



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

Da-Inn Lee and Sushmita Roy

Genome Res. published online March 20, 2025

Access the most recent version at doi:[10.1101/gr.279930.124](https://doi.org/10.1101/gr.279930.124)

P<P	Published online March 20, 2025 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Creative Commons License	This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see https://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

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

Da-Inn Lee¹ and Sushmita Roy^{1,2*}

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

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

*To whom correspondence should be addressed.

Abstract

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

1 Introduction

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

20 Reliable detection of 3D genome dynamics at different units of organization is a significant com-
21 putational challenge. Current computational approaches to examine dynamics in the 3D genome can
22 be grouped into those that identify large-scale or compartmental-level changes (Fotuhi Siahpirani et al.,
23 2016; Chakraborty et al., 2022), those that can identify TAD-scale changes or “differential TADs” (Wang
24 et al., 2020; Cresswell and Dozmorov, 2020), and those that examine changes at the level of loops or
25 interactions (Ardakany et al., 2019; Lun and Smyth, 2015; Djekidel et al., 2018; Galan et al., 2020;
26 Stansfield et al., 2019). Compared to methods for detecting differences at the interaction level, there are
27 relatively few approaches to detect TAD or compartment changes. The most common approach to study
28 TAD dynamics across multiple conditions is to first apply a TAD-calling method to data from each con-
29 dition, followed by post-processing to identify TAD boundaries in one condition but not another (Zhang
30 et al., 2019; Bonev et al., 2017; Stadhouders et al., 2018; Wang et al., 2022; Emerson et al., 2022).
31 While such a two-step approach can identify some meaningful differences, the unsupervised nature of

32 TAD finding could make these approaches more susceptible to finding non-biological differences and
33 technical artifacts (Fletez-Brant et al., 2024; Kobets et al., 2023; Zheng et al., 2022). Numerous studies
34 have shown that Hi-C count profiles obey cell type, timepoint and species relationships, where datasets
35 from nearby contexts are more similar than those that are far away (Bonev et al., 2017; Zhang et al.,
36 2019; Yang et al., 2017; Vietri Rudan et al., 2015). An approach that constrains the TAD and compart-
37 ment finding based on such prior information about the relationships between the input datasets could
38 be less prone to spurious differences. A few methods have been developed to directly identify TAD
39 boundary differences, but they are focused on pairs of conditions (Wang et al., 2020) or limited in their
40 ability to compare more than two conditions (Cresswell and Dozmorov, 2020).

41 To address the dearth of methods for identifying large-scale organizational changes, especially when
42 considering more than two datasets, we developed Tree-Guided Integrated Factorization (TGIF). TGIF
43 is a multi-task, Non-negative Matrix Factorization (NMF) framework enabling joint embedding of mul-
44 tiple Hi-C matrices and identification of compartments and TADs across multiple conditions. NMF is
45 a popular dimensionality reduction approach for analyzing non-negative, genome-scale datasets (Stein-
46 O'Brien et al., 2018; Kotliar et al., 2019; Lee and Roy, 2021), where the low-dimensional factors capture
47 the biologically meaningful structure of the data. Multi-task NMF frameworks factorize multiple in-
48 put matrices simultaneously to yield low-dimensional embeddings in a shared latent space. They have
49 been successfully applied to single-cell omics data to integrate matrices from multiple samples, exper-
50 iments, and modalities by removing batch effect and technical noise (Welch et al., 2019; Liu et al.,
51 2020; Kriebel and Welch, 2022; Luecken et al., 2022; Hu et al., 2024). TGIF incorporates a hierarchical
52 multi-task NMF formulation to simultaneously factorize multiple Hi-C matrices from related biological
53 conditions and constrain the factors from closely related conditions to be similar. The factors represent
54 low-dimensional embeddings of genomic regions in an aligned latent space, representing their global
55 or local chromatin architecture. We use such embeddings to identify changes at both the compartment
56 and TAD levels. When applied to simulated and real timecourse Hi-C matrices, TGIF identifies fewer
57 false positive differences in TAD boundaries, and produces a more reproducible set of boundaries across
58 biological replicates, normalization methods, depths, and resolutions compared to other methods. Dif-
59 ferential boundaries and compartmental regions identified by TGIF show significant changes in relevant
60 biological signals such as gene expression, histone modification, and chromatin accessibility. Finally,
61 persistent boundaries identified by TGIF are enriched in sequence variants associated with cardiovascu-
62 lar disease. Together, these results demonstrate the versatility and utility of TGIF to examine changes in
63 higher-order 3D genome organization across diverse types of dynamic processes.

Results

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

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

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

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

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

96 In cases where the relationship between the input Hi-C data is not available (e.g. integrating Hi-
97 C datasets from multiple studies or pseudo-bulk single-cell Hi-C data from cell clusters; Zhou et al.,
98 2019; Zhang et al., 2022) TGIF can infer a tree structure based on the pairwise similarity of the input
99 Hi-C matrices measured by stratum-adjusted correlation coefficient (SCC; Yang et al., 2017, **Methods**,
100 **Figure S2**) or a similar distance-stratified metric.

101 **TGIF-DB identifies fewer false-positive differential boundaries in simulated and** 102 **real Hi-C data.**

103 TGIF-DB was benchmarked against four other TAD calling methods: three methods designed for calling
104 TADs and boundaries from a single Hi-C matrix (which we refer to as single-task methods), and one
105 designed specifically for differential boundary identification (**Methods**). The three single-task methods
106 were GRINCH (Lee and Roy, 2021), SpectralTAD (Cresswell et al., 2020), and TopDom (Shin et al.,
107 2016, **Supplemental Methods**). TADCompare (Cresswell and Dozmorov, 2020) is a method designed
108 for differential boundary detection.

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

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

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

129 We next measured the Jaccard score between the boundary sets from a pair of biological replicates
130 of H1 human embryonic stem cell line (Zhang et al., 2019, **Methods**), TADCompare and TGIF-DB had
131 the highest scores, recovering more similar set of boundaries between the biological replicates compared
132 to other methods (**Figure 2C**). Furthermore, differential boundaries between two timepoints (day 0 and
133 day 2 of cardiomyocyte differentiation) identified by TADCompare and TGIF-DB had the fewest false
134 positives based on overlap with differential boundaries between two biological replicates of the same
135 timepoint (**Supplemental Methods, Figure S3B**). We also benchmarked the degree of false-positive,
136 non-biological differences identified by each method in: (1) Hi-C datasets with different depths, (2)
137 datasets from biological replicates, (3) datasets normalized using different methods, and (4) datasets in
138 different bin resolutions. To this end, we downsampled a high-depth Hi-C dataset from the GM12878
139 cell line with 4.01 billion reads (Rao et al., 2014; Reiff et al., 2022) by subsampling 5, 10, 25, 50% of
140 the reads (**Methods**). We measured the Jaccard score between the boundary sets from the original high-
141 depth input and the downsampled counterpart. The higher the Jaccard index, the fewer the false-positive
142 differences identified by a method. Across all downsampled depths, TADCompare and TGIF-DB were
143 the top performing methods with consistently high Jaccard scores (**Figure 2D**). Single-task methods
144 (GRiNCH, SpectralTAD, and TopDom) had much lower Jaccard score with discrepancy increasing with
145 depth differences. We observed similar results with TADCompare and TGIF-DB obtaining the highest
146 Jaccard score between TAD boundary sets from mouse embryonic stem cell (mESC) Hi-C data nor-
147 malized using different methods (**Figure 2E, Methods**). Finally, we measured the stability of TAD
148 boundaries to the changing resolution (10kb, 25kb, 50kb) of input Hi-C matrices using Jaccard score
149 (**Methods**). TGIF-DB and GRiNCH yield the most stable or similar boundaries to changing resolu-
150 tion (**Figure S3C**), with the exception of 25kb-50kb where TopDom also performed well. These results
151 demonstrate the advantages of using TGIF-DB to identify biologically relevant boundaries enriched in
152 known boundary elements while minimizing false positive differences.

