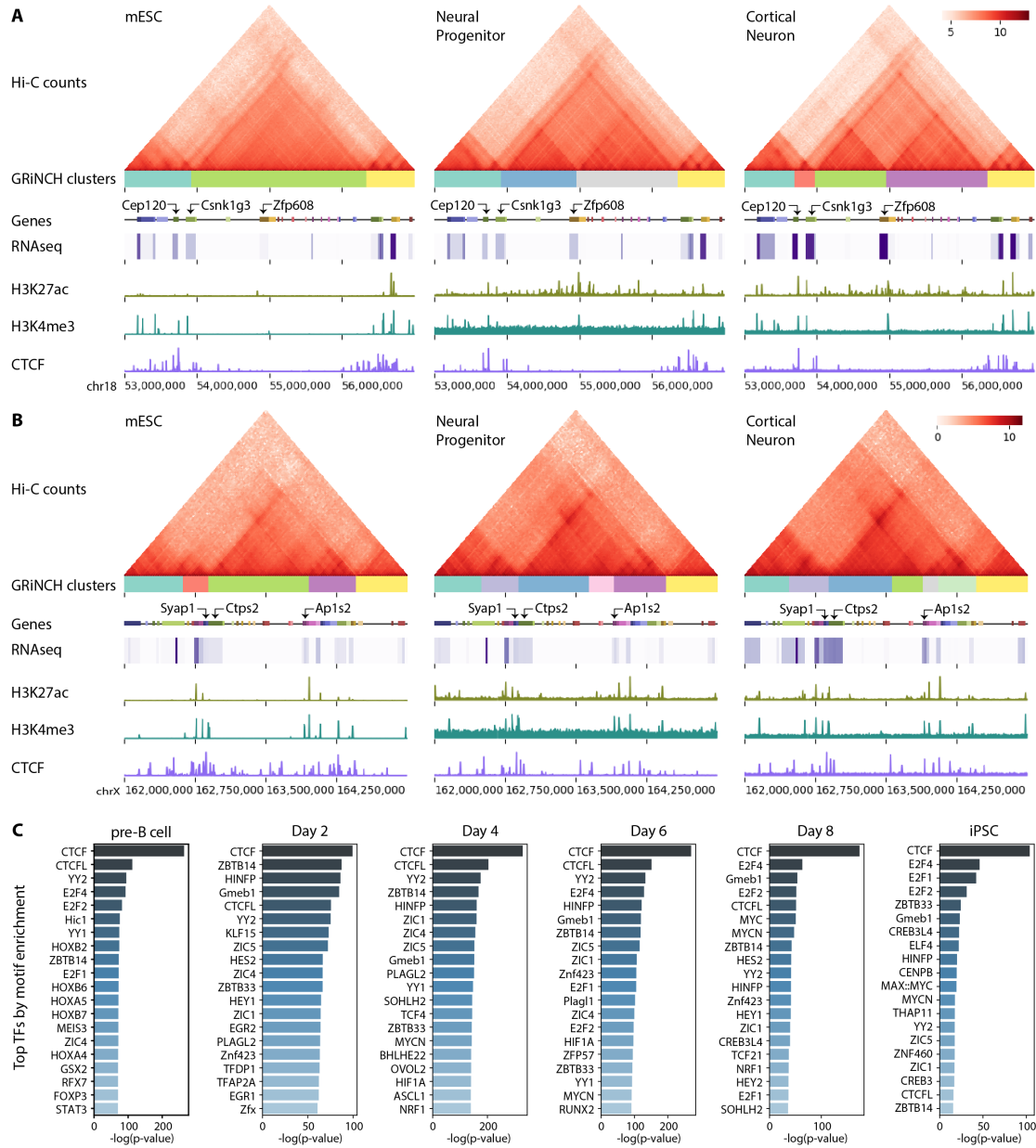


Figure 2.7: GRiNCH applied to Hi-C datasets along developmental time courses.. A. Interaction profile near the Zfp608 gene in mouse embryonic stem cells (mESC), neural progenitors (NPC), and differentiated cortical neurons (CN). Heatmaps are of Hi-C matrices after log2-transformation of interaction counts for better visualization. GRiNCH clusters are visualized as blocks of different colors under the heatmap of interaction counts. Genes in the nearby regions are marked by small boxes, and a heatmap of their corresponding RNA-seq levels (in log-transformed TPM) is shown underneath each gene. ChIP-seq signals from H3K27ac, H3K4me3, and CTCF are shown as separate tracks. **B.** Interaction profile near Syap1 and Ap1s2 in mouse embryonic stem cells (mESC), neural progenitors (NPC), and differentiated cortical neurons (CN). **C.** Top 20 TFs from a collection of 746 TFs ranked based on their motif enrichment in GRiNCH TAD boundaries from the mouse reprogramming time course data. The significance of their fold enrichment was calculated with the hypergeometric test and TFs were ranked by descending negative log p-value.

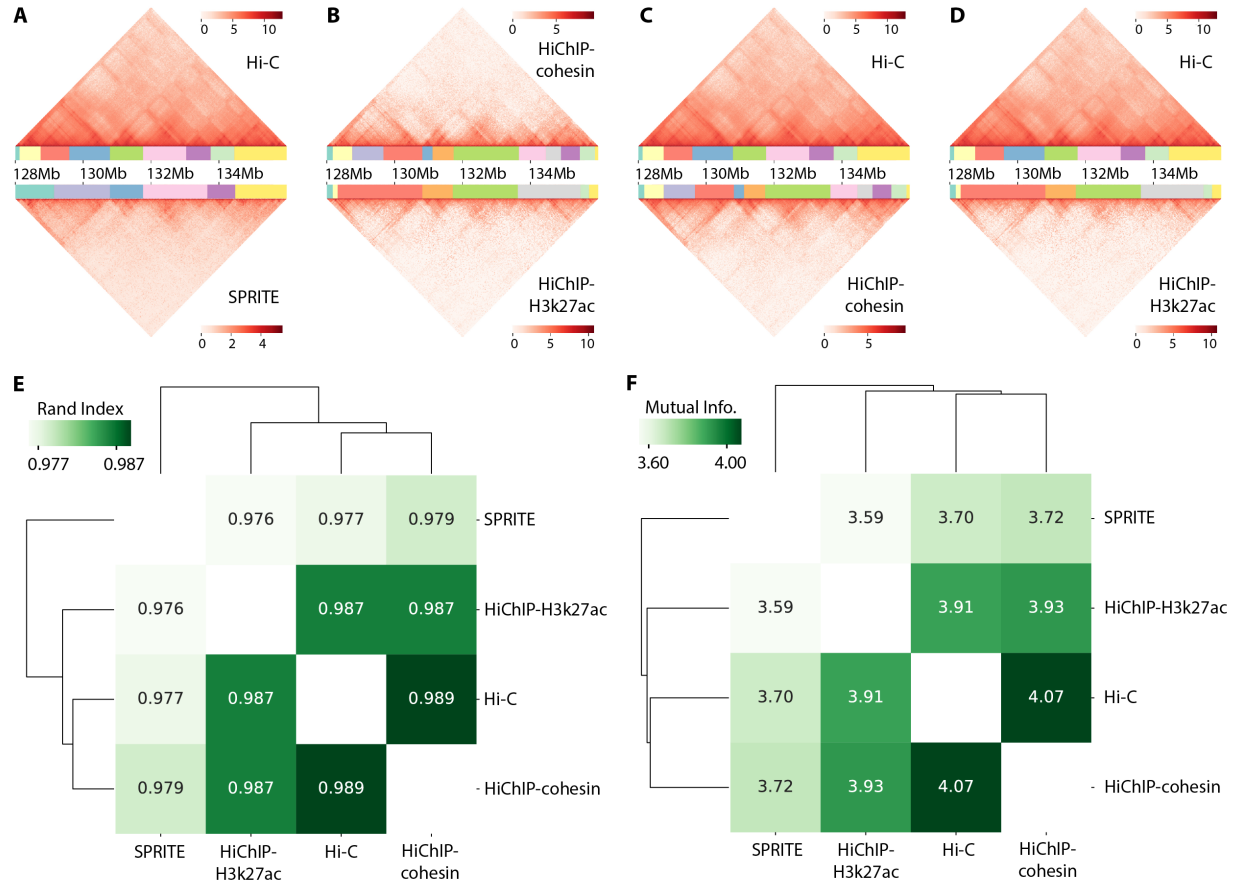


Figure 2.8: Applying GRiNCH to datasets from different 3D genome conformation capture technologies.. Visual comparison of the interaction profile and GRiNCH TADs from a 8Mb region in chr8, GM12878 cell line. GRiNCH TADs are visualized as blocks of different colors under the heatmap of interaction counts. **A.** Hi-C vs SPRITE. The top heatmap and clusters are from Hi-C; bottom from SPRITE. **B.** HiChIP with cohesin (top) vs HiChIP with H3k27ac (bottom). **C.** Hi-C (top) vs HiChIP with cohesin (bottom). **D.** Hi-C (top) vs HiChIP with H3K37ac (bottom). For visualization purposes all interaction counts were log2-transformed. **E.** Measuring the similarity of GRiNCH TADs from Hi-C and other 3D genome conformation capture platform (e.g. SPRITE, HiChIP with cohesin, or HiChIP with H3k27ac) in the same GM12878 cell line, with Rand Index. The dendrogram depicts the relative similarity between samples. **F.** Mutual-Information-based similarity of GRiNCH TADs from Hi-C and other technologies.

2.3 Methods

2.3.1 Graph-regularized Non-negative Matrix Factorization (NMF) and Clustering for Hi-C data (GRiNCH) framework

GRiNCH is based on a regularized version of non-negative matrix factorization (NMF, Cai et al., 2011) that is applicable to high-dimensional chromosome conformation capture data such as Hi-C (**Figure 2.1**). Below we describe the components of GRiNCH: NMF, graph regularization, and clustering for TAD identification.

Non-negative matrix factorization (NMF) and graph regularization

Non-negative matrix factorization is a popular dimensionality reduction method that aims to decompose a non-negative matrix, $X \in \mathbb{R}_{\geq 0}^{(n \times m)}$ into two lower dimensional non-negative matrices, $U \in \mathbb{R}_{\geq 0}^{(n \times k)}$ and $V \in \mathbb{R}_{\geq 0}^{(n \times k)}$, such that the product $X^* = UV^T$, well approximates the original X . We refer to the U and V matrices as factors. Here $k \ll n, m$ is the rank of the factors and is user-specified.

In application of NMF to Hi-C data, we represent the Hi-C data for each chromosome as a symmetric matrix $X = [x_{ij}] \in \mathbb{R}^{(n \times n)}$ where x_{ij} represents the contact count between region i and region j . We note that in the case of a symmetric matrix, U and V are the same or related by a scaling constant.

The goal of NMF is to minimize the following objective: $\|X - UV^T\|_F^2$, s.t. $U \geq 0, V \geq 0$ (Lee and Seung, 2000), where $\|X\|_F$ indicates the Frobenius norm. A number of algorithms to optimize this objective have been proposed; here we used the multiplicative update algorithm, where the entries of U and V are updated in an alternating manner in each iteration:

$$u_{ik} \leftarrow u_{ik} \frac{(XV)_{ik}}{(UV^T V)_{ik}}, \quad v_{jk} \leftarrow v_{jk} \frac{(X^T U)_{jk}}{(VU^T U)_{jk}} \quad (2.1)$$

Here u_{ik} corresponds to the i^{th} row of column $U(:, k)$ and v_{jk} corresponds to the j^{th} row of column $V(:, k)$.

Standard application of NMF to Hi-C data is ignorant of the strong distance dependence of the count matrix, that is, genomic regions that are close to each other tend to interact more with each other. To address this issue we apply a constrained version of NMF with graph regularization, where the graph represents additional constraints on the row (and/or column) entities (Cai et al., 2011). Graph regularization enables the learned columns of U and V to be smooth over the input graph. In our application of NMF to Hi-C data, we define a graph composed of genomic regions as nodes, with edges connecting neighboring regions in the linear chromosome, where the size of the neighborhood is an input parameter. Specifically, we define a symmetric nearest-neighbor graph, W :

$$W_{ij} = \begin{cases} 1, & \text{if } x_i \in N_r(x_j) \text{ and } x_j \in N_r(x_i) \\ 0, & \text{otherwise} \end{cases} \quad (2.2)$$

where $N_r(x_i)$ denotes r nearest neighbors in linear distance to region x_i .

Graph regularized NMF has the following objective:

$$\|X - UV^T\|_F^2 + \lambda \text{Tr}(V^T L V) + \lambda \text{Tr}(U^T L U), \quad (2.3)$$

where D is a diagonal matrix whose entries are column (or row, since W is symmetric) sums of W , i.e., $D_{ii} = \sum_j W_{ij}$. $L = D - W$ denotes the graph Laplacian and encodes the graph topology. The second and third terms are the regularization term and measures the smoothness of U and V with respect to the graph. Here λ is the regularization hyperparameter. This new objective has the effect of encouraging the factors to be smooth on the

local neighborhood defined by the graph. Accordingly, the multiplicative update rule from (2.1) gains regularization terms (Cai et al., 2011):

$$u_{ik} \leftarrow u_{ik} \frac{(XV + \lambda WU)_{ik}}{(UV^T V + \lambda DU)_{ik}}, \quad v_{jk} \leftarrow v_{jk} \frac{(X^T U + \lambda WV)_{jk}}{(VU^T U + \lambda DV)_{jk}} \quad (2.4)$$

Both r (neighborhood radius) and λ are parameters that can be specified, with λ setting the strength of regularization ($\lambda = 0$ makes this equivalent to basic NMF). See section on “Selecting GRiNCH hyper-parameters” below.

Chain-constrained k-medoids for clustering and TAD calling

The factors U (or V) can be used to extract clusters of the row (or column) entities of the input matrix. When X is symmetric, e.g., in our application to Hi-C, either U or V can be used to define the clusters (the factors are equivalent up to a scaling constant). Assuming we use U , there are two common approaches for finding clusters from NMF factors: (1) assign each row entity i to its most dominant factor, i.e., assign it to cluster $c_i = \operatorname{argmax}_{j \in \{1, \dots, k\}} u_{ij}$, or (2) apply k-means clustering on the rows of U . However, both approaches fall short in our application. The first approach is sensitive to extreme values which can still be present in the smoother factors, yielding non-informative clusters. Furthermore, neither approach reinforces contiguity of genomic regions in each cluster along their chromosomal position. As a result, a single cluster could potentially contain genomic regions from two opposite ends of the chromosomes instead of being a contiguous local structural unit. To address this problem, we apply chain-constrained k-medoids clustering. k-medoids clustering is similar to k-means clustering, except that the “center” of each cluster is always an actual data point, rather than the mean of the datapoints in the cluster. In its chain-constrained version (**Algorithm A.1**), adopted from spatially connected k-medoids clustering (Soor et al., 2018): each cluster grows outwards from initial medoids along the linear chromosomal

coordinates. The algorithm assigns a genomic region to a valid medoid region either upstream or downstream along the chromosome, ensuring the contiguity of the clusters and resilience to noise or extreme outliers provided by using a robust ‘median’-like cluster center rather than a ‘mean’-like center used in k-means clustering.

Selecting GRiNCH hyperparameters

GRiNCH has three hyper-parameters: (a) k , the rank of the lower-dimensional matrices, which can alternately be viewed as the number of latent features or clusters, (b) r , the radius of the neighborhood in the graph used for regularization, and (c) λ controlling the strength of regularization.

The parameter k determines the number of latent features to recover and the resulting number of GRiNCH TADs. We can obtain subTAD-, TAD-, or metaTAD-scale clusters (**Supp Figure A.14A**) by setting k such that the expected size of a cluster is 500kb, 1Mb, or 2Mb, i.e., k equals the given chromosome’s length divided by the expected size. We find that a larger portion of subTAD-scale clusters (i.e. expected TAD size = 500kb) have significant internal validation metric values (**Supp Figure A.14B**). SubTAD-scale clusters tend to be more stable to depth and sparsity (**Supp Figure A.14C**), and are also more enriched in boundary elements like CTCF (**Supp Figure A.15A**). As a tradeoff, higher proportion of metaTAD-scale clusters (i.e. expected cluster size = 2Mb) are enriched in histone modification marks (**Supp Figure A.15B**). Based on the use case of GRiNCH, k can be set dynamically by the user; by default, GRiNCH sets k such that the expected size of a cluster is 1Mb, or at TAD-scale.

For regularization strength, $\lambda \in \{0, 1, 10, 100, 100\}$ were considered, with $\lambda = 0$ equivalent to standard NMF without regularization. For neighborhood radius, $r \in \{25K, 50K, 100K, 250K, 500K, 1000K\}$ were considered, where $r = 100K$ in a Hi-C dataset of 25Kb resolution will use 4 bins on either side of a given region as its neighbors. We find that some regularization, with

$\lambda = 1$, yields better CTCF enrichment than other λ values (**Supp Figure A.1A**). With regularization, a neighborhood radius of 100Kb or larger yields higher CTCF enrichment (**Supp Figure A.1B**). We also note that the regularization parameters do not discernibly change the TAD size distribution (**Supp Figure A.16**). Based on these results, the default regularization parameters for GRiNCH are set at $\lambda = 1$ and $r = 250\text{kb}$.

Memory consumption and runtime

In graph-regularized NMF, the size of the input matrix n and the reduced dimension k are the main drivers of computational complexity which is $O(kn^2)$ (Cai et al., 2011). We measured memory consumption (maximum resident set size) and runtime of GRiNCH across five cell lines (GM12878, HMEC, HUVEC, NHEK, K562) with different combinations of input matrix size (determined by chromosome length and Hi-C resolution), expected cluster/TAD size (which determines k for a given matrix), and regularization parameters ($\lambda \in \{0, 1, 10, 100, 100\}$ and neighborhood radius $\in \{25\text{kb}, 50\text{kb}, 100\text{kb}, 250\text{kb}, 500\text{kb}, 1\text{Mb}\}$). These runs were completed across a distributed computing platform with machines of varying computing power. We plot the maximum resident set size and runtime against input matrix size in **Supp Figure A.17**. We observe that, in concordance with the computational complexity, time consumption increases in a quadratic fashion with respect to the input matrix size and in a linear fashion to k . Memory consumption increases in a similar manner, i.e. if the input matrix size doubles, the memory requirement approximately quadruples.

Stability and initialization of NMF

The NMF algorithm is commonly initialized with random non-negative values for the entries of U and V . The initial values can significantly impact the final values of U and V (Belford et al., 2018). This leads to instability of the final factors hinging on the randomization schemes or changing seeds. To address the instability, we used Non-Negative

Double Singular Value Decomposition (NNDsVD), which initializes U and V with a sparse SVD approximation of the input matrix X (Boutsidis and Gallopoulos, 2008). Since the derivation of exact singular values can considerably slow down the initialization step, we use a randomized SVD algorithm which derives approximate singular vectors (Voronin and Martinsson, 2015). NNDsVD initialization with randomized SVD results in lower loss, i.e. factors that can better approximate the original Hi-C matrix, in fewer iterations (**Supp Figure A.18A,B**), and more stable results than direct random initialization (**Supp Figure A.18C,D**).

2.3.2 Datasets used in experiments and analysis

High-throughput chromosome conformation capture datasets

We applied GRiNCH to interaction count matrices from *in situ* Hi-C (with MboI as the restriction enzyme) for five cell lines, GM12878, NHEK, HMEC, HUVEC, and K562 at 10kb, 25kb, and 50kb resolution (Rao et al., 2014, GEO accession: GSE63525). From the same source, we additionally used GM12878 25kb-resolution data from *in situ* Hi-C using DpnII as the restriction enzyme, and GM12878 25kb-resolution data from dilution Hi-C using HindIII as the restriction enzyme in our analysis on smoothing.

To demonstrate the applicability of GRiNCH to multiple high-throughput chromosome conformation capture platforms, we applied GRiNCH to datasets from other technologies that capture the 3D genome structure and chromatin interactions: Split-Pool Recognition of Interactions by Tag Extension (SPRITE) (Quinodoz et al., 2018) and HiChIP (Mumbach et al., 2016). We used the SPRITE data for GM12878 cell line (GEO accession: GSE114242). For HiChIP, we applied GRiNCH to the contact matrices from cohesin HiChIP (GEO accession: GSE80820, Mumbach et al., 2016) and H3k27ac HiChIP (GEO accession: GSE101498, Mumbach et al., 2017).

To demonstrate the utility of GRiNCH to study 3D genome organization in dynamic processes we applied GRiNCH to two different mouse developmental time course data: (a) neural differentiation Hi-C data from embryonic stem cells (mESC), neural progenitors (NPC), and cortical neurons (CN) (GEO accession: GSE96107, Bonev et al., 2017) and (b) Hi-C data from reprogramming pre-B cells to induced pluripotent state (Stadhouders et al., 2018, GEO accession: GSE96553). For (a) neural differentiation dataset, Juicer Straw tool (Durand et al., 2016) was used to obtain 25kb Hi-C matrices with vanilla-coverage square-root normalization. For (b) reprogramming, we applied GRiNCH to published normalized Hi-C data from pre-B cells, B α cells, day 2, day 4, day 6, day 8 of reprogramming, and finally, pluripotent cells.

ChIP-seq, DNase-seq, ATAC-seq, and motif datasets

To interpret the GRiNCH results and for comparison to other methods, we obtained a number of ChIP-seq datasets. For CTCF, ChIP-seq narrow-peak datasets available as ENCODE Uniform TFBS composite track (Rosenbloom et al., 2013) were downloaded from the UCSC genome browser (wgEncodeEH000029, wgEncodeEH000075, wgEncodeEH000054, wgEncodeEH000042, wgEncodeEH000063).

As ChIP-seq data for SMC3 and RAD21 are not available in the five cell lines from Rao et al., 2014, we generated a list of cell-line specific accessible motif sites. Accessible motif sites were defined as the intersection of motif match regions and DNase-accessible regions in the given cell line. The SMC3 and RAD21 motif matches to the human genome (hg19) was obtained from Kheradpour and Kellis, 2014. To create a union of DNase accessible regions from replicates within a cell line, BEDtools (Quinlan and Hall, 2010) merge program was used. Finally, the intersection of DNase accessible regions and motif match regions was calculated for each cell line using BEDtools intersect program. DNase accessibility sites were obtained from the ENCODE consortium (ENCODE Project Consortium, 2012; Sloan et al.,

2016): ENCFF856MFN, ENCFF235KUD, ENCFF491BOT, ENCFF946QPV, ENCFF968KGT, ENCFF541JWD, ENCFF978UNU, ENCFF297CKS, ENCFF569UYX.

We obtained ChIP-seq datasets for histone modification marks from the ENCODE consortium (ENCODE Project Consortium, 2012; Sloan et al., 2016). To generate genome-wide histone modification levels for each mark, fastq reads were aligned to the human genome (hg19) with bowtie2 (Langmead and Salzberg, 2012), and aggregated into a base-pair signal coverage profile using SAMtools (Li et al., 2009), and BEDtools (Quinlan and Hall, 2010). The base-pair signal coverage was averaged within each 25kb bin to match the resolution of Hi-C dataset. The aggregated signal was normalized by sequencing depth within each replicate; the replicates were collapsed into a single value by taking the median.

In order to identify novel transcription factors that could play a role in 3D genome organization, we obtained motifs of 746 different transcription factors from JASPAR core vertebrate collection (Fornes et al., 2020). Next, we obtained their accessible motif match sites in hg19 for the five cell lines from (Rao et al., 2014) using the same process that was used for SMC3 and RAD21 motifs. To identify the accessible motif sites for mouse cells during pluripotency reprogramming (Stadhouders et al., 2018), we aligned ATAC-seq fastq reads to the mouse genome (mm10) with bowtie2 (Langmead and Salzberg, 2012) and deduplicated with SAMtools (Li et al., 2009). Accessible peaks were called with MACS2 (Zhang et al., 2008). The ATAC-seq peaks were then used in place of DNase-seq sites to find the accessible motif sites as was done for SMC3 and RAD21 motifs.

2.3.3 TAD-calling methods

GRiNCH was benchmarked against 7 other TAD-calling methods: Directionality Index method (Dixon et al., 2012), Armatus (Filippova et al., 2014), Insulation Score method (Crane et al., 2015), rGMAP (Yu et al., 2017), 3DNetMod (Norton et al., 2018), HiCseg

(Lévy-Leduc et al., 2014) and TopDom (Shin et al., 2016). For all methods, default or recommended parameters values were used when available.

Directionality index

Directionality index uses a hidden Markov model (HMM) on estimated Directionality Index (DI) scores. The DI score for a genomic region is determined by whether the region preferentially interacts with upstream or with downstream regions. A bin can take on one of three states (upstream-biased, downstream-biased, or not biased) based on the interaction profile within a fixed-sized (e.g. 2Mb) window up- and downstream of the bin, with directionally biased bins becoming TAD boundaries. TADs were called using the directionality index method implementation in TADtool (Kruse et al., 2016), version as of April 23, 2018.

Armatus

Armatus uses dynamic programming to find subgraphs in a network where the nodes are the genomic regions, and the edge weights are the interaction counts. The objective is to find the set of dense subgraphs; subgraph density is defined as the ratio of the sum of edge weights to the number of nodes within the subgraph. Armatus predicts a set of overlapping TADs then consolidates them into consensus TADs. The consensus TADs were used in our analysis. Armatus version 2.3 was used for comparison.

Insulation score

In the insulation score method, each bin is assigned an insulation score, calculated as the mean of the interaction counts in the window (of a predefined size) centered on the given bin. Bins corresponding to the local minima in the vector formed by these insulation scores

are treated as TAD boundaries. TADtool (Kruse et al., 2016) implementation of insulation score method, version as of April 23, 2018, was used in our experiments.

3DNetMod

3DNetMod employs a Louvain-like algorithm to partition a network of genomic regions into communities where the edge weights in the network are the interaction counts. It uses greedy dynamic programming to maximize modularity, a metric of network structure measuring the density of intra-community edges compared to random distribution of links between nodes. 3DNetMod outputs a set of overlapping TADs. It was excluded from any analysis that required a unique TAD assignment for each genomic region or involved TAD shuffling. Software version 1.0 (10/06/17) was used in our comparison.

rGMAP

rGMAP trains a two-component Gaussian mixture model to group interactions into intra-domain or inter-domain contacts. Putative TAD boundary bins are identified by those with significantly higher intra-domain counts in its upstream window or downstream window of predefined size. The chromosome is then segmented into TADs flanked by these boundaries. rGMAP outputs a set of hierarchical, overlapping domains and a set of non-overlapping TADs; we used the latter in our analysis. Software version as of April 23, 2018 was used for comparison.

HiCseg

HiCseg treats the Hi-C matrix as a 2D image to be segmented, with each block-diagonal segment corresponding to a TAD. The counts within each block are modeled to be drawn from a certain distribution (e.g. Gaussian distribution for normalized Hi-C data). Using dynamic programming, HiCseg finds a set of block boundaries that would maximize the

log likelihood of counts in each block being drawn from an estimated distribution. Version 1.1 was used in our experiments.

TopDom

TopDom generates a score for each bin along the chromosome, where the score is the mean interaction count between the given bin and a set of upstream and downstream neighbors (neighborhood size is a user-specified parameter). Putative TAD boundaries are picked from a set of bins whose score forms a local minimum; false positive boundaries are filtered out with a significance test. Version 0.0.2 was used in our analysis.

2.3.4 TAD evaluation criteria

We evaluated the quality of TADs using different enrichment metrics as well as internal validation metrics used for comparing clustering algorithms.

Enrichment analysis

Enrichment of known architectural proteins. We estimated the enrichment of three known architectural proteins (CTCF, RAD21 and SMC3) in the TAD boundaries of five cell lines from Rao et al., 2014. TAD boundaries are defined by the starting bin and the ending bin of each predicted TAD, along with one preceding the starting bin and one following the ending bin. Let N be the total number of bins in a chromosome, n_{BIND} be the number of bins with one or more ChIP-seq peaks or accessible motif sites, n_{TAD} be the number of TAD boundary bins, and $n_{\text{TAD-BIND}}$ be the number of TAD-boundary bins with a binding event (ChIP-seq peak or accessible motif match site). The fold enrichment for a particular protein is calculated as: $\frac{n_{\text{TAD-BIND}}/n_{\text{TAD}}}{n_{\text{BIND}}/N}$. Within each cell line, the fold enrichment

across all chromosomes was averaged; then the mean across cell lines was used to rank the TAD-calling methods (**Supp Table A.1H**).

Histone modification enrichment. We used the proportion of predicted TADs that are significantly enriched in histone modification signals (compared to the “null” histone-modification signal distribution of randomly shuffled TADs) as a validation metric to assess the quality of TADs, similar to Zufferey et al., 2018. For each predicted TAD, we calculated the mean histone modification ChIP-seq signal within the TAD. Next, we find the “null” histone-modification signal distribution from randomly shuffled TADs. To generate randomly shuffled TADs, we take the lengths of all predicted TADs within a chromosome, as well as the lengths of interspersed stretches between the TADs (i.e. “non-TAD” stretches) if a TAD-calling method skips over regions of the genome. Next, we randomly move around the TAD and non-TAD stretches within the chromosome to preserve the TAD length distribution. We repeat this procedure 10 times. Then we compute the mean histone modification ChIP-seq signal within each of these randomly shuffled TADs, generating the null or background distribution of histone modification signals. The empirical p-value of a predicted TAD’s histone modification signal was calculated as the proportion of randomly shuffled TADs with higher ChIP-seq signal than that of the given TAD. A TAD was considered significantly enriched if its empirical p-value was less than 0.05, i.e. more than 95% of randomly shuffled TADs had a lower histone modification signal. Finally, for each TAD-calling method, we found the proportion of predicted TADs with significant histone modification signal; this is visualized across cell lines in **Figure 2.5B**. The mean proportion of TADs with significant enrichment across chromosomes and cell lines was used to rank the TAD-calling methods (**Supp Table A.1I**).

Internal validation metrics

Since a TAD represents a cluster of contiguous regions that tend to interact more among each other than with regions from another TAD or cluster, we used two internal validation or cluster quality metrics, Davies-Bouldin Index (DBI) and Delta contact count (DCC), to evaluate the similarity of interaction profiles among regions within a TAD. Specifically, for each method, we generated a background/null distribution of DBI and DCC from randomly shuffled TADs, then measured the proportion of actual TADs called with significant DBI and DCC level (p-value <0.05) against this null distribution (similar procedure to “Histone modification enrichment” above).

Davies-Bouldin Index (DBI). The DBI for a single cluster C_i is defined as its similarity to its closest cluster C_j , where $i, j \in \{1, \dots, k\}, i \neq j$: $DBI_i = \max_{i \neq j} S_{ij}$. The similarity metric, S_{ij} , between C_i and C_j is defined as:

$$S_{ij} = \frac{d_i + d_j}{\text{distance}_{ij}} \quad (2.5)$$

where d_i is the average distance between each data point in cluster C_i and the cluster centroid and distance_{ij} is the distance between the cluster centroids of C_i and C_j . In applying DBI to Hi-C data, a data point consists of a vector of a genomic region’s interaction counts with other regions in the chromosome (e.g. an entire row or column in the Hi-C matrix); a cluster corresponds to a group of regions within the same TAD; the cluster centroid is a mean vector of rows that belong to the same cluster/TAD. The smaller the DBI, the more distinct the clusters are from one another.

For each method, we computed the DBI of each individual TAD. To measure the significance of a TAD’s DBI value, we generated a background/null distribution of DBI values from randomly shuffled TADs (refer to the procedure in “Histone modification enrichment”

above). The empirical p-value of a TAD was calculated as the proportion of randomized TADs with lower DBI (recall a lower DBI means better clustering) than that of the given TAD. A TAD was considered to have a significant DBI if its empirical p-value was less than 0.05; the proportion of TADs with significant DBI was calculated for each method and used for comparing different TAD calling methods (**Supp Table A.1A**).

Delta Contact Count (DCC). DCC for cluster C_i is defined as follows: let in_i denote the mean interaction counts between pairs of regions that are both in C_i , and out_i denote the mean interaction counts between pairs of regions where one region is in cluster C_i and the other region is not. Then $DCC_i = in_i - out_i$.

We expect that for a good cluster, the pairs of regions within the cluster should have higher contact counts. Therefore, the higher the value of DCC, the higher the quality of the cluster. Again, a cluster corresponds to a group of regions within the same TAD. Given the DCC values for each TAD, we determined its significance against the null/background distribution of DCC values from randomly shuffled TADs (refer to procedure in "Histone modification enrichment" and "Davies-Bouldin Index (DBI)" above). The mean proportion of TADs with significant DCC across cell lines was used to compare the TAD-calling methods (**Supp Table A.1B**).

TAD similarity and stability metrics

When assessing the similarity or stability of TADs, we used two cluster comparison metrics, Rand Index and Mutual Information. First, TADs were converted to clusters so that regions in the same TAD were all assigned to the same cluster; all non-TAD regions, if a TAD-calling algorithm should have them, were assigned to a single cluster together. When comparing TADs across different resolutions of Hi-C data, 10kb, 25kb, and 50kb bins were split into a size of lowest common denominator, i.e., 5kb. Then all 5kb bins were assigned to the

same cluster as in the original lower-resolution bin (e.g. a 10kb bin assigned to cluster i would yield two 5kb bins assigned to cluster i). For these comparisons, we computed these metrics at the 5kb resolution.

For Rand Index, each genomic region is treated as a node in a graph; two nodes are connected by an edge if they are in the same cluster. Then, the number of edges that were preserved between clustering result A and clustering result B is divided by the total number of pairs of nodes, i.e. number of edges in a fully connected graph. Rand Index of 1 corresponds to perfect concordance between two clustering results; Rand Index of 0 means no agreement.

Mutual Information (MI) is an information-theoretic metric measuring the dependency between two random variables, where each variable can be a clustering result. Specifically, for two discrete variables A and B , MI is defined as

$$MI(A; B) = \sum_{a \in A} \sum_{b \in B} p_{(A,B)}(a, b) \log \left(\frac{p_{(A,B)}(a, b)}{p_A(a)p_B(b)} \right) \quad (2.6)$$

For clustering comparisons, A and B are cluster assignments to be compared, e.g., A is the cluster assignment corresponding to TADs from high-depth data and B is the cluster assignment based on TADs from downsampled data. Mutual Information is 0 if the joint distribution of A and B equals the product of each marginal distribution, i.e. A and B are independent, or in an information-theoretic sense, knowing A does not provide any information about B . The higher the Mutual Information value, the greater the information conveyed by the variables about each other; in the context of measuring clustering agreement, one clustering result is similar to the other.

Both metrics were used to evaluate the stability of TADs across resolution and depth, the similarity of TADs from different TAD-calling methods, the recovery of TADs from smoothed Hi-C data, the similarity of TADs along the time-course data, and the consistency

of GRiNCH TADs from different 3D genome capturing technologies (e.g. SPRITE, HiChIP). To rank TAD-calling methods based on stability across resolution and depth, the mean Rand Index or Mutual Information across cell lines was used (**Supp Table A.1D,E,F,G**).

Robustness to low-depth data

To assess the robustness or stability of TADs to low-depth input data, the TADs from a high-depth dataset (GM12878, Rao et al., 2014) were compared to the TADs from a downsampled, low-depth dataset. If the the TADs from the downsampled data are similar to TADs from the high-depth dataset, they are considered to be stable to low depth. The similarity metrics, Mutual Information and Rand Index described in the “TAD similarity and stability metrics” section, were used.

In order to downsample a high-depth Hi-C matrix (e.g. from GM12878) to a lower depth one (e.g. from HMEC), a distance-stratified approach was used to match both the mean of non-zero counts and sparsity level between the two datasets. First, for each distance threshold d , let μ_d^h denote the mean of the non-zero counts in the high-depth dataset and μ_d^l denote the mean of non-zero counts in the low-depth dataset. The scaled down value for each non-zero entry of the original high-depth dataset is: $\tilde{x}_{ij} = \frac{x_{ij}^h}{\mu_d^h/\mu_d^l}$. where x_{ij}^h is the value for the i,j bin pair in the high-depth dataset. Then, to increase the sparsity of the high-depth dataset, z_d of the non-zero counts in the high-depth dataset at distance d is randomly set to zero, where z_d is the number of additional entries in the low-depth dataset that are zero compared to the high-depth dataset.

2.3.5 Identification of candidate genomic regions involved in 3D organization changes during mouse neural development

To identify genomic regions potentially involved in local topological changes during the mouse neural development, we took GRiNCH clusters from the Hi-C data of mouse embryonic stem cells (mESC), neural progenitors (NPC), and cortical neurons (CN, Bonev et al., 2017) and looked for cluster merges or splits across the time points. We first performed pairwise cluster matching between time points (e.g. mESC vs CN). For each pair of clusters from time point A (e.g., cluster i from mESC) and time point B (e.g., cluster j from CN), we calculated their overlap in genomic regions with Jaccard Index, i.e. the ratio of the size of the intersection (regions in both clusters) to the size of the union (regions in either cluster). We then considered clusters matched to two or more clusters in another time point with a Jaccard Index of at least 0.2. For example, if cluster 5 from mESC matched to cluster 4, 5, and 6 from CN with Jaccard Index of 0.3, 0.25, and 0.4, respectively, then we considered cluster 5 in mESC as a site of potential topological changes, identified by cluster splits in CN. We selected a random subset of these clusters from different chromosomes, and visualized the interaction profile of the regions belonging to these clusters. The regions and clusters visualized in **Figure 2.7B** and **Supp Figure A.8** are from this subset.

2.3.6 Identification of novel factor enrichment at GRiNCH TAD boundaries

A similar procedure to CTCF boundary enrichment was used to identify novel boundary elements, by assessing whether the accessible motif sites of 746 transcription factors (TFs) from the JASPAR core vertebrate collection (Fornes et al., 2020) are enriched in GRiNCH TAD boundaries. This procedure was applied to the five cell lines from Rao et al., 2014 and the time points from the mouse reprogramming timecourse data (Stadhouders et al.,

2018). One change to the procedure was that instead of calculating fold enrichment per chromosome, all counts were aggregated across all chromosomes within the given cell line or time point. The hypergeometric test was used to calculate the significance of the number of TF sites in the boundaries and were ranked based on their p-value.

2.3.7 Smoothing methods

Smoothing with GRiNCH via matrix completion GRiNCH smooths a noisy input Hi-C matrix by using the matrix completion aspect of NMF. Specifically, the reconstructed matrix $X^s = UV^T$ is the smoothed matrix. The effectiveness of GRiNCH matrix completion as a smoothing method was compared to that of mean filter and Gaussian filter, two methods used in image blurring (Davies, 2004) and Hi-C data pre-processing (Yang et al., 2017; Rowley et al., 2020; Ardakany et al., 2019), as well as HiCNN (Liu and Wang, 2019), a method based on convolutional neural network to impute interaction counts.

Mean filter

Mean filtering is used in HiCRep (Yang et al., 2017) as a preprocessing step to measure reproducibility of Hi-C datasets. To create a smoothed matrix X^s from a given input matrix X with a mean filter, each element in x_{ij}^s is estimated from the mean of its neighboring elements within radius r : $x_{ij}^s = \frac{1}{(2r+1)^2} \sum_{a=i-r}^{i+r} \sum_{b=j-r}^{j+r} x_{ab}$. Three different values for the radius r were considered: $r \in \{3, 6, 11\}$.

Gaussian filter

A Gaussian filter has been used as a preprocessing step to identify chromatin loops and differential interactions from Hi-C Data (Rowley et al., 2020; Ardakany et al., 2019). It uses a weighted mean of the neighborhood of a particular contact count entry, x_{ij} , where the

weight is determined by the distance of the neighbor from the given position:

$$x_{ij}^s = \frac{1}{2\pi\sigma^2} \sum_{a=i-n}^{i+n} \sum_{b=j-n}^{j+n} e^{-\frac{(i-a)^2 + (j-b)^2}{2\sigma^2}} x_{ab} \quad (2.7)$$

Three different values of (σ) were considered, $\sigma \in \{1, 2, 3\}$ and n was set to $4 * \sigma$.

HiCNN

Unlike mean filter, Gaussian filter, and GRiNCH, HiCNN (Liu and Wang, 2019) uses supervised learning to perform smoothing. HiCNN uses a 54-layer convolutional neural network trained to predict high-resolution Hi-C interaction matrices from downsampled lower-resolution matrices. We downloaded three pre-trained models (from dna.cs.miami.edu/HiCNN along with source code) which were trained on GM12878 Hi-C data downsampled to 1/8, 1/16, and 1/25 depth of the original data, respectively. We used these pre-trained models in the smoothing analysis. These models were trained on interactions <2Mb apart and only make predictions for interaction distances <2Mb. To accommodate this limitation, AUPR on significant interaction recovery was measured separately for interactions <2Mb apart (see “Assessment of benefits from smoothing” below). Measuring TAD recovery after smoothing was not affected since the Directionality Index method uses a 2Mb-sized window of interactions (see “Directionality index” above).

2.3.8 Assessment of benefits from smoothing

Recovery of TADs from smoothed downsampled data

To assess whether smoothing helps preserve or recover structure in low-depth data, we first smoothed downsampled low-depth datasets (see “Robustness to low-depth data”) using methods described above (see “Smoothing methods”). The Directionality Index (DI)

TAD finding method was applied to the high and low-depth datasets. Then the similarity of the TADs from the original high depth data and the TADs from the smoothed data were measured (see “TAD similarity and stability metrics”). Higher similarity metric values imply better recovery of structure from smoothing.

Recovery of significant interactions

Fit-Hi-C (Ay et al., 2014) was used to call significant interactions in the original and the smoothed Hi-C datasets, using a $q\text{-value} < 0.05$. Interactions from the original high-depth Hi-C dataset were used as the set of “true” significant interactions. From the downsampled then smoothed matrices, each smoothed interaction count was assigned a “prediction score” of $1 - q$, where q is its Fit-Hi-C q -value. Precision and recall curves were then computed using the “true” interactions and the “prediction scores.” The recovery of true significant interactions was measured with the Area under the Precision-Recall curve (AUPR).

Robustness to different restriction enzymes

In the smoothing analysis of data from Hi-C protocols using different restriction enzymes (HindIII, DpnII, MboI), the overlap of significant interactions was measured with Jaccard Index, which is the ratio of the size of the intersection (i.e. significant interactions called in both datasets compared) to the size of the union (i.e. significant interactions called in either one of the datasets).

2.4 Implementation and availability

GRiNCH source code (in C++), installation instructions (supported in Linux distributions), documentation, and tutorial for visualization (in Python) are publicly available at roy-lab.github.io/grinch with a GNU General Public License (GPL-3.0). The specific version

of GRiNCH used in our experiments and analyses (v1.0.0) has been deposited with DOI 10.5281/zenodo.4540608, along with the following groups of files which were too large to include in the manuscript as supplementary materials:

- Execution scripts containing the parameter values used for benchmarked TAD-calling methods
- Scripts used to analyze the results and generate the figures
- Scripts and files specifically used to generate rankings of TAD-calling methods

2.5 Discussion

We present GRiNCH, a graph-regularized matrix factorization framework that enables reliable identification of high-quality genome organizational units, such as TADs, from high-throughput chromosome conformation capture datasets. GRiNCH is based on a novel constrained matrix factorization and clustering approach that enables recovery of contiguous blocks of genomic regions sharing similar interaction patterns as well as smoothing sparse input datasets.

A lack of gold standards for TADs emphasizes the need to probe both the statistical and biological nature of inferred TADs. Through extensive comparison of GRiNCH to existing methods with good performance in other benchmarking studies, we identified strengths and weaknesses of existing approaches. In particular, methods like Directionality Index and Insulation Score identify TADs that are generally more enriched for signals such as CTCF and cohesin. However, when comparing statistical properties such as stability across resolutions and cluster coherence, these methods do not necessarily perform better. GRiNCH was among the top methods for both criteria, identifying clusters of genomic regions with high degree of similarity in their interaction profiles, stable to low-depth, sparse datasets, and enriched in architectural proteins and histone modification signals with known roles in chromatin organization.

A unique advantage of GRiNCH lies in its smoothing capability via matrix completion. Smoothing has been an independent task from TAD-calling and a key processing step in downstream analysis of Hi-C data (e.g. measuring reproducibility or concordance between Hi-C replicates, Ursu et al., 2018). We find that GRiNCH smoothing outperforms existing unsupervised smoothing methods (mean filter and Gaussian filter) and comparable to supervised models trained on low-depth datasets in its ability to retain TAD-level and interaction-level features of the input Hi-C data. Furthermore, GRiNCH is applicable to

datasets from a wide variety of platforms, including SPRITE and HiChIP. Application of GRiNCH shows that Hi-C and HiChIP datasets capture more similar topological units than SPRITE. Interestingly, TADs from Hi-C and cohesin HiChIP are much closer than the two HiChIP datasets we compared. This shows that GRiNCH is capturing TADs that are reproducible across platforms. To study the ability of GRiNCH to identify dynamic topological changes along a time course, we applied GRiNCH to published developmental time-course datasets. GRiNCH recapitulated global temporal relationships in 3D organization and also transitions in topological units around previously studied and new genomic loci. Thus, GRiNCH should be broadly applicable for analysis of chromosome conformation capture datasets with different experimental design, sequencing depths, and platforms.

The 3D organization of the genome is determined through a complex interplay of architectural proteins such as CTCF, cohesin elements, and other transcription factors such as WAPL (Haarhuis et al., 2017). Application of GRiNCH to Hi-C datasets representing cell lines and temporally related conditions identified known and novel transcription factors that could be important for establishing these boundaries in a cell-type-specific or generic manner. In particular, we recovered YY1/2 proteins that have been shown to interact with CTCF to establish long-range regulatory programs during lineage commitment (Beagan et al., 2017). Among the novel factors that were present in both the cell lines as well as the mouse reprogramming dataset, were several zinc finger proteins, e.g. PLAGL1, ZIC1, ZIC4/5, ZBTB14; such proteins can be investigated for their role in establishing organizational units in mammalian genomes. We also found several factors that were specific to cell lines and time points. For example, FOXI1, a forkhead protein, was ranked highly in K562. Forkhead proteins are involved in genome organization and replication timing in yeast (Knott et al., 2012) and zebra fish (Yan et al., 2006), but their role in mammalian genome organization is not well known. The time course data identified additional unique TFs that are likely involved in determining specific lineages, e.g. STAT3,

MEIS3, FOXP3 and HOX genes in pre-B cells. HOX genes (Alharbi et al., 2013), FOXP3 (Li et al., 2015), and STAT3 (Chou et al., 2006) in particular have been shown to play critical roles in B cell and T cell development. While MEIS1 and MEIS2 are involved in the hematopoietic lineage, MEIS3 specifically is involved in the binding of HOX TFs to target genes in the brain (Uribe and Bronner, 2015). Therefore the simultaneous enrichment of MEIS3 and HOX sites is consistent with HOX proteins requiring MEIS3 for binding; however, its specific role in the hematopoietic lineage is yet unknown. Investigating the interactions of these proteins with well-known architectural proteins such as CTCF and cohesin could provide mechanistic insight into the factors governing 3D genome organization (Cubeñas-Potts and Corces, 2015; de Wit, 2019).

There are several directions of future work that are natural extensions to our framework. Although our current approach of analyzing temporal organization in time-course data extracted interesting biological insights, TADs are identified independently for each time point, making it difficult to study the conservation and specificity of individual TADs. One area of future work is to allow joint identification of TADs or similar structural units across multiple conditions (Fotuhi Siahpirani et al., 2016; Yang et al., 2019). GRiNCH currently infers one level of TADs for a given input set of parameters. Expanding GRiNCH to provide nested or hierarchical TADs is an additional direction of future work. Another direction is to leverage one-dimensional features to potentially inform the TAD-finding algorithm. The GRiNCH framework makes use of a distance dependence graph of regions; however, one could use the similarity of epigenomic profiles to construct an additional graph to constrain the NMF solution.

2.6 Conclusion

GRiNCH offers a unified solution, applicable to diverse platforms, to discover reliable and biologically meaningful topological units, while handling sparse high-throughput chromosome conformation capture datasets. The outputs from GRiNCH applied to time course datasets can be used to study changes in 3D genome organization and predict novel boundary elements, enabling us to test possible hypotheses of other mechanisms for TAD boundary formation. We have made GRiNCH publicly available at roy-lab.github.io/grinch with a GNU General Public License (GPL) and a comprehensive installation and usage manual. As efforts to map the three-dimensional genome organization expand to more conditions, platforms, and species, a method such as GRiNCH will serve as a powerful analytical tool for understanding the role of 3D genome organization in diverse complex processes.

2.7 Acknowledgments

This work is supported by the National Institutes of Health (NIH) through the grant NHGRI R01-HG010045-01. We thank Shilu Zhang and Alireza Fotuhi Siahpirani for providing scripts for data processing and interpretation of results. We also thank the Center for High Throughput Computing at University of Wisconsin - Madison for computational resources.

Chapter 3

Examining dynamics of three-dimensional genome organization with multi-task matrix factorization

Three-dimensional (3D) genome organization, which determines how the DNA is packaged inside the nucleus, has emerged as a key component of the gene regulation machinery. High-throughput chromosome conformation datasets, such as Hi-C, have become available across multiple conditions and timepoints, offering a unique opportunity to examine changes in 3D genome organization and link them to phenotypic changes in normal and diseases processes. However, systematic detection of higher-order structural changes across multiple Hi-C datasets remains a major challenge. Existing computational methods either do not model higher-order structural units or cannot model dynamics across more than two conditions of interest. We address these limitations with Tree-Guided Integrated Factorization (TGIF), a generalizable multi-task Non-negative Matrix Factorization (NMF) approach that can be applied to time series or hierarchically related biological conditions. TGIF can identify large-scale changes at compartment or subcompartment levels, as well

as local changes at boundaries of topologically associated domains (TADs). Compared to existing methods, TGIF boundaries are more enriched in CTCF and reproducible across biological replicates, normalization methods, depths, and resolutions. Application to three multi-sample mammalian datasets shows TGIF can detect differential regions at compartment, subcompartment, and boundary levels that are associated with significant changes in regulatory signals and gene expression enriched in tissue-specific processes. Finally, we leverage TGIF boundaries to prioritize sequence variants for multiple phenotypes from the NHGRI GWAS catalog. Taken together, TGIF is a flexible tool to examine 3D genome organization dynamics across disease and developmental processes.

A version of this work is available as a preprint:

Lee DI and Roy S. 2024. Examining dynamics of three-dimensional genome organization with multi-task matrix factorization

3.1 Introduction

The three-dimensional (3D) organization of the genome refers to the packaging of DNA inside the nucleus. It has emerged as a key regulatory mechanism of cellular function and dysfunction across diverse developmental (Zheng and Xie, 2019), disease (Lupiáñez et al., 2016), and evolutionary contexts (McCord, 2017; Eres et al., 2019). High-throughput chromosomal conformation capture (Hi-C) technologies enable the study of 3D genome organization by experimentally measuring the tendency of genomic regions to spatially interact with one another (Kempfer and Pombo, 2020; Mumbach et al., 2016; Dekker et al., 2023). The 3D genome is organized into structural units at multiple scales: compartments spanning several megabases, Topologically Associated Domains (TADs) spanning hundreds of kilobases scale, and enhancer-promoter loops involving pairs of loci of a few thousand bases. (Bouwman and de Laat, 2015; Rowley and Corces, 2018; Kempfer and

Pombo, 2020). Changes in 3D genome organization at different topological levels have been observed with transitions in both normal (Bonev et al., 2017; Stadhouders et al., 2018; Zheng and Xie, 2019) and disease processes (Lupiáñez et al., 2016; Norton and Phillips-Cremins, 2017; Wang et al., 2021). For example, during differentiation of mouse embryonic stem cells to a neuronal lineage, changes in topological structure are associated with cell fate specification and gene expression changes (Bonev et al., 2017). Changes in 3D genome organization have also been seen in immune response to viral infections (Wang et al., 2021) and diseases such as cancer (Hnisz et al., 2016; Akdemir et al., 2020; Dubois et al., 2022). Through efforts from large-scale consortia such as the 4D Nucleome project, Hi-C measurements are becoming increasingly common from multiple conditions corresponding to time points, cell types and species (Dekker et al., 2017; Reiff et al., 2022; Dekker et al., 2023; Roy et al., 2023). These datasets provide a unique opportunity to examine the dynamics of 3D genome organization across space and time and its impact on disease and normal processes.

Reliable detection of 3D genome dynamics at different units of organization is a significant computational challenge. Current computational approaches to examine dynamics in the 3D genome can be grouped into those that identify large-scale or compartmental-level changes (Fotuhi Siahpirani et al., 2016; Chakraborty et al., 2022), those that can identify TAD-scale changes or “differential TADs” (Wang et al., 2020; Cresswell and Dozmorov, 2020), and those that examine changes at the level of loops or interactions (Ardakany et al., 2019; Lun and Smyth, 2015; Djekidel et al., 2018; Galan et al., 2020; Stansfield et al., 2019). Compared to methods for detecting differences at the loop level, there are relatively few approaches to detect TAD or compartment changes. The most common approach to study TAD dynamics across multiple conditions is to first apply a TAD-calling method to data from each condition, followed by post-processing to identify TAD boundaries in one condition but not another (Zhang et al., 2019; Bonev et al., 2017; Stadhouders et al.,

2018; Wang et al., 2022; Emerson et al., 2022). While such a two-step approach can identify some meaningful differences, the unsupervised nature of TAD finding could make these approaches more susceptible to finding non-biological differences. Numerous studies have shown that Hi-C count profiles obey cell type, timepoint and species relationships, where datasets from nearby contexts are more similar than those that are far away (Bonev et al., 2017; Zhang et al., 2019; Yang et al., 2017; Vietri Rudan et al., 2015). An approach that constrains the TAD and compartment finding based on such prior information about the relationships between the input datasets could be less prone to spurious differences. A few methods have been developed to directly identify TAD boundary differences, but they are focused on pairs of conditions (Wang et al., 2020) or limited in their ability to compare more than two conditions (Cresswell and Dozmorov, 2020).

To address the dearth of methods for identifying large-scale organizational changes, especially when considering more than two datasets, we developed Tree-Guided Integrated Factorization (TGIF), a multi-task learning framework using Non-negative Matrix Factorization (NMF) to enable joint identification of organizational units such as compartments and TADs across multiple conditions. NMF is a popular dimensionality reduction approach that has been used for analyzing genomic data (Stein-O'Brien et al., 2018; Kotliar et al., 2019; Lee and Roy, 2021) as well as images (Lee and Seung, 1999; Kalayeh et al., 2014), where the low-dimensional factors can recover the major patterns in the data. In the case of Hi-C data, the output factors represent the lower-dimensional view of the chromosomal architecture. TGIF can take as input multiple Hi-C matrices from related biological conditions, for example, different time points, treatments, diseases or cell types. TGIF uses hierarchical multi-task learning to constrain the lower-dimensional factors from closely related tasks (e.g. consecutive time points) to be similar. We use the low dimensional factors for each of the conditions to identify changes at both the compartment and TAD levels.

We applied TGIF-DB and TGIF-DC to three different mammalian differentiation time-course Hi-C datasets: a 2 timepoint dataset comprising human pluripotent cell line H1 and differentiated endoderm (Reiff et al., 2022; Dekker et al., 2023), a 3 timepoint dataset of mouse neural differentiation (Bonev et al., 2017), and a 6 timepoint dataset collected during human cardiomyocyte differentiation (Zhang et al., 2019). Compared to existing approaches, TGIF-DB identifies fewer false-positive differences in simulated data. When applied to real Hi-C data, TGIF-DB boundaries are more enriched in CTCF binding and is less susceptible to false differential boundaries that could arise due to non-biological factors such as different replicates, different downsampling depths, normalization methods, and binning resolutions.

At the compartment level, TGIF-DC differential compartmental regions displays significant change in accessibility and gene expression, while recovering compartments of similar quality compared to competing methods, measured by chromatin accessibility and observed-over-expected interaction counts. TGIF-DC can additionally cluster genomic regions into the more granular subcompartments with distinct histone modification patterns. We used TGIF-DB boundaries to assess the extent to which differentially expressed genes localize near significantly differential boundaries. Finally, we use TGIF boundaries identified in cardiomyocyte differentiation data to interpret sequence variants identified from the NHGRI Genome Wide Association Studies (GWAS) catalog and identify cardiovascular disease associated SNPs to be enriched in TGIF boundaries. Together, these results demonstrate the versatility and utility of TGIF to examine changes in higher-order 3D genome organization across diverse types of dynamic processes.

3.2 Results

3.2.1 Tree-guided Integrated Factorization (TGIF) for examining dynamics in 3D genome organization

Tree-guided Integrated Factorization (TGIF) is a general-purpose framework to study 3D genome organization dynamics both at the TAD and compartment levels (**Figure 3.1**). TGIF is based on multi-task non-negative matrix factorization (NMF). It takes as input a set of Hi-C matrices, each representing a biological condition, and a user-specified tree structure that can encode an arbitrary relationship among the conditions, such as time or cell type lineage (**Figure 3.1, Supp Figure B.1**). TGIF uses a novel regularization term in its objective to jointly factorize the matrices such that input matrices from more closely related conditions result in more similar lower-dimensional representations, i.e., factors.

