Regulatory roles of three-dimensional structures of topologically associating domains

Kelly Yichen Li [1], Qin Cao [2], Huating Wang [3,4], Danny Leung [5] and Kevin Y. Yip [1,6,*]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR
[2]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR
[3]Department of Orthopaedics and Traumatology, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR
[4]Center for Neuromusculoskeletal Restorative Medicine (CNRM), CUHK InnoHK Centres, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR
[5]Division of Life Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR
[6]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA 92037, USA

1 Abstract

Transcriptional enhancers usually, but not always, regulate genes within the same topologically associating domain (TAD). We hypothesize that this incomplete insulation is due to three-dimensional (3D) structures of TADs: Whereas enhancers and genes buried inside the "core" of a TAD interact mostly with other genomic regions in the same TAD, those on the "surface" can more easily interact with regions outside the TAD. Here we show that a simple measure, the intra-TAD ratio, can quantify the "coreness" of a genomic region with respect to the TAD that it resides in. We further show that TAD surfaces are permissive for high gene expression, and accordingly cell type-specific active *cis*-regulatory elements (CREs), active histone marks, and transcription factor binding sites are enriched on TAD surfaces, most strongly in chromatin subcompartments typically considered inactive. These findings suggest a "TAD surface CRE" model of gene regulation. In addition, we also find that disease-associated non-coding variants are enriched on TAD surfaces.

2 Introduction

Mammalian genomes have a complex 3D architecture, such that meter-long DNA sequences can be compactly fitted into the micrometer-scale nucleus. The 3D genome architecture enables loci far apart on the same chromosome or even different chromosomes to come to physical proximity, which is important at different functional levels, from gene regulation, cell differentiation, to developmental processes [1–6].

Various experimental methods developed over the last decades have greatly expanded our understanding of chromatin structures. These methods mainly belong to two categories, namely imaging-based methods, such as fluorescence *in situ* hybridization (FISH), and chromosome conformation capture (3C) methods [7, 8]. Imaging methods enable the direct visualization of spatial organization of whole chromosomes or labeled DNA regions within a single cell, which can naturally capture structural variability across different cells. While

*To whom correspondence should be addressed. Email: kyip@sbpdiscovery.org

some existing imaging methods are highly advanced in visualizing large-scale chromosome organizations [9–11] and the simultaneous spatial distribution of RNA species [12–14], high-resolution imaging of whole mammalian genomes with the genomic locations of all imaged loci resolved is still not achievable. In contrast, 3C-based methods have been developed to detect conceptually all chromatin interactions in the whole genome, exemplified by Hi-C methods [15, 16], although these methods provide only information about pairs of genomic loci in close proximity but not the overall structural organization directly.

Using these experimental technologies, previous studies have discovered many types of structural components that constitute the 3D genome architecture, including compartments and subcompartments of megabases long, topologically associating domains (TADs) of tens to hundreds of kilobases long, and individual chromatin loops connecting enhancers and promoters [17].

TADs are consecutive genomic regions with a clearly higher density of chromatin interactions among loci within the TAD than interactions crossing the TAD boundaries [18, 19]. Many computational methods have been developed to identify TADs from Hi-C contact data [20]. Previous studies have proposed that TADs are formed through loop extrusion, where loop-extruding factors form progressively larger loops until stalled by boundary proteins, such as CTCF (CCCTC-binding factor) [21, 22]. Regions near TAD boundaries usually have properties different from other regions within the TAD, such as enrichment of housekeeping genes and a pair of convergent CTCF binding sites [18, 22]. Functionally, TADs provide insulation for transcriptional regulatory elements. For example, enhancers are much more likely to regulate genes within the same TAD than those outside the TAD. However, this insulation is incomplete, that some interactions between enhancers and genes outside their TADs are observed. The underlying mechanisms are not yet fully known [23].

We reason that this incomplete insulation is due to the 3D structures of TADs: whereas genomic loci hidden within the "core" of a TAD 3D structure are well-isolated from loci outside the TAD, those on the "surface" are exposed to them and thus have a higher chance of interacting with them. Indeed, from super-resolution images, TAD-like domains have been observed as globular structures [9, 24–26]. In particular, some chromatin domains (CDs) observed in these images were found to match TAD locations in mammalian cells and their surfaces and cores exhibit distinct functional properties [25]. Some other studies have attempted to analyze 3D structures of genome architecture using computational structural models inferred from Hi-C data of bulk-samples [27–29] or single-cells [30–32], but a systematic analysis of TAD 3D structures has been lacking.

For Hi-C experiments performed on a population of cells in a bulk sample, the resulting data only reflect average interaction frequencies across the cells. Single-cell Hi-C (scHi-C) methods have been developed to overcome this limitation, but as with other single-cell sequencing methods, scHi-C suffers from data sparsity, which causes dropout and a low signal-to-noise ratio, making it difficult to identify subtle cell-to-cell differences reliably [33–35].

A related recent technological development is the detection of more than two genomic loci that are interacting at the same time in the same complex. Chromatin-interaction analysis via droplet-based and barcode-linked sequencing (ChIA-Drop) [36] and split-pool recognition of interactions by tag extension (SPRITE) [37] are two representatives of such "multiplex" sequencing methods. They identify chromatin interactions between multiple loci simultaneously, by chromatin complex partitioning using gel-bead-in-emulsion droplets and by multiple rounds of tagging, respectively. In general, these multiplex methods aim at detecting com-

plexes each of which consists of multiple DNA fragments in close physical proximity within a single cell, and therefore they can be considered an intermediate between bulk-sample methods and single-cell methods. The multiplex methods can also detect genomic loci that are sufficiently close to be captured in the same complex but are not in direct contact necessary for Hi-C to detect.

Considering their semi-single-cell nature and relatively low data sparsity as compared to scHi-C data, in this study we use multiplex sequencing data to investigate 3D structures of TADs and their potential roles in transcriptional regulation. Our main methodology is the use of "intra-TAD ratio", i.e., the ratio of chromatin interactions of a genomic locus that are within the same TAD, as a proxy of the 3D location of the locus relative to the TAD. A larger intra-TAD ratio would indicate the residence of a genomic locus in the interior of the TAD. We show evidence that the intra-TAD ratio is indeed related to TAD 3D structures and it correlates with various molecular signals and activities, and locations of disease variants.

3 Results

3.1 Intra-TAD ratio as a proxy measurement for 3D location relative to TAD structures

If a TAD has a reasonably stable 3D structure, we expect regions that are buried inside the interior of the structure (referred to as the "core regions" hereafter) to mainly interact with other regions within the same TAD, while regions on the exterior of the structure (referred to as the "surface regions" hereafter) to interact more frequently with regions outside the TAD (Figure 1a). This concept can be quantified by the "intra-TAD ratio", defined as the number of sequencing reads in a chromosome conformation capture data set (such as Hi-C) that connect a region in a TAD to another region within the same TAD divided by the total number of sequencing reads connecting this region to any other regions. Core regions are expected to have higher intra-TAD ratios than surface regions. To apply this idea to multiplex data, normalized intra-TAD counts are first computed for each "cluster" (i.e., the set of DNA fragments captured in the same complex), which is expected to come from a single cell, and then the average intra-TAD ratio over all clusters is computed for each genomic locus (Methods).

To evaluate the validity of this idea, we computed intra-TAD ratios for all equal-sized genomic bins that tile the whole genome based on SPRITE data produced from the GM12878 human lymphoblastoid cells and mouse embryonic stem cells (mESCs) [37]. In general, the distribution of intra-TAD ratios remains highly stable for different sizes of the genomic bins (Figure S1). We developed and applied a data pre-processing procedure to ensure the robustness of the intra-TAD ratios and their analyses (Methods).

To verify the relationship between the intra-TAD ratio and TAD 3D structure, we first examined the intra-TAD ratios at the 1D TAD boundaries, which are supposed to be close to the TAD surface by definition. Indeed, bins overlapping 1D TAD boundaries have significantly lower intra-TAD ratios than other bins within TADs, in both GM12878 and mESC (Figures 1b,c, S2) ($p < 2.23 \times 10^{-308}$ in all cases, Mann-Whitney U test). To make sure that low intra-TAD ratios do not appear only at the 1D TAD boundaries, we examined the change of intra-TAD ratio across different bins within a TAD. We found that while the 1D

TAD boundaries consistently had the lowest intra-TAD ratios, as expected, many other bins also had low intra-TAD ratios, including some "valleys" with particularly low intra-TAD ratios (Figures 1d,e, S2), which likely correspond to genomic loci that tend to stay on TAD surfaces.

It has been shown that although chromatin interactions are dynamic and vary across different cells, 1D TAD boundaries tend to be more stable [29, 38, 39]. To see if the 3D structures of TADs are also stable, we computed intra-TAD ratios using bulk-sample Hi-C data produced from GM12878 cells [16] and mESCs [40]. As mentioned above, SPRITE can detect longer-distance chromatin interactions captured by the complexes [37], and thus conceptually the intra-TAD ratios computed from SPRITE and bulk-sample Hi-C data can be different (Figure 1f). However, we still found their values highly consistent (Pearson correlation coefficient (PCC)=0.83 and Spearman correlation coefficient (SCC)=0.79 for GM12878; PCC=0.82 and SCC=0.82 for mESC) (Figure 1g), suggesting that 3D structures of TADs are consistently captured by both multiplex SPRITE data and pairwise Hi-C data. We also computed intra-TAD ratios using scHi-C data produced from GM12878 cells [31] and mESCs [30]. Despite the sparsity of scHi-C data, intra-TAD ratios computed from SPRITE and scHi-C data still correlated positively (Figure 1g), suggesting that 3D TAD structures are fairly stable across single cells.

To gain further support that intra-TAD ratio is related to TAD 3D structures, we used structural models derived from scHi-C data of GM12878 [31] and mESC [32] to determine the fraction of proximal loci of each genomic bin that are within the same TAD in the structural models, which can be considered an independent way of computing intra-TAD ratio directly from a 3D structure. These structurally-derived intra-TAD ratios were found to be highly correlated with the intra-TAD ratios computed from SPRITE data (PCC=0.64 and SCC=0.66 for GM12878; PCC=0.63 and SCC=0.64 for mESC) (Figure 1g). When intra-TAD ratios computed from SPRITE data were used to color genomic loci in the structural models, the interior regions of TADs were indeed enriched for colors of large intra-TAD ratios (Figure 1h).

Taken together, these results provide strong supporting evidence that the intra-TAD ratio is a proxy measurement of the degree that a genomic locus resides in the core of a TAD.

3.2 TAD 3D structures differ across chromosome subcompartments

Previous studies have classified genomic regions roughly into the A (active) and B (inactive) compartments [15], and further divided them into different subcompartments [16, 41]. These subcompartments have been shown to be associated with various epigenomic signals and molecular activities, such as gene expression, histone modifications and DNA replication timing [16, 41]. To see if the 3D structures of TADs in different subcompartments have different properties, we compared the distributions of intra-TAD ratios of genomic bins in five subcompartments in GM12878 (A1, A2, B1, B2 and B3) [41]. We found that the intra-TAD ratio has different distributions in the different subcompartments, generally larger in the B2 and B3 subcompartments than in the A1, A2 and B1 subcompartments (Figure S3a). This is consistent with previous observations that TAD-like domains in A1 and A2 are more dispersed [16, 24, 42], which allow core regions in these TADs to be more accessible to regions outside the TAD and thus their lower intra-TAD ratios. We also found that TADs in B2 and B3 tend to be larger than TADs in the other subcompartments in terms of 1D genomic

span (Figure S3b), which may also contribute to their larger intra-TAD ratios since core regions in a large TAD are expected to be more insulated from outside. This is supported by the fact that intra-TAD ratio is positively correlated with TAD size in all subcompartments (Figure S4). A recent study has demonstrated that DNA in smaller nucleosome clutches is more accessible and active, which is inline with our finding here since TADs may represent clusters of clutches [42]. The generally smaller TADs in the B1 subcompartment as compared to those in B2 and B3 likely explain their different distributions of intra-TAD ratios despite all three subcompartments belong to the inactive B compartment.

3.3 Genes on TAD surfaces tend to have higher expression

Chromatin structures and gene regulation are coordinated during differentiation and other biological processes [2–4, 43], but how such an interplay is mediated by the 3D structures of TADs has not been investigated in detail. To study whether TAD 3D structures are related to gene expression, we downloaded RNA-seq data produced from GM12878 cells and mESCs by ENCODE [44]. As expected, genes in A1 and A2 subcompartments have significantly higher expression than those in B1, B2, and B3 subcompartments (Figure S5). Partitioning genes into equal-size groups based on intra-TAD ratios of their transcription start sites (TSSs), a clear trend is seen for genes with larger intra-TAD ratios to have lower expression levels (Figures 2a, S6a). The overall correlation between intra-TAD ratio and gene expression is -0.24 for GM12878 and -0.13 for mESCs. This finding suggests that genes with high expression levels tend to reside on TAD surfaces. As to be shown below, this may make these genes more exposed to interactions coming from outside the TAD such as transcription factor binding.

To further characterize this dependency, we compared the intra-TAD ratios of the most highly expressed genes (top 1,000) and the most lowly expressed genes (bottom 1,000), and found a significant difference between them (Figures 2b, S6b). Since housekeeping genes tend to have high expression, we compared the intra-TAD ratios of TSSs of housekeeping genes [45] with those of other genes, and found that housekeeping genes have significantly smaller intra-TAD ratios (Figures 2c, S6c), even when TAD 1D boundaries and their neighbouring regions were excluded (Figure S7).

To rule out the possibility that the inverse relationship between intra-TAD ratio and gene expression is simply caused by TAD size or differences between different chromosome subcompartments, we computed the correlation between intra-TAD ratio and gene expression level across different genes in the same TAD for each TAD containing at least 5 genes, such that any effects due to TAD size and chromosome subcompartment would be eliminated. The resulting distribution of correlation values collected from these TADs has a mean significantly smaller than zero (Figures 2d, S6d, $p = 4.61 \times 10^{-89}$ for GM12878 and $p = 3.59 \times 10^{-200}$ for mESC, two-sided t-test). As a further confirmation, when we plotted these correlation values against TAD sizes for each chromosome subcompartment separately, we did not observe any significant correlation between them (Figure S8). These results show that the inverse relationship between gene expression level and intra-TAD ratio is not due to TAD size or subcompartment alone.

We also used an intra-TAD ratio threshold to classify genes into "TAD core genes" and "TAD surface genes" (Methods), and found that TAD surface genes have significantly higher expression levels than TAD core genes ($p = 3.06 \times 10^{-252}$ for GM12878 and $p < 2.23 \times 10^{-308}$ for mESC, Mann-Whitney U test) (Figures 2e, S6e).

We repeated the above analyses in different subcompartments separately and reached the same conclusions highly consistently across subcompartments (Figure S9).

3.4 Chromatin signals are correlated with TAD 3D structures

Chromatin signals such as chromatin accessibility and histone modifications have been shown to characterize chromosome subcompartments [16, 41], but whether they are associated with TAD 3D structures remains unclear. We computed the correlation between intra-TAD ratio and each of 12 chromatin signals in GM12878, including chromatin accessibility signals from ATAC-seq and enrichment of 11 histone marks from ChIP-seq [44, 46] (Figure 3a). ATAC-seq signals (Figure 3a,b) and most active histone marks (Figure 3a,c) are negatively correlated with intra-TAD ratio, while some repressive histone marks are positively correlated with intra-TAD ratio (Figure 3a,d), with stronger correlations seen in the B2 and B3 subcompartments. These observations are consistent with the negative correlation between gene expression and intra-TAD ratio shown above, that active chromatin signals are enriched on TAD surfaces. Similar trends are also observed from the correlations between epigenomic signals and intra-TAD ratio in mESCs in the A and B compartments (Figure S10).

Interestingly, the correlation between the repressive mark H3K27me3 and intra-TAD ratio is slightly positive in A1 and A2 but quite negative in B2 and B3 (Figure 3a, S11). This is different from the universal positive correlation between the repressive mark H3K9me3 and intra-TAD ratio across all subcompartments. This may be related to differences between facultative and constitutive heterochromatin, that H3K27me3-marked regions can more easily be reactivated if they are on TAD surfaces. Alternatively, this may also be due to the different molecular weights (MW) of the enzyme complexes that deposit H3K27me3 and H3K9me3. The H3K27me3 modifier polycomb repressive complex 2 (PRC2) (MW: ∼300 kDa) has a larger molecular weight than H3K9me3 modifiers such as SETDB1 (MW: ∼140 kDa), SETDB2 (MW: ∼80 kDa), SUV39H1 (MW: ∼50 kDa) and SUV39H2(MW: ∼50 kDa) [25]. As a result, the large TADs in B2 and B3 subcompartments may impose stronger hindrance to the deposition of H3K27m3e to their core regions than to H3K9me3.

To further explore the effect of H3K9me3 and H3K27me3 on intra-TAD ratio, we built a linear regression model with H3K9me3 signal, H3K27me3 signal, and the interaction term between them as the explanatory variables, and intra-TAD ratio as the response variable. Consistent with the correlation analysis above, both H3K9me3 and H3K27me3 had positive model coefficients in the A1, A2, and B1 subcompartments, while the coefficients of H3K27me3 were much more negative than H3K9me3 in the B2 and B3 subcompartments (Figure S12). Surprisingly, the coefficients of the interaction term were consistently opposite in sign to the coefficients of the two independent variables, suggesting that the distribution of the two marks when they appear together with respect to TAD 3D structures is different from the distribution when they appear separately. The exact reason for this discrepancy is unclear.

3.5 Cell type-specific *cis*-regulatory elements and transcription factors are enriched on TAD surfaces

As shown above, despite the different properties of individual TADs, TAD surfaces are generally more active than TAD cores, especially for the TADs in the B2 and B3 subcompart-

ments. We hypothesized two alternative models for the TAD surface genes to be regulated by CREs in the same TAD, namely having the CREs (such as enhancers) also residing on TAD surfaces or being buried in the TAD cores (Figure 4a). The former, "TAD surface CRE" model, provides a better explanation for the CREs to be easily accessible by transcription factors, while the latter, "TAD core CRE" model, provides a better explanation for the insulation of the CREs in a TAD from regulating genes outside the TAD. To test these models, we downloaded different categories of candidate CREs (cCREs) active in GM12878 from ENCODE [47] and examined their locations with respect to TAD 3D structures based on intra-TAD ratios (Methods). Several categories of cCREs were found enriched on TAD surfaces (Figure 4b) and depleted from TAD cores (Figure S13a), most notably proximal enhancer-like signatures (pELS) and promoter-like signatures (PLS). This is not simply due to 1D TAD boundaries, as the enrichment of cCREs on TAD surfaces and depletion of them from TAD cores still hold when cCREs at and near TAD boundaries are excluded from the analysis (Figures 4c, S13b). To check whether the enrichment of cCREs on TAD surfaces is related to cell type-specific activities or it simply reflects the locations of CREs regardless of their activities, we also downloaded cell type-agnostic cCREs and cCREs active in other cell types (Methods). When cCREs active in GM12878 were excluded from these sets, the remaining cCREs were no longer enriched on the surfaces of GM12878 TADs (Figure 4d). Together, these results demonstrate that cCREs are enriched on TAD surfaces in a cell type-specific manner.

