Foundation model reveals the shared organization of transcription and topologically associating domains

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, studies of TADs' function and their influence on transcription have been constrained by ambiguities in TAD boundary definitions and challenges in directly measuring their regulatory effects. We overcome these limitations by developing species-level consensus TAD maps for human and mouse using a bag-of-genes approach that exposes emergent regulatory structure. To quantify TAD-mediated gene relationships, we leverage a generative AI foundation model computed from 33 million transcriptomes to define a contextual similarity metric, revealing higher-order gene relationships elusive to co-expression analysis. We find that TADs are regions of elevated gene co-regulation, and our analytical framework directly leads to testable hypotheses about chromatin organization across cellular states. We 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. Enhancing our understanding of cellular plasticity in differentiation and disease, these findings suggest that chromatin organization may act 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 is fundamentally a physical problem, and evolution has found distinct solutions. Bacteria use operons to align co-regulated genes along the genome[1]; mammals fold DNA within the crowded nucleus, employing dynamic three-dimensional chromatin architecture that brings distant genes into contact[2,3]. Topologically associating domains (TADs), unveiled by diverse chromatin conformation capture technologies[4–8], are key components of this hierarchical architecture[2,9–11]—regions where DNA frequently contacts within the domain but rarely crosses it. These domains create distinct neighborhoods within the genome[12–14] in which the transcriptional machinery operates.

TADs present a paradox. Although their importance is clear, their function remains disputed. Do they coordinate or insulate gene activity? How do they relate to the genome's linear organization? Even these basic questions remain unresolved[15]. This uncertainty stems from technical challenges: TAD boundaries vary between cell types and individual cells; studies of CTCF, a protein central to TAD formation, have yielded inconsistent results[16–21]. The complexity of quantifying gene co-transcription exacerbates the challenge. 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]. This raises a central question: are there general principles governing how TAD organization relates to transcription, or do these relationships vary too unpredictably with context to generalize? To answer this question, we zoom out from individual cell-type analyses to the species level.

At the species level, a striking regularity emerges. Genes sharing a TAD show 20.5% higher transcriptional similarity than genes at equivalent distances outside TADs, an enhancement that persists multiplicatively across all distance scales and holds across diverse cell types. This pattern

had previously been difficult to detect. Precise TAD boundaries vary across cell types and whether TADs represent conserved regulatory units or are artifacts of chromatin conformation remains debated[24]. For measuring co-transcription, the standard metric, pairwise Pearson expression correlation, struggles with sparse single-cell data and misses genes that serve equivalent roles without being expressed together. Our species-level analysis, integrating Hi-C data over dozens of cell types with tens of millions of single-cell transcriptomes, exposes a consistent quantitative relationship: transcriptional coordination within TADs is systematically elevated in ways genomic proximity alone cannot explain.

We develop two conceptual advances that reveal this pattern. The *TAD Map* abstracts chromatin structure to the species level, representing each TAD simply as its contained genes, a ”bag of genes” model (Fig. 1A) inspired by set-based representations in language and protein modeling[25]. This abstraction tolerates boundary variation[26–30] and enables genome-wide analysis without requiring cell type–specific Hi-C data. *Contextual transcriptional similarity* (CTS) applies single-cell foundation models to a new purpose: detecting gene relationships that co-expression misses. Derived from scGPT, a foundation model trained on 33 million cells, CTS captures whether genes occupy similar roles within cellular programs. Consider genes as Lego blocks, each fitting into specific transcriptional roles defined by other genes. Many genes have unique shapes, but certain families, such as olfactory receptors and protocadherins, act like blocks of identical shape but different colors. Naive co-expression misses relationships among such genes; CTS exposes their shared context.

Beyond this core finding, our framework generates testable predictions about chromatin organization during development, aging, and cancer, and we evaluate these predictions using diverse single-cell datasets. We also introduce TAD signatures, bringing species-level chromatin structure to single-cell RNA-seq analysis and improving interpretation of individual transcriptomes.

Results

Consensus TAD Maps for human and mouse. The genome's three-dimensional organization presents a substantial consistency across cell types[26–32]. To quantify this consistency, we analyzed TAD definitions from seven human and four mouse cell types using the TADKB database[33] and the Directionality Index technique[26]. The TADKB database provided harmonized TAD calls across diverse data sources (**Methods**). This analysis revealed high conservation (**Fig. 2A**). 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%. We extended this analysis using data from the 4D Nucleome project[34], incorporating ten human and six mouse additional cell types. Analyzing with the Directionality Index as well as the Insulation Score methods for TAD-calling[35], we found broadly consistent results (**Fig. S1**).

While the boundaries of TADs vary, we reason that their gene memberships may be more stable. We therefore represent each TAD simply as its set of protein-coding genes, a "bag-of-genes" abstraction we call the TAD Map. This representation tolerates minor boundary shifts that don't alter gene membership while retaining the features needed for statistical analysis of gene regulation.

The bag-of-genes approach reveals strong 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**). We extended the analysis to gene triplets, which yielded similar findings (**Fig. S2**).

Consensus TAD "scaffolds" organize the genome into gene groupings preserved across cell types. 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**). Together, human and mouse TADs cover approximately 86% and 92% of their respective genomes. 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. CTCF binding validates the TAD Map across hundreds of cell types lacking Hi-C data. CTCF enrichment at TAD boundaries[36,37] (and sometimes inside them[3,38]) is well-established, but prior studies were limited to cell types with available Hi-C. We sourced CTCF ChIP-seq data from the ENCODE database[39], 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.

Genetic variants within TADs show elevated regulatory effects on nearby genes. We analyzed expression quantitative trait loci (eQTLs), spanning all 49 tissues covered in GTEx v8[40]. 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 reported previously[41,42], 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, further validate our TAD Map, indicating that TAD membership influences the regulatory reach of genetic variants.

Single-cell foundation models characterize the transcriptional context of genes. Co-expression analysis misses a fundamental aspect of gene relationships: functional substitutability. In olfactory neurons, only one of several hundred receptors is typically expressed, yet all serve equivalent roles in specifying neuronal identity[22]. Protocadherin genes show similar patterns[43]. These genes often reside in the same TADs, suggesting TAD-based gene groupings imply functional relationships that traditional correlation-based measures are unable to fully capture.

Single-cell foundation models (scFMs) offer an alternative. These models integrate tens of millions of transcriptomes to learn gene representations that reflect broader functional context[44–46]. We leverage scGPT, trained on 33 million human cells, to define "Contextual Transcriptional Similarity" (CTS): the Pearson correlation between two genes' embedding vectors. Unlike co-expression, which requires observing genes active together in the same cells, CTS characterizes a gene's role through its relationships with all other genes across diverse cellular states. Two genes can thus show high CTS despite rarely co-expressing—capturing substitutability that correlation-based metrics miss.

CTS captures greater variation in gene relationships than co-expression. Analyzing 418,425 protein-coding gene pairs (less than 2 megabases apart on the same chromosome) across four atlas-scale scRNA-seq datasets from CELLxGENE[47–51], we found that CTS shows significantly higher variation than co-expression (variances of 0.010 vs. 0.002 after Fisher's transformation, p

$< 10^{-15}$, one-sided F test; **Fig. 3A-B**). The compressed range of co-expression values likely reflects single-cell data sparsity, where many transcripts per cell are reported as zero.

CTS correlates more strongly with functional annotations than co-expression does. While CTS and co-expression show broad agreement (**Fig. 3D**; Pearson correlation 0.185, $p < 10^{-15}$), we compared both against independent representations of gene function. Gene2vec[52] generates embeddings from transcriptional data from Gene Expression Omnibus, supplemented with MSigDB[53] pathway annotations; HiG2vec[54] derives purely from Gene Ontology hierarchies in hyperbolic space. CTS outperforms co-expression on both, but its advantage is substantially greater with HiG2vec (correlation 0.26 vs. 0.06) than Gene2vec (0.262 vs. 0.220) (**Fig. 3C**). Permutation tests confirm that correlations with CTS are significant (p = 0.0001, **Methods**). The comparably stronger agreement with a purely functional representation, HiG2vec, indicates CTS captures gene relationships beyond expression correlation.

The power of CTS is particularly clear in gene families whose members concentrate within a few TADs. In olfactory receptors and protocadherins, CTS substantially exceeds co-expression (olfactory receptors: 0.099 vs -0.002; protocadherins: 0.196 vs 0.001 after median centering; **Fig. 3E-G, Methods**). Histone genes, coordinately regulated during DNA replication[55], show even higher CTS (0.251 vs 0.021) with modestly better CTS–co-expression agreement (0.114). All differences are highly significant ($p < 10^{-15}$; **Fig. 3H**); keratin genes, by contrast, show strong concordance (**Fig. S3**). These patterns suggest TADs may group genes by shared transcriptional context rather than co-expression, a hypothesis we test directly in the next section.

TADs serve as loci of enriched transcriptional context. Gene pairs sharing a TAD show elevated transcriptional similarity. We defined TAD and non-TAD gene pairs as above and computed the CTS and co-expression within each set of pairs (**Fig. 4A**). Both measures were significantly

higher for TAD pairs ($p < 10^{-15}$, one-sided t-test), but the effect was markedly stronger for CTS (Cohen's d = 0.28 vs 0.12), confirming that CTS better captures relationships associated with TAD co-membership.

The TAD effect persists across genomic distance scales, even after accounting for the known tendency of genomically-proximal genes to share functional relationships[56,57]. We stratified our analysis by distance, grouping gene pairs by their pairwise distance: 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. To ensure robustness and mitigate scFM-specific bias, we repeated the CTS analysis using another scFM, Transcriptformer[58], and observed the same pattern (**Fig. S4**). Our findings contrast with Long et al.[59], who reported no TAD-specific effects beyond proximity using co-expression analysis. Indeed, when we repeated our distance-stratified analysis using co-expression (**Fig. S5**), the TAD-specific signal was substantially weaker, suggesting their conclusion reflected the limitations of the co-expression measure. We further extended this co-expression analysis to 700 additional single-cell transcriptomic datasets from scBaseCount[60], spanning 18 million cells across 40 tissues, and found similar results (**Fig. S6**).

