Tracing the Shared Foundations of Gene Expression and Chromatin Structure

Huan Liang[1], Bonnie Berger[3-4,*], Rohit Singh[1-2,*]

[1]Department of Biostatistics and Bioinformatics, Duke University
[2]Department of Cell Biology, Duke University
[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
[4]Department of Mathematics, Massachusetts Institute of Technology
[*]Corresponding authors: bab@mit.edu, rohit.singh@duke.edu

Summary

The three-dimensional organization of chromatin into topologically associating domains (TADs) may impact gene regulation by bringing distant genes into contact. However, many questions about TADs' function and their influence on transcription remain unresolved due to technical limitations in defining TAD boundaries and measuring the direct effect that TADs have on gene expression. Here, we develop consensus TAD maps for human and mouse with a novel "bag-of-genes" approach for defining the gene composition within TADs. This approach enables new functional interpretations of TADs by providing a way to capture species-level differences in chromatin organization. We also leverage a generative AI foundation model computed from 33 million transcriptomes to define contextual similarity, an embedding-based metric that is more powerful than co-expression at representing functional gene relationships. Our analytical framework directly leads to testable hypotheses about chromatin organization across cellular states. We find that TADs play an active role in facilitating gene co-regulation, possibly through a mechanism involving transcriptional condensates. We also discover that the TAD-linked enhancement of transcriptional context is strongest in early developmental stages and systematically declines with aging. Investigation of cancer cells show distinct patterns of TAD usage that shift with chemotherapy treatment, suggesting specific roles for TAD-mediated regulation in cellular development and plasticity. Finally, we develop "TAD signatures" to improve statistical analysis of single-cell transcriptomic data sets in predicting cancer cell-line drug response. These findings reshape our understanding of cellular plasticity in development and disease, indicating that chromatin organization acts through probabilistic mechanisms rather than deterministic rules.

Software availability: `https://singhlab.net/tadmap`

Keywords— Topologically associating domain, 3D genome, gene expression, bag of words, single-cell foundation models, contextual transcriptional similarity, nuclear plasticity

Introduction

Cells must precisely orchestrate the activity of hundreds of genes to perform even basic functions. This orchestration is an information-theoretic problem—how to specify exact gene combinations—that must be solved through physical mechanisms in the crowded confines of the nucleus. Evolution has shaped distinct solutions to this problem. Bacteria elegantly address this challenge by operons, organizing co-regulated genes linearly along their genome [1]. However, mammalian and other animal cells face a more complex task. While the gene ordering is fixed, diverse cell types need to carry out their functions, each with distinct combinations of genes. These organisms have evolved an additional solution to this challenge, developing a dynamic three-dimensional genome organization that uses chromatin structure to bring distant genes into contact [2, 3].

Modern genomic technologies, such as Hi-C, GAM, SPRITE, and ChIA-PET etc. [4–8], have unveiled the cell's three-dimensional chromatin architecture, operating at multiple scales. Topologically associating domains (TADs) are key sub-components in this hierarchical organization [2, 9–11]—regions where DNA frequently contact within the TAD but are restricted externally. These domains create distinct neighborhoods within the genome [12–14] in which the transcriptional machinery operates. A central question emerges: has evolution shaped a synergy between gene location and the chromatin's TAD structure to support the recurring expression patterns needed by diverse cell types?

TADs present a paradox. Although their importance is clear, their function and mechanism remain disputed. Do TADs actively facilitate co-regulation of their resident genes, or do they mainly function as passive insulators, protecting genes from outside regulatory interactions? What is their relationship with linear genomic distance—do they override or complement evolutionary pressures that might have grouped functionally-related genes together? The field lacks consensus, even on these basic questions about TADs role in gene regulation [15]. This uncertainty stems from several technical challenges. TAD boundaries vary somewhat between cell types and individual cells. Studies of CTCF, a protein central to TAD formation, have yielded inconsistent results [16–21], and its lethality when deleted has hindered comprehensive in vivo investigation. The complexity of measuring gene co-transcription adds another layer of difficulty. Bulk RNA-seq datasets have limited sample sizes, while single-cell RNA-seq introduces technical noise and data sparsity. Some gene-dense TADs, such as those containing olfactory receptor genes, show minimal co-expression [22, 23], further complicating our understanding. To address these challenges, we pursue the hypothesis that there are systematic patterns in how TADs work with genomic proximity to organize gene regulation across cell types.

Beyond cell type-specific analyses, we further propose that understanding the systematic role of TADs requires the identification of species-level principles of how chromatin structure and genomic proximity cooperate in gene regulation. One challenge is that, as a measure of transcriptional similarity, co-expression is too narrow to capture these principles. In addition, TADs have been characterized currently for only a subset of cell types. To overcome these limitations, here we develop new foundation model representations of TADs and transcriptional similarity through massive-scale data integration that spans 33 million single cells, cell type-specific Hi-C maps, 49 GTEx tissues, genome-wide perturbation screens, and intergenic transcripts. Our representation learning [24] approach introduces two key advances, the *TAD Map* and *contextual transcriptional similarity*.

The TAD Map gives a species-level view of chromatin structure which holds broadly across cell types. Extensive evidence has shown that TADs maintain their organization across diverse cellular contexts [25–29]. Leveraging this conservation, we identify biologically meaningful gene groupings at the species level. We represent each TAD simply as its contained genes—what we call a “bag of genes” model (**Fig. 1A**), inspired by set-based sentence representations in natural language modeling and “bag of fragments” approach for protein structure modeling [30]. This abstraction, the *TAD Map*, enables systematic analysis of genome organization without requiring Hi-C data for each cell type. More fundamentally, it provides a new framework for understanding how chromatin structure shapes transcriptional regulation.

Traditional co-expression measures fail to capture the full complexity of gene relationships. Although expression correlation can detect simultaneous activation, it misses crucial patterns such as coordinated repression, especially given sparse single-cell data due to low sequencing depth and dropouts. We need a broader view of how genes relate to each other across cellular states. Consider genes as Lego blocks, each fitting into specific transcriptional roles defined by other genes. Most genes have unique shapes, but certain families, such as olfactory receptors and protocadherins, act like blocks of identical shape but different colors. They are relevant to an individual cell’s identity, but their roles vis-à-vis other genes are interchangeable. Such gene families, which comprise some of the most populous TADs, confound simple co-expression analysis because family members are rarely co-expressed. To capture these nuanced relationships, we introduce contextual transcriptional similarity (CTS, or just *contextual similarity*). CTS leverages scGPT [31], a single-cell foundation model trained on over 33 million cells, to reveal gene relationships that evade traditional correlation analysis.

We show that integrating the TAD Map and contextual similarity reveals unexpected principles of gene regulation. We find that TADs function as loci of increased transcriptional context, significantly exceeding what genomic proximity alone would predict, with TADs boosting contextual similarity by 20.5% compared to non-TAD regions. Strikingly, this enhancement persists across genomic distances. Although contextual similarity declines with distance both within and outside TADs, the enhancement remains multiplicative at all scales. This distance-dependent pattern suggests an underlying physical process—possibly diffusion-based—that operates more efficiently within TAD regions. Building on this insight, we explore a hypothesis linking TADs to phase-separated nuclear compartments, particularly transcriptional condensates [32–36].

We posit that TADs work synergistically with these phase-separated compartments to create specialized regulatory environments [37–39]. Within a TAD, the observed effects may arise from either larger-scale condensates spanning significant portions of the TAD, or from the collective action of multiple smaller transcriptional condensates operating at different loci [40, 41]. In this model, TADs serve as organizational scaffolds that integrate the effects of many distinct condensates, leading to emergent regulatory properties. This model provides a unified explanation for our key findings: elevated contextual similarity in TADs, its multiplicative enhancement pattern, and characteristic distance dependence. We find this hypothesis is supported by multiple independent lines of evidence, including perturbational studies of condensate genes, analysis of how gene orientation constrains transcription, and mapping of transcriptional error rates through intergenic transcripts. These findings generate testable predictions about TAD-condensate interactions while opening new directions for understanding nuclear organization.

The TAD Map fundamentally reshapes our understanding of gene regulation across cell types and conditions. By integrating species-wide chromatin data with foundation model representations of transcription, we uncover previously unrecognized principles governing genome organization and transcription. The framework reveals systematic changes in chromatin organization during aging and generates specific, testable predictions about age-related chromatin plasticity. We draw from these insights to develop TAD signatures, which bring new capabilities to single-cell RNA-seq analysis. Our integrated approach exposes previously invisible patterns in gene expression dynamics, from bulk tissue studies to single-cell resolution. When applied to cancer progression data, our framework reveals distinct patterns of chromatin organization in treated versus untreated cells, suggesting new mechanistic hypotheses about treatment resistance. Our findings demonstrate how bridging epigenetics and representation learning can illuminate fundamental synergies between genome architecture and regulation.

Results

Consensus TAD Maps for human and mouse. The genome's three-dimensional organization presents a substantial consistency across cell types [25–29, 42, 43]. To quantify this consistency, we analyzed TAD definitions from seven human and four mouse cell types using the TADKB database [44] and the Directionality Index technique [25]. The TADKB database provided harmonized TAD calls across diverse data sources (**Methods**). This analysis revealed high conservation. In humans, 92.6% of TAD boundaries matched those in at least one other cell type (within 50 kb), though only 31.7% matched across *all* cell types. Mouse showed lower conservation, with corresponding scores of 69.9% and 13.4%. **Fig. 2A** illustrates these patterns in example genomic segments from both species.

While TAD boundaries show variation, we reasoned that their fundamental organization might be more stable when viewed through the genes they contain. This insight led us to develop a "bag-of-genes" representation, where each TAD is defined simply by its protein-coding genes, whether fully or partially contained. This abstraction, which we call the "TAD Map", renders minor boundary variations irrelevant if they don't change the contained genes. Like the bag-of-words model in natural language processing, this approach sacrifices some precision—such as exact enhancer locations—but retains the key features needed for statistical analysis of gene regulation and TAD structure.

The bag-of-genes approach reveals striking consistency across cell types. We assessed this by examining adjacent gene pairs that share a TAD in at least one cell type (**Fig. 2B**). In humans, 72.8% of these pairs remain grouped together across *all* cell types, rising to 92.3% when considering majority agreement. Mouse shows similar patterns (78.9% and 91.6% respectively). This consistency significantly exceeds expectations under a null model of independent boundary dispersion (one-sided binomial test, $p = 7.3 \times 10^{-75}$ for human, 1.7×10^{-4} for mouse; **Fig. 2C, Methods**).

To convert this insight into a practical tool, we developed consensus TAD “scaffolds” for mouse and human genomes. Using maximum likelihood estimation, we identified optimal TAD boundaries by partitioning chromosomes into 50 kb segments and applying dynamic programming to find non-overlapping intervals best supported by TADKB data (**Methods**). The human scaffold comprises 3,036 TADs (mean length 879 kb, median 4 protein-coding genes per TAD), while the mouse scaffold contains 3,181 TADs (mean length 845 kb, median 4 protein-coding genes per TAD) (**Fig. 2D**). From these scaffolds, we derive the complete TAD Map by associating each TAD with all its genes, including those spanning multiple TADs (**Discussion**).

TAD Map agrees with CTCF ChIP-seq data. We reasoned that TAD boundaries, if correctly identified, should align with known mechanisms of chromatin organization. A key protein in this organization is CTCF, crucial for maintaining chromatin’s hierarchical structure. Previous studies have shown CTCF enrichment at TAD boundaries [45, 46], with substantial binding activity also inside TADs [3, 47]. Our analysis confirms and extends these findings. We sourced CTCF ChIP-seq data from the ENCODE database [48], comprising 281 human and 28 mouse studies (665 and 109 replicates, respectively). These studies include many cell types where TAD-CTCF concordance could not be previously studied due to the lack of Hi-C data. We find that in both human and mouse, CTCF binding occurs most frequently at TAD boundaries and remains significantly higher inside TADs than in the general genomic background (**Methods**, **Fig. 2E**). This pattern persists across diverse cell types and disease states (**Fig. 2F**), supporting the broad applicability of the TAD Map.

Expression quantitative trait loci (eQTL) are enriched within TADs. If TADs organize gene regulation, we would expect genetic variants within TADs to have stronger effects on their resident genes' expression. We can test this using expression quantitative trait loci (eQTLs)—genetic variants that influence nearby gene expression. To evaluate this hypothesis in the human TAD Map, we analyzed the entire GTEx v8 compendium [49], covering 49 tissues. For each gene, we estimated the probability $p(d)$ that a variant at distance d from its transcription start site (TSS) functions as an eQTL (**Methods**). As expected from previous studies [50, 51], this probability declines with distance. However, we find this decline is significantly slower inside TADs (**Fig. 2G**). Beyond the immediate vicinity of genes ($d > 30$ kb), variants within TADs show higher probability of being eQTLs (**Fig. 2H**), with median fold changes of 1.27 and 1.37 for upstream and downstream regions, respectively (p-values $< 5 \times 10^{-4}$, Wilcoxon rank sum test). These results, consistent across nearly all tissues (e.g., blood, brain, liver, skin, etc.), further validate our TAD Map and underscore TADs' role in organizing gene regulation.

Single-cell foundation models characterize the transcriptional context of genes. To understand how TADs influence gene regulation, we move beyond simple measures of gene co-expression. Genes do not function in isolation but rather within complex, co-regulated, and co-expressed networks involving interactions with nearby genes and distant ones, sometimes even across different chromosomes [52–54]. Traditional analyses of these relationships, particularly in TADs, have focused on co-expression or regulatory networks. However, these approaches struggle to capture an important aspect of TAD biology: how genes might serve as functional alternatives to each other while maintaining cellular states.

In olfactory neurons, only one of several hundred olfactory receptors is typically expressed, wherein the expressed receptor determines specific neuronal connections [22]. Similar cell-identity specificity is seen for protocadherin genes, with their proteins forming networks with neighboring proteins that enable cell self-recognition [55]. Despite their different expression patterns, genes in these families often reside in the same TADs and their members appear to serve similar roles relative to other genes. Traditional co-expression analysis misses these functional relationships.

Recent advances in single-cell representational learning offer a promising approach to this challenge. Single-cell foundation models, including Geneformer [56], scGPT [31], and scFoundation [57], integrate data from tens of millions of transcriptomes across various organs and cell types to create comprehensive gene representations. Trained in a self-supervised manner, they capture relationships between genes that go beyond traditional pathway annotations. Here, we use scGPT, which generates 512-length vector embeddings for each gene, trained on 33 million human single cells. In their original work, Cui et al. [31] interpreted these embeddings primarily as defining gene regulatory networks, where genes interact through transcription factors and signaling pathways. Others have suggested the embeddings simply reflect co-expression patterns [58]. We find both interpretations incomplete for understanding TAD function.

We propose these embeddings capture a different property: whether genes can serve similar roles in cellular processes relative to other genes. We term this property “Contextual Transcription Similarity” (CTS) and quantify it between two genes by calculating the Pearson correlation of their embedding vectors. Although both CTS and co-expression derive from gene expression data, they capture different aspects of gene function. Traditional co-expression measures require observing genes expressed together in the same cells. Beyond examples like olfactory receptors and protocadherins, where co-expression is intrinsically low, technical limitations of single-cell RNA-sequencing also pose a challenge. For example, sampling sparsity, dropout effects, and limited cell numbers can make it difficult to estimate co-expression reliably. In contrast, CTS leverages the foundation model’s integration of massive datasets to characterize a gene’s role through its relationships with all other genes, not just those co-expressed with it.

To validate this interpretation, we performed multiple complementary analyses. We first examined four organ-specific, atlas-scale scRNA-seq datasets from CELLxGENE [59]: the Human Breast Cell Atlas, Human Brain Cell Atlas, Human Cell Atlas of Fetal Gene Expression, and Human Heart Atlas [60–63]. Analyzing all protein-coding gene pairs less than 2 megabases apart on the same chromosome (418,425 pairs), we found that CTS shows significantly higher dispersion in its scores than co-expression (variances of 0.010 vs. 0.002 after Fisher's transformation, $p < 10^{-15}$, one-sided F test **Fig. 3A, 3B**). This higher information content suggests CTS can overcome the technical limitations in estimating co-expression from sparse single-cell data.

We hypothesized that CTS should capture gene *function* more comprehensively than simple co-expression, by virtue of capturing the gene's transcriptional context. At the same time, the two measures should broadly agree. The input data to train scGPT is single-cell RNA-seq and, if two genes have similar gene expression profiles, their embeddings should therefore be similar. We visualized co-expression and CTS for gene pairs on a scatterplot (**Fig. 3D**). As expected, CTS and co-expression show broad agreement across gene pairs (Pearson correlation 0.185, Spearman correlation 0.17, both $p < 10^{-15}$). To validate that CTS better captures functional relationships, we compared it with two independent approaches to representing gene function. Gene2vec [64] generates embeddings from 984 gene expression datasets on GEO, supplemented with pathway annotations (MSigDB [65]). In contrast, HiG2vec embeddings derive purely from Gene Ontology hierarchies embedded in hyperbolic space [66], capturing functional relationships independent of expression patterns. CTS shows stronger correlation than co-expression with both methods, but notably its advantage is greater with HiG2vec (correlation with CTS = 0.26 vs. co-expression = 0.06) than Gene2vec (CTS = 0.262 vs. co-expression = 0.220) (**Fig. 3C**). Its stronger agreement with the purely functional representation, HiG2vec, suggests CTS captures fundamental aspects of gene function beyond simple expression correlation.

The power of CTS in identifying functional relationships is particularly clear in certain gene families. In olfactory receptors and protocadherins, we find substantially higher CTS than co-expression (olfactory receptors: 0.099 vs -0.002; protocadherins: 0.196 vs 0.001 after median centering, **Fig. 3E-G, Methods**). Beyond the difference in their magnitudes in these families, CTS and raw co-expression are also less in agreement about relative ordering of gene–gene similarity, with low Spearman rank correlation between the two measures on within-family gene pairs: -0.176 for olfactory receptors and 0.022 for protocadherins. We expect such family-specific comparisons of CTS and co-expression may reveal deeper insights into the “substitutability” of family members with each other. For instance, histone genes, which are coordinately regulated during DNA replication [67], also show substantially higher CTS values (0.251 vs 0.021) even as the within-family correlation between CTS and co-expression, at 0.114, is higher than in protocadherins and olfactory receptors. In all these families, the CTS–co-expression differences are highly significant ($p < 10^{-15}$) (**Fig. 3H**). In contrast, keratin genes show strong concordance between CTS and co-expression (**Fig. S1**).

