Lee and Roy *Genome Biology* (2021) 22:164
https://doi.org/10.1186/s13059-021-02378-z

Genome Biology

METHOD **Open Access**

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

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

*Correspondence:
sroy@biostat.wisc.edu
[1]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 53715 Madison, USA
[2]Wisconsin Institute for Discovery, 330 N. Orchard Street, 53715 Madison, USA

Abstract

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.

Keywords: Three-dimensional (3D) genome organization, High-throughput chromosomal conformation capture (Hi-C), Topologically associating domains (TADs), Matrix factorization

Background

The three-dimensional (3D) organization of the genome has emerged as an important layer of gene regulation in developmental processes, disease progression, and evolution [1–6]. High-throughput chromosome conformation capture (3C) assays such as Hi-C [7, 8], SPRITE [9], and GAM [6] 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 [10], and between normal and disease cellular states [4, 11]. 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 [12–14], these units have been shown to be conserved across species [5, 15] and also associated with developmental [16] and disease processes [11, 17–19]. 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, [20, 21], community and subgraph detection within networks [20, 22], Gaussian mixture modeling [23, 24], and signal processing approaches [25]. However, comparison of TAD-finding methods [26–28] 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 [29]. To address the sparsity of datasets, different smoothing-based approaches have been proposed [30–32]; 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 [33–35]. 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 [36]. 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.

Results

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 (Fig. 1; see the "Methods" section). 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, [37], SPRITE [9], and HiChIP [38] datasets. Previously, NMF has been used for bias correction and dimensionality reduction of Hi-C data [39]; 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, where the columns of the U and V matrices are favored to be smooth on a graph

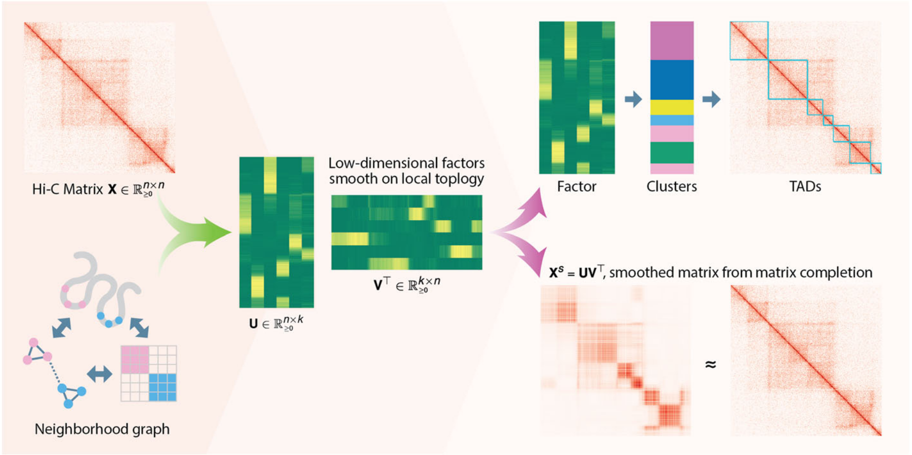

Hi-C Matrix $\mathbf{X} \in \mathbb{R}^{n\times n}_{\geq 0}$
Neighborhood graph
$\mathbf{U} \in \mathbb{R}^{n\times k}_{\geq 0}$
Low-dimensional factors smooth on local toplogy
$\mathbf{V}^\top \in \mathbb{R}^{k\times n}_{\geq 0}$
Factor
Clusters
TADs
$\mathbf{X}^S = \mathbf{U}\mathbf{V}^\top$, smoothed matrix from matrix completion
≈

Fig. 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

of genomic regions (Fig. 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 (Additional File 1, Figure S1). We determined that setting k to identify TADs of size ~1Mb, with a neighborhood size of $r = 250$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., $\lambda = 0$).

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 (see the "Methods" section) and applied them along with GRiNCH to Hi-C data of five different cell lines from Rao et al. [37] 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 (see the "Methods" section); 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 (see the "Methods" section). 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 (Fig. 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 [28]. 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 (Fig. 2B; Additional File 1, Figure S2) 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 (Additional File 1, Figure S2). 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 (see the "Methods" section).

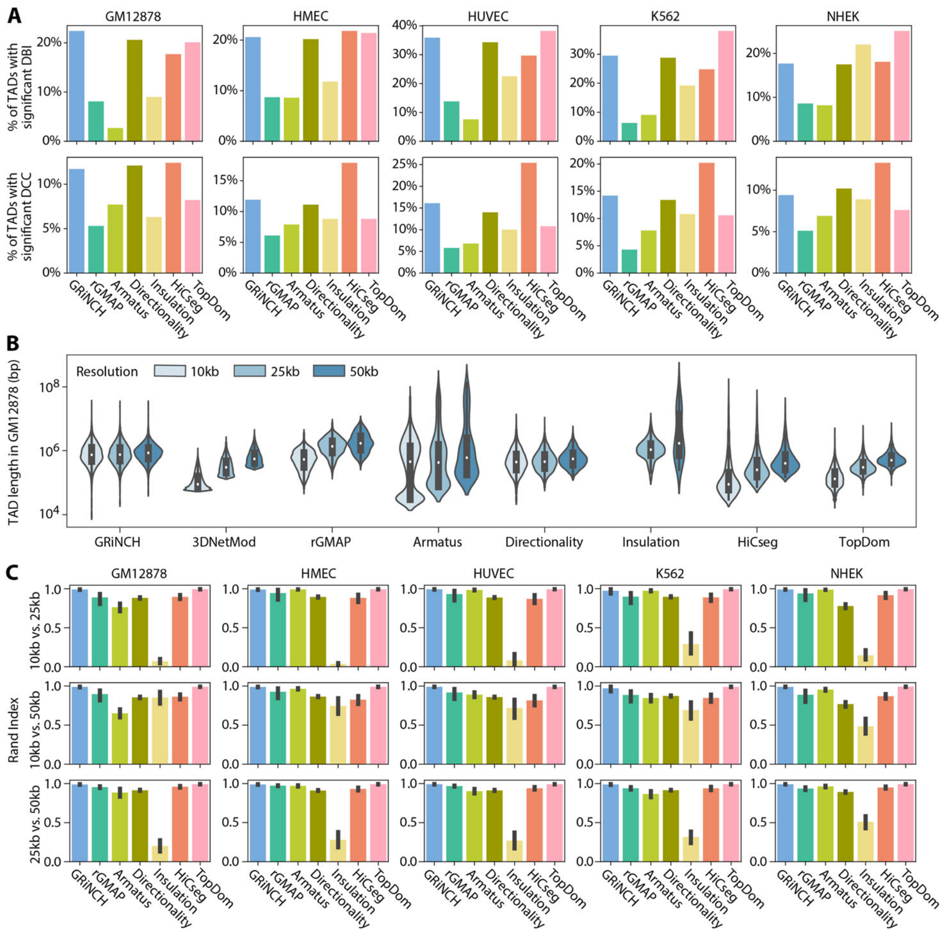
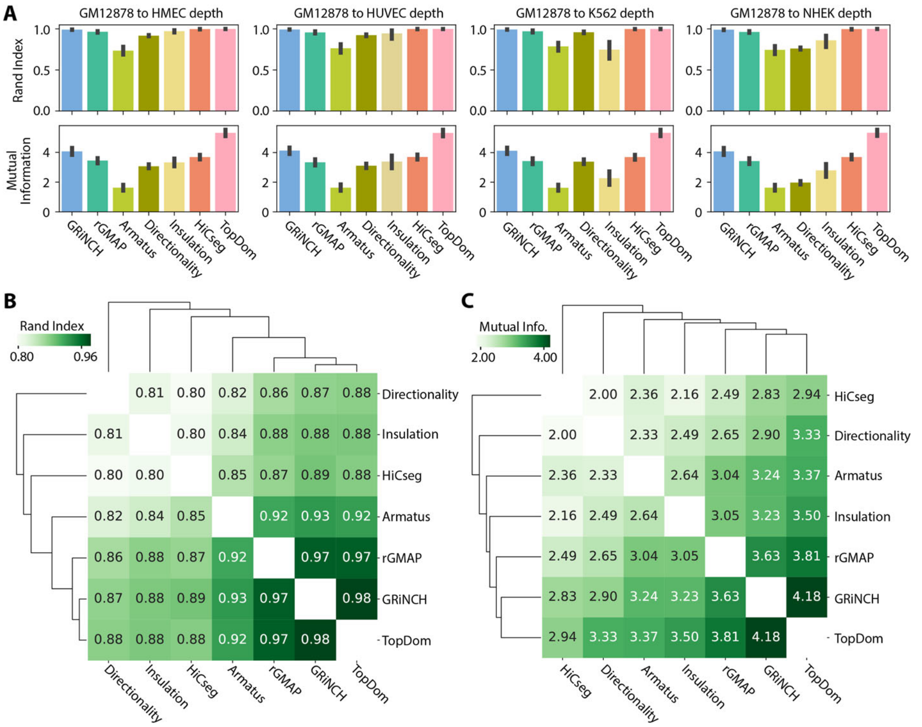