The structure of CTS's distance dependence constrains possible physical mechanisms (**Fig. 4D-E**). If TADs created isolated compartments overriding genomic proximity effects, CTS within TADs would be distance-independent. It is not. If TADs acted solely through chromatin compaction (i.e., they uniformly reduce the effective genomic distance), at very long genomic distances TAD and non-TAD CTS would converge. They do not. Instead, we observe a third pattern: TAD membership multiplicatively enhances CTS while preserving distance dependencies. Non-TAD pairs consistently showed 83% of the CTS observed in TAD pairs across all intervals, corresponding to a multiplicative

enhancement 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**).

What physical process might produce this multiplicative enhancement? One candidate is transcriptional condensates (TCs): dynamic, membrane-less compartments formed by liquid-liquid phase separation that concentrate transcriptional machinery[61–64]. Phase-separated compartments naturally create multiplicative effects on local concentrations[65–68], consistent with our observations. Individual TCs may be too small to span entire TADs; if this model is correct, either larger "TAD condensates" or the collective action of multiple smaller TCs could contribute to the observed effects. We emphasize that this remains a hypothesis requiring direct experimental validation. Nonetheless, if TADs and condensates do interact systematically, indirect signatures should be detectable in existing genomic data, as we investigated next.

Systematic regulation in TAD regions differs mechanistically from that of non-TAD regions. If TADs are associated with specialized transcriptional environments, this should manifest in diverse transcriptional readouts. Leveraging publicly available data, we investigated three specific predictions: (1) perturbing regulatory genes should differentially affect transcription within versus outside TADs; (2) gene orientation should constrain co-regulation more strongly within TADs, where shared machinery may be spatially concentrated; and (3) transcriptional read-through between adjacent genes should similarly depend on orientation within TADs.

For the first analysis, we examined genome-wide Perturb-seq data from Replogle et al.[69], where 9,866 genes were knocked down via CRISPRi with scRNA-seq readout. For each perturbation target, we quantified transcriptional disruption separately for genes inside and outside TADs (**Fig. 4F, Methods**). Extending the analysis to other studies—RPE1, HepG2, K562 essential and Jurkat datasets[70]— yielded broadly consistent patterns across cell types (**Fig. S7, Methods**).

Our analysis framework recapitulates known regulators of transcription and TAD organization. Perturbations causing the most widespread disruption, regardless of genomic location, were enriched for core transcriptional processes such as RNA Polymerase II elongation and initiation (GO:0034243, Adjusted p = 2.39×10^{-6}; GO:0060261, Adjusted p = 6.09×10^{-6})[71]. Mediator complex genes (MED12, MED30, MED21, MED9), which play a central role in transcriptional regulation[72,73], ranked in the 99th percentile (**Fig. S8A**)[73,74]. Perturbations of known TAD-organizing factors showed the expected TAD-specific effects. Cohesin subunits SMC1 and RAD21, which mediate loop extrusion[75–77], produced strong differential disruption between TAD and non-TAD regions (99th and 97th percentiles, respectively). SMC3 perturbation affected both regions strongly, consistent with its genome-wide role. CTCF perturbation yielded broadly TAD-specific effects (75th percentile), in line with its insulator function at boundaries[2,36]. These results confirm the framework's sensitivity to both global transcriptional regulators and TAD-specific organizers.

Among perturbations preferentially affecting TAD regions, we found enrichment for histone acetylation factors (GO:0016573, Adjusted $p = 4.074 \times 10^{-4}$), including TAF12, ACTL6A, EP400, and DMAP1 (**Fig. 4G**; full list in **Fig. S8B**). These findings align with prior work: loss of H4 acetylation increases long-range contacts beyond TADs[78], and TAFs contribute to phase-separated transcriptional hubs[79]. A second category, the Integrator complex (INTS2, INTS5, INTS7, INTS8)[80], also showed TAD-specific enrichment, with roles in snRNA and ncRNA processing (GO:0016180, Adjusted $p = 1.713 \times 10^{-5}$; GO:0034470, Adjusted $p = 1.614 \times 10^{-3}$).

Perturbations disproportionately affecting non-TAD regions were enriched for ribosomal processes (**Fig. S8C**), 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 RPS and RPL subunits among the top hits (**Fig. S8D**). We note that non-TAD genes are fewer and already enriched for ribosomal function, which may inflate these estimates. Nonetheless, the pattern suggests a

possible division of labor, with nucleoli, membrane-less condensates enriched near low-gene-density heterochromatin[81,82], preferentially organizing ribosomal genes outside TADs.

Gene orientation constrains transcriptional relationships within TADs. If adjacent genes share transcriptional machinery more readily within TADs, orientation should matter: same-strand pairs would have easier access to shared machinery than oppositely oriented pairs. We tested this using CTS. Within TADs, same-orientation pairs indeed exhibited significantly higher contextual similarity (0.137 vs. 0.121, $p = 4.5 \times 10^{-14}$); outside TADs, orientation had no effect (0.109 vs. 0.106, $p = 0.508$) (**Fig. 4H, Methods**). Co-expression analysis yielded consistent results: higher correlation for same-orientation pairs within TADs (0.19 vs. 0.15, $p < 10^{-15}$) but no difference outside.

In our third analysis, transcriptional read-through—when intergenic regions are erroneously transcribed—shows the same pattern. Analyzing 11,417 intergenic transcripts from Agostini et al.[83] (**Methods**), we found that within TADs, same-orientation gene pairs showed significantly more read-through than opposite-orientation pairs (11.9% vs 7.6%, $p < 10^{-15}$, one-sided binomial test; **Fig. 4I**). Outside TADs, orientation had no effect, though overall read-through frequencies were higher, suggesting elevated transcriptional noise. These orientation-dependent patterns within TADs, but not outside, are consistent with models where TADs facilitate shared access to transcriptional machinery.

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[84,85]. Prioritized hypotheses can then be experimentally tested.

Cell differentiation: A central question in developmental biology is how cells modulate gene expression while balancing stability and plasticity. Recent work by Pollex et al.[86] shows that

developmental shifts from permissive early enhancer–promoter contacts to more constrained later topologies likely reflect corresponding changes in TAD-level transcriptional organization[12]. We analyzed diverse differentiation trajectories and observed a clear relationship between developmental stage and TAD-mediated gene organization. Using developmental scRNA-seq transcriptomes curated by Gulati et al.[87], 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) and designed a bootstrap-based "TAD usage" score to quantify this effect (**Fig. S9A-B, Methods**). This pattern persisted across stage definitions and tissues (**Fig. S9C**), suggesting that TAD organization is especially prominent during developmental plasticity.

Aging: Motivated by this developmental pattern, we asked whether TAD-level organization also shifts with age. We used the immune system as a model, where aging alters lineage distributions and chromatin states[88–90], to test whether these changes reflect solely shifts in cell type proportions or also intrinsic within–cell-type changes in chromatin regulation. We analyzed scRNA-seq datasets from human immune cells spanning prenatal to late adult stages[91,92] and grouped samples into four age categories (DS1: 28–46 days post-fertilization; DS2: 16–17 weeks post-fertilization; DSmature: 22–61 years; DSold: over 61 years) (**Fig. 5B**) to assess whether CTS between gene pairs changes systematically with age.

As cells age, we observed a decline in the proximal transcriptional programs and machinery coordinating gene pairs using CTS area-under-curve metric. For each curated dataset (DS1, DS2, DSmature, DSold), we fine-tuned the whole-human scGPT model, extracted updated gene embeddings, and computed CTS for TAD and non-TAD gene pairs as before (**Fig. 5C–D, Methods**). CTS was highest 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%). This systematic decline suggests that aging progressively modulates the spatial constraints on transcriptional machinery.

To distinguish between population-level and cell-intrinsic changes, we fine-tuned the model separately on subsets of scRNA-seq data and 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**). Our cell-type–specific analysis suggests that the age-related decline in CTS arises from both shifts in cell type proportions and intrinsic within–cell-type changes, with the strongest effects in T cells and macrophages, highlighting these populations as key targets for studying how aging modulates chromatin organization.

Cancer: If TAD-mediated transcriptional clustering covaries with developmental plasticity, cancer cells should also exhibit the elevated TAD usage typical of progenitor-like states, as they often dedifferentiate or fail to fully differentiate. Indeed, 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[93] (**Fig. 5H**). This difference in TAD usage patterns is robust to adjustments for the higher number of expressed genes in tumor cells (**Methods**).

We also examined how cancer treatment might modulate cell plasticity during disease progression by applying CTS analysis to a colorectal cancer dataset that tracks plasticity through metastasis[94]. 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, consistent with the TCGA analysis above. However, After 5-fluorouracil–based chemotherapy, tumor cells (primary and metastatic)

showed increased CTS relative to untreated tumors, whereas normal cells exhibited a pronounced CTS decrease (**Fig. 5I**).

The divergence of post-treatment CTS patterns between normal and cancer cells align with well-established biological responses to chemotherapy. It 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. 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[95,96]. The elevated CTS in treated cancer populations could reflect this enrichment of stem-like cells, which maintain plastic transcriptional programs even under treatment pressure. Although CTS–plasticity correlations fit known biology, causal links between TAD organization and treatment response will require further investigation. If validated, our findings would address a major gap in clinical diagnostics, with a 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 have had a transformative impact on biological research, challenges such as data sparsity, uneven gene coverage ("dropouts"), and platform effects remain[97]. To complement existing dimensionality reduction techniques, we introduce TAD signatures, an epigenetically-informed scRNA-seq representation that leverages TAD Map gene groupings to mitigate technical noise while preserving biological signal, and that can be inferred for any human or mouse scRNA-seq dataset without requiring Hi-C data (**Methods**). Across 193 cancer cell lines, drug response profiles (4,686 compounds from the PRISM screen) correlate with TAD signature–based measures of higher-order transcriptional heterogeneity, indicating that TAD signatures capture phenotypically relevant variability beyond what is accessible through gene-level expression alone (**Fig. S10A**). In addition,

TAD signatures improve automated cell type inference by enhancing clustering accuracy (**Fig. S10B**) and enabling reliable discrimination of closely related cell subtypes with subtle transcriptional differences that challenge gene-based approaches (**Fig. S10C**).

Discussion

This work seeks to reframe how chromatin structure relates to gene expression. We characterize TADs not as deterministic regulators where membership dictates co-expression but as probabilistic biases: genomic contexts where transcriptional coordination is elevated. Any individual TAD may deviate from this pattern, but across the genome—and at all distance scales—genes sharing a TAD show ~20% higher contextual similarity than genes at equivalent distances outside TADs. The gradual modulation we observe during aging and in cancer is consistent with this probabilistic view: chromatin organization biases transcription without absolutely constraining it.