To handle both compartment and TAD identification, we implemented two versions of TGIF: TGIF-DB and TGIF-DC. TGIF-DB identifies conserved and differential boundaries demarcating TADs under different conditions (**Figure 3.1A, Supp Figure B.1A, Methods**), while TGIF-DC identifies compartment-level changes in 3D genome organization (**Figure 3.1B**). In TGIF-DB, the factorization is performed on sub-matrices along the diagonal of the intrachromosomal Hi-C matrices, as these diagonal sub-matrices capture the TAD-scale, local topology of chromosomes. Each sub-matrix is factorized over a range of k , the hyper-parameter specifying the rank of the lower-dimensional space (**Supp Figure B.1B**). TGIF-DB calculates a boundary score from the factors at each k , which are averaged to provide an overall boundary score (**Supp Figure B.1C**). Considering multiple k allows us to capture structural units or domains of different sizes in the lower dimensional space and removes the need to specify the number of factors (**Methods**). TGIF-DB identifies regions with significant boundary scores by comparing the average boundary scores against a “null

distribution" of boundary scores to calculate an empirical p-value (**Supp Figure B.1D**). TGIF-DB outputs the list of significant boundaries corresponding to each input dataset and a list of significantly differential boundary regions for every pair of input count matrices (**Supp Figure B.1E, Methods**).

TGIF-DC operates at the entire chromosome level and applies its multi-task factorization on the observed-over-expected (O/E) counts matrix as described previously (Lieberman-Aiden et al., 2009; Rao et al., 2014, **Methods**). To identify the two major compartments of active and repressive genomic regions, TGIF-DC factorizes the O/E matrices with parameter $k = 2$. The resulting factors are used to group the genomic regions into 2 different clusters. By specifying a higher parameter value, e.g. $k = 5$, TGIF-DC can also identify more granular subcompartment structures, which can be interpreted using one-dimensional chromatin signals. Similar to TGIF-DB, TGIF-DC identifies significantly differential compartment and subcompartment regions for every pair of input conditions (**Methods**).

In cases where the relationship between the input Hi-C data is not available (e.g. integrating Hi-C datasets from multiple studies or pseudo-bulk single-cell Hi-C data from cell clusters; Zhou et al., 2019; Zhang et al., 2022). TGIF can infer a tree structure based on the pairwise similarity of the input Hi-C matrices measured by stratum-adjusted correlation coefficient (SCC; Yang et al., 2017, **Methods, Supp Figure B.2**) or a similar distance-stratified metric.

3.2.2 TGIF-DB identifies fewer false-positive differential boundaries in simulated and real Hi-C data.

TGIF-DB was benchmarked against four other TAD calling methods: three methods designed for calling TADs and boundaries from a single Hi-C matrix (which we refer to as single-task methods), and one designed specifically for differential boundary identifica-

tion (**Methods**). The three single-task methods were: (1) GRINCH (Lee and Roy, 2021), which uses NMF for TAD identification; (2) SpectralTAD (Cresswell et al., 2020), which also uses dimension reduction; and (3) TopDom (Shin et al., 2016), which uses changes in average contact frequencies upstream and downstream of a given genomic region to determine significant boundaries. TADCompare (Cresswell and Dozmorov, 2020) is a method designed for differential boundary detection and takes as input pairs of input Hi-C matrices and identifies non-differential and differential boundaries between them. Default or recommended parameters were used for all benchmarking experiments (**Methods**).

Since real Hi-C datasets do not have ground-truth set of TAD boundaries, we first evaluated the quality of TAD boundaries identified by each method in simulated datasets. We generated 4 Hi-C matrices each with its own set of ground-truth boundaries based on the count simulation procedure from Forcato et al., 2017 (**Methods**). For every pair of matrices, we calculate the precision on boundaries found only in one matrix ("task-specific" boundaries) and those shared between the two input matrices (shared boundaries). Across the different levels of noise added to the simulated matrices, TGIF-DB has the highest precision on task-specific boundaries (**Figure 3.2A**). With the exception in the lowest level of noise, TGIF-DB is among the methods with the highest precision for shared boundaries along with GRINCH and TopDom (**Figure 3.2A**). For recall of task-specific boundaries (**Supp Figure B.3A**), TGIF-DB is second to TopDom in all noise levels except the lowest. In shared-boundary recall, TGIF-DB is comparable to or outperforms TopDom as the top method again in all noise levels except the lowest. The better performance of TGIF-DB we observe with higher-noise input data is likely due to the fact that significant boundary calls are made against "background scores" which originated from shuffling the mean of the input matrices (**Methods**); if the input matrices have too little noise, the background scores will lack variation for empirical p-value calculation and significance calling.

Next, we evaluated the quality of boundaries from real Hi-C data based on the en-

richment of CTCF binding. CTCF is an architectural protein associated with establishing boundaries (Merkenschlager and Nora, 2016; Gómez-Díaz and Corces, 2014; Cubeñas-Potts and Corces, 2015). We used the time-series dataset of cardiomyocyte differentiation (Zhang et al., 2019) which profiled both genome-wide chromosome conformation with Hi-C and CTCF binding with ChIPseq. The single-task methods, GRiNCH, SpectralTAD, and TopDom, were applied to each of the six timepoints (day 0, 2, 5, 7, 15, 80) from the cardiomyocyte dataset independently. For TADCompare, we gave the algorithm pairs of consecutive timepoints (day 0 vs 2, 2 vs 5, 5 vs 7, 7 vs 15, 15 vs 80) to find non-differential and differential boundaries between each pair of timepoints (**Methods**). Finally, TGIF-DB was applied to all available timepoints together with an input tree structure capturing the temporal dependency (see **Supp Figure B.4** for input tree). The resulting boundaries for each timepoint was used for CTCF enrichment analysis. Fold enrichment of CTCF peaks in the boundary regions was calculated against the genomic background (**Methods**). We find that significant boundaries identified by TGIF-DB has the highest fold enrichment, followed by single-task methods, TopDom and GRiNCH (**Figure 3.2B**).

We next benchmarked the ability to detect true versus false boundary differences between Hi-C datasets, which could arise due to various non-biological reasons such as: (1) datasets from biological replicates, (2) datasets with different depths, (3) different normalization process, (4) different bin resolution. Since the underlying biological process is the same, any differences in boundaries between such datasets are considered as false positives.

We compared methods for their ability to recapitulate TAD structure across biological replicates (**Methods**) by applying them to the day 0 biological replicates of the cardiomyocyte differentiation dataset (Zhang et al., 2019), which represents the H1 ESC state. This time point is expected to have the lowest artificial differences as it is from a cell line and the process of differentiation can introduce additional sources of heterogeneity. We measured

the Jaccard coefficient between the boundaries identified in the pair of replicates. Here again TADCompare and TGIF-DB had the highest Jaccard coefficients recovering most similar set of boundaries across the biological replicates (**Figure 3.2C**).

To compare datasets with different read depths, we took a high-depth Hi-C dataset from the GM12878 cell line with 4.01 billion reads in total (Rao et al., 2014; Reiff et al., 2022) and subsampled 5, 10, 25, 50% of the reads to create downsampled versions of the data (**Methods**). As these datasets represent the same cell line, any differential boundaries can be considered as false positives resulting from the depth difference. We again applied the single-task methods (GRiNCH, SpectralTAD, TopDom) to each of the original and downsampled datasets independently. Differential-TAD methods (TADCompare and TGIF-DB) were applied to a pair of datasets, i.e. the original high-depth and a downsampled. We measured the Jaccard index between the boundaries identified in a pair; the higher the Jaccard index, the fewer the false-positive differences identified by a method. Across all downsampled depths, TADCompare and TGIF-DB were the top performing methods with consistently high Jaccard scores; TADCompare outperformed TGIF-DB at 10 and 25% depth, and TGIF-DB outperformed TADCompare at 50% (**Figure 3.2D**). Single-task methods (GRiNCH, SpectralTAD, and TopDom) had much lower Jaccard score with discrepancy increasing with depth differences.

We also compared the methods for their ability to recover reproducible TADs across different normalization methods: Iterative Correction and Eigenvector decomposition (ICED, Imakaev et al., 2012) and square root vanilla coverage (VCSQRT, Rao et al., 2014) on a mouse embryonic stem cell (mESC) Hi-C dataset (Bonev et al., 2017, **Methods**). We observed similar results with TADCompare and TGIF-DB obtaining the highest Jaccard coefficients between TAD boundaries across different normalizations (**Figure 3.2E**).

Finally, we measured the stability of TAD boundaries to the changing resolution (10kb, 25kb, 50kb) of input Hi-C matrices using Jaccard index (**Methods**). TGIF-DB and GRiNCH

yield the most stable or similar boundaries to changing resolution (**Supp Figure B.3B**), with the exception of 25kb-50kb where TopDom also performed well. Since both TGIF-DB and GRiNCH are based on non-negative matrix factorization, it is possible that NMF-based methods are less susceptible to changes in resolution of the dataset.

Taken together, these results demonstrate the advantages of using a multi-task matrix factorization method such as TGIF-DB to identify biologically relevant boundaries enriched in known boundary elements while minimizing false positive differences.

3.2.3 TGIF-DC identifies compartment dynamics that are significantly enriched for differential regulatory signals.

We compared the TGIF-DC compartments and differential compartments to three existing methods on intra-chromosomal count matrices at 100kb resolution from the H1 hESC and endoderm differentiation dataset. Two were single-task compartment calling methods: PCA-based (Lieberman-Aiden et al., 2009) and Cscore (Zheng and Zheng, 2018). The third method, dcHiC (Chakraborty et al., 2022), is a differential compartment identification method, and was applied in a pairwise comparison mode between H1 hESC and endoderm differentiated from H1 (**Methods**). TGIF-DC was applied to a simple tree with two leaf nodes as H1 and endoderm (**Methods, Supp Figure B.4**).

We first compared the similarity of compartment assignments in H1 hESC from different methods by treating the assignment as clusters and using Rand Index (**Figure 3.3A**). The PCA-based method and dcHiC, which also utilizes PCA, produced the most similar compartments as expected (Rand Index: 0.91), followed by TGIF-DC (Rand Index: 0.79-0.8). Cscore found a substantially different set of compartments (Rand Index: 0.52). We next assessed the quality of compartments by measuring three cluster quality metrics, Silhouette Index (SI, **Figure 3.3B**), Calinski-Harabasz score (CH, **Figure 3.3C**), and Davies-

Bouldin Index (DBI, **Figure 3.3D**), using observed-over-expected (O/E) counts as features of each genomic loci (**Methods**). In all three metrics, TGIF-DC, dcHiC and PCA-based compartments are comparable in their quality and outperformed Cscore.

We also assessed the compartment quality using chromatin accessibility, a key regulatory measurement that characterizes different compartment types (e.g., the active A and repressive B; Lieberman-Aiden et al., 2009; Fortin and Hansen, 2015), as the feature. Briefly, we used SI (**Figure 3.3E**), CH (**Figure 3.3F**) and DBI (**Figure 3.3G**) using the mean basepair ATACseq signal for each 100kb region (**Methods**). For all three metrics, the compartments from TGIF-DC, PCA-based method, and dcHiC are of similar quality and higher than Cscore.

Finally, we compared TGIF-DC exclusively with dcHiC, the only method among the three compared, that identifies *differential* compartment regions. Significantly differential compartmental regions (sigDC) identified by TGIF-DC have significantly higher change in accessibility signal and gene expression compared to regions not part of sigDC (**Supp Figure B.5A,B**). Compared to significantly differential regions identified by dcHiC, sigDC regions from TGIF-DC also have significantly higher change in accessibility signal (t -test p -value $< 1e-2$, **Figure 3.3H**) and are comparable in terms of the change in gene expression levels (**Figure 3.3I**).

3.2.4 TGIF-DC offers a unified framework to identify both compartment and subcompartment dynamics.

While compartments provide a global partitioning of each chromosome, the genome is hierarchically organized with compartments further partitioned into smaller subcompartments that could represent functionally distinct set of regions (Rao et al., 2014; Xiong and Ma, 2019). Unlike existing compartment finding methods that need additional clustering

steps to define subcompartment structure, TGIF-DC has a tunable parameter (k , the rank of NMF factors) that can be used to identify subcompartments within the same framework. TGIF-DC's low-rank dimensionality reduction framework lends itself naturally to identify this subcompartment structure. To demonstrate TGIF-DC's ability to identify both compartments and subcompartments, we applied it to the mouse neural differentiation dataset with 3 timepoints: embryonic stem cell or ES, neural progenitors or NPC, and cortical neurons, CN (**Supp Figure B.4, Supp Figure B.6, Supp Figure B.7**). This dataset additionally measured six different histone modification signals for NPC and CN that were beneficial for additional biological interpretation of TGIF-DC results (**Figure 3.4A, Methods**). We first analyzed the compartment structure from TGIF-DC ($k = 2$) for each chromosome, based on GC content (mean GC percentage for each 100kb bin, **Methods**), annotating the compartment with higher GC content as compartment A and the one with lower GC content as compartment B (**Methods, Supp Figure B.8**). Regions annotated as A compartment by TGIF-DC have significantly higher signal for marks associated with active enhancer (H3K27ac, H3K4me1) or elongation (H3K36me3) than those in B compartment (**Figure 3.4B**).

We next applied TGIF-DC with $k = 5$ to identify subcompartment structure per chromosome, each k corresponding to a different subcompartments (**Methods**). We interpreted these subcompartments based on the mean histone modification signal of the genomic loci assigned to each subcompartment. The subcompartments exhibited distinct histone modification patterns (**Figure 3.4C, chr18**), with subcompartments 1 and 5 associated with repressive marks (H3K9me3, H3K27me3), while the other three (2, 3 and 4) associated with active marks. Within these two groups, each subcompartment had a different signature of marks. For example, subcompartment 3 exhibits relatively lower signal of H3K36me3 compared to 2 and 4, while subcompartment 2 had a higher signal of all three activating marks (H3K27ac, H3K36me3, H3K4me1) compared to 3 and 4. Between the

two subcompartments, 1 and 5, with repressive mark association, one (1) exhibited higher H3K4me3 and H3K9me3 levels compared to the other one (5).

Finally, we assessed TGIF-DC's differential subcompartments by measuring the log fold change in histone modification signals between two timepoints, NPC and CN, and k-means clustering the regions based on this signal difference. We find distinct subgroups of regions with different fold change of the three activating marks, H3K27ac, H3K36me3, and H3K4me1 (**Figure 3.4D**). Interestingly, the repressive marks or the promoter specific mark, H3K4me3, did not vary substantially for these regions.

Taken together, these results demonstrate TGIF-DC's flexible framework to identify both compartment and subcompartment level dynamics that are associated with significant changes in regulatory activity between the timepoints or cell stages compared.

3.2.5 Changes in gene expression are associated with changes in boundaries during differentiation.

Untangling the relationship between 3D genome organization and gene expression remains a key question in regulatory genomics. While a direct mechanistic link between transcription and 3D genome organization has been observed (van Steensel and Furlong, 2019; Heinz et al., 2018) during cell state transitions (Pollex et al., 2024; Chen et al., 2024), other studies did not reveal changes in 3D genome organization to be a strong determinant of gene expression changes (Ing-Simmons et al., 2021; Espinola et al., 2021). To assess the extent to which changes in 3D genome structure are associated with changes in expression, we analyzed differential structural regions identified by TGIF-DC and TGIF-DB with differential gene expression in multiple mammalian differentiation datasets.

We applied TGIF-DC and TGIF-DB to the three different timecourse datasets with both Hi-C and RNAseq measurements (**Supp Figure B.4, Supp Table B.1-B.3**): (1) the two-

timepoint dataset of H1 hESCs differentiated to endoderm state (Reiff et al., 2022; Dekker et al., 2023, **Supp Figure B.5C-F, Supp Figure B.9**), (2) the three-timepoint of mouse neural differentiation time course from mESC to cortical neurons (CN, Bonev et al., 2017, **Supp Figure B.6, Supp Figure B.7, Supp Figure B.10**), and (3) the six-timepoint human cardiomyocyte differentiation timecourse from hESC to ventricular cardiomyocytes (Zhang et al., 2019, **Supp Figure B.11, Supp Figure B.12**). For each time course, we performed pairwise comparison of differential boundary, compartment, and expression between all pairs of time points, e.g. H1 vs endoderm, mESC vs NPC, day 0 vs day 2 of cardiomyocyte differentiation. For each pair of timepoints, we performed a region-centric and gene-centric analysis to ask whether differentially expressed (DE) genes are enriched in three different sets of dynamic regions (**Methods, Figure 3.5A**): (A) regions near (i.e., within 100kb) of significantly differential boundaries (sigDB), (B) regions within a TAD with at least one sigDB, and (C) regions within significantly differential compartmental regions (sigDC). For the region-centric analysis, we measured the fold enrichment of regions overlapping a DE gene among dynamic regions compared to all genomic regions. For the gene-centric analysis, we measured the fold enrichment of DE genes among genes overlapping a dynamic region compared to the proportion of DE genes among all genes. Dynamic regions in set A (within 100kb of sigDB) are consistently enriched for DE genes across all datasets and timepoints compared (**Figure 3.5B top, Supp Table B.5-B.7**). Furthermore, genes in set A are also enriched for DE genes compared to all genes (**Figure 3.5B bottom**). We did not see as much significant enrichment for regions in set B (within a TAD with at least one sigDB), likely because of the permissive inclusion criteria for set B. However, genes in set B were enriched for DE genes as well, although to a lower extent. When examining compartments, we found that regions in set C (within sigDC) are also significantly enriched in DE genes for the H1-endoderm differentiation and majority of the comparisons of the cardiomyocyte differentiation. The enrichment for genes was lower again due to the large number of genes

within compartments.

To assess the biological significance of DE genes near differential boundaries, we examined the biological processes enriched in DE genes near sigDBs compared to processes enriched in other genes (**Methods**). We grouped DE genes into two sets, those near (i.e., within 100kb of) sigDB and those not near sigDB, and tested each set for enrichment of Gene Ontology (GO) biological processes based on FDR-corrected hypergeometric test (**Methods**). In the cardiomyocyte differentiation data, DE genes in both sets showed significant enrichment for generic developmental terms like multicellular organismal development (**Figure 3.5C, Supp Table B.8**). However, DE genes near sigDBs tended to be significant for processes specific to cardiac and heart development (e.g. cardiac cell differentiation, heart development and morphogenesis). DE genes near sigDB between H1 and endoderm also showed significant enrichment in developmental terms (e.g. cell morphogenesis involved in differentiation, cellular component organization or biogenesis) compared to those not near sigDB (**Supp Table B.9**). For the mouse ESC to CN differentiation, DE genes near sigDB were enriched for neuronal processes when comparing ES vs CN and ES vs NPC (**Supp Table B.10**).

Finally, to characterize specific loci with differential 3D organization pattern, we prioritized regions based on the magnitude of change in their boundary scores, then overlapped them with genomic features such as retrotransposons and proximity to DE genes. Human endogenous retrovirus subfamily retrotransposons (HERV-H) in particular have been implicated in chromatin organization (Lawson et al., 2023) as a major determinant of TAD boundaries specific to hESC (i.e., day 0 of cardiomyocyte differentiation) when transcriptionally active (Zhang et al., 2019). We obtained genomic regions with the top 100 most transcriptionally active HERV-H sites and aggregated their TGIF-DB boundary scores at each timepoint. The boundary score at these transcriptionally active HERV-H sites is highest in the hESC (day 0) than in any subsequent timepoints (**Figure 3.6A**). This is further

supported by the presence of a boundary unique to day 0 that disappears in subsequent timepoints at one of the top transcriptionally active HERV-H sites (**Figure 3.6B**). Among the top-ranked sigDBs based on change in boundary scores, we found sigDB regions in which a boundary is present in the pluripotent state (day 0 hESC state of cardiomyocyte differentiation and H1 cell line, **Figure 3.6C,D**, respectively), but absent in differentiated state (day 2 mesoderm and definitive endoderm, respectively). These sigDB instances are proximal to the *ESRG* gene, significantly higher in expression in the pluripotent state compared to the subsequent differentiated states in both datasets. *ESRG* is a HERV-H containing long non-coding RNA (lncRNA, Wang et al., 2014); in addition to demarcating domain boundaries in hESCs, this particular site may effect the pluripotency state on knock-down (Wang et al., 2014) and has known roles in developmental and embryonal carcinoma (Wanggou et al., 2012). Among other top-ranked sigDBs in cardiomyocyte differentiation, we found DE genes with known roles in the cardiac development; for example, a particular boundary was found in primitive cardiomyocytes (day 15) but absent in fully differentiated ventricular cardiomyocytes (day 80, **Supp Figure B.13A**). This boundary overlaps *MYH6*, highly expressed in day 15 compared to day 80, and is adjacent to *MYH7*, which displays the opposite expression change pattern to *MYH6*. Both these genes are involved in cardiac muscle function: *MYH6* is expressed at high levels in developing atria, while *MYH7* in ventricular chambers of the heart (Ching et al., 2005; Warkman et al., 2012). Recently an enhancer region cluster, located downstream to *MYH7* at chr14:23,876,121-23,878,188 , was identified as a switch that can downregulate expression of *MYH7* while upregulating *MYH6* expression upon deletion (Gacita et al., 2021). Further using the same prioritization scheme in mouse neural differentiation data, we found a sigDB close to the *Ncam1* gene, which is differentially expressed between ES and CN (**Supp Figure B.13B**); *Ncam1* has known roles in neuron axon guidance and synapse formation (Hata et al., 2018; Shetty et al., 2013). These examples provide further evidence for TGIF-DB's ability to identify

relevant dynamic boundaries that could impact overall cell state identity.

3.2.6 Persistent boundaries are enriched for SNPs from diverse disease phenotypes.

Single nucleotide polymorphisms (SNPs) identified from genome-wide association studies (GWAS) are frequently found in non-coding regions of the genome and have been implicated in disease phenotypes by affecting the 3D genome organization (Lupiáñez et al., 2015; Orozco et al., 2022). Specifically, such variants could disrupt TAD boundaries and cause promiscuous expression of genes (Lupiáñez et al., 2015; Chakraborty and Ay, 2019). We investigated whether TGIF boundaries from the human cardiomyocyte differentiation data could be used to examine regulatory variants identified for diverse disease phenotypes in GWAS. We considered 17 phenotypic categories from the GWAS catalog and tested the enrichment of SNPs from each category in TGIF boundaries (**Methods**). SNPs across different categories were most enriched in the common set of boundaries across timepoints (i.e., persistent boundaries) than in other timepoint-specific or broader subsets of boundaries, with hematological measurement, cardiovascular disease, and lipid or lipoprotein measurement being the most enriched phenotypic categories (**Figure 3.7A**). Importantly, SNPs associated with cardiovascular disease (CVD) exhibited the second highest enrichment. The traits that had lower enrichment included neurological disorders and non-specific categories. We examined 66 persistent boundaries with at least one CVD-associated SNP. One such boundary had the SNP, rs72705895, which is associated with venous thromboembolism (Lindström et al., 2019, **Figure 3.7B**), and additionally overlaps a CTCF binding site (regulatory feature ENSR00000255184 from Ensembl regulatory build annotations; Zerbino et al., 2015; Cunningham et al., 2022). Another boundary included rs9349379, which is found in the intronic region of *PHACTR1* (**Supp Figure B.14**). Both the

intronic variant and the gene are associated with coronary artery atherosclerotic disease (Kuveljic et al., 2021; Koitsopoulos and Rabkin, 2021), while the SNP itself is on a predicted enhancer region (Ensembl regulatory build annotation ENSR00001107203), suggesting its putative role in disrupting an intronic enhancer. Genome editing experiments of boundary locations harboring these SNPs combined with Hi-C assays could help examine the role of dysregulated 3D genome organization as a possible mechanism by which regulatory variants impact phenotype.

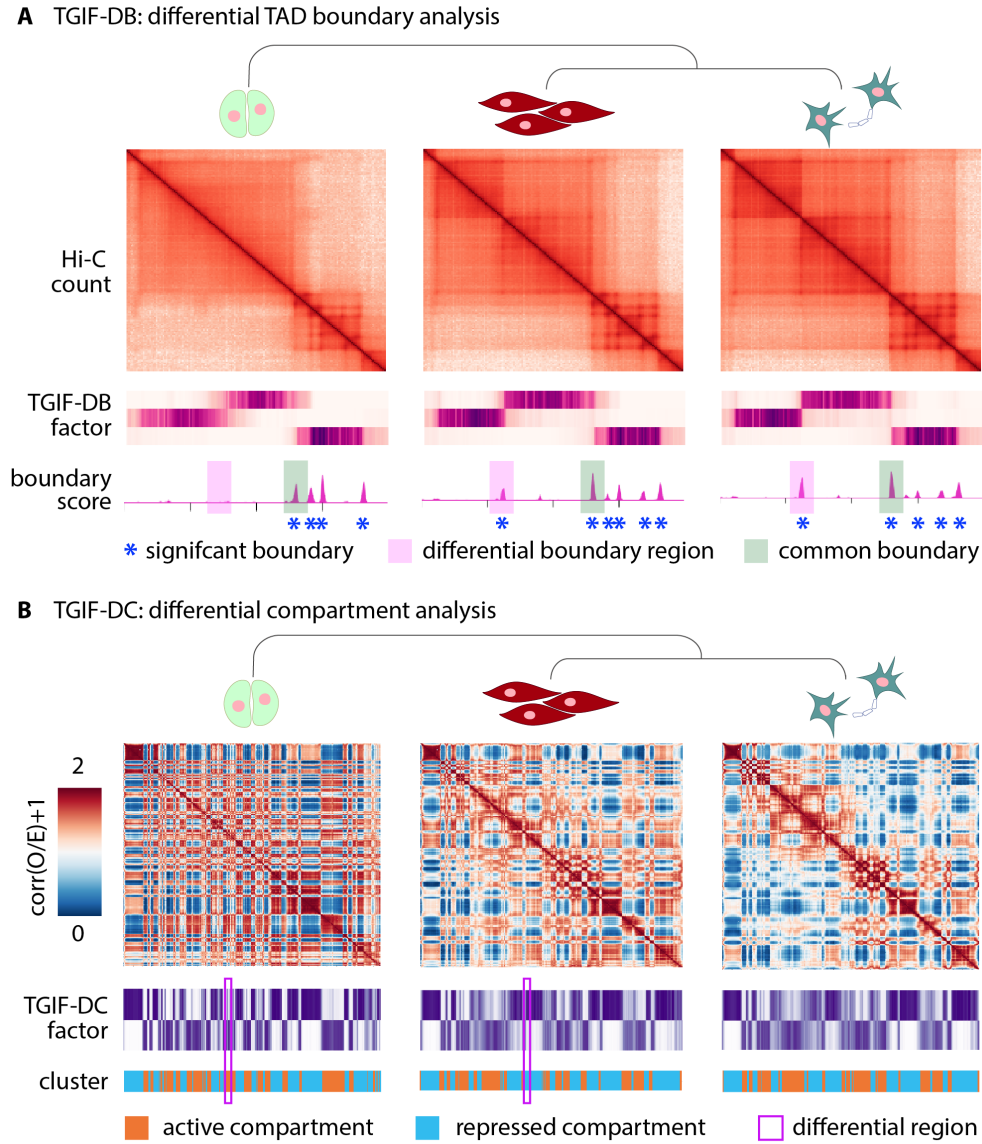


Figure 3.1: Overview of TGIF. (A) TGIF for differential boundary analysis (TGIF-DB). TGIF-DB takes multiple Hi-C count matrices as input and simultaneously learns a lower dimensional representation of genomic regions based on their interaction patterns. The input matrices are from related biological conditions with their relationship encoded as a tree. From the lower-dimensional factors, we measure the boundary score of each region, identify boundaries for each input condition and significantly differential boundaries for every pair of conditions. (B) TGIF for differential compartment analysis (TGIF-DC). TGIF-DC converts input matrices into correlation matrices of observed-over-expected (O/E) counts and factorizes them to yield latent features, which are used to cluster the regions. Each cluster correspond to a compartment or a subcompartment. TGIF-DC also identifies significantly differential compartmental regions for every pairs of input conditions.

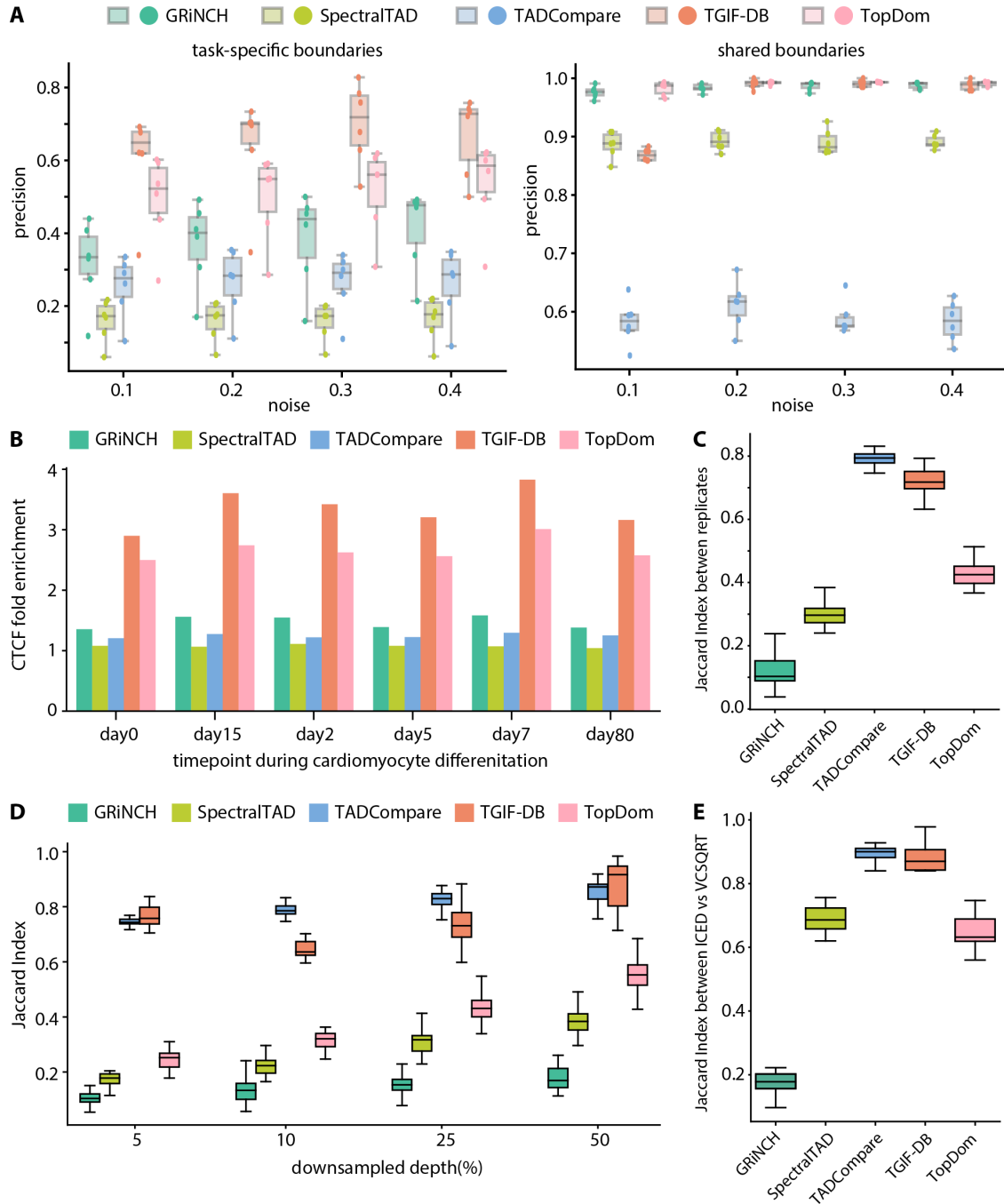


Figure 3.2: Benchmarking TGIF-DB. (A) Precision on simulated data with known task-specific and shared boundaries. Each point represents a pair of simulated matrices whose boundary sets are compared; the box plot the distribution over 6 pairwise comparisons among 4 simulated matrices. On the x-axis is the level of noise added to the simulated matrices. (B) CTCF peak enrichment in boundaries from different TAD-calling and differential-boundary-calling methods. (C) Boundary set similarity measured by Jaccard index between GM12878 data and downsampled data, across different downsampling depths. (D) Boundary set similarity between biological replicates of hESC (from day 0 of cardiomyocyte differentiation data). (E) Boundary set similarity between ICE-normalized and VCSQRT-normalized input matrices of mESC (from mouse neural differentiation data).

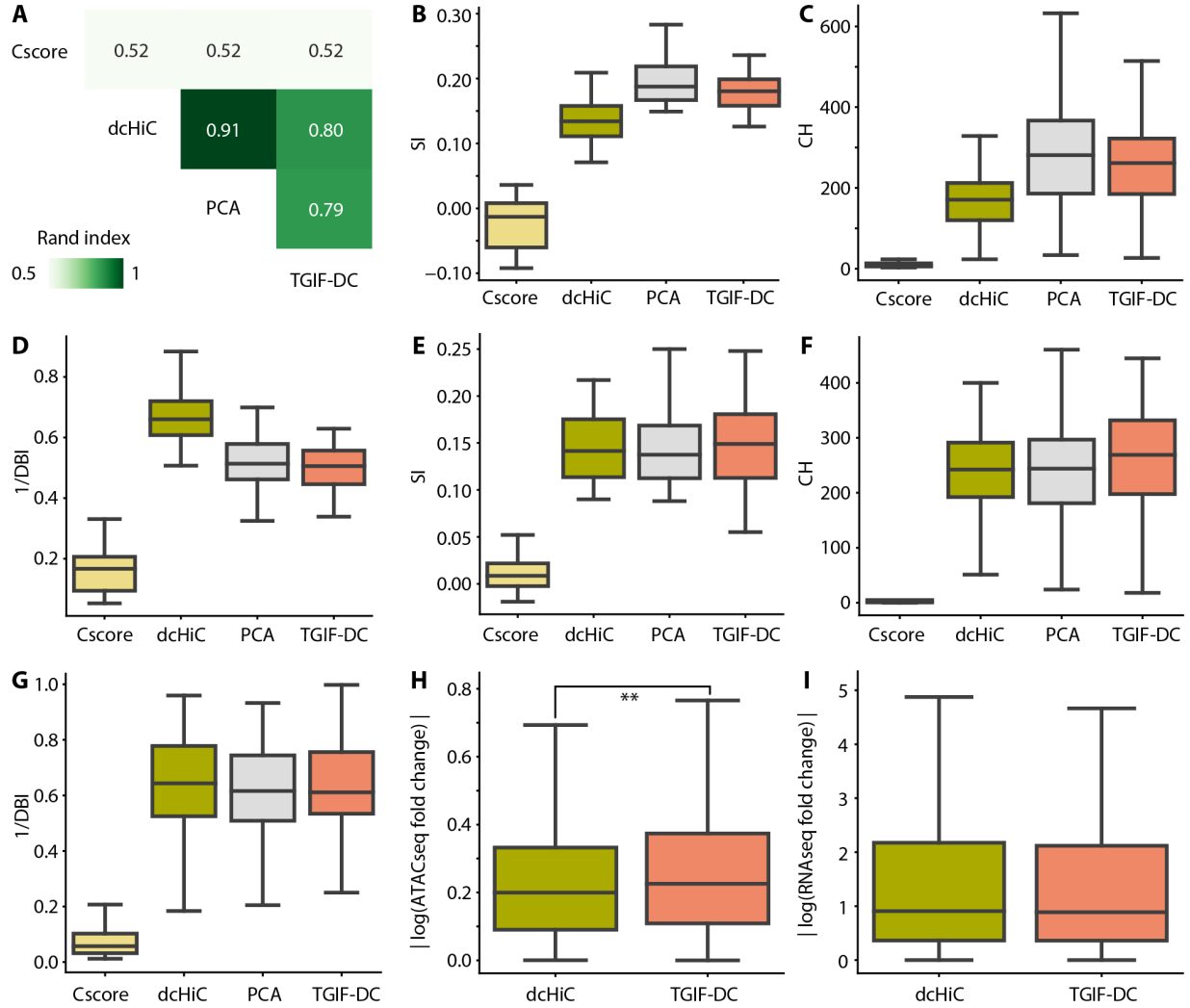


Figure 3.3: Benchmarking TGIF-DC on data from H1 and H1 differentiated to definitive endoderm. (A) Similarity of compartment assignments from different methods measured by Rand Index. (B) Quality of compartments based on O/E counts measured by Silhouette Index (SI), (C) Calinski-Harabasz score (CH), (D) Davies-Bouldin index (DBI). (E) SI, (F) CH, (G) DBI on accessibility (ATACseq) signal. (H) Magnitude of log fold change in accessibility between H1 and endoderm within significantly differential compartmental regions (sigDC) identified by dcHiC and TGIF-DC. (I) Magnitude of log fold change in gene expression between H1 and endoderm within sigDC identified by dcHiC and TGIF-DC.

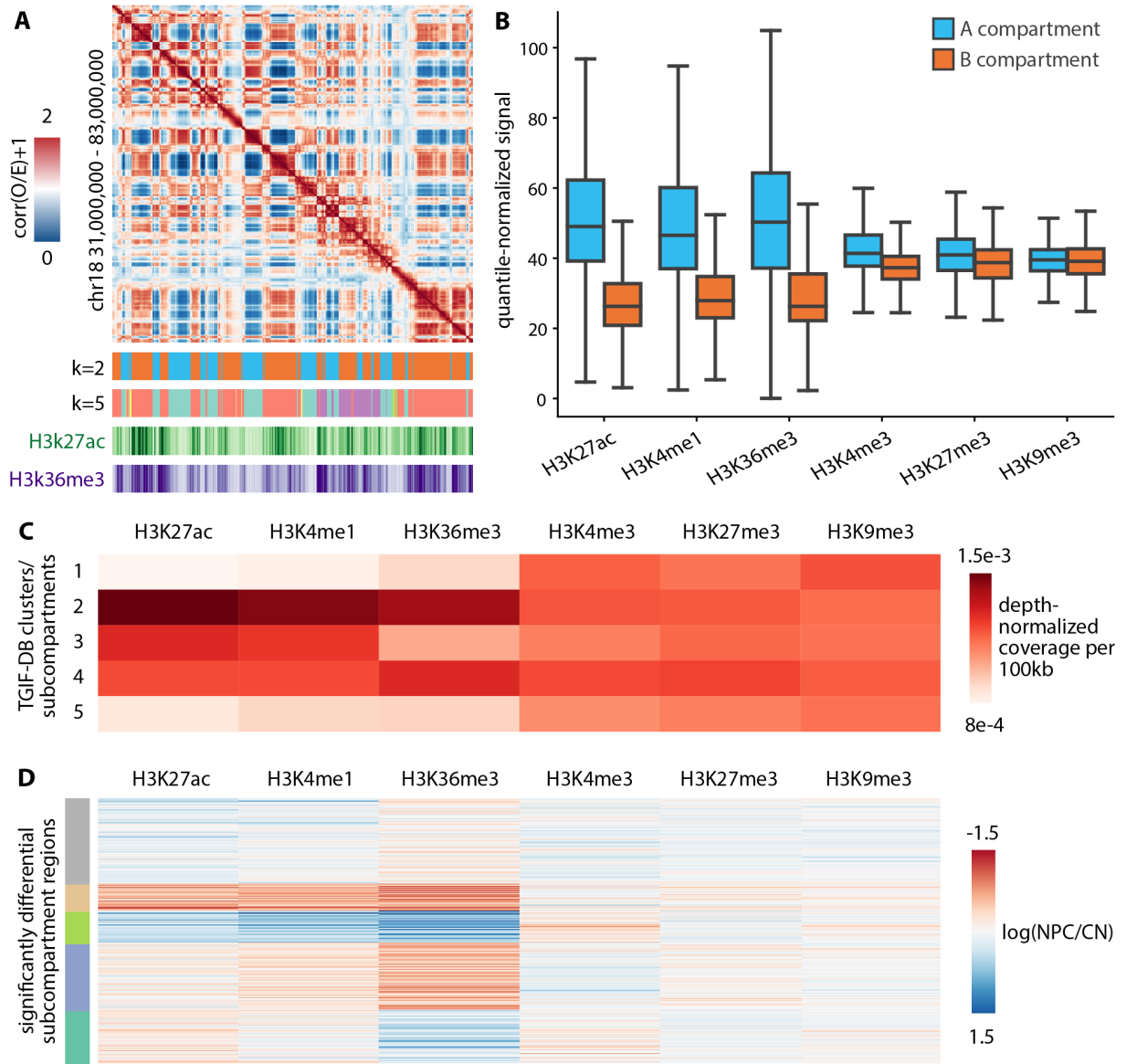


Figure 3.4: Characterizing compartments and subcompartments identified by TGIF-DC in mouse neural differentiation data.. (A) Similarity of compartment assignments from different methods measured by Rand Index. (B) Quality of compartments based on O/E counts measured by Silhouette Index (SI), (C) Calinski-Harabasz score (CH), (D) Davies-Bouldin index (DBI). (E) SI, (F) CH, (G) DBI on accessibility (ATACseq) signal. (H) Magnitude of log fold change in accessibility between H1 and endoderm within significantly differential compartmental regions (sigDC) identified by dcHiC and TGIF-DC. (I) Magnitude of log fold change in gene expression between H1 and endoderm within sigDC identified by dcHiC and TGIF-DC.

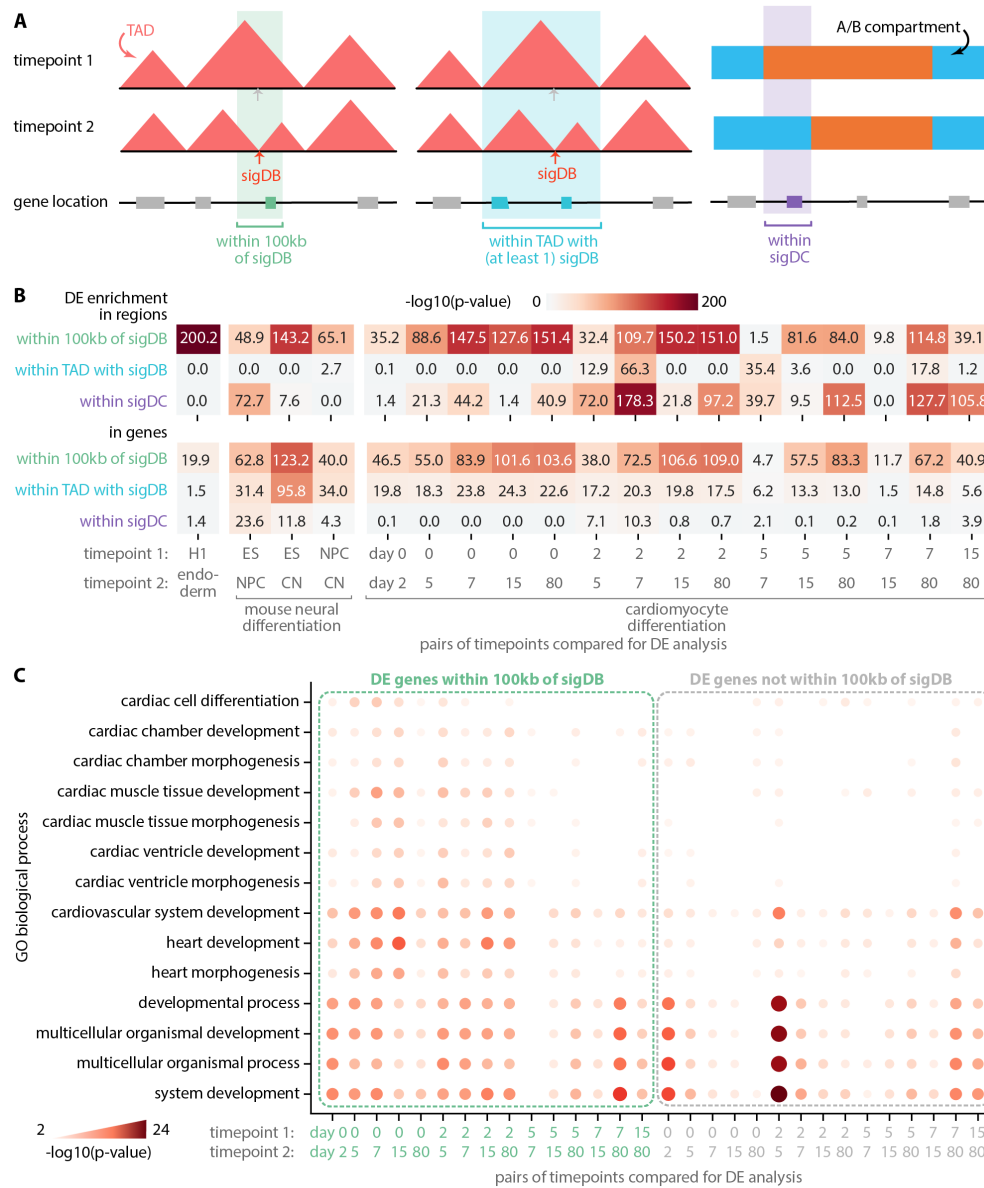


Figure 3.5: Differential gene expression near or within differential structural features..

(A) Differential gene expression (DE) enrichment was measured in regions and genes near or within dynamic regions, i.e. regions within 100kb of significantly differential boundary (sigDB), regions within TAD with at least one sigDB, and regions within significantly differential compartmental regions (sigDC). (B) DE, sigDB, and sigDC were measured and identified in pairwise comparisons of timepoints across 3 mammalian differentiation datasets: H1 differentiated to endoderm, mouse neural differentiation (ES, NPC, CN), and cardiomyocyte differentiation (day 0, 2, 5, 7, 15, 80). Negative log p-value of the enrichment hypergeometric test is visualized here. (C) GO biological process enrichment of genes within 100kb of sigDB from cardiomyocyte differentiation data.

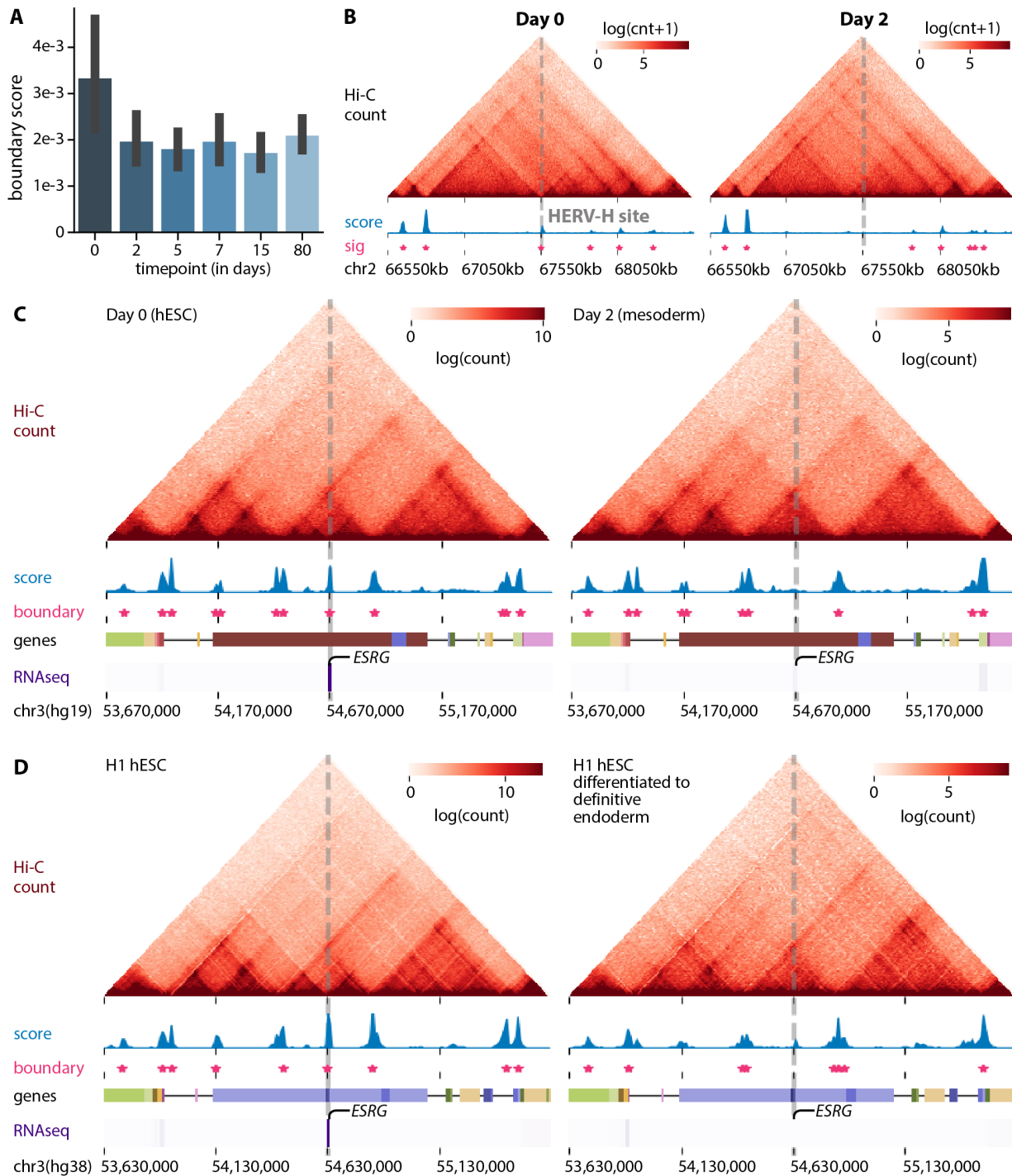


Figure 3.6: Human pluripotency-specific boundary elements.. (A) Boundary scores of transcriptionally active HERV-H retrotransposon sites during each timepoint of cardiomyocyte differentiation. (B) The top HERV-H site based on its transcription level in day 0 pluripotent state (within somatic chromosomes) and the overlapping sigDB identified by TGIF-DB. (C) A sigDB overlapping ESRG, an HERV-H-containing DE gene, in cardiomyocyte differentiation and in (D) H1 differentiated to endoderm.

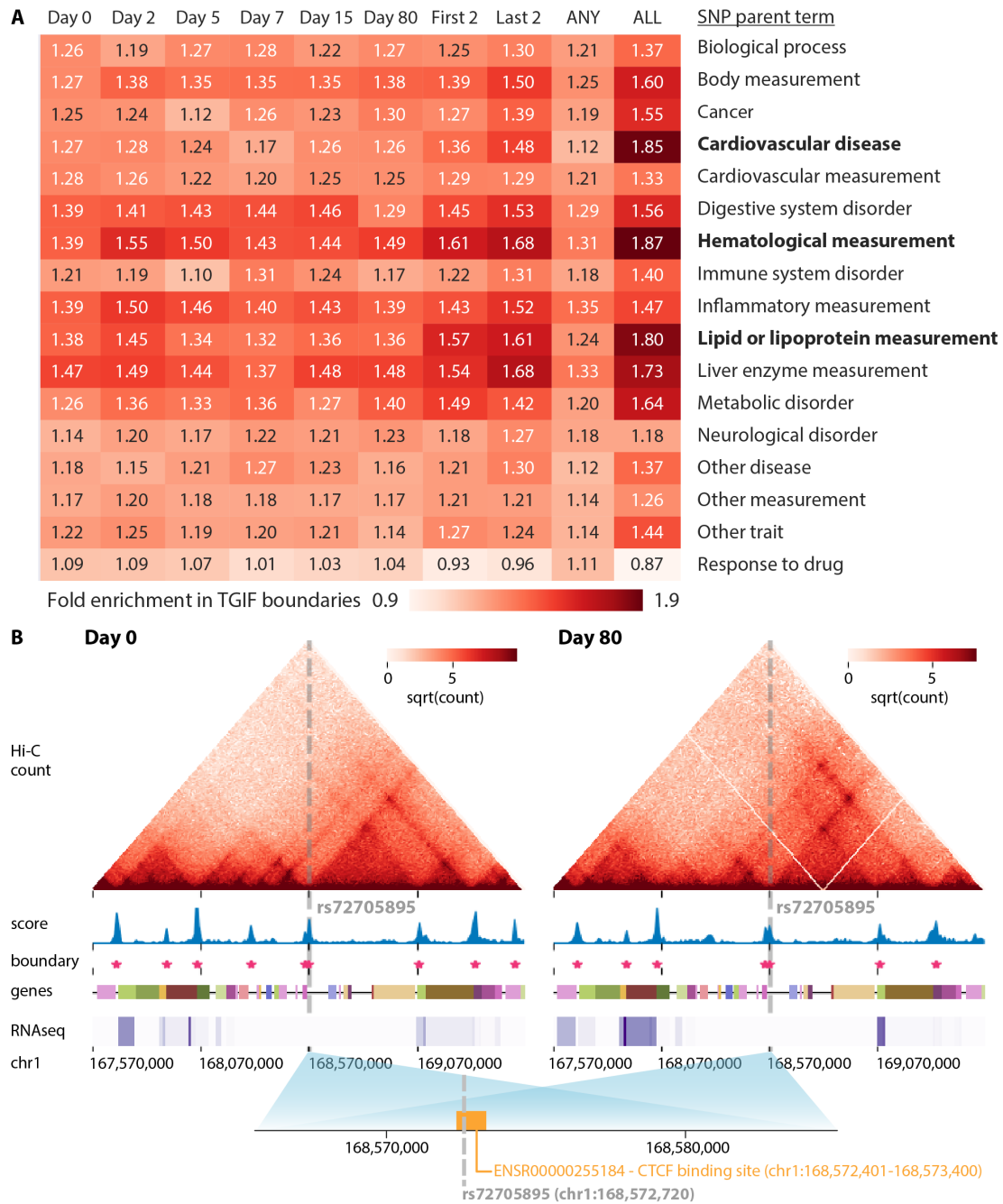


Figure 3.7: SNP enrichment in persistent boundaries.. (A) Fold enrichment of SNPs in different subsets of boundary regions, across different categories (SNP parent terms). We measured the enrichment of SNPs in timepoint-specific boundaries (day 0-80) of cardiomyocyte differentiation, in boundaries common to the first 2 or the last 2 timepoints, in union of boundaries (ANY), and in intersection of boundaries across all timepoints (ALL). (B) A SNP landing in a boundary persistent across all timepoints (only day 0 and day 80 visualized here) and a CTCF binding site.

3.3 Methods

3.3.1 Tree-Guided Integrated Factorization (TGIF)

Tree-Guided Integrated Factorization (TGIF) is based on multi-task non-negative matrix factorization (NMF) that can be used to identify low dimensional structure across multiple Hi-C datasets. Below we describe the TGIF framework in detail, which consists of NMF, hierarchical multi-task learning with tree-based regularization, and optimization with block coordinate descent.

NMF is a powerful dimensionality reduction method that can recover the underlying low-dimensional structure from high-dimensional data (Lee and Seung, 2000). It aims to decompose a non-negative matrix, $X \in \mathbb{R}_{\geq 0}^{(n \times m)}$, into two lower dimensional non-negative matrices, $U \in \mathbb{R}_{\geq 0}^{(n \times k)}$ and $V \in \mathbb{R}_{\geq 0}^{(k \times m)}$, to minimize the following objective: $\|X - UV\|_F^2$, s.t. $U \geq 0, V \geq 0$, where $\|\cdot\|_F$ indicates the Frobenius norm. We refer to the U and V matrices as factors. Here $k \ll n, m$ is the rank of the factors and is an input parameter. As described previously (Lee and Roy, 2021), to apply NMF to Hi-C data, we represent the Hi-C data for each chromosome as a symmetric matrix $X = [x_{ij}] \in \mathbb{R}^{(n \times n)}$ where x_{ij} represents the contact count between region i and region j .

TGIF implements multi-task NMF, where tasks correspond to Hi-C datasets that in turn are from hierarchically related contexts, such as cellular stages, species, timepoints. We note that a hierarchy is a general form for capturing relationships among a set of conditions and can capture both branching and linear relationships. Multi-task NMF has been previously implemented in the multi-view NMF approach (Liu et al., 2013; Baur et al., 2022), where a view and task can be used interchangeably. However, this existing framework assumes that all the tasks are equally related. Formally, in multi-view NMF, given T different datasets $\{X^{(1)} \dots, X^{(T)}\}$ where each dataset $X^{(t)} \in \mathbb{R}_{\geq 0}^{(n_t \times m)}$, the goal is to find view-specific factors $\{U^{(1)}, \dots, U^{(T)}\}$ and $\{V^{(1)} \dots, V^{(T)}\}$, and a consensus factor V^* that minimize the following

objective:

$$\sum_{t=1}^T \left\| X^{(t)} - U^{(t)} V^{(t)} \right\|_F^2 + \lambda \left\| V^{(t)} - V^* \right\|_F^2 \quad (3.1)$$

where $U^{(t)} \in \mathbb{R}_{\geq 0}^{(n_t \times k)}$ and $V^{(t)}, V^* \in \mathbb{R}_{\geq 0}^{(k \times m)}$. This constrains each of the task-specific factors $V^{(t)}$ to be similar to the consensus factor V^* . The hyper-parameter λ controls the strength of this constraint. The key benefit of such a framework is that the latent representation or structure within each task can borrow from other complementary data, as guided by the consensus factor V^* .

TGIF generalizes multi-view NMF to allow for integration of datasets that can come from different biological contexts such as time or developmental stage, and therefore may not all be equally related to each other. Accordingly, instead of requiring all the $V^{(t)}$ to be similar to a single V^* , in TGIF we account for the heterogeneity of the datasets by modeling the tasks to be related by a tree or a hierarchy. This makes TGIF applicable to a wide variety of task collections representing different biological contexts with arbitrary and complex relationships (e.g. Hi-C datasets from different cancer subtypes, cell lineage). In TGIF, the leaves of the tree correspond to the observed dataset while the internal node describe which tasks are most related. The child tasks are then regularized to its immediate parent task.