These findings raise intriguing questions about TAD function. The gene families we analyzed above localize to a limited set of TADs (**Methods**), and high CTS scores we observe among them suggest these TADs might create environments where genes maintain similar roles relative to other genes even when not co-expressed. To investigate this possibility systematically, we next examined whether genes sharing a TAD show higher contextual similarity than those that do not, and how this pattern relates to genomic proximity. This analysis sheds light into how TADs contribute to the organization of gene regulation.

TADs serve as loci of enriched transcriptional context. To understand how TADs influence gene regulation—through simple physical proximity or through broader mechanisms—we systematically analyzed transcriptional relationships between gene pairs. We identified all protein-coding gene pairs within 2 megabases on the same chromosome and classified them as either sharing a TAD (TAD pairs) or not (non-TAD pairs). Comparing both CTS and co-expression between these groups (**Fig. 4A**), we found that both measures were significantly higher for TAD pairs than non-TAD pairs ($p < 10^{-15}$ for each, one-sided t-test). However, the effect was markedly stronger for CTS (Cohen's d = 0.28 vs 0.12). This suggests that CTS better captures TAD-mediated relationships than traditional co-expression metrics.

Since genes within TADs tend to be closer together, and adjacent genes often share functional relationships [68, 69], we next addressed the potential confounding effect of genomic proximity. We stratified our analysis by distance, grouping gene pairs into intervals: less than 20 kb, 20–50 kb, 50–200 kb, 200–500 kb, and 500–1000 kb. TAD pairs showed consistently higher CTS across all intervals (p values of 6.04×10^{-8}, 4.60×10^{-10}, $< 10^{-15}$, $< 10^{-15}$, and $< 10^{-15}$ respectively) (**Fig. 4B**), with the relative magnitude of this difference preserved even as CTS decreased with distance. This finding contrasts with Long et al. [70], who reported no TAD-specific effects beyond proximity using co-expression analysis. Indeed, when we repeated our distance-stratified analysis using co-expression (**Fig. S2**), the TAD-specific signal was substantially weaker, suggesting their conclusion reflected the limitations of co-expression measures.

The remarkably consistent difference between TAD and non-TAD pairs across distances pointed to an underlying mechanistic principle (**Fig. 4D,E**). We posit three possible mechanisms: (1) TADs might create isolated compartments that override genomic proximity effects, but this would predict distance-independent CTS within TADs; (2) TADs could act through chromatin compaction alone, uniformly reducing effective distances between genes; or (3) TADs might multiplicatively enhance gene-gene regulatory relationships while preserving distance dependencies. While we cannot rule out the 2nd mechanism entirely, the observed pattern strongly supports the third mechanism—non-TAD pairs consistently showed 83% of the CTS observed in TAD pairs across all distance intervals, revealing that TAD membership enhances CTS by a multiplicative factor of 20.5% (**Fig. 4C**). We determined this scaling factor by minimizing the area between CTS curves for TAD and appropriately scaled non-TAD pairs (**Methods**).

The consistent multiplicative scaling suggests an underlying physical process that enhances transcriptional context within TADs. We identify transcriptional condensates (TCs), formed by liquid-liquid phase separation (LLPS), as a compelling candidate mechanism. These dynamic, membrane-less compartments concentrate transcriptional machinery and regulatory factors [39, 71], creating environments where genes can more effectively share transcriptional context. LLPS is particularly well-suited to explain our observations because phase-separated compartments naturally create multiplicative effects on local concentrations [72, 73]—precisely the pattern we observed in our distance analysis. We note that individual TCs might be too small to span entire TADs and posit that either larger-scale “TAD condensates” or the collective action of multiple smaller TCs contribute to the observed effects. The spatial proximity of chromatin within TADs may promote local enrichment of TCs. We hypothesize that TADs work synergistically with TCs, systematically enhancing their formation or stability (**Fig. 4E**). To test this hypothesis and better understand TAD–condensate interactions, we performed three complementary analyses.

Systematic regulation in TAD regions differs mechanistically from that of non-TAD regions. Phase-separated condensates create distinct biochemical environments in the nucleus. We hypothesized that the presence and activity of these condensates might be systematically higher in TADs, creating privileged spaces for transcriptional regulation. To systematically evaluate this hypothesis, we analyzed Replogle et al.’s genome-wide Perturb-seq dataset [74] where 9,866 genes were knocked-down using CRISPRi and single-cell RNA-seq measurements were made.

For each perturbation target g, we tested if the magnitude of transcriptional response in TAD and non-TAD regions shows any systematic differences. We applied DESeq2 [75, 76] to the set of single-cell transcriptomes where g was knocked down, comparing them to the baseline non-targeted transcriptomes. To mitigate batch effects, we adjusted for the GEM (gel bead-in-emulsion) group covariate. After estimating β (= log2 fold-change, log2FC) for each protein-coding gene in the transcriptome, we computed the variance of these log2FCs within TAD and non-TAD regions separately. These variances serve as *transcriptional disruption metrics*, quantifying how strongly perturbation of g affects genes in each region (**Fig. 4F, Methods**).

As positive controls, we first examined the perturbations causing the most widespread transcriptional disruption, regardless of gene residence. As expected, these perturbations, summarized in (**Fig. S3A**), are enriched for transcriptional processes such as RNA Polymerase II elongation and initiation (GO:0034243, Adjusted $p = 2.39 \times 10^{-6}$; GO:0060261, Adjusted $p = 6.09 \times 10^{-6}$) [77]. Of particular note were perturbations of Core Mediator Complex genes (GO:0070847), including MED12, MED30, MED21, and MED9, which are central to transcriptional regulation [78, 79]. These genes appeared in the 99th percentile of all perturbations when ranked by their effect on overall log2FC variation, confirming their substantial impact on transcription in perturbed cells.

Having validated our framework on general transcriptional regulators, we next examined its sensitivity to perturbations of known TAD-organizing factors. We looked for excess impact in TAD versus non-TAD regions by comparing their disruption metrics. For the cohesin complex, which mediates loop extrusion [80–82], perturbations of subunits SMC1 and RAD21 showed strong differences between TAD and non-TAD disruptions of 0.0327 (99th percentile) and 0.0132 (97th percentile), respectively (**Methods**). SMC3 perturbation, with a difference of -0.0085, exhibited stronger disruption in non-TAD regions than TAD regions, although its disruption metrics were in the 99th percentile for both regions, indicating strong genome-wide effects. Perturbation of

CTCF, which establishes TAD boundaries and acts as an insulator [45, 83], yielded an excess TAD impact in the 75th percentile. These results demonstrate our framework's ability to detect both TAD-specific and global effects of chromatin organizers. We next sought to systematically identify additional factors that may regulate transcription differently in TAD regions.

Among other perturbations preferentially affecting TAD regions, we found enrichment for factors (**Fig. 4G**) involved in histone acetylation (GO:0016573, Adjusted $p = 4.074 \times 10^{-4}$), such as TAF12, ACTL6A, EP400, and DMAP1 (full list in (**Fig. S3B**). Samata et al. discovered that the loss of acetylation on histone H4 leads to an increase in long-range contacts beyond individual TADs [84]. Additionally, Palacio et al. demonstrated that TAFs (TATA-binding protein Associated Factors), which contribute to the transcription pre-initiation complex, play a major role in the transcriptional phase condensate model that facilitates localized gene transcription [85]. A second major category emerged from our analysis. The Integrator complex genes (INTS2, INTS5, INTS7, and INTS8) [86] showed strong enrichment in TAD-specific perturbations, with roles in snRNA Processing (GO:0016180, Adjusted $p = 1.713 \times 10^{-5}$) and ncRNA Processing (GO:0034470, Adjusted $p = 1.614 \times 10^{-3}$).

Complementing these TAD-specific effects, we identified a distinct set of perturbations leading to disproportionate disruption in non-TAD regions (**Fig. S3C**). These were enriched for ribosomal processes, including Ribosome Biogenesis (GO:0042254, Adjusted $p = 1.254 \times 10^{-15}$) and rRNA Processing (GO:0006364, Adjusted $p = 3.536 \times 10^{-6}$), with multiple targets in the Ribosomal Protein Small (RPS) and Large (RPL) subunits (**Fig. S3D**). We note that the set of non-TAD genes is both smaller than TAD genes and heavily enriched for ribosomal genes; perturbation of ribosomal factors likely affects other ribosomal genes disproportionately, and affecting our enrichment estimates. Resolving this may require additional perturbation studies that cover the remaining human genes. Given our current findings, however, an intriguing possibility is the interplay between nucleoli and TADs. Nucleoli, which are transcriptionally active membrane-less condensates serving as sites for rRNA processes [87], are typically surrounded by heterochromatin and regions of low gene density [88]. A direction of future research could be to investigate the extent of overlap between TADs and nucleoli-associated chromatin. For instance, nucleoli and TADs might play complementary roles in genome organization— with nucleoli preferentially organizing non-TAD genes, particularly those involved in ribosomal biogenesis, while TADs structure other transcriptional programs.

Gene orientation analysis supports our hypothesis: The role of gene order and adjacency in shaping regulatory evolution has been well-documented [89, 90]. If TADs indeed enhance transcriptional regulation through condensate-like mechanisms, we reasoned this should be reflected in how the spatial orientation of adjacent genes relates to their co-transcription. We examined two predictions. First, we hypothesized that if LLPS promotes co-regulation, the transcriptional machinery should more readily be able to access adjacent genes with aligned orientations. Within TADs, same-orientation pairs exhibited significantly higher contextual similarity than oppositely oriented ones (0.137 vs. 0.121, $p = 4.5 \times 10^{-14}$). Outside TADs, orientation had no significant effect (0.109 vs. 0.106, $p = 0.508$) (**Fig. 4H, Methods**). RNA-seq analysis yielded similar results, with higher similarity for same-orientation pairs within TADs (Pearson correlation of 0.19 vs. 0.15, $p < 10^{-15}$, one-sided t-test) but no significant difference outside.

In our second analysis, we examined the frequency of transcriptional read-through between adjacent genes. We analyzed intergenic transcripts curated by Agostini et al. [91] from 38 publicly available datasets, covering over 2.5 billion uniquely mapped reads, and used their annotations to obtain 11,417 intergenic transcripts predicted to have been erroneously transcribed (**Methods**). Within TADs, pairs of similarly-oriented genes showed significantly more intergenic transcripts than pairs with opposing orientations (frequency per pair of 11.9% vs 7.6%, $p < 10^{-15}$, one-sided binomial test) (**Fig. 4I**). In contrast, gene orientation did not significantly affect intergenic transcript frequency between non-TAD gene pairs, though these pairs showed higher overall frequencies, suggesting elevated transcriptional noise in non-TAD regions. Taken together, these analyses provide computational evidence for condensate-like regulation within TADs and highlight specific predictions that can be tested experimentally.

The TAD Map and CTS framework enables new aging and cancer-related investigations. The TAD Map framework enables systematic investigation of chromatin's role in cellular state transitions without requiring additional Hi-C experiments for each cell type. By integrating CTS and co-expression data with our species-wide TAD Map, we can generate and statistically evaluate hypotheses about chromatin organization's role in development, aging, and disease [92, 93]. Prioritized hypotheses can then be experimentally tested.

Cell differentiation: A central question in developmental biology is how cells are able to precisely modulate their gene expression programs, balancing transcriptional stability with plasticity. Recent work by Pollex et al. [94] revealed that this balance involves systematic shifts in enhancer–promoter (E–P) interactions: early embryonic stages use permissive, proximal E–P contacts that enable rapid developmental changes, while later stages establish more constrained, often distal topologies that reinforce tissue-specific expression patterns. Since TADs create the genomic neighborhoods where such E–P interactions occur [12], we hypothesized that these developmental shifts in E–P topology would manifest as systematic changes in TAD-level transcriptional organization.

We started by analyzing diverse differentiation trajectories, revealing a clear relationship between developmental stage and TAD-mediated gene organization. Using developmental scRNA-seq transcriptomes curated by Gulati et al. [95], we found that cells in earlier differentiation stages showed significantly stronger expression clustering within TADs (21 of 24 trajectories, **Fig. 5A**,

p = 0.00019, one-sided t-test). To accommodate the small sizes of some datasets, we designed a bootstrap-based "TAD usage" score to quantify this effect (**Fig. S4A-B, Methods**). The clustering pattern held across different methods of demarcating early versus late differentiation and multiple tissue types (**Fig. S4C**), suggesting that TAD-mediated organization plays a particularly important role in maintaining developmental plasticity.

Aging: This developmentally-linked pattern prompted us to examine whether TAD-level organization changes systematically over longer timescales. The immune system presents an ideal model system for this investigation, as age-related decline affects cell distribution and function across multiple lineages, particularly among myeloid cells, innate lymphocytes, and T cells [96]. While aging-related epigenetic changes, including histone modifications, DNA methylation, and chromatin remodeling, have been extensively documented [97, 98], we wondered if these changes reflect solely on the shifts in cell type proportions during aging, or do they also indicate intrinsic within-cell-type changes in chromatin regulation.

To generate testable predictions about age-related changes in chromatin organization, we analyzed scRNA-seq datasets from human immune cells spanning prenatal to late adult stages [99, 100]. We divided the data into four age groups: DS1 (28–46 days post-fertilization), DS2 (16–17 weeks post-fertilization), DSmature (ages 22–61 years), and DSold (ages > 61 years) (**Fig. 5B**). After downsampling each dataset to 50,000 cells using Geosketch [101], we examined whether CTS between gene pairs shifts systematically with age.

As cells age, we observed a decline in the proximal transcriptional programs and machinery coordinating gene pairs, as measured by a CTS area-under-curve metric. Pre-trained foundation models can be fine-tuned on new datasets, allowing them to capture the "tilt" of these data. For each curated dataset (DS1, DS2, DSmature, DSold), we fine-tuned the whole-human scGPT model using the masked language objective with default hyperparameters. We extracted the updated gene embeddings, calculated CTS for gene pairs on the same chromosome, and stratified them according to the previously-defined distance intervals. For TAD or non-TAD gene pairs, the average CTS was calculated at each interval (**Fig. 5C-D**). Each fine-tuned dataset produced its own CTS curve, and we quantified overall CTS by calculating the area under the curve (AUC) for each dataset (**Methods**).

Quantifying these changes revealed a systematic decline in transcriptional co-regulation with age. CTS analysis showed highest values in DS1, followed by DS2, DSmature, and DSold—a pattern consistent for both TAD and non-TAD gene pairs (**Fig. 5E**). The decline was substantial. In TAD pairs, DS1's CTS (measured by AUC) exceeded DSold by 11.75%, DSmature by 7.66%, and DS2 by 0.90%. Non-TAD pairs showed similar declines (DS1 exceeding DSold by 12.71%, DSmature by 8.80%, and DS2 by 0.51%). Notably, TAD gene pairs maintained higher CTS than non-TAD pairs across all ages, suggesting that the TAD-mediated organization of transcription persists even as it weakens with age. These patterns suggest that aging may systematically modulate the spatial constraints on transcriptional machinery. We wondered if these overall population-level changes were only driven by shifts in cell type proportions (e.g., reduced progenitor prevalence) or if they also reflected changes within individual cell types.

To distinguish between population-level and cell-intrinsic changes, we leveraged scGPT's ability to analyze specific cell types. By fine-tuning the model separately on subsets of RNA-seq data, we examined CTS changes in T cells (general and naive thymus-derived CD4-positive, alpha-beta T cell), macrophages, monocytes, and erythrocytes across early and mature stages (**Methods**). We found that CTS decreased with age within each cell type (**Fig. 5F**), with general T cells and macrophages showing the most pronounced decline (**Fig. 5G**). This suggests that overall population-level decline in CTS is due to a combination of factors: shifts in cell type proportions (fewer progenitor cells in mature populations), and intrinsic within-cell-type changes. While the former adds to with existing observations of plasticity changes during cell differentiation, the latter newly suggests that such changes occur even within terminally-differentiated cell types. Thus, our cell-type specific analysis generates testable predictions about aging's impact on chromatin organization within well-differentiated cell types. Moreover, our observation that general T cells and macrophages display the greatest aging-related changes suggests that the degree of such plasticity changes may vary across cell types and nominates specific cell populations to prioritize for targeted chromatin studies.

Cancer: Our earlier findings that TAD usage marks developmental plasticity suggest examining cell states where plasticity is dysregulated. Cancer presents such a case, where cells often dedifferentiate or fail to differentiate properly, reverting to stem-like states that enable unrestricted proliferation and evasion of terminal differentiation signals. If TAD-mediated transcriptional clustering indeed reflects developmental plasticity, cancer cells should display heightened TAD usage characteristic of progenitor states. Supporting this prediction, analysis of bulk RNA-seq data from The Cancer Genome Atlas (TCGA) across blood, brain, lung, and renal cancers revealed that tumor cells exhibit greater clustering of transcriptionally active genes within TADs compared to normal tissues [102] (**Fig. 5H**). This difference in TAD usage patterns was robust to adjustments for the higher number of expressed genes in tumor cells (**Methods**).

We wondered how cancer treatment might modulate cell plasticity as cancer progresses. To examine this, we leveraged CTS analysis of a recent colorectal cancer dataset tracking plasticity during metastasis [103]. After fine-tuning scGPT on this data, we found that in untreated samples, cancer cells (both primary tumor and metatstatic) show higher CTS than normal cells. These findings are consistent with the TCGA analysis. However, following 5-fluorouracil-based chemotherapy, a notable divergence emerged (**Fig. 5I**). Treated tumor cells (both primary and metastatic) displayed *elevated* CTS exceeding untreated tumor cells, while treated normal cells showed markedly *reduced* CTS. This divergent response suggests fundamental differences in how normal and cancer cells adapt their transcriptional programs to treatment stress.

These CTS patterns align with well-established biological responses to chemotherapy, which indiscriminately targets rapidly dividing cells. In normal tissues, the slightly lower CTS after treatment is consistent with a reduction in the proportion of progenitor cells. On the other hand, in tumors, chemotherapy can select for cancer stem cells (CSCs) that possess inherent resistance mechanisms, including enhanced DNA repair capabilities, drug efflux pumps, and anti-apoptotic pathways [104, 105]. The elevated CTS in treated cancer populations likely reflects this enrichment of stem-like cells, which maintain plastic transcriptional programs even under treatment pressure. While these correlations between CTS patterns and cellular plasticity are supported by known biology, establishing causal relationships between TAD-mediated transcriptional organization and treatment response will require further investigation. Our findings may also address a major gap in clinical diagnostics, by leading to a *new quantitative estimator of post-treatment CSC enrichment*.