Fig. 2 Characterizing TADs with internal validation metrics, TAD size, and composition. **A** Percentage of TADs with significant Davies-Bouldin index (DBI) and Delta Contact Count (DCC) values. Shown are values for GRiNCH and six other methods. The higher the bar, the better a method. Note: 3DNetMod 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

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 information-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 of 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 (see the “Methods” section). 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 (Fig. 2C) and mutual information (Additional File 1, Figure S3). 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 [28]. 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 (Fig. 3A; see the “Methods” section), 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.



	Directionality	Insulation	HiCseg	Armatus	rGMAP	GRiNCH	TopDom
Directionality		0.81	0.80	0.82	0.86	0.87	0.88
Insulation	0.81		0.80	0.84	0.88	0.88	0.88
HiCseg	0.80	0.80		0.85	0.87	0.89	0.88
Armatus	0.82	0.84	0.85		0.92	0.93	0.92
rGMAP	0.86	0.88	0.87	0.92		0.97	0.97
GRiNCH	0.87	0.88	0.89	0.93	0.97		0.98
TopDom	0.88	0.88	0.88	0.92	0.97	0.98	

	HiCseg	Directionality	Armatus	Insulation	rGMAP	GRiNCH	TopDom
HiCseg		2.00	2.36	2.16	2.49	2.83	2.94
Directionality	2.00		2.33	2.49	2.65	2.90	3.33
Armatus	2.36	2.33		2.64	3.04	3.24	3.37
Insulation	2.16	2.49	2.64		3.05	3.23	3.50
rGMAP	2.49	2.65	3.04	3.05		3.63	3.81
GRiNCH	2.83	2.90	3.24	3.23	3.63		4.18
TopDom	2.94	3.33	3.37	3.50	3.81	4.18	

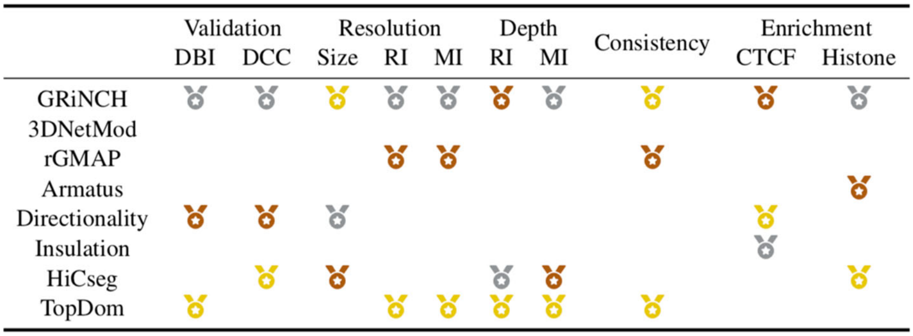
Fig. 3 Evaluating the stability of different TAD-calling methods to datasets of different depths. **A** The mean similarity, across chromosomes, between TADs from high-depth GM12878 dataset and TADs from low-depth GM12878 datasets obtained by downsampling the GM12878 dataset to different depths observed 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

A third hindrance in the interpretation of results from TAD finding methods is the disagreement on the TAD definitions [28, 29]. 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 (Fig. 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 (Additional File 1, Figure S4).

To summarize, our internal validation and stability analysis showed that the top performing methods depends upon the evaluation criteria. However, GRiNCH is among the top performing methods for all the criteria we examined (Fig. 4), producing TADs that are as good or better than existing methods and are stable to varying resolution and depth.

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 [29, 40]. We tested the TAD boundaries from each method for the enrichment of peaks of CTCF, RAD21, and SMC3 in the five Rao et al. cell lines with Hi-C data (Fig. 5A; see the "Methods" section). 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



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

Fig. 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/gray: 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 (Table S1a); DCC: percentage of TADs with significant Delta Contact Counts (Table S1b). *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 (Table S1c); RI, MI: similarity of TADs from high- and low-resolution data, measured by Rand index (RI, Table S1d) and mutual information (MI, Table S1e). *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, Table S1f) and mutual information (MI, Table S1g) *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 (Table S1h); Histone: proportion of TADs with significant mean histone signal (Table S1i). Supplementary Tables S1a-i are available in Additional File 3