Formally, TGIF takes as input $t \in \{1, \dots, T\}$ tasks, each with input matrix $X^{(t)} \in \mathbb{R}^{n \times n}$ representing a symmetric count matrix for a Hi-C dataset over n genomic loci and a user-specified task hierarchy (**Figure 3.1A**) and optimizes the following objective:

$$\sum_{t=1}^T \left\| X^{(t)} - U^{(t)} V^{(t)} \right\|_F^2 + \alpha \sum_c \left\| V^{(c)} - V^{Pa(c)} \right\|_F^2 \quad (3.2)$$

The tree describes parent-child relationships between the tasks. This objective aims to:

1. constrain a task-specific latent factor $V^{(t)}$ in a leaf node of the task hierarchy to be similar to $V^{Pa(t)}$ in its parent node;
2. constrain an internal node's latent factor $V^{(b)}$ to be similar to its direct child nodes' $V^{(c)}$ and its parent node's $V^{Pa(b)}$;
3. constrain the root node's latent factor $V^{(r)}$ to be similar to all of its direct child nodes' $V^{(c)}_S$.

The hyper-parameter α controls the strength of the constraints such that the higher the α , the more the factor $V^{(c)}$ is encouraged to be similar to its parent. If all tasks share a single parent node (the root node), TGIF reduces to multi-view NMF.

Many optimization algorithms exist for learning the latent factors in matrix factorization (Kim et al., 2014). We chose a block coordinate descent (BCD) because it guarantees convergence to a local optimum (Kim et al., 2014). Intuitively, block coordinate descent updates a given block while keeping all other blocks fixed; in TGIF the block is each column or row of $U^{(\cdot)}$ s or $V^{(\cdot)}$ s. Starting at the leaf nodes (i.e., input tasks), we update each column/row of the factors, then move up the tree to update the parent factors. The specific update rules and their derivations can be found in **Supplementary Methods**.

TGIF's factors can be used to define compartments as well as fine-scaled topologically associating domains (TADs). TGIF-DC for compartments and TGIF-DB for boundaries and described in detail in subsequent sections.

3.3.2 TGIF-DB for differential boundary identification

TGIF-DB identifies TAD boundaries in four major steps: (a) multi-task factorization of input Hi-C matrices to obtain low dimensional representations; (b) compute a "boundary" score across different scales, (c) empirical p-value calculation and FDR correction to detect

significant boundaries; and (d) determine significant differential boundaries (sigDB) using z-score of pairwise boundary score differences.

Multi-task factorization of input Hi-C matrices. TGIF-DB operates on small partially overlapping submatrices along the diagonal of the symmetric intra-chromosomal interaction count matrices (**Supp Figure B.1A,B**). This mirrors the approaches taken by existing TAD-calling methods (Lieberman-Aiden et al., 2009; Cresswell and Dozmorov, 2020; Li et al., 2021). By default each submatrix spans $2\text{Mb} \times 2\text{Mb}$ with an overlap “step size” of 1Mb between consecutive submatrices. The exact dimension of the submatrix, namely the number of rows and columns, will depend on the resolution of the Hi-C data. For example, for a Hi-C dataset at 10kb resolution, the submatrix will be 200×200 with an overlapping step size of 100 rows/columns. The minimum size of the submatrices is bound at 100 (and the corresponding step size at 50) to prevent over-fragmentation of the input matrices, especially for lower-resolution input Hi-C matrices. Regions with interaction values missing for more than half of its neighbors in the radius defined by the window size in any of the input matrices are filtered out from the original input intra-chromosomal matrices before any submatrices are formed.

In NMF, usually the rank k of the lower dimensional factors is user-specified. However, TGIF does not require this since a single k value may not be appropriate across all task-specific input submatrices. Instead TGIF scans a range of k values, with $k \in \{2, \dots, 8\}$ to recover lower dimensional factors at multiple resolutions and defines boundaries based on a consensus of these factors (as described below). Because the submatrix size is small, it is computationally tractable to scan a range of k .

Boundary score calculation. After factorization, the next step is to identify genomic regions representing conserved or dynamic TAD boundaries across conditions. We define

a boundary as a region whose low-dimensional representation changes significantly compared to its immediate preceding neighbor bin. To this end we define a boundary score $S_i^{(t)}$ using the output factors for each of the T tasks from TGIF. Since $\mathbf{X}^{(t)}$ is symmetric, either $\mathbf{U}^{(t)}$ or $\mathbf{V}^{(t)}$ could be used to estimate these boundary scores. Assuming we use $\mathbf{U}^{(t)}$, the score $S_i^{(t)}$ for each region i in task t is the cosine distance between the low dimensional representation of region i and region $i - 1$:

$$S_i^{(t)} = 1 - \frac{\mathbf{U}^{(t)}[i, :] \cdot \mathbf{U}^{(t)}[(i - 1), :]}{\|\mathbf{U}^{(t)}[i, :]\| \|\mathbf{U}^{(t)}[(i - 1), :]\|} \quad (3.3)$$

While Euclidean distance could be used to compute the score, we used cosine distance since it normalizes the score regardless of magnitude differences across the tasks (which can arise from Hi-C sequencing depth differences). The final boundary score for region i in task t is the mean of $S_i^{(t)}$ estimated from factors across the range of $k \in \{2, \dots, 8\}$. For regions that are in the overlapping window between two consecutive count submatrices, the final boundary score is averaged from both submatrix factors. This allows us to capture the “boundary-ness” of each region at different structural scales (**Supp Figure B.1C**).

Empirical p-value calculation and FDR correction. Once the scores are calculated, we estimate a “null” distribution of boundary scores and use it to determine the empirical p-value of boundary scores and find significant boundaries. The null distribution is computed from a randomized background matrix. We first calculate the element-wise mean across T input submatrices to yield M , then create randomized background matrix by shuffling the entries of M . The shuffling is done in a distance-stratified manner; that is, we obtain all pairs of genomic regions that are at a distance d and permute them independent of the region pairs that are at different distance than d . The distance d ranges from 0 to the size of each window (e.g. 2Mb), incremented by the size of each bin (e.g. 10kb). We next performed

single-task NMF on this shuffled M matrix for the same range of k factors and derive boundary scores for all regions in the same way described in **Equation 3.3**. We treat this set of boundary scores as the samples from the null distribution. We calculate the empirical p-value for each the region i in task t as the proportion of “null” background scores higher than the given region’s boundary score. Finally, to find significant boundaries and to correct for multiple significant testing, we perform the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

The output of the p-value and FDR estimation step is a binary value for each region i and task t , indicating whether the region has a significant boundary score (1) or not (0). The significant boundaries identified in this manner may still be susceptible to noisy, low count regions of the genome. Therefore, we additionally filter the boundaries to find “summit-only” version of the significant boundaries, i.e. if there are more than one consecutive significant boundary regions along the linear genome, only the region with the highest significant score is called a boundary.

Significantly differential boundary regions in pairwise comparison of conditions. We provide a statistically significant subset of pairwise differential boundary regions (sigDB). For a pair of conditions with input Hi-C matrices, A and B, and for each genomic region i , we calculate the absolute difference in boundary scores $d_i^{(A,B)}$ between the two conditions. We estimate a null Gaussian distribution using the absolute difference of boundary scores of genomic regions which do not have significant boundaries in either A and B. We calculate the z-score and corresponding p-value of $d_i^{(A,B)}$ for all regions using this null distribution. After FDR correction, we report the regions with adjusted p-value < 0.05 as significantly differential boundary regions.

Hyper-parameter selection for TGIF-DB. To determine the setting of α , we examined the agreement between boundary assignments from a pair of input matrices for a given α value and the similarity of the input matrices themselves (SCC, Yang et al., 2017, **Supp Figure B.15A**). We used the cardiomyocyte differentiation data and scanned multiple hyper-parameter values in $\alpha \in \{10^2, 10^4, 10^6, 10^8\}$ across all chromosomes and compared the resulting boundary sets between every pair of timepoints using Jaccard index. In parallel, we measured the similarity of interaction count matrices between every pair of timepoints across all chromosomes using SCC, which is a weighted mean of correlation between interaction counts, stratified by genomic distance. SCC enables unbiased measurement of similarity between Hi-C datasets which are heavily distance-dependent (i.e., closer genomic regions tend to have higher interactions). Finally, we measured the correlation between the Jaccard index and the SCC within each chromosome, across all pairs of timepoints (**Supp Figure B.15B,C**). We find a slight, though not significant, improvement with $\alpha = 10^6$, which we set as default for TGIF-DB.

As BCD is a stochastic algorithm that can reach different local optima depending on the initialization point, we also experimented with multiple random initialization seeds. We used Jaccard index to measure the agreement between pairs of boundary sets from two different seeds, with α fixed at the default value 10^6 . We found that the resulting boundary sets from different initialization are relatively consistent with pairwise Jaccard index 0.76-0.77 (**Supp Figure B.15D**). We also estimated the memory usage and run time trend of TGIF-DB on 10kb input matrices from the three different timecourse datasets (**Supp Figure B.16A,B**). TGIF-DB's submatrix factorization framework with fixed set of k makes it scale linearly with the size of the input matrices.

3.3.3 TGIF-DC for differential compartment and subcompartment identification

Identification of compartments with TGIF-DC. In order to identify compartments, we apply TGIF to a 100kb resolution intrachromosomal Hi-C matrix that is first converted into an observed-over-expected (O/E) count correlation matrix as described previously (Rao et al., 2014, **Datasets and preprocessing**). To obtain the O/E matrix, we first remove rows and columns corresponding to regions with zero interactions, then distance-normalize every entry of the input matrices $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(t)} \in \mathbb{R}_{\geq 0}^{n \times n}$:

$$\mathbf{X}_{\text{OE}}^{(t)}[i, j] = \frac{\mathbf{X}^{(t)}[i, j]}{\frac{\sum_{d(p,q)=d(i,j)} \mathbf{X}^{(t)}[p, q]}{n - |i - j|}} \quad (3.4)$$

where the denominator is the expected count for pairs of loci i and j at distance genomic distance $d(i, j)$. Next we get the correlation matrix as:

$$\mathbf{X}_{\text{corr}}^{(t)}[i, j] = \text{corr}(\mathbf{X}_{\text{OE}}^{(t)}[i, :], \mathbf{X}_{\text{OE}}^{(t)}[j, :]) \quad (3.5)$$

where corr corresponds to the Pearson's correlation.

We upshift the correlation matrix by 1 so that all values are non-negative. To identify compartments, we apply TGIF with the input tree structure to these matrices with rank $k = 2$. After factorization, we infer each region i 's cluster assignment, $c_i^{(t)}$, for each task t , such that $c_i^{(t)} = \text{argmax}_{j \in \{1,2\}} U[i, j]$. We refer to these clusters as compartments. To identify subcompartments and differential subcompartment regions, a higher k value, e.g. 5, can be used and TGIF-DC will generate more granular cluster assignments, e.g. 5 clusters of regions instead of 2 clusters. Each of these clusters corresponds to a subcompartment.

Detecting differential compartments with TGIF-DC. While the cluster assignment switch is a straightforward way of identifying differential compartment regions, we further provide a statistically significant subset of pairwise differential compartment regions. We utilize the lower-dimensional representation of each genomic region from the factors in this step. For a pair of conditions or timepoints being compared, A and B, we calculate the cosine distance $d_i^{(A,B)}$ between $U^{(A)}[i, :]$ and $U^{(B)}[i, :]$ for each genomic region i . Using the cosine distance of regions that do not change their cluster assignment between the conditions (i.e., static regions), we estimate the mean and standard deviation of a Gaussian null distribution. The null distribution is used to calculate the z-score and p-value for the remaining (dynamic/differential) regions. Statistically significant differential regions are those with an FDR < 0.05 . Significantly differential subcompartment regions are identified in the same way as the differential compartment regions.

Hyper-parameter selection for TGIF-DC. To determine the best setting of α , we used a similar approach as for TGIF-DB, measuring the agreement (correlation) between the input matrix similarity and the agreement of TGIF-DC compartment assignments for the same inputs. Specifically, we measured the similarity of observed-over-expected (O/E) count matrices between every pair of timepoints across all chromosomes by flattening them into a vector and measuring correlation. Note, O/E counts are already normalized by distance. In parallel, we measured the similarity of the output clusters (compartments) with Rand index (**Supp Figure B.17A**). We used the mouse neural differentiation data to study this parameter, scanning $\alpha \in \{10^2, 10^4, 10^6, 10^8\}$ across chromosomes (**Supp Figure B.17B**). TGIF-DC consistently yields cluster assignments that are well correlated (~ 0.9) with the input matrix similarity across wide a range of α values (**Supp Figure B.17C**). Our results are from $\alpha = 10^4$, which is the default for TGIF-DC.

Similar to TGIF-DB, we also examined the stability of compartment assignments with

multiple random initialization seeds with $\alpha = 10^4$ using Rand Index between pairs of cluster assignments from two different seeds. At $k = 2$, the compartment setting of TGIF-DC, the Rand Index ranged from 0.99-1 for all pairs of random initializations (**Supp Figure B.17D**). At $k = 5$, the subcompartment setting, Rand Index of resulting subcompartments was 0.7-0.8 (**Supp Figure B.17E**), showing that TGIF-DC yields a stable set of cluster assignments regardless of random initialization. Finally, we report the memory usage and run time trend of TGIF-DC on 100kb input matrices from the three different timecourse datasets (**Supp Figure B.18A,B**). TGIF-DC can analyze 6 datasets needing no more 0.5GB of memory and 400 seconds of run time.

Post-hoc annotation of TGIF-DC clusters into A and B compartments TGIF-DC by default uses $k=2$ and segments the given chromosome into 2 clusters of regions. In our analysis, we use GC content and chromatin accessibility to annotate each cluster as A or B compartment. Note that identification of significant differential compartmental region (sigDC) does *not* depend on this post-hoc annotation step, as sigDC identification only looks at cluster assignment changes and absolute change in latent features of each genomic region.

For the H1 to endoderm differentiation dataset we used the available compendium ATACseq data from 4D Nucleome (Reiff et al., 2022; Dekker et al., 2023, **Supp Table B.1**) as a measurement of chromatin accessibility. For each chromosome, we measure the mean ATACseq reads per 100kb bin from H1 within each of the 2 clusters. The cluster with higher mean ATACseq signal is assigned to compartment A and the other to compartment B for both H1 and endoderm (this is possible due to the cluster correspondence across timepoints in TGIF-DC; **Supp Figure B.19A,B,C**). To validate the compartment annotation, we also measured the GC percentage within each 100kb bin for each compartment, and found that regions assigned to A compartment have higher GC content than those in B

compartment, as expected (**Supp Figure B.19D**).

We proceeded similarly with cardiomyocyte differentiation data, except we used DNase-seq data for chromatin accessibility from day 0 (H9 cell line, hESC state, **Supp Table B.3**). We measure the mean DNase-seq reads per 100kb bin for both clusters in each chromosome. The cluster with higher mean DNase-seq signal is labeled compartment A and the other one is labeled compartment B across all timepoints (**Supp Figure B.20A,B**). To validate the compartment annotation, we also measured GC percentage and found that regions assigned to the A compartment has higher GC content (**Supp Figure B.20C**). Since H3K27ac was available for this dataset, we also compared the mean H3K27ac reads within each 100kb bin of each compartment and found higher H3K27ac signal in the A compartment compared to B (**Supp Figure B.20D**). This is consistent with the definitions of A and B compartments (Nichols and Corces, 2021; Bouwman et al., 2023) and provides further support for TGIF-DC's compartmentalization.

For mouse neural differentiation data, we only used GC content for compartment annotation. We performed this annotation for the ES timepoint and transferred it to the other timepoints. Briefly, for each chromosome, we measure the mean GC% per 100kb bin in each compartment. The compartment with higher mean GC content is called A and the other one B across all timepoints (**Supp Figure B.8**).

3.3.4 Estimating tree structure from input Hi-C matrices for unknown inter-dataset relationships

Typically, prior information about the relatedness of the biological contexts from which the Hi-C data matrix originate is available. However, if this information is not available (e.g. integrating Hi-C datasets from different laboratories or single-cell Hi-C datasets), a tree can be estimated using pairwise similarity of the input Hi-C matrices, converting to

distance followed by hierarchical clustering. We suggest the use of a distance-stratified similarity measure, such as the stratum-adjusted correlation coefficient (SCC, Yang et al., 2017, **Supp Figure B.2A**), that we have also used for our hyper-parameter analysis. Once SCC is calculated for each pair of input matrices, it is converted to a distance by subtracting from 1 (**Supp Figure B.2B**) which in turn is used as input to hierarchical clustering with average linkage. We tested this approach for the mouse neural differentiation dataset and found that the output tree of hierarchical clustering is similar to the known biological relatedness of this dataset (**Supp Figure B.2C**, Bonev et al., 2017) and is identical to the tree we used as input to TGIF for our experiments. The current implementation of TGIF offers this functionality as a pre-processing script (See Section on **Data access and code Availability**).

3.3.5 Datasets and preprocessing

We applied TGIF to three Hi-C timecourse datasets: H1 hESC differentiated to endoderm (Reiff et al., 2022; Dekker et al., 2023), mouse neural differentiation data from Bonev et al., 2017, and human cardiomyocyte differentiation data from Zhang et al., 2019 (**Supp Figure B.4**). The Hi-C interaction count matrices for pluripotent H1 hESC cell line and for endoderm differentiated from H1 was downloaded from 4D Nucleome consortium (**Supp Table B.1**). 100kb intra-chromosomal count matrices were used for comparison of compartment-calling methods. Additionally, ATACseq data for H1 and endoderm was also downloaded and used to measure the accessibility of each 100kb region (i.e. mean ATACseq reads per base in the region) in the comparison of compartment-calling methods. 10kb VCSQRT-normalized intra-chromosomal count matrices was used in the analysis of differential boundary and expression analysis.

The mouse neural differentiation data (Bonev et al., 2017) included 3 timepoints during

mouse neural differentiation: embryonic stem cell (ES) stage, neural progenitor stage (NPC), and the differentiated cortical neuron (CN) stage. For TGIF-DB, we used intra-chromosomal count matrices at a resolution of 10kb resolution with vanilla-coverage square-root (VCSQRT) normalization as input. When benchmarking boundary-calling methods, we also used 25kb and 50kb VC-SQRT-normalized data, as well as 25kb ICE-normalized data. For TGIF-DC, intra-chromosomal interaction count matrices at 100kb resolution without normalization was used as input since TGIF-DC computes the O/E correlation matrices. In addition to the Hi-C measurements, this dataset also included ChIPseq data for 6 histone modification marks (H3K27ac, H3K28me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3), for both CN and NPC (**Supp Table B.2**). This data was used to characterize and validate the chromatin structure inferred by TGIF-DC. For each 100kb bin, the ChIPseq signal in reads-per-million was summed up within each bin, and the signal in each bin was divided by the total number of reads to first normalize by read depth. Subsequently, signals in NPC and CN were quantile normalized to each other to enable log fold change comparison across the two timepoints. The log fold change for each of the 6 marks in each bin was calculated by log transforming ($\log(x+1)$) the normalized signal in NPC divided by the signal in CN. The set of log fold change signals were used as input to a k-means clustering, with $k = 5$ clusters applied to sigDC regions.

The cardiomyocyte differentiation dataset from Zhang et al., 2019 (**Supp Table B.3**) measured Hi-C counts at 6 different timepoints (day 0, 2, 5, 7, 15, 80) from the human embryonic stem cell stage (hESC, day 0) to the ventricular cardiomyocyte stage (day 80). Two replicates from each timepoint were first merged and intra-chromosomal interaction count matrices were generated at 10kb resolution with ICE normalization using the Juicer tool (Durand et al., 2016). The 10kb resolution matrix was provided as input to TGIF-DB. The merged dataset was also used to generate 100kb resolution intra-chromosomal count matrices for input to TGIF-DC. **Benchmarking with downsampled data to assess**

robustness to depth section contains details on how GM12878 cell line data was processed and used.

3.3.6 Benchmarking methods for identifying differential domain boundaries

Description of existing methods. TGIF-DB was benchmarked against four other methods for identifying differential TAD boundaries: GRINCH (Lee and Roy, 2021), SpectralTAD (Cresswell et al., 2020), TADCompare (Cresswell and Dozmorov, 2020), and TopDom (Shin et al., 2016). GRiNCH, SpectralTAD, and TopDom are single-task TAD identification methods accepting a single input matrix individually followed by pairwise comparison of identified boundaries.

GRiNCH is a TAD identification method that also utilizes a variation of NMF. It applies graph-regularized NMF to an input intra-chromosomal matrix as a whole and uses the output factor to cluster the genomic regions; each cluster represents a TAD. The boundary of such clusters were used as TAD boundaries in our analysis. For pairwise differential analysis, boundaries found in both input matrices were considered shared boundaries, while those found in one and not the other were considered differential boundaries. Version 1.0.0 was used with default values for all optional hyperparameters.

SpectralTAD treats an input Hi-C matrix as a graph of interacting genomic regions, and applies eigen decomposition to its graph Laplacian. The eigenvectors are used as latent features of each genomic region to cluster them, with each cluster representing a TAD. Similar to GRiNCH, for pairwise differential analysis, boundaries found in both input matrices were considered shared boundaries; those found in one and not the other differential boundaries. Version 1.16.1 was used in our analysis.

TopDom is a TAD identification method which was shown to be robust to noise and

yielded TADs enriched in structural proteins such as CTCF (Dali and Blanchette, 2017; Lee and Roy, 2021). It generates a score for each bin along the chromosome, where the score is the mean interaction count between the given bin and a set of upstream and downstream neighbors (neighborhood size is a user-specified parameter). Putative TAD boundaries are picked from a set of bins whose score forms a local minimum; false positive boundaries are filtered out with a significance test. Differential boundaries are identified in a pairwise manner similar to GRiNCH and SpectralTAD. Version 0.10.1 was used in our analysis, with the window size hyper-parameter set to the recommended value of 5 (Shin et al., 2016).

TADCompare is a differential TAD identification method that can take as input a pair of Hi-C matrices as well as a time series of Hi-C matrices. It treats each Hi-C matrix as a graph where each genomic region is a node and the pairwise interaction is a weighted edge with the weight corresponding to the count. It performs eigen decomposition on the Graph Laplacian of each Hi-C matrix. Boundary scores are calculated between a pair of regions by computing the Euclidean distance between their corresponding rows in the eigenvectors. Differential boundary scores are calculated by taking the difference in the boundary scores for a pair of conditions and converting it into a z-score and using a threshold of 2 to define a differential boundary. Although TADCompare accepts time-series data differential boundaries are computed for only pairs of matrices at a time. Version 1.2.0 was used in our benchmarking analysis.

In addition to the above mentioned tools, we initially considered the TADsplimer (Wang et al., 2020), and HiCExplorer (Wolff et al., 2020) methods. TADsplimer was specifically developed to detect differential TAD identification; however, its implementation has an unresolved issue that fails to return any output or differential boundaries and was excluded from further analysis. HiCExplorer allows TAD finding followed by differential TAD analysis using its hicDifferentialTAD tool. However, hicDifferentialTAD expects the same TAD for different conditions and outputs TADs with significantly different interaction

counts rather than detecting boundary changes. Due to these reasons, they were excluded from subsequent benchmarking.

Benchmarking on simulated data with known boundaries. In order to benchmark methods that can detect TAD-level changes, we generated simulated contact matrices with known TADs and TAD changes for four hierarchically related conditions (**Supp Figure B.21**). The TAD changes can fall into one of three categories: TAD split creating a new boundary, TAD merge removing a boundary, and TAD shift where the location of a boundary is moved up or down the linear chromosome (**Supp Figure B.21A**, Cresswell and Dozmorov, 2020). We first generate a set of TADs with known change patterns, then populate contact matrices following the Hi-C count simulation procedure in the benchmarking study by Forcato et al., 2017.

Forcato et al., 2017 originally simulated 171 TADs, which at 40kb resolution resulted in a target size of the simulated matrix similar to the size of the human chromosome 5 (180.92Mb). To generate the TADs for each of the four tasks or conditions, we start with these 171 TADs (referred to as A) at the root. Of these, 40 TADs are kept unchanged during the TAD change simulation process. We proceed down the tree branch keeping A unchanged for the left most branch (**Supp Figure B.21B**), and performing different TAD change operations on the other branches and at each level. This results in a tree structure where TAD sets A and A₂ are most similar, followed by B₁ and B₂. Pairwise differences between the TAD sets are quantified in **Supp Figure B.21C**. The resulting TAD sets are considered our "gold-standard".

Given a TAD structure, we follow the simulation procedure from Forcato et al., 2017 to generate the counts. Each interaction count for a pair of regions (i, j) is sampled from a negative binomial distribution defined by two parameters: (1) the dispersion parameter is fixed at 0.01 and (2) the mean parameter $\mu_{i,j} = K(i - j + 1)^{-0.69}$, which is dependent

on the distance $(i - j)$ between the two interacting regions i and j . The mean parameter $\mu_{i,j}$ decays as the distance between i and j grows to reflect how Hi-C interaction counts decay as a function of pairwise distance. A prior value of 1 is added to this distance when calculating $\mu_{i,j}$. Here -0.69 is the rate of decay parameter estimated from a real contact matrix (chr5 from IMR90 cell line). Additionally, $K = 28$ if a pair of regions falls within the same TAD and $K = 4$ otherwise. Finally, we add noise to a random proportion of counts by adding 2 to the mean parameter $\mu_{i,j}$ to the randomly sampled counts. We added different levels of noise, 0.1, 0.2, 0.3, and 0.4 to each of the matrices, corresponding to randomly sampled 10, 20, 30, 40% of the interactions with noise added.

We applied GRiNCH, SpectralTAD, TADCompare, TGIF-DB, and TopDom to the simulated datasets to assess their ability to recover shared and differential boundaries. We applied TADCompare to each pair of the 4 simulated matrices. TADCompare outputs differential boundaries, including the task in which the boundary is significant, and non-differential boundaries, which we consider as shared boundaries. We applied single-task methods (GRiNCH, SpectralTAD, TopDom) to each of the four simulated matrices independently to identify the TADs for each input matrix. The resulting TAD boundaries for each pair of input matrices were compared to identify task-specific and shared boundaries. TGIF was applied to all four simulated matrices together with the known tree structure used to generate the simulated data (**Supp Figure B.21C**).

We calculated precision and recall of task-specific and shared boundaries in every pair of simulated matrices. Briefly, let $T_{\text{shared}}(A, A_2)$ be the true number of shared boundaries from matrices or tasks A and A_2 , $P_{\text{shared}}(A, A_2)$ be the number of predicted shared boundaries, and $I_{\text{shared}}(A, A_2)$ be the overlap between the true and predicted shared boundaries. The precision, p is defined as $\frac{I_{\text{shared}}(A, A_2)}{P_{\text{shared}}(A, A_2)}$, recall, r is defined as $\frac{I_{\text{shared}}(A, A_2)}{T_{\text{shared}}(A, A_2)}$.

Measuring CTCF enrichment in boundaries. To evaluate the boundaries identified by various TAD-calling methods, we measured CTCF peak enrichment in those boundaries on the cardiomyocyte differentiation dataset which had CTCF ChIPseq data and as such represented a dataset with the largest number of time points and different types of assays (**Supp Table B.3**). Using macs2 (Zhang et al., 2008), we first called peaks on CTCF ChIPseq data from each of the 6 timepoints (day 0, 2, 5, 7, 15, 80) of the cardiomyocyte differentiation time course. Replicates from each timepoint were collapsed by intersecting overlapping peaks with Bedtools (Quinlan and Hall, 2010). Each peak was then assigned to a 10kb uniform bin again using Bedtools. TAD-calling methods GRiNCH, SpectralTAD, and TopDom were applied to 10kb Hi-C matrices from each of the 6 timepoints to yield 6 timepoint-specific sets of boundary bins. TGIF-DB was applied to Hi-C matrices from all 6 timepoints using the tree structure as in **Supp Figure B.4**, and significant boundaries from each timepoint were used for enrichment analysis. TADCompare was applied to each pair of consecutive timepoints: day 0 vs 2, 2 vs 5, 5 vs 7, 7 vs 15, 15 vs 80. As TADCompare outputs both non-differential and differential boundaries for every pairwise comparison, we define a boundary set specific to a timepoint as follows: (1) for day 0, union of differential boundaries in day 0 and non-differential boundaries between day 0 and 2; (2) for day 80, union of differential boundaries in day 80 and non-differential boundaries between day 15 and 80; (3) for all intermediate timepoints t , union of differential boundaries in t , non-differential boundaries between day t and t_{previous} , and non-differential boundaries between day t and $t_{\text{following}}$.

The CTCF peak fold enrichment ratio for a given timepoint was calculated as $\frac{q}{M} / \frac{s}{N}$, where q is the number of boundaries with at least one CTCF peak, M is the number of boundary regions, s is the number of regions with at least one CTCF peak, and N is the total number of genomic regions. We also used these estimates to compute a p-value using a hypergeometric test; however, the fold enrichment was more informative for comparing

methods.

Benchmarking with downsampled data to assess robustness to depth We downloaded the high-depth Hi-C dataset of GM12878 cell line (Rao et al., 2014) with 4.01 billion total reads from the 4D Nucleome data portal (Reiff et al., 2022; Dekker et al., 2017, **Supp Table B.4**). We then subsampled 5, 10, 25, 50% of the reads, generated 10kb-resolution intra-chromosomal Hi-C matrices using Juicer (Durand et al., 2016), and ICE-normalized the intra-chromosomal interaction matrices from each downsampled dataset (see scripts on <https://doi.org/10.5281/zenodo.13323899> for details). We used Jaccard index to measure agreement of boundaries identified by different methods despite depth difference. The Jaccard index is calculated by dividing the number of boundaries found at both depths by the number of boundaries identified in either depths for the GM12878 dataset. The higher the Jaccard Index, the fewer the false-positive differences.

Three TAD-calling methods, GRiNCH, SpectralTAD, and TopDom were applied individually to 5 datasets: original high-depth GM12878 data and four low-depth GM12878 data downsampled to 5, 10, 25, 50% depths, respectively. Jaccard index was calculated between the boundaries identified from the original dataset and those from each of the downsampled datasets. TADCompare and TGIF-DB were applied to 4 pairs of datasets, each pair including the original high-depth GM12878 dataset and the downsampled low-depth dataset (e.g., GM12878 data downsampled to 50% depth, **Supp Figure B.4**). For TADCompare, the Jaccard index was calculated for each pair of datasets as the ratio of number of non-differential boundaries and the number of differential and non-differential boundaries. Similarly, for TGIF-DB, the Jaccard index was calculated as the ratio of the number of non-significantly differential boundary regions divided by the size of the union of boundary regions from original-depth and subsampled dataset.

Measuring stability of boundary sets across multiple resolutions of input data We used the mouse neural differentiation dataset (Bonev et al., 2017) to assess the stability of boundary sets identified by different TAD boundary identification methods at different resolutions, namely, 10kb, 25kb and 50kb, since this dataset was readily available at these resolutions. We focused our comparisons only for the mouse embryonic stem cell (mESC) time point. The single-task boundary calling methods (GRiNCH, SpectralTAD, TopDom) were applied individually to 10kb, 25kb, and 50kb intra-chromosomal matrices from mESC. TADCompare was applied to a pair of timepoints including mESC, both at the same resolution: mESC vs neural progenitors (NPC), and mESC vs cortical neurons (CN). In order to find mESC boundaries from the outputs of the pairwise TADCompare comparisons at each resolution, we took the union of non-differential boundaries and differential boundaries enriched in mESC. We applied TGIF-DB to a tree with all three timepoints from mouse neural differentiation dataset at a specific resolution, and took the significant boundaries from mESC (**Supp Figure B.4**). This was repeated for each resolution. To allow for comparison of boundaries from different resolutions, we project the higher resolution bins to the coarsest resolution, namely 50kb. For instance, in the 25kb vs 50kb comparison, each 50kb bin is composed of two 25kb bins and is considered to have a boundary if either of the 25kb bins had a boundary. Similarly, for the 10kb vs 50kb comparison, any of the 5 comprising 10kb bins would be used to define a boundary in the 50kb bin spanning them. In the 10kb vs 25kb comparison, if any of the 10kb bins or the 25kb bins have a boundary in the shared 50kb bin, we define the 50kb bin as a boundary. We then measure Jaccard index of boundaries at this lowest resolution.

3.3.7 Comparison of TGIF-DC to existing compartment-calling methods

We compared TGIF-DC to two established methods for calling compartments, i.e. principal component analysis (PCA) based method (Lieberman-Aiden et al., 2009) and Cscore (version 1.1, Zheng and Zheng, 2018), as well as a method designed specifically for differential compartment analysis, dcHiC (version 2.1, Chakraborty et al., 2022). We applied all four methods to 100kb intra-chromosomal count matrices from H1 hESC cell line. TGIF-DC and dcHiC were additionally applied to 100kb intra-chromosomal count matrices from H1 differentiated to endoderm. Both datasets were downloaded from 4D Nucleome consortium (Reiff et al., 2022; Dekker et al., 2023).

The PCA-based method applies PCA to the observed over expected count (O/E) correlation matrices. The first principal component (PC1) is used to assign genomic regions to two compartments; values equal to or above zero in PC1 are assigned to one compartment and those below zero are assigned to the other compartment. The actual annotation of each compartment as the active “A” or repressed “B” compartments is done by correlating the compartment structure to one-dimensional regulatory signals such as ATACseq assays or histone modifications. Cscore outputs a score that specifies the compartment of a region by modeling the interaction counts and genomic distance with a probabilistic model. Regions with Cscore values above or equal to zero were clustered into one compartment and those below another compartment.

Finally, dcHiC is a framework that can identify differential compartment regions. dcHiC performs fast PCA on distance-normalized correlation matrices, quantile-normalizes the PC values so they can be compared across multiple conditions, and identifies a set of genomic regions whose compartment scores (normalized PC values) are significantly different in any of the conditions using a chi-square test (Chakraborty et al., 2022). As the current version of dcHiC requires at least 2 replicates per condition or timepoint, we provide dcHiC

with interaction counts for two replicates per condition. For the other methods we provide a replicate-merged count matrix per condition available from the 4D Nucleome consortium.

To compare the compartment results across the different methods, we measured the Rand index between compartment assignments to each genomic region. To measure the quality of the compartments, we used three well-known cluster quality metrics: Silhouette Index, Calinski-Harabasz Score, and Davies-Bouldin Index, measured on the observed-over-expected (O/E) matrices for each chromosome, as well as the accessibility signal for each 100kb genomic region. The accessibility signal was defined as the mean ATACseq reads per basepair.

Finally, to compare dcHiC and TGIF-DC for significantly differential compartments between H1 and endoderm, we calculated the log ratio of the accessibility signal and gene expression (from RNAseq, in TPM) in H1 over that of endoderm for each significantly differential region.

3.3.8 Assessing differential gene expression near or within significantly differential boundaries and compartments

We used RSEM (Li and Dewey, 2011) on the raw RNA-seq data from the cardiomyocyte differentiation and the mouse neural differentiation time course to obtain expected counts for each replicate at each timepoint. We also downloaded the RNAseq data for H1 hESC cell line and endoderm differentiated from H1 from 4D Nucleome (Reiff et al., 2022; Dekker et al., 2023). We used these values as input to DESeq2 (Love et al., 2014) to identify differentially expressed (DE) genes for every pair of timepoints in each dataset (e.g., H1 vs endoderm; mESC vs NPC; day 0 vs. day 2 in cardiomyocyte differentiation). DE genes were defined by using a threshold of adjusted p-value < 0.05 .

For every pair of timepoints, we tested the enrichment of these DE genes within regions

of interest (**Figure 3.5A**): (A) regions near (i.e., within 100kb) significantly differential boundaries (sigDB), (B) regions within a TAD with at least one sigDB, and (C) regions within significantly differential compartmental regions (sigDC). For (B), we define all regions bounded within a pair of shared boundaries and containing at least one sigDB within those bounds as belonging to a “TAD with at least one sigDB”.

The fold enrichment of DE genes in these regions was computed as $\frac{q}{M} / \frac{s}{N}$, where N is number of all regions, $s = |\text{set of regions with at least one DE gene}|$, $M = |\text{a subset regions of interest as defined above, e.g., regions near sigDB}|$, $q = |\text{regions of interest with at least one DE gene, e.g., regions near sigDB with a DE gene}|$. We also performed gene-centric fold enrichment calculations: $\frac{q_g}{M_g} / \frac{s_g}{N_g}$, where N_g is total number of genes with expression, $s_g = |\text{DE genes}|$, $M_g = |\text{genes overlapping with a regions of interest, e.g., region near sigDB}|$, $q_g = |\text{DE genes overlapping with a regions of interest}|$. Hypergeometric test was additionally performed to calculate the significance of this fold enrichment value for each pair of timepoints.

Gene ontology (GO) term enrichment analysis was performed for two different subsets of genes based on their DE status and whether they were close to (within 100kb of) sigDB: (1) DE genes not close to a sigDB, (2) DE genes close to a sigDB. The significance of enrichment was determined with an FDR-corrected hypergeometric test p-value < 0.05 . To select candidate differential boundaries for visualization, we ranked a sigDB based on two criteria: (1) adjusted p-value of the change in TGIF-DB boundary score, and (2) the significance of the nearby differential expression measured by the nearest DE gene's adjusted p-value. We converted these values into ranks and used the mean rank of a boundary to select top 10 regions with promising differential boundaries.

3.3.9 SNP enrichment within TGIF boundaries from cardiomyocyte differentiation data

We downloaded SNPs in the GWAS catalog (Sollis et al., 2023) and mapped each SNP's associated trait to its parent phenotype, based on Experimental Factor Ontology (EFO). We refer to these parent terms as SNP categories in our analysis. In total we had 17 such categories (e.g. cardiovascular disease) for which we tested enrichment of SNPs in TGIF-DB boundaries. For each category, we calculated the fold enrichment of associated SNPs in different subsets of TGIF-DB boundaries across different timepoints: boundaries found in a specific timepoint, boundaries found in the first two or the last two timepoints, boundaries found across all timepoints (ALL, **Figure 3.7A**), boundaries found in any of the timepoints (ANY, **Figure 3.7A**). We used the following formula to calculate fold enrichment: $\frac{q}{M} / \frac{s}{N}$. Here, q is the number of boundaries of a particular type (e.g. ANY) with at least one SNP of interest, M is the number of boundaries of a particular type (e.g. ANY), s is the number of regions containing at least one SNP, and N is the total number of genomic regions.

3.4 Implementation and availability

TGIF-DC and TGIF-DB are available at <https://github.com/Roy-lab/tgif>. All scripts used for evaluation, analysis, and visualization have been deposited at <https://doi.org/10.5281/zenodo.13323899>.

3.5 Discussion

Systematic characterization of the dynamics of three dimensional genome organization is important to improve our understanding of how this layer of regulation impacts phenotypic and molecular changes across different biological contexts, such as species, time, and developmental stage. Advances in genomic tools and concerted consortia-level efforts have produced a growing compendia of high-throughput chromosome conformation capture datasets (Dekker et al., 2017, 2023; Reiff et al., 2022). However, systematic analysis of these datasets to quantify the extent of change is a challenge because of the multiple layers at which the 3D genome is organized and the paucity of tools to analyze datasets from a large number of contexts. To address this challenge, we developed Tree-guided Integrated Factorization (TGIF) that combines multitask learning with matrix factorization to examine dynamics of 3D genome organization across multiple structural scales and biological conditions.

TGIF's design is motivated by a number of considerations: (a) TAD and compartment identification are unsupervised learning problems with no ground truth for real Hi-C datasets. Since Hi-C data can be sparse, identification of such structures and assessing how much they change could be susceptible to statistical, non-biological differences. (b) Several studies from multiple cell types, time points, and species have shown that TAD and compartment is conserved across species (Dixon et al., 2012; Vietri Rudan et al., 2015). TGIF's hierarchical, multi-task learning framework exploits this prior information to constrain the identification of organizational structures while being sensitive to the extent of relatedness of the datasets by using a tree structure. Thus datasets that are further apart would be constrained less to share similar TAD structure compared to more closely related ones. (c) Finally, TGIF is motivated by a dimensionality reduction (matrix factorization) framework to reduce the noisy, high-dimensional count profile of each genomic locus into

a low dimensional space of different ranks. This enables TGIF to be a general framework that identifies TADs, compartments, as well subcompartments and their dynamics.

In contrast, existing approaches to identifying differences across multiple Hi-C datasets involve defining structural units of interest independently in each dataset, followed by a post-hoc comparison, or considering pairwise differences. Application of TGIF and existing methods to three mammalian differentiation time course datasets showed that TGIF can accurately recover structural units such as compartments and topologically associated domains (TADs), while being robust to technical differences between datasets such as depth, normalization, and resolution. TGIF also identifies biologically meaningful differences in 3D genome organization that are supported by numerous one-dimensional features such as architectural protein enrichment, histone modification, and differential expression. By allowing users to specify the extent to which the datasets are related, TGIF does not overly impose similarity between datasets. In situations where the datasets are not closely related, we expect TGIF-DB and TGIF-DC to perform similarly to methods that detect TADs (or compartments) per condition and identify differential boundaries and compartments as a post-processing step.

An open question with topological domain changes is how they relate to changes in gene expression (Greenwald et al., 2019; Ghavi-Helm et al., 2019; Cavalleiro et al., 2021; McArthur and Capra, 2021). At the TAD level, fusion or inversion of TADs could result in gene expression change although the extent to which such changes are genome wide or are specific to disease-associated genes is still unclear (Cavalleiro et al., 2021). Evidence suggests that RNA polymerase elongation or the binding of pre-initiation complex to the DNA during transcription can give rise to domain structures, providing a direct mechanistic link between transcription and 3D genome organization (van Steensel and Furlong, 2019; Heinz et al., 2018). This relationship can further depend upon the developmental stage or differentiation status of cells (Pollex et al., 2024; Chen et al., 2024). However, this has

been debated in other studies, for example, during *Drosophila* development (Ing-Simmons et al., 2021; Espinola et al., 2021).

Using multi-sample mammalian datasets, we examined the propensity of differentially expressed genes to be close to differential boundaries and compartments. The enrichment of differentially expressed genes near differential boundaries is indicative of the impact of TAD changes to gene expression changes; furthermore, DE genes that were near differential boundaries were more significantly enriched for context-specific processes which could indicate that such changes are associated with fine tuning of gene expression during cellular differentiation. The types of TAD changes we investigated were gain and loss of boundaries between pairs of time points and thus the expansion or contraction of a particular domain. Finally, we observe a similar trend to hold for regions participating in differential compartments, though to a lesser extent than TAD changes. Follow up experiments that perturb boundaries and compartment structures coupled with gene expression measurements would be beneficial for teasing apart causal versus correlational relationships between chromatin organization and gene expression changes.

Regulatory sequence variants can mis-regulate gene expression by disrupting TAD boundaries (Lupiáñez et al., 2015; Chakraborty and Ay, 2019). We used our TAD boundaries to examine the impact of this variation. Interestingly, when considering different types of boundaries based on whether they were time-point specific, or conserved to different extents, we found the greatest enrichment in boundaries that did not change over time, namely the persistent boundaries. Furthermore, we found several cardiovascular and metabolic disease trait SNPs to be enriched in these boundaries. These persistent boundaries may be specific to the entire cardiac tissue as a whole rather than a specific developmental time or stage. As future work, it would be worth investigating persistent boundaries in other developmental lineages and their propensity to prioritize SNPs for diseases in tissue-specific manner. Additionally, this provides a way to prioritize variants for downstream

functional experiments that could be important to identify the mechanisms by which variants disrupt gene regulatory processes.

There are a number of directions in which TGIF could be extended. TGIF expects the relatedness of the datasets to be provided as user input. This information may be available for specific problem settings or can be estimated from the input count matrices in a data-driven way. However, the same hyper-parameter is used for all branches of the tree. For situations in which a more granular control is needed between datasets, a direction of future research is to consider the hyper-parameter to vary depending upon the position in the hierarchy. Such information could be informed by auxiliary information such as phylogenetic branch length across species or gene expression similarity across cell types. A second direction of research is to consider auxiliary measurements, including sequence, to inform the inference of the topological units using techniques such as semi-supervised clustering (Bair, 2013; Bondell and Reich, 2008).

Overall, TGIF is a flexible and robust framework to examine changes in genome organization at the compartment and TAD level across a large number of Hi-C datasets. As more datasets across diverse biological contexts become available, methods like TGIF are expected to be increasingly helpful to examine 3D genome organization dynamics and its impact on normal and disease processes.

3.6 Acknowledgements

This work is supported by the National Institutes of Health (NIH) through the grant NIH NHGRI R01-HG010045-01 and by the Computation and Informatics in Biology and Medicine (CIBM) training program (NLM 5T15LM007359). We thank the Center for High Throughput Computing at University of Wisconsin - Madison for computational resources. We also thank Yanxiao Zhang and Bing Ren for providing the list of HERV-H

retrotransposon site coordinates and their expression levels.

Chapter 4

Analysis of evolutionarily conserved features in 3D genome organization with multi-task matrix factorization

4.1 Introduction

One of the key building blocks of 3D genome organization is topological associating domains (TADs), which represent a stretch of neighboring genomic regions with high degree of contact and interactions in 3D space as captured by Hi-C technologies (Bonev and Cavalli, 2016; Rowley and Corces, 2018; Szabo et al., 2019). Understanding the conservation of architectural protein function across species or the emergence of species-specific sequence elements that enforce TADs can provide a link between 3D genome organization and evolution. TAD boundaries, which insulate neighboring domains, are enriched for binding of conserved architectural proteins such as CTCF (Bell et al., 1999; Xie et al., 2007), tend to be evolutionarily constrained from sequence variation, and provide breakpoints for syntenic blocks (Vietri Rudan et al., 2015; Harmston et al., 2017; Lazar et al., 2018; Krefting

et al., 2018; McArthur and Capra, 2021). On the other hand, species-specific boundary elements have also been validated through Hi-C experiments. For example, Human-specific Endogenous Retrovirus H (HERV-H), a retrotransposon element that was incorporated into primate genome about 35 million years ago, form boundaries in human pluripotent cells when transcriptionally active (Zhang et al., 2019; Sexton et al., 2022). However, its boundary-demarcating role dissipates in evolutionarily distant species from humans, as observed in non-ape species' Hi-C data (Zhang et al., 2019). A systematic approach for analyzing such loci of interest across species, both in terms of its sequence and structural conservation and divergence could provide insight for how the evolution of 3D genome contributes to the evolution of gene regulation across species.

In the boundary-related analysis of conservation or divergence across species, sequence-based approaches using LiftOver remain dominant (Zhang et al., 2019; Luo et al., 2021; Zemke et al., 2023; Keough et al., 2023). LiftOver is a tool for aligning sequence from the query species to the target species (Hinrichs et al., 2006). A typical workflow entails first identifying TAD boundaries (or sequence elements and protein binding sites associated with boundary formation) in each species of interest, and lifting over the boundary region from one species to another while accounting for sequence divergence, gaps, and breakage. If a boundary is found in the lifted-over region in the target species, the boundary is treated as conserved in the query and target species; otherwise, it is typically annotated as species-specific. Here we demonstrate a similar pipeline using TGIF, which we use to identify TAD boundaries at and around a locus of interest for multiple species, accompanied by a comparative analysis of boundary evolutionary dynamics using sequence and structure conservation. We apply our pipeline to 2mbp regions centered around CTCF, a known, conserved boundary element. This serves as an exploratory analysis on multi-species Hi-C data collected from the aortic endothelial cells (AEC) of human, rat, pig, cow, and dog. AEC found in the vascular endothelium is the key site of leukocyte traffic and regulation during

inflammatory processes; their dysfunction is implicated in heart attacks, stroke, and other coronary artery diseases (Hajra et al., 2000; Adelus et al., 2024). As the endothelium is a key feature of vertebrate blood vascular systems but is entirely absent in invertebrates, studying its conserved cellular and genetic characteristics is also of interest from the vertebrate evolutionary standpoint (Monahan-Earley et al., 2013). To explore the conservation of 3D genome organization within AECs across species, we measure the sequence similarity in the neighborhood of CTCF peaks found in human AEC and their mapped regions in other species, as well as the structural similarity based on the local Hi-C counts in the neighborhood. Finally, we apply a simplified version of TGIF across species in these neighborhoods to look for and quantify the conservation pattern of boundaries marked by CTCF in AEC across species. Our analysis shows that while sequence similarity in the vicinity of CTCF binding sites across species is low, we can identify conserved boundary structures with our multi-task matrix factorization approach.

4.2 Methods

4.2.1 Single-window TGIF

Single-window TGIF (swTGIF) is a simplified version of TGIF-DB ([section 3.3.2](#)), which identifies shared and differential boundaries from a set of input Hi-C matrices from related biological conditions. Just like TGIF-DB, swTGIF takes as input a tree encoding the relationship among the input datasets or conditions and a set of Hi-C matrices for each leaf node ([Figure 4.1A](#)). While TGIF-DB typically takes in entire chromosomal matrices and slides a window down the diagonal to extract significant TAD boundaries, swTGIF directly factorizes the potentially smaller input matrices without further breaking it down to block-diagonal submatrices. The rest of the framework is exactly the same: $k \in \{2, \dots, 8\}$ is scanned to yield the mean boundary scores from the resulting factors ([Figure 4.1B](#)). A randomized “background” matrix is generated and used in the same manner as in TGIF-DB, with the background scores from the randomized matrix also used to calculate the empirical p-value of each boundary score and to call significant boundaries ([Figure 4.1C](#)).

4.2.2 Hi-C and CTCF ChIP-seq data from aortic endothelial cell of 5 mammalian species

Our collaborators from Dr. Michael Wilson’s team at the Hospital for Sick Children in Toronto collected in situ Hi-C data from aortic endothelial cells (AEC) of human, rat, pig, cow, and dog (manuscript in preparation). Two biological replicates were collected from each species. For each species, reads from the two replicates were merged and aligned to Ensembl release 102 reference genomes with HiC-Pro (Cunningham et al., 2022; Servant et al., 2015) to generate 10kbp-resolution count matrices. These matrices were normalized by HiC-Pro with ICE method prior to boundary analysis. CTCF ChIPseq data was also

collected for AEC of the same species, and peaks were called using MACS2 (Zhang et al., 2008). To collapse the peaks across two biological replicates, the intersection of peaks was obtained using BEDTools (Quinlan and Hall, 2010).

4.2.3 Assessing across-species sequence similarity in the vicinity of human CTCF peaks

Multi-way whole genome alignment across our 5 species of interest (human, rat, pig, cow, dog) and mouse was performed with Progressive Cactus using the same reference genomes as in the Hi-C data (Armstrong et al., 2020). For every pair of query and target species among our species of interest (e.g. human and pig), their pairwise alignment was extracted in both directions (human sequence aligned to pig's, pig's aligned to human's) from the Progressive Cactus outputs. Uniform-sized 10kbp bins in the query species were mapped to the target species using LiftOver (Hinrichs et al., 2006) with sequence overlap threshold of 0.5 (described in detail below). In the target species, each stretch of mapped regions was overlapped with its own uniform-sized 10kbp bins using BEDTools; repeating this process with the target and query reversed (e.g. pig to human this time) completes the 10kbp-bin-level mapping between the two species.

Sequence overlap is calculated as: basepair length of lifted-over region in target species/-target uniform-sized bin size (**Figure 4.2**). It is calculated in both direction for a pair of species, followed by computing the mean. For example, if 10kbp bin A in human was mapped to 6000bps in pig which in turn completely overlaps with bin B in pig, the sequence overlap from human to pig for bin A and B is 0.6; if bin B maps to 5000bps in human which completely overlaps with bin A, the sequence overlap from pig to human for bin A and B is 0.5; finally, the mean 0.55 is treated as the sequence overlap between bin A and bin B.

For each 10kbp bin in human with a CTCF peak (referred here as hCB), its mapped bin

in other species with the highest mean sequence overlap was chosen for further analysis and for the application of swTGIF. The mapped bins in rat, pig, cow, dog are referred to as rMB, pMB, cMB, dMB, respectively. To account for inversion, when we extract 2mbp windows surrounding the hCB and the mapped region, we invert the sequence in the mapped species if the mean sequence overlap in the window is higher in inverse direction. Then for every hCB and mapped regions, we extract the 2mbp-sized Hi-C submatrices centered at hCB and the mapped regions: $X_h[a : b, a : b]$, $a = i - 1000000/10000$, $b = i + 1000000/10000$ and $X_s[c : d, c : d]$, $c = j - 1000000/10000$, $d = j + 1000000/10000$, where X_h is the full human intra-chromosomal Hi-C matrix for the chromosome in which the hCB is located; i is the bin index for the hCB; X_s is the full intra-chromosomal Hi-C matrix in species s for the chromosome in which the mapped region is located; and j is the bin index for the mapped region. These matrices are used as inputs to swTGIF, along with the species tree for rat, pig, cow, dog, and human.

4.2.4 Application of TGIF to multi-species AEC data

We apply single-window TGIF (swTGIF) to Hi-C count matrices from aortic endothelial cells (AEC). The input tree for swTGIF is the species tree among human, rat, pig, cow, and dog (**Figure 4.4A**). Each of the 5 leaf nodes in the tree represents the Hi-C count matrix from each species. For human, the matrix is a 2mbp window centered at a CTCF-containing bin (hCB). For other species, the 2mbp window is centered at the hCB mapped region. In total, swTGIF was applied to 27522 sets of 2mbp windows from the 5 species.

4.2.5 Measuring structural similarity of Hi-C count matrices across species

We measure structural similarity for a pair of HiC matrices using two metrics: stratum-adjusted correlation coefficient (SCC, Yang et al., 2017) and cosine similarity between swTGIF boundary scores. SCC is a weighted mean of Pearson's correlation coefficients between sets of interactions stratified by pairwise distance of the regions involved in the interaction; each correlation is weighed by the number of interactions at the given distance and the variance of the interaction counts. Cosine similarity is measured for a pair of boundary-score vectors outputted by swTGIF for the pair of species being compared.

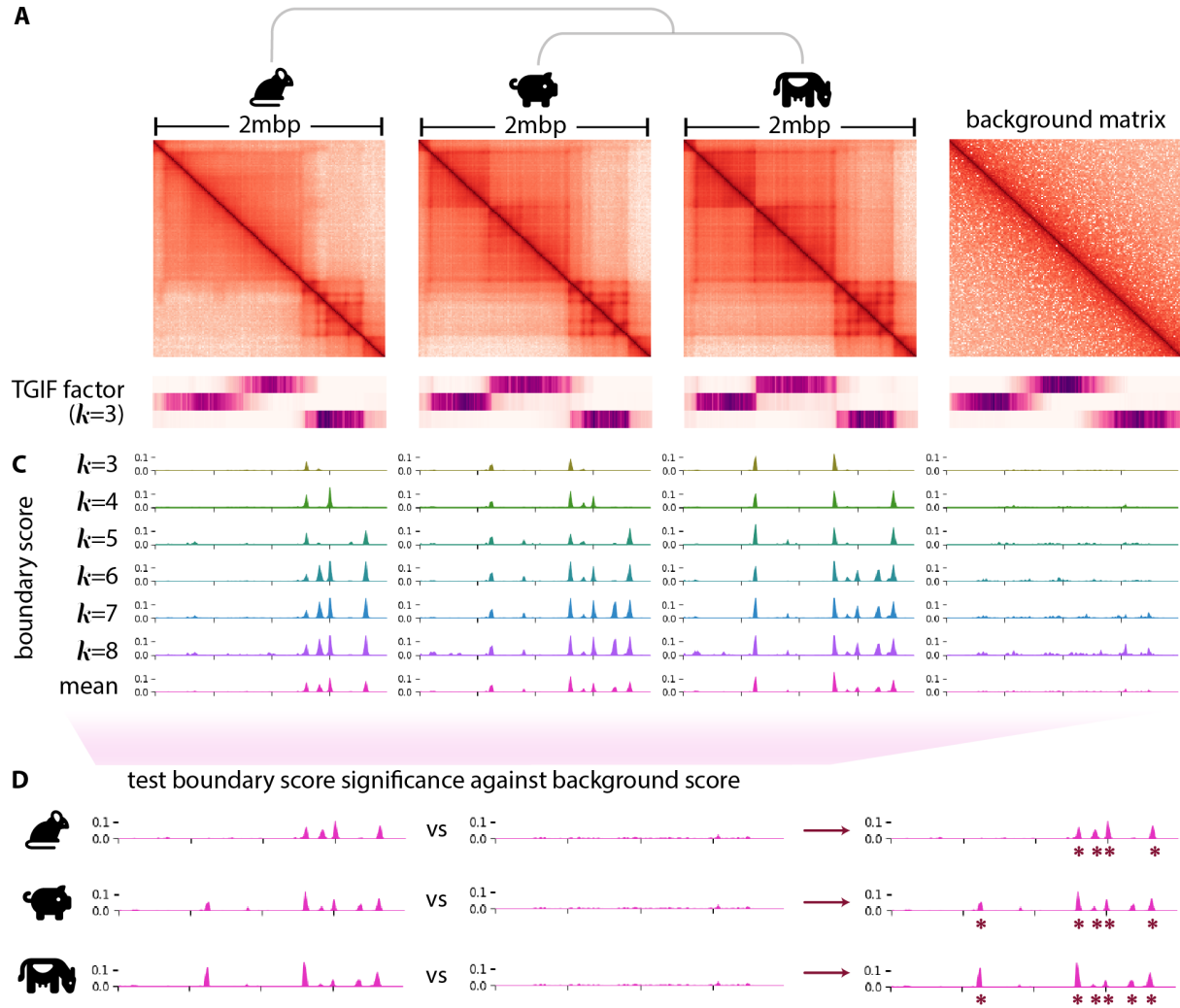
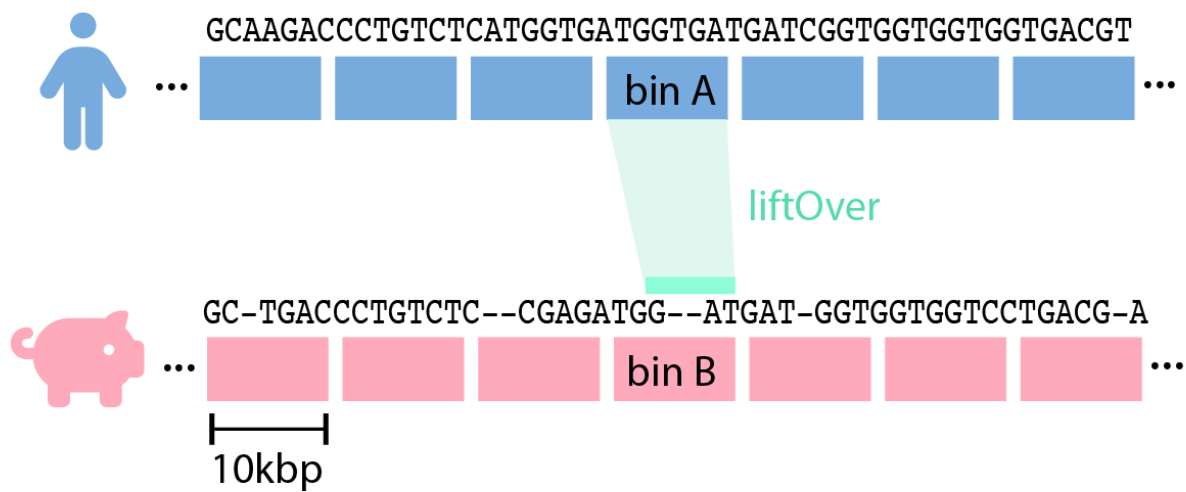


Figure 4.1: Single-window TGIF (swTGIF). **A.** The input to swTGIF consist of Hi-C matrices of fixed sizes and the tree structure encoding the relationship among the input matrices. **B.** Just as in TGIF, all matrices are factorized multiple times using a range of k . From the output factor U , the boundary score for region i is calculated as the cosine distance between its latent representation ($U[i, :]$) and its immediate upstream neighbor's representation ($U[i - 1, :]$). The background score is derived from a randomized matrix in the same manner. **C.** We calculate the empirical p-value of each boundary score against the background scores from the randomized matrix, and after FDR correction, call the final set of significant boundaries from each input data/context (represented by the brown asterisks).



sequence overlap between A and B from human to pig =
 intersection of and bin B / size of bin B

Figure 4.2: Calculating sequence overlap between uniform-sized bins. A uniform-sized (e.g. 10kbp) bin is mapped from the query species (human) to target species (pig) using LiftOver. The sequence overlap between the query bin and the target bin is calculated using the size of the intersection between the lifted-over stretch in the target species and the target bin itself. This is repeated in the other direction (pig to human) and the mean is taken to arrive at the final sequence overlap for bin A and B between human and pig.