TAD signatures highlight heterogeneity of scRNA-seq data. While the unprecedented detail and volume of scRNA-seq data has had a transformative impact on biological research, challenges of data sparsity, uneven gene coverage (“dropouts”), and platform effects in such data remain a challenge [106]. To complement existing dimensionality reduction techniques, we introduce *TAD signatures*: a precise method leveraging the TAD Map’s gene groupings to mitigate technical noise while preserving the biological signal in scRNA-seq data. Specifically, we define a cell’s TAD signature as the set of its TAD activation probabilities. A feature vector of length n (= number of TADs in the scaffold that contain at least one gene), with the i-th feature quantifying the expression activity of TAD i on a scale from 0 (inactive) to 1 (fully active) (**Methods**). Unlike traditional approaches, TAD signatures provide an epigenetically-informed dimensionality reduction that can be inferred for any human or mouse scRNA-seq dataset without requiring Hi-C data. By integrating prior knowledge of chromatin's overall structure (**Fig. 1B**), this approach offers a supplementary lens for accentuating biological signal in complex transcriptomic data.

Understanding cellular heterogeneity within cancer tissues remains a critical challenge in designing effective treatments [107, 108]. With existing methods limited in their accuracy of linking transcriptional variability with phenotypic outcomes [109], we hypothesized that TAD signatures could provide a more nuanced view of cellular diversity. Across 193 cancer cell lines, we discovered that a line's drug response profile (4,686 drugs from the PRISM drug screen [110]) correlates with TAD signature-based measures of cellular heterogeneity [111].

TAD signatures effectively expose the transcriptional variability that drives drug response. While the first moment (mean) of gene expression captures the overall cell-type profile, it fails to reflect the within-cell-type heterogeneity that exists within seemingly uniform populations. Higher moments are needed for that. With transcript counts typically following a Poisson or negative binomial distribution, the first and second moments contain similar information. Thus, skew, represented by the third moment, becomes crucial. By applying principal component regression to TAD signature-derived population statistics, we sought to capture the nuanced variability within cell types. Analyzing the top 50 principal components—which capture 43.7% of drug response variability—we found that higher-order moments of TAD signatures were substantially more predictive than conventional approaches. Notably, TAD signature-based skew measurements demonstrated significantly higher R-squared values ($p = 0.0014$, one-sided Wilcoxon rank-sum test), suggesting these signatures can highlight phenotypically meaningful cellular heterogeneity that gene-level analyses cannot reach (**Fig. S5A, Methods**) and thus be be a powerful supplement to the latter.

Cell type inference remains a critical yet challenging task in single-cell genomics, often requiring extensive manual curation that limits scalability [106, 112]. To complement existing automated clustering methods, we introduced TAD signatures as a supplementary approach to potentially improve inference accuracy. Applying the Leiden clustering algorithm to three large scRNA-seq datasets from the CELLxGENE portal [59]—covering breast [113], lung [114], and T cells [115]—we systematically compared clustering methods (**Methods**). Notably, TAD signatures not only matched but outperformed traditional approaches, with combined RNA-seq and TAD signature representations robustly improving clustering accuracy across all tested datasets (**Fig. S5B**).

Drilling deeper, we discovered TAD signatures' particular strength in distinguishing cell subtypes with subtle transcriptional differences. In breast tissue data, our approach showed a notable ability to differentiate between progenitor and mature luminal cells—a distinction that had previously challenged automated techniques [113]. This capability stems from TAD signatures' sensitivity to transcriptional variations across cell differentiation stages, suggesting a powerful new approach to capturing cellular heterogeneity that goes beyond traditional marker-based methods (**Fig. S5C**).

Discussion

Through two fundamental advances, this work reframes our understanding of how chromatin structure controls gene expression. First, we move beyond viewing TADs as either essential or irrelevant, instead characterizing them as creating statistical biases in transcriptional relationships. Second, we introduce contextual transcriptional similarity (CTS), leveraging single-cell foundation models to capture gene relationships that traditional co-expression analyses miss. The TAD Map—our low-resolution but robust representation of chromatin structure—enables systematic analysis across diverse cell types without requiring additional Hi-C experiments. This approach reveals that TADs enhance contextual similarity between genes by approximately 20%, with the enhancement persisting multiplicatively across genomic distances. The consistency of this pattern suggests an underlying physical mechanism, and we hypothesize TADs and transcriptional condensates synergize to regulate transcription. By bridging foundation models and chromatin biology, our framework reveals fundamental principles governing transcriptional regulation across the genome.

Contextual transcriptional similarity (CTS) transforms our ability to detect gene relationships that evade traditional analyses. For instance, while genes in certain families (e.g., protocadherins and olfactory receptors) show minimal co-expression, they exhibit high contextual similarity, reflecting their shared roles in cellular identity. The ability of CTS to detect additional modes of transcriptional co-regulation helps resolve apparent contradictions in the literature. While Long et al. [70] were limited to simple co-expression analysis in two cell types and could not disentangle TAD-specific effects from genomic proximity effects, CTS reveals systematic TAD-mediated enhancement patterns that hold across distance scales and cell types.

Single-cell foundation models (scFMs) offer a conceptual breakthrough for detecting gene relationships, as demonstrated by scGPT's role in CTS. While scFMs have faced some skepticism over their advantages on traditional single-cell tasks [58, 116], our results suggest they excel at capturing broad-but-subtle biological patterns—particularly when combined with prior knowledge of genomic organization. As with protein language models, the biggest advantage of scFMs over traditional approaches may be in offering general-purpose representations that power downstream analyses. This advantage is particularly evident in "few-shot" settings like our analysis of age-specific patterns in cell types. While each sub-dataset was too small for robust correlation estimation, scGPT could still be fine-tuned on them individually, revealing clear distinctions in the learned representations. However, leveraging scFM representations requires care and further research, as they may not directly map to traditional biological concepts. Specifically, in the original study, Cui et al. interpreted gene representation similarity as implying a traditional gene regulatory network, but we find that CTS instead provides a broader statistical view of a gene's role relative to all other genes.

Our findings suggest that TADs create privileged zones of gene regulation, reshaping our understanding of genome organization. Rather than serving as strict compartments, TADs may work by enhancing co-transcriptional potential. This view may address apparent contradictions in the field where some studies report strong TAD effects on gene regulation and others find them dispensable. Our analysis suggests TADs systematically bias—but do not absolutely constrain—co-transcriptional relationships. Notably, this enhancement persists even as the baseline level of contextual similarity declines with distance, suggesting TADs don't simply redefine effective distances between genes but instead create environments that multiplicatively enhance existing regulatory relationships. This statistical framework resolves long-standing paradoxes about TAD function, explaining both cross-boundary interactions and the subtle effects of boundary disruption.

A physical model emerges from our analysis: TADs and transcriptional condensates work together to amplify gene regulation. Both contextual similarity and eQTL effects show a baseline decline with genomic distance. The consistent enhancement of CTS in TADs suggests an underlying physical process operating with higher efficiency within TAD regions, though future work will be needed to determine if eQTL effects follow similar scaling. While multiple mechanisms might explain this pattern, transcriptional condensates represent one compelling possibility. Supporting this hypothesis, our analysis reveals that both CTS and erroneous transcriptional read-through are specifically enhanced for same-orientation gene pairs within TADs, consistent with shared access to phase-separated transcriptional machinery. The observations can be explained by a unified model that suggests that rather than functioning as independent organizational features, TADs and condensates work together to create specialized transcriptional environments. This unified model bridges previously disparate observations about nuclear organization while generating specific, testable predictions about how TADs and condensates jointly control transcription.

The TAD Map framework uncovers fundamental principles of cellular plasticity across development and aging. We find cells in early differentiation stages show stronger TAD-mediated organization, consistent with recent observations about enhancer-promoter interactions in development. Our aging analysis extends this pattern. TAD-mediated organization systematically declines with age, and notably, this decline persists within specific cell types like macrophages. This suggests age-related chromatin reorganization involves not just shifts in cell populations but also intrinsic changes within individual cell types.

Cancer cells provide further validation of these patterns while revealing new complexities. While tumor samples show elevated TAD usage compared to normal tissue—consistent with our aging-related observations about chromatin plasticity—their response to treatment is distinct. Our analysis of colorectal cancer progression shows that chemotherapy induces divergent responses: treated tumor cells display increased TAD-level coordination while treated normal cells show reduced levels. This bifurcation suggests fundamental differences in how normal and cancer cells adapt their transcriptional programs under stress, potentially reflecting selection for cancer stem cells during treatment and suggesting diagnostic directions. While these correlations between TAD usage and cellular plasticity are supported by known biology, establishing causal relationships will require directed experimental investigation. Our work follows in the line of previous integrative analyses linking chromatin organization with transcriptional signatures in cancer [117, 118].

Our approach has some limitations. While the TAD Map enables species-wide analysis without requiring cell type-specific Hi-C data, this convenience comes at the cost of resolution—we cannot capture fine-grained structural variations between cell types or detect subtle boundary shifts. Fortunately, several excellent methods have been developed that apply machine learning and statistical methods to infer cell-specific or sequence-specific 3D genome structure, at both bulk and single cell levels [98, 119–125]. The growing availability of advanced sequencing technologies, including multimodal methods, further enhances our ability to infer how TADs influence gene regulation [126–131]. Future work could integrate the TAD Map with diverse datasets and computational approaches to investigate the mechanisms underlying cell type-specific deviations from the TAD Map. Other epigenetic modalities, such as histone modifications, have also been reported to dynamically interact with chromatin topology [132, 133]. Furthermore, the current framework focuses primarily on protein-coding genes, leaving open questions about the regulation of non protein-coding genes. In constructing our TAD scaffold, we associate genes that cross TAD boundaries with both TADs. We believe this is a conservative approach that balances statistical robustness with biological accuracy, though the specific regulatory dynamics of boundary-spanning genes merit further study. Future work could also investigate TAD sub-structure and heterogeneity—sub-TADs, large versus small domains, gene-dense versus gene-sparse regions, and structural variations may influence regulatory patterns differently. The heterogeneity of TADs likely also affects the number and size of condensates that form and function within and around them. Additionally, while our findings suggest general principles about TAD-mediated regulation, exceptions likely exist for specific genomic contexts or cell states. Targeted functional studies will be crucial for understanding these nuances.

Our work represents a paradigm shift in studying the link between chromatin structure and transcriptional regulation, moving beyond binary inferences drawn from isolated cell types to a probabilistic framework inferred from millions of cells. This shift enables systematic investigation of cellular transitions without requiring DNA structural data for each state. By bridging genome organization and functional genomics, our approach opens new avenues for studying development, aging, and disease. The TAD Map framework, combined with foundation model representations, demonstrates how integrating prior knowledge of genome organization with modern representation learning can uncover fundamental principles of cellular regulation. Lastly, our work advocates for reimagining transcriptional regulation as a convolution of probabilistic mechanisms—chromatin

3D structure, chromatin accessibility, transcriptional condensates etc.—that form the background against which discrete actors like transcription factors perform.

Methods

Cell type-specific TAD definitions. We sourced cell type-specific TAD architectures from Liu et al.'s TADKB database [44], selecting TAD definitions inferred with the Directionality Index (DI) technique at 50 kb resolution. We chose TADKB because it ensured a consistent "Hi-C to TAD" mapping across experimental data from multiple studies. We retrieved data for all the human and mouse cell lines available in the database: seven for human (GM12878, HMEC, NHEK, IMR90, KBM7, K562, and HUVEC) and four for mouse (CH12-LX, ES, NPC, and CN). These TAD definitions were computed by Liu et al. using source Hi-C data from Rao et al. [2] (Gene Expression Omnibus, GEO, accession GSE63525) and Bonev et al. [134] (GEO accession GSE96107).

Reference genome versions. All genomic coordinates and gene names correspond to Ensembl v102, with the human and mouse reference genomes being hg38 and mm10, respectively. With the *liftOver* program [135], TAD definitions for human cell types were mapped to the hg38 reference genome from hg19, the reference genome used in TADKB. TAD definitions for mouse cell types already corresponded to the mm10 reference genome.

Maximum likelihood estimation of the consensus TAD scaffold. We infer TADs independently for each chromosome: our algorithm takes as input a list of cell type-specific TAD architectures for the chromosome and outputs the consensus TAD definition. Both the input and output definitions specify TADs as a set of non-overlapping genomic intervals along the chromosomes. Our algorithm currently operates at a 50 kb resolution with both the input and output defined at that granularity.

We divide the entire chromosome into segments of size R (currently, R = 50 kb) and for every possible pairwise combination of segments, compute the likelihood that the two segments share a TAD. More formally, given segments at loci i and j, we define X_{ij} as the number of input cell type-specific TAD architectures in which i and j share a TAD. Our goal is to infer C_{ij}, where $C_{ij} \in \{0, 1\}$ indicates if i and j share a consensus TAD. Additionally, C_{ij} need to obey integrity constraints that correspond to a valid TAD architecture; we describe these later. To accommodate the differing amounts of input data available for human and mouse in a unified framework, we discretized X_{ij} into three levels: $X_{ij} \geq T$, $T > X_{ij} > 0$, and $X_{ij} = 0$. With data for 7 human cell types and 4 mouse cell types, we chose T as 4 (for human) and 2 (for mouse), so that the three levels of X_{ij} express the intuition that it receives support from the majority, at least one, or none of the input cell types, respectively. We then parameterize our likelihood model $P(X_{ij}|C_{ij})$ as follows:

$P(X_{ij}\|C_{ij})$	$C_{ij} = 0$	$C_{ij} = 1$
$X_{ij} \geq T$	0	$1 - \phi - \epsilon$
$T > X_{ij} > 0$	θ	ϕ
$X_{ij} = 0$	$1 - \theta$	ϵ

where $0 < \theta, \phi, \epsilon, (1 - \phi - \epsilon) < 1$. Our parameterization implies that a segment pairs i, j which share a TAD in a majority of the input cell types ($X_{ij} \geq T$) must be present in a consensus TAD. On the other hand, we provide for the possibility ($\epsilon > 0$) that a segment pair i, j with no support from any of the input architectures might still share a consensus TAD— this allows us to stitch together overlapping TAD ranges across cell type-specific inputs if strong overall support exists for a broad TAD at that locus.

Under our model, the likelihood of the observations $\mathbf{X} = \{X_{ij}\}$ given the consensus TAD scaffold $\mathbf{C} = \{C_{ij}\}$ is

$$\begin{aligned} P(\mathbf{X}|\mathbf{C}) &\propto \prod_{i,j} P(X_{ij}|C_{ij}) \\ &= \prod_{X_{ij}\geq T} 0^{(1-C_{ij})}(1-\phi-\epsilon)^{C_{ij}} \prod_{T>X_{ij}>0} \theta^{(1-C_{ij})}\phi^{C_{ij}} \prod_{X_{ij}=0} (1-\theta)^{(1-C_{ij})}\epsilon^{C_{ij}} \end{aligned}$$

The first sub-product imposes the binding constraint that $C_{ij} = 1$ for all segment pairs $\{i, j \mid X_{ij} \geq T\}$. Given that, we can focus on the remaining two sub-products to maximize the log-likelihood:

$$\begin{aligned} l(\theta, \phi, \epsilon) &= \sum_{T>X_{ij}>0} \big(\log\theta + C_{ij}\log(\phi/\theta)\big) + \sum_{X_{ij}=0} \big(\log(1-\theta) + C_{ij}\log(\epsilon/(1-\theta))\big) \\ &= h(\theta, \mathbf{X}) + \sum_{T>X_{ij}>0} w_1 C_{ij} + \sum_{X_{ij}=0} w_2 C_{ij} \end{aligned}$$

where $h(\theta, \mathbf{X})$ is not a function of C_{ij}, and the terms w_1 and w_2 represent more convenient combinations of the parameters θ, ϕ, and ϵ. Maximizing the log-likelihood thus requires solving

$$\arg\max_{\mathbf{C}} \sum_{T>X_{ij}>0} w_1 C_{ij} + \sum_{X_{ij}=0} w_2 C_{ij}$$

We note that integrity constraints on C_{ij} link the terms: $\mathbf{C}$ needs to be transitive, i.e., if segments pairs (i, j) and (j, k) share a TAD then so must (i, k). Also, a TAD must be contiguous: if $C_{ij} = 1$ then $C_{iv} = C_{vj} = 1$ for all segments v between i and j. Intuitively, this formulation describes a trade-off between biasing towards long TADs, which cover more true positive C_{ij}s (guided by $T > X_{ij} > 0$ cases) but also have more false positives (driven by $X_{ij} = 0$ cases), and short TADs where the false positives will be fewer but the risk of false negatives increases.

To maximize l and compute $\mathbf{C}$, we formulated a dynamic programming algorithm that splits the chromosome into recursively smaller ranges and finds the globally optimal combination of C_{ij} assignments. We chose $w_1 = 0.5, w_2 = -1$ for human and $w_1 = 0.05, w_2 = -3$ for mouse; these choices produced consensus TAD scaffolds where the number of TADs and the distribution of their lengths was in line with the corresponding statistics for cell type-specific TAD architectures.

TAD boundaries As an alternative quantification, we compare these estimates to a null hypothesis where TAD boundaries are independently dispersed between gene boundaries in each cell type while preserving the number of TADs. As (**Fig. 2C**) shows, for a gene pair that co-occupies a TAD in one of the architectures, the expected number of other cell-type architectures that agree with this grouping under the null hypothesis is just 0.85 (human) and 0.29 (mouse), implying that the extent of agreement actually observed is highly significant (one-sided binomial test, $p = 7.3 \times 10^{-75}$ for human, 1.7×10^{-4} for mouse).