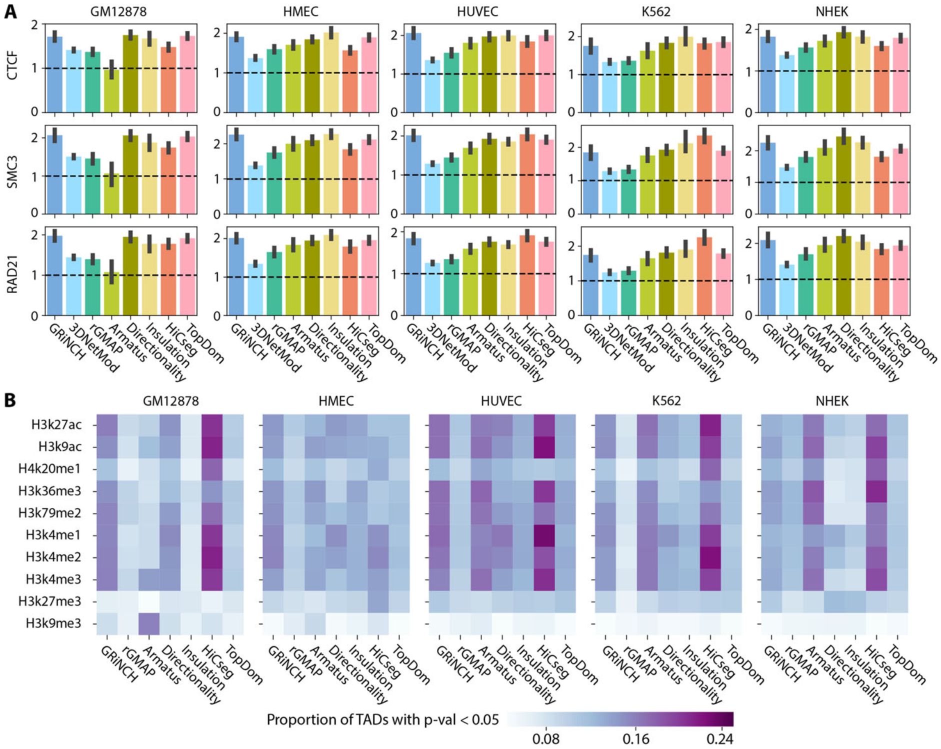


Fig. 5 Evaluating TAD-calling methods with enrichment of boundary elements and regulatory signals. **A** Fold enrichment of binding signals of architectural protein in TAD boundaries. Shown are the mean fold enrichment of CTCF ChIP-seq peaks and accessible motif instances of cohesin proteins, RAD21 and 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

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, and 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 [41], we next measured the proportion of TADs with significant levels of mean histone modification signals (Fig. 5B) compared to randomly shuffled TADs (see the "Methods" section). 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 (Additional File 1, Figure S5). 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 [40]. These enrichment patterns show that when considering existing methods, there is a tradeoff 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 (Fig. 4) suggesting that GRiNCH TADs capture a diverse type of TADs.

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 [30, 31], improve the agreement among biological replicates [30], and assist in identifying loops and differential interactions [42, 43]. 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 techniques such as mean filter [30] and Gaussian filter [43], which have been used for Hi-C data pre-processing [30, 42, 43]. We additionally compared against a supervised learning method, HiCNN [32], 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 (see the "Methods" section). 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.

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 (Fig. 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-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 (Fig. 6B). When comparing interactions less than 2Mb apart, the HiCNN model

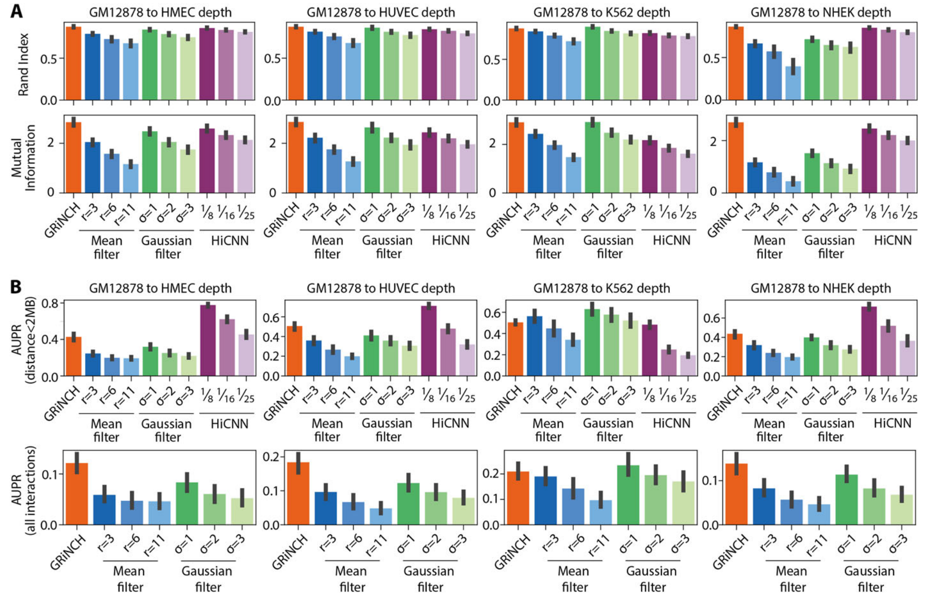


Fig. 6 Evaluating the benefits of smoothing in GRiNCH. Recovery of topology and significant interactions from downsampled then smoothed data. **A** Rand index and mutual information were used to measure the similarity between TADs from high-depth GM12878 dataset and TADs from downsampled datasets smoothed 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

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 25-kb 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 (see the "Methods" section). 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 (Additional File 1, Figure S6a,b). We next used Fit-Hi-C [44] to identify significant interactions (q-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 (Additional File 1, Figure S6c). 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.

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 [45] and (b) pluripotency reprogramming in mouse [46]. Bonev et al. [45] 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 (Additional File 1, Figure S7). 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. 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 (Fig. 7A). To identify novel genomic regions associated with changing 3D structure, we compared GRiNCH TADs across the time points (see the "Methods" section) 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 (Fig. 7B). Syap1-deficient mice have been shown to display motor and movement defects [47]; Ap1s2 has been associated with intellectual disability, basal ganglia disease, and seizures accompanying Pettigrew syndrome [48]. Another example of dynamic 3D organization identified by GRiNCH was near the Arl6ip1 and Foxp1 genes (Additional File 1, Figure S8). These genes are involved in glutamate neurotransmitter transport [49] and neural differentiation [50], respectively. Visual inspection of results from other top-performing TAD-calling methods in the corresponding regions (Additional File 1, Figure S9-11) 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.