153 **TGIF-DC identifies compartment dynamics that are significantly enriched for dif-** 154 **ferential regulatory signals**

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

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

158 We first compared the similarity of compartment assignments between different methods using Rand
159 Index (**Figure 3A**). The PCA-based method and dcHiC, which also utilizes PCA, produced the most
160 similar compartments (Rand Index: 0.91), followed by TGIF-DC (Rand Index: 0.79-0.8). Cscore found
161 a substantially different set of compartments (Rand Index: 0.52). We assessed the quality of compart-
162 ments with three cluster quality metrics, Silhouette Index (SI, **Figure 3B**), Calinski-Harabasz score (CH,
163 **Figure 3C**), and Davies-Bouldin Index (DBI, **Figure 3D**), using observed-over-expected (O/E) counts as
164 features of each genomic loci (**Methods**). In all three metrics, TGIF-DC, dcHiC, and PCA-based com-
165 partments are comparable in their quality and outperformed Cscore. We also measured compartment
166 quality using chromatin accessibility, a key regulatory measurement that characterizes different com-
167 partment types (e.g., the active A and repressive B; Lieberman-Aiden et al., 2009; Fortin and Hansen,
168 2015). Briefly, we measured SI (**Figure 3E**), CH (**Figure 3F**) and DBI (**Figure 3G**) using the mean
169 basepair ATAC-seq signal for each 100kb region as the feature (**Methods**). For all three metrics, the
170 compartments from TGIF-DC, PCA-based method, and dcHiC are of similar quality.

171 Finally, we compared TGIF-DC exclusively with dcHiC, the only other method that specifically iden-
172 tifies *differential* compartment regions. Significantly differential compartmental regions (sigDC) identi-
173 fied by TGIF-DC have significantly higher change in accessibility signal and gene expression compared
174 to regions not part of sigDC (**Figure S5A,B**). Compared to significantly differential regions identified by
175 dcHiC, sigDC regions from TGIF-DC also have significantly higher change in accessibility signal (t -test
176 p -value $< 1e-2$, **Figure 3H**) and are comparable in terms of the change in gene expression levels (**Figure**
177 **3I**).

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

181 **TGIF-DC offers a unified framework to identify both compartment and subcom-** 182 **partment dynamics**

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

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

197 We next applied TGIF-DC with $k = 5$ to identify subcompartment structure per chromosome, each k
198 corresponding to a different subcompartments (**Methods**). We interpreted these subcompartments based
199 on the mean histone modification signal of the genomic loci assigned to each subcompartment. The
200 subcompartments exhibited distinct histone modification patterns (**Figure 4C, Chr18**), with subcom-
201 partments 1 and 5 associated with repressive marks (H3K9me3, H3K27me3), while the other three (2,
202 3 and 4) associated with active marks. Within these two groups, each subcompartment had a different
203 signature of marks. For example, subcompartment 3 exhibits relatively lower signal of H3K36me3 com-
204 pared to 2 and 4, while subcompartment 2 had a higher signal of all three activating marks (H3K27ac,
205 H3K36me3, H3K4me1) compared to 3 and 4. Between the two subcompartments, 1 and 5, with repres-
206 sive mark association, one (1) exhibited higher H3K4me3 and H3K9me3 levels compared to the other
207 one (5).

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

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

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

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

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

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

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

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

268 Among other top-ranked sigDBs in cardiomyocyte differentiation, we found DE genes with known
269 roles in the cardiac development. For example, a boundary was found in primitive cardiomyocytes (day
270 15) but absent in ventricular cardiomyocytes (day 80, **Figure S13A**). This boundary overlaps *MYH6*,
271 highly expressed in day 15 compared to day 80, and is adjacent to *MYH7*, displaying the opposite ex-
272 pression change pattern to *MYH6*. Both genes are involved in cardiac muscle function (Ching et al., 2005;
273 Warkman et al., 2012). Recently an enhancer cluster located downstream to *MYH7* at Chr14:23,876,121-
274 23,878,188, was identified as a switch that can downregulate expression of *MYH7* while upregulating
275 *MYH6* (Gacita et al., 2021). We also identified a sigDB close to the *Ncam1* gene which is differentially
276 expressed between ES and CN (**Figure S13B**); *Ncam1* has known roles in neuron axon guidance and
277 synapse formation (Hata et al., 2018; Shetty et al., 2013). These examples provide further evidence for
278 TGIF-DB's ability to identify relevant dynamic boundaries that could impact overall cell state identity.

Persistent boundaries are enriched for SNPs from diverse disease phenotypes

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

Discussion

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

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

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

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

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

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

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

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

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

Materials and Methods

Tree-Guided Integrated Factorization (TGIF)

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

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

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

The objective aims to:

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

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

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

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

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

411 **TGIF-DB for differential boundary identification**

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

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

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

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

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

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

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

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

465 < 0.05 as significantly differential boundary (sigDB) regions. We further annotate the type of change
 466 represented by each sigDB between conditions A and B: a boundary created in B, deleted in B, or shifted
 467 in B within 5 genomic bins (**Supplemental Methods**).

468 **TGIF-DC for differential compartment and subcompartment identification**

469 **Identification of compartments with TGIF-DC.** In order to identify compartments, we apply
 470 TGIF to a 100kb resolution intrachromosomal Hi-C matrix that is first converted into an observed-over-
 471 expected (O/E) count correlation matrix as described previously in Rao et al., 2014. We upshift the
 472 correlation matrix by 1 so that all values are non-negative. To identify compartments, we apply TGIF
 473 with the input tree structure to these matrices with rank $k = 2$. After factorization, we infer each
 474 region i 's cluster assignment, $c_i^{(t)}$, for each task t , such that $c_i^{(t)} = \operatorname{argmax}_{j \in \{1,2\}} U[i, j]$. We refer to
 475 these clusters as compartments. To identify subcompartments and differential subcompartment regions,
 476 a higher k value, e.g. 5, can be used and TGIF-DC will generate more granular cluster assignments, e.g.
 477 5 clusters of regions instead of 2 clusters. Each of these clusters corresponds to a subcompartment.

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

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

494 **Estimating tree structure from input Hi-C matrices for unknown inter-dataset** 495 **relationships**

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

507 **Datasets used in analysis**

508 We applied TGIF to three Hi-C timecourse datasets: H1 hESC differentiated to endoderm (Reiff et al.,
509 2022; Dekker et al., 2023), mouse neural differentiation data from Bonev et al., 2017, and human car-
510 diomyocyte differentiation data from Zhang et al., 2019. See **Supplemental Methods, Figure S4**, and
511 **Table S1-Table S4** for processing, application of TGIF, and list of accession numbers.

512 **Benchmarking methods for identifying differential domain boundaries**

513 **Existing methods used in benchmarking TGIF-DB** TGIF-DB was benchmarked against four
514 other methods for identifying differential TAD boundaries: GRINCH (Lee and Roy, 2021), Spectral-
515 TAD (Cresswell et al., 2020), TADCompare (Cresswell and Dozmorov, 2020), and TopDom (Shin et al.,
516 2016). GRiNCH, SpectralTAD, and TopDom are single-task TAD identification methods accepting a
517 single input matrix individually followed by pairwise comparison of identified boundaries. TADCom-
518 pare is a differential TAD identification method that can take as input a pair of Hi-C matrices as well as
519 a time series of Hi-C matrices. These methods are described in more detail in **Supplemental Methods**,
520 **Figure S22**.

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

540 **Measuring CTCF enrichment in boundaries** To evaluate the boundaries identified by various
541 TAD-calling methods, we measured CTCF peak enrichment in boundaries found in the cardiomyocyte
542 differentiation dataset (**Table S3**). Using MACS2 (Zhang et al., 2008), we first called peaks on CTCF
543 ChIP-seq data from each of the 6 timepoints (day 0, 2, 5, 7, 15, 80) of the cardiomyocyte differentiation
544 time course. Replicates from each timepoint were collapsed by intersecting overlapping peaks with
545 BEDTools (Quinlan and Hall, 2010). Each peak was then assigned to a 10kb uniform bin again using
546 BEDTools. TAD-calling methods GRiNCH, SpectralTAD, and TopDom were applied to 10kb Hi-C
547 matrices from each of the 6 timepoints. TGIF-DB was applied to Hi-C matrices from all 6 timepoints
548 using the tree structure as in **Figure S4**, and significant boundaries from each timepoint were used for
549 enrichment analysis. TADCompare was applied to each pair of consecutive timepoints: day 0 vs 2, 2 vs
550 5, 5 vs 7, 7 vs 15, 15 vs 80. As TADCompare outputs both non-differential and differential boundaries
551 for every pairwise comparison, we define a boundary set specific to a timepoint as follows: (1) for day

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

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

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

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

595 **Comparison of TGIF-DC to existing compartment-calling methods**

596 We compared TGIF-DC to two established methods for calling compartments, i.e. principal component
597 analysis (PCA) based method (Lieberman-Aiden et al., 2009) and Cscore (version 1.1, Zheng and Zheng,
598 2018), as well as a method designed specifically for differential compartment analysis, dcHiC (version
599 2.1, Chakraborty et al., 2022). Each method is described in detail in **Supplemental Methods, Figure**
600 **S23**. We applied all four methods to 100kb intra-chromosomal count matrices from H1 hESC cell
601 line. TGIF-DC and dcHiC were additionally applied to 100kb intra-chromosomal count matrices from
602 H1 differentiated to endoderm. Both datasets were downloaded from 4D Nucleome consortium (Reiff
603 et al., 2022; Dekker et al., 2023). To compare the compartment results across the different methods,
604 we measured the Rand index between compartment assignments to each genomic region. To measure
605 the quality of the compartments, we used three well-known cluster quality metrics: Silhouette Index,
606 Calinski-Harabasz Score, and Davies-Bouldin Index, measured on the observed-over-expected (O/E)
607 matrices for each chromosome, as well as the accessibility signal for each 100kb genomic region. The
608 accessibility signal was defined as the mean ATAC-seq reads per basepair. Finally, to compare dcHiC
609 and TGIF-DC for significantly differential compartments between H1 and endoderm, we calculated the
610 log ratio of the accessibility signal and gene expression (from RNA-seq, in TPM) in H1 over that of
611 endoderm for each significantly differential region.