To see whether residing on TAD surfaces allows CREs to be easily bound by transcription factors, we next examined the enrichment of cell type-specific transcription factor binding sites on TAD surfaces and in TAD cores, based on ChIP-seq peaks of 153 TFs in GM12878 produced by ENCODE [44, 46]. Strikingly, the binding sites of almost all the TFs are enriched on TAD surfaces and depleted from TAD cores (Figures 4e, S14), consistent with the finding that enhancers and promoters are enriched on TAD surfaces. The two most prominent exceptions are CBX3, which is associated with heterochromatin marked with H3K9me3, and EZH2, which is the enzymatic subunit of the polycomb repressive complex 2 that catalyzes H3K27me3. Interestingly, we found that the fold of enrichment of TFs on TAD surfaces had a positive correlation with molecular weights of TF subunits [48] while the correlation was negative for TAD cores (Figure S15, S16). Again, this may indicate that core regions of TADs, especially those of the large TADs in the B2 and B3 subcompartments, are physically less accessible to the large TF proteins and thus their binding sites are enriched on TAD surfaces [49]. In turn, this may create competitions for the small TF proteins at TAD surfaces and thus their binding sites are relatively enriched in TAD cores.

To check whether the enrichment of TFs on TAD surfaces is cell type-specific, we also downloaded ChIP-seq peak locations of TFs in four other cell types. When the TF peaks in GM12878 were excluded from these sets, the remaining peaks had reduced enrichment on the GM12878 TAD surfaces; Furthermore, when the cCREs in GM12878 were also excluded from these sets, the enrichment largely disappeared (Figures S17, S18, S19, S20). These results further support the "TAD surface CRE" model, that the localization of CREs on TAD surfaces allows TFs to bind them more easily in a cell type-specific manner.

3.6 TAD 3D structures correlate with replication timing and nuclear spatial compartmentalization

Previous studies have demonstrated that DNA replication timing has strong association with spatial genome organization. In general, loci in the active A compartment tend to replicate early, whereas loci in the repressive B compartment tend to replicate late [50]. Although TAD boundaries often align with boundaries of replication domains, the relationship between TAD 3D structures and replication timing remains obscure [50].

To see whether replication timing is correlated with TAD 3D structures, we analyzed Repli-seq data produced from GM12878 cells by ENCODE [44] and found that replication timing signal (higher value corresponding to earlier replication) negatively correlated with intra-TAD ratio when considering all the TADs genome-wide, and even more so for the TADs in the B subcompartments (Figures 5a, S21). These results suggest that 3D TAD structures play a stronger role in replication timing for genomic loci in the B subcompartments and that early replication tends to occur in regions on the surfaces of these TADs.

Nuclear compartments, such as nuclear speckles and lamina, participate in chromosome packaging and potentially associate with genome function [37, 50, 51]. Given the correlation between 3D TAD structure and replication timing, we wondered whether 3D TAD structure is also related to nuclear compartments. A recent study has identified several chromatin states (called Spatial Position Inference of the Nuclear genome, "SPIN" states) that correspond to nuclear compartmentalization [52]. We compared the intra-TAD ratios of genomic loci belonging to different SPIN states in human K562 myelogenous leukemia cells (Figure 5b). In general, the inactive and heterochromatic lamina regions have the highest intra-TAD ratios while the active interior regions have lower intra-TAD ratios than the repressive interior regions. It has been shown that an active inter-chromosomal interaction hub is organized around nuclear speckles [37]. We found that genomic regions near nuclear speckles have higher intra-TAD ratios than some repressive interior regions. This may be due to spatial constraints imposed on chromatin structures by the nuclear speckles, which require further investigations.

3.7 3D TAD structure is related to disease-associated genetic variants

Beyond the correlation of 3D TAD structures with the various molecular- and cellular-level features shown above, we further explored whether they are also related to higher-level phenotypes. Specifically, we studied the locations of disease-associated genetic variants relative to 3D TAD structures. Genome-wide association studies (GWAS) have identified millions of variants associated with various diseases. A large proportion of these variants are outside coding sequences and tend to locate in the regulatory regions [53]. As such, they may tend to reside on TAD surfaces in the relevant cell types. To test this hypothesis, we downloaded lymphoma-related variants from the GWAS Catalog [54]. We first found that these variants are strongly enriched in the cCREs in GM12878 (2.84 fold of enrichment, $p = 1.60 \times 10^{-35}$, hypergeometric test), as expected. We then specifically focused on the lymphoma-associated non-coding variants, and found that they have significantly smaller intra-TAD ratios as compared to all other non-coding regions (Figure 5c), indicating that indeed these variants tend to locate on TAD surfaces in this B lymphocyte cell line.

4 Discussion

In this study, we have shown that the intra-TAD ratio serves as a proxy for 3D TAD structures. It enabled us to discover various molecular features that are correlated with 3D TAD structures, including the enrichment of highly expressed genes, CREs, active histone marks, and TF binding sites on TAD surfaces. Most of these correlations are cell type-specific and are more prominent in the B2 and B3 sub-compartments, which contain larger and more inactive TADs. We have also shown that genomic loci with different replication timing and near different nuclear compartments have different preferences of their localization relative to 3D TAD structures. Finally, we have also shown that disease-associated non-coding variants tend to reside more on TAD surfaces.

The plots of gene expression against intra-TAD ratio (Figures 2a, S6a) show that there is a clear depletion of genes with both a high intra-TAD ratio and a high gene expression level. On the other hand, genes with a low intra-TAD ratio is not guaranteed to have a high expression level. This relationship resembles findings about promoter DNA methylation in gene expression control: whereas a gene with high promoter methylation is repressed, a gene with low promoter methylation is permitted to express but its expression level depends on other factors such as transcription factor binding [55]. Analogously, a low intra-TAD ratio appears to be a condition permissive for a gene to be highly expressed, but whether it actually has a high expression level depends on additional factors.

The strong enrichment of cell type-specific CREs and TF binding sites on TAD surfaces led us to propose the "TAD surface CRE" model, in which CREs and genes are all on TAD surfaces, which make it easy for them to interact with each other and with transcription factors coming from outside the TAD. Indeed, we found that in general most TFs have their binding sites enriched on TAD surfaces, although variations were observed across chromatin subcompartments. We also found that the degree of enrichment for a TF to bind in TAD cores is inversely correlated with its molecular weight, suggesting that physical hindrance caused by TAD 3D structures may play a role in determining the binding locations of TFs. One limitation of this analysis is that we only considered single subunits of the TFs when calculating their molecular weights. A more detailed analysis that considers TF complexes will be needed to confirm the inverse correlation observed in this study.

One question associated with the "TAD surface CRE" model is why CREs tend to regulate genes in the same TAD, given that they are also exposed to genes outside (Figure 4a). One possibility is that genes in the same TAD as a CRE have more stable proximity with it, while the proximity between the CRE and genes outside the TAD could be more dynamic. As a result, a CRE can occasionally regulate genes outside its TAD when proximity requirements are satisfied, resulting in the incomplete insulation of CREs [23].

In this study, we have demonstrated the use of multiplex data from SPRITE to study 3D TAD structures, which offer both semi-single-cell information (by having each cluster expected to come from a single cell but different clusters in the same data set coming from a population of cells) and longer-range interactions that are not captured by the proximity ligation of Hi-C. The fairly strong correlation between the intra-TAD ratios computed from SPRITE and bulk Hi-C suggest that 3D TAD structures are reasonably stable across single cells and that binary interactions from Hi-C could be sufficient for computing the intra-TAD ratio. This opens the opportunity for large-scale study of intra-TAD ratios in different cell types, given the much more abundant published bulk Hi-C data than multiplex data. On

the other hand, to further investigate the stability of 3D TAD structures across single cells, single-cell SPRITE data [56] could be utilized. By comparing intra-TAD ratios computed from bulk Hi-C, scHi-C, SPRITE, and single-cell SPRITE, differences in intra-TAD ratios caused by cell-cell variation and those caused by multiplexing can be separately analyzed.

Our analysis of lymphoma-associated non-coding variants shows that there is a trend for them to reside on TAD surfaces. A difficulty of this analysis was the small number of variants available from the GWAS Catalog, which prohibited us from further focusing on only the non-coding variants within GM12878 cCREs since that would leave too few variants to remain and thus decrease statistical power. Nevertheless, our initial results here suggest that similar analyses can also be performed for other diseases and corresponding cell types. By pooling data from related diseases, such as a pan-cancer analysis, more significant enrichment of disease-associated non-coding variants on TAD surfaces may be found.

In some previous work, efforts have been devoted to deriving useful 1D features (i.e., a single numeric value for each genomic locus) from Hi-C contact maps and to linking them with chromatin organization and gene regulation [57]. As the most well-known examples, the first principal component of the normalized Hi-C contact map was introduced to partition genomic loci into the A and B compartments [15], and both the directionality index (DI) [18] and insulation score (IS) [58] were used to identify TAD boundaries. The intra-TAD ratio we introduce in this work can also be considered a type of 1D feature. The previously proposed open chromatin index (OCI) [59] and distal-to-local ratio (DLR) [60] bear some similarity to the intra-TAD ratio, in that they are both based on the comparison of Hi-C interactions that involve proximal regions versus distal regions. Yet to the best of our knowledge, the current study is the first one that uses 1D features derived from Hi-C contact maps to study 3D TAD structures.

5 Methods

5.1 Calculation of intra-TAD ratio using SPRITE data

SPRITE multi-contact read clusters from GM12878 cells and mESCs were downloaded from the 4D Nucleome data portal (`https://data.4dnucleome.org/`) [61] (data processed in GRCh38 and GRCm38) and Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo/`) [62] (GEO; accession: GSE114242) (data processed in GRCh37). All sequencing reads within the read clusters were overlapped with genomic bins of a fixed size (5kb, 10kb, 25kb, or 50kb). TAD calls were downloaded from the 3D Genome Browser (`http://3dgenome.fsm.northwestern.edu/`) [63]. For a genomic bin i involved in a read cluster j, its normalized intra-TAD ratio in the read cluster, r_{ij}, was defined as $N_{ij}/(C_j - 1)$, where N_{ij} is the number of intra-TAD interactions between bin i and other bins involved in this cluster and C_j is the total number of bins in cluster j. The overall intra-TAD ratio of bin i, r_i, is then the average of all its intra-TAD ratios in the read clusters that it is involved in. For calculations that considered only intra-chromosomal interactions, we first partitioned the reads in each cluster into sub-clusters according to their chromosomes, and then computed intra-TAD ratios of each sub-cluster separately.

In order to compare with intra-TAD ratios calculated from bulk Hi-C data, scHi-C data, and structural models, which were processed using different reference genome builds in the

case of GM12878, we computed intra-TAD ratios using SPRITE data based on both GRCh38 and GRCh37. In the case of mESC, all calculations were based on GRCm38.

In the analyses that involved classifying bins into "TAD surface bins" and "TAD core bins", TAD surface bins were defined in each TAD separately as bins with an intra-TAD ratio smaller than or equal to the q_k-th quantile of intra-TAD ratio within TAD k, where $q_k = n_k^{-\frac{1}{3}}$ and n_k is the total number of bins in TAD k. This definition of q_k is based on the ratio between surface area (S) and volume (V) of a sphere, which is proportional to $V^{-\frac{1}{3}}$. The remaining TAD regions were defined as TAD core bins.

5.2 Pre-processing of SPRITE data when computing intra-TAD ratios

Since SPRITE clusters involving too few or too many reads are potentially problematic, we tested different ranges of cluster size (i.e., number of reads per cluster) following the original SPRITE paper [37], and found 2-1,000 to be a reasonable range that provides a balance between filtering out problematic clusters and maintaining the diversity of intra-TAD ratios (Figure S1). Previous studies have shown that inter-chromosomal interactions provide additional information about chromatin organizations [37, 41] and considering them led to more consistent intra-TAD ratios across different ranges of cluster sizes compared with using intra-chromosomal interactions alone to compute intra-TAD ratios (Figure S1). Considering these observations and the fact that intra-TAD ratios remain fairly stable across bin sizes (Figure S1), in our default setting for computing the intra-TAD ratio, we used a small bin size of 10kb to obtain fine-grained information and considered both intra-chromosomal and inter-chromosomal interactions based on SPRITE clusters with 2-1,000 reads. We also filtered out genomic bins with less than 100 SPRITE clusters involving them.

5.3 Calculation of intra-TAD ratio using bulk Hi-C data

Processed in situ Hi-C data (in .hic format) produced from GM12878 cells [16], K562 cells [16], and mESCs [40] were downloaded from the 4D Nucleome data portal (`https://data.4dnucleome.org/`). To extract whole-genome intra- and inter-chromosomal contact matrices, we first added SCALE normalization vectors to the original .hic data using the addNorm function of Juicer Tools (version 1.22.01). The contact matrices were generated by Juicer Tools (version 1.22.01) dump at 10kb, 25kb, and 50kb resolution. The intra-TAD ratio of each genomic bin was then calculated as the intra-TAD contacts divided by the total contacts involving this bin in a Hi-C contact matrix.

5.4 Calculation of intra-TAD ratio using single-cell Hi-C data

Single-cell Hi-C contacts from GM12878 cells and mESCs were downloaded from GEO (accessions: GSE117876 and GSE80280, respectively). The Hi-C contacts were converted to .pairs format (`https://github.com/4dn-dcic/pairix/blob/master/pairs_format_specification.md#standard-format`) if they were originally not. The intra-TAD contacts and total contacts of a given genomic bin were extracted using the Pyhton package pypairix (v0.3.7), and its intra-TAD ratio was then calculated as their ratio. We used 10kb bins for GM12878

scHi-C data and 50kb bins for mESC scHi-C data due to the sparsity of the latter data set. Finally, for each bin, the average intra-TAD ratio across all single cells was compared with its intra-TAD ratio computed using SPRITE data.

5.5 Calculation of intra-TAD ratio using structural models

Dip-C [31] structures at 20kb resolution for GM12878 cells were downloaded from GEO (accession: GSE117876) and Si-C [32] structures at 10kb resolution for mESCs were downloaded from `https://github.com/TheMengLab/Si-C_3D_structure`. To calculate the intra-TAD ratio of a genomic bin according to a 3D structure, we first found the $k = 10$ nearest neighbors of it in the structure based on their Euclidean distances using the Python package sklearn (v0.23.2). Among these k nearest neighbors, we counted the number of them belonging to the same TAD as the genomic bin of interest and divided it by k to become the intra-TAD ratio. For Dip-C structures, which also contained the diploid chromosome information, genomic bins were regarded as intra-TAD neighbors only if they were from the same TAD and the same chromosome haplotype. To obtain the intra-TAD ratio at 10kb resolution for the 20kb bins in Dip-C data, we separated each 20kb bin into two virtual 10kb bins and assigned the same intra-TAD ratio to them. Finally, for each genomic bin, its intra-TAD ratios over all the available structural models (including cells and configurations) were averaged and compared with its intra-TAD ratio computed using SPRITE data.

5.6 RNA-seq data analysis

Processed RNA-seq data from GM12878 cells were downloaded from GEO (accessions: GSE88583 and GSE88627). GENCODE [64] (v24) was used to annotate the genes and extract TSSs. Replicates were averaged and gene expression levels were quantified by fragments per kilobase of exons per million fragments (FPKM). Processed RNA-seq data from mESCs were downloaded from GEO (accession: GSE90277) and processed in a similar way with GENCODE (vM4) annotations.

5.7 ATAC-seq and histone ChIP-seq data analysis

Processed ATAC-seq data and ChIP-seq data for H2AFZ, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2, and H4K20me1 (in .bigwig format) from GM12878 cells were downloaded from the ENCODE data matrix (`https://www.encodeproject.org/`). For each type of data, the averaged signal (fold change over control) in each genomic bin was computed using bigWigAverageOverBed.

5.8 Enrichment of *cis*-regulatory elements

Candidate *cis*-regulatory elements (cCREs) in GM12878, HeLaS3, HepG2, IMR90, and K562 cells as well as cell type-agnostic cCREs were downloaded from ENCODE (`https://www.encodeproject.org/`). All cCREs were overlapped with genomic bins at 5kb, 10kb, 25kb, and 50kb resolution. Regions either with no cCRE labels or labeled with "Low-DNase" were filtered out since they are unlikely CREs, and the remaining cCREs were intersected with all TAD regions, TAD surface regions, and TAD core regions. Suppose the number of all

TAD bins is a, the number of all TAD surface bins is b, the number of all filtered cCRE bins within TAD regions is c, and the number of cCRE bins on TAD surfaces is d, the TAD surface enrichment fold was defined as $\left(\frac{d}{c}\right) / \left(\frac{b}{a}\right)$. Enrichment p-values of cCREs on TAD surfaces were calculated by fitting a hypergeometric distribution using the hypergeom.sf() function inside the scipy.stats module of Python [65]. TAD core enrichment folds and p-values were calculated similarly.

For each of the four categories of cCREs (dELS, pELS, PLS and CTCF-bound), a genomic bin was labeled with a certain category if it overlapped with a region of this subtype, regardless of its overlap with regions of other categories. To test the effect of TAD boundaries on surface cCRE enrichment, we excluded TAD boundary bins and the neighboring 10% bins of the whole TAD from the TAD surface bins and repeated the analysis. To test cell-type specificity of surface cCRE enrichment, we excluded all cCREs defined in GM12878 from cell type-agnostic cCREs or cCREs in HeLaS3, HepG2, IMR90, or K562, and repeated the analysis.

5.9 Enrichment of transcription factors

ChIP-seq peaks of 153 transcription factors (TFs) obtained from GM12878 cells were downloaded from ENCODE (`https://www.encodeproject.org/`). All TF peaks were overlapped with genomic bins at 5kb, 10kb, 25kb, and 50kb resolution and intersected with TAD surfaces and TAD cores. Enrichment of TFs on TAD surfaces and in TAD cores was calculated in the same way as described above for cCREs.