4.3 Results

4.3.1 Sequence similarity between species rapidly decreases with distance from human CTCF peak.

We first measure the sequence similarity between human and each other mammalian species (rat, pig, cow, dog) within windows surrounding genomic regions of interest. Each window is 2 million base pairs (mbp) in length, and is binned to 10kbp to match the resolution of the Hi-C matrices. Each 2mbp window is centered at a 10kbp bin containing at least 1 CTCF peak in human (hCB), and the corresponding 10kbp bin in the species to which the sequence is mapped (i.e. mapped bin, or rMB, pMB, cMB, dMB in rat, pig, cow, dog, respectively). For each pair of hCB and its mapped bin, we measure sequence overlap between the species (**Methods, Figure 4.3A**). We find that the mean sequence overlap between hCB and its mapped bin in each species is ~ 0.6 and it rapidly decreases in its neighboring bins as distance grows (**Figure 4.3C, top row**). Similar trend is observed even for a subset of pairs where both hCB and the mapped bin have at least 1 CTCF peak in their respective species (**Figure 4.3B, bottom row**).

4.3.2 Structural similarity across species in the neighborhood of human CTCF peak

We next apply single-window TGIF (swTGIF) to 2mbp windows (i.e. Hi-C submatrices) centered at hCB and its mapped bins in other species, and measure structural similarity between species with two different metrics. The species tree between human, rat, pig, cow, and dog was used as input to swTGIF (**Figure 4.3A**), with different sets of leaf-node input matrices for each of the 27522 hCBs mapped to all 4 species (**Figure 4.3B**). We measure stratum-adjusted correlation coefficient (SCC, **Methods**) for each human 2mbp window

and the corresponding window in other species which were inputs to swTGIF. We find that, across all species, SCC between the Hi-C submatrix centered at hCB and the matrix centered at the mapped bin is low, with the median around 0 (**Figure 4.4B**). For the subset of pairs where both matrices were centered at a CTCF-containing bin, a higher proportion of them have higher SCC. When we measure the similarity of the boundary scores outputted by swTGIF between human and other species with cosine similarity, we find that pig and cow have more similar boundary scores with human than dog and rat, regardless of the presence of CTCF-peak at the center of each mapped window (**Figure 4.4C**).

4.3.3 Identification of conserved boundaries with swTGIF

Finally, we identify swTGIF significant and conserved boundaries at the center of each input Hi-C matrices, i.e. hCB and the mapped bins. Out of 27522 total input windows or matrices, we find that all 5 species have similar number of significant boundaries found at the center; ≈ 9000 out of 27522 have a significant boundary found at the hCBs or the mapped bins (\pm one 10kbp bin up- and downstream, **Figure 4.5A**). 3622 of them, or slightly more than a third of such boundaries in each species, is conserved across all 5 species. We visualize one such boundary in **Figure 4.5B**: the 2mbp window in human is centered at the 10kbp bin starting at chr1 117060000 which contains a CTCF peak and a significant boundary according to swTGIF. The hCB's mapped bin in all 4 other species are also found to have a swTGIF significant boundary.

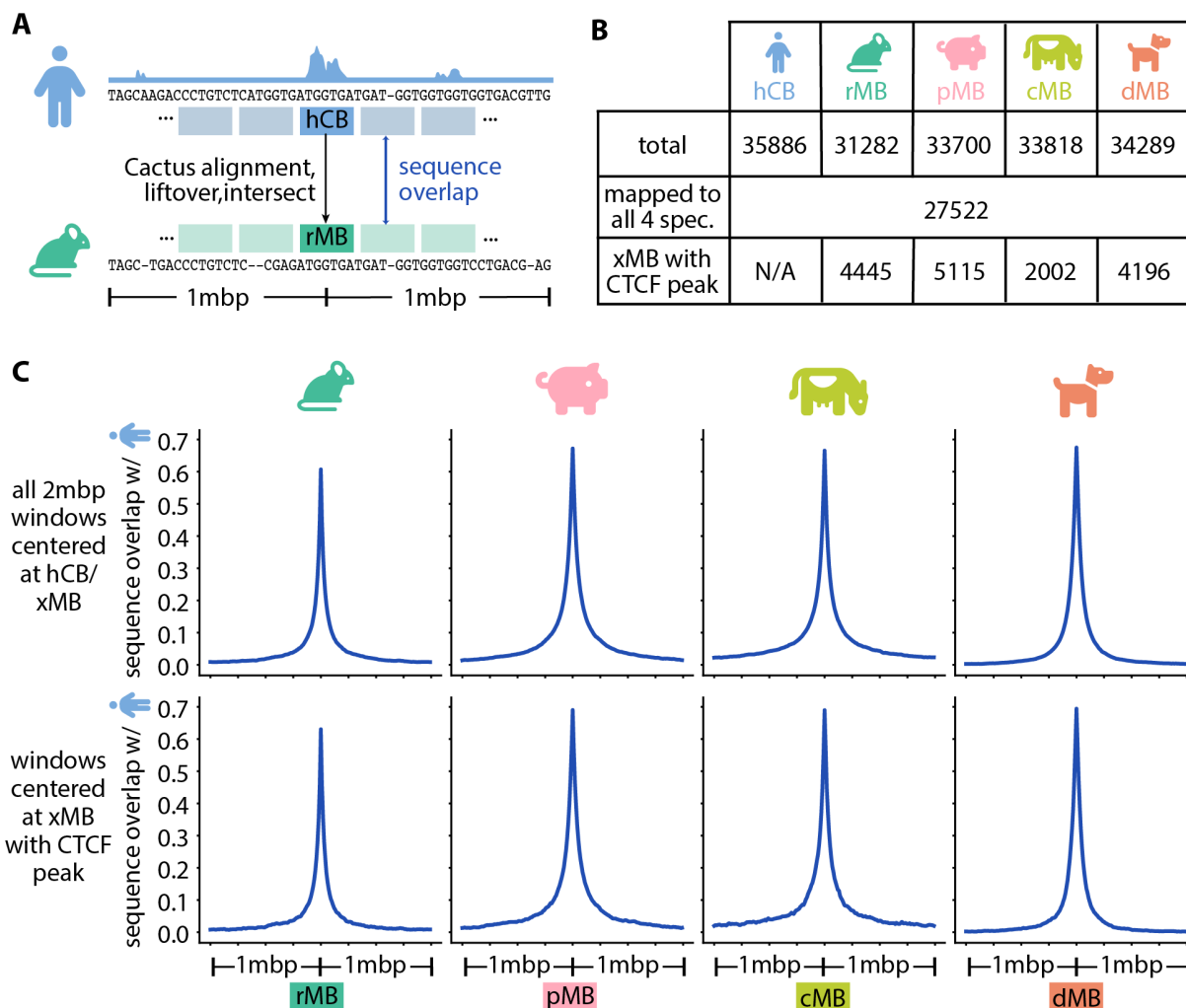


Figure 4.3: Sequence overlap between species around human CTCF peak. **A.** Each CTCF-peak-containing bin in human (hCB) is mapped to a 10kbp bin in another species using Cactus alignment and liftover. For each pair of 10kbp bin in human (including hCB) and its mapped bin, we measure sequence overlap between the species (Methods). **B.** Total count of hCBs and their mapped bins in rat, pig, cow, and dog (rMB, pMB, cMB, and dMB, respectively); count of hCBs mapped in all 4 species; and count of mapped bins also with at least 1 CTCF peak in the given species. **C.** Mean sequence overlap between human 10kbp bin and the mapped bin in rat, pig, cow, and dog, within 2mbp windows centered at hCB and corresponding mapped bin. Top row: all windows; bottom row: windows centered at hCB and mapped bin with also with CTCF peak.

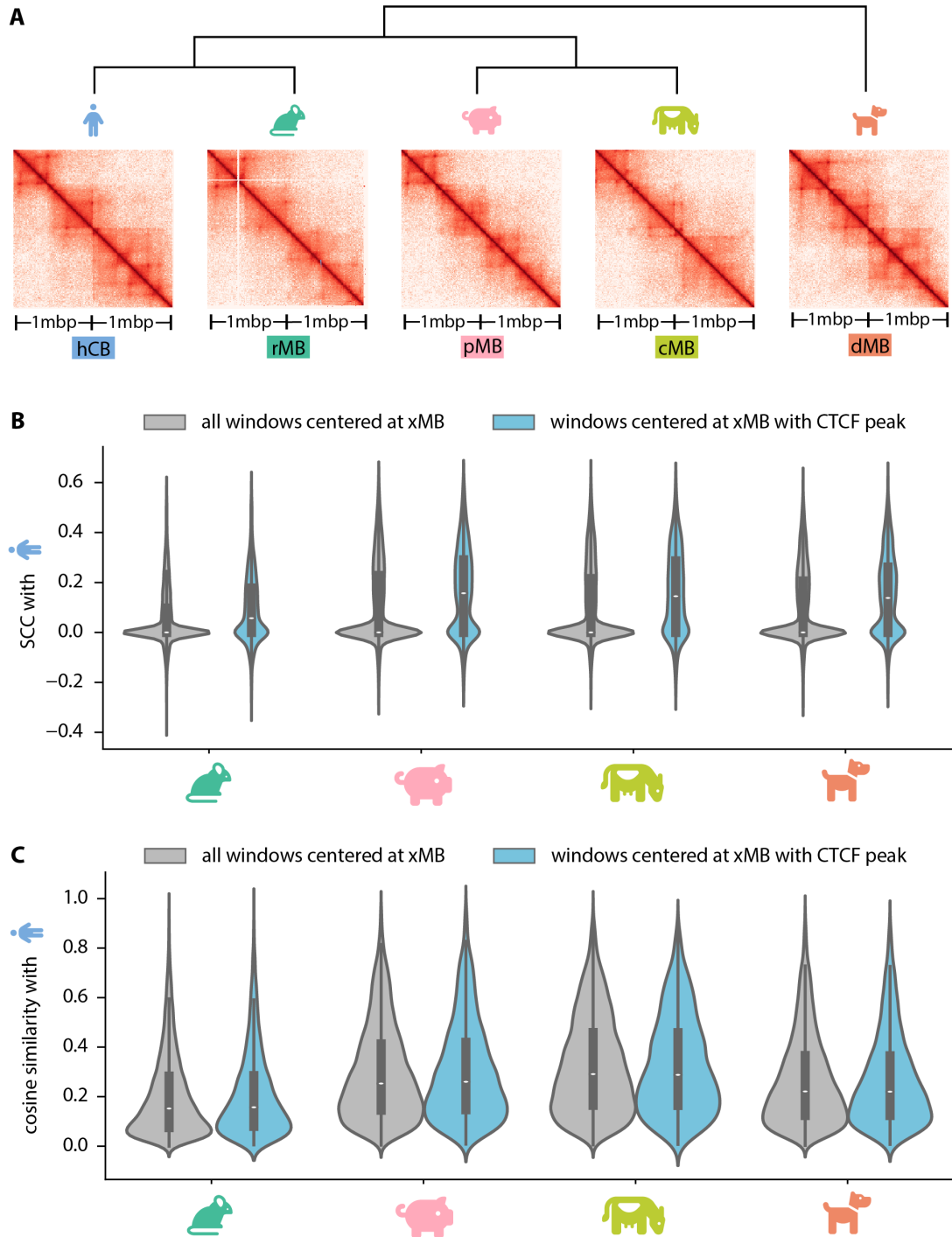


Figure 4.4: Structural similarity between species around human CTCF peak. **A.** Input species tree to swTGIF. **B.** Stratum-adjusted correlation coefficient (SCC) between human Hi-C submatrices centered at hCB and submatrices centered at the mapped bin in other species. **C.** Cosine similarity between the swTGIF output boundary scores of human and other species.

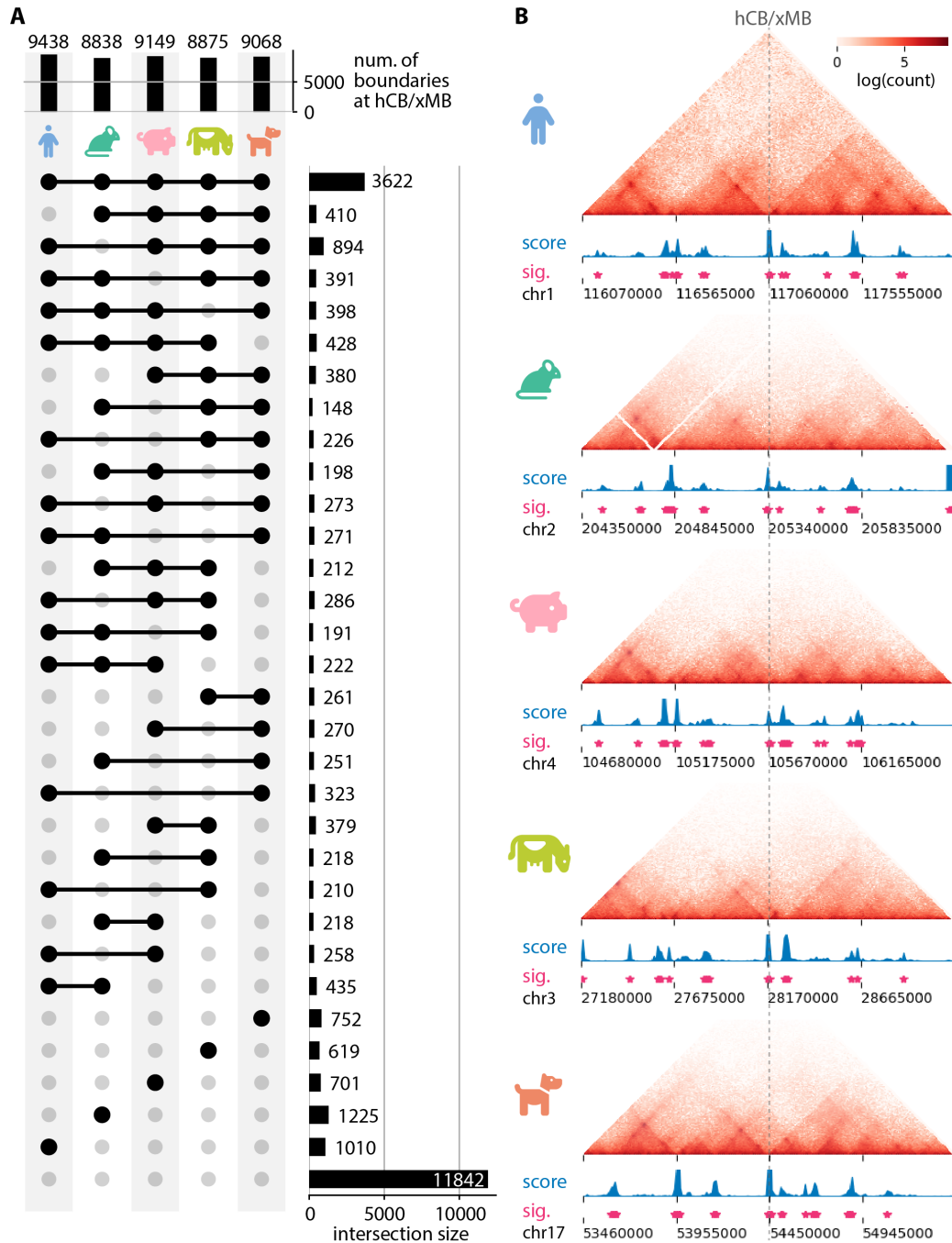


Figure 4.5: Identification of conserved boundaries with swTGIF. **A.** Count of swTGIF significant boundaries at center of input Hi-C matrices, i.e., hCB or the mapped bins \pm one 10kbp bin up- and downstream. **B.** Visualization of a conserved boundary across 5 species (human, rat, pig, cow, dog). At the center is the human CTCF-containing bin (hCB) or the corresponding mapped bin (rMB, pMB, cMB, dMB). The heatmap displays the log Hi-C counts in the 2mbp window surrounding the hCB and the mapped bin; the blue lineplot visualizes the swTGIF boundary score; the pink asterisks mark 10kbp bins with significant boundary scores.

4.4 Discussion

We have applied single-window TGIF, a multi-task matrix factorization method, to Hi-C count matrices representing 2mbp-sized windows centered at human CTCF peaks and their mapped regions in rat, pig, cow, and dog. We found that: (1) sequence similarity decays rapidly as the genomic distance from the center CTCF peak (and corresponding mapped regions) grows, (2) the structural similarity based on the interactions in the neighborhood increases in cases where the mapped region also has a CTCF peak in its respective species, and (3) about a third of boundaries found at the center of the 2mbp windows in each species is conserved in all species of interest.

There is inherent limitation to such boundary-level analysis centered at pre-determined loci of interest: namely that there is no new biological insight that can be gained, such as identification of novel boundary elements that are species-specific or conserved. Furthermore, boundary conservation (or divergence) does not necessarily mean the entire domain that the boundary demarcates is also conserved or divergent; more powerful domain-level alignment tools are needed for such analysis. Since only using sequence alignment to map megabase-scale structures like TADS across species is a challenge, we have started developing an alternate approach incorporating syntenic blocks, small-scale sequence similarity at 10kbp-bin resolution, and a manifold alignment framework (Wang and Mahadevan, 2009) that can simultaneously identify TAD-like structures with near 1-to-1 mapping across multiple species.

Despite the limitations discussed above, we can use the CTCF-peak-centered analysis as the 'baseline' pattern of conservation to which other putative species-specific or conserved boundary elements can be tested for significance. We plan to also further characterize other epigenetic signatures and gene expression patterns near conserved CTCF peaks and boundaries identified by single-window TGIF.

Taken together, single-window TGIF and subsequent sequence-based and structural analysis provide a framework to quantify the degree of conservation for genomic elements involved in organizing the 3D genome.

4.5 Acknowledgments

This work is supported by the National Institutes of Health (NIH) through the grant R01HG012349 and H.I Romnes Faculty Fellowship, both awarded to Sushmita Roy. Liangxi (Dale) Wong, Mohamed Hawash, Sana Alvi, and Michael Wilson at the Hospital for Sick Children in Toronto performed the Hi-C and CTCF ChIPseq experiments, processed the Hi-C and ChIPseq data, and provided documentation on the experimental procedure and the processing pipeline. Sara Knaack and Prakriti Garg performed the Progressive Cactus alignment, pairwise LiftOver, and bin-level sequence overlap calculations, and provided documentation on the processing pipeline. Sushmita Roy guided and provided feedback on the overall approach and the analysis.

Chapter 5

Multi-task matrix factorization leveraging gene orthology for cross-species integration of single-cell transcriptomics data

5.1 Introduction

Single-cell transcriptomics technologies, such as single-cell or single-nucleus RNAseq (scRNAseq, snRNAseq, respectively), measure transcript levels in individual cells, and reveal the diversity of gene expression programs across a mix of heterogeneous cell subpopulations (Aldridge and Teichmann, 2020; Jovic et al., 2022). Among the key tasks in analyzing single-cell transcriptomics (shortened as scRNAseq) data are: (1) clustering cells with similar expression patterns, and (2) identifying sets of genes driving expression in each cell cluster for further annotation (Kiselev et al., 2019; Luecken and Theis, 2019). Non-negative matrix factorization (NMF) is well-suited to handle both tasks by simultaneously learning

the cell and gene embeddings in the form of factor matrices (Stein-O'Brien et al., 2018; Iyer et al., 2022). The resulting factors or embeddings can be used to co-cluster the cells and genes. Multi-task variations of NMF are already in use for integrating and comparing scRNAseq data from multiple samples, experiments, and biomedical contexts (Ryu et al., 2023, **section 1.2**).

Cross-species analysis of single-cell transcriptomics data typically employs integration methods meant to address batch effects, or to reduce spurious segregation of cells (and genes) due to different experimental protocols and other technical artifacts (Luecken et al., 2022; Song et al., 2023). Two key challenges face batch-correction methods deployed for species-integration task: stronger "batch effects" may be found among data from multiple species, and over-correction may reduce the valid signals of heterogeneity within each species. A benchmarking study of batch-correction methods for species integration (Song et al., 2023) found that incorporating gene lineage information (e.g. paralogs) can help integration over large evolutionary distance. SAMap, a method specifically designed for species integration using a gene-gene graph based on sequence similarity (Tarashansky et al., 2021), was found to be effective in whole-atlas-scale integration across species with complex gene lineages. However, SAMap is computationally expensive as it specifically targets integration of entire cell atlases (Song et al., 2023).

Here we propose a multi-task matrix factorization approach called TIMBER, intended for cross-species integration utilizing gene orthology information. TIMBER is an extension of TGIF (**section 3.3.1**) which uses tree-based regularization of the factors representing the feature embedding, or the gene embedding in our primary use case here. While TGIF requires the same number of column entities or features across all input datasets, TIMBER can handle multiple input datasets with varying number of features or genes. This is accomplished by mapping the features in the child node to the features in the parent node, i.e., putative ancestral genes. We use orthology groups (orthogroups) of genes for the

mapping of features from the leaf nodes of a species tree (extant species) to the root node (last common ancestor).

We apply TIMBER to scRNAseq data from three plant species (maize, medicago, and sorghum) to study their nitrogen fixing process. Nitrogen is a crucial nutrient for plant growth, often supplemented by a fertilizer (Graham and Vance, 2000; Reinprecht et al., 2020). As heavy fertilizer usage is expensive and comes with environmental costs as well, a better understanding of plants' inherent nitrogen-fixing process could lead to enhancing its efficiency (Graham and Vance, 2000; Soumare et al., 2020). Sorghum in particular produces mucilage in its aerial roots after rain and water treatment; the mucilage is a host to symbiotic bacteria that can capture nitrogen from the air, which can then be utilized by the plant (Hara et al., 2019; Wolf et al., 2024). Our goal is to understand both the conserved program behind nitrogen fixation across the three species, and to identify sorghum-specific gene expression modules and underlying regulatory programs giving rise to such distinct nitrogen-fixing strategy.

In the subsequent sections, we go over the TIMBER framework in detail and the preliminary results from its application to the multi-species plants data.

5.2 Methods

5.2.1 Tree-guided integrated matrix factorization with branch-specific regularization (TIMBER)

TIMBER is a multi-task matrix factorization approach that can simultaneously factorize multiple input matrices. The relationship among the input matrices is represented as a tree, with closely-related conditions and their corresponding input matrices sharing a parent. For example, two species sharing the same ancestral species can be represented as the child nodes to the ancestral species' parent node in the tree. The factor embedding for the column entities of the input matrix (i.e. genes) in the leaf node are constrained to be similar to that of its parent node, adapting the tree-regularization scheme of TGIF ([section 3.3.1](#)). TIMBER extends TGIF by allowing different number of column entities (features/genes) among the input matrices. In the case of factorizing matrices from multiple species with variable number of genes, we use a mapping matrix to identify which parent features will "influence" the embedding of its child features. The input to TIMBER is the set of input matrices, the tree structure encoding the relationship among them, and a set of such mapping matrices. Below we formalize the objective of TIMBER, and derive its update rules.

Given $t \in \{1, \dots, T\}$ species, each with gene-expression matrix $X^{(t)} \in \mathbb{R}^{n_t \times m_t}$, and a species tree with a root node r (representing the last common ancestor) and a set of nodes $c \in \mathcal{B} \cup \mathcal{T}$ where \mathcal{B} is a set of internal (or branch) nodes $b \in \mathcal{B}$ and \mathcal{T} a set of the species-specific leaf nodes, the following is the objective that TIMBER optimizes/minimizes:

$$O = \sum_{t=1}^T \|X^{(t)} - U^{(t)}V^{(t)\top}\|_F^2 + \sum_c \alpha_{c, P_{\mathcal{A}}(c)} \|V^{(c)} - M_{c, P_{\mathcal{A}}(c)} \cdot V^{P_{\mathcal{A}}(c)}\|_F^2 \quad (5.1)$$

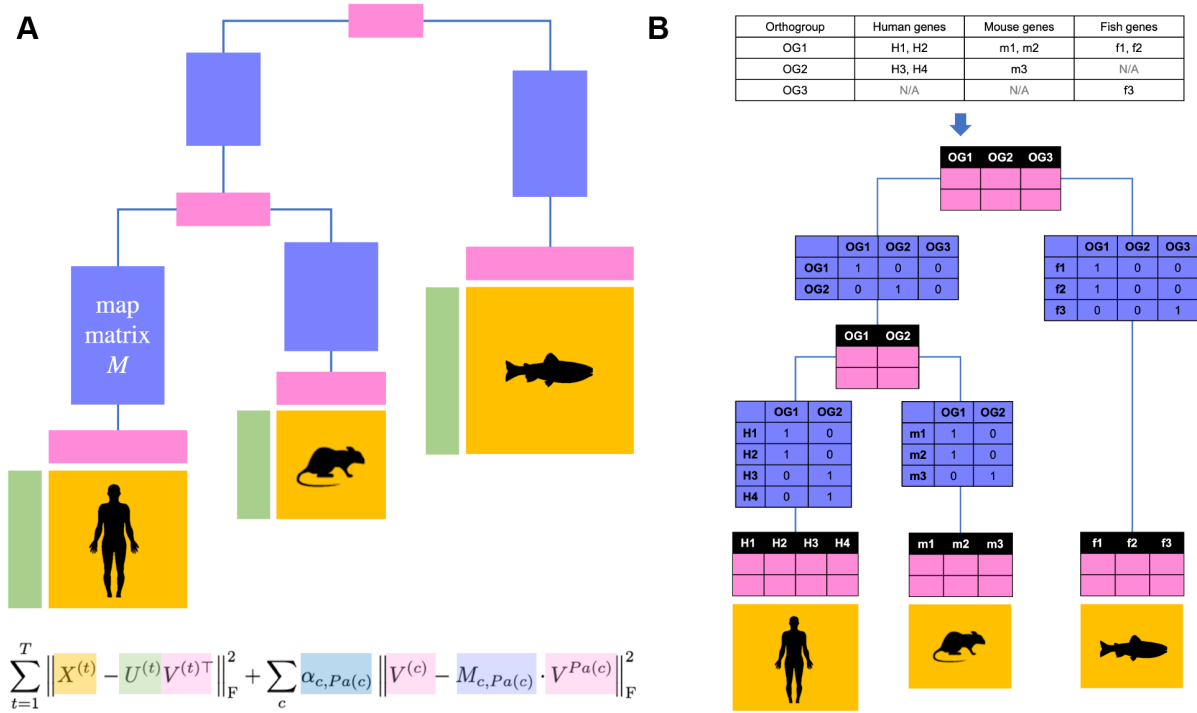


Figure 5.1: Overview of TIMBER. **A.** Visualization of TIMBER objective. Within each species t , we want to find factors $U^{(t)}$ (green) and $V^{(t)}$ (pink) such that (1) their product is a good approximation of the input matrix $X^{(t)}$ (yellow) and (2) the latent representation of the feature set (e.g. genes) is constrained to be similar to the latent representation of its parent feature set. The child node's features are mapped to its parent using the map matrix M (purple). **B.** Mapping genes to their ancestral genes using orthogroups. Given a set of orthogroups composed of genes from different species, we create map matrix M (purple) from leaf node to parent node by assigning all extant genes to the orthogroup they belong to. The orthogroup in the parent node acts as an ancestral gene/feature. The pink matrices represent the V factors in each node, and their columns correspond to extant or ancestral genes. For an internal node or the root node, its gene set is a union of its child nodes' ancestral genes.

where $U^{(t)} \in \mathbb{R}^{n_t \times k}$, $V^{(c)} \in \mathbb{R}^{m_c \times k}$, $k \ll n, m$ (**Figure 5.1A**).

$M_{c, Pa(c)} \in \mathbb{R}^{m_c \times m_{Pa(c)}}$ maps the m_c genes in species c to its ancestral genes represented in its parent node $Pa(c)$ in the species tree. The parameter $\alpha_{c, Pa(c)}$ can be set to represent the phylogenetic distance between child and ancestral species, e.g., the inverse of phylogenetic distance multiplied by a factor. When all $\alpha_{c, Pa(c)}$ values across all branches are fixed to the

same value and the mapping matrices M s are all identity matrices, TIMBER's objective reduces to TGIF's objective. We update the factors $U^{(t)}$ s and $V^{(c)}$ s based on the optimization scheme of block coordinate descent (Kim et al., 2014).

Breaking down the objective to subproblems per column of each factor The TIMBER objective in 5.1 can be written as:

$$O = \sum_{t=1}^T \left\| X^{(t)} - \sum_k u_k^{(t)} v_k^{(t)\top} \right\|_F^2 + \sum_c \sum_k \alpha_{c, Pa(c)} \left\| v_k^{(c)} - M_{c, Pa(c)} \cdot v_k^{Pa(c)} \right\|_2^2 \quad (5.2)$$

Where $u_k^{(t)} \in \mathbb{R}^{n_t}$ is the k th column vector of $U^{(t)}$ and $v_k^{(t)} \in \mathbb{R}^m$ is the k th column vector of $V^{(t)}$. Now we 'pull out' terms involving the k th column in all factors (and for convenience drop the suffix $c, Pa(c)$ from α and M but note that they are specific to a pair of parent-child species.)

$$O = \sum_{t=1}^T \left\| X^{(t)} - u_k^{(t)} v_k^{(t)\top} - \sum_{j \neq k} u_j^{(t)} v_j^{(t)\top} \right\|_F^2 \quad (5.3)$$

$$+ \sum_c \alpha \left(\left\| v_k^{(c)} - M v_k^{Pa(c)} \right\|_2^2 + \sum_{j \neq k} \left\| v_j^{(c)} - M v_j^{Pa(c)} \right\|_2^2 \right) \quad (5.4)$$

Now we'll substitute with $R_k^{(t)} = X^{(t)} - \sum_{j \neq k} u_j^{(t)} v_j^{(t)\top}$:

$$O = \sum_{t=1}^T \left\| R_k^{(t)} - u_k^{(t)} v_k^{(t)\top} \right\|_F^2 + \sum_c \alpha \left\| v_k^{(c)} - M v_k^{Pa(c)} \right\|_2^2 + \sum_c \sum_{j \neq k} \alpha \left\| v_j^{(c)} - M v_j^{Pa(c)} \right\|_2^2 \quad (5.5)$$

We can now optimize $u_k^{(t)}, v_k^{(t)}, v_k^{(b)}, v_k^{(r)}$, fixing all other parameters to be constant.

Optimize $v_k^{(t)}$ To find $v_k^{(t)}$ for each leaf node task t that minimizes the objective, we find the derivative of the objective with respect to $v_k^{(t)}$ and set it to 0, then solve. First we expand the objective into matrix multiplications ¹:

$$O = \left\| R_k^{(t)} - u_k^{(t)} v_k^{(t)\top} \right\|_F^2 + \alpha \left\| v_k^{(t)} - M v_k^{Pa(t)} \right\|_2^2 + C \quad (5.6)$$

$$= \text{Tr} \left[\left(R_k^{(t)} - u_k^{(t)} v_k^{(t)\top} \right)^\top \left(R_k^{(t)} - u_k^{(t)} v_k^{(t)\top} \right) \right] \quad (5.7)$$

$$+ \alpha \left(v_k^{(t)} - M v_k^{Pa(t)} \right)^\top \left(v_k^{(t)} - M v_k^{Pa(t)} \right) + C \quad (5.8)$$

Here C subsumes all elements of the objective that does not involve $v_k^{(t)}$ (including terms involving tasks other than t), since they will be zeroed out when the derivative is taken with respect to $v_k^{(t)}$. Now we keep expanding:

$$O = \text{Tr} \left[R_k^{(t)\top} R_k^{(t)} - 2 R_k^{(t)\top} u_k^{(t)} v_k^{(t)\top} + \left(u_k^{(t)} v_k^{(t)\top} \right)^\top \left(u_k^{(t)} v_k^{(t)\top} \right) \right] \quad (5.9)$$

$$+ \alpha \left(v_k^{(t)\top} v_k^{(t)} - 2 v_k^{(t)\top} M v_k^{Pa(t)} + v_k^{Pa(t)\top} M^\top M v_k^{Pa(t)} \right) + C \quad (5.10)$$

$$= \text{Tr} \left(R_k^{(t)\top} R_k^{(t)} \right) - 2 \text{Tr} \left(R_k^{(t)\top} u_k^{(t)} v_k^{(t)\top} \right) + \text{Tr} \left(v_k^{(t)} u_k^{(t)\top} u_k^{(t)} v_k^{(t)\top} \right) \quad (5.11)$$

$$+ \alpha v_k^{(t)\top} v_k^{(t)} - 2 \alpha v_k^{(t)\top} M v_k^{Pa(t)} + \alpha v_k^{Pa(t)\top} M^\top M v_k^{Pa(t)} + C \quad (5.12)$$

$$= \text{Tr} \left(R_k^{(t)\top} R_k^{(t)} \right) - 2 \left(R_k^{(t)\top} u_k^{(t)} \right)^\top v_k^{(t)} + \left(u_k^{(t)\top} u_k^{(t)} \right) \left(v_k^{(t)\top} v_k^{(t)} \right) \quad (5.13)$$

$$+ \alpha v_k^{(t)\top} v_k^{(t)} - 2 \alpha v_k^{(t)\top} M v_k^{Pa(t)} + \alpha v_k^{Pa(t)\top} M^\top M v_k^{Pa(t)} + C \quad (5.14)$$

¹See Section 2.7.1 Eqn (132) in Petersen and Pedersen, 2006 for expansion of Eqn (5.7).

We continue with derivation ².

$$\frac{\partial O}{\partial v_k^{(t)}} = 0 - 2R_k^{(t)\top} u_k^{(t)} + 2v_k^{(t)} u_k^{(t)\top} u_k^{(t)} + 2\alpha v_k^{(t)} - 2\alpha M v_k^{Pa(t)} + 0 + 0 \quad (5.15)$$

$$0 = -R_k^{(t)\top} u_k^{(t)} + \left(u_k^{(t)\top} u_k^{(t)} + \alpha \right) v_k^{(t)} - \alpha M v_k^{Pa(t)} \quad (5.16)$$

$$v_k^{(t)} = \frac{R_k^{(t)\top} u_k^{(t)} + \alpha M v_k^{Pa(t)}}{\left\| u_k^{(t)} \right\|_2^2 + \alpha} \quad (5.17)$$

With the non-negativity constraint $v_k^{(t)} \geq 0$, we want $R_k^{(t)\top} u_k^{(t)} + \alpha M v_k^{Pa(t)} \geq 0$, because if $R_k^{(t)\top} u_k^{(t)} + \alpha M v_k^{Pa(t)} < 0$, O will increase in Eqn (5.13, 5.14). The finalized update rule is:

$$v_k^{(t)} = \frac{\left[R_k^{(t)\top} u_k^{(t)} + \alpha M v_k^{Pa(t)} \right]_+}{\left\| u_k^{(t)} \right\|_2^2 + \alpha} \quad (5.18)$$

Optimize $u_k^{(t)}$ We can derive the update rule for $u_k^{(t)}$ in leaf node task t similarly but much more simply, in the same manner as TGIF (section 3.3.1). From Eqn (5.13), we take the derivative of O with respect to $u_k^{(t)}$. All regularization terms will zero out since they do not involve $u_k^{(t)}$. Hence the final update rule for $u_k^{(t)}$ is:

$$u_k^{(t)} = \frac{\left[R_k^{(t)} v_k^{(t)} \right]_+}{\left\| v_k^{(t)} \right\|_2^2} \quad (5.19)$$

Optimize $v_k^{(r)}$ For the root node factor loadings $v_k^{(r)}$, we can again ignore terms that do not involve $v_k^{(r)}$ in the objective, Eqn (5.5). Note that we collect the terms involving nodes c

²See Section 2.4 Eqn (69-72), Section 2.7.1 Eqn (132) in Petersen and Pedersen, 2006.

whose parent is the root node, i.e. $\text{Pa}(c) = r$:

$$O = \sum_{c \in \text{Child}(r)} \alpha_{c,r} \left\| \mathbf{v}_k^{(c)} - \mathbf{M}_{c,r} \cdot \mathbf{v}_k^{(r)} \right\|_2^2 + C \quad (5.20)$$

This is equivalent to:

$$O = \sum_{c \in \text{Child}(r)} \alpha_{c,r} \left\| \mathbf{M}_{c,r}^\top \mathbf{v}_k^{(c)} - \mathbf{v}_k^{(r)} \right\|_2^2 + C \quad (5.21)$$

$$= \sum_{c \in \text{Child}(r)} \alpha_{c,r} \left(\mathbf{M}_{c,r}^\top \mathbf{v}_k^{(c)} - \mathbf{v}_k^{(r)} \right)^\top \left(\mathbf{M}_{c,r}^\top \mathbf{v}_k^{(c)} - \mathbf{v}_k^{(r)} \right) + C \quad (5.22)$$

$$= \sum_{c \in \text{Child}(r)} \alpha_{c,r} \left[\mathbf{v}_k^{(c)\top} \mathbf{M}_{c,r} \mathbf{M}_{c,r}^\top \mathbf{v}_k^{(c)} - 2 \mathbf{M}_{c,r}^\top \mathbf{v}_k^{(c)\top} \mathbf{v}_k^{(r)} + \mathbf{v}_k^{(r)\top} \mathbf{v}_k^{(r)} \right] + C \quad (5.23)$$

$$= C - \sum_{c \in \text{Child}(r)} \alpha_{c,r} \left[2 \mathbf{M}_{c,r}^\top \mathbf{v}_k^{(c)\top} \mathbf{v}_k^{(r)} - \mathbf{v}_k^{(r)\top} \mathbf{v}_k^{(r)} \right] \quad (5.24)$$

Now we take the derivative, set to 0, and solve³:

$$\frac{\partial O}{\partial \mathbf{v}_k^{(r)}} = 0 - \sum_{c \in \text{Child}(r)} 2 \alpha_{c,r} \mathbf{M}_{c,r}^\top \mathbf{v}_k^{(c)} + \sum_{c \in \text{Child}(r)} 2 \alpha_{c,r} \mathbf{v}_k^{(r)} \quad (5.25)$$

$$0 = - \sum_{c \in \text{Child}(r)} \alpha_{c,r} \mathbf{M}_{c,r}^\top \mathbf{v}_k^{(c)} + \sum_{c \in \text{Child}(r)} \alpha_{c,r} \mathbf{v}_k^{(r)} \quad (5.26)$$

$$\mathbf{v}_k^{(r)} = \frac{\sum_{c \in \text{Child}(r)} \alpha_{c,r} \mathbf{M}_{c,r}^\top \mathbf{v}_k^{(c)}}{\sum_{c \in \text{Child}(r)} \alpha_{c,r}} \quad (5.27)$$

Optimize $\mathbf{v}_k^{(b)}$ For the internal node factor loadings $\mathbf{v}_k^{(b)}$, we ignore terms that do not involve $\mathbf{v}_k^{(b)}$ for the particular node b of interest in the objective, Eqn (5.5). Note that just as with the root node, internal nodes do not have their own observed data, i.e. \mathbf{X} . We collect terms involving the parent node of b , i.e. $\text{Pa}(b)$, and nodes c whose parent is b , i.e.

³See Section 2.4.1 Eqn (69), Section 2.6 Eqn (129-131) in Petersen and Pedersen, 2006.

$\text{Pa}(c) = b$.

$$O = \alpha_{b, \text{pa}(b)} \left\| \mathbf{v}_k^{(b)} - \mathbf{M}_{b, \text{pa}(b)} \cdot \mathbf{v}_k^{\text{pa}(b)} \right\|_2^2 + \sum_{c \in \text{Child}(b)} \alpha_{c, b} \left\| \mathbf{M}_{c, b}^\top \mathbf{v}_k^{(c)} - \mathbf{v}_k^{(b)} \right\|_2^2 + C \quad (5.28)$$

Taking the hybrid of derivations for $\mathbf{v}_k^{(r)}$ and $\mathbf{v}_k^{(t)}$ from Eqn (5.15-17) and Eqn (5.25-27), we arrive at:

$$\mathbf{v}_k^{(b)} = \frac{\alpha_{b, \text{pa}(b)} \mathbf{M}_{b, \text{pa}(b)} \mathbf{v}_k^{\text{pa}(b)} + \sum_{c \in \text{Child}(b)} \alpha_{c, b} \mathbf{M}_{c, b}^\top \mathbf{v}_k^{(c)}}{\alpha_{b, \text{pa}(b)} + \sum_{c \in \text{Child}(b)} \alpha_{c, b}} \quad (5.29)$$

5.2.2 From factors to clusters

Once we have the factors $\mathbf{U}^{(t)}$ and $\mathbf{V}^{(t)}$ for $t \in \{1, \dots, T\}$ species, we can treat them as cell embeddings and gene embeddings, respectively, i.e. $\mathbf{U}^{(t)}[i, :]$ is the embedding for cell i , and $\mathbf{V}^{(t)}[j, :]$ the embedding for gene j in species t . We can apply a clustering algorithm like k-means clustering to find cell and gene clusters from such embeddings. However, to easily find correspondence between cell clusters and gene clusters, we can simply assign each cell or gene to its most dominant factor loading. In other words, the cluster assignment c_i for cell i in species t is:

$$c_i = \text{argmax}_k \mathbf{U}^{(t)}[i, k] \quad (5.30)$$

And the cluster assignment c_j for gene j in species t :

$$c_j = \text{argmax}_k \mathbf{V}^{(t)}[j, k] \quad (5.31)$$

We refer to this as "maxdim" cluster assignment.

In order to measure the similarity of clustering results between a pair of conditions (e.g.

gene clusters from $\alpha = 100$ vs. those from $\alpha = 500$ for the same k), we use adjusted Rand index (ARI). ARI ranges from 0 to 1 with a higher value representing higher degree of concordance between the pair of clustering results.

5.2.3 From gene orthogroups to gene mapping matrices

We utilize orthogroups to map extant genes to their ancestral genes. An orthogroup refers to a set of genes descended from the same gene in the last common ancestor of all species of interest. Our collaborator Rafael Venado used Orthofinder (Emms and Kelly, 2019) to find orthogroups among 6 plant species including sorghum, maize, and medicago. The following is the list of the species and the exact genome assemblies used:

- Arabidopsis: Athaliana_447_Araport11.protein.fa
- Indian colza: Brapassp_trilocularisR500_795_v2.1.protein.fa
- Medicago: MtrunA17r5.0-ANR-EGN-r1.9.prot.fasta
- Sorghum: SbicolorRTx430_552_v2.1.protein.fa
- Tomatoes: Slycopersicum_796_ITAG5.0.protein.fa
- Maize: Zmays_833_Zm-B73-REFERENCE-NAM-5.0.55.protein.fa

From the initial 21695 orthogroups, we narrowed downed the list to those with at least one gene from sorghum, maize, and medicago. For a set of extant genes within the same orthogroup, they are mapped to the same ancestral gene, creating a binary mapping matrix to the parent species' node. For an internal node or the root node, its features or "genes" are a union of the child nodes' ancestral genes (**Figure 5.1B**). In the resulting mapping, 26131 maize genes and 21124 sorghum genes mapped to 14257 ancestral genes in the parent

species of maize and sorghum. Those ancestral genes, along with 25126 medicago genes, in turn map to a total of 16601 ancestral genes in the root of the species tree (**Figure 5.2A**).

5.2.4 Single cell RNAseq data from maize, sorghum, medicago

Our collaborator Rafael Venado collected single-nucleus RNAseq data from the aerial roots of sorghum (accession IS23992) and maize landrace (CB017456), and the underground roots of medicago. He also provided the list of marker genes for different plant tissue and cell types (e.g. border cells, epidermis). After removal of cells with low expression and normalization with PAGODA (Fan et al., 2016) by Marina Kotvanova and Saptarshi Pyne (Roy lab), the resulting expression matrices consist of 7553 cells from maize, 11526 cells from sorghum, and 4183 cells from medicago (**Figure 5.2A**).

TIMBER was applied with different $k \in \{10, 15, 20\}$ and $\alpha \in \{50, 100, 500\}$ to this dataset. We measured marker gene fold enrichment within each gene cluster for different cell types with hypergeometric test and performed multiple test correction ($FDR < 0.05$). Within each species and gene cluster, the fold enrichment of marker genes for a given cell type was calculated as $(s/M)/(q/N)$ where N = total number of genes in given species, q = number of marker genes for given cell type, M = number of genes in given cluster, and s = number of marker genes in given cluster.

5.3 Preliminary results

We applied TIMBER to scRNAseq data collected from maize, sorghum, and medicago, with different number of cells and genes within each species (**Figure 5.2A, Methods**). The genes in maize and sorghum (26131 and 21224 genes, respectively) were mapped to 14257 ancestral genes (i.e. orthogroups) in their ancestral species node. Those ancestral genes, along with 25126 medicago genes, in turn were mapped to 166001 ancestral genes or orthogroups in the root node representing the last common ancestor. TIMBER was run with different combination of parameter values: $k \in \{10, 15, 20\}$, $\alpha = \{50, 100, 500\}$. Within a single experiment, all branches of the tree used the same fixed α value, although TIMBER can handle different strengths of regularization based on the distance to similarity to the parent node (e.g. phylogenetic distance). We find that for $k = 10$ across all α values, majority of sorghum cells and genes end up with a very similar latent feature pattern, which could imply batch effect and prevents meaningful distinction among a large number of cells and genes (**Figure 5.2B,C**). As we increase k to 15 and 20, both sorghum cell and gene embeddings capture more variation and diversify.

We next project the cell embeddings in the U factors of TIMBER to 2D space using UMAP, and look for the integration of cells across species (**Figure 5.3A**). We visually confirm that, across all α values scanned and especially for $\alpha \geq 100$, specific corners of the UMAP latent space is shared by different subset of species while other corners show clustering of cells from 1 dominant species. To quantify the stability of the clustering results to changing alpha's, we calculate the cell and gene cluster similarity between pairs of alpha values using adjusted Rand index (ARI, **Figure 5.3B**). For $k = 10, 15$, both cell and gene clusters are shown to be stable across different α values with ARI values very close to 1 between most pairs of α 's compared. For $k = 20$, higher $\alpha = 500$ results in a more dissimilar set of cell and gene clusters, but overall there is still a high degree of cluster

stability with ARI above 0.9 and 0.86 for cells and genes, respectively.

We further explore the TIMBER cell clusters across species. In the UMAP projection of TIMBER cell embeddings (i.e. U factors) color-coded by cell clusters, we are able to visually discern cell clusters with different degrees of contribution from each species (**Figure 5.4A**). When we break down each cell cluster by species membership (**Figure 5.4B**), we confirm TIMBER identifies cell clusters to which a single species predominantly contributes (e.g., cluster 0, 3, 11-13), as well as cell clusters with fairly even contribution from all 3 species (e.g., cluster 1, 9, 10). We next explored how cells from each species are distributed across clusters (**Figure 5.4C**). When we hierarchically cluster the species based on cluster membership similarity (and resulting distance among them using average linkage), the resulting species hierarchy groups sorghum to be closer to medicago rather than to maize, which is a deviation from the original TIMBER input species tree; this hierarchical clustering pattern holds even with higher α . It warrants further experiments and analysis to understand how the original orthogroup-based mapping affects downstream cell cluster similarity across species.

Finally, we performed marker gene enrichment analysis for different plant cell types in each gene cluster. Within each species and each TIMBER gene cluster, we calculate the fold enrichment of marker genes for different tissues or cell types, e.g., epidermis, border cells, stele, and its significance using the hypergeometric test (**Methods**). TIMBER gene cluster 1 is significantly enriched for border cell marker genes in both sorghum and medicago, although not in maize (**Figure 5.5**). Border cells facilitate mucilage formation in aerial roots and beneficial bacterial colonization in underground roots (Pankievicz et al., 2022), both critical processes for nitrogen fixation in the aerial roots of sorghum and in the underground root nodules of medicago (Tkacz et al., 2022; Wolf et al., 2024). Further experimentation is needed to integrate the border cells of maize aerial roots (with similar nitrogen fixation strategy as sorghum) into the border cell subpopulation of sorghum and

medicago.

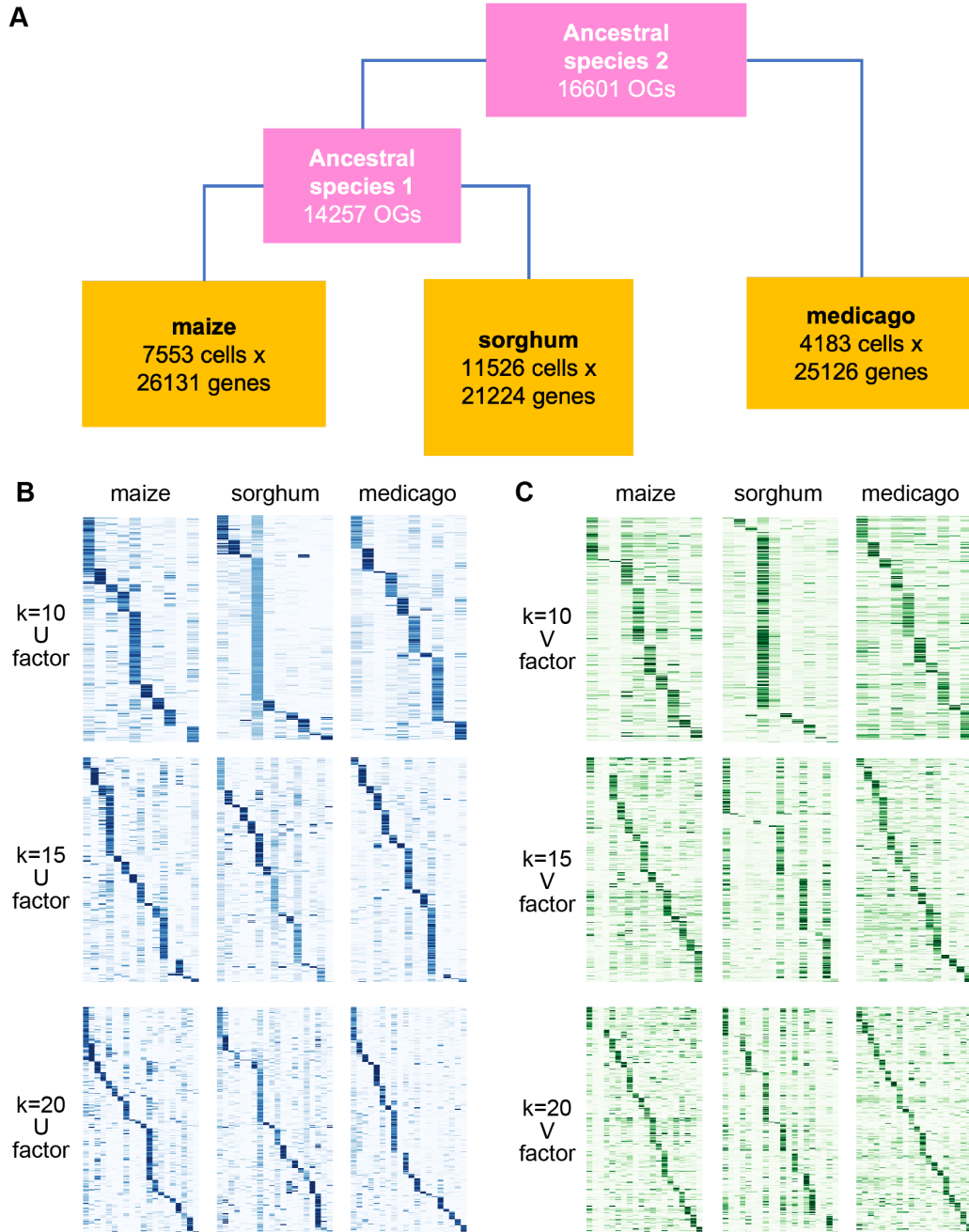


Figure 5.2: Applying TIMBER on scRNAseq data from maize, sorghum, medicago. **A.** The input species tree for TIMBER. The ancestral node for maize and sorghum have 14257 orthogroups (OGs) as ancestral gene features, to which maize and sorghum genes are mapped. The last common ancestor node similarly has 16601 OGs as ancestral gene features, to which child nodes' ancestral genes and extant genes are mapped. **B.** Heatmap visualization of the U factors, i.e. cell embeddings, for each species and for $k \in \{10, 15, 20\}$, $\alpha = 100$. **C.** Heatmap visualization of the V factors, i.e. gene embeddings, for each species and for $k \in \{10, 15, 20\}$, $\alpha = 100$.