612 **Assessing differential gene expression near or within significantly differential bound-** 613 **aries and compartments**

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

621 For every pair of timepoints, we tested the enrichment of these DE genes within regions of inter-
 622 est (**Figure 5A**): (A) regions near (i.e., within 100kb) significantly differential boundaries (sigDB), (B)
 623 regions within a TAD with at least one sigDB, and (C) regions within significantly differential compart-
 624 mental regions (sigDC). For (B), we define all regions bounded within a pair of shared boundaries and
 625 containing at least one sigDB within those bounds as belonging to a “TAD with at least one sigDB”.

626 The fold enrichment of DE genes in these regions was computed as $\frac{q}{M} / \frac{s}{N}$, where N is number of all
 627 regions, $s = |\text{set of regions with at least one DE gene}|$, $M = |\text{a subset regions of interest as defined}$
 628 $\text{above, e.g., regions near sigDB}|$, $q = |\text{regions of interest with at least one DE gene, e.g., regions near}$
 629 $\text{sigDB with a DE gene}|$. We also performed gene-centric fold enrichment calculations: $\frac{q_g}{M_g} / \frac{s_g}{N_g}$, where
 630 N_g is total number of genes with expression, $s_g = |\text{DE genes}|$, $M_g = |\text{genes overlapping with a}$
 631 $\text{regions of interest, e.g., region near sigDB}|$, $q_g = |\text{DE genes overlapping with a regions of interest}|$.
 632 Hypergeometric test was additionally performed to calculate the significance of this fold enrichment
 633 value for each pair of timepoints. We additionally examined the correlation between a gene’s RNA-seq
 634 fold change and the raw boundary score change (positive or negative) of its nearest sigDB in the H1-
 635 endoderm dataset. We found little to no correlation in the magnitude or the direction of change in gene
 636 expression and boundary strength (Pearson’s corr = 0.03, **Figure S17**).

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

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

645 **SNP enrichment within TGIF boundaries from cardiomyocyte differentiation data**

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

657 **Software availability**

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

661 **Competing Interests**

662 Authors have no competing interests.

663 **Acknowledgements**

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

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

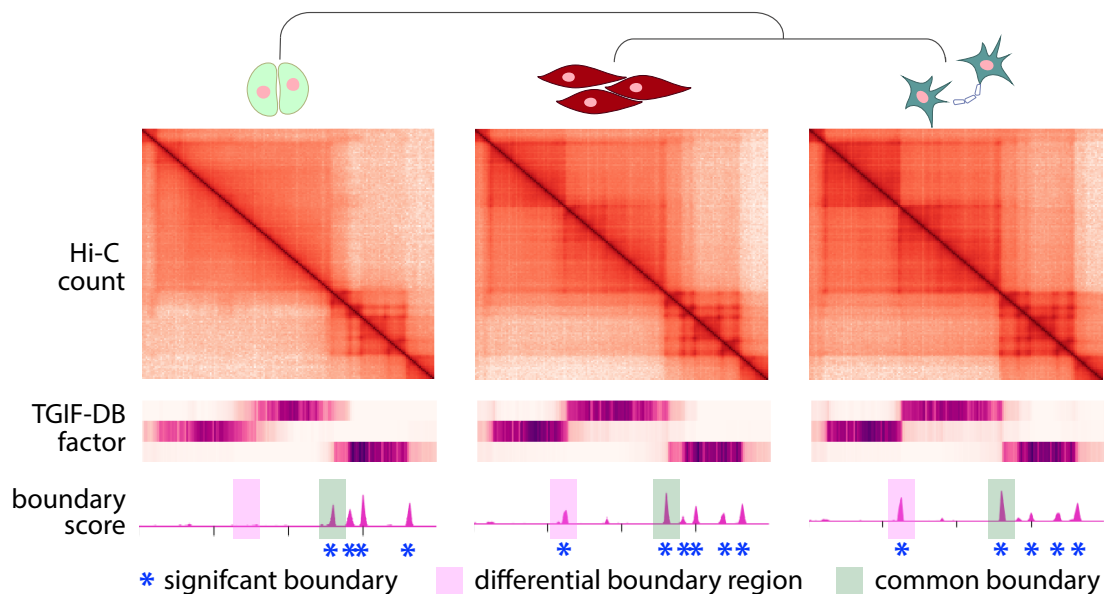
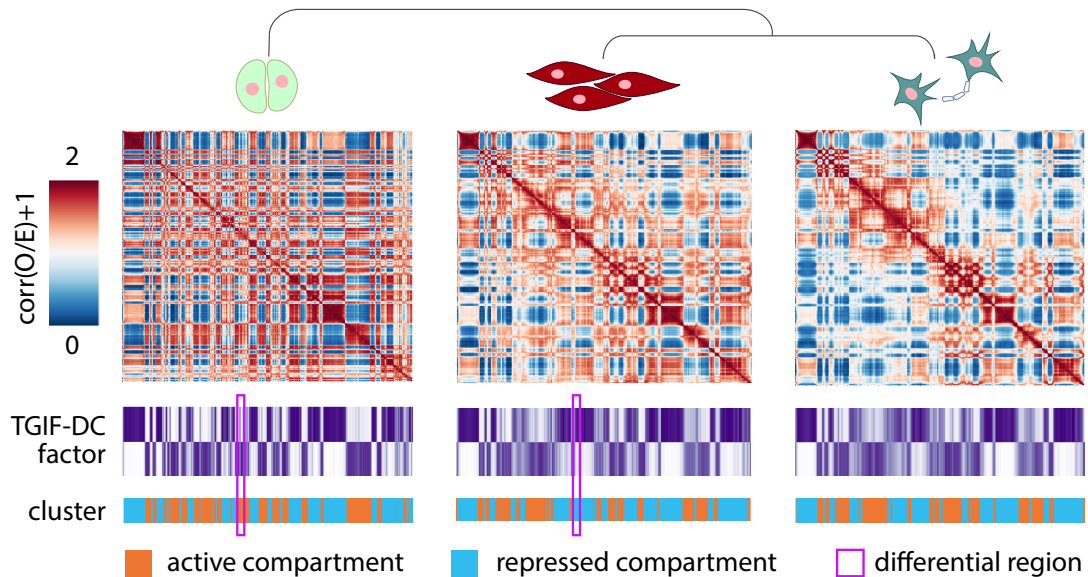
669 **Author contributions**

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

673

Figures

674

Figure 1**A** TGIF-DB: differential TAD boundary analysis**B** TGIF-DC: differential compartment analysis

