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

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

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53715, USA

[2]Wisconsin Institute for Discovery, 330 N. Orchard Street, Madison, WI 53715, USA

[*]To whom correspondence should be addressed.

Abstract

Three-dimensional (3D) genome organization, which determines how the DNA is packaged inside the nucleus, has emerged as a key component of the gene regulation machinery. High-throughput chromosome conformation datasets, such as Hi-C, have become available across multiple conditions and timepoints, offering a unique opportunity to examine changes in 3D genome organization and link them to phenotypic changes in normal and diseases processes. However, systematic detection of higher-order structural changes across multiple Hi-C datasets remains a major challenge. Existing computational methods either do not model higher-order structural units or cannot model dynamics across more than two conditions of interest. We address these limitations with Tree-Guided Integrated Factorization (TGIF), a generalizable multi-task Non-negative Matrix Factorization (NMF) approach that can be applied to time series or hierarchically related biological conditions. TGIF can identify large-scale changes at compartment or subcompartment levels, as well as local changes at boundaries of topologically associated domains (TADs). Based on benchmarking in simulated and real Hi-C data, TGIF boundaries are more accurate and reproducible across differential levels of noise and sources of technical artifacts, and more enriched in CTCF. Application to three multi-sample mammalian datasets shows TGIF can detect differential regions at compartment, subcompartment, and boundary levels that are associated with significant changes in regulatory signals and gene expression enriched in tissue-specific processes. Finally, we leverage TGIF boundaries to prioritize sequence variants for multiple phenotypes from the NHGRI GWAS catalog. Taken together, TGIF is a flexible tool to examine 3D genome organization dynamics across disease and developmental processes.

Introduction

The three-dimensional (3D) organization of the genome refers to the packaging of DNA inside the nucleus. It has emerged as a key regulatory mechanism of cellular function and dysfunction across diverse developmental (Zheng and Xie, 2019), disease (Lupiáñez et al., 2016), and evolutionary contexts (McCord, 2017; Eres et al., 2019). High-throughput chromosomal conformation capture (Hi-C) technologies enable the study of 3D genome organization by experimentally measuring the tendency of genomic regions to spatially interact with one another (Kempfer and Pombo, 2020; Mumbach et al., 2016; Kempfer and Pombo, 2019; Dekker et al., 2023). The 3D genome is organized into structural units at multiple scales: compartments spanning several megabases, Topologically Associated Domains (TADs) spanning hundreds of kilobases scale, and enhancer-promoter loops involving pairs of loci of a few thousand bases. (Bouwman and de Laat, 2015; Rowley and Corces, 2018; Kempfer and Pombo, 2020). Changes in 3D genome organization at different topological levels have been observed with transitions in both normal (Bonev et al., 2017; Stadhouders et al., 2018; Zheng and Xie, 2019) and disease processes (Lupiáñez et al., 2016; Norton and Phillips-Cremins, 2017; Wang et al., 2021). Through efforts from large-scale consortia such as the 4D Nucleome project, Hi-C measurements are becoming increasingly common from multiple conditions corresponding to time points, cell types and species (Dekker et al., 2017; Reiff et al., 2022; Dekker et al., 2023; Roy et al., 2023). These datasets provide a unique opportunity to examine the dynamics of 3D genome organization across space and time and its impact on disease and normal processes.

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

TAD finding could make these approaches more susceptible to finding non-biological differences and technical artifacts (Fletez-Brant et al., 2024; Kobets et al., 2023; Zheng et al., 2022). Numerous studies have shown that Hi-C count profiles obey cell type, timepoint and species relationships, where datasets from nearby contexts are more similar than those that are far away (Bonev et al., 2017; Zhang et al., 2019; Yang et al., 2017; Vietri Rudan et al., 2015). An approach that constrains the TAD and compartment finding based on such prior information about the relationships between the input datasets could be less prone to spurious differences. A few methods have been developed to directly identify TAD boundary differences, but they are focused on pairs of conditions (Wang et al., 2020) or limited in their ability to compare more than two conditions (Cresswell and Dozmorov, 2020).

To address the dearth of methods for identifying large-scale organizational changes, especially when considering more than two datasets, we developed Tree-Guided Integrated Factorization (TGIF). TGIF is a multi-task, Non-negative Matrix Factorization (NMF) framework enabling joint embedding of multiple Hi-C matrices and identification of compartments and TADs across multiple conditions. NMF is a popular dimensionality reduction approach for analyzing non-negative, genome-scale datasets (Stein-O'Brien et al., 2018; Kotliar et al., 2019; Lee and Roy, 2021), where the low-dimensional factors capture the biologically meaningful structure of the data. Multi-task NMF frameworks factorize multiple input matrices simultaneously to yield low-dimensional embeddings in a shared latent space. They have been successfully applied to single-cell omics data to integrate matrices from multiple samples, experiments, and modalities by removing batch effect and technical noise (Welch et al., 2019; Liu et al., 2020; Kriebel and Welch, 2022; Luecken et al., 2022; Hu et al., 2024). TGIF incorporates a hierarchical multi-task NMF formulation to simultaneously factorize multiple Hi-C matrices from related biological conditions and constrain the factors from closely related conditions to be similar. The factors represent low-dimensional embeddings of genomic regions in an aligned latent space, representing their global or local chromatin architecture. We use such embeddings to identify changes at both the compartment and TAD levels. When applied to simulated and real timecourse Hi-C matrices, TGIF identifies fewer false positive differences in TAD boundaries, and produces a more reproducible set of boundaries across biological replicates, normalization methods, depths, and resolutions compared to other methods. Differential boundaries and compartmental regions identified by TGIF show significant changes in relevant biological signals such as gene expression, histone modification, and chromatin accessibility. Finally, persistent boundaries identified by TGIF are enriched in sequence variants associated with cardiovascular disease. Together, these results demonstrate the versatility and utility of TGIF to examine changes in higher-order 3D genome organization across diverse types of dynamic processes.

Results

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

Tree-guided Integrated Factorization (TGIF) is a general-purpose framework to study 3D genome organization dynamics both at the TAD and compartment levels (**Figure 1**). TGIF is based on multi-task non-negative matrix factorization (NMF). It takes as input a set of Hi-C matrices, each representing a biological condition, and a user-specified tree structure that can encode an arbitrary relationship among the conditions, such as time or cell type lineage (**Figure 1**, **Figure S1**). TGIF uses a novel regularization term in its objective to jointly factorize the matrices such that input matrices from more closely related conditions result in more similar lower-dimensional representations, i.e., factors.

To handle both compartment and TAD identification, we implemented two versions of TGIF: TGIF-DB and TGIF-DC. TGIF-DB identifies conserved and differential boundaries demarcating TADs under different conditions (**Figure 1A**, **Figure S1A**, **Methods**), while TGIF-DC identifies compartment-level changes in 3D genome organization (**Figure 1B**). In TGIF-DB, the factorization is performed on sub-matrices along the diagonal of the intrachromosomal Hi-C matrices, as these diagonal sub-matrices capture the TAD-scale, local topology of chromosomes. Each sub-matrix is factorized over a range of k, the hyper-parameter specifying the rank of the lower-dimensional space (**Figure S1B**). TGIF-DB calculates a boundary score from the factors at each k, which are averaged to provide an overall boundary score (**Figure S1C**). Considering multiple k allows us to capture structural units or domains of different sizes in the lower dimensional space and removes the need to specify the number of factors (**Methods**). TGIF-DB identifies regions with significant boundary scores by comparing the average boundary scores against a "null distribution" of boundary scores to calculate an empirical p-value (**Figure S1D**). TGIF-DB outputs the list of significant boundaries corresponding to each input dataset and a list of significantly differential boundary regions for every pair of input count matrices (**Figure S1E**, **Methods**).

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

compartment and subcompartment regions for every pair of input conditions (**Methods**).

In cases where the relationship between the input Hi-C data is not available (e.g. integrating Hi-C datasets from multiple studies or pseudo-bulk single-cell Hi-C data from cell clusters; Zhou et al., 2019; Zhang et al., 2022) TGIF can infer a tree structure based on the pairwise similarity of the input Hi-C matrices measured by stratum-adjusted correlation coefficient (SCC; Yang et al., 2017, **Methods**, **Figure S2**) or a similar distance-stratified metric.

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

TGIF-DB was benchmarked against four other TAD calling methods: three methods designed for calling TADs and boundaries from a single Hi-C matrix (which we refer to as single-task methods), and one designed specifically for differential boundary identification (**Methods**). The three single-task methods were GRINCH (Lee and Roy, 2021), SpectralTAD (Cresswell et al., 2020), and TopDom (Shin et al., 2016, **Supplemental Methods**). TADCompare (Cresswell and Dozmorov, 2020) is a method designed for differential boundary detection.

Since real Hi-C datasets do not have ground-truth set of TAD boundaries, we first evaluated the quality of TAD boundaries identified by each method in simulated datasets. We generated 4 Hi-C matrices each with its own set of ground-truth boundaries based on the count simulation procedure from Forcato et al., 2017 and noise added to 10, 20, 30, 40% of interaction counts (**Supplemental Methods**). For every pair of matrices, we calculate the precision and recall of boundaries found only in one matrix (”task-specific” boundaries) and those shared between the two input matrices (shared boundaries). Across the different levels of noise, TGIF-DB has the highest precision on task-specific boundaries (**Figure 2A**). With the exception in the lowest level of noise (10%), TGIF-DB is among the methods with the highest precision for shared boundaries along with GRiNCH and TopDom (**Figure 2A**). For recall of task-specific boundaries (**Figure S3A**), TGIF-DB is second to TopDom in all but the lowest noise level. For shared boundaries, TGIF-DB and TopDom also have the highest recall in all but the lowest noise level.

Next, we evaluated the quality of TAD boundaries identified by each method based on the enrichment of CTCF binding. CTCF is an architectural protein associated with establishing boundaries (Merkenschlager and Nora, 2016; Gómez-Díaz and Corces, 2014; Cubeñas-Potts and Corces, 2015). We used the time-series dataset of cardiomyocyte differentiation (Zhang et al., 2019) which profiled both genome-wide chromosome conformation with Hi-C and CTCF binding with ChIP-seq. The boundary regions

predicted by each method for each timepoint was used to calculate their fold enrichment of CTCF peaks against the genomic background (**Methods**). Significant boundaries identified by TGIF-DB has the highest fold enrichment, followed by single-task methods, TopDom and GRiNCH (**Figure 2B**).

We next measured the Jaccard score between the boundary sets from a pair of biological replicates of H1 human embryonic stem cell line (Zhang et al., 2019, **Methods**), TADCompare and TGIF-DB had the highest scores, recovering more similar set of boundaries between the biological replicates compared to other methods (**Figure 2C**). Furthermore, differential boundaries between two timepoints (day 0 and day 2 of cardiomyocyte differentiation) identified by TADCompare and TGIF-DB had the fewest false positives based on overlap with differential boundaries between two biological replicates of the same timepoint (**Supplemental Methods**, **Figure S3B**). We also benchmarked the degree of false-positive, non-biological differences identified by each method in: (1) Hi-C datasets with different depths, (2) datasets from biological replicates, (3) datasets normalized using different methods, and (4) datasets in different bin resolutions. To this end, we downsampled a high-depth Hi-C dataset from the GM12878 cell line with 4.01 billion reads (Rao et al., 2014; Reiff et al., 2022) by subsampling 5, 10, 25, 50% of the reads (**Methods**). We measured the Jaccard score between the boundary sets from the original high-depth input and the downsampled counterpart. The higher the Jaccard index, the fewer the false-positive differences identified by a method. Across all downsampled depths, TADCompare and TGIF-DB were the top performing methods with consistently high Jaccard scores (**Figure 2D**). Single-task methods (GRiNCH, SpectralTAD, and TopDom) had much lower Jaccard score with discrepancy increasing with depth differences. We observed similar results with TADCompare and TGIF-DB obtaining the highest Jaccard score between TAD boundary sets from mouse embryonic stem cell (mESC) Hi-C data normalized using different methods (**Figure 2E**, **Methods**). Finally, we measured the stability of TAD boundaries to the changing resolution (10kb, 25kb, 50kb) of input Hi-C matrices using Jaccard score (**Methods**). TGIF-DB and GRiNCH yield the most stable or similar boundaries to changing resolution (**Figure S3C**), with the exception of 25kb-50kb where TopDom also performed well. These results demonstrate the advantages of using TGIF-DB to identify biologically relevant boundaries enriched in known boundary elements while minimizing false positive differences.