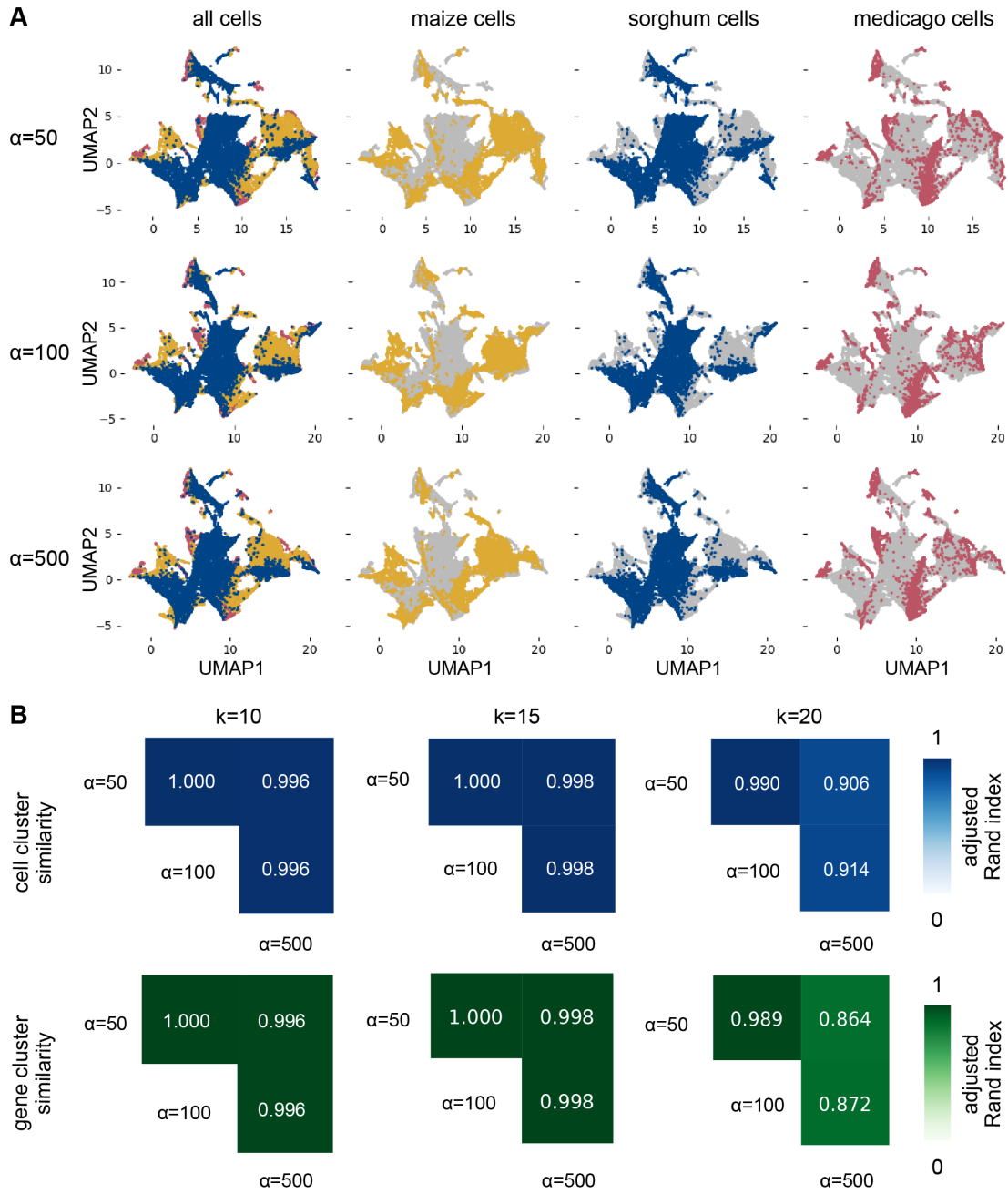


Figure 5.3: Effect of α on embeddings and clusters.. **A.** Visualization of cells using UMAP 2D embeddings of the U factors from each species for $k = 15, \alpha \in \{50, 100, 500\}$. Each row represents different α values and each column represents results from different subsets of species. Each point in each plot is a cell. **B.** Similarity of cell clusters and gene cluster results between pairs of different α values, measured by adjusted Rand index. Each column represents clustering results for different k values.

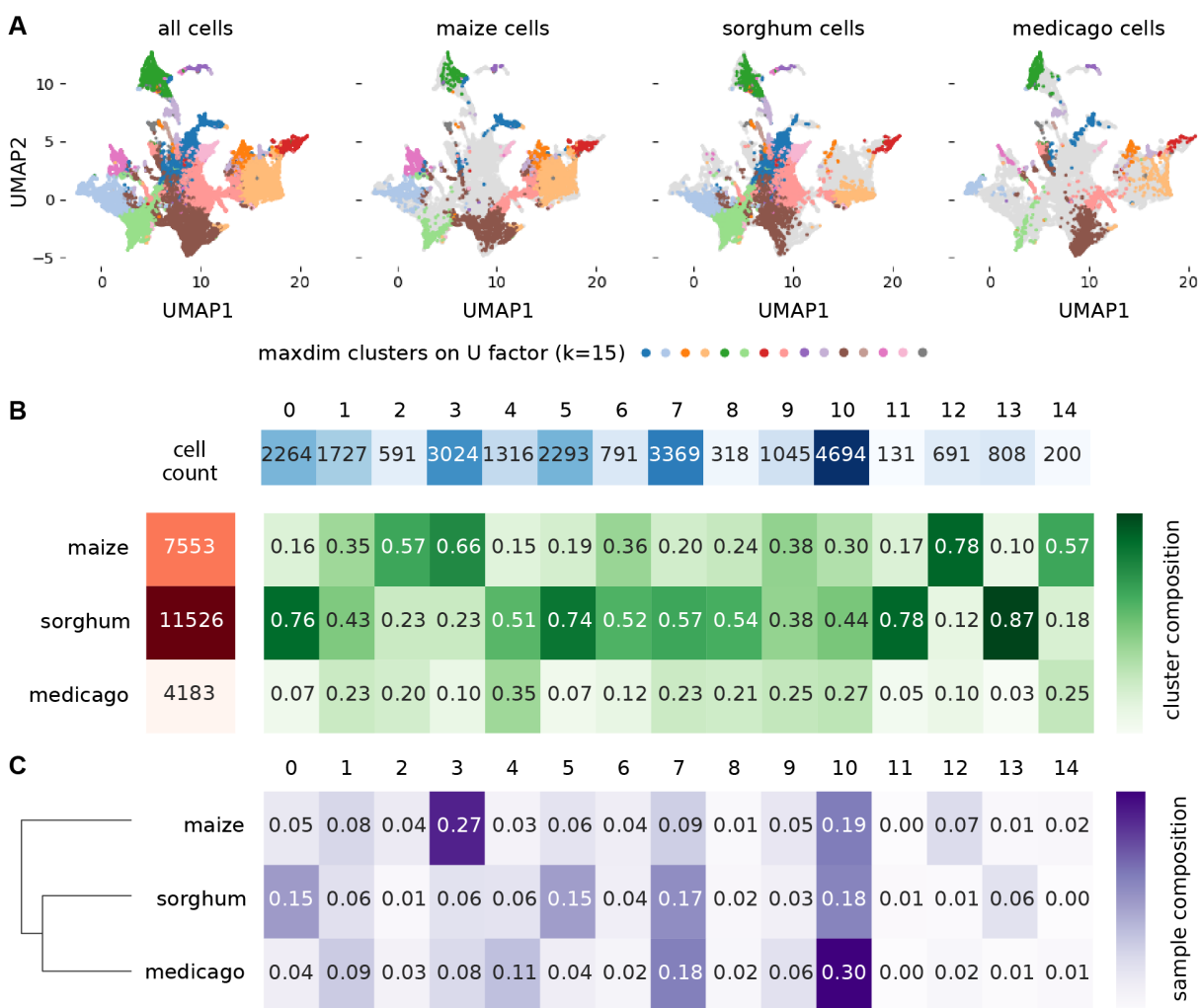


Figure 5.4: TIMBER cell cluster exploration. **A.** UMAP visualization of TIMBER cell clusters from $k = 15, \alpha = 100$. **B.** Composition of each TIMBER cell cluster from $k = 15, \alpha = 100$. The red heatmap to the left is the total number of cells from each species. The blue heatmap on top is the number of cells in each TIMBER cluster. Each column of the cluster composition heatmap (green) sums up to 1. **C.** Composition of each species based on clusters from $k = 15, \alpha = 100$. Each row of the sample/species composition heatmap (purple) sums up to 1. The dendrogram to the left is from hierarchical clustering of the species based on their cluster makeup similarity.

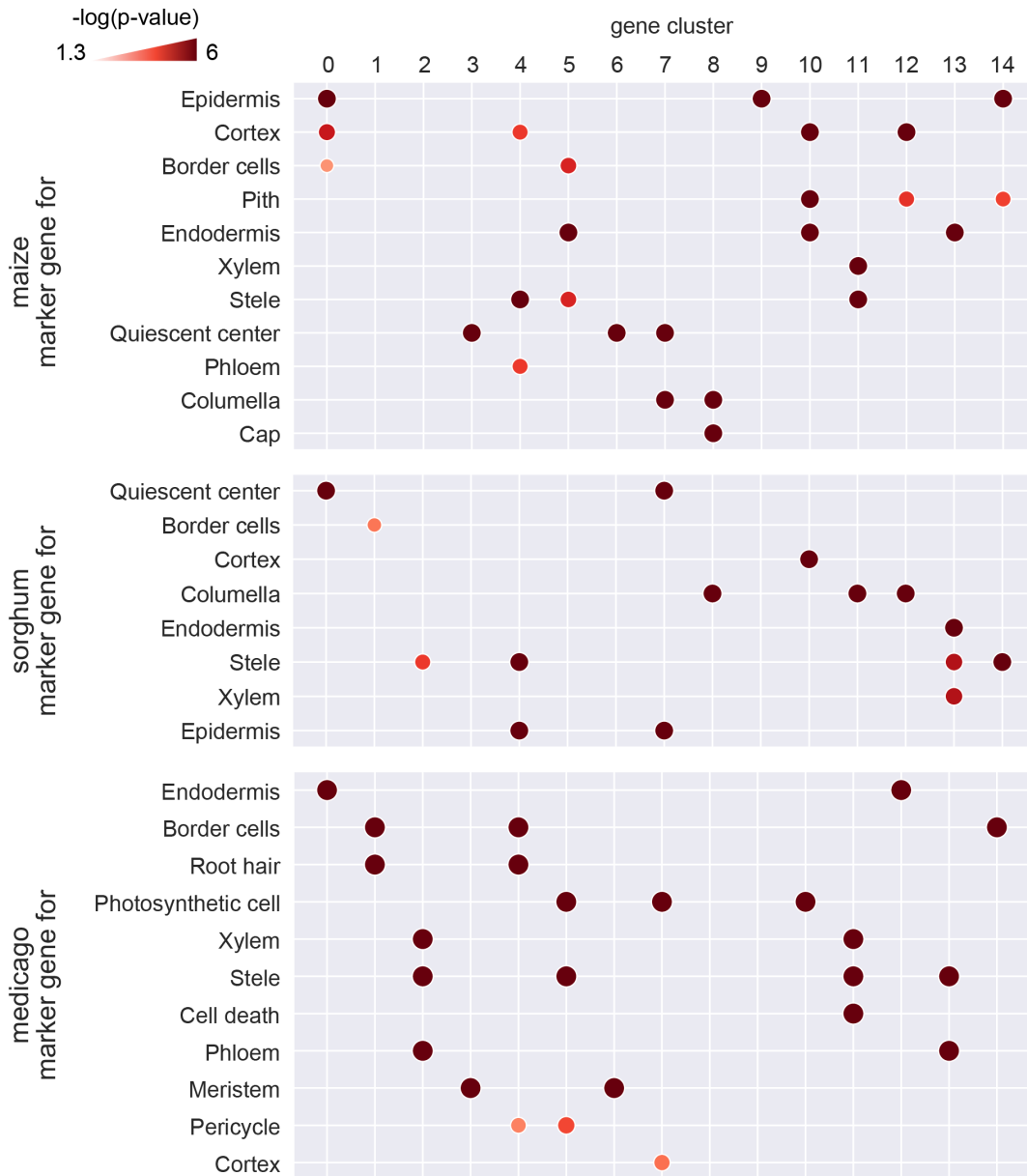


Figure 5.5: Marker gene enrichment within each TIMBER gene cluster ($k = 15$, $\alpha = 100$). Within each species and each gene cluster, we perform hypergeometric test to see if there is significant fold enrichment of marker genes for different plant tissues or cell types ($FDR < 0.05$). Each column is a gene cluster, and each row the plant tissue or cell type whose marker genes have been tested for enrichment in the given cluster. Each bubble represents the negative log (adjusted) p-value of the fold enrichment. Cluster and cell type combinations with adjusted p-value < 0.05 (equivalently $-\log(p\text{-value}) > 1.3$) have been plotted with a bubble.

5.4 Future direction

We have so far applied TIMBER to scRNAseq data from maize, sorghum, and medicago to understand the evolutionarily divergent or conserved programs governing nitrogen fixation in plants. Preliminary results suggest TIMBER is able to integrate cells across species, and recover cell clusters with specific marker gene enrichment patterns for different types of plant tissue and cell types. Further analysis of marker gene expression and differential gene expression is needed to identify the key differences and similarities in transcription associated with mucilage formation, symbiotic relationship with microbiome, and nitrogen fixation.

Benchmarking against existing methods for cell clustering and cross-species integration is needed to evaluate TIMBER's ability to address batch effect and to yield high-quality clusters. Integration metrics, internal cluster quality metrics, and biological conservation metrics will be used to compare TIMBER to a comprehensive list of cross-species integration approaches, including NMF-based methods and batch-correction methods (Luecken et al., 2022; Song et al., 2023).

Current implementation of TIMBER uses a simple orthogroup-based mapping to assign all orthologous genes to the same ancestral gene in the ancestral species node. In-depth analysis of this approach is needed to understand why phylogenetically more distant species (i.e. sorghum and medicago) yield more similar cluster membership pattern according to TIMBER and hierarchical clustering. We also plan to investigate how we can encode more complex sets of gene duplication and speciation events into the mapping matrices.

Hyperparameter selection is a key decision step when applying matrix factorization methods to real data with an unknown rank. We will measure internal cluster quality metrics to select the most suitable rank of the lower-dimensional space (k) for the maize,

sorghum, medicago data, and suggest a selection strategy for *de novo* datasets. While currently α , the strength of regularization between child and parent nodes, is set to a single constant value across all branches of the tree, our TIMBER implementation can handle different values depending on whether the user wants to encode the more nuanced relationship between the input datasets. We plan to experiment with a dynamic set of α values for different branches of the tree based on phylogenetic distance, and measure the stability and quality of the resulting cell and gene clusters.

Finally, TIMBER can be applied to any context where the number of features between input datasets is different, as long as a mapping matrix can be defined. We could exploit this and apply TIMBER to multi-model data integration, e.g. integrating scATACseq and scRNAseq data collected from a system of interest. Because the mapping matrix is flexible in how it encodes feature relationship, we could use Hi-C interactions for mapping to link genes and accessibility to shared genomic regions.

Taken together, TIMBER shows a promising path towards cross-species integration of single-cell transcriptomics data.

5.5 Acknowledgments

This work is supported by the National Institutes of Health (NIH) through the grant R01HG012349 and H.I Romnes Faculty Fellowship, both awarded to Sushmita Roy. Rafael Venado performed single-nucleus RNAseq experiments on maize, medicago, and sorghum, and generated gene orthogroups for the species. Marina Kotvanova and Saptarshi Pyne processed and normalized the snRNAseq dataset for all three species. Marina Kotvanova also processed the output for gene orthogroups. Prakriti Garg provided documentation on experimental procedures and processing pipelines, as well as feedback on orthgroup-to-mapping-matrix logic and algorithm. Sushmita Roy directed the algorithm development

and analysis.

Chapter 6

Concluding remarks

In this dissertation, we have presented a suite of NMF-based methods to understand the underlying structure in 3D genome and single-cell transcriptomics data and their dynamics across multiple biological conditions and evolutionary timescales. In this chapter, we summarize the key findings from each of our frameworks.

GRiNCH

Graph-regularized NMF and clustering of Hi-C data for simultaneous smoothing and TAD identification GRiNCH factorizes an input Hi-C count matrix and uses the factor as an embedding of genomic regions to cluster them into contiguous blocks, which correspond to TADs. It constrains the factorization process with a neighborhood graph, such that regions close to each other in linear distance along the genome are encouraged to have similar output factor loadings or embeddings. Multiplying the factors back together yields a smoothed version of the input Hi-C matrix, with missing values imputed.

GRiNCH outperforms existing TAD-calling and smoothing methods. We benchmarked GRiNCH to 7 other existing TAD-calling methods and 3 existing smoothing methods.

GRINCH shows superior performance in yielding TADs with high internal cluster quality metrics, CTCF enrichment in boundaries, and enrichment of various chromatin modification signals, while being robust to depth and sparsity of the input matrix. Downsampled matrices smoothed by GRINCH recover more similar structures and significant interactions when compared to their original, high-depth counterparts.

GRINCH can identify putative boundary elements and can be applied to data generated by different 3D genome capturing platforms. By analyzing the enrichment of motif sites in GRINCH TAD boundaries in mouse neural differentiation data, we are able to hypothesize a set of novel, context-specific boundary-enforcing elements. We also show that GRINCH can be applied to other types of interaction count matrices generated from a variety of 3D genome capturing technologies outside of Hi-C, such as HiChIP and SPRITE, and identify reproducible sets of TADs from them.

Future Direction: Although the current implementation of GRINCH uses a simple neighborhood graph to regularize the output factors, it is a flexible framework that can adapt any type of graph that encodes the relationship among genomic regions. An adjacency graph based on the similarity of the chromatin states is another natural choice to inform and enforce the low-rank structures found in Hi-C matrices. Another key extension is to adopt a multi-task learning approach so that multiple Hi-C matrices can be analyzed and compared together, which has been implemented in the following method, TGIF.

TGIF

Tree-guided integrated factorization across multiple input Hi-C matrices TGIF takes as input multiple Hi-C matrices from related biological conditions, and a tree that maps the relationship between those conditions as a hierarchy. It simultaneously factorizes the input

matrices in a way such that factors for datasets that are more closely related in the tree are constrained to be more similar to each other. Using the output factors, TGIF identifies a set of significantly differential boundaries, subcompartments, and compartments, revealing the regions contributing to the dynamic 3D genome patterns across the input datasets.

TGIF identifies fewer false positive differential boundaries in both simulated and real Hi-C data. TGIF identifies task-specific boundaries with higher precision than other TAD-calling or differential-boundary-calling methods in multiple simulated datasets. When applied to a cardiomyocyte differentiation timecourse dataset, TGIF boundaries from each timepoint is more enriched for CTCF, a known boundary element, than other methods as well. TGIF also performs well in identifying fewer false positive differences between pairs of biological replicates, or pairs of the same underlying count matrices normalized differently or downsampled to different depths.

TGIF identifies differential compartments and subcompartments with significant changes in other one-dimensional signals along the genome. Significantly differential compartmental regions identified by TGIF between a pluripotent cell and a differentiated cell have significant changes in gene expression and chromatin accessibility. Significantly differential subcompartment regions between different cell states during mouse neural differentiation show diverse patterns of log fold change in histone modification marks.

Differential structure identified by TGIF are enriched for differential gene expression; persistent TGIF boundaries are enriched in disease-associated variants. Differentially expressed genes are significantly enriched in differential compartment regions and in the vicinity of differential boundaries, across 3 different mammalian developmental timecourse datasets we applied TGIF to. We also found that persistent TGIF boundaries across all

timepoints of cardiomyocyte differentiation are significantly enriched for SNPs associated with cardiovascular disease.

Application of a simplified TGIF to Hi-C matrices from multiple species allows quantification of boundary structure conservation. We applied a simplified version of TGIF to 2mbp-by-2mbp Hi-C matrices from 5 different species: human, rat, pig, cow, and dog. The human Hi-C matrices were centered at CTCF peaks; in other species, the matrices were centered at the region to which the corresponding human CTCF peak was lifted over. We find that roughly a third of significant boundaries found in the center of the 2mbp windows were conserved across all species.

Future Direction: A generalization of the input tree structure into a graph would enable even more flexible encoding of the relationship among the input matrices. Allowing different number of features (i.e., columns) in each of the input matrices would make it more suitable for cross-species analysis, where the number of genes or genomic regions of interest are variable across different species. This latter extension is implemented in TIMBER.

TIMBER

Tree-guided integrated factorization with branch-specific feature mapping and regularization TIMBER is an extension to TGIF: (1) it allows for different levels of regularization at each level and branch of the tree, and (2) input matrices can have different number of features (columns), as long as they can be mapped to the parental features.

TIMBER identifies shared and species-specific cell types across three plants species. TIMBER was applied to single-cell gene expression datasets from maize, sorghum, and

medicago, each plant with its own distinct mechanism for nitrogen fixation. Preliminary analysis shows that TIMBER is able to identify clusters of cells shared across all species, as well as clusters whose membership is driven by one dominant species.

Future Direction: Extensive benchmarking of TIMBER is needed to quantify its effectiveness in integrating cell types across multiple species. In-depth analysis of gene clusters identified along with the cell clusters is essential for understanding the underlying gene expression programs driving different cell subpopulations. Finally, TIMBER could easily be adopted for multi-omic or multi-model integration across single-cell gene expression, chromatin accessibility, methylation, and/or proteomics measurements.

Glossary

3D genome organization: Often referred to as 3D genome for simplicity, it refers to how the DNA is folded inside the cell's nucleus.

Batch effect: A type of technical noise that causes two or more datasets to have systematic shift in their values due to differences in experimental protocol, location, etc.

Cluster: A group of data points with similar pattern in their measurements, or to group data points according to such similarity.

Co-clustering: Clustering two different entities (e.g., genes and cells) in a dataset simultaneously.

Compartment: Large segments of the genome that are either transcriptionally active and accessible (A compartment) or repressed and coiled up (B compartment) due to various chromatin modifications.

Chromatin accessibility/modification: Different states of the physical strands of DNA intertwined with proteins (i.e., the chromatin) inside the cell. Accessibility measures whether a stretch of the chromatin is open or is inaccessible because another cellular machinery (e.g. a protein) is already bound to it. Chromatin modification, or histone modification,

measures altered chemical states of proteins (histones) found on the chromatin from different processes like phosphorylation, acetylation, and methylation. These different physical and chemical states correspond to distinct functional states of the chromatin (e.g. the transcriptionally active state).

Dimensionality reduction: A statistical or machine learning method that can describe a dataset more compactly. Clustering, for instance, can describe a large number of data points or samples with a much smaller number of groups whose members share similar patterns in the dataset. More commonly, dimensionality reduction refers to finding a more compact or "lower-dimensional" representation of the data points or samples while retaining the underlying variation or structure in the data. Many data visualization methods (e.g. PCA, UMAP) are in fact reducing the dimension of the data to 2 dimensions, which can then be plotted in a 2-dimensional grid.

Embedding: A "translated" representation of data that machine learning models can more easily and efficiently process. It is often a more compact (or "lower-dimensional") representation of the data than in the original input. It could also refer to the process of generating such representation.

Hi-C: High-throughput chromosomal conformation capture technology; measures pairwise contact or interaction level between genomic regions. Hi-C may refer to and encompass other 3D genome technology and platforms that measure pairwise interactions (e.g. HiChIP) in this dissertation unless distinguished explicitly.

Integration: The computational task of removing batch effect from two or more datasets.

Graph: A network in which nodes (or vertices) are connected with an edge. The edge can be a binary value (0 = not connected; 1 = connect) or have a weight representing the strength of connection or similarity between the nodes.

Matrix factorization: An unsupervised machine learning method that takes a matrix as input and finds two factor matrices that can, when multiplied together, reconstruct the original input matrix. The factors are embeddings of the row entities and the column entities of the matrix, and are often used to co-cluster them.

Multi-omic: Capturing more than one type of cellular readouts, e.g., gene expression, chromatin accessibility, methylation, protein abundance, 3D genome interactions.

Multi-task learning: A machine learning approach that optimizes multiple objectives simultaneously, for example, factorizing multiple matrices at the same time.

Non-negative matrix factorization (NMF): Factorizing a matrix whose entries are all equal to or above zero (common in biomedical domain with count values).

Orthogroup: A group of genes descended or evolved from a single gene in the last common ancestral species among all species under consideration.

Orthology: An evolutionary phenomenon in which different parts of an organism's DNA sequence evolved from the common ancestor, and then diverged or separated after different species split from the ancestor.

Phylogenetic tree: An interchangeable term with species tree; a representation of the evolutionary history and relationship among a group of species, connecting them to their common ancestral species.

Regularization: A machine learning strategy for augmenting an optimization process; providing an additional "regulation", rule, or constraint that the optimization process should obey so that the output may have a desirable quality.

scRNAseq: Single-cell RNAseq, which measures RNA transcript levels at single-cell resolution; often used interchangeably with other single-cell gene expression technology and datasets, e.g., single-nucleus RNAseq (snRNAseq).

Smoothing: Reducing noise in a dataset.

SNP: Single nucleotide polymorphism, or a sequence variant at a single base pair position in the DNA.

Topologically Associating Domain (TAD): A stretch of neighboring genomic regions with high levels of interaction or contact inside the cell.

Transcriptomic: Pertaining to the transcriptome (the complete set of RNA transcribed from the DNA inside the cell).

Bibliography

Adelus ML, Ding J, Tran BT, Conklin AC, Golebiewski AK, Stolze LK, Whalen MB, Cusanovich DA, and Romanoski CE. 2024. Single-cell 'omic profiles of human aortic endothelial cells in vitro and human atherosclerotic lesions ex vivo reveal heterogeneity of endothelial subtype and response to activating perturbations. *eLife* **12**: RP91729. Publisher: eLife Sciences Publications, Ltd.

Adossa N, Khan S, Rytönen KT, and Elo LL. 2021. Computational strategies for single-cell multi-omics integration. *Computational and Structural Biotechnology Journal* **19**: 2588–2596.

Akdemir KC, Le VT, Chandran S, Li Y, Verhaak RG, Beroukhi R, Campbell PJ, Chin L, Dixon JR, and Futreal PA. 2020. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nature Genetics* **52**: 294–305.

Akiduki S and Ikemoto MJ. 2008. Modulation of the neural glutamate transporter EAAC1 by the adducin-interacting protein ARL6IP1. *The Journal of Biological Chemistry* **283**: 31323–31332.

Aldridge S and Teichmann SA. 2020. Single cell transcriptomics comes of age. *Nature Communications* **11**: 4307. Publisher: Nature Publishing Group.

Alharbi RA, Pettengell R, Pandha HS, and Morgan R. 2013. The role of HOX genes in normal hematopoiesis and acute leukemia. *Leukemia* **27**: 1000–1008.

Andrey G, Schöpflin R, Jerković I, Heinrich V, Ibrahim DM, Paliou C, Hochradel M, Timmermann B, Haas S, Vingron M, et al.. 2017. Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. *Genome Research* **27**: 223–233.

Ardakany AR, Ay F, and Lonardi S. 2019. Selfish: discovery of differential chromatin interactions via a self-similarity measure. *Bioinformatics* **35**: i145–i153.

- Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et al.. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**: 246–251. Publisher: Nature Publishing Group.
- Ay F, Bailey TL, and Noble WS. 2014. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research* **24**: 999–1011.
- Bair E. 2013. Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **5**: 349–361.
- Baur B, Lee DI, Haag J, Chasman D, Gould M, and Roy S. 2022. Deciphering the Role of 3D Genome Organization in Breast Cancer Susceptibility. *Frontiers in Genetics* **12**.
- Beagan J, Duong M, Titus K, Zhou L, Cao Z, Ma J, Lachanski C, Gillis D, and Phillips-Cremens J. 2017. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Research* **27**: 1139–1152.
- Belford M, Mac Namee B, and Greene D. 2018. Stability of topic modeling via matrix factorization. *Expert Systems with Applications* **91**: 159–169.
- Bell AC, West AG, and Felsenfeld G. 1999. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**: 387–396.
- Benjamini Y and Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289–300.
- Boltsis I, Grosveld F, Giraud G, and Kolovos P. 2021. Chromatin Conformation in Development and Disease. *Frontiers in Cell and Developmental Biology* **9**.
- Bondell HD and Reich BJ. 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64**: 115–123.
- Bonev B and Cavalli G. 2016. Organization and function of the 3D genome. *Nature Reviews Genetics* **17**: 661–678.
- Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, et al.. 2017. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**: 557–572.e24.
- Boutsidis C and Gallopoulos E. 2008. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition* **41**: 1350–1362. Place: New York, NY, USA Publisher: Elsevier Science Inc.
- Bouwman BA, Crosetto N, and Bienko M. 2023. A GC-centered view of 3D genome organization. *Current Opinion in Genetics & Development* **78**: 102020.

- Bouwman BAM and de Laat W. 2015. Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biology* **16**: 154–9. Publisher: BioMed Central Ltd.
- Braccioli L, Vervoort SJ, Adolfs Y, Heijnen CJ, Basak O, Pasterkamp RJ, Nijboer CH, and Coffey PJ. 2017. FOXP1 Promotes Embryonic Neural Stem Cell Differentiation by Repressing Jagged1 Expression. *Stem Cell Reports* **9**: 1530–1545.
- Brunet JP, Tamayo P, Golub TR, and Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **101**: 4164–4169. Publisher: Proceedings of the National Academy of Sciences.
- Cacciagli P, Desvignes JP, Girard N, Delepine M, Zelenika D, Lathrop M, Lévy N, Ledbetter DH, Dobyns WB, and Villard L. 2014. AP1S2 is mutated in X-linked Dandy–Walker malformation with intellectual disability, basal ganglia disease and seizures (Pettigrew syndrome). *European Journal of Human Genetics* **22**: 363–368.
- Cai D, He X, Han J, and Huang TS. 2011. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**: 1548–1560. Place: Washington, DC, USA Publisher: IEEE Computer Society.
- Cai T, Cai TT, and Zhang A. 2016. Structured Matrix Completion with Applications to Genomic Data Integration. *Journal of the American Statistical Association* **111**: 621–633. Publisher: ASA Website _eprint: <https://doi.org/10.1080/01621459.2015.1021005>.
- Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, Melino G, and Raschella G. 2017. Zinc-finger proteins in health and disease. *Cell Death Discovery* **3**: 1–12.
- Cavalheiro GR, Pollex T, and Furlong EE. 2021. To loop or not to loop: what is the role of TADs in enhancer function and gene regulation? *Current Opinion in Genetics & Development* **67**: 119–129.
- Chakraborty A and Ay F. 2019. The role of 3D genome organization in disease: From compartments to single nucleotides. *Seminars in Cell & Developmental Biology* **90**: 104–113.
- Chakraborty A, Wang JG, and Ay F. 2022. dcHiC detects differential compartments across multiple Hi-C datasets. *Nature Communications* **13**: 6827.
- Chang LH, Ghosh S, and Noordermeer D. 2019. TADs and their borders: free movement or building a wall? *Journal of Molecular Biology* p. S0022283619307429.
- Chen Z, Snetkova V, Bower G, Jacinto S, Clock B, Dizhechi A, Barozzi I, Mannion BJ, Alcaina-Caro A, Lopez-Rios J, et al.. 2024. Increased enhancer–promoter interactions during developmental enhancer activation in mammals. *Nature Genetics* **56**: 675–685.

Ching YH, Ghosh TK, Cross SJ, Packham EA, Honeyman L, Loughna S, Robinson TE, Dearlove AM, Ribas G, Bonser AJ, et al.. 2005. Mutation in myosin heavy chain 6 causes atrial septal defect. *Nature Genetics* **37**: 423–428.

Chou WC, Levy DE, and Lee CK. 2006. STAT3 positively regulates an early step in B-cell development. *Blood* **108**: 3005–3011.

R von Collenberg C, Schmitt D, Rüllicke T, Sendtner M, Blum R, and Buchner E. 2019. An essential role of the mouse synapse-associated protein Syap1 in circuits for spontaneous motor activity and rotarod balance. *Biology Open* **8**.

Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, and Meyer BJ. 2015. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**: 240–244. Publisher: Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

Cresswell KG and Dozmorov MG. 2020. TADCompare: An R Package for Differential and Temporal Analysis of Topologically Associated Domains. *Frontiers in Genetics* **11**.

Cresswell KG, Stansfield JC, and Dozmorov MG. 2020. SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. *BMC Bioinformatics* **21**: 319.

Cubeñas-Potts C and Corces VG. 2015. Architectural Proteins, Transcription, and the Three-dimensional Organization of the Genome. *FEBS letters* **589**: 2923–2930.

Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amodè MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, et al.. 2022. Ensembl 2022. *Nucleic Acids Research* **50**: D988–D995.

Dali R and Blanchette M. 2017. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Research* **45**: 2994–3005.

Davies ER. 2004. *Machine Vision: Theory, Algorithms, Practicalities*. Elsevier.

Dekker J, Alber F, Aufmkolk S, Beliveau BJ, Bruneau BG, Belmont AS, Bintu L, Boettiger A, Calandrelli R, Disteche CM, et al.. 2023. Spatial and temporal organization of the genome: Current state and future aims of the 4D nucleome project. *Molecular Cell* **83**: 2624–2640.

Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O’Shea CC, Park PJ, Ren B, et al.. 2017. The 4D nucleome project. *Nature* **549**: 219–226.

Devarajan K. 2008. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. *PLoS Computational Biology* **4**: e1000029.

- Dixon J, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu J, and Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–380.
- Djekidel MN, Chen Y, and Zhang MQ. 2018. FIND: differential chromatin INteractions Detection using a spatial Poisson process. *Genome Research* **28**: 412–422.
- Dubois F, Sidiropoulos N, Weischenfeldt J, and Beroukhim R. 2022. Structural variations in cancer and the 3D genome. *Nature Reviews Cancer* **22**: 533–546.
- Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, and Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**: 95–98.
- Duren Z, Chen X, Zamanighomi M, Zeng W, Satpathy AT, Chang HY, Wang Y, and Wong WH. 2018. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences* **115**: 7723–7728. Publisher: Proceedings of the National Academy of Sciences.
- Elmentaite R, Domínguez Conde C, Yang L, and Teichmann SA. 2022. Single-cell atlases: shared and tissue-specific cell types across human organs. *Nature Reviews Genetics* **23**: 395–410. Publisher: Nature Publishing Group.
- Emerson DJ, Zhao PA, Cook AL, Barnett RJ, Klein KN, Saulebekova D, Ge C, Zhou L, Simandi Z, Minsk MK, et al.. 2022. Cohesin-mediated loop anchors confine the locations of human replication origins. *Nature* **606**: 812–819.
- Emms DM and Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**: 238.
- ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. Publisher: Nature Research.
- Eres IE, Luo K, Hsiao CJ, Blake LE, and Gilad Y. 2019. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLOS Genetics* **15**: e1008278. Publisher: Public Library of Science.
- Espinola SM, Götz M, Bellec M, Messina O, Fiche JB, Houbbron C, Dejean M, Reim I, Cardozo Gizzi AM, Lagha M, et al.. 2021. Cis -regulatory chromatin loops arise before TADs and gene activation, and are independent of cell fate during early *Drosophila* development. *Nature Genetics* **53**: 477–486.
- Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan JB, Zhang K, Chun J, et al.. 2016. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature methods* **13**: 241.

- Filippova D, Patro R, Duggal G, and Kingsford C. 2014. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology* **9**: 14+.
- Flavahan WA, Drier Y, Liao BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, Suvà ML, and Bernstein BE. 2016. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**: 110–114.
- Forcato M, Nicoletti C, Pal K, Livi C, Ferrari F, and Bicciato S. 2017. Comparison of computational methods for Hi-C data analysis. *Nature Methods* **14**: 679–685.
- Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, et al.. 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **48**: D87–D92.
- Fortin JP and Hansen KD. 2015. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology* **16**: 180.
- Fotuhi Siahpirani A, Ay F, and Roy S. 2016. A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. *Genome Biology* **17**: 114.
- Gacita AM, Fullenkamp DE, Ohiri J, Pottinger T, Puckelwartz MJ, Nobrega MA, and McNally EM. 2021. Genetic Variation in Enhancers Modifies Cardiomyopathy Gene Expression and Progression. *Circulation* **143**: 1302–1316.
- Galan S, Machnik N, Kruse K, Díaz N, Marti-Renom MA, and Vaquerizas JM. 2020. CHESSE enables quantitative comparison of chromatin contact data and automatic feature extraction. *Nature Genetics* **52**: 1247–1255.
- Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korbel JO, and Furlong EEM. 2019. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics* **51**: 1272–1282.
- Álvarez González L, Arias-Sardá C, Montes-Espuña L, Marín-Gual L, Vara C, Lister NC, Cuartero Y, Garcia F, Deakin J, Renfree MB, et al.. 2022. Principles of 3D chromosome folding and evolutionary genome reshuffling in mammals. *Cell Reports* **41**: 111839.
- Graham PH and Vance CP. 2000. Nitrogen fixation in perspective: an overview of research and extension needs. *Field Crops Research* **65**: 93–106.
- Greenwald WW, Li H, Benaglio P, Jakubosky D, Matsui H, Schmitt A, Selvaraj S, D'Antonio M, D'Antonio-Chronowska A, Smith EN, et al.. 2019. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature Communications* **10**: 1054.

- Gómez-Díaz E and Corces VG. 2014. Architectural proteins: regulators of 3D genome organization in cell fate. *Trends in Cell Biology* **24**: 703–711.
- Haarhuis JHI, van der Weide RH, Blomen VA, Yáñez-Cuna JO, Amendola M, van Ruiten MS, Krijger PHL, Teunissen H, Medema RH, van Steensel B, et al.. 2017. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**: 693–707.e14.
- Hafner A and Boettiger A. 2023. The spatial organization of transcriptional control. *Nature Reviews Genetics* **24**: 53–68.
- Hajiveisheh A, Seyedi SA, and Akhlaghian Tab F. 2024. Deep asymmetric nonnegative matrix factorization for graph clustering. *Pattern Recognition* **148**: 110179.
- Hajra L, Evans AI, Chen M, Hyduk SJ, Collins T, and Cybulsky MI. 2000. The NF- κ B signal transduction pathway in aortic endothelial cells is primed for activation in regions predisposed to atherosclerotic lesion formation. *Proceedings of the National Academy of Sciences* **97**: 9052–9057. Publisher: Proceedings of the National Academy of Sciences.
- Hamamoto R, Takasawa K, Machino H, Kobayashi K, Takahashi S, Bolatkan A, Shinkai N, Sakai A, Aoyama R, Yamada M, et al.. 2022. Application of non-negative matrix factorization in oncology: one approach for establishing precision medicine. *Briefings in Bioinformatics* **23**: bbac246.
- Hara S, Morikawa T, Wasai S, Kasahara Y, Koshiba T, Yamazaki K, Fujiwara T, Tokunaga T, and Minamisawa K. 2019. Identification of Nitrogen-Fixing Bradyrhizobium Associated With Roots of Field-Grown Sorghum by Metagenome and Proteome Analyses. *Frontiers in Microbiology* **10**: 407.
- Harmston N, Ing-Simmons E, Tan G, Perry M, Merckenschlager M, and Lenhard B. 2017. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nature Communications* **8**: 441. Publisher: Nature Publishing Group.
- Hata K, Maeno-Hikichi Y, Yumoto N, Burden SJ, and Landmesser LT. 2018. Distinct Roles of Different Presynaptic and Postsynaptic NCAM Isoforms in Early Motoneuron-Myotube Interactions Required for Functional Synapse Formation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **38**: 498–510.
- Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, Rialdi A, White KM, Albrecht RA, Pache L, et al.. 2018. Transcription Elongation Can Affect Genome 3D Structure. *Cell* **174**: 1522–1536.e22.

Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, Lücken MD, Strobl DC, Henao J, Curion F, et al.. 2023. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* **24**: 550–572. Publisher: Nature Publishing Group.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al.. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research* **34**: D590–598.

Hnisz D, Weintraub A, Day D, Valton A, Bak R, Li C, Goldmann J, Lajoie B, Fan Z, Sigova A, et al.. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**: 1454–1458.

Hong S and Kim D. 2017. Computational characterization of chromatin domain boundary-associated genomic elements. *Nucleic Acids Research* **45**: 10403–10414.

Hu X, Shi CH, and Yip KY. 2016. A novel method for discovering local spatial clusters of genomic regions with functional relationships from DNA contact maps. *Bioinformatics* **32**: i111–i120.

Hu Y, Wan S, Luo Y, Li Y, Wu T, Deng W, Jiang C, Jiang S, Zhang Y, Liu N, et al.. 2024. Benchmarking algorithms for single-cell multi-omics prediction and integration. *Nature Methods* **21**: 2182–2194. Publisher: Nature Publishing Group.

Hug CB and Vaquerizas JM. 2018. The Birth of the 3D Genome during Early Embryonic Development. *Trends in Genetics* **0**.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, and Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods* **9**: 999–1003.

Ing-Simmons E, Vaid R, Bing XY, Levine M, Mannervik M, and Vaquerizas JM. 2021. Independence of chromatin conformation and gene regulation during *Drosophila* dorsoventral patterning. *Nature Genetics* **53**: 487–499.

Iyer NR, Shin J, Cuskey S, Tian Y, Nicol NR, Doersch TE, Seipel F, McCalla SG, Roy S, and Ashton RS. 2022. Modular derivation of diverse, regionally discrete human posterior CNS neurons enables discovery of transcriptomic patterns. *Science Advances* **8**: eabn7430. Publisher: American Association for the Advancement of Science.

Jannesari V, Keshvari M, and Berahmand K. 2024. A novel nonnegative matrix factorization-based model for attributed graph clustering by incorporating complementary information. *Expert Systems with Applications* **242**: 122799.

- Jovic D, Liang X, Zeng H, Lin L, Xu F, and Luo Y. 2022. Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine* **12**: e694.
- Kalayeh MM, Idrees H, and Shah M. 2014. NMF-KNN: Image Annotation Using Weighted Multi-view Non-negative Matrix Factorization. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 184–191.
- Kempfer R and Pombo A. 2020. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics* **21**: 207–226.
- Keough KC, Whalen S, Inoue F, Przytycki PF, Fair T, Deng C, Steyert M, Ryu H, Lindblad-Toh K, Karlsson E, et al.. 2023. Three-dimensional genome rewiring in loci with human accelerated regions. *Science (New York, N.Y.)* **380**: eabm1696.
- Kheradpour P and Kellis M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research* **42**: 2976–2987.
- Kim J, He Y, and Park H. 2014. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization* **58**: 285–319.
- Kim S, Yu NK, and Kaang BK. 2015. CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental & Molecular Medicine* **47**: e166–e166.
- Kiselev VY, Andrews TS, and Hemberg M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20**: 273–282. Publisher: Nature Publishing Group.
- Kleinjan DA and Lettice LA. 2008. Long-range gene control and genetic disease. *Advances in genetics* **61**: 339–388.
- Knott SRV, Peace JM, Ostrow AZ, Gan Y, Rex AE, Viggiani CJ, Tavaré S, and Aparicio OM. 2012. Forkhead Transcription Factors Establish Origin Timing and Long-Range Clustering in *S. cerevisiae*. *Cell* **148**: 99–111.
- Koch C, Konieczka J, Delorey T, Lyons A, Socha A, Davis K, Knaack SA, Thompson D, O’Shea EK, Regev A, et al.. 2017. Inference and evolutionary analysis of genome-scale regulatory networks in large phylogenies. *Cell systems* **4**: 543.
- Koitsopoulos PG and Rabkin SW. 2021. The association of polymorphism in PHACTR1 rs9349379 and rs12526453 with coronary artery atherosclerosis or coronary artery calcification. A systematic review. *Coronary Artery Disease* **32**: 448–458.

- Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, and Sabeti PC. 2019. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**: e43803.
- Krefting J, Andrade-Navarro MA, and Ibn-Salem J. 2018. Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biology* **16**: 87.
- Kriebel AR and Welch JD. 2022. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nature Communications* **13**: 780. Publisher: Nature Publishing Group.
- Krijger PHL, Di Stefano B, de Wit E, Limone F, van Oevelen C, de Laat W, and Graf T. 2016. Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming. *Cell Stem Cell* **18**: 597–610.
- Krijger PHL and de Laat W. 2016. Regulation of disease-associated gene expression in the 3D genome. *Nature Reviews Molecular Cell Biology* **17**: 771–782.
- Kruse K, Hug CB, Hernández-Rodríguez B, and Vaquerizas JM. 2016. TADtool: visual parameter identification for TAD-calling algorithms. *Bioinformatics* **32**: 3190–3192.
- Kuang D, Ding C, and Park H. 2012. Symmetric Nonnegative Matrix Factorization for Graph Clustering. In *Proceedings of the 2012 SIAM International Conference on Data Mining (SDM)*, Proceedings, pp. 106–117. Society for Industrial and Applied Mathematics.
- Kuveljic J, Djuric T, Stankovic G, Dekleva M, Stankovic A, Alavantic D, and Zivkovic M. 2021. Association of PHACTR1 intronic variants with the first myocardial infarction and their effect on PHACTR1 mRNA expression in PBMCs. *Gene* **775**: 145428.
- Langmead B and Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.
- Lawson HA, Liang Y, and Wang T. 2023. Transposable elements in mammalian chromatin organization. *Nature Reviews Genetics* pp. 1–12.
- Lazar NH, Nevonen KA, O’Connell B, McCann C, O’Neill RJ, Green RE, Meyer TJ, Okhovat M, and Carbone L. 2018. Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Research* **28**: 983–997. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
- Lee DD and Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**: 788–791.

- Lee DD and Seung HS. 2000. Algorithms for Non-negative Matrix Factorization. In *In NIPS*, volume 13, pp. 556–562.
- Lee DI and Roy S. 2021. GRiNCH: simultaneous smoothing and detection of topological units of genome organization from sparse chromatin contact count matrices with matrix factorization. *Genome Biology* **22**: 164.
- Lee DI and Roy S. 2024. Examining dynamics of three-dimensional genome organization with multi-task matrix factorization.
- Li B and Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078–2079.
- Li X, Zeng G, Li A, and Zhang Z. 2021. DeTOKI identifies and characterizes the dynamics of chromatin TAD-like domains in a single cell. *Genome Biology* **22**: 217.
- Li Z, Li D, Tsun A, and Li B. 2015. FOXP3 + regulatory T cells and their functional regulation. *Cellular & Molecular Immunology* **12**: 558–565.
- Liao Y, Zhang X, Chakraborty M, and Emerson JJ. 2021. Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*. *Genome Research* **31**: 397–410.
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al.. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**: 289–293. Publisher: American Association for the Advancement of Science.
- Lindström S, Wang L, Smith EN, Gordon W, van Hylckama Vlieg A, de Andrade M, Brody JA, Pattee JW, Haessler J, Brumpton BM, et al.. 2019. Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood* **134**: 1645–1657.
- Liu J, Gao C, Sodico J, Kozareva V, Macosko EZ, and Welch JD. 2020. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nature Protocols* **15**: 3632–3662. Publisher: Nature Publishing Group.
- Liu J, Wang C, Gao J, and Han J. 2013. Multi-View Clustering via Joint Nonnegative Matrix Factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (eds. J Ghosh, Z Obradovic, J Dy, ZH Zhou, C Kamath, and S Parthasarathy), pp. 252–260. Society for Industrial and Applied Mathematics, Philadelphia, PA.

- Liu T and Wang Z. 2019. HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data. *Bioinformatics* **35**: 4222–4228.
- Love MI, Huber W, and Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 550.
- Lu L, Barbi J, and Pan F. 2017. The regulation of immune tolerance by FOXP3. *Nature Reviews Immunology* **17**: 703–717.
- Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, et al.. 2022. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods* **19**: 41–50. Publisher: Nature Publishing Group.
- Luecken MD and Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15**: e8746. Publisher: John Wiley & Sons, Ltd.
- Lun AT and Smyth GK. 2015. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**: 258.
- Luo X, Liu Y, Dang D, Hu T, Hou Y, Meng X, Zhang F, Li T, Wang C, Li M, et al.. 2021. 3D Genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell* **184**: 723–740.e21.
- Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al.. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**: 1012–1025.
- Lupiáñez DG, Spielmann M, and Mundlos S. 2016. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics* **32**: 225–237.
- Lévy-Leduc C, Delattre M, Mary-Huard T, and Robin S. 2014. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics (Oxford, England)* **30**: i386–392.
- McArthur E and Capra JA. 2021. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *American Journal of Human Genetics* **108**: 269–283.
- McCord R. 2017. Chromosome biology: How to build a cohesive genome in 3D. *Nature* .
- Melo US, Schöpflin R, Acuna-Hidalgo R, Mensah MA, Fischer-Zirnsak B, Holtgrewe M, Klever MK, Türkmen S, Heinrich V, Pluym ID, et al.. 2020. Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. *The American Journal of Human Genetics* **106**: 872–884.

- Merkenschlager M and Nora EP. 2016. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annual Review of Genomics and Human Genetics* **17**: 17–43.
- Misteli T and Finn EH. 2021. Chromatin architecture is a flexible foundation for gene expression. *Nature Genetics* **53**: 426–427.
- Monahan-Earley R, Dvorak AM, and Aird WC. 2013. Evolutionary origins of the blood vascular system and endothelium. *Journal of Thrombosis and Haemostasis* **11**: 46–66.
- Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, and Chang HY. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* **13**: 919–922.
- Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR, et al.. 2017. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature Genetics* **49**: 1602–1612.
- Nguyen LT, Kim J, and Shim B. 2019. Low-Rank Matrix Completion: A Contemporary Survey. *IEEE Access* **7**: 94215–94237. Conference Name: IEEE Access.
- Nichols MH and Corces VG. 2021. Principles of 3D compartmentalization of the human genome. *Cell Reports* **35**: 109330.
- Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, Bassett DS, and Phillips-Cremens JE. 2018. Detecting hierarchical genome folding with network modularity. *Nature Methods* **15**: 119–122.
- Norton HK and Phillips-Cremens JE. 2017. Crossed wires: 3D genome misfolding in human disease. *Journal of Cell Biology* **216**: 3441–3452.
- Orozco G, Schoenfelder S, Walker N, Eyre S, and Fraser P. 2022. 3D genome organization links non-coding disease-associated variants to genes. *Frontiers in Cell and Developmental Biology* **10**.
- Pankievicz VCS, Delaux PM, Infante V, Hirsch HH, Rajasekar S, Zamora P, Jayaraman D, Calderon CI, Bennett A, and Ané JM. 2022. Nitrogen fixation and mucilage production on maize aerial roots is controlled by aerial root development and border cell functions. *Frontiers in Plant Science* **13**. Publisher: Frontiers.
- Petersen KB and Pedersen MS. 2006. *The Matrix Cookbook*. Technical University of Denmark.

- Pollex T, Rabinowitz A, Gambetta MC, Marco-Ferreres R, Viales RR, Jankowski A, Schaub C, and Furlong EEM. 2024. Enhancer–promoter interactions become more instructive in the transition from cell-fate specification to tissue differentiation. *Nature Genetics* **56**: 686–696.
- Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
- Quinodoz S, Ollikainen N, Tabak B, Palla A, Schmidt J, Detmar E, Lai M, Shishkin A, Bhat P, Takei Y, et al.. 2018. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* .
- Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, and Manke T. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications* **9**.
- Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al.. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–1680.
- Reiff SB, Schroeder AJ, Kırılı K, Cosolo A, Bakker C, Mercado L, Lee S, Veit AD, Balashov AK, Vitzthum C, et al.. 2022. The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nature Communications* **13**: 2365.
- Reinprecht Y, Schram L, Marsolais F, Smith TH, Hill B, and Pauls KP. 2020. Effects of Nitrogen Application on Nitrogen Fixation in Common Bean Production. *Frontiers in Plant Science* **11**. Publisher: Frontiers.
- Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, et al.. 2013. ENCODE Data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Research* **41**: D56–D63.
- Rowley J, Nichols M, Lyu X, Ando-Kuri M, Rivera, Hermetz K, Wang P, Ruan Y, and Corces V. 2017. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Molecular Cell* **67**: 837–852.e7.
- Rowley MJ and Corces VG. 2018. Organizational principles of 3D genome architecture. *Nature Reviews Genetics* p. 1.
- Rowley MJ, Poulet A, Nichols MH, Bixler BJ, Sanborn AL, Brouhard EA, Hermetz K, Linsenbaum H, Csankovszki G, Aiden EL, et al.. 2020. Analysis of Hi-C data using SIP effectively identifies loops in organisms from *C. elegans* to mammals. *Genome Research* **30**: 447–458.

- Roy AL, Conroy RS, Taylor VG, Mietz J, Fingerman IM, Pazin MJ, Smith P, Hutter CM, Singer DS, and Wilder EL. 2023. Elucidating the structure and function of the nucleus—The NIH Common Fund 4D Nucleome program. *Molecular Cell* **83**: 335–342.
- Roy S, Wapinski I, Pfiffner J, French C, Socha A, Konieczka J, Habib N, Kellis M, Thompson D, and Regev A. 2013. Arboretum: Reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Research* **23**: 1039.
- Ryu Y, Han GH, Jung E, and Hwang D. 2023. Integration of Single-Cell RNA-Seq Datasets: A Review of Computational Methods. *Molecules and Cells* **46**: 106–119.
- Schoenfelder S and Fraser P. 2019. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics* p. 1.
- Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, and Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**: 259.
- Sexton CE, Tillett RL, and Han MV. 2022. The essential but enigmatic regulatory role of HERVH in pluripotency. *Trends in Genetics* **38**: 12–21. Publisher: Elsevier.
- Shetty A, Sytnyk V, Leshchyns'ka I, Puchkov D, Haucke V, and Schachner M. 2013. The neural cell adhesion molecule promotes maturation of the presynaptic endocytotic machinery by switching synaptic vesicle recycling from adaptor protein 3 (AP-3)- to AP-2-dependent mechanisms. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **33**: 16828–16845.
- Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, and Zhou XJ. 2016. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research* **44**: e70.
- Shin J, Marx H, Richards A, Vaneechoutte D, Jayaraman D, Maeda J, Chakraborty S, Sussman M, Vandepoele K, Ané JM, et al.. 2021. A network-based comparative framework to study conservation and divergence of proteomes in plant phylogenies. *Nucleic Acids Research* **49**: e3.
- Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, et al.. 2016. ENCODE data at the ENCODE portal. *Nucleic acids research* **44**: gkv1160+. Publisher: Oxford University Press.
- Snetkova V, Ypsilanti AR, Akiyama JA, Mannion BJ, Plajzer-Frick I, Novak CS, Harrington AN, Pham QT, Kato M, Zhu Y, et al.. 2021. Ultraconserved enhancer function does not require perfect sequence conservation. *Nature Genetics* **53**: 521–528. Publisher: Nature Publishing Group.

- Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al.. 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51**: D977–D985.
- Song Y, Miao Z, Brazma A, and Papatheodorou I. 2023. Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nature Communications* **14**: 6495. Publisher: Nature Publishing Group.
- Soor S, Challa A, Danda S, Sagar BSD, and Najman L. 2018. Extending K-Means to Preserve Spatial Connectivity. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6959–6962.
- Soumare A, Diedhiou AG, Thuita M, Hafidi M, Ouhdouch Y, Gopalakrishnan S, and Kouisni L. 2020. Exploiting Biological Nitrogen Fixation: A Route Towards a Sustainable Agriculture. *Plants* **9**: 1011. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.
- Stadhouders R, Vidal E, Serra F, Di Stefano B, Le Dily F, Quilez J, Gomez A, Collombet S, Berenguer C, Cuartero Y, et al.. 2018. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics* **50**: 238–249.
- Stansfield JC, Cresswell KG, and Dozmorov MG. 2019. multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics* **35**: 2916–2923.
- van Steensel B and Furlong EEM. 2019. The role of transcription in shaping the spatial organization of the genome. *Nature Reviews Molecular Cell Biology* **20**: 327–337.
- Stein-O’Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y, Ngom A, Ochs MF, et al.. 2018. Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends in Genetics* **34**: 790–805.
- Stuart T and Satija R. 2019. Integrative single-cell analysis. *Nature Reviews Genetics* **20**: 257–272. Publisher: Nature Publishing Group.
- Szabo Q, Bantignies F, and Cavalli G. 2019. Principles of genome folding into topologically associating domains. *Science Advances* **5**: eaaw1668.
- Tarashansky AJ, Musser JM, Khariton M, Li P, Arendt D, Quake SR, and Wang B. 2021. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *eLife* **10**: e66747. Publisher: eLife Sciences Publications, Ltd.
- Tkacz A, Ledermann R, Martyn A, Schornack S, Oldroyd GED, and Poole PS. 2022. Nodulation and nitrogen fixation in *Medicago truncatula* strongly alters the abundance of its root microbiota and subtly affects its structure. *Environmental Microbiology* **24**: 5524–5533.

- Uribe RA and Bronner ME. 2015. Meis3 is required for neural crest invasion of the gut during zebrafish enteric nervous system development. *Molecular Biology of the Cell* **26**: 3728–3740.
- Ursu O, Boley N, Taranova M, Wang YXR, Yardimci GG, Stafford Noble W, and Kundaje A. 2018. GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* **34**: 2701–2707.
- Valton AL and Dekker J. 2016. TAD disruption as oncogenic driver. *Current opinion in genetics & development* **36**: 34–40.
- Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, and Hadjur S. 2015. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports* **10**: 1297–1309.
- Voronin S and Martinsson PG. 2015. RSVDPACK: An implementation of randomized algorithms for computing the singular value, interpolative, and CUR decompositions of matrices on multi-core and GPU architectures.
- Wang C and Mahadevan S. 2009. A General Framework for Manifold Alignment. In *AAAI Fall Symposium: Manifold Learning and Its Applications*.
- Wang G, Meng Q, Xia B, Zhang S, Lv J, Zhao D, Li Y, Wang X, Zhang L, Cooke JP, et al.. 2020. TADsplimer reveals splits and mergers of topologically associating domains for epigenetic regulation of transcription. *Genome Biology* **21**: 84.
- Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, et al.. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**: 405–409.
- Wang R, Lee JH, Xiong F, Kim J, Hasani LA, Yuan X, Shivshankar P, Krakowiak J, Qi C, Wang Y, et al.. 2021. SARS-CoV-2 Restructures the Host Chromatin Architecture.
- Wang W, Chandra A, Goldman N, Yoon S, Ferrari EK, Nguyen SC, Joyce EF, and Vahedi G. 2022. TCF-1 promotes chromatin interactions across topologically associating domains in T cell progenitors. *Nature Immunology* **23**: 1052–1062.
- Wanggou S, Jiang X, Li Q, Zhang L, Liu D, Li G, Feng X, Liu W, Zhu B, Huang W, et al.. 2012. HESRG: a novel biomarker for intracranial germinoma and embryonal carcinoma. *Journal of Neuro-Oncology* **106**: 251–259.
- Warkman AS, Whitman SA, Miller MK, Garriock RJ, Schwach CM, Gregorio CC, and Krieg PA. 2012. Developmental expression and cardiac transcriptional regulation of Myh7b, a third myosin heavy chain in the vertebrate heart. *Cytoskeleton (Hoboken, N.J.)* **69**: 324–335.