Two conceptual advances enable this perspective. The TAD Map abstracts chromatin structure to a species-level "bag of genes" representation, tolerating boundary variation while enabling genome-wide analysis without cell type–specific Hi-C. Contextual transcriptional similarity (CTS) leverages single-cell foundation models to capture gene relationships that co-expression misses—particularly functional substitutability among genes that rarely co-express.

CTS exposes a class of gene relationships invisible to traditional analysis. Long et al., limited to co-expression analysis in only two cell types[59], could not disentangle TAD-specific effects from genomic proximity. Leveraging single-cell foundation models (scFMs), CTS reveals systematic TAD-mediated enhancement that holds across distance scales and cell types. More broadly, scFMs excel at capturing broad-but-subtle biological patterns when combined with prior knowledge of genomic organization. This is particularly powerful in small-data settings. While individual

age-stratified datasets were too small for robust correlation estimation, fine-tuning scGPT on each revealed clear distinctions. Our work demonstrates that scFM representations can be repurposed beyond their original framing. While the original scGPT study interpreted gene embedding similarity as simply recapitulating gene regulatory networks[45], our CTS-based view reveals new patterns the network interpretation misses.

Our observations are consistent with a physical model in which TADs interact with condensates, phase-separated compartments that concentrate transcriptional machinery. Three independent lines of evidence support this possibility: perturbation of condensate-associated genes disproportionately affects transcription within TADs; same-orientation gene pairs show enhanced contextual similarity specifically within TADs; and transcriptional read-through between adjacent genes follows the same orientation-dependent pattern within TADs but not outside them. We note these patterns are correlational. Establishing causality (and its direction) will require direct experimental manipulation, e.g., perturbing condensate formation within defined TADs or imaging studies to localize phase separation in vivo[98,99].

Our framework opens new avenues for studying cellular transitions. If TAD-associated organization reflects regulatory coordination, it should track with cellular plasticity. It does. Cells in early differentiation show stronger TAD-level coordination; this organization declines systematically with aging, even within specific cell types like macrophages. Cancer cells show elevated TAD usage, with divergent responses to treatment between tumor and normal populations. Establishing causality will require direct experimental manipulation, but the key advantage of our approach is hypothesis-generation: it reveals these large-scale patterns across diverse conditions without requiring Hi-C data for each cellular state.

Our approach has limitations that suggest future directions. There remains active debate regarding the conservation of TADs across tissues and cell types[24]. While the bag-of-genes model

mitigates boundary variability, it does not resolve this broader controversy; our findings should be interpreted within this context. The TAD Map enables species-wide analysis but sacrifices resolution—it cannot capture fine-grained structural variations between cell types, subtle boundary shifts, or chromatin sub-structure such as sub-TADs. Targeted functional studies could probe how specific genomic contexts or cell states deviate from the general patterns we report. Our focus on protein-coding genes leaves open questions about non-coding RNA regulation within TADs. Our treatment of boundary-spanning genes (currently assigned to both flanking TADs) is conservative but deserves dedicated investigation; these genes may experience distinct regulatory environments.

Our systematic framework can serve as the base for more refined models of chromatin–transcription synergy. Computational methods for inferring cell-specific 3D genome structure[90,100–106] and multimodal sequencing approaches[107–112] could integrate with scFM representations fine-tuned on individual cell types, allowing hierarchical chromatin organization to inform CTS while accommodating structural variation. Such a hierarchical scheme may also help identify alternative physical mechanisms, such as the nested enhancer network described by Lin et al.[113]. Other epigenetic modalities, such as histone modifications, dynamically interact with chromatin topology[114,115] and could further refine the framework.

By bridging genome organization with foundation model-based representation learning, this work enables a systematic investigation of chromatin's role in cellular transitions without requiring structural data for each state. The TAD Map and CTS framework demonstrates how integrating prior knowledge with foundation models can reveal patterns obscured by cell type–specific analyses. Lastly, our work advocates for reimagining transcriptional regulation as a convolution of probabilistic mechanisms—chromatin 3D structure, chromatin accessibility, transcriptional condensates—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[33], 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.[116] (GEO accession GSE96107).

Expanded Cell Type Coverage. We expanded our analysis 2.5-fold by increasing coverage from 11 to 27 cell types, resulting in 17 human and 10 mouse cell types drawn from diverse tissues, developmental stages, and both in vivo and in vitro systems. These additional datasets were sourced from the 4D Nucleome project[34]. For human, the updated compendium of cell types includes GM12878 (B lymphoblastoid), K562 (erythroleukemia), KBM7 (haploid leukemia), Hap1 (haploid myeloid), bone-marrow lymphocytes (primary lymphocytes), HMEC (mammary epithelial), NHEK (epidermal keratinocyte), IMR90 (fetal lung fibroblast), HFF-hTERT (immortalized foreskin fibroblast), HUVEC (umbilical vein endothelial), HepG2 (hepatocellular carcinoma), CyT49-derived gut-tube endoderm, H1-derived pancreatic progenitor, olfactory receptor cells, left-atrium cardiomyocytes, HeLa (cervical adenocarcinoma), and H9-derived cardiomyocytes. For mouse, the expanded set comprises CH12-LX (B-lymphocyte lymphoma), ES (embryonic stem cell), NPC (neural progenitor cell), CN (cortical neuron), cardiac muscle, Sertoli (testicular support) cells, trophoblast (placental lineage), granulosa (ovarian somatic) cells, double-negative alpha-beta

thymocytes (immature T-cell precursors), and olfactory receptor cells.

Reconstruction of TAD Maps. We regenerated both human and mouse TAD Maps using these expanded panels of cell types. For all newly added cell types, TADs were called independently using the Directionality Index (DI) method[26] and the Insulation Score (IS) method[35]. The resulting architectures were used as inputs to construct revised species-specific TAD Maps. Broadly, these new maps remain in close agreement with our originally reported TAD Maps, indicating that increased cellular diversity and variation in TAD-calling methodology do not materially alter the large-scale TAD architecture.

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[117], 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:

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

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[39], 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[40], 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. To assess whether the association between HiG2Vec and CTS could be explained by chance, we performed a permutation test in which CTS values were shuffled 10,000 times and correlations were recalculated. The p value was then computed using the observed correlation. 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.[69], 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[118,119] 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.

Expanded Perturb-seq analysis across multiple cell lines. To broaden our analysis of perturbation-induced transcriptional effects, we incorporated additional Perturb-seq datasets from HepG2 and Jurkat cell lines[70], as well as from K562 and RPE1 cell lines[69]. We restricted all analyses to the subset of perturbations and genes that were shared across the four datasets, allowing a unified comparison across cell lines. Using the same analytical framework described above, we quantified, for each perturbation and in each cell line, the relative disruption of transcription in genes residing within TADs versus those outside TADs. We then identified perturbations that showed consistent TAD-specific or non–TAD-specific transcriptional effects and focused on those that recurred in at least three of the four cell lines.

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.[87]. 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.[91] and the single-cell transcriptomic atlas[92], 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[120].

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[93]. 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[121–123], there are diverse epigenetic mechanisms underpinning tumorigenesis[124–126], 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.[94], 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:

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

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[127], limiting ourselves to cell lines for which drug response data from the PRISM study was also available[128]; 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[129] datasets from single-cell RNA-seq studies of the human lung (10x sub-study; European Genome-Phenome Archive accession EGAS00001004344; 65,662 cells;[130]), T-cells (GEO accession GSE126030; 51,876 cells;[131]), and breast epithelial cells (GEO accession GSE164898; 31,696 cells;[132]). 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[133] 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 (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] Eres, I. E. & Gilad, Y. A TAD Skeptic: Is 3D Genome Topology Conserved? *Trends Genet* **37**, 216–223 (2021).

[25] 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).

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

[27] 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).

[28] 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).

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

[30] 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).

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

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

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

[34] Dekker, J. *et al.* The 4d nucleome project. *Nature* **549**, 219–226 (2017).

[35] Crane, E. *et al.* Condensin-driven remodelling of x chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).

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

[37] 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).

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

[39] Davis, C. A. *et al.* The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research* **46**, D794–D801 (2018).

[40] Consortium, G. *et al.* The gtex consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

[41] 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).

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

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

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

[45] Cui, H. *et al.* scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods* 1–11 (2024).

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

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

[48] 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).

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

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

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

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

[53] 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).

[54] Kim, J., Kim, D. & Sohn, K.-A. Hig2vec: hierarchical representations of gene ontology and genes in the poincaré ball. *Bioinformatics* **37**, 2971–2980 (2021).

[55] MacAlpine, D. M. & Almouzni, G. Chromatin and dna replication. *Cold Spring Harbor perspectives in biology* **5**, a010207 (2013).

[56] 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 (2011).

[57] 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).

[58] Pearce, J. D. *et al.* A cross-species generative cell atlas across 1.5 billion years of evolution: The transcriptformer single-cell model. *bioRxiv* 2025–04 (2025).

[59] Long, H. S. *et al.* Making sense of the linear genome, gene function and tads. *bioRxiv* (2020).

[60] Youngblut, N. D. *et al.* scbasecount: an ai agent-curated, uniformly processed, and continually expanding single cell data repository. *bioRxiv* 2025–02 (2025).

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

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

[63] 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).

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

[65] 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).

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

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

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

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

[70] Nadig, A. *et al.* Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics* 1–10 (2025).

[71] Chen, E. Y. *et al.* Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics* **14**, 1–14 (2013).

[72] 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).

[73] Soutourina, J. Transcription regulation by the mediator complex. *Nature reviews Molecular cell biology* **19**, 262–274 (2018).

[74] Allen, B. L. & Taatjes, D. J. The mediator complex: a central integrator of transcription. *Nature reviews Molecular cell biology* **16**, 155–166 (2015).

[75] Davidson, I. F. *et al.* Dna loop extrusion by human cohesin. *Science* **366**, 1338–1345 (2019).

[76] Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J. & Yu, H. Human cohesin compacts dna by loop extrusion. *Science* **366**, 1345–1349 (2019).

[77] Vian, L. *et al.* The energetics and physiological impact of cohesin extrusion. *Cell* **173**, 1165–1178 (2018).