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

We compared the TGIF-DC compartments and differential compartments to three existing methods on the H1 hESC and endoderm differentiation dataset (**Supplemental Methods**): PCA-based (Lieberman-

Aiden et al., 2009),Cscore (Zheng and Zheng, 2018), anddcHiC (Chakraborty et al., 2022).

We first compared the similarity of compartment assignments between different methods using Rand Index (**Figure 3A**). The PCA-based method and dcHiC, which also utilizes PCA, produced the most similar compartments (Rand Index: 0.91), followed by TGIF-DC (Rand Index: 0.79-0.8). Cscore found a substantially different set of compartments (Rand Index: 0.52). We assessed the quality of compartments with three cluster quality metrics, Silhouette Index (SI, **Figure 3B**), Calinski-Harabasz score (CH, **Figure 3C**), and Davies-Bouldin Index (DBI, **Figure 3D**), using observed-over-expected (O/E) counts as features of each genomic loci (**Methods**). In all three metrics, TGIF-DC, dcHiC, and PCA-based compartments are comparable in their quality and outperformed Cscore. We also measured compartment quality using chromatin accessibility, a key regulatory measurement that characterizes different compartment types (e.g., the active A and repressive B; Lieberman-Aiden et al., 2009; Fortin and Hansen, 2015). Briefly, we measured SI (**Figure 3E**), CH (**Figure 3F**) and DBI (**Figure 3G**) using the mean basepair ATAC-seq signal for each 100kb region as the feature (**Methods**). For all three metrics, the compartments from TGIF-DC, PCA-based method, and dcHiC are of similar quality.

Finally, we compared TGIF-DC exclusively with dcHiC, the only other method that specifically identifies *differential* compartment regions. Significantly differential compartmental regions (sigDC) identified by TGIF-DC have significantly higher change in accessibility signal and gene expression compared to regions not part of sigDC (**Figure S5A,B**). Compared to significantly differential regions identified by dcHiC, sigDC regions from TGIF-DC also have significantly higher change in accessibility signal (*t*-test p-value $<$1e-2, **Figure 3H**) and are comparable in terms of the change in gene expression levels (**Figure 3I**).

Taken together, TGIF-DC captures compartment structure consistent with established compartment-calling methods, while pinpointing differential regions with significant changes in regulatory signals such as chromatin accessibility.

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

While compartments provide a global partitioning of each chromosome, the genome is hierarchically organized with compartments further partitioned into smaller subcompartments that could represent functionally distinct set of regions (Rao et al., 2014; Xiong and Ma, 2019). TGIF-DC has a tunable parameter (k, the rank of factors) that can be used to identify such subcompartments. To demonstrate TGIF-DC's ability to identify both compartments and subcompartments, we applied it to the mouse neu-

ral differentiation dataset with 3 timepoints: embryonic stem cell or ES, neural progenitors or NPC, and cortical neurons, CN (**Figure S4,6,7**). This dataset additionally measured six different histone modification signals for NPC and CN that were beneficial for additional biological interpretation of TGIF-DC results (**Figure 4A**, **Methods**). We first analyzed the compartment structure from TGIF-DC ($k = 2$) for each chromosome, based on GC content (mean GC percentage for each 100kb bin, **Methods**), annotating the compartment with higher GC content as compartment A and the one with lower GC content as compartment B (**Methods**, **Figure S8**). Regions annotated as A compartment by TGIF-DC have significantly higher signal for marks associated with active enhancer (H3K27ac, H3K4me1) or elongation (H3K36me3) than those in B compartment (**Figure 4B**).

We next applied TGIF-DC with $k = 5$ to identify subcompartment structure per chromosome, each k corresponding to a different subcompartments (**Methods**). We interpreted these subcompartments based on the mean histone modification signal of the genomic loci assigned to each subcompartment. The subcompartments exhibited distinct histone modification patterns (**Figure 4C**, Chr18), with subcompartments 1 and 5 associated with repressive marks (H3K9me3, H3K27me3), while the other three (2, 3 and 4) associated with active marks. Within these two groups, each subcompartment had a different signature of marks. For example, subcompartment 3 exhibits relatively lower signal of H3K36me3 compared to 2 and 4, while subcompartment 2 had a higher signal of all three activating marks (H3K27ac, H3K36me3, H3K4me1) compared to 3 and 4. Between the two subcompartments, 1 and 5, with repressive mark association, one (1) exhibited higher H3K4me3 and H3K9me3 levels compared to the other one (5).

Finally, we assessed TGIF-DC's differential subcompartments by measuring the log fold change in histone modification signals between two timepoints, NPC and CN, and *k*-means clustering the regions based on this signal difference. We find distinct subgroups of regions with different fold change of the three activating marks, H3K27ac, H3K36me3, and H3K4me1 (**Figure 4D**). The repressive marks or the promoter specific mark, H3K4me3, did not vary substantially for these regions.

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

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

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

We applied TGIF to the three timecourse datasets with both Hi-C and RNA-seq measurements (**Figure S4**, **Table S1-3**): (1) H1 hESCs differentiated to endoderm(Reiff et al., 2022; Dekker et al., 2023, **Figure S5C-F**, **Figure S9**), (2) mouse neural differentiation time course from mESC to cortical neurons (CN, Bonev et al., 2017, **Figure S6,7,10**), and (3) human cardiomyocyte differentiation timecourse from hESC to ventricular cardiomyocytes (Zhang et al., 2019, **Figure S11,12**). We performed pairwise comparison of differential boundary, compartment, and gene expression, e.g. H1 vs endoderm, mESC vs NPC, day 0 vs day 2 of cardiomyocyte differentiation. Within each pairwise comparison, we asked whether differentially expressed (DE) genes are enriched in three different sets of dynamic regions (**Methods**, **Figure 5A**): (A) regions near (i.e., within 100kb) of significantly differential boundaries (sigDB), (B) regions within a TAD with at least one sigDB, and (C) regions within significantly differential compartmental regions (sigDC). Differential regions within 100kb of sigDB are consistently enriched for DE genes (**Figure 5B** top, **Table S5-7**). Furthermore, genes within 100kb of sigDB are also enriched for DE when compared to all genes (**Figure 5B** bottom). Regions in set B (within a TAD with at least one sigDB) do not show consistent enrichment, likely because of the permissive inclusion criteria for set B. Regions within sigDC are significantly enriched in DE genes for the H1-endoderm differentiation and majority of the comparisons in the cardiomyocyte differentiation dataset. The enrichment for genes was lower, possibly due to the large number of genes within compartments.

To assess the biological significance of DE genes near differential boundaries, we examined the biological processes enriched in DE genes near sigDBs compared to processes enriched in other genes (**Methods**). In the cardiomyocyte differentiation data, DE genes in general showed significant enrichment for generic developmental terms like multicellular organismal development (**Figure 5C**, **Table**

S8). However, DE genes near sigDBs tended to be significant for processes specific to cardiac and heart development (e.g. cardiac cell differentiation, heart development and morphogenesis). DE genes near sigDB between H1 and endoderm also showed significant enrichment in developmental terms (e.g. cell morphogenesis involved in differentiation, cellular component organization or biogenesis) compared to those not near sigDB (**Table S9**). For the mouse ESC to CN differentiation, DE genes near sigDB were enriched for neuronal processes when comparing ES vs CN and ES vs NPC (**Table S10**).

Finally, to characterize specific loci with differential 3D organization pattern, we prioritized regions based on the magnitude of change in their boundary scores, then overlapped them with genomic features such as retrotransposons. Human endogenous retrovirus subfamily H retrotransposons (HERV-H) in particular have been implicated in chromatin organization (Lawson et al., 2023) as a major determinant of TAD boundaries specific to hESC (i.e., day 0 of cardiomyocyte differentiation) when transcriptionally active (Zhang et al., 2019). Boundary scores at the top 100 transcriptionally active HERV-H sites is higher in hESC (day 0) compared to subsequent timepoints (**Figure 6A**). We observe presence of such boundary unique to day 0 that disappears in subsequent timepoints at one of the top transcriptionally active HERV-H sites (**Figure 6B**). Among the top-ranked sigDBs based on change in boundary scores, we found sigDB regions where a boundary is present in the pluripotent state, but absent in differentiated state (**Figure 6C,D**). These sigDB instances are proximal to the *ESRG* gene, highly expressed in the pluripotent state compared to the subsequent differentiated states. *ESRG* is a HERV-H containing long noncoding RNA (lncRNA, Wang et al., 2014); in addition to demarcating domain boundaries in hESCs, this particular site may effect the pluripotency state on knockdown (Wang et al., 2014) and has known roles in developmental and embryonal carcinoma (Wanggou et al., 2012).

Among other top-ranked sigDBs in cardiomyocyte differentiation, we found DE genes with known roles in the cardiac development. For example, a boundary was found in primitive cardiomyocytes (day 15) but absent in ventricular cardiomyocytes (day 80, **Figure S13A**). This boundary overlaps *MYH6*, highly expressed in day 15 compared to day 80, and is adjacent to *MYH7*, displaying the opposite expression change pattern to *MYH6*. Both genes are involved in cardiac muscle function (Ching et al., 2005; Warkman et al., 2012). Recently an enhancer cluster located downstream to *MYH7* at Chr14:23,876,121-23,878,188, was identified as a switch that can downregulate expression of *MYH7* while upregulating *MYH6* (Gacita et al., 2021). We also identified a sigDB close to the *Ncam1* gene which is differentially expressed between ES and CN (**Figure S13B**); *Ncam1* has known roles in neuron axon guidance and synapse formation (Hata et al., 2018; Shetty et al., 2013). These examples provide further evidence for TGIF-DB's ability to identify relevant dynamic boundaries that could impact overall cell state identity.

Persistent boundaries are enriched for SNPs from diverse disease phenotypes

Single nucleotide polymorphisms (SNPs) identified from genome-wide association studies (GWAS) are frequently found in noncoding regions of the genome and have been implicated in disease phenotypes by affecting the 3D genome organization (Lupiáñez et al., 2015; Orozco et al., 2022). Specifically, such variants could disrupt TAD boundaries and cause promiscuous expression of genes (Lupiáñez et al., 2015; Chakraborty and Ay, 2018). We investigated whether TGIF boundaries from the human cardiomyoctye differentiation data could be used to examine regulatory variants identified for diverse disease phenotypes in GWAS. We considered 17 phenotypic categories from the GWAS catalog and tested the enrichment of SNPs from each category in TGIF boundaries (**Methods**). SNPs across different categories were most enriched in the common set of boundaries across timepoints (i.e., persistent boundaries) than in other timepoint-specific or broader subsets of boundaries, with hematological measurement, cardiovascular disease, and lipid or lipoprotein measurement being the most enriched phenotypic categories (**Figure 7A**). Importantly, SNPs associated with cardiovascular disease (CVD) exhibited the second highest enrichment. The traits that had lower enrichment included neurorological disorders and non-specific categories. We examined 66 persistent boundaries with at least one CVD-associated SNP. One such boundary had the SNP, *rs72705895*, which is associated with venous thromboembolism (Lindström et al., 2019, **Figure 7B**), and additionally overlaps a CTCF binding site (regulatory feature *ENSR00000255184* from Ensembl regulatory build annotations; Zerbino et al., 2015; Cunningham et al., 2022). Another boundary included *rs9349379*, which is found in the intronic region of *PHACTR1* (**Figure S14**). Both the intronic variant and the gene are associated with coronary artery atherosclerotic disease (Kuveljic et al., 2021; Koitsopoulos and Rabkin, 2021), while the SNP itself is on a predicted enhancer region (Ensembl regulatory build annotation *ENSR00001107203*), suggesting its putative role in disrupting an intronic enhancer. Genome editing experiments of boundary locations harboring these SNPs combined with Hi-C assays could help examine the role of dysregulated 3D genome organization as a possible mechanism by which regulatory variants impact phenotype.

