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. The availability of high-throughput chromosome conformation datasets, such as Hi-C, across multiple conditions and time points offer 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 is 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 applied to time series or hierarchically related biological conditions. TGIF can identify large-scale compartmental changes as well as smaller topologically associated domain-level changes. Compared to existing methods, TGIF identifies has fewer false positive TAD boundary changes. Application to two mammalian developmental time courses provides multi-scale characterization of genome dynamics that we validate with enrichment of one-dimensional regulatory signals from histone modifications, accessibility and architectural proteins. 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 determines the packaging of DNA inside the nucleus and 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 (Eres et al., 2019; McCord, 2017). 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 (Dekker et al., 2023; Kempfer and Pombo, 2019, 2020; Mumbach et al., 2016). 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; Kempfer and Pombo, 2020; Rowley and Corces, 2018). Changes in 3D genome organization at different topological levels have been observed with transitions in both normal (Bonev et al., 2017; Stadhouders et al., 2018; Zheng and Xie, 2019) and disease processes (Lupiáñez et al., 2016; Norton and Phillips-Cremins, 2017; Wang et al., 2021). For example, during differentiation of mouse embryonic stem cells to a neuronal lineage, changes in topological structure are associated with cell fate specification and gene expression changes (Bonev et al., 2017). Changes in 3D genome have also been seen in immune response to viral infections (Wang et al., 2021) and diseases such as cancer (Akdemir et al., 2020; Dubois et al., 2022; Hnisz et al., 2016). 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., 2023, 2017; Roy et al., 2023). These datasets provide a unique opportunity to examine 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 (Chakraborty et al., 2022; Fotuhi Siahpirani et al., 2016), those that can identify TAD-scale changes or "differential TADs" (Cresswell and Dozmorov, 2020; Wang et al., 2020), and those that examine changes at the level of loops (Ardakany et al., 2019; Djekidel et al., 2018; Galan et al., 2020; Lun and Smyth, 2015; Stansfield et al., 2019). Compared to methods for detecting difference at the loop level, there are relatively few approaches that can be used 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 con-

dition followed by post-processing to identify TAD boundaries in one condition but not another (Bonev et al., 2017; Emerson et al., 2022; Stadhouders et al., 2018; Wang et al., 2022; Zhang et al., 2019). Such two-step approaches may have high false positive rates for detecting true 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) and do not extend to large number of conditions which might have arbitrary relationships among the conditions (Cresswell and Dozmorov, 2020).

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

We evaluate TGIF against other methods on real and simulated multi-condition Hi-C datasets. Compared to existing approaches for identifying TAD dynamics, TGIF recovers ground-truth differential TAD boundaries with higher precision in simulated data and is more robust to calling false positive boundary changes arising due to differences in depth. Application of TGIF to human cardiomyocyte differentiation (Zhang et al., 2019) and mouse neural differentiation data (Bonev et al., 2017) showed that cell-stage-specific TAD boundaries are enriched for known and novel boundary elements, and that differentially expressed genes localize near differential boundaries. At the compartment level, TGIF identified time-point-specific association of regions in the active ("A") and repressive ("B") compartments to active and repressive epigenetic signals respectively. Finally, we use TGIF boundaries identified in cardiomyocyte differentiation dataset to interpret sequence variants identified from Genome Wide Association Studies (GWAS). Together, these results show the versatility and utility of TGIF to examine changes in 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 1A**). 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 1B**, **Supp Figure 1**, **Methods**). In TGIF-DB, the factorization is performed on submatrices along the diagonal of the intrachromosomal Hi-C matrices, as these diagonal sub-matrices capture the TAD-scale, local topology of chromosomes. Each submatrix is factorized over a range of k, the hyperparameter specifying the rank of the lower-dimensional space (**Supp Figure 1B**). TGIF-DB calculates a boundary score from the factors at each k, which are averaged to provide an overall boundary score (**Supp Figure 1C**). Considering multiple k allows us to capture structural units or domains of different sizes in the lower dimensional space and removes the need for the user to specify the number of factors. Finally, TGIF-DB identifies regions with significant boundary scores by comparing the boundary scores averaged across multiple ks against a "null distribution" of boundary scores which allows us to calculate an empirical p-value (**Supp Figure 1D**). The significant boundaries across each of tasks are annotated or categorized into task-specific and shared boundaries based on the pattern of conservation across the different tasks (**Supp Figure 1E**).

TGIF-DC is used to identify compartment-level changes in 3D genome organization (**Figure 1C**). 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, TGIF-DB factorizes the processed intrachromosomal matrices with parameter $k = 2$. The resulting factors are used to group the genomic regions into 2 different clusters. We further interpret these clusters based on their association with histone modification signals. From these clusters, we can examine global compartment-level changes as well as

specific regions exhibiting compartment changes across tasks.

TGIF-DB identifies fewer false positive differential boundaries in simulated and downsampled data

We first examined the performance of TGIF-DB for identifying conserved and differential TAD boundaries. We first simulated Hi-C datasets with known TAD boundaries to benchmark TGIF-DB with two other methods: TADCompare, a differential TAD identification method for pairs or sequence of datasets (Cresswell and Dozmorov, 2020), and TopDom (Shin et al., 2016), a baseline "single-task" TAD-calling method shown to identify high quality boundaries (Dali and Blanchette, 2017) while being robust to sparsity of input data (Lee and Roy, 2021). Following the simulation procedure from Forcato et al., 2017, we generated 4 Hi-C matrices. Each pair of matrices had 20 to 78 known differential boundaries between them, i.e., boundaries present in one matrix but not in both, which we considered as ground truth differences (**Supp Figure 2**, **Methods**). We applied TopDom individually to each simulated matrix and then performed pairwise comparison to identify shared and differential boundaries while TADCompare was applied for each pair of datasets. TGIF was applied to all 4 datasets together (**Methods**).

We measured precision, recall, and f-score of the predicted shared and differential boundaries from each of the methods against ground truth from the simulated data. TGIF had the highest precision for detecting differential TAD boundaries for every pair of simulated datasets, yielding fewer false positives compared to the other methods (**Figure 2A**, **Supp Figure 2**). Although TopDom (independent TAD calling method) had high recall, it had much lower precision resulting in an overall lower f-score than TGIF. On shared boundaries, TGIF and TopDom had comparable precision, recall and f-score. TADCompare generally had much lower precision and f-score than either TGIF or TopDom for detecting shared and differential boundaries.

We also examined the robustness of these methods to low sequencing depth. We took the high-depth Hi-C dataset from the GM12878 cell line (Rao et al., 2014) and downsampled it to four lower-depth datasets corresponding to the other cell lines (K562, HUVEC, HMEC, NHEK) using an approach described previously (Lee and Roy, 2021). As these datasets represent the same cell line, any differential boundaries can be considered as false positives resulting from the depth difference. We applied TGIF, TADCompare, and TopDom on each pair of high-depth and low-depth datasets and measured the Jaccard coefficient between the boundaries identified for the two datasets; the higher the Jaccard Index, the fewer the false-positive differences identified by a method. In three of the four pairwise comparisons between the original high-depth data and the downsampled counterpart, TGIF yields the most consistent

boundary sets with highest Jaccard coefficients (**Figure 2B**). The only exception was with GM12878 dataset downsampled to HMEC depth, for which all three methods perform similarly. Taken together, these results show that TGIF accurately detects differential boundaries in both simulated and real Hi-C datasets and is less susceptible to detecting spurious differential boundaries due to depth differences.

Characterizing boundary and compartment dynamics during cardiomyocyte differentiation with TGIF

During developmental processes such as cellular differentiation, the 3D genome undergoes topological changes, which can impact downstream gene expression programs (Hug and Vaquerizas, 2018; Pelham-Webb et al., 2020; Zheng and Xie, 2019). To examine the extent of change in 3D genome organization during a developmental process, we applied TGIF-DB to a human cardiomyocyte differentiation time-course dataset, with Hi-C interaction counts measured at 6 different timepoints from human embryonic stem cell to ventricular cardiomyocyte at day 0, 2, 5, 7, 15, 80 (Zhang et al., 2019, **Figure 3A**). TGIF identified 21592 boundaries across the 6 timepoints, 2408 of which are shared in all 6 timepoints, which we refer to as "persistent" boundaries (**Figure 3B**). Within each timepoint, TGIF found 8000 to 10000 boundaries, of which about a tenth are unique to each timepoint (**Figure 3C**). We quantified the extent of change at the TAD boundary level by computing the Jaccard coefficient for each pair of time points (**Figure 3D**). Immediately neighboring timepoints share larger proportions of their boundaries as detected by a higher Jaccard coefficient (**Figure 3D**). The boundary set from day 80, when cells are fully differentiated to ventricular cardiomyoctyes, were the least similar to those from the preceding timepoints, which in turn were sampled more closely over time.

We next examined changes in 3D genome organization at the coarser compartment level using TGIF-DC. Briefly, we applied TGIF-DC to 100kb-sized regions to assign them into one of two compartments (**Methods**). We compared the change in compartment organization across time using Rand index and found a similar trend as with our boundary analysis, with neighboring time points exhibiting the greatest similarity and increasingly different compartment structure as a function of the difference in time (**Figure 3E**). However, unlike the boundary analysis which showed that the Day 80 time point was substantially different from the other time points, the compartment analysis exhibited a more gradual change. Out of 27979 100kb-sized regions, 20102 regions (72%) had persistent compartment assignments throughout the 6 timepoints; the other 7877 regions changed their compartment assignments at least once during the time course. We validated the compartment assignments from TGIF-DC in two ways: (1) similarity to compartments based on the original principal components analysis (PCA)-based method by Lieberman-

Aiden et al., 2009, and (2) H3K27ac and accessibility signal of regions assigned to different compartments by TGIF-DC (**Methods**). We found high similarity between the compartments from TGIF-DC and the PCA method, measured by Rand index across chromosomes and time points (mean Rand index of 0.848 ± 0.110, **Supp Figure 3A**). Examining the TGIF-DC inferred compartment structure for a specific region of chr10 showed high concordance between the compartment assignment and both accessibility and H3K27ac levels (**Figure 3F**). Accordingly, we labeled the compartment with high accessibility and H3K27ac as the active "A" compartment and the other compartment as the repressive "B" compartment. Examining these levels more globally, showed that nearly all chromosomes, except chr7 and chr9, exhibited a dramatic shift in accessibility with regions in the A compartment exhibiting nearly twice as high accessibility compared to the other one (**Figure 3G**, **Supp Figure 4**). Similarly, TGIF-DC's A compartment exhibited much higher H3K27ac levels for the same chromosomes across all timepoints compared to the other compartment (**Figure 3H**, **Supp Figure 5**). Interestingly, the compartment that exhibited high accessibility also had higher mean observed over expected (O/E) interaction counts, with the exception of chr7 and chr9 (**Supp Figure 4**) Thus, the mean O/E Hi-C counts of the compartments could be another indicator of A and B compartments in case other one-dimensional epigenetic signals are not available.

Examining regulatory signatures of dynamic boundaries during cardiomyocyte differentiation

We next characterized the persistent and dynamic boundaries identified by TGIF-DB during cardiomyocyte differentiation by overlapping these boundaries with two different types of genomic elements: retrotransposons and transcription factor (TF) binding sites. Retrotransposons are a type of transposable element, which have been implicated in chromatin organization (Lawson et al., 2023). In particular, Zhang et al., 2019, found that the human endogenous retrovirus subfamily retrotransposon (HERV-H) is a major determinant of TAD boundaries specific to hESC (i.e., day 0 of differentiation) when transcriptionally active. We obtained genomic regions with the top 100 most transcriptionally active HERV-H sites and aggregated their TGIF-DB boundary scores at each timepoint. The boundary score at these transcriptionally active HERV-H sites is highest in the hESC (day 0) timepoint than in any subsequent timepoints (**Figure 4A**). This is further supported by the presence of a boundary unique to day 0 that disappears in subsequent timepointsat one of the top transcriptionally active HERV-H sites (**Figure 4B**). This suggests that TGIF-DC can successfully identify biologically meaningful context-specific boundaries.