[78] Samata, M. *et al.* Intergenerationally maintained histone h4 lysine 16 acetylation is instructive for future gene activation. *Cell* **182**, 127–144 (2020).

[79] 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).

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

[81] 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).

[82] 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).

[83] 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 (2021).

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

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

[86] 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).

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

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

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

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

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

[92] Consortium*, T. T. S. *et al.* The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).

[93] Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology* **19**, A68 (2015).

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

[95] 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).

[96] 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).

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

[98] Gui, T. *et al.* Targeted perturbation of signaling-driven condensates. *Molecular Cell* **83**, 4141–4157 (2023).

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

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

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

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

[103] Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3d genome folding from dna sequence with akita. *Nature methods* **17**, 1111–1117 (2020).

[104] Xiong, K., Zhang, R. & Ma, J. scghost: Identifying single-cell 3d genome subcompartments. *Nature methods* 1–9 (2024).

[105] Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell hi-c analysis with higashi. *Nature biotechnology* **40**, 254–261 (2022).

[106] Gao, V. R. *et al.* Chromafold predicts the 3d contact map from single-cell chromatin accessibility. *Nature Communications* **15**, 9432 (2024).

[107] 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).

[108] Zhou, T. *et al.* Gage-seq concurrently profiles multiscale 3d genome organization and gene expression in single cells. *Nature Genetics* 1–11 (2024).

[109] Arrastia, M. V. *et al.* Single-cell measurement of higher-order 3d genome organization with scsprite. *Nature biotechnology* **40**, 64–73 (2022).

[110] Liu, Z. *et al.* Linking genome structures to functions by simultaneous single-cell hi-c and rna-seq. *Science* **380**, 1070–1076 (2023).

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

[112] 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).

[113] Lin, X. *et al.* Nested epistasis enhancer networks for robust genome regulation. *Science* **377**, 1077–1085 (2022).

[114] 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).

[115] Brackley, C. A. *et al.* Polymer physics predicts the effects of structural variants on chromatin architecture. *Nature Genetics* **50**, 662–667 (2018).

[116] Bonev, B. *et al.* Multiscale 3d genome rewiring during mouse neural development. *Cell* **171**, 557–572 (2017).

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

[118] Muzellec, B., Telenczuk, M., Cabeli, V. & Andreux, M. Pydeseq2: a python package for bulk rna-seq differential expression analysis. *Bioinformatics* (2023).

[119] 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).

[120] 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).

[121] Flavahan, W. A. *et al.* Altered chromosomal topology drives oncogenic programs in sdh-deficient gists. *Nature* **575**, 229–233 (2019).

[122] 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).

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

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

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

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

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

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

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

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

[131] 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).

[132] 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).

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

Figures

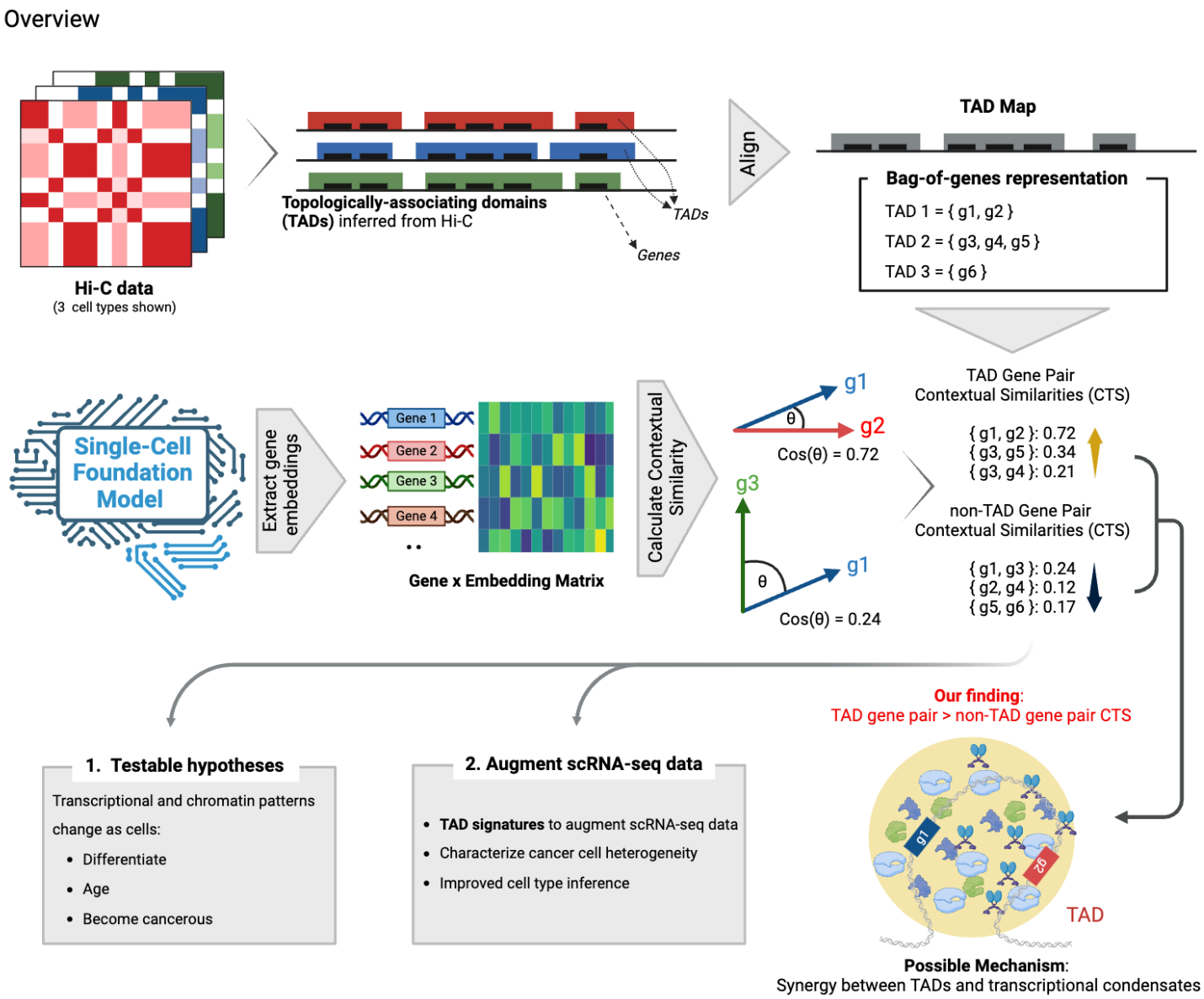

Figure 1: **Overview.** Topologically associating domains (TADs) are a key folding unit of the chromatin. While TADs are currently inferred separately for each cell type, TAD architectures have been observed to have good agreement across cell types. We find this agreement to be even stronger when TADs 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).

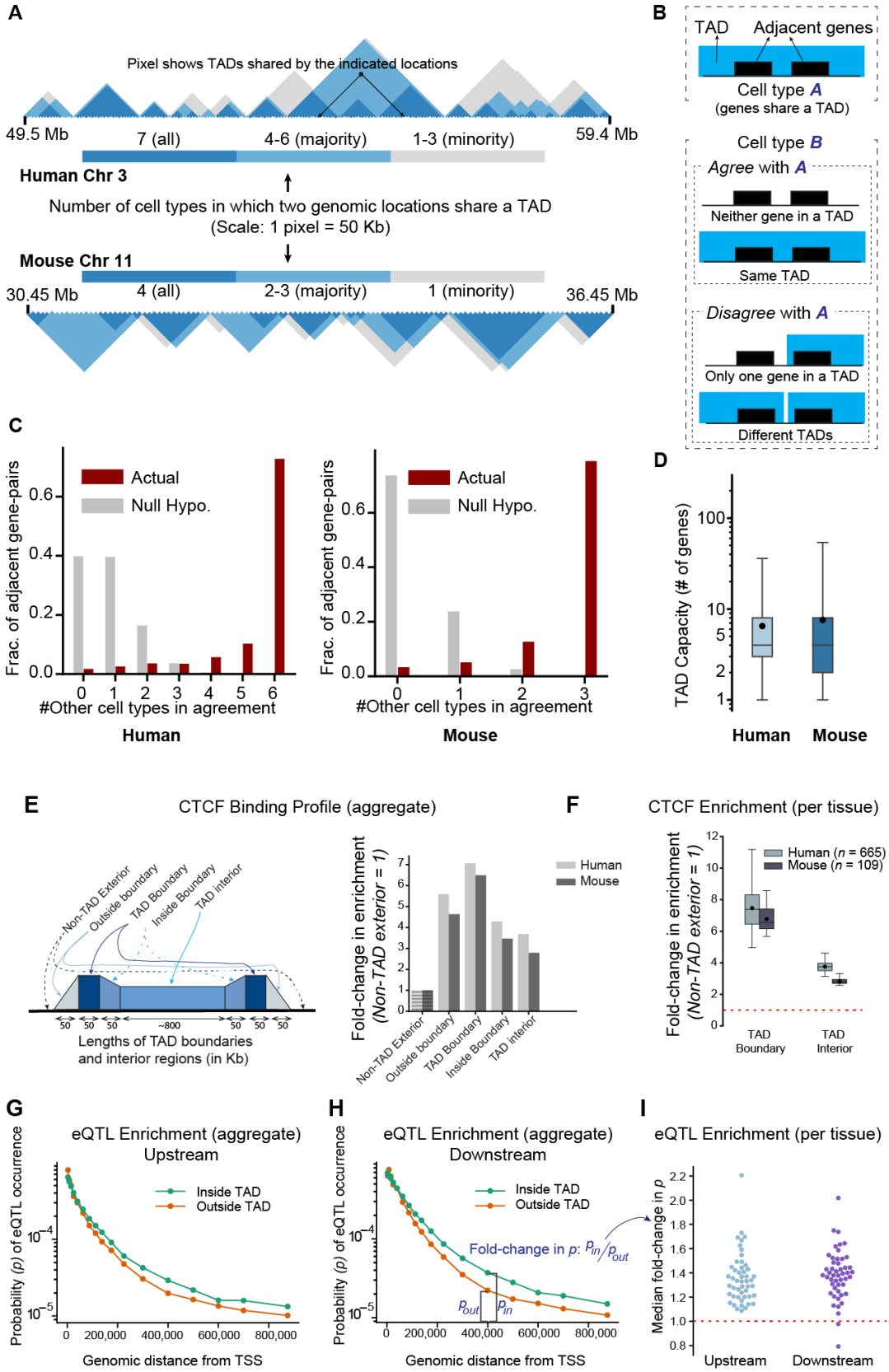

Figure 2: **Agreement between cell-type specific TAD architectures and the constructed consensus TAD scaffold.**

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.