Discussion

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

TGIF's design is motivated by a number of considerations: (a) TAD and compartment identification are unsupervised learning problems with no ground truth for real Hi-C datasets. Since Hi-C data can be sparse, identification of such structures and assessing how much they change could be susceptible to statistical, non-biological differences. (b) Several studies from multiple cell types, time points, and species have shown that TAD and compartment is conserved across species (Dixon et al., 2012; Vietri Rudan et al., 2015). TGIF's hierarchical, multi-task learning framework exploits this prior information to constrain the identification of organizational structures while being sensitive to the extent of relatedness of the datasets by using a tree structure. (c) Finally, TGIF is motivated by a dimensionality reduction (matrix factorization) framework to reduce the noisy, high-dimensional count profile of each genomic locus into a low dimensional space of different ranks. This enables TGIF to be a general framework that identifies TADs, compartments, as well subcompartments and their dynamics. Application of TGIF and existing methods to simulated and read Hi-C time course datasets showed that TGIF can accurately recover structural units such as compartments and topologically associated domains (TADs), while having lower false positive rate and greater robustness to technical differences between datasets such as depth, normalization, and resolution. TGIF also identifies biologically meaningful differences in 3D genome organization that are supported by numerous one-dimensional features such as architectural protein enrichment, histone modification, and differential expression.

An open question with topological domain changes is how they relate to changes in gene expression (Greenwald et al., 2019; Ghavi-Helm et al., 2019; Cavalheiro et al., 2021; McArthur and Capra, 2021). At the TAD level, fusion or inversion of TADs could result in gene expression change although the

extent to which such changes are genome wide or are specific to disease-associated genes is still unclear (Cavalheiro et al., 2021). Evidence suggests that RNA polymerase elongation or the binding of pre-initiation complex to the DNA during transcription can give rise to domain structures, providing a direct mechanistic link between transcription and 3D genome organization (van Steensel and Furlong, 2019; Heinz et al., 2018). This relationship can further depend upon the developmental stage or differentiation status of cells (Pollex et al., 2024; Chen et al., 2024). However, this has been debated in other studies, for example, during *Drosophila* development (Ing-Simmons et al., 2021; Espinola et al., 2021).

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

Regulatory sequence variants can mis-regulate gene expression by disrupting TAD boundaries (Lupiáñez et al., 2015; Chakraborty and Ay, 2018). We used our TAD boundaries to examine the impact of this variation. We found the greatest enrichment in boundaries that did not change over time, namely the persistent boundaries. Furthermore, we found several cardiovascular and metabolic disease trait SNPs to be enriched in these boundaries. These persistent boundaries may be specific to the entire cardiac tissue as a whole rather than a specific developmental time or stage. As future work, it would be worth investigating persistent boundaries in other developmental lineages and their propensity to prioritize SNPs for diseases in tissue-specific manner. Additionally, this provides a way to prioritize variants for downstream functional experiments that could be important to identify the mechanisms by which variants disrupt gene regulatory processes.

There are a number of directions in which TGIF could be extended. One direction is to consider our benchmarking results and identify areas of improvement where TGIF-DB is currently not the best method. For example, TGIF-DB has higher precision for task-specific boundaries in simulated data but at the cost of lower recall compared to TopDom. TGIF also finds higher percentage of false positive differential boundaries between biological replicates than TADCompare, and does not significantly out-

perform dcHiC when comparing gene expression change within differential compartments. Such non-optimal performance could be due to TGIF's regularization scheme which shares information across contexts, but does not explicitly capture differences between them. To address this limitation, TGIF's loss function could be extended to include a contrastive term. Another direction is to enable greater flexibility in capturing dataset relatedness. Currently, TGIF uses the same hyper-parameter value for all branches of the tree, which could be limiting when a more granular control is desirable to define the relationship between the datasets. An extension to TGIF could allow varying hyper-parameter values depending upon the position in the hierarchy, as informed by auxiliary information such as phylogenetic branch length across species or gene expression similarity across cell types. A third direction of research is to consider auxiliary measurements, including sequence, to inform the inference of the topological units using techniques such as semi-supervised clustering (Bair, 2013; Bondell and Reich, 2008).

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

Materials and Methods

Tree-Guided Integrated Factorization (TGIF)

Tree-Guided Integrated Factorization (TGIF) is based on multi-task Non-negative Matrix Factorization (NMF, Lee and Seung, 2000) and can be used to identify low-dimensional structures across multiple Hi-C datasets. The tasks in TGIF correspond to Hi-C datasets that in turn are from hierarchically related contexts, such as cellular stages, species, timepoints. TGIF extends an existing framework, multi-view NMF (Liu et al., 2013; Baur et al., 2022, **Supplemental Methods**) which assumes all the tasks are equally related. TGIF generalizes multi-view NMF to allow for integration of datasets from different biological contexts such as time or developmental stage, and therefore may not all be equally related to each other.

Formally, TGIF takes as input $t \in \{1, \ldots, T\}$ matrices representing T tasks. Each matrix $\mathrm{X}^{(t)} \in \mathbb{R}^{n \times n}$ is a symmetric Hi-C count matrix over n genomic loci. TGIF also requires as input a task tree which describes parent-child relationships between the tasks. (**Figure 1A**). Given these inputs, TGIF optimizes the following objective:

$$\sum_{t=1}^{T} \left\| \mathrm{X}^{(t)} - \mathrm{U}^{(t)}\mathrm{V}^{(t)} \right\|_{\mathrm{F}}^{2} + \alpha \sum_{c} \left\| \mathrm{V}^{(c)} - \mathrm{V}^{\mathrm{Pa}(c)} \right\|_{\mathrm{F}}^{2} \tag{1}$$

The objective aims to:

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

The hyper-parameter α controls the strength of the constraints such that the higher the α, the more the factor $\mathrm{V}^{(c)}$ is encouraged to be similar to its parent. Selection of α is discussed in **Supplemental Methods, Figures S18, S19, S20, S21**.

TGIF uses block coordinate descent (BCD) optimization scheme to learn these factors because BCD guarantees convergence to a local optimum (Kim et al., 2014). Additional details of the TGIF algorithm can be found in **Supplemental Methods**.

TGIF's factors can be used to find changes in compartments as well as changes in boundaries of finer-

scaled topologically associating domains (TADs). TGIF-DB for differential boundary identification and TGIF-DC for identifying differential compartment regions are described in detail in subsequent sections.

TGIF-DB for differential boundary identification

TGIF-DB identifies TAD boundaries in four major steps: (a) multi-task factorization of input Hi-C matrices (b) boundary score computation (c) empirical p-value calculation and FDR correction to detect significant boundaries; and (d) identification of significant differential boundaries (sigDB).

Multi-task factorization of input Hi-C matrices. TGIF-DB applies TGIF to small partially overlapping submatrices along the diagonal of the symmetric intra-chromosomal interaction count matrices (**Figure S1A,B**). This mirrors the approaches taken by existing TAD-calling methods (Lieberman-Aiden et al., 2009; Cresswell and Dozmorov, 2020; Li et al., 2021). By default each submatrix spans 2Mb$\times$2Mb with an overlap "step size" of 1Mb between consecutive submatrices. The exact dimension of the submatrix, namely the number of rows and columns, will depend on the resolution of the Hi-C data. The minimum size of the submatrices is bound at 100 (and the corresponding step size at 50) genomic regions to prevent over-fragmentation of the input matrices, especially for lower-resolution input Hi-C matrices. Regions with interaction values missing for more than half of its neighbors in the radius defined by the window size in any of the input matrices are filtered out from the original input intra-chromosomal matrices before any submatrices are formed. In NMF, usually the rank k of the lower dimensional factors is user-specified. However, TGIF does not require this since a single k value may not be appropriate across all task-specific input submatrices. Instead TGIF scans a range of k values, with $k \in \{2, \cdots, 8\}$ to recover lower dimensional factors at multiple resolutions and defines boundaries based on a consensus of these factors (as described below). Because the submatrix size is small, it is computationally tractable to scan a range of k.

Boundary score calculation. After factorization, the next step is to identify genomic regions representing conserved or dynamic TAD boundaries across conditions. We define a boundary as a region whose low-dimensional representation changes significantly compared to its immediate preceding neighbor bin. To this end we define a boundary score $S_i^{(t)}$ using the output factors for each of the T tasks from TGIF. Since $\mathbf{X}^{(t)}$ is symmetric, either $\mathrm{U}^{(t)}$ or $\mathrm{V}^{(t)}$ could be used to estimate these boundary scores. Assuming we use $\mathrm{U}^{(t)}$, the score $S_i^{(t)}$ for each region i in task t is the cosine distance between the low

dimensional representation of region i and region $i-1$:

$$S_i^{(t)} = 1 - \frac{\mathrm{U}^{(t)}[i,:] \cdot \mathrm{U}^{(t)}[(i-1),:]}{\|\mathrm{U}^{(t)}[i,:]\| \|\mathrm{U}^{(t)}[(i-1),:]\|} \quad (2)$$

The final boundary score for region i in task t is the mean of $S_i^{(t)}$ estimated from factors across the range of $k \in \{2, \cdots, 8\}$ (**Figure S1C**). For regions that are in the overlapping window between two consecutive count submatrices, the final boundary score is averaged from across all submatrix factors.

Empirical p-value calculation and FDR correction. Once the scores are calculated, we estimate a "null" distribution of boundary scores and use it to determine the empirical p-value of boundary scores and find significant boundaries. The null distribution is computed from a randomized background matrix (**Supplemental Methods**). We calculate the empirical p-value for each the region i in task t as the proportion of "null" background scores higher than the given region's boundary score. Finally, to find significant boundaries and to correct for multiple significant testing, we perform the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The output of the p-value and FDR estimation step is a binary value for each region i and task t, indicating whether the region has a significant boundary score (1) or not (0). The significant boundaries identified in this manner may still be susceptible to noisy, low count regions of the genome. Therefore, we additionally filter the boundaries to find "summit-only" version of the significant boundaries, i.e. if there are more than one consecutive significant boundary regions along the linear genome, only the region with the highest significant score is called a boundary. To characterize the boundaries excluded by the summit-only approach, we compared both summit and non-summit boundaries in H1-endoderm data. Non-summit boundaries tend to be shared across cell types and comprise about 50% of the total significant boundaries (**Figure S15**). The current implementation of TGIF-DB outputs both summit-only and all significant boundaries allowing the user to use their own cut-offs.

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

< 0.05 as significantly differential boundary (sigDB) regions. We further annotate the type of change represented by each sigDB between conditions A and B: a boundary created in B, deleted in B, or shifted in B within 5 genomic bins (**Supplemental Methods**).

TGIF-DC for differential compartment and subcompartment identification

Identification of compartments with TGIF-DC. In order to identify compartments, we apply TGIF to a 100kb resolution intrachromosomal Hi-C matrix that is first converted into an observed-over-expected (O/E) count correlation matrix as described previously in Rao et al., 2014. We upshift the correlation matrix by 1 so that all values are non-negative. To identify compartments, we apply TGIF with the input tree structure to these matrices with rank $k = 2$. After factorization, we infer each region i's cluster assignment, $c_i^{(t)}$, for each task t, such that $c_i^{(t)} = \text{argmax}_{j \in \{1,2\}} \text{U}[i, j]$. We refer to these clusters as compartments. To identify subcompartments and differential subcompartment regions, a higher k value, e.g. 5, can be used and TGIF-DC will generate more granular cluster assignments, e.g. 5 clusters of regions instead of 2 clusters. Each of these clusters corresponds to a subcompartment.

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

Post-hoc annotation of TGIF-DC clusters into A and B compartments TGIF-DC by default uses k=2 and segments the given chromosome into 2 clusters of regions. In our analysis, we use GC content and chromatin accessibility to annotate each cluster as A or B compartment, in a manner similar to existing analysis and tools (Fortin and Hansen, 2015; Kruse et al., 2020). Briefly, the cluster with higher mean accessibility signal (measured by ATAC-seq or DNase-seq) or GC content is assigned to A compartment, and the other cluster to B compartment. Detailed annotation process for each of the developmental timecourse datasets can be found in **Supplemental Methods**, **Figure S24**, **Figure S25**.