In addition to retrotransposons, TAD boundaries are often demarcated by architectural proteins such as CTCF, RAD21 and YY1 (Szabo et al., 2019). However, a comprehensive catalog of architectural proteins and their interacting partners, especially those contributing to dynamic 3D genome organization, is not yet known. Hence we set out to study enrichment of TF binding sites in dynamic TAD boundaries, specifically those that are unique to each timepoint. We downloaded the human genome (hg19) motif match sites for 745 TFs from JASPAR (Fornes et al., 2020), and found accessible motif sites by intersecting them with DNase-seq broadpeaks from hESC, specifically the H9 cell line (ENCODE Project Consortium, 2012; Luo et al., 2020). We measured the fold enrichment of these accessible motif sites in TGIF-DB boundaries found in hESC (day 0) against the background of all genomic regions (**Figure 4C**, **Methods**). One of the most enriched TFs was CTCF (**Figure 4D**), a well-known architectural protein of mammalian TAD boundaries (Rao et al., 2014; Szabo et al., 2019). Among other top TFs, were YY1 and MYC. YY1 is known to guide pluripotency-specific regulatory landscape in tandem with CTCF (Beagan et al., 2017; Dong et al., 2022); MYC is a key TF regulating genes in a number of cell-fate, differentiation, and growth-related process (Amjadi-Moheb et al., 2021), and an oncogene associated with many cancers that can shape the transcriptional landscape through genome-wide chromatin remodeling and decompaction (Amjadi-Moheb et al., 2021; Kieffer-Kwon et al., 2017; Knoepfler et al., 2006).

We repeated this analysis for all six time points but using motif sites overlapping H3K27ac peaks, since accessibility measurements were available only for the hESC (day 0) time point (**Methods**). Compared to the TF rankings based on accessibility for the hESC timepoint, H3K27ac-based rankings identified a different but relevant set of TFs (**Figure 4E**). For example, ASCL1 and MAZ were found among the most significant TFs in day 0 (p-value $<$ 1e-8). ASCL1 is a pioneer factor that can bind nucleosomal DNA to activate downstream gene expression programs and was shown to interact with the SWI/SNF chromatin remodeling complex to change chromosome conformation (Păun et al., 2023). MAZ has been shown to co-bind with CTCF to shape genome structure (Gómez-Díaz and Corces, 2014; Xiao et al., 2021). In the subsequent timepoints, we found similarly biologically relevant TFs with some specialization towards the muscle and cardiac lineage. For example, we found TBX5 (day 2), HES1 (day 5), HEY2 (Day 7, 80), KLF15 (day 15), HEY1 (day 80) all of which have been implicated in various programs of cardiac development and congenital heart defect (Fisch et al., 2007; Fischer et al., 2007; Gibb et al., 2018; Ihara et al., 2018; Miao et al., 2018; Rochais et al., 2009; Steimle and Moskowitz, 2017; Waldron et al., 2016; Zou et al., 2020). Of particular interest is YY2, emerging as one of the most significant TFs enriched boundaries unique to day 15 and 80 (primitive and ventricular cardiomyocyte). YY2 was found to be constitutively expressed in mouse heart cells during embryogenesis into adulthood (Drews

et al., 2009) and drives differentiation of mouse embryonic stem cells into cardiovascular lineages (Tahmasebi et al., 2016); its increased expression is observed in hearts with cardiomyopathy from patients with end-stage heart failure (Zhang et al., 2020). The tendency of enrichment of heart, muscle, and cardiac TFs for the later time point is further supportive of TGIF's ability to identify biologically relevant dynamic boundaries. These regulators could be operating immediately downstream or in concert with architectural proteins such as CTCF and YY1 to drive context-specific expression.

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

Changes in TAD structure have been associated with changes in gene expression, though the magnitude of change and which genes change depends upon the local spatial context (Cavalheiro et al., 2021). The cardiomyocyte differentiation dataset included time series of both Hi-C and gene expression measurements matched for each time point, allowing us to evaluate whether a change in TAD boundary is associated with change in gene expression. We first identified differential boundaries (DB) for each pair of time points from TGIF-DB's outputs, defined as a significant boundary present in but not both timepoints. In parallel, we performed differential gene expression analysis for each pair of time points (**Methods**). We asked whether differentially expressed (DE) genes are enriched within 50kb radius of DB using a hypergeometric test (p-value < 0.05). We found that for all pairs of time points, there is a significant enrichment of DE genes within the 50kb neighborhood of a DB (**Figure 5A, B**). Additionally, the further the timepoints, the more significant is the enrichment (**Figure 5B**).

To examine the biological processes associated with DE genes near DB, we partitioned our genes based on whether they were DE and whether they were near a DB. We tested these gene sets for enrichment of Gene Ontology (GO) biological processes based on FDR-corrected hypergeometric test (**Methods**). At a corrected p-value$<$1e-5, both sets of DE genes (near or not near DB) were enriched for several processes for different pairs of timepoints. However, DE genes near DBs were enriched for more biological processes (**Supp Table 1**). Furthermore, these genes were enriched for terms relevant to cardiac development e.g. heart morphogenesis, cardiovascular system development, cardiac ventricle morphogenesis and development (**Figure 5C**), as well as cell cycle and chromatin organization (**Supp Table 1**). While DE genes not close to DB are also enriched for developmental terms (**Supp Table 1**), it is to a lesser degree in heart- or cardiac-specific developmental processes; they are also enriched for more general processes such as those involved in translation (**Figure 5C**).

To prioritize DBs, we ranked them based on the magnitude of change in boundary scores across the

pairs of timepoints as well as the significance of the nearby DE gene (**Methods**). Among the top-ranked boundaries, we found DE genes with known roles in the cardiac development. One particular boundary was found in primitive cardiomyocytes (day 15) but absent in fully differentiated ventricular cardiomyocytes (day 80, **Figure 5D**). This boundary overlaps *MYH6*, highly expressed in day 15 compared to day 80, and is adjacent to *MYH7*, which displays the opposite expression change pattern to *MYH6*. Both these genes are involved in cardiac muscle function: *MYH6* is expressed at high levels in developing atria, while *MYH7* in ventricular chambers of the heart (Ching et al., 2005; Warkman et al., 2012). Recently an enhancer region cluster, located downstream to *MYH7* at chr14:23,876,121-23,878,188 , was identified as a switch that can downregulate expression of *MYH7* while upregulating *MYH6* expression upon deletion (Gacita et al., 2021). Another DB found in pluripotency day 0 but absent in day 2 overlaps *ESRG*, which is highly expressed in day 0 compared to day 2 (**Supp Figure 6**). *ESRG* is also a HERV-H containing long non-coding RNA (lncRNA, Wang et al., 2014), which we showed to be enriched in the day 0-specific boundaries (**Figure 4A**). *ESRG* may effect the pluripotency state on knockdown (Wang et al., 2014) and has known roles in developmental and embryonal carcinoma (Wanggou et al., 2012). This provides further evidence for TGIF-DB's ability to identify relevant dynamic boundaries that could impact overall cell state identity.

TGIF identifies boundary and compartment changes during mouse neural differentiation

To test the applicability of TGIF to multiple species and other dynamic systems, we applied TGIF-DC and TGIF-DB to a mouse neural differentiation dataset (Bonev et al., 2017) comprising three time points: mouse embryonic stem cell (ES) state, neural progenitors (NPC), and cortical neurons (CN). As in our human cardiomyocyte study, TGIF-DC compartment assignments are consistent with those from the PCA-based approach (rand index 0.72-0.95, **Supp Figure 3B**). Across the 3 time points, the majority (75.2%) of the genomic regions maintain their compartment assignments (**Figure 6A**), with the rest switching compartments at least once. NPC and CN showed higher degree of similarity in compartment assignments with a Rand index of 0.792 than with ES (**Figure 6B**), followed by ES and NPC which had an Rand index of 0.750; ES and CN were most distinct in their compartment structure (Rand index of 0.647). These results are consistent with the Bonev et al. study, which showed Hi-C measurements for NPC and CN have higher correlation among them than with the Hi-C measurements from the ESC state (Bonev et al., 2017).

As this dataset had multiple activating and repressive histone modifications available, we performed

a more in-depth characterization of each compartment and changes between compartments. In both NPC and CN, the mean depth- and quantile-normalized signals for activating histone marks such as H3K27ac, H3k36me3, and H3k4me1 were nearly 2-fold higher in regions assigned to one compartment than in the other compartment (**Figure 6C**). We refer to the compartment with higher H3K27ac, H3k36me3, H3k4me1 as the "A" compartment, and the other compartment as the "B" compartment. As in the cardiomyocyte differentiation, we also found that the regions in A compartment had higher mean observed over expected interaction counts compared to those assigned to B (**Supp Figure 7**). To examine more fine-grained chromatin patterns associated with compartment dynamics, we combined the fold change of each histone mark in NPC over CN into a single feature vector for each region. Next, we applied a kmeans clustering algorithm to each set of regions displaying static or differential compartment assignment between NPC and CN (**Methods**, **Figure 6D**). Among regions that change compartment assignment from A in NPC to B in CN (second heatmap, **Figure 6D**), one cluster had higher activating (H3K27ac and H3k4me1) and elongation (H3K36me3) marks in NPC than in CN (gray band, **Figure 6D**), and another cluster had higher H3k36me3 and H3K4me1 signals in NPC (mustard cluster, **Figure 6D**). In contrast, a cluster of regions switching from compartment B in NPC to A in CN displays lower activating marks (H3K27ac/H3k4me1) in NPC (teal and blue clusters in the third heatmap of **Figure 6D**). Static regions, which stayed in the A compartment in both NPC and CN, included a subgroup displaying lower signal in three activating marks (H3K27ac, H3k36me3, and H3k4me1) in NPC but higher levels of H3k4me3, another mark associated with active transcription (grey and green cluster in the first heatmap of **Figure 6D**). Static regions that remained in the B compartment exhibited clusters with high H3K36me3 and H3K4me1 (teal and blue, fourth heatmap, **Figure 6D**). The patterns of histone modifications suggest that not all activating marks behave in the same way between compartments and there are more fine-grained dynamics at chromatin remodelers that influence the compartment structure.

In addition to differential compartment analysis, we applied TGIF-DB to characterize boundary-level changes. Across all 3 time points, we identified a total of 5111 boundaries across the 3 time points, of which 1689 were common in all time points (persistent boundaries) (**Figure 7A**). About 3000 boundaries were found within each time point, with the highest number of unique boundaries found in the ESC state (**Figure 7B**). Accordingly the Jaccard coefficient-based similarity of these boundary sets showed that the ES state was most distinct from the other two time points (**Figure 7C**). This is consistent with the overall compartment dynamics, however, as in the cardiomyocyte analysis indicates compartments change more gradually compared to boundaries.

To examine whether the enrichment of differential gene expression (DE) near differential boundaries

(DB) generalizes to other dynamic systems beyond human cardiomyocyte development, we performed the same DE and enrichment analysis for this dataset as well (**Methods**). Here too, genes close to DBs (within 100kb radius tend to be enriched for DE genes compared to non-DE genes (**Figure 7D, E**). Enrichment of GO biological processes showed that DE genes close to a DB were enriched for nervous system development processes (**Supp Figure 8**). While DE genes not close to a differential boundary were also enriched for developmental processes, they were also enriched for more generic processes such as splicing, cell cycle, and ribosome biogenesis.

Finally, we prioritized boundaries as described in the cardiomyocyte study, and identified DBs close to DE genes with known roles in neuronal development (**Methods**). In particular, immediately upstream of a top-ranked DB is the *Ncam1* gene which is DE between CN and NPC; *Ncam1* has known roles in neuron axon guidance and synapse formation (Hata et al., 2018; Shetty et al., 2013)). Another gene, *Tenm2*, also near a top-ranked DB (**Supp Figure 9**), is involved in transynaptic signaling synaptogenesis (Silva et al., 2011; Vysokov et al., 2018). Between ES and NPC states, we also found high ranking DBs close to relevant DE genes including a microRNA gene, *Mir124-2hg* (**Supp Figure 10A**), and *Maf* (**Supp Figure 10B**) *Mir124-2hg* was the most significant DE gene near a DB and has been shown to regulate neurogenesis (Cheng et al., 2009) and commitment to neural lineage (Makeyev et al., 2007), while *Maf* has been shown to play a role in regulation of glutamatergic neurons in the mouse cortex (Ito and Imamura, 2022). Taken together, TGIF identified biologically meaningful compartment and boundary-level changes that were associated with corresponding chromatin and gene expression changes during mouse neural differentiation.