675

676

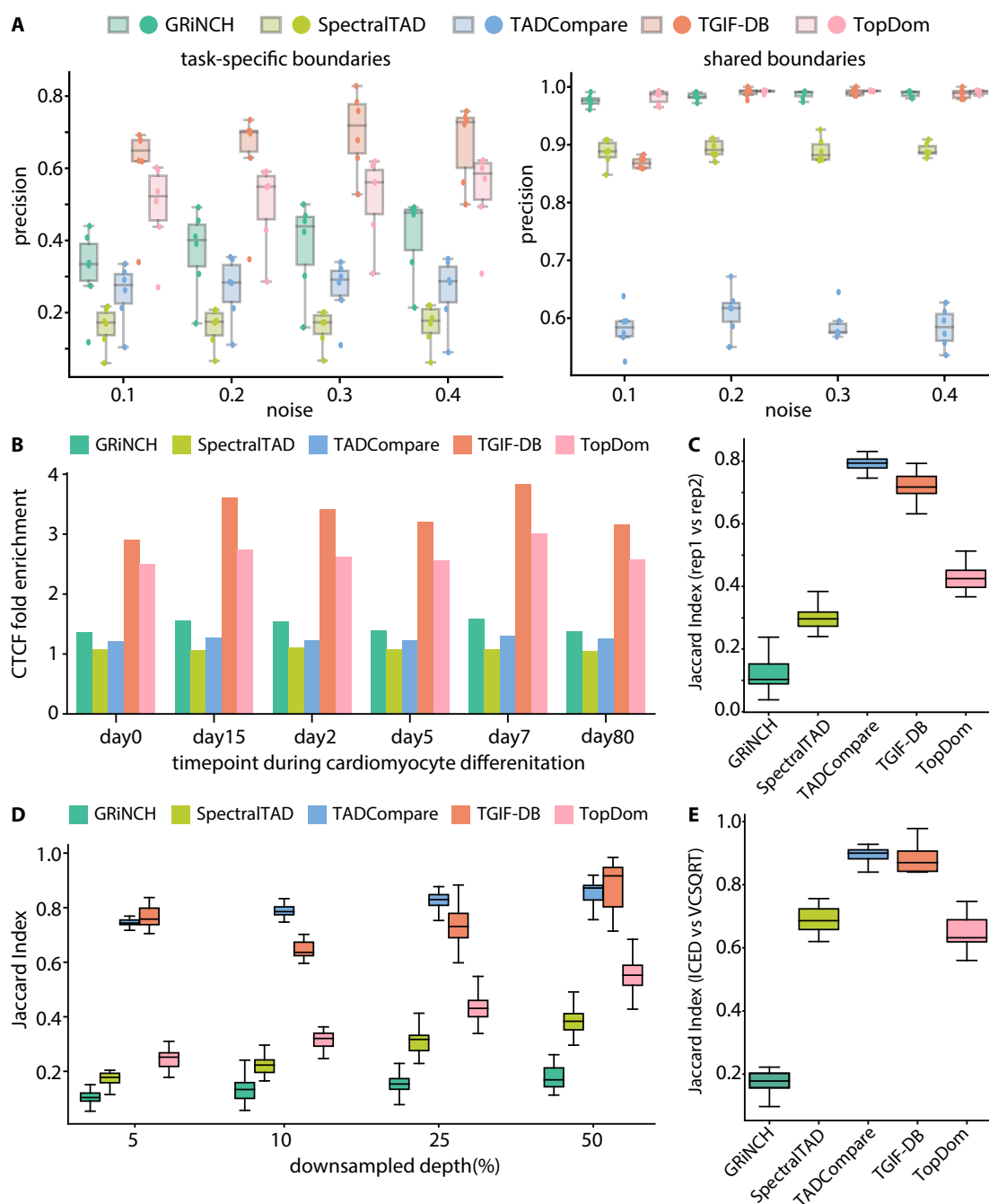
677

678

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

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

686

Figure 2

687

688

689

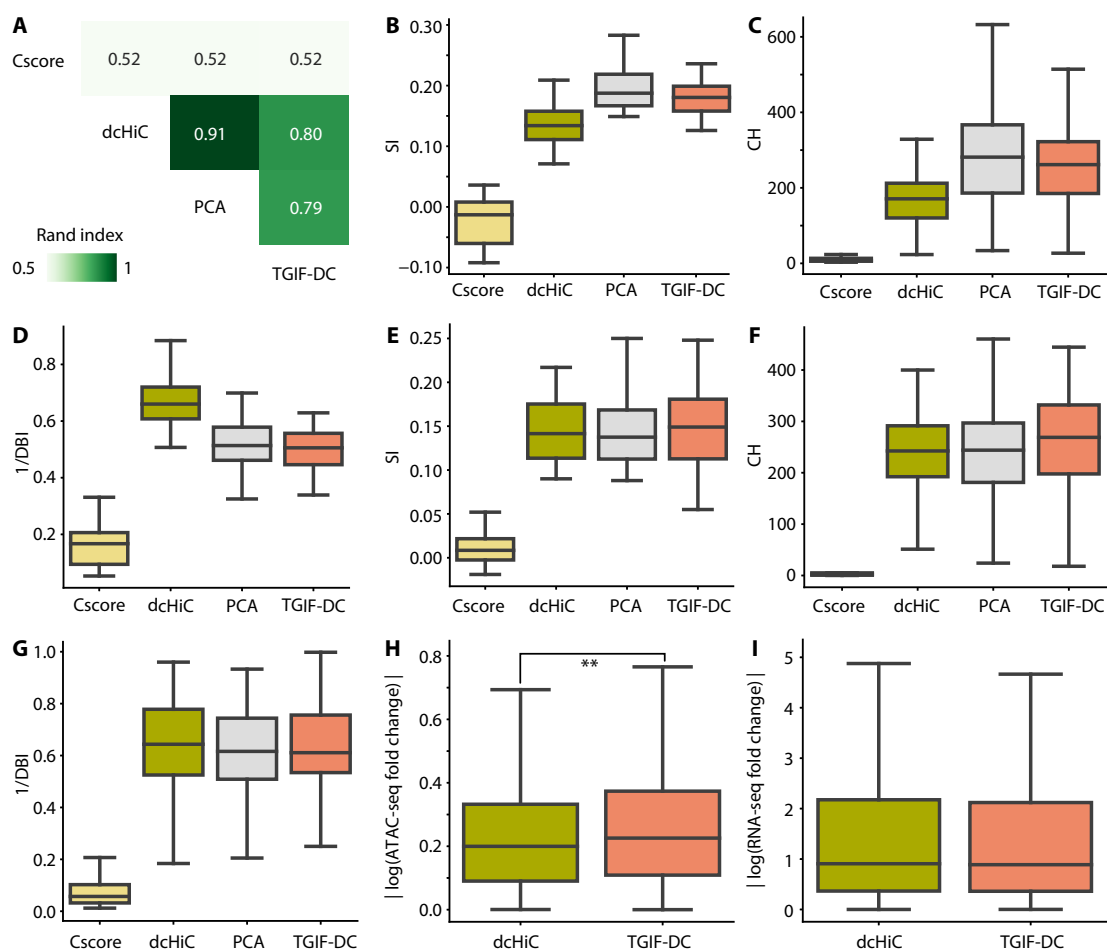
690

691

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

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

697

Figure 3

698

699

700

701

702

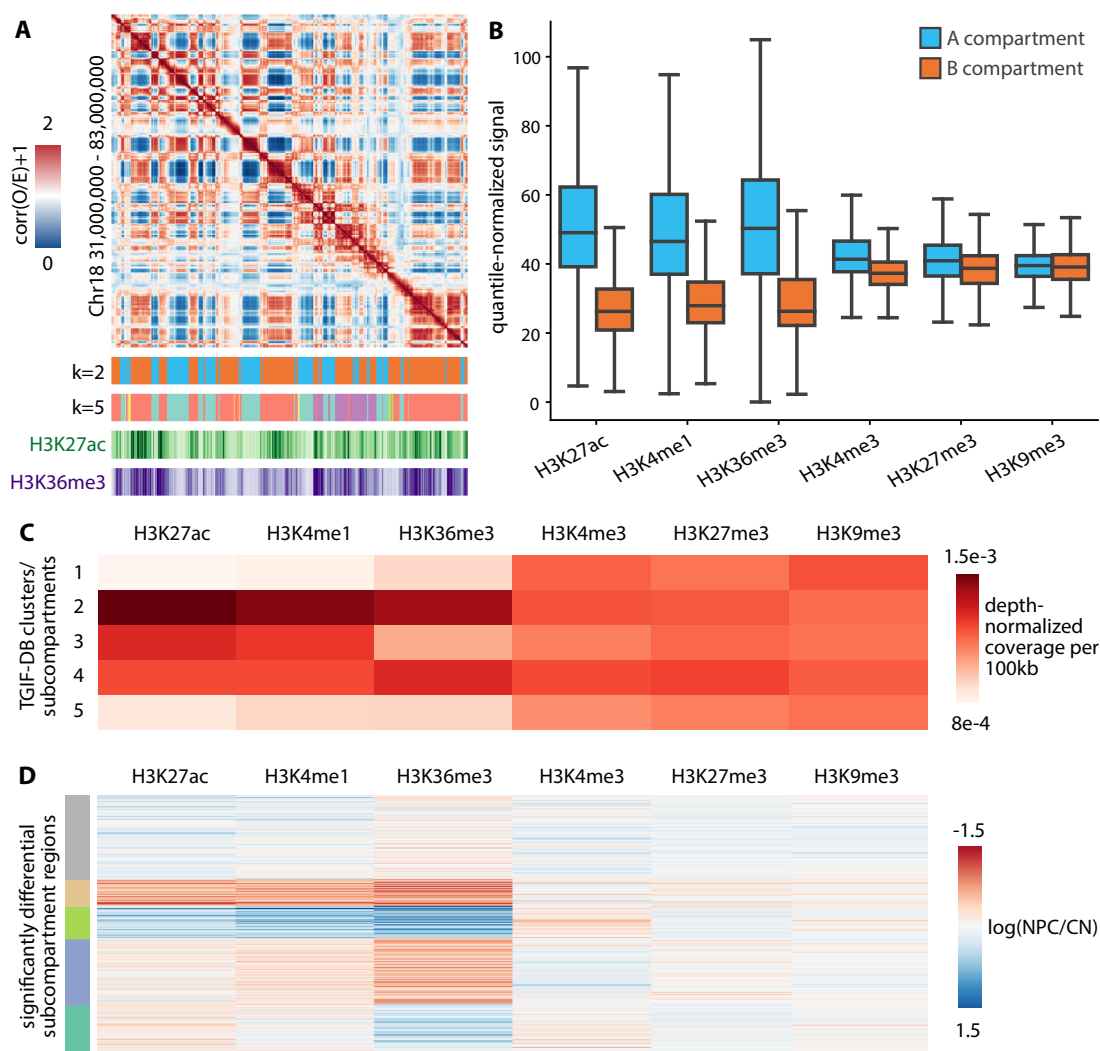
703

704

705

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

706

Figure 4

707

708

709

710

711

712

713

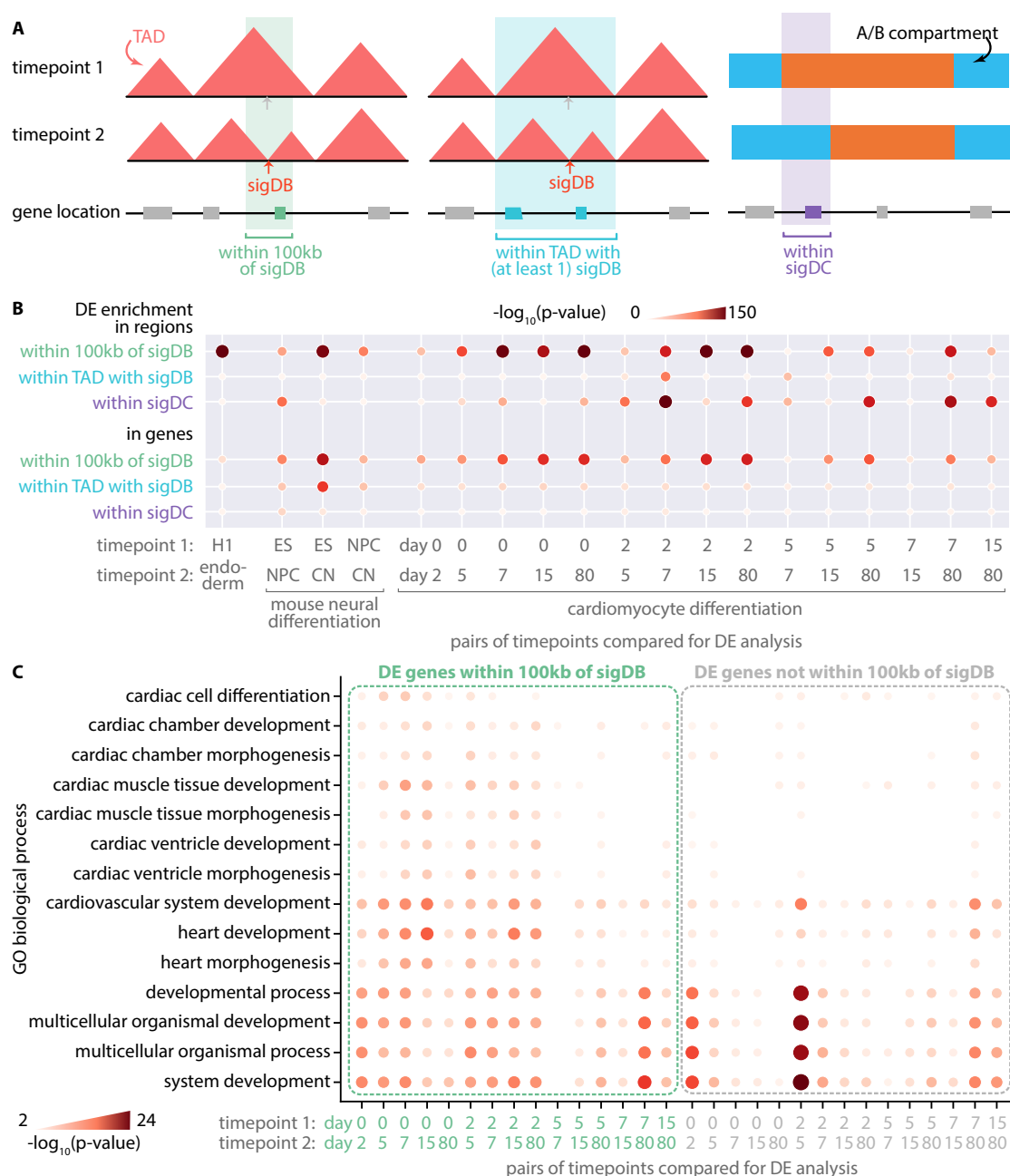
714

715

716

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

717

Figure 5

718

719

720

721

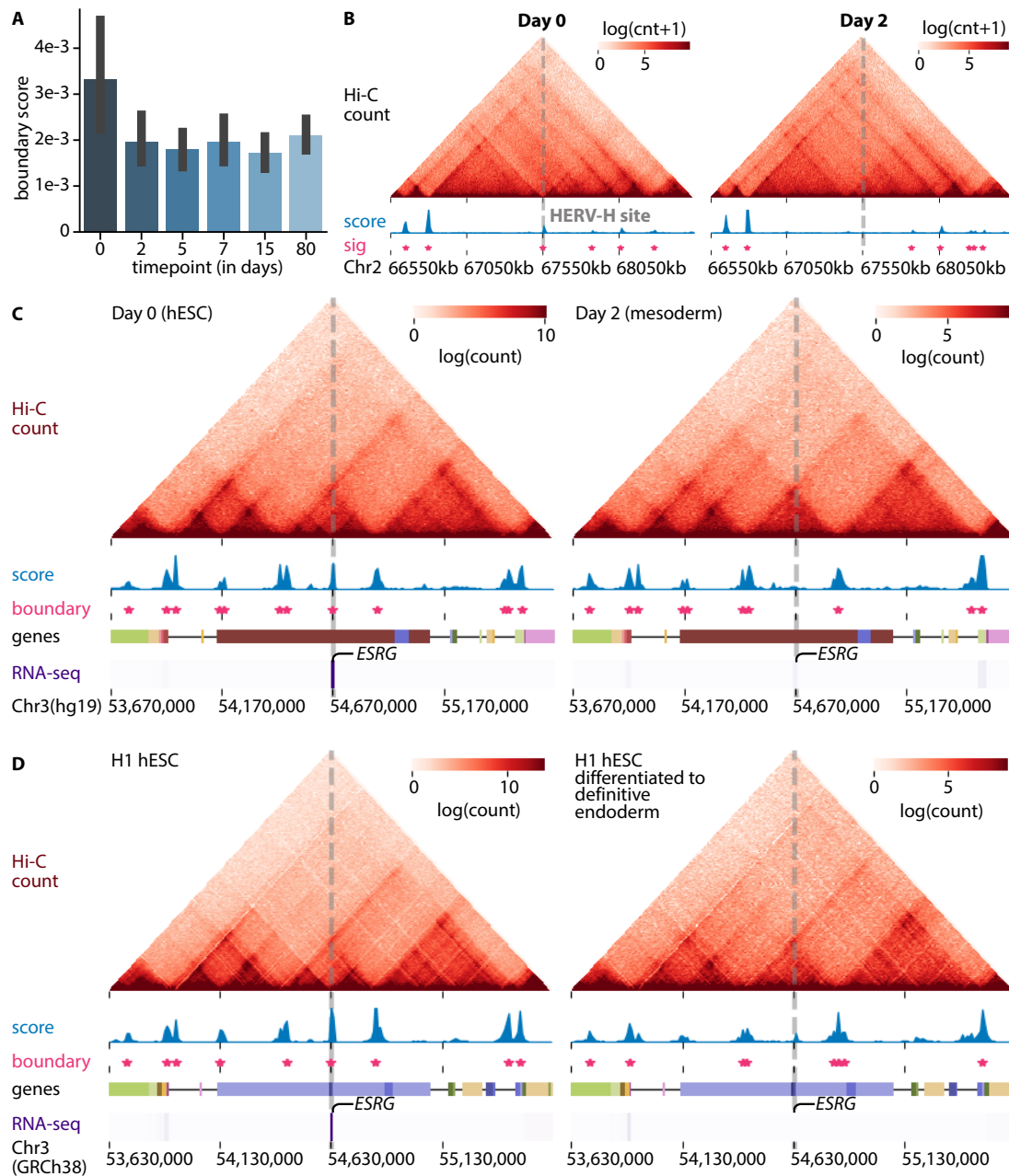
722

723

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

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

728

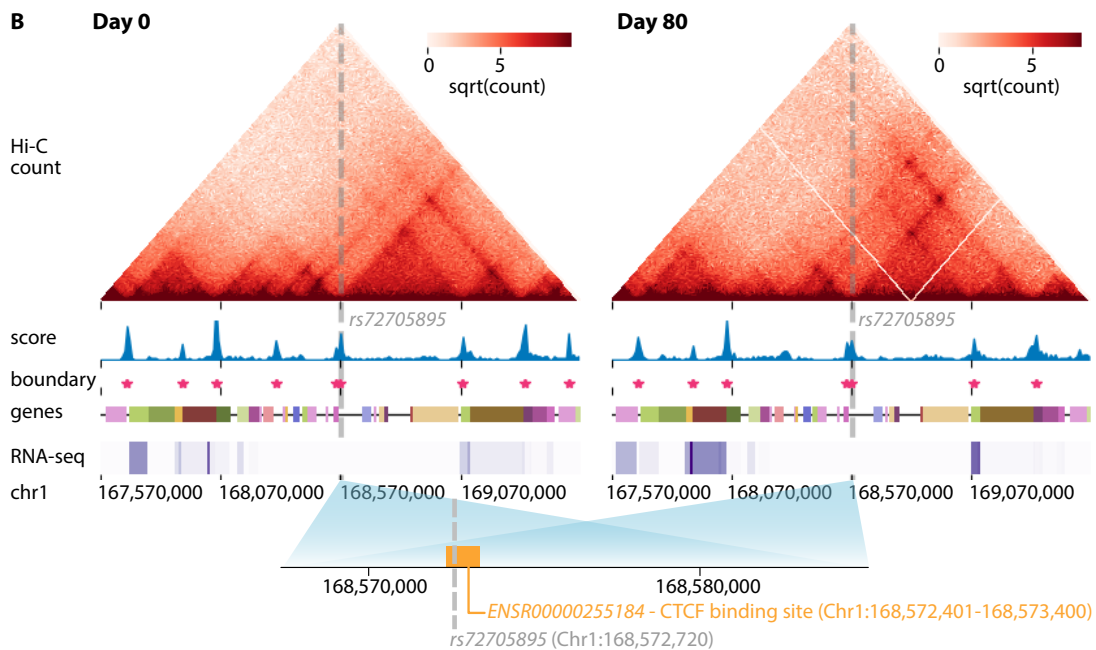
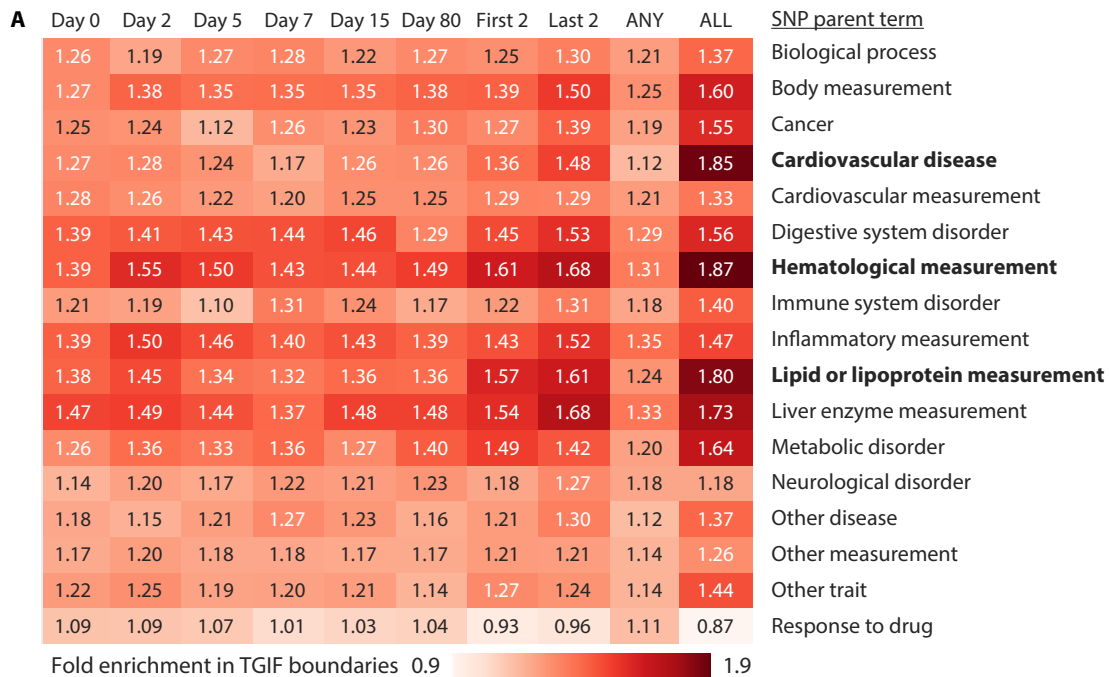
Figure 6

729

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

734

735

Figure 7

736

737

738

739

740

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

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

References

743

744

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

745

746

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

747

748

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

749

750

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].