The molecular weights of TFs were downloaded from UniProt (`https://www.uniprot.org/` [48]). Only the "reviewed" entry was kept if a TF had multiple entries. For TFs with multiple "reviewed" entries, we manually selected one of them to define the molecular weight. We only considered each TF protein as a single sub-unit without considering other proteins in the possible TF complexes.

5.10 Repli-seq data analysis

Wavelet-smoothed signals of Repli-seq data (in .bigWig format) from GM12878 cells were downloaded from ENCODE (`https://www.encodeproject.org/`). Averaged signals in each genomic bin at 5kb, 10kb, 25kb, and 50kb resolution were obtained using bigWigAverageOverBed.

5.11 SPIN state analaysis

SPIN annotations of genomic regions in K562 cells were downloaded from `https://github.com/ma-compbio/SPIN` [52]. Since SPRITE data were not available for K562 cells, intra-TAD ratios were calculated using bulk Hi-C data downloaded from the 4D Nucleome data portal (`https://data.4dnucleome.org/` and processed in the same way for other Hi-C data, as described above.

5.12 Disease variants analysis

Lymphoma-related variants were downloaded from GWAS Catalog [54] with trait ID of "EFO_0000574". The variants overlapped with any human transcript defined by GENCODE (v29) were excluded. The remaining variants were considered non-coding variants and were overlapped with 10kb bins. Intra-TAD ratios (calculated in GM12878) of these lymphoma-related non-coding variants were compared with those of all the non-coding 10kb bins.

References

[1] Clemens B. Hug and Juan M. Vaquerizas. The birth of the 3D genome during early embryonic development. *Trends in Genetics*, 34(12):903–914, 2018.

[2] Seungsoo Kim and Jay Shendure. Mechanisms of interplay between transcription factors and the 3D genome. *Molecular Cell*, 76(2):306–319, 2019.

[3] Ralph Stadhouders, Guillaume J. Filion, and Thomas Graf. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature*, 569(7756):345–354, 2019.

[4] Hui Zheng and Wei Xie. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology*, 20(9):535–550, 2019.

[5] A. Marieke Oudelaar and Douglas R. Higgs. The relationship between genome structure and function. *Nature Reviews Genetics*, 22(3):154–168, 2020.

[6] Erin Aboelnour and Boyan Bonev. Decoding the organization, dynamics, and function of the 4D genome. *Developmental Cell*, 56(11):1562–1573, 2021.

[7] Jinlei Han, Zhiliang Zhang, and Kai Wang. 3C and 3C-based techniques: The powerful tools for spatial genome organization deciphering. *Molecular Cytogenetics*, 11:21, 2018.

[8] Rieke Kempfer and Ana Pombo. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics*, 21(4):207–226, 2019.

[9] Bogdan Bintu, Leslie J. Mateo, Jun Han Su, Nicholas A. Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N. Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413):eaau1783, 2018.

[10] Andrés M. Cardozo Gizzi, Diego I. Cattoni, Jean Bernard Fiche, Sergio M. Espinola, Julian Gurgo, Olivier Messina, Christophe Houbron, Yuki Ogiyama, Giorgio L. Papadopoulos, Giacomo Cavalli, Mounia Lagha, and Marcelo Nollmann. Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. *Molecular Cell*, 74(1):212–222, 2019.

[11] Liangqi Xie and Zhe Liu. Single-cell imaging of genome organization and dynamics. *Molecular Systems Biology*, 17(7):e9653, 2021.

[12] Miao Liu, Yanfang Lu, Bing Yang, Yanbo Chen, Jonathan S.D. Radda, Mengwei Hu, Samuel G. Katz, and Siyuan Wang. Multiplexed imaging of nucleome architectures in single cells of mammalian tissue. *Nature Communications*, 11(1):1–14, 2020.

[13] Jun Han Su, Pu Zheng, Seon S. Kinrot, Bogdan Bintu, and Xiaowei Zhuang. Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*, 182(6):1641–1659, 2020.

[14] Yodai Takei, Jina Yun, Shiwei Zheng, Noah Ollikainen, Nico Pierson, Jonathan White, Sheel Shah, Julian Thomassie, Shengbao Suo, Chee Huat Linus Eng, Mitchell Guttman, Guo Cheng Yuan, and Long Cai. Integrated spatial genomics reveals global architecture of single nuclei. *Nature*, 590(7845):344–350, 2021.

[15] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, 2009.

[16] Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–80, 2014.

[17] M. Jordan Rowley and Victor G. Corces. Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, 19(12):789–800, 2018.

[18] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[19] Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, 2012.

[20] Marie Zufferey, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*, 19(1):1–18, 2018.

[21] Adrian L. Sanborn, Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian Li, Kristopher P. Geeting, Andreas Gnirke, Alexandre Melnikov, Doug McKenna, Elena K. Stamenova, Eric S. Lander, and Erez Lieberman Aiden. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47):E6456–E6465, 2015.

[22] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A Mirny. Formation of chromosomal domains by loop extrusion. *Cell reports*, 15(9):2038–49, 2016.

[23] Stefan Schoenfelder and Peter Fraser. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, 20(8):437–455, 2019.

[24] Melike Lakadamyali and Maria Pia Cosma. Visualizing the genome in high resolution challenges our textbook understanding. *Nature Methods*, 17(4):371–379, 2020.

[25] Ezequiel Miron, Roel Oldenkamp, Jill M. Brown, David M.S. Pinto, C. Shan Xu, Ana R. Faria, Haitham A. Shaban, James D.P. Rhodes, Cassandravictoria Innocent, Sara De Ornellas, Harald F. Hess, Veronica Buckle, and Lothar Schermelleh. Chromatin arranges in chains of mesoscale domains with nanoscale functional topography independent of cohesin. *Science Advances*, 6(39):eaba8811, 2020.

[26] Yubao Cheng, Miao Liu, Mengwei Hu, and Siyuan Wang. TAD-like single-cell domain structures exist on both active and inactive X chromosomes and persist under epigenetic perturbations. *Genome Biology*, 22(1):1–26, 2021.

[27] Harianto Tjong, Wenyuan Li, Reza Kalhor, Chao Dai, Shengli Hao, Ke Gong, Yonggang Zhou, Haochen Li, Xianghong Jasmine Zhou, Mark A. Le Gros, Carolyn A. Larabell, Lin Chen, and Frank Alber. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences of the United States of America*, 113(12):E1663–E1672, 2016.

[28] Simeon Carstens, Michael Nilges, and Michael Habeck. Bayesian inference of chromatin structure ensembles from population-averaged contact data. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14):7824–7830, 2020.

[29] Marco Di Stefano, Ralph Stadhouders, Irene Farabella, David Castillo, François Serra, Thomas Graf, and Marc A. Marti-Renom. Transcriptional activation during cell reprogramming correlates with the formation of 3D open chromatin hubs. *Nature Communications*, 11(1):1–12, 2020.

[30] Tim J. Stevens, David Lando, Srinjan Basu, Liam P. Atkinson, Yang Cao, Steven F. Lee, Martin Leeb, Kai J. Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, Julie Cramard, Andre J. Faure, Meryem Ralser, Enrique Blanco, Lluis Morey, Miriam Sansó, Matthieu G.S. Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, Brian Hendrich, Dave Klenerman, and Ernest D. Laue. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):59–64, 2017.

[31] Longzhi Tan, Dong Xing, Chi Han Chang, Heng Li, and X. Sunney Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, 2018.

[32] Luming Meng, Chenxi Wang, Yi Shi, and Qiong Luo. Si-C is a method for inferring super-resolution intact genome structure from single-cell Hi-C data. *Nature Communications*, 12(1):1–11, 2021.

[33] Takashi Nagano, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[34] Ilya M. Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B. Brandão, Sergey V. Ulianov, Nezar Abdennur, Sergey V. Razin, Leonid A. Mirny, and Kikuë Tachibana-Konwalski. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110–114, 2017.

[35] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L. Gunderson, Frank J. Steemers, Christine M. Disteche, William S. Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3):263–266, 2017.

[36] Meizhen Zheng, Simon Zhongyuan Tian, Daniel Capurso, Minji Kim, Rahul Maurya, Byoungkoo Lee, Emaly Piecuch, Liang Gong, Jacqueline Jufen Zhu, Zhihui Li, Chee Hong Wong, Chew Yee Ngan, Ping Wang, Xiaoan Ruan, Chia Lin Wei, and Yijun Ruan. Multiplex chromatin interactions with single-molecule precision. *Nature*, 566(7745):558–562, 2019.

[37] Sofia A. Quinodoz, Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M. Lai, Alexander A. Shishkin, Prashant Bhat, Yodai Takei, Vickie Trinh, Erik Aznauryan, Pamela Russell, Christine Cheng, Marko Jovanovic, Amy Chow, Long Cai, Patrick McDonel, Manuel Garber, and Mitchell Guttman. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, 174(3):744–757, 2018.

[38] Jingtian Zhou, Jianzhu Ma, Yusi Chen, Chuankai Cheng, Bokan Bao, Jian Peng, Terrence J. Sejnowski, Jesse R. Dixon, and Joseph R. Ecker. Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation. *Proceedings of the National Academy of Sciences of the United States of America*, 116(28):14011–14018, 2019.

[39] Ruochi Zhang, Tianming Zhou, and Jian Ma. Multiscale and integrative single-cell Hi-C analysis with higashi. *Nature Biotechnology*, 40(2):254–261, 2022.

[40] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L. Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean Philippe Hugnot, Amos Tanay, and Giacomo Cavalli. Multiscale 3D genome rewiring during mouse neural development. *Cell*, 171(3):557–572, 2017.

[41] Kyle Xiong and Jian Ma. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nature Communications*, 10(1):5069, 2019.

[42] Jason Otterstrom, Alvaro Castells-Garcia, Chiara Vicario, Pablo A. Gomez-Garcia, Maria Pia Cosma, and Melike Lakadamyali. Super-resolution microscopy reveals how histone tail acetylation affects DNA compaction within nucleosomes in vivo. *Nucleic Acids Research*, 47(16):8470–8484, 2019.

[43] Jesse R. Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E. Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V. Lobanenkov, Joseph R. Ecker, James A. Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, 2015.

[44] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[45] Bidossessi Wilfried Hounkpe, Francine Chenou, Franciele de Lima, and Erich Vinicius de Paula. HRT Atlas v1.0 database: Redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Research*, 49(D1):D947–D955, 2021.

[46] Carrie A. Davis, Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, Kriti Jain, Ulugbek K. Baymuradov, Aditi K. Narayanan, Kathrina C. Onate, Keenan Graham, Stuart R. Miyasato, Timothy R. Dreszer, J. Seth Strattan, Otto Jolanki, Forrest Y. Tanaka, and J. Michael Cherry. The encyclopedia of dna elements (ENCODE): Data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, 2018.

[47] Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J. Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, Zhiping Weng, and The ENCODE Project Consortium. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

[48] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2020.

[49] Kazuhiro Maeshima, Kazunari Kaizu, Sachiko Tamura, Tadasu Nozaki, Tetsuro Kokubo, and Koichi Takahashi. The physical size of transcription factors is key to transcriptional regulation in chromatin domains. *Journal of Physics: Condensed Matter*, 27(6):064116, 2015.

[50] Claire Marchal, Jiao Sima, and David M. Gilbert. Control of DNA replication timing in the 3d genome. *Nature Reviews Molecular Cell Biology*, 20(12):721–737, 2019.

[51] Bas van Steensel and Andrew S. Belmont. Lamina-associated domains: Links with chromosome architecture, heterochromatin, and gene repression. *Cell*, 169(5):780–791, 2017.

[52] Yuchuan Wang, Yang Zhang, Ruochi Zhang, Tom van Schaik, Liguo Zhang, Takayo Sasaki, Daniel Peric-Hupkes, Yu Chen, David M. Gilbert, Bas van Steensel, Andrew S.

Belmont, and Jian Ma. SPIN reveals genome-wide landscape of nuclear compartmentalization. *Genome Biology*, 22(1):1–23, 2021.

[53] Ekta Khurana, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A. Rubin, and Mark Gerstein. Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*, 17(2):93–108, 2016.

[54] Annalisa Buniello, Jacqueline A.L. Macarthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L. Whetzel, Ridwan Amode, Jose A. Guillen, Harpreet S. Riat, Stephen J. Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A. Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 2019.

[55] Shaoke Lou, Heung Man Lee, Hao Qin, Jing Woei Li, Zhibo Gao, Xin Liu, Landon L. Chan, Vincent K.L. Lam, Wing Yee So, Ying Wang, Si Lok, Jun Wang, Ronald C.W. Ma, Stephen Kwok Wing Tsui, Juliana C.N. Chan, Ting Fung Chan, and Kevin Y. Yip. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biology*, 15(7):1–21, 2014.

[56] Mary V. Arrastia, Joanna W. Jachowicz, Noah Ollikainen, Matthew S. Curtis, Charlotte Lai, Sofia A. Quinodoz, David A. Selck, Rustem F. Ismagilov, and Mitchell Guttman. Single-cell measurement of higher-order 3d genome organization with scSPRITE. *Nature Biotechnology*, 40(1):64–73, 2022.

[57] Jiankang Wang and Ryuichiro Nakato. HiC1Dmetrics: Framework to extract various one-dimensional features from chromosome structure data. *Briefings in Bioinformatics*, 23(1):bbab509, 2022.

[58] Emily Crane, Qian Bian, Rachel Patton McCord, Bryan R. Lajoie, Bayly S. Wheeler, Edward J. Ralston, Satoru Uzawa, Job Dekker, and Barbara J. Meyer. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559):240–244, 2015.

[59] Tamir Chandra, Philip Andrew Ewels, Stefan Schoenfelder, Mayra Furlan-Magaril, Steven William Wingett, Kristina Kirschner, Jean Yves Thuret, Simon Andrews, Peter Fraser, and Wolf Reik. Global reorganization of the nuclear landscape in senescent cells. *Cell Reports*, 10(4):471–483, 2015.

[60] Sven Heinz, Lorane Texari, Michael G.B. Hayes, Matthew Urbanowski, Max W. Chang, Ninvita Givarkes, Alexander Rialdi, Kris M. White, Randy A. Albrecht, Lars Pache, Ivan Marazzi, Adolfo García-Sastre, Megan L. Shaw, and Christopher Benner. Transcription elongation can affect genome 3D structure. *Cell*, 174(6), 2018.

[61] Job Dekker, Andrew S. Belmont, Mitchell Guttman, Victor O. Leshyk, John T. Lis, Stavros Lomvardas, Leonid A. Mirny, Clodagh C. O'Shea, Peter J. Park, Bing Ren,

Joan C. Ritland Politz, Jay Shendure, and Sheng Zhong. The 4D nucleome project. *Nature*, 549(7671):219–226, 2017.

[62] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: Archive for functional genomics data - update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

[63] Yanli Wang, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, Mayank N.K. Choudhary, Yun Li, Ming Hu, Ross Hardison, Ting Wang, and Feng Yue. The 3D genome browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biology*, 19(1):1–12, 2018.

[64] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773, 2018.

[65] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, 2020.

6 Acknowledgment

We thank Aniruddha Deshpande, Geneviève Fourel, Chiara Nicoletti, and Pier Lorenzo Puri for their comments on this work. This project was supported by Hong Kong Research Grants Council General Research Fund 14107420. QC is supported by National Natural Science Foundation of China Youth Program 32100515. KYY was additionally supported by Hong Kong Research Grants Council Collaborative Research Funds C4015-20E, C4045-18W, C4057-18E and C7044-19G, General Research Fund 14203119, and Theme-based Research

Scheme T12-712/21-R, the Hong Kong Epigenomics Project (EpiHK), and the Chinese University of Hong Kong Young Researcher Award, Outstanding Fellowship and Project Impact Enhancement Fund.

7 Contributions

KYY conceived the study. KYL, QC and KYY developed the methodology and designed the analyses. KYL performed data processing and analyses. KYL, QC, HW, DL and KYY interpreted the results. KYL and KYY wrote the manuscript. All authors read and approved the final version of the manuscript.

8 Figures

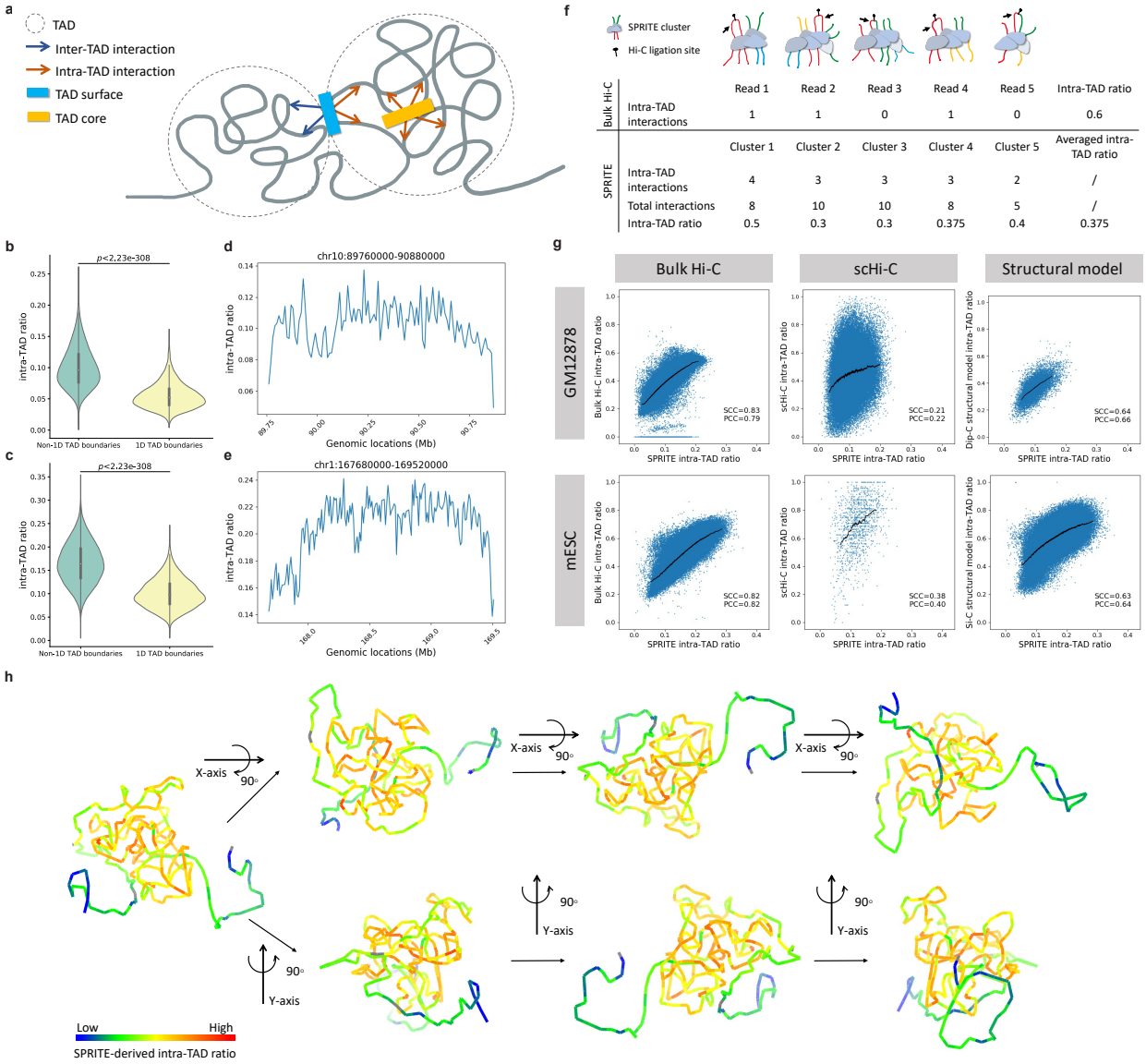

Figure 1: (Figure caption on next page.)