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

When prior information about the relationship among the input matrices is not available, a tree structure can be estimated using pairwise similarity of the input Hi-C matrices, converting to distance followed by hierarchical clustering. We suggest the use of a distance-stratified similarity measure, such as the stratum-adjusted correlation coefficient (SCC, Yang et al., 2017, **Figure S2A**), that we have also used for our hyper-parameter analysis (**Supplemental Methods**). Once SCC is calculated for each pair of input matrices, it is converted to a distance by subtracting from 1 (**Figure S2B**) which in turn is used as input to hierarchical clustering with average linkage. We tested this approach for the mouse neural differentiation dataset and found that the output tree of hierarchical clustering is similar to the known biological relatedness of this dataset (**Figure S2C**, Bonev et al., 2017) and is identical to the tree we used as input to TGIF for our experiments. The current implementation of TGIF offers this functionality as a pre-processing script (See Section on **Software Availability**).

Datasets used in analysis

We applied TGIF to three Hi-C timecourse datasets: H1 hESC differentiated to endoderm (Reiff et al., 2022; Dekker et al., 2023), mouse neural differentiation data from Bonev et al., 2017, and human cardiomyocyte differentiation data from Zhang et al., 2019. See **Supplemental Methods**, **Figure S4**, and **Table S1**-**Table S4** for processing, application of TGIF, and list of accession numbers.

Benchmarking methods for identifying differential domain boundaries

Existing methods used in benchmarking TGIF-DB TGIF-DB was benchmarked against four other methods for identifying differential TAD boundaries: GRINCH (Lee and Roy, 2021), SpectralTAD (Cresswell et al., 2020), TADCompare (Cresswell and Dozmorov, 2020), and TopDom (Shin et al., 2016). GRiNCH, SpectralTAD, and TopDom are single-task TAD identification methods accepting a single input matrix individually followed by pairwise comparison of identified boundaries. TADCompare is a differential TAD identification method that can take as input a pair of Hi-C matrices as well as a time series of Hi-C matrices. These methods are described in more detail in **Supplemental Methods**, **Figure S22**.

Benchmarking on simulated data with known boundaries. In order to benchmark methods that can detect TAD-level changes, we generated simulated contact matrices with known TADs and TAD changes for four hierarchically related conditions (**Figure S16**). The TAD changes can fall into one of three categories: TAD split creating a new boundary, TAD merge removing a boundary, and TAD shift where the location of a boundary is moved up or down the linear chromosome (**Figure S16A**, Cresswell and Dozmorov, 2020). We first generate a set of TADs with known change patterns, then populate contact matrices following the Hi-C count simulation procedure in the benchmarking study by Forcato et al., 2017. The simulation procedure is detailed in **Supplemental Methods**. We applied GRiNCH, SpectralTAD, TADCompare, TGIF-DB, and TopDom to the simulated datasets to assess their ability to recover shared and differential boundaries. We applied TADCompare to each pair of the 4 simulated matrices. TADCompare outputs differential boundaries, including the task in which the boundary is significant, and non-differential boundaries, which we consider as shared boundaries. We applied single-task methods (GRiNCH, SpectralTAD, TopDom) to each of the four simulated matrices independently to identify the TADs for each input matrix. The resulting TAD boundaries for each pair of input matrices were compared to identify task-specific and shared boundaries. TGIF was applied to all four simulated matrices together with the known tree structure used to generate the simulated data (**Figure S16C**). We calculated precision and recall of task-specific and shared boundaries in every pair of simulated matrices. Shared boundaries between simulated matrices A and B are boundaries found or identified in both A and B. Task-specific boundaries are boundaries found in A but not in B, and vice versa.

Measuring CTCF enrichment in boundaries To evaluate the boundaries identified by various TAD-calling methods, we measured CTCF peak enrichment in boundaries found in the cardiomyocyte differentiation dataset (**Table S3**). Using MACS2 (Zhang et al., 2008), we first called peaks on CTCF ChIP-seq data from each of the 6 timepoints (day 0, 2, 5, 7, 15, 80) of the cardiomyocyte differentiation time course. Replicates from each timepoint were collapsed by intersecting overlapping peaks with BEDTools (Quinlan and Hall, 2010). Each peak was then assigned to a 10kb uniform bin again using BEDTools. TAD-calling methods GRiNCH, SpectralTAD, and TopDom were applied to 10kb Hi-C matrices from each of the 6 timepoints. TGIF-DB was applied to Hi-C matrices from all 6 timepoints using the tree structure as in **Figure S4**, and significant boundaries from each timepoint were used for enrichment analysis. TADCompare was applied to each pair of consecutive timepoints: day 0 vs 2, 2 vs 5, 5 vs 7, 7 vs 15, 15 vs 80. As TADCompare outputs both non-differential and differential boundaries for every pairwise comparison, we define a boundary set specific to a timepoint as follows: (1) for day

0, union of differential boundaries in day 0 and non-differential boundaries between day 0 and 2; (2) for day 80, union of differential boundaries in day 80 and non-differential boundaries between day 15 and 80; (3) for all intermediate timepoints t, union of differential boundaries in t, non-differential boundaries between day t and t_{previous}, and non-differential boundaries between day t and $t_{\text{following}}$. The CTCF peak fold enrichment ratio for a given timepoint was calculated as $\frac{q}{\text{M}} / \frac{s}{\text{N}}$, where q is the number of boundaries with at least one CTCF peak, M is the number of boundary regions, s is the number of regions with at least one CTCF peak, and N is the total number of genomic regions.

Benchmarking with downsampled data to assess robustness to depth We downloaded the high-depth Hi-C dataset of GM12878 cell line (Rao et al., 2014) with 4.01 billion total reads from the 4D Nucleome data portal (Reiff et al., 2022; Dekker et al., 2017, **Table S4**). We then subsampled 5, 10, 25, 50% of the reads, generated 10kb-resolution intra-chromosomal Hi-C matrices using Juicer (Durand et al., 2016), and ICE-normalized the intra-chromosomal interaction matrices from each downsampled datase. We calculated Jaccard index by dividing the number of boundaries found at both depths by the number of boundaries identified in either depths for the GM12878 dataset. The higher the Jaccard Index, the fewer the false-positive differences. Three TAD-calling methods, GRiNCH, SpectralTAD, and TopDom were applied individually to 5 datasets: original high-depth GM12878 data and four low-depth GM12878 data downsampled to 5, 10, 25, 50% depths, respectively. TADCompare and TGIF-DB were applied to 4 pairs of datasets, each pair including the original high-depth GM12878 dataset and the downsampled low-depth dataset (e.g., GM12878 data downsampled to 50% depth, **Figure S4**). For TADCompare, the Jaccard index was calculated for each pair of datasets as the ratio of number of non-differential boundaries and the number of differential and non-differential boundaries. Similarly, for TGIF-DB, the Jaccard index was calculated as the ratio of the number of non-significantly differential boundary regions divided by the size of the union of boundary regions from original-depth and subsampled dataset.

Measuring stability of boundary sets across multiple resolutions of input data We used the mouse neural differentiation dataset (Bonev et al., 2017) to assess the stability of boundary sets identified by different TAD boundary identification methods at different resolutions, 10kb, 25kb and 50kb, since this dataset was readily available at these resolutions. We focused our comparisons only for the mouse embryonic stem cell (mESC) time point. The single-task boundary calling methods (GRiNCH, SpectralTAD, TopDom) were applied individually to 10kb, 25kb, and 50kb intra-chromosomal matrices from

mESC. TADCompare was applied to a pair of timepoints including mESC, both at the same resolution: mESC vs neural progenitors (NPC), and mESC vs cortical neurons (CN). In order to find mESC boundaries from the outputs of the pairwise TADCompare comparisons at each resolution, we took the union of non-differential boundaries and differential boundaries enriched in mESC. We applied TGIF-DB to a tree with all three timepoints from mouse neural differentiation dataset at a specific resolution, and took the significant boundaries from mESC (**Figure S4**). This was repeated for each resolution. To allow for comparison of boundaries from different resolutions, we project the higher resolution bins to the coarsest resolution, namely 50kb. For instance, in the 25kb vs 50kb comparison, each 50kb bin is composed of two 25kb bins and is considered to have a boundary if either of the 25kb bins had a boundary. Similarly, for the 10kb vs 50kb comparison, any of the 5 comprising 10kb bins would be used to define a boundary in the 50kb bin spanning them. In the 10kb vs 25kb comparison, if any of the 10kb bins or the 25kb bins have a boundary in the shared 50kb bin, we define the 50kb bin as a boundary. We then measure Jaccard index of boundaries at this lowest resolution.

Comparison of TGIF-DC to existing compartment-calling methods

We compared TGIF-DC to two established methods for calling compartments, i.e. principal component analysis (PCA) based method (Lieberman-Aiden et al., 2009) and Cscore (version 1.1, Zheng and Zheng, 2018), as well as a method designed specifically for differential compartment analysis, dcHiC (version 2.1, Chakraborty et al., 2022). Each method is described in detail in **Supplemental Methods**, **Figure S23**. We applied all four methods to 100kb intra-chromosomal count matrices from H1 hESC cell line. TGIF-DC and dcHiC were additionally applied to 100kb intra-chromosomal count matrices from H1 differentiated to endoderm. Both datasets were downloaded from 4D Nucleome consortium (Reiff et al., 2022; Dekker et al., 2023). To compare the compartment results across the different methods, we measured the Rand index between compartment assignments to each genomic region. To measure the quality of the compartments, we used three well-known cluster quality metrics: Silhouette Index, Calinski-Harabasz Score, and Davies-Bouldin Index, measured on the observed-over-expected (O/E) matrices for each chromosome, as well as the accessibility signal for each 100kb genomic region. The accessibility signal was defined as the mean ATAC-seq reads per basepair. Finally, to compare dcHiC and TGIF-DC for significantly differential compartments between H1 and endoderm, we calculated the log ratio of the accessibility signal and gene expression (from RNA-seq, in TPM) in H1 over that of endoderm for each significantly differential region.

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

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

For every pair of timepoints, we tested the enrichment of these DE genes within regions of interest (**Figure 5A**): (A) regions near (i.e., within 100kb) significantly differential boundaries (sigDB), (B) regions within a TAD with at least one sigDB, and (C) regions within significantly differential compartmental regions (sigDC). For (B), we define all regions bounded within a pair of shared boundaries and containing at least on sigDB within those bounds as belonging to a "TAD with at least one sigDB".

The fold enrichment of DE genes in these regions was computed as $\frac{q}{\mathrm{M}} / \frac{s}{\mathrm{N}}$, where N is number of all regions, $s = |$set of regions with at least one DE gene$|$, $M = |$a subset regions of interest as defined above, e.g., regions near sigDB$|$, $q = |$regions of interest with at least one DE gene, e.g., regions near sigDB with a DE gene$|$. We also performed gene-centric fold enrichment calculations: $\frac{q_g}{\mathrm{M}_g} / \frac{s_g}{\mathrm{N}_g}$, where N_g is total number of genes with expression, $s_g = |$DE genes$|$, $M_g = |$genes overlapping with a regions of interest, e.g., region near sigDB$|$, $q_g = |$DE genes overlapping with a regions of interest$|$. Hypergeometric test was additionally performed to calculate the significance of this fold enrichment value for each pair of timepoints. We additionally examined the correlation between a gene's RNA-seq fold change and the raw boundary score change (positive or negative) of its nearest sigDB in the H1-endoderm dataset. We found little to no correlation in the magnitude or the direction of change in gene expression and boundary strength (Pearson's corr = 0.03, **Figure S17**).

Gene Ontology (GO) term enrichment analysis was performed for two different subsets of genes based on their DE status and whether they were close to (within 100kb of) sigDB: (1) DE genes not close to a sigDB, (2) DE genes close to a sigDB. The significance of enrichment was determined with an FDR-corrected hypergeometric test p-value $<$0.05. To select candidate differential boundaries for visualization, we ranked a sigDB based on two criteria: (1) adjusted p-value of the change in TGIF-DB boundary score, and (2) the significance of the nearby differential expression measured by the nearest

DE gene's adjusted p-value. We converted these values into ranks and used the mean rank of a boundary to select top 10 regions with promising differential boundaries.