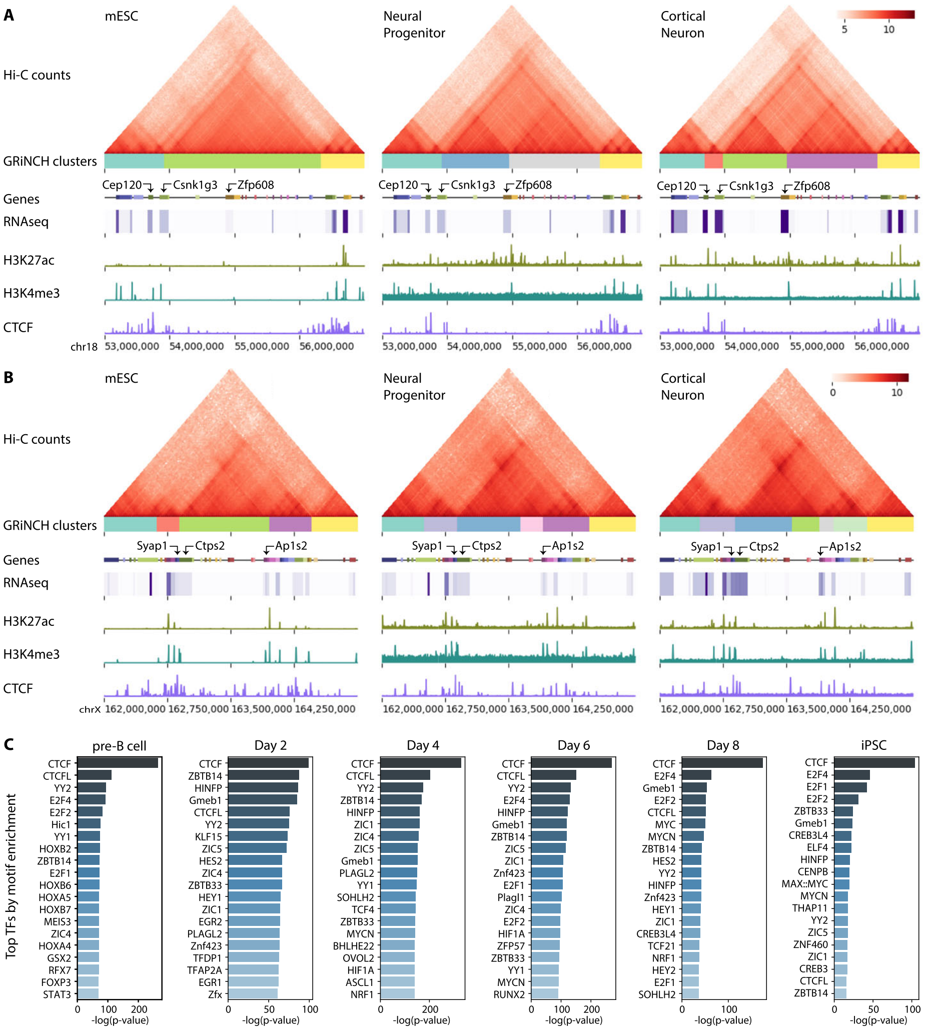


Fig. 7 GRiNCH applied to Hi-C datasets along developmental time courses. **A** Interaction profile near the Zfp608 gene in mouse embryonic stem cells (mESC), neural progenitors (NPC), and differentiated cortinal neurons (CN). Heatmaps are of Hi-C matrices after log2-transformation of interaction counts 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 cortinal 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

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 (days 2, 4, 6, and 8; see the "Methods" section). 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 (Additional File 1, Figure S12). We examined the interaction profile in the 1.3 Mb around the Sox2 gene, a known pluripotency gene (Additional File 1, Figure S13). 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.

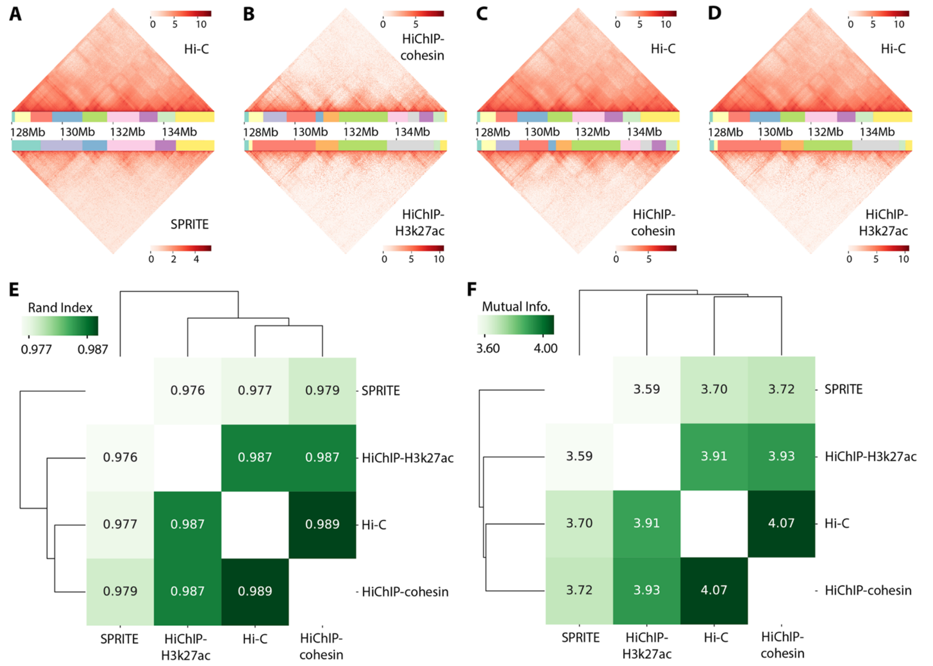
While architectural proteins such as CTCF and cohesin play important roles in establishing TAD boundaries, it is currently unclear if there are additional DNA binding proteins that could, independently or in concert with CTCF, contribute towards establishing these boundaries, especially in a cell type-specific manner. Previous work to identify such regulatory proteins has focused on a single time point [51] or stage [52]. 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 (see the "Methods" section). 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 [53]. We ranked the TFs based on their significant enrichment in each cell type (Fig. 7C, Additional File 3, Table S2). The top-ranking TF across the cell types was CTCF, which is consistent with its role as an architectural protein in establishing TADs (Fig. 7C). We also found other factors in the same zinc finger protein family as CTCF [54], such as ZBTB14, Plagl2/1, ZIC1/3/4/5, CTCFL, and 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 [55], are of interest as YY1 has been identified as an enforcer of long-range enhancer-promoter loops [56]. 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 [57]. FOXP3 is a master regulator of T cells [58], 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 [59] to bind to their targets, supporting the simultaneous enrichment of these factors.

We repeated this analysis for the Rao et al. cell lines (Additional File 3, Table S3). 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.

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 [6]. 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) [9] and HiChIP [38]. 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 [38]. 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[9], cohesin HiChIP [38], and H3k27ac HiChIP [60]. A visual comparison between these datasets for an 8Mb region of chr8 shows regions of good concordance between datasets (Fig. 8A–D). We quantified the global similarity of GRiNCH TADs from the four different datasets, for all chromosomes with Rand index and mutual information (Fig. 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 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 applied to different experimental platforms for measuring genome-wide chromosome conformation.



	SPRITE	HiChIP-H3k27ac	Hi-C	HiChIP-cohesin
SPRITE		0.976	0.977	0.979
HiChIP-H3k27ac	0.976		0.987	0.987
Hi-C	0.977	0.987		0.989
HiChIP-cohesin	0.979	0.987	0.989	

	SPRITE	HiChIP-H3k27ac	Hi-C	HiChIP-cohesin
SPRITE		3.59	3.70	3.72
HiChIP-H3k27ac	3.59		3.91	3.93
Hi-C	3.70	3.91		4.07
HiChIP-cohesin	3.72	3.93	4.07	

Fig. 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

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 [31]). 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 [61]. 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 [62]. Among the novel factors that were present in both the cell lines and the mouse reprogramming dataset were several zinc finger proteins, e.g., PLAGL1, ZIC1, ZIC4/5, and 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 [63] and zebra fish [64], 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 [65], FOXP3 [66], and STAT3 [57] 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 is specifically involved in the binding of HOX TFs to target genes in the brain [59]. 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 [29, 67].

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 [68, 69]. 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.

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.