Figure 1: Calculation and validation of the intra-TAD ratio. (a) Schematic figure showing how the proportions of intra- and inter-TAD chromatin interactions are related to the location of a genomic locus relative to the 3D TAD structure. (b,c) Violin plots comparing the intra-TAD ratios of genomic bins that overlap 1D TAD boundaries and those that do not, in GM12878 (b) and mESC (c). (d,e) Change of intra-TAD ratio along the genomic span of example TADs, in GM12878 (d) and mESC (e). (f) Calculations of intra-TAD ratio from SPRITE and bulk Hi-C data. Black arrows indicate the target DNA fragment of which the intra-TAD ratio is computed. DNA fragments from the same TAD are shown in the same color. Hi-C captures only one DNA fragment ligated to the target fragment at a time, while SPRITE captures the whole complex of multiple DNA fragments at the same time. (g) Scatterplots comparing intra-TAD ratios calculated from SPRITE data with those calculated from bulk Hi-C, scHi-C, and structural models. Black curves represent moving averages. (h) Visualization of intra-TAD ratios in a structural model. The structural model of a TAD (chr18:77920000-80200000) was taken from Meng et al. [32]. Each genomic bin is colored by its SPRITE-derived intra-TAD ratio. Gray indicates missing data.

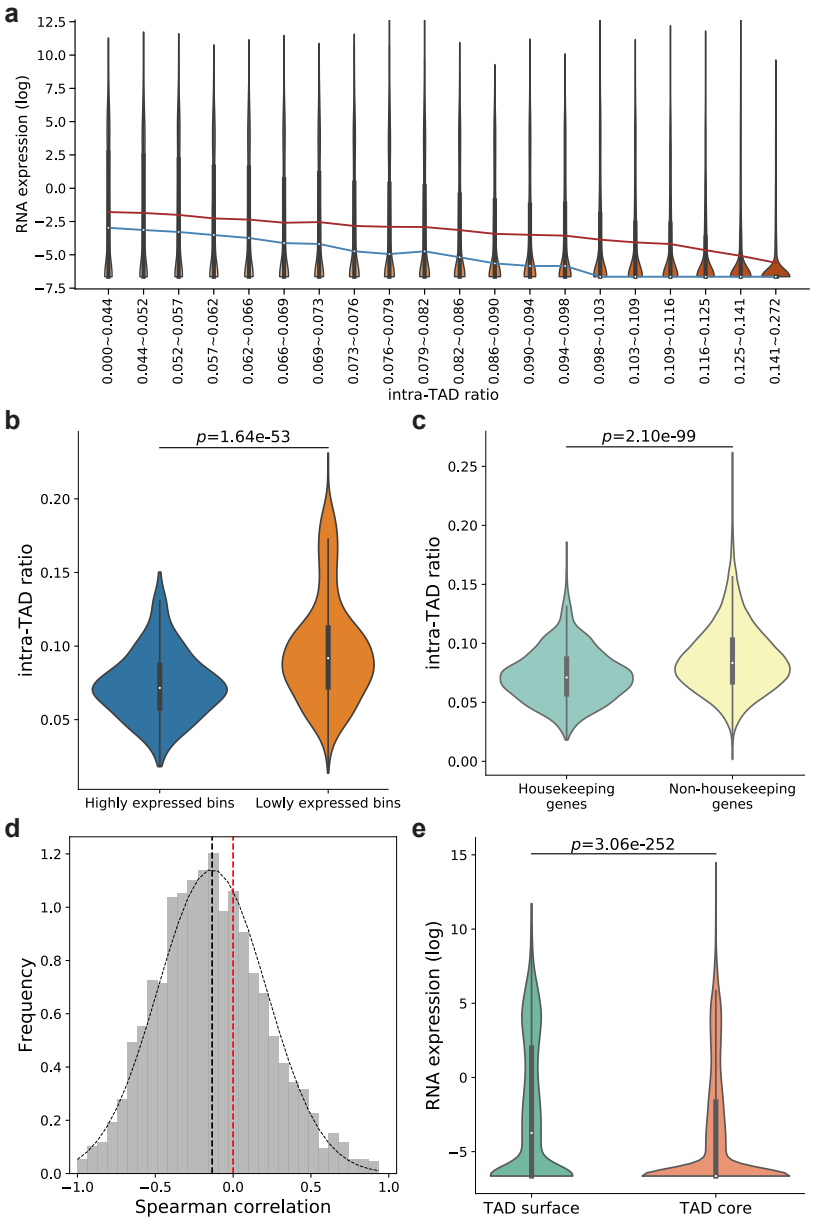

Figure 2: Inverse relationship between intra-TAD ratio and gene expression in GM12878 cells. (a) Violin plots of gene expression ($\log_2$(FPKM+0.01)) of TSS bins in groups with increasing intra-TAD ratio and similar number of bins. The blue and red lines connect median and mean values of the different groups, respectively. (b) Violin plots of intra-TAD ratios of the highly expressed bins and lowly expressed bins, defined as the genomic bins containing TSSs of genes whose FPKM values rank within top and bottom 1,000, respectively. (c) Violin plots of intra-TAD ratio of housekeeping genes and non-housekeeping genes. (d) Distribution of Spearman correlations between gene expression and intra-TAD ratio of genomic bins in individual TADs. Black dotted curve shows the fitted normal distribution. The vertical black dotted line shows the mean of Spearman correlations. The vertical red dotted line shows zero Spearman correlation. (e) Violin plots of gene expression levels ($\log_2$(FPKM+0.01)) in TAD surface TSS bins and TAD core TSS bins. p-values in all panels are calculated using the Mann-Whitney U test.

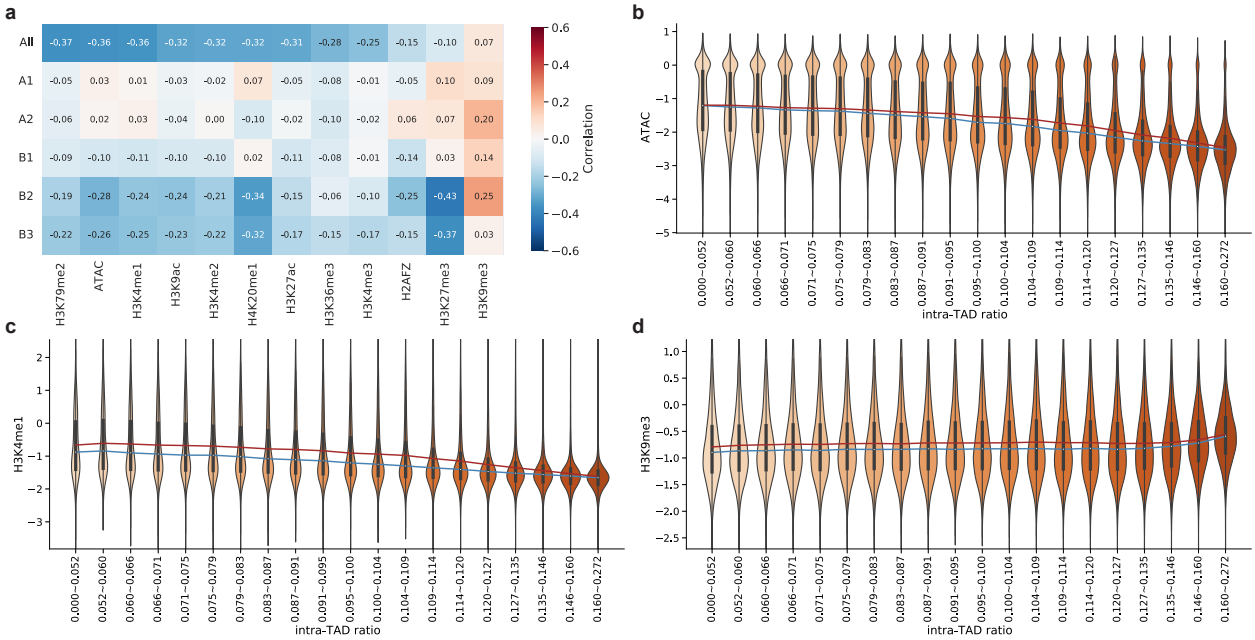

Figure 3: Correlation between intra-TAD ratio and chromatin signals. (a) Spearman correlation between intra-TAD ratio and ATAC-seq signals and 11 histone modifications across either all genomic bins or only bins in a genomic subcompartment. The columns are sorted by the correlation values when all TADs are considered (in the "All" row). (b)-(d) Violin plots of ATAC-seq signals (b), H3K4me1 (c), and H3K9me3 (d) in groups with increasing intra-TAD ratio and similar number of bins. y-axis shows $\log_2$(signal+0.01). The blue and red lines connect median and mean values of the different groups, respectively.

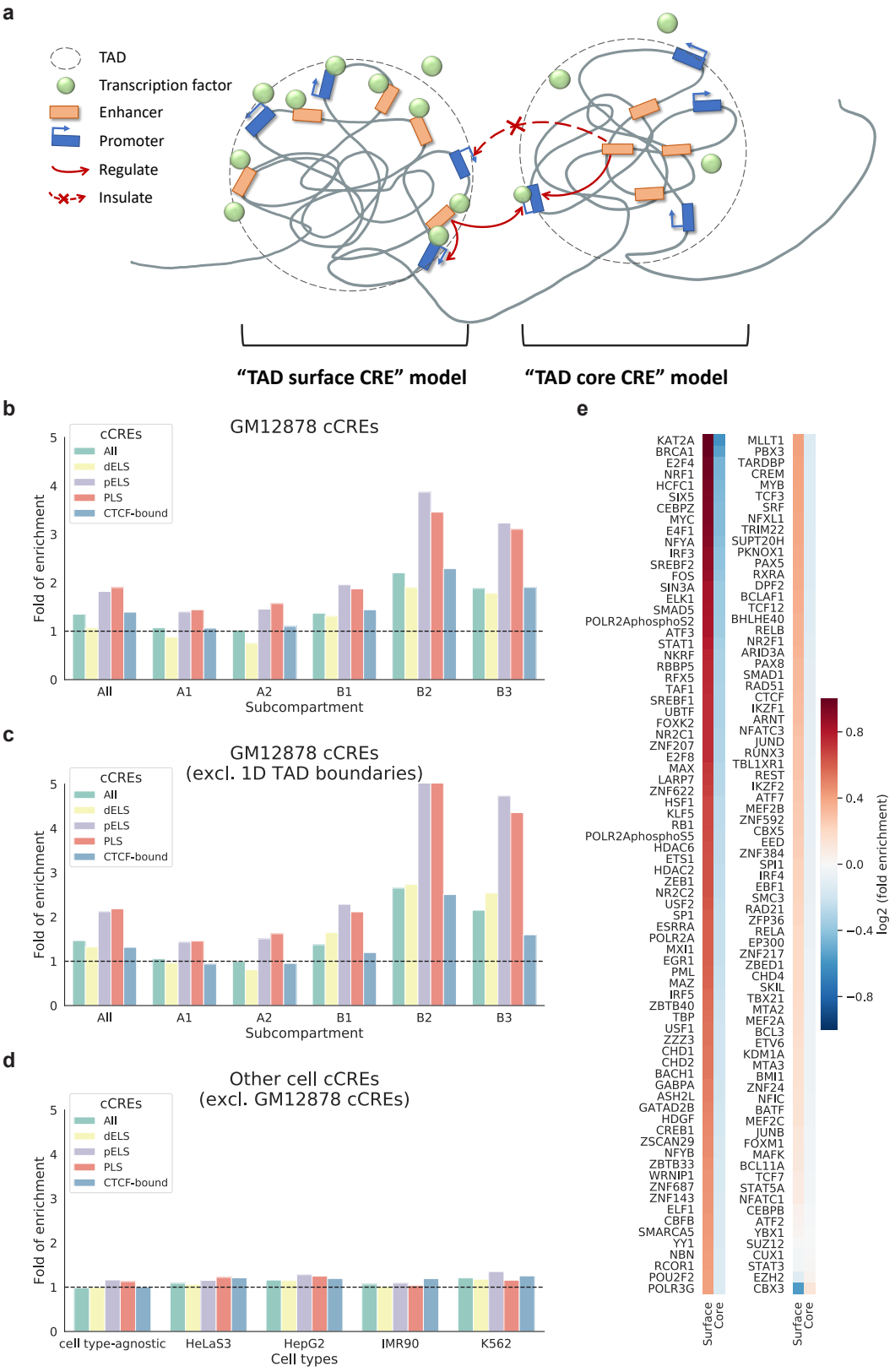

Figure 4: (Figure caption on next page.)

Figure 4: Enrichment of CREs and transcription factors on TAD surfaces. (a) A schematic diagram showing the "TAD surface CRE" model and the "TAD core CRE" model. (b) Fold of enrichment of cCREs on TAD surfaces in GM12878. (c) Fold of enrichment of cCREs on TAD surfaces in GM12878 when regions around TAD boundaries are excluded. (d) Fold of enrichment of cCREs on TAD surfaces when the cCREs are cell type-agnostic or defined in other cell types and the GM12878-specific cCREs are excluded. (e) Heatmap showing log2(enrichment fold) of transcription factor binding sites on TAD surfaces and in TAD cores in GM12878. cCREs All: all categories of cCREs; dELS: distal enhancer-like signatures; pELS: proximal enhancer-like signatures; PLS: promoter-like signatures; CTCF-bound: bound by CTCF.

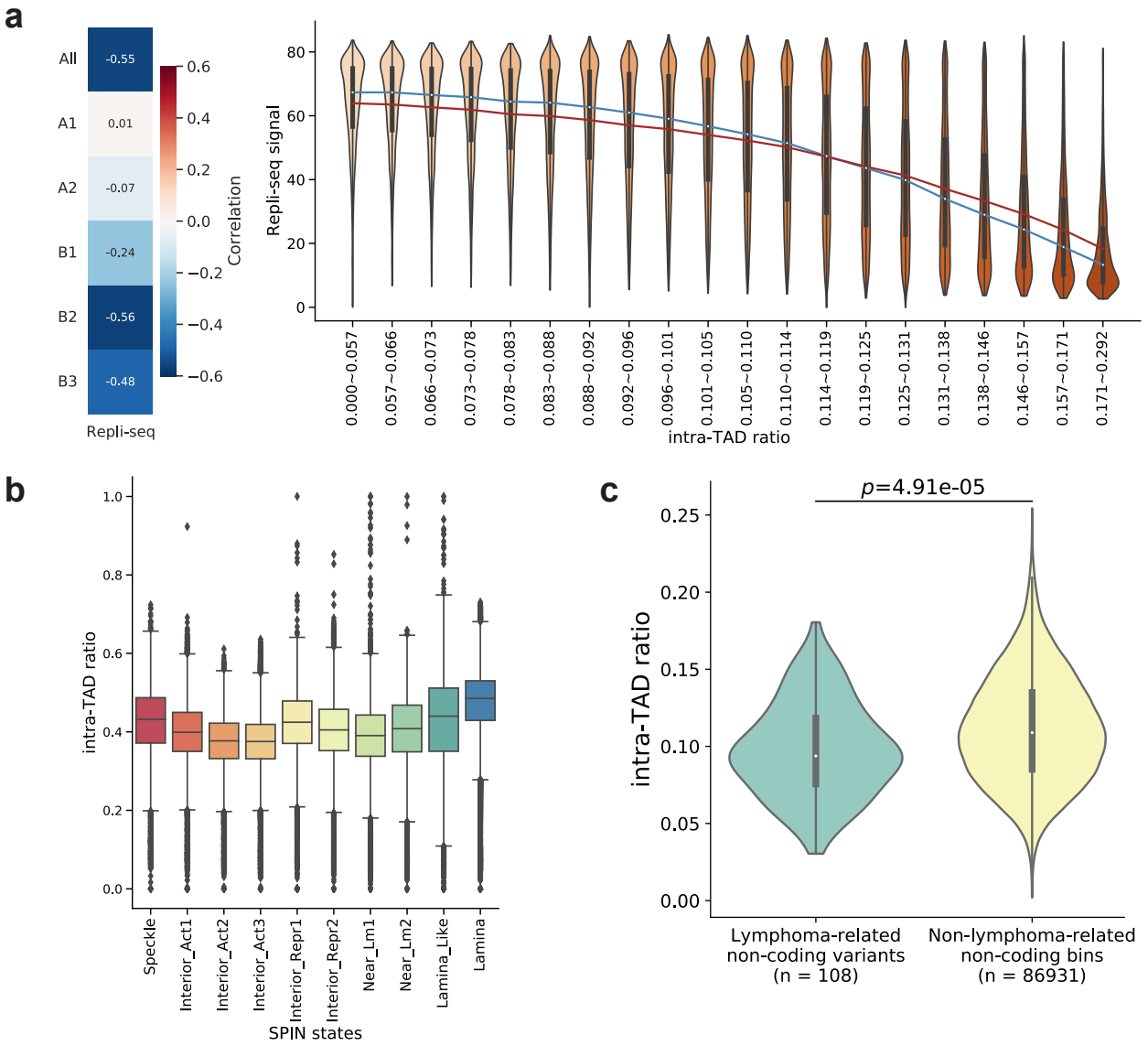

Figure 5: Intra-TAD ratio correlates with replication timing, nuclear compartments, and disease variants. (a) Replication timing is correlated with intra-TAD ratio. left panel: Spearman correlation between intra-TAD ratio and Repli-seq signals; right panel: Violin plots of Repli-seq signals in groups with increasing intra-TAD ratio and similar number of bins. The blue and red lines connect median and mean values of the different groups, respectively. (b) Distribution of intra-TAD ratios of genomic bins in individual SPIN states. Speckle: genomic regions predicted to be associated with nuclear speckles; Interior_Act1, Interior_Act2, Interior_Act3: predicted active interior regions; Interior_Repr1, Interior_Repr2: predicted repressive interior regions; Near_Lm1, Near_Lm2: regions predicted to be near nuclear lamina; Lamina_Like, and Lamina: regions predicted to be associated with nuclear lamina. (c) Violin plots of intra-TAD ratio in GM12878 of lymphoma-related non-coding variants and non-lymphoma-related non-coding bins.