SNP enrichment within TGIF boundaries from cardiomyocyte differentiation data

We downloaded SNPs in the GWAS catalog (Sollis et al., 2023) and mapped each SNP's associated trait to its parent phenotype, based on Experimental Factor Ontology (EFO). We refer to these parent terms as SNP categories in our analysis. In total we had 17 such categories (e.g. cardiovascular disease) for which we tested enrichment of SNPs in TGIF-DB boundaries. For each category, we calculated the fold enrichment of associated SNPs in different subsets of TGIF-DB boundaries across different timepoints: boundaries found in a specific timepoint, boundaries found in the first two or the last two timepoints, boundaries found across all timepoints (ALL, **Figure 7A**), boundaries found in any of the timepoints (ANY, **Figure 7A**). We used the following formula to calculate fold enrichment: $\frac{q}{\mathrm{M}} / \frac{s}{\mathrm{N}}$. Here, q is the number of boundaries of a particular type (e.g. ANY) with at least one SNP of interest, M is the number of boundaries of a particular type (e.g. ANY), s is the number of regions containing at least one SNP, and N is the total number of genomic regions.

Software availability

TGIF-DC and TGIF-DB, along with scripts used for evaluation, analysis, and visualization are available as **Supplemental Code S1, S2, and S3**, at GitHub (`https://github.com/Roy-lab/tgif`), and at Zenodo (`https://doi.org/10.5281/zenodo.13323898`).

Competing Interests

Authors have no competing interests.

Acknowledgements

This work is supported by the National Institutes of Health (NIH) through the grant NIH NHGRI R01-HG010045-01 and by the Computation and Informatics in Biology and Medicine (CIBM) training program (NLM 5T15LM007359). We thank the Center for High Throughput Computing at University of

667 Wisconsin - Madison for computational resources. We also thank Yanxiao Zhang and Bing Ren for
668 providing the list of HERV-H retrotransposon site coordinates and their expression levels.

669 Author contributions

670 Lee and Roy conceptualized the overall framework and algorithm. Lee implemented the algorithm,
671 designed and performed experiments, and wrote the manuscript. Roy designed the experiments and
672 wrote the manuscript.

Figures

Figure 1

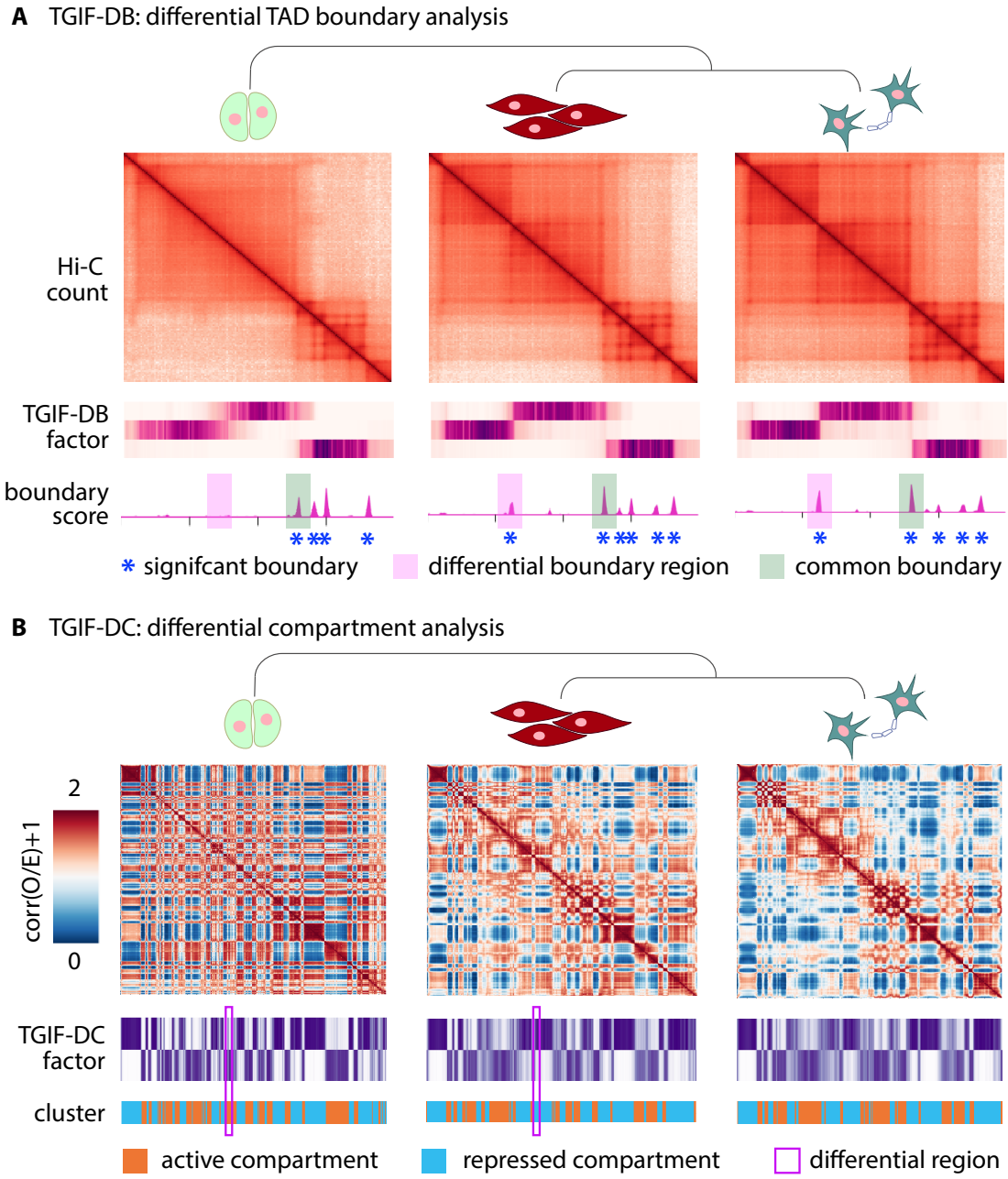
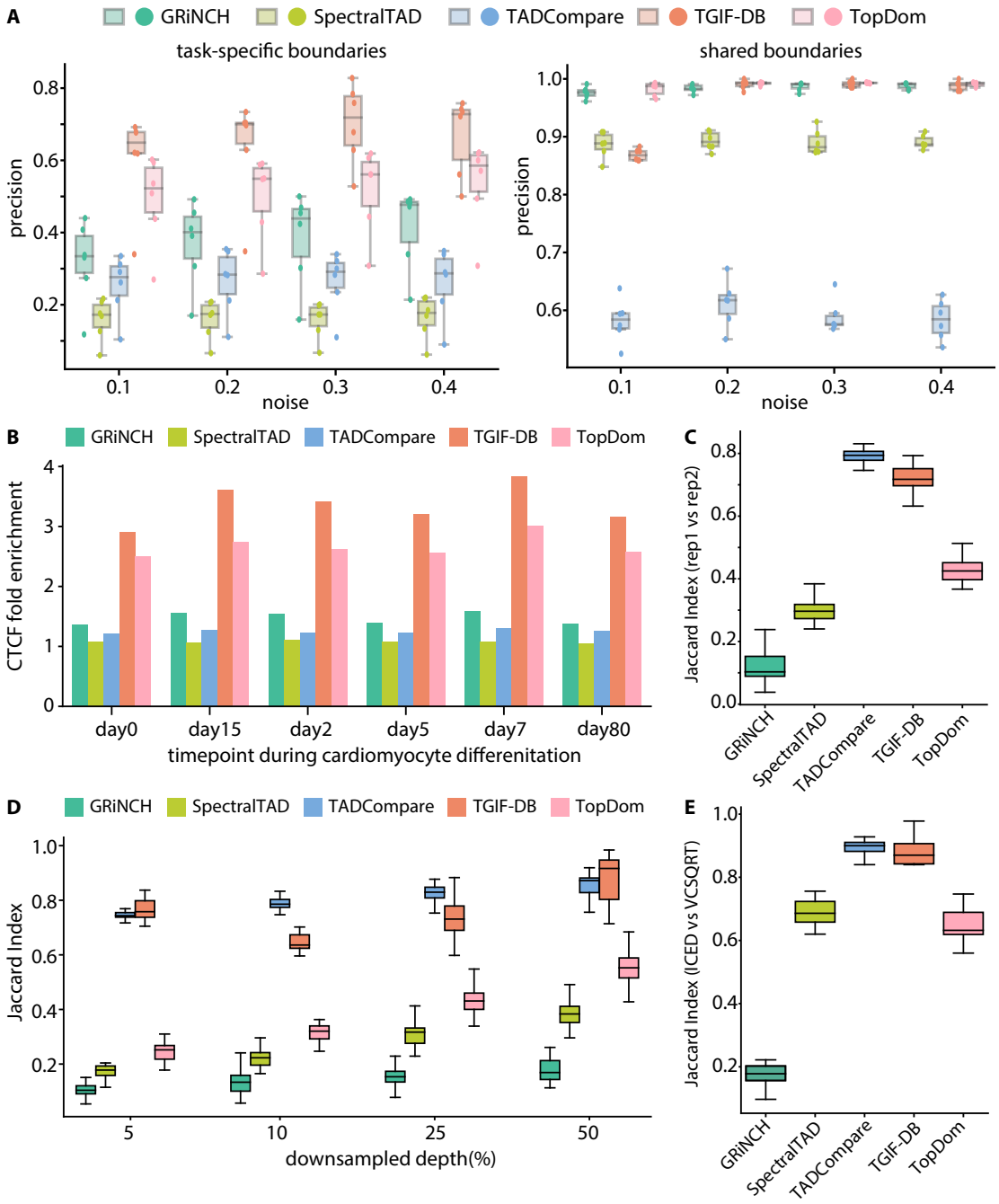


Figure 1. Overview of TGIF. **(A)** TGIF for differential boundary analysis (TGIF-DB). TGIF-DB takes multiple Hi-C count matrices as input and simultaneously learns a lower dimensional representation of genomic regions based on their interaction patterns. The input matrices are from related biological con-

ditions with their relationship encoded as a tree. From the lower-dimensional factors, we measure the boundary score of each region, identify boundaries for each input condition and significantly differential boundaries for every pair of conditions. **(B)** TGIF for differential compartment analysis (TGIF-DC). TGIF-DC converts input matrices into correlation matrices of observed-over-expected (O/E) counts and factorizes them to yield latent features, which are used to cluster the regions. Each cluster correspond to a compartment or a subcompartment. TGIF-DC also identifies significantly differential compartmental regions for every pairs of input conditions.

686 **Figure 2**



687

688 **Figure 2.** Benchmarking TGIF-DB. **(A)** Precision on ground-truth boundaries in simulated Hi-C ma-
689 trices. Each point represents the precision from a pair of input simulated datasets compared to yield
690 task-specific boundaries (i.e. boundaries found in one input dataset but not in the other) and shared
691 boundaries. **(B)** CTCF peak enrichment in boundaries from different TAD-calling and differential-

692 boundary-calling methods. **(C)** Boundary set similarity between biological replicates of hESC (from
693 day 0 of cardiomyocyte differentiation data). **(D)** Boundary set similarity measured by Jaccard index
694 between GM12878 data and downsampled data, across different downsampling depths. **(E)** Boundary
695 set similarity between ICE-normalized and VCSQRT-normalized input matrices of mESC (from mouse
696 neural differentiation data).