TGIF boundaries are enriched for SNPs from diverse disease phenotypes

Single nucleotide polymorphisms (SNPs) identified from genome-wide association studies (GWAS) are frequently found in non-coding regions of the genome and have been implicated in disease phenotypes by affecting the 3D genome organization (Lupiáñez et al., 2015; Orozco et al., 2022). Specifically, such variants could disrupt TAD boundaries and cause promiscuous expression of genes (Chakraborty and Ay, 2018; Lupiáñez et al., 2015). We investigated whether TGIF boundaries from the human cardiomyocytye 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 time points (persistent boundaries) with hematological measurement, cardiovascular disease, and lipid or lipoprotein measurement being the

most the most enriched phenotypic categories (**Figure 8A**, **Supp Figure 11**). The traits that had lower enrichment included neurorological disorders and non-specific categories. Importantly, we found that SNPs associated with cardiovascular disease (CVD) exhibited the second highest enrichment. We examined 66 persistent boundaries with at least one CVD-associated SNPs found within. One such boundary had the SNP, rs72705895, which is associated with venous thromboembolism (Lindström et al., 2019) (**Figure 8B**), and additionally overlaps a CTCF binding site (regulatory feature ENSR00000255184 from Ensembl regulatory build annotations (Cunningham et al., 2022; Zerbino et al., 2015)). Another boundary included rs9349379, which is found in the intronic region of *PHACTR1* (**Supp Figure 12**). Both the intronic variant and the gene is associated with coronary artery atherosclerotic disease (Koitsopoulos and Rabkin, 2021; Kuveljic et al., 2021), while the SNP itself lands on a predicted enhancer region (Ensembl regulatory build annotation ENSR00001107203), suggesting its putative role in disrupting an intronic enhancer. Genetic perturbations of the boundary locations harboring these SNPs and examining their impact on boundary and compartment structure during a developmental process could help examine the role of dysregulated 3D genome organization as a possible mechanism by which regulatory variants impact phenotype.

Discussion

Systematic characterization of dynamics of three dimensional genome organization is important to improve our understanding of the impact of this layer of regulation across different biological contexts, such as species, time, and developmental stage. Advances in genomic tools (Dekker et al., 2023, 2017) and concerted efforts from large consortia are leading to a growing compendia of high-throughput chromosome conformation capture datasets. However, systematic analysis of these datasets to quantify the extent of change and its impact on phenotype is a challenge because of the multiple layers at which the 3D genome is organized and the paucity of tools to analyze multiple datasets representing a large number of conditions. To address this challenge, we developed Tree-guided Integrated matrix Factorization (TGIF) that combines multitask learning with matrix factorization to examine dynamics of 3D genome organization across multiple structural scales and biological conditions. TGIF can further exploit the relatedness of the datasets over time or across biological conditions. Application of TGIF to real and simulated datasets showed that TGIF can accurately recover topological units such as compartments and topologically associated domains (TADs) during developmental processes and predict dynamic boundaries that are supported by numerous one-dimensional features such as architectural protein enrichment and differential expression.

To examine differences and similarities across multiple Hi-C datasets, one could potentially define units of interest independently followed by a post-processing step of comparing them. Another alternative is to consider pairs of datasets and then consolidate differences across pairs to gain a more holistic picture of the overall dynamics. We compared both these approaches to TGIF on simulated data and found that while all types of approaches captured the differences, a non-multi-task approach tended to have a high false positive rate (low precision) suggesting such approaches may call a number of spurious differences. Furthermore, TGIF also was less susceptible to calling artificial boundaries due to depth differences.

An open question with topological domain changes is how this relates to changes in gene expression (Cavalheiro et al., 2021; Ghavi-Helm et al., 2019; Greenwald et al., 2019; 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 specific to disease genes or genome wide is still unclear (Cavalheiro et al., 2021). Using the cardiomyocyte and neuronal differentiation data we examined the propensity of differentially expressed genes to be close to differential boundaries. 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 enriched for context-specific processes which could indicate that such changes are associated with fine tuning of gene expression during cellular differentiation. TAD changes can happen due to *cis* or *trans* factors (Cavalheiro et al., 2021). In our setting, given that these changes are happening during a developmental process, they are likely mediated by trans factors including cardiac-specific transcription factors that were enriched in the timepoint-specific boundaries. The types of TAD changes we investigated were gain and loss of boundaries between pairs of time points and thus the expansion or contraction of a particular domain. In-depth characterization of several of these loci identified genes that have known lineage-specific roles. Follow up experiments that perturb the TAD boundaries and measure the impact of gene expression would be very beneficial for assessing whether the gene associations we observed are causal.

Regulatory sequence variants can mis-regulate gene expression by disrupting TAD boundaries (Chakraborty and Ay, 2018; Lupiáñez et al., 2015). We used our TAD boundaries to examine the impact of variation. Interestingly, when considering different types of boundaries based on whether they were timepoint specific, or conserved to different extents, we found the greatest enrichment in boundaries that did not change over time, namely the persistent TADs. Furthermore, we found several cardiovascular and metabolic disease trait SNPs to be enriched in these TADs. This persistent set of TADs 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 TADs in other developmental lineages and their propensity to represent tissue-specific boundaries. Additionally, this provides a way to prioritize variants for downstream functional experiments that could be important to identify the mechanisms by which variants disrupt gene regulatory processes.

There are a number of directions in which TGIF could be extended. TGIF expects the relatedness of the datasets to be provided as user input. However, for some situations this information may not be known or may not be captured by a hierarchy/tree. One direction of future research is to handle handle more complex relationships, e.g., as a graph, as well as to automatically learn the relationship among the datasets. A second direction of research is to consider auxiliary measurements, including sequence to inform the inference of the topological units using techniques such as semi-supervised clustering (Bair, 2013; Bondell and Reich, 2008).

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

examine 3D genome organization dynamics and its impact on normal and disease processes.

Materials and Methods

Tree-Guided Integrated Factorization (TGIF)

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

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

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

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

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

TGIF generalizes multi-view NMF to allow for integration of datasets that can come from different

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

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

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

The tree describes parent-child relationships between the tasks. This 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 hyperparameter 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. If all tasks share a single parent node (the root node), TGIF reduces to multi-view NMF.

Many optimization algorithms exist for learning the latent factors in matrix factorization (Kim et al., 2014). We chose a block coordinate descent (BCD) because it guarantees convergence to a local optimum (Kim et al., 2014). Intuitively, block coordinate descent updates a given block while keeping all other blocks fixed; in TGIF the block is each column or row of $\mathrm{U}^{(\cdot)}$s or $\mathrm{V}^{(\cdot)}$s. Starting at the leaf nodes (i.e. input tasks), we update each column/row of the factors, then move up the tree to update the parent factors. The specific update rules and their derivations can be found in **Supp Material 1**.

TGIF's factors can be used to define compartments as well as fine-scaled topologically associating domains (TADs). We now describe TGIF-DC for compartments and TGIF-DB for boundaries.

TGIF-DC for differential compartment identification

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 count (O/E) correlation matrix (See Section on **Datasets and preprocessing**). To obtain the O/E matrix, we distance-normalize every entry of the input matrices $\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(t)} \in \mathbb{R}_{\geq 0}^{n \times n}$ as follows:

$$\mathbf{X}_{\text{OE}}^{(t)}[i,j] = \frac{\mathbf{X}^{(t)}[i,j]}{\frac{\sum_{d(p,q)=d(i,j)|} \mathbf{X}^{(t)}[p,q]}{n - |i-j|}} \tag{3}$$

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

$$\mathbf{X}_{\text{corr}}^{(t)}[i,j] = \text{corr}(\mathbf{X}_{\text{OE}}^{(t)}[i,:], \mathbf{X}_{\text{OE}}^{(t)}[j,:]) \tag{4}$$

where corr corresponds to the Pearson's correlation.

We upshift the correlation matrix by 1 so that all values are non-negative and apply TGIF with the input tree structure on 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 and compare the cluster assignments across the tasks to identify regions switching compartments or maintaining them.

We scanned multiple hyperparameter values in $\alpha \in \{10^2, 10^4, 10^6, 10^8\}$ across all chromosomes in the mouse neural differentiation data and compared the resulting compartment assignments using Rand Index, which is a measure of cluster similarity. Rand Index ranges from 0 to 1, with 1 indicating perfect agreement. We find that the compartment assignments remain relatively stable or similar across the hyperparameter settings with a Rand Index ranging from 0.84-0.98 (**Supp Figure 13A**). Our results are from $\alpha = 10^4$. As BCD is a stochastic algorithm that can reach different local optima depending on the initialization point, we also experimented with multiple random initialization seeds. We use Rand Index to measure the cluster assignment similarity between pairs of cluster assignments from two different seeds. For all pairs of random initializations, the Rand Index was 1 (perfect correspondence, **Supp Figure 13B**), showing that TGIF-DC yields a stable set of compartment assignments regardless of random initialization.

TGIF-DB for differential boundary identification

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

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

In NMF, usually the 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 low 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.

To determine the hyperparameter α for TGIF-DB, we experimented with different values of $\alpha \in \{10^2, 10^4, 10^6, 10^8\}$ on the cardiomyocyte differentiation data and measured the Jaccard index-based similarity between the identified boundaries for each hyperparameter setting (**Supp Figure 14A**). The boundaries identified for different α settings were similar (Jaccard index 0.9-0.75), although 10^8 resulted in relatively lower Jaccard index compared to the other settings. We used results from $\alpha = 10^6$ as this represented a good balance between the extent of regularization and stability of boundaries to different hyperparameter settings. As in the compartment analysis, we also experimented with different random initialization seeds and found that the resulting boundary sets from different initialization are relatively consistent with pairwise Jaccard index 0.76-0.77 (**Supp Figure 14B**).

Boundary score calculation. After factorization, the next step is to identify genomic regions represented 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 for task t is the cosine distance between the low dimensional representation of region i and region $i-1$ is defined as follows:

$$S_i^{(t)} = 1 - \frac{\mathrm{U}^{(t)}[i,:] \cdot \mathrm{U}^{(t)}[(i-1),:]}{\|\mathrm{U}^{(t)}[i,:]\| \|\mathrm{U}^{(t)}[(i-1),:]\|} \tag{5}$$

While Euclidean distance could be used to compute the score, we used cosine distance since it normalizes the score regardless of magnitude differences across the tasks (which can arise from Hi-C sequencing depth differences). The final boundary score for region i in task t is the mean of $S_i^{(t)}$ estimated from factors across the range of $k \in \{2, \cdots, 8\}$. and take their mean. This allows us to capture the "boundaryness" of each region at different structural scales (**Supp Figure 1C**).

Empirical p-value calculation and FDR correction. Once the scores are calculated, we estimate a "null" distribution of boundary scores and used it to determine the empirical p-value of boundary scores to find significant boundaries. The null distribution is computed from a randomized background matrix. We first calculate the element-wise mean across T input submatrices to yield M, then create randomized background matrix by shuffling the entries of M. The shuffling is done in a distance-stratified manner; that is, we obtain all pairs of genomic regions that at a distance d and permute them independently of the interaction counts for pairs of regions which are d' apart, where $d' \neq d$. We next perform single-task NMF on this shuffled M matrix for the same range of k and derive boundary scores for all regions in the same way noted in the previous section. We treat this set of boundary scores as the samples from the null distribution. We calculate the empirical p-value for each the region i in task t as the proportion of "null" background scores higher than the given region's boundary score. Finally, to find significant boundaries and to correct for multiple significant testing, we perform the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

The output of the p-value and FDR estimation step is a binary flag for each region i and task t, indicating whether the region has a significant boundary score. 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.

TGIF outputs a bed file containing the significant boundary status of each region-by-region annotation and whether the region includes a shared or task-specific significant boundary. For instance, with two input datasets representing two conditions, A and B, a region is annotated one of four labels: non-boundary, A-specific boundary, B-specific boundary, and shared boundary). We annotate each region based on its boundary patterns for any combination of context-specificity based on the number of input datasets or conditions.

Hi-C data simulation

In order to benchmark methods that can detect TAD-level changes, we generated simulated contact matrices with known TADs and TAD changes for four hierarchically related conditions (**Supp Figure 2**). The TAD changes can fall into one of three categories: TAD split creating a new boundary, TAD merge removing a boundary, and TAD shift where the location of a boundary is moved up or down the linear chromosome (**Supp Figure 2A**, Cresswell and Dozmorov, 2020). We first generate a set of TADs with known change patterns, followed by populating the contact matrices with simulated read counts. We follow the Hi-C matrix simulation procedure in the benchmarking study by Forcato et al. (Forcato et al., 2017). Forcato et al. sampled TAD sizes from a uniform distribution with minimum value 3 (3 bins TAD size) and maximum value 50 (50 bins TAD size). They simulated 171 TADs, which at 40kb resolution resulted in a target size of the simulated matrix similar to the size of the human chromosome 5 (180.92Mb).

To generate the TADs for each of the four tasks, we start with these 171 TADs (referred to as A) at the root. Of these, 40 TADs are kept unchanged during the TAD change simulation process. We proceed down the tree branch keeping A unchanged for the left most branch (**Supp Figure 2B**), and performing different TAD change operations on the other branches. At the level immediately below the root, we introduce 30 changes (10 merges, 10 splits, and 10 shifts) for the right branch to create a TAD set we refer to as B. At the next level, we have the TAD sets A (left) and B (right). We again proceed down the tree keeping A unchanged for the left child (this is now a leaf node), and making 15 changes (5 merges, 5 splits and 5 shifts) to the right branch. The resulting TAD set is referred to as A_2. For the right intermediate node (B), we perform the 15 changes (5 merges, 5 splits and 5 shifts) to both children producing TAD sets, B_1 and B_2. This results in a tree structure where TAD sets A and A_2 are most similar, followed by B_1 and B_2. The remaining pairs have increasing differences as quantified

in **Supp Figure 2C**. These resulting TAD sets are considered our "gold-standard" ground truth set of TADs. Given this TAD structure we use the Hi-C count simulation procedure described next to populate these TADs.