S1 Supplementary information

S1 Supplementary figures

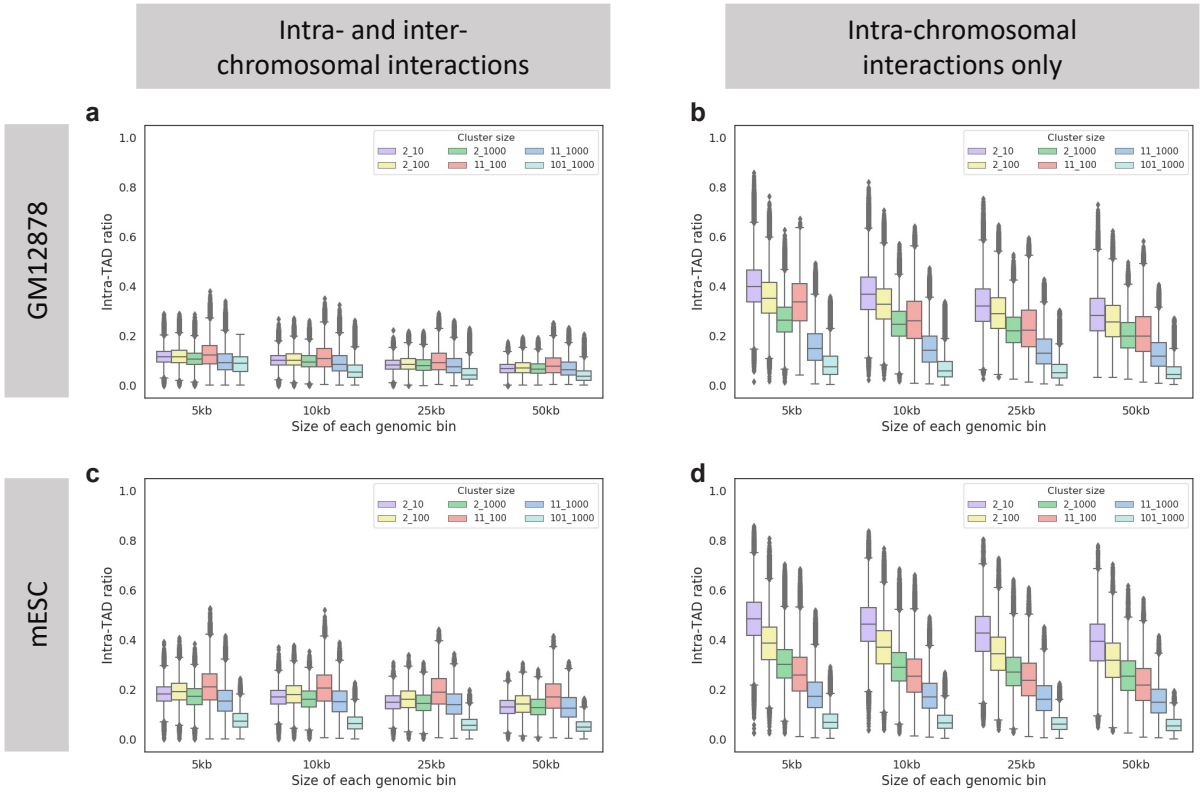

Figure S1: Effects of technical parameters on the intra-TAD ratio. The different panels show the intra-TAD ratios calculated using data from GM12878 cells (a,b) and mESCs (c,d), with both intra- and inter-chromosomal interactions (a,c) or only intra-chromosomal interactions (b,d). In each panel, the different groups show the intra-TAD ratios calculated with a specific size of genomic bins (5kb, 10kb, 25kb, or 50kb). Within a group, the different box plots correspond to different subsets of SPRITE clusters involved in the calculation of intra-TAD ratios selected based on their sizes, where x_y means the inclusion of only SPRITE clusters with at least x and at most y reads.

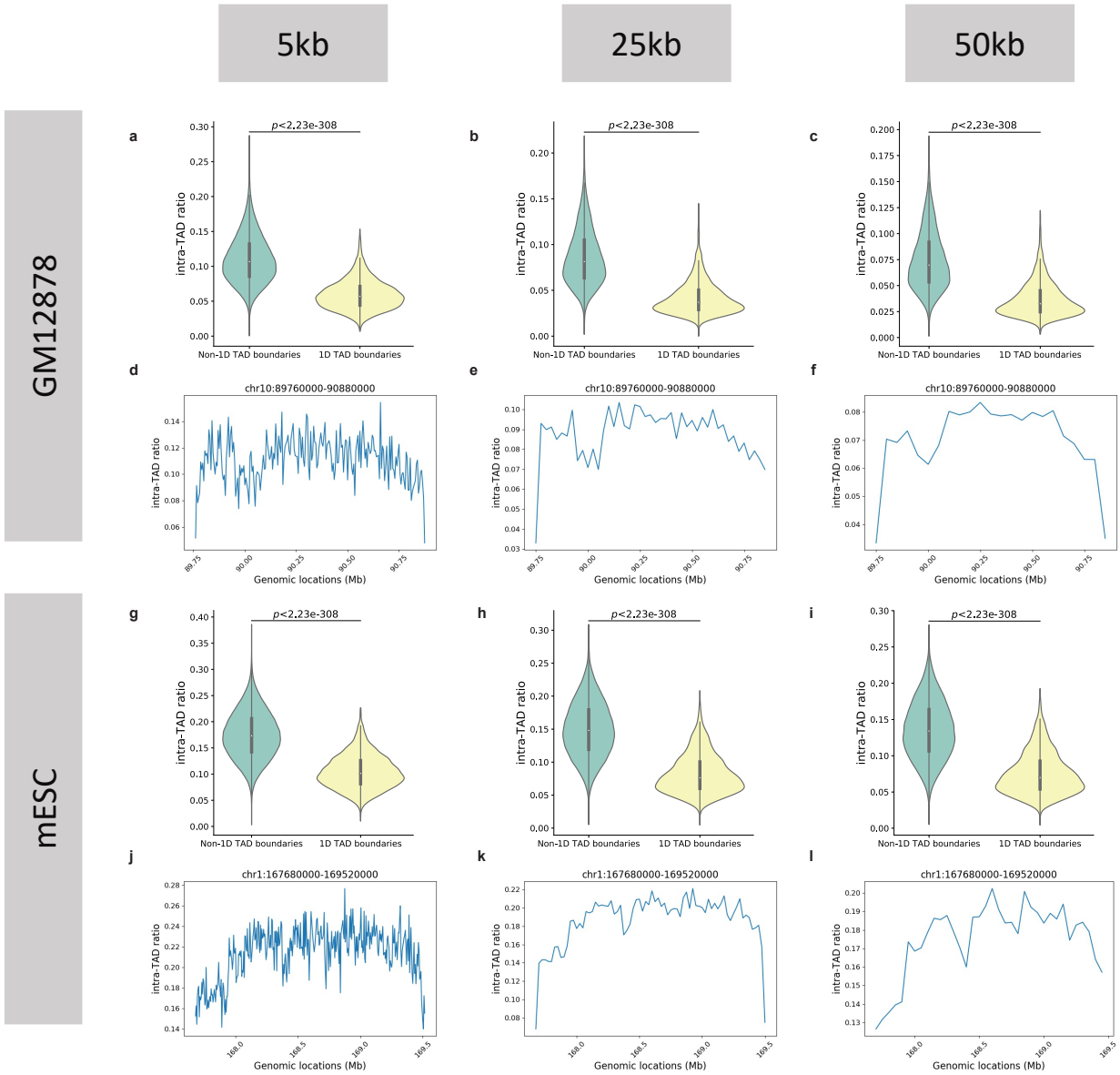

Figure S2: Comparing the intra-TAD ratios of 1D TAD boundaries and other TAD bins. The different panels show the intra-TAD ratios in GM12878 cells (a-f) and mESCs (g-l), at 5kb (a,d,g,j), 25k (b,e,h,k), and 50kb (c,f,i,l) bin sizes. The violin plots (a-c, g-i) compare the distributions of intra-TAD ratios of genomic bins that overlap with 1D TAD boundaries and those that do not. The line plots (d-f, j-l) show the change of intra-TAD ratio along the genomic span of example TADs.

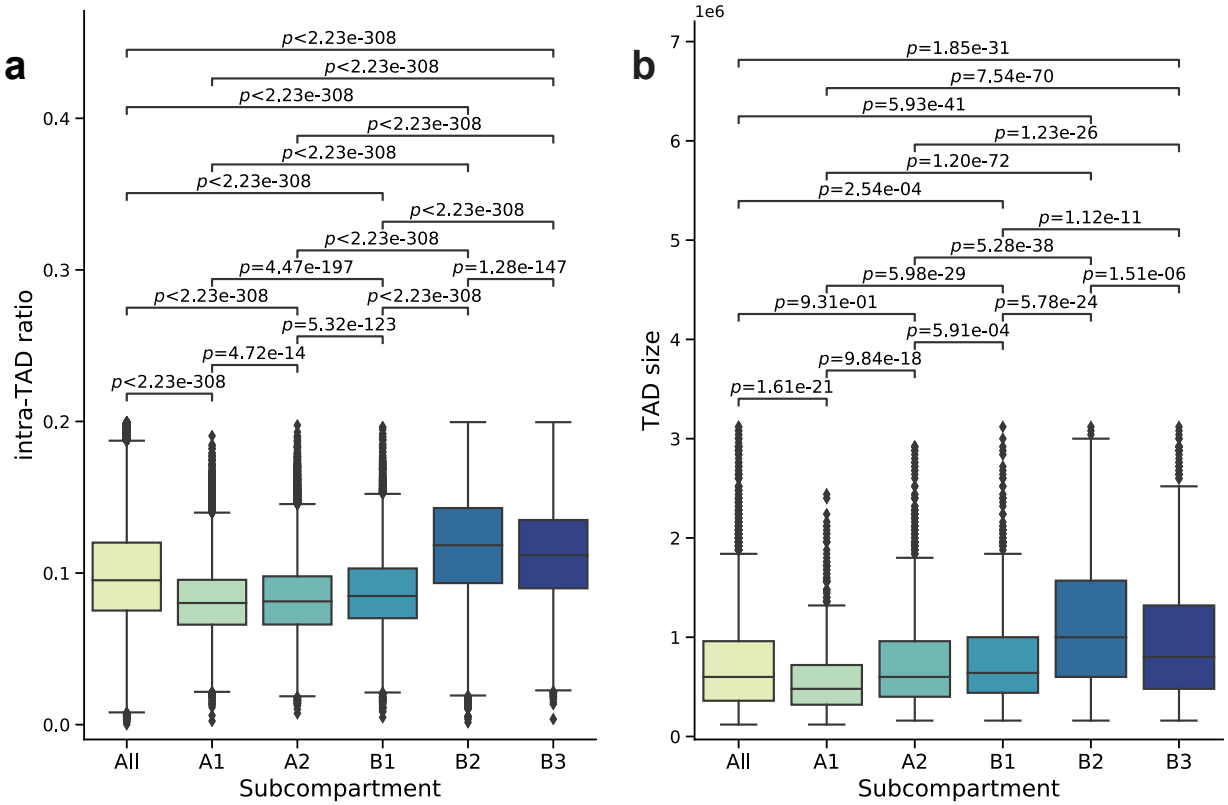

Figure S3: Distribution of intra-TAD ratio (a) and TAD size (b) of either all TADs or only TADs in a genomic subcompartment. The size of a TAD is defined as the length of its 1D genomic span.

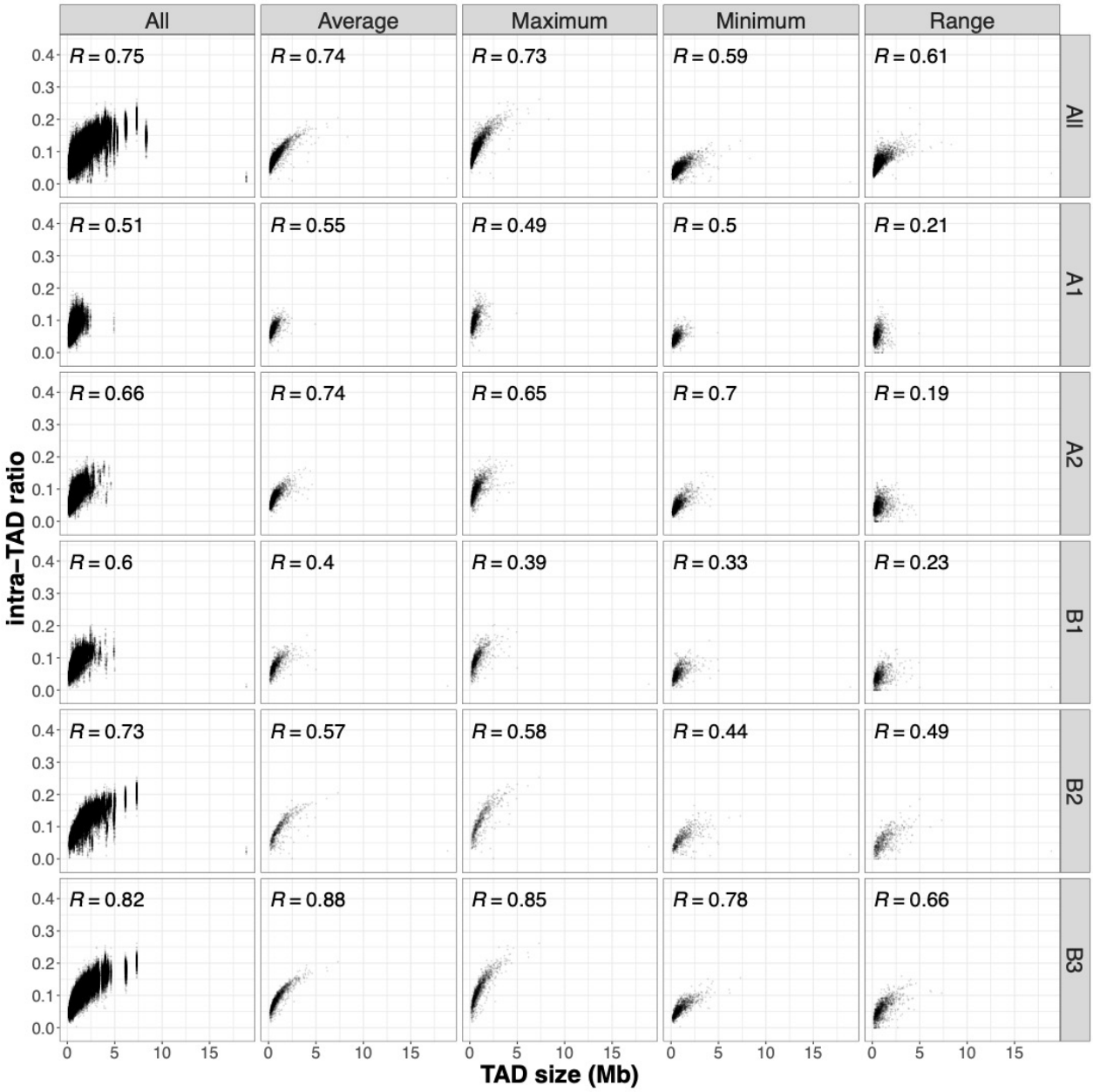

Figure S4: Relationship between intra-TAD ratio and TAD size. Each scatterplot compares TAD size with intra-TAD ratio based on either all TADs or only TADs in a genomic sub-compartment (rows) and a specific way to aggregate intra-TAD ratios of all genomic bins in a TAD (columns).

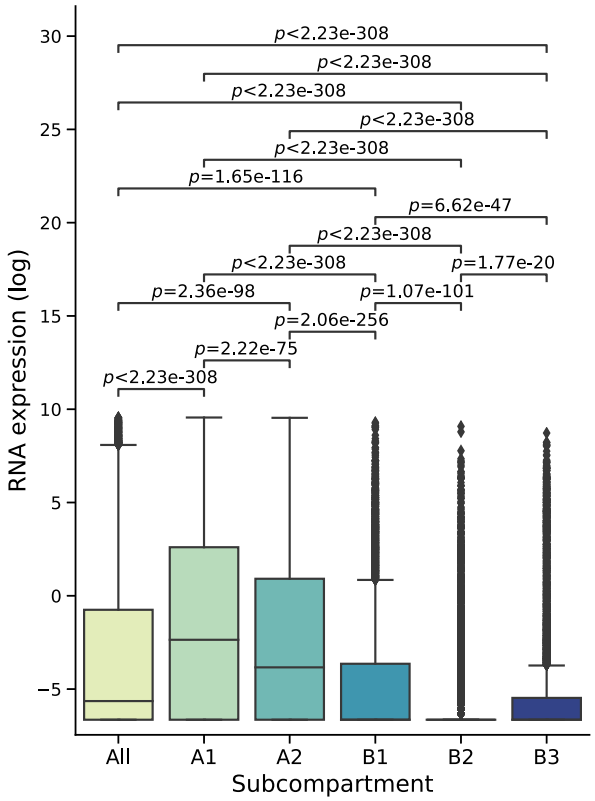

Figure S5: Distribution of expression levels of genes either in the whole genome or in a genomic subcompartment. RNA expression is quantified by $\log_2$(FPKM+0.01).