751

752

753

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

754

755

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.

756

757

758

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

759

760

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.

761

762

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

763

764

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.

765

766

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

767

768

- 769 Ching YH, Ghosh TK, Cross SJ, Packham EA, Honeyman L, Loughna S, Robinson TE, Dearlove AM,
770 Ribas G, Bonser AJ, et al.. 2005. Mutation in myosin heavy chain 6 causes atrial septal defect. *Nature*
771 *Genetics* **37**: 423–428. Number: 4 Publisher: Nature Publishing Group.
- 772 Cresswell KG and Dozmorov MG. 2020. TADCompare: An R Package for Differential and Temporal
773 Analysis of Topologically Associated Domains. *Frontiers in Genetics* **11**. Publisher: Frontiers.
- 774 Cresswell KG, Stansfield JC, and Dozmorov MG. 2020. SpectralTAD: an R package for defining a
775 hierarchy of topologically associated domains using spectral clustering. *BMC Bioinformatics* **21**:
776 319.
- 777 Cubeñas-Potts C and Corces VG. 2015. Architectural Proteins, Transcription, and the Three-dimensional
778 Organization of the Genome. *FEBS letters* **589**: 2923–2930.
- 779 Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O,
780 Azov AG, Barnes I, Bennett R, et al.. 2022. Ensembl 2022. *Nucleic Acids Research* **50**: D988–D995.
- 781 Dekker J, Alber F, Aufmkolk S, Beliveau BJ, Bruneau BG, Belmont AS, Bintu L, Boettiger A, Calan-
782 drelli R, Distèche CM, et al.. 2023. Spatial and temporal organization of the genome: Current state
783 and future aims of the 4D nucleome project. *Molecular Cell* **83**: 2624–2640. Publisher: Elsevier.
- 784 Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O’Shea CC, Park
785 PJ, Ren B, et al.. 2017. The 4D nucleome project. *Nature* **549**: 219–226. Number: 7671 Publisher:
786 Nature Publishing Group.
- 787 Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, and Ren B. 2012. Topological domains
788 in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376–380.
- 789 Djekidel MN, Chen Y, and Zhang MQ. 2018. FIND: difFerential chromatin INteractions Detection using
790 a spatial Poisson process. *Genome Research* **28**: 412–422.
- 791 Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, and Aiden EL. 2016. Juicer
792 Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**:
793 95–98. Publisher: Elsevier.
- 794 Emerson DJ, Zhao PA, Cook AL, Barnett RJ, Klein KN, Saulebekova D, Ge C, Zhou L, Simandi Z,
795 Minsk MK, et al.. 2022. Cohesin-mediated loop anchors confine the locations of human replication
796 origins. *Nature* **606**: 812–819. Number: 7915 Publisher: Nature Publishing Group.