Given a TAD structure, we follow the simulation procedure from Forcato et al., 2017 to generate the counts. Each interaction count for a pair of regions (i, j) is sampled from a negative binomial distribution defined by two parameters: (1) the dispersion parameter is fixed at 0.01 as in Forcato et al., 2017 and (2) the mean parameter $\mu_{i,j} = K(i - j + 1)^{-0.69}$, which is dependent on the distance $(i - j)$ between the two interacting regions i and j. The mean parameter $\mu_{i,j}$ decays as the distance between i and j grows to reflect how Hi-C interaction counts decay as a function of pairwise distance. As in Forcato et al., 2017, a prior value of 1 is added to this distance when calculating $\mu_{i,j}$. Here -0.69 is the rate of decay parameter estimated by Forcato et al from a real contact matrix (chr5 from IMR90 cell line). Additionally, following Forcato et al., 2017, $K = 28$ if a pair of regions falls within the same TAD and $K = 4$ otherwise. Finally, we add noise and sparsity to the simulated matrices as follows. To add noise, we randomly sample 20% of unique interactions in each simulated matrix and add 2 to the mean parameter $\mu_{i,j}$ before sampling. To introduce sparsity, we again randomly sample 20% of the entries and set them to zero.

Methods for identifying differential domain boundaries

TGIF was benchmarked against two other methods for identifying differential TAD boundaries: (1) TADCompare (Cresswell and Dozmorov, 2020) and (2) TopDom (Shin et al., 2016). TopDom is a TAD identification method for a single condition, applied to each dataset individually followed by pairwise comparison of identified boundaries. We also initially considered the TADsplimer method (Wang et al., 2020), which was specifically developed to detect differential TAD identification; however, its implementation has an unresolved issue that fails to return any output or differential boundaries and was excluded from further analysis.

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

fine a boundary. Although TADCompare accepts time-series data, differential boundaries are computed for only pairs of matrices at a time.

TopDom is a TAD identification method which was shown to be robust to noise and yielded TADs enriched in structural proteins such as CTCF (Dali and Blanchette, 2017; Lee and Roy, 2021). It generates a score for each bin along the chromosome, where the score is the mean interaction count between the given bin and a set of upstream and downstream neighbors (neighborhood size is a user-specified parameter). Putative TAD boundaries are picked from a set of bins whose score forms a local minimum; false positive boundaries are filtered out with a significance test. If a boundary was found for one input matrix and not the other, it was considered a task-specific boundary region. A boundary found in both inputs shared boundary region, and the rest non-boundaries.

Benchmarking with simulated Hi-C data with ground truth TADs

We applied TGIF, TopDom, and TADCompare to the simulated datasets (described in Section **Hi-C data simulation**) with ground truth TADs to assess their ability to recover shared and differential boundaries. We applied TADCompare (version 0.99.36) to each pair of the 4 simulated matrices at a time. This was done because our simulated dataset follows a branching pattern rather than a linearly changing pattern. TADCompare outputs differential boundaries, including the task in which the boundary is present and non-differential, which we consider as shared boundaries. We applied TopDom (version 0.10.1) 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 then compared to identify task-specific and shared boundaries.

TGIF was applied to all four simulated matrices together with the known tree structure used to generate the simulated data (**Supp Figure 2C**). Although TGIF can jointly annotate more than two boundaries in all input datasets, as the other methods require pairwise comparisons, we only use the differential and shared boundaries between pairs of tasks for TGIF as well.

We used precision, recall and f-score as metrics to compare the agreement of predicted shared and context-specific boundaries with the true shared and task-specific boundaries. We computed these metrics for every pair of simulated datasets used to identify pairwise task-specific or shared boundaries. Briefly, let $T_{\text{shared}(A,A_2)}$ be the true number of shared boundaries from matrices or tasks A and A_2, $P_{\text{shared}(A,A_2)}$ be the number of predicted shared boundaries, and $I_{\text{shared}(A,A_2)}$ be the overlap between the true and predicted shared boundaries. The precision, p is defined as $\frac{I_{\text{shared}(A,A_2)}}{P_{\text{shared}(A,A_2)}}$, recall, r is defined as $\frac{I_{\text{shared}(A,A_2)}}{T_{\text{shared}(A,A_2)}}$. Then the f-score is the harmonic mean of p and r, defined as $\frac{2*p*r}{p+r}$. For task-specific bound-

aries, we use a similar definition of precision, recall and f-score and also compute these for every pair of tasks/matrices.

Benchmarking with downsampled data to assess robustness to depth

We downsampled the high-depth Hi-C dataset from the GM12878 cell line (Rao et al., 2014) to the depth of lower-depth datasets from four other cell lines, K562, HUVEC, HMEC, NHEK, using an approach previously described (Lee and Roy, 2021). We used Jaccard index to measure how well the three methods, TGIF, TADCompare, TopDom, are able to recover a consistent set of boundaries despite depth difference. The Jaccard index is calculated by dividing the number of boundaries found at both depths and by the number of boundaries identified in either depths for the GM12878 dataset. The higher the Jaccard Index, the fewer the false-positive differences.

TADCompare was 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 K562 depth). The Jaccard index was calculated for each pair of datasets as the number of non-differential boundaries divided by the number of differential and non-differential boundaries. TopDom was applied individually to 5 datasets: original high-depth GM12878 data, GM12878 data downsampled to K562, HUVEC, HMEC, NHEK depths, respectively. Jaccard index was calculated between the original dataset and each of the downsampled datasets. TGIF was applied to all five datasets together (original GM12878 data and the four low-depth GM12878 datasets downsampled to K562, HUVEC, HMEC, NHEK depths) using a "flat" tree with a single parent at the root. From TGIF outputs, we calculated the pairwise Jaccard index between the original Gm12878 dataset and each of the four downsampled datasets.

Datasets and preprocessing

We applied TGIF to two Hi-C timecourse datasets: mouse neural differentiation data from Bonev et al., 2017, and human cardiomyocyte differentiation from Zhang et al., 2019. The mouse neural differentiation data (Bonev et al., 2017) included 3 timepoints during mouse neural differentiation: embryonic stem cell (ES) stage, neural progenitor stage (NPC), and the differentiated cortical neuron (CN) stage. For TGIF-DB, we used intrachromosomal count matrices at a resolution of 25kb resolution with vanilla-coverage square-root normalization as input. For TGIF-DC, intrachromosomal interaction count matrices at 100kb resolution without normalization was used as input as TGIF-DC computes the observed/expected and correlation matrices. In addition to the Hi-C measurements, this dataset also in-

cluded ChIP-seq data for 6 histone modification marks (H3k27ac, H3k28me3, H3k36me3, H3k4me1, H3k4me3,H3k9me3), for both CN and NPC. These data were used to characterize and validate the chromatin structure inferred by TGIF-DC. For each 100kb bin, the ChIP-seq signal in reads-per-million was summed up within each bin, and the signal in each bin was divided by the total number of reads to first normalize by read depth. Subsequently, signals in NPC and CN were quantile normalized to each other to enable log fold change comparison across the two timepoints. The log fold change for each of the 6 marks in each bin was calculated by log transforming (log(x+1)) the normalized signal in NPC divided by the log transformed signal in CN. The set of log fold change signals were used as input to a k-means clustering, with $k = 5$ clusters applied to each subset of regions (e.g. regions remaining in compartment A) or dynamic regions (e.g. regions switching from A in NPC to B in CN).

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

Comparison of TGIF-DC to existing compartment calling method

A commonly used method for segmenting the genome into compartments is to apply principal component analysis (PCA) to the observed over expected count (O/E) correlation matrices (see Section**TGIF-DC for differential compartment identification**, Lieberman-Aiden et al., 2009). The first principal component (PC1) is used to assign genomic regions to two compartments; values equal to or above zero in PC1 are assigned into one compartment and those below zero are assigned to the other compartment. The actual annotation of each compartment as the active “A” or repressed “B” compartments is done by correlating the compartment structure to one-dimensional regulatory signals such as ATAC-seq assays or histone modifications. We measured the similarity of TGIF-DC compartments with that of PCA method with Rand Index.

Accessible and active motif site enrichment in TGIF boundaries

To identify putative architectural proteins we obtained the hg19 motif match sites of 746 different transcription factor (TF) motifs from the JASPAR database (Fornes et al., 2020), and intersected them to hESC DNase-seq broadpeaks downloaded from ENCODE (ENCODE Project Consortium, 2012; Luo et al., 2020). For each TF, we calculated the fold enrichment of its accessible motif match sites in TGIF boundaries specific to each time point from the cardiomyocyte dataset. The fold enrichment ratio for a given TF motif was calculated as $\frac{q}{\mathrm{M}}/\frac{s}{\mathrm{N}}$, where q is the number of boundaries with at least one accessible motif site, M is the number of boundary regions, s is the number of regions with at least one accessible motif site, and N is the total number of genomic regions. We also used these estimates to compute a p-value using a hypergeometric test. Because many of the p-values were 0, the TFs were ordered by their actual fold enrichment value to identify putative architectural proteins and co-factors. We performed a similar enrichment analysis of active TF motif sites by intersecting the JASPAR motif match sites with H3K27ac broadpeaks available for each of the 6 timepoints (Zhang et al., 2019). Fold enrichment and a hypergeometric test p-value was calculated in a similar manner as the accessible motif sites.

Assessing differential gene expression near differential boundaries

We used RSEM (Li and Dewey, 2011) on the RNA-seq data from the cardiomyocyte differentiation time course to obtain expected counts for each replicate at each time point. We used these values as input to DESeq2 (Love et al., 2014) to identify differentially expressed (DE) genes for every pair of timepoints (e.g. day 0 vs. day 2). DE genes were defined by using a threshold of $<$0.05 on the adjusted p-value. For every pair of timepoints, we also identified the list of pairwise differential boundaries (DB), defined as boundaries present in one of the timepoints but not both. To test the enrichment of DE genes near (i.e. within 50kb of) differential boundaries, we used the fold enrichment ratio $\frac{q}{\mathrm{M}}/\frac{s}{\mathrm{N}}$, where q is the number of DE genes near DB, M is the total number of DE genes, s is the number of genes near a DB and N is the set of all genes. Hypergeometric test was additionally performed to calculate the significance of this fold enrichment value for each pair of timepoints.

Gene ontology (GO) term enrichment analysis was performed for different subsets of genes based on their DE status and whether they were near DB. In total, 4 different subsets of genes were assessed for enrichment of GO biological processes, against the background of all genes: (1) non-DE genes not close to (i.e. not within 50kb of) a DB, (2) DE genes not close to a DB, (3) non-DE genes close to a DB, and (4) DE genes close to a DB. The significance of enrichment was determined with an FDR-corrected

hypergeometric test. We repeated the same analysis for every pair of timepoints in both the mouse neural differentiation and cardiomyocyte differentiation dataset.

To select candidate differential boundaries for visualization and detailed interpretation, we ranked differential boundaries based on two criteria: (1) the degree of change in the boundary signal between the pairs of timepoints being compared measured by absolute difference in TGIF-DB boundary score, and (2) the significance of the nearby differential expression measured by the most significant p-value of differentially expressed gene within 50kb of the boundary. Since the scale of the two metrics are different, we converted these values into ranks and used the mean rank of a boundary to select promising differential boundaries.

SNP enrichment within TGIF boundaries from cardiomyocyte differentiation data

For all SNPs in the GWAS catalog (Sollis et al., 2023), we mapped each SNP's associated trait to its parent phenotype, i.e., its Experimental Factor Ontology (EFO) parent term. 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 boundaries. For each category, we calculated the fold enrichment of associated SNPs in different subsets of TGIF 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, **Supp Figure 11**), boundaries found in any of the timepoints (ANY, **Supp Figure 11**). 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.

Data access and code availability

Implementation of TGIF is available at `https://github.com/Roy-lab/tgif`. Hi-C and H3K27ac ChIP-seq data from the cadiomyocyte differentiation time course (Zhang et al., 2019) were downloaded from GEO (accession: GSE116862). Hi-C, RNAseq, and histone modification ChIP-seq data from the mouse neural differentiation (Bonev et al., 2017) was obtained from GSE96107. HESC (H9 cell line) and HUVEC DNase-seq broadpeak datasets were downloaded from ENCODE (ENCFF166OGN, ENCFF541JWD, ENCFF978UNU; ENCODE Project Consortium, 2012; Luo et al., 2020).