Figure 3

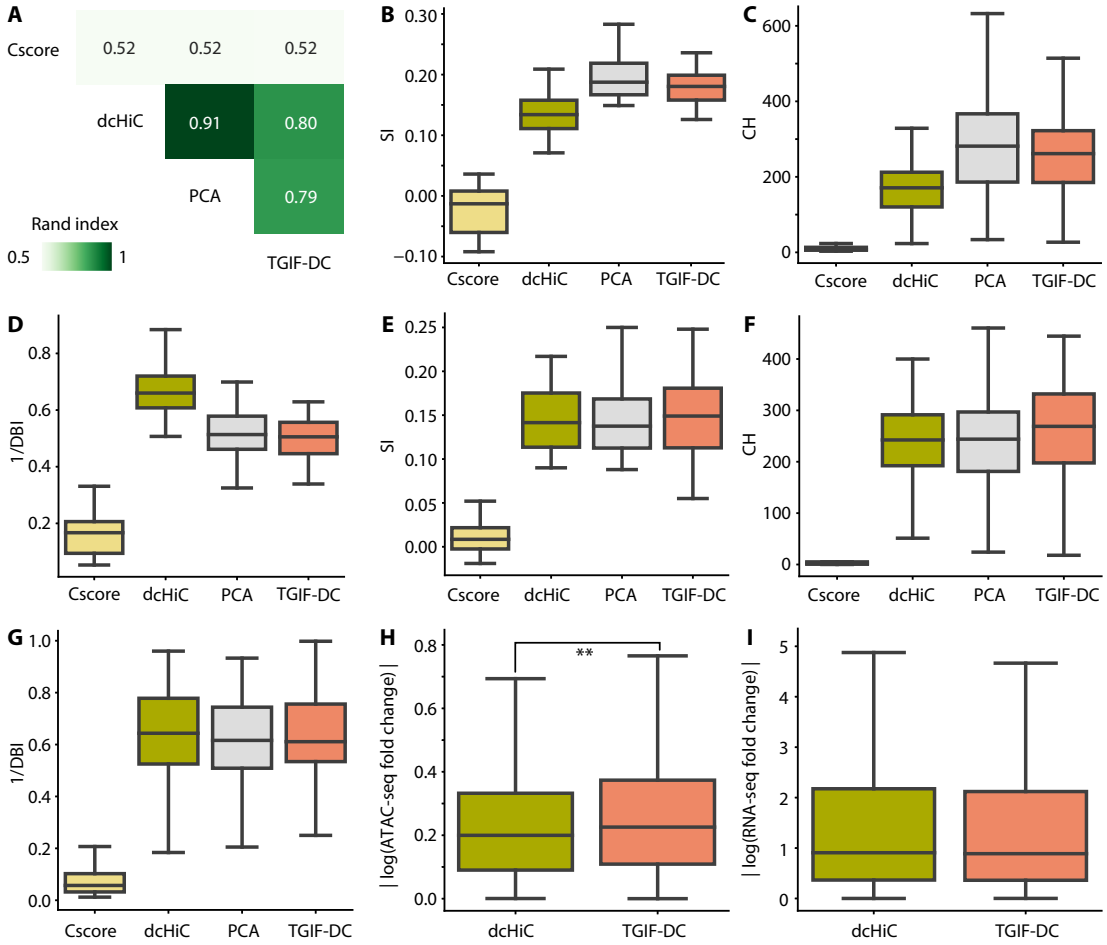


Figure 3. Benchmarking TGIF-DC on data from H1 and H1 differentiated to definitive endoderm. **(A)** Similarity of compartment assignments from different methods measured by Rand Index. **(B)** Quality of compartments based on O/E counts measured by Silhouette Index (SI), **(C)** Calinski-Harabasz score (CH), **(D)** Davies-Bouldin index (DBI). **(E)** SI, **(F)** CH, **(G)** DBI on accessibility (ATAC-seq) signal. **(H)** Magnitude of log fold change in accessibility between H1 and endoderm within significantly differential compartmental regions (sigDC) identified by dcHiC and TGIF-DC. **(I)** Magnitude of log fold change in gene expression between H1 and endoderm within sigDC identified by dcHiC and TGIF-DC.

Figure 4

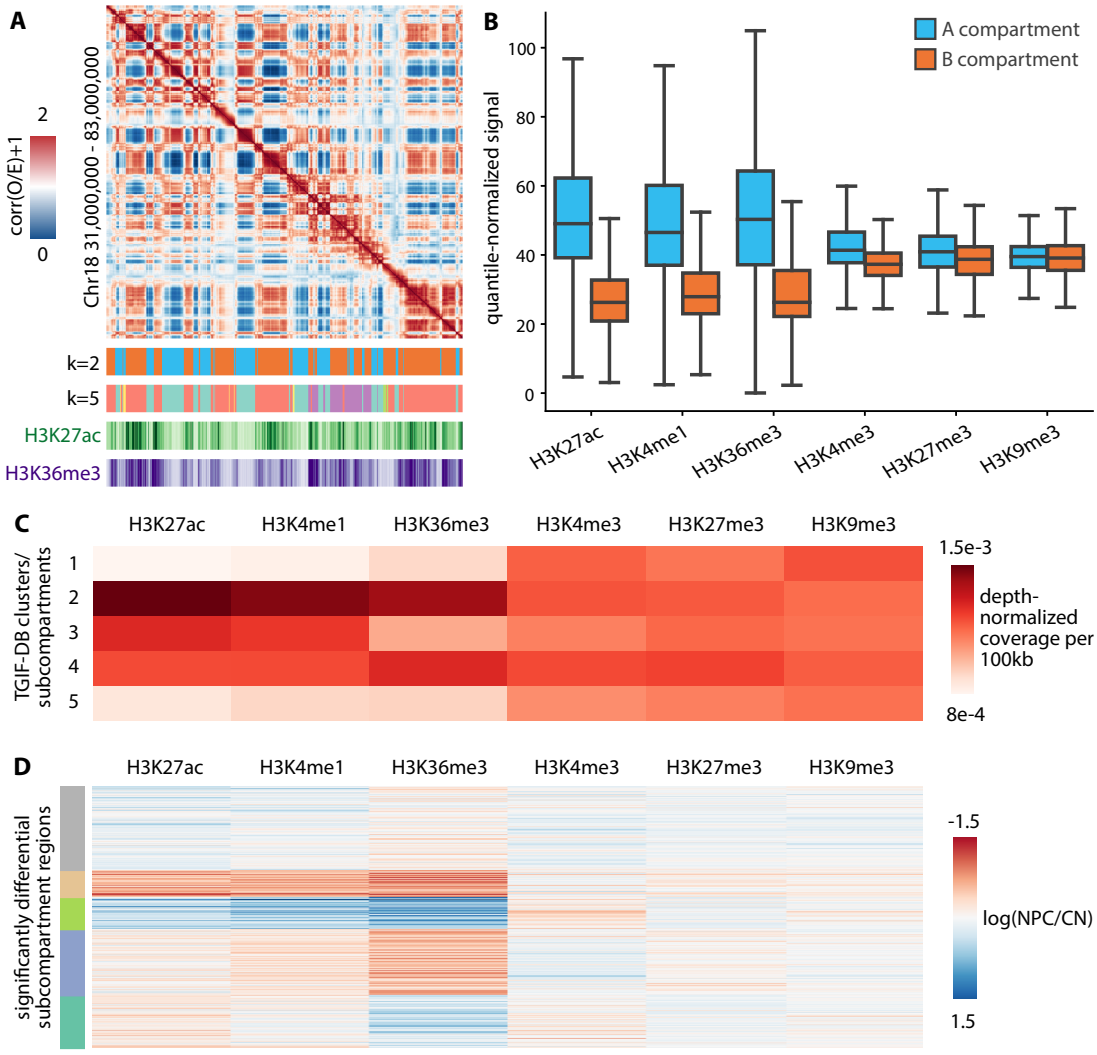


Figure 4. Characterizing compartments and subcompartments identified by TGIF-DC in mouse neural differentiation data. **(A)** A heatmap visualization of correlation matrix of O/E counts from cortical neuron (CN) Chr18 regions at 100kb resolution, followed by TGIF compartment assignments (i.e. clusters from $k = 2$) and subcompartments (e.g. clusters from $k = 5$), and H3K27ac/H3K36me3 ChIP-seq signal heatmaps. **(B)** Distribution of histone modification signal in A and B compartments in neural progenitors (NPC) and CN. **(C)** Mean histone modification signals across different subcompartments in NPC and CN. **(D)** Log fold change of histone modification signals between NPC and CN within significantly differential subcompartment regions identified by TGIF-DC. These regions were grouped based on their histone modification signal fold change patterns using *k*-means clustering and visualized here.

Figure 5

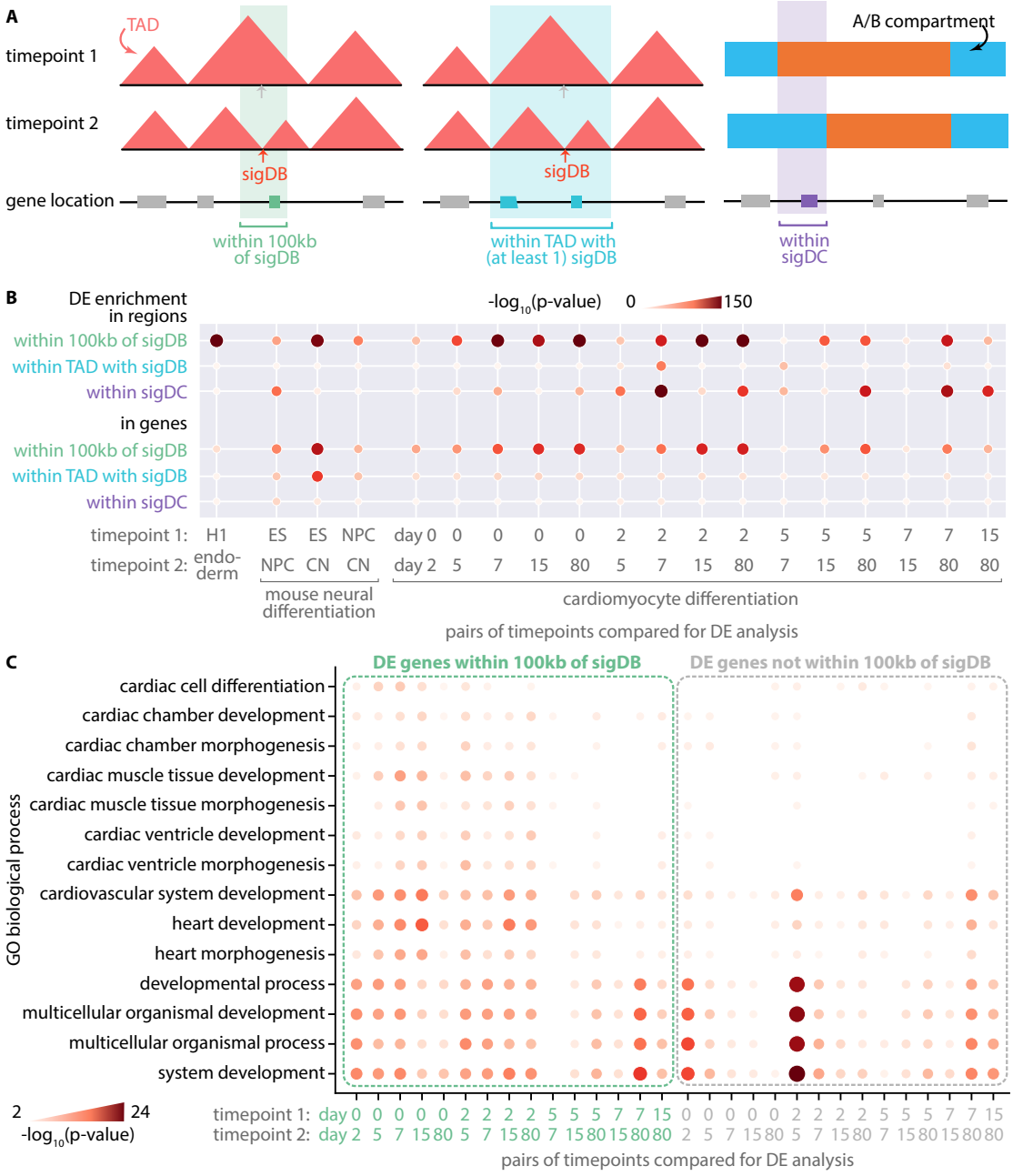


Figure 5. Differential gene expression near or within differential structural features. **(A)** Differential gene expression (DE) enrichment was measured in regions and genes near or within dynamic regions, i.e. regions within 100kb of significantly differential boundary (sigDB), regions within TAD with at least one sigDB, and regions within significantly differential compartmental regions (sigDC). **(B)** DE, sigDB, and sigDC were measured and identified in pairwise comparisons of timepoints across 3 mam-

malian differentiation datasets: H1 differentiated to endoderm, mouse neural differentiation (ES, NPC, CN), and cardiomyocyte differentiation (day 0, 2, 5, 7, 15, 80). Negative log p-value of the enrichment hypergeometric test is visualized here. **(C)** GO biological process enrichment of genes within 100kb of sigDB from cardiomyocyte differentiation data.