Agreement with CTCF ChIP-seq. We sourced CTCF ChIP-seq data from the ENCODE database [48], consisting of 281 human and 28 mouse studies (comprising 665 and 109 replicates, respectively). These studies covered many cell types where concordance of TADs and CTCF binding sites was previously uncharacterized because Hi-C data for the cell type was unavailable. We filtered to only keep peaks with Irreproducible Discovery Rate (IDR) less than 0.05 and mapped these peaks to our estimated TAD scaffold. We then grouped the peaks by their location relative to TADs, partitioning the entire genome into the following disjoint categories: *TAD Boundary* (50 kb segments on each side of the TAD), *Inside Boundary* (two 50 kb segments inside the TAD, just interior to the TAD boundaries), *Outside Boundary* (two 50 kb regions outside the TAD, just exterior to the TAD boundaries), *TAD Interior* (the part of the TAD that's not in the *Inside Boundary* segments), and *TAD Exterior* (all other parts of the genome). With the CTCF peak widths being much smaller (median width = 273 bp) than the granularity of our TAD scaffold (50 kb resolution), we assumed that a peak would not span two segments and assigned each peak to its genomic segment based only on the peak's midpoint locus. Finally, we counted the number of peaks in each segment category, normalizing that count by the aggregate length of genomic segments in that category.

Using the ChIP-seq data, we also assess the precision of our inferred TAD boundaries. We evaluate binding prevalence in 50 kb regions on either side of our predicted TAD boundary (we recall that our TAD scaffold is inferred at a 50 kb resolution). Supporting our inference, we find that the regions adjoining the boundary have substantially lower CTCF binding rates than at the TAD boundary itself. However, the CTCF binding rate in these adjoining regions is still higher than the rate observed in the TAD exterior or interior. This could be explained by minor boundary variations across cell types (as is indeed known to happen) or if the ”effective” TAD boundary is wider than our 50 kb definition. We note that the bag-of-genes model for TADs is designed to handle such ambiguity: as long as the gene memberships in a TAD are unchanged, minor variations in its boundaries are immaterial.

Expression quantitative trait loci. From the GTEx Analysis Release V8 [49], we sourced data for all 49 tissues with available single-tissue eQTL data. We filtered the data, limiting ourselves to eQTLs with p-value less than 10^{-5}. For each gene and eQTL pair, we computed the genomic distance between the transcription start site (TSS) of the gene and the eQTL locus; based on it, we assigned the $\langle$ TSS, eQTL $\rangle$ pair to one of the genomic-distance bins partitioned by the following cut-points (all units in kb): [0, 5, 10, 20, 30, 50, 75, 100, 150, 250, 350, 450, 550, 650, 750, 1000]. We limited the upper bound of our evaluation range to 1 Mb, since the GTEx corpus limits single-tissue eQTL reports to this range. We counted the number of observed pairs in each bin, separately tracking pairs where the TSS and eQTL loci shared a TAD and pairs where they did not;

we additionally separated pairs where the eQTL locus was upstream of TSS from those where it was downstream. The probability $p(d)$ of an eQTL occurring at a distance d from the TSS was then estimated by dividing the number of observed $\langle$TSS, eQTL$\rangle$ pairs in each bin by the midpoint of the bin's genomic range.

Comparison of contextual transcriptional similarity (CTS) with co-expression. We computed CTS by leveraging the whole-human scGPT model, accessed from its GitHub repository (https://github.com/bowang-lab/scGPT). Gene embeddings were extracted from the "encoder.embedding.weight" matrix, using gene row mappings provided in the accompanying vocab JSON file. Chromosome and gene location data were obtained from the hg38 reference genome. For all gene pairs located on the same chromosome and within 2 megabases of genomic distance, we calculated Pearson correlations between their embedding vectors, defining these correlations as the CTS for each pair.

To calculate co-expression, we analyzed cell atlas datasets downloaded from CELLxGENE (https://cellxgene.cziscience.com/datasets), focusing on the Human Breast Cell Atlas, Human Brain Cell Atlas, Human Cell Atlas of Fetal Gene Expression, and Human Heart Atlas. Raw transcript counts for each cell were normalized to a total of 10,000 and subsequently log-transformed using log1p. For genes present across all datasets, we computed Pearson correlations of expression values for gene pairs on the same chromosome within 2 megabases. Finally, we compared CTS and co-expression, analyzing how these measures differ in capturing gene relationships.

Gene family analysis. To investigate gene family-specific patterns, we filtered gene pairs such that both genes belonged to one of the families of interest: olfactory receptors, protocadherins, histones, or keratins. After calculating CTS and co-expression, we highlighted gene pairs from specific families and compared their values to the broader distribution (**Fig. 3E-H**).

We also found genes in certain gene families primarily reside in the same few TADs. Olfactory receptors are plentiful in TADs on chr11-4150000-5450000, chr11-55000000-56850000, and chr11-123750000-124600000. Protocadherins primarily reside in TADs on chr5-140750000-141600000. Histones primarily reside in TADs on chr6-26000000-26600000 and chr6-27150000-27950000. Lastly, keratins primarily reside in TADs on chr17-40600000-41650000, chr21-30300000-31100000 and chr12-52150000-53050000.

Comparison of CTS to other gene function embedding similarity scores. We compared CTS to other gene function embedding similarity scores to determine whether scGPT gene embeddings capture functional information beyond gene expression. We included Gene2vec, which generates 200-dimensional vectors using co-expression patterns and functional gene sets, and HiG2Vec, which creates 1000-dimensional Poincaré embeddings to represent the hierarchical structure of Gene Ontology and gene semantics. We obtained Gene2vec embeddings from https://github.com/jingcheng-du/Gene2vec and HiG2Vec embeddings from https://github.com/JaesikKim/HiG2Vec. For both methods, we computed gene pair similarities using Pearson correlation and related these similarities to CTS, focusing on gene pairs common to both subsets.

CTS and co-expression for TAD vs non-TAD gene pairs and across genomic distance intervals. We categorized gene pairs as TAD pairs if they resided in the same TAD based on the TAD Map; otherwise, we classified them as non-TAD pairs. To determine genomic distance, we averaged the distances between the start and end positions of the two genes. We stratified all gene pairs into genomic distance intervals of 0–20 kb, 20–50 kb, 50–200 kb, 200–500 kb, and 500–1000 kb, with intervals defined as up to but not including the upper bound. We grouped gene pairs by TAD status and interval category and calculated their average CTS and co-expression.

Determining the scaling factor that aligns TAD gene pairs with non-TAD gene pairs. We examined how uniformly scaling CTS for TAD gene pairs best aligns with CTS for non-TAD gene pairs. Across all genomic distance intervals, CTS for TAD gene pairs consistently exceeded that for non-TAD gene pairs. We connected average CTS values per interval to form curves and calculated the area under the curve (AUC) using trapezoidal integration. To evaluate the impact of TADs, we scaled the average CTS for TAD gene pairs iteratively by factors ranging from 0 to 1 while keeping CTS for non-TAD gene pairs fixed. For each scaled TAD CTS, we computed the difference between its AUC and the fixed non-TAD AUC. The scaling factor that minimized the difference in AUCs quantified the effect of TADs on gene pair CTS.

Perturb-seq to identify targets impacting gene expression within and outside TADs. We analyzed the genome-wide Perturb-seq dataset (https://gwps.wi.mit.edu/) from Replogle et al. [74], generated using the K562 chronic myeloid leukemia cell line. We used this dataset to identify genetic perturbations that most disrupt overall transcription, as well as transcription specifically in genes located within TADs and those outside TADs. The dataset has 9866 unique CRISPRi gene-targeted perturbations and 1,914,250 total gene-targeted perturbed cells, with 75,328 non-targeting cells. For each perturbation, we used DESeq2 [75, 76] to calculate fold changes in all captured gene expression relative to a randomly downsampled group of non-targeting cells. We used TAD Map and Ensembl v102 gene annotations to categorize all protein-coding genes as either TAD genes (those residing in a TAD) or non-TAD genes (those outside TADs). Transcriptional disruption was quantified as the variance of DESeq2 log2 fold-change (β) values for protein-coding genes, calculated separately for TAD and non-TAD genes. We repeated this analysis five times with different randomly downsampled non-targeting groups and averaged the variance metrics to evaluate the perturbation-specific transcriptional impacts on all genes, TAD genes and non-TAD genes.

CTS and co-expression in adjacent genes. We use gene information from Ensembl v102 to denote if gene pairs are adjacent or neighoring gene pairs, as well as if they reside on the same strand or if they reside on opposite strands. For analysis with CTS, we used the same gene pairs from previous analysis. For our analysis with adjacent gene pairs For co-expression, we tested on 794 (human) and 69 (mouse) bulk RNA-seq datasets sourced from the ENCODE database (again encompassing a variety of cell types).

Intergenic transcripts. We sourced data from Agostini et al.'s study of intergenic transcripts. They collected and analyzed data from 38 publicly available datasets, covering over 2.5 billion uniquely mapped reads. We limited ourselves to 11,417 intergenic transcripts that are currently unannotated (the others corresponded most frequently to non-coding RNA fragments). As in the original study, each transcript was mapped to its adjoining genes as per the Gencode 27 reference. We further annotated each intergenic transcript by whether its adjoining genes were a) shared a TAD, and b) if they were oriented similarly.

Bootstrap test for clustering of expressed genes into TADs. The bootstrap test operates on data from a single cell (in scRNA-seq data) or a single tissue (in bulk RNA-seq data). Given scRNA-seq readout from any cell, we compute k, the number of genes with non-zero transcript counts in the cell. Treating gene activity as a binary event, we then generate 500 bootstrap samples of single-cell gene expression in each of which we randomly choose k protein-coding genes to be active. For both the actual observation and the bootstrap samples, we map these genes to the TAD Map, computing the number of TADs $n(p, k)$ which have p or more active genes; here, $p = 1$ corresponds to identifying the set of TADs with non-zero usage. We estimate the mean and standard deviation of the distribution $n(p, k)$ from the bootstrap samples and, using that, compute the z-score for the actual observation. In bulk RNA-seq data, which includes gene expressed aggregate from a collection of cells, almost all genes have some non-zero expression. There, we pre-set a threshold k (say, 5000) and limit ourselves to the top-k genes by transcript count; the bootstrap test and z-score are computed for this k.

In the test above, if a gene spans two TADs we count it in both TADs. This avoids us having to assign the gene to one TAD or the other arbitrarily. We also believe this to be the conservative choice for our clustering test: it will lead to more TADs per gene— i.e., lower clustering of expressed genes into TADs— than if we were to assign the gene to just one TAD or the other.

scRNA-seq cell differentiation datasets. We sourced scRNA-seq data from the CytoTrace database made available by Gulati et al. [95]. The study collected and curated scRNA-seq cell differentiation studies across multiple species, covering a variety of protocols and tissues. We chose this corpus as our scRNA-seq testbed, since it covers a diversity of protocols and tissues and allowed us to extend our analysis of also study TAD usage during cell differentiation. Of the 43 datasets available on `cytotrace.stanford.edu` (while their webpage lists 47 entries, 4 rows are blank), we filtered out studies with fewer than 200 cells and those that did not originate from human or mouse tissue, leaving us with 33 studies. We had difficulty converting 3 of these from the original *RDS* format to a *Scanpy*-compatible format and limited ourselves to the remaining 30 (which covered 70,243 cells); these formed our scRNA-seq corpus.

Designation of early vs. late-stage cells during differentiation. When analyzing TAD usage during cell differentiation, we further limited ourselves to the 24 scRNA-seq datasets where the putative differentiation trajectory did not have any branches, allowing us to reliably order cells along a differentiation time course. We used Gulati et al.'s annotations of differentiation stage in each study and considered two measures of early versus late differentiation: 1) consider only the cells at the first differentiation stage (*order* = 0 in CytoTrace) as "early", with all other cells comprising the "late" stage, or 2) partition the cells in each dataset by *order* so that cells are divided roughly equally between the first ("early") and second ("late") halves.

CTS for immune cell aging datasets. We analyzed immune cell scRNA-seq data from Suo et al. [99] and the single-cell transcriptomic atlas [100], both sourced from CELLxGENE. From Suo et al.'s dataset, we selected cells whose development stage is in Carnegie stages 13, 19, and 20, designating this subset as DS1. Cells from the 16th and 17th weeks post-fertilization development stage were filtered and grouped as DS2. For the mature immune cell atlas, we included cells from donors aged 22 to 61 years, labeled as DS Mature, and those from donors older than 62 years, labeled as DS Old. To ensure consistency across datasets, we downsampled each to 50,000 cells using Geosketch [101].

For each dataset (DS1, DS2, DS Mature, DS Old), we fine-tuned pre-trained whole-human scGPT model on a masked gene expression prediction task for 10 epochs. During fine-tuning, all scGPT parameters were frozen except for the gene embeddings ("encoder.embedding.weight"). We adopted default hyperparameters provided in the scGPT tutorials (https://github.com/bowang-lab/scGPT/tree/main/tutorials). Following fine-tuning, we extracted updated gene embeddings and calculated contextual transcriptional similarity (CTS) for the same gene pairs analyzed earlier. To compare overall CTS across datasets, we evaluated CTS trends across genomic distance intervals (**Fig. 5C-D**) and computed the area under the curve (AUC) for TAD and non-TAD gene pairs (**Fig. 5E**) using trapezoidal integration.

CTS for cell-type specific immune cell datasets We combined DS1 and DS2 datasets to define the "Early" developmental stage cells and combined DS Mature and DS Old datasets to define the "Mature" stage cells. Within these aggregated stages, we identified five cell types containing more than 2,500 cells: general T cells, naive thymus-derived CD4-positive, alpha-beta T cells, macrophages, monocytes, and erythrocytes. For each cell type in both Early and Mature stages, we fine-tuned the scGPT model for 10 epochs with masked gene expression prediction. Following fine-tuning, we calculated contextual transcriptional similarity (CTS) for gene pairs from fine-tuned gene embeddings and quantified overall CTS by computing the area under the curve (AUC) for TAD and non-TAD gene pairs (**Fig. 5F**).

To assess developmental changes in transcriptional organization, we calculated the difference in overall AUC (averaged across TAD and non-TAD gene pairs) between Early and Mature stages for each cell type. This analysis allowed us to identify which cell types exhibited the greatest change in AUC as they transitioned from Early to Mature stages, providing insights into the dynamics of transcriptional regulation during cellular maturation (**Fig. 5G**).

Transcriptional clustering regarding TAD usage in Cancer. For TAD usage in cancer, we compare bulk RNA-seq measurements of normal and primary-tumor tissue in blood, brain, lung, and renal cancers from The Cancer Genome Atlas (TCGA) database [102]. As a caveat, we note that our analysis can only infer correlation, not causation: while mutations leading to the mis-specification of TAD boundaries have been associated with certain cancers [136–138], there are diverse epigenetic mechanisms underpinning tumorigenesis [139–141], and the increased expression clustering we observe in the TADs of tumor cells could either be a cause or an effect of these mechanisms. Another caveat regarding our analysis is that the gene groupings were inferred from the consensus TAD scaffold, but in some cancer cells the TAD architecture may have changed. However, the statistical result that these gene groups are over-represented in tumor cells' transcriptional profiles nonetheless remains valid.

CTS in cancer datasets To assess how cancerous states and treatment conditions affect contextual transcriptional similarity (CTS), we analyzed scRNA-seq data from a recent colorectal cancer dataset by Moorman et al. [103], which tracks cellular plasticity during metastasis (data available at https://github.com/dpeerlab/progressive-plasticity-crc-metastasis). The dataset includes non-cancerous cells, primary cancer cells, and metastatic cancer cells, with each cell type categorized further by treatment status—either treated with 5-fluorouracil-based chemotherapy or untreated. This totals six distinct datasets.

For each dataset, we fine-tuned the pre-trained scGPT model for 10 epochs, using the same settings as described for cell type-specific immune datasets. Following fine-tuning, we calculated CTS for gene pairs and quantified overall CTS by computing the area under the curve (AUC) for both TAD and non-TAD gene pairs. This analysis enabled a systematic comparison of transcriptional dynamics across cancerous and non-cancerous states, as well as between treated and untreated conditions.

TAD signatures: probabilistic model. We define a 2-component mixture model to infer TAD activation probabilities of a single-cell dataset. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the gene expression values for a single-cell dataset with n cells and p genes, with a particular cell c's expression being $\boldsymbol{x}^{(c)} \in \mathbb{R}^{p}$. We assume that transcriptionally active TADs (“ON”) correspond to a higher rate of per-gene expression while inactive TADs (”OFF”) correspond to lower rates of gene expression. As mentioned earlier, our model allows inactive TADs to also generate non-zero gene expression, albeit at a lower rate than the “active” TADs. Doing so increases our robustness to noise and allows for one-off gene expression in a TAD. In the mixture model, the probability of $\boldsymbol{x}^{(c)}$'s expression in c is

$$P(\boldsymbol{x}^{(c)}; \boldsymbol{\lambda}) = \prod_{t \in \mathcal{T}} \prod_{g \in t} \Big(P(S_t = 1) P(x_g^{(c)} | S_t = 1; \boldsymbol{\lambda}) + P(S_t = 0) P(x_g^{(c)} | S_t = 0; ; \boldsymbol{\lambda}) \Big)$$

where $\boldsymbol{\lambda} = \{\lambda_{ON}, \lambda_{OFF}\}$; $\mathcal{T}$ is the set of all TADs with one or more genes; each TAD $t \in \mathcal{T}$ is a set of genes, with g being one such gene; S_t is the Bernoulli random variable indicating the activation state of t, with $S_t = 1$ or 0 corresponding to t being “ON” or “OFF”, respectively; and $x_g^{(c)}$ is the expression of gene g in the cell c. Here, $\mathcal{T}, t$ and the gene memberships in TADs are sourced from the species-specific TAD Map. We model that gene expression values are Poisson-distributed counts, though this assumption can be relaxed:

$$P(x_g^{(c)}|S_t = 1; \boldsymbol{\lambda}) = \frac{\lambda_{ON}^{x_g^{(c)}} e^{-\lambda_{ON}}}{x_g^{(c)}!}$$

$$P(x_g^{(c)}|S_t = 0; \boldsymbol{\lambda}) = \frac{\lambda_{OFF}^{x_g^{(c)}} e^{-\lambda_{OFF}}}{x_g^{(c)}!}$$

To infer the TAD signature for a particular dataset, we fit this model with the expectation maximization (EM) algorithm, seeking to maximize the log-likelihood over all cells:

$$\ell(\boldsymbol{X}; \boldsymbol{\lambda}) = \sum_{c} \sum_{t \in \mathcal{T}} \sum_{g \in t} \log \left(P(S_t = 1) P(x_g^{(c)}|S_t = 1; \boldsymbol{\lambda}) + P(S_t = 0) P(x_g^{(c)}|S_t = 0; \boldsymbol{\lambda}) \right)$$

Since the maximum likelihood estimator of the Poisson rate parameter is just the sample mean, the implementation of the EM algorithm is simplified: each maximization round assigns λ_{ON} and λ_{OFF} as averages of observed gene expression values across all genes, weighted by the containing TAD’s activation probability $P(t = 1)$. Also, we note that quasi-Poisson generalizations result in the same maximum likelihood estimator for the expected value of the rate parameter, suggesting that even in cases where the Poisson assumptions do not hold, the corresponding estimate is reasonable.