Methods

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) [36] that is applicable to high-dimensional chromosome conformation capture data such as Hi-C (Fig. 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 $k << n, m$ 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 $X = \left[x_{ij}\right] \in \mathbb{R}^{(n\times n)}$ 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: $\left\|X - UV^\top\right\|_F^2$, s.t.$U \geq 0, V \geq 0$ [33], 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:

$$u_{ik} \leftarrow u_{ik}\frac{(XV)_{ik}}{\left(UV^\top V\right)_{ik}}, \; v_{jk} \leftarrow v_{jk}\frac{\left(X^\top U\right)_{jk}}{\left(VU^\top U\right)_{jk}} \tag{1}$$

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 [36]. 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:

$$W_{ij} = \begin{cases} 1, \text{ if } x_i \in N_r(x_j) \text{ and } x_j \in N_r(x_i) \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

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

Graph regularized NMF has the following objective:

$$\|X - UV^\top\|_F^2 + \lambda \text{Tr}\left(V^\top LV\right) + \lambda \text{Tr}\left(U^\top LU\right), \tag{3}$$

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 terms and measure 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 (1) gains regularization terms [36]:

$$u_{ik} \leftarrow u_{ik} \frac{(\mathrm{XV} + \lambda \mathrm{WU})_{ik}}{\left(\mathrm{UV}^{\top}\mathrm{V} + \lambda \mathrm{DU}\right)_{ik}},\ v_{jk} \leftarrow v_{jk} \frac{\left(\mathrm{X}^{\top}\mathrm{U} + \lambda \mathrm{WV}\right)_{jk}}{\left(\mathrm{VU}^{\top}\mathrm{U} + \lambda \mathrm{DV}\right)_{jk}} \tag{4}$$

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 $c_i = \mathrm{argmax}_{j \in \{1,\ldots,k\}} u_{ij}$, 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 (Additional File 2, Algorithm 1), adopted from spatially connected k-medoids clustering [70], 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 (Additional File 1, Figure S14a) 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 (Additional File 1, Figure S14b). SubTAD-scale clusters tend to be more stable to depth and sparsity (Additional File 1, Figure S14c) and are also more enriched in boundary elements like CTCF (Additional File 1, Figure S15a). As a tradeoff, higher proportion of metaTAD-scale clusters (i.e., expected cluster size = 2Mb) are enriched in histone modification marks (Additional File 1, Figure S15b). 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,1M} were considered, where $r = 100K$ in a Hi-C dataset of 25-Kb 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 (Additional File 1, Figure S1a). With regularization, a neighborhood radius of 100Kb or larger yields higher CTCF enrichment (Figure S1B). We also note that the regularization parameters do not discernibly change the TAD size distribution (Additional File 1, Figure S16). Based on these results, the default regularization parameters for GRiNCH are set at $\lambda = 1$ and r =250kb.

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 $O(kn^2)$ [36]. 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 $\in$ { 25kb, 50kb, 100kb, 250kb, 500kb, 1Mb}). 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 Figure S17 (Additional File 1). 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 [71]. 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 [72]. 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 [73]. NNDSVD initialization with randomized SVD results in lower loss, i.e., factors that can better approximate the original Hi-C matrix, in fewer iterations (Additional File 1, Figure S18a,b), and more stable results than direct random initialization (Additional File 1, Figure S18c,d).

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 10-kb, 25-kb, and 50-kb resolution [37] (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 25-kb 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) [9] and HiChIP [38]. 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) [38] and H3k27ac HiChIP (GEO accession: GSE101498) [60].

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) [45] and (b) Hi-C data from reprogramming pre-B cells to induced pluripotent state [46] (GEO accession: GSE96553). For (a) neural differentiation dataset, Juicer Straw tool [74] 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, Bα cells, day 2, day 4, day 6, and 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 [75] 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. [37], 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) were obtained from [76]. To create a union of DNase accessible regions from replicates within a cell line, BEDtools [77] 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 [78, 79]: ENCFF856MFN, ENCFF235KUD, ENCFF491BOT, ENCFF946QPV, ENCFF968KGT, ENCFF541JWD, ENCFF978UNU, ENCFF297CKS, and ENCFF569UYX.

We obtained ChIP-seq datasets for histone modification marks from the ENCODE consortium [78, 79]. To generate genome-wide histone modification levels for each mark, fastq reads were aligned to the human genome (hg19) with bowtie2 [80] and aggregated into a base-pair signal coverage profile using SAMtools [81] and BEDtools [77]. The base-pair signal coverage was averaged within each 25-kb 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 [53]. Next, we obtained their accessible motif match sites in hg19 for the five cell lines from [37] using the same process that was used for SMC3 and RAD21 motifs. To identify the accessible motif sites for mouse cells during pluripotency reprogramming [46], we aligned ATAC-seq fastq reads to the mouse genome (mm10) with bowtie2 [80] and deduplicated with SAMtools [81]. Accessible peaks were called with MACS2 [82]. 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.

TAD-calling methods

GRiNCH was benchmarked against 7 other TAD-calling methods: directionality index method [23], Armatus [20], insulation score method [25], rGMAP [24], 3DNetMod [22], HiCseg [83], and TopDom [84]. For all methods, default or recommended parameter values were used when available. Execution scripts containing the parameter values used for these methods are available to download ("Availability of data and materials" section).

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 [85], 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 [85] 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.

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. [37]. 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: $\frac{n_{\text{TAD-BIND}}/n_{\mathrm{TAD}}}{n_{\mathrm{BIND}}/N}$. 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 (Additional File 3, Table S1h).

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. [28]. 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 Fig. 5B. The mean proportion of TADs with significant enrichment across chromosomes and cell lines was used to rank the TAD-calling methods (Additional File 3, Table S1i).

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 $i, j \in \{1, \ldots, k\}, i \neq j$: $\mathrm{DBI}_i = \max_{i \neq j} S_{ij}$. The similarity metric, S_{ij}, between C_i and C_j is defined as:

$$S_{ij} = \frac{d_i + d_j}{\mathrm{distance}_{ij}} \tag{5}$$

where d_i is the average distance between each data point in cluster C_i and the cluster centroid and $\mathrm{distance}_{ij}$ 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 (Additional File 3, Table S1a).

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 $\text{DCC}_i = \text{in}_i - \text{out}_i$.

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 (Additional File 3, Table S1b).

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, 10-kb, 25-kb, and 50-kb bins were split into a size of lowest common denominator, i.e., 5kb. Then all 5-kb bins were assigned to the same cluster as in the original lower-resolution bin (e.g., a 10-kb bin assigned to cluster i would yield two 5-kb bins assigned to cluster i). For these comparisons, we computed these metrics at the 5-kb 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

$$\text{MI}(A;B) = \sum_{a \in A} \sum_{b \in B} p_{(A,B)}(a,b) \log \left(\frac{p_{(A,B)}(a,b)}{p_A(a) p_B(b)} \right) \tag{6}$$

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 $p_{(A,B)}(a,b)$ 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 (Additional File 3, Table S1d-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) [37] were compared to the TADs from a downsampled, low-depth dataset. If 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: $\widetilde{x}_{ij} = \frac{x_{ij}^h}{\mu_d^h/\mu_d^l}$. 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.

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) [45] 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 clusters 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 Fig. 7b and Additional File 1, Figure S8 are from this subset.

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 [53] are enriched in GRiNCH TAD boundaries. This procedure was applied to the five cell lines from Rao et al. [37] and the time points from the mouse reprogramming timecourse data [46]. 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.