Figure 6

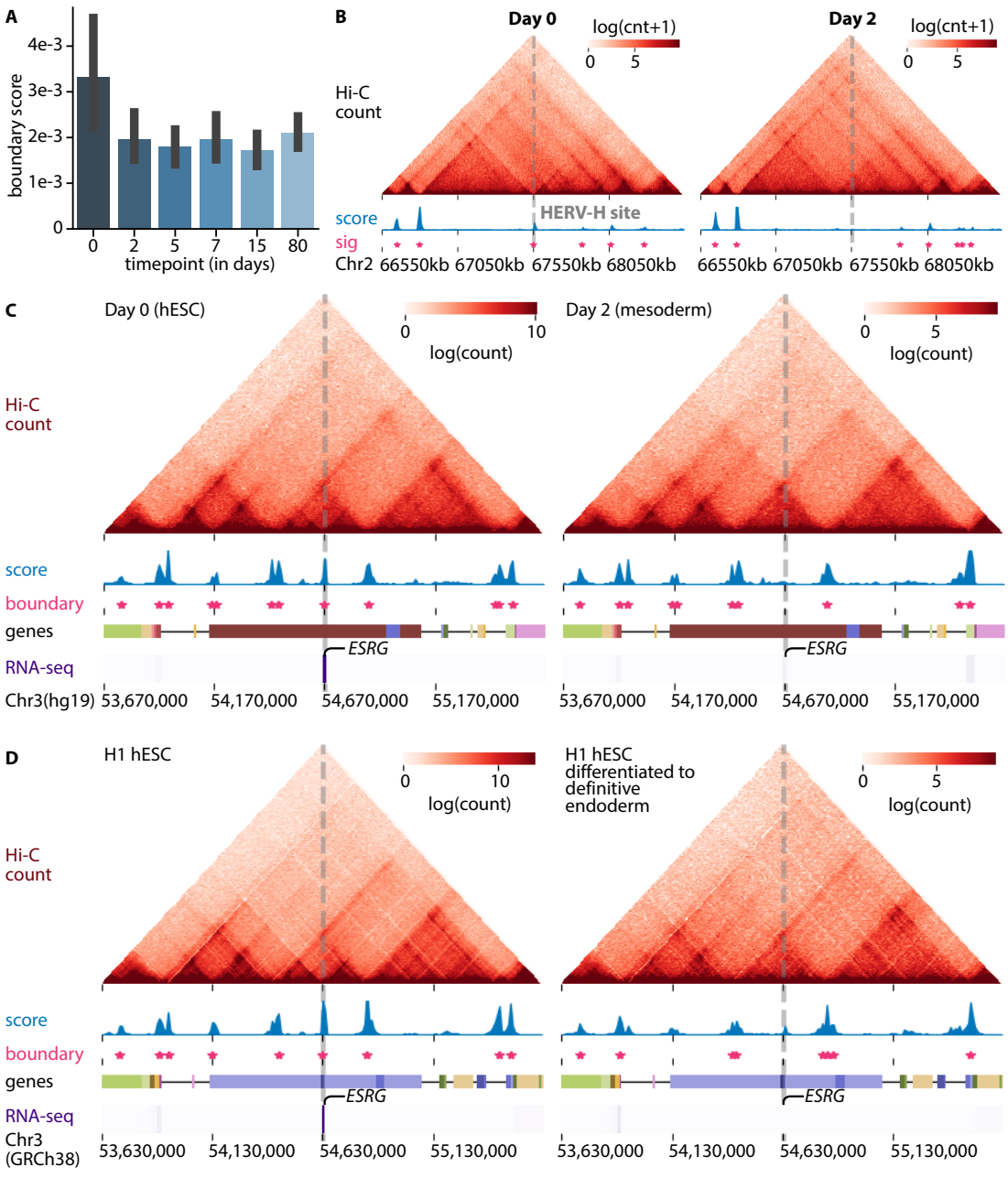


Figure 6. Human pluripotency-specific boundary elements. **(A)** Boundary scores of transcriptionally acive HERV-H retrotransposon sites during each timepoint of cardiomyocyte differentiation. **(B)** The top HERV-H site based on its transcription level in day 0 pluripotent state (within somatic chromosomes) and the overlapping sigDB identified by TGIF-DB. **(C)** *ESRG*, a HERV-H-containing DE gene, overlapping a sigDB in cardiomyocyte differentiation and in **(D)** H1 differentiated to endoderm.

Figure 7

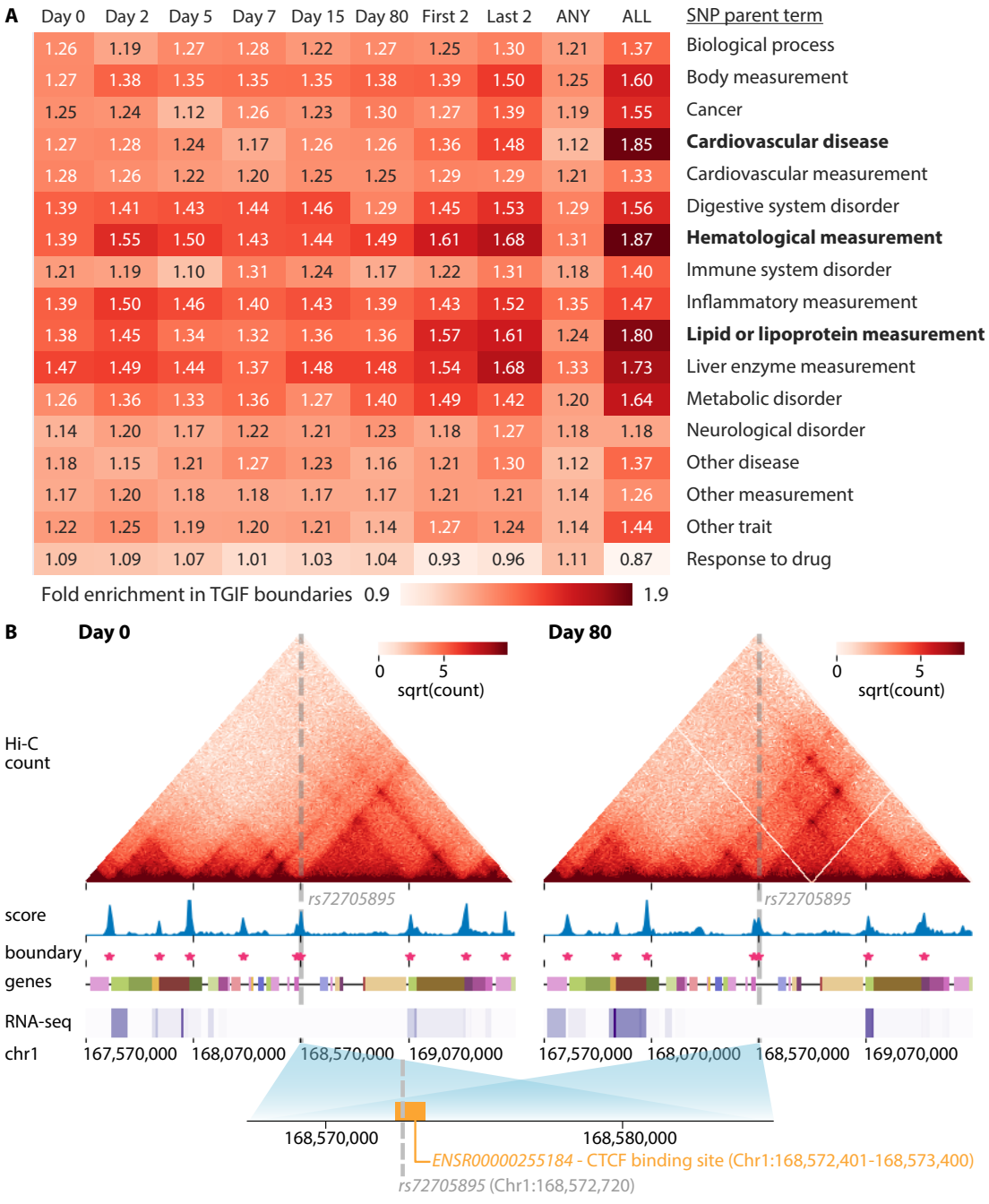


Figure 7. SNP enrichment in persistent boundaries. **(A)** Fold enrichment of SNPs in different subsets of boundary regions, across different categories (SNP parent terms). We measured the enrichment of SNPs in timepoint-specific boundaries (day 0-80) of cardiomyocyte differentiation, in boundaries common to the first 2 or the last 2 timepoints, in union of boundaries (ANY), and in intersection of boundaries across

741 all timepoints (ALL). **(B)** A SNP landing in a boundary persistent across all timepoints (only day 0 and
742 day 80 visualized here) and a CTCF binding site.

References

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

Bair E. 2013. Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **5**: 349–361.

Baur B, Lee DI, Haag J, Chasman D, Gould M, and Roy S. 2022. Deciphering the Role of 3D Genome Organization in Breast Cancer Susceptibility. *Frontiers in Genetics* **12**.

Benjamini Y and Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289–300. Publisher: [Royal Statistical Society, Wiley].

Bondell HD and Reich BJ. 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64**: 115–123.

Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, et al.. 2017. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**: 557–572.e24.

Bouwman BAM and de Laat W. 2015. Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biology* **16**: 154–9.

Cavalheiro GR, Pollex T, and Furlong EE. 2021. To loop or not to loop: what is the role of TADs in enhancer function and gene regulation? *Current Opinion in Genetics & Development* **67**: 119–129.

Chakraborty A and Ay F. 2018. The role of 3d genome organization in disease: From compartments to single nucleotides. *Seminars in Cell & Developmental Biology* .

Chakraborty A, Wang JG, and Ay F. 2022. dcHiC detects differential compartments across multiple Hi-C datasets. *Nature Communications* **13**: 6827. Number: 1 Publisher: Nature Publishing Group.

Chen Z, Snetkova V, Bower G, Jacinto S, Clock B, Dizehchi A, Barozzi I, Mannion BJ, Alcaina-Caro A, Lopez-Rios J, et al.. 2024. Increased enhancer–promoter interactions during developmental enhancer activation in mammals. *Nature Genetics* **56**: 675–685. Publisher: Nature Publishing Group.

Ching YH, Ghosh TK, Cross SJ, Packham EA, Honeyman L, Loughna S, Robinson TE, Dearlove AM, Ribas G, Bonser AJ, et al.. 2005. Mutation in myosin heavy chain 6 causes atrial septal defect. *Nature Genetics* **37**: 423–428. Number: 4 Publisher: Nature Publishing Group.

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

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

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

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

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

Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O'Shea CC, Park PJ, Ren B, et al.. 2017. The 4D nucleome project. *Nature* **549**: 219–226. Number: 7671 Publisher: Nature Publishing Group.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, and Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–380.

Djekidel MN, Chen Y, and Zhang MQ. 2018. FIND: difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Research* **28**: 412–422.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, and Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**: 95–98. Publisher: Elsevier.

Emerson DJ, Zhao PA, Cook AL, Barnett RJ, Klein KN, Saulebekova D, Ge C, Zhou L, Simandi Z, Minsk MK, et al.. 2022. Cohesin-mediated loop anchors confine the locations of human replication origins. *Nature* **606**: 812–819. Number: 7915 Publisher: Nature Publishing Group.

Eres IE, Luo K, Hsiao CJ, Blake LE, and Gilad Y. 2019. Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLOS Genetics* **15**: e1008278.

Espinola SM, Götz M, Bellec M, Messina O, Fiche JB, Houbron C, Dejean M, Reim I, Cardozo Gizzi AM, Lagha M, et al.. 2021. Cis -regulatory chromatin loops arise before TADs and gene activation, and are independent of cell fate during early Drosophila development. *Nature Genetics* **53**: 477–486. Number: 4 Publisher: Nature Publishing Group.