- Weinreb C and Raphael BJ. 2015. Identification of hierarchical chromatin domains. *Bioinformatics* pp. btv485+. Publisher: Oxford University Press.
- Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, et al.. 2017. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**: 1573–1588.e28.
- Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, and Macosko EZ. 2019. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**: 1873–1887.e17.
- de Wit E. 2019. TADs as the Caller Calls Them. *Journal of Molecular Biology* p. S0022283619305923.
- Wolf ESA, Vela S, Wilker J, Davis A, Robert M, Infante V, Venado RE, Voiniciuc C, Ané JM, and Vermerris W. 2024. Identification of genetic and environmental factors influencing aerial root traits that support biological nitrogen fixation in sorghum. *G3 Genes|Genomes|Genetics* **14**: jkad285.
- Wolff J, Rabbani L, Gilsbach R, Richard G, Manke T, Backofen R, and Grüning BA. 2020. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Research* **48**: W177–W184.
- Wu Xn, Shi Tt, He Yh, Wang Ff, Sang R, Ding Jc, Zhang Wj, Shu Xy, Shen Hf, Yi J, et al.. 2017. Methylation of transcription factor YY2 regulates its transcriptional activity and cell proliferation. *Cell Discovery* **3**: 1–22.
- Wu Y, Tamayo P, and Zhang K. 2018. Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding. *Cell Systems* .
- Xiao Q, Luo J, Liang C, Cai J, and Ding P. 2018. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* **34**: 239–248.
- Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, and Lander ES. 2007. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 7145–7150.
- Xiong K and Ma J. 2019. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nature Communications* **10**: 5069.
- Yan J, Xu L, Crawford G, Wang Z, and Burgess SM. 2006. The Forkhead Transcription Factor FoxI1 Remains Bound to Condensed Mitotic Chromosomes and Stably Remodels Chromatin Structure. *Molecular and Cellular Biology* **26**: 155–168.

- Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, Yue F, and Li Q. 2017. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Research* p. gr.220640.117.
- Yang Y, Zhang Y, Ren B, Dixon JR, and Ma J. 2019. Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF. *Cell Systems* 8: 494–505.e14.
- Yu W, He B, and Tan K. 2017. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. *Nature Communications* 8.
- Zemke NR, Armand EJ, Wang W, Lee S, Zhou J, Li YE, Liu H, Tian W, Nery JR, Castanon RG, et al.. 2023. Conserved and divergent gene regulatory programs of the mammalian neocortex. *Nature* 624: 390–402. Publisher: Nature Publishing Group.
- Zerbino DR, Wilder SP, Johnson N, Juettemann T, and Flicek PR. 2015. The Ensembl Regulatory Build. *Genome Biology* 16: 56.
- Zhang J and Xie M. 2022. Graph regularized non-negative matrix factorization with prior knowledge consistency constraint for drug–target interactions prediction. *BMC Bioinformatics* 23: 564.
- Zhang R, Zhou T, and Ma J. 2022. Ultrafast and interpretable single-cell 3D genome analysis with Fast-Higashi. *Cell Systems* 13: 798–807.e6.
- Zhang Y, Li T, Preissl S, Amaral ML, Grinstein J, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, et al.. 2019. Transcriptionally Active HERV-H Retrotransposons Demarcate Topologically Associating Domains in Human Pluripotent Stem Cells. *Nature genetics* 51: 1380–1388.
- Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, Nussbaum C, Myers R, Brown M, Li W, et al.. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* 9: R137.
- Zheng H and Xie W. 2019. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology* 20: 535–550.
- Zheng X and Zheng Y. 2018. CscoreTool: fast Hi-C compartment analysis at high resolution. *Bioinformatics* 34: 1568–1570.
- Zhou J, Ma J, Chen Y, Cheng C, Bao B, Peng J, Sejnowski TJ, Dixon JR, and Ecker JR. 2019. Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proceedings of the National Academy of Sciences* 116: 14011–14018.
- Zhu R, Liu JX, Zhang YK, and Guo Y. 2017. A Robust Manifold Graph Regularized Nonnegative Matrix Factorization Algorithm for Cancer Gene Clustering. *Molecules : A Journal of Synthetic Chemistry and Natural Product Chemistry* 22: 2131.

Zufferey M, Tavernari D, Oricchio E, and Ciriello G. 2018. Comparison of computational methods for the identification of topologically associating domains. *Genome Biology* **19**: 217.

Appendix A

GRiNCH supplementary materials

A.1 Supplementary figures

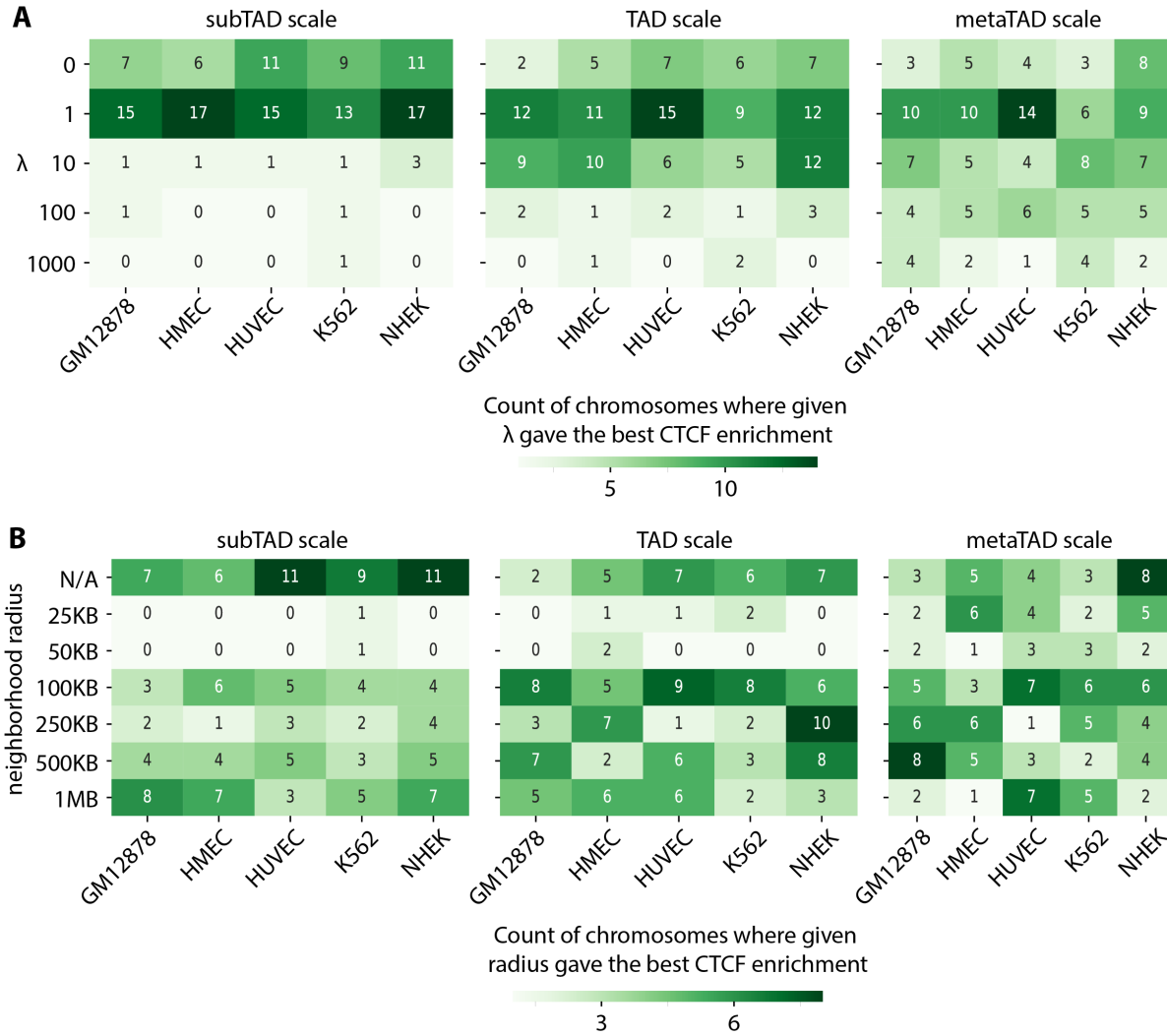


Figure A.1: Selecting graph regularization parameters λ and neighborhood radius, r . λ controls the strength of regularization. r determines how many neighboring genomic regions will be used to influence a given region during the regularization process. **A.** Shown are the count of chromosomes in which the given λ value (row) gave the best CTCF enrichment within the given cell line (column), for different k settings denoted by subTAD, TAD, and metaTAD scale (see **Figure A.14**, **Figure A.15**). Within each cell line, for each chromosome, we ranked the tested parameter combinations of λ and r , and then counted the times a given λ yielded the best CTCF enrichment (regardless of r value). Due to ties in ranking, each column can add up to 23 or more. $\lambda = 0$ corresponds to vanilla NMF without regularization. **B.** Shown are the count of chromosomes in which the given neighborhood radius value (row) gave the best CTCF enrichment in each cell line (column) at different k settings. As in **A**, we rank the parameter combinations of λ and r based on the CTCF enrichment, then count the number of times a particular value of r yielded the best fold enrichment. N/A corresponds to vanilla NMF without regularization.

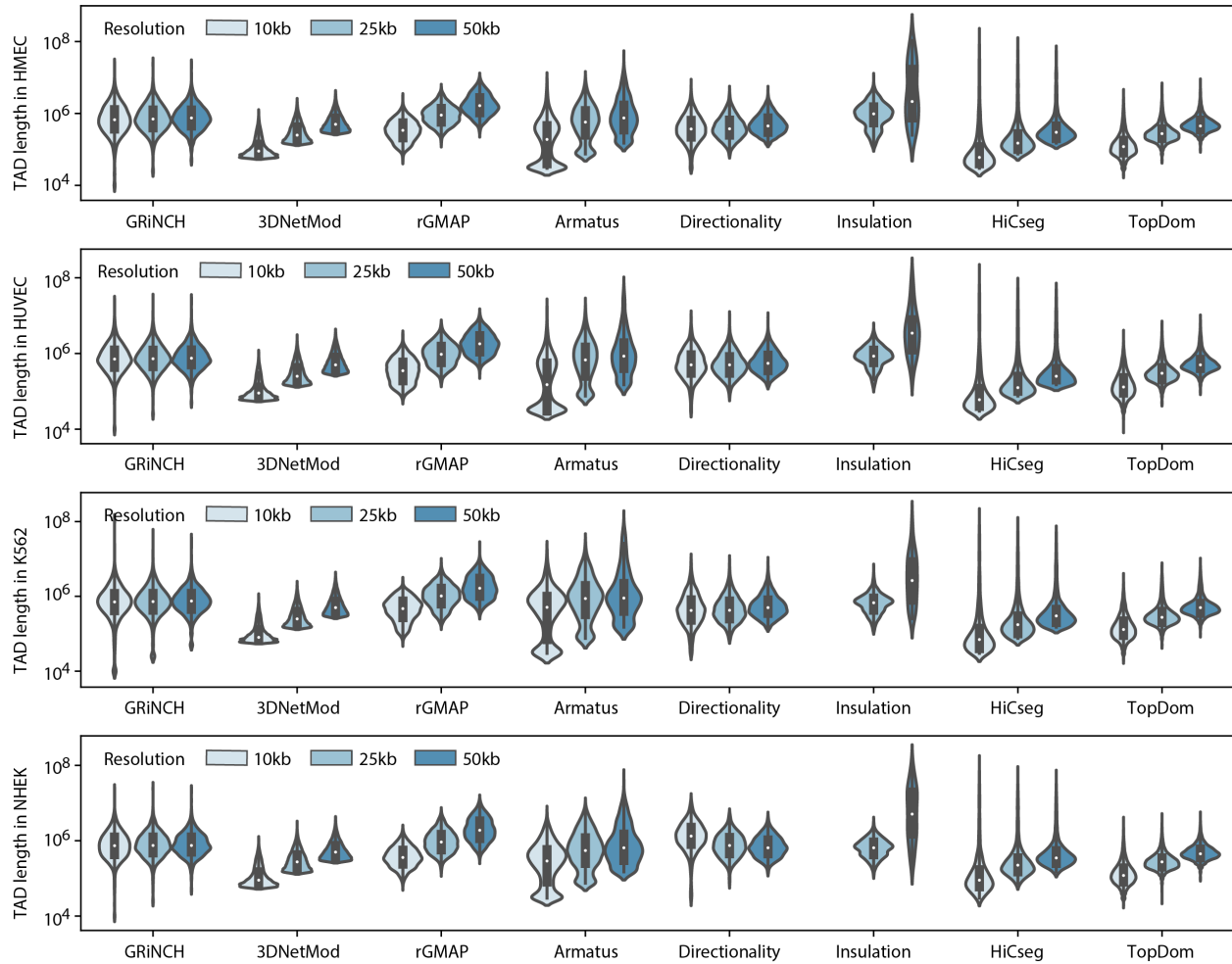


Figure A.2: The size distribution of TADs identified by different methods from different resolutions of Hi-C data. Y-axis is in log10 scale of base pairs. The white dot in inside each violin represents the median; the black box inside each violin stretches from Q1 (25th percentile) to Q3 (75th percentile). Insulation method is missing TAD distributions from 10kb data because it did not return any TADs when using the same hyperparameters as in from 25kb and 50kb data.

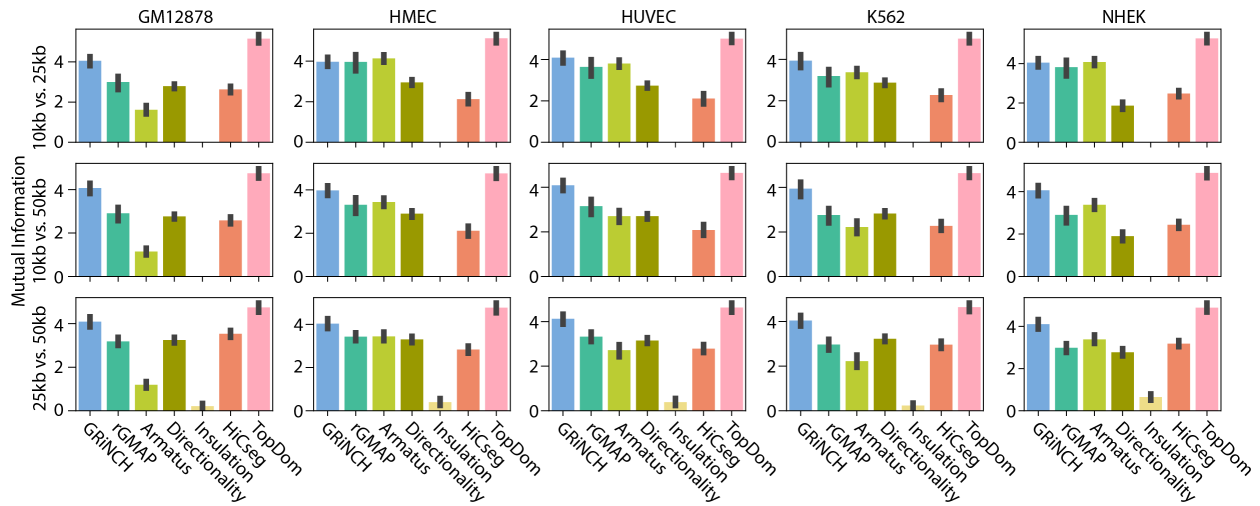
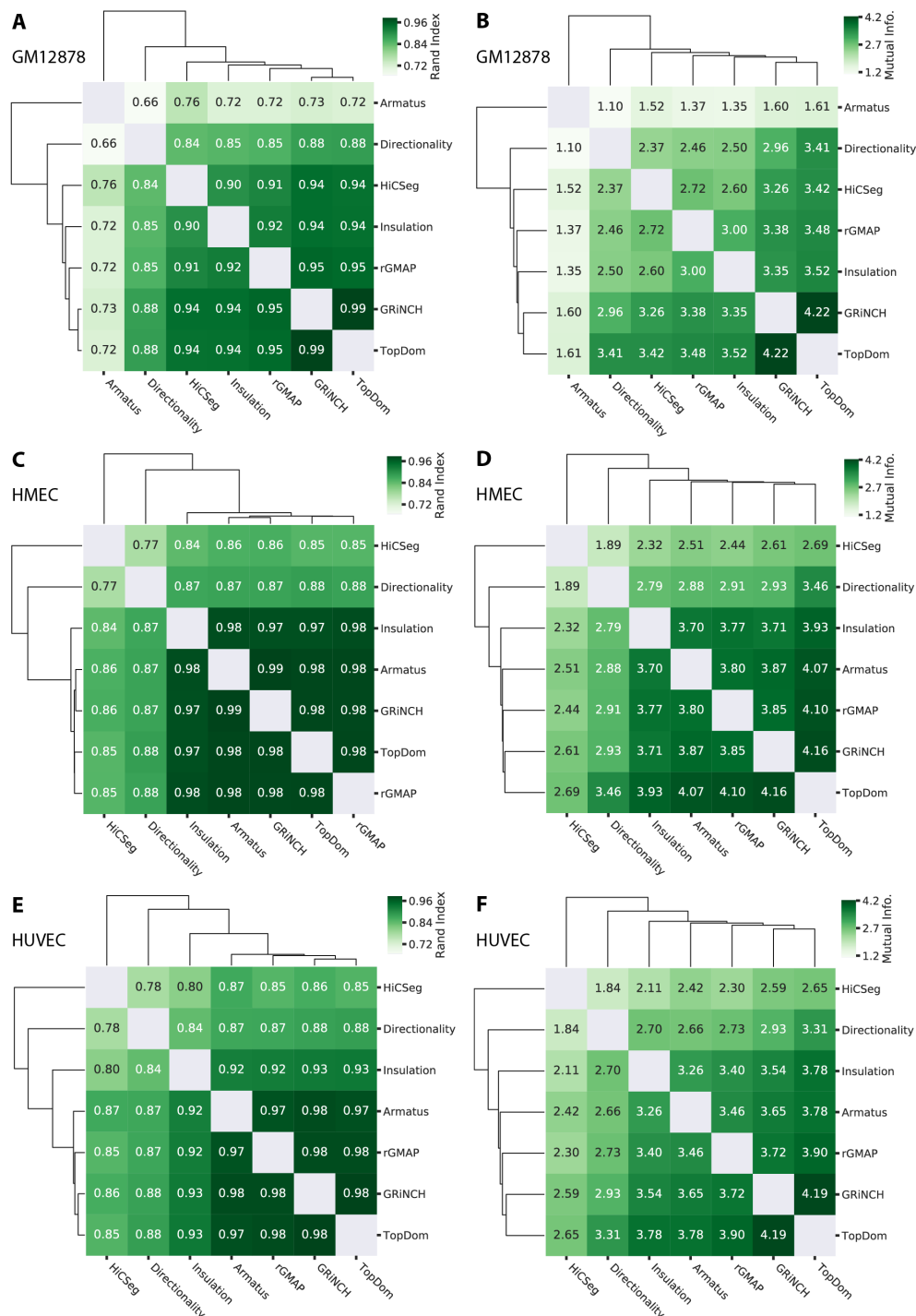


Figure A.3: Similarity of TADs across resolutions for different methods, measured by mutual information. TADs were first converted to clusters, by assigning regions within the same TAD into a single cluster. To compare across resolutions, 10kb, 25kb, and 50kb bins were split into a size of lowest common denominator, i.e., 5kb. Then all 5kb bins were assigned to the same cluster as in the original lower-resolution bin (e.g. a 10kb bin assigned to cluster A would yield two 5kb bins assigned to cluster A). Finally, cluster assignments in these split, higher-resolution bins were compared for their similarity with Rand Index (Figure 3 in main text) and Mutual Information.



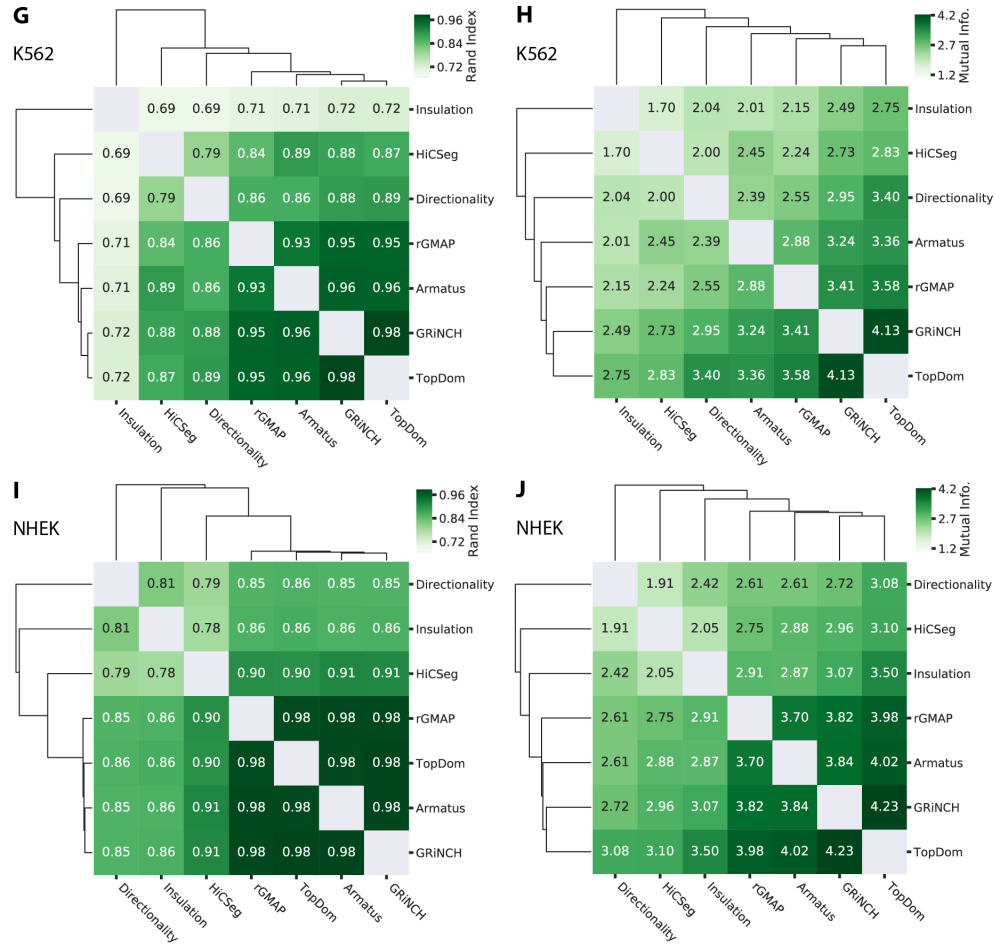


Figure A.4: Evaluating similarity of TADs from different TAD-calling methods using Rand Index (A, C, E, G, I) and Mutual Information (B, D, F, H, J) for five cell lines from Rao et al., Gm12878 (A, B), HMEC (C, D), HUVEC (E, F), K562 (G, H), NHEK (I, J).

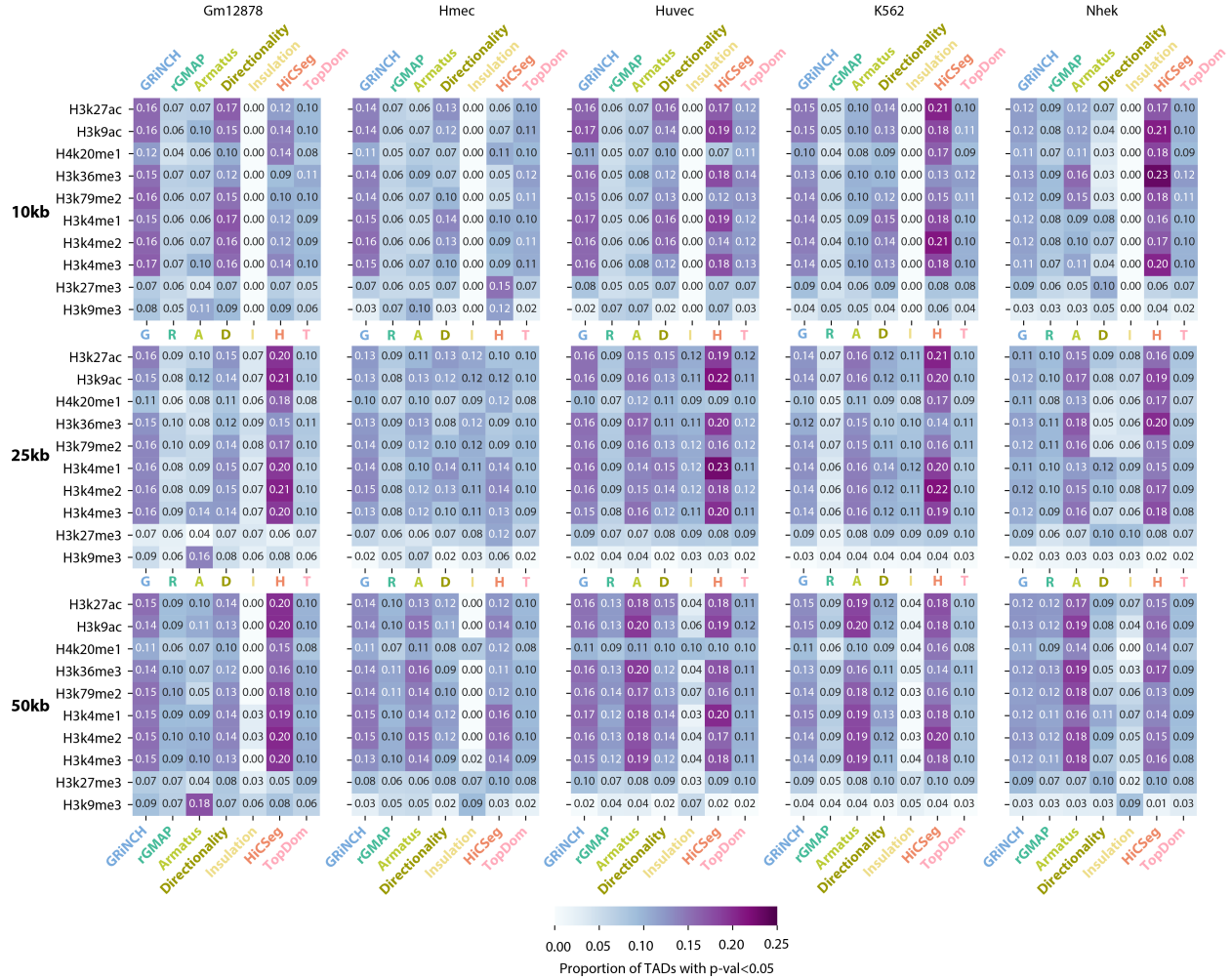


Figure A.5: Proportion of TADs, from different Hi-C resolutions, with significant mean histone modification signal (i.e. empirical p-value < 0.05). The darker the entry the higher the proportion of TADs with significant histone enrichment. The average ChIP-seq signal for each histone modification mark was taken from within each TAD; the p-value of each TAD is derived from an empirical null distribution of mean signals in randomly shuffled TADs. Note: 3DNetMod outputted overlapping TADs and was excluded from this analysis as it involves TAD randomization/shuffling.

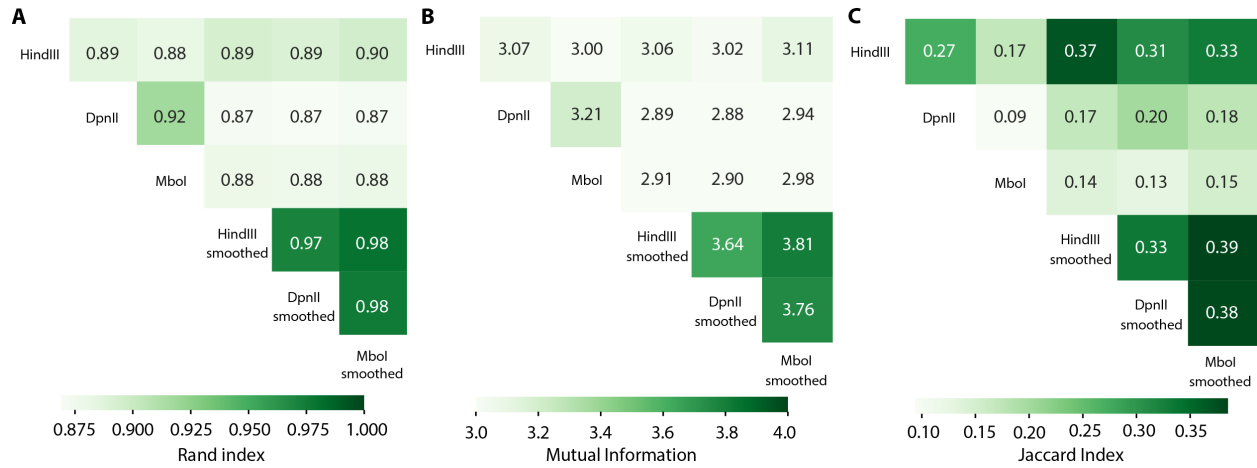


Figure A.6: Similarity of structure and significant interactions among Hi-C data using different restriction enzymes and Hi-C data smoothed by GRiNCH. **A.** Similarity of Directionality TADs, measured by Rand index, from Gm12878 Hi-C data using different restriction enzymes (HindIII, DpnII, MboI) and smoothed by GRiNCH (HindIII smoothed, DpnII smoothed, MboI smoothed). **B.** Similarity of Directionality TADs, measured by mutual information. **C.** Similarity or overlap in significant interactions, measured by Jaccard Index, called by FitHiC from Gm12878 Hi-C data using different restriction enzymes (HindIII, DpnII, MboI) and smoothed by GRiNCH (HindIII smoothed, DpnII smoothed, MboI smoothed).

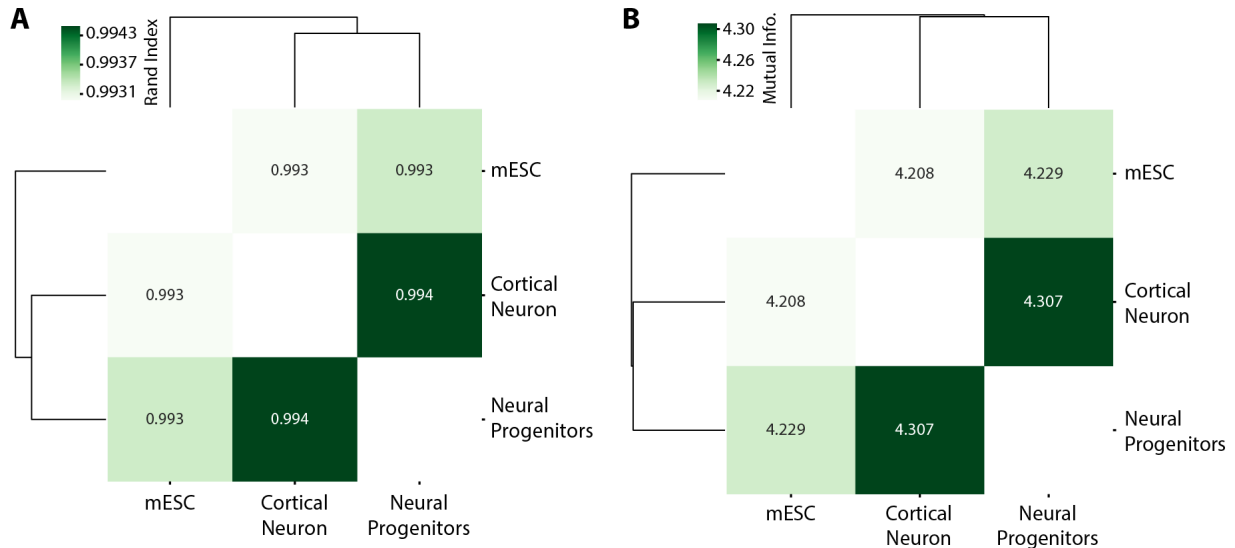


Figure A.7: Similarity of GRiNCH TADs from mouse neural development time-course Hi-C data measured for three stages: mESC, Cortical Neuron, Neural Progenitors. The order of the developmental stages is mESC, Cortical Neurons, Neural Progenitors. **A.** Similarity of TADs by measured by Rand index. **B.** Similarity of TADs measured by Mutual Information.

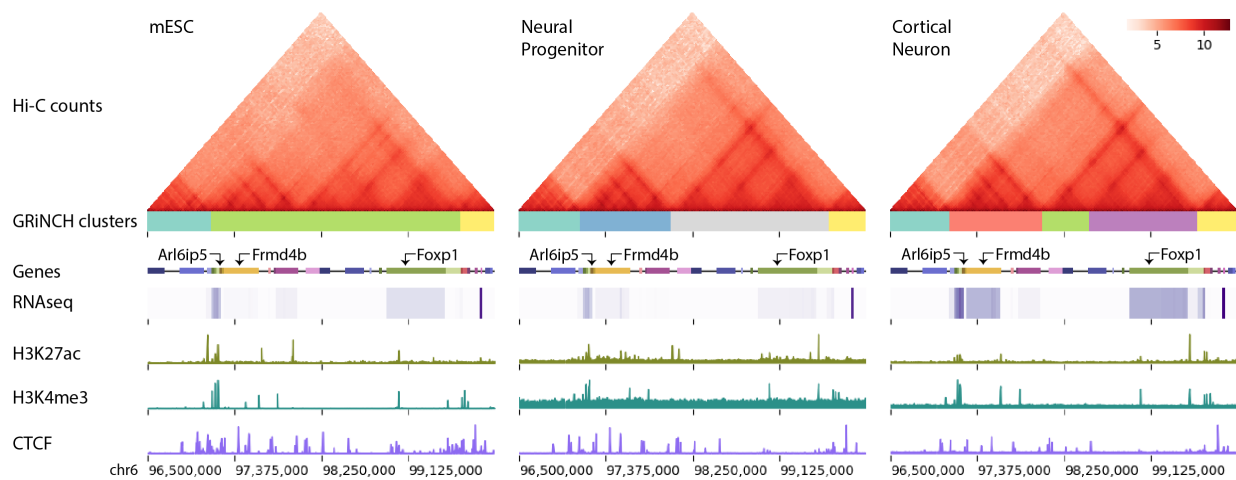


Figure A.8: Interaction profile near *Arl6ip5* and *Foxp1* in mouse embryonic stem cells (mESC), neural progenitors (NPC), and differentiated cortical neurons (CN). Heatmaps are of Hi-C matrices after log2-transformation of interaction counts for better visualization. GRiNCH clusters are visualized as blocks of different colors under the heatmap of interaction counts. Genes in the nearby regions are marked by small boxes, and a heatmap of their corresponding RNA-seq levels (in TPM) is shown underneath each gene. ChIP-seq signals from H3K27ac, H3K4me3, and CTCF are shown as separate tracks.

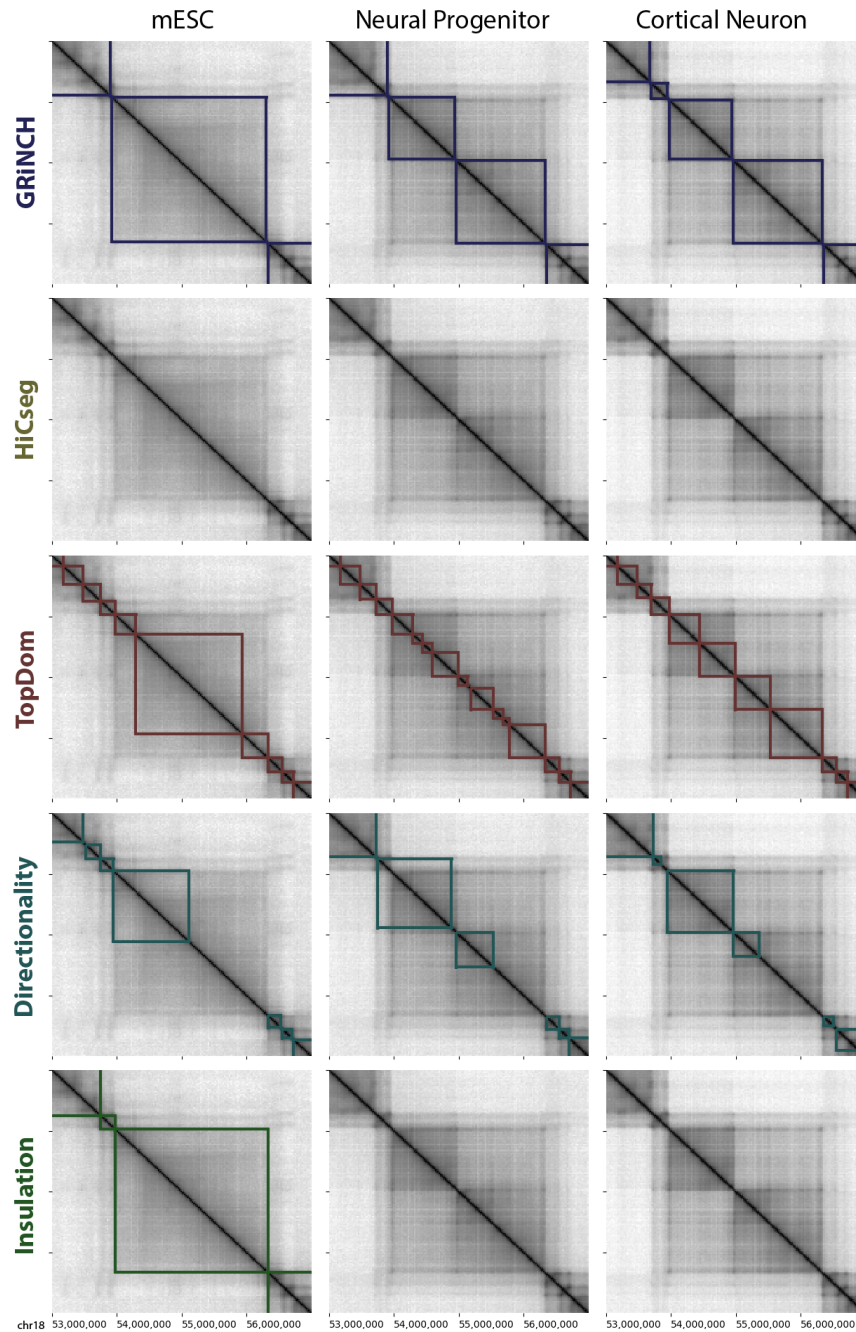


Figure A.9: Visual comparison of TADs identified by different TAD-calling methods in mouse neural development time-course data. The grey heatmap visualizes the interactions surrounding Zfp608 in chr18 (same region visualized in Figure 7A in main text). Interactions counts were log2-transformed for better visualization. The boxes represent TAD boundaries.

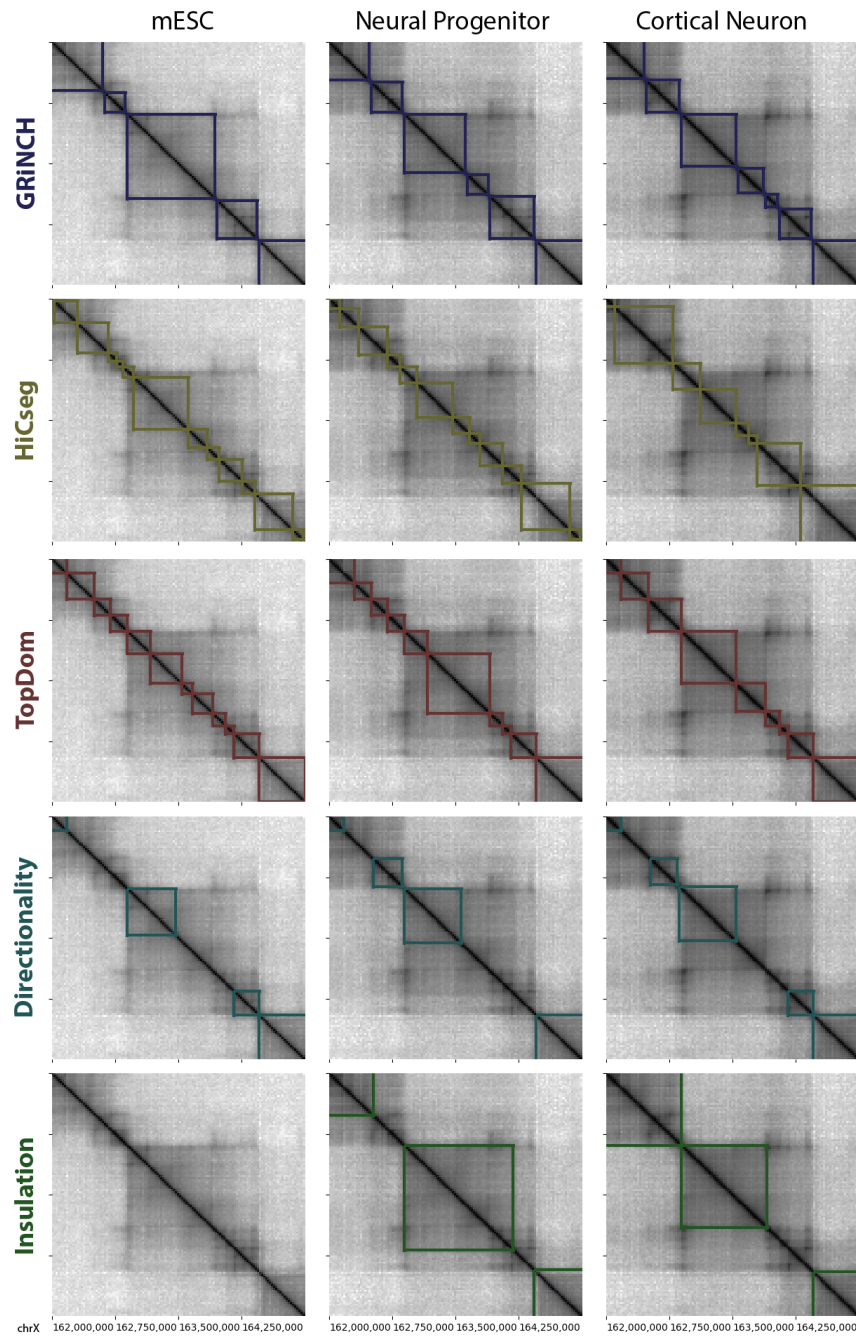


Figure A.10: Visual comparison of TADs identified by different TAD-calling methods in mouse neural development time-course data. The grey heatmap visualizes the interactions surrounding *Syp1* and *Ap1s2* in chrX (same region visualized in Figure 7B in main text). Interactions counts were log2-transformed for better visualization. The boxes represent TAD boundaries.

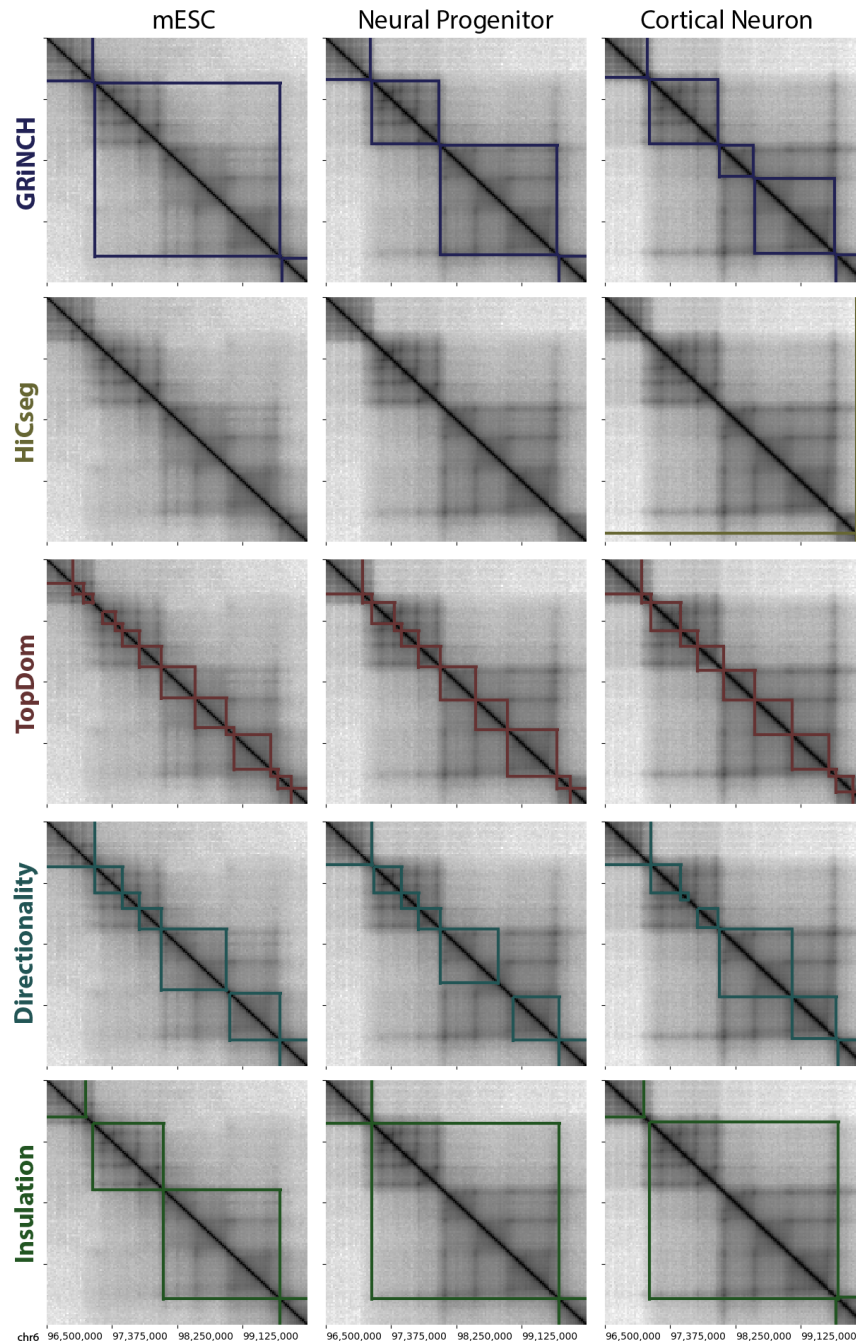


Figure A.11: Visual comparison of TADs identified by different TAD-calling methods in mouse neural development time-course data. The grey heatmap visualizes the interactions surrounding *Arl6ip5* and *Foxp1* in chr6 (same region visualized in Figure S10 above). Interactions counts were log2-transformed for better visualization. The boxes represent TAD boundaries.

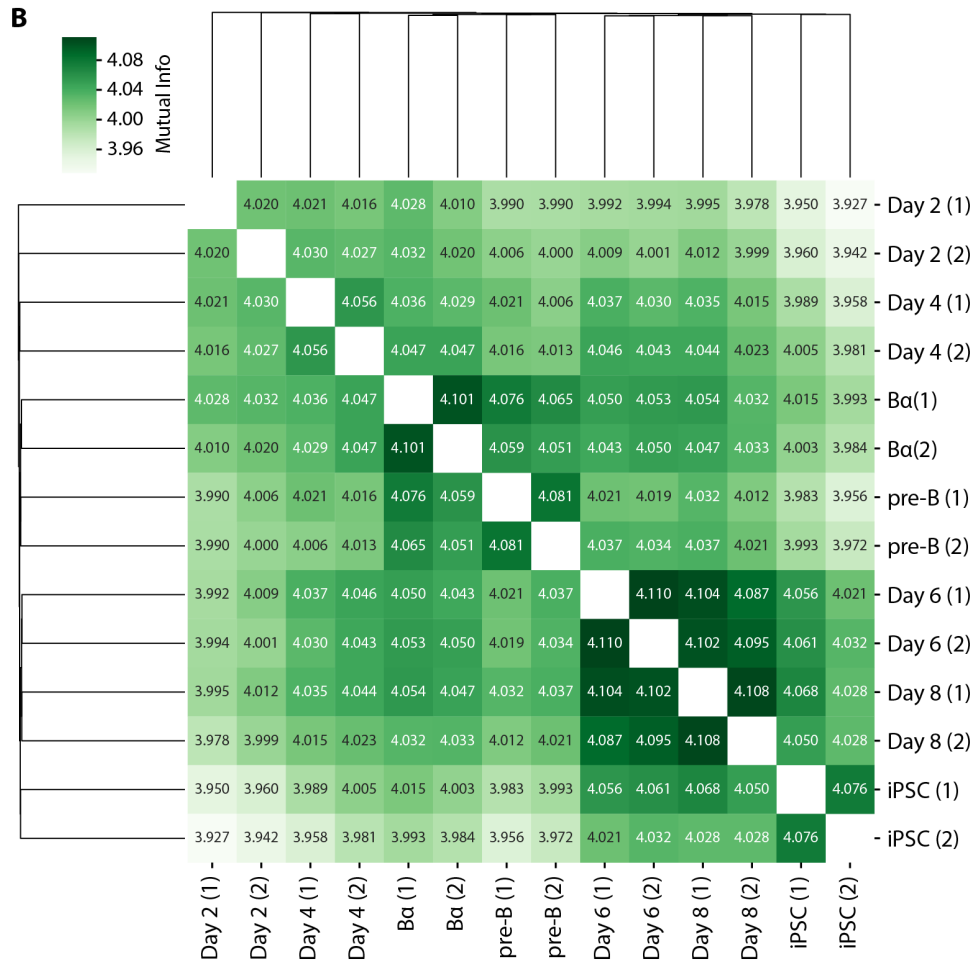


Figure A.12: Similarity of GRiNCH TADs from pluripotency reprogramming time-course Hi-C data measured for the starting pre-B cells and ending induced pluripotent (iPSC) state and five intermediate time points, Bα, Day2, Day4, Day6, Day8. (1) and (2) suffixes represent replicate 1 and 2 respectively. **A.** Similarity of TADs measured by Rand index. **B.** Similarity of TADs measured by Mutual Information.

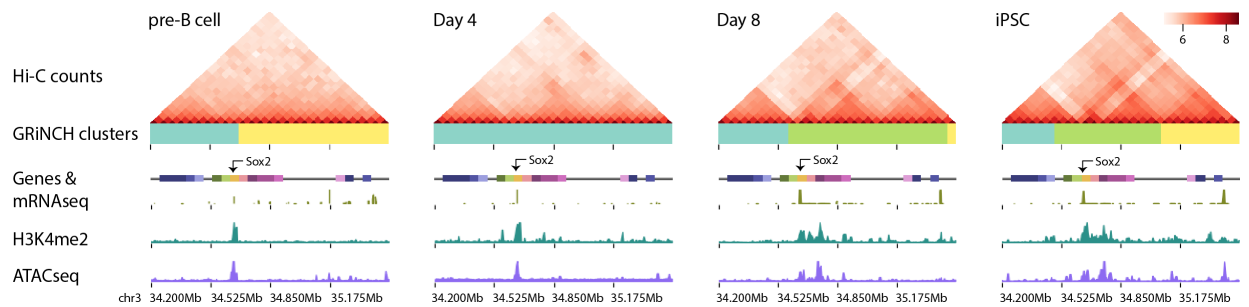


Figure A.13: Interaction profile near the Sox2 gene in mouse pre-B cells, in day 4 and day 8 of reprogramming, and in induced pluripotent stem cell (iPSC). Heatmaps are of Hi-C matrices after log2-transformation of interaction counts for better visualization. GRiNCH clusters are visualized as blocks of different colors under the heatmap of interaction counts. Genes in the nearby regions are marked by small boxes, and peaks of their corresponding RNA-seq levels are shown underneath each gene. ChIP-seq signals from H3K4me2 and ATAC-seq signals are shown as separate tracks.

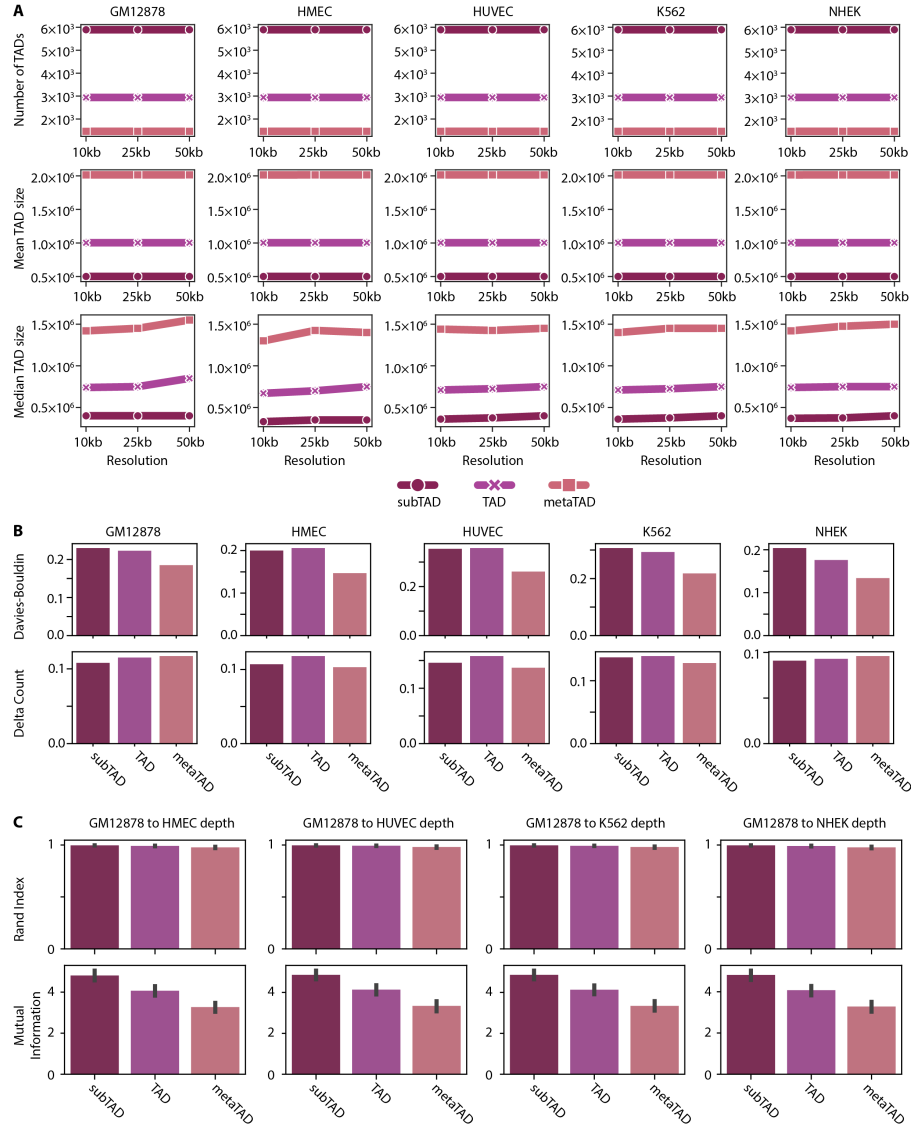


Figure A.14: Characterizing GRiNCH clusters of different size scales, i.e., subTADs, TADs, metaTADs. Shown are different statistics for different settings of the number of clusters in GRiNCH, k . We set k based on the expected size of the clusters and consider three expected sizes: subTADs (500kb), TADs (1MB), metaTADs (2MB). TADs are known to be ~1MB and therefore, for the “TAD scale” regions, $k = \frac{n_c}{1\text{Mb}}$, where n_c is the length of chromosome c . **A.** The number of subTADs, TADs, and metaTADs, and their median size for Hi-C datasets from five cell lines at three different resolutions: 10kb, 25kb, 50kb. **B.** Proportion of subTADs, TADs, or metaTADs with significantly better ($p\text{-val} < 0.05$) cluster quality metrics than random clusters, as measured by Davies-Bouldin Index and Delta Count. Results here are shown for 25kb-resolution data. **C.** The similarity between subTADs, TADs, and metaTADs from high-depth GM12878 dataset and those from datasets downsampled to other lower depths available in different cell lines. Similarity is measured by Rand Index and Mutual Information. Results here are shown for 25kb-resolution data.