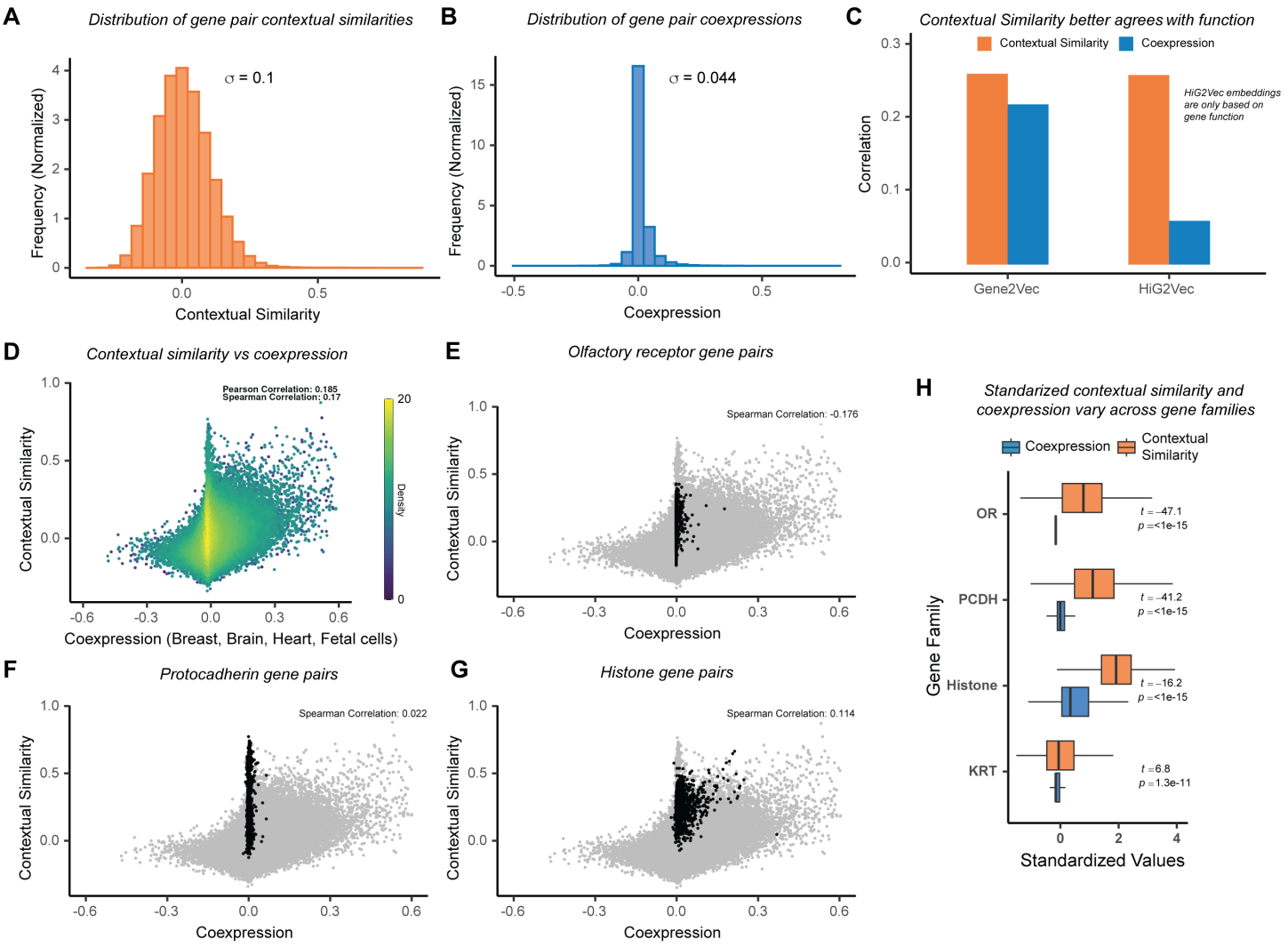

Figure 3: **Contextual similarity offers a more holistic perspective in evaluating gene pair transcriptional patterns. A, B**) Using Pearson correlations of pre-trained scGPT gene embeddings, we calculated contextual transcriptional similarities (CTS) for all gene pairs on the same chromosome within 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).

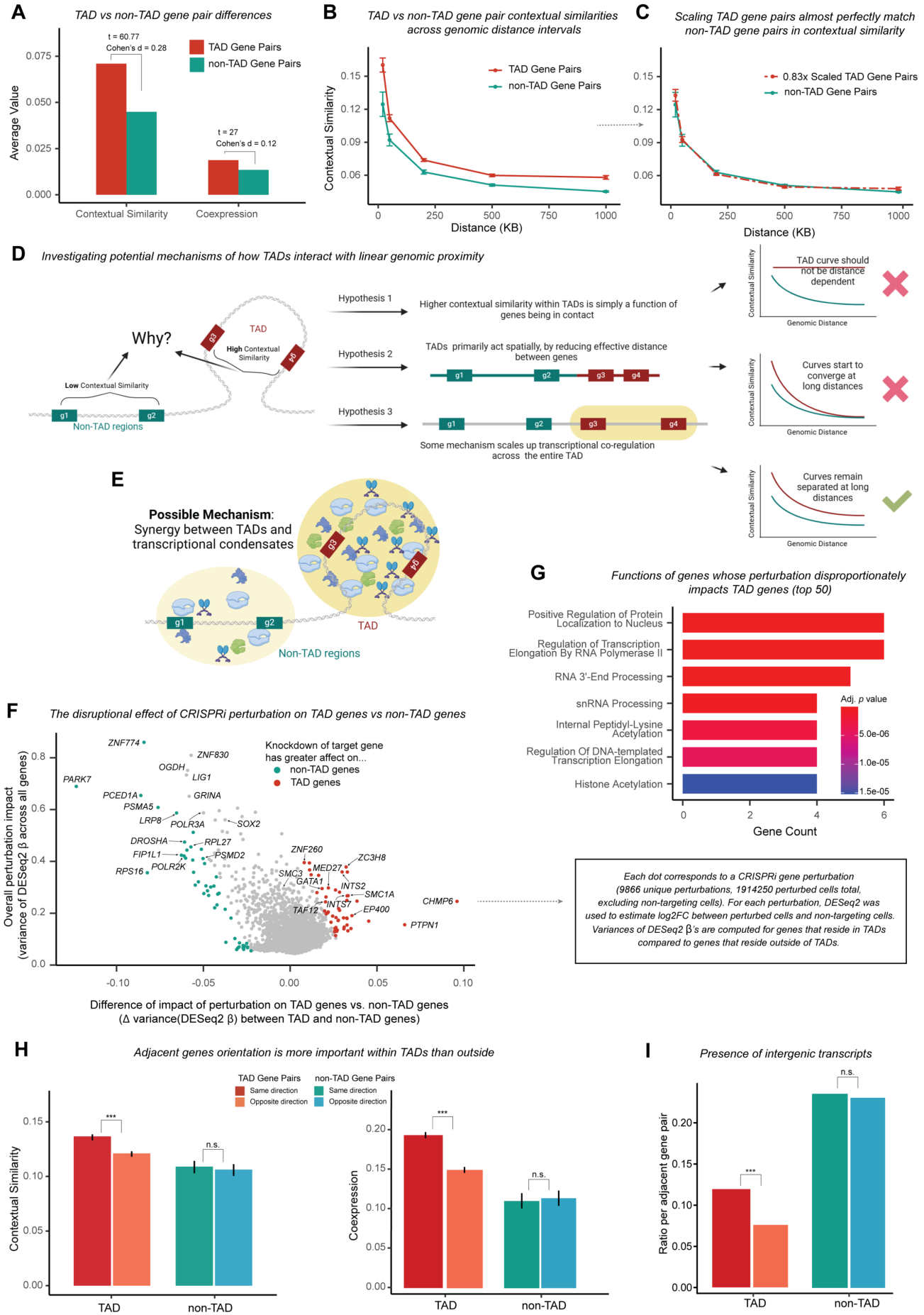

Figure 4: **Comparing CTS in TAD and non-TAD gene pairs provides clues into TADs' role in transcription.**

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[69]. 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.