Log-odds transformation of TAD signatures. We recommend a log-odds transformation when using TAD signatures to generate data representations for clustering, visualization, or predictive analysis. Many such analyses implicitly or explicitly rely on Euclidean distances between observations.

The log-odds transformation converts probabilities (which are in $[0, 1]$ range) to the full range of values in $\mathbb{R}$, making it more amenable to such distance measures.

TAD signatures for cancer cell line heterogeneity. We acquired scRNA-seq data for cancer cell lines from Kinker et al.'s pan-cancer study [111], limiting ourselves to cell lines for which drug response data from the PRISM study was also available [110]; this resulted in data on 51,321 cells spanning 193 cancer cell lines. For each cell line, we computed six population statistics: mean, standard deviation and skew, based on either TAD signatures or log-and-count-normalized transcript counts. We reduced each cell line's drug repsonse profile to the top 50 PCs and considering each PC as an independent regression target, we performed a principal component regression using per-cell-line statistics. Thus a total of 300 (50×6) regressions were performed, each with 193 observations. In each regression, the $\mathbf{x}$ values were reduced to the top 10 principal components, ensuring that all regressions had identical complexity. An R^2 was computed for each regression, indicating the predictive power of the scRNA-seq summary statistic against the drug response measure.

TAD signatures for cell type inference. We acquired single-cell data from the `https://cellxgene.cziscience.com/` portal, obtaining *AnnData*-formatted [142] datasets from single-cell RNA-seq studies of the human lung (10x sub-study; European Genome-Phenome Archive accession EGAS00001004344; 65,662 cells; [114]), T-cells (GEO accession GSE126030; 51,876 cells; [115]), and breast epithelial cells (GEO accession GSE164898; 31,696 cells; [113]). Cells with fewer than 20 active genes and genes active in less than 10 cells were removed. For the breast tissue data, the dataset annotations seemed to suggest samples were grouped in two broad batches and, to reduce batch effects, we limited ourselves to the larger batch (17,153 cells). The data was then count normalized (to 10^6) and log transformed using *Scanpy*. Gene identifiers were converted to Ensemble v102, genes were mapped to the human TAD Map inferred in this work, and TAD signatures were estimated. For each dataset, we then generated the following representations: i) principal component analysis (PCA) with 50 components, ii) log-odds ($= \log \frac{p}{1-p}$) transformation of the TAD signatures computed on the dataset, followed by a 50-component PCA, and iii) a concatenation of the previous two representations. Leiden clustering [112] using *Scanpy* was performed on each of these representations. The datasets from cellxgene portal contained expert-annotated cell-type labels for each cell and we computed the adjusted Rand index (ARI) of the overlap between computed Leiden clusterings and the expert labels.

Quantification and Statistical Analysis

Statistical tests were conducted using version 1.3.1 of the SciPy Python package and R version 4.1.1.

Software availability, utility, and efficiency

Pre-computed TAD Maps and consensus TAD boundary estimates for human and mouse genomes are available at `http://singhlab.net/tadmap`. TAD signatures can be computed using the Python package `tadmap`, available via `pip`, `conda` or `GitHub`. The package also provides direct programmatic access to the TAD Map. Documentation for the package is available at `https://tadmap.readthedocs.io/`. TAD signature computations do not require a GPU and can be performed on a personal computer: processing of a scRNA-seq dataset comprising 51,321 cells required 7 minutes of run-time and 8 GB of memory when using a single Intel Xeon 3.47 GHz processor.

Acknowledgements

We thank Anders Sejr Hansen and Brian Hie for helpful discussions. RS and BB were partially supported by the NIH grants R01GM081871 and R35GM141861. RS and HL were also supported by the Chan-Zuckerberg Initiative.

Author Contributions

All authors conceived of, contributed to, and wrote the paper. RS developed the TAD Map method with BB; HL led the scGPT integration and the related analysis.

Declaration of Interests

None

References

1. Tanouchi, Y. *et al.* A noisy linear map underlies oscillations in cell size and gene expression in bacteria. *Nature* **523,** 357–360 (2015).
2. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
3. Gibcus, J. H. & Dekker, J. The hierarchy of the 3D genome. *Mol Cell* **49,** 773–782 (Mar. 2013).
4. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat Rev Genet* **21,** 207–226 (2020).
5. Grob, S. & Cavalli, G. Technical Review: A Hitchhiker's Guide to Chromosome Conformation Capture. *Methods Mol Biol* **1675,** 233–246 (2018).
6. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543,** 519–524 (2017).
7. Quinodoz, S. A. *et al.* Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174,** 744–757 (2018).
8. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462,** 58–64 (2009).
9. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
10. De Wit, E. & De Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes & development* **26,** 11–24 (2012).
11. Pope, B. D. *et al.* Topologically associating domains are stable units of replication-timing regulation. *Nature* **515,** 402–405 (2014).
12. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167,** 1170–1187 (2016).
13. Nora, E. P., Dekker, J. & Heard, E. Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? *Bioessays* **35,** 818–828 (2013).
14. Chang, L.-H., Ghosh, S. & Noordermeer, D. TADs and their borders: free movement or building a wall? *Journal of molecular biology* **432,** 643–652 (2020).
15. Ibrahim, D. M. & Mundlos, S. The role of 3D chromatin domains in gene regulation: a multi-facetted view on genome organization. *Current opinion in genetics & development* **61,** 1–8 (2020).

16. Barutcu, A. R., Maass, P. G., Lewandowski, J. P., Weiner, C. L. & Rinn, J. L. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nature communications* **9,** 1444 (2018).

17. Ushiki, A. *et al.* Deletion of CTCF sites in the SHH locus alters enhancer–promoter interactions and leads to acheiropodia. *Nature communications* **12,** 2282 (2021).

18. Nora, E. P. *et al.* Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169,** 930–944 (2017).

19. Kubo, N. *et al.* Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. *Nature structural & molecular biology* **28,** 152–161 (2021).

20. Huang, H. *et al.* CTCF mediates dosage-and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nature genetics* **53,** 1064–1074 (2021).

21. Gunsalus, L. M., Keiser, M. J. & Pollard, K. S. In silico discovery of repetitive elements as key sequence determinants of 3D genome folding. *Cell Genomics* **3** (2023).

22. Bashkirova, E. V. *et al.* Opposing, spatially-determined epigenetic forces impose restrictions on stochastic olfactory receptor choice. *Elife* **12,** RP87445 (2023).

23. Goldman, A. L., van Naters, W. V. d. G., Lessing, D., Warr, C. G. & Carlson, J. R. Coexpression of two functional odor receptors in one neuron. *Neuron* **45,** 661–666 (2005).

24. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **35,** 1798–1828 (2013).

25. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

26. Harmston, N. *et al.* Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nature communications* **8,** 1–13 (2017).

27. Schmitt, A. D. *et al.* A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports* **17,** 2042–2059 (2016).

28. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Molecular cell* **62,** 668–680 (2016).

29. McArthur, E. & Capra, J. A. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am J Hum Genet* **108,** 269–283 (2021).

30. Budowski-Tal, I., Nov, Y. & Kolodny, R. FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proceedings of the National Academy of Sciences* **107,** 3481–3486 (2010).

31. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods,* 1–11 (2024).

32. Rippe, K. Liquid–liquid phase separation in chromatin. *Cold Spring Harbor perspectives in biology* **14,** a040683 (2022).

33. Hansen, J. C., Maeshima, K. & Hendzel, M. J. The solid and liquid states of chromatin. *Epigenetics & Chromatin* **14,** 50 (2021).

34. Shrinivas, K. *et al.* Enhancer features that drive formation of transcriptional condensates. *Molecular cell* **75,** 549–561 (2019).

35. Pei, G., Lyons, H., Li, P. & Sabari, B. R. Transcription regulation by biomolecular condensates. *Nature Reviews Molecular Cell Biology,* 1–24 (2024).

36. Bhat, P., Honson, D. & Guttman, M. Nuclear compartmentalization as a mechanism of quantitative control of gene expression. *Nature Reviews Molecular Cell Biology* **22,** 653–670 (2021).

37. Sharp, P. A., Chakraborty, A. K., Henninger, J. E. & Young, R. A. RNA in formation and regulation of transcriptional condensates. *Rna* **28,** 52–57 (2022).

38. Henninger, J. E. & Young, R. A. An RNA-centric view of transcription and genome organization. *Molecular Cell* **84,** 3627–3643 (2024).

39. Wei, M.-T. *et al.* Nucleated transcriptional condensates amplify gene expression. *Nature cell biology* **22,** 1187–1196 (2020).

40. Sabari, B. R., Dall'Agnese, A. & Young, R. A. Biomolecular condensates in the nucleus. *Trends in biochemical sciences* **45,** 961–977 (2020).

41. Boija, A. *et al.* Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* **175,** 1842–1855 (2018).

42. Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. *Cell* **164,** 1110–1121 (2016).

43. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).

44. Liu, T. *et al.* TADKB: Family classification and a knowledge base of topologically associating domains. *BMC genomics* **20,** 1–17 (2019).

45. Braccioli, L. & de Wit, E. CTCF: a Swiss-army knife for genome organization and transcription regulation. *Essays Biochem* **63,** 157–165 (Apr. 2019).

46. Gomez-Marín, C. *et al.* Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proc Natl Acad Sci U S A* **112,** 7542–7547 (2015).

47. Mirny, L. A., Imakaev, M. & Abdennur, N. Two major mechanisms of chromosome organization. *Current opinion in cell biology* **58,** 142–152 (2019).

48. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research* **46,** D794–D801 (2018).

49. Consortium, G. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369,** 1318–1330 (2020).

50. Strunz, T. *et al.* A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. *Sci Rep* **8,** 5865 (2018).

51. Strunz, T. *et al.* A mega-analysis of expression quantitative trait loci in retinal tissue. *PLoS Genet* **16,** e1008934 (2020).

52. Wang, J. *et al.* Single-cell co-expression analysis reveals distinct functional modules, co-regulation mechanisms and clinical outcomes. *PLoS computational biology* **12,** e1004892 (2016).

53. Williams, A., Spilianakis, C. G. & Flavell, R. A. Interchromosomal association and gene regulation in trans. *Trends in genetics* **26,** 188–197 (2010).

54. Yin, W., Mendoza, L., Monzon-Sandoval, J., Urrutia, A. O. & Gutierrez, H. Emergence of co-expression in gene regulatory networks. *PloS one* **16,** e0247671 (2021).

55. Flaherty, E. & Maniatis, T. The role of clustered protocadherins in neurodevelopment and neuropsychiatric diseases. *Current opinion in genetics & development* **65,** 144–150 (2020).

56. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618,** 616–624 (2023).

57. Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nature Methods,* 1–11 (2024).

58. Kedzierska, K. Z., Crawford, L., Amini, A. P. & Lu, A. X. Assessing the limits of zero-shot foundation models in single-cell biology. *bioRxiv,* 2023–10 (2023).

59. Megill, C. *et al.* cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv.* eprint: `https://www.biorxiv.org/content/early/2021/04/06/2021.04.05.438318.full.pdf`. `https://www.biorxiv.org/content/early/2021/04/06/2021.04.05.438318` (2021).

60. Reed, A. D. *et al.* A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nature Genetics* **56,** 652–662 (2024).

61. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382,** eadd7046 (2023).

62. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370,** eaba7721 (2020).

63. Reichart, D. *et al.* Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies. *Science* **377,** eabo1984 (2022).

64. Du, J. *et al.* Gene2vec: distributed representation of genes based on co-expression. *BMC genomics* **20,** 7–15 (2019).

65. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550 (2005).

66. Kim, J., Kim, D. & Sohn, K.-A. HiG2Vec: hierarchical representations of gene ontology and genes in the Poincaré ball. *Bioinformatics* **37,** 2971–2980 (2021).

67. MacAlpine, D. M. & Almouzni, G. Chromatin and DNA replication. *Cold Spring Harbor perspectives in biology* **5,** a010207 (2013).

68. Woo, Y. H. & Li, W. H. Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. *Proc Natl Acad Sci U S A* **108,** 3306–3311 (Feb. 2011).

69. Ibn-Salem, J., Muro, E. M. & Andrade-Navarro, M. A. Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic acids research* **45,** 81–91 (2017).

70. Long, H. S. *et al.* Making sense of the linear genome, gene function and TADs. *bioRxiv* (2020).

71. Demmerle, J., Hao, S. & Cai, D. Transcriptional condensates and phase separation: condensing information across scales and mechanisms. *Nucleus* **14,** 2213551 (2023).

72. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169,** 13–23 (2017).

73. Wang, B. *et al.* Liquid–liquid phase separation in human health and diseases. *Signal Transduction and Targeted Therapy* **6,** 290 (2021).

74. Replogle, J. M. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185,** 2559–2575 (2022).

75. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 1–21 (2014).

76. Muzellec, B., Telenczuk, M., Cabeli, V. & Andreux, M. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics* (2023).

77. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* **14,** 1–14 (2013).

78. El Khattabi, L. *et al.* A pliable mediator acts as a functional rather than an architectural bridge between promoters and enhancers. *Cell* **178,** 1145–1158 (2019).

79. Soutourina, J. Transcription regulation by the Mediator complex. *Nature reviews Molecular cell biology* **19,** 262–274 (2018).

80. Dong, P. *et al.* Cohesin prevents cross-domain gene coactivation. *Nature Genetics,* 1–11 (2024).

81. Kojic, A. *et al.* Distinct roles of cohesin-SA1 and cohesin-SA2 in 3D chromosome organization. *Nature structural & molecular biology* **25,** 496–504 (2018).

82. Rao, S. S. *et al.* Cohesin loss eliminates all loop domains. *Cell* **171,** 305–320 (2017).

83. Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* **15,** 234–246 (2014).

84. Samata, M. *et al.* Intergenerationally maintained histone H4 lysine 16 acetylation is instructive for future gene activation. *Cell* **182,** 127–144 (2020).

85. Palacio, M. & Taatjes, D. J. Merging established mechanisms with new insights: condensates, hubs, and the regulation of RNA polymerase II transcription. *Journal of Molecular Biology* **434,** 167216 (2022).

86. Wagner, E. J., Tong, L. & Adelman, K. Integrator is a global promoter-proximal termination complex. *Molecular cell* **83,** 416–427 (2023).

87. Lafontaine, D. L., Riback, J. A., Bascetin, R. & Brangwynne, C. P. The nucleolus as a multiphase liquid condensate. *Nature reviews Molecular cell biology* **22,** 165–182 (2021).

88. Dillinger, S., Straub, T. & Nemeth, A. Nucleolus association of chromosomal domains is largely maintained in cellular senescence despite massive nuclear reorganisation. *PLoS One* **12,** e0178821 (2017).

89. Avni, E. & Snir, S. Horizontal Gene Transfer Phylogenetics: A Random Walk Approach. *Molecular Biology and Evolution* **37,** 1470–1483 (2020).

90. Makarova, K. S., Wolf, Y. I., Snir, S. & Koonin, E. V. Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems. *Journal of Bacteriology* **193,** 6039–6056 (2011).

91. Agostini, F., Zagalak, J., Attig, J., Ule, J. & Luscombe, N. M. Intergenic RNA mainly derives from nascent transcripts of known genes. *Genome Biol* **22,** 136 (May 2021).

92. Gorkin, D. U., Leung, D. & Ren, B. The 3D genome in transcriptional regulation and pluripotency. *Cell stem cell* **14,** 762–775 (2014).

93. Franchini, L. F. & Pollard, K. S. Genomic approaches to studying human-specific developmental traits. *Development* **142,** 3100–3112 (2015).

94. Pollex, T. *et al.* Enhancer–promoter interactions become more instructive in the transition from cell-fate specification to tissue differentiation. *Nature Genetics* **56,** 686–696 (2024).

95. Gulati, G. S. *et al.* Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367,** 405–411 (2020).

96. Mogilenko, D. A., Shchukina, I. & Artyomov, M. N. Immune ageing at single-cell resolution. *Nature Reviews Immunology* **22,** 484–498 (2022).

97. Tan, L. *et al.* Lifelong restructuring of 3D genome architecture in cerebellar granule cells. *Science* **381,** 1112–1119 (2023).

98. Zhao, Y. *et al.* Multiscale 3D genome reorganization during skeletal muscle stem cell lineage progression and aging. *Science Advances* **9,** eabo1360 (2023).

99. Suo, C. *et al.* Mapping the developing human immune system across organs. *Science* **376,** eabo0510 (2022).

100. Consortium*, T. T. S. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376,** eabl4896 (2022).

101. Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell systems* **8,** 483–493 (2019).

102. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* **19,** A68 (2015).

103. Moorman, A. *et al.* Progressive plasticity during colorectal cancer metastasis. *Nature,* 1–8 (2024).

104. Rezayatmand, H., Razmkhah, M. & Razeghian-Jahromi, I. Drug resistance in cancer therapy: the Pandora's Box of cancer stem cells. *Stem cell research & therapy* **13,** 181 (2022).

105. Fatma, H. & Siddique, H. R. Cancer cell plasticity, stem cell factors, and therapy resistance: how are they linked? *Cancer and Metastasis Reviews* **43,** 423–440 (2024).

106. Hie, B. *et al.* Computational methods for single-cell RNA sequencing. *Annual Review of Biomedical Data Science* **3,** 339–364 (2020).

107. Sun, X. X. & Yu, Q. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacol Sin* **36,** 1219–1227 (Oct. 2015).

108. Adam, G. *et al.* Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis Oncol* **4,** 19 (2020).

109. Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine* **52,** 1452–1465 (2020).

110. Corsello, S. M. *et al.* Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer* **1,** 235–248 (Feb. 2020).

111. Kinker, G. S. *et al.* Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat Genet* **52,** 1208–1218 (Nov. 2020).

112. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* **9,** 1–12 (2019).

113. Bhat-Nakshatri, P. *et al.* A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Rep Med* **2,** 100219 (2021).

114. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587,** 619–625 (2020).

115. Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat Commun* **10,** 4706 (2019).

116. Boiarsky, R., Singh, N. M., Buendia, A., Getz, G. & Sontag, D. A deep dive into single-cell RNA sequencing foundation models. *bioRxiv,* 2023–10 (2023).

117. Kim, Y.-A., Cho, D.-Y. & Przytycka, T. M. Understanding Genotype-Phenotype Effects in Cancer via Network Approaches. *PLOS Computational Biology* **12,** e1004747 (2016).

118. Dao, P. *et al.* BeWith: A Between-Within Method to Discover Relationships Between Cancer Modules via Integrated Analysis of Mutual Exclusivity, Co-occurrence, and Functional Interactions. *PLOS Computational Biology* **13,** e1005695 (2017).

119. Gilbertson, E. N. *et al.* Machine learning reveals the diversity of human 3D chromatin contact patterns. *Molecular Biology and Evolution* **41,** msae209 (2024).

120. Kuang, S. & Pollard, K. S. Exploring the roles of RNAs in chromatin architecture using deep learning. *Nature Communications* **15,** 6373 (2024).

121. Zhang, Y. *et al.* Computational methods for analysing multiscale 3D genome organization. *Nature Reviews Genetics* **25,** 123–141 (2024).

122. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nature methods* **17,** 1111–1117 (2020).

123. Xiong, K., Zhang, R. & Ma, J. scGHOST: Identifying single-cell 3D genome subcompartments. *Nature methods,* 1–9 (2024).

124. Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nature biotechnology* **40,** 254–261 (2022).

125. Gao, V. R. *et al.* ChromaFold predicts the 3D contact map from single-cell chromatin accessibility. *Nature Communications* **15,** 9432 (2024).

126. Pancaldi, V. *et al.* Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs. *Cell Reports* **14,** 1246–1257 (2016).

127. Zhou, T. *et al.* GAGE-seq concurrently profiles multiscale 3D genome organization and gene expression in single cells. *Nature Genetics,* 1–11 (2024).

128. Arrastia, M. V. *et al.* Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nature biotechnology* **40,** 64–73 (2022).

129. Liu, Z. *et al.* Linking genome structures to functions by simultaneous single-cell Hi-C and RNA-seq. *Science* **380,** 1070–1076 (2023).

130. Wei, X. *et al.* HiCAR is a robust and sensitive method to analyze open-chromatin-associated genome organization. *Molecular cell* **82,** 1225–1238 (2022).

131. Li, W. *et al.* scNanoHi-C: a single-cell long-read concatemer sequencing method to reveal high-order chromatin structures within individual cells. *Nature Methods* **20,** 1493–1505 (2023).

132. Schilhabel, M. & Vingron, M. Inference of Interactions Between Chromatin Modifiers and Histone Modifications: From ChIP-Seq Data to Chromatin-Signaling. *Nucleic Acids Research* **42,** 13689–13695 (2014).

133. Brackley, C. A. *et al.* Polymer Physics Predicts the Effects of Structural Variants on Chromatin Architecture. *Nature Genetics* **50,** 662–667 (2018).

134. Bonev, B. *et al.* Multiscale 3D genome rewiring during mouse neural development. *Cell* **171,** 557–572 (2017).

135. Navarro Gonzalez, J. *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res* **49,** D1046–D1057 (Jan. 2021).

136. Flavahan, W. A. *et al.* Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs. *Nature* **575,** 229–233 (2019).

137. Hnisz, D., Schuijers, J., Li, C. H. & Young, R. A. Regulation and dysregulation of chromosome structure in cancer. *Annual Review of Cancer Biology* **2,** 21–40 (2018).

138. Lupianez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161,** 1012–1025 (2015).

139. Ruggero, K., Farran-Matas, S., Martinez-Tebar, A. & Aytes, A. Epigenetic regulation in prostate cancer progression. *Current molecular biology reports* **4,** 101–115 (2018).

140. Kumar, R., Li, D.-Q., Müller, S. & Knapp, S. Epigenomic regulation of oncogenesis by chromatin remodeling. *Oncogene* **35,** 4423–4436 (2016).

141. Pelham-Webb, B., Murphy, D. & Apostolou, E. Dynamic 3D chromatin reorganization during establishment and maintenance of pluripotency. *Stem Cell Reports* (2020).

142. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19,** 15 (Feb. 2018).

Figures

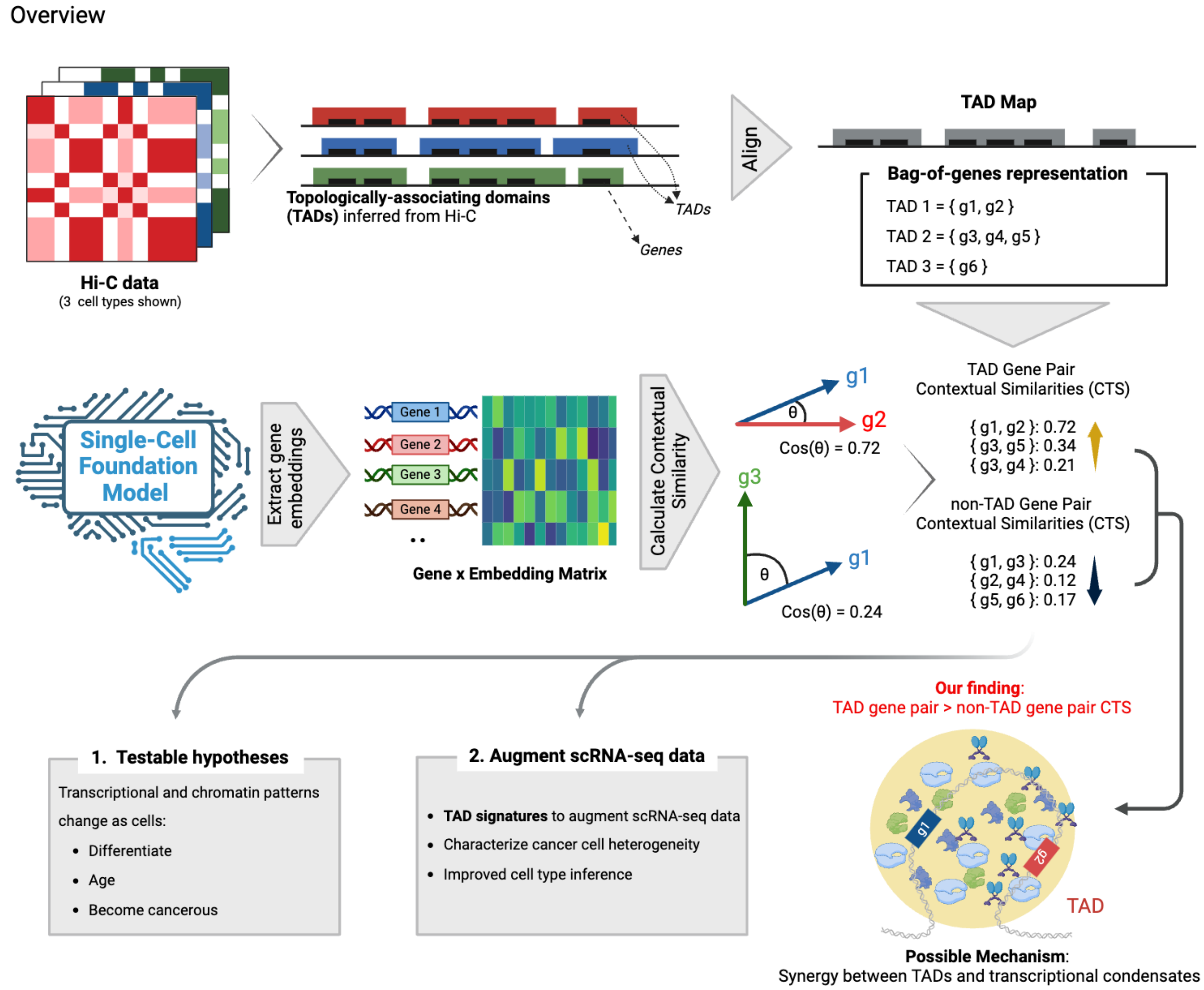

Overview
Hi-C data
(3 cell types shown)
Topologically-associating domains
(TADs) inferred from Hi-C
TADs
Genes
Align
TAD Map
Bag-of-genes representation
TAD 1 = { g1, g2 }
TAD 2 = { g3, g4, g5 }
TAD 3 = { g6 }
Single-Cell Foundation Model
Extract gene embeddings
Gene 1
Gene 2
Gene 3
Gene 4
Gene x Embedding Matrix
Calculate Contextual Similarity
g1
θ
g2
Cos(θ) = 0.72
g3
θ
g1
Cos(θ) = 0.24
TAD Gene Pair
Contextual Similarities (CTS)
{ g1, g2 }: 0.72
{ g3, g5 }: 0.34
{ g3, g4 }: 0.21
non-TAD Gene Pair
Contextual Similarities (CTS)
{ g1, g3 }: 0.24
{ g2, g4 }: 0.12
{ g5, g6 }: 0.17
Our finding:
TAD gene pair > non-TAD gene pair CTS
1. Testable hypotheses
Transcriptional and chromatin patterns change as cells:
• Differentiate
• Age
• Become cancerous
2. Augment scRNA-seq data
• TAD signatures to augment scRNA-seq data
• Characterize cancer cell heterogeneity
• Improved cell type inference
g1
g2
TAD
Possible Mechanism:
Synergy between TADs and transcriptional condensates

Figure 1: **Overview.** Topologically associating domains (TADs) are a key folding unit of the chromatin. While TADs are currently inferred separately for each cell type, TAD architectures have been observed to have good agreement across cell types. We find this agreement to be even stronger when TADs are represented simply as sets of genes, suggesting a bag-of-genes representation would have species-wide applicability. Applying maximum likelihood estimation, we compute a consensus TAD scaffold from Hi-C data and the corresponding bag-of-genes representation (the *TAD Map*, Fig. 2).

The TAD Map enables us to "impute" the high-level chromatin structure in each cell/tissue. We demonstrate that the genome partitioning implied by the TAD scaffold agrees with functional genomic data across a variety of cell/tissue types. (ii) We leverage gene representations from single-cell foundation models, pretrained with more than 33 million cells, to derive contextual transcriptional similarities (CTS) between gene pairs. These measures capture co-regulation beyond simple expression-correlation patterns between two genes (Fig. 3). (iii) We find CTS in TAD gene pairs significantly exceeds that in non-TAD gene pairs and hypothesize that a synergy between TADs and transcriptional condensates may explain it (Fig. 4). The TAD Map framework enables novel, testable hypothesis: We examine how TAD usage and CTS change as cells develop and age, or become cancerous (Fig. 4). (iv) We introduce *TAD signatures*, a probabilistic model of TAD activation inferred from single-cell RNA-seq (scRNA-seq) readouts, showing how they facilitate greater accuracy and robustness in downstream scRNA-seq analyses (Fig. 5).

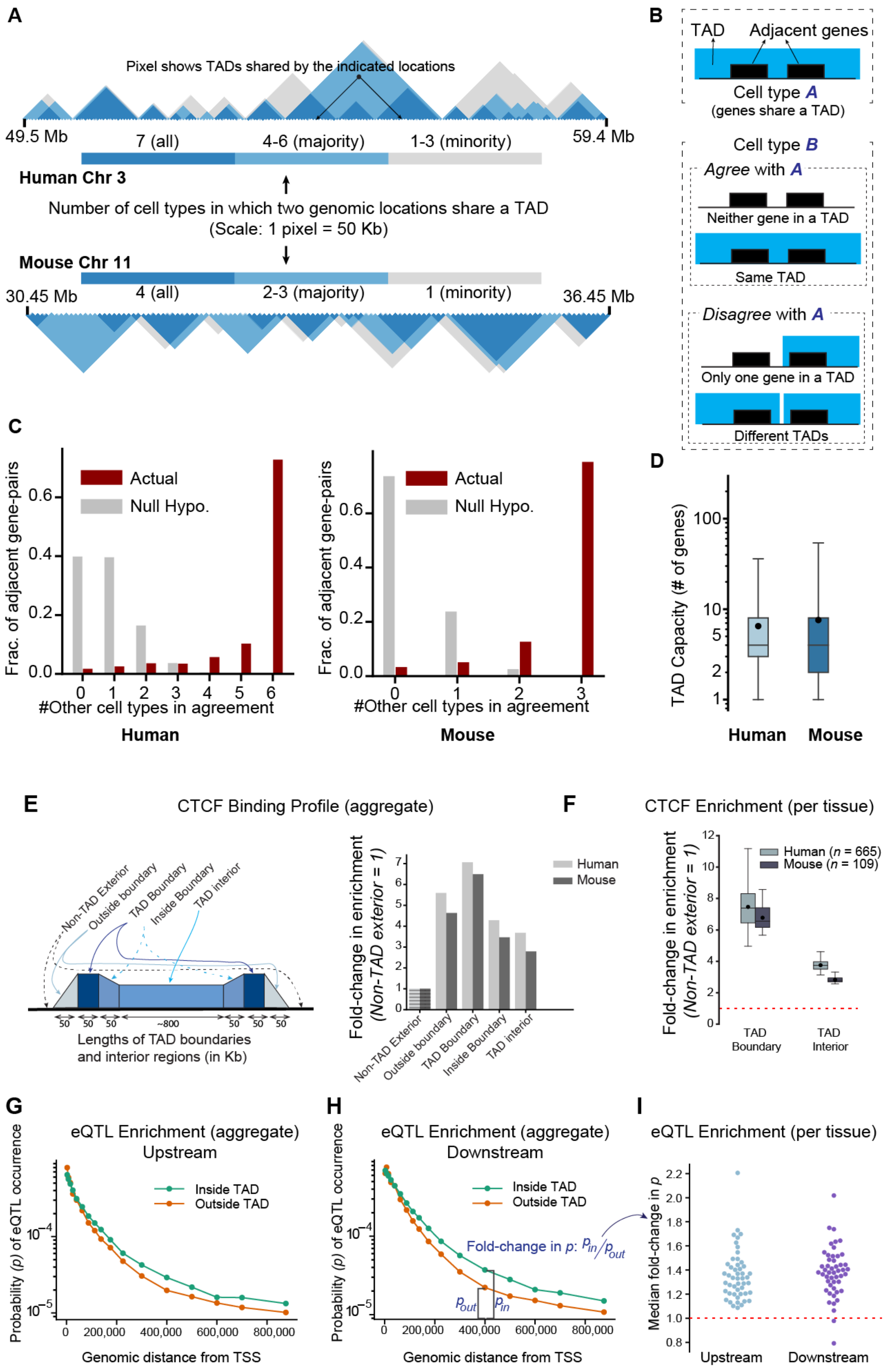


Figure 2: **Agreement between cell-type specific TAD architectures and the constructed consensus TAD scaffold.**

A) On representative segments of the human and mouse genomes, overlap of cell-type specific TAD architectures (7 cell types for human, 4 four mouse) are shown. Genomic positional range (in Mb) is shown on the horizontal axis and the number of cell types in which a pair of genomic loci co-occupy a TAD is indicated by the shaded triangles. In each species, the majority of the cell-type specific TAD architectures are in agreement on most genomic loci pairs; in many cases, *all* TAD architectures are in agreement. **B**) Schematic for the statistical test to evaluate if two bag-of-genes TAD representations are identical: given an adjacently-located gene pair that shares a TAD in at least one cell type's TAD architecture, we mark other cell type(s) to be in agreement if the gene pair is not separated by a TAD boundary. **C**) For each adjacent gene pair that shares a TAD in at least one cell type, the number of other cell types where the genes share a TAD. Under this test, 72.7% of adjacent gene pairs in human (78.9% in mouse) show complete agreement across all cell-type specific TAD architectures. The expected count under a simple null hypothesis (Methods) is below 1 in both species. **D**) Our consensus TAD scaffolds for each of human and mouse, computed by maximum likelihood estimation, yield 3036 (3181) TADs with an average length of 879 kb (845 kb) in the human (mouse) genome, with the median TAD in both species containing 4 genes. In this figure, the box represents the 25-75th percentile range, and the whiskers represent 1-99th percentile range. **E, F**) Analysis of 665 human and 109 mouse CTCF ChIP-seq assays (ENCODE) revealed significant enrichment of CTCF binding at TAD boundaries and within TADs. Binding strength declined on either side of the predicted boundaries, supporting the accuracy of the scaffold. F) In individual tissues, CTCF enrichment at TAD boundaries was highly significant (Bonferroni-corrected $p < 10^{-50}$, one-sided binomial test). The box in the figure represents the 25-75th percentile, while the whiskers show the 1-99th percentile range. **G, H, I**) To assess the functional relevance of the TAD scaffold, we compared eQTL prevalence inside and outside TADs using single-tissue data from GTEx v8. Loci within the same TAD as a gene's transcription start site (TSS) were associated with significantly higher eQTL frequencies ($p = 0.0001$ for both upstream and downstream loci, one-sided Wilcoxon rank sum test). This trend held true for most tissues, demonstrating the scaffold's informativeness across diverse tissue types.