Fletez-Brant K, Qiu Y, Gorkin DU, Hu M, and Hansen KD. 2024. Removing unwanted variation between samples in Hi-C experiments. *Briefings in Bioinformatics* **25**: bbae217.

Forcato M, Nicoletti C, Pal K, Livi C, Ferrari F, and Bicciato S. 2017. Comparison of computational methods for Hi-C data analysis. *Nature Methods* **14**: 679–685.

Fortin JP and Hansen KD. 2015. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology* **16**: 180.

Fotuhi Siahpirani A, Ay F, and Roy S. 2016. A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions. *Genome Biology* **17**: 114.

Gacita AM, Fullenkamp DE, Ohiri J, Pottinger T, Puckelwartz MJ, Nobrega MA, and McNally EM. 2021. Genetic Variation in Enhancers Modifies Cardiomyopathy Gene Expression and Progression. *Circulation* **143**: 1302–1316. Publisher: American Heart Association.

Galan S, Machnik N, Kruse K, Díaz N, Marti-Renom MA, and Vaquerizas JM. 2020. CHESS enables quantitative comparison of chromatin contact data and automatic feature extraction. *Nature Genetics* **52**: 1247–1255.

Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korbel JO, and Furlong EEM. 2019. Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature Genetics* **51**: 1272–1282. Number: 8 Publisher: Nature Publishing Group.

Greenwald WW, Li H, Benaglio P, Jakubosky D, Matsui H, Schmitt A, Selvaraj S, D'Antonio M, D'Antonio-Chronowska A, Smith EN, et al.. 2019. Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature Communications* **10**: 1054.

Gómez-Díaz E and Corces VG. 2014. Architectural proteins: regulators of 3D genome organization in cell fate. *Trends in Cell Biology* **24**: 703–711.

Hata K, Maeno-Hikichi Y, Yumoto N, Burden SJ, and Landmesser LT. 2018. Distinct Roles of Different Presynaptic and Postsynaptic NCAM Isoforms in Early Motoneuron-Myotube Interactions Required for Functional Synapse Formation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **38**: 498–510.

Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, Rialdi A, White KM, Albrecht RA, Pache L, et al.. 2018. Transcription Elongation Can Affect Genome 3D Structure. *Cell* **174**: 1522–1536.e22.

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

Ing-Simmons E, Vaid R, Bing XY, Levine M, Mannervik M, and Vaquerizas JM. 2021. Independence of chromatin conformation and gene regulation during Drosophila dorsoventral patterning. *Nature Genetics* **53**: 487–499. Number: 4 Publisher: Nature Publishing Group.

Kempfer R and Pombo A. 2019. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics* .

Kempfer R and Pombo A. 2020. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics* **21**: 207–226. Number: 4 Publisher: Nature Publishing Group.

Kim J, He Y, and Park H. 2014. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization* **58**: 285–319.

Kobets VA, Ulianov SV, Galitsyna AA, Doronin SA, Mikhaleva EA, Gelfand MS, Shevelyov YY, Razin SV, and Khrameeva EE. 2023. HiConfidence: a novel approach uncovering the biological signal in Hi-C data affected by technical biases. *Briefings in Bioinformatics* **24**: bbad044.

Koitsopoulos PG and Rabkin SW. 2021. The association of polymorphism in PHACTR1 rs9349379 and rs12526453 with coronary artery atherosclerosis or coronary artery calcification. A systematic review. *Coronary Artery Disease* **32**: 448–458.

Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, and Sabeti PC. 2019. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**: e43803. Publisher: eLife Sciences Publications, Ltd.

Kriebel AR and Welch JD. 2022. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nature Communications* **13**: 780. Publisher: Nature Publishing Group.

Kruse K, Hug CB, and Vaquerizas JM. 2020. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biology* **21**: 303.

Kuveljic J, Djuric T, Stankovic G, Dekleva M, Stankovic A, Alavantic D, and Zivkovic M. 2021. Association of PHACTR1 intronic variants with the first myocardial infarction and their effect on PHACTR1 mRNA expression in PBMCs. *Gene* **775**: 145428.

Lawson HA, Liang Y, and Wang T. 2023. Transposable elements in mammalian chromatin organization. *Nature Reviews Genetics* pp. 1–12. Publisher: Nature Publishing Group.

Lee DD and Seung HS. 2000. Algorithms for Non-negative Matrix Factorization. In *In NIPS*, volume 13, pp. 556–562.

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

Li B and Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Li X, Zeng G, Li A, and Zhang Z. 2021. DeTOKI identifies and characterizes the dynamics of chromatin TAD-like domains in a single cell. *Genome Biology* **22**: 217.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al.. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**: 289–293.

Lindström S, Wang L, Smith EN, Gordon W, van Hylckama Vlieg A, de Andrade M, Brody JA, Pattee JW, Haessler J, Brumpton BM, et al.. 2019. Genomic and transcriptomic association studies identify 16 novel susceptibility loci for venous thromboembolism. *Blood* **134**: 1645–1657.

Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, and Welch JD. 2020. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nature Protocols* **15**: 3632–3662. Publisher: Nature Publishing Group.

Liu J, Wang C, Gao J, and Han J. 2013. Multi-View Clustering via Joint Nonnegative Matrix Factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (eds. J Ghosh, Z Obradovic, J Dy, ZH Zhou, C Kamath, and S Parthasarathy), pp. 252–260. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Love MI, Huber W, and Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 550.

Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, et al.. 2022. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods* **19**: 41–50. Publisher: Nature Publishing Group.

Lun AT and Smyth GK. 2015. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**: 258.

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, et al.. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**: 1012–1025.

Lupiáñez DG, Spielmann M, and Mundlos S. 2016. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics* **32**: 225–237.

McArthur E and Capra JA. 2021. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *American Journal of Human Genetics* **108**: 269–283.

McCord R. 2017. Chromosome biology: How to build a cohesive genome in 3D. *Nature* .

Merkenschlager M and Nora EP. 2016. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annual Review of Genomics and Human Genetics* **17**: 17–43.

Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, and Chang HY. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* **13**: 919–922.

Norton HK and Phillips-Cremins JE. 2017. Crossed wires: 3D genome misfolding in human disease. *Journal of Cell Biology* **216**: 3441–3452.

Orozco G, Schoenfelder S, Walker N, Eyre S, and Fraser P. 2022. 3D genome organization links non-coding disease-associated variants to genes. *Frontiers in Cell and Developmental Biology* **10**.

Pollex T, Rabinowitz A, Gambetta MC, Marco-Ferreres R, Viales RR, Jankowski A, Schaub C, and Furlong EEM. 2024. Enhancer–promoter interactions become more instructive in the transition from cell-fate specification to tissue differentiation. *Nature Genetics* **56**: 686–696. Publisher: Nature Publishing Group.

Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al.. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–1680.

Reiff SB, Schroeder AJ, Kırlı K, Cosolo A, Bakker C, Mercado L, Lee S, Veit AD, Balashov AK, Vitzthum C, et al.. 2022. The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nature Communications* **13**: 2365. Number: 1 Publisher: Nature Publishing Group.

Rowley MJ and Corces VG. 2018. Organizational principles of 3d genome architecture. *Nature Reviews Genetics* p. 1.

Roy AL, Conroy RS, Taylor VG, Mietz J, Fingerman IM, Pazin MJ, Smith P, Hutter CM, Singer DS, and Wilder EL. 2023. Elucidating the structure and function of the nucleus—The NIH Common Fund 4D Nucleome program. *Molecular Cell* **83**: 335–342. Publisher: Elsevier.

Shetty A, Sytnyk V, Leshchyns'ka I, Puchkov D, Haucke V, and Schachner M. 2013. The neural cell adhesion molecule promotes maturation of the presynaptic endocytotic machinery by switching synaptic vesicle recycling from adaptor protein 3 (AP-3)- to AP-2-dependent mechanisms. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **33**: 16828–16845.

Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, and Zhou XJ. 2016. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research* **44**: e70.

Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al.. 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* **51**: D977–D985.

Stadhouders R, Vidal E, Serra F, Di Stefano B, Le Dily F, Quilez J, Gomez A, Collombet S, Berenguer C, Cuartero Y, et al.. 2018. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics* **50**: 238–249.

Stansfield JC, Cresswell KG, and Dozmorov MG. 2019. multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics* **35**: 2916–2923.

van Steensel B and Furlong EEM. 2019. The role of transcription in shaping the spatial organization of the genome. *Nature Reviews Molecular Cell Biology* **20**: 327–337. Number: 6 Publisher: Nature Publishing Group.

Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y, Ngom A, Ochs MF, et al.. 2018. Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends in Genetics* **34**: 790–805. Publisher: Elsevier.

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, and Hadjur S. 2015. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell reports* .

Wang G, Meng Q, Xia B, Zhang S, Lv J, Zhao D, Li Y, Wang X, Zhang L, Cooke JP, et al.. 2020. TADsplimer reveals splits and mergers of topologically associating domains for epigenetic regulation of transcription. *Genome Biology* **21**: 84.

Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, et al.. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516**: 405–409. Number: 7531 Publisher: Nature Publishing Group.

Wang R, Lee JH, Xiong F, Kim J, Hasani LA, Yuan X, Shivshankar P, Krakowiak J, Qi C, Wang Y, et al.. 2021. SARS-CoV-2 Restructures the Host Chromatin Architecture. Pages: 2021.07.20.453146 Section: New Results.

Wang W, Chandra A, Goldman N, Yoon S, Ferrari EK, Nguyen SC, Joyce EF, and Vahedi G. 2022. TCF-1 promotes chromatin interactions across topologically associating domains in T cell progenitors. *Nature Immunology* **23**: 1052–1062. Number: 7 Publisher: Nature Publishing Group.

Wanggou S, Jiang X, Li Q, Zhang L, Liu D, Li G, Feng X, Liu W, Zhu B, Huang W, et al.. 2012. HESRG: a novel biomarker for intracranial germinoma and embryonal carcinoma. *Journal of Neuro-Oncology* **106**: 251–259.

Warkman AS, Whitman SA, Miller MK, Garriock RJ, Schwach CM, Gregorio CC, and Krieg PA. 2012. Developmental expression and cardiac transcriptional regulation of Myh7b, a third myosin heavy chain in the vertebrate heart. *Cytoskeleton (Hoboken, N.J.)* **69**: 324–335.

Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, and Macosko EZ. 2019. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**: 1873–1887.e17.

Xiong K and Ma J. 2019. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nature Communications* **10**: 5069. Publisher: Nature Publishing Group.

Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, Yue F, and Li Q. 2017. HiCRep: assessing the reproducibility of Hi-C data using a stratum- adjusted correlation coefficient. *Genome Research* p. gr.220640.117.

Zerbino DR, Wilder SP, Johnson N, Juettemann T, and Flicek PR. 2015. The Ensembl Regulatory Build. *Genome Biology* **16**: 56.

Zhang R, Zhou T, and Ma J. 2022. Ultrafast and interpretable single-cell 3D genome analysis with Fast-Higashi. *Cell Systems* **13**: 798–807.e6.

Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, et al.. 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics* **51**: 1380–1388. Number: 9 Publisher: Nature Publishing Group.

Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, Nussbaum C, Myers R, Brown M, Li W, et al.. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.

Zheng H and Xie W. 2019. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology* **20**: 535–550. Number: 9 Publisher: Nature Publishing Group.

Zheng X and Zheng Y. 2018. CscoreTool: fast Hi-C compartment analysis at high resolution. *Bioinformatics* **34**: 1568–1570.

Zheng Y, Shen S, and Keleş S. 2022. Normalization and de-noising of single-cell Hi-C data with BandNorm and scVI-3D. *Genome Biology* **23**: 222.

Zhou J, Ma J, Chen Y, Cheng C, Bao B, Peng J, Sejnowski TJ, Dixon JR, and Ecker JR. 2019. Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation. *Proceedings of the National Academy of Sciences* **116**: 14011–14018. Publisher: Proceedings of the National Academy of Sciences.

Examining dynamics of three-dimensional genome organization with multitask matrix factorization

Da-Inn Lee and Sushmita Roy

Genome Res. published online March 20, 2025
Access the most recent version at doi:10.1101/gr.279930.124