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 $\mathrm{X}^s = \mathrm{U}\mathrm{V}^\mathsf{T}$ 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 [86] and Hi-C data pre-processing [30, 42, 43], as well as HiCNN [32], a method based on convolutional neural network to impute interaction counts.

Mean filter

Mean filtering is used in HiCRep [30] 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^s_{ij} is estimated from the mean of its neighboring elements within radius r: $x^s_{ij} = \frac{1}{(2r+1)^2}\sum_{a=i-r}^{i+r}\sum_{b=j-r}^{j+r} x_{ab}$. 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 [42, 43]. 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:

$$x^s_{ij} = \frac{1}{2\pi\sigma^2}\sum_{a=i-n}^{i+n}\sum_{b=j-n}^{j+n} e^{-\frac{(i-a)^2+(j-b)^2}{2\sigma^2}} x_{ab} \tag{7}$$

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 [32] 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 the "Assessment of benefits from smoothing" section below). Measuring TAD recovery after smoothing was not affected since the directionality index method uses a 2-Mb-sized window of interactions (see the "Directionality index" section above).

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 the "Robustness to low-depth data" section) using methods described above (see the "Smoothing methods" section). 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 the "TAD similarity and stability metrics" section"). Higher similarity metric values imply better recovery of structure from smoothing.

Recovery of significant interactions

Fit-Hi-C [44] was used to call significant interactions in the original and the smoothed Hi-C datasets, using a q-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).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02378-z.

Additional file 1: Supplementary_figures.pdf. PDF file containing Figures S1-S18.

Additional file 2: Algorithm.pdf. A PDF file containing Algorithm 1, the pseudocode for chain-constrained *k*-medoids clustering.

Additional file 3: Supplementary_tables.xlsx. An Excel file containing Tables S1-3. Each spreadsheet tab includes table name and legend in the top rows.

Additional file 4: Review history

Acknowledgements

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.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional File 4.

Authors' contributions

Lee and Roy conceptualized the overall framework and algorithm. Lee implemented the algorithm, designed and performed experiments, and wrote the manuscript. Roy designed the experiments and wrote the manuscript. The authors read and approved the final manuscript.

Funding

This work is supported by the National Institutes of Health (NIH) through the grant NHGRI R01-HG010045-01.

Availability of data and materials

We applied GRiNCH to published Hi-C datasets from [37] (GEO accession: GSE63525), [45] (GEO accession: GSE96107), and [46] (GEO accession: GSE96553); SPRITE data from [9] (GEO accession: GSE114242); and HiChIP data from [38] (GEO accession: GSE80820) and [60] (GEO accession: GSE101498). In our analysis, we used CTCF ChIP-seq narrow-peak datasets available as ENCODE Uniform TFBS composite track [75], downloaded from the UCSC genome browser (wgEncodeEH000029, wgEncodeEH000075, wgEncodeEH000054, wgEncodeEH000042, wgEncodeEH000063). DNase accessibility sites and histone modification ChIP-seq datasets were obtained from the ENCODE consortium [78, 79]: ENCFF856MFN, ENCFF235KUD, ENCFF491BOT, ENCFF946QPV, ENCFF968KGT, ENCFF541JWD, ENCFF978UNU, ENCFF297CKS, ENCFF569UYX. A detailed description of how these datasets were processed can be found in the "Datasets used in experiments and analysis" section.

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 [87] 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 [87], 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

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 3 September 2020 Accepted: 10 May 2021

References

1. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;17(11):661–78. https://doi.org/10.1038/nrg.2016.112. Accessed 16 Jan 2019.
2. Hug CB, Vaquerizas JM. The birth of the 3D genome during early embryonic development. Trends Genet. 2018;34(12):903–14. https://doi.org/10.1016/j.tig.2018.09.002. Accessed 05 Oct 2018.
3. Rowley J, Nichols M, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, Wang P, Ruan Y, Corces V. Evolutionarily conserved principles predict 3D chromatin organization. Mol Cell. 2017;67(5):837–52. https://doi.org/10.1016/j.molcel.2017.07.022.
4. Krijger PHL, de Laat W. Regulation of disease-associated gene expression in the 3D genome. Nat Rev Mol Cell Biol. 2016;17(12):771–82. https://doi.org/10.1038/nrm.2016.138. Accessed 14 Jan 2019.
5. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019;5(4):1668. https://doi.org/10.1126/sciadv.aaw1668. Accessed 31 Dec 2019.
6. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. Nat Rev Genet. 2019;21(4):207–26. https://doi.org/10.1038/s41576-019-0195-2. Accessed 30 Dec 2019.
7. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93. https://doi.org/10.1126/science.1181369.
8. Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat Rev Genet. 2018;19(12):789–800. https://doi.org/10.1038/s41576-018-0060-8. Accessed 29 Oct 2018.
9. Quinodoz S, Ollikainen N, Tabak B, Palla A, Schmidt J, Detmar E, Lai M, Shishkin A, Bhat P, Takei Y, Trinh V, Aznauryan E, Russell P, Cheng C, Jovanovic M, Chow A, Cai L, McDonel P, Garber M, Guttman M. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. Cell. 2018;174(3):744–57. https://doi.org/10.1016/j.cell.2018.05.024.

10. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. Nat Rev Mol Cell Biol. 2019;20(9):535–50. https://doi.org/10.1038/s41580-019-0132-4. Accessed 19 May 2020.
11. Chakraborty A, Ay F. The role of 3D genome organization in disease: from compartments to single nucleotides. Semin Cell Dev Biol. 2018;90:104–13. https://doi.org/10.1016/j.semcdb.2018.07.005. Accessed 29 Jan 2019.
12. Kim S, Yu N-K, Kaang B-K. CTCF as a multifunctional protein in genome regulation and gene expression. Exp Mol Med. 2015;47(6):166. https://doi.org/10.1038/emm.2015.33. Accessed 08 Aug 2020.
13. Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korbel JO, Furlong EEM. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. Nat Genet. 2019;51(8):1272–82. https://doi.org/10.1038/s41588-019-0462-3. Accessed 08 Aug 2020.
14. van Steensel B, Furlong EEM. The role of transcription in shaping the spatial organization of the genome. Nat Rev Mol Cell Biol. 2019;20(6):327–37. https://doi.org/10.1038/s41580-019-0114-6. Accessed 08 Aug 2020.
15. Eres IE, Luo K, Hsiao CJ, Blake LE, Gilad Y. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. PLoS Genet. 2019;15(7):e1008278. https://doi.org/10.1371/journal.pgen.1008278.
16. Stadhouders R, Filion GJ, Graf T. Transcription factors and 3D genome conformation in cell-fate decisions. Nature. 2019;569(7756):345–54. https://doi.org/10.1038/s41586-019-1182-7. Accessed 08 Aug 2020.
17. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, Suvà ML, Bernstein BE. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature. 2016;529(7584):110–4.
18. Kleinjan DA, Lettice LA. Long-range gene control and genetic disease. Adv Genet. 2008;61:339–88.
19. Valton A-L, Dekker J. TAD disruption as oncogenic driver. Curr Opin Genet Dev. 2016;36:34–40. https://doi.org/10.1016/j.gde.2016.03.008. Accessed 08 Aug 2020.
20. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithm Mol Biol. 2014;9(1):14. https://doi.org/10.1186/1748-7188-9-14.
21. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics. 2015;485:1601–9. https://doi.org/10.1093/bioinformatics/btv485.
22. Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, Bassett DS, Phillips-Cremins JE. Detecting hierarchical genome folding with network modularity. Nat Methods. 2018;15(2):119–22. https://doi.org/10.1038/nmeth.4560.
23. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80. https://doi.org/10.1038/nature11082.
24. Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian mixture model and proportion test. Nat Commun. 2017;8(1):535. https://doi.org/10.1038/s41467-017-00478-8.
25. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015;523(7559):240–4. https://doi.org/10.1038/nature14450.
26. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. Nat Methods. 2017;14(7):679–85. https://doi.org/10.1038/nmeth.4325.
27. Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools,. Nucleic Acids Res. 2017;45(6):2994–3005.
28. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 2018;19(1):217. https://doi.org/10.1186/s13059-018-1596-9. Accessed 11 Dec 2018.
29. de Wit E. TADs as the caller calls them. J Mol Biol. 2019;0022283619305923:638–42. https://doi.org/10.1016/j.jmb.2019.09.026. Accessed 30 Dec 2019.
30. Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, Yue F, Li Q. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. 2017;27(11):1939–49. https://doi.org/10.1101/gr.220640.117. Accessed 29 Aug 2018.
31. Ursu O, Boley N, Taranova M, Wang YXR, Yardimci GG, Stafford Noble W, Kundaje A. GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. Bioinformatics. 2018;34(16):2701–7. https://doi.org/10.1093/bioinformatics/bty164. Accessed 23 Jan 2019.
32. Liu T, Wang Z. HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data. Bioinformatics. 2019;35(21):4222–8. https://doi.org/10.1093/bioinformatics/btz251. Accessed 05 Feb 2021.
33. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: NIPS; 2000. p. 556–62. https://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf.
34. Wu Y, Tamayo P, Zhang K. Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. Cell Syst. 2018;7(6):656–66. https://doi.org/10.1016/j.cels.2018.10.015. Accessed 26 Dec 2018.
35. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y, Ngom A, Ochs MF, Xu Y, Fertig EJ. Enter the matrix: factorization uncovers knowledge from omics. Trends Genet. 2018;34(10):790–805. https://doi.org/10.1016/j.tig.2018.07.003. Accessed 22 Aug 2018.
36. Cai D, He X, Han J, Huang TS. Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell. 2011;33(8):1548–60. https://doi.org/10.1109/tpami.2010.231.
37. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D Map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80. https://doi.org/10.1016/j.cell.2014.11.021.
38. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods. 2016;13(11):919–22. https://doi.org/10.1038/nmeth.3999. Accessed 28 Nov 2018.
39. Hu X, Shi CH, Yip KY. A novel method for discovering local spatial clusters of genomic regions with functional relationships from DNA contact maps. Bioinformatics. 2016;32(12):111–20. https://doi.org/10.1093/bioinformatics/btw256. Accessed 14 Feb 2021.

40. Chang L-H, Ghosh S, Noordermeer D. TADs and their borders: free movement or building a wall? J Mol Biol. 2020;432(3):643–52. https://doi.org/10.1016/j.jmb.2019.11.025. Accessed 31 Dec 2019.
41. Andrey G, Schöpflin R, Jerković I, Heinrich V, Ibrahim DM, Paliou C, Hochradel M, Timmermann B, Haas S, Vingron M, Mundlos S. Characterization of hundreds of regulatory landscapes in developing limbs reveals two regimes of chromatin folding. Genome Res. 2017;27(2):223–33. https://doi.org/10.1101/gr.213066.116. Accessed 01 June 2020.
42. Rowley MJ, Poulet A, Nichols MH, Bixler BJ, Sanborn AL, Brouhard EA, Hermetz K, Linsenbaum H, Csankovszki G, Aiden EL, Corces VG. Analysis of Hi-C data using SIP effectively identifies loops in organisms from C. elegans to mammals. Genome Res. 2020;30(3):447–58. https://doi.org/10.1101/gr.257832.119. Accessed 05 Feb 2021.
43. Ardakany AR, Ay F, Lonardi S. Selfish: discovery of differential chromatin interactions via a self-similarity measure. Bioinformatics. 2019;35(14):145–53. https://doi.org/10.1093/bioinformatics/btz362. Accessed 05 Feb 2021.
44. Ay F, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 2014;24(6):999–1011. https://doi.org/10.1101/gr.160374.113. Accessed 14 Jan 2019.
45. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A, Cavalli G. Multiscale 3D genome rewiring during mouse neural development. Cell. 2017;171(3):557–72. https://doi.org/10.1016/j.cell.2017.09.043. Accessed 30 Aug 2019.
46. Stadhouders R, Vidal E, Serra F, Di Stefano B, Le Dily F, Quilez J, Gomez A, Collombet S, Berenguer C, Cuartero Y, Hecht J, Filion G, Beato M, Marti-Renom M, Graf T. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. Nat Genet. 2018;50(2):238–49. https://doi.org/10.1038/s41588-017-0030-7.
47. R von Collenberg C, Schmitt D, Rülicke T, Sendtner M, Blum R, Buchner E. An essential role of the mouse synapse-associated protein Syap1 in circuits for spontaneous motor activity and rotarod balance. Biol Open. 2019;8(6):bio042366. https://doi.org/10.1242/bio.042366.
48. Cacciagli P, Desvignes J-P, Girard N, Delepine M, Zelenika D, Lathrop M, Lévy N, Ledbetter DH, Dobyns WB, Villard L. AP1S2 is mutated in X-linked Dandy-Walker malformation with intellectual disability, basal ganglia disease and seizures (Pettigrew syndrome). Eur J Hum Genet. 2014;22(3):363–8. https://doi.org/10.1038/ejhg.2013.135. Accessed 06 Feb 2021.
49. Akiduki S, Ikemoto MJ. Modulation of the neural glutamate transporter EAAC1 by the addicsin-interacting protein ARL6IP1. J Biol Chem. 2008;283(46):31323–32. https://doi.org/10.1074/jbc.M801570200.
50. Braccioli L, Vervoort SJ, Adolfs Y, Heijnen CJ, Basak O, Pasterkamp RJ, Nijboer CH, Coffer PJ. FOXP1 promotes embryonic neural stem cell differentiation by repressing Jagged1 expression. Stem Cell Rep. 2017;9(5):1530–45. https://doi.org/10.1016/j.stemcr.2017.10.012. Accessed 06 Feb 2021.
51. Hong S, Kim D. Computational characterization of chromatin domain boundary-associated genomic elements. Nucleic Acids Res. 2017;45(18):10403–14. https://doi.org/10.1093/nar/gkx738. Accessed 14 Feb 2021.
52. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9:189. https://doi.org/10.1038/s41467-017-02525-w. Accessed 14 Feb 2021.
53. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48(D1):87–92. https://doi.org/10.1093/nar/gkz1001. Accessed 03 July 2020.
54. Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, Melino G, Raschellà G. Zinc-finger proteins in health and disease. Cell Death Discov. 2017;3(1):1–12. https://doi.org/10.1038/cddiscovery.2017.71. Accessed 07 July 2020.
55. Wu X, Shi T, He Y, Wang F, Sang R, Ding J, Zhang W, Shu X, Shen H, Yi J, Gao X, Liu W. Methylation of transcription factor YY2 regulates its transcriptional activity and cell proliferation. Cell Discov. 2017;3(1):1–22. https://doi.org/10.1038/celldisc.2017.35. Accessed 07 July 2020.
56. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, Guo YE, Hnisz D, Jaenisch R, Bradner JE, Gray NS, Young RA. YY1 is a structural regulator of enhancer-promoter loops. Cell. 2017;171(7):1573–88. https://doi.org/10.1016/j.cell.2017.11.008. Accessed 07 July 2020.
57. Chou W-C, Levy DE, Lee C-K. STAT3 positively regulates an early step in B-cell development. Blood. 2006;108(9): 3005–11. https://doi.org/10.1182/blood-2006-05-024430. Accessed 10 Aug 2020.
58. Lu L, Barbi J, Pan F. The regulation of immune tolerance by FOXP3. Nat Rev Immunol. 2017;17(11):703–17. https://doi.org/10.1038/nri.2017.75. Accessed 10 Aug 2020.
59. Uribe RA, Bronner ME. Meis3 is required for neural crest invasion of the gut during zebrafish enteric nervous system development. Mol Biol Cell. 2015;26(21):3728–40. https://doi.org/10.1091/mbc.E15-02-0112. Accessed 10 Aug 2020.
60. Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR, Wei Y, Nguyen T, Greenside PG, Corces MR, Tycko J, Simeonov DR, Suliman N, Li R, Xu J, Flynn RA, Kundaje A, Khavari PA, Marson A, Corn JE, Quertermous T, Greenleaf WJ, Chang HY. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet. 2017;49(11):1602–12. https://doi.org/10.1038/ng.3963. Accessed 18 May 2020.
61. 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, Brummelkamp TR, de Wit E, Rowland BD. The cohesin release factor WAPL restricts chromatin loop extension. Cell. 2017;169(4):693–707. https://doi.org/10.1016/j.cell.2017.04.013. Accessed 06 July 2020.
62. Beagan J, Duong M, Titus K, Zhou L, Cao Z, Ma J, Lachanski C, Gillis D, Phillips-Cremins J. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. 2017;27(7): 1139–52. https://doi.org/10.1101/gr.215160.116.
63. Knott SRV, Peace JM, Ostrow AZ, Gan Y, Rex AE, Viggiani CJ, Tavaré S, Aparicio OM. Forkhead transcription factors establish origin timing and long-range clustering in S. cerevisiae. Cell. 2012;148(1):99–111. https://doi.org/10.1016/j.cell.2011.12.012. Accessed 08 Aug 2020.