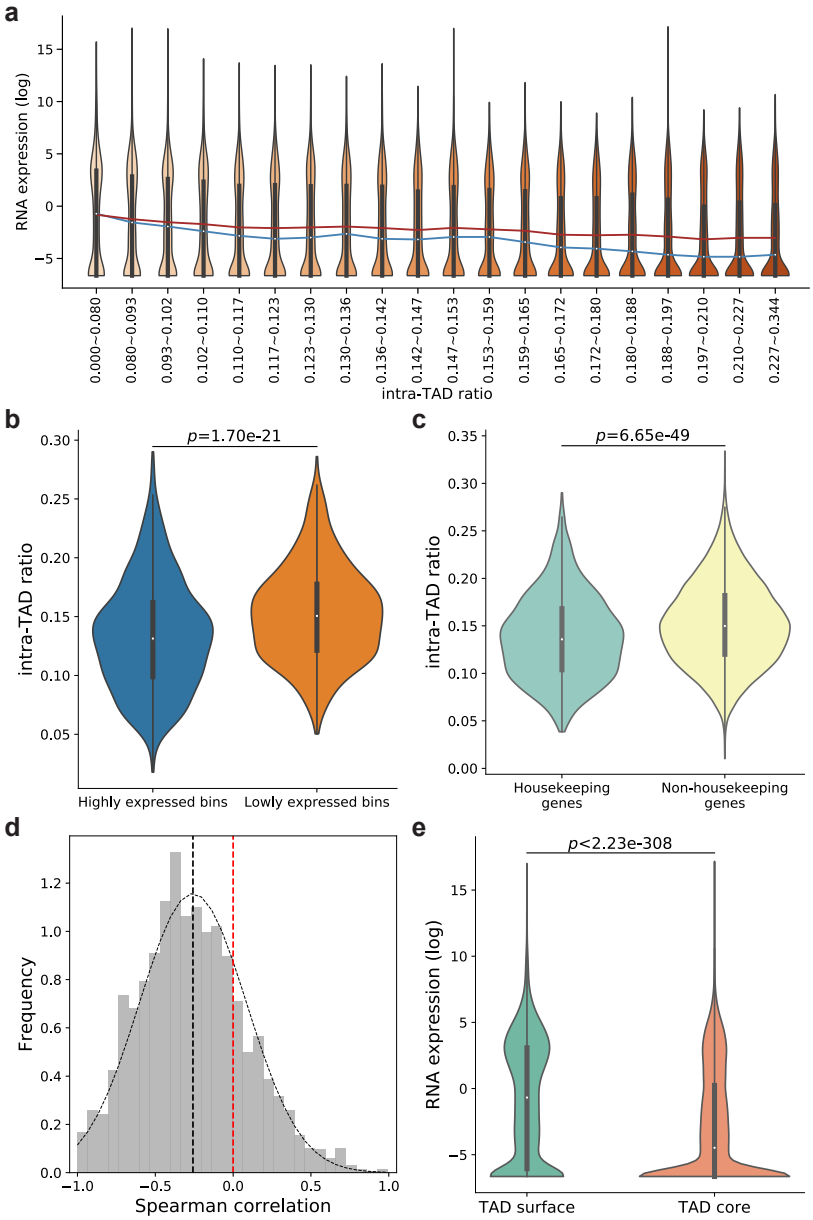

Figure S6: Inverse relationship between intra-TAD ratio and gene expression in mESCs. (a) Violin plots of gene expression ($\log_2$(FPKM+0.01)) of TSS bins in groups with increasing intra-TAD ratio and similar number of bins. The blue and red lines connect median and mean values of the different groups, respectively. (b) Violin plots of intra-TAD ratios of the highly expressed bins and lowly expressed bins, defined as the genomic bins containing TSSs of genes whose FPKM values rank within top and bottom 1,000, respectively. (c) Violin plots of intra-TAD ratio of housekeeing genes and non-housekeeping genes. (d) Distribution of Spearman correlations between gene expression and intra-TAD ratio of genomic bins in individual TADs. Black dotted curve shows the fitted normal distribution. The vertical black dotted line shows the mean of Spearman correlations. The vertical red dotted line shows zero Spearman correlation. (e) Violin plots of gene expression levels ($\log_2$(FPKM+0.01)) in TAD surface bins and TAD core bins. p-values in all panels are calculated using the Mann-Whitney U test.

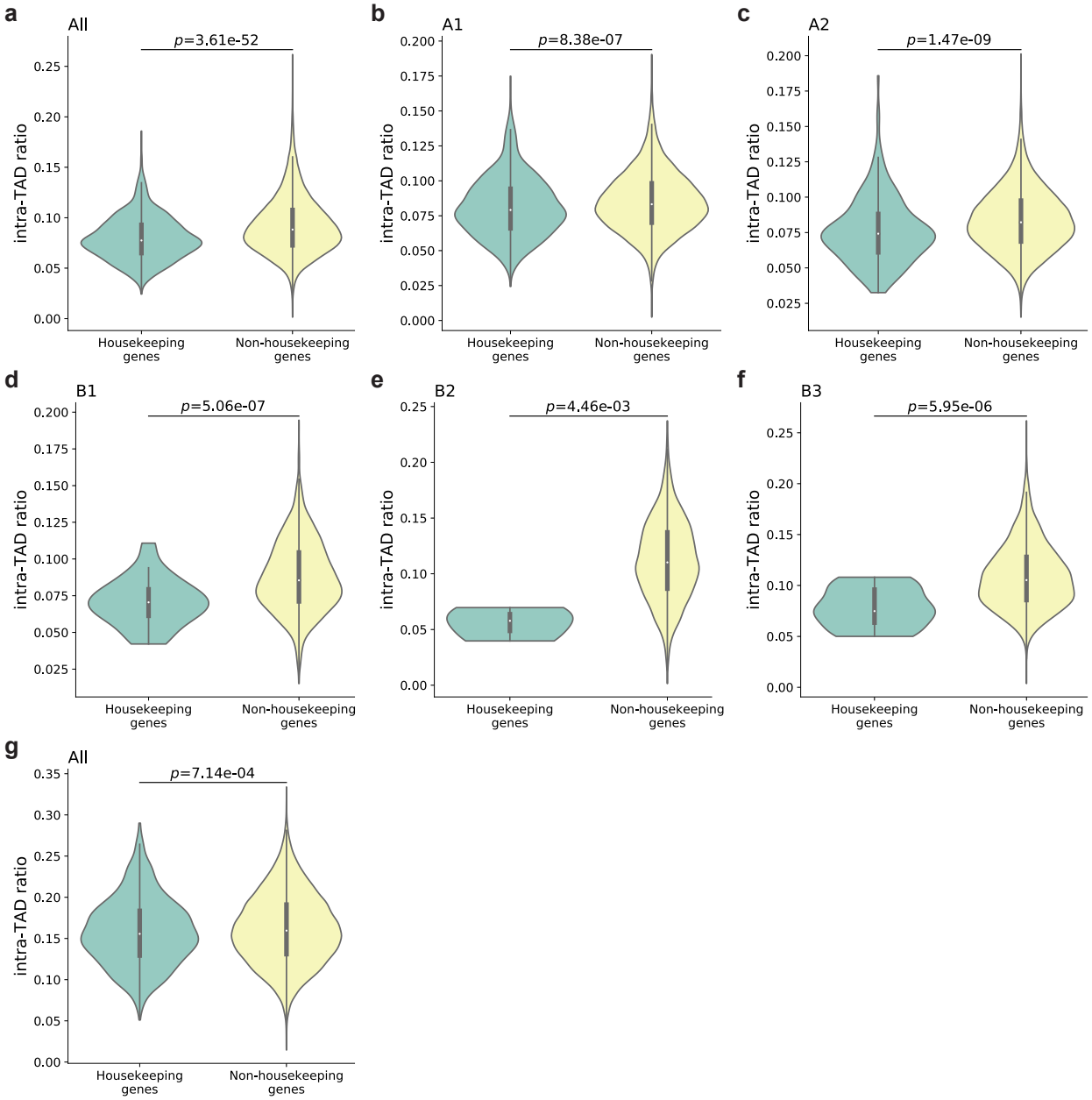

Figure S7: Violin plots of intra-TAD ratio of housekeeping genes and non-housekeeping genes when TAD boundaries and their neighboring regions were excluded in GM12878 (a-f) and mESC (g). p-values in all panels are calculated using the Mann-Whitney U test.

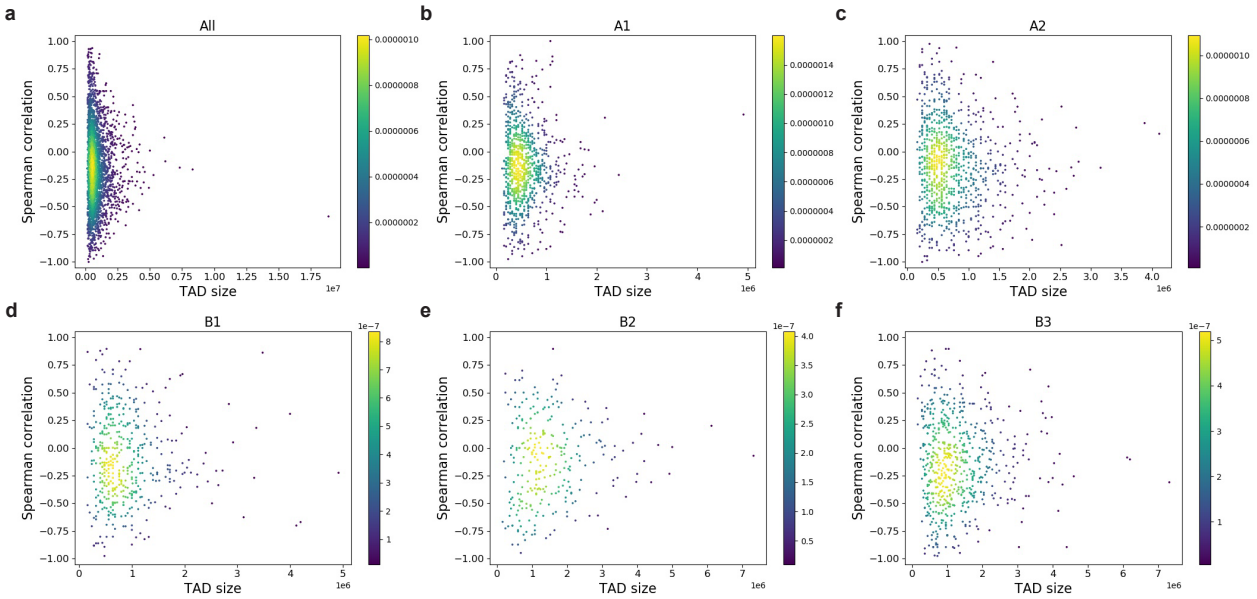

Figure S8: Density plots showing the Spearman correlation between intra-TAD ratio and gene expression of individual TADs versus TAD size for either all genomic genes (a) or genes in a genomic subcompartment (b-f).

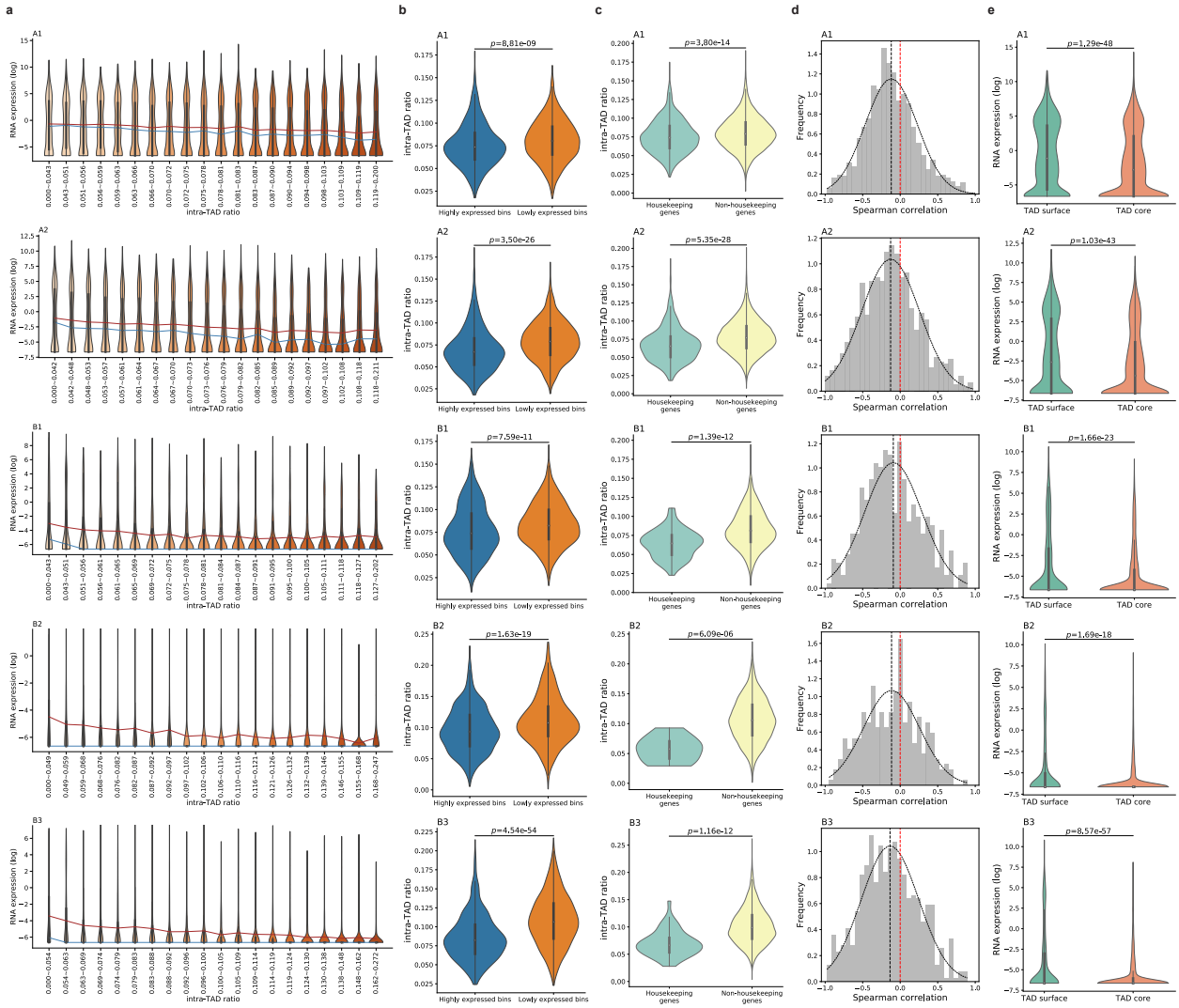

Figure S9: Intra-TAD ratio and gene expression have an inverse relationship in GM12878 for genes in individual subcompartments. (a) Violin plots of gene expression ($\log_2$(FPKM+0.01)) of TSS bins in groups with increasing intra-TAD ratio and similar number of bins. The blue and red lines connect median and mean values of the different groups, respectively. (b) Violin plots of intra-TAD ratios of the highly expressed bins and lowly expressed bins, defined as the genomic bins containing TSSs of genes whose FPKM values rank within top and bottom 1,000, respectively. (c) Violin plots of intra-TAD ratio of housekeeing genes and non-housekeeping genes. (d) Distribution of Spearman correlations between gene expression and intra-TAD ratio of genomic bins in individual TADs. Black dotted curve shows the fitted normal distribution. The vertical black dotted line shows the mean of Spearman correlations. The vertical red dotted line shows zero Spearman correlation. (e) Violin plots of gene expression levels ($\log_2$(FPKM+0.01)) in TAD surface bins and TAD core bins. p-values in all panels are calculated using the Mann-Whitney U test.

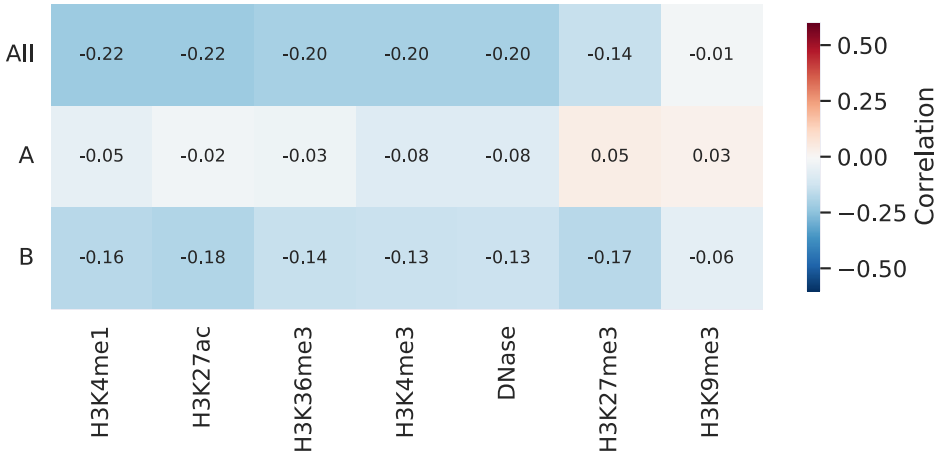

Figure S10: Spearman correlation between intra-TAD ratio and DNase-seq signals and 6 histone modifications across either all genomic bins or only bins in a genomic compartment in mESC. The columns are sorted by the correlation values when all TADs are considered (in the "All" row).