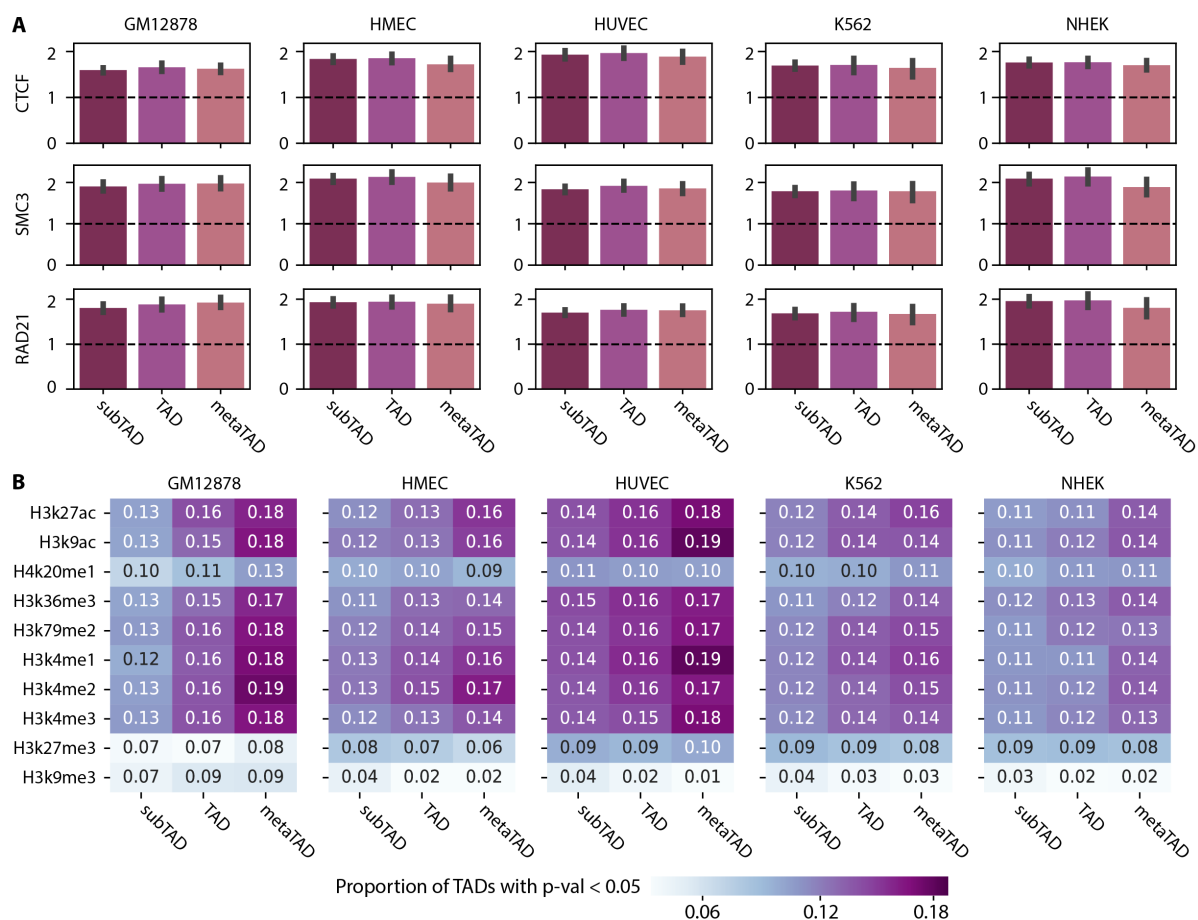
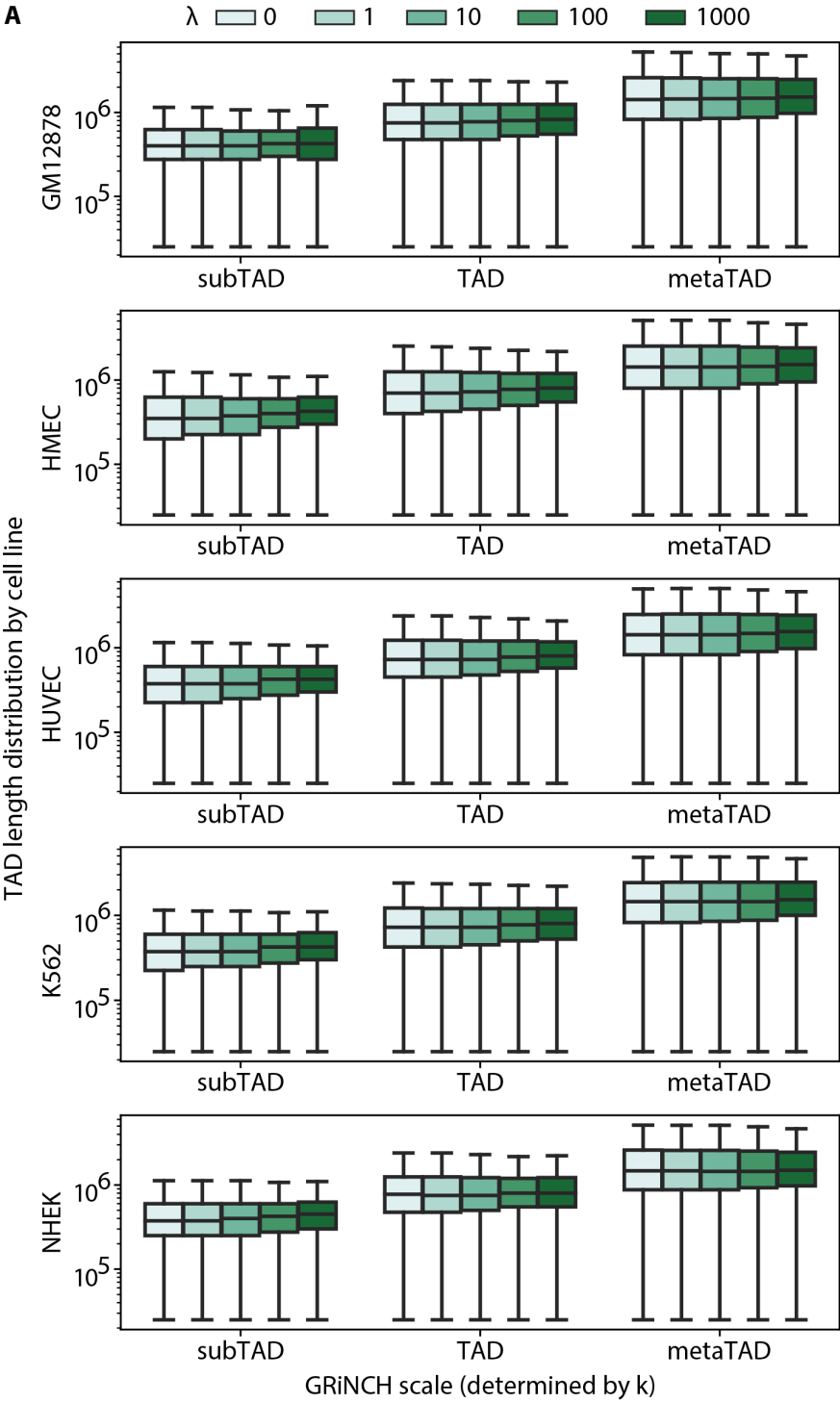


Figure A.15: Enrichment of regulatory signals in GRiNCH clusters of different expected sizes: subTAD (500kb), TAD (1Mb), and metaTAD (2Mb). The expected size is used to set the GRiNCH parameter k , the number of clusters. Results are shown for 25kb resolution data. **A.** Fold enrichment of architectural protein binding signals in subTAD, TAD, and metaTAD boundaries. **B.** Proportion of subTADs, TADs, and metaTADs with significantly higher values ($p\text{-val} < 0.05$) of mean histone modifications compared to random clusters.



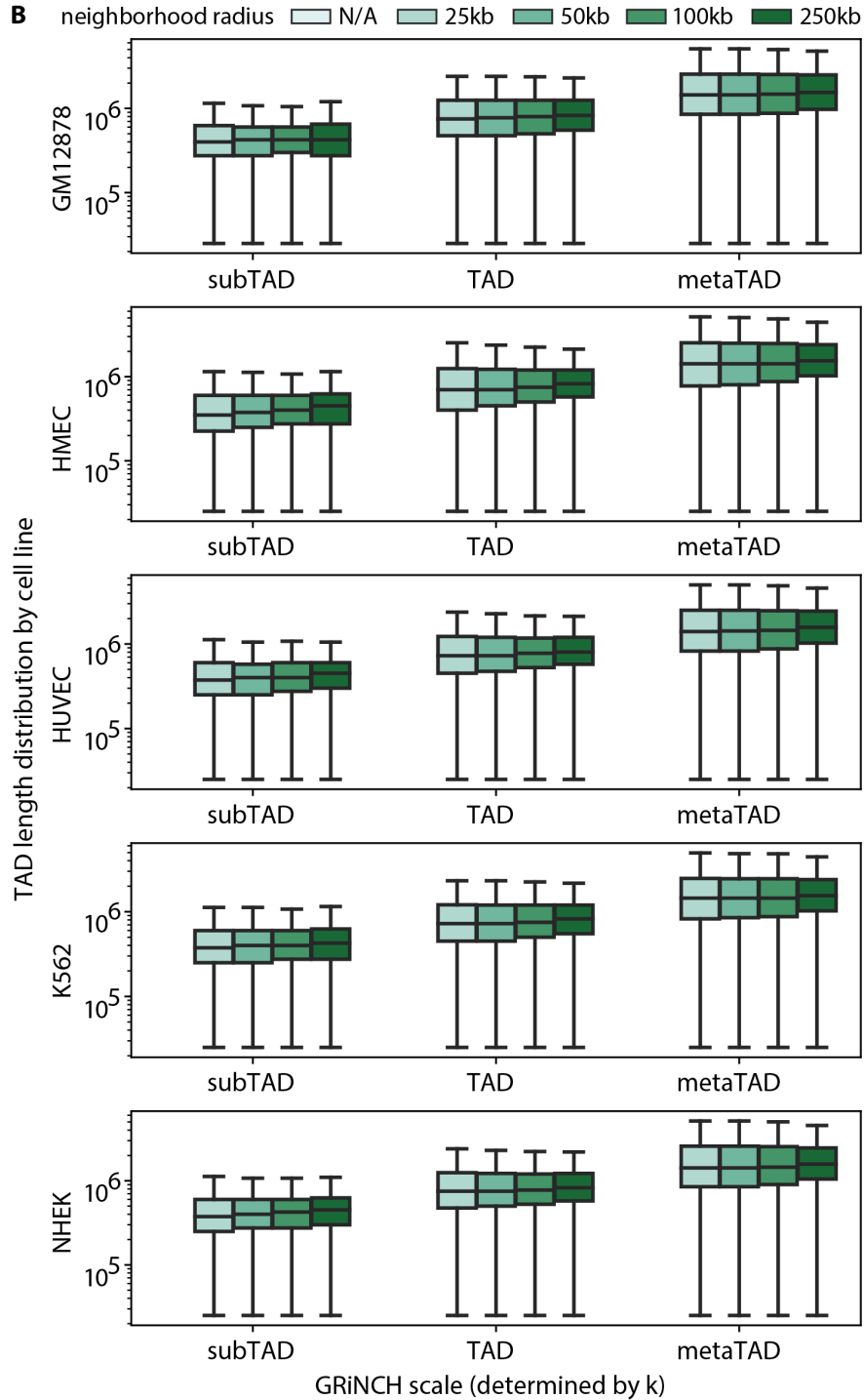


Figure A.16: GRiNCH TAD size distribution by graph regularization parameters **A.** λ and **B.** neighborhood radius. $\lambda = 0$ and neighborhood radius of “N/A” correspond to no regularization. Shown are distributions for different settings of the number of clusters in GRiNCH, k . We set k based on the expected size of the clusters and consider three expected sizes: subTADs (500kb), TADs (1MB), metaTADs (2MB). TADs are known to be ~1MB and therefore, for the “TAD scale” regions, $k = \frac{n_c}{1\text{Mb}}$, where n_c is the length of chromosome c .

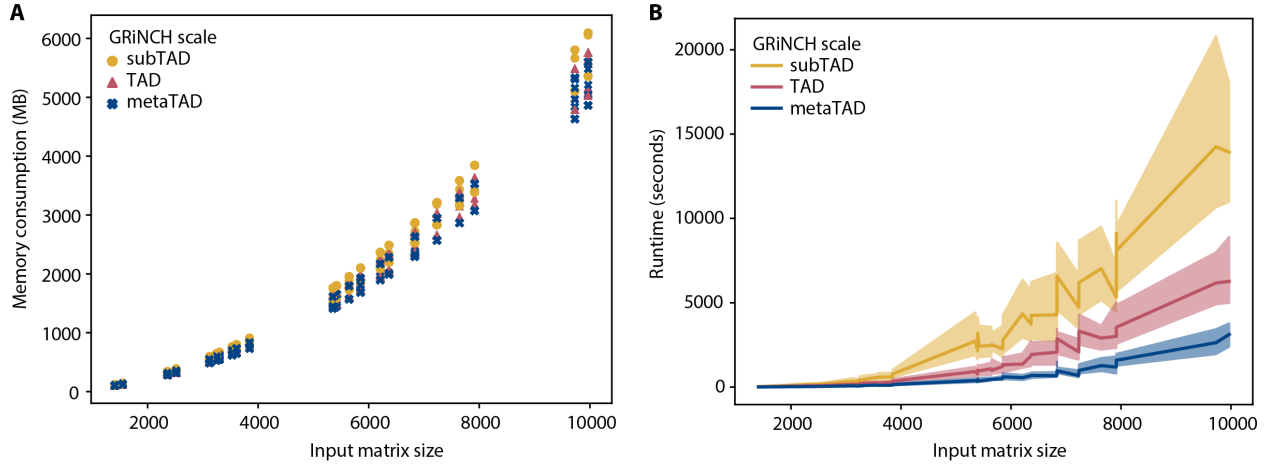


Figure A.17: Memory consumption and runtime trend of the GRiNCH algorithm. **A.** Memory consumption plotted against the input Hi-C matrix size (determined by the size of the chromosome and Hi-C resolution). Each point represents the maximum resident set size for a run of GRiNCH for a given input matrix size. The GRiNCH scales (subTAD, TAD, and metaTAD) represent the chromosome-specific k parameter value, which is set based on the expected size of an output TAD/cluster (500kb, 1Mb, 2Mb respectively); for the same input matrix size, metaTAD setting uses k value that is half of the TAD setting and TAD uses k half of subTAD setting. For a given matrix size and k combination, GRiNCH was run with every combination of regularization parameters $\lambda \in \{1, 10, 100, 100\}$ and neighborhood radius $\in \{25\text{kb}, 50\text{kb}, 100\text{kb}, 250\text{kb}, 500\text{kb}, 1\text{Mb}\}$ as well as without any regularization ($\lambda = 0$). These runs were completed across a distributed computing platform with machines of varying computing power. **B.** Runtime distribution of GRiNCH against the input Hi-C matrix size. GRiNCH runtime was measured for different matrix sizes, TAD scales, and regularization parameters (see A), from start to successful termination of program. The dark line in the middle represents the median runtime of GRiNCH. The top of the shaded area represents the 75th percentile, the bottom 25th percentile of runtimes.

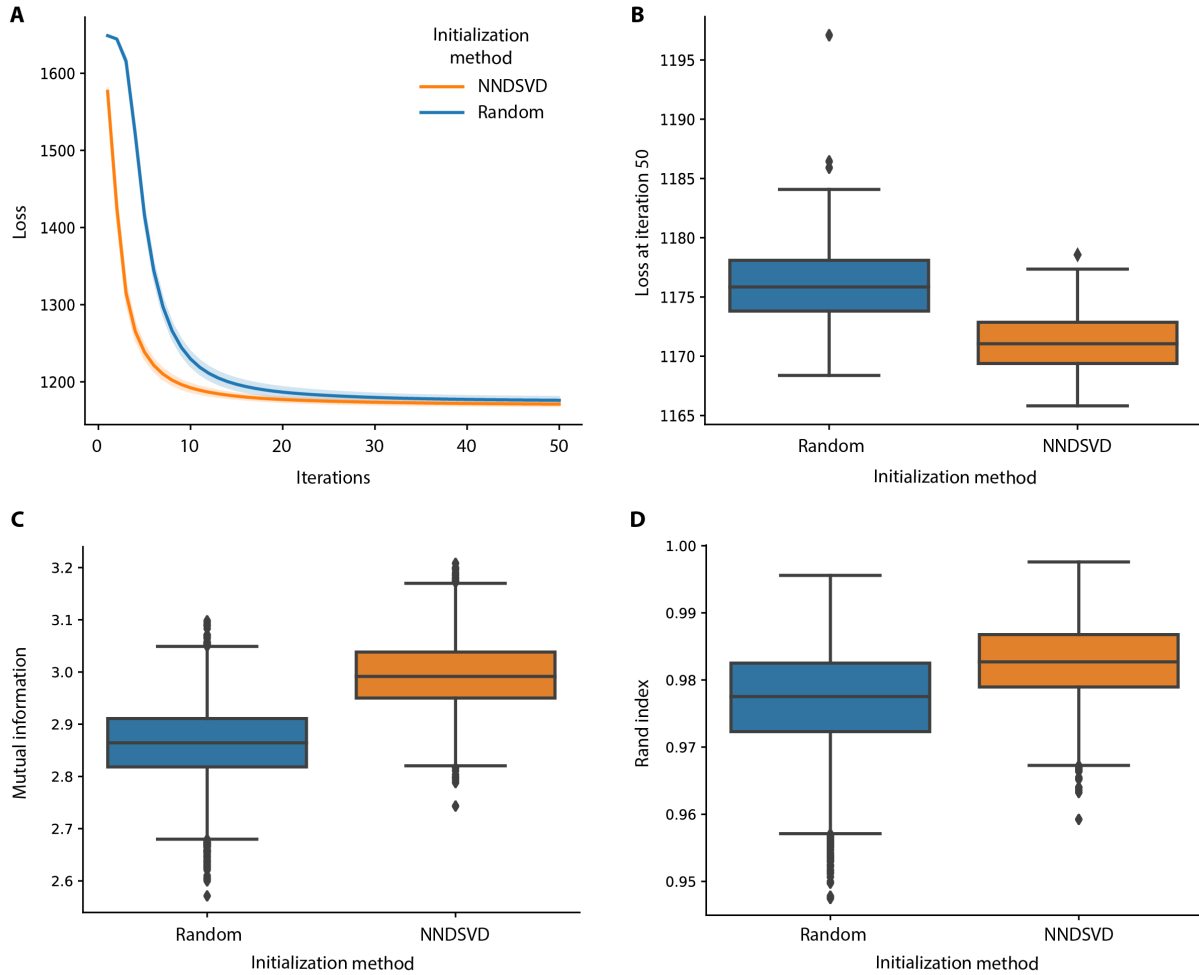


Figure A.18: Comparison of NNDSVD initialization versus random initialization. **A.** Loss as a function of iterations of the GRiNCH algorithm with NNDSVD initialization versus random initialization. For each type of initialization, we tracked the value of the objective/loss over 50 iterations of the GRiNCH algorithm, using 100 different seeds for randomization. The solid line represent the mean loss at a given iteration, and the lightly colored bands around each line is the standard deviation of the loss at the given iteration. **B.** The distribution of the loss after 50 iterations for each type of initialization. Each point on the box plot corresponds to one of 100 different random seeds. **C.** Measuring stability of GRiNCH TADs from NNDSVD initialization versus random initialization. Each box plot shows the distribution of similarity score of GRiNCH TADs using different initialization methods. GRiNCH TADs after 50 iterations from each run with different seeds and initialization methods were converted to clusters. Pairwise similarity of clustering results from seed x and seed y using the same initialization method was measured with mutual information. Higher mutual information means the GRiNCH TADs tend to be more similar and therefore more stable across different seeds. **D.** Measuring stability of GRiNCH TADs with Rand Index. GRiNCH TADs were generated in the same as in **C.** Higher Rand index means the GRiNCH TADs tend to be more similar and therefore more stable across different seeds.

A.2 List of supplementary tables

Table A.1(A) Ranking TAD-calling methods by the percentage of TADs with significant Davies-Bouldin Index across 5 cell lines

Table A.1(B) Ranking TAD-calling methods by the percentage of TADs with significant Delta Contact Count across 5 cell lines

Table A.1(C) Ranking TAD-calling methods by mean (absolute) change in median TAD size as input data resolution changes from 10kb to 25kb to 50kb

Table A.1(D) Ranking TAD-calling methods by mean Rand Index between TADs from two different resolutions of Hi-C data.

Table A.1(E) Ranking TAD-calling methods by mean Mutual Information between TADs from two different resolutions of Hi-C data.

Table A.1(F) Ranking TAD-calling methods by mean Rand Index between TADs from high-depth Gm12878 data and TADs from Gm12878 data downsampled to other cell lines' depth.

Table A.1(G) Ranking TAD-calling methods by mean mutual information between TADs from high-depth Gm12878 data and TADs from Gm12878 data downsampled to other cell lines' depth.

Table A.1(H) Ranking TAD-calling methods by mean fold enrichment of known boundary elements (CTCF, SMC3, RAD21) across 5 cell lines.

Table A.1(I) Ranking TAD-calling methods by mean proportion of TADs with significant histone modification signals across 5 cell lines

Table A.2 Ranking of transcription factors by significant motif enrichment in TAD boundaries across all cell types or time points during mouse pluripotency reprogramming (mouse pre-B cell, D2, D4, D6, D8, iPSC)

Table A.3 Ranking of transcription factors by significant motif enrichment in TAD boundaries across all cell lines (GM12878, HUVEC, HMEC, NHEK, K562) from Rao et al.

A.3 Supplementary method

Below is the pseudocode for chain-constrained k-medoids clustering.

Algorithm A.1: Chain-constrained k-medoids clustering

Input: $U \in \mathbb{R}^{n \times k}$, one of the factors from NMF, and maxIter , the maximum number of iterations

Output: The cluster assignments, $\mathcal{C} \in \{c_1, c_2, \dots, c_n\}$, for each of the chromosomal bins

```

1 Initialize k medoids to be the rows with the largest value from each column of U
2 Initialize an empty priority queue Q
3 while numIter < maxIter do
4     Add current medoids to priority queue Q, with priority value of 0
        // Q orders bins by ascending priority values.
5     while Q is not empty do
6         Pop bin b from Q
7         if b is not assigned to a cluster yet then
            /* First, assign bin to cluster */
8             if b is a medoid then
9                 Assign b to its own cluster
10            else
11                Assign b to either: the same cluster as its nearest upstream neighbor along
                    the chromosome already assigned to a cluster, u, or the same cluster as its
                    nearest downstream neighbor along the chromosome already assigned to a
                    cluster, d, based on the similarity between the latent feature vectors of b
                    and the cluster medoids, i.e.,  $\min_{c \in \{u, d\}} \|U[b, :] - U[\text{medoid of } c, :]\|$ 
12                /* Next, add any unassigned neighbor to priority queue: */
13                for each immediate upstream or downstream neighbor i of b not assigned to a cluster do
                    Add i to Q with priority = priority of b +  $\|U[b, :] - U[i, :]\|$ 
14        Update medoids
15        if sum of distances between each bin and its cluster medoid didn't change from last iteration then
16            Break

```

Appendix B

TGIF supplementary materials

B.1 Supplementary figures

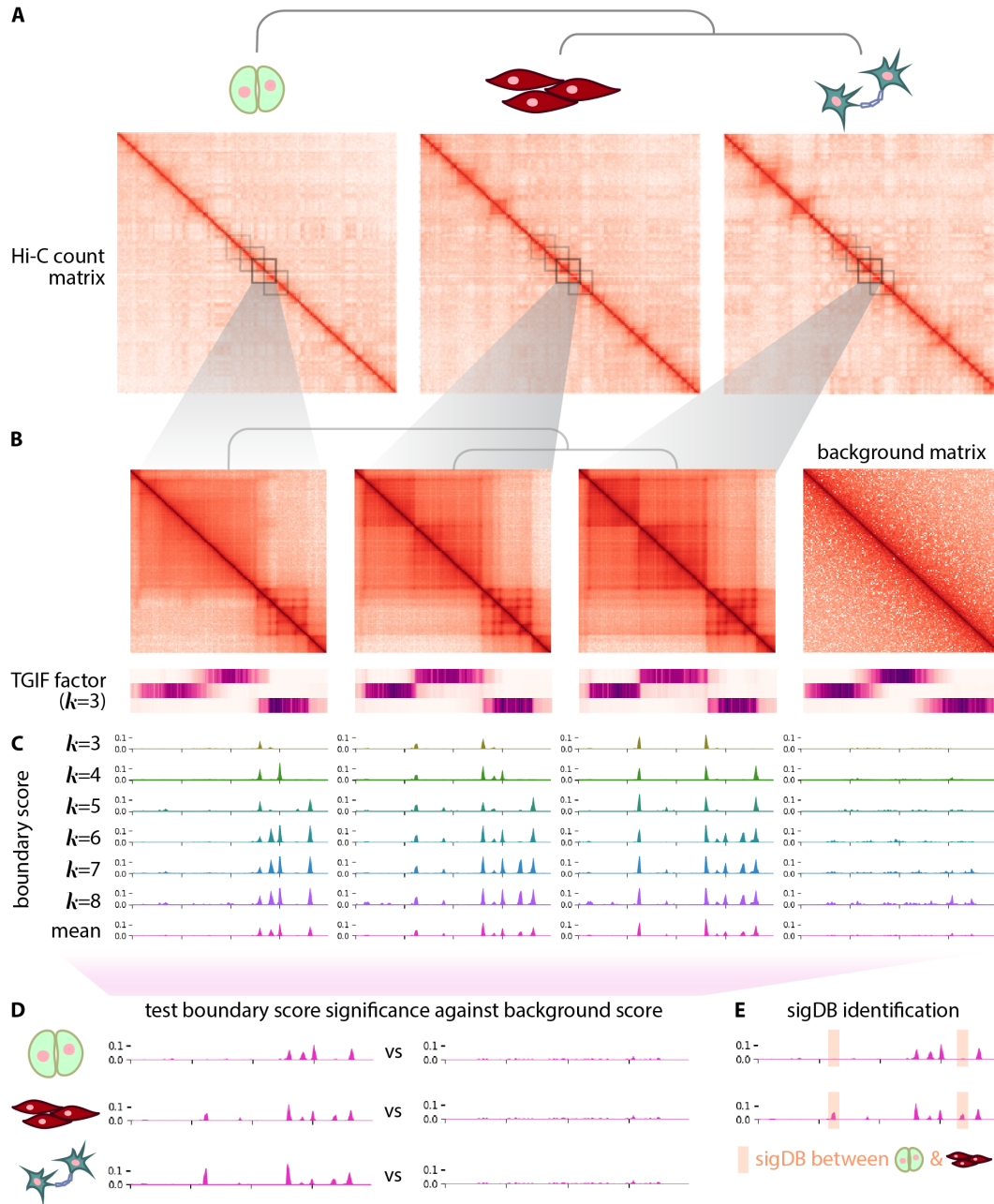


Figure B.1: Overview of TGIF-DB. (A) TGIF-DB takes as input a user-specified tree structure that encodes the relationship among the input Hi-C datasets. Shown is an example of a tree encoding a cell lineage across 3 cell types. (B) Multitask matrix factorization is performed for submatrices along the diagonal of the input intra-chromosomal Hi-C matrices. (C) Within each submatrix, a range of hyper-parameter k values is used to generate boundary scores for TADs at different scales. For each genomic region, the mean boundary score is taken across the range of k s. (D) The mean boundary scores are compared to a “null-distribution” boundary scores generated from a randomly shuffled matrix in order to calculate an empirical p-value for each genomic region and to identify significant boundaries. (E) Significantly differential boundaries (sigDB) are identified for every pair of input matrices (only 1 pair shown here as example).

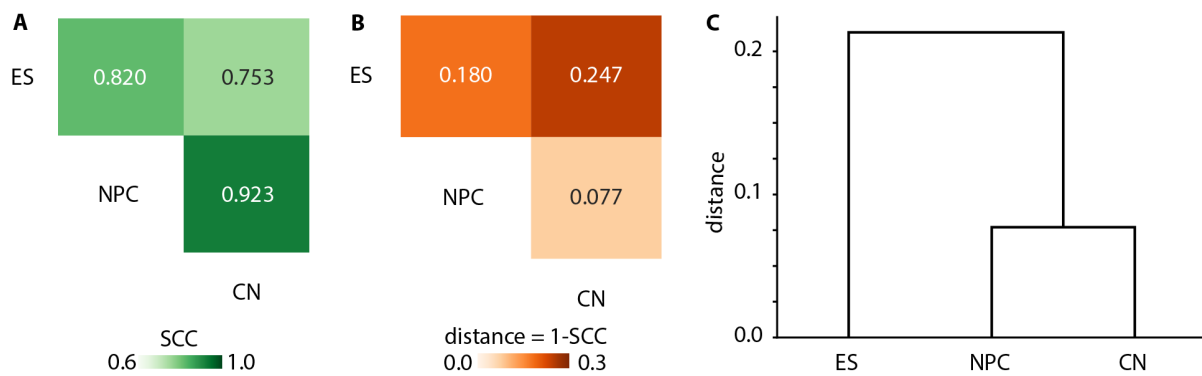


Figure B.2: Inferring a tree structure between input conditions based on the similarity of the input matrices. Here we use mouse chr19 intra-chromosomal matrices (25kb resolution) from neural differentiation data with 3 timepoints (ES, NPC, CN) as an example. (A) Similarity between input matrices for each pair of timepoints, measured by stratum-adjusted correlation coefficient (SCC). **(B)** Similarity converted to distance = 1-SCC. **(C)** Dendrogram of hierarchy constructed using the distance and average linkage.

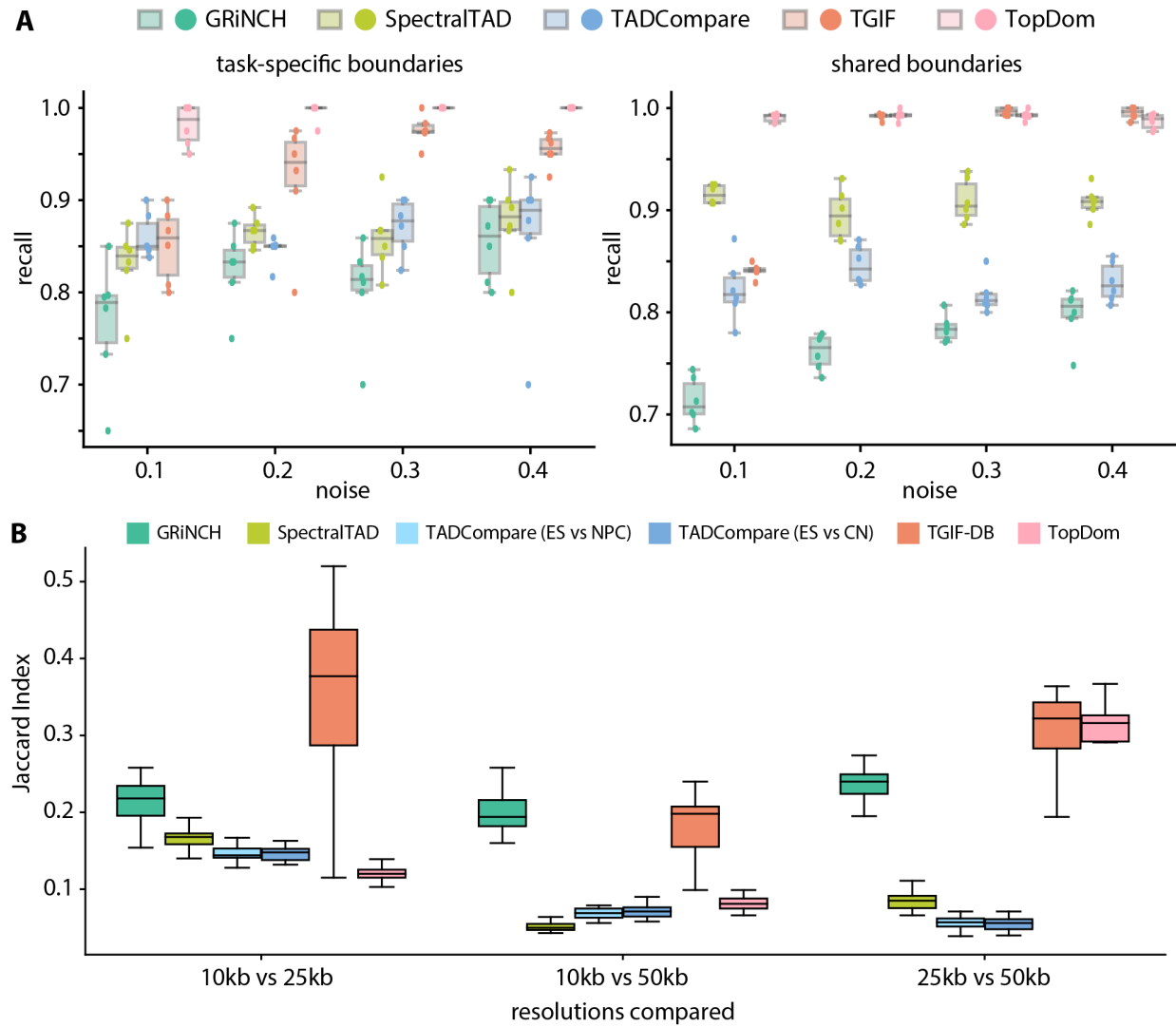


Figure B.3: Benchmarking TGIF-DB. (A) Recall on simulated data. (B) Similarity of boundary sets from input data of different resolutions (10kb, 25kb, and 50kb).

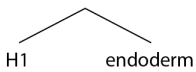
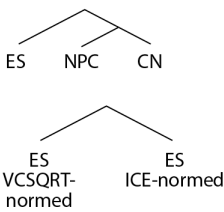
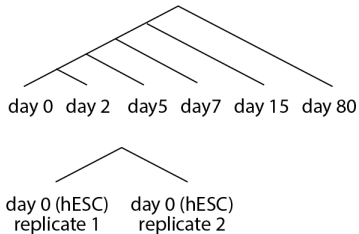
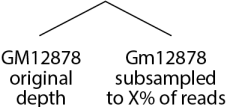
dataset	inputs/timepoints	tree structure	used in analysis
H1-endoderm (Reiff et al., 2022, Dekker et al., 2023)	H1 (hESC), definitive endoderm differentiated from H1		benchmarking: compartment -calling & differential compar- tment methods, DE analysis
mouse neural differentiation (Bonev et al., 2017)	ES (mESC), NPC (neural progenitors), CN (cortical neurons)		benchmarking: TAD stability to resolution, compartment histone mark characterization, DE analysis benchmarking: TAD stability to normalization method
cardiomyocyte differentiation (Zhang et al., 2019)	day 0 (hESC), day 2 (mesoderm), day 5 (cardiac mesoderm), day 7 (cardiac progenitors), day 15 (primitive cardiomyocytes), day 80 (ventricular cardiomyocytes)		benchmarking: CTCF enrich- ment in TAD boundaries, DE analysis, SNP analysis benchmarking: TAD stability across biological replicates
GM12878 cell line (Rao et al.,2014)	GM12878 reads subsampled to 5, 10, 25, 50% of original depth		benchmarking: TAD stability to read depth

Figure B.4: Overview of datasets used in benchmarking and analysis.

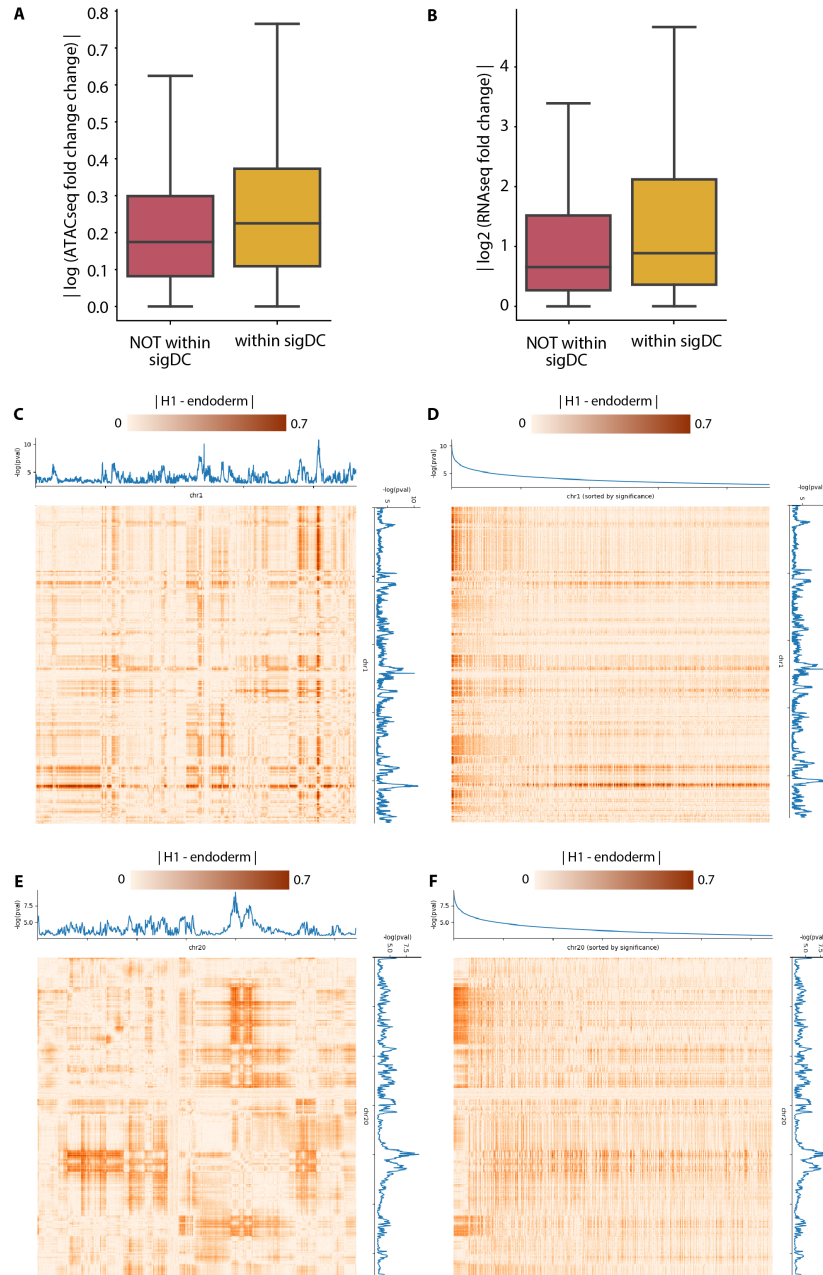


Figure B.5: Characterizing sigDC in H1-endoderm differentiation. (A) Changes in accessibility within significantly differential compartment regions (sigDC). (B) Changes in gene expression level within sigDC. (C) Visualization of the difference in the input matrices (heatmap) and the significance of differences estimated with TGIF-DC (lineplot). Each row and column of the heatmap is a 100kb genomic region of chr1 and each entry in the heatmap = $\text{corr}(\text{O/E})_{\text{H1}} - \text{corr}(\text{O/E})_{\text{endoderm}}$. The lineplot shows $-\log(\text{adjusted p-value})$ from TGIF-DC used for detecting significantly differential compartment regions between H1 and endoderm. (D) Heatmap shows the same information as in (C), but with columns and rows sorted in decreasing significance, i.e., TGIF-DC's $-\log(\text{adjusted p-value})$. The sorting of regions by p-value highlights greater differences in count for regions with higher negative log p-values (high significance). (E) Same visualization as (C), but for chr20. (F) Same visualization as (D), but for chr20.

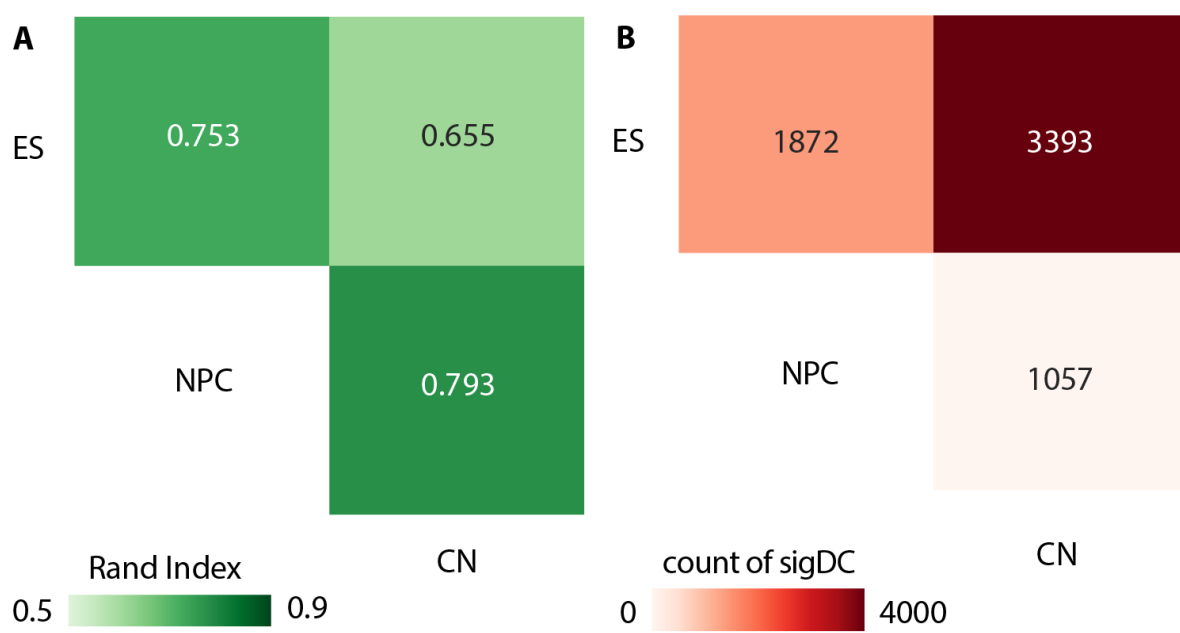


Figure B.6: Characterizing TGIF-DC clusters and sigDC on mouse neural differentiation data. (A) Similarity of cluster assignments for every pair of timepoints/states measured by Rand index. (B) Count of significantly differential compartmental regions (sigDC) for every pair of timepoints/states.

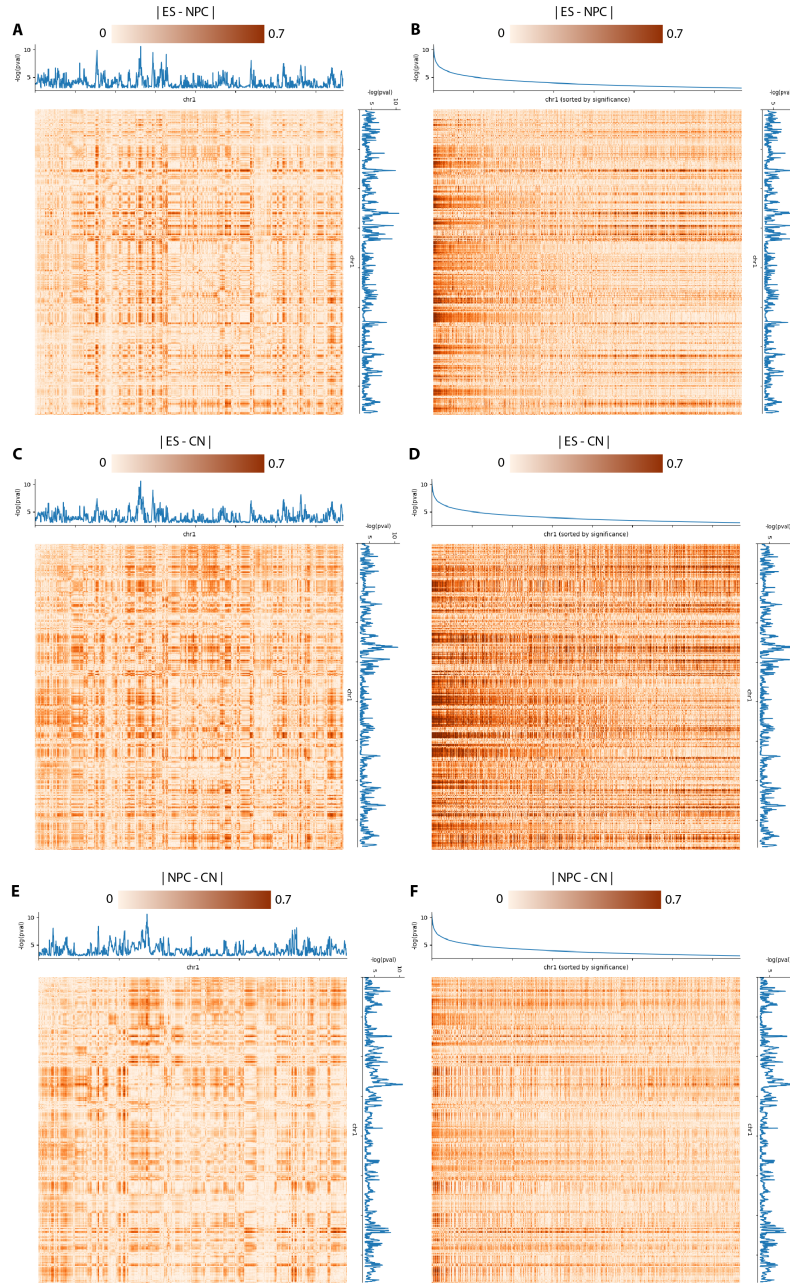


Figure B.7: Heatmap visualization of pairwise difference in O/E counts. (A) Visualization of the difference in the input matrices (heatmap) and the significance of differences estimated with TGIF-DC (lineplot). Each row and column of the heatmap is a 100kb genomic region of chr1 and each entry in the heatmap = $\text{corr}(\text{O/E})_{\text{ES}} - \text{corr}(\text{O/E})_{\text{NPC}}$. The lineplot shows $-\log(\text{adjusted p-value})$ from TGIF-DC used for detecting significantly differential compartment regions between ES and NPC. (B) Same information as in (A), but only the columns are sorted in descending significance. The sorting of regions by p-value highlights greater differences in count for regions with higher negative log p-values (high significance). (C) Same visualization as (A), but for finding sigDC between ES and CN. (D) Same visualization as (B), but for finding sigDC between ES and CN. (E) Same visualization as (A), but for finding sigDC between NPC and CN. (F) Same visualization as (B), but for finding sigDC between NPC and CN.

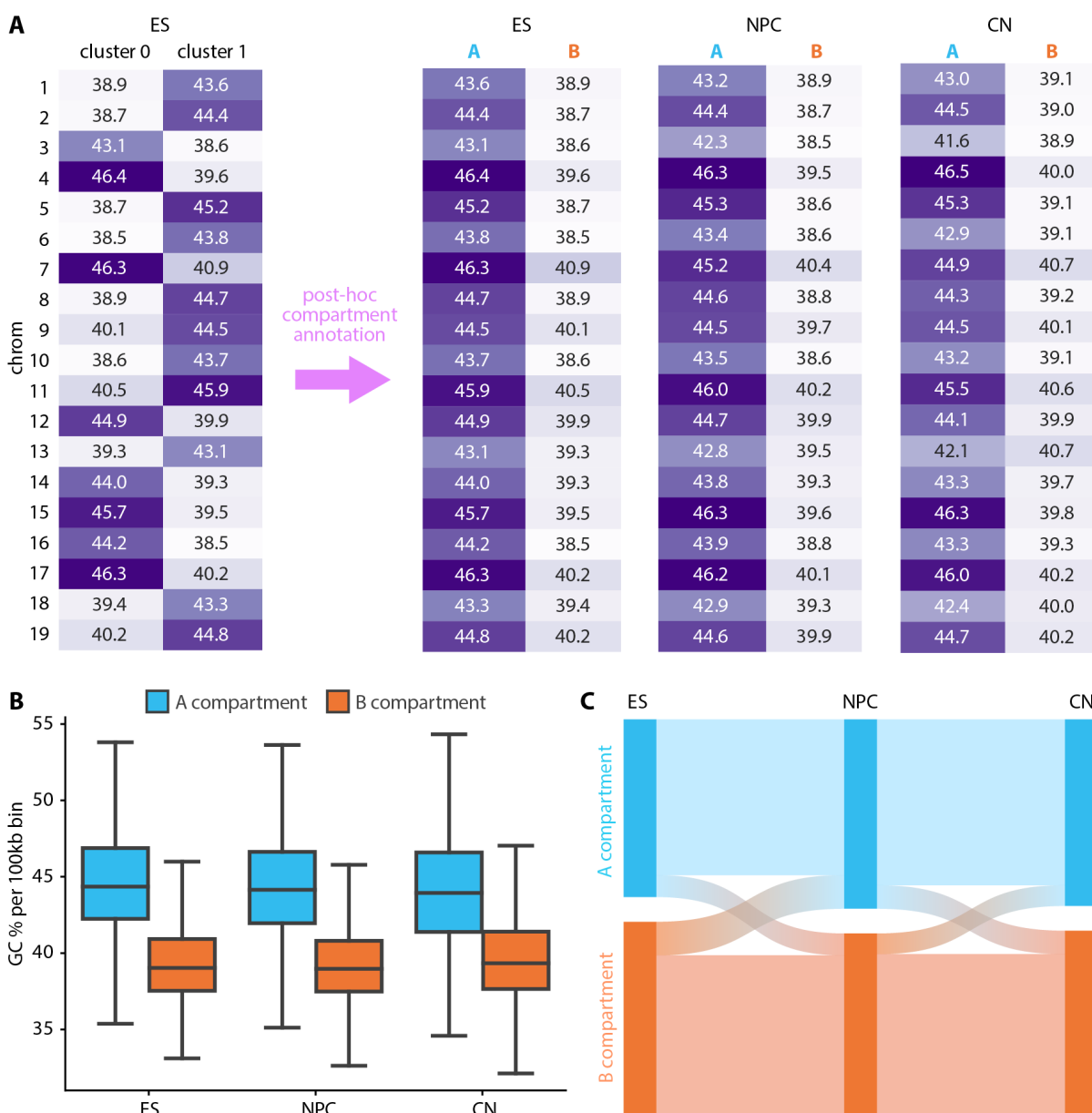


Figure B.8: Post-hoc annotation of TGIF-DC clusters into A and B compartments in mouse neural differentiation data. (A) Within each chromosome, the mean GC percentage is measured for regions in each TGIF-DC cluster for ES. Cluster with higher mean GC content is assigned to compartment A; the other to compartment B across all timepoints. (B) Genome-wide distribution of GC percentage in each 100kb bin by compartment assignment and timepoint. (C) Dynamic compartment assignment patterns for all genomic regions from ES to NPC to CN state.

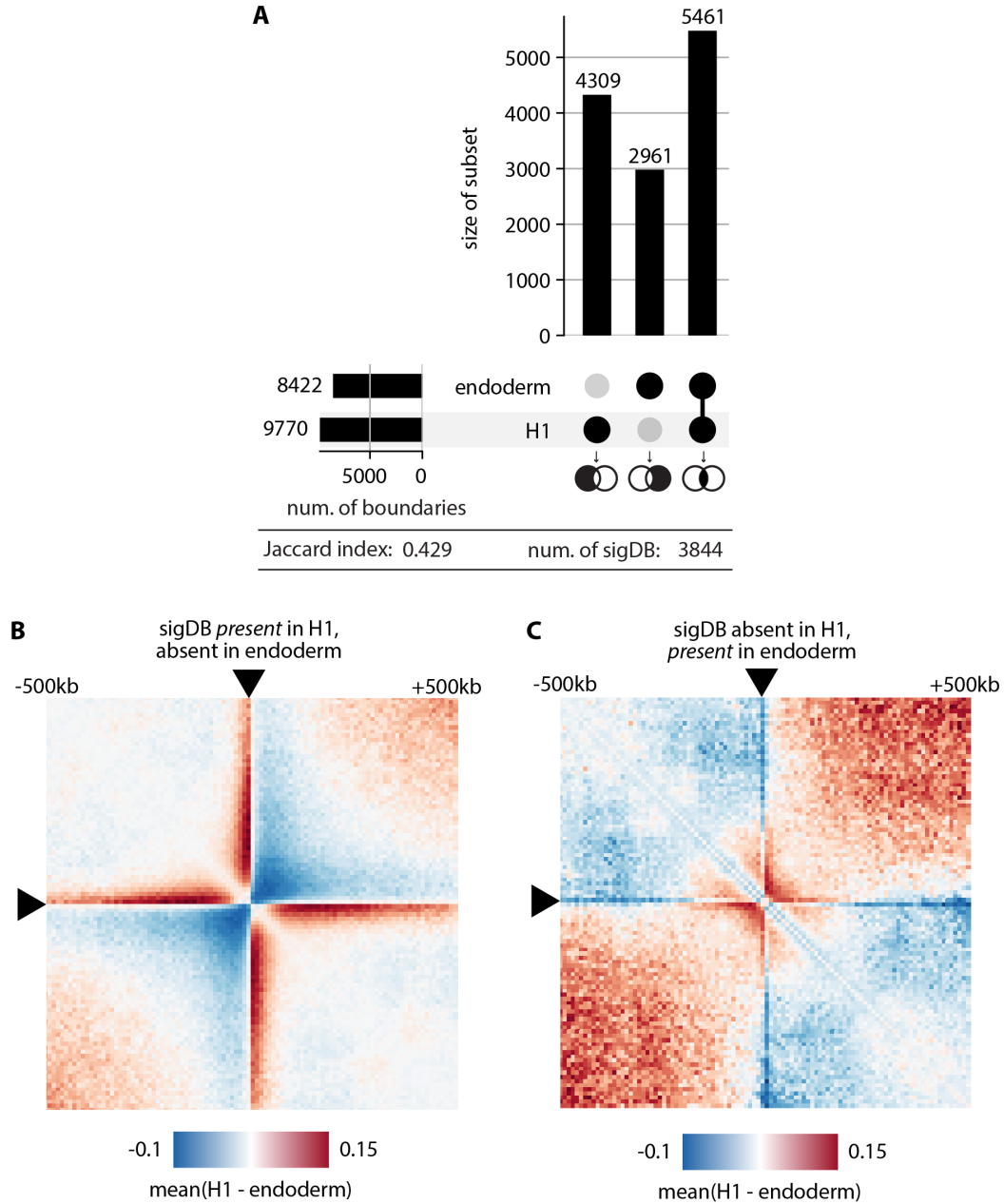


Figure B.9: Characterizing boundaries and sigDB in H1-endoderm differentiation. (A) Number of significant boundaries identified by TGIF-DB in H1 and differentiated endoderm. The vertical bars represent specific subsets only belonging to each category, i.e. boundaries unique to H1; those unique to endoderm; intersection of H1 and endoderm boundaries. The horizontal bars are total counts of boundaries identified in each state. Similarity of the boundary sets between H1 and endoderm is measured by Jaccard index. (B) Mean interaction count difference between H1 and endoderm near sigDB regions (surrounding 1MB window), present in H1 and absent in endoderm. To offset the depth difference between H1 and endoderm, interaction counts were first normalized to O/E matrices. (C) Mean interaction count difference between H1 and endoderm near sigDBs absent in H1 and present in endoderm.

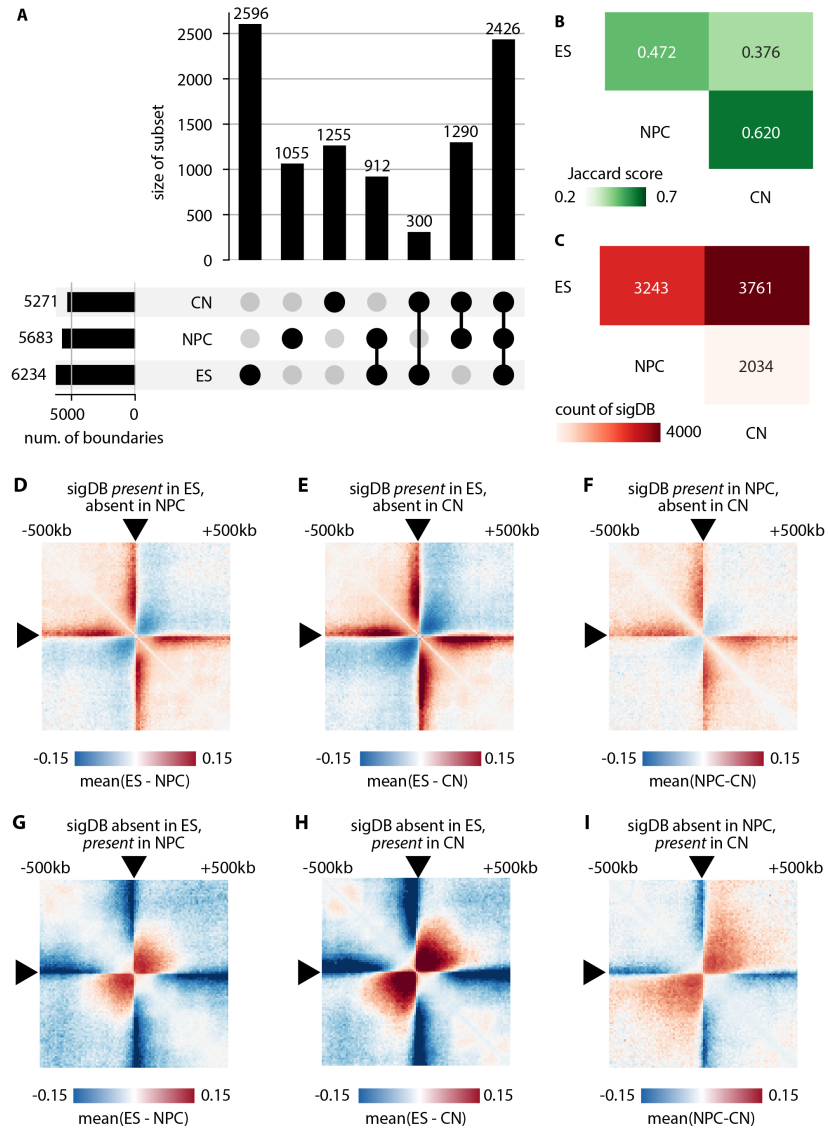


Figure B.10: Characterizing boundaries and sigDB in mouse neural differentiation.

(A) Number of significant boundaries. The vertical bars represent specific subsets only belonging to each category, e.g. boundaries unique to ES, NPC, CN; intersection of ES and NPC boundaries. The horizontal bars are total counts of boundaries identified in each state. (B) Similarity of boundary sets between pairs of timepoints/states, measured by Jaccard index. (C) Count of significantly *differential* boundaries between pairs of timepoints/states. (D) Mean interaction count difference between ES and NPC near sigDB regions (the surrounding 1MB window), present in ES and absent in NPC. To offset the depth difference between ES and NPC, interaction counts were first normalized to O/E matrices. (E) Same visualization as in (D), but for sigDBs present in ES but absent in CN. (F) Same visualization as in (D), but for sigDBs present in NPC but absent in CN. (G) Same visualization as in (D), but for sigDBs absent in ES but present in NPC. (H) Same visualization as in (D), but for sigDBs absent in ES but present in CN. (I) Same visualization as in (D), but for sigDBs absent in NPC but present in CN.