- 797 Eres IE, Luo K, Hsiao CJ, Blake LE, and Gilad Y. 2019. Reorganization of 3D genome structure may
798 contribute to gene regulatory evolution in primates. *PLOS Genetics* **15**: e1008278.
- 799 Espinola SM, Götz M, Bellec M, Messina O, Fiche JB, Houbron C, Dejean M, Reim I, Cardozo Gizzi
800 AM, Lagha M, et al.. 2021. Cis -regulatory chromatin loops arise before TADs and gene activation,
801 and are independent of cell fate during early *Drosophila* development. *Nature Genetics* **53**: 477–486.
802 Number: 4 Publisher: Nature Publishing Group.
- 803 Fletez-Brant K, Qiu Y, Gorkin DU, Hu M, and Hansen KD. 2024. Removing unwanted variation between
804 samples in Hi-C experiments. *Briefings in Bioinformatics* **25**: bbae217.
- 805 Forcato M, Nicoletti C, Pal K, Livi C, Ferrari F, and Bicciato S. 2017. Comparison of computational
806 methods for Hi-C data analysis. *Nature Methods* **14**: 679–685.
- 807 Fortin JP and Hansen KD. 2015. Reconstructing A/B compartments as revealed by Hi-C using long-
808 range correlations in epigenetic data. *Genome Biology* **16**: 180.
- 809 Fotuhi Siahpirani A, Ay F, and Roy S. 2016. A multi-task graph-clustering approach for chromosome
810 conformation capture data sets identifies conserved modules of chromosomal interactions. *Genome*
811 *Biology* **17**: 114.
- 812 Gacita AM, Fullenkamp DE, Ohiri J, Pottinger T, Puckelwartz MJ, Nobrega MA, and McNally EM.
813 2021. Genetic Variation in Enhancers Modifies Cardiomyopathy Gene Expression and Progression.
814 *Circulation* **143**: 1302–1316. Publisher: American Heart Association.
- 815 Galan S, Machnik N, Kruse K, Díaz N, Marti-Renom MA, and Vaquerizas JM. 2020. CHESSE enables
816 quantitative comparison of chromatin contact data and automatic feature extraction. *Nature Genetics*
817 **52**: 1247–1255.
- 818 Ghavi-Helm Y, Jankowski A, Meiers S, Viales RR, Korb J, and Furlong EEM. 2019. Highly re-
819 arranged chromosomes reveal uncoupling between genome topology and gene expression. *Nature*
820 *Genetics* **51**: 1272–1282. Number: 8 Publisher: Nature Publishing Group.
- 821 Greenwald WW, Li H, Benaglio P, Jakubosky D, Matsui H, Schmitt A, Selvaraj S, D’Antonio M,
822 D’Antonio-Chronowska A, Smith EN, et al.. 2019. Subtle changes in chromatin loop contact propen-
823 sity are associated with differential gene regulation and expression. *Nature Communications* **10**:
824 1054.

- 825 Gómez-Díaz E and Corces VG. 2014. Architectural proteins: regulators of 3D genome organization in
826 cell fate. *Trends in Cell Biology* **24**: 703–711.
- 827 Hata K, Maeno-Hikichi Y, Yumoto N, Burden SJ, and Landmesser LT. 2018. Distinct Roles of Different
828 Presynaptic and Postsynaptic NCAM Isoforms in Early Motoneuron-Myotube Interactions Required
829 for Functional Synapse Formation. *The Journal of Neuroscience: The Official Journal of the Society
830 for Neuroscience* **38**: 498–510.
- 831 Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, Rialdi A, White KM, Albrecht
832 RA, Pache L, et al.. 2018. Transcription Elongation Can Affect Genome 3D Structure. *Cell* **174**:
833 1522–1536.e22.
- 834 Hu Y, Wan S, Luo Y, Li Y, Wu T, Deng W, Jiang C, Jiang S, Zhang Y, Liu N, et al.. 2024. Benchmarking
835 algorithms for single-cell multi-omics prediction and integration. *Nature Methods* **21**: 2182–2194.
836 Publisher: Nature Publishing Group.
- 837 Ing-Simmons E, Vaid R, Bing XY, Levine M, Mannervik M, and Vaquerizas JM. 2021. Independence
838 of chromatin conformation and gene regulation during *Drosophila* dorsoventral patterning. *Nature
839 Genetics* **53**: 487–499. Number: 4 Publisher: Nature Publishing Group.
- 840 Kempfer R and Pombo A. 2019. Methods for mapping 3D chromosome architecture. *Nature Reviews
841 Genetics* .
- 842 Kempfer R and Pombo A. 2020. Methods for mapping 3D chromosome architecture. *Nature Reviews
843 Genetics* **21**: 207–226. Number: 4 Publisher: Nature Publishing Group.
- 844 Kim J, He Y, and Park H. 2014. Algorithms for nonnegative matrix and tensor factorizations: a unified
845 view based on block coordinate descent framework. *Journal of Global Optimization* **58**: 285–319.
- 846 Kobets VA, Ulianov SV, Galitsyna AA, Doronin SA, Mikhaleva EA, Gelfand MS, Shevelyov YY, Razin
847 SV, and Khrameeva EE. 2023. HiConfidence: a novel approach uncovering the biological signal in
848 Hi-C data affected by technical biases. *Briefings in Bioinformatics* **24**: bbad044.
- 849 Koitsopoulos PG and Rabkin SW. 2021. The association of polymorphism in PHACTR1 rs9349379 and
850 rs12526453 with coronary artery atherosclerosis or coronary artery calcification. A systematic review.
851 *Coronary Artery Disease* **32**: 448–458.

- 852 Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, and Sabeti PC. 2019. Identifying gene
853 expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**:
854 e43803. Publisher: eLife Sciences Publications, Ltd.
- 855 Kriebel AR and Welch JD. 2022. UINMF performs mosaic integration of single-cell multi-omic datasets
856 using nonnegative matrix factorization. *Nature Communications* **13**: 780. Publisher: Nature Publish-
857 ing Group.
- 858 Kruse K, Hug CB, and Vaquerizas JM. 2020. FAN-C: a feature-rich framework for the analysis and
859 visualisation of chromosome conformation capture data. *Genome Biology* **21**: 303.
- 860 Kuveljic J, Djuric T, Stankovic G, Dekleva M, Stankovic A, Alavantic D, and Zivkovic M. 2021. Associ-
861 ation of PHACTR1 intronic variants with the first myocardial infarction and their effect on PHACTR1
862 mRNA expression in PBMCs. *Gene* **775**: 145428.
- 863 Lawson HA, Liang Y, and Wang T. 2023. Transposable elements in mammalian chromatin organization.
864 *Nature Reviews Genetics* pp. 1–12. Publisher: Nature Publishing Group.
- 865 Lee DD and Seung HS. 2000. Algorithms for Non-negative Matrix Factorization. In *In NIPS*, volume 13,
866 pp. 556–562.
- 867 Lee DI and Roy S. 2021. GRiNCH: simultaneous smoothing and detection of topological units of
868 genome organization from sparse chromatin contact count matrices with matrix factorization. *Genome*
869 *Biology* **22**: 164.
- 870 Li B and Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without
871 a reference genome. *BMC Bioinformatics* **12**: 323.
- 872 Li X, Zeng G, Li A, and Zhang Z. 2021. DeTOKI identifies and characterizes the dynamics of chromatin
873 TAD-like domains in a single cell. *Genome Biology* **22**: 217.
- 874 Lieberman-Aiden E, van Berkum NL, Williams L, Imaikaev M, Ragooczy T, Telling A, Amit I, Lajoie BR,
875 Sabo PJ, Dorschner MO, et al.. 2009. Comprehensive Mapping of Long-Range Interactions Reveals
876 Folding Principles of the Human Genome. *Science* **326**: 289–293.
- 877 Lindström S, Wang L, Smith EN, Gordon W, van Hylckama Vlieg A, de Andrade M, Brody JA, Pattee
878 JW, Haessler J, Brumpton BM, et al.. 2019. Genomic and transcriptomic association studies identify
879 16 novel susceptibility loci for venous thromboembolism. *Blood* **134**: 1645–1657.