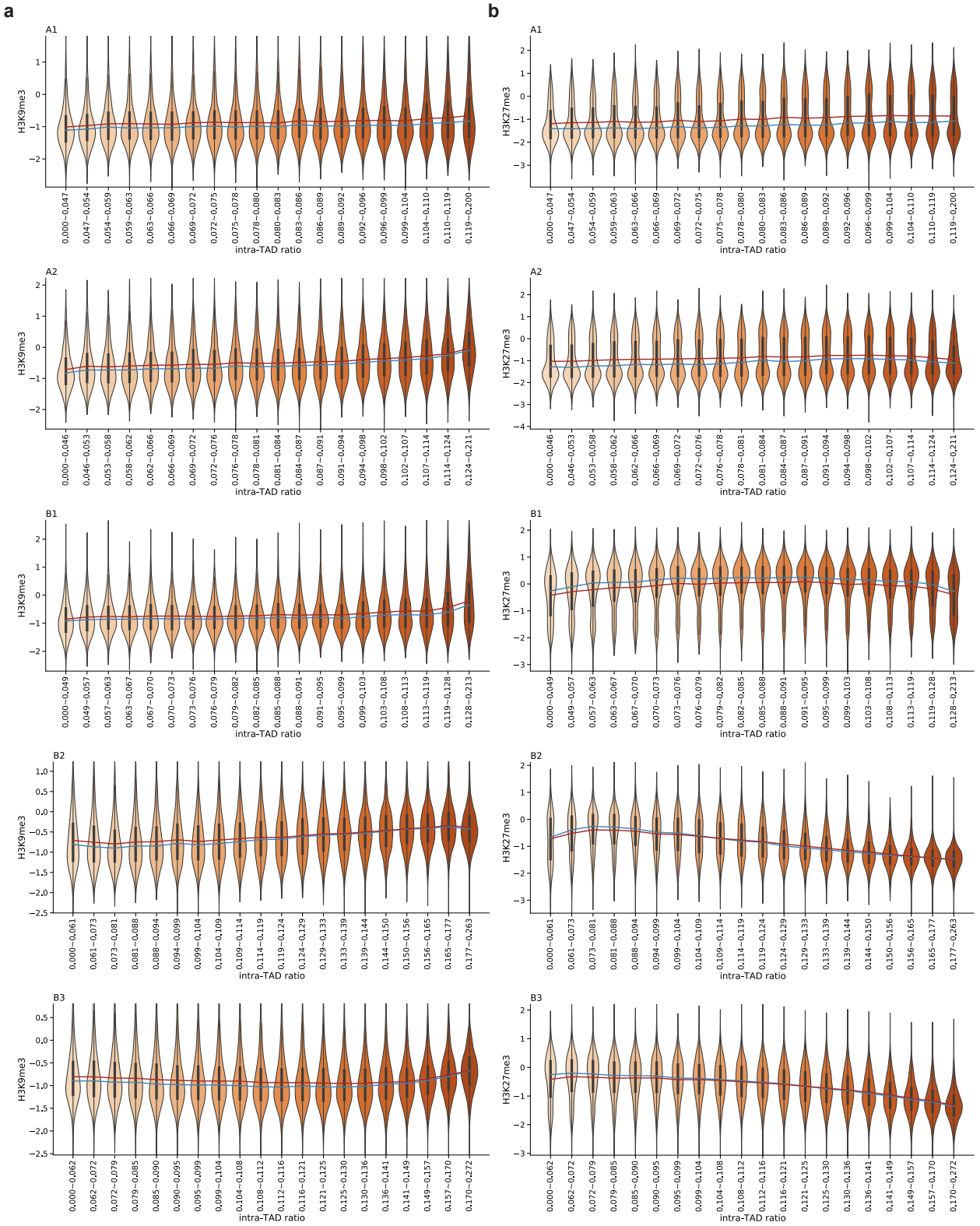

Figure S11: Violin plots of H3K9me3 (a) and H3K27me3 (b) signals in groups with increasing intra-TAD ratio and similar number of bins in individual subcompartments. y-axis shows $\log_2$(signal+0.01). The blue and red lines connect median and mean values of the different groups, respectively.

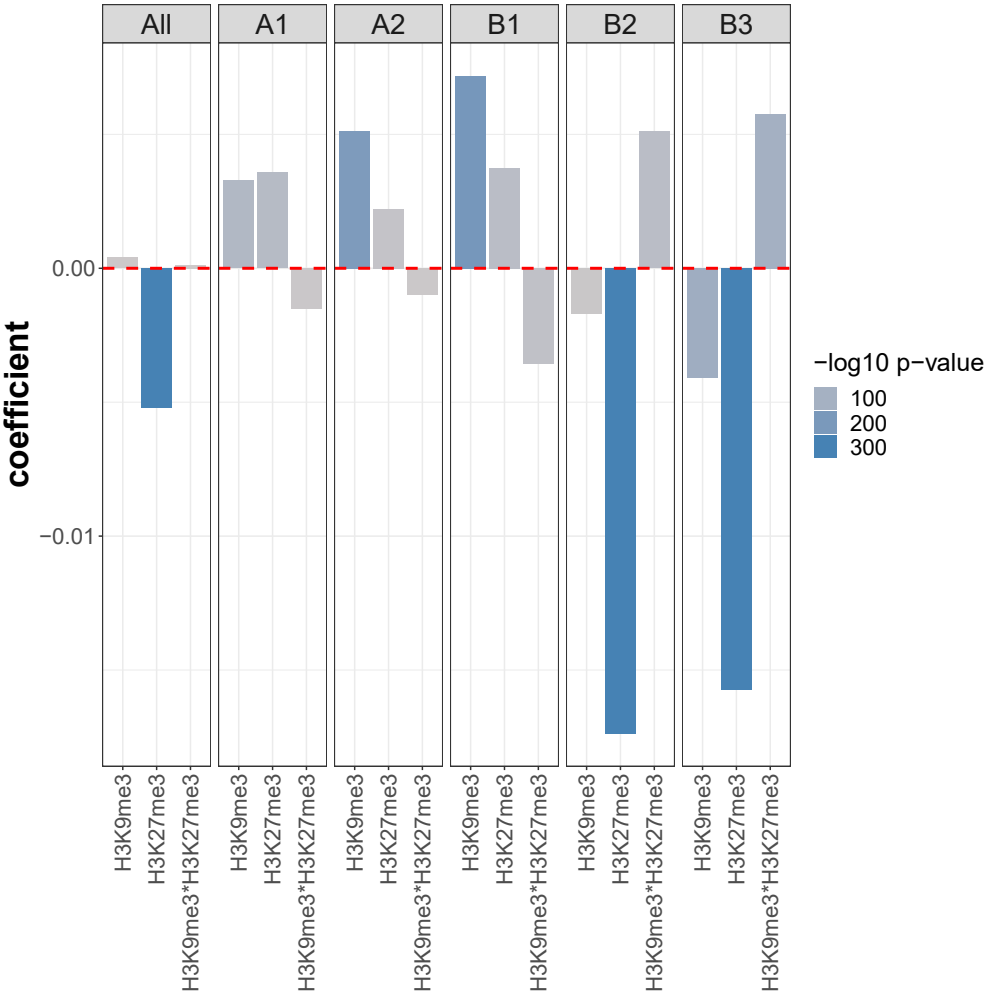

Figure S12: Coefficients of the linear regression models with H3K9me3, H3K27me3, and their interaction term as explanatory variables and intra-TAD ratio as the response variable, considering either all TADs or only TADs in a genomic subcompartment. The explanatory variables were all standardized individually before fitting the linear regression models.

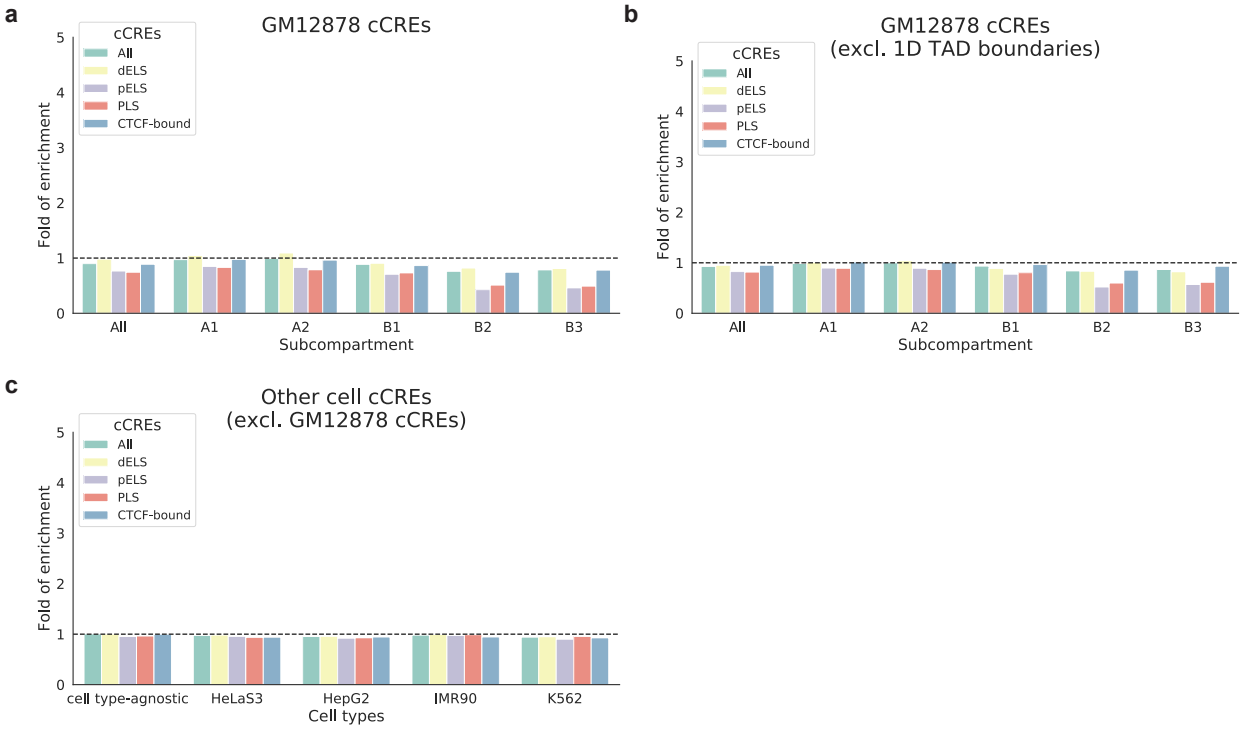

Figure S13: Depletion of CREs in TAD cores. (a) Fold of enrichment of cCREs on TAD cores in GM12878. (b) Fold of enrichment of cCREs on TAD cores in GM12878 when the regions around TAD boundaries are excluded. (c) Fold of enrichment of cCREs on TAD cores when the cCREs are cell type-agnostic or defined in other cell types and the GM12878-specific cCREs are excluded. cCREs All: all categories of cCREs; dELS: distal enhancer-like signatures; pELS: proximal enhancer-like signatures; PLS: promoter-like signatures; CTCF-bound: bound by CTCF.

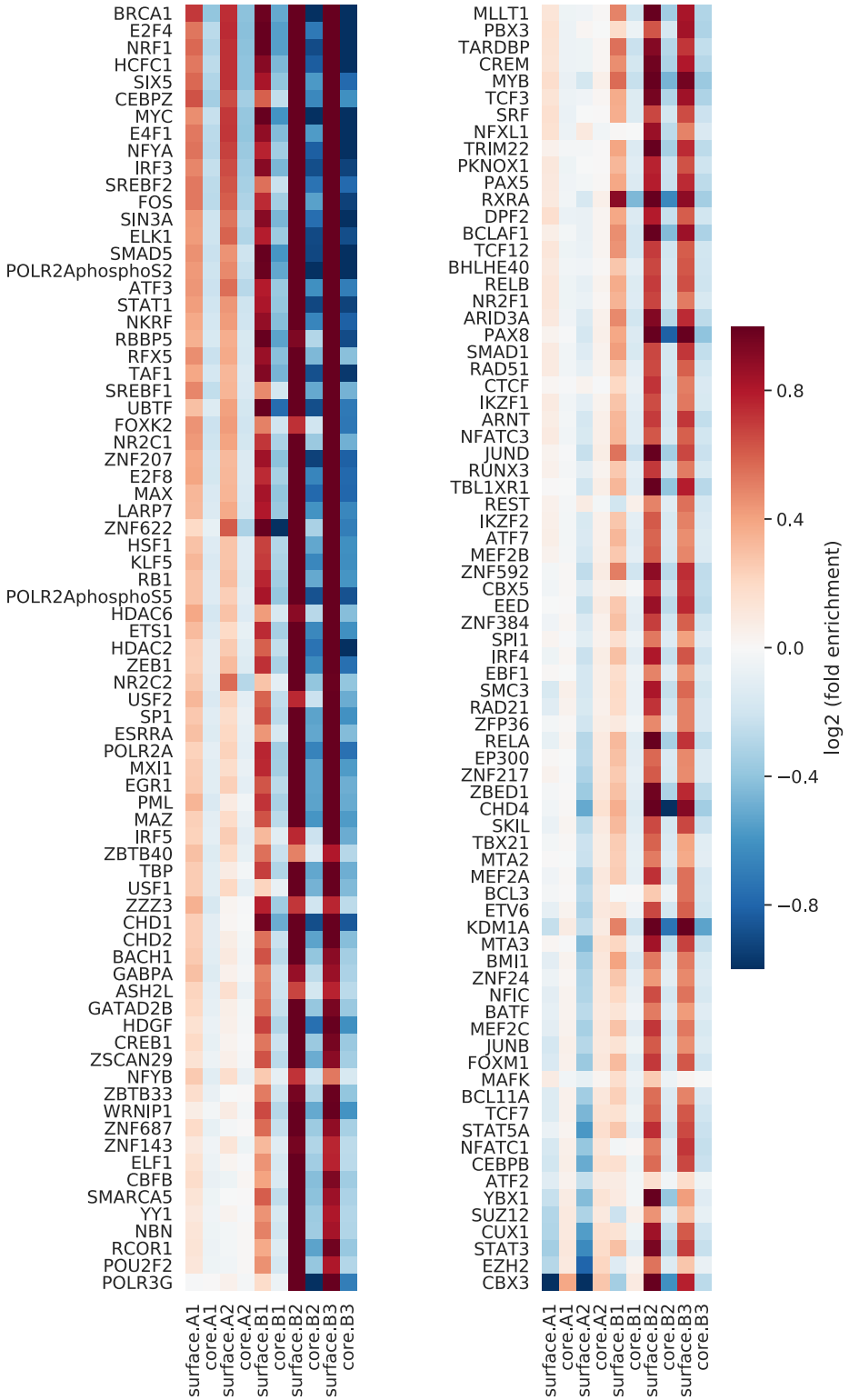

Figure S14: Heatmap showing log2(enrichment fold) of transcription factor binding sites in individual subcomaprtments on TAD surfaces (odd-numbered columns) and TAD cores (even-numbered columns) in GM12878.

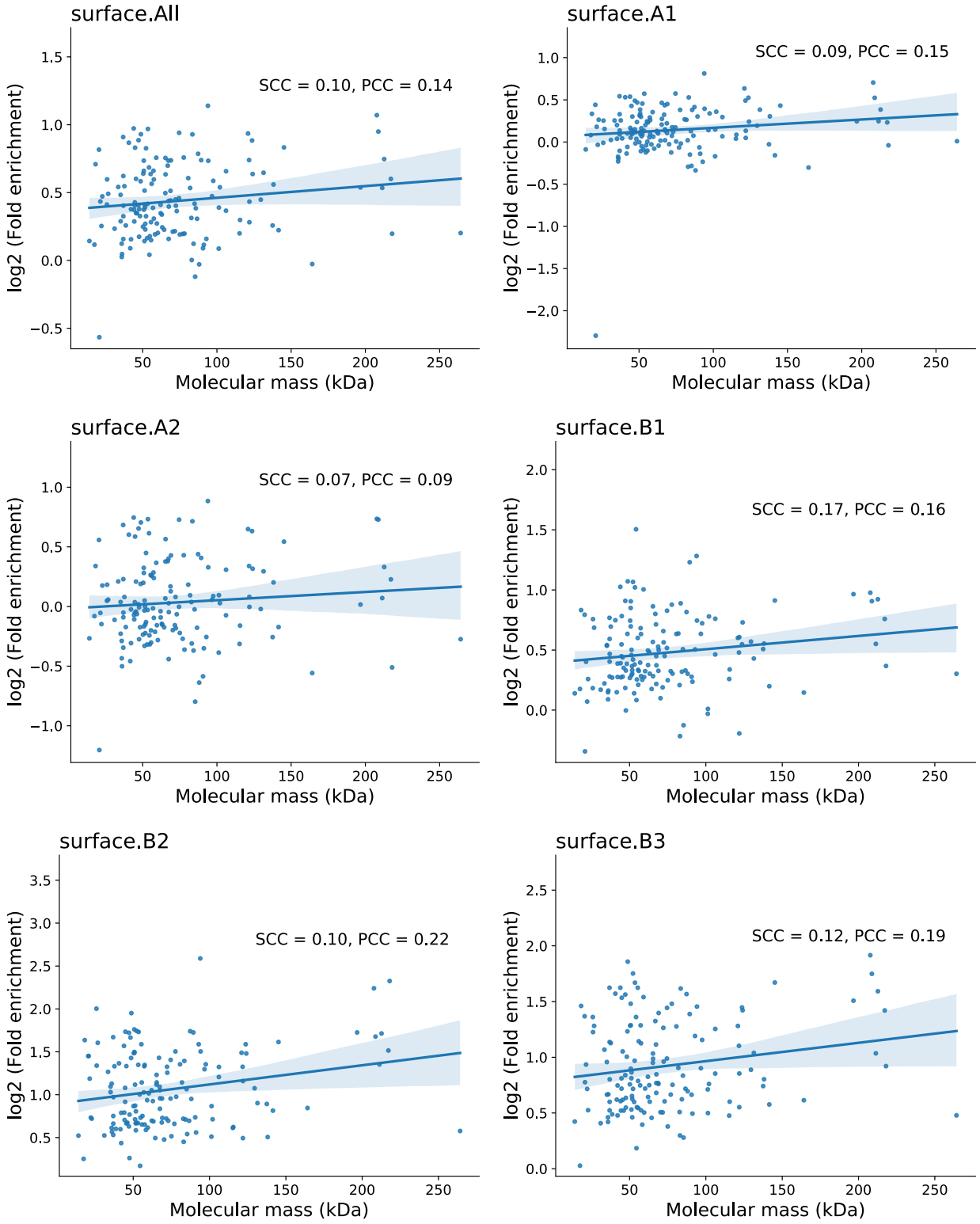

Figure S15: Scatter plots of enrichment of TFs on TAD surfaces versus molecular weights of TFs.