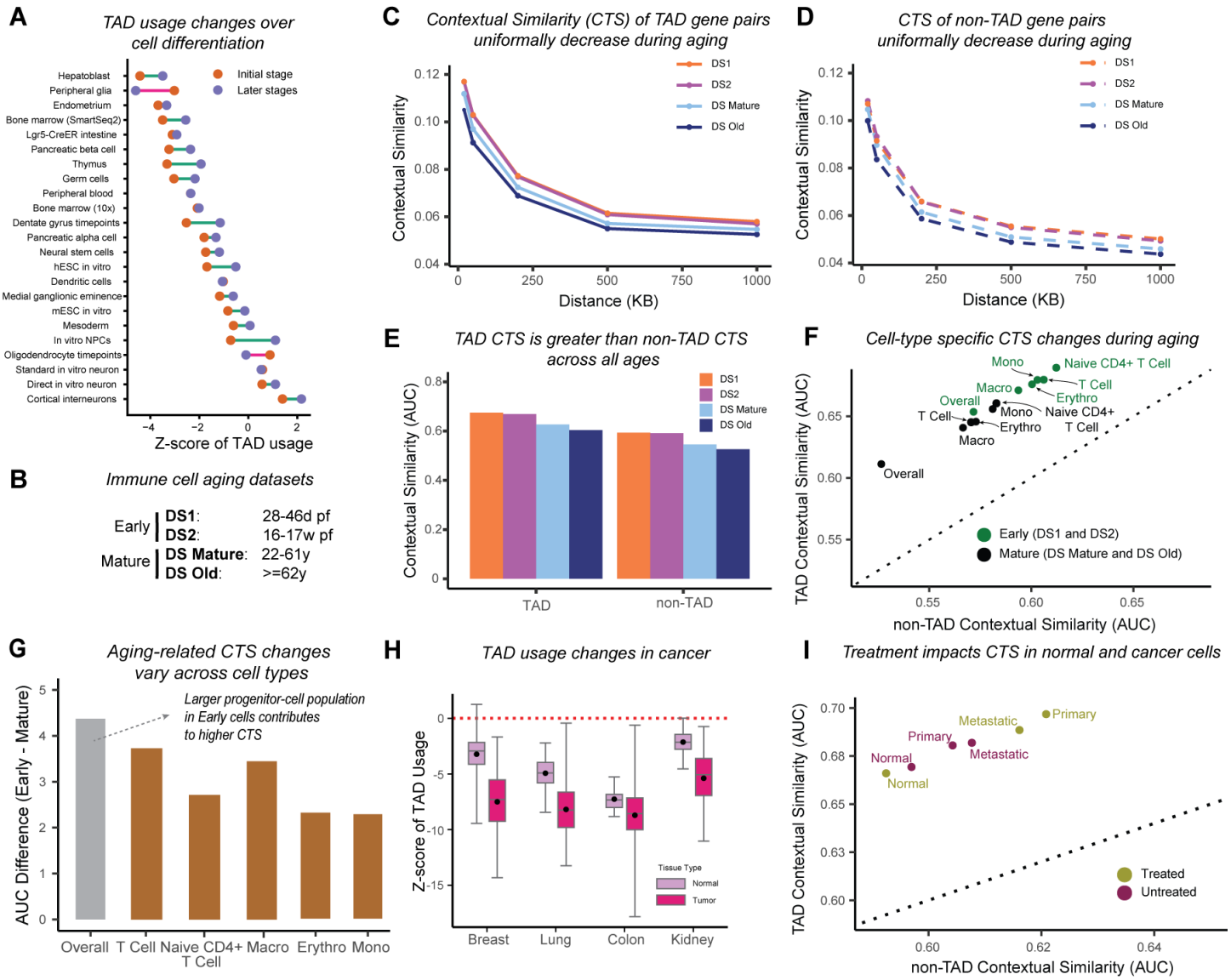

Figure 5: **TAD-mediated transcriptional organization changes systematically during development, aging, and cancer.** **A**) Analysis of 24 non-branching cell differentiation trajectories[87] reveals that early-stage cells consistently show higher TAD usage (more TADs with non-zero expression activity), indicating stronger transcriptional clustering (also see Fig. S4). **B**) To study aging effects, we analyzed immune cells from two large datasets[91,92], 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

Metric	Human DI-Tads	Human IS-Tads	Mouse DI-Tads	Mouse IS-Tads
Base-pair Jaccard Index	0.93	0.85	0.94	0.89
Overlapping TAD %	98%	97%	96%	97%
Reciprocal Best Match %	70%	65%	64%	63%

Figure S1: **TAD Maps with expanded cell types and TAD-calling algorithms strongly agree with originals**: Base-pair Jaccard Index measures nucleotide-level agreement between expanded and original TAD maps as the ratio of overlapping TAD base pairs to their union; TAD-level overlap reflects the fraction of expanded TADs overlapping any original TAD; reciprocal best matches indicates one-to-one correspondence using a greedy one-to-one matching scheme based on genomic distance, which enforces unique pairings between TADs.

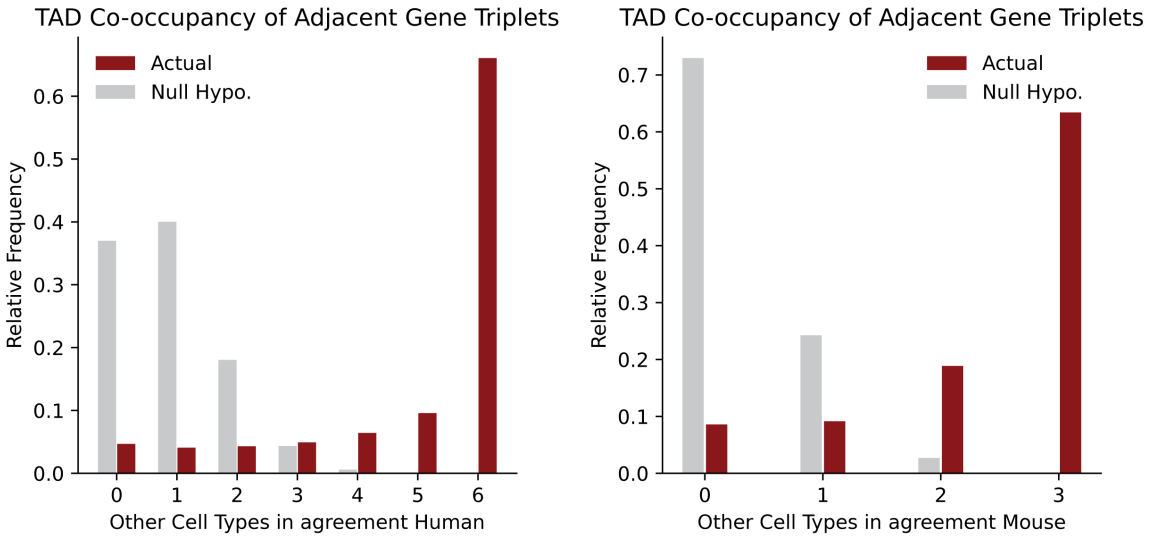

Figure S2: **Adjacent gene triplets yield consistent results with gene pairs:** Like pairs, triplets that co-occupy the same TAD across all cell types exceed expectations under a null model. This figure complements Fig. 2C.

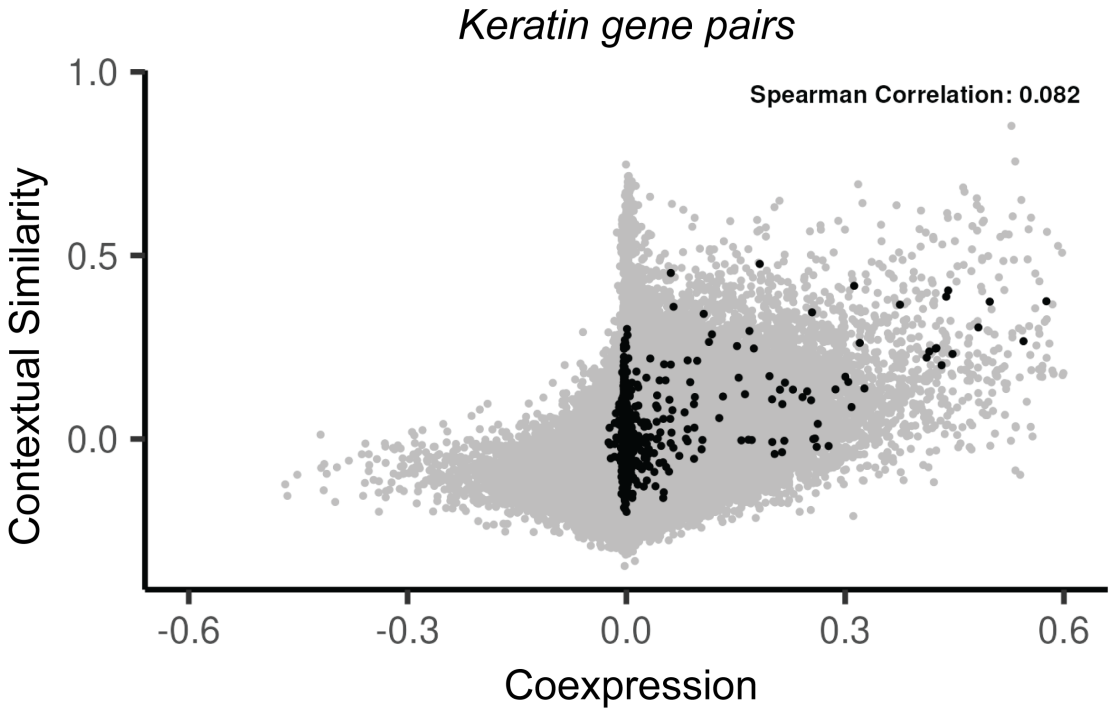

Figure S3: **Contextual similarities (CTS) vs co-expression for keratin genes pairs**: This figure complements Fig. 3E-G, where it visualizes CTS and co-expression for all gene pairs in the keratin family.

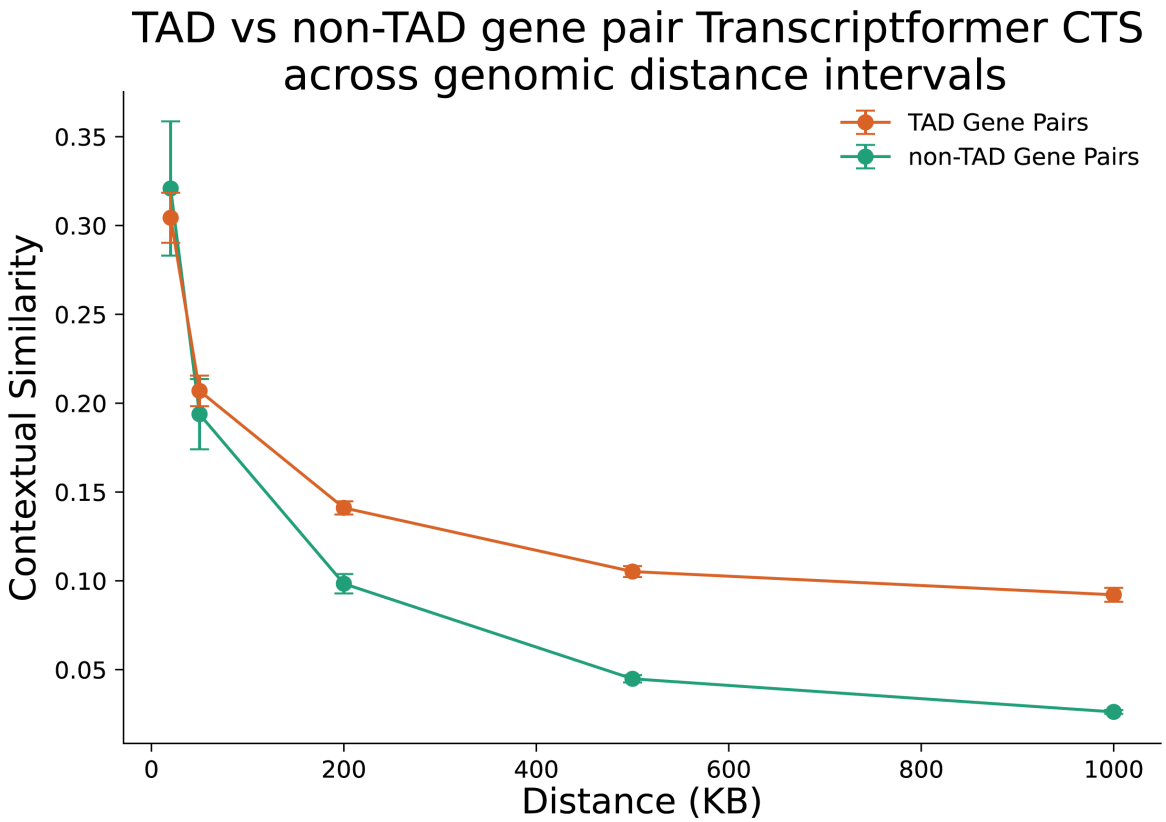

Figure S4: **Contextual similarities (CTS) derived from Transcriptformer show consistent, if not stronger, patterns compared to scGPT CTS across the same genomic distance intervals.** This figure complements Fig. 4B.