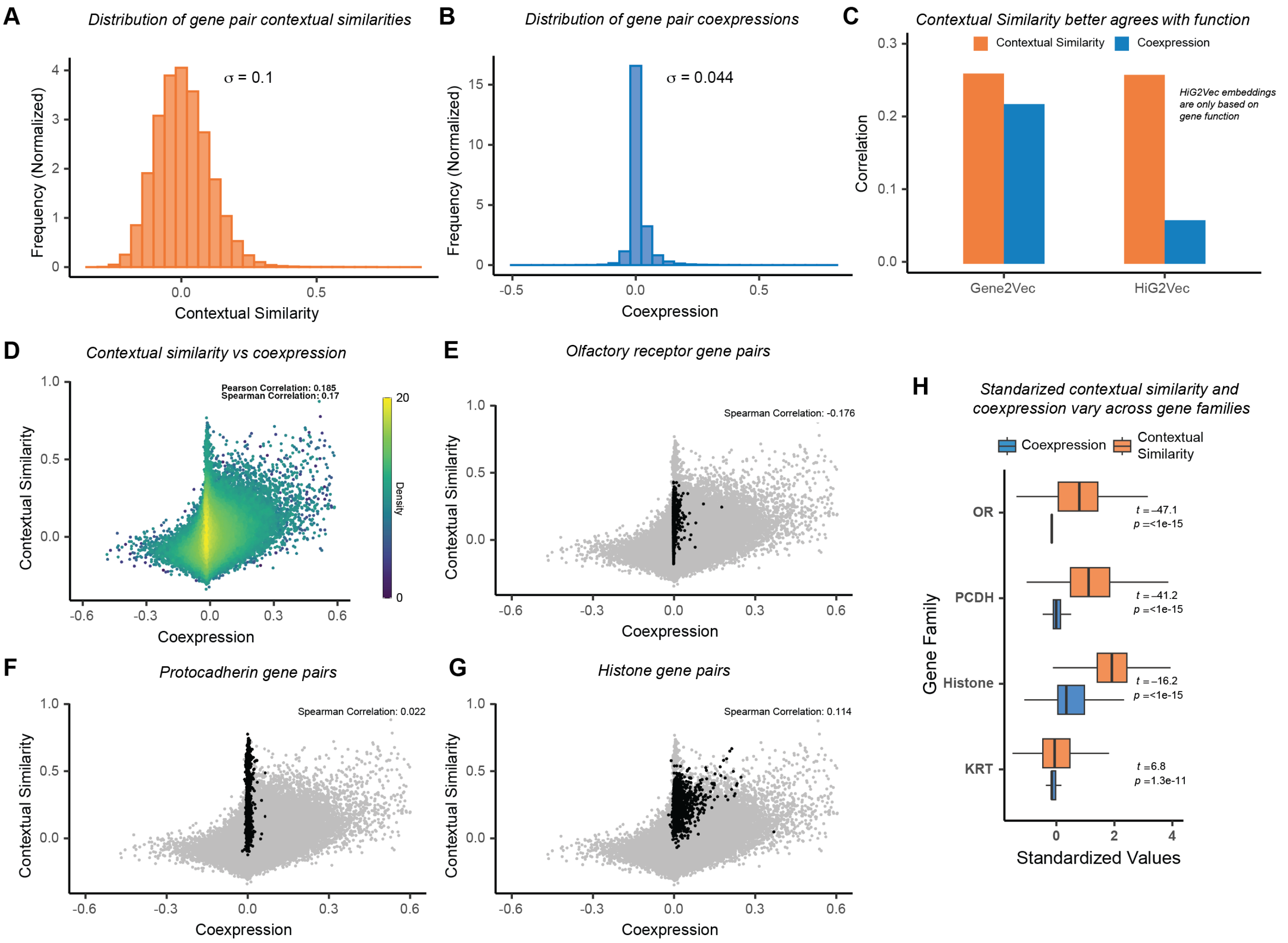
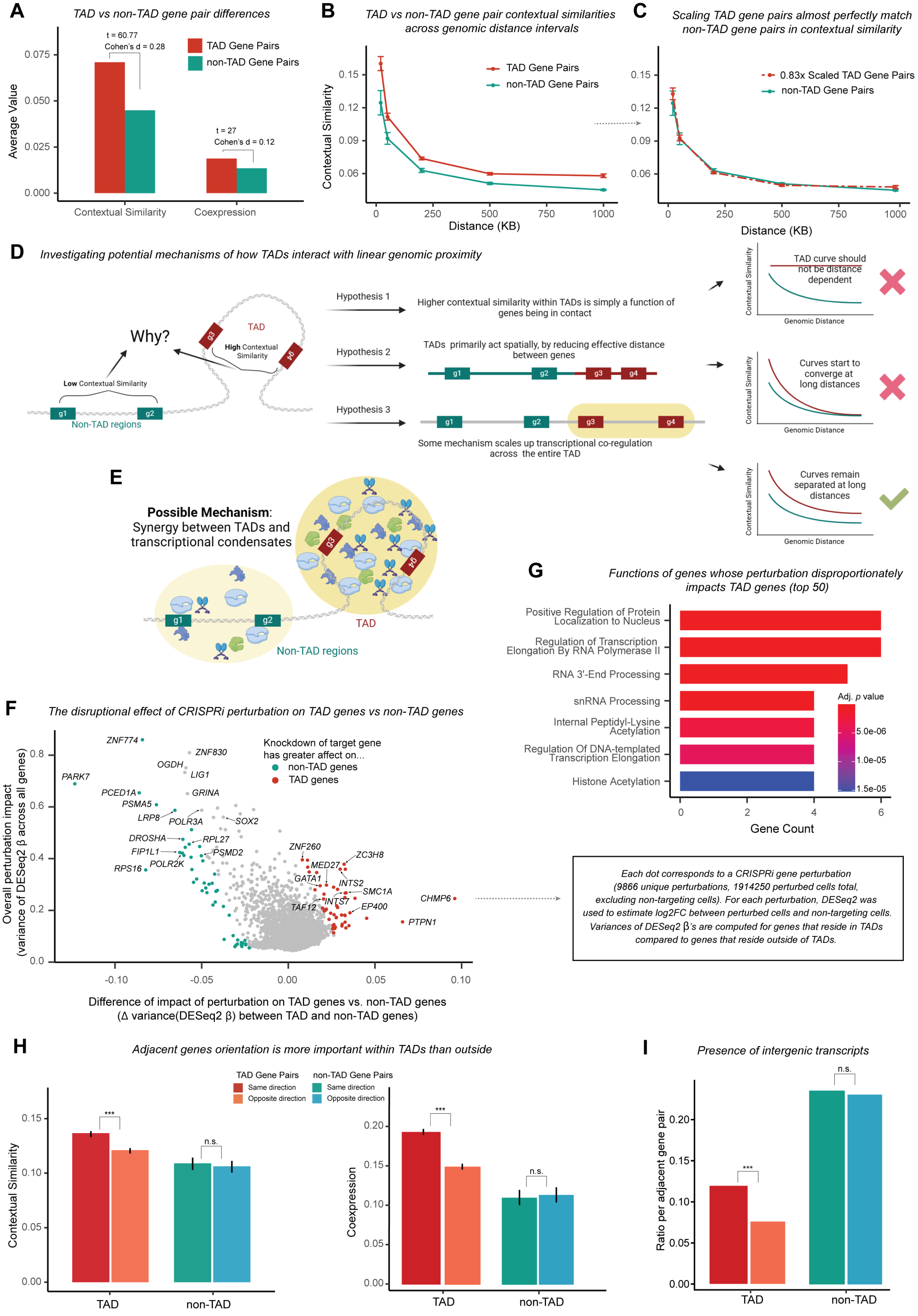


Figure 3: **Contextual similarity offers a more holistic perspective in evaluating gene pair transcriptional patterns. A, B**) Using Pearson correlations of pre-trained scGPT gene embeddings, we calculated contextual transcriptional similarities (CTS) for all gene pairs on the same chromosome within 2 megabases. Gene pair co-expression was similarly computed using tissue-specific single-cell data. CTS proved to be a more diverse and statistically powerful metric for evaluating gene interactions than co-expression. **C**) We compared CTS and co-expression against similarities derived from two popular gene embedding methods: Gene2Vec and HiG2Vec. Gene2Vec incorporates both gene expression and functional annotations, while HiG2Vec is based solely on functional annotations. CTS correlates more strongly than co-expression with both measures, particularly with the gene ontology-based HiG2Vec. This suggests CTS can effectively distill raw expression data to capture gene function in its embeddings.

D) CTS and co-expression show positive correlation across all gene pairs, confirming that both metrics capture shared aspects of gene expression patterns. **E, F, G**) Analysis of three gene families—olfactory receptors (ORs), protocadherins (PCDHs), and histones—reveals CTS's ability to capture functional relationships beyond co-expression. ORs and PCDHs show minimal co-expression by design (ORs are expressed one gene at a time, PCDHs stochastically), yet CTS identifies their shared functional roles. In contrast, histones, which are coordinately expressed, show high values in both metrics. **H**) Quantification across gene families shows CTS captures a broader range of relationships than co-expression, with the strongest signal in histones, followed by PCDHs and ORs, matching their known biological relationships. Additional analysis of keratin genes further supports this pattern (Fig. S2).



Each dot corresponds to a CRISPRi gene perturbation (9866 unique perturbations, 1914250 perturbed cells total, excluding non-targeting cells). For each perturbation, DESeq2 was used to estimate log2FC between perturbed cells and non-targeting cells. Variances of DESeq2 β's are computed for genes that reside in TADs compared to genes that reside outside of TADs.

Figure 4: **Comparing CTS in TAD and non-TAD gene pairs provides clues into TADs' role in transcription.**

A) Using the TAD Map, we classified gene pairs on the same chromosome within 2 megabases as TAD or non-TAD pairs based on whether they share a TAD. TAD gene pairs show significantly higher CTS and co-expression than non-TAD pairs, with CTS showing a more pronounced difference. **B, C**) To control for genomic distance effects, we binned gene pairs by distance and compared their average CTS. While TAD gene pairs consistently show higher CTS than non-TAD pairs across all distances (B), both groups show similar distance-dependent decline. Scaling TAD pair CTS by 0.83x matches non-TAD levels (C), revealing that TADs enhance CTS by 20.5% multiplicatively across all genomic distances. **D, E**) We considered three hypotheses for how TADs enhance gene co-regulation: (1) TADs might create distance-independent contacts between all genes, which would result in constant CTS across distances; (2) TADs might reduce effective distances between genes, which would lead to convergence of TAD and non-TAD CTS at long distances; or (3) TADs might work through a mechanism that multiplicatively enhances co-regulation across all distances. Our observations support the third hypothesis, suggesting a TAD-wide regulatory mechanism. Phase-separated transcriptional condensates, which can create specialized compartments spanning large genomic regions, represent a compelling candidate for such a mechanism, leading us to hypothesize that TADs work synergistically with condensates to create specialized regulatory environments. **F, G**) To identify factors regulating TAD-specific transcription, we analyzed a genome-wide Perturb-seq dataset [74]. For each CRISPRi perturbation, we used DESeq2 to estimate log2-fold-changes (β) between perturbed and control cells. We then identified perturbations that induced significantly different transcriptional variance in TAD versus non-TAD genes (F). Among the strongest hits were condensate-relevant genes including TAF12, INTS2, and INTS7. Gene Ontology analysis of TAD-specific regulators revealed enrichment for RNA-processing and transcriptional control pathways, consistent with condensate-mediated regulation (G, also see Fig. S3). **H**) A condensate-based mechanism of transcriptional regulation would predict easier co-transcription of same-orientation versus opposite-orientation adjacent genes. Indeed, within TADs, both CTS and co-expression are significantly higher for same-orientation gene pairs ($p < 10^{-11}$), while no such difference exists outside TADs. **I**) Similarly, intergenic transcripts, indicators of errorneous transcriptional read-through, are significantly more frequent between same-orientation genes within TADs ($p < 10^{-11}$), but show no orientation bias outside TADs.

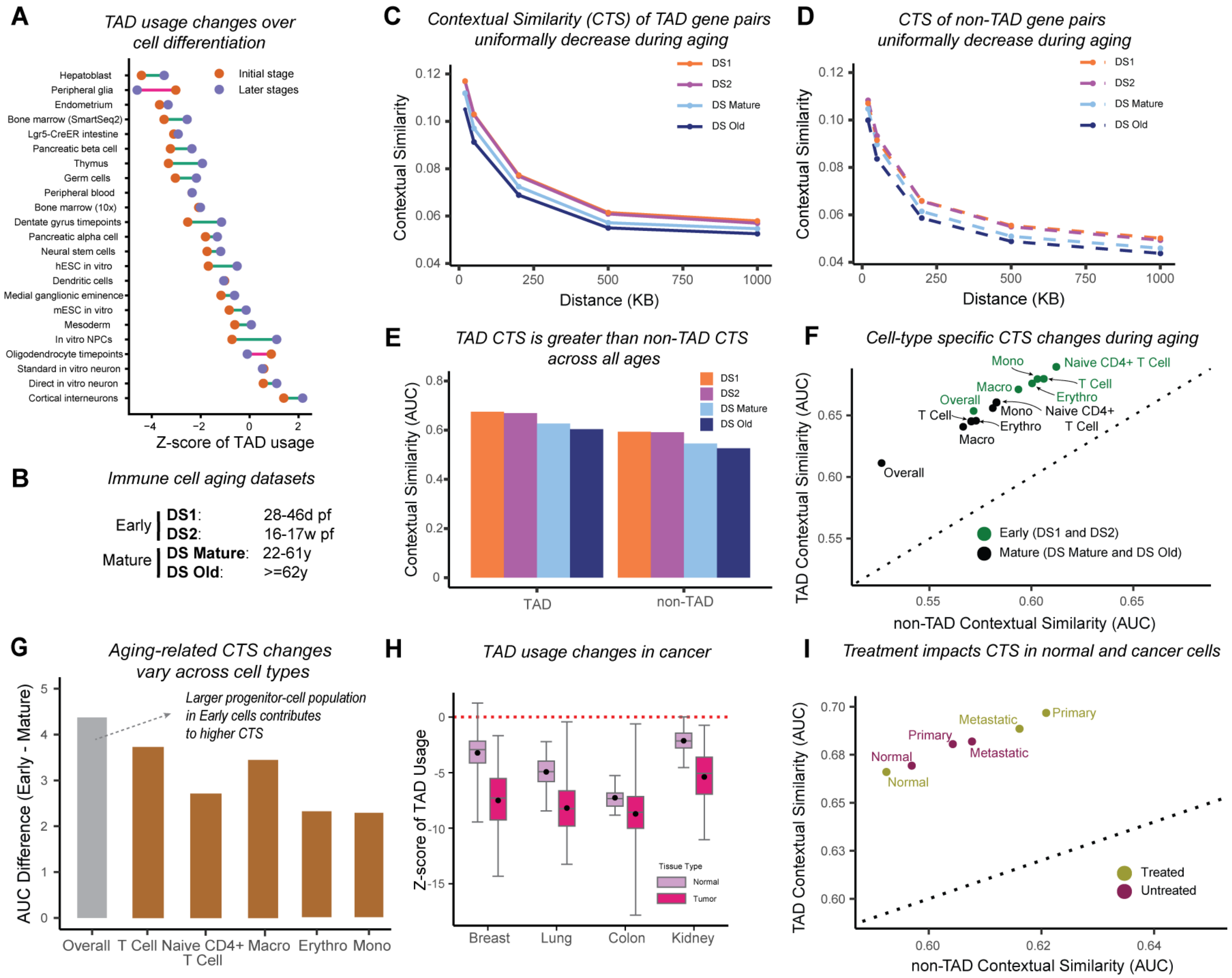


Figure 5: **TAD-mediated transcriptional organization changes systematically during development, aging, and cancer.** **A**) Analysis of 24 non-branching cell differentiation trajectories [95] reveals that early-stage cells consistently show higher TAD usage (more TADs with non-zero expression activity), indicating stronger transcriptional clustering (also see Fig. S4). **B**) To study aging effects, we analyzed immune cells from two large datasets [99, 100], spanning prenatal (DS1, DS2), mature adult (DS Mature), and older adult (DS Old) stages. **C, D, E**) Fine-tuning scGPT on each age group and recalculating CTS revealed systematic changes with age. Both TAD (C) and non-TAD (D) gene pairs show decreasing CTS with age across all genomic distances. Area under the curve (AUC) analysis confirms this decline while showing that TAD pairs maintain higher CTS than non-TAD pairs across all age groups (E).

F, G) Cell-type specific analysis reveals that aging-related chromatin changes occur not just through shifts in cell populations but also within specific cell types (particularly T cells and macrophages). Combining early (DS1, DS2) and mature (DS Mature, DS Old) datasets, we found that while early cells generally show higher CTS (F), the magnitude of age-related decline varies substantially across cell types (G), with T cells and macrophages showing the strongest changes. **H**) Analysis of The Cancer Genome Atlas (TCGA) data shows that, like early developmental stages, tumor cells exhibit increased TAD-based expression clustering across multiple cancer types. **I**) In colorectal cancer, 5-fluorouracil chemotherapy induces opposing changes in chromatin organization between normal and cancer cells: treated tumor cells show increased CTS while treated normal cells show decreased CTS, suggesting distinct adaptive responses to treatment stress.

Supplementary Materials

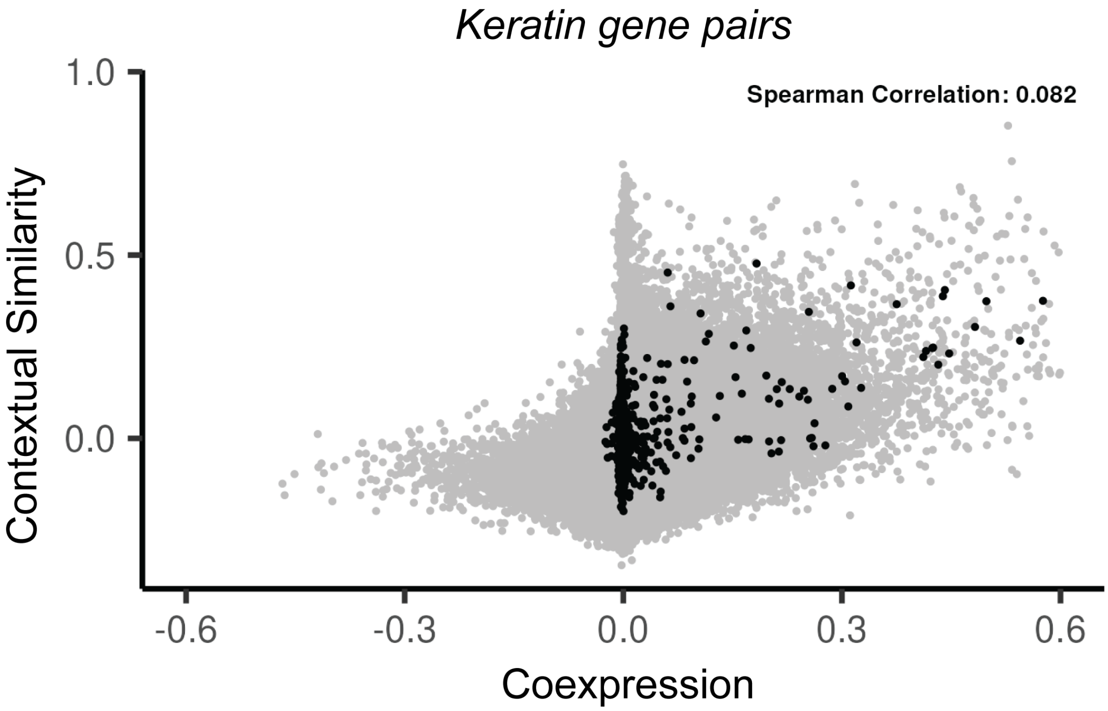


Figure S1: **Contextual similarities (CTS) vs co-expression for keratin genes pairs**: This figure complements Fig. 3E-G, where it visualizes CTS and co-expression for all gene pairs in the keratin family.