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

Author contributions

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

Figures

Figure 1

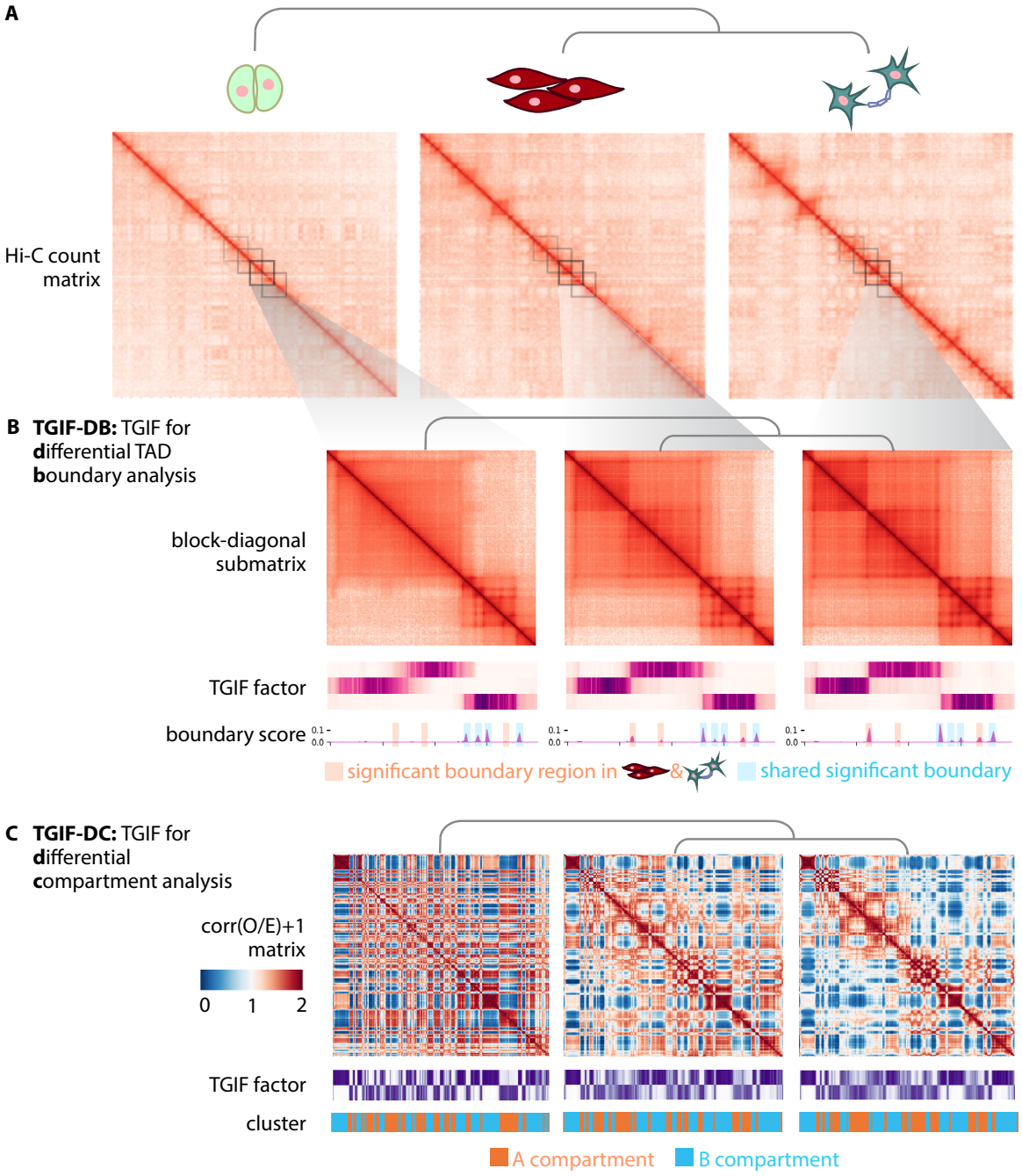

Figure 1. Overview of TGIF **A.** TGIF is a multi-task matrix factorization framework that takes as input multiple Hi-C matrices, each representing a biological condition or a cellular state, and a user-specified tree structure that encodes the relationship among the input Hi-C datasets. Shown is an example of a tree encoding the cell lineage of three tissues. **B.** TGIF-DB: one version of TGIF identifies conserved and differential TAD boundaries. Using the factors resulting from the multi-task factorization of the block-diagonal submatrices, a boundary score and its significance is calculated for each genomic region, and each region is annotated for its context-specific or persistent "boundary" nature. TGIF-DB pipeline is visualized in more detail in **Supp Figure 1**. **C.** TGIF-DC: another version TGIF identifies compartmental patterns across multiple conditions. From the input intrachromosomal interaction count matrices, the correlation matrices of observed over expected counts is calculated. These correlation matrices are factorized into two low dimensional components, and genomic regions are clustered using their representation in these factors. The clusters are annotated as A or B compartment based on their association with accessibility or histone modifications.

Figure 2

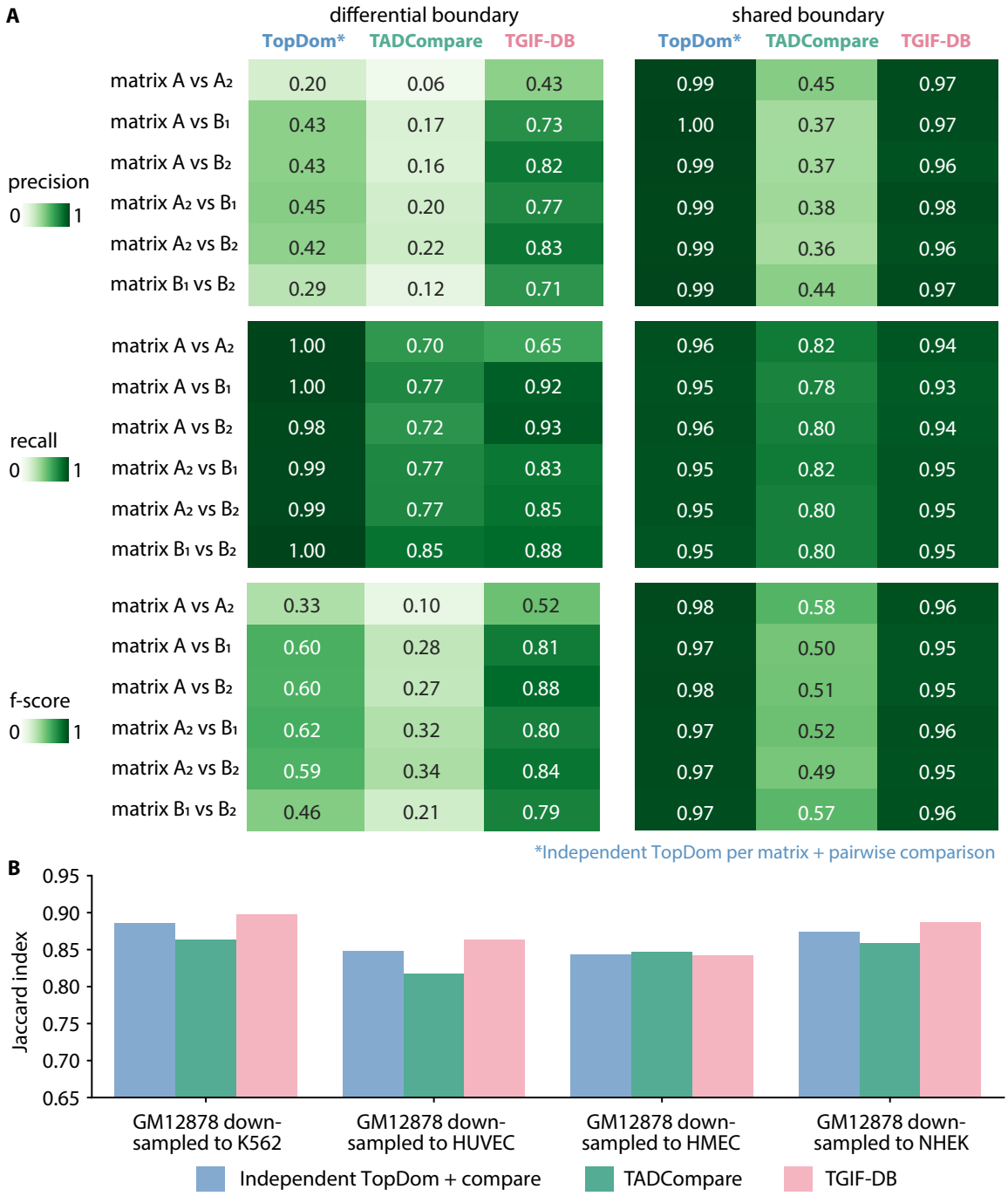

Figure 2. Benchmarking TGIF-DB with simulated and downsampled data. TGIF was benchmarked against two existing methods: TopDom (a single-task TAD-calling method applied independently to each

input matrix, followed by comparing the pairs of outputs to identify differential or shared boundariess) and TADCompare, a differential boundary detection method. **A.** Recovery of shared and differential boundaries on simulated data. The three methods were benchmarked on simulated data with known boundary differences. The output of each method is considered as predictions; precision, recall, f-score are measured for these predictions against the ground truth, for every pair of simulated datasets used to label true differential or shared boundaries. Simulation process is detailed in **Methods**. **B.** Sensitivity of boundary estimation to dataset depth. The three methods were applied to a high-depth Hi-C dataset from the GM12878 cell line, and a downsampled counterpart (downsampled to the depth of 4 lower-depth cell line datasets). The similarity of the boundary sets from the original dataset and the downsampled counterpart is measured with Jaccard index. Higher Jaccard index means fewer false positive differences identified from depth difference.

809

Figure 3

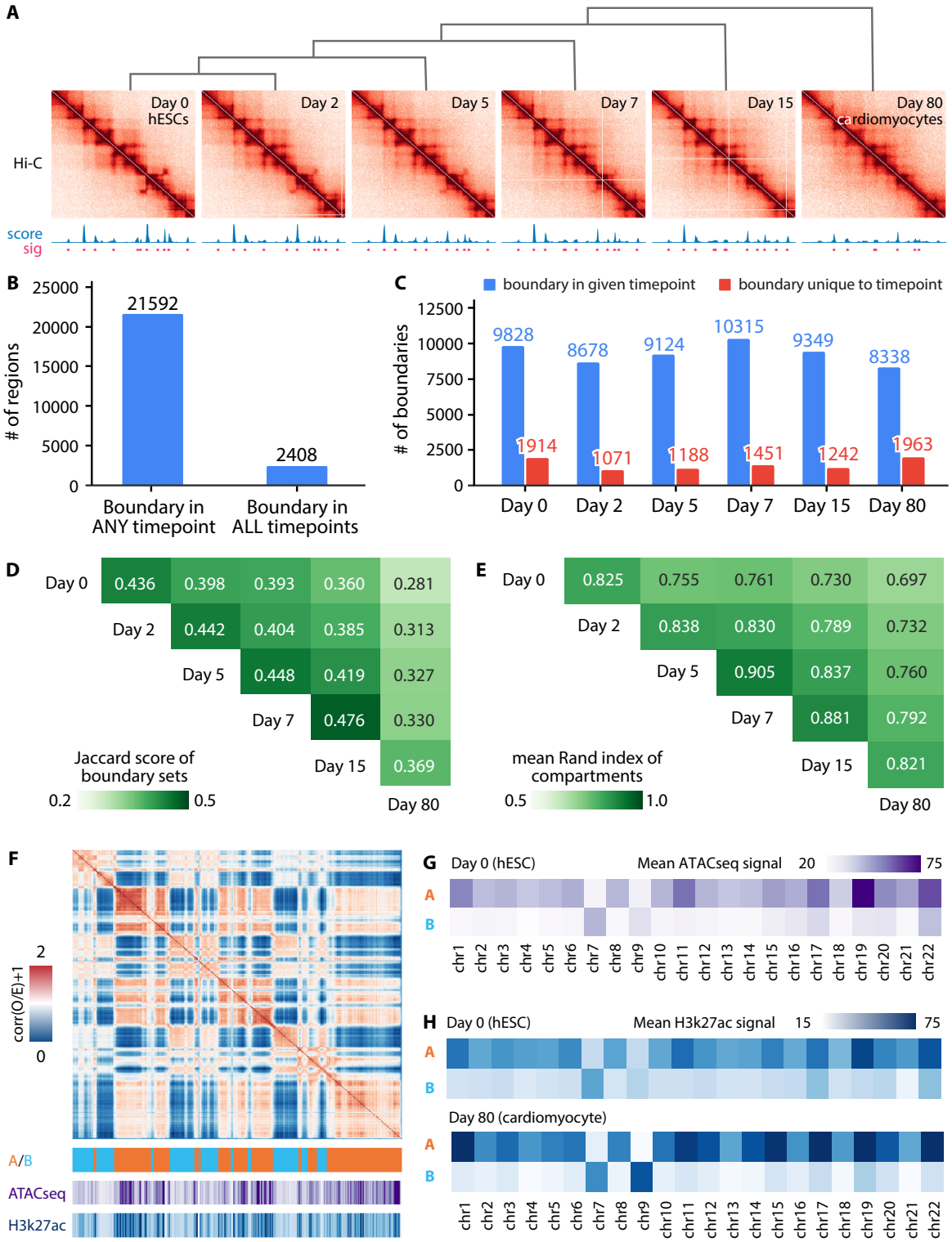
A
Hi-C
Day 0 hESCs
Day 2
Day 5
Day 7
Day 15
Day 80 cardiomyocytes
score
sig
B
# of regions
25000
20000
15000
10000
5000
0
21592
2408
Boundary in ANY timepoint
Boundary in ALL timepoints
C
boundary in given timepoint
boundary unique to timepoint
# of boundaries
12500
10000
7500
5000
2500
0
9828
1914
8678
1071
9124
1188
10315
1451
9349
1242
8338
1963
Day 0
Day 2
Day 5
Day 7
Day 15
Day 80
D
Day 0 0.436 0.398 0.393 0.360 0.281
Day 2 0.442 0.404 0.385 0.313
Day 5 0.448 0.419 0.327
Day 7 0.476 0.330
Day 15 0.369
Day 80
Jaccard score of boundary sets
0.2
0.5
E
Day 0 0.825 0.755 0.761 0.730 0.697
Day 2 0.838 0.830 0.789 0.732
Day 5 0.905 0.837 0.760
Day 7 0.881 0.792
Day 15 0.821
Day 80
mean Rand index of compartments
0.5
1.0
F
corr(O/E)+1
2
0
A/B
ATACseq
H3k27ac
G
Day 0 (hESC)
Mean ATACseq signal
20
75
A
B
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22
H
Day 0 (hESC)
Mean H3k27ac signal
15
75
A
B
Day 80 (cardiomyocyte)
A
B
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22