64. Yan J, Xu L, Crawford G, Wang Z, Burgess SM. The forkhead transcription factor FoxI1 remains bound to condensed mitotic chromosomes and stably remodels chromatin structure. Mol Cell Biol. 2006;26(1):155–68. https://doi.org/10.1128/MCB.26.1.155-168.2006. Accessed 08 Aug 2020.
65. Alharbi RA, Pettengell R, Pandha HS, Morgan R. The role of HOX genes in normal hematopoiesis and acute leukemia. Leukemia. 2013;27(5):1000–8. https://doi.org/10.1038/leu.2012.356.
66. Li Z, Li D, Tsun A, Li B. FOXP3 + regulatory T cells and their functional regulation. Cell Mol Immunol. 2015;12(5): 558–65. https://doi.org/10.1038/cmi.2015.10. Accessed 13 Aug 2020.
67. Cubeñas-Potts C, Corces VG. Architectural proteins, transcription, and the three-dimensional organization of the genome. FEBS letters. 2015;589:2923–0. https://doi.org/10.1016/j.febslet.2015.05.025. Accessed 08 Aug 2020.
68. Fotuhi Siahpirani A, Ay F, Roy S. A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. Genome Biol. 2016;17(1):114. https://doi.org/10.1186/s13059-016-0962-8. Accessed 08 Aug 2020.
69. Yang Y, Zhang Y, Ren B, Dixon JR, Ma J. Comparing 3D genome organization in multiple species using Phylo-HMRF. Cell Syst. 2019;8(6):494–505. https://doi.org/10.1016/j.cels.2019.05.011.
70. Soor S, Challa A, Danda S, Sagar BSD, Najman L. Extending k-means to preserve spatial connectivity. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium; 2018. p. 6959–62. https://doi.org/10.1109/IGARSS.2018.8518643.
71. Belford M, Mac Namee B, Greene D. Stability of topic modeling via matrix factorization. Expert Syst Appl. 2018;91: 159–69. https://doi.org/10.1016/j.eswa.2017.08.047.
72. Boutsidis C, Gallopoulos E. SVD based initialization: a head start for nonnegative matrix factorization. Pattern Recogn. 2008;41(4):1350–62. https://doi.org/10.1016/j.patcog.2007.09.010.
73. Voronin S, Martinsson P-G. RSVDPACK: an implementation of randomized algorithms for computing the singular value, interpolative, and CUR decompositions of matrices on multi-core and GPU architectures. arXiv preprint arXiv:1502.05366. 2016.
74. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3(1):95–8. https://doi.org/10.1016/j.cels.2016.07.002. Accessed 04 Aug 2020.
75. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, Lee BT, Barber GP, Harte RA, Diekhans M, Long JC, Wilder SP, Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. 2013;41:56–63. https://doi.org/10.1093/nar/gks1172. Accessed 13 Jan 2019.
76. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2014;42(5):2976–87. https://doi.org/10.1093/nar/gkt1249. Accessed 03 July 2020.
77. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6): 841–2. https://doi.org/10.1093/bioinformatics/btq033.
78. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome,. Nature. 2012;489(7414):57–74. https://doi.org/10.1038/nature11247.
79. Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, Rowe LD, Dreszer TR, Roe G, Podduturi NR, Tanaka F, Hong EL, Cherry JM. ENCODE data at the ENCODE portal,. Nucleic Acids Res. 2016;44(D1):1160. https://doi.org/10.1093/nar/gkv1160.
80. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923. Accessed 13 Jan 2019.
81. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
82. Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, Nussbaum C, Myers R, Brown M, Li W, Liu S. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):137. https://doi.org/10.1186/gb-2008-9-9-r137.
83. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics (Oxford, England). 2014;30(17):386–92. https://doi.org/10.1093/bioinformatics/btu443.
84. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2016;44(7):70. https://doi.org/10.1093/nar/gkv1505 Accessed 25 March 2019.
85. Kruse K, Hug CB, Hernández-Rodríguez B, Vaquerizas JM. TADtool: visual parameter identification for TAD-calling algorithms. Bioinformatics. 2016;32(20):3190–2. https://doi.org/10.1093/bioinformatics/btw368.
86. Davies ER. Machine Vision: Theory, Algorithms, Practicalities: Elsevier; 2004.
87. Lee D-I, Roy S. GRiNCH: Graph-Regularized NMF and Clustering for Hi-C. Github. 2021. https://roy-lab.github.io/grinch. https://doi.org/10.5281/zenodo.4540608.

Publisher's Note