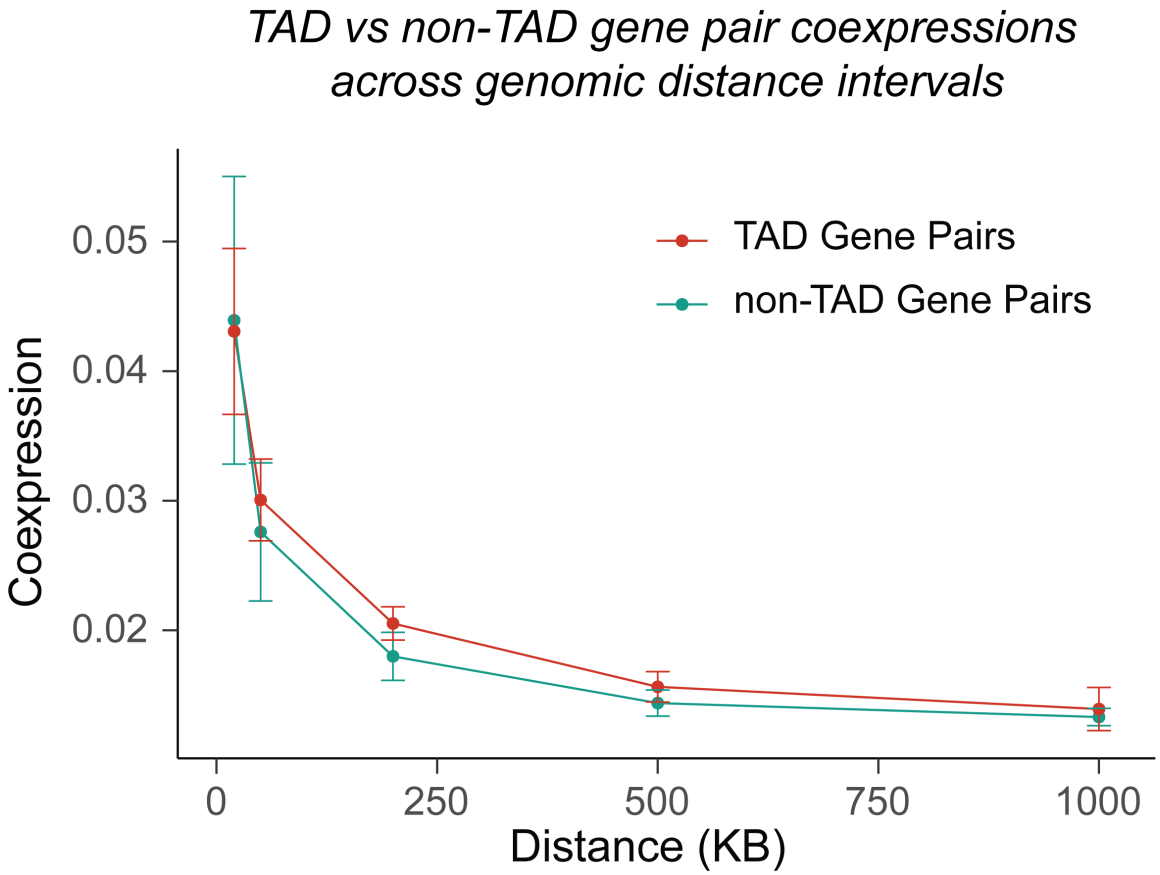


Figure S2: **Gene pair co-expression exhibits inconsistent patterns with respect to distance**: This figure contrasts the analysis with CTS in Fig. 4B, highlighting that average co-expression shows irregular and inconsistent trends between TAD and non-TAD gene pairs across distance intervals.

A

Perturbations causing the most widespread transcriptional disruption (top 100)

KNL1, CPSF2, LIG1, TYW3, SARS, MED9, HSPE1, POLR3A, TFR2, ASB8, RPL27, ZNF830, GRINA, C11ORF74, CARS, NAPA, PARK7, CAAP1, LRP8, TIAL1, SNRPD2, CDC26, PDCD11, RUVBL2, WDR70, DROSHA, PSMD2, FDPS, AQR, INTS2, ZNF260, INTS4, NUP155, CTPS1, DNAJC17, MED23, SEM1, PRPF4, MED21, ZNF774, OGDH, ZNF497, TUBGCP3, COX19, GSTCD, FAM216A, ACTG1, NR2E1, DPY30, ANAPC10, SMC3, RPS16, RPS15, SOX2, MED30, PPAT, HOXC6, ZBED2, PPP1R10, SEC62, CDT1, SEC13, ETV5, ITFG1, COQ3, NAT10, GART, PHF19, AARS, RBM22, FECH, TFCP2L1, DMAC1, ZC3H8, CD99L2, RPAP3, NPAT, MED12, UTP14A, LIN52, ZNF585A, SEPTIN2, STN1, UTP3, METTL4, MCOLN1, PSMA5, ANAPC4, KIAA1958, RNASEK, CAMKK2, NADSYN1, FIP1L1, PCED1A, DLAT, ANKRD52, CHML, POLR2K, NEPRO, RPL7A

B

Perturbations causing the most transcriptional disruption to TAD genes (top 50)

CHMP6, PTPN1, ZC3H8, ACTL6A, INTS2, HSPE1, KDM1A, INTS5, INTS7, EP400, SMC1A, PAM16, BAP1, DMAP1, EIF2B3, GINS4, GINS1, TSG101, NUDT21, GTF2B, GSTCD, GAB2, CCT7, ARPC1B, SRRT, INTS8, TCP1, TAF12, GATA1, MED27, NARS, ZNF260, DBR1, PSMD4, RBM22, PSMD13, LEO1, WDR12, CASP8AP2, COG3, CHMP3, MCM2, VPS28, NOL9, TPR, MARCH3, ATP5PD, AHCY, ETV5, AARS

C

Perturbations causing the most transcriptional disruption in non-TAD genes (top 50)

PARK7, RPS16, FIP1L1, PCED1A, TOMM5, BMS1, POLR2K, RNASEK, NGDN, PSMA5, CDC26, NAT10, ANAPC4, PDCD11, RPL27A, ZNF774, TRPT1, ITFG1, SEC62, ASB8, MKI67, UTP3, NOL6, ENO1, DDX5, RPS6, TYW3, PPP1R2, ZIC5, NEPRO, NDUFAF5, GP1BB, RPS8, ALG8, OXLD1, CZIB, ABCF2, DROSHA, HOXC6, STMP1, FECH, RPL27, TRMT2B, PSMD9, PNO1, RPS27A, C1orf43, ATP13A2, AP1S1, LRP8

D

Functions of genes whose perturbation disproportionately impacts non-TAD genes (top 50)

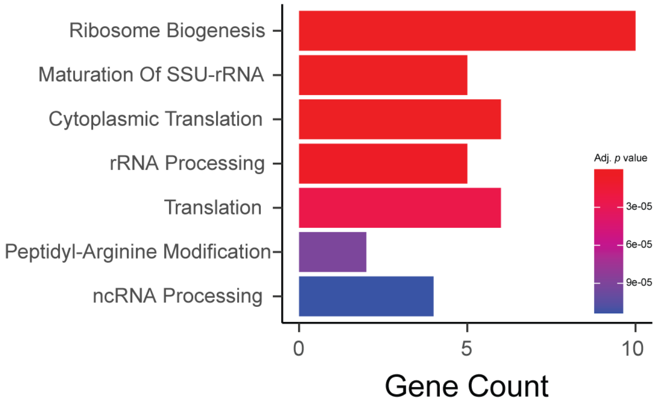


Figure S3: **List of specific genes whose perturbation induces transcriptional disruption across TAD and non-TAD regions:** **A**) This gene group highlights the genes whose perturbation causes the most widespread transcriptional disruption overall, without distinguishing between TAD and non-TAD regions. **B**) This gene group highlights genes whose knockdown disproportionally disrupts transcription within TADs. Corresponding Enrichr analysis of this gene set is shown in Fig. 4G. **C**) This gene group highlights genes whose knockdown leads to the greatest influence on transcriptional disruption specifically within TADs. **D**) This panel displays the Enrichr analysis of genes whose knockdown disproportionately disrupts transcription in non-TAD regions.

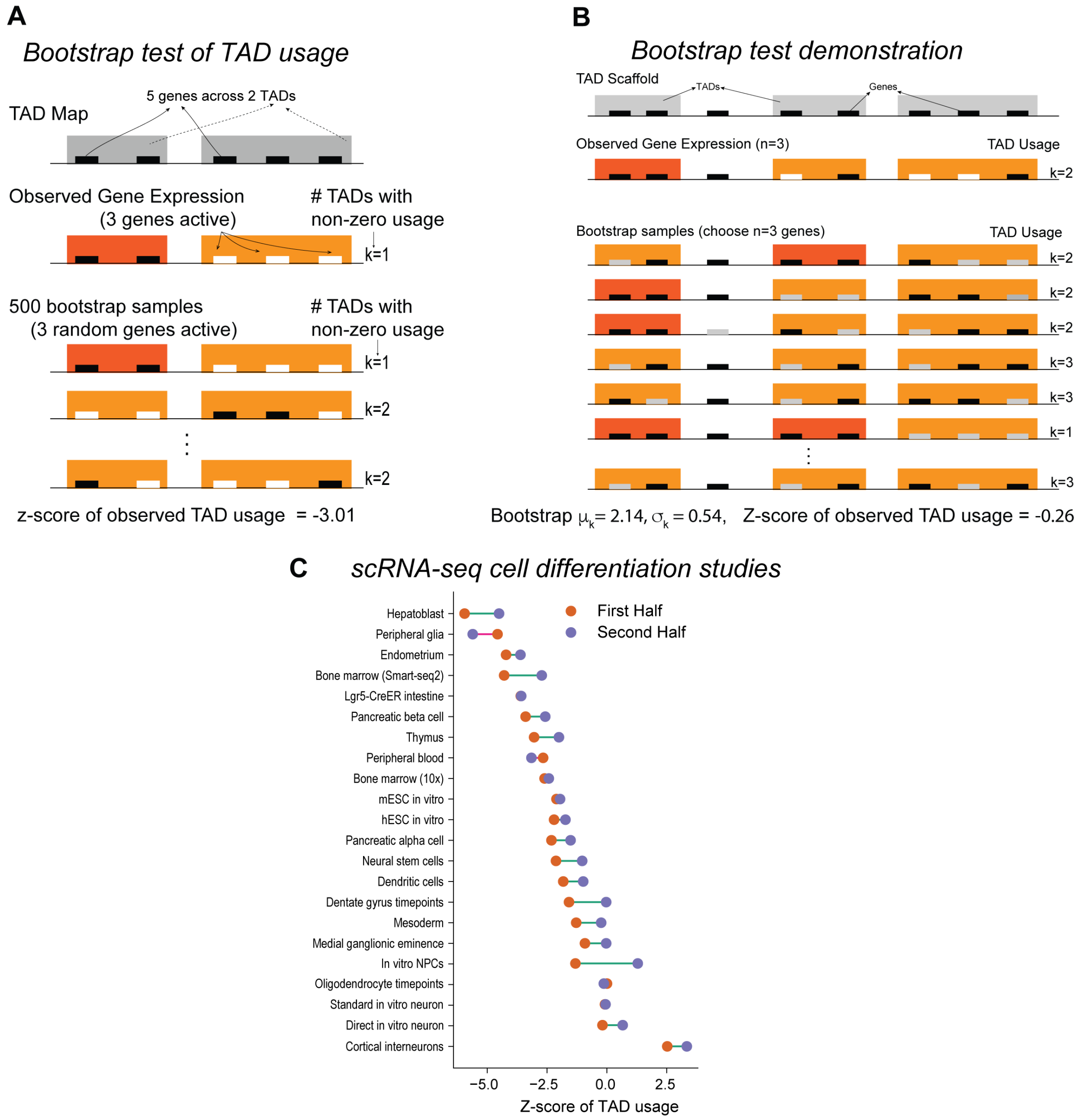


Figure S4: **Bootstrap test to assess gene clustering in TADs**: **A**) We define a TAD's *usage* as the number of transcriptionally active genes within it. Our bootstrap test assumes gene activity is independent of TADs under the null hypothesis. For any transcriptional profile, we randomize gene activity while preserving the total number of active genes, expressing TAD usage as a z-score relative to the expected distribution. This approach is robust to noise, sparsity, and platform effects, particularly in scRNA-seq data. Bootstrap z-scores can be aggregated across datasets or used to distinguish highly versus moderately active TADs. **B**) The bootstrap test is demonstrated here with a more complex artificial example. **C**) Gene clustering into TADs decreases during cell differentiation, as shown in Figure 5A. Early differentiation stages display stronger clustering of genes into TADs. In Figure 5A, cells were divided into "first stage vs. the rest" based on the CytoTrace Order variable [**65**]. Here, cells are partitioned into two groups with equal sizes. Regardless of partitioning, the clustering trend remains consistent.

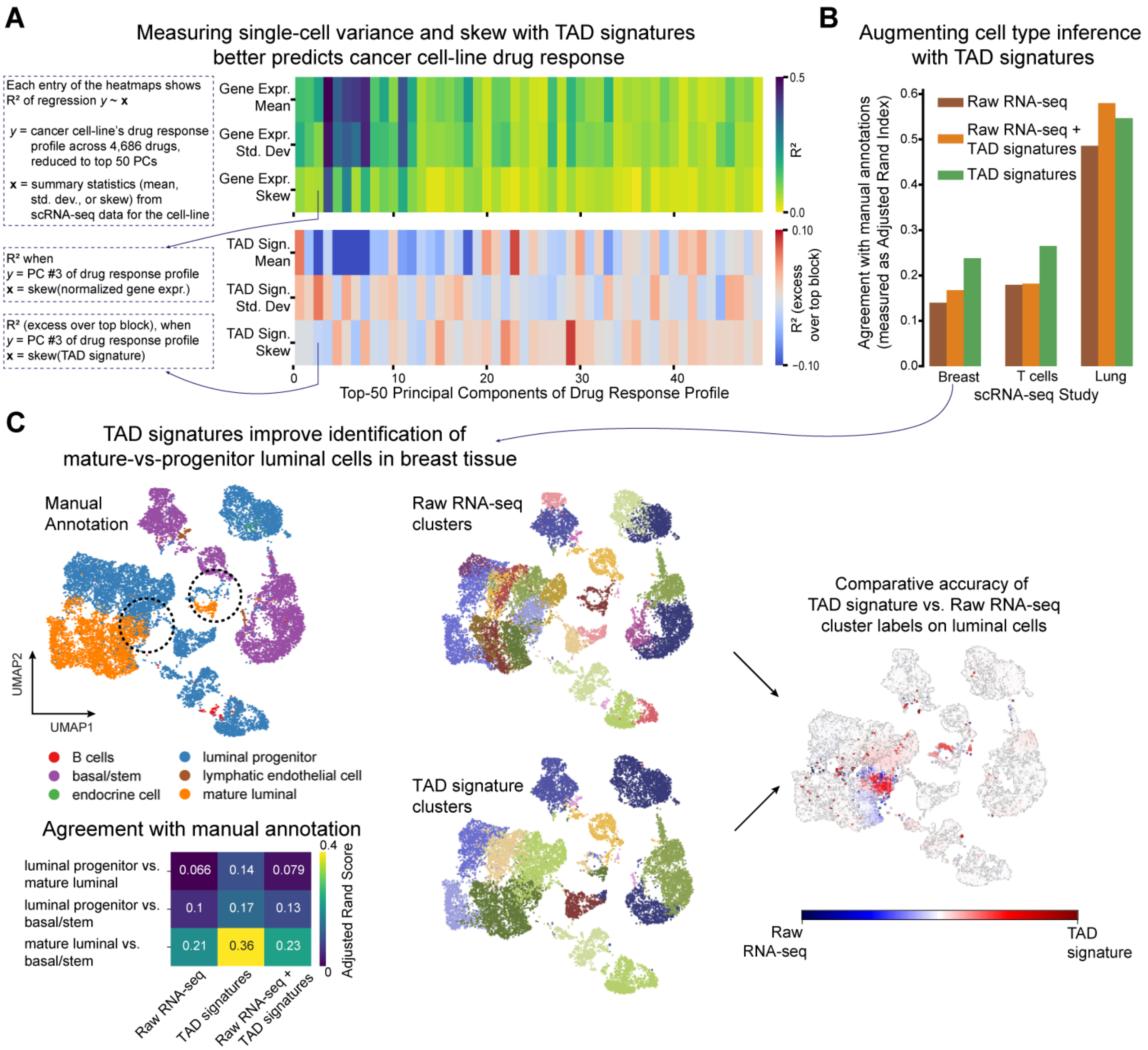


Figure S5: **TAD signatures augment standard scRNA-seq analysis**:**A**) TAD signatures reveal biologically meaningful single-cell heterogeneity. For 193 cancer cell lines, we analyzed drug response profiles from the PRISM study [110], reduced to 50 principal components. We computed the first three moments (mean, standard deviation, skew) of scRNA-seq data, either as log-normalized gene expression (top block) or log-odds of TAD signatures (bottom block). The second and third moments capture within-line heterogeneity. We used principal component regression ($k = 10$) to predict drug responses from per-gene or per-TAD statistics and measured predictive power as R^2. For the second and third moments, TAD signatures were significantly more predictive than gene expression (e.g., skew: $p = 0.0014$, one-sided Wilcoxon rank-sum test).**B**) TAD signatures improve automated cell-type inference. Compared to Leiden clusters from scRNA-seq alone, clusters using only TAD signatures or combining TAD and scRNA-seq better matched manual cell-type annotations in three scRNA-seq datasets. For breast and T cell data, TAD signatures alone outperformed combined clustering, likely due to RNA-seq noise. **C**) In breast tissue data, TAD signatures more accurately distinguished progenitor from mature luminal cells, highlighting their value in studying cell differentiation. Middle-column plots show Leiden clusters; RNA-seq clustering produced more clusters, aligning less with biological distinctions.