Figure 3. Application of TGIF to human cardiomyocyte differentiation data. **A.** TGIF-DB was applied to cardiomyocyte differentiation time course dataset with Hi-C interaction counts collected from 6 different timepoints (visualized here are submatrices from chr15: 88560000-90560000). For the input tree structure, each timepoint is added on a branch. TGIF-DB outputs the boundary score for each genomic region (plotted in blue), as well as the list of significant boundaries (marked by pink asterisks). **B.** The size of the union and intersection of TGIF boundaries found across the 6 timepoints. **C.** Number of TGIF boundaries found in each timepoint (blue), and those found only in each timepoint (red). **D.** Similarity of boundary sets from TGIF-DB between every pair of timepoints, measured by Jaccard index. **E.** Similarity of compartment assignments from TGIF-DC between every pair of timepoints, measured by mean Rand index across chromosomes. **F.** Heatmap visualizing compartment structure from TGIF-DC for a submatrix (chr10: 60000000-135600000) from day 0 of cardiomyocyte differentiation. The input Hi-C count matrices are converted to distance-stratified (Observed over Expected or O/E) correlation matrix and upshifted by 1 to make the values non-negative. TGIF-DC outputs compartment assignments for each genomic region at each timepoint. Compartments are assigned "A" or "B" based on their association to one-dimensional signals such as ATAC-seq or H3K27ac. "A" compartment is visualized in orange and "B" in blue). The ATAC-seq and H3K27ac signal heatmaps are visualized in purple and blue heatmaps. **G.** Heatmap visualizing the mean ATACseq signal in regions annotated as compartment A or B according to TGIF-DC, across all chromosomes, in day 0. **H.** Heatmap visualizing the mean H3k27ac ChIPseq signal in regions annotated as compartment A or B according to TGIF-DC, across all chromosomes, in day 0 and day 80.

830

Figure 4

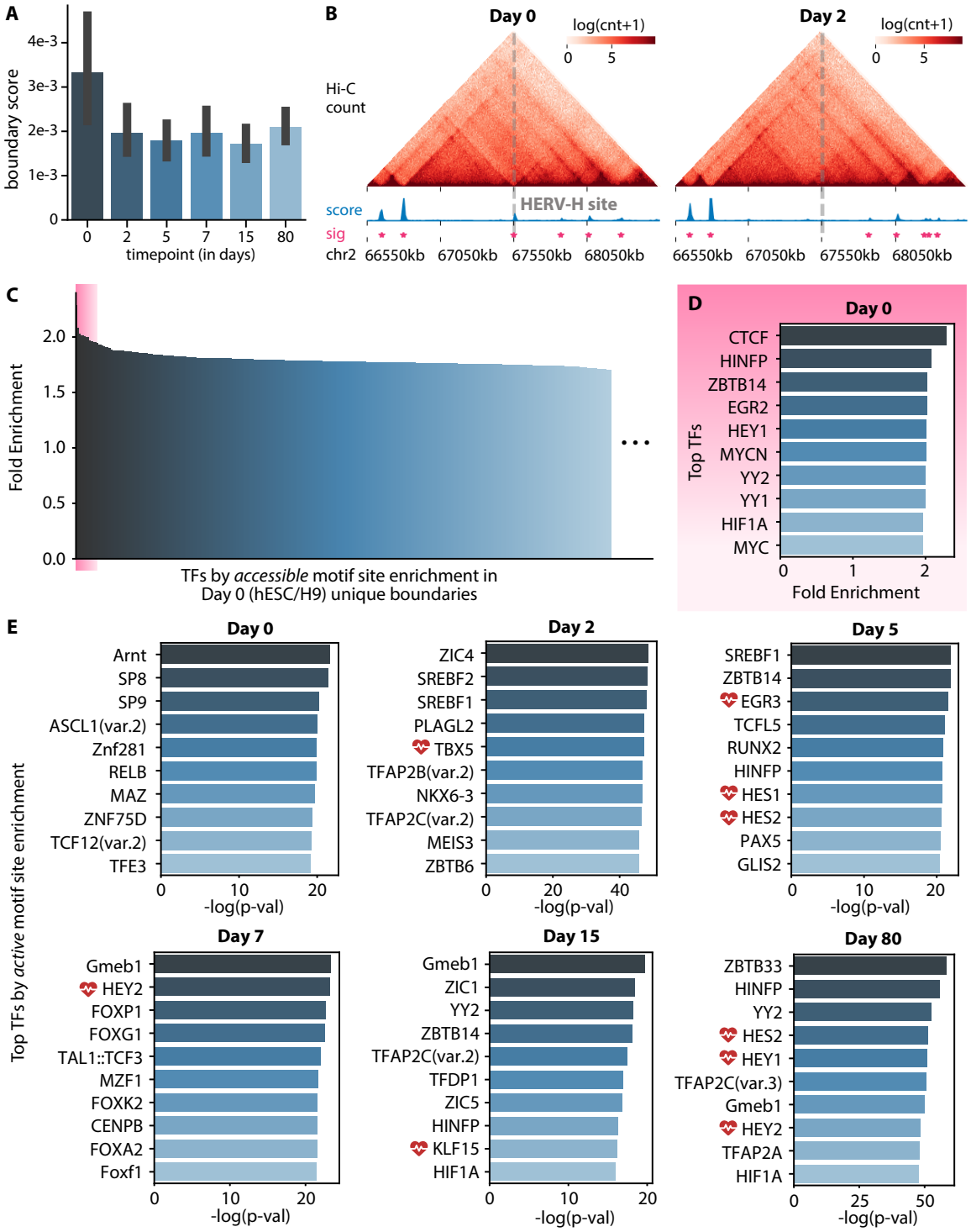
A
boundary score
4e-3
3e-3
2e-3
1e-3
0
0
2
5
7
15
80
timepoint (in days)
B
Day 0
Day 2
log(cnt+1)
0
5
Hi-C count
score
sig
HERV-H site
chr2
66550kb
67050kb
67550kb
68050kb
C
Fold Enrichment
2.0
1.5
1.0
0.5
0.0
TFs by *accessible* motif site enrichment in Day 0 (hESC/H9) unique boundaries
D
Day 0
Top TFs
CTCF
HINFP
ZBTB14
EGR2
HEY1
MYCN
YY2
YY1
HIF1A
MYC
0
1
2
Fold Enrichment
E
Top TFs by *active* motif site enrichment
Day 0
Arnt
SP8
SP9
ASCL1(var.2)
Znf281
RELB
MAZ
ZNF75D
TCF12(var.2)
TFE3
-log(p-val)
Day 2
ZIC4
SREBF2
SREBF1
PLAGL2
TBX5
TFAP2B(var.2)
NKX6-3
TFAP2C(var.2)
MEIS3
ZBTB6
Day 5
SREBF1
ZBTB14
EGR3
TCFL5
RUNX2
HINFP
HES1
HES2
PAX5
GLIS2
Day 7
Gmeb1
HEY2
FOXP1
FOXG1
TAL1::TCF3
MZF1
FOXK2
CENPB
FOXA2
Foxf1
Day 15
Gmeb1
ZIC1
YY2
ZBTB14
TFAP2C(var.2)
TFDP1
ZIC5
HINFP
KLF15
HIF1A
Day 80
ZBTB33
HINFP
YY2
HES2
HEY1
TFAP2C(var.3)
Gmeb1
HEY2
TFAP2A
HIF1A

Figure 4. Identifying context-specific boundary elements from human cardiomyocyte differentiation data. **A.** Mean TGIF-DB boundary score in the top 100 transcriptionally active HERV-H sites. The error bar is the 95 CI. **B.** Visualization of the boundary structure near a top transcriptionally active HERV-H site. Shown are Hi-C interaction counts (heatmap), TGIF boundary score (blue line plot), significant boundaries (pink star), near the HERV-H site (vertical dotted line). The pluripotency-specific boundary observed in day 0 is absent in subsequent timepoints (day 2). **C.** Fold enrichment of accessible TF motif sites (motif sites intersected with ATAC-seq peaks) in boundaries unique to day 0 pluripotent state. **D.** Zoomed in top 10 TFs based on their accessible motif site enrichment (the fold enrichment for these TFs all had hypergeometric test p-value of 0; hence they were ordered by their fold enrichment value.) **E.** Most significant TFs based on the fold enrichment of their active motif sites (motif sites intersected with H3K27ac peaks) in TGIF boundaries unique to each timepoint. Red heart highlights TFs associated with cardiovascular function.