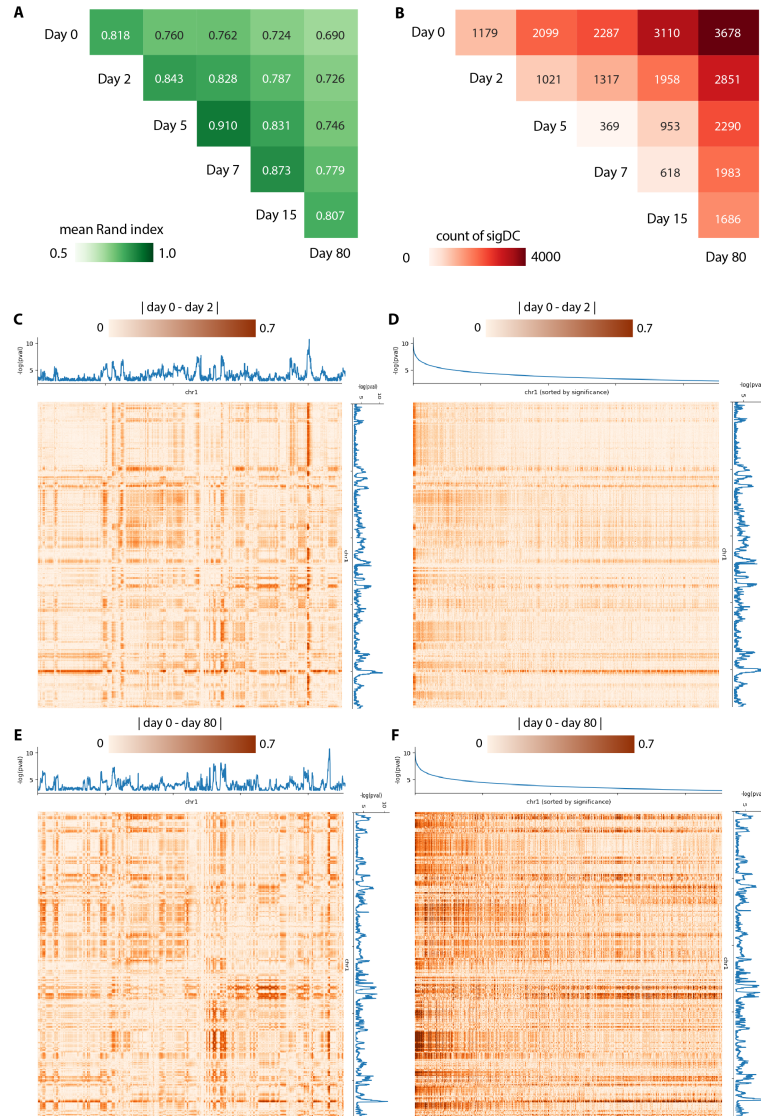


Figure B.11: Characterizing TGIF clusters and sigDC from applying TGIF-DC on cardiomyocyte differentiation data. (A) Similarity of compartment assignments for every pair of timepoints/states measured by Rand index. (B) Count of significantly differential compartmental regions (sigDC) for every pair of timepoints/states. (C) Visualization of the difference in the input matrices (heatmap) and the significance of differences estimated with TGIF-DC (lineplot). Each row and column of the heatmap is a 100kb genomic region of chr1 and each entry in the heatmap = $\text{corr}(\text{O/E})_{\text{day 0}} - \text{corr}(\text{O/E})_{\text{day 2}}$. The lineplot shows $-\log(\text{adjusted p-value})$ from TGIF-DC used for detecting significantly differential compartment regions between day 0 and day 2. (D) Same information as in (C), but only the columns are sorted in descending significance. The sorting of regions by p-value highlights greater differences in count for regions with higher negative log p-values (high significance). (E) Same visualization as (C), but for finding sigDC between day 0 and day 80. (F) Same visualization as (D), but for finding sigDC between day 0 and day 80.

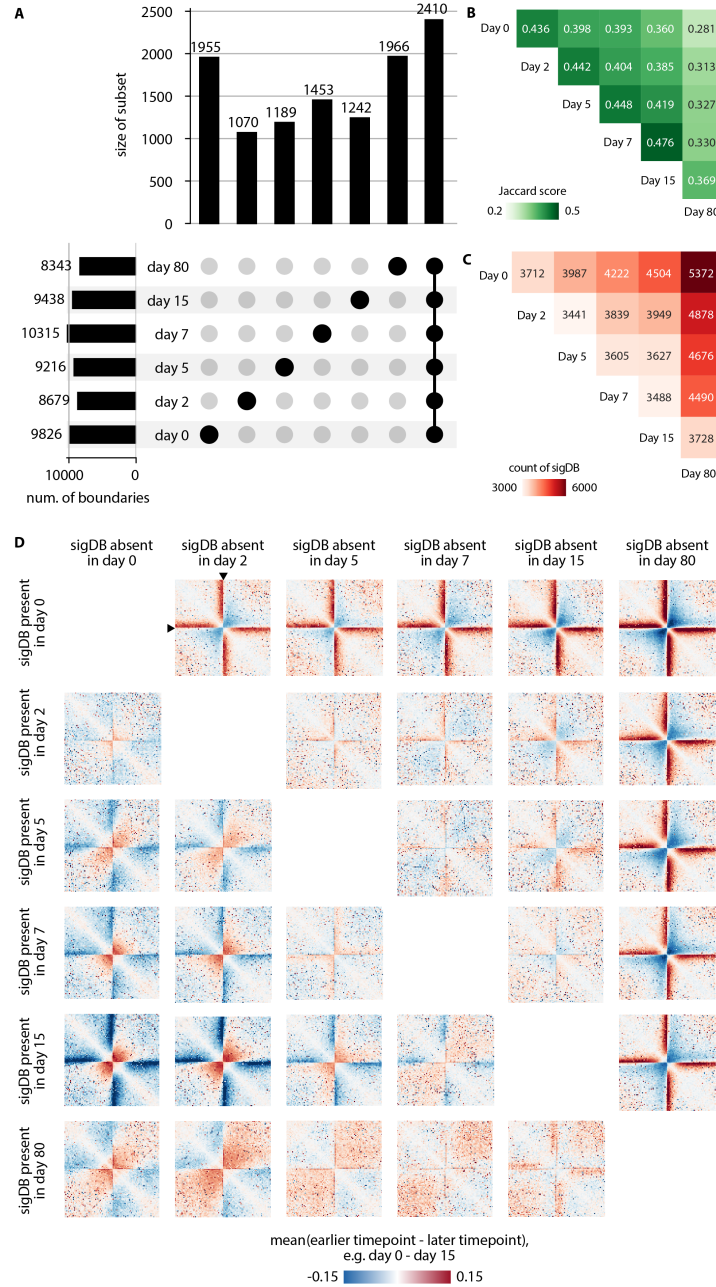


Figure B.12: Characterizing boundaries and sigDB in cardiomyocyte differentiation.

(A) Number of significant boundaries. The vertical bars represent specific subsets only belonging to each category, e.g. boundaries unique to day 0, day 2, etc., intersection of day 0 and 2 boundaries. etc. The horizontal bars are total counts of boundaries identified in each timepoint. (B) Similarity of boundary sets between pairs of timepoints/states, measured by Jaccard index. (C) Count of significantly *differential* boundaries between pairs of timepoints/states. (D) Mean interaction count difference between two timepoints near sigDB regions (the surrounding 1MB window). Each row of the grid represents the timepoint in which the boundary is present and each column the timepoint in which it is absent. To offset the depth difference between the two timepoints, interaction counts were first normalized to O/E matrices. Each heatmap is the count different matrix where the counts from the earlier timepoint is subtracted by those from the later timepoint in the pair as indicated.

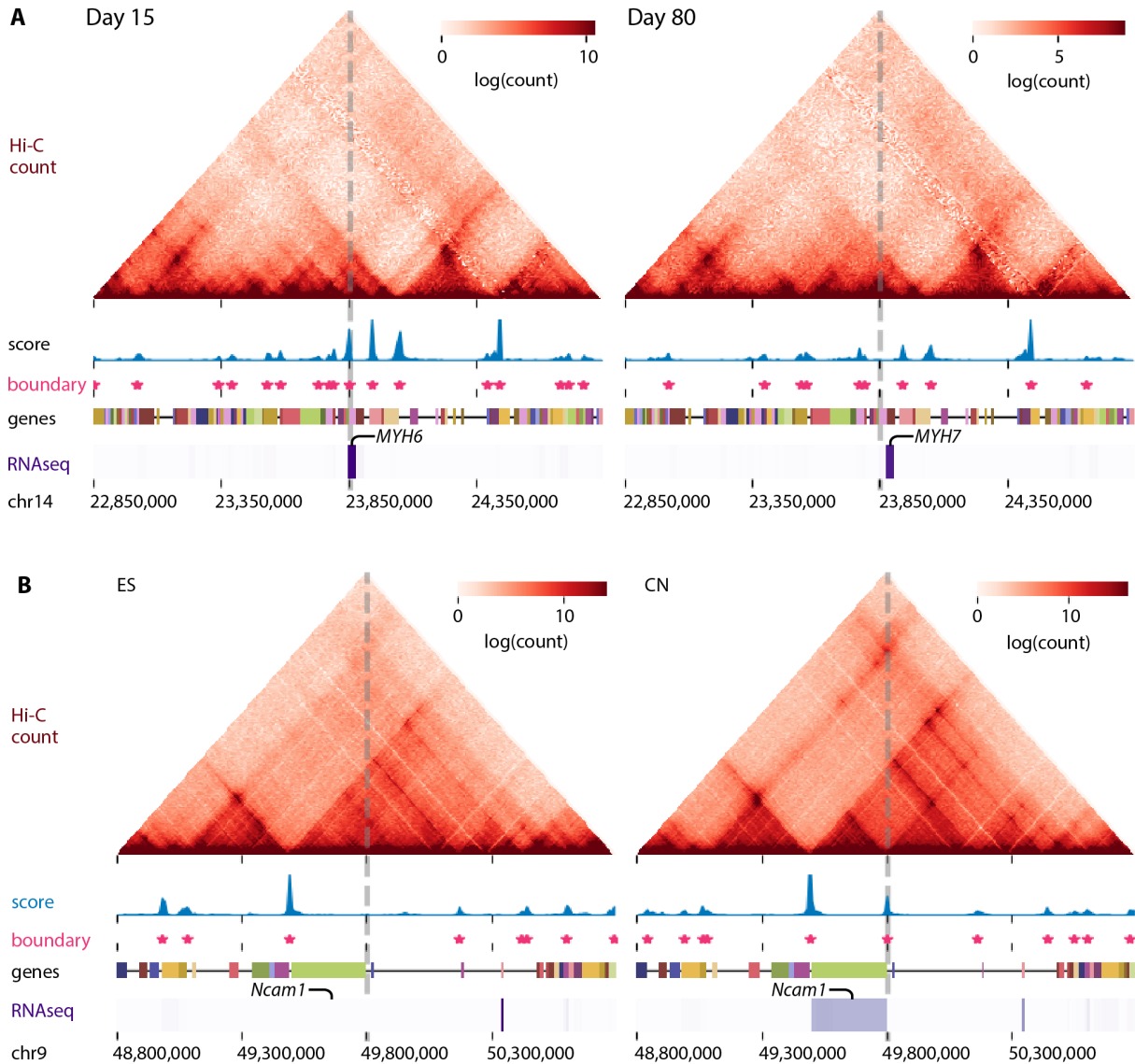


Figure B.13: Examples of differentially expression (DE) gene near significantly differential boundary (sigDB). (A) A highly ranked sigDB (loci marked with dotted vertical line) close to a DE gene, *MYH6*. Shown are the Hi-C interaction matrices for the ES and CN stages of mouse neural differentiation, the boundary score (blue), significant boundary (asterisk), a gene track and the expression heatmap from RNA-seq data. Differential expression of *MYH6* near a sigDB present in day 15 of cardiomyocyte differentiation but absent in day 80 is shown. (B) A sigDB close to a DE gene, *Ncam1*. Differential expression of *Ncam1* near the sigDB that is absent in ES state but appears in CN is shown.

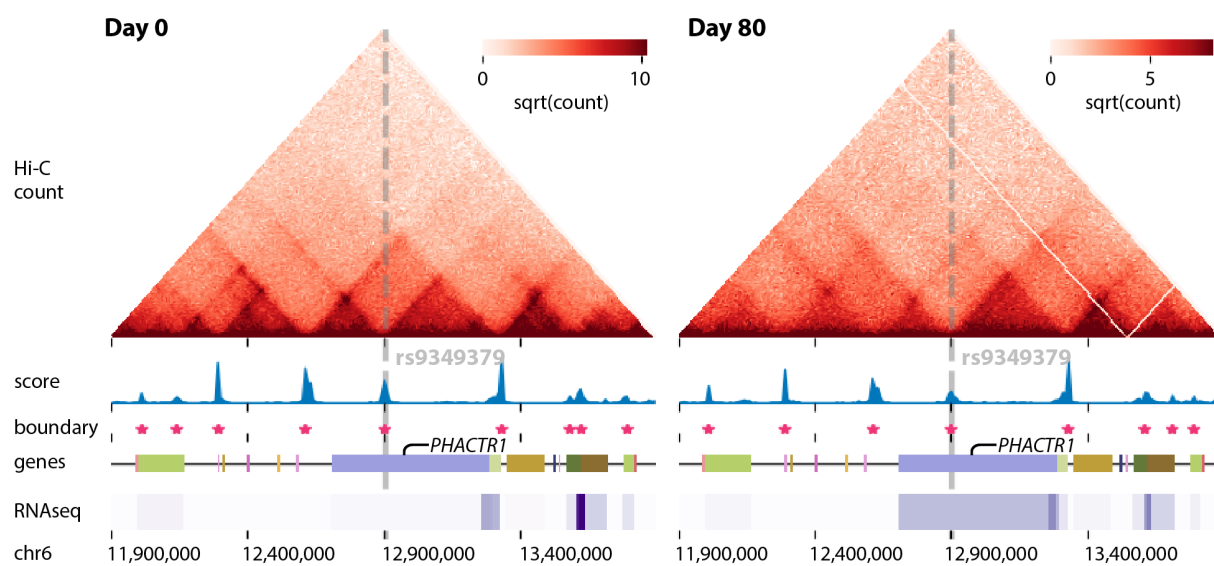


Figure B.14: Persistent TGIF boundaries identified from human cardiomyocyte differentiation containing the cardiovascular disease associated GWAS SNP rs9349379.

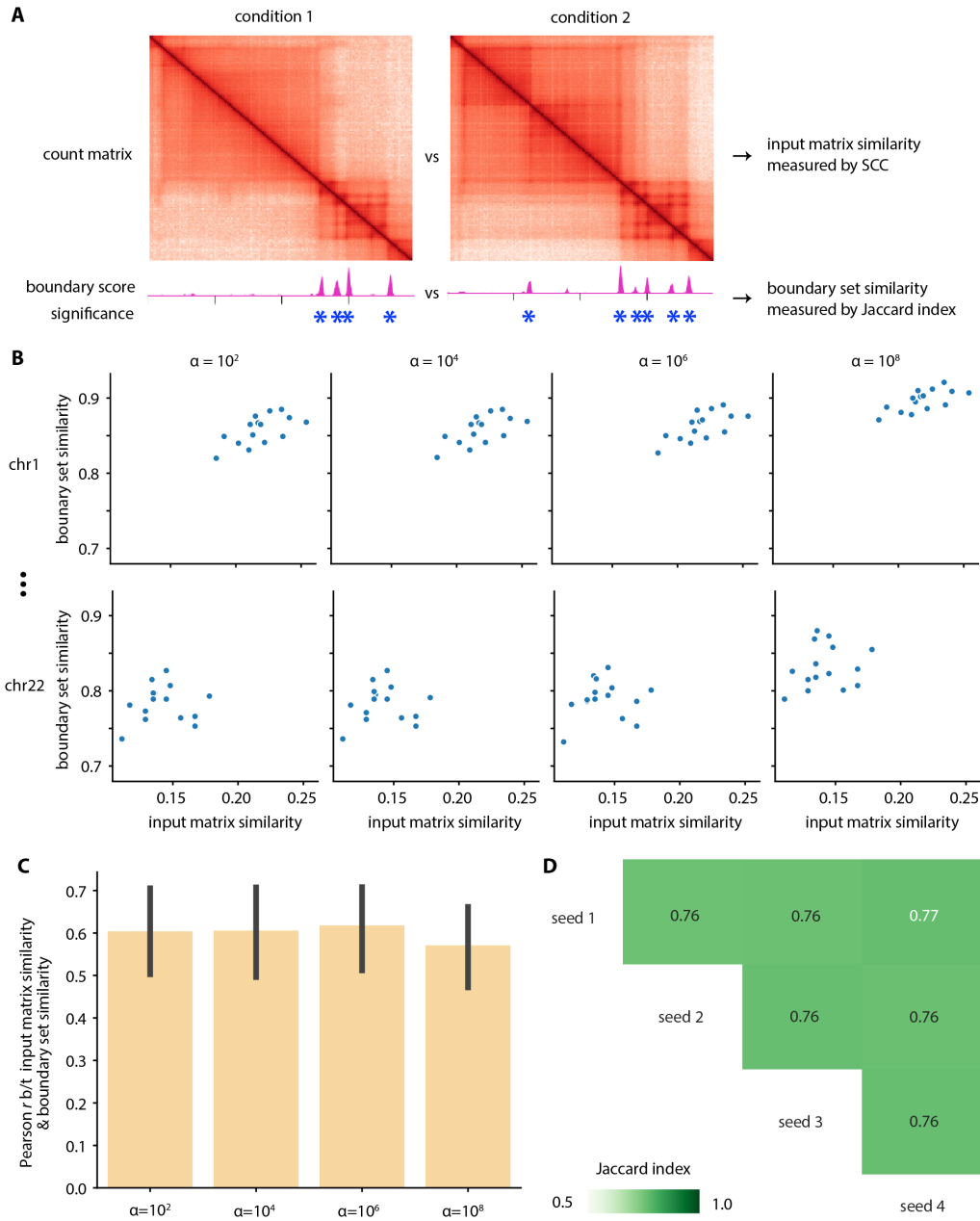


Figure B.15: Hyperparameter α selection for TGIF-DB. (A) The similarity between the input matrices is measured by SCC and the output boundary set agreement measured by Jaccard index. (B) Plotting input matrix similarity vs output boundary set agreement for each α and each chromosome. Each dot represents one pairwise comparison. (C) Correlation between the input matrix similarity and the output boundary set agreement for different values of α . (D) Similarity of boundary sets from different random initialization seeds (with $\alpha = 10^6$), measured by Jaccard index.

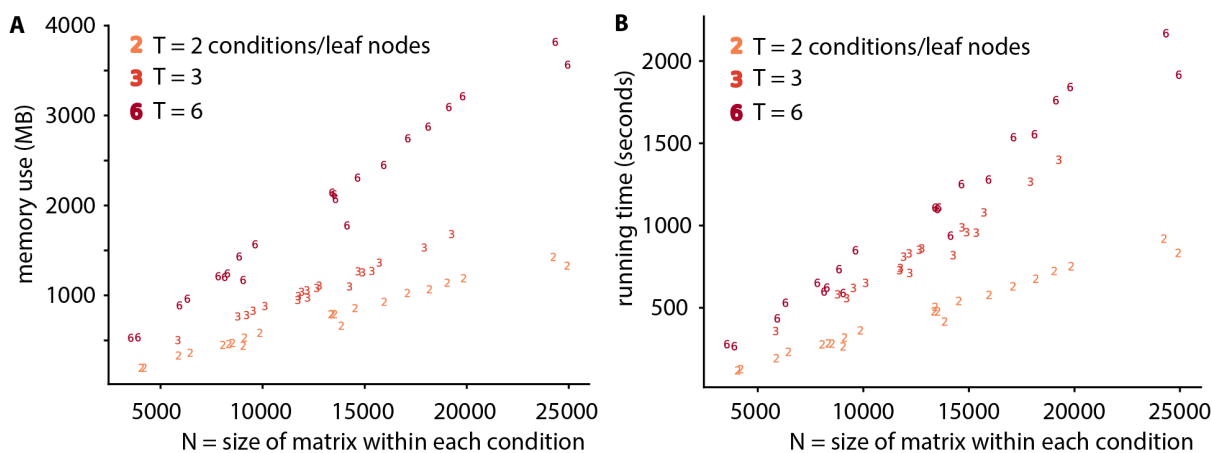


Figure B.16: Resource use by TGIF-DB. (A) Size of each input matrix (100kb resolution) vs memory use in MB. (B) Size of each input matrix vs running time in seconds. T = the number of conditions/states/timepoints in a given run of TGIF-DB.

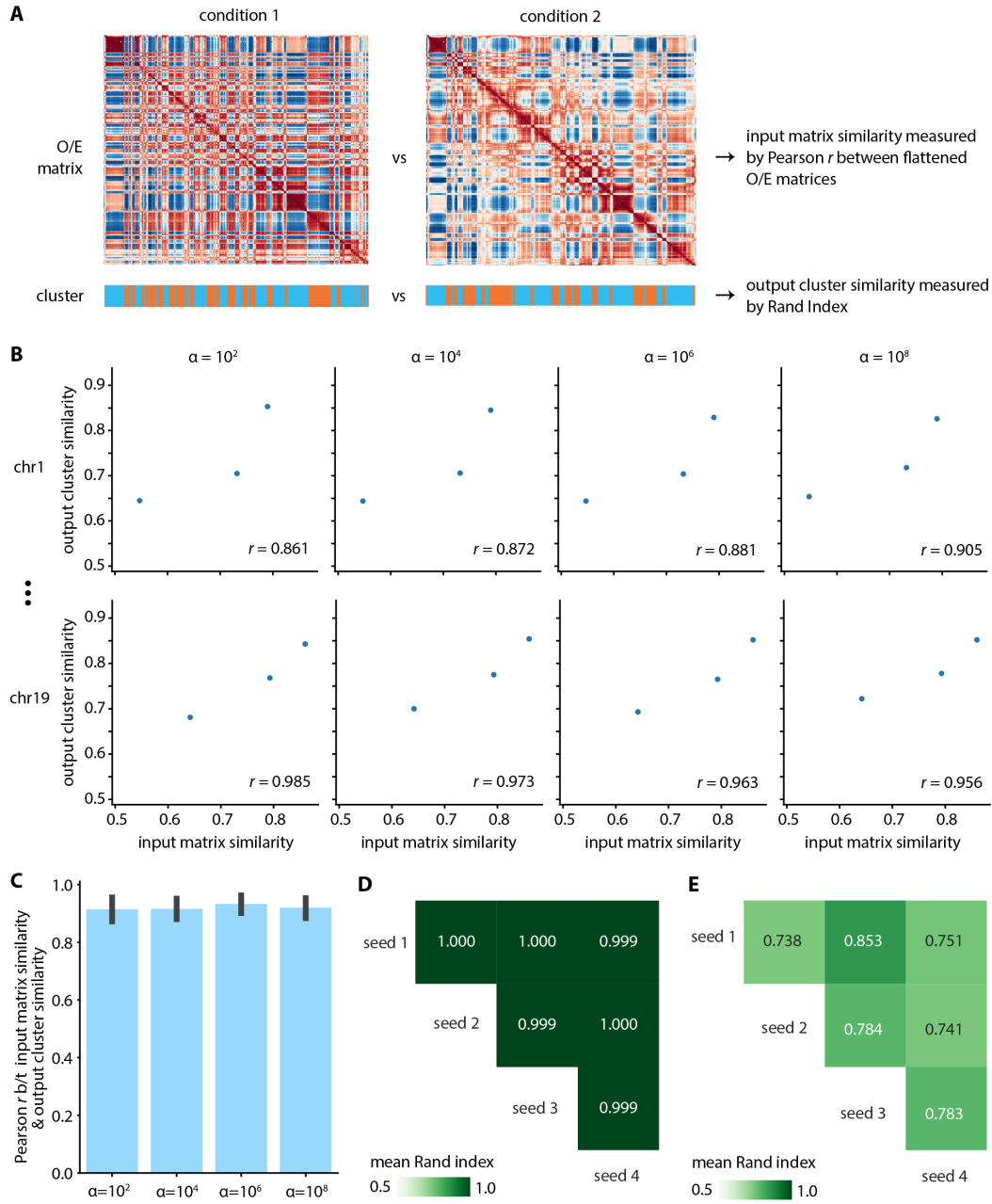


Figure B.17: Hyperparameter α selection for TGIF-DC. (A) The similarity between the O/E count matrices is measured by correlation of the flattened matrices the output cluster similarity measured by Rand index. (B) Plotting input matrix similarity vs output cluster similarity for each α and each chromosome. Each dot represents one pairwise comparison. (C) Correlation between the input matrix similarity and the output cluster similarity for different values of α . (D) Similarity of cluster assignments from different random initialization seeds (with $\alpha = 10^4$), measured by Rand index.

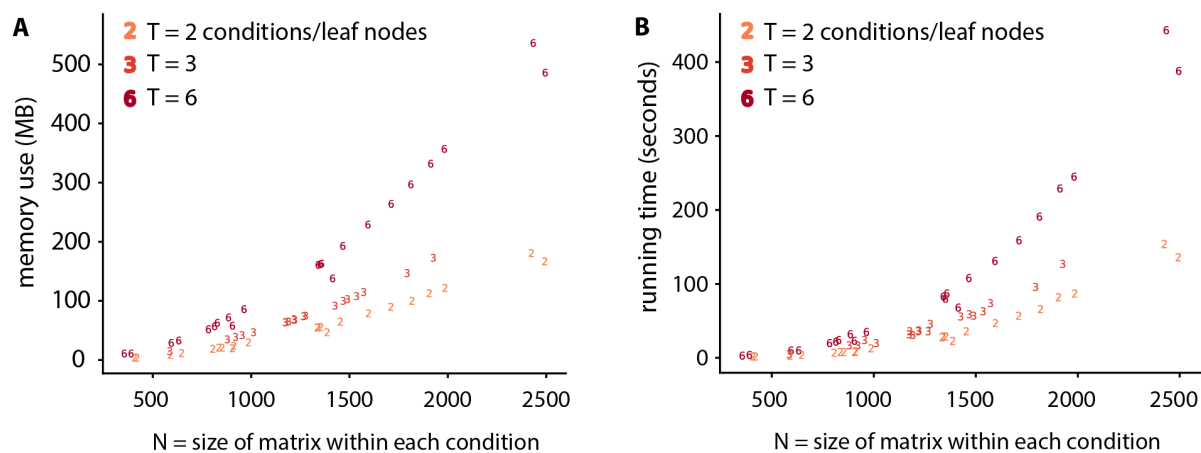


Figure B.18: Resource use by TGIF-DC. (A) Size of each input matrix (100kb resolution) vs memory use in MB. (B) Size of each input matrix vs running time in seconds. T = the number of conditions/states/timepoints in a given run of TGIF-DC.

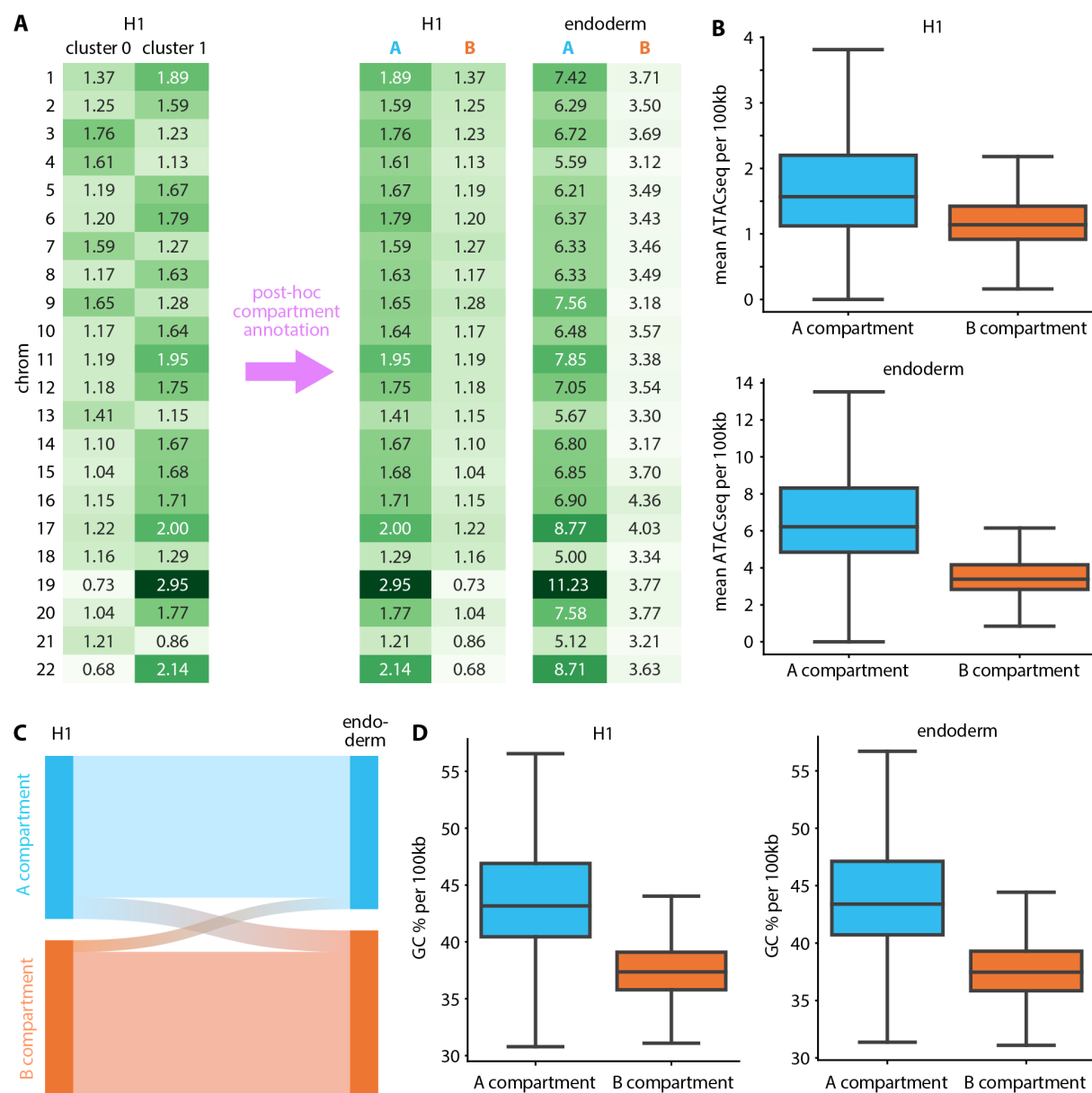


Figure B.19: Post-hoc annotation of TGIF-DC clusters into A and B compartments in H1-endoderm dataset using accessibility. (A) For each chromosome, mean ATACseq signal in each 100kb bin is measured for each TGIF-DC cluster in H1. The cluster with higher mean ATACseq signal is annotated as A compartment; the other cluster B compartment. (B) Genome-wide mean ATACseq signal distribution by compartment in H1 (top) and endoderm (bottom). (C) Dynamic compartment assignment pattern from H1 to endoderm. (D) Genome-wide mean GC content per 100kb bin by compartment in H1 and endoderm.

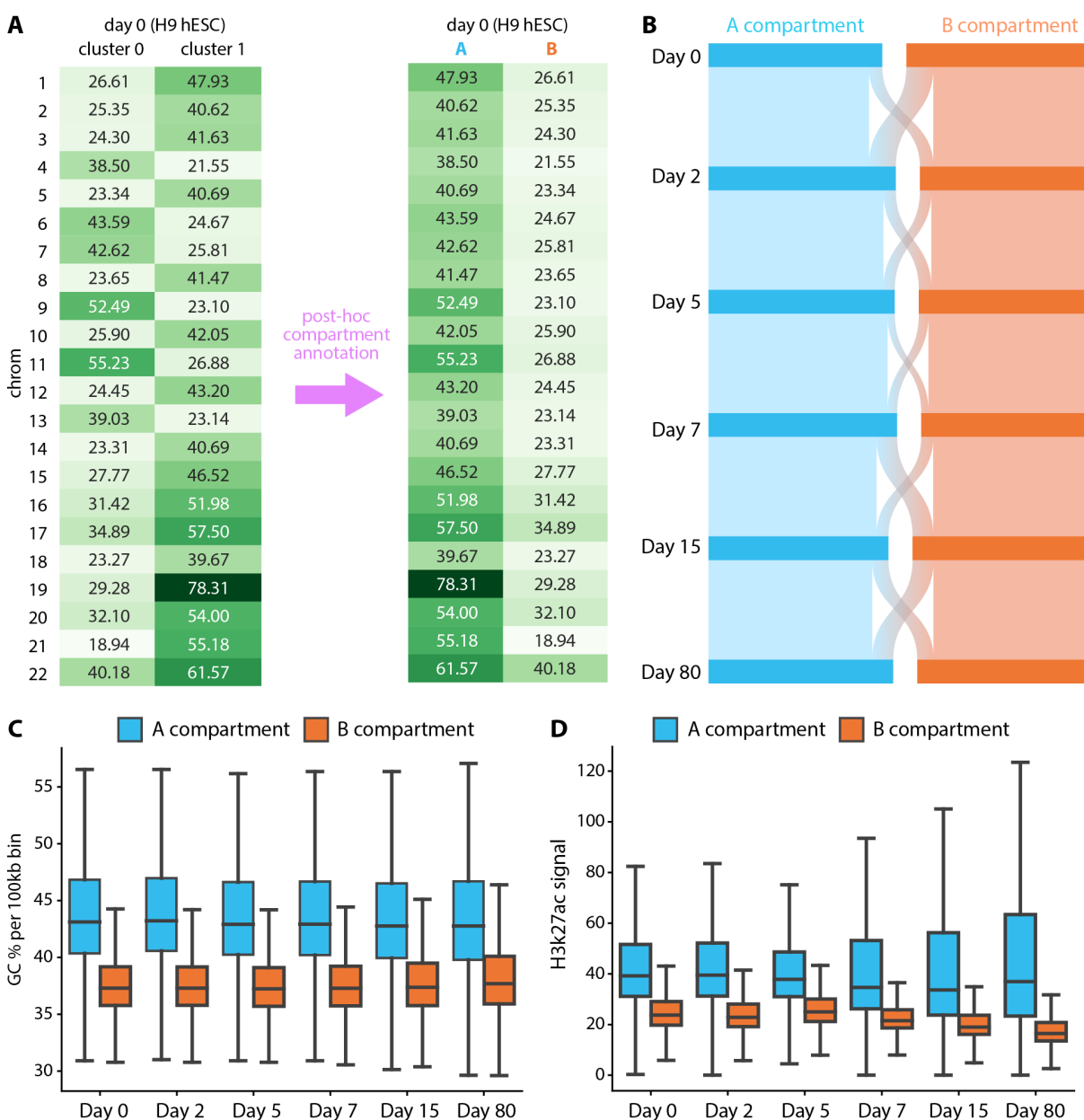


Figure B.20: Post-hoc annotation of TGIF-DC clusters into A and B compartments in cardiomyocyte differentiation dataset using accessibility. (A) For each chromosome, mean DNaseq signal in each 100kb bin is measured for each TGIF-DC cluster in day 0. The cluster with higher mean DNaseq signal is annotated as A compartment; the other cluster B compartment. (B) Dynamic compartment assignment pattern from day 0 to day 80. (C) Genome-wide mean GC content per 100kb bin by compartment across all timepoints. (D) Genome-wide mean H3k27ac signal distribution by compartment across all timepoints.

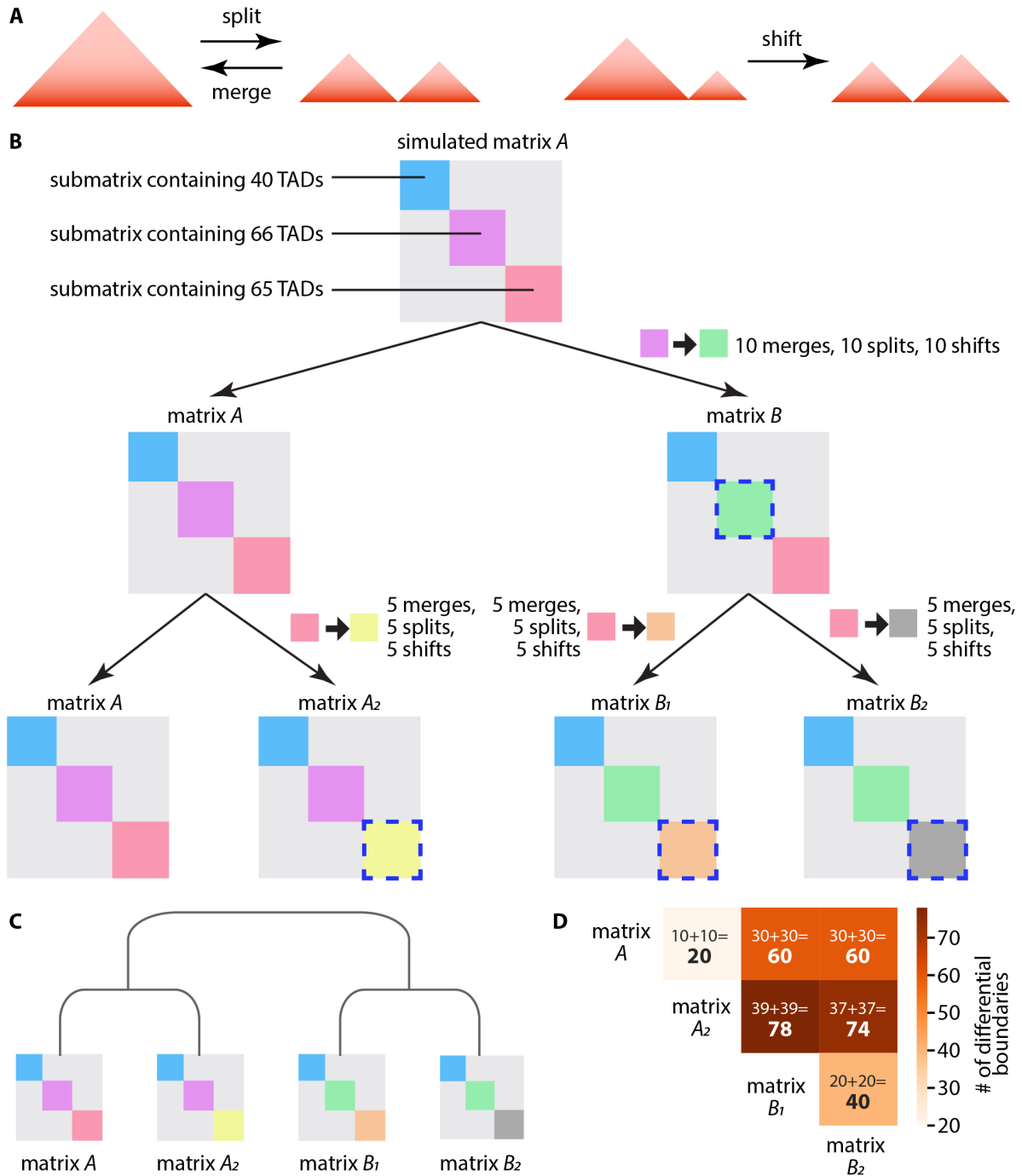


Figure B.21: TAD boundary perturbation procedure for Hi-C data simulation.. **A.** Three different types of TAD boundary changes used in simulation: TAD split, merge, and shift. **B.** Different sets of TADs were split, merged, or shifted within the original simulated TAD set to yield four different simulated matrices. **C.** Tree structure input to TGIF for its application to the simulated matrices. **D.** Number of different TADs between pairs of simulated matrices.

B.2 List of supplementary tables

Table B.1 4D Nucleome accession numbers for H1 and endoderm data.

Table B.2 Data sources and GEO accession numbers for mouse neural differentiation data.

Table B.3 GEO and ENCODE accession numbers for cardiomyocyte differentiation data.

Table B.4 4D Nucleome accession number for GM12878.

Table B.5 Differential expression fold enrichment analysis on H1-endoderm dataset. Each row represents one fold enrichment hypergeometric test.

Table B.6 Differential expression fold enrichment analysis on mouse neural differentiation dataset. Each row represents one fold enrichment hypergeometric test.

Table B.7 Differential expression fold enrichment analysis on cardiomyocyte dataset. Each row represents one fold enrichment hypergeometric test.

Table B.8 GO enrichment analysis on cardiomyocyte differentiation dataset. Each entry is the negative log p-value of GO biological processes for two subsets of differentially expressed (DE) genes: DE genes close to sigDB, and DE genes not close to sigDB. Each column is a pair of timepoints used to determine DB and DE genes. Each row is a GO term (only GO terms with p-value $< 1e-5$ in at least one pair of timepoints is listed). Non-blank entries are the negative log p-value of the GO term enrichment.

Table B.9 GO enrichment analysis on H1-endoderm dataset. Each entry is the negative log p-value of GO biological processes for two subsets of differentially expressed (DE)

genes: DE genes close to sigDB, and DE genes not close to sigDB. Each column is a pair of timepoints used to determine DB and DE genes. Each row is a GO term (only GO terms with $p\text{-value} < 1e-5$ in at least one pair of timepoints is listed). Non-blank entries are the negative log p -value of the GO term enrichment.

Table B.10 GO enrichment analysis on mouse neural differentiation dataset. Each entry is the negative log p -value of GO biological processes for two subsets of differentially expressed (DE) genes: DE genes close to sigDB, and DE genes not close to sigDB. Each column is a pair of timepoints used to determine DB and DE genes. Each row is a GO term (only GO terms with $p\text{-value} < 1e-5$ in at least one pair of timepoints is listed). Non-blank entries are the negative log p -value of the GO term enrichment.

B.3 Supplementary methods

Below we walk through the derivation for the Block Coordinate Descent (BCD) optimization and update rules for Tree-Guided Integrated Factorization (TGIF).

Notation and objective

Given $t \in \{1, \dots, T\}$ tasks, each with input matrix $X^{(t)} \in \mathbb{R}^{n_t \times m}$, related to each other in a task hierarchy/tree with a set of nodes $c \in \{r\} \cup \mathcal{B} \cup \mathcal{T}$ where r is the root node, \mathcal{B} a set of internal (or branch) nodes $b \in \mathcal{B}$, and \mathcal{T} a set of the task-specific leaf nodes, the objective is:

$$O = \sum_{t=1}^T \|X^{(t)} - U^{(t)} V^{(t)\top}\|_F^2 + \alpha \sum_c \|V^{(c)} - V^{\text{Pa}(c)}\|_F^2 \quad (\text{B.1})$$

where $U^{(t)} \in \mathbb{R}^{n_t \times k}$, $V^{(\cdot)} \in \mathbb{R}^{m \times k}$, $k \ll n, m$. The regularization term will:

1. constrain a task-specific latent feature factor $V^{(t)}$ in a leaf node of the task hierarchy to be similar to $V^{\text{Pa}(t)}$ in its parent node;
2. constrain an internal node's latent feature factor $V^{(b)}$ to be similar to its direct child nodes' $V^{(c)}$ and its parent node's $V^{\text{Pa}(b)}$; and
3. constrain the root node's latent feature factor $V^{(r)}$ to be similar to all of its direct child nodes' $V^{(c)}$ s.

Breaking down to task-level and column-level subproblems

The objective can be re-written as:

$$O = \sum_{t=1}^T \left\| X^{(t)} - \sum_k u_k^{(t)} v_k^{(t)\top} \right\|_F^2 + \alpha \sum_c \sum_k \|v_k^{(c)} - v_k^{\text{Pa}(c)}\|_2^2 \quad (\text{B.2})$$

Where $\mathbf{u}_k^{(t)} \in \mathbb{R}^{n_t}$ is the k th column vector of $\mathbf{U}^{(t)}$ and $\mathbf{v}_k^{(t)} \in \mathbb{R}^m$ is the k th column vector of $\mathbf{V}^{(t)}$. Now we ‘pull out’ terms involving the k th column in all factors:

$$O = \sum_{t=1}^T \left\| \mathbf{X}^{(t)} - \mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} - \sum_{j \neq k} \mathbf{u}_j^{(t)} \mathbf{v}_j^{(t)\top} \right\|_F^2 + \alpha \sum_c \left(\left\| \mathbf{v}_k^{(c)} - \mathbf{u}_k^{p_a(c)} \right\|_2^2 + \sum_{j \neq k} \left\| \mathbf{v}_j^{(c)} - \mathbf{v}_j^{p_a(c)} \right\|_2^2 \right) \quad (\text{B.3})$$

Now we will substitute with $\mathbf{R}_k^{(t)} = \mathbf{X}^{(t)} - \sum_{j \neq k} \mathbf{u}_j^{(t)} \mathbf{v}_j^{(t)\top}$:

$$O = \sum_{t=1}^T \left\| \mathbf{R}_k^{(t)} - \mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} \right\|_F^2 + \alpha \sum_c \left\| \mathbf{v}_k^{(c)} - \mathbf{v}_k^{p_a(c)} \right\|_2^2 + \alpha \sum_c \sum_{j \neq k} \left\| \mathbf{v}_j^{(c)} - \mathbf{v}_j^{p_a(c)} \right\|_2^2 \quad (\text{B.4})$$

We can now attempt to optimize $\mathbf{u}_k^{(t)}$ and $\mathbf{v}_k^{(\cdot)}$, fixing all other parameters to be constant.

Optimize $\mathbf{v}_k^{(t)}$

To find $\mathbf{v}_k^{(t)}$ for each leaf node task t that minimizes the objective, we find the derivative of the objective with respect to $\mathbf{v}_k^{(t)}$ and set it to 0, then solve. First we expand the objective into matrix multiplications:

$$O = \left\| \mathbf{R}_k^{(t)} - \mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} \right\|_F^2 + \alpha \left\| \mathbf{v}_k^{(t)} - \mathbf{v}_k^{p_a(t)} \right\|_2^2 + C \quad (\text{B.5})$$

$$= \text{Tr} \left[\left(\mathbf{R}_k^{(t)} - \mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} \right)^\top \left(\mathbf{R}_k^{(t)} - \mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} \right) \right] + \alpha \left(\mathbf{v}_k^{(t)} - \mathbf{v}_k^{p_a(t)} \right)^\top \left(\mathbf{v}_k^{(t)} - \mathbf{v}_k^{p_a(t)} \right) + C \quad (\text{B.6})$$

Here C subsumes all elements of the objective that does not involve $\mathbf{v}_k^{(t)}$ (including terms involving tasks other than t), since they will be zeroed out when the derivative is

taken with respect to $\mathbf{v}_k^{(t)}$. Now we keep expanding:

$$O = \text{Tr} \left[\mathbf{R}_k^{(t)\top} \mathbf{R}_k^{(t)} - 2\mathbf{R}_k^{(t)\top} \mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} + \left(\mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} \right)^\top \left(\mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} \right) \right] \quad (\text{B.7})$$

$$+ \alpha \left(\mathbf{v}_k^{(t)\top} \mathbf{v}_k^{(t)} - 2\mathbf{v}_k^{(t)\top} \mathbf{v}_k^{\text{Pa}(t)} + \mathbf{v}_k^{\text{Pa}(t)\top} \mathbf{v}_k^{\text{Pa}(t)} \right) + C \quad (\text{B.8})$$

$$= \text{Tr} \left(\mathbf{R}_k^{(t)\top} \mathbf{R}_k^{(t)} \right) - 2 \text{Tr} \left(\mathbf{R}_k^{(t)\top} \mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} \right) + \text{Tr} \left(\mathbf{v}_k^{(t)} \mathbf{u}_k^{(t)\top} \mathbf{u}_k^{(t)} \mathbf{v}_k^{(t)\top} \right) \quad (\text{B.9})$$

$$+ \alpha \mathbf{v}_k^{(t)\top} \mathbf{v}_k^{(t)} - 2\alpha \mathbf{v}_k^{(t)\top} \mathbf{v}_k^{\text{Pa}(t)} + \alpha \mathbf{v}_k^{\text{Pa}(t)\top} \mathbf{v}_k^{\text{Pa}(t)} + C \quad (\text{B.10})$$

$$= \text{Tr} \left(\mathbf{R}_k^{(t)\top} \mathbf{R}_k^{(t)} \right) - 2 \left(\mathbf{R}_k^{(t)\top} \mathbf{u}_k^{(t)} \right)^\top \mathbf{v}_k^{(t)} + \left(\mathbf{u}_k^{(t)\top} \mathbf{u}_k^{(t)} \right) \left(\mathbf{v}_k^{(t)\top} \mathbf{v}_k^{(t)} \right) \quad (\text{B.11})$$

$$+ \alpha \mathbf{v}_k^{(t)\top} \mathbf{v}_k^{(t)} - 2\alpha \mathbf{v}_k^{(t)\top} \mathbf{v}_k^{\text{Pa}(t)} + \alpha \mathbf{v}_k^{\text{Pa}(t)\top} \mathbf{v}_k^{\text{Pa}(t)} + C \quad (\text{B.12})$$

Now we take the derivative of O w.r.t. $\mathbf{v}_k^{(t)}$:

$$\frac{\partial O}{\partial \mathbf{v}_k^{(t)}} = 0 - 2\mathbf{R}_k^{(t)\top} \mathbf{u}_k^{(t)} + 2\mathbf{v}_k^{(t)} \mathbf{u}_k^{(t)\top} \mathbf{u}_k^{(t)} + 2\alpha \mathbf{v}_k^{(t)} - 2\alpha \mathbf{v}_k^{\text{Pa}(t)} + 0 + 0 \quad (\text{B.13})$$

$$0 = -\mathbf{R}_k^{(t)\top} \mathbf{u}_k^{(t)} + \left(\mathbf{u}_k^{(t)\top} \mathbf{u}_k^{(t)} + \alpha \right) \mathbf{v}_k^{(t)} - \alpha \mathbf{v}_k^{\text{Pa}(t)} \quad (\text{B.14})$$

$$\mathbf{v}_k^{(t)} = \frac{\mathbf{R}_k^{(t)\top} \mathbf{u}_k^{(t)} + \alpha \mathbf{v}_k^{\text{Pa}(t)}}{\left\| \mathbf{u}_k^{(t)} \right\|_2^2 + \alpha} \quad (\text{B.15})$$

With the non-negativity constraint $\mathbf{v}_k^{(t)} \geq 0$, we want $\mathbf{R}_k^{(t)\top} \mathbf{u}_k^{(t)} + \alpha \mathbf{v}_k^{\text{Pa}(t)} \geq 0$, because if $\mathbf{R}_k^{(t)\top} \mathbf{u}_k^{(t)} + \alpha \mathbf{v}_k^{\text{Pa}(t)} < 0$, O will increase in (B.11) and (B.12). So the finalized update rule is:

$$\mathbf{v}_k^{(t)} = \frac{\left[\mathbf{R}_k^{(t)\top} \mathbf{u}_k^{(t)} + \alpha \mathbf{v}_k^{\text{Pa}(t)} \right]_+}{\left\| \mathbf{u}_k^{(t)} \right\|_2^2 + \alpha} \quad (\text{B.16})$$

Optimize $u_k^{(t)}$

We can derive the update rule for $u_k^{(t)}$ in leaf node task t similarly but much more simply. From (B.12), we take the derivative of O_t with respect to $u_k^{(t)}$; all regularization terms will zero out since they do not involve $u_k^{(t)}$. Hence the final update rule for $u_k^{(t)}$ is:

$$u_k^{(t)} = \frac{\left[R_k^{(t)} v_k^{(t)} \right]_+}{\left\| v_k^{(t)} \right\|_2^2} \quad (\text{B.17})$$

Optimize $v_k^{(r)}$

For the overall consensus factor in the root of the task hierarchy, $v_k^{(r)}$, we can again ignore terms that do not involve $v_k^{(r)}$ in the objective (B.4). Note that we're going to collect the terms involving nodes c whose parent is the root node, i.e. $\text{Pa}(c) = r$:

$$O = \alpha \sum_{c \in \text{Child}(r)} \left\| v_k^{(c)} - v_k^{(r)} \right\|_2^2 + C \quad (\text{B.18})$$

$$= \alpha \sum_{c \in \text{Child}(r)} \left(v_k^{(c)} - v_k^{(r)} \right)^\top \left(v_k^{(c)} - v_k^{(r)} \right) + C \quad (\text{B.19})$$

$$= \alpha \sum_{c \in \text{Child}(r)} \left[v_k^{(c)\top} v_k^{(c)} - 2v_k^{(c)\top} v_k^{(r)} + v_k^{(r)\top} v_k^{(r)} \right] + C \quad (\text{B.20})$$

$$= C - \sum_{c \in \text{Child}(r)} 2\alpha v_k^{(c)\top} v_k^{(r)} + \sum_{c \in \text{Child}(r)} \alpha v_k^{(r)\top} v_k^{(r)} \quad (\text{B.21})$$

Now we take the derivative, set to 0, and solve:

$$\frac{\partial O}{\partial \mathbf{v}_k^{(r)}} = 0 - \sum_{c \in \text{Child}(r)} 2\alpha \mathbf{v}_k^{(c)} + \sum_{c \in \text{Child}(r)} 2\alpha \mathbf{v}_k^{(r)} \quad (\text{B.22})$$

$$0 = - \sum_{c \in \text{Child}(r)} \mathbf{v}_k^{(c)} + |\text{Child}(r)| \cdot \mathbf{v}_k^{(r)} \quad (\text{B.23})$$

$$\mathbf{v}_k^{(r)} = \frac{\sum_{c \in \text{Child}(r)} \mathbf{v}_k^{(c)}}{|\text{Child}(r)|} \quad (\text{B.24})$$

where $|\text{Child}(r)|$ is the number of direct child nodes of the root node r .

Optimize $\mathbf{v}_k^{(b)}$

For the latent feature factor in an internal/branch node of the task hierarchy, $\mathbf{v}_k^{(b)}$, same drill as before: we ignore terms that do not involve $\mathbf{v}_k^{(b)}$ for the particular node b of interest in the objective (B.4). This time we collect terms involving the parent node of b , i.e. $\text{Pa}(b)$, and nodes c whose parent is b , i.e. $\text{Pa}(c) = b$:

$$O = \alpha \left(\left\| \mathbf{v}_k^{(b)} - \mathbf{v}_k^{\text{Pa}(b)} \right\|_2^2 + \sum_{c \in \text{Child}(b)} \left\| \mathbf{v}_k^{(c)} - \mathbf{v}_k^{(b)} \right\|_2^2 \right) + C \quad (\text{B.25})$$

$$= \alpha \left(\mathbf{v}_k^{(b)} - \mathbf{v}_k^{\text{Pa}(b)} \right)^\top \left(\mathbf{v}_k^{(b)} - \mathbf{v}_k^{\text{Pa}(b)} \right) + \alpha \sum_{c \in \text{Child}(b)} \left(\mathbf{v}_k^{(c)} - \mathbf{v}_k^{(b)} \right)^\top \left(\mathbf{v}_k^{(c)} - \mathbf{v}_k^{(b)} \right) + C \quad (\text{B.26})$$

$$= \alpha \left[\mathbf{v}_k^{(b)\top} \mathbf{v}_k^{(b)} - 2\mathbf{v}_k^{(b)\top} \mathbf{v}_k^{\text{Pa}(b)} + \mathbf{v}_k^{\text{Pa}(b)\top} \mathbf{v}_k^{\text{Pa}(b)} \right] + \alpha \sum_{c \in \text{Child}(b)} \left[\mathbf{v}_k^{(c)\top} \mathbf{v}_k^{(c)} - 2\mathbf{v}_k^{(c)\top} \mathbf{v}_k^{(b)} + \mathbf{v}_k^{(b)\top} \mathbf{v}_k^{(b)} \right] + C \quad (\text{B.27})$$

$$= \alpha \mathbf{v}_k^{(b)\top} \mathbf{v}_k^{(b)} - 2\alpha \mathbf{v}_k^{(b)\top} \mathbf{v}_k^{\text{Pa}(b)} - \sum_{c \in \text{Child}(b)} 2\alpha \mathbf{v}_k^{(c)\top} \mathbf{v}_k^{(b)} + \sum_{c \in \text{Child}(b)} \alpha \mathbf{v}_k^{(b)\top} \mathbf{v}_k^{(b)} + C \quad (\text{B.28})$$

Now we take the derivative, set to 0, and solve:

$$\frac{\partial O}{\partial v_k^{(b)}} = 2\alpha v_k^{(b)} - 2\alpha v_k^{\text{Pa}(b)} - \sum_{c \in \text{Child}(b)} 2\alpha v_k^{(c)} + \sum_{c \in \text{Child}(b)} 2\alpha v_k^{(b)} \quad (\text{B.29})$$

$$0 = v_k^{(b)} - v_k^{\text{Pa}(b)} - \sum_{c \in \text{Child}(b)} v_k^{(c)} - |\text{Child}(b)| \cdot v_k^{(b)} \quad (\text{B.30})$$

$$= (1 + |\text{Child}(b)|)v_k^{(b)} - v_k^{\text{Pa}(b)} - \sum_{c \in \text{Child}(b)} v_k^{(c)} \quad (\text{B.31})$$

$$v_k^{(b)} = \frac{v_k^{\text{Pa}(b)} + \sum_{c \in \text{Child}(b)} v_k^{(c)}}{1 + |\text{Child}(b)|} \quad (\text{B.32})$$

where $|\text{Child}(b)|$ is the number of direct child nodes of b .