- 880 Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, and Welch JD. 2020. Jointly defining cell types
881 from multiple single-cell datasets using LIGER. *Nature Protocols* **15**: 3632–3662. Publisher: Nature
882 Publishing Group.
- 883 Liu J, Wang C, Gao J, and Han J. 2013. Multi-View Clustering via Joint Nonnegative Matrix Factor-
884 ization. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (eds. J Ghosh,
885 Z Obradovic, J Dy, ZH Zhou, C Kamath, and S Parthasarathy), pp. 252–260. Society for Industrial
886 and Applied Mathematics, Philadelphia, PA.
- 887 Love MI, Huber W, and Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-
888 seq data with DESeq2. *Genome Biology* **15**: 550.
- 889 Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L,
890 Dugas M, Colomé-Tatché M, et al.. 2022. Benchmarking atlas-level data integration in single-cell
891 genomics. *Nature Methods* **19**: 41–50. Publisher: Nature Publishing Group.
- 892 Lun AT and Smyth GK. 2015. diffHic: a Bioconductor package to detect differential genomic interac-
893 tions in Hi-C data. *BMC Bioinformatics* **16**: 258.
- 894 Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM,
895 Laxova R, et al.. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of
896 gene-enhancer interactions. *Cell* **161**: 1012–1025.
- 897 Lupiáñez DG, Spielmann M, and Mundlos S. 2016. Breaking TADs: How Alterations of Chromatin
898 Domains Result in Disease. *Trends in Genetics* **32**: 225–237.
- 899 McArthur E and Capra JA. 2021. Topologically associating domain boundaries that are stable across
900 diverse cell types are evolutionarily constrained and enriched for heritability. *American Journal of*
901 *Human Genetics* **108**: 269–283.
- 902 McCord R. 2017. Chromosome biology: How to build a cohesive genome in 3D. *Nature* .
- 903 Merckenschlager M and Nora EP. 2016. CTCF and Cohesin in Genome Folding and Transcriptional Gene
904 Regulation. *Annual Review of Genomics and Human Genetics* **17**: 17–43.
- 905 Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, and Chang HY. 2016. HiChIP:
906 efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* **13**: 919–922.

- 907 Norton HK and Phillips-Cremins JE. 2017. Crossed wires: 3D genome misfolding in human disease.
908 *Journal of Cell Biology* **216**: 3441–3452.
- 909 Orozco G, Schoenfelder S, Walker N, Eyre S, and Fraser P. 2022. 3D genome organization links non-
910 coding disease-associated variants to genes. *Frontiers in Cell and Developmental Biology* **10**.
- 911 Pollex T, Rabinowitz A, Gambetta MC, Marco-Ferreres R, Viales RR, Jankowski A, Schaub C, and
912 Furlong EEM. 2024. Enhancer–promoter interactions become more instructive in the transition from
913 cell-fate specification to tissue differentiation. *Nature Genetics* **56**: 686–696. Publisher: Nature
914 Publishing Group.
- 915 Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
916 *Bioinformatics* **26**: 841–842.
- 917 Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I,
918 Omer AD, Lander ES, et al.. 2014. A 3D map of the human genome at kilobase resolution reveals
919 principles of chromatin looping. *Cell* **159**: 1665–1680.
- 920 Reiff SB, Schroeder AJ, Kırılı K, Cosolo A, Bakker C, Mercado L, Lee S, Veit AD, Balashov AK,
921 Vitzthum C, et al.. 2022. The 4D Nucleome Data Portal as a resource for searching and visualizing
922 curated nucleomics data. *Nature Communications* **13**: 2365. Number: 1 Publisher: Nature Publishing
923 Group.
- 924 Rowley MJ and Corces VG. 2018. Organizational principles of 3d genome architecture. *Nature Reviews*
925 *Genetics* p. 1.
- 926 Roy AL, Conroy RS, Taylor VG, Mietz J, Fingerman IM, Pazin MJ, Smith P, Hutter CM, Singer DS, and
927 Wilder EL. 2023. Elucidating the structure and function of the nucleus—The NIH Common Fund 4D
928 Nucleome program. *Molecular Cell* **83**: 335–342. Publisher: Elsevier.
- 929 Shetty A, Sytnyk V, Leshchyns’ka I, Puchkov D, Haucke V, and Schachner M. 2013. The neural cell ad-
930 hesion molecule promotes maturation of the presynaptic endocytotic machinery by switching synaptic
931 vesicle recycling from adaptor protein 3 (AP-3)- to AP-2-dependent mechanisms. *The Journal of Neu-
932 roscience: The Official Journal of the Society for Neuroscience* **33**: 16828–16845.
- 933 Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, and Zhou XJ. 2016. TopDom: an efficient and
934 deterministic method for identifying topological domains in genomes. *Nucleic Acids Research* **44**:
935 e70.

- 936 Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, et al..
937 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids*
938 *Research* **51**: D977–D985.
- 939 Stadhouders R, Vidal E, Serra F, Di Stefano B, Le Dily F, Quilez J, Gomez A, Collombet S, Berenguer
940 C, Cuartero Y, et al.. 2018. Transcription factors orchestrate dynamic interplay between genome
941 topology and gene regulation during cell reprogramming. *Nature Genetics* **50**: 238–249.
- 942 Stansfield JC, Cresswell KG, and Dozmorov MG. 2019. multiHiCcompare: joint normalization and
943 comparative analysis of complex Hi-C experiments. *Bioinformatics* **35**: 2916–2923.
- 944 van Steensel B and Furlong EEM. 2019. The role of transcription in shaping the spatial organization
945 of the genome. *Nature Reviews Molecular Cell Biology* **20**: 327–337. Number: 6 Publisher: Nature
946 Publishing Group.
- 947 Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, Goff LA, Li Y, Ngom
948 A, Ochs MF, et al.. 2018. Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends*
949 *in Genetics* **34**: 790–805. Publisher: Elsevier.
- 950 Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, and Hadjur S. 2015. Com-
951 parative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell*
952 *reports* .
- 953 Wang G, Meng Q, Xia B, Zhang S, Lv J, Zhao D, Li Y, Wang X, Zhang L, Cooke JP, et al.. 2020.
954 TADsplimer reveals splits and mergers of topologically associating domains for epigenetic regulation
955 of transcription. *Genome Biology* **21**: 84.
- 956 Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV,
957 et al.. 2014. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells.
958 *Nature* **516**: 405–409. Number: 7531 Publisher: Nature Publishing Group.
- 959 Wang R, Lee JH, Xiong F, Kim J, Hasani LA, Yuan X, Shivshankar P, Krakowiak J, Qi C, Wang Y,
960 et al.. 2021. SARS-CoV-2 Restructures the Host Chromatin Architecture. Pages: 2021.07.20.453146
961 Section: New Results.
- 962 Wang W, Chandra A, Goldman N, Yoon S, Ferrari EK, Nguyen SC, Joyce EF, and Vahedi G. 2022. TCF-
963 1 promotes chromatin interactions across topologically associating domains in T cell progenitors.
964 *Nature Immunology* **23**: 1052–1062. Number: 7 Publisher: Nature Publishing Group.

- 965 Wanggou S, Jiang X, Li Q, Zhang L, Liu D, Li G, Feng X, Liu W, Zhu B, Huang W, et al.. 2012. HESRG:
966 a novel biomarker for intracranial germinoma and embryonal carcinoma. *Journal of Neuro-Oncology*
967 **106**: 251–259.
- 968 Warkman AS, Whitman SA, Miller MK, Garriock RJ, Schwach CM, Gregorio CC, and Krieg PA. 2012.
969 Developmental expression and cardiac transcriptional regulation of Myh7b, a third myosin heavy
970 chain in the vertebrate heart. *Cytoskeleton (Hoboken, N.J.)* **69**: 324–335.
- 971 Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, and Macosko EZ. 2019. Single-Cell Multi-
972 omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**: 1873–1887.e17.
- 973 Xiong K and Ma J. 2019. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin
974 interactions. *Nature Communications* **10**: 5069. Publisher: Nature Publishing Group.
- 975 Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, Yue F, and Li Q. 2017. HiCRep:
976 assessing the reproducibility of Hi-C data using a stratum- adjusted correlation coefficient. *Genome*
977 *Research* p. gr.220640.117.
- 978 Zerbino DR, Wilder SP, Johnson N, Juettemann T, and Flicek PR. 2015. The Ensembl Regulatory Build.
979 *Genome Biology* **16**: 56.
- 980 Zhang R, Zhou T, and Ma J. 2022. Ultrafast and interpretable single-cell 3D genome analysis with
981 Fast-Higashi. *Cell Systems* **13**: 798–807.e6.
- 982 Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, et al..
983 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating do-
984 mains in human pluripotent stem cells. *Nature Genetics* **51**: 1380–1388. Number: 9 Publisher:
985 Nature Publishing Group.
- 986 Zhang Y, Liu T, Meyer C, Eeckhoutte J, Johnson D, Bernstein B, Nussbaum C, Myers R, Brown M, Li
987 W, et al.. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.
- 988 Zheng H and Xie W. 2019. The role of 3D genome organization in development and cell differentiation.
989 *Nature Reviews Molecular Cell Biology* **20**: 535–550. Number: 9 Publisher: Nature Publishing
990 Group.
- 991 Zheng X and Zheng Y. 2018. CscoreTool: fast Hi-C compartment analysis at high resolution. *Bioinfor-*
992 *matics* **34**: 1568–1570.

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

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