843 **Figure 5**

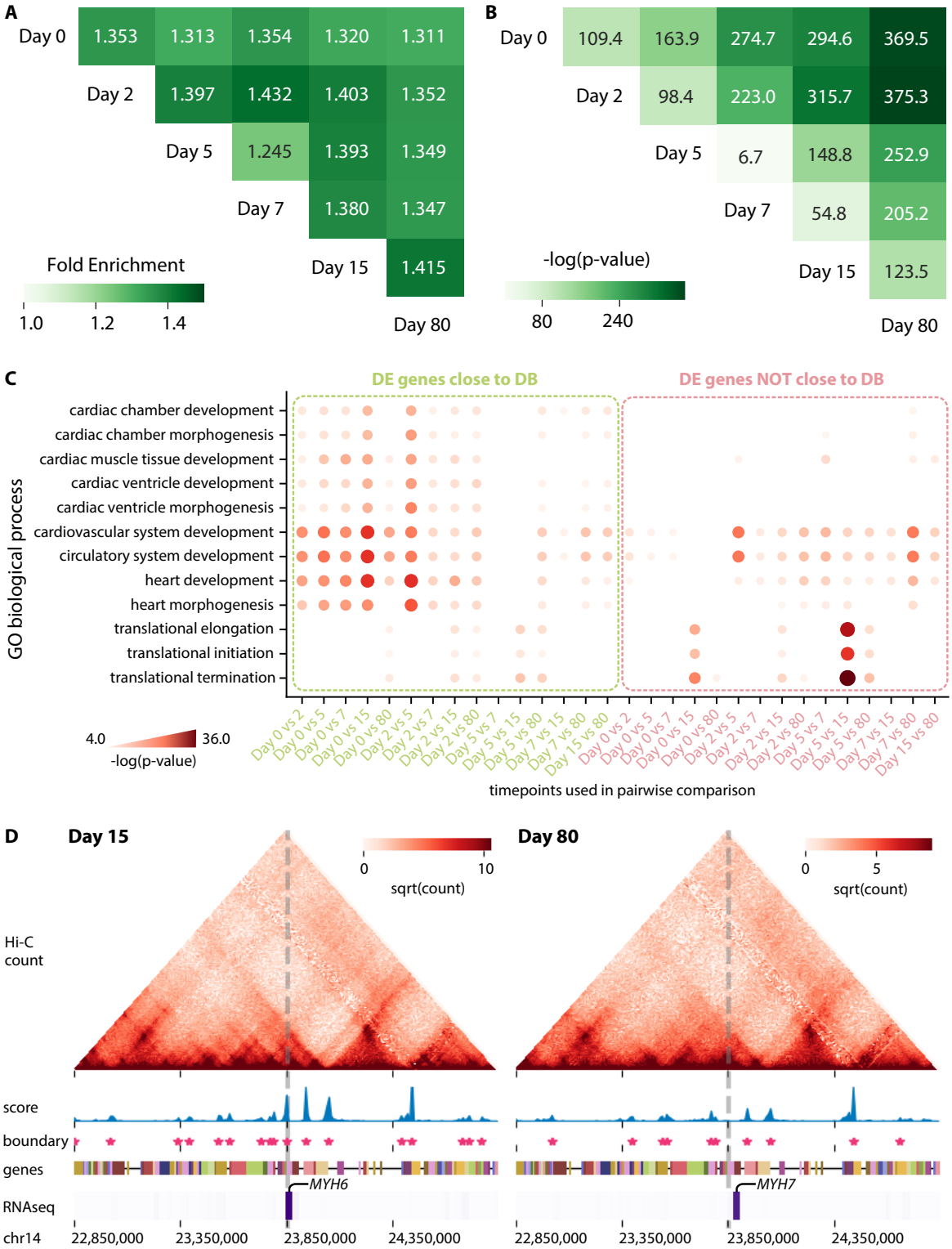
A
Day 0
Day 2
Day 5
Day 7
Day 15
Day 80
1.353 1.313 1.354 1.320 1.311
1.397 1.432 1.403 1.352
1.245 1.393 1.349
1.380 1.347
1.415
Fold Enrichment
1.0 1.2 1.4
B
109.4 163.9 274.7 294.6 369.5
98.4 223.0 315.7 375.3
6.7 148.8 252.9
54.8 205.2
123.5
-log(p-value)
80 240
C
DE genes close to DB
DE genes NOT close to DB
GO biological process
cardiac chamber development
cardiac chamber morphogenesis
cardiac muscle tissue development
cardiac ventricle development
cardiac ventricle morphogenesis
cardiovascular system development
circulatory system development
heart development
heart morphogenesis
translational elongation
translational initiation
translational termination
4.0 36.0
-log(p-value)
Day 0 vs 2
Day 0 vs 5
Day 0 vs 7
Day 0 vs 15
Day 0 vs 80
Day 2 vs 5
Day 2 vs 7
Day 2 vs 15
Day 2 vs 80
Day 5 vs 7
Day 5 vs 15
Day 5 vs 80
Day 7 vs 15
Day 7 vs 80
Day 15 vs 80
timepoints used in pairwise comparison
D
Day 15
Day 80
0 10
0 5
sqrt(count)
Hi-C count
score
boundary
genes
RNAseq
MYH6
MYH7
chr14
22,850,000 23,350,000 23,850,000 24,350,000

Figure 5. Differential expression near differential TGIF boundaries in human cardiomyocyte differentiation. **A.** Fold enrichment of differentially expressed (DE) genes within 50kb of differential TGIF boundaries from pairwise comparison of every pair of timepoints. **B.** Negative log p-value of significance from **A**. **D.** Negative log p-value of GO biological process enrichment for two groups of differentially expressed (DE) genes: DE genes close to (i.e., within 50kb of) differential boundaries (DB) for every pair of timepoints, and DE genes not close to DB. Only pairs of timepoints with at least one GO term with p-value $<$ 1e-5 is visualized here. **D.** A highly ranked differential boundary found in day 15 (primitive cardiomyocyte) but absent in day 80 (ventricular cardiomyocyte). This boundary is near *MYH6* and *MYH7* genes which are differentially expressed across the two timepoints.

Figure 6

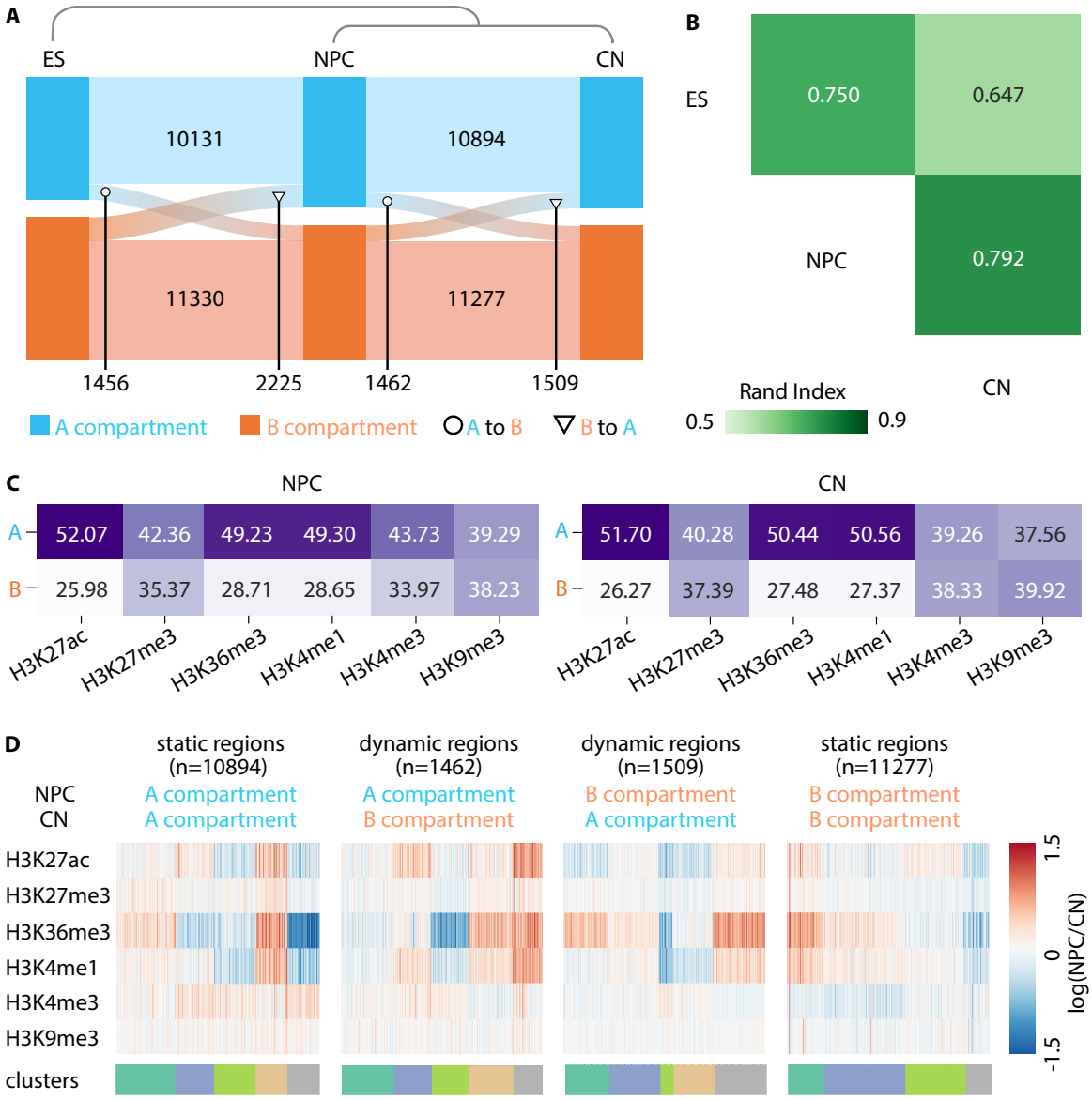

Figure 6. TGIF-DC for differential compartment analysis of mouse neural differentiation. **A.** Patterns of compartment switches throughout differentiation from mouse embryonic stem cell (ESC) state, to neural progenitors (NPC), and finally to cortical neurons (CN). **B.** Similarity of compartment assignments for every pair of timepoints measured by Rand index. **C.** Mean histone modification signal (read-depth and quantile normalized) in regions assigned to A and B compartment in NPC and in CN. **D.** Log fold change from NPC to CN, in different histone modification marks, for genomic regions grouped by their com-

partment switch pattern from NPC to CN. Within each group, regions are further clustered with k-means clusters ($k = 5$) to show patterns of histone modification changes associated with each compartment change patterns.

Figure 7

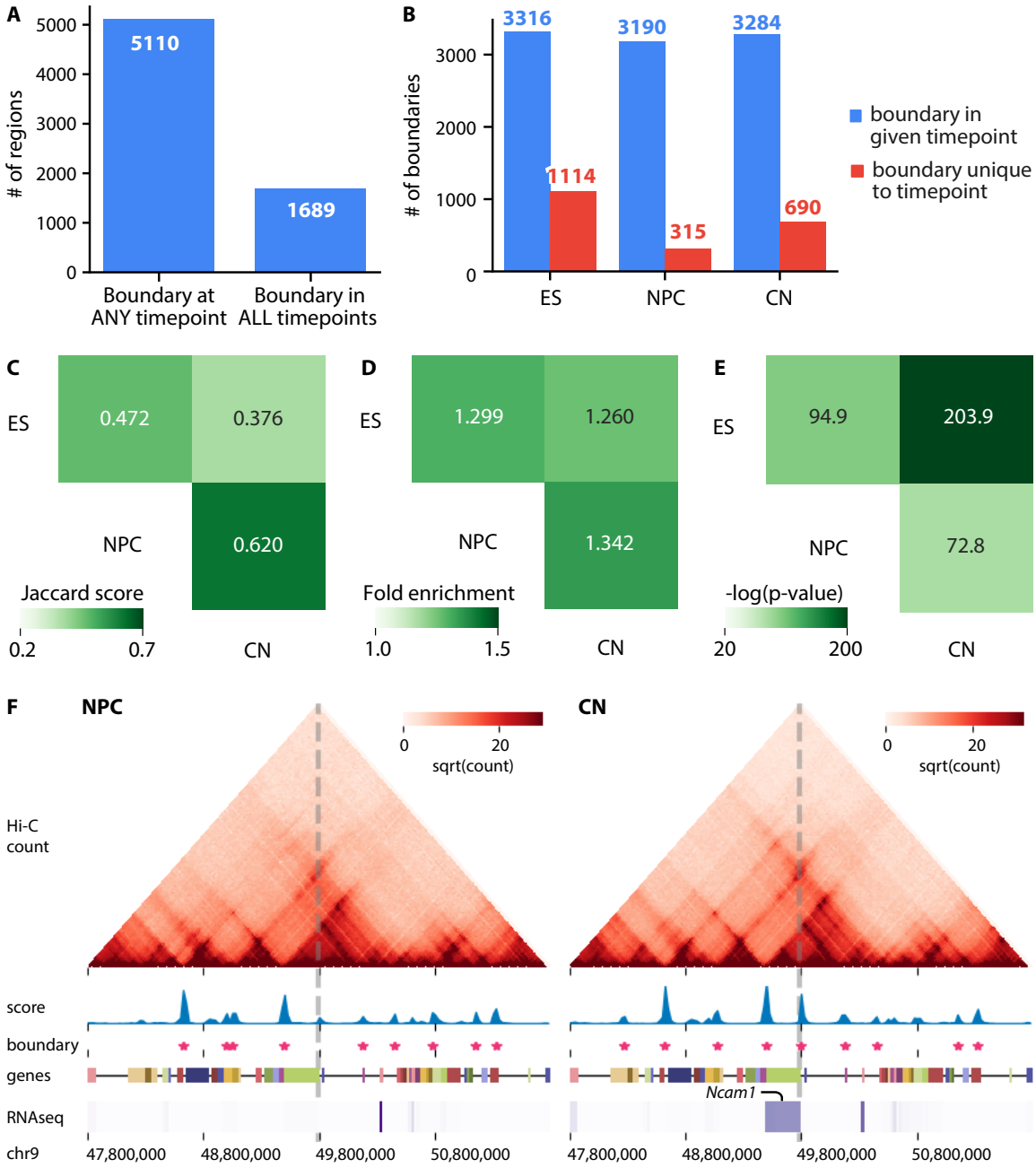

Figure 7. TGIF-DB boundary characterization in mouse neural differentiation. **A.** Number of boundaries in the union and intersection from all 3 timepoints. **B.** Total number of boundaries found in each timepoint (blue), and the unique to each timepoint (red). **C.** Overlap of boundary sets for every pair of

timepoints, as measured by Jaccard index. **D.** Fold enrichment of differentially expressed (DE) genes near differential TGIF boundaries from pairwise comparison of every pair of timepoints. **E.** Negative log p-value of significance from **D**. **F.** Example of a highly ranked differential boundary which is close to a differentially expressed gene, *Ncam1*. Shown are the Hi-C interaction matrices for the NPC and CN stages, the boundary score (blue), significant boundary (asterisk), a gene track and the expression heatmap from RNA-seq data. Differential expression of Ncam1 near the differential TGIF boundary that is absent in NPC state but appears in CN is shown.