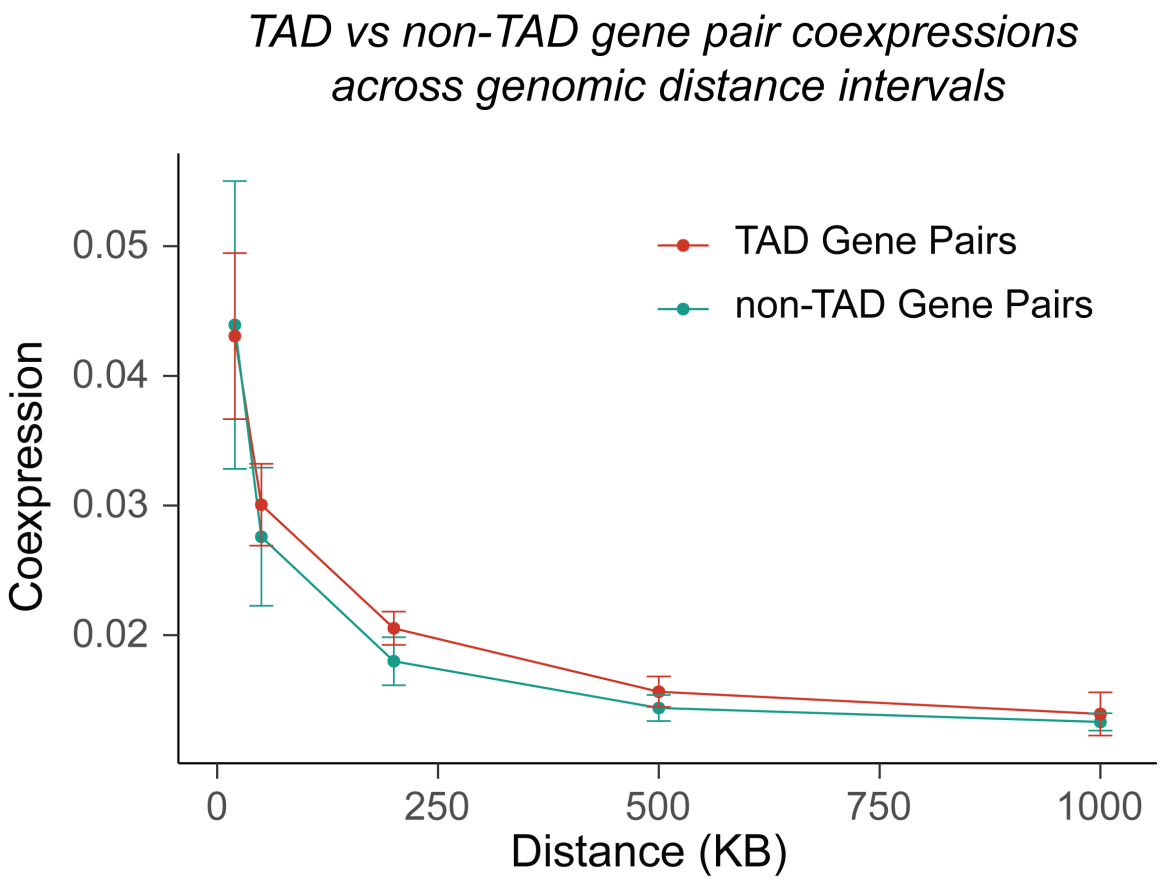

Figure S5: **Gene pair co-expression exhibits inconsistent patterns with respect to distance**: This figure contrasts the analysis with CTS in Fig. 4B, highlighting that average co-expression shows irregular and inconsistent trends between TAD and non-TAD gene pairs across distance intervals.

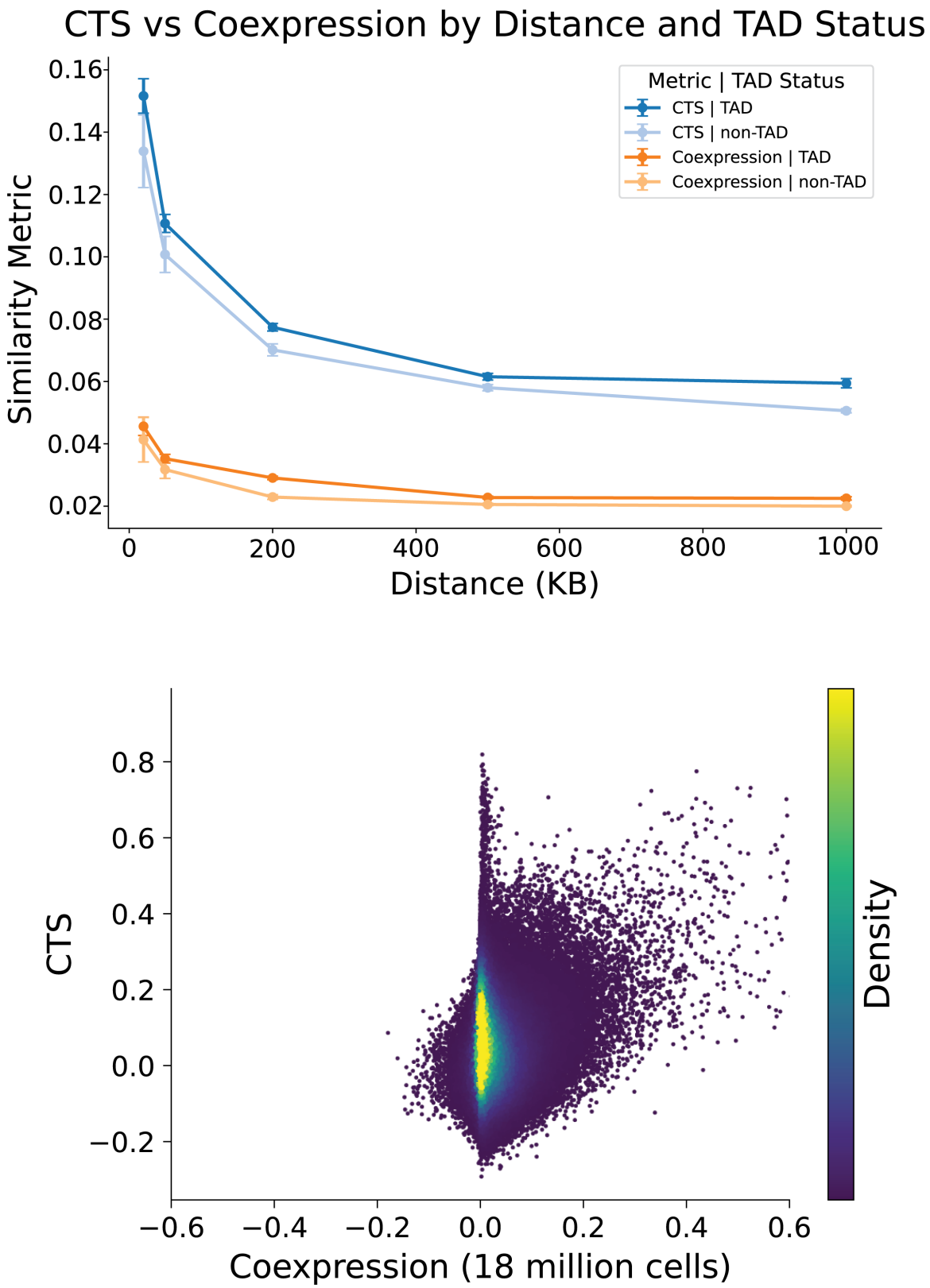

Figure S6: **Genomic distance–binned and unbinned co-expression analyses across 700 additional single-cell datasets from scBaseCount yield patterns consistent with the primary co-expression results** The top figure complements Fig. S5 and the bottom figure complements Fig. 3D, extending both analyses to 18 million cells across 40 tissues.

Gene perturbations that strongly disrupt TAD genes and their significant Gene Ontology terms

GO Term (Name)	GO ID	Perturbation Associated
mRNA Splicing, via Spliceosome	GO:0000398	RBM17, DBR1, PPWD1, CDC5L, PLRG1, LSM4, SRRM1, SNRPD2, ZMAT2, MFAP1, SNRPG, DHX15, SNIP1, SRSF3, SNRNP200
RNA Splicing, via Transesterification Reactions With Bulged Adenosine as Nucleophile	GO:0000377	DBR1, PPWD1, CDC5L, PLRG1, LSM4, SRRM1, SNRPD2, ZMAT2, MFAP1, SNRPG, DHX15, SNIP1, SNRNP200
mRNA Processing	GO:0006397	DBR1, PPWD1, CDC5L, PLRG1, LSM4, SRRM1, SNRPD2, ZMAT2, MFAP1, SNRPG, DHX15, SNIP1, SNRNP200

Gene perturbations that strongly disrupt non-TAD genes and their significant Gene Ontology terms

GO Term (Name)	GO ID	Genes Involved
Ribosome Biogenesis	GO:0042254	LTV1, UTP6, RPS19BP1, DIMT1, PNO1, PAK1IP1, WDR3, FCF1, NOL7, DNTTIP2, NOL6, LAS1L, PES1, RIOK2, EXOSC3
rRNA Processing	GO:0006364	TRMT112, LAS1L, WDR3, PES1, NOL7, RIOK2, EXOSC3, NOL6
Maturation of SSU-rRNA	GO:0030490	TRMT112, UTP6, WDR3, NOL7, RIOK2
Maturation of LSU-rRNA	GO:0000470	TRMT112, LAS1L, PAK1IP1, PES1

Figure S7: **Extended Perturb-seq analysis with additional datasets to assess perturbed genes that most disrupt transcription in TAD and non-TAD regions:** The top panel mirrors the original analysis, highlighting genes involved in transcriptional regulation and mRNA processing (complementing Fig. 4G), while the bottom panel similarly highlights genes associated with ribosomal components and rRNA processing (complementing Fig. S8D).