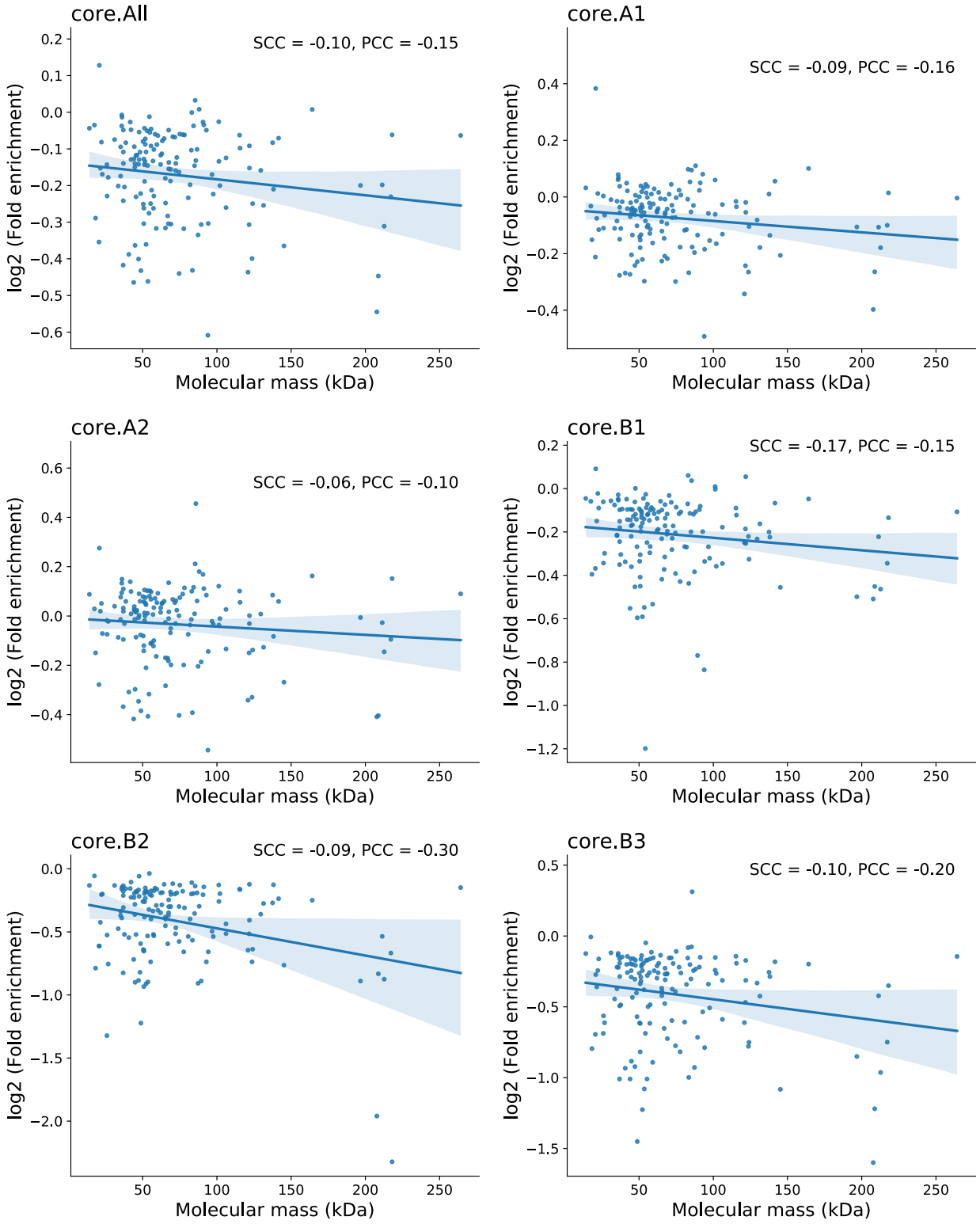

Figure S16: Scatter plots of enrichment of TFs in TAD cores versus molecular weights of TFs.

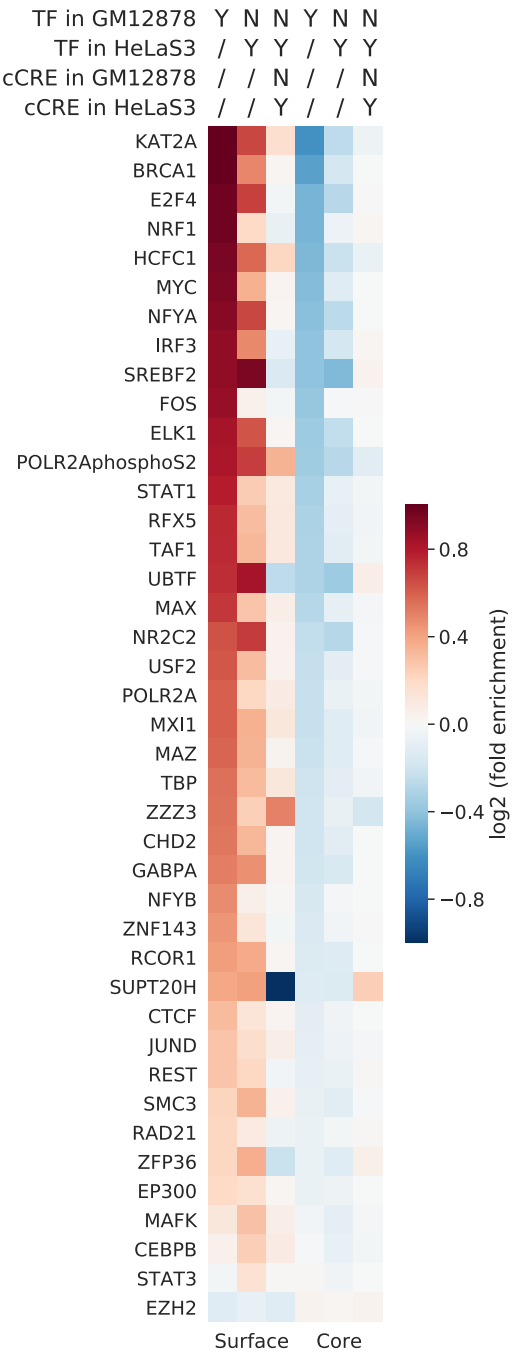

Figure S17: Heatmap comparing the locations of transcription factor binding sites and CREs in HeLaS3 and GM12878 relative to the TADs in GM12878. The first three columns show the enrichment of these locations on the GM12878 TAD surfaces. The last three columns show the enrichment of these locations in the GM12878 TAD cores. In the header, "Y" means only regions satisfying the condition are included, "N" means only regions not satisfying the condition are included, and "/" means the condition is not involved in defining the regions and thus no regions are excluded. The final set of regions is the intersection of the regions satisfying the different specified conditions.

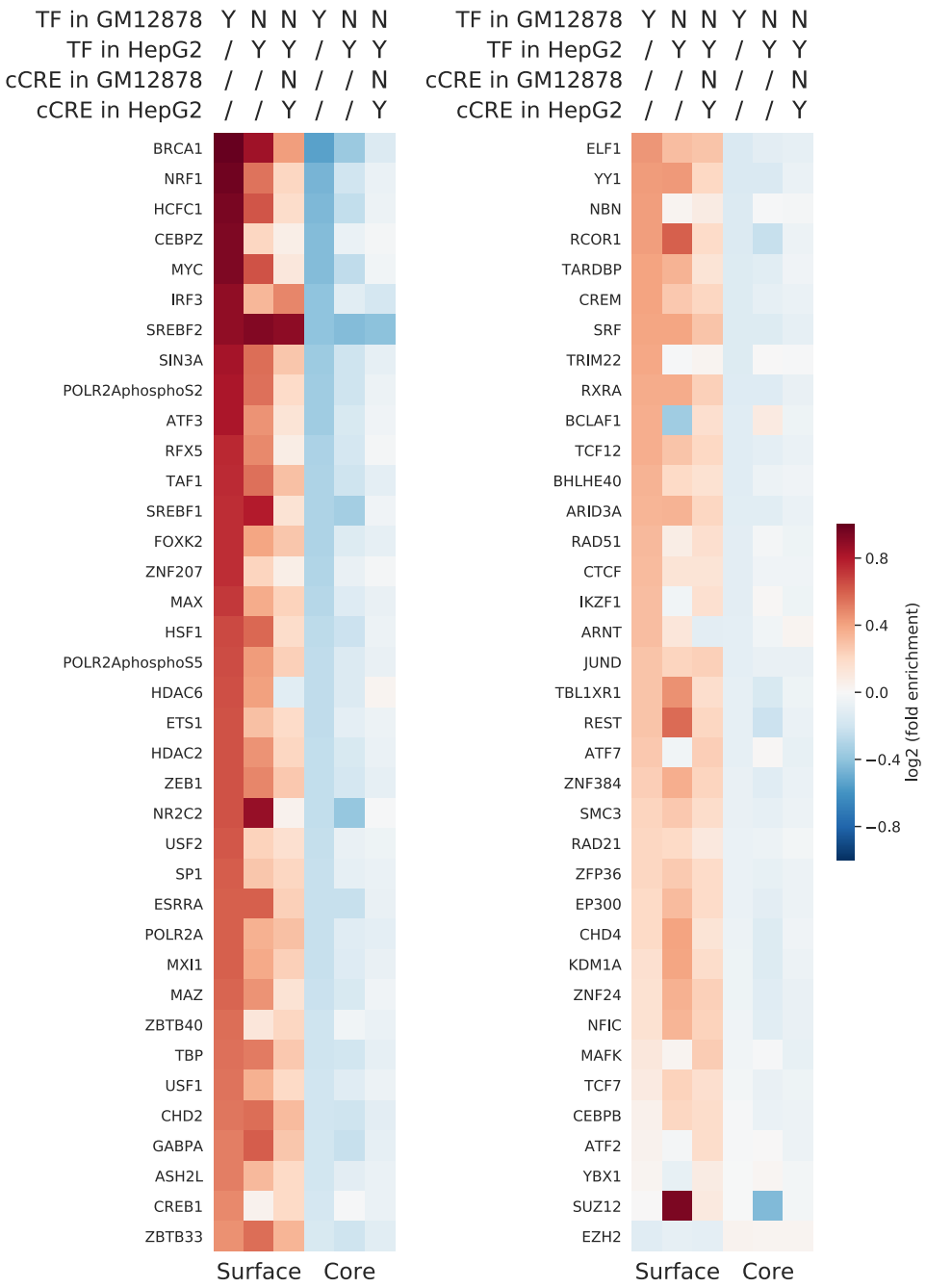

Figure S18: Heatmap comparing the locations of transcription factor binding sites and CREs in HepG2 and GM12878 relative to the TADs in GM12878. The first three columns show the enrichment of these locations on the GM12878 TAD surfaces. The last three columns show the enrichment of these locations in the GM12878 TAD cores. In the header, "Y" means only regions satisfying the condition are included, "N" means only regions not satisfying the condition are included, and "/" means the condition is not involved in defining the regions and thus no regions are excluded. The final set of regions is the intersection of the regions satisfying the different specified conditions.

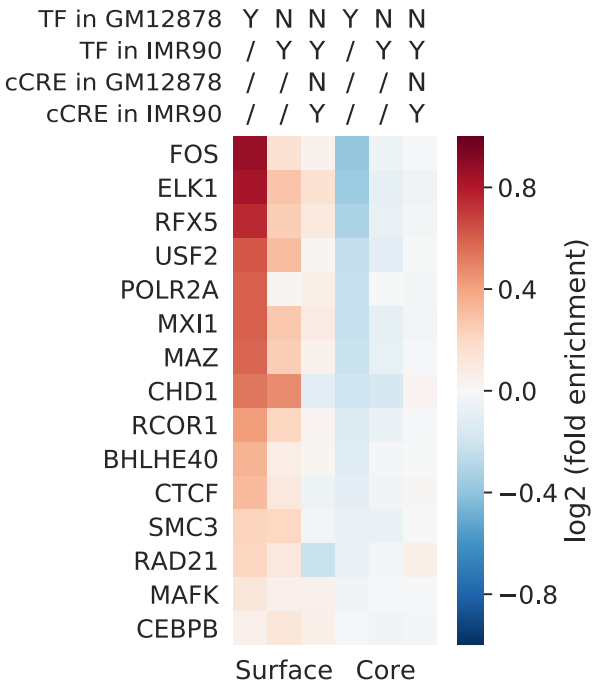

Figure S19: Heatmap comparing the locations of transcription factor binding sites and CREs in IMR90 and GM12878 relative to the TADs in GM12878. The first three columns show the enrichment of these locations on the GM12878 TAD surfaces. The last three columns show the enrichment of these locations in the GM12878 TAD cores. In the header, "Y" means only regions satisfying the condition are included, "N" means only regions not satisfying the condition are included, and "/" means the condition is not involved in defining the regions and thus no regions are excluded. The final set of regions is the intersection of the regions satisfying the different specified conditions.

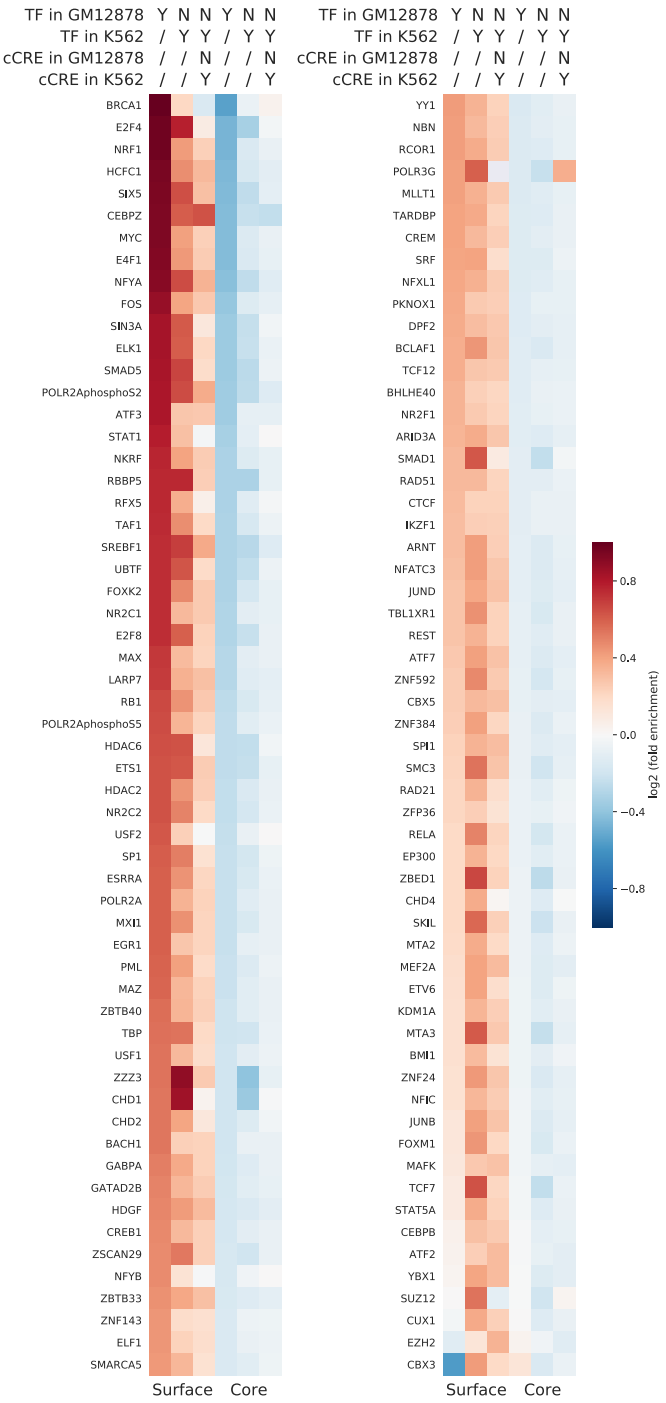

Figure S20: Heatmap comparing the locations of transcription factor binding sites and CREs in K562 and GM12878 relative to the TADs in GM12878. The first three columns show the enrichment of these locations on the GM12878 TAD surfaces. The last three columns show the enrichment of these locations in the GM12878 TAD cores. In the header, "Y" means only regions satisfying the condition are included, "N" means only regions not satisfying the condition are included, and "/" means the condition is not involved in defining the regions and thus no regions are excluded. The final set of regions is the intersection of the regions satisfying the different specified conditions.

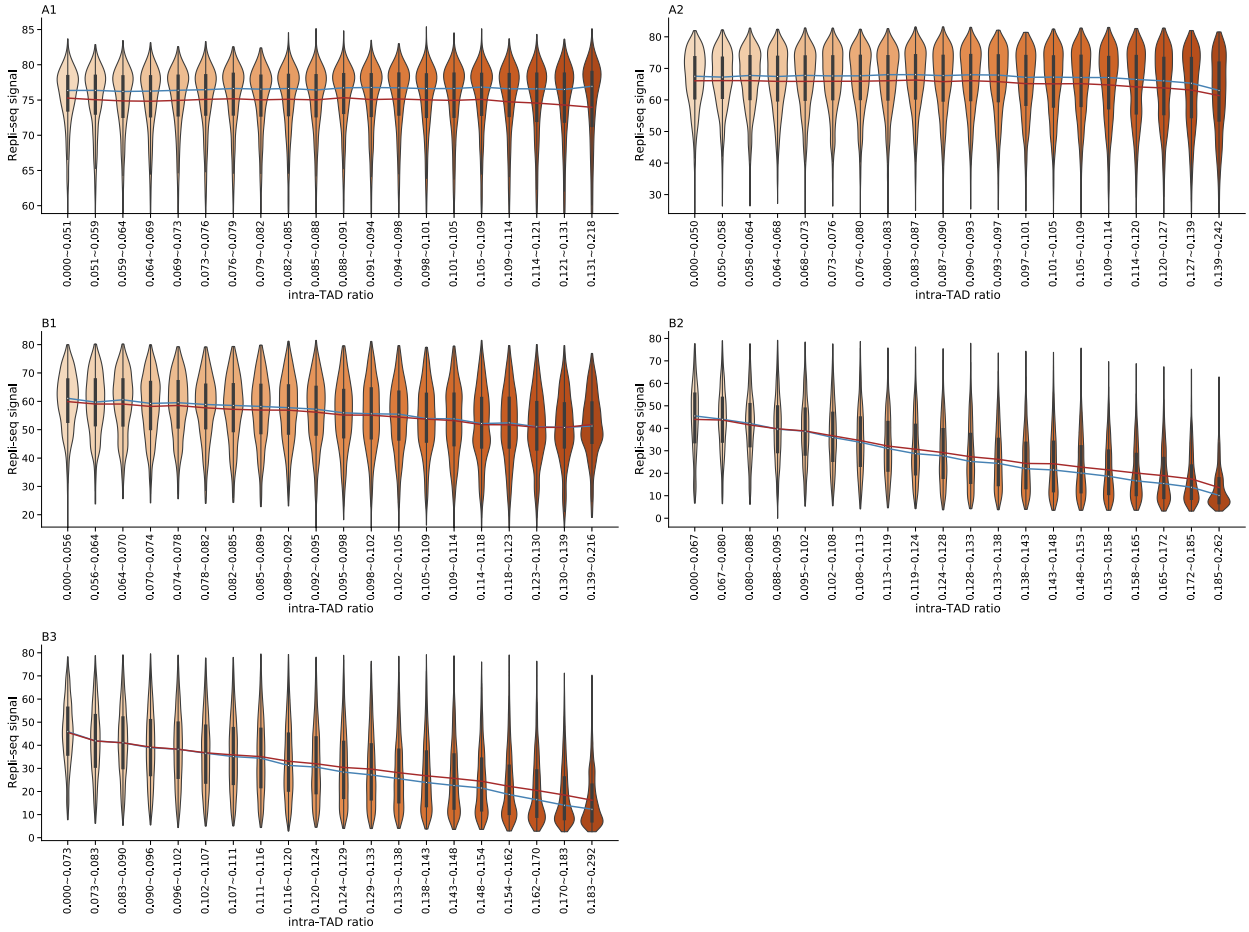

Figure S21: Violin plots of Repli-seq signals in groups with increasing intra-TAD ratio and similar number of bins in individual subcompartments. The blue and red line connects median and mean values, respectively, in each group.