Figure 8

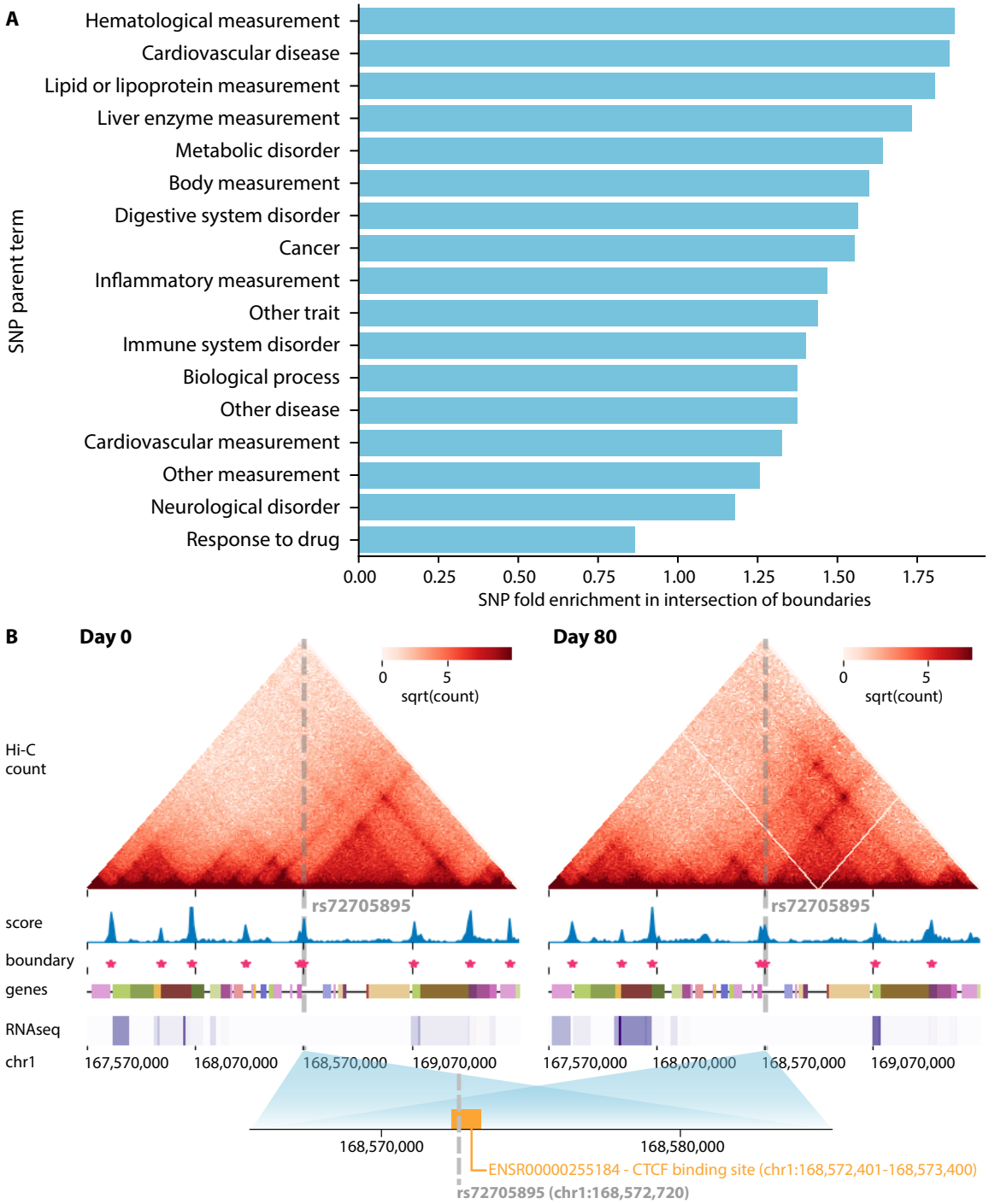

Figure 8. GWAS SNP enrichment in TGIF boundaries persistent throughout human cardiovascular differentiation. **A.** Fold enrichment of GWAS SNPs from different categories (i.e., parent term of the

associated trait) in persistent TGIF boundaries found in all timepoints during cardiovascular differentiation. **B.** A cardiovascular disease associated SNP (rs72705895) found in a shared TGIF boundary across all time points (day 0 - day 80). The SNP overlaps a region annotated with CTCF binding (zoomed in blue). The boundary score, significant boundary, gene track and RNA-seq expression are shown.

References

Akdemir KC, Le VT, Chandran S, Li Y, Verhaak RG, Beroukhim R, Campbell PJ, Chin L, Dixon JR, and Futreal PA. 2020. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nature Genetics* **52**: 294–305. Number: 3 Publisher: Nature Publishing Group.

Amjadi-Moheb F, Paniri A, and Akhavan-Niaki H. 2021. Insights into the Links between MYC and 3D Chromatin Structure and Epigenetics Regulation: Implications for Cancer Therapy. *Cancer Research* **81**: 1925–1936.

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. ArXiv:1307.0252 [cs, stat].

Beagan J, Duong M, Titus K, Zhou L, Cao Z, Ma J, Lachanski C, Gillis D, and Phillips-Cremins J. 2017. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Research* **27**: 1139–1152.

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.

Cheng LC, Pastrana E, Tavazoie M, and Doetsch F. 2009. miR-124 regulates adult neurogenesis in the subventricular zone stem cell niche. *Nature Neuroscience* **12**: 399–408.

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.

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.

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

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.

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

Dong X, Guo R, Ji T, Zhang J, Xu J, Li Y, Sheng Y, Wang Y, Fang K, Wen Y, et al.. 2022. YY1 safeguard multidimensional epigenetic landscape associated with extended pluripotency. *Nucleic Acids Research* p. gkac230.

Drews D, Klar M, Dame C, and Bräuer AU. 2009. Developmental expression profile of the yy2 gene in mice. *BMC Developmental Biology* **9**: 45.

Dubois F, Sidiropoulos N, Weischenfeldt J, and Beroukhim R. 2022. Structural variations in cancer and the 3D genome. *Nature Reviews Cancer* **22**: 533–546. Number: 9 Publisher: Nature Publishing Group.

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.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

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.

Fisch S, Gray S, Heymans S, Haldar SM, Wang B, Pfister O, Cui L, Kumar A, Lin Z, Sen-Banerjee S, et al.. 2007. Kruppel-like factor 15 is a regulator of cardiomyocyte hypertrophy. *Proceedings of the National Academy of Sciences* **104**: 7074–7079. Publisher: Proceedings of the National Academy of Sciences.

Fischer A, Steidl C, Wagner TU, Lang E, Jakob PM, Friedl P, Knobeloch KP, and Gessler M. 2007. Combined Loss of Hey1 and HeyL Causes Congenital Heart Defects Because of Impaired Epithelial to Mesenchymal Transition. *Circulation Research* **100**: 856–863. Publisher: American Heart Association.

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

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, et al.. 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **48**: D87–D92. Publisher: Oxford Academic.

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.

Gibb N, Lazic S, Yuan X, Deshwar AR, Leslie M, Wilson MD, and Scott IC. 2018. Hey2 regulates the size of the cardiac progenitor pool during vertebrate heart development. *Development (Cambridge, England)* **145**: dev167510.

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.

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

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

Ihara D, Watanabe Y, Seya D, Harada Y, Nakagawa O, and Kawamura T. 2018. The Mechanism of Hey2 Expression in Cardiac Development. *European Cardiology Review* **13**: 128.

Ito A and Imamura F. 2022. Expression of Maf family proteins in glutamatergic neurons of the mouse olfactory bulb. *Developmental neurobiology* **82**: 77–87.

Kalayeh MM, Idrees H, and Shah M. 2014. NMF-KNN: Image Annotation Using Weighted Multi-view Non-negative Matrix Factorization. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 184–191. ISSN: 1063-6919.

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.

Kieffer-Kwon KR, Nimura K, Rao SSP, Xu J, Jung S, Pekowska A, Dose M, Stevens E, Mathe E, Dong P, et al.. 2017. Myc Regulates Chromatin Decompaction and Nuclear Architecture during B Cell Activation. *Molecular Cell* **67**: 566–578.e10.

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.

Knoepfler PS, Zhang Xy, Cheng PF, Gafken PR, McMahon SB, and Eisenman RN. 2006. Myc influences global chromatin structure. *The EMBO Journal* **25**: 2723–2734.

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.

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. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**: 788–791. Number: 6755 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, 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.

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

Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, Myers Z, Sud P, Jou J, Lin K, et al.. 2020. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research* **48**: D882–D889.

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.

Makeyev EV, Zhang J, Carrasco MA, and Maniatis T. 2007. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Molecular Cell* **27**: 435–448.

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* .

Miao L, Li J, Li J, Tian X, Lu Y, Hu S, Shieh D, Kanai R, Zhou By, Zhou B, et al.. 2018. Notch signaling regulates Hey2 expression in a spatiotemporal dependent manner during cardiac morphogenesis and trabecular specification. *Scientific Reports* **8**: 2678. Number: 1 Publisher: Nature Publishing Group.

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**.

Pelham-Webb B, Murphy D, and Apostolou E. 2020. Dynamic 3D Chromatin Reorganization during Establishment and Maintenance of Pluripotency. *Stem Cell Reports* **15**: 1176–1195.

Păun O, Tan YX, Patel H, Strohbuecker S, Ghanate A, Cobolli-Gigli C, Sopena ML, Gerontogianni L, Goldstone R, Ang SL, et al.. 2023. Pioneer factor ASCL1 cooperates with the mSWI/SNF complex at distal regulatory elements to regulate human neural differentiation. *Genes & Development* **37**: 218–242. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory

Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

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.

Rochais F, Dandonneau M, Mesbah K, Jarry T, Mattei MG, and Kelly RG. 2009. Hes1 Is Expressed in the Second Heart Field and Is Required for Outflow Tract Development. *PLoS ONE* **4**: e6267.

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.

Silva JP, Lelianova VG, Ermolyuk YS, Vysokov N, Hitchen PG, Berninghausen O, Rahman MA, Zangrandi A, Fidalgo S, Tonevitsky AG, et al.. 2011. Latrophilin 1 and its endogenous ligand Lasso/teneurin-2 form a high-affinity transsynaptic receptor pair with signaling capabilities. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 12113–12118.

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.

Steimle JD and Moskowitz IP. 2017. TBX5: A Key Regulator of Heart Development. *Current Topics in Developmental Biology* **122**: 195–221.

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.

Szabo Q, Bantignies F, and Cavalli G. 2019. Principles of genome folding into topologically associating domains. *Science Advances* **5**: eaaw1668.

Tahmasebi S, Jafarnejad SM, Tam IS, Gonatopoulos-Pournatzis T, Matta-Camacho E, Tsukumo Y, Yanagiya A, Li W, Atlasi Y, Caron M, et al.. 2016. Control of embryonic stem cell self-renewal and differentiation via coordinated alternative splicing and translation of YY2. *Proceedings of the National Academy of Sciences* **113**: 12360–12367. Publisher: Proceedings of the National Academy of Sciences.

Vysokov NV, Silva JP, Lelianova VG, Suckling J, Cassidy J, Blackburn JK, Yankova N, Djamgoz MB, Kozlov SV, Tonevitsky AG, et al.. 2018. Proteolytically released Lasso/teneurin-2 induces axonal attraction by interacting with latrophilin-1 on axonal growth cones. *eLife* **7**: e37935.

Waldron L, Steimle JD, Greco TM, Gomez NC, Dorr KM, Kweon J, Temple B, Yang XH, Wilczewski CM, Davis IJ, et al.. 2016. The Cardiac TBX5 Interactome Reveals a Chromatin Remodeling Network Essential for Cardiac Septation. *Developmental Cell* **36**: 262–275.

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.

Xiao T, Li X, and Felsenfeld G. 2021. The Myc-associated zinc finger protein (MAZ) works together with CTCF to control cohesin positioning and genome organization. *Proceedings of the National Academy of Sciences of the United States of America* **118**: e2023127118.

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

Zhang Y, Beketaev I, Segura AM, Yu W, Xi Y, Chang J, Ma Y, and Wang J. 2020. Contribution of Increased Expression of Yin Yang 2 to Development of Cardiomyopathy. *Frontiers in Molecular Biosciences* **7**.

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.

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.

Zou SF, Yu Y, Peng Y, Xiao YB, Jian Z, Jiang YH, Chen S, Tang FQ, and Chen L. 2020. KLF15 is a protective regulatory factor of heart failure induced by pressure overload. *Molecular Medicine Reports* **21**: 1336–1345.