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)

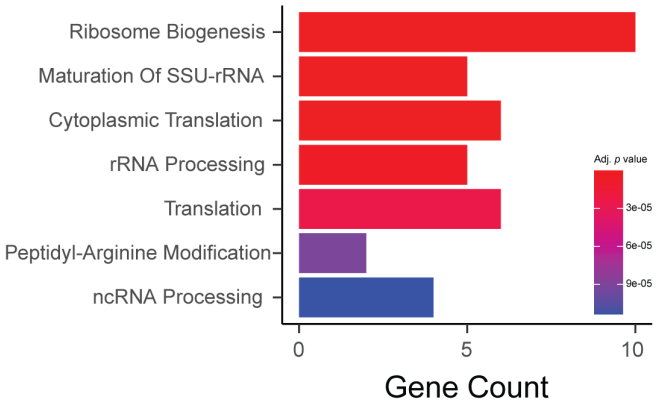

Figure S8: **List of specific genes whose perturbation induces transcriptional disruption across TAD and non-TAD regions: A**) This gene group highlights the genes whose perturbation causes the most widespread transcriptional disruption overall, without distinguishing between TAD and non-TAD regions. **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.

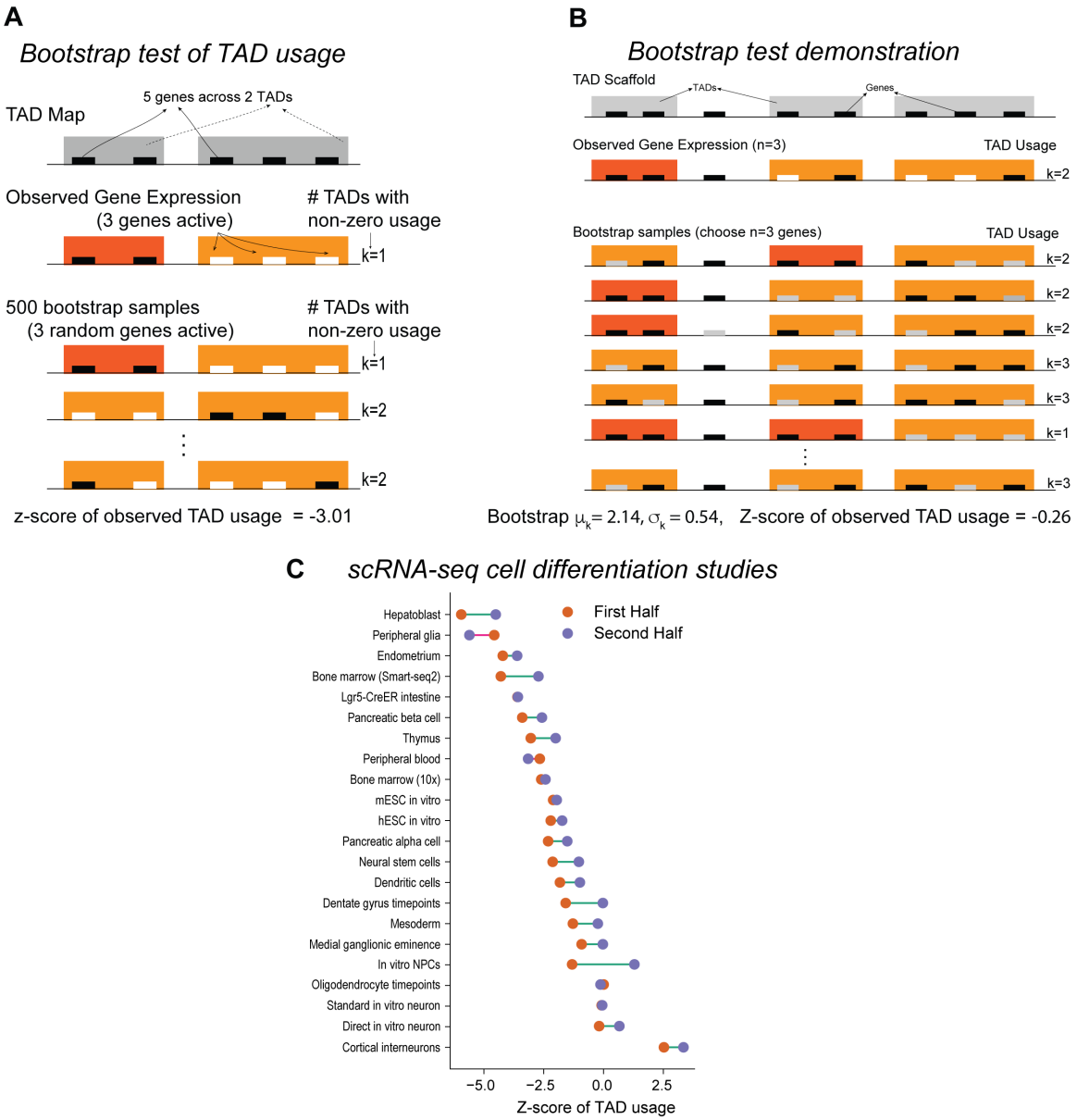

Figure S9: **Bootstrap test to assess gene clustering in TADs**: **A**) We define a TAD's *usage* as the number of transcriptionally active genes within it. Our bootstrap test assumes gene activity is independent of TADs under the null hypothesis. For any transcriptional profile, we randomize gene activity 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[?]. Here, cells are partitioned into two groups with equal sizes. Regardless of partitioning, the clustering trend remains consistent.

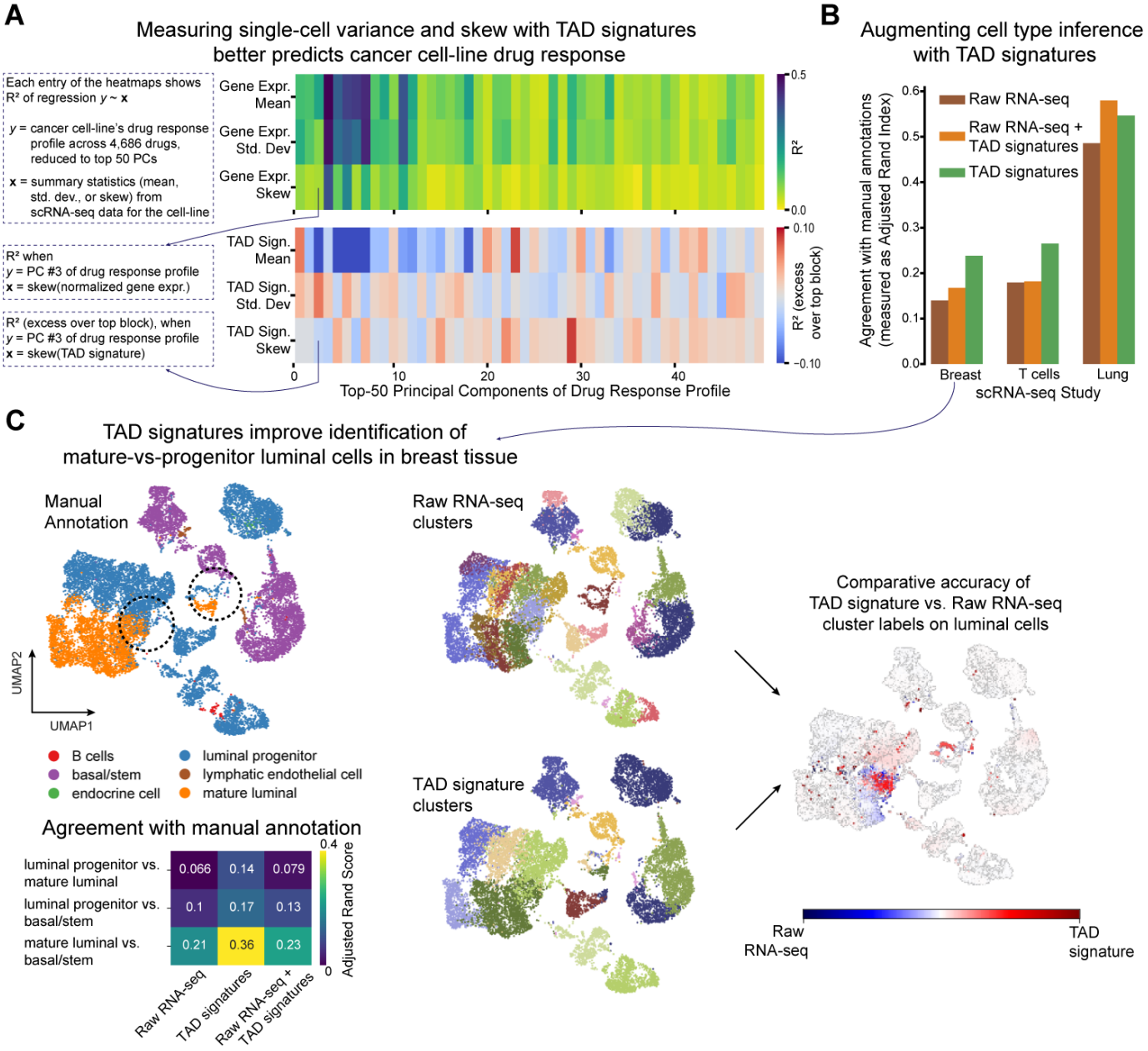

Figure S10: **TAD signatures augment standard scRNA-seq analysis**:**A**) TAD signatures reveal biologically meaningful single-cell heterogeneity. For 193 cancer cell lines, we analyzed drug response profiles from the PRISM study[128], reduced to 50 principal components. We computed the first three moments (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.